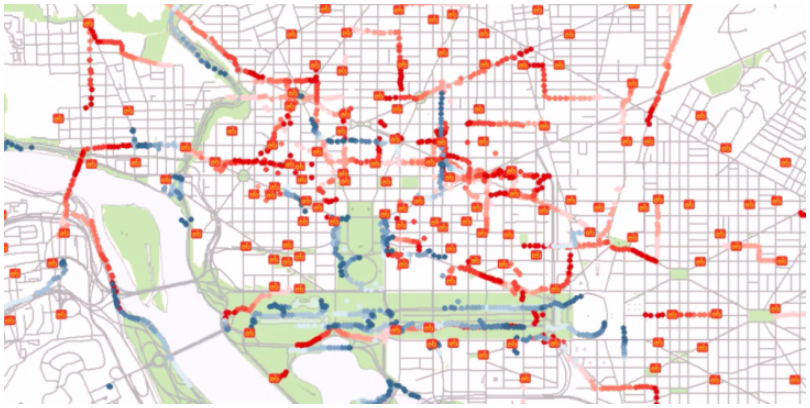
Unit 6: Linear Regression with Bike Sharing Data



The goal of this session is gain working exposure to linear regression by developing models for bike rental counts based on environmental and seasonal factors. You are expected to work through these problems in preparation for the discussion in the live session.

- Click link: Capital Bike Share GPS Study - YouTube video
- Click link: Bikeshare GPS insights

Part 1: Exploring the data

The file `bikes.csv` contains data provided by Capital Bikeshare, a local bike sharing program in the Washington, DC area. The file consists of information measured on a daily basis. For each day, the following variables were recorded:

- `Day`: day number in chronological order, starting with 1.
- `Count`: the number of bike rentals on that day.
- `Temperature`, `Windspeed` and `Humidity`: weather measurements on that day. The temperature is given in °F, the wind speed is given in miles per hour (mph), and the humidity is given in percentage.
- `Month`: the month of the year (1 through 12).
- `DayOfWeek`: Day of the week this data was collected (1 is Monday, 2 is Tuesday, etc.)

Question Set 1:

Read the data into a data frame called `bikes`.

1. How many observations are in the data set?
2. What is the average number of daily bike rentals?
3. What is the maximum number of bikes rented on any given day?
4. What is the average temperature in DC for the days recorded?
5. Create a graph showing the distribution of daily bike rentals. Describe noteworthy features of the distribution.

Part 2: Simple linear regression

We first investigate whether outdoor temperature has an impact on the number of bike rentals on a given day.

Question Set 2:

1. Create a scatter plot of `Count` versus `Temperature`. Does this relationship appear linear?
2. What is the correlation between these two variables?
3. Run a linear regression of `Count` on the predictor variable `Temperature`. What is the coefficient for `Temperature`? Interpret this coefficient.
4. Is temperature a statistically significant predictor of the mean bike rental in DC on a given day?
5. What is the R^2 of this regression? How do you interpret this value?

Part 3: Are hot days different from cool days?

Here we investigate whether the relationship between rentals and temperature is different for hot days compared to cool days. In particular, we can examine the effect of temperature on counts separately for days above and below 75°F.

Question Set 3:

1. Separate the data into two data frames, one consisting of days with temperature less than or equal to 75°F ("cool" days), and the other consisting of days with temperatures greater than 75°F ("hot" days). What is the number of hot days? What is the number of cool days?
2. Run a linear regression of `Count` on `Temperature` for cool days. Interpret the results.
3. Is the relationship different for hot days? Interpret the results of a linear regression derived from the hot days data frame.
4. What are the R^2 values from each regression? How do they compare to the R^2 value from Part 2?
5. Given the results from each data frame separately, do you trust the results from Part 2?

Part 4: Predicting daily rentals

Suppose we are interested in predicting the number of bike rentals for particular days depending on the temperature.

Question Set 4:

1. Based on the linear regression constructed in Part 2, estimate the number of bike rentals when the temperature is 68°F. Determine a 95% prediction interval. How do you interpret this interval?
2. Based on the results from Part 3, estimate the number of bike rentals when the temperature is 68°F. Determine a 95% prediction interval.
3. Are the prediction intervals consistent with each other? Which prediction are you more inclined to trust?
4. Determine prediction intervals from the models in Parts 2 and 3 when the temperature is 105°F. What do you make of the resulting intervals?

Part 5: Covering more ground

In addition to temperature, other factors such as windspeed, humidity, and month may play a role in predicting bike rentals.

Question Set 5:

1. Construct pairwise scatter plots among counts, temperature, humidity, wind speed and month using the entire dataset. What variables are linearly related to bike rental counts? Are any variables nonlinearly related to bike rental counts?
2. Would you be concerned about running a multiple linear regression including `Month` as a predictor variable?
3. Run a multiple linear regression model predicting daily rentals as a function of `Temperature`, `Humidity`, `Windspeed` and `Month`. Determine 95% confidence intervals for the coefficients of this multiple regression. Based on the intervals, which predictors are statistically significant?
4. Interpret the estimated effect of `Temperature`. How does this compare to the simple linear regression result in Part 2?
5. What is the R^2 for this regression? Again, how does this compare to the value obtained in Part 2?
6. Predict the number of bike rentals on a June day with a temperature of 68°F, a wind speed of 12 mph, a humidity of 70%. Also calculate a 95% prediction interval. How useful is this interval?
7. Now predict and calculate a 95% prediction interval for the number of bike rentals on a January day with a temperature of 90°F, a wind speed of 30 mph, a humidity of 2%. What do you make of the results?

Further exposure to linear regression:

Faraway, J. J. (2014). Linear models with R. CRC press.