Work Completed:

Setup of working environment – We have set up a GitHub repository for our project. We've utilized GitHub Pages to easily display the contents of our code as a website. We stored our data using Azure Blob Storage for smooth retrieval of data.

Data Cleaning – We have cleaned and prepared our data by checking for and removing any rows with missing or inconsistent data. We've also removed entries and columns that are irrelevant to our research.

PCA – We performed a principal component analysis on our data to only include the most influential columns. This also narrows down the large quantity of variables in our dataset.

Graph Creation – We have completed the edge creation function which creates edges between nodes based on shared column values. Using the edge creation function we've produced a small graph of 50 nodes as a proof of concept.

Sampling – We have implemented stratified sampling to our dataset and included reasoning for the need and validity of our sampling method.

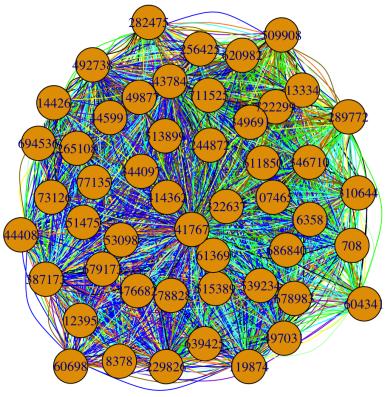
Research Question 1 - Initial approach and strategy thought out, early implementation still in progress.

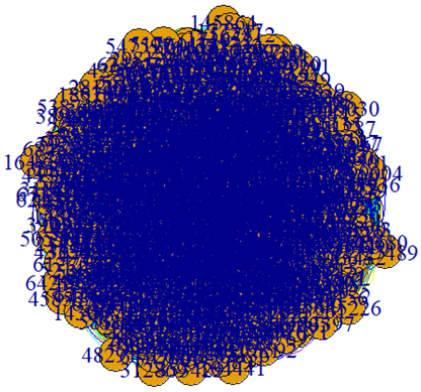
Research Question 2 - Initial approach and strategy thought out, early implementation still in progress.

Research Question 3 - Initial approach and strategy thought out, early implementation still in progress.

Results:

We have been able to create several undirected graphs using samples of our large dataset. The edge colors correspond to the variable which two connected nodes share. These graphs form the basis for our analysis and exploration of research questions.





Challenges:

The main challenge we have faced is the size of our data set. After data cleaning and PCA, our data set contains 743,169 nodes. The computation time required to process this data is immense.

The time required to form edges for a sample of 1000 nodes is one to two minutes and creates approximately 1,500,000 edges. The time required to form edges for a 2000 node sample is approximately 10 minutes and creates around 6,000,000 edges. This increase in computation time is not feasible for processing the entire dataset. Therefore, we have chosen to analyze the dataset using samples in the 1000-2000 node range.

We have utilized stratified sampling and the Chi-Squared test to ensure that any sample we use is representative of the entire dataset.

What is left?

We have yet to implement the code for answering the research questions fully, but we have a start which we will continue work on alongside the final report and presentation.