MA30091 Coursework 1

22719

Analysing the Airbnb Price Data in European Cities

```
#loading relevant libraries
library(ggplot2)
library(gridExtra)
library(boot)
```

Task 1: *Are there any missing data in any of the datasets? Comment on if there is/are any variable/variables that may not be useful for further analysis. Calculate the average listing price per person per night for each room type for the weekdays data in Barcelona.* [3 marks]

```
#reading all the data sets
londonwe <- read.csv("london_weekends.csv")
londonwd <- read.csv("london_weekdays.csv")
barcewe <- read.csv("barcelona_weekends.csv")
barcewd <- read.csv("barcelona_weekdays.csv")
```

We can combine all the datasets to prevent excessive code when summarising. This is only to look for instances of missing values and would not be used for anything further.

```
Airbnbs <- rbind(londonwd,londonwe,barcewe,barcewd)
summary(Airbnbs)
```

```
##        X           realSum          room_type         room_shared
##  Min.   :   0   Min.   :   54.33   Length:12826       Length:12826
##  1st Qu.: 801   1st Qu.:  167.80   Class :character   Class :character
##  Median :1790   Median :  245.07   Mode  :character   Mode  :character
##  Mean   :2115   Mean   :  347.29
##  3rd Qu.:3393   3rd Qu.:  418.64
##  Max.   :5378   Max.   :15499.89
##  room_private       person_capacity host_is_superhost      multi
##  Length:12826       Min.   :2.000   Length:12826       Min.   :0.0000
##  Class :character   1st Qu.:2.000   Class :character   1st Qu.:0.0000
```

```
##  Mode  :character   Median :2.000   Mode  :character   Median :0.0000
##                     Mean   :2.795                     Mean   :0.2993
##                     3rd Qu.:4.000                     3rd Qu.:1.0000
##                     Max.   :6.000                     Max.   :1.0000
##       biz          cleanliness_rating guest_satisfaction_overall    bedrooms
##  Min.   :0.0000   Min.   : 2.000     Min.   : 20.00              Min.   :0.000
##  1st Qu.:0.0000   1st Qu.: 9.000     1st Qu.: 87.00              1st Qu.:1.000
##  Median :0.0000   Median :10.000     Median : 93.00              Median :1.000
##  Mean   :0.3741   Mean   : 9.201     Mean   : 90.75              Mean   :1.136
##  3rd Qu.:1.0000   3rd Qu.:10.000     3rd Qu.: 99.00              3rd Qu.:1.000
##  Max.   :1.0000   Max.   :10.000     Max.   :100.00              Max.   :8.000
##       dist             metro_dist        attr_index      attr_index_norm
##  Min.   : 0.04055   Min.   :0.01299   Min.   :  68.74   Min.   :  3.198
##  1st Qu.: 2.44712   1st Qu.:0.30001   1st Qu.: 188.25   1st Qu.: 11.966
##  Median : 4.28881   Median :0.48014   Median : 279.77   Median : 16.650
##  Mean   : 4.61752   Mean   :0.88091   Mean   : 331.96   Mean   : 19.676
##  3rd Qu.: 6.13409   3rd Qu.:0.89133   3rd Qu.: 407.41   3rd Qu.: 24.425
##  Max.   :17.32121   Max.   :9.28623   Max.   :2934.13   Max.   :100.000
##    rest_index     rest_index_norm        lng                lat
##  Min.   : 140.5   Min.   :  2.515   Min.   :-0.251700   Min.   :41.35
##  1st Qu.: 399.0   1st Qu.:  7.342   1st Qu.:-0.151370   1st Qu.:51.44
##  Median : 575.3   Median : 10.679   Median :-0.087050   Median :51.49
##  Mean   : 683.5   Mean   : 13.033   Mean   : 0.389705   Mean   :49.27
##  3rd Qu.: 830.6   3rd Qu.: 15.660   3rd Qu.: 0.004855   3rd Qu.:51.52
##  Max.   :5587.1   Max.   :100.000   Max.   : 2.225520   Max.   :51.58
```

The summary above shows that there are no missing values in any of our 4 data sets.

The variable X will not be useful for further analysis because it is just the row numbers.

Three variables relate to room type: room_type, room_shared and room_private. Room shared states whether the room is shared, and room_private shows whether it is private. However, room type is a combination of these two variables and also whether it is the entire apartment:

```
levels(factor(Airbnbs$room_type))
```

```
## [1] "Entire home/apt" "Private room"    "Shared room"
```

This means room_shared and room_private are unnecessary for further analysis because we can look at room_type to gather all the necessary information. We shall disregard them. Longitude and latitude are irrelevant variables because we have already categorised the data by city. The difference in longitude and latitude between the Airbnbs distributed around each city will be minimal. Therefore, these two variables will not be used for further analysis. With the restaurant and attraction indexes, we have both the regular and normalised data.

Including both of these will not be necessary, but determining which one will depend on the analysis, we carry out. We shall decide which to use further on. The multi variable shows whether the listing belongs to a host with 2-4 offers. This variable is difficult to analyse because a 0 can respond to the host having only one or over four offers. This is a significant difference and can impact the effectiveness of this data. Because we have the variable biz, which shows whether the host has over four offers, we can use this variable to analyse the relationship between offers and other variables, so multi will be discarded.

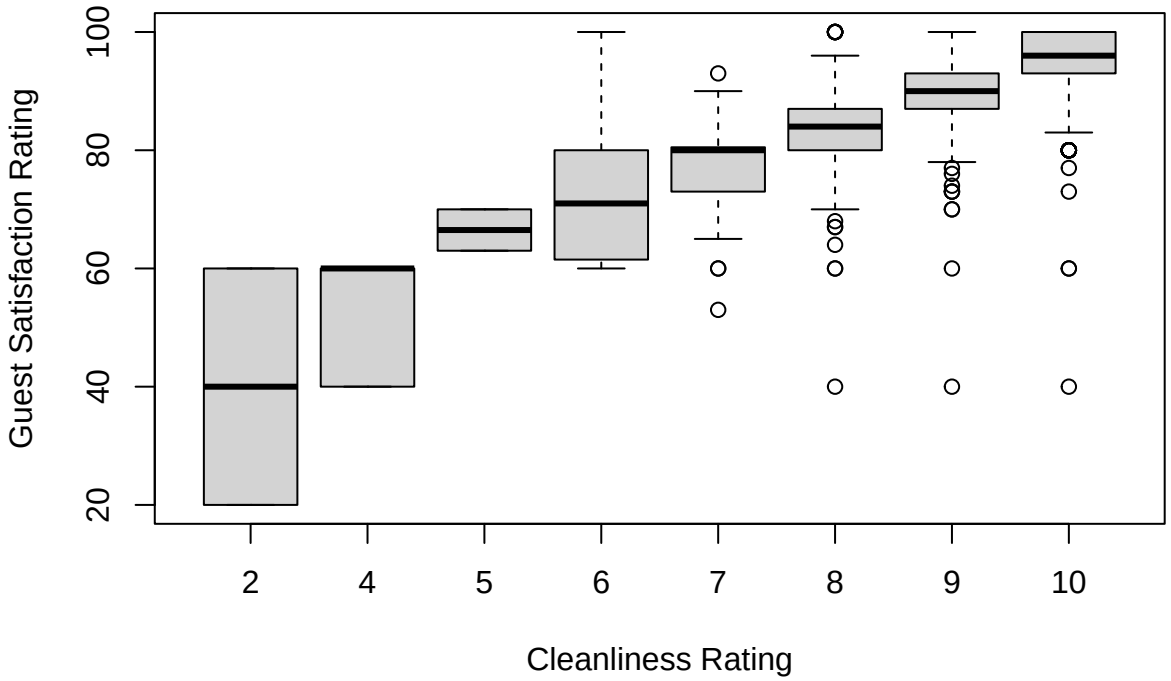
Using the below code, we can calculate the average listing price per person per night for each room type in Barcelona on weekdays.

```
#Creating vectors containing only the realSum data for each certain room type.

aptbarce <- barcewd[barcewd$room_type == "Entire home/apt",]
avlisting1 <- (aptbarce$realSum)/4
x <- sum(avlisting1)/length(avlisting1)

privbarce <- barcewd[barcewd$room_type =="Private room",]
avlisting2 <- (privbarce$realSum)/4
y <- sum(avlisting2)/length(avlisting2)

sharebarce <- barcewd[barcewd$room_type == "Shared room",]
avlisting3 <-(sharebarce$realSum)/4
z <- sum(avlisting3)/length(avlisting3)

a <- c(x,y,z)
avlisting <- matrix(a, nrow = 1, ncol =3)
colnames(avlisting) = c("Entire home/apt", "Private room", "Shared room")

avlisting
```

```
##      Entire home/apt Private room Shared room
## [1,]          143.14     50.68847    28.69937
```

Task 2: *Using appropriate exploratory tools such as tables/graphs/summary statistics comment on the relationship between cleanliness and guest satisfaction in the weekdays data in Barcelona. Also comment on the relationship between superhost and guest satisfaction using exploratory analysis on the weekdays data in London.* [4 marks]

```
boxplot(barcewd$guest_satisfaction_overall~barcewd$cleanliness_rating,
     xlab = "Cleanliness Rating", ylab = "Guest Satisfaction Rating")
```



Based on the boxplots above, there appears to be a positive linear relationship between cleanliness rating and guest satisfaction rating in Airbnbs. As the cleanliness rating increases, the guest satisfaction rating also tends to increase, indicating that guests are more satisfied with their stay when the accommodations are cleaner.

```
#looking at the summary statistics
by(barcewd$guest_satisfaction_overall,barcewd$cleanliness_rating, summary)
```

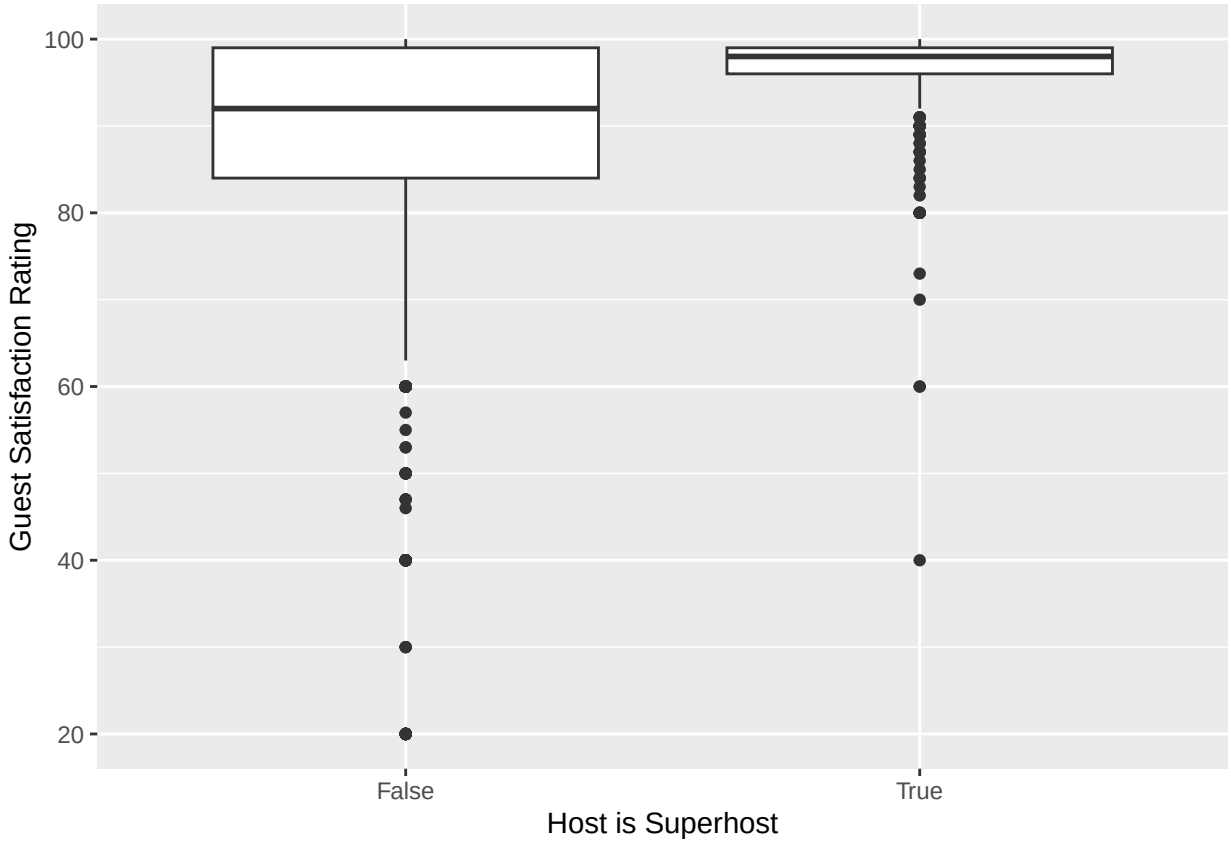
```
## barcewd$cleanliness_rating: 2
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      20      20      40      40      60      60
## ------------------------------------------------------------
## barcewd$cleanliness_rating: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      40      40      60      52      60      60
## ------------------------------------------------------------
## barcewd$cleanliness_rating: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   63.00   64.75   66.50   66.50   68.25   70.00
## ------------------------------------------------------------
## barcewd$cleanliness_rating: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   60.00   61.50   71.00   72.11   80.00  100.00
## ------------------------------------------------------------
## barcewd$cleanliness_rating: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   53.00   73.00   80.00   77.12   80.25   93.00
```

```
## ------------------------------------------------------------
## barcewd$cleanliness_rating: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   40.00   80.00   84.00   83.12   87.00  100.00
## ------------------------------------------------------------
## barcewd$cleanliness_rating: 9
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   40.00   87.00   90.00   89.46   93.00  100.00
## ------------------------------------------------------------
## barcewd$cleanliness_rating: 10
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   40.00   93.00   96.00   95.15  100.00  100.00
```

The summary statistics above confirm the conclusion from the plot. As we increase the cleanliness rating, the average (mean and median) guest satisfaction rating increases.

For the relationship between superhost and guest satisfaction rating, a box plot is the most ideal exploratory tool to use because superhost is a categorical variable.

```
#boxplot
ggplot(londonwd, aes(x = host_is_superhost, y = guest_satisfaction_overall)) +
  geom_boxplot() + xlab("Host is Superhost") + ylab("Guest Satisfaction Rating")
```



These boxplots show that the range of guest satisfaction data is smaller for super hosts

because there is a smaller sample size for super hosts than non-super hosts. Despite this, it is still clear that super hosts have higher guest satisfaction, and this is because the median guest satisfaction is a lot higher for super hosts.

```
#table of super-host variable and summary statistics
table(barcewd$host_is_superhost)
```

```
##
## False  True
##  1274   281
```

```
by(barcewd$guest_satisfaction_overall,barcewd$host_is_superhost,summary)
```

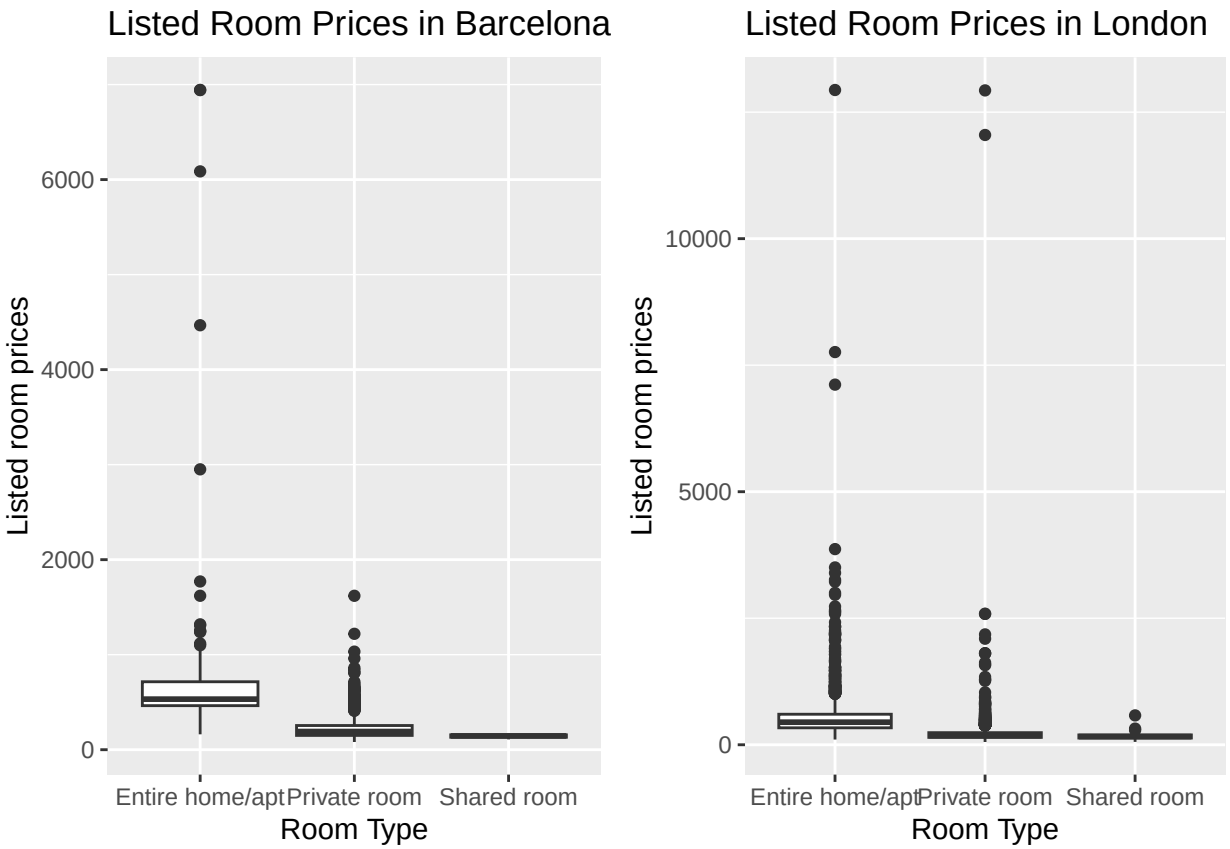
```
## barcewd$host_is_superhost: False
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   20.00   86.00   91.00   89.73   95.00  100.00
## ------------------------------------------------------------
## barcewd$host_is_superhost: True
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   86.00   95.00   96.00   96.35   98.00  100.00
```

When looking at the summary statistics, we come to the same conclusion. The median and mean are much higher for super hosts, and the lower quartile is also higher, suggesting that no super hosts get a low rating. However, the upper quartile guest satisfaction for non-super hosts is high, so they can still get a high rating despite not having super host status. This could suggest that super hosts become super hosts because of their high guest satisfaction ratings, meaning instead of super host status influencing guest satisfaction, guest satisfaction influences whether someone becomes a super host.

Task 3: *Use an appropriate plot to illustrate the distribution of listed room prices per room type for the Barcelona and London weekends datasets. You should provide separate plots for each city. Comment on what you observe.* [3 marks]

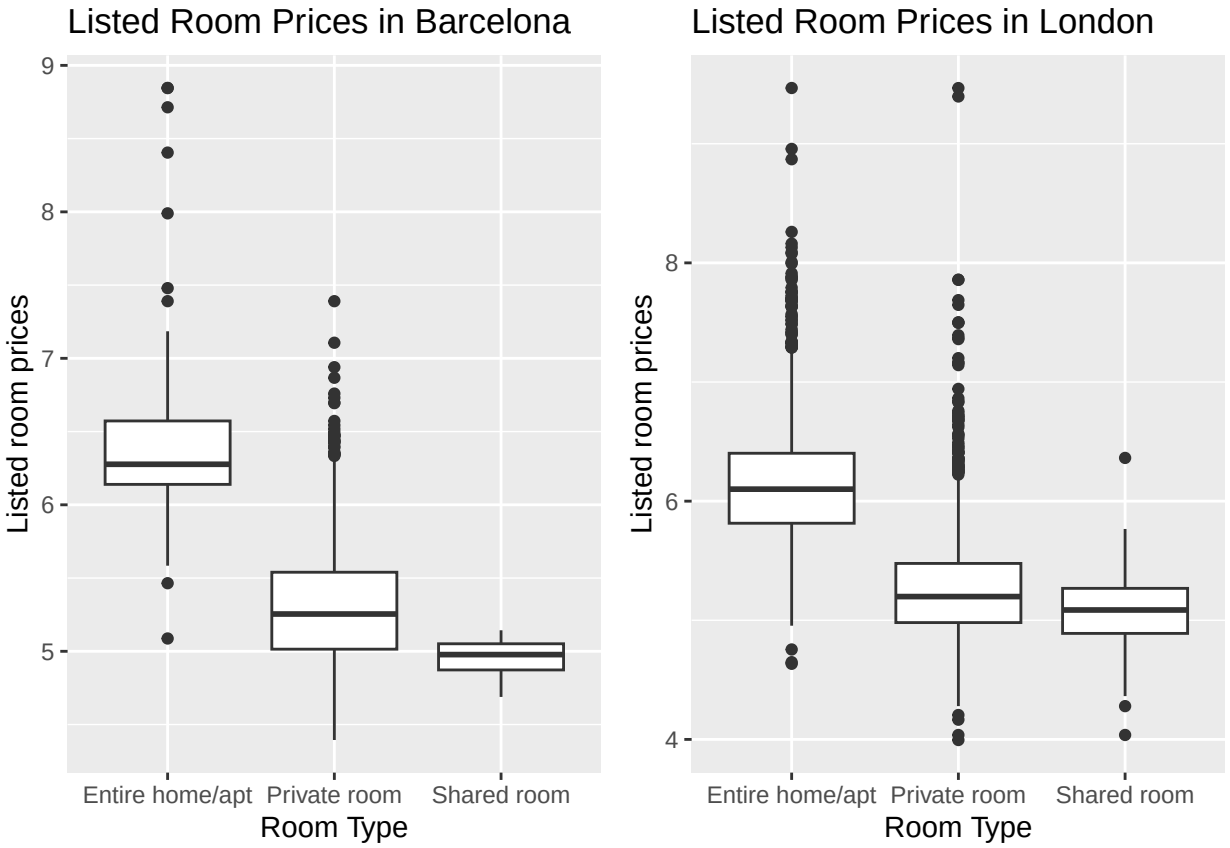
As we are comparing the three room types, a box plot would be the most suitable plot.

```
library(gridExtra)
#box plots
p1 <- ggplot(barcewe, aes(x = room_type, y = realSum)) +
  geom_boxplot() + xlab("Room Type") + ylab("Listed room prices") + ggtitle("Listed Room
p2 <- ggplot(londonwe, aes(x = room_type, y = realSum)) +
  geom_boxplot() + xlab("Room Type") + ylab("Listed room prices") + ggtitle("Listed Room

#arranging the box plots in to one plot.
grid.arrange(p1,p2,ncol=2)
```



There is an issue with this plot. Because of many significant outliers, our plots are skewed and hard to compare. Because there are many outliers, removing them from the data is inappropriate. Instead, we can look for a way to transform the realSum data so that their impact is less extreme. The ideal transformation is using log transformation.

```
p3 <- ggplot(barcewe, aes(x = room_type, y = log(realSum))) +
  geom_boxplot() + xlab("Room Type") + ylab("Listed room prices") + ggtitle("Listed Room
p4 <- ggplot(londonwe, aes(x = room_type, y = log(realSum))) +
  geom_boxplot() + xlab("Room Type") + ylab("Listed room prices") + ggtitle("Listed Room

grid.arrange(p3,p4,ncol=2)
```



This plot is much more readable. Both cities have a clear relation between room type and the listed room price. Entire home listings are more expensive than private and shared homes. The median, upper quartile, and lower quartile of entire homes are higher than those of private and shared rooms, indicating that the prices of entire homes are generally higher than those of the other two types of accommodations. Shared rooms are the cheapest listings for both cities. However, there is less variability between the prices of private rooms and shared rooms in London than in Barcelona, suggesting London is a more expensive city, so even shared rooms are expensive. Nevertheless, entire apartments are the most expensive, and shared rooms are the cheapest.

Task 4: *Perform an appropriate statistical test to compare the listed room price in London vs Barcelona on weekends. You may combine the data across room types to perform the statistical test. Discuss what assumptions you make to perform the test. Comment on the appropriateness of the assumptions made.* [6 marks]

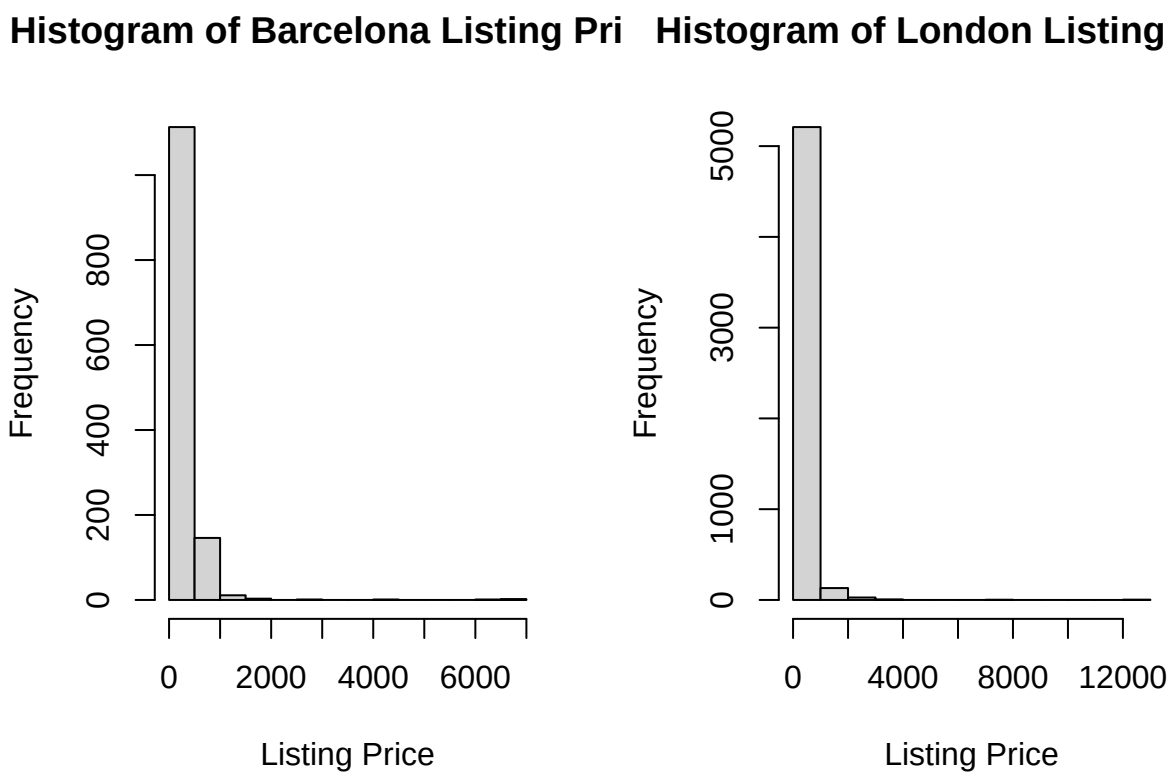
To compare two samples, we can use an independent t-test. The main assumptions for this test are that the variances are equal.

```
var.test(barcewe$realSum,londonwe$realSum)
```

The variance test gives a p-value of 1.963e-06 which is significant. Therefore, we reject the null hypothesis, and equal variance is not assumed. A more appropriate test is Welch's test, as this test allows for unequal variance. Welch's test also assumes that the two samples are

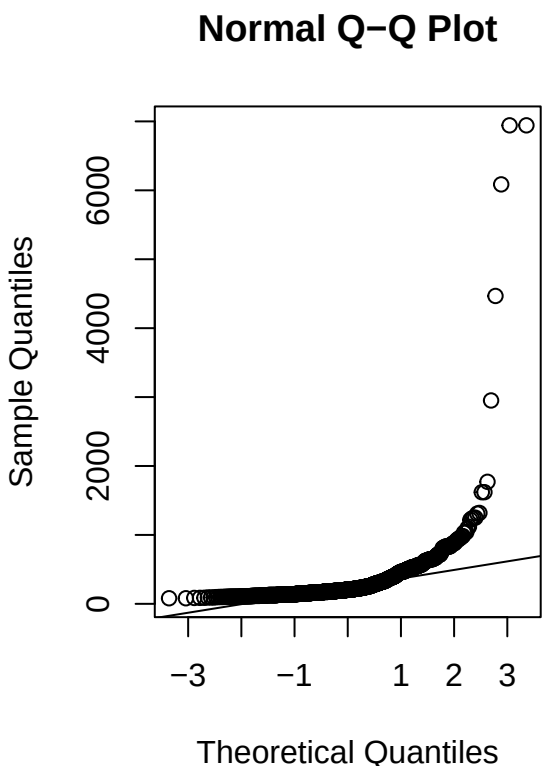
independent, which is true here because the two samples are from different cities and are both continuous. The final assumption is normality.

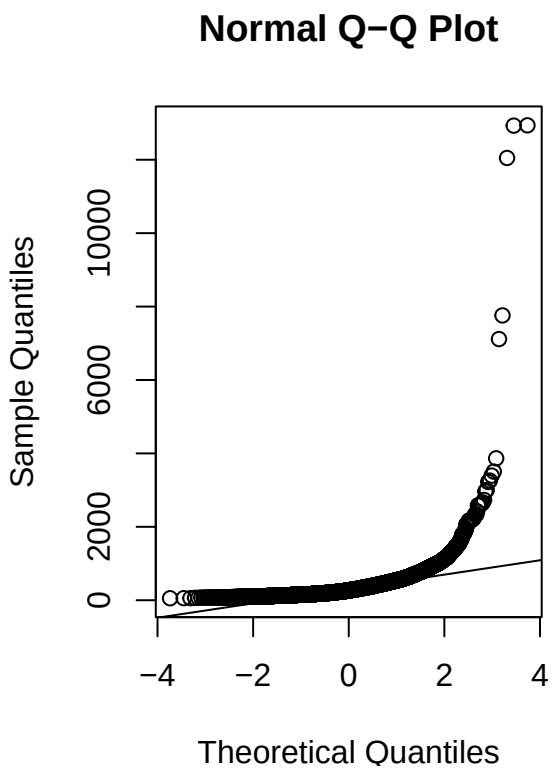
```
par(mfrow=c(1,2))
hist(barcewe$realSum, xlab = "Listing Price", main = "Histogram of Barcelona Listing Pri
hist(londonwe$realSum,xlab = "Listing Price", main = "Histogram of London Listing Prices
```



We can see from the histograms that neither sample follows the normal distribution and is right-skewed, which suggests normality assumption is unlikely to be met. However, both sample sizes are substantial, 1555 and 5379, meaning that the central limit theorem may apply and the normality assumption is less critical.

```
par(mfrow=c(1,2))

qqnorm(barcewe$realSum)
qqline(barcewe$realSum)
qqnorm(londonwe$realSum)
qqline(londonwe$realSum)
```





Unfortunately, the above Q-Q Plot suggests that the data is severely non-normal due to the large tails and that using Welch's two-sample test is inappropriate. Instead, we can use a nonparametric test like the Wilcoxon test, and this test does not require the normality assumption.

```
wilcox.test(londonwe$realSum,barcewe$realSum, alternative="two.sided")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  londonwe$realSum and barcewe$realSum
## W = 4054396, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

We have a significant p-value (<2.2e-16), which means that we shall reject the null hypothesis and conclude that there is a statistically significant difference between listing prices in London and Barcelona. We can look at the medians to see which city has the higher listing prices, as this is what the Wilcoxon test compares.

```
median(barcewe$realSum)
```

```
## [1] 204.69
```

```
median(londonwe$realSum)
```

```
## [1] 268.12
```

London weekends have a higher median. Therefore, London has a statistically higher average listing price than Barcelona.

Task 5: *Use a generalised linear model (GLM) to study the differences between the listed room prices on weekdays and weekends for Barcelona. Check your modelling assumptions.* [10 marks]

```
#combining the Barcelona weekends and weekdays data sets.
weekends <- data.frame(barcewe, days_of_week="Weekend")
weekdays <- data.frame(barcewd, days_of_week="Weekdays")
barcelona <- rbind(weekends,weekdays)

#making sure categorical variables are set as factors
barcelona$days <- as.factor(barcelona$days_of_week)
barcelona$type <- as.factor(barcelona$room_type)
barcelona$host <- as.factor(barcelona$host_is_superhost)
barcelona$bed <- as.factor(barcelona$bedrooms)
```

As we focus on one critical explanatory variable, days_of_week, we start by fitting a GLM with this variable as the only explanatory variable.

```
#initial glm
mod0 <- glm(realSum ~ days, family = Gamma(identity), data = barcelona)
mod1 <- glm(log(realSum) ~ days, family=Gamma(identity), data = barcelona)

AIC(mod0,mod1)
```

```
##      df       AIC
## mod0  3 36811.222
## mod1  3  4658.974
```

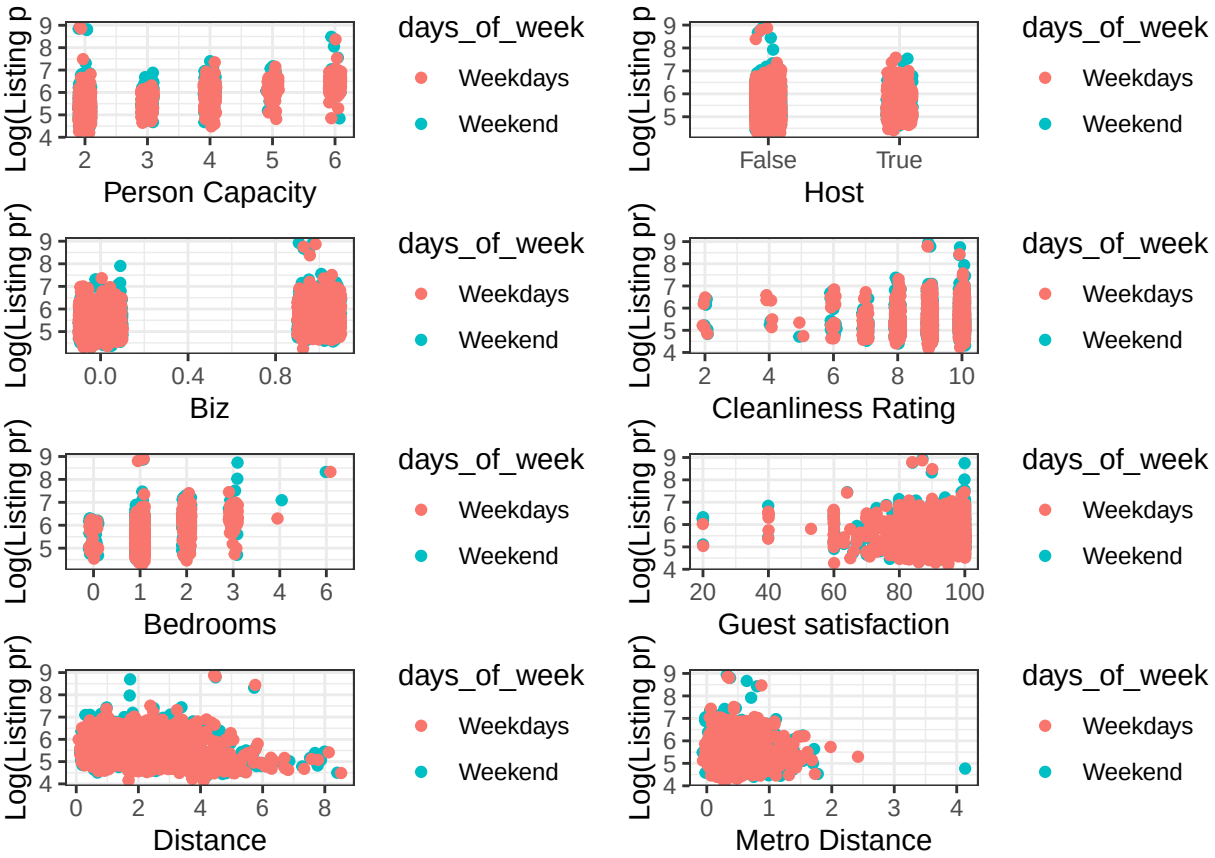
As we did in task 3, a log transformation on the response variable may be helpful because there are outliers in the response variable. Looking at both AICs, the AIC for mod1 is drastically smaller, suggesting that the log transformation gives a much better model. The family gamma is used with the identity link because realSum is continuous.

```
summary(mod1)
```

This initial model, mod1, suggests that the days of the week variable is insignificant because of the p-value, 0.768. However, as our key explanatory variable, we shall keep it in the model. As we add more variables to the GLM, the p-value should change.

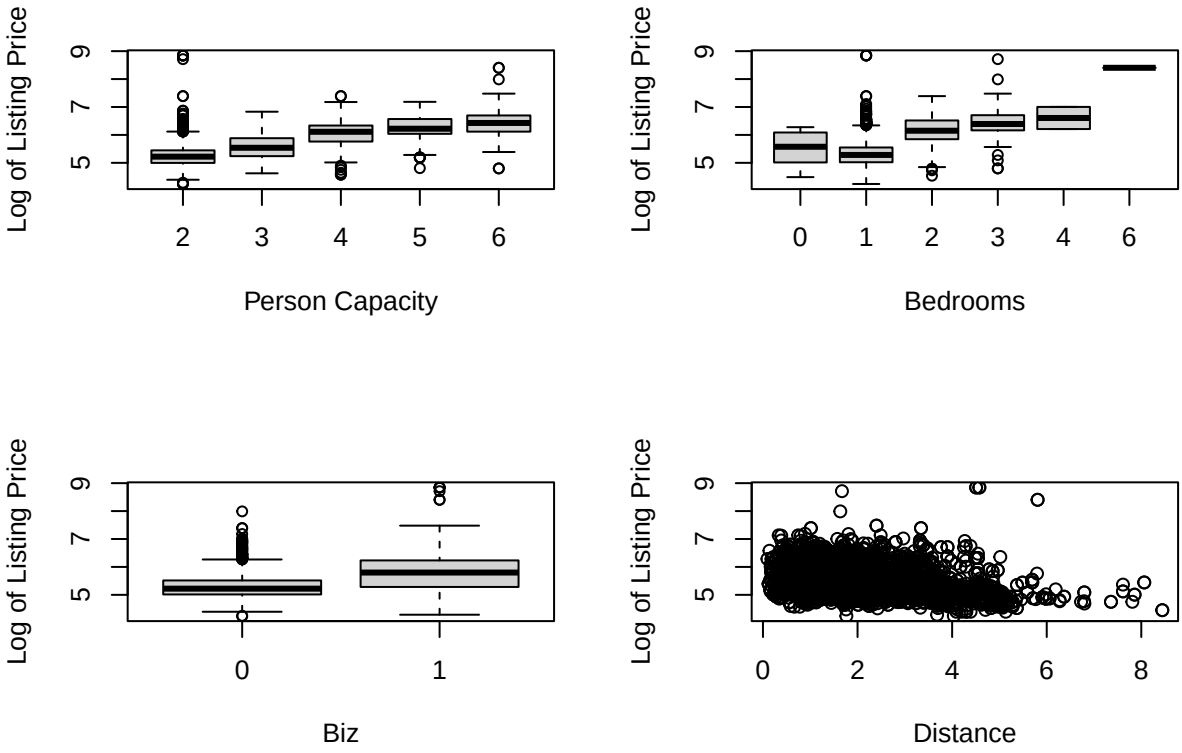
We are focusing on one key explanatory variable, days_of_week so we need to look for potential confounders through exploratory analysis. Potential confounders include: type, person_capacity, host, biz, cleanliness_rating, guest_satisfaction_overall, bedrooms, dist, metro_dist, attr_index, attr_index_norm, rest_index_norm, rest_index_norm. Room type is a very important variable in listing price, as we have seen in previous tasks. Therefore, it will definitely be added to the model. For person capacity, superhost, biz, cleanliness rating, bedrooms, guest satisfaction, distance and metro distance, we can look at plots of the data and compare to see if there are notable trends to suggest association with both days of the week and listing price.

```
library(gridExtra)
#person capacity
p1 <- ggplot(barcelona) + theme_bw()+
  geom_jitter(aes(x=person_capacity ,y=log(realSum),col=days_of_week),width = 0.1, heigh
#superhost
p2 <- ggplot(barcelona) + theme_bw()+
  geom_jitter(aes(x=host ,y=log(realSum),col=days_of_week),width = 0.1, height = 0.1) +
#biz
p3 <- ggplot(barcelona) + theme_bw()+
  geom_jitter(aes(x=biz ,y=log(realSum),col=days_of_week),width = 0.1, height = 0.1) + x
#cleanliness rating
p4 <- ggplot(barcelona) + theme_bw()+
  geom_jitter(aes(x=cleanliness_rating ,y=log(realSum),col=days_of_week),width = 0.1, he
#bedrooms
p5 <- ggplot(barcelona) + theme_bw()+
  geom_jitter(aes(x=bed ,y=log(realSum),col=days_of_week),width = 0.1, height = 0.1) + x
#guest satisfaction
p6 <- ggplot(barcelona) + theme_bw()+
  geom_jitter(aes(x=guest_satisfaction_overall ,y=log(realSum),col=days_of_week),width =
#distance
p7 <- ggplot(barcelona) + theme_bw()+
  geom_jitter(aes(x=dist ,y=log(realSum),col=days_of_week),width = 0.1, height = 0.1) +
#metro distance
p8 <- ggplot(barcelona) + theme_bw()+
  geom_jitter(aes(x=metro_dist,y=log(realSum),col=days_of_week),width = 0.1, height = 0.

grid.arrange(p1,p2,p3,p4,p5,p6,p7,p8, ncol=2, nrow=4)
```



Although there is little evidence to suggest that listing prices change between weekend and weekdays for each of these variables, some still show a trend in listing price as the variable changes. These are shown in the plots below.

```
par(mfrow=c(2,2))
boxplot(log(barcelona$realSum)~barcelona$person_capacity, xlab = "Person Capacity",
        ylab = "Log of Listing Price")
boxplot(log(barcelona$realSum)~barcelona$bed, xlab = "Bedrooms",
        ylab = "Log of Listing Price")
boxplot(log(barcelona$realSum)~barcelona$biz, xlab = "Biz",
        ylab = "Log of Listing Price")
plot(barcelona$dist,log(barcelona$realSum), xlab = "Distance",
     ylab = "Log of Listing Price")
```

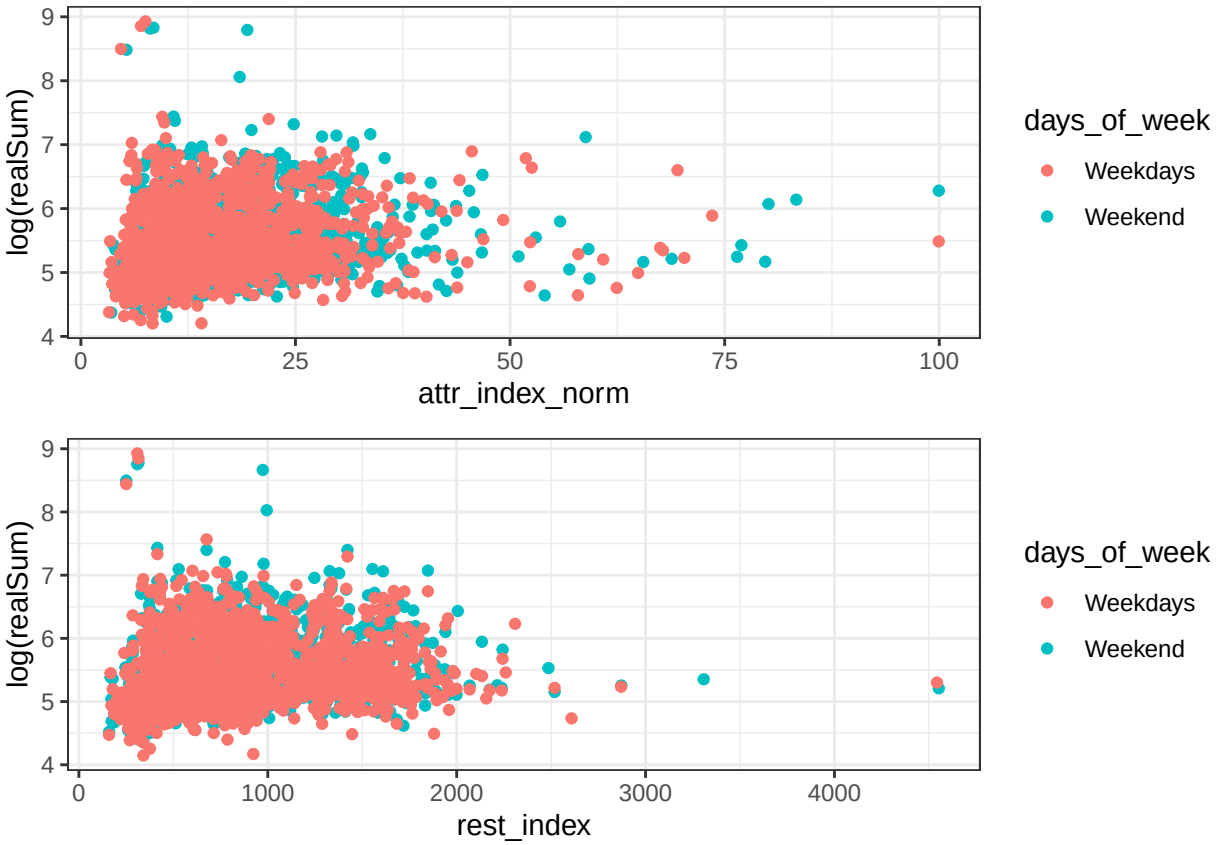


We will therefore consider these as confounders and include them in the potential model.

When looking at the plots of the regular and normalised attraction index against realSum, the normalised attraction index is preferable. However, for the restaurant index, there is minimal difference between the two plots so we can use the regular restaurant index.

However, we now need to see if attr_index_norm and rest_index are confounders.

```
#normalised attraction index
plot1 <- ggplot(barcelona) + theme_bw()+
  geom_jitter(aes(x=attr_index_norm ,y=log(realSum),col=days_of_week),width = 0.1, heigh
#restaurant index
plot2 <- ggplot(barcelona) + theme_bw()+
  geom_jitter(aes(x=rest_index ,y=log(realSum),col=days_of_week),width = 0.1, height = 0

grid.arrange(plot1,plot2,nrow=2)
```



Plot 1 and 2 suggest that attraction and restaurant index do not impact listing prices changing between weekend and weekdays. However, the plot does show a potential relationship between the restaurant index and listing price. Therefore, we will keep restaurant index in the model.

Now, we can build a better model including the confounders.

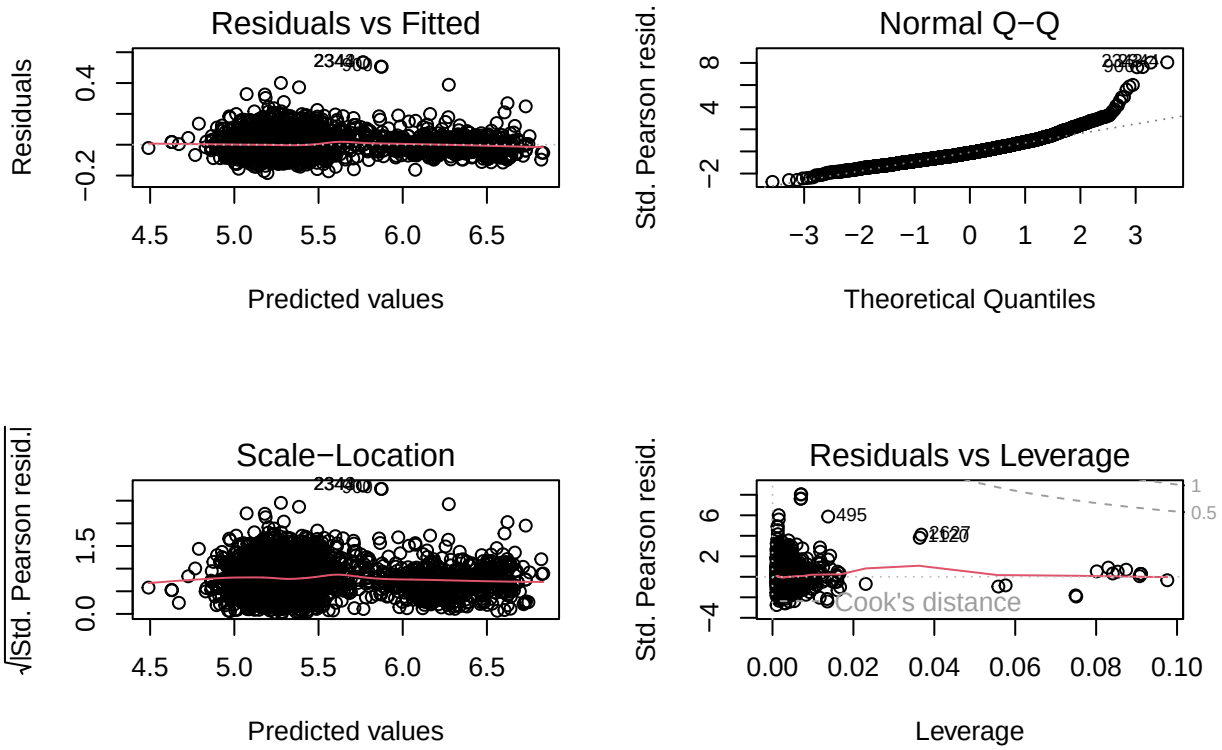
```
mod2 <- glm(log(realSum) ~ days + type + person_capacity + bedrooms + biz + dist
            + rest_index, family=Gamma(identity), data=barcelona)
AIC(mod2)
```

```
## [1] 2160.523
```

By including the confounders, we have reduced the AIC to 2160.5 suggesting a much better model. Also, all of the variables are significant which is another sign of a good model. When checking with the step function to see if removing any more variables improves the AIC, it showed that nothing should be removed. From our GLM, we can see that days of the week is a significant variable due to its small p value, suggesting differences between the listing prices for weekends and weekdays in Barcelona.

Now we can check modelling assumptions by looking at the diagnostic plots.

```
par(mfrow=c(2,2))
plot(mod2)
```



The first assumption we need to check is the linearity between the transformed expected response and the explanatory variables, and this can be checked in the residual vs fitted plot. There is no noticeable trend in the plot, suggesting no missing predictors. From this, we can also see that the independence assumption is met. Also, there is constant variance, as shown in the Scale-Location plot. This shows that the homoscedasticity assumption is also met. The Q-Q plot allows us to comment on the normality assumption. There is a slight tail at the end, but it follows the normal distribution quite well. In the final plot, there are no significant outliers impacting our data.

```
summary(mod2)
```

```
##
## Call:
## glm(formula = log(realSum) ~ days + type + person_capacity +
##     bedrooms + biz + dist + rest_index, family = Gamma(identity),
##     data = barcelona)
##
## Deviance Residuals:
##       Min         1Q     Median         3Q        Max
## -0.19617  -0.04407  -0.00911   0.03258   0.46137
##
```

```
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        5.540e+00  5.990e-02  92.499  < 2e-16 ***
## daysWeekend        1.079e-01  1.376e-02   7.841 6.30e-15 ***
## typePrivate room -6.411e-01  2.774e-02 -23.110  < 2e-16 ***
## typeShared room  -1.110e+00  9.589e-02 -11.579  < 2e-16 ***
## person_capacity    1.139e-01  1.076e-02  10.594  < 2e-16 ***
## bedrooms           9.528e-02  2.047e-02   4.653 3.41e-06 ***
## biz                1.286e-01  1.660e-02   7.745 1.32e-14 ***
## dist              -5.533e-02  8.671e-03  -6.382 2.04e-10 ***
## rest_index         7.754e-05  2.618e-05   2.961  0.00309 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.004444152)
##
##     Null deviance: 28.889  on 2832  degrees of freedom
## Residual deviance: 11.912  on 2824  degrees of freedom
## AIC: 2160.5
##
## Number of Fisher Scoring iterations: 4
```

From the summary, we can see that the coefficient for Weekend in the days variable is 0.107799 which indicates the average difference between the log of weekday listing prices and weekend listing prices. As we have used log of the realSum, we can convert this coefficient to a percentage difference in listing prices using the below equations.

```
#distance
(exp(-0.05533)-1)*100
```

```
## [1] -5.382714
```

```
#room type
#private room
(exp(-0.6411)-1)*100
```

```
## [1] -47.32873
```

```
#shared room
(exp(-1.110)-1)*100
```

```
## [1] -67.0441
```

```
#days of the week
(exp(0.1079)-1)*100
```

```
## [1] 11.39363
```

This means that, on average, weekend listings have a 11.39% higher price than weekday listings, holding all other variables constant.

Task 6: *Fit a GLM to the listed room price on weekdays in Barcelona. Use this model to predict the listed room prices for Barcelona on weekends. Calculate the prediction error and the cross validation error (perform 10-fold cross validation). Comment on your findings.* [5 marks]

To fit the GLM, we are going to start by including all the variables and then refine the model. Variables that were deemed not useful in task 1 are not included. The same family and link function is used as in task 5. We shall also still use log transformation on the response variable.

```
m0 <- glm(log(realSum) ~ room_type + person_capacity + host_is_superhost + biz
    + cleanliness_rating + guest_satisfaction_overall + bedrooms + dist + metro_dist
    + attr_index_norm + rest_index, family=Gamma(identity), data=barcewd)
summary(m0)
```

```
##
## Call:
## glm(formula = log(realSum) ~ room_type + person_capacity + host_is_superhost +
##     biz + cleanliness_rating + guest_satisfaction_overall + bedrooms +
##     dist + metro_dist + attr_index_norm + rest_index, family = Gamma(identity),
##     data = barcewd)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.19976  -0.03937  -0.00683   0.03243   0.47603
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 4.997e+00  1.193e-01  41.891  < 2e-16 ***
## room_typePrivate room      -6.151e-01  3.297e-02 -18.656  < 2e-16 ***
## room_typeShared room       -1.122e+00  1.062e-01 -10.568  < 2e-16 ***
## person_capacity             1.190e-01  1.263e-02   9.417  < 2e-16 ***
## host_is_superhostTrue       7.451e-02  2.298e-02   3.242  0.00121 **
## biz                         1.048e-01  2.027e-02   5.173 2.61e-07 ***
## cleanliness_rating          2.080e-02  1.205e-02   1.726  0.08459 .
## guest_satisfaction_overall  3.112e-03  1.447e-03   2.151  0.03160 *
```

```
## bedrooms                    7.641e-02  2.418e-02   3.160  0.00161 **
## dist                       -5.187e-02  1.127e-02  -4.604 4.48e-06 ***
## metro_dist                  8.493e-02  3.357e-02   2.529  0.01152 *
## attr_index_norm            -2.102e-04  1.326e-03  -0.159  0.87404
## rest_index                  8.408e-05  3.556e-05   2.364  0.01819 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.003663967)
##
##     Null deviance: 15.4511  on 1554  degrees of freedom
## Residual deviance:  5.3528  on 1542  degrees of freedom
## AIC: 890.36
##
## Number of Fisher Scoring iterations: 5
```

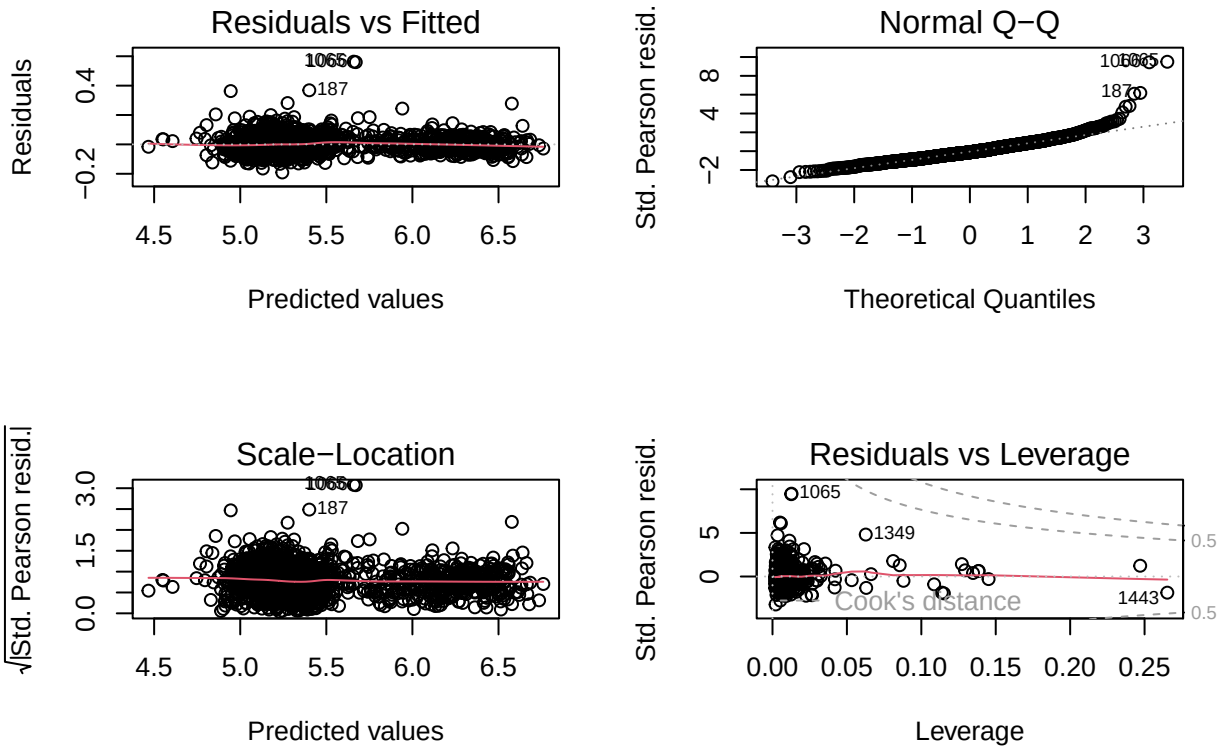
The summary suggests that attraction index is not a significant variable so we shall remove it. In task 3, we saw that there is a relationship between guest satisfaction and cleanliness so adding the interaction could be significant.

```
m1 <- glm(log(realSum) ~ room_type + person_capacity + host_is_superhost + biz +
            cleanliness_rating*guest_satisfaction_overall +  + dist + bedrooms
          + metro_dist + rest_index, family=Gamma(identity), data=barcewd)
AIC(m1)
```

```
## [1] 845.8336
```

We get a much better AIC suggesting that the interaction is significant. The step function suggests that no other variables need to be removed.

```
par(mfrow=c(2,2))
plot(m1)
```



The diagnostic plots look good with no significant outliers and pretty constant variance. There is a slight tail in the Q-Q plot but it is not significant.

Now we can predict the listing prices on weekends in Barcelona using this GLM.

```
pred <- predict(m1, newdata= barcewe, type="response")
#exp because we used the log transformation.
predicted <- exp(pred)

#prediction error

pred.err=function(obs,out,modeltrn)
{
  pred=predict(modeltrn,obs,type="response")
  predicted=exp(pred)
  mean((out-round(predicted))^2)
}

pred.err(barcewe,barcewe$realSum,m1)
```

```
## [1] 129886.3
```

The prediction error is very high. This suggests that modelling Barcelona weekend prices from weekday prices is not appropriate.

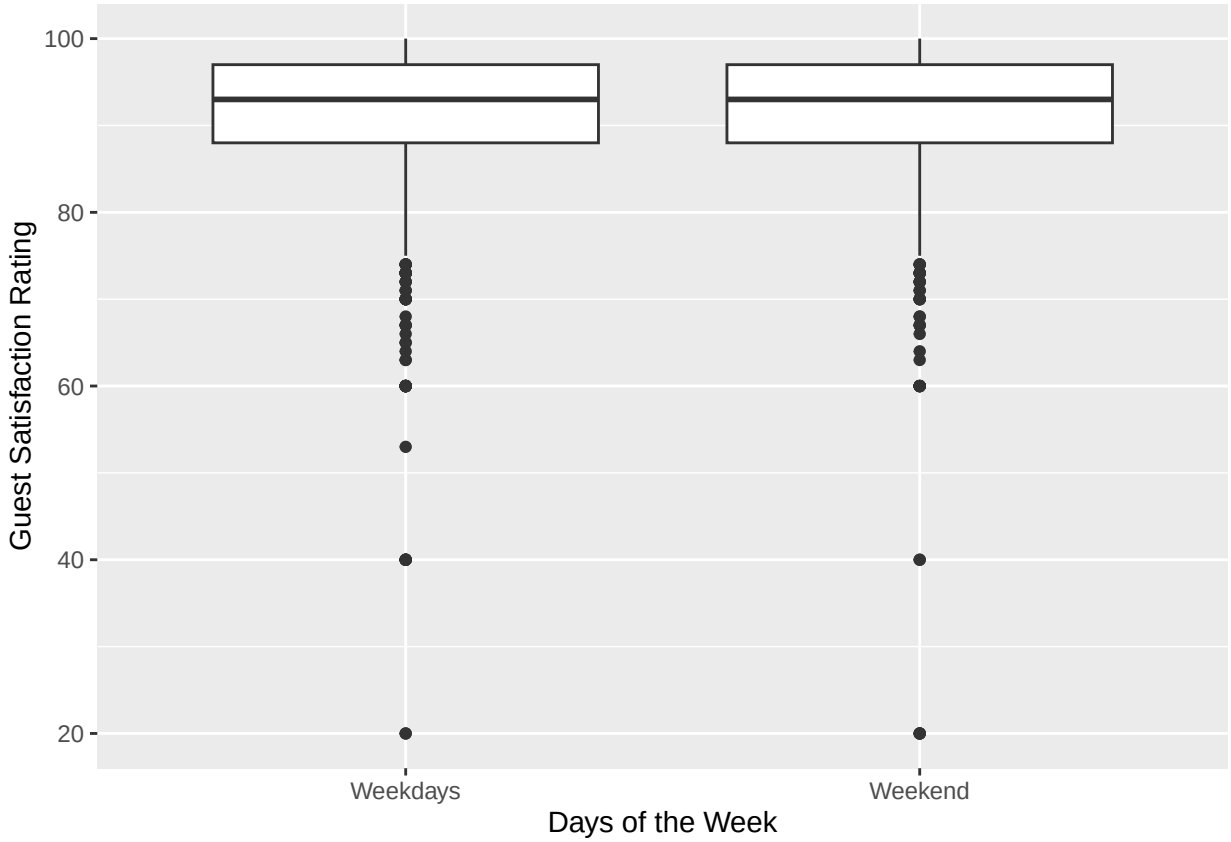
```
#10-fold cross validation

cost <- function(r, pi = 0) mean(abs(r-pi) > 0.5)
cv.glm(barcewd, m1, cost, K=10)$delta[1]
```

```
## [1] 0.07845659
```

Task 7: *Use plots or a statistical test to comment on whether the guest satisfaction varies between the weekdays and the weekends in Barcelona. Further, define a GLM that may be used to predict guest satisfaction.* [6 marks]

Using the Barcelona data set that we created in task 5, we can look at how guest satisfaction varies depending on the days of the week variable.

```
ggplot(barcelona, aes(x = days_of_week, y = guest_satisfaction_overall)) +
  geom_boxplot() + xlab("Days of the Week") + ylab("Guest Satisfaction Rating")
```



We can see from the plot that there is negligible difference between guest satisfaction at the weekend and on weekdays, they are almost the exact same. This shows no relationship between days of the week and guest satisfaction.

Now we can create a GLM to predict guest satisfaction using the Barcelona data. Because we have found from the plot above that days_of_week does not impact guest satisfaction,

we can disregard it for this model. We will not include the variables that are not useful from task 1.

```
glm1 <- glm(guest_satisfaction_overall ~ realSum + type + person_capacity
            + host + biz + cleanliness_rating + bed + dist + metro_dist +
              attr_index_norm + rest_index,family=gaussian, data=barcelona)
AIC(glm1)
```

```
## [1] 17889.42
```

Using the step function, we can refine the above model to one with reduced AIC.

```
glm2=step(glm1)
```

Below gives the refined model from the step function.

```
glm2 <- glm(guest_satisfaction_overall ~ type + host + biz + cleanliness_rating
            + bed + metro_dist + attr_index_norm,
family = gaussian, data = barcelona)
AIC(glm2)
```

```
## [1] 17883.3
```

This model can now be used to predict guest satisfaction.

Task 8: *Predict the London weekend prices for different room types based on the weekends price model for Barcelona. Calculate the prediction error and comment on what you observe.* [5 marks]

We can use a similar GLM to the one created in task 6 so we shall start by fitting that model to the Barcelona weekends data. We expect the same relationship between cleanliness rating and guest satisfaction in the weekends model too so the interaction will remain.

```
#changing variables to factors
barcewe$type <- as.factor(barcewe$room_type)
barcewe$host <- as.factor(barcewe$host_is_superhost)

g0 <- glm(log(realSum) ~ type + person_capacity + host + biz +
            cleanliness_rating*guest_satisfaction_overall + dist + bedrooms
          + metro_dist + rest_index, family=Gamma(identity), data=barcewe)
AIC(g0)
```

```
## [1] 1099.172
```

```
summary(g0)
```

From the summary, metro distance had an insignificant p-value of 0.524. The model may be better if we removed this variable.

```
g1 <- glm(log(realSum) ~ type + person_capacity + host + biz
          + cleanliness_rating*guest_satisfaction_overall + dist + bedrooms
          + rest_index, family=Gamma(identity), data=barcewe)
AIC(g1)
```

```
## [1] 1097.604
```

This is a slightly better model. Now we can predict London weekend prices for each room type using this model.

```
#factors
londonwe$type <- as.factor(londonwe$room_type)
londonwe$host <- as.factor(londonwe$host_is_superhost)

#separating the london weekend prices by room types
apt_london_we <- londonwe[londonwe$room_type == "Entire home/apt",]
priv_london_we <- londonwe[londonwe$room_type == "Private room",]
share_london_we <- londonwe[londonwe$room_type == "Shared room",]

#predicted data set of listing prices for each room type
pred_apt <- predict(g1, newdata= apt_london_we, type="response")
predicted_apt <- exp(pred_apt)

pred_priv <- predict(g1, newdata= priv_london_we, type="response")
predicted_priv <- exp(pred_priv)

pred_share <- predict(g1, newdata= share_london_we, type="response")
predicted_apt <- exp(pred_share)
```

To calculate prediction error, we can look at the London weekend data set as a whole. Using the prediction error function created in task 6, we get the following:

```
pred.err(londonwe,londonwe$realSum,g1)
```

```
## [1] 148240.3
```

From this, we can see that modelling London weekend prices using Barcelona weekend prices is not an ideal model because of the large prediction error.

Task 9: *Provide a non-scientific summary of your analysis in Task 5 (300 words maximum).* [4 marks]

Data from Airbnbs in different European cities were collected to try and determine what attributes impact the listing prices of these accommodations. Through initial exploration, certain variables were found to be unimportant attributes in determining listing prices. In further exploratory analysis, more variables were concluded as insignificant determinants.

Taking a sample from data collected in Barcelona, the aim was to explore the difference between listing prices on the weekend and during weekdays. Using a generalised linear model, it was found that the difference in weekend prices vs weekday prices was statistically significant. Other variables were included in the model and these were room type, person capacity, bedrooms, distance from the city centre and if the host had four or more listings. As an outcome, these variables' impacts on prices were discovered. Every kilometre further from the city centre that an Airbnb is situated led to a 5.38% decrease in the listing price, whether it was weekend or weekday. An extremely significant variable in the listing price was room type. Private rooms were found to be 47.33% cheaper than entire home rentals, and shared rooms were 67.04% cheaper than entire home rentals. The analysis concluded that weekend listing prices are, on average, 11.39% higher than weekday prices.