<u>Assignment-based Subjective Questions</u>

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

There were around 7 categorical variables in the data set – season, year, month, workingday, holiday, weekday, weathersit. From the box plot, we can infer the following :

- The bike demand is less in the month of spring when compared with other seasons
- There is a significant increase in the bike demand from year 2018 to 2019..
- From June till September the bike demand keeps increasing after which it slowly declines the rest of the months.
- There is no significant change in bike demand between a working and non-working day.
- The demand seems to be consistent across all weekdays
- There is no significant change in bike demand between a holiday and non-holiday.
- The demand for the bikes are the highest when the weather looks clear/Few clouds/Partly cloudy. Days with Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds are the next most rented days.

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

In case of categorical variable which have a fixed set of values, we can change them into indicator variables.

For eg: if we have a variable say education and the values for the same as graduate and post graduate. Now, we can create 2 indicator variables graduate and post graduate and have values as 0 or 1 for each record. These variables are called as dummy variables. But when we create these dummy variable, instead of 2 variables, we could have created only one with post graduate and have values as 0 or 1. A 0 value would mean the record has value graduate while a 1 indicates it is post graduate.

So, with drop_first=True in the dummy variable creation actually does essentially the same. For variables, with n possible values, it creates n-1 indicator variables rather than n and drops one of the variables. If we don't use, it leads us to dummy variable trap.

The dummy variable trap refers to the scenario where two or more dummy variables representing a categorical variable are highly correlated, which can lead to multicollinearity issues. This can result in an unstable and unreliable model with inaccurate coefficient estimates.

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

The numerical **variables temp (temperature) and atemp (adjusted temperature)** have the highest correlation with the target variable with a value of **0.63**.

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

   The following steps were done in order to validate the assumptions of Linear regression after building the model:

   - Plotted the error terms or residuals and verified that it was normally distributed.
   - Checked the multicollinearity using VIF and ensured that the values are well within the range of less than 5.
   - Linearity was visible among certain variables when we plotted the data set using scatter 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)*

The top 3 features that contributed significantly towards the demands are as below with the coefficient values
Temperature (positively correlated) : 4272.1902
Light snow/rain/thunderstorm (weathersit) (negatively correlated) : -2478.5705
yr (positively correlated) : 2029.4295

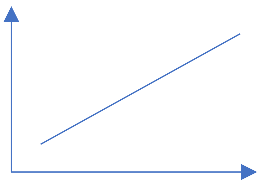
General Subjective Questions

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

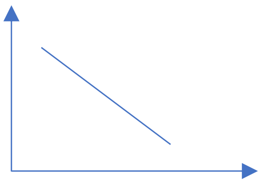
Linear regression is a type of supervised learning algorithm. Linear regression algorithm is a an approach to find the relationship between a dependent variable and one or more independent variables. In linear regression the relationship between the dependent and independent variables cane plotted in a straight line. The slope of the line indicates how much the dependent variables changes for each unit of independent variable. The intercept of the line tells the value of the dependent variable when the independent variable is zero. Mathematically, It can be represented in the form of an equation.

$$Y = \underset{\substack{\downarrow \\ \text{Intercept}}}{\beta_0} + \underset{\substack{\downarrow \\ \text{Slope}}}{\beta_1}X$$

And if we plot a linear regression, the graph would look something as below :



Positive linear relationship



Negative linear relationship

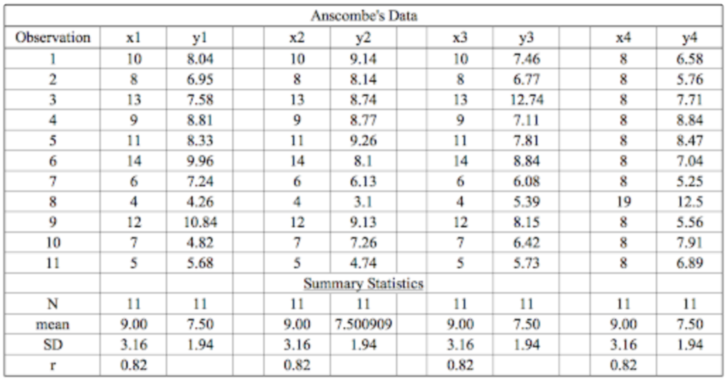
The linear regression can be classified into 2 types – Simple Linear regression and multiple linear regression depending upon the number of independent variables.

Linear regression algorithm aims to find the best fit line given the set of independent and dependent variables. The line can be used to predict the values of the dependant variables for new values of the independent variables. The best fit line is obtained by having the minimum value for sum of the squared distances between the observed and predicted values.

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

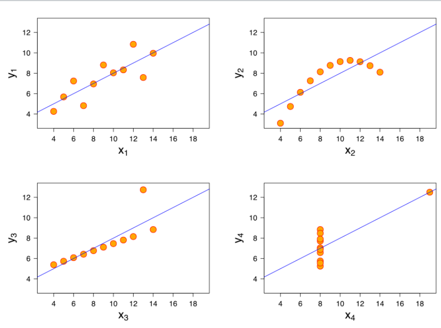
Anscombe's quartet is a set of four datasets that have been created to highlight the importance of data visualization in statistical analysis. Each dataset consists of 11 (x,y) pairs with nearly identical statistical properties, such as mean, variance, and correlation coefficient. However, when plotted, they showed visually different patterns, ranging from linear to non-linear to outlier-driven relationships, and even an absence of relationship.

The quartet was created by Francis Anscombe, a statistician, to show that relying only on summary statistics can be misleading, and that visualizing data is essential for understanding its underlying patterns and relationships. Anscombe's quartet is often used as an example in statistics courses to demonstrate the importance of data visualization and how summary statistics can sometimes fail to capture important characteristics of a dataset.



Anscombe's Data

Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

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 between x and y is 0.816 for each dataset

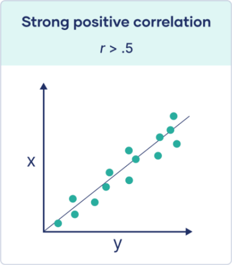
When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:

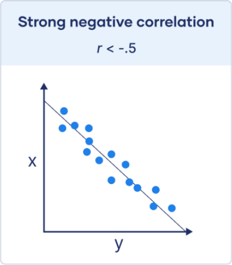
The quartet illustrates that visualizing data is essential for understanding its underlying patterns and relationships, which can lead to more informed decision-making and better statistical analysis. Therefore, statisticians and data scientists must not only rely on summary statistics but also use data visualization techniques to explore and interpret data.

3. What is Pearson’s R? (3 marks)

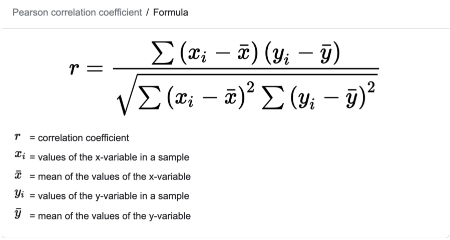
Pearson's R is a statistical measure that quantifies the linear correlation between two continuous variables. It is named after the statistician Karl Pearson, who developed the formula for its calculation in the late 19th century.

Pearson's R takes a value between -1 and +1, where values close to +1 indicate a strong positive correlation, values close to -1 indicate a strong negative correlation, and a value close to zero indicates no correlation or a weak correlation.





The formula for calculating Pearson's R involves dividing the covariance of the two variables by the product of their standard deviations. The covariance measures how the two variables vary together, while the standard deviation measures how much the variables vary individually.



Pearson correlation coefficient / Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

While Pearson's R is widely used in statistical analysis, it is important to note that it only measures the strength and direction of a linear relationship between two variables. It does not capture non-linear relationships or other types of associations that may exist between variables. Therefore, it is essential to use other statistical measures and data visualization techniques to understand the full picture of the relationship between variables.

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

When there are a lot of independent variables in a model, the variables might be on very different scales which will lead a model with very strange coefficients that might be challenging to interpret. So during the data pre-processing we need to standardize these for the following reasons.

1. Ease of interpretation
2. Faster convergence for gradient descent methods.

For example, suppose we have a dataset that contains two features, one measuring weight in pounds and the other measuring height in inches. If we were to apply a machine learning algorithm to this dataset without scaling the features, the weight feature would dominate the analysis because it has much larger values than the height feature. This is the main reason for scaling to be done.

	Normalized Scaling	**Min-Max scaling**
Range	Rescales data to have a value between 0 and 1	Rescales data to have a mean of 0 and a standard deviation of 1
Formula	(x-min) / (max-min)	(x-mean) / std
Outliers	Sensitive to outliers	Less sensitive
Tools	Scikit-Learn provides a transformer called StandardScaler for standardization.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.

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

If the value of the Variance Inflation Factor (VIF) is infinite, it usually indicates a perfect multicollinearity between the predictor variables. If the VIF is 10, this means that the variance of the model coefficient is inflated by a factor of 10 due to the presence of multicollinearity.

If the VIF value is infinite, it indicates that the variable in question is perfectly correlated with the other variables in the model, and it should be removed. . In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. Once the redundant variable have been removed, the VIF values for the remaining predictor variables should be rechecked to ensure that there are no remaining issues with multicollinearity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

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. You can tell the type of distribution using the power of the Q-Q plot just by looking at the plot.

The importance of a Q-Q plot in linear regression lies in its ability to help determine whether the residuals are normally distributed. Residuals are the difference between the observed and predicted values of the dependent variable.In linear regression, it is important that the residuals are normally distributed. By plotting the residuals on the y-axis and the expected quantiles of a normal distribution on the x-axis, a Q-Q plot can reveal whether the residuals are approximately normally distributed. If the residuals follow a straight line, it suggests that the residuals are normally distributed. Conversely, if the plot shows significant deviations from a straight line, it indicates that the residuals are not normally distributed and may require further investigation.

In summary, a Q-Q plot is an important tool in linear regression as it helps to verify the normality assumption of the residuals and provides a visual representation of how well the data fit the model.