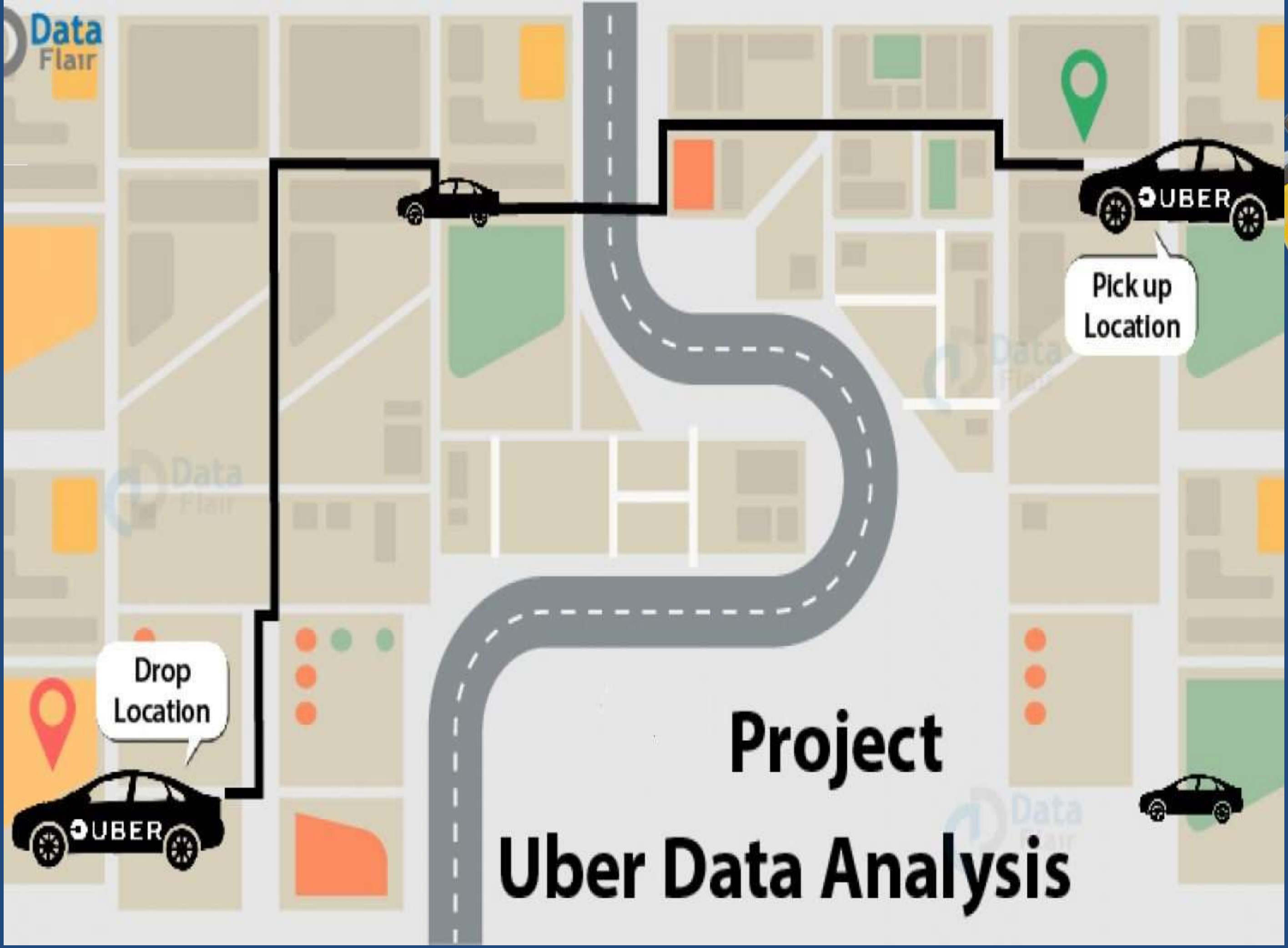
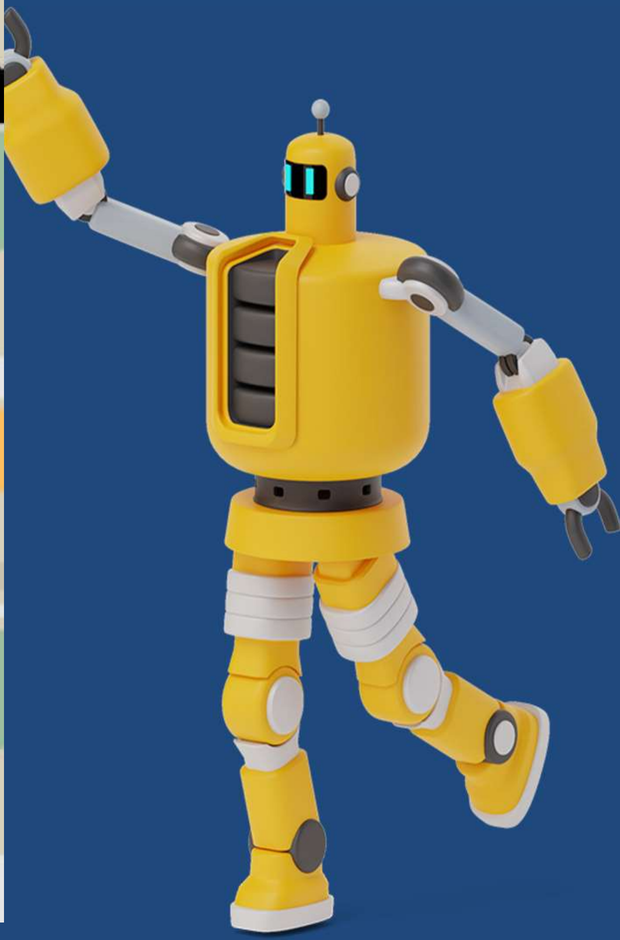
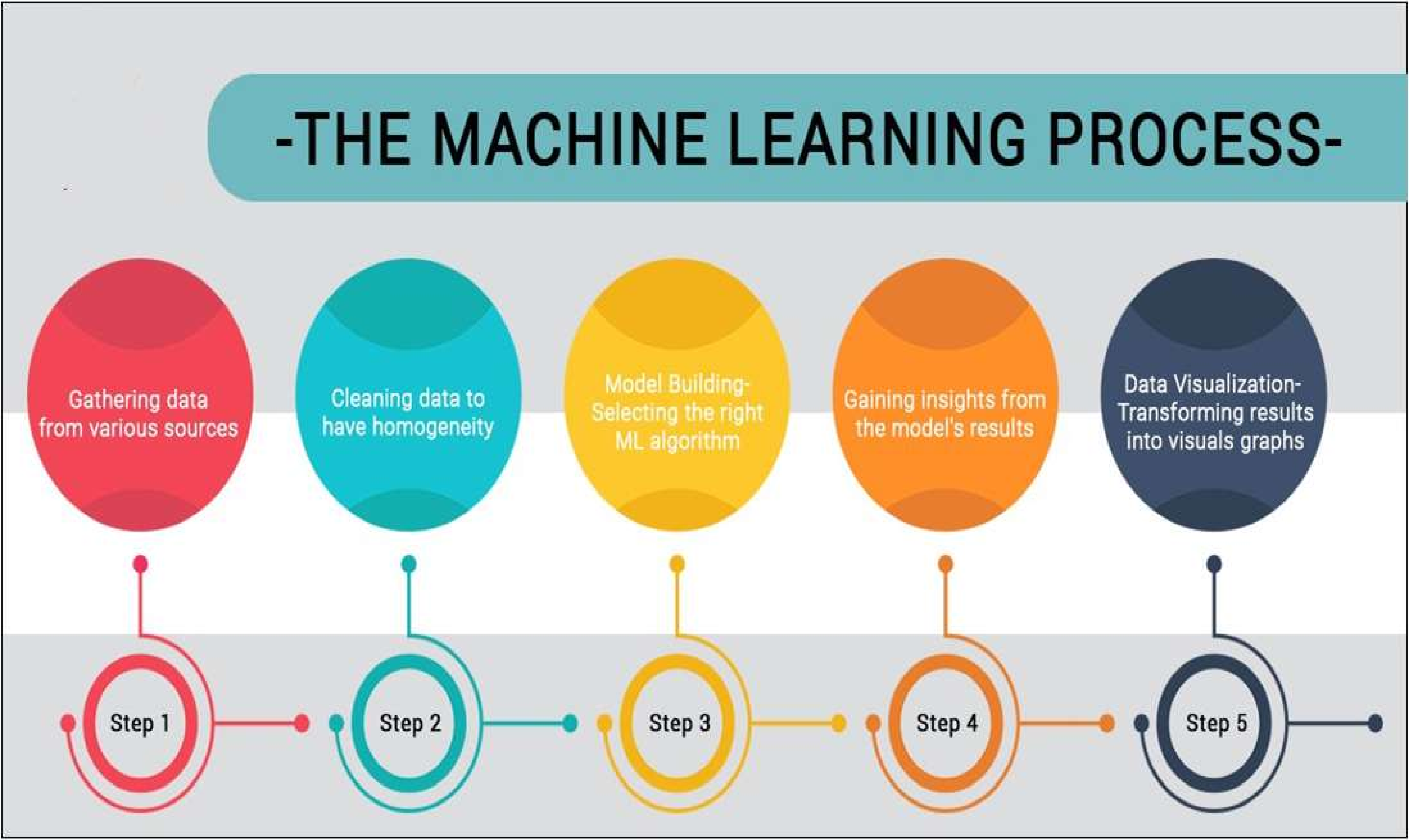

Data Flair
UBER
Pick up Location
Drop Location
UBER
Project
Uber Data Analysis



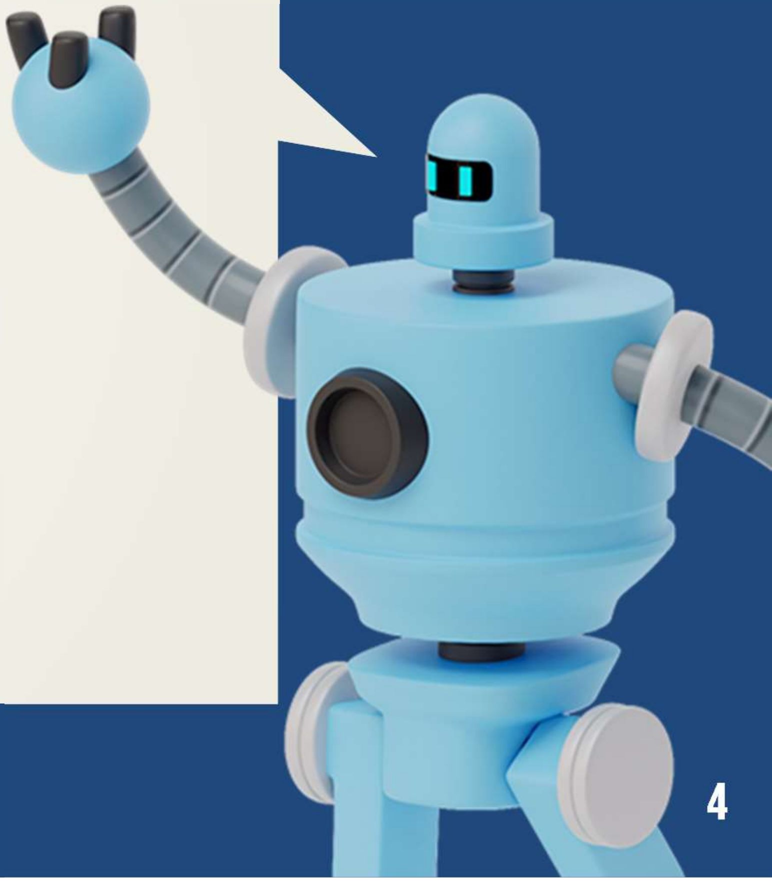


# -THE MACHINE LEARNING PROCESS-

- Step 1: Gathering data from various sources
- Step 2: Cleaning data to have homogeneity
- Step 3: Model Building- Selecting the right ML algorithm
- Step 4: Gaining insights from the model's results
- Step 5: Data Visualization- Transforming results into visuals graphs

## Aim

Complete Data Analysis and Exploration of Uber Dataset

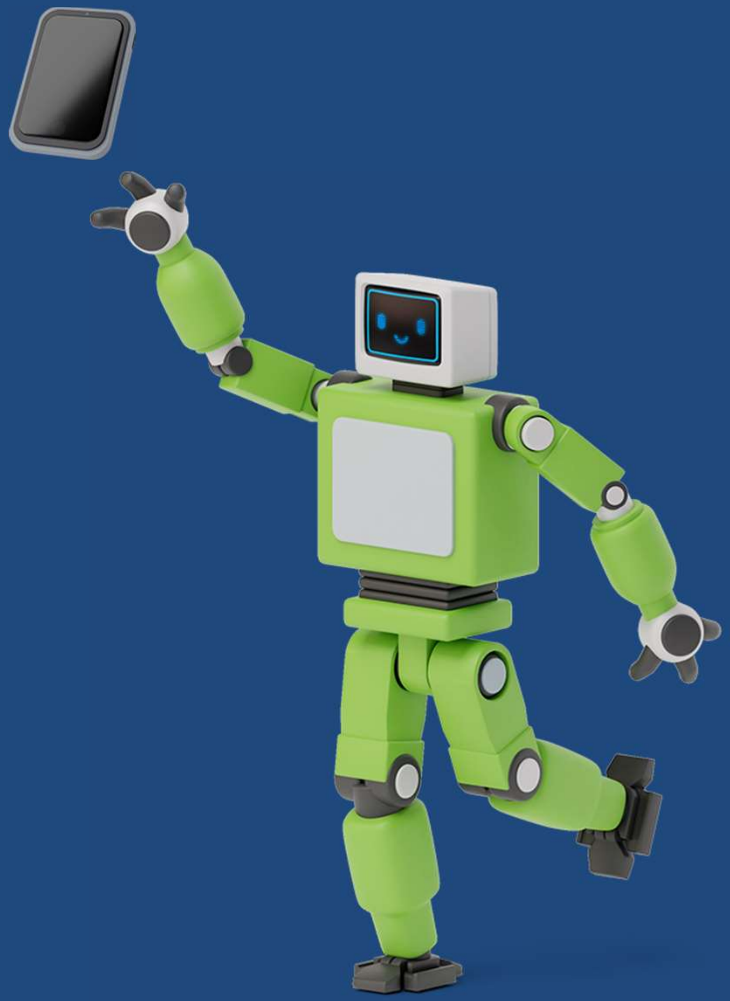


## Data Set:

Kaggle
CSV Format

## Shape:

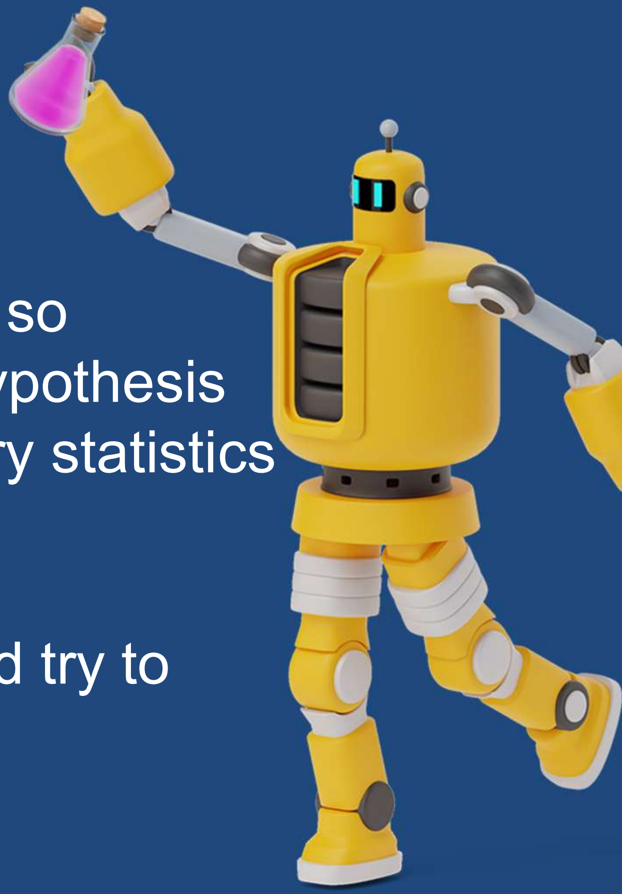
322844,56



# Libraries

NumPy
matplotlib
Pandas
scikit
learn
Seaborn

# Exploratory Data Analysis



Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns to spot anomalies to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

It is a good practice to understand the data first and try to gather as many insights from it.

EDA is all about making sense of data in hand.

# Feature Engineering

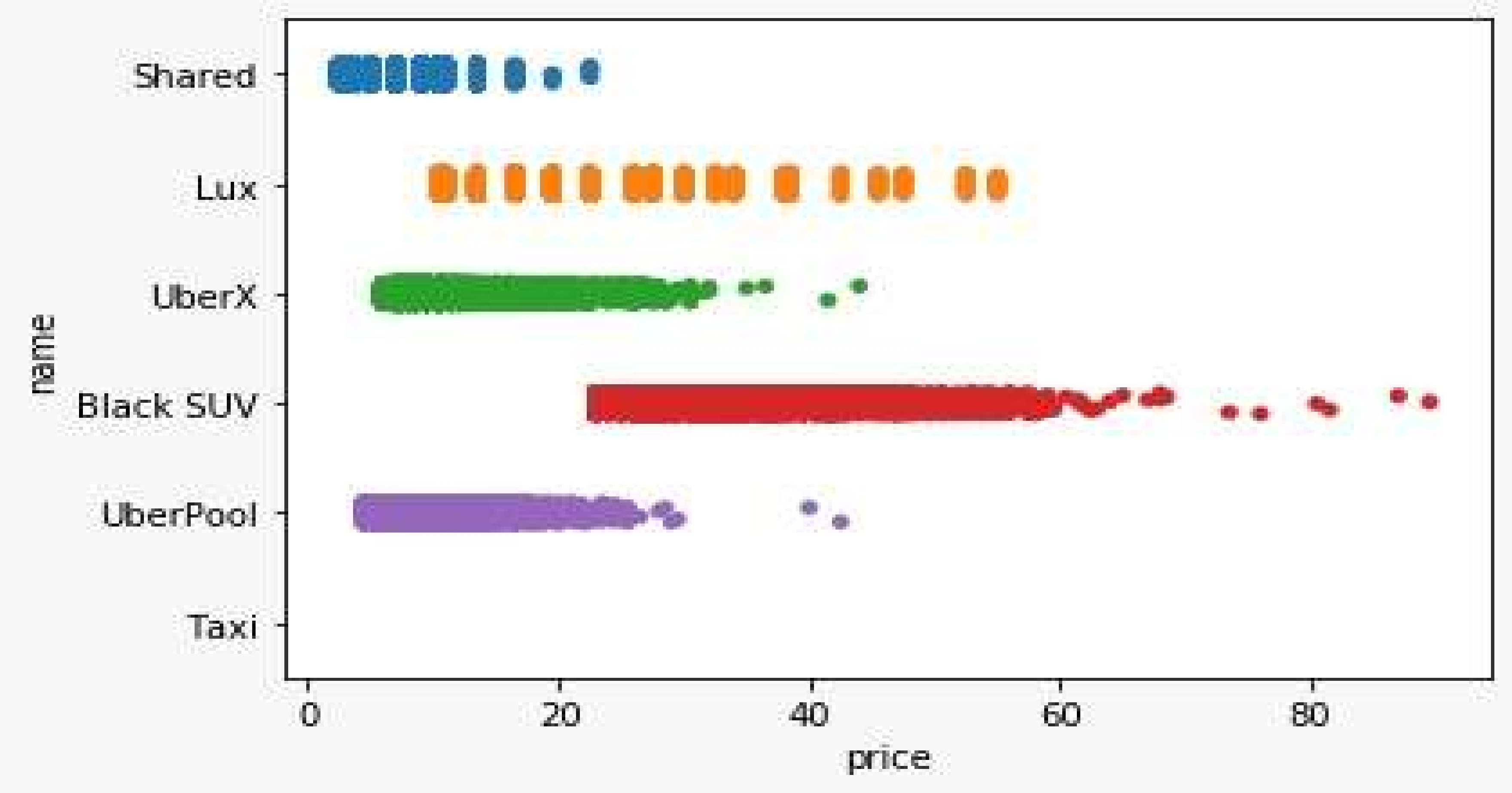
All machine learning algorithms use some input data to create outputs. This input data comprise features, which are usually in the form of structured columns. Algorithms require features with some specific characteristics to work properly and so the need for feature engineering arises.

I think feature engineering efforts mainly have two goals:

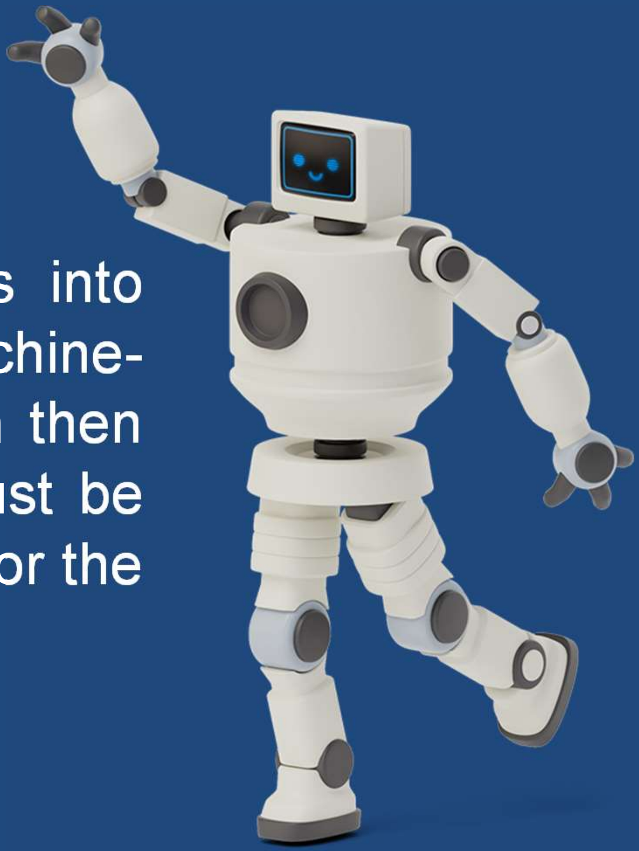
➢Preparing the proper input dataset, compatible with the machine learning algorithm requirements.

➢Improving the performance of machine learning models.

# Strip-plot between Name and Price



# Label Encoding



Label Encoding refers to converting the labels into numeric form so as to convert it into the machine-readable form. Machine learning algorithms can then decide in a better way on how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning.

# **NANs**(missing values)

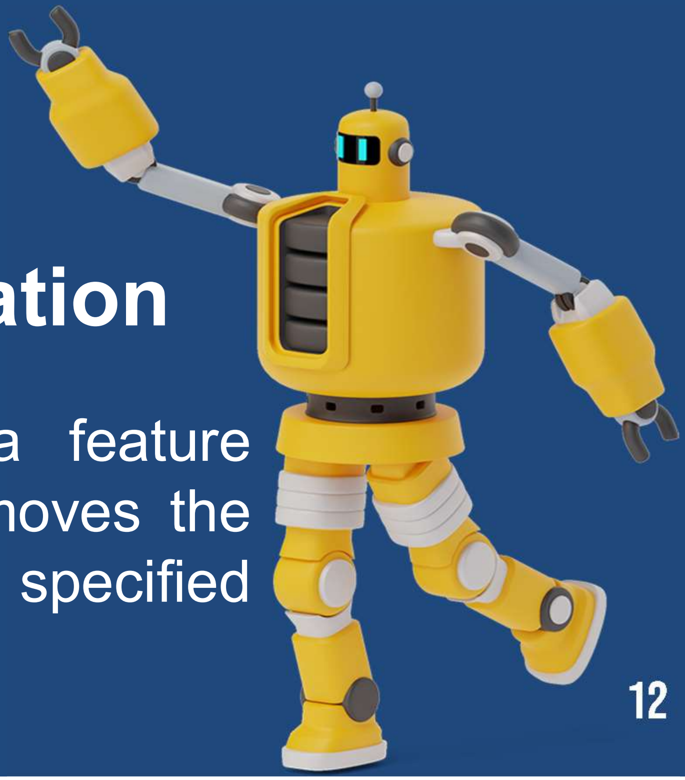
Our data set contain NANs only in Price column.

The count of nans is: 55095

The nans is filled with the median of other values.

# Feature Selection

Feature Selection is the process of selecting a subset of relevant feature (variables, predictors) for use in model construction.



## Recursive Feature Elimination

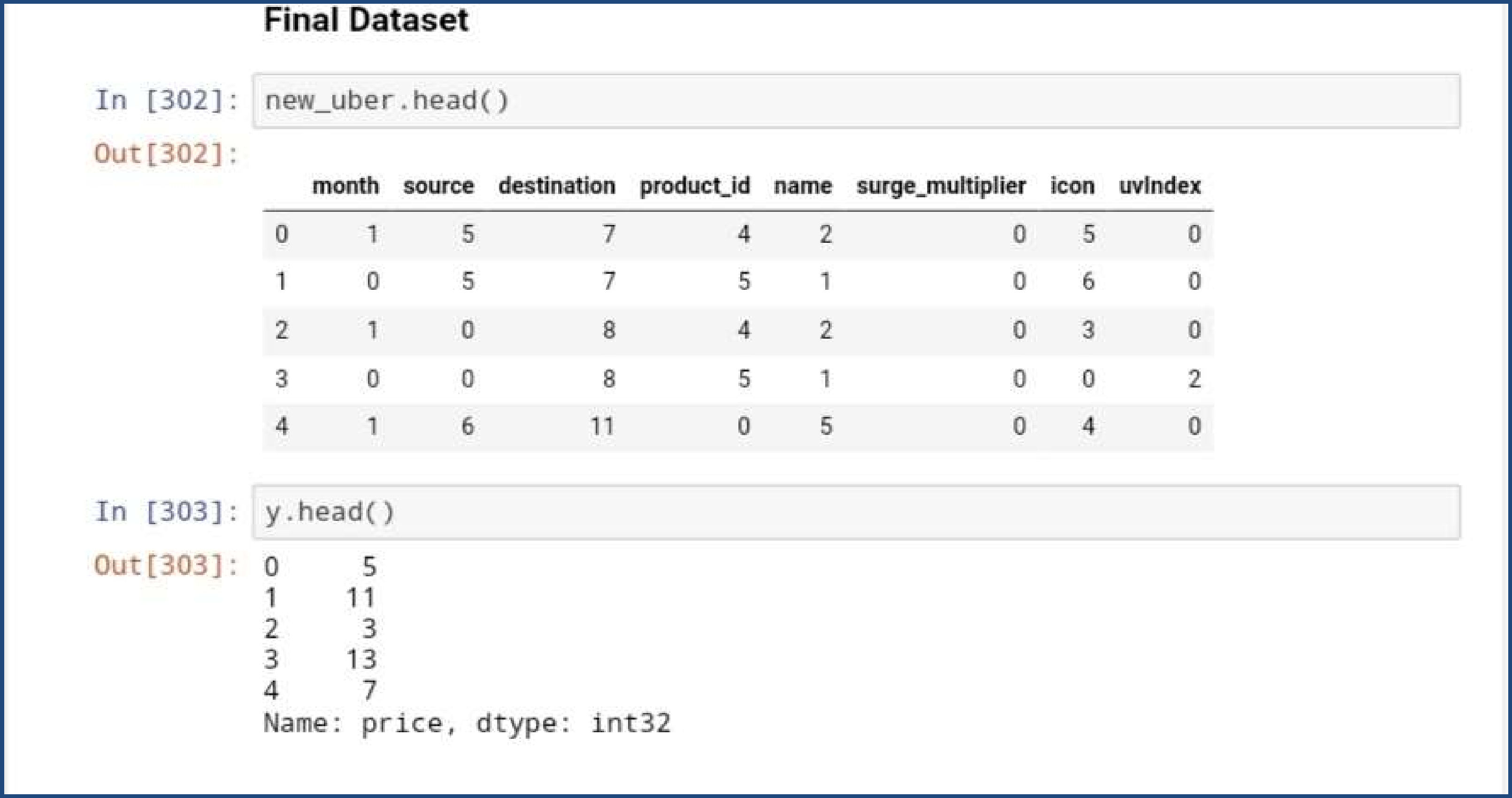
:

Recursive feature elimination (RFE) is a feature selection method that fits a model and removes the weakest feature (or features) until the specified number of features is reached.

After applying RFE on given data set with Linear Regression, we found accuracy with different no of columns as follows:-

| Serial No. | No. of Feature | Accuracy |
|---|---|---|
| 1 | 56 | 0.805483422 |
| 2 | 40 | 0.8050662132 |
| 3 | 25 | 0.80553551515 |
| 4 | 15 | 0.8050457819 |

# Final Dataset



### Final Dataset

```
In [302]: new_uber.head()
```

Out[302]:

| | month | source | destination | product_id | name | surge_multiplier | icon | uvIndex |
|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 5 | 7 | 4 | 2 | 0 | 5 | 0 |
| 1 | 0 | 5 | 7 | 5 | 1 | 0 | 6 | 0 |
| 2 | 1 | 0 | 8 | 4 | 2 | 0 | 3 | 0 |
| 3 | 0 | 0 | 8 | 5 | 1 | 0 | 0 | 2 |
| 4 | 1 | 6 | 11 | 0 | 5 | 0 | 4 | 0 |

```
In [303]: y.head()
```

```
Out[303]: 0      5
          1     11
          2      3
          3     13
          4      7
          Name: price, dtype: int32
```

# Modeling

After Completion of Recursive Feature Elimination process, we can done Modeling on final dataset.

**Linear Regression**: Linear Regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. It’s used to predict values within a continuous range.

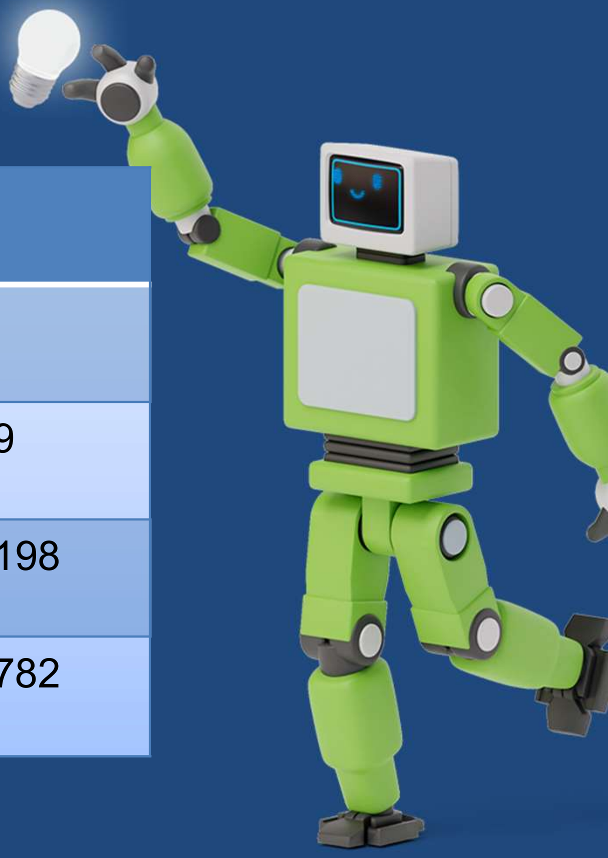
**Decision Tree**: A decision tree is a graphical representation of all the possible solutions to a decision based on certain conditions.

**Random Forest:** Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of **ensemble learning,** which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

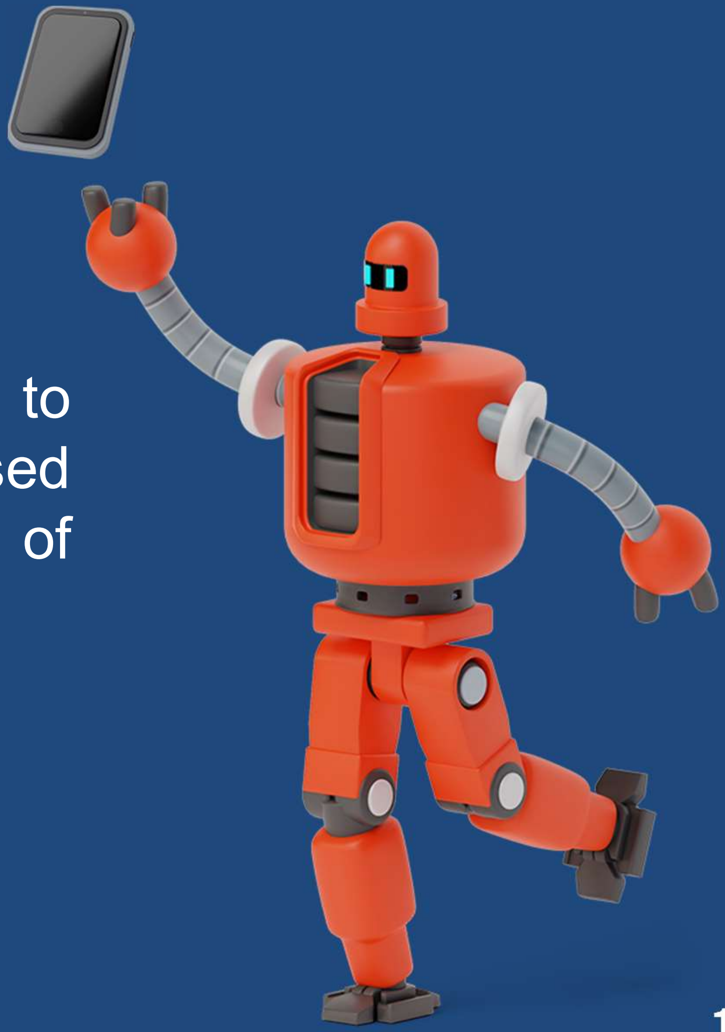
**Gradient Boosting Regressor** :Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

After applying different models on final dataset, we found different accuracy as given below :-

| Serial No. | Models | Accuracy |
|---|---|---|
| 1 | Linear Regression | 0.74754507316 |
| 2 | Decision Tree | 0.961791729999 |
| 3 | Random Forest | 0.96226947434198 |
| 4 | Gradient Boosting Regressor | 0.96318719462782 |



# Testing



The usage of the word "testing" in relation to machine learning models is primarily used for testing the model performance in terms of accuracy/precision of the model.

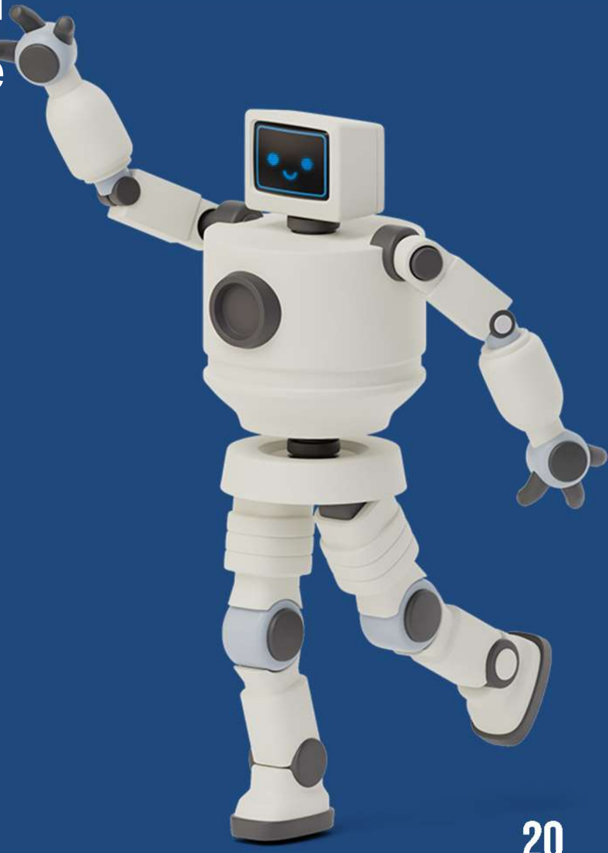
With the help of linear regression and random forest models, we predict the price, plot a graph between actual and predicted values and find the following errors:-

For linear regression:-

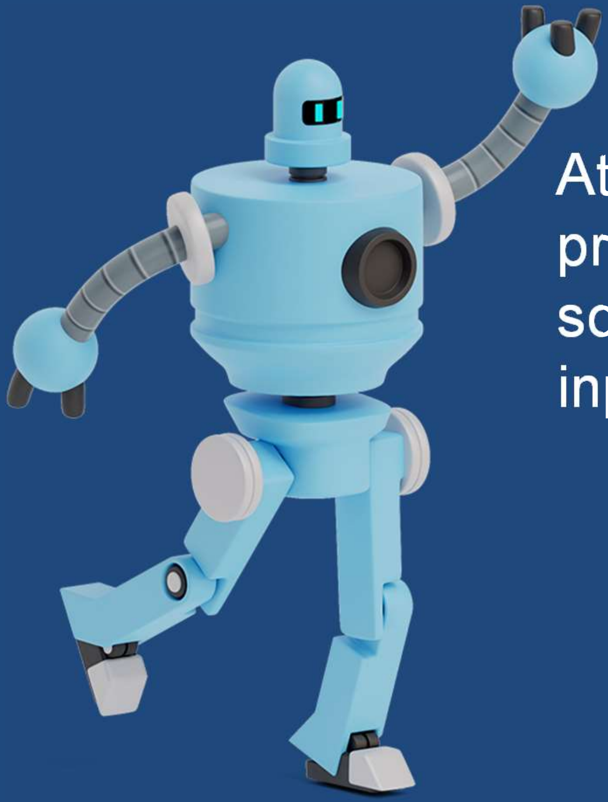
- MAE : 3.406077219
- MSE : 20.03343709
- RMAE : 4.47587277

For random forest:-

- MAE : 0.998137009
- MSE : 2.944653619
- RMAE : 1.71599930



# Price Prediction Function



At last, we create a function for price prediction which take cab name, source, surge multiplier and icon as input and predict the price.

Thank you