Our workflow utilizes three scripts written in R and python to complete the analysis. The first script titled 'preprocess.R' is to clean the data by removing select columns, reformatting other columns, and creating new features. The 'clean.fun' function does most of the data cleaning for the analysis. The 'FeatureCreation.fun' is to create new features based on a factor analysis of mixed data. We create 5 new dimensions to help explain variability among the observed values in the dataset. These 5 dimensions are then used to create rolling averages for variables relating to the horse, trainer, jockey, and other variables. The purpose of this analysis is to balance the influence of both continuous and categorical variables in the analysis. By computing rolling averages over the horse, sire, dam, jockey, and trainer we hope to improve model performance in scenarios when one of the variables has no prior information.

The second script titled 'fnn_modelling.R' is an R script to import the parquet files, apply any additional preprocessing and then pass the data into the python script for model training. This script also analyzes the output of the python script.

The third and final script is titled 'SiameseFNN.py'. This file contains python code for create a variable length siamese feed forward neural network as well as testing the performance of this network on new data. The reason for using a variable length siamese FNN is to be flexible as to the number of horses in a race. This is done by fitting the same network base to each horse and then concatenating the outputs of these networks into a centralized FNN to amalgamate the outputs and provide a final prediction. The siamese FNNs are flexible to the number of horses in a race as instead of each FNN being treated as unique and updating weights independently, the input FNNs share weights and weight updates. This allows for each horse to be analyzed individually by the same network and to be assigned a unique value of the probability of the horse winning. From the output of the siamese networks we obtain pairwise distances between the embeddings of individual horses. These distances are then used to measure the dissimilarity or similarity between horses. The L1 distance, also known as the Manhattan distance, is used in this case. The embeddings represent the essential characteristics of the horse that are relevant for predicting the race outcome. These embeddings are then used by the central FNN to identify the relative strength of horses within a race and identify which may be most likely to win.