# Report for R script

**Monika Aulich-Olszewska**

[1]Data used for this report comes from Kaggle

monika.aulich.olszewska@gmail.com

## 1. Introduction

This file is a report with proper plots, information and tables from Superstore_data.Rmd file.
This report is made only for training purposes.
The script was made to show some of my skills in R and statistics.
Right now my goal is to find an interesting job as Data Analyst, Data Scientist and related.

## 2. Description

Here will be listed code, plots and some resume. Whole script with commentary is on the repository.
Repository can be found here: https://github.com/senpai-ao/training
Data that was used can be found here: https://www.kaggle.com/datasets/vivek468/superstore-dataset-final

```
1 library(ggplot2)
2
3 csv_file <- read.csv("Sample-Superstore.csv", header = TRUE)
4 df <- data.frame(csv_file)
5
6 #to check if data is loaded one can use head() function
7 head(df)
```

Next step is to choose interesting data and what to do with them. At first I want to see what product is the most bought

```
1 ggplot(df, aes(x=Product.ID, y=Row.ID)) +
2 geom_bar(stat = "identity", color="lightpink") +
3 labs(x="Product ID",y="Number of products",
4     title="Bar plot with numbers of sold products"))
5
6 #to check if data is loaded one can use head() function
7 head(df)
```

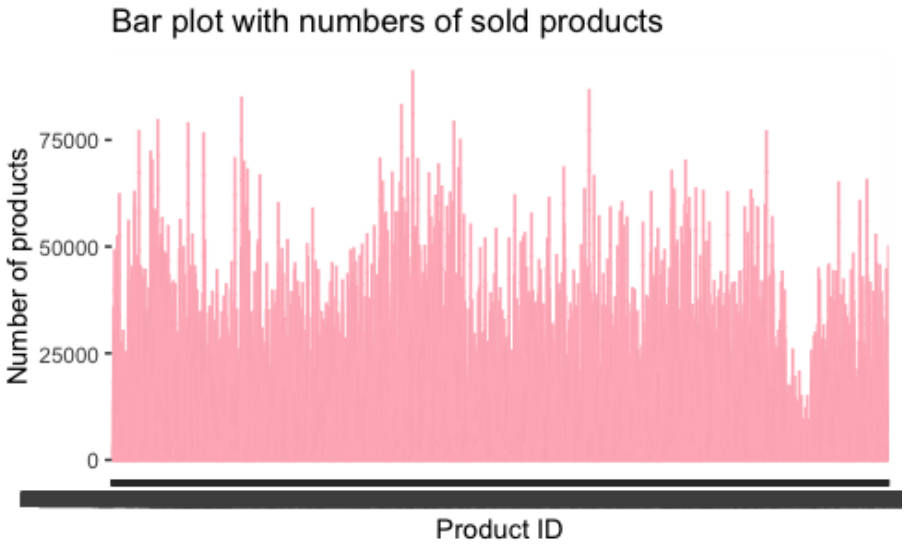


**Figura 1. Plot 1**

As we can see, this plot won't tell us much. There is too many different data in one category. Let's try wider category - Segment.

```
ggplot(df, aes(x=Segment, y=Row.ID)) +
geom_bar(stat = "identity", color="lightblue") +
labs(x="Cathegory name",y="Number of
    sold products in segment",
     title="Bar plot with numbers of sold products")
```

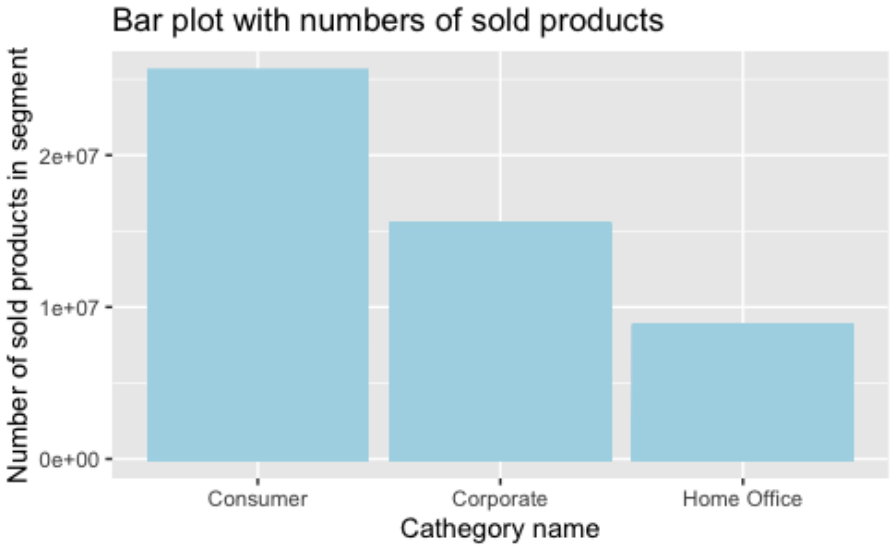


**Figura 2. Plot 2**

So right now one can see, that the biggest group in Superstore is Consumer group. Let's see what info we can get from Home Office data. First step here is to get data from Home office group only.

```
df_ho <- df[df$Segment=="Home Office",]
head(df_ho)
print(paste("Number of sold products for Home Office
    Group is",nrow(df_ho)))

> "Number of sold products for Home Office Group is 1783"

ggplot(df_ho, aes(x=Ship.Mode, y=Row.ID)) +
geom_bar(stat = "identity", color="lightpink") +
labs(x="Product ID",y="Number of products",
     title="Bar plot with chosen ship mode")
```

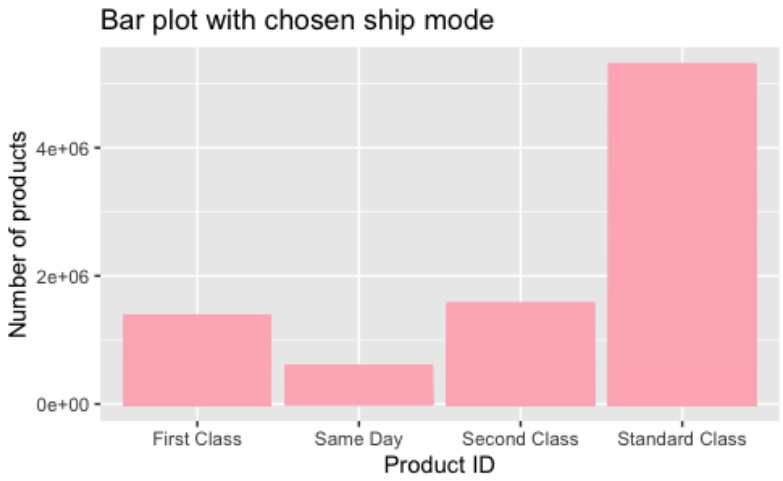


**Figura 3. Plot 3**

As we can see the most used ship mode is Standard Class. I would like to see if there is a correlation between used ship mode and product category. I used easy function cor().

```
df_corr <- data.frame(df_ho$Ship.Mode,df_ho$Category)
colnames(df_corr) <- c("Ship.Mode","Category")

#only numeric variables can be used to computate correlation
df_corr$Ship_num <- as.numeric(as.factor(df_corr$Ship.Mode))
df_corr$Categ_num <- as.numeric(as.factor(df_corr$Category))

cor_ship_cat <- cor(df_corr$Ship_num,df_corr$Categ_num)

print(paste("Correlation is", round(cor_ship_cat,3)))
```

>"Correlation is -0.016"
The correlation is very close to 0, so one can say that there is no correlation between used

ship mode and product category. What interesting is for me, I would also like to see, how many orders were in what year in chosen group. At first I need to change data format.

```
df_ho$newdate <- strptime(as.character(df_ho$Order.Date),
    "%m/%d/%Y")
df_ho$Date <- format(df_ho$newdate, "%Y-%m-%d")
splitted_data <- split(df_ho,
    format(as.Date(df_ho$Date,format="%Y-%m-%d"),
    format = "%Y"))
df_dates <- data.frame()
for(i in 1:length(splitted_data)){
  data <- splitted_data[[i]]
  year <- format(as.Date(data$Date[1],
    format="%Y-%m-%d"),"%Y")
  print(paste("In year",year[1],"there was",
    nrow(data),"solds"))
  line <- c(year,nrow(data))
  df_dates <- rbind(df_dates,line)}
colnames(df_dates) <- c("Year","Solds")
ggplot(data=df_dates, aes(x=Year, y=Solds, group=1)) +
geom_line(color="pink")+ geom_point() +
labs(x="Years",y="Number of sold products",
    title="Plot with number of sold products per year")
```

>[1] "In year 2014 there was 312 solds"

>[1] "In year 2015 there was 341 solds"

>[1] "In year 2016 there was 466 solds"

>[1] "In year 2017 there was 664 solds"

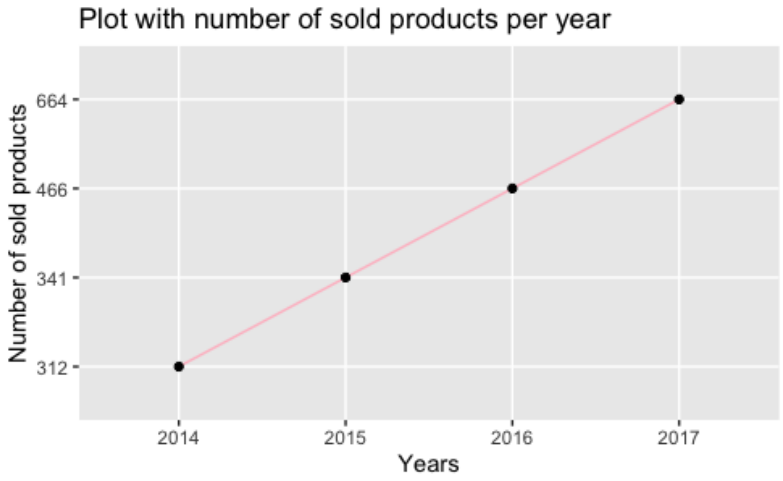


**Figura 4. Plot 4**