A1_DSC424

Alex Teboul

September 16, 2019

Problem 1

Problem 1(5 points – Due Wednesday, September 11th, 2019 at 5PM) Introduce yourself on D2L by posting to the Class Introductions forum on D2L. Include a bit of information about yourself including some of the following. Note, this:

- **Name**: Alex Teboul
- **Undergraduate Degree**: Biomedical Engineering
- **Major/Degree Program(Concentration)/Time in Program (e.g. 3rd quarter, 2nd yr, graduating this quarter)**: MS Data Science, (Computational Methods), 4th Quarter
- **Position at Work, if applicable**: Data Scientist at Cicero School District 99
- **What is your experience with R? Have you used it for any courses? For work?**: Minimal R work, used it in DSC 423 and DSC 465.
- **What interests you about Advanced Data Analysis?**: Excited to learn more sophisticated methods of finding strategic information in large datasets and hopefully work on an interesting, portfolio-building final project.
- **Field(s) of Interest and/data**: Computer Vision applications and healthtech/care (also edu to an extent because it's the field I'm working in currently).
- **Hobbies**: Reading, cryptocurrency tracking and investing, cooking, and working out.

Problem 2

Problem 2(10 points) Perform in R, the following calculations from linear algebra. For the following matrices and vectors. Submit both R code and the solution for credit.

$$Z = \begin{bmatrix} 1 & 9 \\ 1 & 5 \\ 1 & -3 \\ 1 & 11 \end{bmatrix}, Y = \begin{bmatrix} -1 \\ 6 \\ 0 \\ 8 \end{bmatrix}, M = \begin{bmatrix} 1 & 11 & 0 \\ 42 & 52 & 35 \\ 0 & 9 & 3 \end{bmatrix}, N = \begin{bmatrix} -10 & -10 & 0 \\ 0 & 10 & 20 \\ 10 & 20 & 10 \end{bmatrix}, v = \begin{bmatrix} -11 \\ 11 \\ 22 \end{bmatrix}, w = \begin{bmatrix} 8 \\ -2 \\ 4 \end{bmatrix},$$

a. $v \cdot w$ (dot product)
b. $-3 * w$
c. $M * v$
d. $M + N$
e. $M - N$
f. $Z^T Z$
g. $(Z^T Z)^{-1}$
h. $Z^T Y$
i. $\beta = (Z^T Z)^{-1} Z^T Y$
j. $\det(Z^T Z)$

```
Z = matrix(c(1, 9, 1, 5, 1, -3, 1, 11), nrow=4, ncol=2, byrow=T)
Y = matrix(c(-1, 6, 0, 8), nrow=4, ncol=1, byrow=T)
M = matrix(c(1, 11, 0, 42, 52, 35, 0, 9, 3), nrow=3, ncol=3, byrow=T)
N = matrix(c(-10, -10, 0, 0, 10, 20, 10, 20, 10), nrow=3, ncol=3, byrow=T)
v = matrix(c(-11, 11, 22), nrow=3, ncol=1, byrow=T)
w = matrix(c(8, -2, 4), nrow=3, ncol=1, byrow=T)
```

a)

```
# dot product = multiplying two vectors. Answer is scalar
a = t(v)%*%w
a
```

```
##      [,1]
## [1,]  -22
```

b)

```
# multiply vector by scalar
b = -3*w
b
```

```
##      [,1]
## [1,]  -24
## [2,]    6
## [3,]  -12
```

c)

```
# multiply matrix by vector
c = M%*%v
c
```

```
##      [,1]
## [1,]  110
## [2,]  880
## [3,]  165
```

d)

```
# Matrix Addition
d = M + N
d
```

```
##      [,1] [,2] [,3]
## [1,]   -9    1    0
## [2,]   42   62   55
## [3,]   10   29   13
```

e)

```
# Matrix Subtraction
e = M - N
e
```

```
##      [,1] [,2] [,3]
## [1,]   11   21    0
## [2,]   42   42   15
## [3,]  -10  -11   -7
```

f)

```
# Gramian Matrix
f = t(Z)%*%Z
f
```

```
##      [,1] [,2]
## [1,]    4   22
## [2,]   22  236
```

g)

```
# inverse matrix calculation
g = solve(t(Z)%*%Z)
g
```

```
##             [,1]         [,2]
## [1,]  0.51304348 -0.047826087
## [2,] -0.04782609  0.008695652
```

h)

```
# Transpose and multiply matrices
h = t(Z)%*%Y
h
```

```
##      [,1]
## [1,]   13
## [2,]  109
```

i)

```
# inverse f times h
beta = solve(f)%*%h
beta
```

```
##          [,1]
## [1,] 1.456522
## [2,] 0.326087
```

j)

```
# det(f)
j = det(f)
j
```

```
## [1] 460
```

Problem 3

Problem 3 (10 points – other types of regression models): There are other types of regression models outside of linear and logistic regression. Using Google Scholar, locate a journal article, which utilizes one of the types of regressions listed below or another regression outside of linear/logistic that interests you. Write a summary of the journal article and how it utilizes the regression model in two to three paragraphs. Cite the paper in APA format. Choose one of the following regressions:

1. Ridge Regression

2. Lasso Regression

3. Elastic Net Regression

4. Poisson Regression

5. Negative Binomial Regression

6. Cox Regression

7. Robust Regression

8. Jackknife Regression

9. Time Series Regression

10. Polynomial Regression

11. Bayesian Linear Regression

Chosen Regression Model: Bayesian Linear Regression

Article Summary:

In this article, S. Baldwin and M. Larson describe the mechanics of Bayesian Linear Regression and provide an example of its use on EEG/anxiety study data. The article is split into 5 segments, which discuss how Bayesian Regression works, a sample Bayesian model on the EEG/anxiety data, results analysis, additional models, and best practices for researchers looking to apply these methods (Baldwin & Larson, 2017). First, Bayesian Linear Regression offers an alternative to traditional frequentist methods like standard linear regression and ANOVA, which are at the heart of the authors' field of psychology. As the authors describe it, the Bayesian method requires a 'prior' or prediction of parameter probabilities before a model is fit to data. The prior or priors are the researchers' subjective beliefs about parameter distributions, which help inform the model. Once data is acquired, the priors can be updated using Bayes Theorem to give what is called the 'posterior distribution' (Baldwin & Larson, 2017, p. 60). An advantage of this approach is that the result is a distrubion or density rather than single number or point estimate which is obtained with frequentist methods. When, as is often the case, posterior distributions turn out to be non-normal, Markov Chain Monte Carlo methods are used to obtain random pulls from the distribution - basically simulation is used over thousands of iterations to arrive at results.

The sample Bayesian Regression analysis the authors perform uses electroencephalogram (EEG) data, specifically event-related potential (ERPs) called error-related negativity (ERN). That is quite a mouthful, but the crux of it is that signals are recorded on the scalp that correlate with brain activity in response to certain thoughts or actions. The model was constructed for use with a scores from a state-trait anxiety survey. The general linear model was given by ERN = b0 + b1Anxiety + e. They go on to extend the model to men and women, and in both cases their method provides a useful analysis for relating ERNs with anxiety (Baldwin & Larson, 2017). They conclude with an optimistic take on using Bayesian methods to augment clinical research in the field of psychology.

Citation:

Baldwin, S. A., & Larson, M. J. (2017). An introduction to using Bayesian linear regression with clinical data. *Behavior Research and Therapy*, *98*, 58–75. https://doi.org/10.1016/j.brat.2016.12.016 (https://doi.org/10.1016/j.brat.2016.12.016)

Problem 4

Problem 4 (10 points-Data Ethics or Data Integrity): Using Google Scholar, locate a journal article, which discusses data ethics or data integrity in terms of big data in your field of interest. Write a summary of the journal article and how it utilizes data ethics or data integrity in two to three paragraphs. Cite the paper in APA format.

Article Summary:

Chen, Biglari-Abhari, and Wang (2017) explore the usefulness of computer vision systems in protecting privacy using what they refer to as a 'privacy-affirming' approach. The authors propose that, "in an ideal surveillance scenario, no person should ever visually see any image footage, only processed output," (Chen et al., 2017, p. 57). In this way, private information, which could potentially be used for nefarious purposes or leaked out would be shielded from human eyes. The authors essentially argue that by adhering to certain design principles, surveillance systems with embedded computer vision applications, can actually improve privacy over current systems. This notion seems to run counter to many of the fears associated with such systems and computer vision in general.

In terms of data integrity, which has to do with the accuracy, completeness, and reliability of data, the authors stress that in order for such systems to work, trust in the computer vision component needs to increase. By eliminating portions of human input in surveillance tasks, and replacing it with automated processes, one has to trust the accuracy of the computer vision system fully, say Chen et al (2017). To better understand how one such system might work, they provide an example of a smart camera surveillance setup to monitor a room. If the system was to be privacy-aware then people entering the room could can blurred out faces, such that they are anonymous to the human watching the footage. But to be privacy-affirming, the authors suggest taking it a step further to simply alert a human in the loop with text information like "person 3 in Zone B" (Chen et al., 2018, p. 58). Layers of computer processing protect privacy, because a human viewer could gain additional unintended information from the video, information that is private and not the objective of the system in the first place. The authors do not use this example, but one could image a camera set up to count the number of passers by on a street corner wearing a certain brand of tshirt. If all the video is recorded, and looked at by a human, the human might see a passer by walking into an AA meeting or discussing something private. This type of privacy breach brings data ethics into play.

The authors discuss the data ethics surrounding surveillance systems concerning how data is collected, transformed, protected, and ultimately used. They take a cautious approach, pushing for design that obfuscates personal information as often as possible, and prioritizes automated-processes to add layers of privacy protection. The surveillance networks the Chen et al are describing are also, "always watching but do not record the footage until trigger conditions occur, such as a target individual being identified or a violent action being detected," (2017, p. 59). This would fit in with many other systems that, supposedly, only activate when triggers are observed - like calling out 'Alexa' to you Amazon Alexa-enabled devices. While more information and video recordings could prove useful in certain situations, the privacy protections offered by more carefully considered privacy-affirming systems could help prevent privacy breaches in the future.

Citation:

Chen, A. T., Biglari-Abhari, M., & Wang, K.I. (2017).Trusting the Computer in Computer Vision: A Privacy-Affirming Framework.

Published in 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) in Honolulu, Hawaii.

https://doi.org/10.1109/CVPRW.2017.178 (https://doi.org/10.1109/CVPRW.2017.178)

Problem 5

Problem 5: (15 pts – regression analysis, visualization, and interpretation): This dataset includes quantitative and categorical features from online reviews from 21 hotels located in Las Vegas Strip, extracted from TripAdvisor. All the 504 reviews were collected between January and August of 2015 and

there are 19 extracted features in the revised dataset.

Feature name	Origin	Source type	Data type	Description	Status
Username	Extracted (1)	User	Categorical	Username as registered in TripAdvisor	Excluded
User country	Extracted (2)	User	Categorical	User's nationality	Included
Nr. Reviews	Extracted (3)	User	Numerical	Number of reviews	Included
Nr. Hotel reviews	Extracted (4)	User	Numerical	Total hotel reviews	Included
Helpful votes	Extracted (5)	User	Numerical	Helpful votes regarding reviews's info	Included
Score	Extracted (6)	Review	Numerical	Review score {1,2,3,4,5}	Included
Review date	Extracted (7)	Review	Date	Date when the review was written	Transformed
Review text	Extracted (8)	Review	Text	Textual content of the review	Excluded
Review language	Extracted (9)	Review	Categorical	Language of the review	Excluded
Period of stay	Extracted (10)	Review	Categorical	Period of stay: {Dec-Feb, Mar-May, Jun-Aug, Sep-Nov}	Included
Traveler type	Extracted (11)	Review	Categorical	{Business, Couples, Families, Friends, Solo}	Included
Member registered year	Extracted (12)	User	Date (year)	Year the user has registered in TripAdvisor	Transformed
Pool	Extracted (13)	Hotel	Categorical	If the hotel has outside pool	Included
Gym	Extracted (14)	Hotel	Categorical	If the hotel has gym	Included
Tennis court	Extracted (15)	Hotel	Categorical	If the hotel has tennis court	Included
Spa	Extracted (16)	Hotel	Categorical	If the hotel has spa	Included
Casino	Extracted (17)	Hotel	Categorical	If the hotel has a casino inside	Included
Free internet	Extracted (18)	Hotel	Categorical	If the hotel provides free internet	Included
Hotel name	Extracted (19)	Hotel	Categorical	Hotel's name	Included
Hotel stars	Extracted (20)	Hotel	Categorical	Hotel's number of stars	Included
Nr. Rooms	Extracted (21)	Hotel	Numerical	Hotel's number of rooms	Included
User continent	Computed	User	Categorical	Continent where the user's country is located	Included
Member years	Computed	User	Numerical	Number of years the user is member of TripAdvisor	Included
Review month	Computed	Review	Categorical	Month when the review was written (from review date)	Included
Review weekday	Computed	Review	Categorical	Day of the week the review was written (from review date)	Included

We are interested in which independent variables are significant for predicting the score of the hotel reviews by the other predictors excluding date and hotel name variables.

a. (5 points) Before running any regressions make sure to check for multicollinearity. How did you check for multicollinearity? If there is multicollinearity, how do you plan to resolve it? Are there any other issues with the dataset we have to consider before running the regressions?

First, we have to get all the data, check the structure, fix names, check for missing values, and drop unneeded columns. Then, in order to check for multicollinearity, I followed the method explained in class with VIF. As was desribed in class, VIF>10 indicates multicollinearity, and informs which variables should remain in the model. Because many variables are categorical, I couldn't show correlations either

a)

```
#Libraries
library(Hmisc) #Describe Function
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
library(psych) #Multiple Functions for Statistics and Multivariate Analysis
```

```
##
## Attaching package: 'psych'
```

```
## The following object is masked from 'package:Hmisc':
##
##     describe
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
library(GGally) #ggpairs Function
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(ggplot2) #ggplot2 Functions
library(vioplot) #Violin Plot Function
```

```
## Loading required package: sm
```

```
## Package 'sm', version 2.2-5.6: type help(sm) for summary information
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(corrplot) #Plot Correlations
```

```
## corrplot 0.84 loaded
```

```
library(DescTools) #VIF Function
```

```
##
## Attaching package: 'DescTools'
```

```
## The following objects are masked from 'package:psych':
##
##     AUC, ICC, SD
```

```
## The following objects are masked from 'package:Hmisc':
##
##     %nin%, Label, Mean, Quantile
```

```
library(leaps) #Best Set Linear Regression Functions
library(Amelia) #missmap
```

```
## Loading required package: Rcpp
```

```
## ##
## ## Amelia II: Multiple Imputation
## ## (Version 1.7.5, built: 2018-05-07)
## ## Copyright (C) 2005-2019 James Honaker, Gary King and Matthew Blackwell
## ## Refer to http://gking.harvard.edu/amelia/ for more information
## ##
```

```
library(car) #to get vif
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:DescTools':
##
##     Recode
```

```
## The following object is masked from 'package:psych':
##
##     logit
```

```
#Set Working Directory
setwd("C:/Users/ateboul/Desktop")

#Read in Datasets
data <- read.csv(file="Las_Vegas_Hotel_Reviews2.csv", header=TRUE, sep=",")
```

```
#Data Dimensions
dim(data)
```

```
## [1] 504  19
```

```
#First rows
head(data)
```

```
##   ID Nr..reviews Nr..hotel.reviews Helpful.votes Score Period.of.stay
## 1  1          11                 4            13     5        Dec-Feb
## 2  2         119                21            75     3        Dec-Feb
## 3  3          36                 9            25     5        Mar-May
## 4  4          14                 7            14     4        Mar-May
## 5  5           5                 5             2     4        Mar-May
## 6  6          31                 8            27     3        Mar-May
##   Traveler.type Pool Gym Tennis.court Spa Casino Free.internet
## 1       Friends   NO YES           NO  NO    YES           YES
## 2      Business   NO YES           NO  NO    YES           YES
## 3      Families   NO YES           NO  NO    YES           YES
## 4       Friends   NO YES           NO  NO    YES           YES
## 5          Solo   NO YES           NO  NO    YES           YES
## 6       Couples   NO YES           NO  NO    YES           YES
##                             Hotel.name Hotel.stars Nr..rooms
## 1 Circus Circus Hotel & Casino Las Vegas           3      3773
## 2 Circus Circus Hotel & Casino Las Vegas           3      3773
## 3 Circus Circus Hotel & Casino Las Vegas           3      3773
## 4 Circus Circus Hotel & Casino Las Vegas           3      3773
## 5 Circus Circus Hotel & Casino Las Vegas           3      3773
## 6 Circus Circus Hotel & Casino Las Vegas           3      3773
##   Member.years Review.month Review.weekday
## 1            9      January       Thursday
## 2            3      January         Friday
## 3            2     February       Saturday
## 4            6     February         Friday
## 5            7        March        Tuesday
## 6            2        March        Tuesday
```

So, we can drop Hotel.name, Review.month, and Review.weekday because they are the Date or Hotel Name variables. We can also remove ID because it's just the index.

```
df = subset(data, select=-c(Hotel.name, Review.month, Review.weekday, ID))
head(df)
```

```
##   Nr..reviews Nr..hotel.reviews Helpful.votes Score Period.of.stay
## 1          11                 4            13     5        Dec-Feb
## 2         119                21            75     3        Dec-Feb
## 3          36                 9            25     5        Mar-May
## 4          14                 7            14     4        Mar-May
## 5           5                 5             2     4        Mar-May
## 6          31                 8            27     3        Mar-May
##   Traveler.type Pool Gym Tennis.court Spa Casino Free.internet Hotel.stars
## 1       Friends   NO YES           NO  NO    YES           YES           3
## 2      Business   NO YES           NO  NO    YES           YES           3
## 3      Families   NO YES           NO  NO    YES           YES           3
## 4       Friends   NO YES           NO  NO    YES           YES           3
## 5          Solo   NO YES           NO  NO    YES           YES           3
## 6       Couples   NO YES           NO  NO    YES           YES           3
##   Nr..rooms Member.years
## 1      3773            9
## 2      3773            3
## 3      3773            2
## 4      3773            6
## 5      3773            7
## 6      3773            2
```

```
#structure
str(df)
```

```
## 'data.frame':    504 obs. of  15 variables:
##  $ 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 "Dec-Feb","Jun-Aug",..: 1 1 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             : Factor w/ 2 levels "NO","YES": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Gym              : Factor w/ 2 levels "NO","YES": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Tennis.court     : Factor w/ 2 levels "NO","YES": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Spa              : Factor w/ 2 levels "NO","YES": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Casino           : Factor w/ 2 levels "NO","YES": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Free.internet    : Factor w/ 2 levels "NO","YES": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Hotel.stars      : Factor w/ 5 levels "3","3,5","4",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Nr..rooms        : int  3773 3773 3773 3773 3773 3773 3773 3773 3773 3773 ...
##  $ Member.years     : int  9 3 2 6 7 2 4 0 3 5 ...
```

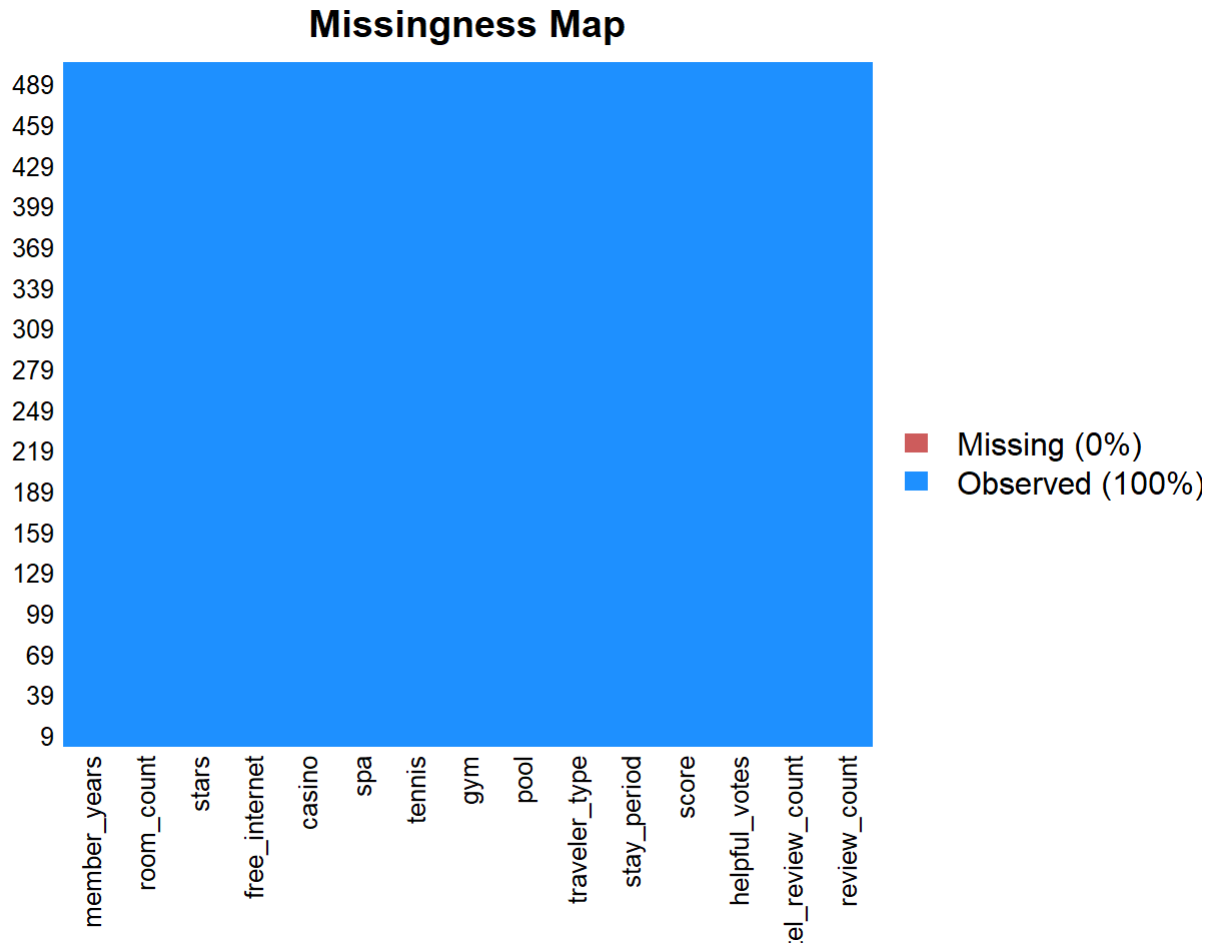
This is good. R already recognized a lot of the variables as Factor so it can run categorical regression later on.

```
#let's change the hotel stars. TripAdvisor has half stars. I checked 'The Cromwell' and it had
 4.5 stars on tripadvisor where here it says 4,5.
#df$Hotel.stars = as.numeric(df$Hotel.stars)
```

```
#fix the names
names(df) = c("review_count","hotel_review_count","helpful_votes", "score", "stay_period", "trav
eler_type", "pool","gym","tennis","spa","casino","free_internet","stars","room_count","member_ye
ars")
head(df)
```

```
##   review_count hotel_review_count helpful_votes score stay_period
## 1           11                  4            13     5     Dec-Feb
## 2          119                 21            75     3     Dec-Feb
## 3           36                  9            25     5     Mar-May
## 4           14                  7            14     4     Mar-May
## 5            5                  5             2     4     Mar-May
## 6           31                  8            27     3     Mar-May
##   traveler_type pool gym tennis spa casino free_internet stars room_count
## 1       Friends   NO YES     NO  NO    YES           YES     3       3773
## 2      Business   NO YES     NO  NO    YES           YES     3       3773
## 3      Families   NO YES     NO  NO    YES           YES     3       3773
## 4       Friends   NO YES     NO  NO    YES           YES     3       3773
## 5          Solo   NO YES     NO  NO    YES           YES     3       3773
## 6       Couples   NO YES     NO  NO    YES           YES     3       3773
##   member_years
## 1            9
## 2            3
## 3            2
## 4            6
## 5            7
## 6            2
```

```
#missing data
missmap(df)
```




```
#VIF Calculation

M1 <- lm(score ~.,data=df)
summary(M1)
```

```
##
## Call:
## lm(formula = score ~ ., data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4147 -0.4943  0.2541  0.6794  1.9935
##
## Coefficients: (1 not defined because of singularities)
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            2.081e+00  6.469e-01   3.217 0.001383 **
## review_count          -8.508e-04  9.426e-04  -0.903 0.367174
## hotel_review_count     1.687e-03  2.822e-03   0.598 0.550330
## helpful_votes          4.428e-04  1.704e-03   0.260 0.795093
## stay_periodJun-Aug    -2.224e-02  1.215e-01  -0.183 0.854868
## stay_periodMar-May    -1.221e-01  1.200e-01  -1.017 0.309470
## stay_periodSep-Nov    -1.158e-01  1.212e-01  -0.955 0.339915
## traveler_typeCouples   4.230e-01  1.305e-01   3.242 0.001269 **
## traveler_typeFamilies  1.717e-01  1.470e-01   1.168 0.243482
## traveler_typeFriends   4.809e-01  1.554e-01   3.095 0.002083 **
## traveler_typeSolo      1.431e-01  2.296e-01   0.624 0.533225
## poolYES                9.128e-01  3.566e-01   2.560 0.010786 *
## gymYES                 2.894e-01  3.317e-01   0.873 0.383325
## tennisYES              1.582e-01  1.132e-01   1.397 0.162963
## spaYES                -3.411e-01  2.647e-01  -1.289 0.198030
## casinoYES              3.716e-01  2.633e-01   1.411 0.158862
## free_internetYES       4.933e-01  2.285e-01   2.159 0.031313 *
## stars3,5               2.629e-01  1.817e-01   1.447 0.148481
## stars4                -2.597e-02  1.621e-01  -0.160 0.872778
## stars4,5                     NA         NA      NA       NA
## stars5                 5.366e-01  1.507e-01   3.560 0.000408 ***
## room_count            -7.229e-05  5.014e-05  -1.442 0.150039
## member_years          -6.082e-04  5.377e-04  -1.131 0.258531
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9446 on 482 degrees of freedom
## Multiple R-squared:  0.1573, Adjusted R-squared:  0.1206
## F-statistic: 4.284 on 21 and 482 DF,  p-value: 1.656e-09
```

```
#vif(M1)
```

```
#alias(lm(score ~., data=df))
#cor(df, method="spearman")
#corrplot(cor(econ,method="spearman"), method = "number")
```

b. Run a multiple regression of price on the variables listed above.

*** i. (5 points) Run the model using an automatic method (i.e. stepwise, forward, backward). Explain why you chose the method. Comment on the overall significance of the regression fit. Which predictors have coefficients that are significantly different from zero at the .05 level?**

```
M2 <- lm(score ~., data=df)
#Backward Regression
train_backward = step(M2, direction="backward")
```

```
## Start:  AIC=-35.92
## score ~ review_count + hotel_review_count + helpful_votes + stay_period +
##     traveler_type + pool + gym + tennis + spa + casino + free_internet +
##     stars + room_count + member_years
##
##
## Step:  AIC=-35.92
## score ~ review_count + hotel_review_count + helpful_votes + stay_period +
##     traveler_type + pool + tennis + spa + casino + free_internet +
##     stars + room_count + member_years
##
##                       Df Sum of Sq    RSS     AIC
## - stay_period          3    1.4659 431.56 -40.209
## - helpful_votes        1    0.0602 430.15 -37.853
## - hotel_review_count   1    0.3188 430.41 -37.550
## - review_count         1    0.7270 430.82 -37.072
## - member_years         1    1.1418 431.23 -36.587
## - spa                  1    1.4825 431.57 -36.189
## <none>                             430.09 -35.923
## - tennis               1    1.7422 431.83 -35.886
## - casino               1    1.7768 431.87 -35.846
## - room_count           1    1.8546 431.95 -35.755
## - free_internet        1    4.1608 434.25 -33.071
## - pool                 1    5.8456 435.94 -31.119
## - traveler_type        4   14.2869 444.38 -27.453
## - stars                4   24.1363 454.23 -16.405
##
## Step:  AIC=-40.21
## score ~ review_count + hotel_review_count + helpful_votes + traveler_type +
##     pool + tennis + spa + casino + free_internet + stars + room_count +
##     member_years
##
##                       Df Sum of Sq    RSS     AIC
## - helpful_votes        1    0.0753 431.63 -42.121
## - hotel_review_count   1    0.2791 431.84 -41.883
## - review_count         1    0.6574 432.21 -41.441
## - member_years         1    1.2575 432.81 -40.742
## - spa                  1    1.4843 433.04 -40.478
## <none>                             431.56 -40.209
## - tennis               1    1.7636 433.32 -40.153
## - casino               1    1.7868 433.34 -40.126
## - room_count           1    1.8388 433.40 -40.066
## - free_internet        1    4.1807 435.74 -37.350
## - pool                 1    6.0048 437.56 -35.244
## - traveler_type        4   13.8679 445.43 -32.268
## - stars                4   24.0717 455.63 -20.852
##
## Step:  AIC=-42.12
## score ~ review_count + hotel_review_count + traveler_type + pool +
##     tennis + spa + casino + free_internet + stars + room_count +
##     member_years
##
##                       Df Sum of Sq    RSS     AIC
```

```
## - hotel_review_count  1    0.6453 432.28 -43.368
## - review_count        1    0.6512 432.28 -43.361
## - member_years        1    1.2506 432.88 -42.662
## - spa                 1    1.5180 433.15 -42.351
## <none>                            431.63 -42.121
## - casino              1    1.8215 433.45 -41.998
## - room_count          1    1.8267 433.46 -41.992
## - tennis              1    1.8295 433.46 -41.989
## - free_internet       1    4.1822 435.81 -39.261
## - pool                1    6.0211 437.65 -37.139
## - traveler_type       4   14.1816 445.81 -33.828
## - stars               4   24.2833 455.92 -22.535
##
## Step:  AIC=-43.37
## score ~ review_count + traveler_type + pool + tennis + spa +
##     casino + free_internet + stars + room_count + member_years
##
##                 Df Sum of Sq    RSS     AIC
## - review_count   1    0.1445 432.42 -45.199
## - member_years   1    1.2210 433.50 -43.946
## - spa            1    1.6256 433.90 -43.476
## <none>                       432.28 -43.368
## - room_count     1    1.7750 434.05 -43.302
## - tennis         1    1.7898 434.07 -43.285
## - casino         1    1.9036 434.18 -43.153
## - free_internet  1    4.3137 436.59 -40.363
## - pool           1    6.4982 438.78 -37.848
## - traveler_type  4   13.9767 446.25 -35.330
## - stars          4   23.9859 456.26 -24.150
##
## Step:  AIC=-45.2
## score ~ traveler_type + pool + tennis + spa + casino + free_internet +
##     stars + room_count + member_years
##
##                 Df Sum of Sq    RSS     AIC
## - member_years   1    1.2303 433.65 -45.767
## - spa            1    1.5647 433.99 -45.379
## - room_count     1    1.7171 434.14 -45.202
## <none>                       432.42 -45.199
## - tennis         1    1.8091 434.23 -45.095
## - casino         1    1.8439 434.27 -45.055
## - free_internet  1    4.4916 436.91 -41.991
## - pool           1    6.3995 438.82 -39.795
## - traveler_type  4   14.3171 446.74 -36.783
## - stars          4   23.9036 456.33 -26.082
##
## Step:  AIC=-45.77
## score ~ traveler_type + pool + tennis + spa + casino + free_internet +
##     stars + room_count
##
##                 Df Sum of Sq    RSS     AIC
## - room_count     1    1.5703 435.22 -45.946
## - spa            1    1.6645 435.32 -45.837
## <none>                       433.65 -45.767
```

```
## - casino         1    1.8876 435.54 -45.578
## - tennis         1    2.0334 435.69 -45.410
## - free_internet  1    4.6763 438.33 -42.362
## - pool           1    6.5643 440.22 -40.195
## - traveler_type  4   13.9570 447.61 -37.802
## - stars          4   23.4862 457.14 -27.185
##
## Step:  AIC=-45.95
## score ~ traveler_type + pool + tennis + spa + casino + free_internet +
##     stars
##
##                 Df Sum of Sq    RSS     AIC
## - casino         1    1.7281 436.95 -45.948
## <none>                       435.22 -45.946
## - spa            1    2.4635 437.69 -45.101
## - tennis         1    2.5541 437.78 -44.997
## - free_internet  1    5.9475 441.17 -41.105
## - traveler_type  4   13.2248 448.45 -38.859
## - pool           1   10.5558 445.78 -35.868
## - stars          4   22.0113 457.23 -29.080
##
## Step:  AIC=-45.95
## score ~ traveler_type + pool + tennis + spa + free_internet +
##     stars
##
##                 Df Sum of Sq    RSS     AIC
## - spa            1    0.7715 437.72 -47.059
## - tennis         1    1.7155 438.67 -45.974
## <none>                       436.95 -45.948
## - free_internet  1    6.4074 443.36 -40.612
## - traveler_type  4   13.0168 449.97 -39.154
## - pool           1    9.8292 446.78 -36.737
## - stars          4   20.9943 457.95 -30.296
##
## Step:  AIC=-47.06
## score ~ traveler_type + pool + tennis + free_internet + stars
##
##                 Df Sum of Sq    RSS     AIC
## - tennis         1    1.7225 439.45 -47.080
## <none>                       437.72 -47.059
## - free_internet  1    6.3943 444.12 -41.750
## - traveler_type  4   12.9259 450.65 -40.392
## - pool           1    9.3648 447.09 -38.390
## - stars          4   20.3979 458.12 -32.104
##
## Step:  AIC=-47.08
## score ~ traveler_type + pool + free_internet + stars
##
##                 Df Sum of Sq    RSS     AIC
## <none>                       439.45 -47.080
## - traveler_type  4   13.2597 452.70 -40.097
## - free_internet  1    8.6079 448.05 -39.303
## - pool           1   10.9097 450.35 -36.720
## - stars          4   18.7237 458.17 -34.051
```

```
summary(train_backward)
```

```
##
## Call:
## lm(formula = score ~ traveler_type + pool + free_internet + stars,
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5054 -0.5001  0.3155  0.7414  1.9487
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            2.233772   0.308391   7.243  1.7e-12 ***
## traveler_typeCouples   0.393175   0.127794   3.077  0.00221 **
## traveler_typeFamilies  0.146362   0.143009   1.023  0.30660
## traveler_typeFriends   0.457180   0.152630   2.995  0.00288 **
## traveler_typeSolo      0.158450   0.223147   0.710  0.47800
## poolYES                0.783789   0.224038   3.498  0.00051 ***
## free_internetYES       0.671151   0.215974   3.108  0.00200 **
## stars3,5               0.423510   0.158164   2.678  0.00766 **
## stars4                -0.008463   0.147920  -0.057  0.95440
## stars4,5               0.104556   0.222613   0.470  0.63879
## stars5                 0.418176   0.130605   3.202  0.00145 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9441 on 493 degrees of freedom
## Multiple R-squared:  0.139,  Adjusted R-squared:  0.1215
## F-statistic: 7.957 on 10 and 493 DF,  p-value: 6.346e-12
```

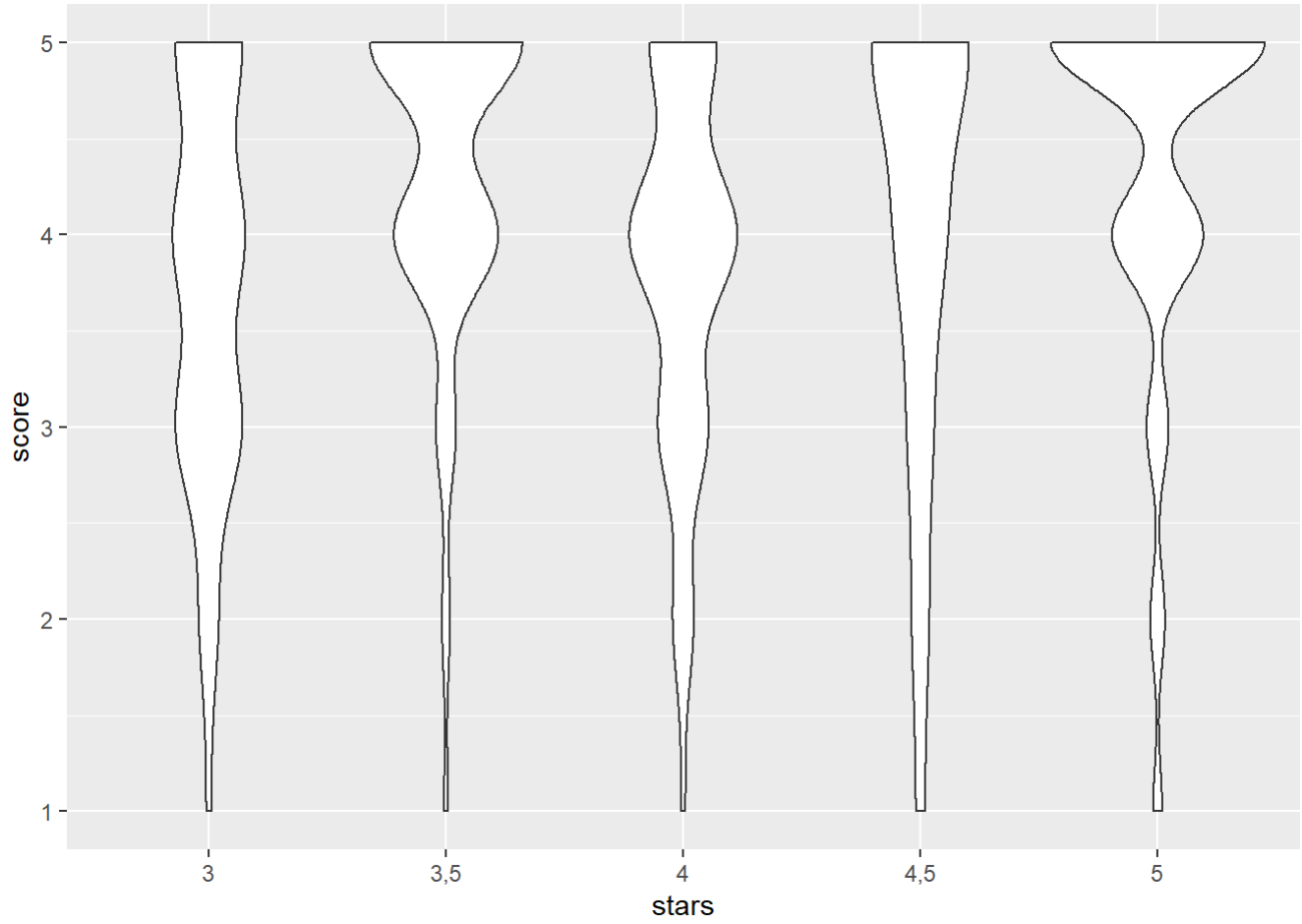
As a result of the backward regression, the following predictors have coefficients that are significantly different from zero at the 0.05 level: traveler_typeCouples, traveler_typeSolo, poolYes, free_internetYES, stars3,5 , and stars5. From the original variables that is: traveler_type, pool, free_internet, and stars

What this suggests is that the type of traveler, whether the hotel has a pool, free internet, and how many stars a hotel has have the greatest predictive power for the score given to a particular establishment. One other thing to note is that the 3,5 in stars actually means 3.5 stars. I double checked the ratings on TripAdvisor and they have half star rankings. Im not sure how to change this in R and hope we cover what the appropriate method to deal with this is.

Adj R-squared shows that this model does not fit the data well. There is likely some unresolved multicollinearity or maybe a different model would fit the data better than a linear one.

*** ii. (5 points) Using the variables above, create a visualization, which will provide an interesting story or insight within this data.**

```
library(ggplot2)
# Basic violin plot
p <- ggplot(df, aes(x=stars, y=score)) +
  geom_violin()
p
```



I created a violin plot to look at the score received by a hotel and the stars the hotel has. Note the the 3,5 and 4,5 are actually 3.5 and 4.5 respectively on TripAdvisor. This may explain why my model fails so badly, and may source of some multicollinearity. It's strange that there isn't much difference between scores given to a hotel despite having different stars. A 5 star and a 3.5 star hotel tend to have similar distributions of scores. Lets see what the effect of removing stars altogether has on the model fit.

```
df2 = subset(df, select=-c(stars))

Mend <- lm(score ~., data=df2)
#Backward Regression
train_backward = step(Mend, direction="backward")
```

```
## Start:  AIC=-17.77
## score ~ review_count + hotel_review_count + helpful_votes + stay_period +
##     traveler_type + pool + gym + tennis + spa + casino + free_internet +
##     room_count + member_years
##
##                      Df Sum of Sq    RSS      AIC
## - stay_period         3    1.4207 452.63 -22.1814
## - room_count          1    0.0016 451.21 -19.7641
## - hotel_review_count  1    0.0258 451.23 -19.7371
## - tennis              1    0.1066 451.32 -19.6468
## - helpful_votes       1    0.2822 451.49 -19.4509
## - review_count        1    0.6868 451.90 -18.9993
## - member_years        1    0.7452 451.95 -18.9342
## <none>                            451.21 -17.7659
## - casino              1    2.7431 453.95 -16.7112
## - spa                 1    2.8894 454.10 -16.5488
## - gym                 1    3.0193 454.23 -16.4046
## - traveler_type       4   12.2652 463.47 -12.2485
## - free_internet       1   16.9917 468.20  -1.1349
## - pool                1   18.1807 469.39   0.1435
##
## Step:  AIC=-22.18
## score ~ review_count + hotel_review_count + helpful_votes + traveler_type +
##     pool + gym + tennis + spa + casino + free_internet + room_count +
##     member_years
##
##                      Df Sum of Sq    RSS      AIC
## - room_count          1    0.0022 452.63 -24.1790
## - hotel_review_count  1    0.0163 452.65 -24.1633
## - tennis              1    0.1118 452.74 -24.0569
## - helpful_votes       1    0.3115 452.94 -23.8347
## - review_count        1    0.6244 453.25 -23.4867
## - member_years        1    0.8339 453.46 -23.2538
## <none>                            452.63 -22.1814
## - casino              1    2.7423 455.37 -21.1371
## - spa                 1    2.8733 455.50 -20.9921
## - gym                 1    2.9998 455.63 -20.8522
## - traveler_type       4   11.8976 464.53 -17.1046
## - free_internet       1   17.0351 469.66  -5.5612
## - pool                1   18.4463 471.08  -4.0491
##
## Step:  AIC=-24.18
## score ~ review_count + hotel_review_count + helpful_votes + traveler_type +
##     pool + gym + tennis + spa + casino + free_internet + member_years
##
##                      Df Sum of Sq    RSS      AIC
## - hotel_review_count  1    0.0160 452.65 -26.1612
## - tennis              1    0.1111 452.74 -26.0553
## - helpful_votes       1    0.3149 452.95 -25.8284
## - review_count        1    0.6333 453.26 -25.4743
## - member_years        1    0.8394 453.47 -25.2452
## <none>                            452.63 -24.1790
## - casino              1    2.7405 455.37 -23.1367
```

```
## - gym                1    3.0453 455.68 -22.7994
## - spa                1    3.3112 455.94 -22.5054
## - traveler_type      4   11.9808 464.61 -19.0120
## - free_internet      1   17.1345 469.77  -7.4522
## - pool               1   22.7935 475.42  -1.4170
##
## Step:  AIC=-26.16
## score ~ review_count + helpful_votes + traveler_type + pool +
##     gym + tennis + spa + casino + free_internet + member_years
##
##                 Df Sum of Sq    RSS      AIC
## - tennis         1    0.1077 452.76 -28.0413
## - helpful_votes  1    0.5551 453.20 -27.5435
## - review_count   1    0.6187 453.27 -27.4728
## - member_years   1    0.8373 453.48 -27.2297
## <none>                       452.65 -26.1612
## - casino         1    2.7397 455.39 -25.1199
## - gym            1    3.0721 455.72 -24.7521
## - spa            1    3.3473 455.99 -24.4479
## - traveler_type  4   11.9945 464.64 -20.9798
## - free_internet  1   17.1767 469.82  -9.3897
## - pool           1   22.9956 475.64  -3.1859
##
## Step:  AIC=-28.04
## score ~ review_count + helpful_votes + traveler_type + pool +
##     gym + spa + casino + free_internet + member_years
##
##                 Df Sum of Sq    RSS      AIC
## - helpful_votes  1    0.5729 453.33 -29.4039
## - review_count   1    0.6368 453.39 -29.3329
## - member_years   1    0.8921 453.65 -29.0492
## <none>                       452.76 -28.0413
## - casino         1    2.6333 455.39 -27.1184
## - gym            1    3.0776 455.83 -26.6270
## - spa            1    3.2471 456.00 -26.4395
## - traveler_type  4   12.0870 464.84 -22.7627
## - free_internet  1   17.8904 470.65 -10.5093
## - pool           1   22.9930 475.75  -5.0746
##
## Step:  AIC=-29.4
## score ~ review_count + traveler_type + pool + gym + spa + casino +
##     free_internet + member_years
##
##                 Df Sum of Sq    RSS      AIC
## - review_count   1    0.1082 453.44 -31.2836
## - member_years   1    0.8713 454.20 -30.4361
## <none>                       453.33 -29.4039
## - casino         1    2.7228 456.05 -28.3858
## - gym            1    3.1316 456.46 -27.9342
## - spa            1    3.3422 456.67 -27.7017
## - traveler_type  4   12.4336 465.76 -23.7667
## - free_internet  1   18.1635 471.49 -11.6042
## - pool           1   23.3491 476.68  -6.0913
##
```

```
## Step:  AIC=-31.28
## score ~ traveler_type + pool + gym + spa + casino + free_internet +
##     member_years
##
##                 Df Sum of Sq    RSS     AIC
## - member_years   1    0.8819 454.32 -32.304
## <none>                       453.44 -31.284
## - casino         1    2.6600 456.10 -30.336
## - gym            1    3.0958 456.53 -29.854
## - spa            1    3.2429 456.68 -29.692
## - traveler_type  4   12.7506 466.19 -25.307
## - free_internet  1   18.5215 471.96 -13.106
## - pool           1   23.2912 476.73  -8.038
##
## Step:  AIC=-32.3
## score ~ traveler_type + pool + gym + spa + casino + free_internet
##
##                 Df Sum of Sq    RSS     AIC
## <none>                       454.32 -32.304
## - casino         1    2.6590 456.98 -31.363
## - gym            1    3.1107 457.43 -30.865
## - spa            1    3.2362 457.55 -30.727
## - traveler_type  4   12.4318 466.75 -26.698
## - free_internet  1   18.6599 472.98 -14.018
## - pool           1   23.3369 477.66  -9.058
```

```
summary(train_backward)
```

```
##
## Call:
## lm(formula = score ~ traveler_type + pool + gym + spa + casino +
##     free_internet, data = df2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3164 -0.3622  0.1077  0.6836  1.9197
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.0659     0.5565   1.916  0.05600 .
## traveler_typeCouples    0.3965     0.1297   3.057  0.00235 **
## traveler_typeFamilies   0.1825     0.1454   1.256  0.20987
## traveler_typeFriends    0.4424     0.1548   2.858  0.00445 **
## traveler_typeSolo       0.1209     0.2264   0.534  0.59366
## poolYES                 1.4032     0.2786   5.037 6.63e-07 ***
## gymYES                  0.5105     0.2776   1.839  0.06650 .
## spaYES                 -0.3811     0.2032  -1.876  0.06127 .
## casinoYES               0.4086     0.2403   1.700  0.08969 .
## free_internetYES        0.9127     0.2026   4.504 8.31e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.959 on 494 degrees of freedom
## Multiple R-squared:  0.1098, Adjusted R-squared:  0.09361
## F-statistic: 6.772 on 9 and 494 DF,  p-value: 3.433e-09
```

Removing stars altogeth makes the model worse though, so it should stay in at least. Other techniques would likely bear more fruit for this particular dataset.