XI'AN JIAOTONG-LIVERPOOL UNIVERSITY

西 交 利 物 浦 大 学

YEAR 4

COURSE WORK SUBMISSION

Name	Osmond	Wilbert
ID Number	*1926308*	
Programme	*Exchange (non-UoL)*	
Module Title	*Big Data Analytics*	
Module Code	*CSE313*	
Assignment Title	*Lab 1*	
Submission Deadline	*(Mitigating circumstances)*	
Lecturer Responsible	*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own words except where I have clearly indicated and acknowledged that I have quoted or used figures from published or unpublished sources (including the web);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  Date …3rd January 2020……

For Academic Office use:	Date Received	Days Late	Penalty

1 What I have found

I have found that adults are significantly least interested in the advertisements shown than either youngsters and middle-aged adults, along the first 10 days of the month. On the other hand, youngers are as interested in the advertisements shown as middle-aged adults.

The data was divided in 3 subgroups: youngsters (age of below 30), adults (age of between 30 and 45), and middle-aged adults (age of between 45 and 60). The threshold lines were set according to the age concepts and small adjustments to ensure the three subgroups have similar population sizes.

As the measurement, I have created a new property called click-through-rate (CTR), which is the rate of clicks over impressions, which I consider also as the response rate to viewing the advertisements. I have decided to consider positive CTR (i.e. CTR > 0) as a sign that a user is interested in the advertisements proposed.

The results clearly show how the different age groups are attracted differently by the advertisements, despite the similar population sizes. The results are consistent throughout the span of 10 days.

2 Source code with comments on my BDA process

```
# libraries used
install.packages("doBy")
library("doBy")
library(ggplot2)

# This is my source code to demonstrate the BDA process
# The data will be processed according to these steps:
#   1. Data Acquisition: the data will be loaded from the dataset provided
#   2. Data Understanding: the data will be looked at given the business
environment of the advertisement shown/clicked
#   3. Data Pre-processing: data will be cleaned and categorized and impossible
records will be removed from the analysis
#   4. Data Analysis: the processed data will be represented in graphs and tables
#   5. Data Interpretation: the results of the data analysis and summary will
be analyzed looking for one pattern

### 1. Data acquisition ---------------------------------------------------
## Load all available datasets
setwd("C:/Users/wilbert osmond/Documents/XJTLU/CSE313 Big Data
Analytics/Assignment 1/Lab 1") # Set to the same directory
filenames <- list.files (path = "C:/Users/wilbert osmond/Documents/XJTLU/CSE313
Big Data Analytics/Assignment 1/Lab 1", pattern = "*.csv") # list all csv files
main_data <- NULL
```

```
## Add the first 10 files to the main dataset
for (day in 1:10) {
  cat("Loading file nyt", day, ".csv\n", sep="")
  dat = read.csv(filenames[day], header=TRUE)

  dat$Day <- day # add day column to plot the graphs

  if (is.null(main_data)) # first frame is initialized
    main_data <- dat
  else                  # all following frames are joined
    main_data <- rbind(main_data, dat)
}
cat("Data read in successfully\n\n")

### 2a. Data understanding
-----------------------------------------------------
# Check if there is any data with age below 0
is.null(subset(main_data, Age<0))
cat("There is data with age below 0. We will remove this later\n\n")

# Add a new property (click-through-rate), which represents the rate of how much
a user clicks on the advertisement after they view it
main_data$CTR <- main_data$Clicks / main_data$Impressions
cat("Added Click-Through-Rate attribute to the dataset\n\n")

### 3. Data pre-processing
-----------------------------------------------------
## Data cleaning
# Non-signed-in users have assigned default values that will affect the analysis
outcome. They will therefore be removed to maintain the analysis quality
main_data <- subset(main_data, Signed_In==1) # extract subset of data with
signed-in users as the main data
cat("Non signed-in users have been removed\n\n")

# Remove data with no impressions, because CTR with no impressions must be a mistake
since it is not possible for users to click on advertisements that don't show
(i.e. wrong data)
main_data <- subset(main_data, Impressions>0) # extract subset of main_data which
has impressions above 0
cat("No-impressions data have been removed\n\n")

# Remove data with ages below 0, as this is impossible and unrealistic
main_data <- subset(main_data, Age>0) # extract subset of main_data which has
age above 0
cat("Negative-age data have been removed\n\n")

# Data discretization: grouping age ranges to 0-30 (youngsters), 30-45 (adults),
```

```
45-60 (middle-aged adults)
main_data$agecat <- cut(main_data$Age,c(0, 30, 45, 60))
cat("Age has been discretized to three age categories\n\n")

# Create a separate subset, along with the same age category, for CTR > 0 later
on
dataCTR <- subset(main_data, CTR>0)

### 2b. Data understanding
-----------------------------------------------------
# Take a step back and look at the data again to understand it better

# Print the number of users in the three age groups
cat("Youngsters (0-30):", nrow(subset(main_data, Age>0 & Age<=30)), "\n")     #
860667
cat("Adults (30-45):  ", nrow(subset(main_data, Age>30 & Age<=45)), "\n")     #
923952
cat("Middle-aged adults (45-60): ", nrow(subset(main_data, Age>45 & Age<=60)),
"\n")      # 841378

cat("The populations of the three age categories are similar. This enables us
to more fairly compare their properties. \n\n")

### 4. Data analysis -----------------------------------------------------
# There are three age categories: youngsters, adults, and middle-aged adults.
The thresholds are arbitrarily set to ensure similar population sizes and related
to general age category concepts.
# Since the three population sizes are similar, we will compare the three to see
which of the age category has the highest response rate to advertisements.

# The best metric is the Click-Through-Rate: the rate of which a user clicks on
the advertisement when they see it (impressions)
# I have chosen to consider CTR>0 as being a sign that a user is interested in
the advertisements proposed.

# Plot over 10 days
ggplot(subset(main_data, CTR>0), aes(x=factor(Day), fill=factor(agecat))) +
  geom_histogram(stat="count", position="dodge") +
  ggtitle("Interested Users", subtitle = "Users of different age groups interested
in the advertisements shown") +
  scale_fill_manual("Gender\n", values = c("green", "hotpink","turquoise"),
labels = c("<30", "30-45", "45-65")) +
  xlab("Day") + ylab("Number of users")

# Create function metrics that summarize the data for CTR attribute, and use it
to compare against the three age categories
siterange <- function(x){c(length(x), quantile(x), mean(x), var(x))}
summaryBy(CTR~agecat, data=dataCTR, FUN=siterange)
```

```
# CTR quantiles are 0 because clicks quantiles are also 0 (too small)

### 5. Data Interpretation
-----------------------------------------------------
# The graph above show how, between ages 30 and 45, users are least interested
in the advertisements shown.
# On the other hand, youngsters' (0-30) and middle-aged adults' (45-60) interests
do not significantly differ.
cat("\nBetween ages 30 and 45, users are least interested in the advertisements
shown.On the other hand, youngsters' (0-30) and middle-aged adults' (45-60)
interests do not significantly differ.")
```

3 Plots and data coming from the code output

First, in Figure 1, we see an overview of the population sizes of the three age categories. It can be seen that the population sizes between age groups are similar to each other. Therefore, this enables us to fairly compare them with respect to a new attribute (i.e. the CTR) with a histogram.

Age category	Age range	Population size
Youngsters	0-30	860667
Adults	30-45	923952
Middle-aged adults	45-60	841378

Figure 1: Population sizes of the three age categories: youngsters (0-30), adults (30-45), middle-aged adults (45-60).

Using the three age categories, we plot a graph in Figure 2 to compare their frequency of who are interested in the advertisements shown. The x-axis represents the 10 days period in which the data is investigated, and for each there is a green bin for youngsters, a pink bin for adults, and a blue bin for middle-aged adults. The y-axis shows the height of the bars, which represents the population size of the group. Each group is defined as a subset of the whole dataset with each specific day and age category, after being filtered by the data pre-processing and CTR above 0. For example, the first bar on the left represents the number of youngster users below the age of 30 who have a CTR of more than 0, on day 1. The positive CTR is the threshold for data to be included in the plot, as it is defined as #Clicks / #Impressions, thus at least a click on any views of the advertisement shown considers the user to be interested in the advertisement. It is clear how the pink bins are constantly significantly lower than the green and blue bins, thus showing how there are fewer interested adults than either youngsters and middle-aged adults. On the other hand, it is also visible how the green bins tend to differ only insignificantly with the blue bins, hence showing how there are similar numbers of interested youngsters and middle-aged adults.

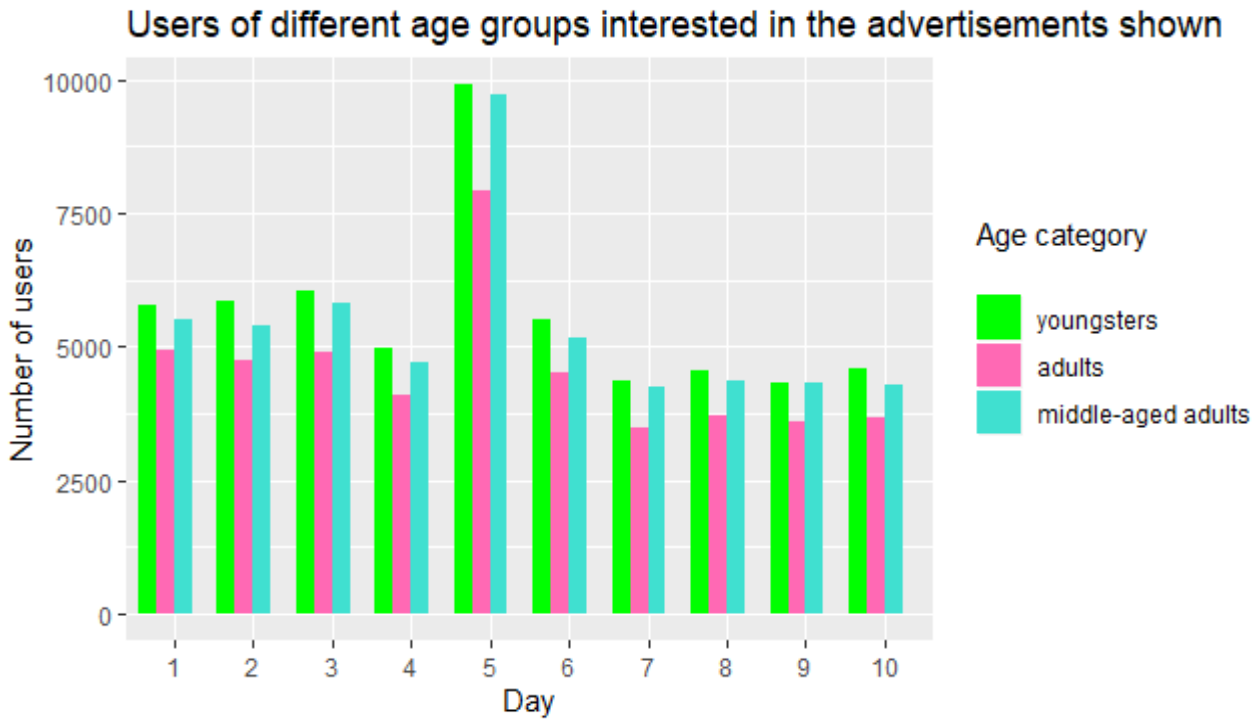

Figure 2: Users interested in the advertisements shown, categorized by the three age groups (youngsters, adults, and middle-aged adults), in the 10 days period.

Furthermore, I created a metric function that summarizes the data for the CTR attribute in the 10 days period and use it to compare against the three different age categories, as shown in Figure 3. The function contains metrics of the population size (i.e. the length), the quantiles (including minimum and maximum), mean, and variance. The results (particularly the length and mean) are consistent with the pattern shown in Figure 2. The adults (30,45] age category has the lowest length and mean, whereas the length and mean do not significantly differ between youngsters (0,30] and middle-aged adults (45,60]. A more interesting observation is that if the original population sizes of the three age groups (as shown in Figure 1) are not considered to be similar, so then adults would have the highest population size, and yet when filtered with CTR>0, the length and mean of the adults category is the lowest.

Age category	Population size (length)	Min (0^{th} quantile)	25^{th} quantile	Median (50^{th} quantile)	75^{th} quantile	Max (100^{th} quantile)	Mean	Variance
(0,30]	55913	0.05000000	0.1428571	0.1666667	0.25	1	0. 2071237	0.01289731
(30,45]	45521	0.05882353	0.1428571	0.1666667	0.25	1	0.2034026	0.01242704
(45,60]	53504	0.05000000	0.1428571	0.1666667	0.25	1	0.2067463	0.01307326

Figure 3: Table of metrics function comparing the three different age categories.

In conclusion, the results show that adults are least interested in the advertisements. An intuitive explanation for this may be that in the age of between 30 and 45, users are in their productive years and too busy working to consider purchasing things. Another explanation could be that since they are working, they may value money more as they are working hard to earn it, so they may be less interested to purchase things. Youngster may be more interested in the advertisements, as they may be more open to trying new things

and following the trend. Similarly, for middle-aged adults, their productive years may be declining, thus they have more time to settle down so they have more time to click on and further read through advertisements. They may also be more interested in the advertisements, if they feel like they have missed out on not enjoying their adult years as much since they were too busy working. These factors may explain why youngsters and middle-aged adults' interests in the advertisements are similar, and larger than the interest of adults.