

## Assignment-based Subjective Questions

**Question 1.** From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

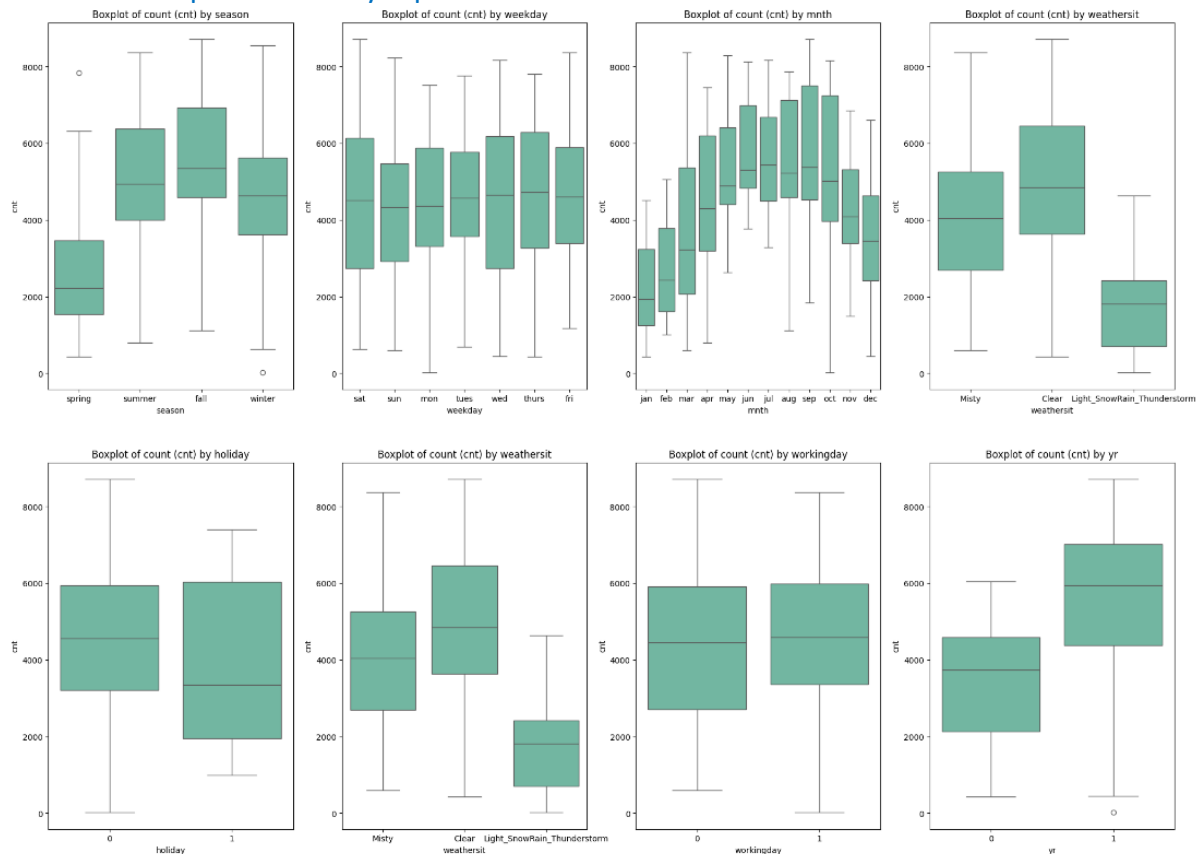
**Total Marks:** 3 marks (Do not edit)

**Answer:** <Your answer for Question 1 goes below this line> (Do not edit)

The Categorical variables such as season, weathersit, yr, mnth, holiday and weekday, based on the analysis below points summarizes what I could infer about their effect on the dependent variables.

- season: Fall Season has the highest demand but Spring Season has the lowest demand for bikes
- weathersit:
  - “Clear, Few clouds” weather Situation has highest demand
  - “Light Snow, Light Rain” weather situation has lowest demand
  - “Mist-Cloudy\_Mist-Broken” weather situation demand is in medium
- yr: The year 2018 had low demand but there was high demand in the year 2019
- mnth: September month has the highest demand followed by October, August and June but the lowest demand is in January.
- holiday: There is high demand in holidays compared to a non- holiday
  - weekday: There is not high variation in demand in the weekdays.

Below the boxplot to visually represent the data.



---

**Question 2.** Why is it important to use **drop\_first=True** during dummy variable creation? (Do not edit)

**Total Marks:** 2 marks (Do not edit)

**Answer:** <Your answer for Question 2 goes below this line> (Do not edit)

drop\_first=True drops the first column during dummy variable creation.

This is important to use because it helps in reducing the extra column created during dummy variable creation. Hence, it reduces the correlation created among dummy variables.

Let's say we have 3 types of values in a categorical column and we want to create dummy variable for that column. If one variable is not furnished and second is semi furnished, then it is obvious the third variable is unfurnished; so we do not need 3rd variable to identify the unfurnished.

Therefore, if we have categorical values of 'n' levels come on then we need to use 'n -1' columns to represent the dummy variable.

---

**Question 3.** Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

**Total Marks:** 1 mark (Do not edit)

**Answer:** <Your answer for Question 3 goes below this line> (Do not edit)

Both 'temp' and 'atemp' have high correlation with cnt variable as can be seen in the pair plot.

---

**Question 4.** How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** <Your answer for Question 4 goes below this line> (Do not edit)

Below are the assumptions for simple linear regression:

- Linear relationship between X and Y
- Error terms are normally distributed (not X, Y)
- Error terms are independent of each other
- Error terms have constant variance (homoscedasticity)

With these assumptions we can go ahead and make inferences about the model which, otherwise, we wouldn't have been able to.

Also note that, there is NO assumption on the distribution of X and Y, just that the error terms have to have a normal distribution.

To meet above assumptions, we ensured to build the model on training set with below values;

- Co-Efficient Values – Non-Zero Co-Efficient indicate that all there is relationship between independent and dependent variables.
- p-Value – p-values are less than 0.05 which indicate they are significant to model.

- VIF – VIF is less than 5, so that there is no multicollinearity between predictor variables.
- F-statistic and Prob(F-statistic) – F-statistic high and low Prob(F-statistic) indicates that overall model fit is significant and not just by chance or only predictor variables are significant

**Question 5.** Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)

**Total Marks:** 2 marks (Do not edit)

**Answer:** <Your answer for Question 5 goes below this line> (Do not edit)

Below are the top 3 features contributing significantly towards the demand of the shared bikes:

1. atemp
2. yr
3. weathersit

## General Subjective Questions

**Question 6.** Explain the linear regression algorithm in detail. (Do not edit)

**Total Marks:** 4 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

Linear regression may be defined as the statistical model that analyses 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).

Mathematically the relationship can be represented with the help of following equation –

$$Y = mX + c$$

Here, Y is the dependent variable we are trying to predict.

X is the independent variable we are using to make predictions.

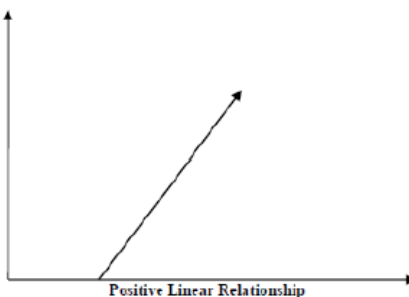
m is the slope of the regression line which represents the effect X has on Y

c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

Furthermore, the linear relationship can be positive or negative in nature as explained below–

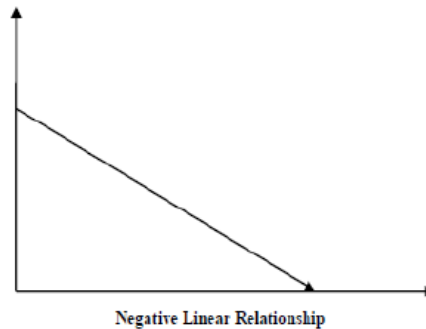
### Positive Linear Relationship:

- A linear relationship will be called positive if both independent and dependent variable increases. It can be understood with the help of following graph –



### Negative Linear Relationship:

- A linear relationship will be called positive if independent increases and dependent variable decreases. It can be understood with the help of following graph –



Linear regression is of the following two types –

1. Simple Linear Regression
2. Multiple Linear Regression

### Assumptions -

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

- a. Multi-collinearity –

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

- b. Auto-correlation –

Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.

- c. Relationship between variables –

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

- d. Normality of error terms –

Error terms should be normally distributed

- e. Homoscedasticity –

There should be no visible pattern in residual values.

---

**Question 7.** Explain the Anscombe's quartet in detail. (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their similar

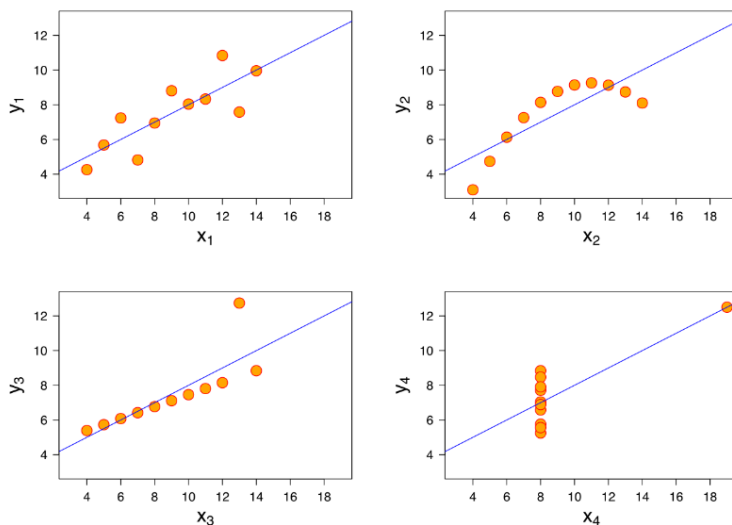
summary statistics.

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

The summary statistics show that the means and the variances were identical for x and y across the groups:

- Mean of x is 9 and mean of y is 7.50 for each dataset.
- Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset
- The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:



- Dataset I appears to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

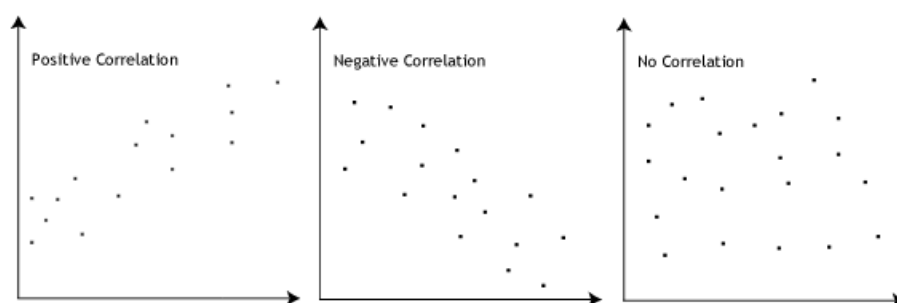
This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

**Question 8.** What is Pearson's R? (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

Pearson's  $r$  is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative. The Pearson correlation coefficient,  $r$ , can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



**Question 9.** What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values. Example: If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

S.NO.	Normalized scaling	Standardized scaling
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

---

**Question 10.** You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables. For example, we would fit the following models to estimate the coefficient of determination  $R_1$  and use this value to estimate the VIF:

$$X_1 = C + \alpha_2 X_2 + \alpha_3 X_3 + \dots$$

$$VIF_1 = 1 / (1 - R_1^2)$$

Next, we fit the model between  $X_2$  and the other independent variables to estimate the coefficient of determination  $R_2$ :

$$X_2 = C + \alpha_1 X_1 + \alpha_3 X_3 + \dots$$

If all the independent variables are orthogonal to each other, then  $VIF = 1.0$ . If there is perfect correlation, then  $VIF = \text{infinity}$ . Orthogonality can refer to the independence of predictors:

**Independent Variables:** If two independent variables are orthogonal, it means they do not share any variance. This can reduce multicollinearity, leading to more stable and interpretable models. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. This would mean that the standard error of this coefficient is inflated by a factor of 2 (square root of variance is the standard deviation). The standard error of the coefficient determines the confidence interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity. A general rule of thumb is that if  $VIF > 10$  then there is multicollinearity. Note that this is a rough rule of thumb, in some cases we might choose to live with high VIF values if it does not affect our model results such as when we are fitting a quadratic or cubic model or depending on the sample size a large value of VIF may not necessarily indicate a poor model.

VIF	Conclusion
1	No Multicollinearity
4-5	Moderate
10 or greater	Severe

---

**Question 11.** What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it

helps to determine if two data sets come from populations with a common distribution. This helps in a scenario of 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.

Few advantages:

- a) It can be used with sample sizes also
- b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios:

If two data sets —

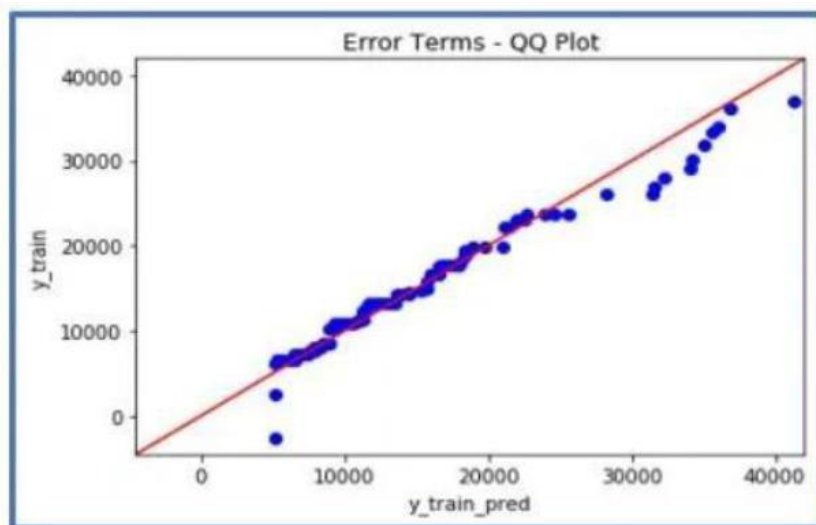
- i. come from populations with a common distribution
- ii. have common location and scale
- iii. have similar distributional shapes
- iv. have similar tail behavior

Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

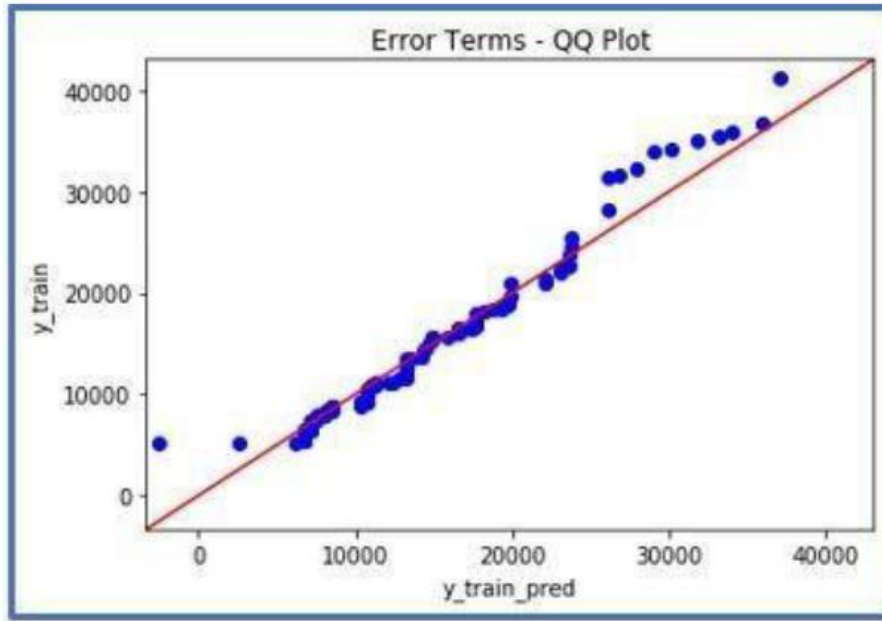
Below are the possible interpretations for two data sets.

- a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis
- b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.



- c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.





d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis

Python: statsmodels.api provide qqplot and qqplot\_2samples to plot Q-Q graph for single and two different data sets respectively.

---