The goal of this assignment is to get you started with predictive analytics. You will first prepare and explore the data, and run a basic regression. You will then predict the variable COUNT as a function of the other variables. You will also determine the effect of bad weather on the number of bikes rented. Finally, you will build alternative models, measure and compare their predictive performance, make *data-informed* and *data-driven* inferences for a business case.

Assignment Instructions

You will use data from DC's Capital Bikeshare (also serves Maryland and Virginia). Capital Bikeshare has about 30K members, and served about 23.6 million trips through its 550 stations. In this dataset, we combined the Capital Bikeshare data with weather data to gather insights.

Data Dictionary:

1. DATE -*You'll also create a* MONTH *variable using this*
2. HOLIDAY: Whether the day is a U.S. holiday or not.
3. WEEKDAY: If a day is neither a weekend nor a holiday, then WEEKDAY is YES.
4. WEATHERSIT: The values are (1) Clear/Few clouds (2) Misty (3) Light snow or light rain (4) Heavy rain, snow, or thunderstorms.
5. TEMP: Average temperature in Celsius.
6. ATEMP: "Feels like" temperature in Celsius.
7. HUMIDITY: Humidity out of 100 (not divided by 100).
8. WINDSPEED: Wind speed in km/h.
9. CASUAL: Count of bikes rented by casual bikeshare users.
10. REGISTERED: Count of bikes rented by registered bikeshare members.
    COUNT: Total count of bikes rented by both casual users and members **-You'll create this**

Before you start:

- Load the following four libraries in the given order: *tidyverse, tidymodels, plotly, skimr*
- Load the bikeshare data and call it *dfbOrg*
- Explore the dataset using `skim()` etc.

1) Data preparation

a) **Create the additional variables:**

i) Create the COUNT variable and add it to the data frame.

ii) Extract MONTH from the DATE variable and add it to the data frame. **This time, do NOT use lubridate. Use the native `months()` function instead.**

b) **Scale the data (and save it as *dfbStd*):** Start by standardizing the four variables, TEMP, ATEMP, HUMIDITY, WINDSPEED. If you don't remember what it

means to standardize a variable, see the link. Surely, you don't need to do this manually!

2) **Basic regression in R:** In `dfbStd`, run a regression model `fitAll` using COUNT as the DV, and all the variables as independent variables. [Don't forget to use summary(fitAll)]
   a) Does this appear to be a good model? Why or why not?

Ans - It appears to fit the data too perfectly, causing overfitting. Both R^2 and Adjusted R^2 being 1 implies that there is an overfitting issue as it is not possible to meaningfully obtain a model with R^2 = 1. This suggests that there is 0 variance in the model which is not practically possible.

   b) According to your model, what is the effect of humidity on the total bike count in a formal interpretation? Does this finding align with your answer to Part (a)?

Ans - As humidity increases by 1 unit, there is an increase in the number of bikes used (count) by 1.4 * 10^-13. It is not significant and doesn't make much sense. Thus, it does align with the answer to part a - that the model is not a good one.

In the rest of the assignment, use the original data frame `dfbOrg`:

3) **Working with data and exploratory analysis:**
   a) Add a new variable and call it **BADWEATHER**, which is "YES" if there is light or heavy rain or snow (if WEATHERSIT is 3 or 4), and "NO" otherwise (if WEATHERSIT is 1 or 2). You know what functions to use at this step.
   b) Present a scatterplot of COUNT (y-axis) and ATEMP (x-axis). Use different colors or symbols to distinguish "bad weather" days. Briefly describe what you observe.

Ans - When Badweather condition is true, the count decreases drastically. Within this cluster, as the temperature remains below 10 degrees, the count remains constant, then it increases as temperature reaches 20 and then shows a flat trend again. Within the bad weather cluster, when temperatures are extreme, count shows decreasing trend.
If badweather condition is not true, the count goes on increasing as "feels-like" temperature (ATEMP) increases till about 30 degrees and then shows a decreasing trend. Extreme temperatures are associated with less bike usage and moderate temperature shows maximum usage.

   c) Make two more scatterplots (and continue using the differentiated coloring for BADWEATHER) by keeping ATEMP on the x-axis and changing the variable on the y-axis: One plot for CASUAL and another for REGISTERED. Given the plots:
      i) How is *temperature* associated with casual usage? Is that different from how it is associated with registered usage?

Ans - As temperature increases, bike usage increases up to a point. After that the usage dips for both casual & registered users. However, the rate of change seems to be more in case of casual users. In case of registered users, the points are clustered together more, indicating less rate of change in usage with temperature. Points are more scattered for casual users, indicating more effect of temperature on usage. Overall, the effect of temperature on casual users is more as compared to registered users.

ii) How is *bad weather* associated with casual usage? Is that different from how it is associated with registered usage?

Ans - There are very few casual users when there is bad weather. It also show a very flat trend when there is bad weather. As temperature increases there is a slight increase.

There are still a significant number of users when the weather is bad in case of registered users. They also show a more prominent increase in usage when there is bad weather and temperature increases.

Overall, this suggests the effect of bad weather on casual users is more as compared to registered users.

iii) Do your answers in (i) and (ii) make logical sense? Why or why not?

Ans - It makes logical sense to see that registered users would continue to use bikes if the weather conditions are manageable and since they have already paid for registration. However, casual users pay per ride and if the weather is bad or temperatures are too high or low, they would explore other options as they have not paid anything yet.

iv) Keep ATEMP in the x-axis, but change the y-axis to COUNT. Remove the color variable and add a geom_smooth() without any parameters. How does the overall relationship between temperature and bike usage look? Does this remind you of Lab 2? Why do you think the effects are similar?

Ans - As the temperature increases, the overall bike usage increases up to a point and then starts decreasing. Extreme temperatures show less usage & moderate temperatures show high usage.

It is similar to the relationship between weekly sales and temperature.

This could be because people are more likely to go out and be more active when the temperature is moderate (pleasant climate) rather than when it is too high or too low. Both shopping and bike usage are activities requiring users to go out and thus the effect of temperature is similar.

4) More linear regression: Using `dfbOrg`, run another regression for COUNT using the variables MONTH, WEEKDAY, BADWEATHER, TEMP, ATEMP, and HUMIDITY.

a) What is the resulting adjusted R^2? What does it mean?

Ans - Adjusted R^2 is 52.1%, meaning around 52% of the observed variation in bike usage can be explained by the model.

b) State precisely how BADWEATHER is associated with the predicted COUNT.

Ans - On average, when there is bad weather (BADWEATHER = YES), the count decreases by 1954 as compared to when the weather is not bad (BADWEATHER = no), all else being held constant.

c) What is the predicted count of rides on a weekday in January, when the weather is BAD, and the temperature is 20^o and feels like 18^o, and the humidity is 60%?

Ans - Equation :

3967.981 + (-858.334)*(1) + (69.745)*(1) + (-1954.835)*(1) + (184.596)*(20) + (-48.640)(18) + (-25.431)(60)
Predicted count of rides = 2520.497 = 2520

d) Do you have any concerns about this model or your predicted COUNT in **Q4-c**? Why or why not?

Ans - The model considers temperature, feels like temperature as well as bad weather. These variables might not be independent of each other as these are all interrelated aspects of the same entity - weather. Thus, it might not be very accurate in predicting the count value above, since the interaction between variables has not been considered.

5) **Regression diagnostics:** Run the regression diagnostics for the model developed in **Q4**. Discuss whether the model complies with the assumptions of multiple linear regression. ***If you think you can mitigate a violation, take action,*** and check the diagnostics again. **Hint:** The Q-Q plot and the other diagnostics from the plot() function look fine to me!

Ans - By checking the diagnostics using plot() function, four critical diagnostics have been verified (linearity, normality, homodesckacity and influential cases) and this model doesn't seem to violate any of these conditions.
However, on checking the multicollinearity (inter-association among independent variables) assumption using vif function, gives very high gvif values for temp and atemp. This indicates there is considerable multicollinearity between both.
Upon trying various models - 1. with only each of temp, atemp 2. Temp and interaction variable, 3. Atemp and interaction variable, the third model gave best performance measures and diagnostics. Thus, considering this as the best model - with WEEKDAY + MONTH + ATEMP + BADWEATHER + HUMIDITY + ATEMP*BADWEATHER as independent variables.

6) **Even more regression:** Run a simple linear regression to determine the effect of bad weather on COUNT when **none** of the other variables is included in the model.
   a) Compare the coefficient with the corresponding value in **Q4**. Are they different? Why or why not?

Ans - Coefficient of bad weather is -2780.95. It is different from Q4, which was -1954.8.
This could be because BADWEATHER has a correlation with some other independent variable and we have not considered this in the model, violating the independence of independent variables assumption.

   b) A consultant has indicated that bike use is affected differently by bad weather on weekdays versus non-weekdays, as people go to work on weekdays. How can you add this domain knowledge to the regression model you built in (a)? Why?

Ans - This information can be incorporated by adding WEEKDAY categorical variable in the model. This is because it will be able to capture the effect of a day being weekday or non-weekday on bike usage.

c) Run a new model with your addition from (b). Is this a better or worse model than your original model in (a)? How do you decide?

Ans - This is slightly better than the above model as R^2 and Adjusted R^2 values have decreased slightly implying the variation has been explained slightly better by this model. However, p value has increased, implying the above model is slightly better as there is less probability of getting such a result due to randomness in the earlier models.

d) Using your model from (c),
   i) interpret the average effect of bad weather on the COUNT depending on whether it is a weekday or not, and

Ans - On an average every unit increase/decrease in Bad weather is associated with a 2637 decrease/increase in count as compared to when the weather is not bad and it is not a weekday, holding all else constant.

   ii) quantify the effect of bad weather on the COUNT in different scenarios (be sure to calculate *all* effect sizes for the **four alternatives (2x2)** here).

*[In calculating the effects here, do **not** worry about the statistical significance]*

	BADWEATHERYES	BADWEATHERNO
WEEKDAYYES	4452.5 + (-2637.1) +185.3 + (-201.2) = 1799.5	4452.5 + 185.3 = 4637.8
WEEKDAYNO	4452.5 + (-2637.1) = 1815.4	4452.5

7) Predictive analytics: Follow the steps below to build two predictive models. Which model is a better choice for predictive analytics purposes? Why? Does your conclusion remain the same for explanatory analytics purposes? Please copy and paste the predictive and explanatory performance levels of both models into your response.

a) Set the seed to **333** (Always set the seed and split your data in the same chunk!).
b) Split your data into two: 80% for the training set, and 20% for the test set
   i) Call the training set *dfbTrain* and the test set *dfbTest*
c) Build two different models, calculate, and compare performance.
   i) The first model will include the variables in **Q4 with any adjustments you may have made during the diagnostics tests in Q5** (call this one `fitOrg`). The second model will add WINDSPEED to this model -*Call it* `fitNew`.

Hint: Remember, every time you build a new model, there are three steps you need to follow to be able to calculate the predictive performance of the model:

i. Build the model and store it as *fitXxx*
ii. Create a new copy of the test dataset *dfbTest* by adding the predicted values as a new column. Name this new dataframe as *resultsXxx*

iii. Calculate the performance measures (RMSE and MAE) using the actual and predicted values stored in the results dataframe *resultsXxx*

-*You'll replace* Xxx *with the model names you use (*Org *&* New *are suggestions)*

You may have trouble with the metric_set() function if you used `modelr` in Q5 for the diagnostics test. Trouble means learning. If you run the following code, you can simply ask R to unload `modelr` and you'll be fine: **`detach('package:modelr', unload=TRUE)`**

Ans - Model 2 (with WINDSPEED) seems to be better for both explanatory & predictive purposes.
This is because it has a better R^2 value (more relevant for explanatory models as it explains the variation in COUNT based on independent variables).
The values of RMSE and MAE are also lower in Model 2, which indicates this model has less as compared to Model 1, which is an important parameter in predictive models.

Explanatory performance measures:

1. Model 1 -

```
Residual standard error: 1354 on 568 degrees of freedom
Multiple R-squared:  0.5229,     Adjusted R-squared:  0.5094
F-statistic: 38.91 on 16 and 568 DF,  p-value: < 2.2e-16
```

2. Model 2 -

```
Residual standard error: 1322 on 567 degrees of freedom
Multiple R-squared:  0.546,      Adjusted R-squared:  0.5324
F-statistic: 40.11 on 17 and 567 DF,  p-value: < 2.2e-16
```

Predictive Performance Measures:

1. Model 1 (Without WINDSPEED)-

.metric <chr>	.estimator <chr>	.estimate <dbl>
rmse	**standard**	**1385.959**
mae	**standard**	**1174.862**

2. Model 2 (With WINDSPEED)-

.metric <chr>	.estimator <chr>	.estimate <dbl>
rmse	**standard**	**1340.903**
mae	**standard**	**1149.881**

8) **More predictive analytics:** In this final question, experiment with the time component. In a way, you will almost treat the data as a time series. We will cover time series data later, so this is just a little experiment. Taking into account date, you can't split your data randomly (well, evidently, you would not want to use future data to predict the past). Instead, you have to split your data by time. Start with `dfbOrg` and **use the variables you used in `fitOrg` from Q7c**. Split your data into training using the year "2011" data, and test using the "2012" data. Has the performance improved over the random split that assumed cross-sectional data (which you did in the previous questions)? Why do you think so? Split again by assigning 1.5 years of data starting from January 1st, 2011 to the training set and the remaining six months of data (the last six months) to the test set. Does this look any better? Discuss your findings.

Ans - The performance has significantly improved in the first case (training data with 2011 and test with 2012) as compared to the random split model. Current Adjusted R^2 is 78.23%. Adjusted R^2 value for random split was in the 50-55% range.
This is because when data is split according to the time range, we are using historical data to predict future data. This will help in capturing the relationship between subsequent observations, which is very significant in time series data. When time series data is split randomly, we lose this aspect, thus resulting in an inferior model.

When more training data is added (considering 1.5 years instead of 1 year for training data), the errors increase and R^2 value decreases, indicating the earlier model is a better predictive and explanatory model.
This could be because, when we use a year's data to train the model and predict subsequent year's count, it might be able to capture the seasonality and trends better. However, when we increase the training data set, it leads to loss of these seasonality and trend factors, leading to decrease in model performance.

9) **Data-*informed* decision making:** Based on your quick analysis of the Capital Bikeshare data, what are some actions you would take if you were managing Capital Bikeshare's pricing and promotions? How do you think you would use your predictions?

Ans - Possible Business Decisions based on above models:

1. Decrease prices during less demand - During weekends, holidays , decrease the price of usage, especially for casual users as they are more affected by inclement weather conditions.

2. Decrease prices for casual users during inclement weather conditions (based on weather predictions) as there is a lot of variation in usage of casual users during bad weather conditions.
3. Reduce the price of registration and more targeted marketing focusing on casual users - As we know users are more likely to stick once they join, even when the weather conditions are not ideal or it's not a high-demand day (weekday), focus should be on obtaining more registrations. This would lead to an increase in the number of registered users, thus leading to more regular usage and higher and sustained revenues. The slight drop in prices should be such that it can be compensated by increased revenue due to more users.

10) Data-*driven* solutions to "the" big challenge of bikeshare: As shown in the visuals on the next page, Capital Bikeshare (like most other shared services) has an inherent challenge. In the morning, people use bikes to commute to their workplaces, leaving the bike racks empty in residential areas (this is called *rush-hour surge*). In the evening, the same phenomenon repeats in the opposite direction. Shared-service companies attempt to resolve this problem by *rebalancing*, which is basically moving bikes manually during the off-peak hours using trucks (which you may have seen on the streets) and other means. **Assuming you have access to all the data Capital Bikeshare collects, and you can collect new data**, what is a data-driven solution you would pursue? Be specific about the data you would collect (if any) and the analytics project/model you would use.

Ans -

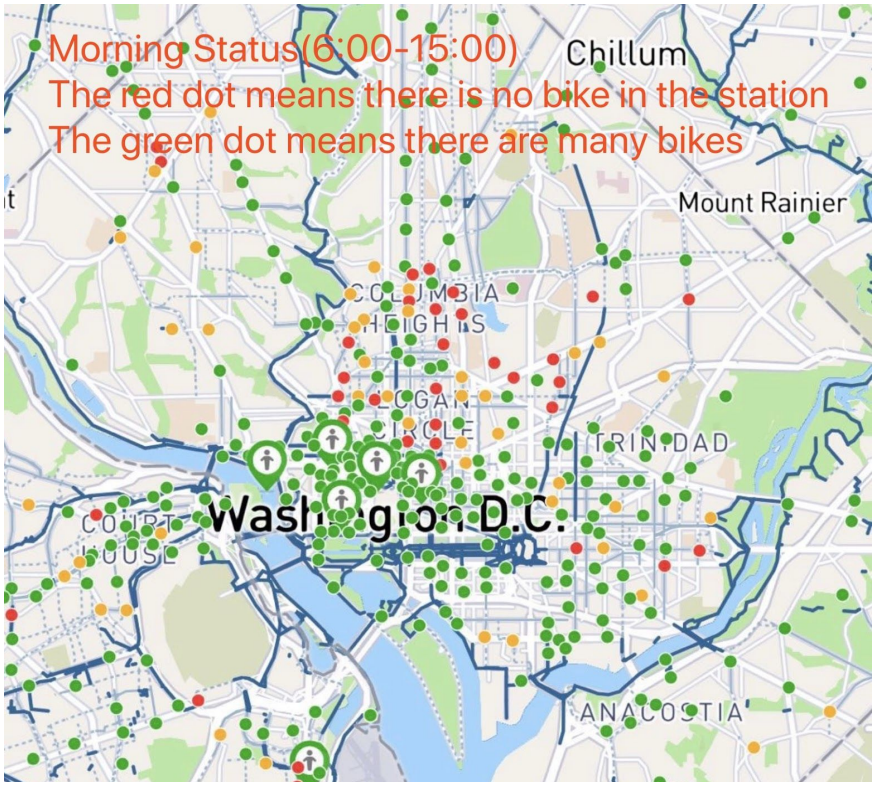
Analysis of problem - We need to check if people need bikes during off-peak hours, the scale of loss due to the unavailability of bikes during off-peak hours due to this issue, and how much rebalancing costs.

Additional Data Collection - For this, we need the following data:

1. Location
2. Proximity to office/residential area
3. Trip durations
4. Start Times
5. End Times
6. Data on user demographics - office goes, regular users, casual users etc.

If it is found that rebalancing is needed, and it is costing a significant amount, then looking at an alternate way to redistribute the bikes during off-peak hours would be an option. By reducing the cost of usage during off-peak hours, targeting users who use the service at these locations during off-peak hours regularly, customers can themselves be used to redistribute the bikes. Rebalancing costs given to truck companies can be saved by using it as incentive to customers for using services during off-peak hours, leading to some cost savings, more regular users, thus leading to long term wins as well as more popularity and better user experience.

Morning -Green dots are stations with many bikes, red ones are those with no bikes:



Evening -Green dots are stations with many bikes, red ones are those with no bikes: