Group Project Part 1 Analysis (C and D)

c.

Here are the results for the linear, kNN, and decision tree regression algorithms.

Linear Regression:

```
[1] "cor= 0.959206463271685"
[1] "mse= 36101.9618087527"
[1] "rmse= 190.005162584475"
```

kNN Regression:

```
[1] "cor= 0.957191866567774"
[1] "mse= 38144.5006673951"
[1] "rmse= 195.306171606007"
```

Decision Tree Regression (without RF/bagging):

```
[1] "cor:  0.859246727417666"
[1] "mse:  116650.137250333"
[1] "rmse:  341.54082808697"
```

Decision Tree Regression (with bagging):

```
[1] "cor:  0.957201775647217"
[1] "mse:  37384.8144143277"
[1] "rmse:  193.351530674902"
```

As shown above, the linear regression algorithm outperformed all other algorithms in both correlation and calculated rmse values. The decision tree algorithm (with bagging) performed second best, narrowly beating out kNN regression in correlation and rmse. kNN performed third best, and the regular decision tree without random forest prediction/bagging performed the worst by far.

d.

We can see that the linear regression and kNN regression algorithms performed very similarly, with a difference of correlation measuring to be 0.2%. kNN often outshines linear regressions when the data is more messy or complex, or if there is a prominent non-linear trend to the data. Because these performed so similarly, it means that our data is likely to be non-complex and somewhere between linear and non-linear.

The decision tree regression performed much worse than the other two algorithms on its own, resulting in a correlation of 0.8592 after pruning. By predicting on a random forest, we we're able to greatly increase the correlation to 0.9561, and bagging brought our correlation up to 0.9572, roughly the same as our kNN and linear regressions.

The regular decision tree regression likely performed so poorly due to overfitting. This is further proven through the high improvement of correlation after implementing random forest and bagging, which are designed to combat overfitting by running the algorithms on multiple different generated trees.