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)**
Answer:
I have done analysis on categorical columns using the boxplot and Heatmap. There are a few points from the visual analysis –

- Most of the bookings has been made during the month of may, june, july, aug, sep
- and oct.
- Clear weather attracted more bookings than other weather.
- Thu, Fir, Sat, and Sun have more number of bookings as compared to the other days of the week.
- When it’s not a holiday, the number of booking decreases.
- Bike Booking seemed to be almost equal on working days and non-working days also.

2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**
Answer:
drop_first = True is important to us because during dummy variables creation it helps in reducing the extra column creation on that particular data set.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**
Answer:
Column ‘temp’ variable has the highest correlation with the target variable among other numeric variables.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**
Answer:
The validated assumption of the Linear Regression Model is based on some important assumptions -

- o Error terms have to be normally distributed.
- o There has to be insignificant multicollinearity among those variables.
- o some amount of linearity should be visible among data set variables.
- o No auto-correlation is present between variables.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**
Answer:
Below are the top 3 features are - temp , atemp , winter.

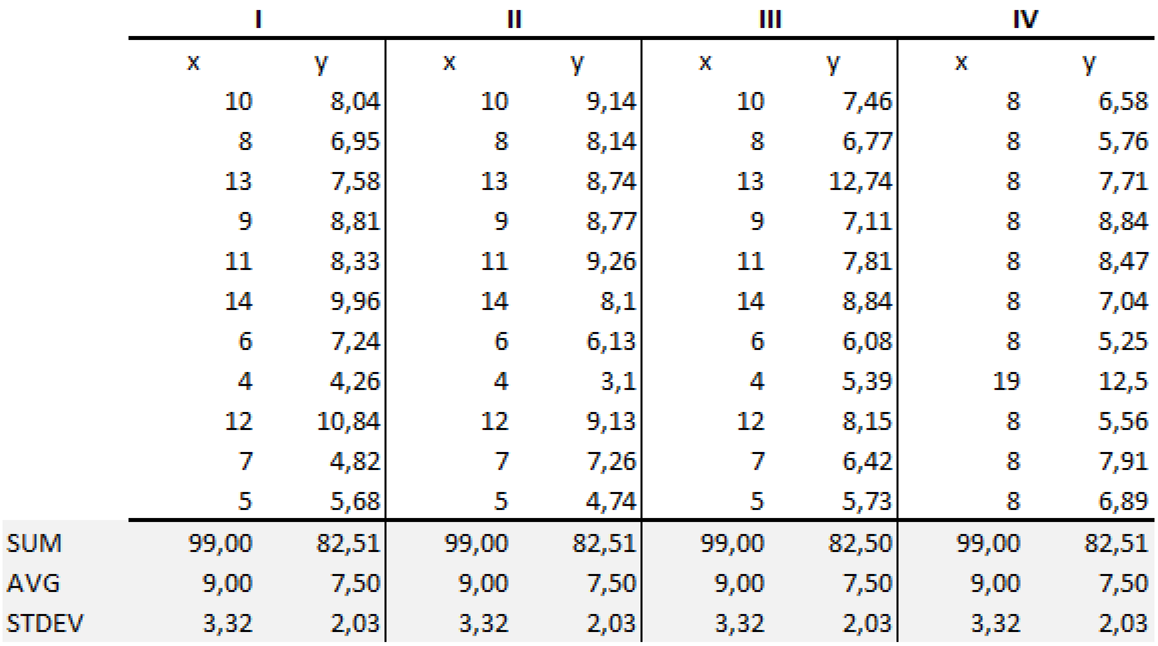
General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**
Answer: Linear regression can be defined as the statistical model that analyses the linear relationship between a dependent variable and one or many given sets of independent variables.
A linear relationship between variables means that when the value of one or more independent variables will be affected (increase or decrease), the value of the dependent variable will also be affected accordingly (increase or decrease).

Mathematically the relationship can be represented with the help of the 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.
o Positive Linear Relationship:
A linear relationship will be called positive if both the independent and dependent variable increases.
o Negative Linear relationship:
A linear relationship will be called positive if independence increases and the dependent variable decreases.
Linear regression is two types −
Simple Linear Regression
Multiple Linear Regression

2. **Explain the Anscombe's quartet in detail. (3 marks)**
Answer:
Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing 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.
The summary statistics show that the means and the variances were identical for x and y across the groups:



	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

- Mean of x is 9 and the mean of y is 7.50 for each dataset.
- Similarly, the variance of x is 11, and the variance of y is 4.13 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 will appear to have clean and well-fitting linear models.
- Dataset II is not distributed normally in the model.
- In Dataset III the distribution is linear, but an outlier throws off the calculated regression.
- 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)**

Answer:

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 to the low values of one variable associated with the high values of the other, the correlation coefficient will be negative.

The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.

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

Answer:

The Scaling feature is a technique to standardize the independent features present in the data between 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, and higher and consider smaller values as the lower values, regardless of the unit of the values.

Normalized scaling and Standardized scaling -

1. The minimum and maximum values of features are used for scaling Mean and standard deviation are used for scaling.
2. Normalized scaling is used when features are of different scales. Standardized scaling 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. Normalized scaling is affected by outliers. On the other hand, Standardized scaling is much less affected by outliers.
5. Scikit-Learn provides a feature transformer called MinMaxScaler for Normalization. Scikit-Learn also provides another feature transformer called StandardScaler for standardization.

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

Answer:

If there is a 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 leads 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 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 falls 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.