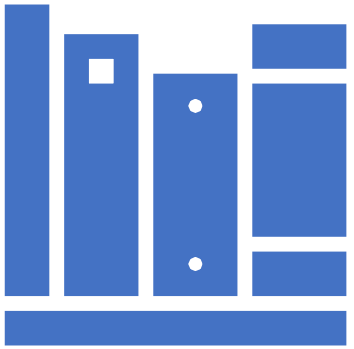


Exploratory Data Analysis

IDS Project Presentation

N Pruthvi ~ PES1201801672
G Deepank ~ PES1201800395

Dataset Description

1. Los Angeles Metro Bike Share Trip Data

2. Hosted by the city of Los Angeles. Open data platform. Information constantly updated according to the amount of data that is brought in

3. Dataset consists of 22 columns and 132427 rows with many nans and few outliers

4. Analysis on this data to draw meaningful conclusions supported by visualizations and outputs

Data Pre-processing

```
% of missing data in each feature:
LA Specific Plans                 88.066633
Zip Codes                         25.551436
Council Districts                 25.551436
Neighborhood Councils (Certified) 25.551436
Starting Lat-Long                 25.527272
Census Tracts                      0.829891
Precinct Boundaries                0.829891
Ending Lat-Long                    0.793645
Ending Station Longitude           0.793645
Ending Station Latitude            0.793645
Plan Duration                      0.578432
Ending Station ID                  0.072493
Starting Station Longitude         0.036246
Starting Station Latitude          0.036246
Starting Station ID                0.014348
Bike ID                            0.007551
Trip Route Category                0.000000
Passholder Type                    0.000000
End Time                           0.000000
Start Time                         0.000000
Duration                           0.000000
Trip ID                            0.000000
dtype: float64
```

- Data provided was clean but consisted of missing values.
- Columns with more than 40% of the missing values were dropped.
- In many cases, rows with nan were dropped.
- Columns where no significant insights could be drawn upon were deemed as unnecessary and hence ignored.
- Data cleaning was performed while elucidating the possible insights from the dataset.

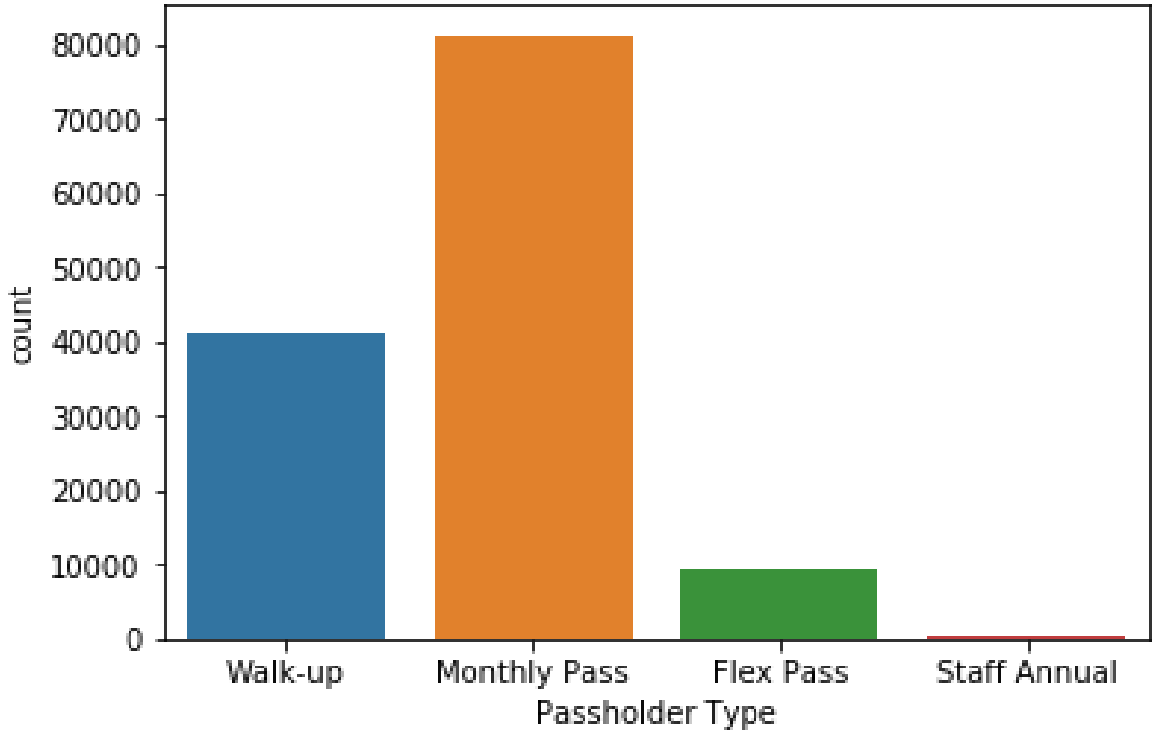
Insights

- Which is the busiest starting and ending station
- Which is the busiest route
- Relation between passholder type and duration
- Which passholder type has done more trips
- Breaking down the monthly usage of bikes
- Relation between trip route category and passholder type
- Which were the bikes that had logged maximum hours
- Determining the distance travelled and subsequent pricing
- Regression analysis and Hypothesis testing

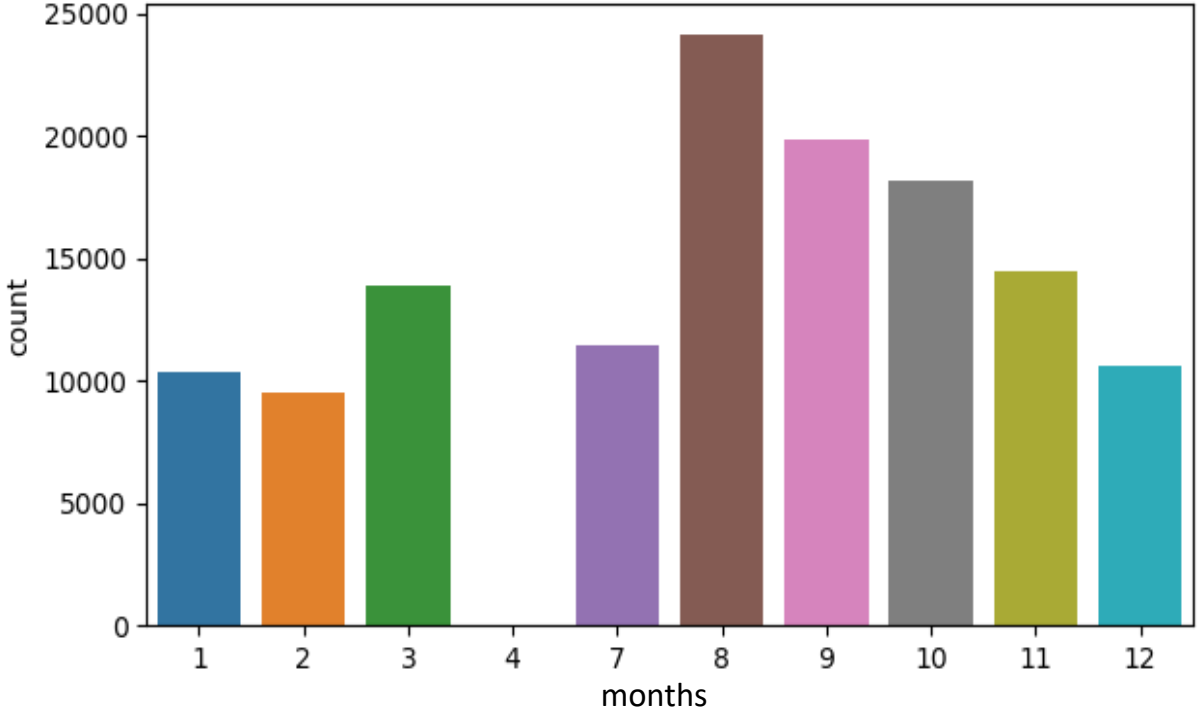
Data Visualization

Important aspect of EDA. Enables decision makers to see analytics presented visually, so they can grasp difficult concepts or identify new patterns. Few graphs along with their observations are shown below.

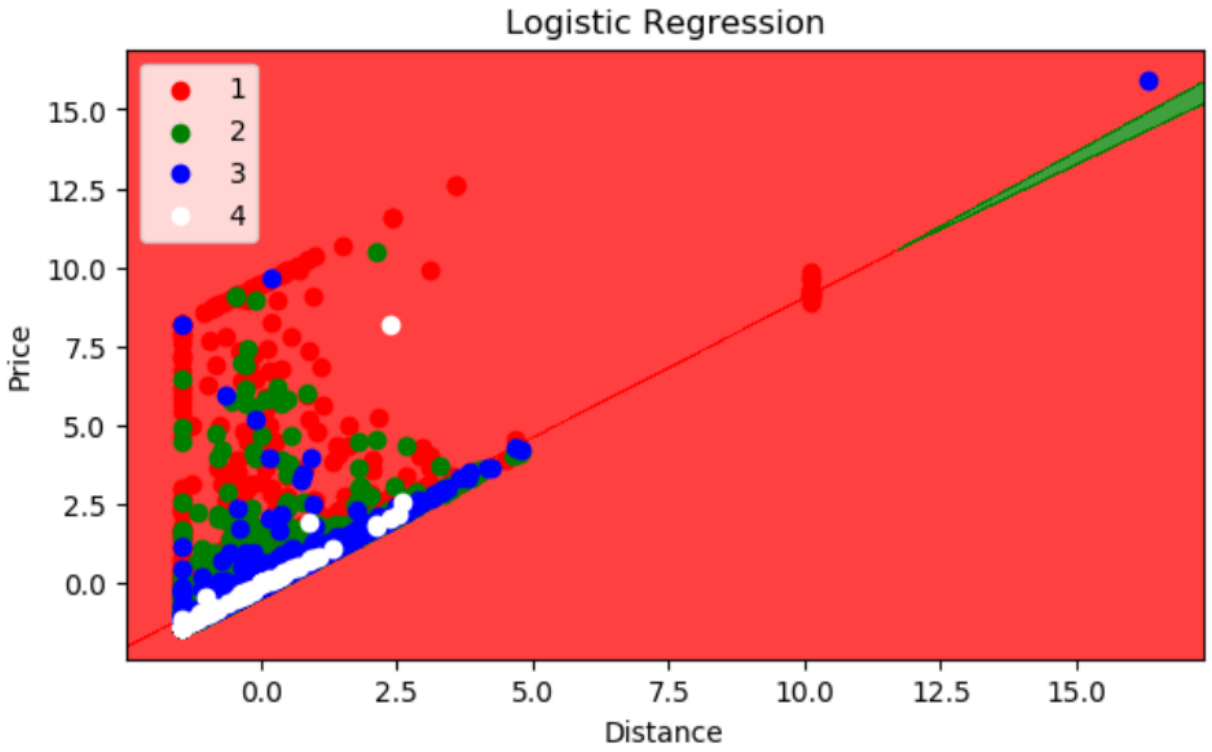
From the graph below, **monthly pass** looks to be the popular choice among the users



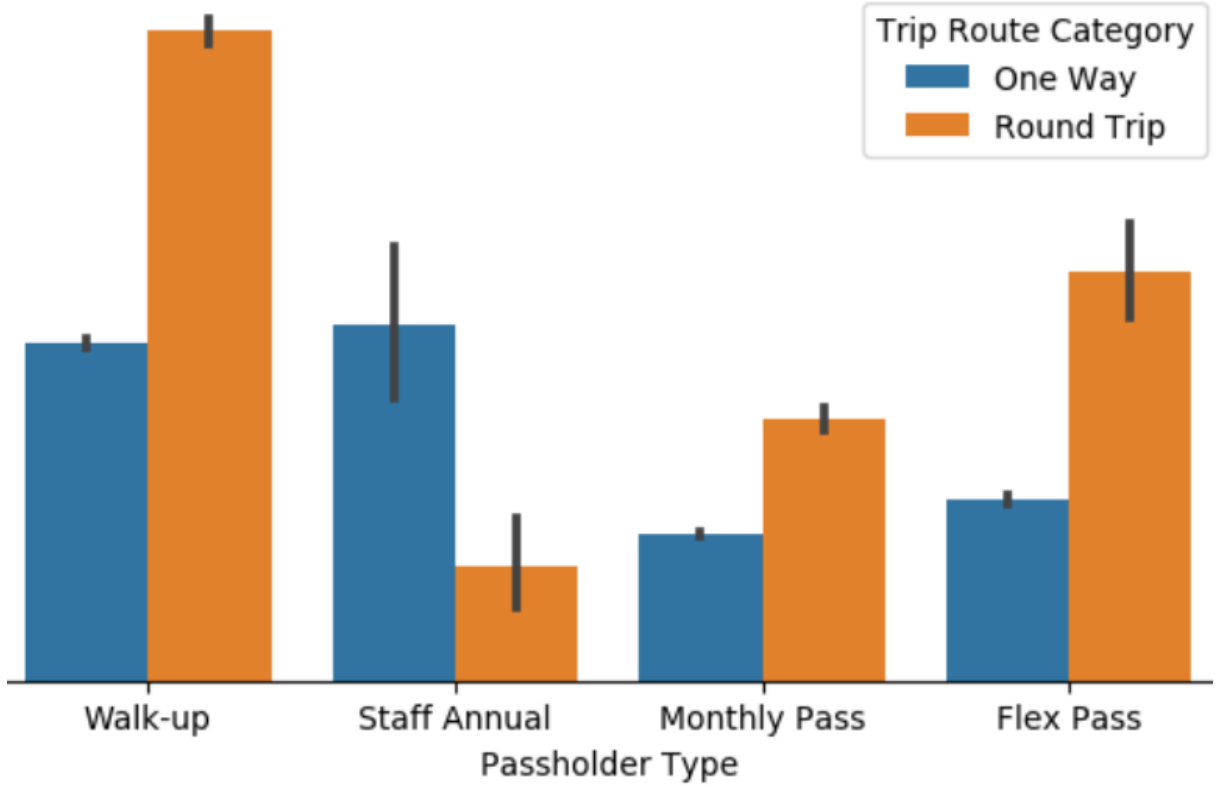
For examining the monthly usage of the bikes, the below graph plotted shows that **August** seems to the busiest month



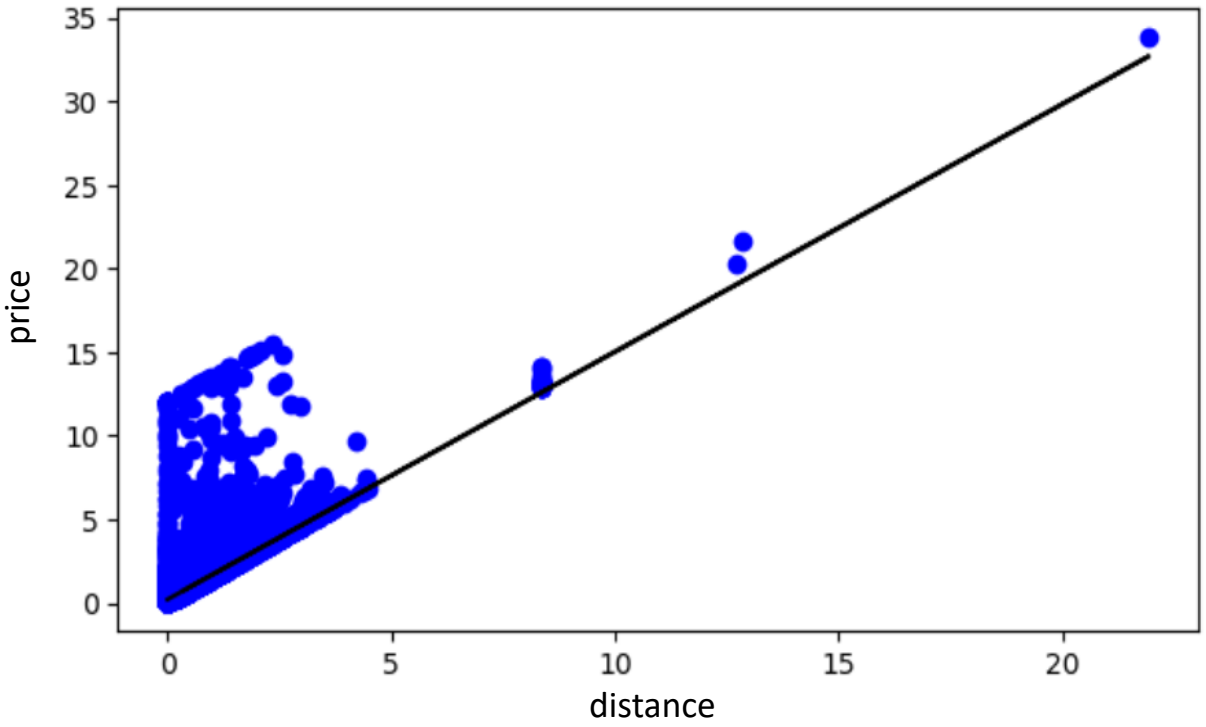
Logistic Regression was performed between Distance and Price data with Passholder Type as the categorical variable. An accuracy of **66%** was obtained



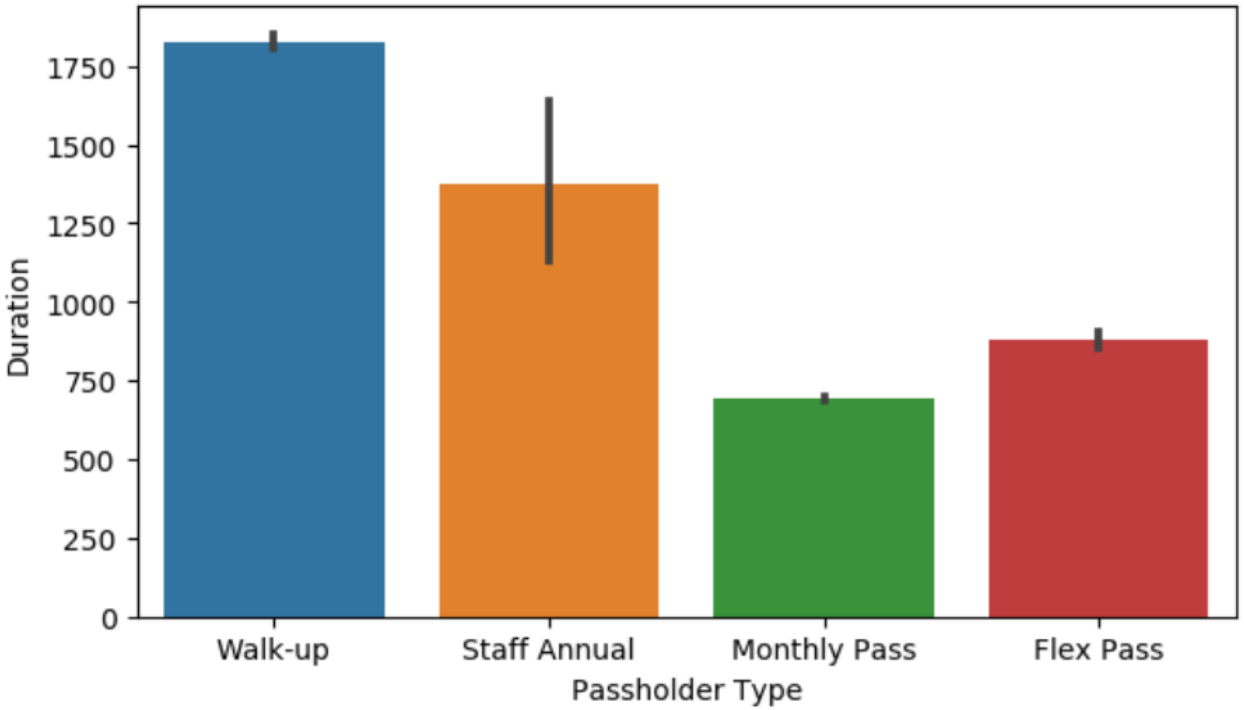
Below graph depicts the relation between Trip Route Category and Passholder Type. As evident, **Round Trips** are more than One way Trips



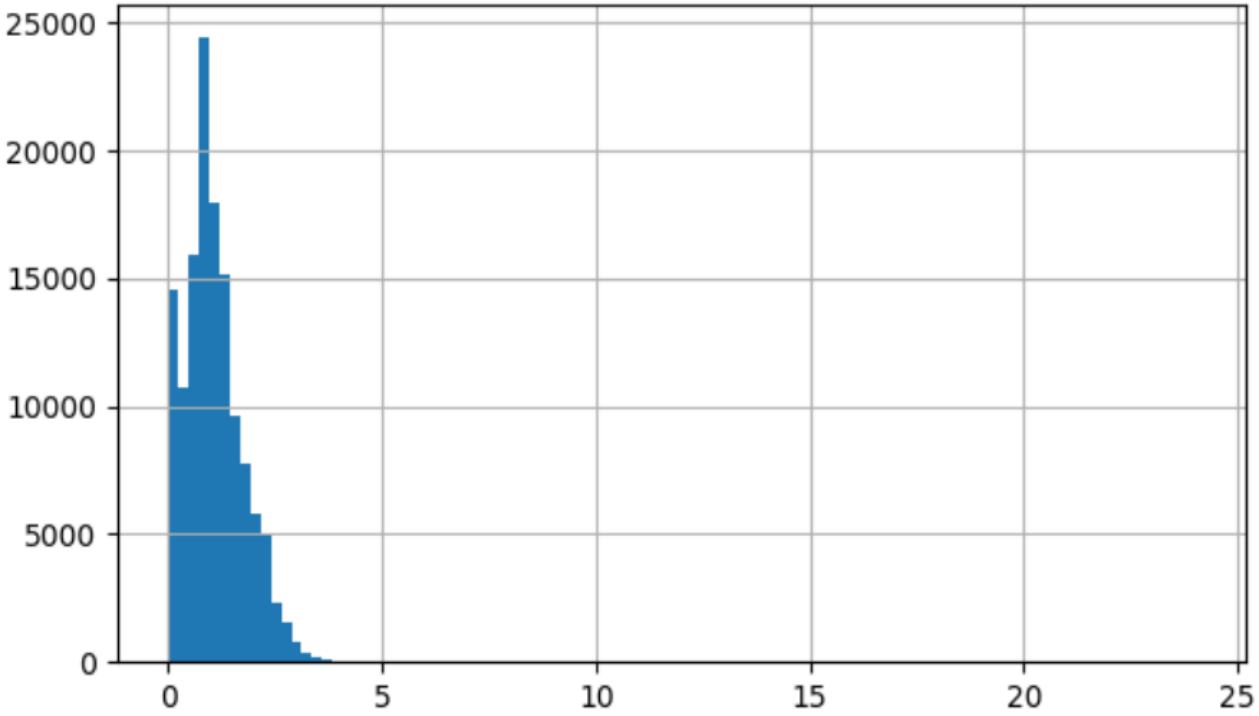
Linear Regression between Distance and Price data was implemented. An accuracy of **74%** was obtained



Graph illustrating the relation between Duration and Passholder Type. **Walk up** passholders tend to ride for longer durations



Histogram of the distance travelled by the users, calculated from the location coordinates using the Haversine Formula



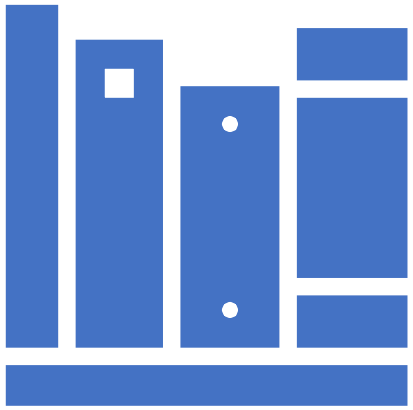
Output for the busiest route is depicted below. This implies that station IDs **3030** and **3014** are the busiest with 931 trips between the two

```
concat.value_counts().head()

3014.03030.0    931
3030.03014.0    677
3005.03031.0    610
3048.03048.0    568
3031.03005.0    513
dtype: int64
```

Conclusion

Data Cleaning and Visual Data Representation represent the main aspects of data analysis. Uncleaned data may give wrong conclusions. It is also easier and faster to analyse from pictorial data.




Project Presentation End

Thank You