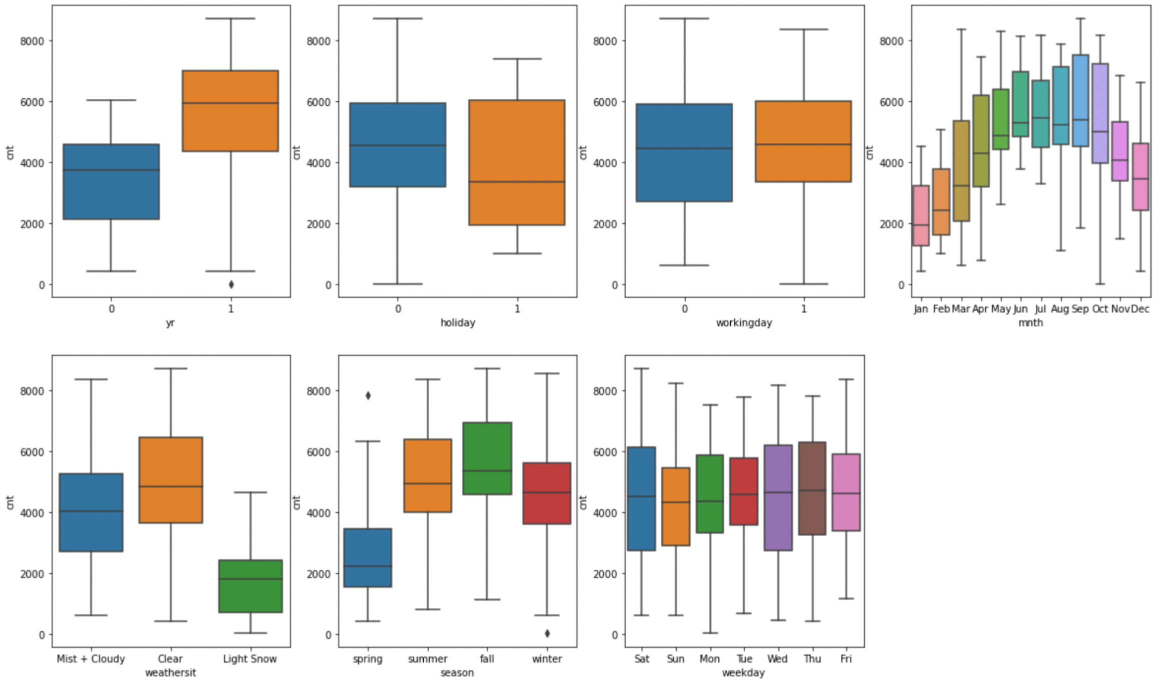
Assignment-based Subjective Questions

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

Some of the categorical variables have a significant effect on the dependable variable.



Season- Summer and Fall are the most desired season for biking.

Year(yr)- We see an increase in bike hire from 2018 to 2019

Month(mnth)- Bike hire is prominently seen around Sep and Oct

Weekdays(Weekday) - Saturday is the peak day for bike hire while Monday is the worst.

Weather Situation(weathersit)- Clearly on Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog weather the bike hire slows down. As the weather situation goes terrible and extreme we see less bike hire.

2. **Why is it important to use drop_first=True during dummy variable creation?**
   drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

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

   Steps to validate assumptions of the linear regression model are:

- **Linear Relationship:** A scatter plot was plotted between one independent and one dependent variable, a straight line passing through the points could be observed.
- **Homoscedasticity:** Variance of error terms was observed and found that the variance of error terms is constant.
- **Absence of Multicollinearity:** Heatmap and VIF were used.
- **Independence of residuals:** Durbin Watson test was conducted.
- **Normality of Errors:** Histogram and QQ plots were used.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   a. Temperature (0.4923)
   b. Year (0.2337)
   c. Weather Sit - Light Snow (-0.3007)

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Linear regression attempts to model the best relationship between variables by fitting a linear equation to observed data. One variable is considered to be a dependent variable, and the others are considered to be an explanatory variables.

The linear regression algorithm consists of the following steps:

1. **Analysis and conversion of variables:** Variables must be converted to the required format, ie, Conversion of Categorical variables. Analysis of variables, to understand

correlation and directionality of the data.

2. **Dividing the model into test and train dataset:** The data set must be divided ideally in 70-30 proportion. This is done to check the predictive capacity of final regression model.

3. **Estimating the model, i.e., fitting the line:** A final model is estimated which has the best representation of maximum points in a linear line. After developing the model, we check the assumptions of linear regression model to determine usefulness of the model.

4. **Evaluating the validity and accuracy of the model:** The model is run of the test dataset to obtain the R_2 and other factors.

2. **Explain the Anscombe's quartet in detail.**

2.

Anscombe's quartet : These are four datasets that have nearly identical simple statistical description, yet have very different distributions and appear very different when graphed.

Anscombe's quartet tells us about the importance of visualizing the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.

The four datasets can be defined as:

1. One which fits the linear regression model well.

2. One could not fit linear regression model on the data quite well as the data is non-linear.

3. One which shows the outliers involved in the dataset which can be handled by linear regression model.

4. One which shows the outliers involved in the dataset which cannot be handled by linear regression model.

Key points from **Anscombe's quartet**:

1. Plotting the data is very important and a good practice before analyzing the data.
2. Outliers should be removed while analyzing the data.

3. Descriptive statistics do not fully depict the data set in its entirety.

3. **What is Pearson’s R?**

The Pearson’s R (aka Pearson’s correlation coefficients) measures the strength between the different variables and the relation with each other. The Pearon’s R returns values between -1 and 1. The interpretation of the coefficients are:

- -1 coefficient indicates strong inversely proportional relationship.
- 0 coefficient indicates no relationship.
- 1 coefficient indicates strong proportional relationship.

$r= n(\Sigma x*y)-(\Sigma x)*(\Sigma y)\surd[n\Sigma x2-(\Sigma x)2]*[n\Sigma y2-(\Sigma y)2]$

Where:

N = the number of pairs of scores
Σxy = the sum of the products of paired scores Σx = the sum of x scores
Σy = the sum of y scores
Σx2 = the sum of squared x scores
Σy2 = the sum of squared y scores

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

The scaling is done in the data preparation step for ML model. The scaling normalizes the varied scaled datatypes to a particular data range.

Two basic types of scaling:

1. Standardization:

   In this the features will be rescaled so that they'll have the properties of a standard normal distribution with
   μ=0 and σ=1
   where μ is the mean (average) and σ is the standard deviation from the mean.

   $Standardization: x = x - mean(x)/sd(x)$

2. Normalization:
   The Min max scaling normalizes the data within the range of 0 and 1. The Min max scaling helps to normalize the outliers as well.

   $MinMaxScaling: x = x - \min(x)/\max(x) - \min(x)$

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

From formula we can say, if the R2 is 1 then the VIF is infinite. The reason for R2 to be 1 is that there is a perfect correlation between 2 independent variables, i.e the independent variables are orthogonal to each other.

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

Q-Q plots are used to find the type of distribution for a random variable whether it be a Gaussian Distribution, Uniform Distribution, Exponential Distribution or even Pareto Distribution, etc.

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.