Lyft Baywheels: An Analysis and Machine Learning Project by Riley Predum

Problem Statement

- How can I better understand who rides Baywheels? What are the demographic patterns?
- Can I predict a rider type by features such as age, gender, ride duration?

Background Context

- Looked at Lyft's repository of bikeshare trips
- Data spans 2017 to present thanks to a real-time API
- Wanted to explore these data to understand trends and make predictions with machine learning

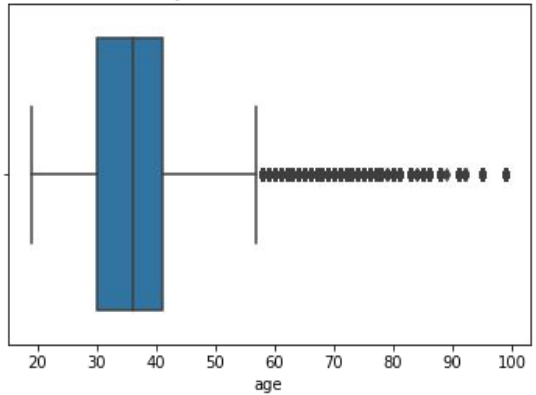
Data Science Workflow Steps: Data Acquisition and Cleaning

- Scraped repository and unzipped files programmatically
- Created a master dataframe from all the different files
- Removed unnecessary columns, transformed strings to datetime, created new features (feature engineering)

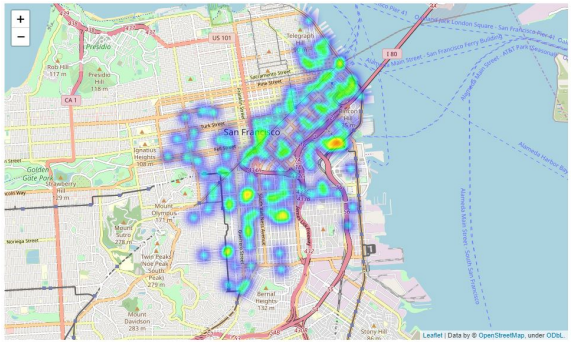
Data Science Workflow Steps: Exploratory Data Analysis

- Explored distributions of variables to answer my first question
  - Median age ~38 years
  - Men use them the most (blue in far right chart)
  - Rides start in Financial District, South of Market, Mission

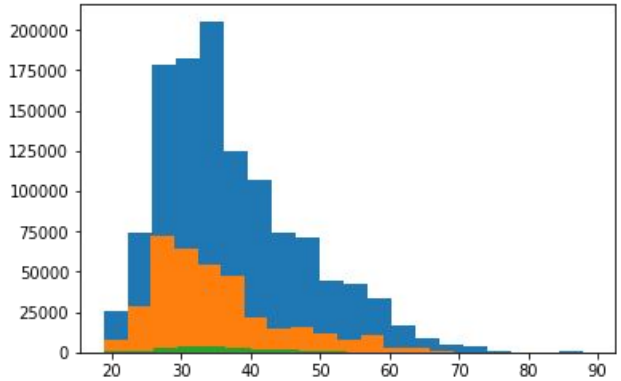
Age distribution



Ride start locations

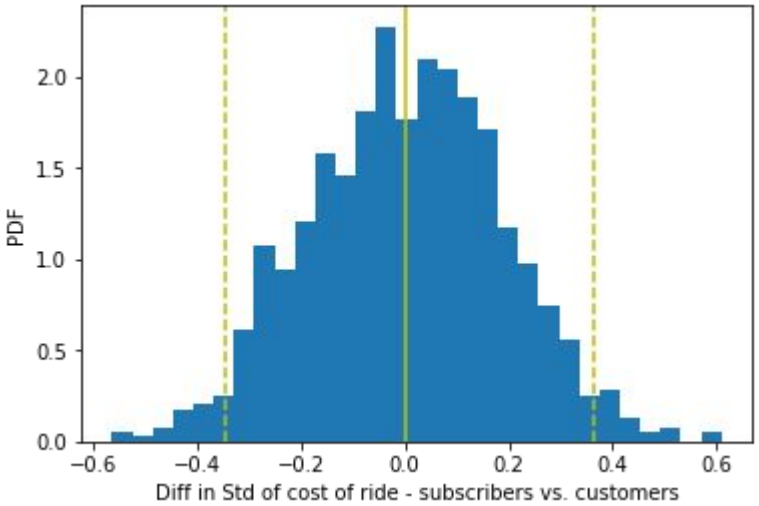


Age by gender



Data Science Workflow Steps: Statistical Tests

- Tested subscriber vs. customer and cost of ride
  - No significant difference between user type and how much they spend on rides



Data Science Workflow Steps: Machine Learning

- Performed linear regression to predict duration of ride
  - Irrelevant results
- Performed logistic regression to predict user type
  - Strong performance but need to balance classes in future iteration of project

```
[[   18  1075]
 [   19 13238]]
              precision    recall  f1-score   support

           0       0.49      0.02      0.03      1093
           1       0.92      1.00      0.96     13257

    accuracy                           0.92     14350
   macro avg       0.71      0.51      0.50     14350
weighted avg       0.89      0.92      0.89     14350
```

Summary and Conclusion

- Late 30s millennials use bikeshare the most
- Men use the bikes the most, followed by women, followed by other
- Most rides originate in Financial District, South of Market, or the Mission
- User type (subscriber vs. customer) predictable based on ride duration, age, gender
- Linear regression seems to make less sense (possibly because ride duration is exponentially distributed - most rides are less than 30 min long)