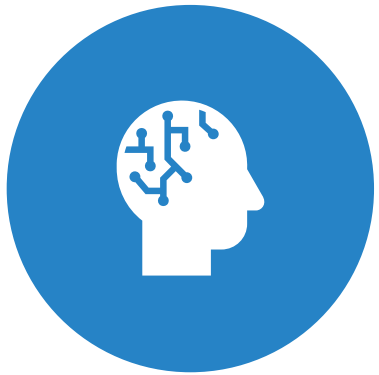
BIKE RENTAL CASE

Milos Randic

ON THE AGENDA

BUSINESS CONTEXT

KEY FINDINGS



MACHINE LEARNING MODEL

BUSINESS CONTEXT

BUSINESS CONTEXT

Wheelie Wonka, a rental bike company from Boston, has increased need for data driven understanding for their customers needs and behavior.

To enhance their mobile app experience, they would like to show bike availability in the future to their users.

An important component in achieving this goal is to predict, at the beginning of a bike trip, the trip duration in Boston, MA, USA.

The requirement from bike company is:

1. to explore customer base behavior from provided data
2. build a data product to support estimates of trip durations before starting the service.

KEY FINDINGS

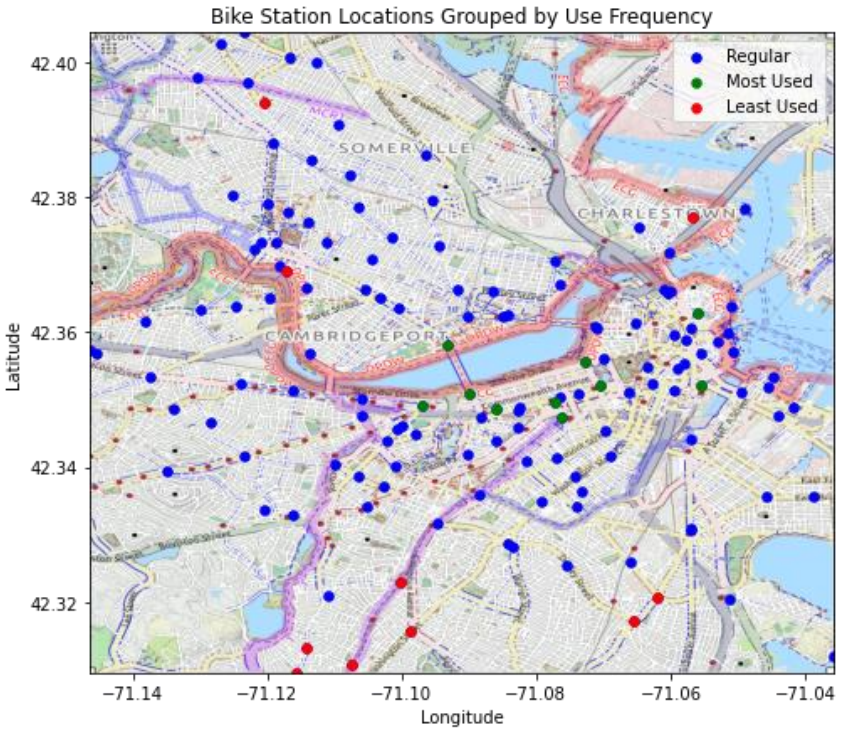
KEY FINDINGS: OVERALL ANALYSIS

- Data presents 3 years (2011-2013) of bike trip recorded usage behavior.
- Bike company provides total 1.163 bikes for use across 142 different stations within the city (12 removed stations).
- Bike service model applies for both registered users (1.095.689 trips) and non-regisered, casual users (417.138 trips).
- Bike trips are usualy intiated within Boston municipality (in 77% of cases). Brookline has only 1% of initiated bike trips.

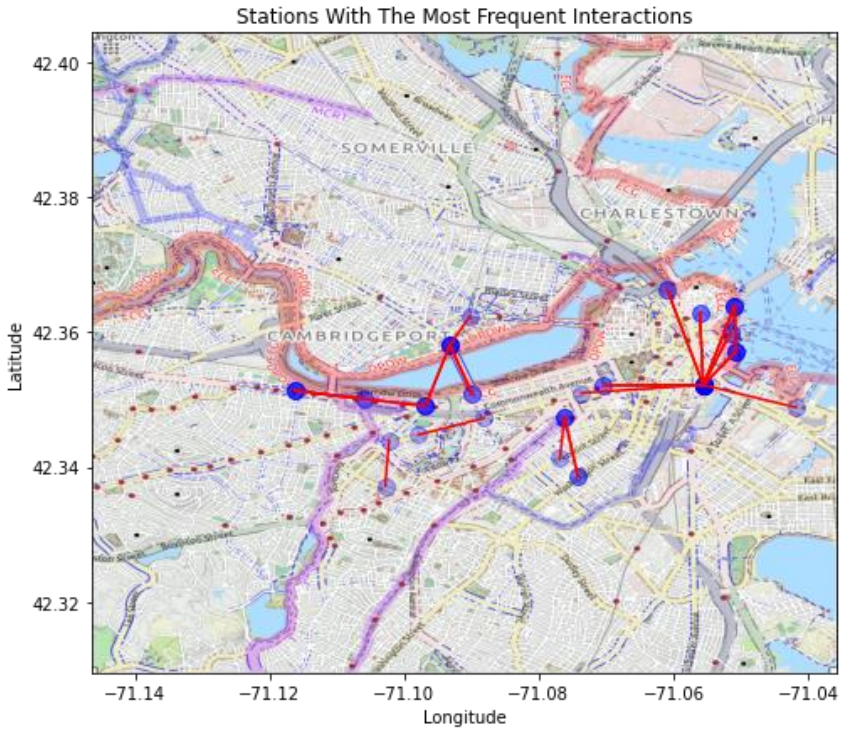
Bike Ride Share Per Municipality

Municipality	Bike Ride Count	Share (%)
Boston	1.165.665	77
Cambridge	288.706	19
Somerville	38.457	3
Brookline	19.987	1

KEY FINDINGS: GEO SPATIAL ANALYSIS ON BOSTON CYCLING MAP

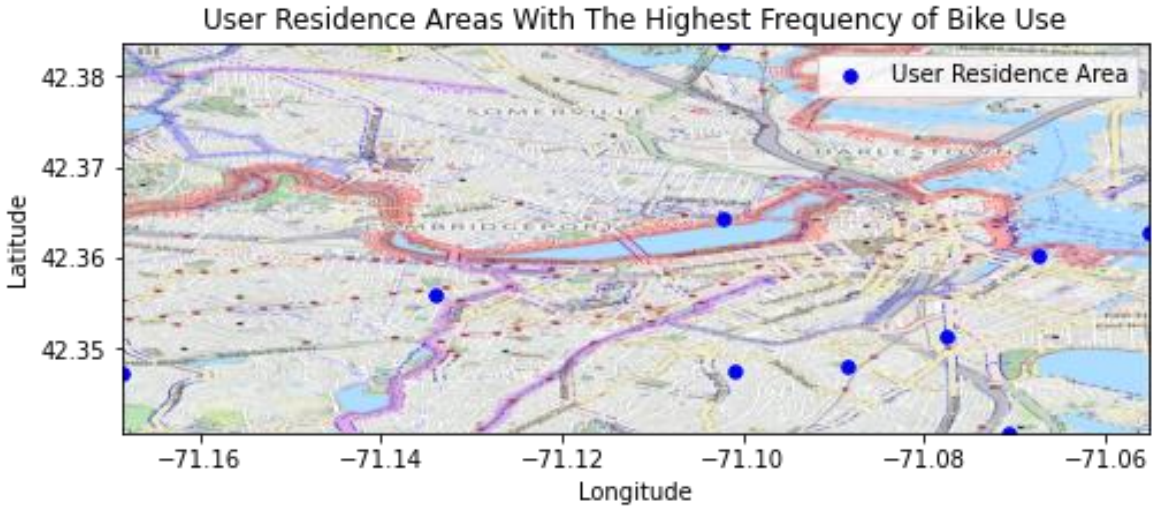


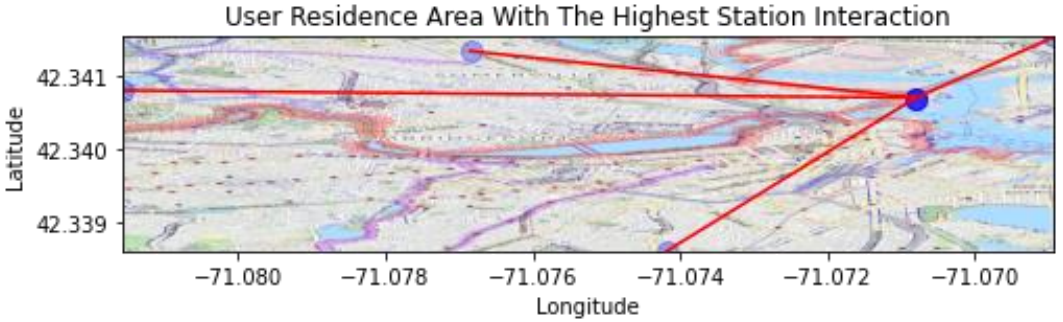
- Bike stations are more frequently used in Boston city center.
- In out-of-center areas bikes are used much less frequently.



- In most cases, users tend to drop off bikes to stations closer to starting station.
- Map shows 3 most common interaction areas, and these are located in central-east part of Boston.

KEY FINDINGS: GEO SPATIAL ANALYSIS ON BOSTON CYCLING MAP

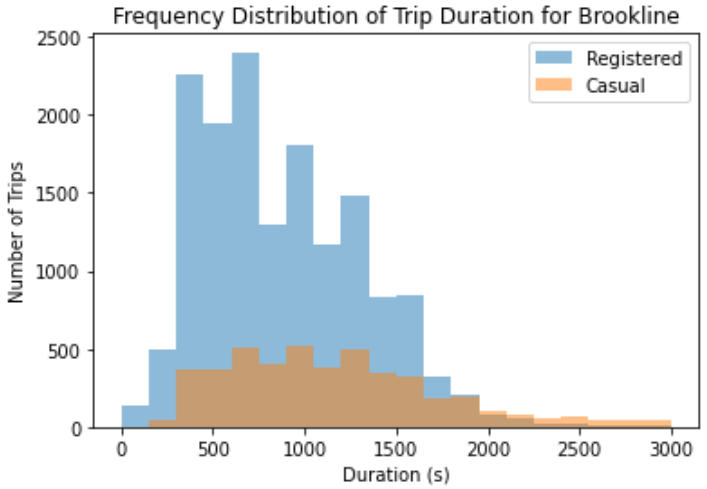


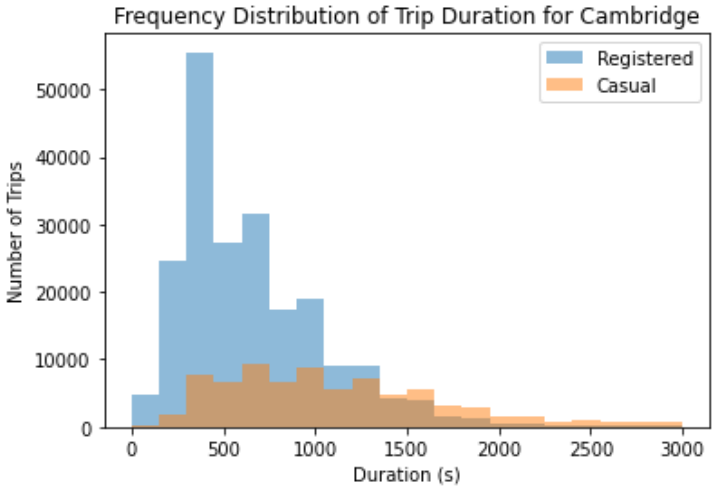


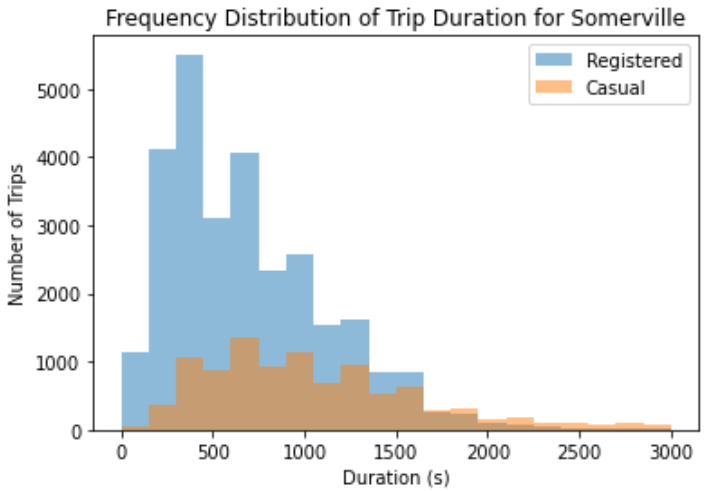
- Registered users living primarily in east-central areas tend to use bike service the most.
- User residence location is approximately estimated based on zip code they've provided.

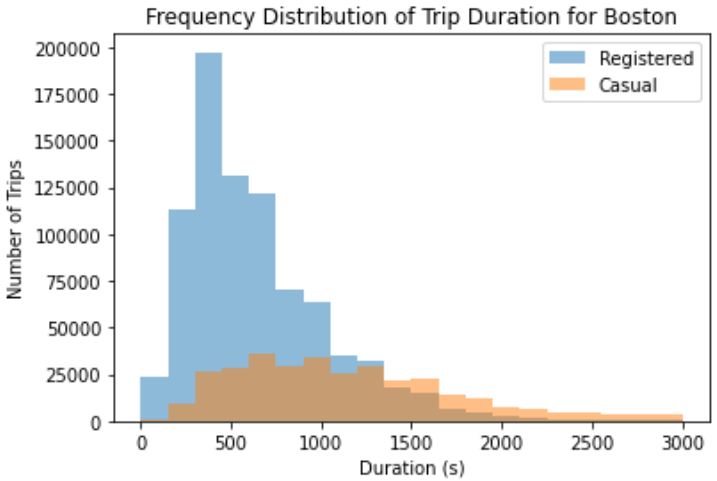
- Out of all engaged city areas, the area in eastern part has the highest engagement in using bike services.
- Customers living in these areas tend to take most of the bike trips from these bike stations.

KEY FINDINGS: TRIP DURATION ACROSS MUNICIPALITIES



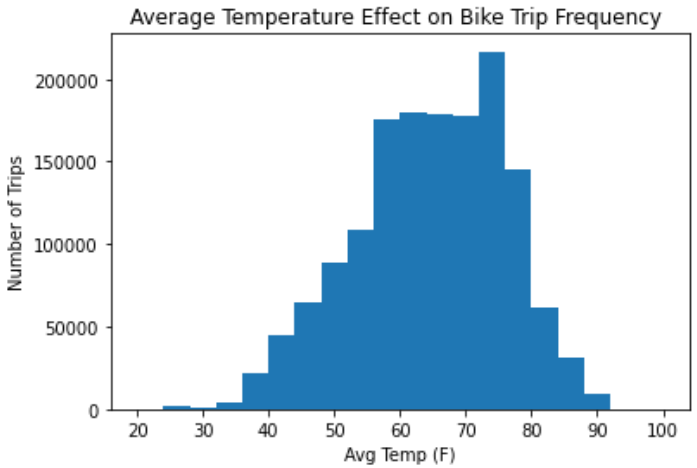


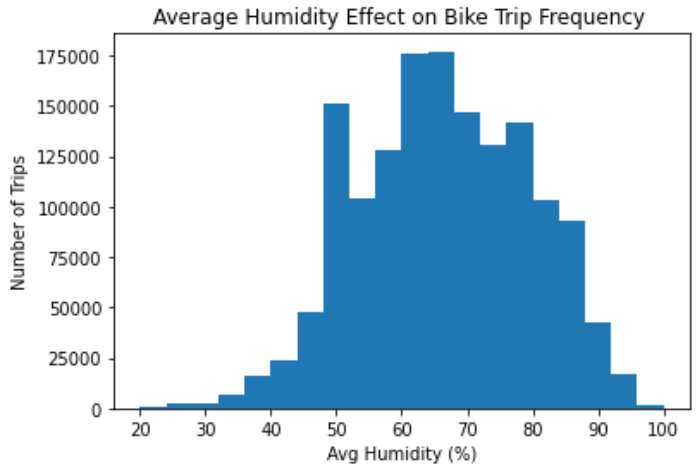


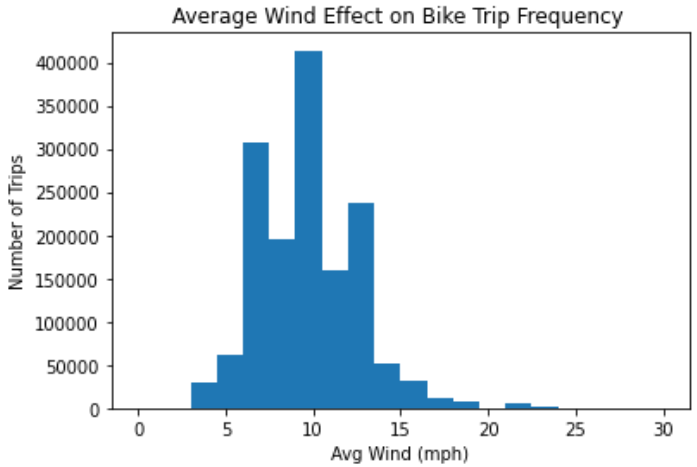


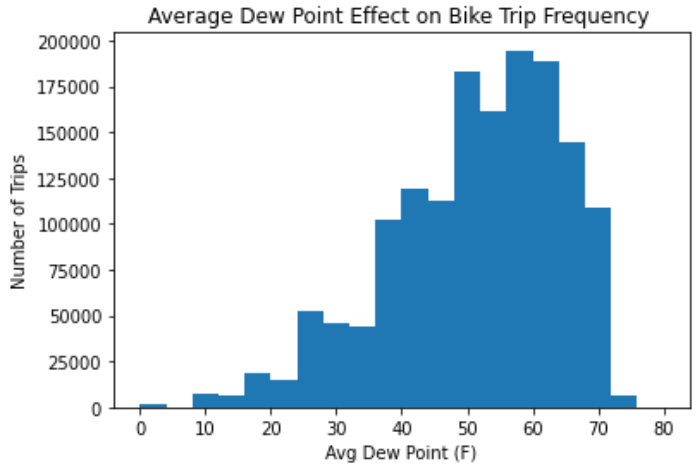
- Casual users tend to take longer bike trips comparing to registered bike users.
- The rationale behind this potentially lies in need to maximize service use for one time fee these users might have to pay.
- Registered users from Brookline and Somerville tend to take longer bike trips. This fact makes sense since these 2 municipalities are a bit far away from city center and users might want to have a bike trip closer to city center.

KEY FINDINGS: WEATHER IMPACT





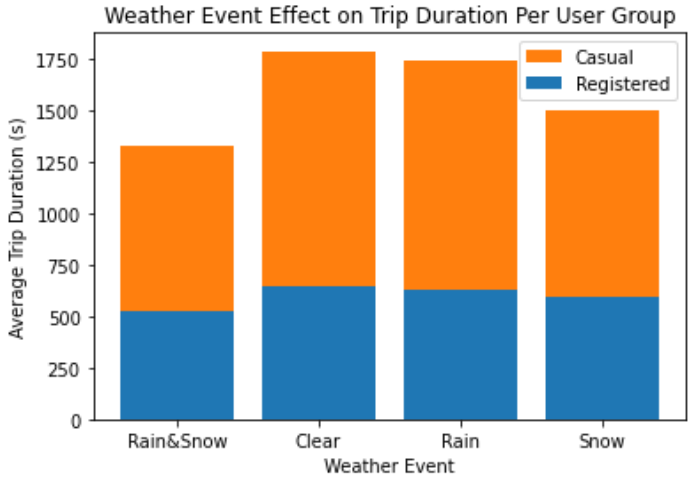


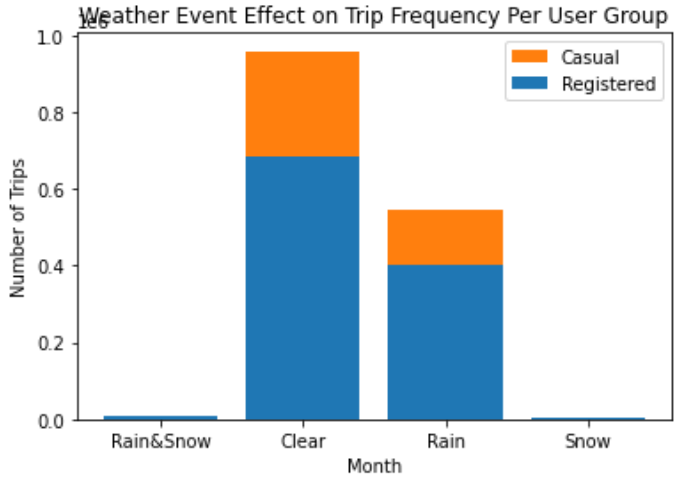


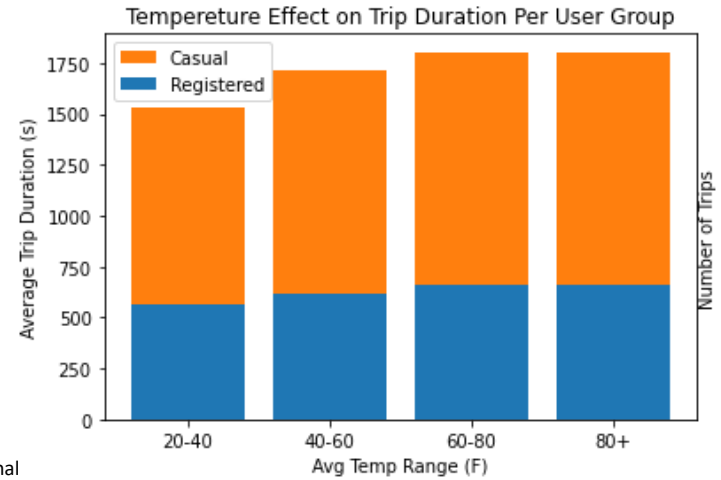
Weather impact on users increased tendency for bike use fits into expected intervals:

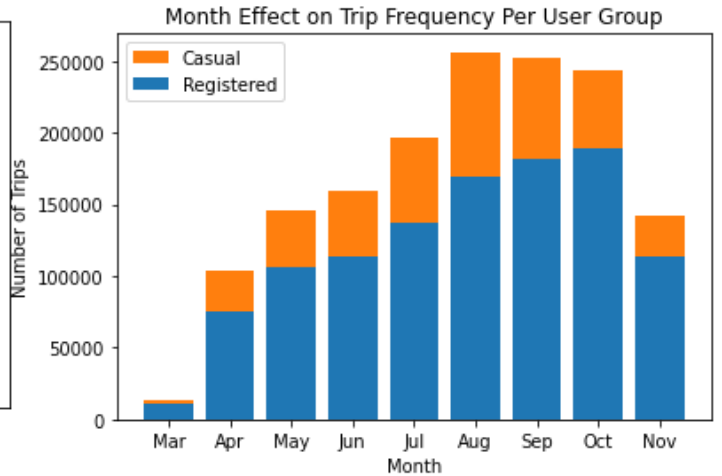
- Temperature (F): 60-80
- Humidity (%): 50-70
- Wind (mph): 5-10
- Dew Point (F): 50-60
- Visibility (mi): 9-10
- Precip (in): 0-0.1

KEY FINDINGS: WEATHER IMPACT AND SEASONALITY PATTERNS PER USER TYPE



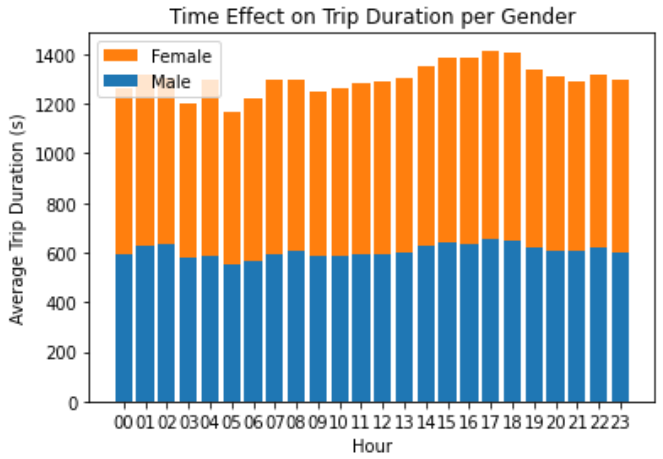


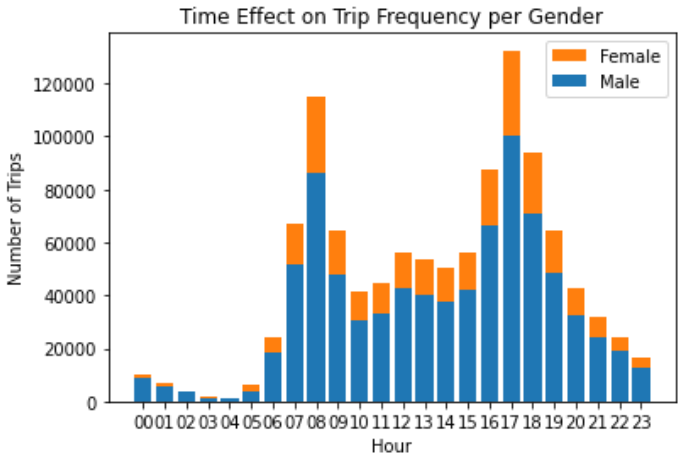


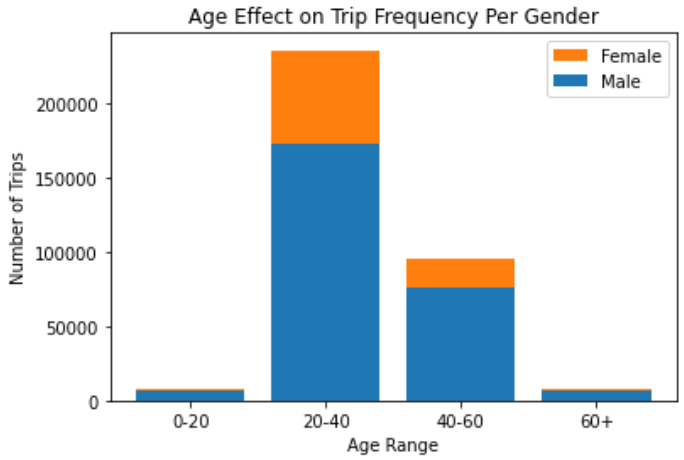


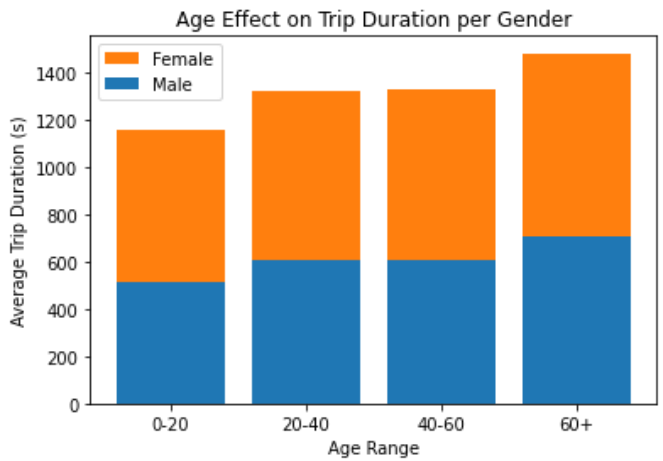
- Casual users have increased tendency to use bikes on clear weather, while on poor weather conditions they tend to decrease trip duration, relative to registered users.
- Both user groups slightly tend to take longer bike rides when temperature increases.
- Bike service utilization is the highest during summer-autumn season.

KEY FINDINGS: WEATHER IMPACT AND SEASONALITY PATTERNS PER GENDER





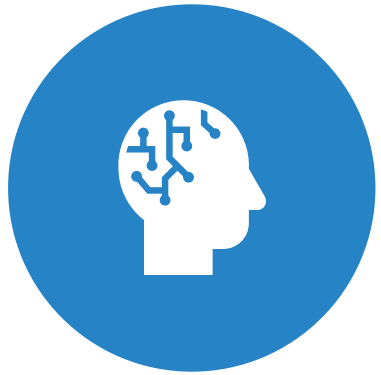




Key findings:

- Men have significantly higher tendency to use bike comparing to women, apprx. 80%.
- There is no significant influence of hour of day on trip duration among both genders.
- Both genders tend to use bikes around rush hour peaks (07-10h and 16-19h). This could indicate the commute need for bike during the week. (Tip: It could be interesting to observe hourly distributions over different months (i.e. vacation seasonality patterns)).
- As age increases, both genders tend to take longer trips and less frequently.
- Most of bike users belong to 20-40 age group.
- After age 40, women tend to decrease more in bike use frequency comparing to men.

MACHINE LEARNING MODEL



DATA PRE-PROCESSING AND FEATURE CREATION

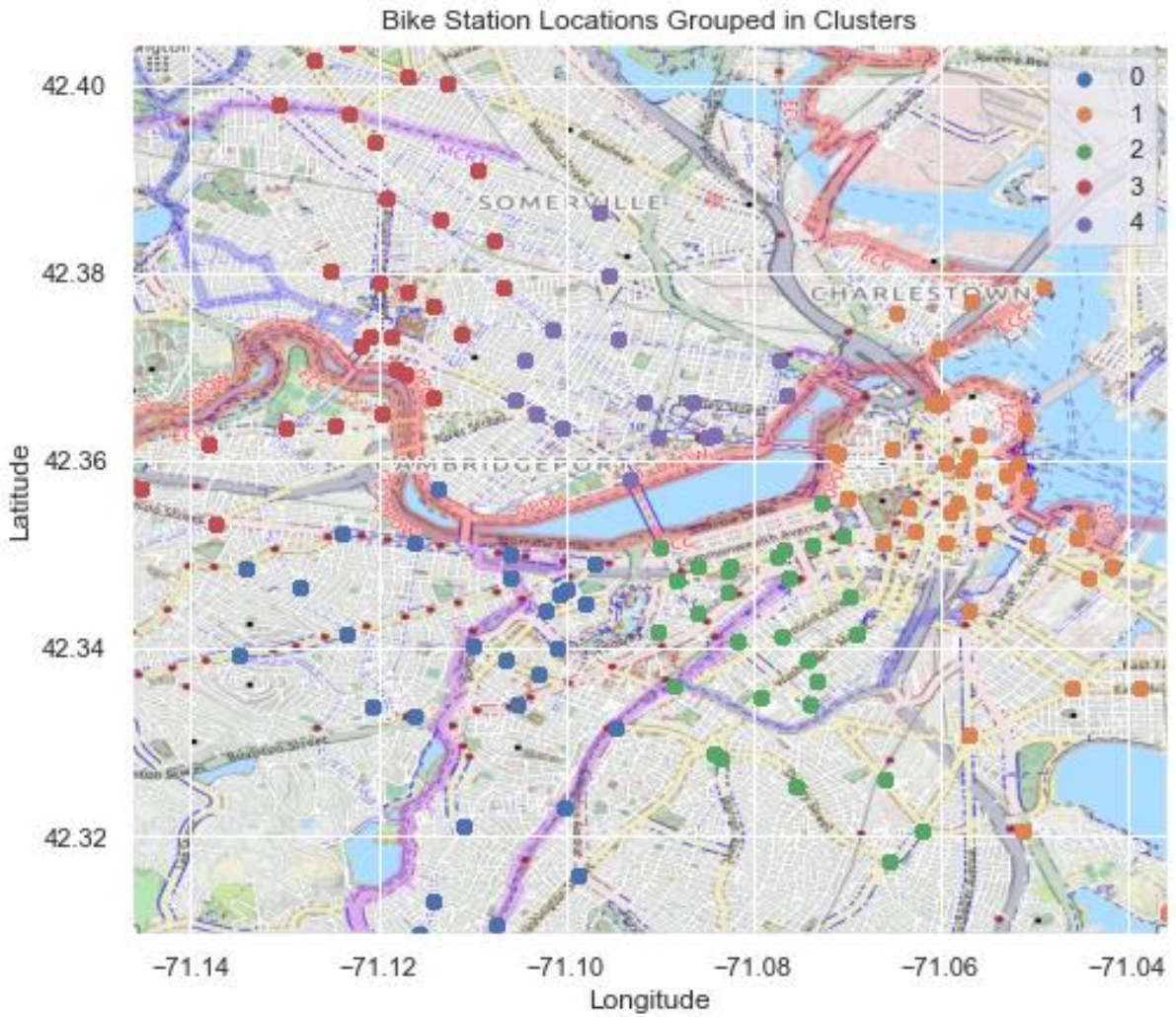
Data pre-processing

- A subset of variables from initial dataset has been selected for feature preparation.
- Numerical features are checked for distribution patterns and potential outliers.
- Outliers have been removed and missing values imputed with average population values.
- Categorical features are one-hot encoded in order to be converted to numerical format.

Feature engineering

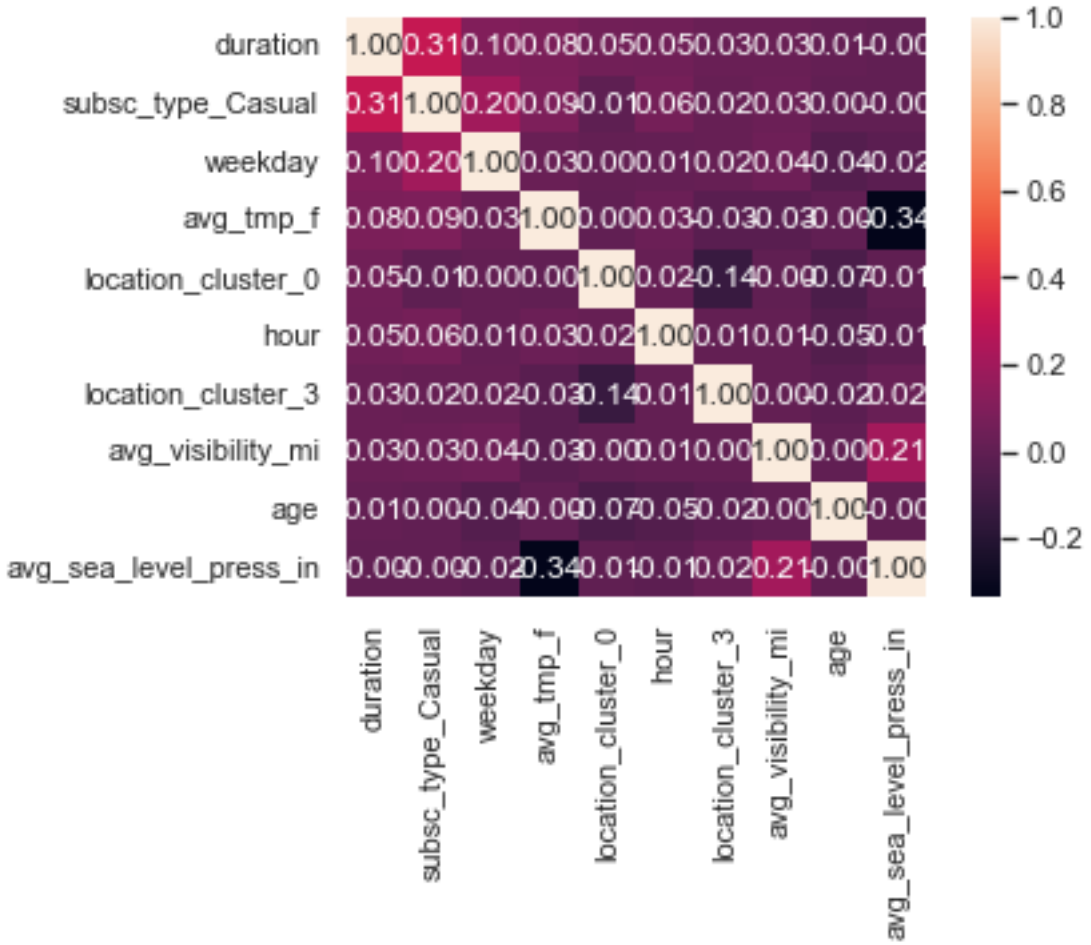
From dataset, following additional features are created:

- Age of registered user
- Avg Visibility Range (mi)
- Avg Temperature Range (F)
- Avg Wind Range (mph)
- Location Cluster (5 clusters using K-Means clustering technique) – displayed on the map.



FEATURE SELECTION

- Feature selection is performed using Pearson correlation.
- Top 10 most correlated features with target variable (trip duration) are selected as model input.
- The heat-map shows casual user as the most correlated feature to trip duration followed by 9 other features.



MODEL SELECTION AND CONFIGURATION

For this type of problem (continuous variable prediction), regression family of models are prefered choice.

Initial set of models has been selected for competition (with default parameters applied):

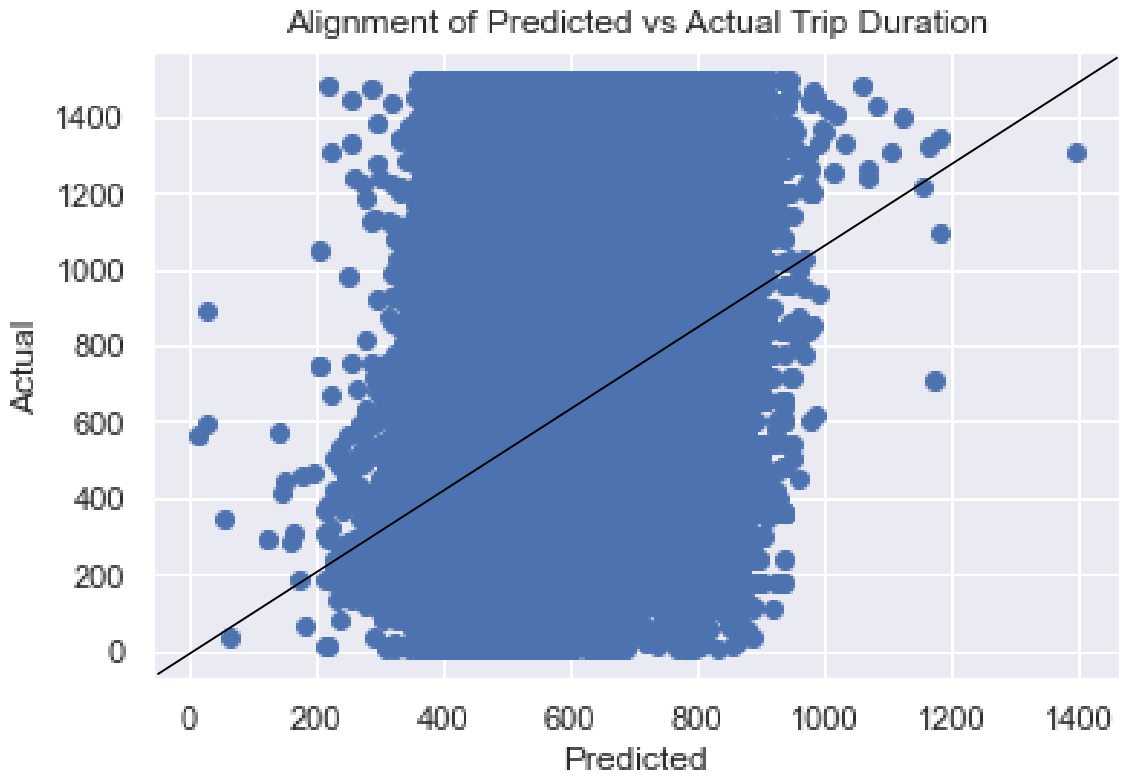
- LinearRegression
- Lasso
- ElasticNet
- KNeighborsRegressor
- DecisionTreeRegressor
- XGBRegressor
- GradientBoostingRegressor

Cross-Validation comparison	mean	std
XGBM	-0.416756	0.009928
LR	-0.414483	0.008387
LASSO	-0.491182	0.010600
EN	-0.491182	0.010600
KNN	-0.483043	0.008868
CART	-0.694894	0.013379
GBM	-0.408345	0.008669

- The winning model is XGBoost Regression (based on max MEAN and min STD of cross-validation using 5 k-folds this is GBM, but XGBM is more efficient).
- Additional reason for choosing XGBoost model is large number of observations in training data.
- Train set covers 70% of the whole dataset, and test set contains 30% of data.
- Further, winning model has been hyperparameter tuned on number of estimators and constant L2 regularization (alpha=0.2).

MODEL TEST AND EVALUATION

- After multiple trials and experiments, selected model resulted with **MSE=0.35**.
- This value uses log() scaled *trip duration*, since large trip duration numbers may affect the absolute numbers of the regression model.
- Apparently, model does not satisfy fitting criteria.
- For a good result, we should get positive correlation trend pattern fitting to 45 degree line.
- There is a slight indication for better performance on predicting shorter trip durations.
- The advice is not to use the model in final production.



MODEL TEST AND EVALUATION: EXAMPLE PREDICTION

Expected Trip Duration	Predicted Trip Duration	Difference
119	542	422
180	465	285
540	496	43
419	441	21
1500	840	659
299	507	207
1224	796	428
509	762	252
1500	763	736
1080	544	535