

44008 STATS 201/8 Assignment 3

your name

2022-08-20

Contents

Question 1	1
Question 2	8

Question 1

Question of interest/goal of the study

We are interested in the typical sale prices for 3-bedroom and 4-bedroom houses in Auckland. In particular, we want to see whether buying a 4- bedroom house increases the cost by more than 15% as compared to buying a 3-bedroom house.

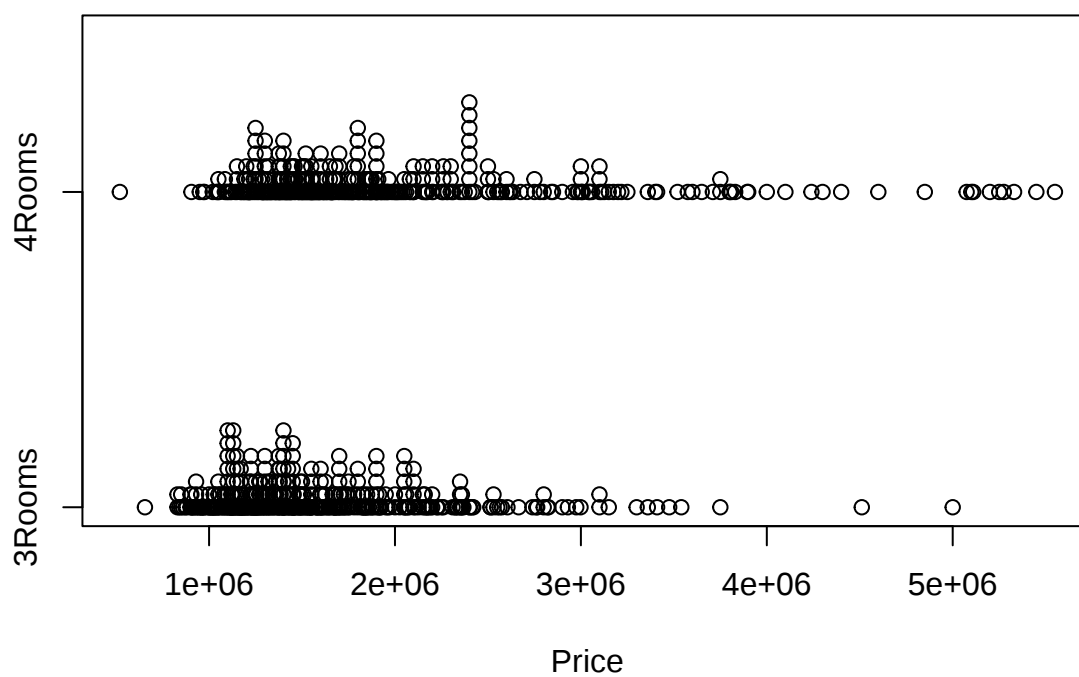
Read the data.

```
house.df=read.csv(file='house.csv', header = TRUE, stringsAsFactors = TRUE)
```

Inspect the data:

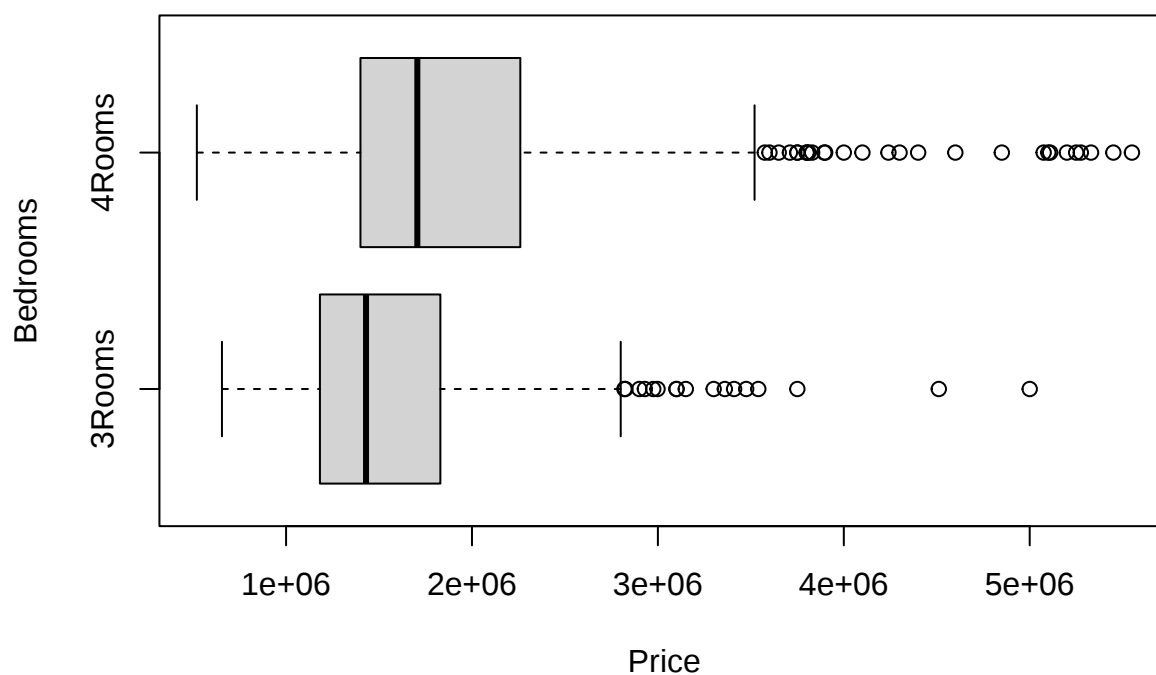
```
stripchart(Price~Bedrooms,main="Price by Number of Bedrooms",method="stack",pch=1,data=house.df)
```

Price by Number of Bedrooms



```
boxplot(Price~Bedrooms,main="Price by Number of Bedrooms",horizontal=TRUE,data=house.df)
```

Price by Number of Bedrooms

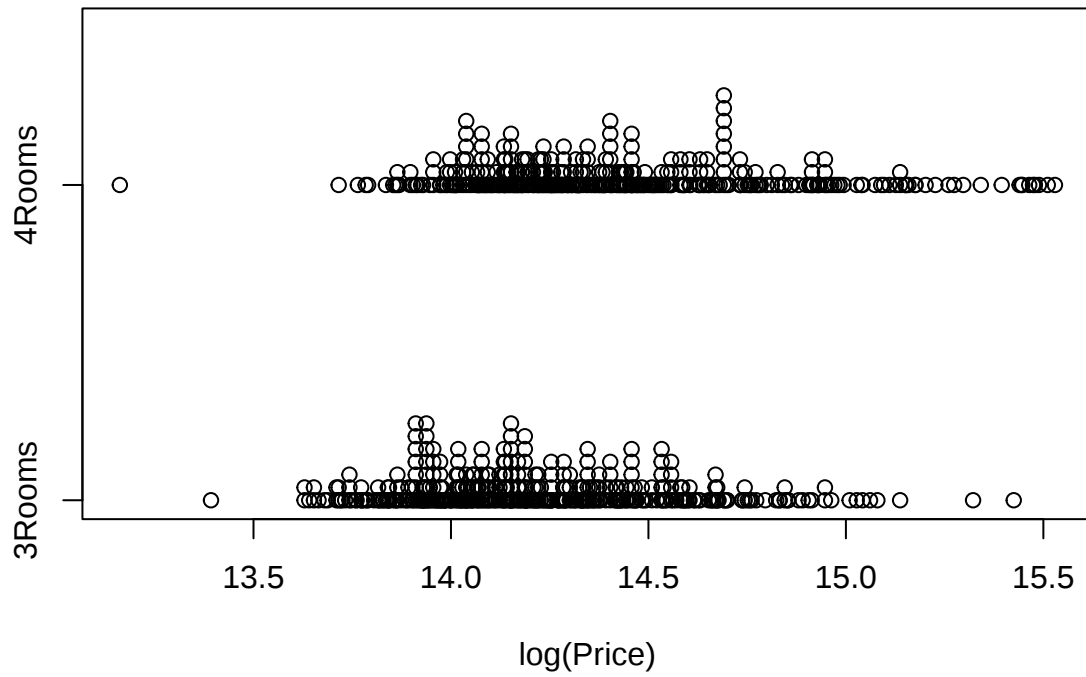


```
library(s20x)
summaryStats(Price~Bedrooms, data=house.df)
```

```
##      Sample Size   Mean  Median Std Dev Midspread
## 3Rooms      419 1582918 1430000 573595.8   648000
## 4Rooms      398 1982567 1706000 872275.7   858000
```

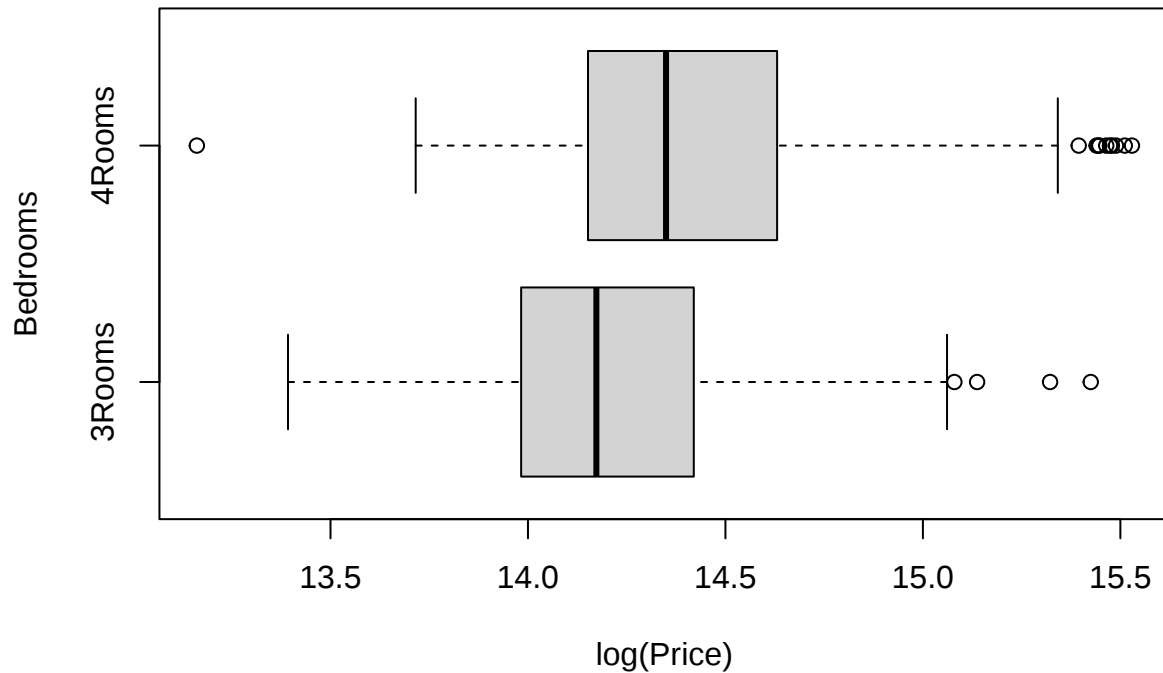
```
stripchart(log(Price)~Bedrooms,main="log(Price) by Number of Bedrooms",method="stack",pch=1,data=house.df)
```

log(Price) by Number of Bedrooms



```
boxplot(log(Price)~Bedrooms,main="log(Price) by Number of Bedrooms",horizontal=TRUE,data=house.df)
```

log(Price) by Number of Bedrooms



```
summaryStats(log(Price)~Bedrooms, data=house.df)
```

##	Sample Size	Mean	Median	Std Dev	Midspread
## 3Rooms	419	14.22053	14.17319	0.3190121	0.4371084
## 4Rooms	398	14.42410	14.34966	0.3724359	0.4774681

Comment on plots and summary statistics

By inspecting the data, we can intuitively see that the average price and median price of a 4-bedroom apartment are more expensive than that of a 3-bedroom apartment, which is what we care about. For the logarithmic price distribution, we can see that the overall distribution of prices is more relevant, convenient for our further data analysis.

1.5 We will fit models to the log-transformed data and to the untransformed data. Why aren't we concerned about the lack of Normality for using either an untransformed or log model?

First of all, we have enough data. During the modeling process, we pay more attention to the relationship between the two sets of data. Logarithmic transformation tends to be normal. After completing the modeling, we can use non-parametric testing to complete hypothesis testing.

1.6 Fit linear model on the untransformed data and check assumptions. Generate confidence intervals for the model.(DO NOT use the Welch test here.)

```
house.df$Price[which(house.df$Bedrooms=="3Rooms")]=1.15*house.df$Price[which(house.df$Bedrooms=="3Rooms")]
shapiro.test(house.df$Price)

##
##  Shapiro-Wilk normality test
##
## data:  house.df$Price
## W = 0.82952, p-value < 2.2e-16

bartlett.test(Price~Bedrooms, house.df)

##
##  Bartlett test of homogeneity of variances
##
## data:  Price by Bedrooms
## Bartlett's K-squared = 31.499, df = 1, p-value = 1.995e-08

aov1=aov(Price~Bedrooms,house.df,alternative="greater")
summary(aov1)

##              Df      Sum Sq   Mean Sq F value    Pr(>F)
## Bedrooms      1 5.371e+12 5.371e+12   9.045 0.00272 **
## Residuals    815 4.839e+14 5.938e+11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

model.tables(aov1,"means")

## Tables of means
## Grand mean
##
## 1899377
##
## Bedrooms
##      3Rooms  4Rooms
## 1820356 1982567
## rep    419    398

predict(aov1,list(1820356,Bedrooms='3Rooms'),interval = "confidence")

##      fit      lwr      upr
## 1 1820356 1746462 1894249

predict(aov1,list(1982567,Bedrooms='4Rooms'),interval = "confidence")

##      fit      lwr      upr
## 1 1982567 1906750 2058385
```

It can be seen that the two data distributions are significantly different, and it can be considered that the price of a 4-bedroom is more than 15% higher than that of a 3-bedroom

1.7 Interpret BOTH confidence intervals from above model (as if for an Executive Summary)

The above two confidence intervals reflect the value interval of the mean value of the two types of houses under the condition of 95% confidence. This reflects the price intervals in which we normally buy a house

1.8

```
shapiro.test(log(house.df$Price))

##
##  Shapiro-Wilk normality test
##
## data:  log(house.df$Price)
## W = 0.96237, p-value = 1.194e-13

bartlett.test(log(Price)~Bedrooms,house.df )

##
##  Bartlett test of homogeneity of variances
##
## data:  log(Price) by Bedrooms
## Bartlett's K-squared = 9.738, df = 1, p-value = 0.001805

aov2=aov(log(Price)~Bedrooms,house.df,alternative="greater")
summary(aov2)

##              Df Sum Sq Mean Sq F value Pr(>F)
## Bedrooms      1   0.83  0.8309    6.938 0.0086 **
## Residuals    815  97.61  0.1198
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

model.tables(aov2,"means")

## Tables of means
## Grand mean
##
## 14.39137
##
## Bedrooms
##      3Rooms 4Rooms
##      14.36  14.42
## rep 419.00 398.00

predict(aov2,list(log(14.36),Bedrooms='3Rooms'),interval = "confidence")

##          fit          lwr          upr
## 1 14.36029 14.32711 14.39348

predict(aov2,list(log(14.42),Bedrooms='4Rooms'),interval = "confidence")

##          fit          lwr          upr
## 1 14.4241 14.39005 14.45815
```

1.9 Interpret BOTH confidence intervals from above model (as if for an Executive Summary).

The above two confidence intervals are the value intervals of the logarithm of two different house prices under the condition of 95% confidence. This reflects the logarithm of the price intervals in which we normally buy a house

1.10 Write the equation for each of the two models fitted (as if for Methods and Assumption Checks).

Methods:ANOVA Assumption: There was no significant difference in the distribution of prices for a 15% increase in the price of a three-bedroom and a four-bedroom equation: $Y = AX + e$ where $A = \text{diag}(I_{\{N1\}}, I_{\{N2\}}), I_{\{N\}}$ means $[1, 1 \dots 1]^T$

```
coefficients(aov1)
```

```
##      (Intercept) Bedrooms4Rooms  
##      1820355.7      162211.7
```

```
coefficients(aov2)
```

```
##      (Intercept) Bedrooms4Rooms  
##      14.36029430      0.06380113
```

1.11 Does buying a four bedroom house increase the cost by more than 15% as compared to buying a three bedroom house? Justify your answer.

We can compare the confidence intervals and find that the obvious confidence interval of the four-bedroom apartment is higher than the confidence interval of the three-bedroom apartment in the previous session, which means that the price of a four-bedroom apartment is 15% higher than that of a three-bedroom apartment.

1.12 To address all the questions the lecturer has, do you prefer to use the log-transformed model or not transformed model? Justify your choice with at least two reasons.

log-transformed model. Because the data has better normality and homogeneity of variance after log transformation, which ensures the rationality of our model. The fitted model has better effect, have a stronger linear correlation.

Question 2

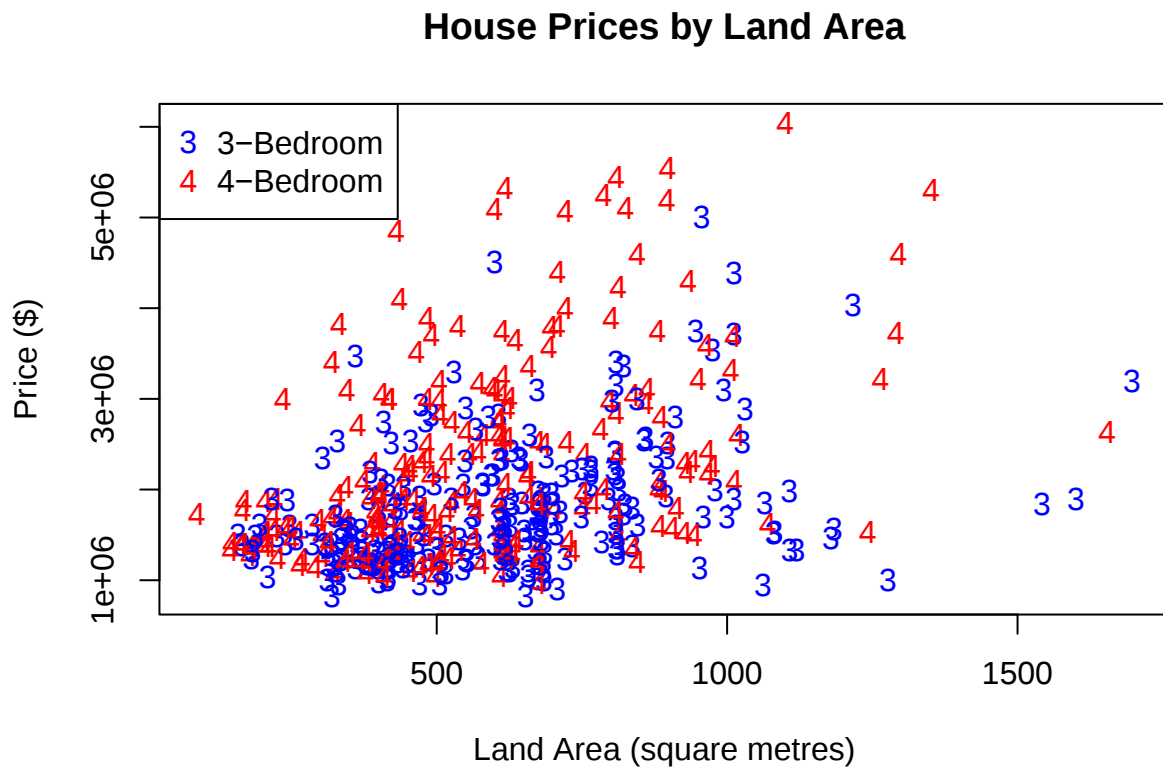
2.1 Question of interest/goal of the study

We wish to see whether sale price of the house increases as the land area increase, and we also wish to see whether the effect of land area on sale price is the same for both 3- and 4-bedroom houses.

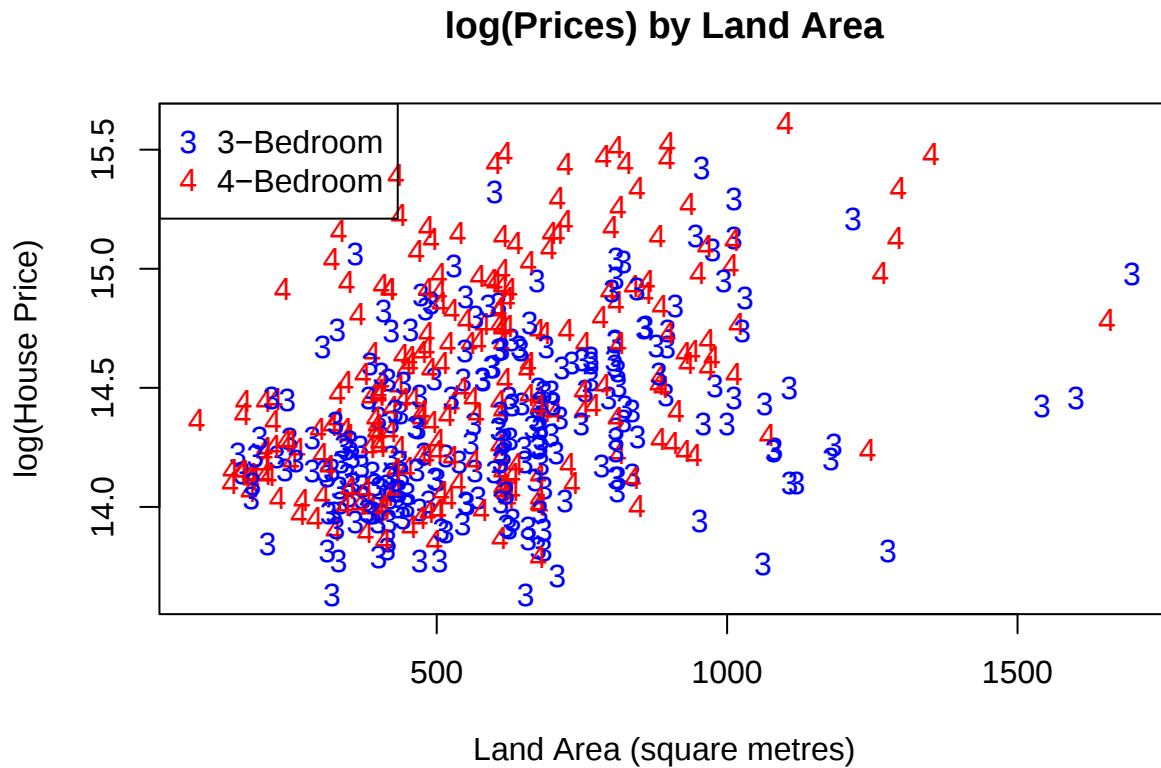
2.2 Read in and inspect the data:

```
land.df = read.csv("land.csv", header = TRUE, stringsAsFactors = TRUE)  
plot(Price ~ Land, data = land.df, ylab = 'Price ($)', xlab = 'Land Area (square metres)',  
     main = "House Prices by Land Area",  
     pch = ifelse(Bedrooms == '3Rooms', '3', '4'),
```

```
col = ifelse(Bedrooms == '3Rooms', 'blue', 'red'))
legend('topleft',c('3-Bedroom','4-Bedroom'),col=c("blue","red"),pch=c('3', '4'))
```



```
plot(log(Price)~Land,data=land.df,ylab='log(House Price)',xlab='Land Area (square metres)',
     main = "log(Prices) by Land Area",
     pch = ifelse(Bedrooms == '3Rooms', '3', '4'),
     col = ifelse(Bedrooms == '3Rooms', 'blue', 'red'))
legend('topleft',c('3-Bedroom','4-Bedroom'),col=c("blue","red"),pch=c('3', '4'))
```



2.3 Comment on plots

Through the plot, we can see that the sale price increases with the increase of the land area and the 4-bedroom increases linearly faster, which is what we are concerned about.

2.4 What is the main reason we should be using log(Price) instead of Price for this analysis?

Eliminate heteroscedasticity and facilitate our calculations

2.5 Fit an appropriate linear Model and Check Assumptions

```
aov3=aov(log(Price)~Land+Bedrooms, land.df)
summary(aov3)
```

```
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Land         1  10.07   10.074    80.47 < 2e-16 ***
## Bedrooms     1   5.80    5.803    46.35 2.94e-11 ***
## Residuals  483   60.47    0.125
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
coefficients(aov3)
```

```
##      (Intercept)          Land Bedrooms4Rooms
```

```
## 1.396992e+01 5.955542e-04 2.192708e-01
```

Through the model coefficients, we find the linear coefficient is small.

2.6 Method and Assumption Checks

Method:ANCOVA

```
cor.test(log(land.df$Price),land.df$Land,method = "pearson")
```

```
##
## Pearson's product-moment correlation
##
## data: log(land.df$Price) and land.df$Land
## t = 8.5774, df = 484, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## 0.2834643 0.4380442
## sample estimates:
## cor
## 0.3632515
```

We can see that as land size has a very weak or non-linear correlation with house price

```
cor.test((log(land.df$Price[which(land.df$Bedrooms=="4Rooms")])))
, (land.df$Land[which(land.df$Bedrooms=="4Rooms")]),method = "pearson")
```

```
##
## Pearson's product-moment correlation
##
## data: (log(land.df$Price[which(land.df$Bedrooms == "4Rooms")]))) and (land.df$Land[which(land.df$Bedrooms == "4Rooms")])
## t = 7.5425, df = 242, p-value = 9.28e-13
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## 0.3286965 0.5326739
## sample estimates:
## cor
## 0.4362731
```

```
cor.test((log(land.df$Price[which(land.df$Bedrooms=="3Rooms")])))
, (land.df$Land[which(land.df$Bedrooms=="3Rooms")]),method = "pearson")
```

```
##
## Pearson's product-moment correlation
##
## data: (log(land.df$Price[which(land.df$Bedrooms == "3Rooms")]))) and (land.df$Land[which(land.df$Bedrooms == "3Rooms")])
## t = 5.8437, df = 240, p-value = 1.656e-08
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## 0.2373935 0.4586255
## sample estimates:
## cor
## 0.3529326
```

By comparing the 4-bedroom and three-bedroom, it can be seen that the 4-bedroom is more affected by the land area

2.7 Executive Summary (Remember to answer ALL the questions asked.)

Through our modeling and analysis, we can get the following conclusions, there is no obvious linear relationship between the land area and house price we care about. However, the 4-bedroom and 3-bedroom are not affected by the change in land area, and the 4-bedroom is more significantly affected