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)**

   ➔ From the analysis of the categorical variables from the dataset, we can infer the following effect on the dependent variables:

   - More bookings are most likely to happen in the end days of the week like Friday, Saturday and Sunday.
   - More bookings happen when it is not a holiday.
   - More booking were made on the months between March and November while on peak between June to September
   - More booking happened on the Summer and Fall season
   - More booking we made when the weather was clear and least if the weather was light snow rain
   - More booking happened when the temperature was neither high nor low.

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

   ➔ drop_first=True is important to use, this flag commands to reduce the extra column created during dummy variable creation. Thus it reduces the correlations created among the dummy variables.

   For example, we have 4 types of values in a column and want to create dummy variable for it. If one variable is not Male and Other, then it obviously Female. So we do not need 3rd variable to identify the Demale. To do this, will have to use drop_first=True.

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 among the numerical variables, temp has the highest correlation with the target variable i.e. cnt. Temperature of the location is directly proportional to the total count of bookings in a day.

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

   ➔ After building the model on the training set, the validation of assumptions of linear regression was taken place by:

   - Homoscedasticity: No visible pattern found
   - Normality of error: Was normally distributed
   - Multicollinearity test: All the VIF value was below 5 and thus approved
   - Linearity: Visible
   - Independence of all residuals: Durbin-Watson value of final model lr_6 is 2.111, which signifies there is no autocorrelation.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**

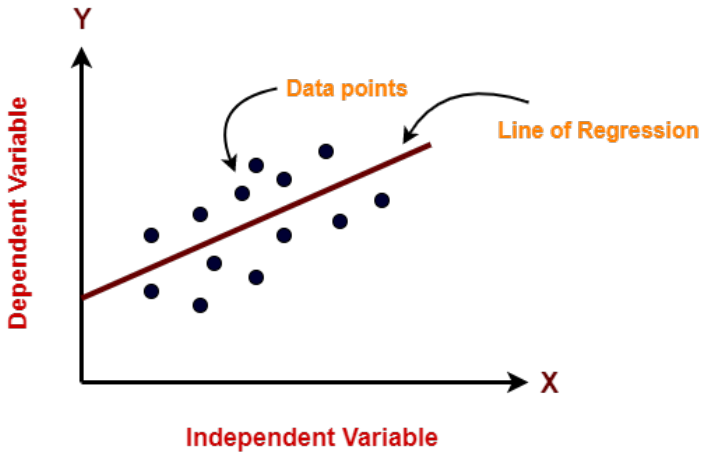
   ➔ Based on the final model, the top 3 features contributing significantly towards explaining the demand of the shared bikes are:

   - Temp: Whenever there is good temperature, bike booking are more likely to increase
   - Windspeed: Whenever there is comforting windspeed, bike booking are more likely to increase
   - Summer: Bike booking tends to increase more in the summer season of the year.

General Subjective Questions

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

➔ Linear regression is a supervised learning-based machine learning algorithm. It carries out a regression task. Based on the independent variables, regression models a goal prediction value. It's usually used to figure out how variables relate to one another and for forecasting. The type of link between dependent and independent variables that each regression model considers, as well as the quantity of independent variables that are employed, are two factors that distinguish different regression models.



Linear regression is used to predict the value of a dependent variable (y) given an independent variable (x). As a result of this regression technique, a linear relationship between x (input) and y (output) is established (output). As a result, it's called Linear Regression.

There are two types of linear regression and they are:

- *Simple Linear Regression*: Simple Linear Regression is a Linear Regression approach that uses a single independent variable to predict the value of a numerical dependent variable.
- *Multiple Linear Regression*: Multiple Linear Regression is a Linear Regression approach that uses more than one independent variable to predict the value of a numerical dependent variable.

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

➔ Anscombe's quartet compromises of four different data sets which have nearly identical simple descriptive statistics but have a different distribution that appear very differently when visualised graphically. Each dataset consists of eleven x and y points.

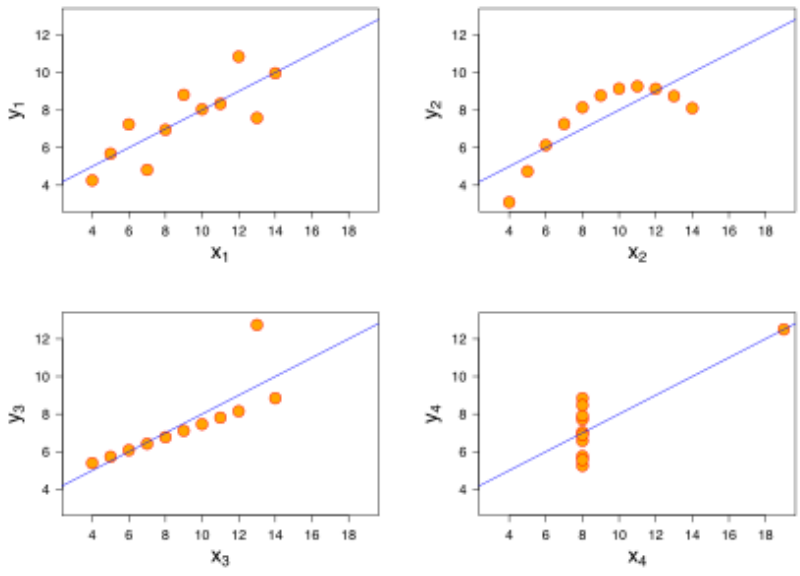
It helps demonstrate the importance of graphical visualization, effect of outliers and other few observation on statistical properties.

For example in the dataset as provided below:

Anscombe's quartet

I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

The graph visualization looks like:



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

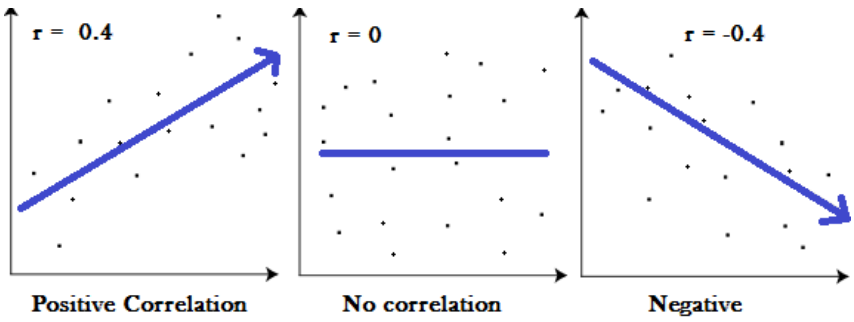
➔ Pearson's R or also known as correlation coefficient is a measurement of linear correlation between two sets of data. It is also the ratio between the covariance of two variable and the product of their standard deviation. It is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

- A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, clothes sizes go up in (almost) perfect correlation with age.
- A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. For example, the amount of battery charge in a mobile decreases in (almost) perfect correlation with mobile usage.
- Zero means that for every increase, there isn't a positive or negative increase. The two just aren't related.

The formula for Pearson's R is as follow

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

The graphical representation of correlation is as follow:



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

➔ Scaling or also feature scaling in ML is a technique of standardizing the feature available in the data in a fixed range. Scaling is performed during pre-processing data in order to handle highly fluctuating values or units. ML algorithm weighs some greater value, higher and also consider smaller value as the values, regardless the unit if feature scaling procedure is not done.

For instance: If an algorithm does not use scaling, it will consider few unit values like 1 hour and 60 minutes, 5000ml and 5L as the same value, and will result in error in the algorithm. Thus we use feature scaling to fix this issue in units.

Difference between normalized scaling and standardized are:

Normalization	Standardization
Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.	Standardization is a scaling technique where the values are centred around the mean with a unit standard deviation.
Normalization is calculated by: $X' = \frac{X - X_{min}}{X_{max} - X_{min}}$	Standardization is calculated by: $X' = \frac{X - \mu}{\sigma}$
It scales between [-1,1]	It does not have a fixed range
It is highly affected by outliers	It is very less affected by outliers

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

➔ If there is a perfect correlation or r^2 is equals to 1, then in that case the value of VIF is infinite. Higher r^2 or a large VIF indicates that there is high correlation between the variables. For example, if the VIF of a column in 2, it means that the model coefficient variance is inflated by a 2 factor because of the presence of multicollinearity.

Perfect correlation between two independent variables results in VIF value to be infinite. Since VIF = 1/(1-r^2), and when r^2 is 1 the equation becomes 1/(1-1) = 1/0 which is infinite

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

➔ A Q-Q plot also known as quantile-quantile plot is a probability plot and is a graphical representation/way of determining if two datasets come from populations with a common distribution. The pattern of points in the plot is used to compare the two distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, measurement and skewness are similar or different in the two distributions.

QQ plot is useful to determine:

- Know whether two populations are of the same distribution or not
- know whether the residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it using Q-Q Plot.
- Know the skewness of distribution