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. Before you start, load the following libraries in the given 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?
   -732.7
   On an average, with every unit change in CPI, the weekly sale changes by $732.7.

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? The model is not good at explaining the variance in Weekly_Sales as the R-squared value is very low. Yes the coefficient of CPI is still significant as the p-value is very low and the f-statistic is large.

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?

   The graph shows a slight decrease in weekly sales with increase in CPI. The variance is more or less constant. Yes, it aligns with Q1, according to which the weekly sales must decrease with one unit change in CPI, on average. Store 11 have a constant regression line while stores 12 and 13 have an increasing regression line. Even when the individual stores have different regression trends, the overall regression model gives an average of this, hence the difference in the individual stores models and the combined model.

```
plot <- dfw %>%
               filter(...fill in here...) %>%
               ggplot(...fill in here...) +
               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? The graph shows that as the CPI increases there is a decrease in the weekly sales. The vertical clusters are because of the range of different Walmart stores present in different locations with different weekly sales for almost similar CPIs. One can conclude that some of these stores have a higher footfall than the others.

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? Yes, the range of CPI for store 1 is bigger because the number of data points for this graph is limited to a single store and a single year unlike the previous plots which had a large number of data points. The plot tends to stretch in areas with less number of data points.

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**.

   This model is better at explaining weekly sales as the anova result has a low p-value. Also, the second model had high f-statistic and f-values meaning the coefficients of the variables are highly significant.

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

   Yes, because now another dependent variable has been introduced to the equation which, according to the model has a significant contribution to the independent variable.

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.

   This model is also statistically significant as it has a high R-squared value and low p-value. Most of the dependent variables in this model are significant. Also, size holds a particular significance in this model as it has a high f-value. Given the very low p-value of the anova result, this model is much better than the model in Q5.

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?

   -19.82 is the coefficient of the squared term. Yes, it is statistically significant as the p-value for the variable is less than 0.5. This means that the weekly sales of a store is dependent on the temperature of the area where the store is located. Based on this information I would keep different prices for the products depending on whether it is a hot day, a cold one or a pleasant one, to maximize sales, after interpreting how exactly sales differ on these days.

   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 <- setdiff(dfw, dfwTrain)
   ```

   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?]*

   Converting a regression to tibble would give a structured result providing ease of interpretation.

   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...)
   ```

   They tell us about how efficient was the model in predicting the values when compared to the actual test data. Yes, they are interpretable. The predicted values of weekly sales is off by $240942 as compared to the actual values. Also, the residuals are on an average $180045 in magnitude.

   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?

Yes, the model has slightly improved as the errors are slightly lower as compared to the previous model. This is because date has a small role to play in predicting the weekly sales of a store.

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

Since the p-value of the anova test is not less than 0.05, I would not keep Date in the model as this shows that the improvement in not statistically significant.

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?

No, the model hasn't improved as the errors have gone up. This is because Unemployment is a significant statistic in determining the weekly sales.

In the explanatory sense, the conclusion will be the same i.e. to not choose this model as the p-value of the anova result is very low indicating that the full model performs better than the reduced 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.

Yes, the coefficient estimates and variance has improved in fitLog. FitLog has a greater r-squared value indicating that it is more statistically significant than fitOrg. Also, MAE is smaller in case of the fitLog model.

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

The 2 plots show that both models are good at capturing the nonlinear relationship however fitLog captures the nonlinear relationship better than fitOrg as fitOrg has a slight curve whereas fitLog has an almost straight line.

FitOrg violates the normality assumption of linear regression while the residuals in fitLog line up well on the line.

FitOrg has more heteroscedasticity than the fitLog model as in the fitLog model the points are spread equally and the line is approximately flat which is not the case for fitOrg.

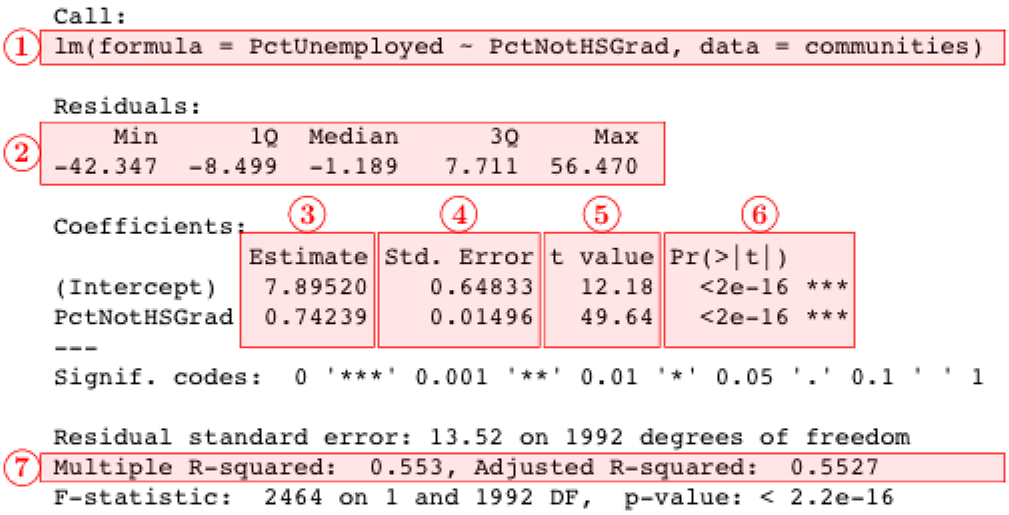
FitOrg has more outliers affecting the regression line than fitLog.

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.

The r-squared value is not that good and RMSE and MAE are greater than the other models which tells us that this model is not good and efficient in predicting the weekly sales per size of a store. Removing the Size variable worsens the r-sqaured value and the errors which shows that the variable should be present in the model as it is significant.

Appendix: How to run a linear regression in R and read the regression output

The output from summary() may seem overwhelming at first, so let's break it down one item at a time:



```
Call:
lm(formula = PctUnemployed ~ PctNotHSGrad, data = communities)

Residuals:
    Min      1Q  Median      3Q     Max 
-42.347  -8.499  -1.189   7.711  56.470 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   7.89520    0.64833   12.18   <2e-16 ***
PctNotHSGrad  0.74239    0.01496   49.64   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.52 on 1992 degrees of freedom
Multiple R-squared:  0.553, Adjusted R-squared:  0.5527 
F-statistic:  2464 on 1 and 1992 DF,  p-value: < 2.2e-16
```

#	Item	Description
1	formula	The formula describes the relationship between the dependent and independent variables.
2	residuals	The differences between the observed values and the predicted values are called residuals (errors).

3	coefficients	The coefficients for all the independent variables and the intercept. Using the coefficients we can write down the relationship between the dependent and the independent variables as: `PctUnemployed = 7.90 + ( 0.74 * PctNotHSGrad )` This tells us that each unit increase in the variable `PctNotHSGrad` is associated with the increase of the variable `PctUnemployed` by 0.74, on average.
4	standard error	The standard error estimates the standard deviation of the sampling distribution of the coefficients in the model. Think of the standard error as a measure of precision for the estimated coefficients.
5	t-statistic	The t-statistic is obtained by dividing the coefficients by the standard error.
6	p-value	The p-value for each of the coefficients in the model. Recall that according to the null hypothesis, the value of the coefficient of interest is zero. The p-value tells us whether we can reject the null hypothesis or not.
7	R^2 and adj-R^2	R-squared and adjusted R-squared tell us how much of the variance in our model is accounted for by the independent variables. The adjusted R^2 is always smaller than R^2 as it takes into account the number of independent variables (and penalized accordingly).

Source for the annotated output: Blumenau and Ali of University College London (2019)

Important note: The data used in this lab is from multiple stores and has a time component. Ideally, you would want to take into account this fact. One way of doing so is to include *Store* and *Date* in your models as categorical variables. However, because the sample size is small, adding so many categorical variables would reduce the power of your analysis. Try adding them if you like, and you will see a warning: “prediction from a rank-deficient fit may be misleading.” The models are informative for educational purposes, but the results may require some caution.