Income Prediction using Multiple Regression

Kartik Sharma
Statistics For Data Analysis
National College Of Ireland
Dublin, Ireland
x21125813@student.ncirl.ie

***Abstract*—This is a project report explaining how the income of an individual can be predicted using a multiple regression. This project used Python and SPSS for the cleaning, EDA, selecting the critical features for the prediction of income and then finally creating and evaluating the model.**

***Keywords*— multiple regression, EDA**

I. Introduction

The regression is method/statistical way to explain a relationship between a dependent variable and one or more predictors/independent variables using an equation. The Multiple regression or multiple linear regression model creates an equation / relationship between dependent variable (Y) and more than one independent variables ($X_1 \ldots X_n$) so that we can predict the Y by providing the X(s).

For the prediction of income, the dataset of 4508 individuals is used where for every individual data was recorded over 13 data points ranging from years spent in education to owning a house. Out these 12 (as one of the features is income), we'll be finding most critical features that can predict an individual's income.

For this project we have kept the minimum accuracy of the model to 70% and a minimum correlation of 0.5 to conclude that the variables are correlated.

II. Methodology

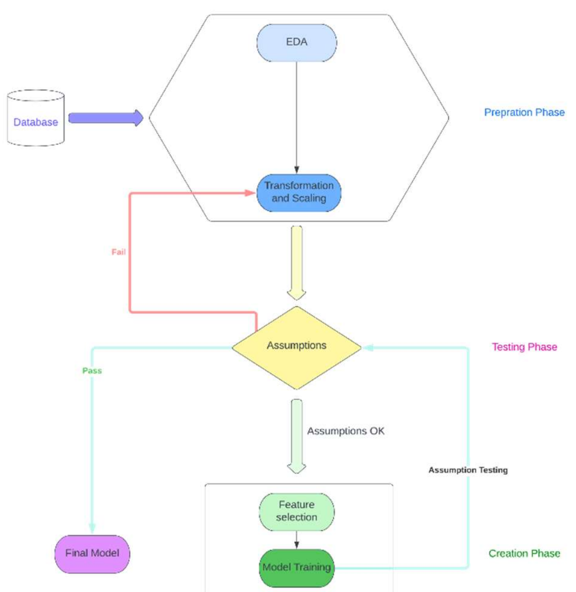

Fig. 1. Methodology Diagram [1]

The Methodology is referenced from [1] but is re-designed in this project.

A. Prepration Phase

A.1. Exploratory Data Analysis (EDA)

The Exploratory Data Analysis is an initial analysis done on the whole raw data in order to get information about the nature and characteristics of the variables/features present in the data set. It also tells how the different features are distributed in the dataset.

How does it help us? based on this analysis we can make a list of ordinals, non-ordinals, temporal, discrete and continuous features. We can use this list to transform features based on there type and the distribution.

"Skewness refers to a distortion or asymmetry that deviates from the symmetrical bell curve, or normal distribution, in a set of data. If the curve is shifted to the left or to the right, it is said to be skewed. Skewness can be quantified as a representation of the extent to which a given distribution varies from a normal distribution."[2]

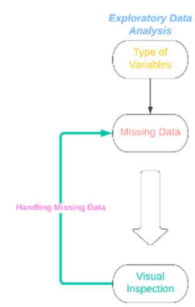

Fig. 2. Steps in EDA

Step 1: Types of Variables

Table 1

Type of Variables

Feature	Type	Sub-Type
edcat	Categorical	Non-Ordinal
default	Categorical	Binary
jobsat	Categorical	Ordinal
homeown	Categorical	Binary
cars	Discrete	
yrsed	Discrete	
yrsempl	Discrete	
address	Discrete	
age	Discrete	
income	Discrete	
creddebt	Continuous	
othdebt	Continuous	
carvalue	Continuous	

Step 2: Checking Missing Values

Below code snippet return a list of features containing missing values: -

```
features_miss_val= [features for features in dataset.columns if dataset[features].isnull().sum()>1]
print(features_miss_val)
```

The List returned is empty, hence we can conclude that data contains no missing values.

Step 3: Visual Inspection

a. Income: -

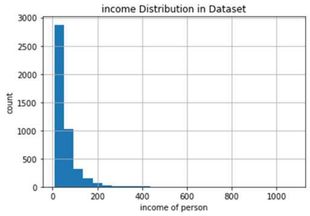

Fig. 3. skewness in Income

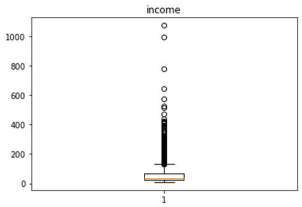

Fig. 4. outliers in Income

Result of inspection: -
1. Highly Right Skewed
2. Skewness - 5.233689457953158
3. Large Number of Outliers

b. Credit card debt (in thousands) creddebt: -

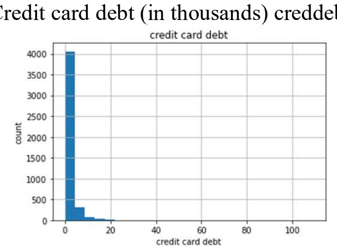

Fig. 5. skewness id creddebt

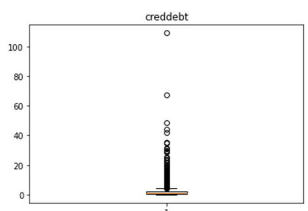

Fig. 6. outliers in creddebt

Result of inspection: -
1. Highly Right Skewed
2. Skewness - 10.962120273419949
3. Large Number of Outliers

c. Other Debt (in thousands) othdebt: -

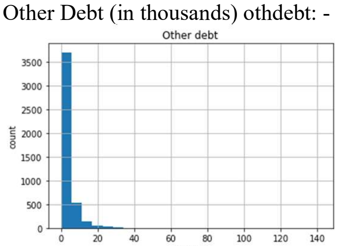

Fig. 7. skewness in othdebt

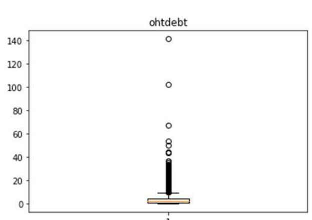

Fig. 8. outliers in othdebt

Result of inspection: -
1. Highly Right Skewed
2. Skewness - 7.69338390729621
3. Large Number of Outliers

d. Value of the primary Vehicle (carvalue): -

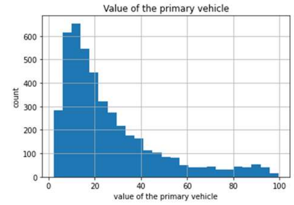

Fig. 9. Skewness in carvalue

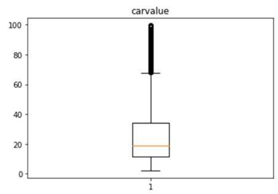

Fig. 10. outliers in carvalue

Result of inspection: -
1. Highly Right Skewed
2. Skewness - 1.5301894027033907
3. Large Number of Outliers

Summary of the Data: -

Table 2

Statistic Summary of Dataset

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
age	4508	18	79	46.93	17.665	.094	.036	-1.175	.073
yrsed	4508	6	23	14.53	3.286	.020	.036	-.606	.073
edcat	4508	1	5	2.67	1.214	.257	.036	-.993	.073
yrsempl	4508	0	52	9.72	9.651	1.249	.036	1.066	.073
income	4508	9	1073	55.41	56.514	5.234	.036	57.644	.073
creddebt	4508	.000000	109.072596	1.89786636	3.542646400	10.962	.036	237.337	.073
othdebt	4508	.000000	141.459150	3.69144686	5.378583009	7.693	.036	133.643	.073
default	4508	0	1	.24	.426	1.225	.036	-.500	.073
jobsat	4508	1	5	2.96	1.377	.031	.036	-1.231	.073
homeown	4508	0	1	.63	.483	-.532	.036	-1.718	.073
address	4508	0	57	16.37	12.368	.717	.036	-.183	.073
cars	4508	1	8	2.37	1.158	.875	.036	.799	.073
carvalue	4508	2.2	99.6	26.082	20.8626	1.530	.036	1.884	.073
Valid N (listwise)	4508								

A.2. Transformation and Scaling

From EDA, we can conclude that the variables are skewed and have high number of outliers.

In this step we'll be doing Box-Cox transformation on the dependent variable (income). "We won't be doing transformation on the independent variables, as Linear Regression Model is not affected by the distribution of the predictors." [4]

"A box-cox transformation is a commonly used method for transforming a non-normally distributed dataset into a more normally distributed one." [3]

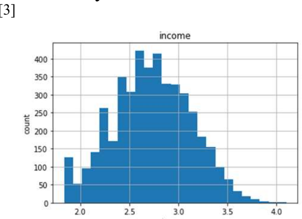

Fig. 11. Income After Transformation

Now, we'll be handling the outliers.

"An *outlier* is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal. Before abnormal observations can be singled out, it is necessary to characterize normal observations." [5]

If a data point is less than the value of 1st Quartile or is greater than the value of 3rd Quartile, then we consider that data point as an Outlier. The outliers need to handle as they can cause drastic shift in the mean.

Consideration while handling outliers: -

1. Check if that outlier is necessary or not,
   E.g., Age of person = 180 years is an outlier in income dataset.
2. We can trim the insignificant outliers.
3. We can perform Quantile-based Flooring and Capping [6]

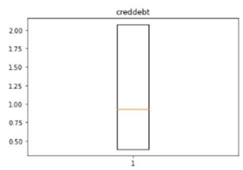

Fig. 12. creddebt after handling outliers

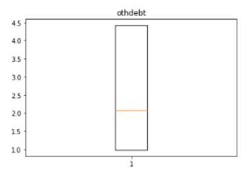

Fig. 13. othdebt after handling outliers

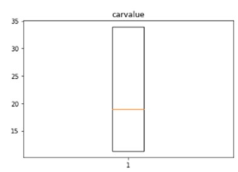

Fig. 14. carvalue after handling outliers

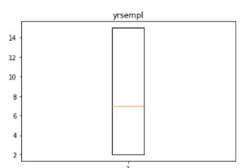

Fig. 15. yrsempl after handling outliers

How Quantile-based Flooring and Capping [6] was applied?

1. For each variable the Q1, IQR and Q3 was calculated.
2. The Data points greater than Q3 or less than Q1 were replaced by the Q3 & Q1 respectively.

The Encoding is required for the level of education (edcat) as it is a non-Ordinal Categorical variable. The best encoding in this case will be One-Hot [7] encoding.
The One-Hot [7] encoding takes every value from the variable and creates a binary data column for each type
E.g., if x has values (1,2,3), we'll have 2 new columns with binary data in it, like this (for 1 all columns will be 0): -

Table 3

One-Hot Encoding Example

X	2	3
1	0	0
2	1	0
3	0	1

B. Testing Phase (Before Model Creation)

B.1. Assumptions

1. Correlation between the variables: -

Null Hypothesis 1: - The Independent features are Highly corelated with each other
Null Hypothesis 2: - The Independent features are multi-Correlated

Table 4

Correlation Matrix

	yrsempl	income	creddebt	othdebt	default	jobsat	homeown	address	cars	carvalue	2	3	4	5
yrsempl	1.000000	0.311253	0.218660	0.245371	-0.363322	0.503632	0.024256	0.611117	0.006647	0.368083	0.019866	-0.045318	-0.094877	-0.077902
income	0.311253	1.000000	0.570192	0.645215	-0.033584	0.250681	0.156806	0.196321	0.056096	0.889347	-0.103246	0.022602	0.114828	0.123884
creddebt	0.218660	0.570192	1.000000	0.600492	0.160191	0.180391	0.070693	0.135896	0.022499	0.530280	-0.031156	-0.005258	0.075184	0.060650
othdebt	0.245371	0.645215	0.600492	1.000000	0.094991	0.197541	0.093721	0.176155	0.041540	0.604885	-0.068500	-0.004209	0.094432	0.082026
default	-0.363322	-0.033584	0.160191	0.094991	1.000000	-0.203444	-0.027361	-0.357760	0.013890	-0.070019	-0.001450	0.010645	0.064227	0.040252
jobsat	0.503632	0.250681	0.180391	0.197541	-0.203444	1.000000	0.022143	0.350661	0.013886	0.272611	-0.004409	-0.011538	-0.039797	-0.032055
homeown	0.024256	0.156806	0.070693	0.093721	-0.027361	0.022143	1.000000	0.137432	0.017491	0.140499	-0.023626	0.023809	0.036212	0.001901
address	0.611117	0.196321	0.135896	0.176155	-0.357760	0.350661	0.137432	1.000000	-0.000072	0.252117	-0.051968	-0.023364	-0.008046	-0.000447
cars	0.006647	0.056096	0.022499	0.041540	0.013890	0.013886	0.017491	-0.000072	1.000000	0.055792	0.000054	0.013717	0.008527	-0.017404
carvalue	0.368083	0.889347	0.530280	0.604885	-0.070019	0.272611	0.140499	0.252117	0.055792	1.000000	-0.094809	0.005893	0.121886	0.103692
2	0.019866	-0.103246	-0.031156	-0.068500	-0.001450	-0.004409	-0.023626	-0.051968	0.000054	-0.094809	1.000000	-0.342529	-0.359345	-0.188303
3	-0.045318	0.022602	-0.005258	-0.004209	0.010645	-0.011538	0.023809	-0.023364	0.013717	0.005893	-0.342529	1.000000	-0.268046	-0.140461
4	-0.094877	0.114828	0.075184	0.094432	0.064227	-0.039797	0.036212	-0.008046	0.008527	0.121886	-0.359345	-0.268046	1.000000	-0.147356
5	-0.077902	0.123884	0.060650	0.082026	0.040252	-0.032055	0.001901	-0.000447	-0.017404	0.103692	-0.188303	-0.140461	-0.147356	1.000000

Please Note: - 2, 3, 4, 5 are the new columns created after One-Hot encoding of the edcat (Level of education) and the edcat is dropped from the dataset.

The variables with Pearson's correlation value >=0.5 between them are considered highly correlated and we can either remove one of them or we can aggregate them during model training. For those variables (except *vs income*) our assumption of *Multicorrelation* among the independent variables *Fails*.

(The Assumption of Multicorrelation states "there is no correlation between the Independent Variables")

Since the Assumption of Multicorrelation failed, therefore we fail to reject the Null Hypothesis 1.

Though income and carvalue depict very high correlation, but since income is dependent variable, we can say that carvalue is essential in predicting the income.

To check for *Null Hypothesis 2,* we'll look for VIF (Variance Inflation Factor). The table below only shows features with VIF>=10

Table 5

VIF of the features

Feature	VIF
age	32.790978
yrsed	57.512400
carvalue	12.639053

The feature with VIF >=10 should be considered for the removal from the dataset as these features are not much independent.
Before considering a feature for the removal, we checked how much it is good in predicting the income. To decide this, we compared the correlation of each feature with high VIF with income and found age and yrsed *(years of education)* have least correlation of 0.1 approx. Therefore, these features (age & yrsed *(years of education)*) were removed from the dataset.

Table 6

Correlation of Income with High VIF features

Feature vs Income	correlation
income & age	0.12833155470874583
income & yrsed	0.19798343138251495
income & carvalue	0.8893470860662386

C. Creation Phase

C.1. Feature Selection and Model Training

In Feature Selection we try select only the most relevant feature for our model, this process helps in avoiding the overfitting and underfitting of the linear regression model.
In Model Training we use the combination of the selected features, to create a highly accurate model.

We combined the Feature selection and Model training phase together to generate a model with best subset of features.
For this we applied Subset Algorithm [8]

Subset Algorithm for Model Training and Selection: -

1. The Dataset is split in Test and Training data
2. The Train Data is used to train various models based on different combinations of the features.
3. The Test Data is used to test these models in every iteration
4. All the models with their respective adj. R^2 and RSS value are stored in list.
5. The model with the respective subset that has highest adj. R^2 is selected.

Total no. of combination on which models are created and tested is 2^n
Where n = no. of features, therefore we had $2^{14} = 16384$ combinations

Out of these combinations only 2 models fitting the criteria (given below) were selected with an approx. accuracy of 81.9%.
One of the models was dropped as it contained more feature than the other.

The selected model is then tested for various assumptions of linear regression model, if selected model is satisfactorily satisfying the assumptions, we consider the relationship is accurate enough for prediction.

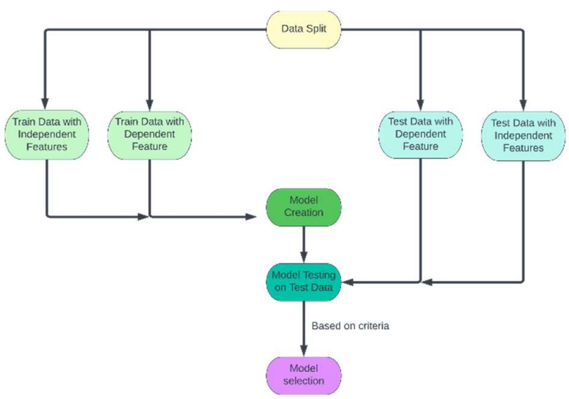

Fig. 16. Model Training and Testing

The criteria based on which the model is selected: -

1. The model with accuracy >70% (adj. R^2>0.70)
2. The model with least features
3. The model with least skewness {*in Testing Phase (After model creation)*}

Equation from the selected model: -
*income = 1.8627 + credebt *0.0520 + othdebt *0.0380 – default *0.0210 + homeown *0.0297 – address *0.0014 + cars *0.0026 + carvalue *0.0331 + edcat(level-3) *0.0202 + edcat(level-4) *0.0062 + edcat(level-5) *0.0511*

D. Testing Phase(After model creation)

D.1. Assumptions of Linear Regression Model

The Gauss Markov Theorem [9]

Some important information regarding Gauss-Markov Theorem

- ➢ We say that an estimator is linear if it is a linear function of y1, ..., yn. the OLS estimators b1, b2 are linear estimators.
- ➢ We say that an unbiased estimator is more efficient than another unbiased estimator if it has a smaller variance
- ➢ We say that an estimator is Blue (Best Linear Unbiased Estimator) if it is linear and unbiased and more efficient than any other linear and unbiased estimator.

"The Gauss-Markov theorem states that if your linear regression model satisfies the first six classical assumptions, then ordinary least squares (OLS) regression produces unbiased estimates that have the smallest variance of all possible linear estimators"[9].

The Assumptions of Linear Regression: -

1. *Linear Relationship*
   Aims at finding a linear relationship between the independent and dependent variables.
   Scatter plots are used to visually determine this assumption.

Fig. 17. Scatterplots

From above plot we can conclude that the income is satisfactorily linearly related with the independent variables

2. *Variables follow a normal distribution*
   This assumption ensures that for each value of independent variable, the dependent variable is a random variable following a normal distribution and its mean lies on the regression line.
   Quantile-Quantile plot is visual way to inspect this assumption.

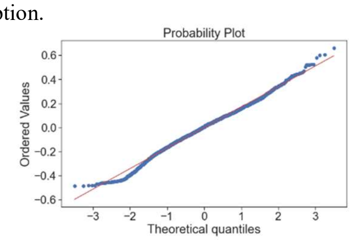

Fig. 18. Q-Q plot

Form the above plot we can infer that the Assumption 2 is almost satisfied

3. *Little or no multicollinearity*
   It tests the correlation between the independent variables.
   If multicollinearity exists between them, they are no longer independent.
   We checked the VIF of each independent variable and the correlation between each independent variable.

Table 7

Correlation Matrix

	creddebt	othdebt	default	homeown	address	cars	carvalue	3	4	5
creddebt	1	0.61	0.17	0.067	0.12	0.011	0.51	0.0086	0.072	0.048
othdebt	0.61	1	0.087	0.087	0.17	0.026	0.6	-0.011	0.095	0.066
default	0.17	0.087	1	-0.025	-0.35	0.0098	-0.077	0.034	0.042	0.028
homeown	0.067	0.087	-0.025	1	0.14	0.028	0.14	0.019	0.039	0.0022
address	0.12	0.17	-0.35	0.14	1	0.01	0.24	-0.034	-0.017	0.0018
cars	0.011	0.026	0.0098	0.028	0.01	1	0.059	0.0064	-0.027	-0.0022
carvalue	0.51	0.6	-0.077	0.14	0.24	0.059	1	-0.0028	0.12	0.086
3	0.0086	-0.011	0.034	0.019	-0.034	0.0064	-0.0028	1	-0.27	-0.14
4	0.072	0.095	0.042	0.039	-0.017	-0.027	0.12	-0.27	1	-0.14
5	0.048	0.066	0.028	0.0022	0.0018	-0.0022	0.086	-0.14	-0.14	1

By observing the above table, we can conclude that there are no signs of multicollinearity as no correlation is >=0.8.

4. *Little or no Autocorrelation*

This assumption is like the above assumption, only the exception is, it applies to the residuals of the linear regression model.

We can test the assumption with Durbin-Watson test.

Values from Durbin-Watson test are in range 0-4 where if d = 2, we accept that there is no autocorrelation.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 income   R-squared:                       0.816
Model:                            OLS   Adj. R-squared:                  0.816
Method:                 Least Squares   F-statistic:                     1338.
Date:                Sun, 06 Mar 2022   Prob (F-statistic):               0.00
Time:                        14:13:39   Log-Likelihood:                 1054.4
No. Observations:                3019   AIC:                            -2087.
Df Residuals:                    3008   BIC:                            -2021.
Df Model:                          10
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      1.8627      0.011    165.153      0.000       1.841       1.885
creddebt       0.0520      0.006      8.695      0.000       0.040       0.064
othdebt        0.0380      0.003     12.180      0.000       0.032       0.044
default       -0.0210      0.008     -2.568      0.010      -0.037      -0.005
homeown        0.0297      0.007      4.537      0.000       0.017       0.043
address       -0.0014      0.000     -5.053      0.000      -0.002      -0.001
cars           0.0025      0.003      0.917      0.359      -0.003       0.008
carvalue       0.0331      0.000     72.823      0.000       0.032       0.034
edcat3         0.0202      0.008      2.503      0.012       0.004       0.036
edcat4         0.0062      0.008      0.758      0.449      -0.010       0.022
edcat5         0.0511      0.013      4.060      0.000       0.026       0.076
==============================================================================
Omnibus:                       22.863   Durbin-Watson:                   2.005
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               27.825
Skew:                          -0.133   Prob(JB):                     9.08e-07
Kurtosis:                       3.388   Cond. No.                         126.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Fig. 19. Model Summary

By observing the value from Durbin-Watson test(d=2.005), we can conclude that there is no autocorrelation.

5. *Data is homoscedastic*

According to this assumption, the error terms along the regression line are equal.
It is also applied to the residuals of the linear regression model.
This assumption can be tested visually using a scatter plot of the residuals.

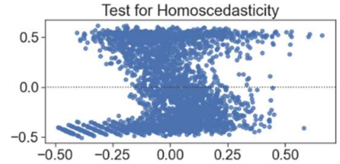

Fig. 20. Scatter plot for residuals

The above plot fails to provide any signs of heteroscedastic pattern in residual; therefore, we can evidently conclude that the data is Homoscedastic.

Since all the Assumptions of Linear Regression Model satisfied to an extent where we can say our model follows Gauss-Markov theorem with an accuracy of 81.6% in predicting the Income of an Individual.

E. *Abbreviations and Acronyms*

1. LRM - Linear Regression Model
2. E.g. - For Example
3. VIF - Variance Inflation Factor
4. Corr - Correlation
5. creddebt - Credit Card Debt (in thousands)
6. edcat - Level of education (1,2,3,4,5)
7. yrsed - Years of education

References

[1] Migliorini, Matteo "Machine Learning Pipelines with Modern Big Data Tools for High Energy Physics"
Retrieved from :- Machine Learning Pipelines with Modern Big Data Tools for High Energy Physics - CERN Document Server

[2] Skewness by James Chen , Reveiwed by Charles Potters
Retrieved from :- Skewness Definition, Formula, & Calculation (investopedia.com)

[3] Zach – 'How to Perform a Box-Cox Transformation in Python'
Retrieved from :- How to Perform a Box-Cox Transformation in Python - Statology

[4] Songhao Wu "Is Normal Distribution Necessary in Regression?"
Retrieved from :- Is Normal Distribution Necessary in Regression? How to track and fix it? | by Songhao Wu | Towards Data Science.

[5] "Engineering Statistics Handbook"
Retrieved from :- 7.1.6. What are outliers in the data? (nist.gov)

[6] Deepika Singh "Cleaning up Data from Outliers"
Retrieved from :- Cleaning up Data Outliers with Python | Pluralsight

[7] Jason Browniee "Why One-Hot Encode Data in Machine Learning?"
Retrieved from :- Why One-Hot Encode Data in Machine Learning? (machinelearningmastery.com)

[8] Xavier Bourret Sicotte "Choosing the optimal model: Subset selection"
Retrieved from :- Choosing the optimal model: Subset selection — Data Blog (xavierbourretsicotte.github.io)

[9] Jim Frost "*The Gauss-Markov Theorem and BLUE OLS Coefficient Estimates*"
Retrieved from :- The Gauss-Markov Theorem and BLUE OLS Coefficient Estimates - Statistics By Jim