



DS & ML Project

- Statement: Train ticket Price Prediction within Spanish Rail setup

- Algorithm used: Polynomial Regression based on MLR.





Problem Statement

In the given data, a price monitoring system is described with columns such as insert_date, origin, destination, travel dates, train type, class, fare, and price. The objective is to predict the best train ticket prices by performing data wrangling to clean the data , data visualization , and applying machine learning algorithms like Linear and Polynomial Regression to derive accurate results.



Analysis on Acquiring and Injestion/ETL

A blue-toned background image of a financial candlestick chart. The chart features several green candlesticks representing price movements. Overlaid on the chart are technical analysis tools: a green parabolic curve, a green line with a label '61.6%: 99.19' indicating a retracement, and two green boxes with numerical values '104.19' and '86.72'.

Exploratory Data Analysis

DATA WRANGLING

Here , we have cleaned the data by removing all the nulls values and also the unwanted columns.

Exploratory Data Analysis

VISUALIZATION

By visualizing different graphs we are able to find the maximum no. of trains , majority of people preferred origin and destination places and also the ticket price and fare type.



Question 01: Explain the summary statistics for the data set

Answer:

In this dataset there are more categorical variables than numerical variables. Hence the numerical summary is given for price column. We can also notice that the majority of the origin and destination place is MADRID. The top train_class is Turista and fare is Promo.



Question 02: What insights do you get from the above plot?

Answer :

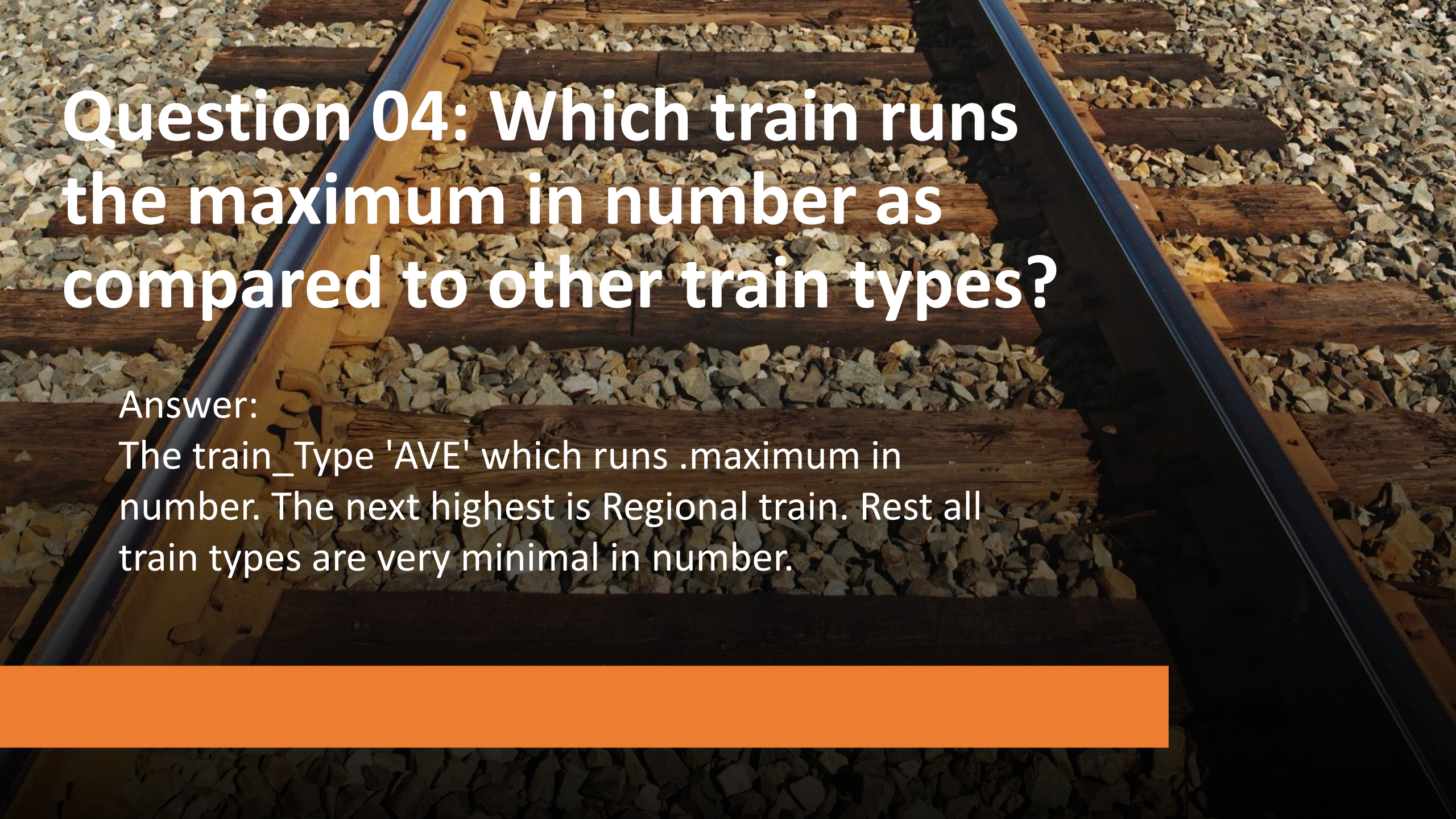
There are more people boarding from Madrid Station. The next highest is Barcelona station , Valencia station and Sevilla station. Ponferrada Station has less number of people boarding.



Question 03: What insights do you get from the graphs?

Answer:


Majority of people's destination is Madrid. Next is Barcelona and Valencia which is almost equal. Least is Ponferrada station.



Question 04: Which train runs the maximum in number as compared to other train types?

Answer:

The train_Type 'AVE' which runs .maximum in number. The next highest is Regional train. Rest all train types are very minimal in number.



Question 05: Which the most common train class for traveling among people in general?

Answer:

The train_class 'Turista' is the most common class for travelling among people in general.

A blurred high-speed train, likely a Spanish AVE, is shown in motion at a train station platform. The train is red and white with a yellow and blue stripe. The platform is visible on the right, and the background is a blurred station scene.

Question 06: Which type of trains cost more as compared to others?

Answer:

The train types such as 'AVE', 'REGIONAL', 'ALVIA' costs more compared to others.



**Which model gives
the best result for
price prediction?
Find out the
complexity using R2
score and give your
answer.**

Answer:

The R2 score for the testing
and training data is almost
same around 88 percent.
Comparatively the testing
data is best for price
prediction

CONCLUSION:

We have cleaned the dataset using the data wrangling techniques. We have viewed the dataset using different graphs using data visualization method. Then the dataset is separated into training and testing data. Finally we have applied the machine learning algorithm(Linear Regression) and predicted the R^2 score.