**D208 Performance Assessment – Task 1**

NBM3 – NBM3 Task 1: Linear Regression Modeling

Table of Contents

[**Part I: Research Question** 1](#_Toc130385913)

[A.  Describe the purpose of this data analysis by doing the following: 1](#_Toc130385914)

[**Part II: Method Justification** 2](#_Toc130385915)

[B.  Describe multiple linear regression methods by doing the following: 2](#_Toc130385916)

[**Part III: Data Preparation** 4](#_Toc130385917)

[C.  Summarize the data preparation process for multiple linear regression analysis by doing the following: 4](#_Toc130385918)

[**Part IV: Model Comparison and Analysis** 10](#_Toc130385919)

[D.  Compare an initial and a reduced linear regression model by doing the following: 10](#_Toc130385920)

[E.  Analyze the data set using your reduced linear regression model by doing the following: 13](#_Toc130385921)

[**Part V: Data Summary and Implications** 15](#_Toc130385922)

[F.  Summarize your findings and assumptions by doing the following: 15](#_Toc130385923)

[**Part VI: Demonstration** 16](#_Toc130385924)

[G.  Provide a Panopto video recording that includes the presenter and a vocalized demonstration of the functionality of the code used for the analysis of the programming environment, including the following elements: 16](#_Toc130385925)

[H.  List the web sources used to acquire data or segments of third-party code to support the application. Ensure the web sources are reliable. 17](#_Toc130385926)

[I.  Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized. 17](#_Toc130385927)

[J.  Demonstrate professional communication in the content and presentation of your submission. 17](#_Toc130385928)

# **Part I: Research Question**

## A.  Describe the purpose of this data analysis by doing the following:

1.  Summarize **one** research question that is relevant to a real-world organizational situation captured in the data set you have selected and that you will answer using multiple linear regression in the initial model.

Which numerical variables influence or have a positive correlation to a patient’s additional charges (Additiona\_charges) accrued during hospitalization?

H0 – There are no explanatory variables within the medical dataset that show any significant sway of additional charges (Additional\_charges) that a patient accrues during their hospitalization.

H1 – Some explanatory variables have been shown to have significant sway of the additional charges (Additional\_charges) a patient accrues during their hospitalization.

2.  Define the goals of the data analysis.

Note: Ensure that your goals are within the scope of your research question and are represented in the available data.

The goal of the analysis it to use MLR (multiple linear regression) to assess which variables, if any, of the medical dataset have a positive correlation to ‘Additional Charges’ of a patient’s hospitalization. With that a model can be created using OLS (Ordinary Least Squares) as the means to create a model based on predictions and test if it’s a fit.

Through this analysis and model creation you have a tool that allows you to calculate the potential additional cost of patients who get certain values that fit the model of the explanatory variables with positive correlation to the dependent variable.

# **Part II: Method Justification**

## B.  Describe multiple linear regression methods by doing the following:

1.  Summarize **four** assumptions of a multiple linear regression model.

|  |  |
| --- | --- |
| Linearity/Linear Relationship | “The line of best fit through the data points is a straight line, rather than a curve or some sort of grouping factor.”  (Bevans, 2022)  With this straight line the relationship between a dependent variable and all its independent variables can be assessed. Scatter plotting is a great tool for visualizing this relationship. |
| Independence | “The next assumption of linear regression is that the residuals are independent.” (Zach, 2021).  This means that we want to make sure that there is actually independence and to do that we need to use a time series. With that time series we are confirming that the residuals of the variable are not growing larger as time goes on. |
| Homoscedasticity | Means that we need the variance of x to be consistent at every level. This will allow us to be assured that when a variable is found to be significant that we can trust it.  It the variance of x is not consistent it falls into what is called heteroscedasticity.  “When heteroscedasticity is present in a regression analysis, the results of the analysis become hard to trust. Specifically, heteroscedasticity increases the variance of the regression coefficient estimates, but the regression model doesn’t pick up on this. This makes it much more likely for a regression model to declare that a term in the model is statistically significant, when in fact it is not.” (Zach, 2021). |
| Normality | Residuals should be normally distributed to rule out anomalies. This can be checked with a Q-Q Plot. |

2.  Describe **two** benefits of using Python or R in support of various phases of the analysis.

Two benefits include:

1. A great benefit of using languages like Python and R (*I chose Python*) is that they are easy programming languages to pick up and have a wide scientific community support for packages geared toward analysis, science, and machine learning.
2. Another benefit of using these programming languages is reusability. Because your work is done in code form it can be reproduced and run again and again with small tweaks without having to start from scratch.

3.  Explain why multiple linear regression is an appropriate technique to use for analyzing the research question summarized in part I.  
  
Because multiple leaner regression is a great technique for answering the question stated above, because it allows us to test multiple explanatory/independent variable(s) against a single response/dependent/target variable. This allows us to see which variables have positive or negative correlation to the response variable in question.   
  
As it pertains to the objective stated above MLR lends itself as tool that allows us to check the significance of large amounts of potential explanatory variables together and reviewed quickly.

# **Part III: Data Preparation**

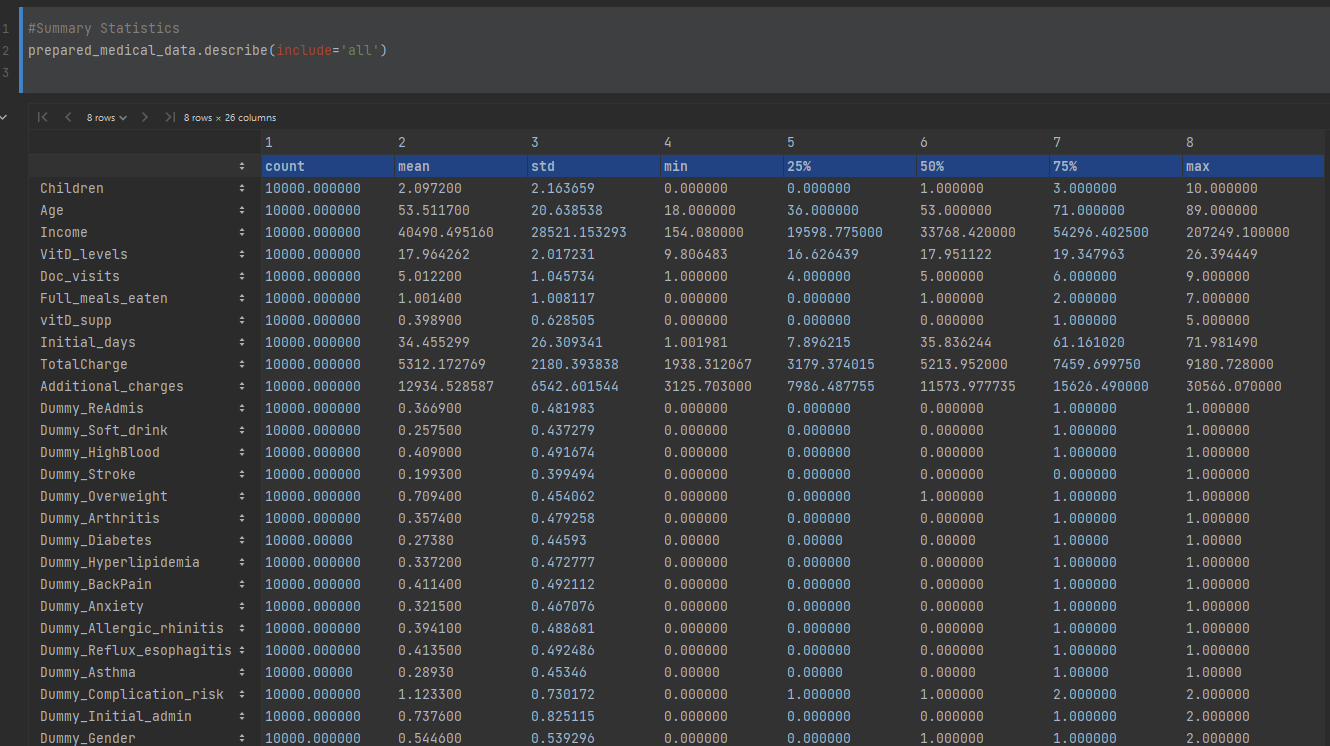
## C.  Summarize the data preparation process for multiple linear regression analysis by doing the following:

1.  Describe your data cleaning goals and the steps used to clean the data to achieve the goals that align with your research question including your annotated code.

The preparation will be minor as this dataset is a continuation of the medical\_data that was cleaned in D206. Below will still be done to clean a little further.

The data preparation process steps used for MLR (multiple linear regression) included:

1. Importing dataset into Python as a Pandas Data Frame
2. Check for missing values.
3. Compare poorly named series (columns) with the WGU provided data dictionary. Rename ones that need clarity.
4. Convert chr and str datatypes to Category.
5. ReExpress logical yes/no binaries categorical data to 1,0 for processing
6. Calculate and remove Categoricals with high cardinality values as these can be explanatory variables that lead to over fitting of models.  
     
   Code for Data Preparation:  
     
   original\_medical = medical\_clean\_data = pd.read\_csv('./Data/Medical/medical\_clean.csv')  
   #del medical\_clean\_data[medical\_clean\_data.columns[0]]  
     
   any\_missing\_values = medical\_clean\_data.isna().values.any()  
   if not any\_missing\_values:  
    print('Medical data does NOT contain any missing values\n')  
   else:  
    print('Medical data CONTAINS missing values.\n')  
     
   column\_renames = {  
    'Item1': 'Timely\_Admission'  
    ,'Item2': 'Timely\_Treatment'  
    ,'Item3': 'Timely\_Visits'  
    ,'Item4': 'Reliability'  
    ,'Item5': 'Options'  
    ,'Item6': 'Hours\_Of\_Treatment'  
    ,'Item7': 'Courteous\_Staff'  
    ,'Item8': 'Listening' #Evidence of active listening from Doctor  
   }  
   medical\_clean\_data.rename(columns=column\_renames, inplace=True)  
   #medical\_clean\_data  
     
   removal\_attributes = ['CaseOrder', 'Customer\_id', 'Interaction', 'UID',  
    'Zip', 'Lat', 'Lng', 'City', 'State', 'County',  
    'Area', 'Job', 'Marital', 'Population', 'TimeZone']  
     
   medical\_clean\_data = medical\_clean\_data.drop(columns=removal\_attributes)  
     
   category\_dtype = 'category'  
   convert\_to\_category = {  
    'Gender': category\_dtype,  
    'ReAdmis': category\_dtype,  
    'Soft\_drink': category\_dtype,  
    'Initial\_admin': category\_dtype,  
    'HighBlood': category\_dtype,  
    'Stroke': category\_dtype,  
    'Complication\_risk': category\_dtype,  
    'Overweight': category\_dtype,  
    'Arthritis': category\_dtype,  
    'Diabetes': category\_dtype,  
    'Hyperlipidemia': category\_dtype,  
    'BackPain': category\_dtype,  
    'Anxiety': category\_dtype,  
    'Allergic\_rhinitis': category\_dtype,  
    'Reflux\_esophagitis': category\_dtype,  
    'Asthma': category\_dtype,  
    'Services': category\_dtype,  
    'Timely\_Admission': category\_dtype,  
    'Timely\_Treatment': category\_dtype,  
    'Timely\_Visits': category\_dtype,  
    'Reliability': category\_dtype,  
    'Options': category\_dtype,  
    'Hours\_Of\_Treatment': category\_dtype,  
    'Courteous\_Staff': category\_dtype,  
    'Listening': category\_dtype  
   }  
     
   medical\_clean\_data = medical\_clean\_data.astype(convert\_to\_category)  
     
   #Logical categorical variables converted to numerical  
   columns\_to\_reexpress = ['ReAdmis', 'Soft\_drink', 'HighBlood', 'Stroke',  
    'Overweight', 'Arthritis', 'Diabetes', 'Hyperlipidemia',  
    'BackPain', 'Anxiety', 'Allergic\_rhinitis', 'Reflux\_esophagitis',  
    'Asthma']  
   for column in columns\_to\_reexpress:  
    medical\_clean\_data[f'Dummy\_{column}'] = medical\_clean\_data[column].map({'Yes': 1, 'No': 0 }).astype(np.int64)  
     
   categorical\_medical\_data = medical\_clean\_data[convert\_to\_category.keys()]  
   high\_cardinalities = categorical\_medical\_data.nunique() > 3 #(> 3-5 Levels)  
   high\_cardinalities = high\_cardinalities[high\_cardinalities == True]  
   high\_cardinalities = list(high\_cardinalities.index.values)  
   print('Categoricals with high cardinality to be removed:')  
   print(high\_cardinalities)  
   print('\n')  
     
   medical\_clean\_data = medical\_clean\_data.drop(columns=high\_cardinalities)  
     
   low\_cardinalities = [item for item in list(convert\_to\_category.keys()) if item not in high\_cardinalities]  
   print('Categoricals remaining:')  
   print(low\_cardinalities)  
   print('\n')  
     
   #Re-level ordinal/nominal categoricals  
   medical\_clean\_data['Dummy\_Complication\_risk'] = medical\_clean\_data['Complication\_risk'].map({'Low': 0, 'Medium': 1, 'High': 2}).astype(np.int64)  
   medical\_clean\_data['Dummy\_Initial\_admin'] = medical\_clean\_data['Initial\_admin'].map({'Emergency Admission': 0, 'Elective Admission': 1, 'Observation Admission': 2}).astype(np.int64)  
   medical\_clean\_data['Dummy\_Gender'] = medical\_clean\_data['Gender'].map({'Male': 0, 'Female' : 1, 'Nonbinary': 2}).astype(np.int64)  
     
   prepared\_medical\_data = medical\_clean\_data.drop(columns=medical\_clean\_data.select\_dtypes(include=category\_dtype))

2.  Describe the dependent variable and all independent variables using summary statistics that are required to answer the research question, including a screenshot of the summary statistics output for each of these variables.  
  


1. Generate univariate and bivariate visualizations of the distributions of the dependent and independent variables, including the dependent variable in your bivariate visualizations.  
     
   Code is available in the Jupyter Notebook included in submission.   
   File: `*D208 - nbm3-task1-linear-regression-modeling.ipynb*`  
     
   **Univariate Visualizations:**

Chart

Description automatically generated

**Bivariate Visualizations:**

A picture containing diagram

Description automatically generated

4.  Describe your data transformation goals that align with your research question and the steps used to transform the data to achieve the goals, including the annotated code.  
  
The transformation of the original dataset to the prepared dataset for the Linear Regression modeling was done in a way to help reduce variables that could potentially lead to over fitting. The dataset is prepared so that it can be used for Multiple Linear Regression as well as Logical Linear Regression.

5.  Provide the prepared data set as a CSV file.  
See submitted file: ‘*initial-medical-model-dataset.csv*’

# **Part IV: Model Comparison and Analysis**

## D.  Compare an initial and a reduced linear regression model by doing the following:

1.  Construct an initial multiple linear regression model from all independent variables that  
 were identified in part C2.  
  
We are going to be using *ALL* the numerical variables from the prepared dataset. This is going to be a longer equation. The maximum number of predictors that are related to the dependent (target).  
  
We will be using the Multiple Linear Regression Model, sample   
  
*Initial Model:*

y = 4014.0409798534274 + 9.9515\*Children + 224.3642\*Age + 0.0005\*Income + 63.2296\*VitD\_levels + 66.9367\*Doc\_visits + 6.2086\*Full\_meals\_eaten + 22.5048\*vitD\_supp + 35.1320\*Initial\_days + 0.4705\*TotalCharge + 146.1032\*Dummy\_ReAdmis + 9.6810\*Dummy\_Soft\_drink + 8665.8292\*Dummy\_HighBlood + 345.9470\*Dummy\_Stroke + 5.3773\*Dummy\_Overweight + 42.8769\*Dummy\_Arthritis + 68.3028\*Dummy\_Diabetes + 26.6751\*Dummy\_Hyperlipidemia + 2.9351\*Dummy\_BackPain + 56.5823\*Dummy\_Anxiety + 4.3332\*Dummy\_Allergic\_rhinitis + 28.1193\*Dummy\_Reflux\_esophagitis + 35.7533\*Dummy\_Asthma + 348.3569\*Dummy\_Complication\_risk + 445.6781\*Dummy\_Initial\_admin + 148.0374\*Dummy\_Gender

2.  Justify a statistically based feature selection procedure or a model evaluation metric to reduce the initial model in a way that aligns with the research question.  
  
Using the OLS (Ordinary Least Squares) model it was calculate that the R2 value is >90 with a value of 98. This indicates that the linearity between the medical datasets and the model is significant.  
  
From the OLS results provided, we can see that the coefficients of "Children", "Income", "Full\_meals\_eaten", "vitD\_supp", "Dummy\_Soft\_drink", "Dummy\_Overweight", "Dummy\_Arthritis", "Dummy\_Hyperlipidemia", "Dummy\_BackPain", "Dummy\_Anxiety", "Dummy\_Allergic\_rhinitis", "Dummy\_ReAdmis", "Dummy\_Diabetes", "Dummy\_Reflux\_esophagitis", and "Dummy\_Asthma" are not statistically significant at the 5% level (P>|t| > 0.05), which suggests that these variables may not be contributing significantly to the model.

This leaves us with a reduced set of explanatory variables which include: Age, VitD\_levels, Doc\_visits, Initial\_days, TotalCharge, Dummy\_HighBlood, Dummy\_Stroke, Dummy\_Complication\_risk, Dummy\_Initial\_admin and Dummy\_Gender

3.  Provide a reduced linear regression model that follows the feature selection or model evaluation process in part D2, including a screenshot of the output for each model.  
  
*Reduced Model:*

y = 3611.932336539983 + 224.3247\*Age + 65.3868\*VitD\_levels + 69.5676\*Doc\_visits + 34.4714\*Initial\_days + 0.4350\*TotalCharge + 8662.0148\*Dummy\_HighBlood + 346.2801\*Dummy\_Stroke + 340.2887\*Dummy\_Complication\_risk + 439.3655\*Dummy\_Initial\_admin + 148.7851\*Dummy\_GenderSummary Output of Both

Models:  
  
Initial Model Summary:  
Text

Description automatically generated  
  
Reduced Model Summary:  
Text

Description automatically generated

## E.  Analyze the data set using your reduced linear regression model by doing the following:

1.  Explain your data analysis process by comparing the initial multiple linear regression model and reduced linear regression model, including the following element:

•   a model evaluation metric

*Model Evaluation Metrics:*

|  |  |  |
| --- | --- | --- |
|  | Initial Model | Reduced Model |
| Dependent Variable | Additional\_charges | Additional\_charges |
| R-Squared | 0.987 | 0.987 |
| Adj. R-Squared | 0.987 | 0.987 |
| Prob (F-Statistic) | 3.097e+04 | 7.740e+04 |

When deal with the initial model which included all variables and the reduced model the value of the evaluation matrix that stuck out was the F-Statistic. The comparison is an F-Statistic comparison of 3.097e+04 (Initial) < 7.740e+04 (Reduced). Because the reduced modes F-Statistic more than double the initial model it’s clear that the model as a whole has a much stronger relationship to the dependent variable which is ‘Additional\_charges’. Normally, more than just the F-Statistic would be needed to determine which model is better but in this case each models values are identical other than the F-Statistic, which is why it’s the focus for which is better.

2.  Provide the output and all calculations of the analysis you performed, including the following elements for your reduced linear regression model:

•   a residual plot

•   the model’s residual standard error

Chart, line chart

Description automatically generated

3.  Provide an executable error-free copy of the code used to support the implementation of the linear regression models using a Python or R file.  
  
Please see submitted file: *`D208 - nbm3-task1-linear-regression-modeling.ipynb`*

# **Part V: Data Summary and Implications**

## F.  Summarize your findings and assumptions by doing the following:

1.  Discuss the results of your data analysis, including the following elements:

•   a regression equation for the reduced model  
  
y = 3611.932336539983 + 224.3247\*Age + 65.3868\*VitD\_levels + 69.5676\*Doc\_visits + 34.4714\*Initial\_days + 0.4350\*TotalCharge + 8662.0148\*Dummy\_HighBlood + 346.2801\*Dummy\_Stroke + 340.2887\*Dummy\_Complication\_risk + 439.3655\*Dummy\_Initial\_admin + 148.7851\*Dummy\_Gender

•   an interpretation of the coefficients of the reduced model

|  |  |
| --- | --- |
| **Coefficient** | **Interpretation of statistical Significance** |
| Age | Additional\_charges will increase by a factor of 224 units for every year Age increases. |
| VitD\_levels | Additional\_charges will decrease by a factor of 63 units for every VitD\_levels unit |
| Doc\_visits | Additional\_charges will increase by a factor of 224 units for every Doc\_visits unit. |
| Initial\_days | Additional\_charges will increase by a factor of 34 units for every Initial\_days unit. |
| TotalCharge | Additional\_charges will decrease by a factor of 0.4 units for every TotalCharge unit. |
| Dummy\_HighBlood | Additional\_charges will increase by a factor of 8662 units for every HighBlood unit. |
| Dummy\_Stroke | Additional\_charges will increase by a factor of 346 units for every Stroke unit. |
| Dummy\_Complication\_risk | Additional\_charges will increase by a factor of 340 units for every Complication\_Risk unit. |
| Dummy\_Initial\_admin | Additional\_charges will decrease by a factor of 439 units for every Initial\_admin unit. |
| Dummy\_Gender | Additional\_charges will decrease by a factor of 148 depending on ones Gender. |

•   the statistical and practical significance of the reduced model  
  
The reduced model was reduced to only the statistically significant values that are below 0.05 for P>|t|. Each of these variables have some kind of significant sway over the dependent Additional\_charges variable.   
  
When we started with the null-hypothesis of “H0 – There are no explanatory variables within the medical dataset that show any significant sway of additional charges (Additional\_charges) that a patient accrues during their hospitalization.”, which we were able to reject given that we found variables that do in fact have statistically significant p-values. To prevent over fitting and a more practical model we reduced from modelling all the values down to the significant ones.

•   the limitations of the data analysis  
  
In terms of building a great linear regression model the medical\_data provided is minimal. To get more accurate results it would be better to have long standing data throughout the hospitals history.

2.  Recommend a course of action based on your results.  
  
There are quite a few significant factors. From a hospital perspective quite a few of these factors cannot be controlled such as Age and how a Patient is initial admitted because we can’t control accidents.  
  
For the factors that can be compensated for is having doctors pay great attention earlier during initial days for high blood pressure, stroke risk and major complication risks. Finding and addressing these earlier would have effects on reducing additional charges.

# **Part VI: Demonstration**

## G.  Provide a Panopto video recording that includes the presenter and a vocalized demonstration of the functionality of the code used for the analysis of the programming environment, including the following elements: Video Link: [https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=d84239f9-e49c-41fd-b4ee-afcd01763592#](https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=d84239f9-e49c-41fd-b4ee-afcd01763592%23)

•   an identification of the version of the programming environment

Windows 11 + MacOS, DataSpell IDE powered by Python3, and Jupyter Notebook

•   a comparison of the initial multiple linear regression model you used and the reduced linear regression model you used in your analysis  
  
- Test all variables against Additional\_charges to see if any variables have any significant sway over it.  
- Reduced model down to only variables that have P>|t||0.005

•   an interpretation of the coefficients of the reduced model  
  
For the factors that can be compensated for is having doctors pay great attention earlier during initial days for high blood pressure, stroke risk and major complication risks. Finding and addressing these earlier would have effects on reducing additional charges.

Note: The audiovisual recording should feature you visibly presenting the material (i.e., not in voiceover or embedded video) and should simultaneously capture both you and your multimedia presentation.

Note: For instructions on how to access and use Panopto, use the "Panopto How-To Videos" web link provided below. To access Panopto's website, navigate to the web link titled "Panopto Access," and then choose to log in using the “WGU” option. If prompted, log in using your WGU student portal credentials, and then it will forward you to Panopto’s website.

To submit your recording, upload it to the Panopto drop box titled “Regression Modeling – NBM3 | D208.” Once the recording has been uploaded and processed in Panopto's system, retrieve the URL of the recording from Panopto and copy and paste it into the Links option. Upload the remaining task requirements using the Attachments option.

## H.  List the web sources used to acquire data or segments of third-party code to support the application. Ensure the web sources are reliable.

None used.

## I.  Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

Bevans, R. (2022, November 15). *Multiple linear regression: A quick guide (examples)*. Scribbr. Retrieved March 13, 2023, from https://www.scribbr.com/statistics/multiple-linear-regression/

Zach. (2021, January 21). The Four Assumptions of Linear Regression. Statology. Retrieved March 13, 2023, from https://www.statology.org/linear-regression-assumptions/

## J.  Demonstrate professional communication in the content and presentation of your submission.