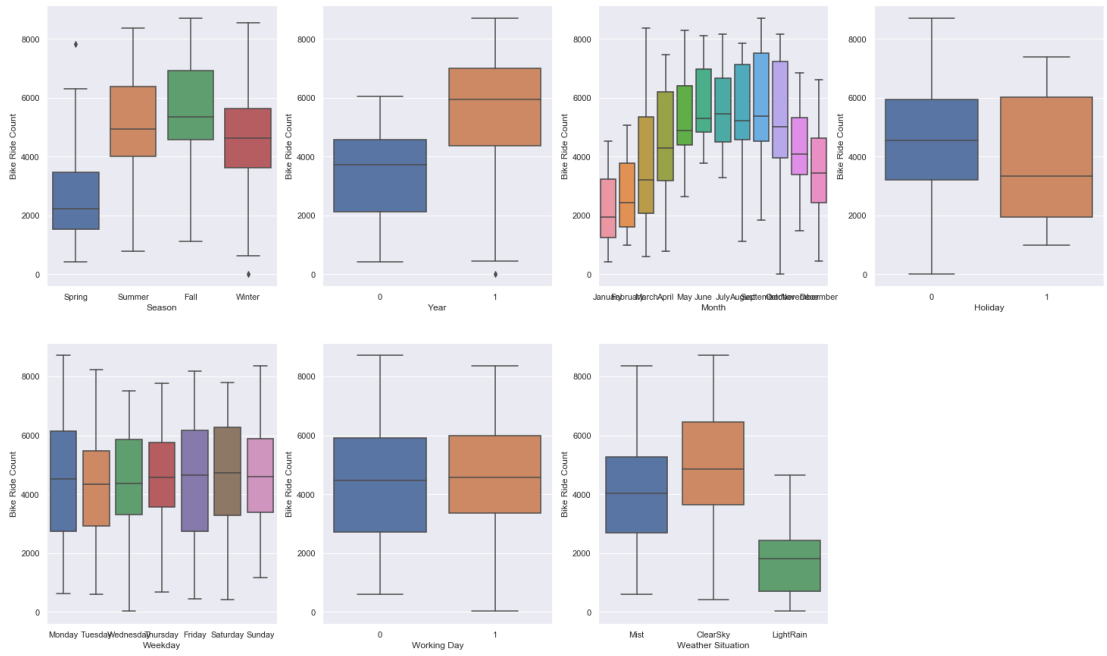
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Boxplot is one of the way to understand the effect of categorical variable. From the following boxplots of categorical variables against target variable, it can be seen that,

- Seasons play a role in deciding the bike demand. The bike ride count is relatively high during Summer and Fall seasons.
- Months also have trend similar to the seasons
- Weather Situation plays important role in bike demand, especially when there is rain lesser is the bike demand.



2. Why is it important to use drop_first=True during dummy variable creation?

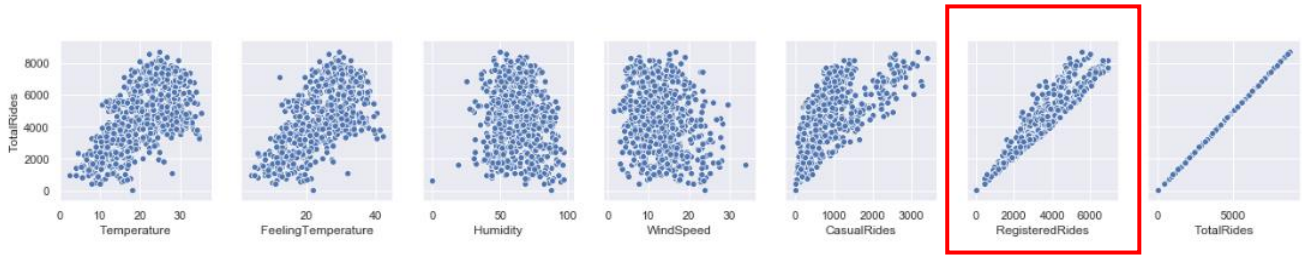
In order to avoid dummy variable trap that arises due to Multicollinearity of independent variables. When encoding using pandas.get_dummies(), multicollinearity will be encountered. This problem can be avoided by using drop_first=True.

It will be difficult to explain a model and to evaluate how much an independent variable affects the target variable, if there exists multicollinearity. Hence multicollinearity should be avoided in regression analysis.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

The pair-plots of the numerical variables present in the dataset with the target variable (TotalRides / cnt, as

given in data dictionary) are shown below:

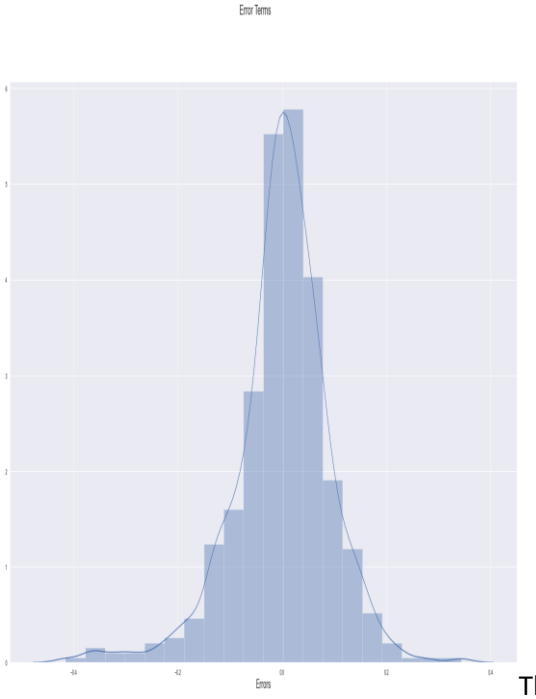


It can be seen that the RegisteredRides (registered, as provided in data dictionary) has high correlation with the target variable.

(However this variable is not used in the model (since TotalRides=CasualRides + RegisteredRides). Among the variables used in the model Temperature has high correlation with the target variable. This is also observed in the pearson correlation matrix.)

4. How did you validate the assumptions of Linear Regression after building the model on the training set?

1. Residual Analysis:
Normality of the residuals is checked



The residual terms are pretty much normally distributed for the training dataset.

2. Mean of the Residuals = -1.9026174971027191e-16 is close to zero.
3. It is also verified that there is no multicollinearity in the independent variables used in the built model. The VIF values are found to be less than 5.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

The final model for the demand of the shared bikes is

TotalRides = **(0.496 × Temperature) + (0.231 × Year) - (0.249 × WeatherSituation3)** - (0.182 × WindSpeed) - (0.127 × Humidity) + (0.089 × Winter) + (0.063 × Monday) + (0.054 × Workingday) + (0.043 × Summer) - (0.074 × Spring)- (0.058 × WeatherSituation2) +0.237

The **features (i) Temperature, (ii) Year, (iii) WeatherSituation3** (i.e. Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) **contributes significantly**. The fit coefficients of Temperature and Year are positive whereas that of WeatherSituation3 is negative.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a statistical model that analyzes the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).
In general, the linear regression equation is of the form,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon$$

where, for $i = n$ observations:

y_i = dependent variable

x_i = expanatory variables

β_0 = y-intercept (constant term)

β_p = slope coefficients for each explanatory variable

ϵ = the model's error term (also known as the residuals)

The following are some assumptions about dataset that is made by Linear Regression model :

Multicollinearity – Linear regression model assumes that there is very little or no multicollinearity in the data. Basically, multicollinearity occurs when the independent variables or features have dependency in them.

Auto-correlation – There is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.

Relationship between variables – Linear regression model assumes that the relationship between response and feature variables must be linear.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when plotted. These peculiarities in the dataset that fools the regression model if built, just with the information from summary statistics. Hence, all the important features in the dataset must be visualised before implementing any machine learning algorithm on them which will help to make a good fit model.

3. What is Pearson's R?

Pearson's R or Pearson correlation coefficient is a measure of linear correlation between two variables.
It is given by,

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = Pearson's R

x_i = values of the variable x in the sample

$\bar{x}$ = mean of the variable x

y_i = values of the variable y in the sample

$\bar{y}$ = mean of the variable y

It lies between –1 and 1 that measures the strength and direction of the relationship between two variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling is a method used to normalize the range of independent variables in the dataset.
Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization. If one of the features has a broad range of values, the objective functions will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the objective functions. Another reason why feature scaling is applied is that gradient descent converges much faster with feature scaling than without it.

Normalized Scaling:

i. Bring data into common range such as [0,1]
ii. Normalized scaling are sensitive to outliers

Standardized Scaling:

i. Brings data with zero mean and unit variance
ii. Standardized Scaling are less sensitive to outliers

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

When two independent variables in a regression model are perfectly correlated, the variance inflation factor (VIF) equals infinity. This indicates a strong linear relationship between the two variables, resulting in a situation of perfect multicollinearity. In such cases, the coefficient estimates of the regression model become highly unstable, making it difficult to interpret the impact of each independent variable on the dependent variable.

Perfect correlation between two variables results in an R-squared value of 1, which leads to a VIF value of

infinity, calculated as 1/(1-R2). To resolve the issue of perfect multicollinearity, one of the variables causing the issue must be removed from the dataset.

An infinite VIF value suggests that the corresponding variable can be expressed exactly as a linear combination of other variables, each of which also exhibits an infinite VIF value. Identifying and removing one or more of these variables can help to address the problem of multicollinearity and stabilize the coefficient estimates of the regression model.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. The quantiles of the first data set is plotted against the quantiles of the second data set. If the two sets come from a population with the same distribution, the points should fall approximately along 45^{o} reference line.

Q-Q plots (Quantile-Quantile plots) are very useful to determine to verify whether the residuals follow a normal distribution. In regression it is assumed that the error terms are following the normal distribution. Also, these plots can be used to check if the distribution is the same for the training and testing datasets.