Pittsburgh Bike Share Ridership Demand Prediction

Milestone Project

Sneha Krishnamurthy

Introduction:

Bike sharing systems have been increasing in demand over the past two decades as a result of rapid advancements in technology. Healthy Ride is a public bicycle sharing system that serves parts of Pittsburgh to fulfill the growing need for changes in mobility pattern. Healthy Ride is operated by Pittsburgh Bike Share and has plans for expansion to reach new neighborhoods by adding more stations, including several electric bikes to help riders navigate Pittsburgh's hilly geography, located throughout the city.

Problem:

Bike-sharing systems are used world-wide. Given that the system tends to be unbalanced, there are challenging analytical issues such as accurately predicting the demand. This project explores on predicting the total number of bikes rented from individual stations on any given day.

Clients:

Bike sharing operators can use this model to proactively shape the mobility market by forecasting demand prediction and to meet customer expectations.

Datasets:

Healthy Ride operated by Pittsburgh Bike Share data:
https://data.wprdc.org/dataset/healthyride-trip-data
This dataset includes bike number, membership type, trip start and end timestamp, and origin and destination information as features. A trip is defined as any valid rental one minute or longer that begins and ends at a Healthy Ride station. The combined dataset has roughly about 285455 rows of data.

Bike Station data:
https://healthyridepgh.com/data/
This dataset includes StationNum , StationName, RackQnty, Latitude, Longitude as columns.

Weather data: and **Storm data:**:
NOAA makes available their daily weather station data (I used station ID FIPS:42003) to extract the data.

Bike score, Transit Score, Walk score Data:
https://www.walkscore.com

Bike Score service measures whether a location is good for biking on a scale from 0 - 100 based on four equally weighted components: -Bike lanes, Hills, Destinations and road connectivity. **Transit Score** is a patented measure of how well a location is served by public transit on a scale from 0 to 100. **Walk Score** measures the walkability of any address.

Data Cleaning and Data Wrangling:

Step 1:

The csv files downloaded as Zip files from **Healthy Ride operated by Pittsburgh Bike Share data** shows trips taken using the Healthy Ride system by quarter. I have selected csv files from 2015, Quarter 2 to 2018, Quarter 4. The consolidated csv files were loaded as pandas dataframe and saved as "Rentals" dataframe did not include Latitude and longitudes as features. However, the **Bike Station data** had those data fields as shown in the snippet below:

	StationNum	StationName	RackQnty	Latitude	Longitude
0	1000	Liberty & Stanwix	16.0	40.441326	-80.004679

I extracted unique matching longitude and Latitude columns based on station id and merged the coordinates with "Rentals" dataset.

Step 2:

To fetch the scores such as Bike Score, Transit score, and Walk score information from this website for specific bike stations, the python code had to perform multiple HTTP requests from a single query. The latitudes and longitudes extracted from "Healthy Ride Bike Stations" dataset as explained in Step 1 were first saved as a text file. The responses from API were extracted by looping through every single line in the text file containing matching latitudes and longitudes.

Step3:

The Transit score and bike score column had irrelevant data and only the numeric score was retained. "Rentals" data frame was joined with "Transit score" data based on the common station ids. I extracted these station ids by merging "Station's" dataset with "Transit Score" dataset using common latitudes & longitudes (rounding it off to 4 digits after the decimal point) in both datasets.

Step 4:

Seasons (fall, summer, winter, spring) were segregated based on the fixed dates of the solstices and equinoxes.

Step 5:

Federal holidays (1" if it is a holiday or "0") were mapped to the main data frame. This was done by creating an instance of class custom business day and using this as frequency to extract federal holidays

Step 6:

Distances between stations are not included in Healthy Ride Bike share's data. To find distances based on Latitude and Longitudes, I used the Haversine's formula . Later, the trips with '0' distance were dropped since the start station id and the end station id were same and only the person who rented the bike knows about the distance covered and the duration of the trip taken per rental.

Step 7:

A separate "weather" column was added to the main data frame after scoring, based on windspeed, rain, fog, lightening, temperature, and also event types such as tornados, hurricanes as below:

0 = Worst weather including all the event types listed in the Storm dataset
1 = Moderate weather including
2 = Good weather not listed in either 1 or 2

Step 8:

What is the number of bike rentals per day? This would be our target variable. To do this, a pandas group-by function was performed for counting the number of trips per day.

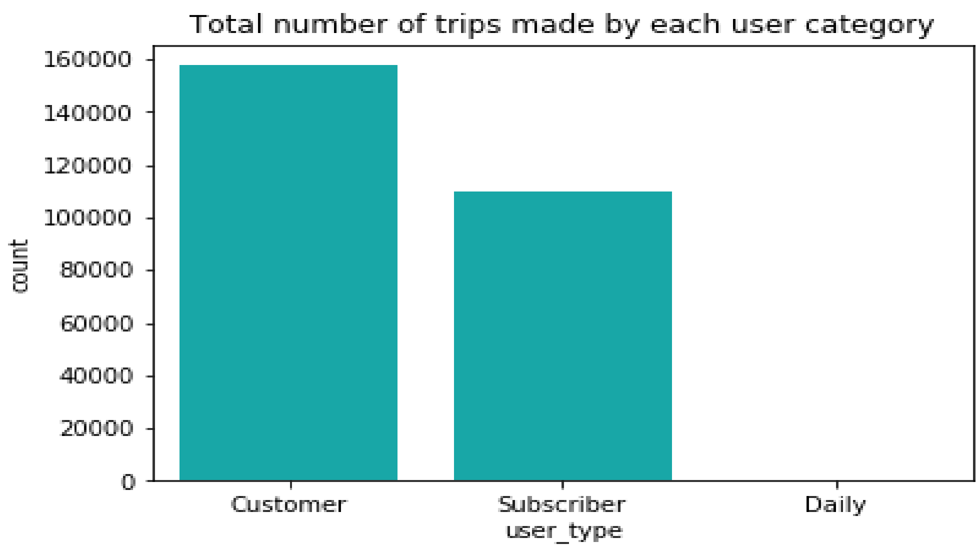
The link to *IPYNB* files to perform these data wrangling and data cleaning steps is provided below:
https://docs.google.com/document/d/1JQyyyJKpjudR04VyMZ8NL-8Sc49_jiWi5CqlHyhfHN0/edit?usp=sharing

After ensuring that all features are of the correct data type, and dropping duplicate rows, the dataframe was saved as a csv file to perform further Exploratory Data Analysis.

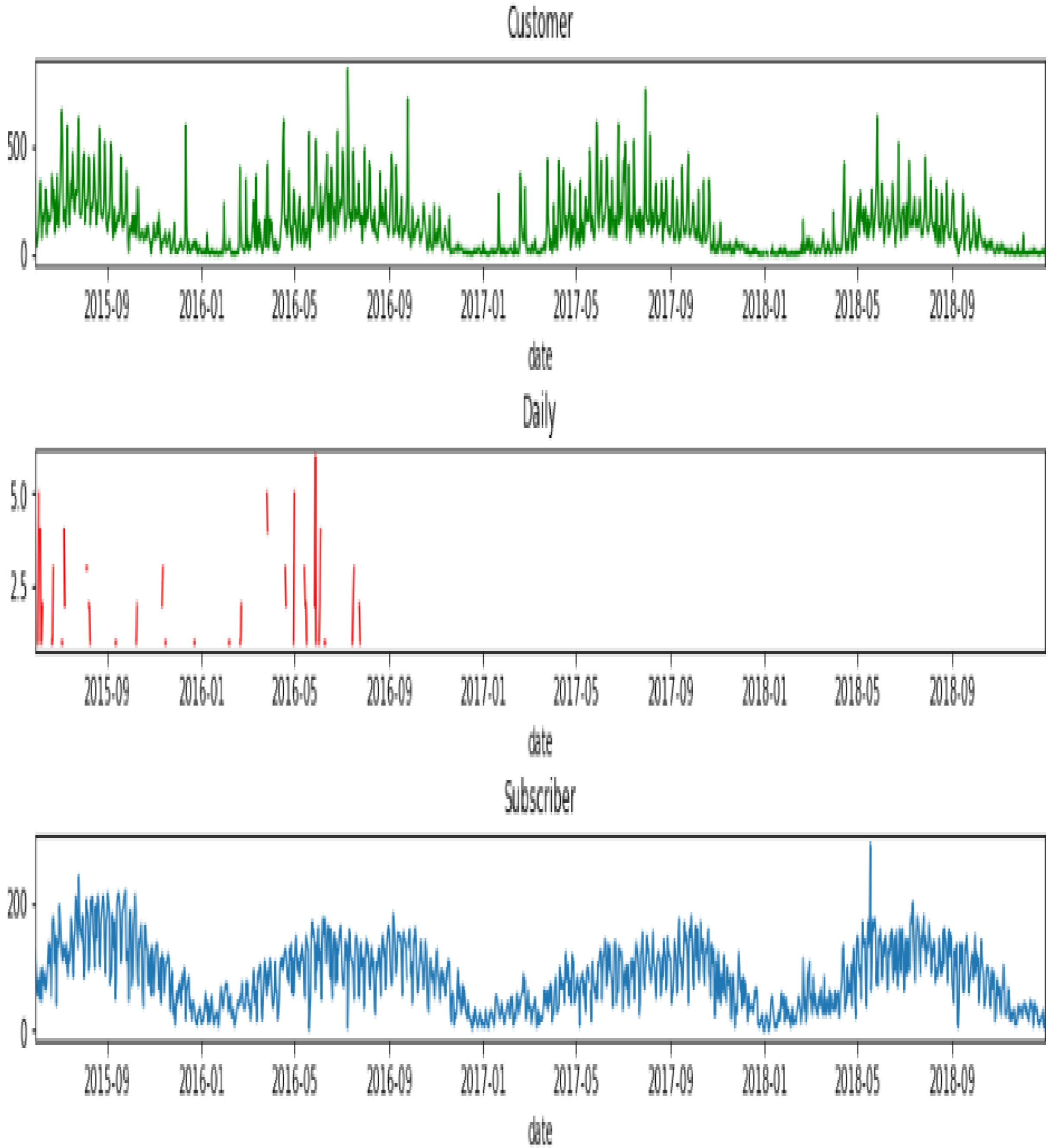
Data Visualization and EDA:

Total number of trips made by each user category:

The total number of trips made by each user category over the course of the time period from when the program went live on July 1st, 2015 to Dec 31st, 2018 are depicted in the chart below:



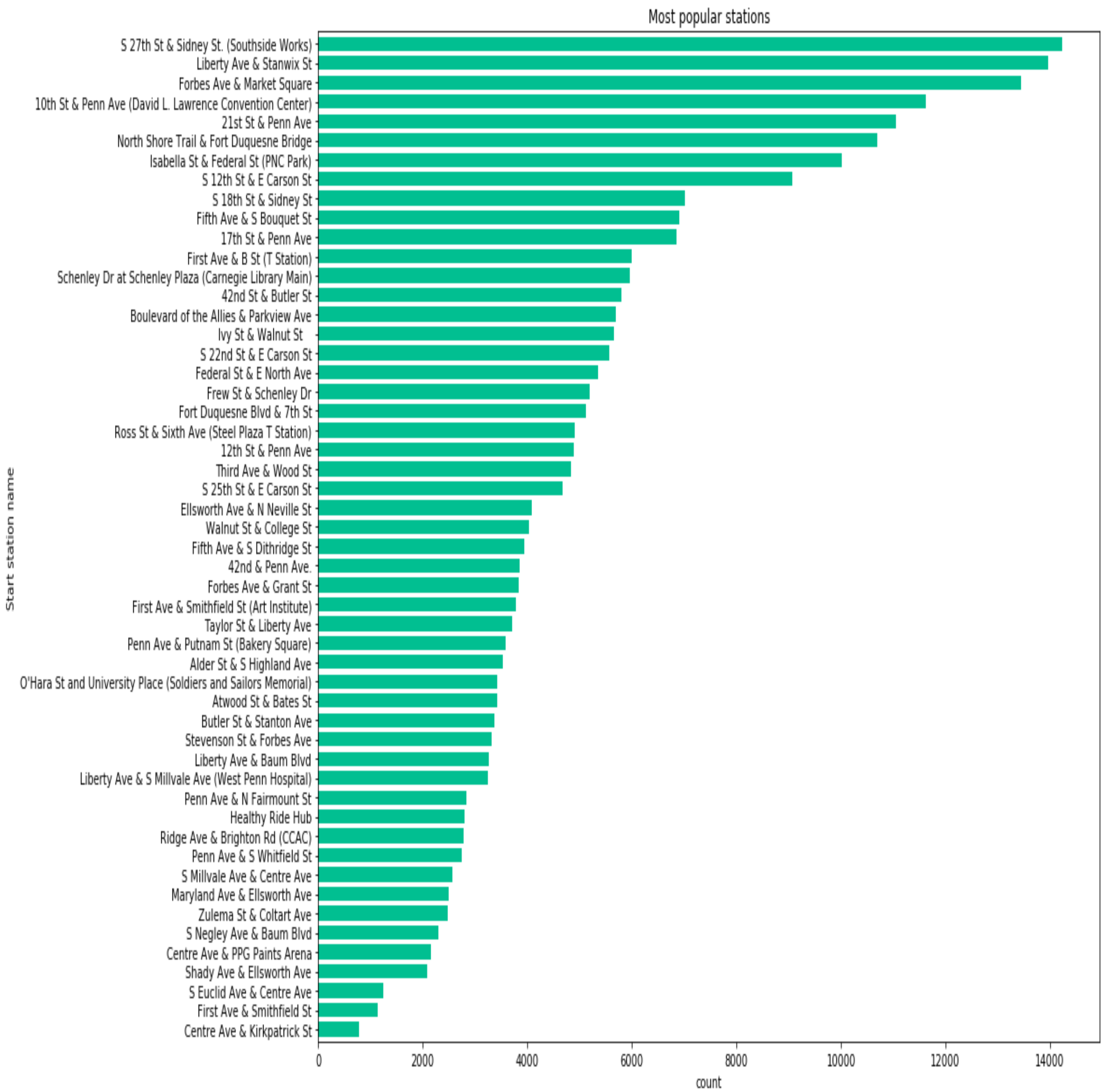
Compared to other user types such as Customer passes and Subscriber passes, Daily pass user type counts looks negligible. The plot below shows the daily trends (trip counts by date), separated between Customers(top), Daily(middle), and Subscribers(bottom).



According to Healthy rides Bike Share website, Subscribers are all standard and deluxe monthly members and Customers are Pay-As-You-Go riders. There is no corresponding category for daily passes on Healthy ride website. It can be concluded from the graph above that the daily passes were issued inconsistently from 2015, Quarter 2 to 2016, Quarter 2. Since there were only 186 instances out of total 267260 rider counts, it was excluded from further analysis.

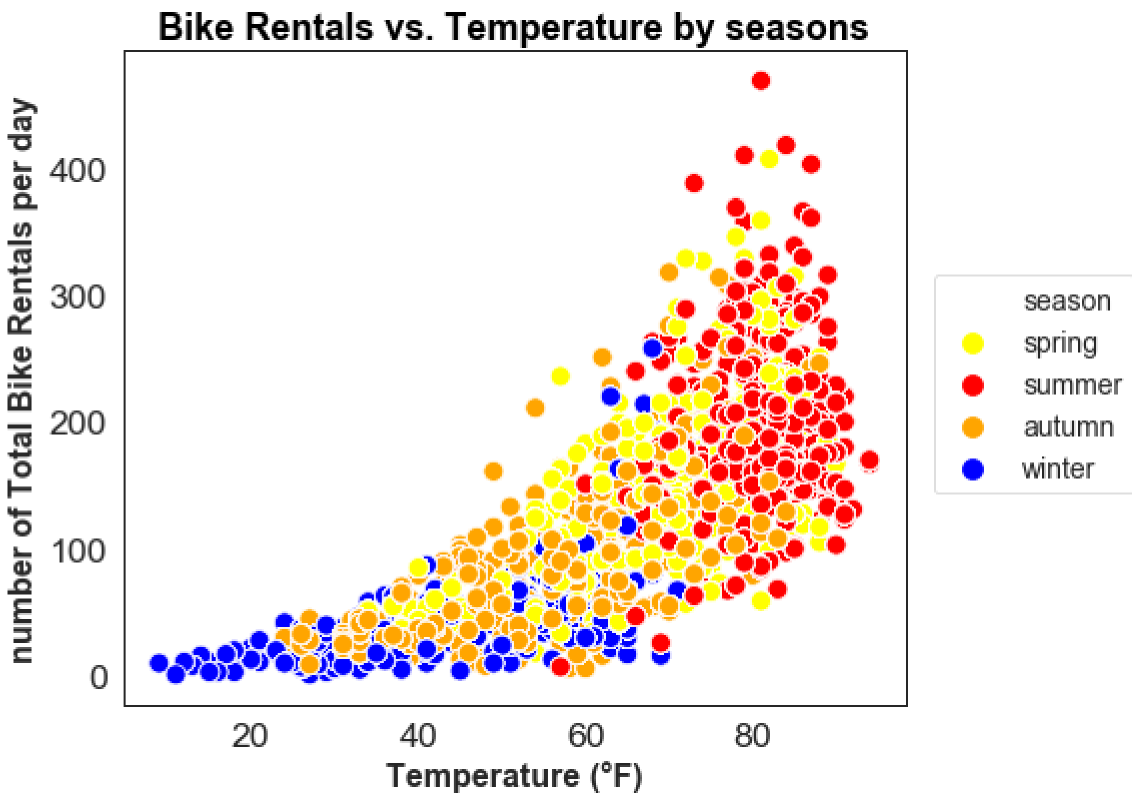
Most popular start stations:

The horizontal bar plot below helps us to visualize the top 15 most popular start stations based on the trip counts. “S 27th St & Sidney St. (Southside Works)” seems to be most popular followed by” Liberty Ave & Stanwix St”. “Centre Ave & Kirkpatrick St” seems to be the least popular.



Centre Ave & Kirkpatric St , First Ave $ Smithfield St, S Euclid Ave & Center Ave seems to be the least popular checkout stations.

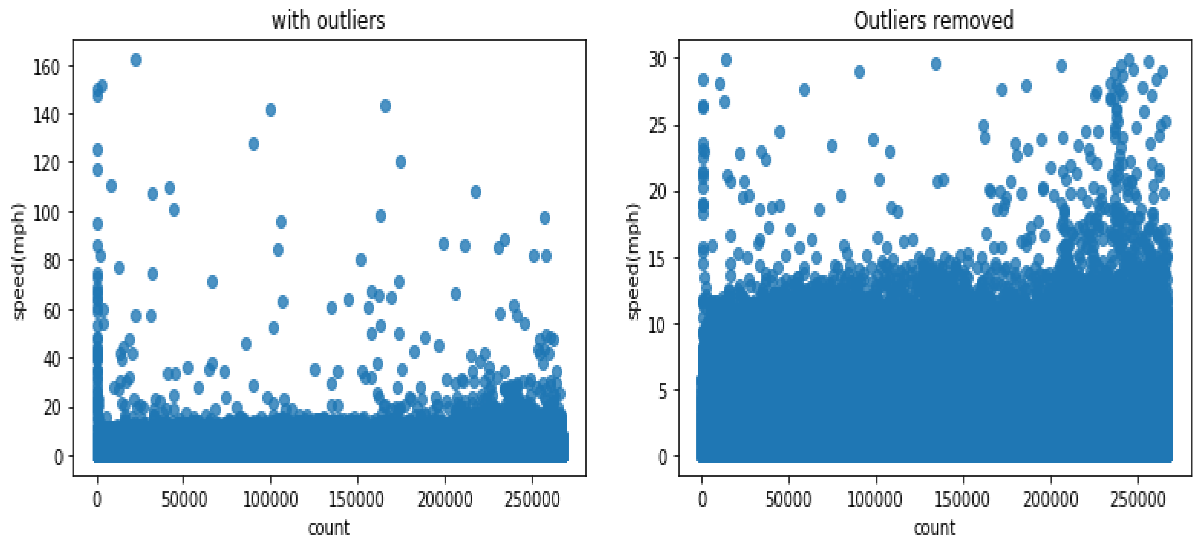
Bike Rentals and Temperature:



The multivariate scatter plot above shows that as the temperature increases, the count i.e. the number of total rentals per day also increases. There is a strong linear relationship between temperature and bike rentals. The maximum number of rental counts seems to be when the temperature is between 65°F to 90°F**.** There is a clear seasonal trend where the total rental bikes seem to decrease during Winter and increase during summers.

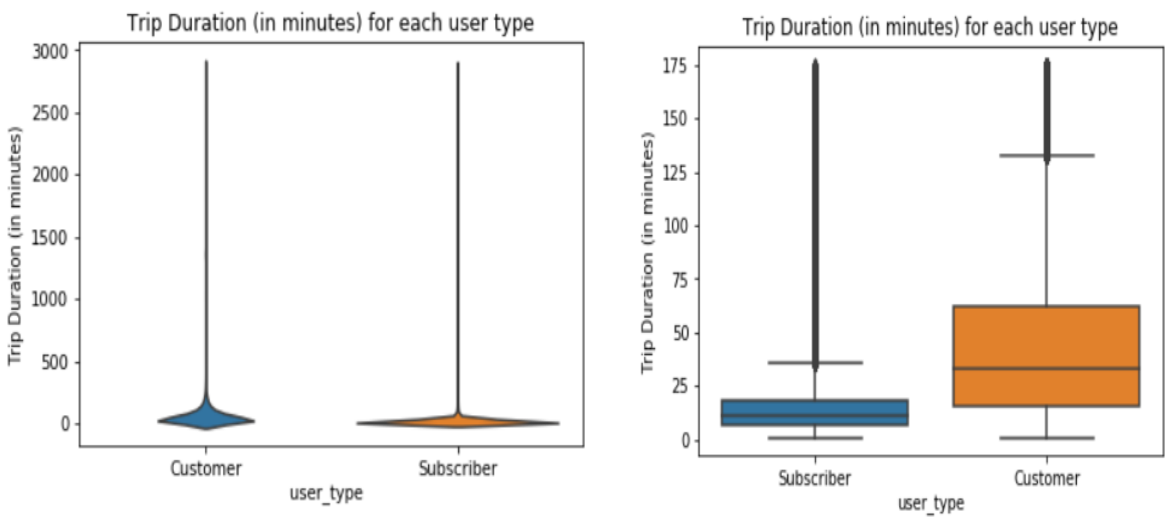
Speed & Outliers:

I added a new column "speed" to find any outliers.70771 zero values recorded where start station and stop station are same and distance was 0. To find outliers, seaborn plot was used by plotting speed in miles/hour in y-axis and count in horizontal axis. The plot below (left) depicts that the maximum speed ever recorded is 161 m/h. Even professional bicycle racers can usually maintain 25-28 mph on flat ground. Obviously, values above 30 mph are outliers as shown in the seaborn regplot below:



After dropping rows with values over 30 mph speed as shown in the plot above(right), the dataset reduced to 196427 rows of data.

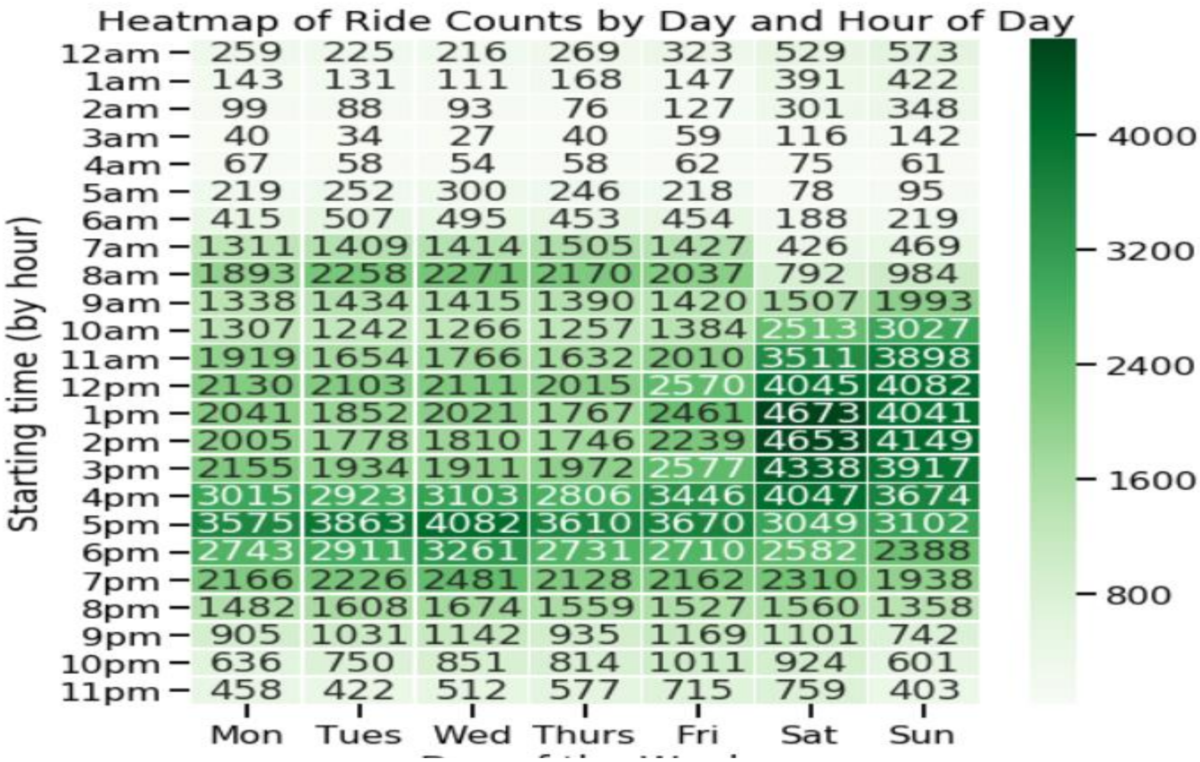
Trip duration (in mins) before and after removing outliers:

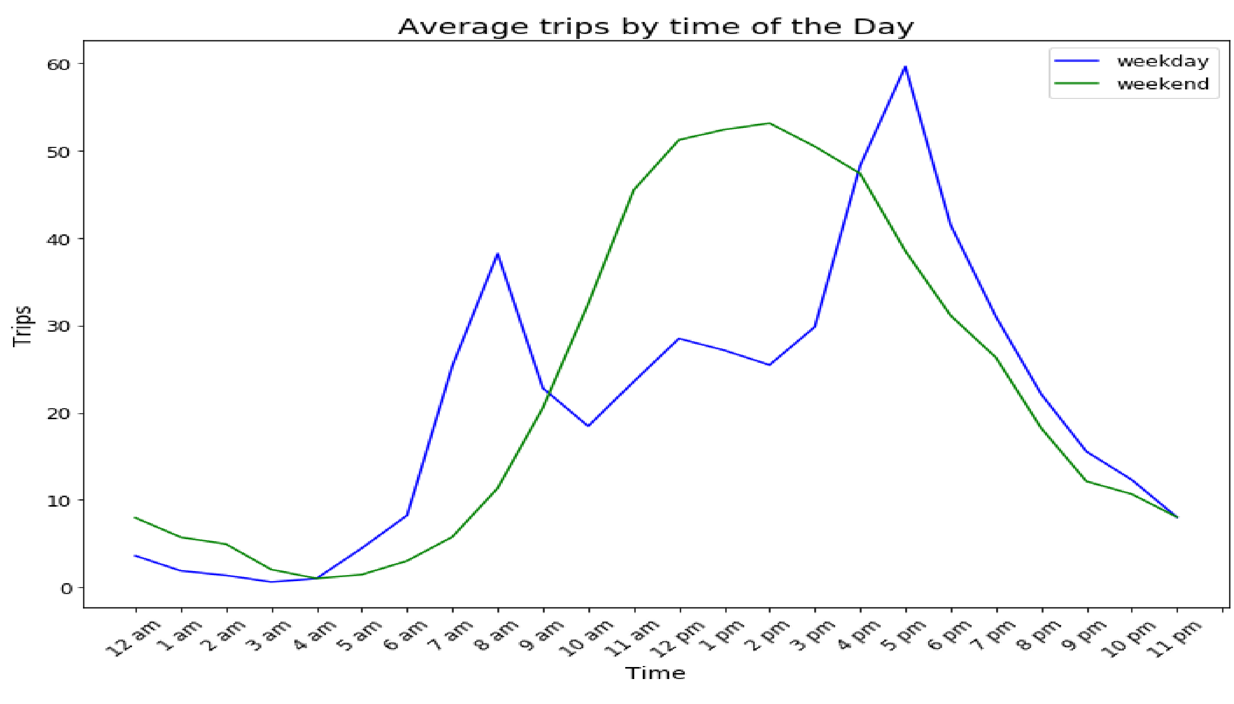


The box plot of trip duration in minutes for each user type has a lot of anomalies as depicted above(left). Trip duration over 175 minutes do not represent the vast majority of users. Hence, trips over 175 minutes long will be excluded from further analysis. Trip duration in minutes based on user type after removing few anomalies is shown in the box plot above(right).

Trips by time of the day:

The heat map below shows the total ride counts by hour of the day and Day of the week. On weekdays, Monday to Friday, most rides are taken from 7am to 7pm.On weekends, most rides are taken from 9am - 7pm.

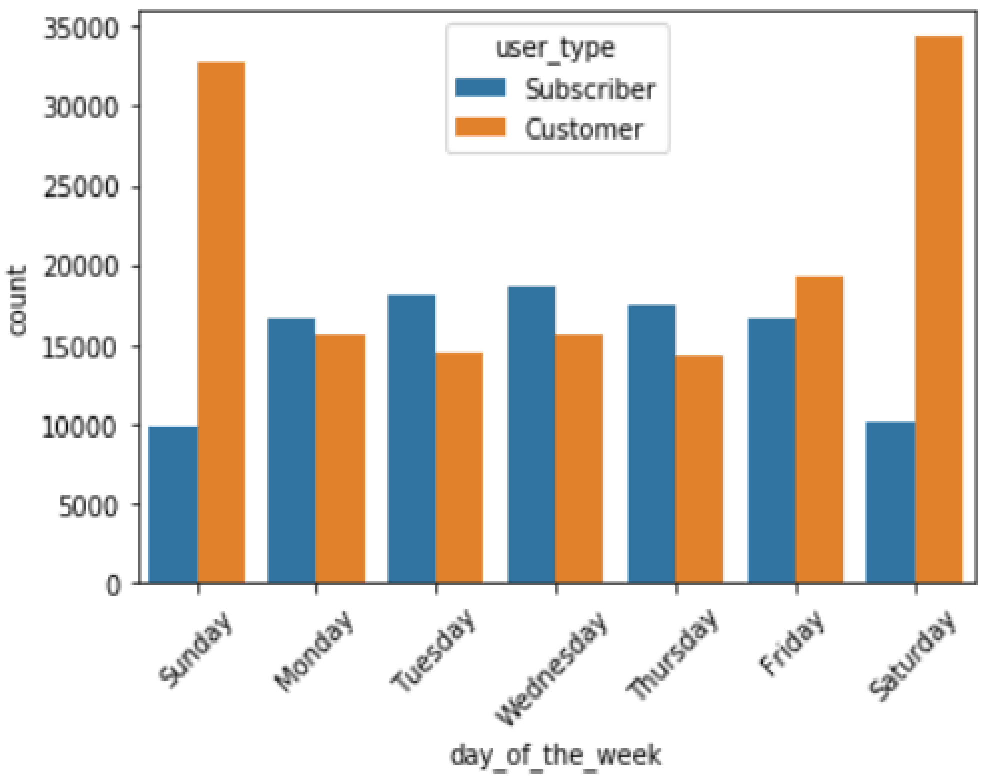




The graph above shows the usage of bikes by hour of the day for an average day and also weekends. We see that the usage takes off at about 6 a.m. during work days in the morning till it climbs to its first peak at around 7 and then drops. This corresponds well with people going to work in the morning. Then, at about 3 p.m., usage starts to increase steadily. The maximum is reached at about 5 p.m. Again, this fits the fact that people use the bikes to return back home after work during work days. Thereafter, there is a decline until the minimum at around 5 a.m. On weekends however, there is a steady increase from 7 am. It peaks from 11 a.m. to 4 p.m. which corresponds well with people using bikes for recreation during weekends.

User types and day of the week:

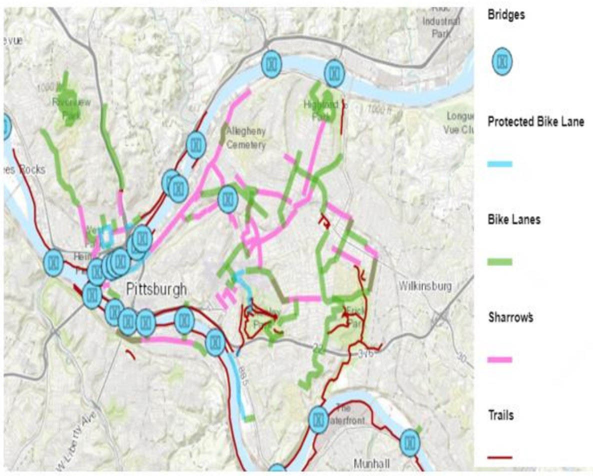
It is evident from the plot below that more customers rent bikes than subscribers on Saturdays and Sundays. It fits well with the fact that subscribers use bikes more on weekdays than on weekends since they might be renting bikes to commute to office.



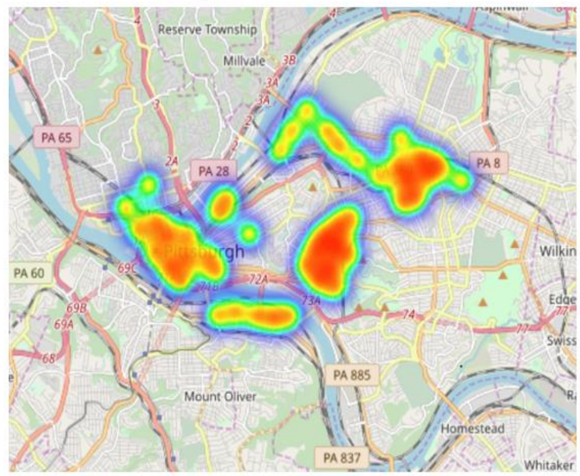
Heatmap of bike usage in check-out stations in Pittsburgh:

The geographic map below on the left shows the concentration of bridges , shared lane markings (sharrows), trails (brown lanes), bike lanes (green lanes) is extracted from this [website.](#) The geographic plot below (right) shows the concentration of bike checkouts demonstrated using Folium to visualize the number of bike checkouts using latitudes and longitudes.

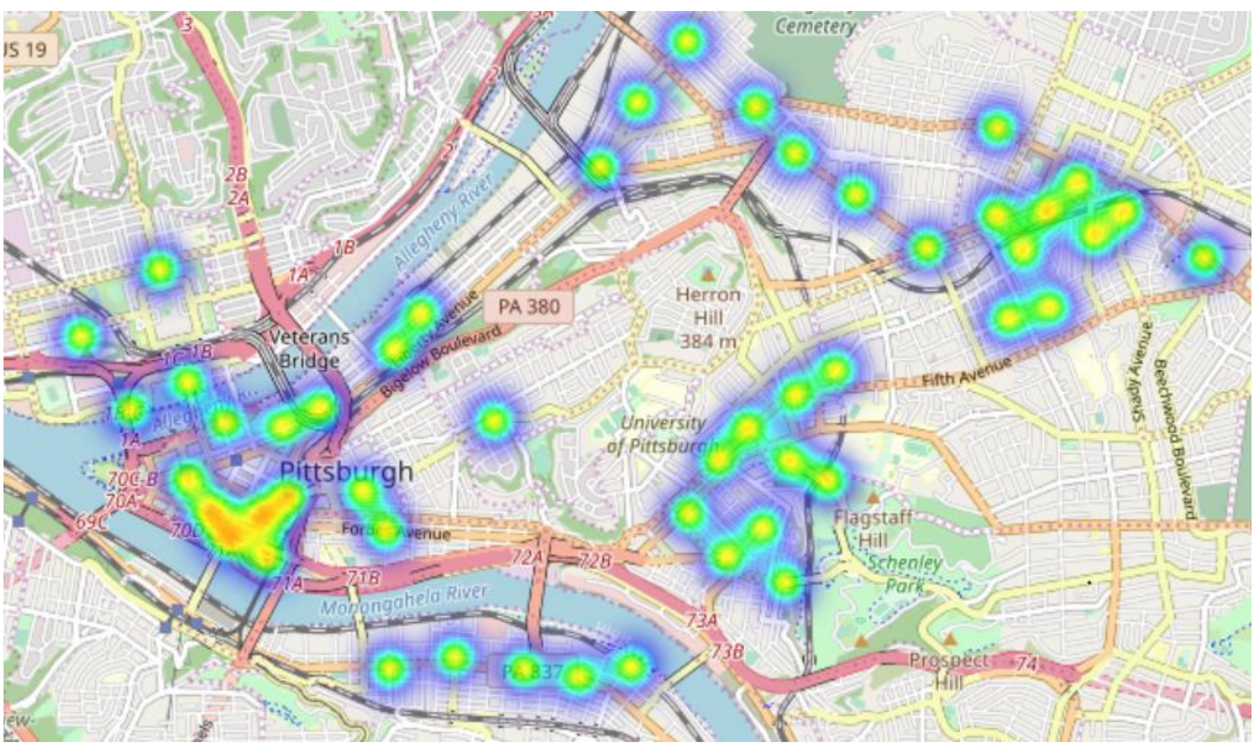
Pittsburgh Bike Infrastructure



Heatmap of bike check-outs

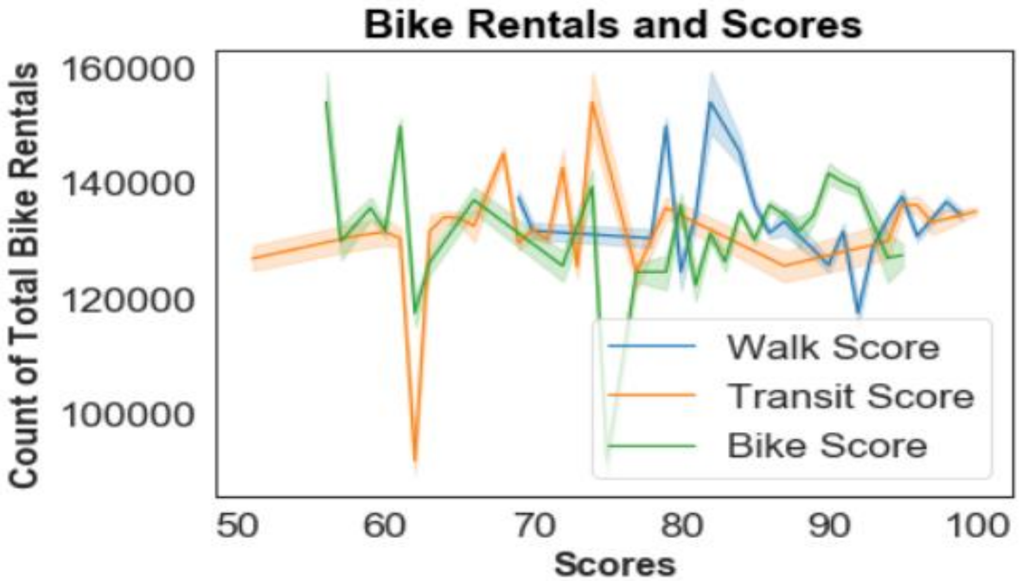


And if we zoom into the map, concentrations of checkouts are evident in the three distinct hotspots near the lake where Gateway Station, First Avenue and Steel Plaza are located.

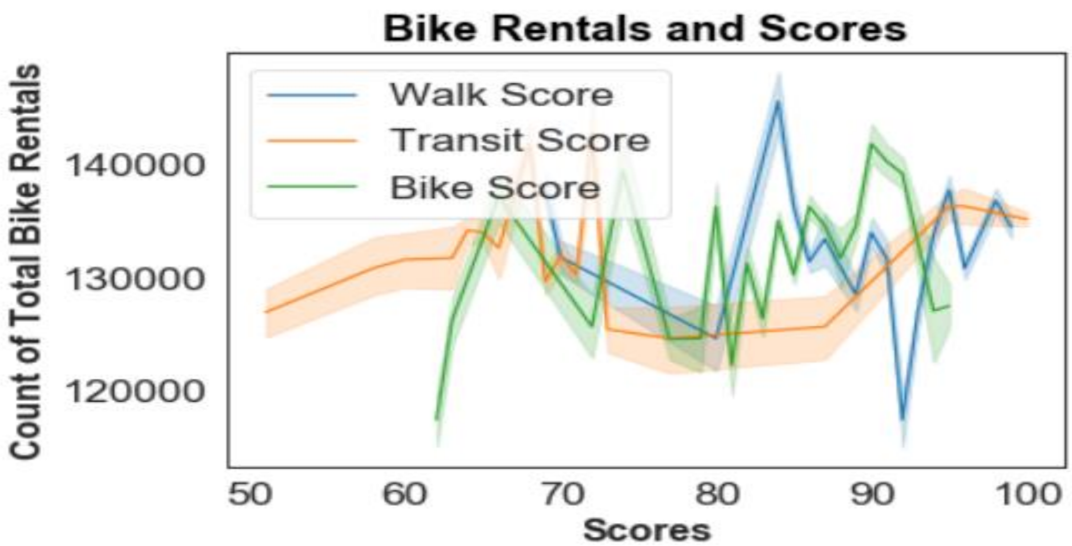


Bike Rentals and Scores:

The graph below shows the relation between the rental counts to their respective bike,transit and walk scores.



As per [website](), the location with bike score between 70 - 89 is considered as very bikeable. But the graph above shows anomalies at 75 and the bike rental counts drops drastically. Hence I will be deleting rows containing bike score of 75, and transit scores less than or equal 61. The plot below shows the relation between rental counts and score after removing anomalies. A total of 163377 rows will be retained for further analysis.



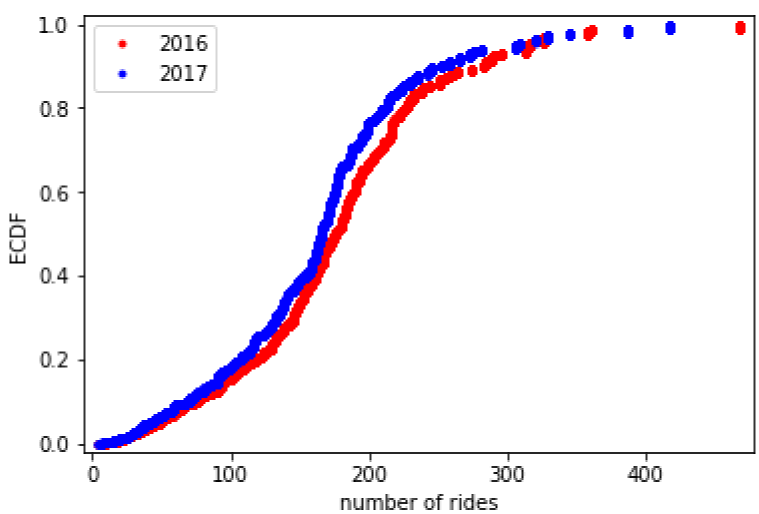
Inferential Statistics and Exploratory Data Analysis:

As per the article posted in Pittsburgh Post-Gazette, published on DEC 3, 2018; there was a noticeable drop in ridership from 2016 to 2017. I wanted to test whether this hypothesis is commensurate with the data. My consolidated dataset has data from 2015, Quarter 2 to 2018, Quarter 4.

Before conducting any statistical analyses, conditions of inference for comparing two means has to be met which are:
1) The two samples are random and they come from two distinct populations. The samples are independent.
2) Both populations are Normally distributed.

Our samples are independent of each other and are random. I performed a two-sample bootstrap hypothesis test to find if the difference in means is actually significant. I also performed a two sample, one-tailed upper test Z test by setting the alpha level to be 5%. Alpha level is the probability of making the wrong decision when the null hypothesis is true.

Let's perform exploratory analysis using ECDF plots first. ECDF is Empirical Cumulative Distribution Function. x- value of ECDF is the quantity you are measuring. The y value (0 to 1 probability) is the fraction of datapoints that have a value smaller than the corresponding x value.



Notice that the Ecdfs do not overlap except a few of suggesting that the hypothesis is not commensurate with the data.

They are normally distributed and the sample size (n_1) of first sample is 46112 and the sample size (n_2) of second sample is 42020

The average number of trips ($\bar{x}_1$) of the year 2016 was 177.53, and that of year 2017 ($\bar{x}_2$) was 164.87 with a difference of 12.66. It is possible this observed difference in mean of trip counts was by chance. We will compute the probability of getting at least a 12.66 difference in average number of trips under the hypothesis that the average number of trips counts in both years are identical. For our hypothesis to be true, we are going to figure out the probability of getting the actual difference of means between 2016 and 2017 to be zero under the assumption that our hypothesis is correct. It is going to be a one-tailed upper test. I will set alpha to be 0.05.

Null Hypothesis: - H_0: $\mu_1 - \mu_2$

Alternate Hypothesis: - H_a: $\mu_1 > \mu_2$

Our data comes from two random samples or two groups in a randomized experiment, since the population means ($\mu_1 > \mu_2$) are unknown, the sample means ($\bar{x}_1$ - $\bar{x}_2$)) are used to make inferences.

I performed a two sample, one-tailed upper test Z test. The z-score or z- statistic is calculated using the formula:

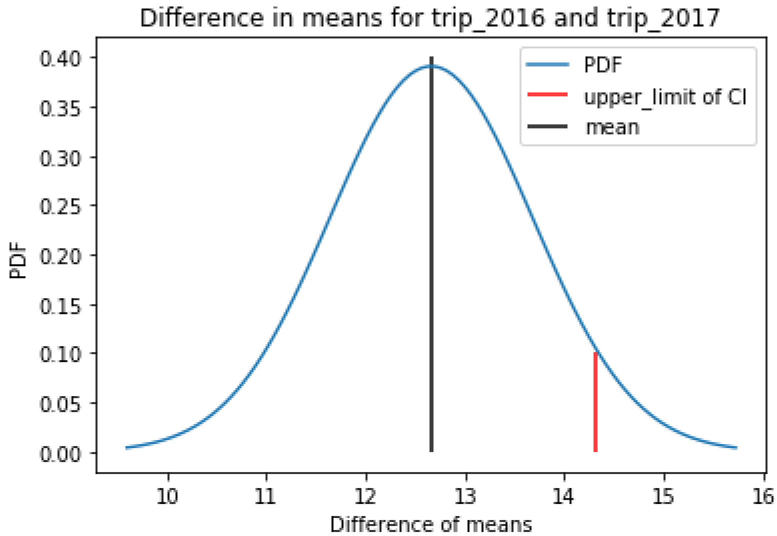
$$z_{stat} = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_0}{\sqrt{\frac{({\sigma_1}^2 - {\sigma_2}^2)}{n_1 - n_2}}}$$

The sample means are denoted as $\bar{x}_1$ and $\bar{x}_2$). The sample sizes are denoted as n_1 and n_2. The parameter of interest is μ_0, the hypothesized population mean, which in our case is 0 (the difference between μ_1 *and* μ_2). σ_1 and σ_2 are the standard deviations of first and second samples.

$$SE = \sqrt{\frac{({\sigma_1}^2 - {\sigma_2}^2)}{n_1 - n_2}}$$

SE is the standard error of the distribution of differences which is also the standard deviation of that sampling distribution

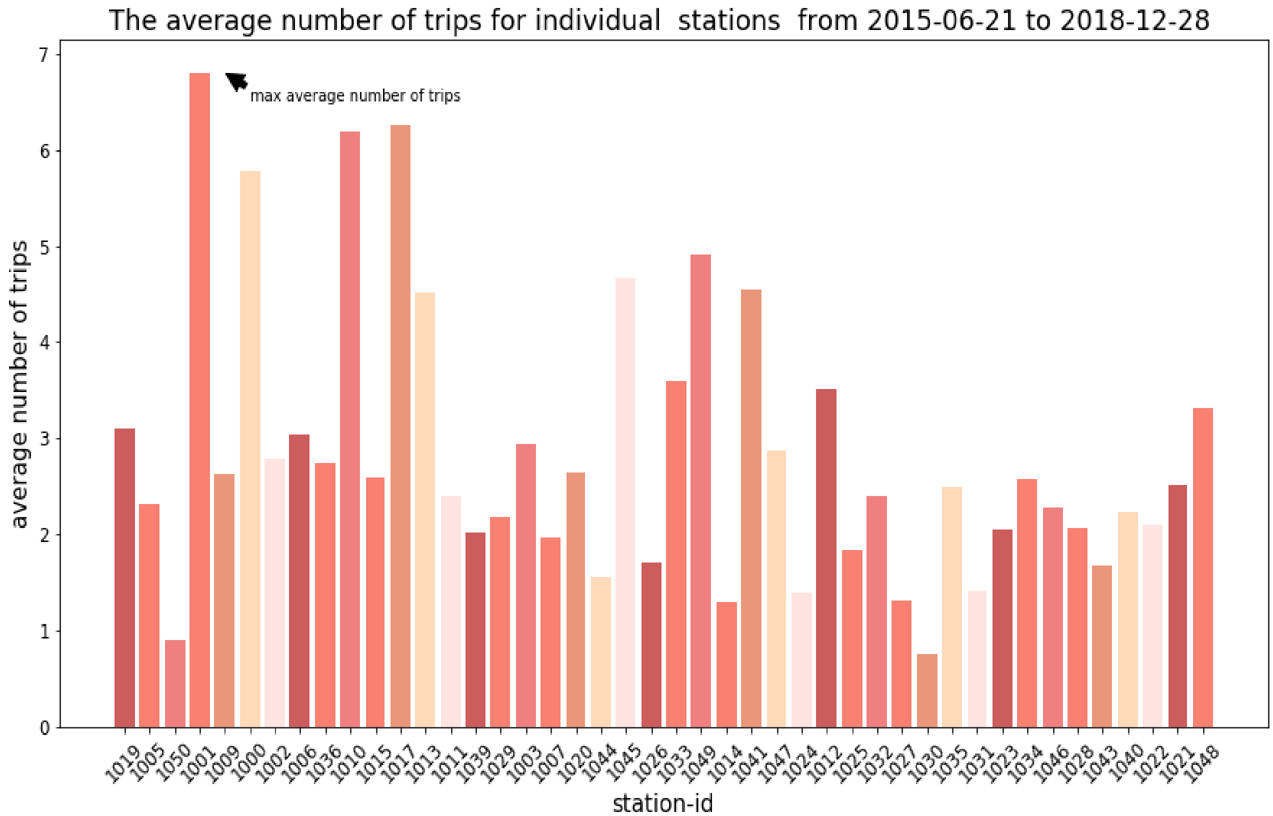
Calculating p-value using scipy.stats.norm.sf(abs(z)) yielded a value less than 0.05. Critical value (Z-crit) or the boundary of 95% confidence interval of one-tailed upper right test when alpha (rejection region) is 5% using stats.norm.isf (0.05) yielded a result of +1.6448. The z-score we got is 24.78. The pyplot below can be used to draw meaningful standardized meaningful conclusions.



The z-score calculated using frequentist approach falls in the rejection region, and since the p-value is less than the threshold significance level, which in our case is 5%, we can conclude that the results in difference of means was in fact statistically significant and we can reject the null hypothesis.

Average number of trips for individual stations:

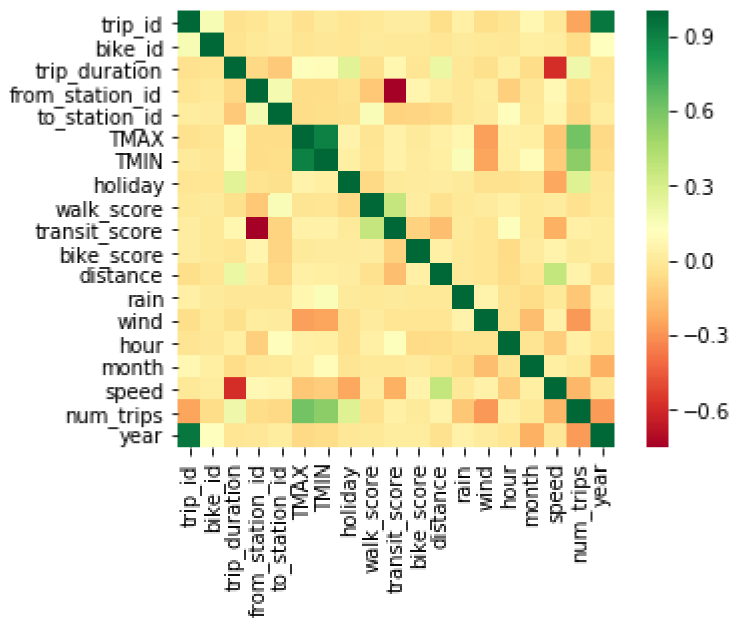
The plot below shows the average number of rentals per day including zero days at individual stations



Looks like the average number of rentals per day at individual station including zero days is maximum at station #1001 (Forbe's Ave & Market Square) with the mean average of 6.8 out of 1307 trips.

Correlation between the numerical features:

The heatmap shown below showing the correlation between the different numerical features of the dataset. Cells that are in green show positive correlation, while cells that are in red show negative correlation. Features are positively correlated with our target variable "num_trips" (number of trips per day) are temperature, holiday and the ones that are negatively correlated are rain, wind, speed and even year which makes sense intuitively as well.



Future Approach:

Build a linear regression model to predict the number of trips per day based on a variety of factors which helps in expansion of Pittsburgh Healthy ride bike share stations.
Build a model to predict the number of trips per hour at each individual stations.

Appendix:

Station ids and their respective names:

station_id	station_name
1019	42nd St & Butler St
1005	Forbes Ave & Grant St
1050	Healthy Ride Hub
1001	Forbes Ave & Market Square
1009	12th St & Penn Ave
1000	Liberty Ave & Stanwix St
1002	Third Ave & Wood St
1006	Ross St & Sixth Ave (Steel Plaza T Station)
1036	Schenley Dr at Schenley Plaza (Carnegie Librar...
1010	10th St & Penn Ave (David L. Lawrence Conventi...
1015	Federal St & E North Ave
1017	21st St & Penn Ave
1013	Isabella St & Federal St (PNC Park)
1011	Fort Duquesne Blvd & 7th St
1039	Atwood St & Bates St
1029	Alder St & S Highland Ave
1003	First Ave & Smithfield St (Art Institute)
1007	Stevenson St & Forbes Ave
1020	42nd & Penn Ave.
1044	Zulema St & Coltart Ave
1045	S 27th St & Sidney St. (Southside Works)
1026	Penn Ave & S Whitfield St
1033	Ivy St & Walnut St
1049	S 12th St & E Carson St
1014	Ridge Ave & Brighton Rd (CCAC)
1041	Fifth Ave & S Bouquet St
1047	S 22nd St & E Carson St
1024	S Negley Ave & Baum Blvd
1012	North Shore Trail & Fort Duquesne Bridge
1025	Penn Ave & N Fairmount St
1032	Walnut St & College St
1027	Shady Ave & Ellsworth Ave
1030	S Euclid Ave & Centre Ave

1035	Fifth Ave & S Dithridge St
1031	Maryland Ave & Ellsworth Ave
1023	Liberty Ave & Baum Blvd
1034	Ellsworth Ave & N Neville St
1046	S 25th St & E Carson St
1028	Penn Ave & Putnam St (Bakery Square)
1043	S Millvale Ave & Centre Ave
1040	O'Hara St and University Place (Soldiers and S...
1022	Liberty Ave & S Millvale Ave (West Penn Hospital)
1021	Taylor St & Liberty Ave
1048	S 18th St & Sidney St
1003	First Ave & Smithfield St