



Tutorial – JMP

A SAS SOFTWARE

Priyanka Kalmane | Technology Fundamentals for Business | 12/9/2016

Introduction

JMP, pronounced as JUMP is a data analysis tool. This is a product of SAS and can be used as a stand-alone package. It provides a wide range of functionalities for understanding, visualizing and communicating the contents of a chosen dataset.

JMP was developed with a purpose of being an interactive visualization and discovery tool in the late 1980's.

In general, JMP provides the following functionalities:

- Data exploration and display
- Experiment design
- Quality control
- Qualitative analysis
- Statistical modeling, Report building.

WHY USE JMP?

JMP is a visual discovery tool which can be installed directly on a desktop or a server. It has multiple advantages which may be listed as below:

- Ease of use
- Used in association with SAS. Allows for easy visualization even without enough knowledge of statistical complexities.
- Can be used without training or prior programming knowledge
- Point and click interface.

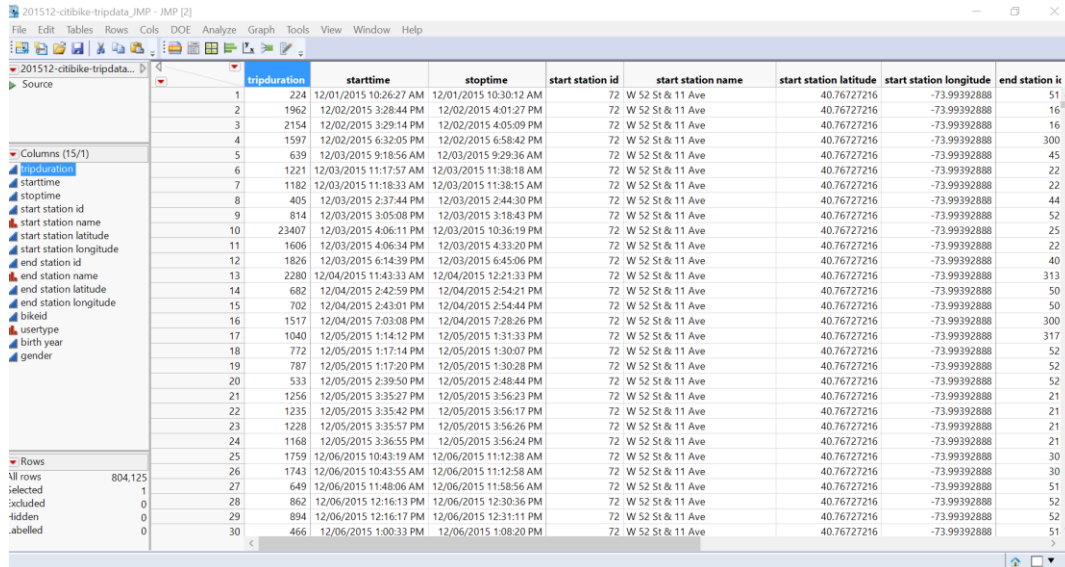
Experiment with JMP

In this tutorial, I have covered the basic functionalities of JMP on an example dataset : Citi Bike System Data(Only for the month of Dec,2015). It covers the following topics:

- Importing Data
- Data Understanding and Preparation
- Using simple statistical functions for simple data analysis
- Plotting Graphs

IMPORTING DATA

JMP being a highly intuitive interface, allows the user to import Excel files or CSV files. It also allows us to work with SAS data sets. Once imported, the data looks like this:

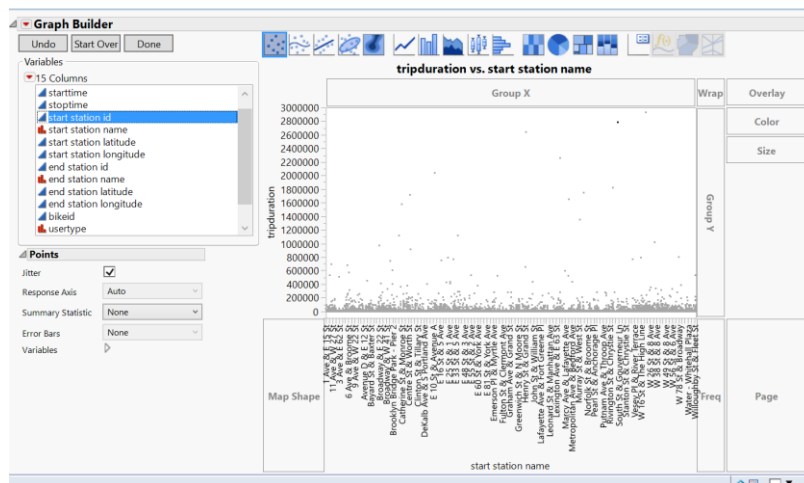


	tripduration	starttime	stoptime	start station id	start station name	start station latitude	start station longitude	end station id
1	224	12/01/2015 10:26:27 AM	12/01/2015 10:30:12 AM	72	W 52 St & 11 Ave	40.76727216	-73.99392888	51
2	1962	12/02/2015 3:28:44 PM	12/02/2015 4:01:27 PM	72	W 52 St & 11 Ave	40.76727216	-73.99392888	16
3	2154	12/02/2015 3:29:14 PM	12/02/2015 4:05:09 PM	72	W 52 St & 11 Ave	40.76727216	-73.99392888	16
4	1597	12/02/2015 6:32:05 PM	12/02/2015 6:58:42 PM	72	W 52 St & 11 Ave	40.76727216	-73.99392888	300
5	639	12/03/2015 9:18:56 AM	12/03/2015 9:29:36 AM	72	W 52 St & 11 Ave	40.76727216	-73.99392888	45
6	1221	12/03/2015 11:17:57 AM	12/03/2015 11:38:18 AM	72	W 52 St & 11 Ave	40.76727216	-73.99392888	22
7	1182	12/03/2015 11:18:33 AM	12/03/2015 11:38:15 AM	72	W 52 St & 11 Ave	40.76727216	-73.99392888	22
8	405	12/03/2015 2:37:44 PM	12/03/2015 2:44:30 PM	72	W 52 St & 11 Ave	40.76727216	-73.99392888	44
9	814	12/03/2015 3:05:08 PM	12/03/2015 3:18:43 PM	72	W 52 St & 11 Ave	40.76727216	-73.99392888	52
10	23407	12/03/2015 4:06:11 PM	12/03/2015 10:36:19 PM	72	W 52 St & 11 Ave	40.76727216	-73.99392888	25
11	1606	12/03/2015 4:06:34 PM	12/03/2015 4:33:29 PM	72	W 52 St & 11 Ave	40.76727216	-73.99392888	22
12	1826	12/03/2015 6:14:39 PM	12/03/2015 6:45:06 PM	72	W 52 St & 11 Ave	40.76727216	-73.99392888	40
13	2280	12/04/2015 11:43:33 AM	12/04/2015 12:21:33 PM	72	W 52 St & 11 Ave	40.76727216	-73.99392888	313
14	682	12/04/2015 2:42:59 PM	12/04/2015 2:54:41 PM	72	W 52 St & 11 Ave	40.76727216	-73.99392888	50
15	702	12/04/2015 2:43:01 PM	12/04/2015 2:54:44 PM	72	W 52 St & 11 Ave	40.76727216	-73.99392888	50
16	1517	12/04/2015 7:03:08 PM	12/04/2015 7:28:26 PM	72	W 52 St & 11 Ave	40.76727216	-73.99392888	300
17	1040	12/05/2015 1:14:12 PM	12/05/2015 1:31:33 PM	72	W 52 St & 11 Ave	40.76727216	-73.99392888	317
18	772	12/05/2015 1:17:14 PM	12/05/2015 1:30:07 PM	72	W 52 St & 11 Ave	40.76727216	-73.99392888	52
19	787	12/05/2015 1:17:20 PM	12/05/2015 1:30:28 PM	72	W 52 St & 11 Ave	40.76727216	-73.99392888	52
20	533	12/05/2015 2:39:50 PM	12/05/2015 2:48:44 PM	72	W 52 St & 11 Ave	40.76727216	-73.99392888	52
21	1256	12/05/2015 3:35:27 PM	12/05/2015 3:56:23 PM	72	W 52 St & 11 Ave	40.76727216	-73.99392888	21
22	1235	12/05/2015 3:35:42 PM	12/05/2015 3:56:17 PM	72	W 52 St & 11 Ave	40.76727216	-73.99392888	21
23	1228	12/05/2015 3:35:57 PM	12/05/2015 3:56:26 PM	72	W 52 St & 11 Ave	40.76727216	-73.99392888	21
24	1168	12/05/2015 3:36:55 PM	12/05/2015 3:56:24 PM	72	W 52 St & 11 Ave	40.76727216	-73.99392888	21
25	1759	12/06/2015 10:43:19 AM	12/06/2015 11:12:38 AM	72	W 52 St & 11 Ave	40.76727216	-73.99392888	30
26	1743	12/06/2015 10:42:55 AM	12/06/2015 11:12:58 AM	72	W 52 St & 11 Ave	40.76727216	-73.99392888	30
27	649	12/06/2015 11:40:06 AM	12/06/2015 11:58:56 AM	72	W 52 St & 11 Ave	40.76727216	-73.99392888	51
28	862	12/06/2015 12:16:13 PM	12/06/2015 12:30:36 PM	72	W 52 St & 11 Ave	40.76727216	-73.99392888	52
29	894	12/06/2015 12:16:17 PM	12/06/2015 12:31:11 PM	72	W 52 St & 11 Ave	40.76727216	-73.99392888	52
30	466	12/06/2015 1:00:33 PM	12/06/2015 1:08:20 PM	72	W 52 St & 11 Ave	40.76727216	-73.99392888	51

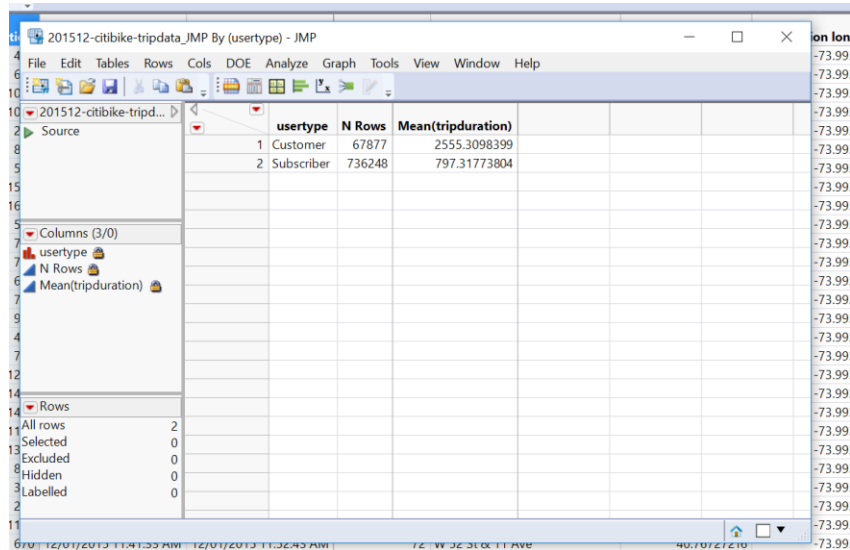
The data above consists of 804,125 observations with 15 columns. Our main focus here would be to understand how the trip duration varies across different stations, user types, birth years and gender.

DATA UNDERSTANDING AND PREPARATION

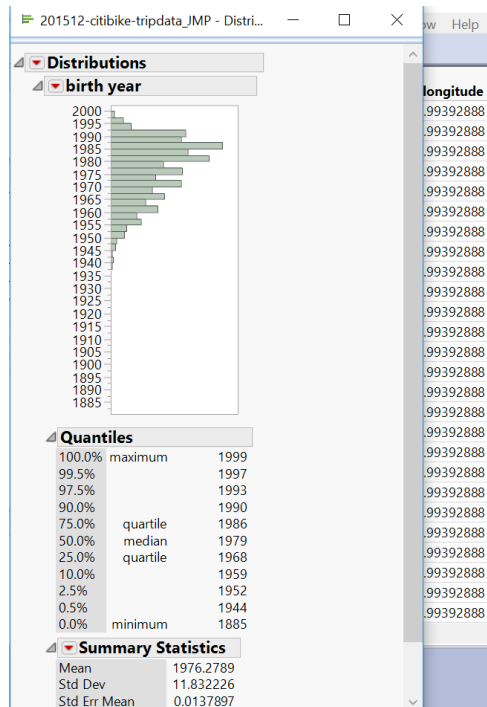
Before preparing the data for effective analysis that assist in making informed choices and inferences, it is always a good practice to visualize the data. Thus, using the Graph builder functionality in JMP, the following is obtained:



We see that there are quite a few outliers and we may need to deal with them as per the demands of analysis. Thus, we try to understand what exactly our data comprises of. JMP provides the functionality of analyzing the data by a simple point-and-click. We can learn about the different categories and the number of records that belong to each category. Here's an example of that:



To better understand the distribution, we can use the Analyze->Distribution functionality of JMP. An example of that is as follows:



We see that from above that though Birth year does not follow a perfect normal distribution, it does follow an almost normal distribution.

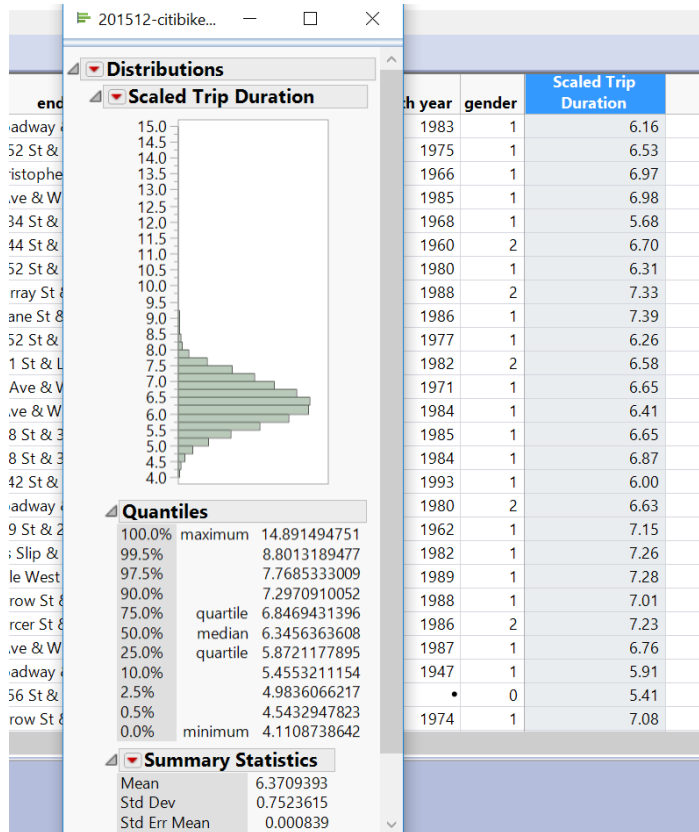
If we want to obtain a view of only partial data, JMP provides us with a facility to do so. It looks like the following:

The screenshot shows the JMP software interface with a data table titled 'Untitled 13'. The table has columns for 'start station name', 'end station name', 'bikeid', and 'usertype'. The data is filtered to show rows 1 through 19. The 'start station name' column contains values like '2 St & 11 Ave', '3 St & 11 Ave', etc. The 'end station name' column contains values like 'W 56 St & 10 Ave', 'W 18 St & 6 Ave', etc. The 'bikeid' column contains values like 16419, 17242, etc. The 'usertype' column contains values like 'Customer', 'Subscriber', etc.

For a few variables such as Trip Duration where it is not possible to obtain a normal distribution, we can scale the data by adding additional columns. By creating formulas for the new column (such as transcendental form – logarithmic value of any column), we can plot to check if they follow normal distribution. Also, different formulas such as date-time, Trigonometric functions etc. may be applied to the data. Example:

The screenshot shows the JMP software interface with a data table titled '201512-ctibike-tripdata...'. The table has columns for 'time', 'stop time', 'start station name', 'end station name', 'bikeid', 'usertype', 'birth year', 'gender', and 'Scaled Trip Duration'. A dialog box is open for creating a new column named 'Scaled Trip Duration'. The dialog box shows the 'Column Name' as 'Scaled Trip Duration', the 'Data Type' as 'Numeric', the 'Modeling Type' as 'Continuous', and the 'Formula' as 'Log (tripduration)'. The 'Format' is set to 'Fixed Dec' with a width of 12 and a decimal of 2. The 'Use thousands separator' checkbox is unchecked. The 'Column Properties' dropdown is set to 'Formula'.

By choosing the Distribution function in JMP, we get:



APPLYING SIMPLE STATISTICAL FUNCTIONS:

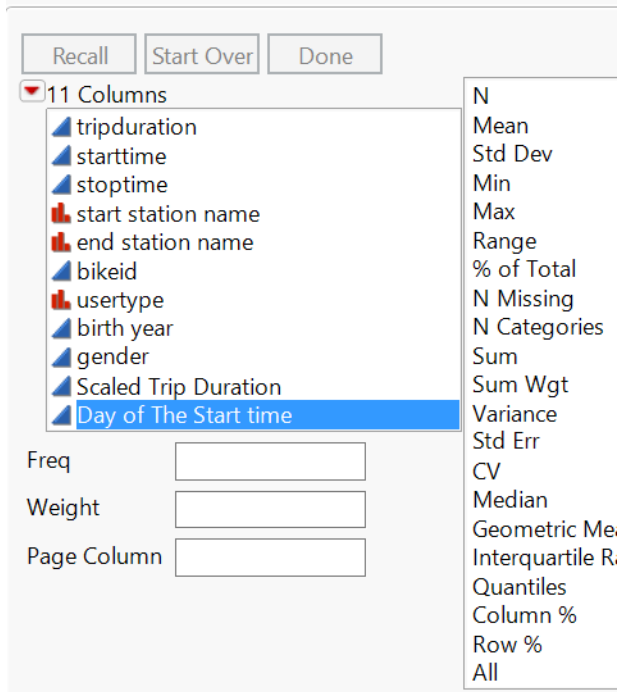
JMP provides the flexibility for non-programmers to further understand the data by using some basic statistical functions. Here's an example of how the Grand Mean is computed using the "Tabulate" functionality:

201512-citibike-tripdata_JMP - Tabulate - JMP

Tabulate

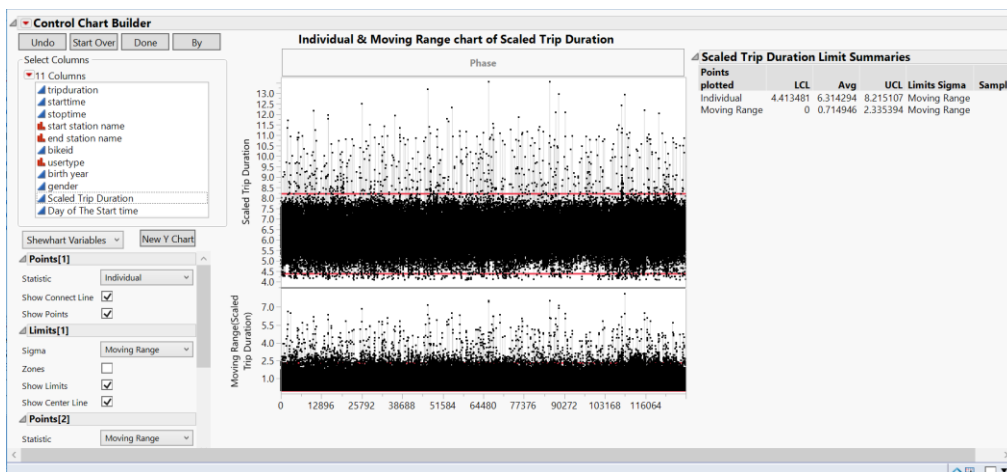
start station name	usertype		
	Customer	Subscriber	All
1 Ave & E 15 St	2300.33	613.39	660.94
1 Ave & E 18 St	2145.90	677.21	727.13
1 Ave & E 30 St	1169.82	739.34	755.15
1 Ave & E 44 St	1678.55	835.70	905.80
1 Ave & E 62 St	1871.45	831.36	897.65
1 Ave & E 68 St	4368.25	783.36	972.09
1 Ave & E 78 St	2871.27	851.19	953.02
10 Ave & W 28 St	2269.57	874.33	944.09
11 Ave & W 27 St	1921.60	885.51	933.34
11 Ave & W 41 St	1708.42	783.00	819.17
11 Ave & W 59 St	2068.28	926.26	999.31
12 Ave & W 40 St	2067.32	1156.97	1366.17
2 Ave & E 31 St	1232.67	691.94	712.92
2 Ave & E 58 St		5069.00	5069.00
21 St & 41 Ave	1652.23	1144.53	1252.73
21 St & 43 Ave	2424.20	1036.48	1226.58
3 Ave & E 62 St	6285.25	740.57	1197.35
3 Ave & Schermerhorn St	2241.47	778.90	941.41
31 St & Thomson Ave	1288.33	1281.20	1281.40
44 Dr & Jackson Ave	1948.87	747.19	833.65
45 Rd & 11 St	1027.46	616.61	642.09
46 Ave & 5 St	1482.52	462.63	522.46
47 Ave & 31 St	4858.00	1187.48	1372.32
48 Ave & 5 St	1831.16	719.84	909.21
5 Ave & E 29 St	1749.12	627.21	704.87
5 Ave & E 63 St	2779.29	808.72	1756.93
5 Ave & E 73 St	2548.90	1099.13	1845.47
5 Ave & E 78 St	3189.06	975.11	2258.77
6 Ave & Broome St	3292.63	819.22	1025.33
6 Ave & Canal St	2143.08	1215.35	1354.26
6 Ave & W 33 St	1323.89	754.76	802.83
8 Ave & W 31 St	1525.29	786.85	829.08

In the above figure, we see how the Grand mean of the Trip duration is computed for different user types for the third day of the week. Similar to this, JMP provides other features within the Tabulate functionality as follows:

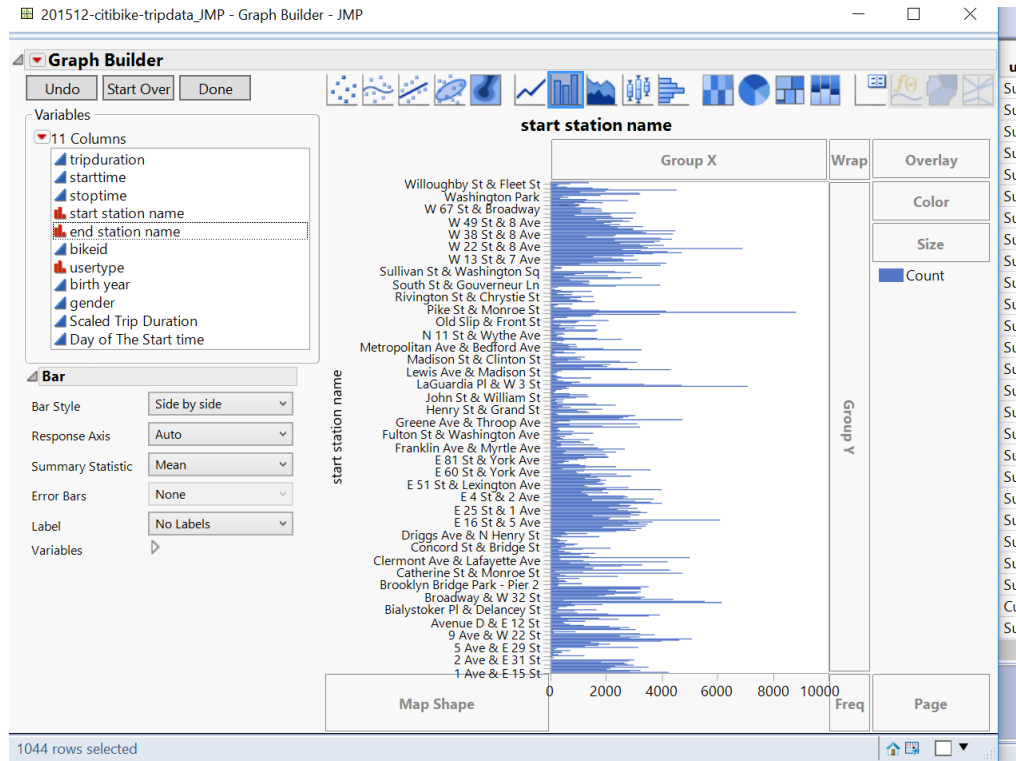


PLOTTING GRAPHS:

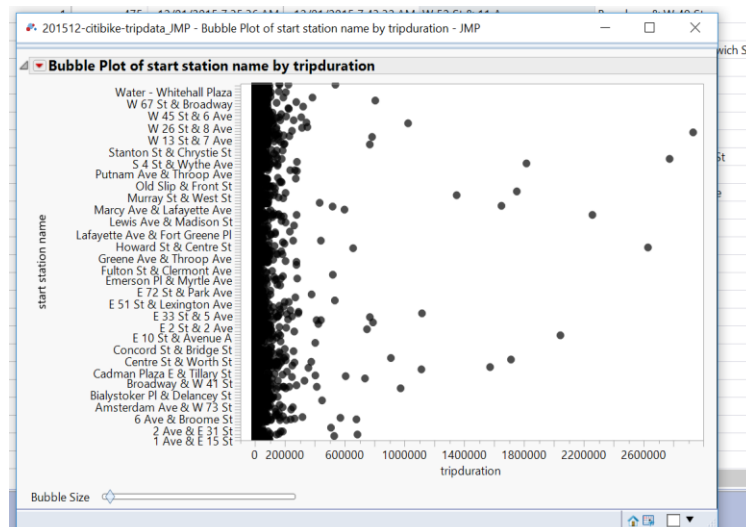
Data Visualization is one of the most important steps in data analysis. It is something that it carried out in every step of the CRISP model. JMP provides us with many different interesting ways to plot and visualize the aggregated data and data before analysis. This helps us in making decisions on the go. Also, with simple drag and drop functionalities, it makes the user's life a lot easier. Here are a few examples of how that is done:



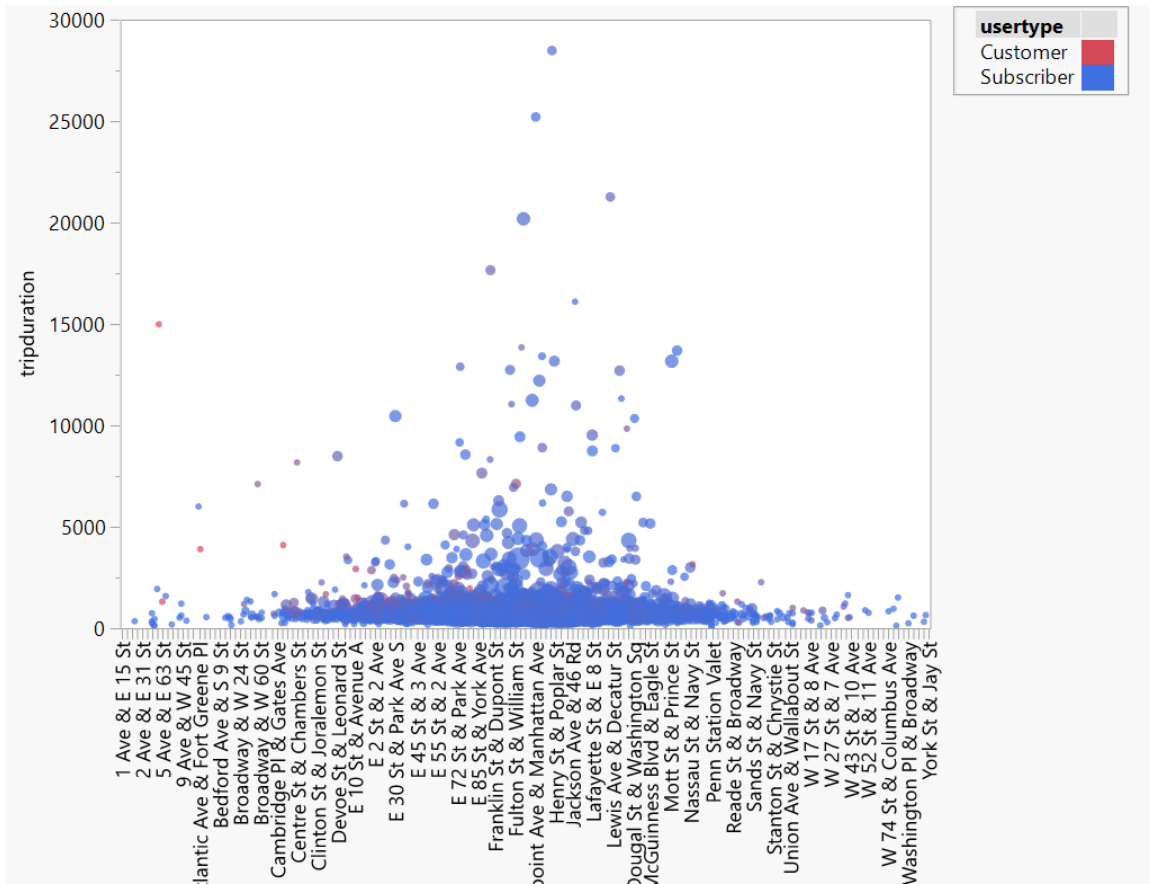
The above mentioned chart is called the Control Chart which summarizes the Scaled Trip Duration and enables the users in easily understanding the various outliers within the data. This kind of a Control Chart is particularly useful when we need to plot Price of a particular product over time.



The above graph gives us a picture of which of the Start Stations had the highest frequency. This helps us know which place was the most in demand. With this, we may be further able to optimize our resources to regulate costs and forecast future demands. The same data is expressed in Bubble Chart form.



Using Bubble Chart, we can plot the Start station name against the trip duration for different user types grouped by different days of the week. The Graph builder functionality in JMP is used here. It is as shown below:



Conclusion

JMP is capable of providing a lot more powerful features. Although it may not be able to handle as many records as SAS, it can still be used for initial analysis and basic understanding of the data set. After learning the general intuition, it is possible to delve deeper into the data and carry out more rigorous analysis either in JMP or by setting up a connection with SAS. Thus, both JMP and SAS together would make for a very useful data analysis tool.

With this tutorial, I have tried to outline the basic analysis capabilities of JMP. While JMP is capable of performing a lot more complex tasks than the ones described here, this tutorial is the place to learn about what kind of a tool JMP is, especially for beginners.