# Xi’an JIAOTONG-LIVERPOOL UNIVERSITY

**西 交 利 物 浦 大 学**

# Year 4

Course Work Submission

|  |  |  |
| --- | --- | --- |
| Name | Sabour | Sahand |
| ID Number | *1614650* | |
| Programme | *Computer Science and Technology* | |
| Module Title | *Big Data Analytics* | |
| Module Code | *CSE313* | |
| Assignment Title | *Lab tutorial 1* | |
| Submission Deadline | *11th October, 5:00PM* | |
| Lecturer Responsible | *Dr. Gangmin Li* | |

I certify that:

* I have read and understood the University’s definitions of COLLUSION and PLAGIARISM (available in the Student Handbook of Xi’an Jiaotong-Liverpool University).

With reference to these definitions, I certify that:

* I have not colluded with any other student in the preparation and production of this work;
* this document has been written solely by me and in my own words except where I have clearly indicated and acknowledged that I have quoted or used figures from published or unpublished sources (including the web);
* where appropriate, I have provided an honest statement of the contributions made to my work by other people including technical and other support staff.

I understand that unauthorised collusion and the incorporation of material from other works without acknowledgement (plagiarism) are serious disciplinary offences.

Signature ……*Sahand Sabour*…… Date ……*……2019/10/11…*……………

|  |  |  |  |
| --- | --- | --- | --- |
| For Academic Office use: | Date Received | Days Late | Penalty |
|  |  |  |

**Introduction**

Big Data Analytics (BDA) has become one of the most essential parts of any successful organization. It can be defined as a field of study on how to acquire valuable information from major storages of data with an aim; particularly business wise. For instance, achieving business objectives, cost reduction, and development of new services and products.

The process of BDA can be divided into the following main parts:

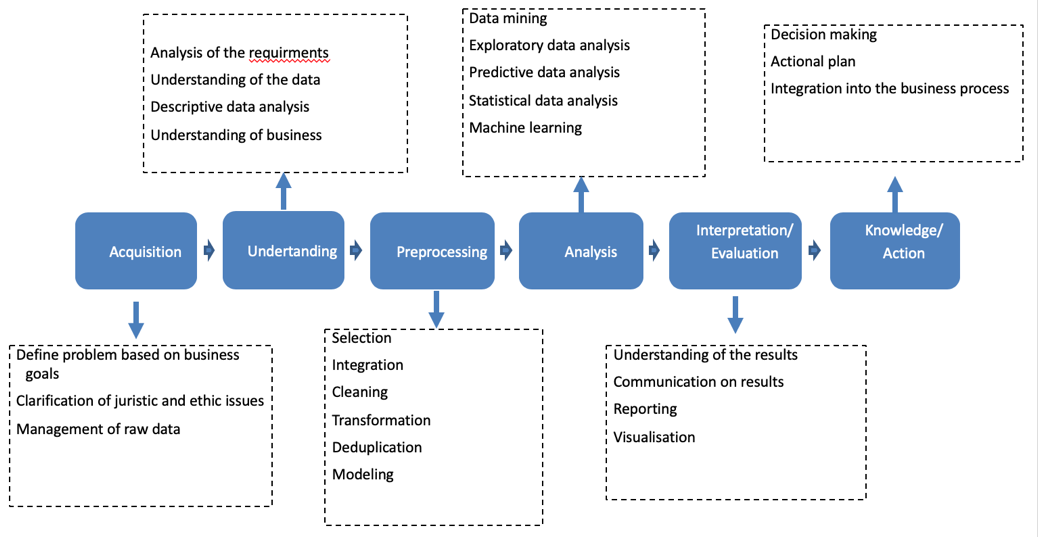


Figure 1: BDA analysis process [1]

In this lab tutorial, the programming language R was used to process, analyse, and/or plot sets of given data. With this practice, the above steps regarding BDA (Figure 1) were integrated, studied, and utilized. This report will demonstrate the findings regarding the analysis of the required dataset over specific periods of time. In addition, the related source codes as well as the corresponding graphs are provided respectively.

**Findings**

By thorough analysis of a week’s dataset, it was noticed that the youngest group of users (under 18), such as children and teenagers, as well as the oldest age groups of users, more than 55 and more specifically 65 and over, are much more likely to click a displayed advertisement compared to users from the age groups in between. This pattern is believed to be highly beneficial as it highlights the demographic that is more likely to show interest in a displayed advertisement and therefore, most of the advertisements can be utilized to target these age groups in order to achieve probable business objectives. Moreover, it was noticed that male users have slightly higher average CTR in the youngest age group while female users have a slightly higher average CTR in the oldest age group. Hence, in addition to the target age group, the target gender can also be chosen in order to ensure better achievement of the mentioned objectives.

**Source codes**

The source code for BDA process of a week’s dataset with comments, which sufficiently explain each step of the process, is provided below:

|  |  |  |
| --- | --- | --- |
| |  |  | | --- | --- | | 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52 53  54  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99  100  101102103104105106107108109110111 | # The source code for this project is divided into the 5 main steps of the BDA process; excluding the ‘knowledge/action’ step  library(doBy)  library(ggplot2)  # Data Acquisition  # Read the data for 7 days  day1<-read.csv("Datasets/nyt1.csv")  day2<-read.csv("Datasets/nyt2.csv")  day3<-read.csv("Datasets/nyt3.csv")  day4<-read.csv("Datasets/nyt4.csv")  day5<-read.csv("Datasets/nyt5.csv")  day6<-read.csv("Datasets/nyt6.csv")  day7<-read.csv("Datasets/nyt7.csv")  # Data Understanding  # The dataset was analysed and each column’s meaning was realized  # Data PreProcessing  # Adding a column to distinguish each day's data from others  day1$whichDay <- "Day1"  day2$whichDay <- "Day2"  day3$whichDay <- "Day3"  day4$whichDay <- "Day4"  day5$whichDay <- "Day5"  day6$whichDay <- "Day6"  day7$whichDay <- "Day7"  # Storing all the seven days' data into a single variable  allDays <- rbind(day1, day2, day3, day4, day5, day6, day7)  # Setting a level for each day  levels(allDays$whichDay) <- list(  'Day1' = 'Day1',  'Day2' = 'Day2',  'Day3' = 'Day3',  'Day4' = 'Day4',  'Day5' = 'Day5',  'Day6' = 'Day6',  'Day7' = 'Day7'  )  # Setting the gender values to Male or Female (from 0 or 1)  allDays$Gender <- ifelse(allDays$Gender == 0, "M", "F")  # Assigning users to seven age categories (Excluding invalid)  allDays$ageCategory <- cut(  allDays$Age,  c(-Inf,0,17,24,34,44,54,64,Inf),  labels = c("Invalid age","Under 18","18-24","25-34","35-44","45-54", "55-64", "65 and over")  )  # Calculating the Click-through-rate (CTR) for each user across all days  allDays$CTR <- ifelse(allDays$Impressions != 0, allDays$Clicks / allDays$Impressions, 0)  # Removing errors (Age is less than 3) and useless information (users who have not seen any ads; 0 impressions) from the data  # It is assumed that average children would have an understanding of simple software and internet tools by the age of 4  allDays <- subset(allDays, Age>3 & Impressions!=0)  # Defining the set of functions for obtaining required metrics  siterange <- function(x){c(length(x), min(x), max(x), mean(x))}  # Data Analysis  # Creating tables for number impressions based on gender, age groups and days respectively  impsTable\_Age <-summaryBy(Impressions~ageCategory,data=allDays, FUN=siterange)  impsTable\_Gender <-summaryBy(Impressions~Gender,data=allDays, FUN=siterange)  impsTable\_Day <-summaryBy(Impressions~whichDay,data=allDays, FUN=siterange)  # Drawing the plot for number impressions based on gender, age groups and days  png(file = "Impressions.png")  ggplot(data = allDays, aes(x=Gender, y=Impressions, fill=whichDay, color= whichDay))+geom\_bar(stat = "summary",fun.y="mean")+ labs(title = "Comparison of number impressions during a week between differnet genders and age groups")+ facet\_grid(.~ ageCategory)  dev.off()  # Creating a table for number of clicks based on gender, age groups and days respectively  clicksTable\_Age <-summaryBy(Clicks~ageCategory,data=allDays, FUN=siterange)  clicksTable\_Gender <-summaryBy(Clicks~Gender,data=allDays, FUN=siterange)  clicksTable\_Day <-summaryBy(Clicks~whichDay,data=allDays, FUN=siterange)  # Drawing the plot for number of clicks based on gender, age groups and days  png(file = "Clicks.png")  ggplot(data = allDays, aes(x=Gender, y=Clicks, fill=whichDay, color= whichDay))+geom\_bar(stat = "summary",fun.y="mean")+ labs(title = "Comparison of clicks during a week between differnet genders and age groups")+ facet\_grid(.~ ageCategory)  dev.off()  # Creating a table for CTR based on gender, age groups and days respectively  ctrTable\_Age <-summaryBy(CTR~ageCategory,data=allDays, FUN=siterange)  ctrTable\_Gender <-summaryBy(CTR~Gender,data=allDays, FUN=siterange)  ctrTable\_Day <-summaryBy(CTR~whichDay,data=allDays, FUN=siterange)  # Drawing the plot for CTR based on gender, age groups and days  png(file = "CTR.png")  ggplot(data = allDays, aes(x=Gender, y=CTR, fill=whichDay, color= whichDay))+geom\_bar(stat = "summary",fun.y="mean")+ labs(title = "Comparison of CTR during a week between differnet genders and age groups")+ facet\_grid(.~ ageCategory)  dev.off()  # Data Interpretation  # This step of the BDA process is included in the following section | |

**Tables and Graphs**

By compiling and running the above source code, the following tables and graphs would be obtained. Below, thorough explanation for each table and graph is provided respectively. It should be noted that the analysis is for a week’s dataset.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Gender | Length | Min | Max | Mean |
| F | 1220906 | 1 | 20 | 5.034731 |
| M | 1128656 | 1 | 18 | 5.031322 |

Table 1: Distribution of number impressions based on gender

The above table (Table 1) shows that in a period of seven days, on average, approximately the same number of advertisements (5 ads) are displayed to both male and female users. This is highly beneficial as it allows us to compare the Click-Through-Rate (CTR) distribution with less error. With the same number of impressions, CTR would be directly correlated to the number of clicks, which would more clearly distinguish the distribution of interest in advertisements between users.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Day number | Length | Min | Max | Mean |
| Day1 | 319198 | 1 | 20 | 5.044129 |
| Day2 | 313231 | 1 | 18 | 5.032238 |
| Day3 | 306188 | 1 | 19 | 5.031105 |
| Day4 | 307515 | 1 | 17 | 5.034584 |
| Day5 | 257207 | 1 | 18 | 5.034630 |
| Day6 | 531584 | 1 | 18 | 5.028078 |
| Day7 | 314639 | 1 | 19 | 5.030444 |

Table 2: Distribution of number impressions based on day

Both the above and below table (Tables 2&3) also clearly display that on average, the distribution of impressions between different days of the dataset as well as different age groups is approximately the same. Hence, the result of the CTR analysis would be valuable as the impressions are equally distributed (displayed) to different range of users; different genders, age groups, and on different days.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Age group | Length | Min | Max | Mean |
| Under 18 | 101733 | 1 | 16 | 5.039849 |
| 18-24 | 297061 | 1 | 18 | 5.028119 |
| 25-34 | 424296 | 1 | 18 | 5.031747 |
| 35-44 | 515698 | 1 | 20 | 5.032345 |
| 45-54 | 470467 | 1 | 18 | 5.030797 |
| 55-64 | 328978 | 1 | 19 | 5.039036 |
| 65 and over | 211329 | 1 | 18 | 5.037226 |

Table 3: Distribution of number impressions based on age group

In order to illustrate the approximately equal distribution of the displayed advertisements (impressions), the following graph was plotted (Figure 2). As it is clearly displayed in the figure, the impressions are equally distributed amongst users of different genders as well as users of different age groups. In addition, the distribution of impressions between days of the week is also displayed to be approximately the same.

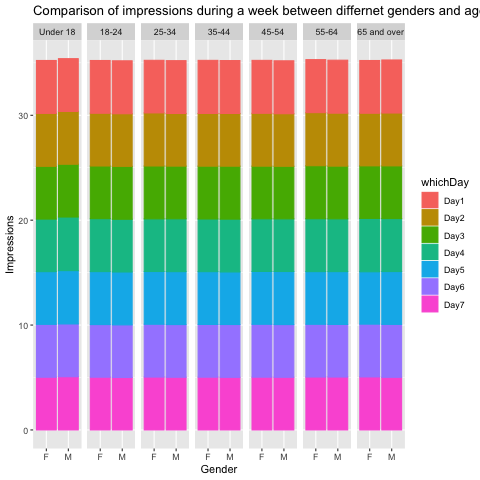


Figure 2: Comparison of number impressions for a week based on gender and age

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Gender | Length | Min | Max | Mean |
| F | 1220906 | 0 | 4 | 0.07019541 |
| M | 1128656 | 0 | 4 | 0.07331729 |

Table 4: Distribution of clicks based on gender

Both the above and the below tables (Tables 4&5) display similar results to those of impressions based on gender and day of the week; the distribution of clicks amongst different genders and days of the dataset is approximately the same (7 clicks per 100 displayed advertisements on average).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Day number | Length | Min | Max | Mean |
| Day1 | 319198 | 0 | 3 | 0.07195847 |
| Day2 | 313231 | 0 | 3 | 0.07222784 |
| Day3 | 306188 | 0 | 4 | 0.07169451 |
| Day4 | 307515 | 0 | 3 | 0.07195747 |
| Day5 | 257207 | 0 | 4 | 0.07122668 |
| Day6 | 531584 | 0 | 3 | 0.07137160 |
| Day7 | 314639 | 0 | 4 | 0.07157091 |

Table 5: Distribution of clicks based on day

The below table (Table 6) displays that the distribution of clicks amongst different age groups, dissimilar to the previous two comparisons, is not equally distributed. Based on the data from Table 6, it can be realized that the youngest age group (Under 18) and the oldest age group (65 and over) possess the largest average of clicks, with the second oldest group also possessing a large average of clicks compared to the other age groups.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Age group | Length | Min | Max | Mean |
| Under 18 | 101733 | 0 | 4 | 0.15206472 |
| 18-24 | 297061 | 0 | 3 | 0.05539266 |
| 25-34 | 424296 | 0 | 3 | 0.05009946 |
| 35-44 | 515698 | 0 | 3 | 0.05106089 |
| 45-54 | 470467 | 0 | 3 | 0.05049451 |
| 55-64 | 328978 | 0 | 4 | 0.10109187 |
| 65 and over | 211329 | 0 | 4 | 0.15106777 |

Table 6: Distribution of clicks based on age group

The obtained results from the above tables is clearly displayed in the below graph (Figure 3). As it is illustrated in Figure 3, the youngest as well as the oldest users are more likely to click a displayed advertisement as the average number of clicks from the users belonging to these age groups are considerably larger than the others.

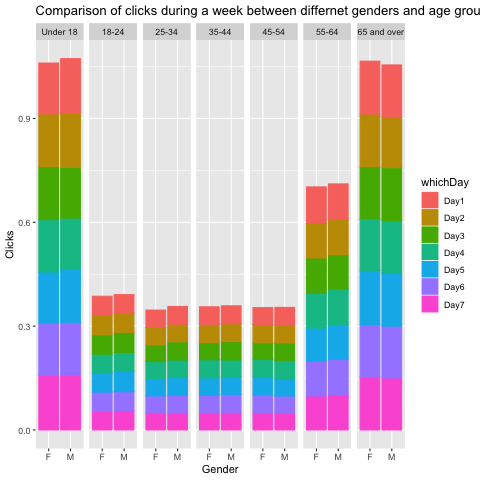


Figure 3: Comparison of clicks for a week based on gender and age

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Gender | Length | Min | Max | Mean |
| F | 1220906 | 0 | 1 | 0.01395261 |
| M | 1128656 | 0 | 1 | 0.01454922 |

Table 7: Distribution of CTR based on gender

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Day number | Length | Min | Max | Mean |
| Day1 | 319198 | 0 | 1 | 0.01425364 |
| Day2 | 313231 | 0 | 1 | 0.01430813 |
| Day3 | 306188 | 0 | 1 | 0.01430315 |
| Day4 | 307515 | 0 | 1 | 0.01420272 |
| Day5 | 257207 | 0 | 1 | 0.01407678 |
| Day6 | 531584 | 0 | 1 | 0.01426051 |
| Day7 | 314639 | 0 | 1 | 0.01422615 |

Table 8: Distribution of CTR based on day

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Age group | Length | Min | Max | Mean |
| Under 18 | 101733 | 0 | 1 | 0.03014247 |
| 18-24 | 297061 | 0 | 1 | 0.01097226 |
| 25-34 | 424296 | 0 | 1 | 0.01001837 |
| 35-44 | 515698 | 0 | 1 | 0.01018204 |
| 45-54 | 470467 | 0 | 1 | 0.01000797 |
| 55-64 | 328978 | 0 | 1 | 0.02010337 |
| 65 and over | 211329 | 0 | 1 | 0.02984153 |

Table 9: Distribution of CTR based on age group

Evidently, the same results as the distribution of clicks is obtained and displayed in the above tables; same distribution of average CTR between different gender groups and days of dataset as well as a clear pattern in the distribution of average CTR between different age groups. Similar results are also illustrated in the following graph (Figure 4), where the youngest and the oldest age groups (Under 18 and 65 and over) have significantly larger average CTR compared to the other age groups.



Figure 4: Comparison CTR for a week based on gender and age

**Conclusion**

In this lab tutorial, the programming language R was used to display, analyse and explore a given set of datasets. The given dataset files were acquired; Acquisition, and primarily analysed and understood; Understanding. Accordingly, certain measures were taken to modify the data in order to classify the data into different categories, create new valuable attributes and remove redundant and useless records; Pre-processing. In addition, the ordered and classified data was analysed to find patterns within the given datasets over the span of a single day as well as span of ten consecutive days; Analysis. The understanding of the given analysis was then visualized using various tables and plus; Interpretation. Hence, it is believed that all the steps of BDA procedure were implemented successfully. In addition, this lab tutorial was a valuable experience as well as an intriguing learning method.

**References**

[1] *Lecture 4 – Process and Tasks of BDA*, XJTLU, Department of Computer Science and Software Engineering, Suzhou, p.10.

[2] *Practice BDA Process with R (With Lab Tasks)*, XJTLU, Department of Computer Science and Software Engineering, Suzhou.