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)**

   Based on analysis done on categorical variables using the boxplot.

   - Fall and Summer season seems to have attracted more booking. And, in each season the booking count has increased drastically from 2018 to 2019.
   - Most of the bookings has been done during the month of August, September, and October after which it started decreasing as we approached the end of year.
   - Good/Clear weather attracted more booking.
   - Wednesday, Thursday, and Saturday had a greater number of bookings as compared to the beginning of the week.
   - Holiday did not make much of difference in the bookings as it showed same 75% on both holidays and non-holidays
   - Booking seemed to be almost equal either on working day or non-working day.
   - 2019 attracted a greater number of bookings from the previous year, which shows good progress in terms of business.

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

   drop_first = True is important to use, as it helps in reducing the extra column created during dummy variable creation.

   If we do not use drop_first = True, then n dummy variables will be created, and these predictors (n dummy variables) are themselves correlated which is known as multicollinearity and it, in turn, leads to Dummy Variable Trap.

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

   The **temp** and **atemp** variable had the highest correlation with the target variable.

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

   I have validated the assumption of Linear Regression Model based on below 5 assumptions -
   - Normality of error terms
     - Error terms should be normally distributed.
   - Multicollinearity check
     - There should be insignificant multicollinearity among variables.
   - Linear relationship validation
     - Linearity should be visible among variables.
   - Homoscedasticity
     - There should be no visible pattern in residual values.
   - Independence of residuals
     - No autocorrelation

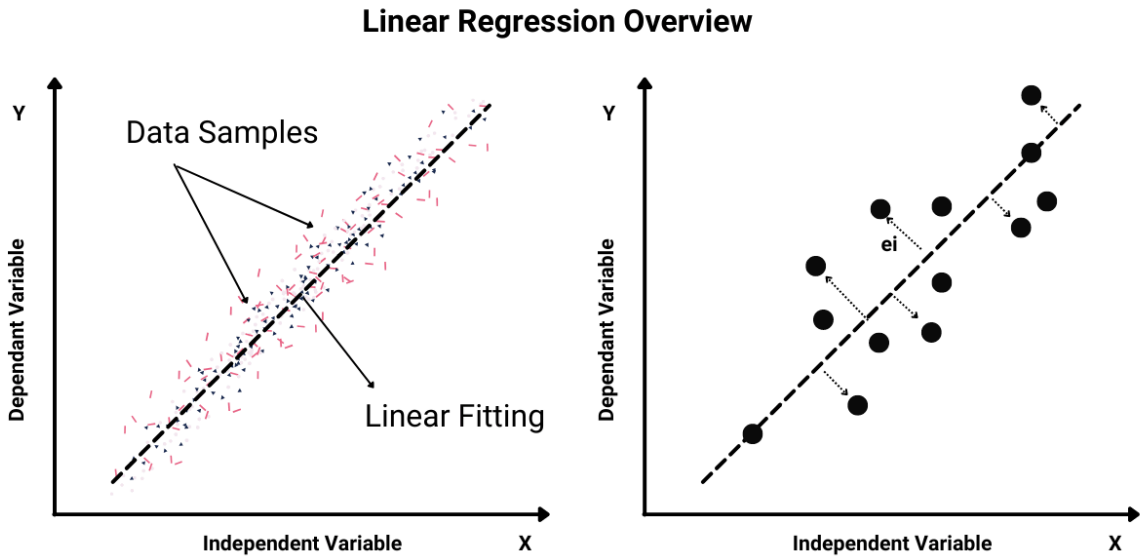
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans: Top 3 features contributing towards demand of bikes were temperature, winter and September.

General Subjective Questions

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

   Linear Regression is a supervised machine learning algorithm that learns a linear relationship between one or more independent variable and a single dependent variable.



   Mathematical representation

   The linearity is defined as the linearly dependent nature of a set of independent features **X** and the dependent quantity **y**. If X = [x1, x2,…,xn] is a set of independent features, and y is a dependent quantity, we try to find a function that maps y to X as,

   $y = \beta0 + \beta1*x1 + \beta2*x2 + ... + \beta p*xp + \epsilon$

   The βi's are the variables that Linear Regression algorithm learns while mapping X to Y using supervised historical data. ϵ is the error due to fitting imperfection, as we cannot assume that all the data samples will perfectly follow the expected function

   There are different types of Linear regression.

   1. Simple Linear Regression ($y = \beta0 + \beta1*x1$
      - If the number of independent features in X is just one, then it becomes a simple linear regression where we try to fit a linear line

        $Y = m*X + c$

        where "m" is the slope, and "c" is the intercept. For example, suppose we want to predict the house price by knowing the area.

   2. Multiple Linear Regression ($y = \beta0 + \beta1*x1 + \beta2*x2$)

- If the number of independent features in X is more than 1, it becomes a multiple linear regression.

Loss Function in Linear Regression

In the image shown above, the imperfection is shown as **ei.** Suppose the actual value for input X1 is Y, and our linear regression model predicted Y' for the same input X1. Then the error (also known as residual) can be calculated as,

ei = (y – y') ^2

The cumulative error for all the samples present in the dataset, it will be called the **Loss function** for our model. The Sum of Squared Residuals is one of the most common loss functions, where we sum up the squares of all the errors.

Ordinary Least Squares (OLS)

The loss function averaged over all the samples is called Cost function for linear regression. Ordinary Least Squares (OLS) method estimates the parameters in a regression model by minimizing the sum of the squared residuals. This method draws a line through the data points that minimizes the sum of the squared differences between the observed values and the corresponding fitted values.

Best - fit line

R^2 score, can help determine the goodness of fit.

R^2 = 1 – sum squared regression (SSR) total sum of squares (SST)
R^2= 1 – ∑ (Yi – Yi) ^2 / ∑ (Yi – $\bar{Y}$) ^2

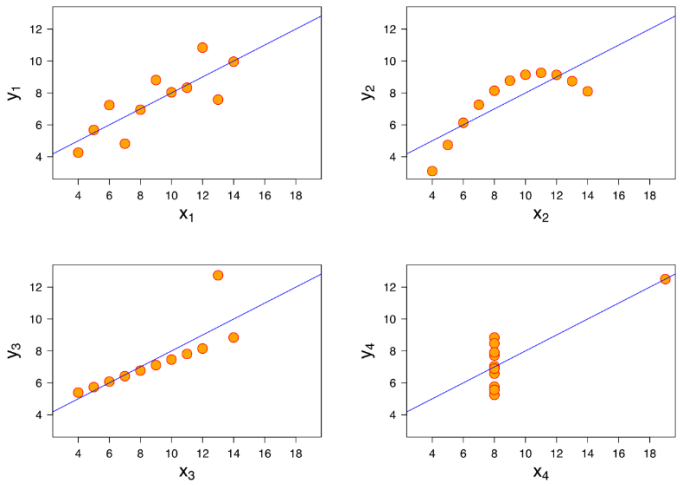
2. Explain the Anscombe's quartet in detail? (3 marks)

   Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. All these datasets share the same descriptive statistics. But we see that this is not the case when they are plotted. Each graph looks very different 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 3 and variance of y is 2 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:



- 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.

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

The Pearson correlation coefficient (*r*) is the most widely used correlation coefficient and is known by many names:

- Pearson's *r*
- Bivariate correlation
- Pearson product-moment correlation coefficient (PPMCC)
- The correlation coefficient

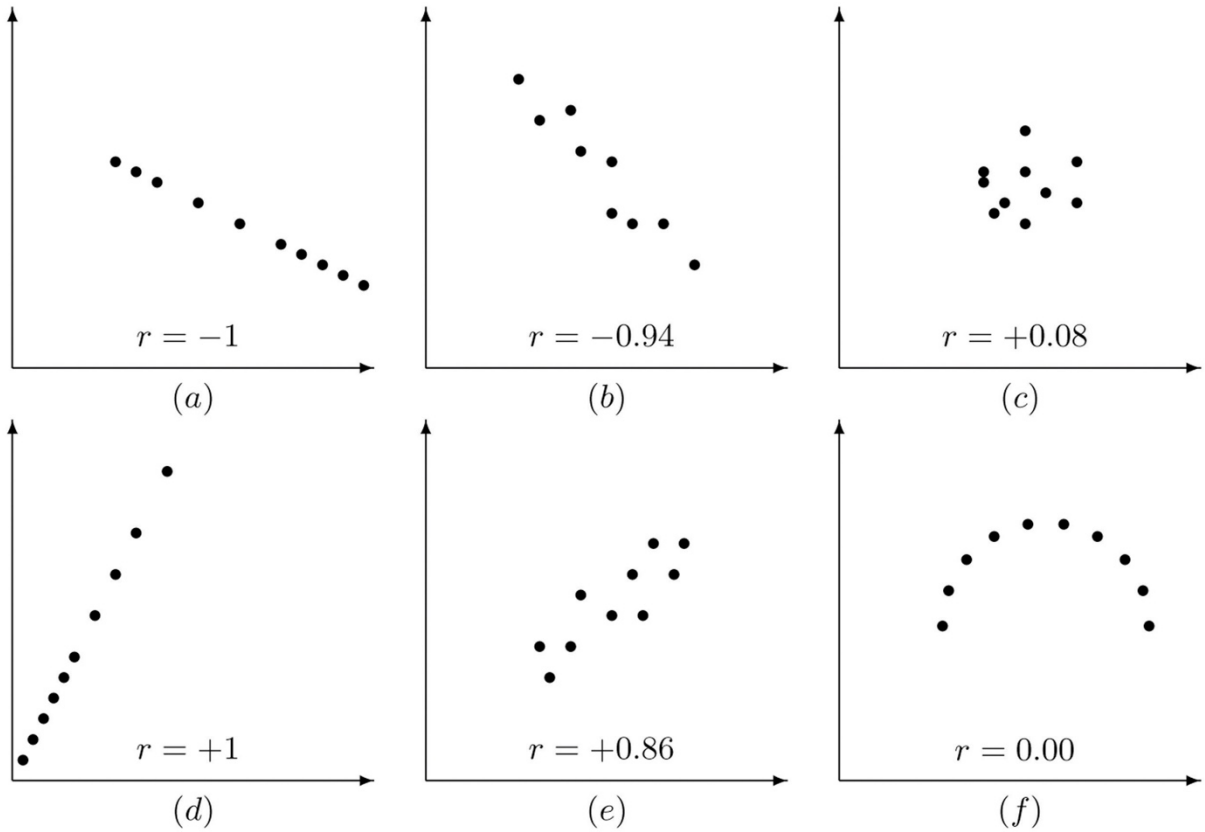
The Pearson correlation coefficient is a descriptive statistic, meaning that it summarises the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables. This table below shows the general rule of thumb:

Pearson correlation coefficient (*r*) value	**Strength**	**Direction**
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

The Pearson correlation coefficient is also an inferential statistic, meaning that it can be used to test statistical hypotheses. Specifically, we can test whether there is a significant relationship between two variables.

Visualizing the Pearson correlation coefficient

Pearson correlation coefficient (r) is as a measure of how close the observations are to a line of best fit. It also tells us whether the slope of the line of best fit is negative or positive. When the slope is negative, r is negative. When the slope is positive, r is positive. When r is 1 or –1, all the points fall exactly on the line of best fit:



When to use Pearson correlation coefficient

The Pearson correlation coefficient is a good choice when the following are true:

1. The variables need to be quantitative.
2. The variables are normally distributed, it's not a problem if the variables are a little non-normal.
3. The data have no outliers.
4. The relationship between two variables is linear.

<u>Calculate the Pearson correlation coefficient (r):</u>

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

Where,

r = Pearson Correlation Coefficient

x_i = x variable samples $\quad$ y_i = y variable sample

$\bar{x}$ = mean of values in x variable $\quad$ $\bar{y}$ =mean of values in y variable

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

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.

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

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

   When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this, 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. (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.

Use of Q-Q plot:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence

for the conclusion that the two data sets have come from populations with different distributions.

Importance of Q-Q plot:

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.