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: After analysis of seasons and counts we can say that most bike booking happened in Fall, followed by Summer & Winter respectively.

- After analysis of year and counts we can say that There has been a significant growth in demand during 2019 compared to 2018.

- Most of the bike booking happened during months of April to October. September has most number of bike booking and January has minimum bike booking.

- During clear weather most of the booking happened followed by misty weather and light rain. Very few bookings happened during Light rain and snow due to obvious reasons

- Bike booking on weekdays are evenly distributed.

- Working day has most number of bike bookings.

- Most bike booking happened during non holiday time.Very few booked bike on holiday.

2. **Why is it important to use drop_first=True during dummy variable creation?**
Drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.
Like if we have 3 types of variable (High, medium and low) of categorical column. So if one variable is high other medium then it is obvious that other is low. Sp we do not need third variable low to identify low. So if we have categorical variable of n levels then we need to have n-1 dummy variables.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
Temperature 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?**
Assumption 1:
The Dependent variable and Independent variable must have a linear relationship.
Using scatterplot we can check if there is any linearity b/w dependent variables and target variable.
Assumption 2:
Multicollinearity check - There should be insignificant multicollinearity among variables.
In case of very less variables, one could use heatmap, but that isn't so feasible in case of large number of columns.

Another common way to check would be by calculating VIF (Variance Inflation Factor) values.

If VIF=1, Very Less Multicollinearity

VIF<5, Moderate Multicollinearity

VIF>5 , Extreme Multicollinearity (This is what we have to avoid)

Assumption 3:

Residuals must be normally distributed.

Assumption 4:

Homoscedasticity - There should be no visible pattern in residual values.

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

Temp , year and Light_Rain or snow.

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.

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 –

1. Simple Linear Regression

2. Multiple Linear Regression

Assumptions -

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

• 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 among them.

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

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

• Normality of error terms – Error terms should be normally distributed

• Homoscedasticity –There should be no visible pattern in residual value.

2. Explain the Anscombe's quartet in detail.

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, and I must emphasize 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:
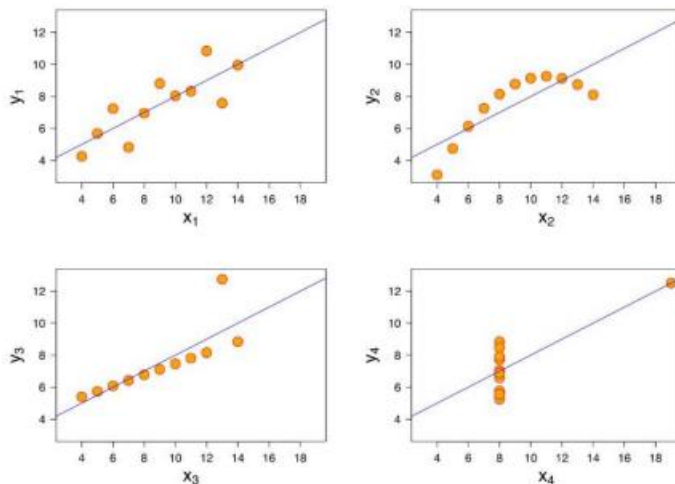

3. What is Pearson's R?

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. Its value ranges between -1 to +1.

r = 1 means the data is perfectly linear with a positive slope

r = -1 means the data is perfectly linear with a negative slope

r = 0 means there is no linear association.

This is shown in the diagram below:
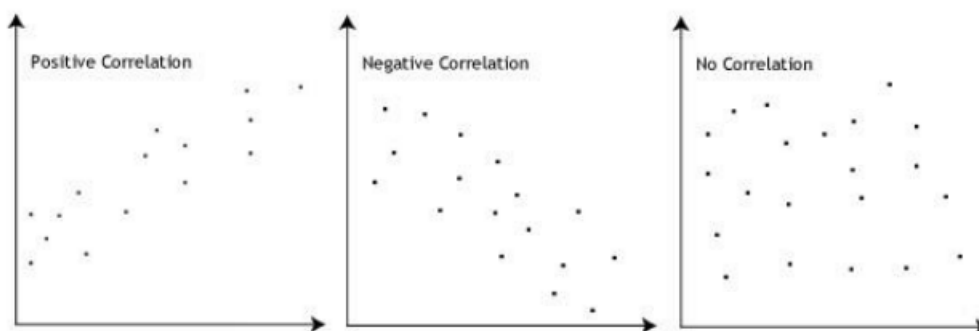
4.What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

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 10000 meter to be greater than 5 km but that's 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.

Normalized scaling

- Minimum and maximum value of features are used for scaling.
- It is used when features are on different scale.
- Scales values between [0, 1] or [-1, 1].
- It is really affected by outliers.
- Scikit-Learn provides a transformer called MinMaxScaler for Normalization.

Standardized scaling

- Mean and standard deviation is used for scaling.
- It is used when we want to ensure zero mean and unit standard deviation.
- It is not bounded by a certain range.
- It is much less affected by outliers.
- 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?

VIF - the variance inflation factor -The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity. (VIF) =1/ (1-R_1^2). 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.

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.