Data Science

1. Math and Statistics
   1. Basic Statistics
      1. Mean
      2. Median
      3. Standard deviation
      4. Minimum, maximum, quartiles, and interquartile range
   2. Observational study/experiment
      1. Sample
      2. Census – entire population
      3. Observational study
      4. Experiment – the ideal way to test the data; parameters are controlled
      5. Correlation does not imply causation!
   3. Probability
   4. Normal Distribution
   5. Binomial, Geometrical, Poisson, Exponential, and Gamma
   6. Sample Mean and Central Limit Theorem
   7. F-distribution
   8. Chi-Squire
   9. Linear Regression
   10. Multivariable Linear Regression
   11. Correlation
   12. Bayes Estimate; conditional probability
2. R
3. Install and load libraries
4. Rmarkdown
5. HTML
6. PDF
7. Dplyr
8. Filter – SQL where (select rows)
9. Arrange – SQL order by
10. Select – SQL select (select columns)
11. Spread – transpose rows
12. Gather – transpose columns
13. Mutate – ads new variable/column
14. Group\_by – SQL by; group\_by(data, variable)
15. Summarize – mean, record count, sum, median, etc.
16. Distinct – dedup
17. Top\_n – the same as head()
18. Count – record count
19. Rename – rename columns
20. Separate – similar to mutate; break a column into multiple columns
21. Inner\_join/outer\_join/left\_join/right\_join
22. %>% - pipe
23. Long vs wide data
24. Tidyr
25. Ggplot2
    1. Histogram
    2. Boxplot
    3. Scatterplot
    4. Bar graphs
    5. Ggplot() – creates graph
    6. Aes() – declaration of graph parameters; x and y axis
    7. + - to add additional components
    8. Example: ggplot(data=dat, aes(x=time, y=total\_bill)) + geom\_bar(stat="identity")
26. Stringr
    1. Str\_sub
    2. Str\_length
    3. Str\_to\_upper
    4. Str\_sub
    5. Str\_detect
27. Subset
28. Import data
    1. CSV – read.csv
    2. Text
    3. Excel
    4. SQL (mySQL)
29. 8. Vectors, list, and dataframe
30. C – vector
31. List – a[2]
32. Data.frame
33. Dataframe$column
34. 9. Summary function
35. 10. Head and tail functions
36. 11. Data types – factors, strings, integers, etc.
37. 12. Plot function – scatterplot
38. 13. Table function – frequencies
39. 14. as.numeric – converting data types
40. 15. is.numeric – testing data types
41. 16. class function – object type
42. 17. ifelse – ifelse(data, condition, 1st output, 2nd output)
43. 18. read\_html library
    1. Html\_table() – to read html tables
    2. 19. names – to set column names
44. GitHub
45. SQL/MySQL
46. Create/Delete database, table, and view
47. Select data from the table
48. Update data
49. Where – condition to limit data
50. Case when … then … else … end – the same as if
51. Web Scraping
    * 1. HTML – text/markup language; HTML tables
      2. Tags – head, title, and body
      3. Read\_html – r command to read a html page
      4. Html\_table – r command to read a html table
      5. JSON - data
      6. XML – data/markup language
      7. Java – automation and links
      8. API – interface to extract web data
      9. HTTP – protocol to connect to websites
         1. Get
         2. Put
         3. Post
         4. Delete
      10. CSS – language to format web sites
      11. RCurl – R library to scrap web
      12. Xpath – language to work with XML
      13. Rvest – R library to read websites
      14. Jsonlite – R library to extract and work with Json files
52. MongoDB – NoSQL Database for unstructured data
53. Visualization
54. Python
55. Data Science Process
56. One version (“R for Data Science” book)
    1. Import Data
       1. CSV/Text files
       2. Web
       3. Using R/Python/MySQL
    2. Tidy Data
       1. Using R; libraries – tidyr and dplyr
    3. Transform Data
       1. Outliers
    4. Visualize
       1. Using R – ggplot2 library
       2. Using Python – matplotlib library
    5. Model
       1. Regressions – linear and logistical; LASSO and Ridge
       2. K-neighbors
       3. SVM – support vector modeling
       4. Random Forest
       5. Clustering
       6. Neural Net
    6. Communicate/Presentation
       1. Rmarkdown
       2. PowerPoint
       3. Youtube
57. Version 2 (R class; Geron Project checklist)
    1. Frame the problem and look at the big picture
58. NLP – Natural Language Processing
    1. Sentiment Analysis
       1. Tidytext – r library to work with text
    2. Document classification
       1. Corpus – collection of string with additional information
59. Recommender System
60. Graph Database –
61. Cloud Computing – distributed computing
    1. Hadoop; Apache Hive – SQL like language for Hadoop
    2. AWS
    3. Microsoft Azure
    4. Apache Spark
62. Machine Learning