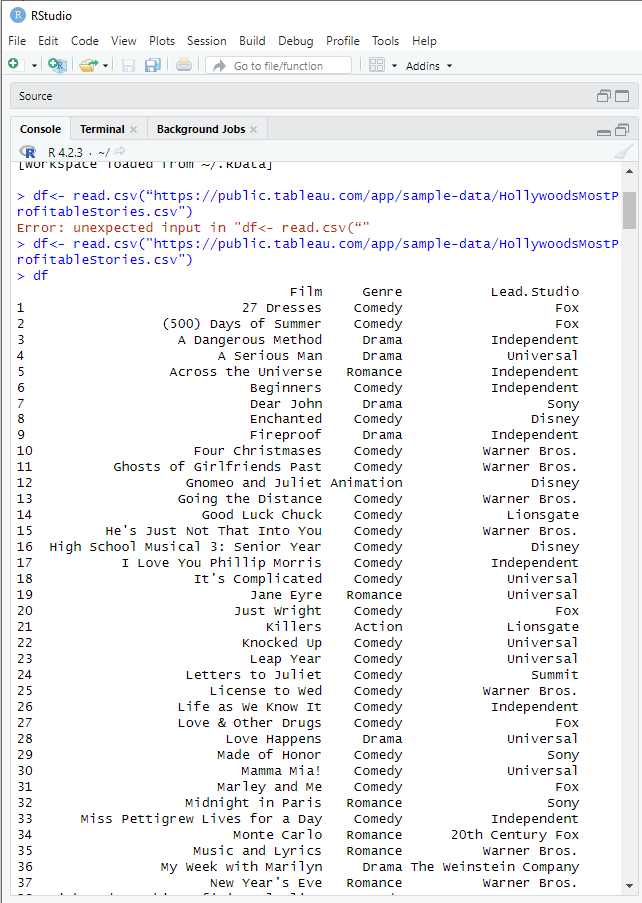
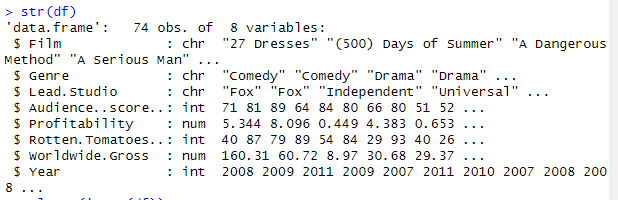
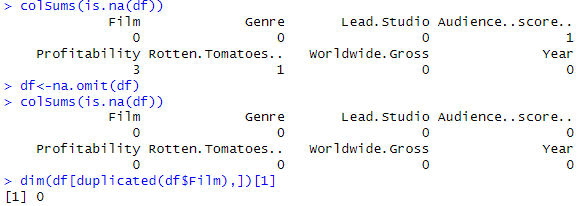
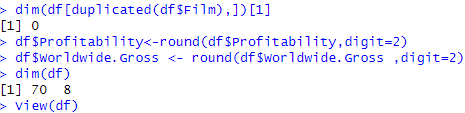
**TASKS**

1. **I first run df<- read.csv(“https://public.tableau.com/app/sample-data/HollywoodsMostProfitableStories.csv") to load the data that I’m about to use and use the df to read it**

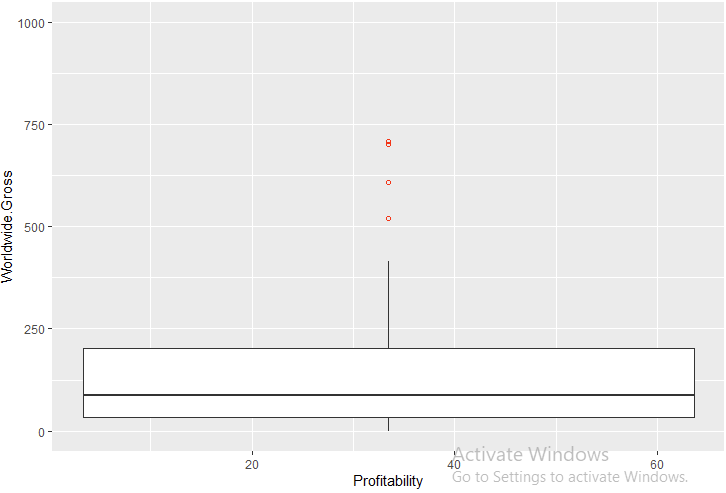
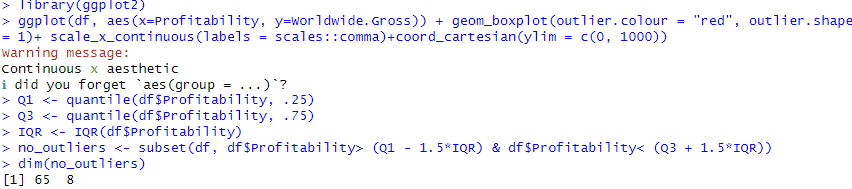


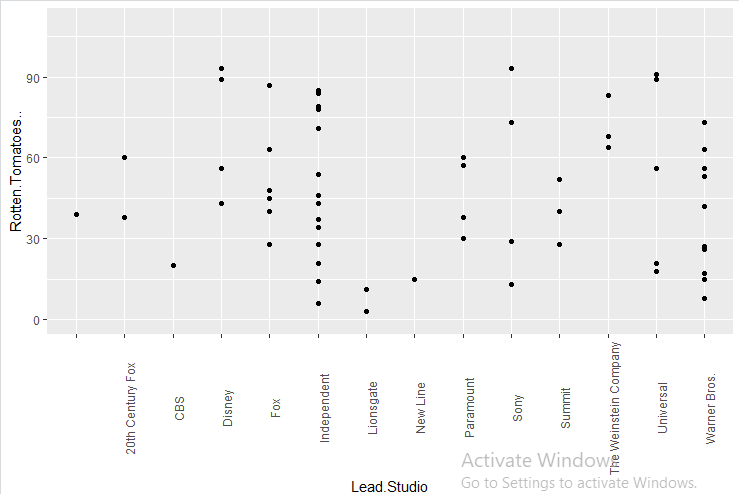
**After words I run install.packages(“tidyverse”) for the tools I’m going to use. After words I run str(df) to see the structure of the data.**

2. **After I’ve done this I go and check if there is any missing values and I get rid of them.**

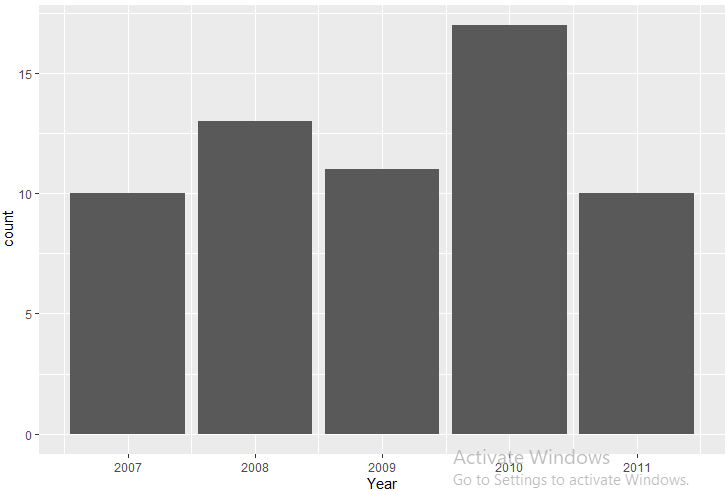
3. **I then check for any duplicates.** 

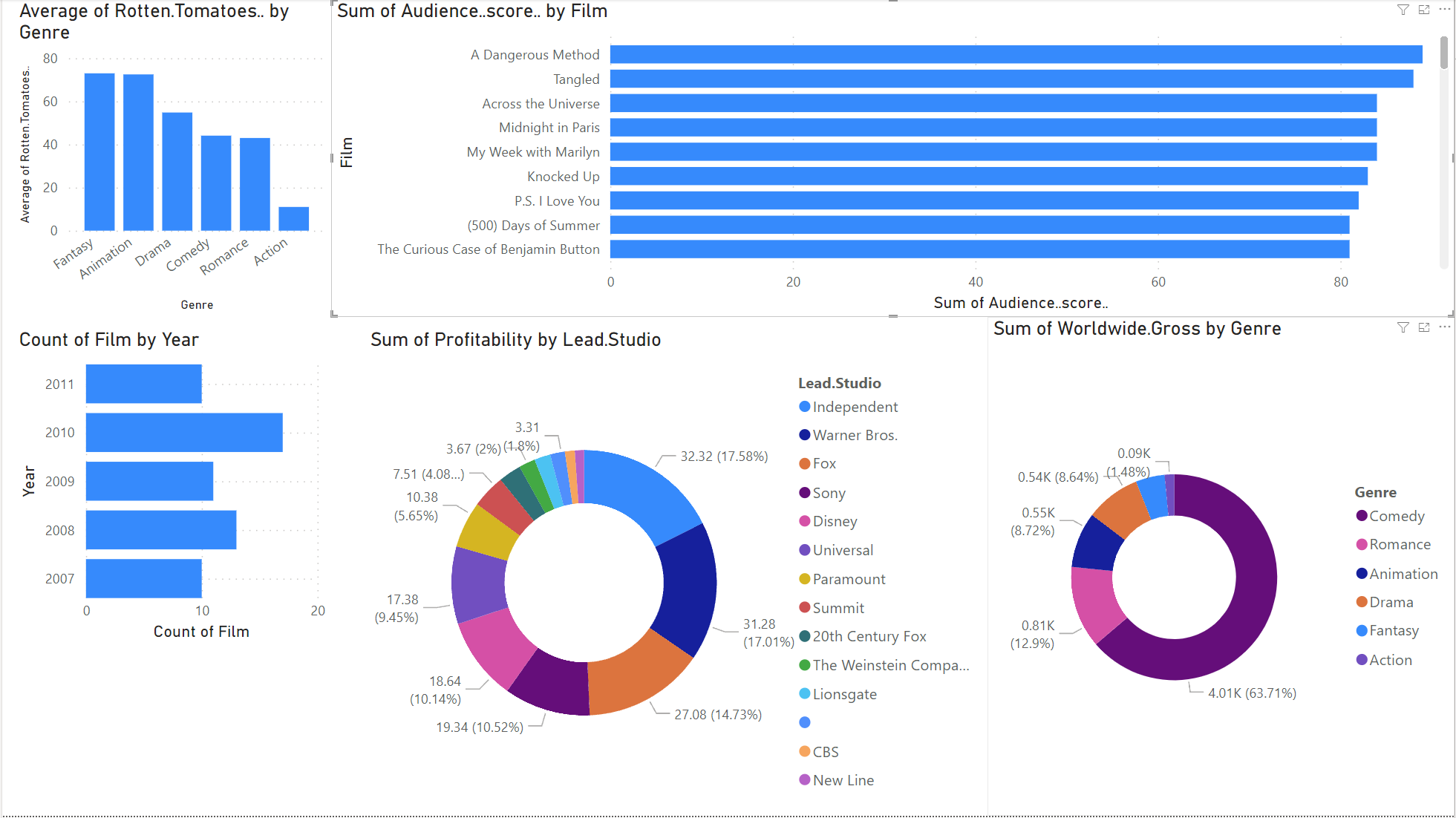
**Charts**

4. **I make a plot chart and remove outliers in Profitability**







5. **After I have done this and clean the data I would put in write.csv(df1, "clean\_df.csv") to save the file and import it to powerbi where I start to follow what the client wants to see.** 

https://app.powerbi.com/groups/me/reports/1fba8dac-600b-4553-bff4-9ed5414eb17a/ReportSection

**INTERVIEW QUESTIONS**

Q1:What is the difference between Univariate, Bivariate, and multivariate analysis?

A1: Univariate statistics summarise only one variable at a time. Bivariate statistics compare two variables. Multivariate statistics compare more than two variables.

Q2: What is an outlier and how to identify them?

A2: Outliers, as the name suggests, are the data points that lie away from the other points of the dataset. That is the data values that appear away from other data values and hence disturb the overall distribution of the dataset. This is usually assumed as an abnormal distribution of the data values.One of the easiest ways to identify outliers in R is by visualizing them in boxplots.

Q3: What is correlation?

A3: Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate). It's a common tool for describing simple relationships without making a statement about cause and effect.