HW5 Zihao Li

Question 1: The impact on traffic [50 points]

a) Please use traffic as dependent variables, all other variables except for *storeid, training, date,* and *sales* as independent variables, and build NNM with decay=0.5 and size= 1,2,3,4,5. Fill the table below with the MAE for both training and test data. [5 points]

size	**Training data**	**Test data**
1	178.043	191.430
2	160.510	175.346
3	149.794	163.229
4	160.667	175.842
5	160.648	175.807

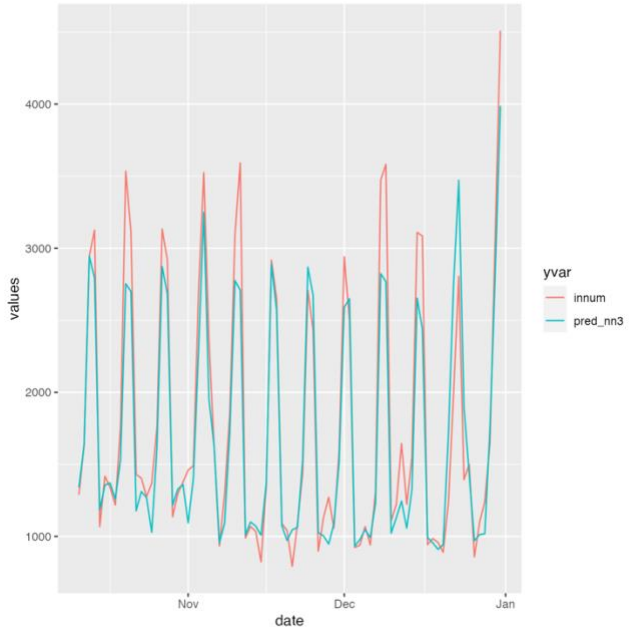
b) What is the optimal size? Why? [5 points]
I would say size 3 is the optimal size because it has lowest MAE for both training and test data.

c) Build the model with decay=0.5 and size equals to the optimal size in b). Show the MAE and RMSE for All data. Calculate MAE%. Is this model acceptable regarding MAE%? Why? [10 points]

MEAN	RMSE	MAE	MAE%
1,749.786	215.989	155.972	8.914%

I would say it is acceptable because there are lots of thing might affect the result especially in and around the mall so this model should be acceptable.

d) Transform year, month, day, weekday into factor variable in out sample data, and use the model in c) to predict the traffic in out sample data. Show daily trend of real traffic and predicted traffic in out sample data. [combined both line in one plot] [Do not filter data] [5 points]



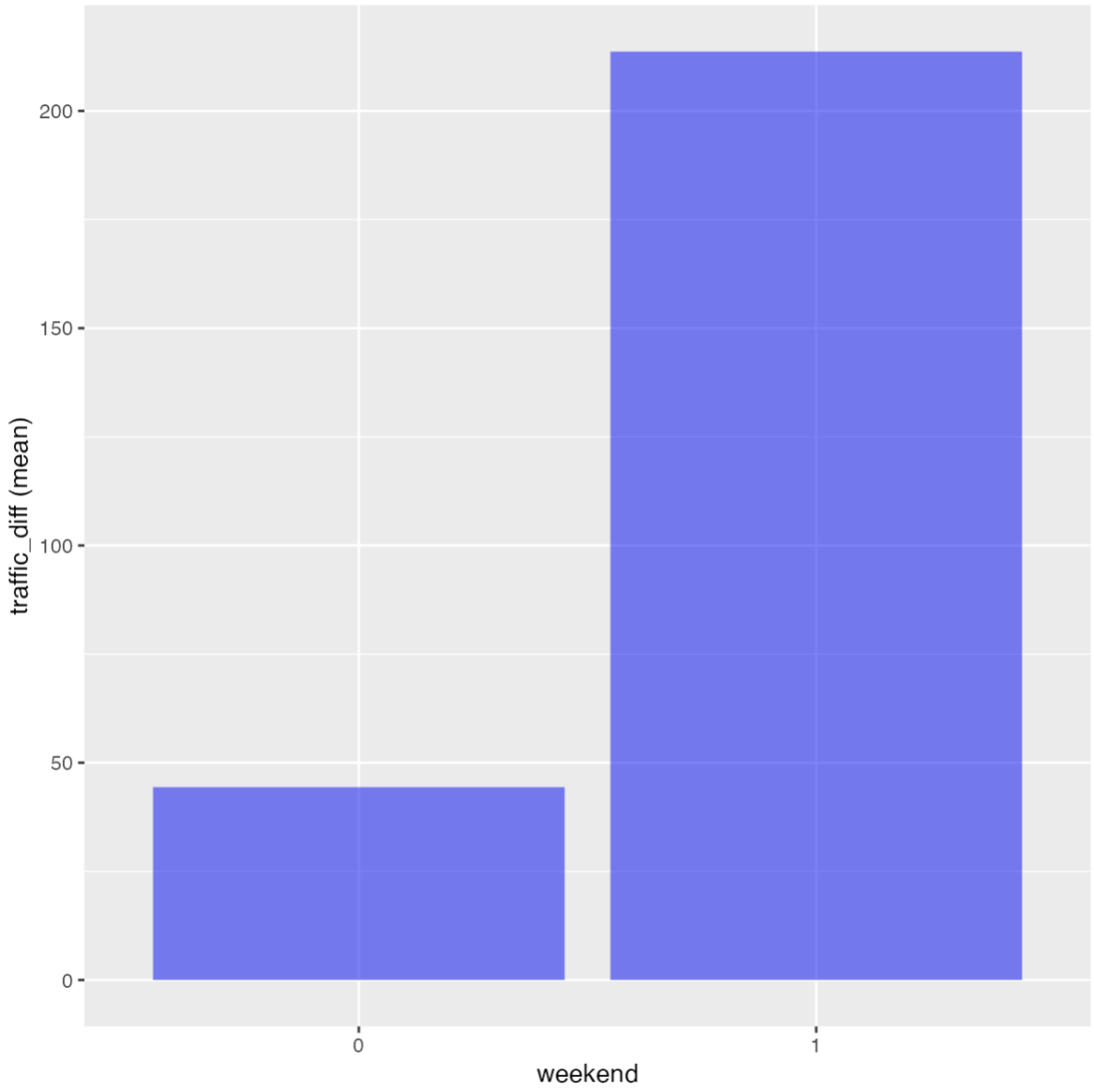
e) Create a new variable, *traffic_diff*, indicating the difference between real traffic and predicted traffic in the out sample data (traffic_diff=real traffic-predicted traffic). Show the mean of *traffic_diff* during campaign period. Please explain the meaning of this number. [10 points]

It means the real traffic is higher than our prediction in average during the campaign period.

f) Please show the 95% prediction interval for the effect of GAP campaign on the traffic of Guess in Chicago. [5 points]

lower	upper
-338.065	525.897

g) Show the average *traffic_diff* between weekend and weekday. What do you observe? [5 points]



During the weekend, there are way more traffic compare to weekday.

h) Conduct a hypothesis test to confirm the observation in g). Show hypothesis test results and explain the results. [Hint: transform weekend into factor variable in this step] [5 points]

```
Pairwise mean comparisons (t-test)
Data      : GUESS_OUTSAMPLE
Variables : weekend, traffic_diff
Samples   : independent
Confidence: 0.95
Adjustment: None

 weekend    mean  n n_missing      sd     se      me
       0  44.376 58         0 182.802 24.003  48.065
       1 213.648 24         0 422.031 86.147 178.208

 Null hyp. Alt. hyp. diff     p.value
 0 = 1     0 < 1     -169.272 0.035   *

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is less than 0.05 which mean we can reject the null hypothesis, which means impact of the campaign is likely greater on weekends compared to weekdays.

Question 2: The impact on sales [30 points] [Note: change "weekend" back to a numeric variable]

a) Please use sale as dependent variables, all other variables except for *storeid, date*, *training* and *traffic* as independent variables, and build NNM with decay=0.5 and size= 2. Show the MAE and RMSE for All data. Calculate MAE%, and is this model acceptable regarding MAE%? Why? [hint: keep year, month, day, weekday as factor variable]. [10 points]

```
Evaluate predictions for regression models
Data        : GUESS_MODEL
Filter      : training==1
Results for : All
Predictors  : pred_nn2
Response    : sales

 Type Predictor     n   Rsq      RMSE       MAE
  All  pred_nn2 1,011 0.854 6,888.877 4,823.243
```

	Function
variable	mean
All	All
sales	26,609.514

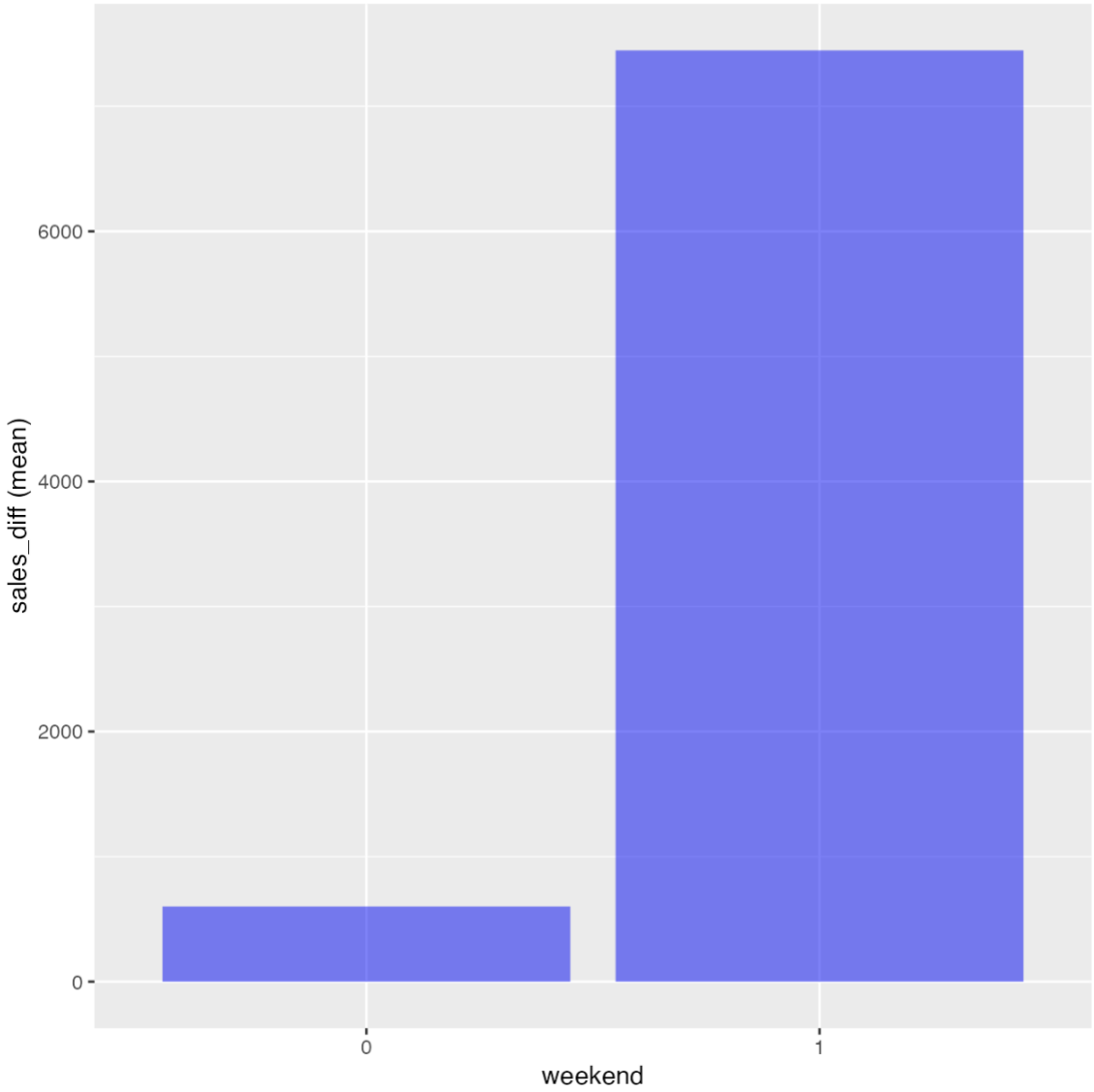
MAE% = 18.13%
It might not best model because the MAE% is too high close to 20 percent but it still can be acceptable because it's under 20%.

b) Based on the model in a), predict the sales in out sample data. Create a new variable, *sales_diff*, indicating the difference between real sales and predicted sales in the out sample data (sales_diff=real sales-predicted sales). Show the mean of *sales_diff* during campaign period. Please explain the meaning of this number. [10 points]

	Function
variable	mean
All	All
sales_diff	2,604.636

It shows the actually sales is more than predicted sales.

c) Show *sales_diff* between weekend and weekday. What do you observe? [5 points]



It shows model predict less accurate for the weekend compare to the weekday and also these are much less sales in weekday compare to weekend.

d) Based on the results in e) of question 1 and b) of question 2, what do you observe regarding the impact of GAP's campaign on the traffic and sales of Guess in Chicago respectively? Please come up with an explanation for this observation. [5 points]

Because the traffic we predict is close to real which cause the real sales is close to what we predict in weekday. And for weekend, there are more traffic than we predict, which also cause more sales. We can see more traffic will brings more sales.

e) [Bonus question] Based on the results in g) of question 1 and c) of question 2, what do you observe regarding the effect variations between weekday and weekend for the traffic and sales respectively? Please come up with an explanation for this observation. [5 points]

More traffic bring more Sales. When the traffic significant increase during the weekend, and the sales_diff is increased which means the actually sale is increased when the traffic increased. Which mean weekend cause higher traffic and higher traffic cause higher sales.