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**

[**Introduction** 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 main purpose of our project is to try to predict the house prices of a given state or country according to the existing data that we have. This results in giving the buyers some perspective on how to select their house and compare between different houses. Also, this project also gives the real estate agencies some insight into how different features affect the houses price.

# **Data Description**

## Data source:

We got this dataset from Kaggle.

## Size and format:

A screenshot of a computer

Description automatically generated with low confidence

This dataset has 1460 rows and 81 columns.

## 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

## Missing values:

We made a table which contained the percentage of null values in each feature by using isnull.sum() function and then we divided this by the length of the data frame. We handled the categorical missing values in this dataset by using the mode of the column. And the numerical values of the dataset by using the KNN (K nearest neighbours) which we will explain later.

## Outliers:

We used data transformation by the log transform method which is explained in the data cleaning section.

## Data quality:

The quality of the data was good overall. The null values were not many and easily handled.

**Data Cleaning:**

**Handling missing values:**

We use KNN to fill in missing values for numerical features, and the mode to fill in missing values for categorical features.

## Data transformation:

We use several data transformations in our dataset:

* + - **Target Transformation:** We use log transformation to the target variable. The first subplot of the figure shows the distribution of targets without the log transformation.

A picture containing text, screenshot, plot, line

Description automatically generated

The second subplot of the figure shows the distribution of target of np.log(target) with the log transformation.

A picture containing text, screenshot, plot, font

Description automatically generated

The logarithmic transformation is used to make the distribution of the target variable more symmetric and reduce the impact of extreme values.

**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:

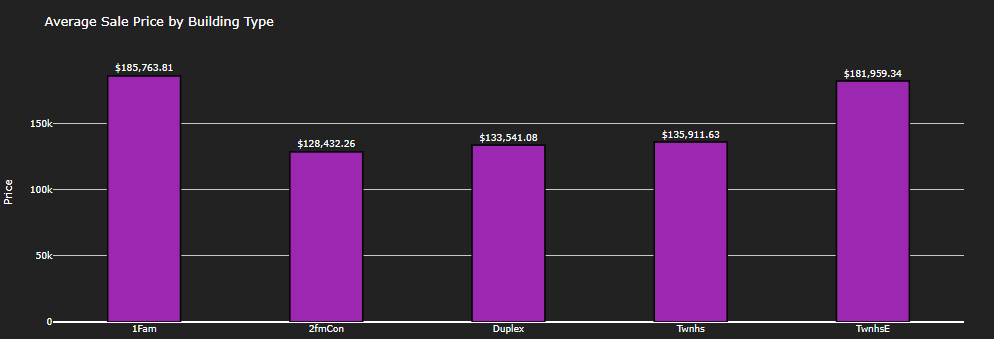
1. Distribution of dwelling types and their relation to sale prices?
2. Does zoning impact sale prices?
3. What is the street and alley access types of effect on sale price?
4. What is the Average sale price by property shape?
5. Is there a Correlation between Property Age and Sale Price
6. Is there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

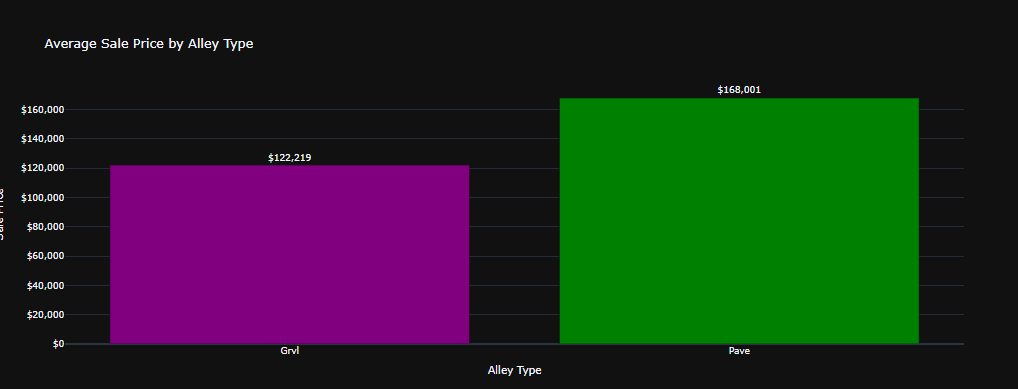
### The effect of street and ally on house prices

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

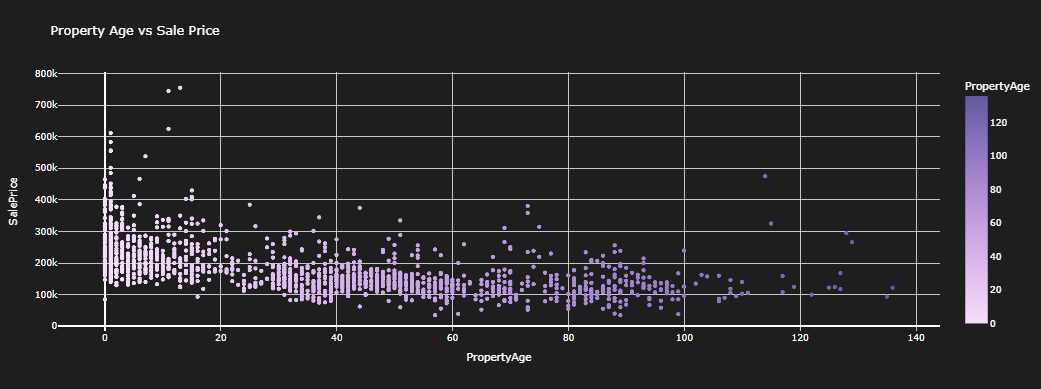
Here we tried to find the relationship between the property shape and the average price.

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.