Exploring Urban Interaction

Rafael Garcia Cano da Costa, Varun D. N.

Abstract: This work explored a weather data set from NOAA (metadata) and another with information on users of the Citi Bike service where both data sets had data from the New York City area in 2015. Among different plottings, correlation tables, and data joins, some interactions could be found showing the influence of the weather on the usage of Citi Bike over the year.

Introduction:

The motivation for this work is to get a better understanding about the impact of weather on the usage of the Citi Bike service. The objective is achieved by defining certain usage metrics and plotting them over the year and by seasons (spring, summer, fall and winter) in order to visualize the impact of the weather conditions on the usage by joining the two data sets. From the visualization of these joins, correlations between the weather condition and the bike usage can be derived. Once having a clearer idea of how both systems interact with each other, a myriad of implementations can take advantage of the findings. As a result, questions such as "when is the best period of time to carry out Citi Bike's scheduled maintenance?" and "who is a good target market for a new advertising campaign?" can have more effective answers. These outcomes come at a cost because the granularity and size of the data sets are huge. To address these specificities, this work applies techniques and technologies available in the open source world among Hadoop, Python, and D3.js which are famous for supporting an endeavor in Big Data analysis.

Experimental Techniques and Methods:

One of the techniques implemented in this work was MapReduce. This used the Hadoop HPC cluster infrastructure made available by New York University. The mappers and reducers were coded in Python. All the map-reduce tasks involved only 1 mapper and 1 reducer. As the goal of Big Data tools is to speedup processes, some small tasks like splitting a file were performed on the command line which consume more time if implemented on Hadoop. The weather data was recorded for every hour during every day of the year in GMT. Hence, the time stamp in the weather data was converted from GMT to EST. As a consequence, the data from the first five hours of January, 1st 2016 was inserted to the last five hours of December, 31st 2015. In the weather data, there were multiple entries for the same hour. The decision was to use only the last entry for the hour. Even though the Citi Bike data had a continuous time frame, a conversion was implemented in order to have both data sets with the same time granularity (hours). This approach set both data sets to the same granularity level having 8760 (365x24) data points for the year. After that, a sequence of plotting of distribution and joins were done for understanding the data and finding correlations. The log of all steps taken will follow the end of this report.

Results and Discussion:

The results were split by four tasks as a means to have a more understandable and modular discussion.

1) *Summarize the number of rides taken per day over the entire year and correlate it with the weather condition. Also, look for outliers in the usage rate*:

Initially, the attempt to make inference based on an unique plot having the number of rides for the whole year was not the best option. The solution was to plot (code) by the seasons of the year. This approach helped to identify that the number of rides is low starting in January, and it keeps that way during the winter. After the winter, the number of rides increases up to middle August, which is the middle of the fall. The mean of the number of rides per season starting in January was 142.196, 393.532, 537.327, and 455.167. An interesting fact is that the highest number of rides happened in the fall (3932) instead of in the summer (3774). This calls attention to the fact that a more pleasant weather can attract more users than just a very hot weather. Also, the minimum number of rides in the winter was zero while the other season had at least one.

Trying to establish a correlation between the number of rides and the weather, it is possible to see a positive correlation between temperature and number of rides. This is showed by the correlational analysis where the correlation between temperature and number of rides for the entire year is +.474702. Therefore, if the temperature increases, the number of rides increases, and the opposite is also true. However, within each season, this correlation values is smaller, which shows that the variance in temperature within each season impacts less the number of rides. The correlation between the number of rides and temperature per season starting in January was +.198974, +.247079, +.180363, and +.224204. Another positive correlation happens between the number of rides and the dew point temperature. For the whole year, the correlation is +.359986. However, this correlation is stronger in the spring (+.131701) while in the other season is below +.08. This explains why a windy day in the spring can be very freezing. An interesting conclusion is that the dew point temperature is very correlated with the regular temperature (at least +.741426), and as a result, the regular temperature on its own can convey good information on the number of rides oscillation. Finally, the correlation between the number of rides and a snowing day is -.106122 for the entire year even though this is only registered in the winter. When checking the specific correlation value for the winter, it is +.058131.

Another approach taken by this work is the direct comparison between the plotting for number of rides and weather condition by season. This is helpful for finding some hidden correlation. One of them is found in the winter more specifically in the beginning of February. In this period there is a big drop in the number of rides, but even though the correlation between the number of rides and the snow depth is much smaller than between number of rides and temperature, the steep drop for that period is related to the snow occurrence. This fact is clear when comparing the temperature and snow depth plotting against the number of rides. Also, the snow depth seems to have more impact when it reaches 3 inches, and the delay in the number of rides getting back to normal for the period seems to be directly related to the time taken to have the affected areas cleaned. For verifying the sequence of events, it is helpful to take a look at the precipitation plot. Once there is precipitation, it is probable to find presence of snow on the snow depth plot. As a result, the number of rides starts declining when the precipitation happens, it keeps the trend after the snow accumulation up to the cleaning of the affected areas. For the year of 2015, no one special event called attention, and any outlier was identified.

January 19th presents the number of rides lower than the normal trend for a Monday, and this is possibly due to the Martin Luther King's holiday and the contribution of some precipitation between 6 a.m. and 5p.m. January 27th shows a not very common low number of rides for Tuesday, and this is probably for being a day with precipitation and accumulation of snow above 3 inches. February 2nd was a Monday with intense precipitation, presence of more than 3 inches of snow and the Groundhog's Day, which may be enough to account for the low number of rides for that day of the week. March 5th was an atypical Thursday having presence of precipitation from midnight to 5p.m., which may have caused a big drop in the number of rides. March 11th was a Wednesday and had a peak in the number of rides probably due to the temperature reaching values over 60 F. Surprisingly, on March 28th, there were no rides recorded from 2 AM till 11 AM. It could be because of a scheduled maintenance by Citi Bike or due to some events like NYCRUNS Half Marathon in Brooklyn which saw upto 25,000 runners (article). May 25th had lower number of rides compared with the other Mondays in May, and it was possibly influenced by the Memorial day. Even though June 1st and 2nd were Monday and Tuesday, the number of rides were lower than expected influenced by occurrence of precipitation for a long period. June 27th was a Saturday, but it had lower number of rides than expected for a weekend, and this might have been caused by a precipitation starting at 1p.m. and making intermittent presence for fifteen hours. July 3rd, a Friday, had the number of rides typical for a weekend, and this was possibly related to the Independence day. September 7th, a Monday, had users behaving as in weekends, and this was possibly caused by the Labor day. October 1st and 2nd had a big drop in the number of rides for regular weekdays, and this was caused by a prolonged period of precipitation. October 28th, a Wednesday, showed a drop in the number of rides for a regular weekday, and this was possibly caused by a long term precipitation. November 10th had a low number of rides for a Tuesday, and this can be explained by the occurrence of precipitation. November 26th and 27th presented a lower number of rides for weekdays, but this can be explained for the Thanksgiving celebration. The presence of precipitation on December 1st, 2nd, and 3rd might have caused a different pattern in the number of rides for that month. On December 17th, a Thursday, might have its number of rides dropped by precipitation presence in the end of the day. From December 23rd to 31st, the change in the trend for the number of rides might be associated with the end of the year when people tend to be enjoying the holiday season in different ways.

2) Identify the best period over a year for doing maintenance on bikes based on lowest service demand:

Based on the initial analysis made on the interaction between the number of rides and the weather condition, the recommendation is to schedule the maintenance on the infrastructure for keeping the bikes running during the winter. Being more specific, February would be the best month for concentrating the most part of snowing days and for having the lowest number of rides over the year. Outside the winter period, the best hour for conducting small repairs would be between 2 a.m. and 4 a.m. when the number of rides tend to be below 300. During the winter this can be extended from 0 a.m. to 5 a.m.

3) Summarize the bike usage rates per day for males and females over the year and look for potential correlations with the weather condition:

Males:

Initially, the attempt to make inference based on an unique plot having the number of rides for the whole year was not the best option. The solution was to plot by the season of the year. This approach helped to identify that the number of rides is low starting in January, and it keeps that way during the winter. After the winter, the number of rides increases up to middle August, which is the middle of the fall. The mean of the number of rides per season by males starting in January was 303.322, 772.52, 1003.87,and 928.456. An interesting fact is that the highest number of rides happened in the fall (3932) instead of in the summer (3774). This calls attention to the fact that a more pleasant weather can attract more male users than just a very hot weather. Also, the minimum number of rides in the winter was zero while the other season had at least five.

Trying to establish a correlation between the number of rides and the weather, it is possible to see a positive correlation between temperature and number of rides. This is showed by the correlational analysis where the correlation between temperature and number of rides on males for the entire year is +.416520. Therefore, if the temperature increases, the number of rides increases, and the opposite is also true. However, within each season, the correlation value is smaller, which shows that the variance in temperature within each season impacts less the number of rides. The correlation between the number of rides and temperature on males per season starting in January was +.341877, +.333888, +.238523, and +.309230. Another positive correlation happens between the number of rides and the dew point temperature. For the whole year, the correlation is +.315347. However, this correlation is stronger in the spring (+.173323) while in the other season is below +.12. This explains why a windy day in the spring can be very freezing. An interesting conclusion is that the dew point temperature is very correlated with the regular temperature (at least +.751941), and as a result, the regular temperature on its own can convey good information on the number of rides oscillation. Finally, the correlation between the number of rides and a snowing day is -.098486 for the entire year even though this is only registered in the winter. When checking the specific correlation value for the winter, it is -.095591. For the year of 2015, no one special event called attention, and any outlier was identified.

Females:

Initially, the attempt to make inference based on an unique plot having the number of rides for the whole year was not the best option. The solution was to plot by the season of the year. This approach helped to identify that the number of rides is low starting in January, and it keeps that way during the winter. After the winter, the number of rides increases up to middle August, which is the middle of the fall. The mean of the number of rides per season by females starting in January was 68.6106, 230.403, 277.646,and 271.594. An interesting fact is that the highest number of rides happened in the summer (1363) instead of in the fall (1248). This calls attention to the fact that a hotter weather can attract more female users. Also, the minimum number of rides in all season was zero.

Trying to establish a correlation between the number of rides and the weather, it is possible to see a positive correlation between temperature and number of rides. This is showed by the correlational analysis where the correlation between temperature and number of rides on males for the entire year is +.463361. Therefore, if the temperature increases, the number of rides increases, and the opposite is also true. However, within each season, the correlation value is smaller, which shows that the variance in temperature within each season impacts less the number of rides. The correlation between the number of rides and temperature on females per season starting in January was +.406709, +.388047, +.260977, and +.345270. Another positive correlation happens between the number of rides and the dew point temperature. For the whole year, the correlation is +.352939. However, this correlation is stronger in the spring (+.210083) while in the other season is below +.119. This explains why a windy day in the spring can be very freezing. An interesting conclusion is that the dew point temperature is very correlated with the regular temperature (at least +.751941), and as a result, the regular temperature on its own can convey good information on the number of rides oscillation. Finally, the correlation between the number of rides and a snowing day is -.105179 for the entire year even though this is only registered in the winter. When checking the specific correlation value for the winter, it is -.121248.

As aforementioned, it is possible to conclude that the number of rides is higher among males, and women as a whole could be targeted by advertisement as a means to increase revenue from that segment. Also, it is possible to derive that women are more influentiable by the temperature for the first three seasons of the year when thinking about using the Citi Bike system. Besides that, a snowing day has more impact on the frequency of rides taken by women. Another curious aspect noticed was the difference between the two peaks on the number of rides for a given day for males and females in November, December, January and February. The second peak of a given weekday is lower for females while males tend to have both peaks very similar. This can be related to people commuting back home when the night took place, and women can be more cautious about riding a bicycle during the night when days get darker earlier. Even though some of them used the citi bike service earlier in the day, they would prefer another option at night. For the year of 2015, no one special event called attention, and no outlier was identified.

4) Understand the bike usage rates per day for people from different age groups over the year. The goal is to use this information for better marketing. Also, look for correlations in the bike usage rate of the different age groups with the weather conditions:

Age group below 21:

Initially, the attempt to make an inference based on an unique plot having the number of rides for the whole year was not the best option. The solution was to plot by the season of the year. This approach helped to identify that the number of rides is low starting in January, and it keeps that way during the winter. After the winter, the number of rides increases up to middle August, which is the middle of the fall. The mean of the number of rides per season for users below twenty one years old starting in January was 6.34352, 15.9684, 25.9973,and 32.0634. An interesting fact is that the highest number of rides happened in the fall (124) instead of in the summer (128). This calls attention to the fact that a hotter weather may be more attractive to users below 21 years old than just a pleasant weather in the fall. Also, the minimum number of rides was zero for all season.

Trying to establish a correlation between the number of rides and the weather, it is possible to see a positive correlation between temperature and number of rides. This is showed by the correlational analysis where the correlation between temperature and number of rides on people below 21 for the entire year is +0.349305. Therefore, if the temperature increases, the number of rides increases, and the opposite is also true. However, within each season, the correlation value was higher in the winter, which shows that the variance in temperature during the winter impacts the number of rides at a higher level. The correlation between the number of rides and temperature on people below 21 per season starting in January was +0.391662, +0.282416, +0.111655, and +0.335076. Another positive correlation happens between the number of rides and the dew point temperature. For the whole year, the correlation is +0.267934. However, this correlation is stronger in the spring (+0.125466) while in the other season is below +.12. This explains why a windy day in the spring can be very freezing. An interesting conclusion is that the dew point temperature is very correlated with the regular temperature (at least +0.751941), and as a result, the regular temperature on its own can convey good information on the number of rides oscillation. Finally, the correlation between the number of rides and a snowing day is --0.098772 for the entire year even though this is only registered in the winter. When checking the specific correlation value for the winter, it is --0.080391. For the year of 2015, no one special event called attention, and any outlier was identified.

Age group above 20 and below 36:

Initially, the attempt to make inference based on an unique plot having the number of rides for the whole year was not the best option. The solution was to plot by the season of the year. This approach helped to identify that the number of rides is low starting in January, and it keeps that way during the winter. After the winter, the number of rides increases up to middle August, which is the middle of the fall. The mean of the number of rides per season for users between twenty and thirty six years old starting in January was 163.278, 486.516, 687.168,and 596.65. An interesting fact is that the highest number of rides happened in the summer (2717) instead of in the fall (2713). This calls attention to the fact that a hotter weather may be more attractive to that group of users than just a pleasant weather in the fall. Also, the minimum number of rides was zero in the winter while at least one for the other season.

Trying to establish a correlation between the number of rides and the weather, it is possible to see a positive correlation between temperature and number of rides. This is showed by the correlational analysis where the correlation between temperature and number of rides on people between twenty and thirty six years old for the entire year is +0.427267. Therefore, if the temperature increases, the number of rides increases, and the opposite is also true. However, within each season, the correlation value is smaller, which shows that the variance in temperature within each season impacts less the number of rides. The correlation between the number of rides and temperature on that age group per season starting in January was +0.348979, +0.321453, +0.212139, and +0.294934. Another positive correlation happens between the number of rides and the dew point temperature. For the whole year, the correlation is +0.338004. However, this correlation is stronger in the spring (+0.183486) while in the other season is below +.11. This explains why a windy day in the spring can be very freezing. An interesting conclusion is that the dew point temperature is very correlated with the regular temperature (at least +0.751941), and as a result, the regular temperature on its own can convey good information on the number of rides oscillation. Finally, the correlation between the number of rides and a snowing day is -0.100142 for the entire year even though this is only registered in the winter. When checking the specific correlation value for the winter, it is -0.093876. Snowing days seemed to drop the number of rides to zero many times, which was not common among other group ages. One example was January 28th. On March 11th, there was a peak when the number of rides reached 45. This may be related to the fact that that day was warmer being the only one in March having temperature above 60 F. For the year of 2015, no one special event called attention, and any outlier was identified.

Age group above 35 and below 51:

Initially, the attempt to make inference based on an unique plot having the number of rides for the whole year was not the best option. The solution was to plot by the season of the year. This approach helped to identify that the number of rides is low starting in January, and it keeps that way during the winter. After the winter, the number of rides increases up to middle August, which is the middle of the fall. The mean of the number of rides per season for users between thirty five and fifty years old starting in January was 134.399, 338.415, 423.728,and 398.025. An interesting fact is that the highest number of rides happened in the fall (1925) instead of in the summer (1807). This calls attention to the fact that a more pleasant weather may be more attractive to that group of users than just a hotter weather in the summer. Also, the minimum number of rides was zero in the winter while at least one for the other season.

Trying to establish a correlation between the number of rides and the weather, it is possible to see a positive correlation between temperature and number of rides. This is showed by the correlational analysis where the correlation between temperature and number of rides on people between thirty five and fifty years old for the entire year is +0.408899. Therefore, if the temperature increases, the number of rides increases, and the opposite is also true. However, within each season, the correlation value is smaller, which shows that the variance in temperature within each season impacts less the number of rides. The correlation between the number of rides and temperature on that age group per season starting in January was +0.332164, +0.348351, +0.259663, and +0.318974. Another positive correlation happens between the number of rides and the dew point temperature. For the whole year, the correlation is +0.300357. However, this correlation is stronger in the spring (+0.176294) while in the other season is below +.11. This explains why a windy day in the spring can be very freezing. An interesting conclusion is that the dew point temperature is very correlated with the regular temperature (at least +0.751941), and as a result, the regular temperature on its own can convey good information on the number of rides oscillation. Finally, the correlation between the number of rides and a snowing day is -0.094812 for the entire year even though this is only registered in the winter. When checking the specific correlation value for the winter, it is -0.099486. For the year of 2015, no one special event called attention, and any outlier was identified.

Age group above 50:

Initially, the attempt to make inference based on an unique plot having the number of rides for the whole year was not the best option. The solution was to plot by the season of the year. This approach helped to identify that the number of rides is low starting in January, and it keeps that way during the winter. After the winter, the number of rides increases up to middle August, which is the middle of the fall. The mean of the number of rides per season for users between thirty five and fifty years old starting in January was 67.9977, 162.504, 195.959,and 190.951. An interesting fact is that the highest number of rides happened in the fall (827) instead of in the summer (782). This calls attention to the fact that a more pleasant weather may be more attractive to that group of users than just a hotter weather in the summer. Also, the minimum number of rides was zero for all season.

Trying to establish a correlation between the number of rides and the weather, it is possible to see a positive correlation between temperature and number of rides. This is showed by the correlational analysis where the correlation between temperature and number of rides on people between thirty five and fifty years old for the entire year is +0.430839. Therefore, if the temperature increases, the number of rides increases, and the opposite is also true. However, within each season, the correlation value is smaller, which shows that the variance in temperature within each season impacts less the number of rides. The correlation between the number of rides and temperature on that age group per season starting in January was +0.374118, +0.381625, +0.304801, and +0.346879. Another positive correlation happens between the number of rides and the dew point temperature. For the whole year, the correlation is +0.302270. However, this correlation is stronger in the spring (+0.175185) while in the other season is below +.12. This explains why a windy day in the spring can be very freezing. An interesting conclusion is that the dew point temperature is very correlated with the regular temperature (at least +0.751941), and as a result, the regular temperature on its own can convey good information on the number of rides oscillation. Finally, the correlation between the number of rides and a snowing day is -0.099903 for the entire year even though this is only registered in the winter. When checking the specific correlation value for the winter, it is -0.110086. For the year of 2015, no one special event called attention, and any outlier was identified.

After observing the parameters for each age group, it is clear that most rides are taken by people between twenty and thirty six years old. Also, the people below 21 and above 50 could be a good target market for an advertising campaign as a means to increase revenue once they comprise the smallest share of users. Another difference that called attention is the pattern for the number of rides between the two common peaks during weekdays among people below the age of 21. This group presents a proper trend, showing that their usage pattern might not be influenced by any rigorous driver like older people going to work in the morning and going back home in the end of the day. Their number of rides between the two main peaks for weekdays seemed more variable instead of having a basic huge decline as for the other age groups.

Individual Contributions:

The members of this team partnered and worked jointly for the accomplishment of all tasks.

Summary and Conclusions:

The Citi Bike and Weather data sets on the New York City area for 2015 conveyed valuable information referent to statistical and correlational data. Both information being helpful in finding ways to provide better support to the Citi Bike system’s operation both on the logistic and financial sides. Moreover, the possibility to visualize the interaction between the data sets offered a better understanding of sequential events involving temperature change, precipitation, snow depth, and number of rides taken by users throughout the year.

References:

Dalessandro, Brian. "DS-GA 1001 Intro to Data Science." DS-GA 1001 Intro to Data Science. New York University, New York, NY. 2015. Lecture.

Freire, Juliana. "DS-GA 1004 BigData." DS-GA 1004 BigData. New York University, New York, NY. 2016. Lecture.

Watson, Gregory. "DS-GA 3001 Advanced Python." DS-GA 3001 Advanced Python. New York University, New York, NY. 2016. Lecture.

1) FINAL LOG:

This project comprises some steps in order to be accomplished. Each step is described below in a ordered way.

Reproducible Steps:

1- Downloaded 2015 citi bike data from
https://s3.amazonaws.com/tripdata/index.html

2- Downloaded 2015 NYC weather data from
https://nyu.app.box.com/s/41krxs9rdsfjzpu1gh9nwen89jxtc9dd

3- Downloaded 2015 NYC weather metadata from
https://nyu.box.com/s/461edjve0obbsefcpaygb1h9z8yc49c4

4- Compiled metadata for citi bike from
https://www.citibikenyc.com/system-data

5- Appended all 2015 citi bike data into one single file keeping only one header on the top using "cat file.csv >> result.csv". The size is 1.86GB.

6- Created google-drive folder to store data with shareable link
https://drive.google.com/open?id=0BzMI4SEa4vLeZTBWclEwdWo4SnM

7- Created file to register all steps for having a reproducible project with link
https://drive.google.com/open?id=1hFTvG2zYPxeUpFa8Z-L-1R3TkeTtlNtGytsgmtfEE64

8- Created github repo to store code on:
https://github.com/rgc292/dsga1004finalproject

9- Wrote a map-reduce code to extract the number of rides taken during an hour of window for every day in the year 2015.

> **Example output: 1/1/2015,3 100** - implies that on 1/1/2015 during the 3rd hour of the day (2 AM - 3 AM) there were 100 rides taken or rather the start time of these 100 rides were between 2 AM and 3 AM.

10- On the weather data coded the first line where there was "USAF" became "USUSAF". Wrote a map-reduce code available on github in folder map1_reduce1_on_weatherdata. The output is the weather data as a key-value pair having the latest data for each hour each day in 2015.

Example output:KEY (DATE, HOUR) and VALUES (DIR,SPD,GUS,SKC,TEMP,DEWP,SLP,STP,MAX,MIN,PCP01,PCP06,PCP24,SD)

11- Wrote a map-reduce code available on github in folder map2_reduce2_on_output_map1_reduce. The output has the same data, but the time was converted from GMT into EST.

Issues: The conversion moved the first five hours on January 2015 to 2014. As a result, it is necessary to get the respective data from January 2016 which is going to be introduced into December 2015. It was not possible to get the data from 2016 online.

12- The dates that were used as the keys in the map-reduce phase were strings and we want them as a timestamp object. We could have done this modification in the map-reduce code itself. But, the first version of the code is really long to edit.

Hence, wrote a small code to convert the strings to dates and sort them according to the timestamp. This sorting will help in further analysis.

13- Split the sorting by season having for spring (April to June), summer (July to September), fall (October to December), and winter (January to March). The code and splitted file are in weather_by_seasons.

14- Add the map3_reduce3_join_on_final_num_ride as the output file from the join between final_num_rides_data.txt and final_weather_data.txt.

15- Added the 5 first hours from January first 2016 in the end final_weather_data.txt of by hand.

16- Run the program to segregate the data into weekdays and weekends. Follow the instructions in the file to feed the inputs.

17 - Run the file plot_num_rides.py to get the plot of number of rides taken during all 4 seasons. Follow the instructions in the file to feed the inputs.

18 - Run the file plot_total_rides_every_hour.py to get the plot of the total number of rides taken every hour during all 4 seasons.

19 - Join final_gender_total_rides.txt and final_weather_data.txt using mapreduce_gender_weather_join.

20 - Split gender_weather_joined by season in genders_joined_by_season.

21 - The female data had some “missing” days, but this was filled out with the respective dates having zero rides applying mapreduce_female_weather_join.

22 - Separate females and males, and split by season as in males_joined_by_season and females_joined_by_season.

23- Plotted correlational analysis as in Correlational Analysis

24- Wrote initial final report on the three first tasks.

25- Wrote map6.py and reduce6.py to obtain joined_duration_weather. Also, split by season as in Split by season. Everything done following join_duration_weather_steps.txt.

26- Add visual correlational analysis on total rides and weather at total_rides_weather_correlation.

27- Add visual correlational analysis on male rides and weather at males_rides_weather_correlation.

28- Add visual correlational analysis on female rides and weather at females_rides_weather_correlation.

29- Add visual correlational analysis on female rides and weather at total_rides_length_weather_correlation.

30- Wrote “Abstract”, “Introduction”, “Experimental Techniques and Methods”, “Results and Discussion” on the report.

31 - Wrote map7_reduce7_on_citibike for extracting rides by four age groups (<=20, >20 and <=35, >35 and <=50, >50). Among 9937969 trips, 1311378 did not have date of birth reported. Split files by age group using bash_ages.txt. Remove the age group field from files creating a new file like rides_age_20_clean.txt. Using format_time_for_sorting.py, format the date of each age group file having as outcome files like final_rides_age_20_clean.txt for joining with weather data.

32 - Join final_weather_data.txt and respective age group file like final_rides_age_20_clean.txt using map6.py and reduce6.py. The output are in join_age_weather.

32 - Split by season in joined_rides_age_20.output using ride_age_weather_by_season.txt and do the same for the other age groups.

33 - Plot correlational analysis on age groups and weather by season in age_above_20_below_35_weather_correlation, age_above_35_below_50_weather_correlation, age_above_50_weather_correlation, and age_below_20_weather_correlation.

34 - Add correlational analysis tables in rides_ages_weather.

35 - Add last parts to the report.

36 - All the plots needed some basic formatting on the specific data set used. The removal of the spacing between the date and hour key and the respective value keeping only a comma between them was one approach. Another was the sorting of the dataset on the command line. Finally, the introduction of headers for each column separated by comma in the first line.

37 - Submission of the report.