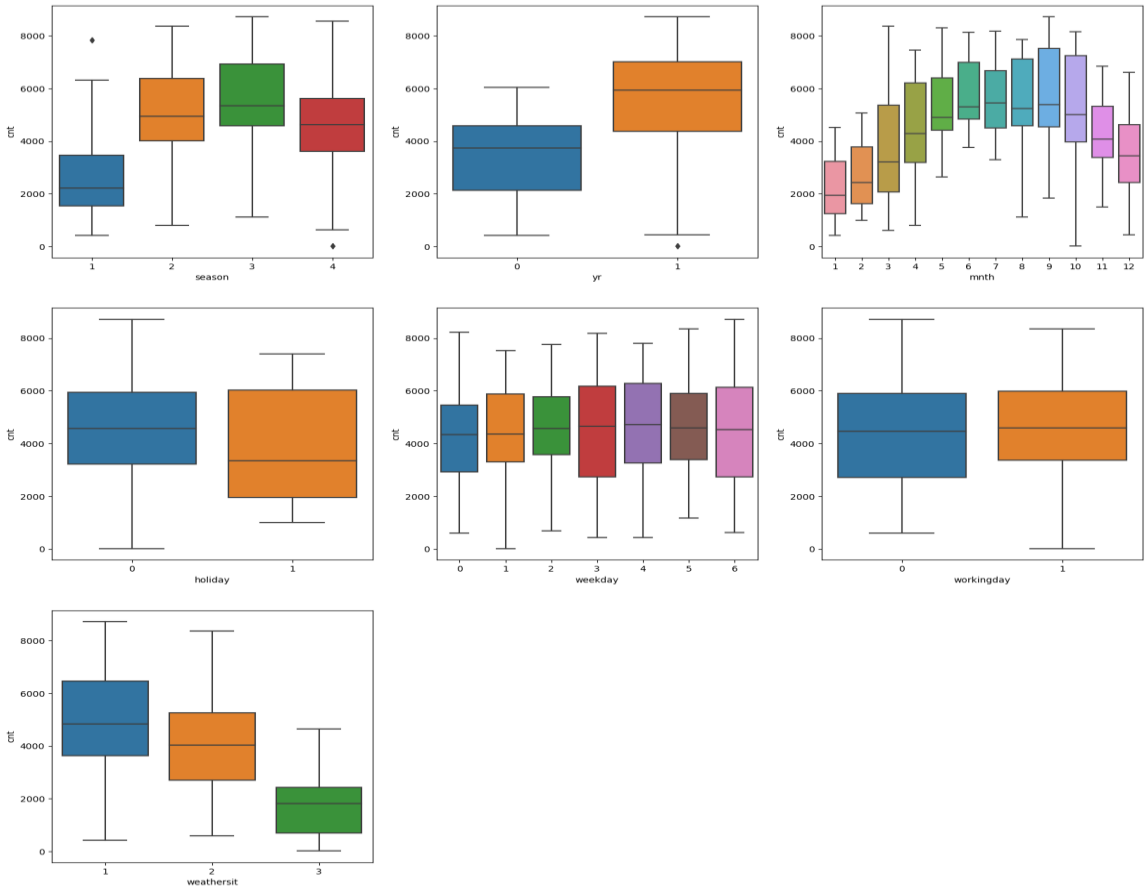
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

   Inference that can be drawn from categorical variables are

   1. Summer, fall, winter has more bike rentals when compared to spring.
   2. 2019 have more rentals when compared to 2018.
   3. Month wise analysis also reveal the same inferences summer and fall months have more bike rentals when compared to winter and spring.
   4. Working days, weekend and holidays bike rentals were at same pace and no significant pattern.
   5. Weather- more rentals when it is 'clear' when compared to other weather conditions. No rentals at bad weather like Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog



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

   Drop_first= True will drop the first dummy variable created because for n levels of a variable only n-1 dummy variables are needed so it will be unnecessary variable can be avoided in the analysis.

   Example in our case study scenario for creating dummy variables for season which has 4 levels i.e., spring, summer, fall, winter.

   Spring -1000

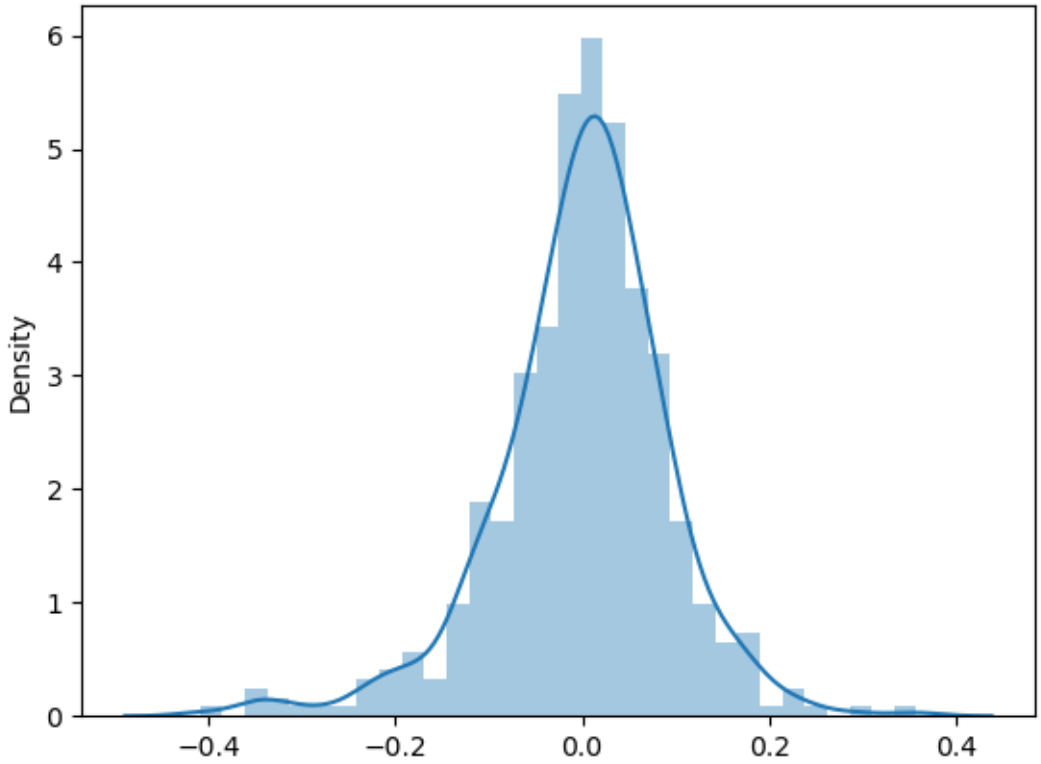
   Summer- 0100

   Fall – 0010

Winter – 0001

Can be created as below Spring can be identified with Spring -000
Summer- 100
Fall – 010
Winter – 001

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
   Looking at the pair-plot cnt (target variable) has highest correlation with temp/atemp.
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   Residue analysis of train data is used to validate the assumptions of Linear Regression after building the model and check if the error terms are also normally distributed around zero mean.



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

   Final Model:
   cnt = 0.217 + 0.230yr - 0.086holiday + 0.496temp - 0.138hum - 0.182windspeed + 0.116summer + 0.074fall + 0.162winter - 0.053cloudy - 0.240light rain

   R-squared for lm4-fouth model - 0.827 R-squared for test data set -0.806

   Based on my model temp, yr and light rain are top 3 features.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Linear Regression is a supervised machine learning algorithm that predicts a dependent variable (target) based on one or more independent variables. The relationship between the dependent and independent variables is assumed to be linear.

Example:

If we have a dataset of houses for a city, where we are given the number of bedrooms for each house and the price of each house. We can create a linear function (a straight line) that models the relationship between these two parameters (bedroom count and price). If a new house enters the dataset, we take the number of bedrooms in that house and read off the price that our line marks at that specific number of bedrooms.

Types of Linear Regression: When the number of independent features is 1, it is known as Simple Linear Regression. In the case of more than one feature, it is known as Multivariate Linear Regression.

1. Assumptions of simple linear regression
    - Linear relationship between X and y.
    - Normal distribution of error terms.
    - Independence of error terms.
    - Constant variance of error terms.
2. Hypothesis testing in linear regression.
    - To determine the significance of beta coefficients.
    - H0:β1=0;HA:β1≠0.
    - T-test on the beta coefficient.
    - t score=^βi/SE(^βi).
3. Building a linear model
    - OLS (Ordinary Least Squares) method in statsmodels to fit a line.
    - Summary statistics
        - F-statistic, R-squared, coefficients and their p-values.
4. Residual Analysis
    - Histogram of the error terms to check normality.
    - Plot of the error terms with X or y to check independence.
5. Predictions
    - Making predictions on the test set using the 'predict()' function.

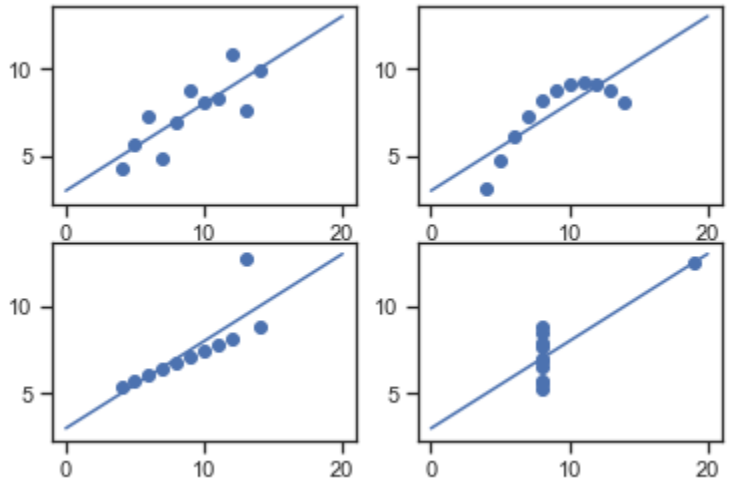
6. Linear Regression using SKLearn

- A second package apart from statsmodels for linear regression.
- A more hassle-free package to just fit a line without any inferences.

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's Quartet is a group of four datasets that were constructed by statistician Francis Anscombe in 1973. The quartet is used to illustrate the importance of visualizing data before analyzing it and building models.

Anscombe's Quartet consists of four datasets, each with 11 x-y pairs. These datasets are nearly identical in simple descriptive statistics, including the mean, variance, correlation, and linear regression line. However, when these datasets are visualized on scatter plots, they appear very different from one another.



The quartet demonstrates the importance of visualizing data before applying various algorithms to build models. It shows that relying solely on summary statistics can be misleading. Visualizing the data can help identify anomalies such as outliers, diversity of the data, linear separability of the data, etc.

3. What is Pearson's R? (3 marks)

A coefficient of correlation is generally applied in statistics to calculate a relationship between two variables. The correlation shows a specific value of the degree of a linear relationship between the X and Y variables, say X and Y. There are various types of correlation coefficients. However, Pearson's correlation (also known as Pearson's R) is the correlation coefficient that is frequently used in linear regression.

Karl Pearson's coefficient of correlation is an extensively used mathematical method in which the numerical representation is applied to measure the level of relation between linearly related variables. The coefficient of correlation is expressed by "r".

$$r = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2}\sqrt{(Y-\overline{Y})^2}}$$

Where, $\overline{X}$ = mean of X variable

$\overline{Y}$ = mean of Y variable

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling is the process to normalize the data within a particular range. It is performed to bring multiple variables in different ranges to a single range. The two most discussed scaling methods are Normalization and Standardization.

The difference between normalized scaling and standardized scaling is:

Normalizing can either mean applying a transformation so that you transformed data is roughly normally distributed, but it can also simply mean putting different variables on a common scale.

Standardizing means subtracting the mean and dividing by the standard deviation.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data came from which type of distribution such as a Normal, exponential or Uniform distribution. It helps to determine if two data sets come from populations with a common distribution.

In linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.