Midterm

Coding in R Language

Vian Tran

Summer 2024

**Data Analysis Report**

The data science team wants to create a predicative model that will predict the number of bikes leaving each station and the number of bikes being returned to each station in the next three days. The following report will provide insight that will help the maintenance teams plan their bike maintenance operations.

**Summary of Data**

The report consists of three datasets containing data that represents stations where users can pick up or return bikes (station), individual bike trips (trip) and data about the weather on a specific day for certain cities (weather). There are 70 unique stations spanning across Mountain View, Palo Alto, Redwood City, San Francisco and San Jose, with majority of the stations residing in San Francisco and San Jose, as shown in Figure 1a. The dock count for each station ranged from 11 to 27 with most stations having around 15 bike docks, as shown in Figure 1b.

If we quickly look at the distribution of the users’ start and end stations, the preference of stations remain relatively the same for starting and ending stations. For example, the station at San Francisco Caltrain (Townsend at 4th) remains the top starting and ending station. The start and end stations are ordered in alphabetical order, there are 4 checkpoints that lines up the data including 2nd at Folsom, Franklin at Maple, Ryland Park and Speak at Folsom as shown in Figure 2ab. Furthermore, majority of the users of Bay Area Bike Rental are subscribers indicating they use the services often. The bike rides for each customer ranged from 60 seconds (1 minute) to 17270400 seconds (200 days), with a median of 511 seconds (8.5 minutes). If we loge the duration and display it in a graph as shown in Figure 2d, the graph ranges from loge(duration) of 4 to 16 with much of the bike ride around 6. Most bike rides last for a few minutes but there are outliers that can last for days. This is an area we need to explore when preparing the data for analysis.

Throughout the year of 2014, the mean temperature ranged from 44˚F to 84˚F. The range of fluctuation between the minimum and maximum temperature per city can be seen in Figure 3, 4, 5, 6, and 7 a-c for San Francisco, Redwood City, Mountain View, Palo Alto and San Jose, respectively. The mean visibility ranged from 4 to 20 miles. The range of fluctuation between the minimum and maximum visibility per city can be seen in Figure 3, 4, 5, 6, and 7 d-f. There are a few NAs in max, mean and min visibility variable that should be noted. The mean wind speed ranged from 0 to 19 mph. The range of fluctuation between maximum win speed per city can be seen in Figure 3, 4, 5, 6, and 7 g-h. The max wind speed variable contains wind speed above 50 mph which indicates a strong gale to hurricane force level (over 75 mph) winds. During this time there were no reports of high dangerous winds in those cities in 2014. Therefore, for future analysis, max wind speed above 50 mph were set as NA. The max gust speed also is displayed in Figure 3, 4, 5, 6, and 7 i with a range between 6 and 114 mph. There is quite a bit of NAs in max gust speed, up to 450s Nas and should be noted. The precipitation mostly remained around 0 inches throughout 2014 as in it did not rain a lot, the max it got was 3.36 inches in precipitation as shown in Figure 3, 4, 5, 6, and 7 j. There were a few Nas in precipitation variable. Furthermore, the cloud cover ranged between 0 to 8, with 0 being no cloud cover to 8 being high cloud cover as shown in Figure 3, 4, 5, 6 and 7 k.

**Pre-Processing**

Preparing the datasets for future analysis requires the removal of any duplicate entries in each dataset. For comparability, any date column was change to POSIX using the “lubridate” package. Furthermore, the ZIP codes in the trip dataset were changed to only include USA ZIP codes. This includes codes that start with 0-9 digits and is exactly 5 digits longs, otherwise it was set to NA. Additionally, the T’s entries in the precipitate (inches) variables, the purpose of the T’s was not explained and should be changed to NA.

It is necessary to remove any cancelled trips from the dataset before analyzing it. A trip less than 3 minutes long (180 seconds) was considered cancelled. We can also ensure the trip did not leave the station if the starting and ending station IDs are the same. There were a total of 1082 cancelled trip and the trip IDs can be summarized in the trip\_cancelled CSV file. The cancelled trips were removed from the dataset. Additionally, outliers that are skewing the data should be removed. We are interested in seeing the return of bikes to each station within the next 3 days. Therefore, outliers should be considered as those above 3 days long and should be removed from the dataset. Considering the median duration is around 8.5 minutes, most bike users do not use the bikes for a long time. Therefore, setting the outliers as above 3 days is reasonable and generous. There were 27 outliers that can be summarized in the trip\_outlier CSV file. The outliers were removed from the dataset and the loge(duration) frequencies plot can be seen in Figure 2e.

**Findings**

To create a strong predictive model to predict the number of bikes leaving each station and the number of bikes being returned we need to gather some information from the dataset such as rush hours times, frequent starting and ending stations used during rush hours and weekends, average utilization of bikes for each month and the impact of weather conditions on bike rental patterns that can influence the predictive model.

*Rush Hour and Frequencies*

Isolating for the weekdays and extracting the hours of each day, we can see the frequencies of the highest volume hours on weekdays also known as rush hour for bike rental. The rush hour during the week is from 7 to 9 am and 4 to 6 pm (16 to 18 hours) as shown in Figure 8). Furthermore, we can determine the top 10 most frequently used start and end stations during the rush hour. The top 10 most frequently used start and end station can be summarized in Table 1 and 2. The start and the end stations remained relatively the same but in different order of top frequencies. The only different station in the start and end is 2nd at South Park and Market at 10th. We can see this general trend in Figure 2 a-b as the popular stations remain relatively the same for the start and end stations. The following can provide helpful insight that can help the maintenance team plan their bike maintenance during times of high customer volume.

*Weekend Frequencies*

We can apply the same concept as above to find the top 10 most frequently used start and end stations during the weekend. The top 10 most frequently used start and end station can be summarized in Table 3 and 4. The start and the end stations remained relatively the same but in different order of top frequencies. The only different station in the start and end is Steuart at Market and Powell at Post (Union Square). We can see this general trend in Figure 2 a-b as the stations remain relatively the same for the start and end stations.

*Average Use per Month*

Finding the monthly use of the bikes can provide helpful insight to update the predictive model on the popular and active months. We can find the average utilization per month by dividing total time used by total time available. We can find the total time used for each month by adding up all the bike ride durations for each month. The total time available is the total time in each month, and ensuring the total time available is in the same units as total time used. For example, there are 744 hours in October. The average use per month can be summarized in Table 5. It is common for people to have a higher interest in bike rides in the summer, fall and spring months. There is lower use in bikes in the winter months where it is colder and may be snowing.

*Weather Conditions*

The data science team assumes that the weather conditions have an impact on the bike rental patterns. Currently, they are unsure whether they should use temperature, weather events, visibility or other weather measurements metrics available. Therefore, a correlation matrix between the trip and weather dataset was created to test the relationship between the number of trips each day in each city correlates to the different weather metric. The different weather metric tested were max/mean/min temperature, max/mean/min visibility, max/mean wind speed, max wind speed and precipitation.

The highest correlation tended to correspond with temperature (weather it was max, mean or min temperature) and visibility (weather it was max, mean or min visibility)

1. The data science team assumes that weather conditions probably have an impact on the bike rental patterns, but they are not sure whether they should use temperature, weather events, visibility or other weather measurements available. Help them decide by creating a new dataset combining trip data with the  
   weather data. (Note that the weather data is available for each city and date. Join your datasets accordingly). Create a correlation matrix for the new dataset using the cor() function from the corrplot package. Flag the highest correlations for the data science team.

You team lead expects a Data Analysis Report with all your findings for the next meeting in two weeks.

**Appendix**

a.

b.

A graph of a number of blue squares

Description automatically generatedA graph of different cities

Description automatically generated

**Figure 1: Categorical variable summary of the station dataset including the frequencies of dock counts and cities.**

A graph of a number of stations

Description automatically generatedA graph of different colored columns

Description automatically generated with medium confidence

a.

b.

c.

d.

e.

A blue rectangles with white text

Description automatically generatedA graph of a logger

Description automatically generatedA graph of a log

Description automatically generated

**Figure 2: Categorical and numerical summary of the trip dataset including start stations, end stations, subscription type, and loge(duration).** The start and end stations are ordered in alphabetical order, there are 4 checkpoints that lines up the data including 2nd at Folsom, Franklin at Maple, Ryland Park and Speak at Folsom.

A graph with blue dots

Description automatically generated A graph with blue dots

Description automatically generated A graph with blue dots

Description automatically generated A graph with numbers and lines

Description automatically generated A graph with numbers and dots

Description automatically generated A graph with blue dots

Description automatically generated A graph with blue dots

Description automatically generated A graph with blue dots

Description automatically generated A graph with blue dots

Description automatically generated A graph of precipitation with blue dots

Description automatically generated A diagram of data with blue circles

Description automatically generated with medium confidence

b.

i.

h.

k.

j.

g.

d.

e.

f.

c.

a.

**Figure 3: Weather metrics of San Francisco including temperature, visibility wind speed, gust speed, precipitation and cloud cover.**

**A graph with blue dots

Description automatically generated** **A graph with blue dots

Description automatically generated** A graph with blue dots

Description automatically generated A graph with numbers and lines

Description automatically generated A graph with numbers and dots

Description automatically generated A graph with blue dots

Description automatically generated A graph with blue dots

Description automatically generated A graph with blue dots

Description automatically generated A graph of blue dots

Description automatically generated A graph of precipitation and date

Description automatically generated A diagram of a cloud cover

Description automatically generated

a.

b.

c.

d.

e.

f.

g.

h.

i.

j.

k.

**Figure 4: Weather metrics of Redwood City including temperature, visibility wind speed, gust speed, precipitation and cloud cover.**

**A graph with blue dots

Description automatically generated** **A graph with blue dots

Description automatically generated** A graph with blue circles

Description automatically generated A graph with numbers and lines

Description automatically generated A graph with numbers and dots

Description automatically generated A graph with blue dots and white text

Description automatically generated A graph with blue dots

Description automatically generated A graph with blue dots

Description automatically generated A graph with blue dots

Description automatically generated A graph with blue dots

Description automatically generated A graph with blue dots

Description automatically generated

a.

b.

c.

d.

e.

f.

g.

h.

i.

j.

k.

**Figure 5: Weather metrics of Mountain View including temperature, visibility wind speed, gust speed, precipitation and cloud cover.**

**A graph with blue dots

Description automatically generated** **A graph with blue dots

Description automatically generated** A graph with blue dots

Description automatically generated A graph with numbers and a number of months

Description automatically generated with medium confidence A graph with blue dots and white text

Description automatically generated A graph with blue dots

Description automatically generated A graph with blue dots

Description automatically generated A graph with blue dots

Description automatically generated A graph with blue dots

Description automatically generated A graph with numbers and lines

Description automatically generated A diagram of a cloud cover

Description automatically generated

a.

b.

c.

d.

e.

f.

g.

h.

i.

j.

k.

**Figure 6: Weather metrics of Palo Alto including temperature, visibility wind speed, gust speed, precipitation and cloud cover.**

**A graph with blue dots

Description automatically generated** **A graph with numbers and dots

Description automatically generated** A graph with blue dots

Description automatically generated A graph with numbers and lines

Description automatically generated A graph with numbers and dots

Description automatically generated A graph with blue dots and white text

Description automatically generated A graph with numbers and dots

Description automatically generated A graph with numbers and blue dots

Description automatically generated A graph with blue dots

Description automatically generated A graph of precipitation with numbers and a line of blue dots

Description automatically generated with medium confidence A graph with blue dots

Description automatically generated

a.

b.

c.

d.

e.

f.

g.

h.

i.

j.

k.

**Figure 7: Weather metrics of San Jose including temperature, visibility wind speed, gust speed, precipitation and cloud cover.**

**A graph of blue bars

Description automatically generated**

**Figure 8: Rush hours during the weekday.**

**Table 1: Rush Hour Top 10 Frequently Used Start Stations**

|  |  |  |
| --- | --- | --- |
| Start Station ID | Name | Trip Count |
| 70 | San Francisco Caltrain (Townsend at 4th) | 17908 |
| 69 | San Francisco Caltrain 2 (330 Townsend) | 10198 |
| 55 | Temporary Transbay Terminal (Howard at Beale) | 9519 |
| 50 | Harry Bridges Plaza (Ferry Building) | 7887 |
| 61 | 2nd at Townsend | 7288 |
| 74 | Steuart at Market | 7280 |
| 77 | Market at Sansome | 6527 |
| 65 | Townsend at 7th | 6398 |
| 67 | Market at 10th | 5519 |
| 60 | Embarcadero at Sansome | 5265 |

**Table 2: Rush Hour Top 10 Frequently Used End Stations**

|  |  |  |
| --- | --- | --- |
| End Station ID | Name | Trip Count |
| 70 | San Francisco Caltrain (Townsend at 4th) | 23611 |
| 69 | San Francisco Caltrain 2 (330 Townsend) | 10483 |
| 77 | Market at Sansome | 8376 |
| 61 | 2nd at Townsend | 8195 |
| 55 | Temporary Transbay Terminal (Howard at Beale) | 7753 |
| 50 | Harry Bridges Plaza (Ferry Building) | 7668 |
| 65 | Townsend at 7th | 7270 |
| 74 | Steuart at Market | 7142 |
| 60 | Embarcadero at Sansome | 5223 |
| 64 | 2nd at South Park | 4775 |

**Table 3: Weekend Top 10 Frequently Used Start Stations**

|  |  |  |
| --- | --- | --- |
| Start Station ID | Name | Trip Count |
| 50 | Harry Bridges Plaza (Ferry Building) | 3164 |
| 60 | Embarcadero at Sansome | 3116 |
| 76 | Market at 4th | 1661 |
| 54 | Embarcadero at Bryant | 1603 |
| 61 | 2nd at Townsend | 1546 |
| 39 | Powell Street BART | 1486 |
| 70 | San Francisco Caltrain (Townsend at 4th) | 1361 |
| 73 | Grant Avenue at Columbus Avenue | 1298 |
| 77 | Market at Sansome | 1095 |
| 71 | Powell at Post (Union Square) | 1090 |

**Table 4: Weekend Top 10 Frequently Used End Stations**

|  |  |  |
| --- | --- | --- |
| End Station ID | Name | Trip Count |
| 60 | Embarcadero at Sansome | 3368 |
| 50 | Harry Bridges Plaza (Ferry Building) | 3174 |
| 76 | Market at 4th | 1877 |
| 39 | Powell Street BART | 1676 |
| 70 | San Francisco Caltrain (Townsend at 4th) | 1660 |
| 61 | 2nd at Townsend | 1591 |
| 54 | Embarcadero at Bryant | 1384 |
| 74 | Steuart at Market | 1223 |
| 77 | Market at Sansome | 1110 |
| 73 | Grant Avenue at Columbus Avenue | 1097 |

**Table 5: Average Use per Month**

|  |  |  |  |
| --- | --- | --- | --- |
| Month | Total Time Used (hours) | Total Time Available (hours) | Average Utilization |
| August | 9676.478 | 744 | 13.006019 |
| June | 9180.204 | 720 | 12.750283 |
| July | 9362.854 | 744 | 12.584481 |
| September | 8901.983 | 720 | 12.363865 |
| October | 8850.099 | 744 | 11.895295 |
| May | 8839.435 | 744 | 11.880961 |
| April | 7695.530 | 720 | 10.688236 |
| March | 7796.084 | 744 | 10.478608 |
| January | 6553.868 | 744 | 8.808962 |
| November | 6150.414 | 720 | 8.542241 |
| February | 5536.062 | 672 | 8.238187 |
| December | 6100.219 | 744 | 8.199219 |

**Table 6: Correlation between the Number of Trips each Day to the Weather Metric**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Mountain View | Palo Alto | Redwood City | San Francisco | San Jose |
| Max Temperature | 0.35322639 | **0.39334372** | 0.13802499 | 0.33833220 | **0.4289523174** |
| Mean Temperature | **0.38885842** | 0.37015775 | **0.14536614** | **0.34877956** | 0.3998805260 |
| Min Temperature | 0.35239322 | 0.25612388 | 0.12060684 | 0.25897335 | 0.2801188607 |
| Max Visibility | NA | 0.21121117 | 0.01432151 | 0.09984789 | NA |
| Mean Visibility | 0.18297241 | **0.25676173** | **0.16116266** | **0.17443069** | **0.2319322992** |
| Min Visibility | **0.22220406** | 0.24056987 | 0.08595961 | 0.16246166 | 0.2511889400 |
| Max Wind Speed | 0.11951473 | 0.08852608 | -0.03098037 | -0.06332080 | 0.0063324166 |
| Mean Wind Speed | 0.15619448 | 0.10619730 | 0.02380014 | -0.04138217 | -0.000420177 |
| Max Gust Speed | -0.03999753 | -0.03429523 | 0.01890057 | -0.01478028 | -0.042940467 |
| Precipitation | -0.14070440 | -0.08040558 | -0.10447120 | -0.23029968 | -0.252821360 |