Los Angeles Bike Sharing Network

Author: Sathish Manthani

Date: 2/6/2021

Contents

Introduction

Image source: Jorge

Bike sharing is the shared use of a fleet of bicycles which provides users with on-demand access to bicycles for one-way (point-to-point) or roundtrip travel. Bikes can be e-bikes or manual.

Bike sharing programs are an elegantly simple answer to urban gridlock, air pollution, and healthier lifestyles. It sounds like a simple system but its not. Keeping hundreds of bikes tuned up at the right stations and making them available to users requires the help of technology.

Bike sharing networks are present in cities of all sizes, they allow everyone (including visitors) a fun, cheap and easy way to explore the city they are based in and have even become part of regular commuter routines. Bike in itself doesn't use a lot of tech, however, technology plays a major role in the rapid expansion of the bike sharing networks. Technology helps in building the

smart bike sharing networks. The bike sharing companies use GPS sensors to track the bikes, and smartphones. They also keep the credit card on the file till the bike is returned (same like rental car) and to penalize the customer if the wheels of the bike go missing. Riders, meanwhile, can use apps to track down available rides or bike-share stations when they need them

In this project, I will be exploring the various features of the bike sharing network and the relationships among them. I would then predict the number of trips from each station. I will split the dataset into two parts, one is training dataset and the other one is testing dataset. I will use the training dataset to build the model. I will use decision tree and other algorithms for modeling. I will use the test dataset to test the accuracy of the target variable and conclude the best model based on the accuracy.

Research Questions

Machine learning is key in building smart bike sharing network. Data collection and analysis also plays vital role. I'm hoping this data can provide us more insights and visualizations into bike sharing network.

Below are my research questions for this project:

1. I want to analyze the spread of the bike stations across the city. How closely are they located.
2. How the ridership changed over the years. Did the number of riders increase over time?
3. Which day of the week the riders preferred the most for bike sharing. How is the usage during weekend?
4. Also, the usage of the bikes throughout the day. Is morning time preferred more by riders than evening time?

5. How the ride duration varied among riders
6. How the ride duration varied by membership type
7. What are the top 5 busiest stations in LA metro
8. Do the trend analysis of the trips throughout the year
9. Is there any impact on bike rentals due to holidays
10. What are the features influencing the trip counts most?
11. Can you build a model to predict the number of trips from a given station and test its accuracy.

Data Sources

The datasets are provided by the LA-Metro Bike Share Network.

https://bikeshare.metro.net/about/data/

I downloaded Trips data for 4 quarters of the year 2020 and the latest stations data.

Features available in the Trip dataset are:

Feature name	Description
trip_id	Locally unique integer that identifies the trip
duration	Length of trip in minutes
start_time	The date/time when the trip began, presented in ISO 8601 format in local time
end_time	The date/time when the trip ended, presented in ISO 8601 format in local time
start_station	The station ID where the trip originated (for station name and more information on each station see the Station Table)
start_lat	The latitude of the station where the trip originated
start_lon	The longitude of the station where the trip originated
end_station	The station ID where the trip terminated (for station name and more information on each station see the Station Table)
end_lat	The latitude of the station where the trip terminated
end_lon	The longitude of the station where the trip terminated

bike_id	Locally unique integer that identifies the bike
plan_duration	The number of days that the plan the passholder is using entitles them to ride; 0 is used for a single ride plan (Walk-up)
trip_route_category	"Round Trip" for trips starting and ending at the same station or "One Way" for all other trips
passholder_type	The name of the passholder's plan
bike_type	The kind of bike used on the trip, including standard pedal-powered bikes, electric assist bikes, or smart bikes.

Features in the station dataset are:

Feature name	Description
Station ID	Unique integer that identifies the station (this is the same ID used in the Trips and Station Status data)
Station name	The public name of the station. "Virtual Station" is used by staff to check in or check out a bike remotely for a special event or in a situation in which a bike could not otherwise be checked in or out to a station.
Go live date	The date that the station was first available
Region	The municipality or area where a station is located, includes DTLA (Downtown LA), Pasadena, Port of LA, Venice
Status	"Active" for stations available or "Inactive" for stations that are not available as of the latest update

Date dimension data is taken from

https://data.world/cegomez22/dimdate

This is a typical date dimension data where it has fields like date, week of day, month, year etc.,

Exploratory Analysis

The first step is to analyze the dataset by looking at few sample records.

I got bike sharing data for 4 different files for each quarter in year 2020. I concatenated the datasets using pandas and created a single dataframe.

Bike trips dataset:

	trip_id	duration	start_time	end_time	start_station	start_lat	start_lon	end_station	end_lat	end_lon	bike_id	plan_duration	trip_route_cate
0	148179433	14	10/1/2020 0:06	10/1/2020 0:20	3042	34.049301	-118.238808	3074	34.044170	-118.261169	6378	30	One
1	148179933	18	10/1/2020 0:25	10/1/2020 0:43	4404	34.048130	-118.271027	4444	34.061619	-118.305573	12440	30	One
2	148182639	141	10/1/2020 0:29	10/1/2020 2:50	4482	34.094372	-118.331009	4482	34.094372	-118.331009	19803	1	Round
3	148182739	141	10/1/2020 0:29	10/1/2020 2:50	4482	34.094372	-118.331009	4482	34.094372	-118.331009	18915	1	Round
4	148180233	15	10/1/2020 0:42	10/1/2020 0:57	3074	34.044170	-118.261169	3042	34.049301	-118.238808	20062	30	One

Stations dataset:

	Station_ID	Station_Name	Go_live_date	Region	Status
0	3000	Virtual Station	7/7/2016	NaN	Active
1	3005	7th & Flower	7/7/2016	DTLA	Active
2	3006	Olive & 8th	7/7/2016	DTLA	Active
3	3007	5th & Grand	7/7/2016	DTLA	Active
4	3008	Figueroa & 9th	7/7/2016	DTLA	Active

Next, I merged bike trips data with stations data.

I looked at the descriptive statistics of the dataset and is shown below

	trip_id	duration	start_station	start_lat	start_lon	end_station	end_lat	end_lon	plan_duration
count	2.099740e+05	209974.000000	209974.000000	165305.000000	165305.000000	209974.000000	201033.000000	201033.000000	209974.000000
mean	1.426180e+08	38.773396	3536.158472	34.052192	-118.257198	3670.051454	34.050258	-118.270163	56.332170
std	4.884567e+06	128.805149	661.597057	0.312124	2.236606	676.529117	0.248230	1.774130	105.574509
min	1.348675e+08	1.000000	3000.000000	33.928459	-118.491341	3000.000000	33.928459	-118.491341	1.000000
25%	1.381835e+08	7.000000	3006.000000	34.040600	-118.290092	3033.000000	34.039982	-118.291443	1.000000
50%	1.422025e+08	15.000000	3054.000000	34.048401	-118.260948	3076.000000	34.048038	-118.260948	30.000000
75%	1.466776e+08	27.000000	4335.000000	34.056610	-118.252441	4390.000000	34.056610	-118.252441	30.000000
max	1.517697e+08	1440.000000	4583.000000	55.705528	37.606541	4583.000000	55.705528	37.606541	999.000000

I also converted start_time and end_time columns to Date datatype from Object datatype.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 209974 entries, 0 to 209973
Data columns (total 15 columns):
trip_id                 209974 non-null int64
duration                209974 non-null int64
start_time              209974 non-null datetime64[ns]
end_time                209974 non-null datetime64[ns]
start_station           209974 non-null int64
start_lat               165305 non-null float64
start_lon               165305 non-null float64
end_station             209974 non-null int64
end_lat                 201033 non-null float64
end_lon                 201033 non-null float64
bike_id                 209974 non-null object
plan_duration           209974 non-null int64
trip_route_category     209974 non-null object
passholder_type         205248 non-null object
bike_type               209974 non-null object
dtypes: datetime64[ns](2), float64(4), int64(5), object(4)
memory usage: 24.0+ MB
```

Correlation matrix

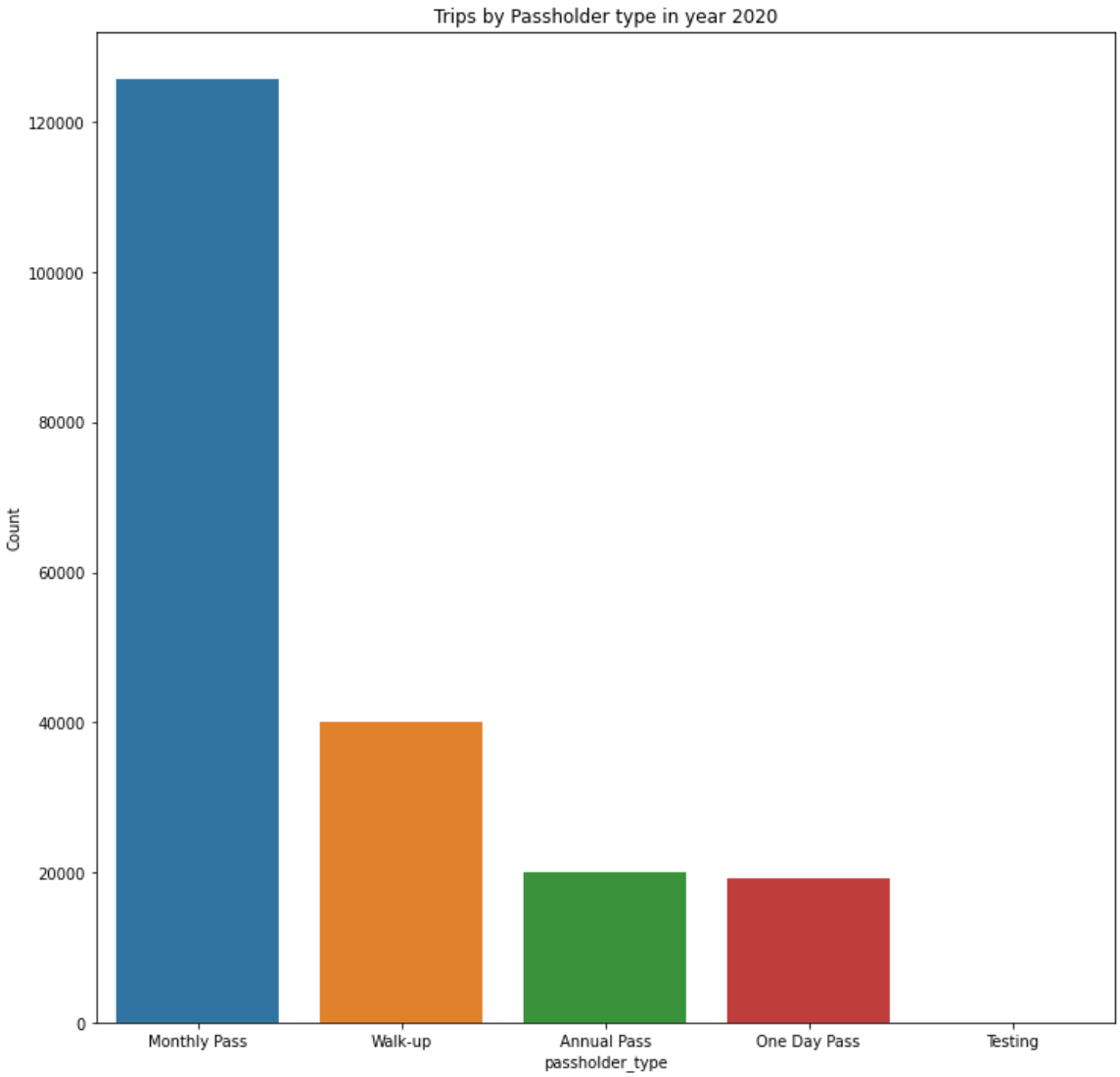
I drew the correlation matrix among the variables to see what features are correlated.

I see the correlation is strong between end_lat and end_long. Also, start_lat and start_lon. But that is expected since they are part of geographical coordinates. We need to dig further to derive more insights.

Exploratory Data Analysis

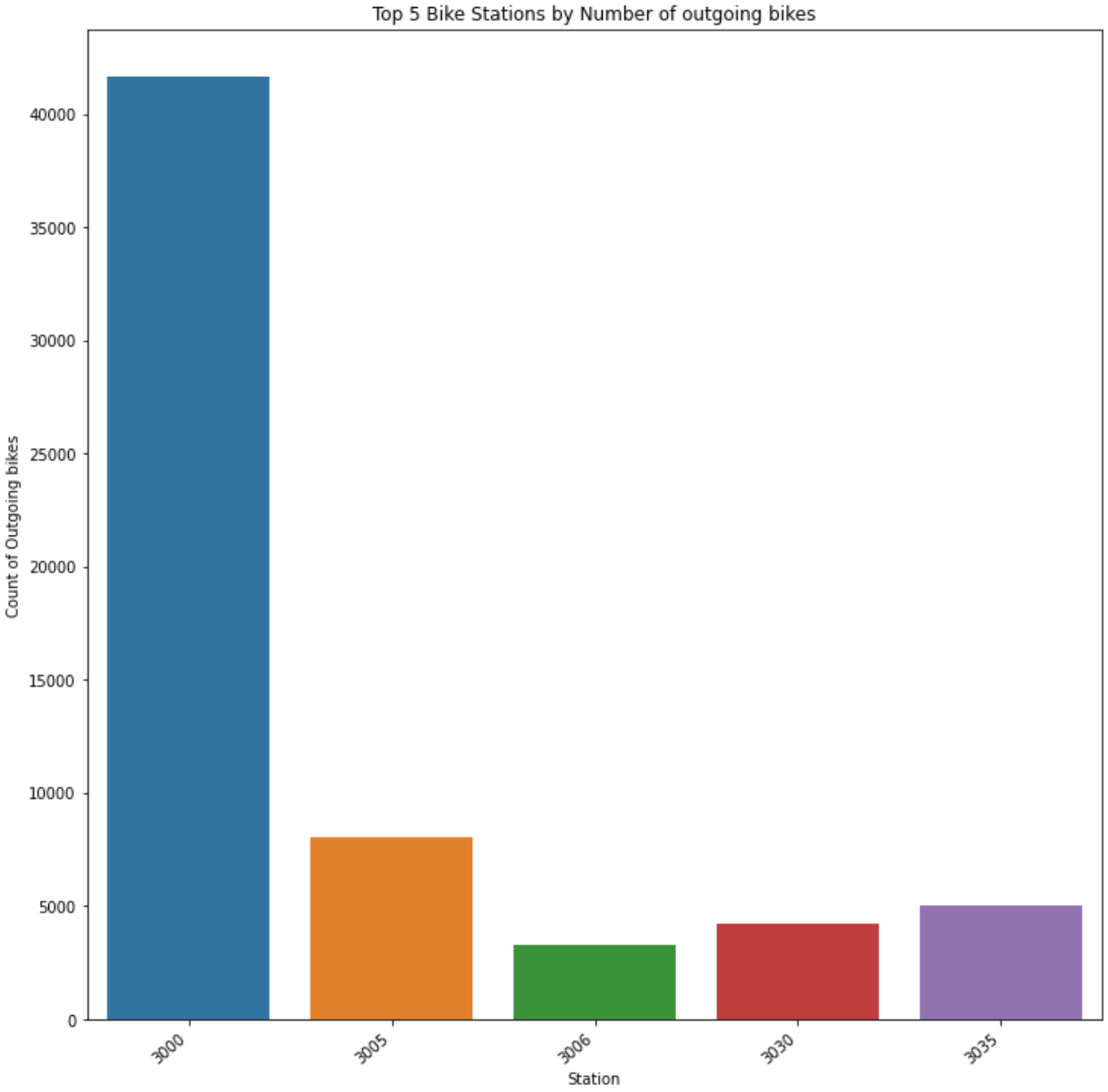
I did exploratory analysis on the data and transformed some of the features along the way.

As part of my research questions, I wanted to see the passholder type of the users in 2020. Below is the count of trips made by the users and their passholder type.



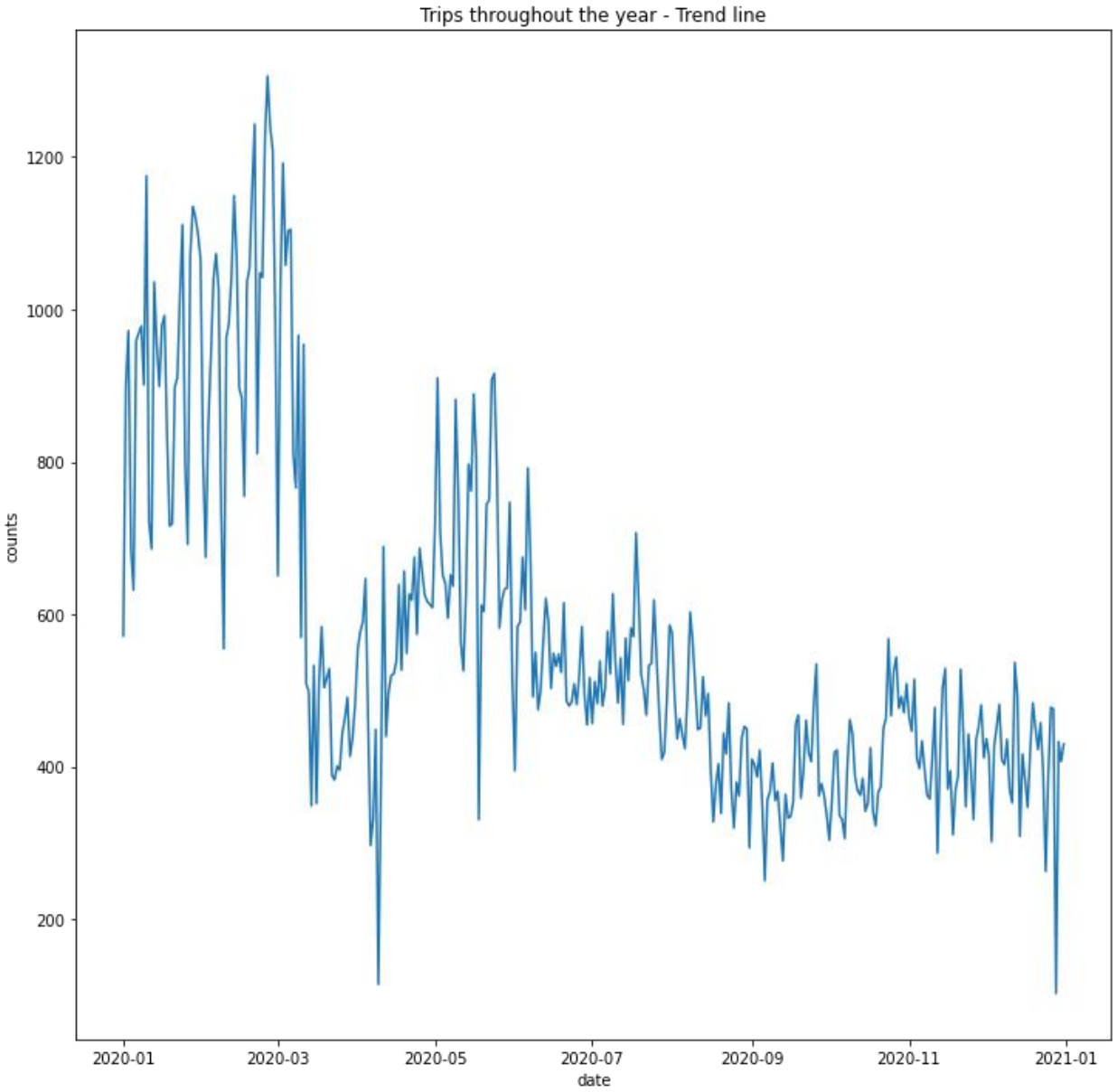
As you can see, users with monthly pass made the most trips. Its interesting to note that users are interested more in possessing monthly pass compared to annual pass.

Next I drew the chart for top 5 bike stations with highest number of outgoing bikes.



One particular station (station id: 3000) is has highest number of outgoing bikes compared to all others.

I then drew trend line for Bike trips throughout the year.



From above trend line, you can see that the trips were more in Feb-March timeframe and then during summer. Bike sharing demand has dropped during winter holidays.

After the exploratory data analysis and data preparatory steps, I transformed the data to be able to build predictive model for the question – predicting the count of outgoing bikes for each station.

Test/Train data split

After exploratory analysis and data transformation, we also have to split the dataset into training and test sets. Where I will use Training dataset to build the machine learning model and test dataset to test the accuracy of the machine learning model.

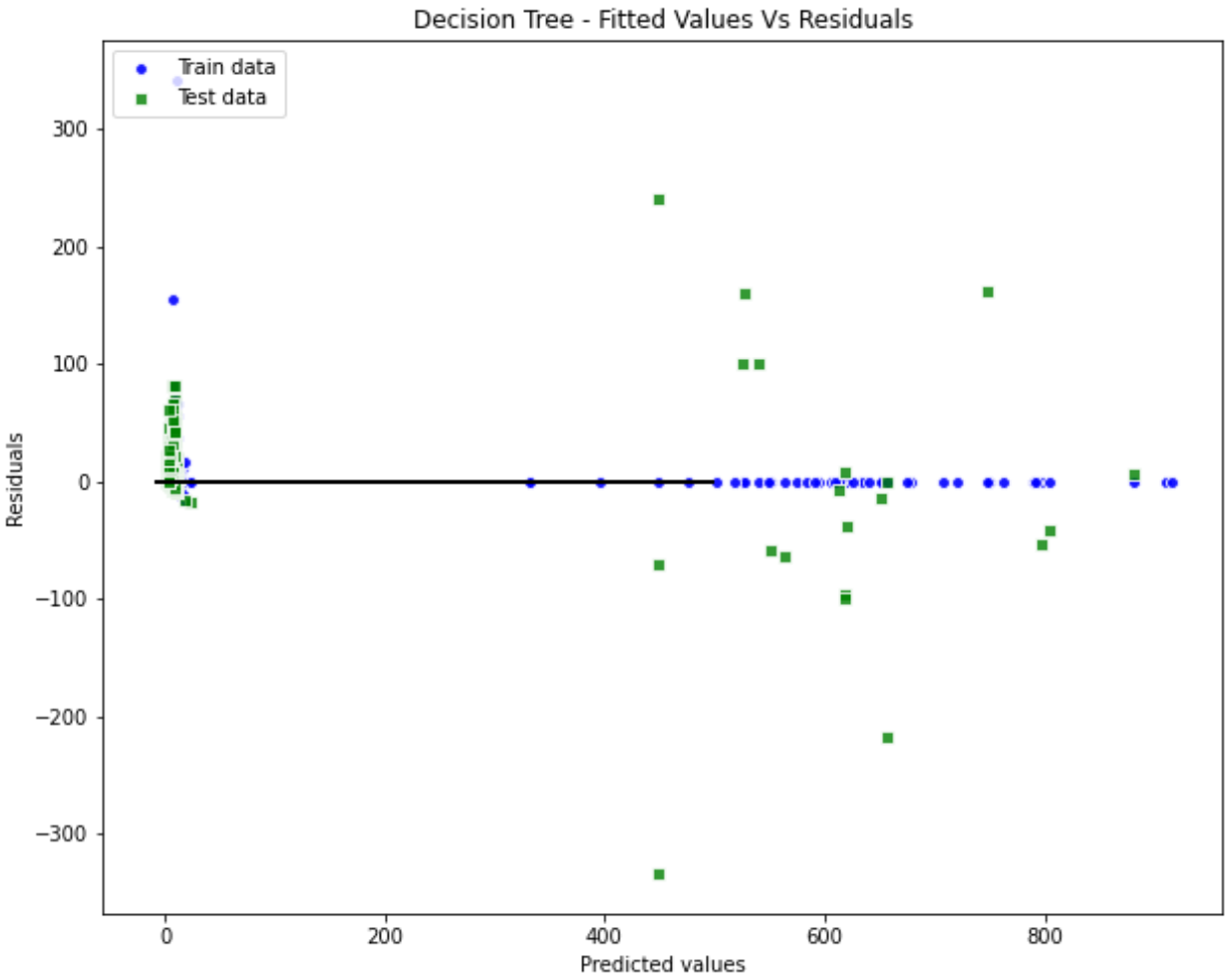
```
from sklearn.model_selection import train_test_split
# Split into train and test sets
X_train, X_test, y_train, y_test = train_test_split(features_df,
                                    target_df,
                                    test_size=0.30,
                                    random_state=42)
```

I split the given dataset into 2 parts. 70% of it into train data and 30% into test data.

Modeling and evaluation

The dependent variable (outgoing bikes count) is numerical and continuous in nature. And it is also non-linear in nature in relation to the predictors, so it is better to go with non-linear regression analysis. After data preparation activities including data cleansing, merging and feature engineering - I have tried Decision tree modeling and Random forest modeling.

Since I'm dealing with regression problem, I have used Mean Absolute Value, Mean Square Value to check the error in predictions based on both training and test dataset. R-square also gives us the variance around the mean. I'm also planning to plot residuals vs predicted values to see whether there is any pattern in the residuals.



Decision Tree gave me decent accuracy when compared with Test data with R-squared value of 91%. R-squared value basically tells us how well the fit line explains the data.

I also tried Random forest regressor, the fitted model gave me R-square value of 94% for training dataset and 92% with test dataset.

Summary/Conclusion

The goal of this project is to predict the count of outgoing bikes from each station while also exploring the dataset to find valuable insights. I started with data extraction using pandas library and did the exploratory analysis on the data. I got the datasets for each quarter separately, so I

merged them into one dataset. After looking at the descriptive statistics, I felt the need to bring the data values in few features to uniform range – so I used Standard Scaler to do the scaling.

The dependent variable (outgoing bikes) is not a variable that's available in the dataset readily, so I created it using transformation. Later, I split the dataset into two parts – Training set and Test Set. I used the training set to build the machine learning model. I tried Decision tree model and random forest models. Later, I evaluated each of the models using different evaluation techniques. I got decent accuracy and R-square using Random forest regression model.

References

[1] LA metro bike share network:

https://bikeshare.metro.net/about/

[2] Dataset codebook for LA metro bike share:

https://bikeshare.metro.net/about/data/

[3] Bike sharing system:

https://en.wikipedia.org/wiki/Bicycle-sharing_system

[4] What Is Bike Sharing? How Bike-Share Programs Work, Pros & Cons

https://www.moneycrashers.com/bike-sharing-best-bike-share-programs/

[5] Analysis of Network Structure of Urban Bike-Sharing System:

https://www.mdpi.com/2071-1050/11/19/5425/pdf

[6] LA Bike sharing programs: https://www.metro.net/projects/tod-toolkit/bike-share-programs/

[7] How bike-sharing works:

https://www.economist.com/the-economist-explains/2017/10/03/how-bike-sharing-works

[8] How Metro DC's Bikeshare System Works

https://www.capitalbikeshare.com/how-it-works

[9] How do bike sharing programs work?

https://www.quora.com/How-do-bike-sharing-programs-work

[10] Here’s what bike-sharing programs need to succeed

https://theconversation.com/heres-what-bike-sharing-programs-need-to-succeed-85969