What is the business problem in need of solving and what are the main questions at hand?

I plan to analyze Capital bike share program data alongside historical User Type and weather data, in order to predict bike rental count for any duration based on the dataset given.

Few questions for the analysis: -

- How does User Type: Registered vs. Casual influence Bike Rental volume?
- How does weather - temperature, precipitation, humidity, and wind speed - affect bike rental count?
- How does bike use vary on each weekday?

Who is your client and why do they care about this problem? In other words, what will your client DO or DECIDE based on your analysis that they wouldn't have otherwise?

The client is Capital Bikeshare System and this research to predict bike rental count will be useful to them in knowing:

- What features in the dataset influence the bike rental count
- When is the demand for bike share program maximum during the day, season, quarter or year.

- Does weather conditions like temperature, humidity, windspeed have any impact on the demand? If yes, then is it to advantage or adverse.
- Are bike users whether Registered or Casual drive Bike Rental Count? If yes, do they have similar influence on Bike Rental Count distribution?

What important fields and information does the data set have?

Important fields in our data include daily observations for casual and registered member riders trip duration and distance, along with weather variables such as average temperature, precipitation, and occuring weather types (i.e. hail, snow, thunder). Exploration of the data shows differing use behavior for casual and member customers based on day of the week, time of day and responses to weather scenarios. Analyzing bike use behavior and correlations to weather scenarios allows for insights related to optimal maintenance scheduling and any potential price restructuring for strategies related to revenue and growth.

What kind of cleaning and wrangling did you need to do?

The Capital bikeshare datasets required little data wrangling other than renaming a few columns based on preference, formatting the date and time columns to match with the weather data.The data was formatted based on hourly weather and riding observations.

Any preliminary exploration you've performed and your initial findings.

1. **Renaming the columns as below:**

    'instant': 'rec_id',

    'dteday':'datetime',

    'holiday': 'is_holiday',

    'workingday': 'is_workingday',

    'weathersit': 'weather_condition',

    'hum': 'humidity',

    'mnth': 'month',

    'cnt': 'total_count',

    'hr': 'hour',

    'yr': 'year'

2. **There were not any missing values to drop or replace. Type casting the attributes as 'datetime' or 'category' shown below**

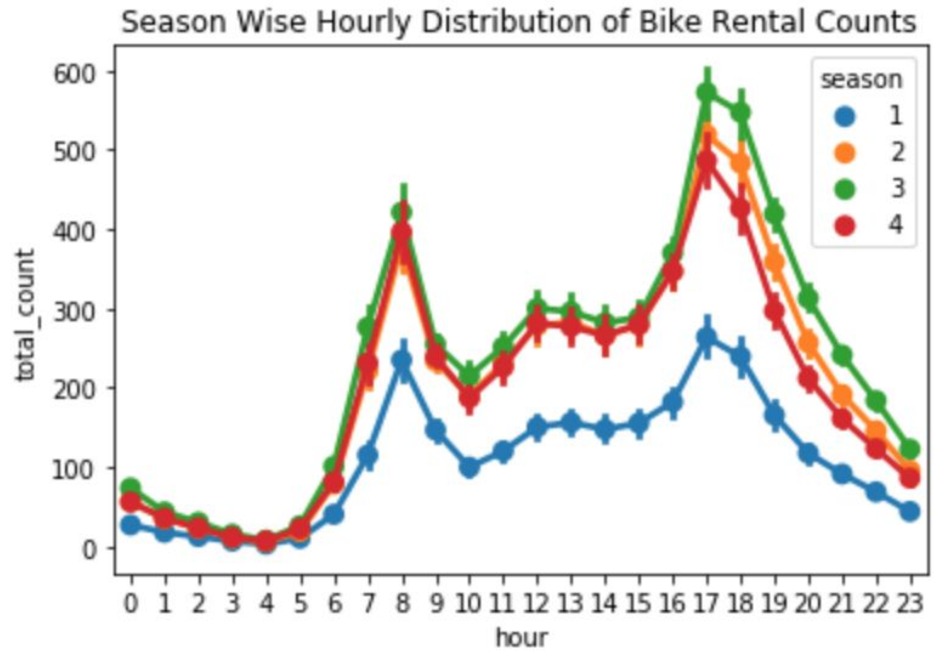
```
stats['datetime'] = pd.to_datetime(stats.datetime)#dae time conversion

# categorical variables

stats['season'] = stats.season.astype('category')

stats['is_holiday'] = stats.is_holiday.astype('category')

stats['weekday'] = stats.weekday.astype('category')

stats['weather_condition'] = stats.weather_condition.astype('category')

stats['is_workingday'] = stats.is_workingday.astype('category')

stats['month'] = stats.month.astype('category')

stats['year'] = stats.year.astype('category')

stats['hour'] = stats.hour.astype('category')
```

3. **Exploratory Data Analysis**

Visualized distribution of bike rental counts across

- Four seasons
- Hour of a day
- User types: Registered or Casual
- Weather conditions: Temperature, wind speed, humidity

Based on EDA similar trend was observed like peaks in rentals occurred during 7-9 AM and another peak between 4-6 PM, Business commute hours mainly.



Similar pattern was observed by Months and Seasons.

4. **Outlier Analysis**

At first look, "count" variable contains lot of outlier data points which skews the distribution towards right (as there are more data points beyond Outer Quartile Limit).But in addition to that, following inferences can also been made from the simple boxplots on Season, Hour, Weekday by Counts.

Spring season has got relatively lower count.The dip in median value in boxplot gives evidence for it. The boxplot with "Hour Of The Day" is quiet interesting.The median value are relatively higher at 7AM - 8AM and 5PM - 6PM. It can be attributed to regular school and office users at that time. Most of the outlier points are mainly contributed from "Working Day" than "Non Working Day".

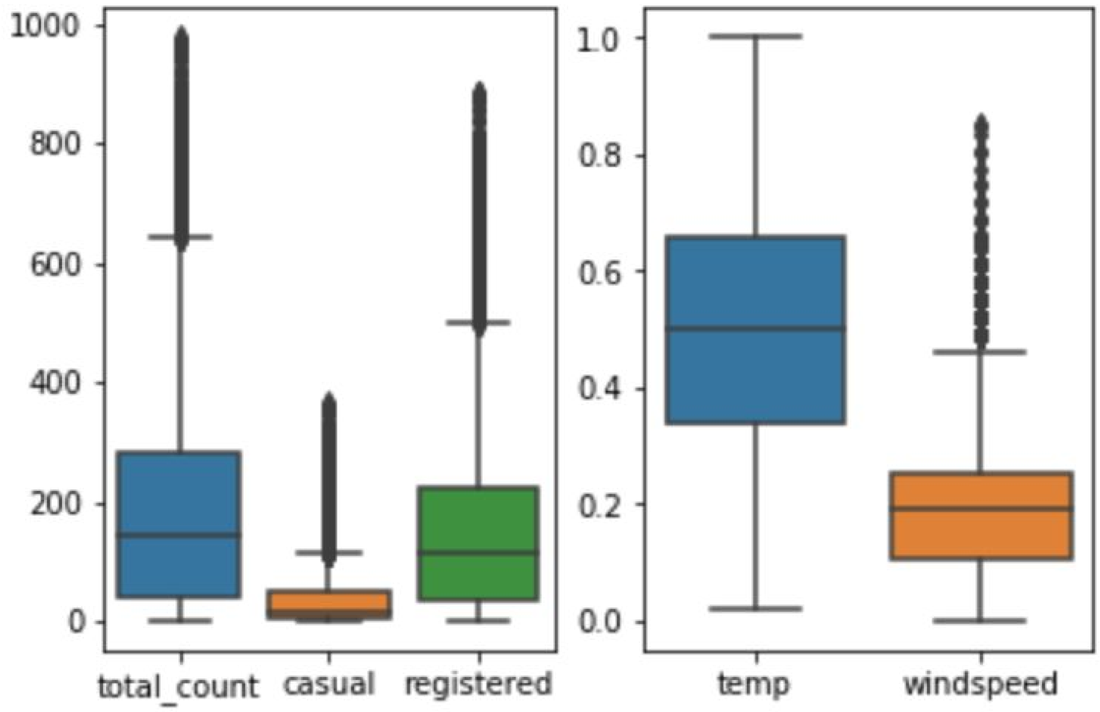
I removed the outliers in the Count column.
One common to understand how a dependent variable is influenced by features (numerical) is to fibd a correlation matrix between them. I plot a correlation plot between "count" and ["temp","atemp","humidity","windspeed"].

temp and humidity features has got positive and negative correlation with count respectively.Although the correlation between them are not very prominent still the count variable has got little dependency on "temp" and "humidity". windspeed is not gonna be really useful numerical feature and it is visible from it correlation value with "count" "atemp" is variable is not taken into since "atemp" and "temp" has got strong correlation with each other. During model building any one of the variable has to be dropped since they will exhibit multicollinearity in the data. "Casual" and "Registered" are also not taken into account since they are leakage variables in nature and need to dropped during model building.

Based on statistical analysis exploring the strength of relationships between bike volume and weather variables such as temperature, precipitation, wind speed, and humidity, we uncovered the following insights:

In comparison to casual riders, registered riders appear more willing to ride in average weather variables of the following manner:

- Lower temperatures
- Higher wind speed
- Higher precipitation
- Higher relative humidty
- Lower heat index



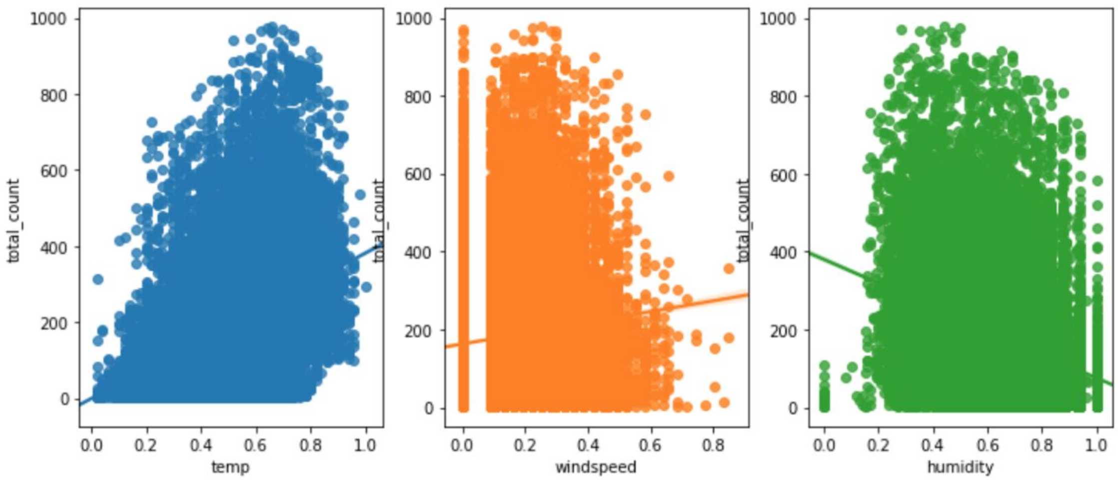
The total, casual & registered type users show sizeable number of outlier values, however casual show lower numbers though. For weather attributes of temperature and wind speed, we see outliers only in the case of windspeed.

Based on these findings, what approach are you going to take? How has your approach changed from what you initially proposed, if applicable?

The next step in our analysis will be to apply inferential statistics: Continuous Distribution, Normal distribution test, Bootstrap sample mean, Confidence Interval to assess Bootstrap replicates, Paired Bootstrap test, Null Hypothesis, Alternate Hypothesis, regression modeling to a majority proportion of the historical data, T-Test, p-value. We will be seeking to determine a best line of fit over the data based on selective application of the variables described earlier. Having chosen the optimal arrangement of our variables, we will test the predictive strength of this model on the remaining portion of our data. This will serve as a secondary check and ensure a minimal amount of model predictions are false positives or negatives. Once this testing phase has validated our model, we can confidently plan to apply the model to future bike observations for the upcoming 2018 season.

One common to understand how a dependent variable is influenced by features (numerical) is to find a correlation matrix between them.

temp and humidity features has got positive and negative correlation with count respectively.Although the correlation between them are not very prominent still the count variable has got little dependency on "temp" and "humidity". windspeed is not gonna be really useful numerical feature and it is visible from it correlation value with "count" "atemp" is variable is not taken into since "atemp" and "temp" has got strong correlation with each other. During model building any one of the variable has to be dropped since they will exhibit multicollinearity in the data. "Casual" and "Registered" are also not taken into account since they are leakage variables in nature and need to dropped during model building. Regression plot in seaborn is one useful way to depict the relationship between two features. Here we consider "count" vs "temp", "humidity", "windspeed".



total_count
temp
total_count
windspeed
total_count
humidity