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/
│   │   ├── cleaned_house_data.csv
│   │   ├── X_train.csv
│   │   ├── X_test.csv
│   │   ├── y_train.csv
│   │   └── y_test.csv
│   └── new/
│       └── incoming_house_data.csv  # ✅ For dashboard input testing
│
├── 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_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:                       0.794
Model:                            OLS   Adj. R-squared:                  0.791
Method:                 Least Squares   F-statistic:                     249.0
Date:                Sat, 19 Apr 2025   Prob (F-statistic):           3.03e-66
Time:                        17:24:40   Log-Likelihood:                -436.96
No. Observations:                 198   AIC:                             881.9
Df Residuals:                     194   BIC:                             895.1
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -9.5423      1.744     -5.472      0.000     -12.982      -6.103
area           0.0346      0.002     17.111      0.000       0.031       0.039
distance      -1.8704      0.162    -11.564      0.000      -2.189      -1.551
schools        1.3187      0.371      3.552      0.000       0.586       2.051
==============================================================================
Omnibus:                       12.632   Durbin-Watson:                   1.558
Prob(Omnibus):                  0.002   Jarque-Bera (JB):               13.367
Skew:                          -0.629   Prob(JB):                      0.00125
Kurtosis:                       3.191   Cond. No.                     1.16e+04
==============================================================================
```

EXPLANATION:

Variable	Coefficient	P-value	Interpretation
Carpet Area	0.0346	<0.05	For each additional **sq. ft.**, the selling price increases by **0.035 million Rs**, holding other variables constant.
Distance to Nearest Metro	-1.8704	<0.05	For each additional **km away from metro**, the price **decreases by 1.87 million Rs**, all else equal.
Schools	1.3187	<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 = 0.794** → ~79% of the variation in house price can be explained by this model.

Adjusted R^2 = 0.791 → 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.

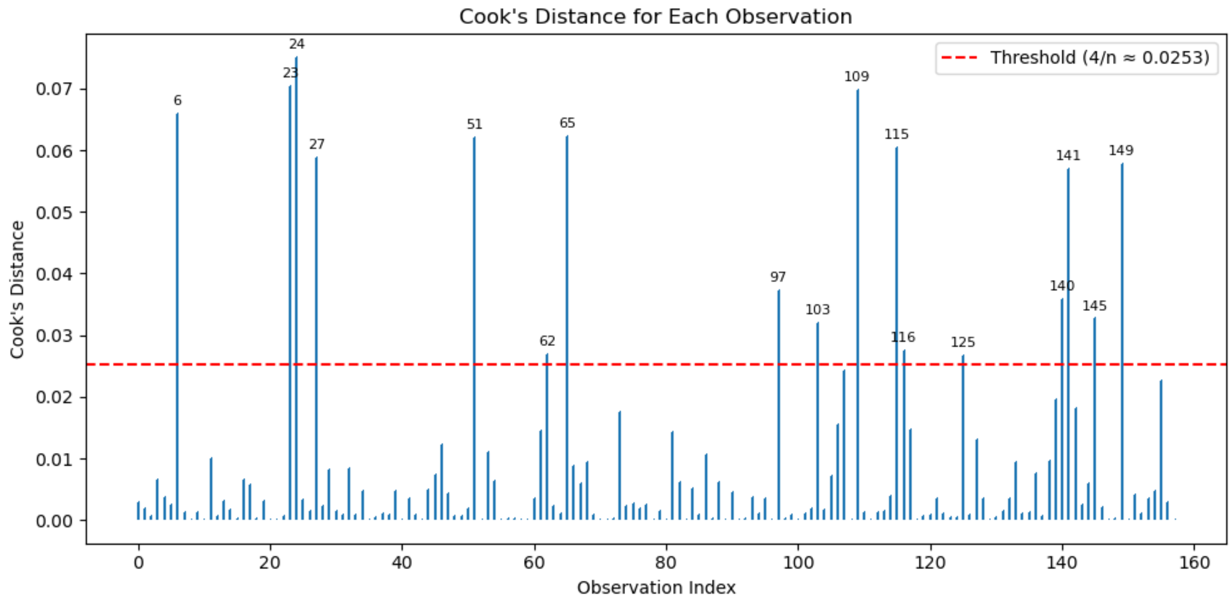
VIF Output:

	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:                       0.797
Model:                            OLS   Adj. R-squared:                  0.794
Method:                 Least Squares   F-statistic:                     229.4
Date:                Sat, 19 Apr 2025   Prob (F-statistic):           2.17e-60
Time:                        17:06:32   Log-Likelihood:                -391.07
No. Observations:                 179   AIC:                             790.1
Df Residuals:                     175   BIC:                             802.9
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -9.7951      1.827     -5.362      0.000     -13.401      -6.190
area           0.0349      0.002     16.370      0.000       0.031       0.039
distance      -1.8288      0.168    -10.898      0.000      -2.160      -1.498
schools        1.2644      0.390      3.240      0.001       0.494       2.035
==============================================================================
Omnibus:                       10.245   Durbin-Watson:                   1.721
Prob(Omnibus):                  0.006   Jarque-Bera (JB):               10.722
Skew:                          -0.599   Prob(JB):                      0.00470
Kurtosis:                       3.069   Cond. No.                     1.17e+04
==============================================================================
```

Model Summary Comparison Original vs Cleaned(Influential points removed)

Variable	Coefficient		P-value		Conclusion
	Original	Cleaned	Original	Cleaned	
Carpet Area	0.0346	0.0349	< 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	-1.8288	< 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	1.2644	< 0.05	< 0.05	There is no significant difference between the two models as it relates to schools nearby

R^2 and Adjusted R^2 Value comparison

Metric	Original Model	Cleaned Model
R^2	0.794	**0.797**
Adjusted R^2	0.791	**0.794**

Key Takeaway:

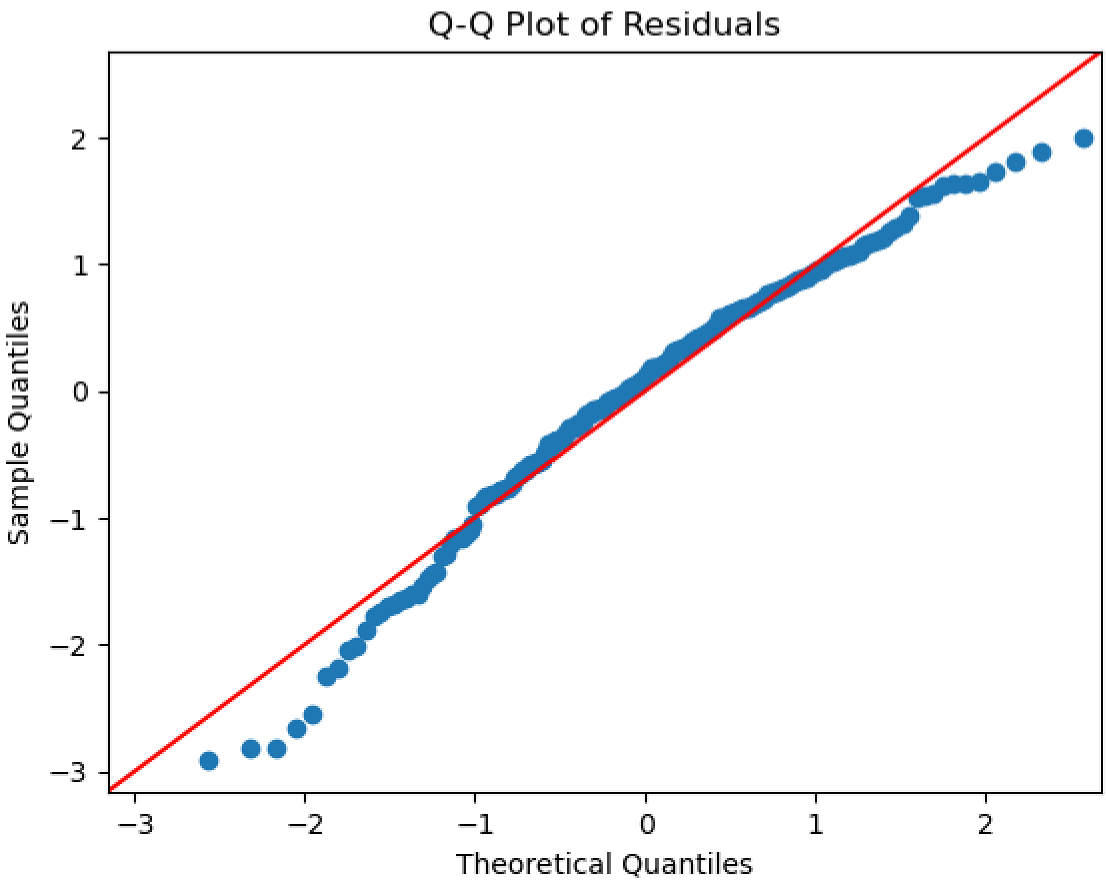
There is no significant difference in the model fit between the two models

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
0.0004	p-value <0.05 therefore we fail to reject normality

QQ-Plot of Residuals.

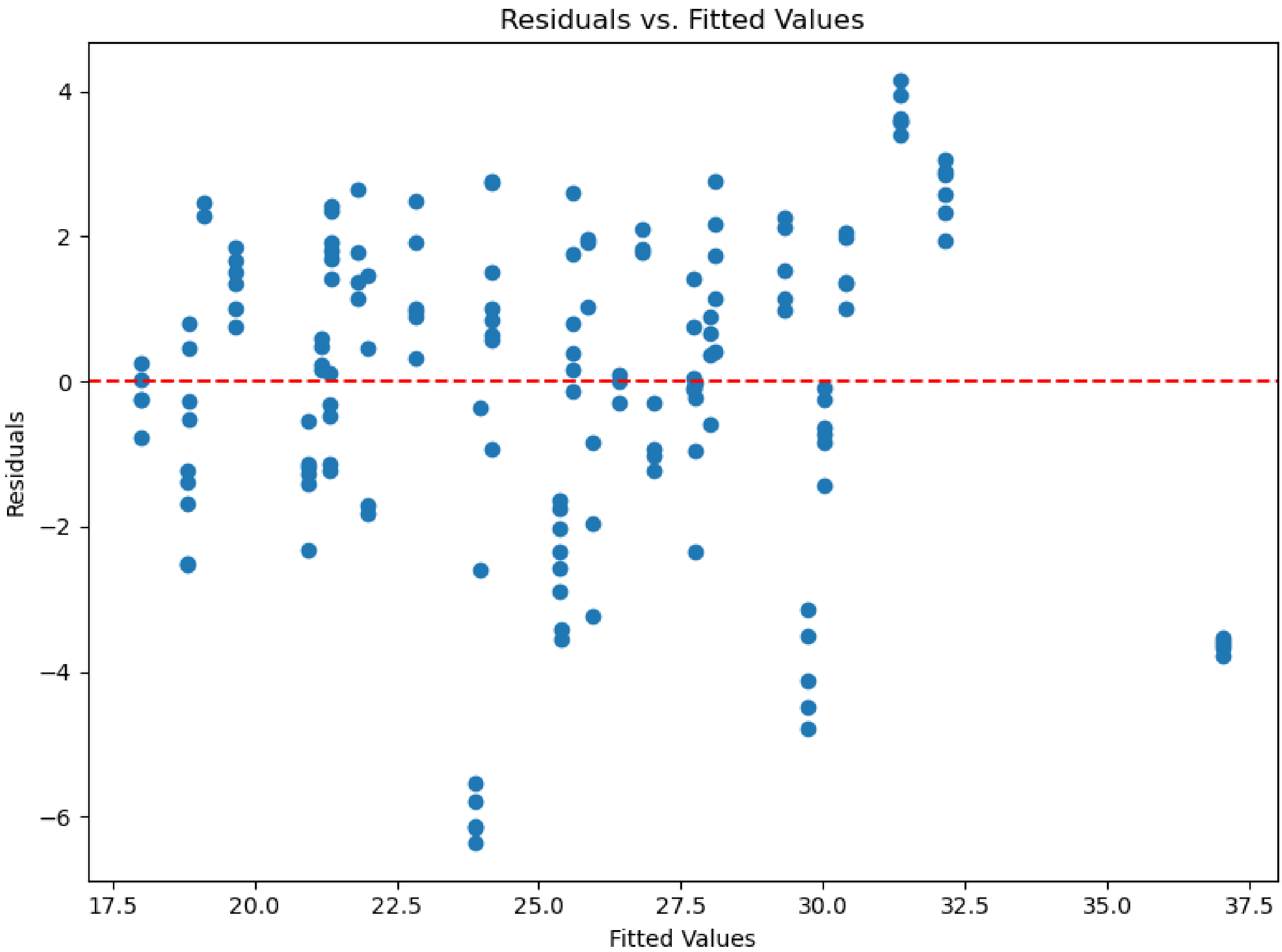


The points in the QQ plot fall along the 45^{0} line, residuals(errors) are likely to be normally distributed.

Both the Shapiro-Wilk test and QQ plot indicate that the residuals follow normal distribution.

Question 7: Is there a Heteroscedasticity problem?

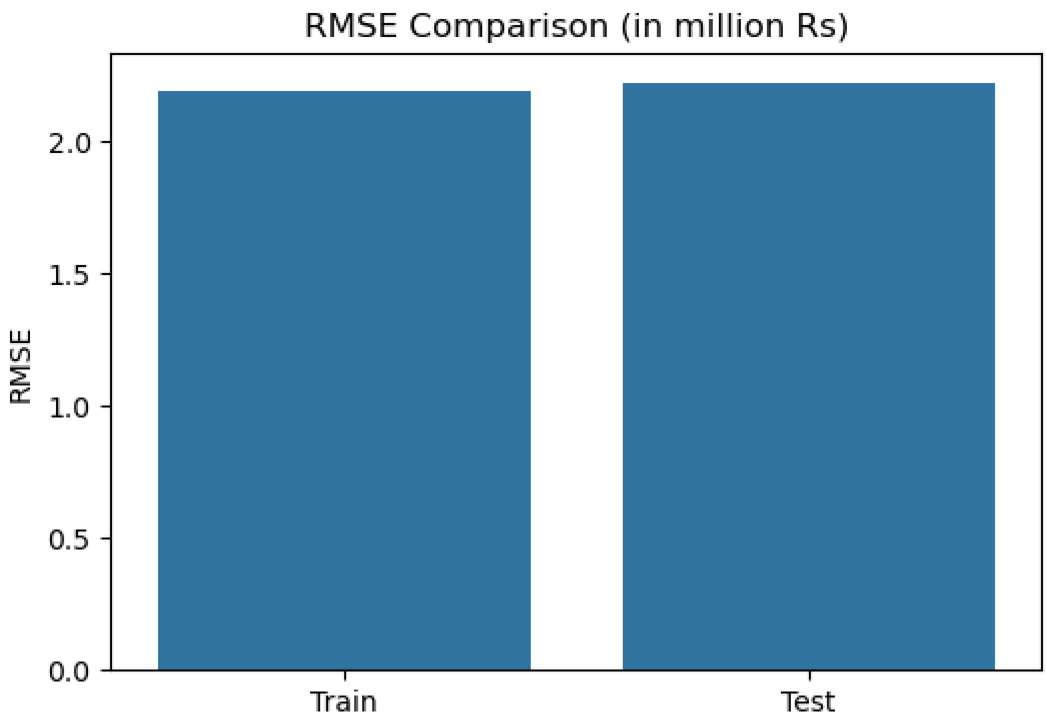
Heteroscedasticity was analysed using the Residuals vs Fitted values plot.



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)	**2.193**
RMSE(Test Data)	**2.222**



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.