UNSW Business School/ Economics



DATA1001 Assignment 1

# Cover Sheet

Student details

|  |  |
| --- | --- |
| Surname: Ekdawi | Given name: Mina |
| Student number: 5163723 |  |
| Tutorial day/time: Tuesday 12:00pm – 1:00pm and Wednesday 4:00pm-5:00pm |  |
| Date due: 5pm Friday August 24 | Date submitted: 24th August 2018 |

Declaration

|  |  |
| --- | --- |
| I declare that this assessment item is my own work, except where acknowledged, and acknowledge that the assessor of this item may, for the purpose of assessing this item:     1. Reproduce this assessment item and provide a copy to another member of the University; and/or 2. Communicate a copy of this assessment item to a plagiarism checking service (which may then retain a copy of the assessment item on its database for the purpose of future plagiarism checking)     I certify that I have read and understood the University Rules in respect of Student Conduct. | |
| Signed: | Date: |

Mark/comments

|  |
| --- |
| Mark: |
| Comments: |

business.unsw.edu.au

Last Updated 16 January 2017 CRICOS Code 00098G

# Assignment 1

## Question 1.a.

Outliers should not be removed from the data as it may indicate trends which may otherwise go unnoticed. If an outlier is portrayed as too extreme to be believable, such as due to measurement or a typographical error, then the outliers should be included in the data set but may be excluded from the statistical analysis of the data. If a point(s) lies outside of 1.5 Interquartile Range of the mean, it is classified as an outlier(s) and it may be best to analyse the data both with and without the outlier(s).

Outliers can have a significant impact on linear regression models and the value of the mean and if it is proven to be an error, the outliers should be excluded from the analysis. A footnote should be included in the documenting of the statistical analyses to explain why the outliers were excluded from the regression or calculation of the mean. This will prevent the data from being removed without just cause, for example, to enable researchers to prove their hypothesis. However, the inclusion of outliers in statistical analysis may have little to no impact on the results such as the computation of the median, where the outlier may be included in the analysis.

Winsorizing the outliers may also be used instead of deleting them from the data set. This involves changing data points at the end of the tails of the distribution to the next highest/lowest values within the distribution that are not suspected to be outliers.

In summary, outliers should not be removed from the data set but may be excluded from analysis if it is proven that the outlier is caused by an error during the collection of the data.

Question 1.b.

The above statement is false. The sign of the co-efficient of education (𝛽1) determines the correlation between the two variables of education and electricity used by the households. The fact that highly educated members tend to consume less electricity because they have a stronger awareness of conservation and environmental concerns would suggest that as your level of education increases, your electricity usage decreases.

This suggests that there is a negative correlation between the level of education and the electricity used by the household. Hence, based on this observation, the sign of 𝛽1 would be negative.

The fact that higher education is normally associated with higher income, which in turn is associated with high electricity usage would mean that the 𝛽2 value would be positive but would not affect the 𝛽1 value. After controlling for the variable of income, the partial correlation between education and electricity is negative.

This is assuming that none of the variables are associated with conservation and environmental concerns, in which case the value of 𝛽1 is unknown.

## Question 2.a.

This survey is only taking the opinions of those who are staying at the hotel at a time and hence selection bias may affect the results obtained from this survey. If the customers’ opinions about the hotel services and pricing is different to that of the general population, then the results of the survey may be irrelevant in deciding strategies to entice new customers. For example, if the hotel is very expensive, people of high socioeconomic status currently residing in the hotel may consider the nightly rate appropriately priced whilst new potential customers may consider it expensive. Therefore, the survey may not be representative of the views of potential customers.

The survey is voluntary and hence may overrepresent those who have biased views (either positive or negative) as they would be more inclined to participate.

By only handing out a survey to the customers who are currently staying in their hotels, the hotel is likely to only receive current feedback on the services they provide and their pricing strategies. This may be biased depending on the time of year in which the survey is taken. For example, the prices that the hotel sets during holidays such as Christmas may be considerably more in comparison to other times of the year.

## Question 2.b.

The fact that the survey is not compulsory may overrepresent customers who have polarised views (either positive or negative) as they would be more inclined to fill in the online survey and therefore the survey may not accurately portray the opinions of all customers.

Similarly, to the above question, the feedback is limited to those who are currently residing in the hotel and therefore may not accurately depict the opinions of the general population. Therefore, taking measures to act on the feedback of the hotel guests may not be efficient in increasing the market share of the hotels. This reflects convenience sampling, the practice of sampling subjects that the researcher can reach easily but results in groups of the population (in this case those who have not recently stayed at the hotel) being left out.

The results may be subject to recall bias since the hotel guests may not accurately recall their hotel experience and may be unable to clearly articulate their experience at the hotel.

## Question 2.c.

This survey could be biased in that the residents in the cities served by their hotels may have not actually been to their hotel and therefore may be unable to comment on the service or prices offered by the hotel.

However, since the questions are not only directed at those staying at the hotel but the general population, it may be more representative of the opinions of potential customers as opposed to approaches taken in part a and part b.

Also, since the calls are taken from 9am to 4pm it could be possible that those who are working are unable to take the phone call and hence the survey may be biased towards the opinion of the unemployed. For example, since they do not have a job, they may think that the prices that the hotel charges may be too expensive.

It may be that the people who receive the call are very busy or uninterested in the call and hence may answer hastily, giving different responses to the survey that they may actually believe.

## Question 2.d.

The Google Review responses may overrepresent customers who have polarised views (either positive or negative) as they would be more inclined to partake in the Google Review responses since it is not compulsory and therefore may not accurately portray the opinions of all customers.

It may be possible that the staff working at the hotel may have surveyed the hotel and hence the reviews may be biased.

It may not represent the current opinions of the hotel as the Google Review responses may be outdated and therefore may represent biased views. For example, maybe the prices at the hotel were considerably higher 3 years ago which led to customers complaining about the expensive prices. This in turn may have resulted in the reduction in the prices of the hotel rates and the negative reviews that customers may have posed regarding expensive prices may no longer be relevant.

The Google Review responses are also likely to represent the views of those who have stayed at the hotel and therefore their feedback may be irrelevant regards to attracting new customers to their hotel.

## Question 3.a.

Note: The below statistics are in thousands of dollars.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min | 1Q | Median | Mean | 3Q | Max |
| 0.03 | 2.20 | 7.50 | 21.73 | 21.00 | 1700.00 |

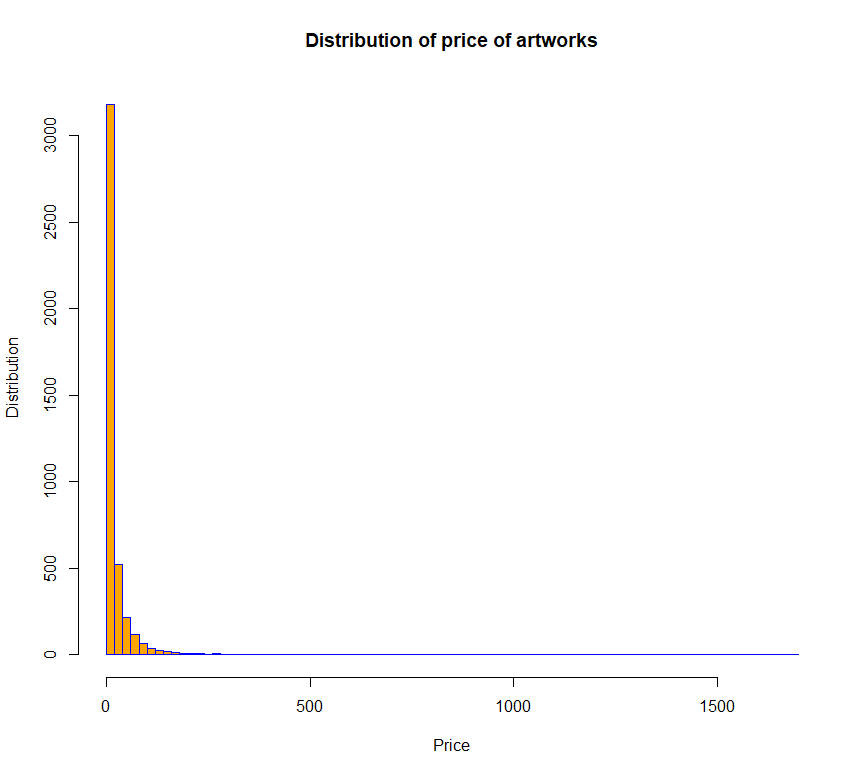
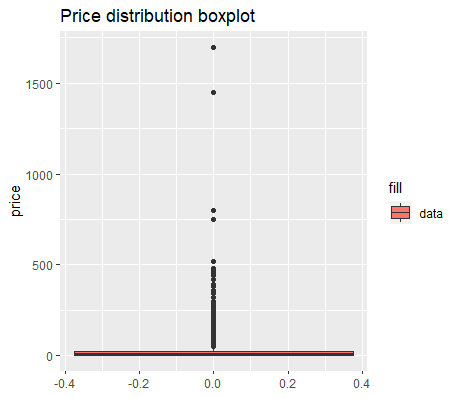
It is noted that the mean is greater than the median, suggesting that the data is positively skewed. The distribution is skewed to the right (that is most of the data points are bunched to the left with a tail of data points stretching toward the right).

The Interquartile Range is 3rd Quartile – 1st Quartile = 21.00 – 2.20 = 18.80 thousand dollars.

The Range is equal to 1700.00 – 0.03 = 1699.97 thousand dollars.

The Variance of the given data = 3039.037.

The Standard Deviation of the given data = 55.12746.

The artwork sold for 1700 thousand dollars is most likely an outlier as it lies a relatively large distance from the other results in the sample.

## Question 3.b.

Again, note that the below statistics is recorded in thousands of dollars.

## Residuals:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Min | 1Q | Median | 3Q | Max |
| -22.40 | -18.96 | -14.43 | -0.43 | 1680.66 |

## Coefficients:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate Std. | Error | t value | Pr( >|t|) |
| (Intercept) | 19.336 | 1.777 | 10.881 | <2e-16\*\*\* |
| Assignment\_1$sold | 3.093 | 2.020 | 1.531 | 0.126 |

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 005 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 55.12 on 4255 degrees of freedom

Multiple R-squared: 0.0005507, Adjusted R-squared: 0.0003158

F-statistic: 2.344 on 1 and 4255 DF, p-value: 0.1258

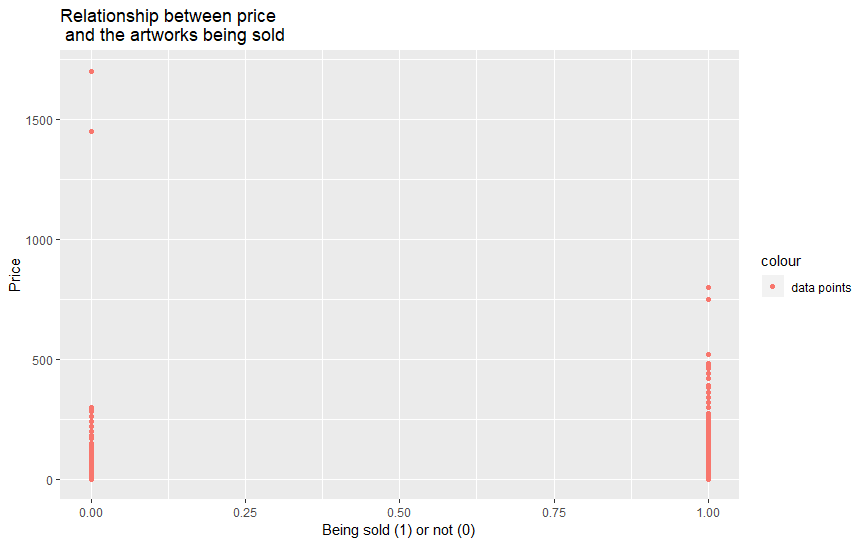
Covariance = 0.5411058

Correlation = 0.02346668

B1 (3.093 thousand) represents the average difference between the artworks which are sold and those which are not.

At first glance, the data suggests that there is a linear relationship between the price and the state of being sold or unsold. Since there is a positive relationship, denoted by the positive gradient, being sold is correlated with a higher price. However, the p-value for the gradient of the line (0.126) is greater than 0.05 and therefore the relationship between the two variables is considered statistically insignificant. Coupled with the fact that the correlation and covariance values, although positive, are relatively low is suggestive that there is little to no relationship between the two variables.

The intercept (B0) is positive, which represents the probable highest bidding price of an artwork which is not sold (19.336 thousand dollars). It seems that the artworks which have the highest bid for 1700 thousand dollars and 1450 thousand dollars are outliers as they lie a relatively large distance from the other data points.



1. Now consider the relationship between *price* and *low\_est*. Using relevant summary statistics, plots, correlation and regression write a short report describing this relationship. **(5 marks)**

## Residuals:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Min | 1Q | Median | 3Q | Max |
| -206.45 | -3.99 | -3.23 | -0.16 | 379.07 |

## Coefficients:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate Std. | Error | t value | Pr( >|t|) |
| (Intercept) | 3.635740 | 0.334146 | 10.88 | <2e-16\*\*\* |
| Assignment\_1$sold | 0.918232 | 0.005655 | 162.37 | <2e-16\*\*\* |

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.55 on 4255 degrees of freedom

Multiple R-squared: 0.861, Adjusted R-squared: 0.861

F-statistic: 2.636e+04 on 1 and 4255 DF, p-value: < 2.2e-16

Regression line: Price = 0.918232(B1) \*low\_est + 3.635740(B0) + (ui)

The graph estimates that an artwork with a zero lowest price estimate would have a selling price of 3.635740 thousand dollars (since this value refers to the y-intercept of the graph or the B0 value).

There is a positive association between the value of the lowest price estimate and the price the artwork is sold, as denoted by the positive gradient B1, which in this case is 0.918232. This means that an above-average value of the lowest price estimate is associated with an above average value of the price and below-average values also tend to occur together.

The standard errors in the value of the intercept and gradient are 0.334146 and 0.005655 and represent the variability of the data set (denoted by ui).

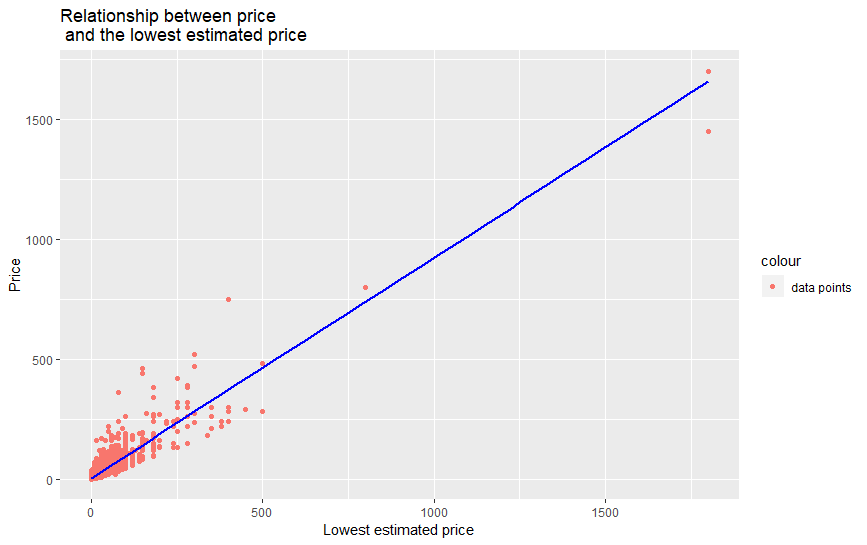
The t value is calculated by dividing the estimated values of the intercept and gradient by the standard errors for each respective value.

The p values for the intercept and gradient are less than 2\*10^(-16) and therefore the information obtained regarding the intercept and gradient are statistically significant (p-value is less than 0.05). This means that it scores three stars on the signif. codes.

Covariance = 2849.711, confirming the hypothesis that there is a positive relationship between the lowest estimated price and the price. The correlation value 0.927916 suggests that the relationship is very strong.

It appears that the data point with a price of 750 thousand dollars is an outlier as it lies furthest from the line of best fit.

The line of best fit plotted in ggplot2 demonstrates the linear relationship existing between the price and the lowest estimated price.



Appendix

//Code for importing the conart text file into the R program

assignment.data <- read.table(file=file.choose(),header=TRUE,sep="\t",dec=".")

//After executing this code, I selected the conart text file

//Code for installing packages

install.packages(“dplyr”)

//Executing the package

library(dplyr)

install.packages(“ggplot2”)

library(“ggplot2”)

a.

//This code provides summary statistics of the variable price

summary(Assignment\_1$price)

//This code is used to calculate the variance of price

var(Assignment\_1$price)

//This code is used to calculate the standard deviation of price

sd(Assignment\_1$price)

//Plotting the distribution of price

x <- hist(Assignment\_1$price,

breaks = seq(from = 0, to = 1700, by = 20), # Changes the endpoints of each bin

main = "Distribution of price of artworks",

xlab = "Price",

ylab = "Distribution",

ylim = c(0,3200),

col = "orange",

border = "blue"

)

//Plotting the boxplot

ggplot(Assignment\_1,aes(y=price, fill = 'data')) +

geom\_boxplot() + ggtitle("Price distribution boxplot")

b.

//Opening the library for ggplot2

library(ggplot2)

//Running the linear regression

|  |
| --- |
| best\_fit <- lm(Assignment\_1$price ~ Assignment\_1$sold)  summary(best\_fit)  //Calculating co-variance between the variables of price and sold  cov(Assignment\_1$price,Assignment\_1$sold)  //Calculating the correlation between price and sold  cor(Assignment\_1$price,Assignment\_1$sold) |
| //Plotting the price against sold  qplot(x = Assignment\_1$sold, y = Assignment\_1$price,  main = "Relationship between price \n and the artworks being sold",  xlab = "Being sold (1) or not (0)",  ylab = "Price",  col = 'data points',  ) |
| |  | | --- | |  | |

c.

best\_fit <- lm(price ~ sold, data = Assignment\_1)

summary(best\_fit)

//Calculating co-variance between the variables of price and lowest estimated price

cov(Assignment\_1$price,Assignment\_1$low\_est)

//Calculating the correlation between price and lowest estimated price

cor(Assignment\_1$price,Assignment\_1$low\_est)

//Plotting the scatterplot of price against lowest estimated price

qplot(x = Assignment\_1$low\_est, y = Assignment\_1$price,

main = "Relationship between price \n and the lowest estimated price",

xlab = "Lowest estimated price",

ylab = "Price",

col = 'data points',

) + geom\_smooth(method = "lm", se = FALSE, colour = 'blue')