Anh Nguyen

Course: IS6030

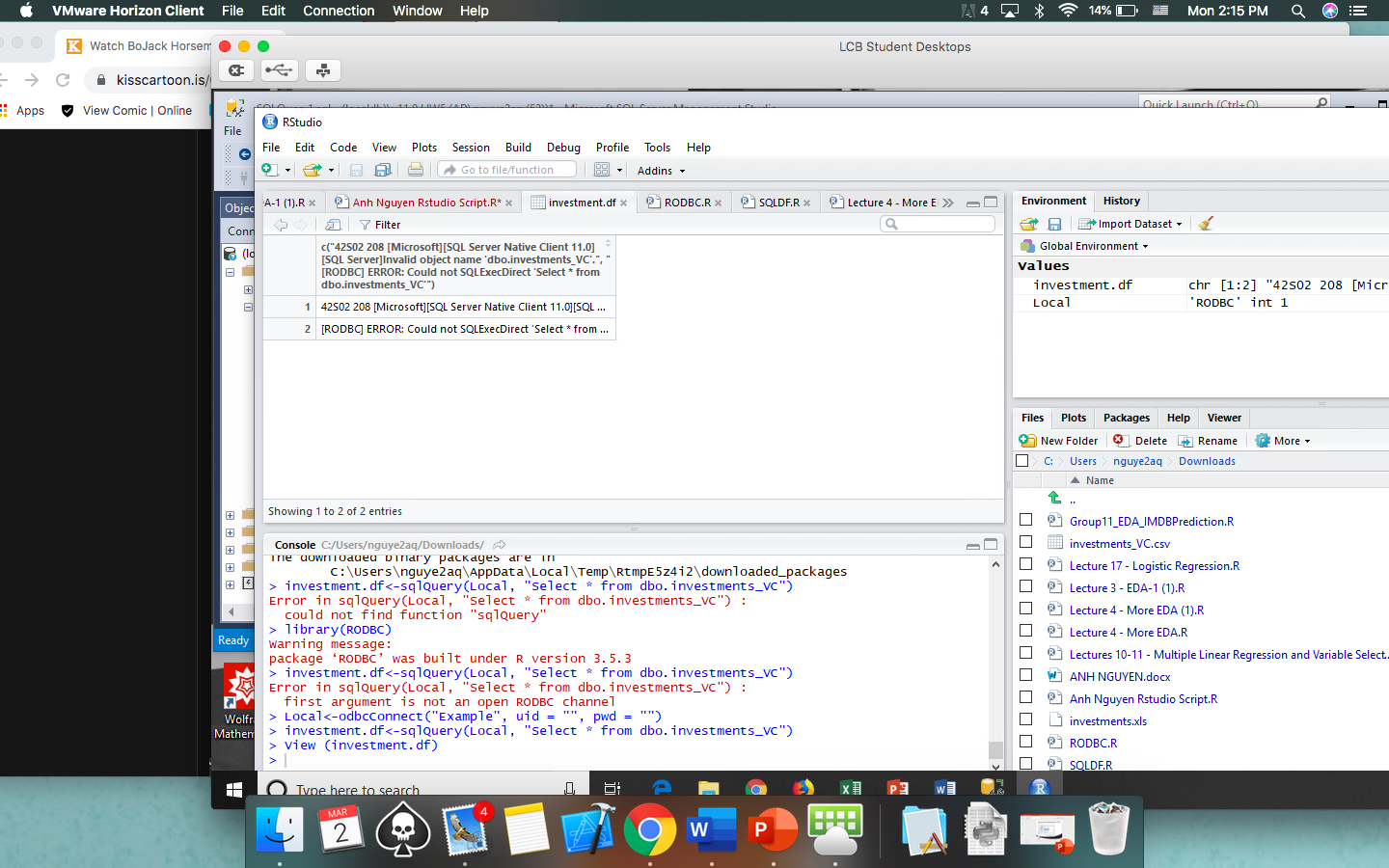
Instructor: Lichei Wen

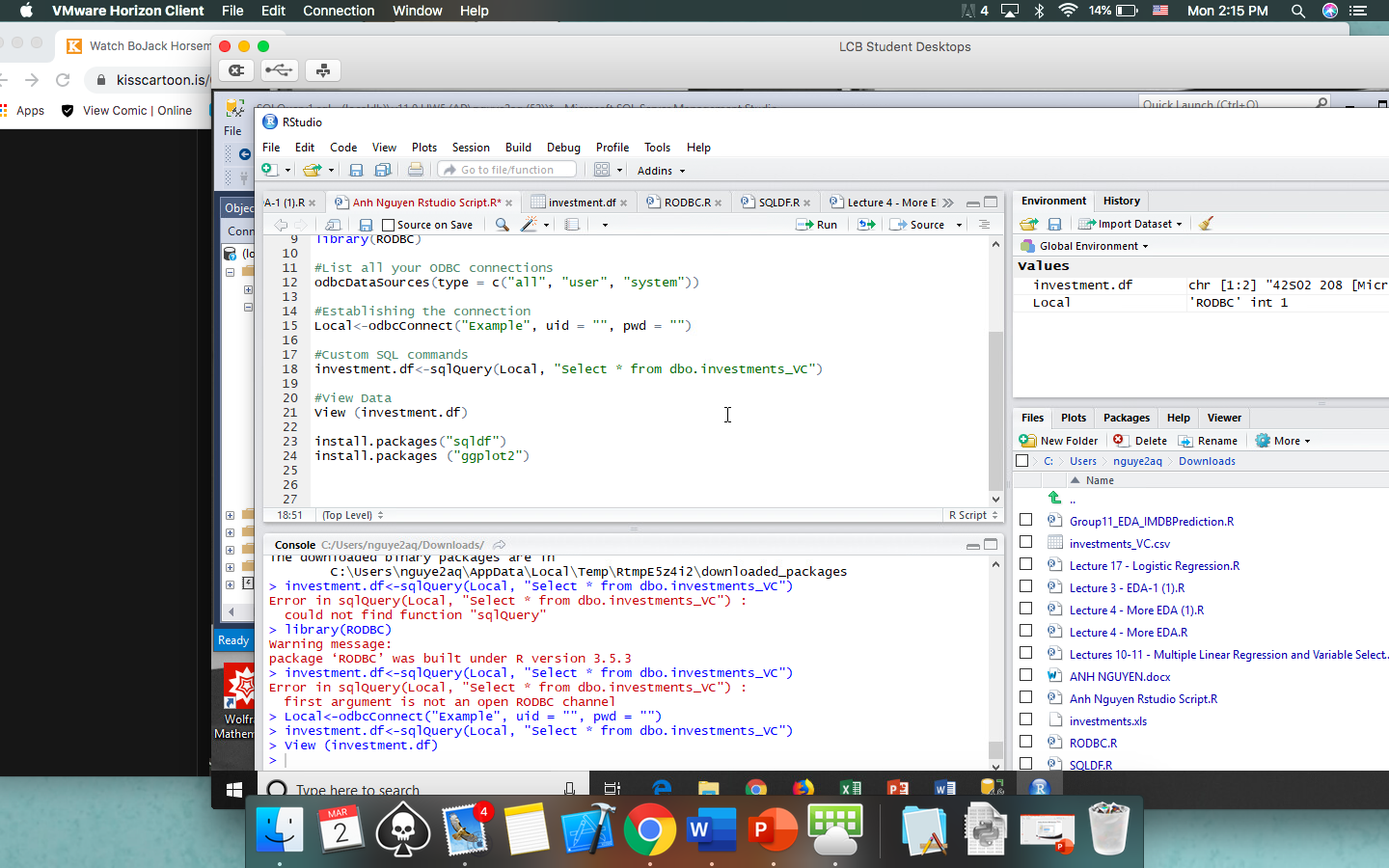
Homework 5

**ANALYSIS ON STARTUP INVESTMENTS**

Source: Kraggle – Andy\_M

Note: I have a mac and I use VMWare, and since I couldn’t connect the Control Panel to ODBC SQL Server 11.0, or it could be some other issues, I cannot connect SQL with R.

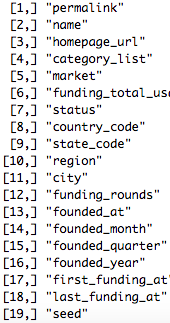




Therefore I will code SQL in as well as performing R code on the regular dataset. There will be 2 sections to every action ( Pure R and SQL R).

**I. Overview**

The dataset is from <https://www.kaggle.com/arindam235/startup-investments-crunchbase>, containing the series of investment on 54294 companies. It’s extracted by user Andy\_M on Kraggle, from the Crunchbase website. There are 39 variables describing the journeys of these companies, shown below:

: string

: string

: string

: string

: string

: string

: string

: string

: string

: string

: string

: numeric

: datetime

: datetime

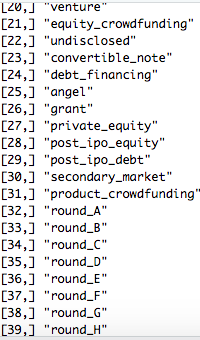
: string

: numeric

: datetime

: datetime

: numeric

: numeric

: numeric

: numeric

: numeric

: numeric

: numeric

: numeric

: numeric

: numeric

: numeric

: numeric

: numeric

: numeric

: numeric

: numeric

: numeric

: numeric

: numeric

: numeric

**II. Pre-processing the dataset**

\* Comments about the dataset:

- There’s a lot of NULL / NA / Missing data cell

=> For the continuous variables (funding\_total\_usd) we add into the missing data with the median of the rest of the data for that variable.

=> For specific analysis, we’ll create a table just with those variables, delete all the missing cells, and perform analysis on the remaining data.

- The column for funding \_total\_usd has weird comma scheme. I will cast it into a numeric variable

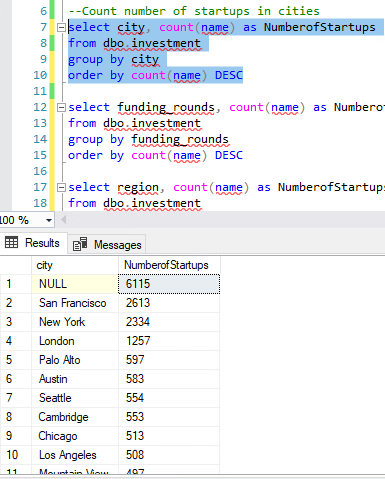
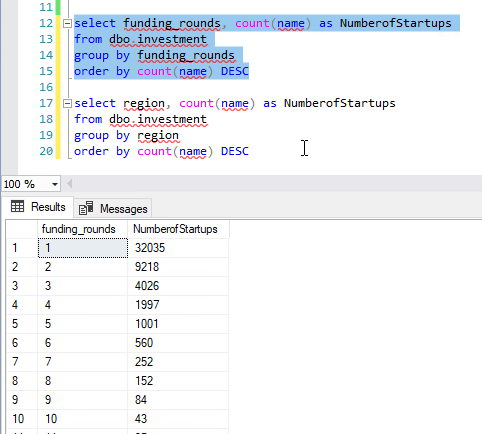
(I could not find a way to do it in R purely, so I modified it in excel first, by find and replace all commas with nothing, and replace all ‘ - ‘ cells with no values. Even with the SQL method, I doubt it will work)

\* Any other Pre-processing will be discussed in the R file.

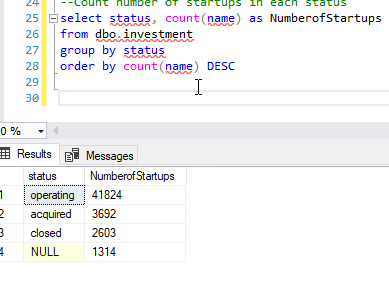
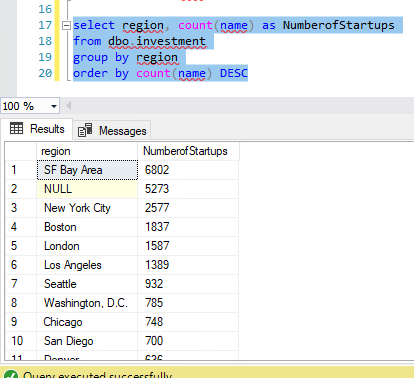
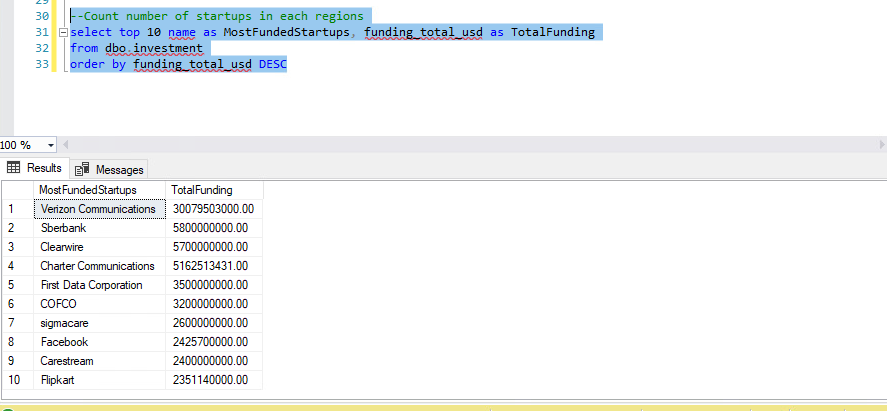
**III. EDA**

1. Explore the data set

To better understand the data we’re dealing with, I find out the number of startups in a few categories. (Again, I can’t access SQL through VMWare, so I did it natively and code them in R)

=> Although there’s a lot of data missing, the cities with the most startups are predictably San Francisco, New York, London, and Palo Alto. 

=> Also understandable is that most startups made first round funding, while few made more than 10.

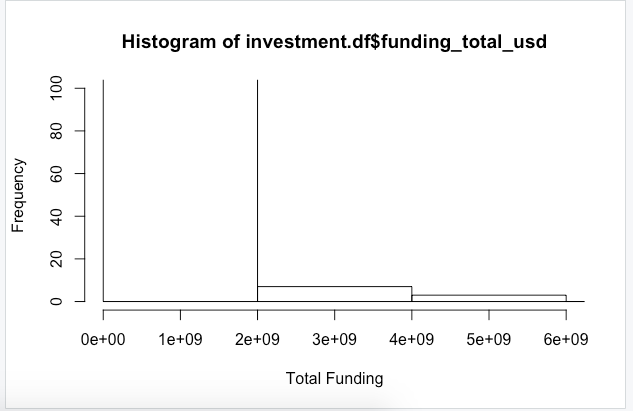
  
=> We also found out that most startups are in the SF Bay Area and North East Coast of America, and for this data set, most startups are still operating. 

=> For the most funded startups, Verizon tops at roughly 30 billions dollars. Facebook stands at number 8 with 2.4 billion dollars.

**IV. Visualization**

\* I lost my Tableau key, so I can’t use that. I’m sorry. I’ll be using Excel for this

1. Total Funding of 15 cities with the most funding in the dataset.
2. Total Funding of the top 20 industries in startups
3. Histogram of the frequency of total funding



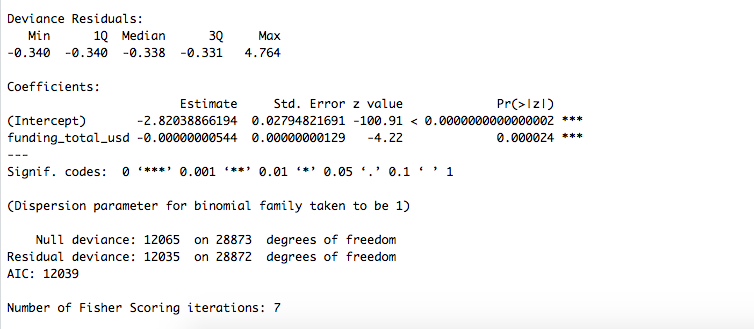
=> Most companies get funded for less than $2,000,000,000, while there are only about 10 companies that are funded more and even fewer funded more than 4 billions.

**V. Analytics**

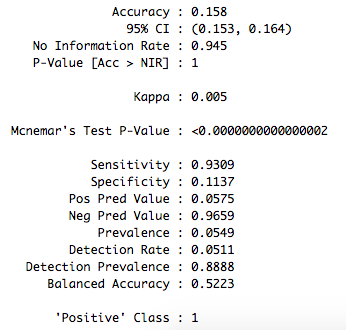
I will be do a linear regression to determine if the amount of total funding decide whether the companies will still be Operating, Shut down or Acquired.

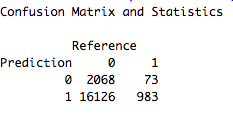
1. Creating dummy variables for these 3 statuses and Delete other variables that are irrelevant
2. Partition the data 60-40 for training set and validation set
3. Running Logistic Regression on this model

=> Since the variable we’re trying to predict is a binary variable, we should use Logistic Regression.

=> We found that although the P value is significantly less than any alpha of any confidence level. Therefore we can disprove that the variable funding total is an ineffective predictor.

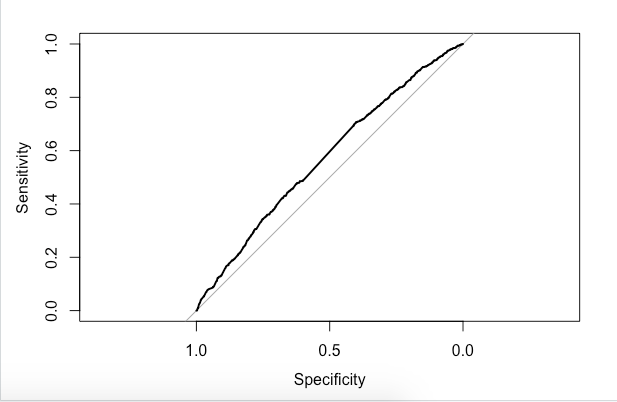
1. Model Fitting and Confusion Matrix



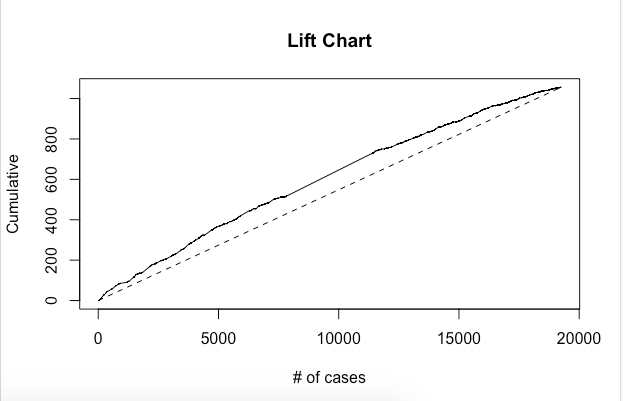
=> If we choose our cutoff point as 0.05 (reasonable, since most companies here are still operating or acquired, and only good companies would be included in the dataset, so the companies here are more likely to be still active than closed), we have these following statistics.

=>It suggest that the Accuracy is not very good, and Sensitivity is really good. That means choosing another cutoff point might be beneficial. Someone with more industry knowledge will be the better person to do this.

1. Performance Evaluation
2. ROC Curve

Area Under the Curve: 0.571

=> Both the shape and the AUC suggest that this model is closed to useless.



1. Lift Chart

=> Once again, the lift chart suggests that there’s not much more effective to use the model than just randomly pick cases to pick out a certain number of closed startups.

**Conclusion: Total Funding to a startup alone cannot predict effectively whether the startup is going to be closed**