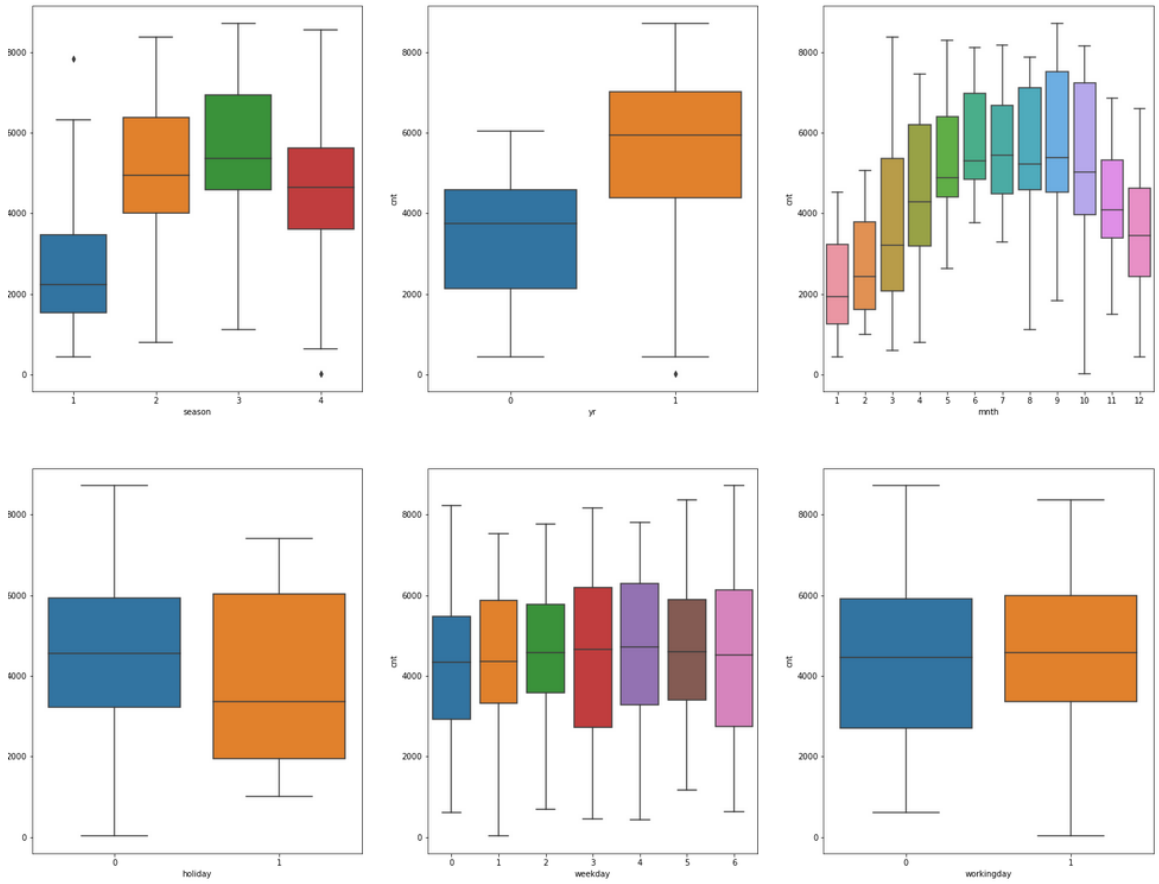
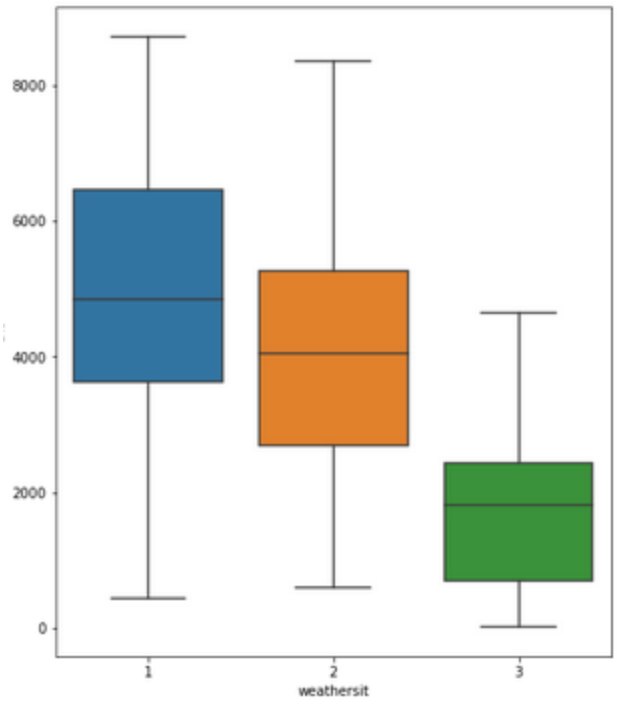
Assignment-based Subjective Questions

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

   1. Summer and Fall are the popular seasons for bike sharing while winter is the least popular season. The months data obviously backs up this conclusion, January and February are the months with least demands.
   2. If we compare the years, the demands increased significantly in the year 2019 compared to year 2018.
   3. The mean value for demand does not change much depending on the weekday. There does not seem to have much impact of whether it is a working day or not on the bike sharing demand overall.
   4. on the other hand clear skies, or partial cloudy atmosphere seem to be a favorite weather condition for bike sharing than any other weather condition. People do not prefer bike ride sharing in extreme weather conditions such as heavy rain, thunderstorm or snow pallets which is evident from the fact that there are absolutely no observations for such weather conditions.





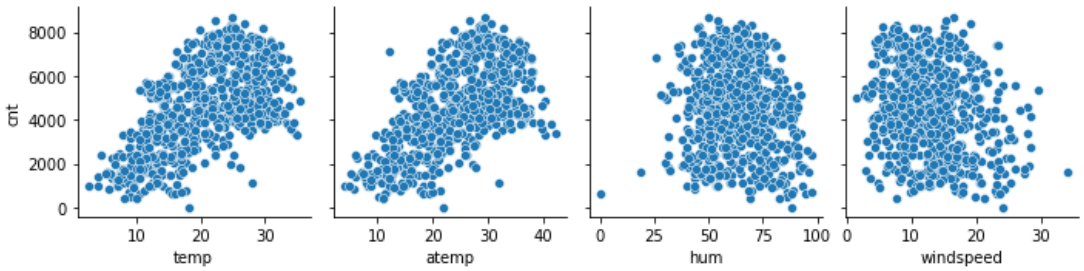
2. Why is it important to use **drop_first=True** during dummy variable creation?

   Machine learning models can only work with numerical variables. In order to convert categorical variables to numerical, we can use pd.get_dummies function which performs one hot encoding.

   To remove multicollinearity in the dataset after one-hot encoding using pd.get_dummies we can drop one of the dummy variables using drop_first = True.

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

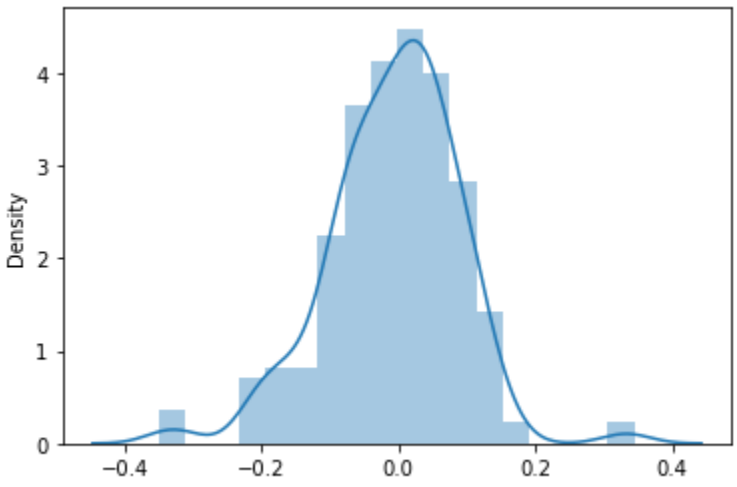
   Looking at the pair plot, temp and atemp variables have the highest correlation with the target variable.

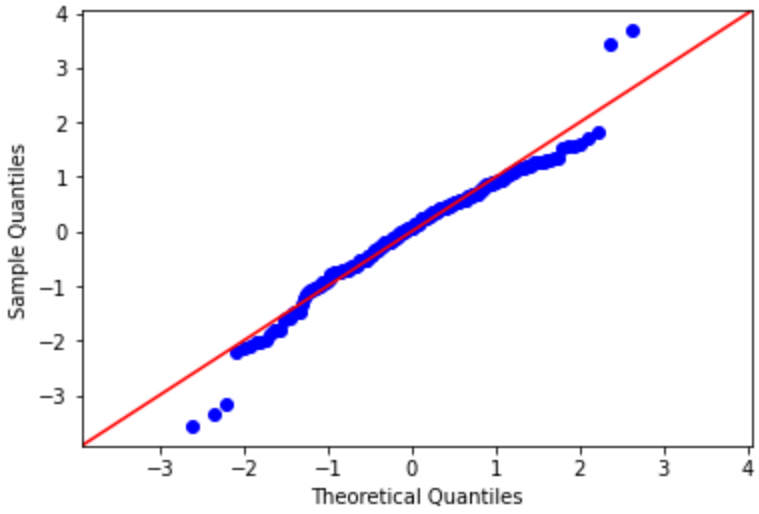


4. How did you validate the assumptions of Linear Regression after building the model on the training set?

   Validated the assumptions by plotting the Residual distribution and Q-Q Plot. The assumption of linear regression is that the error term has normal distribution, based on the distplot and Q-Q plot we can clearly see that it has normal distribution.

   R2 score for the model is also very high 0.840 on train set and 0.78 on the test set.





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

Based on the final model, following are the top 3 features:

From the model we can conclude that following variables are significant in predicting the bike demands:

- atemp - Actual temperature is a good predictor for the bike demand with 0.57 coefficient
- yr - Has coefficient of 0.24, demand in 2019 was more than 2018
- Weathersit - Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds: Has negative coefficient showing, that the demand is low in this season

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a very simple approach for supervised learning, it is useful tool for predicting a quantitative response. Linear regression can be used to answer some questions such as:

1. Is there a relationship between predictor variables and target variable?
2. How strong is the relationship between predictor variables and target variable?
3. How large is the association between each predictor variable and target variable?
4. How accurately can we predict target variable?
5. Is the relationship linear?

Linear regression assumes that there is a linear relationship between one or more predictor variables (denoted by X) and target variable (denoted by Y) and is given by following equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

Where, X_j represents j^{th} predictor and β_j gives the association between the variable and the response. β_0 represents the intercept.

Aim of Linear regression is to estimate the regression coefficients $\beta_0, \beta_1, \beta_2 \ldots \beta_p$.

If we have only one predictor variable then it is called Simple Linear Regression and given by following equation:

$$Y \approx \beta_0 + \beta_1 X$$

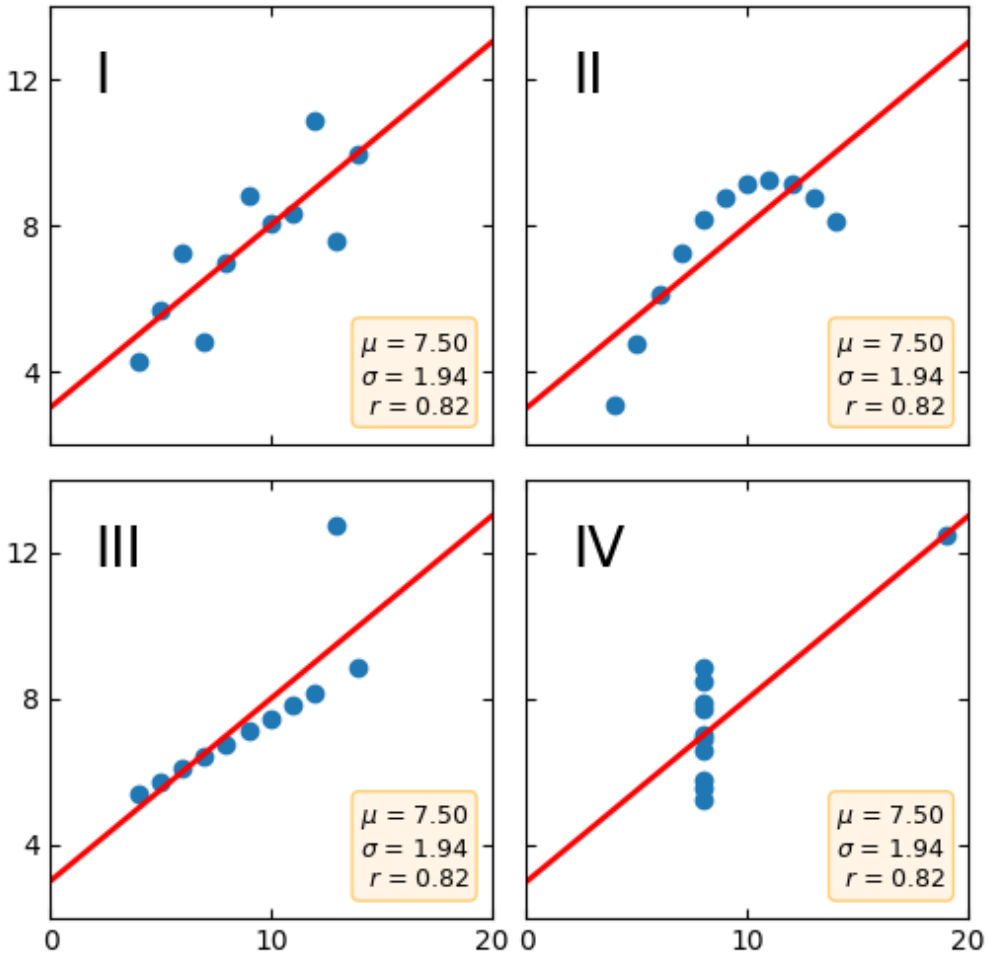
Linear regression has following assumptions:

a. Linear relationship exists between dependent and independent variables.
b. The error terms are independent of each other
c. Independent variables are not correlated
d. The error term has constant variance at every X (Homoscedasticity)
e. Error terms have normal distribution

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet was constructed by statistician Francis Anscombe to demonstrate the importance of visualizing data before analyzing and model building. It comprises of four data sets that have nearly similar descriptive statistics, but have very different distributions and appear very different when visualized.

Below are the plots of four distinct datasets (Reference: Anscombe's quartet — Matplotlib 3.5.0 documentation) which same descriptive statistics and linear regression line but the data is very different when graphed.



3. What is Pearson's R?

Pearson's R is also known as Pearson correlation coefficient is a measure of linear correlation between two sets of data. Correlation coefficient ranges from -1 to 1. The correlation sign is determined by the slope, where:

- Value of +1 implies that all data points lie on a line for which increase in X increases Y also
- Value of -1 implies that increase in X causes decrease in Y

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

Scaling is a technique to standardize the features present in the data in a fixed range. This is an important pre-processing step to handle highly varying values.

If scaling is not performed, then machine learning algorithms will consider the higher values to be more important than lower values irrespective of the unit of the values. This will cause algorithm to give wrong predictions. So, scaling is performed to bring all values to the same magnitude or range.

Normalized scaling

Calculated as
$(X - X_{min}) / (X_{max} - X_{min})$

This scales range to [0, 1] or sometimes [-1, 1]. Normalization is useful when there are no outliers.

Standardized Scaling

Calculated as: (X – mean) / Stddev

Often called as Z-score. It is not bounded to any range and is less affected by outlier.

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

VIF is a measure of multicollinearity in set of regression variables. In order to determine VIF, regression model is fit between independent variables and calculated using following formula:

$VIF = 1 / (1 - R^2)$

If there is perfect correlation between the variables, then R^2 will be very high equal to 1. If R^2 is 1 then VIF will be infinite.

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

Q-Q (Quantile – Quantile) plot is probability plot for comparing two probability distributions by plotting their quantiles against each other. If two distributions are similar, then Q-Q plot will approximately lie on the line y = x. Q-Q plot is also useful in identifying the skewness of the distribution in the data. If the left side of the Q-Q plot is deviating from the central line (x=y line) then the data is left-skewed, while right side deviation in the Q-Q plots denotes that the data is right-skewed.
There are machine learning models that work best with particular kind of distribution, and if we know the distribution of our data, we can select the right model of the right need.

In Linear regression, residual error term should follow normal distribution. Having a normal error term is an assumption and we can verify if it is met using this.