Online Cryptography Course Analysis

Patrick Atak

Research Question

A Kenyan entrepreneur has created an online cryptography course and would want to advertise it on her blog. She currently targets audiences originating from various countries. In the past, she ran ads to advertise a related course on the same blog and collected data in the process. She would now like to employ your services as a Data Science Consultant to help her identify which individuals are most likely to click on her ads.

Metric of Success

In order to work on the above problem, I need to do the following:

1.Find and deal with outliers, anomalies, and missing data within the dataset.

2.Perform univariate and bivariate analysis recording your observations.

3.From your insights provide a conclusion and recommendation.

Reading in the CSV dataset

```
ads <- read.csv('advertising.csv')
head(ads)
```

==========

Checking for Outliers, Anomalies and Missing data within the dataset.

```
# Identifying missing data
is.na(ads)

# Finding the total missing values
colSums(is.na(ads))

# Dealing with the missing
na.omit(ads)

#There seems to be no missing values in this dataset
```

==========

Handling of Outliers

```
# Using a boxplot to visualise any existing outlier.
# Boxplot about the daily internet usage.
boxplot(ads$Daily.Internet.Usage)

# Function boxplot.stats which lists the outliers in the vectors.
boxplot.stats(ads$Daily.Internet.Usage)$out

# From the plot and the stats function, there seems to be no outlier present.
```

==========

Handling the duplicated data

```
# Using duplicated() function to check for duplicates across rows.
dupl_ads_rows = ads[duplicated(ads),]

# Using the unique() function to remove the duplicated rows.
unique_ads_rows = unique(ads)
```

Getting the statistical summary of the data

```
# Statistics of the data
summary(ads)

# Structure of the data
str(ads)
```

==========

Performing Univariate EDA & Graphical EDA

```
# Performing an analysis of a single variable. (Area Income).
# calculating the mean.
x <- unique_ads_rows$Area.Income
avg <- mean(x)

# calculating the median.
mid <- median(x)

# calculating the mode.
getmode <- function(v){
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v,uniqv)))]
}
```

```
most <- getmode(x)


# calculating the maximum.
high <- max(x)

# calculating the minimum.
low <- min(x)

# calculating the range.
rng <- range(x)

# calculating the quantile.
qtile <- quantile(x)

# calculating the variance.
vari <- var(x)

# calculating the standard deviation.
stdd <- sd(x)
```

```
# Performing the Univariate Graphical Plots
# Box Plot
boxplot(x)

# getting the frequency table
area_income <- unique_ads_rows$Area.Income
income_frequecy <- table(area_income)

# barplot of the area income
barplot(income_frequency)

# histogram of the area income
hist(income_frequency)
```

Performing Bivariate EDA & Graphical EDA

```
# using two variables. Area Income & Daily Internet Usage
daily_internet_use <- unique_ads_rows$Daily.Internet.Usage
#
# performing a covariance
cov(area_income, daily_internet_use)

# correlation coefficient
cor(area_income, daily_internet_use)

# creating a scatter plot
plot(area_income, daily_internet_use, xlab="Area Income", ylab="Daily Internet Use")
```

==========

Including Plots

```
# creating a scatter plot of Area Income vs Daily Internet Usage
plot(area_income, daily_internet_use, xlab="Area Income", ylab="Daily Internet Use")

# creating a scatter plot of Area Income vs Daily Time Spent on The Site
plot(ads$Area.Income, ads$Daily.Time.Spent.on.Site, xlab="Area Income", ylab="Daily Time Spent on Site"

# install the GGally package
# load library(GGally)
# visualise the correlation matrix
ggcorr(ads, method = c("everything", "pearson"))
```

==========