**Problem Set 1**

1. Inside the covariance matrix we see that the user score and critic score have a covariance over 100 meaning it has a strong positive relationship and with a correlation of 0.58 suggest that there is general agreement when it comes to assessments of video games. For the red values in the covariance matrix, are just variance values. We also notice that Na sales and Global sales in our covariance matrix is sitting at 1.81 and has a correlation of 0.95 suggesting that when a game sells well in North America, it tends to sell similarly well globally, and vice versa. The ‘1.81’ could mean that the unit is in millions or billions of dollars. A surprising statistic pair is the User score and global sales. We see a positive relation in the covariance matrix at ~2.5 but a very weak correlation at 0.088 indicates an almost negligible linear relationship between User Score and Global Sales. It indicates that, to some extent, as User Scores tend to increase, Global Sales also tend to increase, and when User Scores decrease, Global Sales tend to decrease. However, the magnitude of 2.5 is relatively small, indicating that the relationship is not very strong. Other factors, such as marketing, genre, platform, or regional preferences, could be influencing Global Sales more strongly than User Scores, leading to a weak linear correlation. In the correlation matrix it doesn’t mean anything because it’s correlating the same variable.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Cov Matrix** | Critic score | User score | NA sales | JP sales | Global sales |
| Critic score | 192.3372996 | - | - | - | - |
| User score | 115.8892026 | 207.343200 | - | - | - |
| NA sales | 3.1337726 | 1.195120 | 0.9358343 | - | - |
| JP sales | 0.5875144 | 0.528045 | 0.1303621 | 0.08269625 | - |
| Global sales | 6.4686776 | 2.497828 | 1.8154394 | 0.34642402 | 3.855108 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Cor Matrix** | Critic score | User score | NA sales | JP sales | Global sales |
| Critic score | 1 | - | - | - | - |
| User score | 0.5803184 | 1 | - | - | - |
| NA sales | 0.2335803 | 0.08579601 | 1 | - | - |
| JP sales | 0.1473139 | 0.12752142 | 0.46860735 | 1 | - |
| Global sales | 0.2375557 | 0.08834853 | 0.95579337 | 0.6135456 | 1 |

1. The 1st scatter plot shows a general pattern where, as the number of critics' reviews (Critic\_Count) increases, the Critic\_Score tends to vary widely. There is a cluster of games with varying scores where the Critic\_Count is relatively low, suggesting that less well-known games may have a wider range of critic ratings. However, as the Critic\_Count increases, the spread of Critic\_Score tends to narrow, indicating that more popular games tend to have a more consistent range of critic ratings. Outliers, if present, could represent games that received extreme scores compared to their review counts. In the 2nd scatter plot, there isn't a strong linear relationship between User\_Score and Global\_Sales. This suggests that user ratings alone do not predict global sales directly. There is a spread of data points across the entire range of User\_Scores, regardless of the level of Global\_Sales. Some games with high User\_Scores have relatively low Global\_Sales, while others with lower User\_Scores have high Global\_Sales. This implies that other factors, such as marketing, genre, or platform, may play a significant role in determining a game's global sales.

A graph of a bar chart

Description automatically generated with medium confidence

A graph of a graph

Description automatically generated with medium confidence

1. The values of the class attribute ‘user count’ have been computed as follows [0 – 100] → low, [101 – 1000] → medium, [1001 – 10665] → high. With that being said, for the low attribute histogram, was rightly skewed with the bin from 5-10 having a frequency over 1000. We notice as user count increases, the frequency increases. The exception to this is the bin ‘0-5’ to ‘5-10’ where there was an increase of 2x from 1 bin to another.

For the medium attribute histogram, was strongly right skewed where we see the vast majority of users concentrated in bin ‘101 – 200’ and seeing a drop over 2x to the next bin. This histogram follows the same trend as the previous histogram as well. We also notice that the number of bins has decreased because the variation has increased between the 2 sets. For the high attribute histogram was also strongly right skewed. We also notice that the number of bins is the same because the variation has increased between the low and high sets. But the Outliers are a bit more present in this set versus the medium attribute histogram. The first bin is more than 3x larger the next bin. The last 4 bins are also relatively small because games with high user count are going to be extremely rare and it entails that those are the best games that are being played. When analyzing the histogram for platform frequencies we see its skewed to the right as well. Majority of the platforms had less than 400 users.

A graph of a graph showing a number of levels

Description automatically generated with medium confidence

A graph of a graph

Description automatically generated

A graph of a graph

Description automatically generated

A graph of a graph

Description automatically generated

1. For the boxplot with User\_Score and GS\_Category. All GS\_Category attribute classes (low, medium, and high) had a lot of outliers which could be due to extreme user reactions especially since those outliers were low scores. The user scores across all the box plots landed in the lower half of the dataset meaning the sets were skewed to the left. The medians of the datasets were hovering between 70-80. Meaning users were having a higher satisfaction rating on games across all GS\_Categories. This is also the same in the critic score datasets where the median was also hovering around that range. In general, the critic score follows the same suit with skewness as well. Since the box plots across the critic and user score are very similar, we can conclude that there may be limited variation in the quality or user satisfaction among games in different GS\_Category classes. This might suggest that the GS\_Category classification does not strongly influence how games are received by critics and users. Also means quality and consensus: It could indicate that the perceived quality of games, as assessed by both critics and users, is relatively consistent across different categories. Games in all GS\_Category classes receive comparable ratings from both critics and users.

A group of graphs with text

Description automatically generated with medium confidence

A group of graphs with text

Description automatically generated with medium confidence

1. The relationship between Critic Score, User Score, NA Sales, and predicting Global Sales appears complex, with overlap among instances from different GS\_category classes.

Predicting Global Sales may be challenging using these individual attributes alone, as they do not provide clear separation between the three GS\_category classes.

A graph with different colored dots

Description automatically generatedOther factors not considered in these plots, such as game genre, platform, or marketing strategies, may also play a significant role in determining Global Sales.

A graph with red and blue dots

Description automatically generatedA screen shot of a graph

Description automatically generated

1. A graph of a user score

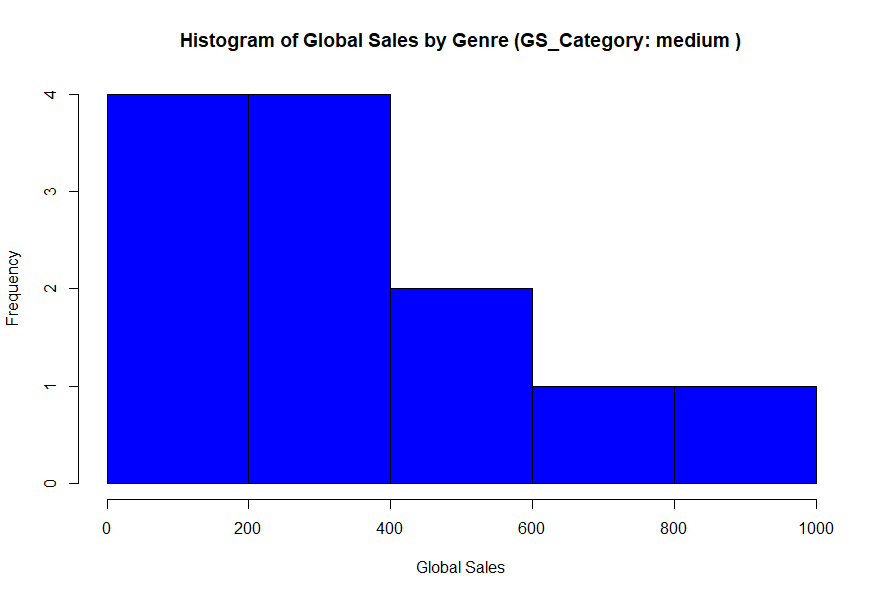
   Description automatically generatedboth plots are skewed to the left with the ‘high’ GS\_Category in both plots showing high proportion of where the data is concentrated meaning the games in those categories tend to have high critic scores. In the critic score plot, both ‘low’ and ‘medium’ GS\_Category curves are not smooth but rather mixed distribution suggesting that the data is not continuous meaning the scores from one title to the next can spike or dip in the beginning but after its peak, they smooth out.. This is the opposite in the user score plot where those categories are smoother. In the critic score plot, we see an overlap of all 3 categories in the mid 70s and also a similar range in the user score plot as we know that the average for those scores are in the 70s based of our boxplots.
2. A graph with different colored lines

   Description automatically generatedBased of the table and histograms we can see that the histograms are skewed to the left and we see breaks in the ‘low’ and ‘high’ genre global sales histograms. By examining the frequency counts in each cell of the table, you can identify which genres are more common or popular in each GS\_Category class. For example, you might find that action and adventure games are prevalent in the High GS\_Category, while sports games dominate the Low GS\_Category. Understanding genre distribution across GS\_Category classes can help in market segmentation. You might identify specific genres that appeal to different segments of the gaming market, allowing for targeted advertising and product development. In the medium global sales histogram, we see a wider range of values as high as 1000 whereas as the other 2 histograms cap out at 350 as games companies are hitting average sales. All 3 histograms also have a peak average between 50-100 global sales with the medium histogram having a tie with 0-50 at 5. We can see game genres across all GS categories most average no more than 100 (million) game sales.

|  |  |  |
| --- | --- | --- |
| "Genre" | “GS\_Category" | “Frequency |
| "Action" | "high" | 307 |
| "Adventure" | "high" | 16 |
| "Fighting" | "high" | 74 |
| "Misc" | "high" | 95 |
| "Platform" | "high" | 93 |
| "Puzzle" | "high" | 21 |
| "Racing" | "high" | 115 |
| "Role-Playing" | "high" | 123 |
| "Shooter" | "high" | 202 |
| "Simulation" | "high" | 52 |
| "Sports" | "high" | 194 |
| "Strategy" | "high" | 15 |
| "Action" | "low" | 347 |
| "Adventure" | "low" | 98 |
| "Fighting" | "low" | 69 |
| "Misc" | "low" | 62 |
| "Platform" | "low" | 83 |
| "Puzzle" | "low" | 46 |
| "Racing" | "low" | 143 |
| "Role-Playing" | "low" | 181 |
| "Shooter" | "low" | 211 |
| "Simulation" | "low" | 84 |
| "Sports" | "low" | 133 |
| "Strategy" | "low" | 144 |
| "Action" | "medium" | 976 |
| "Adventure" | "medium" | 134 |
| "Fighting" | "medium" | 235 |
| "Misc" | "medium" | 227 |
| "Platform" | "medium" | 227 |
| "Puzzle" | "medium" | 51 |
| "Racing" | "medium" | 323 |
| "Role-Playing" | "medium" | 408 |
| "Shooter" | "medium" | 451 |
| "Simulation" | "medium" | 161 |
| "Sports" | "medium" | 616 |
| "Strategy" | "medium" | 108 |

A graph of blue bars

Description automatically generated

A graph of blue rectangular bars

Description automatically generated with medium confidence

1. **Intercept (0.7775897)**: This is the estimated value of Global\_Sales when all the z-scored attributes (Year\_Z, Critic\_Score\_Z, Critic\_Count\_Z, User\_Score\_Z, User\_Count\_Z) are equal to zero. In this context, it represents the baseline level of Global\_Sales when all other factors are zero.

**Year\_Z (-0.1599845):** This coefficient represents the change in Global\_Sales for a one-unit change in the z-score of the "Year" attribute while holding all other variables constant. A negative coefficient suggests that as the year increases (indicating older games), Global\_Sales tend to decrease.

**Critic\_Score\_Z (0.2908886)**: This coefficient represents the change in Global\_Sales for a one-unit change in the z-score of the "Critic\_Score" attribute while holding all other variables constant. A positive coefficient indicates that higher z-scored critic scores are associated with higher Global\_Sales.

**Critic\_Count\_Z (0.3884601):** This coefficient represents the change in Global\_Sales for a one-unit change in the z-score of the "Critic\_Count" attribute while holding all other variables constant. A positive coefficient suggests that games with more critic reviews tend to have higher Global\_Sales.

**User\_Score\_Z (-0.1175949):** This coefficient represents the change in Global\_Sales for a one-unit change in the z-score of the "User\_Score" attribute while holding all other variables constant. A negative coefficient implies that higher z-scored user scores are associated with lower Global\_Sales.

**User\_Count\_Z (0.3333075):** This coefficient represents the change in Global\_Sales for a one-unit change in the z-score of the "User\_Count" attribute while holding all other variables constant. A positive coefficient suggests that games with more user reviews tend to have higher Global\_Sales.

1. Decided to create 3 decision trees using Critic\_Score, Critic\_Count, User\_Count, User\_Score against EU\_Sales, NA\_Sales, and JP\_Sales. The first formula used NA\_Sales followed by EU\_Sales then JP\_Sales. he model was trained using a subset of the dataset, train\_data, which consists of 80% of the original data. This is done to have a separate dataset for training and testing. After building these three decision trees using the training data, they were evaluated on a separate testing dataset (test\_data) to calculate their accuracy in predicting the GS\_Category. The accuracies were as follows: training accuracy: 0.8423077 and testing Accuracy: 0.8527473 for tree 1. Training accuracy: 0.6349817 and testing Accuracy: 0.592674 for tree 2. Training accuracy: 0.752381 and testing accuracy: 0.7377289 for tree 3. It appears that the first decision tree primarily relies on NA\_Sales for classification, with this attribute being the dominant factor in distinguishing between different sales categories. The roles of other attributes in this tree are either less relevant or not considered at all. Tree 2 has a pattern of starting with User\_Count as the most important decision, then moving to JP sales. Mind you these are all in the ‘medium’ GS\_Category. Then moves to User\_Count for high sales over 0.005. Then JP\_Sales becomes the next important decision. In regard to GS\_Category, JP\_Sales and User\_Count are the most important decisions. In Tree 3, EU sales is repeated for decision making until the we factor in sales that are less then 0.045 which leads to Critic Count then moves back to EU sales. Bear in mind that the sales number being used is most likely in the millions. The attributes that were chosen for the data set with the class variable ‘GS\_Categories’ show that NA\_Sales, JP\_Sales, and EU\_Sales have a large determining factor on how well a game will sell compared to the User count, User score, A diagram of a graph

   Description automatically generated with medium confidencecritic count and critic score.

A diagram of a company

Description automatically generated

A diagram of a tree

Description automatically generated with medium confidence

1. In this analysis, we explored a dataset of video game sales in 2016, aiming to forecast game popularity based on various attributes. We began by calculating covariance matrices to understand relationships among the attributes. Global\_Sales showed a strong positive correlation with NA\_Sales (0.96) and JP\_Sales (0.61), highlighting the significance of the North American and Japanese markets. However, Critic\_Score and User\_Score displayed weaker correlations with Global\_Sales (0.24 and 0.09, respectively), indicating that reviews have a moderate impact on sales. Platform choice also influences sales, as different platforms attract distinct user bases. Developers should optimize games for their target platforms to maximize sales. Additionally, the timing of a game's release can affect its success. Launching a game strategically, avoiding competition, and aligning with market trends can boost sales. Quality, assessed through Critic\_Scores and User\_Scores, is pivotal. High-quality games tend to garner positive reviews, building trust and reputation. Positive reviews can lead to word-of-mouth recommendations, enhancing sales. Diversification and staying informed about industry trends are vital for successIn conclusion, understanding regional markets, optimizing for platforms, and focusing on quality are crucial for predicting video game sales. Developers and individuals can use these insights to make informed decisions in the competitive gaming industry, increasing the chances of creating successful video games.