SAS University Edition Challenge

Day 1 of 5 - Monday, 16th 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.

Kenny read an article that SAS University Edition would be the perfect free tool to analyse the data he has collected from Netflix before uploading it to the website's database.

Before you upload the data, you'll need to see what data you're working with. You'll import and clean the data by sorting, filtering and visualising some variables.

What are you waiting for? Let's get to work!

Summary of Skills Demonstrated

- Importing data
- Filtering data
- Sorting data
- Generating a simple bar chart

Submission Details

Submissions close **Monday, 23th 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 link below to download and install SAS University Edition.

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

Guided Exercises

Before we start

Make sure you've downloaded the resources for today's challenge from here.

It is not necessary to save tasks from the guided exercises. However, it is a good idea to get into the habit and will mean the tasks are available, should you want to run the code again.

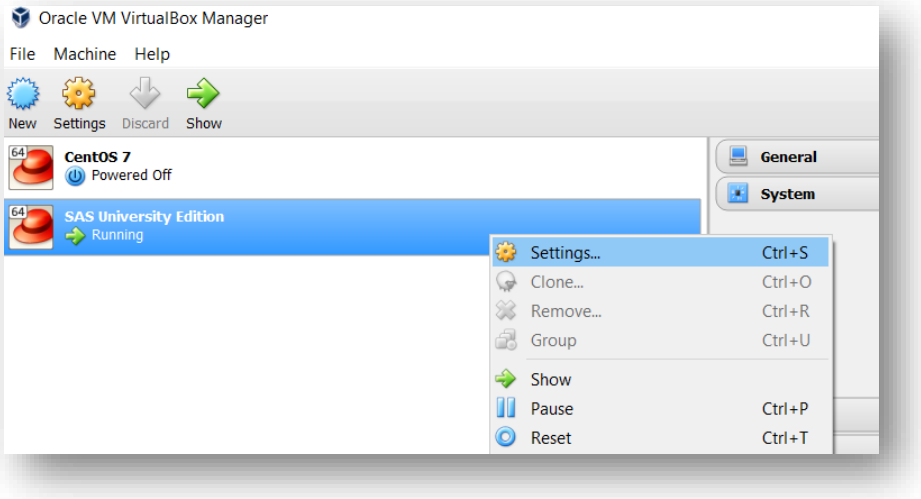
Setting up your library

1. Create a new folder in your main drive (i.e. C:\ or D:\) and name it **SASUniversityEdition**.
2. Inside the **SASUniversityEdition** folder add another folder named **myfolders**.
3. Place the data set you downloaded from the SASbox in **myfolders**.

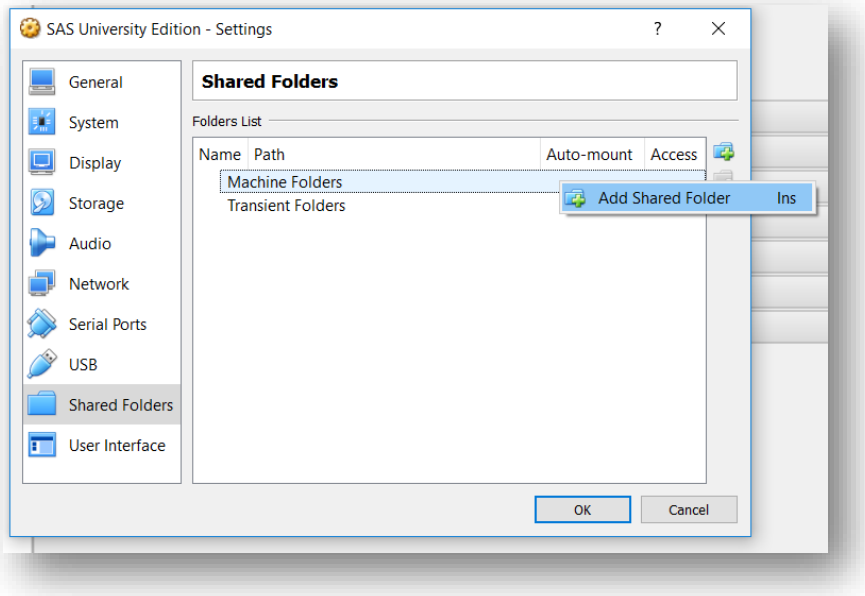
Set up a shared folder

A **shared folder** is basically a network folder that allows us to access files and data from our local machine in **SAS University Edition** image.

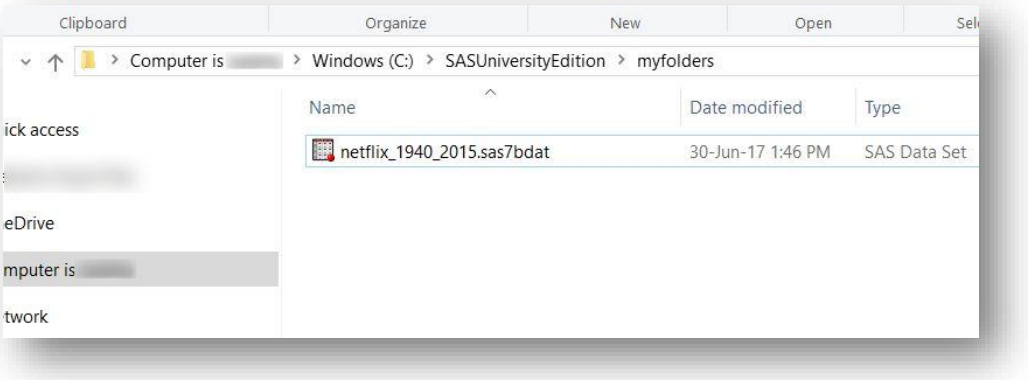
1. Open VirtualBox but do not start your SAS University Edition machine yet.
2. In VirtualBox, right-click the **SAS University Edition** machine and select **Settings**



3. Select **Shared Folders**, right-click **Machine Folders** and select **Add Shared Folder.**

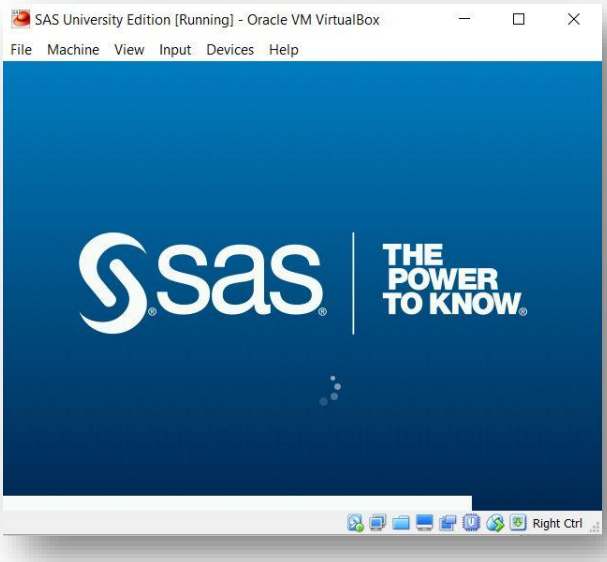


4. Point the **Folder Path** to the folder you created on your desktop earlier. Tick the following options: **Automount** and **Make Permanent**. Click **OK**.
5. Copy the **netflix_1940_2015** data set into the created location: **C:\SASUniversityEdition\myfolders**

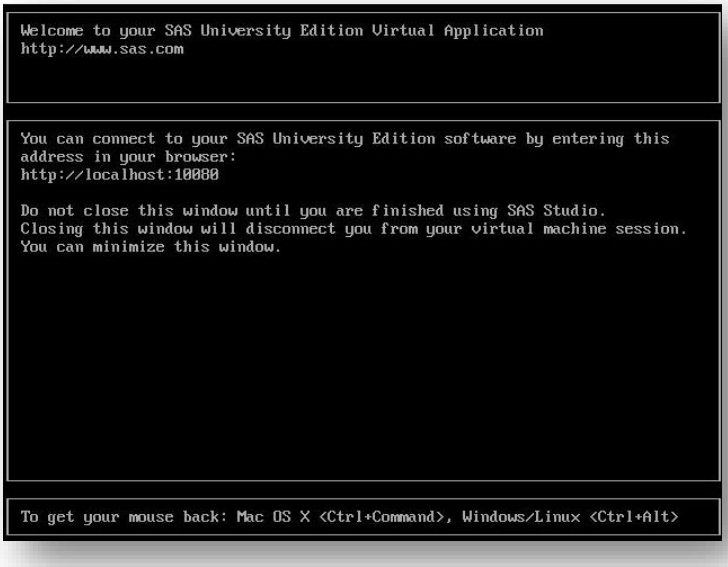


6. Start your SAS University Edition machine: right-click **SAS University Edition** and select **Start** or select **SAS University Edition** and click **Start** button.

   A SAS loading screen should appear as the machine starts up.

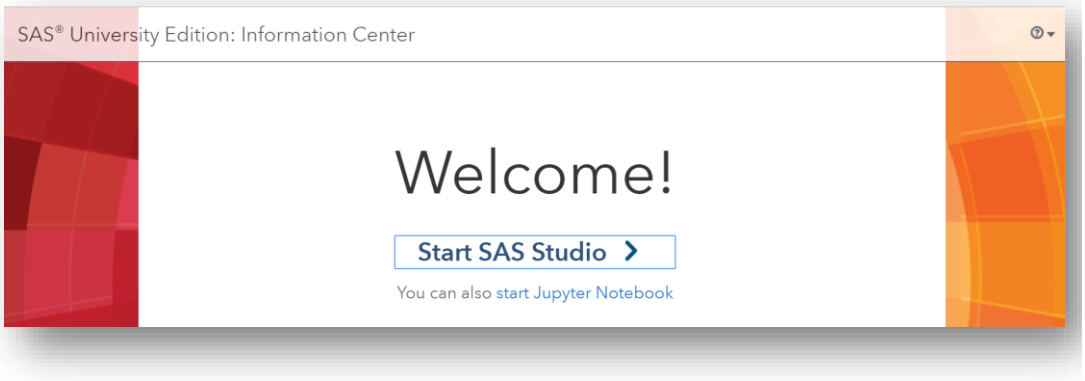


   When the machine is loaded, a welcome screen will appear with the address to connect to the image.



   In this case, **http://localhost:10080**. If you are using VMWare, it will appear in this format: **http://X.X.X.X:PortNumber**.

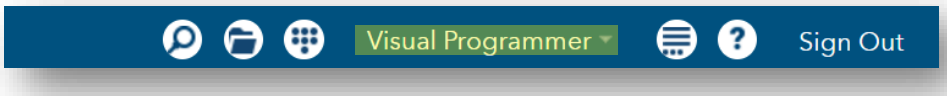
7. Enter the URL that appears in the **command prompt**, in your preferred browser. Click the **Start SAS Studio** button to launch **SAS Studio**.



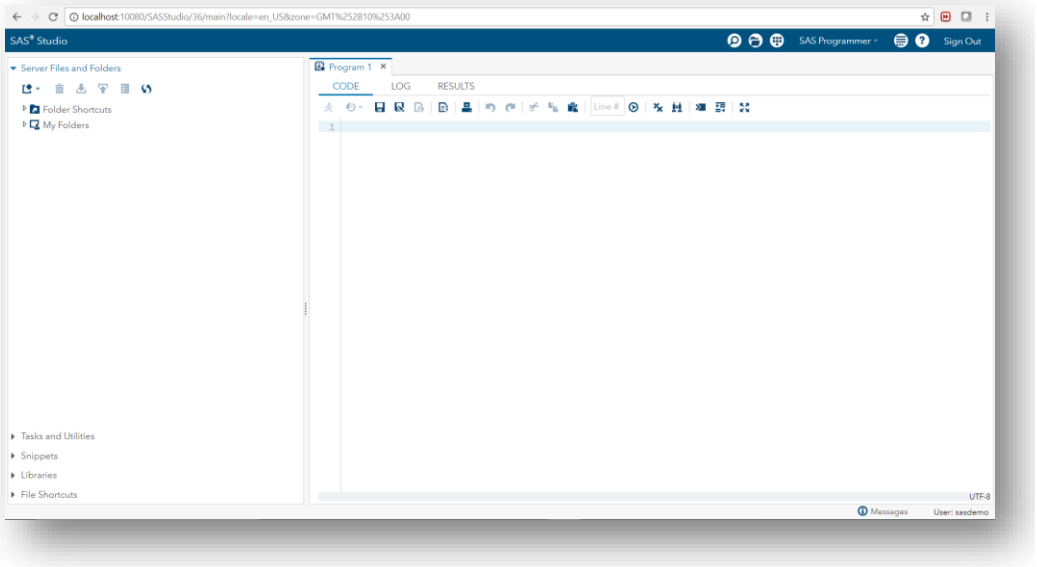
For more information, here are some links to the Oracle VirtualBox Quick Start video and creating a shared folder in VirtualBox.

Set up the library in SAS Studio

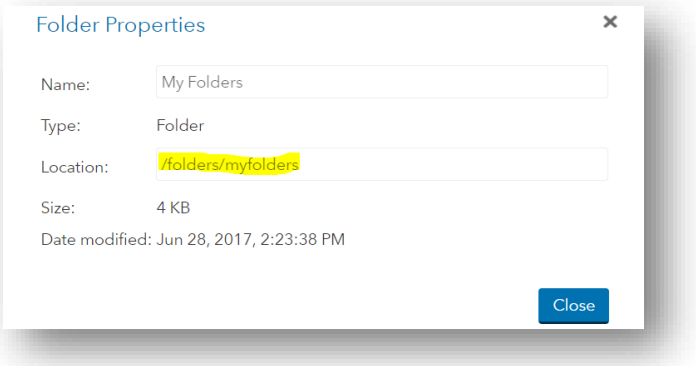
1. In SAS Studio, navigate to the top right corner and ensure **Visual Programmer** is selected.



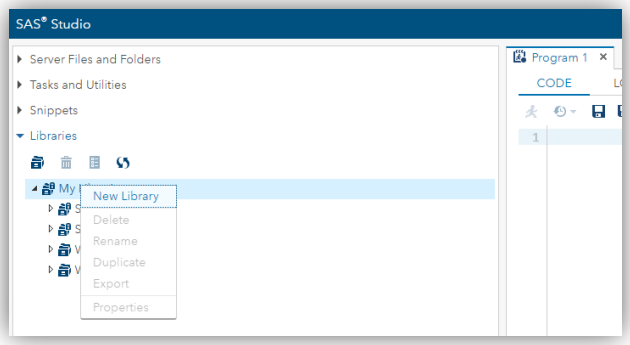
2. Navigate to the left pane and expand **Server Files and Folders.**



3. Right click on **My Folders** and select **Properties**.
4. Copy the value in **Location: /folders/myfolders**

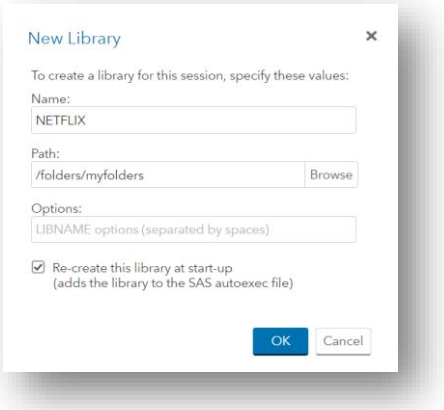


5. Navigate to the **Libraries** menu in the left pane.
6. Right-click **My Libraries** and create a new library named **NETFLIX**.



7. In the **Path** field, paste the value from step 4.

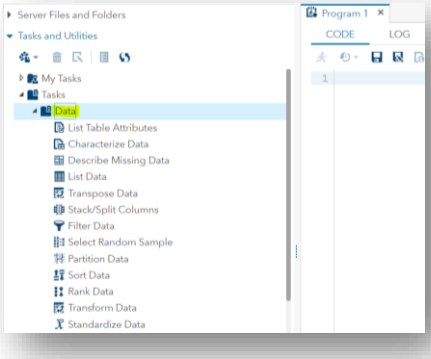
8. Check the **Re-create the library at start-up (adds the library to the SAS autoexec file)**. Click **OK**.



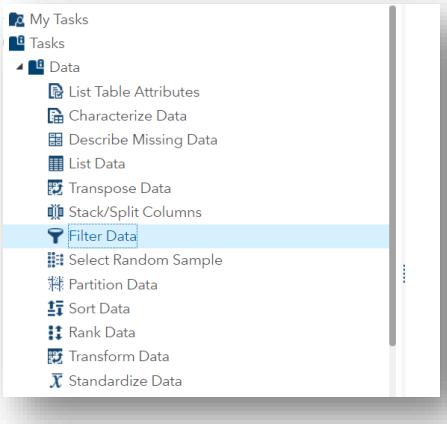
Before we start analysing the Netflix data, let's practice a few tasks on one of the inbuilt SAS data sets.

Filtering Data

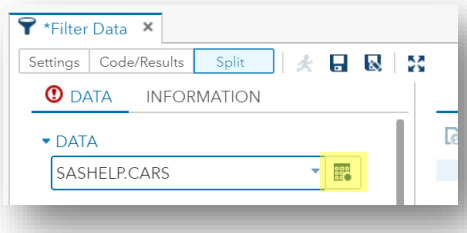
1. In the left pane, expand the **Tasks and Utilities** folder.
2. Expand the **Data** folder.

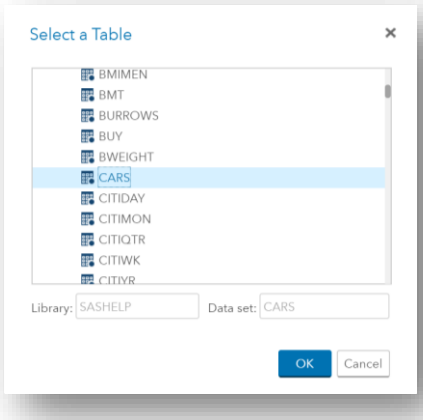


3. Double-click the **Filter Data** to open it.



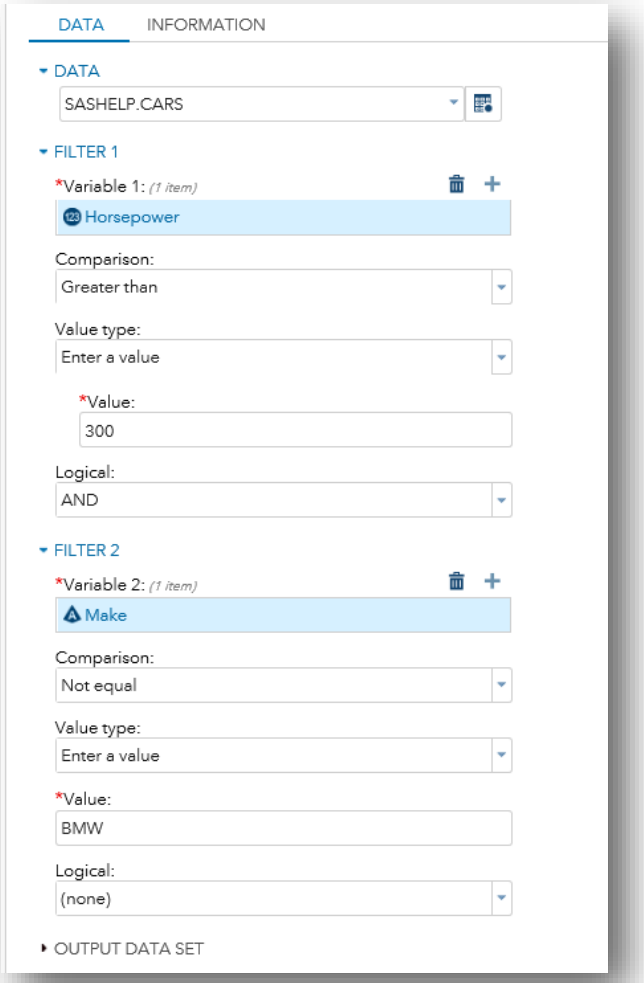
4. In the **Filter Data** tab, click on the **Select a Table** next to the **Data** drop-down menu and expand the **SASHELP** library. Scroll and find the **CARS** data set. Click **OK.**



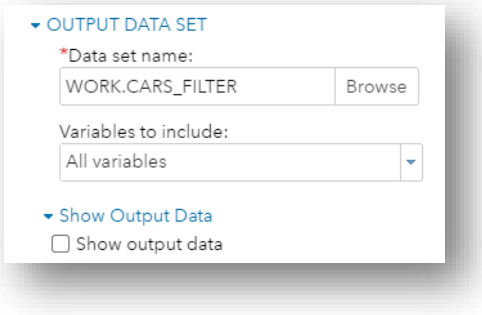


5. Click the '**+**' icon above **Variable 1** to select the variable to filter by. From the column list, select **Horsepower**. Click **OK**.
6. Select **Greater Than** as the **Comparison** operator.
7. Set **Value type** as **Enter a Value**.
8. Type **300** as the **Value**.
9. Select **AND** in the **Logical** drop-down list.
10. Repeat Step 5, instead selecting **Make** as the variable to filter by.
11. Select **NOT Equal** as the **Comparison** operator.
12. Set **Value type** as **Enter a Value**.

13. Enter '**BMW**' in the **Value** field.



14. In the **Data** tab, expand the **Output Data Set** menu. Save the output data set as **CARS_FILTER** in the **WORK** library.



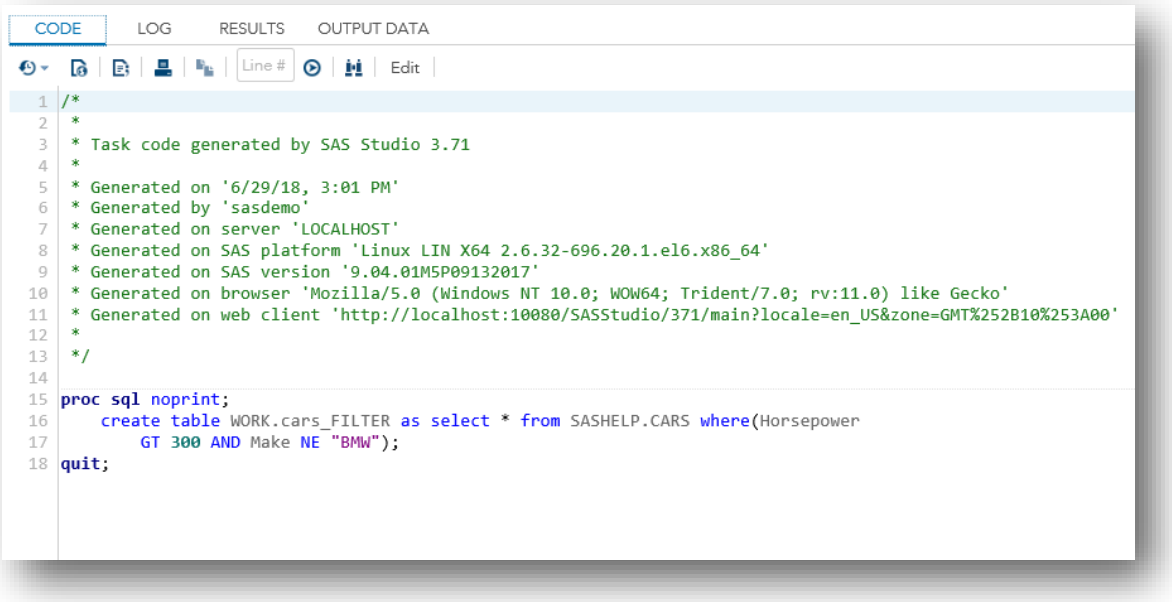
15. Click **Run**

The **Results** output displays cars that have **horsepower greater than** 300 **AND** without BMW in the Make column.

Total rows: 44 Total columns: 15 Rows 1-44

	Make	Model	Type	Origin	DriveTrain	MSRP	Invoice	EngineSize	Cylinders
1	Audi	A8 L Quattro 4dr	Sedan	Europe	All	$69,190	$64,740	4.2	8
2	Audi	S4 Quattro 4dr	Sedan	Europe	All	$48,040	$43,556	4.2	8
3	Audi	RS 6 4dr	Sports	Europe	Front	$84,600	$76,417	4.2	8
4	Audi	S4 Avant Quattro	Wagon	Europe	All	$49,090	$44,446	4.2	8
5	Cadillac	SRX V8	SUV	USA	Front	$46,995	$43,523	4.6	8
6	Cadillac	XLR convertible 2dr	Sports	USA	Rear	$76,200	$70,546	4.6	8
7	Cadillac	Escalade EXT	Truck	USA	All	$52,975	$48,541	6	8
8	Chevrolet	Corvette 2dr	Sports	USA	Rear	$44,535	$39,068	5.7	8
9	Chevrolet	Corvette convertible 2dr	Sports	USA	Rear	$51,535	$45,193	5.7	8
10	Dodge	Viper SRT-10 convertible 2dr	Sports	USA	Rear	$81,795	$74,451	8.3	10
11	Ford	Excursion 6.8 XLT	SUV	USA	All	$41,475	$36,494	6.8	10
12	GMC	Yukon XL 2500 SLT	SUV	USA	All	$46,265	$40,534	6	8
13	Hummer	H2	SUV	USA	All	$49,995	$45,815	6	8
14	Infiniti	M45 4dr	Sedan	Asia	Rear	$42,845	$38,792	4.5	8
15	Infiniti	Q45 Luxury 4dr	Sedan	Asia	Rear	$52,545	$47,575	4.5	8
16	Infiniti	FX45	Wagon	Asia	All	$36,395	$33,121	4.5	8
17	Jaguar	S-Type R 4dr	Sedan	Europe	Rear	$63,120	$57,499	4.2	8
18	Jaguar	XJR 4dr	Sedan	Europe	Rear	$74,995	$68,306	4.2	8
19	Jaguar	XKR coupe 2dr	Sports	Europe	Rear	$81,995	$74,676	4.2	8
20	Jaguar	XKR convertible 2dr	Sports	Europe	Rear	$86,995	$79,226	4.2	8
21	Lincoln	Aviator Ultimate	SUV	USA	Front	$42,915	$39,443	4.6	8
22	Mercedes-Benz	C32 AMG 4dr	Sedan	Europe	Rear	$52,120	$48,522	3.2	6
23	Mercedes-Benz	CL500 2dr	Sedan	Europe	Rear	$94,820	$88,324	5	8
24	Mercedes-Benz	CL600 2dr	Sedan	Europe	Rear	$128,420	$119,600	5.5	12

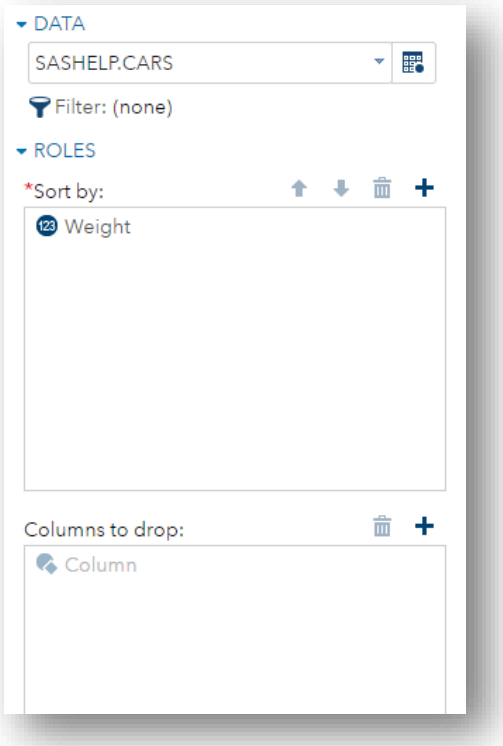
The SAS code to generate the result can be seen below for your reference.



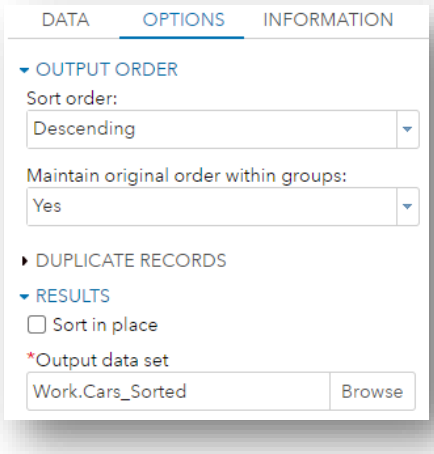
Sorting Data

1. In the left pane, expand the **Tasks and Utilities** folder.
2. Expand the **Data** folder.
3. Double-click the **Sort Data** task to open it.
4. Click on **Select a Table** next to the **Data** drop-down menu and click the **SASHELP** library.
5. Scroll down, select the **CARS** data set. Click **OK**.

6. Click the '+' icon for **Sort by** under the **Roles** heading. Select **Weight**.

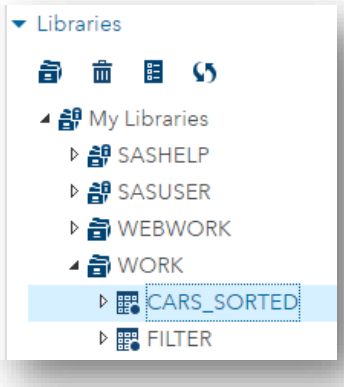


7. Click the **OPTIONS** tab and change **Sort order** to **Descending**.
8. Expand **Results** and save the **Output data set** as **CARS_SORTED** in the **WORK** library (**WORK.CARS_SORTED**).



9. Click **Run**

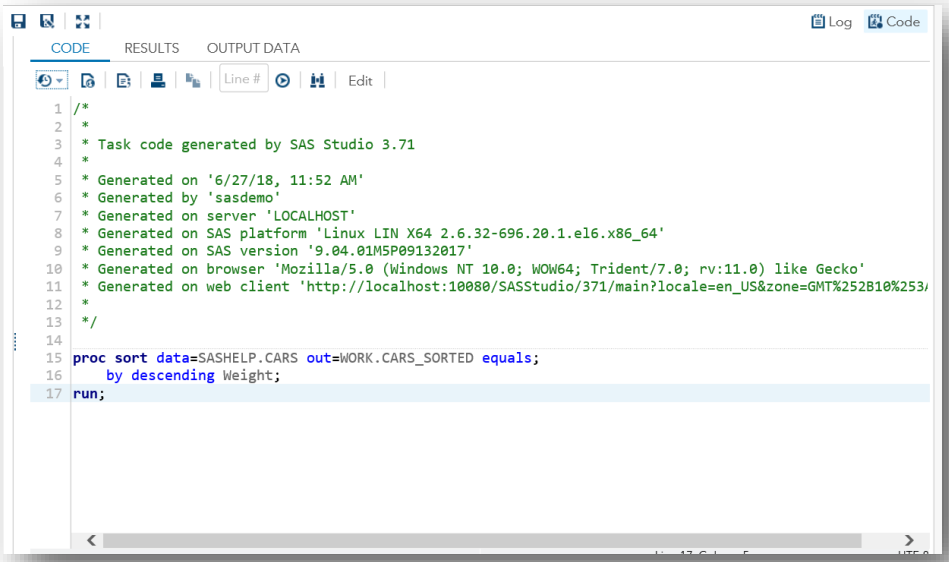
10. In the left pane, expand **Libraries**. Notice that the **CARS_SORTED** data set appears in the list of data sets.



The **Results** output displays all records in the data set, ordered by weight, from heaviest to lightest.

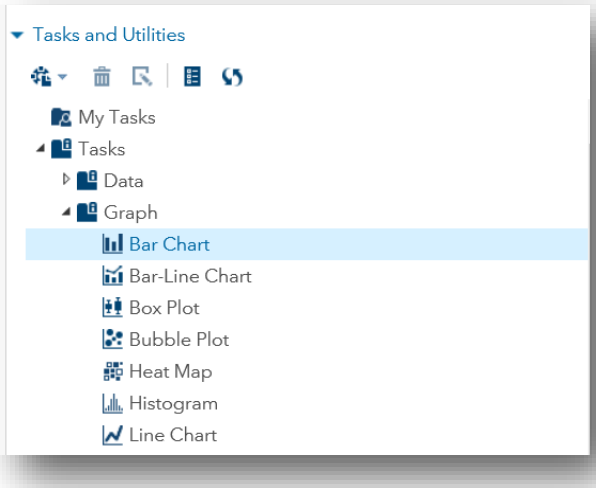
	Make	Model	Type	Origin	DriveTrain	MSRP	Invoice
1	Ford	Excursion 6.8 XLT	SUV	USA	All	$41,475	$36,494
2	Hummer	H2	SUV	USA	All	$49,995	$45,815
3	GMC	Yukon XL 2500 SLT	SUV	USA	All	$46,265	$40,534
4	Lincoln	Navigator Luxury	SUV	USA	All	$52,775	$46,360
5	Cadillac	Escalade EXT	Truck	USA	All	$52,975	$48,541
6	Chevrolet	Avalanche 1500	Truck	USA	All	$36,100	$31,689
7	Lexus	LX 470	SUV	Asia	All	$64,800	$56,455
8	Ford	F-150 Supercab Lariat	Truck	USA	All	$33,540	$29,405
9	GMC	Sierra HD 2500	Truck	USA	All	$29,322	$25,759
10	Mercedes-Benz	G500	SUV	Europe	All	$76,870	$71,540

The SAS Code for filtering can be seen below for your reference.



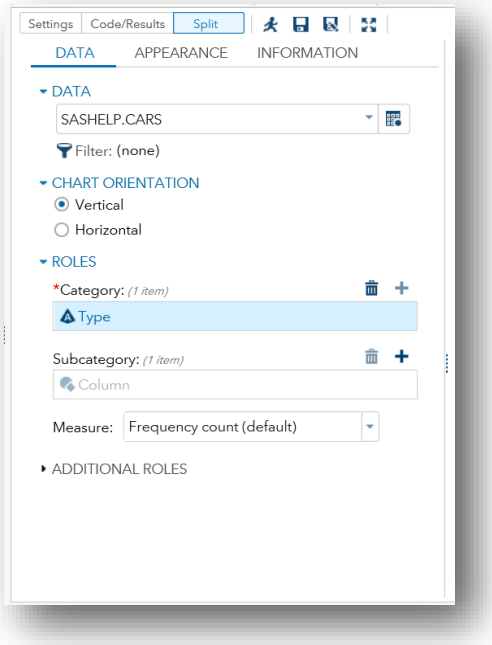
Generating a simple chart

1. In the left pane, expand the **Tasks and Utilities** folder.
2. Expand the **Graph** folder.
3. Double-Click on **Bar Chart**.

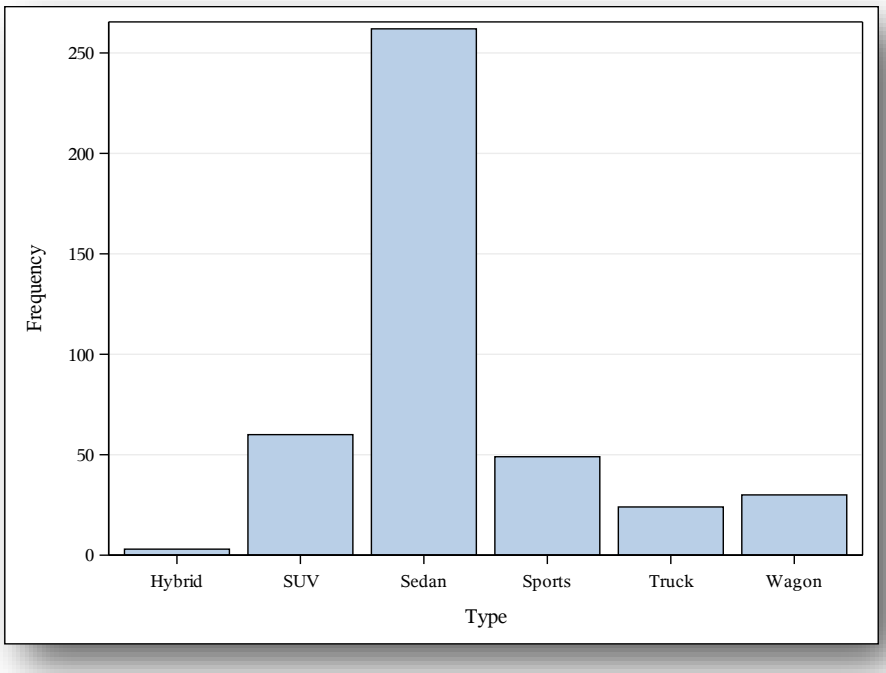


4. Click on **Select a Table** next the **Data** drop-down menu and click the **SASHELPlibrary**.
5. Scroll down, select the **CARS** data set. Click **OK**.

6. Select **Vertical** under **Chart Orientation**.
7. Under **Roles**, select **Type** for **Category**.



8. Select **Frequency Count** for **Measure**. Click **Run**



End of Guided Exercises

Challenge Exercises

Submit your answers here. Answer all 3 questions correctly to gain an entry into the prize draw.

Make sure you've downloaded the resources for today's challenge from here and placed it in **desktop/SASUniversityEdition/myfolders**, the folder you created earlier.

You will need to have completed the guided exercise above before attempting this challenge.

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.

Create a new data set from the **NETFLIX_1940_2015** data set. Ensure that the new data set only contains records which **do not** contain '**NA**' as the **user_rating_score** and only contains movies since the **year 2000**. Save the output data set in the **NETFLIX** library as '**NETFLIX_1940_2015_FILTERED**'.
Hint: *Make sure the* ***Comparison*** *operator is appropriate so that movies from the year 2000 are also included.*

> Question 1: *How many* ***titles*** *were released since the* ***year 2000*** *and* ***do not*** *have '****NA****' as the* ***user_rating_score*** *value?*

Sort the **NETFLIX_1940_2015_FILTERED** data set in **descending order** by **user_rating_score.** Also drop the variables **user_rating_size, ratingDescription, rating, release_year and ratingLevel**. (No output data set is required.)
Hint: *Remember to check the* ***sort order*** *option when answering question 2.*

> Question 2: *What are the* ***3 most popular titles*** *and the* ***3 least popular titles*** *by* ***user_rating_score****?*

Create a bar chart using the **NETFLIX_1940_2015_FILTERED** data set and select the appropriate category and measure to answer Question 3.

> Question 3: *Which* ***year*** *in the data set has the* ***most*** *movie titles?*

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

https://www.facebook.com/SASAustNZ