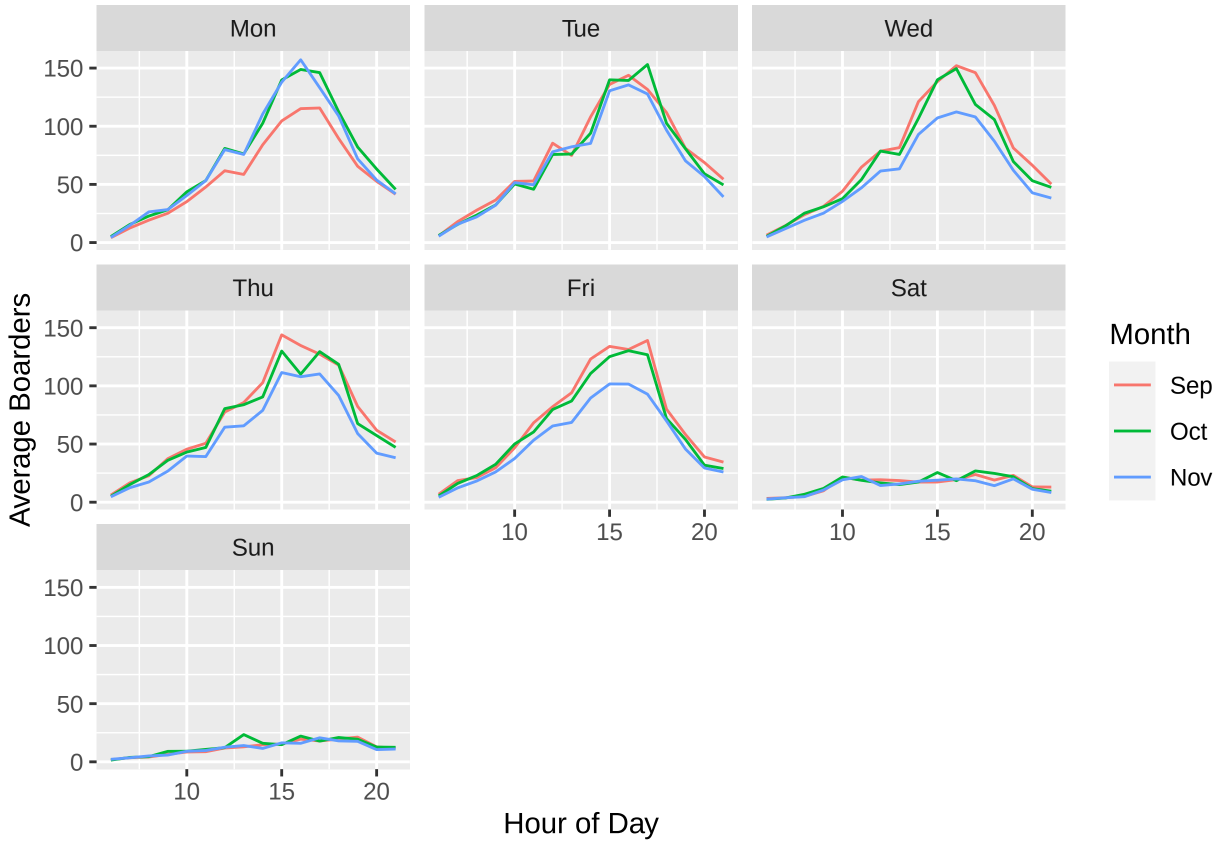
Homework2

Problem 1

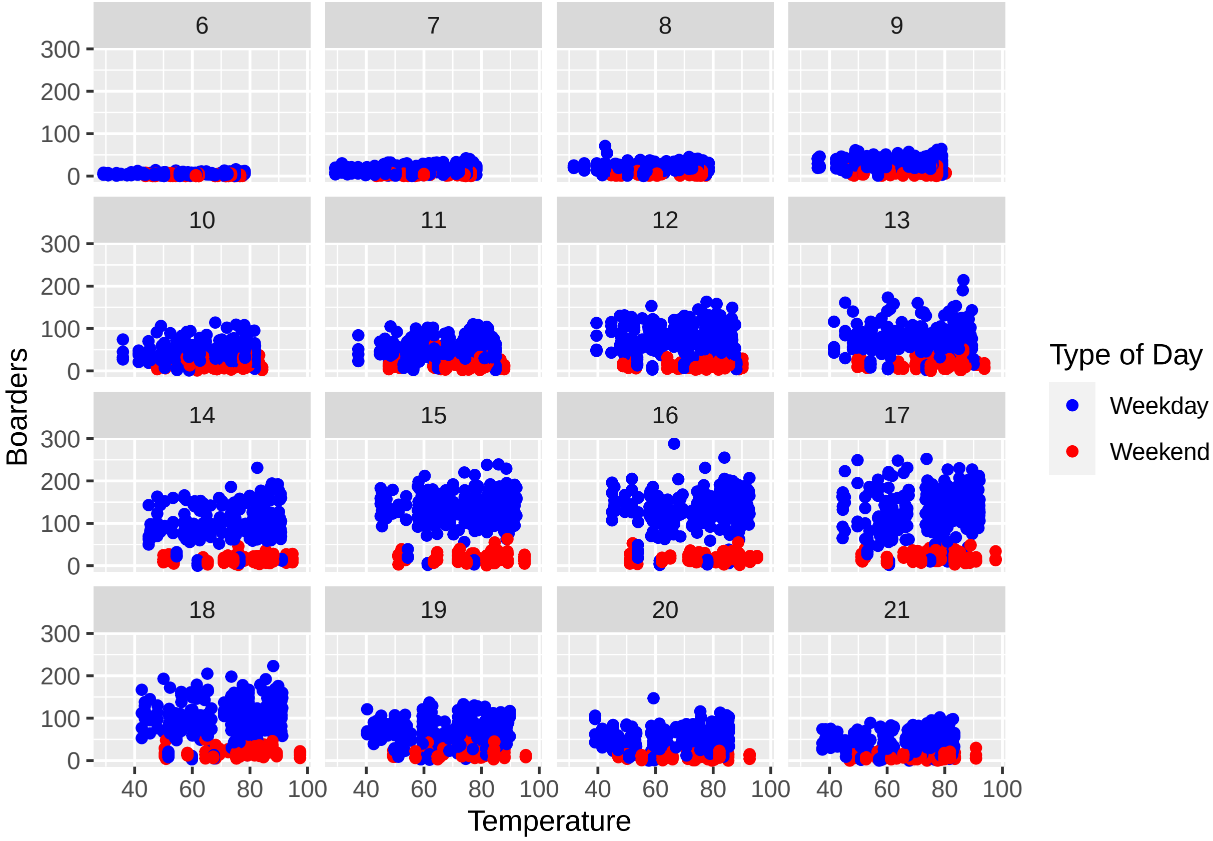
Part A

```
## ‘summarise()‘ has grouped output by ’hour_of_day’, ’day_of_week’. You can override using the ‘.groups
```



What we have here is a series of graphs that show the average boarding size with respect to the time of day. Furthermore, each graph, starting from top left and going rightwards, represents a day of the week beginning with Monday and ending with Sunday. The colored lines represents the average boarding size with respect to the time of day in the given month. As we can see from the graphs, the hour of peak boarding do not seem to change day to day during the weekday as it always happens between the 15th and 17.5th hour of the day. This is not surprising given the fact that most classes have ended by this point and most people have finished their working shift. The best explanation as to why boarding on Mondays in September and the average boarding in Wednesday, Thursday and Friday in November look lower is due to holidays. The first Monday in September is always labor day, meaning that students do not have classes, and some workers have the day off, whereas the days mentioned in November is when the Thanksgiving holidays occur.

Part B



In these series of graphs, we are plotting the average boarding as a function of temperature in 15-minute intervals and showing whether the datapoints represent the weekday or weekend. When we hold hour of day and weekend status constant, there does not seem that temperature has a noticeable effect on the number of UT students riding the bus. In some cases , we can see an uptick in the number of boarders at around 90 degrees Fahrenheit but this could simply be due to unknown factors.

Problem 2

There are two ways that I used to build the best linear model. The first is by looking running a regression of price on all the variables to determine which variables to keep based off their statistical significance. The second way is using the step function and allowing for interactions between the variables. I then estimate the out-of-sample RMSE over many different random train/test splits.

```
## [1] 76754.73
```

```
## [1] 66081.02
```

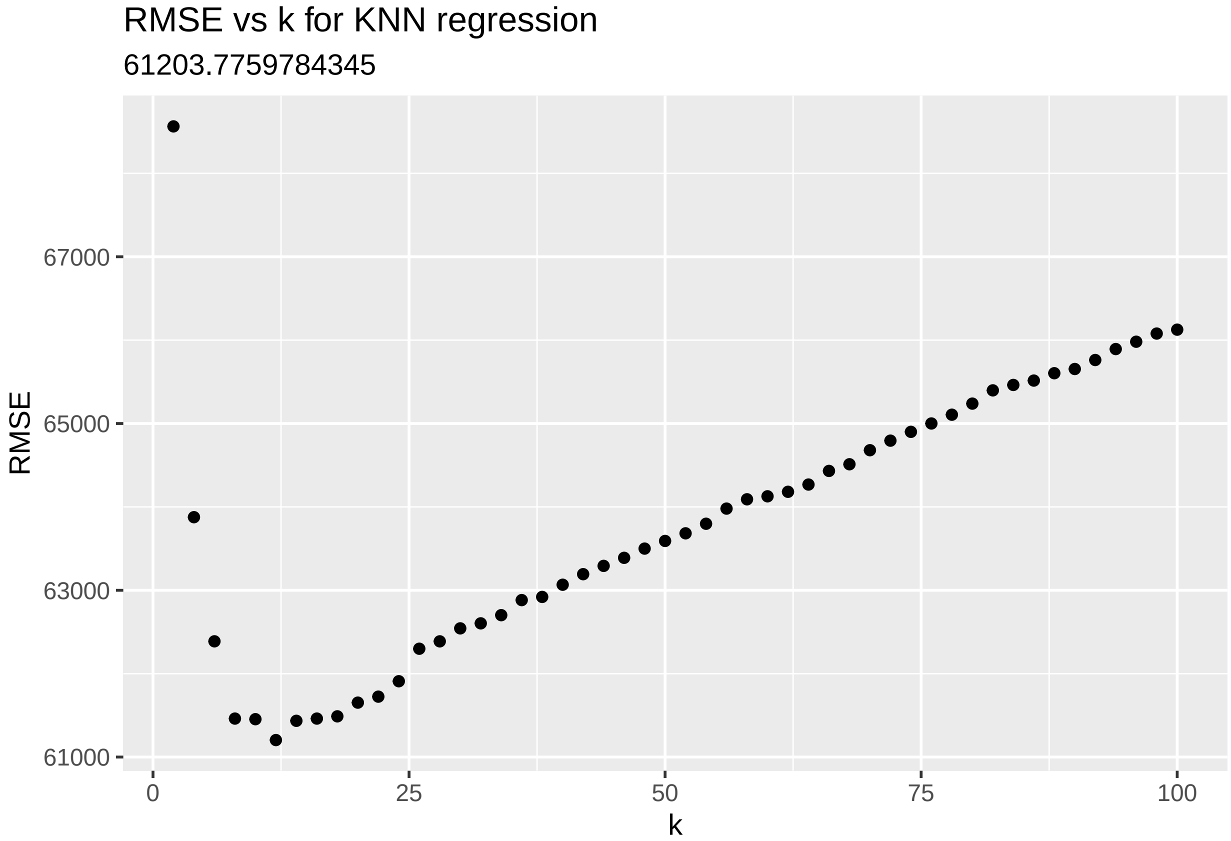
```
## [1] 71453.17
```

```
## [1] 58360.29
```

```
## [1] 64072.04
## [1] 58261.98
```

The RMSE that we are concerned are RMSE 5 and RMSE 6 . The RMSE 5 represents the linear regression based off the result of the step function, whereas the RMSE 6 represents the model that I visually inspected and selected the variables. As we can see the RMSE of my handmade visual model is signficantly better than what the step function found, which is a bit surprising.

KNN



The model that seems to do better at achieving lower out-of-sample mean-squared error is the linear model that I made by hand. This means for the local taxing authority that if they want to predict the market values for properties in order to tax them, they should consider all the variables in the dataset with the exception of sewer, fuel, and heating type, the number of fireplaces that a house has and the percent of college students situated near the property.
## Problem 3