JUST IT, London

DATA TECHNICIAN BOOT CAMP

PROJECT 3

Programming Language R and Power BI

- Registering with Studio R
- Downloading data
- Basic manipulations
- Cleaning Data
- Visualisation

STUDENT: Alexei Kosyhin

Date: 20 Feb 2022

Content

Pages

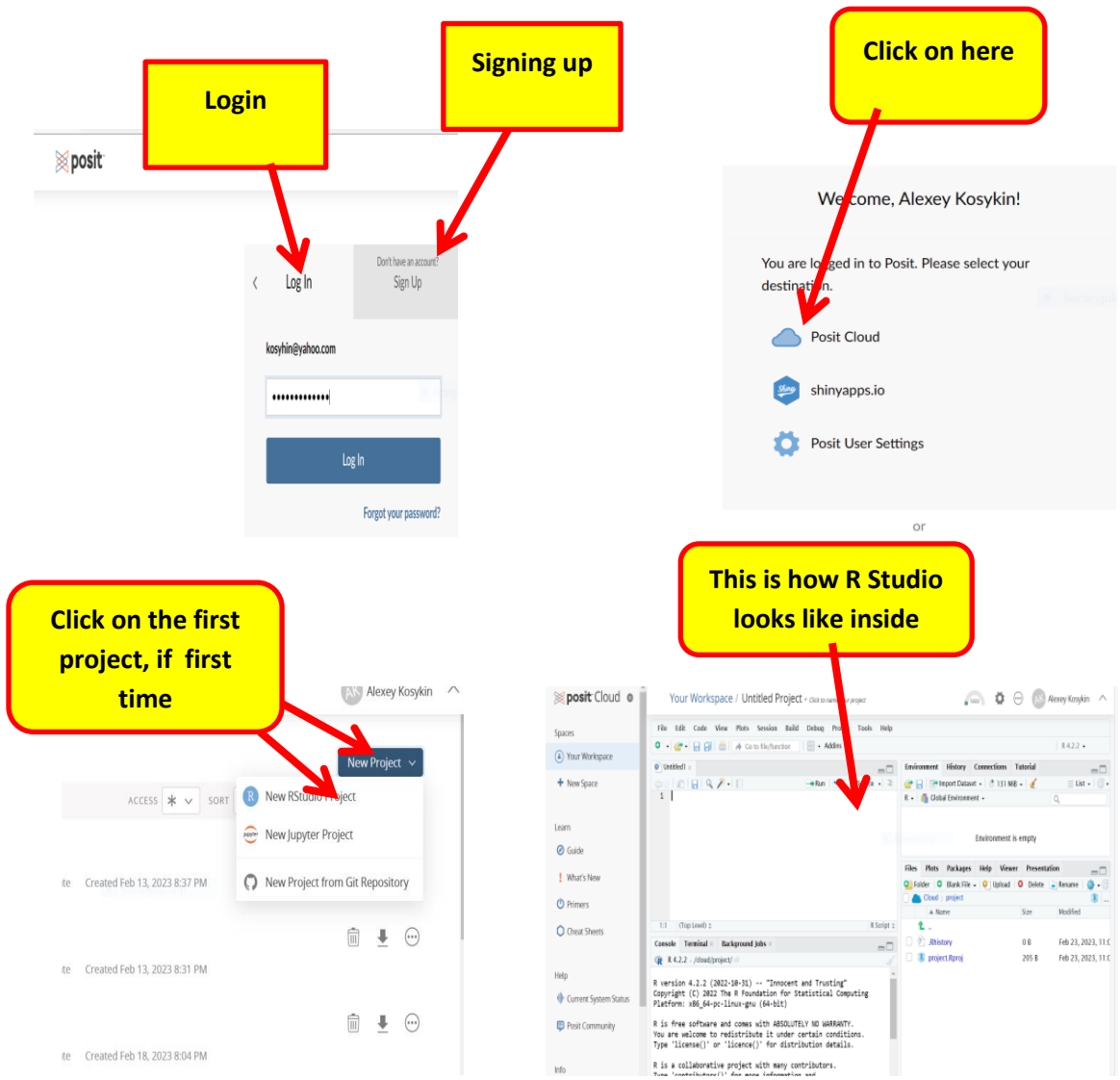
About Programming Language R, Registration, Lanching

From Wikipedia: **R** is a programming language for statistical computing and graphics supported by the R Core Team and the R Foundation for Statistical Computing. Created by statisticians Ross Ihaka and Robert Gentleman, R is used among data miners, bioinformaticians and statisticians for data analysis and developing statistical software.[7] Users have created packages to augment the functions of the R language.

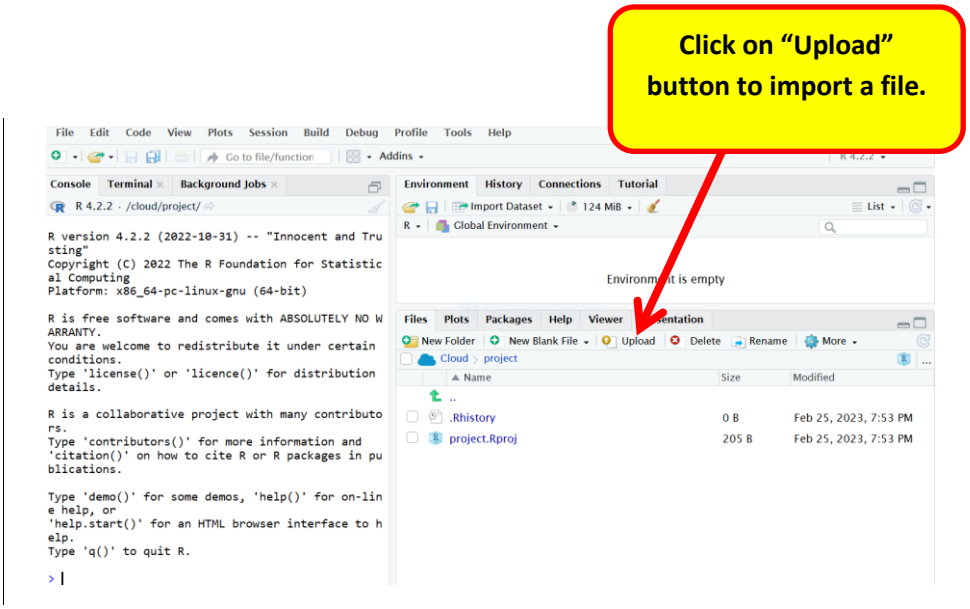
RStudio is an integrated development environment for R, a programming language for statistical computing and graphics. It is available in two formats: R Studio Desktop is a regular desktop application while RStudio Server runs on a remote server and allows accessing R Studio using a web browser.

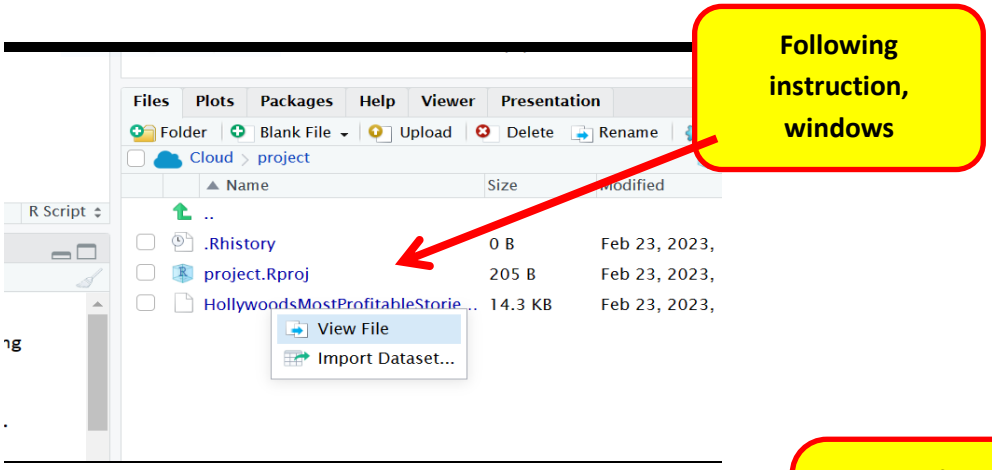
---

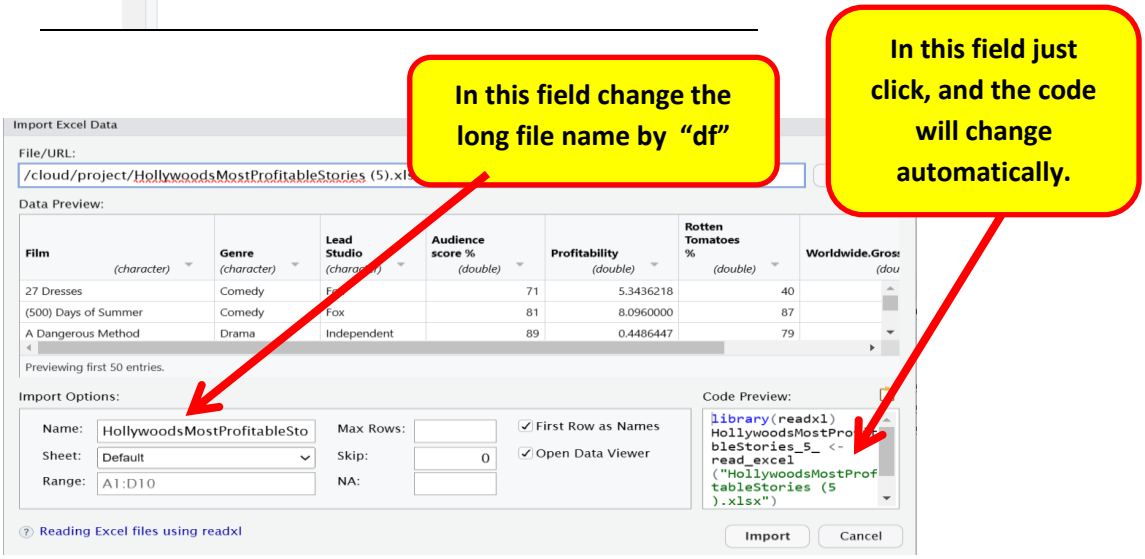
https://login.posit.cloud is the link for using an online R Studio version

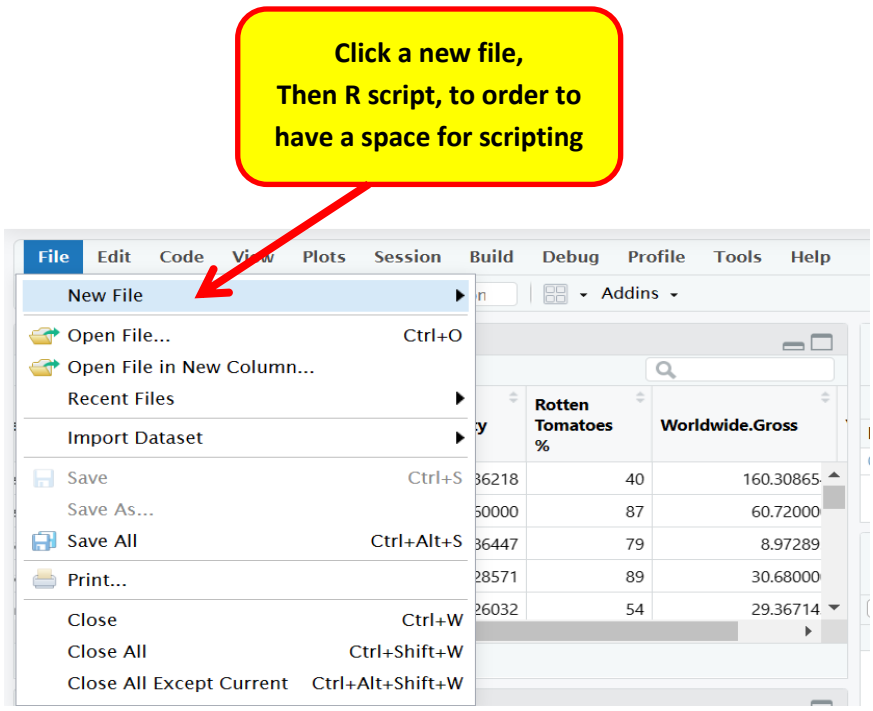


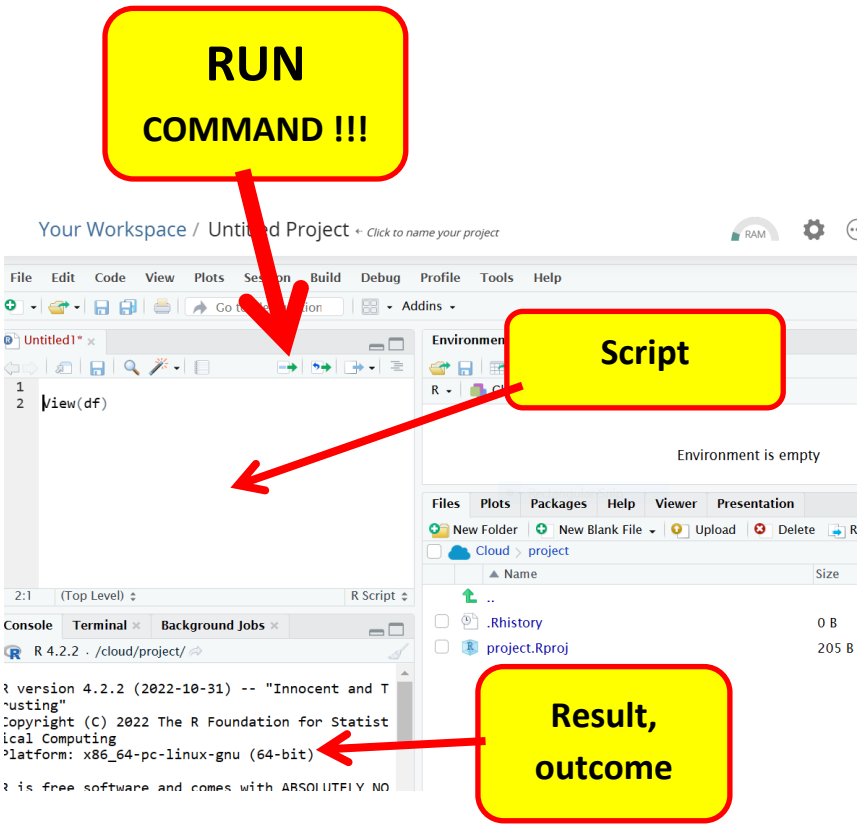
Uploading data and creating data frame





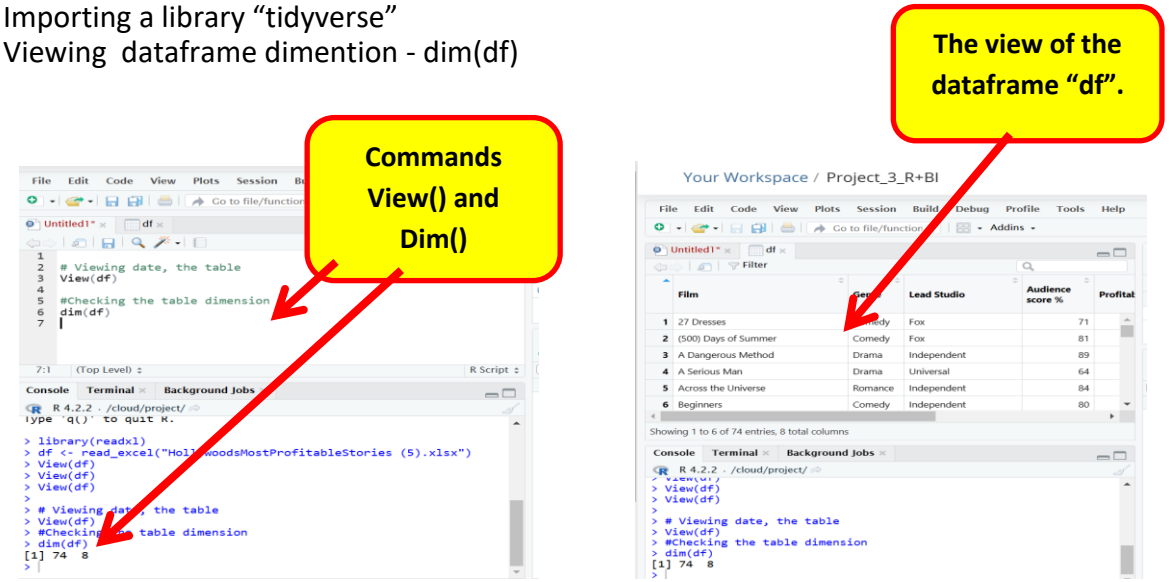


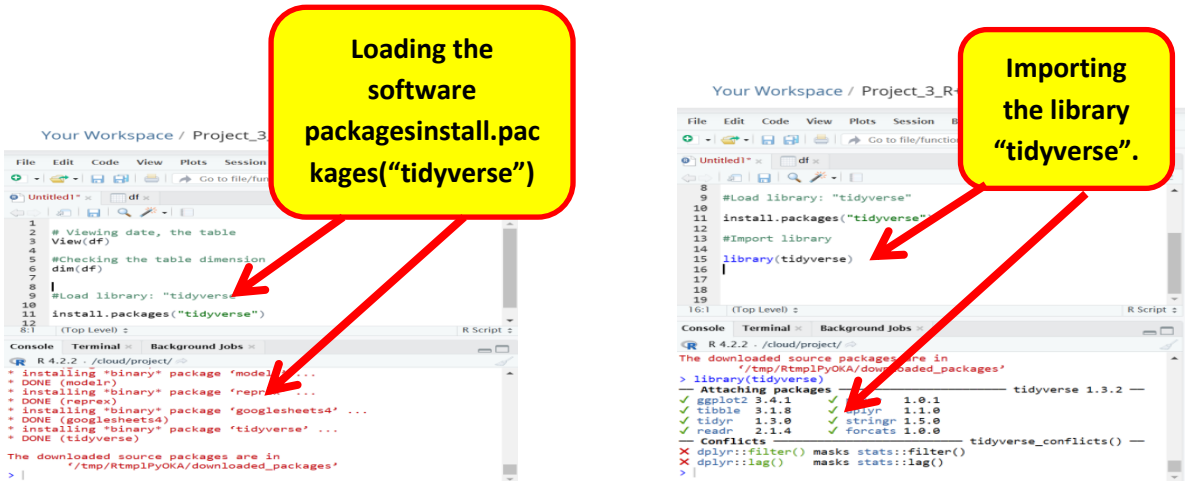

Click a new file,
Then R script, to order to
have a space for scripting
File
Edit
Code
View
Plots
Session
Build
Debug
Profile
Tools
Help
New File
Open File... Ctrl+O
Open File in New Column...
Recent Files
Import Dataset
Save Ctrl+S
Save As...
Save All Ctrl+Alt+S
Print...
Close Ctrl+W
Close All Ctrl+Shift+W
Close All Except Current Ctrl+Alt+Shift+W
Addins
Rotten Tomatoes %
Worldwide.Gross
36218 40 160.30865
60000 87 60.72000
36447 79 8.97289
28571 89 30.68000
26032 54 29.36714


RUN
COMMAND !!!
Your Workspace / Untitled Project
Click to name your project
File
Edit
Code
View
Plots
Session
Build
Debug
Profile
Tools
Help
Addins
Untitled1*
1
2
View(df)
Script
Environment is empty
Files
Plots
Packages
Help
Viewer
Presentation
New Folder
New Blank File
Upload
Delete
Cloud
project
Name
Size
.Rhistory 0 B
project.Rproj 205 B
2:1
(Top Level)
R Script
Console
Terminal
Background Jobs
R 4.2.2 · /cloud/project/
version 4.2.2 (2022-10-31) -- "Innocent and T
usting"
Copyright (C) 2022 The R Foundation for Statist
ical Computing
Platform: x86_64-pc-linux-gnu (64-bit)
Result,
outcome

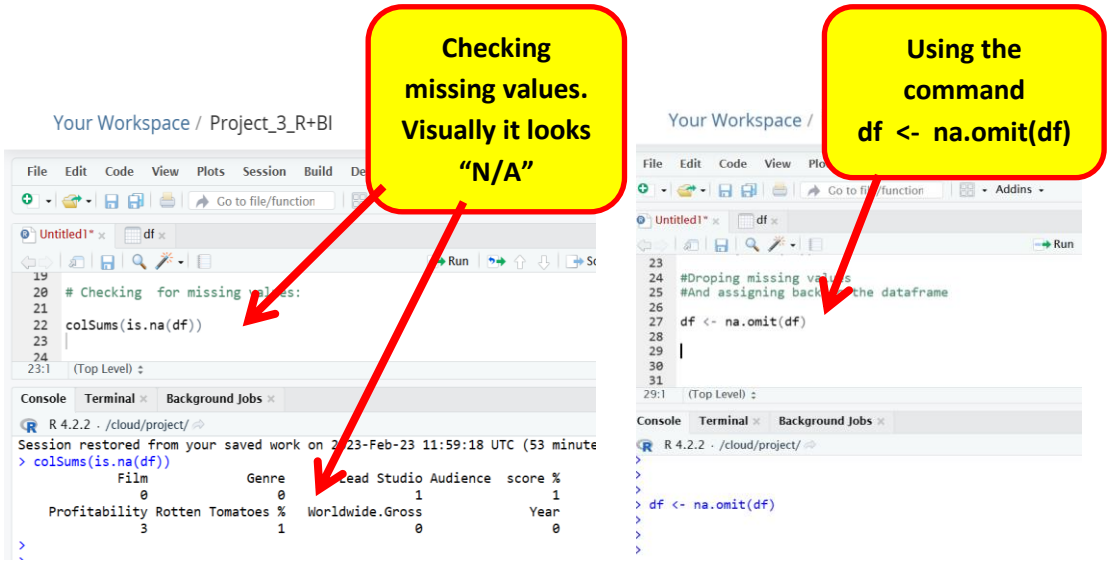
Step1: Initial Exploratory Analysis

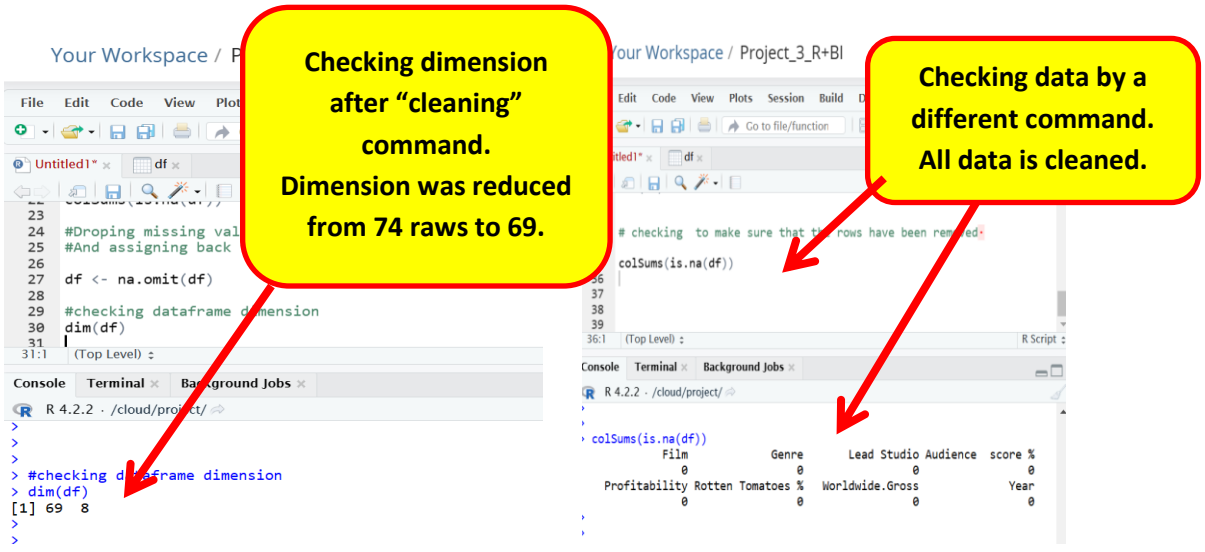
Basic manipulation, installing packages and importing libraries.
Viewing the data frame “df” using a command View(df)
Importing a library “tidyverse”
Viewing dataframe dimention - dim(df)

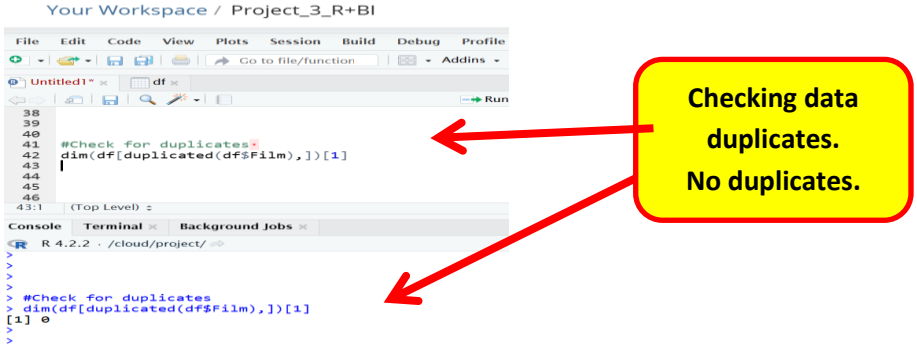




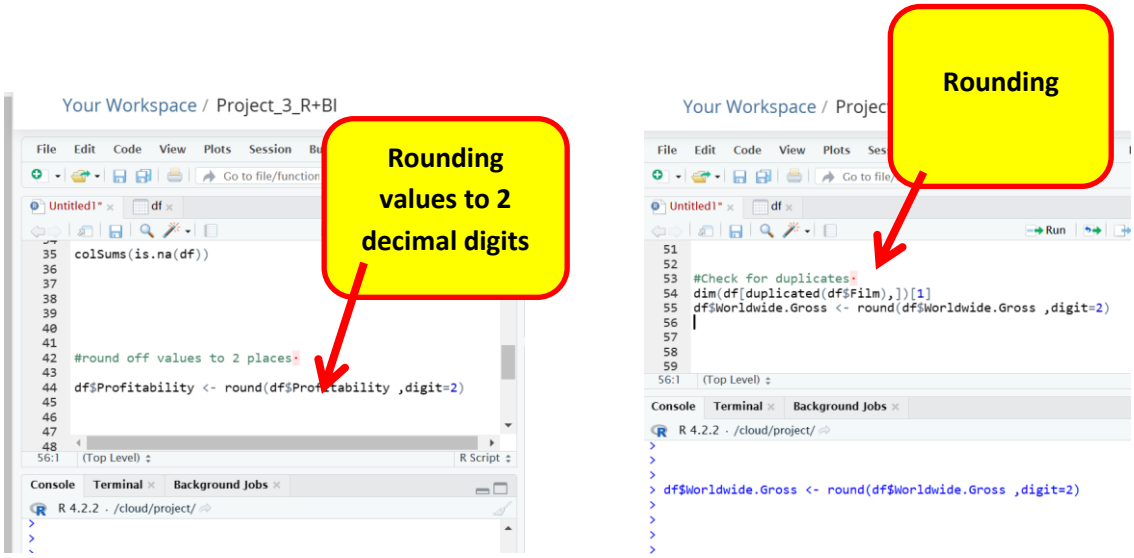
Step 2: Cleaning Data

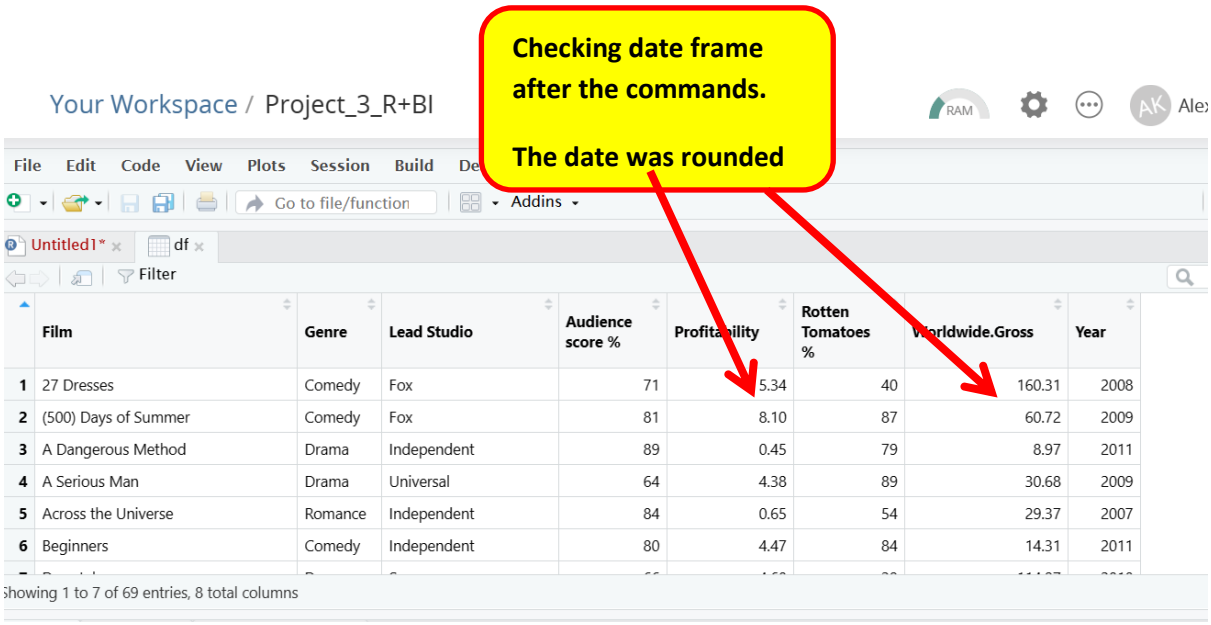






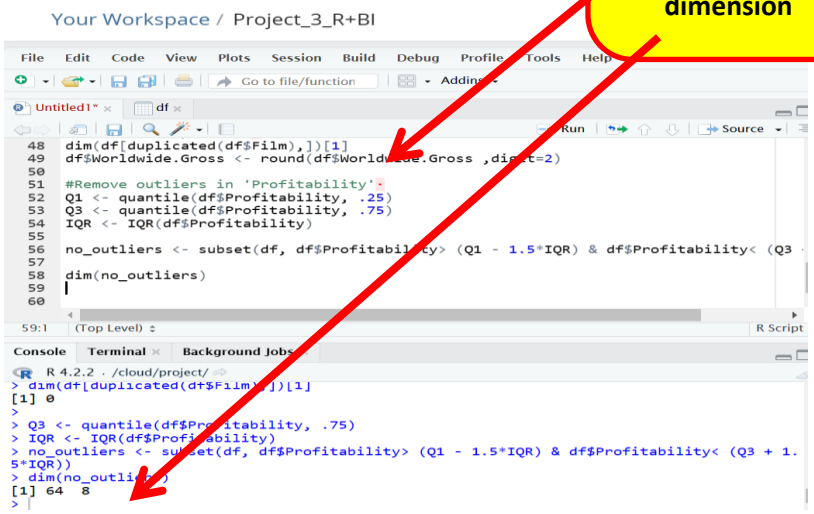
Rounding values

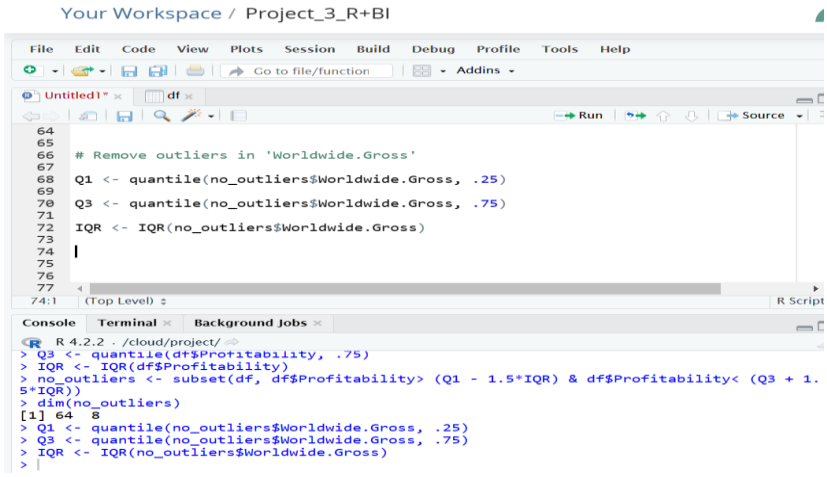




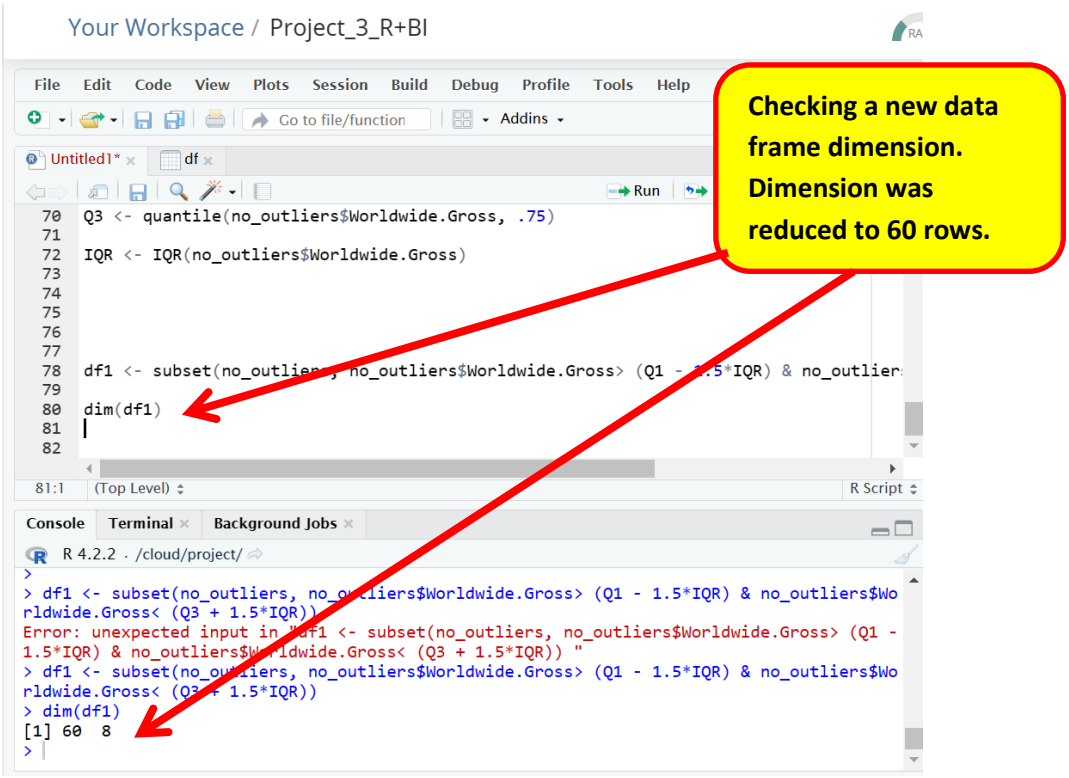
Removing Outliers

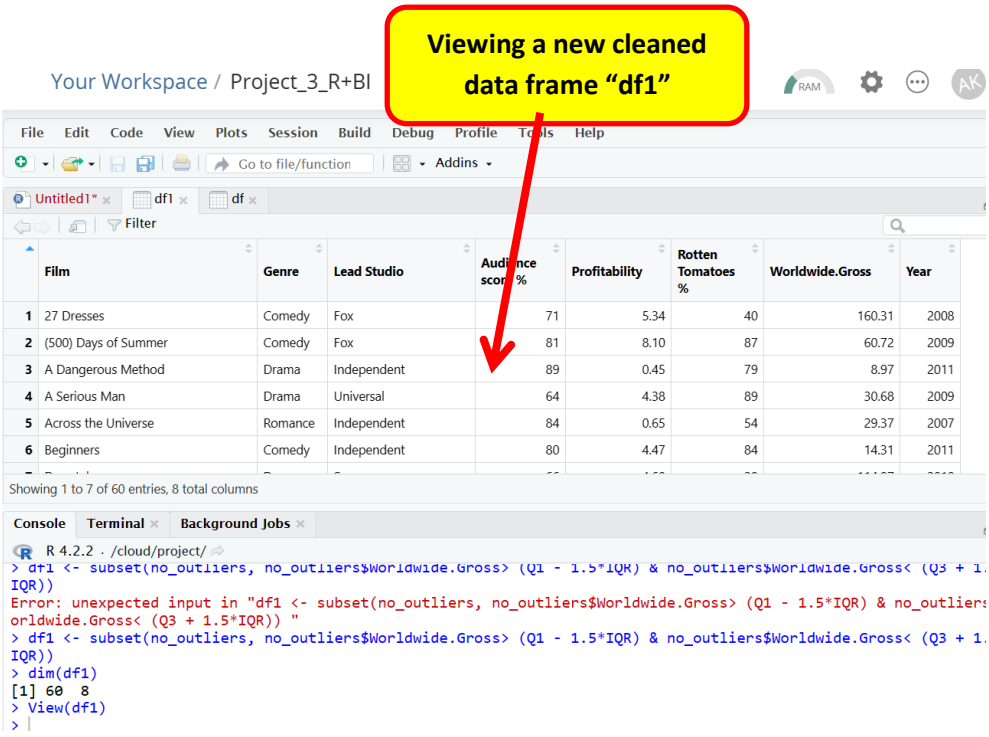
Removing ouliers/ tails And checking dtaframe dimension





Checking a new cleaned data frame

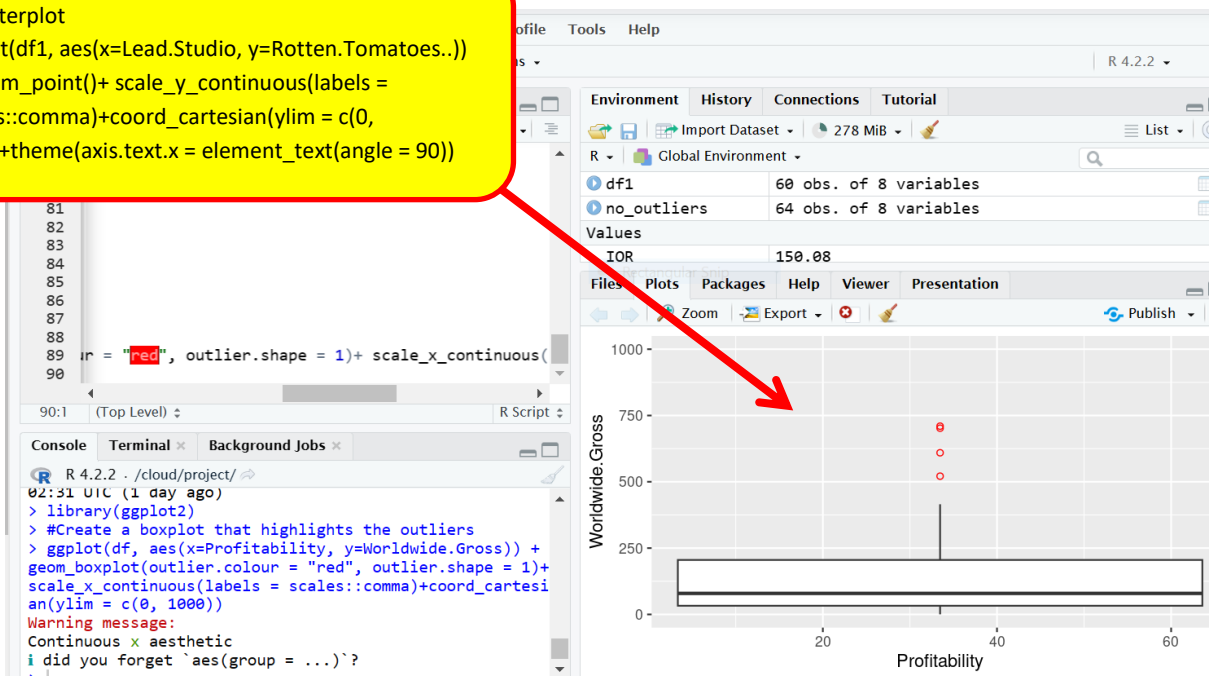




	Film	Genre	Lead Studio	Audience score %	Profitability	Rotten Tomatoes %	Worldwide.Gross	Year
1	27 Dresses	Comedy	Fox	71	5.34	40	160.31	2008
2	(500) Days of Summer	Comedy	Fox	81	8.10	87	60.72	2009
3	A Dangerous Method	Drama	Independent	89	0.45	79	8.97	2011
4	A Serious Man	Drama	Universal	64	4.38	89	30.68	2009
5	Across the Universe	Romance	Independent	84	0.65	54	29.37	2007
6	Beginners	Comedy	Independent	80	4.47	84	14.31	2011

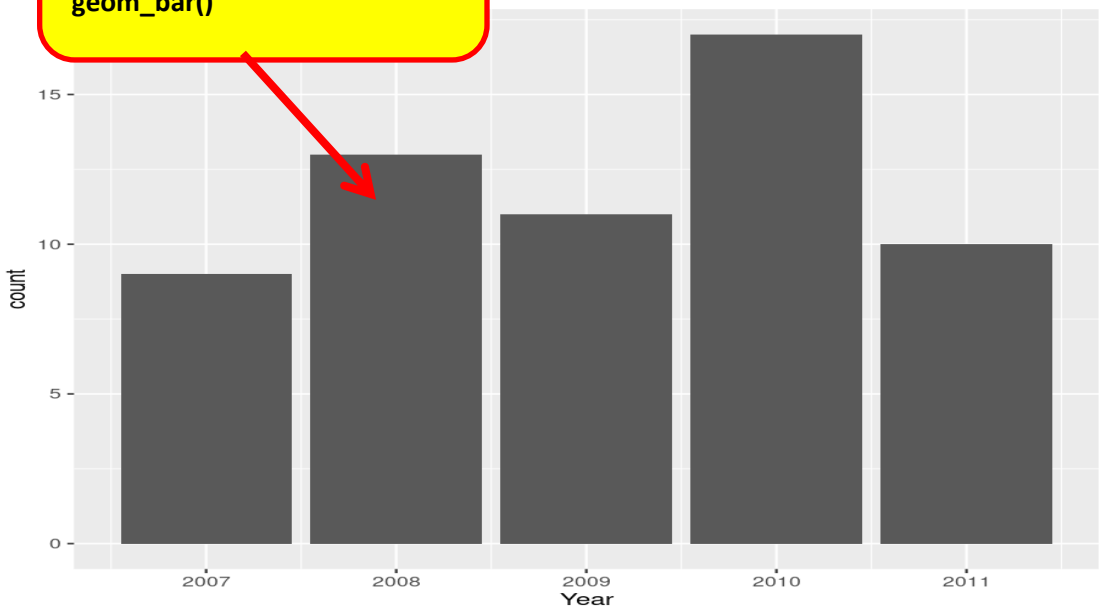
Step 3: Exploratory Data Analysis

#scatterplot
ggplot(df1, aes(x=Lead.Studio, y=Rotten.Tomatoes..)) + geom_point()+ scale_y_continuous(labels = scales::comma)+coord_cartesian(ylim = c(0, 110))+theme(axis.text.x = element_text(angle = 90))



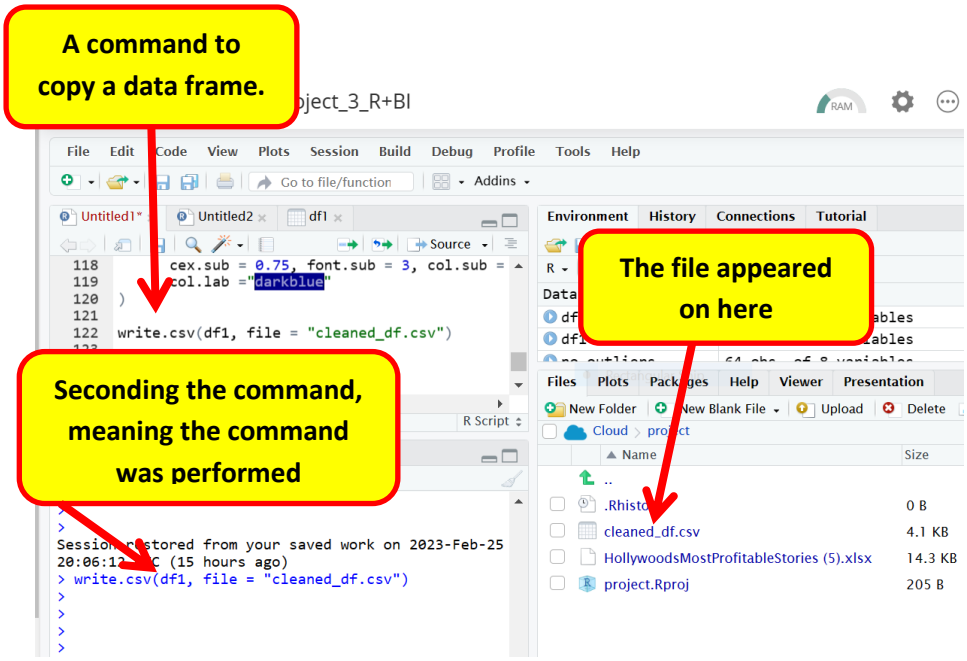
#bar chart

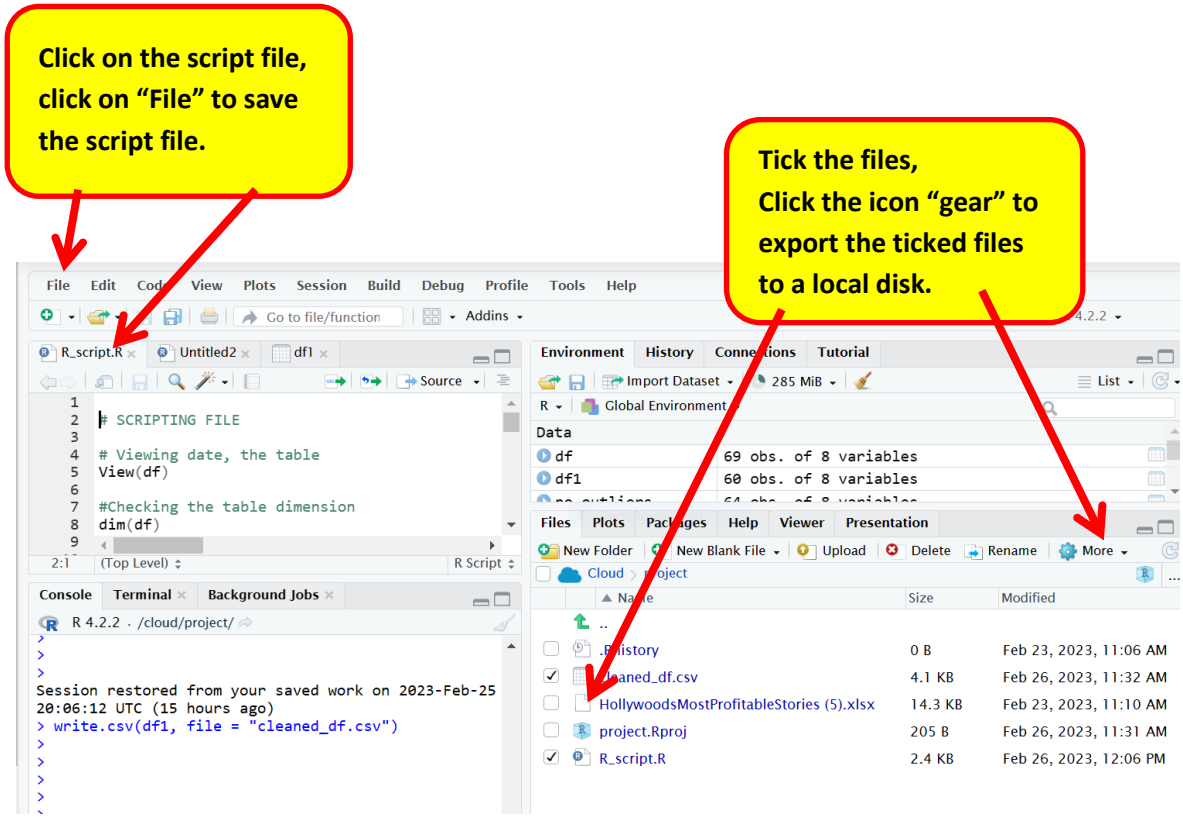
ggplot(df1, aes(x=Year)) + geom_bar()



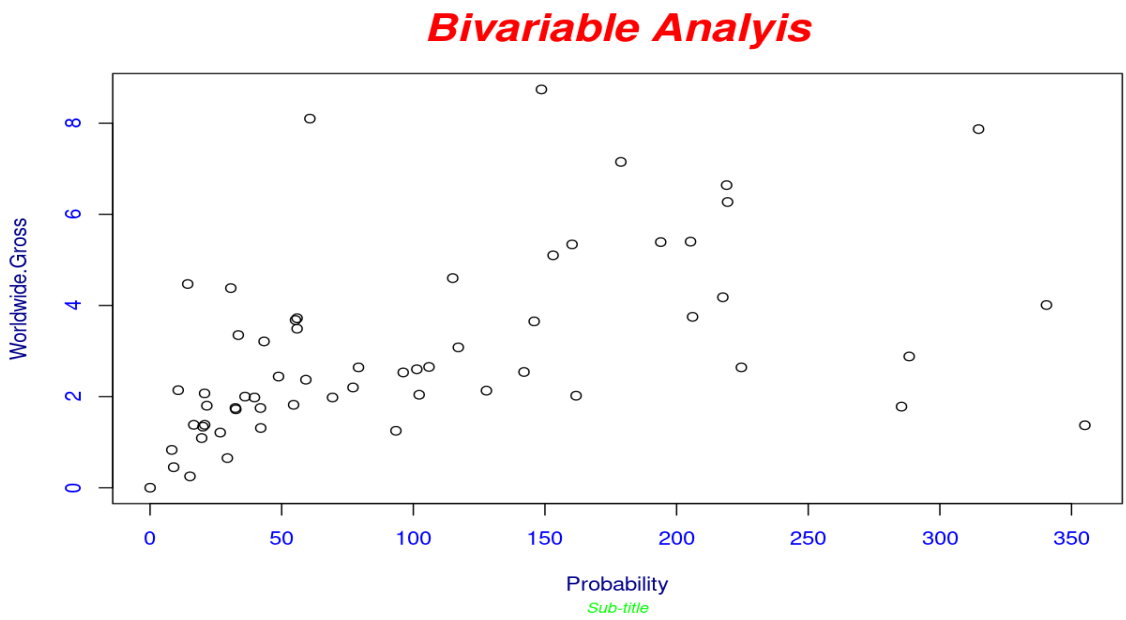
Step 4: Export data

It's vitally important to save the cleaned data and working scripting files to a local disk !!! Because it can be used for next analysis, prof of the work done, repeating scripting again.





Additional plots and analysis

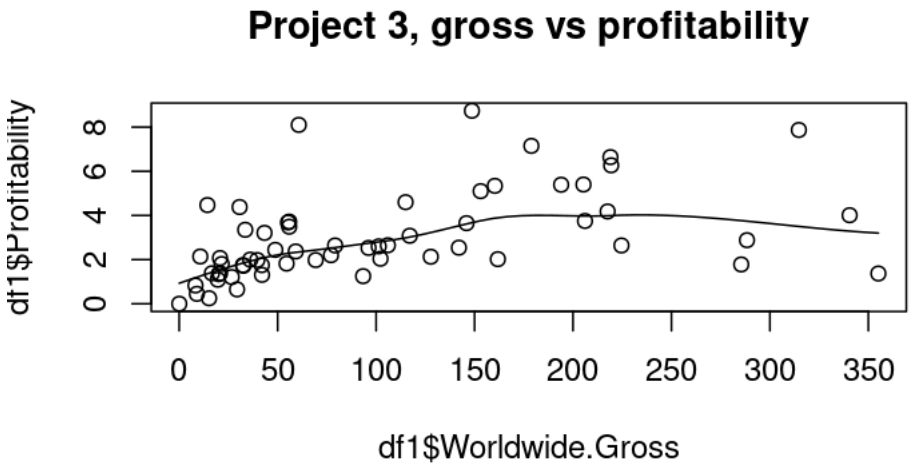


Using this R script:

plot(df1$Worldwide.Gross, df1$Profitability, main ="Probability vs World.Gross")

Plotting scatter using R

scatter.smooth(df1$Worldwide.Gross, df1$Profitability, main='Project 3, gross vs profitability')



Calculating correlations between variables:

```
78  # CALCULATING COROLATIONS:
79
80  #Calculating correlations for the variables
81  cor(df1$Profitability, df1$Worldwide.Gross)
82  #0.47 which moderate.
83
84  cor(df1$Profitability, df1$`Rotten Tomatoes %`)
85  #0.146 which very weak
86
87  cor(df1$Worldwide.Gross, df1$`Rotten Tomatoes %`)
88  # 0.035 no correlation
89
90  cor(df1$`Audience  score %`, df1$`Rotten Tomatoes %`)
91  # 0.60 is a strong correlation
92
```

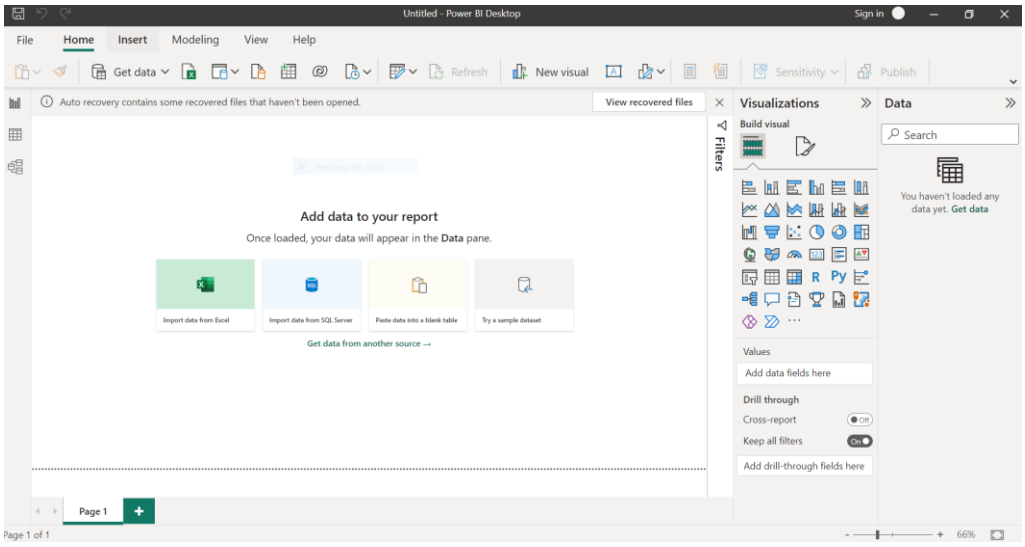
Manipulation and Visualisation in Power BI

From Wikepedia Power BI is an interactive data visualization software product developed by Microsoft with a primary focus on business intelligence.[1] It is part of the Microsoft Power Platform. Power BI is a collection of software services, apps, and connectors that work together to turn unrelated sources of data into coherent, visually immersive, and interactive insights. Data may be input by reading directly from a database, webpage, or structured files such as spreadsheets, CSV, XML, and JSON.

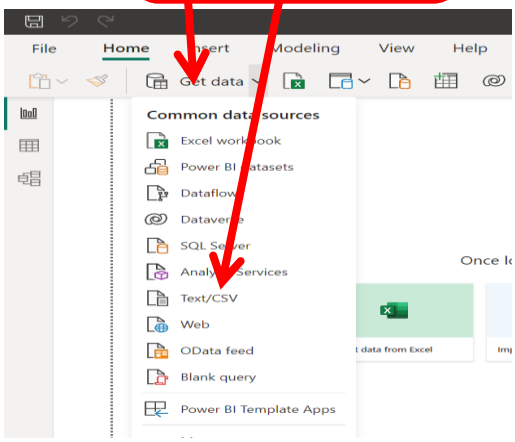
Microsoft Power BI

Developer(s)	Microsoft
Initial release	11 July 2011; 11 years ago
Stable release	March 2021 Update (2.91.383.0) / March 2021; 2 years ago
Operating system	Microsoft Windows
Type	Data visualization Business intelligence
License	Proprietary
Website	powerbi.microsoft.com

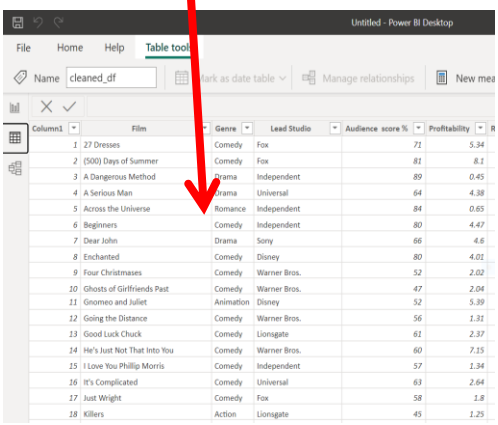
When we open Power BI, the software look like this:



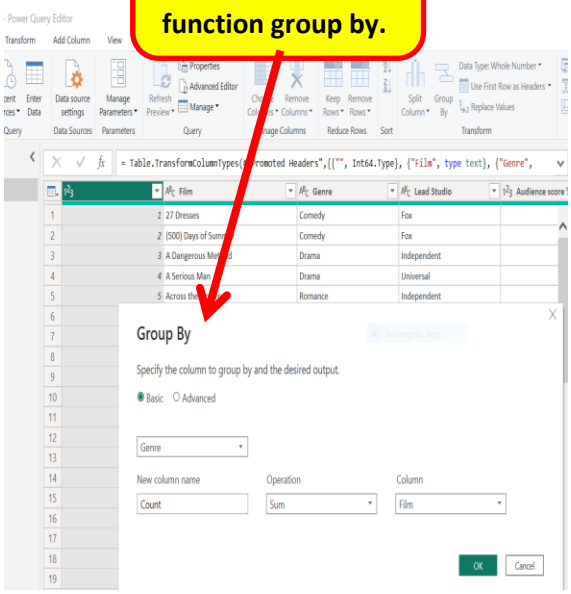
Importing data from a local disk to Powr BI as a text file.



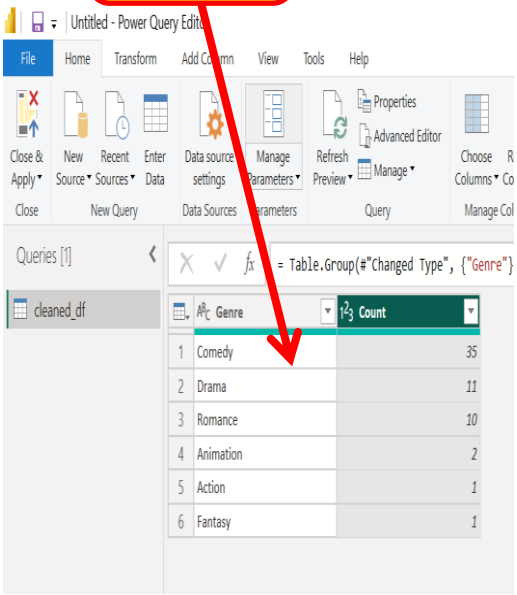
We can view this data as a table in Power BI.

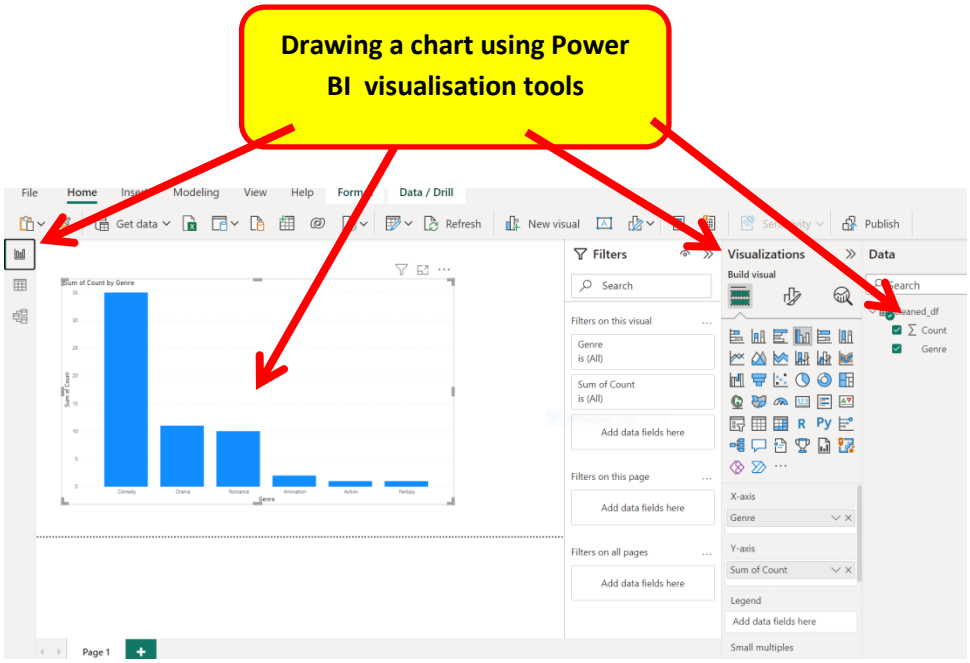


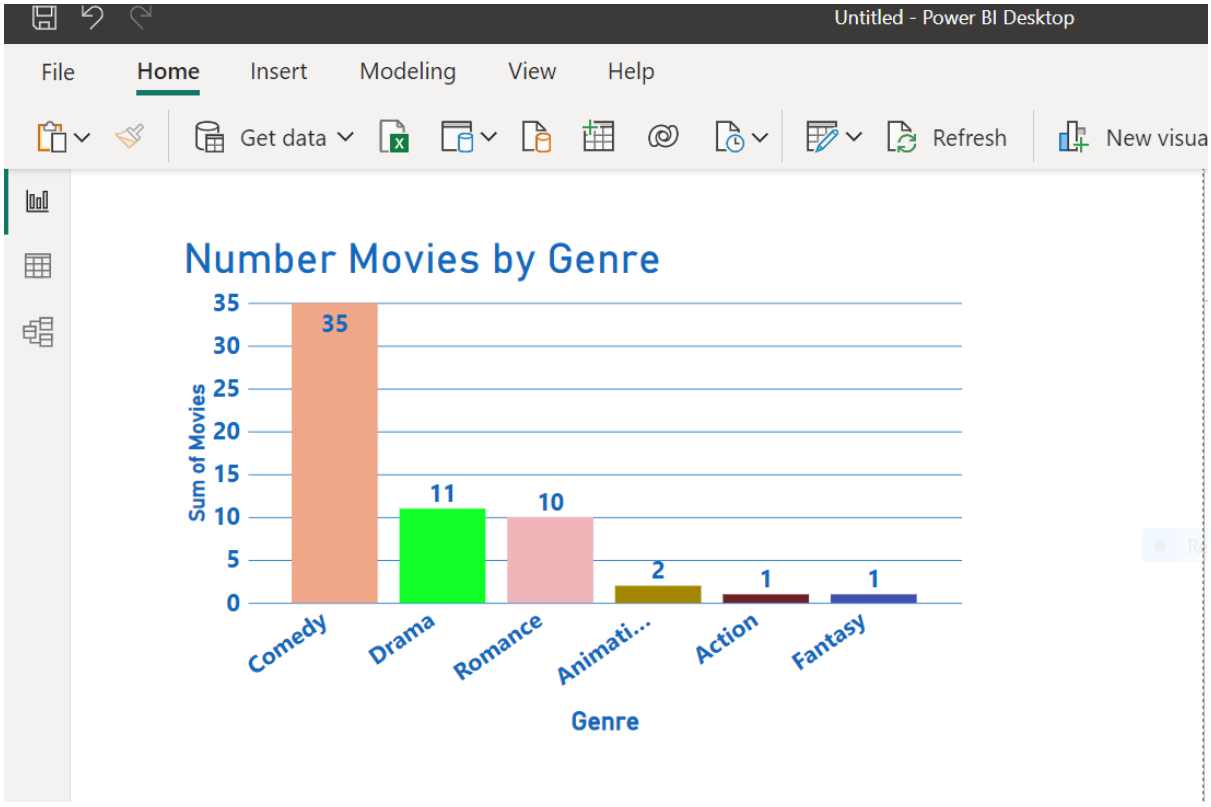
To group by data we use Power BI function group by.



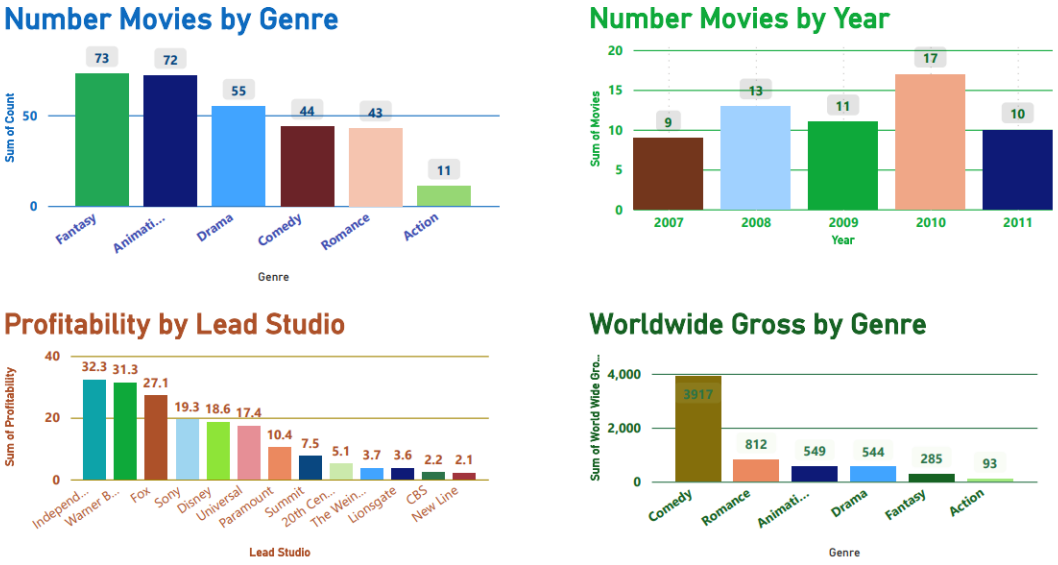
Grouped by data.




Drawing a chart using Power BI visualisation tools
File
Home
Modeling
View
Help
Data / Drill
Get data
Refresh
New visual
Publish
Filters
Search
Filters on this visual
Genre is (All)
Sum of Count is (All)
Add data fields here
Filters on this page
Filters on all pages
Visualizations
Build visual
Data
X-axis
Genre
Y-axis
Sum of Count
Legend
Small multiples
Page 1


Untitled - Power BI Desktop
File
Home
Insert
Modeling
View
Help
Get data
Refresh
New visua
Number Movies by Genre
Sum of Movies
35
30
25
20
15
10
5
0
35
11
10
2
1
1
Comedy
Drama
Romance
Animati...
Action
Fantasy
Genre

Beautiful dashboard



https://app.powerbi.com/groups/me/reports/9db75f58-2430-4a1f-84b8-df477a15d763/ReportSectionbc6680698ec7a709c444

Beautiful Dashboard