**MIDDLE EAST TECHNICAL UNIVERSITY**

**DEPARTMENT OF STATISTICS**

**STAT112 TERM PROJECT**

**CLEANING AND ANALYSIS OF DIRTY MOVIE DATA**

ALİ ALTUNTAŞ - 2646529

MEHMET KARAMAN - 2425783

RUBAR AKYILDIZ - 2561033

RÜMEYSA DAŞTAN - 2561173

**ABSTRACT**

The topic of research here was movies. Movie business is a large and evolving area of work. Changes occur constantly and interpreting those changes bring success to the movie makers.

Data processing consists of data cleaning, visualisation and interpretation. Dirty data were retrieved about movies, consisting of variables such as gross, cost, duration, etc. Aim here was to see the relationship and correlation between these variables. The eneral conclusion was that the newest streaming technologies might not be too profitable for producers.

**INTRODUCTION**

After descriptive data analysis, research questions were produced about the data. In order to answer these questions, statistical visualization methods, were used and that required cleaning the data. Computer programming was used to speed up the process, with tools like Python, Pandas and Seaborn. After cleaning and visualisation, interpretation helped with analyzing different factors about movie business, such as linear relationships between gross and cost, or how publishing movies on streaming platforms affect views.

**DATA TIDYING AND CLEANING STEPS**

Data cleaning was the longest part of this research. It took a lot of creative thinking and persistence through a trial-and-error process. The number of the things to consider and check to clean the data while keeping everything explicit and easy to understand for the reader was a challenge.

The steps of this long, tedious, but well-rewarding process (if it is done right) can be explained as follows:

1. Getting your feet wet the data. Raw and dirty data were examined to see what we are dealing with in terms of variables, the size of the data, data types, empty values, what to expect and what not to expect. In order to achieve this, small portions of the data and general information about the dataset were inspected to see if there were NA values or something wrong with the source file by importing the Excel spreadsheet that contains the dataset via Python and pandas. This way descriptive statistics could be analyzed and we would have an idea of the nature of the data. Also, the spreadsheet is skimmed over to have an insight before going blindly to work.
2. Tidying the data. This step consists of manipulating the already existing data entries rather than

altering their existence. These steps were followed to tidy up the dirty data:

* + - * Checking if the column headers of the Excel spreadsheet are values rather than variable names. Variable names themselves should not interfere with the data values.
      * Fixing the column names if they have any typos. This step might include removing whitespaces, capitalizing the words, or renaming.
      * Checking to see if there are multiple different types of data stored in a column.
      * Checking if there are multiple types of observational units stored in the same table.
      * See whether a single observational unit is stored in multiple tables.
      * Remove unnecessary strings from values.
      * Making sure that all string values have same format, for example lower-case.
      * Checking and ensuring the uniqueness of the values.
      * Checking to see if numerical values are float type and categorical values are in object type in pandas. Dates should be in date format if there is month and day specified with year.

1. Cleaning the data. Now it is time to add, remove and change values for good. This is especially important as whole dataset gets manipulated value by value. Cleaning consists of...

* dropping duplicates, if that conflicts with the behavior of the variable.
* dropping unnecessary columns. This could be applied for different scenarios, such as, for a column that has around 60% of its cells empty or NA. Or maybe, the column has no real use for the researcher.
* examining the descriptive statistics and searching for unusual behavior.
* checking if a column has a specified range, and correcting values that are beyond that range, possibly replacing them with median, mean or mode.
* searching for outliers and replacing them with the median.
* creating uniformity, by making sure numerical values that are in the same column has the same format and units.
* searching for missing values. As mentioned before, columns with more than 60% NA values are dropped. However, if the percentage is lower, NA values can be replaced with either mean, median or mode.

These steps were followed as data were prepared and processed for interpretation. Detailed comments on how to apply and implement those steps with calculations are available in the relevant *.ipynb* file.

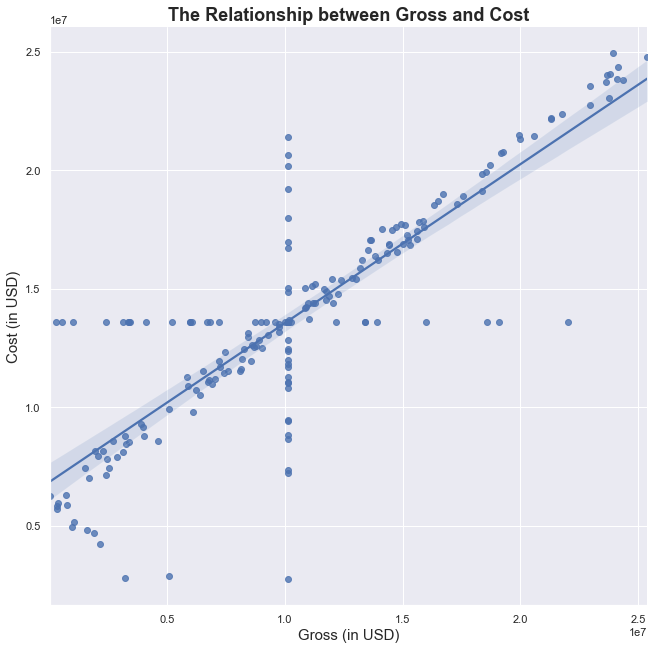
**EXPLORATORY DATA ANALYSIS**

Research questions were prepared, and then graphs were plotted using seaborn and matplotlib to answer them. More visualizations and detailed methods are available on the relevant *.ipynb* file.

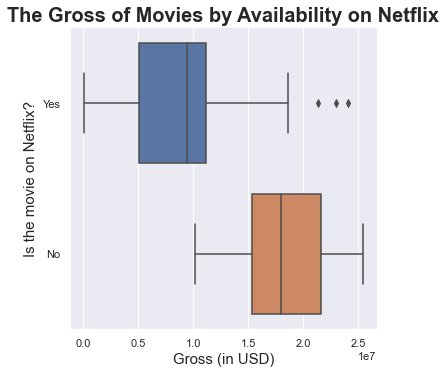
**Research questions:**

**1.** Do larger investments in the movie industry yield more gross? What is the relationship between gross and cost for movies? Should you invest more to earn more?

Answer: Based on the strong linear correlation between cost and gross in the scatter plot below, it can be interpreted that there is a positive relationship between the two variables. This means that as the cost increases, the gross also tends to increase. This suggests that movies with higher budgets tend to generate more money at the box office. To conclude, filmmakers that make a lot of money are probably the ones that invest a lot of money.



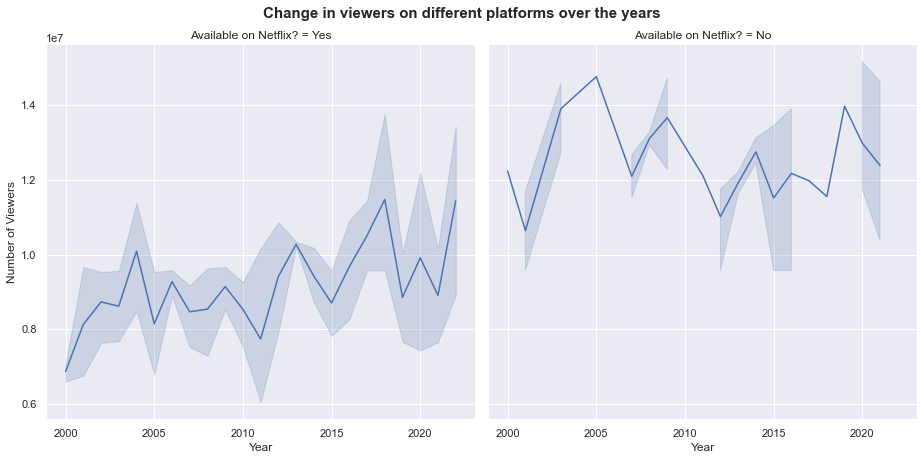
**2.** Do movies on Netflix earn more? How does being available on Netflix affect the gross?

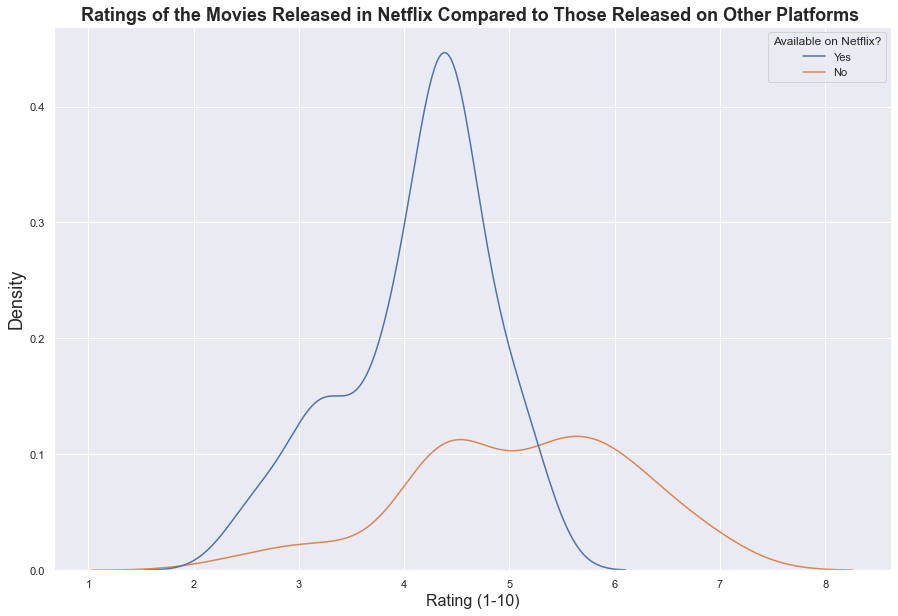


Answer: When we look at these boxplots, while they both seem negatively skewed, Netflix movies have smaller range in their gross. In addition, we can say that Netflix movies gross less. Here, we can interpret that Netflix streams low-grossing movies and it would not be profitable to stream movies on Netflix.

**3.** How do the number of the viewers change for different platforms over the years?

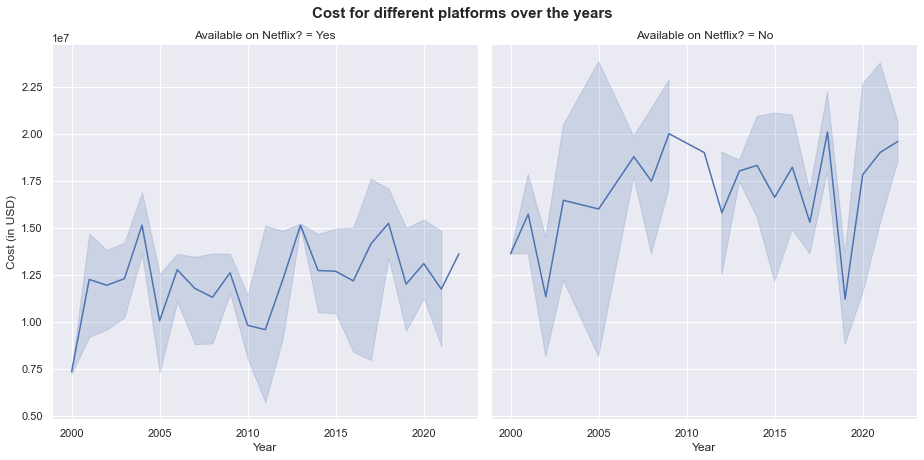
Answer: The increasing line plot below indicates that the number of viewers has increased over the years both for Netflix and other platforms while non-Netflix movies had larger audiences overall. This suggests that there is a growing interest or demand for films among the population in general. The specific reasons for this increase could be due to a variety of factors such as advancements in technology, changes in societal trends, or an increase in the number of films being produced and distributed.



**4.** How do ratings change for different platforms?

Answer: If we look at this density graph, we can see that the movies which are streamed on Netflix has a narrower rating range but are rated more, and the opposite on other platforms; they are rated at higher scores and cover a wider range of ratings. We can say that Netflix movies are rated lower yet more often, and non-Netflix movies are rated higher yet less often.

**5.** Has it become more expensive to make movies for different platforms over the years?



Answer: While costs in both Netflix and non-Netflix movies are risen over the years, non-Netflix movies has been more expensive to produce over the time. Yet, Netflix cost trends has followed a more steady line while non-Netflix movies had their fair share of ups and downs over the years. This instability and costly trend might push producers to make movies to stream on Netflix only.

**CONCLUSION:**

It was interpreted that:

* big investments might bring “big money” in movie industry,
* Netflix is not necessarily a great source of revenue for producers,
* Netflix might not be a good source for quality movies for viewers,
* traditional movie industry has a large audience base despite the technological advancements in streaming,
* Netflix could be considerably cheaper to make movies for, from producers’ perspectives.

**REFERENCES:**

**1.** The outlier function that we used in the cleaning IPython notebook was taken from https://careerfoundry.com/en/blog/data-analytics/how-to-find-outliers/

**2.** pandas documentation

https://pandas.pydata.org/docs/

**3.** seaborn documentation

https://seaborn.pydata.org/examples/index.html

**OUR GITHUB LINKS:**

* Ali Altuntaş Github page: https://github.com/ForxDeven
* Mehmet Karaman Github page: https://github.com/memeth-my-hawk
* Rubar Akyıldız Github page: <https://github.com/rubaraky/>
* Rümeysa Daştan Github page: <https://github.com/juniorcimcime>