

|  |
| --- |
| **PROJECT ON EXCEL**  **(Project Title)** |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Developed by**  **Members:**   |  |  |  | | --- | --- | --- | | **No.** | **Student Name** | **Student ID** | | 1 | LÊ VĂN KHIÊM |  | | 2 | ĐẶNG THÀNH TRUNG |  |  * **Class No.:** * **Start Date:** * **End Date:** * **Name of the Coordinator:** * **Date of Submission:** |
| **1. GET THE DATA** |

* **Who**: Who collected the data? What perspectives or agenda might that person or group bring to the data collection? For example, data collected from partisan political groups may reflect their biases.

**Author name**: BIO

**Khiem & Trung collect data and discuss the story to be told**. The data was posted in Kaggle.

* **What**: What data is included? Is the data complete? Does it address the data questions?

**Data includes (**Poster\_Link, Series\_Title, Released\_Year, Certificate, Runtime, Genre, IMDB\_Rating, Meta\_score, Director, Star1,Star2,Star3,Star4, Noofvotes, Gross**), completed and ready to answer questions.**

* **When**: When was the data collected? Is it current? Is it still relevant? For example, a five-year-old dataset may be incapable of addressing a customer's current problem.

**Data was created at. (2022), still quite new and capable to address current problem.**

* **Where**: If the data is from an external source, where did it come from? Is the source reputable? Trustworthy? For example, user engagement data would likely be more reliable than user reported data on a survey.

**From Kaggle which is a reliable source**

* **Why**: Why was the data collected? Can data collected for one purpose be appropriately used for another purpose? For example, data collected to track student learning may not be useful for evaluating teacher and school quality.

Data was collect to analyze movies history which is useful for improving future development of the movie industry.

**Example**

<https://www.kaggle.com/datasets/harshitshankhdhar/imdb-dataset-of-top-1000-movies-and-tv-shows>

* **People**
  + Top 1000 movies
  + Poster\_Link = Link of the poster that imdb using
  + Series\_Title = Name of the movie
  + Released\_Year = Year at which that movie released
  + Certificate = Certificate earned by that movie
  + Runtime = Total runtime of the movie
  + Genre = Genre of the movie
  + IMDB\_Rating = Rating of the movie at IMDB site
  + Overview = mini story/ summary
  + Meta\_score = Score earned by the movie
  + Director = Name of the Director
  + tar1,Star2,Star3,Star4 - Name of the Stars
  + Noofvotes = Total number of votes
  + Gross = Money earned by that movie

|  |
| --- |
| **2. PREPARE THE DATA** |

* **Extracting the data from its source**: Extract data from csv file.
* **Transforming/organizing the data**:Rename columns, delete columns, extract columns name,, convert data type, remove rows.
* **Cleaning the data**: Check null and empty values, replace null values equal zero, replace empty equal NaN
* Correcting inaccurate or irrelevant data
* Updating missing or incomplete data
* Standardizing the case for text
* Standardizing formats for each data type (dates, numbers, etc.)
* Working with outliers
* Visualizing to verify clean data

|  |
| --- |
| **3. EXPLORE THE DATA (EXPLORATORY DATA ANALYSIS, OR EDA)** |

* Group data
* Apply descriptive statistics to that data
* Explore and analyze the data
* Ask questions and iterate over our exploratory analysis
* Different kinds of distributions:
  + Normal distributions
* Visualizing the spread, including the interquartile range (IQR) and outliers, for a given distribution using boxplots.
* Visualizing boxplots by categories to visually understand the relationship between numerical columns of data and those categories.

|  |
| --- |
| **4. ANALYZE THE DATA** |

**Descriptive analysis** uses past data to answer the question *What happened*? Previously collected data undergoes analysis to answer key questions about past situations or events.

**Diagnostic analysis** seeks reasons or explanations for what has happened, asking and answering the question *Why did this occur*? Because it is often difficult or impossible to prove cause and effect, diagnostic analysis tends to explore correlations and other relationships in data. For example, Robert Di Nero used to be famous in the 70 years but not earn top Dollar due to it was not accessible and attracted as many audiences as the last 20 years movies.

|  |
| --- |
| **5. COMMUNICATE THE RESULTS** |

1. **Design for an audience**: tailor results and visualizations to the appropriate audience. For example, Visualization with graphs, plots, slides to effectively communicate with mass audiences
2. **Less is more**: the takeaway from each visual should be clear. Adding too much data or unnecessary design features can distract our audience from the main point. The reasoning are very cohesive and condensed to bring main ideas clearly
3. **Clear labels and annotations**: The chart titles, axis titles, legends, etc. work together so well that a viewer can read the chart or report and know what the data represents without guessing.
4. **Tell a story with data**: Using elements from storytelling like asking questions, revealing answers, raising possible theories can make our analysis more memorable and engaging.