Roberto Jackson Baeza, 1861128
Sept. 23 2023
Data Science

1. Compute the covariance matrix for each pair of the following attributes: Critic_Score, User_Score, NA_Sales, JP_Sales, Global_Sales. Next, compute the correlations for each of the pairs of attributes. Interpret the statistical findings!

Covariance Table

	Critic_Score	User_Score	NA_Sales	JP_Sales	Global_Sales
Critic_Score	192.337	115.889	3.134	0.588	6.469
User_Score	115.889	207.343	1.195	0.528	2.498
NA_Sales	3.134	1.195	0.936	0.130	1.815
JP_Sales	0.588	0.528	0.130	0.083	0.346
Global_Sales	6.469	2.498	1.815	0.346	3.855

Correlation Table

	Critic_Score	User_Score	NA_Sales	JP_Sales	Global_Sales
Critic_Score	1	0.580	0.234	0.147	0.238
User_Score	0.580	1	0.086	0.128	0.088
NA_Sales	0.234	0.086	1	0.469	0.956
JP_Sales	0.147	0.128	0.469	1	0.614
Global_Sales	0.238	0.088	0.956	0.614	1

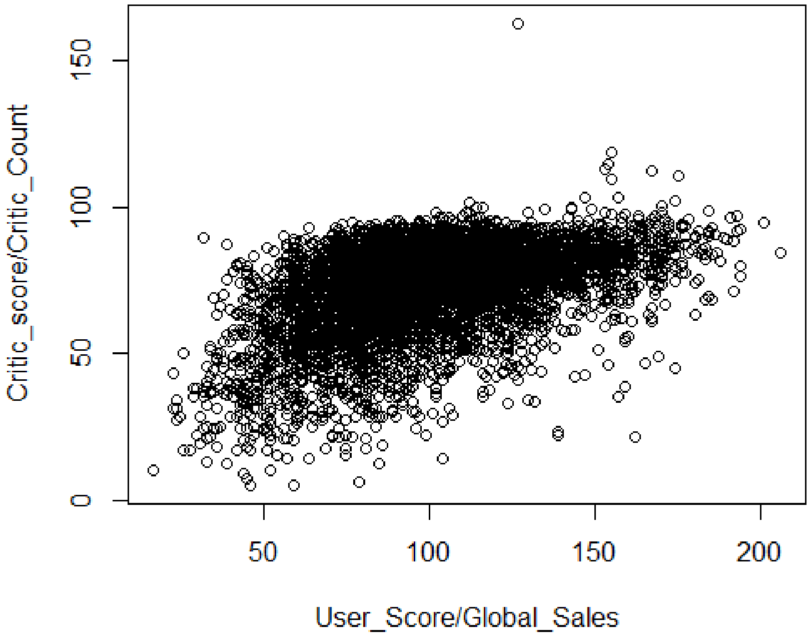
Covariance is a measure used to determine how much that two different variables change alongside each other. All of our variables have positive covariance with each other, this indicates that at the very least all of our variables increase. However most of this data has limitations because many of the variables have different scales.

Correlation is a direct measure of the linear relationship between two variables. Here we can see that obviously every variable has a perfect linear relationship with itself. Most of the others have weak to moderate relationships but a few have a strong relationship.

- Critic_Score / User_Score: with a positive Covariance value of 115.9 we can see that an increase of either Critic Score or User Score correlates to a very large gain of the other. This is backed by a correlation value of 0.58 which implies a moderate positive correlation between the two.
- Critic_Score / NA_Sales: at 3.1 an increase of either variable weakly correlates to an increase of the other. But this measure is less reliable because they have different scales. This is seen in the correlation value of .234 implying a trivial correlation.

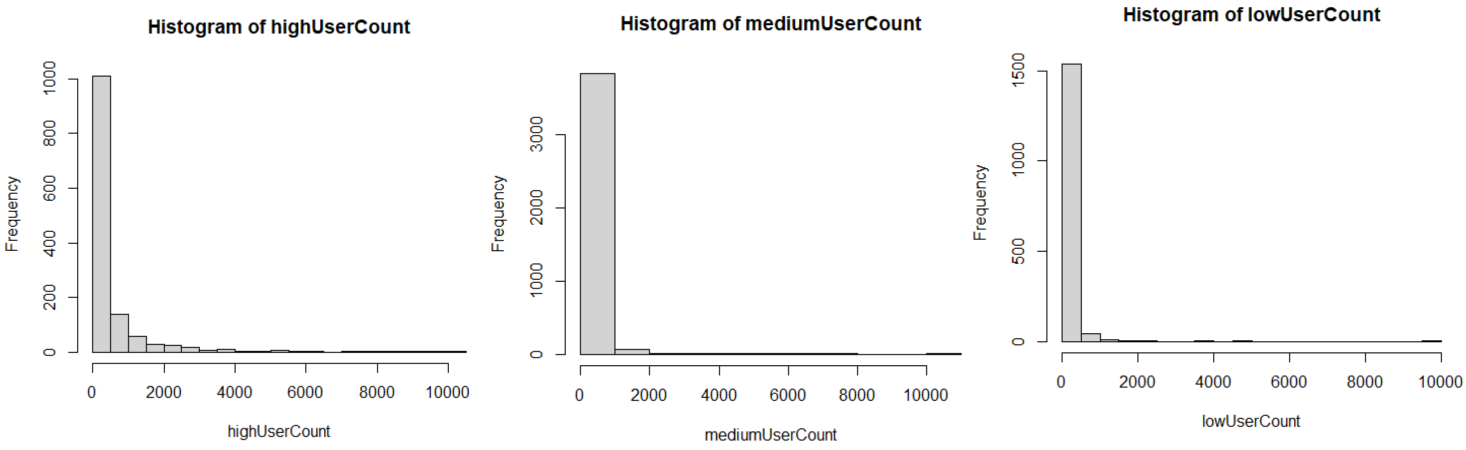
- Critic_Score / JP_Sales: At 0.588 there is a very weak covariance between the two and so no linear relationship is implied. This measure is less reliable because they have different scales. This is seen in the Correlation Value of .147 implying a trivial correlation.
- Critic_Score / Global_Sales: at 6.5 the two have a sizable positive covariance which implies a strong linear relationship. However this measure is less reliable because they have different scales. Which is seen in the correlation value of .238 implying a weak correlation between the two.
- User_Score / NA_Sales: at 1.1 a positive linear relationship is implied however this. measure is less reliable because they have different scales. This is seen in the Correlation value of .086 implying almost no relationship between the two.
- User_Score / JP_Sales: at .528 there is no linear relationship implied however the measure is less reliable because they have different scales. Seen in the Correlation value of .128 implying no relationship between the two.
- User_Score / Global_Sales: A covariance relationship of 1.8 implies a positive linear relationship between the two. The measure is less reliable because they have different scales which is shown in the correlation value of 0.088 implying no relationship.
- NA_Sales / JP_Sales: A covariance value of .130 does not imply a positive relationship and the correlation value of .469 implies only a weak connection between the two.
- NA_Sales / Global_Sales: A covariance value of 1.815 implies a positive linear relationship between the two which is backed up by the correlation coefficient of .956 implying a strong positive relationship between the two.
- JP_Sales / Global_Sales: a Covariance of .614 weakly implies a positive correlation between the two. And the correlation of .614 implies a moderate linear relationship between the two.

2. Create a scatter plot for the attributes `Critic_score/Critic_Count and User_Score / Global_Sales` Interpret the scatter plot**! 3 points**



For this Scatter plot you can see that there is a moderate positive relationship between the attributes. There seems to be a cluster of points in the center of the scatter plot approximately contained in a triangle contained within y = 1/2x, y = 100, and x=50 with only a minimal amount of points outside of that cluster. There is one large outlier with a Critic_Score/Critic_Count of over 150 and a User_Score/Global_Sales of approximately 130.

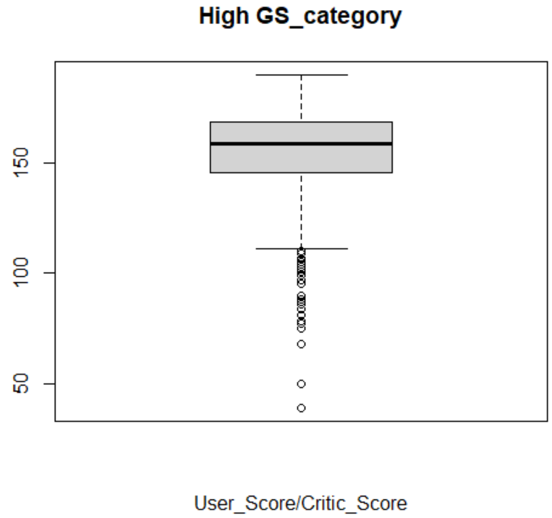
3. Create histograms for `Platform` and `User_Count` attributes for High, Medium and Low GS_category classes; interpret the obtained histograms. **6 points**

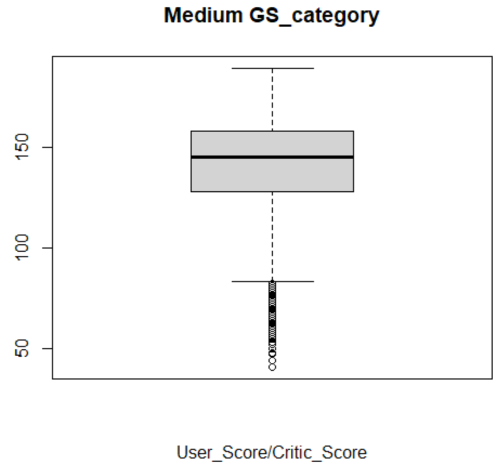


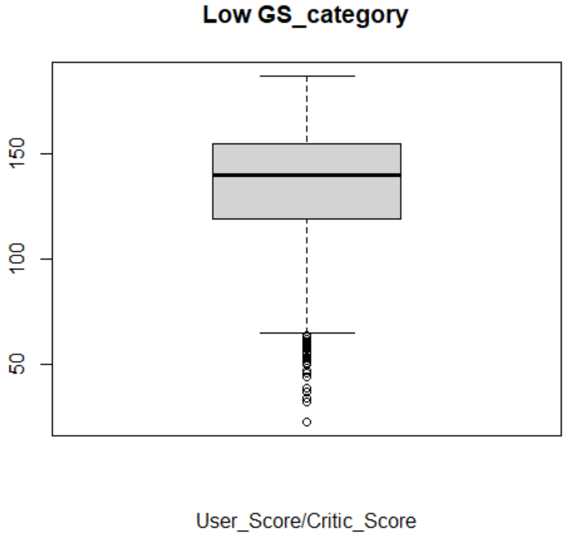
- Overall all of the histograms have 15 breaks to their buckets, they are all unimodal and they all skew right with at least one outlier at or beyond a user count of 1000. This makes sense as most games have very few users while most users play only a few popular games.

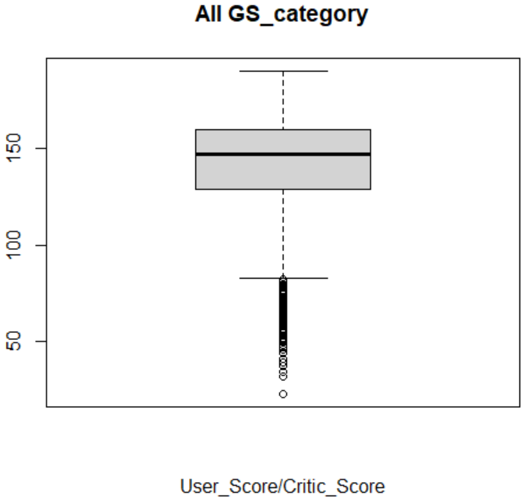
- High User Count: Contains a Mode of around 1000 games between 0 and 500, This graph has the smoothest transition and a presence in most buckets above a 1000. This makes sense as most games they sell high would have more current users.
- Medium User Count: Contains the largest mode of well over 3000 games with less than 1000 users. There are still at least one game in most buckets up to 10000 user count. This widespread between having low and large user counts makes sense for the middle of the pack.
- Low User Count: Contains a large mode of more than 1500 games with less than 500 User Count. This histogram has the least smooth drop off and does not have at least one game in most buckets greater than 2000. However it does contain at least one outlier with a user count just under 10000.

4. Create box plots for the `User_Score/Critic_Score` attributes for the instances of the 3 GS_category class— low/medium/high — and a fourth box plot for all instances in the dataset. Interpret and compare the box plots for each attribute! **4 points**



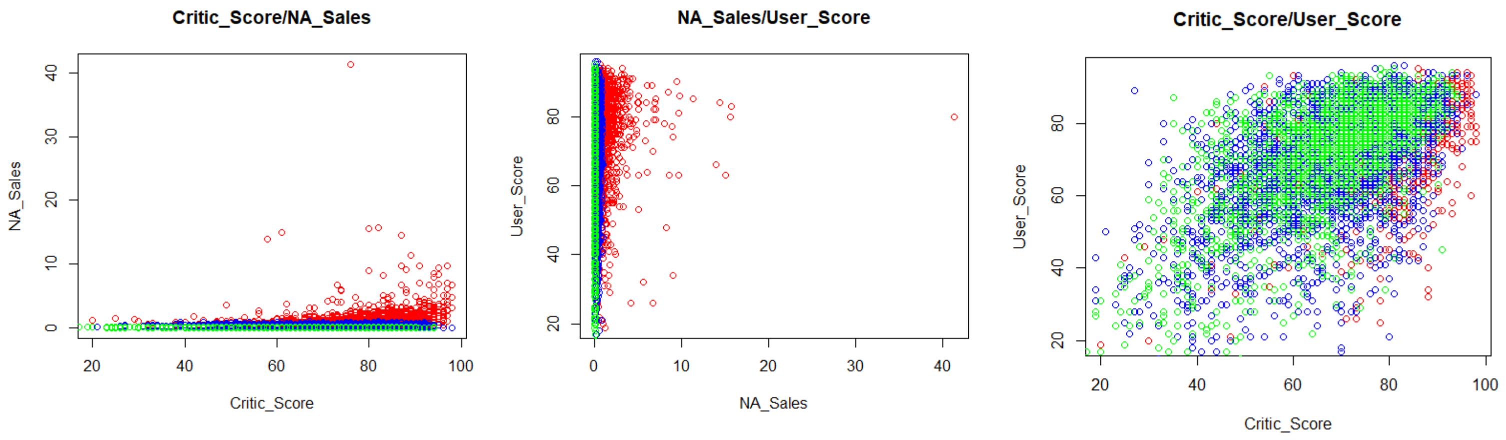






- Overview: All of the Boxplots are roughly similar, they are all roughly symmetrical with a medium around 150 with marginally different IQRs. Their overall ranges do differ drastically, especially the Low GS_Category boxplot. All the boxplots lack any outliers on the larger side and have a significant number of outliers on the smaller size.
- High: has a median value around 160 with a symmetric IQR of 24 meanwhile and an overall range of under 100. This would make sense as games with a very high sales should likely have very high User and Critic Scores and little spread.
- Medium: has a medium value just under 145 with an IQR of around 30, with a marginally larger lower quartile, and an overall range around 100. This would make sense as it is the middle of the pack for global sales, this boxplot looks the most similar to the combined boxplot.
- Low: Has a median of 140 and an IQR of 36, with a larger lower quartile with an overall range well over 100. This makes sense as the Games with Low sales would likely have poor overall sales
- All: Overall the Medium and Low categories are significantly more analogous of the combined data than the High category. This can be seen in the IQR Overlaps of the different plots.
    - High and All have an IQR overlap of 37.5%
    - Medium and All have an IQR overlap of 93.75%
    - Low and All have an IQR overlap of 63.4%

5. Create supervised scatter plots for the following 3 pairs of attributes using the GS_category as a class variable: `Critic Score`/NA_Sales, NA_Sales/`User Score` and `Critic Score/User Score`. Use different colors for the class variable. Interpret the obtained plots; in particular, address what can be said about the difficulty in predicting the `Global Sales Target` and the distribution of the instances of the three classes. **6 points**

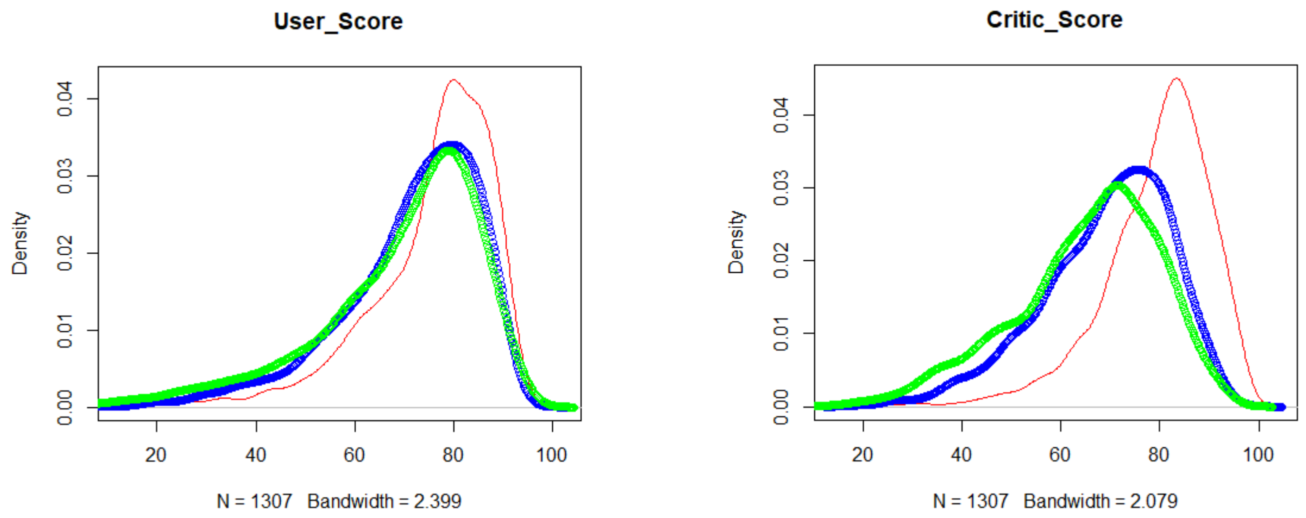


Legend: Red -> High, Blue -> Medium, Green-> Low

- Overall: Predicting a global sales target is very difficult especially with these variables as they are all on different scales which makes reading the scatterplot fairly difficult.
- Critic_Score/NA_Sales: Critic_Score and NA_Sales have almost no linear relationship.
- NA_Sales/User_Score: these variables have almost no linear relationship.

- Critic_Score/User_Score: These variables do have a moderate linear relationship, and cluster in the upper right hand of the scatter plot

6. Create 2 density plots for the instances of the 3 GS_category classes in the `Critic_Score`/`User_score` space. Compare the density plots! **6 points**



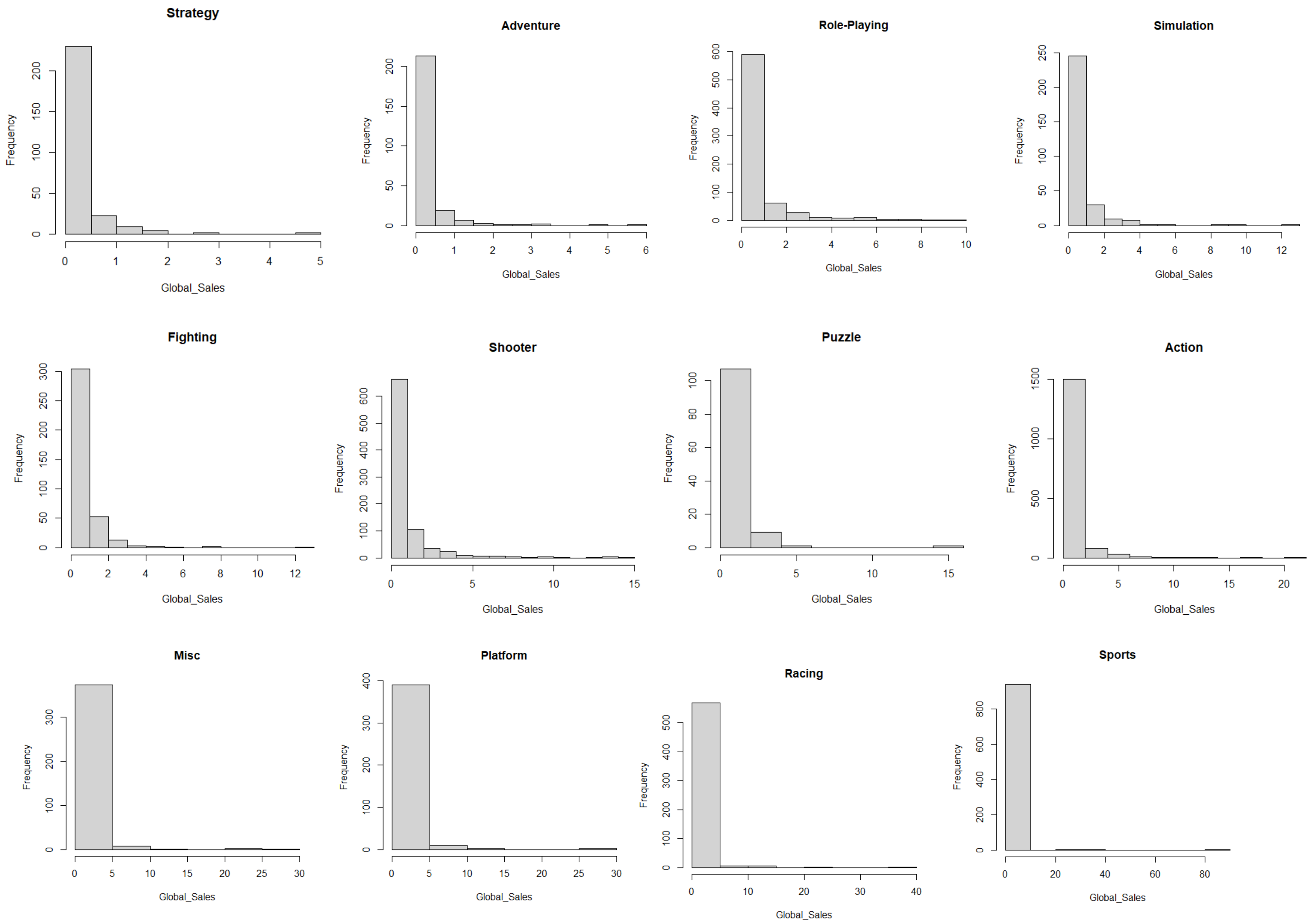
Legend: Red -> High, Blue -> Medium, Green-> Low

- Overall the two density graphs are fairly similar with the majority of data falling in the range between 60 and 90. For both of the graphs the Medium and Low Data makes up relatively even amounts of the overall data, while the High data makes up significantly more of the data where the User and Critic Score is above 80.
- User_Score: For this data the Medium and Low data are essentially equal at every point of the graph sharing a peak of just over .03% at User_Score 80. The High is also similar, it is marginally less until around User_Score of 75 when the frequency of High data increases drastically and is substantially more than the other data. The high data peaks at .04% at around a user_score of 80 and then quickly drops.
- Critic_Score: Here the data differs more. The medium and low data are marginally different with low data consisting of marginally more data than Medium until its peak of just under .03% at User_Score 70, while the Medium peaks at just over .03% at User_Score 75. The High Data is significantly different from the rest of the data. High Data makes up significantly less data when User_Score is less than 70, but quickly climbs to being just under .05% at around User_Score of 85.

7. Create a table which reports the frequency of associations of the 12 genres with the three classes of the GS_Category attribute. Create histograms for the Global_Sales attribute for the instances of each of the 12 genres. Interpret the table and the histograms you created. **8 points**

	High	Medium	Low
Sports	0.21	0.65	0.14

Racing	0.20	0.56	0.25
Platform	0.23	0.56	0.21
Misc	0.25	0.59	0.16
Action	0.19	0.60	0.21
Puzzle	0.18	0.43	0.39
Shooter	0.23	0.52	0.24
Fighting	0.20	0.62	0.18
Simulation	0.18	0.54	0.28
Role-Playing	0.17	0.57	0.25
Adventure	0.06	0.54	0.40
Strategy	0.06	0.40	0.54



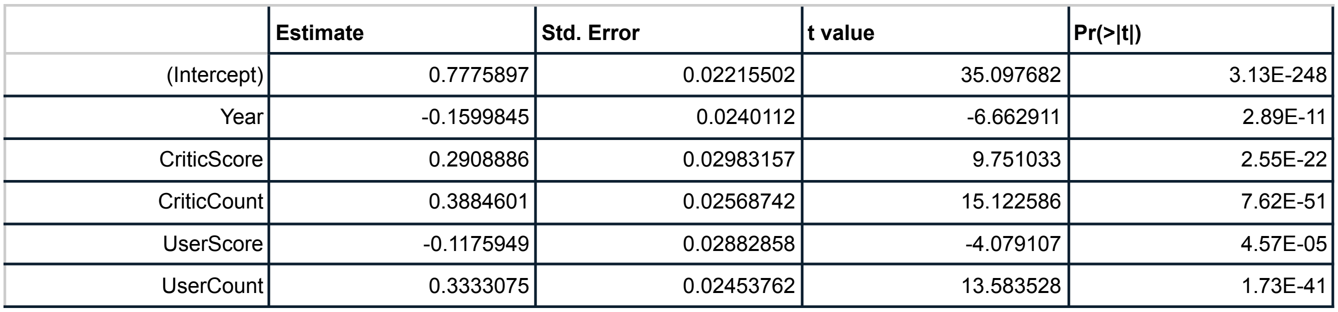
- Table Overview: For every Genre except for Strategy games, the Medium GS_Category makes up the largest portion of the genre. With the Exception of Strategy and Puzzle the Medium

GS_Category makes up over half of the Genre. For all of the categories the Medium and Low GS_Categories make up at least 75% of the games in that genre with Strategy and Adventure leading with 94% of the games being either medium or low.

- All of the Histograms are unimodal with a mode of at least 100 with Puzzle, and Action has the largest mode of 1500 games with less than 2 in Global Sales.
- All of the genre's histograms skew to the right each with at least one major outlier with Sports having the largest with at least one game with a Global_Sales of at least 80
- Overall all of this makes sense as most games get few sales while a few popular games get many sales.

8. Create a new dataset *Z-Processed Video Games* from the *Processed Video Games* dataset by transforming the `Year, Critic_Score, Critic_Count, User_Score, User_Count` attributes into z-scores. Fit a linear model that predicts the values of the Global`_Sales` attribute using the 5 z-scored, continuous attributes as the independent variables. Report the R^2 of the linear model and the coefficients of each attribute in the obtained regression function. What do the obtained coefficients tell you about the importance of each attribute for predicting a successful game? **6points**

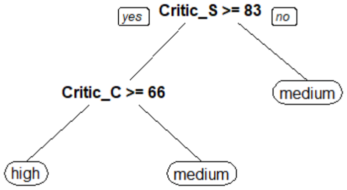
$R^2 = .131$

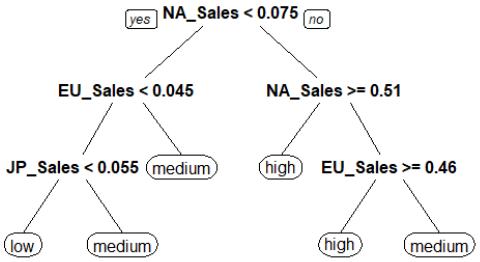


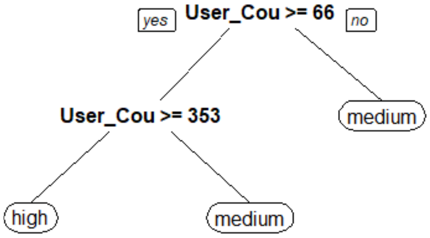
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.7775897	0.02215502	35.097682	3.13E-248
Year	-0.1599845	0.0240112	-6.662911	2.89E-11
CriticScore	0.2908886	0.02983157	9.751033	2.55E-22
CriticCount	0.3884601	0.02568742	15.122586	7.62E-51
UserScore	-0.1175949	0.02882858	-4.079107	4.57E-05
UserCount	0.3333075	0.02453762	13.583528	1.73E-41

- $R^2 = .131$: This R Squared value tells us that largely this linear model is not a very good model to explain the variance in the data.
- Year: here you can see that as the PValue, 2.89E-11, is very low so there is a significant reason to believe that this is a good measure to predict however as the estimate is .29 the effect is extremely low.
- Critic_Score: Here you can see that as the PValue, 2.89E-11, is very low so there is a significant reason to believe that this is a good measure to predict however as the estimate is .29 the effect is extremely low.
- Critic_Count: with a PValue of 7.62E-5, we have significant reason to believe there is an effect however with an estimate of .38 that effect is very small.
- User_Score: with a PValue of 4.57E-0, we have significant reason to believe there is an effect however with an estimate of -0.12 that effect is very small.
- User_Count: with a PValue of 1.73E-41, we have significant reason to believe there is an effect however with an estimate of .33 that effect is very small.

9. Create 3 decision tree models with 20 or less nodes for the dataset (both intermediate and leaf nodes count; do not submit models with more than 20 nodes!); use the `GS_Category` attribute as the class variable, and use the remaining attributes of the dataset excluding attribute Global_Sales of the dataset when building the decision tree model. Explain how the 3 decision tree models were obtained! Report the training accuracy and the testing accuracy of the submitted decision trees. Interpret the learnt decision tree. What does it tell you about the importance of the chosen attributes for the classification problem? **11 points**







- These three decision trees were created by breaking apart different aspects of the Video Game data and building a decision tree from that
- The first one is made by creating the decision tree with the Critic Score and Critic Count data to generate the decision tree predicting GS_Category. I chose these specifically to test how much that critics might predict how well a game does.
- The middle one is generated by using the NA_Sales, EU_Sales, and the JP_Sales to predict the GS_Category. I chose these because given the Sales in North America, Europe, and Japan i would imagine you could make a pretty good guess as to how well Global Sales would be.
- The last one is generated similar to the first one by using the User_Count and User_Sales data to generate a decision tree that predicts the GS_Category, Similar to the Critic Data, I imagine the User Score and User Count would be very important for predicting how many users choose to by the game.
- I think the chosen attributes and what to focus on are very important for classification problems.

10. Write a conclusion (at most 13 sentences!) summarizing the most important findings of this task; in particular, address the findings obtained related to predicting a successful game (high global sales) using attributes 1-14. If possible, write about which attributes seem useful for predicting high video game sales and what you as an individual can learn from this dataset! **6 points (and up to 4 extra points)**

The most Important finding from this task for me was to take this very seriously and to interpret all along the way. I approached this task by writing all of the code and then leaving the last day to interpret the data once I had been able to see everything. This was a losing strategy, however the task was very helpful with getting me acclimated to R again.

Finding and predicting a successful game really rests on the ability to know how much the game is already selling in a diverse number of places. I thought it was quite interesting but also intuitive that in question 1 Critic Score and Japanese Sales were correlated by just 0.147. This makes sense as most critics are probably White American Males, they likely have entirely different cultural attitudes about what makes a great game. But on the same end North American Sales and Global Sales being correlated by 0.956 makes a lot of sense as in the 21st century the United States is still the economic center of the world, and games that do well there will see success in other markets as well.