BIGDATA ANALYTICS

report

under supervision of eng.Hosam eldein ramdan and eng.Sally Mohamed Ahmed

Made by :

حازم محمد غانم 20221460067

عمر عبد القادر محمد 20221311487

عمرو محمود خميس 20221460147

همسه محمد جابر 20221451413

خالد سامي محمد20221311293

علي الدين ياسر حسن20221509779

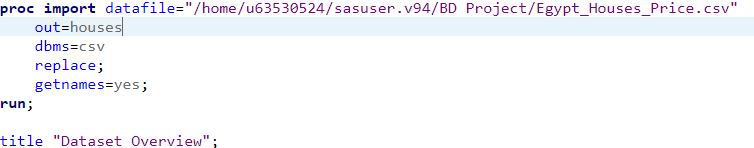
فارس محمد يحيى 20221446302

introduction

This report walks through the steps taken to analyze house prices in Egypt. We started by importing the dataset, cleaned up any messy data, and handled outliers to make sure the data was accurate. After that, we transformed the data, built models to predict prices, and even tried to classify compounds. Along the way, we checked how well the models performed and visualized the results to make sense of the findings.

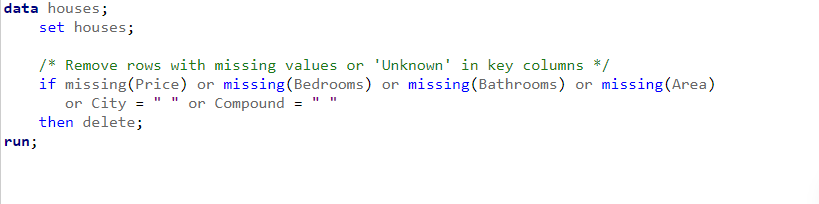
Import Data

The first step involved importing the dataset containing information about house prices in Egypt. We used the PROC IMPORT procedure to bring the data into SAS



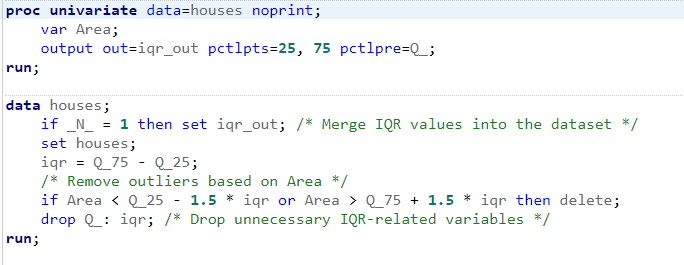
Data Cleaning

The second step involved cleaning the data by removing rows with missing values or 'Unknown' labels in important columns such as Price, Bedrooms, Bathrooms, Area, City, and Compound



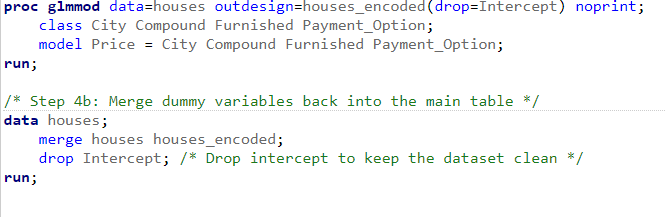
RemovingOutliers

We identified and removed outliers in the Area column using the Interquartile Range (IQR) method, which detects values far from the typical range.



One-Hot Encoding

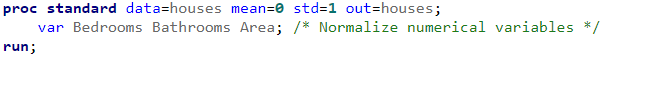
proc glmmod creates dummy variables for categorical variables like City, Compound, Furnished, and Payment\_Option, which are then merged back into the main dataset. This step transforms categorical data into numeric features suitable for regression models.



Scale and Normalize Numerical Data

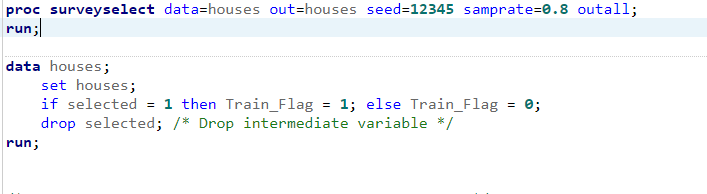
Standardize the numerical variables (Bedrooms, Bathrooms, and Area) by scaling them to have a mean of 0 and a standard deviation of 1.

proc standard is used to normalize the specified variables in the dataset.



Split Data into Train and Test

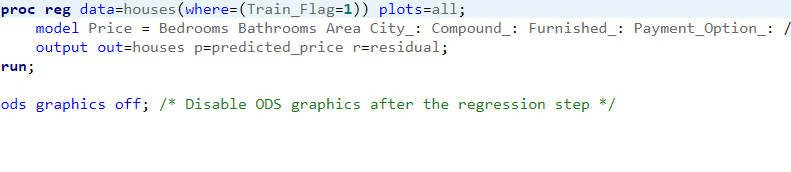
in this step we Split the data into training and testing sets (80% training, 20% testing). proc surveyselect performs random sampling of the dataset, and Train\_Flag is assigned to indicate which rows are part of the training set.



Regression Model - Predict House Price

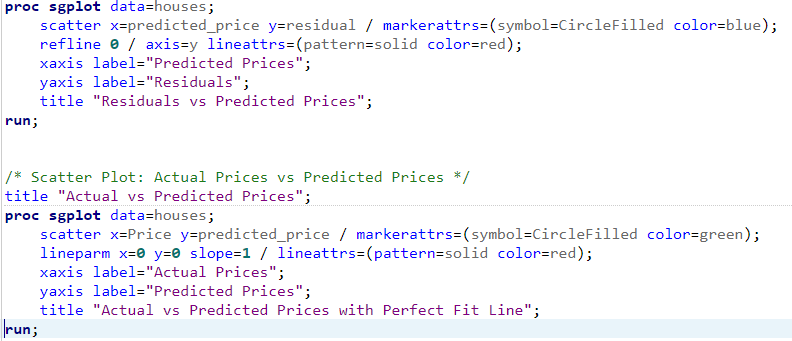
The purpose of this step is to Build a regression model to predict house prices.

proc reg fits a linear regression model to predict Price using features like Bedrooms, Bathrooms, Area, and the one-hot encoded variables. It also calculates predicted prices (predicted\_price) and residuals (residual) for further analysis.



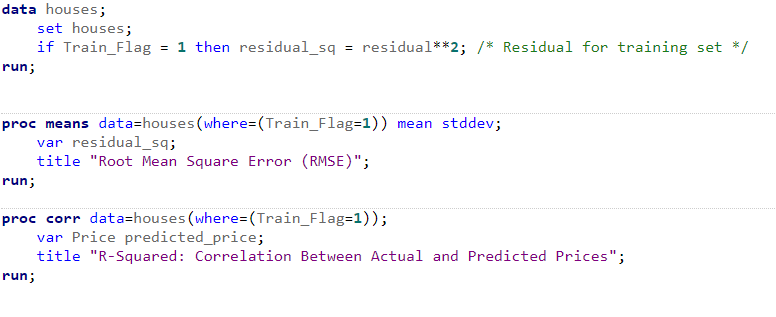
Visualize Residuals and Predicted vs Actual

Visualize the residuals and compare actual vs predicted house prices. The first plot shows residuals (difference between actual and predicted values) against predicted values, with a reference line at 0. The second plot compares actual house prices with predicted values, including a perfect fit line.



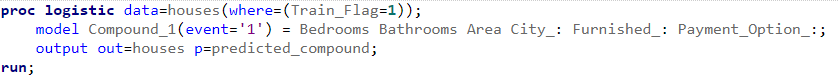
Evaluate Regression Model Performance

Evaluate the model performance using RMSE and R-squared. The first step calculates the RMSE by taking the square root of the mean of squared residuals. The second step computes the correlation between actual and predicted prices, providing an R-squared value.



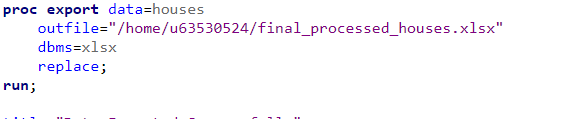
Logistic Model - Predict Compound

In this step we Build a logistic regression model to predict whether a house is part of a compound. This step uses proc logistic to predict the Compound\_1 variable based on the same features used in the regression model. The predicted probabilities are stored in predicted\_compound.

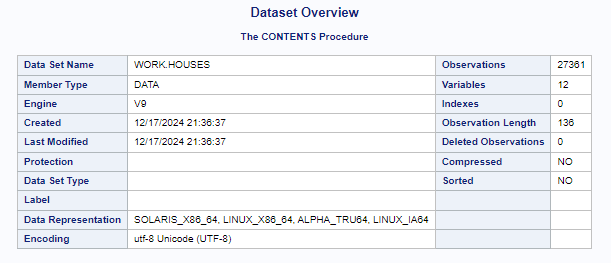


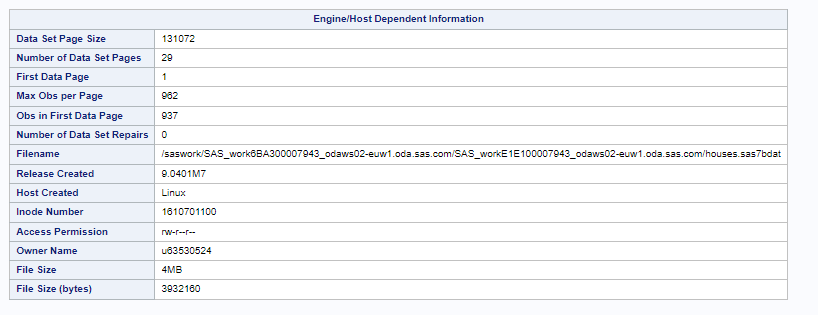
Export Final Processed Table

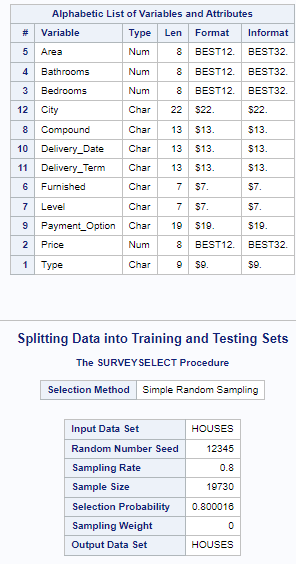
In this step we Export the final processed dataset to an Excel file. The proc export procedure is used to save the processed dataset to an Excel file at the specified location.

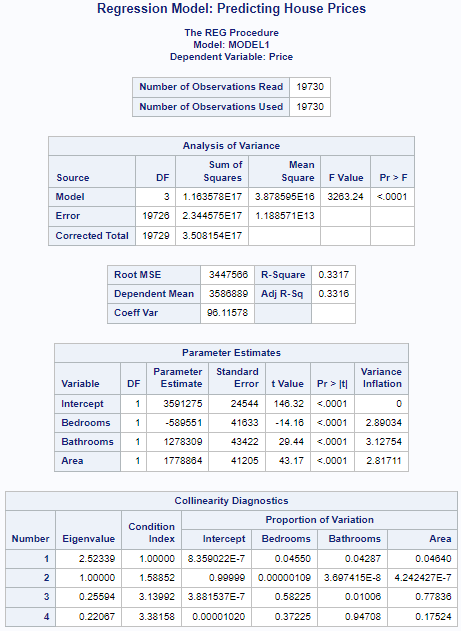


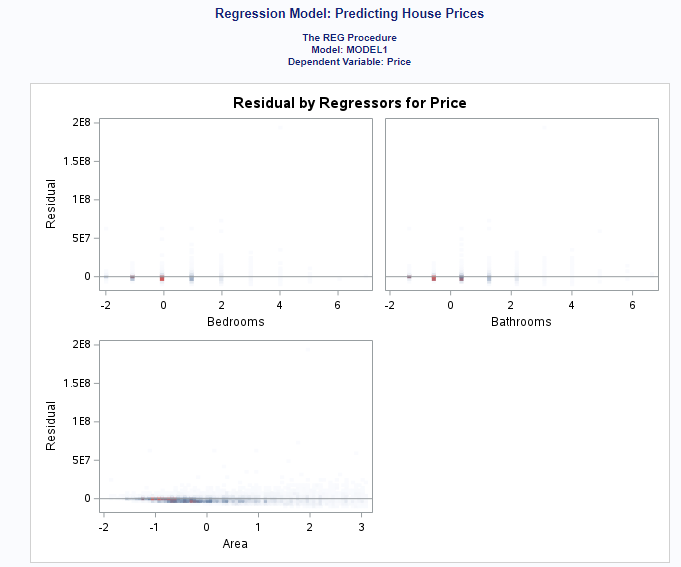
Results

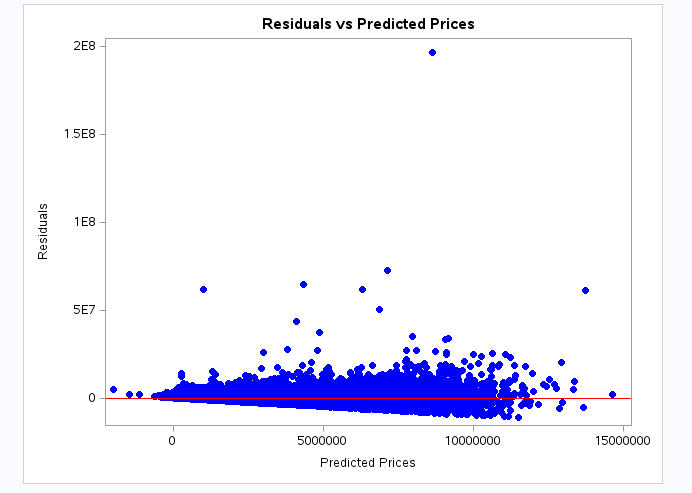


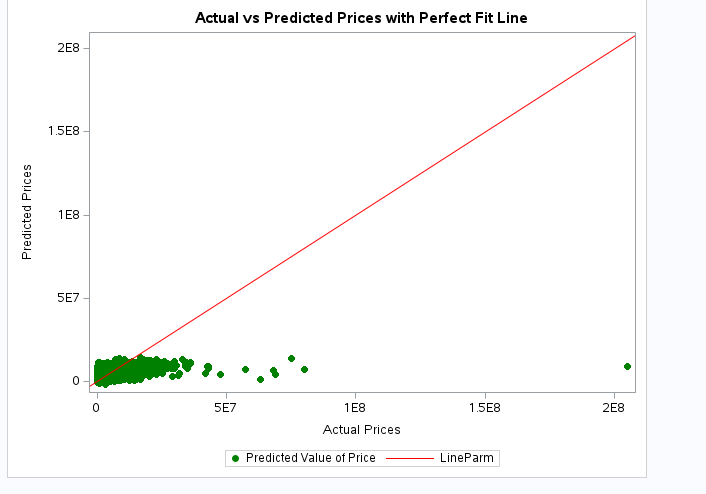


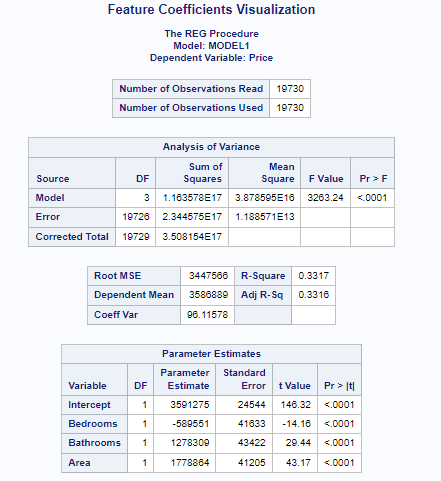


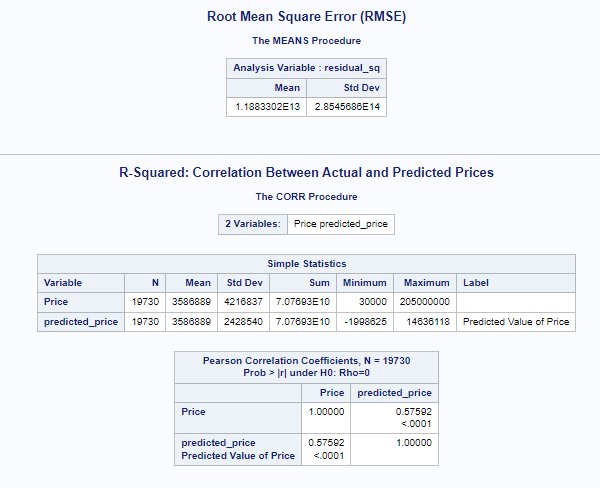




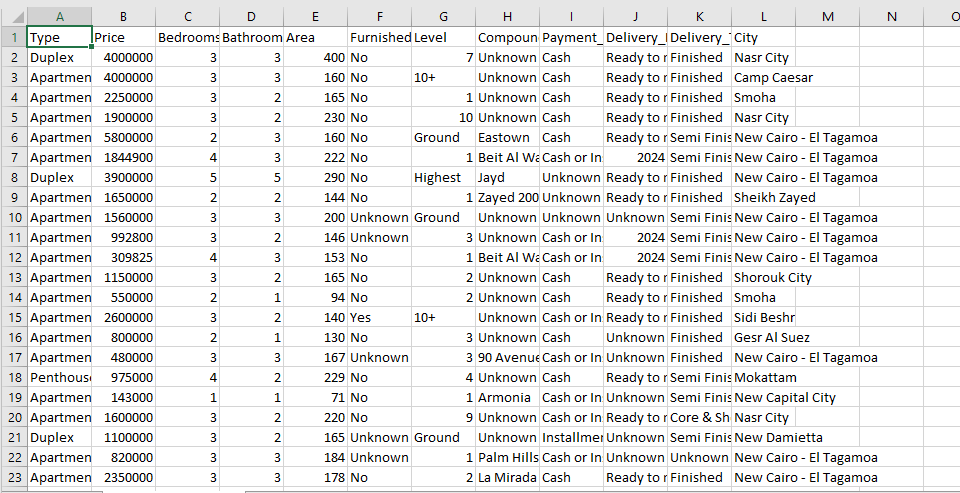








DATA Before



DATA AFTER

