SESUG 2021 Paper 16-2021

Missing Value Imputation

Ross Bettinger, Silver Spring, MD

Abstract

Survey data are typically collected from people who respond to a set of questions presented to them by an interviewer or in response to questions on a website. Surveys are vulnerable to missing data due to nonresponse to specific items on the survey questionnaire. The process of replacing a missing value with a substituted value is called imputation, and many techniques have been developed to "plug the hole in the data" with an imputed value. We briefly summarize a number of imputation techniques and present an example of imputation using the fuzzy c-means clustering algorithm.

Keywords

Bias, census, clustering, complete case, enumeration, fuzzy c-means, imputation, Little's MCAR test, MAR, MCAR, missing data mechanism, missing value, MNAR, nonignorable, SAS® PROC IML, PROC MI, survey questionnaire

Introduction

We observe a process and collect measurements on it. If the process is a laboratory experiment, we can be reasonably assured of obtaining observations for which every variable measured is accurately recorded because the process is under our direct control. If the process is not controllable, there will be a greater likelihood of recording incomplete observations for which some variables are represented by noncomputational values that serve as placeholders. When we have observations that include noncomputational missing values, we may impute representative values to be substituted for them so as to create complete cases for analysis. Imputation is often necessary to create complete cases when an analytic technique, e.g., regression, cannot process missing values, or when there may be too few complete observations in a sample to permit incomplete observations to be omitted from the analysis.

Since the imputation process is by definition an external intervention performed on missing data, the sample data will unavoidably be modified to some degree by our remediation, and the data will be biased as a result. A successful imputation effort will fill the gaps in the data with substituted data that are similarly distributed as existing data and which minimize bias due to introduction of an artifact [1].

Classification of Missing Data

If we have some understanding of the process underlying the creation of a variable's missing value, we may be able to make an informed imputation of an appropriate value. We may classify missing values into three categories, based on the mechanism of their creation [2, 3]. They are: 1) missing completely at random, 2) missing at random, and 3) missing not at random. "The missing data mechanism is simply the equation that expresses the probability of missingness as a function of Y and X" [4] where variable Y has missing data and variable X has no missing data.

Missing Completely at Random (MCAR)

Data that are MCAR have no relationship between missing and nonmissing values. They represent a random subset of the data and there is no *systematic* relationship that can be used to model or predict which values will be missing or nonmissing. The nonmissing data contain no information related to the missing data. The property of missingness is independent of observable values. If missing values can be accurately predicted using domain knowledge, regression, or some other method, then the missing values are not MCAR. A more formal procedure is to use Little's MCAR test [5].

Another approach to detecting MCAR missingness is to use indicator variables where

$$MissingValue = \begin{cases} 1\ if\ variable\ value\ is\ \qquad missing \\ 0\ if\ variable\ value\ is\ not\ missing \end{cases} \tag{1}$$

and then performing a χ^2 test on the hypothesis that the missing data are MCAR. For example, let us assume that we observe missing values for questions related to income in an online webpage survey and we want to know if respondents are sensitive about indicating their income. Perhaps we have also asked their ages earlier in the survey. Then we can create indicators such as MV_AGE and MV_INCOME as described in Equation 1. Let us state the null hypothesis: there is no causal relationship between missing values of age and missing values of income, e.g., the missingness of income given age is MCAR. If we observe that the frequencies of (MV_AGE, MV_INCOME) are disproportionately (1, 1) or (0, 0) then the χ^2 statistic will be significant at, say, the $p < .05$ level, and there is adequate evidence to reject the null hypothesis that the data are MCAR. In this case, we might attempt to build a regression model predicting income from age and other predictors to impute representative values of income.

Structurally Missing Data (Missing by Design)

Data that are structurally missing are absent because they do not exist according to the survey item definition. For example, a family that has no children cannot answer the question, "What is the age in years of the youngest child in your family?" To enter a value of 0 is not correct because the answer implies that there is a child less than a year old when there is no child, so a placeholder missing value is used to indicate nonresponse.

Missing at Random (MAR)

Paul Allison describes MAR involving one variable as follows:

> "Data on *Y* are said to be missing at random if the probability of missing data on *Y* is unrelated to the value of *Y*, after controlling for other variables in the analysis. To express this more formally, suppose there are only two variables *X* and *Y*, where *X* always is observed and *Y* sometimes is missing. MAR means that
>
> $$P(Y\ is\ missing \mid X, Y) = P(Y\ is\ missing \mid X) \tag{2}$$
>
> In words, this expression means that the conditional probability of missing data on *Y*, given both *Y* and *X*, is equal to the probability of missing data on *Y* given *X* alone.[1]
>
> Suppose that *Y* is body weight and *X* is gender. Then if women are less likely to disclose their weight, *Y* can be MAR because the probability of missing *Y* depends on *X*, which is observed. But what if both men and women are less likely to disclose their weight *if they're overweight*? Then values are not MAR, since the probability *Y* is missing depends on *Y* net of *X*" [6].

[1] Thus, for two variables X and Y where only Y has missing values, the missingness of Y is completely independent of any value of Y. If the variables are AGE and INCOME in a particular survey, where AGE is always observed but INCOME has missing values, the probability of a missing value of INCOME depends only on the reported value of AGE, after conditioning the data on AGE.

He describes the case when two or more variables have missing data, citing Example 1.13 in Little and Rubin [7], which we paraphrase below:

> Let us consider the simplest case involving two variables X and Y. Each may have missing data, so there are four patterns to the missingness.
>
> 1. When X and Y are both observed, the probability of observing both X and Y may depend on the values of both X and Y.
>
>   Then $P(observe\ Y\ and\ X \mid X, Y) = P(X, Y)$.
>
> 2. When X is observed but Y is missing, the probability of observing X but not Y may depend on X but may not depend on Y.
>
>   Then $P(Y\ is\ missing \mid X, Y) = P(Y\ is\ missing \mid X)$.
>
> 3. When X is missing but Y is observed, the probability of observing Y but not X may depend on Y but may not depend on X.
>
>   Then $P(X\ is\ missing \mid X, Y) = P(X\ is\ missing \mid Y)$.
>
> 4. When X and Y are both missing, the probability that both X and Y are missing may not depend on either X or Y.
>
>   Then $P(X\ is\ missing\ and\ Y\ is\ missing \mid X, Y) = P(X\ is\ missing\ and\ Y\ is\ missing)$.

The significant concept involved with the MAR mechanism is that it is a statement about the probability of observing a particular pattern of missingness among variables and is not related to the probability of individual variables containing missing values.

The missing data mechanism is called *ignorable* if the data are MAR and the parameters of the model describing the missing data mechanism are distinct from the parameters in the model to be estimated [7]. "As the name suggests, if the missing-data mechanism is ignorable, then it is possible to get valid, optimal estimates of parameters without directly modeling the missing-data mechanism" [8].

Missing Not at Random (Nonignorable)

If the probability that variable Y contains missing values depends on Y itself after controlling for X, then the MAR assumption is violated and the missing values are not missing at random (MNAR). In this case, the missing data mechanism may not be ignored in the estimation process if the parameter estimates are to be unbiased. In this case, the data contain no information for testing the MNAR mechanism, and there is no way to ensure that a particular missing data mechanism is the correct model.

Criteria for Evaluating an Imputation Method

There is general agreement that a good imputation method ought to have the following characteristics [8]:

1. While substituting artefactual values for missing values may introduce bias into parameter estimates, a good method will minimize this induced bias.
2. A good method will maximize the use of available information. Methods that discard observations that contain missing values throw away information about the process being surveyed. We need to use all of the available data to produce efficient parameter estimates that have minimum-sampling variability.

3. A good method will generate good estimates of uncertainty, e.g., accurate estimates of standard errors, confidence intervals, and *p*-values.

In the sequel, evaluation criteria are reported from [8] unless otherwise indicated.

Brief Review of Missing Data Handling and Imputation Methods

There are many missing data handling and imputation methods discussed in the literature. Many are only of historical interest due to advances in statistical science. We include references to descriptions of these methods.

Listwise Deletion (Complete Case Analysis) [8]

The listwise deletion method for missing data simply deletes all observations containing one or more missing values so that only complete cases are used. This method is straightforward to implement and can be used with any statistical method. If the number of incomplete observations is small relative to the size of the sample, “listwise deletion is robust to violations of MCAR or MAR for predictor variables in a regression analysis” [8]. However, if many observations are discarded, standard errors may be larger, confidence intervals will be wider, and there will be a loss of statistical power in testing hypotheses. See [8] for more details.

Evaluation criteria rating of imputation method performance: 1-“so-so”, 2- “terrible”, 3-good.

Pairwise Deletion (Available Case Analysis) [8]

The “minimal sufficient statistics” for a wide class of linear models are the means, variances, and covariances. If the parameters of interest of a linear model can be expressed as functions of the minimal sufficient statistics, then using all of the available data for each variable or pair of variables may be more efficient than listwise deletion because more data are utilized. Once the sample moments have been computed, their values are substituted into the formulas for the population parameters and all observations are used.

Evaluation criteria rating: 1-good, 2-good, 3-poor because the sample size for each sample moment may vary widely.

Single Imputation [9]

The single imputation method replaces a missing value by a value derived from the sample data. This replacement operation produces complete cases so that the number of observations is preserved. However, imputation produces lower estimates of standard errors, which bias test statistics upward and *p*-values downward. It is unrealistic to assume that the replacement values are generated by the same process as the original data, and computer software cannot distinguish between actual data and imputed data so as to compensate for the bias introduced by imputation. Also, the larger the percent of missing data in the sample, the more pronounced this bias will be.

Unconditional Mean Imputation

The unconditional mean imputation method replaces missing values of a specified variable with the mean of the available cases for that variable. Thus, the sample mean of the variable is unchanged. However, the variance of the variable is underestimated because the presumably different missing values have been replaced with a single value. Parameter estimates involving variances will be biased upward. The magnitudes of correlations and covariances are decreased because the variability in the sample data are diminished.

Evaluation criteria rating: 1-poor, 2-good, 3-poor.

Regression/Conditional Mean Imputation

The regression/conditional mean imputation method uses the information in complete observations to predict missing values. It is essentially a linear regression of Y on **X** where Y is the variable containing missing values and **X** is a data matrix of predictors. Missing values of Y are predicted using the complete case data in **X**. The imputed values do not have an error term included in the model, so they fit the regression line exactly. Such a nice result belies the fact that real data are never so well-behaved. The variance of Y is reduced, and parameter estimates are biased upward. Regression/conditional mean imputation overestimates correlations between Y and the various predictors because the error term is omitted in the predicted missing values.

Evaluation criteria rating: 1-poor, 2-good, 3-poor

Stochastic Regression Imputation

Stochastic regression imputation is an improvement over regression/conditional mean imputation in that a residual term is added to each predicted value of the missing variable. This residual term is drawn from a normal distribution with a mean of 0 and a variance equal to the residual variance of the regression of the variable with missing values on the predictors. Adding the residual term preserves the variability of the sample data and creates unbiased parameter estimates with MAR data. However, stochastic regression imputation may produce implausible values, e.g., negative income when income is always a nonnegative quantity. Also, if the data are heteroscedastic, the assumption of constant variance is violated. Standard errors will be underestimated because the residual error associated with the imputed values is not included in the residual variance.

Evaluation criteria rating: 1-good, 2-good, 3-poor

Hot-deck Imputation

The term “hot-deck imputation” is derived from the early days of statistical computing when all data values were recorded onto punched cards (the “deck”) prior to analysis. A variable’s missing value is imputed from the value of a similar complete observation that has been selected from the “deck” representing the sample data. Methodologies for selecting the complete observation that is similar to the observation containing the missing value are described at length in [10]. If the sample size of the distribution of the nonmissing values of a particular variable is small, then the values imputed from the nonmissing data will be limited in diversity and the variable’s sample variance will be underestimated. The measures of uncertainty will be inaccurate: standard errors will be too small, confidence intervals will be too short, and *p*-values will be too small.

Evaluation criteria rating: 1-fair, 2-good, 3-poor

Cold-deck Imputation

“Cold-deck imputation” refers to the selection of donor values from a prior dataset or source external to the survey data of interest. Complete observations in the prior preprocessed dataset are matched with observations in the survey data, and missing values in the latter dataset are replaced with nonmissing values in the former dataset. There is no randomness in this method, so variability of the missing data will be reduced

Evaluation criteria rating: 1-poor, 2-good, 3-poor

Last Observation Carried Forward (for Longitudinal Data)

The LOCF imputation method is relevant to surveys conducted over time in which nonresponses to survey items occur due to drop-outs of responders. If a surveyed person is no longer responding to surveys, the last known response is propagated forward in time as if the respondent were present. While this method imputes a response to missing values, the assumption that the respondent will answer identically for future items after the last recorded response is unrealistic. Substituting the value of the LOCF into missing values creates complete cases but reduces sample variability and introduces bias in the data.

Evaluation criteria rating: 1-poor, 2-fair, 3-poor

Multiple Imputation

A characteristic problem with imputation is the nonrandomness of the imputed values. Nonmissing data are generated by some process that includes intrinsic randomness and it is this natural variation that is absent in imputation algorithms. Stochastic regression imputation includes a random component based on the variance of a variable's nonmissing values but this random error is added *after* missing values have been predicted by a regression equation. The imputed values are estimates and have their own errors that may be drawn from distributions different from those of the nonmissing values. For example, the random error applied to missing values in stochastic regression imputation is created from a normal distribution but errors in estimating nonmissing values may not be normally distributed. Thus, bias may be added to imputed values. Because computer software cannot distinguish between nonmissing values and imputed values substituted for missing values, the estimates of uncertainty, e.g., sample means or parameter estimates, are underestimated so that the F and t-statistics will be too large and hence the p-values of significance will be misleadingly small. So, the likelihood of a Type I error will be increased.

Multiple imputation [8] is an extension of stochastic regression imputation where the distribution of complete case data is used to estimate multiple values that reflect the variation around the actual unobserved missing value of a variable. There are three steps in applying multiple imputation [11]:

1. Imputation: Estimated values are created from complete cases and substituted for the missing values in variables so that a complete-case dataset is produced. This process is repeated m times.
2. Analysis: A model is built using each of the m complete datasets to produce m sets of parameter estimates, e.g., coefficients and standard errors.
3. Pooling: The m sets of parameter estimates are combined for inference.

Reference [12] suggests that the following information be included in a publication or report in which multiple imputation was used:

1. Which statistical software program was used to conduct the imputation, e.g., SAS' PROC MI
2. The type of imputation algorithm used, i.e., multivariate normal (MVN) or fully conditional specification (FCS)
3. Some justification for choosing a particular imputation method
4. The number of imputed datasets (m) created
5. The proportion of missing observations for each imputed variable
6. The variables used in the imputation model and why so your audience will know if you used a more inclusive strategy. This is particularly important when using auxiliary variables.[2]

Reference [12] contains a detailed example of multiple imputation.

Evaluation criteria rating: 1-good, 2-good, 3-good

[2] "Auxiliary variables are variables in your dataset that are either correlated with a missing variable(s) or are believed to be associated with missingness." [12]

Clustering Imputation

The purpose of cluster analysis is to group a collection of data into homogeneous sets that are composed of observations that are maximally-related to other observations in the same set and minimally-related to observations in any other set. Observations within a cluster are more similar to each other than to observations in any other cluster. For example, a dataset of housing observations may contain information on geographic location, number of bedrooms, total number of rooms, median house value, median income of homeowner, &cetera. A clustering analysis would produce groupings of the data based on cluster centers that could be interpreted as prototypes of the characteristics measured, such as "low house value/far from urban center/low income/small house" or "high house value/suburban/high income/multi-generational house" or "high house value/central location/high income/single-family apartment".

"Hard" clustering algorithms produce sets of observations that are disjoint from one another in that an observation in cluster A cannot be a member of cluster B. These clusters are called "crisp". Alternatively, "fuzzy" clustering algorithms partition sets of observations into "fuzzy sets"[3] in which observations in cluster A are allowed to be members of cluster B, with a *degree of membership* in each cluster that represents the strength of similarity in each cluster. For example, winner-take-all elections allow each eligible participant one vote, to be used to select among two or more candidates. A vote cannot be split amongst several candidates, despite the voter's desires. Voting in this case is "crisp". But "fuzzy voting" would allow the voter to split the vote among, say three candidates A, B, and C: if the voter identifies with candidate A very strongly, then A might get 60% of the vote, B might get 30% of the vote, and C would get the remaining 10% of the vote. The winning candidate would accrue the highest sum over all the percentages assigned by voters.

One commonly-used crisp clustering algorithm is k-means clustering. The algorithm iteratively 1) assigns data points to k clusters, 2) computes the centroid of each cluster, and 3) reassigns points to their nearest cluster centroid. This process is repeated until there is convergence of values of an objective function, e.g., within-cluster sum of squares, or an iteration limit is reached. The number of clusters is a parameter that is varied between clustering runs. Many heuristics have been developed to suggest the optimal number of clusters.

An imputation method for k-means clustering might be to use complete cases to compute centroids, then perform nearest-neighbor clustering to find neighbors to the incomplete observations. The missing value(s) would then be computed from weighted combinations of nearest-neighbor complete cases and center of the cluster nearest to the incomplete case.

Fuzzy c-means clustering [11, 12] is an extension of k-means clustering in that an observation can belong to more than one cluster at the same time. If there are two or more clusters that overlap, it may make more sense to assign an observation to more than one cluster, with a membership value that reflects the similarity of the observation to others in each cluster, respectively. The membership value is a nonnegative real number ≤ 1. Fuzzy c -means clustering also requires specification of a fuzzification factor, m, which controls how fuzzy the cluster will be, i.e., the degree of overlap allowed between clusters. The higher the value of m, the greater the overlap possible. It, like the number of clusters, must be chosen through trial and error. Bezdek's algorithm was written in SAS/IML, and the SAS macro implementing it is be described in Appendix A.

[3] Lotfi A. Zadeh wrote the seminal paper on fuzzy set theory, *Fuzzy Sets*, Information and Control 8, 338-353 (1965).

Imputation of missing values using fuzzy c-means clustering means is performed in two phases. In phase 1, using complete-case data, centroids for specified c clusters and fuzzy parameter m are computed using Picard iteration [15]. In phase 2, all missing values are initialized with estimates computed from membership data and centroids computed in phase 1, and the clustering process is repeated with complete cases and estimates. There is an implicit assumption that the estimates of missing values are representative of their actual nonmissing values [12].

Evaluation criteria rating for fuzzy c-means clustering imputation[4]: 1-good, 2-good, 3-good

Case Study of Fuzzy c-Means Clustering Imputation

We demonstrate the use of fuzzy c-means clustering to perform multiple imputation of missing values of variables for California assembly districts from the 1990 US Census[5].

Data

There were 20,640 observations, each consisting of 10 variables. Each observation represents one census block, the smallest geographic unit used by the United States Census Bureau for tabulation of 100-percent data (data collected from all houses, rather than a sample of houses).

Census Attribute	**Description**
Ocean Proximity	Location of the house w.r.t. ocean/bay
Longitude	A measure of how far west a house is. A higher value is farther west.
Latitude	A measure of how far north a house is. A higher value is farther north.
Median Income	Median income for households within a block of houses (US Dollars)
Median House Value	Median house value for households within a block (measured in US Dollars)
Households	Total number of households. A household is a group of people residing within a home unit, for a block
Population	Total number of people residing within a block
Housing Median Age	Median age of a house within a block. A lower number is a newer building.
Total Rooms	Total number of rooms within a block
Total Bedrooms	Total number of bedrooms within a block

Table 1: Census Attributes

Census attributes represent location, wealth, population density, and house characteristics. All the attributes are numeric with the exception of Ocean Proximity, which is a character descriptor.

[4] It is our personal opinion that fuzzy c-means clustering imputation is superior to other methods discussed based on ease of use and performance according to the criteria for evaluating an imputation method.

[5] The data were downloaded from https://www.kaggle.com/abrahamanderson/geographical-analysis-regression-housing-prices, a Kaggle notebook created by Abraham Anderson.

Exploratory Data Analysis

Figure 1 shows the distribution of Median House Values by physical location, which is a key component of house value. The color scale shows lower median house prices in blue proceeding to moderate house prices in green and the highest-priced houses in red.

We see that houses located on the ocean or bay are most highly-valued, while median house values decline as they are farther from water. The urban centers of the San Francisco-Oakland Bay area, Santa Barbara, Los Angeles, and San Diego contain the highest concentration of population and, since demand for housing is high in these geographic regions, the median house values are correspondingly high. Catalina Island, off the coast of Santa Barbara, contains relatively few houses, but their median house values are high.

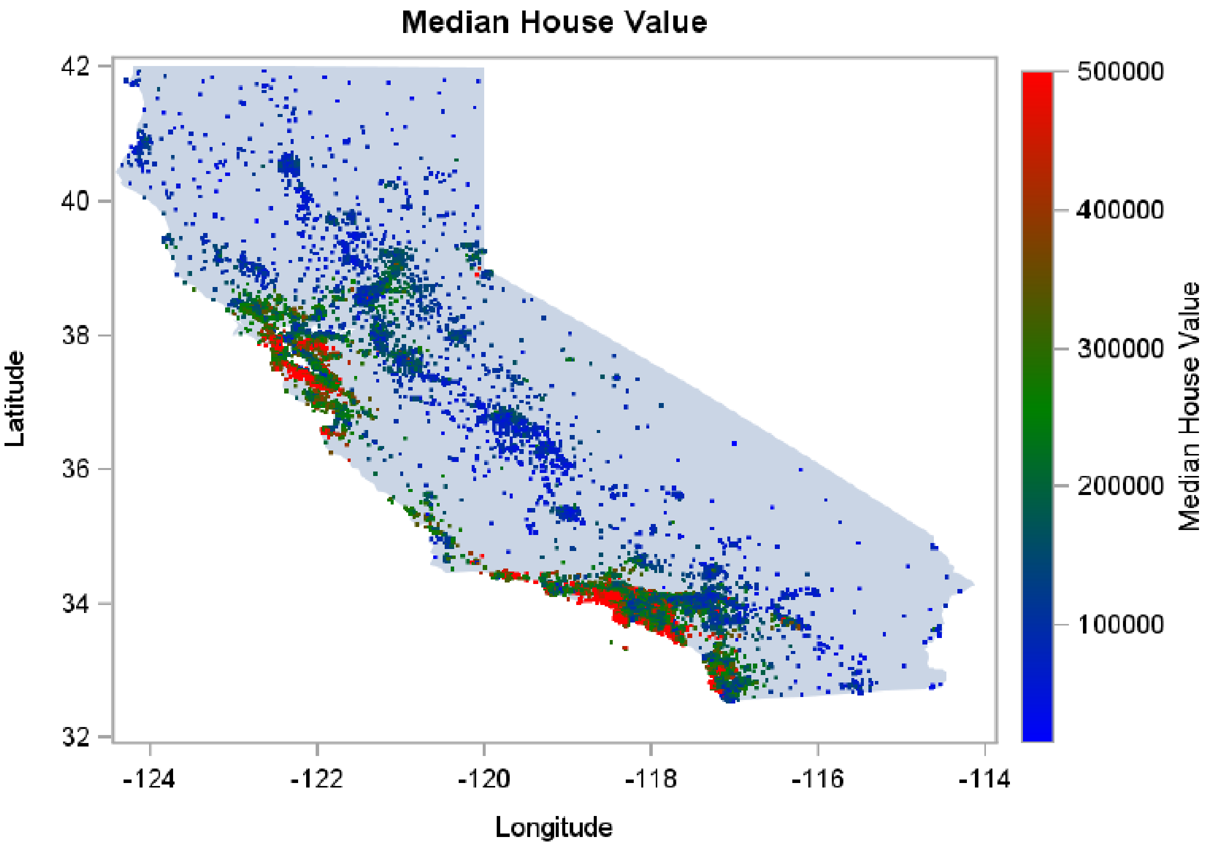


Figure 1: Median House Value by Physical Location

Figure 2 shows a breakdown of median house value by Census designation of physical location. It is evident that house values are strongly related to ocean proximity, with value inversely related to distance from water. Highest values are observed on coastal, bay, or island properties, and lower values are found on inland properties.

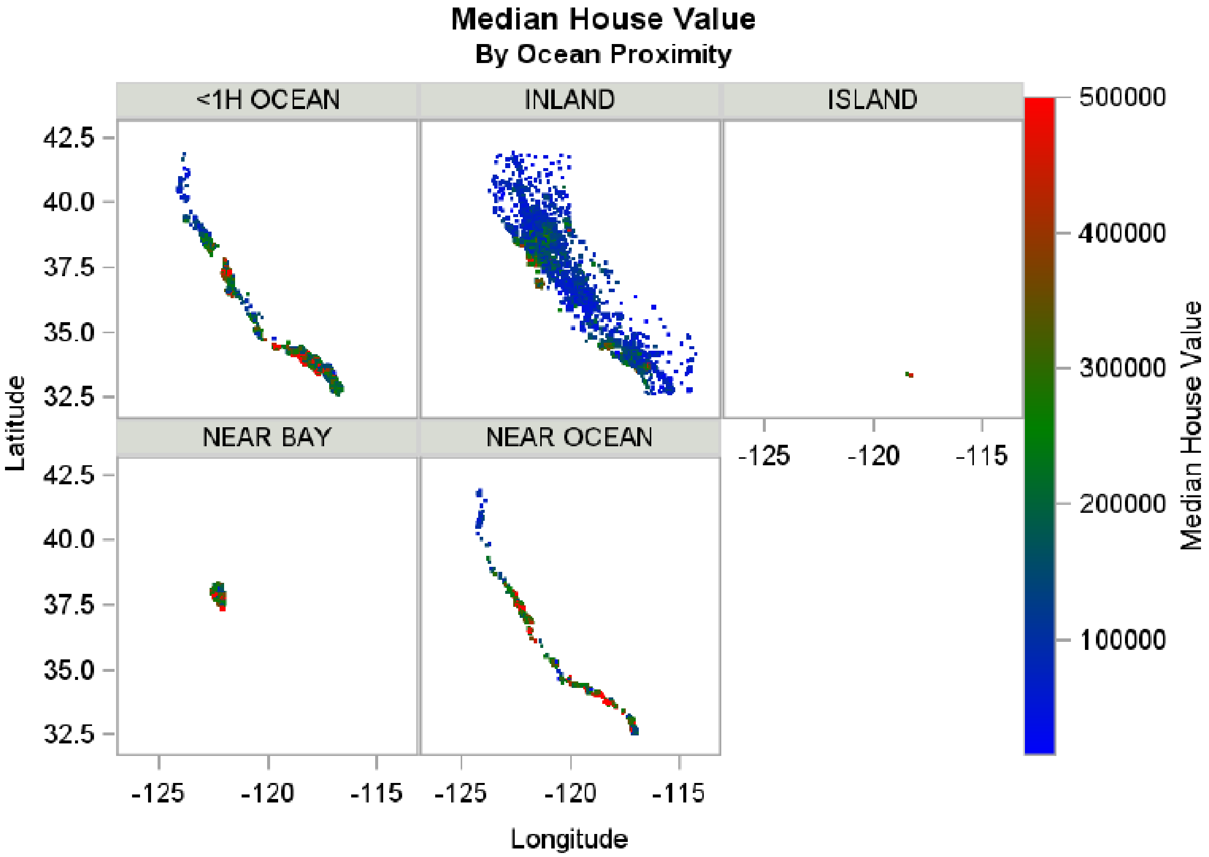
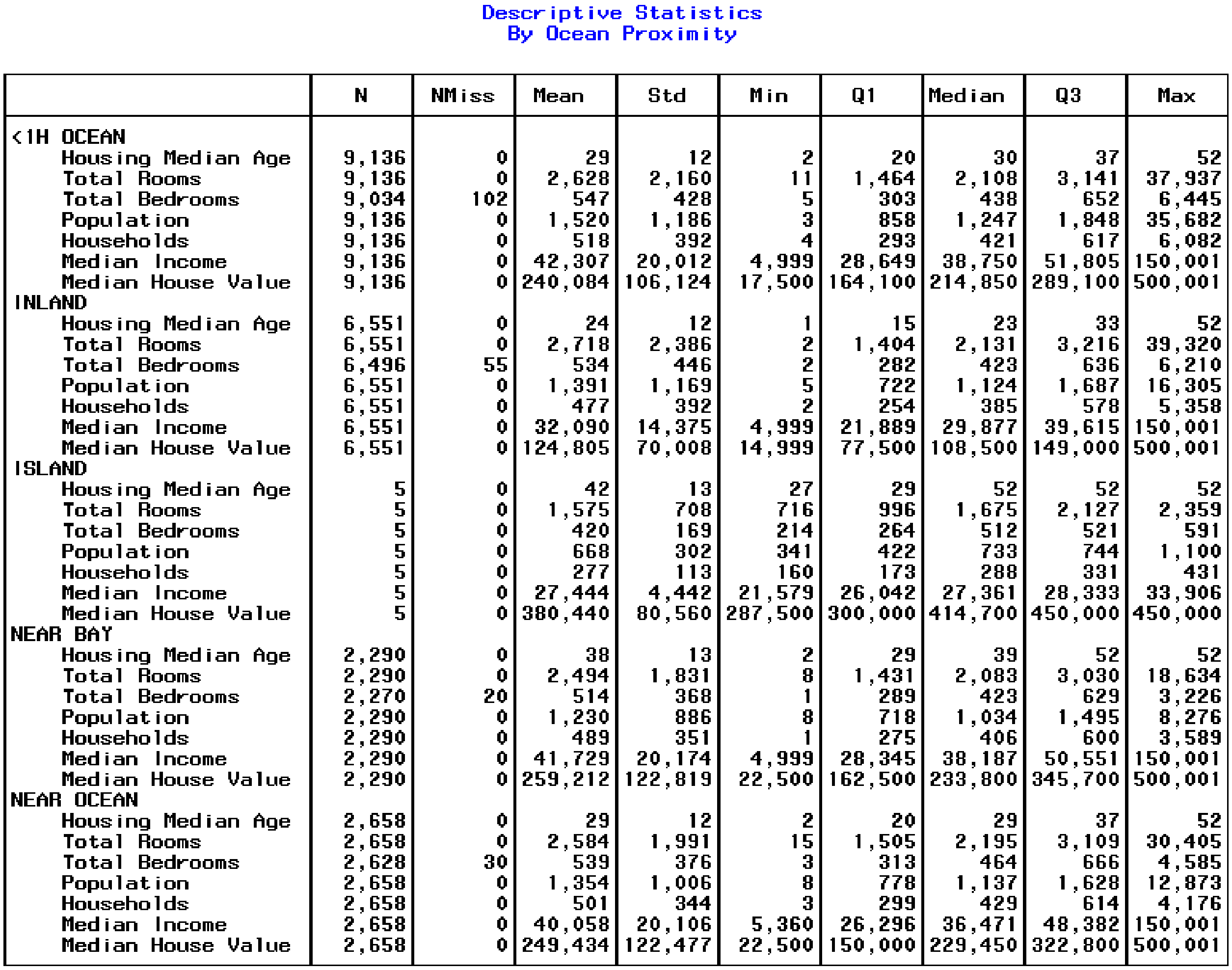


Figure 2: Median House Value By Ocean Proximity

Table 1 contains descriptive statistics of the Census data by ocean proximity. We note that there are only five observations for the class value ISLAND, and that there are 207 missing values for the variable Total Bedrooms in the other four class values. The missing values for Total Bedrooms comprise about 1% of the 20,640 observations in the dataset.

In particular, the median house values of the properties located near water are significantly higher than those located farther inland.



Descriptive Statistics
By Ocean Proximity

	N	NMiss	Mean	Std	Min	Q1	Median	Q3	Max
<1H OCEAN									
Housing Median Age	9,136	0	29	12	2	20	30	37	52
Total Rooms	9,136	0	2,628	2,160	11	1,464	2,108	3,141	37,937
Total Bedrooms	9,034	102	547	428	5	303	438	652	6,445
Population	9,136	0	1,520	1,186	3	858	1,247	1,848	35,682
Households	9,136	0	518	392	4	293	421	617	6,082
Median Income	9,136	0	42,307	20,012	4,999	28,649	38,750	51,805	150,001
Median House Value	9,136	0	240,084	106,124	17,500	164,100	214,850	289,100	500,001
INLAND									
Housing Median Age	6,551	0	24	12	1	15	23	33	52
Total Rooms	6,551	0	2,718	2,386	2	1,404	2,131	3,216	39,320
Total Bedrooms	6,496	55	534	446	2	282	423	636	6,210
Population	6,551	0	1,391	1,169	5	722	1,124	1,687	16,305
Households	6,551	0	477	392	2	254	385	578	5,358
Median Income	6,551	0	32,090	14,375	4,999	21,889	29,877	39,615	150,001
Median House Value	6,551	0	124,805	70,008	14,999	77,500	108,500	149,000	500,001
ISLAND									
Housing Median Age	5	0	42	13	27	29	52	52	52
Total Rooms	5	0	1,575	708	716	996	1,675	2,127	2,359
Total Bedrooms	5	0	420	169	214	264	512	521	591
Population	5	0	668	302	341	422	733	744	1,100
Households	5	0	277	113	160	173	288	331	431
Median Income	5	0	27,444	4,442	21,579	26,042	27,361	28,333	33,906
Median House Value	5	0	380,440	80,560	287,500	300,000	414,700	450,000	450,000
NEAR BAY									
Housing Median Age	2,290	0	38	13	2	29	39	52	52
Total Rooms	2,290	0	2,494	1,831	8	1,431	2,083	3,030	18,634
Total Bedrooms	2,270	20	514	368	1	289	423	629	3,226
Population	2,290	0	1,230	886	8	718	1,034	1,495	8,276
Households	2,290	0	489	351	1	275	406	600	3,589
Median Income	2,290	0	41,729	20,174	4,999	28,345	38,187	50,551	150,001
Median House Value	2,290	0	259,212	122,819	22,500	162,500	233,800	345,700	500,001
NEAR OCEAN									
Housing Median Age	2,658	0	29	12	2	20	29	37	52
Total Rooms	2,658	0	2,584	1,991	15	1,505	2,195	3,109	30,405
Total Bedrooms	2,628	30	539	376	3	313	464	666	4,585
Population	2,658	0	1,354	1,006	8	778	1,137	1,628	12,873
Households	2,658	0	501	344	3	299	429	614	4,176
Median Income	2,658	0	40,058	20,106	5,360	26,296	36,471	48,382	150,001
Median House Value	2,658	0	249,434	122,477	22,500	150,000	229,450	322,800	500,001

Table 2: Descriptive Statistics By Ocean Proximity

Figure 3 presents the distribution of median house values by Ocean Proximity. We note that the unusual spike in median house values at the value $500,001 seems to indicate that values were capped at $500,001. The high frequency of the histogram counts at lower levels of median house value tells us that the homes of the majority of people who live either inland or more than one hour's travel to the ocean are of lower median value than the homes of people who live near water, as indicated in Table 1, *supra.*

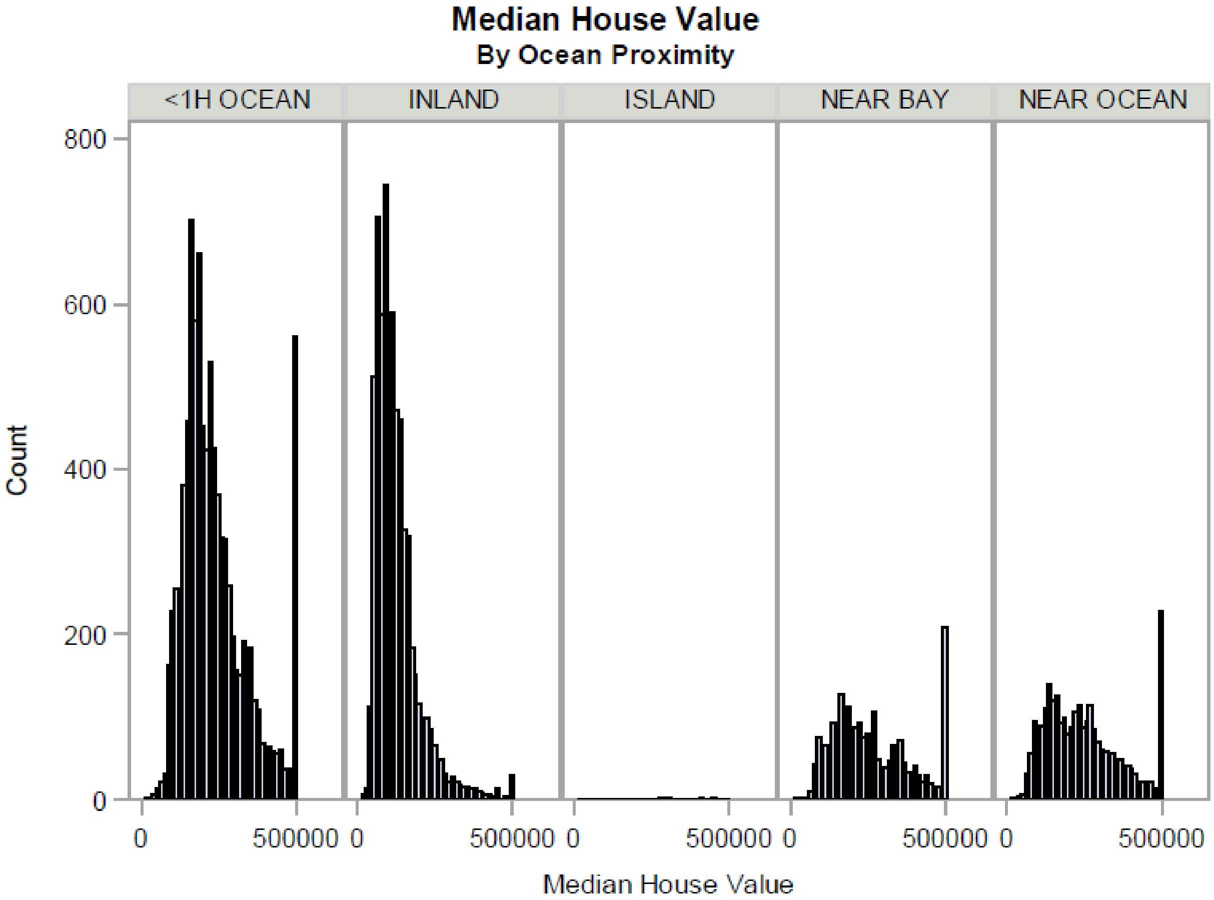
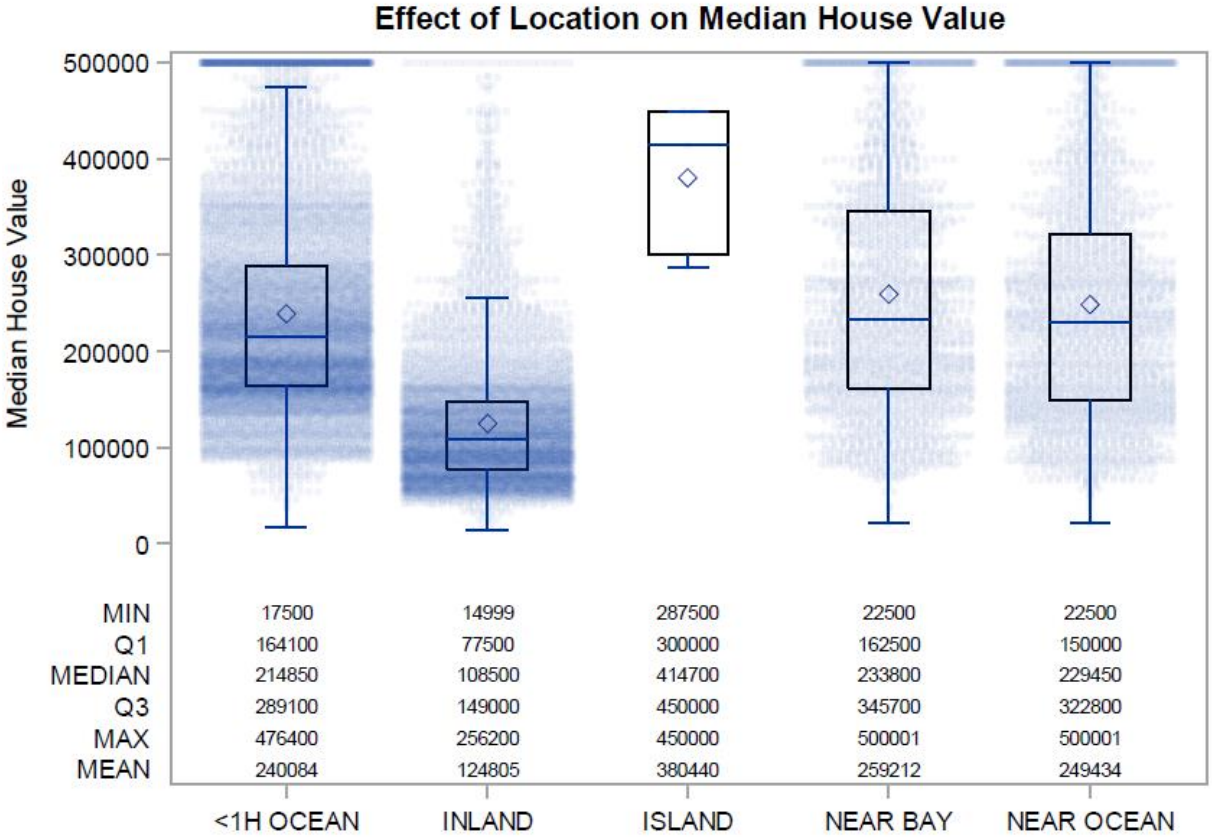


Figure 3: Histogram of Median House Value by Ocean Proximity

The display of box plots of median house value in Figure 4, overlaid with jittered house values, is another way to visualize the impact of location on median house value. The median is represented by the horizontal line inside the box, and the mean is indicated by the diamond marker symbol. When the mean is above the median, the distribution is skewed to the right. We see that this is the case for all of the ocean proximities except ISLAND, where the skew is to the left. Since the average median is less than the median of the medians for ISLAND houses, the average median value of a house is less than 50% of the median house values in the sample data. However, there were only five houses in the ISLAND category of Ocean Proximity, so we cannot with confidence make definitive statements about the effects of location on ISLAND median house values.



	<1H OCEAN	INLAND	ISLAND	NEAR BAY	NEAR OCEAN
MIN	17500	14999	287500	22500	22500
Q1	164100	77500	300000	162500	150000
MEDIAN	214850	108500	414700	233800	229450
Q3	289100	149000	450000	345700	322800
MAX	476400	256200	450000	500001	500001
MEAN	240084	124805	380440	259212	249434

Figure 4: Effect of Location on Median House Value

Figure 5 is a matrix of correlations of median house value and other covariates. The data show us that:

- Housing Median Age is not strongly correlated with any other variable at any location, with the exception of Median Income of ISLAND houses. However, since there are only five ISLAND observations, this relationship may not be indicative of any trend.
- Median House Value is moderately correlated with Median Income of houses <1H OCEAN or NEAR OCEAN houses. Perhaps individuals with high incomes are more able to afford more expensive houses nearer water.
- Median Income is moderately correlated with Housing Median Age for ISLAND locations ($n = 5$) and Median House Value for the remaining locations.
- Number of Households per census block correlates strongly with Population, Total Bedrooms, and Total Rooms, presumably because people live in houses, and houses have rooms and bedrooms.
- Population and Households are highly correlated so it is reasonable to observe that the variables Population, Total Rooms, and Total Bedrooms are highly correlated with each other.

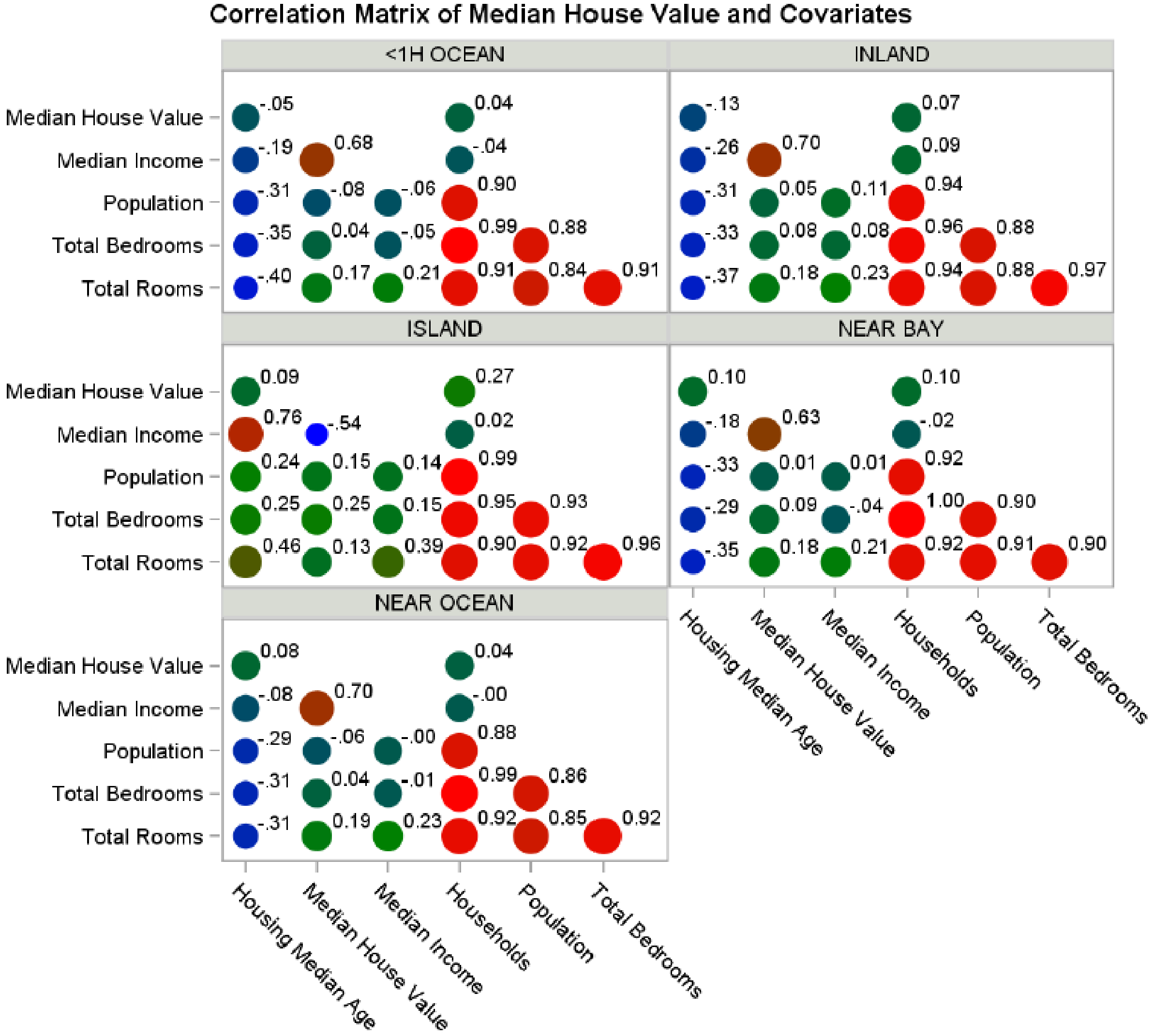


Figure 5: Matrix of Correlations of Median House Value and Other Covariates

Figure 6 is a graph comparing median house value with individual covariates. Median house value is weakly correlated with all covariates but Median Income for all locations except ISLAND. Since there are only five observations for ISLAND, we cannot make definitive statements about the relationship of Median House Value and Median Income for ISLANDers.

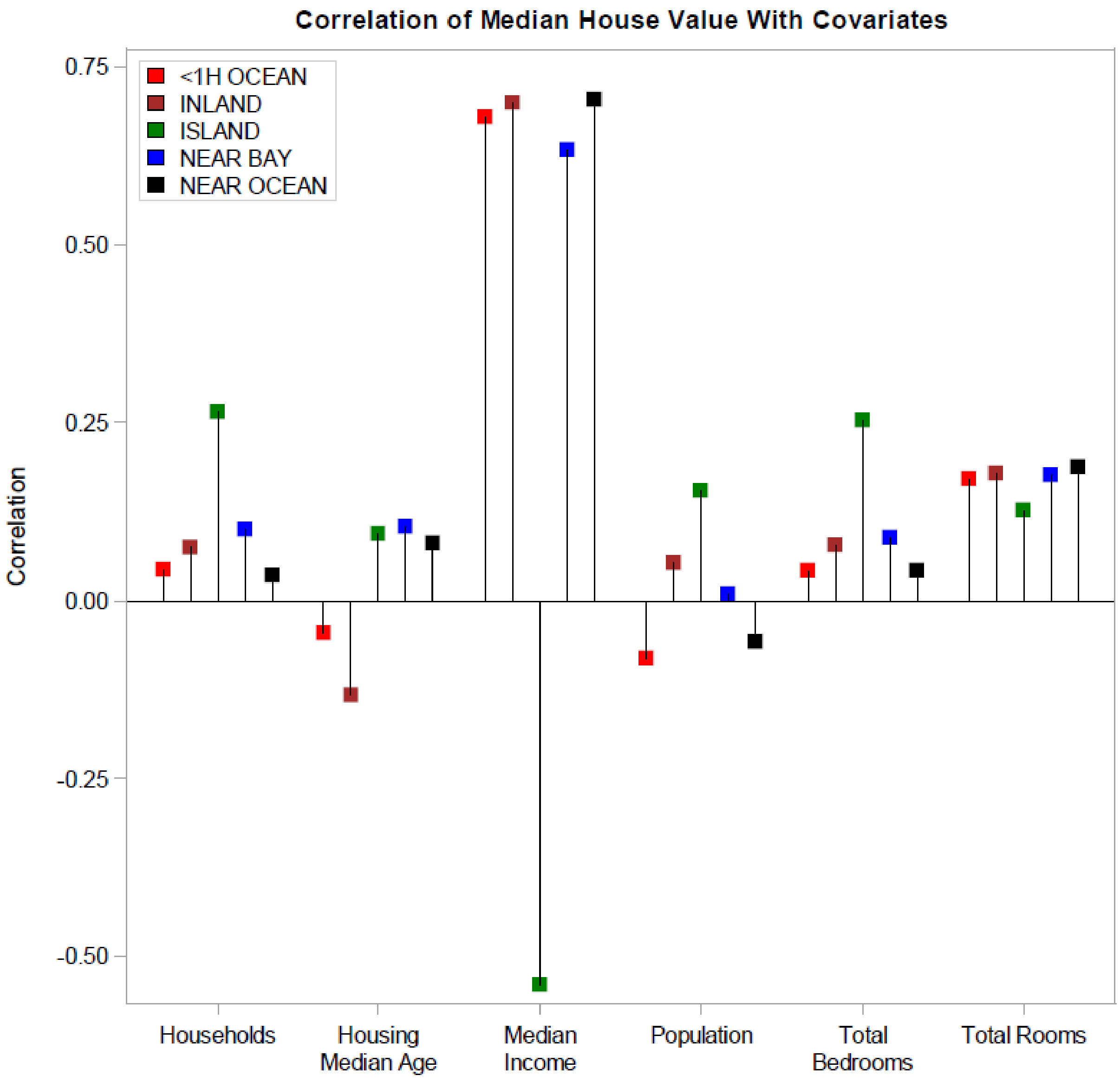


Figure 0: Correlation of Median House Value With Covariates

Analysis of Missingness of Total Bedrooms Variable

Figure 7 is a map of the locations of missing values for Total Bedrooms. Since Total Bedrooms is the only variable with missing values, we cannot perform a χ^2 test to determine whether or not the variable is MCAR. So, we will have to proceed indirectly to test the hypothesis that missingness of Total Bedrooms is caused completely at random with no functional relationship between Total Bedrooms and any other covariate.

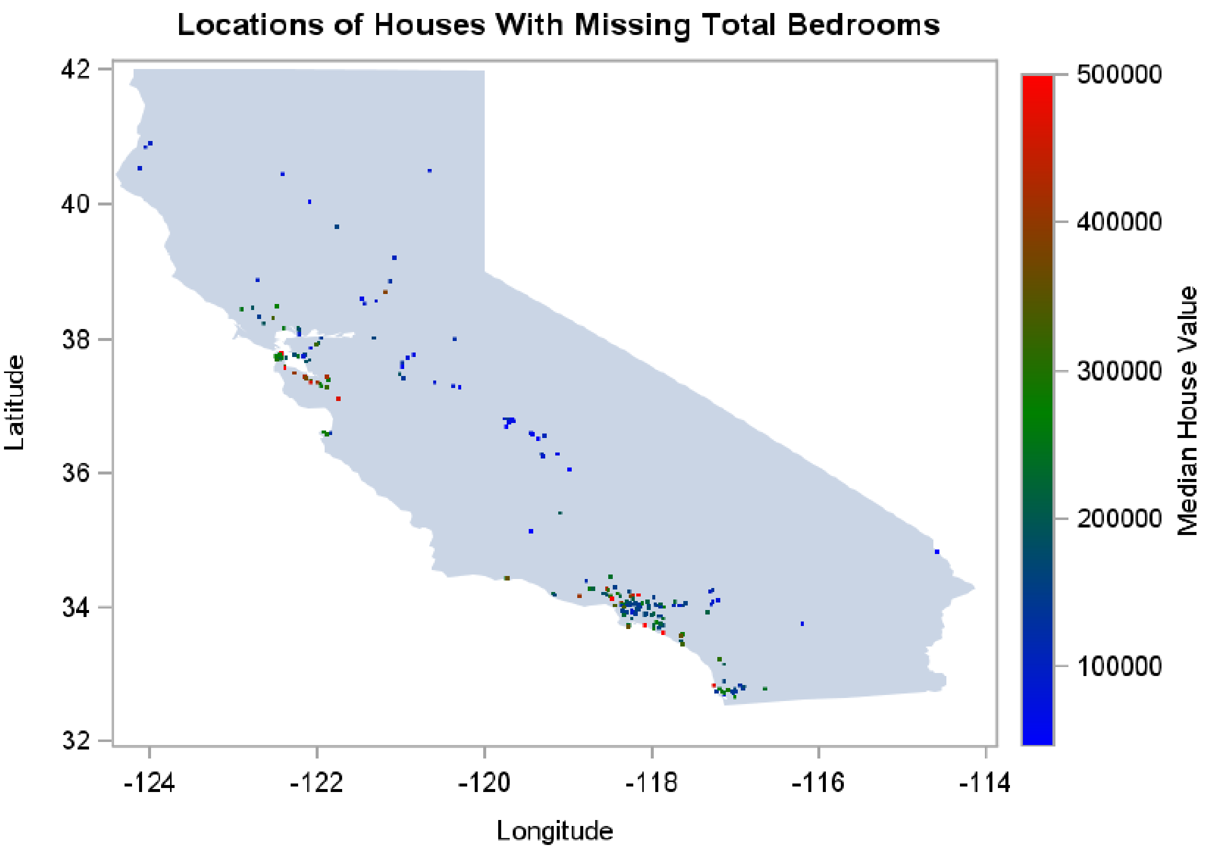


Figure 7: Locations of Houses With Missing Total Bedrooms

Household-level nonresponse to a census survey questionnaire is problematic because there may be many causes of nonresponse, such as outdated information regarding inhabitants of a dwelling or non-compliance of neighbors or building manager in supplying information to a census enumerator. Since we cannot attribute the pattern of nonresponse of Total Bedrooms to any functional relationship in the dataset, we must classify the missing values of Total Bedrooms to be missing completely at random, or MCAR. We know that the correlation between Total Bedrooms and Total Rooms is quite high ($r = 0.9+$) but if we attempt to impute values to the missing values of Total Bedrooms based on Total Rooms or any other combination of variables, we will be introducing bias into the data in the ways described *supra*.

It is very curious that Total Rooms is 100% populated while Total Bedrooms, which is a subset of Total Rooms, has missing values. There must have been a failure in the enumeration process, but we are unable to determine it.

Fuzzy *c*-Means Imputation

Fuzzy *c*-means imputation is a two-stage process. In the first stage, the data are clustered using the fuzzy *c*-means algorithm. This effort requires that we determine the fuzzy exponent with which to build fuzzy membership functions, and also the number of clusters to create. It is iterative in nature and requires subjective thinking regarding the number of clusters to form. The cluster centers produced in this stage are used to impute candidate values for missing values in the second stage. In the second stage, the centers of the clusters formed in the first stage are used to replace the missing values with candidate values and thus create an initial set of complete data. The candidate values are replaced iteratively by calculations involving updated cluster centers and fuzzy membership functions derived from the data to be clustered. The fuzzy *c*-means algorithm is described at length in [13], and its extension to imputation is developed in [14].

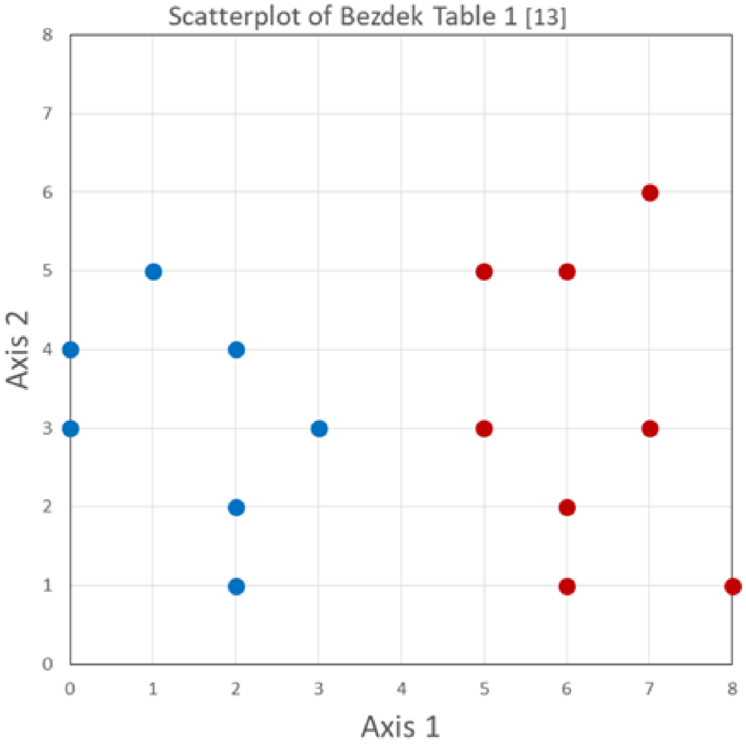
Simple Example of Fuzzy *c*-Means Clustering

This example is taken from [13] and is meant to demonstrate the fuzzy *c*-means SAS macro `%FCM` and validate its implementation. The data and results come from Table 1 in [13]. The SAS program to invoke and execute the `%FCM` macro is given below. A description of the `%FCM` macro parameters and its invocation is given in the Appendix.

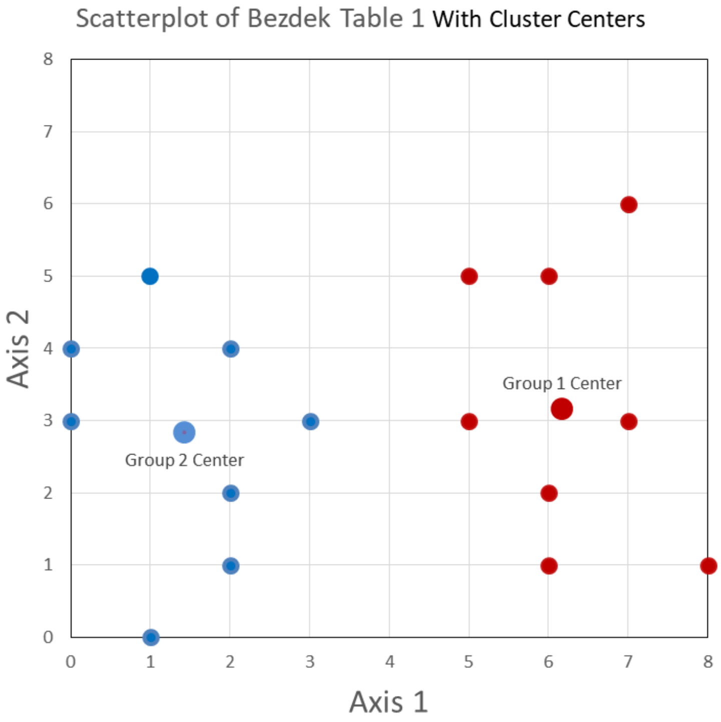
```
data test ;
input group $ y_k1 y_k2 ;
datalines ;
2 0 4   2 0 3   2 1 5   2 2 4   2 3 3   2 2 2   2 2 1   2 1 0
1 5 5   1 6 5   1 7 6   1 5 3   1 7 3   1 6 2   1 6 1   1 8 1
;;;;
run ;

%FCM( test
    , dsnout=test_out_fcm, dsnseed=test_seed_fcm, dsnstat=test_stat_fcm
    , class=group, print_mf=16
    )
```

PROC IML is used to perform fuzzy *c*-means clustering. The results are shown below.

```
/--------------------------\
| Fuzzy c-means clustering |
\--------------------------/

Parameters

Name of input  dataset   = test
Name of output dataset   = test_out_fcm
Name of seed   dataset   = test_seed_fcm
Name of stat   dataset   = test_stat_fcm
Number of clusters       =          2
Number of observations   =         16
Number of variables      =          2
Fuzzification parameter  =       2.00
Maximum # iterations     =        100
Minimum improvement      = .000100000
Random number seed       =       2020
```



```
Fuzzy c-means clustering converged after 12 iterations

       Cluster Centers
         CC_y_k1      CC_y_k2

[1]      6.176596     3.158212
[2]      1.436979     2.827618

 Fuzzy Membership Matrix
       MF_1   MF_2 Cluster

[ 1] 0.0813 0.9187  2.0000
[ 2] 0.0520 0.9480  2.0000
[ 3] 0.1399 0.8601  2.0000
[ 4] 0.0852 0.9148  2.0000
[ 5] 0.1964 0.8036  2.0000
[ 6] 0.0506 0.9494  2.0000
[ 7] 0.1420 0.8580  2.0000
[ 8] 0.1821 0.8179  2.0000
[ 9] 0.7848 0.2152  1.0000
[10] 0.8818 0.1182  1.0000
[11] 0.8241 0.1759  1.0000
[12] 0.9003 0.0997  1.0000
[13] 0.9778 0.0222  1.0000
[14] 0.9400 0.0600  1.0000
[15] 0.8375 0.1625  1.0000
[16] 0.8532 0.1468  1.0000

      Performance Measures
                       Statistic

Within-Cluster SS         60.5644
Total          SS        156.4388
R^2 =  1-WCSS/TSS          0.6129
Partition Coefficient      0.7942
Partition Entropy          0.3518
```



The fuzzy membership matrix represents the fuzzy membership function and the cluster assignment for each observation. For example, the first observation has the membership function (0.0813, 0.9187), and the observation is assigned to cluster 2. By way of interpretation, observation 1 is similar to observations in cluster 1 to fuzzy membership degree 0.0813 and similar to observations in cluster 2 to membership degree 0.9187. Hence, since it is more similar to observations in cluster 2, it is assigned the label ‘2’. The sum of the membership functions per observation is 1, which is a precondition of fuzzy c-means clustering. Bezdek [13] goes into extensive detail concerning fuzzy membership functions. The Partition Coefficient and the Partition Entropy are “cluster validity functionals” which may be used to determine the “goodness” of the clustering. The interested reader is referred to Bezdek [13] for details.

Equivalent results from Table 2 in [13] are reported in Table 2 below. The fuzzification parameter $m = 2.0$, and there are two clusters formed.

Cluster Center				Partition Coefficient	Partition Entropy
$\hat{v}_{11}$	$\hat{v}_{12}$	$\hat{v}_{21}$	$\hat{v}_{22}$	0.80	0.35
6.18	3.15	1.44	2.83		

Table 3: Fuzzy c-means Performance Measures from Bezdek [13]

The SAS %FCM macro reliably replicates Bezdek's implementation of the fuzzy c-means algorithm with results very similar to those reported in Table 2 of [13].

Fuzzy c-Means Clustering and Imputation of California Census Data

The reproduction of Bezdek's results as reported in Table 1 of [13] was a straightforward exercise because the fuzzification parameter and the number of clusters were specified. We must determine the number of clusters and the fuzzy exponent to perform imputation on the missing values of the Total Bedrooms variable.

Phase 1: Determining Number of Clusters and Fuzzy Exponent

As part of the exploratory phase of the project, we invoked the %FCM macro over a grid of (fuzzy exponent, number_of_clusters) pairs.

The SAS code for the pair (2.0, 10) is given below:

```
libname CALHOUS "C:\Users\Username\Documents\My SAS Files\Missing Value Impu-
tation\SASData" ;

%let DSNIN = CALHOUS.calhous_fcm ;

/* USAGE NOTE: the order of the variables must be maintained and not changed
 *             when used with %FCM and %FCM_IMPUTE
 */

%let VARS  = longitude latitude housing_median_age total_rooms total_bedrooms
             population households median_income median_house_value ;

data &DSNIN ;
    /* create complete-case data for Phase 1 of fuzzy c-means clustering */
    set CALHOUS.calhous ;

    /* total_bedrooms has missing values */

    where ^missing( total_bedrooms) ;

    /* create discrete values to serve as class variable */
    group = round( median_house_value, 50000 ) ;
run ;

%FCM( &DSNIN
    , class=group
    , dsnout=CALHOUS.calhous_fcm_out
    , dsnseed=CALHOUS.calhous_fcm_seed
    , dsnstat=CALHOUS.calhous_fcm_stat
    , m=2
    , max_iter=200
    , min_improv=.01
    , n_clus=10
    , print=y
    , print_mf=5
    , vars=&VARS
    )
```

An abridged listing of the `%FCM` output is shown below. Particularly, the iteration history is omitted, and several of the cluster center variables are not shown for reasons of brevity.

```
/--------------------------\
| Fuzzy c-means clustering |
\--------------------------/

---------------------------------------------------------------------------------
Parameters

Name of input  dataset  = CALHOUS.calhous_fcm
Name of output dataset  = CALHOUS.calhous_fcm_out
Name of seed   dataset  = CALHOUS.calhous_fcm_seed
Name of stat   dataset  = CALHOUS.calhous_fcm_stat
Number of clusters      =         10
Number of observations  =     20,433
Number of variables     =          9
Fuzzification parameter =       2.00
Maximum # iterations    =        200
Minimum improvement     = .010000000
Random number seed      =       2020
```

```
Fuzzy c-means clustering converged after 159 iterations
                                                                    Cluster Centers
     CC_LONGITUDE  CC_LATITUDE CC_HOUSING_MEDIAN_AGE CC_TOTAL_ROOMS CC_TOTAL_BEDROOMS

[ 1] -119.855659    36.892478            29.338589    1899.563148        428.169465
[ 2] -119.944815    35.502333            30.807174    2951.673724        545.095635
[ 3] -119.696029    35.382338            29.112855    2811.035682        560.006992
[ 4] -119.691000    36.285741            28.171865    2244.820167        488.755966
[ 5] -119.153864    35.139555            28.351361    2445.084001        529.868732
[ 6] -119.908163    35.513969            28.067366    3141.307275        598.765221
[ 7] -120.020693    35.549007            32.394759    2984.975188        540.953807
[ 8] -119.338603    35.176920            28.329170    2644.842600        553.110608
[ 9] -119.680902    35.205665            33.651908    2989.335932        503.338642
[10] -119.368787    35.636333            26.801518    2583.640099        551.172991

                             Fuzzy Membership Matrix
      MF_1   MF_2   MF_3   MF_4   MF_5   MF_6   MF_7   MF_8   MF_9  MF_10 Cluster

[1] 0.0011 0.0209 0.0041 0.0014 0.0021 0.0070 0.9252 0.0028 0.0337 0.0017  7.0000
[2] 0.0071 0.0102 0.3374 0.0119 0.0628 0.0373 0.0042 0.5027 0.0022 0.0243  8.0000
[3] 0.0022 0.0075 0.8810 0.0033 0.0109 0.0535 0.0025 0.0322 0.0012 0.0057  3.0000
[4] 0.0015 0.0187 0.0344 0.0021 0.0052 0.9188 0.0041 0.0103 0.0016 0.0033  6.0000
[5] 0.0066 0.0091 0.2896 0.0111 0.0604 0.0330 0.0038 0.5615 0.0020 0.0229  8.0000

      Performance Measures
                       Statistic

Within-Cluster SS      6.9927E12
Total          SS     3.09849E14
R^2 =  1-WCSS/TSS         0.9774
Partition Coefficient     0.5895
Partition Entropy         0.8903
```

The highest fuzzy membership value for observation 1 is 0.9252 for cluster 7 and minimal values for the other 9 clusters, so it is assigned to cluster 7. Similarly, the highest fuzzy membership value for observation 5 is 0.5615 for cluster 2 and the observation is assigned to cluster 8.

The result for the full set of grid pairs is graphed in Figure 8 and summarized in Figure 9.

We see that the within-cluster sum of squares, which is a measure of the variation within a cluster, decreases as the number of clusters increases. There are many subspaces within the overall data space that are detected by the algorithm, and increasing the number of clusters that are formed allows observations to be assigned more accurately to a cluster.

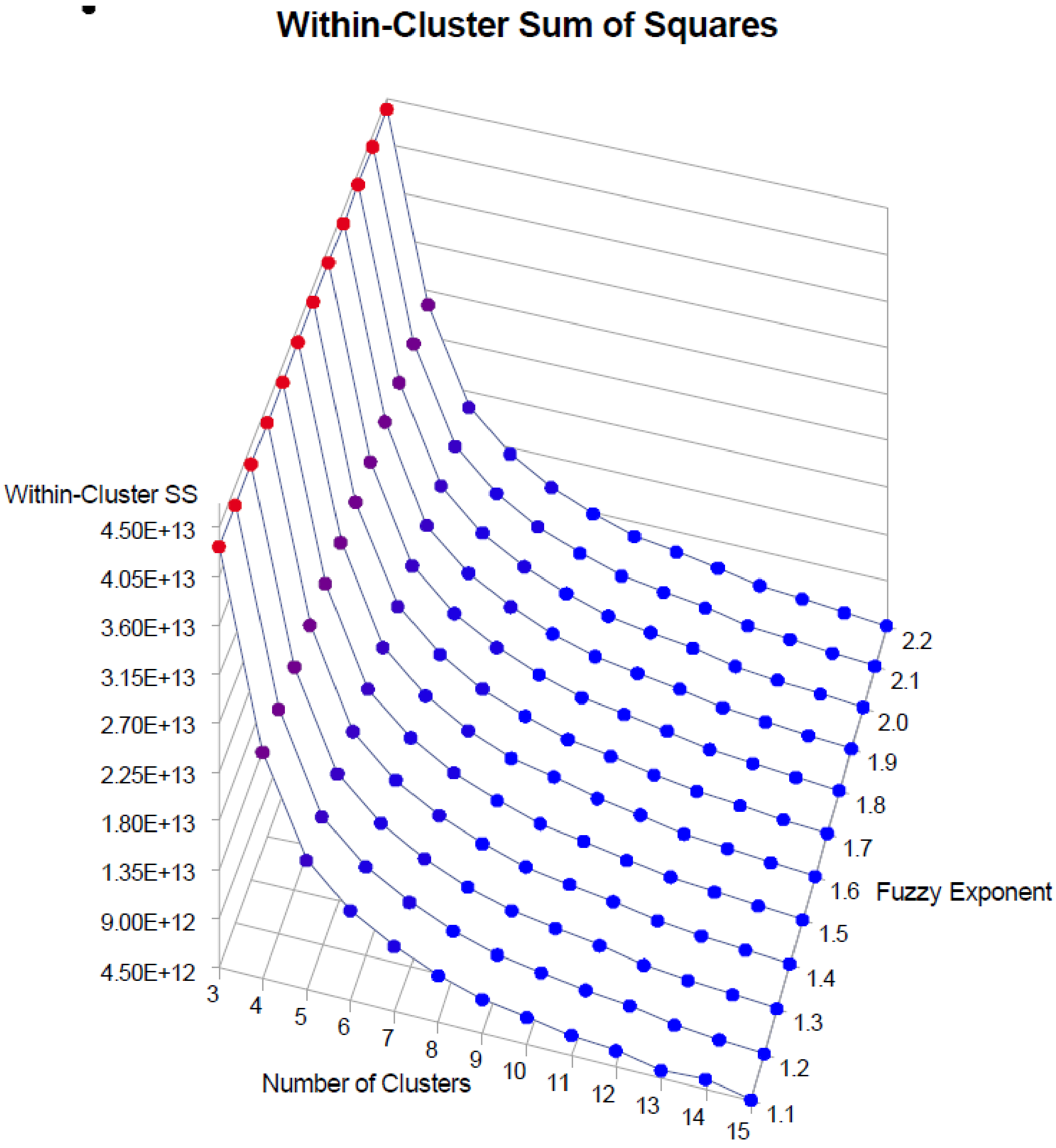


Figure 8: Fuzzy c-Means Performance

Figure 8 shows the heuristic “knee” method of determining the number of clusters. We subjectively determine the number of clusters by examining the decrease in the mean WCSS[6] as a function of the number of clusters and deciding where the curve flattens into approximate linearity. We have superimposed

[6] We computed the mean WCSS as the mean of the WCSS over all WCSS values computed for a given number of clusters. Each (fuzzy exponent, number_of_clusters) pair produced a WCSS, and the mean of these WCSS over all

the percent change in the decrease of the mean WCSS as an aid to suggesting where the flattening starts. In this case, the decrease in the mean WCSS begins to become linear at the 10-cluster point. The decreases are relatively constant after 10 clusters, so we will apply the `%FCM` macro using 10 clusters.

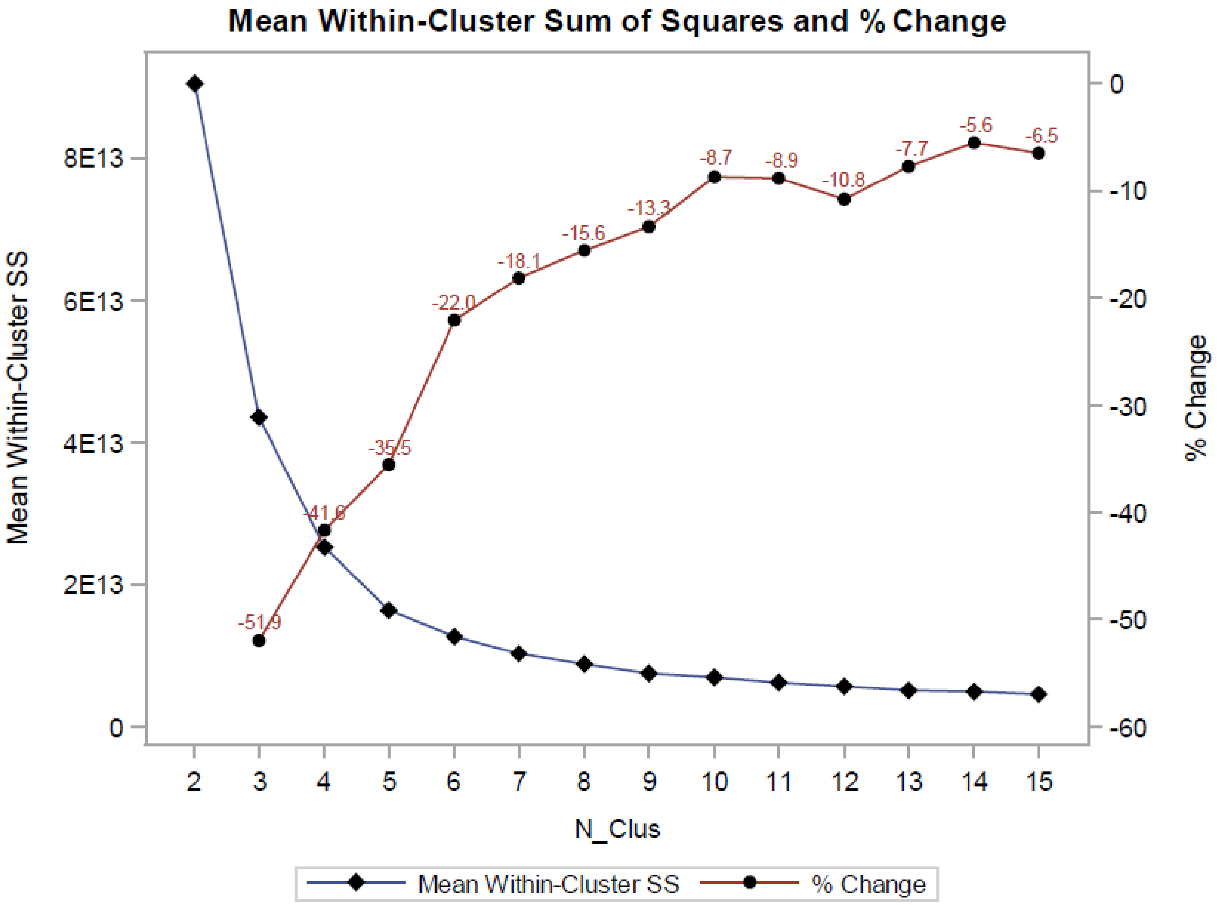


Figure 9: Summary of Fuzzy c-Means Macro Performance

fuzzy exponent runs was computed for each number_of_clusters point, thus collapsing a two-dimensional grid of WCSS values into a one-dimensional set of WCSS values.

We must now answer the question: what is the best fuzzy exponent to use? Figure 9 shows the pattern of decreasing WCSS as the number of clusters increases, but does not clearly indicate the best fuzzy exponent to use. Figure 10 represents the percent change in WCSS for each (fuzzy exponent, number_of_clusters) grid point, and is meant to clarify the effect of the fuzzy exponent on the WCSS for a specified number of clusters. Using the result from Figure 9, if we choose 10 clusters, we see a significant decrease in the WCSS for the fuzzy exponent value = 1.6 . Let us proceed to the imputation phase using 10 clusters and fuzzy parameter = 1.6 .

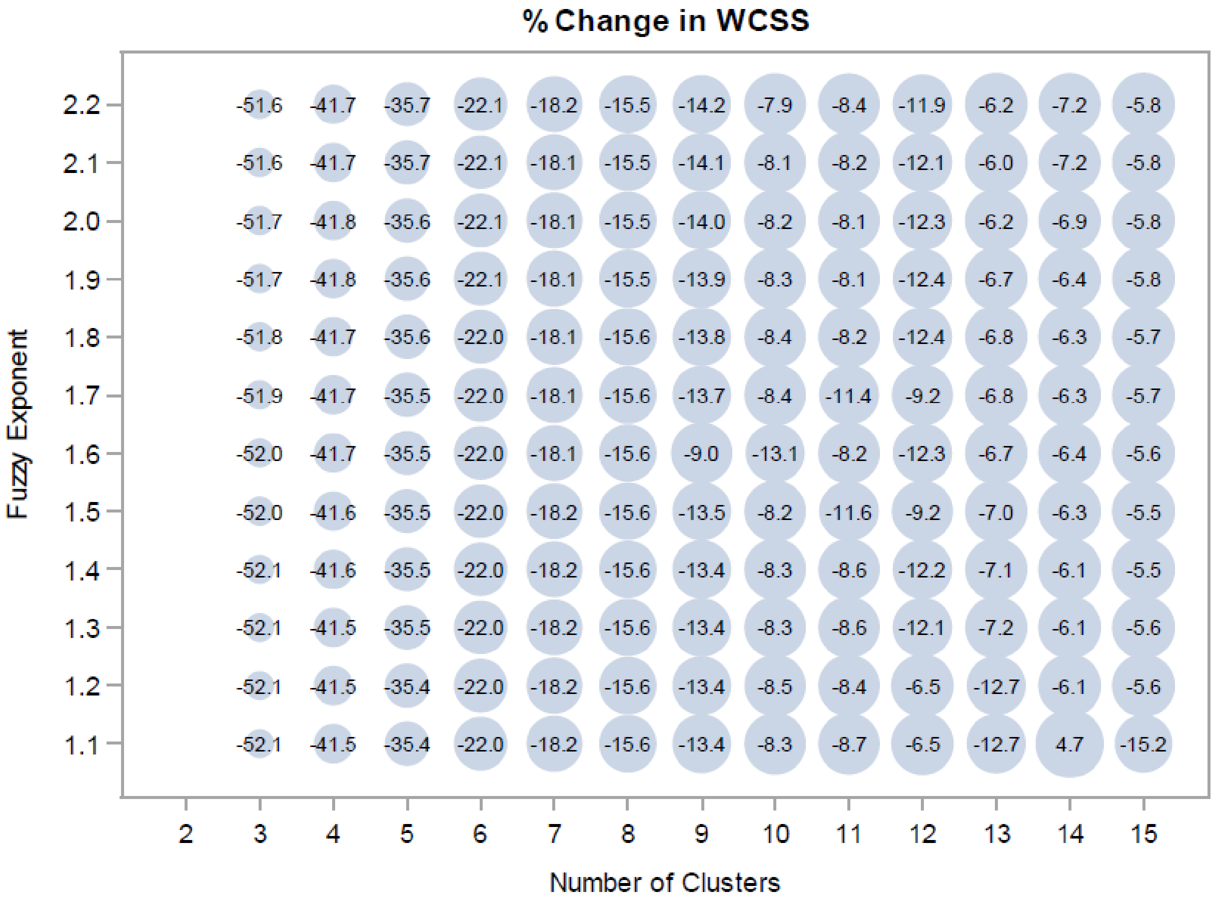


Figure 10: Percent Change in WCSS as Function of Fuzzy Exponent and Number of Clusters

Phase 2: Imputation

The SAS code for the pair (1.6, 10) is given below. The dataset `CALHOUS.calhous_fcm_seed` was created in Phase 1. It contains the cluster centers generated from the data and is used to populate the initial cluster center matrix. The dataset `CALHOUS.calhous_fcm_impute_out` contains the original complete-case variables, the variables with imputed missing values, membership functions, and the cluster assignment.

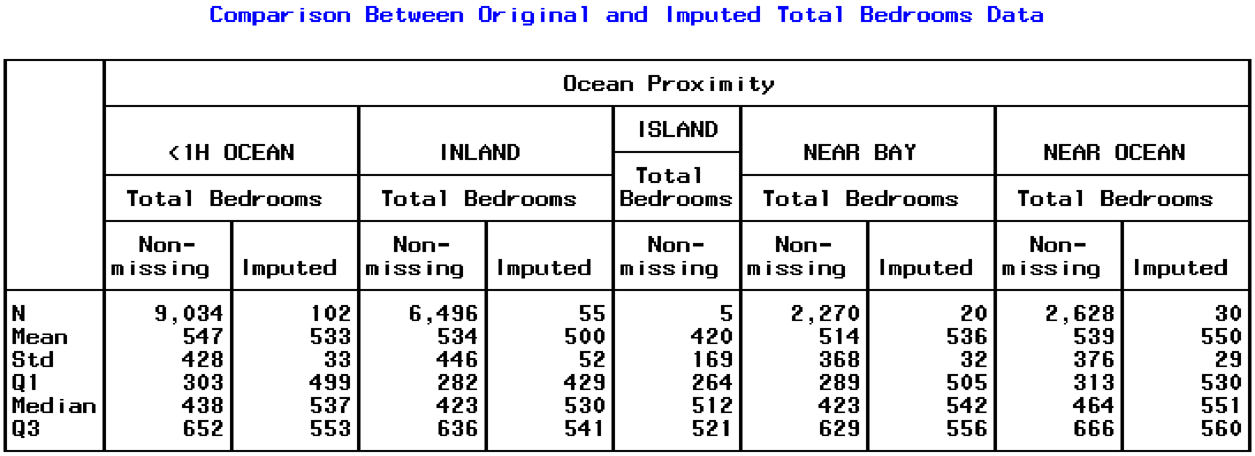
```
data CALHOUS.calhous_fcm_impute ;
    /* include obs with missing values of total_bedrooms */

    set CALHOUS.calhous ;

    group = round( median_house_value, 50000 ) ; /* create cluster labels */
run ;

%FCM_IMPUTE( CALHOUS.calhous_fcm_impute
           , CALHOUS.calhous_fcm_seed
           , dsnout_impute=CALHOUS.calhous_fcm_impute_out
           , dsnseed_impute=CALHOUS.calhous_fcm_impute_seed
           , class=group
           , m=1.6
           , max_iter=200
           , min_improv=.01
           , n_clus=10
           , print=y
           , print_mf=5
           , vars=&VARS
           )
```

An abridged listing of the output is shown below.

```
/-------------------------------------\
| Fuzzy c-Means Clustering Imputation |
| Using Optimal Completion Strategy   |
\-------------------------------------/

--------------------------------------------------------------------------------
Parameters

Name of input  dataset  = CALHOUS.calhous_fcm_impute
Name of output dataset  = CALHOUS.calhous_fcm_impute_out
Name of seed   dataset  = CALHOUS.calhous_fcm_impute_seed
Name of stat   dataset  = CALHOUS.calhous_fcm_impute_stat
Number of clusters      =         10
Fuzzification parameter =       1.60
Maximum # iterations    =        200
Minimum improvement     = .010000000
Random number seed      =       2020
--------------------------------------------------------------------------------

Fuzzy c-means clustering converged after 41 iterations
```

```
                                                                      Cluster Centers
      CC_LONGITUDE   CC_LATITUDE CC_HOUSING_MEDIAN_AGE CC_TOTAL_ROOMS CC_TOTAL_BEDROOMS

[ 1]   -119.846098    36.883775             29.370071    1890.681959         421.648718
[ 2]   -119.317454    35.169573             28.141863    2665.613995         553.806944
[ 3]   -119.690157    35.221560             33.528932    2981.695411         502.189852
[ 4]   -119.147628    35.152518             28.076115    2465.949268         526.110540
[ 5]   -119.354580    35.654250             26.562873    2633.155192         553.883384
[ 6]   -119.626306    35.324228             28.935163    2839.210505         564.346983
[ 7]   -119.680358    36.286874             28.099602    2240.982271         482.208816
[ 8]   -120.003948    35.544412             32.035141    3024.015241         537.978343
[ 9]   -119.912554    35.484552             30.456059    3011.143826         552.355045
[10]   -119.884283    35.498624             28.010253    3132.628479         591.279152

                            Fuzzy Membership Matrix
      MF_1   MF_2   MF_3   MF_4   MF_5   MF_6   MF_7   MF_8   MF_9  MF_10 Cluster

[1] 0.0000 0.0001 0.0082 0.0001 0.0001 0.0002 0.0000 0.9873 0.0034 0.0006  8.0000
[2] 0.0006 0.4022 0.0001 0.0186 0.0041 0.5616 0.0013 0.0002 0.0011 0.0102  6.0000
[3] 0.0001 0.0078 0.0000 0.0015 0.0005 0.9600 0.0002 0.0001 0.0009 0.0288  6.0000
[4] 0.0000 0.0005 0.0000 0.0002 0.0001 0.0038 0.0000 0.0001 0.0019 0.9932  10.000
[5] 0.0005 0.4860 0.0001 0.0189 0.0040 0.4789 0.0013 0.0002 0.0010 0.0090  2.0000

      Performance Measures
                         Statistic

Within-Cluster SS      7.04485E12
Total          SS      3.09586E14
R^2 =  1-WCSS/TSS          0.9772
Partition Coefficient      0.7903
Partition Entropy          0.4087
```

The dataset CALHOUS.calhous_fcm_impute_out contains the original data and imputed data that the %FCM_IMPUTE macro produced. Table 4 contains descriptive statistics of the original and imputed data distributions of the Total Bedrooms variable. If the imputation method applied to missing observations accurately reproduces the distribution of the nonmissing observations, the values of the mean, standard deviation, and quantiles of the Total Bedrooms data ought to be close to the values of the nonmissing data. We see that the means of the nonmissing and imputed observations are reasonably close but the disparity by an order of magnitude between nonmissing and imputed standard deviation indicate that the %FCM_IMPUTE algorithm does not capture the variation inherent in the sample data.



Comparison Between Original and Imputed Total Bedrooms Data

	Ocean Proximity								
	<1H OCEAN		INLAND		ISLAND	NEAR BAY		NEAR OCEAN	
	Total Bedrooms		Total Bedrooms		Total Bedrooms	Total Bedrooms		Total Bedrooms	
	Non-missing	Imputed	Non-missing	Imputed	Non-missing	Non-missing	Imputed	Non-missing	Imputed
N	9,034	102	6,496	55	5	2,270	20	2,628	30
Mean	547	533	534	500	420	514	536	539	550
Std	428	33	446	52	169	368	32	376	29
Q1	303	499	282	429	264	289	505	313	530
Median	438	537	423	530	512	423	542	464	551
Q3	652	553	636	541	521	629	556	666	560

Table 4: Comparison Between Original and Imputed Total Bedrooms Data

Fuzzy c-Means Clustering for Multiple Imputation

Let us extend the exercise to imputing several variables, namely, Median Income and Total Bedrooms. We will assume that homeowners in the upper 1% of median house values are shy about revealing their incomes and that 90% of them decline to answer the census survey item regarding annual income. We will further assume that each Ocean Proximity location constitutes a separate income group so that the median income is different for each value of Ocean Proximity. We exempt ISLAND from this intervention because there are too few houses in that location. Thus, we will have created MAR data because $P(Y\ is\ missing \mid X, Y) = P(Y\ is\ missing \mid X)$. We have studied the Total Bedrooms variable in the section *Analysis of Missingness of Total Bedrooms Variable* (*vide supra*). The SAS code to perform the analysis is contained in Appendix B.

Figure 11 represents a comparison between the original median income data and the imputed median income data by Ocean Proximity. No missingness was introduced into the ISLAND data because there were only five observations, much too sparse a sample with which to tinker. We see that the distributions are very close in that the percentiles are only marginally different (see Table 4, *vide supra)*. In particular, the outliers of the imputed data are more constrained than the original data, as shown by the smaller standard deviation statistic.

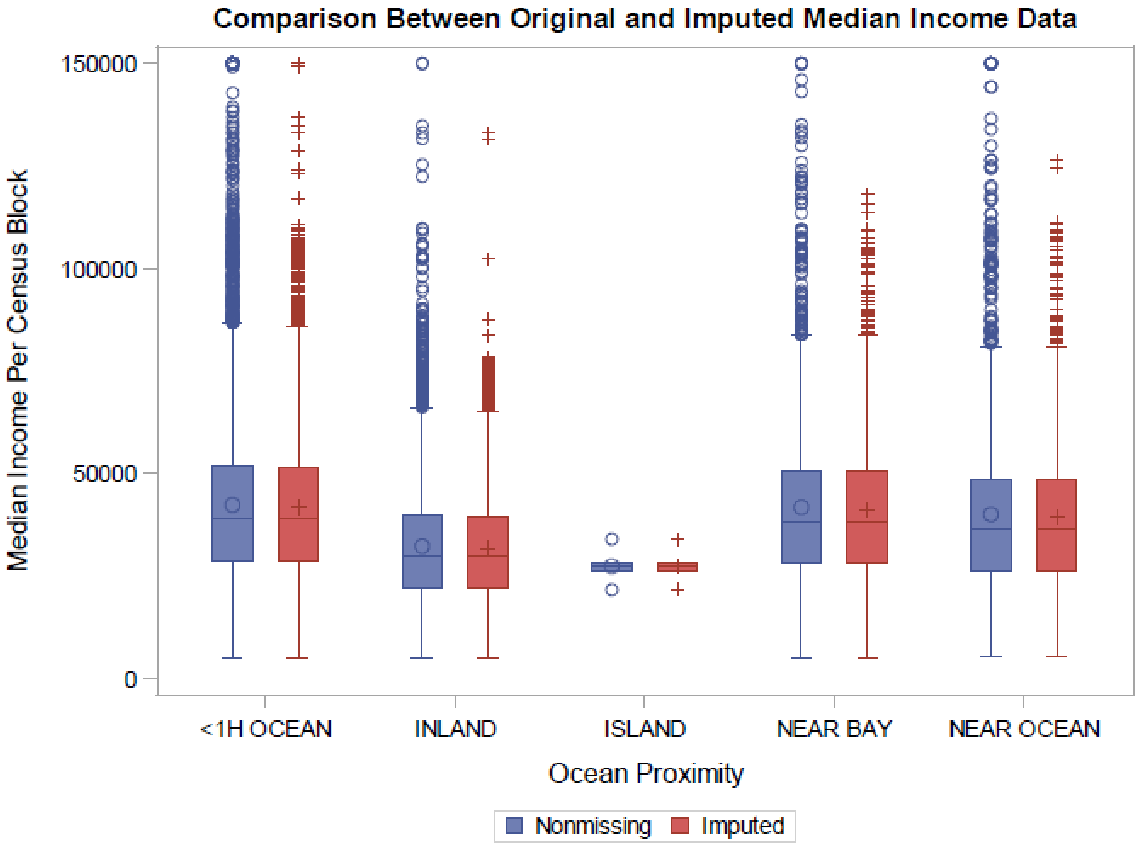
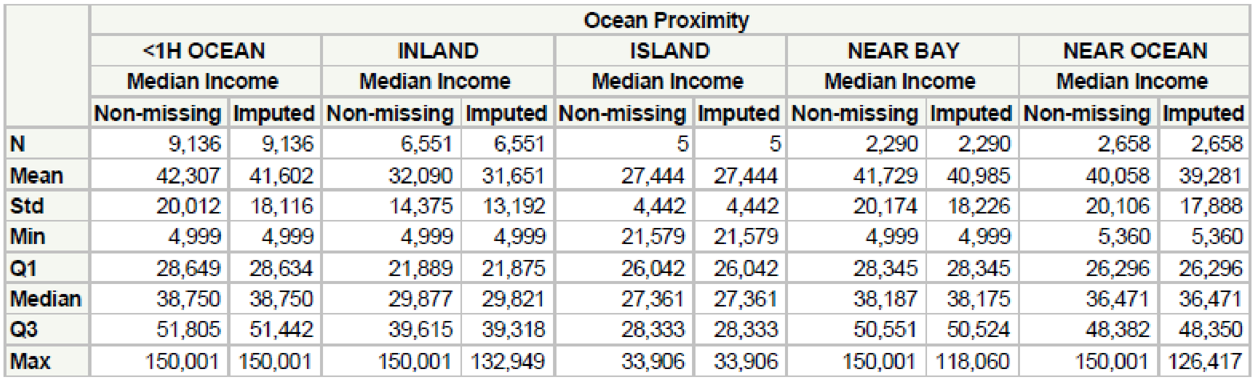


Figure 11: Comparison Between Original and Imputed Median Income Data

Table 5 contains the first two moments and the 0th, 25th, 50th, 75th, and 100th percentiles of the original and imputed median income data[7]. We see that the fuzzy c-means algorithm produces imputed values that very closely resemble the distribution of the original data. The relatively large differences between original and imputed values in the 100th percentile are due to the fact that the fuzzy c-means algorithm does not create imputed values with the exact same variation as is represented in the original data.

Descriptive Statistics of Original and Imputed Median Income
By Ocean Proximity



	Ocean Proximity									
	<1H OCEAN		INLAND		ISLAND		NEAR BAY		NEAR OCEAN	
	Median Income		Median Income		Median Income		Median Income		Median Income	
	Non-missing	Imputed	Non-missing	Imputed	Non-missing	Imputed	Non-missing	Imputed	Non-missing	Imputed
N	9,136	9,136	6,551	6,551	5	5	2,290	2,290	2,658	2,658
Mean	42,307	41,602	32,090	31,651	27,444	27,444	41,729	40,985	40,058	39,281
Std	20,012	18,116	14,375	13,192	4,442	4,442	20,174	18,226	20,106	17,888
Min	4,999	4,999	4,999	4,999	21,579	21,579	4,999	4,999	5,360	5,360
Q1	28,649	28,634	21,889	21,875	26,042	26,042	28,345	28,345	26,296	26,296
Median	38,750	38,750	29,877	29,821	27,361	27,361	38,187	38,175	36,471	36,471
Q3	51,805	51,442	39,615	39,318	28,333	28,333	50,551	50,524	48,382	48,350
Max	150,001	150,001	150,001	132,949	33,906	33,906	150,001	118,060	150,001	126,417

Table 5: Descriptive Statistics of Original and Imputed Median Income

Summary

We have explored several topics related to missing data, an important topic in statistical applications and machine learning. We described the varieties of missing data in terms of the possible mechanisms of creation and indicated criteria for evaluating an imputation method. We briefly reviewed historic and current imputation methods and note that these methods have become more sophisticated as computational resources have become more available and powerful.

We examined in detail one popular imputation method, fuzzy c-means clustering imputation, as applied to California housing data collected in the 1990 US Census. We implemented Bezdek's algorithm for fuzzy c-means clustering in SAS/IML and SAS/MACRO and discussed the concept of clustering in the context of fuzzy set membership. We then applied several graphical techniques to determine suitable parameters for the imputation process, and performed single and multiple imputation on naturally-occurring missing values (Total Bedrooms data) and induced missing values (Median Income data). We compared the original data distribution to the imputed data distribution in each case and observed that the fuzzy c-means imputation process can accurately generate imputed values for observations that are centrally-located, but for outlying observations, the process is more conservative and the values imputed to outlying observations are not as extreme.

Based on our case study results, we conclude that fuzzy c-means clustering imputation is an excellent tool for data scientists and machine learning practitioners to use to deal with missing values in their data.

[7] The original Median Income data were 100% nonmissing, while the imputed Median Income data represent the Median Income data into which missingness was introduced artificially as described in the introduction to this section.

Appendix A – Fuzzy *c*-Means Macro

In Phase 1, the SAS macro %FCM is used to assign observations to clusters and to generate cluster centers. It produces a SAS dataset that is named in the macro variable &DSNSEED that contains the cluster centers used. It requires PROC IML. In Phase 2, the SAS macro %FCM_IMPUTE is used to perform imputation to all observations containing missing values. It uses the &DSNSEED_FCM dataset produced by %FCM in Phase 1. It also requires PROC IML.

The %FCM heading and parameter definitions are:

```
%macro FCM( DSNIN           /* name of input dataset containing observations to be clustered                          */

          , DSNOUT=         /* [optional] name of output dataset containing original vars and membership fcns per obs */
          , DSNSEED=        /* [optional name of output dataset containing cluster centers                            */
          , DSNSTAT=        /* [optional] name of output dataset containing # clus, # obs, iteration history          */
          , CLASS=          /* [optional] name of variable which contains discrete value of cluster                   */
          , M=2             /* [optional] fuzzification parameter, which indicates degree of fuzzy overlap            */
                            /* btwn clusters                                                                          */
          , MAX_ITER=100    /* [optional] maximum # of iterations                                                     */
          , MIN_IMPROV=1e-4 /* [optional] minimum improvement in objective fcn btwn iterations                        */
          , N_CLUS=2        /* number of clusters to create                                                           */
          , PRINT=Y         /* [optional] flag to control printing of final results                                   */
          , PRINT_MF=10     /* [optional] number of membership functions to print after convergence                   */
          , RAN_SEED=2020   /* [optional] random number seed for fuzzy partition matrix initialization                */
          , VARS=           /* [optional] list of numeric variables to use in clustering                              */
          ) ;
```

The %FCM_IMPUTE heading and parameter definitions are:

```
%macro FCM_IMPUTE( DSNIN_IMPUTE    /* name of input dataset containing observations with missing values to be imputed        */
                 , DSNSEED_FCM     /* name of input dataset containing cluster centers from prior %FCM clustering            */
                 , DSNOUT_IMPUTE=  /* [optional] name of output dataset containing original vars and membership fcns per obs */
                 , DSNSEED_IMPUTE= /* [optional name of output dataset containing cluster centers                            */
                 , DSNSTAT_IMPUTE= /* [optional] name of output dataset containing # clus, # obs, # vars, iteration history  */
                 , CLASS=          /* [optional] name of variable which contains discrete value of cluster label             */
                 , M=2             /* [optional] fuzzification parameter, which indicates degree of fuzzy overlap            */
                                   /* btwn clusters                                                                          */
                 , MAX_ITER=100    /* [optional] maximum # of iterations                                                     */
                 , MIN_IMPROV=1e-4 /* [optional] minimum improvement in objective fcn btwn iterations                        */
                 , N_CLUS=2        /* number of clusters to create                                                           */
                 , PRINT=Y         /* [optional] flag to control printing of final results                                   */
                 , PRINT_MF=10     /* [optional] number of membership functions to print after convergence                   */
                 , RAN_SEED=2020   /* [optional] random number seed for fuzzy partition matrix initialization                */
                 , VARS=           /* [optional] list of numeric variables to use in clustering                              */
                 ) ;
```

For imputation, the order of the variables in the %FCM_IMPUTE macro parameter &VAR must be identical to the order in the %FCM macro.

Appendix B – SAS Code to Perform Fuzzy *c*-Means Multiple Imputation

```
/* purpose: perform multiple imputation
 *          total_bedrooms is MCAR: unknown mechanism of missingness
 *          median_income  is MAR :   known m-o-m
 */

options mlogic mprint sgen linesize=200 pagesize=9999 ;

%include "C:\Users\Username\Documents\My SAS Files\Missing Value Imputation\SASCode\fcm.sas" ;
%include "C:\Users\Username\Documents\My SAS Files\Missing Value Imputation\SASCode\fcm_impute.sas" ;

libname CALHOUS "C:\Users\Username\Documents\My SAS Files\Missing Value Imputation\SASData" ;

%let DSNIN = CALHOUS.CALHOUS ;
%let VARS  = longitude latitude housing_median_age total_rooms total_bedrooms population households median_income
             median_house_value ;

proc univariate data=&DSNIN( keep= ocean_proximity median_income ) ;
    /* compute percentiles of median_income by ocean_proximity */

    class ocean_proximity ;
    var median_income ;

    output out=pctiles n=n p1=p1 p5=p5 p10=p10 p25=p25 p50=pt0 p75=p75 p90=p90 p95=p95 p99=p99 ;
run ;

/******************************************************************************/

/* create MAR missing values for median_income: set random sample of top 1% median_income to missing
 * for all proximities but ISLAND
 *
 * note: values of 99th %-ile taken by hand from pctiles, generated above
 */

data missing_calhous_fcm ;
    /* create missing-at-random data: if median_income > 99%-ile, randomly set 90% of them = missing */

    set &DSNIN( keep=missing_flg ocean_proximity &VARS ) ;

    med_inc_missing_flg = 0 ;

    select( ocean_proximity ) ;
    when ( '<1H OCEAN'  ) if median_income >= 111077 then
                          if ranuni( 2020 ) > 0.10   then do ; median_income = . ; med_inc_missing_flg = 1 ; end ;

    when ( 'INLAND'     ) if median_income >=  77876 then
```

```
                          if ranuni( 2020 ) > 0.10   then do ; median_income = . ; med_inc_missing_flg = 1 ; end ;

    when ( 'NEAR BAY'   ) if median_income >= 115706 then
                          if ranuni( 2020 ) > 0.10   then do ; median_income = . ; med_inc_missing_flg = 1 ; end ;

    when ( 'NEAR OCEAN' ) if median_income >= 113074 then
                          if ranuni( 2020 ) > 0.10   then do ; median_income = . ; med_inc_missing_flg = 1 ; end ;
    otherwise ;
    end ;
run ;

proc freq data=missing_calhous_fcm ;
    /* data check: show %-ages of missing by ocean_proximity */

    table ocean_proximity * med_inc_missing_flg ;
    table ocean_proximity * missing_flg         ;
run ;

/******************************************************************************/

data nonmissing_calhous_fcm ;
    set missing_calhous_fcm ;

    /* total_bedrooms has MCAR missing values and median_income has MAR missing values */

    if missing( total_bedrooms ) | missing( median_income ) then delete ;
run ;

proc freq data=nonmissing_calhous_fcm ;
    /* data check: show %-ages of missing by ocean_proximity */

    table ocean_proximity * med_inc_missing_flg ;
    table ocean_proximity * missing_flg         ;
run ;

/******************************************************************************/

ods listing ;
ods html close ;

/* phase 1: exploration */

%FCM( nonmissing_calhous_fcm
    , dsnout=nonmissing_calhous_fcm_out
    , dsnseed=nonmissing_calhous_fcm_seed
    , dsnstat=nonmissing_calhous_fcm_stat
    , m=2
    , max_iter=300
    , min_improv=.01
    , n_clus=10
    , print=y
    , print_mf=5
    , vars=&VARS
    )

/* phase 2: imputation */

ods listing ;
ods html close ;

%FCM_IMPUTE( missing_calhous_fcm
           , nonmissing_calhous_fcm_seed
           , dsnout_impute=missing_calhous_fcm_impute_out
           , dsnseed_impute=missing_calhous_fcm_impute_seed
           , class=
           , m=1.6
           , max_iter=200
           , min_improv=.01
           , n_clus=10
           , print=y
           , print_mf=5
           , vars=&VARS
           )

/******************************************************************************/

data comparison ;
    /* create dataset of original median_income data with imputed median_income data appended */

    set &DSNIN                        ( in=in1 keep=ocean_proximity median_income )
        missing_calhous_fcm_impute_out( in=in2 keep=ocean_proximity med_inc_missing_flg median_income
                                          rename=( median_income=median_income_imput )
                                      )
                                                    ;

    if in1 then med_inc_missing_flg = 0 ; else med_inc_missing_flg = 1 ;

    if in2 then median_income = median_income_imput ;

    drop median_income_imput ;
run ;

/******************************************************************************/

proc format ;
    value missing
    0='Nonmissing'
    1='Imputed'
    ;
run ;

goptions reset=all ;

options papersize=( 8.5in 8.5in ) leftmargin=.1cm rightmargin=.1cm bottommargin=.1cm topmargin=.1cm ;

ods listing close ;
ods pdf file="C:\Users\Username\Documents\My SAS Files\Missing Value Imputation\SASCode\04.1 FCM_IMPUTE CA Census.pdf"
        style=statistical ;
ods graphics on ;

title 'Comparison Between Original and Imputed Median Income Data' ;

proc sgpanel data=comparison  ;
    format med_inc_missing_flg missing. ;
    label med_inc_missing_flg='00'x
          median_income      ='Median Income Per Census Block'
          ocean_proximity    ='Ocean Proximity'
          ;

    panelby ocean_proximity / novarname skipemptycells ;
    vbox median_income / group=med_inc_missing_flg ;
    rowaxis grid logbase=10 logstyle=linear type=log values=( 4000 40000 400000 ) ;
run ;

proc sgplot data=comparison ;
    format med_inc_missing_flg missing. ;
    label med_inc_missing_flg='00'x
          median_income      ='Median Income Per Census Block'
          ocean_proximity    ='Ocean Proximity'
          ;

    vbox median_income / category=ocean_proximity group=med_inc_missing_flg groupdisplay=cluster ;
run ;
title ;

proc format ;
    value missing
    0='Non-missing'
    1='Imputed'
    ;
run ;

proc tabulate data=comparison format=comma7.0 noseps ;
    format med_inc_missing_flg missing. ;
    label ocean_proximity='Ocean Proximity'
          median_income ='Median Income'
          ;

    class med_inc_missing_flg ocean_proximity ;
    var median_income ;

    table ( n mean std min q1 median q3 max ), ocean_proximity * ( median_income ) * med_inc_missing_flg=''*f=comma8.
    / rts=8 indent=4 ;
run ;
title ;

ods graphics off ;
ods pdf close ;
ods listing ;
```

References

[1] https://en.wikipedia.org/wiki/Imputation_(statistics)

[2] https://www.theanalysisfactor.com/missing-data-mechanism/

[3] https://www.displayr.com/different-types-of-missing-data/#_edn1

[4] Allison, Paul D. (2010). "Missing data" Pp. 631-657 in Handbook of Survey Research, (Wright, James D. and Peter V. Marsden, eds.), Bingley, UK. Emerald Group Publishing Ltd.

[5] Roderick J. A. Little, A Test of Missing Completely at Random for Multivariate Data with Missing Values, *Journal of the American Statistical Association*, Vol. 83, No. 404 (Dec., 1988), pp. 1198-1202.

[6] Allison, Paul D. (6 Jun 2017). The Peculiarities of Missing at Random, https://statisticalhorizons.com/missing-at-random

[7] Little, Roderick J.A., and Donald B. Rubin (2002) *Statistical Analysis with Missing Data*. Wiley.

[8] Allison, Paul D. (2009). "Missing Data", The SAGE Handbook of Quantitative Methods in Psychology, Millsap, Roger E and Alberto Maydeu-Olivares (eds), https://statisticalhorizons.com/resources/articles/Allison 2009.

[9] Eekhout, Iris (2019). https://www.iriseekhout.com/missing-data/missing-data-methods/

[10] Andridge, Rebecca R. and Roderick J.A. Little (2010). "A Review of Hot Deck Imputation for Survey Non-response", https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3130338/

[11] https://www.theanalysisfactor.com/multiple-imputation-in-a-nutshell/

[12[https://stats.idre.ucla.edu/sas/seminars/multiple-imputation-in-sas/mi_new_1/

[13] Bezdek, James C., Robert Ehrlich, William Full (1984). "FCM: The Fuzzy C-Means Clustering Algorithm", Computers & Geosciences, Vol. 10, No. 2-3, pp. 191-203.

[14] Hathaway, Richard J., James C. Bezdek (2001). "Fuzzy c-Means Clustering of Incomplete Data", IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics, Vol. 31, No. 5.

[15] http://www.math.wisc.edu/~angenent/519.2014s/coursenotes/picard

Acknowledgement

We thank Jim Bezdek, Bruce Gilsen, Mark Leventhal, and Joseph Naraguma for their generosity in taking the time to review this document.

Contact Information

Your comments and questions are valued and encouraged. The code is available upon request. Contact the author at:

Name: Ross Bettinger
Enterprise: Consultant
E-mail: rsbettinger@gmail.com