First, before analyzing the data, I imported several python modules that I needed to analyze and clean this data. I imported Pandas, Numpy, Pyplot from Matplotlib, Glob, and Date from Datetime.

To combine all of the bike data from all twelve months, I use the Glob module, append each file’s data into a list, then concatenate all of the files’ data in the list into a data frame called **df**.

For the data cleaning, I first check to see if any of the 247,584 rows have any missing data. I noticed that there were 380 rows of missing data from the User Type column, and 18,999 rows from the birth year column. I elect to drop those rows, leaving 228,205 rows for further data cleaning.

Next, I *describe* the data to see if I find other oddities. Two things I found were an odd minimum Birth Year of 1900 and a long maximum Trip Duration of 16,329,808 seconds, so I decide to investigate those further.

For the Birth Year anomaly, I printed a list of unique birth years, and I found that there were no birth years before 1930 except for this 1900 outlier. So, I elected to drop this 1900 outlier since it is very likely that this Birth Year was entered incorrectly by a user.

Next, I investigated the apparent Trip Duration outlier. I printed the median Trip Duration, which was only 370 seconds (6 minutes 10 seconds), so it appeared to make the outlier look more as such. I then looked at the trip durations over 1,000,000 seconds and found that there were seven other trip durations longer than 1,000,000 seconds. Upon investigating this further, I found that all of these users were ‘Subscribers’, meaning that they had a full-year bike rental pass. Given that 16,329,808 seconds is equivalent to just over 189 days, or just over half of one year, I decided to leave this data in.

Next, I thought about which columns could be split away into their own data frame, so we could create a database schema and make our **df** data frame easier to analyze. I saw that the train station information fit this description since I could separate the station data while leaving the Start Station ID and End Station ID in **df** to tie to the station data. I gather all of the unique stations’ information and separate that all into the data frames **start\_stations** and **end\_stations**, which I will combine into one **all\_stations** data frame.

Next, I perform several more data cleaning steps in the **df** and **all\_stations** data frames, such as:

* Renaming columns so we can use them more easily in the database I’ll create.
* Eliminating redundant station information since that is all contained in the **all\_stations** data frame now.
* Ensuring the data types all match their respective columns.
  + I changed birth year from a float data type to an integer data frame.
* Replaced the numbers in the Gender column to more descriptive “Male”, “Female”, or “Unknown”, according to the data documentation.
* Converted the datetime columns to just include the dates.
* Opted to keep just the date columns and eliminate the datetime columns.
  + I did not believe that having the time being more granular than days would be meaningful for our purposes here.

I have performed similar steps with the Newark weather dataset. I first imported the ‘newark\_airport\_data\_2016.csv’ CSV file into a data frame named **Newark**.

Next, I pull a summary of which columns contain missing data. I found that the columns PGTM and TSUN columns were missing data in all 366 rows, so I dropped those columns entirely. I pull the summary again, and this time I find that two of the columns each still have 2 rows of missing data, the WDF5 and WSF5 columns. These columns had missing data on the same days, 2/2/2016 and 11/25/2016. I will deal with these columns in an upcoming step.

Next, I decide to rename these columns since they are not descriptive nor helpful as they currently are. I rename them as follows to follow the data dictionary descriptions for these columns:

* STATION: station
* NAME: name
* DATE: date
* AWND: avg\_daily\_wind\_mph
* PRCP: precip\_in
* SNOW: snow\_accum\_in
* SNWD: snow\_depth\_in
* TAVG: avg\_temp\_f
* TMAX: max\_temp\_f
* TMIN: min\_temp\_f
* WDF2: avg\_2sec\_wind\_dir
* WDF5: avg\_5sec\_wind\_dir
* WSF2: max\_2sec\_wind\_mph
* WSF5: max\_5sec\_wind\_mph

After this column renaming, now I can better understand which columns are still necessary and which are not. I decide to drop the following columns since they are not particularly useful:

* station
* avg\_2sec\_wind\_dir
* avg\_5sec\_wind\_dir
* max\_2sec\_wind\_mph
* max\_5sec\_wind\_mph

Note that the missing values from a few steps ago were included in these columns, particularly the avg\_5sec\_wind\_dir and max\_5sec\_wind\_mph, so those missing values have been cleaned out of the data.

Next, I *describe* the data and print the data types for each column. I did not find anymore weird or outlying data, so I then write the **newark** dataset to a CSV file called ‘newark\_wx.csv’.

In each of the CSV files that I have (newark\_wx.csv, stations.csv, trip\_data.csv), before I upload then to the Postbird software for PostGreSQL, I perform the following data cleaning steps:

* Add an id column to each so that this column will serve as each database’s primary key.
* For any text columns, I add an apostrophe to the beginning and end of each text string so that the software can read those as text when inserted.
  + Example: string becomes ‘string’
* Copy and paste the query to be used into a Text file for easy copying and pasting of the query into Postbird. This comes in especially handy for the large trip\_data.csv dataset.

Once I have cleaned the CSV data well enough to insert into their respective tables, I begin to run queries on this data.

1. To find the days when the most trips started, limit of 10.
   1. These all are within the 9/13/16 to 10/12/16 timeframe, which surrounds the time the fall or autumn season begins.
2. To find the days when the least trips started, limit of 10.
   1. Interestingly, these all fall in January, February, or December of 2016. These first two queries may reveal more information if we include temperature and precipitation data.
3. To find the days, their average temperature (F), max temperature (F), precipitation (inches) , and average daily wind (mph) when the most trips started, limit of 10.
   1. This tells us that the weather was quite pleasant and conducive to bike riding, with temperatures averaging in the 60s to 70s range, and virtually no precipitation (rain).
      1. 9 of 10 days had no precipitation, while one had a nominal 0.01 inches.
   2. Also, every one of these days, with average winds ranging from 5-12 mph, had a light to moderate breeze, which would feel great on warm days like these.
4. To find the days, their average temperature (F), max temperature (F), precipitation (inches), average daily wind (mph), and snow accumulation (inches) when the least trips started, limit of 10.
   1. Since this query had some cold temperatures originally, I decided to add snow accumulation (inches) since snow was possible.
   2. Sure enough, the average temperature was cold, with the absolute least having a bone chilling 8°F average temperature and 18°F maximum temperature. Very unpleasant bike riding conditions indeed.
   3. Most days were in the 20s-40s range of average temperatures, with some days recording snow accumulations. Some days were even colder.
   4. Winds ranged from an average of about 5-20 mph, which makes for terrible wind chills, especially on the coldest days.
   5. The “warmest” day of these 10, at 43°F average and 52°F max, was Christmas Day, 12/25/2016. There was no precipitation or snow accumulation, so the fact that this was a major holiday influenced the ridership more than the weather conditions, which are pleasant for late December in Newark.
5. Gather count by user type.
   1. The vast majority of the users are annual subscribers (228,153 of them!), with only 51 with a 24-hour or 3-day pass.
6. Gather count by gender.
   1. More than 3 male bike riders for every 1 female.
7. Most popular starting stations.
   1. Looks like Grove Street PATH takes the cake, since it’s in densely populated Jersey City, located just a few blocks from the Hudson River, and provides a nice view of the New York City skyline from across the Hudson.
8. Least popular starting stations.
   1. This would be JCBS Depot, which flanks a large empty part of an industrial area and has a better public transportation alternative with city tram stops nearby.

As a result of performing this data analysis, I have come to following conclusions:

* Location, location, location! Stations with multiple redeeming qualities will attract more riders, justifying the stations’ presence. It’s not worth placing stations in unattractive locations or locations with low population density.
* Weather makes a huge difference on ridership. Bike riders are most attracted to pleasant weather conditions. There is a “Goldilocks zone” for temperatures; most bike riders don’t like it too hot or too cold, but instead they like it “just right”. Not to mention, most bike riders don’t like having to ride in the rain or snow, which can make bike riding more dangerous. Also, it feels good to have a gentle breeze blowing on warm days, but those breezes feel terrible on the coldest days.
* Holidays can also influence ridership, with Christmas Day serving as a good example. Not as many riders will venture out on major holidays since they may be busy with holiday-related activities or resting.
* This seems to be most attractive to male riders. There may be many factors inherent within a city that may be more encouraging to male riders, but less encouraging to female riders. To attract more female riders, safety measures may have to be implemented to make female riders feel safer when riding bikes on Newark streets.
* Since Newark and its surrounding areas are relatively flat, it can make for an easy and enjoyable bike ride for people, young and old. There are riders ranging from their teens all the way into their 80s.
* It seems more cost effective for the general public to purchase annual passes. There is much more opportunity than just a mere 24 hours or 3 days, to experience pleasant weather conditions or other attractive qualities of Newark. Most times of the year have pleasant weather conditions, and Newark and its surrounding area cover tens of square miles. Those shorter passes may be good for riders in a pinch, or for short term purposes, such as if they were visiting for a few days.

Well, I hope you have enjoyed this analysis of Newark, New Jersey’s city bike trip data from 2016. Take care!