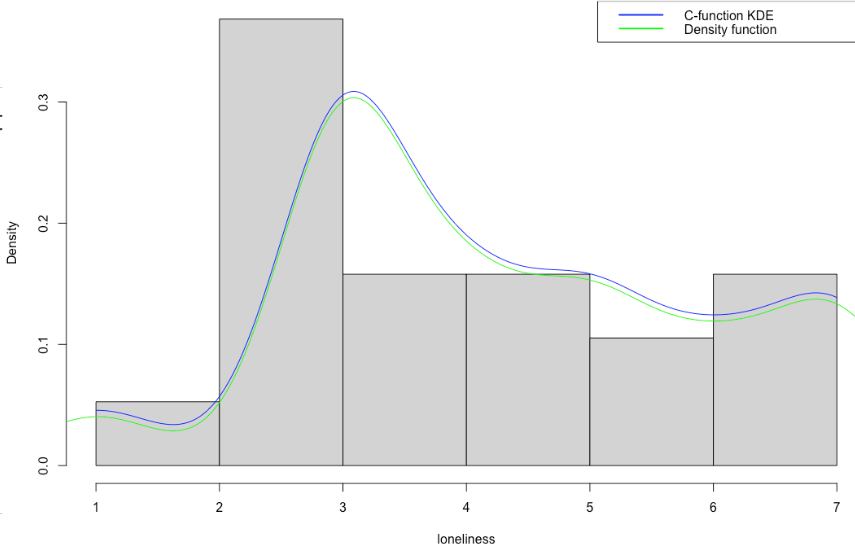
Daniel Smith

Stats 202A

Fall 2023

Professor Schoenberg

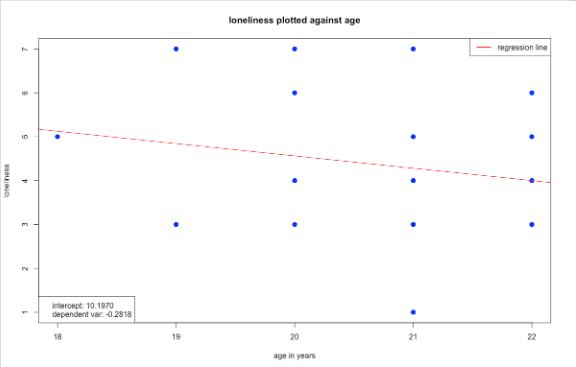
**Final Project**

**Introduction:**

For this project I decided to do an analysis on Dating patterns. The dataset I obtained was from [<https://www.kaggle.com/datasets/annavictoria/speed-dating-experiment>]. This dataset was collected from 2002-2004 from two Colombia Business School Professors, Ray Fisman and Sheena Iyengar. The number of participants totaled to around 8,378, each participant filling out 195 questions. This dataset contained people from all backgrounds, but I decided to narrow down the dataset to only college students, which left the total number of participants to 190. Using some python analysis, after further investigation of the data, all the N/A values were dropped leading to a total number of valid participants to 171. After that I decided to one hot encode the categorical variables and leave the numerical ones alone. Using the cleaned data I now decided to select certain features to predict the frequency of dates an individual goes on, or what I call as “Loneliness” Score. The features I selected were: race, age, major, and gender. Before training the model I decided to do some investigation on the data.

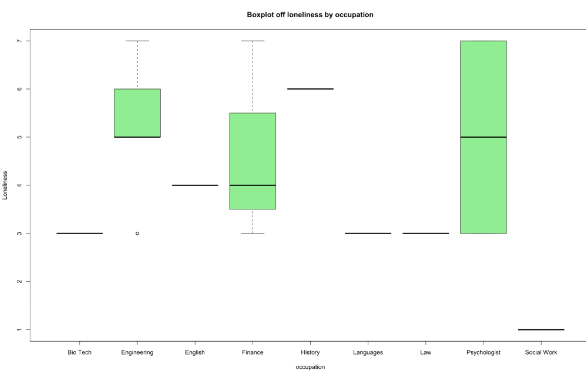
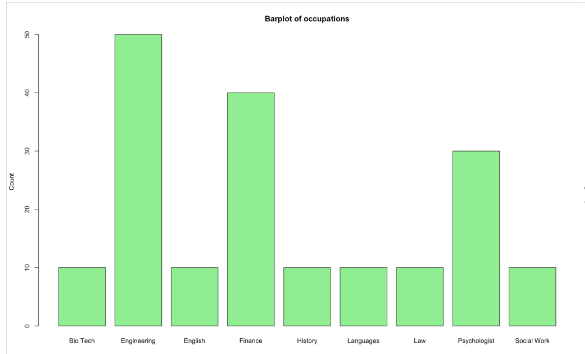
**Data Analysis and Preprocessing:**

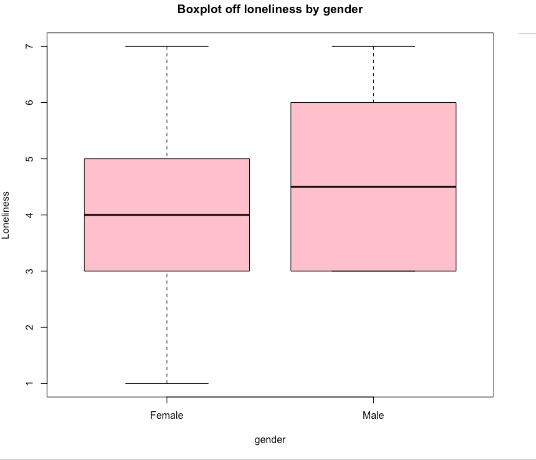
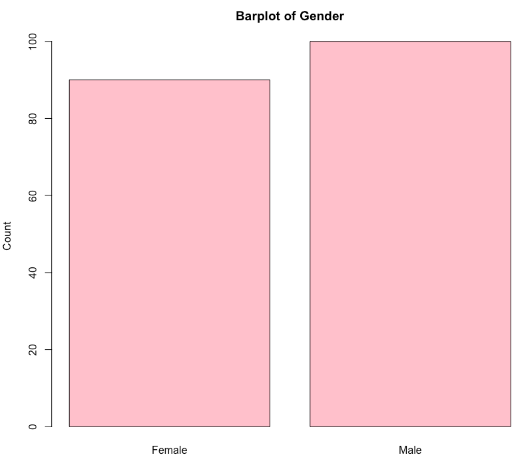
First thing I did was try and plot the distribution of loneliness scores in the college dataset. As seen below the data is right skewed with the plurality of participants ranking from 2-3 in the loneliness factor.This seems intuitive as college students would tend to be very social leading to the right skew we expect.

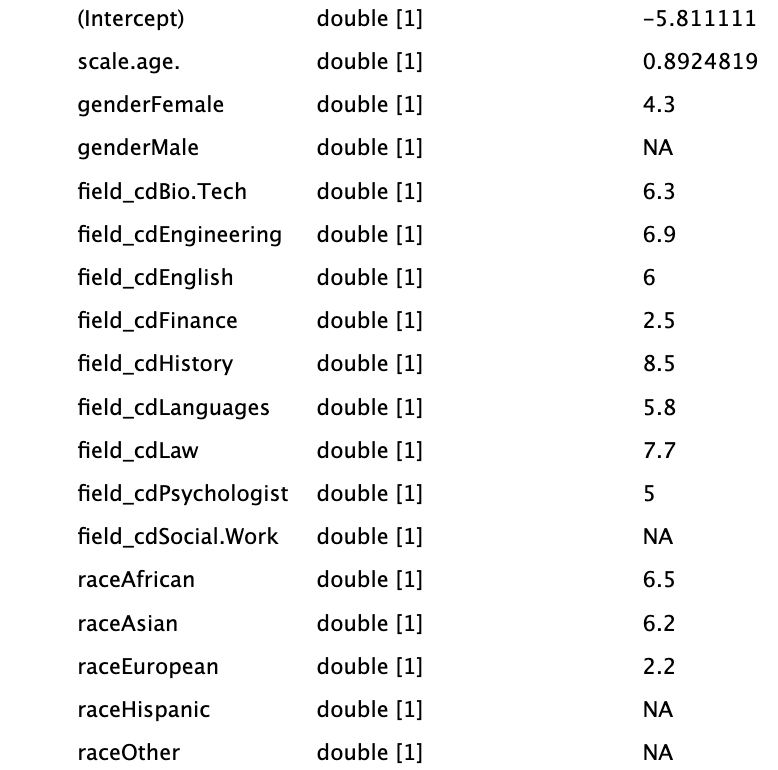
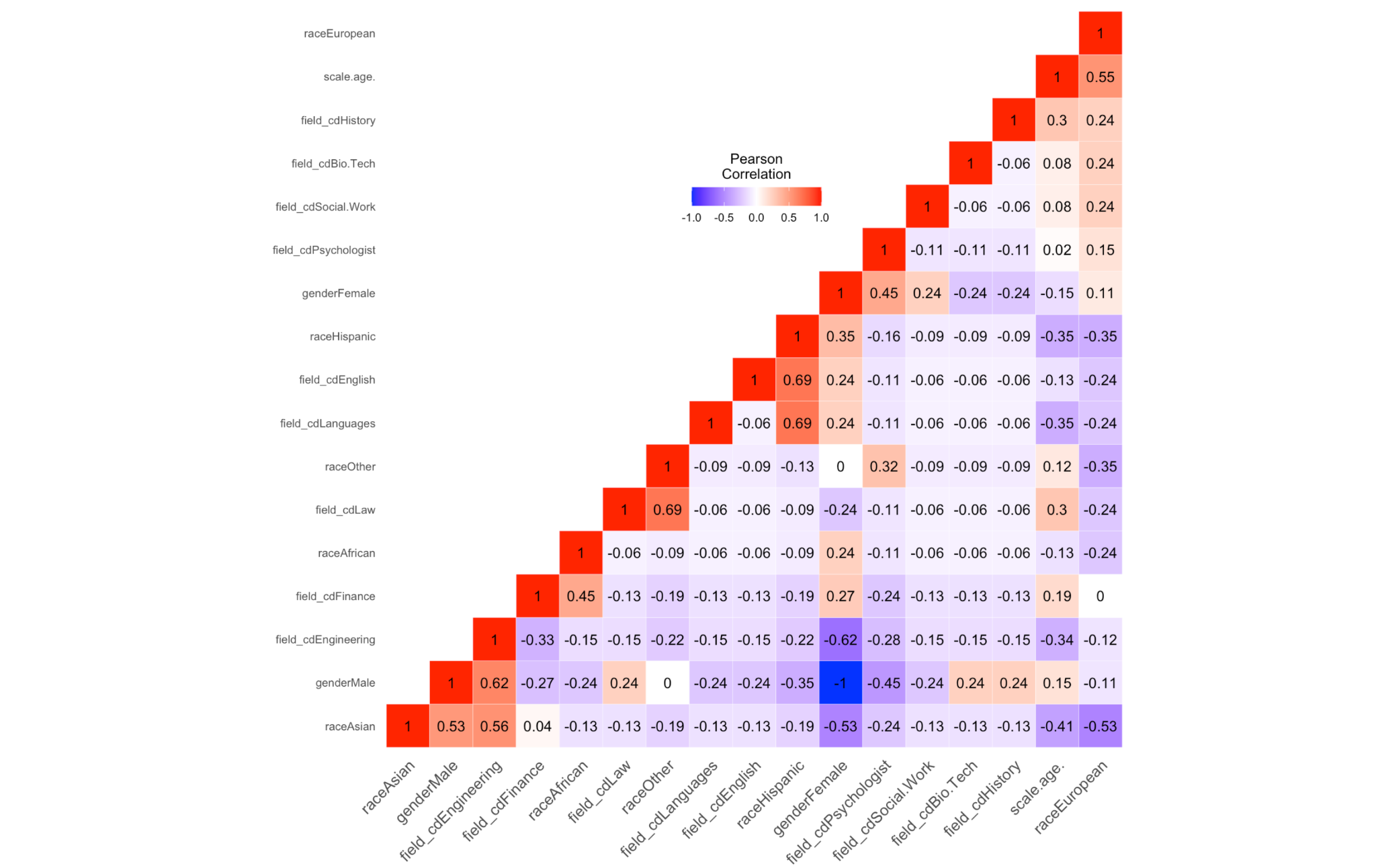


The next thing I wanted to inspect was the relationship between age and loneliness and age. We can see there is a slight negative correlation between the two which makes sense since older students tend to be more experienced and extroverted.

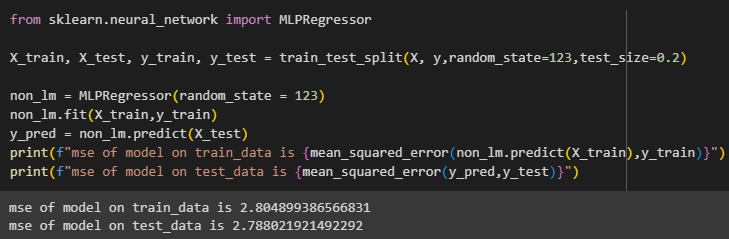
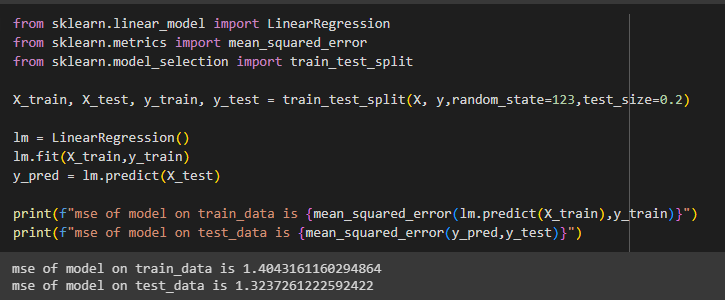
I continued to do some inspection with some of the other variables as shown below. I don't believe there was anything of importance to note. Except the somewhat uneven distribution of majors and their variances when it comes to loneliness

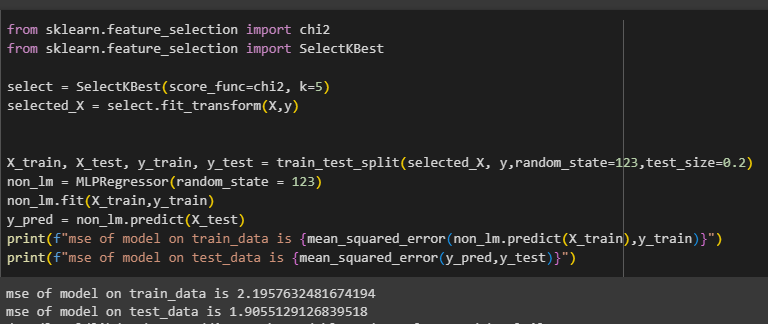




After the inspection is when I decided to create the Model to relate loneliness to all the features we selected earlier. After training the model the weights were shown in the table below. I was initially confused about the N/A values but after further inspection, we can see from the pearson correlation matrix that there were highly correlated variables that influenced the model and as as result some of the were left as N/A

**Model Building and Predictions:**

After the presentation that was given I wanted to further investigate to see if the Linear model was the best that could be done. For this I decided to code in Python a base Linear Model, a Multilayer Perceptron (Neural Network), and Multilayer Perceptron using Feature selection based off of a Chi-Squared Distribution. Snippets of the code are shown as such: 

The first to do was split the college dataset into training and test sets, this way we can reduce the amount of overfitting the models have in general. After that we can train both the Multilayer Perceptron and Linear models and test them using the data the model has not seen from the test set. 

**Conclusion**

The results show that a Linear model is still the best model with a Training MSE of 1.40 and a Test MSE of 1.32. Both MultiLayer Perceptrons performed worse than the Linear model with a Train MSE of 2.80 and 2.19 and a Test MSE of 2.78 and 1.90 for the Vanilla MLP and the Chi-Squared feature selected MLP respectively. The main reason I suspect this happened is because of the highly linearly correlated features we have. While Neural Networks tend to perform better than Linear Model as they can capture Linear and Non Linear relationships, by selecting certain features to compile our dataset I introduced a bias into the dataset which ended up skewing the results of my models. That being said, this dataset was relatively small (171 data points), and had there been more features selected to add and participants, I am confident that the Neural Networks would have performed better. This was a great opportunity to test some of the things we learned in class onto a real life example. For future I would like to try and implement more features to the model, one example would be Sentiment Analysis as each participant wrote a brief description about themselves that would have definitely influenced the outcomes for the model since as anyone who has been in a relationship can attest, that personality makes a difference and using just raw values such as age, race, occupation, and gender can not tell the full story on how successful or “lonely” an individual will be in the dating market. All the code and data are linked in my Github profile [github.com/jakq277].