**Mini Project 1 Report**

1. **The data and its source**

* Dataset name: Uber Request Data.
* Source: Kaggle.
* Link: [Uber Request Data.csv | Kaggle](https://www.kaggle.com/datasets/anupammajhi/uber-request-data)

1. **Description of data exploration and data cleaning steps**

|  |  |  |
| --- | --- | --- |
| **Column name** | **Data type** | **Description** |
| Request id | Int64 | Random unique id generated when requested for a ride. |
| Pickup point | Object | Pick up has 2 options for requesting ride they are:   * Airport * City |
| Driver id | Float64 | Unique id for every driver who are there in Uber data. |
| Status | Object | Status has 3 options that let us know about the trip status:   * Trip Completed * Cancelled * No Cars Available |
| Request timestamp | Datetime64 | Gives the date and time when the ride is requested. |
| Drop timestamp | Datetime64 | Gives the date and time when the ride is completed. |

* The dataset in total has 6 columns and 2001 rows.
* There are 346 null values in Driver id column and 841 null values in Drop timestamp column.
* All the null values are filled with 0.
* There are no duplicate values in the dataset.

1. **Unit of analysis**
2. **Analyzing and segregating the rides to six different categories of ride period based on ride request time.**

* Segregating the rides using if else-if loop.
* Using Request timestamp attribute segregating the rides into six types of periods they are:

1. If request time is before 5:00 AM i.e. Pre Morning.
2. If request time is between 5:00 AM to 9:00 AM i.e. Early Morning.
3. If request time is between 9:00 AM to 12:00 PM i.e. Pre Noon.
4. If request time is between 12:00 PM to 17:00 PM i.e. Day time.
5. If request time is between 17:00 PM and 21:00 PM i.e. Evening.
6. If request time is anything other than this then i.e. Late Night.

* Getting the count of the ride request during different ride periods.
* Using pivot to get the count of three different categories of status column with every request period.

1. **When is the most cancellations are occurring when the ride is requested from city and during which period do we have most Demand-supply gap when the ride is requested from airport.**

* Using pivot to get the count of three different categories of status column with both the pickup points.
* Again using pivot to get the count of the two things that I have mentioned in the question using three different columns they are:

1. Pickup point.
2. Request period.
3. Status.
4. **Description of the program**
5. Importing the packages and accessing the csv file.

**Table

Description automatically generated**

1. Knowing the number of rows and columns, checking for null values, filling null values with 0, checking for duplicate values.

Graphical user interface, text, application, email

Description automatically generated

1. Getting the request timestamp and drop timestamp column into datetime format.

Graphical user interface, text, application, table

Description automatically generated with medium confidence

1. Converting the columns in csv file to list, knowing the datatype of each column in the dataset.

Graphical user interface, text, application, email

Description automatically generated

1. Generating a new column called ride duration, getting the count of columns status with three different categories.

Graphical user interface, text, application, email

Description automatically generated

1. Analyzing and segregating the rides to six different categories of ride period based on ride request time and saving it into new csv file.

Text

Description automatically generated

1. Getting count of the ride requested for every ride period, Using pivot to get the count of three different categories of status column with every ride period.

Table

Description automatically generated with medium confidence

1. Using pivot to get the count of three different categories of status column with both the pickup points.

**A picture containing text

Description automatically generated**

1. Knowing the most cancellations are occurring when the ride is requested from city and knowing during which period do we have most Demand-supply gap when the ride is requested from airport and saving it into new csv file.

**Table

Description automatically generated**

1. **Description of the output files.**
2. Output file name: Uber\_Request\_Data\_V2.csv

|  |  |  |
| --- | --- | --- |
| **Column name** | **Data type** | **Description** |
| Request id | Int64 | Random unique id generated when requested for a ride. |
| Pickup point | Object | Pick up has 2 options for requesting ride they are:   * Airport * City |
| Driver id | Float64 | Unique id for every driver who are there in Uber data. |
| Status | Object | Status has 3 options that let us know about the trip status:   * Trip Completed * Cancelled * No Cars Available |
| Request timestamp | Datetime64 | Gives the date and time when the ride is requested. |
| Drop timestamp | Datetime64 | Gives the date and time when the ride is completed. |
| Ride duration | Datetime64 | Gives the total duration of ride by subtracting Request timestamp value from Drop timestamp value. |
| Request Period | Object | Let us know when the ride was requested out of six different categories they are:   * Pre Morning * Early Morning * Pre Noon * Day time * Evening * Late Night |

Output:

|  |  |
| --- | --- |
| **Ride period** | **Count** |
| Early Morning | 770 |
| Evening | 418 |
| Day time | 331 |
| Pre Noon | 199 |
| Pre Morning | 167 |
| Late Night | 116 |

1. Output file name: Count\_of\_Pickup\_Point\_with\_seperate\_Request\_period.csv

Trips requested from the city early in the morning are the ones that are canceled most frequently. For rides requested from the airport in the evening (17:00 to 21:00) are having the demand-supply gap (no vehicles available) is particularly noticeable.