A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name* | Denisse Guadarrama Tinoco |
| *Student Number* | 2024154 |
| *Module Title* | Machine Learning for Data Analysis |
| *Assessment Title* | CA1 |
| *Assessment Due Date* | 21th April 2024 |
| *Date of Submission* | 23th April 2024 |

**Declaration**

By submitting this assessment, I confirm that I have read the CCT policy on academic misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source.

I declare it to be my own work and that all material from third parties has been appropriately referenced.

I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Contents

[Introduction: iii](#_Toc164635070)

[Objective: iii](#_Toc164635071)

[Problem Definition: iii](#_Toc164635072)

[Data Source: iv](#_Toc164635073)

[Execution: iv](#_Toc164635074)

[References iv](#_Toc164635075)

# Introduction:

Predicting the price of a home could be very useful for anyone who is looking to buy or sell one.

With Machine Learning we can forecast very complex things having a reliable database, but for this Project we used a "simple" database to find more everyday answers such as: "how much will my house with 2 bathrooms, 3 bedrooms and that is in a rural area cost?" Being able to predict the price of a house makes what we have seen easier to understand.

In this project 3 different models will be used with a split of 20%, 25% and 30%:

* Random Forest Regression:

This model takes longer to execute because it creates multiple random decision trees, then averages the results of each one to generate a new result that leads to predictions.

It is commonly used in companies to predict prices of products or services in the future. It's an extremely accurate model thanks to its ‘wisdom of the crowds’ approach, it is easy to use and scales well.

* Linear Regression:

When you know the relationship between the independent and dependent variable have a linear relationship. It is a great tool to analyse the relationships among the variables but it isn’t recommended for most practical applications because it assumes that all variables in a problem have a linear relationship.

It is important before trying to fit a linear model to the observed dataset, one should assess whether or not there is a relationship between the variables. Of course, this doesn't mean that one variable causes the other, but there should be some visible correlation between them.

* K-Neighbors Regressor:

It is one of the simplest models in machine learning. As it will use the entirety of the training dataset to find the “nearest neighbors” according to the average distance of the target variable values of these neighbors.

# Objective:

Predict the price of a house depending on 5 variables, finding the best model with the best split:

1. House Size
2. Number of bathrooms
3. Number of bedrooms
4. Type of Neighborhood
5. Year of construction

# Problem Definition:

Let's suppose we are a new real estate agency in X place, and we are starting to sell our first houses, but we have no idea how to sell it, however there is a database with the values of the houses sold, with this information I could find a price that according to the conditions of the house you could sell. Therefore, the main problem is to find a model that best predicts the price of a house.

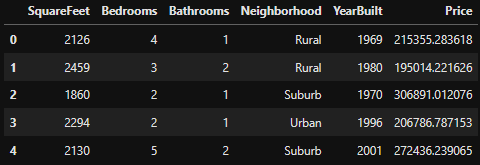
# Data Source:

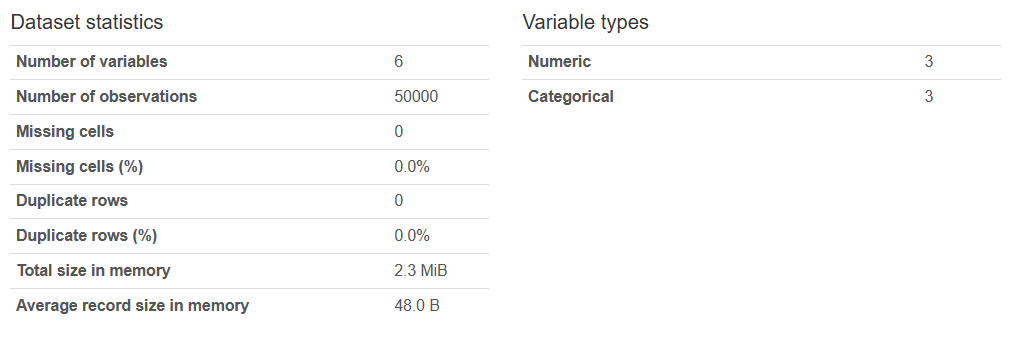
The URL database was taken: [Housing Price Prediction Data (kaggle.com)](https://www.kaggle.com/datasets/muhammadbinimran/housing-price-prediction-data)

# Execution:

The first step is to observe and clean our data, using statistical methods to know its distribution.

This is the visualization of 4 rows of the data set:

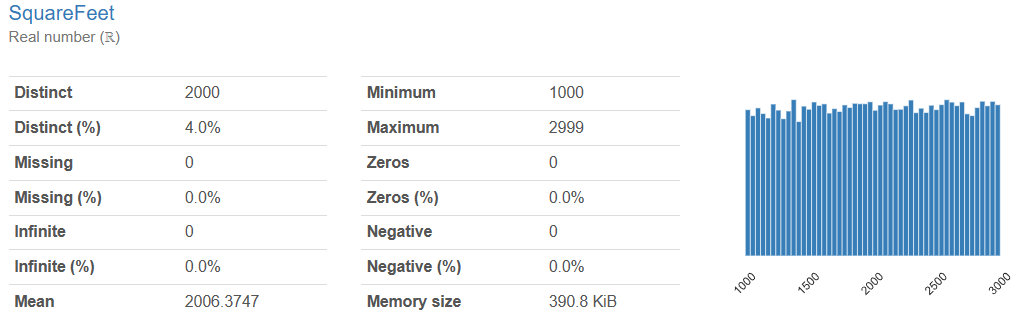


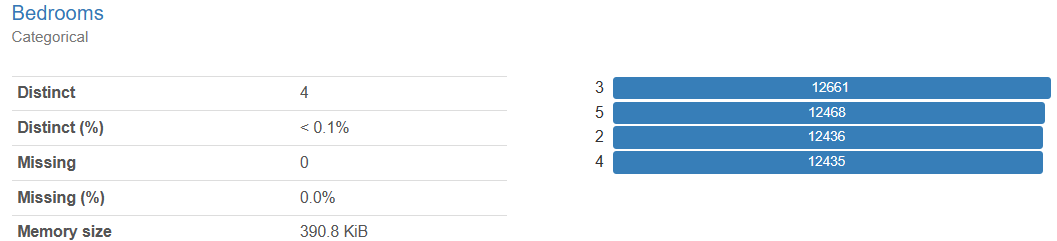


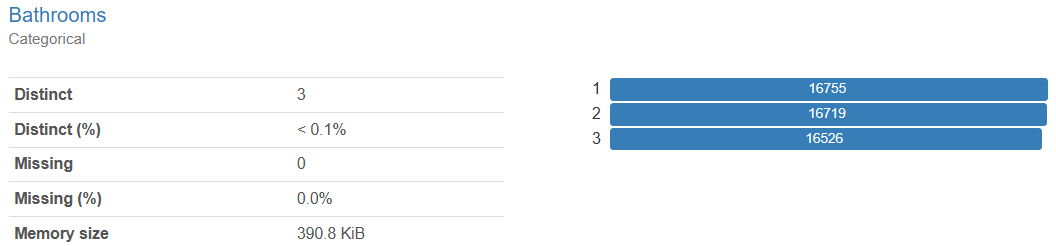
The data set contains 50,000 observations and 6 features.

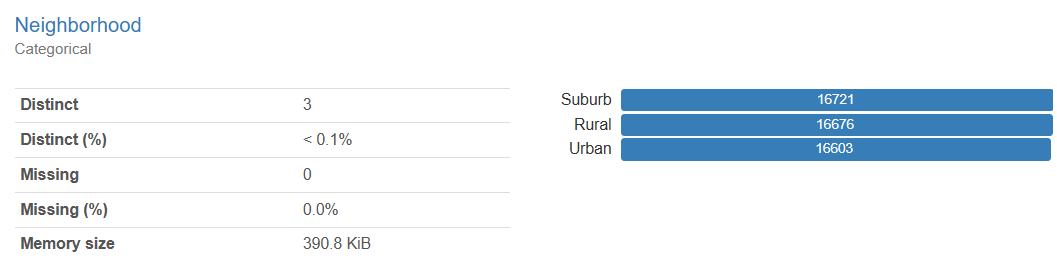
There is no missing data or duplicate rows.

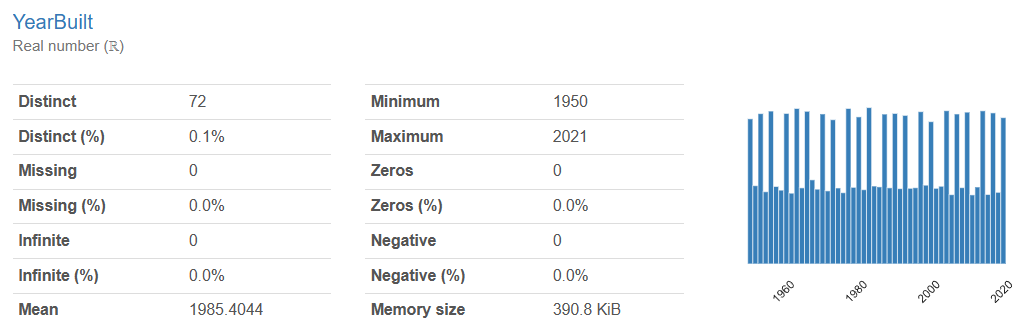
There are five variables:



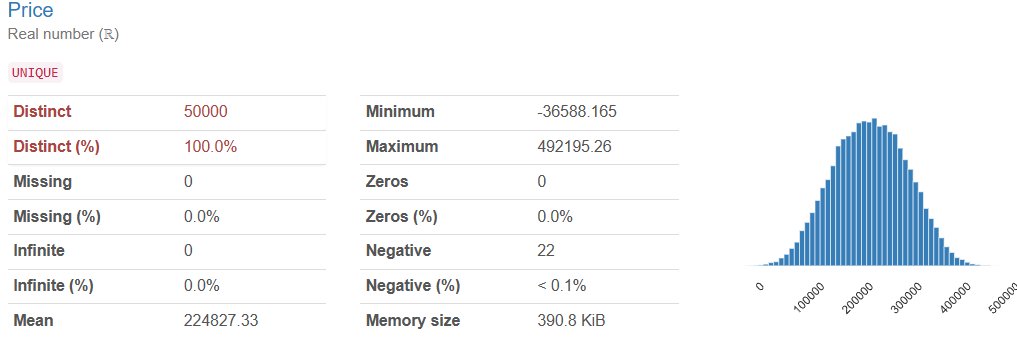






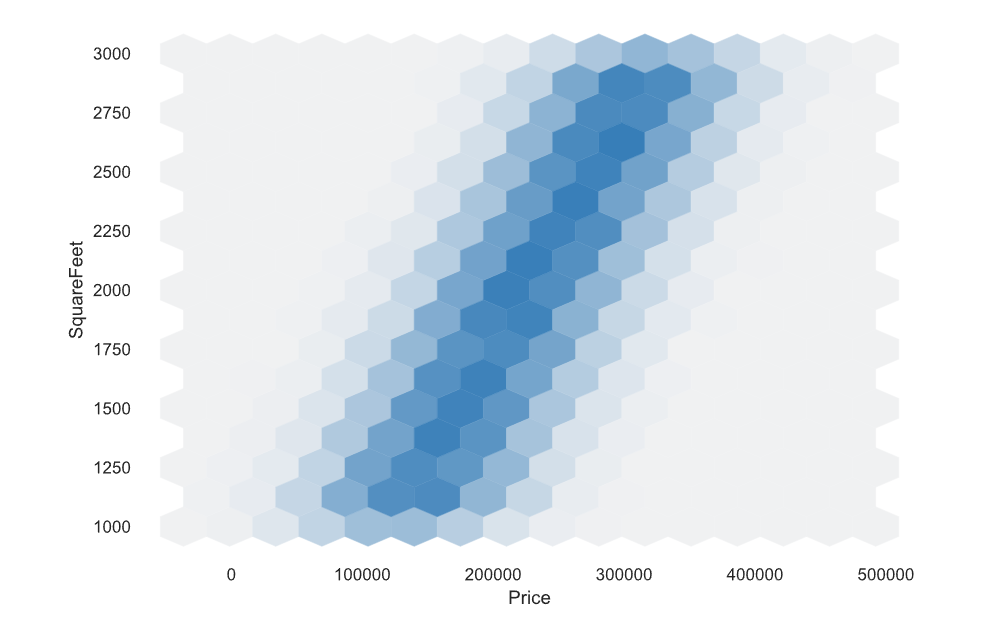


And the target variable:

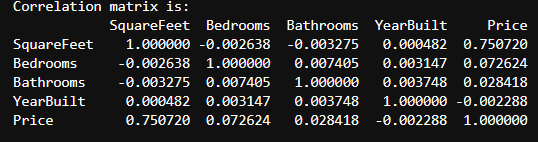


We can see that there are some outliers because there are prices less than 0. The mean, median and mode are not the same values for these outliers, however it can be observed that the price has a normal distribution.

In the following graph you can see the relationship between the squareFeet and price, the price increases as the size of the square increases.



Again, we see the correlation of the variables with the correlation matrix, only the squarefeet comes close to 1, this means that they have a linear relationship:



All of our columns are numeric, except for Neighborhood which contains 3 unique values: Rural, Suburb and Urban, so we are going to create dummies to create 3 new columns, which will be binary, this will make it easier to run our models, because if we just modified the Neighborhood column and assigned numerical values to each word, our model could assume that one value is better than another, therefore it is better to add new columns where only the data is represented as 1 or 0, it is easier to analyse.



We removed the original column and one of the 3 dummies because with the first two we have all the information of the original.

# References