**Final Group Project Report**

**Conestoga College**

Business Analytics

INFO 8136: Data Acquisition, Analysis, and Visualization

Section#1

**Submitted To:**

Professor Gbenga Adeleye

**Group 4:**

Siddhesh Otari

Sanjana Upender

Harshrajsinh Chavda

Vats Sanghvi

**Submission Date:**

April 3, 2024

Table of Contents

[Dataset Summary 3](#_Toc163062931)

[Purpose of Statistical Analysis 3](#_Toc163062932)

[Descriptive Data Mining 4](#_Toc163062933)

[Statistical Inference 5](#_Toc163062934)

[Linear Regression 8](#_Toc163062935)

[Data Visualization 10](#_Toc163062936)

[Predictive Analysis using Python 12](#_Toc163062937)

[Conclusion 17](#_Toc163062938)

[References 18](#_Toc163062939)

## Dataset Summary

Our dataset is about health insurance information of individuals in various parts of USA *(Choi. 2018)*. The data is represented by individuals’ personal information and insurance charges for the. The description of the variables taken into consideration are as follows:

* Age: It is an independent variable and represents the age of an individual in a number of years.
* Sex: Sex is also an independent variable and represents the gender of an individual (male/female). We coded categorical variables for data analysis; therefore, male = 0, female = 1.
* Body Mass Index (BMI): BMI is to some extent. a function of age and sex but not entirely. Therefore, it is an independent variable as well, and it is represented in numeric values.
* Children: This variable collects data on how many children an individual has. It’s again a numeric value and an independent variable as it does not get influenced by age or any other factor.
* Smoker: It’s a categorial and independent variable, too. Hence, coded as Yes = 0 and No = 1. It shows the smoking preference of an individual.
* Region: As the dataset includes individuals residing in the USA, it’s divided into four categories: northeast, northwest, southeast, and southwest. The codes for all are 0, 1, 2, and 3, respectively. It’s an independent variable, too.
* Insurance Charges: This is the only dependent variable and is represented is USD values. At the first glance, these values seem to be function of all other independent variables listed above.

## Purpose of Statistical Analysis

By conducting a statistical analysis of this dataset, we will be able to assess the impact of various variables on insurance charges. This may help insurance companies, corporations, individuals, and governments as well to form various healthcare policies.

Here are some of the objectives we are looking forward to achieving through analysis:

* Risk assessment: Analysis of this dataset will help us create and analyze risk profiles of individuals, which are based on age, gender, BMI, and smoking habits. This is valuable data for insurance companies as it helps them navigate through insurance premium rates as well as coverage plans, options, and policy riders.
* Factors affecting insurance: Statistical analysis will help us assess which variables significantly impact insurance charges and which do not. Identifying major factors is important to insurers as they can decide on policy guidelines.
* Predictive analysis: By incorporating statistical techniques such as regression analysis, we may be able to create predictive models which will provide insights into future utilization of healthcare resources, claims to be filed, and consumer behavior, too. Development of such models helps insurers prepare for the future including budgeting, and also governments in resource management.

We have utilized following statistical methods to achieve our targets objectives:

* Descriptive Data Mining
* Statistical Inference
* Linear Regression
* Data Visualization
* Predictive Analysis

## Descriptive Data Mining

To uncover underlying patterns and characteristics, descriptive data mining entails analyzing and summarizing datasets. In this assignment, we examine a dataset that includes multiple important attributes: age, sex, region, number of children, smoking status, BMI (body mass index), and insurance costs.

Our goal in using descriptive statistics is to draw conclusions that are relevant to this data. We aim to identify patterns and connections between health-related and demographic variables, including age, BMI, smoking status, and insurance costs, through the analysis. These results can provide useful data for resource allocation plans, decision-making procedures, and possibly even predictive modeling related to insurance costs. Understanding the structure and subtleties of the dataset through descriptive data mining is an essential first step towards conducting a more thorough analysis and making well-informed decisions *(Cengage, 2021)*.

**Summary Report**

A screenshot of a computer

Description automatically generated

**Interpretation**

Descriptive statistics shed light on the characteristics of the dataset. People are roughly 39 years old on average, and their median BMI is 30.4. The majority of people do not smoke, and insurance costs range widely, with a cap of more than $63,000. These figures provide an overview of the demographic and health-related variables included in the dataset.

The information supplied shows the average insurance costs for various age groups. Notably, the charges peak between the ages of 60 and 64 and generally show an increasing trend with age. This implies that older people typically have higher insurance costs, maybe as a result of the higher health risks that come with aging. The wide range of prices reflects the complexity of insurance pricing and the variety of factors that affect it, such as coverage requirements, lifestyle decisions, and medical history. The significance of age as a factor in insurance pricing is highlighted by this data, which also reflects the different risk levels associated with different life stages.

## Statistical Inference

Here, we have performed a t-test on two variables, i.e., Body Mass Index and Health Insurance Charges, using the method called hypotheses testing. In this method, there are null hypotheses (H0) and alternative hypotheses (H1). A null hypothesis is an assumption, and an alternative hypothesis is the opposite of null hypothesis. The p-value plays a crucial role in hypothesis testing, particularly in determining whether to reject or fail to reject the null hypothesis *(Cengage, 2021)*.

For our t-test, the following are H0 and H1.

**H0:** There is no significant difference in insurance charges among individuals with different BMI.

**H1:** There is a significant difference.

**Process**

In hypothesis testing, p-value is compared to a predetermined significance level known as alpha, which is normally set to 0.05 (5%). If the p-value is less than or equal to the significance level, you reject the null hypothesis or accept it.

For our t-test, we got following statistics:

|  |  |  |
| --- | --- | --- |
| Data Points | BMI | Charges |
| Mean | 30.66339686 | 13270.42227 |
| Variance | 37.18788361 | 146652372.2 |
| Observations | 1338 | 1338 |
| Pooled Variance | 73326204.67 |  |
| Hypothesized Mean Difference | 0 |  |
| Df | 2674 |  |
| t Stat | -39.99111667 |  |
| P(T<=t) one-tail | 7.6918E-275 |  |
| t Critical one-tail | 1.645423672 |  |
| P(T<=t) two-tail | 1.5384E-274 |  |
| t Critical two-tail | 1.960851541 |  |

**P-values:** The p-values for both the one-tailed (7.6918E-275) and two-tailed (1.5384E-274) tests are extremely small (close to zero), indicating strong evidence against the null hypothesis.

**t Statistic:** The calculated t-statistic (-39.99) is significantly different from zero, indicating that the difference between the means of BMI and insurance charges is not due to random chance.

**t Critical Values:** The calculated t-statistic is much larger (in absolute terms) than the critical t-values for both one-tailed (1.645423672) and two-tailed (1.96) tests. This indicates that the observed difference is unlikely to occur under the null hypothesis.

**Inferences**

Looking at the t-test performed on BMI and Insurance Charges, the p-value suggests that the difference between means of BMI and insurance charges does not look like resulting from any random variable or chance alone, and there must be some relation between these two variables. Hence, we can reject the null hypothesis and conclude that there is a significant difference in Insurance Charges among individuals with different BMI values. The t-test suggests that individuals with different BMIs vary significantly in their insurance charges.

**Practical Implications**

Since there is a strong correlation between BMI and insurance costs, BMI may be taken into account by legislators, insurance companies, and healthcare professionals when setting insurance rates, developing health programs, or distributing funds. Following are the implications of the inferences and findings.

Health Risks:

BMI is frequently used to measure body weight in relation to height and is linked to a number of health hazards. increased BMI levels may indicate an increased susceptibility to obesity-related ailments including diabetes, hypertension, and cardiovascular disorders. Higher BMI levels may result in higher insurance rates from insurance companies because of the higher risk of health problems and related medical expenses.

Healthcare Utilization:

Higher BMI individuals could need medical services, such as prescription drugs, hospital stays, diagnostic testing, and doctor visits, more frequently. Because more people are using healthcare, insurance companies may charge more to cover the expected expenses of treating those with higher BMI levels.

Actuarial Risk Assessment:

To determine the chances and costs of offering coverage to various categories of people, insurance firms employ actuarial risk assessment. In actuarial risk assessment models, BMI could be one of the variables considered when forecasting healthcare expenses. Individuals with higher BMI values may face higher insurance premiums since higher BMI values are linked to higher expected healthcare expenses.

Policy Design:

Health-promoting and health-deterring habits may be rewarded in insurance policy. Higher premiums for those with higher BMI values could be a component of policy designs meant to encourage better lives and lower medical expenses connected to diseases linked to obesity.

The correlation between BMI and insurance costs is complex and varies depending on a number of variables including health risks, healthcare use, risk assessment, and policy design. Because there's a greater chance of health problems and the resulting medical expenses, higher BMI readings are frequently linked to higher insurance rates.

## Linear Regression

This is the Data that represents the medical cost, which is supported by age, sex, BMI, children, smoker, and region. In this table, the independent variable is age, sex, BMI, children, smoker, and region, whereas the dependent variable is the charges data.

**Regression Model**

We have made this scatter chart before the regression model to detect whether the BMI and age are related or not, if the both are related, we will have to eliminate one of the variables from the regression model *(Cengage, 2021)*.

As we can see that the scattergram clearly indicates that the trendline we achieved is a horizontal trendline which indicates that there is no correlation between the BMI and age due to which they both can be used as an independent variable in our Regression Model.



In the regression model, we were able to predict the change in medical cost associated with the change in per unit of Age, Sex, BMI, Children, and Smokers. In this data, we eliminated the region for data analysis as region data was not relevant to the analysis that was carried out.

**Intercept**

Explanation: The intercept value indicates the expected value of the dependent variable (Y). That is, the medical cost is $12181.10 when all independent variables are set to zero.

**Variable 1 (Age)**

Explanation: The value in the Co-efficient column for variable 1 indicates that for every 1-year increase in age, the medical cost would increase by approximately $257.73, having the other variables constant. Getting the p-value= 2.59E-.

**Variable 2 (Sex)**

Explanation: The value for the co-efficient of Sex is $128.63, which indicates that the medical cost for females is $128 more than for males. As we can see the p-value is high which indicates that gender may not be statistically significant in predicting medical cost and the standard error is 333.36 due to which we can not definitely depend on this data achieved.

**Variable 3 (BMI)**

Explanation: The value for the co-efficient of BMI is $322.36, which indicates that each unit increase in BMI increases the medical cost by $322.36, having other variables constant. Here, the p-value is 1.95E-30, which is remarkably close to zero, which indicates that BMI is highly statistically significant in predicting medical cost.

**Variable 4 (Children)**

Explanation: The value of the co-efficient for Children is $474.41, which predicts that each additional child increases the medical cost by $474.41, having the other variables constant. The p-value = 0.0005 indicates that the number of children is statistically significant in predicting medical costs.

**Variable 5 (Smoker)**

Explanation: The value of the co-efficient for Smokers is $23823.39, which predicts that smoking increases medical costs by $23823.39 as compared to non-smokers. The p-value = 0 indicates that smoking is a significant predictor of the medical cost.

**Multiple R**

It represents the strength and direction of the linear relationship between the dependent variable and the independent variable.

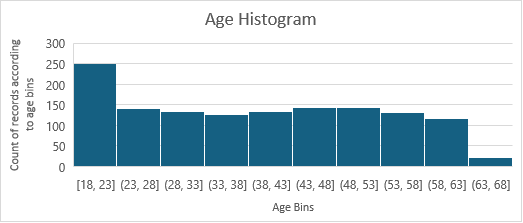
**R Square**

It indicates the proportion of the variance in the dependent variable that is justified by the independent variable in the regression model. A value closer to 1 indicates that the model is fit for the data.

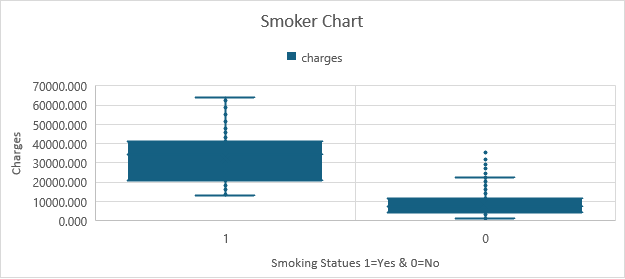
**Standard Error**

It measures the average distance between the observed value and the predicted values from the regression model, which has lower values and indicates that the model is a better fit for the data.

## Data Visualization



* This is a Histogram Chart for the age-related data for the medical cost. Due to the Histogram Chart, we can see that the highest data for the medical cost we have is for the age between 18 to 23 years. Whereas the lowest data we have for the medical cost is between 63 to 68 years.
* We can adjust the bin size by clicking on the x-axis bar and adjust the bin size in the setting.
* We can now the exact count of records of the age bins by moving the cursor to the age bars in the excel sheet.
* This is a scattergram for the BMI data which we can visualize that as we have less data for the higher ages, but we can clearly visualize that increase in BMI tends to increase in medical cost.
* We can visualize that the Linear Trendline indicates that the medical cost increase as the BMI increases.
* The equation y = 393.87x + 1192.9 represents a linear regression model where y is the predicted medical cost, and x is the BMI value. In this equation:
* The coefficient 393.87 represents the estimated increase in medical cost for each unit increase in BMI. The constant term 1192.9 represents the estimated medical cost when BMI is 0.
* The R² value of 0.0393 indicates the proportion of the variance in medical costs that is explained by the linear relationship with BMI. In other words, approximately 3.93% of the variability in medical costs can be attributed to changes in BMI according to this model. This suggests that BMI alone may not be a strong predictor of medical costs, as the model explains only a small portion of the variability *(Cengage, 2021)*.



* For this chart we can see that the mean of the people who smoke is not close to the mean average of people who not smoke.
* This Box and whisker chart indicates that the smoking status of an individual make a huge change in the medical cost on which can use this data for the regression modelling.

## Predictive Analysis using Python

**Data Preparation**

Loading the Dataset: The dataset containing information related to insurance charges was loaded. Attributes such as age, sex, BMI (Body Mass Index), number of children, smoking status, and insurance charges were included.

Understanding the Dataset: A careful examination of the dataset was conducted to understand the nature of each attribute and its potential relevance to predicting insurance charges by using scatter plot, and insights from the earlier method stated.

Selecting Input Variables: After assessing the dataset, input variables (features) were selected based on their potential influence on insurance charges:

* Age: Age is often a significant factor in determining insurance charges, as older individuals may have higher medical needs.
* Sex: Gender may influence insurance charges due to differences in healthcare utilization and risk factors.
* BMI: BMI is an indicator of body weight relative to height and is associated with health risks, potentially impacting insurance costs.
* Number of Children: The number of children may affect insurance charges, reflecting additional dependents and associated healthcare expenses.
* Smoking Status: Smoking is a well-known risk factor for various health conditions, leading to higher insurance premiums for smokers.

The dataset was split into input variables (X) and the target variable (y, insurance charges). Subsequently, the Random Forest Regression model was trained using the entire dataset.

**Partitioning Method**

To ensure the robustness of the predictive model and guard against overfitting, the dataset was partitioned into training, validation, and testing subsets. The partitioning was performed using a combination of static holdout and random selection methods.

Training and Testing Sets:

Most of the dataset was allocated to the training and testing sets using the static holdout method.

Specifically, the train\_test\_split function from scikit-learn was employed to split the dataset into training and testing subsets. A test size of 20% was chosen, ensuring that approximately 80% of the data was allocated to the training set and the remaining 20% to the testing set.

This approach allowed for training the model on a large portion of the data while retaining a separate subset for evaluating the model's performance on unseen data.

**Validation Set**

Additionally, to further assess the model's performance and fine-tune hyperparameters, a validation set was randomly selected from the training data.

Specifically, 30 records were randomly sampled from the training set to form the validation set. The validation set served as an intermediate evaluation step during model development, enabling iterative adjustments and fine-tuning to enhance the model's performance.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| age | sex | bmi | children | smoker | actual charges |
| 18 | 0 | 15.96 | 0 | 0 | 1694.80 |
| 21 | 1 | 16.815 | 1 | 0 | 3167.46 |
| 38 | 0 | 16.815 | 2 | 0 | 6640.54 |
| 26 | 1 | 17.195 | 2 | 1 | 14455.64 |
| 18 | 0 | 17.29 | 2 | 1 | 12829.46 |
| 28 | 1 | 17.29 | 0 | 0 | 3732.63 |
| 34 | 0 | 22.42 | 2 | 0 | 27375.90 |
| 29 | 0 | 22.515 | 3 | 0 | 5209.58 |
| 26 | 1 | 22.61 | 0 | 0 | 3176.82 |
| 48 | 1 | 22.8 | 0 | 0 | 8269.04 |
| 53 | 1 | 22.88 | 1 | 1 | 23244.79 |
| 29 | 0 | 22.895 | 0 | 1 | 16138.76 |
| 54 | 1 | 23 | 3 | 0 | 12094.48 |
| 57 | 1 | 23.18 | 0 | 0 | 11830.61 |
| 39 | 1 | 23.275 | 3 | 0 | 7986.48 |
| 45 | 0 | 23.56 | 2 | 0 | 8603.82 |
| 31 | 1 | 23.6 | 2 | 0 | 4931.65 |
| 47 | 1 | 23.6 | 1 | 0 | 8539.67 |
| 23 | 0 | 23.845 | 0 | 0 | 2395.17 |
| 28 | 1 | 23.845 | 2 | 0 | 4719.74 |
| 49 | 1 | 23.845 | 3 | 1 | 24106.91 |
| 44 | 1 | 23.98 | 2 | 0 | 8211.10 |
| 57 | 1 | 23.98 | 1 | 0 | 22192.44 |
| 25 | 0 | 24.13 | 0 | 1 | 15817.99 |
| 30 | 0 | 24.13 | 1 | 0 | 4032.24 |
| 35 | 0 | 24.13 | 1 | 0 | 5125.22 |
| 44 | 1 | 25 | 1 | 0 | 7623.52 |
| 62 | 1 | 25 | 0 | 0 | 13451.12 |
| 58 | 0 | 25.175 | 0 | 0 | 11931.13 |
| 58 | 1 | 25.2 | 0 | 0 | 11837.16 |

**Model Construction Random Forest Regression**

Model Comparison: Decision Tree vs. Random Forest Regression

A Decision Tree Regression model was initially considered for the predictive modeling task. However, it was observed that the predicted values overall had a huge difference. Random Forest Regression was ultimately chosen as the preferred model structure to address the limitations of Decision Trees. Random Forest Regression extends the concept of Decision Trees by constructing an ensemble of trees and aggregating their predictions.

**Code**

A screenshot of a computer program

Description automatically generated

**Output**

A white paper with numbers

Description automatically generated

**Performance measure: Evaluating the Estimation of Continuous Outcomes**

In assessing the performance of the Random Forest Regression model, the mean absolute error (MAE) was calculated for two different values of n\_estimators: 200 and 100.

Excel Table showing the ForecastError Absolute value and the mean for the validation dataset.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| n estimator = 100 | | |  | n estimator = 200 | | |
| Forecast | Forecast Error | Absolute |  | Forecast | Forecast Error | Absolute |
| 1956.299 | -261.5026 | 261.5026 |  | 1789.26 | -94.46 | 94.4636 |
| 2915.786 | 251.66985 | 251.6699 |  | 2943.005 | 224.45 | 224.4513 |
| 7282.874 | -642.32915 | 642.3292 |  | 7056.271 | -415.73 | 415.7259 |
| 17154.131 | -2698.48695 | 2698.487 |  | 16400.17 | -1944.53 | 1944.528 |
| 14185.365 | -1355.9099 | 1355.91 |  | 13968.78 | -1139.32 | 1139.322 |
| 5082.0182 | -1349.3931 | 1349.393 |  | 5412.105 | -1679.48 | 1679.48 |
| 19315.42 | 8060.486733 | 8060.487 |  | 20197.17 | 7178.74 | 7178.737 |
| 5887.32 | -677.7398843 | 677.7399 |  | 5867.162 | -657.58 | 657.5829 |
| 3128.27 | 48.5500985 | 48.5501 |  | 3243.708 | -66.89 | 66.89253 |
| 8416.41 | -147.368814 | 147.3688 |  | 8397.509 | -128.47 | 128.4654 |
| 23944.11 | -699.3211375 | 699.3211 |  | 23617.39 | -372.60 | 372.6015 |
| 16414.83 | -276.072024 | 276.072 |  | 16339.62 | -200.86 | 200.8619 |
| 12304.00 | -209.5244468 | 209.5244 |  | 12357.16 | -262.68 | 262.6788 |
| 11894.80 | -64.1881781 | 64.18818 |  | 11846.52 | -15.92 | 15.91743 |
| 7864.38 | 122.099372 | 122.0994 |  | 7920.613 | 65.86 | 65.86225 |
| 8605.46 | -1.633092 | 1.633092 |  | 8613.159 | -9.34 | 9.335789 |
| 5194.62 | -262.9680485 | 262.968 |  | 5340.928 | -409.28 | 409.281 |
| 10071.40 | -1531.731542 | 1531.732 |  | 10125.82 | -1586.15 | 1586.145 |
| 3224.08 | -828.9117681 | 828.9118 |  | 3182.485 | -787.31 | 787.3138 |
| 5618.67 | -898.9372743 | 898.9373 |  | 6168.322 | -1448.59 | 1448.586 |
| 23955.34 | 151.575543 | 151.5755 |  | 24040.83 | 66.08 | 66.08245 |
| 8594.15 | -383.0504097 | 383.0504 |  | 8528.969 | -317.87 | 317.8689 |
| 18398.49 | 3793.949184 | 3793.949 |  | 18336.48 | 3855.96 | 3855.962 |
| 17074.54 | -1256.556781 | 1256.557 |  | 17408.75 | -1590.77 | 1590.768 |
| 5807.67 | -1775.428182 | 1775.428 |  | 5524.648 | -1492.41 | 1492.408 |
| 5374.26 | -249.0411226 | 249.0411 |  | 5906.677 | -781.46 | 781.4611 |
| 7686.98 | -63.460445 | 63.46044 |  | 7654.18 | -30.66 | 30.6621 |
| 15863.88 | -2412.754356 | 2412.754 |  | 16009.54 | -2558.42 | 2558.416 |
| 11981.98 | -50.8576966 | 50.8577 |  | 12007.25 | -76.13 | 76.12733 |
| 12176.18 | -339.0233751 | 339.0234 |  | 12224.17 | -387.01 | 387.0128 |
|  |  |  |  |  |  |  |
|  | Mean Absolute Error | 1028.817 |  | Mean Absolute Error | | 994.8333 |

**n\_estimators = 200**

The calculated MAE for n\_estimators = 200 was found to be 994. This indicates that, on average, the model's predictions deviated from the actual insurance charges by approximately $994.

**n\_estimators = 100**

Similarly, for n\_estimators = 100, the calculated MAE was slightly higher, at 1028. This suggests a slightly less accurate prediction performance compared to the model with n\_estimators = 200.

**Insights**

The MAE provides a measure of the average magnitude of errors in the model's predictions. Lower MAE values indicate better predictive performance.

The observed MAE values for both n\_estimators configurations indicate that the model is making reasonably accurate predictions. However, there is room for improvement, as indicated by the relatively high error values.

**Room for Improvement**

While the Random Forest Regression model demonstrates promising predictive performance, there is still potential for enhancement.

Further experimentation with hyperparameters such as n\_estimators, max\_depth, and min\_samples\_split could lead to improved model performance.

Additionally, feature engineering techniques, such as incorporating interaction terms or transforming variables, may help capture additional information and improve prediction accuracy.

Model evaluation on a larger and more diverse dataset could provide valuable insights and potentially lead to better generalization.

## Conclusion

* Using statistical inference, we got to know that BMI significantly affects insurance charges. Hence, BMI is one of the factors that insurers must analyze while calculating premium charges and policy riders.
* The descriptive statistics for the 'age' variable provide valuable insights into the distribution and characteristics of age among individuals in the dataset. Understanding these statistics is crucial for further analysis and modeling tasks, such as identifying patterns or relationships between age and other variables. Age is a strong predictor of future healthcare expenses and insurance claims. Research has shown that age is one of the most influential factors in predicting healthcare costs and utilization. Therefore, including age in the analysis can improve the predictive power of the model.
* The calculated mean absolute error for different values of n\_estimators provide valuable insights into the performance of the Random Forest Regression model. While the model demonstrates reasonable predictive accuracy, there remains room for improvement. By refining model hyperparameters and exploring additional feature engineering strategies, it is possible to enhance the model's predictive capabilities and achieve even better performance. Continued experimentation and evaluation will be crucial in further optimizing the model for the task of predicting insurance charges.

## References

Choi, M. (2018, February 21). *Medical Cost Personal Datasets.* Kaggle. <https://www.kaggle.com/datasets/mirichoi0218/insurance>

Cengage Learning (2021). *The Multiple Regression Model* [PowerPoint Presentation]. Conestoga College. <https://conestoga.desire2learn.com/d2l/le/content/997903/viewContent/21638864/View>

Cengage Learning (2021). *The Spreadsheet Models* [PowerPoint Presentation]. Conestoga College. <https://conestoga.desire2learn.com/d2l/le/content/997903/viewContent/21638869/View>

Cengage Learning (2021). *Linear Regression* [PowerPoint Presentation]. Conestoga College. <https://conestoga.desire2learn.com/d2l/le/content/997903/viewContent/21638863/View>