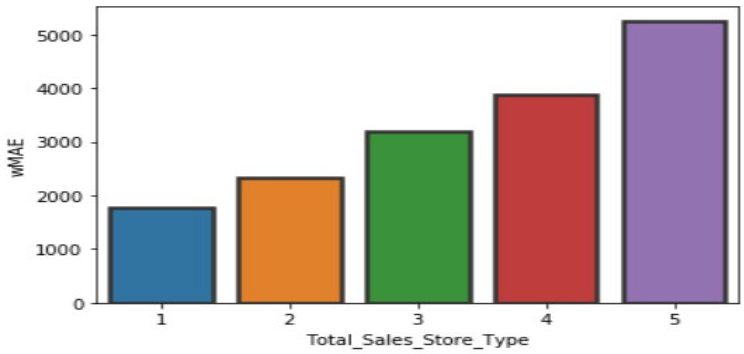
Task-3 Analyzing the resulting model

1. **Analyze the results for different Stores:**
    a) Compute wMAE separately for Stores of different types (`Total_Sales_Store_Type`).
    b) Is there any difference between errors obtained for different types of Stores? Plot a boxplot or bar plot of errors distribution (1 box/bar for every week type).

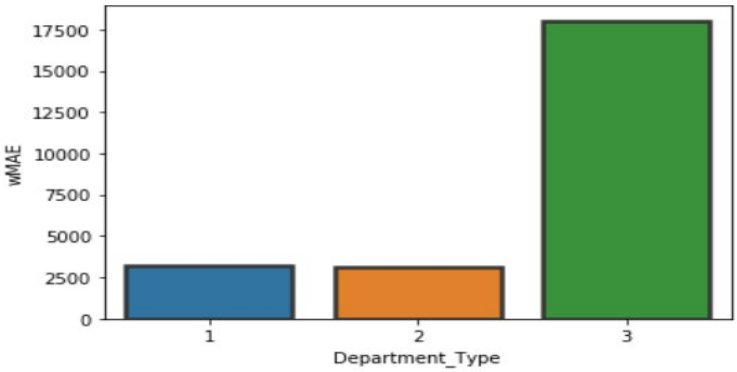
	Number of Instances	Total_Sales_Store_Type	wMAE
0	92233	5	5254.741462
1	69080	1	1775.510362
2	80712	2	2334.333414
3	88634	3	3199.978581
4	90911	4	3871.463128



Observation:
As we can observe that number of instances/examples in the data-set of each store type are different. Among all the store types, type-1 has less number of instances in the data-set on the other hand type-5 has highest number of instances in the data-set. We can also observe that wMAE is different for every type of stores. Store type-5 has highest value of wMAE and type-1 has lowest value of wMAE.

2. **Analyze the results for different Departments (Dpt).**
    a) Compute wMAE separately for different Department types (`Department_Type`).
    b) Is there any difference between errors obtained for different types of departments? Plot a boxplot or bar plot of errors distribution (1 box/bar for every week type).

	Department_Type	Number of Instances	wMAE
0	2	401547	3111.226028
1	1	12319	3211.441797
2	3	7704	18021.163880

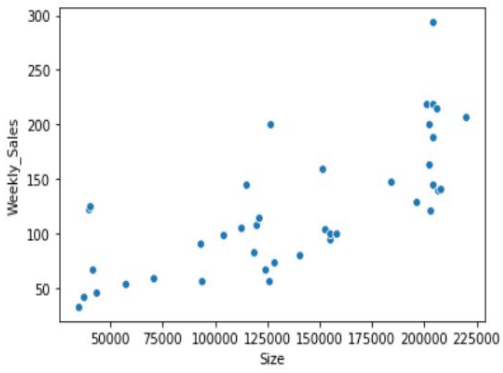


Observation:
In case of different types, we can observe that dept-3 has lowest number of instances and highest wMAE. So, we can conclude that due to less number of instances/examples of dept-3, model couldn't learn very well on that department.

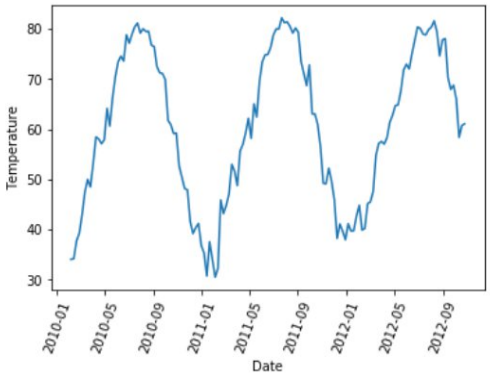
5. Make an overall conclusion. Your conclusion should include answers to (but not limited to) the following questions:

- How good is your final model for small/huge stores?
- What kind of additional features could you suggest for this task?
- Would you suggest building individual models for Stores/Departments/Regular-Holiday weeks/Nearest-Farthest weeks?
- Which features were most useful? Least useful?

Conclusion



Scatter plot between Weekly_Sales (in millions) in a store and Size of the store.
We can notice that Weekly_Sales of the stores is proportional to the Size of the store which is quite reasonable. Both variables have high correlation 0.72 so we can conclude that Size is important feature to predict the Weekly_Sales.



Line graph of temperature over time
We can easily observe that temperature during winter and spring is quite low compare to the summer and fall which is expected. Hence, we can conclude that temperature in the data-set is well recorded.

Grid search computation has been performed on the 10% of the data-set and tried to preserve the distribution of types of the stores.

On predicting the Weekly_Sales on small size and huge size of the stores, we got the around 87% accuracy using the best estimator/model which is quite well as we didn't use too complex feature to predict the Weekly_Sales and also number of feature are quite low.

Additional feature for this task: we consider the 'Coverage Area' for each store as a feature. It is more likely store having high coverage area will have high weekly_sales that's mean there will be high correlation between them. To get the 'Coverage Area' we can use the NearestNeighbors unsupervised leaner.

Building the individual models for Stores/Departments/Regular-Holiday weeks/Nearest-Farthest weeks is not a good approach because models will have less complexity which result into high Bias and low variance.

'Dept' is the most useful features to predict the Weekly_Sales, it has around 52% importance in the best estimator. 'Is_Christmas' is the least useful features which has around 0.2% importance.