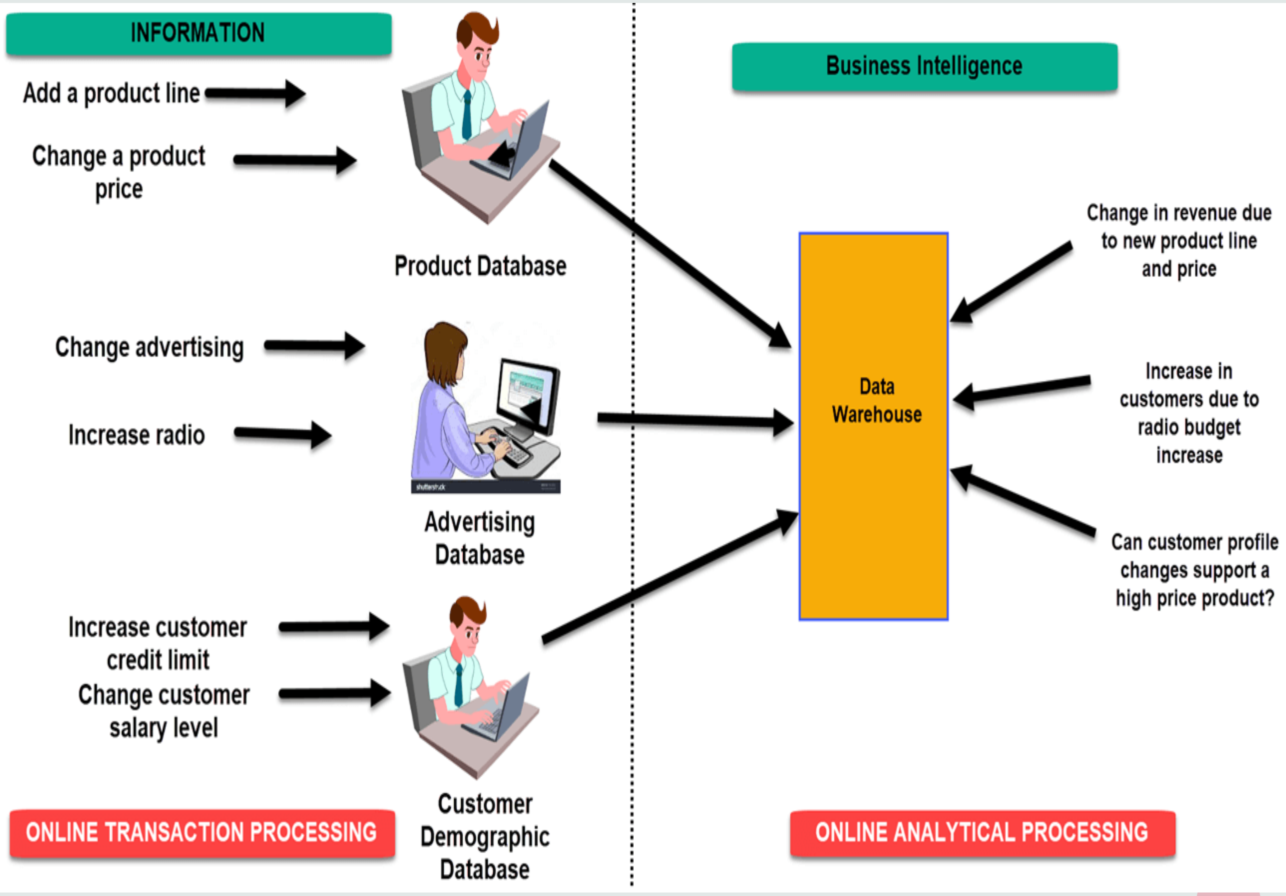
Welcome to my team 3

Table of content

- Introduction BI
- BI techniques and tools
- Dataset
- Design dashboard
- Influence of BI
- Legal Issues

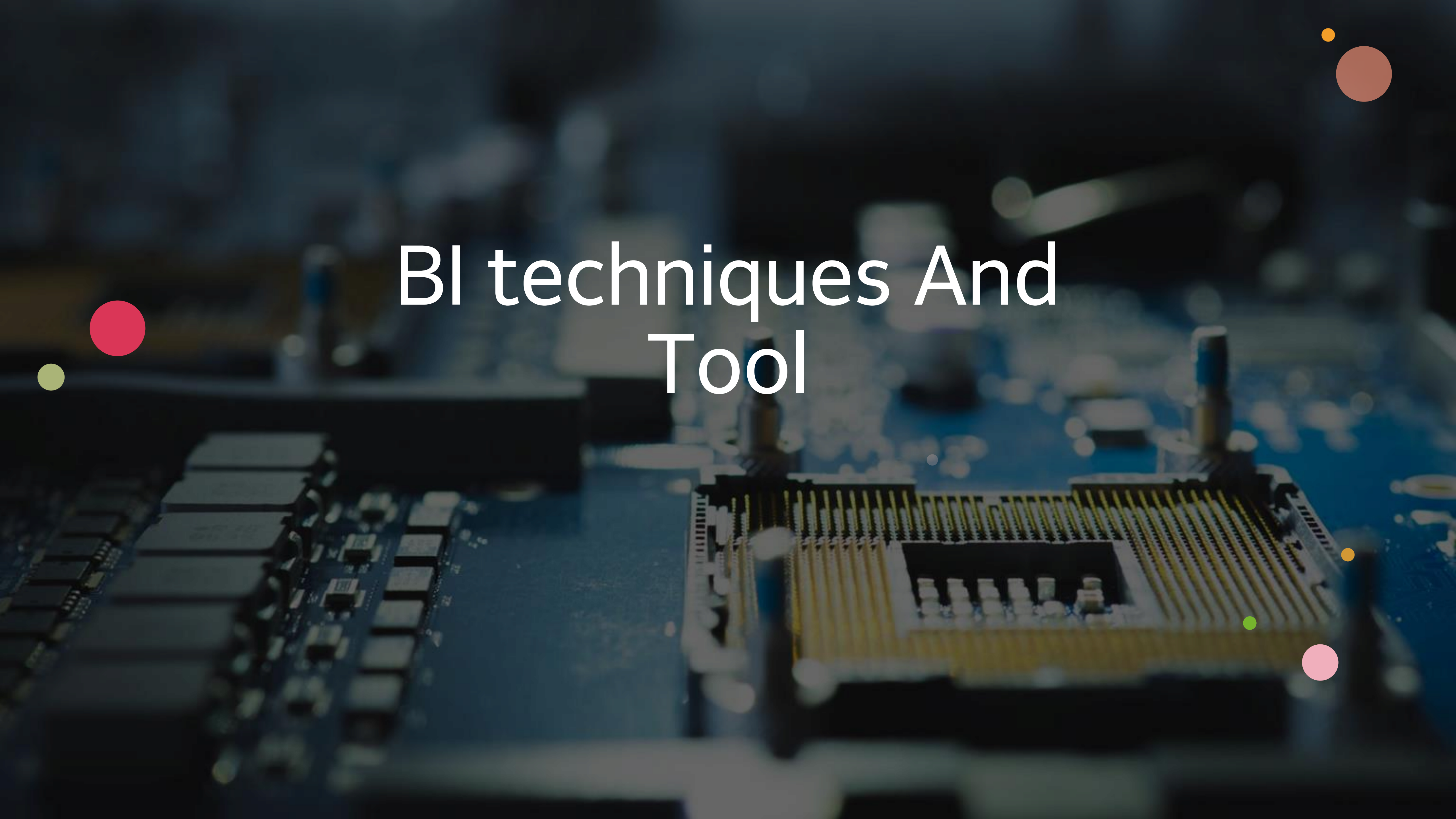
Introduction about BI

- Business Intelligence (BI) is the process of using a variety of tools, software and processes to analyze data, uncover insights and inform decisions within businesses. Through descriptive analytics and predictive models, it enables businesses to derive useful insights from their unprocessed data, assisting them in developing more strategic and informed business decisions.
- Better business decisions that help enterprises boost revenue, boost operational effectiveness, and gain a competitive edge over rival companies are the ultimate goal of BI projects. In order to accomplish that, BI combines analytics, reporting, and data management technologies with a number of different data management and analysis approaches.

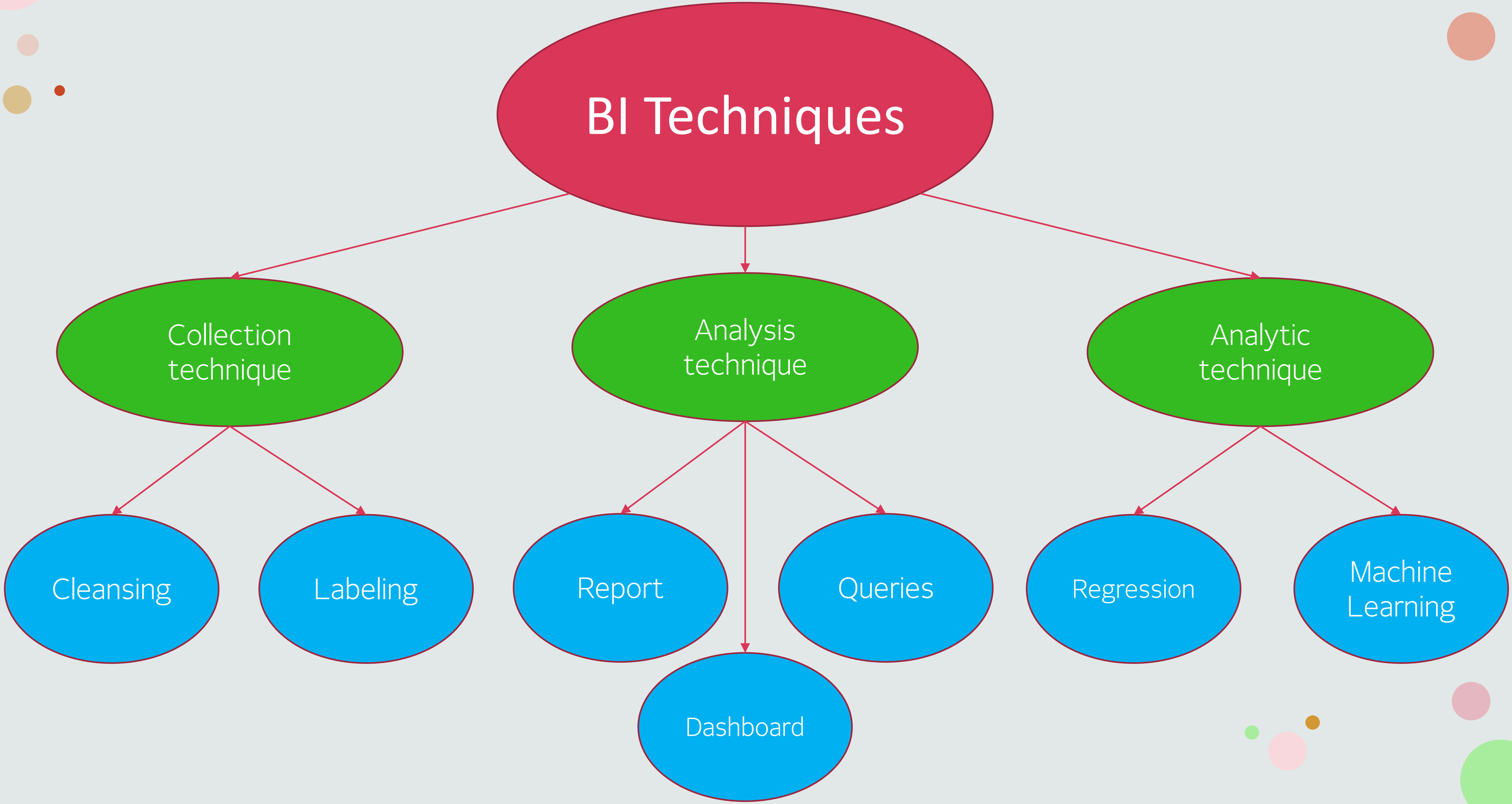

INFORMATION
Add a product line
Change a product price
Product Database
Change advertising
Increase radio
Advertising Database
Increase customer credit limit
Change customer salary level
Customer Demographic Database
ONLINE TRANSACTION PROCESSING
Business Intelligence
Data Warehouse
Change in revenue due to new product line and price
Increase in customers due to radio budget increase
Can customer profile changes support a high price product?
ONLINE ANALYTICAL PROCESSING

Example of Business Intelligence

- Netflix is a company that specializes in providing streaming TV and movie services. The company is aiming to expand the business model by expanding the market from domestic to international. However, Netflix's international expansion is facing many challenges in a world full of competition with other platforms such as Amazon, Microsoft, etc. That's why the solutions the company has used as not to target all markets at once. During this phase, Netflix will focus on understanding its internationalization strategy, improving partnerships with local businesses, and making investments in content geared toward local interests. management as well as data analytics and deep analytics technology. As a result, the company has been successful on this path and is being covered in 190 countries around the world with an incredible amount of revenue from home and abroad.



BI techniques And Tool



BI Techniques

Collection technique

Analysis technique

Analytic technique

Cleansing

Labeling

Report

Dashboard

Queries

Regression

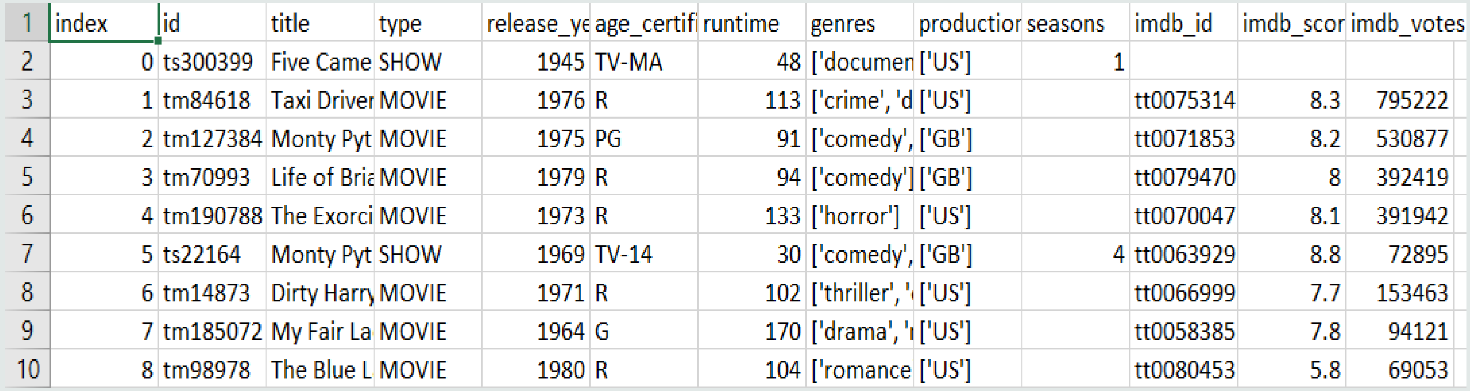
Machine Learning

Collection technique

- Cleansing
  - is the process of changing or removing incorrect, duplicate, corrupted or incomplete data within a database. If the data is incorrect, the algorithms and the results they produce are unreliable (even if they appear to be correct). The Data Cleaning process is not merely concerned with deleting data to increase capacity for new data. But also find a way to maximize the authenticity of the data set without having to delete the information.
  - The engine of the Data Cleaning service is to build standardized and unified data sets. It allows data analytics and business intelligence tools to easily access and perceive the correct data for each issue.

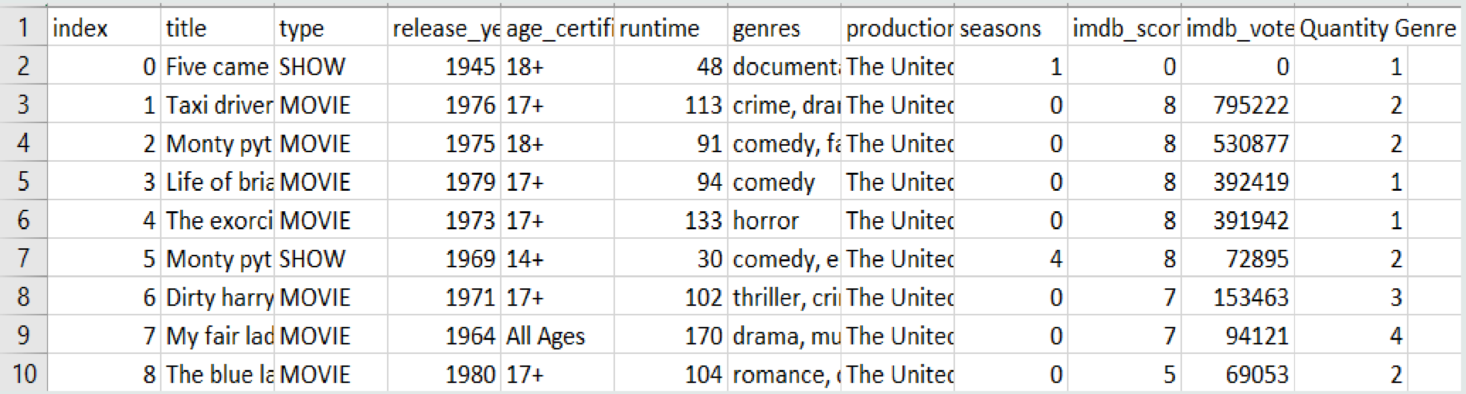
- Labeling
  - are methods used to organize, categorize and identify data in a database or spreadsheet. Examples of label techniques include labels, tags, categories, keywords, taxonomies and code numbers. These tools make it easier to sort and filter large volumes of data quickly and accurately.
  - This is an important aspect of data preparation, as it helps to normalize the data and make it more meaningful for analysis.

Example about Cleansing



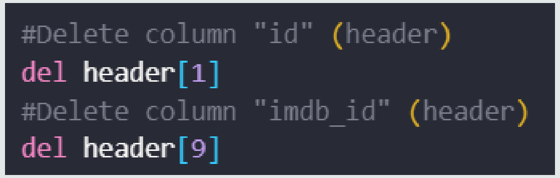
1	index	id	title	type	release_ye	age_certifi	runtime	genres	productior	seasons	imdb_id	imdb_scor	imdb_votes
2	0	ts300399	Five Came	SHOW	1945	TV-MA	48	['documen	['US']	1			
3	1	tm84618	Taxi Driver	MOVIE	1976	R	113	['crime', 'd	['US']		tt0075314	8.3	795222
4	2	tm127384	Monty Pyt	MOVIE	1975	PG	91	['comedy',	['GB']		tt0071853	8.2	530877
5	3	tm70993	Life of Bria	MOVIE	1979	R	94	['comedy']	['GB']		tt0079470	8	392419
6	4	tm190788	The Exorci	MOVIE	1973	R	133	['horror']	['US']		tt0070047	8.1	391942
7	5	ts22164	Monty Pyt	SHOW	1969	TV-14	30	['comedy',	['GB']	4	tt0063929	8.8	72895
8	6	tm14873	Dirty Harry	MOVIE	1971	R	102	['thriller', '	['US']		tt0066999	7.7	153463
9	7	tm185072	My Fair La	MOVIE	1964	G	170	['drama', 'ı	['US']		tt0058385	7.8	94121
10	8	tm98978	The Blue L	MOVIE	1980	R	104	['romance	['US']		tt0080453	5.8	69053

File raw_titles1 has 13 columns, however there are columns that are not relevant to the analysis like the id column and the imdb_id column.



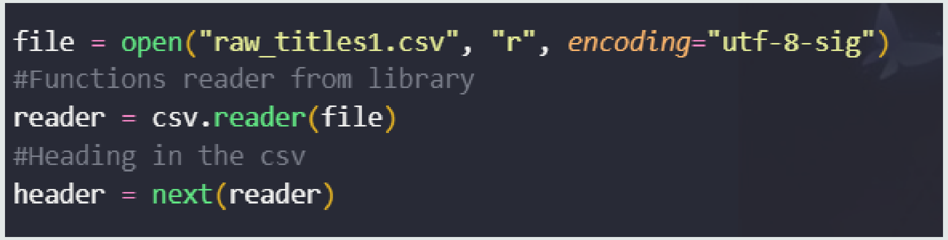
1	index	title	type	release_ye	age_certifi	runtime	genres	productior	seasons	imdb_scor	imdb_vote	Quantity Genre
2	0	Five came	SHOW	1945	18+	48	document	The Unitec	1	0	0	1
3	1	Taxi driver	MOVIE	1976	17+	113	crime, drai	The Unitec	0	8	795222	2
4	2	Monty pyt	MOVIE	1975	18+	91	comedy, fa	The Unitec	0	8	530877	2
5	3	Life of bria	MOVIE	1979	17+	94	comedy	The Unitec	0	8	392419	1
6	4	The exorci	MOVIE	1973	17+	133	horror	The Unitec	0	8	391942	1
7	5	Monty pyt	SHOW	1969	14+	30	comedy, e	The Unitec	4	8	72895	2
8	6	Dirty harry	MOVIE	1971	17+	102	thriller, cri	The Unitec	0	7	153463	3
9	7	My fair lad	MOVIE	1964	All Ages	170	drama, mu	The Unitec	0	7	94121	4
10	8	The blue la	MOVIE	1980	17+	104	romance, (	The Unitec	0	5	69053	2

Data after being clean => more suitable for analysis



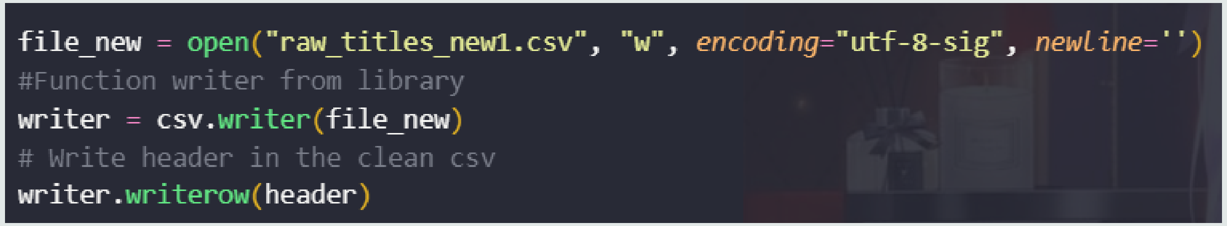
```
#Delete column "id" (header)
del header[1]
#Delete column "imdb_id" (header)
del header[9]
```

Example about Label



```
file = open("raw_titles1.csv", "r", encoding="utf-8-sig")
#Functions reader from library
reader = csv.reader(file)
#Heading in the csv
header = next(reader)
```

Input data for analysis: file raw_titles1.csv



```
file_new = open("raw_titles_new1.csv", "w", encoding="utf-8-sig", newline='')
#Function writer from library
writer = csv.writer(file_new)
# Write header in the clean csv
writer.writerow(header)
```

Output data after being analysis: file raw_title_new1.csv

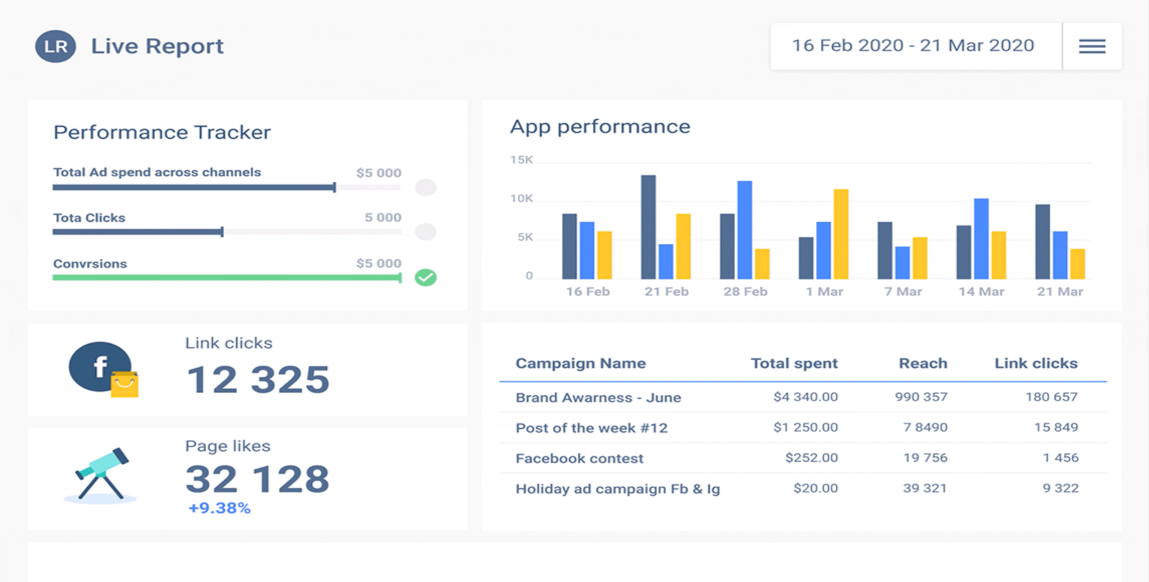
Analysis techniques

- Report
  - An analysis report is an essential business report displaying analysis results and feasible suggestions, and providing valuable information for decision-makers so that they can evaluate current operation status and then make well-informed decisions.
  - According to different analysis objects, purposes, and methods, analysis reports can be categorized into many types, such as comprehensive analysis reports, thematic analysis reports, routine work analysis reports, etc.

- Dashboard
  - is a tool used to perform a variety of tasks, organize, visualize, analyze, and track data. It is designed to connect and help extract important information from a variety of data sources, services and APIs with the support of artificial intelligence and machine learning to save time and automate processes. processes such as collecting, discovering, preparing, copying, and reporting.
  - The overall purpose of data analytics dashboards is to help data analysts, decision makers, and ordinary users understand their data more easily, gain insights, and make data-driven decisions better.

- Queries
  - is a process used in a database to determine how to further optimize queries for performance.
  - is an important aspect of query processing because it helps to improve the overall performance of query processing, which will speed up many aspects and database functionality. To do this, the query optimizer analyzes a particular query statement and generates both local and remote access plans for use on the query fragment, based on the resource cost of each package.

Example about Analysis Technique



LR Live Report

16 Feb 2020 - 21 Mar 2020

Performance Tracker

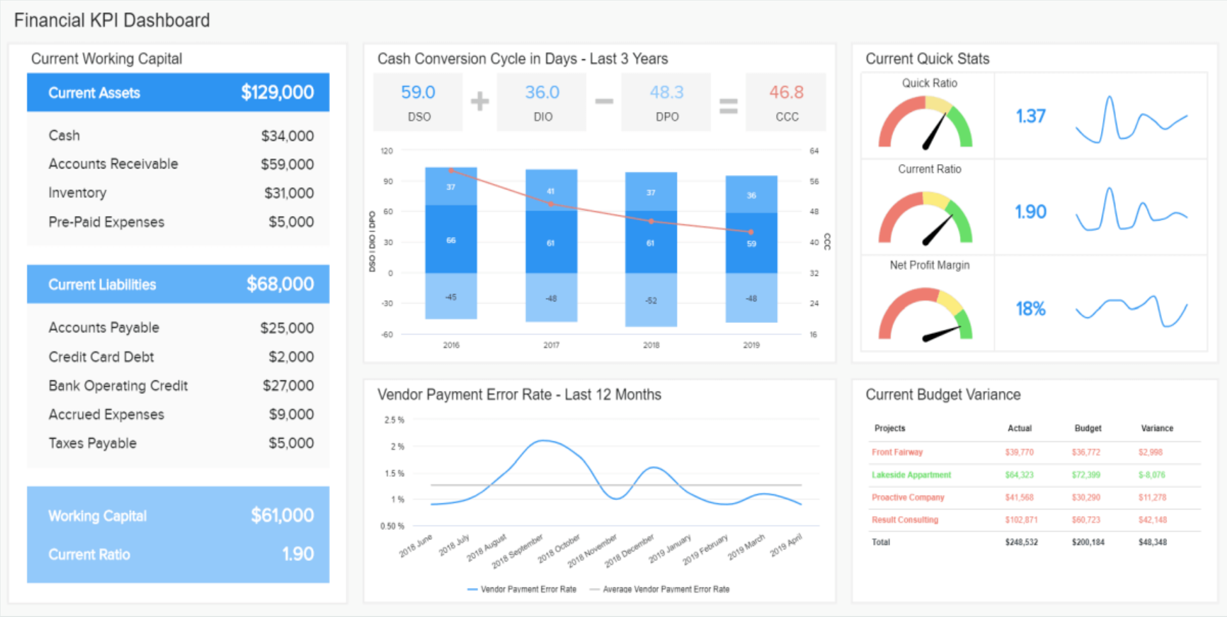
Total Ad spend across channels	$5 000
Tota Clicks	5 000
Convrsions	$5 000

App performance

Link clicks
12 325

Page likes
32 128
+9.38%

Campaign Name	Total spent	Reach	Link clicks
Brand Awarness - June	$4 340.00	990 357	180 657
Post of the week #12	$1 250.00	7 8490	15 849
Facebook contest	$252.00	19 756	1 456
Holiday ad campaign Fb & Ig	$20.00	39 321	9 322



Financial KPI Dashboard

Current Working Capital

Current Assets	$129,000
Cash	$34,000
Accounts Receivable	$59,000
Inventory	$31,000
Pre-Paid Expenses	$5,000
Current Liabilities	$68,000
Accounts Payable	$25,000
Credit Card Debt	$2,000
Bank Operating Credit	$27,000
Accrued Expenses	$9,000
Taxes Payable	$5,000
Working Capital	$61,000
Current Ratio	1.90

Cash Conversion Cycle in Days - Last 3 Years

59.0 DSO + 36.0 DIO − 48.3 DPO = 46.8 CCC

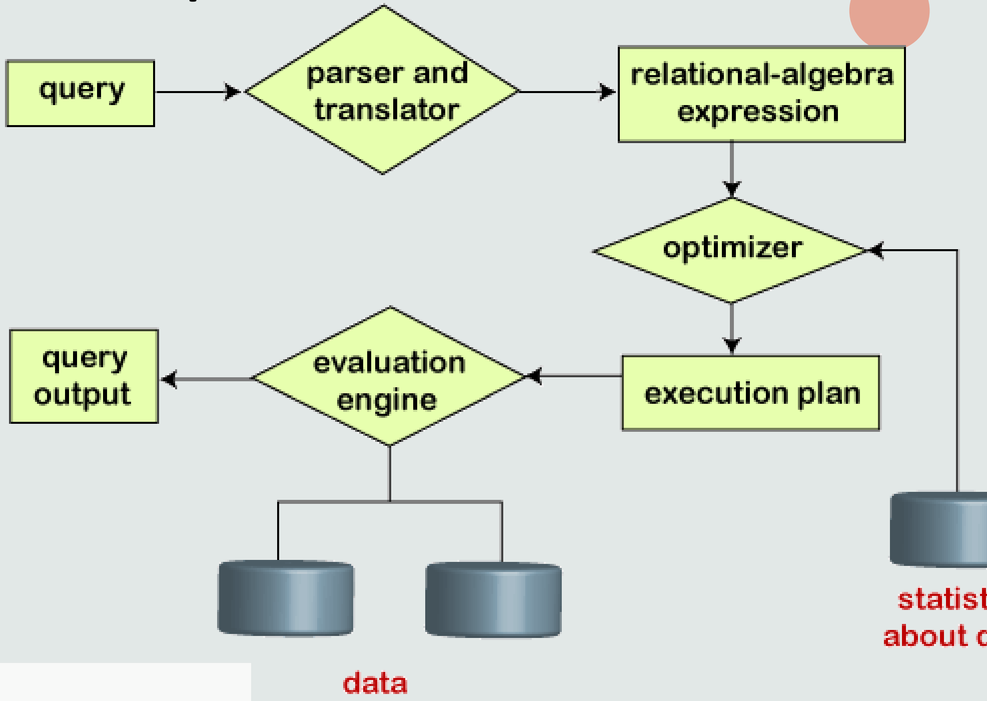
Current Quick Stats

Quick Ratio	1.37
Current Ratio	1.90
Net Profit Margin	18%

Vendor Payment Error Rate - Last 12 Months

Current Budget Variance

Projects	Actual	Budget	Variance
Front Fairway	$39,770	$36,772	$2,998
Lakeside Appartment	$64,323	$72,399	$-8,076
Proactive Company	$41,568	$30,290	$11,278
Result Consulting	$102,871	$60,723	$42,148
Total	$248,532	$200,184	$48,348



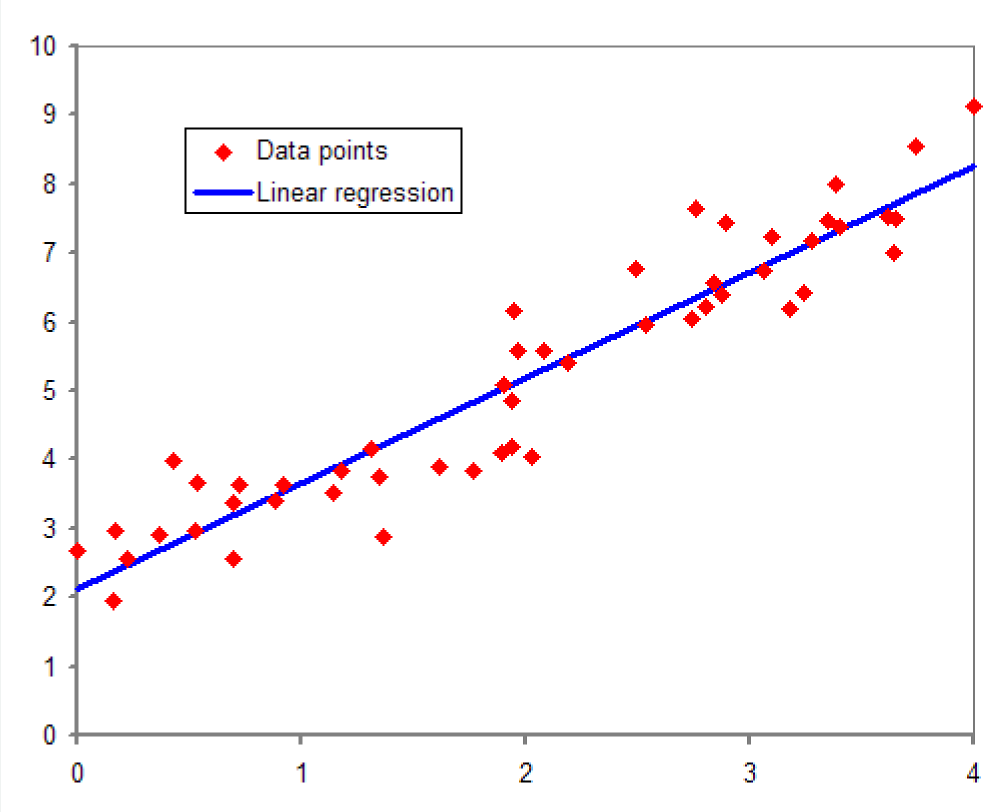
Steps in query processing

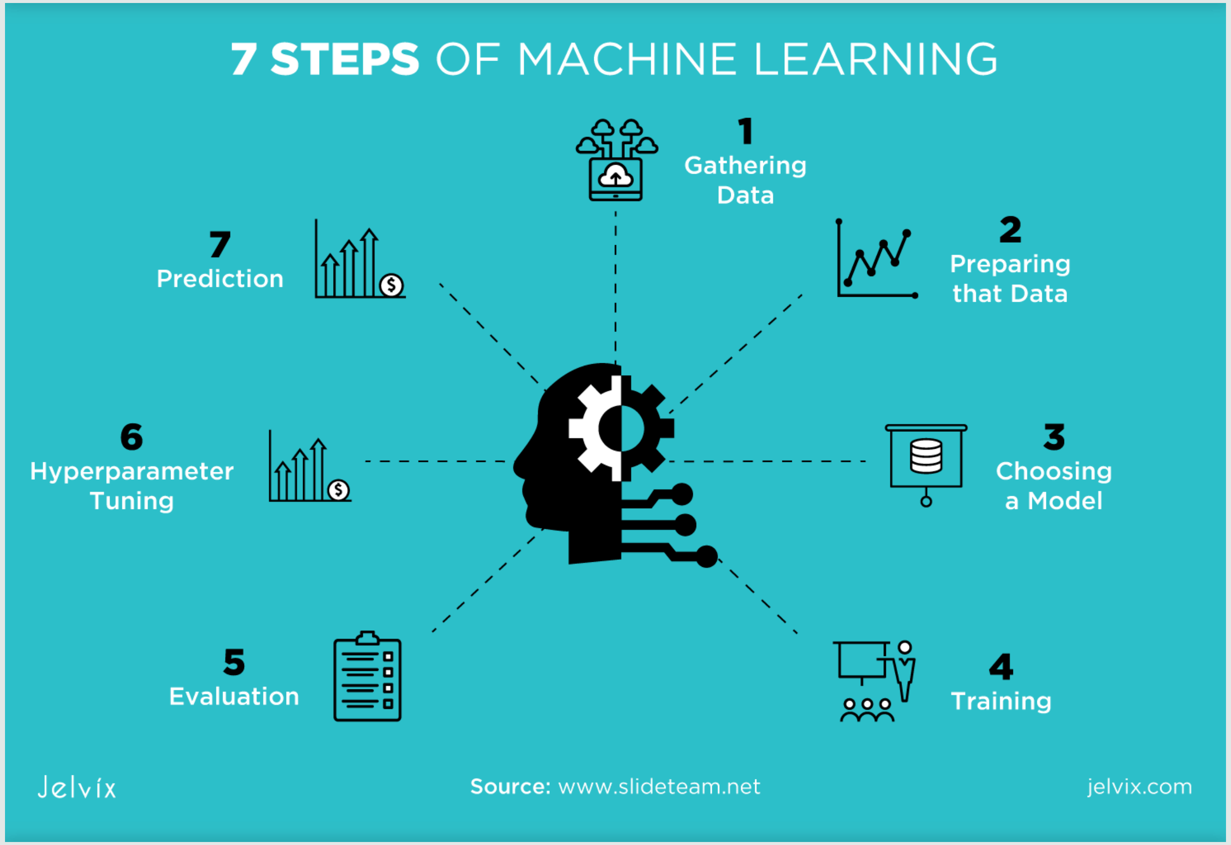
Analytic techniques

- Regression
  - is a statistical technique used to evaluate the relationship between two or more independent variables. Organizations use regression analysis to understand the significance of their data points and use analytical techniques to make better decisions.
  - This type of analysis is used by business analysts and data professionals to remove unwanted variables and select those that are statistically significant.

- Machine learning
  - Machine learning is a subset of AI with the narrow purpose of learning from information (data) as far as possible without explicit programming. ML utilizes numerical and statistical approaches to encode learning in models. Machine learning in data analytics is the new way of designing algorithms that learn on their own from data and adapt with minimal human intervention.
  - They are used by traders and investors to be able to study a field at a more detailed, complete, and maybe interested size, change management, and culture of curiosity.

Example about Analysis Technique





7 STEPS OF MACHINE LEARNING

1 Gathering Data

2 Preparing that Data

3 Choosing a Model

4 Training

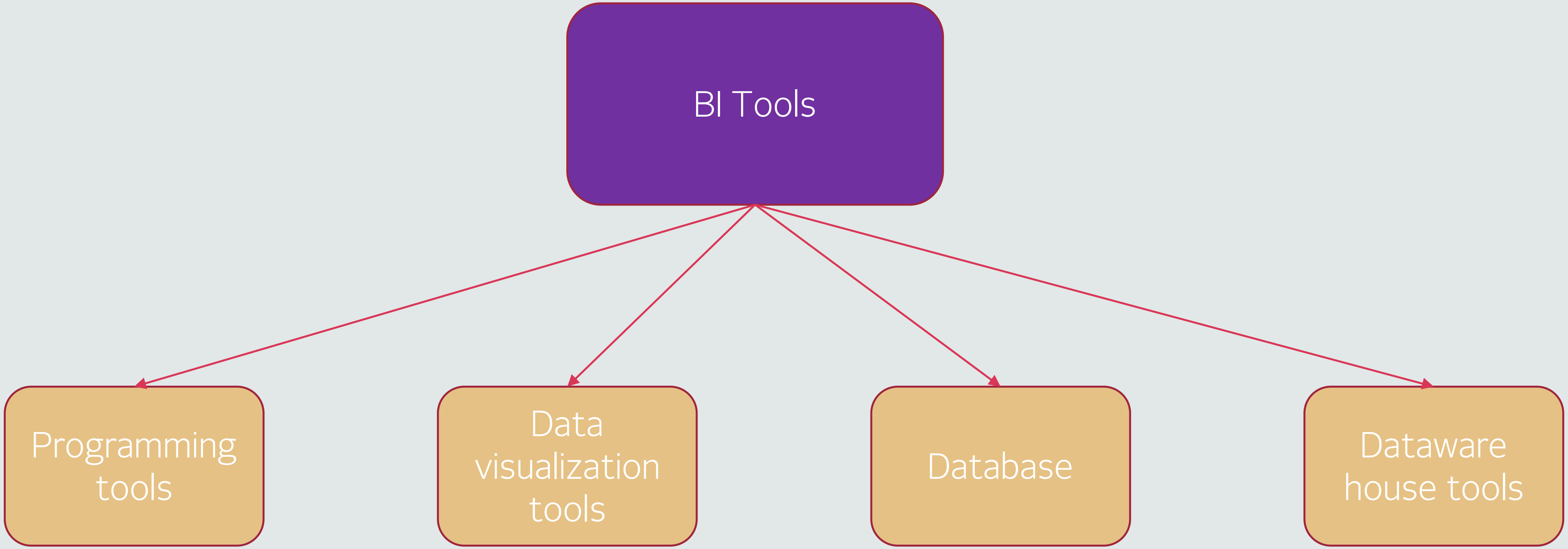
5 Evaluation

6 Hyperparameter Tuning

7 Prediction

Source: www.slideteam.net

BI Tools



BI Tools

Programming tools

Data visualization tools

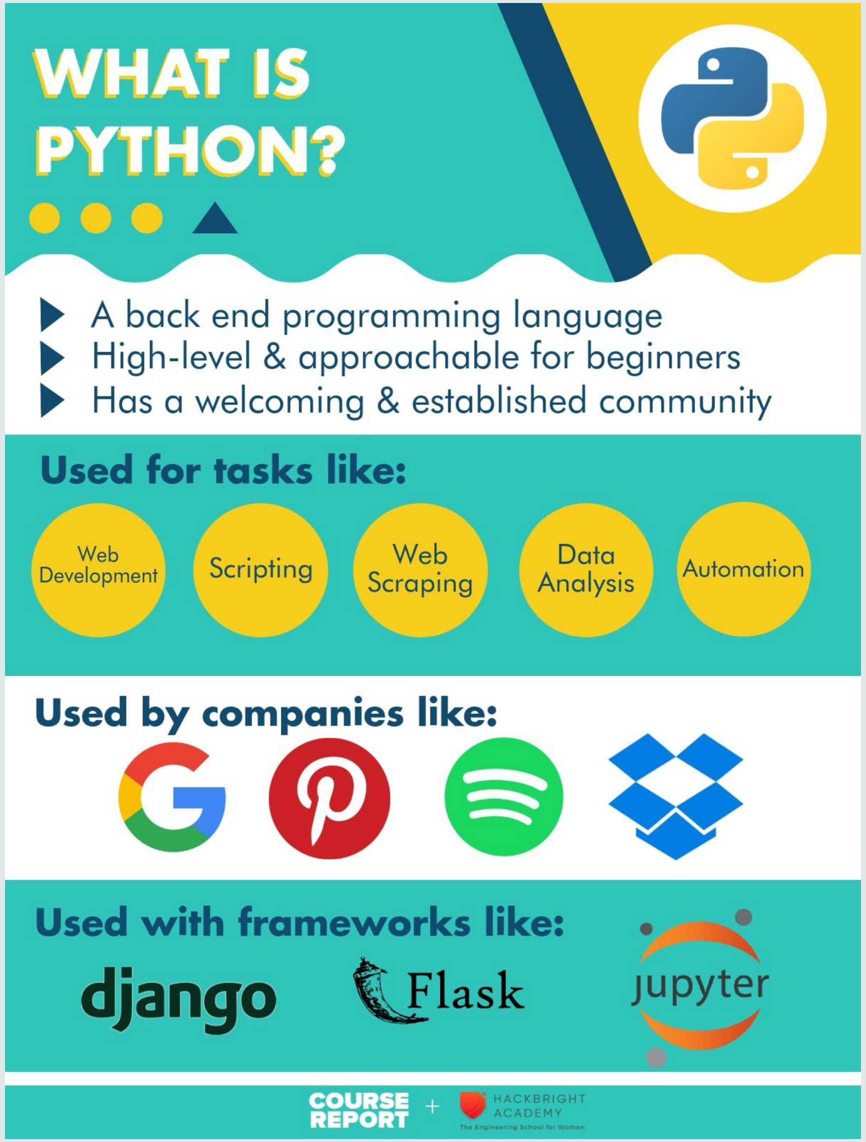
Database

Dataware house tools

Programming tools

Python

- is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together.
- simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms and can be freely distributed.



WHAT IS PYTHON?

- A back end programming language
- High-level & approachable for beginners
- Has a welcoming & established community

Used for tasks like:

Web Development · Scripting · Web Scraping · Data Analysis · Automation

Used by companies like:

Used with frameworks like:

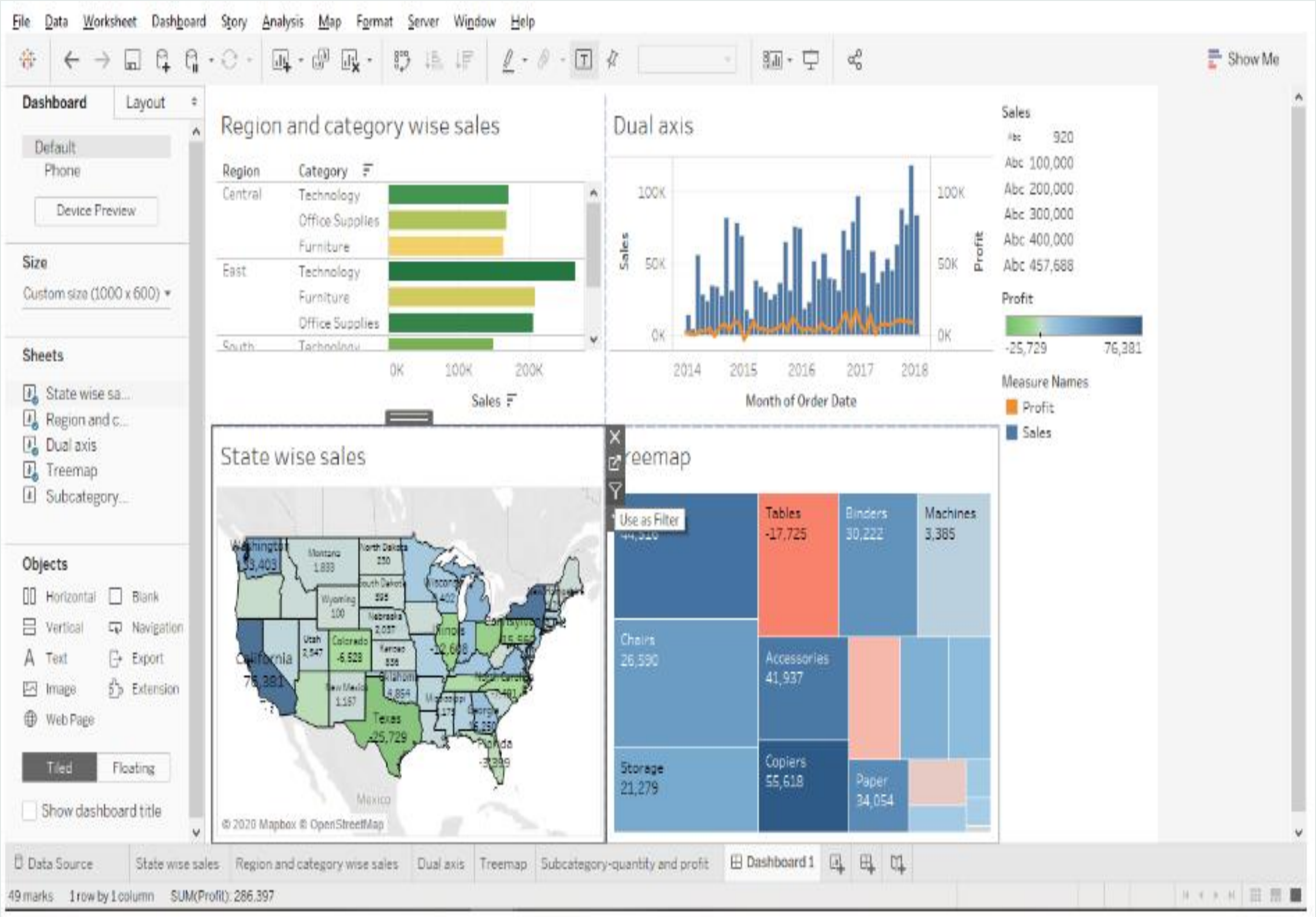
django · Flask · jupyter

COURSE REPORT + HACKBRIGHT ACADEMY The Engineering School for Women

Visualization tools

Tableau

- is a powerful and fastest growing data visualization tool used in the Business Intelligence Industry. It helps in simplifying raw data in a very easily understandable format.
- helps create the data that can be understood by professionals at any level in an organization. It also allows non-technical users to create customized dashboards.



Database tools

Database

- is an organized collection of structured information, or data, typically stored electronically in a computer system. A database is usually controlled by a database management system (DBMS). Together, the data and the DBMS, along with the applications that are associated with them, are referred to as a database system, often shortened to just database.
- Data within the most common types of databases in operation today is typically modeled in rows and columns in a series of tables to make processing and data querying efficient. The data can then be easily accessed, managed, modified, updated, controlled, and organized.

Database tool: Mongodb

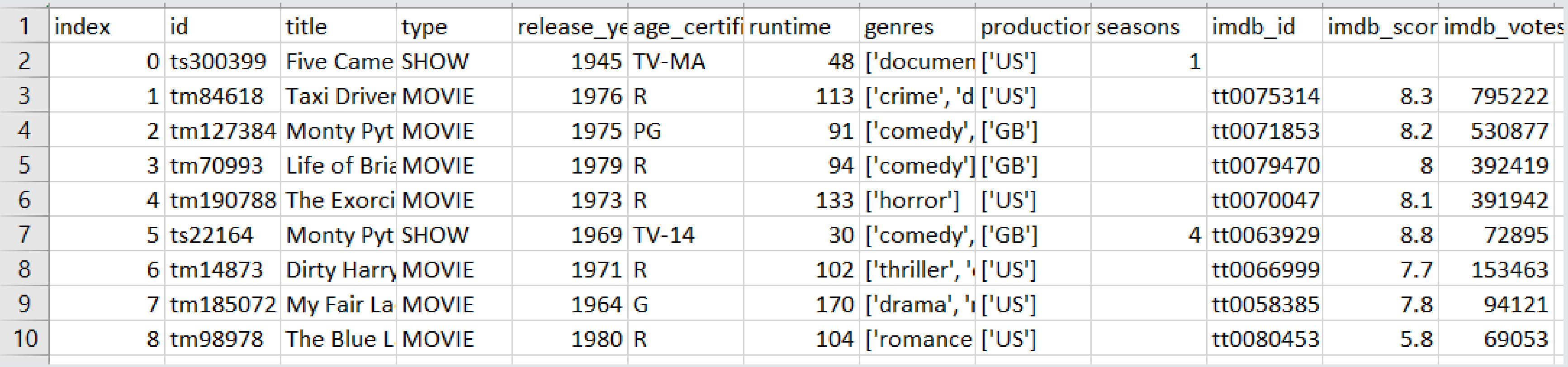
Dataware house

Data warehouse is a type of data management system that is designed to enable and support business intelligence (BI) activities, especially analytics. Data warehouses are solely intended to perform queries and analysis and often contain large amounts of historical data.

The data within a data warehouse is usually derived from a wide range of sources such as application log files and transaction applications.

A data warehouse centralizes and consolidates large amounts of data from multiple sources. Its analytical capabilities allow organizations to derive valuable business insights from their data to improve decision-making. Over time, it builds a historical record that can be invaluable to data scientists and business analysts.

Dataset



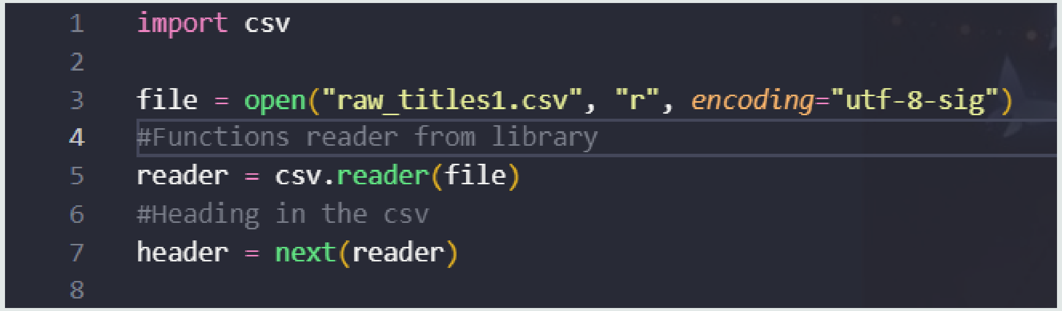
	index	id	title	type	release_ye	age_certifi	runtime	genres	productior	seasons	imdb_id	imdb_scor	imdb_votes
1	index	id	title	type	release_ye	age_certifi	runtime	genres	productior	seasons	imdb_id	imdb_scor	imdb_votes
2	0	ts300399	Five Came	SHOW	1945	TV-MA	48	['documen	['US']	1			
3	1	tm84618	Taxi Driver	MOVIE	1976	R	113	['crime', 'd	['US']		tt0075314	8.3	795222
4	2	tm127384	Monty Pyt	MOVIE	1975	PG	91	['comedy',	['GB']		tt0071853	8.2	530877
5	3	tm70993	Life of Bria	MOVIE	1979	R	94	['comedy']	['GB']		tt0079470	8	392419
6	4	tm190788	The Exorci	MOVIE	1973	R	133	['horror']	['US']		tt0070047	8.1	391942
7	5	ts22164	Monty Pyt	SHOW	1969	TV-14	30	['comedy',	['GB']	4	tt0063929	8.8	72895
8	6	tm14873	Dirty Harry	MOVIE	1971	R	102	['thriller', '	['US']		tt0066999	7.7	153463
9	7	tm185072	My Fair La	MOVIE	1964	G	170	['drama', '	['US']		tt0058385	7.8	94121
10	8	tm98978	The Blue L	MOVIE	1980	R	104	['romance	['US']		tt0080453	5.8	69053

- This dataset shows the votes and ratings for the movies listed above.
- The table above contains 13 columns and 5807 rows.
- The columns contain information such as: index, id, title, type, release_year, age_certificate, runtime, genres, production_country, seasons, imdb_id, imdb_score, imdb_votes
- In this dataset, it has not been scientifically optimized, so we will perform data cleaning steps such as: delete the id column, imdb_id, the type column has the first word in uppercase, age_certification changes to the number according to age, genre changes to "crime, drama", country changes to country-specific names, seasons to integers, votes to integers, add column number of genre.

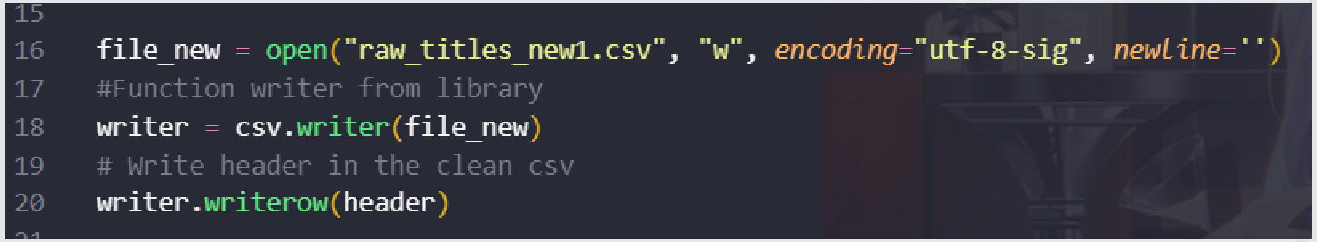
Library used for pre-processing

In this analysis project we will use python to solve the above requirements. At the beginning of the processing we will use the csv library because:

- The data is stored in the form of tables, databases.
- Support read and write csv file



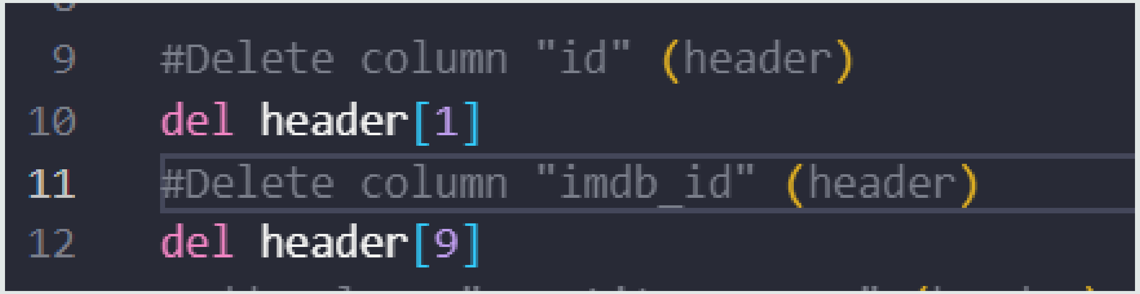
```python
import csv

file = open("raw_titles1.csv", "r", encoding="utf-8-sig")
#Functions reader from library
reader = csv.reader(file)
#Heading in the csv
header = next(reader)

```



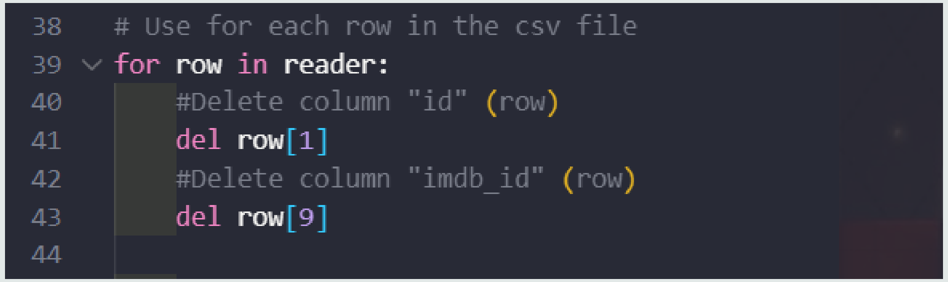
```python

file_new = open("raw_titles_new1.csv", "w", encoding="utf-8-sig", newline='')
#Function writer from library
writer = csv.writer(file_new)
# Write header in the clean csv
writer.writerow(header)
```

Pre-processing

Use del function in list to remove header and an element in row

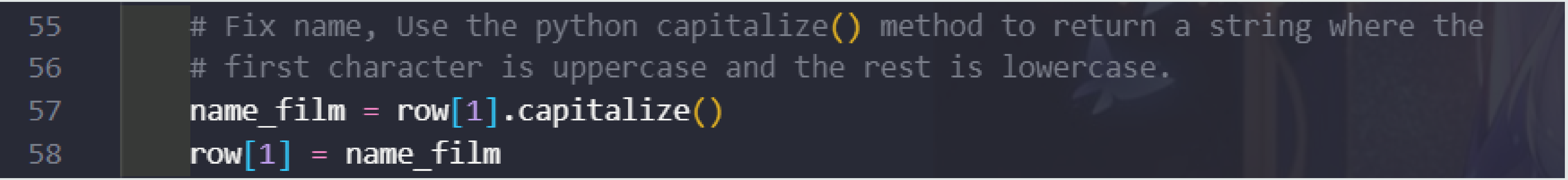


```python
#Delete column "id" (header)
del header[1]
#Delete column "imdb_id" (header)
del header[9]
```



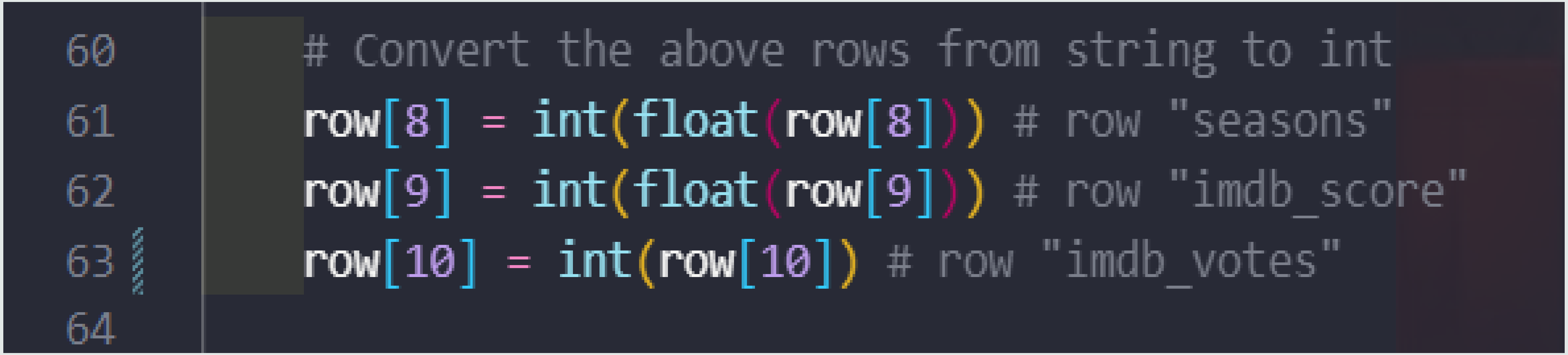
```python
# Use for each row in the csv file
for row in reader:
    #Delete column "id" (row)
    del row[1]
    #Delete column "imdb_id" (row)
    del row[9]

```

Fix name



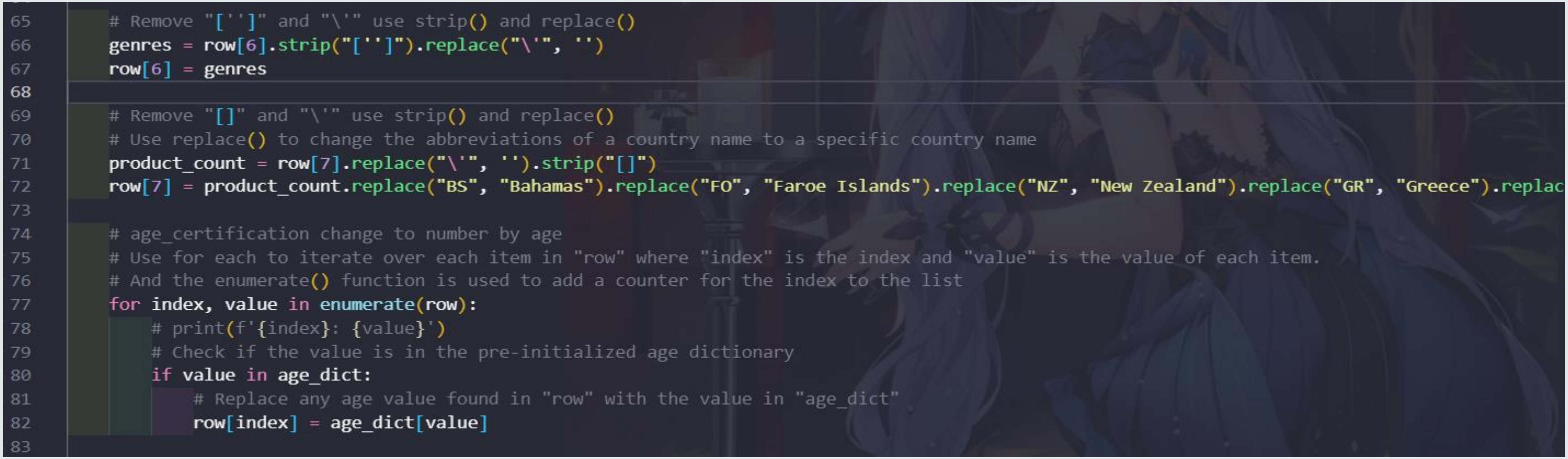
```
    # Fix name, Use the python capitalize() method to return a string where the
    # first character is uppercase and the rest is lowercase.
    name_film = row[1].capitalize()
    row[1] = name_film
```

Convert string to int



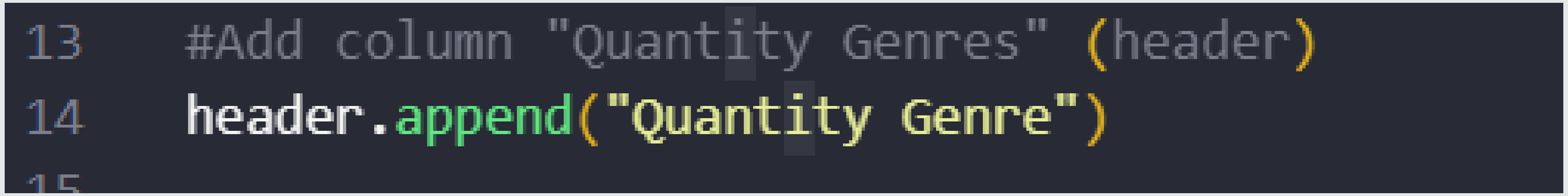
```python
# Convert the above rows from string to int
row[8] = int(float(row[8])) # row "seasons"
row[9] = int(float(row[9])) # row "imdb_score"
row[10] = int(row[10]) # row "imdb_votes"
```

Remove bracket, apostrophe, repair age_certificate and countries

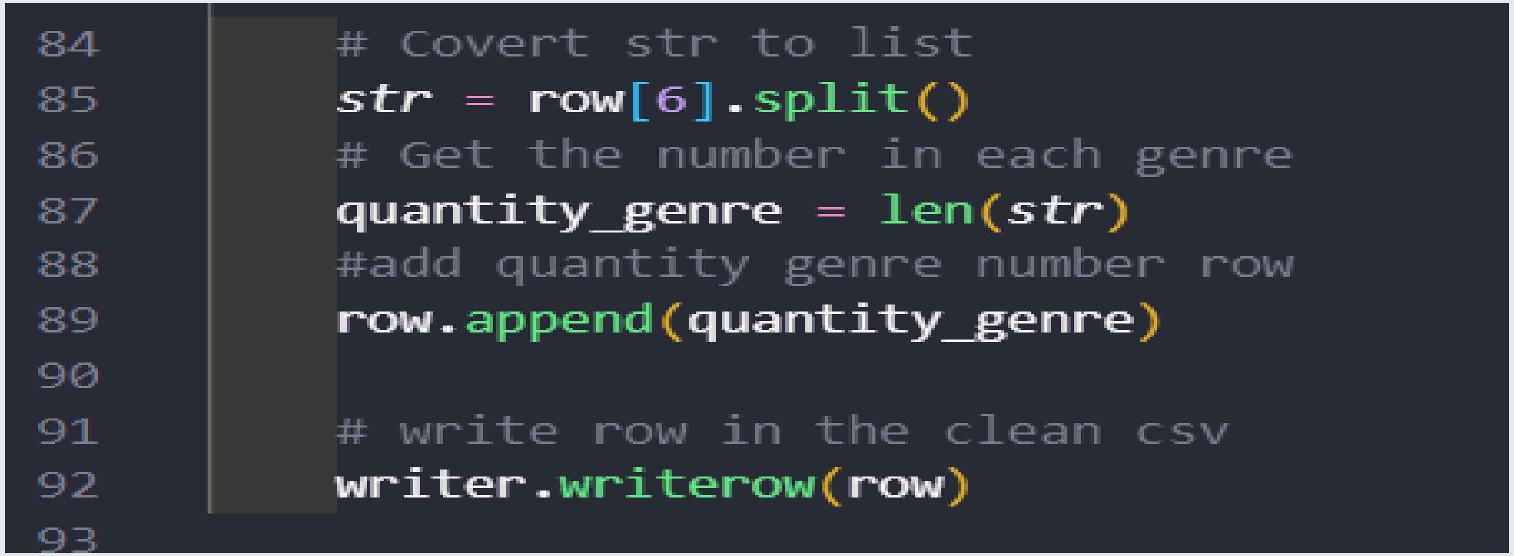


```
# Remove "['']" and "\'" use strip() and replace()
genres = row[6].strip("['']").replace("\'", '')
row[6] = genres

# Remove "[]" and "\'" use strip() and replace()
# Use replace() to change the abbreviations of a country name to a specific country name
product_count = row[7].replace("\'", '').strip("[]")
row[7] = product_count.replace("BS", "Bahamas").replace("FO", "Faroe Islands").replace("NZ", "New Zealand").replace("GR", "Greece").replac

# age_certification change to number by age
# Use for each to iterate over each item in "row" where "index" is the index and "value" is the value of each item.
# And the enumerate() function is used to add a counter for the index to the list
for index, value in enumerate(row):
    # print(f'{index}: {value}')
    # Check if the value is in the pre-initialized age dictionary
    if value in age_dict:
        # Replace any age value found in "row" with the value in "age_dict"
        row[index] = age_dict[value]

```

Add column “Quantity Genres” (header)

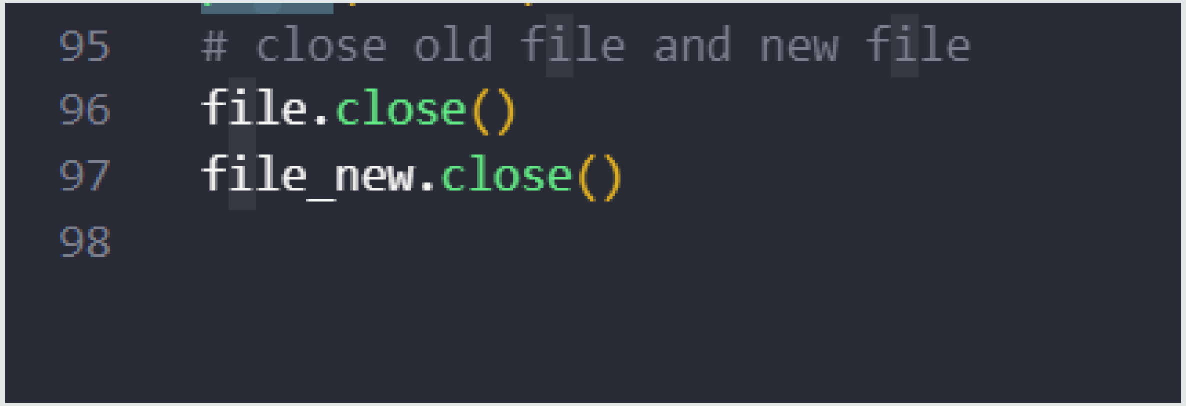


```python
#Add column "Quantity Genres" (header)
header.append("Quantity Genre")
```



```python
        # Covert str to list
        str = row[6].split()
        # Get the number in each genre
        quantity_genre = len(str)
        #add quantity genre number row
        row.append(quantity_genre)

        # write row in the clean csv
        writer.writerow(row)
```

Close file

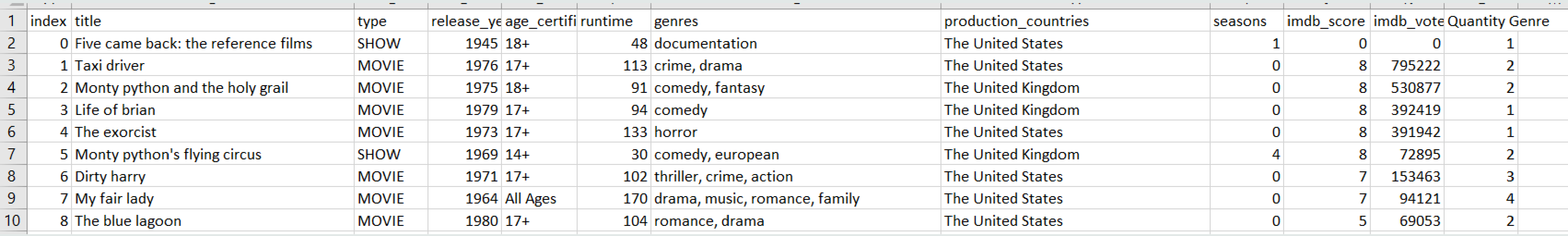


```
# close old file and new file
file.close()
file_new.close()

```

Use the close() function in the csv library to end the program, close the file being processed and free up memory

Clean dataset

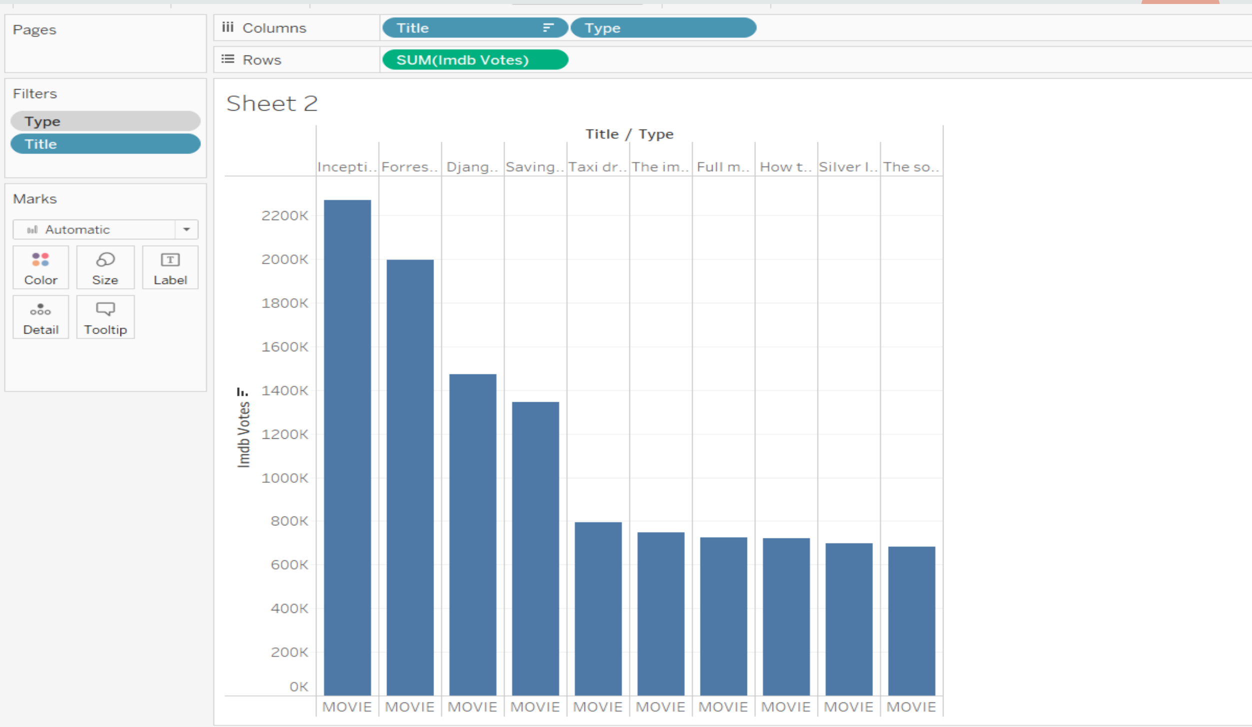
After cleaning the data, they are saved under a new file named "raw_titles_new1.csv"



1	index	title	type	release_ye	age_certifi	runtime	genres	production_countries	seasons	imdb_score	imdb_vote	Quantity Genre
2	0	Five came back: the reference films	SHOW	1945	18+	48	documentation	The United States	1	0	0	1
3	1	Taxi driver	MOVIE	1976	17+	113	crime, drama	The United States	0	8	795222	2
4	2	Monty python and the holy grail	MOVIE	1975	18+	91	comedy, fantasy	The United Kingdom	0	8	530877	2
5	3	Life of brian	MOVIE	1979	17+	94	comedy	The United Kingdom	0	8	392419	1
6	4	The exorcist	MOVIE	1973	17+	133	horror	The United States	0	8	391942	1
7	5	Monty python's flying circus	SHOW	1969	14+	30	comedy, european	The United Kingdom	4	8	72895	2
8	6	Dirty harry	MOVIE	1971	17+	102	thriller, crime, action	The United States	0	7	153463	3
9	7	My fair lady	MOVIE	1964	All Ages	170	drama, music, romance, family	The United States	0	7	94121	4
10	8	The blue lagoon	MOVIE	1980	17+	104	romance, drama	The United States	0	5	69053	2

Top 10 titles of each type with the highest votes.

In this section we are analyzing titles with types selected by users. All of those titles are ranked in the top 10 of the best titles.



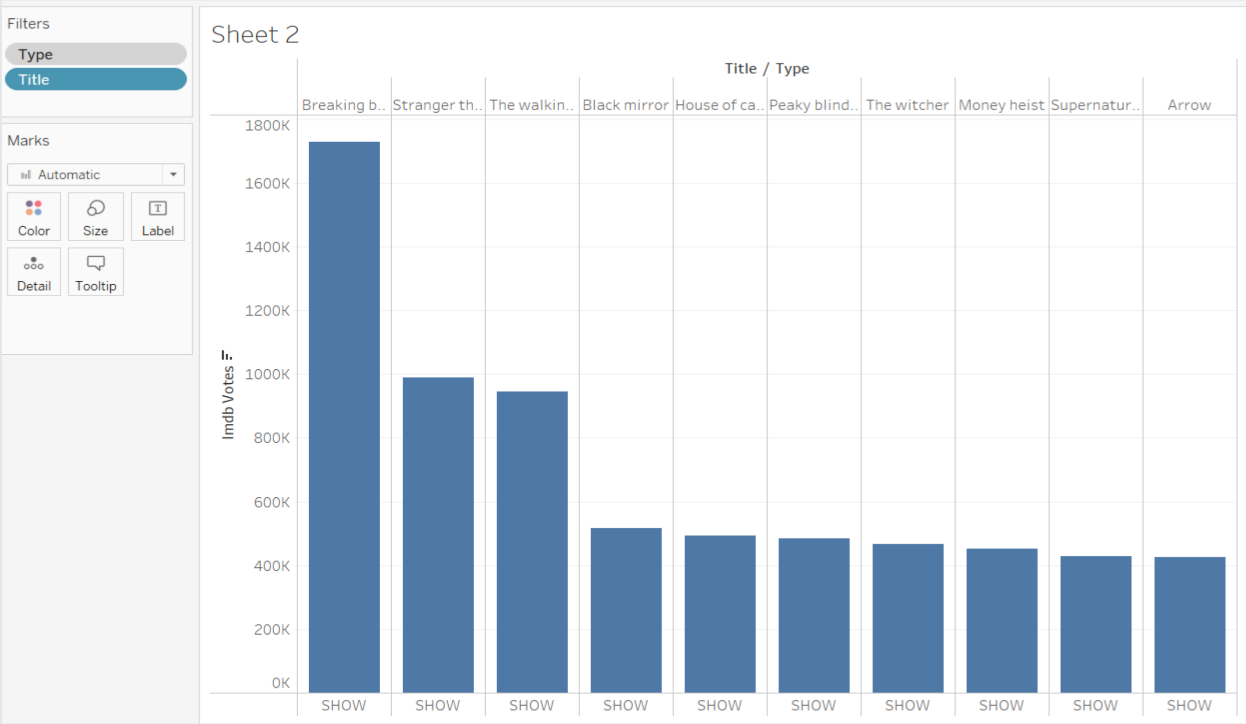
Pages

Columns: Title, Type

Rows: SUM(Imdb Votes)

Filters: Type, Title

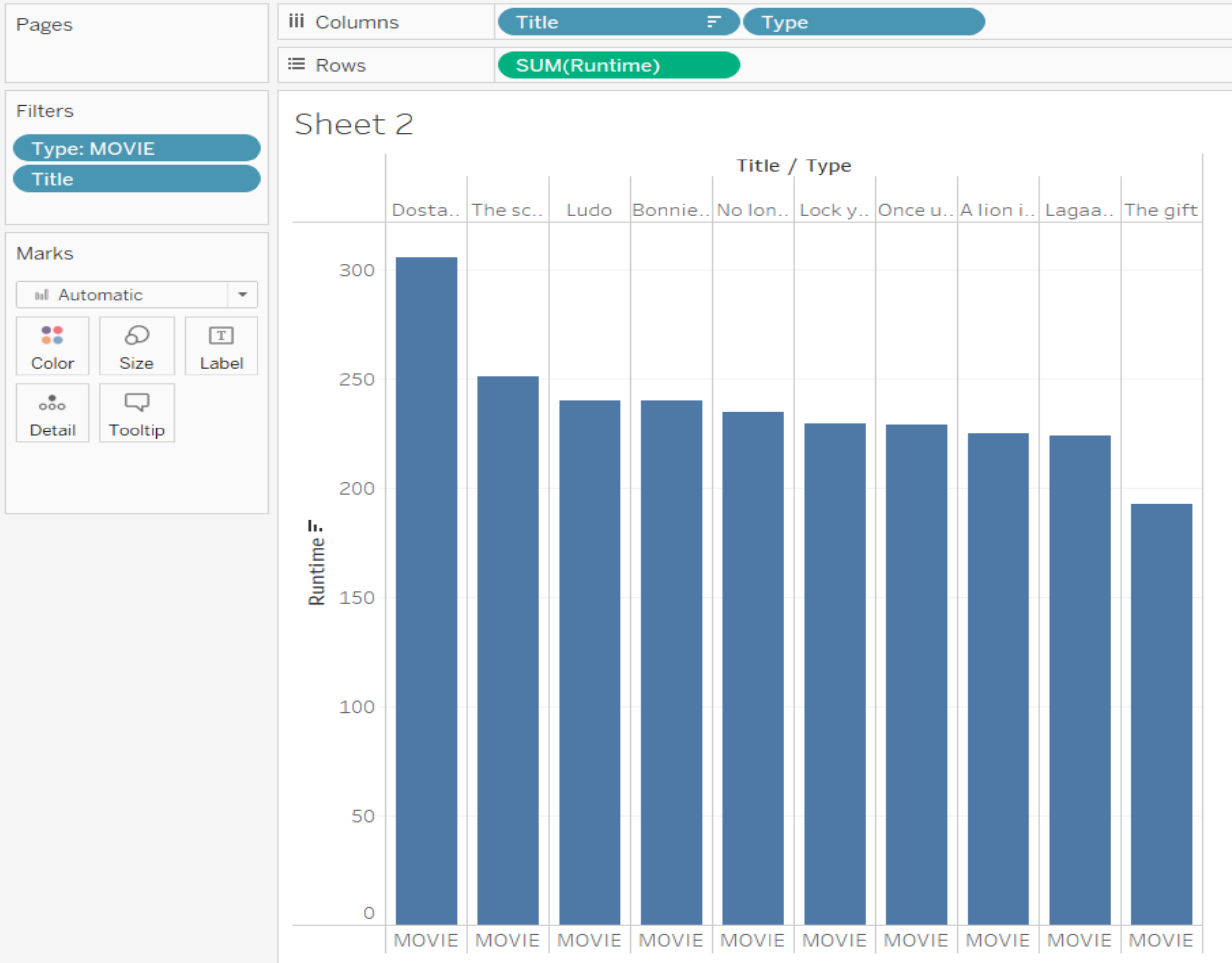
Marks: Automatic



Filters: Type, Title

Marks: Automatic

The most runtime movie and show duration.



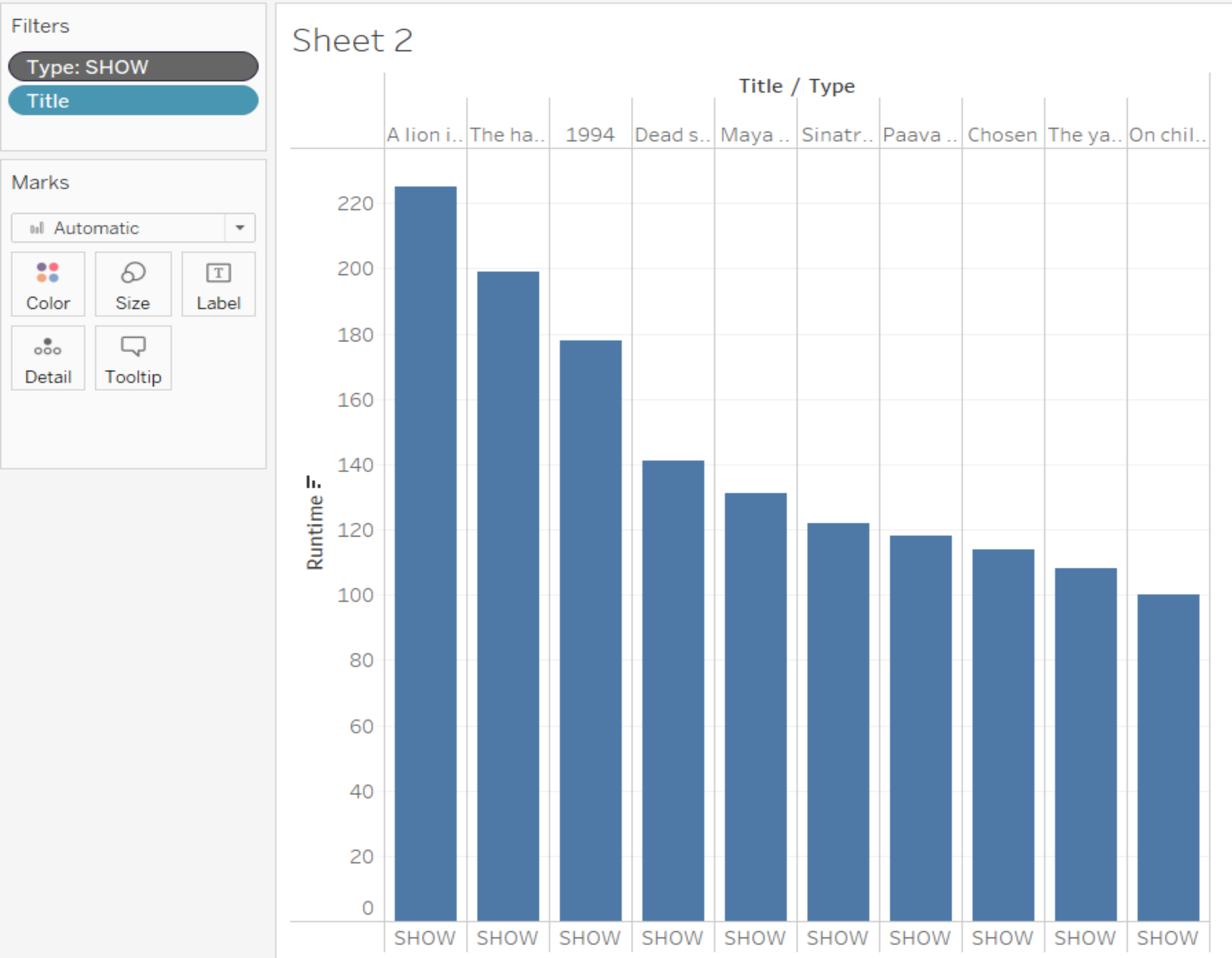
Pages

Columns: Title, Type

Rows: SUM(Runtime)

Filters: Type: MOVIE, Title

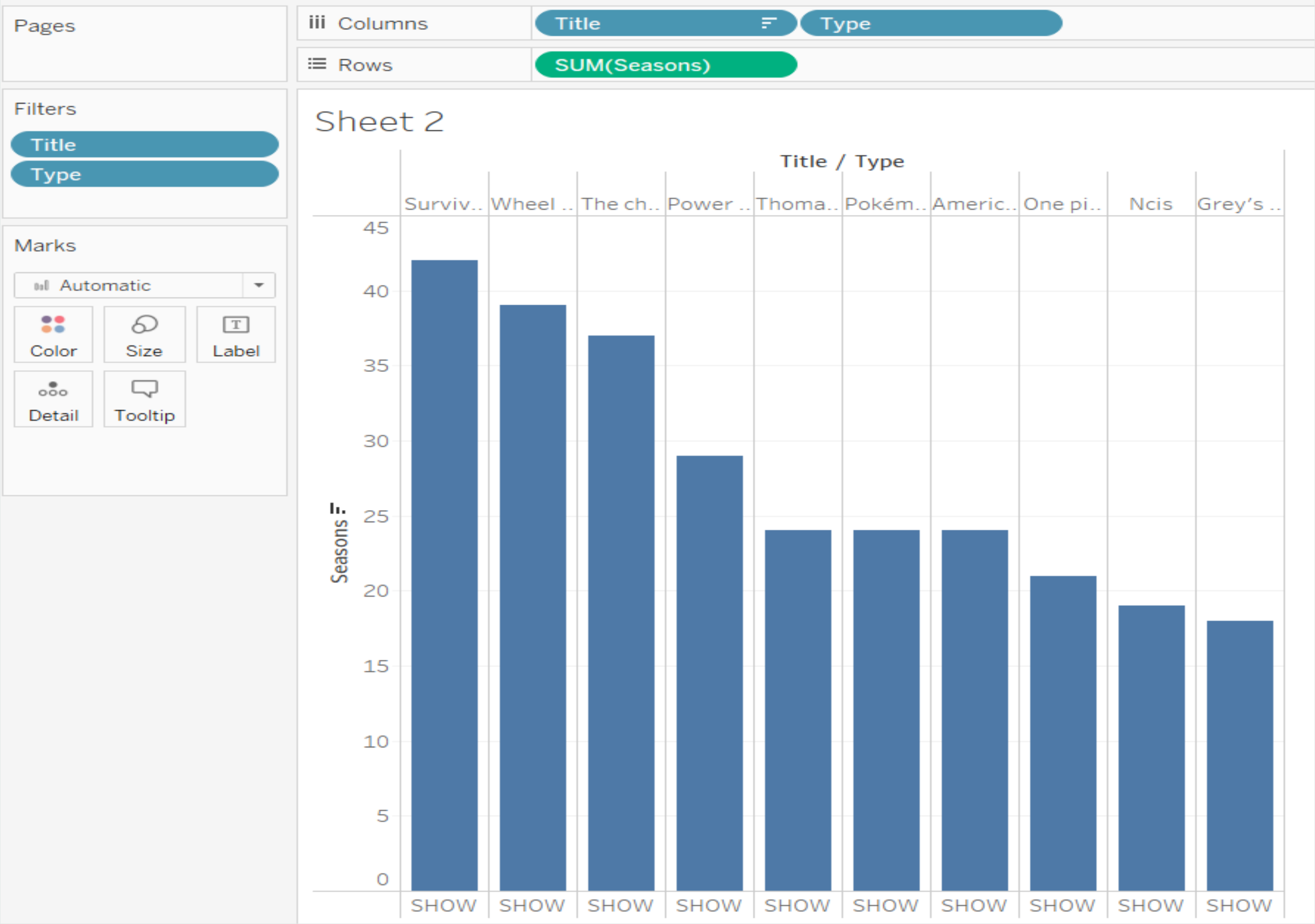
Marks: Automatic



Filters: Type: SHOW, Title

Marks: Automatic

Top 10 shows with the most seasons



Pages

Columns: Title, Type

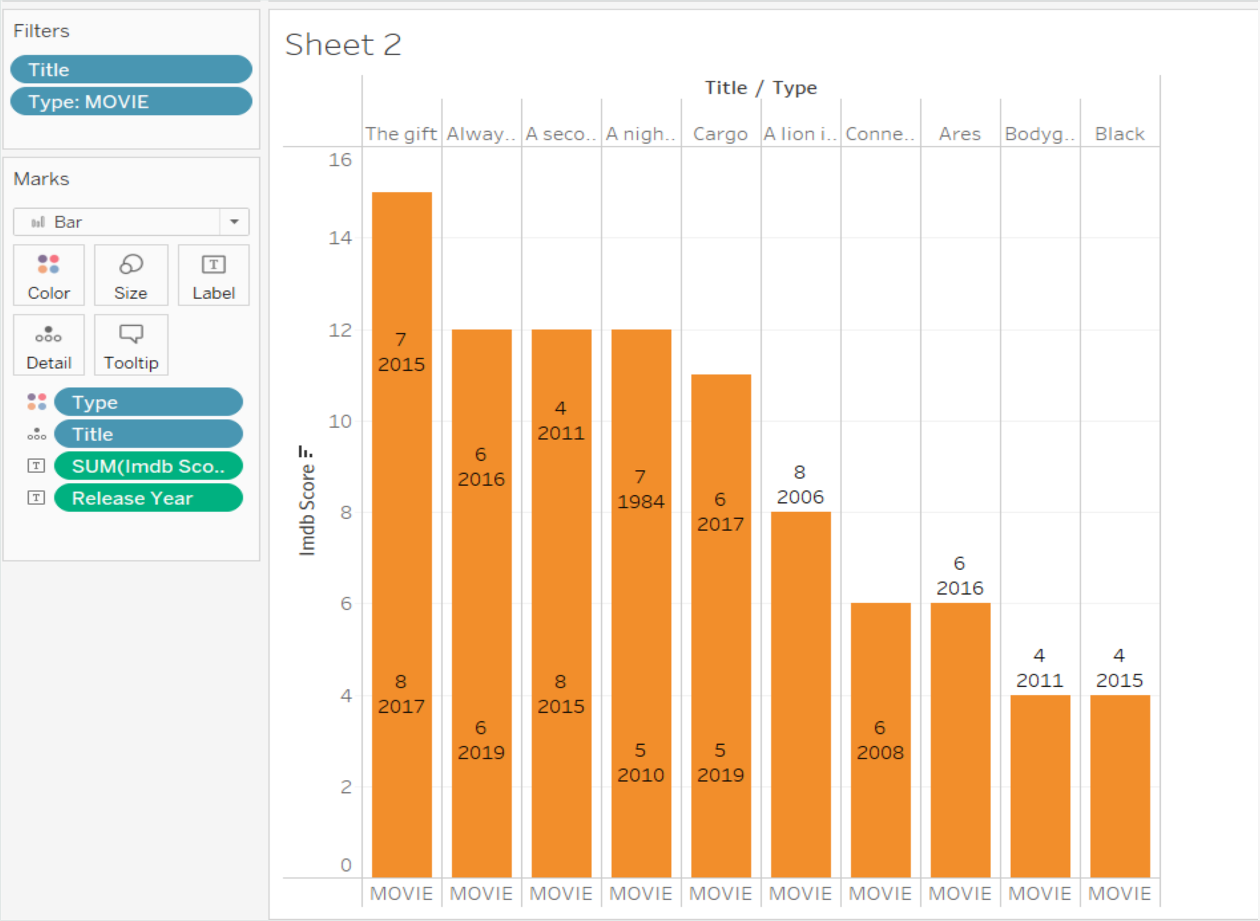
Rows: SUM(Seasons)

Filters: Title, Type

Marks: Automatic

Sheet 2

Statistics of the top 10 movies or shows produced in years.



Filters

Title

Type: MOVIE

Marks

Bar

Color Size Label

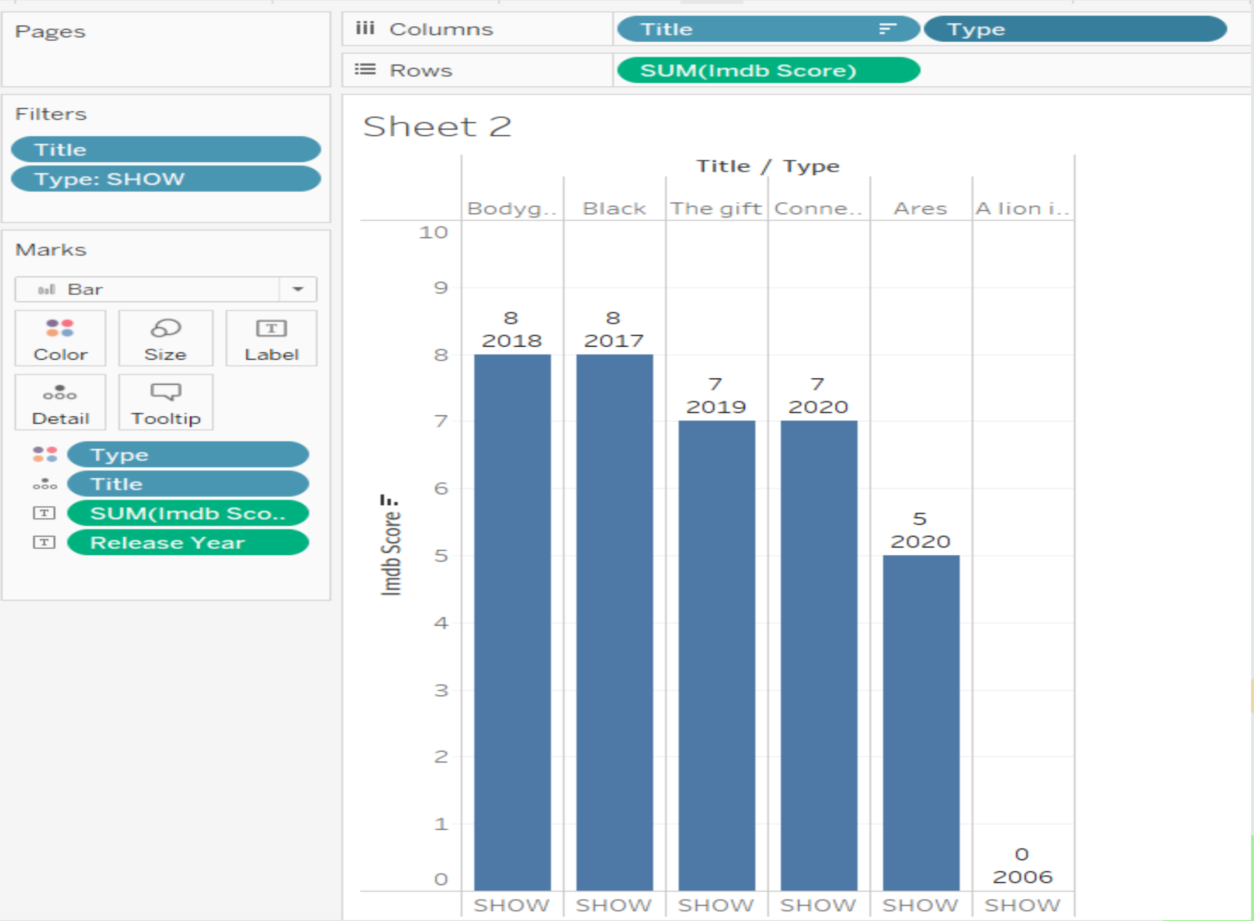
Detail Tooltip

Type

Title

SUM(Imdb Sco..

Release Year



Pages

Columns: Title, Type

Rows: SUM(Imdb Score)

Filters

Title

Type: SHOW

Marks

Bar

Color Size Label

Detail Tooltip

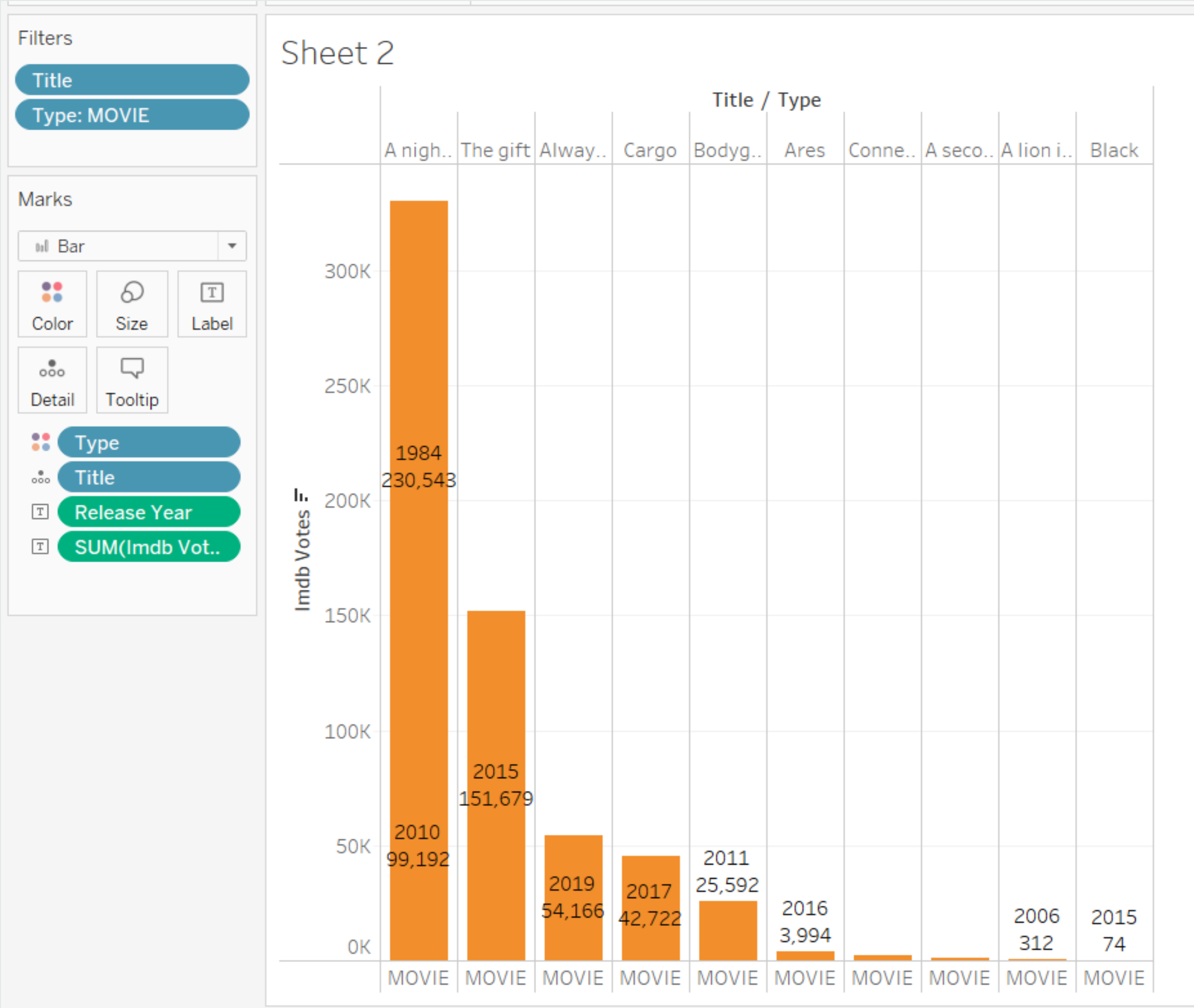
Type

Title

SUM(Imdb Sco..

Release Year

Statistics of votes in the top 10 movies or shows produced in each years.



Filters

Title

Type: MOVIE

Marks

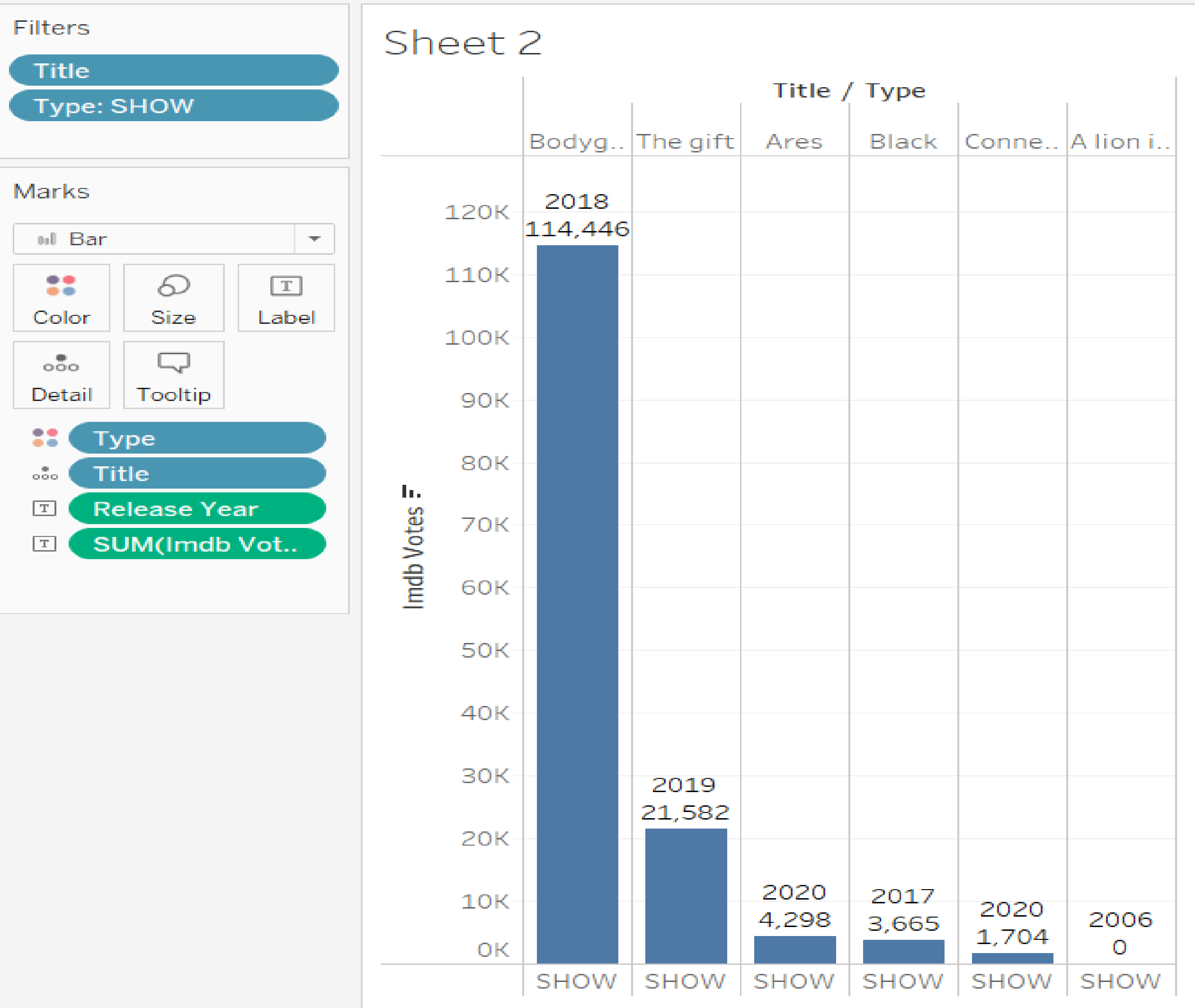
Bar

Type

Title

Release Year

SUM(Imdb Vot..



Filters

Title

Type: SHOW

Marks

Bar

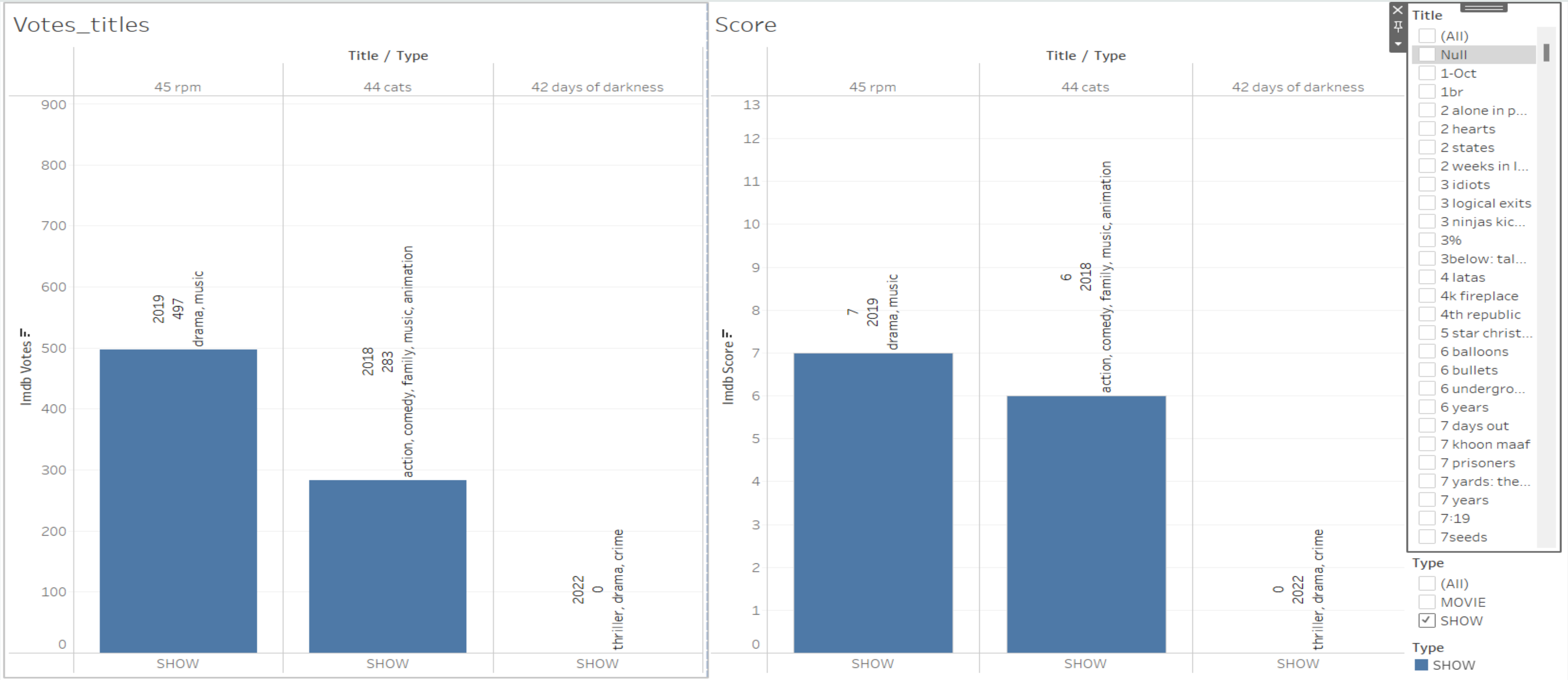
Type

Title

Release Year

SUM(Imdb Vot..

Dashboard 1



Title

- □ (All)
- □ Null
- □ 1-Oct
- □ 1br
- □ 2 alone in p...
- □ 2 hearts
- □ 2 states
- □ 2 weeks in l...
- □ 3 idiots
- □ 3 logical exits
- □ 3 ninjas kic...
- □ 3%
- □ 3below: tal...
- □ 4 latas
- □ 4k fireplace
- □ 4th republic
- □ 5 star christ...
- □ 6 balloons
- □ 6 bullets
- □ 6 undergro...
- □ 6 years
- □ 7 days out
- □ 7 khoon maaf
- □ 7 prisoners
- □ 7 yards: the...
- □ 7 years
- □ 7:19
- □ 7seeds

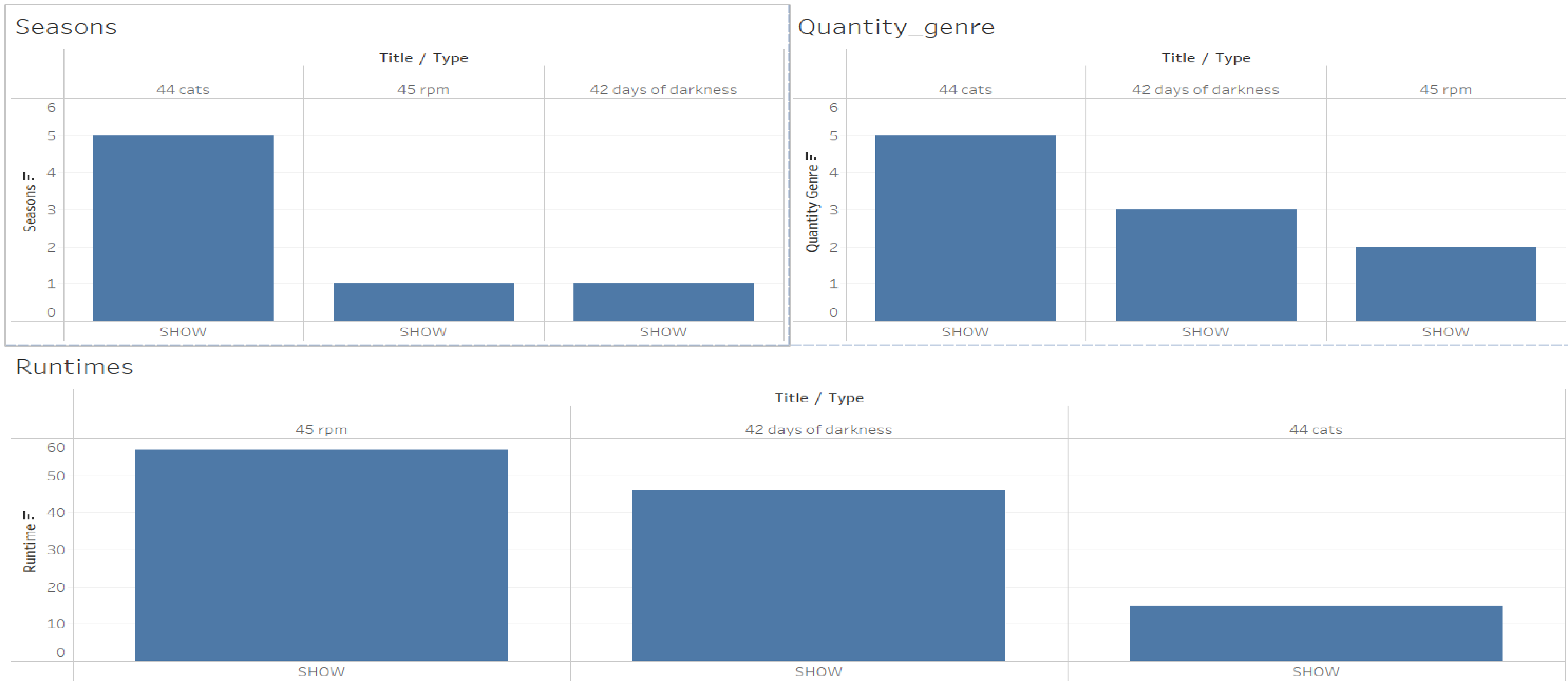
Type

- □ (All)
- □ MOVIE
- ☑ SHOW

Type

🟦 SHOW

Dashboard 2


Seasons
Title / Type
44 cats
45 rpm
42 days of darkness
Seasons
6
5
4
3
2
1
0
SHOW
SHOW
SHOW
Quantity_genre
Title / Type
44 cats
42 days of darkness
45 rpm
Quantity Genre
6
5
4
3
2
1
0
SHOW
SHOW
SHOW
Runtimes
Title / Type
45 rpm
42 days of darkness
44 cats
Runtime
60
50
40
30
20
10
0
SHOW
SHOW
SHOW

Influence of BI