Milestone 2: Exploratory Data Analysis and Regression

**Overview and Objectives**

The primary objective of this project will be to provide an in-depth analysis of data through the lens of regression analysis. Modern research looks at the interaction between a multitude of variables to provide conclusions based on experimental data. You will have the opportunity to explore a few realistic regression models and provide conclusions on them based on your own observations.

You will also observe a few measures of central tendency of the data. You will gain an understanding of the distribution of data through identifying the standard deviation and quartiles of your data set. You will then use these measures of central tendency to identify outliers in your data and determine whether they should be omitted in your regression.

You’ll also make inferences based on your regression data and determine which variables are significant in your model. This will help you identify the independent variables that adequately explain the variance in your model. The techniques you apply in this project will hopefully form the foundation of future research and exploration.

**Part One: Brain Size**

*Excel Tasks:* In the tab labelled brain size, please compute the following on your Excel sheet using the Head Size column.

* Five Number Summary
* Interquartile Range and Lower/Upper Limits for Outlier
* Identify whether any of the values are outliers

Next follow then next steps to create your scatterplot:

* Create a scatterplot that shows the relationship between head size and brain weight. Be sure to include your equation/r-squared value in your scatterplot.
* Use the Excel Regression tool to create a Residual vs. Fitted plot for your data and copy your plot on the original tab. \*don’t forget to adjust your x-axis

*Analysis of Results:* please respond in 2-4 complete sentences to each question.

1. What kind of relationship exists between head size and brain weight?
2. Are the outlier(s) of the data set reasonable? Should you omit them?
3. Do you think the other variables would be significant in predicting brain weight along with head size? Why?

**Part Two: Infection Risk in Hospitals**

*Excel Tasks:* Using the Infection Risk dataset determine the following measures of central tendency.

* Mean
* Median
* Mode
* Standard deviation

Once you’ve found these measures, find a regression model that predicts the infection risk of patients using the other data recorded:

* Determine which columns might influence the chance that a patient is infected while they are in the hospital.
* Run a multiple regression that creates a model that predicts the infection risk of a patient using the columns you indicated.

*Analysis of Results:* please respond in 2-4 complete sentences to each question.

1. What is the typical age of a participant of this study? What is the range of patient ages that are within three standard deviations of the mean?
2. What variables will you use from the data to predict infection risk?
3. Is your regression model a good predictor of infection risk? Are all of the variables you selected statistically significant?

**Part Three: Using Medical Expenses to Project Insurance Rates**

*Excel Tasks:* Using the Medical Expenses dataset, please compute the following in Excel.

* Calculate mean, standard deviation, and the z-score for each individual value.
* Use this data to determine whether the values are outliers.
* Use the Countif function to find how many outliers there are in your data.

Next create a regression model that will predict medical expenses by the other variables listed in each column.

*Analysis of Results:*

1. Are all the variables statistically significant in predicting the medical expenses of a patient?
2. What equation would you use to predict the medical expenses of a patient who is not part of this sample?
3. Use the equation from (2) to predict the medical expenses of someone who is 34, female, 32 BMI, 2 children, and a smoker.
4. Do you think this model would accurately predict the medical costs of a patient given the following information? If not, what additional predictors could the model include to improve the prediction? Could the data be adjusted to be more meaningful?