

# FOUNDATION OF STATISTICS AND PROBABILITY

## CAPSTONE PROJECT

Name – Manuj Mehta

Roll no: B24BS2194

Batch- 02\_Sem01

Date of Submission: Friday, September 12, 2025

## OBJECTIVE

**Video Game Review-** The objective of this project is to analyze player behavior and review scores using data visualization and correlation techniques. The project aims to uncover patterns in gameplay hours and their impact on game ratings to provide insights into player engagement and satisfaction."

# ACKNOWLEDGMENT

## 1. Acknowledgment to Dr. Rajlaxmi Chouhan

We extend our sincere gratitude to Dr. Rajlaxmi Chouhan, our respected professor of Foundations of Statistics and Probability, for her invaluable guidance. Her recorded lectures greatly clarified our concepts and provided us with a strong foundation for learning. Her dedication and efforts ensured that we could revisit the material and grasp the subject with greater confidence.

## 2. Acknowledgment to Mr. Ankit Chouhan and Ms. Nikita Shrivastav

We are deeply thankful to our teaching assistants, Mr. Ankit Chouhan and Ms. Nikita Shrivastav, for their consistent support and encouragement. They not only enhanced our understanding during live sessions but also patiently resolved our queries whenever needed. Their commitment played a vital role in strengthening our knowledge and building our confidence in the subject.

## 3. Acknowledgment to Dr. Afshan Khan and Prof. Abhishek

Our heartfelt thanks go to Dr. Afshan Khan and Prof. Abhishek, who enriched our learning by introducing us to real-world challenges. Their insightful discussions helped us connect theoretical concepts to practical applications, showing us how knowledge can be transformed into solutions. Their guidance broadened our perspective and motivated us to approach problems with analytical and innovative thinking.

# DATA SET OVERVIEW:

The Video Game Review Dataset contains a total of 60 records representing individual players. Each record consists of three attributes: Player ID, Hours Played, and Review Score. The Player ID is a categorical variable uniquely identifying each player. The Hours Played and Review Score are continuous variables that provide insights into gameplay duration and player satisfaction, respectively. The dataset is designed to analyze how many hours players spend on the game and the ratings they assign to the game.

This dataset allows for exploration of player engagement patterns and the relationship between gameplay duration and player satisfaction.

1	Player ID	Hours Played	Review Score		
2	Player_1	11	67		
3	Player_2	24	91		
4	Player_3	19	100		
5	Player_4	15	92		
6	Player_5	12	82		
7	Player_6	11	89		
8	Player_7	23	70		
9	Player_8	15	64		
10	Player_9	15	88		
11	Player_10	8	75		
12	Player_11	12	67		
13	Player_12	7	71		
14	Player_13	6	69		
15	Player_14	16	81		
16	Player_15	10	86		
17	Player_16	6	89		
18	Player_17	5	89		
19	Player_18	16	88		

## TOOLS USED:

- Google Sheets: Used for data entry, calculation of summary statistics, quartiles, interquartile range (IQR), creation of scatter plots, and box plots for data visualization.
- Microsoft Word: Used to organize, write, format, and compile the final report, including embedding tables and images generated in Google Sheets.

## BASIC STATISTICS:

Statistical Measure	Hours Played	Review Score
Count	60	60
Mean	16.33333333	80.615
Median	16	80.5
Mode	16	89
Minimum	5	63.3
Maximum	45	100
Standard Deviation	7.710144466	9.913931731
Variance	58.45555556	96.64794167

The summary statistics show that players, on average, played about 16.33 hours and gave a review score of approximately 80.62, indicating generally favorable experiences. Both variables have moderate variability, with review scores being slightly more spread out (standard deviation ~9.91) than hours played (standard deviation ~7.71), but most gameplay and ratings cluster around the central values. The mode of hours played matches the median (16), suggesting a common gameplay

duration, while the mode for review scores is higher (89), hinting that many players gave top ratings.

## UNDERSTANDING DATA SPREAD

To understand the spread of this data set we will plot a box plot using google sheets. Before plotting a box plot we need some manual calculations of percentile range called quartiles. One example is shown below:

By using formula

$$i = p/100 * (n+1)$$

i= rank, p = percentile position, n= total number of data points after sorting.

$$i = 25/100 * (60+1)$$

$$i = 61/4 = 15.25$$

Q1 or 25%ile lies between 15<sup>th</sup> & 16<sup>th</sup> position. Since both in this case is 11 so 25%ile of dataset is 11.

This is how we compute percentiles through manual calculation but computing percentile is quite easy in google sheets by using formula:

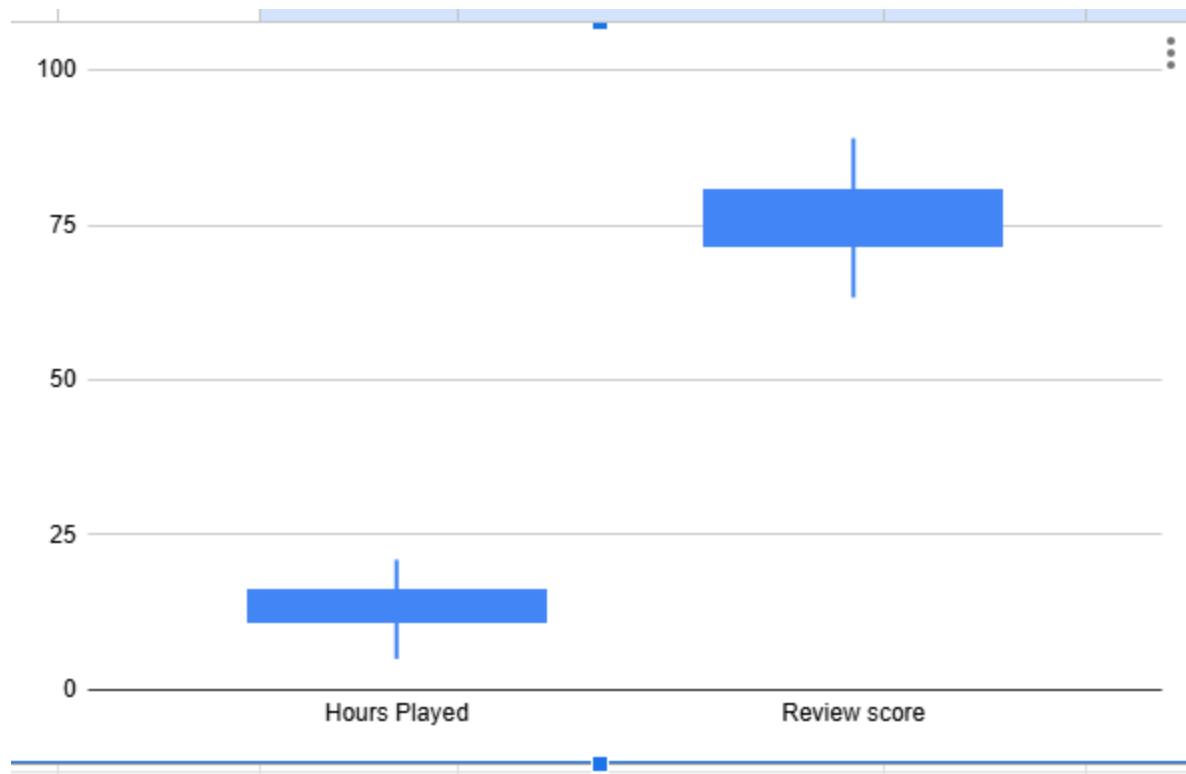
**=PERCENTILE.EXC (range: p)**

Statistical Measure	Hours Played	Review Score	
Minimum	5	63.3	
Q1	11	71.75	
Q2	16	80.5	
Q3	21	89	
Maximum	45	100	

Next step is to compute Interquartile Range which is” **IQR = Q3-Q1**”

Hours Played =  $21 - 11 = 10$

Review Score =  $89 - 71.75 = 17.25$



## Interpretation of Box Plot

### Hours Played

- The central box spans from Q1 (11) to Q3 (21), showing most players play between 11 and 21 hours.
- The median (16) is nearly centered, suggesting a symmetric distribution of playtime.
- The whiskers extend to the minimum (5) and maximum (45), but the clustering around the median means few players play extremely short or long hours.

### Review Score

- The box covers from Q1 (71.75) to Q3 (89), indicating most review scores fall in the higher range.

- The median (80.5) is visibly above the midway mark, suggesting generally high satisfaction.
- The shorter whiskers and higher values further confirm that most players rate the game positively, with scores tightly clustered above 70.

### Insights

- Both variables show moderate spread but a clear central tendency toward higher values.
- There are no significant outliers apparent from the box plot for either measure.

## CORRELATION: HOURS PLAYED V/S REVIEW SCORES

Since both variables are continuous variables, we can easily find their correlation using Pearson correlation method.

Computing correlation manually is a tough task but in google sheet it can be easily computed using formula:

=CORREL(x1:x2,y1:y2)

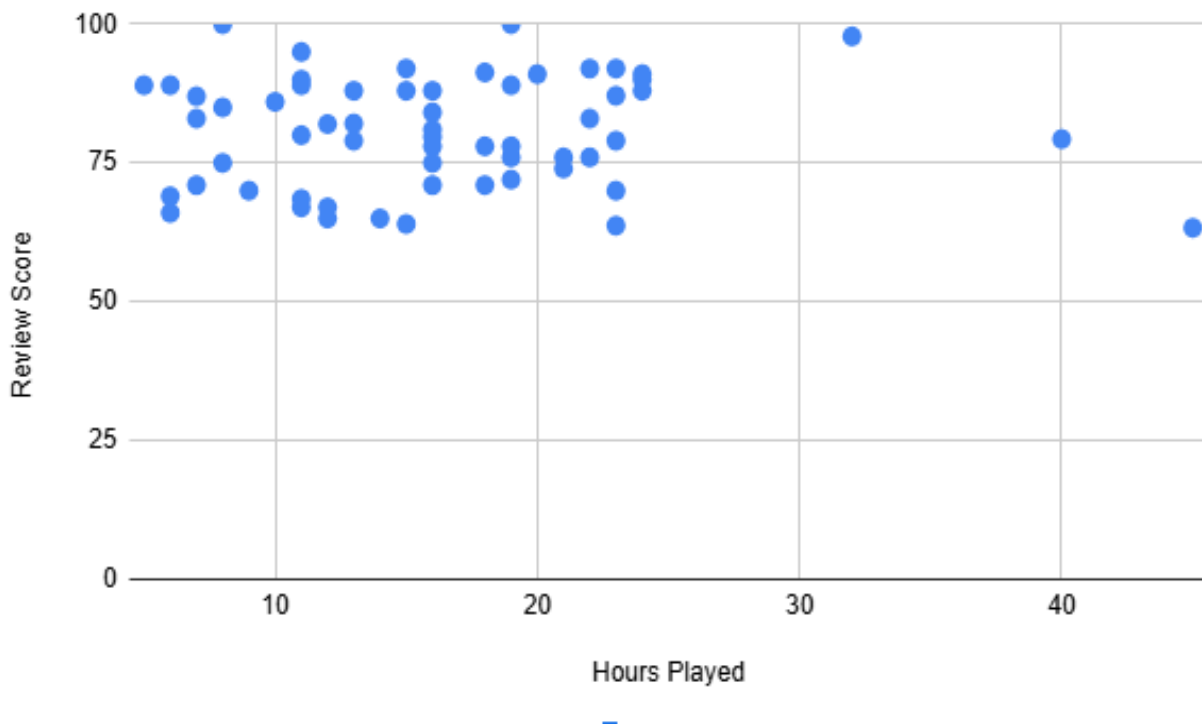
<b>Correlation Value</b>	<b>0.007272996734</b>
--------------------------	-----------------------

The computed Pearson correlation value between Hours Played and Review Score is approximately **0.0073**, which is very close to zero. Screenshot-502.jpg

This indicates that there is virtually **no linear relationship** between the amount of time players spend playing the game and the scores they give in their reviews. In other words, increased gameplay hours do not result in higher or lower review scores, and review scores remain relatively independent of how long someone has played. This suggests player satisfaction is not directly linked with time investment for this game.

However for more clearance we can use a scatter plot to visualize it.

Review Score vs. Hours Played



The scatter plot of Review Score vs. Hours Played shows that the data points are widely dispersed and no clear upward or downward trend is visible. Most review scores are clustered in the upper range (between 60 and 100) regardless of how many hours have been played. This visual result supports the calculated correlation value near zero, reinforcing that there is **no significant linear relationship** between how long a player spends in the game and how high or low they rate it. In summary, additional gameplay hours do not predictably affect a player's review score for this dataset.

## INSIGHTS

- The summary statistics and box plot show most players spend between 11 and 21 hours playing, with a median of 16 hours and play sessions ranging from 5 to 45 hours.
- Review scores are consistently high, with most ratings clustered between 72 and 89 and a median score of 80.5, indicating overall positive feedback about the game.
- The interquartile ranges for Hours Played (IQR = 10) and Review Score (IQR = 17.25) suggest moderate variability among both playtime and player ratings.
- The box plot visually supports these findings, showing compact distribution around central values, with no extreme outliers observed.
- Correlation analysis yields a coefficient of approximately 0.007, confirming a negligible linear relationship between Hours Played and Review Score.

## CONCLUSION

Player satisfaction, as measured by review scores, does not depend on gameplay duration in this dataset. The game receives mostly positive reviews regardless of how much time players spend playing, and there is no evidence of time-based rating bias. These results suggest other factors beyond time played may drive player satisfaction—for example, content quality, game features, or support.

All required visualizations and analysis have been performed, and the report's findings are supported by your data tables, box plots, and scatter plots.

**THANK  
YOU  
FOR  
YOUR  
ATTENTION**

