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)**

   Ans: I have analyzed the categorical columns using the boxplot and bar plot. Below are the few points we can infer from the visualization –

   - Bookings appear to have increased over the Fall season. And, from 2018 to 2019, the number of bookings in each season significantly increased.
   - Most bookings were made in the months of May, June, July, August, September, and October. The trend grew before beginning to decline as the year came to a close.
   - It appears obvious that more bookings were lured by clear weather.
   - Bookings are higher on Thursday, Friday, Saturday, and Sunday than at the beginning of the week.
   - Bookings appear to be less frequent when it's not a holiday, which makes sense given that during holidays, people would want to stay home and enjoy time with their families.
   - Both working days and non-working days looked to have about the same amount of bookings.
   - The amount of reservations for 2019 increased over the prior year, which indicates positive business growth.

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

   Ans: Use of drop_first = True is crucial since it aids in eliminating the excess column produced when a dummy variable is formed. As a result, it reduces the correlations created among dummy variables.

   drop_first: bool, defaulting to False, indicates whether to remove the first level from the k category levels in order to obtain k-1 dummies.

   For example if we want to build a dummy variable for a categorical column that has three different types of data. If one factor is neither A nor B, then it is clear that its C. Thus, we do not require the third variable to identify the C.

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

   Ans: ‘temp’ variable has 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)**

   Ans: Based on the following five assumptions, I have verified the linear regression model's underlying premise. -
   - Error terms should be normally distributed
   - Multicollinearity among variables should be insignificant .
   - Linearity should be visible among variables
   - Homoscedasticity - The residual values should not have any visible pattern.
   - Independence of Residuals

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: Here are the top 3 features that significantly contribute to explaining the demand for shared bikes:
   - temp
   - winter
   - sep

General Subjective Questions

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

   Ans: Linear regression may be defined as the statistical model that analyzes the linear relationship between a dependent variable with a 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, also 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.
   - Negative Linear relationship: A linear relationship will be called negative if independent increases and dependent variable decreases.

Linear regression is of the following two types −

- Simple Linear Regression
- Multiple Linear Regression

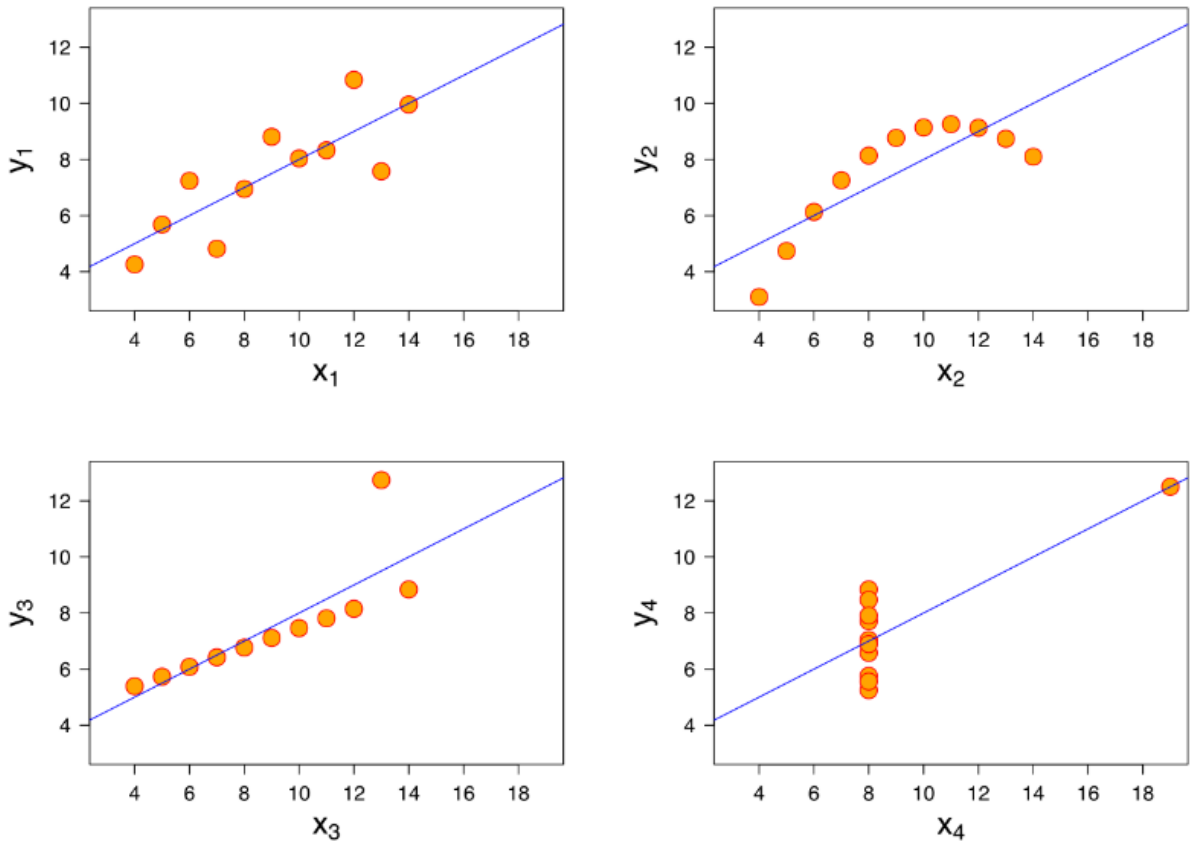
2. **Explain the Anscombe's quartet in detail. (3 marks)**

Ans: Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots. It was constructed in 1973 by statistician Francis Anscombe.

	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 give above show that the means and the variances were identical for x and y across the groups:

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)**

Ans: 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's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)

- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association

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

   Ans: 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.	Normalisation	Standardisation
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.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

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

   Ans: VIF = infinity if there is perfect correlation. A high VIF score denotes a strong correlation between the variables. The presence of multicollinearity causes the variance of the model coefficient to be inflated by a factor of 4 if the VIF is 4.

   VIF displays a complete correlation between two independent variables when its value is infinite. If the correlation is perfect, we have R-squared (R2) = 1, which results in 1/ (1-R2) infinity. The variable that is producing this perfect multicollinearity must be removed from the dataset in order to remedy the problem.

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

   Ans: The quantile-quantile (q-q) plot is a graphical method for assessing if two data sets originate from populations with a common distribution.

   <u>Using the Q-Q plot:</u>

The quantiles of the first data set are plotted against the quantiles of the second dataset in a q-q figure. A quantile is the percentage of points that fall below the specified number. In other words, the 0.3 (or 30%) quantile is the value at which 30% of the data are below it and 70% are above it. Additionally, a 45-degree reference line is plotted. The points should roughly lie along this reference line if the two sets are drawn from a population with the same distribution. The further the two data sets deviate from this reference line, the more evidence there is that they came from populations with different distributions.

Relevance of the Q-Q plot:

It is frequently desirable to determine whether the assumption of a common distribution is supported when there are two data samples. If so, location and scale estimators can combine the two sets of data to derive estimates for the shared position and scale. If two samples do differ, it is also helpful to comprehend the variations. More information about the nature of the difference can be gleaned from the q-q plot than from analytical techniques like the chi-square and Kolmogorov-Smirnov 2-sample tests.