James Shea

Automotive Insurance Interest Prediction

11/1/2020

1. **Research Question**

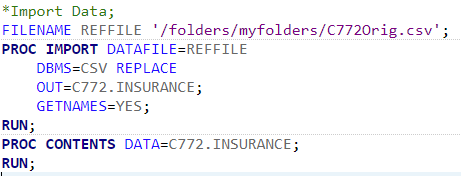
The purpose of this analysis is to determine if data collected on current health insurance customers is sufficient to predict customer interest in acquiring automotive insurance from the same insurance company. This analysis will utilize a dataset containing information recorded about a subset of current health insurance customers along with whether those customers were interested in automotive insurance. The analysis will attempt to determine which factors, if any, best predict interest in automotive insurance. A Generalized Logistic Model finds relationships between predictor variables and the response variable (Date, 2020). A predictive model can be created by using multiple logistic regression. It is hypothesized that health insurance customer data is sufficient to determine likelihood of interest in automotive insurance. If a relationship between the health insurance customer data and interest in automotive insurance is found, a model to predict future response given acquired data should be possible. (Statistical Consulting Group, 2020).

1. **Data Collection**

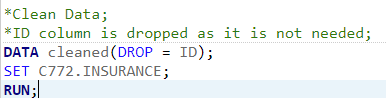
The data set used was acquired from the website Kaggle.com from Anmol Kumar who was provided the data set from a health insurance company and made the information publicly available (Kumar, 2020). The data set is available at https://www.kaggle.com/anmolkumar/health-insurance-cross-sell-prediction. This data set included a total of 12 variables and 381,109 observations. There are 3 continuous variables and 9 categorical variables. The variables in the data set are: ID, Gender, Age, Driving License, Region Code, Previously Insured, Region Code, Previously Insured, Vehicle Age, Vehicle Damage, Annual Premium, Policy Sales Channel, Vintage, and Response. This data set has several advantages. This data set has no missing data. This data set was collected from a single health insurance company on customers from across the United States of America. There are over 300,000 rows and it is necessary to have a sufficiently large quantity of rows for a Logistic Regression Model to fit well (Austin & Steyerberg, 2015). The disadvantages of using this data set include not having more variables that may better predict interest, not having data from multiple time points, and not having data from multiple countries. This disadvantages could be overcome in a future study as discussed at the end of the paper.

1. **Data Extraction and Preparation**

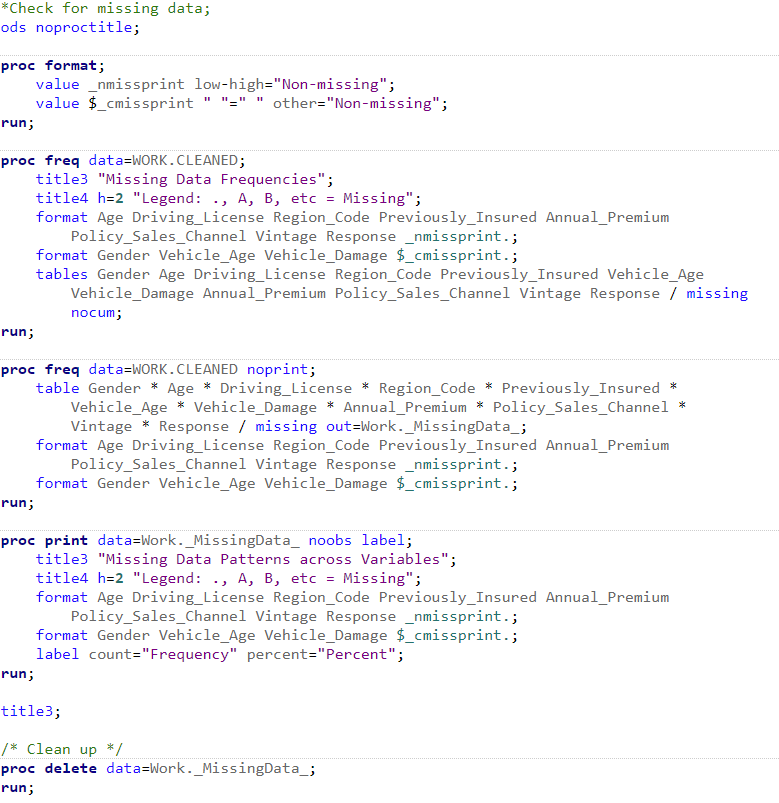
Once the data has been downloaded, the first step is to import it into SAS using the following code.



This takes our original excel file and imports it into our SAS library as a SAS data set. The next step is to review our data and determine what else needs to be done to clean our data. Immediately upon reviewing the data, we can see that the ID column is not needed as it is simply used as an identifier variable, which is not needed for any of the analyses that will be performed.



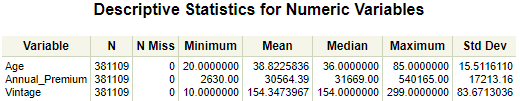
Next, we can check for any missing data using the following code.



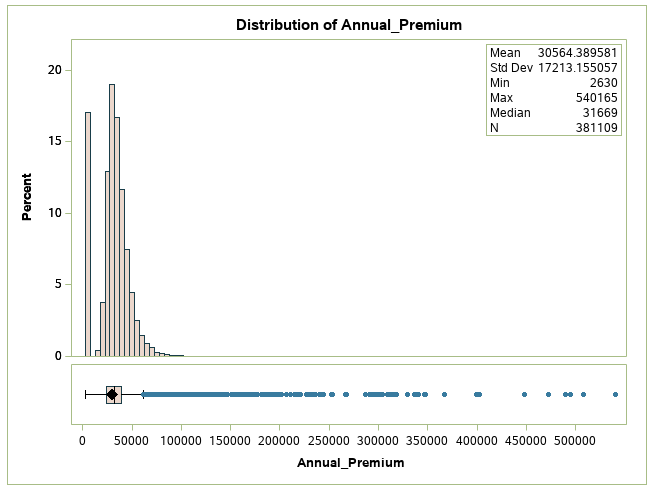
As we can see in the results below, there are no missing values. The lack of missing data is a benefit as missing data can reduce the model accuracy (Swalin, 2018).



Now we can check for any outliers in our variables as shown below.

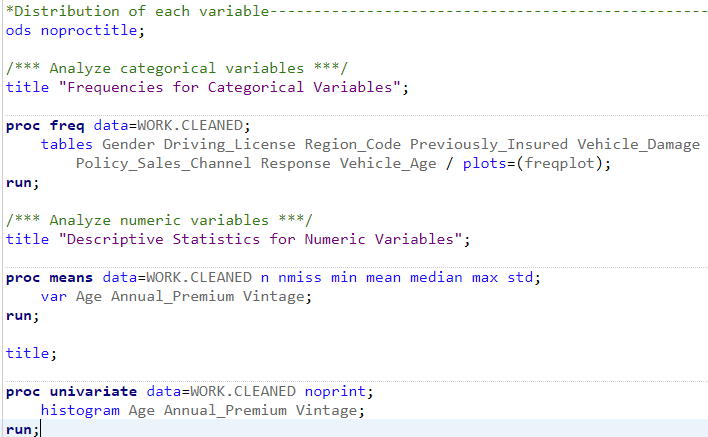


The Annual\_Premium column appears to be the most likely to have outliers, so we can take a closer look at this variable. As shown in the histogram and boxplot below, we can see that while Annual\_Premium is clearly skewed, there does not appear to be any significant outliers as datapoints exist throughout the range. Additionally, the distribution for this variable makes logical sense, as many factors can influence insurance premiums, resulting in a large range. The majority of customers being on the low end also makes sense as most people would tend to attempt to decrease their premiums as much as possible.



1. **Analysis**

To begin the analysis, the first step taken is to look at a univariate analysis of all variables. Frequency tables and histograms are generate look at each variable individually using the following code. This univariate analysis can be used to find interesting information about individual variables as well as help identify outliers or extreme observations.



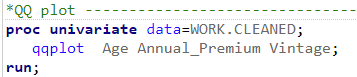
The histograms and frequency tables output are below. These are visual representations of individual variables for the observations in our data set.

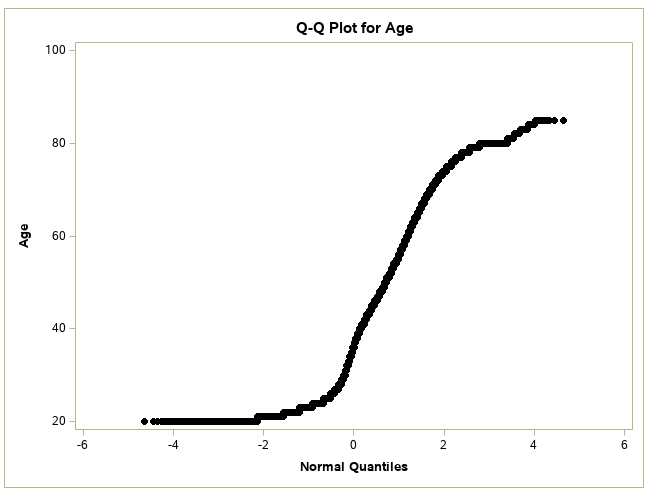


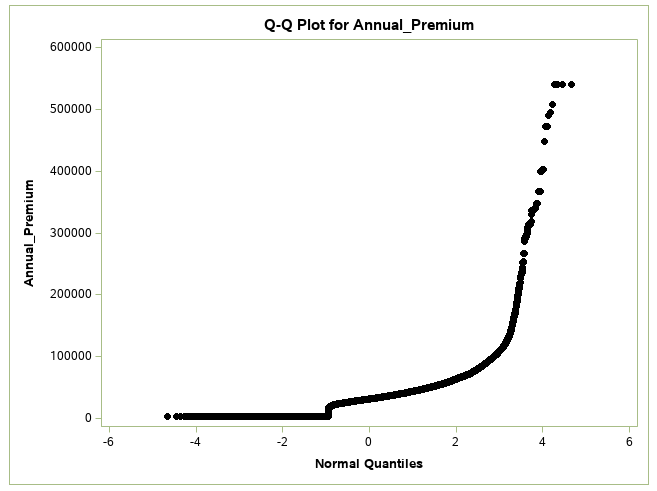
| **Variable** | **N** | **N Miss** | **Minimum** | **Mean** | **Median** | **Maximum** | **Std Dev** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Age Annual\_Premium Vintage | 381109 381109 381109 | 0 0 0 | 20.0000000 2630.00 10.0000000 | 38.8225836 30564.39 154.3473967 | 36.0000000 31669.00 154.0000000 | 85.0000000 540165.00 299.0000000 | 15.5116110 17213.16 83.6713036 |

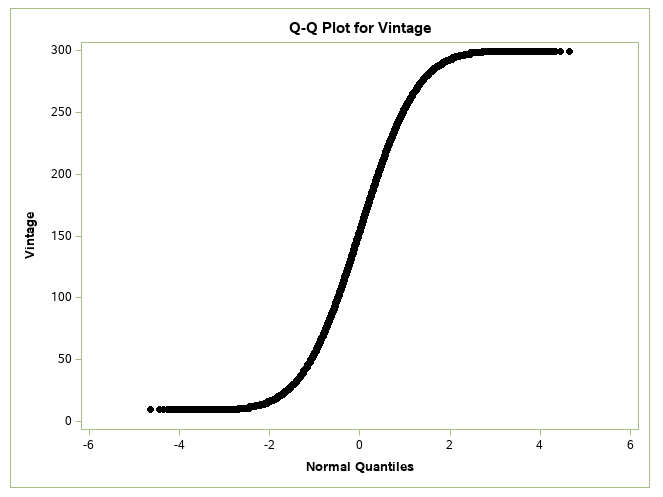
While the univariate analysis does not reveal any outliers or extreme observations, it does reveal some interesting information about our data set. Age appears to have many more observations in the 20 to 25 range than other ages. Most observations were not interested in automotive insurance. Most observations had a driving license.

Next, quantile-quantile plots are created to compared ordered variable values. This allows for the assessment of normality. To further assess normality, additional tests are run. The code to run this assessment and the results are below.

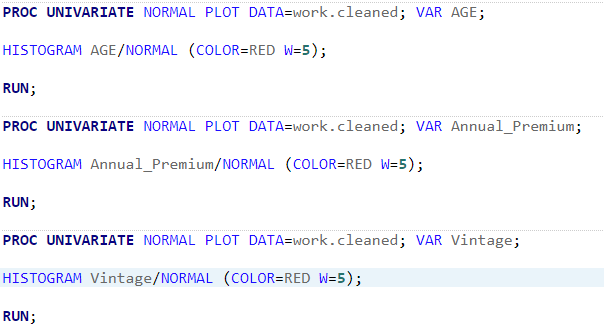




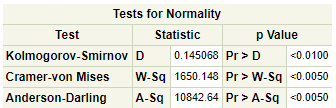




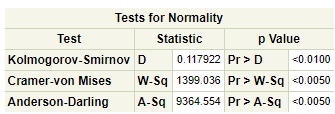
Further tests of normality are run using the following code.



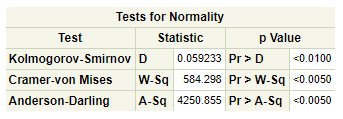
Kolmogorov-Smirnov, Cramer-von Mises, and Anderson-Darling tests of normality for age.



Kolmogorov-Smirnov, Cramer-von Mises, and Anderson-Darling tests of normality for Annual\_Premium.



Kolmogorov-Smirnov, Cramer-von Mises, and Anderson-Darling tests of normality for Vintage.



The results suggest that the data is not all normal. However, as cited above, normality is not a requirement for our analysis.

Next, we can generate stacked bar graphs to view the bivariate data for each of our predictor variables with the outcome variable using the following code. This will allow for visual representations of how the predictor variables each interact with the outcome variable. This can help identify if any variables have observations with extreme values or outliers. Separately from the univariate analysis, this shows interesting information about the interaction between the predictors and the outcome variable, which means it could reveal observations that outliers separate from the univariate analysis.

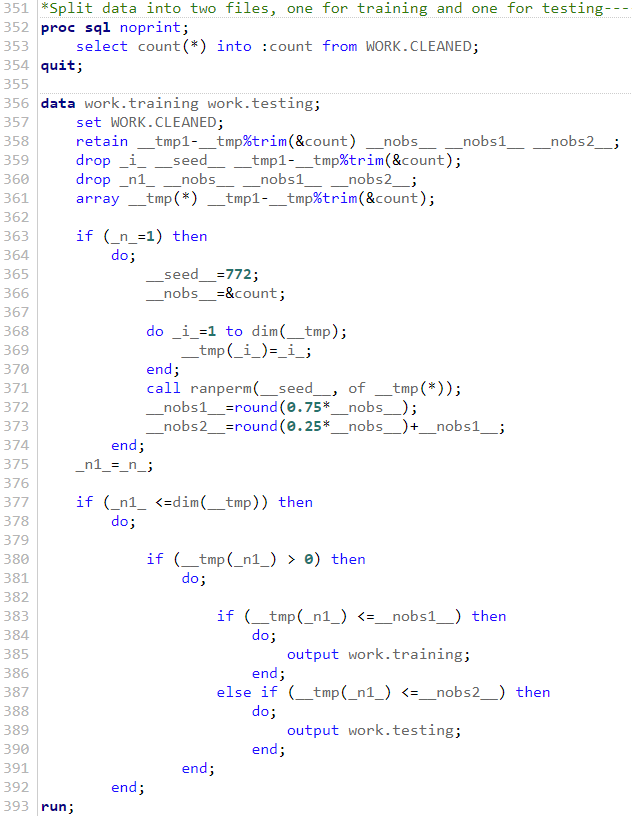


The bivariate stacked bar graphs generated are below.

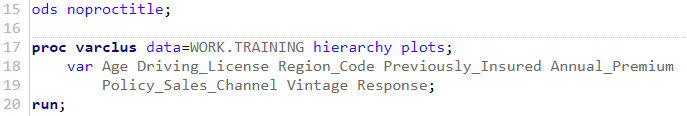


The bivariate analysis shows a lot of interesting information, but does not identify any variables that may have observations that are extreme or outliers. The bivariate analysis reveals interesting information about our data set. Males are proportionately more likely than females to be interested in automotive insurance. Further, customers who are both older and younger than average are less likely to be interested in automotive insurance.

Next, the data is split into two files, one for training and one for testing. This is done so that a model can be created using the observations in the training file which does not contain any observations that are in the testing file. Once the model is created, it can be tested against the testing observations. This helps prevent over fitting the data and allows for verification of the models efficacy. Of the total observations, 75% are used for the training data while the remaining 25% are used for the testing data. A seed is chosen (772) so that the same pseudo random split of data can be replicated. The observations are split into two files using the code below.



Next we can conduct a cluster analysis of the variables using the following code. The cluster analysis is conducted using the training data in order to be consistent with the logistic regression later in the analysis. A hierarchy plot will be created to visualize the plots.



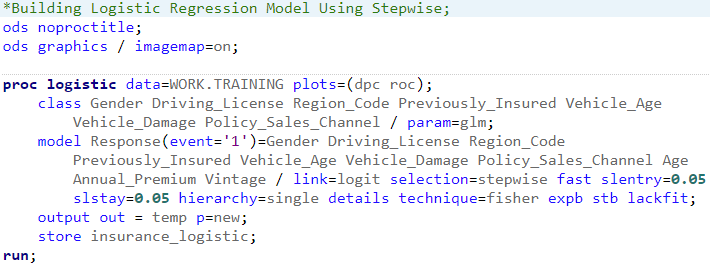
The cluster analysis suggests that there are three clusters. Three clusters are shown to explain approximately 49.5% of the variance. The results are shown below.

| **Number of Clusters** | **Total Variation Explained by Clusters** | **Proportion of Variation Explained by Clusters** | **Minimum Proportion Explained by a Cluster** | **Maximum Second Eigenvalue in a Cluster** | **Minimum R-squared for a Variable** | **Maximum 1-R\*\*2 Ratio for a Variable** |
| --- | --- | --- | --- | --- | --- | --- |
| **1** | 1.872456 | 0.2341 | 0.2341 | 1.129696 | 0.0000 |  |
| **2** | 2.963571 | 0.3704 | 0.3244 | 1.007643 | 0.0002 | 0.9998 |
| **3** | 3.958293 | 0.4948 | 0.4042 | 0.999945 | 0.0002 | 0.9998 |

Below is the hierarchy plot of the cluster analysis. This plot shows the clusters that resulted from the cluster analysis.



The next step in the analysis is to develop the logistic regression model. A logistic regression model can be helpful in estimating the target variable by determining the strength of predictor variables with respect to loan data (Martinson, 2020). A Generalized Logistic Model, which does not require a test for normality (University of Colorado Boulder, 2018), will be run to identify relationships in the data. For this process, the training data step is used in a stepwise selection with fast backward elimination for the logistic regression model. The code used to run this analysis is below.



The results of the analysis are shown below.

| **Summary of Stepwise Selection** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Step** | **Effect** | | **DF** | **Number In** | **Score Chi-Square** | **Wald Chi-Square** | **Pr > ChiSq** |
| **Entered** | **Removed** |
| **1** | **Vehicle\_Damage** |  | 1 | 1 | 35970.1570 |  | <.0001 |
| **2** | **Policy\_Sales\_Channel** |  | 151 | 2 | 5218.3190 |  | <.0001 |
| **3** | **Previously\_Insured** |  | 1 | 3 | 3480.2400 |  | <.0001 |
| **4** | **Age** |  | 1 | 4 | 1469.3584 |  | <.0001 |
| **5** | **Region\_Code** |  | 52 | 5 | 915.0917 |  | <.0001 |
| **6** |  | **Region\_Code** | 52 | 4 |  | 0.0000 | 1.0000 |

Four of our variables are included in the final model created using stepwise logistic regression on the training data set. They include vehicle damage, policy sales channel, previously insured, and age. The last step removed region code and found nothing else significant enough to enter or remove from the model. Below is the ROC curve for the model.



The ROC curve for the model shows an area under the curve of .8454. Below are the ROC curves for the models tests at each step of the stepwise logistic regression analysis.



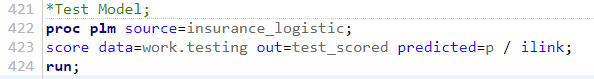
As one would expect, each step of the stepwise logistic regression analysis improves the area under the curve until the final model is created. While the initial steps improve the most, this is to be expected as more variance becomes explained at each step before including more variables. Below are the tables for association of predicted probabilities and observed responses, model convergence status, and type 3 analysis of effects.

| **Association of Predicted Probabilities and Observed Responses** | | | |
| --- | --- | --- | --- |
| **Percent Concordant** | 84.4 | **Somers' D** | 0.691 |
| **Percent Discordant** | 15.3 | **Gamma** | 0.692 |
| **Percent Tied** | 0.2 | **Tau-a** | 0.149 |
| **Pairs** | 8804358655 | **c** | 0.845 |

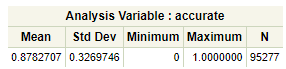
| **Model Convergence Status** | | | |
| --- | --- | --- | --- |
| Convergence criterion (GCONV=1E-8) satisfied. | | | |
| **Type 3 Analysis of Effects** | | | | |
| **Effect** | **DF** | **Wald Chi-Square** | **Pr > ChiSq** | |
| **Previously\_Insured** | 1 | 1730.6491 | <.0001 | |
| **Vehicle\_Damage** | 1 | 2498.7755 | <.0001 | |
| **Policy\_Sales\_Channel** | 151 | 5540.1496 | <.0001 | |
| **Age** | 1 | 1451.0757 | <.0001 | |

The model was concordant with 84.4% of observations in the training data and model convergence was satisfied. In the second table, we can see the effects of each variable included in the final model. These are the unique contributions to the model of each variable for predicting interest in automotive insurance for current health insurance customers.

With the model created using the training data, the model can now be applied to the testing data. Below is the code used to apply the model to the training data and to predict whether each observation would show interest in automotive insurance. A new row is created which records the likelihood the model predicts for each observation being interested. As the model was created with a separate data set (our training set) we could now observe problems such as overfit and see the accuracy of the model to new data.



With the likelihood of each observation showing interest in automotive insurance as predicted by the model, we can now grade the model for accuracy. This was done by assigning a 1 to all observations where the model predicted 50% or greater chance for interest and a 0 to all observations with a lower chance.



The mean shown above is simply the number of observations that were predicted correctly divided by the total number of observations in the test data set. The model was actually more accurate in predicting interest in automotive insurance of current health insurance customers in the test data than in the training data, accurately predicting interest 87.8% of the time, compared to only 84.4% in the training data set. These results suggest that the model is generalized and able to predict interest in automotive insurance using the predictor variables.

1. **Data Summary and Implications**

The analysis of current health insurance customer data has shown that previously collected data on those customers can predict interest in automotive insurance from the same company. The results suggest that the predictor variables Previously Insured, Vehicle Damage, Policy Sales Channel, and Age can be used to predict interest in automotive insurance. The analysis is limited as all data is from a single insurance company and thus, while the regression equation is generalized to this insurance company’s customers, may not generalize to other insurance companies. Further, the results may change in the future with national policy or cultural changes.

Future analyses could expand upon this analysis in several ways. First, additional predictor variables could be used that more accurately predict interest. Further, it may be beneficial to get customer data from multiple insurance companies. It may also be beneficial to see if interest changes over the course of a year, as the data on interest does not show how interest may be different during different periods of the year, such as season differences. Finally, getting customer data from multiple companies may allow the logistic regression equation to generalize to different populations.

1. **Sources**

Analyzing Non-Normal Data with Generalized Linear Models (GLMs). (2018). Retrieved October 18, 2020, from https://www.colorado.edu/lab/lisa/services/short-courses/analyzing-non-normal-data-generalized-linear-models-glms

Austin, P., & Steyerberg, E. (2015, January 22). The number of subjects per variable required in linear regression analyses. Retrieved October 13, 2020, from https://www.sciencedirect.com/science/article/pii/S0895435615000141

Coding Systems for Categorical Variables in Regression Analysis. (n.d.). Retrieved October 22, 2020, from https://stats.idre.ucla.edu/spss/faq/coding-systems-for-categorical-variables-in-regression-analysis-2/

Date, S. (2020, May 10). Generalized Linear Models. Retrieved October 13, 2020, from https://towardsdatascience.com/generalized-linear-models-9ec4dfe3dc3f

Generalized linear model. (2020, September 26). Retrieved October 13, 2020, from https://en.wikipedia.org/wiki/Generalized\_linear\_model

Kumar, A. (2020, August). Health Insurance Cross Sell Prediction. Retrieved October 29, 2020, from https://www.kaggle.com/anmolkumar/health-insurance-cross-sell-prediction

Swalin, A. (2018, March 19). How to Handle Missing Data. Retrieved October 13, 2020, from https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4

What is Linear Regression? (n.d.). Retrieved October 13, 2020, from https://www.statisticssolutions.com/what-is-linear-regression/