Capital Bikeshare Project Data Wrangle Explanation

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 2017 hourly weather dataset was more challenging for various reasons outlined in question/answer format below:

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. **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.