Introduction.

The task is to explore how supervised learning models can be used to predict the sales performance of video games worldwide.

To compare regression models that predict the “global sales” of video games based on a single numerical input feature, regression models that take multiple numerical variables as input features to predict the “global sales” of video games, categorical variables in the dataset that are likely to affect the global sales of video games. This report looked at how we developed an Artificial Neural Network model to predict the global sales of video games based on information from the dataset, how its performance compares to other supervised learning models carried out, and the use of the k-Means clustering algorithm to identify clusters in the video games’ sales data.

Problem Statement.

Video game industry is a very competitive market, there are various factors that affect the sales of games globally. Developers, publishers etc, who can understand these factors and their relationships to the performance of sales can provide some deep insights required. To identify the significant predictors from numerical and categorical features such as developer, genre, platform and publisher, can assist in predicting the global sales of video games correctly, clustering techniques can also help in accessing sale performance as well.

Some Challenges Faced.

1. Categorical Features Encoding: Encoding categorical features properly can be a challenge. Encoding methods like One-Hot-Encoder and Label-Encoder can lead to dimensionality issues which can cause loss of some meaningful relationships among these categories.
2. Nan values or Incomplete Data: Model training can be hindered, or prediction can be inaccurate due to missing values in Rating, Publisher, Developer, and Genre. Therefore, it is very important to handle these missing values properly when building a model.
3. Clustering Evaluation: Evaluating the quality of clusters in unsupervised learning, e.g., k-Means, is subjective and depends on the choice of the variables, the distance, and evaluation methods, e.g., the silhouette score.
4. Overfitting: When the dataset is small or it is unbalanced across categories, models like random forest and neural networks are prone to overfitting. Also, using scatter plot to label qualitative i.e. categorical features like Genre, publisher and developer tend to be difficult.

Work Summary Review.

Necessary libraries were imported (NumPy, Pandas, Seaborn, Matplotlib.pyplot). The data frame has 16719 rows and 16 columns after loading. The data frame was cleaned to have accurate predictions while carrying out the analysis.

Comparing Regression Models.

The dataset has numerical features which consists of NA_Sales, EU_Sales, JP_Sales, Other_Sales, Critic_Score, Critic_Count, User_Score, User_Count, and Year_of_Release. After training the model and regression analysis was

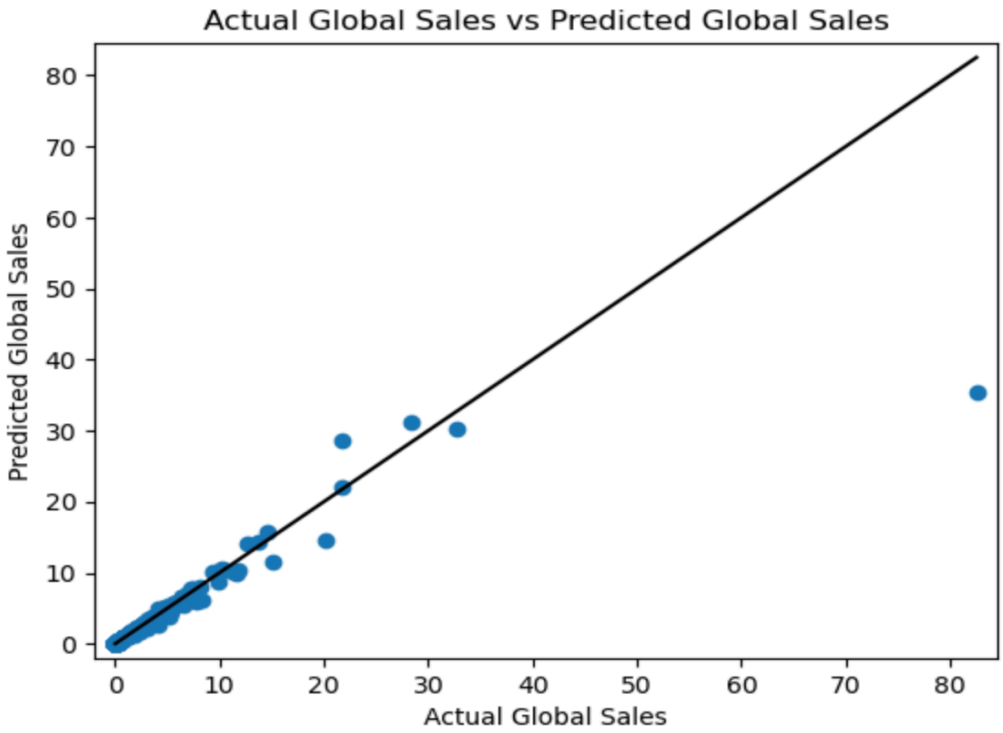
carried out, the relationship between Global_Sales and the numerical variables, the linear regression model for some variables gives better results compared to polynomial, while overall polynomial performed more. The R2 score gives an overall measure of how well the model is performing. An R2 score closer to 1 indicates that the model is doing well. The linear regression analysis in NA_Sales R2 score was 0.939, compared to the polynomial that had an R2 score of 0.833.

Linear regression R2 score		Polynomial regression R2 score
EU_Sales	0.927	0.925
JP_Sales	0.316	0.321
Other_Sales	0.716	0.608
Critic_Score	0.023	0.038
Critic_Count	0.050	0.059
User_Score	0.000	0.001
User_Count	0.019	0.044
Year_of_Release	-0.000	-0.001

From the table above, we can see that in some cases the model performed well in linear regression, but overall polynomial has a slight edge over it.

Regression Models That Take Multiple Numerical Variables.

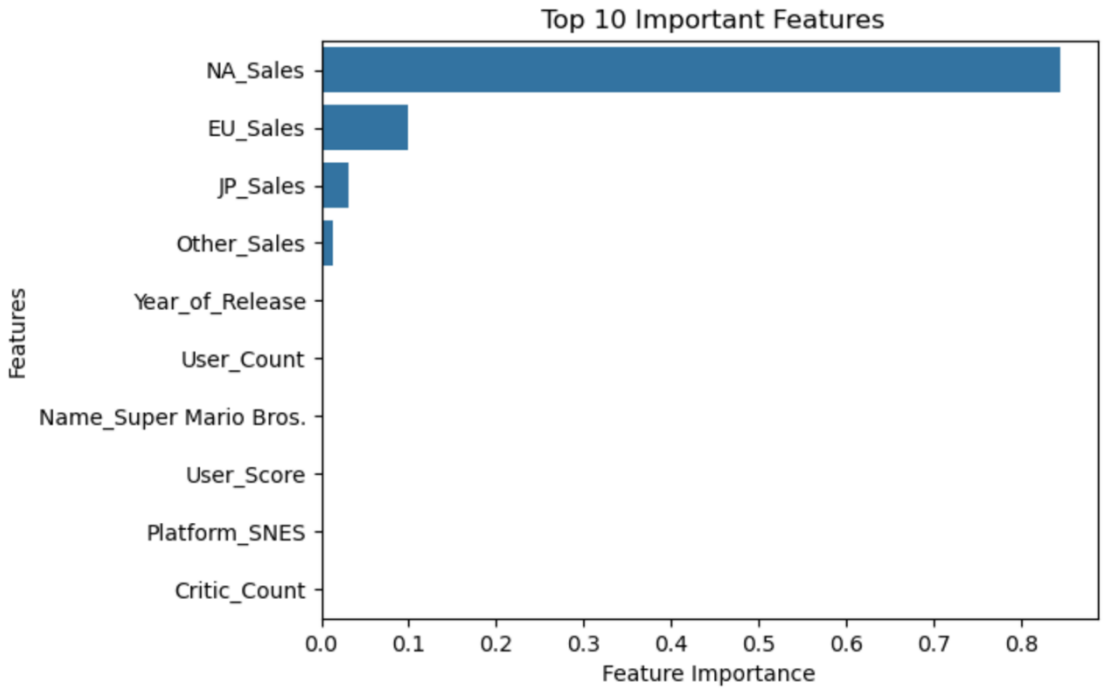
Considering regression models that take multiple numerical variables as input features to predict the global sales of video games, the inclusion of these multiple numerical variables after training had an R2 score of 0.9999, which is better than any model so far that we have used. I went ahead to visualise the predicted global sales from this model by plotting the predicted global sales against the actual global sales from the testing dataset, and added a diagonal straight line to the plot. If the model is performing well, the plotted points are expected to lie along the diagonal line corresponding to the predicted and actual global sales, which shows that it is equal.



The image above shows the plot of a regression model that takes multiple numerical variables.

Regression Model That Uses All Relevant Input Variables.

Most regression models can't handle categorical data directly. To address this, the use of One-Hot Encoding to transform categorical variables into a binary format additionally handles missing or unseen categories by configuring the encoder. Using the random forest regressor model, the R2 score was 0.803. I went further to make a bar-plot visualization for the top 10 important features.

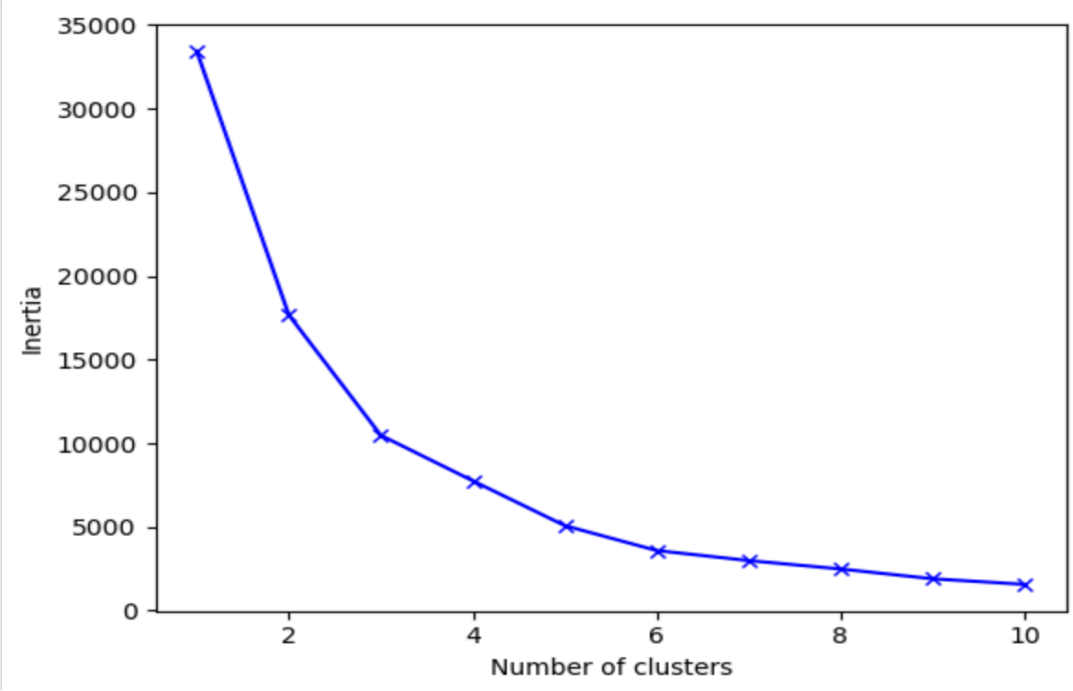


The plot above shows NA_Sales has high feature importance, which affects the global sales of video games.

Unsupervised Learning Techniques Used.

The k-Means clustering algorithm was used to identify clusters in the video games' sales data. First, I extracted the features from the dataset that I want to use as the input to the clustering algorithm. I then re-scaled them using z-score standardisation to re-scale each variable such that they have a mean of zero and a standard deviation of one. Built the k-Means model using the elbow method.

For Na_Sales, the optimal value of k = 3, this is because that is where it starts to flatten out. David Bouldin index = 0.5394, Silhouette Coefficient = 0.8692.



The optimal value of k corresponds to the "elbow" of this plot, i.e., where it starts to flatten out.

For EU_Sales the k value = 4, DBI =0.4673, SC = 0.8606. JP_sales the k value= 3, DBI =0.7054, SC = 0.8794. The optimal value of k for Other_Sales = 4, DBI = 0.6640, SC = 0.8484. For Critic_Score k= 4, DBI =0.6764, SC = 0.6273. For Critic_Count, k=5 DBI =0.6883, SC = 0.6269. For User_Score k = 4, DBI = 0.5756, SC = 0.7247. User_Count k = 5, Davis Bouldin Index =0.6696, Silhouette coefficient = 0.8432.

Using Another Clustering Algorithm: Hierarchical Clustering

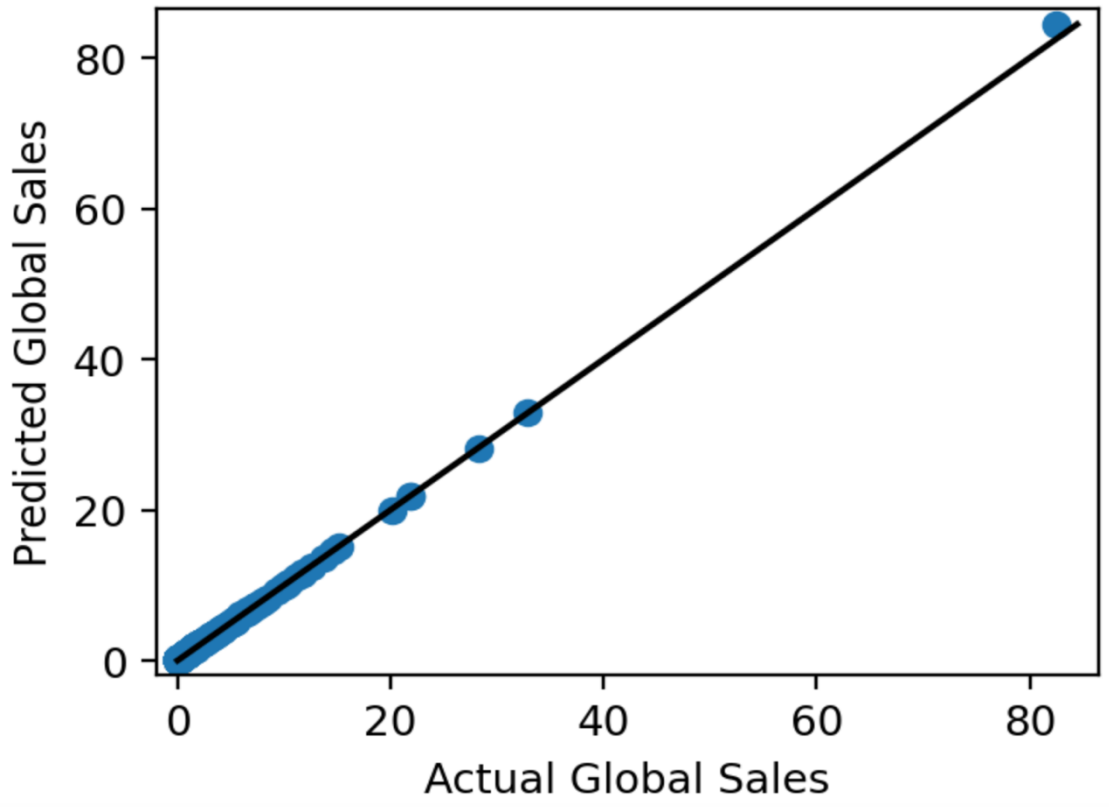
For NA_Sales, considering the model with 3 clusters, so we can easily compare to the other results. I achieved a Davies Bouldin Score of 0.2336 and Silhouette Score of 0.9743. Recall earlier using k-Means, we achieved a DB of 0.5394 and a silhouette Coefficient of 0.8692 using the k-Means model. Compared to that, the hierarchical clustering model is better, however, the DB index suggests that the k-Means model is better. For this comparison,

given that the difference in the DB index is relatively small, and the silhouette score is way higher, I would therefore conclude that the hierarchical model gives a better clustering result overall.

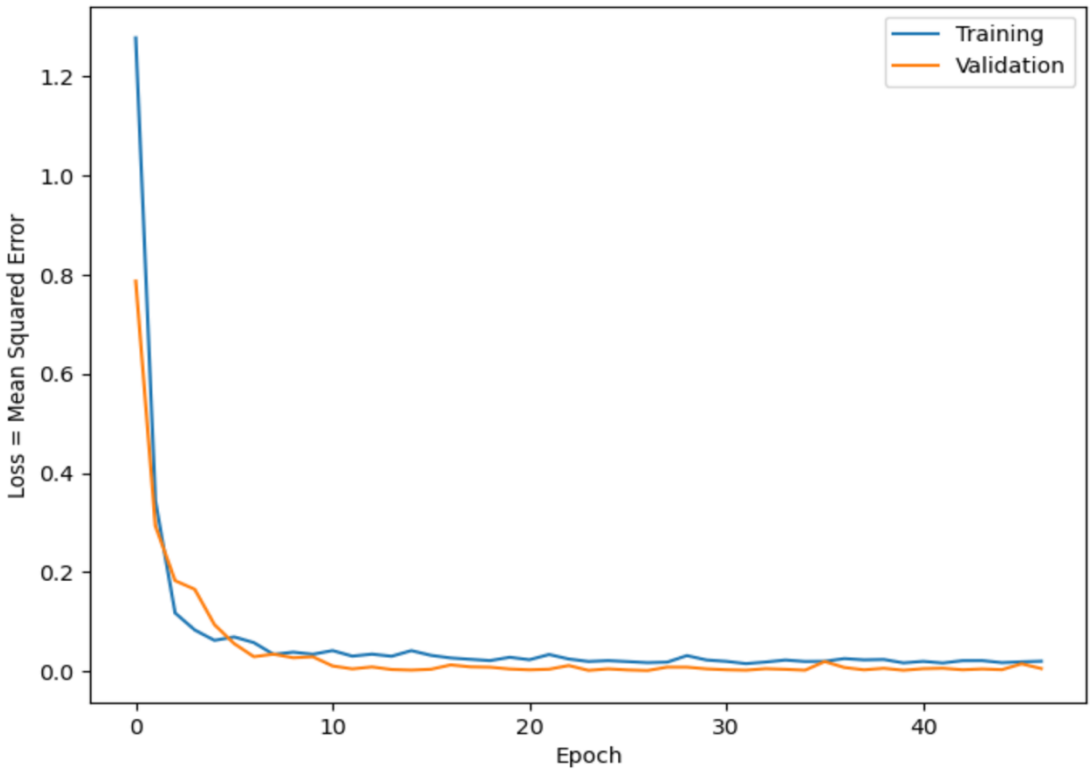
ANN Model That Predicts the Global Sales of Video Games.

The sk_learn function was used to divide the input and output data into training and testing sets, and applied min-max scaling, which rescaled each variable such that they have values between 0 and 1. The neural network model was built with 2 hidden input layers, a dropout layer, and an output layer. It was compiled, and the setting of various options and hyperparameters, such as optimiser, the loss function, and the metrics. Adam Optimiser was used, and the mean squared error was used for the loss and metrics.

The model was trained for 200 epochs but also imposed with a patience early stopping criterion of 20. This means that the training will be stopped if the validation loss has not improved over any 20 consecutive epochs; this helps prevent over-fitting. After training the model, I visualised the result.



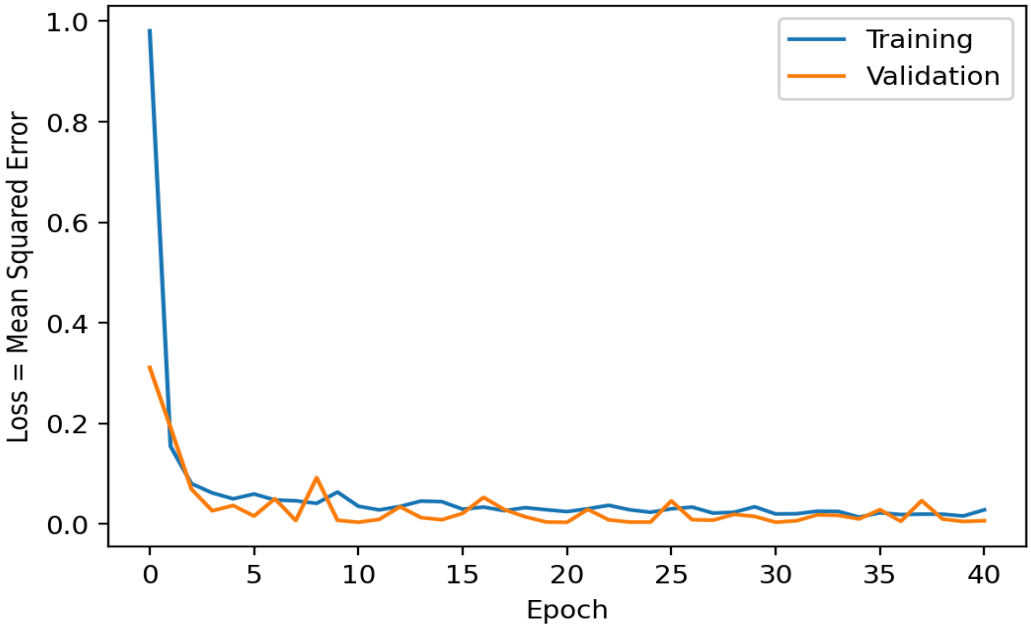
Further plotted how the loss function varies with epoch, for both the training and the validation sets.



From this plot, we can see that the mean squared error loss for training started from 1.3, and the validation sets decrease from 0.8 in the first 5 epochs and then continue to decrease gradually after that. There is no strong evidence of over-fitting; we would see this if the validation loss started increasing at later epochs while the training loss continued to decrease, but this does not happen.

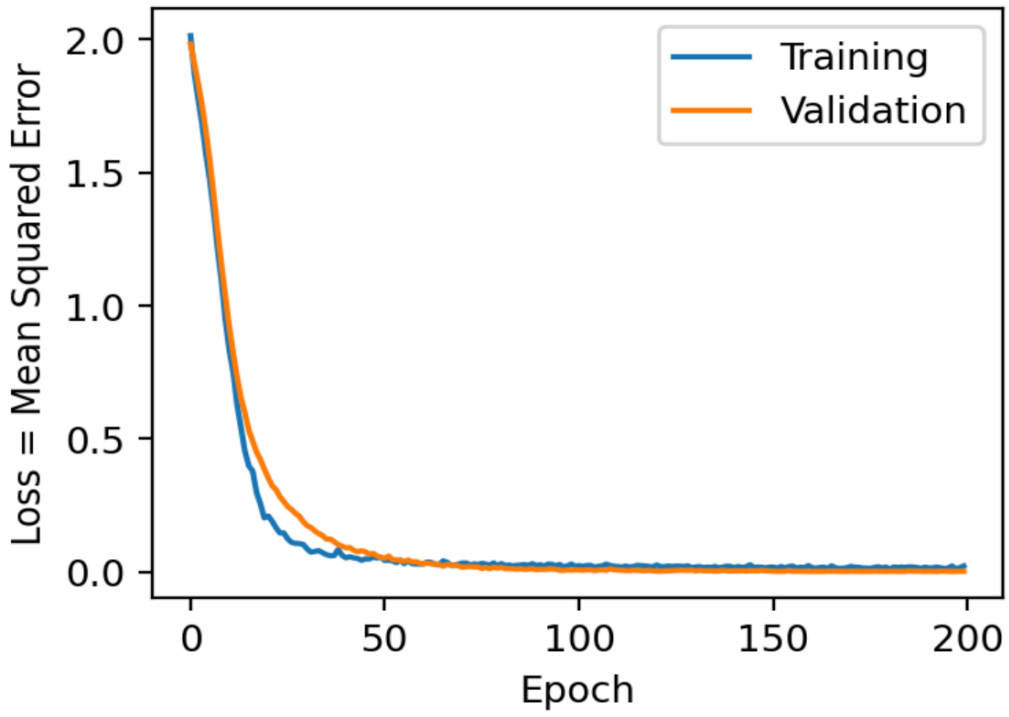
Hyperparameter Tuning.

A new neural model was built that adds a third layer.



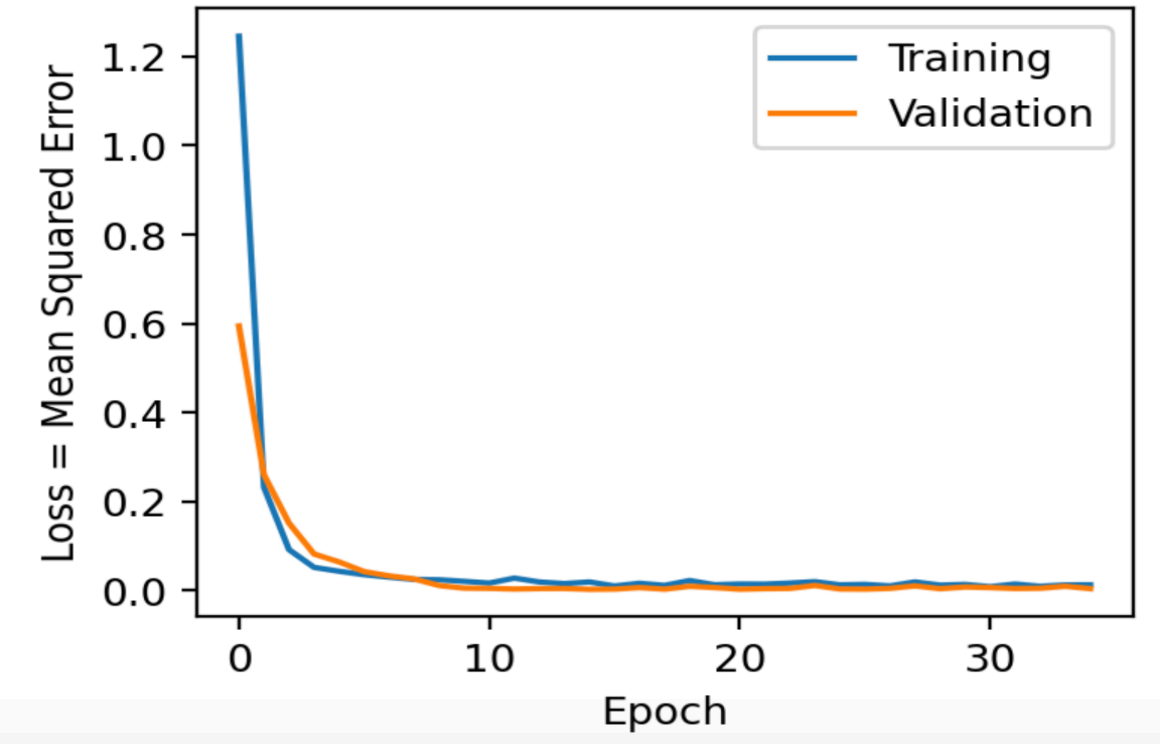
Compared to the plot earlier, there is evidence of over fitting. I continued with this architecture using two (2) hidden layers.

A new model was created, but with a smaller learning rate of 0.0001. With a smaller learning rate, the model takes a lot longer to train.



Changing the Dropout Rate.

The final hyperparameter considered is the dropout rate that was imposed in the first hidden layer. The neural network model was repeated, but with a dropout of 10%



The final mean squared errors are generally lower in this model, the predicted values are somewhat closer to the actual values. Reducing the dropout rate to 10% has improved the model by a small difference.

Conclusion.

After successfully completing the analysis on video game sales, the best model for predicting global sales is the regression model that takes multiple numerical variables. This is because considering the R2 score of 0.999, which

is the highest compared to all other training methods that have been used during the course of this analysis.