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 analysis on categorical columns using the boxplot and bar plot. Below are the few points we can infer from the visualization –

   - Fall season seems to have attracted more booking.
   - Most of the bookings has been done between the months of May to October.
   - Clear weather attracted more booking which seems obvious.
   - Thu, Fir, Sat have more number of bookings as compared to the other days of week.
   - When it's holiday, booking seems to be less in number which seems reasonable as on holidays, people may want to spend time at home and enjoy with family.
   - Booking seemed to be almost equal either on working day or non-working day.
   - 2019 attracted more number of booking from the previous year, which shows good progress in terms of business

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

   Ans:
   The dummy variable is used to convert a categorical variable to numeric in the form of 0 and 1.
   So for different labels present in categorical variable are treated as dummy columns, for an actual column value putting 1 for that dummy column label and zero for others. So for a categorical column having n labels can be converted to n-1 dummy columns. So we are dropping first dummy column using ***drop_first=True*** *parameter****.***

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

   Ans: The "temp" and "atemp" variables have highest correlation

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

Ans : Linear Regression models are validated based on Linearity, No auto-correlation, Normality **of error, Homoscedasticity, Multicollinearity**

5. ***Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?***
   Ans:
   - ***Temp***
   - ***Year***
   - ***Clear weather (***Light_Rain_Snow + Mist_Cloudy)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans:

Linear regression is a form of predictive modeling technique which tells us the relationship between the dependent (target variable) and independent variables (predictors). Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

If there is a single input variable (x), such linear regression is called simple linear regression.

And if there is more than one input variable, such linear regression is called multiple linear regression.

The linear regression model gives a sloped straight line describing the relationship within the variables. A regression line can be a Positive Linear Relationship or a Negative Linear Relationship. The goal of the linear regression algorithm is to get the best values for β 0 and β 1 to find the best fit line and the best fit line should have the least error. In Linear Regression, RFE or Mean Squared Error (MSE) or cost function is used, which helps to figure out the best possible values for β 0 and β 1, which provides the best fit line for the data points.

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

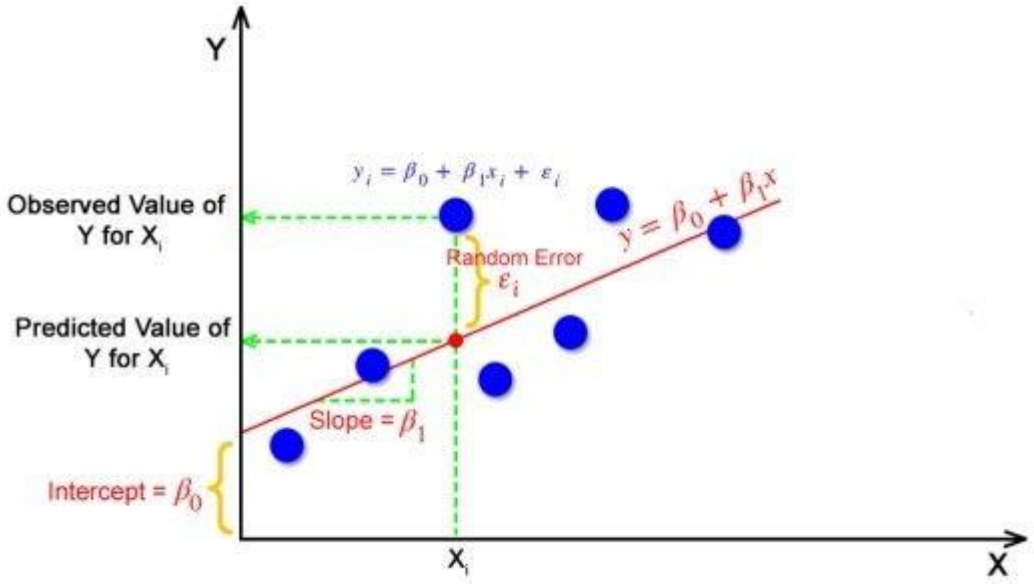
$$Y = \beta 0 + X\beta 1$$

Here, Y is the dependent variable we are trying to predict.

X is the independent variable we are using to make predictions.

β 1 is the slope of the regression line which represents the effect X has on Y

β 0 is a constant, known as the Y-intercept. If X = 0, Y would be equal to β 0.



2. Explain the Anscombe's quartet in detail

Ans:

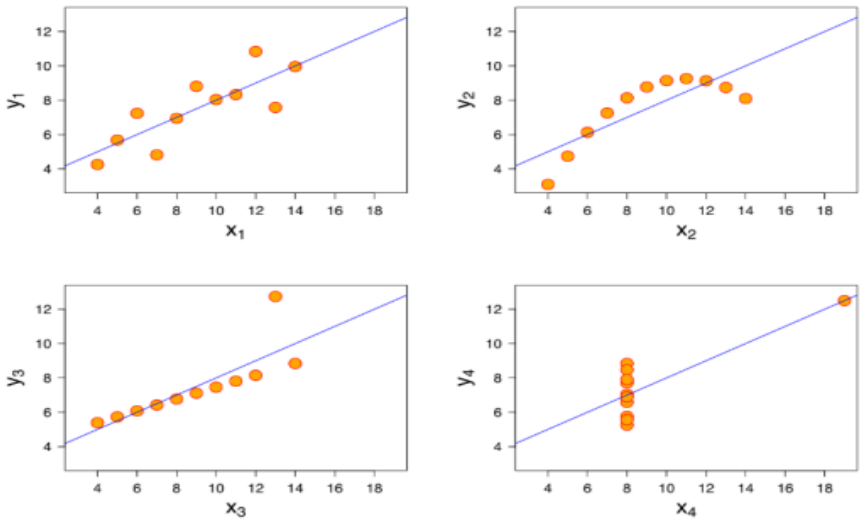
Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when 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:



- 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?

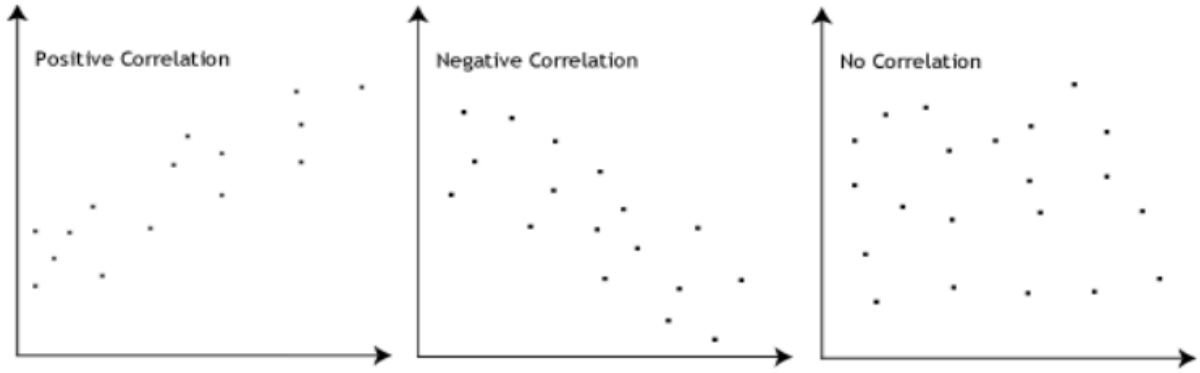
Ans:

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.

In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

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. This is shown in the diagram below:



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

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.	Normalized scaling	Standardized scaling
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.

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

Ans:

VIF(VarianceInflationFactor) basically helps explain the relationship of one independent variable with all the other independent variables. The formulation of VIF is given below: A VIF value of greater than 10 is definitely high, a VIF of greater than 5 should also not be ignored and inspected appropriately. A very high VIF value shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem 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.

Ans:

Q–Q plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data possibly came from some theoretical distribution such as a Normal, exponential or uniform distribution.

QQ plot can also be used to determine whether or not two distributions are similar or not. If they are quite similar you can expect the QQ plot to be more linear. The linearity assumption can best be tested with scatter plots. Secondly, the linear regression analysis requires all variables to be multivariate normal. This assumption can best be checked with a histogram or a Q-Q-Plot.

Importance of QQ Plot in Linear Regression : In Linear Regression when we have a train and test dataset then we can create Q-Q plot by which we can confirm that both the data train and test data set are from the population with the same distribution or not.

Advantages:

- It can be used with sample size also
- Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

Q-Q plot use on two datasets to check

- If both datasets came from population with common distribution
- If both datasets have common location and common scale

- If both datasets have similar type of distribution shape
- If both datasets have tail behavior