

**ASSIGNMENT 2**

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 [bikeShare.csv](https://drive.google.com/file/d/1LVWmJTMyweqjxqDskhWErz0COin7d6GO/view?usp=sharing) from DC’s [Capital Bikeshare](https://en.wikipedia.org/wiki/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 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**
   1. **Create the additional variables:**
      1. Create the COUNT variable and add it to the data frame.
      2. 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.**
   2. **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](https://en.wikipedia.org/wiki/Standard_score#Calculation). 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) ]
   1. Does this appear to be a good model? Why or why not?
   2. 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)?

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

1. **Working with data and exploratory analysis:**
   1. 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.
   2. 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.
   3. 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:
      1. How is *temperature* associated with casual usage? Is that different from how it is associated with registered usage?
      2. How is *bad weather* associated with casual usage? Is that different from how it is associated with registered usage?
      3. Do your answers in (i) and (ii) make logical sense? Why or why not?
      4. 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?
2. **More linear regression:** Using dfbOrg, run another regression for COUNT using the variables MONTH, WEEKDAY, BADWEATHER, TEMP, ATEMP, and HUMIDITY.
   1. What is the resulting adjusted R2? What does it mean?
   2. State precisely how BADWEATHER is associated with the predicted COUNT.
   3. What is the predicted count of rides on a weekday in January, when the weather is BAD, and the temperature is 20o and feels like 18o, and the humidity is 60%?
   4. Do you have any concerns about this model or your predicted COUNT in **Q4-c**? Why or why not?
3. **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!
4. **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.
   1. Compare the coefficient with the corresponding value in **Q4**. Are they different? Why or why not?
   2. 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?
   3. 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?
   4. Using your model from (c),
      1. interpret the average effect of bad weather on the COUNT depending on whether it is a weekday or not, and
      2. 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 ]*

1. **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.
   1. Set the seed to **333** (Always set the seed and split your data in the same chunk!).
   2. Split your data into two: 80% for the training set, and 20% for the test set
      1. Call the training set *dfbTrain* and the test set *dfbTest*
   3. Build two different models, calculate, and compare performance.
      1. 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:

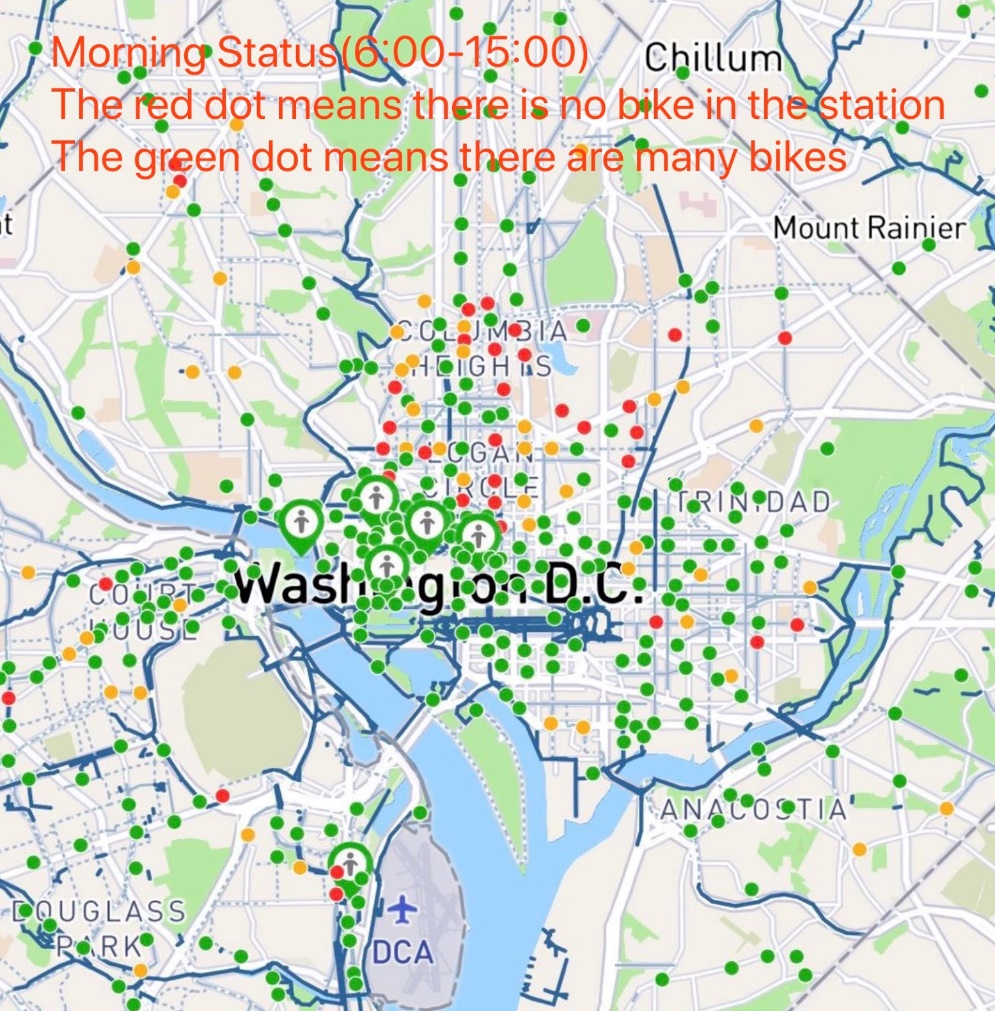
1. Build the model and store it as *fitXxx*
2. Create a new copy of the test dataset *dfbTest* by adding the predicted values as a new column. Name this new dataframe as *resultsXxx*
3. Calculate the performance measures (RMSE and MAE) using the actual and predicted values stored in the results dataframe *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)**

1. **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.
2. **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?
3. **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.

**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:**

