Introductory R Training
October 5, 2017
Region 5

Welcome!

- Please log into Meeting Space: https://meet.gsa.gov/r_statistics (even if you’re in-person).
- Please download the training files (see chat link) and place on your PC Desktop.
- Please mute your phone to prevent echo.

Training Agenda

Background (10-15 min)

- What is R?
- Is learning R hard?
- Business Case for Learning R at GSA

Demo (45-60 min)

- Getting Started
- Loading Data into R
- Variable Exploration and Graphing
  - Box and Whisker Plots
  - Downloading an R package
  - Correlation Matrix
- Variable Formatting
- Building a Predictive Model
  - Model Interpretation & Evaluation

Goals

❑ Learn about using R at GSA.

❑ Provide introduction to R interface.

❑ Learn about predictive modeling.

❑ Empower you to start using R!

❑ Provide basis for continued data science training GSA-wide.

Presenters

Human Capital Analytics Division
Office of Human Resources Management
U.S. General Services Administration

Chicago-based Human Capital Analytics Team presenters:

- **Matt Albucher, Program Analyst**
- **Arunice Wilbon, Operations Research Analyst**
- **Paul Tsagaroulis, Director**

What is R?

R is a GSA IT approved, **free (open-source)**, statistical computing software/language.

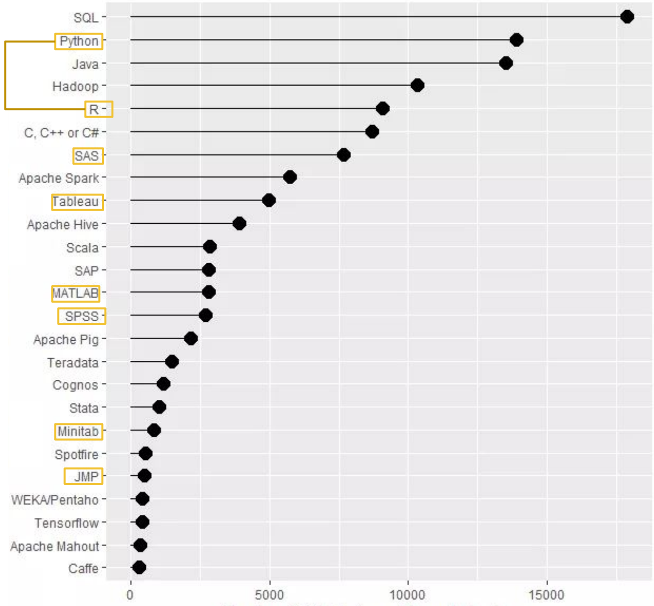
R is the "Wikipedia" of analytics software. >10,000 add-on packages and growing every day.

No GSA IT Ticket needed to download and install an R package.

R maintained by the Comprehensive R Archive Network (CRAN -- www.cran.r-project.org), mirror servers around the globe.

R can help analysts **develop data-driven solutions**, and **predict future trends**.

Data Science Job Postings, Mentions of Software, Indeed.com 2/2017



Source: r4stats.com/articles/popularity/

Checkpoint 1 : Analytics Software Experience

Has anyone used any of the following analytics software before?:

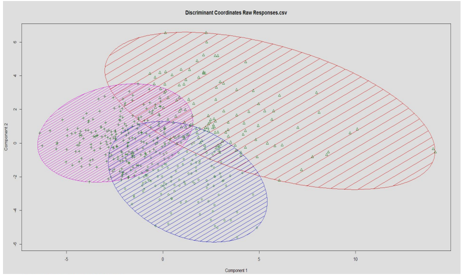
- R
- Python
- SPSS
- SAS
- Microstrategy
- Matlab
- Minitab
- None of the above

What are Practical Uses for R at GSA?

Limitless applications; Just to name a few:

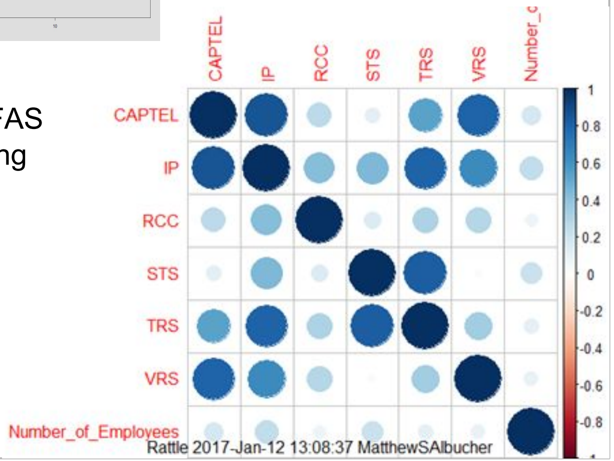
- Basic Statistical Testing
  - Are differences observed between groups of customers’ meaningful (statistically significant), or more likely due to random chance?
- Clustering
  - Understanding which types of goods/services customers tend use together to enhance category management.
- Predictive/Descriptive
  - Excelling for data visualization, testing business hypotheses, and forecasting.

Clustering and Correlation Analysis in R; GSA Examples



K-means clustering in R from GSA - GSA Region 5 survey data on technology preferences, using the R Rattle package.

Correlation analysis of FAS Federal Relay Data, using corrplot package.



Rattle 2017-Jan-12 13:08:37 MatthewSAlbucher

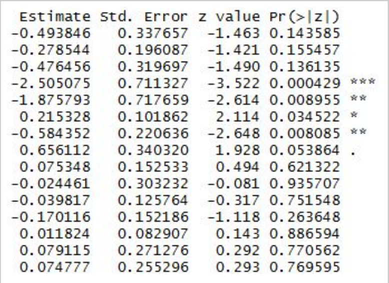
Is R Hard?

Learning R **is not** hard! Many resources within GSA and online. R is logical and intuitive. As with any language, practice is key.

Appropriately applying, understanding, and communicating statistical findings and solutions **is hard**.

R Model Output

Complex information; difficult to communicate



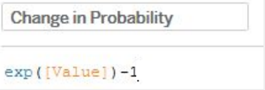
```
 Estimate Std. Error z value Pr(>|z|)
-0.493846   0.337657  -1.463 0.143585
-0.278544   0.196087  -1.421 0.155457
-0.476456   0.319697  -1.490 0.136135
-2.505075   0.711327  -3.522 0.000429 ***
-1.875793   0.717659  -2.614 0.008955 **
 0.215328   0.101862   2.114 0.034522 *
-0.584352   0.220636  -2.648 0.008085 **
 0.656112   0.340320   1.928 0.053864 .
 0.075348   0.152533   0.494 0.621322
-0.024461   0.303232  -0.081 0.935707
-0.039817   0.125764  -0.317 0.751548
-0.170116   0.152186  -1.118 0.263648
 0.011824   0.082907   0.143 0.886594
 0.079115   0.271276   0.292 0.770562
 0.074777   0.255296   0.293 0.769595
```

Model Coefficients
Log of odds ratio

Script
>write.csv

(model$coefficients,
"model.csv")

Exponentiate
in Tableau for bar chart

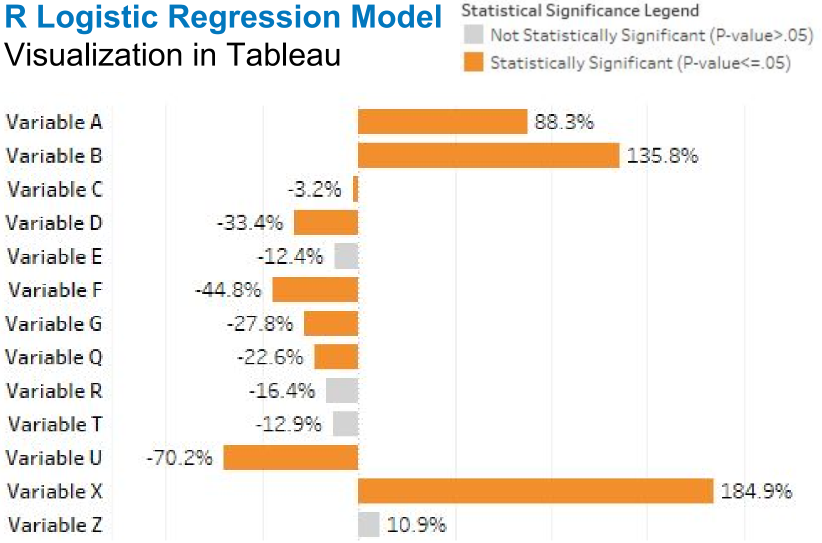


Change in Probability

```
exp([Value])-1
```

R Logistic Regression Model

Visualization in Tableau



Training Part II: Demo
Predictive Modeling - Las Vegas

tripadvisor

UCI Machine Learning Open Data

Dataset of TripAdvisor reviews by hotel variables.

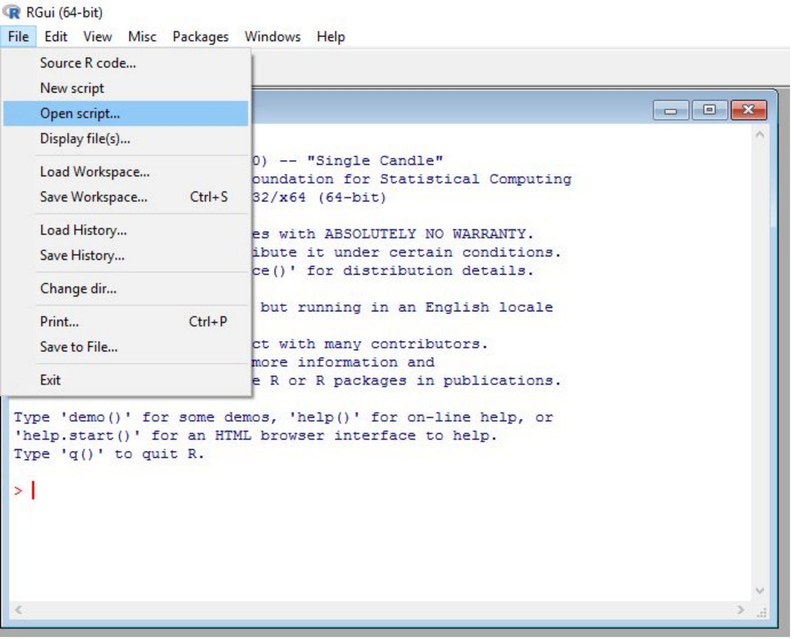
Q: What factors predict a positive hotel review on TripAdvisor?

Getting Started: Interactive R Demo

- Make sure you saved to your **desktop:**
  - R script (.R file)
  - Sample datasource (.csv)

- Open R icon on your desktop, or go to start> search programs> R x64 3.4.1.

Launch R GUI, and Open R Script



```
0) -- "Single Candle"
oundation for Statistical Computing
32/x64 (64-bit)

es with ABSOLUTELY NO WARRANTY.
ibute it under certain conditions.
ce()' for distribution details.

 but running in an English locale

ct with many contributors.
more information and
e R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

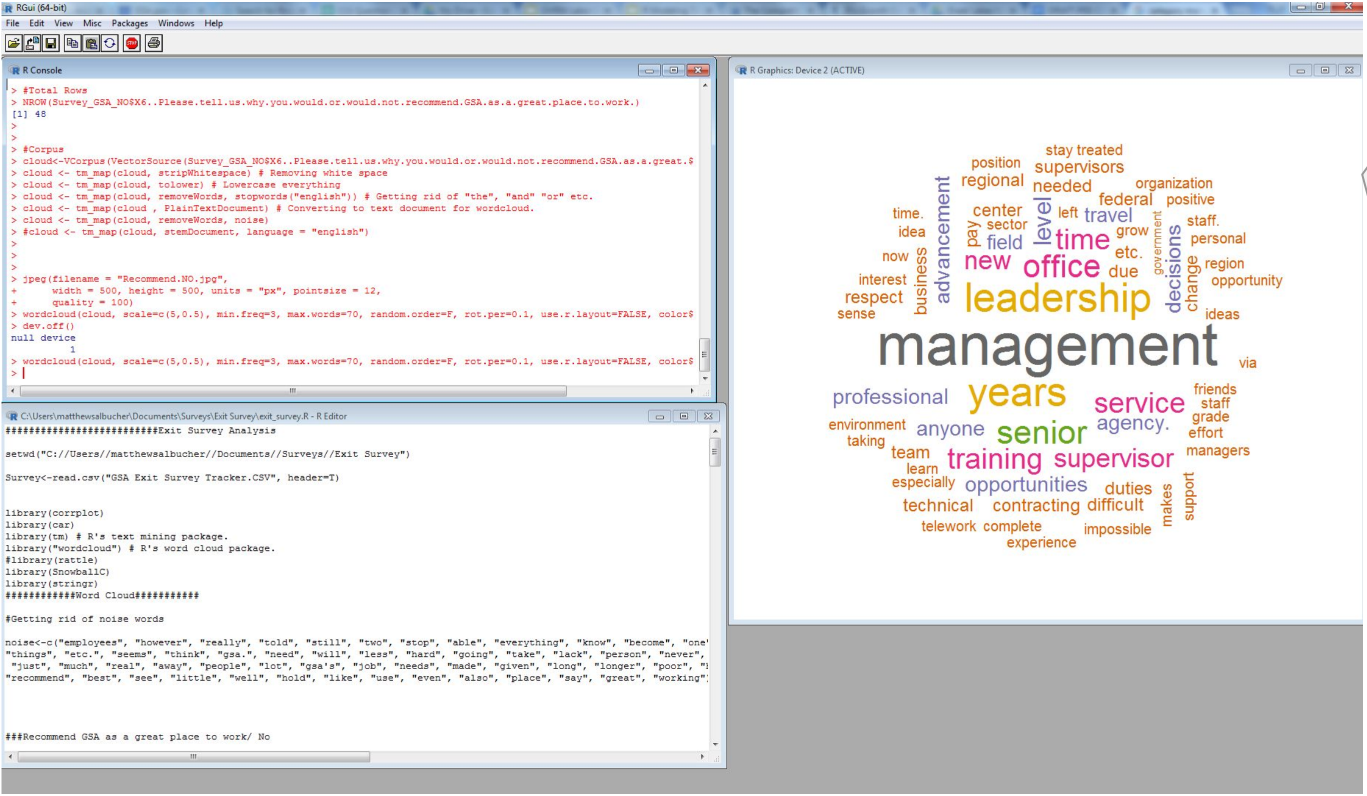
Open the training R script:

- Open R script on your desktop titled R Intro Training.R.
- A secondary window should appear, this is the R script.

Getting Started, Layout of the R GUI

R Console: Shows commands entered, results and error messages, if any.

R Scripts: Where you edit your commands, annotate and save steps for future, repeatable use.



```
> #Total Rows
> NROW(Survey_GSA_NO$X6..Please.tell.us.why.you.would.or.would.not.recommend.GSA.as.a.great.place.to.work.)
[1] 48
>
>
> #Corpus
> cloud<-VCorpus(VectorSource(Survey_GSA_NO$X6..Please.tell.us.why.you.would.or.would.not.recommend.GSA.as.a.great.$
> cloud <- tm_map(cloud, stripWhitespace) # Removing white space
> cloud <- tm_map(cloud, tolower) # Lowercase everything
> cloud <- tm_map(cloud, removeWords, stopwords("english")) # Getting rid of "the", "and" "or" etc.
> cloud <- tm_map(cloud , PlainTextDocument) # Converting to text document for wordcloud.
> cloud <- tm_map(cloud, removeWords, noise)
> #cloud <- tm_map(cloud, stemDocument, language = "english")
>
>
>
> jpeg(filename = "Recommend.NO.jpg",
+      width = 500, height = 500, units = "px", pointsize = 12,
+      quality = 100)
> wordcloud(cloud, scale=c(5,0.5), min.freq=3, max.words=70, random.order=F, rot.per=0.1, use.r.layout=FALSE, color$
> dev.off()
null device
          1
> wordcloud(cloud, scale=c(5,0.5), min.freq=3, max.words=70, random.order=F, rot.per=0.1, use.r.layout=FALSE, color$
> 
```

```
##########################Exit Survey Analysis

setwd("C://Users//matthewsalbucher//Documents//Surveys//Exit Survey")

Survey<-read.csv("GSA Exit Survey Tracker.CSV", header=T)


library(corrplot)
library(car)
library(tm) # R's text mining package.
library("wordcloud") # R's word cloud package.
#library(rattle)
library(SnowballC)
library(stringr)
############Word Cloud###########

#Getting rid of noise words

noise<-c("employees", "however", "really", "told", "still", "two", "stop", "able", "everything", "know", "become", "one'
"things", "etc.", "seems", "think", "gsa.", "need", "will", "less", "hard", "going", "take", "lack", "person", "never",
 "just", "much", "real", "away", "people", "lot", "gsa's", "job", "needs", "made", "given", "long", "longer", "poor", "I
"recommend", "best", "see", "little", "well", "hold", "like", "use", "even", "also", "place", "say", "great", "working"]



###Recommend GSA as a great place to work/ No
```

R Graphics Device: to preview visualizations generated using R code (can also be exported within code to external file)

STEP 1: Setting wd, loading data

```
###STEP 1### - LOADING DATA INTO R
```

##You will need to update the wd address to your ENT username. This is your sign-in username to your PC.

Need to insert your ENT

```
setwd("C://Users//MatthewSAlbucher//Desktop") # Set the working directory to the file location of your data. setwd() will set R to reference the file location for all read and write commands.

getwd() # Check the current working directory.
```

Tells R first column= variable names

```
LV<-read.table("LasVegasTripAdvisorReviews-Dataset.csv", sep=";", header=T) # Read in the data file. The '<-' assigns the csv file to a data.frame object called "LV".  We can then reference that object.
```

Note that # symbol in R comments out text on that line.
##Best practice to comment throughout your script
so that you/others can understand your approach.

If you get stuck at any point...

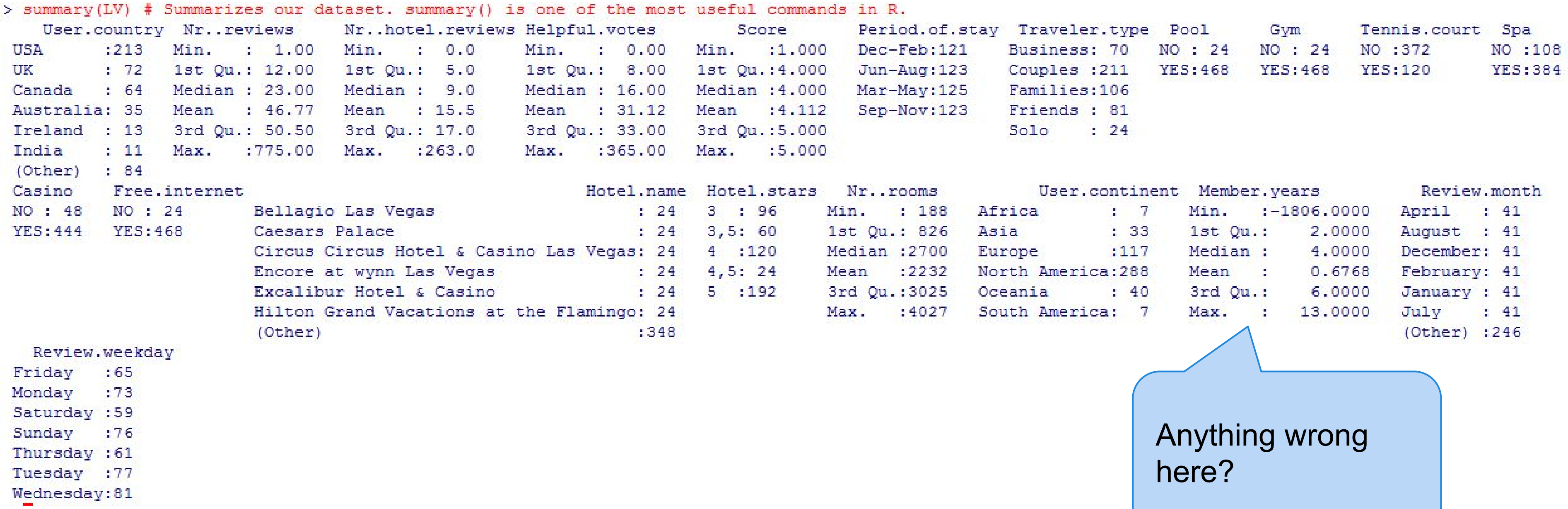
Don’t worry:

1. Just click on the R console, press “Esc”.
2. Highlight and re-run (Ctrl+R) the R script from step 1 to the current step.

Each portion of the training and slides are labeled by the step we are covering.

STEP 2: Data Summary and Cleanup

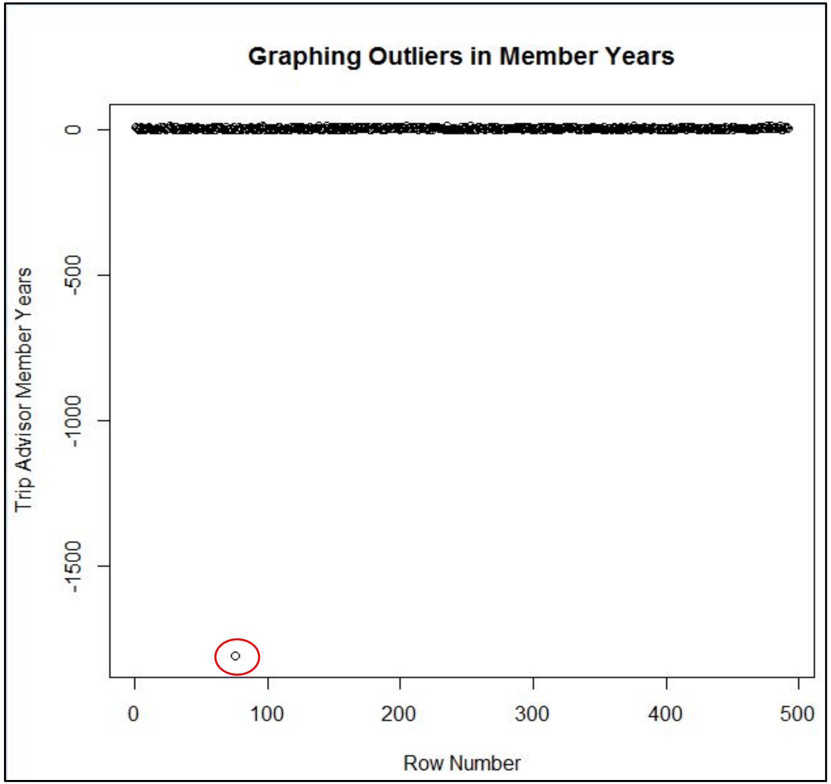
```
###STEP 2### - DATA SUMMARY AND CLEANUP
summary(LV) # Summarizes our dataset. summary() is one of the most useful commands in R.
```



```
> summary(LV) # Summarizes our dataset. summary() is one of the most useful commands in R.
    User.country  Nr..reviews     Nr..hotel.reviews Helpful.votes        Score       Period.of.stay  Traveler.type  Pool      Gym       Tennis.court  Spa
 USA      :213   Min.   :  1.00   Min.   :  0.0     Min.   :  0.00   Min.   :1.000   Dec-Feb:121    Business: 70   NO : 24   NO : 24   NO :372      NO :108
 UK       : 72   1st Qu.: 12.00   1st Qu.:  5.0     1st Qu.:  8.00   1st Qu.:4.000   Jun-Aug:123    Couples :211   YES:468   YES:468   YES:120      YES:384
 Canada   : 64   Median : 23.00   Median :  9.0     Median : 16.00   Median :4.000   Mar-May:125    Families:106
 Australia: 35   Mean   : 46.77   Mean   : 15.5     Mean   : 31.12   Mean   :4.112   Sep-Nov:123    Friends : 81
 Ireland  : 13   3rd Qu.: 50.50   3rd Qu.: 17.0     3rd Qu.: 33.00   3rd Qu.:5.000                  Solo    : 24
 India    : 11   Max.   :775.00   Max.   :263.0     Max.   :365.00   Max.   :5.000
 (Other)  : 84
 Casino    Free.internet                                  Hotel.name  Hotel.stars   Nr..rooms          User.continent  Member.years          Review.month
 NO : 48   NO : 24       Bellagio Las Vegas                    : 24   3  : 96     Min.   : 188   Africa       :  7    Min.   :-1806.0000   April   : 41
 YES:444   YES:468       Caesars Palace                        : 24   3,5: 60     1st Qu.: 826   Asia         : 33    1st Qu.:    2.0000   August  : 41
                         Circus Circus Hotel & Casino Las Vegas: 24   4  :120     Median :2700   Europe       :117    Median :    4.0000   December: 41
                         Encore at wynn Las Vegas              : 24   4,5: 24     Mean   :2232   North America:288    Mean   :    0.6768   February: 41
                         Excalibur Hotel & Casino              : 24   5  :192     3rd Qu.:3025   Oceania      : 40    3rd Qu.:    6.0000   January : 41
                         Hilton Grand Vacations at the Flamingo: 24               Max.   :4027   South America:  7    Max.   :   13.0000   July    : 41
                         (Other)                               :348                                                                        (Other) :246
   Review.weekday
 Friday   :65
 Monday   :73
 Saturday :59
 Sunday   :76
 Thursday :61
 Tuesday  :77
 Wednesday:81
```

Anything wrong here?

STEP 2: Member Years Cleanup



#Member Years Cleanup#

```
outliers<-plot(LV$Member.years, xlab="Row Number", ylab="Trip Advisor Member Years", main="Graphing Outliers in Member Years")
```

#One error can be seen with a negative value

```
LV<-subset(LV, LV$Member.years>=0)
```

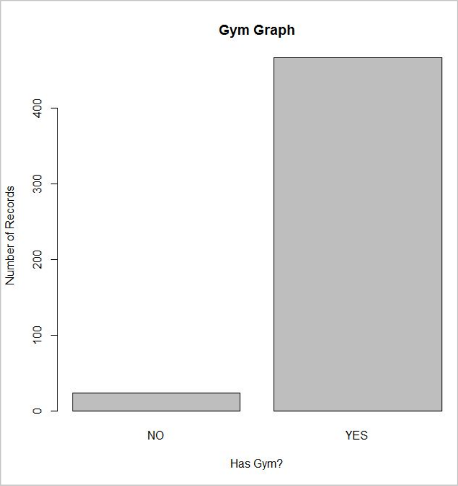
#Subsetting the LV data frame object to remove all values where member years in negative.

#Press up arrow twice to print out plot again after subsetting.
#New plot has no outliers

Types of Variables in our Data

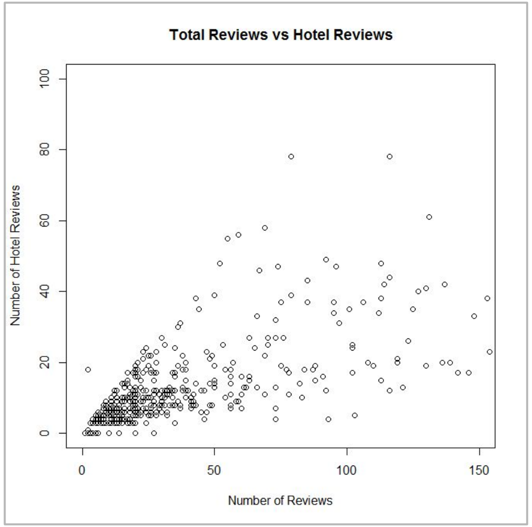
Categorical

- Made of ‘levels’ or categories.
- Ex. Yes/No, Weekday, Has Gym/No Gym
- Nominal level of measurement



Continuous

- Equally spaced, can be added, subtracted, multiplied and divided.
- Ex. Money, number of hotel rooms
- Interval/Ratio levels of measurement



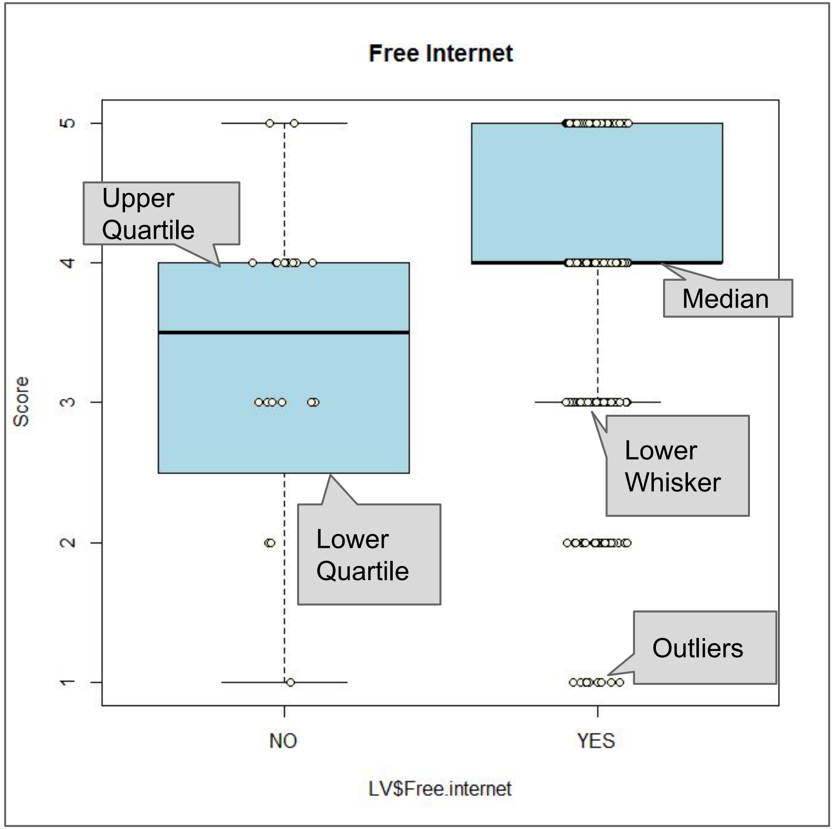
Variable Exploration/Categorization

List current model variables - names(LV) #List the variables in the dataset:

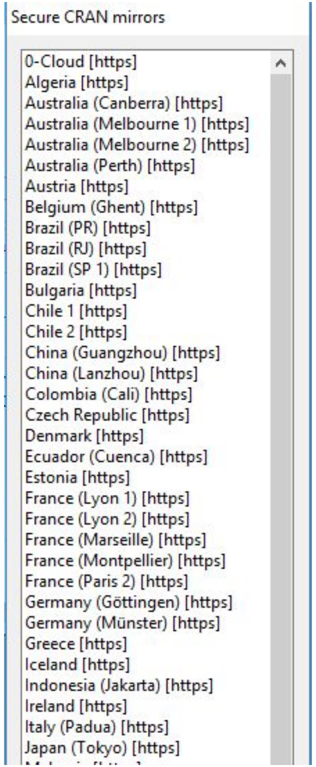
Categorical (Nominal/Ordinal)	**Continuous** (Interval/Ratio)
● User.country (many levels) ● Period.of.stay ● Traveler.type ● Pool ● Gym ● Tennis.court ● Spa ● Casino ● Free.internet ● Hotel.name (Many levels) ● User.continent ● Review.month ● Review.weekday	● Nr..reviews ● Nr..hotel.reviews ● Helpful.votes ● Member.years

- **Score (Dependent Variable)**
- Hotel.stars

STEP 3: Box and Whisker Plots



STEP 4: Correlation Matrix



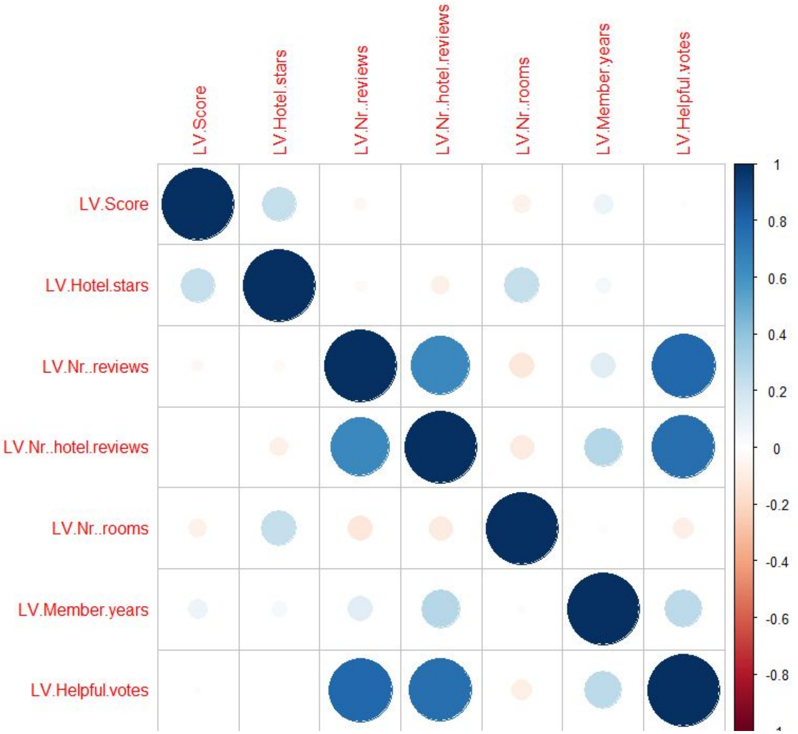
Secure CRAN mirrors

0-Cloud [https]
Algeria [https]
Australia (Canberra) [https]
Australia (Melbourne 1) [https]
Australia (Melbourne 2) [https]
Australia (Perth) [https]
Austria [https]
Belgium (Ghent) [https]
Brazil (PR) [https]
Brazil (RJ) [https]
Brazil (SP 1) [https]
Bulgaria [https]
Chile 1 [https]
Chile 2 [https]
China (Guangzhou) [https]
China (Lanzhou) [https]
Colombia (Cali) [https]
Czech Republic [https]
Denmark [https]
Ecuador (Cuenca) [https]
Estonia [https]
France (Lyon 1) [https]
France (Lyon 2) [https]
France (Marseille) [https]
France (Montpellier) [https]
France (Paris 2) [https]
Germany (Göttingen) [https]
Germany (Münster) [https]
Greece [https]
Iceland [https]
Indonesia (Jakarta) [https]
Ireland [https]
Italy (Padua) [https]
Japan (Tokyo) [https]

After running install.packages() select CRAN mirror server closest to you.

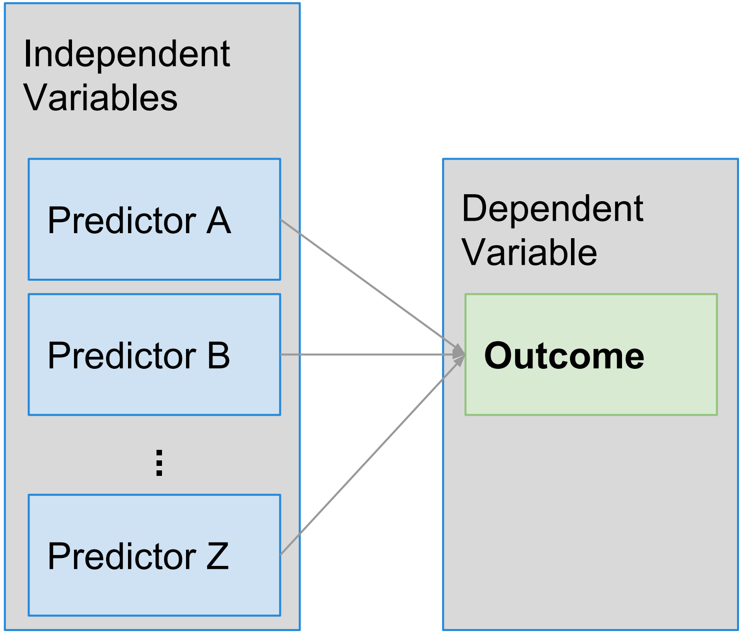
The corrplot package

matrix<-cor(x) # correlation matrix
corrplot(matrix) # Generate a correlation matrix using corrplot



What is a Predictive Model?

An **mathematical representation** of relationships between variables in data, used to better understand cause and effect, and predict future results. Independent variables (or predictors) are mathematically modeled to describe their association a dependent variable (or outcome).



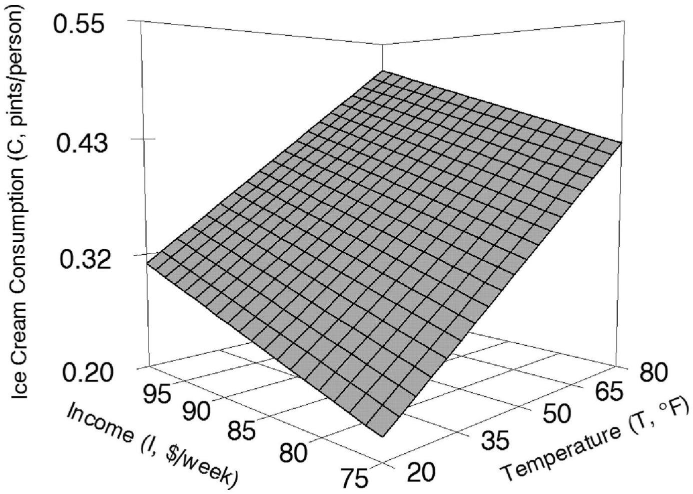
- Evaluate relationship between predictors and outcome.
- ****Hold constant other variables in model**.
- Determine which predictor(s), if any, are significant.
- Outcomes can take many forms, including **nominal, ordinal and interval and ratio.**

Predictive Models

Many types of predictive models, two of the most commonly used are:

- **Multiple Linear Regression**
  - Just like the one you learned in high school (y=mx+b), but now more variables ($y=b_1x_2 + b_2x_2…+b_nx_n$)
  - Predicts how much of something will occur.
  - Needs a continuous dependent variable.
  - Difficult to graph, need a multidimensional plot.

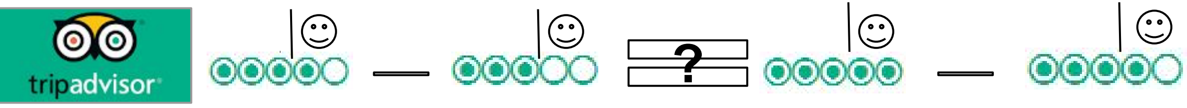
- **Logistic Regression**
  - Probability model
  - Predicts how likely something is to occur.
  - Does not need a continuous dependent variable, normally categorical variable
  - Predicts outcome instead of amount.



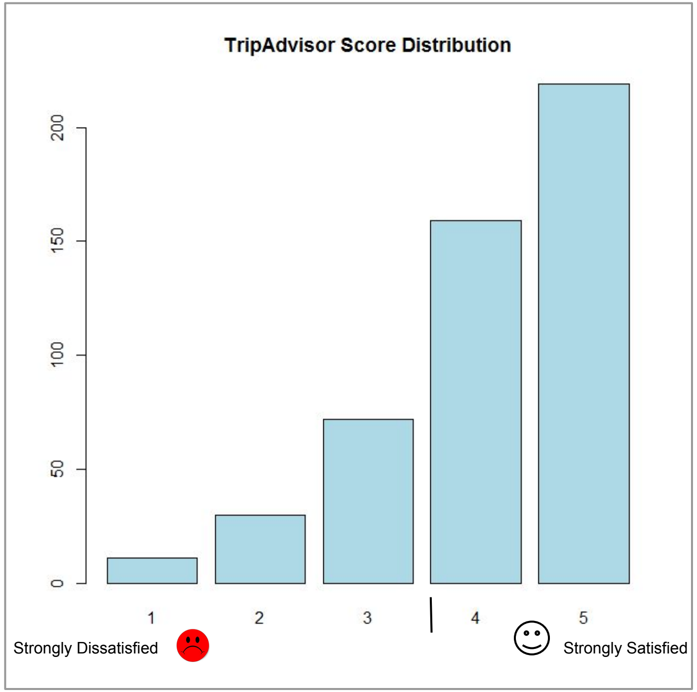
3D Plot of multiple linear regression

Photo Credit: http://circ.ahajournals.org/content/117/13/1732

Checkpoint 2: Which Model?



```
##Evaluate score distribution
plot(as.factor(Score), col="light blue", main= "TripAdvisor Score Distribution")
```



We want to build a model to predict TripAdvisor score. Which type of model should we use?

Model Equation

TripAdvisor Score = Hotel Stars + Gym + Pool, etc.

- Linear regression: Score treated as a continuous variable; spacing between each level is assumed to be the same.
- Logistic Regression: Score split as a binary variable (4 and above =1, 3 and below = 0), we'd be modeling probability of getting a positive review (4 or higher).

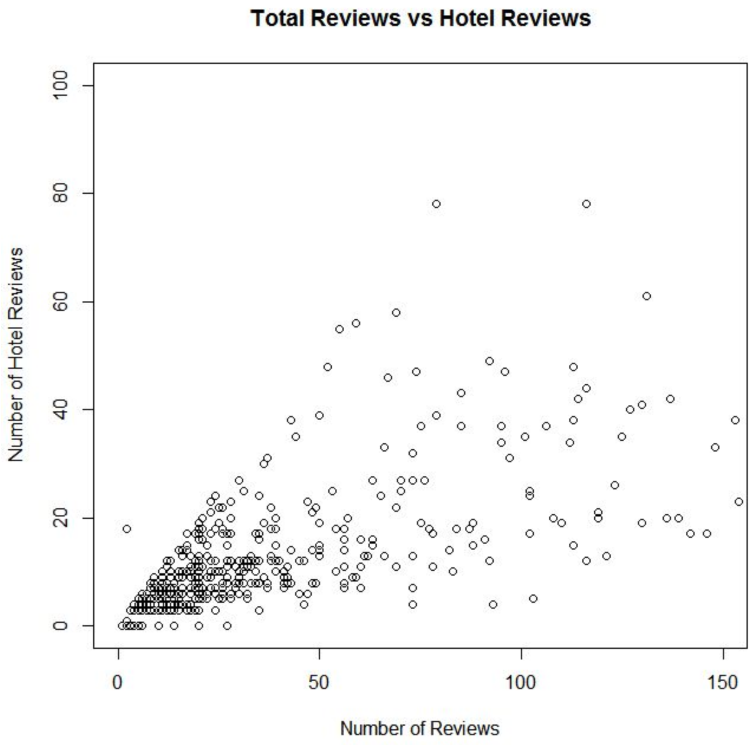
STEP 5: Variable Formatting

Goals

- Evaluate issues of multicollinearity.

Multicollinearity

What's wrong with including both number of reviews and number of hotel reviews in our model?



STEP 6: Model Equations

Goals

- Create a binary variable for positive/negative TripAdvisor review.

  -
    ```
    LV$Score_binary<-as.factor(ifelse(LV$Score>=4, "1", "0")) # Convert Score to binary nominal
    ```
- Build the model equation.

  -
    ```
    equation_logistic<-(Score_binary ~ Nr..rooms + User.continent + Member.years + Free.internet + Casino + Spa + Tennis.court +Period.of.stay+ Traveler.type + Pool + Gym + Tennis.court + Hotel.stars + Review.weekday +Helpful.votes)
    ```
- Split the data randomly 85/15:
  - Train dataset: Used to build the model.
  - Test dataset: Used to test model performance on new data.

STEP 7: Building the Models

Goals

- Run the logistic regression.
    - logistic<-glm(equation_logistic, family=binomial, data=train) #Creating model object
- Use predict() function to graph new values from test data.
    - predicted<-predict(logistic, newdata=test) # Predicting new probabilities on test data.

STEP 8: Interpreting Output

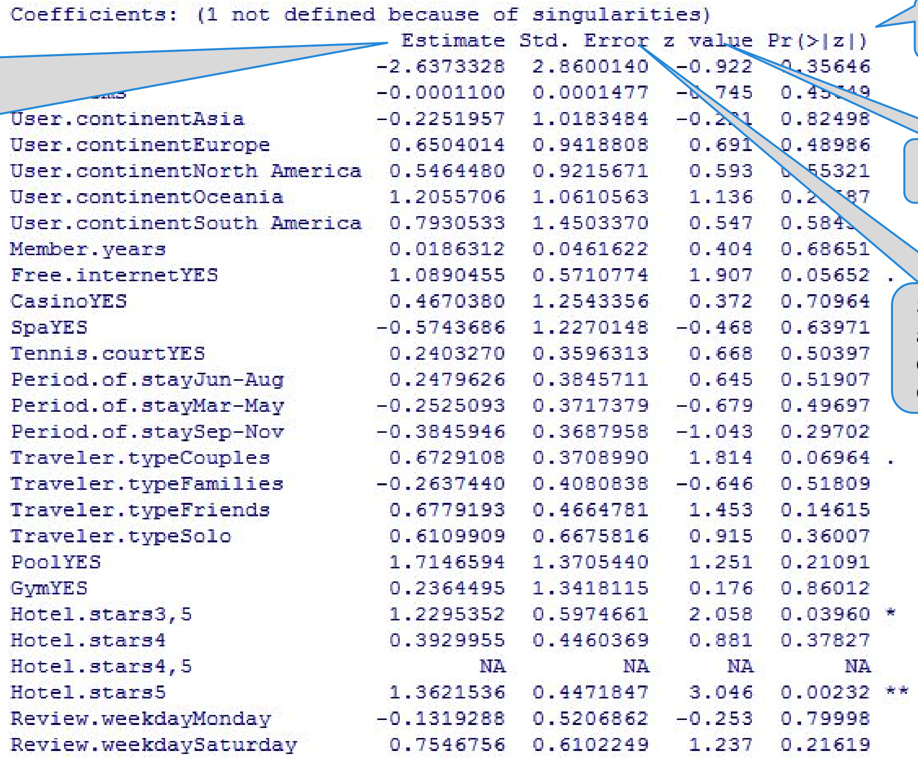
Estimate; log of odds ratio, holding all other variables constant. Positive values indicate more likely to be associated with positive rating.

P Values

# std deviations

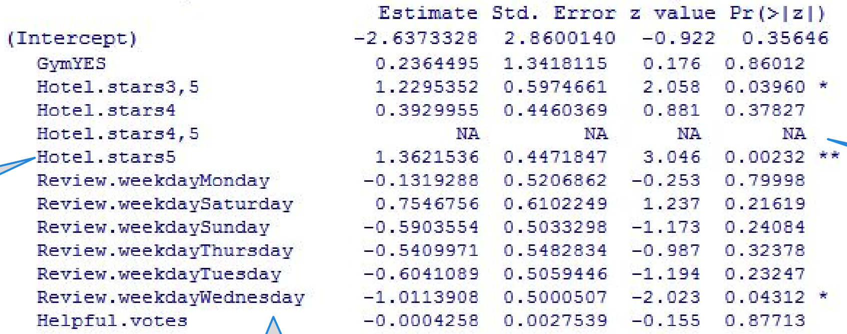
Statistical accuracy of the estimate

Variable levels (compared to baseline, not displayed but built into intercept)



```
Coefficients: (1 not defined because of singularities)
                               Estimate Std. Error z value Pr(>|z|)
                             -2.6373328  2.8600140  -0.922  0.35646
           ms                -0.0001100  0.0001477  -0.745  0.45[illegible]49
User.continentAsia           -0.2251957  1.0183484  -0.2[illegible]1  0.82498
User.continentEurope          0.6504014  0.9418808   0.691  0.48986
User.continentNorth America   0.5464480  0.9215671   0.593  0.55321
User.continentOceania         1.2055706  1.0610563   1.136  0.2[illegible]87
User.continentSouth America   0.7930533  1.4503370   0.547  0.584[illegible]
Member.years                  0.0186312  0.0461622   0.404  0.68651
Free.internetYES              1.0890455  0.5710774   1.907  0.05652 .
CasinoYES                     0.4670380  1.2543356   0.372  0.70964
SpaYES                       -0.5743686  1.2270148  -0.468  0.63971
Tennis.courtYES               0.2403270  0.3596313   0.668  0.50397
Period.of.stayJun-Aug         0.2479626  0.3845711   0.645  0.51907
Period.of.stayMar-May        -0.2525093  0.3717379  -0.679  0.49697
Period.of.staySep-Nov        -0.3845946  0.3687958  -1.043  0.29702
Traveler.typeCouples          0.6729108  0.3708990   1.814  0.06964 .
Traveler.typeFamilies        -0.2637440  0.4080838  -0.646  0.51809
Traveler.typeFriends          0.6779193  0.4664781   1.453  0.14615
Traveler.typeSolo             0.6109909  0.6675816   0.915  0.36007
PoolYES                       1.7146594  1.3705440   1.251  0.21091
GymYES                        0.2364495  1.3418115   0.176  0.86012
Hotel.stars3,5                1.2295352  0.5974661   2.058  0.03960 *
Hotel.stars4                  0.3929955  0.4460369   0.881  0.37827
Hotel.stars4,5                       NA         NA      NA       NA
Hotel.stars5                  1.3621536  0.4471847   3.046  0.00232 **
Review.weekdayMonday         -0.1319288  0.5206862  -0.253  0.79998
Review.weekdaySaturday        0.7546756  0.6102249   1.237  0.21619
```

STEP 8: More Interpretation



```
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -2.6373328  2.8600140  -0.922  0.35646
  GymYES                 0.2364495  1.3418115   0.176  0.86012
  Hotel.stars3,5         1.2295352  0.5974661   2.058  0.03960 *
  Hotel.stars4           0.3929955  0.4460369   0.881  0.37827
  Hotel.stars4,5                NA         NA      NA       NA
  Hotel.stars5           1.3621536  0.4471847   3.046  0.00232 **
  Review.weekdayMonday  -0.1319288  0.5206862  -0.253  0.79998
  Review.weekdaySaturday 0.7546756  0.6102249   1.237  0.21619
  Review.weekdaySunday  -0.5903554  0.5033298  -1.173  0.24084
  Review.weekdayThursday -0.5409971 0.5482834  -0.987  0.32378
  Review.weekdayTuesday -0.6041089  0.5059446  -1.194  0.23247
  Review.weekdayWednesday -1.0113908 0.5000507 -2.023  0.04312 *
  Helpful.votes         -0.0004258  0.0027539  -0.155  0.87713
```

5-star hotels are 1.36 log odds more likely to have a positive rating than 3-star hotels, **holding all other variables constant.**

What happened to Friday?!

What does NA here mean?

The smaller, the more significant

Checkpoint 3: Model Performance

How should we evaluate the performance of the model?

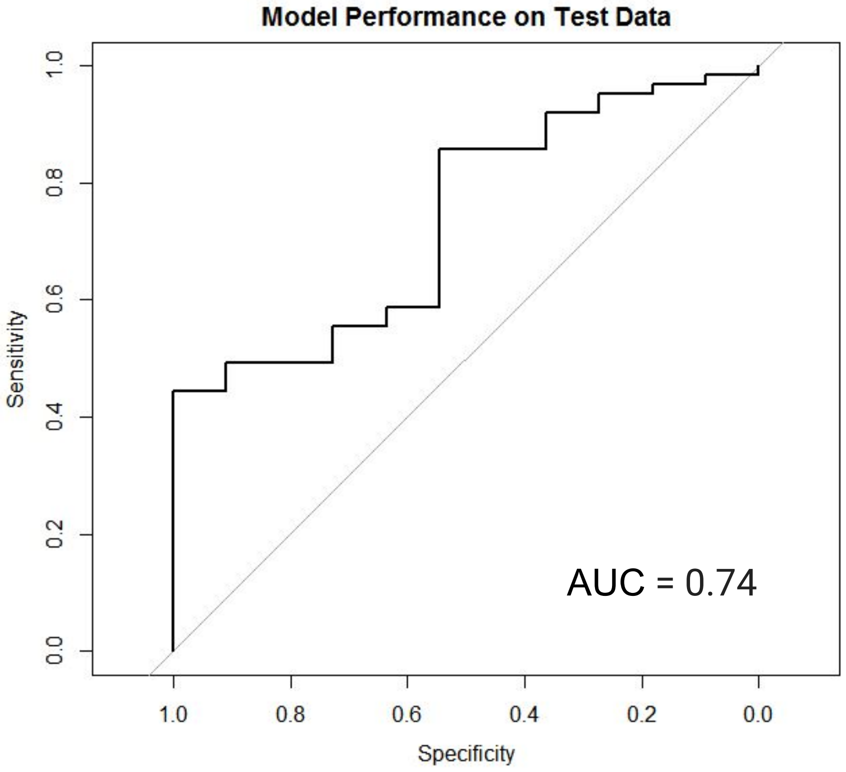
Test model predictions against:

- Train (the 85% used to build the model)
- Test (the 15% the model has not yet seen)
- The entire LV dataset (test+train)

STEP 9: Model Performance

###STEP 9 ### MODEL EVALUATION

```
install.packages("pROC")
library("pROC")
roc(test$Score_binary, predicted, plot=T,
main="Model Performance on Test Data")
```



Model Performance - Area Under the Curve (AUC)

Area Under the Curve (AUC) is an indication of model performance;

Higher AUC indicates higher True Positive Rate (TPR): False Positive Rate (FPR) ratio

Area Under Curve (AUC)	**Model Strength**	
0.9-1.0	**Excellent**	
0.8-0.9	**Good**	
0.7-0.8	**Fair**	⇦ Tripadvisor Model
0.6-0.7	**Poor**	
0.5-0.6	**Failed**	

⇦ Random Chance

Source: Obuchowski NA. Receiver operating characteristic curves and their use in radiology. Radiology. 2003;229:3–8.

Wrap-Up and Review

Today we covered:

- The business applications of learning R.
- Loading data into R (setwd, getwd and read.table).
- Data cleansing (summary and subset).
- Downloading an R package (install.packages, library).
- Variable exploration (corrplot, boxplot).
- Types of predictive models (linear and probability models).
- Building a predictive model (glm function).
- Interpreting model output.
- Testing and training data.
- Evaluating model performance.

Next Steps; Additional Resources

- R Studio--Now GSA IT approved (free version), a GUI for R.

Top 3 online resources (in my opinion):

1. The YouTube lectures of Mike Marin (this is where I got started--He walks you from the beginning most basic functions, to complex statistical testing and predictive models). If you watch these videos and follow along in R, you'll pick it up very quickly.
2. The R codeschool -- an interactive, step-by-step learning tutorial on R syntax.
3. Statistical lectures of Emory Professor Courtney Brown -- this is a bit more advanced, but he does a great job explaining the practical applications of predictive modeling in R, and provides some great code to transform the variable coefficients for practical business use.

Appendix

Building the Regression Model

Step 1: Determine the business question:

- What factors predict a positive/negative hotel review on TripAdvisor?

Step 2: What are we trying to predict (what is the dependent variable?)

- Dependent variable: Influenced by predictor(s); the response variable the model aims to predict.
- Independent variable(s): Used to predict the dependent variable. Model output indicates magnitude, directionality and significance of association between independent and dependent variables.

Step 3: What type of model best addresses the business question?

- Linear model: Measures how predictors influence the **size** of an interval, numerical dependent variable (ex. salary).
- Classification model: Measures how predictors influence the **probability** of an outcome (ex. positive/negative review, probability of earning the next higher rating on an ordinal scale).

Binary Logistic Regression

```
#Setup the model equation
#Best to create an object for the equation for repeatability if placing in multiple models
equation<-(Score ~ Nr..rooms + User.continent + Member.years + Free.internet + Casino + Spa +
Tennis.court +Period.of.stay + Traveler.type + Pool + Gym + Tennis.court + Hotel.stars + Review.weekday
+Helpful.votes )


##Building the model
model<-glm(equation, family=binomial)
#Creating model object G
#glm="generalized linear models"' family=binomial tells R to use binary logistic regression
summary(model) # summarize the model results
```

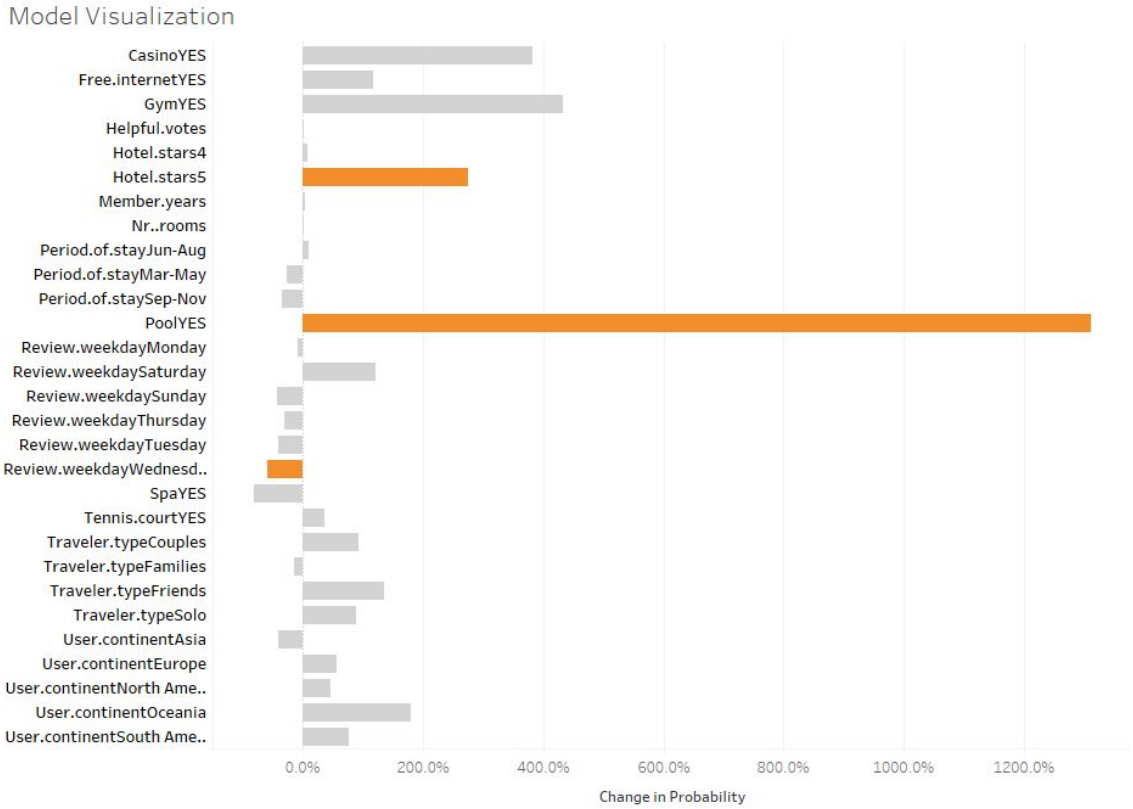
Making Use of Model Output

- The model coefficients indicate the effect of each independent variable on the log of the odds ratio of the dependent variable, holding all other variables constant.
  - Odds ratio (OR) represents odds of a positive hotel review within each variable, divided by the odds of a positive review of the baseline (non displayed) variable.

- To obtain odds ratio, need to exponentiate log odds:
  - OR = exp(log(OR))
  - Can do this in R, or export to Excel, Tableau or other visualization software.
  - `round(exp(cbind(Estimate=coef(model), confint(model))), 2)` # Calculating odds ratios

- OR is much more useful to stakeholders than on the logarithmic scale.
  - OR greater than 1 indicate positive association with target variable.
  - Subtract 1 from exponentiated coefficient to determine change in probability for each variable level.

Example of Model Visualization

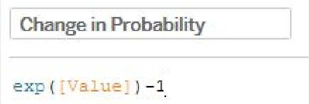


R Script

```
>write.csv
    (model$coefficients,
    "model.csv")
```

Exponentiate

in Tableau for bar chart



Change in Probability

```
exp([Value])-1
```

Real-life Examples at GSA

Predictive models have both explanatory and predictive power.
Explanatory value often overlooked.

Independent Variable Side	**Dependent Variable Side**
What factors are associated with: ● A higher/lower transfer rate from GSA? ● Employees staying longer/shorter past optional retirement date? ● Higher/lower salaries? ● Higher/lower performance evaluations?	*Who is predicted* to be most likely to: ● Transfer from GSA? ● Retire sooner? ● Have higher pay (and by how much)? ● Have higher performance evaluations?

STEP 1: Setting wd, loading data

What have we accomplished in step 1, the most important step?

1. We’ve set working directory to your desktop.
2. We’ve used the read.table() command to load our raw data into R, and create a new object called LV.

*R is an object-oriented language. Language is structured to create, manipulate and produce objects of many types. Differs from a command-oriented language such as SAS.

Common Object Types in R include:

- Data frames (like our LV object).
- Vectors (numerical, character, factor, logical) - Often take the form of variables within a data frame (such as LV$Score).
- Lists
- Matrices - Ex. correlation matrix (which we will demonstrate in this training)