Las Vegas TripAdvisor Reviews

ISE 535: Data Mining
Final Report
Jack Kelly

Presentation Guide

01. Business Case

02. Data Review & Cleanup

03. EDA

04. Hypothesis Testing

05. Logistic Regression

06. Decision Tree

07. Random Forest

08. PCA

09. *Summary and Recommendations*

01. Business Case

Business Case

Dataset Background

- Dataset: *Las Vegas TripAdvisor Reviews* from the UCI Machine Learning Repository
  - https://archive.ics.uci.edu/ml/datasets/Las+Vegas+Strip
- Contents:
  - Responses and background info from raters of Las Vegas hotels
  - Information on hotels rated

Problem Statement

Online reviews from sites like TripAdvisor are a chief tool in prospective hotel customers' lodging choices. The testimonies of past customers represent perhaps the most steadfast metric of quality and value in the eyes of future customers. Consequently, our data mining team will scrutinize the dataset to derive the drivers of high scoring reviews. This will allow hotel management at companies like Caesars Entertainment and MGM Resorts to boost ratings, attract more customers, and increase revenue for their Las Vegas businesses.

Hotels

```
> unique(reviews$Hotel.name)
 [1] Circus Circus Hotel & Casino Las Vegas              Excalibur Hotel & Casino
 [3] Monte Carlo Resort&Casino                           Treasure Island- TI Hotel & Casino
 [5] Tropicana Las Vegas - A Double Tree by Hilton Hotel Caesars Palace
 [7] The Cosmopolitan Las Vegas                          The Palazzo Resort Hotel Casino
 [9] Wynn Las Vegas                                      Trump International Hotel Las Vegas
[11] The Cromwell                                        Encore at wynn Las Vegas
[13] Hilton Grand Vacations on the Boulevard             Marriott's Grand Chateau
[15] Tuscany Las Vegas Suites & Casino                   Hilton Grand Vacations at the Flamingo
[17] Wyndham Grand Desert                                The Venetian Las Vegas Hotel
[19] Bellagio Las Vegas                                  Paris Las Vegas
[21] The Westin las Vegas Hotel Casino & Spa
```

02. Initial Data Review and Cleanup

Initial Data Assessment

Initial data set is relatively clean

- No missing values
- *Score* will be the outcome - represents the reviewer's score of hotel [1-5]

No Unique Identifier

- Not an issue: reviewers need not be uniquely identified

Composition

- 6 measures (numerics)
- 14 categories (qualitative)
    - 6 of the categories appear to be logical

Convert to Logicals

Convert 6 character variables to logicals: Pool, Gym, Tennis.court, Spa, Casino, Free.internet

```
# convert character variables to logical
reviews = reviews %>% mutate(Pool = as.logical(
  case_when(
  Pool == "NO" ~ FALSE,
  Pool == "YES" ~ TRUE)))

reviews = reviews %>% mutate(Gym = as.logical(
  case_when(
    Gym == "NO" ~ FALSE,
    Gym == "YES" ~ TRUE)))

reviews = reviews %>% mutate(Tennis.court = as.logical(
  case_when(
    Tennis.court == "NO" ~ FALSE,
    Tennis.court == "YES" ~ TRUE)))

reviews = reviews %>% mutate(Spa = as.logical(
  case_when(
    Spa == "NO" ~ FALSE,
    Spa == "YES" ~ TRUE)))
```

Convert to Factors

Convert to Factors: 4 Ordered, 4 Unordered

```
# convert character variables to factors
reviews = reviews %>% mutate(Period.of.stay = as.factor(Period.of.stay))
reviews$Period.of.stay = factor(reviews$Period.of.stay, levels = c("Sep-Nov","Dec-Feb","Mar-May","Jun-Aug"))
reviews = reviews %>% mutate(Traveler.type = as.factor(Traveler.type))
reviews = reviews %>% mutate(Hotel.stars = as.factor(Hotel.stars))
reviews$Hotel.stars = factor(reviews$Hotel.stars, levels = c("3","3,5","4","4,5", "5"))
reviews = reviews %>% mutate(User.continent = as.factor(User.continent))
reviews = reviews %>% mutate(Review.month = as.factor(Review.month))
reviews$Review.month = factor(reviews$Review.month, levels = c("January","February","March","April","May","Ju
reviews = reviews %>% mutate(Review.weekday = as.factor(Review.weekday))
reviews$Review.weekday = factor(reviews$Review.weekday, levels = c("Monday","Tuesday","Wednesday","Thursday",
reviews = reviews %>% mutate(User.country = as.factor(User.country))
reviews = reviews %>% mutate(Hotel.name = as.factor(Hotel.name))
```

Initial Data Summary

Summary of Numerics

```
> summarize_numeric(reviews)
          Attribute Missing Values Unique Values         Mean   Min  Max          SD
1        Nr..reviews              0           139   48.1309524     1  775   74.996426
2 Nr..hotel.reviews              0            64   16.0238095     0  263   23.957953
3      Helpful.votes              0           109   31.7519841     0  365   48.520783
4              Score              0             5    4.1230159     1    5    1.007302
5          Nr..rooms              0            21 2196.3809524   188 4027 1285.476807
6       Member.years              0            15    0.7678571 -1806   13   80.692897
```

```
> reviews %>% select(Member.years) %>% arrange(Member.years)
  Member.years
1        -1806
2            0
```

Only one row has a value below zero, so we can assume this is an outlier. Remove.

Additional Conversions

- Convert Hotel Stars to numerics
- Assume “3,5” and “4,5” means 3.5 and 4.5 stars respectively

```
# Convert Hotel Stars to Numeric
reviews = reviews %>% mutate(Hotel.stars = case_when(Hotel.stars == "3,5" ~ 3.5,
                                                     Hotel.stars == "4,5" ~ 4.5,
                                                     Hotel.stars == "3" ~ 3,
                                                     Hotel.stars == "4" ~ 4,
                                                     Hotel.stars == "5" ~ 5))
str(reviews)
```

Summary After Conversion

6 Measures
Interval Numeric: *Nr.reviews, Score, Nr.rooms*
Ratio Numeric: *Nr.hotel.reviews, Helpful.votes, Member.years*

14 Categories
6 Nominal Factors: *Period.of.stay, Traveler.type, Hotel.name, User.continent, User.country, Review.month, Review.weekday*
1 Ordinal Factor: *Hotel.stars*
6 Logicals: *Pool, Gym, Tennis.court, Spa, Casino, Free.internet*

Summary After Conversion

```
> str(reviews)
'data.frame':	503 obs. of  20 variables:
 $ User.country     : Factor w/ 48 levels "Australia","Belgium",..: 48 48 48 46 4 4 46 48 18 4 ...
 $ Nr.reviews       : int  11 119 36 14 5 31 45 2 24 12 ...
 $ Nr.hotel.reviews: int  4 21 9 7 5 8 12 1 3 7 ...
 $ Helpful.votes    : int  13 75 25 14 2 27 46 4 8 11 ...
 $ Score            : int  5 3 5 4 4 3 4 4 4 3 ...
 $ Period.of.stay   : Factor w/ 4 levels "Sep-Nov","Dec-Feb",..: 2 2 3 3 3 3 3 3 3 3 ...
 $ Traveler.type    : Factor w/ 5 levels "Business","Couples",..: 4 1 3 4 5 2 2 3 4 3 ...
 $ Pool             : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ Gym              : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
 $ Tennis.court     : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ Spa              : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ Casino           : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
 $ Free.internet    : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
 $ Hotel.name       : Factor w/ 21 levels "Bellagio Las Vegas",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ Hotel.stars      : Factor w/ 5 levels "1","2","3","4",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ Nr.rooms         : int  3773 3773 3773 3773 3773 3773 3773 3773 3773 3773 ...
 $ User.continent   : Factor w/ 6 levels "Africa","Asia",..: 4 4 4 3 4 4 3 4 2 4 ...
 $ Member.years     : int  9 3 2 6 7 2 4 0 3 5 ...
 $ Review.month     : Factor w/ 12 levels "January","February",..: 1 1 2 2 3 3 4 4 5 5 ...
 $ Review.weekday   : Factor w/ 7 levels "Monday","Tuesday",..: 4 5 6 5 2 2 5 2 6 2 ...
```

Grouping of Attributes

Numeric(6)

Customer(3)

- Nr.reviews
- Member.years
- Helpful.votes

Hotel(3)

- Nr.rooms
- Nr.hotel.reviews
- Score

Factors(14)

Customer(6)

- User.Continent
- User.country
- Period.of.stay
- Traveler.type
- Review.month
- Review.weekday

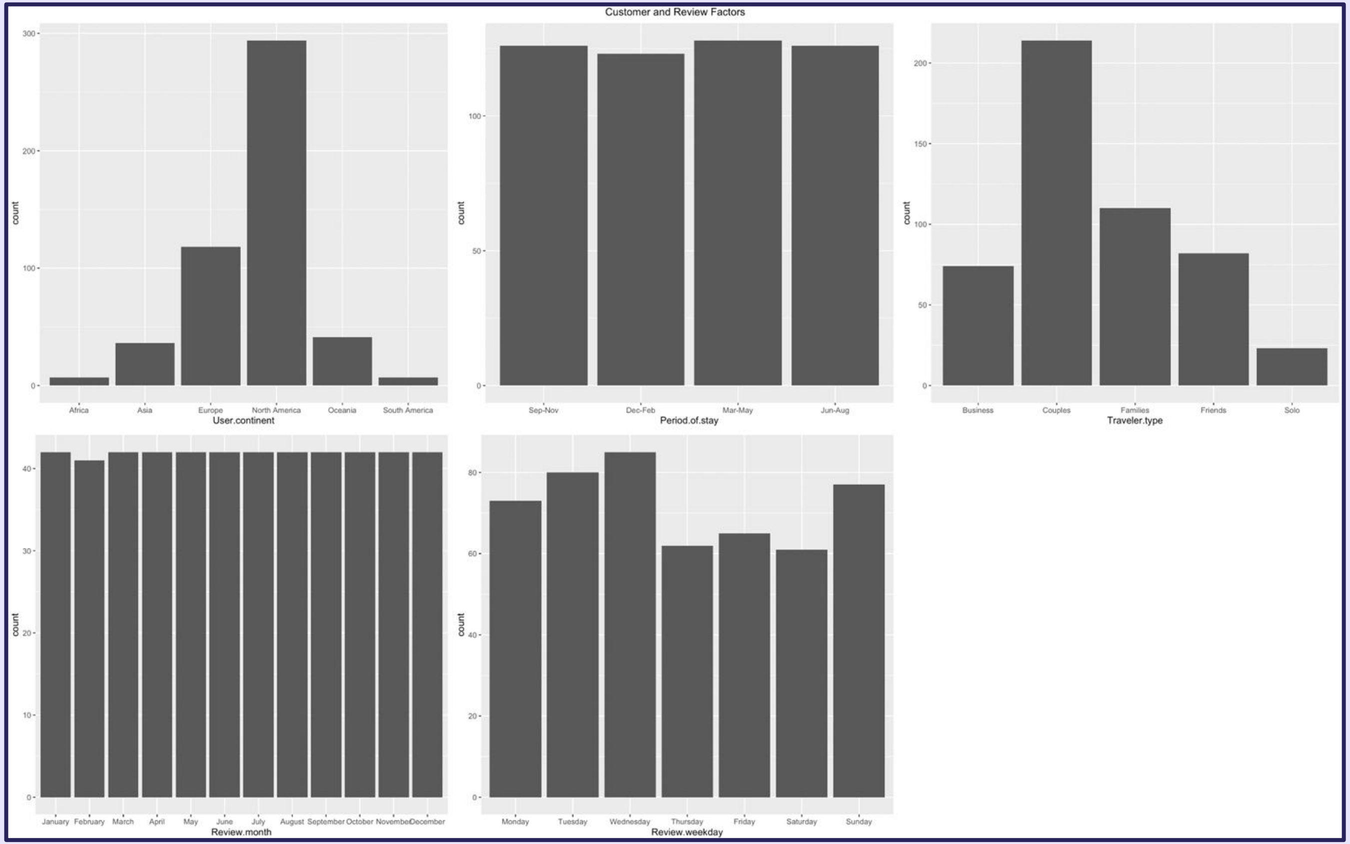
Hotel(8)

- Hotel.name
- Hotel.stars
- Pool
- Gym
- Tennis.court
- Spa
- Casino
- Free.internet

03. Exploratory Data Analysis

Univariate Analysis

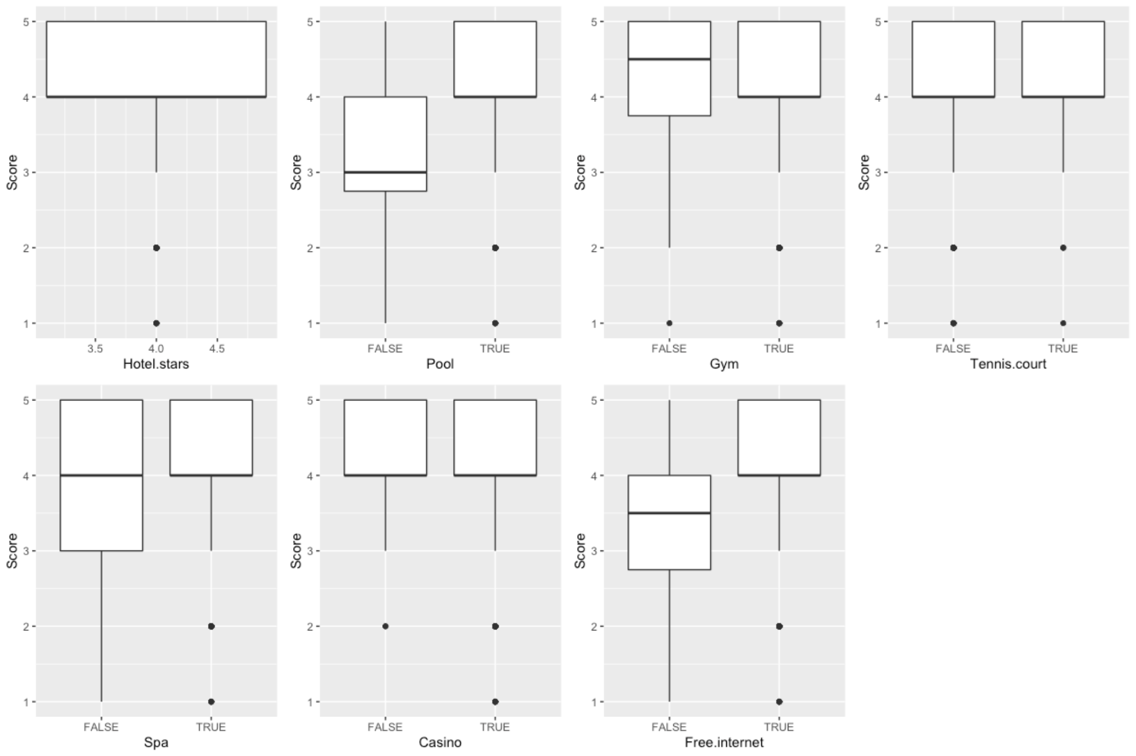
Univariate Summary of Factors - Customer and Review



- Most Travelers are from North America and Europe.
  - Business travelers are minority.
- Travel time, review time distribute uniformly

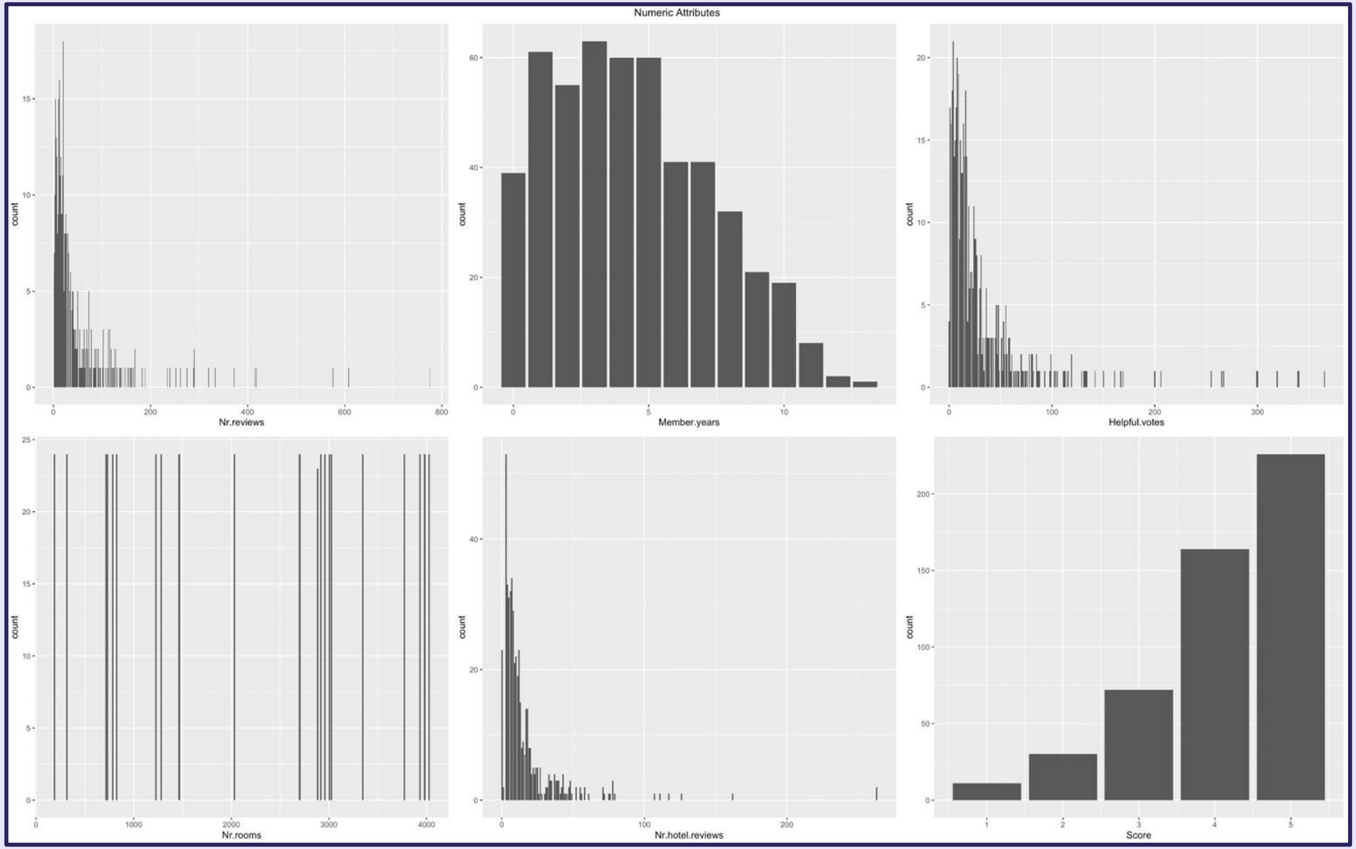
Univariate Analysis

Univariate Summary of Factors - Hotel



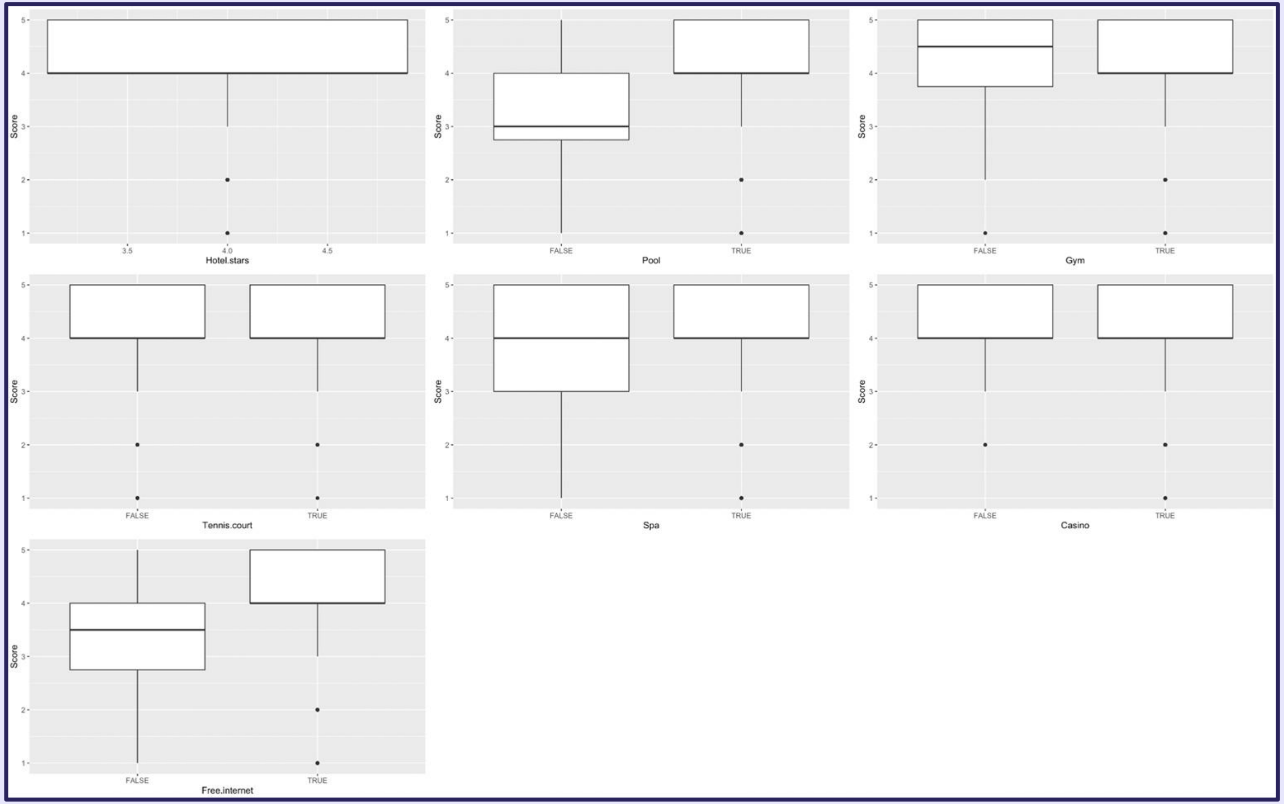
Univariate Analysis

Univariate Summary of Numeric Attributes



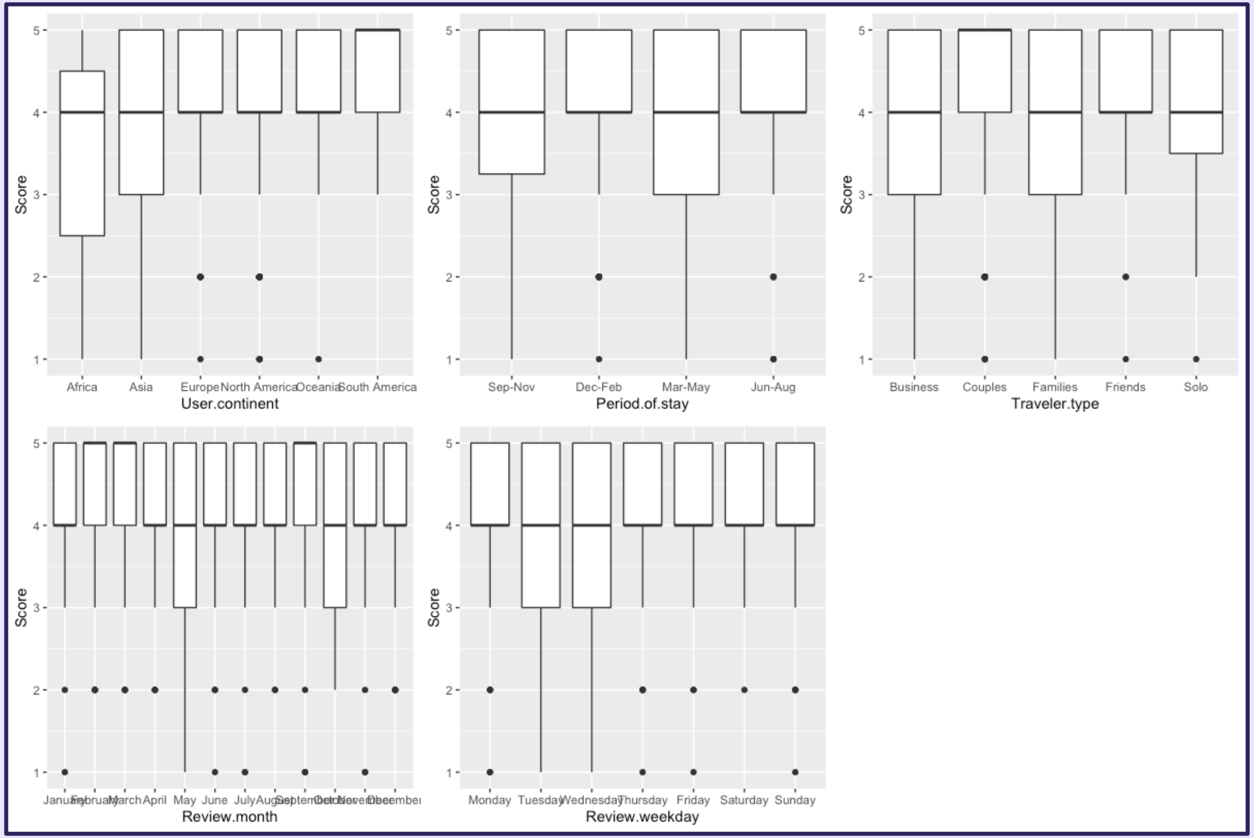
- Majority of reviewers' reviews are less than 100, helpful votes less than 50.
- Number of hotel rooms range from 200 to 400, distributed uniformly.
- Hotel scores range from 1 to 5. score and number of hotels is numerically positively related.

Bivariate Analysis



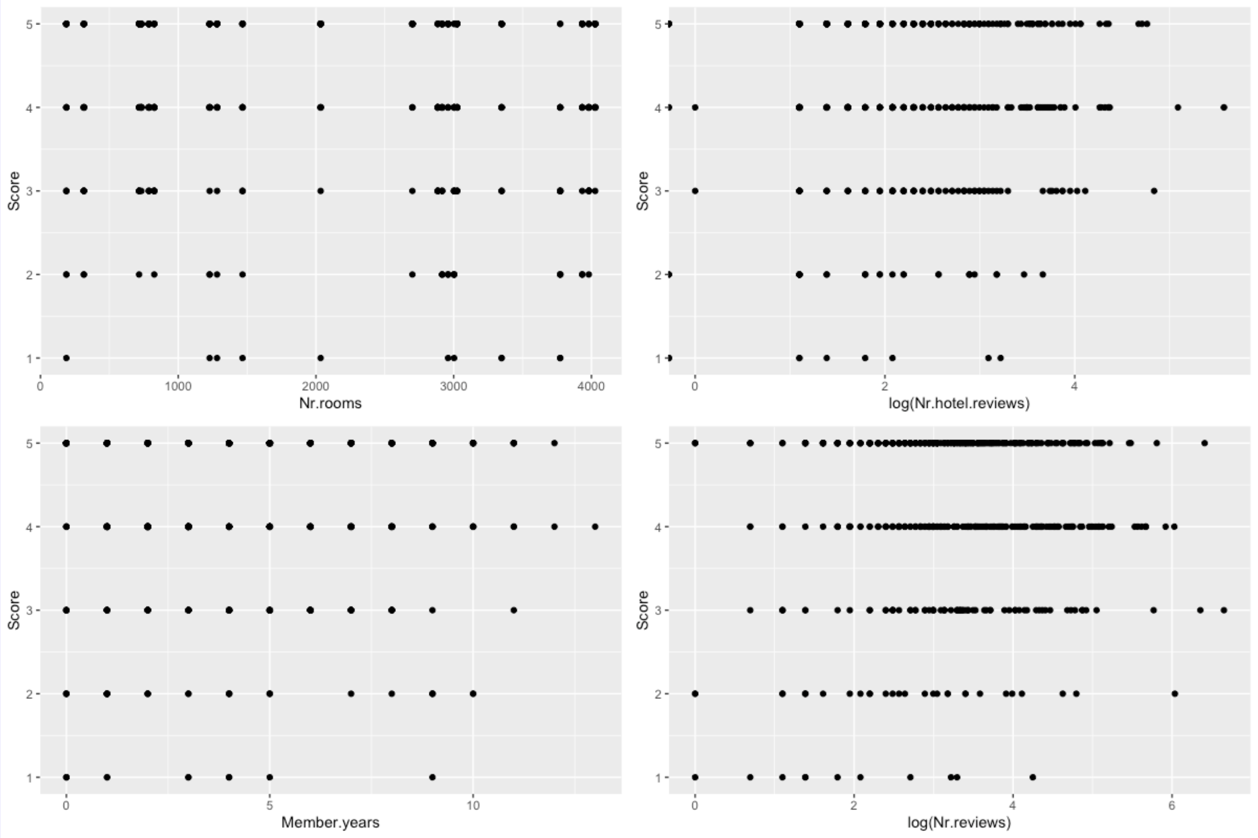
- Pool, Spa and Free internet have impact on hotel's rating.

Bivariate Analysis



- Africa users' reviews tend to give lower scores than users from other continents.
- Reviews given in Spring and Autumn tend to have more fluctuation downward.
- Travelers of type couples and friends tend to give higher scores than other type of travelers.
- In May and October, Tuesday and Wednesday, review tend to have more lower scores.

Bivariate Analysis



- These four factors appears not to have influence on score.

04. Hypothesis Testing

Setup and Bootstrapping

```r
good.reviews <- reviews %>% filter(Score >= 4)
bad.reviews <- reviews %>% filter(Score < 4)
```

```r
sample_means = rep(0, 1000)

for (i in 1:1000) {
  sample_means[i] = mean(sample(good.reviews$Nr.hotel.reviews,
                                length(good.reviews$Nr.hotel.reviews), replace = TRUE))
}
```

```r
library(BSDA)
sample_means_bad <- sample_means_bad %>% as_vector()
sample_means_good <- sample_means_good %>% as_vector()
z.test(x = sample_means_bad, y = sample_means_good,
       alternative = "two.sided", sigma.x = sd(sample_means_bad), sigma.y = sd(sample_means_good))
```

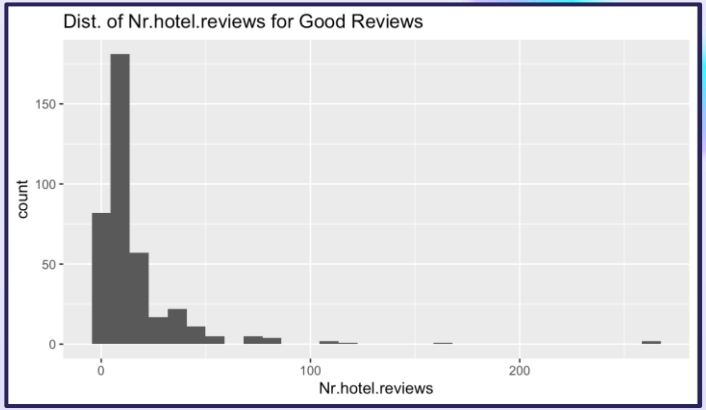
Difference in Means

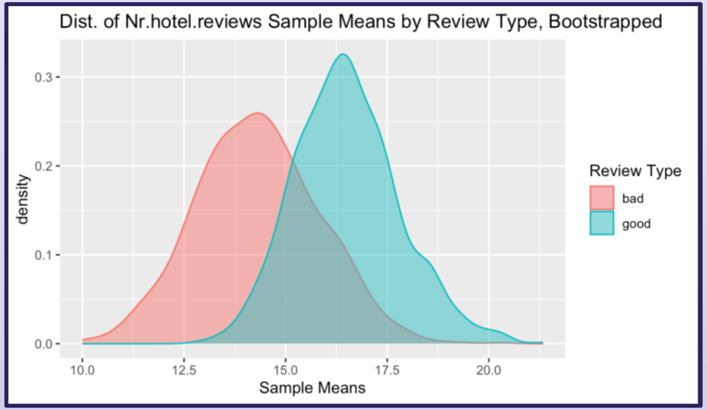
H0: There is *no* significant difference in the mean number of hotel reviews left by those who give a "bad" review [0-3] than by those who leave a "good" review [4-5].

HA: There *is* a significant difference in the mean number of hotel reviews left by those who give a "bad" review [0-3] than by those who leave a "good" review [4-5].

```
        Two-sample z-Test

data:  sample_means_bad and sample_means_good
z = -31.775, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.256496 -1.994296
sample estimates:
mean of x mean of y
 14.37752  16.50292
```

Two-sided z-test for difference in means





Distributions of 1000 bootstrapped sample means for each review type

05. Logistic Regression

Linear Regression

Response Variable: Score

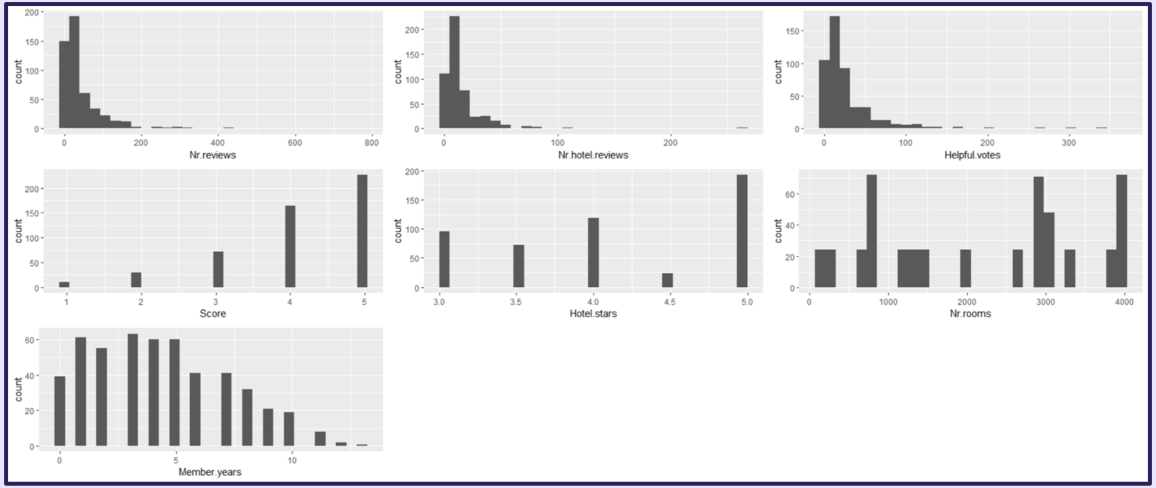
Score given to the hotel by the reviewer

```
> head(reviews)
# A tibble: 6 x 20
  User.country Nr.reviews Nr.hotel.reviews Helpful.votes Score Period.of.stay Traveler.type Pool  Gym
  <fct>             <dbl>            <dbl>         <dbl> <dbl> <fct>          <fct>         <lgl> <lgl>
1 USA                  11                4            13     5 Dec-Feb        Friends       FALSE TRUE
2 USA                 119               21            75     3 Dec-Feb        Business      FALSE TRUE
3 USA                  36                9            25     5 Mar-May        Families      FALSE TRUE
4 UK                   14                7            14     4 Mar-May        Friends       FALSE TRUE
5 Canada                5                5             2     4 Mar-May        Solo          FALSE TRUE
6 Canada               31                8            27     3 Mar-May        Couples       FALSE TRUE
# ... with 11 more variables: Tennis.court <lgl>, Spa <lgl>, Casino <lgl>, Free.internet <lgl>, Hotel.name <fct>,
#   Hotel.stars <fct>, Nr.rooms <dbl>, User.continent <fct>, Member.years <dbl>, Review.month <fct>,
#   Review.weekday <fct>
```

Objective: To better understand what attributes impact a reviewer's score

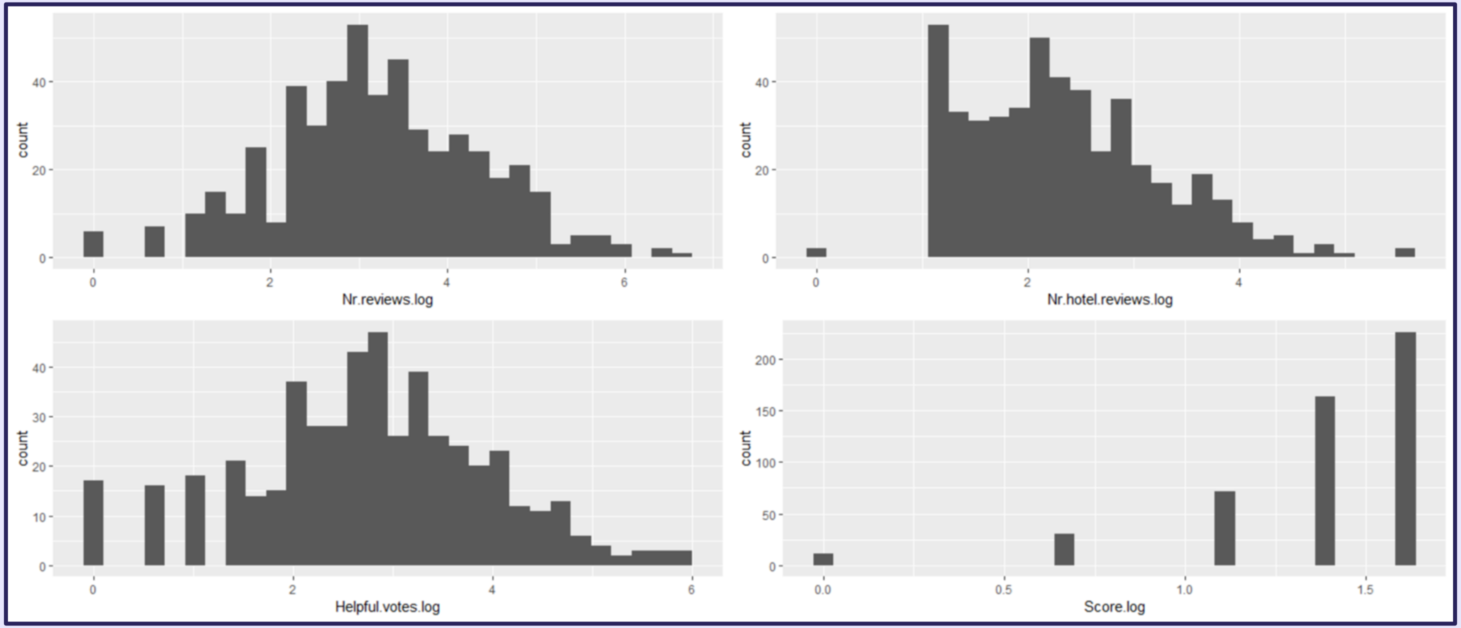
Check Distributions

- We want to check that individual variables have a somewhat normal, not heavily skewed distribution



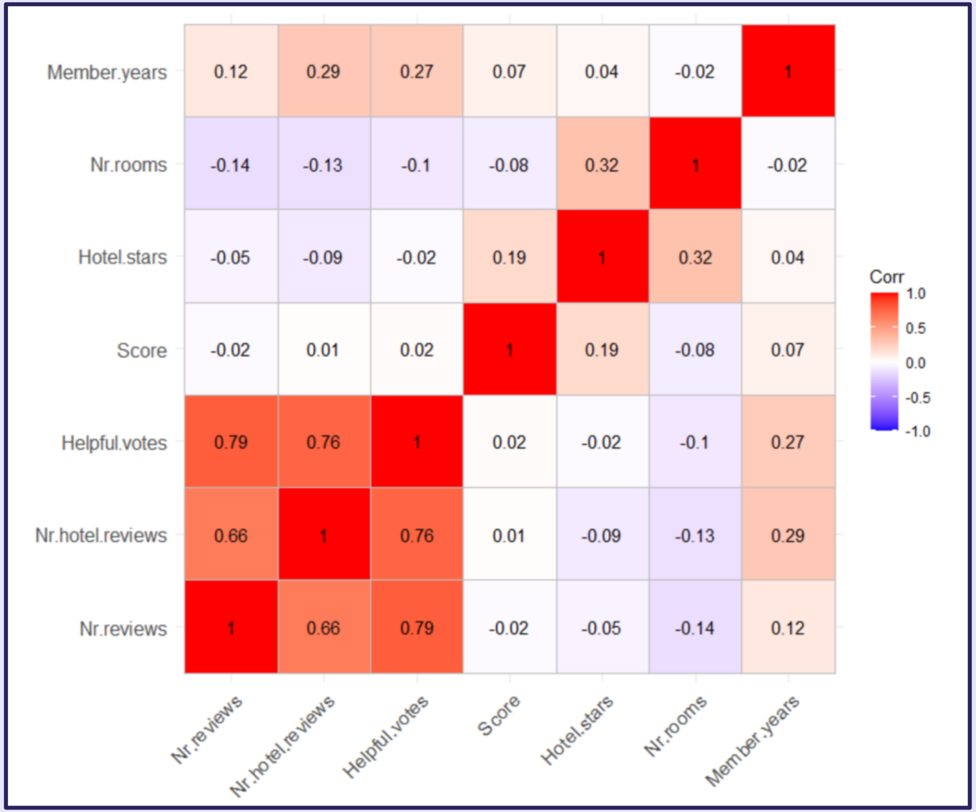
- Nr.reviews, Nr.hotel.reviews, Helpful.votes, Score all have heavy skews
- We will apply a logarithmic transformation

Apply Logarithmic Transformation



- Nr.reviews, Nr.hotel.reviews, and Helpful.votes look good
- Score (our response variable) still heavily skewed.

Inspect Correlations



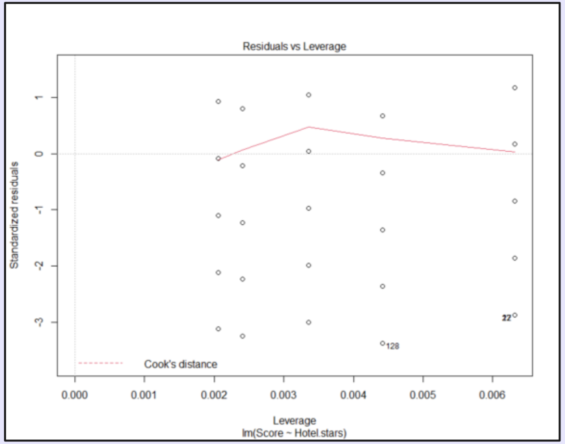
- Some correlation between reviewer attributes, but not high enough to cause concern (< 0.9)

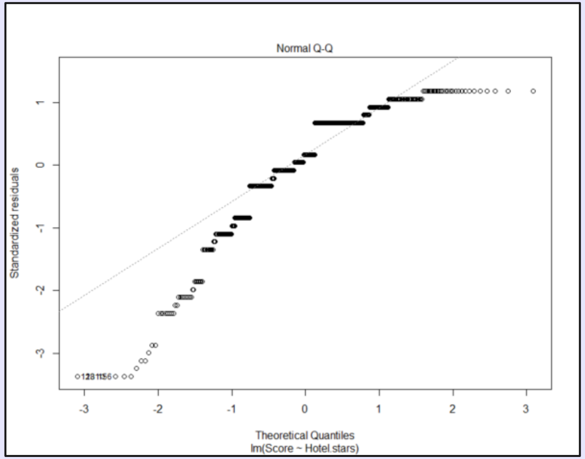
Initial Model

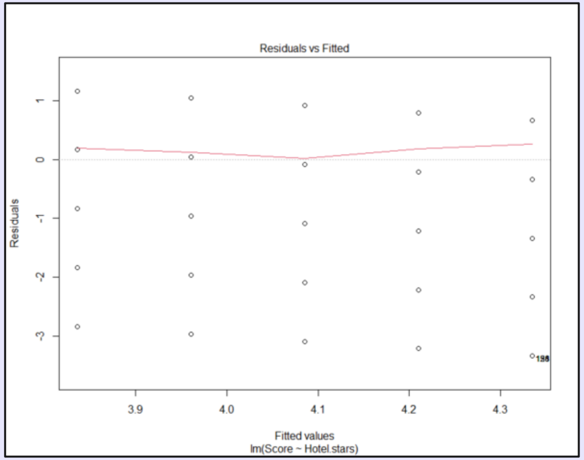
```
> # Create initial model
> m1 <- lm(Score ~ Hotel.stars, data = reviews_numeric)
> summary(m1)

Call:
lm(formula = Score ~ Hotel.stars, data = reviews_numeric)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.3350 -0.3350  0.1639  0.6650  1.1639 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.08762    0.24019  12.855  < 2e-16 ***
Hotel.stars  0.24949    0.05699   4.378 1.46e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9898 on 501 degrees of freedom
Multiple R-squared:  0.03685,	Adjusted R-squared:  0.03493 
F-statistic: 19.17 on 1 and 501 DF,  p-value: 1.458e-05

> confint(m1)
                2.5 %    97.5 %
(Intercept) 2.6157200 3.5595133
Hotel.stars 0.1375269 0.3614451
```

- Roughly centered around zero
- P value looks good
- We can say with strong confidence that this coefficient (Hotel.stars) does influence the response variable (Score)

Initial Model







- Slight pattern in residuals - calls into question independence assumption
- Some deviation at the extremes of the Q-Q plot, especially on the left side. So we may not be close to normality.
- So let's try something else

Logistic Regression Model Using Hotel Amenities as Predictors

- Convert response variable (Score) to a binary and run logistic regression
  - 1-3 = “Bad” = 0
  - 4-5 = “Good” = 1

```
> summary(logistic_model_two)

Call:
glm(formula = score_bi ~ Pool + Gym + Tennis.court + Spa + Casino +
    Free.internet, family = "binomial", data = reviews)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5211   0.2918   0.6647   0.6647   1.2491

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -5.5629     2.2066  -2.521  0.01170 *
PoolTRUE            3.3026     1.1004   3.001  0.00269 **
GymTRUE             2.0369     1.1248   1.811  0.07017 .
Tennis.courtTRUE    0.0790     0.2792   0.283  0.77722
SpaTRUE            -1.7381     1.0324  -1.684  0.09227 .
CasinoTRUE          1.9615     1.0861   1.806  0.07092 .
Free.internetTRUE   1.3974     0.4353   3.210  0.00133 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 535.93  on 502  degrees of freedom
Residual deviance: 507.52  on 496  degrees of freedom
AIC: 521.52
```

Odds Ratios

```
> exp(coef(logistic_model_two))
    (Intercept)        PoolTRUE         GymTRUE  Tennis.courtTRUE         SpaTRUE      CasinoTRUE Free.internetTRUE
    0.003837707    27.181815787     7.666665991       1.082200423     0.175854716     7.110319296       4.044658121
```

- We see that both Pool and Free Internet have a statistically significant impact on the probability of a good review

06. Decision Tree

Decision Tree Setup

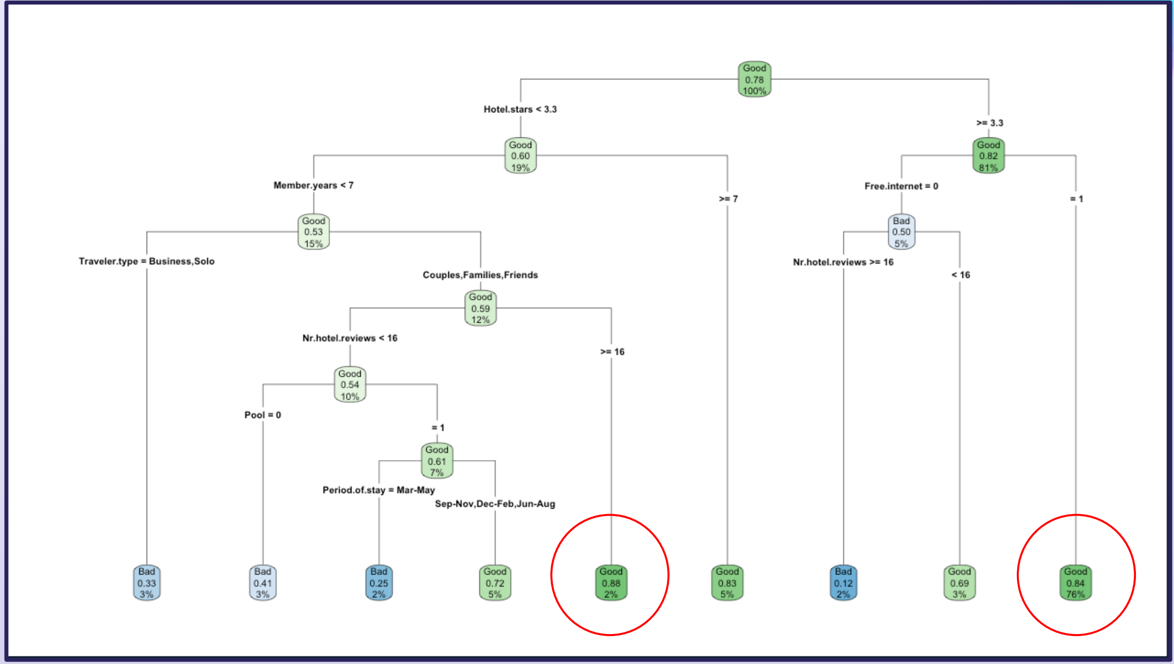
- Separate *Score* into “bad” and “good” reviews.
- Omit several variables: *Hotel.name, Nr. rooms, Review.month, Review.weekday, Score, User.country, Nr.reviews, Helpful.votes*

```
reviews.tree <- select(reviews.tree, -c(Hotel.name, Nr.rooms, Review.month, Review.weekday,
                                         Score, User.country, Nr.reviews, Helpful.votes))

tree <- rpart(review.type ~ ., data = reviews.tree, method = "class")
```

Tree and Variable Importance

- Several unused variables
- Differing factors for high and low star hotels.
  - 4-5 Star
    - Free internet!
  - 1-3 Star
    - More complicated

```
Root node error: 113/503 = 0.22465

n= 503

        CP nsplit rel error xerror     xstd
1 0.019469      0   1.00000 1.0000 0.082834
2 0.010000      8   0.84071 1.0973 0.085540
```




```
> table(pred, reviews.tree$review.type)

pred   Bad Good
  Bad   33   15
  Good  80  375
```

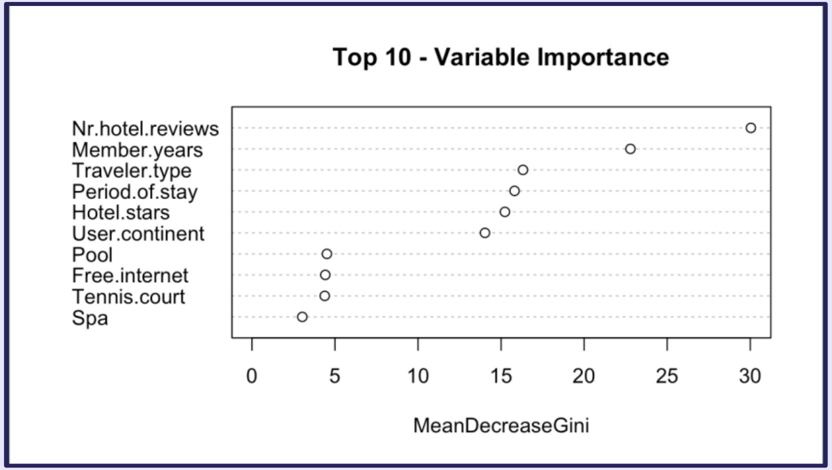
```
> tree$variable.importance
     Hotel.stars Nr.hotel.reviews    Free.internet     Member.years   Period.of.stay
       6.9532392        5.7910790        5.0845709        2.9792039        2.6775758
            Pool    Traveler.type   User.continent              Spa
       2.5854399        1.5241694        0.8861026        0.1993247
```

07. Random Forest

Forest Output

```
forest <- randomForest(formula = review.type ~ ., data = reviews.tree)
```

```
> forest

Call:
 randomForest(formula = review.type ~ ., data = reviews.tree)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 23.86%
Confusion matrix:
     Bad Good class.error
Bad   13  100  0.88495575
Good  20  370  0.05128205
```



08. Principal Components Analysis

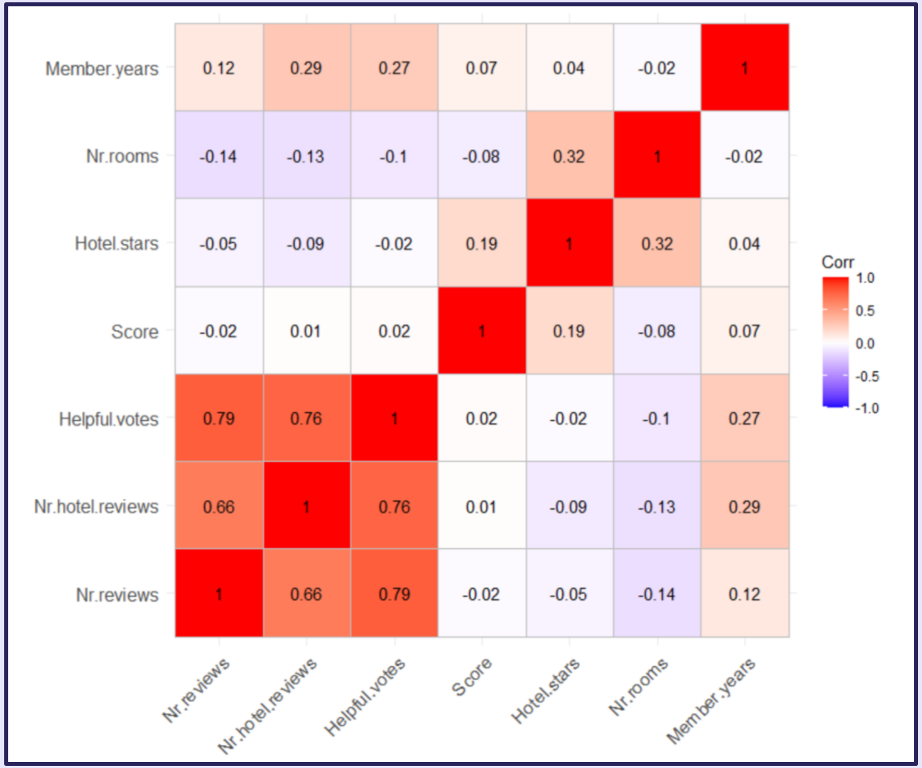
Select Numeric Data

```{r}
df1 = reviews %>% select_if(is.numeric) %>% scale()
head(df1)
```

```
     Nr.reviews Nr.hotel.reviews Helpful.votes      Score Hotel.stars Nr.rooms Member.years
[1,] -0.4955200       -0.5019974    -0.3867743  0.8721485   -1.474581 1.226678    1.5843617
[2,]  0.9433634        0.2069340     0.8898919 -1.1128772   -1.474581 1.226678   -0.4643116
[3,] -0.1624452       -0.2934882    -0.1396776  0.8721485   -1.474581 1.226678   -0.8057572
[4,] -0.4555511       -0.3768919    -0.3661829 -0.1203644   -1.474581 1.226678    0.5600250
[5,] -0.5754580       -0.4602956    -0.6132795 -0.1203644   -1.474581 1.226678    0.9014706
[6,] -0.2290601       -0.3351900    -0.0984948 -1.1128772   -1.474581 1.226678   -0.8057572
```

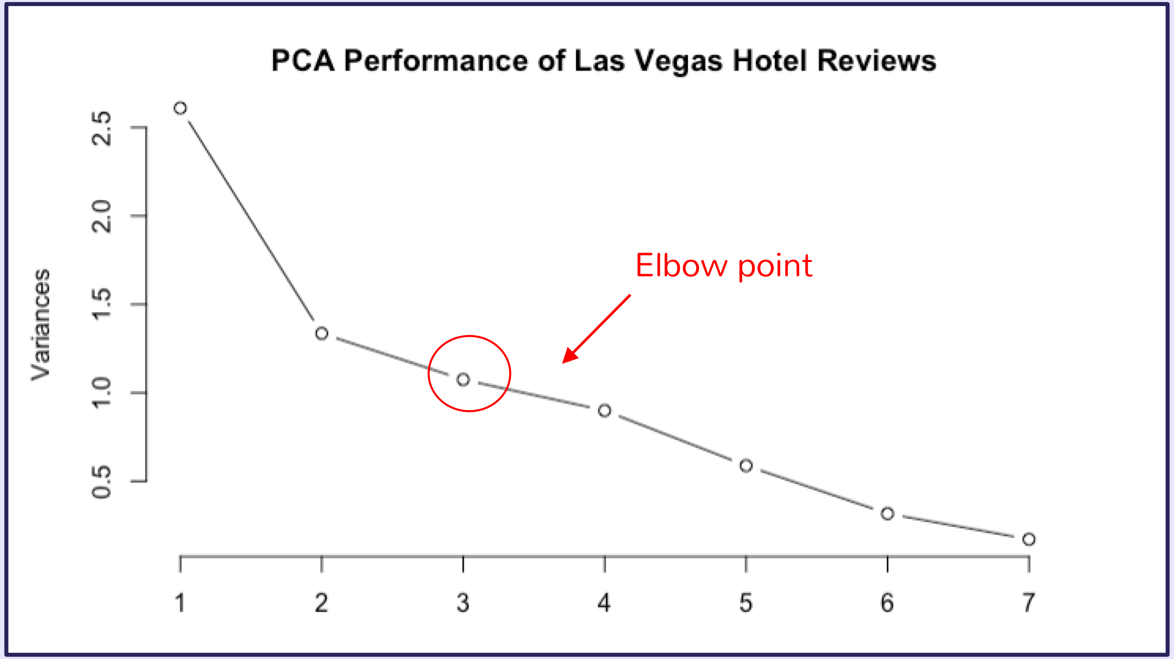
- First of all, as the category data is not recommended in pca, we just want numeric data from our dataset. there are seven in total.

Correlation Matrix



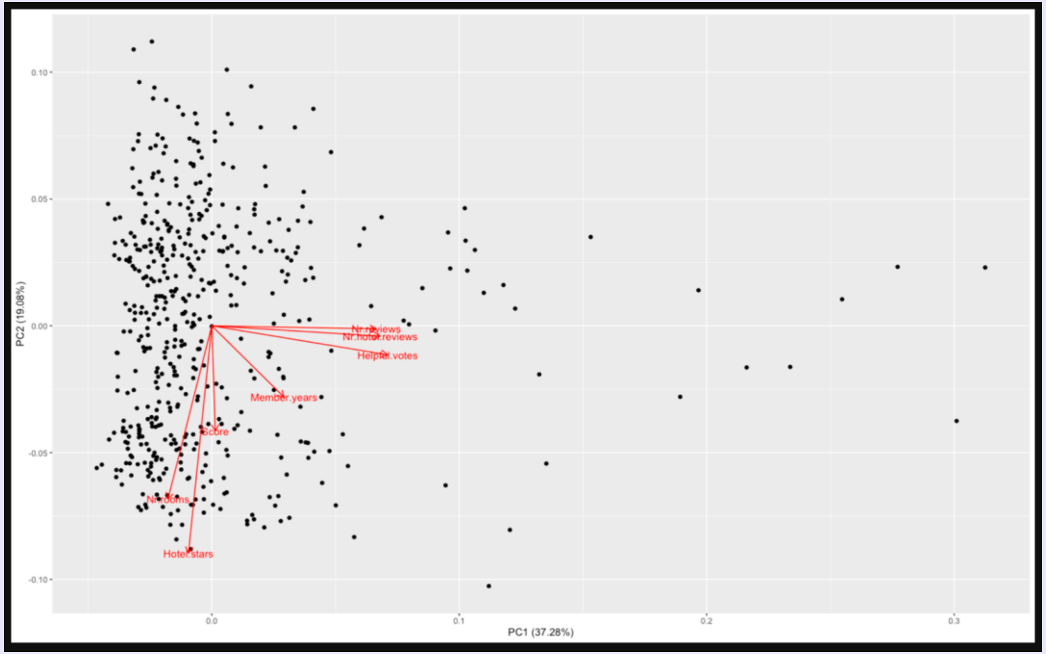
- as the correlation matrix shows, there are correlation between review attributes

PCA Performing



- From the scree plot we got, we can tell from the shape that k = 2 or k = 3 is good enough to be an elbow point.
- When the k value = 3, the eigenvalue is greater than 1. So we will choose k = 3 in our further study.

Biplot



- As the biplot shows, the arrows of Nr.reviews, Nr.hotel.reviews, and Helpful.votes are in the same direction, and the angles are small, it indicates that they are highly correlated. They strongly affect PC1.
- Similarly, Hotel.stars, Nr.rooms, and Score are highly correlated. They strongly affect PC2.
- The attributes that are perpendicular to each other are not likely to be correlated.

PCA Loadings

```
```{r}
pca_reviews <- prcomp(df1)
pca_reviews
```

Standard deviations (1, .., p=7):
[1] 1.6155216 1.1558113 1.0368824 0.9487695 0.7673777 0.5635027 0.4153283

Rotation (n x k) = (7 x 7):
                       PC1   PC2   PC3   PC4   PC5   PC6   PC7
Nr.reviews   0.53647727 -0.009655258 0.13247154 0.282464188 0.03763082 -0.59530895 -0.50887656
Nr.hotel.reviews 0.54918416 -0.032850959 0.03858999 -0.001684426 -0.10819903 0.76506733 -0.31430955
Helpful.votes   0.57343983 -0.092308408 0.08477503 0.105727068 -0.01936170 -0.09503210 0.79679039
Score   0.01153066 -0.334532984 -0.79301259 0.267062180 -0.43031648 -0.04882604 -0.01439787
Hotel.stars   -0.07584413 -0.723345078 0.01149901 0.194770316 0.64919337 0.10347190 -0.02816735
Nr.rooms   -0.14264575 -0.551083731 0.54122463 -0.073395255 -0.61198510 -0.04742046 -0.02955472
Member.years   0.23563224 -0.226876729 -0.22770597 -0.891284682 0.07315698 -0.18949672 -0.07419602
```

- PC1: Nr.reviews, Nr.hotel.reviews, and Helpful.votes are mostly weighted.
- PC2: Hotel.stars, Nr.rooms, and Score are mostly weighted.
- PC3: Score, Nr.rooms are mostly weighted.

- Factor1: All about reviews
- Factor2: Brand of Hotel
- Factor3: Crowds, size of group of people

Possible Factors

- **Factor1: all about reviews.**
  It indicates the importance of reviews and reviewers. I think reviews matter a lot, even for me as a customer. For the hotel's aspect, it might be really helpful if they pay more attention to the reply to the reviews already posted and listen to their advice. And they can also find some influencer to do more good reviews to improve their images

- **Factor2: Brand of Hotel.**
  People usually prefer to book hotels with familiar brands. They might be middle class, big families who want big and clean hotels with good service. For the hotel's aspect, they can think of how to make their hotel more fancy, and provide good and patient service.

- **Factor3: Crowds, size of group of people.**
  Hotel can make improvements in room type that can fit all different sizes of group.

Summary

```
```{r}
summary(pca_reviews)
```

Importance of components:
                          PC1   PC2   PC3   PC4   PC5   PC6   PC7
Standard deviation   1.6155 1.1558 1.0369 0.9488 0.76738 0.56350 0.41533
Proportion of Variance 0.3728 0.1908 0.1536 0.1286 0.08412 0.04536 0.02464
Cumulative Proportion 0.3728 0.5637 0.7173 0.8459 0.93000 0.97536 1.00000
```

- In total, the first three principal components explained (37.28%+19.08%+15.36%) = 71.72% of the total variance, which is the most variance of our dataset. It indicates that hotels should really value these three factors to attract more customers and increase revenues.

Summary and Recommendations

Provide amenities like free internet, hands down.

Logistic Regression, Decision Tree Model

Elevate hotel brand and presence on the strip to boost reputation.

PCA

Gain exposure to experienced reviewers via direct ads, influencers, etc.

PCA, Forest, Hypothesis Test

Identify a more refined target market and get more data on frequent reviewers.

Decision tree, Forest, PCA

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, infographics & images by Freepik and illustrations by Storyset