

Analyze the internet data of www.datadb.com

Business Analytic Foundation with SAS Tools & Excel- Solutions



Solutions

Disclaimer: In Business Analytics, there are different ways of solving the same set of problems, we are just presenting one. Feel free to explore other ways of answering these questions.

1. The team wants to analyze each variable of the data collected through data summarization to get a basic understanding of the dataset and to prepare for further analysis.

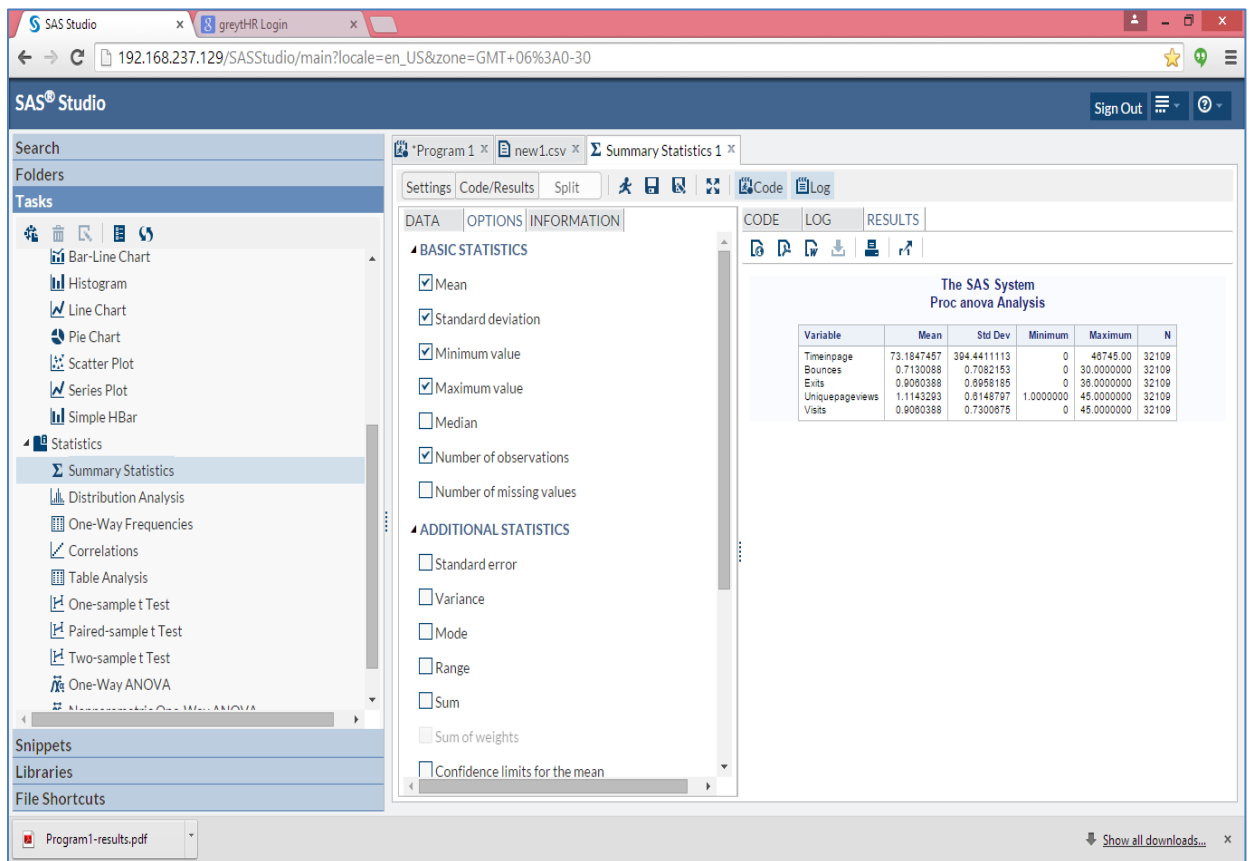
To understand how the data is distributed in the dataset and the kind of variables present along with their count, what is their maximum and minimum value, etc., a summarization of the data is done.

Code:

Click 'summary' and specify the analysis variables—time on page, bounces, exits, unique pageviews, and visits.

Result:

From the result of the summarized dataset, it is observed that the numerical data includes the information of maximum, minimum, mean, standard deviation, and N of the data. N gives the total number of data under each variable. This will help us in further analysis.



The screenshot displays the SAS Studio web interface. The left sidebar shows the 'Tasks' pane with 'Summary Statistics' selected under the 'Statistics' category. The main window shows the 'Summary Statistics 1' results for the dataset 'new1.csv'. The 'BASIC STATISTICS' section is expanded, showing the following statistics:

- ☒ Mean
- ☒ Standard deviation
- ☒ Minimum value
- ☒ Maximum value
- ☐ Median
- ☒ Number of observations
- ☐ Number of missing values

The 'ADDITIONAL STATISTICS' section is also visible, with the following options:

- ☐ Standard error
- ☐ Variance
- ☐ Mode
- ☐ Range
- ☐ Sum
- ☐ Sum of weights
- ☐ Confidence limits for the mean

The 'RESULTS' pane on the right displays the output of the 'Proc anova Analysis'.

Variable	Mean	Std Dev	Minimum	Maximum	N
Timeinpage	73.1847457	394.4411113	0	48745.00	32109
Bounces	0.7130088	0.7082153	0	30.0000000	32109
Exits	0.9080388	0.8958185	0	36.0000000	32109
Uniquepageviews	1.1143283	0.8148797	1.0000000	45.0000000	32109
Visits	0.9080388	0.7300875	0	45.0000000	32109

- As mentioned earlier, a unique page view represents the number of sessions during which that page was viewed one or more times. A visit counts all instances, no matter how many times the same visitor may have been to your site. So the team needs to know whether the unique page view value depend on visits.

To analyze the variables that affect the unique pageviews to the website, it is needed to perform ANOVA on the unique pageviews and the other variables.

Code:

Tasks -> Statistics -> One Way ANOVA

Result:

We can infer from the results that the unique pageviews variable has a significant impact on the variables—source group, bounces, continent, visits, and time on page. So we can conclude that unique page values depend on visits.

The screenshot shows the SAS Studio interface. The left sidebar contains a 'Folders' section with 'My Folders' and a list of files including 'sasuser.v94', 'anova.csv', 'drug.treatment.csv', 'health.csv', 'HOUSEHOLD_PRICE_ANALYSIS.csv', 'ins.csv', 'loan.csv', 'Medication.csv', 'new.csv', 'new1.csv', 'news.source.csv', 'regression.csv', and 'SALES VS TEMP.csv'. Below this is a 'Tasks' section with 'Snippets', 'Libraries', and 'File Shortcuts'. The main window displays the 'Program 1' results, showing the 'Proc ANOVA Analysis' output. The output includes a table of source statistics and a table of variable statistics.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1402	12139.29708	8.65856	Infy	<.0001
Error	30706	0.00000	0.00000		
Corrected Total	32108	12139.29708			

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Bouncesnew	13	6810.55433	754.85803	Infy	<.0001
Continent	6	34.84010	5.77335	Infy	<.0001
Exits	15	11185.56462	745.70431	Infy	<.0001
Sourcegroup	8	137.21818	17.15227	Infy	<.0001
Timeinpage	1344	5150.44904	3.83218	Infy	<.0001
Visits	10	11476.00715	717.25045	Infy	<.0001

- Please find out the probable factors from the dataset which could affect the exits. Exit Page Analysis is usually required to get an idea of why a user is leaving the website for a session and moving to another site. Please keep in mind that exits should not be confused with bounces.

To understand and analyze the factors that are related to exits from the site, we need to use ANOVA.

Code:

Tasks -> Statistics -> One-way ANOVA

Result:

From the result of ANOVA given here, we can see that all the variables here, such as, source group, bounces, unique pageviews, continent, visits, and time on page affect the exits from the page.

The SAS System
Proc anova Analysis
The ANOVA Procedure

Dependent Variable: Exits

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1404	15545.51908	11.07231	Infy	<.0001
Error	30704	0.00000	0.00000		
Corrected Total	32108	15545.51908			

R-Square	Coeff Var	Root MSE	Exits Mean
1.000000	0	0	0.908039

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Bouncesnew	13	11041.71087	895.51622	Infy	<.0001
Continent	6	22.84088	3.80678	Infy	<.0001
Sourcegroup	8	323.99908	40.49988	Infy	<.0001
Timeonpage	1344	7079.49730	5.26525	Infy	<.0001
Uniquepageviews	17	9910.94000	583.52624	Infy	<.0001
Visits	18	10601.87988	675.11749	Infy	<.0001

- Every site wants to increase the time on page for a visitor. This increases the chances of the visitor understanding the site content better and hence there are more chances of a transaction taking place. Find the variables that possibly have an effect on the time on page.

To understand and analyze the factors that are related to the time on page of the site, we need to use ANOVA.

Code:

Tasks -> Statistics -> One-way ANOVA

Result:

From the result of ANOVA given here, we can see that continent, source group, bounces, unique.pageviews, and visits have more significance. Hence we can say that the time on page of the site is affected by the factors of continent, source group, bounces, unique pageviews, and visits.

The screenshot shows the SAS Studio interface with the following components:

- Search Panel:** Displays folders like 'Folder Shortcuts', 'My Folders', and 'sasuser.v94' containing files like 'Cluster_food.csv' and 'new.csv'.
- CODE Panel:** Shows a program named 'new.csv' with a table of data including 'Continent', 'Sourcegroup', 'Uniquepageviews', and 'Visits'.
- RESULTS Panel:** Displays the ANOVA procedure results for the dependent variable 'Timeinpage'.

The SAS System
The ANOVA Procedure

Dependent Variable: Timeinpage

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	74	536925238	7254395	52.15	<.0001
Error	32034	4458059100	139185		
Corrected Total	32108	4995484338			

R-Square	Coeff Var	Root MSE	Timeinpage Mean
0.107462	609.7720	373.0753	73.18475

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Bounces	13	201113989.6	15470306.9	111.15	<.0001
Exits	16	136492888.5	9099524.6	65.38	<.0001
Continent	5	5380609.7	1076121.9	7.73	<.0001
Sourcegroup	8	6693743.1	836717.9	6.01	<.0001
Uniquepageviews	17	103028162.5	6060480.1	43.54	<.0001
Visits	16	84115805.0	5257241.6	37.77	<.0001

5. A high bounce rate is a cause for alarm for websites that depend on visitor engagement. Help the team in determining the factors that are impacting the bounce.

To know the variables that have an impact on the target variable bounces, we need to perform regression. Here we will perform logistic regression.

Since the data for the variable bounces should be between 0 and 1, we will multiply the value of bounces with 0.01 and save it in a variable *BouncesNew*. Then perform logistic regression with BouncesNew and the independent factors.

Code:

```
ods graphics on;
proc logistic data=shoes plots=effect;
class UniquePageviews Exits;
model BouncesNew = UniquePageviews Exits;
run;
ods graphics off;
```

It can be inferred that unique pageviews and exits impact the target variable, bounces. Since every bounce is an exit, it has greater significance.

