DATA SCIENCE INSTITUTE

\WOOLF/

ASSIGNMENT

FUNDAMENTALS IN PREDICTIVE MODELING

Python

REPORT

Pillar1-e

Prepared By

William C. Phiri

PG Dip in Data Science

Student No. 4295158639

+353-87-3502102

chenechoz@gmail.com

Athlone, IRELAND

BACKGROUND:

The data for modelling provided contained information on the selling price of each house in million Rs. It also contained Carpet area in square feet, Distance from nearest metro station and number of schools within a 2 km distance. The data has 198 rows and 5 columns.

ASSIGNMENT OBJECTIVE:

To establish the following below using Python programming.

1. Build a regression model on training data to estimate selling price of a House.

2. List down significant variables and interpret their regression coefficients.

3. What is the R2 and adjusted R2 of the model? Give interpretation.

4. Is there a multicollinearity problem? If yes, do the necessary steps to remove it.

5. Are there any influential observations in the data?

6. Can we assume that errors follow 'Normal' distribution?

7. Is there a Heteroscedasticity problem? Check using residual vs. predictor plots.

8. Calculate the RMSE for the Training and Testing data.

RESULTS:

Question 1: Build a regression model on training data to estimate selling price of a House.

OUTPUT: Refer to code on GitHub at the LINK HERE for the model created

Refer to app location HERE for the prediction app built off the training data

The project structure outline was laid out as follows.

```
FPM_Assignment_PY/
│
├── dashboard/
│   └── app.py                     # ✅ Streamlit app
│
├── data/
│   ├── raw/
│   │   └── House Price Data.csv
│   ├── processed/
│   │   ├── hse_price_cleaned.csv
│   │   ├── hse_price_optimized.csv
│   │   ├── incoming_house_data.csv  # ✅ For dashboard input testing
│   │   ├── X_train.csv
│   │   ├── X_test.csv
│   │   ├── y_train.csv
│   │   └── y_test.csv
├── environment/
│   ├── environment.yml
│   └── requirements.txt
│
├── models/
│   └── house_price_model.pkl
│
├── notebooks/
│   ├── 01_EDA.ipynb
│   ├── 02_Model_Building.ipynb
│   └── 03_Evaluation_Report.ipynb
│
├── reports/
│   ├── summary.txt
│   └── 03_objective_evaluation_report.pdf
│
├── src/
│   ├── data_prep.py
│   ├── train_model.py
│   └── utils.py
│
├── .gitignore
├── main.py
└── README.md
```

Question 2: List down significant variables and interpret their regression coefficients.

OUTPUT:

Model Summary Results.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.808
Model:                            OLS   Adj. R-squared:                  0.804
Method:                 Least Squares   F-statistic:                     215.9
Date:                Sun, 20 Apr 2025   Prob (F-statistic):           6.08e-55
Time:                        14:36:40   Log-Likelihood:                -348.18
No. Observations:                 158   AIC:                             704.4
Df Residuals:                     154   BIC:                             716.6
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -9.5704      1.935     -4.947      0.000     -13.392      -5.748
area           0.0343      0.002     14.867      0.000       0.030       0.039
distance      -1.8737      0.178    -10.552      0.000      -2.224      -1.523
schools        1.4379      0.447      3.216      0.002       0.555       2.321
==============================================================================
Omnibus:                       13.376   Durbin-Watson:                   2.259
Prob(Omnibus):                  0.001   Jarque-Bera (JB):               14.373
Skew:                          -0.726   Prob(JB):                     0.000757
Kurtosis:                       3.270   Cond. No.                     1.15e+04
==============================================================================
```

EXPLANATION:

Variable	Coefficient	P-value	Interpretation
Carpet Area	0.0343	<0.05	For each additional **sq. ft.**, the selling price increases by **0.034 million Rs**, holding other variables constant.
Distance to Nearest Metro	-1.8737	<0.05	For each additional **km away from metro**, the price **decreases by 1.87 million Rs**, all else equal.
Schools	1.4379	<0.05	Each additional **school nearby** increases the price by **1.3 million Rs**, if other factors are held constant.

Question 3: What is the R2 and adjusted R2 of the model? Give interpretation.

OUTPUT: **R^2 = 0.808** → ~80% of the variation in house price can be explained by this model.

Adjusted R^2 = 0.808 → Similar value after adjusting for number of predictors, confirms the model strength.

This is a strong model with a good explanatory power.

Question 4: Is there a multicollinearity problem? If yes, do the necessary steps to remove it.

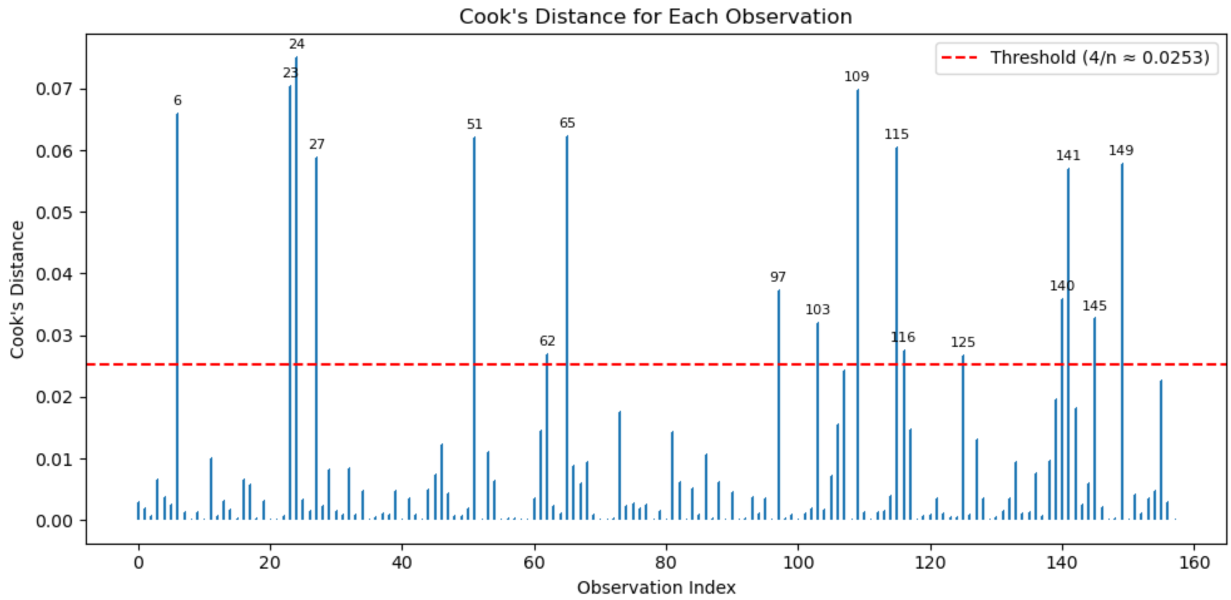
VIF Output:

	Variable	VIF
0	Intercept	119.98940511378272
1	area	1.6985770212861035
2	distance	1.05520609424952
3	schools	1.7641262589958706

Multicollinearity was checked using the variance inflation factor and all variables had VIFs <5 indicating that there was no multicollinearity and hence all variables were retained in the model.

Question 5: Are there any influential observations in the data?

OUTPUT: Cooks distance check revealed influential points above the threshold, these points can disproportionately affect the regression. The model was re-run with the influential observations removed for comparison to initial model.



The influential points were noted as follows

```
Influential points: [  6  23  24  27  51  62  65  97 103 109 115 116 125 140 141 145 149]
```

Model re-fit after removing the influential points

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.858
Model:                            OLS   Adj. R-squared:                  0.855
Method:                 Least Squares   F-statistic:                     275.2
Date:                Sun, 20 Apr 2025   Prob (F-statistic):           8.66e-58
Time:                        14:28:17   Log-Likelihood:                -274.07
No. Observations:                 141   AIC:                             556.1
Df Residuals:                     137   BIC:                             567.9
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    -12.0073      1.899     -6.322      0.000     -15.763      -8.251
area           0.0397      0.002     17.164      0.000       0.035       0.044
distance      -2.1219      0.160    -13.221      0.000      -2.439      -1.805
schools        0.5122      0.399      1.285      0.201      -0.276       1.300
==============================================================================
Omnibus:                        9.670   Durbin-Watson:                   2.280
Prob(Omnibus):                  0.008   Jarque-Bera (JB):               10.140
Skew:                          -0.656   Prob(JB):                      0.00628
Kurtosis:                       3.051   Cond. No.                     1.37e+04
==============================================================================
```

Model Summary Comparison Original vs Cleaned(Influential points removed)

Variable	Coefficient		P-value		Conclusion
	Original	Cleaned	Original	Cleaned	
Carpet Area	0.0346	0.0397	< 0.05	< 0.05	There is no significant difference between the two models as it relates to carpet area
Distance to Nearest Metro	-1.8704	-2.1219	< 0.05	< 0.05	There is no significant difference between the two models as it relates to distance to nearest Metro station
Schools	1.3187	0.5122	< 0.05	**0.201**	***There is a significant difference between the two models as it relates to schools nearby. The cleaned model has a p-value greater >0.05. This suggests that those outlier rows were contributing to the apparent importance of this variable.***

R^2 and Adjusted R^2 Value comparison

Metric	Original Model	Cleaned Model
R^2	0.808	**0.858**
Adjusted R^2	0.804	**0.855**

Key Takeaway:

After the removal of the influential points it was observed that the p-value for schools became >0.05 indicating that the schools variable is not statistically significant and it was the influential observations that were contributing to the apparent importance of this variable. In the light of this new information we need to re-fit the model after removing the schools variable seeing as it is not statistically significant. So next key steps will include;

1- **Drop the schools variable from the data set**
2- **Re-split the remaining data set into train and test sets (80/20)**
3- **Re-fit the model and print model summary**
4- **Assess R^2 and Adjusted R^2 and reconfirm statistical significance of variables.**

Re-fit model summary after influential points removed and statistically insignificant variable removed.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.795
Model:                            OLS   Adj. R-squared:                  0.792
Method:                 Least Squares   F-statistic:                     300.5
Date:                Sun, 20 Apr 2025   Prob (F-statistic):           4.58e-54
Time:                        15:21:18   Log-Likelihood:                -353.32
No. Observations:                 158   AIC:                             712.6
Df Residuals:                     155   BIC:                             721.8
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    -10.6139      1.964     -5.404      0.000     -14.494      -6.734
area           0.0391      0.002     21.416      0.000       0.035       0.043
distance      -2.0031      0.178    -11.249      0.000      -2.355      -1.651
==============================================================================
Omnibus:                       16.129   Durbin-Watson:                   2.289
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               17.806
Skew:                          -0.768   Prob(JB):                     0.000136
Kurtosis:                       3.588   Cond. No.                     1.13e+04
==============================================================================
```

EXPLANATION:

Variable	Coefficient	P-value	Interpretation
Carpet Area	0.0391	<0.05	For each additional **sq. ft.**, the selling price increases by **0.039 million Rs**, holding other variables constant.
Distance to Nearest Metro	-2.0031	<0.05	For each additional **km away from metro**, the price **decreases by 2 million Rs**, all else equal.

Key Takeaway:

Based on the model summary above we can conclude that the remaining variables are significant and we can proceed to re-build the model with the said variables.

R^2 = 0.795 → ~80% of the variation in house price can be explained by this model.

Adjusted R^2 = 0.792 → Similar value after adjusting for number of predictors, confirms the model strength**.**

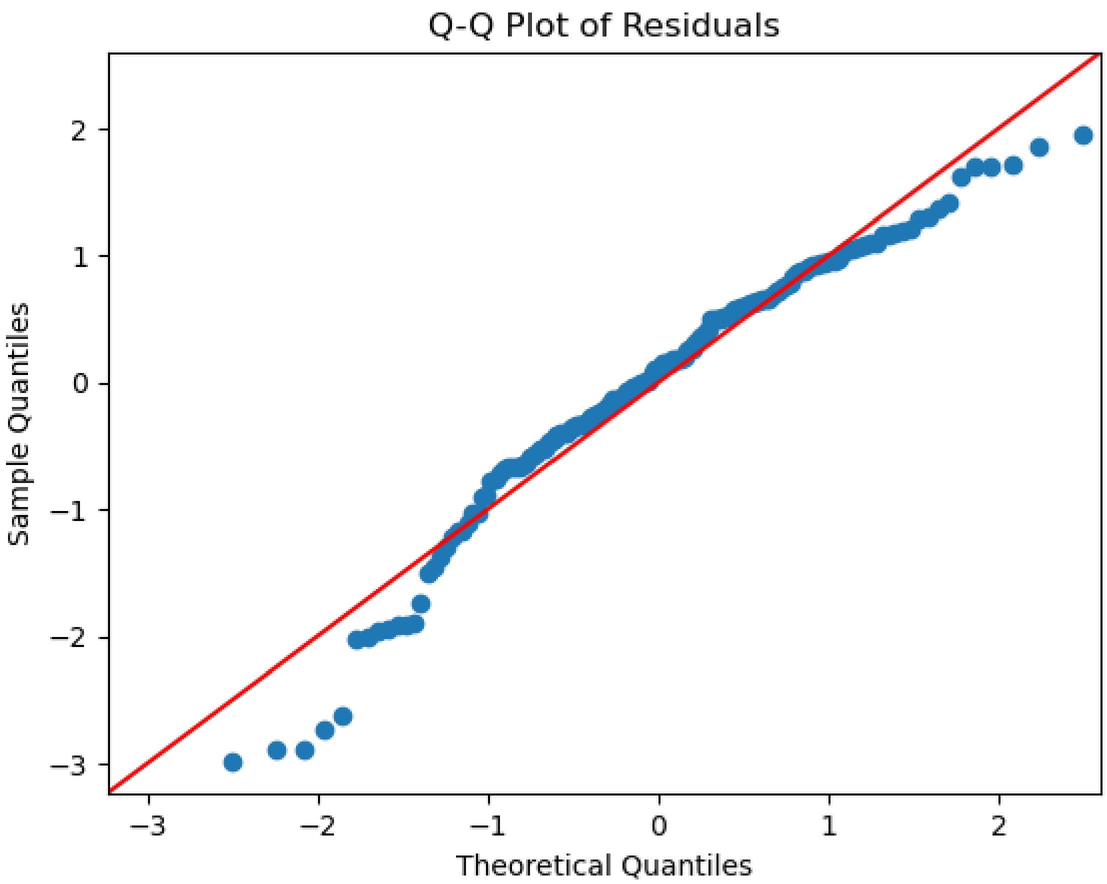
This is a strong model with a good explanatory power.

Question 6: Can we assume that errors follow 'Normal' distribution?

To check for normality among the residuals(errors), we ran the Shapiro-Wilk test and made a QQ plot of the residuals.

Shapiro-Wilk Test p-value:	Conclusion
0.0001	p-value <0.05 therefore we fail to reject normality

QQ-Plot of Residuals.

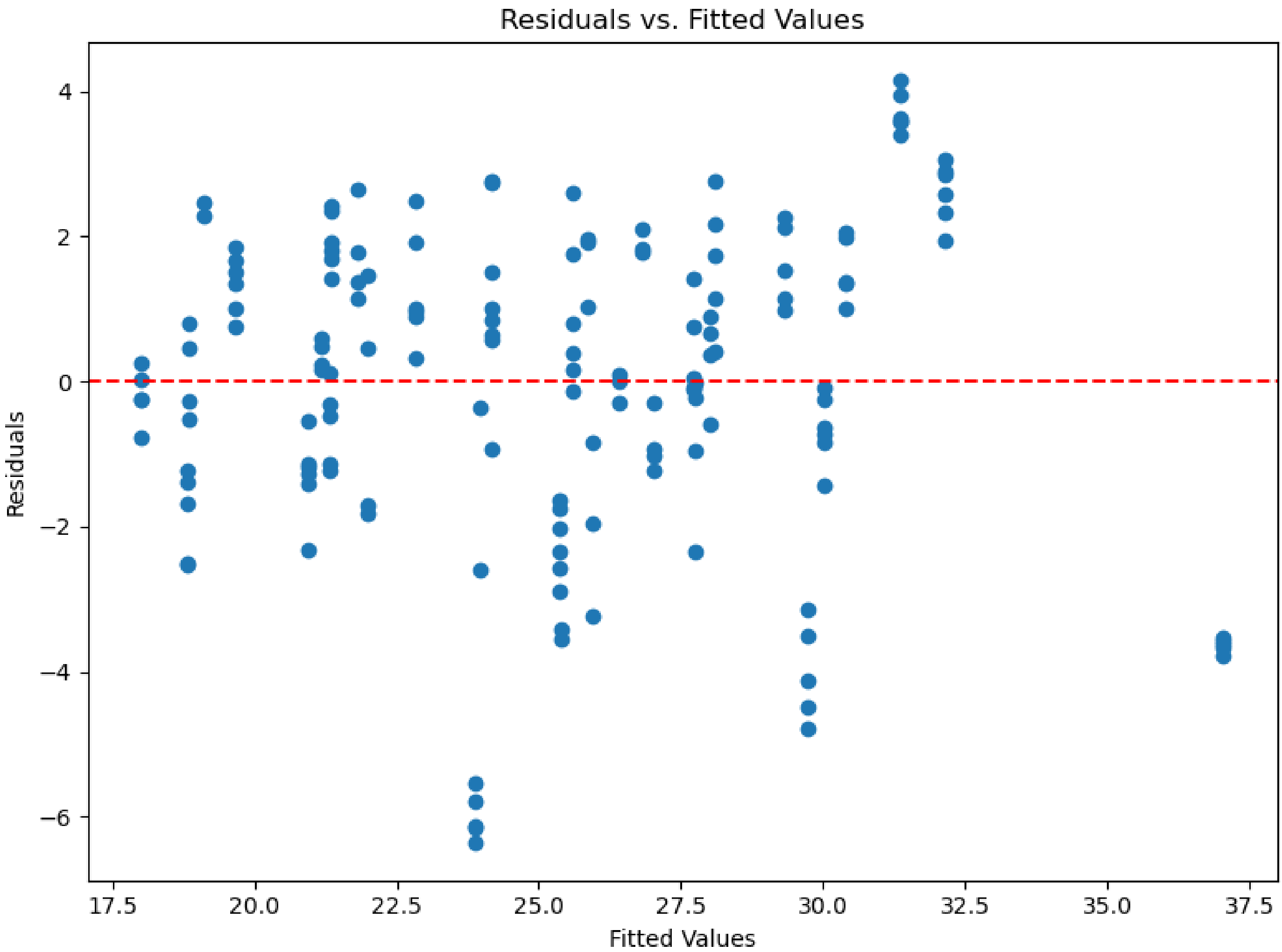


The points in the QQ plot fall along the 45^0 line, residuals(errors) are likely to be normally distributed.

Based on both the Shapiro-Wilk test and QQ plot indicate that the residuals follow normal distribution.

Question 7: Is there a Heteroscedasticity problem?

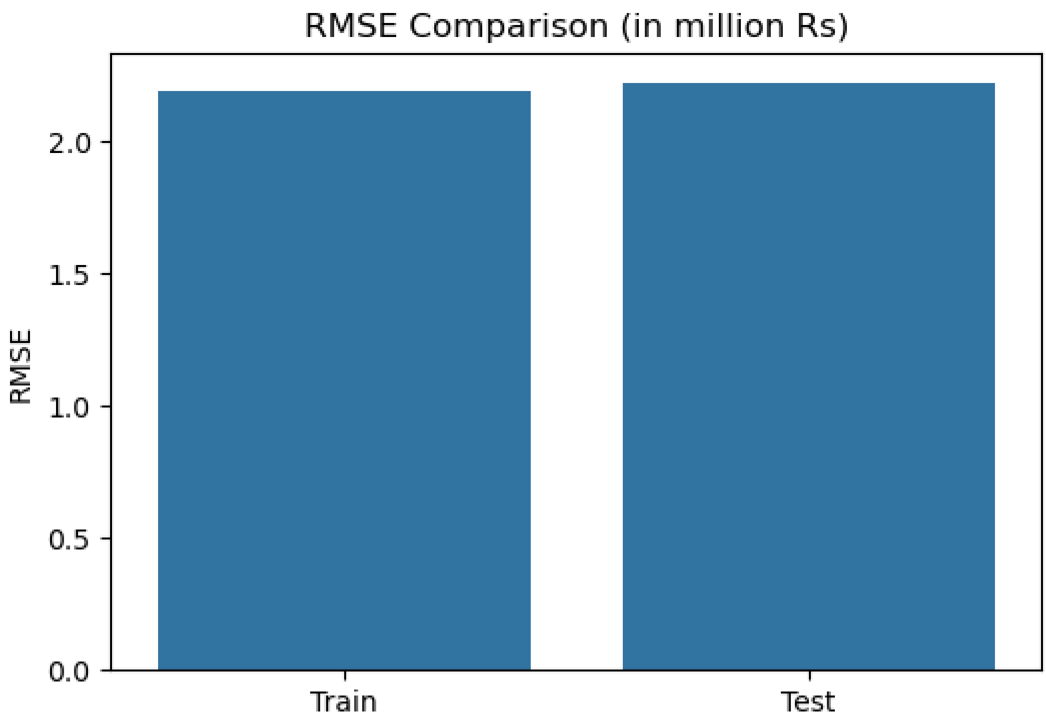
Heteroscedasticity was analysed using the Residuals vs Fitted values plot.



Seeing as the spread of variance is random, we can conclude that there is no heteroscedasticity, and we have homoscedasticity inherent between the residuals and the fitted values.

Question 7: Calculate the RMSE for the Training and Testing data

RMSE(Training Data)	**2.264**
RMSE(Test Data)	**2.291**



The RMSE is consistent between the train and the test data sets, indicating a well generalized model. Ideally you want the training RMSE to be lower than the test data which is the case in point.