Washington DC Transit Data Project

By:

Maddie Warndorf, Cameron Husseini, Marzouq Almohammed, Damien Budnick

Overview

General idea: Traveling around Washington DC

Why are we interested: Because when we travel to DC it seems like we never go at the right time. So when is the right time? We have the answer.

We chose three different datasets because we wanted to be challenged as well as have the opportunity to analyze different attributes.

Who would want this data: Travelers, Marketing agencies, D.O.T (Department of Transportation), WMATA, Bikeshare

Why that time frame: quick answer: a lot of data in 3 years, enough time too

What we gathered from the data: See Conclusion

Software we used: R, Tableau and Excel

Describe the Data

Where was the data from?

- Online sources Bikeshare data: from the actual Bikeshare website archives
- Gas Prices: https://www.eia.gov/opendata/qb.php?sdid=PET.EMM_EPMR_PTE_SWA_DPG.M
- Weather:https://www.wunderground.com/history/airport/KDCA/2014/1/1/MonthlyHistory.html?req_city=&req_state=&req_statename=&reqdb.zip=&reqdb.magic=&reqdb.wmo=
- Bikeshare: https://s3.amazonaws.com/capitalbikeshare-data/index.html

What is the data?
How was it originally cleaned from the raw; which variables were used?

Cleaned the data from the selected time from Q1 2014 to Q1 2017. Also the location of which the data pertained, in this case being the District of Columbia (Aka D.C.).

Capital BikeShare Data

The BikeShare data was downloaded from Capital BikeShare System Data website[1]. From this site, all of the .zip files were downloaded that corresponded to the time range that was analyzed (2014 Q1 to 2017 Q1). The raw data .CSV files were then converted into an Excel file to allow the data to be imported into R. The raw data was broken down by mm/dd/yyyy h:m:s[2].

The raw data was then cleaned in R and the variables that were selected to be added to the data sample were Duration and StartDate. From the Duration variable, the hours, minutes, and seconds were pulled and converted so that the Duration is in minutes instead of h:m:s, (DurationInMinute). The variable StartDate was used to separate out the Month and Year. The new FullDate variable was then created using the format MM/YYYY. Since the data was broken up by year and quarter, the variable Quarter was added to the new data sample.

Gas Price Data

The gas price data was found on the U.S. Energy information Administration (EIA) website[3]. The data was downloaded from the website in by first looking at the time range of the project (Q1 2014 to Q1 2017). Then the Excel sheet file was cleaned to match the time frame that we using for this project. The excel file for the gas data included a lot of years and information that we didn't need in the project, so I eliminated

[1] https://s3.amazonaws.com/capitalbikeshare-data/index.html
[2] mm/dd/yyyy h:m:s means month/day/year hour:minute:second ex: 10/11/2015 20:28:22
[3] https://www.eia.gov/opendata/qb.php?sdid=PET.EMM_EPMR_PTE_SWA_DPG.M

the unwanted data using Microsoft Excel. The data for the gas included the value of the gas, which is dollars per gallon, and it had the frequency set to month. The raw data started from 2003 to 2017 which is really more than what we needed so the clean it an make it start from Q1 2014 to Q1 2017. The variable that we used in our analysis was the gas price and the time of the year was equally matched.

Weather Data

The weather data was found on the website Weather Underground for Washington D.C., Ronald Reagan National Airport[4]. In order for the data to be properly formatted, the data had to be manually entered into an Excel file. This resulted in nine variables; Date (mm/dd/yyyy), Temperature (°F), Dew point, Humidity, Sea level pressure, Visibility, Wind, Precipitation, Events. All of the data was filled in but the events, which would include: Rain, Fog, Snow. So if the data was blank it had no events on each day so we filled in with No Events. The data was cleaned such that only the temperature (°F) would be used for the data sample due to it being the variable of interest to compare to the BikeShare data.

Data Sample

The data sample that was created from the selected variables mentioned above, required some calculations to create the variables in the dataset. Each of the datasets were broken down by different units of time as described above.. To prevent missing data in the data sample, the data from the datasets that were broken down by days were either averaged together by month to find the data shown in the sample. The new combined dataset contains the attributes FullDate (mm/yyyy), Year, Quarter, Month, MonthName, TotalRides, DurationMin (in minutes), DurationMax (in minutes), DurationTotalMinutes (in minutes), GasPrices ($), and TempAvg (°F).

The FullDate, Year, and Month were pulled from the Capital BikeShare dataset's variable StartDate. The MonthName was determined by matching the numerical representation of the month to the name of the month. The attribute Quarter was pulled from the name of the Capital BikeShare dataset file. The attribute TotalRides was calculated by adding all of the records that occured in the specific month and year. DurationMin (in minutes) and DurationMax (in minutes) are the minimum and maximum durations of the rides for that month. DurationTotalMinute (in minutes) is the total duration of all of the rides taken in that month.

The attribute GasPrices ($) was taken directly from the dataset since the Gas Price dataset was broken down by mm/yyyy. This attribute was only altered to fit the

[4]https://www.wunderground.com/history/airport/KDCA/2014/1/1/MonthlyHistory.html?req_city=&req_state=&req_statename=&reqdb.zip=&reqdb.magic=&reqdb.wmo=

time range that was analyzed for this project. The TempAvg (°F) attribute was pulled from the Weather dataset. However, since this dataset was broken down by mm/dd/yyyy, the data shown in the data sample is the average temperature found from the temperature of each day in the specific month and year. The five-number summary for the data sample is shown below. The image of datatable below the Five-Number Summary table is the data sample for this project.

Five-Number Summary					
Attribute	**Minimum**	**1st Quartile**	**Median**	**3rd Quartile**	**Maximum**
TotalRides	9	126312	226303	314140	368096
DurationMin (min)	0	0	1.0004	1.0014	4.2785
DurationMax (min)	19.99	1429.17	1437.07	5674.38	83009.56
DurationTotal Minute (min)	101	1623805	3677496	5783351	8615605
GasPrices ($)	2.044	2.519	2.698	3.241	3.947
TempAvg (°F)	29.61	44.79	57.53	75.65	82.97

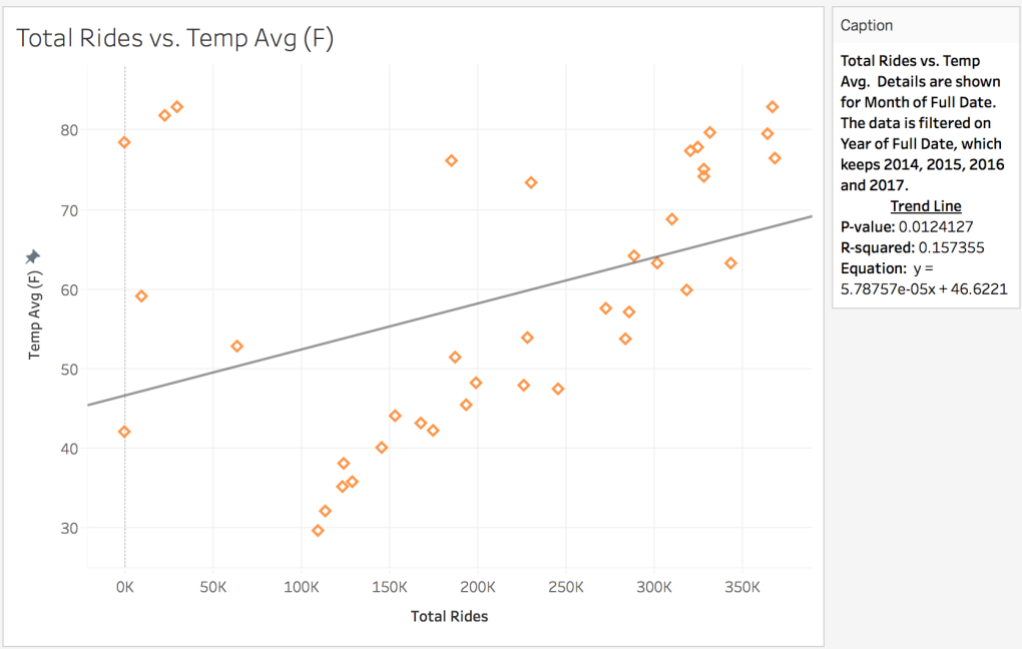
fullDate	year	quarter	month	monthNam	totalRides	durationMi	durationMa	durationTotalMinut	GasPrices	TempAvg
Jan-14	2014	Q1	1	January	114006	0	2883.8667	1318485.55	3.32	32.13
Feb-14	2014	Q1	2	February	124031	0	4208.4167	1548833.117	3.331	38.07
Mar-14	2014	Q1	3	March	167568	0	5879.2	2682490.267	3.573	43.13
Apr-14	2014	Q2	4	April	272490	0	9429.9667	5780149.233	3.783	57.53
May-14	2014	Q2	5	May	310152	0	6444.2667	5987454.717	3.865	68.74
Jun-14	2014	Q2	6	June	320527	0	12925.617	5786552.117	3.947	77.4
Jul-14	2014	Q3	7	July	331159	0	31044.183	6459677.533	3.943	79.61
Aug-14	2014	Q3	8	August	324219	0	24369.617	6217591.333	3.862	77.87
Sep-14	2014	Q3	9	September	327597	0	18394.9	5717735.767	3.773	74.13
Oct-14	2014	Q4	10	October	301244	0	4178.7833	4894477.1	3.414	63.19
Nov-14	2014	Q4	11	November	199298	0	4759.4	2742103.85	3.097	48.17
Dec-14	2014	Q4	12	December	153221	0	12640	2023574.967	2.81	44.03
Jan-15	2015	Q1	1	January	128592	0.0026	83009.557	1608081.786	2.3	35.84
Feb-15	2015	Q1	2	February	109766	0.00213	14911.237	1309053.725	2.237	29.61
Mar-15	2015	Q1	3	March	193107	0.01248	5469.5558	3002383.229	2.737	45.55
Apr-15	2015	Q2	4	April	318127	1.00043	1430.0153	6566657.118	2.731	59.9
May-15	2015	Q2	5	May	230035	1.00088	1413.8345	4521389.103	2.978	73.45
Jun-15	2015	Q2	6	June	9	3.2198	56.37152	251.9606	3.097	78.4
Jul-15	2015	Q3	7	July	22944	1.00948	1072.6996	392508.1176	3.163	81.77
Aug-15	2015	Q3	8	August	364313	1.00153	1434.7764	7037188.937	3.038	79.48
Sep-15	2015	Q3	9	September	328038	1.00003	1406.36	5836440.382	2.697	75.1
Oct-15	2015	Q4	10	October	9338	1.00812	1254.4632	191881.2331	2.44	59.16
Nov-15	2015	Q4	11	November	228296	1	1416.8627	3605309.484	2.432	53.9
Dec-15	2015	Q4	12	December	187357	1.00077	1430.2235	2793105.946	2.392	51.45
Jan-16	2016	Q1	1	January	123252	1.00133	1437.0141	1639527.255	2.329	35.23
Feb-16	2016	Q1	2	February	145654	1.00217	1401.768	1996177.073	2.044	40.1
Mar-16	2016	Q1	3	March	283493	1.00042	1438.4368	5509922.081	2.171	53.77
Apr-16	2016	Q2	4	April	285516	1.0005	1437.196	5539812.454	2.284	57.13
May-16	2016	Q2	5	May	288720	1.00238	1439.2839	5760438.448	2.502	64.13
Jun-16	2016	Q2	6	June	368096	1.00047	1437.0664	7908729.214	2.614	76.5
Jul-16	2016	Q3	7	July	366435	1.00055	1436.7592	8615604.545	2.634	82.97
Aug-16	2016	Q3	8	August	29484	1.01725	1429.7843	565222.0982	2.551	82.94
Sep-16	2016	Q3	9	September	185424	1.0006	1438.9243	3677496.049	2.643	76.2
Oct-16	2016	Q4	10	October	343243	1.00012	1436.991	6494365.77	2.684	63.32
Nov-16	2016	Q4	11	November	64043	1.01085	1426.6996	1163541.881	2.624	52.87
Dec-16	2016	Q4	12	December	9	4.27853	19.99145	100.76927	2.536	42.1
Jan-17	2017	Q1	1	January	174804	1.00077	1433.9379	2592321.913	2.679	42.32
Feb-17	2017	Q1	2	February	226303	1.00205	1428.5594	3948416.01	2.698	47.96
Mar-17	2017	Q1	3	March	245401	1.00007	1434.4424	4352267.267	2.798	47.42

Methodology

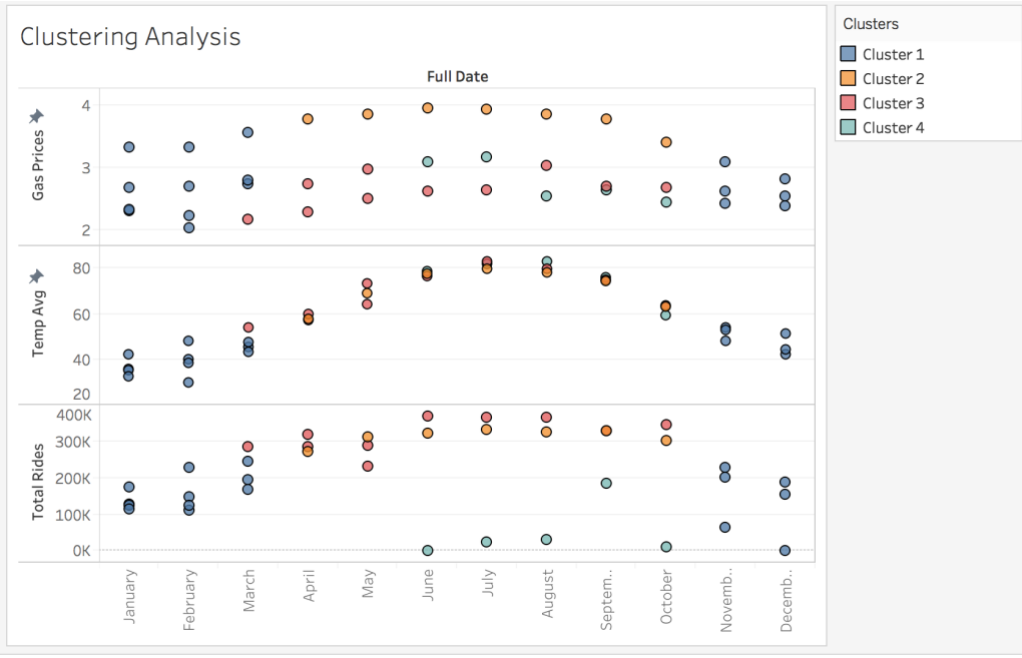
After the data sample was created in R, it was saved as a .CSV file then copied over to a .txt file to be uploaded into Tableau Public. From here, the graphics were made to show the trend lines and time series plot. The trend lines that were analyzed compared the clusterings of Gas Prices given Total Rides as well as Temp Avg (°F) given Total Rides to see if there was any correlation.

The graph below shows the trend line for Gas Prices compared to Total Rides taken. The equation for the trend line, R-squared value, as well as the p-value are shown in the caption of the graph. The trend line for this graph shows that as the number of Total Rides increase, the price of gas also increases.

The graph below shows the trend line for Temp Avg (°F) compared to Total Rides taken. The equation for the trend line, R-squared value, as well as the p-value are shown in the caption of the graph. The trend line for this graph shows that as the number of Total Rides increase, the average temperature (°F) also increases.



To cluster the data, we used Tableau's clustering option and clustered the data using the averages of the variables by month. The graph below shows the data clustered following this format along with the inputs for clustering and analysis of variance shown in the image below the graph. Using the ANOVA F-statistic test with a significance level of α=0.05, the clustering p-values for all three of the variables are less than the significance level. This means that it rejects the null hypothesis, that there is no correlation between the variables, and supports the hypothesis that there is a correlation between the variables.



Project Goal

Seeing the trend of riders/drivers in the DC area in the course of the last three years (Q1 2014 to Q1 2017). Allowing to see this data can be predicted to see the for coming future riders/drivers. Also see market trend on what to expect the heavy amount of people to use during what time of the year or what occasions may have happened to see where traffic occurs. This data can be used for advertisers to capture the most amount of traffic of people to see their ads or allow the D.O.T (Department of Transportation) focus on what people use the most and focus more construction needed in those areas or have more access to those modes of transportation.

Having this data in hand for some people will allow to make predictions to know which is the best mode of transportation to take when in the city or just be curious the best and most popular transportation. Allowing the user with the knowledge to save time when traveling around the city or to and from the city.

Inputs for Clustering

Variables:	Avg. Gas Prices
	Avg. Temp Avg
	Avg. Total Rides
Level of Detail:	Month of Full Date
Scaling:	Normalized

Summary Diagnostics

Number of Clusters:	4
Number of Points:	39
Between-group Sum of Squares:	7.4471
Within-group Sum of Squares:	2.8242
Total Sum of Squares:	10.271

		Centers		
Clusters	**Number of Items**	**Avg. Gas Prices**	**Avg. Temp Avg**	**Avg. Total Rides**
Cluster 1	17	2.7022	42.934	1.5204e+05
Cluster 2	7	3.7981	71.21	3.1248e+05
Cluster 3	10	2.6333	68.575	3.176e+05
Cluster 4	5	2.7788	75.694	49440.0

Analysis of Variance:

			Model		Error	
Variable	**F-statistic**	**p-value**	**Sum of Squares**	**DF**	**Sum of Squares**	**DF**
Avg. Total Rides	9.078	0.0001392	2.77	3	3.56	35
Avg. Temp Avg	8.795	0.0001756	2.718	3	3.606	35
Avg. Gas Prices	7.358	0.0005996	1.959	3	3.106	35

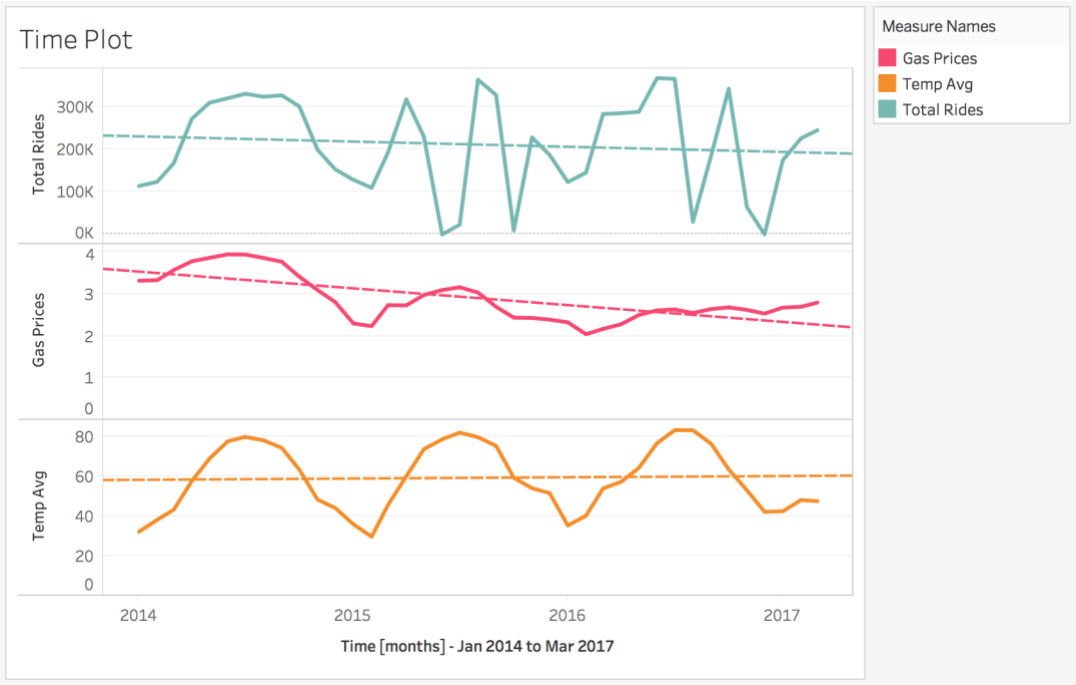
Software Used

Presentation software used: Prezi

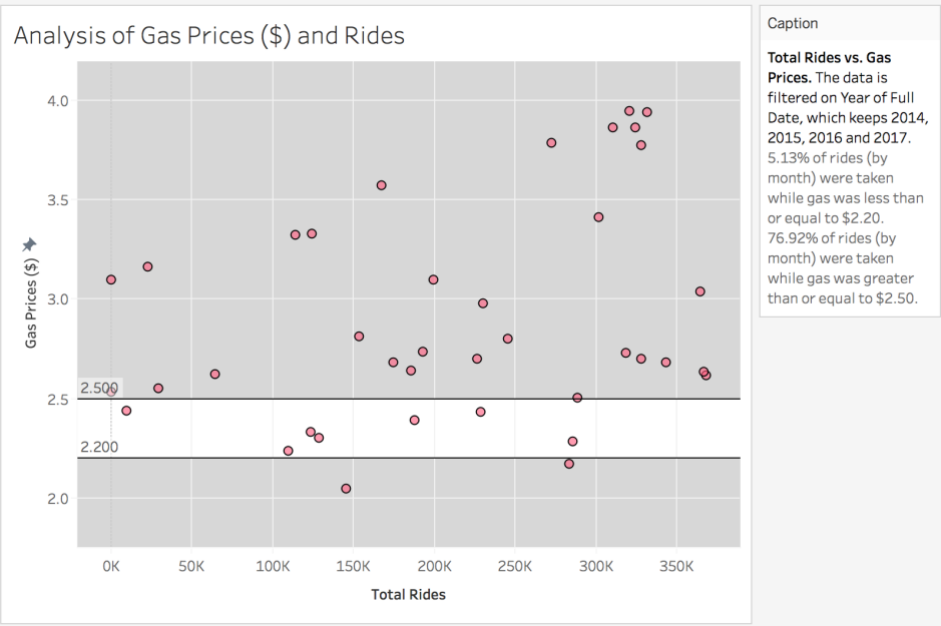
Data compiled into, manipulated, and visualized: R and Tableau
R: https://github.com/madelinew/IT385R
Tableau: https://public.tableau.com/profile/maddie.warndorf#!/vizhome/IT385Real/Dashboard1?publish=yes
File format from the data: CSV files

Analysis and Result

Time Plot



The graph above shows a time plot from January 2014 to March 2017 with all three variables graphed with their own trend line. This graph shows that there is a correlation of as the price of gas increases, the total number of rides increase. This same correlation is shown with the rise and fall of the temperature. The trend lines on each of the graphs allows the individual to see if throughout time if there was an increase or decrease in the variable. The trend line of Total Rides throughout time is slowly decreasing, however this same decrease is shown in the price of gas as well. However, it is not shown in the temperature. The trend line for the temperature throughout the time analyzed has a positive slope which means that the temperature in the Washington D.C. area has slowly increased.



Influence of Gas Prices

Another question that we were interested in analyzing was if more people would use BikeShare if the gas price was above $2.50 or below $2.20. Our hypothesis was that there would be more people choosing BikeShare if it was above $2.50 and less if it was below $2.20. The graph above shows that 76.92% of rides were taken while gas was greater than or equal to $2.50 and 5.13% of rides were taken while gas was less than or equal to $2.20. This proves that our hypothesis was right about gas possibly being a contributing factor to influencing people to choose BikeShare.

Conclusion

The data sample supported the hypothesis that there would be more BikeShare riders during times where the price of gas was high as well as when the temperature was warm. The graphs constructed in Tableau support this statement along with the statistical analysis also conducted.