ENG:Salma Gamal Ahmed ENG: AL -AMIR HASSAN

Omar Morshdy - 211001749Mohamed Ashraf Qushta - 211001221Omar Reda - 19200053Dima Ayad - 211000081Philopater Ayman - 211001925

**Predicting House Prices**

*Dr. Khaled Mohammed Fouad Ibrahim*

**Contents**

[**Intro****duction** 3](#_Toc132491541)

[**Data Description** 3](#_Toc132491542)

[**Data Cleaning** 4](#_Toc132491543)

[**Exploratory Data Analysis (EDA):** 4](#_Toc132491544)

[**Modelling:** 5](#_Toc132491545)

[**Results:** 6](#_Toc132491546)

[**Conclusion:** 6](#_Toc132491547)

# **Introduction**

# The real estate market is a vital component of the global economy, with billions of dollars in transactions taking place annually. A significant challenge for buyers and sellers in this market is accurately predicting the prices of houses. Machine learning models offer a promising solution to this challenge, providing the ability to predict house prices based on existing data. Our project aims to develop a machine-learning model that can accurately predict house prices in each state or country based on the available data. This will offer several benefits, including:

# Providing buyers with valuable information on how to compare and select houses.

# Enabling real estate agencies to gain insights into the factors that affect house prices.

# Addressing the challenge of limited and incomplete data in the real estate market

# Providing a continuous process of data collection, model refinement, and evaluation

# However, developing an accurate and reliable machine learning model is not without challenges, including limited and incomplete data, non-linear relationships, and model interpretability. To overcome these challenges, we aim to develop a solution that provides an effective tool for predicting house prices in each state or country. By doing so, our project can help solve several problems in the real estate market, providing valuable information to buyers and real estate agencies alike**.**

# **Data Description**

## Data source:

Our dataset was obtained from Kaggle, a popular platform for hosting and sharing datasets. The dataset consists of four files, with a total size of 957.39 kB, and is available in both CSV and TXT formats. The Ames Housing dataset was compiled by Dean De Cock specifically for use in data science education.

Link: <https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview>

## Size and format:

The dataset consists of 81 columns, including 79 feature columns and two additional columns for the ID and target variable. The target variable, SalePrice, is the column that the machine learning model must predict using the 79 explanatory variables. The dataset includes a total of 2919 rows, 1460 of which are in the training set and 1459 in the test set. To train the model, we can ignore the ID column, leaving us with 79 explanatory variables that can be used for prediction.

A picture containing text, screenshot, font, line

Description automatically generated

## Data type:

The data contains both categorical columns and numerical columns. The numerical features were 38.

A screenshot of a computer program

Description automatically generated with medium confidence

But some of the numerical features are, in fact, categorical. We concluded this fact from a visualization of the numerical features. The categorical features were 43.

A screenshot of a computer program

Description automatically generated with medium confidence

A screenshot of a computer screen

Description automatically generated with low confidence

## Missing values:

We have 2 types of null values, Numerical null values, and categorical null values.

**Numerical null values:**

Based on the “summary\_num” function, it appears that there are three new columns added to the “desc” DataFrame: “nunique”, “%unique”, “%unique”, and “Null”. The “nunique” column provides the number of unique values in each feature, which can help to identify whether a feature is categorical or continuous. The `%unique` column gives the percentage of unique values in each feature, which is useful for assessing the level of variation in the data. Additionally, the `Null` column provides the number of missing values in each feature, which is crucial for determining how to handle missing data. The function reports that there are 1460 observations in the dataset, with only one duplicated observation. It also states that many numerical variables may be categorical and require special treatment during feature engineering. Furthermore, it notes that most of the features do not have null values, but any features that do have null values should be examined carefully to determine how to handle the missing data. Finally, the function indicates that there are no duplicate values in the data.

A picture containing text, screenshot, number, font

Description automatically generated

A picture containing text, screenshot, number, document

Description automatically generated

**Categorical Null Values:**

The summary\_cat function returns a DataFrame called desc which contains summarized statistics for all categorical features in the input DataFrame. The function also reports that there are 1460 observations in the dataset, with 21 of them being duplicated. The function notes that there are many null values in multiple features that must be checked. It's worth considering that null values may indicate features that are not present in a given house. Therefore, it is essential to investigate the null values to ensure that the data is accurate and reliable. To determine whether null values are important or not, it is necessary to consult the dataset description and understand the context of the data. The description can provide information about the data collection process and the meaning of the features and their values. This information can help to determine whether null values represent a meaningful absence of data or a data processing issue.

A screenshot of a computer

Description automatically generated with medium confidence A screenshot of a computer

Description automatically generated with low confidence

From data description some of the null values which have meaning:

* Alley: Type of alley access to property

Grvl Gravel

Pave Paved

NA No alley access

* BsmtQual: Evaluates the height of the basement.

Ex Excellent (100+ inches)

Gd Good (90-99 inches)

TA Typical (80-89 inches)

Fa Fair (70-79 inches)

Po Poor (<70 inches

NA No Basement

* BsmtCond: Evaluates the general condition of the basement.

Ex Excellent

Gd Good

TA Typical - slight dampness allowed.

Fa Fair - dampness or some cracking or settling.

Po Poor - Severe cracking, settling, or wetness.

NA No Basement

* FireplaceQu: Fireplace quality

Ex Excellent - Exceptional Masonry Fireplace

Gd Good - Masonry Fireplace in main level

TA Average - Prefabricated Fireplace in main living area or Masonry Fireplace in basement

Fa Fair - Prefabricated Fireplace in the basement

Po Poor - Ben Franklin Stove

NA No Fireplace

* GarageType: Garage location

2Types More than one type of garage

Attchd Attached to home

Basment Basement Garage

BuiltIn Built-In (Garage part of house - typically has room above garage)

CarPort Car Port

Detchd Detached from home.

NA No Garage

## Outliers:

## We create a histogram using the matplotlib and seaborn libraries with the purpose of identifying outliers and understanding the characteristics of the dataset. The histogram allows us to visualize the distribution of a numerical variable, providing insights into the range and frequency of its values, and whether it is normally distributed or skewed.

A picture containing screenshot, plot, diagram, line

Description automatically generated

A picture containing text, screenshot, font, line

Description automatically generated

The histogram plot shows that the dataset is skewed to the left, indicating positive skewness. This can impact the validity of statistical tests and modelling techniques, so appropriate data transformation techniques may be necessary. However, the sales prices distribution seems normal for most houses, with only a small number of expensive outliers. Understanding the distribution of the data is crucial for making reliable analyses and modelling decisions.

* To better understand the distribution of the numerical features in the `numerical\_features` dataframe, we can create a histogram for each variable using the `hist` method. The resulting histograms provide valuable insights into the characteristics of the data, including the presence of any potential outliers or patterns. To generate the histograms, we use the `figsize` parameter to specify the size of the resulting grid, while the `bins`, `xlabelsize`, and `ylabelsize` parameters control the number of bins and font sizes of the x and y axis labels, respectively. By visualizing the distribution of each numerical feature, we can gain a better understanding of the data and make informed decisions about how to handle any potential issues or outliers. The resulting histograms can also provide insights into the skewness or kurtosis of the data, which can be useful for building models or conducting further analysis.

A screenshot of a graph

Description automatically generated with low confidence

## Data quality:

The quality of the data was generally good, with a relatively small number of missing values that were easily handled. However, during the analysis, some outliers were identified in some of the features, which could potentially impact the accuracy or validity of any subsequent modeling or analysis. To address this, appropriate outlier detection and treatment techniques should be applied to ensure that the data is of high quality and appropriately processed. By carefully handling outliers and ensuring that the data is of high quality, we can improve the accuracy and validity of any models or analyses that we develop.

**Exploring some of our features**:

Numerical Features:

It appears that different `MSSubClass` values have varying prices, and some types have many outliers, while `LotFrontage` is a continuous feature representing the linear feet of street connected to a property.

A picture containing text, diagram, plan, screenshot

Description automatically generated

**LotFrontage:** **Linear feet of street connected to property.**

The scatter plot displays the relationship between the `LotFrontage` feature, representing the linear feet of street connected to a property, and the target variable, `SalePrice`. It appears that, on average, a higher `LotFrontage` value is associated with a higher `SalePrice`, suggesting a positive correlation between these two variables.

A picture containing diagram, screenshot, plot, text

Description automatically generated

**OverallQual feature: Rates the overall material and finish of the house**

Based on its histogram distribution, it appears that the feature representing the rating of house material and finish is categorical, with each unique number identifying a different level of quality. This feature should be transformed into a categorical variable during the feature engineering process to improve the accuracy and validity of any subsequent modeling or analysis.

A picture containing text, diagram, screenshot, line

Description automatically generated

Additionally, there seems to be a linear relationship between the quality of a house and its sale price, which is not surprising since people typically prefer houses with better quality. This relationship is likely to be an important feature for modelling or analysis purposes.

**MoSold: Month House was sold in**

The line plot of `MoSold` versus `SalePrice` allows us to visualize any potential patterns or trends in the data over time. As `MoSold` represents the month of the year in which a house was sold, it is a categorical variable that may have seasonal effects on the housing market. It appears that houses sold in certain seasons have higher prices than those sold in other seasons, suggesting that `MoSold` may be an important feature for modeling or analysis purposes. By treating `MoSold` as a categorical feature during the feature engineering process, we can improve the accuracy and validity of any subsequent modeling or analysis.

**A picture containing plot, line, diagram, origami

Description automatically generated**

In conclusion for numeric features, the analysis of the numeric features has provided several feature engineering ideas that could be useful for improving the accuracy and validity of any subsequent modeling or analysis:

* Create a new feature that combines OverallQuall & OverallCond
* Create a building age feature (YrSold - YearBuilt)
* Create a building age after being remodeled feature (YrSold - YearRemodAdd)
* Turn MSSubClass feature into a categorical feature.
* Turn OverallQuall & OverallCond into categorical features (and their combined feature)
* Remove YearBuilt feature when you add the building age feature and turn the building age feature into a categorical feature.
* Turn all features that count things present in the house into categorical features.

All the graphs and visualizations are included in the code section rather than the documentation to avoid clutter and make it easier to access and understand the visualizations.

**Category Features**

**Alley: Type of alley access to property**

The `NaN` values in the `Alley` feature indicate that there is no alley access for those houses. When analyzing the `Alley` feature against the `SalePrice` variable, it was found that houses with gravel alley access have the lowest prices compared to those with no alley access or paved alley access. This suggests that the type of alley access may be an important factor to consider when analyzing the price of a house. By understanding the relationship between the `Alley` feature and `SalePrice`, we can gain insights into the characteristics of the data and make informed decisions about how to handle missing values and incorporate this feature into our modeling or analysis.

A picture containing screenshot, diagram, rectangle, line

Description automatically generated

**A screenshot of a computer

Description automatically generated with low confidence**

**LotShape: General shape of property**

When analyzing the relationship between the `LotShape` feature and `SalePrice`, it was found that `Regular` shaped houses are the most popular and the cheapest of the bunch, with a median price of $146,000 and a mean price of $164,754.81 across the 925 instances in the dataset. `IR1` shaped houses are also popular, with a higher average price of $206,101.67 compared to `Regular` shaped houses. `IR2` and `IR3` shaped houses are less common, with only 41 and 10 instances in the dataset, respectively, and have higher median and mean prices than `Regular` shaped houses. These findings can provide valuable insights into the characteristics of the data and may be useful for modeling or analysis purposes.

A picture containing text, diagram, screenshot, line

Description automatically generated

**A picture containing text, screenshot, font, number

Description automatically generated**

**Neighbourhood: Physical locations within Ame’s city limits**

When analyzing the relationship between the `Neighborhood` feature and `SalePrice`, it became apparent that each neighborhood is a significant factor in determining the sale price of a house. This is because each neighborhood has a different average sale price, indicating that location is an important factor in determining the value of a property. These findings suggest that the `Neighborhood` feature may be a crucial variable to consider when building a model or conducting further analysis. By understanding the relationship between `Neighborhood` and `SalePrice`, we can gain insights into the characteristics of the data and make informed decisions about how to incorporate this feature into our analyses.

A picture containing diagram, text, line, plan

Description automatically generated

**HouseStyle: Style of dwelling**

The analysis of the `HouseStyle` feature and `SalePrice` shows significant differences in median and mean sale prices for each house style, with `2.5Fin` and `2Story` houses having the highest median and mean sale prices. These findings suggest that `HouseStyle` is an important factor to consider when analyzing the sale price of a house.

**A picture containing text, diagram, screenshot, line

Description automatically generated**

Based on the analysis of the categorical features, several feature engineering ideas could be considered.

* Combine LotShape & LotConfig into a new feature.
* Combine Condition1 & Condition2 into a new feature.
* combine RoofStyle & RoofMatl into a new feature.
* Combine Exterior1st, Exterior2ndm ExterQual & ExterCond intoa anew feature.
* Combine BsmtQual & BsmtCond into a new feature.
* Combine Heating & HeatingQC into a new feature.
* Combine Heating & CentralAir into a new feature.
* Combine GarageType, GarageCars & GarageArea into a new feature.
* Combine SaleType & SaleCondition into a new feature.

**Data Cleaning:**

**Handling missing values:**

To handle missing values in the dataset, we utilized two different methods based on the type of feature. For numerical features, we used K-Nearest Neighbors (KNN) imputation to fill in the missing values. KNN imputation involves finding the K nearest neighbors to the missing value and using their values to estimate the missing value. This approach can be effective in preserving the underlying distribution of the data and reducing bias in the imputed values. For categorical features, we used the mode (i.e., the most frequent value) to fill in the missing values. The mode is a simple and effective method for imputing missing categorical values, as it preserves the most common value and does not introduce any new categories. By using these methods to fill in missing values, we can ensure that our data is complete and suitable for modeling or analysis purposes.

## Data transformation:

We use several data transformations in our dataset:

* **Target Transformation:**

To improve the accuracy and stability of our models, we performed a log transformation on the target variable. The first subplot of the figure shows the distribution of the target variable before the transformation, while the second subplot displays the distribution of `np.log(target)` after applying the log transformation. By using a log transformation, we can reduce the impact of outliers and skewness in the target variable, making it more suitable for modeling purposes. Additionally, the transformed target variable is more interpretable, as the resulting values represent the percentage change in the original target variable rather than the absolute change. Overall, the target transformation is an important step in preparing the data for modeling and improving the accuracy and reliability of our results.

A picture containing plot, diagram

Description automatically generated

**Trigonometric Transformation:**

This transformation was applied to the MoSold feature using the cosine function. The transformation is designed to capture the cyclic or seasonal pattern in data, if the variable represents the month in which a property was sold and that there is a repeating cycle of 12 months. The transformation involves taking the cosine of a scaled version of the MoSold variable, where the scaling factor is 0.5236. This scaling factor is chosen so that one cycle of the cosine function corresponds to one year (12 months). The negative sign is used to invert the values of the transformed variable, which helps to make it more symmetrical and easier to interpret. The resulting transformed variable will have values between -1 and 1, with peaks at the beginning and end of the 12-month cycle. This can help to capture any seasonal or cyclic patterns in the data and make them more apparent in visualizations or statistical analyses.

**Scaling Transformation**:

This transformation was applied to our data by using scaler.fit() method is then called on the scaler object, which calculates the mean and standard deviation of each feature in the input data. The StandScaler scales the input data so that it has zero mean and unit variance. This transformation ensures that all features have the same scale.

A picture containing screenshot, plot, text, line

Description automatically generated

# 

# **Exploratory Data Analysis (EDA):**

## Main Dependent Variable

In the Exploratory Data Analysis part, we first needed to explore our main dependent variable which is the house price, so we made a histogram and a QQ plot to get a better understanding of this variable. We also normalized the values of the price according to the normal distribution.

A picture containing screenshot, text, diagram

Description automatically generated

A picture containing screenshot, diagram, line, plot

Description automatically generated

The QQ plot represents how far the values of our variable are from the normal distribution which gives us an insight into how much the data is skewed.

We also got the skewness and kurtosis of the variable.

A screenshot of a computer

Description automatically generated with medium confidence

## Descriptive Statistics of some of the continuous variables

We studied three of our continuous variables and got their statistics and checked if they had any outliers. One of these variables were the price of houses.

A screenshot of a computer

Description automatically generated with medium confidence

Then we created a boxplot to check for any outliers

A screenshot of a computer screen

Description automatically generated with low confidence

As we can see there are a significant number of outliers, which we solved in the data transformation part when we applied the log transform.

Another continuous variable we had was the OverallQual, which rates the overall quality of the material and the finishing of the house.

A screenshot of a computer

Description automatically generated with medium confidence

The boxplot shows that there are not many outliers in this feature.

A screenshot of a graph

Description automatically generated with medium confidence

## Visualizations and correlations

For this part, we wanted to answer several questions:

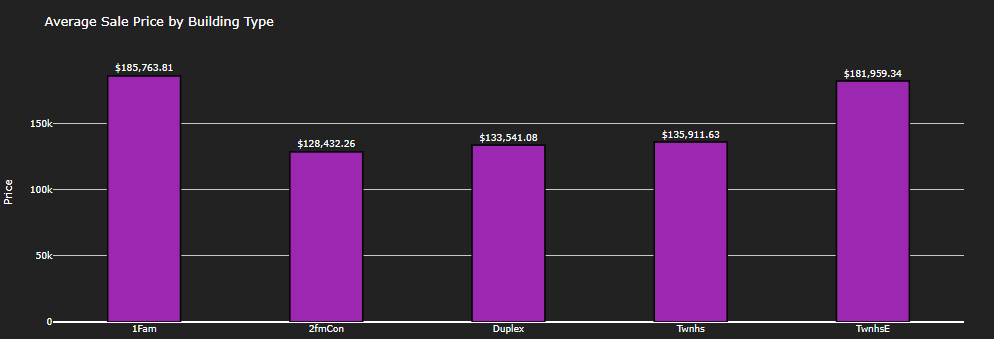
1. Distribution of dwelling types and their relation to sale prices?
2. Does zoning impact sale prices?
3. What is the street and alley access types of effect on sale price?
4. What is the Average sale price by property shape?
5. Is there a Correlation between Property Age and Sale Price
6. Is there a Correlation between Living Area and Sale Price
7. Does the price change year to year?

### Dwelling Types and their relation to sale prices

In this section, we used a bar plot to determine the count of each dwelling typeA screenshot of a video game

Description automatically generated with low confidence

Then, we wanted to find if there’s a correlation between the dwelling type and the house price, so we also used a bar plot with the average house price as the y-axis.



### Does zoning affect house prices.

In this section we wanted to know the correlation between the MSZoning feature and the house price feature.

A picture containing screenshot, design

Description automatically generated

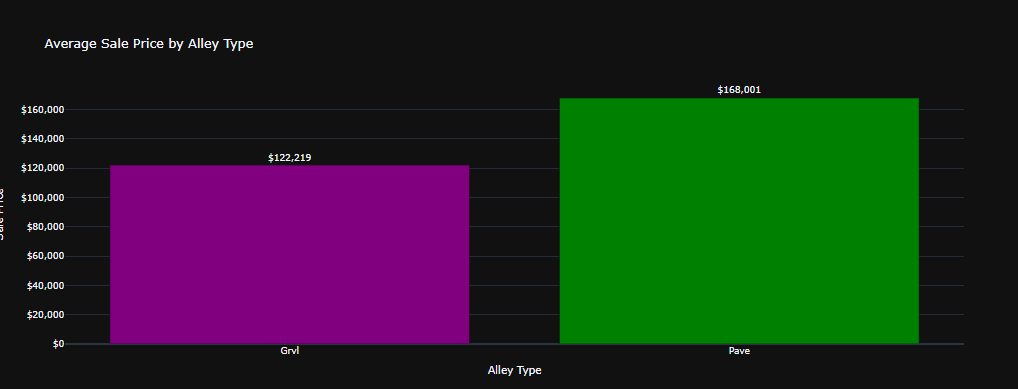
### The effect of street and ally on house prices

In this section we wanted to know the correlation between the Street type feature and the house price feature.

A screenshot of a computer screen

Description automatically generated with low confidence

And the correlation between the ally type and the price feature



### The average sales by property shape

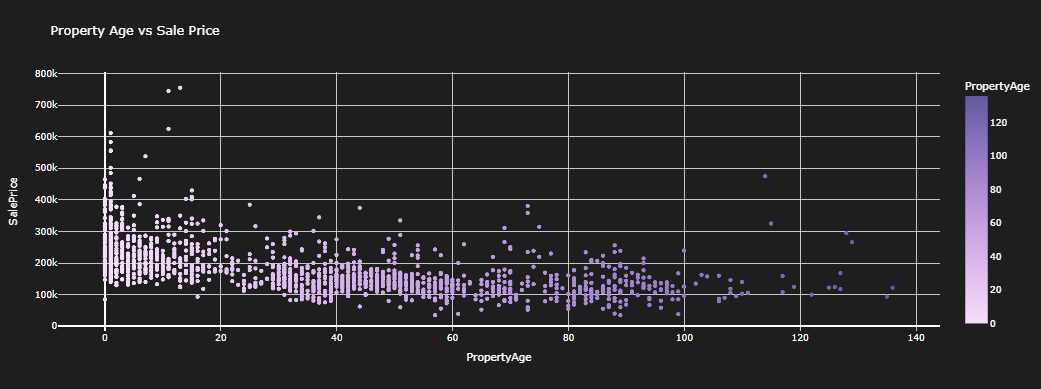
Here we tried to find the relationship between the property shape and the average price.

A picture containing screenshot, multimedia software, colorfulness, graphics software

Description automatically generated

### The relationship between the property age and the sale price

Because both features are numerical, we needed to use a scatter plot instead of a bar plot.



We calculated the correlation coefficient and found it to be = -0.52335, which means that there is a negative correlation.

### The correlation between living area and house price

Again, both of these features are continuous, which made us use the scatter plot again.

A screen shot of a graph

Description automatically generated with medium confidence

Here, there is a clear positive correlation between these two variables and their correlation coefficient is high: 0.7086

### The change of house prices over the years

As for our last question, we wanted to figure whether the house prices increased or decreased over time.

A screenshot of a computer

Description automatically generated with medium confidence

As we can see, the price has not changed much through the period from 2006 and 2010.

# **Modelling:**

When it comes to predicting the best model in a given data set, it is imperative to use a reliable and accurate method. In this regard, the compare model is a useful tool that can be used to efficiently compare and evaluate the performance of different models. Based on the results obtained from this tool, we were able to identify the top 5 models that were predicted to be the best. These models are:

1. CatBoost Regression
2. GradientBoost Regression
3. LGBM Regression
4. Bayesian Ridge
5. Orthogonal Matching Pursuit

After carefully analysing the results obtained from the compare model, we concluded that the CatBoost regressor was the best model for our data set. This decision was based on the fact that CatBoost regressor had the best least mean absolute error and mean square error when compared to the other models. This indicates that the CatBoost regressor was able to accurately predict the output values for our data set, making it the most suitable model for our needs.

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Mean Absolute Error | Mean Squared Error | R^2 |
|  |  |  |  |
| CatBoost | 0.0799 | 0.0146 | 0.9021 |

# **Results:**

In the results section of the project report, you should present the findings of your analysis.

Some of the key elements you can include in the results section are:

* **Model results:**

|  |  |  |  |
| --- | --- | --- | --- |
| * Model | Mean Absolute Error | Mean Squared Error | R^2 |
|  |  |  |  |
| CatBoost | 0.0799 | 0.0146 | 0.9021 |

**Visualizations:**

A picture containing text, screenshot, diagram, plot

Description automatically generated

* **Key results:** we use CatBoost regression to predict housing prices based on features such as location, number of bedrooms, and square footage, a low MAE and MSE and a high R2 would indicate good performance. A low MAE and MSE mean the model's predictions are close to actual values, while a high R2 means the model can explain a significant portion of the variability in the target variable. Future work on CatBoost regression could focus on improving its interpretability by developing methods for visualizing and understanding its internal workings or exploring techniques such as feature importance analysis to better understand how the model makes predictions.

# **Conclusion:**

# 

In conclusion, the initial exploration of the data set involved several important steps, including data cleaning, exploratory data analysis (EDA), and the development of a baseline model. During the data cleaning process, we checked for null values and used different techniques such as calculating the mode for categorical values and using KNN for numerical values. This helped us to ensure the data set was clean and ready for analysis.

In addition to data cleaning, we performed several transformations to prepare the data set for modeling. We also applied log transformations to reduce the impact of outliers and make the data more normally distributed. Additionally, we scaled and normalized numerical features to ensure they were on the same scale. These transformations helped to improve the performance of our models by reducing overfitting and making it easier for algorithms to learn patterns in the data.

Lastly, the development of a baseline model provided a starting point for our analysis. It allowed us to establish a benchmark against which we could compare the performance of more advanced models. Overall, the initial exploration of the data set was a critical step in the machine learning process, as it provided a solid foundation for subsequent analysis and modeling.