**Homework #7 (Due Oct 17 11:59 PM)**

IST 3420 - Fall 2017, Chen

**Name**: \_\_\_\_\_\_\_Adam Forestier\_\_\_\_\_\_\_\_\_\_\_\_\_

**Weblog Analytics (30 points)**

**Instruction:** Finish the following tasks. Then upload this document with your answers and your R Markdown file to “Homework 7” on Canvas.

1. Download the cleansed weblog analytics data file “weblog\_clean.jsonl”.
2. Read the data file into R. (2 points)

Note: The data file is in the format of JSON Lines. If you are not familiar with the reading of JSON Lines, you can refer to below discussion:

<https://mst.instructure.com/courses/11517/discussion_topics/9881>

1. Explain the meaning of all columns in the following table. (2 points)

|  |  |
| --- | --- |
| Column | Description |
| remote.host | Gives the IP Address of host connecting to server |
| request.datetime | Both date and time the request was received |
| weekday | Day of the week the request was received |
| request.method | HTTP Method used for request |
| request.uri | URI stands for uniform resource identifier. It is a string of characters used to identify a resource. This gives the resource on the server that was requested access to |
| status | Checks which HTTP status of the host |
| response.size | Size of the Request that was returned |
| Ip\_number | IP Address converted to a number |
| country\_code | Country code or abbreviation from which the request came |
| country\_name | Country full name from which the request cames |

1. Show the structure of the dataset. Change the data type of “request.datetime” as POSIXct by using the *as.POSIXct()* function. (2 points)
2. Show summary statistics of the dataset. Paste results in the box below. (2 points)

|  |
| --- |
|  |

1. Calculate frequency distribution of weekday.

Which day has the largest number of visits? \_\_\_\_ Wednesday\_\_\_\_\_ (2 points).

Show the frequency of weekday in column format in the following box (2 points).

|  |
| --- |
|  |

1. Crosstabulate weekday and request.uri.

How many visits of the “faq.html” page on Fridays? \_\_\_\_\_\_4407\_\_\_\_\_\_\_\_\_ (2 points).

1. Draw a bar plot to show the distribution of visits by weekday. Paste the bar plot in the following box. (2 points)

|  |
| --- |
|  |

1. Draw a stacked bar plot to show the distribution of visits by request.uri and request.method (use request.method to group visits). Paste the stacked bar plot in the following box. (2 points)

|  |
| --- |
| Notes: robots.txt only makes up 723 which is such a small value that is not visible; post makes up so few values it is hard to see as well |

1. Draw a pie chart to present the distribution of visits by request.uri. Show the count of visits for each URI. Paste the pie chart in the following box. (2 points)

|  |
| --- |
|  |

1. Draw a pie chart to present the distribution of visits by request.uri. Show the percentage of visits for each URI. Paste the pie chart in the following box. (2 points)

|  |
| --- |
|  |

1. Calculate a five-number summary for “response.size”. Paste result in the following box (2 points)

|  |
| --- |
|  |

What is the explanation of the five-number summary? Explain in the following box (2 points)

|  |
| --- |
| A 5 number summary is a set of statistics that provide descriptive information about a data in set. It is made up of the five percentiles including the following: the minimum (smallest observation) the lower quartile or first quartile. the median (middle value), the third quartile, and the maximum (biggest observation)  My picture above stands for the following:  Value 15 at 0% means that 15 is the minimum response size  Value 1590 at 25% means that 1590 is the first quartile response size  Value 11489 at 50% means that 11489 is the median response size  Value 23574 at 75% means that 23574 is the third quartile response size  Value 217164050 at 100% means that 217164050 is the maximum response size |

1. Paste your R Markdown code in the following box (2 points).

|  |
| --- |
| ---  title: "Homework 7"  author: "Adam Forestier"  date: "October 17, 2017"  output: html\_document  ---  ```{r setup, include=FALSE}  knitr::opts\_chunk$set(echo = TRUE, message = F)  ```  Clear the Environment  ```{r}  rm(list = ls())  ```  Load in packages  ```{r}  require(jsonlite)  require(dplyr)  ```  set directory  ```{r}  setwd("D:/College Work/Junior Year/IST 3420/Homework - Code/Homework 7")  ```  # Read in JSON File Data  ```{r}  # Read in data  json\_df <- stream\_in(file("weblog\_clean.jsonl"))  ```  # Change request.datetime as POSIXct  ```{r}  json\_df <- json\_df %>% mutate(request.datetime = as.POSIXct(request.datetime))  ```  # Show summary statistics  ```{r}  summary(json\_df)  ```  # Calculate Frequency distribution of weekday  ```{r}  # Calculate frequency distribution of weekday and put it in column format ordered greatest to least  day\_freq <- cbind(table(json\_df$weekday))  day\_freq <- cbind(day\_freq[order(-day\_freq),])  print(day\_freq)  ```  # Crosstabulate weekday and request.uri.  ```{r}  cross\_tbl <- xtabs(~ request.uri + weekday, data=json\_df)  print(cross\_tbl)  ```  # Bar plot to show the distribution of visits by weekday  ```{r}  # Create table  weekday\_distribution <- table(json\_df$weekday)  # Bar graph  barplot(weekday\_distribution,  las = 2,  main="Distribution of Visits by Weekday",  col=rainbow(7))  ```  # Stacked bar plot to show the distribution of visits by request.uri and request.method  ```{r}  # Create Table  request\_info <- table(json\_df$request.uri, json\_df$request.method)  # Stacked Bar  barplot(request\_info,  main="Distribution of Visits by request.uri and request method",  xlab = "Request Method",  col = rainbow(6),  ylim = c(0, 140000),  legend = rownames(request\_info))  ```  # Pie chart to present the distribution of visits by request.uri by count  ```{r}  # Create Table  uri\_request <- table(json\_df$request.uri)  # Create Label to include on pie chart  pie1\_lbl <- paste(names(uri\_request), " \n", uri\_request, sep="")  pie(uri\_request,  labels = pie1\_lbl,  main = "Pie Chart with URI Visitor Count",  col = rainbow(length(pie1\_lbl)))  ```  # Pie chart to present the distribution of visits by request.uri by perecentage  ```{r}  # Create percent and label  pct <- round(uri\_request/sum(uri\_request)\*100, digits = 1)  pie2\_lbl <- paste(names(uri\_request), " \n", paste(pct,"%"), sep="")  pie(uri\_request,  labels = pie2\_lbl,  main = "Pie Chart with URI Percentage of Visitors",  col = rainbow(length(uri\_request)))  ```  # Five-number summary for “response.size”.  ```{r}  # Create df with the response size column going from smallest to largest  ordered\_size\_df <- json\_df[order(json\_df$response.size),]  quantile(ordered\_size\_df$response.size, type = 7)  ```s |

Upload this document with your answers to “Homework 7” on Canvas. (1 point)

Upload your R Markdown script file to “Homework 7” on Canvas. (1 point)