Dr. Xiyue Liao
STAT 450
Spring 2021

Hotels on Las Vegas Strip

By Noah Gallagher, Alex Gonzalez

Introduction

The data used for this project consists of 504 hotel reviews from guests that stayed at various hotels located among the Las Vegas Strip. This data was collected by the UCI Center for Machine Learning and Intelligent Systems, which was made available to us by the UCI Machine Learning Repository. In the dataset, the data consists of 21 various hotels located on the Strip, each with 24 user reviews associated with them. Each entry represents a single user review, and each hotel has 2 reviews per month which means, that the data was collected over the course of one year exactly between the months of January and August of 2015.

Questions of Interest

By analyzing the attributes and content of this dataset, we were able to develop five research questions that we plan to answer throughout this presentation. These questions include:

- Is there a relationship between a hotel’s rating and stars?
- Is there a significant difference between tourist reviews and US resident reviews for hotels?
- What can we tell about each Period of Stay and the Traveler's type?
- What amenities are normally included across all hotels?
- What are the highest and lowest correlations among hotels’ attributes?

Analysis Methods

Q1:Is there a relationship between a hotel’s rating and stars?

- We asked this question because we wanted to know if a Hotel’s stars were actually based off of User Reviews or if they were made up by the hotel, and this question will allow us to determine where Hotel’s stars actually come from. First, Conducting ANOVA tests to analyze the variables User Review Score and Hotel Stars individually. We also wanted to find the Correlation between User Review Score and Hotel Stars. And lastly, we conducted a Mahalanobis test to check the proportion of points located within our prediction ellipse.

Q2: Is there a significant difference between tourist reviews and US resident reviews for hotel?

- We asked this question because we were interested in finding out if tourist reviews were biased to Las Vegas hotels compared to US residents. First, finding the average Review Score for both Tourists and US residents. Then, visualizing the Review Scores for both groups we get confidence intervals for Review Score as well as checking the correlation between groups. And lastly, conducting ANOVA tests for both groups.

Q3: What can we tell about each Period of Stay and the Traveler's type?

- This question was asked because this can help guests determine when the best time to travel is based on their traveler type. Using R we found the most frequent period of stay, the count of traveler types for each period, and which period of stay is best for certain traveler types. We then used MANOVA testing to analyze multiple variables of interest.

Q4: What amenities are normally included across all hotel?

- This question can help guests determine what amenities to expect when visiting Las Vegas. We answered this question by finding the proportion of amenities offered across all hotels and visualizing each proportion.

Q5: What are the highest and lowest correlations among hotels' attributes?

- This can help us determine what variables are dependent on each other regarding hotel reviews. We answered this by creating a subset of all the numeric variables in the dataset. Then creating a correlation matrix and visualizing this.

Analysis

Q1:Is there a relationship between a hotel's rating and stars?

ANOVA Testing

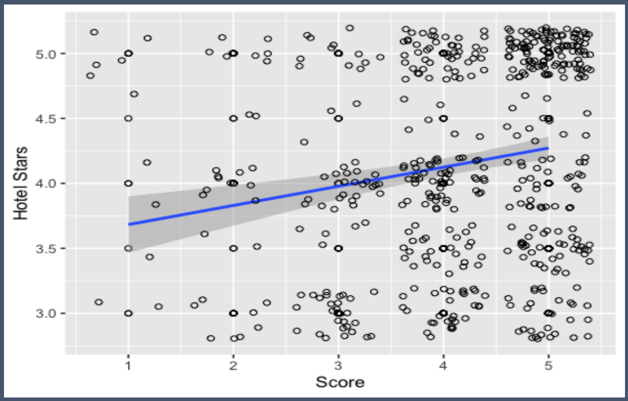
Our first method was ANOVA testing, and we picked this method because it allows us to analyze User Review Score and Hotel Stars independently, in order to determine if the variables differ among all Hotels.

We first developed a Null Hypothesis which analyzed the equality between the Mean vectors for User Review Score, as well as Hotel Stars. The decision rule was to reject the null hypothesis if our F statistic was greater than 2.1. For our Hotel Stars our F statistic was 3.74 hence for Hotel Score we determined that we will Reject the null hypothesis, meaning that all hotel ratings differ significantly among all hotels on the Las Vegas Strip.

For Hotel Stars we determined that we will also reject the null hypothesis since our F statistic is $3.44e^{28}$ meaning that all Hotel Stars differ significantly among all hotels on the Las Vegas Strip.

Correlation

Our next step focused on the correlation between User Review Score and Hotel Stars. We first found the Correlation Coefficient within R and found it to be 0.19, which indicated a weak positive correlation. This was also strengthened by creating a scatter plot between the two variables and plotting a regression curve. As you can see, our regression curve has a positive slope, which can be related to the Correlation Coefficient we found.



Mahalanobis Test

We conducted Mahalanobis testing in order to find the proportion of points within the prediction ellipse relating Hotel Stars and Hotel Ratings. Using R, we found the proportion to be 0.96, which indicates that the proportion of points within the prediction ellipse is 96%.

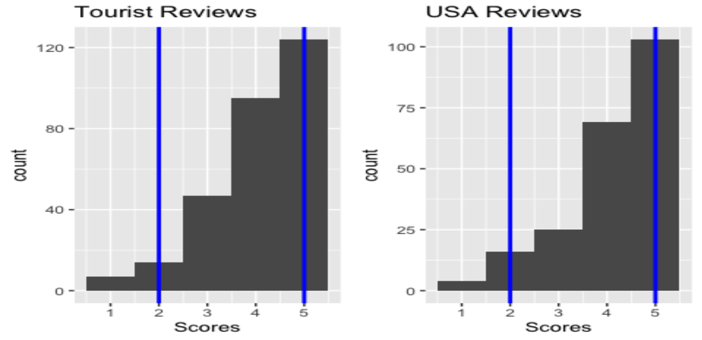
Q2: Is there a significant difference between tourist reviews and US resident reviews for hotel?

Average Review Score

Using data wrangling and manipulation tools within R, we found the Average Review Scores to be 4.098 for Tourists and 4.157 for US Residents. Both of these scores differ by only 0.059 points thus they are both about the same.

Correlation and Confidence Intervals

Next, we were interested in finding the correlation value between Tourist and US Resident Scores. We found this value to be 0.31 which is a weak positive correlation. We also were able to find Confidence Intervals for both groups and found that for both, we can be 95% positive that we can expect to see a review score between 2 and 5.



ANOVA Test

ANOVA testing was our final step, which was used in order to analyze the variable Score independently based off of the type of tourist. We did this by creating two different datasets, one with tourist reviews and one with US resident reviews. From there, we took samples of 200 reviews from each dataset and conducted ANOVA testing for both of these, grouping by Tourists and US Residents.

Our Decision Rule was to reject the null hypothesis when our F Statistic was greater than 2.5. Hence using our F Statistic of 0.023, we fail to reject the null hypothesis, thus all hotel scores do not differ significantly among all tourist types.

Q3: What can we tell about each Period of Stay and the Traveler's type?

Traveler Info

From data wrangling and manipulation, we were able to find that the busiest time of the year at Las Vegas Hotels is between March and May. Out of the different Traveler Types, which include Couples, Families, Friends, Business, and Solo we found that the most common traveler type is Couples. Lastly, we were able to find the times of year where each traveler type is the most frequent.

- The best time for Couples is between September and November.
- The best time for Families is between June and August.
- The best time for Friends is between March and May.
- The best time for Business is between December and February.
- The best time for Individuals is between June and August.

MANOVA Testing

Next step was MANOVA testing, which was used to test if the average of Review Score and the average of Hotel Stars are the same among each Period of Stay. Our Null Hypothesis was to analyze the equality of the mean vectors for Score and Stars.

The decision rule was to reject the null hypothesis when the F statistic is greater than 2.1.
Our result was that our F Statistic is 0.24 hence we fail to reject our null hypothesis, the average of Scores and the average of Stars are the same among each Period of Stay.

Q4: What amenities are normally included across all hotels?

Amenities

For the first step we counted the number of times each amenity was present for each Hotel, and then created 6 bar plots to visualize the types of Amenities offered. The second step was to determine the proportion of hotels that offered certain amenities, our findings were:

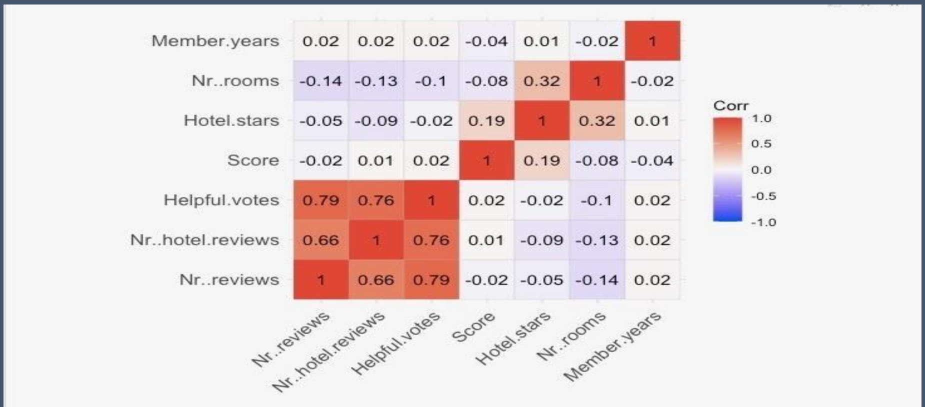
- 95% of hotels have Pools
- 95% of hotels have Gyms
- 24% of hotels have Tennis Courts
- 76% of hotels have Spas
- 90% of hotels have Casinos
- 95% of hotels have Free Internet

Q5: What are the highest and lowest correlations among hotels' attributes?

Correlations

We determined the highest correlations to be number of reviews a user has done and the number of helpful votes they have received, which has a value of 0.79. The second highest correlation is Hotel Stars and the Number of Rooms a hotel has, which has a value of 0.32.

We then determine the lowest correlations to be Hotel Stars and Number Hotel Reviews which is -0.09. The second weakest correlation is Score and Number of Rooms which is -0.08.



	Nr..reviews	Nr..hotel.reviews	Helpful.votes	Score	Hotel.stars	Nr..rooms	Member.years
Member.years	0.02	0.02	0.02	-0.04	0.01	-0.02	1
Nr..rooms	-0.14	-0.13	-0.1	-0.08	0.32	1	-0.02
Hotel.stars	-0.05	-0.09	-0.02	0.19	1	0.32	0.01
Score	-0.02	0.01	0.02	1	0.19	-0.08	-0.04
Helpful.votes	0.79	0.76	1	0.02	-0.02	-0.1	0.02
Nr..hotel.reviews	0.66	1	0.76	0.01	-0.09	-0.13	0.02
Nr..reviews	1	0.66	0.79	-0.02	-0.05	-0.14	0.02

Conclusion

We have now addressed all questions of interest for this project. To conclude our first question, our findings were that all Hotel Ratings and Hotel Stars differ significantly among the Hotels. The correlation between Hotel Rating and Stars is 0.19, which is a weak positive correlation. The proportion of points within the prediction ellipse is 96%. Thus, we can say that a Hotel's Stars are dependent on its Guest's Reviews. The average Review Score between Tourists and US residents differs by 0.059. The Confidence Interval for Review Score between Tourists and US residents is [2, 5] for both and Hotel Scores do not differ significantly among the tourist types. Thus, we can say that there is not a significant difference between tourist reviews and US resident reviews for Hotels.

For our third question we found that the most frequent Period of Stay is from March to May. The most frequent types of guests are couples and the average of Scores and the average of Stars are the same among each Period of Stay. We can conclude that the Traveler Type differs for each Period of Stay at Las Vegas Hotels. Next, looking at amenities, we found that 90% of all hotels located on the Las Vegas Strip have Pools, Gyms, Casinos and Free Internet, 75% of all hotels have Spas, and 20% of all hotels have Tennis Courts. Thus, guests can typically expect to have some type of amenity when visiting Hotels in Las Vegas.

Appendix

```
vegas <- read.csv('vegas2.csv',sep = ',')
library(tidyverse)

glimpse(vegas)

head(vegas)
```

Question 1 : Is there a relationship between a Hotel’s Ratings and Stars?

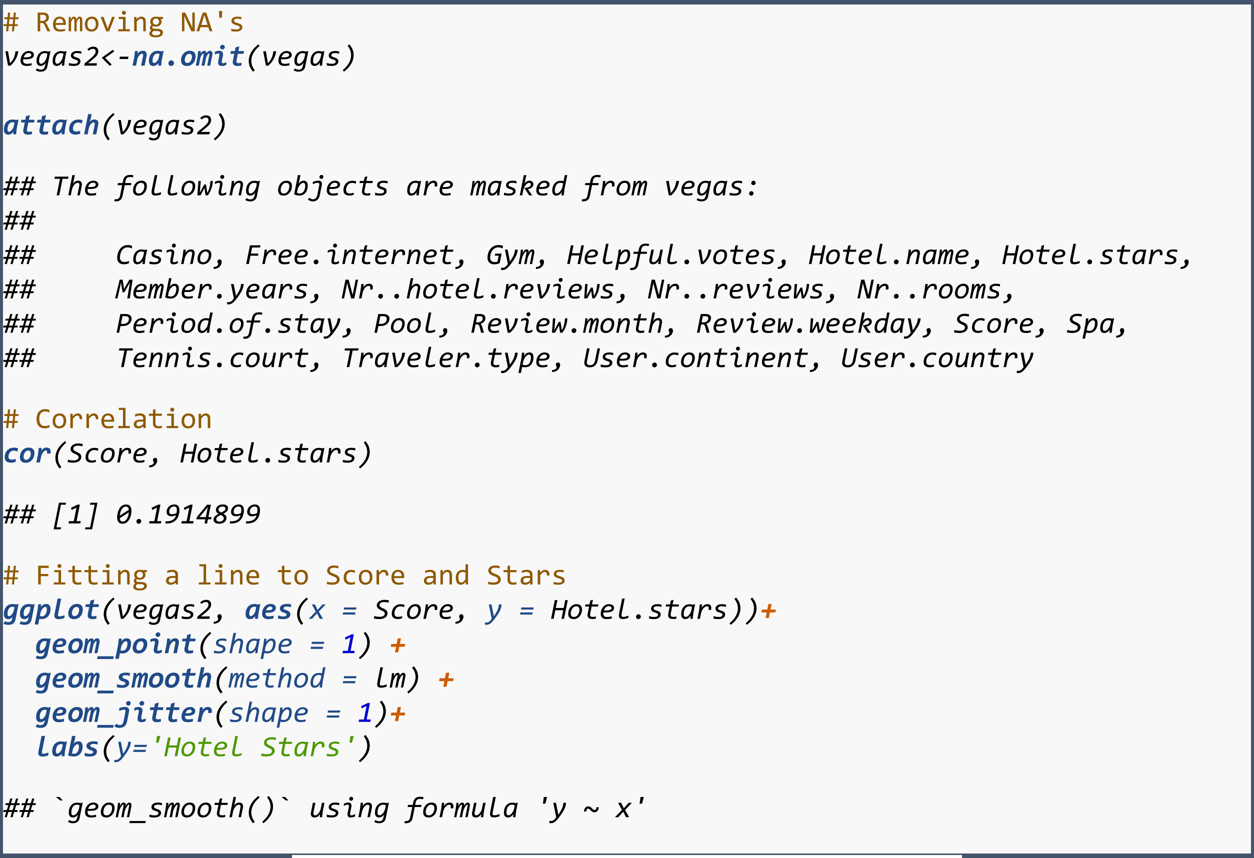
a) First let us conduct ANOVA tests to analyze each variable Score and Hotel.stars

```
attach(vegas)
grp = factor(Hotel.name)

# AOV test for Score
summary(aov(Score ~ grp))

##              Df Sum Sq Mean Sq F value   Pr(>F)
## grp          20   68.4   3.419   3.736 1.38e-07 ***
## Residuals   483  442.0   0.915
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# AOV test for Hotel.stars
summary(aov(Hotel.stars ~ grp))

##              Df Sum Sq Mean Sq   F value Pr(>F)
## grp          20  301.7   15.09 3.446e+28 <2e-16 ***
## Residuals   483    0.0    0.00
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# F stat = 2.10
g=21
p=2
N=504
alpha=0.05
qf(1-(alpha)/p, N-g, g-1)

## [1] 2.103349
```

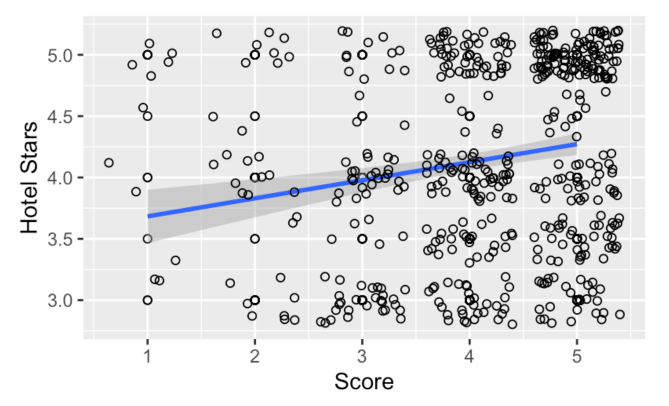
-For Hotel Score: We Reject Ho meaning, all hotel ratings differ significantly among the Hotels and there is no correlation between them.

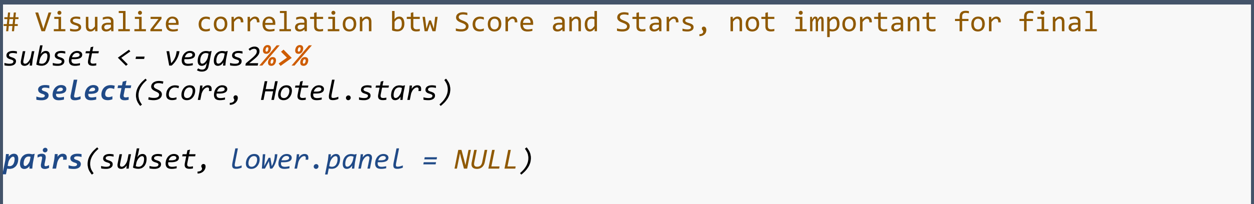
-For Hotel Stars: We Reject Ho, meaning all hotel Stars differ significantly among the Hotels.

b) Now, let us determine the correlation between Hotel Reviews and Hotel Stars.



```
# Removing NA's
vegas2<-na.omit(vegas)

attach(vegas2)

## The following objects are masked from vegas:
##
##     Casino, Free.internet, Gym, Helpful.votes, Hotel.name, Hotel.stars,
##     Member.years, Nr..hotel.reviews, Nr..reviews, Nr..rooms,
##     Period.of.stay, Pool, Review.month, Review.weekday, Score, Spa,
##     Tennis.court, Traveler.type, User.continent, User.country

# Correlation
cor(Score, Hotel.stars)

## [1] 0.1914899

# Fitting a line to Score and Stars
ggplot(vegas2, aes(x = Score, y = Hotel.stars))+
  geom_point(shape = 1) +
  geom_smooth(method = lm) +
  geom_jitter(shape = 1)+
  labs(y='Hotel Stars')

## `geom_smooth()` using formula 'y ~ x'
```

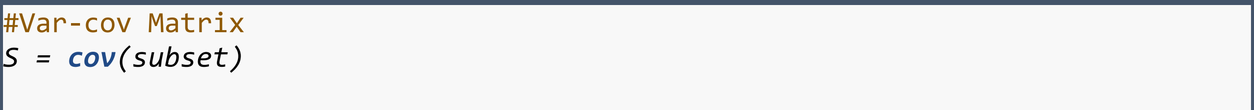




```
# Visualize correlation btw Score and Stars, not important for final
subset <- vegas2%>%
  select(Score, Hotel.stars)

pairs(subset, lower.panel = NULL)
```

The correlation is 0.19, which is a weak positive correlation.

c) Let us check proportion of points within the prediction ellipse



```
#Var-cov Matrix
S = cov(subset)
```

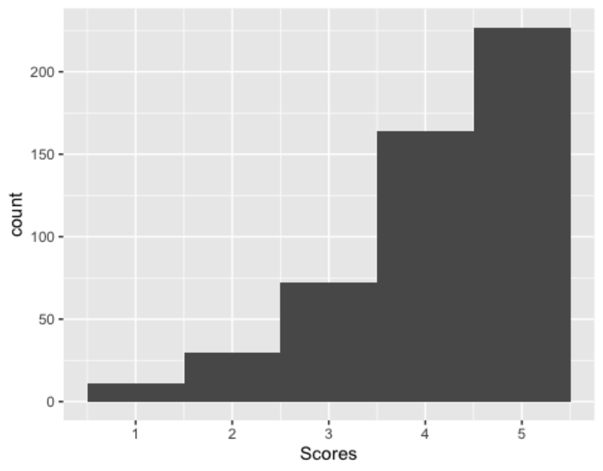
```
# Find column means
xbar = colMeans(subset)

# Mahalanobis Test
D2 = mahalanobis(subset, xbar, S)

# Variable count
p = ncol(subset)

# Proportion
sum(D2 <= qchisq(.95, p))/nrow(subset)

## [1] 0.9623016
```

Thus, the proportion of points within the prediction ellipse is 96.23%

Getting Score means for each hotel

```
#attach(vegas)

score_means<-vegas%>%
  group_by(Hotel.name)%>%
  summarize(
    avg_rating = mean(Score)
  )%>%
  arrange(desc(avg_rating))

## `summarise()` ungrouping output (override with `.groups` argument)

score_means

## # A tibble: 21 x 2
##    Hotel.name                          avg_rating
##    <chr>                                    <dbl>
##  1 Wynn Las Vegas                            4.62
##  2 The Venetian Las Vegas Hotel              4.58
##  3 Encore at wynn Las Vegas                  4.54
##  4 Marriott's Grand Chateau                  4.54
##  5 The Palazzo Resort Hotel Casino           4.38
##  6 Trump International Hotel Las Vegas       4.38
##  7 Wyndham Grand Desert                      4.38
##  8 The Cosmopolitan Las Vegas                4.25
##  9 Bellagio Las Vegas                        4.21
## 10 Tuscany Las Vegas Suites & Casino         4.21
## # … with 11 more rows

cat("The avergae score a hotel is given is: ", mean(score_means$avg_rating),
"\nThe variance across all hotel's Scores is: ", var(score_means$avg_rating))
```

```
## The avergae score a hotel is given is:  4.123016
## The variance across all hotel's Scores is:  0.1424438

#plotting scores
ggplot(vegas, aes(x=Score)) +
            geom_histogram(bins=5) +
            labs(x='Scores', y='count')
```



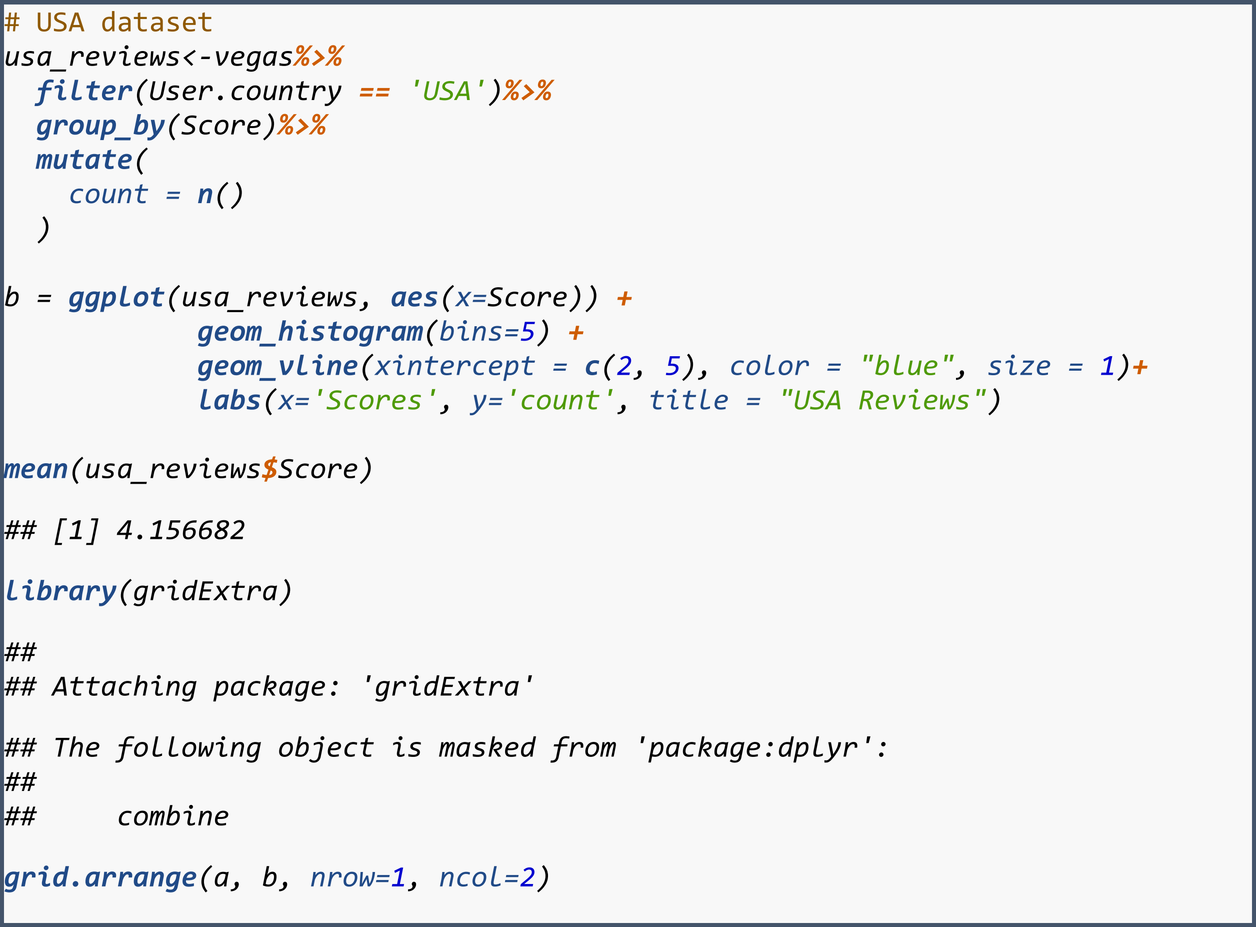
Conclusion:

-all hotel ratings differ significantly among the Hotels and there is no correlation between them. -all hotel Stars differ significantly among the Hotels. -The correlation is 0.19, which is a weak positive correlation. -The proportion of points within the prediction ellipse is 96.23%

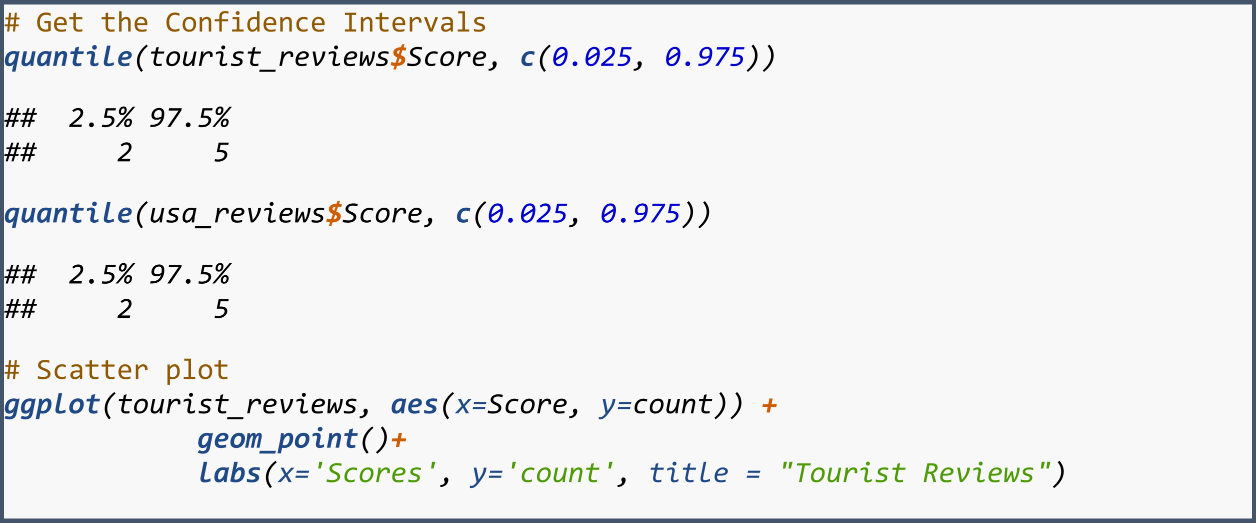
Question 2 - Can we determine if there is a significant difference between tourist reviews and USA resident reviews for hotels?

a) Plotting difference of Means

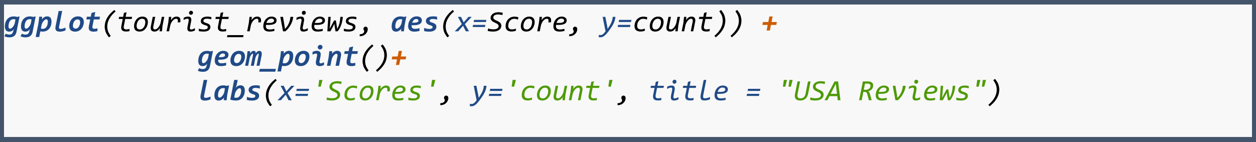
```
# Tourist Dataset
tourist_reviews<-vegas%>%
  filter(User.country != 'USA')%>%
  group_by(Score)%>%
  mutate(
    count = n()
  )

a = ggplot(tourist_reviews, aes(x=Score)) +
            geom_histogram(bins=5)+
            geom_vline(xintercept = c(2, 5), color = "blue", size = 1)+
            labs(x='Scores', y='count', title= "Tourist Reviews")

mean(tourist_reviews$Score)

## [1] 4.097561
```



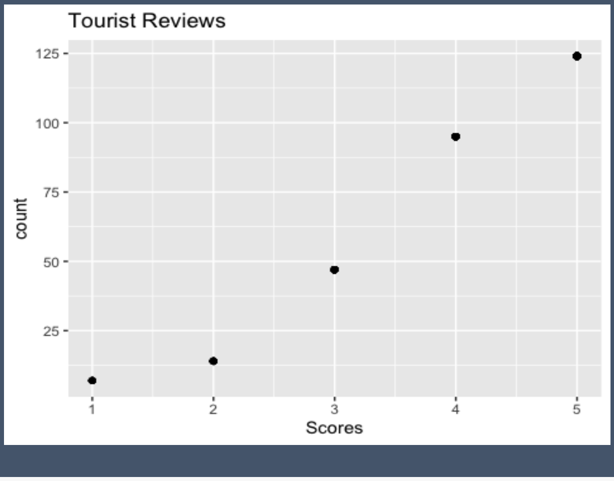
```
# USA dataset
usa_reviews<-vegas%>%
  filter(User.country == 'USA')%>%
  group_by(Score)%>%
  mutate(
    count = n()
  )

b = ggplot(usa_reviews, aes(x=Score)) +
            geom_histogram(bins=5) +
            geom_vline(xintercept = c(2, 5), color = "blue", size = 1)+
            labs(x='Scores', y='count', title = "USA Reviews")

mean(usa_reviews$Score)

## [1] 4.156682

library(gridExtra)

##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine

grid.arrange(a, b, nrow=1, ncol=2)
```

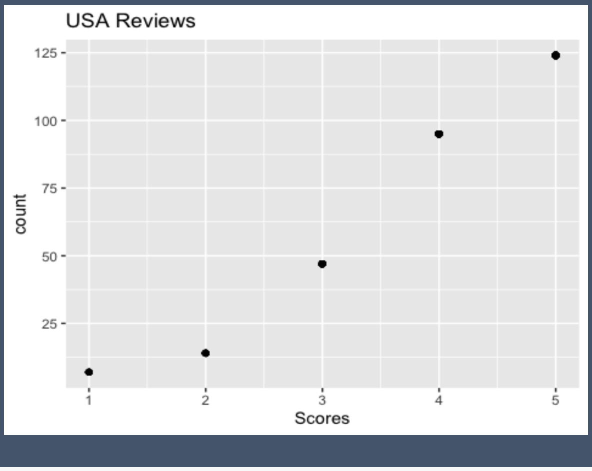


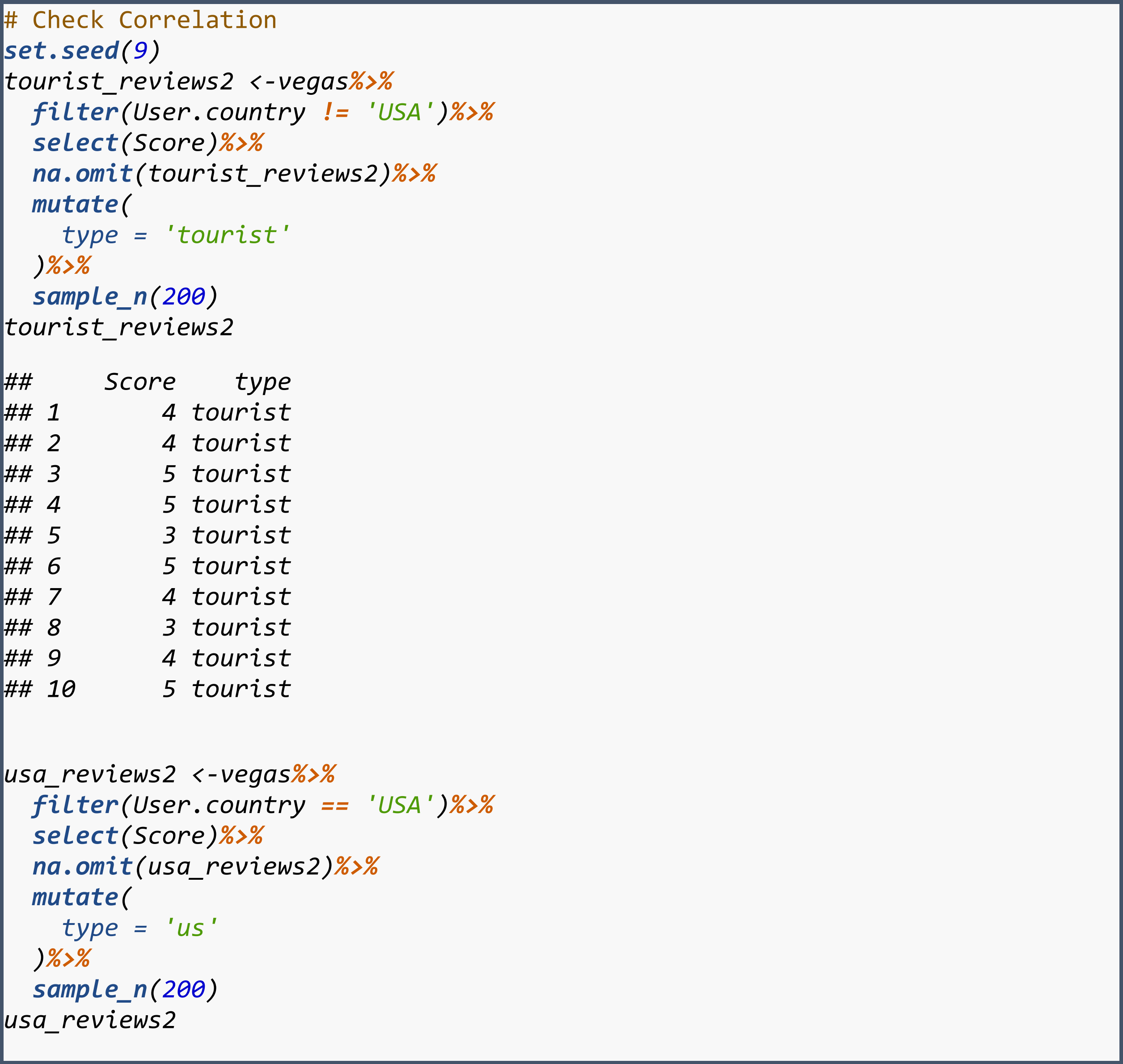
```
# Get the Confidence Intervals
quantile(tourist_reviews$Score, c(0.025, 0.975))

##  2.5% 97.5%
##     2     5

quantile(usa_reviews$Score, c(0.025, 0.975))

##  2.5% 97.5%
##     2     5

# Scatter plot
ggplot(tourist_reviews, aes(x=Score, y=count)) +
            geom_point()+
            labs(x='Scores', y='count', title = "Tourist Reviews")
```



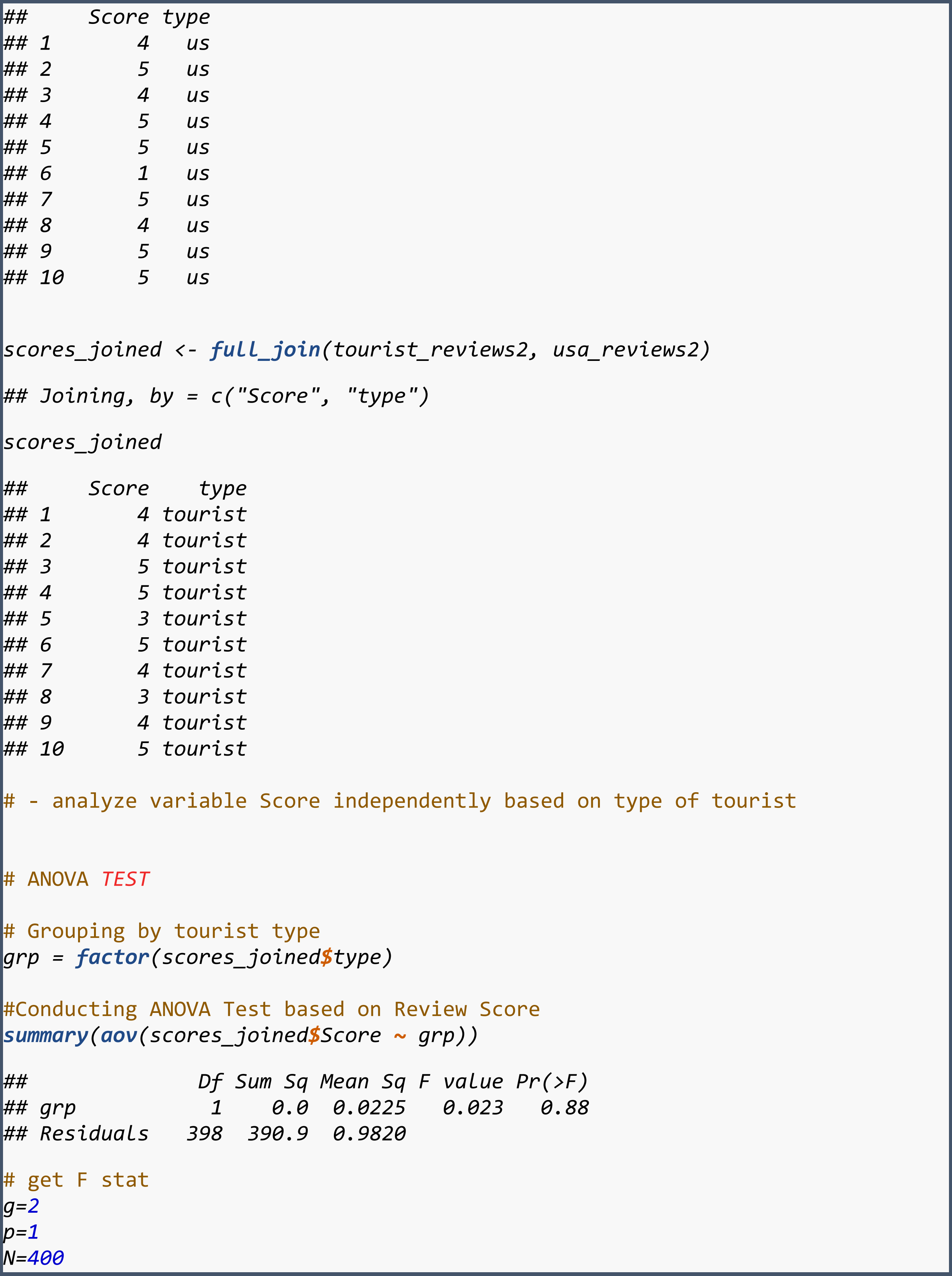
```
ggplot(tourist_reviews, aes(x=Score, y=count)) +
            geom_point()+
            labs(x='Scores', y='count', title = "USA Reviews")
```







```
# Check Correlation
set.seed(9)
tourist_reviews2 <-vegas%>%
  filter(User.country != 'USA')%>%
  select(Score)%>%
  na.omit(tourist_reviews2)%>%
  mutate(
    type = 'tourist'
  )%>%
  sample_n(200)
tourist_reviews2

##      Score    type
## 1        4 tourist
## 2        4 tourist
## 3        5 tourist
## 4        5 tourist
## 5        3 tourist
## 6        5 tourist
## 7        4 tourist
## 8        3 tourist
## 9        4 tourist
## 10       5 tourist


usa_reviews2 <-vegas%>%
  filter(User.country == 'USA')%>%
  select(Score)%>%
  na.omit(usa_reviews2)%>%
  mutate(
    type = 'us'
  )%>%
  sample_n(200)
usa_reviews2
```



```
##     Score type
## 1       4   us
## 2       5   us
## 3       4   us
## 4       5   us
## 5       5   us
## 6       1   us
## 7       5   us
## 8       4   us
## 9       5   us
## 10      5   us
```

```
scores_joined <- full_join(tourist_reviews2, usa_reviews2)
```

```
## Joining, by = c("Score", "type")
```

```
scores_joined
```

```
##     Score    type
## 1       4 tourist
## 2       4 tourist
## 3       5 tourist
## 4       5 tourist
## 5       3 tourist
## 6       5 tourist
## 7       4 tourist
## 8       3 tourist
## 9       4 tourist
## 10      5 tourist
```

```
# - analyze variable Score independently based on type of tourist


# ANOVA TEST

# Grouping by tourist type
grp = factor(scores_joined$type)

#Conducting ANOVA Test based on Review Score
summary(aov(scores_joined$Score ~ grp))
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## grp            1    0.0  0.0225   0.023   0.88
## Residuals    398  390.9  0.9820
```

```
# get F stat
g=2
p=1
N=400
```

```
alpha=0.05
qf(1-alpha, N-g, g-1)

## [1] 253.9939
```

Tourist Mean: 4.0976 Resident Mean: 4.1567

We do NOT reject HO, all hotel scores DO NOT differ significantly among the tourist type.

Question 3: Are the average of scores and average of stars are the same among each Period of Stay?

Lets first find the most frequent period of stay.

```
period_stay <-vegas%>%
  select(Period.of.stay)%>%
  group_by(Period.of.stay)%>%
  summarize(
    num_guests = n()
  )%>%
  arrange(desc(num_guests))

## `summarise()` ungrouping output (override with `.groups` argument)

period_stay

## # A tibble: 4 x 2
##   Period.of.stay num_guests
##   <chr>               <int>
## 1 Mar-May               128
## 2 Jun-Aug               126
## 3 Sep-Nov               126
## 4 Dec-Feb               124
```

Now lets use MANOVA to answer: Are the average of scores and average of stars are the same among each Period of Stay?

```
#attach(vegas)
Y = cbind(vegas$Score, vegas$Hotel.stars)
grp = factor(vegas$Period.of.stay)
fit = manova(Y ~ grp)
summary(fit, test = 'Wilks')

##              Df   Wilks approx F num Df den Df Pr(>F)
## grp           3 0.99703   0.2477      6    998 0.9603
## Residuals 500

# a = 500, b = 2, c = 2
a <- N-g-((p-g+2)/2)
b <- sqrt((p^2*(g-1)^2-4) / (p^2+(g-1)^2-5))
c<- (p*(g-1)-2)/(2)
```

```
# F stat = 13.9
g=4
p=2
N=504
alpha=0.05
qf(1-(alpha), p*(g-1), (a*b)-c)

## [1] 2.121364
```

We do not reject Ho, average of scores and average of stars are the same among each Period of Stay.

Count of travelers

```
popular_traveler <- vegas%>%
  group_by(Traveler.type)%>%
  summarize(count = n())%>%
  arrange(desc(count))

## `summarise()` ungrouping output (override with `.groups` argument)

popular_traveler

## # A tibble: 5 x 2
##   Traveler.type count
##   <chr>         <int>
## 1 Couples         214
## 2 Families        110
## 3 Friends          82
## 4 Business         74
## 5 Solo             24
```

- Most popular Traveler type is Couples.

What period of stay is best for travelers?

```
#get period per traveler type count
ppt <- vegas%>%
  group_by(Traveler.type, Period.of.stay)%>%
  summarize(count = n())%>%
  rename(
    traveler = Traveler.type,
    stay = Period.of.stay
  )%>%
  arrange(traveler)

## `summarise()` regrouping output by 'Traveler.type' (override with
`.groups` argument)

ppt
```

```
## # A tibble: 20 x 3
## # Groups:   traveler [5]
##    traveler stay    count
##    <chr>    <chr>   <int>
##  1 Business Dec-Feb    24
##  2 Business Jun-Aug    10
##  3 Business Mar-May    20
##  4 Business Sep-Nov    20
##  5 Couples  Dec-Feb    51
##  6 Couples  Jun-Aug    50
##  7 Couples  Mar-May    54
##  8 Couples  Sep-Nov    59
##  9 Families Dec-Feb    27
## 10 Families Jun-Aug    37
```

```
library(sqldf)

## Loading required package: gsubfn

## Loading required package: proto

## Loading required package: RSQLite

sqldf("SELECT distinct traveler, stay, max(count) as num_guests
       FROM ppt
       GROUP BY traveler
       ORDER BY num_guests desc")

##   traveler    stay num_guests
## 1  Couples Sep-Nov         59
## 2 Families Jun-Aug         37
## 3  Friends Mar-May         24
## 4 Business Dec-Feb         24
## 5     Solo Jun-Aug          8
```

The best time for couples is: Sep-Nov The best time for families is: Jun-Aug The best time for Friends is: Mar-May The best time for Business is: Dec-Feb The best time for individuals is: Jun-Aug

Question 4: Amenities

First lets determine What amenities are normally included for all hotels?

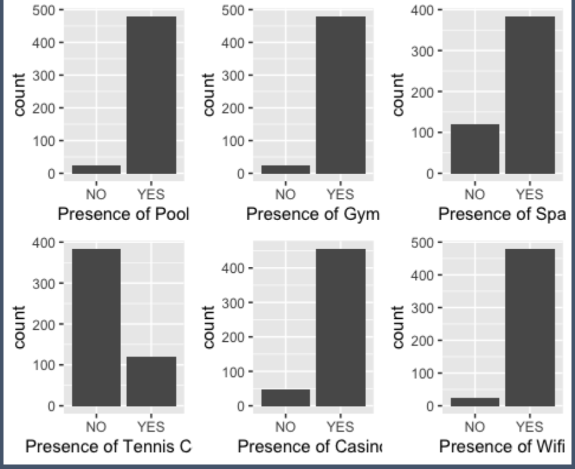
```
# Lets see what hotels have certain amenities
normal_amenities <- vegas%>%
  group_by(Hotel.name)%>%
  summarize(
    pools = sum(Pool == 'YES')/24, #we divide by 24 since there are 24
entries for each hotel
    gyms = sum(Gym == 'YES')/24,
    tennis = sum(Tennis.court == 'YES')/24,
    spas = sum(Spa == 'YES')/24,
    casinos = sum(Casino == 'YES')/24,
    wifi = sum(Free.internet == 'YES')/24
  )

## `summarise()` ungrouping output (override with `.groups` argument)

#normal_amenities

# Get the number of hotels in dataset (21 hotels)
count_hotels <- vegas%>%
  summarize(num_hotel = length(unique(Hotel.name)))
#count_hotels

# Lets get the proportions of amenities
hotel_props<- normal_amenities%>%
  summarize(
    prop_pools = sum(pools) / 21,
    prop_gym = sum(gyms) / 21,
    prop_tennis = sum(tennis) / 21,
    prop_spas = sum(spas) / 21,
    prop_casinos = sum(casinos) / 21,
    prop_wifi = sum(wifi) / 21
  )

hotel_props

## # A tibble: 1 x 6
##   prop_pools prop_gym prop_tennis prop_spas prop_casinos prop_wifi
##        <dbl>    <dbl>       <dbl>     <dbl>        <dbl>     <dbl>
## 1      0.952    0.952       0.238     0.762        0.905     0.952
```

Thus we can conclude: - 95.24% of hotels have Pools - 95.24% of hotels have Gyms - 23.81% of hotels have Tennis Courts - 76.19% of hotels have Spas - 90.48% of hotels have Casinos - 95.24% of hotels have Free Internet

Let’s plot these:

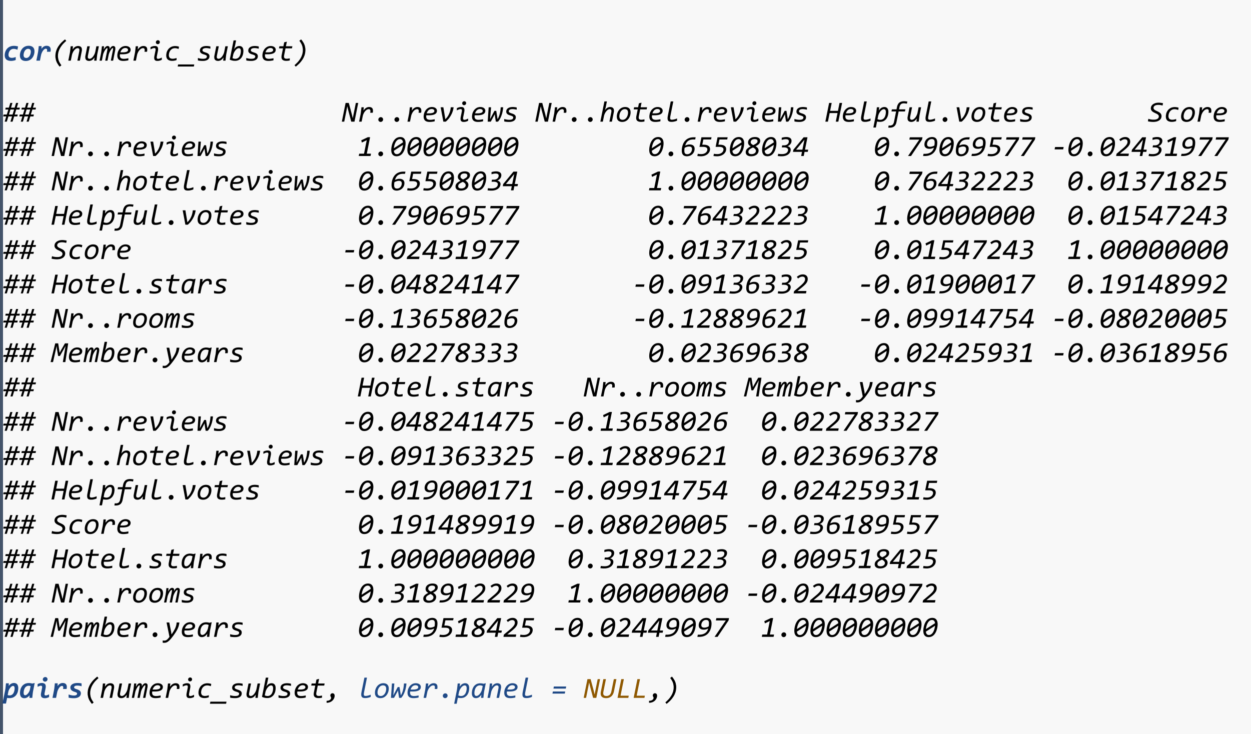
```
# Plotting Pool Presence
pool_plot = ggplot(vegas, aes(x=Pool)) +
            geom_bar() +
            labs(x='Presence of Pool', y='count')
#pool_plot

# Plotting Gym Presence
gym_plot = ggplot(vegas, aes(x=Gym)) +
            geom_bar() +
            labs(x='Presence of Gym', y='count')
#gym_plot
```

```
# Plotting Spa Presence
spa_plot = ggplot(vegas, aes(x=Spa)) +
            geom_bar() +
            labs(x='Presence of Spa', y='count')
#spa_plot

# Plotting Tennis Court Presence
tennis_plot = ggplot(vegas, aes(x=Tennis.court)) +
            geom_bar() +
            labs(x='Presence of Tennis Court', y='count')
#tennis_plot

# Plotting Casino Presence
casino_plot = ggplot(vegas, aes(x=Casino)) +
            geom_bar() +
            labs(x='Presence of Casino', y='count')
#casino_plot

# Plotting Free Wifi
wifi_plot = ggplot(vegas, aes(x=Free.internet)) +
            geom_bar() +
            labs(x='Presence of Wifi', y='count')
#wifi_plot

# Merging
grid.arrange(pool_plot, gym_plot, spa_plot, tennis_plot, casino_plot,
wifi_plot, ncol=3, nrow=2)
```

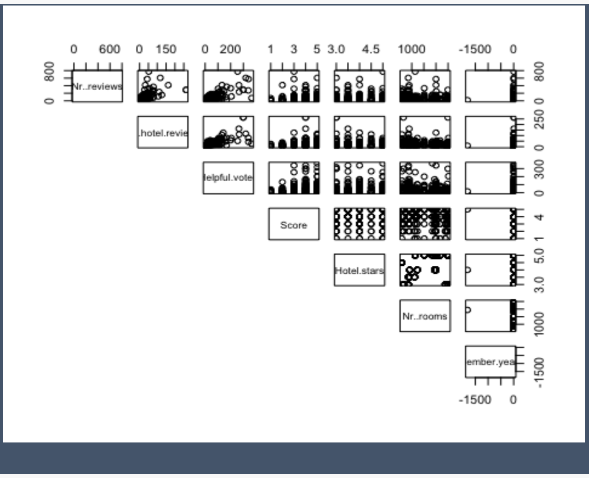


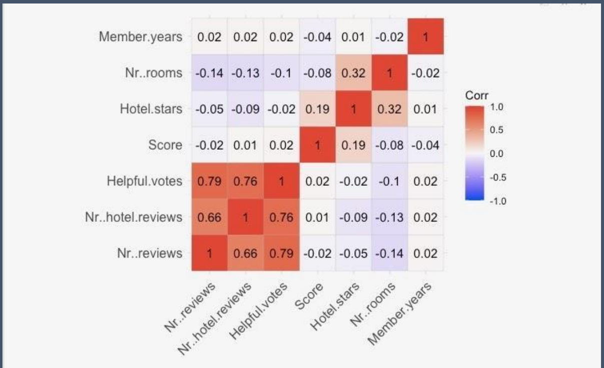
Question 5: What correaltions can we determine?

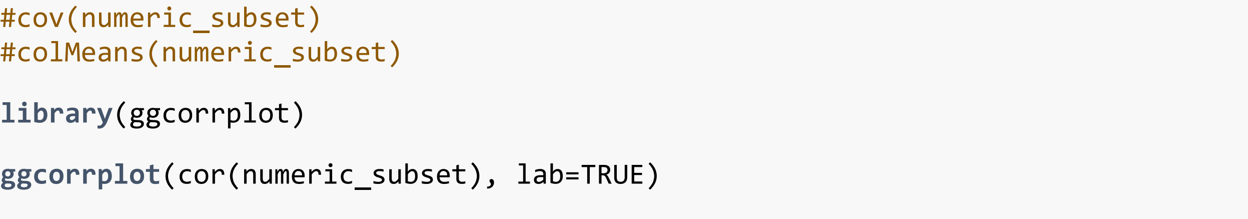
```
numeric_subset<- vegas2%>%
  select(Nr..reviews, Nr..hotel.reviews, Helpful.votes, Score, Hotel.stars,
Nr..rooms, Member.years)
```



```
cor(numeric_subset)

##                    Nr..reviews Nr..hotel.reviews Helpful.votes       Score
## Nr..reviews         1.00000000        0.65508034    0.79069577 -0.02431977
## Nr..hotel.reviews   0.65508034        1.00000000    0.76432223  0.01371825
## Helpful.votes       0.79069577        0.76432223    1.00000000  0.01547243
## Score              -0.02431977        0.01371825    0.01547243  1.00000000
## Hotel.stars        -0.04824147       -0.09136332   -0.01900017  0.19148992
## Nr..rooms          -0.13658026       -0.12889621   -0.09914754 -0.08020005
## Member.years        0.02278333        0.02369638    0.02425931 -0.03618956
##                     Hotel.stars   Nr..rooms Member.years
## Nr..reviews        -0.048241475 -0.13658026  0.022783327
## Nr..hotel.reviews  -0.091363325 -0.12889621  0.023696378
## Helpful.votes      -0.019000171 -0.09914754  0.024259315
## Score               0.191489919 -0.08020005 -0.036189557
## Hotel.stars         1.000000000  0.31891223  0.009518425
## Nr..rooms           0.318912229  1.00000000 -0.024490972
## Member.years        0.009518425 -0.02449097  1.000000000

pairs(numeric_subset, lower.panel = NULL,)
```







```
#cov(numeric_subset)
#colMeans(numeric_subset)

library(ggcorrplot)

ggcorrplot(cor(numeric_subset), lab=TRUE)
```

Highest: Number of reviews a user has done and the number of helpful votes they have received 0.79

Lowest: Hotel stars and Number of Hotel Reviews: -0.09