Contents

1 OVERVIEW

This is the report document for Lab1: Data Clients and Information Servers that deals primarily with data collection. In this lab, we learn about Exploratory Data Analysis its performance gives one a fantastic exposure to a vast domain associated with twitter tweets by learning from the example of new york time from textbook, and various related statistics. This specific data set gives us a great overview of datas related knowledge. The goal of this project will follow a linear progression. Of course we will get familiarize ourself with the data set by cleaning, formatting, and compressing to a manageable size and as we desire. We will also categorize the data as appropriately so that we can apprach to our desire datas more by using different methods and visualizations. We will perform EDA for this lab and which consist of not various summaries and visualizations such as plots and distributions but a lot of background knowledge in many sectors. We will further extend our analysis to get the best data what we desire. We need to have also knowledge with maps google api and provide some other cool and interesting visualizations that express the data in a more easy to understand. This lab need to be all summaries and graphs as informative as possible.

1.1 Description

Some requiement of Exploratory Data Analysis for this lab are as in the following list of objectives, learn and explore statistical modeling, learn R language for data analysis,learn how to utilize tools such as R Studio, learn how to structure data into a more useful format, learn how to read and analyze results from a data set. Become familiar with a few different domains. Get some practice using maps and coordinate systems. Provide useful metrics and visualizations.Provide code and work in an easily reproducible fashion. Provide detailed comments and descriptions within the code. Present results in an easy to understand manner.

1.2 Usage

For this lab, it is better if we run it on jyupter as describe in the lab manual.However, we will aslo import the .csv file from right directory into RStudio for those desktop who does not have Microsoft office to open .csv, or xls file. There are couple of files end with .csv as test case for the data output and please import into RStuio to see the format and text. Mainly, we will complie our code on Jyupter.

1.3 Division of work

This lab is seperated into 3 notebook as report. Please take a look at lab maunal for more informations.

2 PROJECT APPROACH

For this particular lab, we will need to read Chapter 2 of Doing Data Science until we understand the material completely to develop our knowledge and some actual Exploratory Data Analysis. Before we are ready to proceed to the lab, first we need to slove the NY Times example in Chapter 2 of the book. We are going to use both R Studio and Jyputer for this example and the rest of the project. From the first example, we will learn how to explore data from a single file and extend this exploration to multiple files, a skill we will need later on,and aslo hope to delvelop into different functionalities of R such as plots, distributions, summaries, and other complicated functions.

2.1 Practice

After we completed the questions in the New York Times example, we will proceed with RealDirect, where we will apply and reaffirm my new knowledge of statistical analysis using R. We will learn some more through the second example and to fully understand the domains as we have encountered thus far. It is essential that to pay close attention to the distributions, summaries, and methodologies we are utilizing. We cannot forget to streamline and comment the code so that it is easily accessible to any user. Definitely, we will also play around with R and its functionalities to ensure that I have a good understanding of a large set of practices

that will indubitably prove helpful when we start working with my own data set.

2.2 Lab1 work

When we begin work on the lab1 data, we hope to have a good enough understanding of methodologies so that I can efficiently and properly perform Exploratory Data Analysis. This should hopefully be the case after the New York Times and RealDirect examples. Since the data set is rather large and the source site informs us of missing values, I will clean up those missing values. The tables also seem to contain a bunch of redundant and unnecessary columns that we will need to get rid of. Since our analysis will focus on tweets we have to group by locations of tweet on sepcific search term and by different categories, this is why we will categorize the data set as appropriate. I will then use summaries and visualizations, applying only those that seem most relevant, a skill which should have acquired through earlier parts of the assignment. I shall express the scatter plots on a map of the United States. Once again I will comment my code well and make it easily accessible.

2.3 Outcomes

From this lab, we have learned plenty about R, R Studio, exploratory data analysis, the domain of the given data set, and interpreting data through summaries, plots, distributions, and other visualizations. I am also now familiar with extending analysis over multiple files, by compressing the data in a useful manner.

2.4 New York Time Data Set, Questions, and Outcomes

We have learned plenty about R, R Studio, exploratory data analysis, the domain of the given data set, and interpreting data through summaries, plots, distributions, and other visualizations. I am also now familiar with extending analysis over multiple files, by compressing the data in a useful manner.

2.5 RealDirect Questions and Outcomes

We would advice engineers to log as much data as possible. It is certainly true that the more the data, the easier it would be to form a variety of conclusions. To be more specific however, it would be most useful to log data such as sale prices, locations, crime, schools, types of buildings, building age, taxes by type, and perhaps proximity to nearby points of interest. The way we are going to use the data to monitor product usage could be by analyzing what types of places each user is browsing, and learning what they may or may not like. This is very useful if we want to give users suggestions. We can also log what sort of places are being actually sold so that we can infer the likelihood a listing would have of selling at a certain price range. We can build this data back into the production website by providing useful statistics, summaries, and visualizations to each user so that they know exactly what sort of realty they are looking into. We have to evaluate what information is useful to the user and what isn't. It is certainly true that if we put too much on the screen a user could get confused rather than informed.

2.6 Problem 2

We are interested in finding out how the nation reacted to search term. We are NOT interested in sentiment analysis. We are interested in sheer number of tweets on a topic that is associated with search term. We will have to choose a good topic. Understand the Search API that we are using for can give you only limited number of tweets per day and also only a sampling of the all the tweets. You will collect at least 20000 tweets Group them by geo-location as in Google maps API (one more API) and plot them on the map of USA. If you plot the location every tweet then there will be too many points on the map. You can plot all the tweets at a given location (say a city or state) by a single blob, the size of the blob representing the density of tweets. You may need some R programming here. Input: Search word or hash tag related to super bowl. Data client processing: Obtain and group tweets by location. Output: plot the groups by size on a map of USA for visual understanding of the response to an event. Issue 1: Of course, there is an issue with location meta-data. This is not available (N/A) if the user does hide his/her location. This is often the case nowadays with most of us. Many celebrities are especially conscious about this. They don't want people knowing their locations for obvious reasons. Then how can we get "set of locations "Here is a verified approach using function of twitteR (1) Convert search result tweets into dataframe (2) Lookup screen name from this dataframe (3) From Screen names get user info and convert into dataframe (4) Keep only users with location info (5) Get the geo code of the locations from this dataframe

2.7 Problem 3

We summarize trending topics about a location. When we are visiting places for example, as an interview or other official visits, we may want to analysis about topics trending in that place. Instead of reading newspapers and online news, we want just a quick summary. We want to put use your twitter "data client" application development experience. You use the twitteR libraries "trends" function to retrieve 10 top things trending about the place and summarize it appropriately as a complete message.