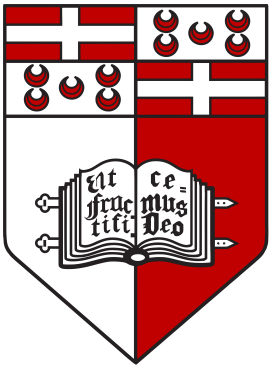

Statistics for Cyclists
-
SOR1231 Assignment

Full name: Miguel Dingli
I.D: 49997M
Course: B.Sc. (Hons) in Computing Science
Lecturer: Mark A. Caruana

Table of Contents

1 – Introduction

This chapter will introduce the data set which was created for this assignment. The data set created is based on data obtained from the 'Mountain Bikes' section of an outdoor activity website [1] which provides information about various mountain bikes from many brands. The selection of mountain bikes was limited to five mountain bike types ('All-Mountain', 'Cross-Country', 'Downhill/Freeride', 'Lifestyle', and 'Trail') and four bike manufacturers ('Yeti', 'Scott', 'Trek', and 'Orbea').

The main reason why I decided to create a data set based on mountain bikes is because of the love I have for cycling. In the process of gathering the data, I got to know more about different brands and the different components that characterize a mountain bike. I also learnt about which parts are more important to look out for when purchasing a mountain bike.

The sample consists of 45 mountain bikes and is characterized by 10 variables. This gives a detailed description of the key elements that make up each mountain bike. The ten variables that characterize the data points will be described below, starting with the fixed factors.

Fixed factors:

- **Brand**: the brand of the mountain bike, which ranges from 1 to 4, corresponding to 'Orbea', 'Trek', 'Yeti', and 'Scott'.
- **Susp**: the type of suspension that the mountain bike has, which ranges from 1 to 2, corresponding to 'Hardtail' and 'Full Suspension'. Hardtail means that the mountain bike only has front suspension, while full suspension means that it also has rear suspension.
- **BkTyp**: the type of mountain bike, which ranges from 1 to 5, corresponding to 'Lifestyle', 'Cross-Country', 'Trail', 'All-Mountain', and 'Downhill/Freeride'. These were ordered according to the terrain for which the mountain bike is designed for, with bikes of type 'Downhill/Freeride' being the ones designed for the roughest terrain.
- **FrmMat**: the material used in the mountain bike frame, which ranges from 1 to 4, corresponding to 'Aluminium', 'Carbon Fiber', 'Steel', and 'Titanium'. The frame material is the component most crucial to the weight and strength of a mountain bike.
- **WhlSz**: the size of the mountain bike's wheel, which ranges from 1 to 3, corresponding to '26', '27.5', and '29'.
- **BrkTyp**: the braking system that the mountain bike has, which ranges from 1 to 2, corresponding to 'Disk Brake' and 'V Brake'. Disk brakes offer superior braking.

Covariates:

- **Year**: the year of manufacture of the mountain bike, ranging from 2010 to 2016.
- **Price** (dependent variable): the price of the mountain bike, ranging from 350 to 11599.
- **FrtTrav**: the travel in mm (i.e. the length by which the front suspension can compress to absorb a shock) of the front suspension, ranging from 63 to 203.
- **Speed**: the amount of possible gear combinations of the mountain bike's gears, ranging from 10 to 30. More gear combinations makes a mountain bike more versatile since a cyclist can switch between gear combinations based on factors such as steepness of the road or the terrain on which the mountain bike is being used.

2 – Aims and Objectives

The analysis of the data set will give importance to the dependent variable of the data set, which is the **'Price'** variable. Since the data set consists of ten variables, this will give ample opportunity to discuss the effects that different values for particular variables have on the price of a mountain bike.

An interesting variable which will be included in various tests and statistics is the **'BkTyp'** (bike type) variable. The bike type indicates the environments and terrain that the mountain bike will be ideal for. Since the mountain bike type is a crucial characteristic of a mountain bike, this means that many variables including **'Susp'** (the suspension type), **'FrtTrav'** (the front travel) and **'Speed'** (the amount of gear combinations) will adapt to the mountain bike type, forming interesting relationships between the bike type and such variables.

It will also be interesting to explore how mountain bikes have changed throughout the years, by testing the **'Year'** variable with other variables, to show any trends in the type of bikes being manufactured by each **'Brand'**, especially when it comes to the **'Price'** of the bike.

Generally, the description of results will often take into consideration a situation where a beginner or a professional is looking to purchase a mountain bike. Such descriptions will serve as the basis for the conclusion, which will aim to specify the ideal mountain bike for a beginner, and the ideal mountain bike for a professional through the results obtained.

3 – Descriptive Statistics & Illustrations

3.1 – Descriptive Statistics

In this section, measures of location and dispersion will first be applied to covariates. Fixed factors will then be discussed afterwards. The measures will be discussed in the following order: Mean, Mode, Median, Range, Variance, Standard Deviation, Skewness, and Kurtosis, assuming all of the measures will be discussed for a particular variable.

3.1.1 – Descriptive Statistics for Covariates

Descriptive statistics will be calculated for the following covariates:

- **Price** (Purchase price of mountain bike)
- **FrtTrav** (Front suspension travel)
- **Speed** (Amount of gear combinations)

Table 1 below shows the results of the descriptive statistics for covariates. These results will be discussed one variable at a time in the following subsections.

Statistics

		Price	FrtTrav	Speed
N	Valid	45	45	45
	Missing	0	0	0
Mean		4225.04	115.24	19.80
Median		3680.00	100.00	20.00
Std. Deviation		3556.165	35.203	6.937
Variance		12646311.316	1239.280	48.118
Skewness		.533	.516	-.076
Std. Error of Skewness		.354	.354	.354
Kurtosis		-1.131	.055	-1.234
Std. Error of Kurtosis		.695	.695	.695
Range		11249	140	20

Table 1 – Descriptive statistics for covariates

The Price Variable

Starting with the measures of location for the **'Price'** variable, the **mean** suggests that the average price of a mountain bike is arguably quite high and may not be low enough for a beginner cyclist. The **median** however gives an indication that mountain bikes of a lower cost may be more common.

Moving on to the measures of dispersion, the **range** clearly shows that the price of a relatively cheap mountain bike is much less than an expensive bike which a professional cyclist is more likely to buy. This gives hope to the beginner cyclist that a mountain bike of a lower cost can be found. The high **standard deviation** suggests that bikes of a price quite lower or quite higher than the mean are available, further indicating that both cheap and expensive mountain bikes are available. The positive **skewness** continues on what the median suggested; i.e. that mountain bikes of lower cost may be more common. Finally, the negative **kurtosis** shows that the price has a platykurtic distribution, meaning that prices are not necessarily close to the mean price, which was found to be quite high. This further strengthens the idea that lower-cost bikes are available for beginners while more expensive ones are also available for cyclists willing to spend more.

The FrtTrav Variable

The descriptive statistics for the **'FrtTrav'** variable produced interesting results. Consulting a table in [2], a **mean** front travel of 115.24 millimetres indicates that front suspensions for cross-country and trail mountain bikes are more common. Thus, manufacturers are more inclined towards selling mountain bikes that are suitable for moderately rough terrain. The **median**, being close to the mean, indicates that the mean is accurate but bikes with lower suspension travel may be more common.

The **range** measure of dispersion for the front travel is very wide, meaning that front travels range from relatively low travels ideal for smoother bike rides to high suspension travels which are common on bikes designed for off-roading. The relatively low **standard deviation** hints that travels do not diverge too much from the mean front travel. The positive **skewness** suggests that bikes with lower front travel are more common. The final measure, the low **kurtosis** is saying that the front travel variable is close to a normal distribution, meaning that most mountain bikes are equipped with a decent front travel for rougher terrains.

The Speed Variable

The final covariate to be discussed is the **'Speed'** variable which is the count of gear combinations on a mountain bike. The **mean** speed of 19.80 combinations shows that the average mountain bike comes with seemingly many possible gear combinations, meaning a more versatile mountain bike. The **median** agrees with the mean since it is almost equal to it.

The **range** measure for the speed variable suggests that there are many possible amounts of gear combinations, ranging from few to much more than the mean indicated. The **standard deviation** suggests that the speed is neither too low, nor too high. The negative **skewness**, on the other hand, suggests that bikes with a higher speed are more common. The negative **kurtosis** shows that the speed variable has a platykurtic distribution, meaning that gear combinations are not necessarily close to the mean and that there are a significant amount of bikes with a lower-than-average or higher-than-average speed.

3.1.2 – Descriptive Statistics for Fixed Factors

Descriptive statistics will be calculated for the fixed factors below. For fixed factors, only the mode and frequencies will be discussed. Information will be derived from frequency tables.

- **Susp** (Suspension type)
- **BkTyp** (Type of mountain bike)
- **FrmMat** (Frame material)

Starting from the frequency table of the **'Susp'** variable, Table 2 shows that hardtail mountain bikes are as common as full-suspension mountain bikes. The **mode**, however, hints that full-suspension mountain bikes may be more popular.

Susp

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Hardtail	22	48.9	48.9	48.9
	Full Suspension	23	51.1	51.1	100.0
	Total	45	100.0	100.0	

Table 2 – Frequency table for Susp

The frequencies for the **'BkTyp'** variable in Table 3 show that cross-country mountain bikes are generally more common, followed by trail mountain bikes. This agrees with what the mean of the 'FrtTrav' variable indicated; i.e. that front suspensions suitable for cross-country or trail mountain bikes were more common. Downhill/freeride mountain bikes are, on the other hand, less common. The **mode** is the cross-country mountain bike type which indicates that cross-country mountain bikes may be more common.

BkTyp

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Lifestyle	6	13.3	13.3	13.3
	Cross-Country	17	37.8	37.8	51.1
	Trail	12	26.7	26.7	77.8
	All-Mountain	8	17.8	17.8	95.6
	Downhill/Freeride	2	4.4	4.4	100.0
	Total	45	100.0	100.0	

Table 3 – Frequency table for BkTyp

The final frequency table is that of the **'FrmMat'** variable. Table 4 shows that mountain bike frames made out of aluminium or carbon fiber equally dominate the market, with only a few instances of mountain bikes with steel or titanium frames. The **mode** is aluminium, suggesting that aluminium frames may be more common.

FrmMat

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Aluminium	21	46.7	46.7	46.7
	Carbon Fiber	20	44.4	44.4	91.1
	Steel	3	6.7	6.7	97.8
	Titanium	1	2.2	2.2	100.0
	Total	45	100.0	100.0	

Table 4 – Frequency table for FrmMat

3.2 – Graphical Forms of Representation

In this section, graphical forms of representation will be used to obtain statistical information about variables and to display visual relations between them. Diagrams will be presented in the following order: Bar Chart, Pie Chart, Histogram, Box Plot, and Scatter Diagram.

3.2.1 – Bar Chart

Figure 1 shows a bar chart based on the **'BrkTyp'** variable. This bar chart shows a great difference between the number of mountain bikes with disc brakes and those with v-brakes. The popularity of disc brakes can be due to the fact that disc brakes are generally more powerful brakes, meaning that they reduce the braking distance.

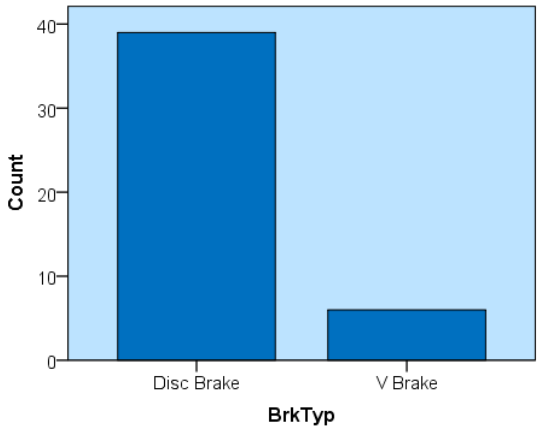

Figure 1 – Bar Chart for BrkTyp

3.2.2 – Pie Chart

Figure 2 shows a pie chart based on the **'Brand'** variable. The brand leading by a significant percentage is 'Trek'. On the other hand, 'Yeti' is the brand with the least amount of mountain bikes. Brands 'Scott' and 'Orbea' have an equal percentage and together account for half of the mountain bikes.

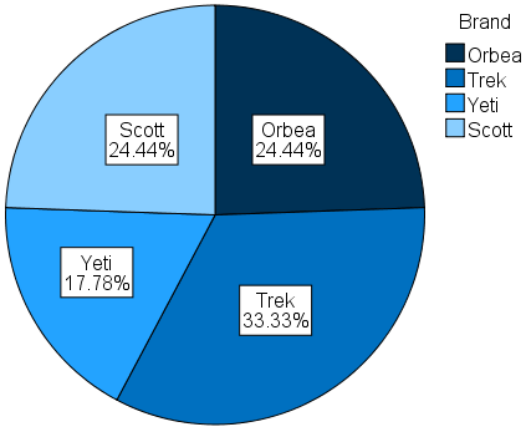

Figure 2 – Pie Chart for Brand

3.2.3 – Histogram

The **'Speed'** variable is represented by a histogram in Figure 3. The histogram was set to have 10 intervals and includes a normal curve. The first most evident element of the histogram is that a majority of counts are situated in the right half of the normal curve. From this, it is deducible that there is a trend for mountain bike with an average or higher-than-average speed. This trend is slightly balanced by the count of 14 mountain bikes with a low speed. However, the trend for higher speeds agrees with the previously-calculated descriptive statistic which showed that the `Speed' variable had a negatively-skewed distribution.

It is also clear that none of the mountain bikes has an excessively low or high speed compared to the mean, which matches with the relatively low standard deviation calculated earlier on for the 'Speed' variable. Finally, it is also worth noting how a significant amount of mountain bikes are situated away from the mean. This is why it was found earlier on (3.1.1) that the 'Speed' variable has a platykurtic distribution.

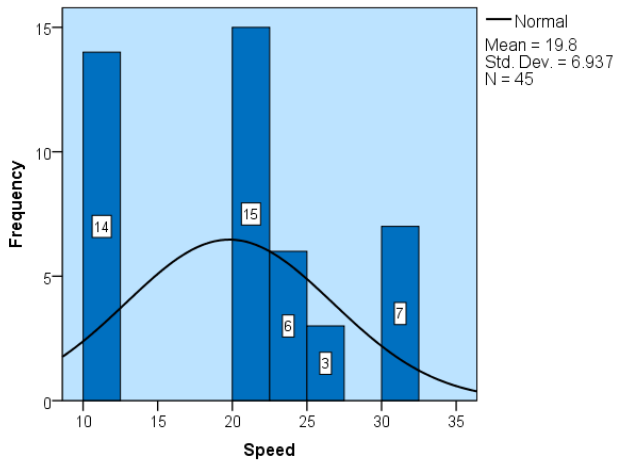

Figure 3 – Histogram for Speed

3.2.4 – Box Plot

Figure 4 is a clustered box plot made up of the **'BkTyp'** and **'FrtTrav'** variables and clustered by the **'Susp'** variable. This gives a very interesting result where relations between the front travel, mountain bike type, and suspension type are revealed.

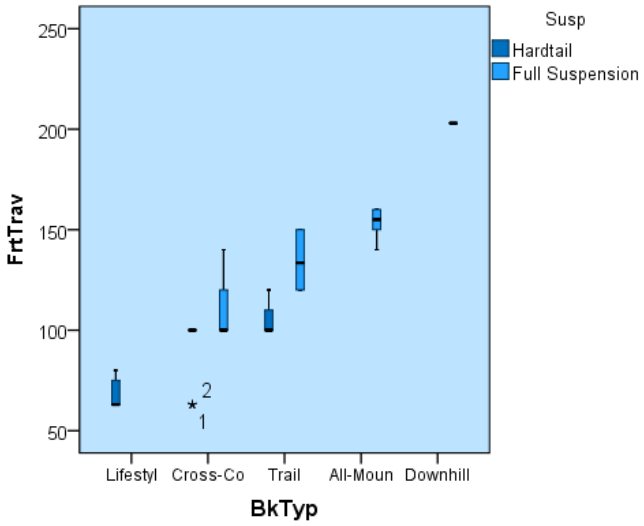

Figure 4 – Box Plot for BkTyp and FrtTrav, by Susp

The immediately noticeable characteristic is a general increase in front travels with each bike type from left to right. This increase is expected since with rougher terrains, a mountain bike needs to have more front travel to absorb greater impacts. It is also interesting that as the front travel increases, more mountain bikes are full-suspension, rather than hardtail.

Starting from the left, 'lifestyle' bikes are only hardtail. Cross-country and trail bikes both see an increase in full-suspension with the latter having higher average front travel. All-mountain and downhill/freeride mountain bikes are all full-suspension, with the latter again having a higher front travel.

3.2.5 – Scatter Plot

Figure 5 is a scatter plot which shows relations between the **'Year'** and **'Price'** variables, with markers categorized by the **'Brand'** variable. A line of best fit for each subgroup was included. This produces a timeline of the price of mountain bikes by different manufacturers.

The plot suggests that the average price of mountain bikes manufactured by 'Orbea' (from c.8500 in 2010) and 'Scott' (from c.7000 in 2010) both went down with each year. On the other hand, the average price of bikes by 'Trek' (from very low prices in earlier years) and 'Yeti' (from c.2500 in 2010) both rose with each year. This suggests that the prices may have still remained balanced due to half the brands lowering prices and the other half increasing it.

The plot also indicates that in the years 2013, 2014, and 2015, *less* mid-priced mountain bikes were manufactured, while 2016 saw a great increase in such mountain bikes. Despite this, due to the increases and decreases in prices, the lines of best fit indicate that in the years 2013, 2014, and 2015, the average mountain bike price seems to be in the mid-range (~5000).

The most expensive mountain bike was manufactured in 2015 by 'Yeti' while the cheapest mountain bike seems to have been manufactured in 2011 by 'Trek'. The latter was confirmed by going through the data set itself.

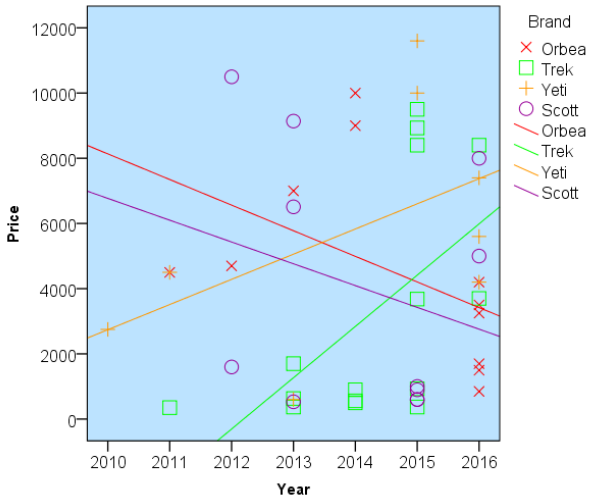

Figure 5 – Scatter Plot for Year and Price, by Brand

4 – Parametric / Non-Parametric Tests

4.1 – Checking for Normality

In this section, all of the covariates will be tested for normality using both the Kolmogorov-Smirnov test and the Shapiro-Wilk test.

- **Year**
  - H_0: **Year** is normally distributed
  - H_1: **Year** is not normally distributed
- **Price**
  - H_0: **Price** is normally distributed
  - H_1: **Price** is not normally distributed
- **FrtTrav**
  - H_0: **FrtTrav** is normally distributed
  - H_1: **FrtTrav** is not normally distributed
- **Speed**
  - H_0: **Speed** is normally distributed
  - H_1: **Speed** is not normally distributed

Table 5 below shows the output of the normality tests:

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Year	.229	45	.000	.876	45	.000
Price	.183	45	.001	.880	45	.000
FrtTrav	.201	45	.000	.923	45	.005
Speed	.209	45	.000	.871	45	.000

a. Lilliefors Significance Correction

Table 5 - Tests for Normality

Since the p-value of all of the variables for both the Kolmogorov-Smirnov test and Shapiro-Wilk test is less than the 0.05 level of significance, then all of the variables are not normally distributed (i.e. H_0 is rejected for all).

4.2 – Non-Parametric Tests

Since none of the variables is normally distributed, only non-parametric tests will be performed in the following subsections. The tests to be performed are the Kruskal-Wallis test, the Chi-Squared test, and the Pearson Correlation Matrix.

4.2.1 – Kruskal-Wallis Test

A Kruskal-Wallis test will be performed on the **'Price'** and **'WhlSz'** variables to find out whether the mean price for mountain bikes with different wheel sizes is equal or varies with different wheel sizes. This test will be performed mainly to find out whether a customer should choose a wheel size based on the budget chosen (if the price and wheel size are related) or based only on comfort (if the price and wheel size are not related). The null and alternative hypotheses can be written as follows:

- $\boldsymbol{H_0}$: All **wheel sizes** have an equal mean **price**.
- $\boldsymbol{H_1}$: Some **wheel sizes** have a greater mean **price** than others.

The output of the Kruskal-Wallis test was as follows:

Test Statistics[a,b]

	Price
Chi-Square	1.378
df	2
Asymp. Sig.	.502

a. Kruskal Wallis Test

b. Grouping Variable: WhlSz

Table 6 – Kruskal-Wallis Test on Price and WhlSz

Since the p-value of 0.502 is greater than the 0.05 level of significance, then the H_0 hypothesis can be accepted, meaning that the mean price for mountain bikes with different wheel sizes does *not* vary. This shows that a customer can choose a wheel size based on comfort, which ultimately means that the wheel size should not be given too much importance when coming up with a budget for purchasing a mountain bike.

4.2.2 – Chi-Squared Test

A Chi-Squared Test will be performed on the **'Susp'** and **'BkTyp'** variables to find out whether there is an association between these fixed factors. The main purpose of this test will be to find out whether the results deduced from the box plot generated for subsection 3.2.4 were valid results. The box plot results indicated that the type of suspension that a mountain bike is manufactured with depends on the type of mountain bike. Naturally, it makes sense that a mountain bike that is designed for rougher terrains is a full-suspension mountain bike rather than a hardtail one. The null and alternative hypotheses can be written as follows:

- $\boldsymbol{H_0}$: There is no association between **Susp** and **BkTyp**.
- $\boldsymbol{H_1}$: There is an association between **Susp** and **BkTyp**.

The output of the Chi-Squared test was as follows:

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	16.515[a]	4	.002
Likelihood Ratio	22.691	4	.000
Linear-by-Linear Association	14.497	1	.000
N of Valid Cases	45		

a. 6 cells (60.0%) have expected count less than 5. The minimum expected count is .98.

Table 7 – Chi-Square Test on Susp and BkTyp

Since the p-value is 0.002 and is smaller than the level of significance of 0.05, then the H_0 hypothesis is rejected, meaning that there *is* an association between the suspension type and the type of mountain bike. Hence, results obtained from the box plot in 3.2.4 were correct. A customer looking to purchase a mountain bike may hence choose to ignore the suspension type and focus on the mountain bike type since the bike will most likely have been manufactured with a suspension type suitable for the terrain that it was designed for.

4.2.3 – Pearson Correlation Matrix Test

The Pearson Correlation Matrix to be generated will include all the four covariates **'Year'**, **'Price**, **'FrtTrav'**, and **'Speed'** so that the results may then be used in the Regression chapter. For convenience, consider $\boldsymbol{X_i}$ and $\boldsymbol{X_j}$ refer to the $\boldsymbol{i^{th}}$ and $\boldsymbol{j^{th}}$ variables, respectively, where $\boldsymbol{i}$ and $\boldsymbol{j}$ both range from 1 (first variable) to 4 (fourth variable). The null and alternate hypotheses for each pair of variables can be written in a general form as follows:

- $\boldsymbol{H_0}$: Correlation between $\boldsymbol{X_i}$ and $\boldsymbol{X_j}$ is 0.
- $\boldsymbol{H_1}$: Correlation between $\boldsymbol{X_i}$ and $\boldsymbol{X_j}$ is significantly different from 0.

The output of the Pearson Correlation Matrix was as follows:

Correlations

		Year	Price	FrtTrav	Speed
Year	Pearson Correlation	1	.097	.151	-.460**
	Sig. (2-tailed)		.527	.321	.001
	N	45	45	45	45
Price	Pearson Correlation	.097	1	.582**	-.649**
	Sig. (2-tailed)	.527		.000	.000
	N	45	45	45	45
FrtTrav	Pearson Correlation	.151	.582**	1	-.321*
	Sig. (2-tailed)	.321	.000		.031
	N	45	45	45	45
Speed	Pearson Correlation	-.460**	-.649**	-.321*	1
	Sig. (2-tailed)	.001	.000	.031	
	N	45	45	45	45

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Table 8 – Pearson Correlation Matrix with Year, Price, FrtTrav, and Speed

From the above table, it can be observed that the correlation between four pairs of variables is strong, while that between the other variables is weak, ignoring correlations of 1 between identical variables.

The correlation between **'Speed'** and all of the remaining three variables is strong, and the correlation between **'FrtTrav'** and **'Price'** is also strong. For these pairs of values, since the p-values are smaller than the 0.05 level of significance, then H_0 is rejected, meaning that the correlation between the pairs of values is significantly different from 0. The following list points out what conclusions can be deduced from the above results;

- Since there is a strong positive correlation between **'Price'** and **'FrtTrav'**, this indicates that mountain bikes with larger front suspension travels are likely to cost more. In subsection 3.2.4, the box plot suggested that the **'FrtTrav'** variable increases with each bike type depending on factors related to terrain. Combining the results, it can be deduced that the price of mountain bikes designed for rougher terrains are likely to be higher.
- Since there is a strong negative correlation between **'Speed'** and **'Year'**, this indicates that there is a decreasing trend in the number of gear combination (speed) of a mountain bike with each consecutive year. The popularity of lower-speed bikes is increasing.
- Since there is a strong negative correlation between **'Speed'** and **'Price'**, this indicates that as the price of a mountain bike increases, the speed decreases, counter-intuitively suggesting that an expensive bike is more likely to have less gear combinations.
- Finally, since there is a strong negative correlation between **'Speed'** and **'FrtTrav'**, this indicates that mountain bikes with higher speeds will most likely have less front travel. This result agrees with the first and third correlations discussed above which related 'Price' to these two variables and which turned out positive and negative, respectively. It can be deduced that, in general, higher-costing mountain bikes are more likely to have lower speeds but also more likely to have higher front travels.

On the other hand, the correlation between **'Year'** and **'Price'** and that between **'Year'** and **'FrtTrav'** are not strong. For these pairs, since the p-values are greater than 0.05, then H_0 is accepted, meaning that the correlation between the pairs of values is 0. The following list points out what conclusions can be deduced from the above results;

- Since there is no correlation between **'Year'** and **'Price'**, then the observation in 3.2.5, which stated that the prices may have still remained balanced with consecutive years due to half the brands lowering prices and the other half increasing it, was valid. A lack of a correlation indicates that the prices of mountain bikes remained relatively the same (due to a balance in increases and decreases in prices) throughout the years.
- Since there is no correlation between **'Year'** and **'FrtTrav'**, then it can be deduced that there was no increasing or decreasing trend in the front travel of a mountain bike throughout the years. This suggests that, since front travel depends on the bike type (as observed in 3.2.4), there was also no change in trends regarding mountain bike types. This will be confirmed by generating another correlation matrix in subsection 4.2.4.

4.2.4 – Pearson Correlation Matrix for 'Year' and 'BkTyp'

The following Pearson Correlation Matrix for **'Year'** and **'BkTyp'** will be generated to confirm that there is in fact no correlation between these two variables. The null and alternative hypotheses can be written as follows:

- $\boldsymbol{H_0}$: Correlation between **Year** and **BkTyp** is 0.
- $\boldsymbol{H_1}$: Correlation between **Year** and **BkTyp** is significantly different from 0.

The output of the Pearson Correlation Matrix was as follows:

Correlations

		Year	BkTyp
Year	Pearson Correlation	1	.141
	Sig. (2-tailed)		.357
	N	45	45
BkTyp	Pearson Correlation	.141	1
	Sig. (2-tailed)	.357	
	N	45	45

Table 9 – Pearson Correlation Matrix with Year and BkTyp

Since the p-value for the pair of values is 0.357 and is larger than the 0.05 level of significance, then H_0 is accepted, meaning that there is no correlation between **'Year'** and **'BkTyp'**. This confirms the deduction stated at the end of subsection 4.2.3; i.e. there was no change in trends regarding mountain bike types throughout the range of years considered.

5 – Regression

In section 4.1, it was discovered that the covariates (importantly, the dependent variable) are not normally distributed. Despite this, Simple Linear Regression and Multiple Linear Regression will still both be done in the following sections for the purpose of including these in the assignment and for the method required to create such models to be better understood.

5.1 – Simple Linear Regression

In this section, a relationship will be formed between the dependent variable **'Price'** and the **'FrtTrav'** variable. A scatter diagram (Figure 6) with a line of best fit for these two variables is presented below:

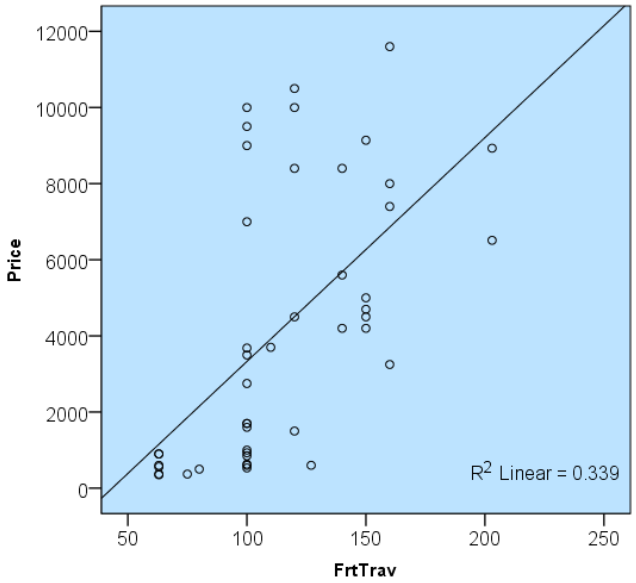

Figure 6 - Scatter Plot for Price and FrtTrav

From the scatter plot above, it is already evident that there are a lot of potential outliers since many of the data points appear to be quite far away from the line of best fit. To find the equation of the line of best fit, a set of tables will be produced.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.582[a]	.339	.324	2924.584

a. Predictors: (Constant), FrtTrav

b. Dependent Variable: Price

Table 10 – Model Summary for Simple Linear Regression

Table 10 shows a correlation R of 0.582 between the **'Price'** and **'FrtTrav'** variables (identical to the discussed result in Table 8) and an R^2 value of 0.339 which indicates that the regression line explains 33.9% of the variability of the data set. The R^2 value is closer to *0* than it is to *1*, meaning that the regression line does not fit the data set well enough.

Table 11 below shows the ANOVA table for the **'Price'** and **'FrtTrav'** variables. For analysis of the table's data, the following hypotheses need to be taken into consideration:

- $\boldsymbol{H_0}$: the model $y = b_0$ is adequate to the data set
- $\boldsymbol{H_1}$: the model $y = b_0 + b_1 x$ fits the data set better than the constant model

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	188650548.729	1	188650548.729	22.056	.000[b]
	Residual	367787149.182	43	8553189.516		
	Total	556437697.911	44			

a. Dependent Variable: Price

b. Predictors: (Constant), FrtTrav

Table 11 – ANOVA table for Simple Linear Regression

The p-value in the ANOVA table is zero, which is less than 0.05. Hence, the H_1 hypothesis is accepted, meaning that values for b_0 and b_1 now have to be found. These are obtained from the Coefficients table (Table 12) shown below:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-2553.524	1507.761		-1.694	.098
	FrtTrav	58.819	12.524	.582	4.696	.000

a. Dependent Variable: Price

Table 12 – Coefficients table for Simple Linear Regression

From the above table, it can be deduced that the values of b_0 and b_1 are -2553.524 and 58.819, respectively. Therefore, the equation of the model is:

$$\boldsymbol{y = -2553.524 + 58.819x}$$

This equation can be re-written as:

$$\mathbb{E}[\boldsymbol{Price|FrtTrav}] = \boldsymbol{-2553.524 + 58.819(FrtTrav)}$$

The created model clearly shows how the price of a mountain bike increases with the front travel. From the model, the expected price for a given particular front travel can be computed. As a result of the recent calculations, four columns were added to the data set and are shown in Table 13 as 'PRE_1', 'SRE_1', 'COO_1', and 'LEV_1', alongside **'Price'** and **'FrtTrav'**.

Price	FrtTrav	PRE_1	SRE_1	COO_1	LEV_1
2750	100	3328.38069	-.20044	.00055	.00426
350	63	1152.07580	-.28474	.00316	.05006
4499	120	4504.76171	-.00199	.00000	.00041
4500	150	6269.33324	-.61887	.00889	.02215
1599	100	3328.38069	-.59931	.00489	.00426
4699	150	6269.33324	-.54927	.00700	.02215
10499	120	4504.76171	2.07320	.04978	.00041
370	63	1152.07580	-.27764	.00300	.05006
530	100	3328.38069	-.96978	.01279	.00426
600	127	4916.49507	-1.49455	.02835	.00253
630	100	3328.38069	-.93512	.01189	.00426
1700	100	3328.38069	-.56431	.00433	.00426
6510	203	9386.74294	-1.07545	.11299	.14123
6999	100	3328.38069	1.27205	.02201	.00426
9140	150	6269.33324	1.00410	.02341	.02215
500	80	2151.99967	-.57802	.00787	.02278
560	63	1152.07580	-.21019	.00172	.05006
900	63	1152.07580	-.08949	.00031	.05006
8999	100	3328.38069	1.96515	.05253	.00426
9999	100	3328.38069	2.31170	.07269	.00426
370	75	1857.90441	-.52250	.00748	.02970
599	100	3328.38069	-.94586	.01217	.00426
599	63	1152.07580	-.19634	.00150	.05006
899	63	1152.07580	-.08984	.00031	.05006
930	100	3328.38069	-.83116	.00940	.00426
999	100	3328.38069	-.80724	.00886	.00426
3680	100	3328.38069	.12185	.00020	.00426

8400	140	5681.14273	.94561	.01548	.01124
8930	203	9386.74294	-.17075	.00285	.14123
9500	100	3328.38069	2.13877	.06222	.00426
9999	120	4504.76171	1.90027	.04182	.00041
11599	160	6857.52375	1.67126	.08749	.03673
849	100	3328.38069	-.85923	.01004	.00426
1499	120	4504.76171	-1.03959	.01252	.00041
1699	100	3328.38069	-.56466	.00434	.00426
3249	160	6857.52375	-1.27192	.05068	.03673
3499	100	3328.38069	.05913	.00005	.00426
3700	110	3916.57120	-.07491	.00007	.00050
4199	140	5681.14273	-.51549	.00460	.01124
4199	150	6269.33324	-.72416	.01218	.02215
5000	150	6269.33324	-.44398	.00458	.02215
5599	140	5681.14273	-.02857	.00001	.01124
7399	160	6857.52375	.19086	.00114	.03673
7999	160	6857.52375	.40234	.00507	.03673
8400	120	4504.76171	1.34723	.02102	.00041

Table 13 – Price, FrtTrav, and four new columns

The **'PRE_1'** column in the above table corresponds to the predicted price values. The rest of the three new columns will be discussed in the following three subsections, followed by a final subsection that will point out how the Linear Regression model can be improved.

5.1.1 – Studentized Residuals

The second column **'SRE_1'** is the Studentized Residuals column. The values in this column correspond to the *vertical* distance between the observed Price values and the predicted Price values **'PRE_1'**. It can be observed that three of the residuals (shaded in blue) exceeded ± 2. The three data points corresponding to these three residuals are considered to be outliers, since they do not agree with the model by a larger degree than the rest of the data points.

5.1.2 – Cook's Distance

The third column **'COO_1'** of the four new columns is the Cook's Distance column. The values in this column correspond to the change in the regression coefficients that occurs when

the respective observation is removed. The values show that none of the cook's distances exceeded 1, meaning that this method did not detect outliers.

5.1.3 – Leverage Values

The final column **'LEV_1'** of the four new columns is the Leverage column. The values in this column correspond to the *horizontal* distance of the expected value from the centroid. To detect observations that highly affect the outcome of the fitting regression models, values that exceed a leverage of $\frac{2(no.of\ estimated\ parameters)}{(size\ of\ the\ data\ set)} = \frac{2(2)}{45} = 0.0\dot{8}$ should be considered. It can be observed that two leverage values (shaded in blue) exceeded a leverage value of $0.0\dot{8}$. The two data points corresponding to these two leverage values can be considered to be outliers, since they do not agree with the model by a larger degree than other data points.

5.1.4 – Improving the Linear Regression model

From the above analysis, it can be deduced that five of the observations have a high chance of being outliers. Removing these observations and recreating the model would certainly produce a model which is more representative of the rest of the observations.

5.2 – Multiple Linear Regression

In this section, three out of four covariates in the data set will be considered. This includes the dependent variable **'Price'**, and the two covariates **'FrtTrav'** and **'Speed'**. A relationship will be established between the dependent variable and the two covariates.

Since the Pearson correlation matrix from subsection 4.2.3 indicated that the **'Year'** covariate is correlated with the dependent variable **'Price'**, then the **'Year'** will not be considered in this section in order to conform with the assumption that all variables must be correlated with **'Price'**. However, as stated in the introduction to the this chapter (chapter 5), the assumption that the dependent variable must be normally distributed will be broken since in section 4.1, it was found that the dependent variable **'Price'** is not normally distributed. The assumption that there cannot be correlations between the independent variables will *also* be broken since in subsection 4.2.3, it was found that the two covariates **'FrtTrav'** and **'Speed'** are significantly correlated.

Despite the fact that two assumptions were not met, Multiple Linear Regression will still be done for the purpose of including it in the assignment and for the method to be understood.

A multiple linear regression model will now be fit over the three covariates **'Price'**, **'FrtTrav'** and **'Speed'**. Table 14 below is the Model Summary generated using SPSS. It shows a correlation R of 0.759, which indicates a strong positive correlation between the three variables based on which the model will be created. It also shows an R^2 value of 0.577, meaning that 57.7% of the variability of the data is described by this model.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.759[a]	.577	.556	2368.466

a. Predictors: (Constant), Speed, FrtTrav

b. Dependent Variable: Price

Table 14 – Model Summary for Multiple Linear Regression

A second table of results (Table 15) is the ANOVA table. Before the values of this table are interpreted, the following hypothesis is stated:

- $\boldsymbol{H_0}$: model $\mathbb{E}[y] = b_0$ is adequate to the data set
- $\boldsymbol{H_1}$: model $\mathbb{E}[y|x_1, x_2] = b_0 + b_1x_1 + b_2x_2$ fits the data set better than the constant model

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	320833101.915	2	160416550.958	28.597	.000[b]
	Residual	235604595.996	42	5609633.238		
	Total	556437697.911	44			

a. Dependent Variable: Price

b. Predictors: (Constant), Speed, FrtTrav

Table 15 –ANOVA table for Multiple Linear Regression

The p-value in the ANOVA table is zero, which is less than the level of significance of 0.05. Hence, the H_1 hypothesis is accepted, meaning that values for b_0, b_1, and b_2 now have to be found. These are obtained from the Coefficients table (Table 16) shown below:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4596.415	1913.245		2.402	.021
	FrtTrav	42.111	10.711	.417	3.932	.000
	Speed	-263.861	54.357	-.515	-4.854	.000

a. Dependent Variable: Price

Table 16 – Coefficients table for Multiple Linear Regression

In Table 16, the p-values are all less than 0.05, meaning that all of the coefficients are significantly different from 0. This means that the two variables **'FrtTrav'** and **'Speed'** both influence the price significantly. This agrees with the results obtained in 4.2.3, which showed that both of these variables are correlated with the dependent variable **'Price'**.

From Table 16, it can be deduced that the values of b_0, b_1, and b_2 are 4596.415, 42.111, and -263.861, respectively. Therefore, the equation of the model is:

$$\boldsymbol{y = 4596.415 + 42.111x_1 - 263.861x_2}$$

This equation can be re-written as:

$$\mathbb{E}[\boldsymbol{Price|FrtTrav, Speed}] = \boldsymbol{4596.415 + 42.111(FrtTrav) - 263.861(Speed)}$$

The created model shows how the price of a mountain bike increases with a greater front travel, and with a smaller speed value (i.e. less gear combinations). This was already clear in subsection 4.2.3 in pairwise correlations between the three variables involved in the model.

From this model, the expected price for a specific combination of front travel and speed values can be computed. As a result of the recent calculations, four columns were added to the data set and are shown in Table 17 as 'PRE_2', 'SRE_2', 'COO_2', and 'LEV_2'.

Price	FrtTrav	Speed	PRE_2	SRE_2	COO_2	LEV_2
2750	100	20	3530.31236	-.33396	.00102	.00457
350	63	21	1708.34198	-.59619	.00955	.05240
4499	120	30	1733.92881	1.21763	.04340	.05850
4500	150	30	2997.26114	.67843	.02199	.10315
1599	100	30	891.70726	.30990	.00246	.04918
4699	150	30	2997.26114	.76827	.02820	.10315
10499	120	20	4372.53391	2.61664	.05318	.00055
370	63	21	1708.34198	-.58741	.00927	.05240
530	100	24	2474.87032	-.83461	.00767	.00977
600	127	20	4667.31146	-1.73936	.02610	.00300
630	100	24	2474.87032	-.79170	.00691	.00977
1700	100	30	891.70726	.35415	.00321	.04918
6510	203	24	6812.31132	-.14393	.00188	.19137

6999	100	20	3530.31236	1.48455	.02022	.00457
9140	150	20	5635.86624	1.51587	.03812	.02519
500	80	24	1632.64877	-.48988	.00395	.02482
560	63	21	1708.34198	-.50402	.00683	.05240
900	63	27	125.17892	.34111	.00338	.05803
8999	100	11	5905.05695	1.35949	.05119	.05449
9999	100	11	5905.05695	1.79890	.08962	.05449
370	75	21	2213.67491	-.79986	.01191	.03066
599	100	24	2474.87032	-.80500	.00714	.00977
599	63	24	916.76045	-.13932	.00051	.05048
899	63	27	125.17892	.34067	.00338	.05803
930	100	27	1683.28879	-.32579	.00174	.02474
999	100	30	891.70726	.04701	.00006	.04918
3680	100	11	5905.05695	-.97770	.02647	.05449
8400	140	11	7589.50005	.35315	.00270	.03879
8930	203	10	10506.35845	-.73184	.03733	.15071
9500	100	11	5905.05695	1.57963	.06911	.05449
9999	120	11	6747.27850	1.41658	.04321	.03846
11599	160	11	8431.72160	1.39246	.05445	.05548
849	100	30	891.70726	-.01871	.00001	.04918
1499	120	20	4372.53391	-1.22730	.01170	.00055
1699	100	20	3530.31236	-.78378	.00564	.00457
3249	160	20	6056.97702	-1.22530	.03411	.04158
3499	100	20	3530.31236	-.01340	.00000	.00457
3700	110	11	6326.16772	-1.14771	.03136	.04443
4199	140	20	5214.75546	-.43660	.00231	.01288
4199	150	10	8274.47134	-1.78892	.08623	.05257
5000	150	20	5635.86624	-.27507	.00126	.02519
5599	140	11	7589.50005	-.86729	.01629	.03879
7399	160	11	8431.72160	-.45403	.00579	.05548
7999	160	11	8431.72160	-.19024	.00102	.05548
8400	120	11	6747.27850	.71999	.01116	.03846

Table 17 – Price, FrtTrav, Speed, and four new columns

The **'PRE_2'** column in the above table corresponds to the predicted price values. The rest of the three new columns will be discussed in the following three subsections, followed by a final subsection that will point out how the Linear Regression model can be improved.

5.2.1 – Studentized Residuals

The second column **'SRE_2'** is the Studentized Residuals column. The values in this column correspond to the *vertical* distance between the observed Price values and the predicted Price values **'PRE_2'**. It can be observed that only one of the residuals (shaded in blue) exceeded ± 2. The data point corresponding to this residual is considered to be an outlier, since it does not agree with the model by a larger degree than the rest of the data points.

5.2.2 – Cook's Distance

The third column **'COO_2'** of the four new columns is the Cook's Distance column. The values in this column correspond to the change in the regression coefficients that occurs when the respective observation is removed. The values show that none of the cook's distances exceeded 1, meaning that this method did not detect outliers.

5.2.3 – Leverage Values

The final column **'LEV_2'** of the four new columns is the Leverage column. The values in this column correspond to the *horizontal* distance of the expected value from the centroid. To detect observations that highly affect the outcome of the fitting regression models, values that exceed a leverage of $\frac{2(no.of\ estimated\ parameters)}{(size\ of\ the\ data\ set)} = \frac{2(3)}{45} = 0.1\dot{3}$ should be considered. It can be observed that two leverage values (shaded in blue) exceeded a leverage value of $0.1\dot{3}$. The two data points corresponding to these two leverage values can be considered to be outliers, since they do not agree with the model by a larger degree than other data points.

5.2.4 – Improving the Linear Regression model

From the above analysis, it can be deduced that three of the observations have a high chance of being outliers. Removing these observations and recreating the model would certainly produce a model which is more representative of the rest of the observations.

6 – General Linear Models

In this chapter, two two-way ANOVA models and an ANCOVA model will be created. A three-way ANOVA model was not created due to a difficulty in obtaining three fixed factors all of which significantly affect the dependent variable.

6.1 – ANOVA Model

In previous sections, only two out of six fixed factors were tested alongside the dependent variable. The scatter plot in subsection 3.2.5 which included the dependent variable **'Price'** and fixed factor **'Brand'** suggested that different brands offer bikes of significantly different prices. The Kruskal-Wallis test performed in subsection 4.2.1 on **'Price'** and fixed factor **'WhlSz'** showed that wheel size does not significantly affect the price of a mountain bike.

In this section, two ANOVA models will be created.

6.1.1 – Two-Way ANOVA Model 1

In this subsection, an attempt will be made to create a two-way ANOVA model based on the **'Price'** dependent variable, and two fixed factors **'Susp'** and **'BkTyp'**, corresponding to the suspension type and mountain bike type. These two fixed factors were specifically picked because of results obtained in subsection 4.2.2, which indicated that there is an association between these two fixed factors. It will be interesting to analyse any interaction effects between the two related fixed factors.

For the purpose of finding out whether the suspension type and mountain bike type significantly affect the price, three pairs of hypothesis are formed:

- $\boldsymbol{H_0}$: The **type of suspension** does not significantly affect the **price**.
- $\boldsymbol{H_1}$: The **type of suspension** significantly affects the **price**.

- $\boldsymbol{H_0}$: The **type of mountain bike** does not significantly affect the **price**.
- $\boldsymbol{H_1}$: The **type of mountain bike** significantly affects the **price**.

- $\boldsymbol{H_0}$: There are no interaction effects between **suspension type** and **bike type**.
- $\boldsymbol{H_1}$: There are interaction effects between **suspension type** and **bike type**.

The above hypotheses will now be tested by generating a set of results using SPSS. The p-values for the above three pairs of hypotheses are found in the next table (Table 18).

Tests of Between-Subjects Effects

Dependent Variable: Price

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	338416164.388[a]	6	56402694.065	9.831	.000
Intercept	582744240.840	1	582744240.840	101.569	.000
Susp	155807683.060	1	155807683.060	27.156	.000
BkTyp	31302875.849	4	7825718.962	1.364	.265
Susp * BkTyp	8741659.788	1	8741659.788	1.524	.225
Error	218021533.523	38	5737408.777		
Total	1359732723.000	45			
Corrected Total	556437697.911	44			

a. R Squared = .608 (Adjusted R Squared = .546)

Table 18 – Tests of effects for Susp, BkTyp, and interaction

Starting from the p-value of **'Susp'**; this p-value is zero which is less than 0.05, meaning that H_1 is accepted. Hence, this means that the type of suspension is significantly affecting the mountain bike price and should therefore be taken into consideration when coming up with a budget when one is thinking about purchasing a mountain bike.

The p-value corresponding to the **'BkTyp'** variable is 0.265 which is greater than 0.05, meaning that H_0 is accepted. Hence, this means that the type of mountain bike does not significantly affect the mountain bike price when **'Susp'** is considered.

In subsection 4.2.2, results indicated that there is an association between **'Susp'** and **'BkTyp'**, and it was concluded that a potential buyer may choose to focus on the bike type since the bike will most likely have been manufactured with a suspension type suitable for the terrain that it was designed for. Interestingly, this means that although **'BkTyp'** does not significantly affect the mountain bike price (as was discovered above), choosing a specific bike type over another *indirectly* affects the price due to a potentially different suspension type (**'Susp'**), which in turn still significantly affects the price of the bike.

Finally, the p-value of the interaction **'Susp*BkTyp'** is 0.225 which is greater than 0.05, meaning that H_0 is accepted. Hence, this means that there are no interaction effects between the **'Susp'** and **'BkTyp'** fixed factors. For this reason, the model has to be run again without this interaction. The third pair of hypotheses will therefore be ignored.

Running the model again, the p-values for the first two hypotheses out of the three hypotheses previously stated are found in the following table (Table 19):

Tests of Between-Subjects Effects

Dependent Variable: Price

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	329674504.601[a]	5	65934900.920	11.340	.000
Intercept	454912437.510	1	454912437.510	78.238	.000
Susp	171850674.349	1	171850674.349	29.556	.000
BkTyp	28023762.466	4	7005940.617	1.205	.324
Error	226763193.310	39	5814440.854		
Total	1359732723.000	45			
Corrected Total	556437697.911	44			

a. R Squared = .592 (Adjusted R Squared = .540)

Table 19 – Tests of effects for Susp, BkTyp

Even though the interaction **'Susp*BkTyp'** was removed, the p-value of the **'BkTyp'** is still greater than 0.05, meaning that H_0 is accepted, and that the type of mountain bike does not significantly affect the mountain bike price. This variable has to be removed.

Running the model without the **'BkTyp'** fixed factor, the p-value for the first pair of hypotheses out of the three previously stated is found in the following table (Table 20):

Tests of Between-Subjects Effects

Dependent Variable: Price

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	301650742.134[a]	1	301650742.134	50.909	.000
Intercept	781174722.223	1	781174722.223	131.838	.000
Susp	301650742.134	1	301650742.134	50.909	.000
Error	254786955.777	43	5925278.041		
Total	1359732723.000	45			
Corrected Total	556437697.911	44			

a. R Squared = .542 (Adjusted R Squared = .531)

Table 20 – Tests of effects for Susp

The p-values in the above table are now all smaller than 0.05, meaning that there are no variables that do not significantly affect the price, allowing the next step to be taken. The next table to be considered is the table of parameter estimates (Table 21, on the next page).

Parameter Estimates

Dependent Variable: Price

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	6757.217	507.564	13.313	.000	5733.618	7780.817
[Susp=1]	-5179.445	725.914	-7.135	.000	-6643.390	-3715.499
[Susp=2]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

Table 21 – Table of parameter estimates

From the above table, the model can be written as follows:

$$\boldsymbol{Price = 6757.217 - 5179.445S_1}$$

…where S_1 corresponds to the 'Hardtail' suspension type. Note that S_1 is equal to 1 if the suspension type is 'Hardtail', and is otherwise equal to 0. Since the coefficient for the 'Hardtail' suspension type (Susp=1) is negative and large, while that of the 'Full Suspension' suspension type (Susp=2) is zero, then this means that predicted price values will be 5179.445 lower in value for mountain bikes with 'Hardtail' suspension type (Susp=1).

The next table (Table 22) shows four new columns which were introduced into the data set and are shown as 'PRE_3', 'SRE_3', 'COO_3', and 'LEV_3'.

Price	Susp	PRE_3	SRE_3	COO_3	LEV_3
2750	1	1577.77	.49	.01	.05
350	1	1577.77	-.52	.01	.05
4499	2	6757.22	-.95	.02	.04
4500	2	6757.22	-.95	.02	.04
1599	1	1577.77	.01	.00	.05
4699	2	6757.22	-.86	.02	.04
10499	2	6757.22	1.57	.06	.04
370	1	1577.77	-.51	.01	.05
530	1	1577.77	-.44	.00	.05
600	2	6757.22	-2.59	.15	.04
630	1	1577.77	-.40	.00	.05
1700	1	1577.77	.05	.00	.05
6510	2	6757.22	-.10	.00	.04

6999	2	6757.22	.10	.00	.04
9140	2	6757.22	1.00	.02	.04
500	1	1577.77	-.45	.00	.05
560	1	1577.77	-.43	.00	.05
900	1	1577.77	-.28	.00	.05
8999	1	1577.77	3.12	.23	.05
9999	2	6757.22	1.36	.04	.04
370	1	1577.77	-.51	.01	.05
599	1	1577.77	-.41	.00	.05
599	1	1577.77	-.41	.00	.05
899	1	1577.77	-.29	.00	.05
930	1	1577.77	-.27	.00	.05
999	1	1577.77	-.24	.00	.05
3680	1	1577.77	.88	.02	.05
8400	2	6757.22	.69	.01	.04
8930	2	6757.22	.91	.02	.04
9500	2	6757.22	1.15	.03	.04
9999	2	6757.22	1.36	.04	.04
11599	2	6757.22	2.03	.09	.04
849	1	1577.77	-.31	.00	.05
1499	1	1577.77	-.03	.00	.05
1699	1	1577.77	.05	.00	.05
3249	2	6757.22	-1.47	.05	.04
3499	2	6757.22	-1.37	.04	.04
3700	1	1577.77	.89	.02	.05
4199	2	6757.22	-1.07	.03	.04
4199	2	6757.22	-1.07	.03	.04
5000	2	6757.22	-.74	.01	.04
5599	2	6757.22	-.49	.01	.04
7399	2	6757.22	.27	.00	.04
7999	2	6757.22	.52	.01	.04
8400	2	6757.22	.69	.01	.04

Table 22 – Price, Susp, and four new columns

The **'PRE_3'** column shows predicted price values based on the ANOVA model that was created. Since there are only a constant and a single binary variable in the created model, the predicted price values are either 6757.22 (for 'Full Suspension') or 1577.77 (for 'Hardtail').

The **'SRE_3'** column is the Studentized Residuals column. Values in this column correspond to the *vertical* distance between the observed Price values and those predicted by the model. It can be observed that three of the residuals (shaded in blue) exceeded ± 2. The three data points corresponding to these three residuals are considered to be outliers, since they do not agree with the model by a larger degree than the rest of the data points.

The third column **'COO_3'** of the four new columns is the Cook's Distance column. The values in this column correspond to the change in the model coefficients that occurs when the respective observation is removed. Since no distance exceeded 1, no outliers were detected.

The final column **'LEV_3'** of the four new columns is the Leverage column. The values in this column correspond to the *horizontal* distance of the expected value from the centroid. To detect observations that highly affect the outcome of the fitting ANOVA models, values that exceed a leverage of $\frac{2(no.of\ estimated\ parameters)}{(size\ of\ the\ data\ set)} = \frac{2(p)}{n} = \frac{2(2)}{45} = 0.0\dot{8}$ should be considered. Since none of the leverage values exceeded $0.0\dot{8}$, then this method did not detect outliers.

Removing the three data points for which the Studentized residuals exceeded ± 2 would improve the model, if it were to be recreated. The new model would be more representative of the rest of the data points.

6.1.2 – Two-Way ANOVA Model 2

In this subsection, another attempt will be made to create a two-way ANOVA model. Even though the data set consists of six fixed factors, a succession of quick tests performed on the data proved that it is not easy to find a set of three fixed factors based on which a three-way ANOVA model could be created due to the p-values always exceeding 0.05. It is because of this reason that another two-way ANOVA model will be created instead.

A two-way ANOVA model will be created based on the **'Price'** dependent variable and two fixed factors **'BkTyp'** and **'BrkTyp'**, corresponding to the mountain bike type and brake type. Bike type was chosen for the previous ANOVA model as well. However, it will be interesting to observe results which may be different from those in the previous subsection.

For the purpose of finding out whether the mountain bike type and brake type significantly affect the price, three pairs of hypotheses are formed:

- $\boldsymbol{H_0}$: The **type of mountain bike** does not significantly affect the **price**.
- $\boldsymbol{H_1}$: The **type of mountain bike** significantly affects the **price**.

- $\boldsymbol{H_0}$: The **type of brake** does not significantly affect the **price**.
- $\boldsymbol{H_1}$: The **type of brake** significantly affects the **price**.

- $\boldsymbol{H_0}$: There are no interaction effects between **mountain bike type** and **brake type**.
- $\boldsymbol{H_1}$: There are interaction effects between **mountain bike type** and **brake type**.

Testing of the above pairs of hypotheses will be done by forming a set of tables using SPSS. The p-values for these three pairs of hypotheses are found in the table below (Table 23).

Tests of Between-Subjects Effects

Dependent Variable: Price

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	220716736.655[a]	6	36786122.776	4.164	.003
Intercept	226781260.981	1	226781260.981	25.669	.000
BkTyp	99092630.934	4	24773157.734	2.804	.039
BrkTyp	26832877.886	1	26832877.886	3.037	.089
BkTyp * BrkTyp	20762557.056	1	20762557.056	2.350	.134
Error	335720961.256	38	8834762.138		
Total	1359732723.000	45			
Corrected Total	556437697.911	44			

a. R Squared = .397 (Adjusted R Squared = .301)

Table 23 – Tests of effects for BkTyp, BrkTyp, and the interaction

Starting from the p-value of **'BkTyp'**; this p-value is 0.039 which is less than 0.05, meaning that H_1 is accepted. Hence, this means that the type of mountain bike is significantly affecting the mountain bike price. This result is different from the one obtained from the first table of p-values for the previous ANOVA model, for which bike type did not significantly affect the mountain bike price. This means that when brake type is considered, the mountain bike type affects the price more significantly than if suspension type was considered instead.

The p-value corresponding to **'BrkTyp'** is 0.089 which is greater than 0.05, meaning that H_0 is accepted. Hence, this means that the type of brake does not significantly affect the mountain bike price when **'BkTyp'** is also taken into consideration.

Finally, the p-value of the interaction **'BkTyp*BrkTyp'** is 0.134 which is greater than 0.05, meaning that H_0 is accepted. Hence, this means that there are no interaction effects between the **'BkTyp'** and **'BrkTyp'** fixed factors. For this reason, the model has to be run again without this interaction. The third pair of hypotheses will therefore be ignored. Running the model again, the p-values for the first two hypotheses out of the three hypotheses previously stated are found in the following table (Table 19):

Tests of Between-Subjects Effects

Dependent Variable: Price

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	199954179.599[a]	5	39990835.920	4.375	.003
Intercept	162302920.820	1	162302920.820	17.756	.000
BkTyp	101119373.688	4	25279843.422	2.766	.041
BrkTyp	42130349.347	1	42130349.347	4.609	.038
Error	356483518.312	39	9140603.034		
Total	1359732723.000	45			
Corrected Total	556437697.911	44			

a. R Squared = .359 (Adjusted R Squared = .277)

Table 24 – Tests of effects for BkTyp and BrkTyp

As a result of the interaction **'BkTyp*BrkTyp'** being removed, the p-value of **'BrkTyp'** is now smaller than 0.05, meaning that H_1 is accepted, and that when the interaction between bike type and brake type is not considered, the type of brake is significantly affecting the price of the mountain bike. The p-values are now all smaller than 0.05, meaning that there are no variables or interactions that do not significantly affect the price, allowing the next step to be taken. The next table to be considered is the parameter estimates table (Table 25).

Parameter Estimates

Dependent Variable: Price

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	4462.607	2621.522	1.702	.097	-839.922	9765.137
[BkTyp=1]	-5453.304	2582.490	-2.112	.041	-10676.883	-229.725
[BkTyp=2]	-2578.578	2275.890	-1.133	.264	-7182.000	2024.844
[BkTyp=3]	-4354.833	2309.116	-1.886	.067	-9025.462	315.795
[BkTyp=4]	-1114.375	2390.163	-.466	.644	-5948.935	3720.185
[BkTyp=5]	0[a]	.	.	.	.	.
[BrkTyp=1]	3257.393	1517.260	2.147	.038	188.445	6326.341
[BrkTyp=2]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

Table 25 – Table of parameter estimates

From Table 25, the model can be written as follows:

$$\boldsymbol{Price = 4462.607 - 5453.304M_1 - 2578.578M_2 - 4354.833M_3 - 1114.375M_4 + 3257.393B_1}$$

…where M_1, M_2, M_3, and M_4 correspond to the 'Lifestyle', 'Cross-country', 'Trail', and 'All-Mountain' mountain bike types, respectively, and where B_1 corresponds to the 'Disk Brake' brake type. Note that M_1, M_2, M_3, and M_4 are equal to 1 when the mountain bike type is of their corresponding mountain bike, and are otherwise equal to 0. On the other hand, B_1 is equal to 1 when the brake type is 'Disk Brake', and is otherwise equal to 0.

Since the coefficients of M_1, M_2, M_3, and M_4 are all negative, and the coefficient for the 'Downhill/Freeride' (BkTyp=5) mountain bike type is zero, then downhill/freeride mountain bikes will have a significantly higher predicted price value compared to other mountain bike types. According to the model, the cheapest mountain bike type is the 'Lifestyle' (BkTyp=1) bike type since it subtracts the most value from the predicted price value. Similarly, the most expensive bike type is the 'Downhill/Freeride' (BkTyp=5) bike type.

On the other hand, since the coefficient for the 'Disk Brake' (B_1) brake type is positive, while that of the 'V Brake' (BrkTyp=2) brake type is zero, then this means that predicted price values will be 3257.393 higher in value for mountain bikes with 'Disk Brake' brakes.

The next table (Table 26) shows four new columns which were introduced and shown as 'PRE_4', 'SRE_4', 'COO_4', and 'LEV_4'.

Price	BkTyp	BrkTyp	PRE_4	SRE_4	COO_4	LEV_4
2750	2	1	5141.42	-.82	.01	.07
350	2	2	1884.03	-.58	.02	.23
4499	3	1	3365.17	.39	.00	.08
4500	4	1	6605.63	-.74	.01	.13
1599	2	1	5141.42	-1.21	.02	.07
4699	4	1	6605.63	-.67	.01	.13
10499	2	1	5141.42	1.83	.04	.07
370	2	2	1884.03	-.57	.02	.23
530	2	2	1884.03	-.51	.01	.23
600	3	1	3365.17	-.96	.01	.08

630	2	1	5141.42	-1.54	.03	.07
1700	2	1	5141.42	-1.18	.02	.07
6510	5	1	7720.00	-.57	.05	.50
6999	2	1	5141.42	.64	.00	.07
9140	3	1	3365.17	2.00	.06	.08
500	1	2	-990.70	.56	.02	.23
560	1	2	-990.70	.58	.02	.23
900	1	1	2266.70	-.52	.01	.23
8999	2	1	5141.42	1.32	.02	.07
9999	2	1	5141.42	1.66	.03	.07
370	1	2	-990.70	.51	.01	.23
599	3	1	3365.17	-.96	.01	.08
599	1	1	2266.70	-.63	.02	.23
899	1	1	2266.70	-.52	.01	.23
930	2	1	5141.42	-1.44	.02	.07
999	3	1	3365.17	-.82	.01	.08
3680	2	1	5141.42	-.50	.00	.07
8400	4	1	6605.63	.63	.01	.13
8930	5	1	7720.00	.57	.05	.50
9500	2	1	5141.42	1.49	.03	.07
9999	2	1	5141.42	1.66	.03	.07
11599	4	1	6605.63	1.77	.07	.13
849	3	1	3365.17	-.87	.01	.08
1499	3	1	3365.17	-.64	.01	.08
1699	3	1	3365.17	-.58	.01	.08
3249	4	1	6605.63	-1.19	.03	.13
3499	2	1	5141.42	-.56	.00	.07
3700	3	1	3365.17	.12	.00	.08
4199	3	1	3365.17	.29	.00	.08
4199	3	1	3365.17	.29	.00	.08
5000	4	1	6605.63	-.57	.01	.13
5599	2	1	5141.42	.16	.00	.07

7399	4	1	6605.63	.28	.00	.13
7999	4	1	6605.63	.49	.01	.13
8400	3	1	3365.17	1.74	.05	.08

Table 26 – Price, BkTyp, BrkTyp, and four new columns

The **'PRE_4'** column shows predicted price values based on the model that was created.

The **'SRE_4'** column is the Studentized Residuals column. Values in this column correspond to the *vertical* distance between the observed Price values and those predicted by the model. It can be observed that none of the residuals exceeded ± 2. No outliers were detected.

The third column **'COO_4'** of the four new columns is the Cook's Distance column. The values in this column correspond to the change in the model coefficients that occurs when the respective observation is removed. Since no distance exceeded 1, no outliers were detected.

The final column **'LEV_4'** of the four new columns is the Leverage column. The values in this column correspond to the *horizontal* distance of the expected value from the centroid. To detect observations that highly affect the outcome of the fitting ANOVA models, values that exceed a leverage of $\frac{2(no.of\ estimated\ parameters)}{(size\ of\ the\ data\ set)} = \frac{2(6)}{45} = 0.2\dot{6}$ should be considered. It can be observed that two leverage values (shaded in blue) exceeded a leverage value of $0.2\dot{6}$. The two data points corresponding to these two leverage values can be considered to be outliers, since they do not agree with the model by a larger degree than other data points.

Removing the two outliers detected by the leverage distances and recreating the model would certainly improve the model since the new model will be more representative of the remaining observations.

6.2 ANCOVA Model

In this section, an ANCOVA model will be created. Attempts to create an ANCOVA model based on all of the ten variables and interactions between them prior to starting this section proved to be somewhat overwhelming due to the resulting 36 interactions, many of which were initially not given a p-value, meaning that these would have to be removed one-by-one.

Instead of creating an ANCOVA model for a very large or a very small set of main effects and interactions, it was decided that instead, an ANCOVA model will first be created using all of the variables, but based only on the main effects. Main effects with a p-value greater than 0.05 will then be removed due to not significantly affecting the dependent variable, and the pairwise interactions between the remaining variables will then be introduced.

Starting off, an attempt will be made to create an ANCOVA model based on the **'Price'** dependent variable, and all of the fixed factors (namely **'Brand'**, **'Susp'**, **'BkTyp'**, **'FrmMat'**, **'WhlSz'**, and **'BrkTyp'**) and covariates (namely **'Year'**, **'FrtTrav'**, and **'Speed'**). For the purpose of finding out whether the **main effects** for these variables significantly affect the price, nine pairs of hypothesis are formed. Since some variables will be removed, hypotheses for **interactions** will be formed further on.

- H_0: **Brand** does not affect the price
- H_1: **Brand** significantly affects the price

- H_0: **BkTyp** does not affect the price
- H_1: **BkTyp** significantly affects the price

- H_0: **WhlSz** does not affect the price
- H_1: **WhlSz** significantly affects the price

- H_0: **Year** does not affect the price
- H_1: **Year** significantly affects the price

- H_0: **Speed** does not affect the price
- H_1: **Speed** significantly affects the price

- H_0: **Susp** does not affect the price
- H_1: **Susp** significantly affects the price

- H_0: **FrmMat** does not affect the price
- H_1: **FrmMat** significantly affects the price

- H_0: **BrkTyp** does not affect the price
- H_1: **BrkTyp** significantly affects the price

- H_0: **FrtTrav** does not affect the price
- H_1: **FrtTrav** significantly affects the price

Testing of the above hypotheses will be done by generating a set of results using SPSS. The p-values for these nine pairs of hypotheses are given in the next table (Table 27).

Tests of Between-Subjects Effects

Dependent Variable: Price

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	465400929.484[a]	17	27376525.264	8.119	.000
Intercept	5337185.182	1	5337185.182	1.583	.219
Brand	8484480.906	3	2828160.302	.839	.484
Susp	2455003.069	1	2455003.069	.728	.401
BkTyp	6408936.203	4	1602234.051	.475	.754
FrmMat	10513078.725	3	3504359.575	1.039	.391
WhlSz	7308023.015	2	3654011.508	1.084	.353
BrkTyp	11005727.117	1	11005727.117	3.264	.082
Year	5269191.615	1	5269191.615	1.563	.222
FrtTrav	1322331.216	1	1322331.216	.392	.536
Speed	46127598.599	1	46127598.599	13.681	.001
Error	91036768.427	27	3371732.164		
Total	1359732723.000	45			
Corrected Total	556437697.911	44			

a. R Squared = .836 (Adjusted R Squared = .733)

Table 27 – Tests of effects for main effects 1

The two variables with the highest p-value are **'BkTyp'** and **'FrtTrav'**, meaning that these variables affect the price the least. These will be removed and the model will be recreated.

Tests of Between-Subjects Effects

Dependent Variable: Price

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	457217662.062[a]	12	38101471.838	12.288	.000
Intercept	8100546.263	1	8100546.263	2.613	.116
Brand	9644642.099	3	3214880.700	1.037	.389
Susp	14221533.254	1	14221533.254	4.587	.040
FrmMat	9993624.192	3	3331208.064	1.074	.374
WhlSz	7110262.624	2	3555131.312	1.147	.330
BrkTyp	12096211.798	1	12096211.798	3.901	.057
Year	7958304.281	1	7958304.281	2.567	.119
Speed	61623802.933	1	61623802.933	19.875	.000
Error	99220035.849	32	3100626.120		
Total	1359732723.000	45			
Corrected Total	556437697.911	44			

a. R Squared = .822 (Adjusted R Squared = .755)

Table 28 – Tests of effects for main effects 2

In Table 28, the three variables with the highest p-value are **'Brand'**, **'FrmMat'** and **'WhlSz'**, meaning that these are the variables that affect the price the least. These variables will be removed and the model will be created again.

Tests of Between-Subjects Effects

Dependent Variable: Price

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	421599939.758[a]	4	105399984.940	31.267	.000
Intercept	25127846.277	1	25127846.277	7.454	.009
Susp	74608011.567	1	74608011.567	22.133	.000
BrkTyp	22566751.915	1	22566751.915	6.694	.013
Year	24755850.623	1	24755850.623	7.344	.010
Speed	109386937.588	1	109386937.588	32.450	.000
Error	134837758.153	40	3370943.954		
Total	1359732723.000	45			
Corrected Total	556437697.911	44			

a. R Squared = .758 (Adjusted R Squared = .733)

Table 29 – Tests of effects for main effects 3

In Table 29, all p-values are less than 0.05. The ANCOVA model will now be recreated, but with the addition of the pairwise interactions between the four variables left (namely **'Susp'**, **'BrkTyp'**, **'Year'**, and **'Speed'**). The hypotheses for the interactions are given below:

- $\boldsymbol{H_0}$: There are no interaction effects between **Susp** and **BrkTyp**.
- $\boldsymbol{H_1}$: There are interaction effects between **Susp** and **BrkTyp**.

- $\boldsymbol{H_0}$: There are no interaction effects between **Susp** and **Year**.
- $\boldsymbol{H_1}$: There are interaction effects between **Susp** and **Year**.

- $\boldsymbol{H_0}$: There are no interaction effects between **Susp** and **Speed**.
- $\boldsymbol{H_1}$: There are interaction effects between **Susp** and **Speed**.

- $\boldsymbol{H_0}$: There are no interaction effects between **BrkTyp** and **Year**.
- $\boldsymbol{H_1}$: There are interaction effects between **BrkTyp** and **Year**.

- $\boldsymbol{H_0}$: There are no interaction effects between **BrkTyp** and **Speed**.
- $\boldsymbol{H_1}$: There are interaction effects between **BrkTyp** and **Speed**.

- $\boldsymbol{H_0}$: There are no interaction effects between **Year** and **Speed**.
- $\boldsymbol{H_1}$: There are interaction effects between **Year** and **Speed**.

Testing of the new hypotheses and will be done by generating a set of results using SPSS. The p-values for these six pairs of hypotheses and four pairs of main effects hypotheses are given in Table 30. Note that, from the set of main effects hypotheses listed at the start of the section, only the four pairs that are still relevant will be tested.

Tests of Between-Subjects Effects

Dependent Variable: Price

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	445297124.664[a]	9	49477458.296	15.581	.000
Intercept	7518763.168	1	7518763.168	2.368	.133
Susp	12947407.185	1	12947407.185	4.077	.051
BrkTyp	853545.824	1	853545.824	.269	.607
Year	7784208.244	1	7784208.244	2.451	.126
Speed	2571110.658	1	2571110.658	.810	.374
BrkTyp * Speed	1077787.314	1	1077787.314	.339	.564
Susp * BrkTyp	.000	0	.	.	.
BrkTyp * Year	830684.042	1	830684.042	.262	.612
Susp * Speed	8668677.460	1	8668677.460	2.730	.107
Year * Speed	2553217.923	1	2553217.923	.804	.376
Susp * Year	12877595.929	1	12877595.929	4.055	.052
Error	111140573.247	35	3175444.950		
Total	1359732723.000	45			
Corrected Total	556437697.911	44			

a. R Squared = .800 (Adjusted R Squared = .749)

Table 30 – Tests of effects for Susp, BrkTyp, Year, Speed, and interactions 1

Since the p-value of interactions in which **'BrkTyp'** is included are greater than 0.05, then H_0 is accepted for the hypotheses concerned. This means that there are no interaction effects between brake type and speed, no interaction effects between brake type and suspension type, and no interaction effects between brake type and the year. All three of these interactions will be removed and the model will be recreated.

In Table 31 which represents the newly recreated model, the p-value of the interaction between the year and speed is greater than 0.05. This means that there are no interaction effects between year and speed, and that H_0 was accepted. This interaction will be removed and the model will be recreated.

Tests of Between-Subjects Effects

Dependent Variable: Price

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	443242880.797[a]	7	63320411.542	20.698	.000
Intercept	9849843.931	1	9849843.931	3.220	.081
Susp	16595371.903	1	16595371.903	5.425	.025
BrkTyp	16322448.884	1	16322448.884	5.335	.027
Year	9776591.082	1	9776591.082	3.196	.082
Speed	2181349.344	1	2181349.344	.713	.404
Susp * Speed	9717974.819	1	9717974.819	3.177	.083
Year * Speed	2153790.737	1	2153790.737	.704	.407
Susp * Year	16513868.452	1	16513868.452	5.398	.026
Error	113194817.114	37	3059319.381		
Total	1359732723.000	45			
Corrected Total	556437697.911	44			

a. R Squared = .797 (Adjusted R Squared = .758)

Table 31 – Tests of effects for Susp, BrkTyp, Year, Speed, and interactions 2

In Table 32, the p-value for the interaction between suspension type and speed is greater than 0.05, meaning that H_0 is accepted for the corresponding pair hypotheses. This means that there are no interaction effects between suspension type and speed. This interaction will be removed and the model will be recreated once again.

Tests of Between-Subjects Effects

Dependent Variable: Price

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	441089090.069[a]	6	73514848.345	24.218	.000
Intercept	36818870.546	1	36818870.546	12.129	.001
Susp	17284245.998	1	17284245.998	5.694	.022
BrkTyp	15527818.309	1	15527818.309	5.115	.030
Year	36375432.644	1	36375432.644	11.983	.001
Speed	126173097.592	1	126173097.592	41.566	.000
Susp * Speed	11401758.163	1	11401758.163	3.756	.060
Susp * Year	17200136.444	1	17200136.444	5.666	.022
Error	115348607.842	38	3035489.680		
Total	1359732723.000	45			
Corrected Total	556437697.911	44			

a. R Squared = .793 (Adjusted R Squared = .760)

Table 32 – Tests of effects for Susp, BrkTyp, Year, Speed, and interactions 3

In Table 33, the p-value for the interaction between suspension type and year is greater than 0.05, meaning that H_0 is accepted. This interaction will be removed and the model will be recreated.

Tests of Between-Subjects Effects

Dependent Variable: Price

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	429687331.908[a]	5	85937466.382	26.442	.000
Intercept	26570099.026	1	26570099.026	8.175	.007
Susp	8131897.584	1	8131897.584	2.502	.122
BrkTyp	17417882.190	1	17417882.190	5.359	.026
Year	26158243.383	1	26158243.383	8.049	.007
Speed	116422064.271	1	116422064.271	35.822	.000
Susp * Year	8087392.149	1	8087392.149	2.488	.123
Error	126750366.003	39	3250009.385		
Total	1359732723.000	45			
Corrected Total	556437697.911	44			

a. R Squared = .772 (Adjusted R Squared = .743)

Table 33 – Tests of effects for Susp, BrkTyp, Year, Speed, and interactions 4

Since all of the p-values in Table 34 (below) are smaller than 0.05, then H_0 is rejected for all of the remaining pairs of hypotheses. This means that suspension type, brake type, year, and speed all significantly affect the price. The next step can now be taken.

Tests of Between-Subjects Effects

Dependent Variable: Price

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	421599939.758[a]	4	105399984.940	31.267	.000
Intercept	25127846.277	1	25127846.277	7.454	.009
Susp	74608011.567	1	74608011.567	22.133	.000
BrkTyp	22566751.915	1	22566751.915	6.694	.013
Year	24755850.623	1	24755850.623	7.344	.010
Speed	109386937.588	1	109386937.588	32.450	.000
Error	134837758.153	40	3370943.954		
Total	1359732723.000	45			
Corrected Total	556437697.911	44			

a. R Squared = .758 (Adjusted R Squared = .733)

Table 34 – Tests of effects for Susp, BrkTyp, Year and Speed

The table to be considered next is that of parameter estimates (Table 35, in the next page). From Table 35, the model can be written as follows:

$$\boldsymbol{Price = 1060375.133 - 3119.554A_1 + 2347.378B_1 - 521.801C - 284.074D}$$

…where A_1 corresponds to the 'Hardtail' suspension type, B_1 corresponds to the 'Disk Brake' brake type, C corresponds to the year, and D corresponds to the speed.A_1 and B_1 are equal to 1 when the suspension type and brake type are 'Hardtail' and 'Disk Brake',

respectively. There are otherwise equal to 0. On the other hand, C and D are equal to the manufacture year and speed of the mountain bike, respectively.

Parameter Estimates

Dependent Variable: Price

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	1060375.133	388061.044	2.732	.009	276074.507	1844675.759
[Susp=1]	-3119.554	663.095	-4.705	.000	-4459.719	-1779.390
[Susp=2]	0[a]	.	.	.	.	.
[BrkTyp=1]	2347.378	907.244	2.587	.013	513.769	4180.988
[BrkTyp=2]	0[a]	.	.	.	.	.
Year	-521.801	192.549	-2.710	.010	-910.957	-132.645
Speed	-284.074	49.868	-5.696	.000	-384.862	-183.286

a. This parameter is set to zero because it is redundant.

Table 35 – Table of parameter estimates

Since the coefficient of $\boldsymbol{A_1}$ is negative, this means that mountain bikes with a 'Hardtail' suspension type cost significantly less than ones with a 'Full Suspension' suspension type (i.e. Susp=2), for which the coefficient is 0.

On the other hand, the coefficient of $\boldsymbol{B_1}$ is positive, meaning that mountain bikes with a 'Disk Brake' brake type cost significantly more than ones with a 'V Brake' brake type (i.e. BrkTyp=2), for which the coefficient is 0.

Finally, both the coefficient of $\boldsymbol{C}$ and that of $\boldsymbol{D}$ are negative. For C, this means that more recently manufactured mountain bikes (i.e. for which year is greater) cost significantly less than older mountain bike models. For D, this means that mountain bikes with more gear combinations (i.e. for which speed is greater) cost significantly less than mountain bikes with less gear combinations.

The next table (Table 36) shows four new columns which were introduced and shown as 'PRE_5', 'SRE_5', 'COO_5', and 'LEV_5'.

Price	Susp	BrkTyp	Year	Speed	PRE_5	SRE_5	COO_5	LEV_5
2750	1	1	2010	20	5101.53	-1.55	.22	.31
350	1	2	2011	21	1948.28	-.99	.06	.23
4499	2	1	2011	30	4858.54	-.22	.00	.17
4500	2	1	2011	30	4858.54	-.21	.00	.17

1599	1	1	2012	30	1217.19	.22	.00	.12
4699	2	1	2012	30	4336.74	.21	.00	.14
10499	2	1	2012	20	7177.48	1.90	.07	.09
370	1	2	2013	21	904.67	-.32	.00	.17
530	1	2	2013	24	52.45	.29	.00	.17
600	2	1	2013	20	6655.68	-3.40	.15	.06
630	1	1	2013	24	2399.83	-1.00	.02	.08
1700	1	1	2013	30	695.39	.57	.01	.09
6510	2	1	2013	24	5519.38	.56	.00	.07
6999	2	1	2013	20	6655.68	.19	.00	.06
9140	2	1	2013	20	6655.68	1.40	.02	.06
500	1	2	2014	24	-469.35	.58	.01	.18
560	1	2	2014	21	382.87	.11	.00	.17
900	1	1	2014	27	1025.81	-.07	.00	.07
8999	1	1	2014	11	5570.99	2.06	.19	.18
9999	2	1	2014	11	8690.55	.74	.01	.08
370	1	2	2015	21	-138.93	.31	.00	.19
599	1	1	2015	24	1356.23	-.43	.00	.07
599	1	1	2015	24	1356.23	-.43	.00	.07
899	1	1	2015	27	504.01	.23	.00	.09
930	1	1	2015	27	504.01	.24	.00	.09
999	1	1	2015	30	-348.22	.78	.02	.12
3680	1	1	2015	11	5049.19	-.81	.02	.15
8400	2	1	2015	11	8168.75	.13	.00	.06
8930	2	1	2015	10	8452.82	.27	.00	.07
9500	2	1	2015	11	8168.75	.75	.01	.06
9999	2	1	2015	11	8168.75	1.03	.01	.06
11599	2	1	2015	11	8168.75	1.93	.05	.06
849	1	1	2016	30	-870.02	1.02	.04	.16
1499	1	1	2016	20	1970.72	-.27	.00	.08
1699	1	1	2016	20	1970.72	-.15	.00	.08
3249	2	1	2016	20	5090.28	-1.05	.02	.09

3499	2	1	2016	20	5090.28	-.91	.02	.09
3700	1	1	2016	11	4527.39	-.49	.01	.14
4199	2	1	2016	20	5090.28	-.51	.01	.09
4199	2	1	2016	10	7931.02	-2.12	.08	.08
5000	2	1	2016	20	5090.28	-.05	.00	.09
5599	2	1	2016	11	7646.95	-1.16	.02	.07
7399	2	1	2016	11	7646.95	-.14	.00	.07
7999	2	1	2016	11	7646.95	.20	.00	.07
8400	2	1	2016	11	7646.95	.43	.00	.07

Table 36 – Price, Susp, BrkTyp, Year, Speed and four new columns

The **'PRE_5'** column shows predicted price values based on the model that was created.

The **'SRE_5'** column is the Studentized Residuals column. Values in this column correspond to the *vertical* distance between the observed Price values and those predicted by the model. It can be observed that three of the residuals exceeded ± 2. The three data points corresponding to these three residuals are considered to be outliers, since they do not agree with the model by a larger degree than the rest of the data points.

The third column **'COO_5'** of the four new columns is the Cook's Distance column. The values in this column correspond to the change in the model coefficients that occurs when the respective observation is removed. Since no distance exceeded 1, no outliers were detected.

The final column **'LEV_5'** of the four new columns is the Leverage column. The values in this column correspond to the *horizontal* distance of the expected value from the centroid. To detect observations that highly affect the outcome of the fitting ANOVA models, values that exceed a leverage of $\frac{2(no.of\ estimated\ parameters)}{(size\ of\ the\ data\ set)} = \frac{2(5)}{45} = 0.\dot{2}$ should be considered. It can be observed that two leverage values (shaded in blue) exceeded a leverage value of $0.\dot{2}$. The two data points corresponding to these two leverage values can be considered to be outliers, since they do not agree with the model as much as other data points.

From the above analysis, it can be deduced that five of the observations have a high chance of being outliers. Removing these observations and recreating the model would create a model that is more representative of the remaining observations.

7 – Conclusion

This assignment, with all the results that were obtained throughout, was an interesting experience. Some tables and outcomes were more important than others.

The most important results are certainly those that have to do with the **Price** dependent variable. At the start, it was found that a mountain bike's price is on the high side, but both cheap and expensive mountain bikes are available. Further on, it was discovered that the price is directly proportional to **FrtTrav** but inversely proportional to **Speed**. These findings were reinforced by the Multiple Linear Regression model created. As for **WhlSz** and **Brand**, it was found that all wheel sizes have the same mean price and that two of the brands have rising prices while the other two have decreasing trend in prices, as years go by.

From the ANOVA models, it was found that mountain bikes with 'Hardtail' **Susp** and those 'V Brake' **BrkTyp** cost much less than ones with 'Full suspension' and with 'Disk Brake', respectively. It was also found that the most expensive mountain bikes are 'Downhill/Freeride' **BkTyp** bikes, while 'Lifestyle' mountain bikes tend to be the cheapest. In the ANCOVA model, ANOVA results and the inverse proportionality to **Speed** were reinforced, but the model also showed how more recent mountain bikes cost less than older models, even though **Price** and **Year** are not correlated, by the Pearson Correlation Matrix.

For a beginner cyclist, a typical mountain bike should not be too expensive and will likely be a 'hardtail' 'cross-country' mountain bike with 'v-brake', high 'speed', decent 'front travel', and an 'aluminium' or 'carbon fiber' frame. The brand, year of manufacture, and wheel size should not be too much of a concern.

On the other hand, for a more professional cyclist, the typical mountain bike will obviously cost much more and will be highly characterized by the bike type, but will likely be a 'full-suspension' 'all-mountain' or 'full-suspension' 'trail' mountain bike with 'disc brake', a moderate 'speed', high 'front travel', and again an 'aluminium' or 'carbon fiber' frame. Similarly, the brand, year, and wheel size should not be too important.

8 – Appendix (The Data Set)

	Brand	Year	Price	Susp	BkTyp	FrtTrav	FrmMat	Speed	WhlSz	BrkTyp
1.	3	2010	2750	1	2	100	1	20	3	1
2.	2	2011	350	1	2	63	3	21	1	2
3.	1	2011	4499	2	3	120	2	30	1	1
4.	3	2011	4500	2	4	150	1	30	1	1
5.	4	2012	1599	1	2	100	1	30	3	1
6.	1	2012	4699	2	4	150	1	30	1	1
7.	4	2012	10499	2	2	120	2	20	1	1
8.	2	2013	370	1	2	63	3	21	1	2
9.	4	2013	530	1	2	100	1	24	1	2
10.	3	2013	600	2	3	127	2	20	3	1
11.	2	2013	630	1	2	100	1	24	3	1
12.	2	2013	1700	1	2	100	1	30	3	1
13.	4	2013	6510	2	5	203	1	24	1	1
14.	1	2013	6999	2	2	100	2	20	3	1
15.	4	2013	9140	2	3	150	2	20	2	1
16.	2	2014	500	1	1	80	4	24	1	2
17.	2	2014	560	1	1	63	1	21	3	2
18.	2	2014	900	1	1	63	1	27	3	1
19.	1	2014	8999	1	2	100	2	11	3	1
20.	1	2014	9999	2	2	100	2	11	1	1
21.	2	2015	370	1	1	75	3	21	1	2
22.	4	2015	599	1	3	100	1	24	2	1
23.	4	2015	599	1	1	63	1	24	3	1
24.	4	2015	899	1	1	63	1	27	3	1
25.	2	2015	930	1	2	100	1	27	2	1
26.	4	2015	999	1	3	100	1	30	3	1
27.	2	2015	3680	1	2	100	2	11	2	1
28.	2	2015	8400	2	4	140	2	11	3	1
29.	2	2015	8930	2	5	203	2	10	2	1
30.	2	2015	9500	2	2	100	2	11	3	1

31.	3	2015	9999	2	2	120	2	11	3	1
32.	3	2015	11599	2	4	160	2	11	2	1
33.	1	2016	849	1	3	100	1	30	2	1
34.	1	2016	1499	1	3	120	1	20	2	1
35.	1	2016	1699	1	3	100	1	20	3	1
36.	1	2016	3249	2	4	160	1	20	2	1
37.	1	2016	3499	2	2	100	2	20	2	1
38.	2	2016	3700	1	3	110	1	11	3	1
39.	1	2016	4199	2	3	140	2	20	2	1
40.	3	2016	4199	2	3	150	1	10	2	1
41.	4	2016	5000	2	4	150	2	20	2	1
42.	3	2016	5599	2	2	140	2	11	3	1
43.	3	2016	7399	2	4	160	2	11	2	1
44.	4	2016	7999	2	4	160	2	11	2	1
45.	2	2016	8400	2	3	120	2	11	3	1

9 – Reference List

[1] http://mountain-bikes.gearsuite.com/.

[2] (03/06/2015). *Front Suspension Explained*. Available: https://www.rei.com/learn/expert-advice/suspension.html.

THE END