SAS University Edition Challenge

Day 4 of 5 – Thursday, 19th July 2018

Challenge Overview

You and your friend (let's call him) Kenny, have decided to start an online website that aggregates movie and TV show data streamed by Netflix.

So far, you have conducted preliminary analysis on the individual data sets and managed to append them into one large data set. Kenny calls to inform you that he is still configuring the website's database. Meanwhile, Kenny wants you to carry out further exploratory analysis of the appended data set.

You tell him that will be easy! Let's go!

Summary of Skills Demonstrated

- Exploring data using summary statistics
- Exploring data using graph tasks

Submission Details

Submissions close **Monday, 23rd July 2018, 12:00pm**.
Submit your answers here.

Be sure to use your SAS account when you submit your solution. If you don't use your SAS account, we can't put your entry in the draw!

If you haven't registered yet, it's not too late. Click here to register now.

Also, make sure to save your answers and tasks/code somewhere safe as a backup.

SAS University Edition

You'll need to download and install SAS University Edition to complete this challenge. Use the links below to install and set up SAS University Edition.

SAS University Edition download link:
https://www.sas.com/en_au/software/university-edition/download-software.html

How do I create a shared folder in VirtualBox?
https://support.sas.com/software/products/university-edition/faq/shared_folder_virtualbox.htm

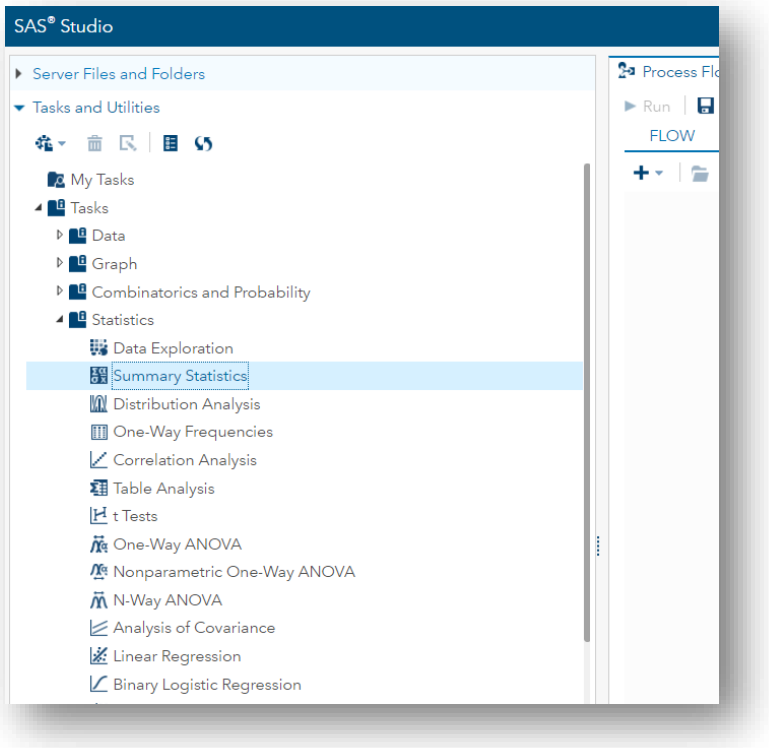
Guided Exercises

Before we start

Make sure you've downloaded the resources for today's challenge from here and launched SAS University Edition.

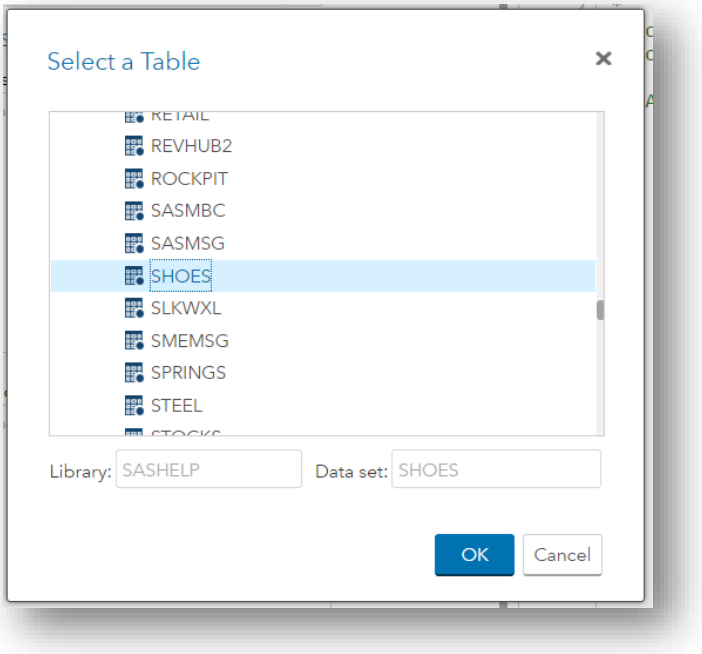
Generate Summary Statistics

1. Navigate to **Tasks and Utilities** in the left pane. Expand the **Statistics** menu.

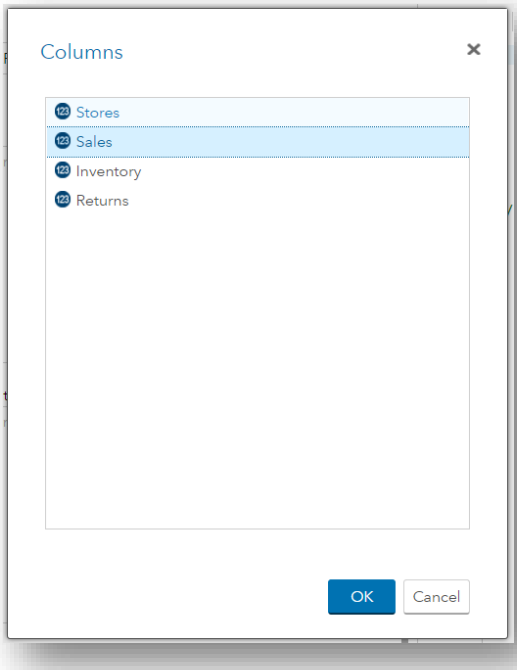


2. Double-click the **Summary Statistics** task to open it.

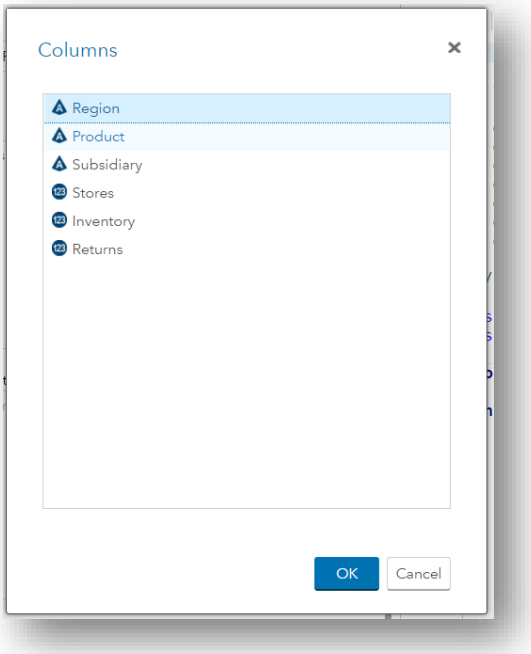
3. Under the **Data** heading, select the **SHOES** data set from the **SASHELP** library.



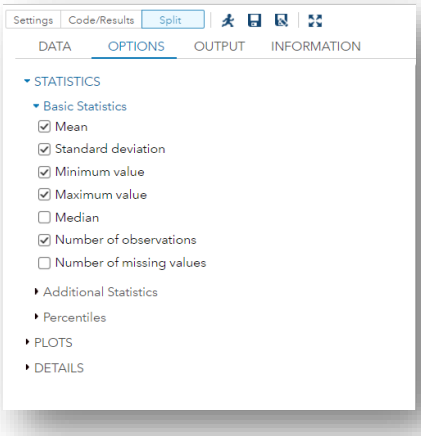
4. Under roles, click the '**+**' icon next to **Analysis variables**. Select **Sales** and click **OK**.



5. Click the '+' icon next to **Classification variables** and select **Region**.



6. Click the **Options** tab and confirm that the options from the below screenshot are checked.



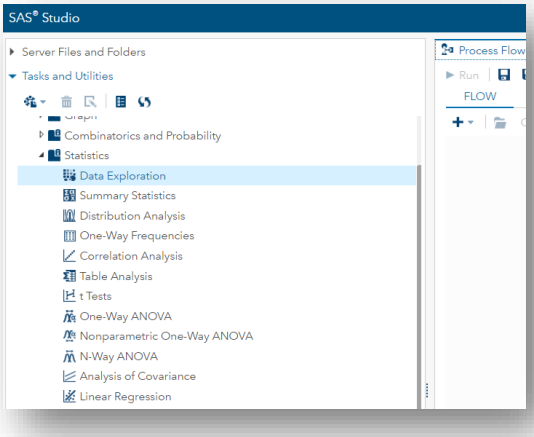
7. Click **Run** .

8. Click the **Results** tab to see the output table.

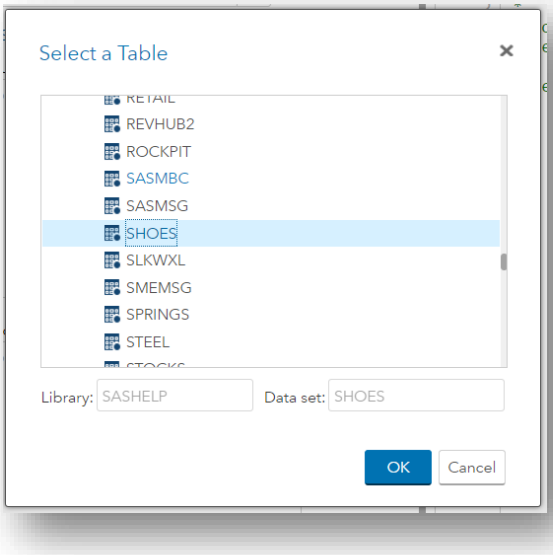
Analysis Variable : Sales Total Sales						
Region	N Obs	Mean	Std Dev	Minimum	Maximum	N
Africa	56	41831.93	65180.31	801.0000000	360209.00	56
Asia	14	32873.64	48880.01	937.0000000	149013.00	14
Canada	37	115019.24	205923.37	1190.00	757798.00	37
Central America/Caribbean	32	114304.78	137309.25	936.0000000	576112.00	32
Eastern Europe	31	77256.13	61066.23	712.0000000	304093.00	31
Middle East	24	234657.46	274458.02	449.0000000	1298717.00	24
Pacific	45	51039.87	73400.53	325.0000000	373908.00	45
South America	54	45088.57	47143.78	1716.00	245757.00	54
United States	40	137599.65	116442.77	554.0000000	456985.00	40
Western Europe	62	78596.77	87787.98	736.0000000	502636.00	62

Generate a Data Exploration

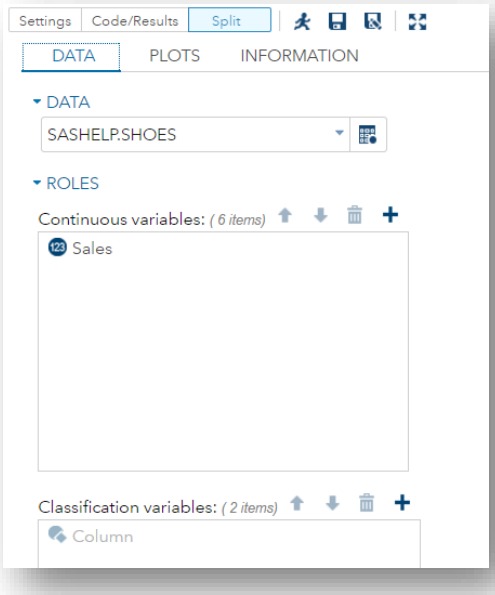
1. In the left pane expand **Tasks and Utilities**, and **Statistics**, double-click the **Data Exploration** task to open it.



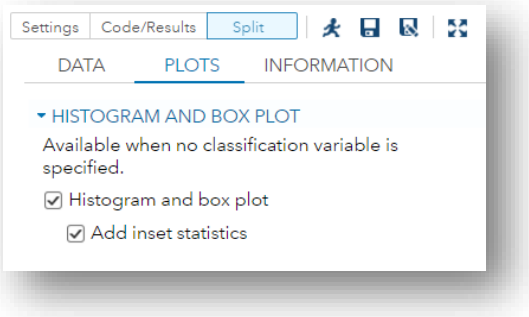
2. Under the **Data** heading, select the **SHOES** data set from the **SASHELP** library.



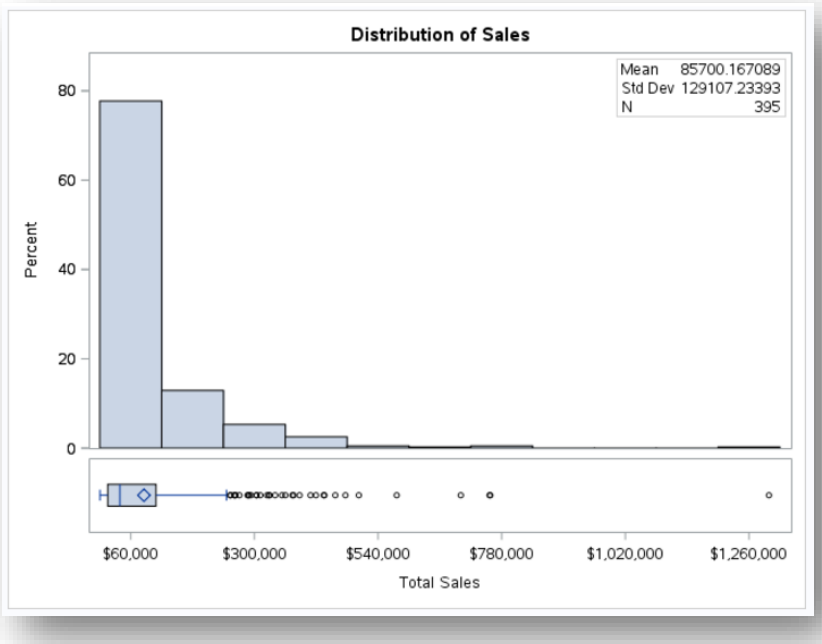
3. Add the **Sales** variable to **Continuous variables**, and do not add any variable to the **Classification variables**.



4. Click on the **Plots** tab and check the **Histogram and box plot** and **Add inset statistics** options.



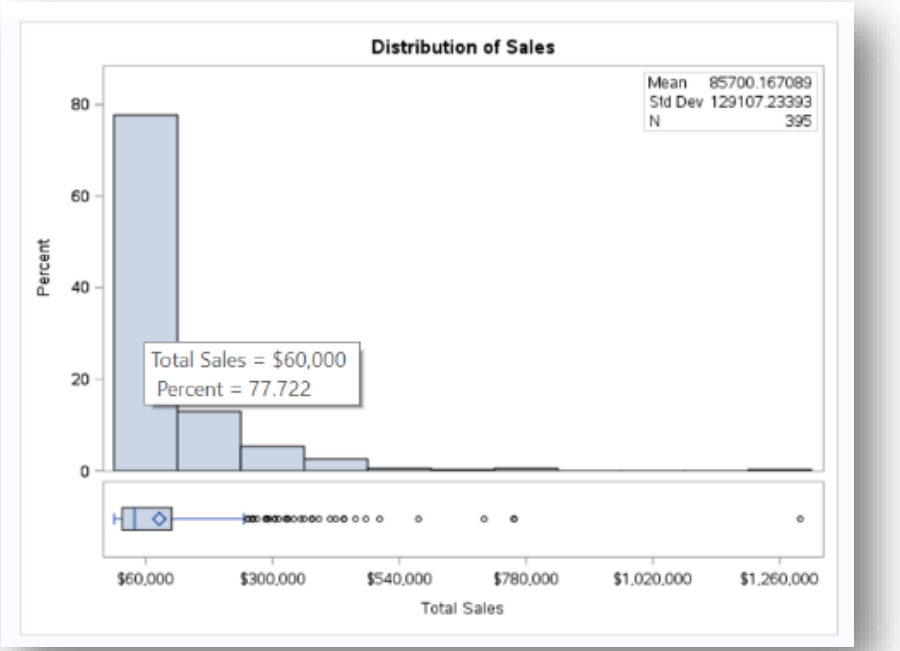
5. Click **Run** .
6. Click the **Results** tab to see the output histogram and boxplot.



As **Sales** is a continuous variable with many different values, you can see that it has been grouped into bins. You'll notice that the majority of the values appear on the left-hand side of the plot. This means this histogram is **positively skewed**.

To learn more about skewness and statistics in SAS, complete our free Statistics 1: Introduction to ANOVA, Regression, and Logistic Regression course.

7. Hover your mouse over the **$60,000 bin** on the plot to see more information. A tool tip also appears when you hover over the box plot.



Approximately 78% of **Sales** values are grouped into the **$60,000 bin**.

End of Guided Exercises

Challenge Exercises

Submit your answers here. Answer all 4 questions correctly to gain an entry into the prize draw.

This challenge requires information from the previous challenges. Be sure to complete the other challenges before completing this one.

Make sure you've downloaded the resources for today's challenge from here and placed it in **C:/SASUniversityEdition/myfolders**, the folder you created on Day 1.

It's a good idea to save your tasks in an accessible location once complete. If anything happens to your data, you can always rerun the tasks.

To complete these exercises, you will need to open and **Run** the SAS program named **Challenge_4_setup2018**. This program will ensure the data set is of the correct variable types. You need to have used the data set names specified in the previous challenge exercises for this code to run properly.

Generate **Summary Statistics** for the **NETFLIX_FINAL**, analysing the **user_rating_score** by **rating**. Use the resulting table to answer the following questions.

Question 1: *What is the average **user_rating_score** for **R**? Round this answer to 2 decimal places.*

Question 2: *What is the highest **user_rating_score** for **G**?*

Generate a **Data Exploration** for the **NETFLIX_FINAL** data set using the **user_rating_score** variable. *Hint: a **classification variable** is unnecessary but **plots** are.*

Question 3: *What is the median **user_rating_score**?*

Question 4: *What percentage of **user_rating_score** values fall into the **97.5 bin**?*

Don't forget to visit our Facebook page tomorrow for the last challenge!

https://www.facebook.com/SASAustNZ