Assignment-based Subjective Questions

Q1: **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**
Answer: From my analysis on categorical variables from the dataset, I infer the below effects on the dependent variables:

- Around 32% of the bike booking happened in fall, followed by Summer & Winter with 28% & 26% of total booking respectively.
- More than 10% of the bike booking happened in the months of May to Sep.
- Around 67% bike booking happens during the Clear weather situation whereas Very few bookings (around 1%) happened during Light_RainSnow weather situation.
- Almost each day has around 14% of bike booking demand hence we can conclude that it is evenly distributed.
- Around 70% bike booking happens on the working day.
- Around 97% of bike bookings took place during non-holiday time.
- There has been a significant growth in demand during 2019 compared to 2018.

Q2: Why is it important to use drop_first=True during dummy variable creation?
Answer: When we have a categorical variable with say 'n' levels then we have to create dummy variable with 'n-1' levels to include a categorical variables in our analysis. Hence we use drop_first = True to get 'n-1' levels. 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.

Q3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Answer: temp and atemp has highest correlation with the target variable.

Q4: **How did you validate the assumptions of Linear Regression after building the model on the training set?**
Answer: To validate the assumptions of Linear Regression is met or not, we creates a scatter plot y vs y_pred. If the data points fall on a straight line in the graph, there is a linear relationship between the dependent and the independent variables, and the assumption is met.

Q5: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
Answer: Top 3 features that have significant impact towards explaining the demand of the bike are year, month and workingday.

General Subjective Questions

Q1: Explain the linear regression algorithm in detail.
Answer: Linear regression is a supervised machine learning method in which past data with labels is used for building a model. It provides a linear relationship between an independent variable (predictor or input) and a dependent (output) variable to predict the outcome of future events.

There are two type of regression. If there is one input variable then it is called simple linear regression and if there is more than one input variable then it is called multiple linear regression.

Equation of straight line for simple linear regression is represented by Y = B0 + B1X (Where B0 is intercept and B1 is Slope).
Equation of straight line for multiple linear regression is represented by Y = B0 + B1X + B2X... BnX (Where B0 is intercept and B1, B2... Bn is Slope).

Goal of the linear regression algorithm is to get the best values for B0 and B1 to find the best fit line and the best fit line should have the least errors. To ensure this, a cost function Like RSS (Residual Sum of Square) or MSE (Mean Squared Error) is used which helps to find out the best possible values for B0 and B1, which provides the best fit line for the data points.

Q2: Explain the Anscombe's quartet in detail.

Answer: Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

To understand this, suppose we have below data set . When you will find statistic of these data set then you will notice that all are same.

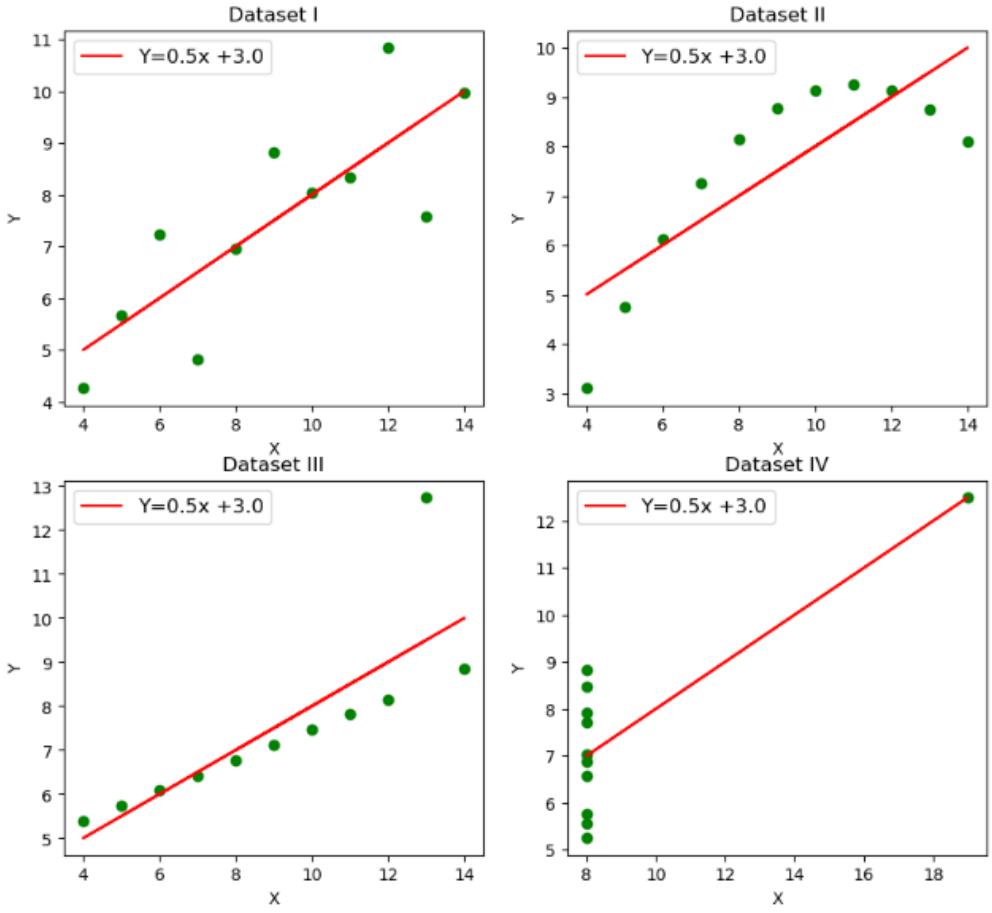
- Mean for x and y for all four datasets.
- Standard deviations for x and y for all four datasets.
- Correlations with their corresponding pair of each datasets.
- Slope and Intercept for each datasets.

	x1	x2	x3	x4	y1	y2	y3	y4
0	10	10	10	8	8.04	9.14	7.46	6.58
1	8	8	8	8	6.95	8.14	6.77	5.76
2	13	13	13	8	7.58	8.74	12.74	7.71
3	9	9	9	8	8.81	8.77	7.11	8.84
4	11	11	11	8	8.33	9.26	7.81	8.47
5	14	14	14	8	9.96	8.10	8.84	7.04
6	6	6	6	8	7.24	6.13	6.08	5.25
7	4	4	4	19	4.26	3.10	5.39	12.50
8	12	12	12	8	10.84	9.13	8.15	5.56
9	7	7	7	8	4.82	7.26	6.42	7.91
10	5	5	5	8	5.68	4.74	5.73	6.89

	I	II	III	IV
Mean_x	9.000000	9.000000	9.000000	9.000000
Variance_x	11.000000	11.000000	11.000000	11.000000
Mean_y	7.500909	7.500909	7.500000	7.500909
Variance_y	4.127269	4.127629	4.122620	4.123249
Correlation	0.816421	0.816237	0.816287	0.816521
Linear Regression slope	0.500091	0.500000	0.499727	0.499909
Linear Regression intercept	3.000091	3.000909	3.002455	3.001727

But when plot scatter plot and linear regression line for each datasets then you will find out that they are appear very different.

- In the first one (top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the second one (top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the third one (bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, the fourth one (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.



Q3: What is Pearson's R?

Answer: Pearson's correlation coefficient, often denoted as Pearson's R or simply "r," is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It ranges from -1 to 1, where:

- -1 indicates a perfect positive linear relationship: as one variable increases, the other also increases proportionally.
- -1 indicates a perfect negative linear relationship: as one variable increases, the other decreases proportionally.
- 0 indicates no linear correlation between the two variables.

The formula for Pearson's correlation coefficient (r) between two variables X and Y is:

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

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

Q4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer: Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step to handle highly varying magnitudes or values or units. It is performed because:

- Larger scale features may dominate the learning process and have an excessive impact on the outcomes. Feature scaling helps to make sure that each feature contributes equally to the learning process.
- The algorithm's performance (like gradient descent) can be enhanced by scaling the features, which can hasten the convergence of the algorithm to the ideal outcome.

There are two major methods to scale the variable that are Standardization and MinMax scaling. Standardization basically brings all of the data into a standard normal distribution with mean zero and standard deviation one. MinMax scaling, on the other hand, brings all of the data in the range of 0 to 1.

Q5: You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer: 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 equal to one which led to VIF equal to infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

Q6: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Answer: A Q-Q plot, or quantile-quantile plot, is a graphical tool used to assess whether a dataset follows a specific theoretical distribution. It compares the quantiles of the observed data against the quantiles of a theoretical distribution, usually the normal distribution. The primary purpose of a Q-Q plot is to visually assess the goodness of fit between the observed data and the expected distribution. If the points in the Q-Q plot fall approximately along a straight line, it suggests that the data follows the theoretical distribution (e.g., normal distribution).Departures from a straight line indicate deviations from the expected distribution.

Q-Q plot is used in linear regression:

- A Q-Q plot can be used in regression models to check some of the assumptions that are required for valid inference. For example, you can use a Q-Q plot to check if the residuals of the model are normally distributed, which is an assumption for many parametric tests and confidence intervals.
- You can also use a Q-Q plot to check if the residuals have a constant variance, which is an assumption for the homoscedasticity of the model. To do this, you need to create a Q-Q plot for the residuals of the model and compare them with the normal distribution.

Checking the normality of residuals is crucial because many statistical tests and confidence intervals in linear regression rely on the assumption that residuals are normally distributed. If the Q-Q plot reveals significant departures from normality, it may indicate the need for further investigation or potential model adjustments.