**Universitat de les Illes Balears**

GRADO DE INGENIERÍA INFORMÁTICA

# DATA MINING

*Portugal Housing Analysis Report*

Juan Arturo Abaurrea Calafell, Marta González Juan, Andreas Manuel Korn, Marc Román Colom, Sergio Vega García, Andrés Borrás Santos

January 28, 2025

# Contents

# I Introduction

The data set Real Estate Listings in Portugal includes detailed information about each property listing such as its price, location, and more. We have been tasked with studying the dataset, understanding its shortcomings, addressing them, and conducting three studies on it. In the following points, we will explain what we have done, what questions we studied, and what results we got.

## I-A Raw Data

The dataset, consisting of information on real estate properties in Portugal, contains 114623 samples and 25 features. The features include numeric variables such as property prices and areas, and categorical variables like property types and location details. The raw data has both numeric and character variables.

For visualizing the dataset structure, here we can see the variable with the most ”normal” distribution and the worst one.

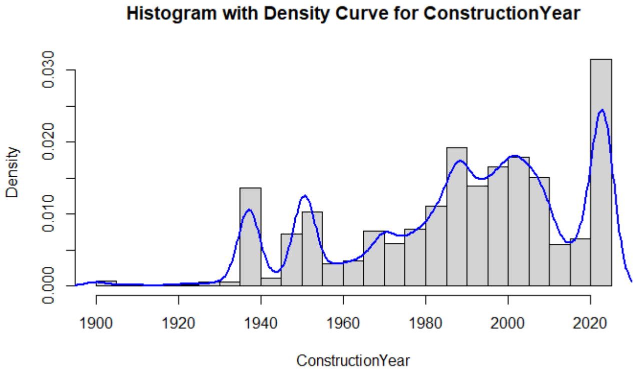

Fig. 1. Most normally distributed variable: construction year

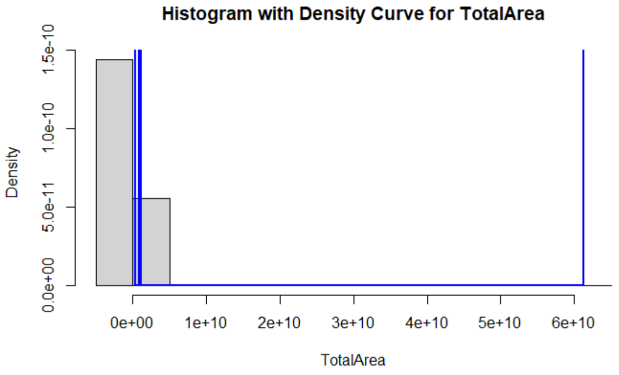

Fig. 2. Least normally distributed variable: total area

### I-A1 Preliminary Study

As seen, the distribution of the dataset variables leave a lot to be desired and present several challenges:

- **Duplicates**: The data set has some duplicate rows. These were retained, as it is common for properties with similar features to be listed multiple times.
- **Negative values**: 13 instances of negative values were found in fields where they are not valid, such as property area. These rows were removed.
- **Inconsistent values**: For example, the total number of rooms was sometimes less than the sum of bedrooms, bathrooms, and other rooms. These inconsistent values were set to NA and processed later.
- **Outliers**: After a deep analysis of the data, we found that some variables had huge outliers, for example houses with low living area but millions of rooms, or houses with one room but worth a billion euros.
- **Missing values**: Those are prevalent across several fields, particularly in fields like `BuildArea`, `GrossPrice`, and `LotSize`. The missing data handling strategy involved considering them as NAs for later removal, merging, or imputation based on context and field importance.
- **Redundant Values**: There are some variables that explain the same, for example `Elevator` and `Lift`, which we assume are the same, or `Parking` and `HasParking`.

### I-A2 Variables

Before delving into the analysis of the dataset, it is important to have a clear understanding of the variables involved. Below is a brief explanation of each of the variables:

- **Price**: Property price (euros).
- **District**, **City**, **Town**: Location details.
- **Type**: Property type (e.g., Apartment, House).
- **EnergyCertificate**: Energy rating.
- **Floor**: Floor number.
- **Lift**, **Elevator**: Building has elevator (True/False).
- **Parking**, **HasParking**: Number of parking spots availability.
- **ConstructionYear**: Year built.
- **TotalArea**, **GrossArea**, **LivingArea**, **BuiltArea**, **LotSize**: Area measurements (m²).
- **PublishDate**: Listing date.
- **Garage**: Garage availability.

| Variable | NºOutlier | NºMissingValues | Redundance? | Inconsistency? |
|---|---|---|---|---|
| Price | 25 | 244 | No | No |
| District | No | 0 | No | No |
| City | No | 0 | No | No |
| Town | No | 2 | No | No |
| Type | No | 16 | No | No |
| EnergyCertificate | No | 14 | No | No |
| Floor | No | 86694 | No | No |
| Lift | No | 46408 | Yes | No |
| Parking | 0 | 194 | Yes | No |
| HasParking | 0 | 46408 | Yes | No |
| ConstructionYear | 261 | 41073 | No | No |
| TotalArea | 3 | 6452 | No | Yes |
| GrossArea | 20 | 86985 | No | Yes |
| PublishDate | No | 99567 | No | No |
| Garage | No | 68247 | Yes | No |
| Elevator | No | 68247 | Yes | No |
| ElectricCarsCharging | No | 68247 | No | No |
| TotalRooms | 27 | 49355 | Yes | Yes |
| NumberOfBedrooms | 363 | 83931 | No | Yes |
| NumberOfWC | 646 | 73097 | No | Yes |
| ConservationStatus | No | 101542 | No | No |
| LivingArea | 145 | 27260 | No | Yes |
| LotSize | 10 | 85694 | No | Yes |
| BuiltArea | 17 | 96234 | No | No |
| NumberOfBathrooms | 1102 | 5355 | No | Yes |

TABLE I
PRELIMINARY STUDY RESULTS

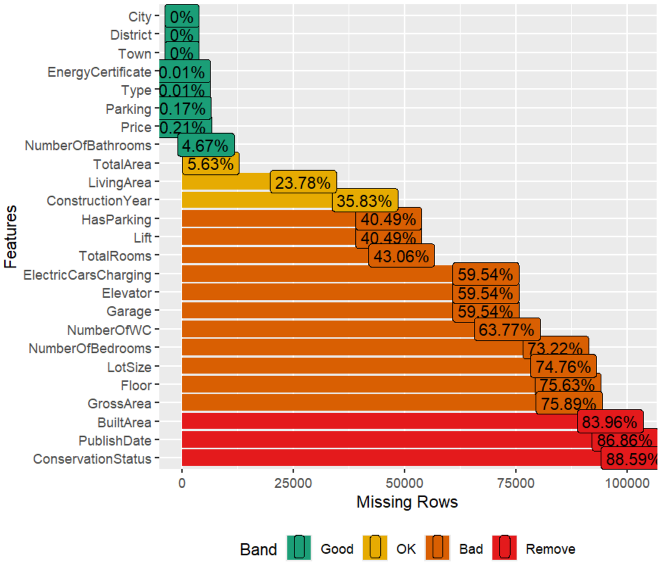

Fig. 3. NAs and empty values in Dataset

- **ElectricCarsCharging**: Electric car charging available.
- **TotalRooms**, **NumberOfBedrooms**, **NumberOfBathrooms**, **NumberOfWC**: Room counts.
- **ConservationStatus**: Condition (e.g., new, needs renovation).

This overview captures the key variables describing the properties. These variables provide detailed insights into the various aspects of each property. Understanding these variables is crucial for conducting further analysis and deriving meaningful conclusions from the dataset.

### I-B Data Expansion

To answer the third question more easily, we incorporated a dataset of Portuguese locations obtained from download.geonames.org, which includes latitude and longitude coordinates for each location.

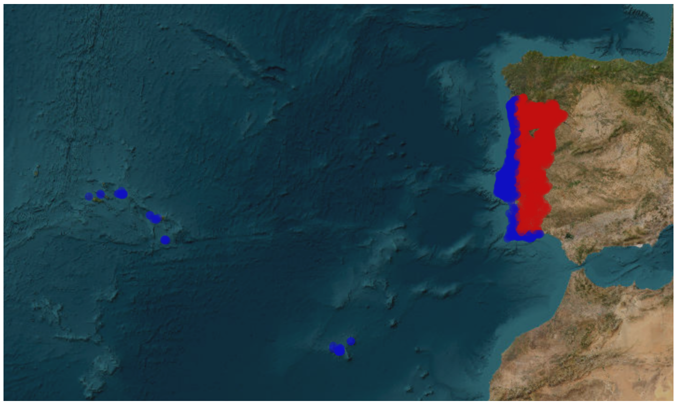

Fig. 4. Scatterplot of towns and cities in Portugal. Blue dots represent coastal locations, while red dots indicate inland locations.

### I-C Research Questions

Our group, based on the qualities and possibilities of the dataset, has decided to investigate the following questions:

1) Can we predict the year of construction based on the other variables?
2) Is there a relationship between the year of construction, energy efficiency and conservation status?
3) Is there a relationship between the property's geographical location and its price? Are properties in coastal districts more expensive?

## II Data Processing

To solve the issues stated in the previous section, we have used the next techniques learned in class.

### II-A Techniques

- **Merging**: This technique consists in merging some variables with equal or very similar description, such as with `Parking` and `HasParking`, and `Lift` and `Elevator`.
- **Grouping**: We have joined some values together in order to reduce the number of factors in some attributes, grouping less significant factors by their covariance in fused subsets, increasing their significance. This was done with the `Type` variable.
- **Outlier Detection (IQR)**: The Interquartile Range (IQR) method was used to detect and handle outliers. By calculating the IQR (difference between the third and first quartiles) and determining the lower and upper bounds (typically using 1.5 times the IQR), we identified values that fell outside these bounds as outliers. These outliers were changed to NAs, later handled.
- **Standardization**: Standardization is the process of transforming numeric data to have a mean of 0 and a standard deviation of 1. This technique ensures that the data follows a standard normal distribution, making it particularly useful for algorithms sensitive to the scale of features, such as gradient-based algorithms. In this project, we applied standardization to all

numeric variables to ensure consistency and improve the performance of our models.

- **Imputation**: Imputation refers to the technique of filling in missing values within the dataset. We used the library `randomForest`, explained at the appendix. We have used this model due to his explainability, internal automatic transformation to factor and robustness to outliers. Making a custom dataset with approximated variables. This was done with all NA values once all the other techniques were applied in order to obtain a dataset ready to use.
- **Feature Engineering**: Feature engineering involves creating new features from existing data to improve model performance. We have added a `is_coastal` property to know if an observation is on a coastal zone, and we have transformed `TotalRooms`, which is a superset of all rooms to a non-related new column, called `OtherRooms`. This increases the independency of attributes and models performance.
- **Variable formatting**: Variable formatting involves ensuring that all variables in the dataset are of the correct type and format, such as converting data types (e.g., character to factor).
- **Dummy Variables**: Dummy variables are binary variables created from categorical data also known in Machine Learning as One-Hot-Encoding. This helps us to know better correlations and perform better models. This was done in question 1.

## II-B Models used

We have used some samples of each type of model: Statistical, Clustering and Hierarchical for this assignment.

- **K-NN - Clustering**: Classifies or predicts values based on the majority or average of the k-nearest neighbours.
- **Random Forest - Hierarchical**: Combines multiple decision trees to make more accurate predictions and reduce overfitting.
- **Linear Model - Statistical**: Models a linear relationship between input features and the target variable.
- **K-medoids - Clustering**: Clusters data by choosing actual points as cluster centres, which are more robust to outliers.
- **Hierarchical - Clustering**: Builds a tree structure of clusters by either merging or splitting them iteratively.
- **T-test - Statistical**: Compares the means of two groups to determine if they are statistically different from each other. It is commonly used to analyze the effect of categorical variables on a continuous outcome.

# III Experimental Results

## III-A First research question

In this process, the goal was to predict the year of construction of a property using other variables in the dataset. First, the data was filtered by selecting specific columns, such as property characteristics, location, and other relevant factors. After cleaning the dataset by removing missing values, one-hot encoding was applied to handle categorical variables with many categories, such as property type and district. This transformation helped in the generation of a correlation matrix to identify which variables were relevant for predicting the construction year.

The correlation analysis revealed which attributes contributed significantly to the prediction of the construction year, and those with little impact were filtered out based on a set threshold. The data was further processed by removing samples with ”noisy” information (e.g., certain property types, districts, and energy certificates).

For handling missing values, Random Forest imputation was chosen to fill gaps in the dataset, using other available attributes as predictors. After imputing the missing values, a refined subset of the data was created, focusing on complete cases and removing columns with too many factor levels, ensuring the dataset was ready for model training.

With the clean and imputed datasets, the data was divided into training and test sets to evaluate the model performance. Three machine learning models —Random Forest, Linear Regression, and K-Nearest Neighbors (K-NN)— were applied to both the clean and imputed datasets. Each model was trained on the respective training data and evaluated based on the mean squared error (MSE) on the test data.

Finally, the results of each model were compared between the clean and imputed datasets, and visualizations such as scatter plots were used to analyze the accuracy of the K-NN model's predictions.

The analysis reveals that predicting a property's construction year is feasible using features like parking availability, energy certificates, and property type, with model performance varying based on data quality and method. K-Nearest Neighbors (K-NN) performed exceptionally well with real data (maybe too much), achieving an error rate of less than 5%, while Random Forest and Linear Regression models showed better accuracy when using imputed data.

Imputation and one-hot encoding improved model performance overall, but excessive reliance on invented data could distort predictions, as seen with K-NN. The study highlights that maintaining data integrity is key to achieving accurate results across clustering models.

### III-A1 K-NN Comparison Table

Hyperparameters are the key to get better performant models, for K-NN this is the table of values and why we have chosen the lowest one.

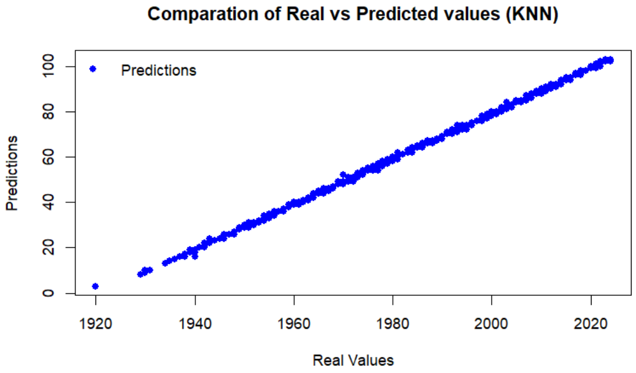

Fig. 5. K=1 Excellent performance

| $k$-value | MSE (Clean Data) | MSE (Imputed Data) |
|---|---|---|
| 1 | 3% | 639.4% |
| 3 | 15.9% | 775% |
| 5 | 20.6% | 811.4% |

TABLE II
K=X RESULTS VARIATION

### III-A2 Random Forest comparation table

The NTrees and NSamples hyper-parameters are linearly positive: bigger values give better results, but also increase computational cost and time. Because of that, we decided to use 100 trees and 15.000 samples, to reach a good accuracy but with finite computation time.

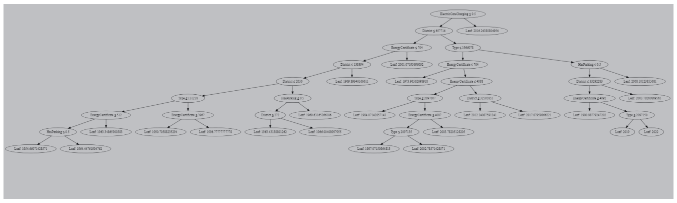

Fig. 6. Random Forest Plot - 1000 samples

| | MSE (Clean Data) | MSE (Imputed Data) |
|---|---|---|
| Normalized | 64.37% | 33.95% |
| Denormalized | 440.94% | 232.57% |

TABLE III
RANDOM FOREST RESULTS

### III-A3 Lineal Model comparation table

| | MSE (Clean Data) | MSE (Imputed Data) |
|---|---|---|
| Normalized | 64.8% | 60.3% |
| Denormalized | 444.29% | 413.46% |

TABLE IV
LINEAR MODEL RESULTS

## III-B Second research question

For this question, we started by filtering the original data set to a subset of the required variables and modified them accordingly.

With the data set prepared, we studied some plots that we found interesting. The plots in Fig. 7 gave some information, such as the fact that many of the properties from 1950 or before need a renovation, while most of the properties from 2010 and after are considered new.

In addition, properties with a conservation status of "new" are spread throughout all years, suggesting that properties in need of a renovation do receive it eventually.

Furthermore, regardless of the year of construction and conservation status, many properties do not have an energy certificate. We can even conclude that some renovations seem to only be superficial instead of comprehensive, as even for properties with good conservation status, energy certificates were bad or non-existent.

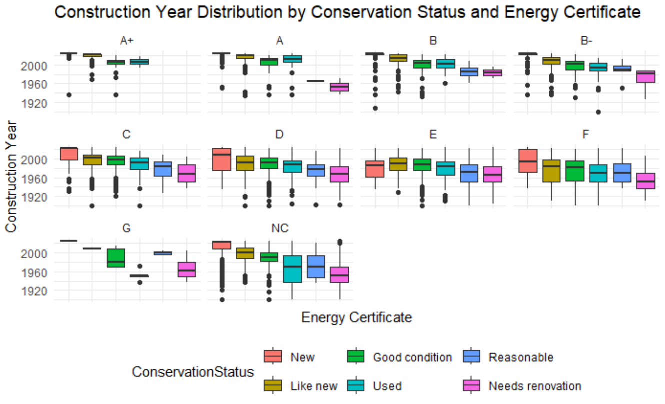

Fig. 7. Box plots of the distribution of year by energy certificate and conversation status

Finally, we developed two clustering models: a partitioning one based on K-medoids and a hierarchical one.

The K-medoids model showed an average silhouette width of 0.69, indicating that samples have a strong similarity to its cluster compared to the other clusters.

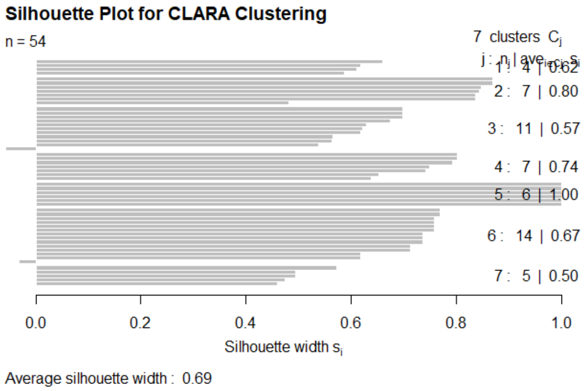

Fig. 8. Silhouette plot for CLARA clustering

The hierarchical model didn't provide any meaningful data as the highly dimensional structure of the dataset returned a hard to analyze dendogram (Fig. 9).

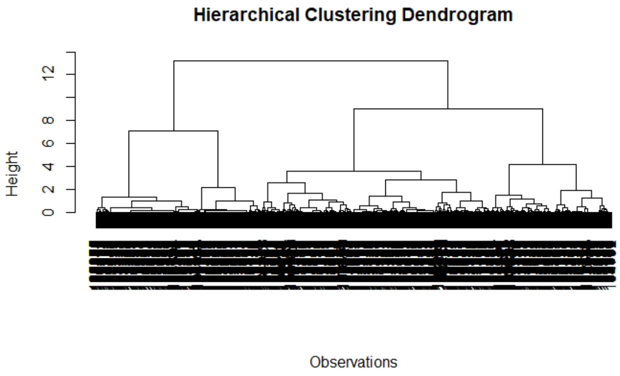

Fig. 9. Dendogram for hierarchical model

In conclusion, the analysis shows a strong relationship between construction year, energy efficiency, and conservation status. Older properties (pre-1950) mostly require renovation, while newer ones (post-2010) are often classified as "New" but frequently lack energy certificates. Surprisingly, some older properties with “New” conservation status still have poor energy ratings, suggesting superficial renovations. K-medoids clustering revealed clear groupings, while hierarchical methods struggled with data size. Overall, energy certification should be prioritized to ensure transparency and efficiency in both renovations and new builds.

## III-C Third research question

This study investigates the relationship between a property’s geographical location and its price, with a particular focus on whether coastal properties are more expensive than their non-coastal counterparts. To address this, we adopted a multi-layered analytical approach that included statistical testing, visual exploration, and regression modeling to uncover patterns and quantify the impact of location on property prices.

The visual analysis revealed clear patterns through both boxplot and scatterplot visualizations. The boxplot (Fig. 10) demonstrated that coastal properties have a notably higher median price compared to non-coastal properties, with distinctive outliers present in both categories while also revealing the higher variability of prices in coastal properties. The scatterplot (Fig. 11), mapping mean prices across different locations, showed a concentration of higher-priced properties along coastal areas, providing a clear geographical representation of the price distribution.

Later, we compared the average prices of coastal and non-coastal properties using a Welch Two Sample t-test. The results showed a highly significant difference, with coastal properties exhibiting consistently higher average prices. The statistical significance of this finding ($p\text{-}value < 2.2 \times 10^{-16}$) strongly supports the hypothesis that coastal properties are more expensive.

To gain a deeper understanding of how location interacts with other property features, we developed a linear regression model. This model incorporated almost all variables alongside a coastal location indicator. The results indicated that being located in a coastal area significantly increases property prices, even after controlling for other factors. Specifically, the analysis showed that proximity to the coast adds an average price of 97.480€, even though coastal properties are usually around 142.410€ more expensive.

Overall, the findings confirm a significant relationship between geographical location and property prices. Coastal properties are not only more expensive on average, but their higher prices persist even when accounting for other characteristics. This emphasizes the premium associated with coastal

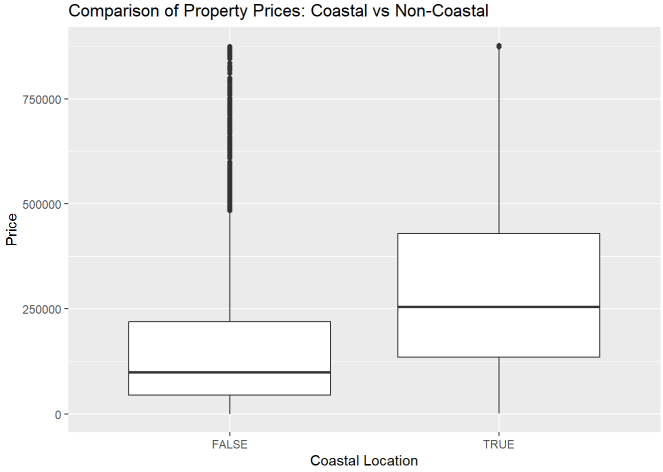

Fig. 10. Price box plot of coastal vs non-coastal

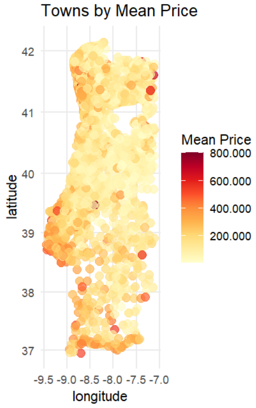

Fig. 11. Price scatter plot by location

locations and underscores the critical role of geographical factors in shaping real estate markets.

# IV Conclusions

This Data Mining project aimed to explore key research questions related to real estate data, specifically: (1) Can we predict the year of construction based on other variables?, (2) Is there a relationship between the year of construction, energy efficiency, and conservation status?, and (3) Is there a relationship between geographical location and property price? Through the application of various data mining techniques, we gained valuable insights into the factors influencing property prices, construction years, and sustainability.

For the first question, we used K-Nearest Neighbors (K-NN) to predict the year of construction. The model performed exceptionally well with errors below 5%, highlighting the effectiveness of K-NN for regression tasks. Features such as property size and the number of bathrooms were strong indicators of the construction year, helping predict property age even without explicit information.

In the second question, we explored how the year of construction affects energy efficiency and conservation status. Regression analysis revealed that older properties generally had lower energy efficiency scores and were more likely to require conservation efforts. This finding underscores the importance of considering a property's age when assessing its environmental performance.

For the third question, we focused on coastal properties and their relationship with price. Statistical tests and regression analysis confirmed that coastal properties are, on average, more expensive, with a 97.480€ price increase caused by the coast influence alone.

In summary, we learned that K-NN is effective for predicting continuous variables like construction year, while regression and clustering provided insights into price and sustainability trends. The project demonstrated the value of data preprocessing, statistical tests, and machine learning in uncovering real estate trends and providing actionable insights. These findings offer valuable tools for real estate professionals and highlight the significance of considering multiple factors, such as location and energy efficiency, in property evaluations.

# V Bibliography

- OpenAI, ChatGPT. (2024): LLM for natural language processing. https://chat.openai.com
- R Documentation: Documentation of the R language. https://www.rdocumentation.org/
- Deepseek R1: LLM for extreme text analysis. https://chat.deepseek.com/
- FINNSTATS: Statistics blog. https://finnstats.com/importance-of-data-cleaning-in-machine-learning/

# Appendix

- **DataExplorer**: Facilitates exploratory data analysis (EDA) with predefined functions that generate plots, tables, and statistical summaries of datasets. Useful for identifying missing values, distributions, and automating reports.
- **dplyr**: A library for data manipulation and transformation. It provides easy-to-use functions to filter, select, mutate, group, and summarize data.
- **ggplot2**: One of the most popular libraries for data visualization. It allows you to create customizable graphics using a "grammar of graphics" approach.
- **randomForest**: Implements the random forest algorithm, a machine learning method based on decision trees. Used for both classification and regression tasks.

- **caret**: Simplifies the workflow for predictive modeling. It provides functions for preprocessing, training, evaluating, and comparing models. Compatible with many machine learning algorithms.
- **class**: Provides functions to perform the k-Nearest Neighbors (k-NN) algorithm, a classification and regression method based on data proximity.
- **maps**: Facilitates creating basic static geographic maps, including regions and political divisions (countries, states, etc.).
- **leaflet**: A library for creating interactive maps. It allows you to add markers, data layers, controls, and customize maps using Leaflet.js in R.
- **ggmap**: Extends ggplot2 to work with geographic data. It enables overlaying data on maps retrieved from services like Google Maps or OpenStreetMap.
- **cluster**: Contains tools for cluster analysis, such as k-means, clara, pam, and hierarchical dendrograms.
- **RColorBrewer**: Used to apply the color palette in some images.
- **mapproj**: Similar to ggmap, used by the cluster library, helps us to plot the data in a map-like format.