Covid-19's Effect on US College Students

By Daniel Jackson, Nischal Panta & Nelson Tran

*** Each author contributed equally to the design, coding & development, analysis, and writing of this project**

Abstract:

This research paper investigates the relationships between various survey predictors and anxiety levels among college students related to the COVID-19 pandemic. We employ regression models to analyze our data, aiming to identify key predictors of anxiety. The insights gained from this analysis are intended to support stakeholders in developing targeted policies and interventions for current and future student populations. By improving the identification of students experiencing COVID-19-related anxiety, our findings can assist in the more effective allocation of resources and funding, ensuring that appropriate mental health programs and support services are available to those in need. This research provides valuable insights into the impact of the pandemic on student well-being and offers evidence-based recommendations for policy decisions aimed at strengthening support systems for students.

Table Of Contents

Background

The research question we will be exploring is: Does COVID-19 impact college students in the US? And if so, is there a way to determine which groups are most affected? We will assume that our stakeholder for this project is a government-funded COVID-19 student relief group. They are responsible for spearheading relief efforts to ensure students affected by COVID-19 have proper resources available to them to help with the stress endured from the pandemic.

We are interested in understanding the impact that COVID-19 had on college students during the global pandemic. It was an unforeseen event that negatively impacted so many around the world. We wanted to focus on college students specifically. These institutions and universities were not prepared to handle the magnitude of the pandemic. Obtaining a college degree is stressful in itself. When combined with a pandemic, college students then had to deal with something out of their control. This was the catalyst for a new form of anxiety that students had never experienced before. This research paper is meant to analyze the effects that COVID-19 had on college students. We want to use the information and data provided by students themselves to be able to predict which students will suffer from anxiety in the future when faced with similar circumstances. To ensure that funds, attention, and services are properly provided to those in need. We will be focusing on US college students specifically, but we feel that our findings can be implemented in other fields and institutions around the world.

For our research proposal, we are going to be creating a quantitative regression model to predict the effects that COVID-19 has on college students' anxiety levels. We are assuming that there are similar studies out there on the general population. However, we want to focus on college students as they were a vulnerable subgroup during the pandemic. The goal of this research project is to help education institutions provide support services and create new policies that will help students who experienced a change in their anxiety levels. Although the question is not unique in itself, some studies have addressed the effects of COVID-19 on the general public. The novelty of our project lies in the fact that it is solely analyzing and studying the effects of COVID-19 on college students. It is worth our time to explore this question, as we want to provide insights into how COVID-19 affected students during the pandemic. This will be very valuable in aiding and informing both educational institutions and policymakers of the psychological effects the pandemic had on students. This information can be used in designing better support systems and policies to help cope with those with increased anxiety levels. This will also be useful for building the base for other areas of mental health and public health research.

As mentioned our research proposal is not an original one. We are taking the improvement approach. We are looking to combine multiple feelings from students into one standalone feeling, which will be our response variable. We will be coding in R for our project.

Our response variable will be quantitative. We will call our response variable “total_anxiety_level”. It will be a quantitative variable of the total anxiety level experienced by a US college student. We will be combining the scores of multiple feelings regarding COVID-19 into one response variable. The scores of feelings that will be combined: **Afraid, Irritable, Guilty, Sad, Preoccupied, and Stressed**. We will take the value of each variable, add them, then take the average and assign that to a new variable in our data set. We will then create a quantitative regression model to predict the response variable. We will use numerous predictor variables from the survey. Most of the predictors in the dataset are qualitative, but there are also some quantitative variables. Using the mean squared errors (MSE) of our predictions, we will test the performance of our models and choose a final one that we will use for our final submission. Our data dictionary, which can be found in the appendix, will have a better summary of each variable used. We will be removing and merging some features before and during our exploratory data analysis. Before we dive into our analysis plan, let us first cover our hypothesis and predictions:

Hypothesis: COVID-19 had a significant impact on the anxiety levels of college students because of the unique stressors associated with the pandemic.

Prediction: College students are more likely to have higher anxiety due to the effects that COVID-19 had on students' emotions and lifestyle changes.

The structure of our analysis plan will be as follows:

Analysis Plan:

1. **Data Acquisition**
2. **Data Exploration**
3. **Model Selection/Pre-Processing**
4. **Model Tuning, Validation, Exploration**
5. **Conclusions**

Data Acquisition

As mentioned, we decided to conduct a quantitative linear regression model to help us predict college students' anxiety levels based on a survey that they would fill out. We are using a data set that we found in a study about the psychological impacts of COVID-19. This data set features 2,534 observations from 14,147 surveys taken from students from 7 universities. These universities are located in the US. The 7 universities: Arizona State University, Clemson

University, North Carolina State, Oregon State University, Pennsylvania State University, and the University of Montana. This data set can be freely downloaded from the study using the link below:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7790395/#sec021

Data Exploration

Our first order of business when it came to cleaning our data, was removing redundant variables from the data set that we found. Some of the variables (ending in _group) were derived from grouping individual variables in the data set from a previous analysis. These were removed as the groups the independent variable belongs to are highly correlated to each other. Some of the other variables were Z-scores for a group. The entries for these variables are not necessary as we will be doing our tests and fitting them into our models.

We also combined our six stress responses from the survey into one response variable that we are looking to predict. The six stress responses are as follows: “covid_afraid,” “covid_irritable,” “covid_guilty,” “covid_sad,” “covid_preoccupied,” and “covid_stressed”. Below are the questions regarding the stress responses. Each response was answered between 0-100 by each student (0 being not at all worried and 100 being extremely worried).

Afraid: "How *afraid* do you feel when you think about coronavirus?"

Irritable: "How *irritable* do you feel when you think about coronavirus?"

Guilty: "How *guilty* do you feel when you think about coronavirus?"

Sad: "How *sad* do you feel when you think about coronavirus?"

Preoccupied: "How *preoccupied* do you feel when you think about coronavirus?"

Stressed: "How *stressed* do you feel when you think about coronavirus?"

These six variables were combined and averaged out to create a new variable called "total_anxiety_level." This variable will become our response variable when we create our regression model. Our idea is that we want to see if we can predict this response variable without having to ask students each of the six questions. We want to minimize direct COVID-19 stressor questions to eliminate as much bias to our study and future predictions. We want to proactively identify students who may need assistance with stress caused by COVID-19 without having to address COVID-19 with them directly. Having one response variable will allow us to fit and train models. We will then analyze both the training and test error predictions to help us select the best model.

Once we have selected our model, we will then test our predictions on the entire data set to simulate our process. We want to predict a student's anxiety level and see if they lie above or below the average value of the sample's mean. If a student's anxiety level lies above that mean, we will deem them as someone who needs assistance with their COVID-19-induced anxiety. We used the mean instead of the median for this new variable as a threshold to determine if a student's anxiety had been altered due to COVID-19 because it would include more students than using the median.

After cleaning out the combined variables in the data set, we investigated more by creating a correlation matrix with all of our numerical variables. That can be seen below in *Figure 1.1* and *Figure 1.2*.

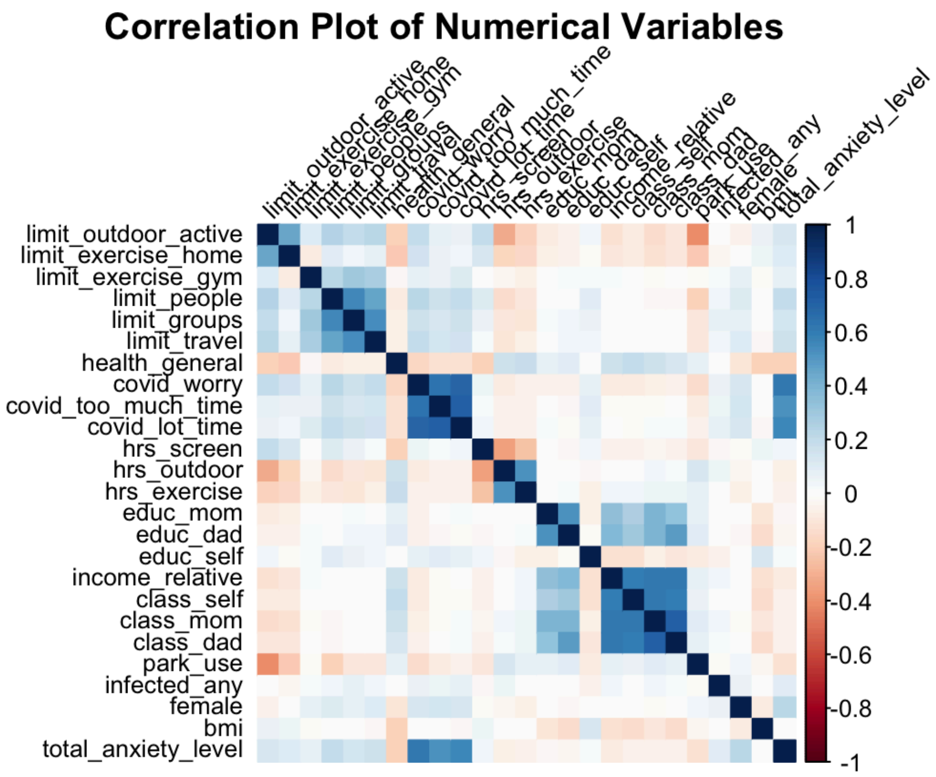


Figure 1.1

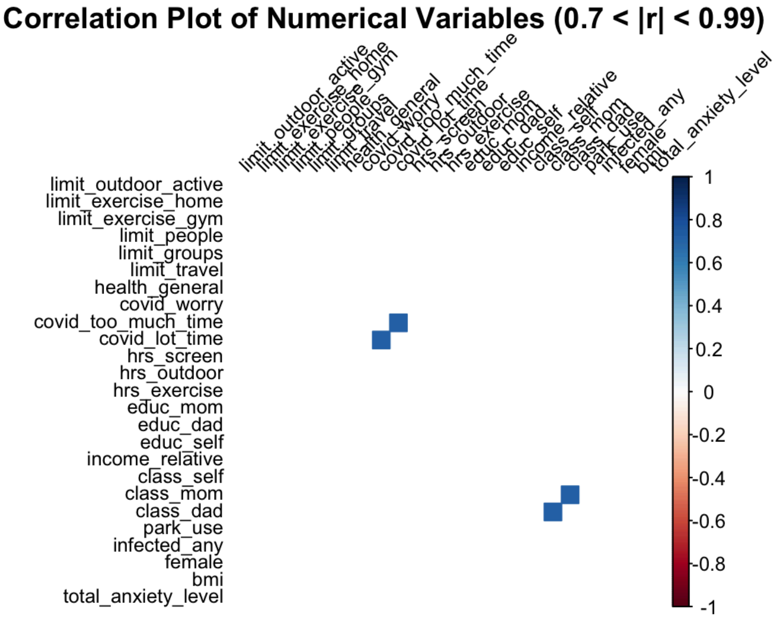


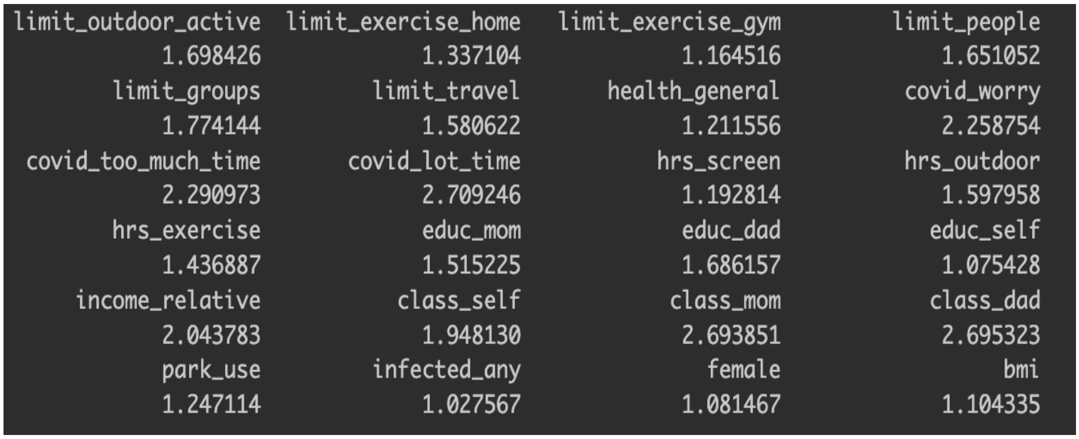
Figure 1.2

Once we created the correlation matrix, we wanted to single out the significantly correlated variables so we chose a range between 0.7 and 0.99. This can be seen in *Figure 1.2* above. From the matrix, we found that 4 variables fit this description. The 2 pairs are “covid_lot_time,” “covid_too_much_time,” and “class_dad,” “class_mom.”. Since the correlation coefficient is not more than 0.75, they are not heavily correlated.

The covid_lot_time variable is a qualitative ordinal variable. Students answered "I spend a lot of time thinking about coronavirus” from 1 (Strongly disagree) – 7 (Strongly agree). The covid_too_much_time variable is also a qualitative ordinal variable. Students answered "I give too much time/thought to coronavirus” again with 1 (Strongly disagree) – 7 (Strongly agree). Based on the correlation matrix and the common wording in each phrase, these variables seem to be asking very similar questions. Therefore, we removed one of these variables. Since they are highly correlated predictors, it does not matter which one we remove, especially since they are very similar statements. With no specific reason, we removed the covid_too_much_time variable.

class_dad and class_mom are both qualitative ordinal variables. The class_dad variable represents a student’s dad’s social class from 1 (Working Class) – 5 (Upper Class) and the class_mom variable represents a student’s mom’s social class from 1 (Working Class) – 5 (Upper Class). It makes sense why these variables have high correlation. If a student’s dad is upper class, then most likely the student’s mom is upper class as well. And vice versa. Similar to the covid_lot_time and covid_too_much_time variables from above, we could argue removing one of the class_dad or class_mom variables. Instead, what we did is create a new variable called class_family which takes the average of class_dad and class_mom, rounds up to the nearest whole number, and gives us a new value that represents the social class of a student’s family.

After removing all grouped variables and others that we did not deem fit for our models we ran the VIF test. For this test, what we were looking for was low VIF values which suggest no multicollinearity. VIF quantifies how much a variance (standard error) is inflated which exists when there is multicollinearity between two variables. When we performed this test, all our values were within the appropriate range. If values were above 5 we would have to further investigate as to which variable is causing this inflation. When collinearity exists it undermines the statistical significance of an independent variable, increasing the model's complexity and causing over-fitting. The VIF helps estimate the inflation that multicollinearity inflates the variance of a regression coefficient. All of the values that we got from the VIF are fairly low which indicates that we are safe from multicollinearity. This can be seen below in *Figure 2.*



```
limit_outdoor_active  limit_exercise_home   limit_exercise_gym         limit_people
            1.698426             1.337104             1.164516             1.651052
        limit_groups         limit_travel       health_general          covid_worry
            1.774144             1.580622             1.211556             2.258754
  covid_too_much_time       covid_lot_time           hrs_screen          hrs_outdoor
            2.290973             2.709246             1.192814             1.597958
        hrs_exercise             educ_mom             educ_dad            educ_self
            1.436887             1.515225             1.686157             1.075428
      income_relative           class_self            class_mom            class_dad
            2.043783             1.948130             2.693851             2.695323
            park_use         infected_any               female                  bmi
            1.247114             1.027567             1.081467             1.104335
```

Figure 2

For our training and test data for our models, we used an optimal split code in RStudio that selected the optimal split of our data. This code can be found in our GitHub repository link. That link can be found in the appendix in the "Code Availability" section. The optimal split code informed us that 80% of our data would be used to train our models, and 20% of our data would be used to test those trained models. Since we have 2,534 observations, 2,029 of the observations will represent the training data and 505 observations will represent our test data.

After splitting the dataset into training and test sets we used K-nearest neighbor to input any missing values for our variables with missing values. Those two variables were the “ethnoracial_group” and “age” predictors. The “ethnoracial_group” predictor only had four levels, so we will be creating dummy variables to separate these levels. We also did the same for both “age” and “source”. The “age” predictor had six levels and the “source” predictor had seven levels.

We wanted to inspect the distribution of our response variable “total_anxiety_level” to see if it was normal or not. We looked at both a histogram and a Q-Q plot to see if we needed to transform our response variable or not. We also conducted the Shapiro-Wilk Test to see if our response variable has a normal distribution or not. If the p-value of the test is less than 0.05, then we reject the null hypothesis that the distribution is normal and assume that a transformation can better normalize the distribution. *Figure 3.1* and *Figure 3.2* below show the distribution and Q-Q plot.

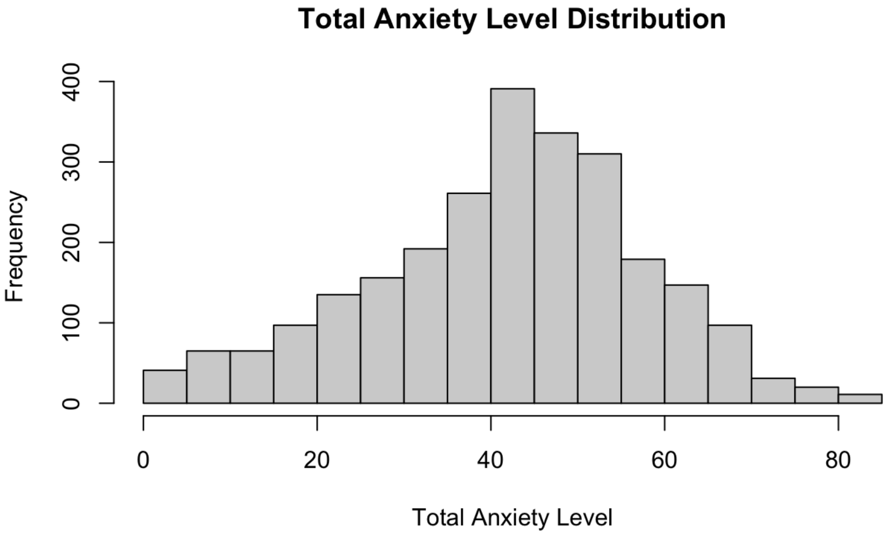


Figure 3.1

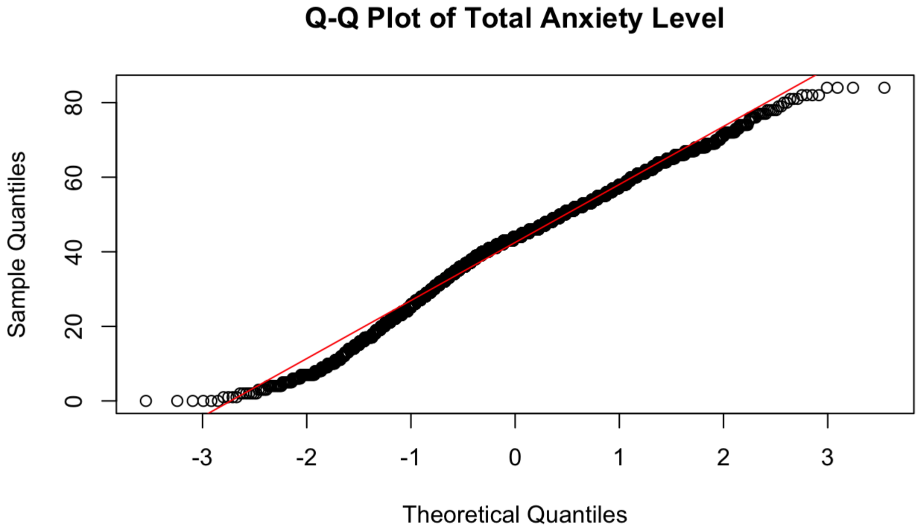


Figure 3.2

The histogram shows that our response variable relatively follows a normal distribution. However, when we look at the Q-Q plot, we see that the plot does not fully follow the reference line from -2 to 2 on the Theoretical Quantiles axis. This is proven by the Shapiro-Wilk Test. With a p-value of 2.787e-14, we reject the null hypothesis and state that the distribution is not normal. Therefore, we needed to transform our response variable. After a squared, square-rooted, log, and Box-Cox transformation, we found that a Box-Cox transformation made the distribution more normal. The Shapiro-Wilk test still produced a p-value of 4.437e-10, which was still less than 0.05. However, the p-value was closer to 0.05 than the p-value of the non-transformed distribution. We can see that the Box-Cox Q-Q plot follows the reference line from -2 to 2 on the Theoretical Quantiles axis. The histogram in *Figure 4.1* and the Q-Q plot in *Figure 4.2* show that the distribution is more normal. The Q-Q plot shows the response variable distribution closely following the reference line. You can find the histograms and Shapiro-Wilk p-values of the other transformations that we attempted in the appendix.

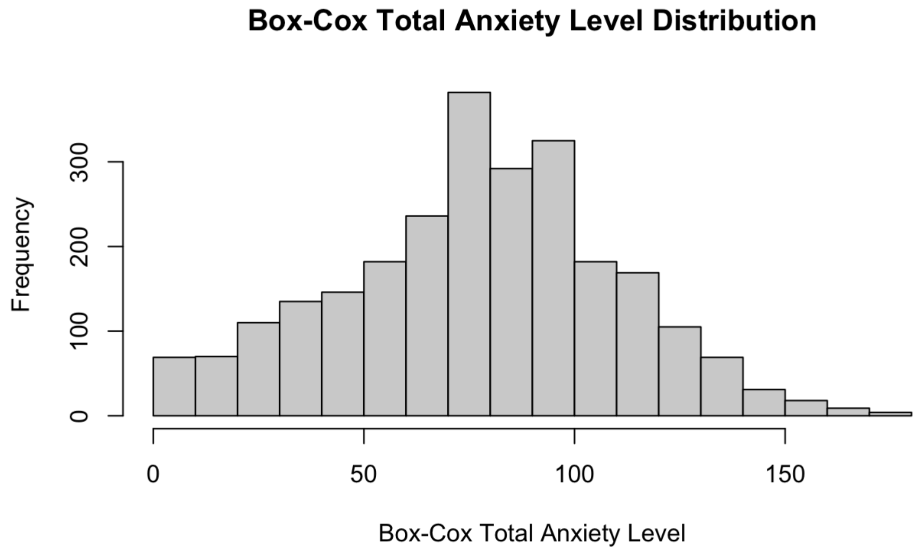


Figure 4.1

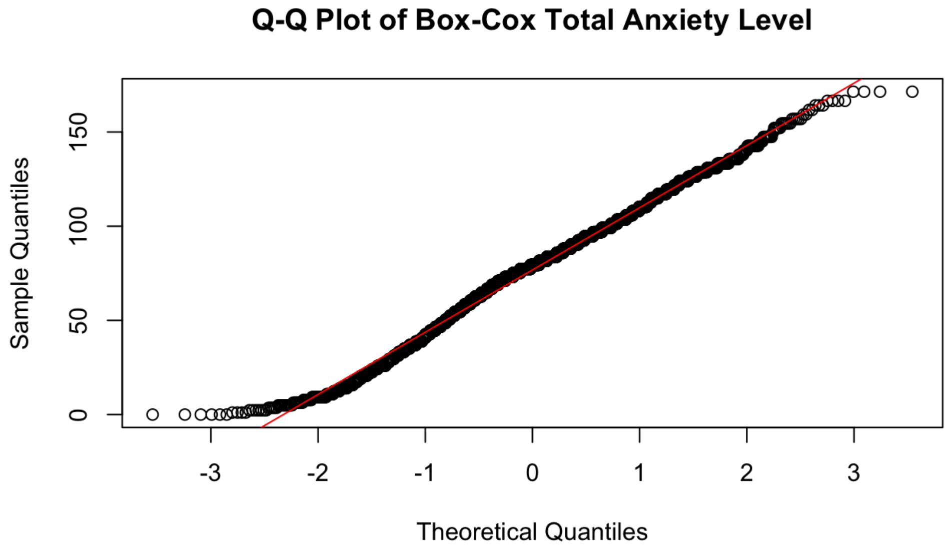


Figure 4.2

Pre-Processing & Initial Model Selection

After our EDA, we were left with 26 predictors and one response variable. We wanted to reduce the dimensionality of our data set before fitting our model. We decided to use Principal Component Analysis (PCA). To do so, we first had to one-hot encode our three qualitative variables, then center and scale all of our predictors. We did not scale or center our response variable, as that has already been Box-Cox transformed to follow a normal distribution. After our one-hot encoding, centering, and scaling, we were left with 35 total predictors (not including our response variable). We then conducted our PCA on those 35 predictors to create 35 principal components. In *Figure 5* below, you will see the Scree Plot of our PCA on our training set.

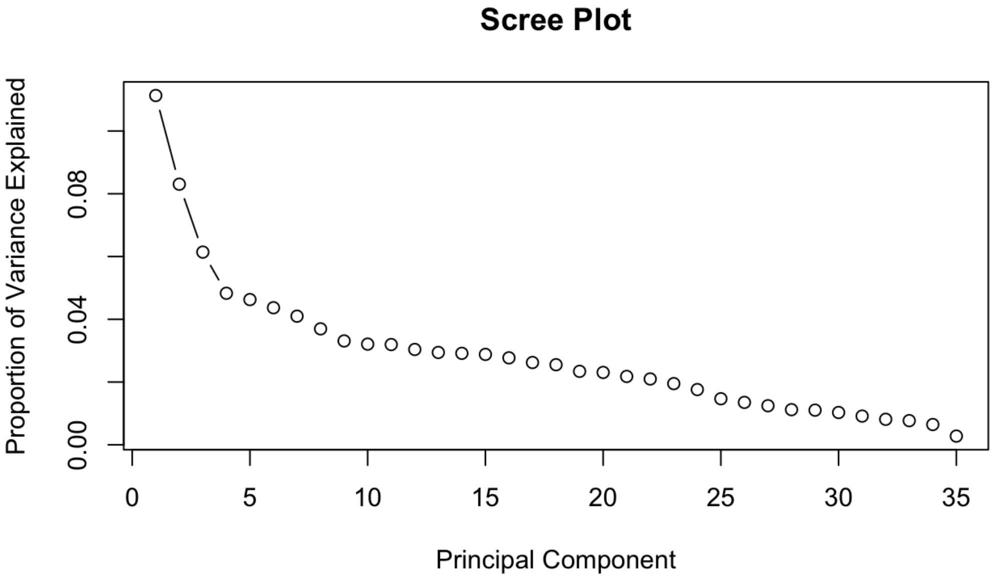
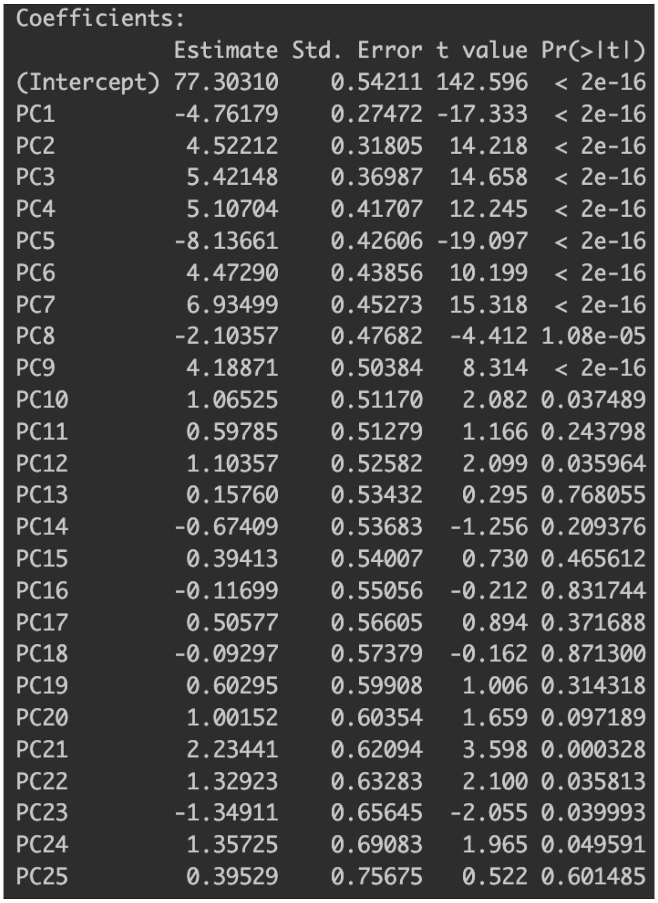


Figure 5

The Scree Plot shows the 35 principal components and the proportion of variance explained in the training data set. Looks like we have a few elbow points that we can point to in our scree plot. We see the proportion of variance explained tends to level out after the 5th, 10th, 15th, 19th, and 25th principal components. The drop-off in the proportion of variance explained is relatively small when comparing those principal components. However, the proportion of variance explained seems to level off after the 25th principal component. We then changed the scaled-trained data set to a new data set with those 25 principal components and added the response variable as well. We also did the same for our test data. Our training and testing data are not ready for input in our models.

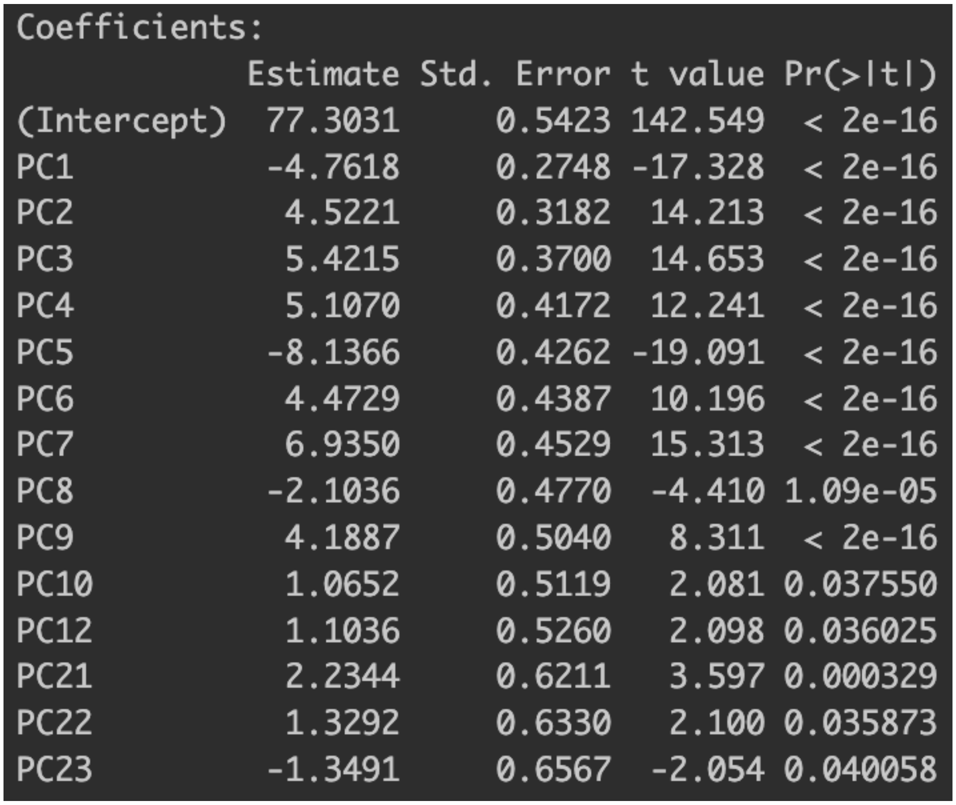
For our initial model, we wanted to fit a multiple linear regression on our new data set comprising the 25 principal components. We chose to do this using the backward selection approach. We then fit a multiple linear regression model using a p-value of 0.05 as our statistically significant threshold. This means that if a principal component predictor has a p-value less than 0.05, we reject the null hypothesis that the predictor has no statistical significance on our predictor. Thus, this means that we consider the predictor to have statistical significance. *Figure 6.1* below represents the summary of our linear model and the corresponding p-values of each principal component.



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 77.30310    0.54211 142.596  < 2e-16
PC1         -4.76179    0.27472 -17.333  < 2e-16
PC2          4.52212    0.31805  14.218  < 2e-16
PC3          5.42148    0.36987  14.658  < 2e-16
PC4          5.10704    0.41707  12.245  < 2e-16
PC5         -8.13661    0.42606 -19.097  < 2e-16
PC6          4.47290    0.43856  10.199  < 2e-16
PC7          6.93499    0.45273  15.318  < 2e-16
PC8         -2.10357    0.47682  -4.412 1.08e-05
PC9          4.18871    0.50384   8.314  < 2e-16
PC10         1.06525    0.51170   2.082 0.037489
PC11         0.59785    0.51279   1.166 0.243798
PC12         1.10357    0.52582   2.099 0.035964
PC13         0.15760    0.53432   0.295 0.768055
PC14        -0.67409    0.53683  -1.256 0.209376
PC15         0.39413    0.54007   0.730 0.465612
PC16        -0.11699    0.55056  -0.212 0.831744
PC17         0.50577    0.56605   0.894 0.371688
PC18        -0.09297    0.57379  -0.162 0.871300
PC19         0.60295    0.59908   1.006 0.314318
PC20         1.00152    0.60354   1.659 0.097189
PC21         2.23441    0.62094   3.598 0.000328
PC22         1.32923    0.63283   2.100 0.035813
PC23        -1.34911    0.65645  -2.055 0.039993
PC24         1.35725    0.69083   1.965 0.049591
PC25         0.39529    0.75675   0.522 0.601485
```

Figure 6.1

Using backward selection, we removed each predictor that had a p-value greater than 0.05. We did this one Principal Component at a time and then refitted our model. In doing so, we were left with the Principal Components shown below in *Figure 6.2*.



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  77.3031     0.5423 142.549  < 2e-16
PC1          -4.7618     0.2748 -17.328  < 2e-16
PC2           4.5221     0.3182  14.213  < 2e-16
PC3           5.4215     0.3700  14.653  < 2e-16
PC4           5.1070     0.4172  12.241  < 2e-16
PC5          -8.1366     0.4262 -19.091  < 2e-16
PC6           4.4729     0.4387  10.196  < 2e-16
PC7           6.9350     0.4529  15.313  < 2e-16
PC8          -2.1036     0.4770  -4.410 1.09e-05
PC9           4.1887     0.5040   8.311  < 2e-16
PC10          1.0652     0.5119   2.081 0.037550
PC12          1.1036     0.5260   2.098 0.036025
PC21          2.2344     0.6211   3.597 0.000329
PC22          1.3292     0.6330   2.100 0.035873
PC23         -1.3491     0.6567  -2.054 0.040058
```

Figure 6.2

Since all of our principal components had p-values less than 0.05, we then tested the MSE rate of both the training and the test data. Our MSE rate for our training data was 592.28 and our MSE rate for our test data was 1,846.93. These seemed quite high for our initial model. Our ultimate goal is to get our MSE rate to be as close to 0 as possible, all while balancing our variance-bias tradeoff.

For our MSE to be so high, we also wanted to test the model assumptions for multiple linear regression. The 4 main assumptions for linear regression are linearity, independence, homoscedasticity, and normality. All 4 assumptions can be tested using diagnostic plots such as residual plots, Q-Q plots, scale-location plots, and residuals vs. leverage plots. *Figure 6.3* shows that the data set agrees with all of the assumptions for linear regression.

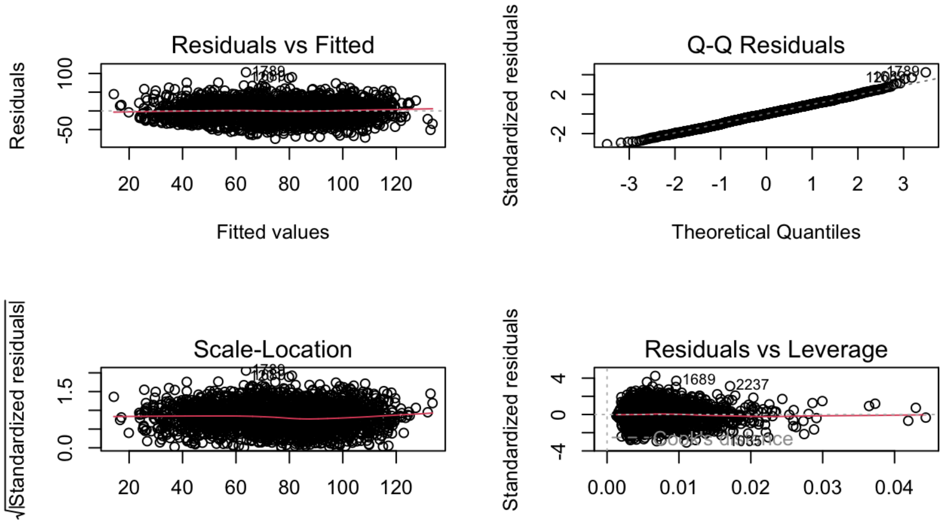


Figure 6.3

Another thing that we can look at is the predicted values versus the actual values. Ideally, our predicted versus actual value plot should follow a linear pattern. Let us now look at the predicted values versus the actual values for both the training and test values. *Figure 7.1* represents the training data and *Figure 7.2* represents the test data.

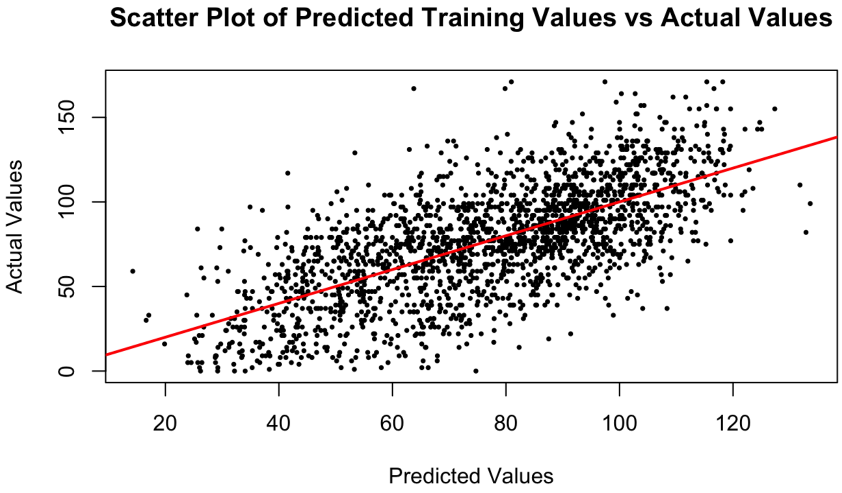


Figure 7.1

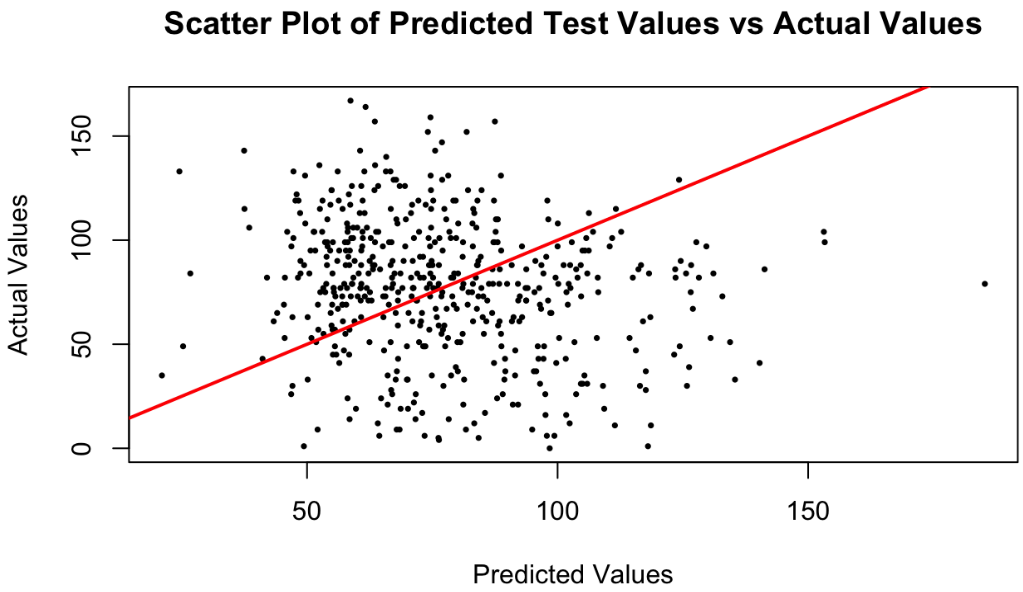


Figure 7.2

We see that our predicted versus actual value plots reflect the respective training and test MSEs. Our linear model seems to be overfitting our training data and is not performing well on our test data. The training plot follows a much more linear pattern than our test plot. Therefore we concluded that our initial linear model may not be the best, but that it was a good starting point. We then wanted to see how we could better our training/test MSEs.

Model Tuning, Validation & Exploration

We wanted to fit a few other models to see how a lasso regression model and a regression tree model would perform on our data. Ultimately, each model produced similar results to our multiple linear regression model. *Figure 8* below shows a snippet of our regression model on our 25 principal components.

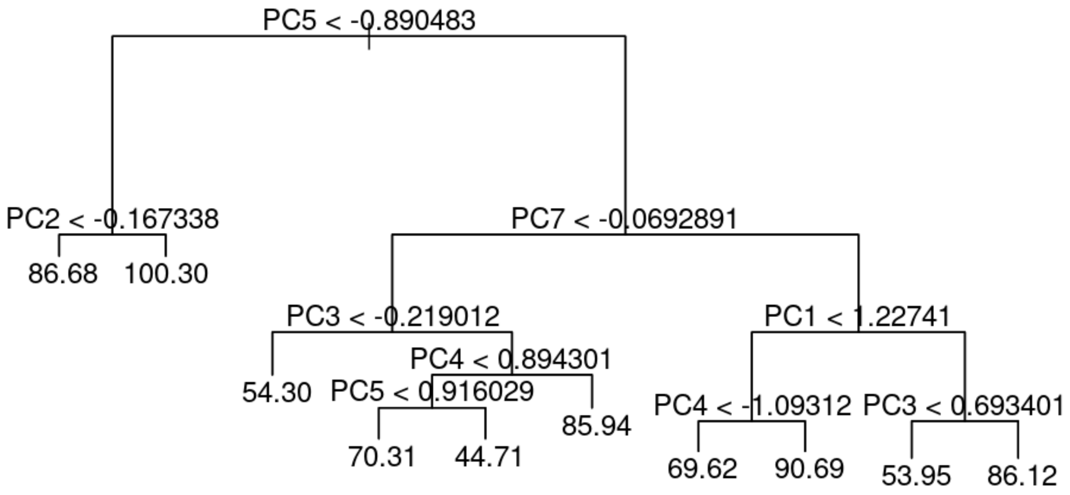


Figure 8

Our regression tree model above shows that the most statistically significant predictors in our model are PC5, PC2, and PC7. Our regression model produced a training MSE of 772.43 and a test MSE of 1417.15.

The lasso regression model was tested to see if regularization would improve upon the performance of multiple linear regression. Our lasso model used cross-validation to select a lambda value of 0.28 to fit the model. The final model then produced a training MSE of 590.35 and a test MSE of 1799.89. Based on these initial results, there could be an issue where the model is severely underfitting the data.

With the models chosen so far performing poorly to the MSE, we decided to pivot over to a feed-forward neural network. Neural networks (NN) are great at reducing dimensionality meaning that performing PCA on our data set beforehand will not be necessary. *Figure 9.1* below shows our mean squared error and loss on our training and validation data. The validation data acts as the test data within the training data.

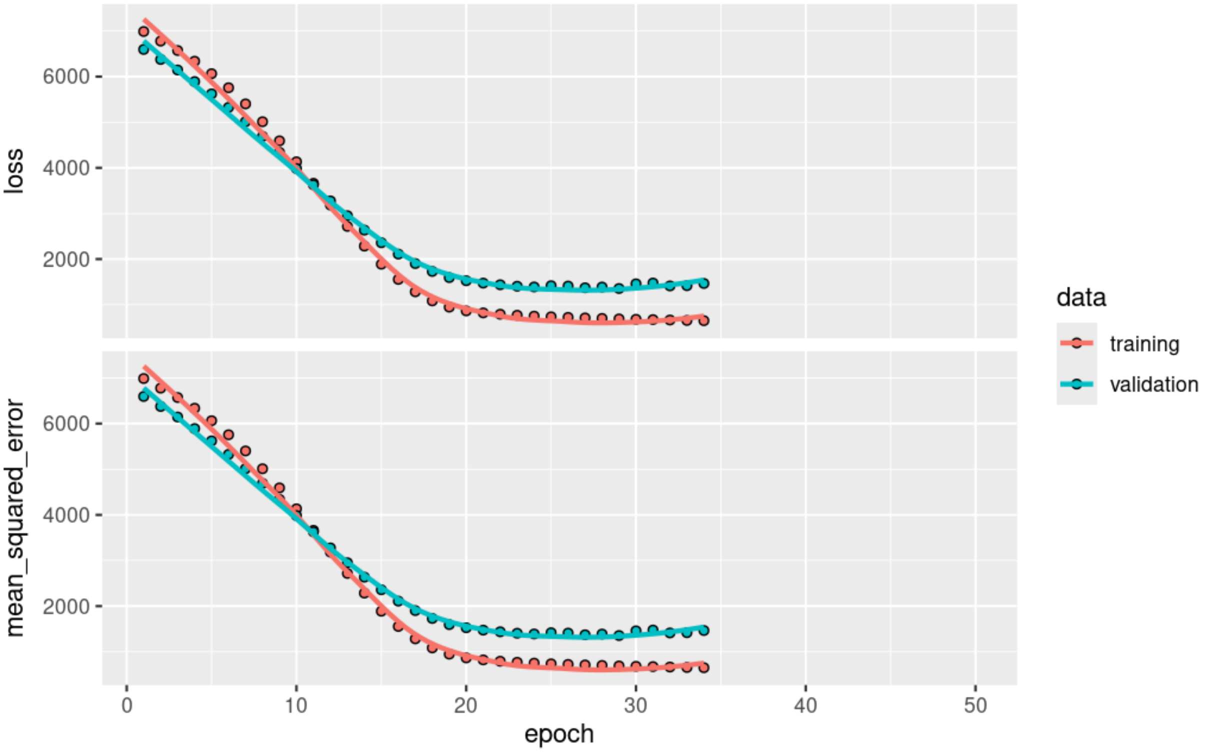


Figure 9.1

The initial NN model used 3 hidden layers with 75, 50, and 25 nodes respectively. We used “relu” for our hidden layer activation function and “linear” for our output layer activation function. The NN included “rmsprop” as the optimizer, “MSE” for our loss, 5 for our patience, 50 epochs, batch size 512, and a validation split of 0.20. This network resulted in a training MSE of 748.92 and a test MSE of 807.09. We used tuning methods to try and get better results. The tuning performed on the model included adding dropout layers at a rate of 0.5 which will help the model from overfitting, Adam optimizer which integrates momentum and RMSprop into one algorithm, learning rate of 0.001, increasing the number of epochs, adjusting batch size to 256, and increasing our patience from 5 to 10 stopping the NN if there isn't an improvement on the MSE within 10 epochs. *Figure 9.2* below shows the mean squared error, loss and learning rate of our fine-tuned NN model. This fine-tuned NN resulted in a training MSE of 841.92 and a test MSE of 866.96.

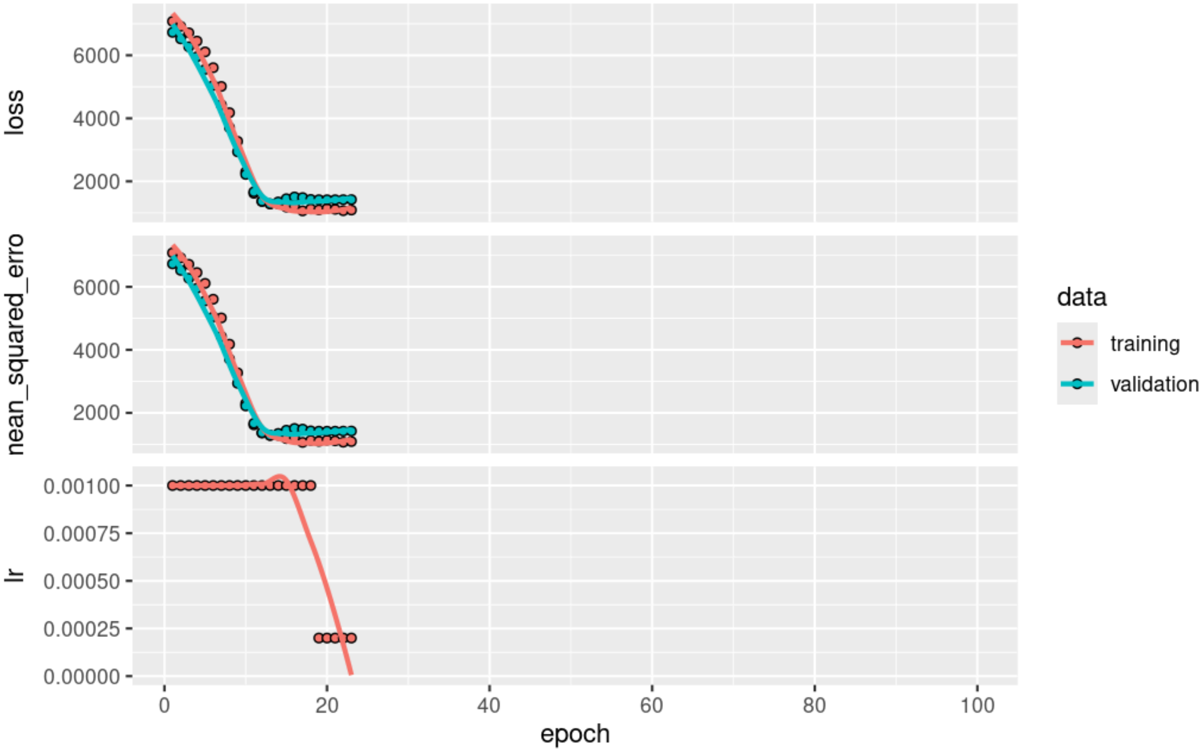


Figure 9.2

The NN showed increased performance on the data which led to our final model, the support vector machine (SVM). This model was chosen because SVM is more interpretable than a feed-forward NN. Tuning the SVM will require understanding the hyperparameters of this model. The cost of the model controls the trade-off between achieving a low error on the training data and minimizing model complexity. The epsilon of the model defines the margin of tolerance where no penalty is applied for error in support vector regression (SVR). The degree is for the polynomial kernels. Gamma determines the shape of the decision boundary. With this understanding of the hyperparameter, we came up with 2 models with the latter model using a cross-validation of 10. The first model had a tune of 75.025 cost, 0.4 epsilon, and 0.1 gamma which resulted in a training MSE of 127 and testing MSE of 761. The second model had a tune of 50.05 cost, 0.4 epsilon, and 0.1 gamma which resulted in 127.69 training MSE and 761.54 testing MSE. Based on the similar performance of both models we will use a model with less cross-validation for faster computing time.

The SVM model has the best performance. This will be our final model used and some interpretations taken from these results are related to model complexity, the risk of overfitting, and the epsilon parameter. The high cost of 75.25 and a relatively low gamma of 0.1 likely means that our model is complex but fits the data closely. The summary function shows that the SVM model had 1242 support vectors which suggests that the model may be complex with data points close to the margin. The combination of parameters suggests that the model could potentially overfit the data, especially with a high cost and relatively large number of support vectors. The epsilon value of 0.4 shows that the model is allowing some deviations in predictions, but the margin is relatively tight, meaning the model is trying to predict quite accurately.

Exploration

Using our SVM we were able to make predictions about our test data set. Using visual and some test statistics we are able to look further into the performance of our model.

Variable	Mean	Median	SD	Correlation Coeff
Total Anxiety Level	42.10099	44.00000	15.93874	0.5411549865
Predictions	77.70883	76.45324	16.16888	

Figure 10.1

As shown above in *Figure 10.1*, we can see the spread of our two variables, total_anxiety_level, and the predicted values of that variable. The mean and median are higher in our predictions and had a correlation coefficient of approximately 0.54115. Ideally, we would want our correlation plot to show a diagonal line. This is shown in *Figure 10.2* below. If the predicted value is 5, the actual value should be close to 5, in this case, it is not.

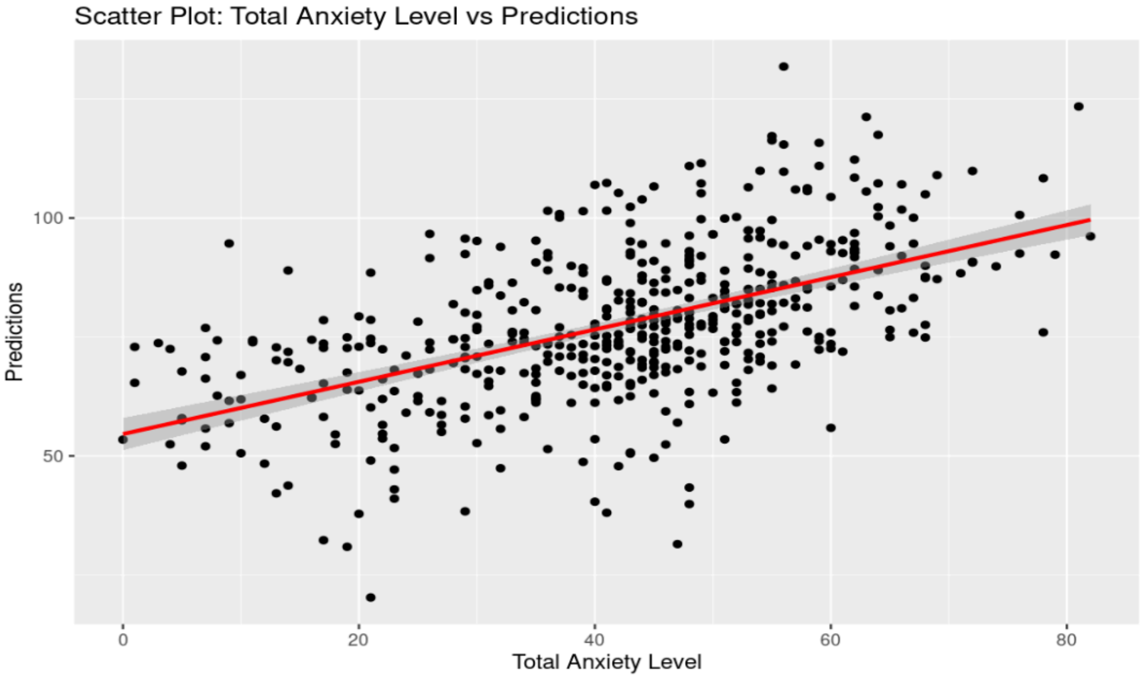


Figure 10.2

Our residual plots show that our predicted values are negative and large. This suggests that our model is severely overpredicting the values for total_anxiety_level. Most of the values are negative and large, for example, values of -60 suggest that the predicted value is 60 units higher than actual values as the residuals are calculated actual-predicted.

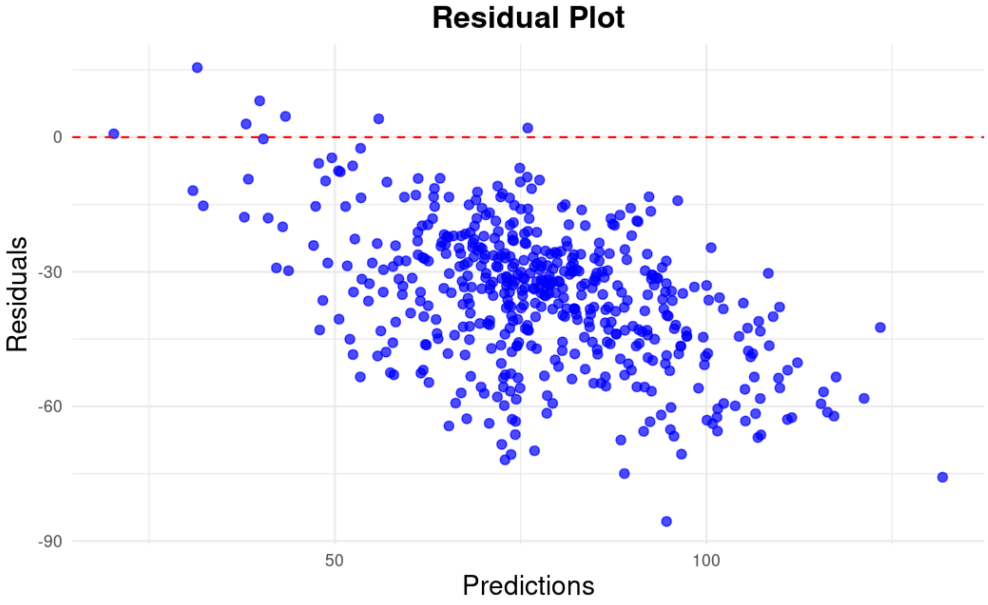
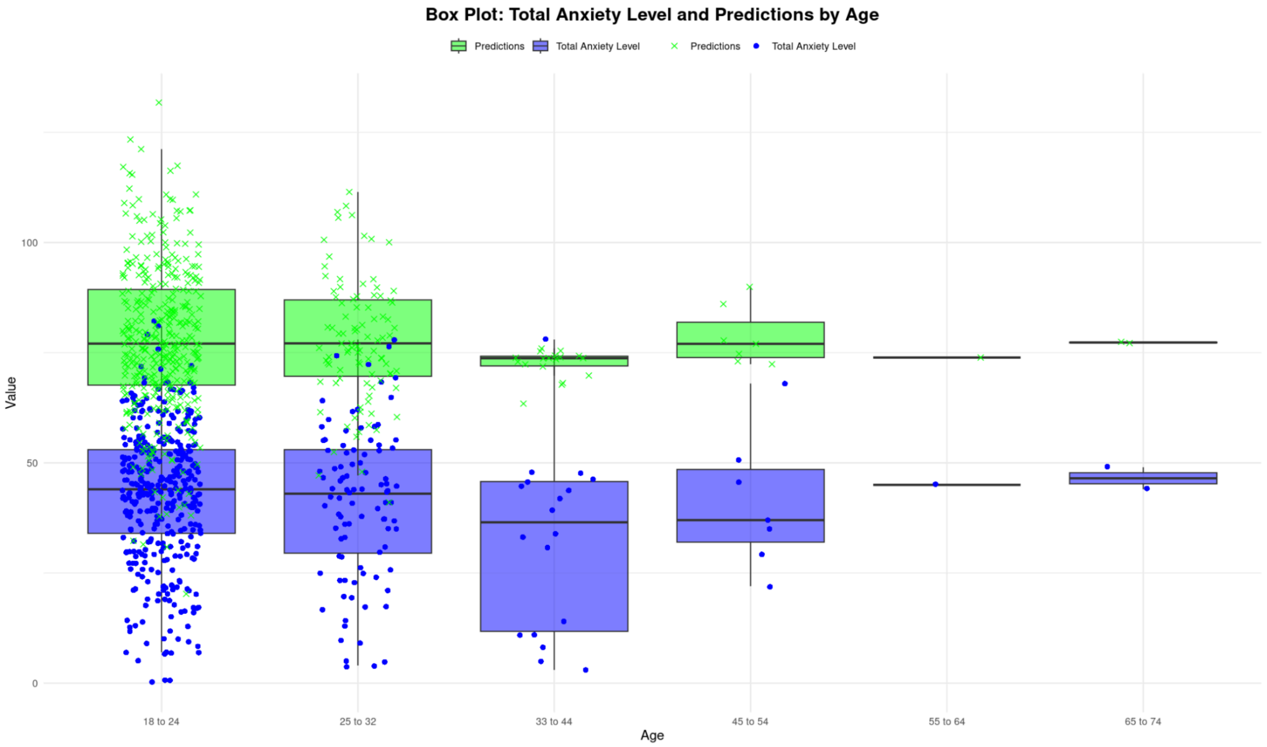
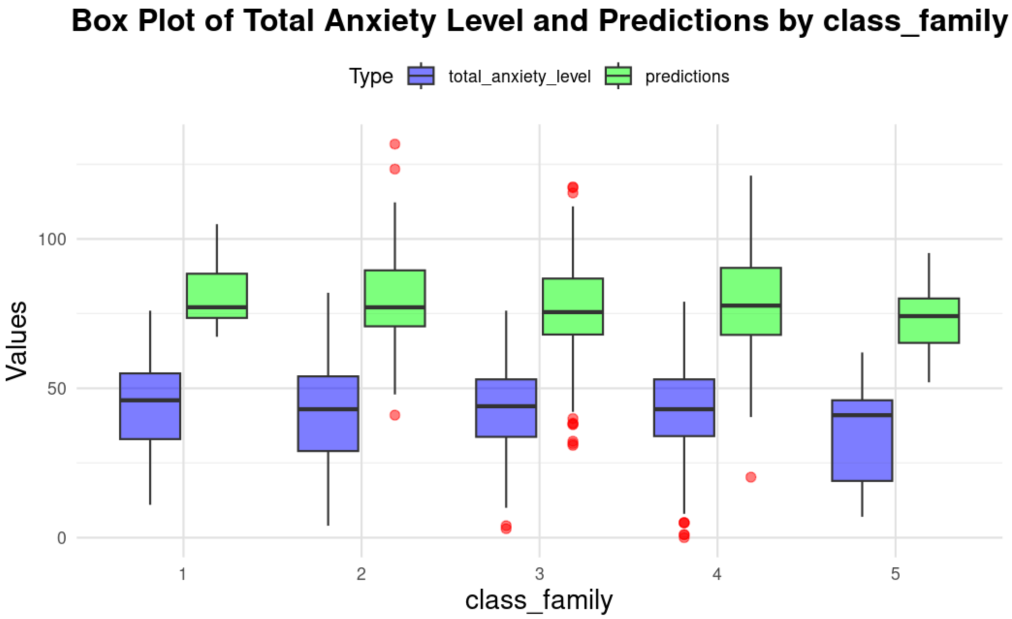
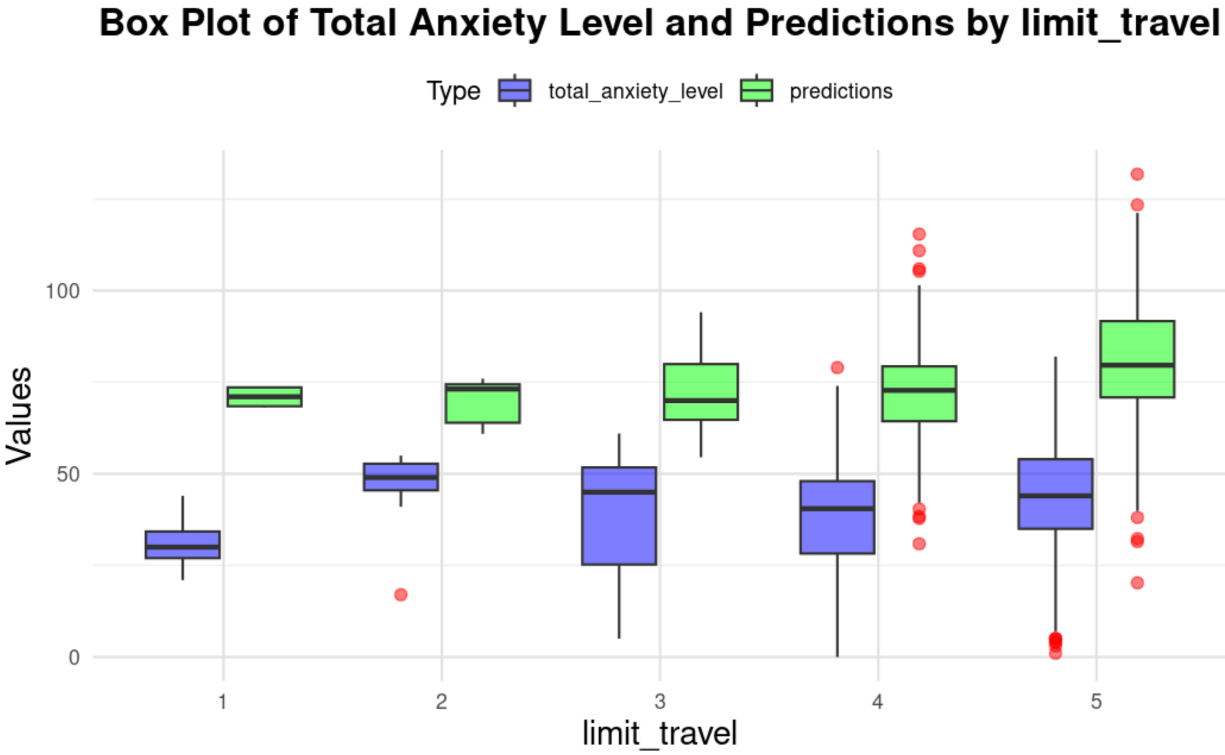


Figure 10.2 above & Figure 10.3 below





Figures 10.4 above & Figure 10.5 below



When we compare the predicted values vs actual values broken down by factors within different variables, we can see the severity of the over-predictions. As seen in *Figures 10.4-10.6,* for variables such as class_family, age, hrs_exercise, and limit_travel, the predicted values are always higher than actual values.

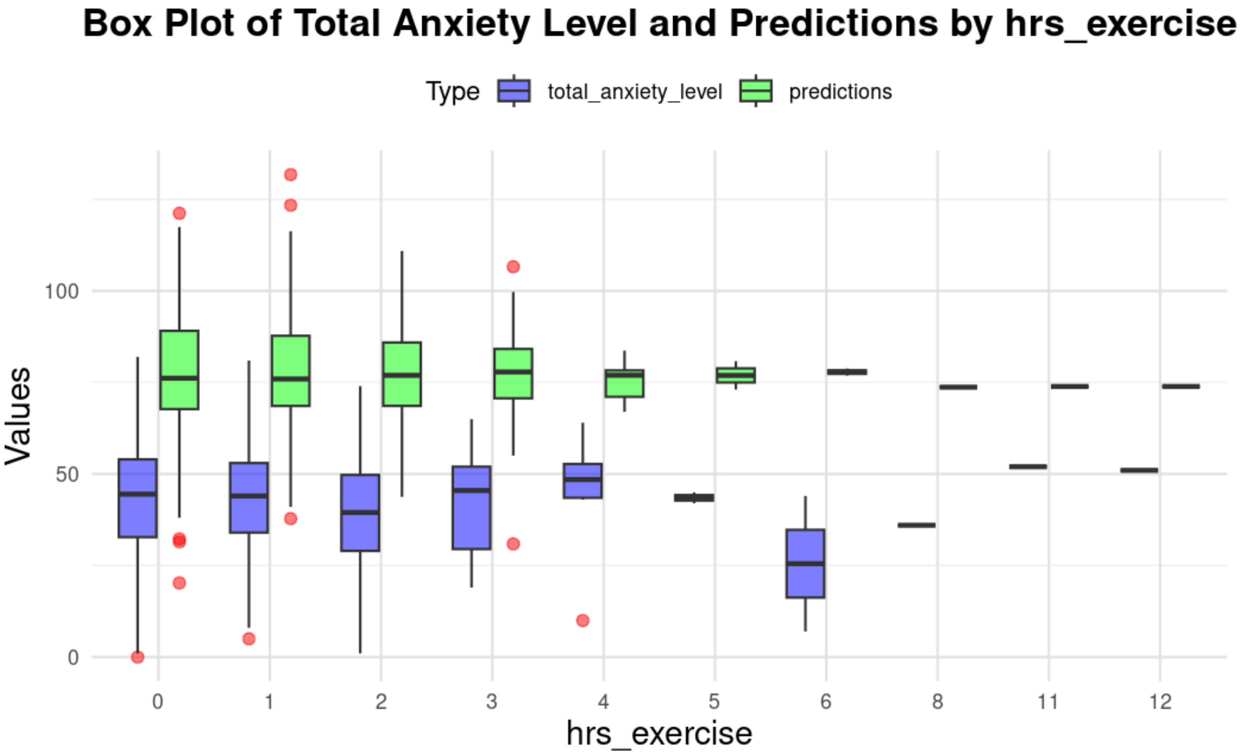


Figure 10.6

To further understand the increase in predicted values, we wanted to compare the mean, median, and standard deviation for all the variables and factors within the variables. For that, we took all values from the actual vs predicted values of total_anixety_level and plotted them against each other. In *Figure 11* we can see an upward trend of correlation between all the plots. As the actual total_anixety_level increases for the mean, median, and standard deviation, the predictions also increase. The approximate correlation coefficient of 0.54115 suggests that there is a moderate linear relationship between the actual and predicted values but a value close to 1 would be ideal. This furthers our understanding of the underfitting of the model. The model is not complex enough to capture the underlying patterns within our dataset, so further hypertunning of our model is required.

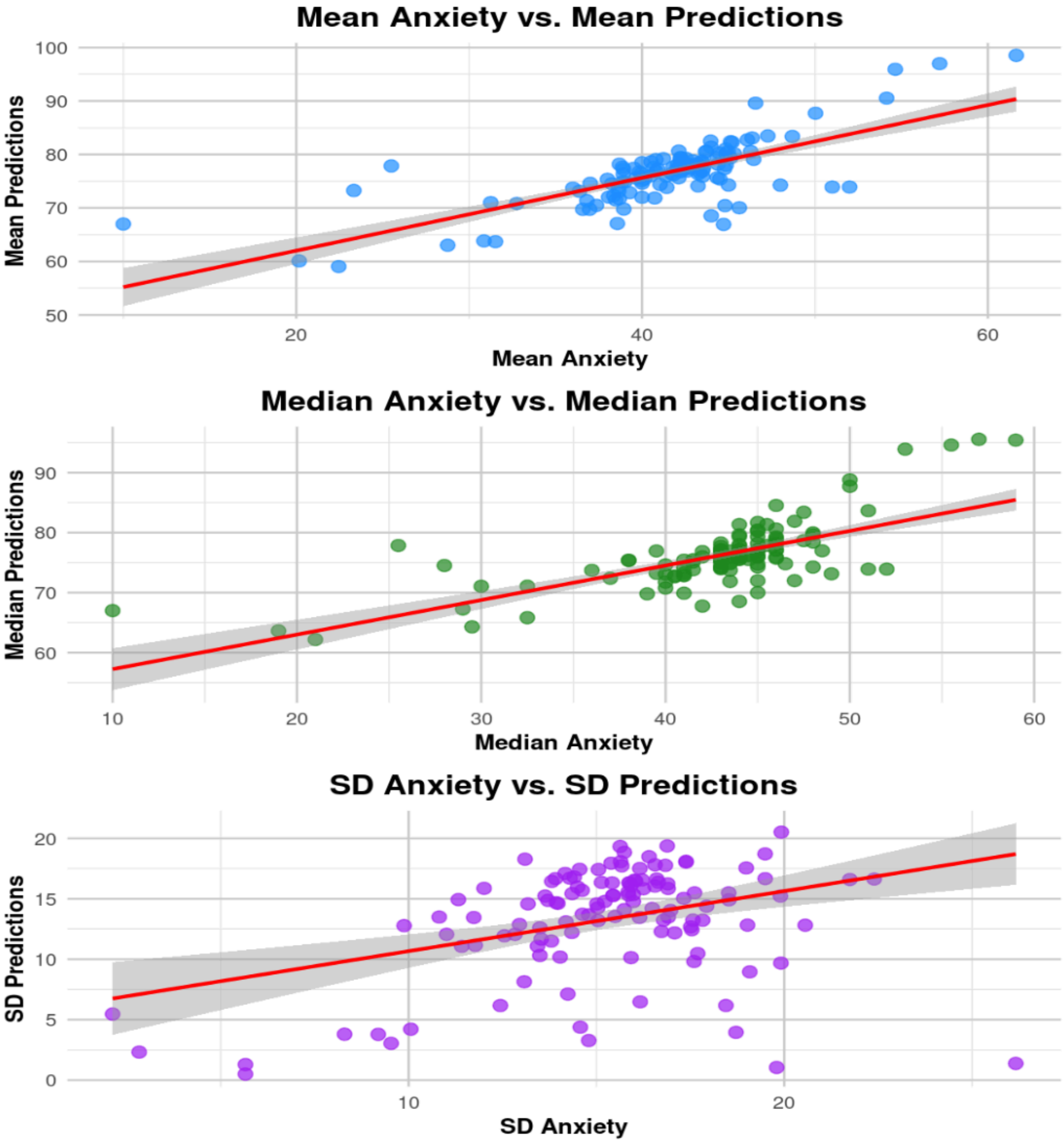


Figure 11

There is also a possibility that the features we used to build our model might not be strong predictors of our target variable. The possibility of adding more details regarding the students would have given more variables to use as additional predictors (degree, year attending university, location of residence (on or off campus), relationship status, etc). A lack of linear relationship between the variables might also have posed an issue for the model in terms of capturing the complexity. During our EDA, we removed the 4 variables that were similar to each other and were the only variables with a correlation coefficient of over 0.7. After removing outliers we still might have noise and highly influential points in our dataset, so further EDA and hypertuning of our models is required.

Our initial idea was to create a model that would be able to predict the anxiety level of students. While we improve our model and reanalyze our dataset while performing EDA we can still use our findings to create protocols and programs to help susceptible during times of crisis such as a pandemic. During our analysis, we can see some factors within a variable show a greater score for total anxiety level such as female vs male, ages 18-24 vs 33 to 44, and limit travel 1 vs 5. This information can be used to create bulletin boards for students to help overcome stress or even programs to help certain groups come together during stressful times. If waiting times are an issue, individuals who fall into these categories can have priority in receiving help.

Conclusion

This project is centered on understanding the psychological impact of COVID-19 on U.S. college students, particularly anxiety levels, that are aimed to predict future anxiety levels in similar crises. The stakeholder for this project is a government-funded relief group focused on mitigating the pandemic's effect on students. The dataset used in this project was obtained from a study involving 2,534 observations from 14,147 surveys across seven U.S. universities. The variable of interest or response variable is "total_anxiety_level," which aggregates scores from six stress-related feelings (afraid, irritable, guilty, sad, preoccupied, and stressed) into one measure. This data set was cleaned by removing redundant variables and combining stress-related variables into the "total_anxiety_level" response variable.

Our exploratory data analysis included correlation matrices, variance inflation factor testing, and distribution analysis using Q-Q plots and Shapiro-Wilk tests, leading to a Box-Cox transformation of the response variable to achieve a more normal distribution. Our research uses a quantitative regression model, initially exploring multiple linear regression, then refining the model through techniques such as linear regression, lasso regression, regression trees, neural networks, and finally support vector machines (SVM).

Multiple linear regression and other models like lasso regression and regression trees were tested, but these had relatively high mean square errors (MSE). The SVM model was found to be the most effective, balancing complexity and predictive accuracy. The final SVM model had a tuned cost and epsilon value, leading to a lower MSE on the test data. This model was preferred for this research based on the performance which allows for accurate prediction of students' anxiety levels based on their responses to non-COVID-19-specific questions, which helps in identifying those at risk without directly addressing COVID-19. This approach proves a valuable tool for institutions to proactively support students in managing anxiety during future crises, enhancing the relevance and applicability of the research in other areas of mental and public health.

Appendix

Code Availability:

https://github.com/DanielJackson2155/DSE6311_Capstone_Project/tree/main

Data Dictionary:

**Response variable

Variable Name	**Variable Type**	**Variable Definition**
limit_outdoor_active	**Qualitative - Ordinal**	**How COVID-19 has limited a student's outdoor activities: 1(Strongly disagree) - 5(Strongly agree)**
limit_exercise_home	**Qualitative - Ordinal**	**How COVID-19 has limited a students exercise at home: 1(Strongly disagree) - 5(Strongly agree)**
limit_exercise_gym	**Qualitative - Ordinal**	**How COVID-19 has limited a students exercise at the gym: 1(Strongly disagree) - 5(Strongly agree)**
limit_people	**Qualitative - Ordinal**	**How COVID-19 has limited a students ability to meet people: 1(Strongly disagree) - 5(Strongly agree)**
limit_groups	**Qualitative - Ordinal**	**How COVID-19 has limited a students ability to meet in groups: 1(Strongly disagree) - 5(Strongly agree)**
limit_travel	**Qualitative - Ordinal**	**How COVID-19 has limited a students travel: 1(Strongly disagree) - 5(Strongly agree)**
health_general	**Qualitative - Ordinal**	**General Health of the Student: 1(poor) - 5(excellent)**

covid_worry	**Qualitative - Ordinal**	**"I worry about the coronavirus all of the time." 1 (Strongly disagree) – 7 (Strongly agree)**
covid_lot_time	**Qualitative - Ordinal**	**"I spend a lot of time thinking about coronavirus." 1 (Strongly disagree) – 7 (Strongly agree)**
total_anxiety_level **Response variable	**Qualitative - Ordinal**	**“covid_afraid,” “covid_irritable,” “covid_guilty,” “covid_sad,” “covid_preoccupied,” and “covid_stressed” were added together than divided by 6 to get the average score**
hrs_screen	**Quantitative - Continuous (might round up/down to make discrete)**	**Number of Screen time (hours)**
hrs_outdoor	**Quantitative - Continuous (might round up/down to make discrete)**	**Number of Outdoor time (hours)**
hrs_exercise	**Quantitative - Continuous (might round up/down to make discrete)**	**Number of Exercise time (hours)**
educ_mom	**Qualitative - Ordinal**	**Education Achievement for Mom: 1(less than highschool) - 7 (Doctorate)**
educ_dad	**Qualitative - Ordinal**	**Education Achievement for Dad: 1(less than highschool) - 7 (Doctorate)**
educ_self	**Qualitative - Ordinal**	**Education Achievement for Self: 1(less than highschool) - 7 (Doctorate)**
income_relative	**Qualitative - Ordinal**	**Relative Family Income**

class_self	**Qualitative - Ordinal**	**Social Class for Self: 1(Working Class) - 5(Upper Class)**
class_family	**Qualitative - Ordinal**	**Social Class for family. Takes the average of class_mom and class_dad rounded up to the nearest whole number to give social class for the entire family.**
park_use	**Quantitative - Continuous**	**Hours of Park Use**
infected_any	**Qualitative - Nominal (Binary)**	**If the student knew if anyone was infected with COVID-19**
female	**Qualitative - Nominal**	**If a Student is a female or not**
bmi	**Quantitative - Continuous**	**Body Mass Index**
age	**Qualitative - Nominal**	**The age group that the student is in**
source	**Qualitative - Nominal**	**Which University that the observation comes from**
ethnoracial_group	**Qualitative - Nominal**	**Ethnoracial Group that student belongs to**

Response Variable Transformation Histograms and Q-Q Plots

Square Root

Shapiro-Wilk Test p-value: p-value < 2.2e-16

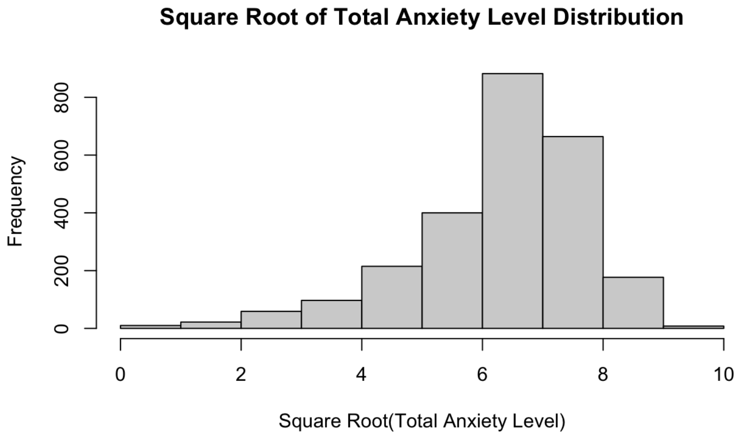


Figure 12.1

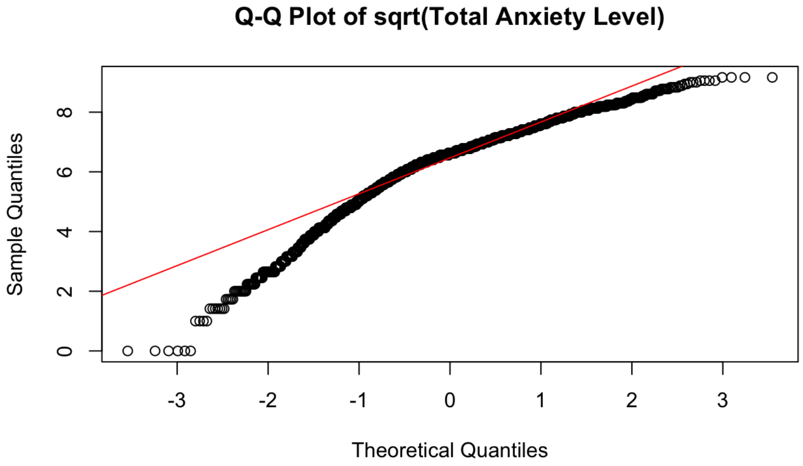


Figure 12.2

Squared

Shapiro-Wilk Test p-value: p-value < 2.2e-16

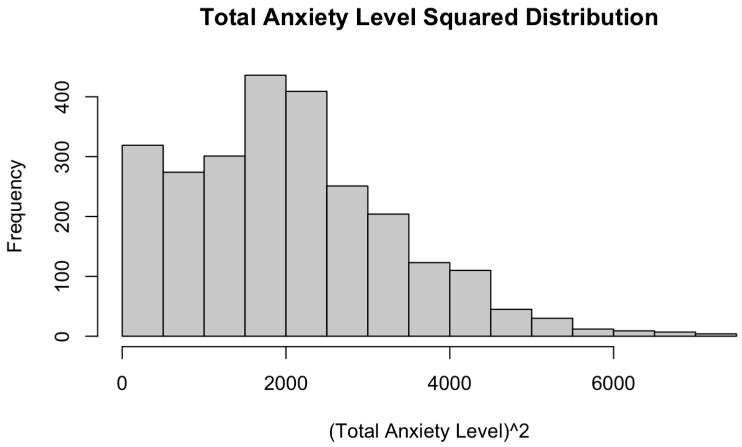


Figure 12.3

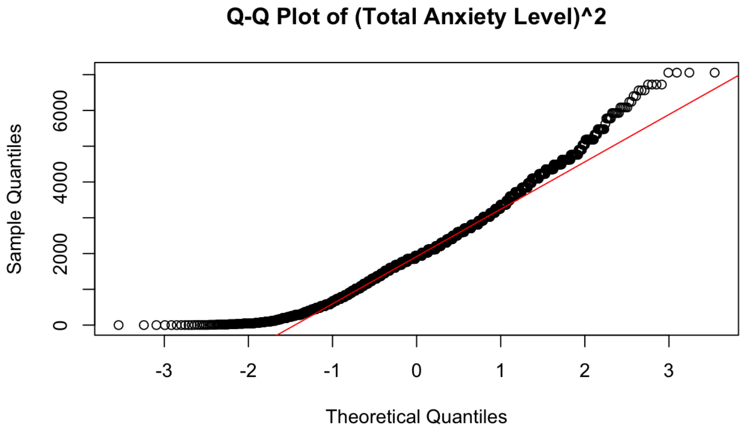


Figure 12.4

Log Transformation

Shapiro-Wilk Test p-value: NA

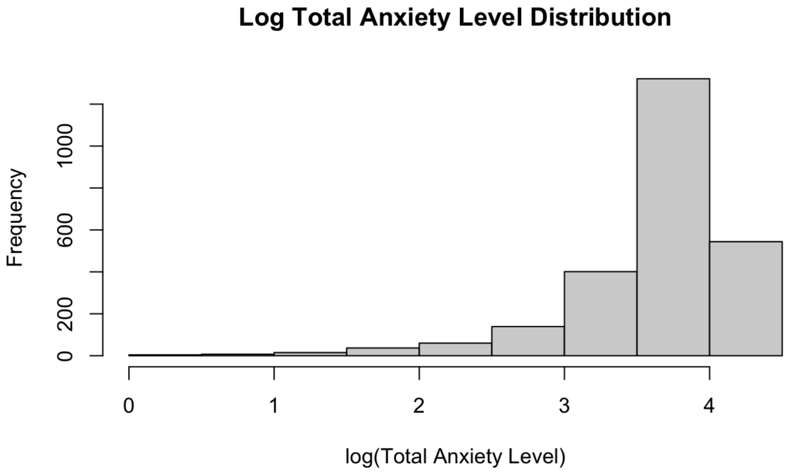


Figure 12.5