

Project Based Learning Report

on

Uber Data Analysis

Submitted in the partial fulfillment of the requirements For
the Project based learning in (ITC-2 Essentials of Data
Science)

in

Electronics & Communication Engineering

By

Name of Students in Alphabetical order with Seat Number /PRN Number

PRN

2114110763

2114110766

Name of the Student

Md Huzaifa Jalal

Vishal Kumar Singh

Under the guidance

of Prof. Vikas

Kaduskar

Department of Electronics & Communication
Engineering Bharati Vidyapeeth
(Deemed to be University)
College of Engineering,
Pune – 4110043

Academic Year: 2023-2024

**Bharati Vidyapeeth (Deemed
to be University) College of
Engineering,
Pune – 411043**

DEPARTMENT OF ELECTRONICS & COMMUNICATION ENGINEERING

CERTIFICATE

Certified that the Project Based Learning report entitled, (**“Uber Data Analysis”**) is work done by

PRN	Name of the Student
2114110763	Md Huzaifa Jalal
2114110766	Vishal Kumar Singh

in partial fulfillment of the requirements for the award of credits for Project Based Learning (PBL) in **“ITC-2 Essentials of Data Science”** of Bachelor of Technology Semester IV, in ECE Div-II.

Date: April,2024

**Prof. Vikas Kaduskar
Course In-chargee**

**Dr. Arundhati A. Shinde
Professor & Head**

INDEX

Sr. No	Title	Page No.
1.	Problem statement	4 -5
2.	Solution to problem statement	6-7
3.	Software Used	8-8
4.	Result	9-13
5.	Outcome & Conclusion	13
6.	Appendix	15

Problem statement:-

Uber Data Analysis

Solution :-

1. Data Collection: Gather information about movies, including attributes like title, genre, director, actors, release year, plot summaries, ratings, etc. Additionally, collect data on user interactions such as ratings, watched history, and explicit preferences.
2. Data Preprocessing: Clean and preprocess the data, handling missing values, removing duplicates, and transforming text or categorical features into a suitable format for analysis.
3. Feature Engineering: Extract relevant features from the movie data, such as genre, director, actors, and plot summaries. These features are used to represent each movie in a feature space.
4. Build the Model: Develop a model that takes into account of dataset.
5. Evaluation: Assess the performance of the recommendation system using metrics such as precision, recall, or mean average precision. Evaluate how well the recommendations align with users' preferences and whether they lead to increased user engagement or satisfaction.
6. Deployment: Deploy the recommendation system in a suitable environment, such as a website, streaming platform, or mobile app. Ensure scalability and performance to handle large volumes of users and data. Overall, a movie recommendation system aims to enhance user experience by providing personalized suggestions that align with users' tastes and preferences, thereby facilitating content discovery and engagement.

What is the role of data science in Uber Data Analysis?

Data science plays a crucial role in Uber data analysis by leveraging various techniques and methodologies to extract valuable insights and drive informed decision-making. Here's how data science contributes to Uber data analysis:

1. **Data Collection and Storage:** Data scientists design and implement data collection systems to gather information from various sources such as user interactions, trip data, driver behavior, and operational metrics. They also work on storing this data efficiently in databases or data lakes.
2. **Data Cleaning and Preprocessing:** Raw data collected by Uber may contain errors, inconsistencies, or missing values. Data scientists preprocess and clean the data to ensure its quality, which involves tasks like handling missing data, removing outliers, and standardizing formats.
3. **Exploratory Data Analysis (EDA):** Data scientists conduct exploratory data analysis to understand the characteristics and patterns present in the data. They use statistical methods and visualization techniques to identify trends, correlations, and anomalies.
4. **Feature Engineering:** Feature engineering involves creating new features or transforming existing ones to improve the performance of machine learning models. Data scientists identify relevant features from Uber data that can enhance the accuracy and interpretability of predictive models.
5. **Machine Learning and Predictive Modeling:** Data scientists develop machine learning models to address various use cases within Uber, such as demand forecasting, surge pricing optimization, driver allocation, and fraud detection. They select appropriate algorithms, train models on historical data, and evaluate their performance using metrics like accuracy, precision, and recall.
6. **Optimization and Experimentation:** Data scientists collaborate with product managers and engineers to design experiments and optimize Uber's platform. They use techniques like A/B testing to evaluate the impact of new features or changes and make data-driven decisions to improve user experience and business outcomes.
7. **Insights and Decision Support:** Data scientists generate actionable insights and recommendations based on their analysis of Uber data. They communicate findings to stakeholders through reports, dashboards, or presentations, helping guide strategic decisions and operational improvements.

What are different data science techniques in the Uber Data Analysis system with examples?

In Uber's data analysis system, various data science techniques are employed to extract meaningful insights and drive decision-making. Here are some examples of these techniques:

1. Time Series Analysis:

- Example: Analyzing historical ride data to forecast future demand for rides in specific regions and time periods. This helps Uber optimize driver allocation and pricing strategies to meet anticipated demand.

2. Geospatial Analysis:

- Example: Using GPS coordinates from trip data to visualize ride patterns on a map. Geospatial analysis helps Uber identify high-demand areas, optimize driver routes, and plan the expansion of its services into new markets.

3. Predictive Modeling:

- Example: Building machine learning models to predict the estimated time of arrival (ETA) for rides. By considering factors such as traffic conditions, route distance, and historical data, Uber can provide accurate ETAs to users and improve overall customer satisfaction.

4. Customer Segmentation:

- Example: Segmenting Uber users based on their ride preferences, frequency of usage, and demographics. This enables targeted marketing campaigns, personalized promotions, and tailored product offerings to different customer segments.

5. Anomaly Detection:

- Example: Identifying unusual patterns or behaviors in ride data that may indicate fraudulent activity or system errors. Anomaly detection helps Uber detect and prevent fraudulent transactions, ensuring the safety and integrity of its platform.

6. Natural Language Processing (NLP):

- Example: Analyzing user feedback and reviews submitted through the Uber app to extract sentiment and identify common issues or concerns. NLP techniques enable Uber to understand customer feedback at scale and take proactive measures to address user needs and improve service quality.

7. Optimization Algorithms:

- Example: Developing algorithms to optimize driver dispatching and routing decisions in real-time. By considering factors such as driver location, ride requests, and traffic conditions, Uber

can minimize wait times for users and increase driver efficiency.

8. Market Basket Analysis:

- Example: Analyzing patterns of co-occurring ride requests to identify common trip routes or travel patterns. Market basket analysis helps Uber understand user behavior and preferences, informing decisions related to service expansion, pricing strategies, and promotional offers.

These techniques, among others, are essential components of Uber's data analysis system, enabling the company to leverage its rich data ecosystem for strategic planning, operational optimization, and enhancing the overall user experience.

DATASET:-

We have downloaded dataset from NYC TLC trip record DATASET

We have performed analysis visulization on VS CODE

Library used:-

For mathematical computation:-

Numpy library - numpy is used to perform various mathematical operations on arrays.

Pandas Library - pandas provides various data structures and operations for manipulating numerical data and time series.

Scipy-stats - All of the statistics functions are located in the sub-package scipy.stats and a fairly complete listing of these functions can be obtained using info(stats) function. A list of random variables available can also be obtained from the docstring for the stats sub-package.

For data visuallisation:-

Matplotlib library from which pyplot module is used for plotting library used for 2D

graphics. Seaborn library - seaborn is a library for making statistical graphics in Python.

Plotly - Plotly is a Montreal based technical computing company involved in development of data analytics and visualisation tools such as Dash and Chart Studio. It has also developed open source graphing Application Programming Interface (API) libraries for Python.

Software used is VS CODE

Visual Studio Code (VS Code) is a popular integrated development environment (IDE) that can be used for developing various types of software applications, including those related to data science and machine learning, such as movie recommendation systems. Here's a brief description of how VS Code can be used for this problem statement:

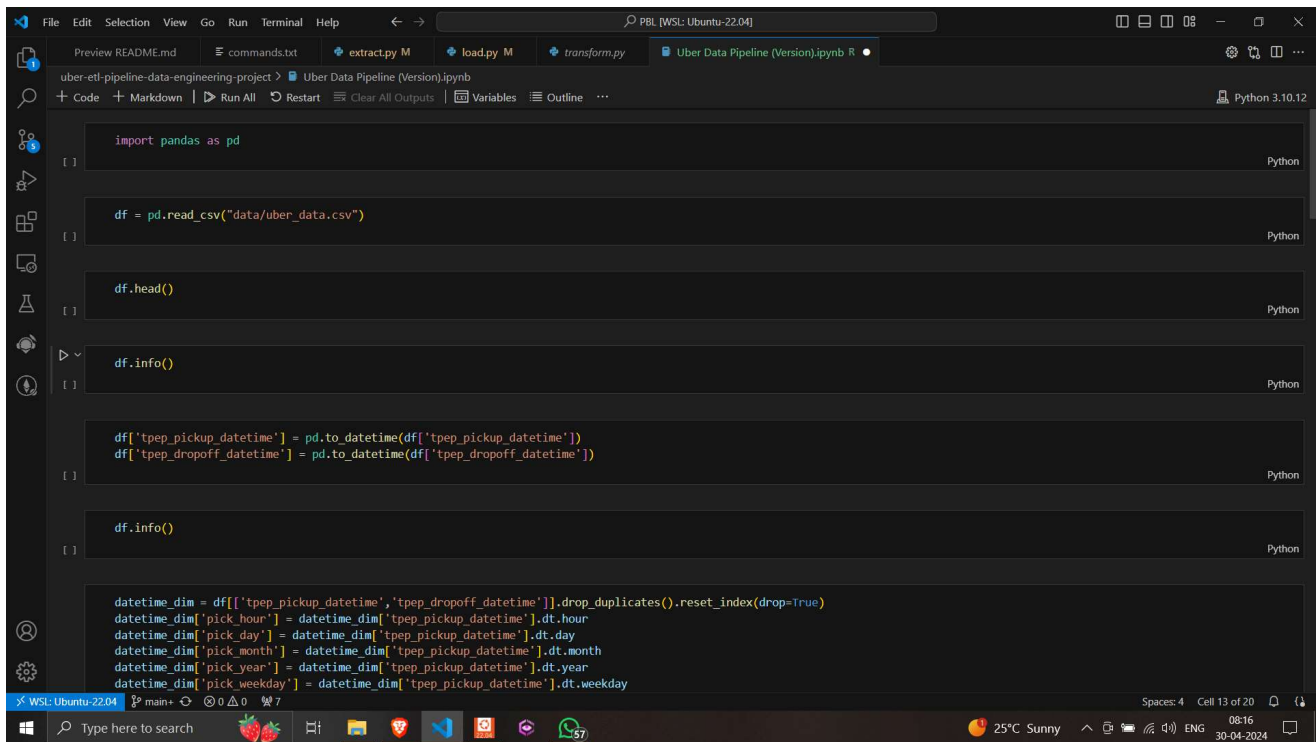
1. **Code Editing:** VS Code provides a feature-rich code editor with syntax highlighting, IntelliSense (code completion), and code navigation features. Data scientists can write and edit Python code for tasks such as data preprocessing, feature engineering, model development, and evaluation.
2. **Integrated Terminal:** VS Code includes an integrated terminal, allowing data scientists to execute Python scripts, run machine learning models, and interact with the Python interpreter directly within the IDE.
3. **Extensions:** VS Code has a vast ecosystem of extensions that can enhance its functionality for data science tasks. Extensions like Python, Jupyter, and GitLens are commonly used by data scientists for Python development, Jupyter notebook support, and version control, respectively.
4. **Debugging:** VS Code provides built-in debugging support for Python code, allowing data scientists to set breakpoints, inspect variables, and step through code to identify and fix issues in their machine learning models or data processing pipelines.
5. **Version Control:** VS Code integrates seamlessly with version control systems like Git, enabling data scientists to manage their code repositories, collaborate with team members, and track changes to their codebase.
6. **Notebook Support:** VS Code supports Jupyter notebooks, which are widely used by data scientists for interactive data analysis and experimentation. Data scientists can create, edit, and run Jupyter notebooks directly within VS Code, leveraging its features for code editing, debugging, and version control.
7. **Visualization:** VS Code supports various data visualization libraries and tools, such as Matplotlib and Plotly, allowing data scientists to create visualizations to explore and analyze data, evaluate model performance, and communicate insights effectively.
8. **Integration with Data Science Libraries:** VS Code can be configured to work seamlessly with popular data science libraries and frameworks like NumPy, pandas, scikit-learn, TensorFlow, and PyTorch, enabling data scientists to leverage these tools for tasks such as data manipulation, machine learning model development, and deep learning.

Overall, VS Code provides a versatile and feature-rich environment for data scientists to develop, debug, and deploy machine learning models and data science applications, including movie recommendation systems. Its extensibility, integration with data science tools, and support for various programming languages make it a preferred choice for many data scientists and machine learning practitioners.

Result with analysis

Analysis of the code: -

- First, we Import the libraries
- Secondly, Download the dataset and add that to the path to load the dataset. we use panda library and used head() function for displaying first five row of dataset.
- We get more information by using df.info().then for checking null values in dataset we used is null() function.
- Then we find the graph of number of time charted by artist by using px.bar() function.
- Then we create a correlation using heatmap()
- Then we use the library plotly to plot the graph of danceability by use px.line().
- Then we plot graph by using px.bar()
- At last we use pandas library to get information about genre and plot the pie chart.



The image shows a Jupyter Notebook titled "Uber Data Pipeline (Version).ipynb" open in VS Code. The notebook contains the following Python code:

```
import pandas as pd

df = pd.read_csv("data/uber_data.csv")

df.head()

df.info()

df['tpep_pickup_datetime'] = pd.to_datetime(df['tpep_pickup_datetime'])
df['tpep_dropoff_datetime'] = pd.to_datetime(df['tpep_dropoff_datetime'])

df.info()

datetime_dim = df[['tpep_pickup_datetime', 'tpep_dropoff_datetime']].drop_duplicates().reset_index(drop=True)
datetime_dim['pick_hour'] = datetime_dim['tpep_pickup_datetime'].dt.hour
datetime_dim['pick_day'] = datetime_dim['tpep_pickup_datetime'].dt.day
datetime_dim['pick_month'] = datetime_dim['tpep_pickup_datetime'].dt.month
datetime_dim['pick_year'] = datetime_dim['tpep_pickup_datetime'].dt.year
datetime_dim['pick_weekday'] = datetime_dim['tpep_pickup_datetime'].dt.weekday
```

The code is executed in a series of cells, with the output of each cell visible below the code. The output of the first cell is "Python". The output of the second cell is "Python". The output of the third cell is "Python". The output of the fourth cell is "Python". The output of the fifth cell is "Python". The output of the sixth cell is "Python". The output of the seventh cell is "Python". The output of the eighth cell is "Python". The output of the ninth cell is "Python". The output of the tenth cell is "Python". The output of the eleventh cell is "Python". The output of the twelfth cell is "Python". The output of the thirteenth cell is "Python". The output of the fourteenth cell is "Python". The output of the fifteenth cell is "Python". The output of the sixteenth cell is "Python". The output of the seventeenth cell is "Python". The output of the eighteenth cell is "Python". The output of the nineteenth cell is "Python". The output of the twentieth cell is "Python". The output of the twenty-first cell is "Python". The output of the twenty-second cell is "Python". The output of the twenty-third cell is "Python". The output of the twenty-fourth cell is "Python". The output of the twenty-fifth cell is "Python". The output of the twenty-sixth cell is "Python". The output of the twenty-seventh cell is "Python". The output of the twenty-eighth cell is "Python". The output of the twenty-ninth cell is "Python". The output of the thirtieth cell is "Python". The output of the thirty-first cell is "Python". The output of the thirty-second cell is "Python". The output of the thirty-third cell is "Python". The output of the thirty-fourth cell is "Python". The output of the thirty-fifth cell is "Python". The output of the thirty-sixth cell is "Python". The output of the thirty-seventh cell is "Python". The output of the thirty-eighth cell is "Python". The output of the thirty-ninth cell is "Python". The output of the fortieth cell is "Python". The output of the forty-first cell is "Python". The output of the forty-second cell is "Python". The output of the forty-third cell is "Python". The output of the forty-fourth cell is "Python". The output of the forty-fifth cell is "Python". The output of the forty-sixth cell is "Python". The output of the forty-seventh cell is "Python". The output of the forty-eighth cell is "Python". The output of the forty-ninth cell is "Python". The output of the fiftieth cell is "Python". The output of the fifty-first cell is "Python". The output of the fifty-second cell is "Python". The output of the fifty-third cell is "Python". The output of the fifty-fourth cell is "Python". The output of the fifty-fifth cell is "Python". The output of the fifty-sixth cell is "Python". The output of the fifty-seventh cell is "Python". The output of the fifty-eighth cell is "Python". The output of the fifty-ninth cell is "Python". The output of the sixtieth cell is "Python". The output of the sixty-first cell is "Python". The output of the sixty-second cell is "Python". The output of the sixty-third cell is "Python". The output of the sixty-fourth cell is "Python". The output of the sixty-fifth cell is "Python". The output of the sixty-sixth cell is "Python". The output of the sixty-seventh cell is "Python". The output of the sixty-eighth cell is "Python". The output of the sixty-ninth cell is "Python". The output of the seventieth cell is "Python". The output of the seventy-first cell is "Python". The output of the seventy-second cell is "Python". The output of the seventy-third cell is "Python". The output of the seventy-fourth cell is "Python". The output of the seventy-fifth cell is "Python". The output of the seventy-sixth cell is "Python". The output of the seventy-seventh cell is "Python". The output of the seventy-eighth cell is "Python". The output of the seventy-ninth cell is "Python". The output of the eightieth cell is "Python". The output of the eighty-first cell is "Python". The output of the eighty-second cell is "Python". The output of the eighty-third cell is "Python". The output of the eighty-fourth cell is "Python". The output of the eighty-fifth cell is "Python". The output of the eighty-sixth cell is "Python". The output of the eighty-seventh cell is "Python". The output of the eighty-eighth cell is "Python". The output of the eighty-ninth cell is "Python". The output of the ninetieth cell is "Python". The output of the ninety-first cell is "Python". The output of the ninety-second cell is "Python". The output of the ninety-third cell is "Python". The output of the ninety-fourth cell is "Python". The output of the ninety-fifth cell is "Python". The output of the ninety-sixth cell is "Python". The output of the ninety-seventh cell is "Python". The output of the ninety-eighth cell is "Python". The output of the ninety-ninth cell is "Python". The output of the hundredth cell is "Python".

Fig.: App Code

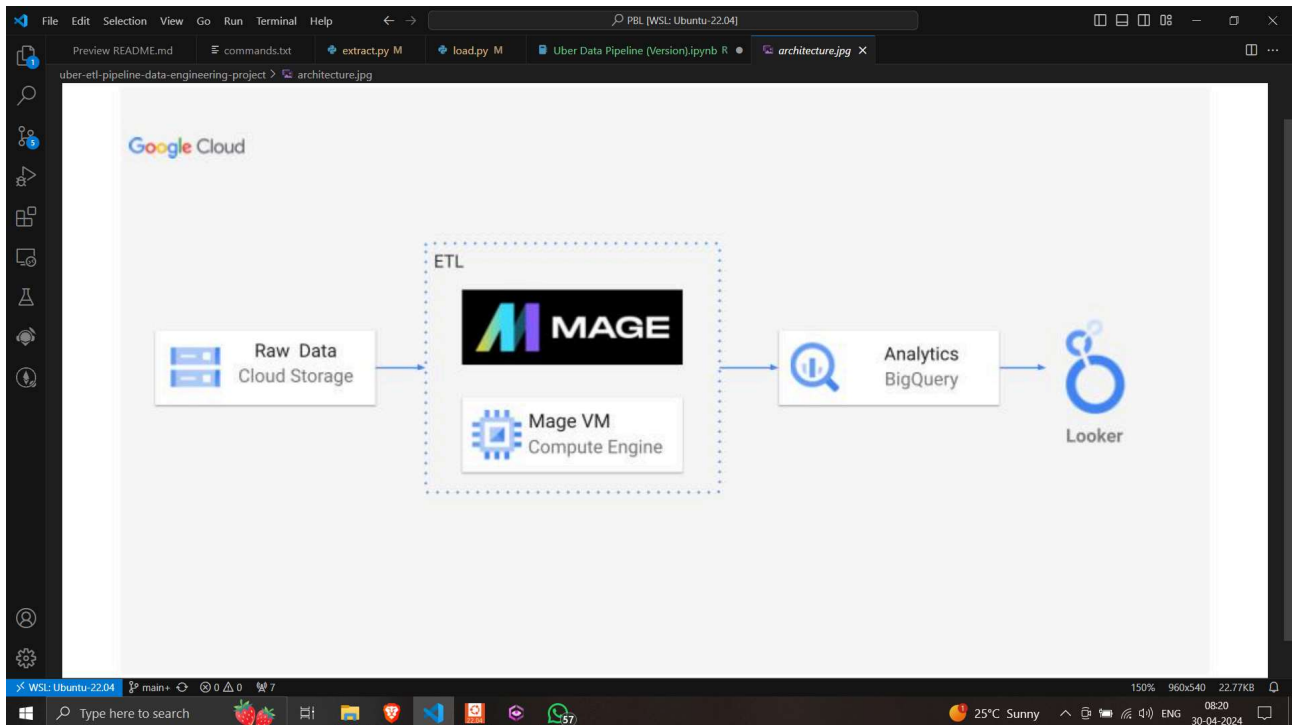


Fig.: Architecture

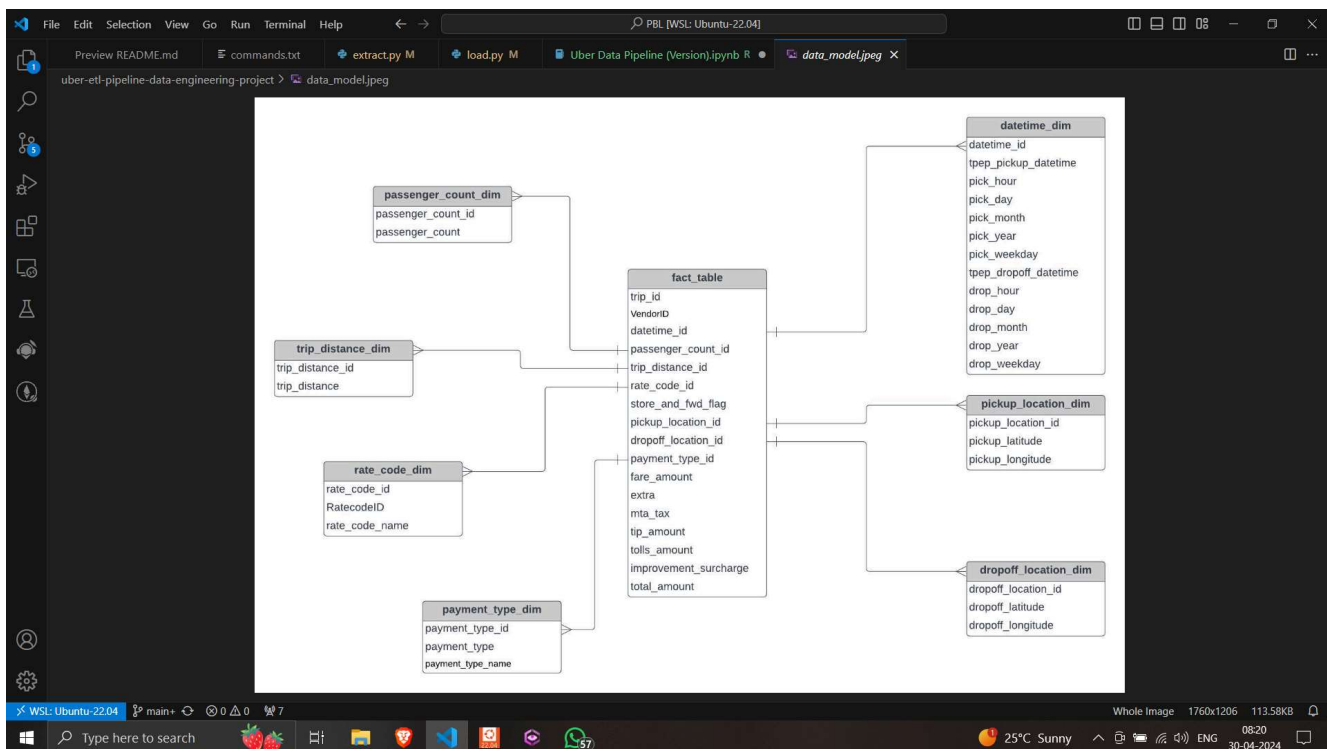


Fig.: Uber Data Set

Output

WSL: Ubuntu-22.04

```
fact_table = df.merge(passenger_count_dim, on='passenger_count') \
               .merge(trip_distance_dim, on='trip_distance') \
               .merge(rate_code_dim, on='RatecodeID') \
               .merge(pickup_location_dim, on=['pickup_longitude', 'pickup_latitude']) \
               .merge(dropoff_location_dim, on=['dropoff_longitude', 'dropoff_latitude']) \
               .merge(datetime_dim, on=['tpep_pickup_datetime', 'tpep_dropoff_datetime']) \
               .merge(payment_type_dim, on='payment_type') \
               [['VendorID', 'datetime_id', 'passenger_count_id',
                 'trip_distance_id', 'rate_code_id', 'store_and_fwd_flag', 'pickup_location_id', 'dropoff_location_id',
                 'payment_type_id', 'fare_amount', 'extra', 'mta_tax', 'tip_amount',
                 'improvement_surcharge', 'total_amount']]
```

fact_table

	VendorID	datetime_id	passenger_count_id	trip_distance_id	rate_code_id	store_and_fwd_flag	pickup_location_id	dropoff_location_id	payment_type_id	fare_amount	extra	mta_tax	tip_amo
0	1	0	0	0	0	N	0	0	0	9.0	0.5	0.5	2
1	1	1	0	1	1	N	1	1	0	11.0	0.5	0.5	3
2	2	2	1	2	0	N	2	2	0	54.5	0.5	0.5	8
3	2	3	2	3	0	N	3	3	0	31.5	0.0	0.5	3
4	2	3	3	4	1	N	4	4	0	98.0	0.0	0.0	0
...
99995	1	99848	0	19	0	N	98050	98412	1	5.0	0.0	0.5	0
99996	1	99849	0	71	0	N	98051	98413	0	14.0	0.0	0.5	2
99997	1	99850	0	295	0	N	98052	98414	0	29.0	0.0	0.5	8
99998	2	99851	0	152	0	N	98053	98415	0	5.5	0.5	0.5	1
99999	1	99852	0	29	0	N	98054	98416	1	6.0	0.0	0.5	0

WSL: Ubuntu-22.04

architecture.jpg

```
graph LR
    RawData[Raw Data Cloud Storage] --> ETL
    subgraph ETL
        MAGE[MAGE]
        MageVM[Mage VM Compute Engine]
    end
    ETL --> Analytics[Analytics BigQuery]
    Analytics --> Looker[Looker]
```

WSL: Ubuntu-22.04

Project Conclusion:

From this project, we gained the knowledge of software VS code. We learnt to analyse the datasets and afterwards, visualizing them. We learnt about various plots .

Outcome:

From this project, we learnt to describe a flow process for data science problems and classified data science problems into standard typology. We also learnt about correlating results to the solution approach followed and assessing the solution approach.

APPENDIX:

1. Github link: <https://github.com/this-vishalsingh/UberDataAnalysis>