MINING FOR RIDESHARE PATTERNS

Ronald Thompson
Jonathan Prindle
Saloni Sharma

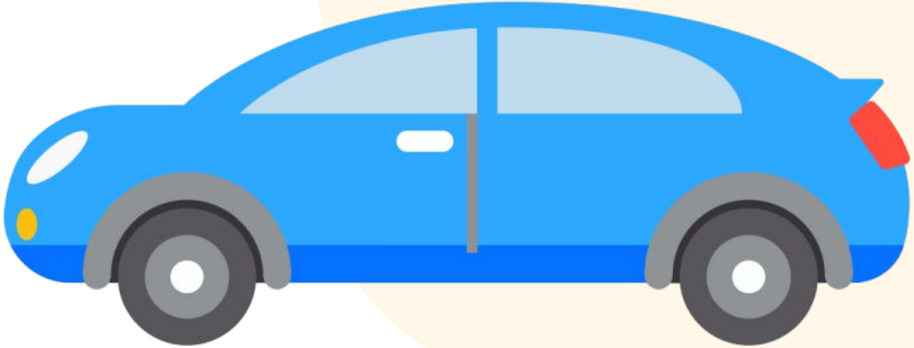


TABLE OF CONTENTS

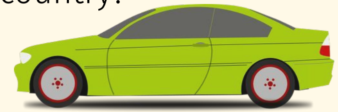
01 What is our goal?

- Description
  - Reveal rideshare patterns that can improve customer experiences and optimize profits

- Understand the patterns that rideshares take
  - Are there recurring patterns that show when peak traffic is hit?
  - What are the average duration of rideshares at different times?
  - When are the better and worse times to take rideshares?
  - When do drivers hit diminishing returns?

- Comparing rideshares to census data, looking at demographic differences in rideshares
  - Are prices and tips different based on neighborhoods? Is there any discrimination?
  - What are the most common areas in a city that rideshares start from and end at?
  - Do the trips demonstrate people going to work or to do other activities?

- Are there patterns that are city specific or more generalizable to all rideshares across the country?
  - Do patterns with similar attribute values produce similar results?

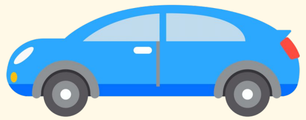


02 What work has been done before in this area?

- Open source problem
  - Many examples on Towards Data Science and Medium blogs
    - Analyzing 1.1 Billion NYC Taxi and Uber Trips, with a Vengeance
  - Kaggle competitions for this problem
    - 538
    - Google/Coursera
  - University departments have done work on it
    - Arizona State
    - MIT

- **Rideshare companies** have done considerable research of their own and publish information on blogs
  - Uber's case studies

- **Local governments** are also interested in learning from the data in order to create better regulation and understand changes in traffic flows

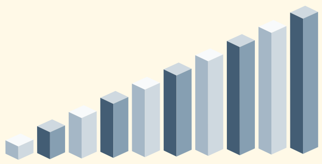


03 What data are we using?

- **Dataset 1:** City of Chicago
  https://data.cityofchicago.org/Transportation/Transportation-Network-Providers-Trips/m6dm-c72p
  - 129 million rides between 2018 and 2020 (Chicago)
  - Rich feature set (21 columns): location of pick-up and dropoff, fare, tip, pooled or not, etc.

- **Dataset 2:** FiveThirtyEight Kaggle Competition
  https://www.kaggle.com/fivethirtyeight/uber-pickups-in-new-york-city#other-Firstclass_B01536.csv
  - Almost 18 million rides between 2014 and 2015 (NYC)
  - Attributes are limited to time and location
  - 538 was able to split out Uber vs Lyft etc

- **Dataset 3:** NYC Taxi and Limousine Commission
  https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page
  - Over 1 million data points per month (NYC)
  - Data from 2015 for for-hire vehicles (which can generally be classified as rideshare)
  - Attributes include timestamps and start and end ‘taxi zones’ for locations
  - Will mainly be used to see rates of pickups

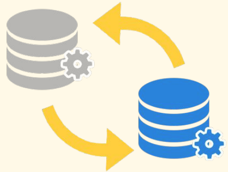


(All datasets downloaded on Saloni’s hard drive.)

04

What will we accomplish and how?

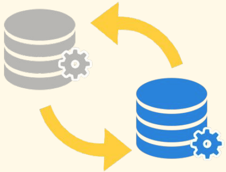
- Data munging
  - Cleaning:
    - Need to clean the various cities datasets and standardize the fields
    - Remove any duplicates and missing or unnecessary values
    - Binning by fare (dataset 1 already rounds fares to every $2.50)
  - Preprocessing:
    - Potentially transform and reduce number of attributes with dimension reducing algorithms such as PCA or backward elimination
  - Integration:
    - Merge 2014 and 2015 data together for dataset 2
    - Combine monthly data to get all of 2015 data for dataset 3
    - Possibly combine dataset 2 & 3 since they are both for NYC



04

What will we accomplish and how?

- Analysis
  - Run a number of different analyses to understand temporal changes and differences between attributes (such as city, neighborhood) with a focus on Chicago's comprehensive data
  - Correlation and Linear regression:
    - Lift measure: between time or location and fares
    - Chi-squared test: whether tip was given with variables like short vs. long trip durations
  - Clustering:
    - Utilize data visualization to aid in analyzing geographical data
  - Predictive modeling:
    - Find main attributes to predict price (time of day, duration of ride, location)



05 What tools will we use?

- **Programming language:** Python (via Jupyter Notebooks, VS Code)
  - SciPy, Pandas, Matplotlib, NumPy
  - GeoPy, GeoPandas, Tableau
- **Data:** three datasets in .csv files
  - All downloaded, but may also pass flat files around with the team
- **Code:** storing and sharing code through GitHub

06 How will we evaluate our results?

- See how our models compare to other analyses
  - Does our analysis show positive or negative correlation of high fares in certain locations and at specific times of day?
  - Does that match with prior work done on rideshare data?

- Have hold outs of data when running models to see how it performs out of sample
  - Can we use them to predict times or locations with high fares or tips?
  - Can we use the analysis of Chicago data to predict information for NYC, and vice versa?

- See if our anecdotal experiences match up with our findings
  - As riders
  - As/from drivers