# ASSIGNMENT 2 FRONT SHEET

| Qualification | BTEC Level 5 HND Diploma in Computing | | |
|---|---|---|---|
| **Unit number and title** | **Unit 14 : Business Intelligence** | | |
| **Submission date** | 13 - 05 - 2022 | **Date Received 1st submission** | |
| **Re-submission Date** | | **Date Received 2nd submission** | |
| **Student Name** | Nguyen Van Duy | **Student ID** | GCH 18819 |
| **Class** | GCH 0901 | **Assessor name** | Doan Trung Tung |

**Student declaration**

I certify that the assignment submission is entirely my own work and I fully understand the consequences of plagiarism. I understand that making a false declaration is a form of malpractice.

| | **Student's signature** | *Duy* |
|---|---|---|

**Grading grid**

| P3 | P4 | P5 | P6 | M3 | M4 | D3 | D4 |
|---|---|---|---|---|---|---|---|
| | | | | | | | |

| ☐ Summative Feedback: | | ☐ Resubmission Feedback: |
|---|---|---|
| **Grade:** | **Assessor Signature:** | **Date:** |
| **Lecturer Signature:** | | |

# Business

# Intelligence

*Group: 4*

# What is BI

Business intelligence (BI) is a technology-driven process for analyzing data and delivering actionable information that helps executives, managers and workers make informed business decisions. As part of the BI process, organizations collect data from internal IT systems and external sources, prepare it for analysis, run queries against the data and create data visualizations, BI dashboards and reports to make the analytics results available to business users for operational decision-making and strategic planning.

# Real examples of how to apply BI on business

## **Amazon use Big Data**

→ Amazon may collect information about your purchases on their website and through their application activities, such as the products you purchase, and how often you do.

→ By analyzing the buying habits of a large number of customers, Amazon can effectively leverage the data. It examines new developments and variations to existing products based on data.

# Real examples of how to apply BI on business

**Cosmetics company uses CRM system to increase sales**

When the company just launched, the total revenue of 572 million, of which single accounts for 87%, combos only 13% and the rate of returning customers is almost only 7%. Then when the company started to go downhill, the company used the CRM system to find out the reason above to let the leaders make decisions.

Set sales goals when selling acne products. forced to sell more face wash combo. Then analyze more deeply about the product use cycle, For example, acne usually has a 30-day usage cycle, if you actively call the customer before it runs out.

Upsell rate increased significantly. sales reached 925 million, simply because at this time, single order only accounts for 40% and combo accounts for 60%. Not only that, for customers who have already purchased, the percentage of customers back to us after the first has increased from 7% to 48%. 15% of our new customers come from referrals from existing customers at no MKT cost.

# Data cleansing

**What is Data cleansing**

When aggregating multiple data sources, it is easy for the data to become inaccurate. This affects the results and the algorithm of the analysis. Therefore, the data needs a cleaning process

Data cleansing is the process of repairing or removing incorrect, corrupted, malformed, duplicate, or omitted data in a data set.

# Business intelligence reporting

Business intelligence reporting is the process of collecting data using various software and tools to extract relevant insights.
→ Reports allow business users to see data trends over time and explore relationships between data.
→ Finally, it provides recommendations and observations on business trends, helping to make decisions.
→ Reports are available in a variety of formats: tables, spreadsheets, PDFS, Images, charts, graphs or dashboards, etc

# BI Dashboards

Business intelligence dashboards are a mainstay in today's enterprise-class data analytics tools, and for good reason. Business users

regardless of skill level can use these capabilities in order to better

understand what is going on inside their organization. Dashboards make

it easy for a company to visualize their data by displaying metrics, graphs,

gauges, maps, percentages and comparisons of all the information that is

streaming in and out of the company

# BI Query

BI Query is an enterprise query, reporting, and analytics application. BI Query provides an easy-to-use, visual way to query databases, integrate data with other applications, and generate reports.

BI Query lets information they need using a data model, a graphical representation of data in a database. By using the data model, business users can easily form queries without needing to know Structured Query Language (SQL) users extract the

# Regression Analysis

**In the domain of statistics, the most used statistical technique is Regression Analysis which is used to estimate relationships among variables. Under this technique, the focus is upon the relationship between dependent variable and any one or more independent variables. There are several techniques within this analysis that are used for modeling and analyzing several variables.**

# Machine Learning

Machine learning is a part of artificial intelligence and the subfield of Data Science. It is a growing technology that enables machines to learn from past data and perform a given task automatically. It can be defined as: Machine Leaning allows the computers to learn from the past experiences by its own, it uses statistical methods to improve the performance and predict the output without being explicitly programmed.

# Tools

## Excel

The team had used Excel to help record, present processed information in tabular form, perform calculations and build statistics, visualization included in the table from Excel.

In the data cleansing process, the team had used Excel spreadsheets to represent data into many cells created by rows and columns, which helped the author to identify the unclean data, or unusable data easily.

For example, the team can observe among all the columns to clarify the necessary columns with the unimportant one in oder to keep the clean data for each column and continue with the chosen column for more
cleaning.

# Tools

## Python

Generally Python is a powerful, high-level, object-oriented programming language. Moreover, it is easy to learn and is emerging as one of the best introductory programming languages for first-timers to programming languages.

When applying on the data cleansing process, Python was showing to be a great tool because it was intergrated with many libraries for file and data processing. For example, the team had used "Pandas" for easily importing and exporting dataset file which had the .csv extension. For data processing, the team can use some Python techniques like array indexing, list comprehension, etc.

# Tools

## Tableau

The purpose that the team uses this tool comes from its outstanding capabilities in the process of visualizing data. Tableau is capable of handling a very large amount of data, so after the data pre-processing, Tableau will be able to easily read and systemize that data and represent it back into visual forms like pie. charts, bar charts, etc. Furthermore, it contains various options of filtering, sorting, and operators which provided a great interactive ability to the dataset, so it helped the team better at analyzing the dataset. By using Tableau, the team can also create dashboards out of charts, so that the team can have an overview of the dataset from different angles, so it would improve the data analyzing while the team can keep track with the changes or the trends when interacting with the dashboards which will enhance the decision making processes.

# Business intelligence tools can contribute to effective decision-making

### How Netflix's Business Intelligence is Quickly changing the Sales and Marketing Landscape?

Netflix uses big data analytics to create different types of user personas based on their previous watches and then serves personalized recommendations in the form of changing thumbnails. Showing changing thumbnails instead of a static poster of a movie, that too with stills carefully picked for different personas, makes recommendations even more exciting and eye-catchy.

### How did Creators of House of Cards Know that the Show was a Hit Before it was a Hit?

BI processes were used for selecting which show to produce, which actors to cast, trailer making, color palette selections, and character personas as well. The most intriguing part is that BI processes gave so much confidence to the company that it committed to two seasons of the show without even watching the pilot. And the bid for those two seasons (26 episodes) was worth $100 million. Jonathan Friedland, Netflix's communications director, told Wired, "We know what people watch on Netflix and we're able with a high degree of confidence to understand how big a likely audience is for a given show based on people's viewing habits.

# Legal Issues Involved

## Privacy

User's privacy is a very sensitive topic in daily's technology industries since using user's information without permission can lead to many legal issues and can destroy business' image. Furthermore, an appropriate privacy policy is very important in collecting information and processing data.

# Legal Issues Involved

## Data Encryption

Some companies use internal control system to prevent information leak, however, privacy and security issues need to be handled both from the point of view of man-made error and technological advancement. As far as handling errors of employees are concerned, any company or organization should first educate them about the work culture and the procedure of handling confidential and sensitive corporate data. Company should have strict law to act against the person indulged in such breaching

# Dataset
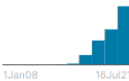


**netflixData.csv** (2.5 MB)

Detail   Compact   Column
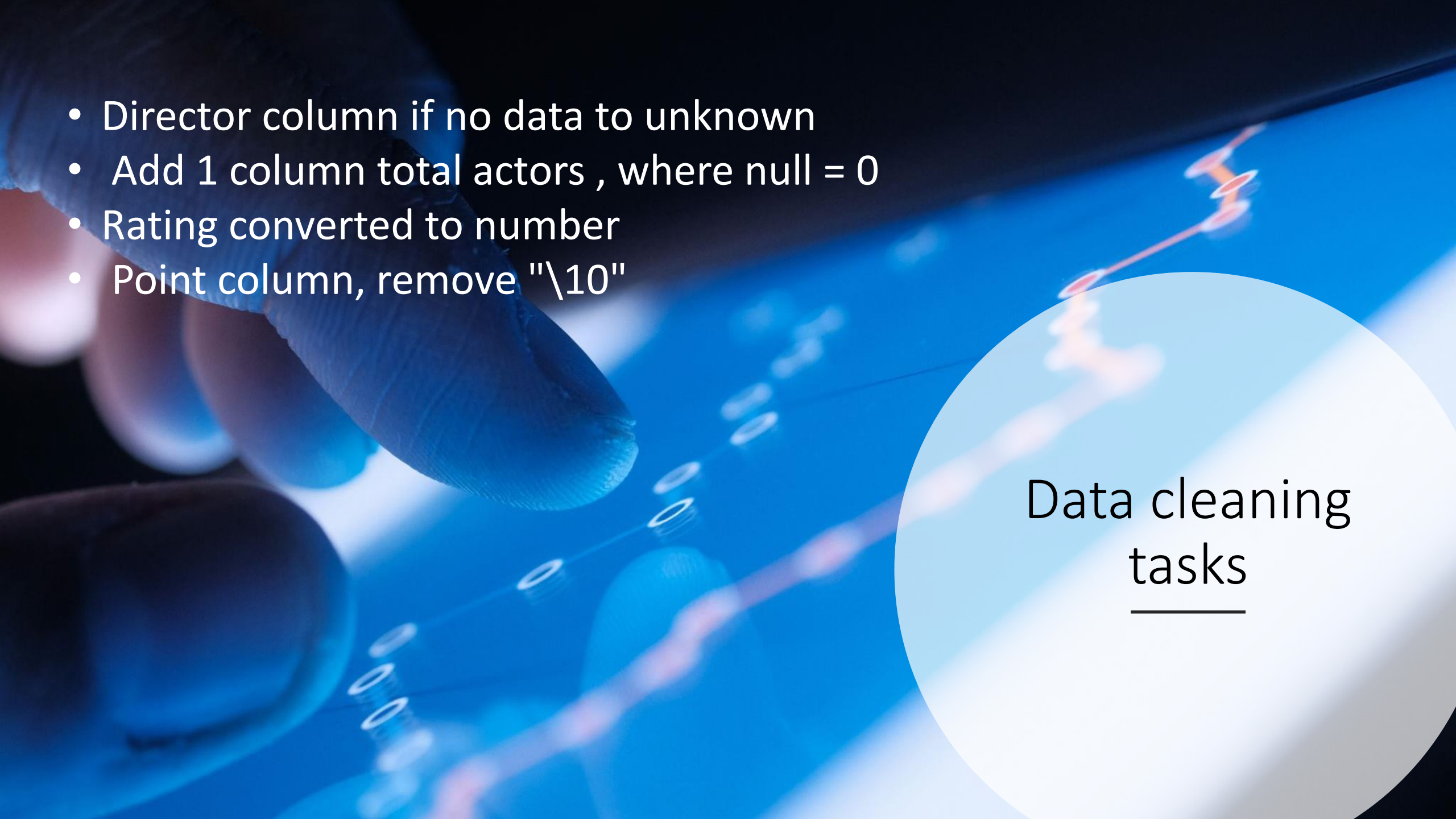
13 of 13 columns ∨

**About this file**

There are 5097 entries in the dataset.

| ⊘ Show Id | A Title | A Description | A Director | A Genres | A Cast | A Production Country | # Release Date | A Rating | A Duration | A Imdb Score | A Content Type | 📅 Date Added |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Unique Id for each entry | Title | Description | Director | Genres (each content belongs to multiple categories. They are seperated by commas) | Cast (seperated by commas) | Production Country | Release Date (year in which it was released) | Rating | Duration (in min for movies and #seasons (eg 2 Seasons) for shows) | IMDb Score | type (movie or TV show) | Date added on Netflix |
| **5967** unique values | **5897** unique values | **5946** unique values | [null] 35% / Raúl Campos, Jan S... 0% / Other (3885) 65% | Stand-Up Comedy 5% / Dramas, Internation... 4% / Other (5398) 90% | [null] 9% / Jeff Dunham 0% / Other (5430) 91% | United States 30% / India 11% / Other (3547) 59% | 1925 — 2021 | TV-MA 43% / TV-14 26% / Other (1875) 31% | 1 Season 23% / 2 Seasons 6% / Other (4236) 71% | [null] 10% / 7.1/10 4% / Other (5126) 86% | Movie 65% / TV Show 35% | 1Jan08 — 16Jul21 |
| cc1b6ed9-cf9e-4057-8303-34577fb54477 | (Un)Well | This docuseries takes a deep dive into the lucrative wellness industry, which touts health and heali... | | Reality TV | | United States | 2020 | TV-MA | 1 Season | 6.6/10 | TV Show | |
| e2ef4e91-fb25-42ab-b485-be8e3b23dedb | #Alive | As a grisly virus rampages a city, a lone man stays locked inside his apartment, digitally cut off f... | Cho Il | Horror Movies, International Movies, Thrillers | Yoo Ah-in, Park Shin-hye | South Korea | 2020 | TV-MA | 99 min | 6.2/10 | Movie | September 8, 2020 |
| b01b73b7-81f6-47a7-86d8-acb63080d525 | #AnneFrank - Parallel Stories | Through her diary, Anne Frank's story is retold alongside those of five Holocaust survivors in this ... | Sabina Fedeli, Anna Migotto | Documentaries, International Movies | Helen Mirren, Gengher Gatti | Italy | 2019 | TV-14 | 95 min | 6.4/10 | Movie | July 1, 2020 |
| b6611af0-f53c-4a08-9ffa-9716dc57eb9c | #blackAF | Kenya Barris and his family navigate relationships, race and culture while grappling with their newf... | | TV Comedies | Kenya Barris, Rashida Jones, Iman Benson, Genneya Walton, Scarlet Spencer, Justin Claiborne, Ravi Ca... | United States | 2020 | TV-MA | 1 Season | 6.6/10 | TV Show | |
| 7f2d4170-bab8-4d75-adc2-197f7124c070 | #cats_the_mewvie | This pawesome documentary explores how our feline friends became online icons, from the earliest tex... | Michael Margolis | Documentaries, International Movies | | Canada | 2020 | TV-14 | 90 min | 5.1/10 | Movie | February 5, 2020 |

- Director column if no data to unknown
- Add 1 column total actors , where null = 0
- Rating converted to number
- Point column, remove "\10"

Data cleaning tasks

# Cleansing Process

```python
import csv
data = []
f = open('netflixData.csv', 'r',encoding="utf-8")
for row in csv.reader(f, delimiter=','):
    data.append(row)

header = data[0]
header.insert(6, "Number of actors")
data = data[1:]
```

- First declare csv library and read csv file(dataset file)

```python
def Director(row):
    x = row[3]
    if x == '': x = 'Unknown'
    row[3] = x
    return row
data = list(map(Director,data))
```

# Cleansing Process

Director column if no data to unknown

# Cleansing Process

Add 1 column total actors , where null = 0

```python
def NumberOfActors(row):
    x = row[5]
    if x == '' :
        x = '0'
        row.insert(6,x)

    else :
        x = x.split(",")
        x = len(x)
        row.insert(6,x)
    return row
data = list(map(NumberOfActors,data))
```

# Cleansing Process

```python
def Rating(row):
    x= row[8]
    if x == 'TV-MA' and 'R' : x = 18
    elif x == 'PG-13' : x = 17
    elif x == 'TV-14': x = 14
    elif x == 'TV-G'and 'TV-Y' : x = 6
    elif x == 'TV-PG' : x = 13
    else : x = 17
    row[8] = x
    return row
data = list(map(Rating,data))
```

- Rating converted to number
-  Point column, remove "\10"

```python
def Score(row):
    x = row[10]
    x =x[:-3]
    row[10] = x
    return row
data = list(map(Score,data))
```

# Dataset

Deleted columns:
- id . column
- URL
- blot, sumeri
- adult movies

Tasks
- type = movie then episodes = 1
- Geners : add 1 column total Geners
- actors to form ',' add 1 column total actors

Data cleaning tasks

# Cleansing Process

```python
import csv
data = []
f = open('june_netflix.csv', 'r',encoding="utf-8")
for row in csv.reader(f, delimiter=','):
    del row[0]
    del row[10]
    del row[10]
    del row[13]
    del row[14]
    data.append(row)
header = data[0]
header.insert(13, "Total of geners")
header.insert(15, "Total of Actors")
data = data[1:]
```

First declare csv library and read csv file(dataset file)

 Deleted columns:

- id . column

- URL

-  blot, sumeri

-  adult movies

# Cleansing Process

type = movie then episodes = 1

```python
def type_move(row):
    x = row[7]
    if x == 'movie': row[5] = 1
    row[7] = x
    return row
data = list(map(type_move,data))
```

# Cleansing Process

category: add 1 column total category

```python
def Geners(row):
    x = row[12]
    x = x.split(",")
    x = len(x)
    row.insert(13,x)
    return row
data = list(map(Geners,data))
```

actors to form ',' add 1 column total actors

# Cleansing Process

```python
def actor(row):
    x = row[14]
    x = x.replace("'", "")
    x =  x = x.replace("[", "")
    x =   x.replace("]", "")
    row[14] = x
    return row
data = list(map(actor,data))

def Total_Actor(row):
    x = row[14]
    x = x.split(",")
    x = len(x)
    row.insert(15,x)
    return row
data = list(map(Total_Actor,data))
```

```python
with open('june_Done.csv', 'w', newline='', encoding='utf-8') as f:
    w = csv.writer(f)
    w.writerow(header)
    for row in data:
        w.writerow(row)
print('Done')
```

# Cleansing Process

Write the processed data to a new file

# Design Dashboard

- Statistics to see which genre is most popular, which movie genre is voted the most

- Movie production country according to the popularity of the genre and according to the popularity of the viewers

- Statistics of the 10 most popular actors, along with movie genres and popularity
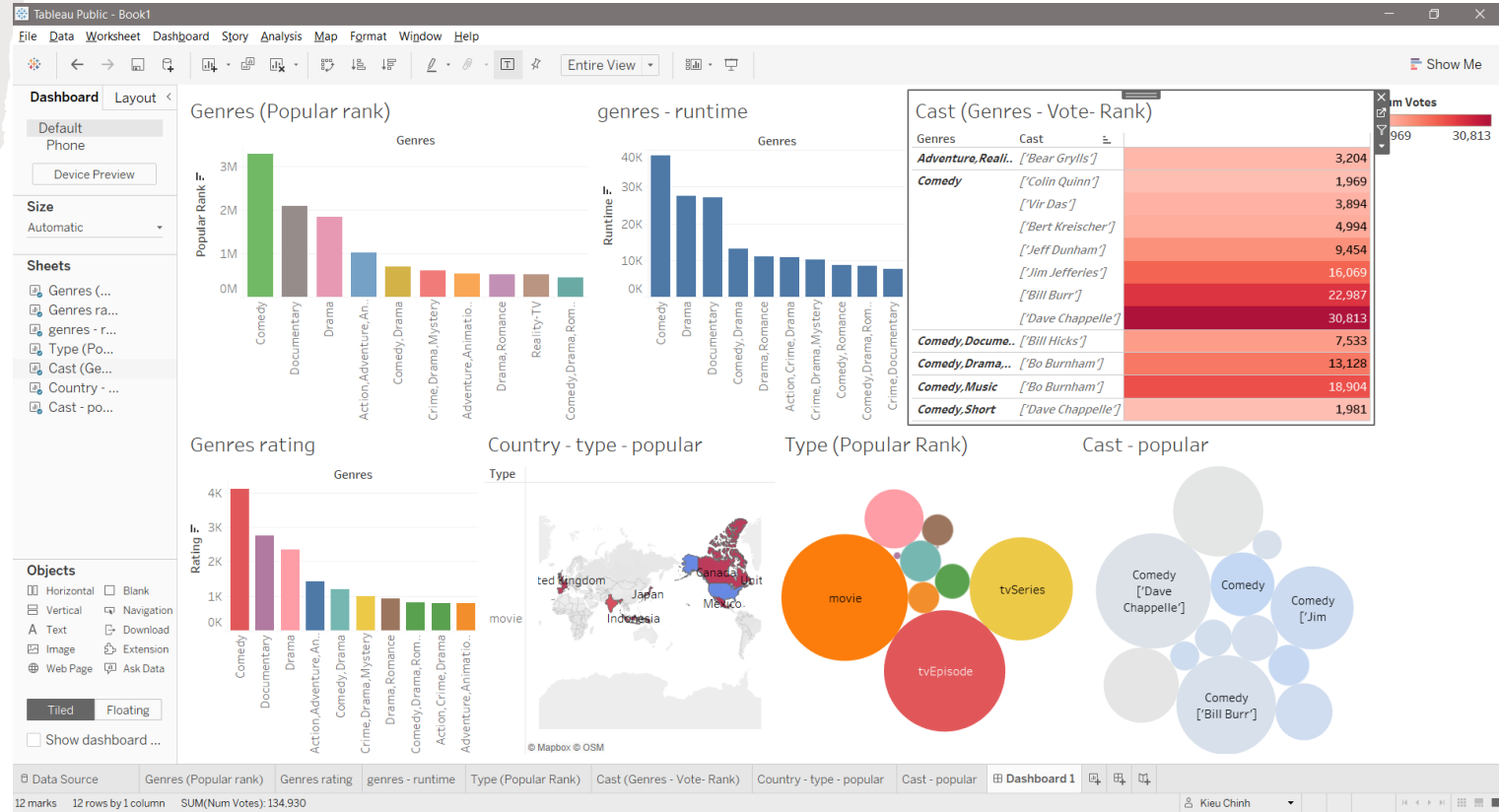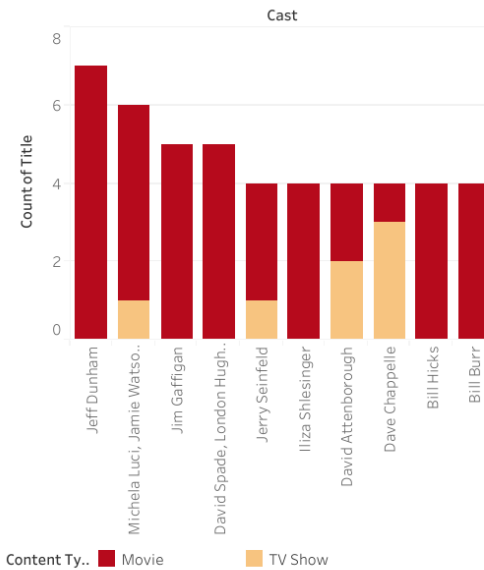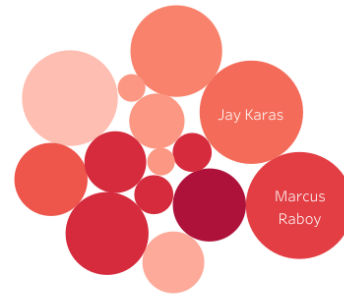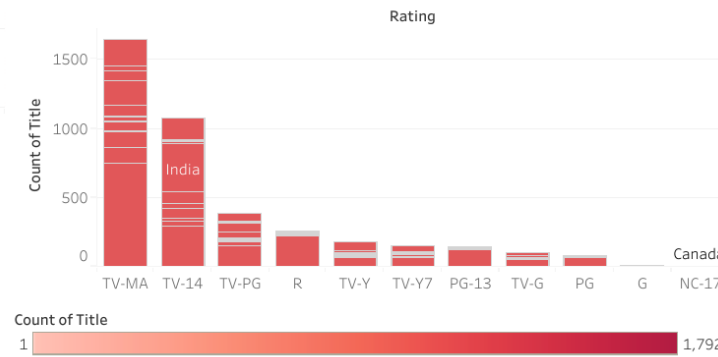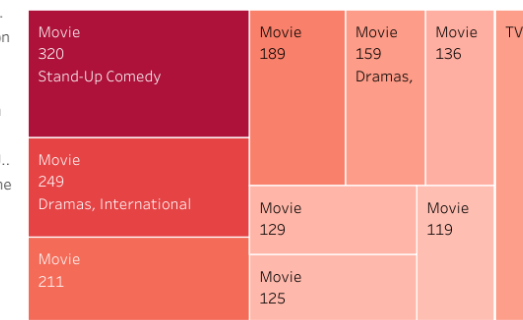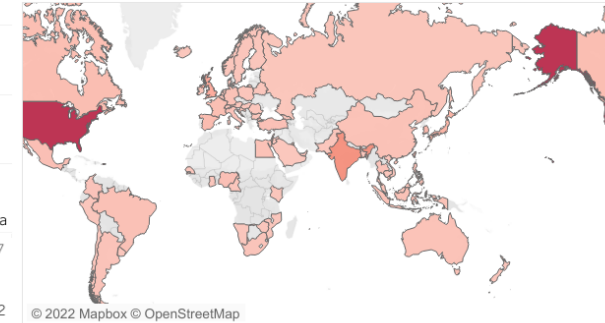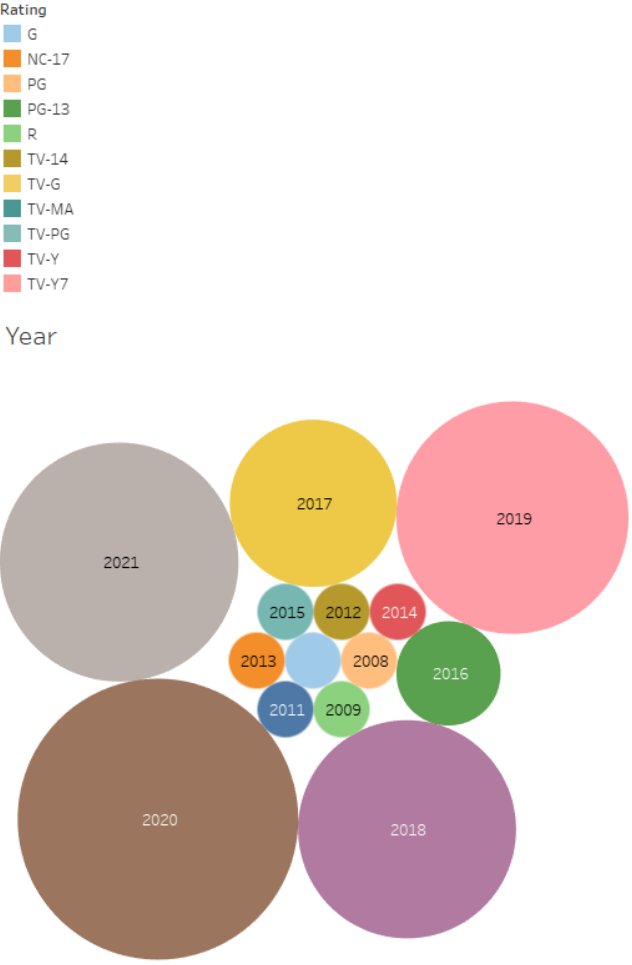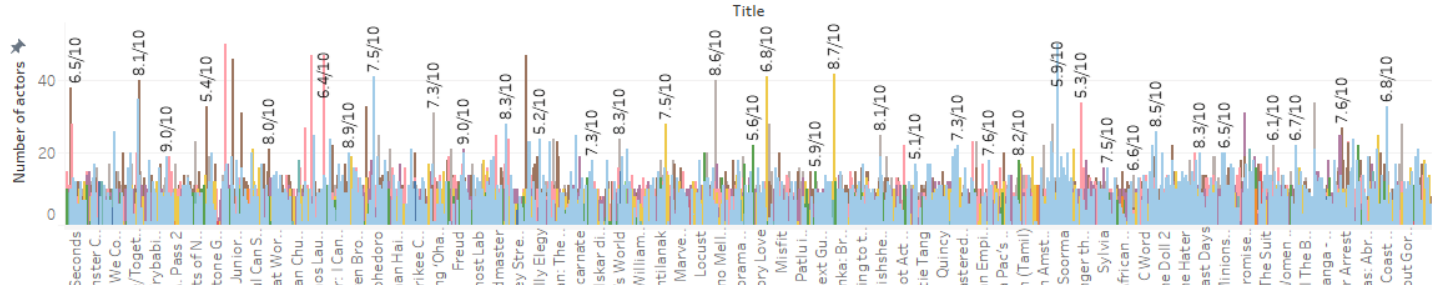
# Design Dashboard

- This chart helps users to see the entire map for each movie genre that the country develops and the number of users for each movie genre, and then shows the top 10 movies of the whole country (tv show). and movies). Based on the top 10 movies, it is possible to determine the top 10 directors and the top 10 cast based on popularity and votes.

- Thereby, it is easy to find a way to develop the business.

# Design Dashboard

Thank you for watching