# Industrial Project Report

*Submitted in partial fulfillment of the degree of*

## B-tech in Electrical Engineering

### By

*ABINAB NAG [11901620011]*
*RANJAN GHOSH [11901621021]*
*PRIYA CHAKRABORTY [11901620009]*
*RUPA SHARMA [11901621026]*
*TSHERINGMA TAMANG [11901621016]*

## Third-year student of

## SILIGURI INSTITUTE OF TECHNOLOGY

*THIS IS SUBMITTED IN FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF*

**AFFILIATED TO**

**Maulana Abul Kalam Azad University of Technology**

## Under the supervision of :- Mr. Ripam Kundu

## Sikharthy Infotech Pvt. Ltd.

# *PROJECT ON :-   UBER DATA ANALYSIS WITH MACHINE LEARNING*

By

## *ABINAB NAG [11901620011]*
## *RANJAN GHOSH [11901621021]*
## *PRIYA CHAKRABORTY [11901620009]*
## *RUPA SHARMA [11901621026]*
## *TSHERINGMA TAMANG [11901621016]*

UNDER THE GUIDANCE OF

**Mr. Ripam Kundu**

**Project Guide**

**Sikharthy Infotech Pvt. Ltd.**



*THIS IS SUBMITTED IN FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF*

**B.Tech**

IN

Electrical Engineering

## SILIGURI INSTITUTE OF TECHNOLOGY

**AFFILIATED TO**

**Maulana  Abul Kalam Azad University of Technology**

# Department of Electrical Engineering

I hereby forward the documentation prepared under my supervision by **Ripam Kundu Sir** entitled **Siliguri Institute Of Technology** to be accepted as fulfillment of the requirement for the Degree of Bachelor of Technology in Electrical Engineering, **Siliguri Institute Of Technology** affiliated to **Maulana Abul Kalam Azad University of Technology** (**MAKAUT**).

_____

**Mr.Ripam Kundu
(Software Developer)
Project Guide
Sikharthy Infotech Pvt. Ltd.**

_____

**HOD**

**Department Of Electrical Engineering, SIT**

_____

**Shilpi Ghosal
(Director)
Sikharthy Infotech Pvt. Ltd.**

_____

**TPO
Siliguri Institute of Technology**

Siliguri Institute of Technology   Hill Cart Road, Salbari, Sukna, West Bengal 734009

# Certificate of Approval

The foregoing project is hereby approved as a creditable study for the B.Tech in Electrical Engineering presented in a manner of satisfactory to warrant its acceptance as a prerequisite to the degree for which it has been submitted. It is understood that by this approval the undersigned do not necessarily endorsed or approved any statement made, opinion expressed or conclusion therein but approve this project only for the purpose for which it is submitted.

Final Examination for
Evaluation of the Project

---------------------------------------

----------------------------------------

-----------------------------------------

**Signatures of Examiners**

# **ABSTRACT**

*The paper explains the working of an Uber dataset, which contains data produced by Uber for Russia . Uber is defined as a P2P platform. The platform links you to drivers who can take you to your destination. The dataset includes primary data on Uber pickups with details including the date, time of the ride as well as longitude-latitude information, Using the information, the paper explains the use of the classification algorithm on the set of data and classify the various parts of Russia . Since the industry is booming and expected to grow shortly. Effective taxi dispatching will facilitate each driver and passenger to reduce the wait time to seek out one another. The model is employed to predict the demand on points of the Russia.*

# ACKNOWLEDGEMENT

It is a great pleasure for me to acknowledge the assistance and participation of a large number of individuals in this attempt. Our project report has been structured under the valued suggestion, support, and guidance of **Mr. Ripam Kundu**. Under his guidance, we have accomplished the challenging task in a very short time.

Finally, we express our sincere thankfulness to our family members for inspiring me all throughout and always encouraging us.

**Group Member Signature**

_____

_____

_____

_____

# TABLE OF CONTENTS

We will use Python and its different libraries to complete the uber data analysis

**WHAT LIBRARIES WE USED**

# Importing Libraries

The analysis will be done using the following libraries :

- **Pandas**:  This library helps to load the data frame in a 2D array format and has multiple functions to perform analysis tasks in one go.
- **NumPy**: NumPy arrays are very fast and can perform large computations in a very short time.
- **Matplotlib** / **Seaborn**: This library is used to draw visualizations.
- **Plotly**: Plotly is a free and open-source graphing library for Python.
- **Matplot3D :** The mplot3d toolkit adds simple 3D plotting capabilities to matplotlib by supplying an axes object that can create a 2D projection of a 3D scene

To importing all these libraries, we can use the  below code :

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.graph_objects as go
from  mpl_toolkits.mplot3D import Axes3D
```

# Importing Dataset

After importing all the libraries , you can import the dataset using the pandas library.

```python
dataset = pd.read_csv("uber_dataset.csv")
dataset.head()
```

**So after importing the datasets the output we get is : -**

| | trip_completed_at | trip_status | ride_hailing_app | | trip_uid | driver_uid | rider_uid | cu |
|---|---|---|---|---|---|---|---|---|
| 0 | May 11, 2015 at 6:55PM | Completed | | Uber | ee89076fd9da9bddf5f096b0ca42f8d5 | 05cfeb269e606247fe9d2b6082942c59 | 3ffa4a71a5aa791a8bc3409f5b15b936 | |
| 1 | May 11, 2015 at 8:12PM | Completed | | Uber | 518be51d403944a03c47e8d1f2c87311 | 4a4e248742f9d5ff517c5bbbb48d0e54 | 3ffa4a71a5aa791a8bc3409f5b15b936 | |
| 2 | May 13, 2015 at 11:38AM | Completed | | Uber | 6e460cc8a12c3c6568d0d4a67ac58393 | cb249a2bd807ca78697b4ed0348c37da | 3ffa4a71a5aa791a8bc3409f5b15b936 | |
| 3 | May 16, 2015 at 1:44AM | Completed | | Uber | 49613a86a04e6c15d72b51d1a2935d81 | d3f73f8151c2e8c34b541f961db7f5fa | 3ffa4a71a5aa791a8bc3409f5b15b936 | |
| 4 | May 16, 2015 at 3:18AM | Completed | | Uber | 9896148fdecdb4c5d977a8691510bdb6 | 1287d21e6455ee40d4861f6b91c680f4 | 3ffa4a71a5aa791a8bc3409f5b15b936 | |

5 rows × 45 columns

# DATA CLEANING :-

Data cleaning is the process of preparing data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted.

```
dataset[dataset['surge_multiplier'].isnull()].head(5)
```

| | trip_completed_at | trip_status | ride_hailing_app | | trip_uid | driver_uid | rider_uid | cu |
|---|---|---|---|---|---|---|---|---|
| 20 | June 12, 2015 at 6:01PM | Completed | | Gett | c9df3a9edb2c7b37c80f9ffa5e1b8c36 | 151944a8f9967b3edad02deb712d61f2 | 3ffa4a71a5aa791a8bc3409f5b15b936 | |
| 135 | March 5, 2016 at 4:35PM | Completed | | Gett | 19564cbd0929d51297a5fce739a5c777 | 524f217db5cb84a0c343cb7f05f40f52 | 3ffa4a71a5aa791a8bc3409f5b15b936 | |
| 137 | March 17, 2016 at 12:57PM | Completed | | Gett | 2f7ca9c163bb40c5e5b49771e929075a | af6d53a1a051b89037613fa74b0a2039 | 3ffa4a71a5aa791a8bc3409f5b15b936 | |
| 142 | April 3, 2016 at 3:15PM | Completed | | Gett | b603f7e2fbda958c778a8e236b369b34 | b700e12aef6fe616f4132b5b10735f26 | 3ffa4a71a5aa791a8bc3409f5b15b936 | |
| 199 | July 29, 2016 at 3:35PM | Completed | | Gett | 1e4023aecad6062b038656b9bec5b433 | e75db42065bc8e071d2b57c0c9977376 | 3ffa4a71a5aa791a8bc3409f5b15b936 | |

5 rows × 45 columns

# DATA PROFILING:-

Data profiling is a technique used to analyse and gain a better understanding of raw data. It is the first step in determining what insights data can yield when you run it through machine learning algorithms in order to make predictions. Through data profiling, you determine whether the dataset is complete and accurate enough to solve a practical business problem. It is the very first step in preparing your data for predictive analytics, and it is essential for clarifying the structure, content (features), and relationships of your dataset for predictive modeling.

**In the <u>Data Profiling</u> section , we were able to take out the output of the trip start address .**



DATA PROFILING

Unique trip_start_address ▶

[10]  dataset.trip_start_address.unique()

```
array(['ulitsa Esenina, 3 корпус 1, Sankt-Peterburg, Russia, 194354',
       'Podrezova ulitsa, 2, Sankt-Peterburg, Russia, 197136',
       'Pushkarskiy pereulok, Sankt-Peterburg, Russia, 197101',
       'Voznesenskiy prospekt, 6, Sankt-Peterburg, Russia, 190000',
       'ulitsa Efimova, 4, Sankt-Peterburg, Russia, 190031',
       'Khersonskiy proyezd, 4A, Sankt-Peterburg, Russia, 191167',
       'ulitsa Esenina, 1 корпус 1, Sankt-Peterburg, Russia, 194354',
       'Furshtatskaya ulitsa, 21, Sankt-Peterburg, Russia, 191028',
       'ulitsa Efimova, 2, Sankt-Peterburg, Russia, 190031',
       "Koltsovo Airport (SVX), ulitsa Bahchivandji, 4, Yekaterinburg, Sverdlovskaya oblast', Russia, 620056",
       "Koltsovo Airport (SVX), ulitsa Bahchivandji, 1, Yekaterinburg, Sverdlovskaya oblast', Russia, 620056",
       "ulitsa Akademika Shvartsa, 2/1, Yekaterinburg, Sverdlovskaya oblast', Russia, 620085",
       "Vokzalnaya ulitsa, 23A, Yekaterinburg, Sverdlovskaya oblast', Russia, 620027",
       "ulitsa Akademika Shvartsa, 2/2, Yekaterinburg, Sverdlovskaya oblast', Russia, 620085",
       "Samotsvetnyy bulvar, Yekaterinburg, Sverdlovskaya oblast', Russia, 620085"
```

# DATA PREPROCESSING :-

**Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.**

**When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data pre-processing task.**



DATA PREPROCESSING

▶  dataset.dropna(inplace=True)

[20]  dataset.drop_duplicates(inplace=True)

# DATA VISUALISATION :-

Data visualization is the graphical representation of information and data. By using **visual elements like charts, graphs, and maps**, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. Additionally, it provides an excellent way for employees or business owners to present data to non-technical audiences without confusion.

In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions.

## INPUT :-

```python
obj = (dataset.dtypes == 'object')
object_cols = list(obj[obj].index)

unique_values = {}
for col in object_cols:
    unique_values[col] = dataset[col].unique().size
unique_values
```

## Output :-

```
{'trip_completed_at': 643,
 'trip_status': 2,
 'ride_hailing_app': 2,
 'trip_uid': 643,
 'driver_uid': 593,
 'rider_uid': 1,
 'customer': 1,
 'trip_start_time': 642,
 'trip_end_time': 642,
 'trip_time': 548,
 'total_time': 78,
 'wait_time': 451,
 'trip_type': 6,
 'vehicle_make_model': 119,
 'vehicle_license_plate': 1,
 'driver_name_en': 174,
 'vehicle_make': 36,
 'vehicle_model': 117,
 'driver_gender': 2,
 'driver_photo_url': 1,
 'driver_phone_number': 1,
 'trip_map_image_url': 1,
 'trip_path_image_url': 1,
 'city': 3,
 'country': 1,
 'trip_start_address': 287,
 'trip_end_address': 250,
 'price_rub': 390,
 'temperature_time': 642,
 'cloudness': 99,
 'weather_main': 9,
 'weather_desc': 13,
 'precipitation': 3}
```

# DATA VISUALISATION USING PLOTTING :-

**INPUT :-**

Make a plot for 👇

    driver_gender vs precipitation
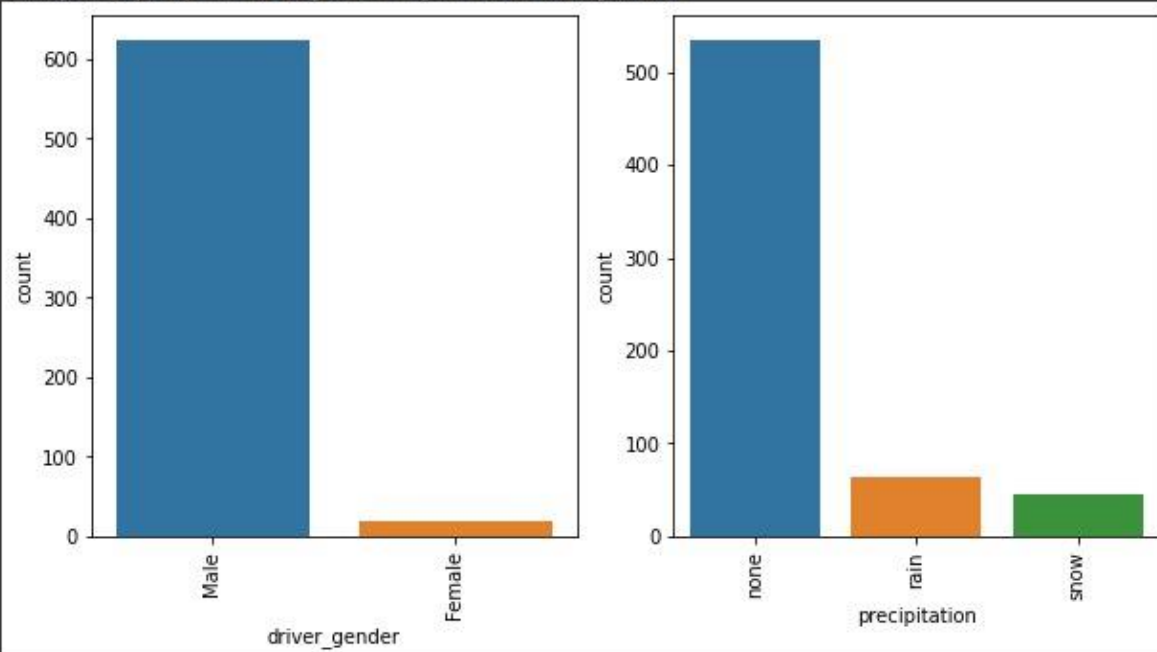
```
[26]  plt.figure(figsize=(10,5))

      plt.subplot(1,2,1)
      sns.countplot(dataset['driver_gender'])
      plt.xticks(rotation=90)

      plt.subplot(1,2,2)
      sns.countplot(dataset['precipitation'])
      plt.xticks(rotation=90)
```

**Output :-**

# DATE TIME OPERATION :-

**By converting the strings into datetimes, this exposes all the pandas dt properties.**

```
DATE TIME OPERATION

[49]  dataset['trip_start_time'] = pd.to_datetime(dataset['trip_start_time'],
              errors='coerce')
      dataset['trip_end_time'] = pd.to_datetime(dataset['trip_end_time'],
              errors='coerce')


[58]  from datetime import datetime

      dataset['date'] = pd.DatetimeIndex(dataset['trip_start_time']).date
      dataset['time'] = pd.DatetimeIndex(dataset['trip_start_time']).hour

      #changing into categories of day and night
      dataset['day-night'] = pd.cut(x=dataset['time'],
              bins = [0,10,15,19,24],
              labels = ['Morning','Afternoon','Evening','Night'])
```

# FUNCTIONAL REQUIREMENTS OF THE SYSTEM

*SOFTWARE:*

- *Operating System*
- Windows OS 11

*WEB BROWSER:*

- Internet Explorer 7
- Google Chrome

*CODING LANGUAGE :*

- Python

# <u>Conclusion :</u>

        Working with different kinds of data poses a unique challenge each time. Issues might crop up in the data values stemming from the data collection stage or the data storing/retrieval stage. One such challenge for the Uber dataset is that many location columns have NULL values or say "Unknown Location." When fewer in number, you can delete these rows. But in our case, "Unknown Location" has a high occurrence in the location columns but does not give us any knowledge or insight about the user's travel patterns. But due to their significance, those rows cannot be ignored unless the rest of the features of those rows are proven to be equally useless.

# REFERENCE

## [www.interview.projectideas](www.interview.projectideas)

## [www.geesksforgeesks.com](www.geesksforgeesks.com)

## [www.3Dsurfaceplot.com](www.3Dsurfaceplot.com)