~~~
---
title: "Analysis for ad target audience"
author: "Cindy Gachuhi"
date: "19 November 2021"
output: html_document
---

# R Week 1 IP
## Defining the 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.

## Loading the dataset

### Dataset: http://bit.ly/IPAdvertisingData

```{r}
# Let us import the library 'Data Table'
#
library("data.table")
ad_data <- fread('http://bit.ly/IPAdvertisingData')

# preview the first 6 values
head(ad_data)

```

## Cleaning the dataset
### Removing missing values
```{r}

```

```{r}
# using the function 'colSums', we can identify the total number of missing values in each column
#
colSums(is.na(ad_data))

```
We can see that no columns in our dataset contain missing values.

### Removing duplicates

```{r}
# To get rid of duplicates, we will identify the unique values from our dataset and assign them to variable 'unique_ad'
unique_ad <- unique(ad_data)

# print the variable and check out the unique values
unique_ad

# let us confirm whether all the duplicates have been removed
duplicated_ad <- ad_data[duplicated(ad_data),]
duplicated_ad

```

It seems that the dataset has no duplicated values. We originally had 1000 rows and even after getting the unique values,
the number of rows remained the same. After printing out variable 'duplicated_ad', we had no output; confirming that
indeed our dataset has no duplicated values.

## Checking for outliers
An outlier is an observation that is numerically distant from the rest of the data. When reviewing a boxplot, an outlier
is defined as a data point that is located outside the fences (“whiskers”) of the boxplot.

```{r}
# using boxplots, we will check for outliers in various columns
# column 'Age'

boxplot(ad_data$Age) # No outliers!
#
boxplot(ad_data$`Daily Time Spent on Site`) # No outliers!
#
boxplot(ad_data$`Daily Internet Usage`) # No outliers!
#
boxplot(ad_data$`Area Income`)

# Since the outliers are not too far from our minimum value, we will keep them to avoid removing important data points

```

## Univariate analysis
### Measures and Central Tendency

```{r}
~~~

~~~~
# to get the descriptive statistics of all the numerical variables, we will use the function 'summary'
summary(ad_data)
#
# since mode doesn't have an inbuilt function, we will create a variable 'mode'
#
mode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}

# let us get the mode in column 'Age'
age.mode <- mode(ad_data$Age)

#print out mode
age.mode

```


### Measures of Dispersion

```{r}
# we already got the quantiles, minimum and maximum values
# for the rest of the measures, we will focus on column 'daily time spent on site'
#
# range
site_time.range <- range(ad_data$`Daily Time Spent on Site`)
site_time.range

# variance
# The variance is a numerical measure of how the data values is dispersed around the mean.
site_time.var <- var(ad_data$`Daily Time Spent on Site`)
site_time.var

# standard deviation
site_time.sd <- sd(ad_data$`Daily Time Spent on Site`)
site_time.sd

```
### Univariate visualizations
#### Frequency distribution using a barplot

```{r}
# fetching the 'Age' column
age <- ad_data$`Age`

# computing the frequency distribution using the table() function
age.fr <- table(age)
age.fr

# using a barplot to visualize this
barplot(age.fr)

# From ages 28-37(the more frequent ages), age 31 is the most frequent
# Age 61 is the least frequent

```


## Bivariate analysis

```{r}

# Correlation Coefficient

# It is a normalized measurement of how the two are linearly related. If the correlation coefficient is close to 1, it would indicate that the variables are positively linearly related. For -1, it indicates that the variables are negatively linearly related and the scatter plot almost falls along a straight line with negative slope. And for zero, it would indicate a weak linear relationship between the variables.

# get the columns
daily_usage <- ad_data$`Daily Internet Usage`
clicked <- ad_data$`Clicked on Ad`
site_time <- ad_data$`Daily Time Spent on Site`

# get the correlation coefficients between the variables using cor()
cor(daily_usage , site_time)
# 0.5; indicates a positive correlation although it is a bit weak.
# However, we can see that people who have a higher daily internet usage spend more time on the site

cor(clicked, daily_usage)
# -0.78; a negative correlation which is relatively strong.
# We see that people who spend more time on the site and on the Internet tend to click less on the ads.

```
### Visualizations

```{r}
# scatterplots
~~~~

```
plot(daily_usage, site_time, xlab="Daily Internet usage", ylab="Time spent on site")

# The points are quite scattered but if you look closely, you see the points show a weak positive linear relationship
# between the two variables

plot(daily_usage, clicked, xlab="Time spent on site", ylab="Clicked on ads")
# With how the points are distributed, we realize a rather stronge negative correlation between the variables

```

From our bivariate analysis, we see that the users who have higher daily internet usage tend to spend more time on the site. Also, we realised that those users who spend more time on the site, actually tend to click on the ads less. This might be because the ads are not in their interest hence they tend to ignore them. A very brief survey should be taken by the user before using the site so as to tailor the ads to the user's interests which might make them click on them more frequently.