ADVERT CONSUMER ANALYSIS & MODELING

Objective:

To create a model that accurately predicts the likelihood of a blog reader clicking on an advert.

Success Metric:

Establish the specific groups of individuals more like to view web adverts on cryptography given the data provided. Create a model with at least 90% accuracy

The context:

The data was collected from a blog the ran ads to advertise a course. The data contains daily time Spent on site, age of the user, area income, daily internet usage of the user, ad Topic Line, the city, gender,country and the time stamp. The data source can be accessed from http://bit.ly/IPAdvertisingData

Experimental Design

R has been used for this analysis. The exploratory data analysis has given graphical presentations of univariate and bivariate analysis.

Reading the data set

```
advert <-read.csv('http://bit.ly/IPAdvertisingData')
advert_df <- data.frame(advert)
rmarkdown::paged_table(head(advert_df,n=4))
#Checking the the last 4 rows of the data set
rmarkdown::paged_table(tail(advert_df,n=4))
#Checking the number of rows and columns in the data set
dim(advert_df)
## [1] 1000   10
#The data set has 1000 rows and 10 columns
#Checking the data types of the columns
sapply(advert_df, class)
```

```
## Daily.Time.Spent.on.Site                      Age
Area.Income
##                "numeric"                "integer"
"numeric"
##     Daily.Internet.Usage            Ad.Topic.Line
City
##                "numeric"              "character"
"character"
##                     Male                  Country
Timestamp
##                "integer"              "character"
"character"
##            Clicked.on.Ad
##                "integer"
#Checking column names
colnames(advert_df)
##  [1] "Daily.Time.Spent.on.Site" "Age"
##  [3] "Area.Income"              "Daily.Internet.Usage"
##  [5] "Ad.Topic.Line"            "City"
##  [7] "Male"                     "Country"
##  [9] "Timestamp"                "Clicked.on.Ad"
```

Data Cleaning

Checking for duplicates

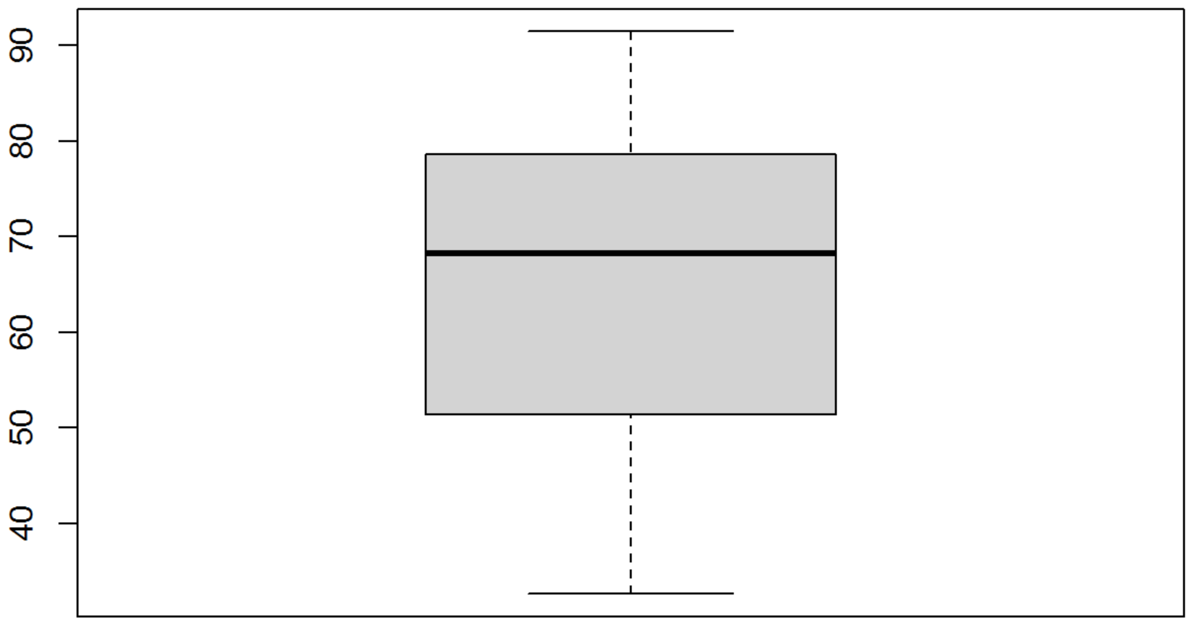
```
sum(duplicated(advert_df))
## [1] 0
#The data has no duplicated rows
```

Checking for missing values

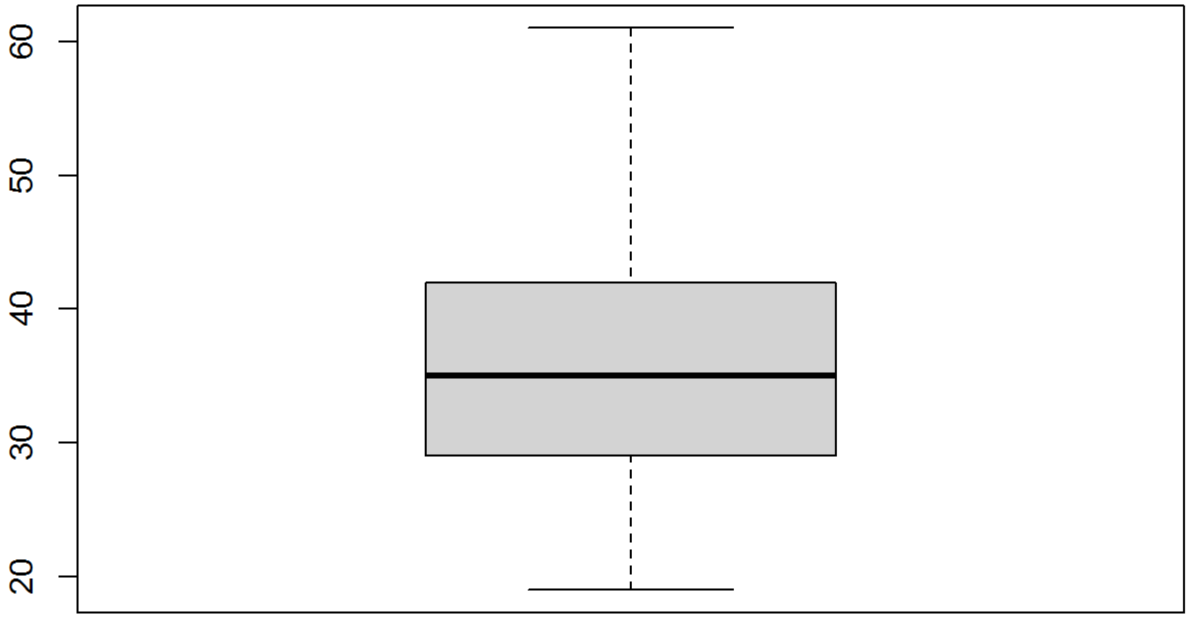
```
colSums(is.na(advert_df))
## Daily.Time.Spent.on.Site                      Age
Area.Income
##                        0                        0
0
##     Daily.Internet.Usage            Ad.Topic.Line
City
##                        0                        0
0
##                     Male                  Country
Timestamp
##                        0                        0
0
##            Clicked.on.Ad
##                        0
#There are no missing values in all columns
```

Checking for outliers

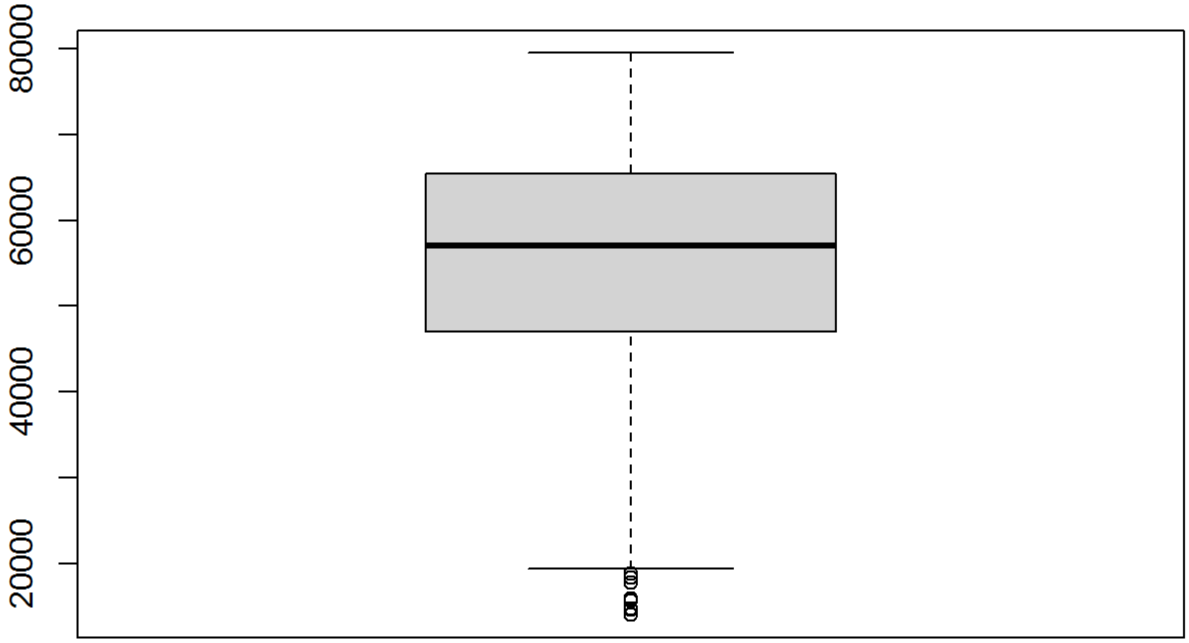
```
boxplot(advert_df$Daily.Time.Spent.on.Site)
```



```
#The are no outliers on the 'daily time spent on site' column
boxplot(advert_df$Age)
```



```
#The are no outliers on the 'age' column
boxplot(advert_df$Area.Income)
```



```
#The are outliers on the 'area income' column. However, the outliers
will not be dropped as they are legitimate  data give the context of
the column data.
```

Correcting column names

```
colnames(advert_df)
##  [1] "Daily.Time.Spent.on.Site" "Age"
##  [3] "Area.Income"              "Daily.Internet.Usage"
##  [5] "Ad.Topic.Line"            "City"
##  [7] "Male"                     "Country"
##  [9] "Timestamp"                "Clicked.on.Ad"
#Renaming the 'male' column to 'gender' where 1 represents male and 0
female
colnames(advert_df)[7] <- 'Gender'
#Removing fullstops in the column names
colnames(advert_df)[1] <- 'Daily Time Spent on Site'
colnames(advert_df)[3] <- 'Area Income'
colnames(advert_df)[4] <- 'Daily Internet Usage'
colnames(advert_df)[5] <- 'Ad Topic Line'
colnames(advert_df)[10] <- 'Clicked on Ad'
#Checking corrected column names
colnames(advert_df)
```

```
##  [1] "Daily Time Spent on Site" "Age"
##  [3] "Area Income"              "Daily Internet Usage"
##  [5] "Ad Topic Line"            "City"
##  [7] "Gender"                   "Country"
##  [9] "Timestamp"                "Clicked on Ad"
```

Correcting Column values

```
#Changing 0 and 1 values in the gender and clicked on Ad columns in
male and female values and yes and no values respectively.It will help
in visualizing the outcomes.
advert_df$Gender <- ifelse(advert_df$Gender == 1,'male','female')
advert_df$`Clicked on Ad` <- ifelse(advert_df$`Clicked on
Ad`==1,'yes','no')
#reading the data to check if the corrections made have reflected
rmarkdown::paged_table(head(advert_df,n=5))
```

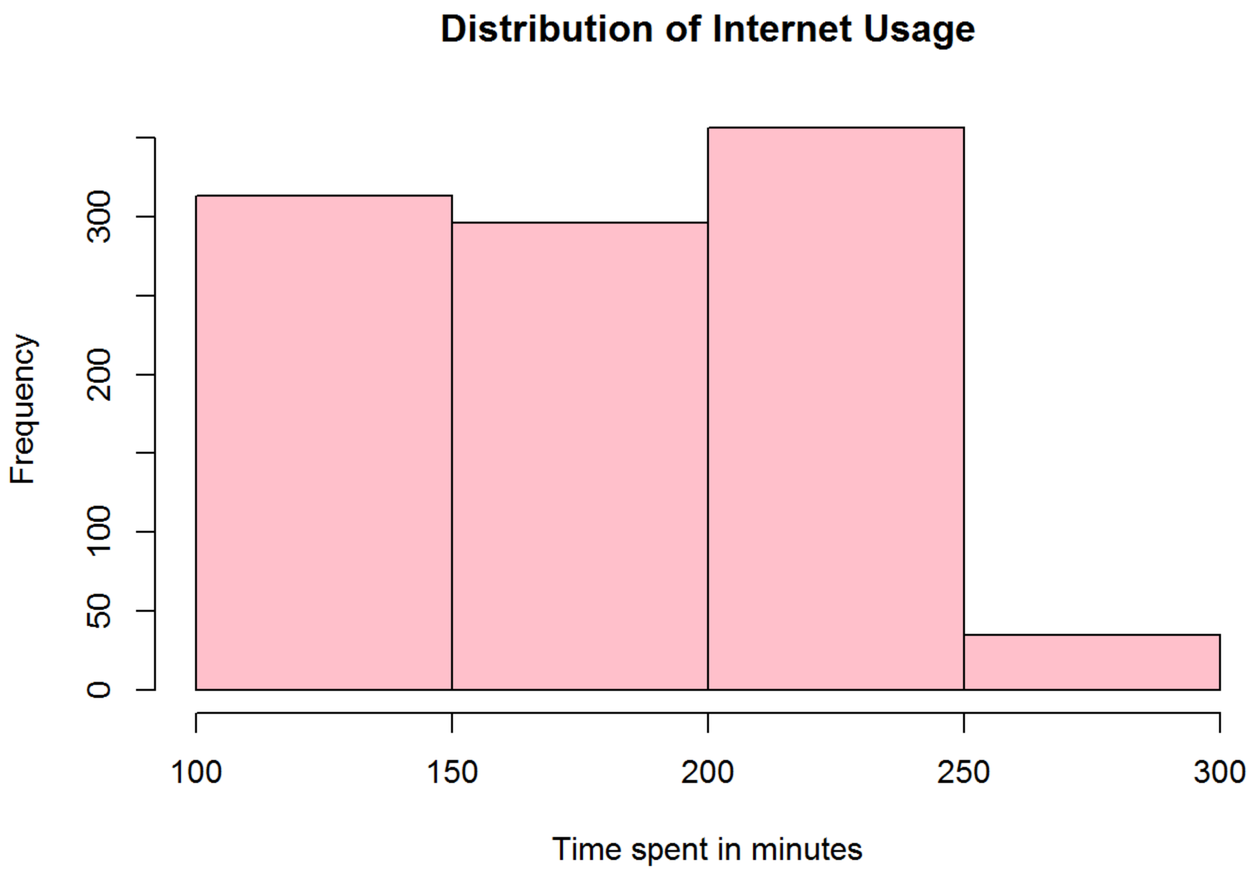
Exploratory Data Analysis

Univariate Analysis

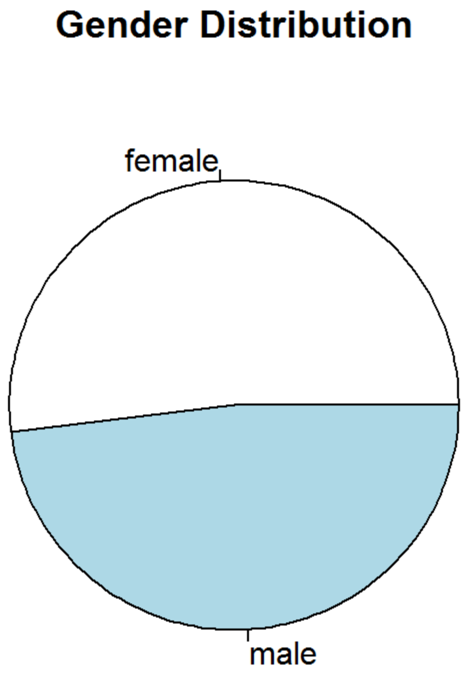
Installing libraries

```
library(tidyverse)
## -- Attaching packages ---------------------------------------
tidyverse 1.3.1 --
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
## -- Conflicts ------------------------------------------
tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
mean(advert_df$`Daily Time Spent on Site`)
## [1] 65.0002
#On average, the respondents spent 65.0002 minutes on the site
mean(advert_df$Age)
## [1] 36.009
#On average, the respondents are aged 36 years
mean(advert_df$`Daily Internet Usage`)
## [1] 180.0001
#On average there is an internet usage of 180.0001
median(advert_df$`Area Income`)
## [1] 57012.3
```

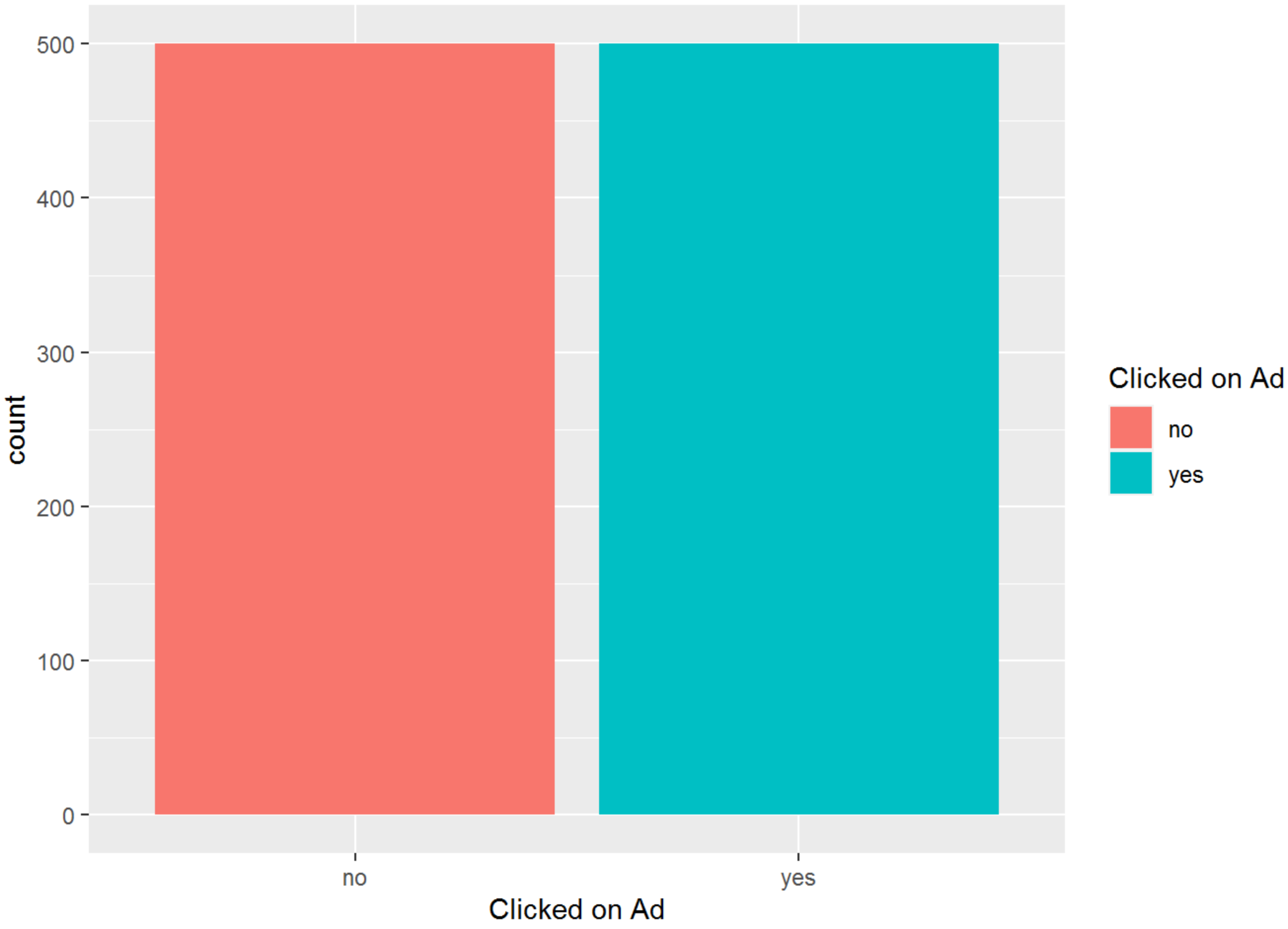
```
#The median area income is 57012.3
hist(advert_df$`Daily Internet Usage`, main = 'Distribution of
Internet Usage',
     xlab = 'Time spent in minutes',col = 'pink',breaks=5)
```



```
#Majority used between 200 and 250 megabytes of data
x <-table(advert_df$Gender)
pie(x, main = 'Gender Distribution',)
```



```
#More females than men visited the blog website.
#Distribution on countries that visited the website
c <-table(advert_df$Country)
s <-sort(c,descreasing=TRUE)[1:5]
as.matrix(s)
##                                                     [,1]
## Aruba                                                  1
## Bermuda                                                1
## British Indian Ocean Territory (Chagos Archipelago)    1
## Cape Verde                                             1
## Germany                                                1
ggplot(data =advert_df, aes(x = `Clicked on Ad`,fill= `Clicked on
Ad`)) +
    geom_bar()
```

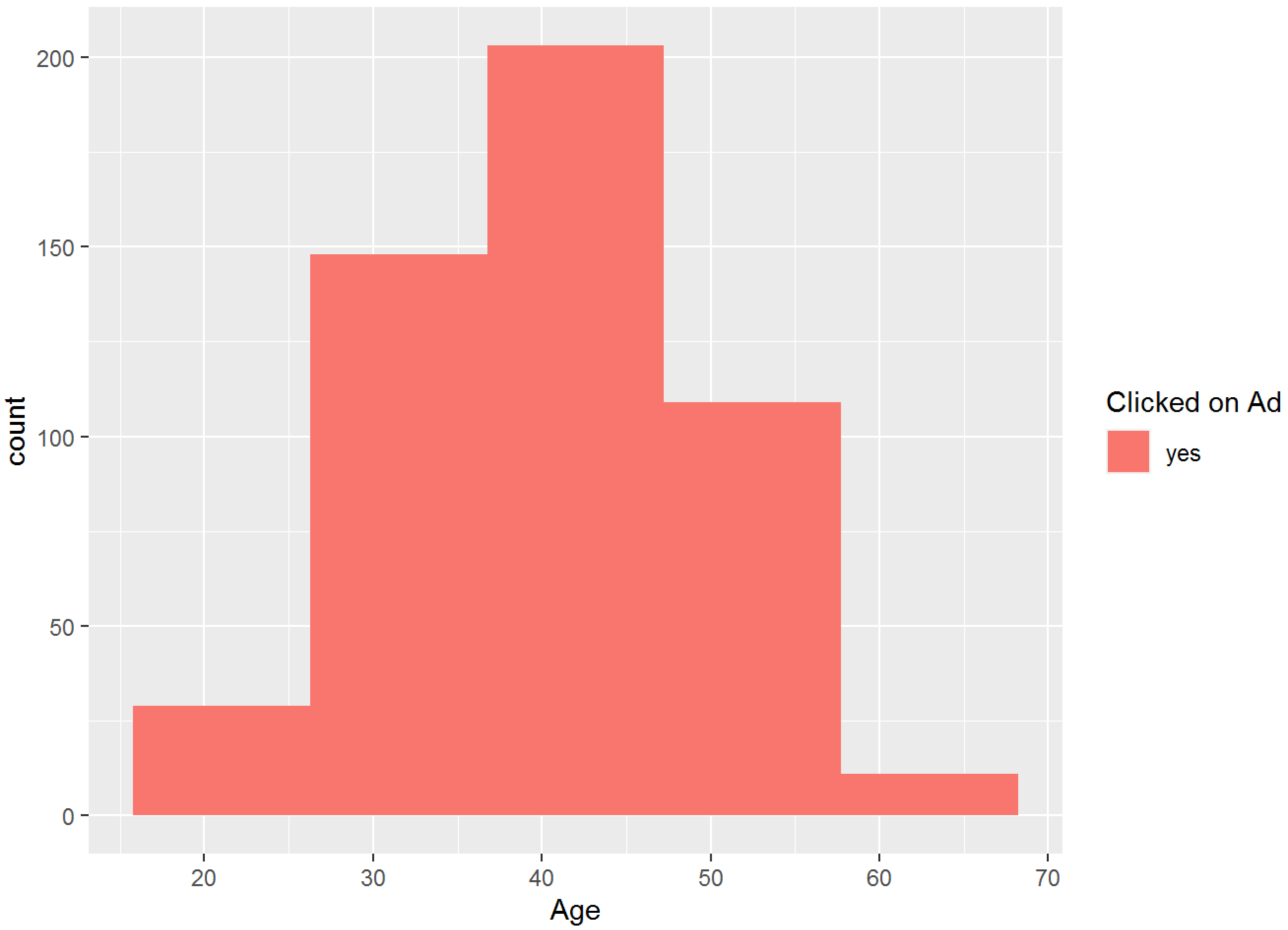


```
#The data set has an equal distribution of ad viewers with 50% having
viewed the ads and 50% who did no. Therefore the data set is balanced
```

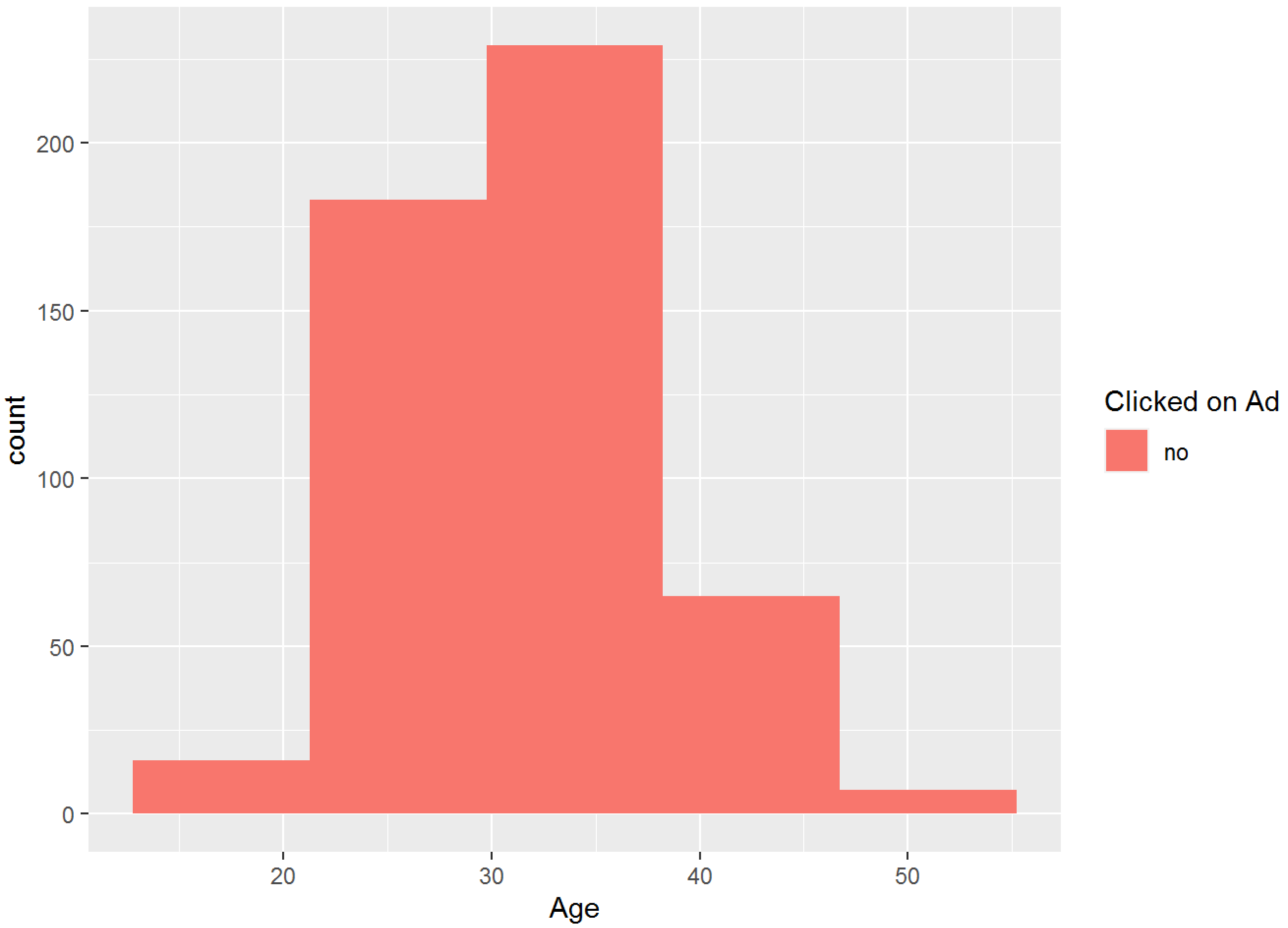
Bivariate Analysis

Clicks on Ad and Age

```
#A data frame with 'clicked on Ad ='Yes'
ad_df<-advert_df[advert_df$`Clicked on Ad`=='yes',]
ggplot(data=ad_df, aes(x=Age,fill=`Clicked on Ad`))+
  geom_histogram(position= 'stack',bins=5)
```



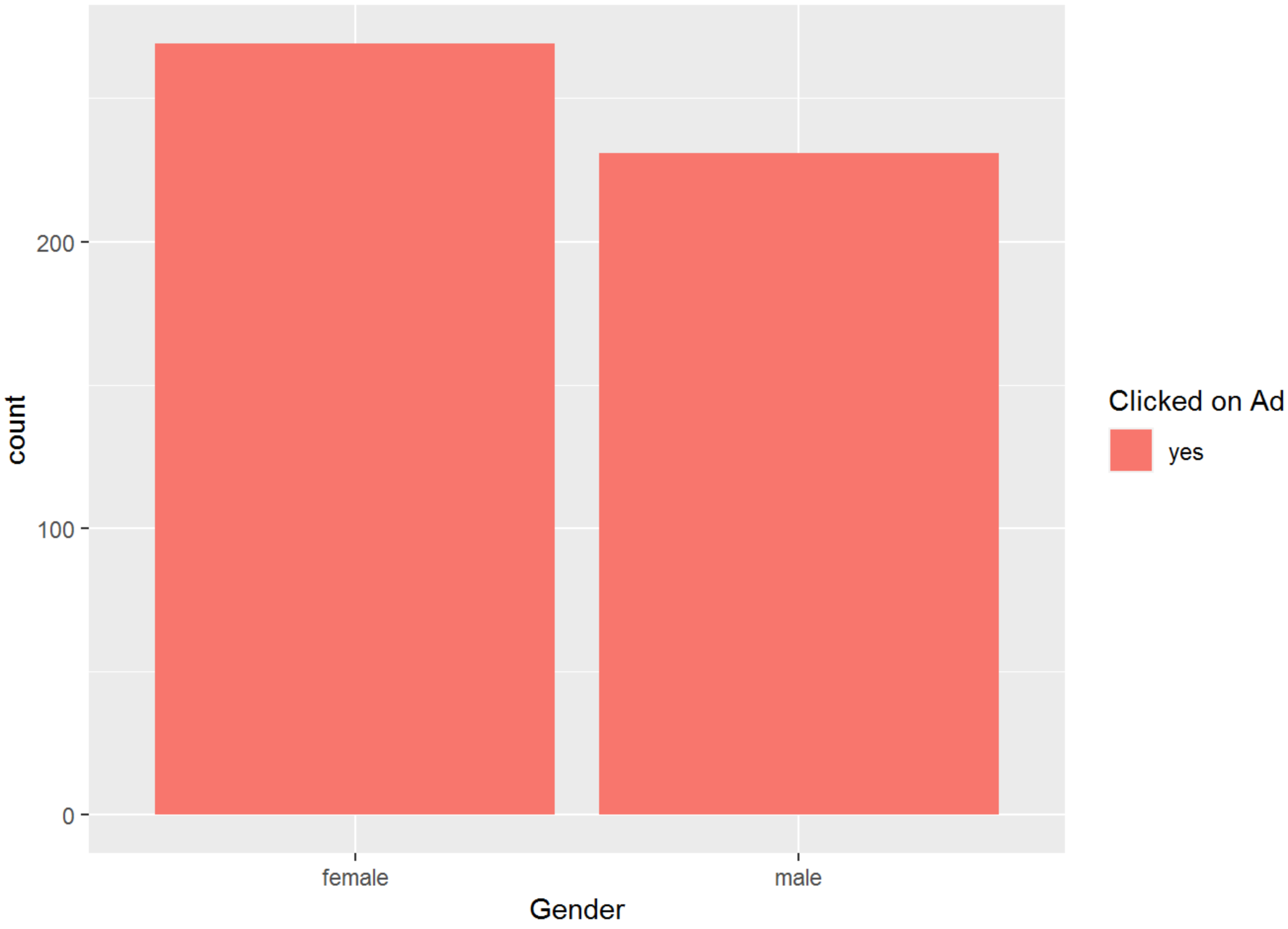
```
#A data frame with 'clicked on Ad ='no'
ad_df1<-advert_df[advert_df$`Clicked on Ad`=='no',]
ggplot(data=ad_df1, aes(x=Age,fill=`Clicked on Ad`))+
  geom_histogram(position= 'stack',bins=5)
```



```
#The age 40 and 50 were more likely to click on the Ad than those in
the aged between(30-40)
```

Clicks on Ad and Gender

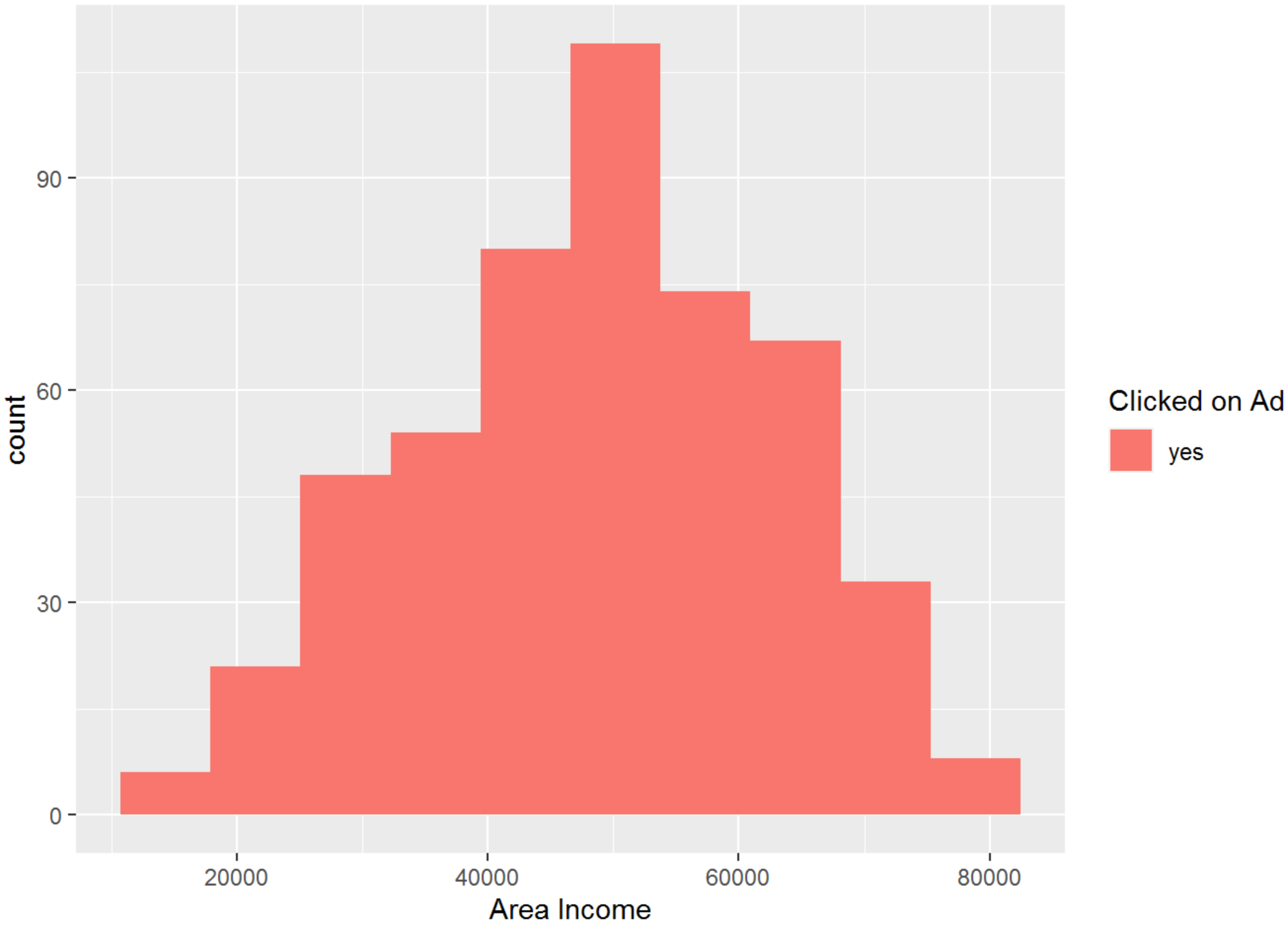
```
ggplot(data=ad_df, aes(x=Gender,fill=`Clicked on Ad`))+
  geom_bar(position= 'dodge')
```



```
#More women clicked on the ads than men.
```

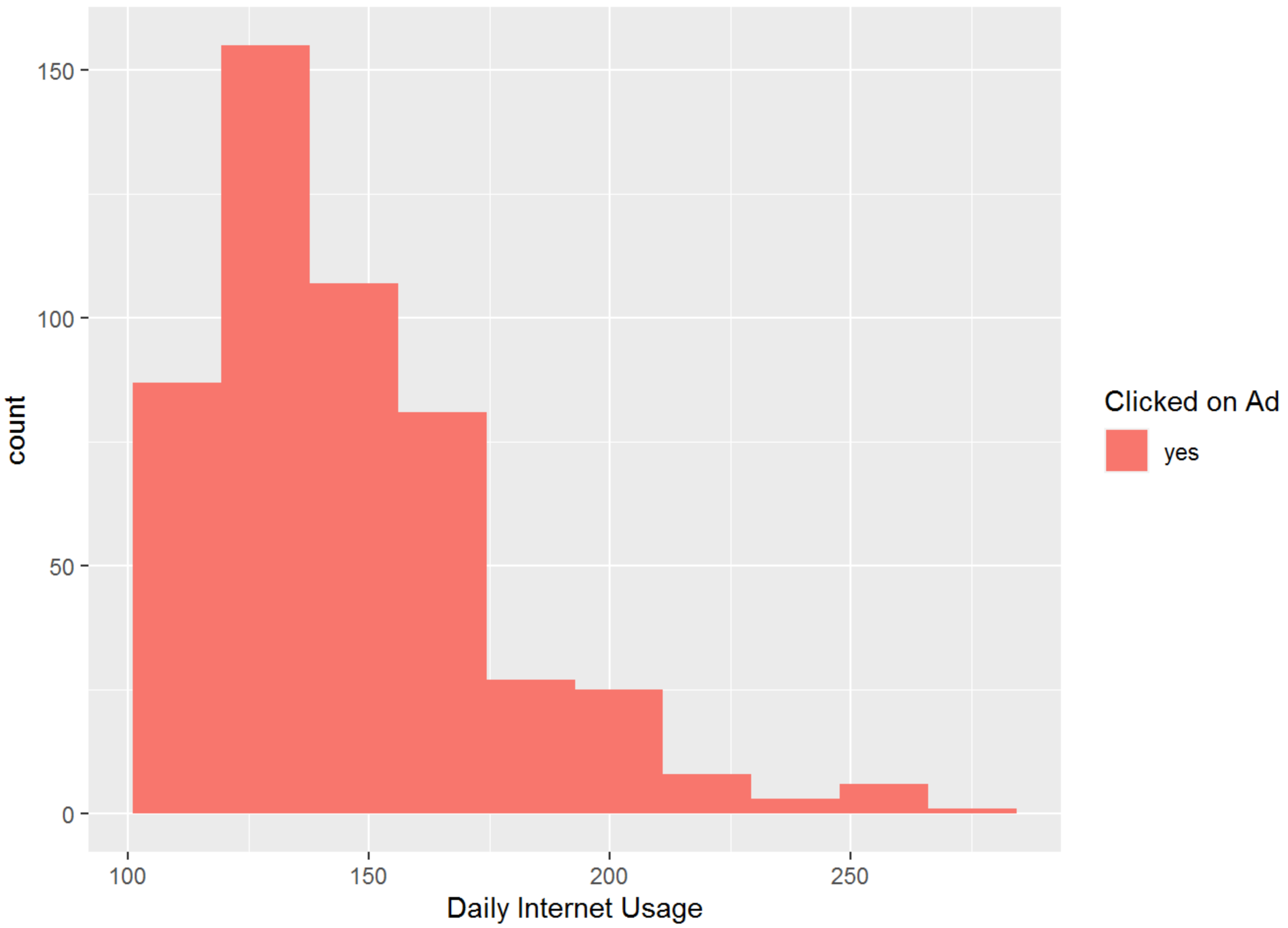
Clicks on Ad and Area income

```
ggplot(data=ad_df, aes(x=`Area Income`,fill=`Clicked on Ad`))+
  geom_histogram(position= 'stack',bins=10)
```



Clicks on Ad and Daily Internet Usage

```
ggplot(data=ad_df, aes(x=`Daily Internet Usage`,fill=`Clicked on
Ad`))+
  geom_histogram(position= 'stack',bins=10)
```



```
sort(table(ad_df$Country),descreasing= FALSE)[1:5]
##
##     Angola  Argentina    Armenia    Austria Azerbaijan
##          1          1          1          1          1
```

Modeling

Data Preparation

```
#Encoding Gender and Clicked on ad columns
advert_df$Gender <- ifelse(advert_df$Gender == 'male',1,0)
advert_df$`Clicked on Ad` <- ifelse(advert_df$`Clicked on
Ad`=='yes',1,0)
#Dropping unnecessary columns
advert_df$`Ad Topic Line` <- NULL
```

```
advert_df$City <- NULL
advert_df$Country <- NULL
advert_df$Year <- NULL
advert_df$Timestamp <- NULL
#renaming column gender back to male
colnames(advert_df)[5] <- 'Male'
#Checking the dataset
rmarkdown::paged_table(head(advert_df,n=4))
```

Splitting data

```
set.seed(5)
sample<- caret::createDataPartition(advert_df$`Clicked on Ad`, p =
0.8,
                                  list = FALSE,
                                  times = 1)
train <- advert_df[sample,]
test <- advert_df[-sample,]
advert_df$`Clicked on Ad` <-factor(advert_df$`Clicked on Ad`)
#Checking dimensions
dim(train)
## [1] 800   6
dim(test)
## [1] 200   6
```

Data Scalling

```
train[1:5]<-scale(train[1:5])
xtest<-scale(test[1:5])
```

SVM Model

```
#Training the model
library(e1071)
model= svm(formula =`Clicked on Ad` ~ .,data = train,
           type = 'C-classification',kernel = 'linear')
# Making predictions using the test data
prediction <-predict(model,xtest)
cm(table(test$`Clicked on Ad`,prediction))
##    prediction
##          0      1
##   0 251.46   2.54
##   1   5.08 248.92
```

Naive Bayes Model

```
x<-train[,1:5]
y<-train$`Clicked on Ad`

# fiting naive bayes
set.seed(42)
model1 <- naiveBayes(`Clicked on Ad` ~ ., data = train)
model
##
## Call:
## svm(formula = `Clicked on Ad` ~ ., data = train, type =
"C-classification",
##     kernel = "linear")
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  linear
##        cost:  1
##
## Number of Support Vectors:  81
prediction1 <-predict(model1,xtest)
cm(table(test$`Clicked on Ad`,prediction1))
##    prediction1
##          0      1
##   0 248.92   5.08
##   1   5.08 248.92
```

Observations

1. 38 more women than men click on the ad
2. People from Czech Republic and France contributed the highest among those who visited the blog website, however people from Turkey,Ethiopia and Australia contributed the highest among those who clicked on the Ads.
3. On average , people spent 65 minutes on the blog site and an average of 180 megabyte on daily internet usage
4. The average age is 36 years
5. Persons falling under the average area income (57012.3) were more likely to click on ads.
6. Majority of those who clicked on the Ads were between the ages of 40-50 while majority of those who skipped Ads were between ages 30-40.

Recommendations

1. Increase Ads friendly to people from Czech Republic and France since they visit the blog the most.
2. Increase female and male friendly content in the blog in increase female consumption to increase number of Ad consumption
3. Make more targeted ads for people between the ages of 30 - 40 years

Conclusion

The models has an accuracy of 98.5% at 95% confidence interval.It is statistically significant at a p value of 0.00 which is less than alpha= 0.5. However,the SVM model classifies better with only 3 misclassifications.