PES UNIVERSITY, Bangalore

(Established under Karnataka Act No. 16 of 2013)

UE18CS203

B.Tech, Sem III
Session: Aug-Dec, 2019

UE18CS203 – INTRODUCTION TO DATA SCIENCE

REPORT
ON
EXPLORATORY ANALYSIS ON LOS ANGELES METRO BIKE SHARE TRIP DATA

SECTION:

#	SRN	Name	Contact No.	Email ID	Sign
1	PES1201800395	G Deepank	8217582948	deepankgrandslam@gmail.com	
2	PES1201801672	N Pruthvi	9686595359	pruthviniranjan@gmail.com	

ABOUT THE DATA SET

We were assigned the Los Angeles Metro Bike Share Trip Data for exploratory data analysis. The data gives us information regarding which bikes were used, locations of start and end stations, and the length of each ride. This is hosted by the city of Los Angeles. The organization has an open data platform and they update their information according to the amount of data that is brought in. The dataset had initially 22 columns and 132427 rows with a size of 33 MB.

ABSTRACT

The purpose of this assignment was to increase our awareness, fervor and encourage exploration of fields and topics which are overlapping with the fields of Data Science. Its main focus was to augment the theoretical learning of the class with invaluable hands-on knowledge on the topic of data analysis. This entailed us in making the most of the knowledge acquired from the class sessions to plot visualizations which would help us in understanding the dataset better, and derive insights from it. LA Metro Bike Share Trip Data provides the trip duration and origin/destination location, types of passes sold i.e. annual flex or monthly passes for each ride. Data were pre-processed by removing columns with many missing values. A few insights that we contemplated were: “Which is the busiest starting and ending station?” “Which is the busiest route?” “Relation between Passholder Type and Duration?” “What was the monthly usage of bikes?” “Relation between Trip Route Category and Passholder Type?” “Which were the bikes that had logged maximum hours?”. We were able to come up with conclusive answers for some of these questions, supported by visual representations.

EXPLORATORY ANALYSIS

Data Cleaning

The dataset provided to us was in most parts clean, and many columns with integer values (trip, station IDs, duration) were cleaned of any nuances (special characters, tabs, heading/trailing/multiple spaces, among others) using the regex functionality. Columns with missing values greater than 40% were dropped. Rows with missing values were also dropped instead of replacing them with mean/median, as they were comparatively very small compared to the total no. of rows. Outliers were found based on z-score using the ‘SciPy.Stats’ library. Values with z-score greater than 3 were labeled as outliers and hence dropped. Most of the data cleaning was performed while delving deeper into the insights formed from the data.

Data visualization

Data visualization describes the presentation of abstract information in graphical form. Data visualization allows us to spot patterns, trends, and correlations that otherwise might go unnoticed in traditional reports, tables, or spreadsheets. Descriptive Analysis was carried out to draw meaningful insights. We have used a few plots from our project in this section to illustrate the visual analysis. Conclusions and inferences about the graphs have been mentioned later.

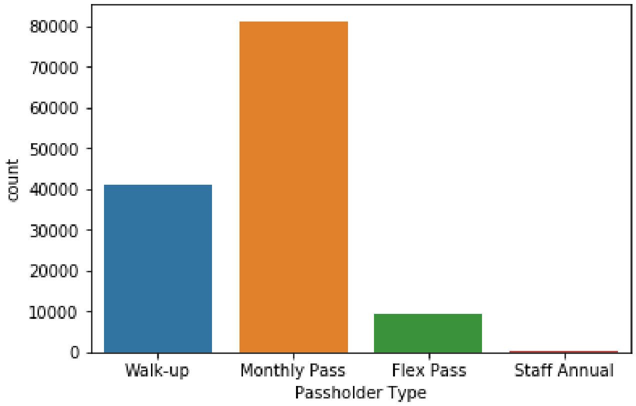
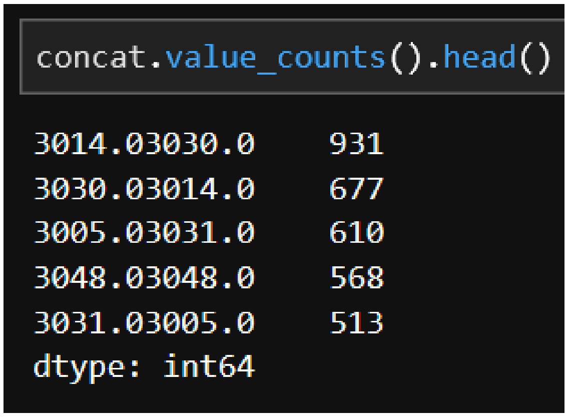


Fig. 1



```
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
```

Fig. 2

"*Which passholder type has done more trips*". We have plotted a count plot using the ‘Seaborn’ library in Fig. 1. From the graph, it is evident that monthly pass holders have done the most trips. They are almost twice the no. of walk-up riders. This gives us a perception of its popularity among the locals.

"Which is the busiest route". In Fig. 2 we have used a series of python string and mapping operations to produce the output depicted above. We can infer that starting station ID 3030 and ending station ID 3014 is the busiest route with over 931 trips between the 2 stations. This result can be used to solve inventory problems and enhance user experience.

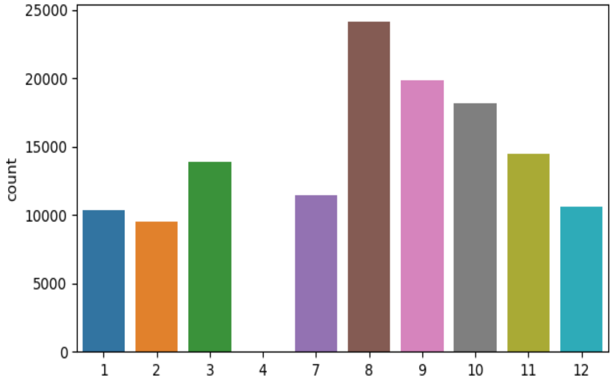


Fig. 3

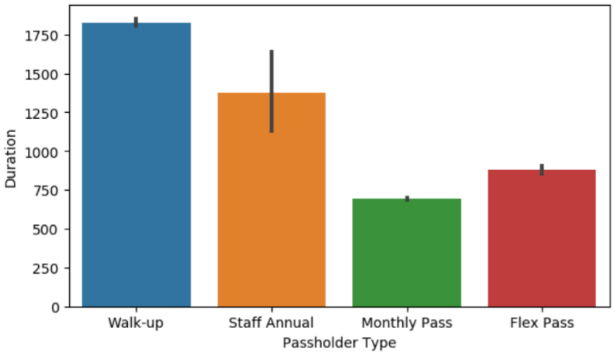


Fig. 4

"*What is the monthly usage of bikes*". Fig. 3 is a count plot showing the monthly usage of bikes. It can be observed that August is the busiest month of the year. A possible explanation for the spike in the rides during Q3 and Q4 could be due to the fact that it marks the beginning of festive periods and quarter cycles in companies. Due to a lack of data from Q2 (from both years) in the dataset, the graph shows almost no rides between April – June.

"*What is the relation between passholder type and duration*". From Fig. 4, it is clear that walk-up riders travel for longer durations compared to monthly and flex pass holders. This trend could be due to the fact that a large percentage of walk-up users maybe downtown Los Angeles visitors or tourists who are not price-sensitive. Monthly pass riders might be occasional commuters.

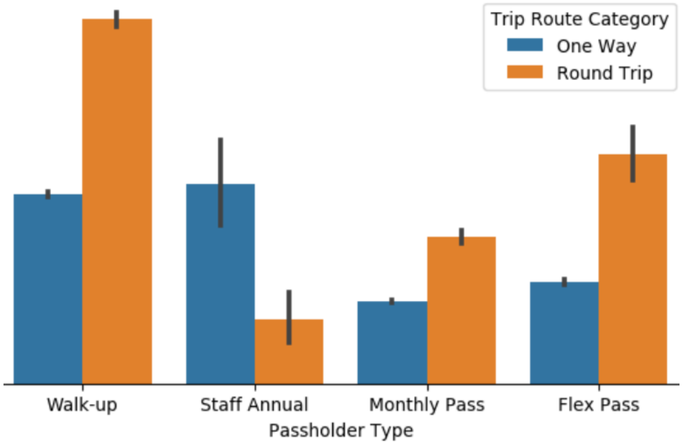


Fig. 5

"*Relation between Passholder Type and Trip Route Category*". Fig. 5 is a grouped bar plot illustrating the relation between the passholder type and the trip route category. Visibly, round trips are more than one-way trips. Such a pattern may be due to round trips being more economical than one-way trips and offering riders a convenient mode of commutation between workplaces.

Hypothesis Testing

The goal of hypothesis testing is to establish the viability of a hypothesis about a parameter of the population (often the mean). We have performed hypothesis testing in our presentation on the price value using the 'statsmodels.api' library.

Null Hypothesis: LA Metro Bike Service believes that about 4% of their users ride the bike more than 3.5 m/s (assume this as more of a threshold value).

Alternate Hypothesis: More than 4% of the users ride the bike greater than 3.5 m/s.

We attained a z-statistic value of **-2.284** and a p-value of **0.988**. Since the p-value is greater than the significance level, we fail to reject the null hypothesis.

CONCLUSION

It was the first time that we worked with a dataset as large as the one assigned to us. We spent a good amount of time cleaning the dataset. Having cleaned the data, we moved on to plotting relevant graphs for the same. We came to the conclusion that most of the riders had monthly passes. They also spent the least duration riding and there were more round trips than one-way trips among all riders. August was the busiest month in terms of bike usage. Traveling from station 3030 to 3014 seemed to be the busiest route. We also noticed that distance and price data had a linear relationship and performing regression on those two variables gave us an accuracy score of 0.73. The key takeaway for us through this project is that we can now appreciate how affluent and meaningful visual representation is. We can also conclude that performing data analysis helped us in finding hidden trends and patterns present in the dataset and this gave us a completely different outlook about the data.