VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

SCHOOL OF ADVANCED SCIENCE - (SAS)

Winter Semester - 2021-2022

PREDICTION ON UBER RIDES

USING PYTHON

PROJECT REVIEW FINAL

Submitted fulfilment for the J-component of CSC5007

EXPLORATORY DATA ANALYSIS

CAL COURSE
In

MSC. DATASCIENCE

BY

DHIVYA V - 21MDT0037

GEETHA LAKSHMI B - 21MDT0057

SHAKTHIVEL D – 21MDT0070

SWATHI E – 21MDT0127

Under the guidance of

Prof. PALLAVI MISHRA

ABSTRACT

India is a developing country with more than 1.38 billion people. Majority on them are working and belong to middle class or upper middle class. It is also well-known fact that majority of the Indian women living in cities are working and contribute to the GDP of the country. The working women mostly depend on the public services finding it to be the safest and the cheapest mode of transportation to their work places. The introduction of "UBER DRIVES" services has replaced the usage of public transportation because of ease of booking, cheaper fares and safe mode of transport. Online taxi booking not only helps you with best prices but also helps you with the convenience of paying through multiple payment options (like Debit Card, Credit Card, E-Wallets etc.). In this work, dataset is collected from an internet source and cleaned. Using Python programming, data pre-processing is done where we identify, handle and treat the missing values from the taken dataset. This project is based on the trips made by Uber Drivers. Different Aspects of the trip are analysed by using different functions in Python and using Data Interpretation techniques. Then, we visualize a graphical representation, where insights behind our dataset on uber drives are explained with first and last 10 records of uber drives, information about the variables and the summary of the dataset.

Keywords: uber drives , missing values, visualization, bar graph and summary

INTRODUCTION

Uber is a technology company that connects riders and partner drivers using a smartphone application (**Uber** Partner App, **Uber** App, the App). . By using technology and focusing on safety and customer service, we aim to increase urban mobility, create economic opportunities and support Malaysian cities. . Your journey begins now.

GENERAL

Exploring data is certainly one of the most important stages in Data Science processes. Despite its simplicity, it can be a powerful tool to put you ahead on data and business context, as well as to determine crucial treatments before creating machine learning models. Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques which includes to maximize insight into a data set, to uncover underlying structure, to extract important variables, to detect outliers and anomalies, to test underlying assumptions, to develop parsimonious models and to determine optimal factor settings.

SCOPE

The obtained results give the purpose for which uber usage is maximum along with the counts and mileage covered.

OBJECTIVES

The main purpose of EDA is to help look at the dataset before making any assumptions. It can help identify obvious errors, as well as to understand patterns better within the data, detect outliers or anomalous events, find interesting relations among the variables through Visualization.

PROBLEM STATEMENTS

From the dataset taken, we need to detect missing values, to find the summary and other information on the dataset taken. To visualize the most frequent category of the trip , the total number of unique START and stop destination .The most popular starting and dropping point and the most frequent route taken by uber driver .

And to find the purpose of maximum number of trips taken according to the respective categories.

PROPOSED SYSTEM

In this work, the dataset is collected, cleaned and generated EDA reports, and visualized analysed the results using Python programming.

SOFTWARE REQUIREMENTS SPECIFICATIONS

REQUIREMENTS: ENVIRONMENTS AND TOOLS:

1. Pandas
2. Numpy
3. Seaborn
4. Matplotlib
5. Building

SOFTWARE CONFIGURATION:

Operating system – Window, Linux.

IDE: Jupyter Notebook

Coding Language: R programming

It is a web-based interactive development environment for notebooks, code, and data. It is flexible to configure and arrange the user interface to support a wide rangeof workflows in data science, scientific computing and machine learning. It exists to develop open-source software, open-standards, and services for interactive computing across dozens of programming languages.

EXISTING WORK:-

The Big 3 Cab Services In India:

1. **Savaari car rentals:**

   Savaari Car Rentals is an online cab booking aggregator that aims to provide affordable and safe taxi services to travelers. With operations across 60 cities in India, Savaari is uniquely placed as the largest car rental company in terms of geographical reach. Savaari provides competitive Airport transfers which includes toll, parking and waiting charges, cabs for outstation travel as well as intra-city local cabs.

2. **Ola Cabs**

OlaCabs is an online cab aggregator based out of Bengaluru and among the fastest growing taxi hiring firms. Taxi booking facility can be availed through app, website or through calls. It was founded on 3rd December 2010 by BhavishAggarwal (CEO) and AnkitBhati. By 2014, the company has expanded to a network of more than 18,000 cars across more than 65 cities. Today, Ola has more than 1,50,000 plus cabs registered on its platform and is present in more than 100 cities across the country . It claims to clock an average of more than 150,000 bookings per day and commands 60 percent of the market share in India.

3. **Meru Cabs**

Meru Cabs is a taxi aggregator company based in Mumbai, India. It provides cab booking facilities through calls, website or through their mobile application and payment through cash, card or wallet Meru Cabs integrated their cab service with Google Now which will send passengers remainders for cab pickups, alerting them if they wish to book a cab based on their location and other information through Now Cards within the Google app.

LITERATURE SURVEY

PAPER I

Authors: M. Wang and L. Mu

Article title: Spatial disparities of Uber accessibility: An exploratory analysis inAtlanta, USA

Journal: Computers, Environment and Urban Systems

Description

[1] Inequality of accessibility in transportation systems is a constant concern, whichis intensified by the transportation economization process and the digital divide. How should the accessibility of crowd sourced transportation be measured and understood? Without any prior assumption, this paper openly explores spatialdisparities of accessibility in the city of Atlanta, USA using both the UberX (the most popular Uber product) and the Uber BLACK (the premium Uber product) data. Accessibility is measured by both the expectation and variability of Uber wait time.With spatial autoregressive models, we find that after controlling for other socioeconomic factors, wealth and race do not have significant associations with Uber accessibility. Additionally, higher road network density, population density, and less commuting time to work correlate with greater Uber accessibility. More public transport stops are related to better accessibility of UberX but worse accessibility of Uber BLACK. Finally, implications for policy- makers are provided

PAPER-II

Author: A. R. L. John T. Behrens, Kristen E. Dicerbo, Nedim YelArticle title:
Exploratory data analysis
Journal: Handbook of Psychology
Description

[2] Exploratory data analysis (EDA) is a conceptual frame- work with a core set of ideas and values aimed at providing insight into data as it is presented to the workingresearcher (regardless of its origin), and to encourage understanding probabilistic.

And no probabilistic models in a way that guards against erroneous conclusions. Because this set of goals covers experimental and no experimental data, clean and messy data, and data in forms that may not be properly statistically modeled, Tukeydistinguished these goals from the more specific probabilistic goals of traditional "statis- tics," which he referred to as "confirmatory data analysis" (CDA). Clearly these practice-based and pragmatic goals are well aligned with the needs of active researchers in the psychological community (Behrens, 1997a).

PAPER-III

Author: J. Cramer and A. B. Krueger

Article title: Disruptive change in the taxi business: The case of uberJournal:
American Economic Review
Description

[3] The innovation of ride sharing services, such as Uber and Lyft, which use Internet-based mobile technology to match passengers and drivers, is providing unprecedented competition in the taxi industry. Weighted by hours worked, there were about half as many Uber and Lyft drivers as taxi and limo drivers operating inthe United States at the end of 2015.1 This paper examines the efficiency of the ridesharing and service Uber by comparing the capacity utilization rate of UberX driversto that of taxi drivers.

PAPER-IV

Authors: K. Abt

Article title: Descriptive data analysis: A concept between confirmatory andexploratory data analysis

Journal: Methods of Information in Medicine

Description

[4] Confirmatory Data Analysis (CDA) in randomized comparative ("controlled") studies with many variables and/or time points of interest finds its limitations in themultiplicity of desired inferential statements which leads to unfeasibly small adjusted significance levels ("Bonferronization") and, thereby, to unduly increased risks of not rejecting false hypotheses. In general, analytical models adequate for such complex data structures and suitable for practical use do not exist as yet. Exploratory Data Analysis (EDA), on the other hand, is usually intended to generatehypotheses and not to lead to final conclusions based on the results of the study. In this paper, it is proposed to fill the conceptual gap between CDA and EDA by "Descriptive Data Analysis" ("DDA") which concept is mainly based on descriptive inferential statements. The results of a DDA in a controlled study are interpreted simultaneously on the basis of the investigator's experience with respect to numerically relevant treatment effect differences and on "descriptive significances"as they appear in "near regular" patterns corresponding to the resulting relevant effect differences. A DDA may also contain confirmatory parts and/or tests on global hypotheses at a prechosen maximum risk a of erroneously rejecting true hypotheses.The paper is in parts expository and is addressed to investigators as well asstatisticians

SURVEY OF THE EXISTING MODEL/WORKS:-

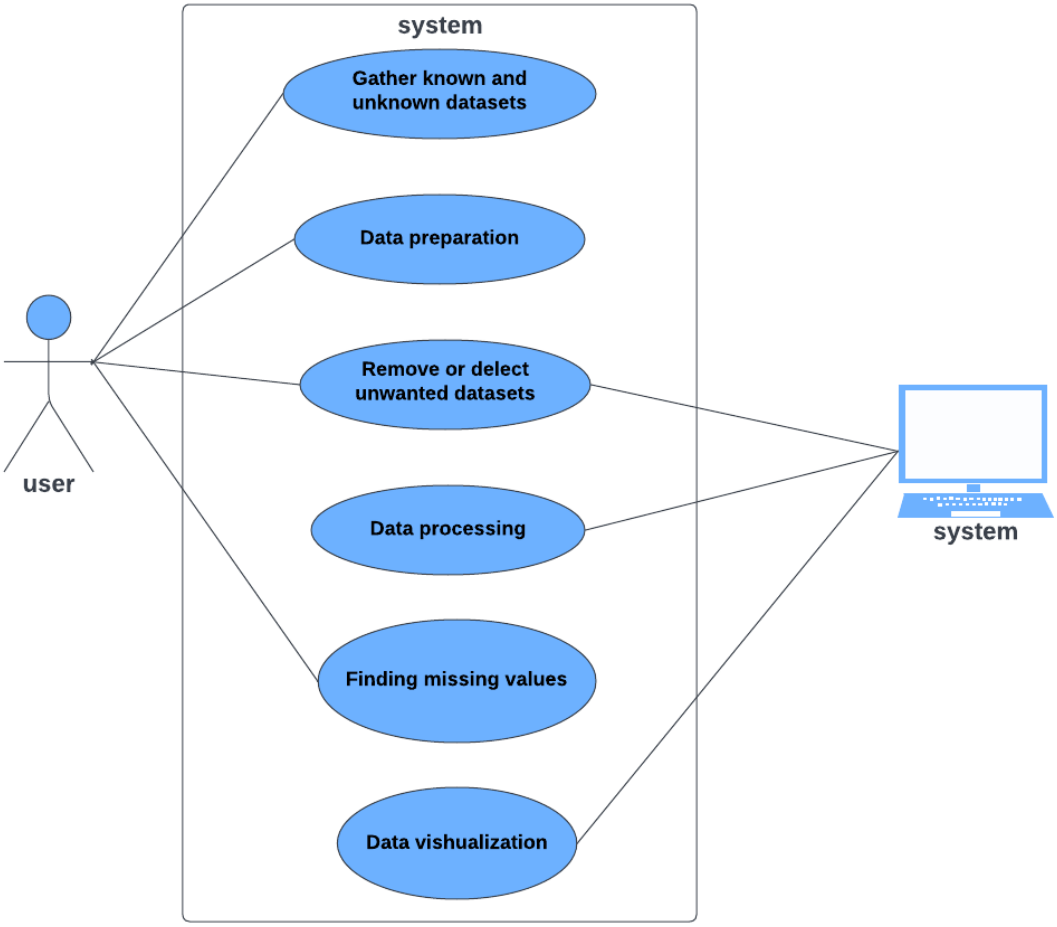
NAME OF THE PAPER	AUTHOR/YEAR	RELAVENT FINDING	ADVANTAGES	DISADVANTAGES
Antecedents Of E-Trust In Cab Services Market	Ms. Sharon Sophia. J, J. Clement Sudhahar, Joseph Varghese **YEAR : MARCH 2019**	The study reveals that there is an importance of customer trust towards online services influenced by website services and quality of website characteristics.	• To assess the user profile of the customers and perception on select cab aggregators. • To identify the determinants of usefulness, information quality and securit	• The study did not represent the entire population. • The study is limited by time bound and it does not address the broad area on the various factors of customer satisfaction since it focused only on trust and relative components of trust.
Research Paper on Increasing Preference to "Ola Cab Service"	Snehal Nikam, Surbhi Deshmukh, Dr. Priyanka Kokatnur **YEAR:APRIL 2020**	Uber is a provider of a mobile application connecting passengers with drivers for hire.	• Ola is first of its kind taxi aggregator service provider in the country. • It achieved the no.1 rank in the sector after acquiring Taxi for sure.	• Brand image can be easily influenced by the misbehaviour of the drivers as they are the ne in direct contact with the customers
A study on online cab services with special references.	Irene elsa mani **YEAR: SEPTEMBER 2018**	The Uber app allows consumers to submit a trip request, which is routed to crowd-sourced taxi drivers.	• To identify the problems encountered by the customers while availing the services.	• Monetization becomes difficult due to the demand.

A study on customer's preference for online cab services	Mr. Sathish Kumar J AND MRS.TR. Kalai Lakshmi **YEAR: MAY 2021**	This study shows the global interference of technology advancement in cab hailing services in smart cities which enables customers to hail taxis through their smart phones, become popular worldwide.	• To analyse the issues faced by the customers towards booking process • To analyse customer needs and comforts.	• Sometimes the drivers cancel the booking in the nick of time.
Consumers' Perspective on Cab Services	aibal Kumar Saha , Jupitara Kalita and Sangita Saha **YEAR: AUG 2018**	The companies Ola and Uber are the major players. Prime Cabs' owner Pallav Bagaria started the business with 30 cars of Indigo and due to the growing demand of cab services 70 more vehicles were added.	• It achieved the no.1 rank in the sector after acquiring Taxi for sure. The services offered by • ola are well appreciated by the public	• Drivers don't reach on time • The management of the company is not cooperative
Service Quality and Customer Satisfaction	A.SIVAKUMAR & Mrs. ANURADHA	The study stated that to determine relationship between the	• Ola leads a perfect balance in their excellent service and good quality cab	• This system does not have performance problem since it built

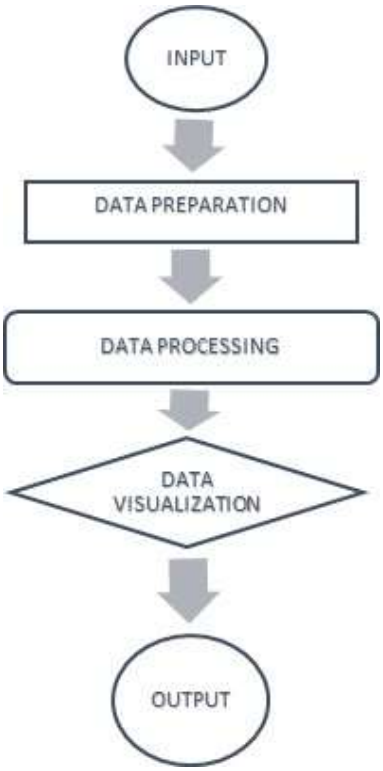
Towards Cab Service Providers	**YEAR:NOVEMBER 2021**	quality of service and customer satisfaction for cab services in Bangalore city.	provided followed by Red taxi for their customers	the recommendations offline
A study on factors influencing the consumers in selection of cab services.	Shipra Jain, Ms. Ekata Gupta, Ramandeep Kaur YEAR:MAY 2018	As data is maintained electronically, it's easy for a person to update the details, which overcomes the tedious updating task in the previous system.	Ÿou Can be guided by GPS or different apps to mobilized by the country Can have a car that	The cab management is a Web application and it is restricted to only limited type of Users i.e. only managers and admin.
Customer Satisfaction towards Call Taxi Services A study with reference	V. Hemanth Kumar and K. Sentamilselvan **YEAR: SEPTEMBER 2018**	The primary purpose of Descriptive research is to provide an accurate description or picture of the status or characteristics of a situation or phenomenon and hence the same is adopted in this study.	To ascertain the customer view towards the driver behaviour and courtesy. To provide inputs to enhance the services to delight the customers.	The precise effect still should be verified through long-term trial

DIAGRAMS

UML CLASS DIAGRAM:



FLOW CHART



METHODOLOGY

(i) Data Collection

Data collection is a systematic approach of collecting data. Irrespective of acting research for various purposes, data collection allows you to gain knowledge and original insights into a research problem. Here, data is downloaded containing start date, end date, miles, category, purpose and other fields from Kaggle in .csv format.

(ii) Data Cleaning

Data cleaning is the method of ensuring that the data taken is correct, consistent and usable. It is the main work that has to be done after we collect the data. Since data could be structured and unstructured, it is necessary to clean the data for our process.

(iii) Exploratory Data Analysis

Exploratory Data Analysis, also known as EDA, investigates the data and discovers patterns, spot anomalies and test hypotheses. EDA techniques are graphical with a few quantitative techniques. EDA's objectives include extracting important variables, identifying outliers, missing values, or human error and understanding the relationship(s), or lack of, the relationship between variables. Ultimately, maximizing your insights into a dataset and minimizing potential error. Here, using Python coding, we have visualized using bar plot and found necessary summary and information related to our requirement.

(iv) Data Visualization

Data visualization means presenting raw data through graphical representations that allow viewers (business analysts and executives) to explore the data and uncover deep insights. With this visualization, the process can be studied effectively or flow's explanation compared to the raw report. Data visualization benefits include analysis, quick action, identifying patterns, finding errors, understanding the story, exploring business insights and grasping the latest trends. We could get different types of bar plots between the variables namely purposes, counts and miles.

IMPLEMENTATION:

Import datasets (2016 – uber drives datas)

```
import numpy as np
import pandas as pd

import matplotlib.pyplot as plt
from builtins import list
import matplotlib
matplotlib.style.use('ggplot')

import datetime
import seaborn as sns
%matplotlib inline
```

```
sns.set(color_codes=True)
uber_df=pd.read_csv('uberdrive.csv')
```

Arrange START, STOP, MONTH, WEEK, HOURS

```
sns.set(color_codes=True)
uber_df=pd.read_csv('uberdrive.csv')
```

```
uber_drives['START_DATE*']=pd.to_datetime(uber_drives['START_DATE*'])
uber_drives['END_DATE*']=pd.to_datetime(uber_drives['END_DATE*'])

uber_drives['HOUR']=[x.hour for x in uber_drives['START_DATE*']]
uber_drives['DAY']=[x.hour for x in uber_drives['START_DATE*']]
uber_drives['MONTH']=[x.hour for x in uber_drives['START_DATE*']]
uber_drives['WEEKDAY']=[x.hour for x in uber_drives['START_DATE*']]
uber_drives['DAY_OF_WEEK']=[x.hour for x in uber_drives['START_DATE*']]
```

```
uber_drives.head(10)
```

	START_DATE*	END_DATE*	CATEGORY*	START*	STOP*	MILES*	PURPOSE*	HOUR	DAY	MONTH	WEEKDAY	DAY_OF_WEEK
0	2016-01-01 21:11:00	2016-01-01 21:17:00	Business	Fort Pierce	Fort Pierce	5.1	Meal/Entertain	21	21	21	21	21
1	2016-01-02 01:25:00	2016-01-02 01:37:00	Business	Fort Pierce	Fort Pierce	5.0	NaN	1	1	1	1	1
2	2016-01-02 20:25:00	2016-01-02 20:38:00	Business	Fort Pierce	Fort Pierce	4.8	Errand/Supplies	20	20	20	20	20
3	2016-01-05 17:31:00	2016-01-05 17:45:00	Business	Fort Pierce	Fort Pierce	4.7	Meeting	17	17	17	17	17
4	2016-01-06 14:42:00	2016-01-06 15:49:00	Business	Fort Pierce	West Palm Beach	63.7	Customer Visit	14	14	14	14	14
5	2016-01-06 17:15:00	2016-01-06 17:19:00	Business	West Palm Beach	West Palm Beach	4.3	Meal/Entertain	17	17	17	17	17
6	2016-01-06 17:30:00	2016-01-06 17:35:00	Business	West Palm Beach	Palm Beach	7.1	Meeting	17	17	17	17	17
7	2016-01-07 13:27:00	2016-01-07 13:33:00	Business	Cary	Cary	0.8	Meeting	13	13	13	13	13
8	2016-01-10 08:05:00	2016-01-10 08:25:00	Business	Cary	Morrisville	8.3	Meeting	8	8	8	8	8
9	2016-01-10 12:17:00	2016-01-10 12:44:00	Business	Jamaica	New York	16.5	Customer Visit	12	12	12	12	12

Check that all data is fixed and ready to work on it

```
# check that all data is fixed and ready to work on it
uber_drives.info()
```

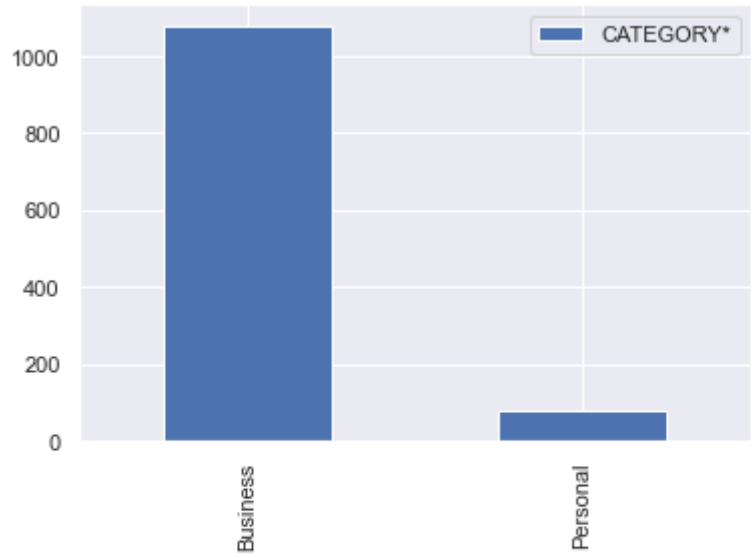
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1155 entries, 0 to 1154
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype         
---  ------       --------------  -----         
 0   START_DATE*  1155 non-null   datetime64[ns]
 1   END_DATE*    1155 non-null   datetime64[ns]
 2   CATEGORY*    1155 non-null   object        
 3   START*       1155 non-null   object        
 4   STOP*        1155 non-null   object        
 5   MILES*       1155 non-null   float64       
 6   PURPOSE*     653 non-null    object        
 7   HOUR         1155 non-null   int64         
 8   DAY          1155 non-null   int64         
 9   MONTH        1155 non-null   int64         
 10  WEEKDAY      1155 non-null   int64         
 11  DAY_OF_WEEK  1155 non-null   int64         
dtypes: datetime64[ns](2), float64(1), int64(5), object(4)
memory usage: 108.4+ KB
```

Plot number of trip at each category

```
uber_drives['CATEGORY*'].value_counts(normalize=True)*100
```

```
Business    93.333333
Personal     6.666667
Name: CATEGORY*, dtype: float64
```

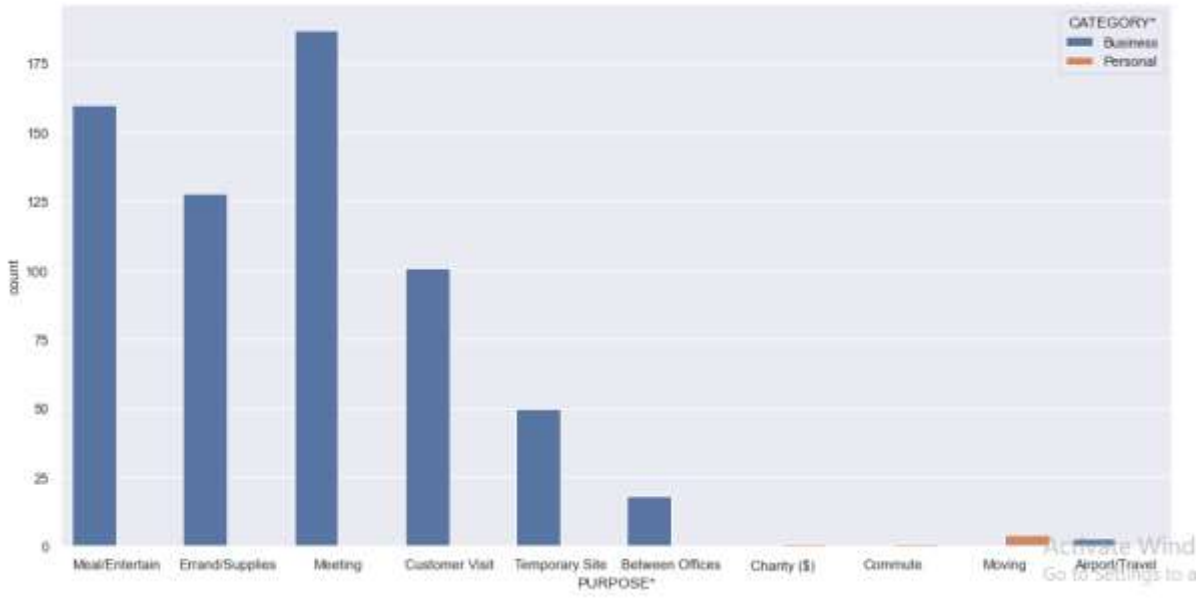
```
#PLOT NUMBER OF TRIP AT EACH CATEGORY

#another way
uber_drives.head()

visual_df2 = pd.DataFrame(uber_drives['CATEGORY*'].value_counts())
visual_df2.reset_index()

visual_df2.plot(kind = 'bar')
plt.show()
visual_df2
```



	CATEGORY*
Business	1078
Personal	77

```
plt.figure(figsize=(16,8))
sns.countplot(uber_drives['PURPOSE*'],hue=uber_drives['CATEGORY*'])
```

```
<AxesSubplot:xlabel='PURPOSE*', ylabel='count'>
```



Check that all data is fixed and ready to work on it

```
# check that all data is fixed and ready to work on it
uber_drives.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1155 entries, 0 to 1154
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   START_DATE*  1155 non-null   datetime64[ns]
 1   END_DATE*    1155 non-null   datetime64[ns]
 2   CATEGORY*    1155 non-null   object
 3   START*       1155 non-null   object
 4   STOP*        1155 non-null   object
 5   MILES*       1155 non-null   float64
 6   PURPOSE*     653 non-null    object
 7   HOUR         1155 non-null   int64
 8   DAY          1155 non-null   int64
 9   MONTH        1155 non-null   int64
 10  WEEKDAY      1155 non-null   int64
 11  DAY_OF_WEEK  1155 non-null   int64
dtypes: datetime64[ns](2), float64(1), int64(5), object(4)
memory usage: 108.4+ KB
```

Extract month from start date

```
#extract month from start date
count = 0
month=[]
while count < len(uber_drives):
    month.append(uber_drives['START_DATE*'][count].month)
    count = count+1
uber_drives['Month'] = month
```

```
#Total number of unique START are STOP destination
uber_drives['START*'].nunique()
```

```
177
```

```
uber_drives['STOP*'].nunique()
```

```
188
```

```
uber_drives.groupby(['START*','STOP*'])['MILES*'].sum().sort_values(ascending=False).head()
```

```
START*            STOP*
Unknown Location  Unknown Location    1360.8
Morrisville       Cary                395.7
Cary              Durham              390.0
                  Morrisville         380.0
Raleigh           Cary                365.7
Name: MILES*, dtype: float64
```

```
uber_drives.size
```

```
15015
```

Information of missing values

```
#Information of missing values
pd.isna(uber_drives)
```

	START_DATE*	END_DATE*	CATEGORY*	START*	STOP*	MILES*	PURPOSE*	HOUR	DAY	MONTH	WEEKDAY	DAY_OF_WEEK	Month
0	False	False	False	False	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	True	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...
1150	False	False	False	False	False	False	False	False	False	False	False	False	False
1151	False	False	False	False	False	False	False	False	False	False	False	False	False
1152	False	False	False	False	False	False	False	False	False	False	False	False	False
1153	False	False	False	False	False	False	False	False	False	False	False	False	False
1154	False	False	False	False	False	False	False	False	False	False	False	False	False

Checking for null values from data

```
#checking for null values from data
uber_drives.isnull().sum()
```

```
START_DATE*       0
END_DATE*         0
CATEGORY*         0
START*            0
STOP*             0
MILES*            0
PURPOSE*        502
HOUR              0
DAY               0
MONTH             0
WEEKDAY           0
DAY_OF_WEEK       0
Month             0
dtype: int64
```

Summary of the data

```
#Summary of the data
uber_drives.describe()
```

	MILES*	HOUR	DAY	MONTH	WEEKDAY	DAY_OF_WEEK	Month
count	1155.000000	1155.000000	1155.000000	1155.000000	1155.000000	1155.000000	1155.000000
mean	10.566840	14.696104	14.696104	14.696104	14.696104	14.696104	6.982684
std	21.579106	4.575226	4.575226	4.575226	4.575226	4.575226	3.544915
min	0.500000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000
25%	2.900000	12.000000	12.000000	12.000000	12.000000	12.000000	3.500000
50%	6.000000	15.000000	15.000000	15.000000	15.000000	15.000000	7.000000
75%	10.400000	18.000000	18.000000	18.000000	18.000000	18.000000	10.000000
max	310.300000	23.000000	23.000000	23.000000	23.000000	23.000000	12.000000

The most popular STARTING and DROPPING destination

```
#The most popular STARTING and DROPPING destination
uber_drives['START*'].value_counts()
```

```
Cary               201
Unknown Location   148
Morrisville         85
Whitebridge         68
Islamabad           57
                  ...
Florence             1
Ridgeland            1
Daytona Beach        1
Sky Lake             1
Gampaha              1
Name: START*, Length: 177, dtype: int64
```

```
uber_drives['STOP*'].value_counts()
```

```
Cary                 203
Unknown Location     149
Morrisville           84
Whitebridge           65
Islamabad             58
                    ...
Daytona Beach          1
Sand Lake Commons      1
Sky Lake               1
Vista East             1
Ilukwatta              1
Name: STOP*, Length: 188, dtype: int64
```

The most frequent route

```
#The most frequent route
Route=uber_drives['START*'].astype(str)+','+ uber_drives['STOP*']
Route.value_counts()
```

```
Unknown Location,Unknown Location    86
Morrisville,Cary                     75
Cary,Morrisville                     67
Cary,Cary                            53
Cary,Durham                          36
                                     ..
Chessington,Chessington               1
Meredith Townes,Harden Place          1
Cary,Holly Springs                    1
Meredith,Cedar Hill                   1
Gampaha,Ilukwatta                     1
Length: 363, dtype: int64
```

Unique START destination

```
#Unique START destination
uber_drives['START*'].unique()
```

```
array(['Fort Pierce', 'West Palm Beach', 'Cary', 'Jamaica', 'New York',
       'Elmhurst', 'Midtown', 'East Harlem', 'Flatiron District',
       'Midtown East', 'Hudson Square', 'Lower Manhattan',
       "Hell's Kitchen", 'Downtown', 'Gulfton', 'Houston', 'Eagan Park',
       'Morrisville', 'Durham', 'Farmington Woods', 'Whitebridge',
       'Lake Wellingborough', 'Fayetteville Street', 'Raleigh',
       'Hazelwood', 'Fairmont', 'Meredith Townes', 'Apex', 'Chapel Hill',
       'Northwoods', 'Edgehill Farms', 'Tanglewood', 'Preston',
       'Eastgate', 'East Elmhurst', 'Jackson Heights', 'Long Island City',
       'Katunayaka', 'Unknown Location', 'Colombo', 'Nugegoda',
       'Islamabad', 'R?walpindi', 'Noorpur Shahan', 'Heritage Pines',
       'Westpark Place', 'Waverly Place', 'Wayne Ridge', 'Weston',
       'East Austin', 'West University', 'South Congress', 'The Drag',
       'Congress Ave District', 'Red River District', 'Georgian Acres',
       'North Austin', 'Coxville', 'Convention Center District', 'Austin',
       'Katy', 'Sharpstown', 'Sugar Land', 'Galveston', 'Port Bolivar',
       'Washington Avenue', 'Briar Meadow', 'Latta', 'Jacksonville',
       'Couples Glen', 'Kissimmee', 'Lake Reams', 'Orlando',
       'Sand Lake Commons', 'Sky Lake', 'Daytona Beach', 'Ridgeland',
       'Florence', 'Meredith', 'Holly Springs', 'Chessington', 'Burtrose',
       'Parkway', 'Mcvan', 'Capitol One', 'University District',
       'Seattle', 'Redmond', 'Bellevue', 'San Francisco', 'Palo Alto',
       'Sunnyvale', 'Newark', 'Menlo Park', 'Old City', 'Savon Height',
       'Kilarney Woods', 'Townes at Everett Crossing', 'Huntington Woods',
       'Seaport', 'Medical Centre', 'Rose Hill', 'Soho', 'Tribeca',
       'Financial District', 'Oakland', 'Emeryville', 'Berkeley',
       'Kenner', 'CBD', 'Lower Garden District', 'Lakeview', 'Storyville',
       'New Orleans', 'Metairie', 'Chalmette', 'Arabi',
       'Pontchartrain Shores', 'Marigny', 'Covington', 'Mandeville',
       'Jamestown Court', 'Summerwinds', 'Parkwood',
       'Pontchartrain Beach', 'St Thomas', 'Banner Elk', 'Elk Park',
```

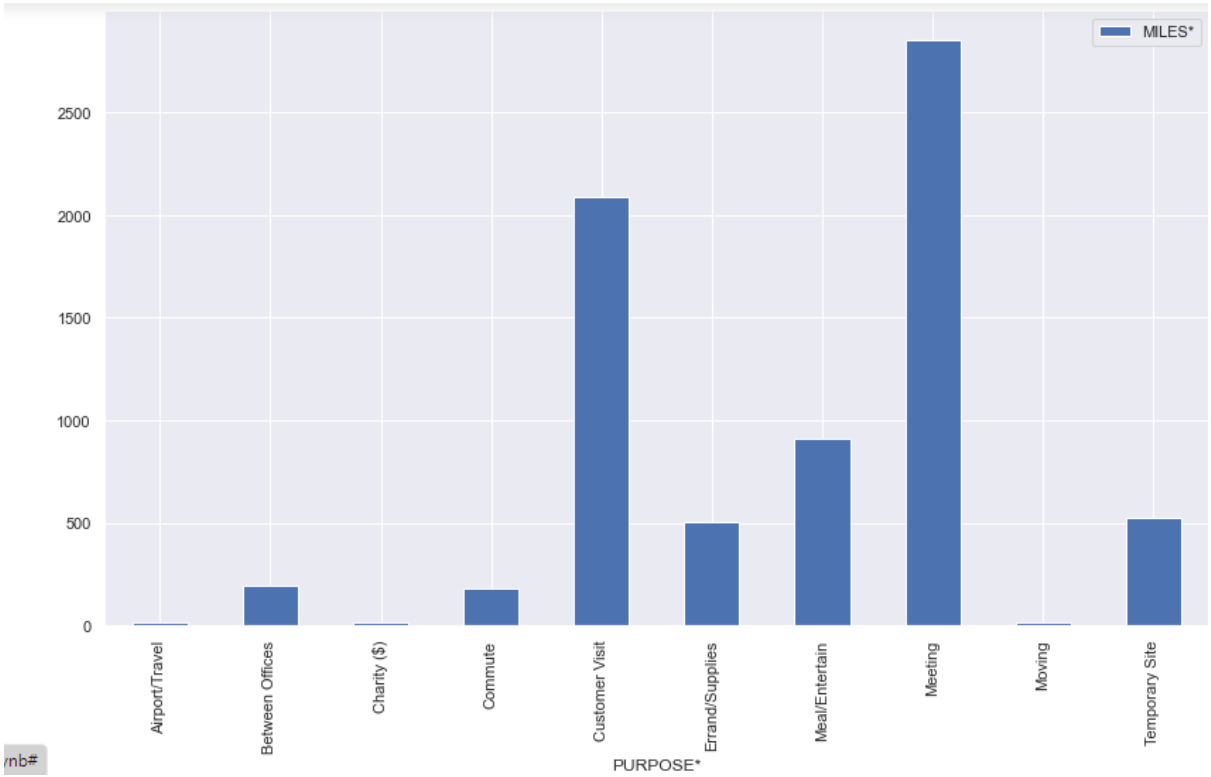
```
#Unique STOP destination
uber_drives['STOP*'].unique()
```

```
array(['Fort Pierce', 'West Palm Beach', 'Palm Beach', 'Cary',
       'Morrisville', 'New York', 'Queens', 'East Harlem', 'NoMad',
       'Midtown', 'Midtown East', 'Hudson Square', 'Lower Manhattan',
       "Hell's Kitchen", 'Queens County', 'Gulfton', 'Downtown',
       'Houston', 'Jamestown Court', 'Durham', 'Whitebridge',
       'Lake Wellingborough', 'Raleigh', 'Umstead', 'Hazelwood',
       'Westpark Place', 'Meredith Townes', 'Leesville Hollow', 'Apex',
       'Chapel Hill', 'Williamsburg Manor', 'Macgregor Downs',
       'Edgehill Farms', 'Northwoods', 'Tanglewood', 'Preston',
       'Walnut Terrace', 'Jackson Heights', 'East Elmhurst',
       'Midtown West', 'Long Island City', 'Jamaica', 'Unknown Location',
       'Colombo', 'Nugegoda', 'Katunayaka', 'Islamabad', 'R?walpindi',
       'Noorpur Shahan', 'Heritage Pines', 'Waverly Place', 'Wayne Ridge',
       'Depot Historic District', 'Weston', 'West University',
       'South Congress', 'Arts District', 'Congress Ave District',
       'Red River District', 'The Drag', 'Convention Center District',
       'North Austin', 'Coxville', 'Katy', 'Alief', 'Sharpstown',
       'Sugar Land', 'Galveston', 'Port Bolivar', 'Washington Avenue',
       'Briar Meadow', 'Greater Greenspoint', 'Latta', 'Jacksonville',
       'Kissimmee', 'Isles of Buena Vista', 'Orlando', 'Lake Reams',
       'Vista East', 'Sky Lake', 'Sand Lake Commons', 'Daytona Beach',
       'Ridgeland', 'Florence', 'Cedar Hill', 'Holly Springs',
       'Harden Place', 'Chessington', 'Burtrose', 'Parkway',
       'Capitol One', 'University District', 'Redmond', 'Bellevue',
       'Seattle', 'Mcvan', 'Palo Alto', 'Sunnyvale', 'Newark',
       'Menlo Park', 'San Francisco', 'Parkway Museums', 'Hog Island',
       'Savon Height', 'Kildaire Farms', 'Kilarney Woods',
```

Plot between purpose, miles and number of trips

```
#Plot between purpose,miles and number of trips
pivot=pd.pivot_table(uber_drives,index='PURPOSE*',aggfunc='sum',values='MILES*')
pivot
pivot.plot(kind='bar',figsize=(14,8))
```

```
<AxesSubplot:xlabel='PURPOSE*'>
```



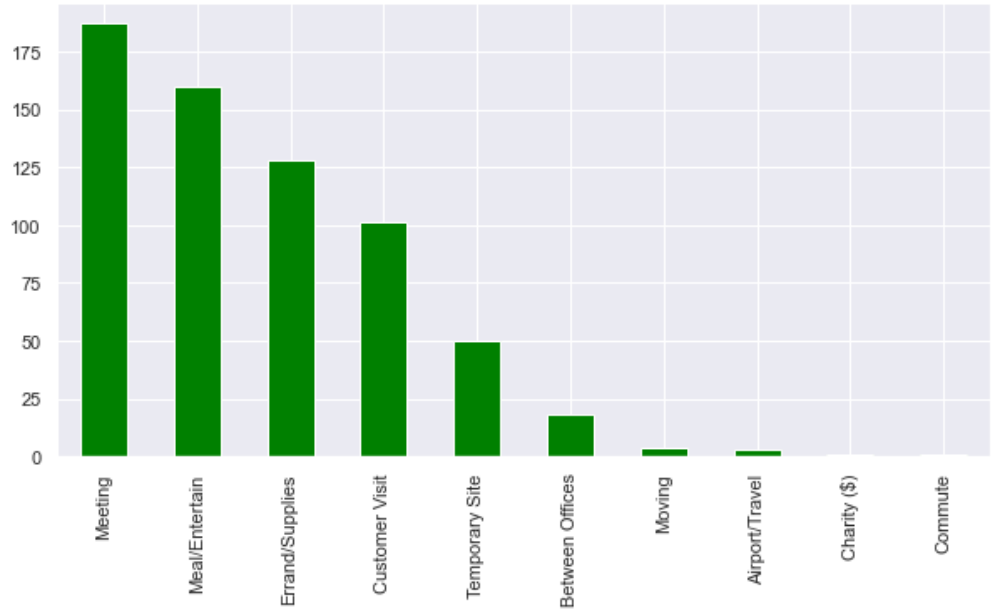
See how many trips made by each purpose

```
#How many miles was earned per  purpose ?
uber_drives.groupby('PURPOSE*').sum()['MILES*'].sort_values(ascending = False)
```

```
PURPOSE*
Meeting            2851.3
Customer Visit     2089.5
Meal/Entertain      911.7
Temporary Site      523.7
Errand/Supplies     508.0
Between Offices     197.0
Commute             180.2
Moving               18.2
Airport/Travel       16.5
Charity ($)          15.1
Name: MILES*, dtype: float64
```

```
# see how many trips made by each purpose
purpose_time = uber_drives['PURPOSE*'].value_counts()
purpose_time.plot(kind='bar',figsize=(10,5),color='green')
```

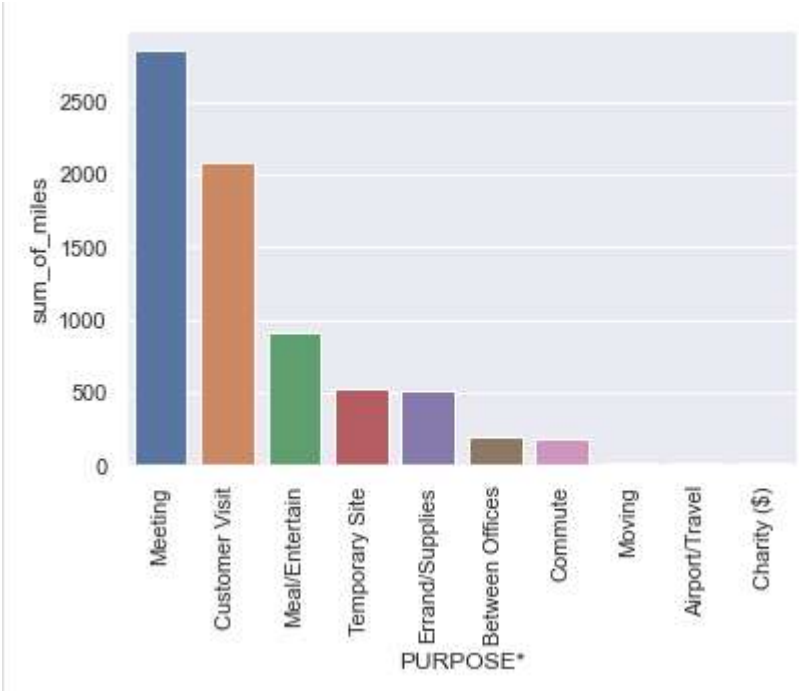
```
<AxesSubplot:>
```



#Plot a bar graph of Purposes vs Distance.

```
#Plot a bar graph of Purposes vs Distance.
k3 = uber_drives.groupby('PURPOSE*')['MILES*'].sum().sort_values(ascending=False).head(10)
k3
k3= k3.reset_index() # flatten the dataframe
k3
k3.columns = ['PURPOSE*' ,'sum_of_miles']
k3
%matplotlib inline
import seaborn as sns
sns.barplot(data= k3 , x= 'PURPOSE*' , y ='sum_of_miles')
plt.xticks(rotation=90)
```

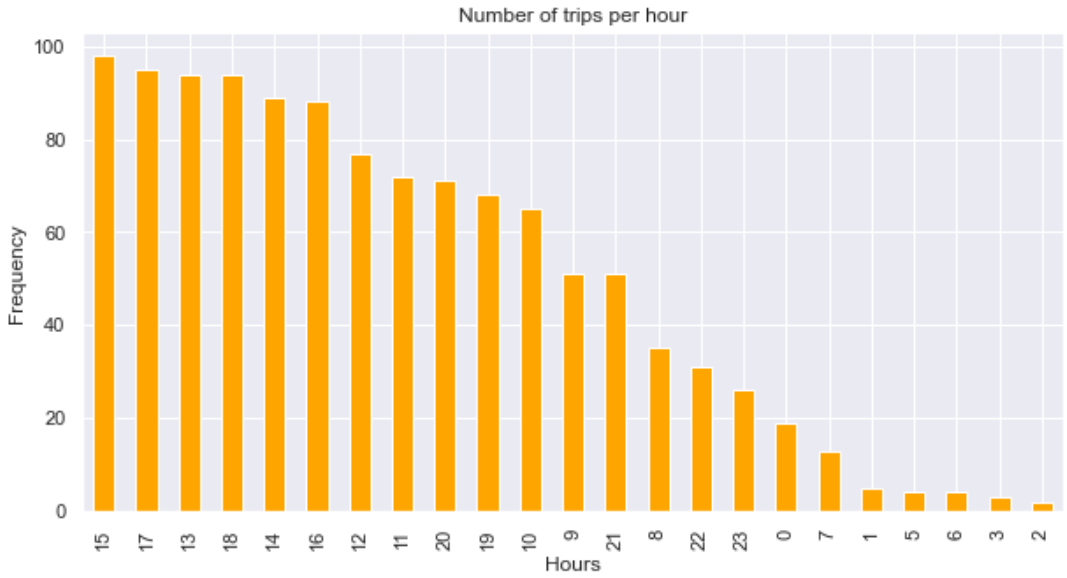
```
(array([0, 1, 2, 3, 4, 5, 6, 7, 8, 9]),
 [Text(0, 0, 'Meeting'),
  Text(1, 0, 'Customer Visit'),
  Text(2, 0, 'Meal/Entertain'),
  Text(3, 0, 'Temporary Site'),
  Text(4, 0, 'Errand/Supplies'),
  Text(5, 0, 'Between Offices'),
  Text(6, 0, 'Commute'),
  Text(7, 0, 'Moving'),
  Text(8, 0, 'Airport/Travel'),
  Text(9, 0, 'Charity ($)')])
```



I need to see how many trip made at each clock and as you see the clock which has the highest number of trips is 3:00PM

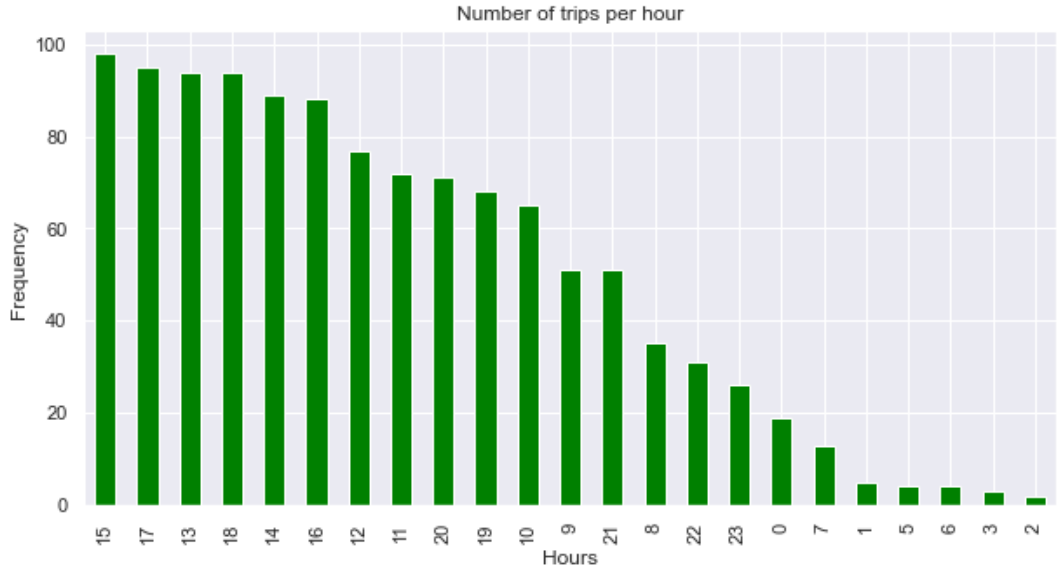
```
# I need to see how many trip made at each clock and as you see the clock which has the higest number of trips is 3:00PM
hours = uber_drives['HOUR'].value_counts()
hours.plot(kind='bar',color='orange',figsize=(10,5))
plt.xlabel('Hours')
plt.ylabel('Frequency')
plt.title('Number of trips per hour')
```

```
Text(0.5, 1.0, 'Number of trips per hour')
```



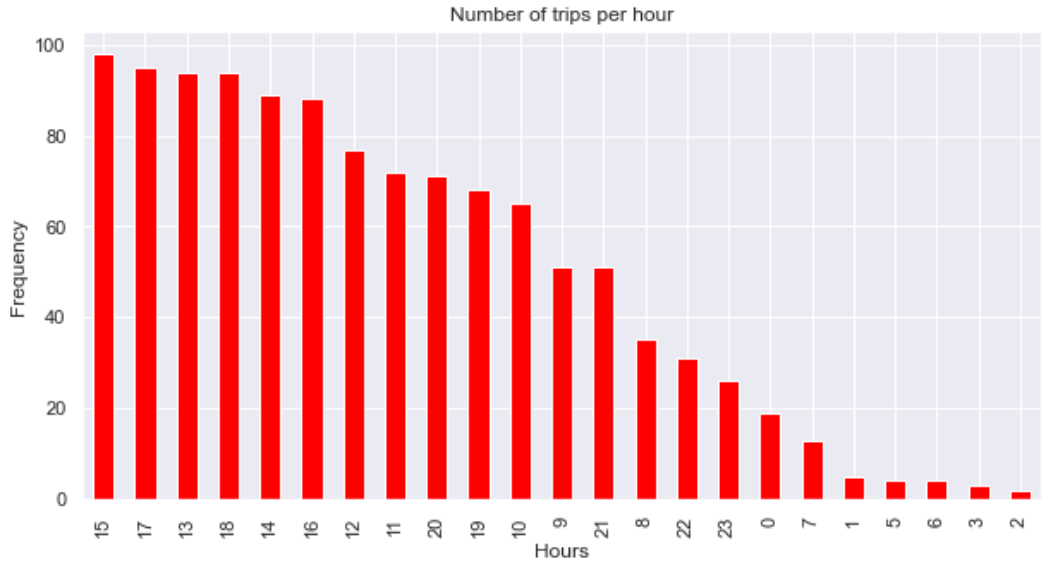
```
days = uber_drives['DAY'].value_counts()
hours.plot(kind='bar',color='green',figsize=(10,5))
plt.xlabel('Hours')
plt.ylabel('Frequency')
plt.title('Number of trips per hour')
```

```
Text(0.5, 1.0, 'Number of trips per hour')
```




```
month = uber_drives['MONTH'].value_counts()
hours.plot(kind='bar',color='red',figsize=(10,5))
plt.xlabel('Hours')
plt.ylabel('Frequency')
plt.title('Number of trips per hour')
```

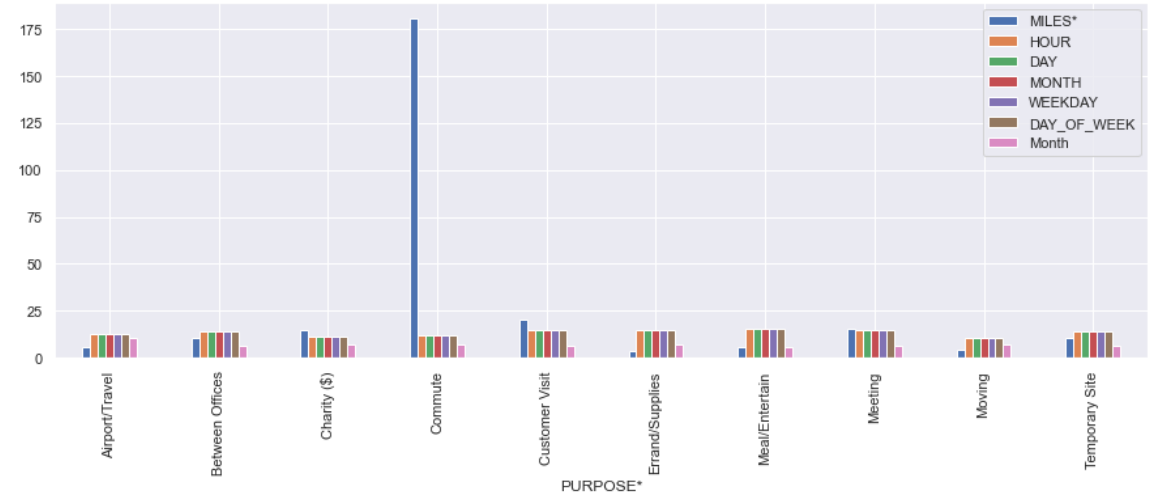
```
Text(0.5, 1.0, 'Number of trips per hour')
```



Average of each trip according to purpose

```
# aveverage of each trip according to purpose
purpose = uber_drives.groupby('PURPOSE*').mean()
purpose.plot(kind = 'bar',figsize=(15,5))
```

```
<AxesSubplot:xlabel='PURPOSE*'>
```



BENEFITS:

- **Largest Ride Sharing Technology:** In 2020, Uber is available in more than 93 countries and over 900 cities, with 103 million monthly users served by a total of 5 million drivers.
- **Strong Brand Recognition:** Uber has maintained a strong brand recognition in over 50 countries. It has already overtaken GM, Honda, and Ford regarding brand value framework.
- **Low Fixed Investment:** Uber operates on low fixed investment (low operational cost) and has easily accessed more cites in its communicative network. Because there is no fixed infrastructure or investment in place, the company continues to expand at a fast pace.
- **Low Prices as Compared to Taxis and Other Commute apps:** Uber offers low prices as compared to traditional taxis. The biggest difference between taxis and Uber is that Taxis charge per mile (while traveling) and per minute (when not traveling).

- **Customer to Driver Interaction:** The business model of Uber is ideal for a customer to driver interaction. Uber has created a rating system that helps customers rate their traveling experience as well as the driver. This rating system helps identify the best drivers and monitors the performance of the drivers.

LIMITATIONS:

- **Multiple Scandals**: Uber's brand has received negative coverage over numerous scandals and controversies. Cases such as sexual harassment and targeted attacks have defamed the company. It came to the point that its co-founder Travis Kalanick had to resign**.**
- **Substantial Losses**: Although it has increased its revenues, Uber has been facing significant losses since 2009. In order to beat out its growing competition, the company began providing bonuses to its drivers and discounts to its customers.
- **Public Backlash**: Uber faced a severe public backlash over its high pricing during Hurricane Sandy. This forced the company to revise its policy.
- **Poor Working Condition:** Most companies invest heavily to support their employees. On the other hand, Uber's drivers are almost entirely on their own, which exposes them to security risks. Also, they have to bear expenses like insurance, repairs, and gas.

FUTURE WORK:

The biggest bet Uber has made is on autonomous vehicle- the self-driving cars. And that's a big one, because really, it's just like at the edge of living…… If you have the money you can dream the big wish.. And Uber is pretty good in that way, very visionary.

1. The road to self-driving vehicle:

This innovative team is dedicated to building safe, reliable, and cost-effective self-driving Technologies. With teams in Detroit, Pittsburgh, San Francisco, Tempe, and Toronto, the Group is bringing self-driving cars and freight trucks to the Uber network.

2. Uber Elevate:

The future of urban air transport Uber is developing shared air transportation—planned for 2023—between suburbs and cities, and ultimately within cities. We're working with our

Elevate Network partners to launch fleets of small, electric VTOL (vertical take-off and landing) aircraft in Delhi, India

CONCLUSION:

We conclude that from the taken uber drives (2016) dataset, we have detected values the missing values, dimension, size and information about all variables in the dataset so that we derived the summary of original data.

Also we have found the most frequent category of the trip is business compared to personal category of proportion 93.33: 6.67. The total number of unique START and STOP destination are found to be 177 and 188. The most popular starting and dropping point in carry and the most frequent route taken by uber driver is carry to Morris vive and vice versa.

Also we have plotted the purpose of each trip according to the respective categories with their miles covered.

From this we have concluded that the maximum number of trips travelled for the meeting purpose and the longest miles covered is also for meeting purpose.

REFERENCES

1 Parasuraman, Valarie Zeithaml and Leonard berry(1988)Literature review-measuring service qualitySERVQUALto compare service quality performance with customer service quality needs.

2 Zhigany Yao and Xiao-dong Ding 2011, Changan University, "A study on measuring passenger's perceptions of taxi service quality with weighted SERVPERF". Applied Mechanics and Material, Volume-97-98, pp-1181-1184, Trans Tech Publications, Switzerland.

3 Tazyn Rahman (2014), they stated that "a study on radio taxi service providers in Guwahati", International Journal of advance and innovative research, volume-1, issue-1, pp 19-25. Dr. Mona Bhalla (2018),"An empirical study of the Indian taxi market- With a brief analysis of the emerging taxi aggregator sectors". Shanlax International Journal of Commerce. Volume:6 issue:1 ISSN: 2320-4168.

4 Sabena Tika Arsyawati and Mahir Pradana, 2017, "A study on service quality factors that determine consumer satisfaction uber according to user's perception

in Jakarta", International Journal of Scientific Engineering and Applied Science, volume-3, issue-1, ISSN-2395- 3470.

5 Mehul Mehta and Riddhish Joshi 2015, says that "Radio taxi industry, current scenario industry forces and challenges in India", International journal of Research in Management, and social science, volume-3, issue-1, pp100.

6 Mr. Worapong Archarworarit 2015, "The study on consumer behavior and selection criteria on alternative taxi service in Bangkok" Digital collections, Thammasat university thesis.

7 Pajaree Ackaradejruangsri, 2015, "A study about grab taxi service an alternative ride service improving satisfaction in Thailand". Society of interdisciplinary business research, ISSN: 2304-1013, volume-4, issue-3, pp: 49-61

8 Keder Gokhale 2021, Cab service quality influence on customer satisfaction- A study conducted on major cab aggregators in Bangaluru. IT in industry, Vol-9, issue-2, 2021, pp-338-344