MATH2201 – How Are Insurance Charges Calculated?

Group 03 – Final Project Report

Active Project Members

Name	Week Joined
Samuel Macintyre	1
Karthikeya Reddy Mopur	1
Rahul Raj	1
Momitha Yepuri	2
Shivani Pillai	1

Executive Summary

We studied a dataset that contains 1338 rows of data, where health insurance premium charges were recorded alongside sex, age, BMI, number of children, smoking status, and region to answer four research questions.

- Is there a difference in charges between males and females?
- Is there a difference in charges between smokers and non-smokers?
- Is there a correlation between the other variables and charges?
- How do charges differ based on all variables?

Alongside introductory exploratory data analysis we performed hypothesis tests, 95% confidence intervals, chi-square goodness-of-fit tests and chi-square test for association to explore our first two research questions. We used linear and logistical regression to explore our other research questions.

We found that there was a difference in charges between males and females, and smokers and non-smokers. We also found that region and sex were not significant in predicting charges, and BMI and age would be better predictors of charges.

Aims and Introduction

In 2019, the US health consumption expenditures market was approximately US$3.6 trillion, accounting for 16.8% of the country's GDP (CRS, 2021). Out of an estimated population of 323 million, 90.8% had some form of health insurance or federal coverage and 9.2% were uninsured with no access to federal or other coverage (CRS, 2021). Additionally, the private health insurance expenditure accounted for USD$1.195 trillion, representing 33.3% of total health consumption expenditure (CRS, 2021). This means that the private health insurance expenditure alone represented 5.6% of USA's GDP.

A study in 2008 by non-profit 'Families USA', discovered that more than 26,000 individuals in the US die every year due to a lack of health insurance (PMC, 2008). From 2000 -2006 it was estimated that 162,700 individuals died from this cause, with lack of insurance being the third leading cause of death in US for individuals aged 55 to 64 (PMC, 2008). Some studies have shown that the likelihood of death is 25% higher in uninsured individuals aged 25-64. (PMC, 2008).

Health insurance functions by collecting up front premium from individuals on a regular basis. When an insured individual has medical expenses, these are then paid for by the insurer. In a bid to remain profitable, insurers will use a series of factors to determine how much to charge an individual as a premium based on the perceived level of risk for insuring that person. This represents a trade-off for the individual, they must sacrifice a fixed premium to insulate themselves from further financial risk through medical expenses, though in some cases these premiums can be quite costly.

Given the size of the industry and the impact being uninsured can have on mortality rates, we are setting out to identify the relationship between different demographic and lifestyle factors and the premiums charged. Identifying these relationships, if they exist, provides us with an insight into how insurance companies calculate premiums, which in turn can be useful in predicting the expected premiums of different individuals and allowing people to identify if any lifestyle choices can reduce their premiums.

This report will look at some of these factors as well as the premiums charged for 1338 individuals by examining a dataset of insurance premium charges in the US with important details for risk underwriting, provided by Anirban Datta, a data scientist at Krish Mark Infotech. The dataset contains 1338 rows of insured data where charges are given against age, sex, BMI, number of children, smoking status and region.

The following are the variables in the dataset:

Name	**Data Type**	**Units**	**Description**
Age	Numerical (Discrete)	Years	Age of the individual in years
Sex	Binary	NA	If the individual is Male/Female
BMI	Numerical(continuous)	kg/m^2	Body mass Index of the Individual
Children	Numerical (Discrete)	NA	Number of children covered by health insurance / Number of dependents.
Smoker	Binary	NA	If the individual smokes or not.
Region	Categorical (Nominal)	NA	The beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
Charges (target variable)	Numerical (continuous)	USD	Individual medical costs billed by health insurance.

This report will set out to answer the following questions:

1. Is there a difference in charges between males and females?
2. Is there a difference in charges between smokers and non-smokers?
3. Is there a correlation between age, sex, BMI, number of children, smoker vs non-smoker and region compared to charges and if so, how strong is the correlation?
4. How do charges differ based on age, sex, BMI, number of children, smoker vs non-smoker and region?

Data Preparation and Exploratory Data Analysis

Data Preparation

The following is a full list of the tasks that were completed and in the same order.

Task Description

1. Downloaded the insurance.csv file from https://www.kaggle.com/teertha/ushealthinsurancedataset
2. Imported the CSV file into R studio
3. Checked the dimensions of the dataset and the variable types of each column
4. Converted the incorrectly assigned “sex”, “smoker” and “region” character variables into categorical variables.
5. Scanned the dataset for missing values
6. Scanned the dataset for infinite values
7. Scanned the dataset for obvious errors and inconsistencies
8. Made boxplots of all numeric variables to check the presence of outliers
9. Located and removed rows with outlier values from the “bmi” and “charges” columns using the Z-score method
10. Created new numerical “sex_num”, “smoker_num” and “region_num” columns from the categorical variables “sex”, “smoker” and “region” to be utilized for linear regression
11. Saved the dataset as insurance_clean_csv

The dataset was thoroughly checked for all the missing values, infinite values, outliers and obvious errors and inconsistencies like negative number of children, negative charges etc. There were no missing values and errors but however, a total of 11 outliers were identified from the charges and BMI variables. After a deep assessment based on the z-scores and the importance of the outliers, it was decided that it would be best to remove the outliers since they would distort the graphs in the exploratory data analysis especially when considering the Mean. Moreover, to see impact of the presence of outliers, all the statistical tests such as hypothesis tests, confidence intervals, linear regression and logistic regression were done twice, once with and once without outliers.

Exploratory Data Analysis

Descriptive Statistics of Charges grouped by Male vs Female -

Statistics

Variable	sex	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
charges	female	662	0	12570	433	11129	1608	4871	9413	14460	63770
	male	676	0	13957	499	12971	1122	4564	9370	19024	62593

Descriptive Statistics of Charges grouped by Smoker vs Non-Smoker -

Statistics

Variable	smoker	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
charges	no	1064	0	8434	184	5994	1122	3983	7345	11363	36911
	yes	274	0	32050	697	11542	12829	20767	34456	41050	63770

Univariate Analysis

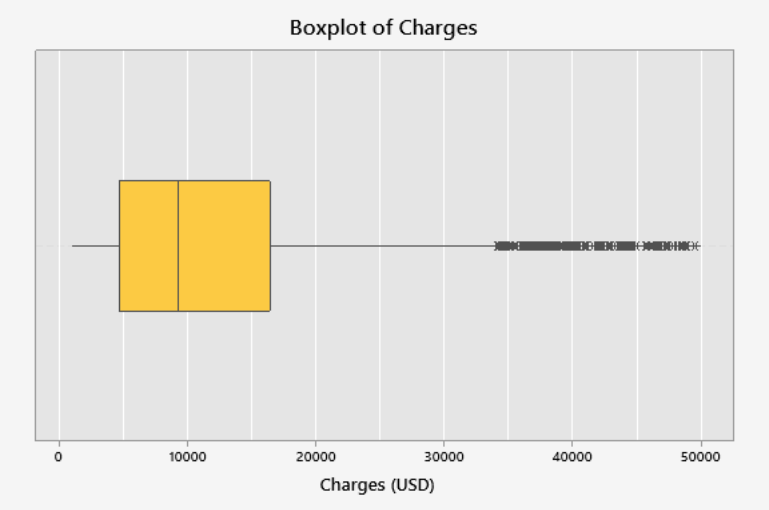

Figure 1.1 Boxplot of Charges

In ***Figure 1.1*** we notice that the values are highly right-skewed, and the distribution is more dispersed thus has a wider spread. This suggests that the upper 25% have much higher charges compared to the upper quartile value (75%).

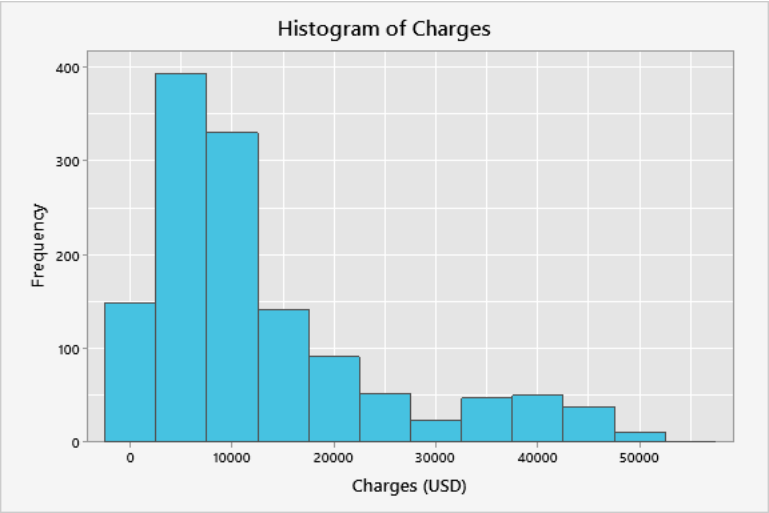

Figure 1.2 Histogram of Charges

The shape of ***Figure 1.2*** is right-skewed and bimodal, where we see that the most common charges occur between 4,000 to 12,000(USD).

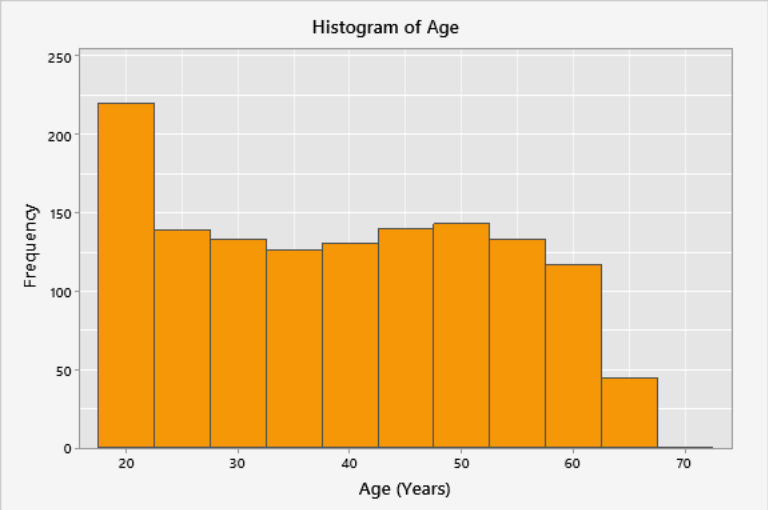

Figure 1.3 Histogram of Age

In ***Figure 1.3,*** the Age range seems to be representative of the true age distribution of the adult population. The shape appears uniformly distributed although an evident peak suggests greater frequency between the ages 18-22. While the age ranges from 18 to 64, the median age is 39.

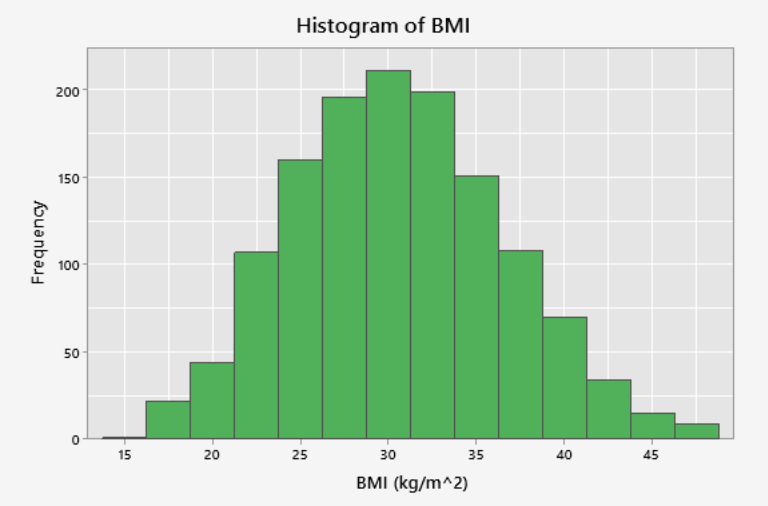

Figure 1.4 Histogram of Age

Figure 1.4 represents a normal distribution, and the shape appears to be fairly symmetric. The single peak indicates it is unimodal and the higher frequency seems to be between 26 - 34 kg/m^2. BMI ranges from 16 to 48 kg/m^2. The mean and median BMI are 30 kg/m^2 and IQR falls between 6 and 35kg/m^2.

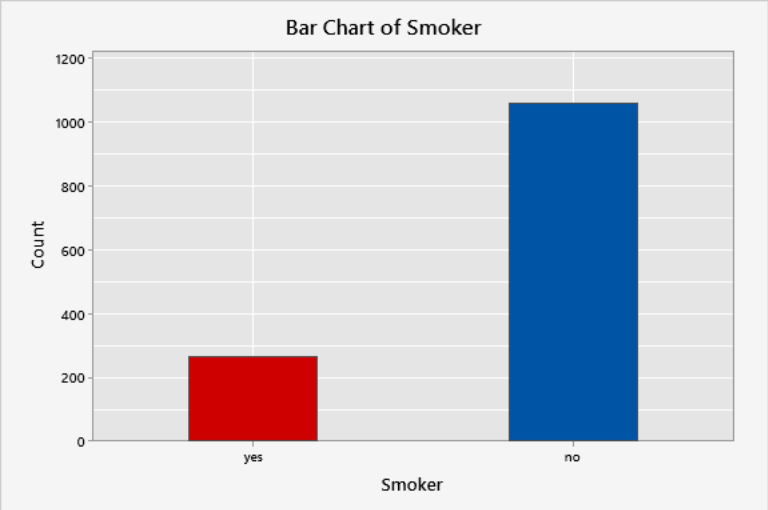

Figure 1.5 Bar Chart of Smoker

Figure 1.5 indicates that a large proportion of the population are non-smokers (80%) than smokers (20%).

Multivariate Analysis

Box Plot Analysis

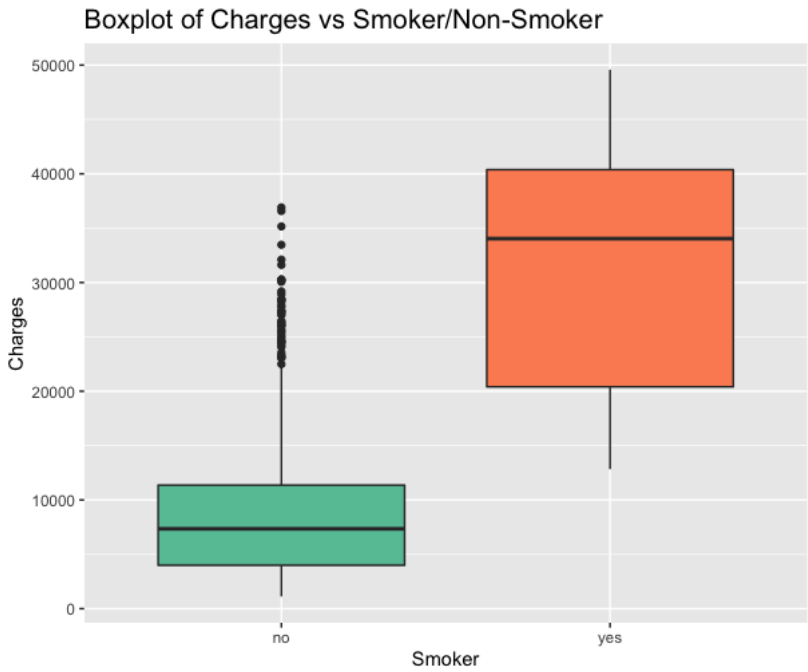

Figure 2.1 Box Plot of Charges vs Smoker

We observe that in ***Figure 2.1***, there's a major difference between the smoker/non-smokers as the median line of smoker falls out of the IQR range of non-smokers. The smokers' minimum value is starting at a higher rate than non-smokers minimum value and there appears to be significantly higher charges are claimed overall for smokers compared to non-smokers.
The distribution of the smoker box plot is negatively skew whereas the non-smoker plot is having slightly positively skew with a broader spread in the upper 25%.

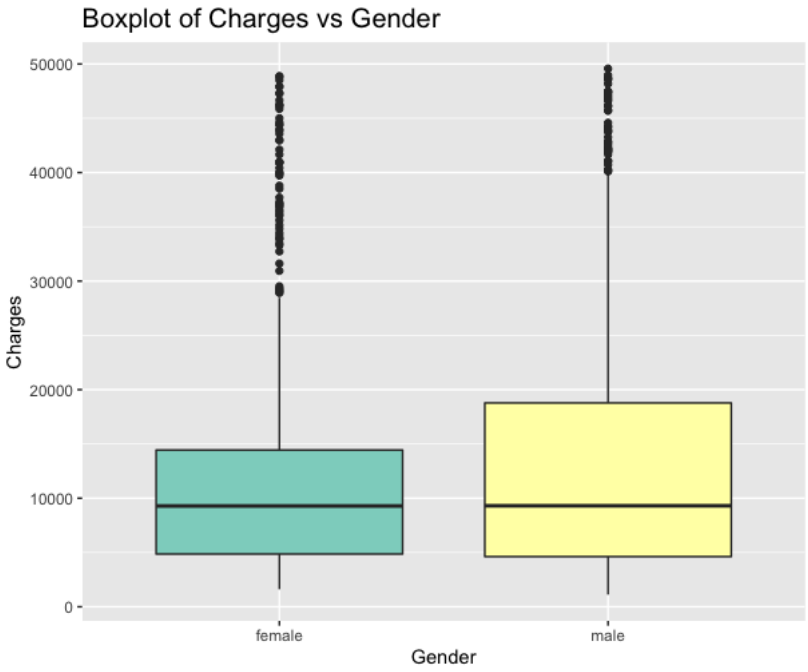

Figure 2.2 Boxplot of Charges vs Gender

In ***Figure 2.2,*** we notice that although males and females have precisely the same average charges, a larger percentage of males have higher charges than females as the IQR and the upper 25% is dispersed. The max value for charges is higher for males by USD 10,000.

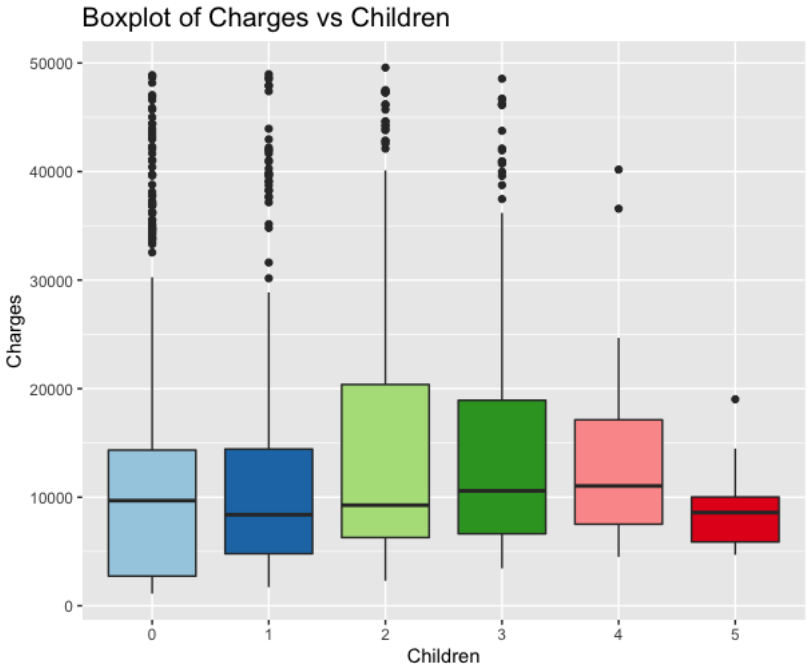

Figure 2.3 Box Plot of Charges vs Children

In ***Figure 2.3,*** we notice that people with 2-4 children have fairly high averages and people with 2 children have a wider spread of range in contrast to people with 5 children concentrated around the USD 10,000 range with one outlier. There is a greater concentration of higher charges with people with 2 children as they have more dispersed IQR as well as the wider upper 25%.

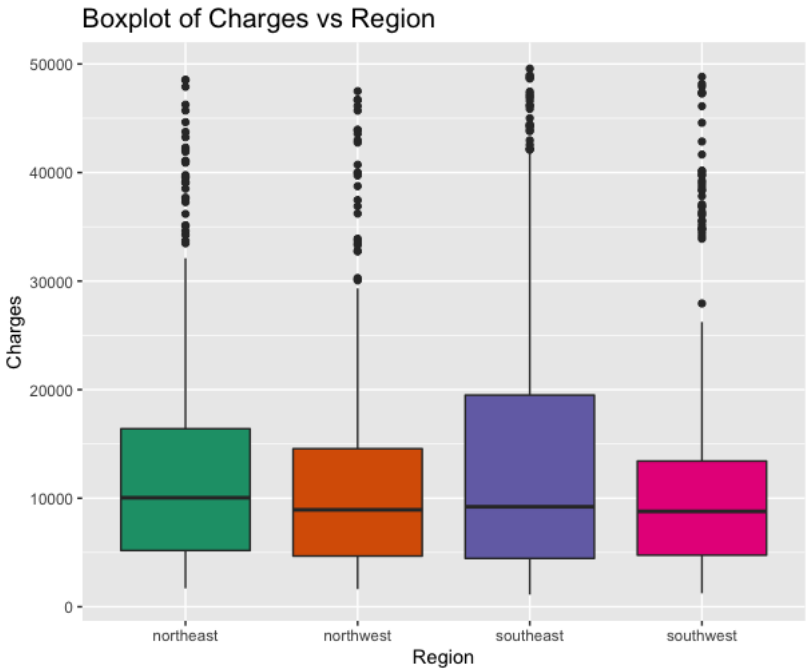

Figure 2.4 Boxplot of Charges vs Region

In ***Figure 2.4,*** The distribution by region is relatively similar, though we see a greater density of high-charge values in the southeast region plus a broader spread. The median seems to be roughly similar for people living in northwest, southeast and southwest compared to northeast's region where median is somewhat higher.

Scatter Plot Analysis

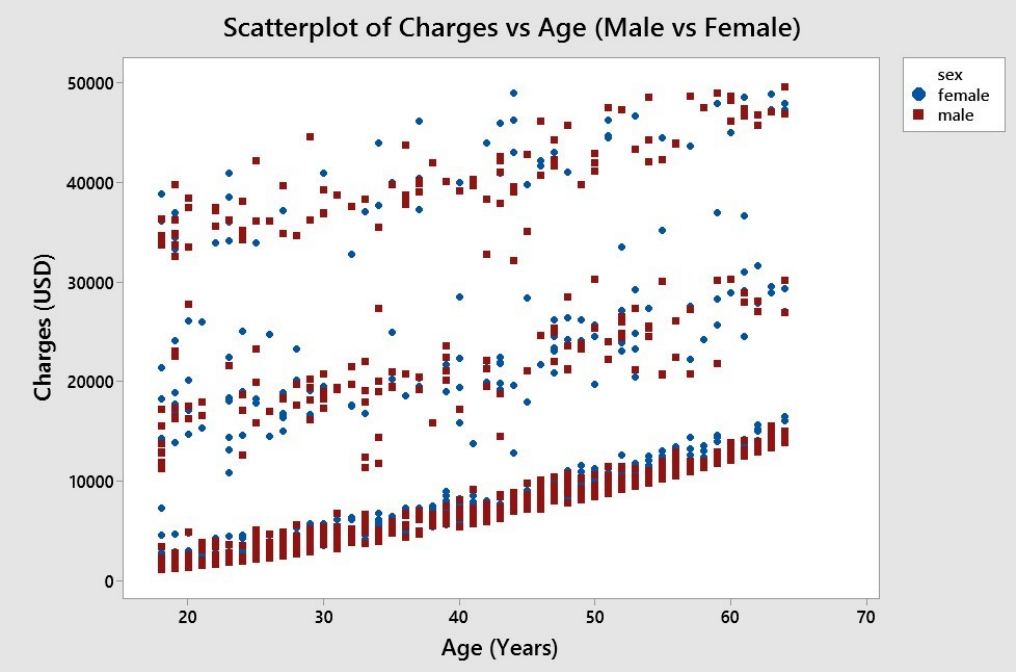

Figure 3.1 Scatter plot of Charges vs Age by Sex

In ***Figure 3.1,*** We see an interesting pattern where older males are charged more than younger males. However, there is no apparent pattern between age and charges for both males/ females claiming above 10,000(USD).

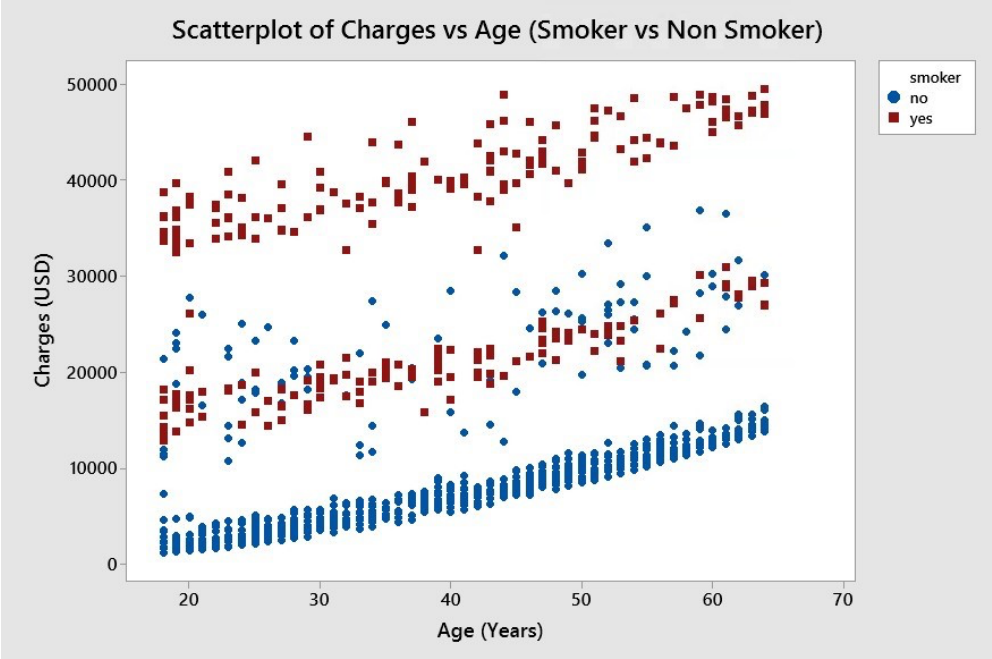

Figure 3.2 Scatter plot of Charges vs Age by Smoker/Non-Smoker

In ***Figure 3.2,*** The difference between the charges of smokers and non-smokers is evident. Although, Smokers and non-smokers are at the same age, the smokers seem to claim significantly more money than non-smokers. There's a gradual increase in charges for older individuals who are non-smokers.

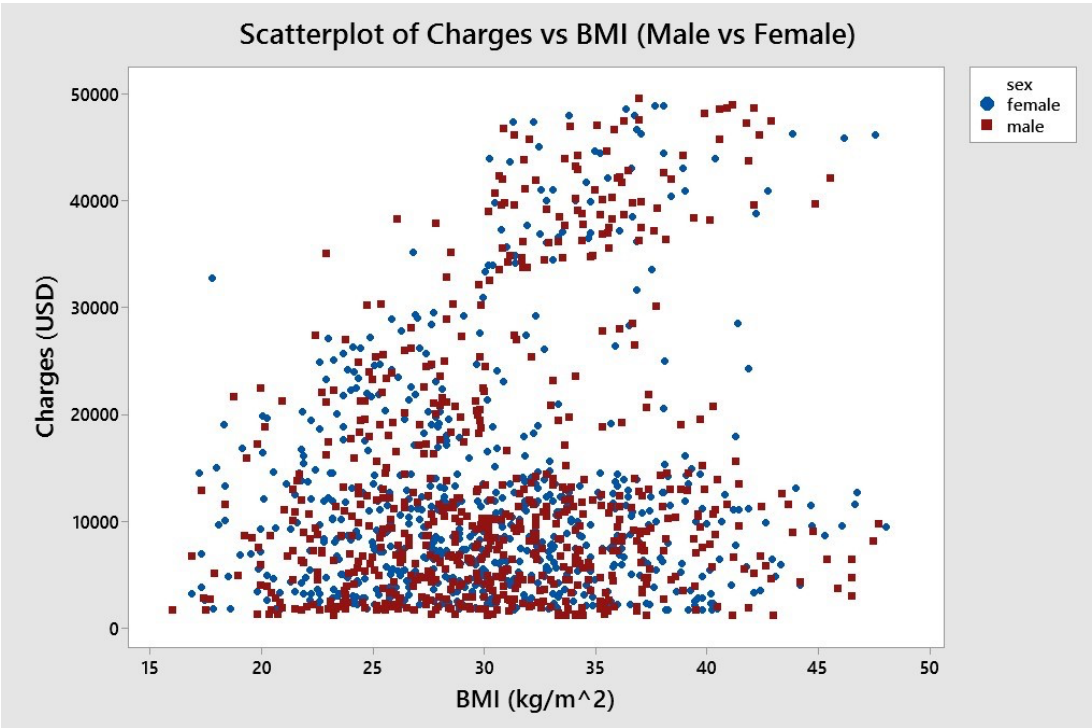

Figure 3.3 Scatter plot of Charges vs BMI by Sex

In ***Figure 3.3,*** It appears that as the BMI exceeds 30 for both males/females in roughly similar proportions, higher charges are claimed. A great density of people with BMI below 35 claimed lower charges mostly under USD 30,000.

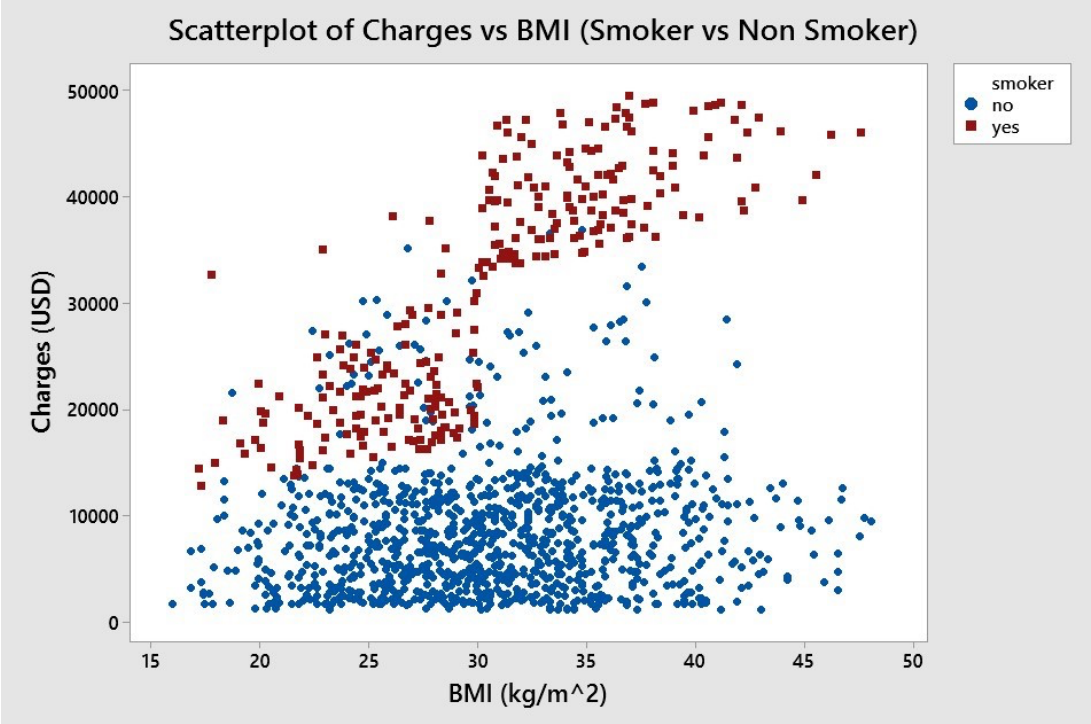

Figure 3.4 Scatter plot of Charges vs BMI by Smoker/Non-Smoker

Evidently, In ***Figure 3.4*** smokers tend to be charged significantly more compared to non-smokers overall. we see that as the BMI of an individual exceeds above 30 for smokers, the charges are remarkably high (above 30,000 USD). A blend of both smokers and non-smokers with BMI ranging from 18-42(kg/m^2) claim charges between 15,000-30,000(USD).

Bar Chart Analysis

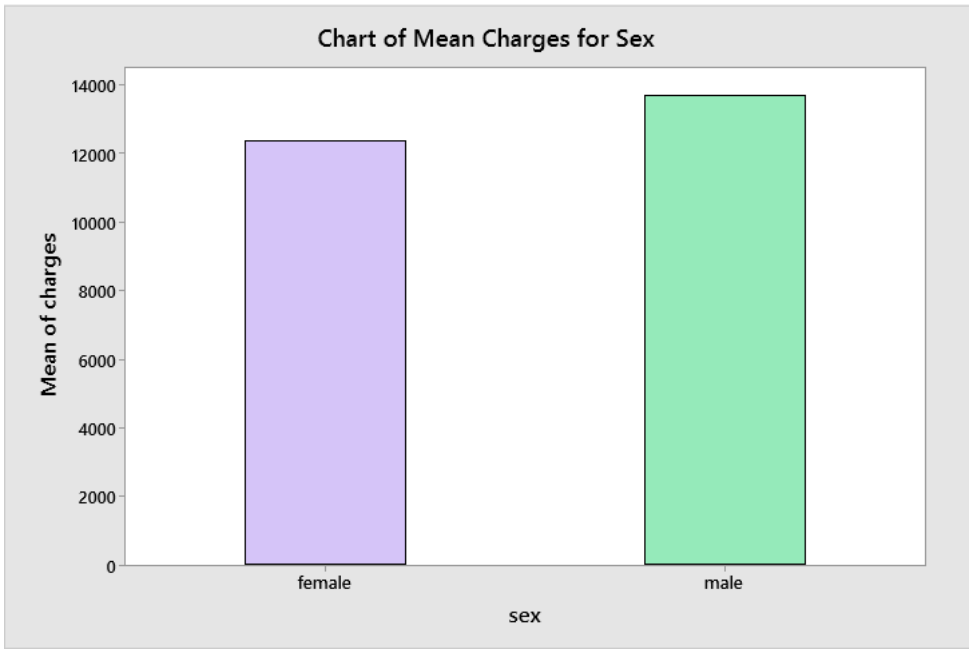

Figure 4.1 Bar Chart of Mean of Charges for Sex

In ***Figure 4.1,*** Males on average are charged $1,338.40 USD greater than females. So, while there is a difference in the charges it is not by a very large amount.

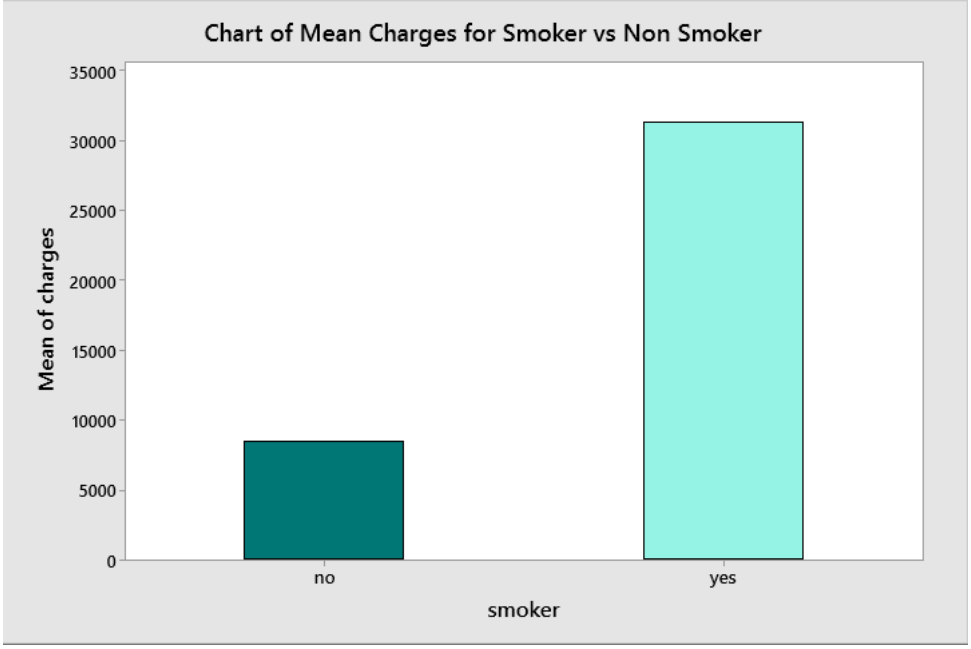

Figure 4.2 Bar Chart of Mean of Charges for Smoker vs Non-Smoker

In **Figure *4.2,*** We can very clearly see smokers on average are charged at a much higher rate than those who do not smoke. With non-smokers paying on average $22,884.51 USD less than their smoker counterparts.

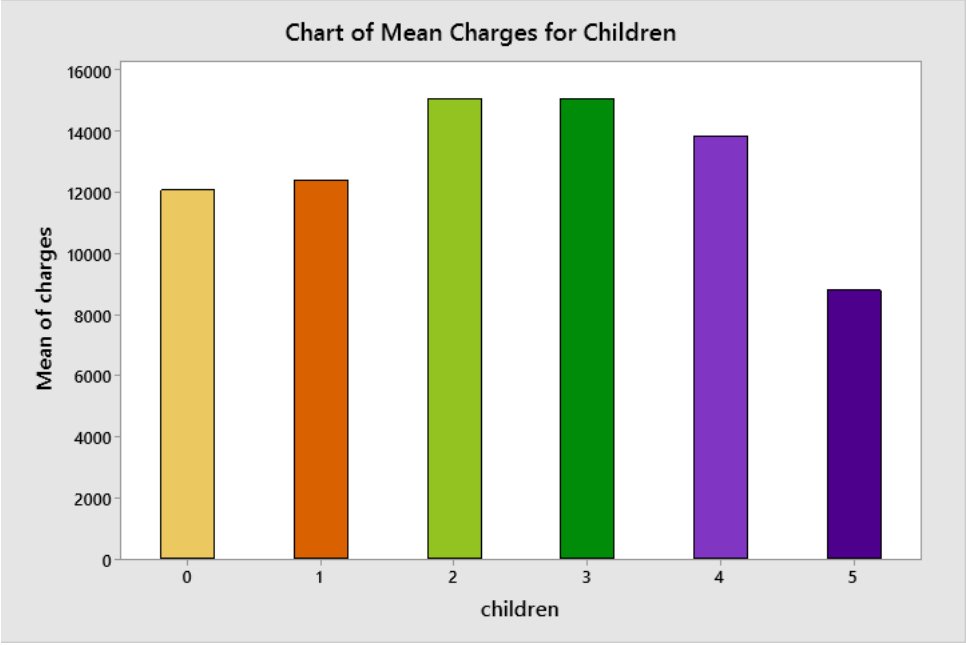

Figure 4.3 Bar Chart of Mean of Charges for Children

In **Figure *4.3,*** Those with 0-1 children are charged at a similar rate and we can see that these charges increase for 2-3 children, then decrease slightly for those with 4 children and then decrease even further so that those with 5 children are charged on average the least by $3,608.56 USD.

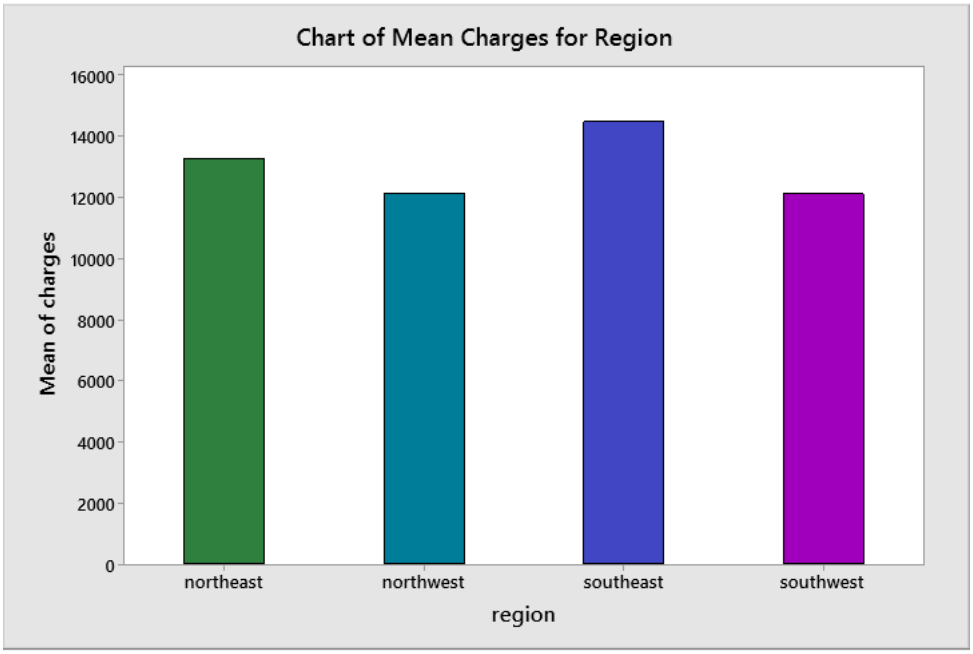

Figure 4.4 Bar Chart of Mean of Charges for Region

In **Figure *4.4,*** Both the western regions (northwest and southwest) have a very similar average charge which is less than the two eastern (northeast and southeast) regions. With the southeast region having the highest charges by $1,196.60 USD.

Methodology

Linear Regression

In order to answer question 3 ("Is there a correlation between age, sex, BMI, number of children, smoker vs non-smoker and region compared to charges and if so, how strong is the correlation?") we build a linear regression model.

Linear regression is used to model the relationship between 1 or more independent variables (input variables) and a dependent variable (predicted variable), which can then be used to predict other dependent variables if the independent variables are already known. Linear regression models are simple and provide an easily interpretable formula that can be used to identify relationships and predict future values.

As the name suggests linear regression models will fit a linear equation creating a line of best fit to a set of data. The line of best fit is determined by the method of least squares, minimising the Sum of Square Errors (SSE).

The accuracy of the model is provided by an R^2 value.

Figure 4.5 shows a graphical representation of how linear regression is calculated.

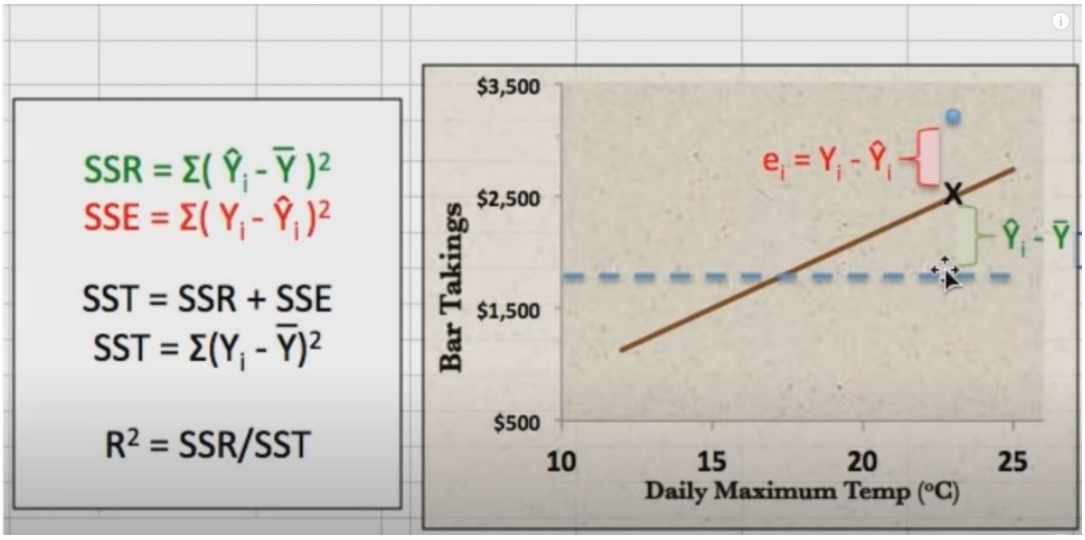

Figure 4.5 Method for Linear Regression (Source Unknown)

Linear regression models can also be generated with several additional analyses such as:

- Variance Inflation Factor: how much the variance of an independent variable is influenced by other independent variables (essentially a measure of multicollinearity).
- P-value: measuring the statistical significance of a variable
- Residual plots: there are a few residual plots such as:
    - Normal Probability Plot: identifying if the error terms are normally distributed (linear indicates normally distributed.)
    - Residual vs Fitted: used to detect unequal error variances, outliers on non-linearity of the relationships.
    - Histogram of the residual frequencies.
    - Residual vs Order plot: Looks for correlations between the error terms based on location.

Hypothesis Testing –

We decided to investigate RQ 1: Is there a difference in charges between males/females and RQ2: Is there a difference in charges between smokers/non-smokers. As we are comparing the means of two different groups here, hypothesis testing allows us to evaluate the strength of evidence from the sample and draw inferences about a population parameter. Then, we propose our two hypotheses to investigate which claim is true. We will be performing two tailed tests for each RQ twice, (i.e with & without outliers) to determine whether the difference between the two populations is statistically significant.

H_0: The null hypothesis is the prediction of no relationship between variables, that is assumed to be true. Our null hypothesis states that there is no difference in charges between the between the specified populations.

H_a: The alternate hypothesis predicts there is a relationship between the variables. Our alternative hypothesis states that there is a difference in charges between specified, populations.

The next step is to compute the test statistic, it shows how closely our observed data matches the distribution expected under the null hypothesis of the statistical test.

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Using the formula above, where, sample size **n**, with sample mean $\bar{\mathbf{x}}$, population mean $\boldsymbol{\mu}$, and with standard deviation **s**. A p-value describes how likely it is that out data would have occurred by random chance (i.e. that the null hypothesis is true). We use the value of the test statistic to decide regarding the null hypothesis.

Using a significance level of 0.05%. We reject the null hypothesis if the p-value is less than the alpha level (P-value < α) suggesting that the results are statistically significant. Or state that the null hypothesis is plausible if the p-value is higher than the alpha level, suggesting we don't have enough evidence to reject the null hypothesis and that the results are not statistically significant.

Confidence Intervals -

As we used the sample means to investigate RQ1: Is there a difference in charges between males and females? and RQ2: Is there a difference in charges between smoker and non-smokers? We now need to see how our results translate to the larger population.
We decided to use 95% confidence intervals to investigate what likely range the population mean would lie in with 95% certainty. We will be calculating the confidence intervals for each research question with and without outliers as well, to once again see if the addition of outliers notably impacts the results.

We shall be using the formula $CI = \bar{x} \pm z\frac{s}{\sqrt{n}}$, where $\bar{x}$ is the sample mean, s is the sample standard deviation, n is the sample size, and as we are calculating 95% confidence intervals our z value will be 1.96.

Logistic Regression –

Logistic regression is a type of statistical analysis is often used for predictive analytics and modelling and extends to applications in machine learning. In this analytics approach, the dependent variable is finite or categorical. It is used in statistical software to understand the relationship between the dependent variable and one or more independent variables by estimating probabilities using a logistic regression equation.

Logistic analysis can help you predict the likelihood of an event happening or a choice being made, Predictive models built using this approach can make a positive difference in a business or organization. Because logistic analysis can help understand relationships and predict outcomes and improve decision-making. In our report we have chosen our categorical variable to be smokers vs non-smoker. Choosing the variable smoker vs non-smoker, we are able to determine the relationship with other variables such as age, bmi, charges, sex, region, and children identifying the significance of a smoker when compared to other variables, the accuracy of the dataset and finding the coefficients for each variable against smokers. We also have done this method twice once with the dataset with outliers and another without outliers too see the difference.

Chi-Squared Tests –

To further explore RQ1 and RQ2 we performed a chi-squared goodness of fit on our variables – sex and smoker. The chi-squared goodness of fit is used to determine if our variables, which are taken from a sample, are an accurate representation of the true population. For our chi-squared goodness of fit test for sex, we used equal proportions because as per the population we would expect our sample to be an even split between males and females. Then for our chi-squared goodness of fit test for the smoker variable, we used the latest data from the Centers of Disease Control and Prevention (CDC) from 2019 that showed 14% of adults in the US smoke. Therefore, we set our expected proportions as 84% for non-smokers and 14% for smokers.

Chi-squared test for association is used to determine if two factors in a single population are independent or not. We will be using it to even further explore RQ1 and RQ2. However, chi-squared tests for association required both variables to be categorical, our sex and smoker variables are categorical, but we needed to test them against charges, which is a numerical value. To be able to proceed with the tests, we organised our data by charges in ascending order, we then split it into three group – Low, Medium and High. Since our charges are now categorical, we could use them in the chi-square test for association for sex vs charges and smoker vs charges.

However, when changing numerical values to categorical we lose the accuracy of the data. So, to supplement the chi-squared tests for associations, we need to also refer to the hypothesis testing as our hypothesis testing compares the means of two different groups to determine if there is a statistically significant difference in the means for both groups.

Results

Hypothesis Testing – Males/Females & Smoker/Non-Smokers

In this section, we want to investigate our Research Question 1: Is there a difference in charges between Males and Females. To answer this, we'll conduct a two-sample z-test on data with and without outliers to see if it's making any significant impact.

Test 1- Is there a difference in charges between Males and Females?

Minitab Steps:

1. Separated by the charges of Male and Female
2. Pasted in individual Columns in Minitab
3. Stat → 2-Sample t-test → Each Sample in its own Column → Confidence Level: 95.0 → Alt Hypothesis: Difference ≠ Hypothesized Difference.
4. Performed steps 1-3 with the outliers' dataset.

Part 1 – With Outliers

Our null hypothesis is that there is no difference in the mean charges between males and females.

μ_M = Population mean of chargers for males

μ_F = Population mean of chargers for females

H_0: $\mu_M - \mu_F = 0$

The alternative hypothesis is that there is a difference in the mean charges between males and females.

H_A: $\mu_M - \mu_F \neq 0$

For this test we will consider that for the null hypothesis to be rejected the p-value must be less than 0.05.

$\alpha = 0.05$

Method

μ_1: mean of Male
μ_2: mean of Female
Difference: $\mu_1 - \mu_2$

Equal variances are not assumed for this analysis.

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
Male	676	13957	12971	499
Female	662	12570	11129	433

Estimation for Difference

Difference	95% CI for Difference
1387	(92, 2682)

Test

Null hypothesis H_0: $\mu_1 - \mu_2 = 0$
Alternative hypothesis H_1: $\mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
2.10	1313	0.036

Figure 5.1 Hypothesis Testing for Males/Females

P-value < α
0.036 < 0.05

As the **P-value is substantially less than α**, the null hypothesis is rejected and we can state that, there is a statistically significant difference in charges between males and females.

Part 2 – Without Outliers

Method

μ_1: mean of Male
μ_2: mean of Female
Difference: $\mu_1 - \mu_2$

Equal variances are not assumed for this analysis.

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
Male	668	13696	12542	485
Female	659	12357	10696	417

Estimation for Difference

Difference	95% CI for Difference
1338	(84, 2593)

Test

Null hypothesis $H_0: \mu_1 - \mu_2 = 0$
Alternative hypothesis $H_1: \mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
2.09	1297	0.037

Figure 5.2 Hypothesis Testing for Males/Females

P-value < α
0.037 < 0.05
As the **P-value is substantially less than α**, the null hypothesis is rejected and we can state that, there is a statistically significant difference in charges between males and females.

Test 2 - Is there a difference in charges between Smokers and Non-smokers?

The second test will investigate our Research Question 2: Is there a difference in charges between Smokers and Non-smokers. To answer this, we'll conduct a two-sample z-test on data with and without outliers to see if it's making any significant impact.

Minitab Steps:

1. Separated by the charges of Smokers and Non-Smokers
2. Pasted in individual Columns in Minitab
3. Stat → 2-Sample t-test → Each Sample in its own column → Confidence Level: 95.0 → Alt Hypothesis: Difference ≠ Hypothesized Difference.
4. Performed steps 1-3 with the outliers' dataset.

Part 1 (With Outliers)

Our null hypothesis is that there is no difference in the mean charges between smokers and non-smokers.
μ_S = Population mean of charges for smokers
μ_N = Population mean of charges for non-smokers
$H_0: \mu_S - \mu_N = 0$

The alternative hypothesis is that there is a difference in the mean charges between smokers and non-smokers.
$H_A: \mu_S - \mu_N \neq 0$

For this test we will consider that for the null hypothesis to be rejected the p-value must be less than 0.05.
α = 0.05

Method

μ_1: mean of Smoker
μ_2: mean of Non-Smoker
Difference: $\mu_1 - \mu_2$

Equal variances are not assumed for this analysis.

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
Smoker	274	32050	11542	697
Non-Smoker	1064	8434	5994	184

Estimation for Difference

Difference	95% CI for Difference
23616	(22197, 25035)

Test

Null hypothesis	$H_0: \mu_1 - \mu_2 = 0$
Alternative hypothesis	$H_1: \mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
32.7519	311	0.0000000000000000

Figure 5.3 Hypothesis Testing for Smoker/Non-Smokers

P-value < α
0.00 < 0.05
As the **P-value is substantially less than α**, the null hypothesis is rejected and we can state that, there is a statistically significant difference in charges between smokers and non-smokers.

Part 2 (Without Outliers)

Method

μ_1: mean of Smoker
μ_2: mean of Non-Smoker
Difference: $\mu_1 - \mu_2$

Equal variances are not assumed for this analysis.

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
Smoker	266	31329	10872	667
Non-Smoker	1061	8444	5995	184

Estimation for Difference

Difference	95% CI for Difference
22885	(21524, 24245)

Test

Null hypothesis	$H_0: \mu_1 - \mu_2 = 0$
Alternative hypothesis	$H_1: \mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
33.0906	306	0.0000000000000000

Figure 5.4 Hypothesis Testing for Smoker/Non-Smokers

P-value < α
0.00 < 0.05
As the **P-value is substantially less than α**, the null hypothesis is rejected and we can state that, there is a statistically significant difference in charges between smokers and non-smokers.

Confidence Intervals

Using 95% confidence intervals we can provide an estimate of the means and their difference for smoker pressed against charges and for non-smoker pressed against charges.

Statistics

Variable	smoker	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
charges	no	1061	0	8444	184	5995	1122	3989	7346	11363	36911
	yes	266	0	31329	667	10872	12829	20281	34037	40494	49578

95% Confidence intervals for Smoker vs Non-smoker

95% Confidence intervals for Smoker

$$95\%CI = \overline{x} \pm \frac{1.96sx}{\sqrt{n}} \qquad 31329 \pm \frac{1.96 \times 10872}{\sqrt{266}}$$

95% CI = (30022.45, 32635.55)

The 95% confident that the charges mean for Smokers falls between 30022.45 and 32635.55.

95% Confidence intervals for Non-Smoker

$$95\%CI = \overline{x} \pm \frac{1.96sx}{\sqrt{n}} \qquad 8444 \pm \frac{1.96 \times 5995}{\sqrt{1061}}$$

95% CI = (8083.26, 8804.73)

The 95% confident that the charges mean for non-Smokers falls between 8083.26 and 8804.73.

95% CI for difference in means

$$31329 - 8444 \pm \sqrt{\frac{10872^{2\cdot}}{266} + \frac{5995^2}{1061}}$$

95% CI for difference in smoker vs non-smoker against charges = (22193.45, 23576.55)

95% Confidence intervals for Smoker vs Non-smoker with Outliers

Statistics

Variable	smoker	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
charges	no	1064	0	8434	184	5994	1122	3983	7345	11363	36911
	yes	274	0	32050	697	11542	12829	20767	34456	41050	63770

95% Confidence intervals for Smoker

$$95\%CI = \overline{x} \pm \frac{1.96sx}{\sqrt{n}} \qquad 32050 \pm \frac{1.96 \times 11542}{\sqrt{274}}$$

95% CI = (30683.34, 33416.66)

The 95% confident that the charges mean for Smokers falls between 30683.34 and 33416.66.

95% Confidence intervals for Non-Smoker

$$95\%CI = \overline{x} \pm \frac{1.96sx}{\sqrt{n}} \qquad 8434 \pm \frac{1.96 \times 5994}{\sqrt{1064}}$$

95% CI = (8073.93, 8794.17)

The 95% confident that the charges mean for non-Smokers falls between 8073.93 and 8794.17.

95% CI for difference in means

$$32050 - 8434 \pm \sqrt{\frac{11542^2}{274} + \frac{5994^2}{1064}}$$

95% CI for difference in smoker vs non-smoker against charges = (22894.92, 24337.08)

95% Confidence intervals for Male vs Female

Statistics

Variable	sex	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
charges	female	659	0	12357	417	10696	1608	4831	9284	14450	48885
	male	668	0	13696	485	12542	1122	4564	9303	18797	49578

95% CI for Males

$$95\%CI = 13696 \pm 1.96\frac{12542}{\sqrt{668}}$$
$$95\%CI = (12744.88, 14647.12)$$

There is 95% confidence that the mean charge for males falls between USD$12,744.88 and USD$14,647.12.

95% CI for Females

$$95\%CI = 12357 \pm 1.96\frac{10696}{\sqrt{659}}$$
$$95\%CI = (11540.35, 13173.65)$$

There is 95% confidence that the mean charge for females falls between USD$11,540.35 and USD$13,173.65.

95% CI for difference in mean

$$95\%CI = 13696 - 12357 \pm 1.96\sqrt{\frac{12542^2}{668} + \frac{10696^2}{659}}$$

$$95\%CI = (85.39, 2592.61)$$

We can say with 95% confidence that the charges for males are between USD$85.39 and USD$2,592.61 higher than the charges for females.

95% Confidence intervals for Male vs Female with Outliers

Statistics

Variable	sex	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
charges	female	662	0	12570	433	11129	1608	4871	9413	14460	63770
	male	676	0	13957	499	12971	1122	4564	9370	19024	62593

95% CI for Males with Outliers

$$95\%CI = 13957 \pm 1.96\frac{12971}{\sqrt{676}}$$

$$95\%CI = (12979.19, 14934.81)$$

There is 95% confidence that the mean charge for males including the outliers falls between USD$12,979.19 and USD$14,934.81.

There is an overlap of USD$12,979.19 and USD$14647.12 for mean charges for males between the data containing outliers and that without outliers.

95% CI for Females with Outliers

$$95\%CI = 12570 \pm 1.96\frac{11129}{\sqrt{662}}$$

$$95\%CI = (11722.22, 13417.78)$$

There is 95% confidence that the mean charge for females including the outliers falls between USD$11,722.22 and USD$13,417.78.

There is an overlap of USD$11.722.22 and US$13,173.65 for mean charges for females between the data containing outliers and that without outliers.

95% CI for difference in Mean with Outliers

$$95\%CI = 13957 - 12570 \pm 1.96\sqrt{\frac{12971^2}{676} + \frac{11129^2}{662}}$$

$$95\%CI = (92.84, 2681.16)$$

We can say with 95% confidence that the charges for males are between USD$92.84 and USD$2,681.16 higher than the charges for females.
There is an overlap of USD$92.84 and US$2,592.61 for the difference in mean charges for data containing outliers and that without outliers.

Chi-Squared Goodness-of-Fit for Sex

To perform a chi-squared goodness of fit we did the following steps in Minitab.

1. Loaded the clean data into Minitab worksheet
2. Stat → Tables→Chi-Square Goodness-of-Fit Test (One Variable)
3. Set variables
    a. Categorical Data: Sex
    b. Test: Equal Proportions

***H*$_0$**: $p_1 = 0.5$, $p_2 = 0.5$

***H*$_1$**: not all the p_i's are as specified

Our null hypothesis states that our proportions for males and females will be equal while our alternative hypothesis states that our proportions for males and females will not be equal. To calculate the critical value for the rejection region we are using **α = 0.05.**

Observed and Expected Counts

Category	Observed	Test Proportion	Expected	Contribution to Chi-Square
female	659	0.5	663.5	0.0305200
male	668	0.5	663.5	0.0305200

Chi-Square Test

N	N*	DF	Chi-Sq	P-Value
1327	0	1	0.0610399	0.805

For this variable we get a chi-square value of 0.061. We then found the critical value for the rejection region, using $X^2_{0.05,1} = 3.843$. Since our value of 0.061 does not fall in the rejection region we cannot reject ***H*$_0$**, which states that are proportions are equal and therefore an accurate representation of the entire population.

When this test was run again with outliers added, we get a chi-square value of 0.146 which still means our ***H*$_0$** is not rejected.

Chi-Squared Goodness of Fit for Smoker

To perform a chi-squared goodness of fit we did the following steps in Minitab.

1. Loaded the clean data into Minitab worksheet
2. Stat → Tables→Chi-Square Goodness-of-Fit Test (One Variable)
3. Set variables
    a. Categorical Data: Smoker
    b. Proportions specified by historical counts: Input Constants
    c. Historical Counts: No – 86%, Yes- 14%

***H*$_0$**: $p_1 = 0.86$, $p_2 = 0.14$

***H*$_1$**: not all the p_i's are as specified

Our null hypothesis states that our proportions for those who don't smoke will be 0.86 to those who smoke at 0.14, while our alternative hypothesis states that our proportions for smokers and non-smokers will not be as specified. To calculate the critical value for the rejection region we are using **α = 0.05.**

Observed and Expected Counts

Category	Observed	Historical Counts	Test Proportion	Expected	Contribution to Chi-Square
no	1061	86	0.86	1141.22	5.6389
yes	266	14	0.14	185.78	34.6391

Chi-Square Test

N	N*	DF	Chi-Sq	P-Value
1327	0	1	40.2780	0.000

For this variable we get a chi-square value of 40.278. We then found the critical value for the rejection region, using $X^2_{0.05,1}$ = 3.843. Since our value of 40.278 falls in the rejection region we reject H_0, which states that are proportions are 0.86 for non-smokers, and 0.14 for smokers and accept H_1, which states that the proportions are not as specified. This shows that our sample for this variable is not a true representation of the full population.

When this test was run again with outliers added, we get a chi-square value of 46.640 which still means our H_0 is rejected, and that the proportions are still not as specified.

Chi-Squared Test for Association of Sex vs Charges

To perform a chi-squared test for association we did the following steps in Minitab.

1. Loaded the data with the charges categorized into Minitab worksheet
2. Stat → Tables→Chi-Square Test for Association
3. Raw Data (categorical values)
    a. Row: Sex
    b. Columns: Charges Categorical

H_0: Sex and Charges are independent factors.

H_1: Sex and Charges are dependent factors

We conducted the test using **α = 0.05.**

Rows: sex Columns: Charges Categorical

	High	Low	Medium
female	209	221	229
male	232	222	214

Cell Contents
Count

Chi-Square Test

	Chi-Square	DF	P-Value
Pearson	1.649	2	0.439
Likelihood Ratio	1.649	2	0.438

For this test we get Chi-Square = 1.649 and P-value = 0.439

P-value > α

0.439 > 0.05

As the **P-value is greater than α**, the null hypothesis is not rejected and we can state that the variables of sex and charges are independent variables, and therefore do not influence each other.

When this test was run again with outliers, we got Chi-Square = 1.979 and P-value = 0.372 so the same inferences can be made.

Chi-Squared Test for Association of Smoker vs Charges

To perform a chi-squared test for association we did the following steps in Minitab.

1. Loaded the data with the charges categorized into Minitab worksheet
2. Stat → Tables→Chi-Square Test for Association
3. Raw Data (categorical values)
    a. Row: Smoker
    b. Columns: Charges Categorical

H_0: Smoker and Charges are independent factors.

H_1: Smoker and Charges are dependent factors

We conducted the test using **α = 0.05.**

Rows: smoker Columns: Charges Categorical

	High	Low	Medium
no	175	443	443
yes	266	0	0

Cell Contents
Count

Chi-Square Test

	Chi-Square	DF	P-Value
Pearson	668.394	2	0.000
Likelihood Ratio	737.287	2	0.000

For this test we get Chi-Square = 668.394 and P-value = 0.000

P-value < α

0.000 < 0.05

As the **P-value is lesser α**, the null hypothesis is rejected, and the alternative hypothesis is accepted at a statistically significant level. Therefore, we can state that the variables of smoker and charges are dependent variables and do influence each other.

When this test was run again with outliers, we got Chi-Square = 689.120 and P-value = 0.000 so the same inferences can be made.

Regression Analysis

Regression Analysis Without Outliers

Performing a multivariate linear regression, we took the following steps in Minitab.

1. Loaded the clean data into Minitab worksheet
2. Stat → Regression →Fit Regression Model
3. Set variables
    a. Responses: charges
    b. Continuous Predictors: age, bmi
    c. Categorical Predictors: sex, children, smoker, region
    d. Graphs: Four in one

This provided us with the following model shown in Figure 6.1.

Regression Equation

charges = -11358 + 255.4 age + 320.1 bmi + 0.0 sex_female - 86 sex_male + 0.0 children_0
+ 365 children_1 + 1807 children_2 + 1003 children_3 + 3050 children_4
+ 1120 children_5 + 0.0 region_northeast - 472 region_northwest
- 974 region_southeast - 985 region_southwest + 0.0 smoker_no + 23151 smoker_yes

Figure 6.1 Linear Regression Equation (Outliers Removed)

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-11358	965	-11.76	0.000	
age	255.4	11.5	22.21	0.000	1.02
bmi	320.1	28.0	11.42	0.000	1.11
sex					
male	-86	321	-0.27	0.789	1.01
children					
1	365	407	0.90	0.369	1.19
2	1807	449	4.03	0.000	1.17
3	1003	528	1.90	0.058	1.13
4	3050	1190	2.56	0.010	1.03
5	1120	1398	0.80	0.423	1.03
region					
northwest	-472	459	-1.03	0.303	1.52
southeast	-974	461	-2.11	0.035	1.64
southwest	-985	460	-2.14	0.033	1.53
smoker					
yes	23151	403	57.50	0.000	1.02

Figure 6.2 Linear Regression Coefficients (Outliers Removed)

There are a few important factors here to note. Firstly, is that the categorical variables for sex, region, smoker and children hold individual coefficients for each of the values, with the value itself being binary. As such if an individual is of sex female, then sex_female = 1 and sex_male = 0. Another observation is that sex_female, children_0, region_northeast and smoker_no all hold coefficients of 0, we can assume that these are factored into the regression equation for charges = constant + x*age + y*bmi, essentially we can consider this the baseline from which alternate values will cause the charges to deviate.

Looking at a breakdown of the coefficients in ***Figure 6.2***, we note that the baseline values for categorical variables are absent as to be expected. For this section of the analysis, we are most interested in the Variance Inflation Factor (VIF) to measure the multicollinearity amongst variables. Here we can see that the VIF is relatively low, giving us confidence that they are independent (e.g., region does not influence sex etc.).

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
5815.93	75.42%	75.19%	74.89%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	12	1.36352E+11	11362637132	335.92	0.000
age	1	16679978344	16679978344	493.13	0.000
bmi	1	4411830028	4411830028	130.43	0.000
sex	1	2433172	2433172	0.07	0.789
children	5	748595144	149719029	4.43	0.001
region	3	208322635	69440878	2.05	0.105
smoker	1	1.11851E+11	1.11851E+11	3306.75	0.000
Error	1314	44446034975	33824989		
Lack-of-Fit	1311	44184698652	33703050	0.39	0.948
Pure Error	3	261336323	87112108		
Total	1326	1.80798E+11			

Figure 6.3 Linear Regression Model Summary and ANOVA (Outliers Removed)

As evident ***in Figure 6.3*** the coefficient of determination (R-sq) for this model is 75.42%, meaning 75.42% of values can be explained by the model, given that there is a wide array of other factors and personal circumstances that insurance companies would use to determine charges, it is fair to say that this a reasonable R-sq value for the data set we are operating on.

It is important to also note the P-values of our variables, with the null hypothesis that a variable has no significance, and our alternate hypothesis is that they are significant. Using a significance level of 0.05 we can see that both sex and region hold a greater P-value thus we fail to reject the notion that they are statistically insignificant variables. For our age, BMI, children and smoker variables, our P-values are all under 0.05 so we can reject the null that they are insignificant and accept the alternate hypothesis that these variables have a statistical significance in predicting charges.

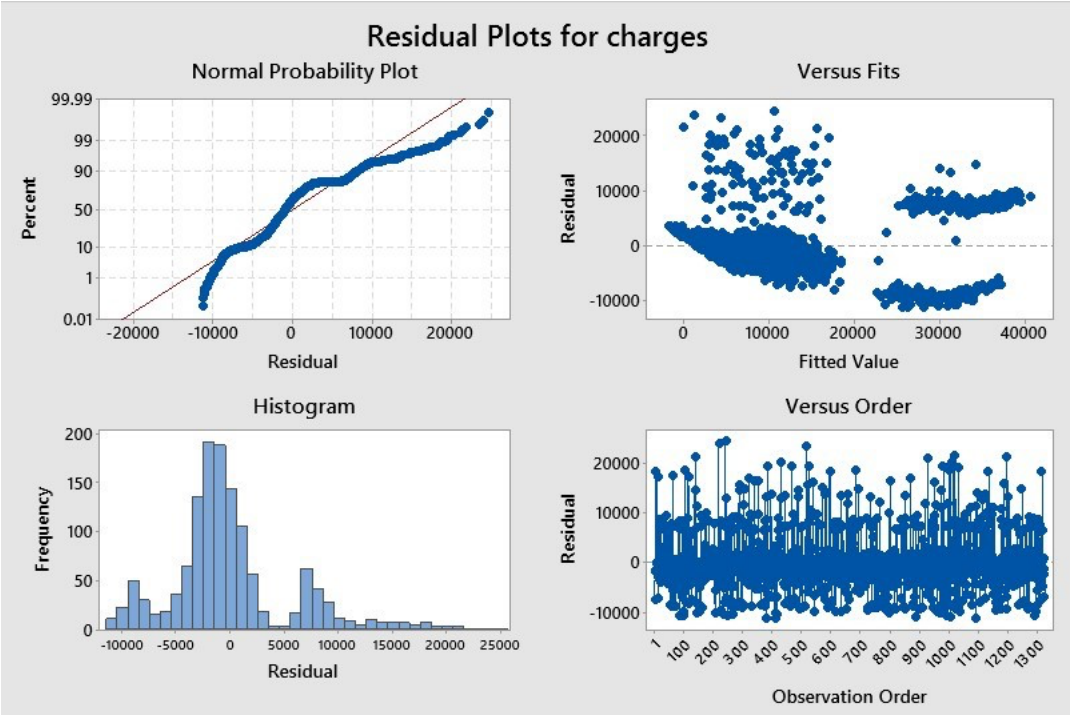

Figure 6.4 Linear Regression Residual Plots (Outliers Removed)

Looking further at the residual plots in ***Figure 6.4,*** the observation that is immediately noticeable is that there is a clear grouping in the residual vs fitted value plot. Based on the grouping to either side of 0 residual as charges increase, one observation we can make is that accuracy of the model decreases with

higher charges and is heteroscedastic in nature. This suggests that perhaps linear regression may not be the most suitable form of modelling for this specific problem.

Regression Analysis with Outliers

Regression Equation

charges = -11927 + 257.2 age + 336.9 bmi + 0.0 sex_female - 128 sex_male + 0.0 children_0
+ 391 children_1 + 1636 children_2 + 964 children_3 + 2947 children_4
+ 1116 children_5 + 0.0 region_northeast - 380 region_northwest
- 1033 region_southeast - 953 region_southwest + 0.0 smoker_no + 23836 smoker_yes

Figure 6.5 Linear Regression Equation (Outliers Included)

A quick comparison to see how the outliers in the original dataset effected the regression analysis. Steps for generating this analysis were the same as the prior but with the original data set. We generated but omitted the four in one as there was no visible difference in the residual plots.

As we see in Figure 6.5 the regression equation is like Figure 6.1, with perhaps the largest change being to the coefficient applies to gender and the equation constant.

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-11927	994	-12.00	0.000	
age	257.2	11.9	21.59	0.000	1.02
bmi	336.9	28.6	11.77	0.000	1.11
sex					
male	-128	333	-0.39	0.700	1.01
children					
1	391	421	0.93	0.354	1.19
2	1636	467	3.51	0.000	1.17
3	964	548	1.76	0.079	1.13
4	2947	1239	2.38	0.018	1.03
5	1116	1456	0.77	0.444	1.03
region					
northwest	-380	477	-0.80	0.425	1.52
southeast	-1033	479	-2.16	0.031	1.66
southwest	-953	478	-1.99	0.046	1.53
smoker					
yes	23836	414	57.56	0.000	1.02

Figure 6.6 Linear Regression Coefficients (Outliers Included)

In ***Figure 6.6***, we see no noteworthy changes to VIF by comparison to the regression analysis with the outliers removed in ***Figure 6.2***.

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
6058.77	75.19%	74.97%	74.67%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	12	1.47435E+11	12286261543	334.70	0.000
age	1	17105754278	17105754278	465.99	0.000
bmi	1	5089678622	5089678622	138.65	0.000
sex	1	5442885	5442885	0.15	0.700
children	5	637996402	127599280	3.48	0.004
region	3	226406038	75468679	2.06	0.104
smoker	1	1.21607E+11	1.21607E+11	3312.76	0.000
Error	1325	48639083056	36708742		
Lack-of-Fit	1322	48377746733	36594362	0.42	0.932
Pure Error	3	261336323	87112108		
Total	1337	1.96074E+11			

Figure 6.7 Linear Regression Model Summary and ANOVA (Outliers Included)

In ***Figure 6.7,*** we see that the outliers cause a slight decrease in R-sq -0.23% by comparison to Figure 6.3 and have an impact on the P values for children though not by enough to change the statistical insignificance of the variable.
Overall, the outliers appear to have very little impact on the regression analysis. This is most likely because the outliers only made up 11 data points out of the original 1338.

Logistic Regression

Minitab Steps:

1. Loaded the cleaned dataset into Minitab.
2. Stat → Regression → Binary Logistic Regression → Fit Binary Logistic Model → Response in Binary Response → Response is Smoker → Response event is Yes → Continuous Predictors Selected → Categorical Predictors Selected → OK
3. Performed steps 1 & 2 with the outliers' dataset.

Part 1 – Without Outliers

Response Information

Variable	Value	Count	
smoker	yes	266	(Event)
	no	1061	
	Total	1327	

Figure 7.1 Logistic Regression Response Info (Outliers Removed)

Model Summary

Deviance R-Sq	Deviance R-Sq(adj)	AIC	AICc	BIC
77.91%	77.01%	319.74	320.02	387.22

Figure 7.2 Logistic Regression Model Summary (Outliers Removed)

Coefficients

Term	Coef	SE Coef	VIF
Constant	5.87	1.10	
age	-0.1052	0.0138	1.66
bmi	-0.3838	0.0484	2.44
charges	0.000407	0.000033	3.36
sex			
male	0.531	0.305	1.03
children			
1	-0.372	0.382	1.24
2	-1.337	0.440	1.32
3	-0.046	0.487	1.16
4	-1.81	1.03	1.06
5	-1.11	1.93	1.03
region			
northwest	0.189	0.408	1.49
southeast	0.599	0.429	1.49
southwest	0.317	0.450	1.44

Figure 7.3 Logistic Regression Coefficients (Outliers Removed)

Analysis of Variance

		Wald Test	
Source	DF	Chi-Square	P-Value
Regression	12	160.02	0.000
age	1	58.23	0.000
bmi	1	62.89	0.000
charges	1	152.95	0.000
sex	1	3.02	0.082
children	5	11.78	0.038
region	3	2.05	0.562

Figure 7.8 Logistic Regression Analysis of Variance (Outliers Included)

Using Minitab and logistic regression, we have been able to calculate the probability that a subject is a smoker or not.
We have been given four pieces of analysis as our stats.

1. ***Figure 7.1*** represents how many subjects are smokers in our dataset and explain which value is being considered as the event for probability calculation. In the pre-processed dataset, the number of subjects that are smokers are 266 and a total of 1026 that do not smoke.

2. Our second piece of analysis coefficients, ***Figure 7.3*** present the coefficient for each variable in the probability calculation. For example, an increase in age by one year affects the probability calculation of the subject being a smoker by -0.1052. It is the same for each variable but with different coefficients.
3. The information given in ***Figure 7.3*** presents our accuracy of the regression. We use R-sq(adj) to measure our accuracy and the result is 77.01, which can be considered as a strong value.
4. In ***Figure 7.4***, the results we can determine which variable is significant or not significant in predicting if the subject is a smoker or not. The variables age, BMI, charges, and children are significant variables as their p values are below 0.05. As region and sex have a p value that is above 0.05, they are not significant in predicting that a subject is a smoker.

Part 2 – With Outliers

Model Summary

Deviance R-Sq	Deviance R-Sq(adj)	AIC	AICc	BIC
78.33%	77.45%	319.93	320.21	387.52

Figure 7.6 Logistic Regression Model Summary (Outliers Included)

Response Information

Variable	Value	Count	
smoker	yes	274	(Event)
	no	1064	
	Total	1338	

Figure 7.5 Logistic Regression Model Summary (Outliers Included)

Coefficients

Term	Coef	SE Coef	VIF
Constant	5.82	1.09	
age	-0.1055	0.0138	1.65
bmi	-0.3816	0.0478	2.49
charges	0.000407	0.000033	3.38
sex			
male	0.534	0.305	1.04
children			
1	-0.365	0.381	1.24
2	-1.339	0.440	1.32
3	-0.047	0.487	1.16
4	-1.82	1.03	1.06
5	-1.09	1.92	1.02
region			
northwest	0.186	0.408	1.49
southeast	0.603	0.429	1.49
southwest	0.312	0.450	1.44

Figure 7.8 Logistic Regression Analysis of Variance (Outliers Included)

Analysis of Variance

		Wald Test	
Source	DF	Chi-Square	P-Value
Regression	12	160.25	0.000
age	1	58.70	0.000
bmi	1	63.71	0.000
charges	1	152.97	0.000
sex	1	3.06	0.080
children	5	11.81	0.037
region	3	2.07	0.557

Figure 7.7 Logistic Regression Coefficients (Outliers Included)

We have computed the logistic regression analysis once again in the presence of the outliers to see if there is any difference in the results. The data presented does have differences but however, the changes are extremely minor due to the size of the dataset.

1. In ***Figure 7.5,*** Comparing the response information of both the logistic regression analysis, there is an increase of 11 observations. 8 of the new subjects are smokers and 3 are non-smokers.
2. Compared to the results of the logistic regression without outliers in ***Figure 7.7***, the values of the coefficients are practically the same with minimal changes. To exemplify the interpretation of these coefficients, an increase in age by one year affects the probability calculation of the subject being a smoker by –0.1055.
3. The information represented in ***Figure 7.6*** explains the accuracy of the regression. We use R-sq(adj) to measure the accuracy and the result is 77.45 when we included outliers and 77.01 without outliers. This can be considered as a minor difference.
4. In the results of the presented in ***Figure 7.8***, there is no difference at all. Just like the logistic regression without outliers, the only significant variables are age, BMI, charges and children as their p values are less than 0.05.

Overall, there is no significant difference when comparing the results of both the logistic regressions.

Discussion and Conclusions

We examined 1338 rows of insurance data to determine the difference in insurance premium charges amongst our sample and which variables would be the most utilitarian in predicting charges, with a specific focus on charges for males compared to females, and smokers to non-smokers.

Upon completing our exploratory data analysis, it was evident that smokers, although only 20% of our sample (fig.1.5), have higher charges than their non-smoking counterparts. With smokers with a BMI of greater than 30kg/m^2 being amongst the highest charged demographic (fig.3.4). On the other hand, difference in charges between males and females were a lot more subtle, as the two groups have the same median charge (fig.2.2), and no discernible pattern to differentiate the charges between them can be seen in the scatterplots (fig.3.1 and fig.3.3).

We tested our hypotheses with and without outliers, the addition of outliers did not impact the data in any significant way. With our P-value for the male/female hypothesis test being 0.036 with outliers and 0.037 without, and 0.00 for both smoker/non-smoker tests. All our hypothesis testing confirmed with statistical significance that our theory that smokers are charged a higher premium than non-smokers, and that males are charged at a higher rate than females, was correct.

There were a few difficulties we faced in our study. One of the limitations was the chi-square goodness-of-fit test that was conducted for our smoker variable showed that our sample was not a true representation of the population and therefore may not be a relevant factor if the results from this study is applied for the whole population.

Another limitation we faced was, even though our hypothesis testing states that the difference in the mean charges between males and females are significantly different, our chi-square test for association inferred that charges and sex were actually two independent variables and didn't have an effect on each other. This was also suggested in our regression analysis, which showed sex is not a statistically significant variable, with a P-value of 0.789, and variables of age and BMI would be a better predictor of premium charges.

References

1. 2021. U.S. Health Care Coverage and Spending. 9th version. [pdf] Congressional Research Service. Available at: <https://fas.org/sgp/crs/misc/IF10830.pdf> [Accessed 25 July 2021].

2. National Centre for Biotechnology Information, 2008. More than 26,000 Americans die each year because of lack of health insurance. Available at <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2323087/ >BMJ Publishing Group Ltd 2008. [Accessed 25 July 2021].

3. Sirohi, K. (2018, December 29). Simply Explained Logistic Regression with Example in R. Medium. https://towardsdatascience.com/simply-explained-logistic-regression-with-example-in-r-b919acb1d6b3

4. Interpret the key results for Binary Logistic Regression - Minitab Express. (n.d.). (C) Minitab, LLC. All Rights Reserved. 2019. Retrieved August 18, 2021, from https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/how-to/binary-logistic-regression/interpret-the-results/key-results/

5. Figure 4.5 Source is unknown, extracted from Samuel Macintyre's notes from Introduction to Probability and Statistics Semester 1 2020. Originally a screenshot from a video of unknown origin.

6. Centers of Disease Control and Prevention (CDC). (2019). *Current Cigarette Smoking Among Adults in the United States* [Dataset]. Centers of Disease Control and Prevention (CDC).