Final Project

Matt Imberman and Luke Liao

May 22, 2021

Introduction

Horse racing is a huge business in Hong Kong, the gambling aspect of the sport has attracted many people thus creating one of the biggest money markets. The goal of this project is to use the data from horse racing in Hong Kong, which we download from **Kaggle**, to predict what factors are most crucial to make a winning horse, or a horse with the best finish time. There are many elements that need to be considered when mapping the prediction, such as difference in a horse's age, country of origin, breed, etc. All play a critical part when evaluating. We focused on age, horse type, declared weight and draw as predictors for finish time.

Explanation of subgroups, n, number of variables and limitation of the data

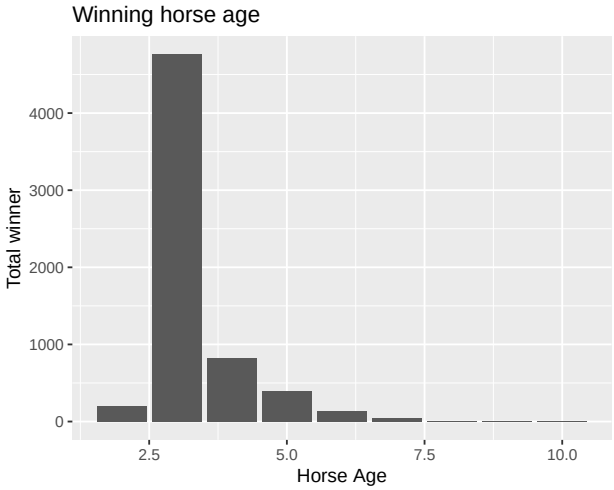
Within our data set, there are certain subgroups that we feel are necessary to clarify what they mean. Other than the obvious age, country, "horse_type" is sex of the horse (e.g. 'Gelding', 'Mare', 'Horse', 'Rig', 'Colt', 'Filly'); "declared_weight" is the weight of the horse and jockey combined, in lbs; "draw" is the post position of the horse in the race; "finish_time" is the horse's completed race time in seconds. The data has a total of 79448 rows and 35 columns, with total of 5799631 variables.

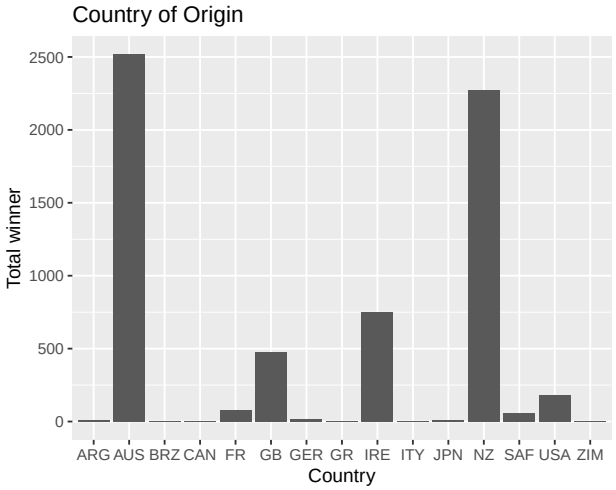
HYPOTHESIS

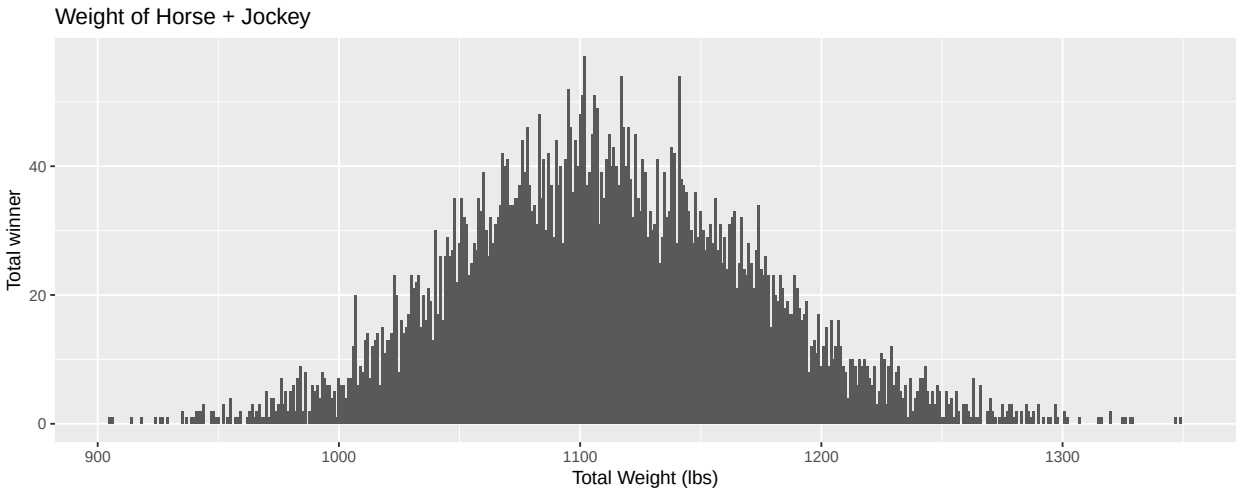
A younger, lighter horse will have better finish times than other horses.

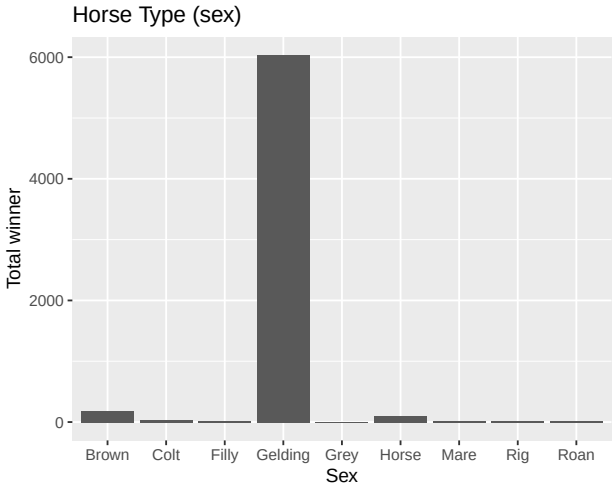
Exploratory Summary and Graphs

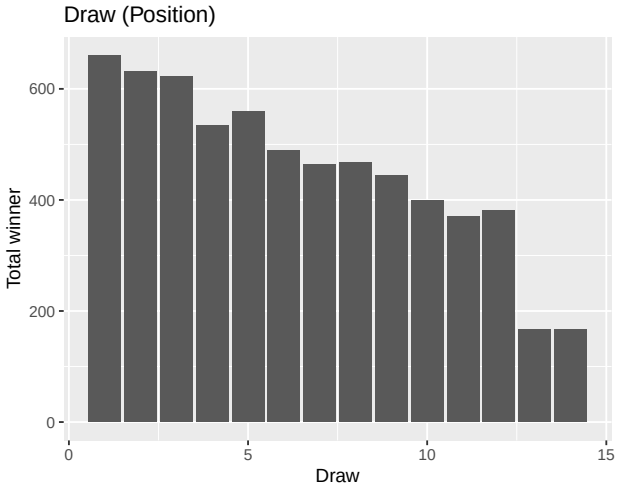
We start exploring our data sets by comparing differences in factors, especially trying to find out what is the most common element in a winning horse.









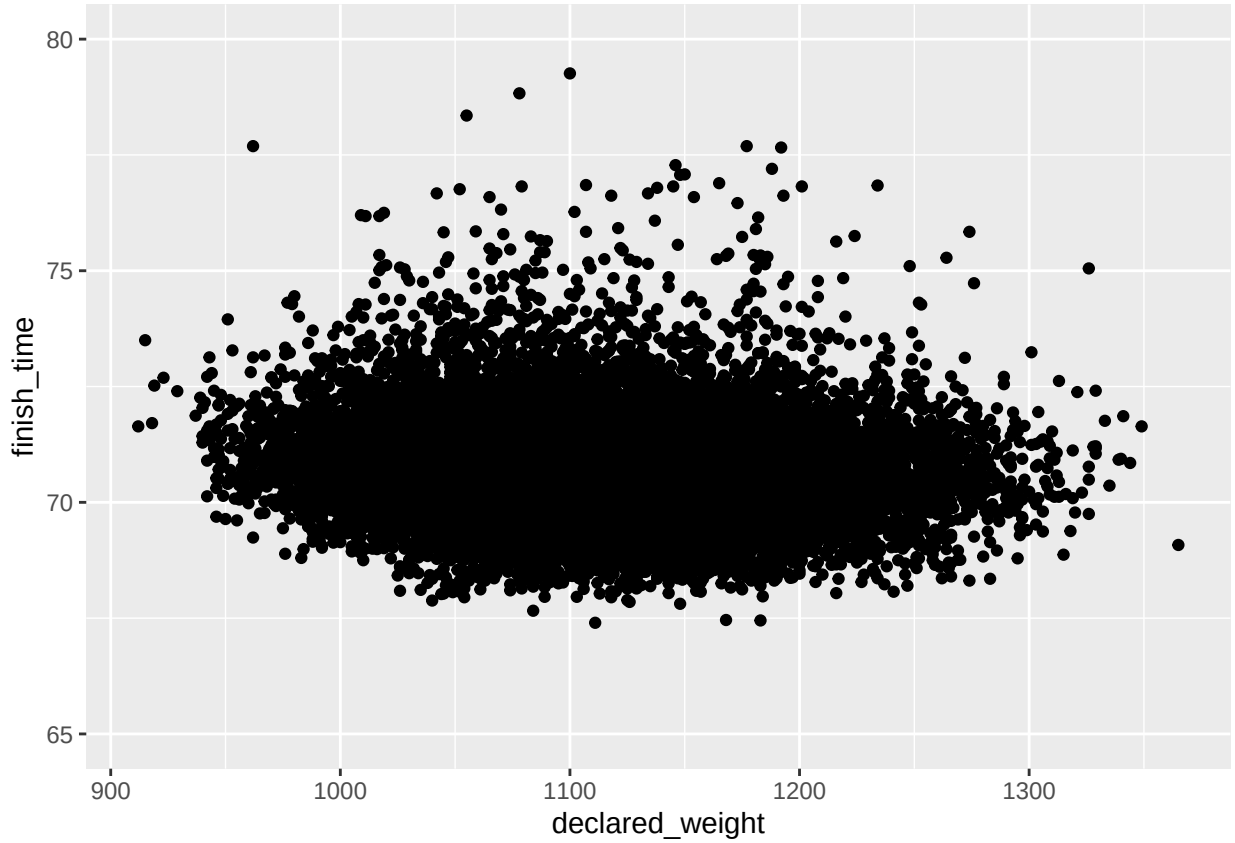


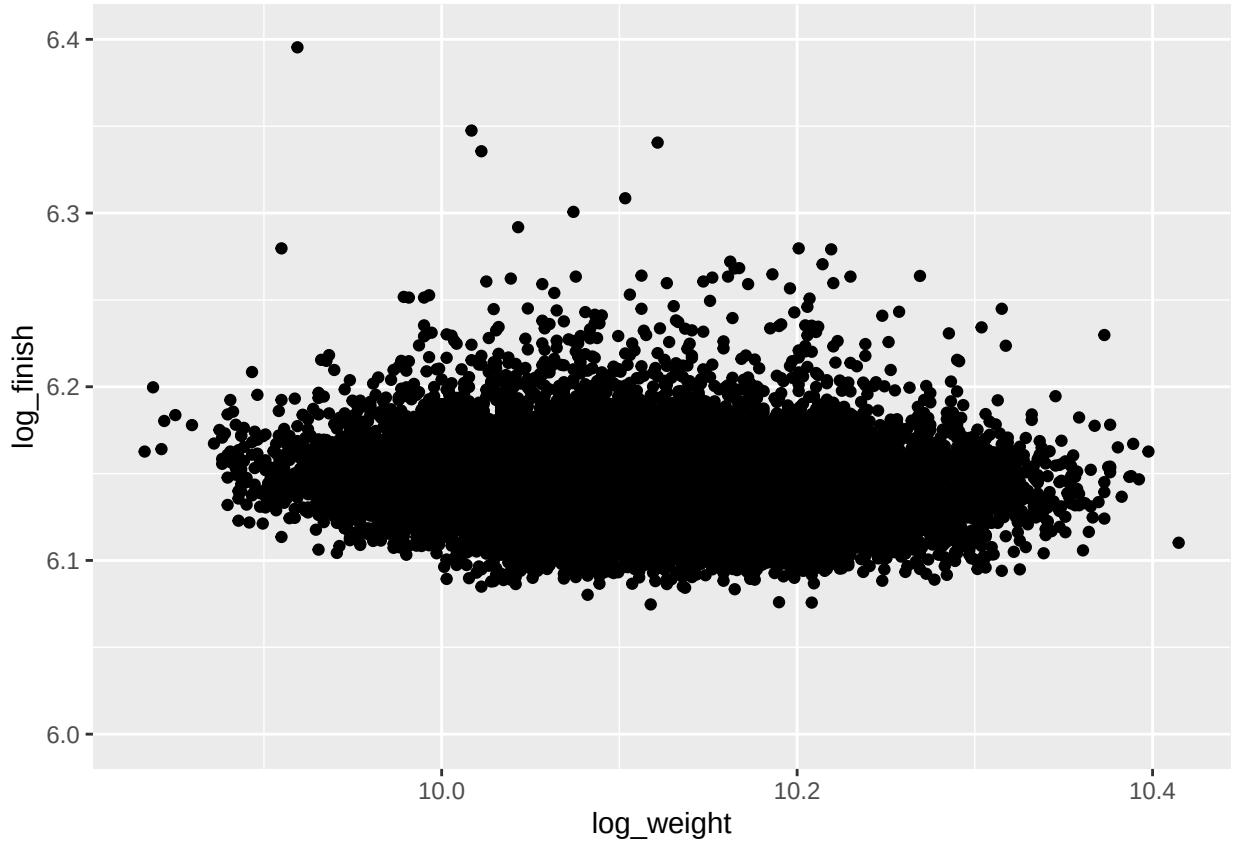
As the above graphs show, most of the winning horses are at age 3; majority of them were either from Australia or New Zealand; Gelding seems to be the most winning horse, which is just a another name for

a castrated male horse; In terms of draw (position) in a race, lane number between 1-3 produced the most winners and finally there were no guaranteed winner in terms of combined weight of horse and jockey. These findings are rather interesting since it somewhat verified our hypothesis, we predicted the younger the horse the faster it can run however this is only partially true since age 3 seems to be the golden age. In terms of weight, we thought the lighter the horse the better however this was not true as the distribution shows the middle pack around 1100lbs produced the best time.

Linear models and multiple regression to predict finish time

Not all the races in our data are the same distance. We decided to focus on races that are 1200 meters in distance. This was the distance that had the most races. We start off our analysis by plotting declared weight vs finish times and then another plot of the log of these values





After looking at the plotted data, we do not see too much linearity. We ran multiple linear/regression models below to see if there was any predictive qualities in our data

Model Outputs

```
##
## Call:
## lm(formula = finish_time ~ declared_weight, data = data1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.321 -0.621 -0.049  0.529 33.039
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     72.601773   0.117743  616.61   <2e-16 ***
## declared_weight -0.001693   0.000106  -15.97   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.027 on 24695 degrees of freedom
## Multiple R-squared:  0.01022,    Adjusted R-squared:  0.01018
## F-statistic: 254.9 on 1 and 24695 DF,  p-value: < 2.2e-16

##
```

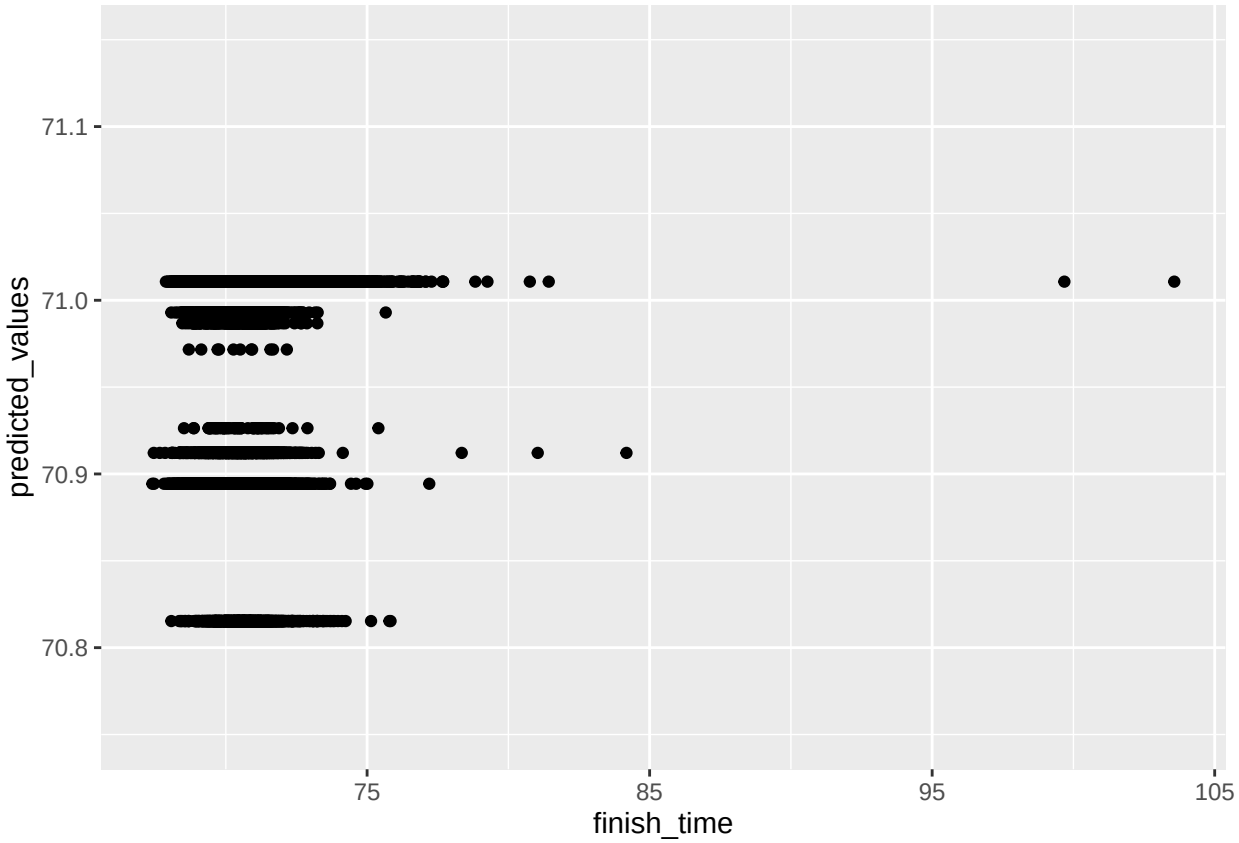
```
## Call:
## lm(formula = log_finish ~ log_weight, data = data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.06917 -0.01256 -0.00084  0.01091  0.55439
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.41981    0.01658  387.20   <2e-16 ***
## log_weight  -0.02727    0.00164  -16.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02056 on 24695 degrees of freedom
## Multiple R-squared:  0.01108,    Adjusted R-squared:  0.01104
## F-statistic: 276.7 on 1 and 24695 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = finish_time ~ declared_weight, data = data1, subset = horse_age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5571 -0.0214 -0.0214 -0.0214  2.9018
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     71.5091631  0.0408887 1748.87   <2e-16 ***
## declared_weight -0.0004896  0.0000383  -12.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3574 on 24695 degrees of freedom
## Multiple R-squared:  0.006573,   Adjusted R-squared:  0.006533
## F-statistic: 163.4 on 1 and 24695 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = finish_time ~ declared_weight, data = data1, subset = horse_age,
##     weights = draw)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6511 -0.0687 -0.0687 -0.0687 10.4679
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      7.193e+01  4.804e-02 1497.40   <2e-16 ***
## declared_weight -8.882e-04  4.424e-05  -20.07   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.167 on 24695 degrees of freedom
```

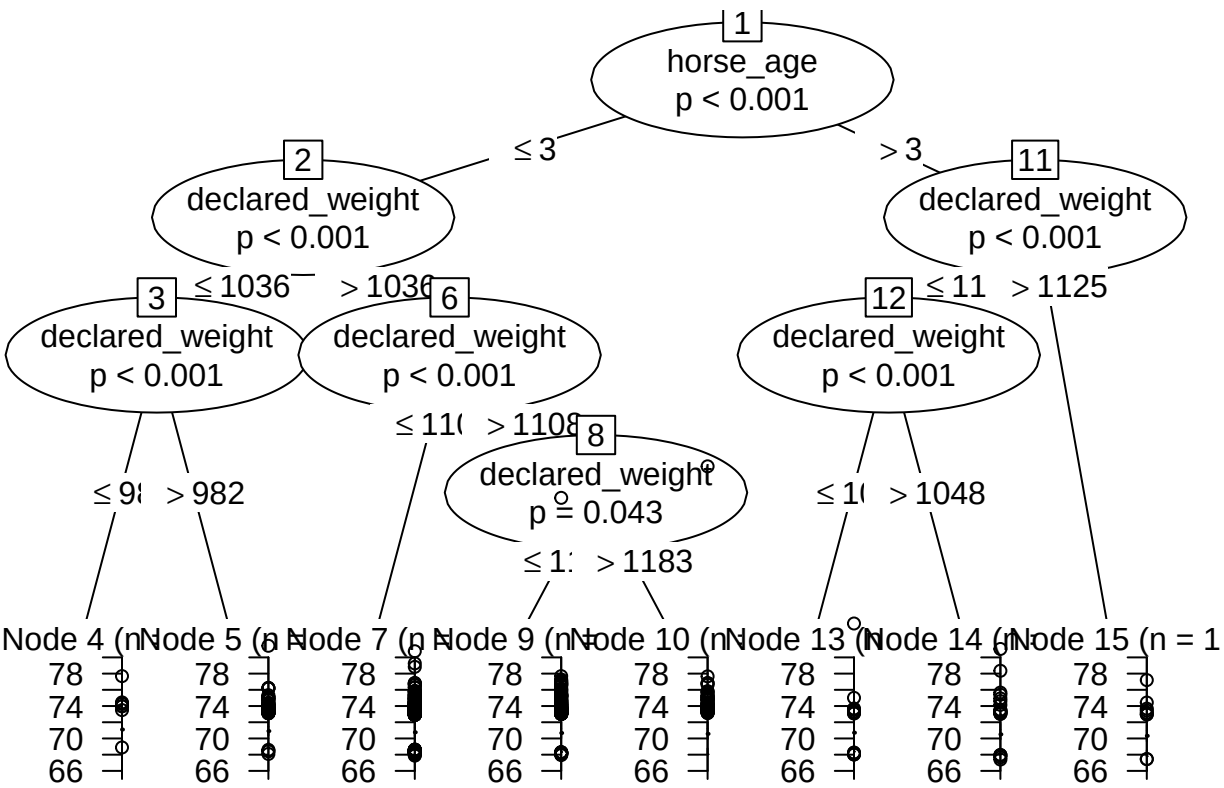
```
## Multiple R-squared:  0.01606,    Adjusted R-squared:  0.01602
## F-statistic:   403 on 1 and 24695 DF,  p-value: < 2.2e-16
```

Though each model has a small r squared only averaging **0.01108**, we decided to use predictive data from mult_model2, which uses the most variables. Below, we see a plot of all predicted values. It appears that the predictions are clustering around certain finish times, which will not have much predictive value



Conditional inference tree analysis

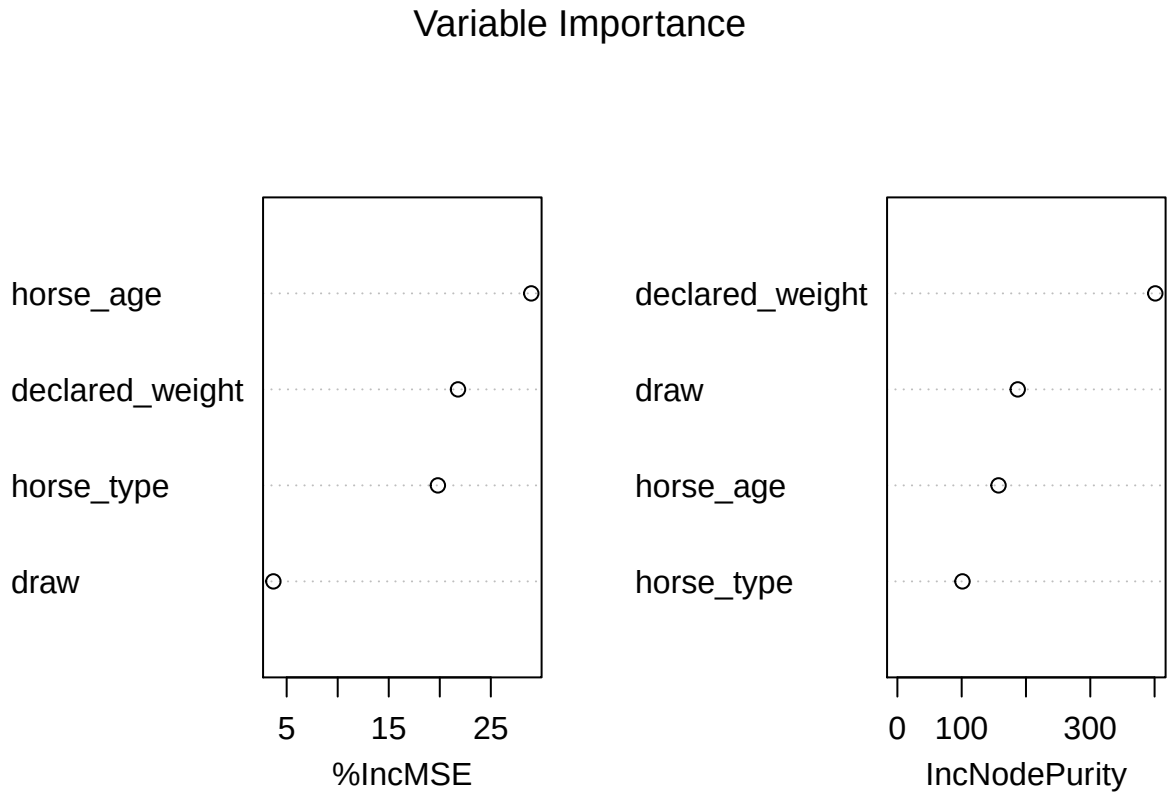
We decided to use a conditional inference tree analysis to see if we could predict the finish time of each horse. We selected declared weight and horse age as predictor values, as these are very important for the speed of the horse. We then selected the distance with the most races. Races at **1200** meters were by far the most prevalent.



After running the conditional tree we see that horse age is the most important predictor of finish time. As you travel down through the nodes, we see small differences in finish times amongst various weight classes. We can see that some of the quickest horses (lowest finish time) are horses that are younger than or equal to3 years old and have a declared weight of less than 1183 pounds.

Random Forest

To expand on our tree analysis, we decided a random forest analysis would be most appropriate. We separated our data as 75% training data and 25% testing data. We used a random forest model with 500 trees.



Our random forest analysis clearly shows what variables are most important when predicting finish time. On the first graph, we can see that removing horse age from our model would significantly increase our mean squared error. This shows a high importance of horse age in our model. when we look at the node purity graph, we can see that declared weight is one of the most important variables in our model. This number might be skewed by the large amount of different weights, but it is nonetheless, important. The node impurity graph also shows a high importance for horse age.

CONCLUSION

We began our analysis of horse racing data in an attempt to find out what factors are most critical in making a winning horse. We discovered that although age and weight of horse + jockey are important (proving our hypothesis was correct), the data we possess did not show any valuable predictions. The data did not show a linear path and our linear models had very low r squared. In conclusion, we think the horse racing data set did prove our hypothesis and was heading to the right direct. A horse's age and weight are important factors when trying to predict the finish time, however, our, models came up short in predictive value. In the end, we learned that THE HOUSE ALWAYS WINS. Making acurate predictions for finishing times of horses races is not feasible. Many have tried in the past and have come up short.