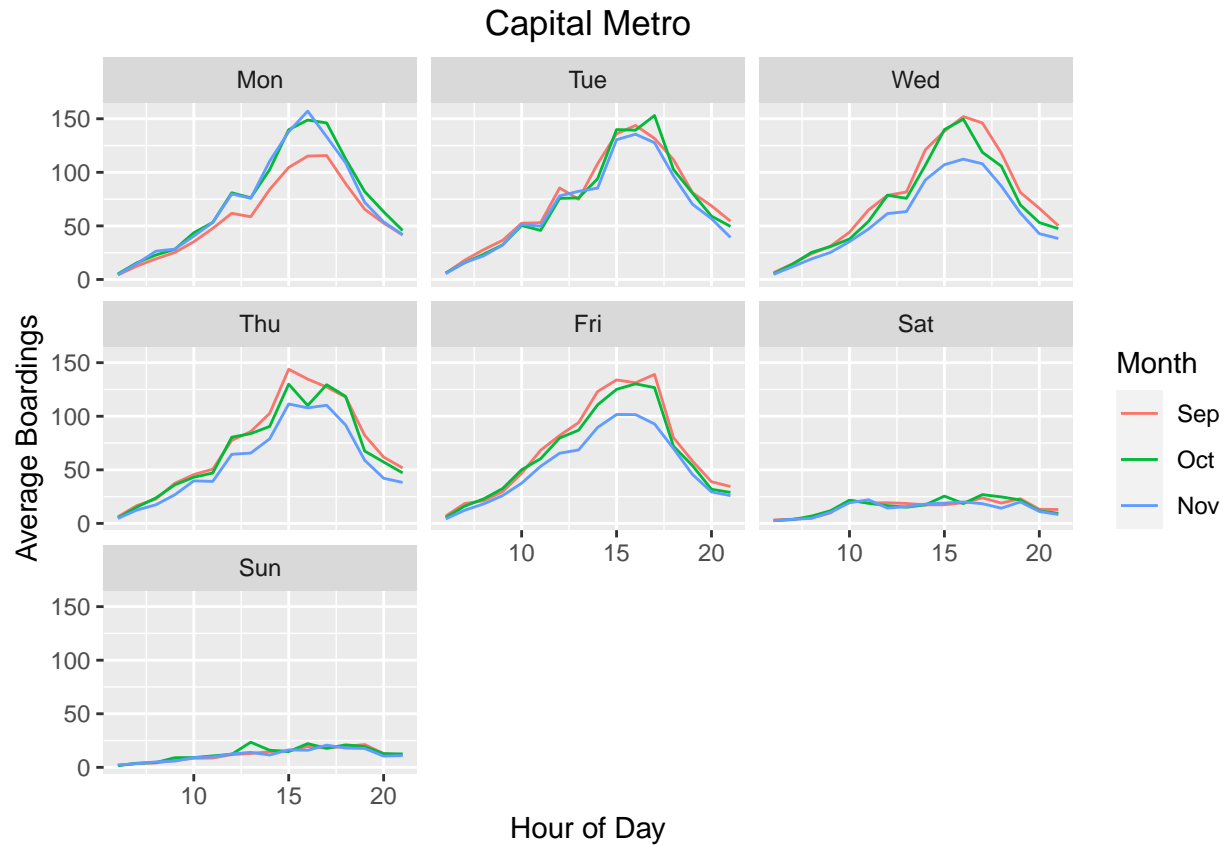


## Exercise 2

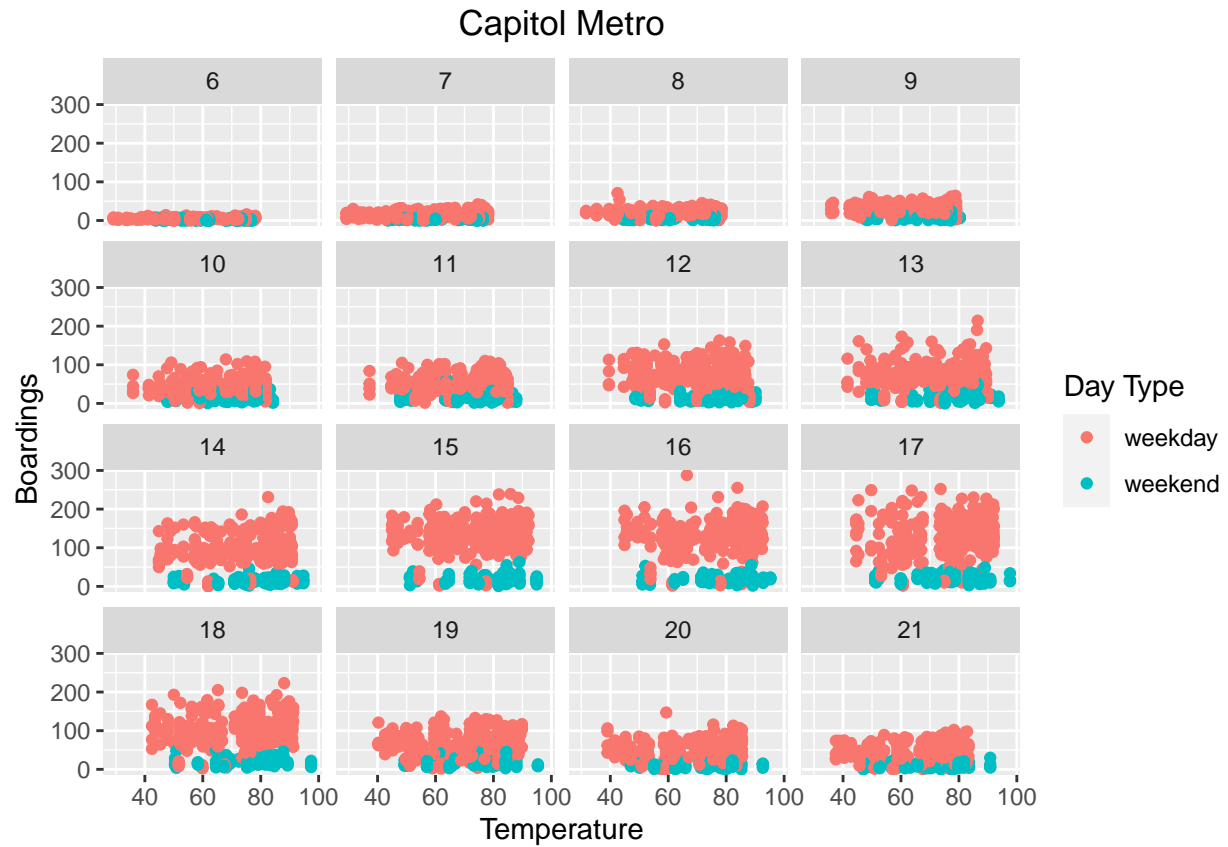
### Question 1



Excluding the weekends, peak boardings are broadly similar, which is during evening rush hour.

The average boardings on Monday are lower in September because of Labor Day. Since it's a national holiday, the university is closed and UT students would stay home and would not need to use the bus.

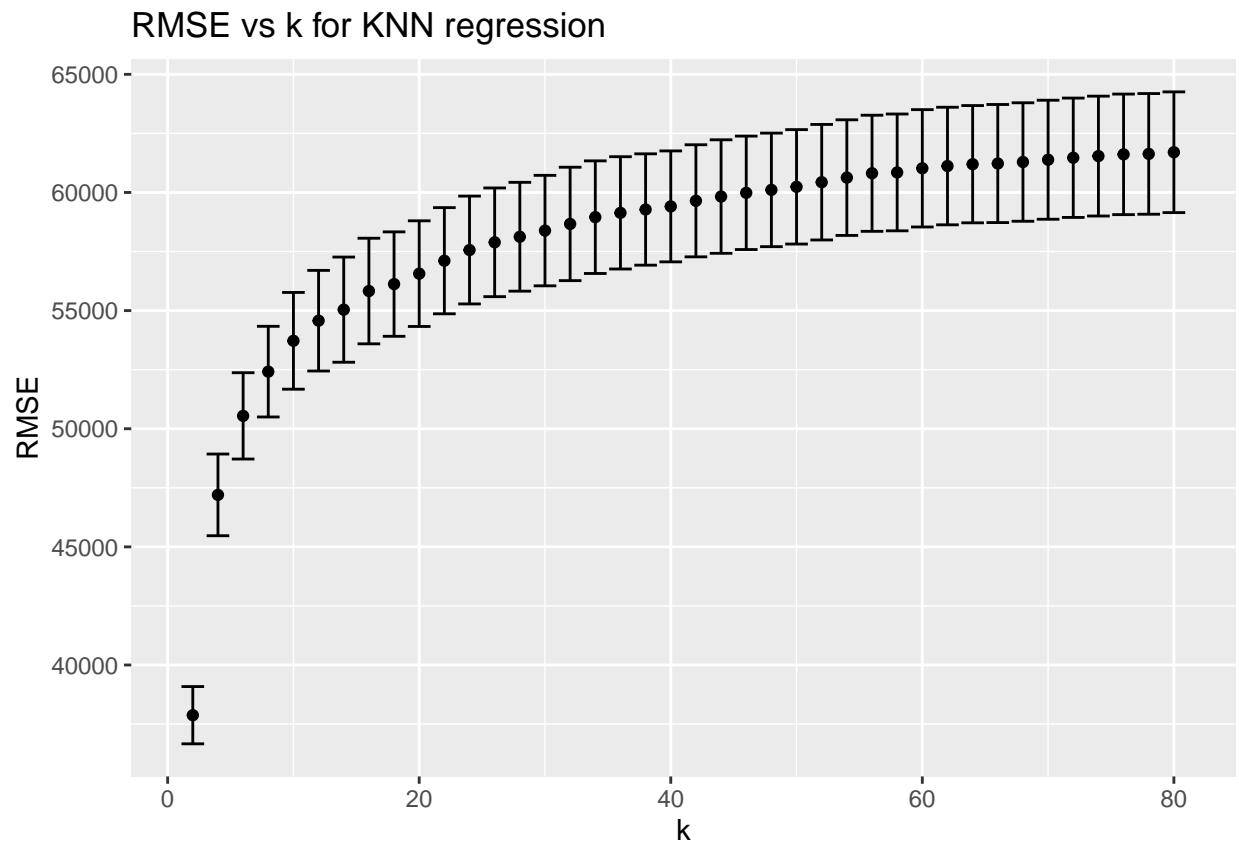
Additionally, average boarding is lower on Wed/Thurs/Fri for November because Thanksgiving lands on a Thursday in November, and UT students would most likely not travel during days of that particular week, but instead spending time with family/friends.



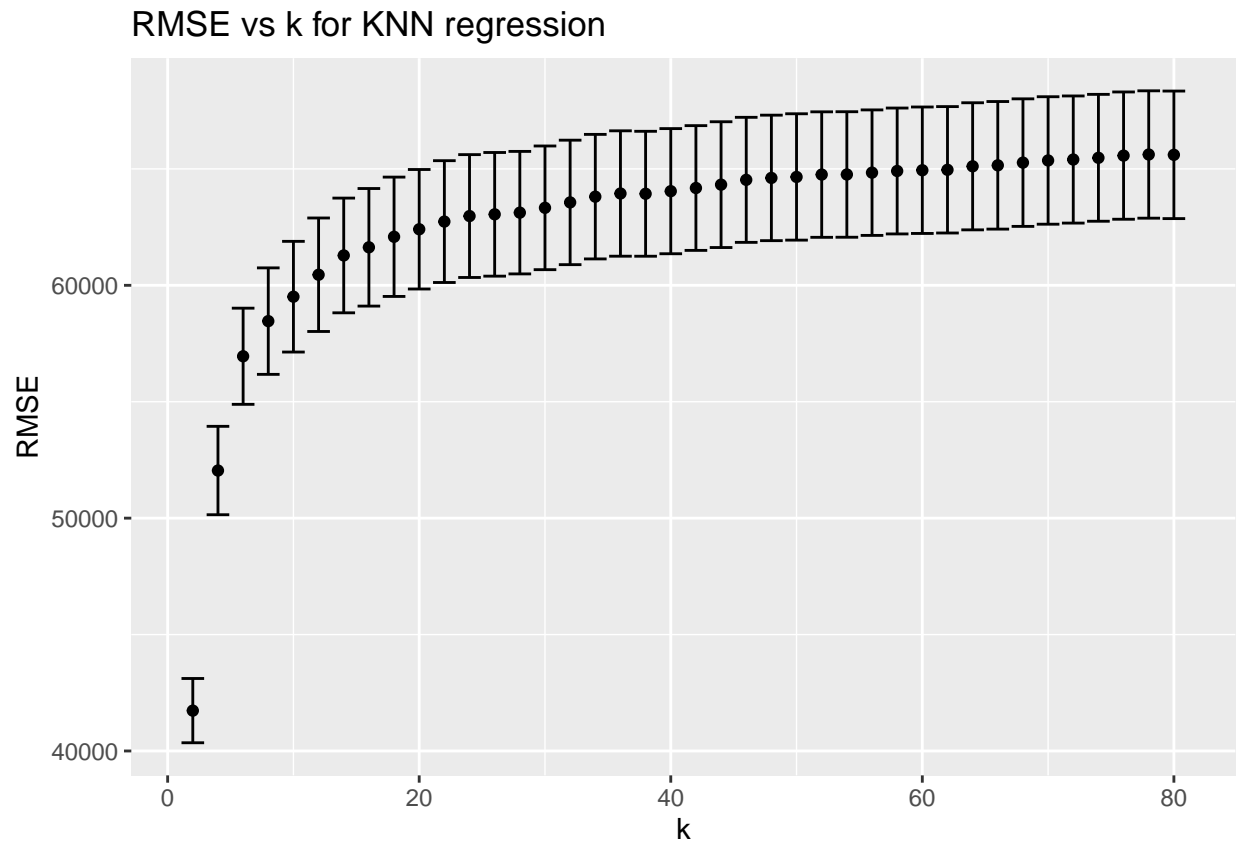
When holding both hour of day and weekend status constant, it does not appear that temperature has a noticeable effect on the number of UT students riding the bus. This is determined by the observable changes in ridership throughout the day. Within each hour of the day, the quantity of boardings remain similar as the temperature changes.

## Question 2

### Best Linear Model

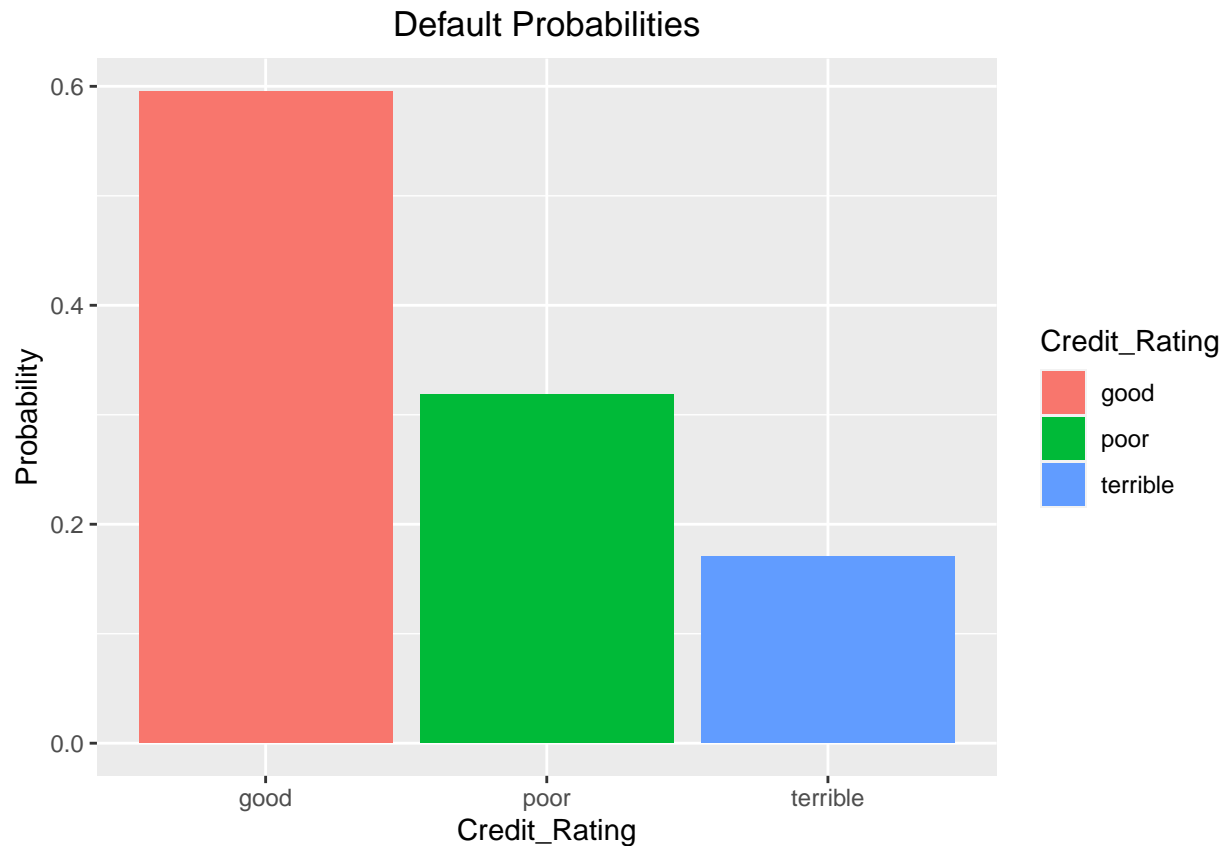


## Medium Linear Model



Relative to the Medium Linear Model, the Best Linear Model has both a lower AIC and RMSE, while also a higher Adjusted R-squared. For these reasons, the Best Linear Model is the optimal choice since it seems to do better at achieving lower out-of-sample mean-squared error.

### Question 3



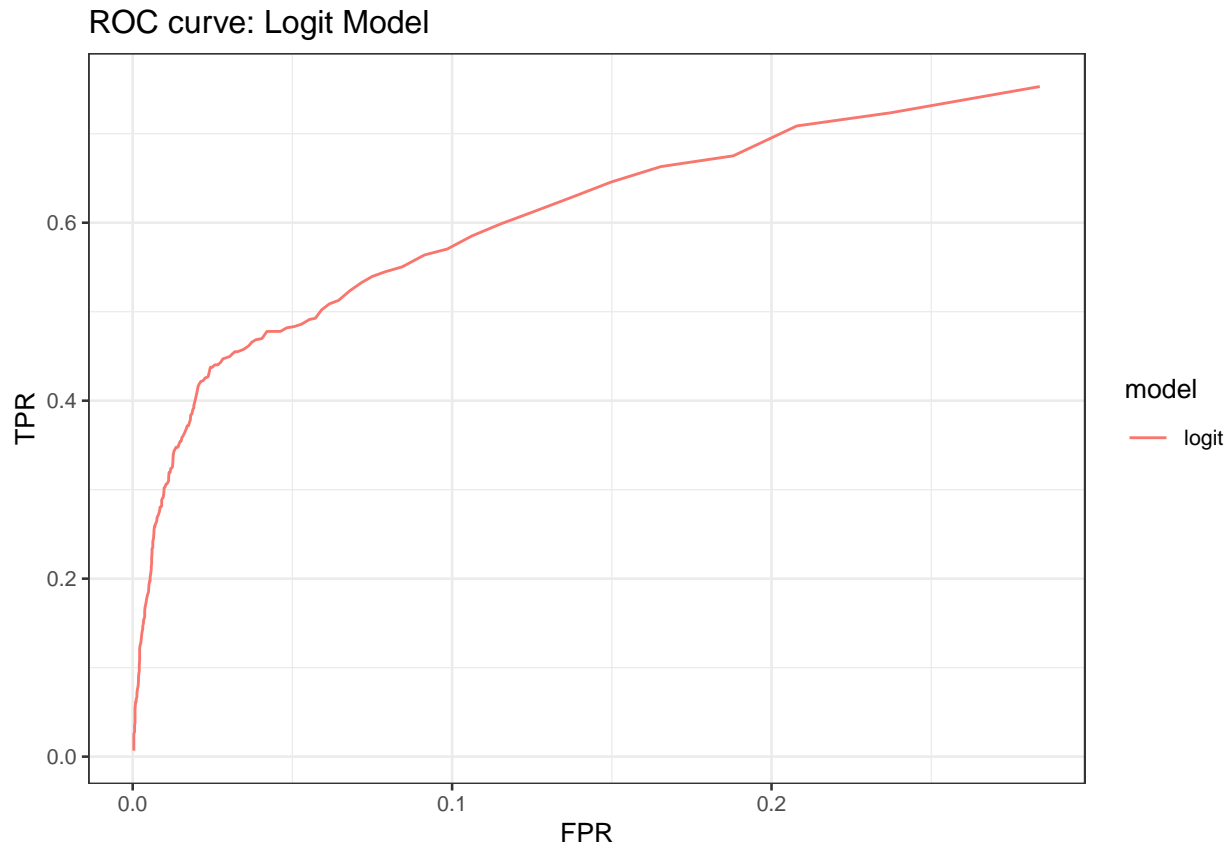
```
##      (Intercept)      duration      amount      installment
##      -7.075258e-01  2.525834e-02  9.596288e-05  2.216019e-01
##           age      historypoor  historyterrible  purposeedu
##      -2.018401e-02  -1.107586e+00  -1.884675e+00  7.247898e-01
## purposegoods/repair  purposenewcar  purposeusedcar  foreigngerman
##      1.049037e-01   8.544560e-01   -7.959260e-01  -1.264676e+00
```

It should be noted that the default probabilities are much lower for individuals with either a poor or terrible credit history relative to individuals with a good history. Although, when observing the glm, individuals with either a poor or terrible credit history have a negative effect in regards to defaulting on a loan. I believe that these conflicting observations are due to the bank having a more strict loan process for individuals with either a poor or terrible credit history. Individuals that are in either category are more likely to repay their loans back on time. In contrast, individuals with a good credit history are given a more lenient loan process and as a result, have a higher probability in defaulting on their loans.

If the purpose of the model is to screen prospective borrowers to classify them into “high” versus “low” probability of default, then no, I do not believe this data set accomplished its goal. The conclusion from this data set should not be to loan less to individuals to a good credit history. Instead, the end result should be for the bank to apply the same strict loan standards that are applied to others that may have either a poor or terrible credit history. This would give a better source of data and better inform the bank on the probability of default.

## Question 4

### Model Validation 1



### Model Validation 2

```
# 20 folds
knn_k20 = caret::knn3(children ~ hotel + lead_time+ reserved_room_type + assigned_room_type +
                      booking_changes + adults + required_car_parking_spaces + booking_changes +
                      average_daily_rate + is_repeated_guest + arrival_date, data = val, k = 20)

val_pred = val %>%
  mutate(children_pred_k20 = predict(knn_k20, val))

val_pred1 = val_pred %>%
  mutate(ifelse(children_pred_k20 > 0.5, 1, 0))

expected_children = val_pred1%>%
  summarize(num = sum(ifelse(children_pred_k20 > 0.5, 1, 0)))
expected_children

## # A tibble: 1 x 1
```

```
##      num
##    <dbl>
## 1  4975
```

```
actual_children = val_pred1%>%
  summarize(num = sum(children ==1))
actual_children
```

```
## # A tibble: 1 x 1
##      num
##    <int>
## 1    402
```

My model did fairly poorly relative to the number of bookings made with children. Across all 20 folds, my model predicted that 4975 children would be present, while 402 actual children were present in the data.