# ReadMe

By: Eric Lu

# Preface

1. Acknowledgements
   1. Credit any sources used for ideas. Credit any code snippets where you collaborated with someone else.

* http://www.datasciencemadesimple.com/binning-or-bucketing-of-column-in-pandas-python-2/
* https://stackoverflow.com/questions/44517191/how-to-find-the-python-list-item-that-start-with
* https://stackoverflow.com/questions/37528373/how-to-remove-all-text-between-the-outer-parentheses-in-a-string
* https://thispointer.com/python-check-if-a-list-contains-all-the-elements-of-another-list/
* https://eulertech.wordpress.com/2017/11/28/pandas-valueerror-if-using-all-scalar-values-you-must-pass-an-index/
* https://stackoverflow.com/questions/36606930/delete-an-element-in-a-json-object
* https://stackoverflow.com/questions/34582729/python-how-to-only-retain-elements-from-all-previous-sublists-while-iterating-o
* <https://github.com/cjhutto/vaderSentiment>
* Natural Language Text Processing with Python by Jonathan Mugan, O'Reilly Online Learning

1. Code environment
   1. What tools did you use for your solution? (e.g. Python 3.7, R 3.5, SPSS, etc.

I used Python 3.7 running with Jupyter Notebook

* 1. If your code needs to be run to be graded, list any versions of packages, tools, etc that a grader would need:

Tools: Python 3.7, Jupyter Notebook

Packages/Libraries:

* Pandas
* os
* Matplotlib
* Numpy
* Haversine
* Sklearn
* SciPy
* Re
* Json
* Wordcloud
* Nltk
* vanderSentiment
* genism
* pyLDAvis

# Part 1: Divvy Bike Data

1. Data exploration
   1. What is the source of this data?

Divvy Bike’s public historical trip data (<https://www.divvybikes.com/system-data>) for 2017 Q1 & Q2 and 2017 Q3 & Q4.

* 1. Overall data structure: how many rows and columns are in this dataset?

The ride data is split into 4 quarters (Q1, Q2, Q3, Q4)

* Q1 has 431691 rows, 12 columns
* Q2 has 1119814 rows, 12 columns
* Q3 has 1608270 rows, 12 columns
* Q4 has 669239 rows, 12 columns

The station data is split into 2 files (Divvy\_Stations\_2017\_Q1Q2 and Divvy\_Stations\_2017\_Q3Q4) Q3Q4’s data is the most updated version

* Divvy\_Stations\_2017\_Q1Q2 has 582 rows, 7 columns
* Divvy\_Stations\_2017\_Q3Q4 has 585 rows, 7 columns

1. Data Visualization.   
   For the following 5 questions, create your charts and copy them into your report. Make sure to choose the appropriate chart type, and to include labels and titles on your graphs. Describe any additional insights in your report.   
   Include your numerical results here.
   1. Top 5 stations with the most starts (showing # of starts)

|  |  |  |
| --- | --- | --- |
| Station ID | Station Name | # of Trips |
| 35 | Streeter Dr & Grand Ave | 97567 |
| 76 | Lake Shore Dr & Monroe St | 53396 |
| 192 | Canal St & Adams St | 50911 |
| 91 | Clinton St & Washington Blvd | 49832 |
| 177 | Theater on the Lake | 47908 |

* 1. Trip duration by user type

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Trip Duration by type | | | | |
|  |  |  | Minutes | |
| User type | Trips | Total Hours | Average | Median |
| Subscriber | 2992135 | 584952.1 | 11.73 | 9.53 |
| Customer | 836850 | 431465.58 | 30.93 | 22.82 |
| Dependent | 7 | 1.46 | 12.53 | 13.4 |

* 1. Most popular trips based on start station and stop station

|  |  |  |
| --- | --- | --- |
| Start Station | End Station | # of Trips |
| Lake Shore Dr & Monroe St | Streeter Dr & Grand Ave | 12171 |
| Streeter Dr & Grand Ave | Streeter Dr & Grand Ave | 10042 |
| Streeter Dr & Grand Ave | Theater on the Lake | 8180 |
| Streeter Dr & Grand Ave | Lake Shore Dr & North Blvd | 7993 |
| Lake Shore Dr & North Blvd | Streeter Dr & Grand Ave | 7226 |

* 1. Rider performance by Gender and Age based on avg trip distance (station to station), median speed (distance traveled / trip duration)

Assumption: Removed rides that start and end in the same station

|  |  |  |  |
| --- | --- | --- | --- |
| **Gender** | **Age Group** | **Median Speed (miles/hr)** | **Avg. trip distance (miles)** |
| Female | 17-25 | 6.12 | 1.25 |
| 25-35 | 6.30 | 1.31 |
| 35-45 | 6.07 | 1.28 |
| 45-55 | 5.77 | 1.19 |
| 55-65 | 5.39 | 1.11 |
| 65-70 | 4.93 | 1.15 |
| Male | 17-25 | 6.60 | 1.12 |
| 25-35 | 6.79 | 1.24 |
| 35-45 | 6.60 | 1.22 |
| 45-55 | 6.30 | 1.12 |
| 55-65 | 5.83 | 1.04 |
| 65-70 | 5.50 | 1.02 |

* 1. What is the busiest bike in Chicago in 2017? How many times was it used? How many minutes was it in use?

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Time in Use | | |  | Trip Count | | |
| Bike ID | Trip Duration (Minutes) | Times Used |  | Bike ID | Times Used | Trip Duration (Minutes) |
| 2565 | 22526.42 | 1489 |  | 2565 | 1489 | 22526.42 |
| 5880 | 20693.9 | 1177 |  | 3308 | 1234 | 17662.63 |
| 5293 | 19865.42 | 1049 |  | 3489 | 1225 | 17617.47 |
| 5479 | 19569.18 | 1044 |  | 3128 | 1210 | 18244.3 |
| 5731 | 19356.55 | 1027 |  | 5880 | 1177 | 20693.9 |

1. Data cleaning  
   Before building your model, you will need to clean your data. Describe the steps you take here.
   1. What is the approach to deal with missing values, outliers, skewed data, mixed data type, etc:

A big issue was that trip duration varied a lot. For example, a trip from station A to B could range from taking 10 minutes to 30 minutes to over 12 hours. Even with the removal of outliers, the data showed similar trips with a huge range of trip duration variation.

* 1. What assumptions did you make as you cleaned your data:

I removed all the rides that started and ended at the same station: while rides that start and end at the same station makes sense, because I am basing my trip duration prediction on distance traveled, keeping trips that had a start and end trip distance of 0 would ruin the data.

1. Modelling  
   Build a model that can predict how long a trip will take given a starting point and destination
   1. Feature engineering: What new features did you create?

I used the haversine formula to calculate the distance between two stop. This distance is what I used to plot and train my model.

* 1. External data sources: Describe any external data sources you found and how you merged it into your data:

I used the haversine formula?

* 1. Model selection: How did you select or tune the algorithm you used?)

I selected a linear regression algorithm as there was a distinct linear trend after looking past the outliers.

* 1. Feature selection: What are the features you use in your model? How did you select them?

I removed all the features other distance and usertype because the other data seemed to subjective to use to create an all-inclusive model for users to potentially use at a kiosk. In the future, creating a model that takes into account the user’s age, gender, type of user would be very interesting.

* 1. Model validation: What steps did you take to validate your model? How large are your training/test sets? What is your training error? Testing error?

I created a train and test set with an 80/20 split.

* Train set size: 2298043
* Test set size: 574711
* Training error: 196.56
* Testing error: 196.82  
  1. Improve baseline model: What steps did you take to improve your model?

The baseline model had a score of 0.09. I plotted a box plot and realized that there were too many outliers skewing the model. I removed all the outliers but the model still only had a score of 0.51. I saw that the distribution for subscribers were much more uniform compared to regular customers. I based my model on subscribers only as I presumed it would give a more accurate representation of users trying to get from station A to station B. This allowed my model to reach a score of 0.72.

1. Conclusion and next steps:
   1. What recommendations can you make based on this model? How accurate are your travel time predictions?

This model can pretty accurately predict travel times from one station to another. This model cannot predict the intentions of a user. For example, if a user wanted to stop at a breakfast café to grab some food before heading to the office, this model would be not a good indicator on the users travel time.

R^2 of model on test set: 0.717

# Part 2: NLP

1. Data exploration
   1. What is the source of this data?  
      Yelp Dataset Challenge <https://www.yelp.com/dataset/challenge>
   2. Overall data structure: how many rows and columns are in this dataset?

\* Files were originally JSON files so rows and columns count is based on post DataFrame conversion

* Yelp Business data has 161950 rows, 14 columns
* Yelp Checkin data has 161950 rows and 2 columns
* Yelp Review data has 6685900 rows and 9 columns
  1. List any data cleaning steps and assumptions you make in your NLP analysis

I am only looking at 2017 data as this is an Operating Report for 2017

1. Data visualization  
   Create a word cloud to show the most popular keywords or phrases that reviewers use for each cuisine. Include this image in your operating report.
   1. What are the top 10 Cuisine types (Mexican, American, Thai, etc) based on the number of restaurants and number of check ins?

* I combined American (New) and American (Traditional) into a single cuisine “American”
* I selected the top 10 cuisines from restaurant category by hand
* Did not take into account of overlapping businesses as that would be subjective to each business

|  |  |
| --- | --- |
| Cuisine | # of check-ins |
| American | 437747 |
| Mexican | 148515 |
| Japanese | 134682 |
| Asian Fusion | 107330 |
| Italian | 102731 |
| Chinese | 88369 |
| Korean | 49618 |
| Mediterranean | 44528 |
| Thai | 42237 |
| Vietnamese | 36642 |

* 1. For the 10 most popular cuisines, what are the top keywords or phrases used by reviewers?

|  |  |
| --- | --- |
| Cuisine | Top keywords & phrases |
| American | us, chicken, even, come, nice, best, always, love, menu, first |
| Mexican | tacos, mexican, taco, best, salsa, chicken, us, love, even, always |
| Japanese | sushi, ramen, come, us, try, best, love, i've, always, even |
| Asian Fusion | chicken, sushi, try, come, us, rice, love, even, i've, best |
| Italian | pizza, us, italian, best, pasta, even, nice, love, always, sauce |
| Chinese | chicken, chinese, fried, rice, soup, beef, come, even, noodles, always |
| Korean | korean, chicken, come, us, pork, meat, try, spicy, rice, side |
| Mediterranean | chicken, best, love, try, i've, greek, nice, always, salad, us |
| Thai | thai, pad, chicken, curry, rice, i've, best, fried, love, always |
| Vietnamese | pho, vietnamese, beef, chicken, broth, come, spring, best, i've, try |

Common keywords: got, definitely, service, came, food, go, really, one, ordered, place, time, restaurant, like, get, also, would, order, back, good, great

* I parsed through the Yelp review.json keeping only 2017 reviews
* I masked the DataFrame to show only reviews to restaurant businesses
* I mapped all the businesses to each of the top 10 cuisines
* I calculated the word frequency of all reviews for each of the cuisines
* I generated a word cloud based on the word frequencies of each cuisine
* I found common keywords that all 10 cuisines reviews used, removing these common keywords leaves us with more relevant top keywords for each type of cuisine

1. Topic modeling  
   Define a set of topics by applying topic modeling algorithms such as LDA on textual reviews. Choose an optimal number of topics in a data-driven fashion such as by using a figure that plots perplexity versus number of topics.
   1. What topic modelling techniques did you apply?  
      I used LDA to classify the topics from the review text
   2. How did you select the optimal number of topics? How many topics did you define?

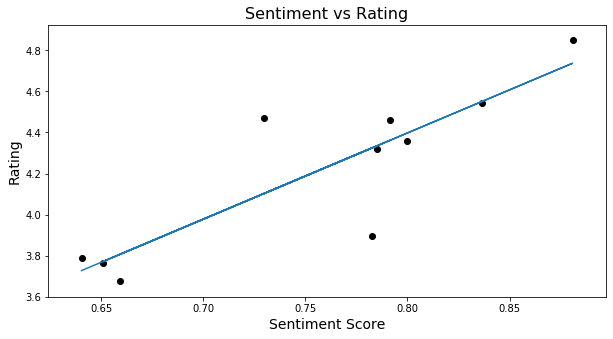
I defined 2, 4, 6, 8, 10 topics and plotted them on an intertopic distance map.

Based on the distance map, 4 topics is the most optimal. At 2, the topics are 2 extremities but do not encompass enough topics. Past 4, the topics begin to overlap.

1. Sentiment & Correlation  
   Calculate sentiment score on each review to answer the question: how strong is the correlation between star rating and number of reviews?
   1. What are the top 10 restaurants by number of check ins? What is the sentiment score of their reviews?

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Check-ins | Avg. Star Rating | Sentiment Score |
| Gangnam Asian BBQ Dining | 11081 | 4.458689 | 0.791647 |
| Kung Fu Tea | 8715 | 4.468468 | 0.729564 |
| Nacho Daddy | 6901 | 4.31733 | 0.785049 |
| Bacchanal Buffet | 4693 | 3.674051 | 0.659237 |
| Brew Tea Bar | 3956 | 4.849138 | 0.880873 |
| Egg & I | 3938 | 4.541889 | 0.836465 |
| Nacho Daddy Downtown | 3226 | 4.356643 | 0.799597 |
| The Cosmopolitan of Las Vegas | 3143 | 3.761803 | 0.650749 |
| Momofuku Las Vegas | 2576 | 3.897059 | 0.78264 |
| The Venetian Las Vegas | 2450 | 3.78934 | 0.640318 |

* 1. What steps did you take to calculate sentiment score?
* I selected the top 10 restaurants based on the number of check ins
* I masked out only reviews related to the top 10 restaurants
* I calculated the sentiment on each review with VADER Sentiment Analysis
* I calculated the average star rating and sentiment score for each business based on 2017 reviews and not the overall star rating of the business across all the years it has been open
* I plotted the sentiment vs rating on a scatter plot and found the correlation between the two
  1. How strong is the correlation between sentiment score and star rating?



* r^2 : 0.7488331150937708
* correlation : 0.8653514402217002

There is a high correlation between average sentiment score and rating of a restaurant