# Preface

1. Acknowledgements
   1. Stack Overflow and the IBM documentation on SPSS
2. Code environment
   1. Python 3.7 and SPSS
   2. Should not need to be run… but if so, the grader would need the haversine package for Python 3.7

# Part 1: Divvy Bike Data

1. Data exploration
   1. The Divvy website, 2017 Q1-Q4
   2. Overall data structure: how many rows and columns are in this dataset?
      1. Divvy trips – 3,829,014 rows; 12 columns
      2. Divvy stations – 369,381 rows; 7 columns
2. 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 | 97569 |
| 76 | Lake Shore Dr & Monroe St | 53398 |
| 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 |  |  |  |  |
| Customer |  |  |  |  |
| Dependent |  |  |  |  |

* 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)

|  |  |  |  |
| --- | --- | --- | --- |
| **Gender** | **Age Group** | **Median Speed (miles/hr)** | **Avg. trip distance (miles)** |
| Female | 17-25 | 3.763444143 | 0.808191571 |
| 25-35 | 4.222750727 | 0.913970182 |
| 35-45 | 3.924982 | 0.878464545 |
| 45-55 | 3.774649182 | 0.820846364 |
| 55-65 | 3.352927545 | 0.760800909 |
| 65-70 | 2.475776 | 0.66448475 |
| Male | 17-25 | 3.938338556 | 0.701571444 |
| 25-35 | 4.654489091 | 0.870021 |
| 35-45 | 4.806502364 | 0.872742909 |
| 45-55 | 4.643323909 | 0.802413818 |
| 55-65 | 4.445533 | 0.767918636 |
| 65-70 | 3.763947167 | 0.7200355 |

* 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 |  | 5880 | 1177 | 20693.9 |
| 5293 | 19865.42 | 1049 |  | 5293 | 1049 | 19865.42 |
| 5479 | 19569.18 | 1044 |  | 5479 | 1044 | 19569.18 |
| 5731 | 19356.55 | 1027 |  | 5731 | 1027 | 19356.55 |

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:
      1. I dropped null values and merged all the files into dataframes, which I then left joined trips on stations to get latitude and longitude for each trip
   2. What assumptions did you make as you cleaned your data:
2. 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?
      1. Added a model that will classify tripduration based on other values
   2. External data sources: Describe any external data sources you found and how you merged it into your data:
      1. None
   3. Model selection: How did you select or tune the algorithm you used?)
      1. Used only classification models, and ensembled four (neural net, random forest, random tree, and linear regression)
   4. Feature selection: What are the features you use in your model? How did you select them?
      1. Ensembling, to increase accuracy
   5. 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?
      1. Using SPSS, specified to use 75% training set, 25% testing set
   6. Improve baseline model: What steps did you take to improve your model?
      1. Did not have time
3. Conclusion and next steps:
   1. What recommendations can you make based on this model? How accurate are your travel time predictions?

# Part 2: NLP – I did not have time to complete this portion

1. Data exploration
   1. What is the source of this data?
   2. Overall data structure: how many rows and columns are in this dataset?
   3. List any data cleaning steps and assumptions you make in your NLP analysis
2. 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?

|  |  |
| --- | --- |
| Cuisine | # of check-ins |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

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

|  |  |
| --- | --- |
| Cuisine | Top keywords & phrases |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |



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?
   2. How did you select the optimal number of topics? How many topics did you define?
2. 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 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

* 1. What steps did you take to calculate sentiment score?
  2. How strong is the correlation between sentiment score and star rating?