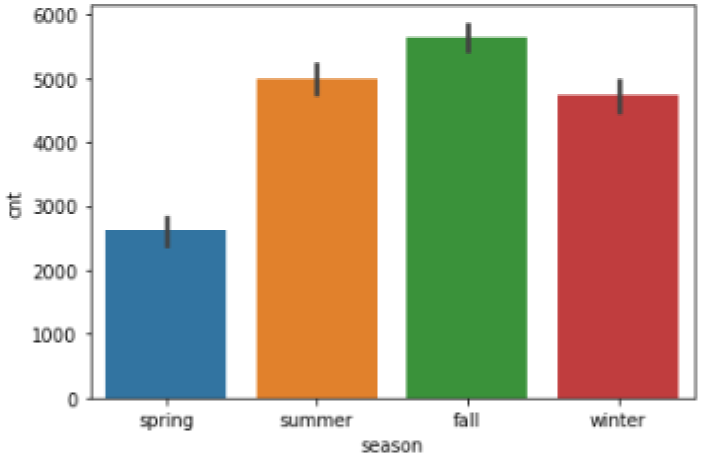
Assignment-based Subjective Questions

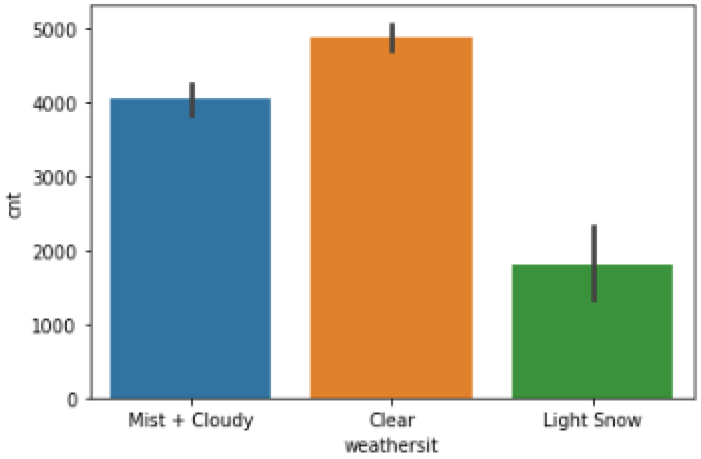
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Ans.: Based on the categorical variable analysis, below are the inferences which I have made.

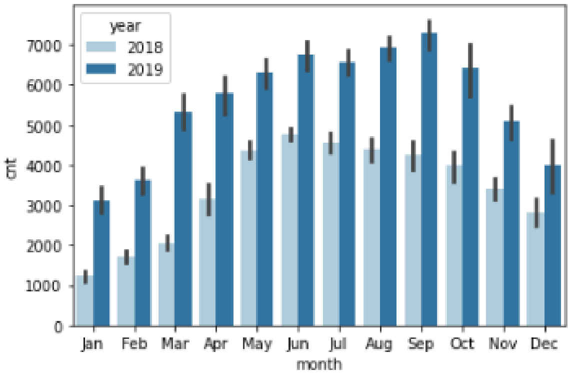
I. Bike rental is more during Fall season



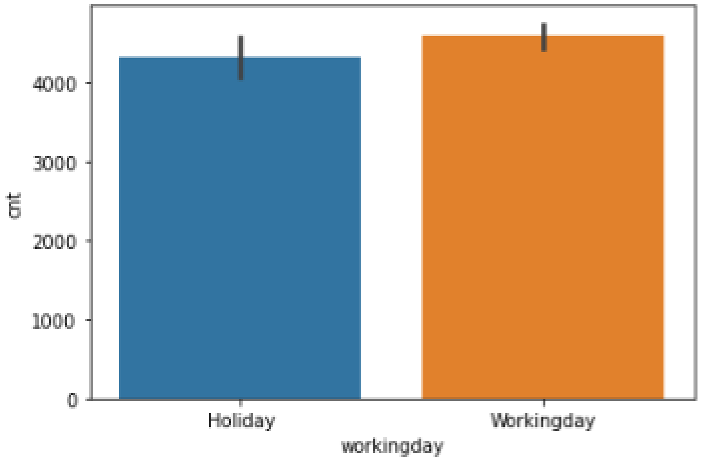
II. It is observed that bike rental demand is more when weather is clear



III. Bike rental demand was increased in 2019 as compared to 2018



IV. Bike rental demand is more on working day compared to holiday.



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

Ans.: By using drop_first=True while creating dummy variables, we can represent n variable with n-1 dummy variables by dropping first level.

e.g: We are having variable House_Type with 3 levels 'Rented', 'Own' and 'Mortgaze', then we can represent the dummy variable as:

House_Type	Own	Mortgaze
Rented	0	0
Own	1	0
Mortgaze	0	1

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans.: Form pair-plot, '**temp'** variable is having strongest relationship with target variable '**cnt**'

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans.: From all below parameters, we can conclude that our model is good

I. P-value of all variables almost equal to 0
II. R-squared and Adj R-squared are 83.7% and 83.50% respectively
III. VIF of all variables is less than 5 except temp and we are not removing it, as temp is having string correlation with 'cnt' which we have already identified earlier in pair plot

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
Ans.: As per my analysis below are the features contributing significantly explaining the demand of the shared bikes.

I. Temperature
II. Season
III. Workingday

General Subjective Questions

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

Ans.: Linear regression is a machine learning algorithm for supervised learning.

It performs tasks to predict the dependent variable based on the input independent variables.

This algorithm used to find out the relationship between the variables.

There are 2 types of linear regression:

I. Simple Linear Regression: It is used to find out the relationship between dependent variable and one independent variable.

II. Multilinear Linear Regression: It is used to find out the relationship between dependent variable and more than one independent variable.

Formula to calculate best fit line in linear regression is:

y= mx+c

y: dependent variable

x: independent variable

m: slope

c: intercept

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

Ans.: Anscombe's Quartet is the modal example to demonstrate the importance of data visualization which was developed by the statistician Francis Anscombe in 1973 to signify both the importance of plotting data before analyzing it with statistical properties.

With the different datasets but having same descriptive properties of that sets like mean, variance standard deviation etc. are same for all data sets, but showing different graphical representation for all date sets and it fools the regression model if built.

Basically, Anscombe's Quartet tells us about the importance of visualizing the data before applying various algorithm out there to build model and this help to identify anomalies in the data set such as like outliers, diversity of the data, linear separability of the data, etc.

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

Ans.: Pearson's correlation coefficient is also known as Pearson's R, which is measure of linear correlation between two set of data.

It is ratio of between the covariance of two variables and product of their standard deviations, thus it is essentially a normalized measurement of the covariance such that result always has value between -1 and 1

For given pairing of random variable (X,Y), formula for Pearson's R is:

PX,Y= cov(X,Y)/sd(X)sd(Y)

where,

- cov is the covariance
- sd (X) is standard deviation of X
- sd (Y) is standard deviation of Y

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

Ans.: It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modeling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

- Min-Max Scaling:

It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

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

- Standardization Scaling:

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

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

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

Ans.: If there is perfect correlation, then VIF = infinity. 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.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Ans.: 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.
If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x.
If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x.
Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions
A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.