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?**

   Ans: I have done EDA of the categorical variables from the dataset. By doing EDA, I can infer below points. Please check below points.

   a. **Season**: More bookings had been done on "Fall Season". Also booking number had increased drastically from 2018 to 2019.
   b. **Month**: If we observe the "cnt" for all the months, May to Oct booking count was on higher side than rest of the months. Booking trend had increased from Jan to May, then May to Oct it was on average higher side then after Oct, it was again showing decreasing trend. For both the year, this trend was observed.
   c. **Weather**: Most number of bookings had been done when weather was "Clear" then followed by "Misty" and then "Light Snow Rain".
   d. Day: If We observe the Day wise trend, for 2018, most of the days, booking count was near about same with small fluctuation but for 2019, Sunday to Tuesday, increasing trend in booking count had been observed, and from Wednesday to Saturday, booking count was on higher side, and not much difference in booking count was observed for these days.
   e. Holiday: When its holiday a smaller number of bookings had been observed.
   f. Working Day: Higher number of bookings had been observed in working day.

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

   Ans: when we are doing one hot encoding for categorical data, using pandas.get_dummies(), we create dummy variables for each discrete categorical variable for a feature. So it will create n dummies out of n discrete categorical levels for the feature. If we don't use, drop_first=True , these n dummy variables are themselves correlated which in turn leads to dummy variable Trap.

   Using drop_first = True helps in creating the n-1 dummies out of n categorical levels by removing first level.

   Let's say we have 3 types of values in Categorical column, and we want to create dummy variable for that column. If one variable is not A and B, then It is obvious C. So, we do not need 3rd 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?**

   'temp' and 'atemp' both the variable has the highest correlation (0.63) with the target variable, and it is a multicollinearity problem here as 'temp' and 'atemp' both are highly correlated.

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

Ans:

- 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?**

   Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes –

   1. Atemp
   2. Year
   3. Sep

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

   Linear regression may be defined as the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables.

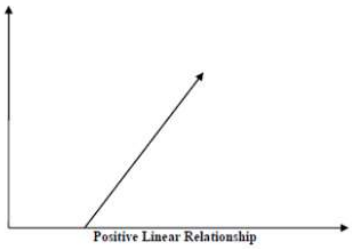
   Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

   Mathematically the relationship can be represented with the help of following equation –

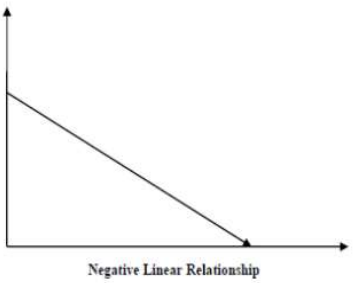
   Y = mX + c Here, Y is the dependent variable we are trying to predict.
   X is the independent variable we are using to make predictions.
   m is the slope of the regression line which represents the effect X has on Y.
   c is a constant, known as the Y-intercept.
   If X = 0, Y would be equal to c.

   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. It can be understood with the help of following graph



   - Negative Linear relationship:
     A linear relationship will be called positive if independent increases and dependent variable decreases.



Linear regression is of the following two types

- Simple Linear Regression
- Multiple Linear Regression

Assumptions - The following are some assumptions about dataset that is made by Linear Regression model

a. Multi-collinearity

Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.

b. Auto-correlation

Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.

c. Relationship between variables

Linear regression model assumes that the relationship between response and feature variables must be linear.

d. Normality of error terms

Error terms should be normally distributed

e. Homoscedasticity

There should be no visible pattern in residual values.

2. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's Quartet was developed by statistician Francis Anscombe.

It comprises four datasets, each containing eleven (x, y) pairs.

The essential thing to note about these datasets is that they share the same descriptive statistics.

But things change completely when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

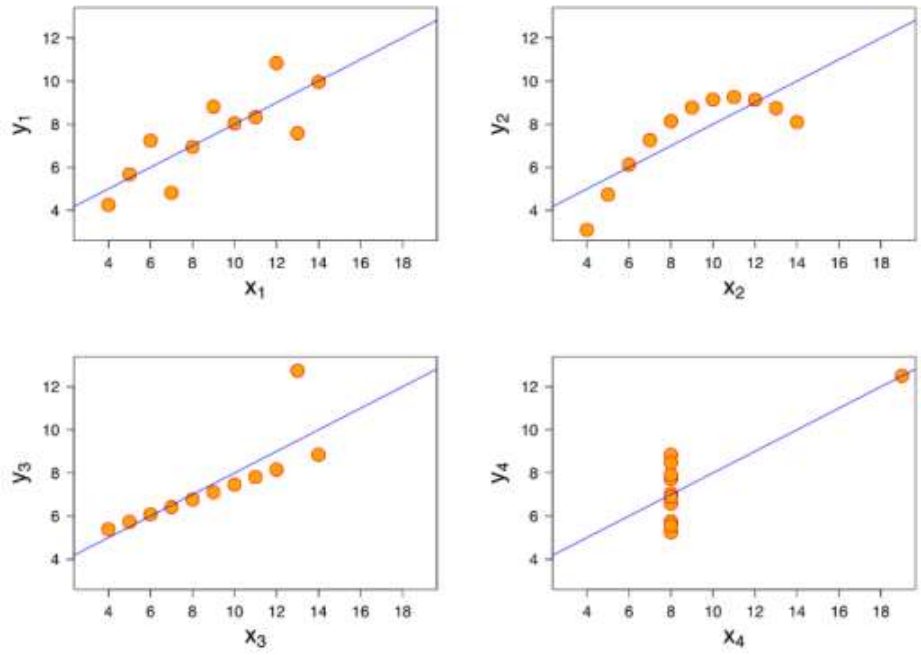
	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	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 11 and variance of y is 4.13 for each dataset

• The correlation coefficient (how strong a relationship is between two variables) 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 appear 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.

Anscombe's quartet is used to show the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also says the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

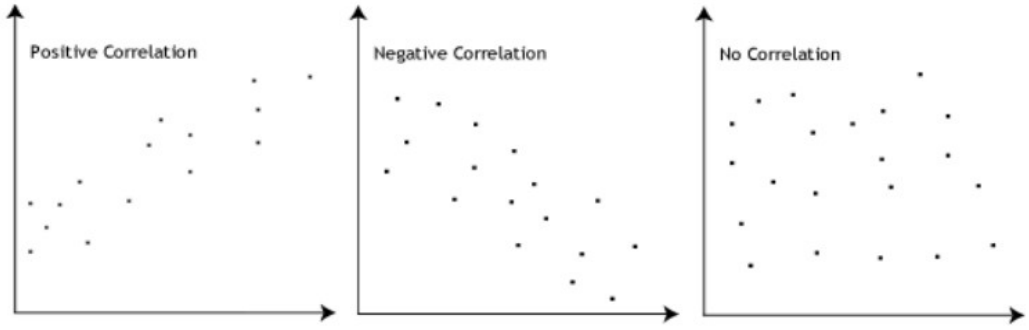
3. **What is Pearson's R? (3 marks)**

   Answer:

   The Pearson coefficient measures the strength of association between two continuous variables on a ratio scale, representing the relationship between them.

   The Pearson coefficient, also known as the Pearson correlation coefficient or product-moment correlation coefficient, is calculated by plotting two variables, X and Y, on a scatter plot.

- A linear relationship is necessary for calculation, and a plot with a straight line indicates stronger association. The Pearson coefficient is represented numerically as a correlation coefficient, ranging from -1 to +1. Positive correlations indicate the same direction, while negative correlations indicate inverse relationships. A zero indicates no correlation. The Pearson coefficient shows correlation, not causation.



4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

   Feature scaling is a data pre-processing technique that transforms features' values to a similar scale, ensuring equal contribution to the model. Common techniques include standardization, normalization, and min-max scaling. Applying feature scaling improves accuracy and effectiveness in machine learning models.

   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.	Normalization	Standardization
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 often called as Scaling Normalization	It is 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)**

   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 (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)

Answer: 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 compares the quantiles of the first data set against the quantiles of the second dataset. Quantiles represent the fraction of points below a given value, while a 45-degree reference line is plotted. If the two sets have the same distribution, the points should fall along this reference line. A larger deviation from this line indicates 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