Investigation Report on Housing Price

Title: **Relationship between Size of Family House and Its Selling Price**
Prepared by: Hui Shan Tan

INTRODUCTION

The housing market in Australia had been falling from its 2017 peak, following the tighten of credit restriction introduced by APRA. The housing price was in free fall, notably in the December quarter 2018. After almost two years of downtrend, in early of 2019, the housing market was surging as credit restriction loosening. However, Emily Cadman's article (June 2020) stated that, after an uptrend lasting for almost one year, house price fall due to pandemic Covid-19 shutdown. Theoretically, there are multiple factors influencing the housing market in a region. A research is conducted to investigate the influence of housing size to the prices in Melbourne. The locations researched is fixed to two suburbs in Melbourne.

LITERATURE REVIEW

Despite of the range of determining factors such as overall economic growth, geographic, size of houses, and demographics, the main topics to be discussed in this report will be the number of rooms (size) and location.

Location

Kiel & Zabel (2008) stated in their article, 'The immobility of houses means that their location affects their values'. They claimed that the prices are mainly affected by 3L, location, location and location. They claimed that the neighbourhood environment, such as local surrounding, district functionality, the local resident behaviours and criminal rate will devote impact on the housing market in the area.

House size/ Number of rooms

Maude Toussaint-Comeau and Jin Man Lee (2018) included the data come from the Cook County Assessor and the Northwest Illinois MLS, showing the positive impact of house size (number of rooms) for different purposes and lot size to the housing prices. Geographic reason is mentioned in the article as well.

ANALYSIS/ DISCUSSION

Sampling

The sampling methods selected to be applied in this researched are cluster sampling and simple randomization sampling, where two suburbs, Sunshine West and Dandenong on East were selected as two clusters and 10 samples was collected randomly from list of houses on sales from each suburb. However, the quality of the samples collected was questioned, whether those two suburbs can be representative of the entire housing market in Melbourne. There are 14 suburbs in Melbourne. The possibility of cluster selection bias in this survey with only two suburbs is high. Within the budget on research cost, it is suggested to randomly select more suburb clusters, then carry on with simple randomization sampling strategy in each cluster. This is expected to reduce the sampling bias and hence draw a more accurate result on the analysis.

Visualization

The rough idea of possible relationship between the explanatory/independent variables number of rooms and dependent variable, housing price through scatter plot. Figure 1, 9 out of 10 houses in East Dandenong have 8 or more rooms, and these 8 houses are priced for more than AUD 600 000. While for West Sunshine, 7 out of 10 houses have less than 8 rooms and all samples were sold at price of less than AUD 600 000. The overall pattern shown in the scatter plot is exhibiting a positive association between house sizes and selling price of houses, despite considering the geographic factor, with the possible presence of outliers on the sample. The plot indicated that the more the number of rooms, the higher the selling price.

Boxplot is opted as spreading, some statistics and distribution such as skewness of the housing price in each area can be spotted. Figure 2, the boxplot shows that house prices in West Sunshine spread around AUD 200 000 to AUD 450 000, which is much narrower than in Dandenong, around AUD 350 000 – AUD 1000 000. Outlier is present in Dandenong samples. The distribution of samples for Dandenong is symmetric, there is slight negative skewness for Sunshine samples.

Modeling

Linear regression is conducted to quantify the association between the variables.

Multiple linear regression equation:

$$\begin{aligned} P &= \beta_1 X_1 + \beta_2 X_2 + \beta_0 \\ &= 54.8981\,\mathrm{N} - 222.0446\,\mathrm{L} + 155.2885 \end{aligned}$$

P – Selling Price
N – Number of Rooms / House Size
L – Location / Suburbs (0 for Dandenong, 1 for Sunshine)

The slope, β_i for each independent variable denotes the strength and type of association with the dependent variable. The equation obtained above implies that price is positively associated with house sizes and negatively associated with location. Association of location is much stronger than the size.

For a house with 9 rooms located in east of Melbourne is predicted to be AUD 649 371 using the model.

$$H_0: \beta_i = 0 \;\; vs \;\; H_1: \beta_i \neq 0 \;\; for \; i = 1,2$$

The F-test gives F-statistic = 42.95 with p-value < 0.05. Therefore, at 5% significant level, we reject the null hypothesis, H_0. Hence, there are significant relationships between the variables in the model.

At 5% significant level, number of rooms and location resulted t-value = 5.177 with p-value < 0.05 and t-value = -4.222 with p-value < 0.05, respectively. Coefficients of independent variables are not zero.

Multiple linear regression equation with joint term:

$$P_r = \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_1 X_2 + \alpha_0$$
$$= 64.516\,\text{N} - 5.011\,\text{L} - 26.569\,\text{NL} + 61.997$$

P_r – Selling Price
N – Number of Rooms / House Size
L – Location / Suburbs (0 for Dandenong, 1 for Sunshine)

The joint term, NL has t-value = -1.221 with p-value = 0.23962 (> 0.05). This implies the coefficient of joint term is not significantly different from 0.

The most appropriate model is the second model with joint term NL. The summary of the regression models (attached in Appendix: Summary 1, Summary 2) show that the model with and without joint term, NL, have R^2 of 0.8348 and 0.8489 respectively. The higher R^2 denotes the larger proportion of dependent variable (prices) variance is explained by the model. With higher adjusted R^2 the additional joint term improves the model more than expected by chance. Comparing the residuals plots (Figure 3 and 5) and normal Q-Q plots (Figure 4 and 6), the residuals for second model scattered more randomly and fit the normality assumption much better.

CONCLUSION

With the set of samples, the analysis and modelling done deduce that model with joint term performs slightly better. The individual independent variables of geographic factor (suburbs) and the joint term with p-value > 0.05 indicate that they do not imply significant influence on the prices in these two suburbs with the possible high similarity between them. Therefore, it can be concluded, the larger the house, the higher the selling price. But, the validity of the result is expected to be improved, by applying a better sampling method and collecting a larger sample set.

REFERENCE

1. Sharpe, Norean R., et al. Business Statistics, Global Edition, Pearson Education Limited, 2015. ProQuest Ebook Central, https://ebookcentral.proquest.com/lib/unsw/detail.action?docID=5176644.
2. Zabel, Jeffrey & Kiel, Katherine. (2008). Location, location, location: The 3L Approach to house price determination. Journal of Housing Economics. 17. 175-190. 10.1016/j.jhe.2007.12.002.
3. Jin Man Lee & Maude Toussaint-Comeau, 2018. "Determinants of Housing Values and Variations in Home Prices Across Neighborhoods in Cook County," Profitwise, Federal Reserve Bank of Chicago, issue 1, pages 1-23.
4. Dirk Wittowsky, Josje Hoekveld, Janina Welsch & Michael Steier (2020) Residential housing prices: impact of housing characteristics, accessibility and neighbouring apartments – a case study of Dortmund, Germany, Urban, Planning and Transport Research, 8:1, 44-70, DOI: 10.1080/21650020.2019.1704429
5. Emily Cadman (2020). Are you a robot? Article. Bloomberg. [ONLINE] Available at: https://www.bloomberg.com/news/articles/2020-06-01/australia-house-prices-fall-as-shutdowns-hit-property-market. [Accessed 05 June 2020].

APPENDIX

Housing Prices

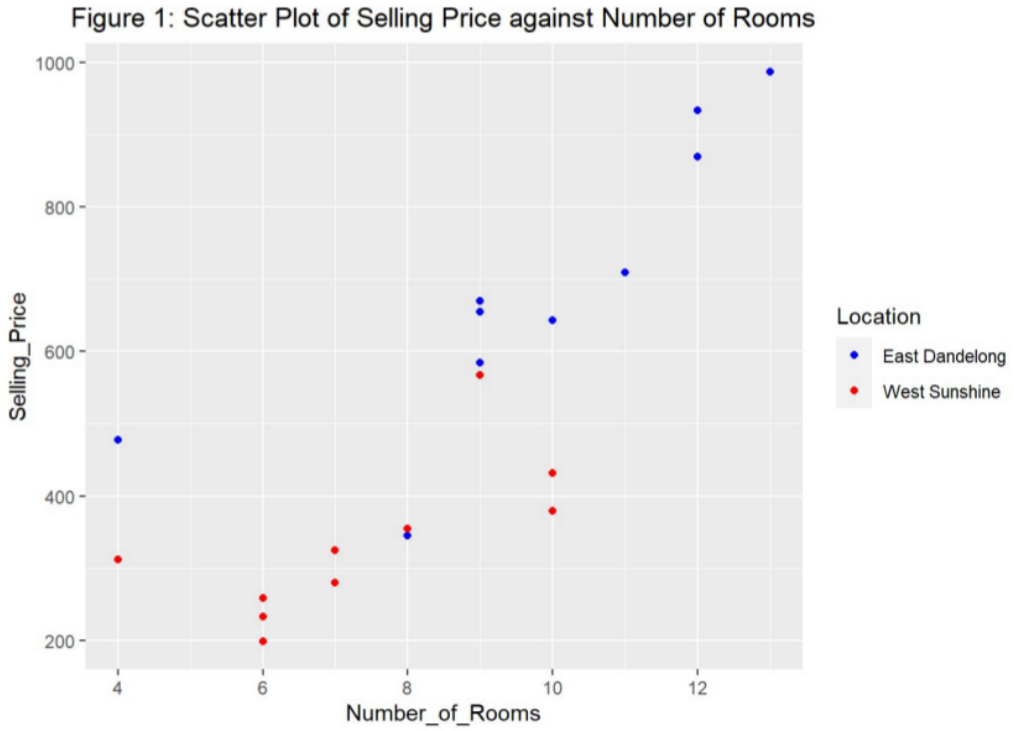
Visualization

Scatter Plot

```
#install.packages("readxl")
#install.packages("ggplot2")

library("readxl")
library('ggplot2')
data <- read_excel('C:/Users/Hshan/Desktop/Housing prices_revised.xlsx')
names(data) <- gsub(" ", "_",names(data))

ggplot(data, aes(x = Number_of_Rooms, y = Selling_Price)) +
    geom_point(aes(color = factor(Location))) +
    theme(plot.title = element_text(hjust = 0.5)) +
    labs(title = "Figure 1: Scatter Plot of Selling Price against Number of Rooms",
color = "Location") +
    scale_color_manual(labels = c("East Dandelong ", "West Sunshine"), values = c
("blue", "red"))
```



Boxplot

```
fill <- '#12B7F7' #hex code for color
ggplot(data, aes(x = as.factor(Location), y = Selling_Price)) +
    geom_boxplot(fill = fill) +
    ggtitle('Figure 2: Boxplot of Selling Price by Location' )  +
    theme(plot.title = element_text(hjust = 0.5)) +
    scale_x_discrete(name = 'Location',labels=c('0' = 'East Dandelong', '1' = 'West
Sunshine'))
```

Modeling

Multiple Linear Regression

```
# data$Location_factor = factor(data$Location)

model <- lm(Selling_Price~Number_of_Rooms+ Location, data=data); model
```

```
##
## Call:
## lm(formula = Selling_Price ~ Number_of_Rooms + Location, data = data)
##
## Coefficients:
##     (Intercept)  Number_of_Rooms         Location
##           155.3             54.9           -222.0
```

```
paste('Selling_Price = ', model$coefficients['Number_of_Rooms'], '*Number of Rooms
+ ',
      model$coefficients['Location'], '*Location + ', model$coefficients['(Intercep
t)'], sep='')
```

```
## [1] "Selling_Price = 54.8980891719746*Number of Rooms + -222.044585987261*Locati
on + 155.288535031847"
```

```
summary(model)
```

```
##
## Call:
## lm(formula = Selling_Price ~ Number_of_Rooms + Location, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -249.47  -52.99  -11.03   67.73  159.16
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       155.29     107.88   1.439 0.168191
## Number_of_Rooms    54.90      10.60   5.177 7.58e-05 ***
## Location         -222.04      52.59  -4.222 0.000574 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 102.9 on 17 degrees of freedom
## Multiple R-squared:  0.8348, Adjusted R-squared:  0.8153
## F-statistic: 42.95 on 2 and 17 DF,  p-value: 2.257e-07
```

```
predict_data <- data.frame(Location= 0, Number_of_Rooms = 9)
prediction <- predict(model,newdata = predict_data)
paste('Selling price for house with 9 rooms in east of Melbourne is ', prediction,
sep='')
```

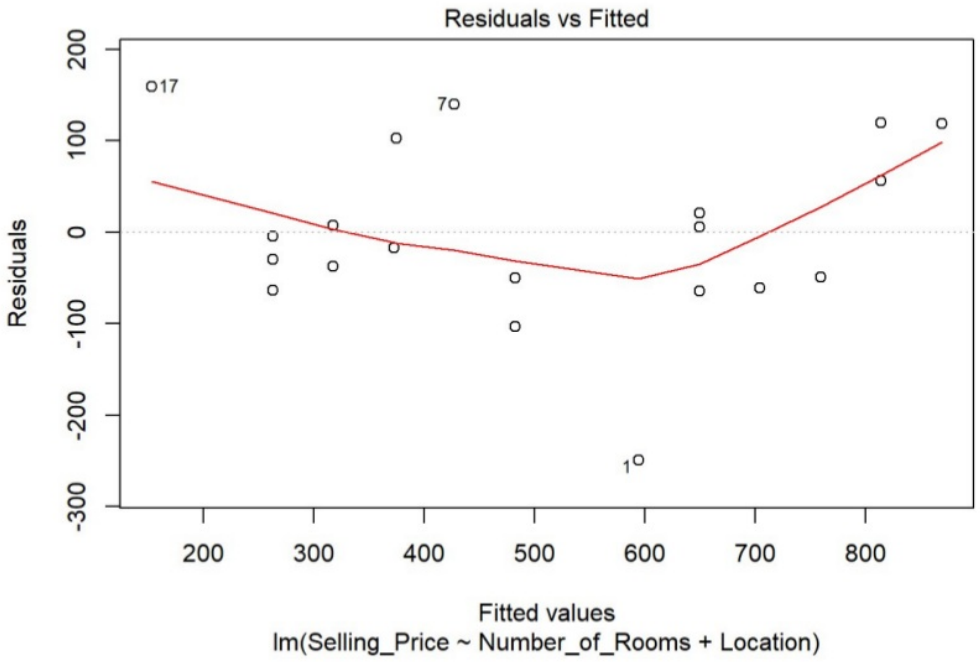
```
## [1] "Selling price for house with 9 rooms in east of Melbourne is 649.3713375796
18"
```

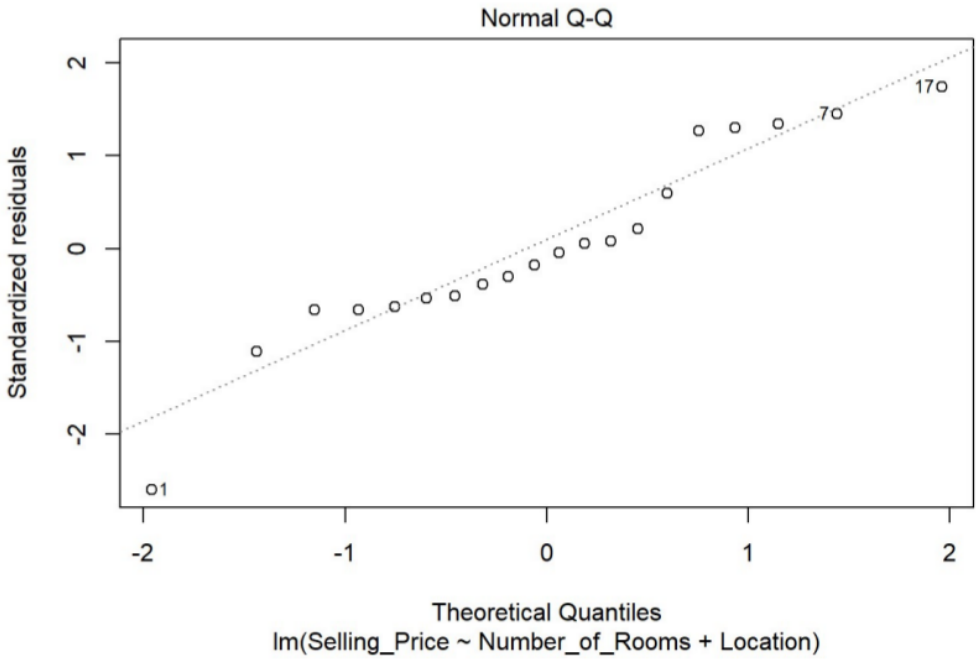
Frist Model without Interaction

Figure 3: Residuals vs Fitted

Figure 4: Normal Q-Q Plot

```
plot(model,which=c(1,2))
```





Multiple Linear Regression with Interaction Term

```
model_interaction <- lm(`Selling_Price`~`Number_of_Rooms`*Location, data=data); mod
el_interaction
```

```
##
## Call:
## lm(formula = Selling_Price ~ Number_of_Rooms * Location, data = data)
##
## Coefficients:
##              (Intercept)          Number_of_Rooms                  Location
##                   61.997                   64.516                   -5.011
## Number_of_Rooms:Location
##                  -26.569
```

```
paste('Selling_Price = ',
      model_interaction$coefficients['Number_of_Rooms'], '*Number of Rooms + (',
      model_interaction$coefficients['Location'], '*Location) + (',
      model_interaction$coefficients['Number_of_Rooms:Location'],'*Number of Rooms
*Location) + ',
      model_interaction$coefficients['(Intercept)'], sep='')
```

```
## [1] "Selling_Price = 64.5158069883528*Number of Rooms + (-5.01133496957667*Locat
ion) + (-26.5685929121064*Number of Rooms *Location) + 61.9966722129781"
```

```
summary(model_interaction)
```

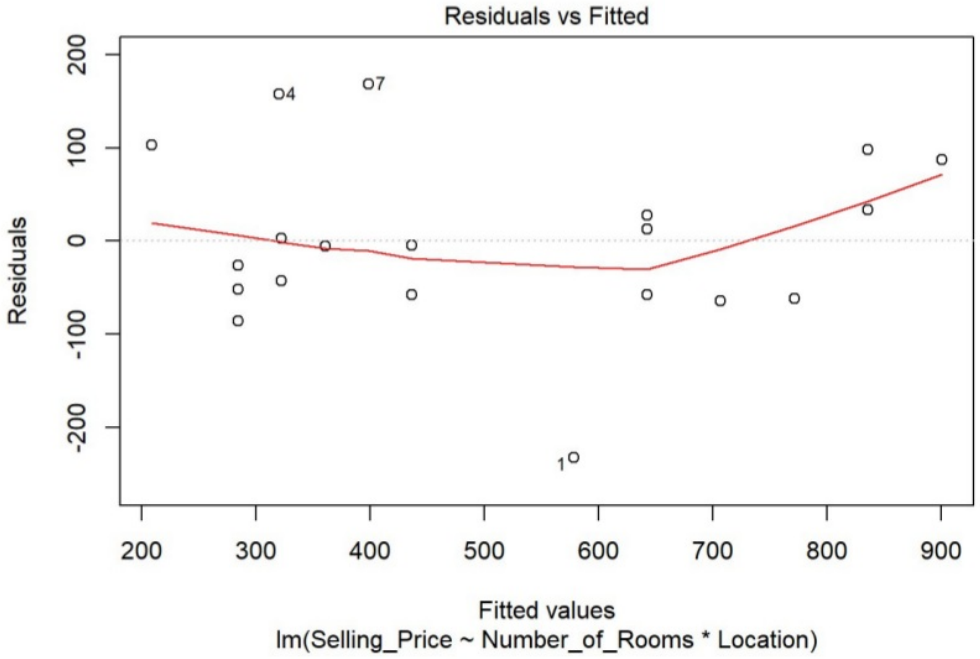
```
##
## Call:
## lm(formula = Selling_Price ~ Number_of_Rooms * Location, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -233.12  -57.50   -5.01   47.19  168.49
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 61.997    130.940   0.473 0.642271
## Number_of_Rooms             64.516     13.087   4.930 0.000151 ***
## Location                    -5.011    185.099  -0.027 0.978736
## Number_of_Rooms:Location   -26.569     21.752  -1.221 0.239620
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 101.5 on 16 degrees of freedom
## Multiple R-squared:  0.8489, Adjusted R-squared:  0.8205
## F-statistic: 29.96 on 3 and 16 DF,  p-value: 8.452e-07
```

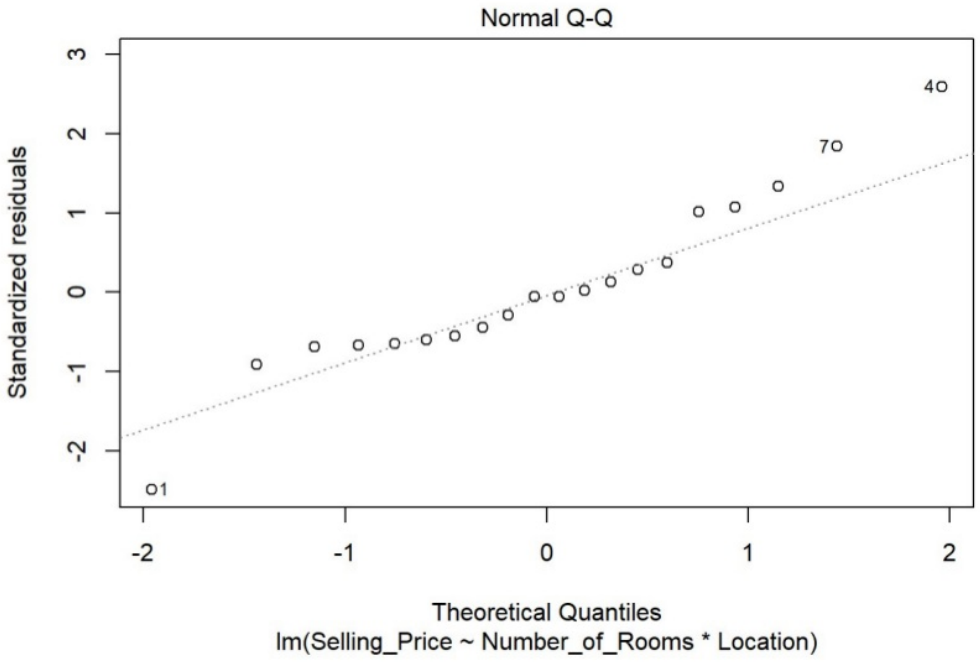
Second Model with Interaction

Figure 5: Residuals vs Fitted

Figure 6: Normal Q-Q Plot

```
plot(model_interaction,which=c(1,2))
```


Residuals vs Fitted
Residuals
Fitted values
lm(Selling_Price ~ Number_of_Rooms * Location)
4
7
1


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Selling_Price ~ Number_of_Rooms * Location)
4
7
1

Report

GRADEMARK REPORT

FINAL GRADE

90/100

GENERAL COMMENTS

Instructor

Great understanding about statistic concepts. Well motivated introduction and comprehensive literature review which is very catchy. Good utilisation of data display. Research questions are addressed within context supported by data process skills. Well organised structure and articulated analysis. Conclusion part may require further beefing up (summarise the key findings and point out future research direction). A comprehensive reference list is provided. Overall a great attempt.

Edward

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

RUBRIC: ZZSC9001 A2 MARKING RUBRIC

PRESENTATION	
EXCEEDS EXPECTATIONS (D/	Excellent understanding and professional presentation of data and graphs. Ideas expressed effectively: well labeled graphs and clear instruction on key statistics. Evidence of thorough editing: consistent font and size, uniform reference style with negligible errors.
MEETS EXPECTATIONS (P/CR)	Generally good understanding and professional presentation of key data and graphs, although could be more effective. An attempt at professional presentation and editing: sufficient instruction with tolerable mistakes.
BELOW EXPECTATIONS (F)	Absent of/ unprofessional presentation of data/graphs, e.g., missing/misunderstanding key aspects of the key statistics. Little evidence of editing (e.g., No labels on the graphs, absence of instructions on important data, inconsistent font and size). Verbose, expressed in more words than are needed.

STRUCTURE OF	
EXCEEDS EXPECTATIONS (D/	Ideas developed logically and coherently. Clear focus; no irrelevant material. Well structured.
MEETS EXPECTATIONS (P/CR)	Sufficiently clear focus. Mostly logical sequence of ideas. Adequately structured.
BELOW EXPECTATIONS (F)	Difficult to follow sequence of ideas; unclear focus. Text not clearly structured, e.g. paragraphs not clearly developed.

CRITICAL THIN	
EXCEEDS EXPECTATIONS (D/	Excellent understanding of key research topic. Insightful analysis and critical evaluation of key data, using appropriate statistical methods. Great ability to address research questions and draw conclusion with supporting analysis.
MEETS EXPECTATIONS (P/CR)	Identifies and defines key research topic with good depth and coverage. Appropriate usage of data and statistical methods within the context. Some analysis of key data and application of statistical knowledge. Exhibit an ability to draw conclusion with supporting analysis.
BELOW EXPECTATIONS (F)	Does not clearly/correctly identify or understand the research topic. Little analysis or critical evaluation of ideas or information. Limited usage of data and statistical methods. Struggles to appropriately analyse data.