Data Science / Analytics network meeting - 13/3-2024

“Good decisions require good data”

Pia Skare Rønnevik, Customer Success Manager, FANS
pia.roennevik@sas.com

(1) I just discovered that the data system we have been working on for the last five years has major data quality problems

(2) That is why I treat data systems the same way I do sausage – I do not want to know what is inside either one.

(3) Ouch!! That's why I am a vegetarian!

Copilot – What to think of in data preparation for analytics?

1. Data Collection and Sourcing:
- Gather relevant data from various sources. Ensure data quality by addressing missing values, duplicates, and inconsistencies.
- Verify the data's accuracy and reliability.

2. Data Cleaning and Preprocessing:
- Handle missing data: Impute missing values using techniques like mean, median, or regression.
- Remove duplicates and outliers.
- Standardize or normalize features to a common scale.
- Address data inconsistencies (e.g., typos, variations in formatting).

3. Feature Engineering:
- Create new features from existing ones (e.g., aggregations, transformations).
- Select relevant features based on domain knowledge and statistical significance.

4. Exploratory Data Analysis (EDA):
- Visualize data distributions, correlations, and patterns.
- Identify relationships between variables.
- Detect anomalies or unexpected behavior.

5. Data Transformation:
- Apply necessary transformations (e.g., log, square root) to achieve normality.
- Encode categorical variables (one-hot encoding, label encoding).
- Consider dimensionality reduction techniques (PCA, t-SNE) if needed.

6. Data Splitting:
- Divide the dataset into training, validation, and test sets.
- Ensure that the split maintains the distribution of target variables.

7. Handling Imbalanced Data:
- Address class imbalance using techniques like oversampling, undersampling, or synthetic data generation.

8. Feature Selection:
- Use statistical tests (e.g., chi-squared, ANOVA) or feature importance scores (from models) to select relevant features.

9. Data Validation and Quality Assurance:
- Validate data against business rules and domain knowledge.
- Monitor data quality over time.

10. Document the Process:
- Maintain clear documentation of data preparation steps.
- Record assumptions, transformations, and any decisions made during the process

Copilot – is it preferable to use categorical variables instead of continuous

1. Continuous Variables:

- Continuous data represents measurements that can take any value within a range (e.g., height, temperature, weight).

- Advantages:

- Provides more information due to the fine-grained nature of measurements.
- Allows for precise modeling of relationships.
- Well-suited for linear regression and other models that assume linearity.

- Disadvantages:

- Assumes a linear relationship, which may not hold in all cases.
- Sensitive to outliers.
- Requires more data points for accurate estimation.

- When to use:

- Use when the relationship with the response variable is expected to be linear.
- Suitable for regression models, correlation analysis, and ANOVA.

2. Categorical Variables:

- Categorical data represents distinct categories or groups (e.g., gender, color, program type).

- Advantages:

- Simplifies complex relationships by grouping data into meaningful categories.
- Handles non-linear relationships more effectively.
- Robust to outliers.

- Disadvantages:

- May lose information due to grouping.
- Requires encoding (e.g., one-hot encoding) for modeling.
- Limited to the defined categories.

- When to use:

- Use when the relationship is non-linear or when there are clear groupings.
- Useful for decision trees, random forests, and logistic regression.
- Especially valuable when capturing interactions or nonlinear effects.

Considerations:

- Interaction Effects: Categorical variables allow modeling of interactions (e.g., gender and age).
- Data Context: Consider the context and domain knowledge. Some variables may naturally be categorical (e.g., program type).
- Data Exploration: Explore relationships visually (scatter plots, histograms) to decide how to treat variables.

Copilot - is it best to have continuous or binary target?

1. Continuous Target (Regression):

- Definition: A continuous target variable can take any numeric value within a range. It has a natural order and can be measured with precision.

- Use Cases:

- When you want to predict a numerical value, such as predicting house prices, temperature, or stock prices.
- In scenarios where the relationship between predictors and the response is linear or can be approximated by linear models.
- When you need to estimate the mean change in the dependent variable given changes in independent variables.

- Advantages:

- Provides more information due to the greater number of possible values.
- Allows for a wider variety of analyses.
- Can handle non-linear relationships using techniques like polynomial regression1.

2. Binary Target (Classification):

- Definition: A binary target variable has only two possible outcomes (e.g., yes/no, true/false, 1/0).

- Use Cases:

- When you want to perform classification tasks, such as spam detection, disease diagnosis, or customer churn prediction.
- In scenarios where you need to assign instances to predefined classes.
- When the outcome of interest is binary (e.g., whether a customer will buy a product or not).

- Advantages:

- Simplicity: Binary targets are straightforward to interpret.
- Commonly used in practical applications.
- Suitable for algorithms like logistic regression, support vector machines, and decision trees

3. Hybrid Approaches:

Sometimes, you might encounter situations where a continuous target can be transformed into a binary one. For instance:

- Threshold-based classification: Convert a continuous target (e.g., probability) into binary outcomes by setting a threshold (e.g., 0.5).
- Risk prediction: Predicting the likelihood of an event (e.g., defaulting on a loan) using continuous scores and then categorizing into low/high risk.
- Ordinal regression: When the target has ordered categories (e.g., low/medium/high), you can use ordinal regression.

Consideration: If you can collect continuous data, it’s often better to do so because it provides more information.

VIMO!

- Valid
- Invalid
- Missing
- Outlier

Missing values and outliers

Missing values:

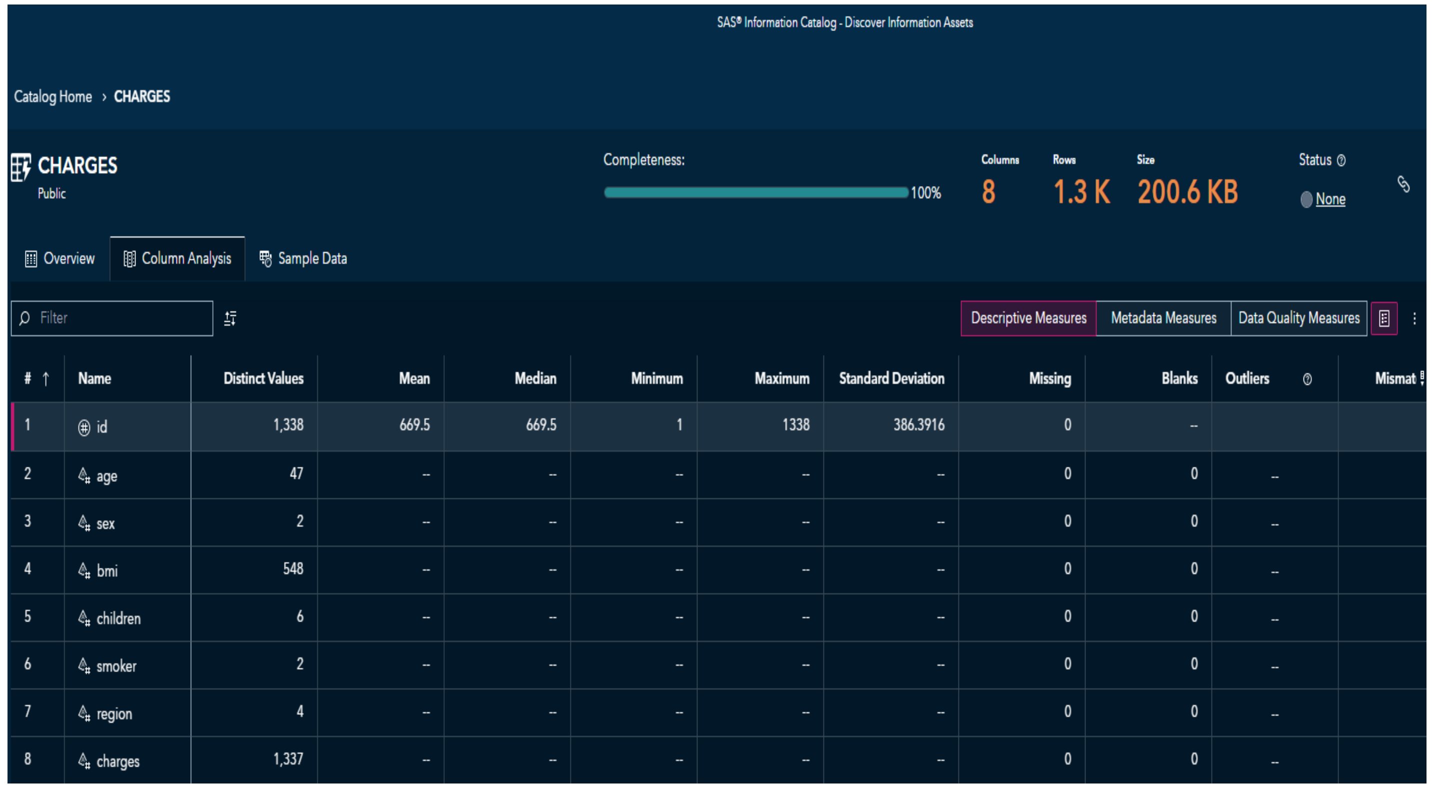
- is due to information loss, dropouts and none-responses.
- give smaller sample size, could compromises the study and entail bias.
- could be ignored or imputed based on the present values.

Outliers:

- are extreme values that's outside the overall pattern of a distribution of variables.
- arises from respons and data errors.
- Could lead to under- or over-estimated values
- If you are sure the outliers are wrong, then ignore them or impute.

How to check VIMO!

Information catalog on Viya


SAS® Information Catalog - Discover Information Assets

Catalog Home › CHARGES

CHARGES
Public

Completeness: 100%

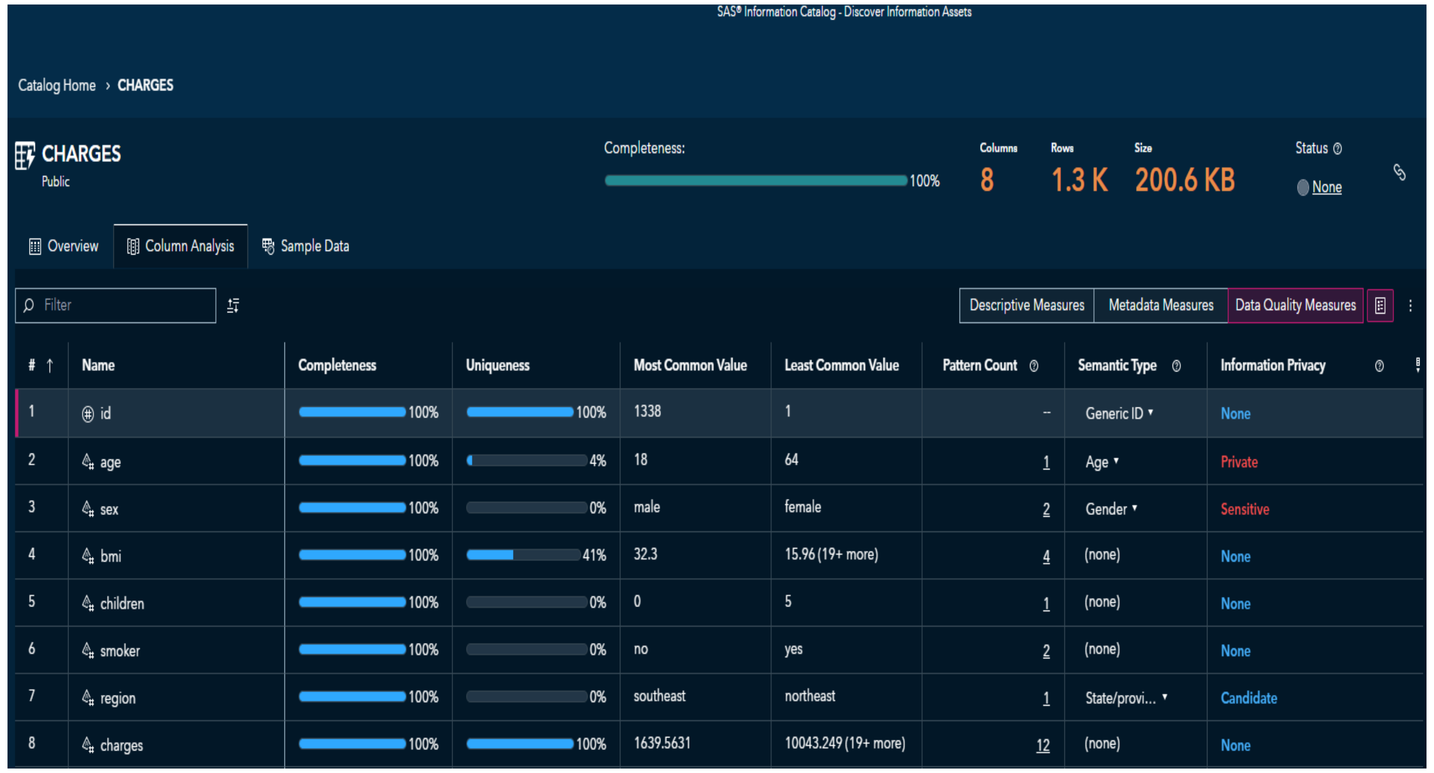
Columns 8 | Rows 1.3 K | Size 200.6 KB | Status None

Overview | Column Analysis | Sample Data

Filter

Descriptive Measures | Metadata Measures | Data Quality Measures

#	Name	Distinct Values	Mean	Median	Minimum	Maximum	Standard Deviation	Missing	Blanks	Outliers	Mismat
1	id	1,338	669.5	669.5	1	1338	386.3916	0	--		
2	age	47	--	--	--	--	--	0	0	–	
3	sex	2	--	--	--	--	--	0	0	–	
4	bmi	548	--	--	--	--	--	0	0	–	
5	children	6	--	--	--	--	--	0	0	–	
6	smoker	2	--	--	--	--	--	0	0	–	
7	region	4	--	--	--	--	--	0	0	–	
8	charges	1,337	--	--	--	--	--	0	0	–	


SAS® Information Catalog - Discover Information Assets

Catalog Home › CHARGES

CHARGES
Public

Completeness: 100%

Columns 8 | Rows 1.3 K | Size 200.6 KB | Status None

Overview | Column Analysis | Sample Data

Filter

Descriptive Measures | Metadata Measures | Data Quality Measures

#	Name	Completeness	Uniqueness	Most Common Value	Least Common Value	Pattern Count	Semantic Type	Information Privacy
1	id	100%	100%	1338	1	--	Generic ID	None
2	age	100%	4%	18	64	1	Age	Private
3	sex	100%	0%	male	female	2	Gender	Sensitive
4	bmi	100%	41%	32.3	15.96 (19+ more)	4	(none)	None
5	children	100%	0%	0	5	1	(none)	None
6	smoker	100%	0%	no	yes	2	(none)	None
7	region	100%	0%	southeast	northeast	1	State/provi...	Candidate
8	charges	100%	100%	1639.5631	10043.249 (19+ more)	12	(none)	None

Visual Analytics

- Could check for missing values by just looking on histogram for the measure variables, and bar chart for the categorical variables. Can easily impute missing variables with calculated item.
- Could check for outliers by using box-plot

The data - medical charges of a specific patient

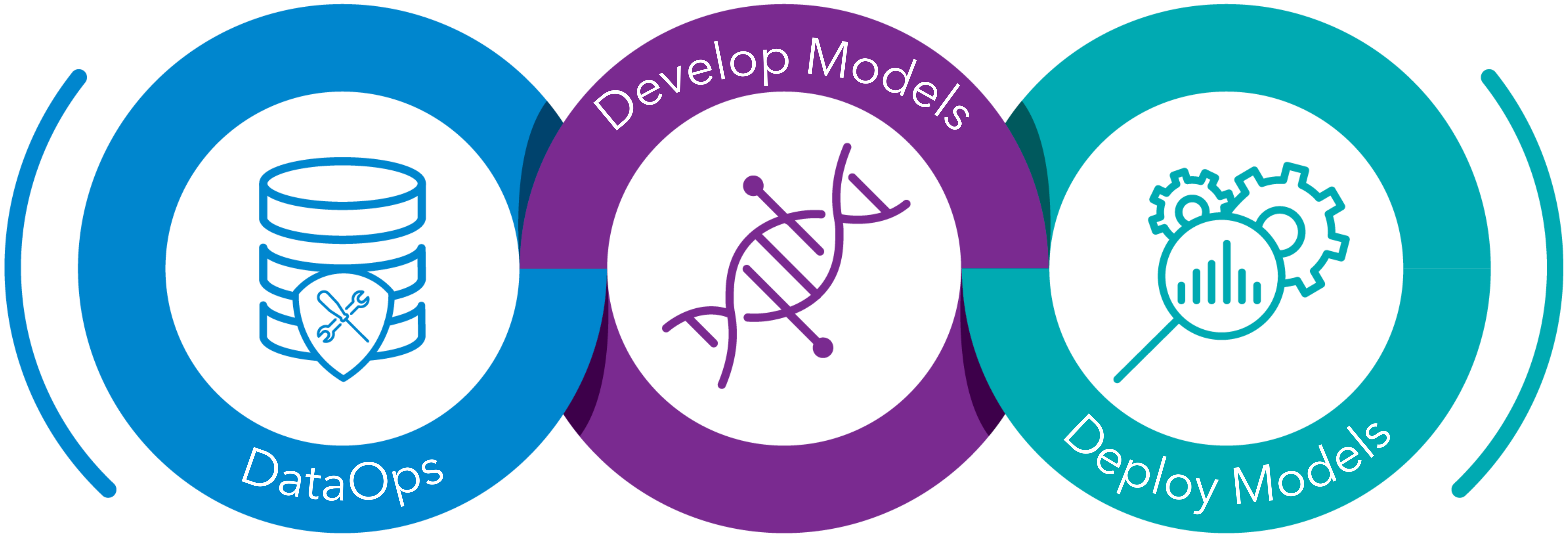
- **age:** age of primary beneficiary
- **sex:** insurance contractor gender, female, male
- **bmi:** body mass index, providing an understanding of body, weights that are relatively high or low relative to height objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
- **children:** number of children covered by health insurance / Number of dependents
- **smoker:** smoking
- **region:** the beneficiary's residential area in the US, northeast, southeast, southwest, northwest
- **charges (insurance price):** Individual medical costs billed by health insurance

https://www.kaggle.com/datasets/nanditapore/medical-cost-dataset

The Analytics Life Cycle

QUESTION

What is the medical expenses of specific patient?



Value

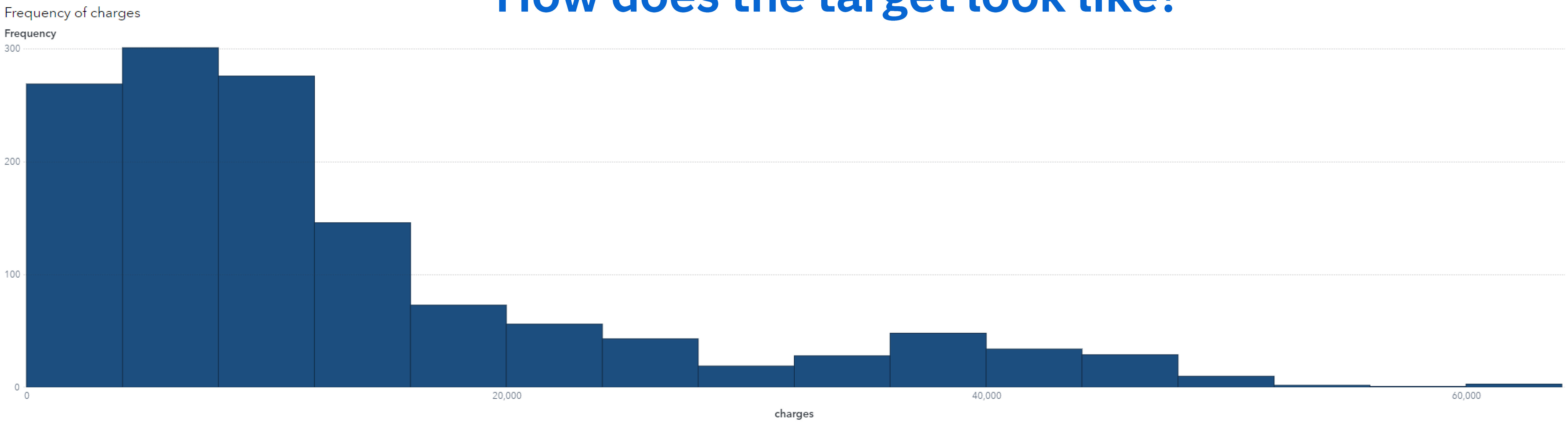
Given certain factors, we could say what the medical expenses is for a patient

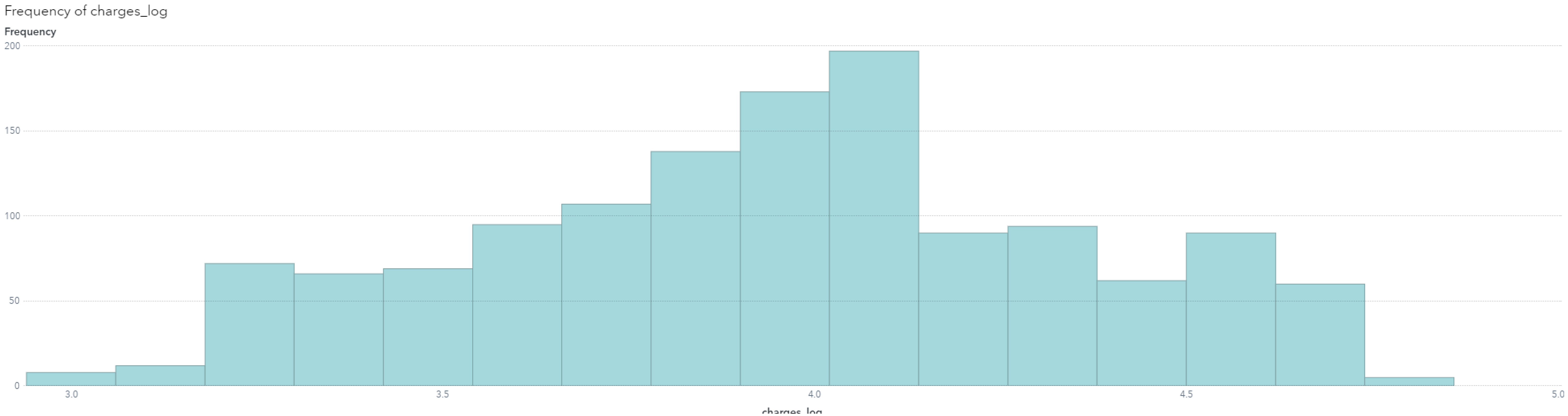
- Access
- Prepare
- Visualize
- Govern

- Build
- Train
- Evaluate
- Select

- Validate
- Deploy
- Monitor
- Embed

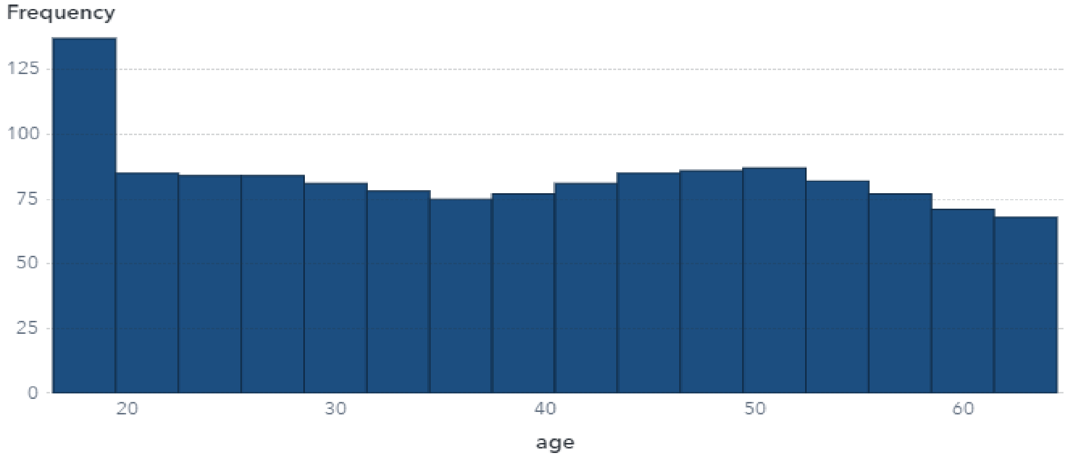
Data preparation in VA –
How does the target look like?



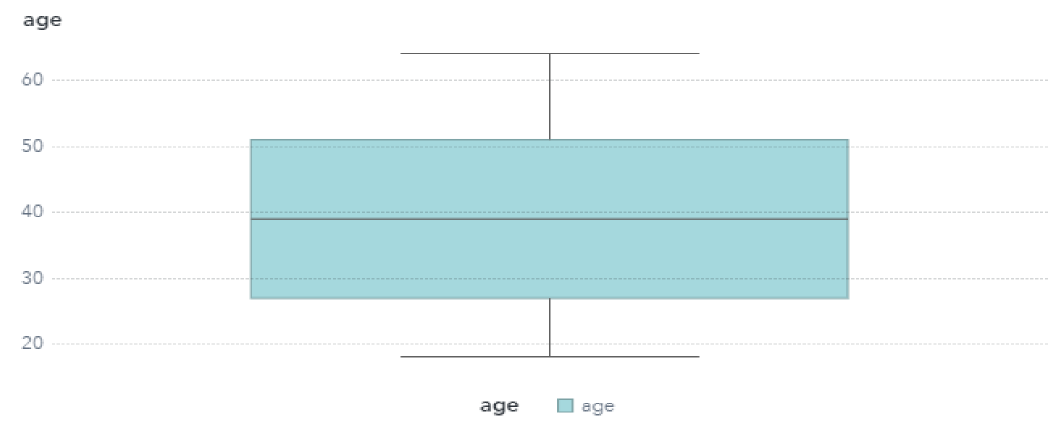


Data preparation in VA –
How does the measure variables look like?

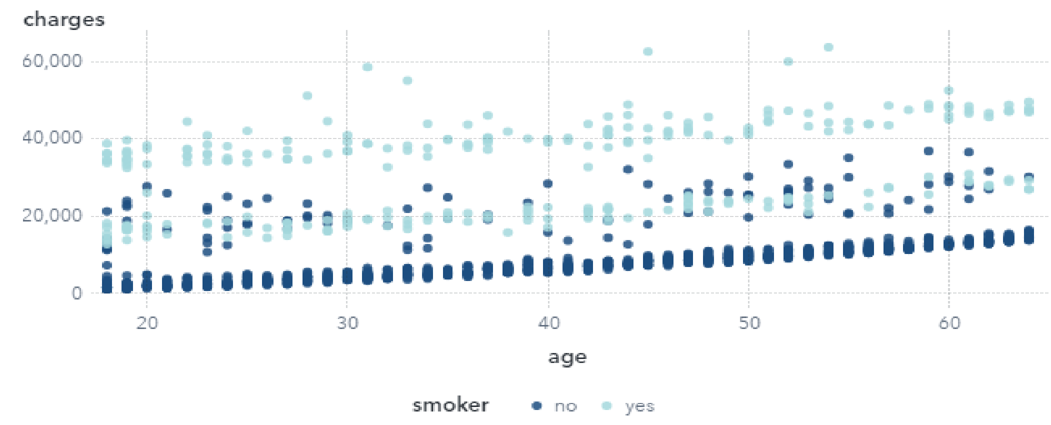
Frequency of age



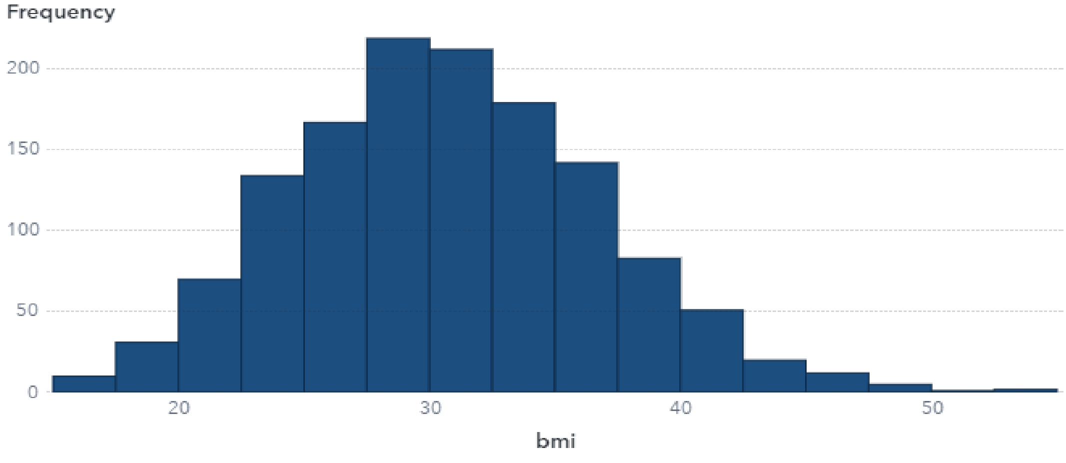
age



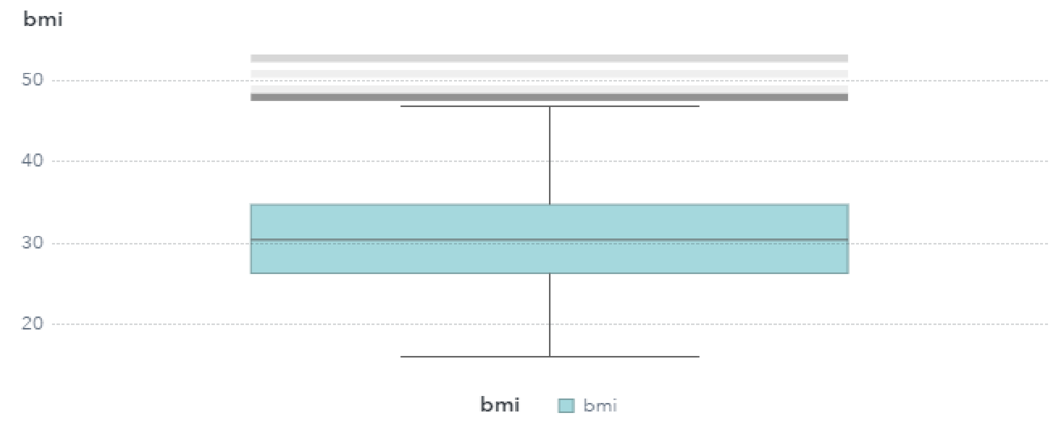
Scatter Plot of Selected Measures



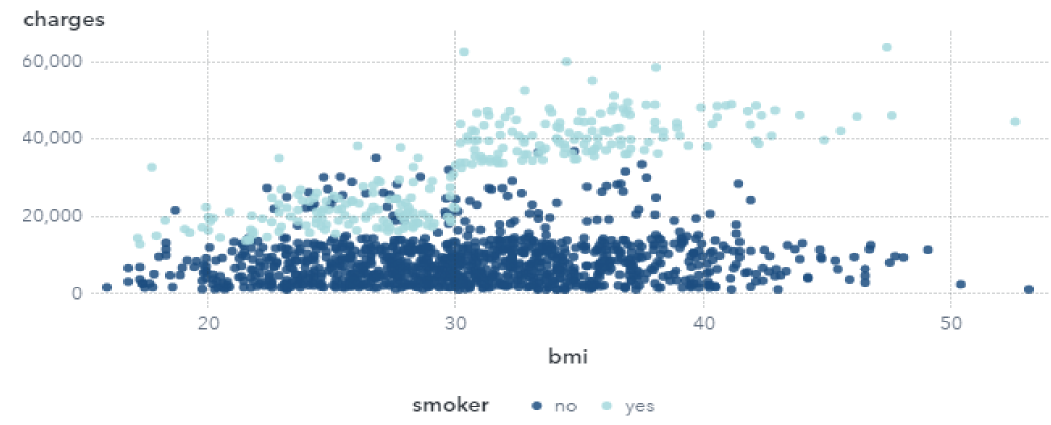
Frequency of bmi



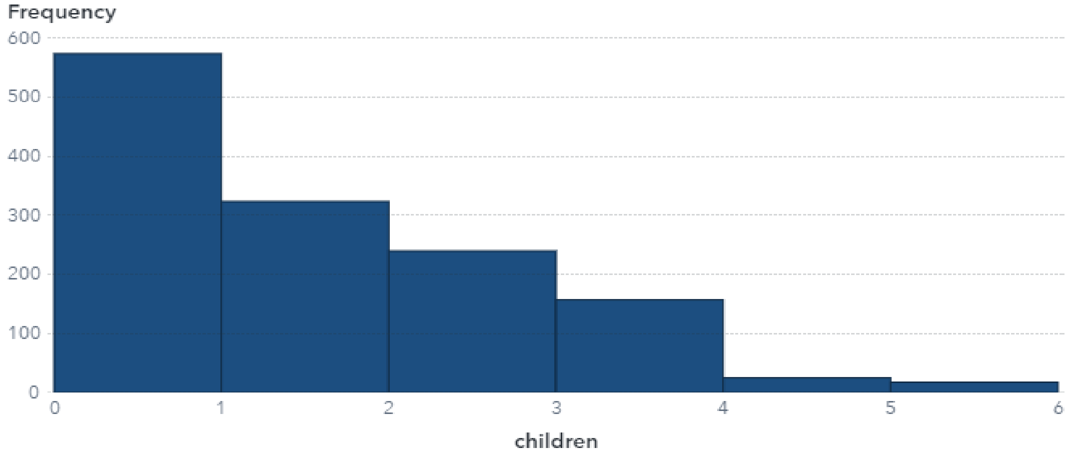
bmi



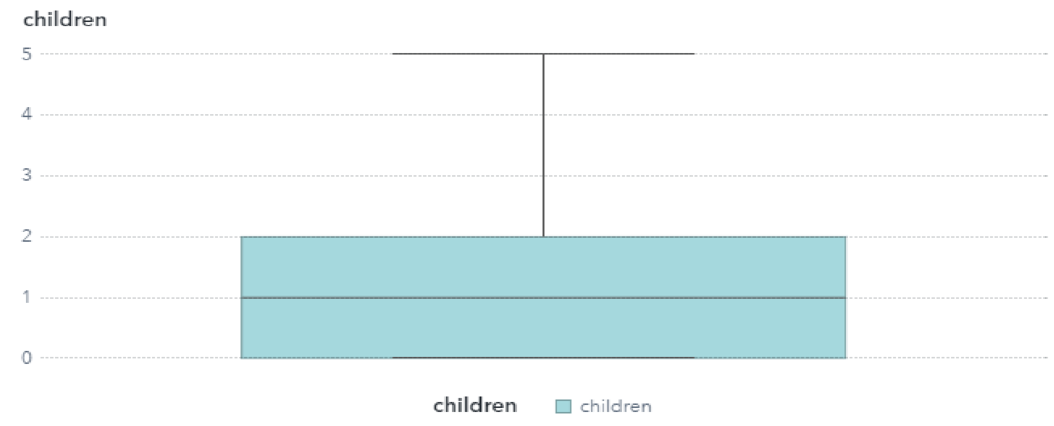
Scatter Plot of Selected Measures



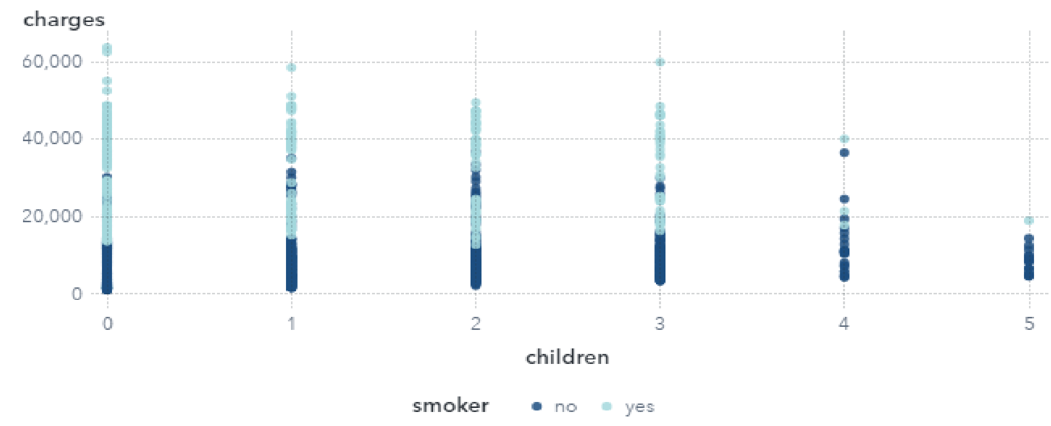
Frequency of children



children

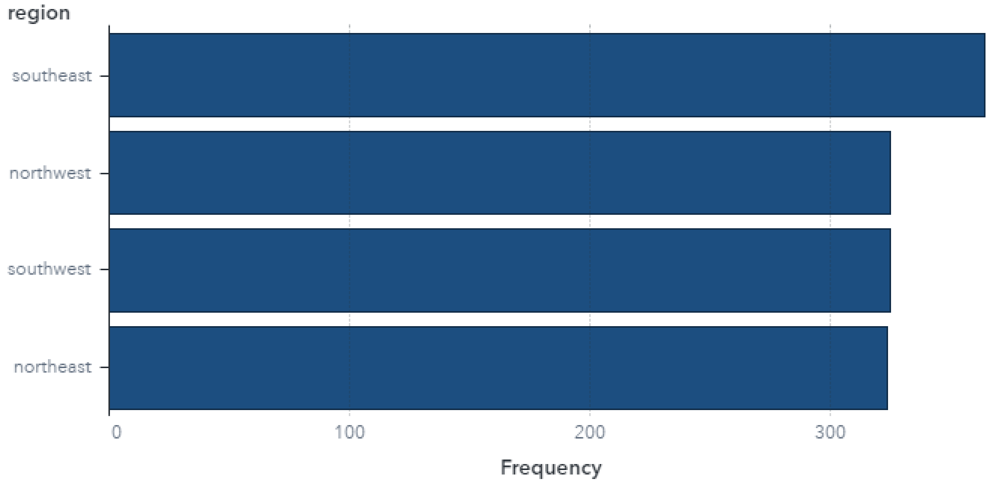


Scatter Plot of Selected Measures

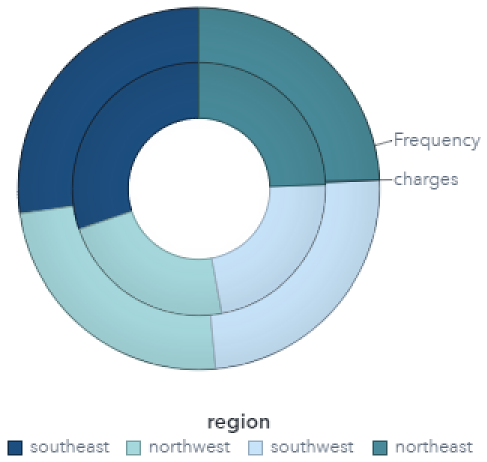


Data preparation in VA –
How does the categorical variables look like?

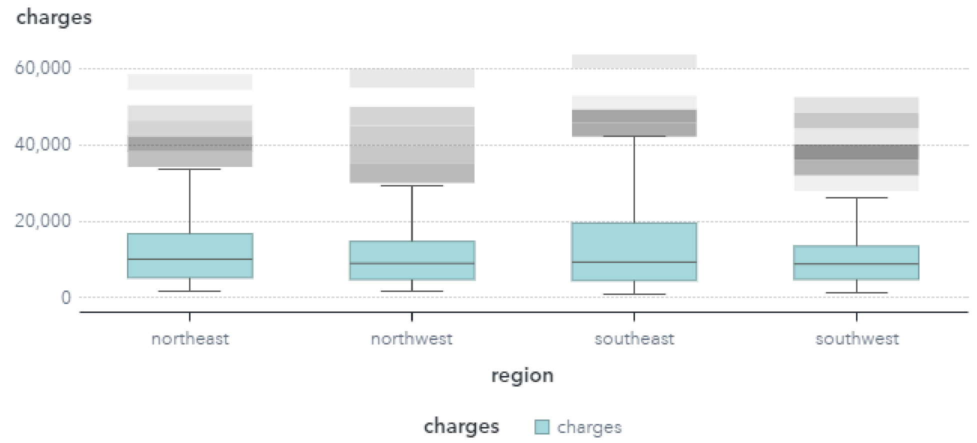
Frequency of region



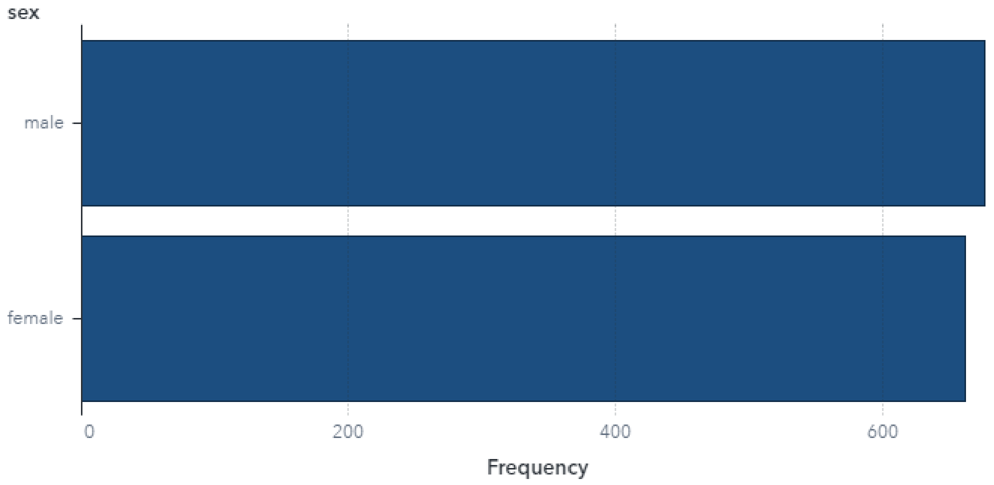
Frequency, charges by region



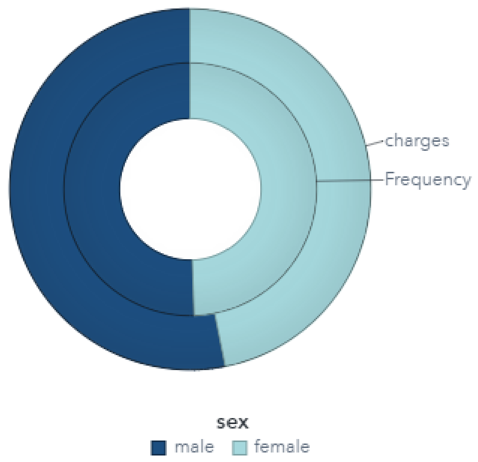
charges by region



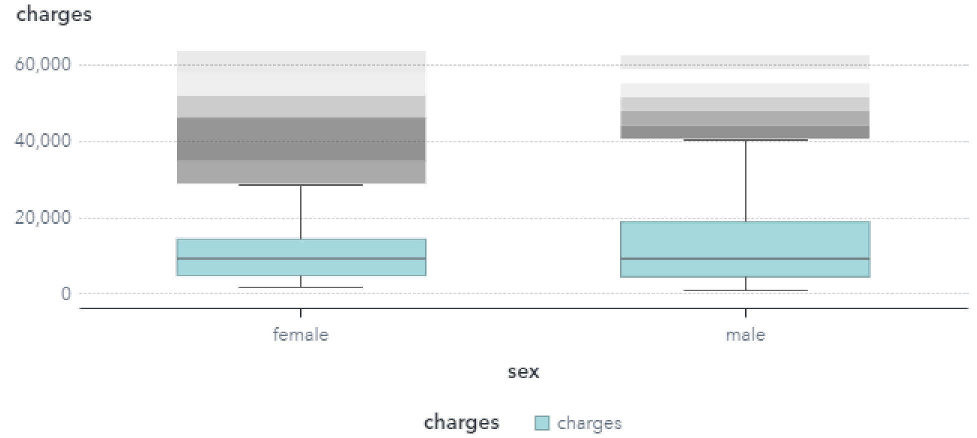
Frequency of sex



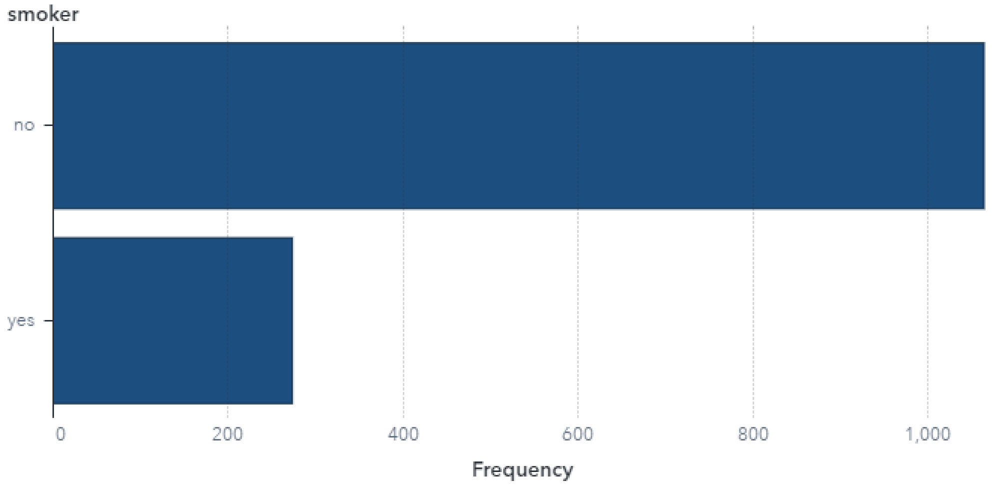
charges, Frequency by sex



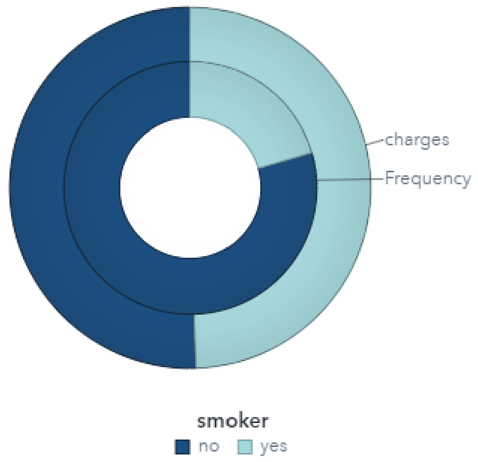
charges by sex



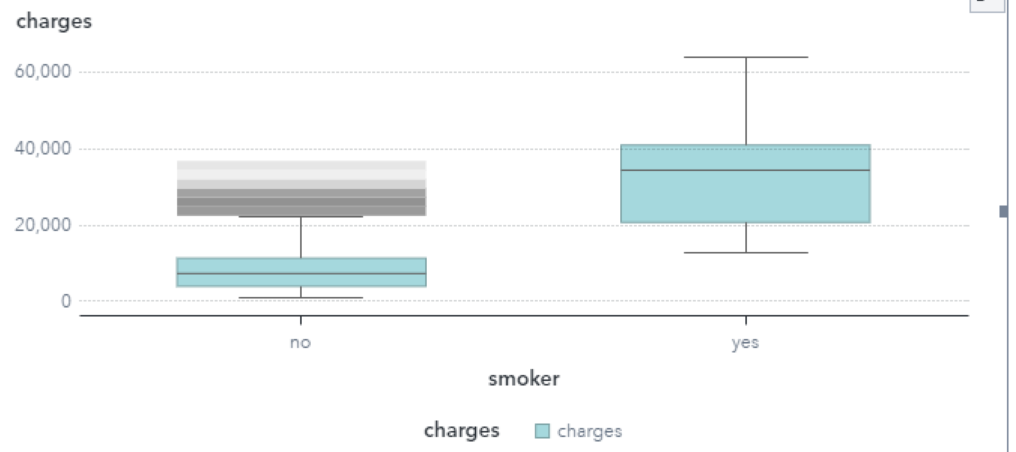
Frequency of smoker



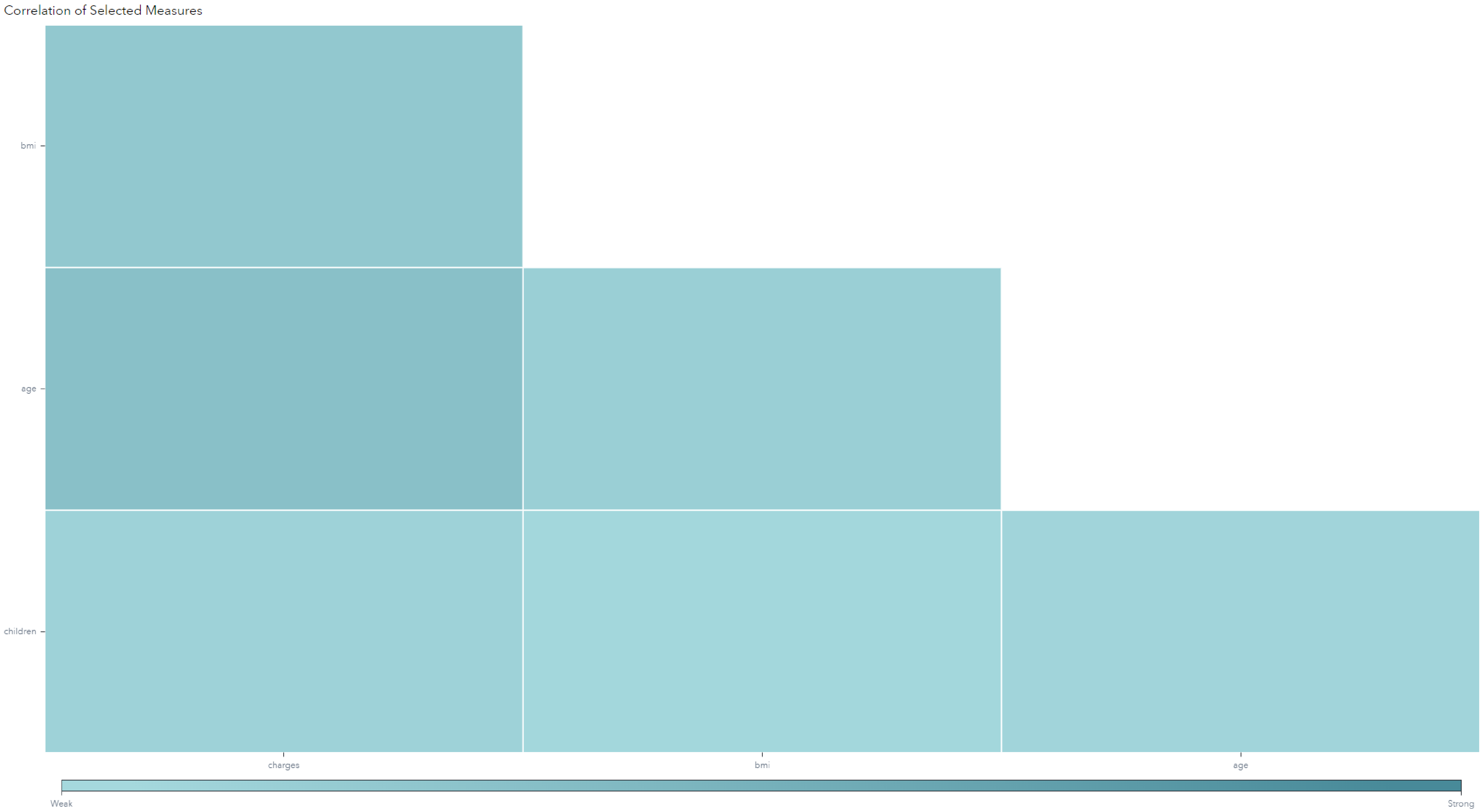
charges, Frequency by smoker



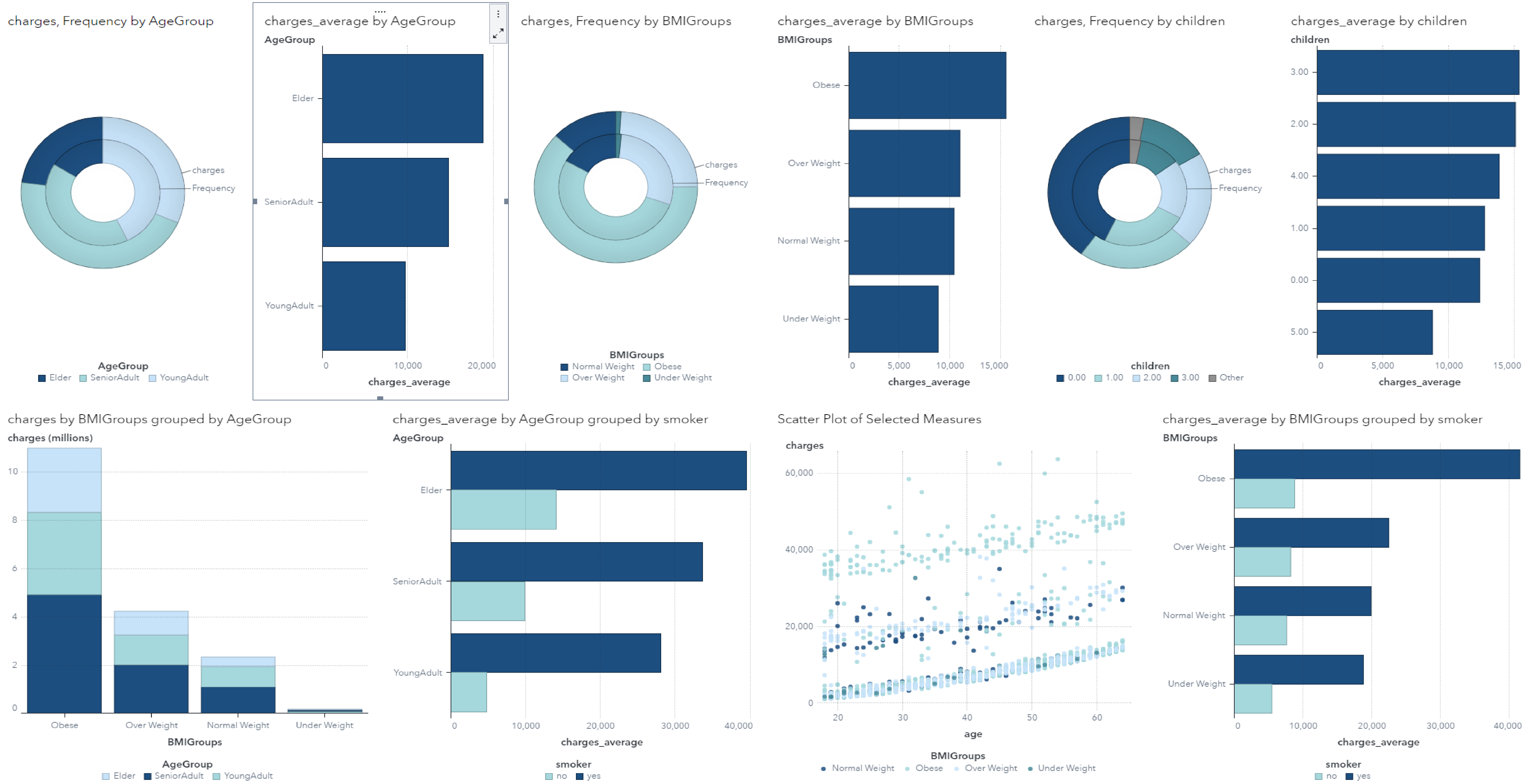
charges by smoker



Data preparation in VA –
The correlation between the measure variables?



Data preparation in VA –
Re-define the measure variables as categorical



References

- Process for Data Quality Assurance at Manitoba Centre for Health Policy (MCHP): https://umanitoba.ca/manitoba-centre-for-health-policy/sites/manitoba-centre-for-health-policy/files/2021-11/data-quality-framework-document.pdf

- Statistical data preparation: management of missing values and outliers: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5548942/