Lab 2

You are provided with the weekly historical sales revenue from 45 Walmart stores (real data):

Variable	**Definition**
Weekly_Sales	Weekly sales for a given store ($)
CPI	Consumer Price Index *(Google it!)*
Size	Size of the store
IsHoliday	Whether the week is a special holiday week
Temperature	Average temperature of that region
Fuel_Price	Cost of fuel in that region
Unemployment	Unemployment rate in that region
Store	Store ID
Date	Start date of the particular week

This lab will provide you with the opportunity to work on a business case using real data. Start with the template, load the following libraries in the order: *tidyverse*, *tidymodels*, *plotly*, *skimr,* ***lubridate, car***. Remember, you load the libraries in the first chunk: `library('library_name')`

After loading the libraries, load the **walmartSales.csv** in the **data** folder as *dfw*. Then, explore the dataset using ***any of the following functions*** *as you see fit* to understand the data first: `head()`, `str()`, `glimpse()`, `nrow()`, `dim()`, `summary()`, and `skim()` -> Use wide screen!

1. Create a regression model using Weekly_Sales as the DV (Dependent Variable, outcome variable), and CPI as the IV (Independent Variable, feature, predictor, explanatory variable). *[If you don't remember how to run and interpret a linear model in R, see the appendix]*

   ```
   fitCPI <- lm(...fill in here...)
   summary(fitCPI)
   ```

   What is the coefficient of CPI, and what does it mean in plain English?

   A. The CPI coefficient is -732.7. When all the other variables are constant, a unit increase in CPI value results in the weekly sales to reduce by 732.7 units on average.

Based on the output, how good is the model explaining the variance in Weekly_Sales? Why? Given the fact, do you think your interpretation of the coefficient of CPI is still useful? Why?

A. The model is a pretty good one. The p-value is 3.33e-09 which is less than 0.001, that is good, and also when the CPI changes, we see a change in the weekly sales. Well, the interpretation of the co-efficient of CPI is useful because a change in CPI causes a change in Weekly Sales so it can be used accordingly.

2. For Store 10, create a scatter plot of the relationship between CPI and Weekly_Sales. Add a regression line to this plot. What do you observe? Does it align with your interpretation in Q1? Now, try it for Store 11, Store 12, and Store 13. What do you think is going on here?

   A. We can observe that the points start to spread out slightly from stores 10 through 13. Talking about store 10, the weekly sales come down slightly as the CPI increases. There are clusters observed at 3 different areas on the graph. Yes, it does align with the interpretation made previously. Apart from the fact that the points start to slightly spread out, there is no significant change observed. This differs the analysis performed before.

```
plot <- dfw %>%
                filter(...fill in here...) %>%
                ggplot(...fill in here...) +    => Set Weekly_Sales as y!
                geom_point() +
                geom_smooth(method=lm)

plot                 => For the static plot

ggplotly(plot)       => For the dynamic plot
```

3. Now, filter for the year 2012 instead of a store (so, you'll plot data from all stores in a year). For this, you will need to (install and) load the `lubridate` library. Check the cheat sheet for lubridate here. *[Start by copying/pasting your code from Q2 into a new chunk and reuse]*

   What do you observe? Why do you think there are almost vertical clusters of observations?

   A. The Consumer Price Index (CPI) is a measure of the average change over time in the prices paid by urban consumers for a market basket of consumer goods and services. How much the consumer might be willing to pay changes throughout the year. Say during black Friday or before Christmas, a consumer may be willing to pay more. Here, we have filtered by year 2012 so it shows data only for that year with all the different stores of Walmart. So, we can see the clusters formed because the CPI would be fixed during a particular time whereas there are changes in Weekly Changes. Hence the vertical line like structures. If weekly sales are very close in their spread, then the vertical lines will shorten.

4. Now, create a plot of sales in Store 1 in the year 2010. Did you know that you can use multiple arguments in one filter function as follows: *filter(argument_1, argument_2,…)?*

   Compared to the earlier plots, do you notice a difference in the range of CPI? Why is it so?
   A. So, it is bound to happen that as the CPI value increases, there are more scattered points in the graph as the buyers will show a range of characteristics. But, when the CPI values come down, the varied habits reduce.

5. Build another regression model but this time include both CPI and Size as independent variables and call it *fitCPISize*. Compare this model with the model you built in Q1.

   Which model is better at explaining Weekly Sales? Why? **Hint:** Use *anova()* **as well**.

   A. When we compare the model from Q1 to the model in Q5, we see that the p-value is very low in model of Q5. Hence, adding Size has really helped our model and made it a better fit with more clarity. It is kind of obvious that when we factor in the Size of the store, we can get more information as to why the sales are how they are for that store. Hence, it helps our model as there is a correlation and so adding in the Size variable results in more relevant information and a lower p-value.

6. Has the estimated coefficient for CPI changed? If so, why do you think it has changed?

   A. The CPI co-efficient became -657. Since we add Size which is another independent variable to the mix, we will experience a change in the dependence value on CPI, which is the case here, because Size also factors in.

7. Let's build a full model now and call it *fitFull*. This time, include all the variables in the dataset **(EXCEPT Store AND Date)** and report your observations. You can **also** use *anova()* to compare the reduced model in Q5 with the full model you have just built in this question.

   A. The use of anova points towards the fact that the RSS value of fitFull is very low as compared to that of fitCPISize. The adjusted R square value for fitFull is 0.6206 and that of fitCPISize is 0.6155. Independent variables other than CPI and Size have an effect on the Weekly Sales here.

8. The output of Q7 shows that temperature is positively associated with weekly sales. However, is that relationship really linear? Test it out by adding a squared transformation of temperature into the model using the following *I(Temperature^2)* and call it *fitFullTemp*

   What is the coefficient of the squared term? Is it statistically significant? What does it mean? Based on this, what would you do differently if you were managing Walmart's promotions?

A. The coefficient of the squared term is negative 19.822. Yes, it is statistically significant. From the visualization, it is observed that the sales go upward as the temperature starts reaching optimal values. But if the temperature increases too much, the sales go down. Either way, if the temperature is too low or too high if affects the sales negatively. Sales appear to be maximum at approximately 47 degrees and fall on both the sides identically. So, if managing Walmart's promotions, it should be kept in mind that promotions should be more aggressive in the periods of harsh weather so as to increase the sales as we see a drop during that time. Walmart can roll out the concepts of Summer Sales or Winter Sales depending on the region. Other tactics like encouraging online shopping, providing coupon codes for online purchases only during a particular season, giving away free goodies on a particular amount of purchase, etc. could be used.

Let's visualize the relationship of temperature with sales to understand what it means:

```
dfw %>% ggplot(aes(...fill in here...)) +
         geom_smooth(...fill in here..., formula = y ~ x + I(x^2))
```

This is to visualize the shape of the relationship between temperature and sales. Note that the values shown in this plot are not accurate (but the curved shape is) because the model defined in the ggplot() above includes only Temperature and its squared form as the two independent variables.

9. In a true predictive analytics exercise, we need to split the dataset, train the model using the training dataset and make predictions using the test set. Now, let's do it the predictive way. *[Do **not** hesitate to copy and paste **your own code** from above, change, and reuse here]*

   a. Set the seed to **333** [Always set the seed and split your data in the same chunk!]
   b. Randomly sample 80% of the data for training, and assign the difference to the test set

      ```
      dfwTrain <- dfw %>% sample_frac(...fill in here...)

      dfwTest <- dplyr::setdiff(dfw, dfwTrain)  =>Make sure dplyr is there!
      ```

   c. Run the model from **Q8 using only the training set** now, and store the model as *fitOrg*
      [By the way, as a prep for more advanced analysis work, try running `tidy(fitOrg)`
      Can you imagine the benefits of being able to convert a regression output into a tibble?]

   d. Create a new copy of the test data frame *dfwTest* by adding the predicted values as a new column. Name this new dataframe as *resultsOrg*

      ```
      resultsOrg <- dfwTest %>%
                    mutate(predictedSales = predict(fitOrg, dfwTest))
      ```

      Before you press on, check out *resultsOrg*, the new data frame you have just created.

   e. Calculate the performance measures by calling the *rmse(resultsOrg, truth=..., estimate=...)* and *mae(...)* functions and inputting the values stored in *resultsOrg*

      What do they mean? Are they interpretable? If so, how do you make sense of them?

Hint: Instead of running the rmse() and mae() functions separately every time, you can:

```
performance <- metric_set(rmse, mae) => You need to run this only once!
performance(...fill..., truth=...fill..., estimate=...fill...)
```

A. The rmse value is 236686.8 and the mae value is 177862.6. Essentially, they depict the error in the models. They are interpretable. RMSE can be interpreted as the standard deviation of the unexplained variance. Lower values of RMSE indicate better fit. MAE gives the absolute values of error and do not explode with a higher variance like RMSE does.

f. Now, add the variable Date to create a new model *fitOrgDate*, repeat the process in (c)-(d)-(e) to create a new results table *resultsOrgDate*, and calculate the performance of the new model using the new predictions. Has the model improved? Why or why not?

A. The adjusted R-square value is 0.6181. The adjusted R-square value for fitOrg model is 0.6174. We can say that the model is showing slight improvement as higher the value of adjusted R-square, the better it is. Also, the errors come down since the values of RMSE and MAE see a drop in the second model. Thus, the second model is better.

Now, take the same question from an explanatory perspective. Would you keep Date?

A. Date is more often than not an important part. It helps put up a timeline to data. Date here will give us a detailed perspective and hence I would keep the date.

g. Remove Date and go back to your original model *fitOrg* from Q9c. This time, remove Unemployment and build a new model *fitOrgNoUn*. As you did in (f), make predictions using the test set and calculate performance. Has the model improved? Why or why not? Is your conclusion about Unemployment the same for both to predict and explain?

A. The adjusted R-square value is 0.6175. The adjusted R-square value for *fitOrgDate* model is 0.6181. So, we can say that the model has regressed. Also, the values of RMSE and MAE increase in the second case as compared to the first. Hence, it backs up the conclusion that the second model is not better. The omission of Unemployment might have impacted the model is a bad way since it is an important aspect to consider while building our model.

10. The finale has to be sweet, right? Instead of using sales, create a log-transformed version, set the seed, split the data, run the model *fitLog*, make predictions, calculate performance.

a. Have the coefficient estimates and variance explained in DV improved? Compare the model output and performance of *fitLog* with that of *fitOrg* from Q9c, and discuss.

A. The adjusted R square value improved. It became 0.7078. for fitLog. The co-efficients have improved. Also, the RMSE and MAE values have reduced. SO, the new model is better.

b. Check and compare the diagnostics from *fitLog* with those from *fitOrg*, and discuss.

A. The line is more linear in the *fitLog* model as compared to a curved line in *fitOrg* model in the residuals vs fitted plot. Also, in the Q-Q plot, fitLog behaves more normally because we can see a linear trend. In residual vs leverage plot, the spread is even. Hence, the fitLog model is a better fit for the data.

Bonus question: Instead of predicting sales, you may also want to create a new dependent variable by dividing the Weekly Sales by store Size ("Sales per square foot" -makes sense if you focus on the utilization of store space, for example). Call it *fitSalesSqFoot*. For this exercise, like in Q10, create a variable, set the seed, split the data, make predictions, calculate performance. What do you think is going on here? Discuss. In addition, in this model, you may want to try removing the variable Size, because your DV is a function of it now. Explore the differences.

A. The model has a very low RMSE and MAE values which indicates that this is good and fairly accurate model. First, Size is considered in the model as an independent variable alongside other variables. Eliminating Size increases the RMSE and MAE values which initially might tell that it is not good but when further investigation is done, it can be seen the model was overfitting. So, increase in RMSE and MAE actually makes the model a better fit.