CHANAKYA UNIVERSITY, BENGALURU

PROJECT REPORT – SEPTEMBER 2024

Data Mining – DAT 502

Title of the Paper	Statistical Analysis of Film Data: Measures of Central Tendency
Name	Kumari Yachana
Registration no.	CU23MSD0013A
Program	MSc Data Science
Semester	Second
Name of the examiner	Prof. Usha Subramanian

PROJECT REPORT

Statistical Analysis of Film Data : Measures of Central Tendency

Introduction :-

The film industry is a dynamic sector with substantial variations in box office performance, influenced by factors like distribution strategies, release timing, market competition, and audience reception. Understanding these variations is crucial for stakeholders, including producers, distributors, and cinema owners, who rely on data-driven insights to make strategic decisions. This project aims to analyze the performance metrics of a collection of films through statistical analysis, focusing on measures of central tendencies. The primary objective of this report is to evaluate key numerical data points, such as Rank, Weekend Gross, Percentage Change on Last Week, Weeks on Release, Number of Cinemas, Site Average, and Total Gross to Date, using statistical measures like mean, median, mode, mid-range, range, variance, and standard deviation. These measures provide insights into the distribution, central positioning, and variability of the data, revealing trends and patterns that might otherwise go unnoticed. By analyzing these central tendencies, we can identify typical performance levels, detect anomalies, and understand the spread of the data, offering valuable benchmarks for future film releases. The findings are intended to guide strategic decisions in film marketing, distribution, and cinema management, ensuring that stakeholders can optimize their approaches based on empirical evidence. This report will present a detailed analysis of each measure, supported by visual aids like graphs, box plots and Q-Q plots to enhance comprehension. Through this statistical approach, the project seeks to provide a clear and concise evaluation of the data, highlighting the factors that influence film success in the marketplace.

Data Description:-

Columns Overview:

- **Rank** - The ranking of films.
- **Film** – Name of the movie.
- **Weekend Gross**: Earnings over the weekend.
- **% Change on Last Week**: Weekly percentage change in earnings.
- **Weeks on Release**: Duration of release.
- **Number of Cinemas**: Number of cinemas showing the film.
- **Site Average**: Average earnings per cinema.
- **Total Gross to Date**: Cumulative earnings of the film.

Methodology:-

Statistical Measures:

- **Mean**: The average value.

- **Median**: The middle value when data is sorted.
- **Mode**: The most frequently occurring value.
- **Mid-Range**: The average of the maximum and minimum values.
- **Range**: The difference between the highest and lowest values.
- **Variance**: A measure of data dispersion around the mean.
- **Standard Deviation**: A measure of data spread or variability.
- **Mean Deviation**: The average absolute deviation from the mean.
- **Quartile Deviation (IQR)**: The spread of the middle 50% of the data.

Analysis and Results :-

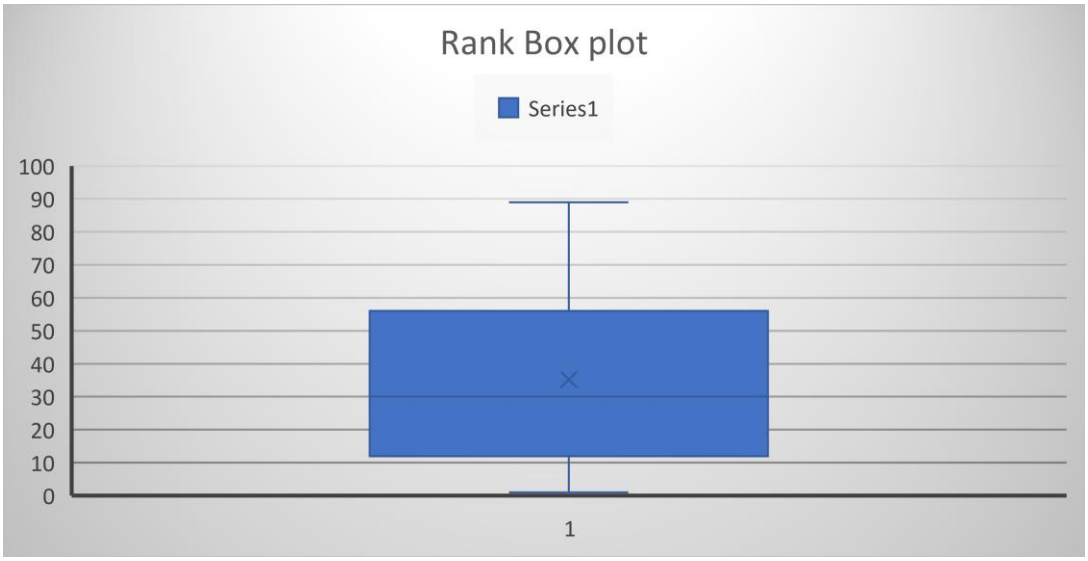
Measure	Rank	Weekend Gross	% change	Weeks on release	No. of cinemas	Site average	Total gross
Mean	28.98	173,888.65	0.14	33.39	194.51	795.02	6,543,426.32
Median	20.0	23,668.5	-0.405	5.0	61.0	464.5	1,425,784.0
Mode	3.0	36.0	-0.78	2.0	1.0	325.0	25,568.0
Mid-Range	41.0	1,092,804.0	7.56	288.5	356.0	2,826.0	31,763,192.5
Range	78.0	2,185,536.0	17.09	573.0	710.0	5,594.0	63,475,249.0
Variance	531.90	148,256,463,481.14	6.01	6,672.92	55,443.10	1,079,025.57	142,084,746,029,530.56
Standard Deviation	23.06	385,040.86	2.45	81.69	235.46	1,038.76	11,919,930.62
Mean Deviation	20.66	226,486.04	1.05	46.57	201.24	624.44	7,452,271.56
Quartile Deviation (IQR)	40.5	114,287.5	0.55	8.25	366.75	499.75	7,496,386.5

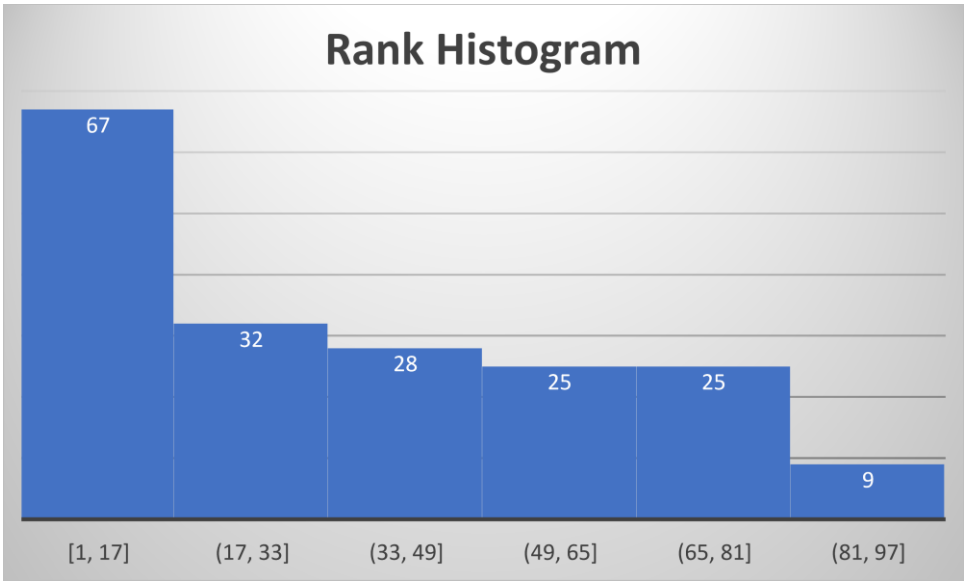
Insights gained :

1. Rank

- The mean rank of 28.98 indicates the average position of the movies in the ranking list. Since lower ranks (closer to 1) are better, a mean of about 29 suggests that, on average, the movies are not at the very top but are relatively well-positioned.
- The median rank of 20 shows that half of the movies are ranked at 20 or better. This is notably lower than the mean, indicating that a significant number of movies are clustered in lower (better) positions, with fewer movies in higher (worse) ranks skewing the mean upwards.
- The mode of 3 is the most frequently occurring rank, indicating that there are multiple instances of movies being ranked very high. This is a strong indicator that several movies have performed particularly well compared to others.
- The mid-range is calculated as the average of the maximum and minimum ranks, providing a sense of the central position between the best and worst ranks. A mid-range of 41 reflects that while some movies are near the top, others are significantly lower.
- Range (78.0): The range, which is the difference between the highest (worst) and lowest (best) rank, is quite large. This suggests a broad dispersion in the ranks, with some movies performing very well and others much worse.
- Variance (531.90) and Standard Deviation (23.06) : A high variance and standard deviation indicate that the ranks are widely spread around the mean, confirming significant variability in movie performance. This spread suggests that while some movies consistently rank well, others do not.
- Mean Deviation (20.66): The mean deviation, which measures the average absolute deviation of each rank from the mean, shows considerable dispersion. This indicates that ranks vary substantially around the average position.
- The IQR measures the spread of the middle 50% of the data. A relatively high IQR of 40.5 suggests that even the central ranks are widely spread, indicating variability even among movies that perform relatively similarly.
- Q-Q Plots Analysis: Deviations from normality, especially in tails.

Graphs :-





Overall Interpretation:

The lower median compared to the mean suggests that many movies rank well, with a few poorly ranked movies pulling the mean upwards.The large range, variance, and standard deviation reflect a significant disparity in performance. Some movies are doing exceptionally well, while others rank much lower.The mode being close to the top ranks (3.0) emphasizes that several movies are repeatedly finding themselves in high positions, which could indicate popular franchises or highly anticipated films consistently drawing audiences. Understanding this variability can help in identifying factors contributing to top performance (e.g., genre, cast, marketing) and addressing what might be dragging lower-ranked movies down.

2. Film :-

The ‘Film’ column contains the titles of the movies, which is categorical data. As such, measures of central tendency like mean, median, mode, and other statistical measures (variance, standard deviation, etc.) are not applicable because these measures require numerical or ordinal data.

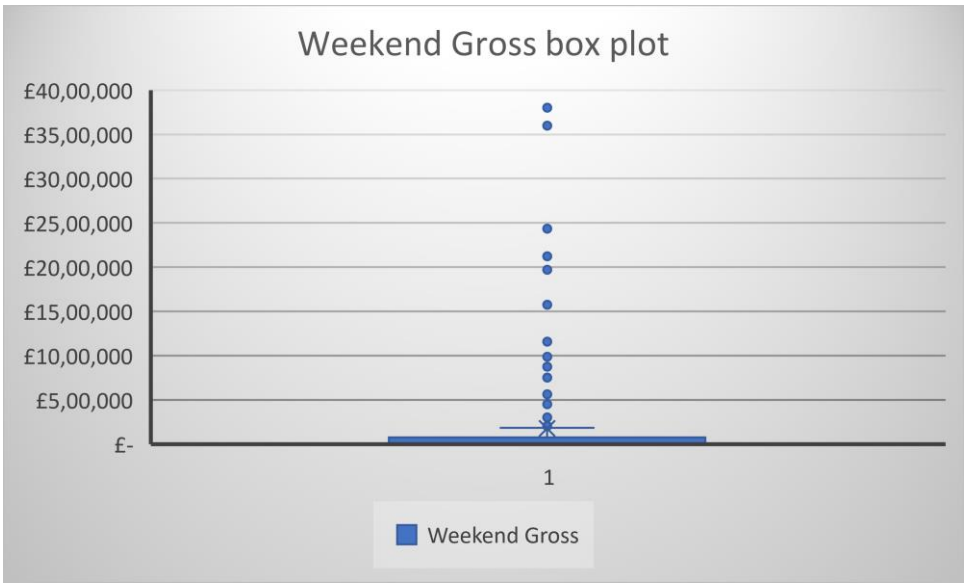
However, we can still derive insights from the ‘Film’ column by analyzing it in other ways.

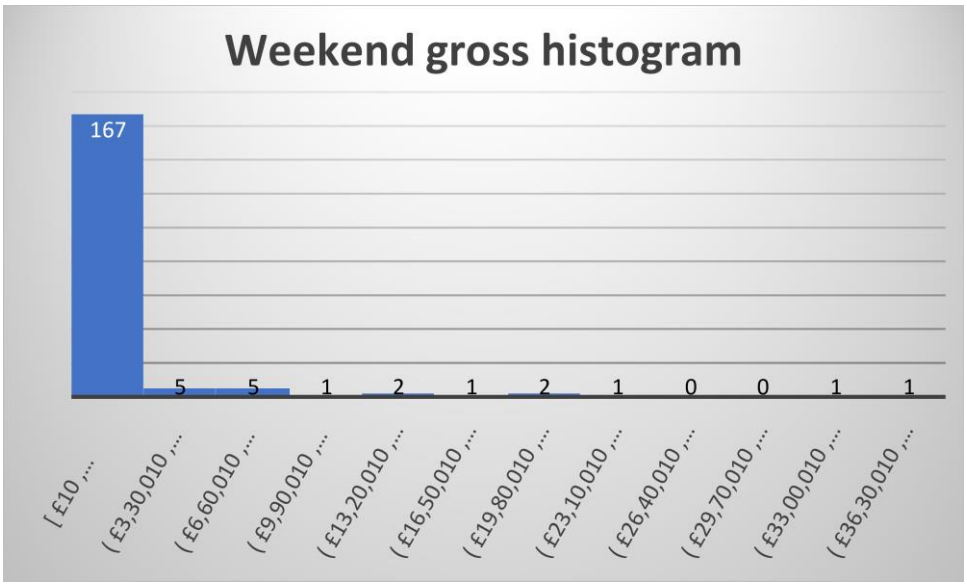
- Mode (Most Frequent Film) - Not applicable: Since the ‘Film’ column consists of unique titles, the mode will not be meaningful unless some films are listed multiple times due to re-releases or different versions.
- Frequency Distribution - We can analyze the frequency of films if there are any duplicates, like a film appearing in different weeks due to being re-released or still in theaters. This could give insight into films that have had long-lasting appeal or are being marketed differently over time. If a film appears multiple times, it can indicate extended performance over weeks, popularity spikes after re-releases, or different versions (e.g., special editions, director's cuts).
- Categorical Grouping and Aggregation - By grouping films, we can aggregate associated numerical data like ‘Weekend Gross’, ‘Total Gross’, or ‘Number of cinemas’. For example: Sum or average the ‘Total Gross’ for each film to compare their overall box office performance. Check how many films were distributed by different distributors. Grouping and aggregating the film titles can highlight top-grossing films or those with the widest cinema distribution.
- Word Cloud or Word Frequency Analysis - For an exploratory view, a word cloud could be generated from the film titles to visualize common words in film names (e.g., frequent use of words like "Star," "War," etc.). This could provide fun, high-level insights into trends or common themes in film naming.
- Comparative Analysis - We can compare the performance of films based on other columns like ‘Total Gross’ or ‘Number of cinemas’. You could identify the highest-grossing films and compare them to lower-performing ones to find potential factors influencing their success (e.g., country of origin, distributor).
- Why Traditional Measures Aren’t Applicable - Mean, median, range, variance, standard deviation, etc.: These measures require numerical data, and since the ‘Film’ column consists of text (film titles), these statistical methods cannot be applied.
- Summary of Insights for the *Film* Column
- Aggregating data by the film can reveal patterns in gross revenue, theatre distribution, or release duration. Frequency analysis may highlight films with longer runs or multiple releases. Word clouds or other visualizations can show trends in film titles.

3. Weekend Gross :-

- The Weekend Gross column represents the revenue a film earned over a weekend. Since this is numerical data, various measures of central tendency, dispersion, and distribution can be applied to understand the performance of films in terms of box office revenue.
- The mean weekend gross of approximately 173,889 suggests that, on average, movies are generating significant revenue over the weekend. However, given the other measures, this average appears to be influenced by a few high-performing outliers.
- The median is substantially lower than the mean, which indicates a highly skewed distribution. This suggests that while most movies earn around 23,669 over the weekend, a few movies are making significantly more, pulling the average up.
- The mode, being very low compared to both the mean and median, highlights that the most frequently occurring weekend gross is only 36. This stark difference suggests that many movies have low earnings, while only a few manage to make high amounts.
- The mid-range, calculated as the average of the maximum and minimum values, is extremely high. This further supports the presence of extreme values (outliers) that heavily influence the data distribution, indicating the presence of blockbuster movies making exceptionally high weekend grosses.
- Range (2,185,536.0) : The very large range suggests an enormous disparity in weekend grosses among different movies, indicating that the performance of movies varies dramatically. Some movies make almost no money, while others make millions.
- Variance (148,256,463,481.14) and Standard Deviation (385,040.86) : The extremely high variance and standard deviation reflect a broad spread of weekend gross values, confirming the significant variability in movie performance. This high variability suggests that predicting weekend grosses would be challenging without further segmentation (e.g., by genre or budget).
- Mean Deviation (226,486.04) : A high mean deviation shows that individual weekend grosses are on average far from the mean, reinforcing the idea of high dispersion and inconsistency in movie earnings.
- Quartile Deviation (IQR: 114,287.5): The IQR value indicates that the middle 50% of movies have weekend grosses that are still quite spread out, albeit much less than the overall data. This spread in the central range still points to varying performance among typical movies, without considering the extreme outliers.
- **Highly Skewed Data:** The large difference between the mean and median highlights a right-skewed distribution, indicating that while most movies have relatively modest earnings, a few blockbusters earn substantially more.
- **Presence of Outliers**: The high mid-range, range, variance, and standard deviation confirm the presence of outliers (exceptionally high-grossing movies) that significantly affect the overall average.
- **Disparities in Performance**: The data reveals significant disparities in weekend grosses, likely driven by factors such as marketing budget, franchise strength, star power, and release timing.
- **Insights for Strategic Decision-Making**: Identifying the characteristics of high-performing movies (e.g., genre, audience appeal) can help studios and distributors strategize future releases, potentially improving the overall weekend gross.
- **Q-Q Plots Analysis :** Highly skewed and does not follow a normal distribution.

Graphs :-





4. Distributor :-

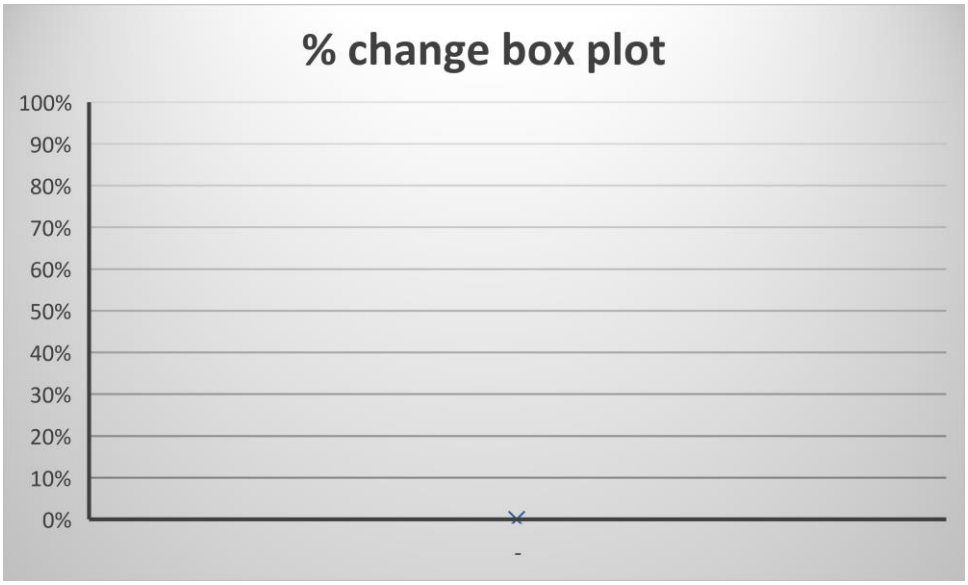
- The Distributor column contains the distributors of the movies, which is categorical data. As such, measures of central tendency like mean, median, mode, and other statistical measures (variance, standard deviation, etc.) are not applicable because these measures require numerical or ordinal data.

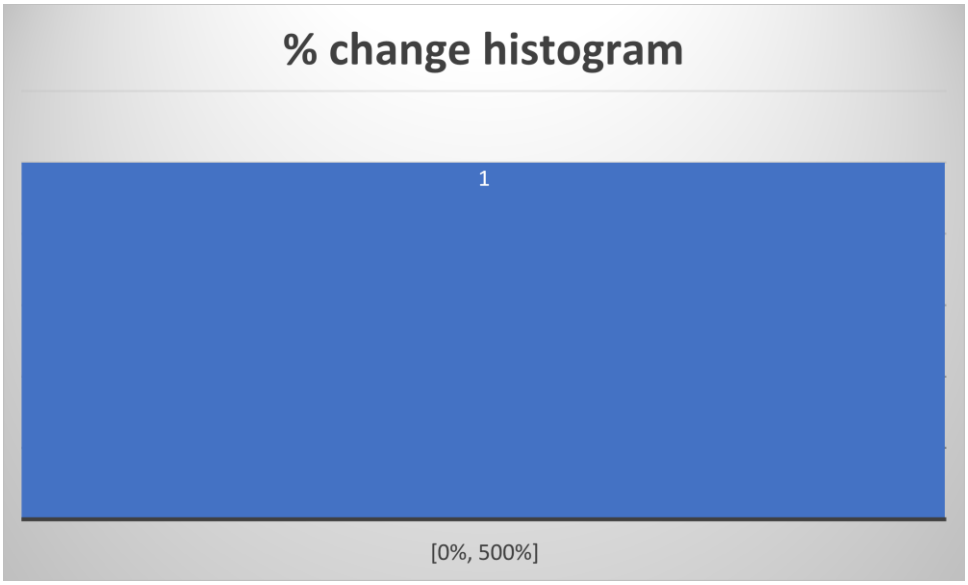
5. % change on last week :-

- The %Change column likely represents the percentage change in box office revenue (or another metric) from one weekend to the next. This is a useful metric for assessing how a film's performance is evolving over time—whether it is improving or declining in terms of weekend grosses.
- The mean value of 0.14 suggests a small average increase in performance from the previous week, indicating that, on average, movies slightly improved their earnings week over week. However, the small mean value close to zero also implies a balance between movies that gain and lose revenue week to week.
- The median of -0.405 shows that more than half of the movies experienced a decline in their earnings compared to the previous week. This indicates that a typical movie tends to lose revenue as its run progresses, a common pattern in box office performance.
- The mode, being the most frequent value at -0.78, suggests that the most common scenario is a significant drop in earnings compared to the prior week. This aligns with the typical behaviour of movies as audience interest wanes after initial screenings.
- The mid-range value of 7.56 suggests that there are extreme cases of both massive gains and losses, which indicates highly volatile changes in revenue from week to week. This wide value range suggests that certain movies either recover dramatically or crash unexpectedly.
- The large range value of 17.09 reflects the substantial disparity in week-over-week performance changes. It implies that while some movies experience significant revenue drops, others might achieve exceptional gains, possibly due to factors like expanding to more cinemas, marketing pushes, or holiday effects.
- **Variance (6.01) and Standard Deviation (2.45)**: A relatively high variance and standard deviation compared to the mean indicate that the week-over-week changes in performance are highly variable. This suggests that the revenue trajectory of movies is unpredictable, with many experiencing sharp increases or decreases rather than gradual changes.
- The mean deviation of 1.05 shows that the changes in percentage from the previous week deviate quite significantly from the mean, highlighting the inconsistent nature of box office performance week to week.
- The interquartile range of 0.55 indicates that the middle 50% of movies have relatively consistent performance changes compared to the entire range. However, this still points to noticeable variability even within the central half of the data.
- **Dominant Decline in Performance**: The negative median and mode values indicate that most movies tend to lose revenue over time, which is typical as initial interest wanes and newer releases take center stage.
- **High Variability**: The large range and high standard deviation reflect a highly volatile box office, where some movies manage to buck the trend with significant gains, while others drop sharply.
- **Insights for Scheduling and Marketing**: Understanding these patterns can help studios plan marketing efforts and release schedules. For example, sudden gains could be tied to promotional pushes, critical acclaim, or expansions into new theaters.

- **Impact of External Factors**: Drastic week-over-week changes could be influenced by external factors such as holidays, competing releases, or awards buzz, which can cause spikes or drops in performance.
- **Q-Q Plots Analysis**: Skewness and deviations, particularly in the tails.

Graphs :-

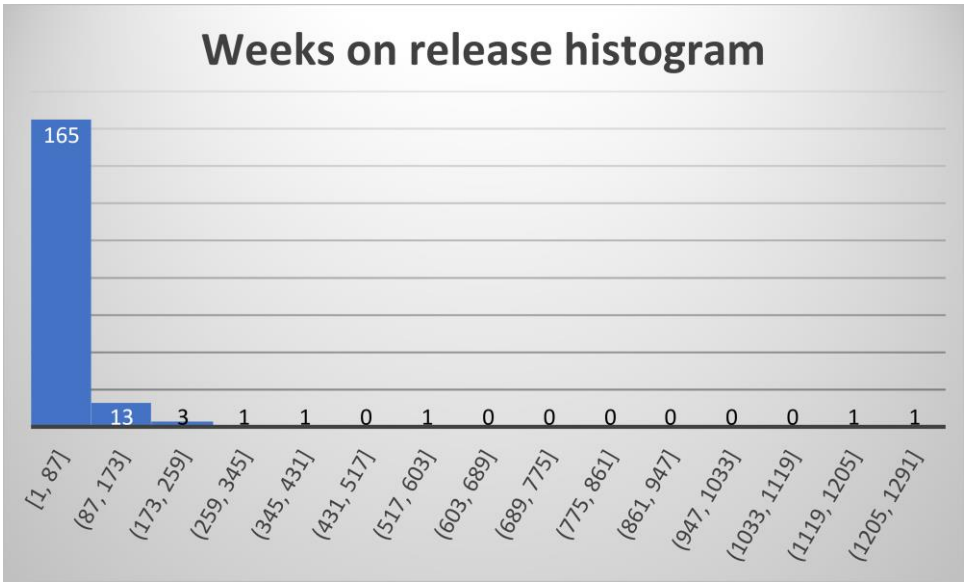




6. Weeks on release :-

- The 'Weeks on Release' column represents the number of weeks a film has been in theaters. This is a numerical column that reflects how long a movie has been running, and various measures of central tendency, dispersion, and other analyses can be applied. Here's an analysis of how these measures work for the 'Weeks on Release' column and the insights they provide.
- The mean of 33.39 weeks suggests that, on average, movies are staying in theaters for a considerable amount of time. However, given the presence of a much lower median, this average is likely skewed by a few long-running movies.
- The median value of 5 weeks shows that half of the movies are in theaters for only about 5 weeks. This indicates that a typical movie run is relatively short, with most movies exiting theaters soon after their initial release.
- The mode of 2 weeks is the most frequently occurring value, indicating that many movies have a very short lifespan in theaters, often leaving after only a couple of weeks. This highlights the competitive nature of the box office, where most films quickly make way for new releases.
- The mid-range value of 288.5 weeks suggests the presence of extreme values, where some movies run for exceptionally long periods. This could include special re-releases, cult classics, or movies with a significant fan base that keep them in theaters much longer than usual.
- The enormous range of 573 indicates a substantial disparity in how long movies stay in theaters. While most films have short runs, a few stay for extended periods, which heavily skews the overall data.
- **Variance (6,672.92) and Standard Deviation (81.69)**: The high variance and standard deviation reflect a wide spread of data around the mean, underscoring the significant differences in release durations. This suggests that while some movies leave theaters quickly, others, often due to popularity or special circumstances, remain for much longer.
- The mean deviation of 46.57 weeks shows that there is considerable variability in how long movies remain on release compared to the average duration.
- The relatively low interquartile range (IQR) of 8.25 weeks suggests that the middle 50% of movies have a more consistent, though still variable, release duration compared to the overall data. This indicates that the core group of movies doesn't stay in theaters nearly as long as the mean might suggest.
- **Short Typical Run**: The low median and mode values indicate that the typical movie stays in theaters for a brief period, emphasizing the rapid turnover in the film industry, where only a few movies have extended runs.
- **Outliers Driving the Mean**: The much higher mean, driven by a few long-running movies, reflects outliers that can distort the perception of average performance. This can include re-releases, holiday specials, or films that gain momentum over time.
- **High Variability**: The wide range and high standard deviation illustrate the unpredictability of a movie's theater run, with some movies finding extended success and others exiting quickly due to poor performance or competition.
- **Insights for Release Strategies**: Understanding the typical short duration of most movie runs can help studios and theaters optimize scheduling, promotional efforts, and strategic planning to maximize box office returns during the crucial early weeks of a film's release.
- **Q-Q Plots Analysis:** Not normally distributed, with notable skewness.

Graphs :-



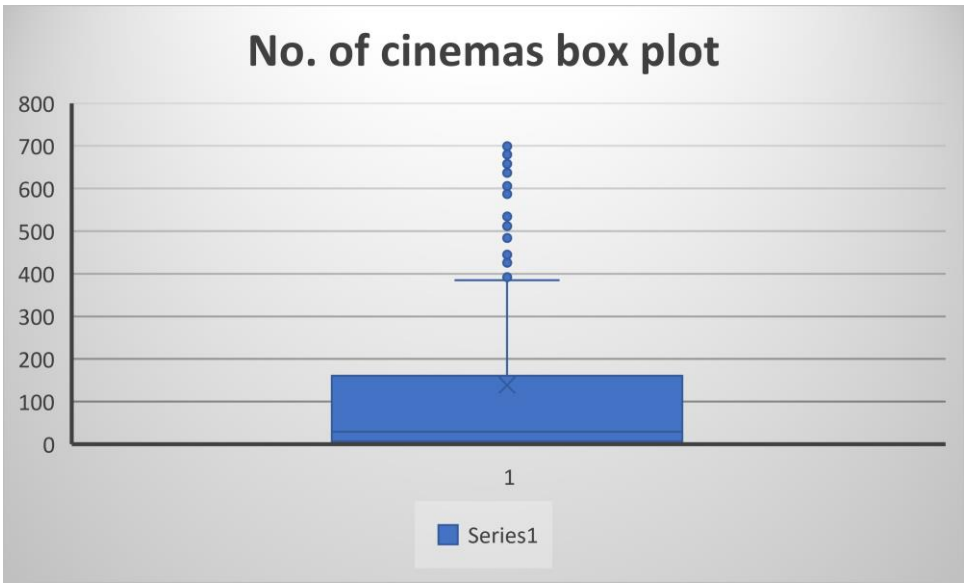
7. Number of Cinemas :-

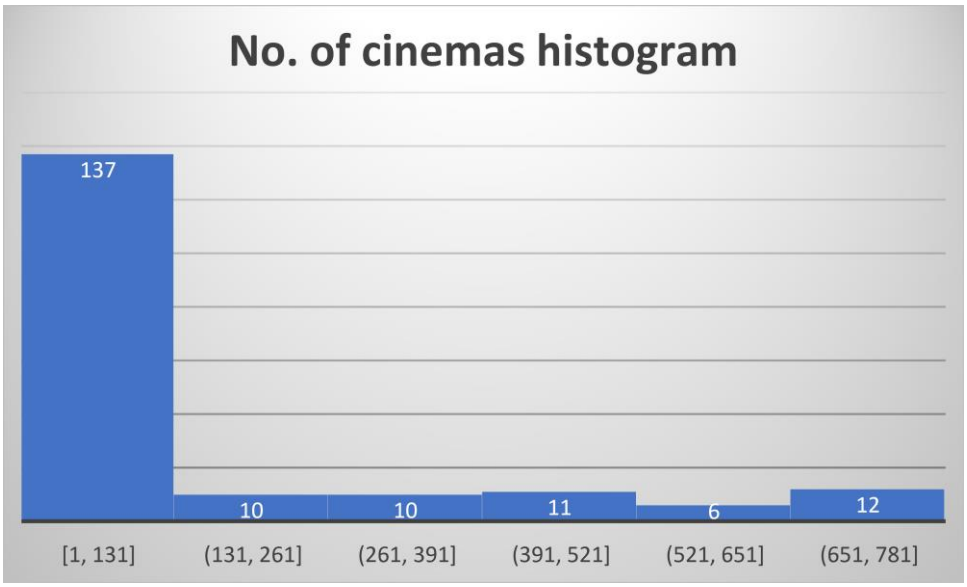
- The 'Number of Cinemas' column represents the number of theaters or cinemas in which a film is currently showing. This is a crucial factor for understanding the distribution and availability of a film, which directly impacts its box office performance. Here's how various measures of central tendency, dispersion, and other statistical analyses apply to this column and the insights they provide.
- The mean of 194.51 cinemas indicates that, on average, movies are shown in about 195 theaters. This average suggests a reasonable distribution reach for many movies, but the

large difference between the mean and other measures indicates variability influenced by extreme values.

- The median value of 61 cinemas shows that half of the movies are shown in 61 or fewer theaters. This is significantly lower than the mean, highlighting that the majority of movies have a more limited release, with only a few movies enjoying much wider distribution.
- The mode of 1 suggests that the most common scenario is for movies to be shown in only one cinema, possibly indicating limited or special screenings, niche films, or smaller releases with targeted audiences.
- The mid-range of 356 indicates that the distribution reach varies widely, with some movies being shown in only a few cinemas and others enjoying very broad releases. This number reflects the presence of both limited releases and nationwide rollouts.
- The wide range of 710.0 shows the substantial difference in the number of cinemas where movies are shown, highlighting the disparity between the least and most distributed films. This variability suggests a broad spectrum of release strategies, from small independent films to major blockbuster releases.
- **Variance (55,443.10) and Standard Deviation (235.46)**: The high variance and standard deviation values indicate significant dispersion around the mean. This large spread confirms that some movies are released in a vast number of theaters, while others have very limited exposure, contributing to the overall variability.
- **Mean Deviation (201.24)**: The mean deviation shows that the average deviation from the mean is quite high, reinforcing the observation that the data points are widely spread around the average, with many movies differing significantly in their number of screens.
- **Quartile Deviation (IQR: 366.75)**: The very high IQR suggests that even among the middle 50% of movies, there is considerable variability in the number of cinemas. This indicates that distribution is not consistent even for the central majority of films, reflecting diverse release strategies.
- **Highly Skewed Distribution**: The lower median and mode compared to the mean highlight a right-skewed distribution, where most movies are released in a relatively small number of cinemas, and a few are widely released.
- **Presence of Limited and Wide Releases**: The data suggests that most movies do not receive wide distribution, likely due to factors such as marketing budgets, target audiences, and distributor resources. On the other hand, the few movies that are widely released can significantly affect overall averages.
- **High Variability in Distribution**: The large range and high standard deviation emphasize the stark differences in release strategies, with some movies only playing in a handful of cinemas and others reaching hundreds. This reflects a mix of limited indie releases and major studio blockbusters.
- **Strategic Insights for Film Distribution**: Understanding this variability can help studios and distributors decide on the most effective release strategy based on the movie's potential audience size, expected demand, and competition. For example, a limited release strategy may build word-of-mouth for indie films, while wide releases maximize initial audience reach for blockbusters.
- **Q-Q Plots Analysis:** Some deviations from normality, indicating skewness.

Graphs :-





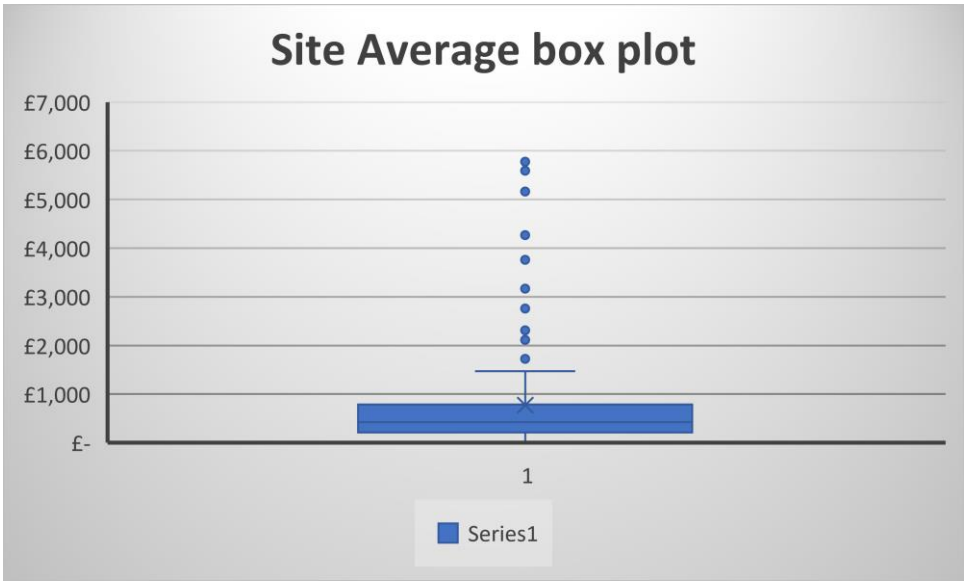
8. Site Average :-

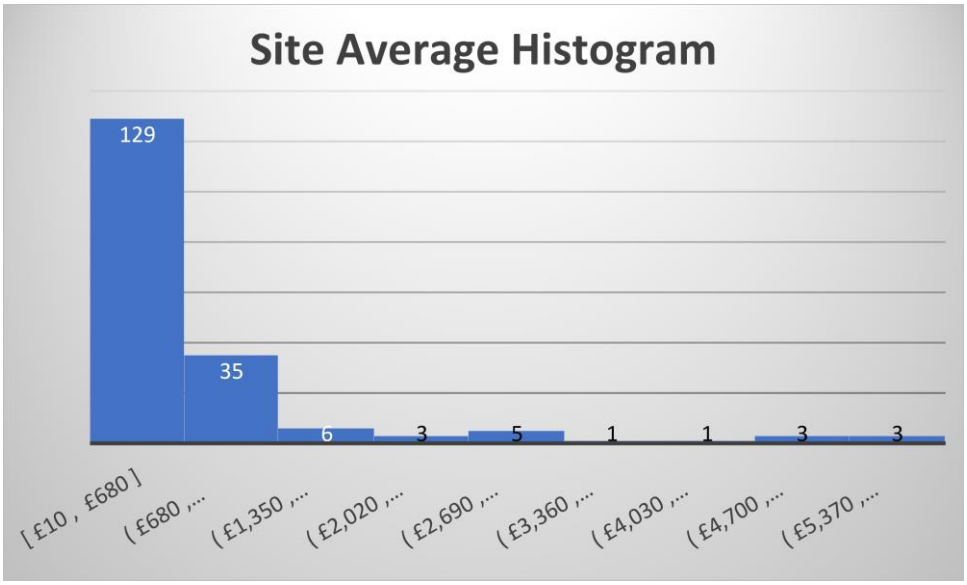
- The mean value of 795.02 indicates that, on average, each cinema's site performance is moderately high. This suggests a decent overall turnout per movie, but it's important to consider this in the context of the wide variability observed in other measures.
- **Median (464.5):** The median is significantly lower than the mean, which indicates that the distribution of site averages is skewed to the right (positively skewed). This suggests that there are a few very high-performing cinemas pulling the mean upward, while the majority are performing closer to the median.
- **Mode (325.0) :** The mode is even lower than the median and mean, showing that the most common site average is relatively low compared to the overall average. This further

reinforces the idea of a positively skewed distribution, where the bulk of cinemas perform below the mean.

- **Mid-Range (2,826.0) :** The mid-range, calculated as the average of the maximum and minimum values, is very high. This indicates that there are extreme values (outliers) in the data set—some cinemas have extremely high site averages, contributing to the skewness.
- **Range (5,594.0) :** A large range suggests significant disparities in cinema performance, indicating that some cinemas are doing exceptionally well while others are performing poorly. This disparity may be due to differences in cinema location, size, audience demographics, or other operational factors.
- **Variance (1,079,025.57) and Standard Deviation (1,038.76) :** The high variance and standard deviation values indicate a wide spread of data points around the mean. This confirms that cinema performance varies greatly, with significant deviations from the average.
- **Mean Deviation (624.44) :** This measure reflects a high average deviation from the mean, further illustrating the variability in site performance.
- **Quartile Deviation (IQR: 499.75) :** The IQR of 499.75 indicates that the middle 50% of the data points are relatively close to each other compared to the overall range. However, when compared to the mean and other measures, it suggests that the real variability lies outside this central range, driven by outliers.
- **Q-Q Plots Analysis :** Deviations, especially in the tails.
- The large difference between the mean, median, and mode highlights a distribution that is heavily influenced by a few high-performing cinemas, suggesting that a minority of sites are driving up the overall average.
- The high range and variance suggest the presence of outliers, meaning performance is not consistent across all cinemas.
- Stakeholders might need to investigate why certain cinemas perform significantly better than others, potentially identifying best practices, marketing strategies, or location advantages that could be replicated elsewhere.
- Understanding these disparities can help in tailoring strategies for underperforming cinemas and leveraging the success factors of high performers.
- These insights provide a detailed understanding of site performance, highlighting areas where intervention or strategic adjustments could improve overall outcomes.

Graphs :-





9. Total Gross to Date :-

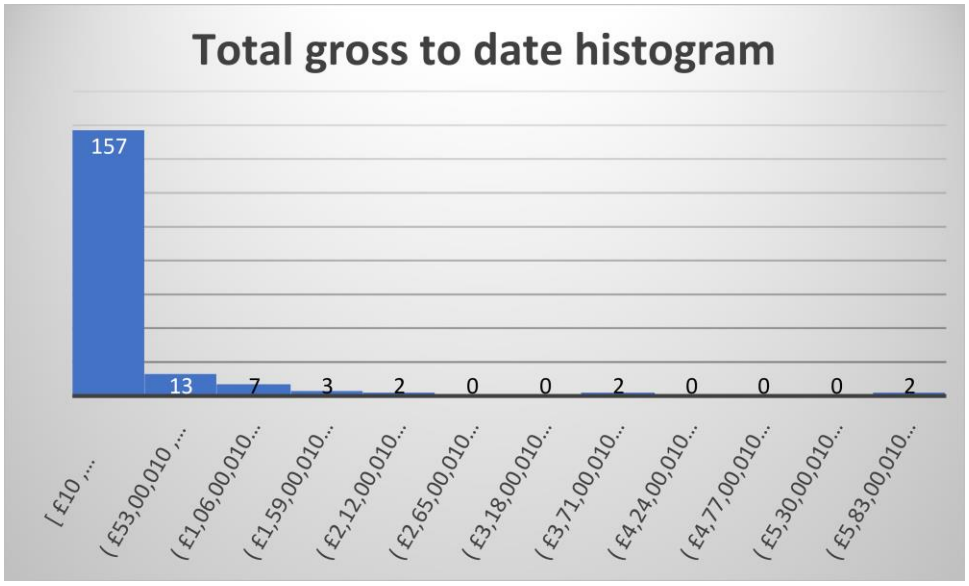
- The mean total gross of approximately 6.54 million suggests that, on average, movies are generating substantial revenue. However, the mean is likely skewed by a few extremely high-grossing movies, given the large disparity with the median.
- The median value of around 1.43 million indicates that half of the movies have grossed this amount or less, suggesting that the typical movie earns significantly less than the average. This highlights the presence of a few very successful movies driving up the overall mean.
- The median value of around 1.43 million indicates that half of the movies have grossed this amount or less, suggesting that the typical movie earns significantly less than the

average. This highlights the presence of a few very successful movies driving up the overall mean.

- The mode, being just 25,568, shows that the most frequently occurring gross is relatively low, suggesting that many movies have underperformed at the box office compared to their peers. This reflects the common occurrence of films that achieve limited financial success.
- **Mid-Range (31,763,192.5)** : The mid-range, which averages the highest and lowest total grosses, is very high, emphasizing the presence of extreme values on both ends of the spectrum. This reinforces the idea of a few standout blockbusters and some films with very low earnings.
- **Range (63,475,249.0)**:The enormous range indicates a vast difference between the lowest and highest-grossing movies, pointing to the variability in movie success. This large disparity suggests that while some movies perform exceptionally well, others struggle to make a significant impact.
- **Variance (142,084,746,029,530.56) and Standard Deviation (11,919,930.62)**: The extremely high variance and standard deviation reflect a vast spread in total grosses, confirming that there is considerable variability among movie earnings. This wide dispersion underscores the unpredictability of box office performance, with a few movies generating exceptionally high revenue while many earn much less.
- **Mean Deviation (7,452,271.56)** :A high mean deviation shows that individual total grosses deviate substantially from the mean, highlighting significant inconsistency in the financial performance of movies.
- **Quartile Deviation (IQR: 7,496,386.5)** : The high interquartile range (IQR) indicates that even within the central 50% of movies, there is considerable variation in total earnings. This suggests that the middle range of movies still experiences a broad spectrum of success.
- **Q-Q Plots Analysis** : Highly skewed with large deviations.
- **Highly Skewed Distribution with Outliers**: The large gap between the mean and median, along with the high range and standard deviation, suggests a right-skewed distribution. A few blockbuster movies earn substantial revenue, while the majority have much lower total grosses.
- **Presence of Highly Successful Movies and Underperformers**: The data reflects the common "winner-takes-all" scenario in the movie industry, where a small number of films achieve exceptional financial success, dramatically impacting overall averages.
- **High Variability in Movie Earnings**: The significant spread in total gross values points to the challenges and unpredictability of box office performance. Factors such as genre, star power, marketing, and competition play crucial roles in determining a movie's success.
- **Strategic Implications**: Studios should be aware of the high variability and potential for both high returns and losses. Investment in high-quality content, targeted marketing, and strategic release timing can help mitigate risks and capitalize on potential success.
- **Target Audience and Niche Markets**: For smaller or niche films, managing expectations and focusing on targeted marketing could help maximize earnings, even if not reaching blockbuster levels. Understanding these insights allows for better planning and resource allocation.

Graphs :-


Total gross to date box plot
£7,00,00,000
£6,00,00,000
£5,00,00,000
£4,00,00,000
£3,00,00,000
£2,00,00,000
£1,00,00,000
£-
1
Series1


Total gross to date histogram
157
13
7
3
2
0
0
2
0
0
0
2
[£10 ,...
(£53,00,010 ,...
(£1,06,00,010...
(£1,59,00,010...
(£2,12,00,010...
(£2,65,00,010...
(£3,18,00,010...
(£3,71,00,010...
(£4,24,00,010...
(£4,77,00,010...
(£5,30,00,010...
(£5,83,00,010...

Conclusion:-

The analysis of the given dataset using measures of central tendency and variability provides valuable insights into the performance of movies across various metrics such as rank, weekend gross, percentage change on last week, weeks on release, number of cinemas, site average, and total gross to date. The rank data showed a skewed distribution with most movies ranking lower, indicating intense competition. The high variability suggests that while a few movies consistently perform well, most struggle to maintain a high position, emphasizing the highly dynamic nature of box office performance. The weekend gross values highlighted significant disparities in earnings, with a few blockbusters driving up the average. The high range and standard deviation underline the unpredictable nature of opening weekend performances, often impacted by factors such as marketing effectiveness, competition, and audience reception. Analysis of week-over-week changes showed that most movies experience a decline in earnings after their initial release. The high variability suggests a mix of factors influencing these changes, including market saturation, competing releases, and changes in audience interest, which collectively underscore the volatility of movie earning. The distribution of weeks on release demonstrated that most movies have a short theatrical lifespan, with only a few exceptions staying longer due to special circumstances. The site average data showed high variability, reflecting diverse performance across different cinemas. This suggests that localized factors such as demographics, cinema location, and competing local events can significantly impact how well a movie performs in specific theaters. The total gross earnings were highly skewed, with a small number of high-grossing films inflating the average. The analysis underscores the highly competitive and unpredictable nature of the movie industry, where performance can vary drastically based on numerous factors including release strategy, competition, marketing, and audience reception. For studios, understanding these dynamics is crucial for optimizing release schedules, tailoring marketing efforts, and making informed decisions about distribution to maximize box office potential. The data emphasizes the importance of focusing on early earnings and identifying strategies to prolong a movie's run in theaters to enhance total revenue. These insights can help guide future decision-making, improve market positioning, and ultimately enhance the financial performance of movies in a rapidly evolving entertainment landscape.

References :-

1. Practical Statistics for Data Scientists by Peter Bruce, Andrew Bruce and Peter Gedeck.
2. The Elements of Statistical Learning by Trevor Hastie, Robert Tibshirani and Jerome Friedman.
3. Geeks for Geeks.
4. ChatGPT4.o
5. Referred Digicampus slides to understand the concept.