Uber Analytic

Project Introduce:

The project is about on world's largest taxi company Uber inc. In this project, we're looking to predict the customer's behavior for their future transactional cases. Now it becomes important to manage their data properly to come up with new business ideas to get best results. Eventually, it becomes significant to estimate when and where to pick up and drop off the customer accurately.

What is the Uber?

Uber Technologies, Inc., commonly known as Uber, is an American technology company. Its services include ride-hailing, food delivery (Uber Eats and Postmates), package delivery, couriers, freight transportation, and, through a partnership with Lime, electric bicycle, and motorized scooter rental. The company is based in San Francisco and has operations in over 900 metropolitan areas worldwide. It is one of the largest firms in the gig economy. Uber is estimated to have over 93 million monthly active users worldwide. In the United States, Uber has a 71% market share for ride-sharing and a 22% market share for food delivery.

Problem statement:

Uber is a platform where those who drive and deliver can connect with riders, eaters, and restaurants. In cities where Uber is available, we will analyze the different time series, and average hours of working and growth of uber and will calculate the price of distance travel and will analyze different companies' growth with uber and check which one is best.

The primary methodology behind this study is to analyze and find the accuracy of the most frequent category of trip among all trips taken by a customer in a region using data analysis. Uber Data Analysis task permits us to recognize the complicated factual visualization of this large organization.

Dataset:

The dataset contains Start Date, End Date, Start Location, End Location, Miles Driven and Purpose of drive (Business, Personal, Meals, Errands, Meetings, Customer Support etc.).
Moreover, separating the 'START_DATE*', and 'END_DATE*' into five variables each (day, name of day, hour, minute, and week). In addition, including the 'time different' column so that we could know how much it took from location to destination.

Geography: USA, Sri Lanka and Pakistan

Time period: January - December 2016

Unit of analysis: Drives

Total Drives: 1,155

Total Miles: 12,204

Website: https://www.kaggle.com/datasets/zusmani/uberdrives

Inspiring:

- Which area is most popular place to pick up an drop off
- Total number of Uber pick up on each day
- The mean miles by day
- The mean time and distance to destination from most popular pick up location
- Get the federal holidays for the periods
- The effect of time on demand for Uber rides: distribution per hour and week

Data Preprocessing and Exploration

Figure 1. Missing value:

Find the missing value

```
[5]: df.isnull().sum()

[5]: START_DATE*      0
     END_DATE*        1
     CATEGORY*        1
     START*           1
     STOP*            1
     MILES*           0
     PURPOSE*       503
     dtype: int64
```

Firstly, after checking for missing observation, it is evident that the there is one row includes the missing values, and the 'PURPOSE*' includes too many missing values which is more than 30%.

Figure 2. Drop off the column and row:

```
[6]: # more than 30% of the data is missing
     # drop the 'purpose' columns

     df.drop(['PURPOSE*'], axis = 1, inplace = True)
     df
```

[6]:

	START_DATE*	END_DATE*	CATEGORY*	START*	STOP*	MILES*
0	1/1/2016 21:11	1/1/2016 21:17	Business	Fort Pierce	Fort Pierce	5.1
1	1/2/2016 1:25	1/2/2016 1:37	Business	Fort Pierce	Fort Pierce	5.0
2	1/2/2016 20:25	1/2/2016 20:38	Business	Fort Pierce	Fort Pierce	4.8
3	1/5/2016 17:31	1/5/2016 17:45	Business	Fort Pierce	Fort Pierce	4.7
4	1/6/2016 14:42	1/6/2016 15:49	Business	Fort Pierce	West Palm Beach	63.7
...	...	...	...	...	...	...
1151	12/31/2016 13:24	12/31/2016 13:42	Business	Kar?chi	Unknown Location	3.9
1152	12/31/2016 15:03	12/31/2016 15:38	Business	Unknown Location	Unknown Location	16.2
1153	12/31/2016 21:32	12/31/2016 21:50	Business	Katunayake	Gampaha	6.4
1154	12/31/2016 22:08	12/31/2016 23:51	Business	Gampaha	Ilukwatta	48.2
1155	Totals	NaN	NaN	NaN	NaN	12204.7

1156 rows × 6 columns

```
[7]: # drop the row that include 4 missing data.
     #df[df['END_DATE*'].isnull()]
     df.drop(labels = [1155], axis = 0, inplace = True)
```

Drop off the column call ‘PURPOSE*’, and row located at 1155.

Figure 3. check the final missing value:

```
[8]: # check there is any missing value left

     df.isnull().sum()

[8]: START_DATE*    0
     END_DATE*      0
     CATEGORY*      0
     START*         0
     STOP*          0
     MILES*         0
     dtype: int64
```

Figure 4. summary of MILES:

```
[9]: df.describe()
```

[9]:

	MILES*
count	1155.000000
mean	10.566840
std	21.579106
min	0.500000
25%	2.900000
50%	6.000000
75%	10.400000
max	310.300000

The range of riding is between 0.5 miles (Minimum) to 310 miles (Maximum). The miles rightly skewed as we can see average (10.5) of miles is bigger than median (6.0). the Average of miles of ride is 10.5 miles

Analyzing Data

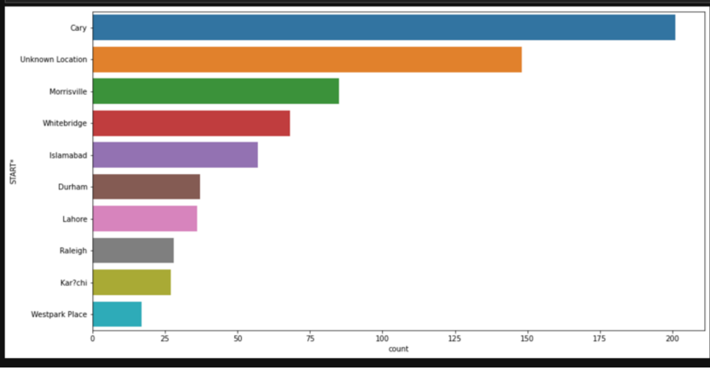
1. **Which area is most popular place to pick up and drop off?**

Figure 5. Top 10 places picking up and dropping off

```
[11]:
      df['START*'].value_counts().head(10)

[11]: Cary                201
      Unknown Location    148
      Morrisville          85
      Whitebridge          68
      Islamabad            57
      Durham               37
      Lahore               36
      Raleigh              28
      Kar?chi              27
      Apex                 17
      Name: START*, dtype: int64
```

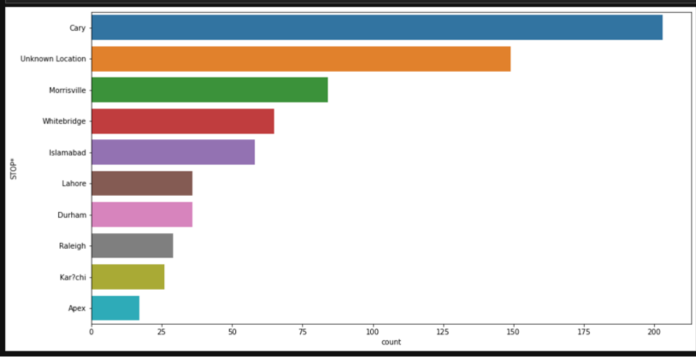
Picking places:

```
[16]: # top 10 places picking up

plt.figure(figsize=(15,8))
sns.countplot(y="START*",order= pd.value_counts(df['START*']).iloc[:10].index, data = df)
plt.show()
```



Dropping off places:

```
[17]: # top 10 places dropping off

plt.figure(figsize=(15,8))
sns.countplot(y="STOP*",order= pd.value_counts(df['STOP*']).iloc[:10].index, data = df)
plt.show()
```



Majority of people are picked up and dropped off at the ‘Cary’. Top 1 to 5 are same places

2. Total number of Uber pick up on each day

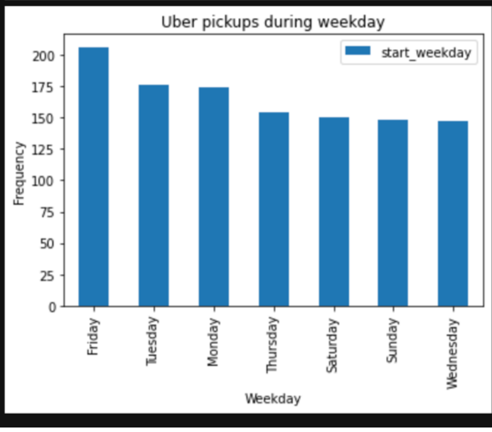
Figure 6. Uber pickup during weekday

```
[18]: week_day_fre = pd.DataFrame(df['start_weekday'].value_counts())
      week_day_fre
```

[18]:

	start_weekday
Friday	206
Tuesday	176
Monday	174
Thursday	154
Saturday	150
Sunday	148
Wednesday	147

```
[19]: week_day_fre.plot.bar()
      plt.title('Uber pickups during weekday')
      plt.xlabel('Weekday')
      plt.ylabel('Frequency')
      plt.show()
```



Friday has the maximum frequency of Uber pickups whereas Wednesday is the least busy.

3. The mean miles by day

Figure 7. Average Miles by day

```
[20]: print('Mean miles by day')
      week_day_mile = df[['start_weekday','MILES*']].groupby(['start_weekday']).mean()
      week_day_mile
```

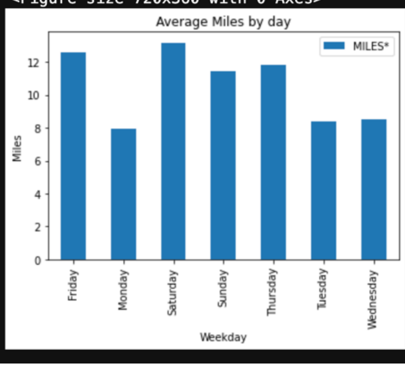
Mean miles by day

[20]:

start_weekday	MILES*
Friday	12.597087
Monday	7.973563
Saturday	13.175333
Sunday	11.462162
Thursday	11.805195
Tuesday	8.418750
Wednesday	8.502721

```
[21]: plt.figure(figsize=(10, 5))
      week_day_mile.plot.bar()
      plt.title('Average Miles by day')
      plt.xlabel('Weekday')
      plt.ylabel('Miles')
      plt.show()
```

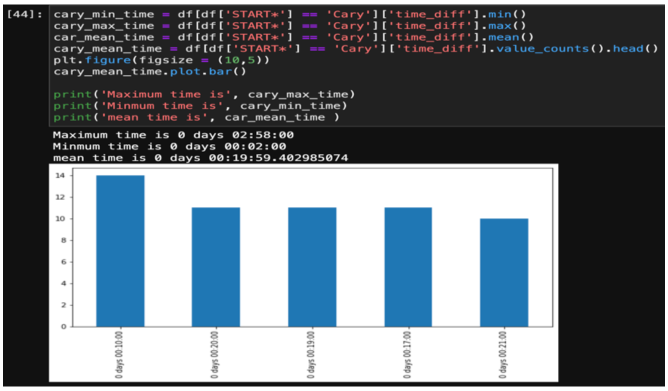
<Figure size 720x360 with 0 Axes>



Compared to other day, Friday and Saturday have longer trip than others.

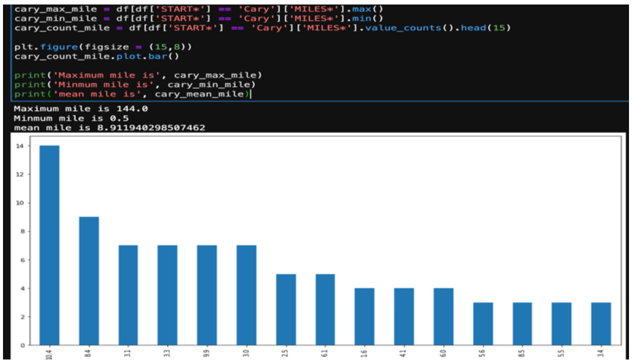
4. The mean time and distance to destination from most popular picking location

Figure 8. the mean time



the uber rider takes between 2 min and 2 hour 58 min from cary to destination. the average of riding is about 20 min. the most frequency occurrence is from 10 min to 20 min.

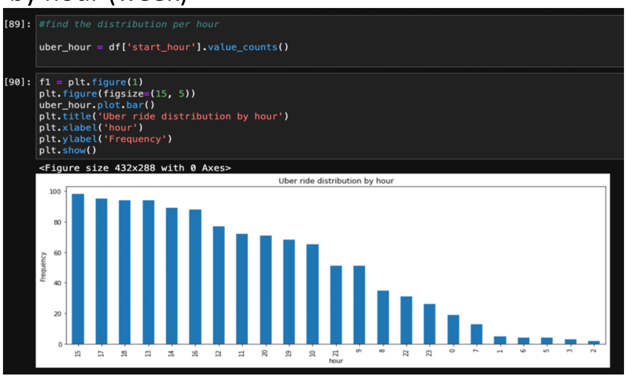
figure 9. mean miles from 'Cary'



the maximum distance from Cary is 144 miles, and on the other hand the minimum distance is 0.5 miles. The average distance is about nine miles, and the distance people often go can be seen as one to ten miles

5. The effect of time on demand for Uber rides: distribution per hour and week

Figure 10. Uber ride by hour (week)



Uber pickups tend to be maximum around 1-6 pm, when the most people actively move around. This trend can vary for weekends, thus separately checking for weekdays and weekends

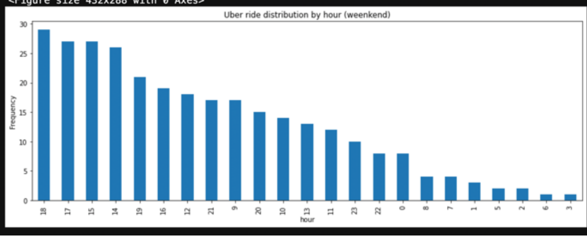
Figure 11. Uber ride by hour (weekend)

```
[91]: weekend=df[(df['start_weekday']=='Sunday') | (df['start_weekday']=='Saturday')]
weekend.head()
```

[91]:

	START_DATE*	END_DATE*	CATEGORY*	START*	STOP*	MILES*	start_date	start_weekday	start_hour	start_min	start_weekno	end_date	end_weekday	end_hour	end_min	end_weekno	time_diff
1	2016-01-02 01:25:00	2016-01-02 01:37:00	Business	Fort Pierce	Fort Pierce	5.0	2	Saturday	1	25	52	2	Saturday	1	37	52	0 days 00:12:00
2	2016-01-02 20:25:00	2016-01-02 20:38:00	Business	Fort Pierce	Fort Pierce	4.8	2	Saturday	20	25	52	2	Saturday	20	38	52	0 days 00:13:00
8	2016-01-10 08:05:00	2016-01-10 08:25:00	Business	Cary	Morrisville	8.3	10	Sunday	8	5	0	10	Sunday	8	25	0	0 days 00:20:00
9	2016-01-10 12:17:00	2016-01-10 12:44:00	Business	Jamaica	New York	16.5	10	Sunday	12	17	0	10	Sunday	12	44	0	0 days 00:27:00
10	2016-01-10 15:08:00	2016-01-10 15:51:00	Business	New York	Queens	10.8	10	Sunday	15	8	0	10	Sunday	15	51	0	0 days 00:43:00

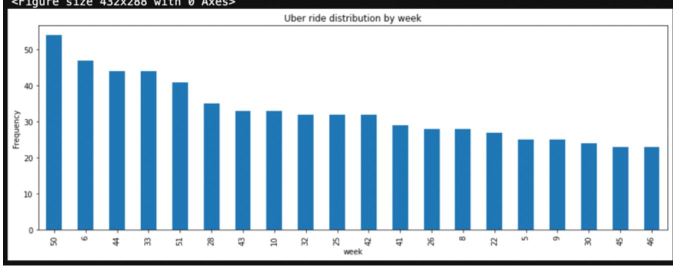
```
[97]: uber_weekend_hour = weekend['start_hour'].value_counts()

[100… f1 = plt.figure(1)
plt.figure(figsize=(15, 5))
uber_weekend_hour.plot.bar()
plt.title('Uber ride distribution by hour (weenkend)')
plt.xlabel('hour')
plt.ylabel('Frequency')
plt.show()

<Figure size 432x288 with 0 Axes>
```



compared to weekday, we can observe that the nighttime is more likely active than the day.

Figure 12. Uber ride by week

```
f2 = plt.figure(2)
plt.figure(figsize=(15, 5))
uber_weekday.plot.bar()
plt.title('Uber ride distribution by week')
plt.xlabel('week')
plt.ylabel('Frequency')
plt.show()

<Figure size 432x288 with 0 Axes>
```



You can see that a lot happened at the beginning of the year or at the end of the year