# What factors are most influential in predicting a Las Vegas Hotel's online rating?

By: Omer Canca, Ben Caggiano, Sarvjot Baxi, Ray Chen

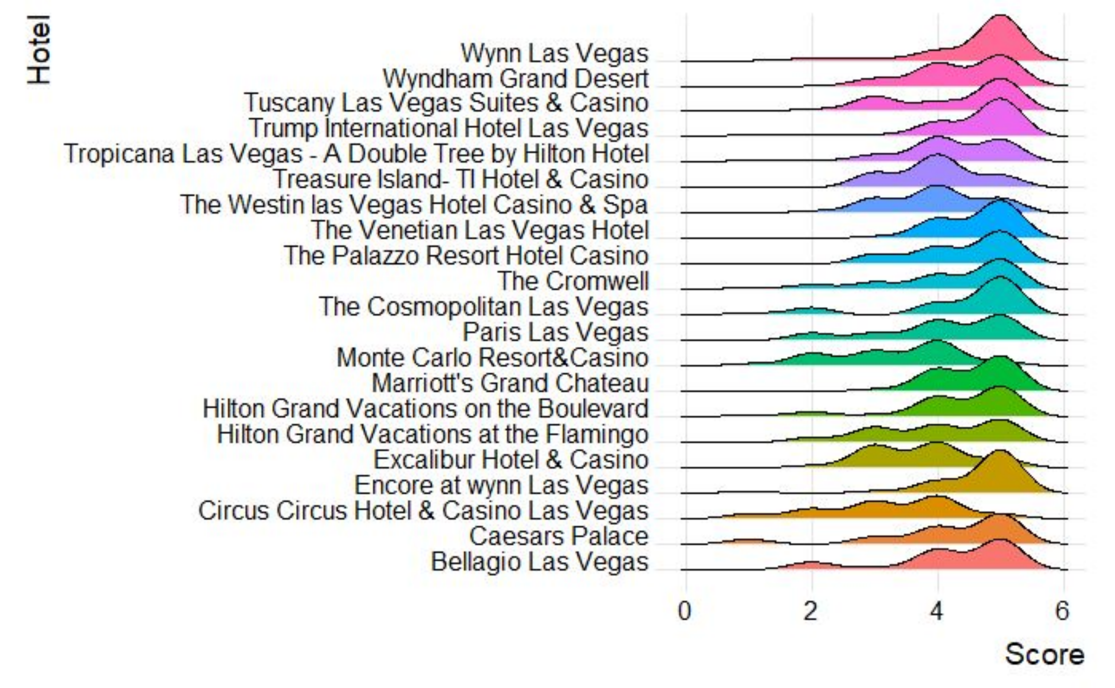
# About the Data Set

- Reviews taken from 21 hotels on the Las Vegas Strip
- Two reviews selected per month from 2015
    - 24 reviews per hotel, 504 total reviews
- 20 features


**Figure 1** - Review and user features extracted.

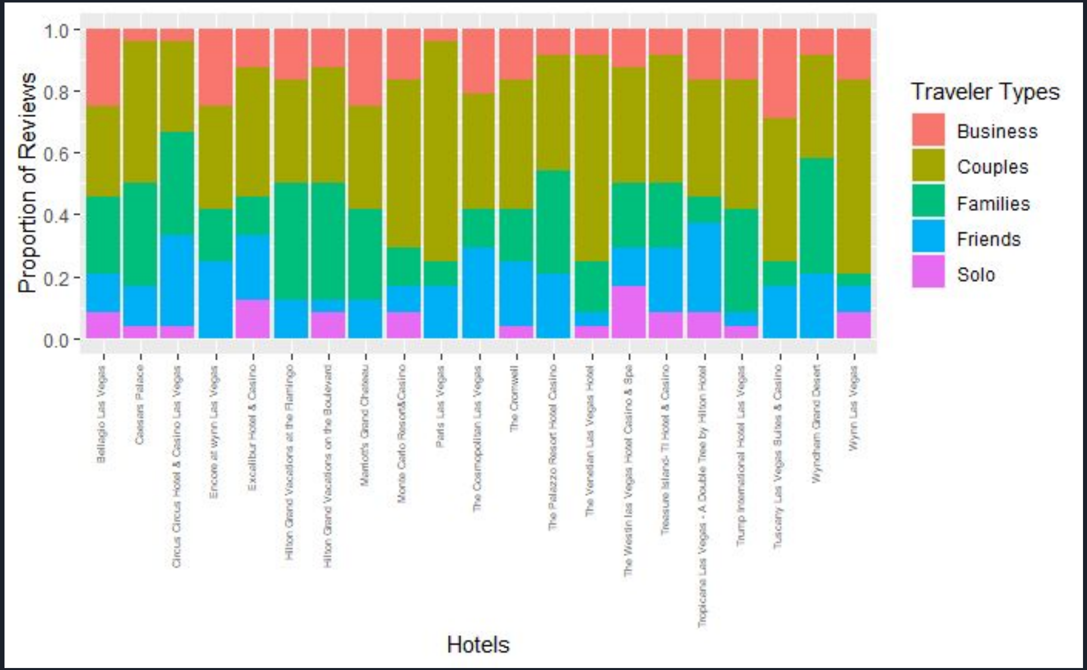
# Score Distribution of All Hotels

```
ggplot(hotel_review_df,
    aes(x = Score,
      y = Hotel.name,
      fill = Hotel.name)) +
  geom_density_ridges() +
  theme_ridges() +
  labs("Hotel Rating Distribution") +
  ylab("Hotel")+
  theme(legend.position = "none")
```



# Score Distribution of All Traveler Types

```
ggplot(hotel_review_df,
    aes(x = Hotel.name,
        fill = Traveler.type)) +
  geom_bar(position = "fill") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.1,hjust=1, size = 5))+
  scale_y_continuous(breaks = seq(0, 1, .2)) +
  labs(y = "Proportion of Reviews", x = "Hotels", fill = "Traveler Types")
```



# Simple Linear Regression

**Coding Variables**

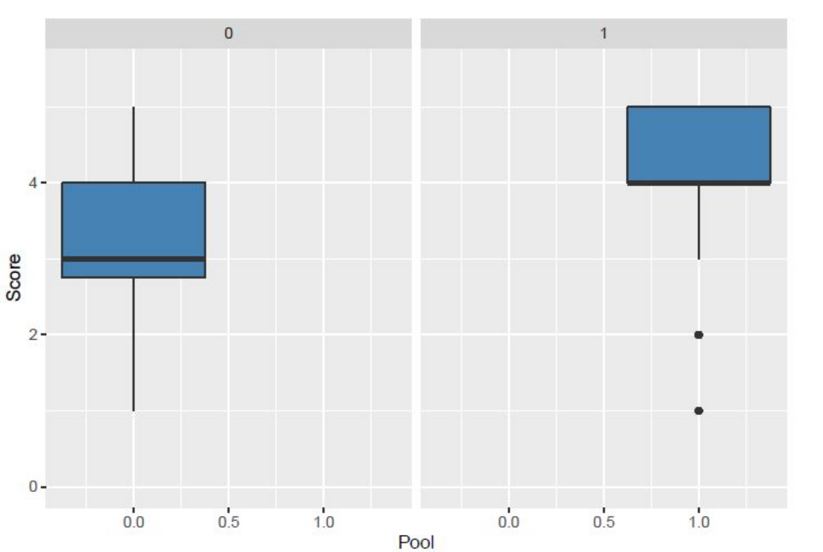
Spa = 1 for yes; 0 for no Gym = 1 for yes; 0 for no Pool = 1 for yes; 0 for no Casino = 1 for yes; 0 for no Free.internet = 1 for yes; 0 for no Tennis.court = 1 for yes; 0 for no

```
data$Spa<-ifelse(data$Spa=="YES",1,0)
data$Gym<-ifelse(data$Gym=="YES",1,0)
data$Pool<-ifelse(data$Pool=="YES",1,0)
data$Casino<-ifelse(data$Casino=="YES",1,0)
data$Free.internet<-ifelse(data$Free.internet=="YES",1,0)
data$Tennis.court<-ifelse(data$Tennis.court=="YES",1,0)
df <- dplyr::select_if(data, is.numeric)
```

```
## Subset selection object
## Call: regsubsets.formula(Score ~ ., data = df, nvmax = 1, method = "backward")
## 11 Variables  (and intercept)
##                  Forced in Forced out
## Nr..reviews          FALSE      FALSE
## Nr..hotel.reviews    FALSE      FALSE
## Helpful.votes        FALSE      FALSE
## Pool                 FALSE      FALSE
## Gym                  FALSE      FALSE
## Tennis.court         FALSE      FALSE
## Spa                  FALSE      FALSE
## Casino               FALSE      FALSE
## Free.internet        FALSE      FALSE
## Nr..rooms            FALSE      FALSE
## Member.years         FALSE      FALSE
## 1 subsets of each size up to 1
## Selection Algorithm: backward
##          Nr..reviews Nr..hotel.reviews Helpful.votes Pool Gym Tennis.court Spa
## 1  ( 1 ) " "         " "               " "           "*"  " " " "          " "
##          Casino Free.internet Nr..rooms Member.years
## 1  ( 1 ) " "    " "           " "       " "
```

Our results tell us that Pool is the most significant variable.

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.2083     0.2015  15.920  < 2e-16 ***
## Pool          0.9604     0.2065   4.651 4.23e-06 ***
```



# Multiple Linear Regression

```
## Casino                 FALSE      FALSE
## Free.internet          FALSE      FALSE
## Nr..rooms              FALSE      FALSE
## Member.years           FALSE      FALSE
## 1 subsets of each size up to 5
## Selection Algorithm: forward
##          Nr..reviews Nr..hotel.reviews Helpful.votes Pool Gym Tennis.court Spa
## 1  ( 1 ) " "         " "               " "           "*"  " " " "          " "
## 2  ( 1 ) " "         " "               " "           "*"  " " " "          " "
## 3  ( 1 ) " "         " "               " "           "*"  " " " "          " "
## 4  ( 1 ) " "         " "               " "           "*"  "*" " "          " "
## 5  ( 1 ) " "         " "               " "           "*"  "*" " "          "*"
##          Casino Free.internet Nr..rooms Member.years
## 1  ( 1 ) " "    " "           " "       " "
## 2  ( 1 ) " "    "*"           " "       " "
## 3  ( 1 ) " "    "*"           " "       "*"
## 4  ( 1 ) " "    "*"           " "       "*"
## 5  ( 1 ) " "    "*"           " "       "*"
```

```
model2 = lm(Score ~ Pool + Free.internet, data = df)
model3 = lm(Score ~ Pool + Free.internet + Member.years, data = df)
model4 = lm(Score ~ Pool + Free.internet + Member.years + Gym, data = df)
model5 = lm(Score ~ Pool + Free.internet + Member.years+ Gym + Spa, data = df)
```

**AIC**

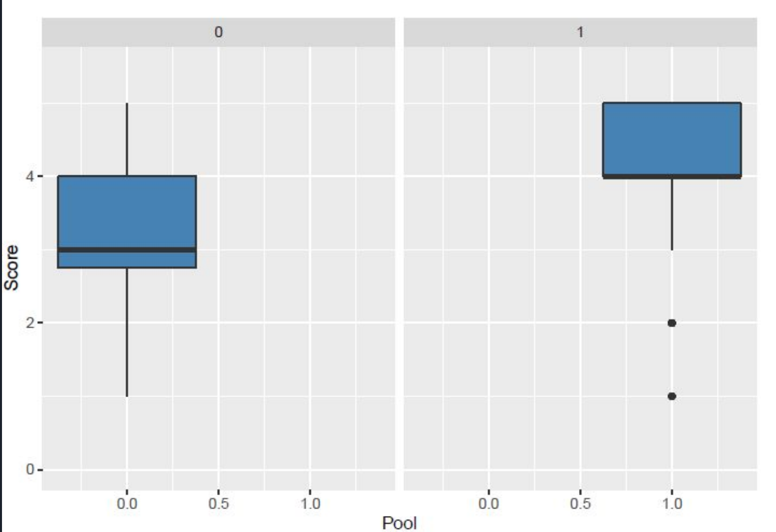
We will use AIC to determine which model is the best of the three. AIC is a score that is used to determine which model is best based on prediction error. A lower AIC is better

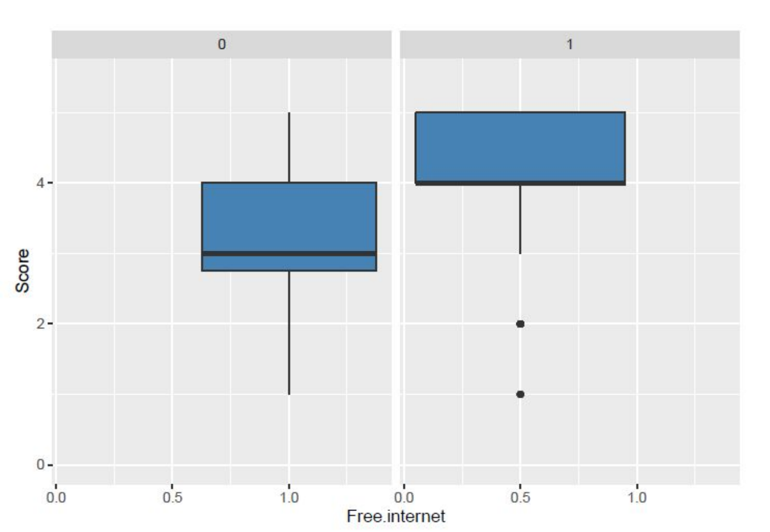
```
AIC(model2)
## [1] 1402.934
AIC(model3)
## [1] 1404.342
AIC(model4)
## [1] 1405.88
AIC(model5)
## [1] 1407.411
```

```
## Analysis of Variance Table
##
## Model 1: Score ~ Pool + Free.internet
## Model 2: Score ~ Pool
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    501 469.86
## 2    502 489.29 -1   -19.434 20.723 6.671e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Our results tell us that the extra predictor in model2 is significant.

# Multiple Linear Regression 2



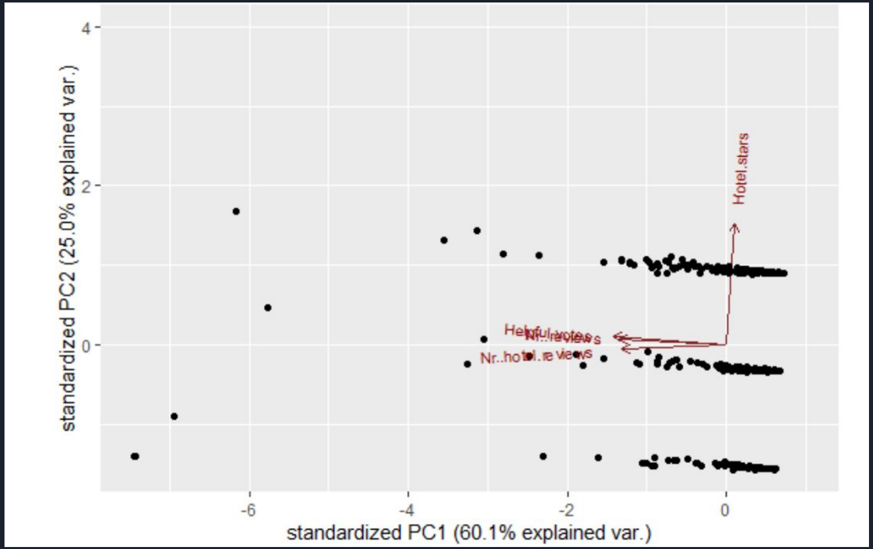

## Conclusion

Our final model includes free internet and pool as the best predictors for score. This tells us that when looking for a hotel in Vegas, we should look for these two predictors to find the hotels with the best experience.

# Dimensionality reduction through PCA

We did some dimensionality reduction and clustering too

Most of the predictors contribute to PC2 and only Hotel.Stars contributes to PC1




Here we can see the actual PCA we get, there are 4

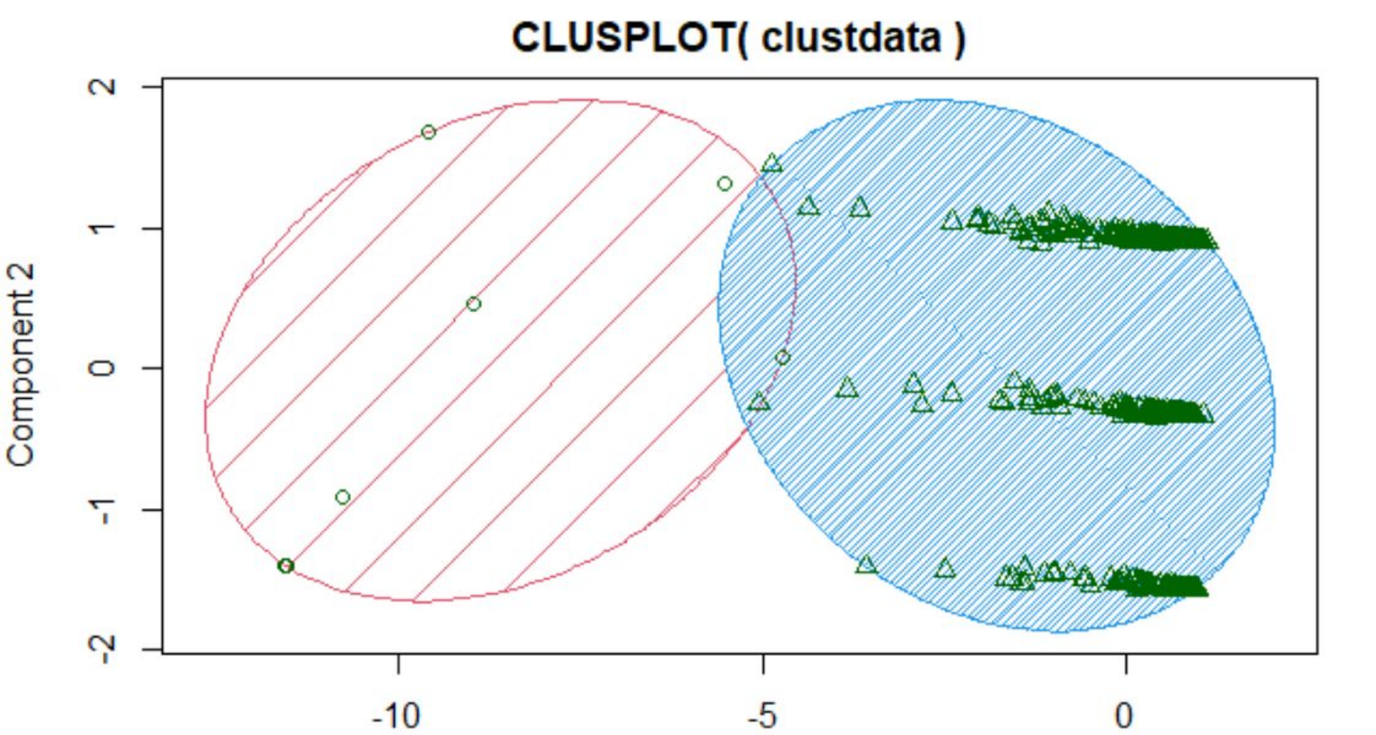
```
[1] 0
Importance of components:
                          PC1    PC2    PC3     PC4
Standard deviation     1.5501 1.0007 0.6369 0.43597
Proportion of Variance 0.6007 0.2504 0.1014 0.04752
Cumulative Proportion  0.6007 0.8511 0.9525 1.00000
```

1-variablity_explained,  within 3 PCA we get 90% of variability in the dataset

```
[1] 0.3992821 0.7496315 0.8986044 0.9524820
```

# Some clustering

We also tried k-means clustering on our data set to see if there were nay more trends or patterns within it




We found that 2 clusters can be clearly identified both of which when combined can explain 85.11% point variability in the data

# Feature Selection (Remove Irrelevant Variables)

- Model <- lm(data = LasVegas, Score ~  Pool + Gym + Tennis_court + Spa + Casino + Free_internet)
- We can use Adjusted R squared, AIC and BIC to see which model is the best fit.
- Backward Stepwise Selection: Begins with the full least squares model containing all p predictors, and then iteratively removes the least useful predictor, one-at-a-time.
- Forward  Stepwise Selection: Starts with a model with no predictors and then we add predictors to the model one-at-time until getting the complete model (all the predictors). At each step we add the variable that gives the greatest additional improvement to the fit: usually R2 or RSS.

# Backward Stepwise Selection

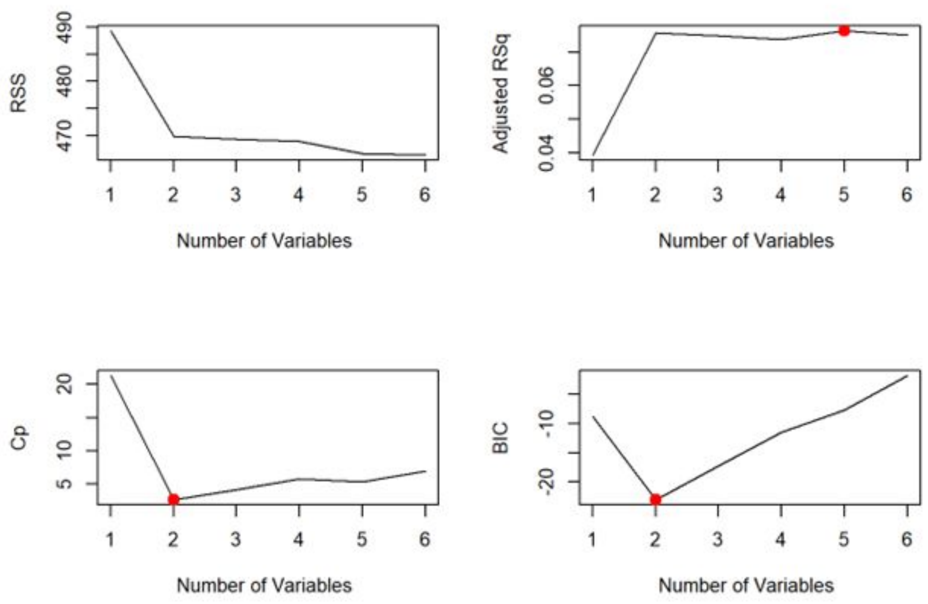
```
regfit.bwd = regsubsets(Score ~  Pool + Gym + Tennis_court + Spa + Casino + Free_internet, data = LasVegas,  nvma
x = 6, method="backward")
reg.summary <- summary(regfit.bwd) #get the summary

par(mfrow=c(2,2))
#rss plot -  NOT USEFUL
plot(reg.summary$rss ,xlab="Number of Variables ",ylab="RSS",type="l")

#adjr2 plot
plot(reg.summary$adjr2 ,xlab="Number of Variables ", ylab="Adjusted RSq",type="l")
max_adjr2 <- which.max(reg.summary$adjr2)
points(max_adjr2,reg.summary$adjr2[max_adjr2], col="red",cex=2,pch=20)

# AIC criterion (Cp) to minimize
plot(reg.summary$cp ,xlab="Number of Variables ",ylab="Cp", type='l')
min_cp <- which.min(reg.summary$cp )
points(min_cp, reg.summary$cp[min_cp],col="red",cex=2,pch=20)

# BIC criterion to minimize
plot(reg.summary$bic ,xlab="Number of Variables ",ylab="BIC",type='l')
min_bic <- which.min(reg.summary$bic)
points(min_bic,reg.summary$bic[min_bic],col="red",cex=2,pch=20)
```




Adjusted R-Square highest at 5 variables. ( 2 ~ 6 variables are about the same)
C(p) lowest at 2 variables.
BIC  lowest at 2 variables.

# Forward Stepwise Selection

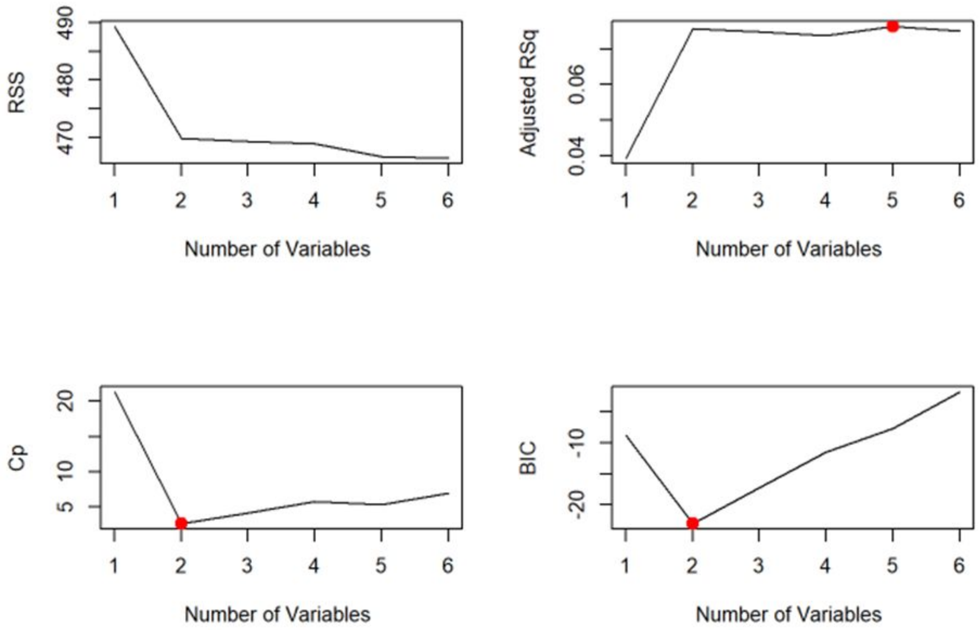
```r
regfit.fwd = regsubsets(Score ~  Pool + Gym + Tennis_court + Spa + Casino + Free_internet, data = LasVegas,  nvma
x = 6, method="forward")
reg.summary <- summary(regfit.fwd) #get the summary

par(mfrow=c(2,2))
#rss plot -  NOT USEFUL
plot(reg.summary$rss ,xlab="Number of Variables ",ylab="RSS",type="l")

#adjr2 plot
plot(reg.summary$adjr2 ,xlab="Number of Variables ", ylab="Adjusted RSq",type="l")
max_adjr2 <- which.max(reg.summary$adjr2)
points(max_adjr2,reg.summary$adjr2[max_adjr2], col="red",cex=2,pch=20)

# AIC criterion (Cp) to minimize
plot(reg.summary$cp ,xlab="Number of Variables ",ylab="Cp", type='l')
min_cp <- which.min(reg.summary$cp )
points(min_cp, reg.summary$cp[min_cp],col="red",cex=2,pch=20)

# BIC criterion to minimize
plot(reg.summary$bic ,xlab="Number of Variables ",ylab="BIC",type='l')
min_bic <- which.min(reg.summary$bic)
points(min_bic,reg.summary$bic[min_bic],col="red",cex=2,pch=20)
```




Adjusted R-Square highest at 5 variables.
( 2 ~ 6 variables are about the same)
C(p) lowest at 2 variables.
BIC  lowest at 2 variables.

# Summary

Best Subsets Regression

| Model Index | Predictors |
| --- | --- |
| 1 | Pool |
| 2 | Pool Free_internet |
| 3 | Pool Gym Free_internet |
| 4 | Pool Gym Spa Free_internet |
| 5 | Pool Gym Spa Casino Free_internet |
| 6 | Pool Gym Tennis_court Spa Casino Free_internet |

Model 2 will be the best model based on Adj.R-Square, C(p), AIC, and SBIC. The predictor variables that are relevant to the hotel score are Pool and Free internet.

Subsets Regression Summary

| Model | R-Square | Adj. R-Square | Pred R-Square | C(p) | AIC | SBIC | SBC | MSEP | FPE | HSP | APC |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 0.0413 | 0.0394 | 0.0327 | 21.3393 | 1421.3607 | -9.0646 | 1434.0284 | 491.2390 | 0.9785 | 0.0019 | 0.9663 |
| 2 | 0.0794 | 0.0757 | 0.0666 | 2.6319 | 1402.9336 | -27.3161 | 1419.8240 | 472.6707 | 0.9434 | 0.0019 | 0.9316 |
| 3 | 0.0802 | 0.0747 | 0.062 | 4.1646 | 1404.4630 | -25.7659 | 1425.5758 | 473.1758 | 0.9463 | 0.0019 | 0.9345 |
| 4 | 0.0811 | 0.0737 | 0.06 | 5.7207 | 1406.0154 | -24.1892 | 1431.3509 | 473.7051 | 0.9492 | 0.0019 | 0.9374 |
| 5 | 0.0855 | 0.0763 | 0.0626 | 5.3233 | 1405.5917 | -24.5376 | 1435.1497 | 472.3811 | 0.9484 | 0.0019 | 0.9366 |
| 6 | 0.0861 | 0.0750 | 0.0598 | 7.0000 | 1407.2640 | -22.8293 | 1441.0446 | 473.0258 | 0.9516 | 0.0019 | 0.9397 |

$$\widehat{\text{Score}} = 2.29 + 1.01(\text{Pool}) + 0.92(\text{Free\_internet})$$

Yes = 1
No = 0

For example, a hotel with free internet but no pool, the score would be 2.29+0.92 = 3.21