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?

   - Fall season seems to have attracted more booking. And, in each season the booking count has increased drastically from 2018 to 2019.
   - Most of the bookings has been done during the month of May, June, July, Aug, Sept and Oct. Trend increased starting of the year till mid of the year and then it started decreasing as we approached the end of year.
   - There are no users when there is heavy rain/ snow indicating that this weather is extremely unfavorable. Highest count was seen when the weathersit was' Clear, Partly Cloudy'.
   - Thu, Fir, Sat and Sun have a greater number of bookings as compared to the start of the week.
   - When it's a 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 a greater 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?

   drop_first = True is important to use because it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. If we don't drop the first column, then the dummy variables will be correlated (redundant). This may affect some models adversely and the effect is stronger when the cardinality is smaller.

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

   Temperature(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?

   I have validated the assumption of Linear Regression Model based on below 5 assumptions -

   - Normality of error terms: Residuals distribution should follow normal distribution and centered around 0 (mean = 0). We validated this assumption about residuals by plotting a Histogram of residuals and saw if residuals are following normal distribution or not. The diagram below shows that the residuals are distributed about mean = 0.

- Multicollinearity check: We calculated the VIF (Variance Inflation Factor) to get the quantitative idea about how much the feature variables are correlated with each other in the new model.
- Linear relationship validation: I visualized the numeric variables using a pair plot to see if the variables are linearly related or not.
- Homoscedasticity: I visualized the residuals and actual values(y_train) on a scatter plot to see whether variance is constant or not
- Independence of residuals: I plotted a graph with error terms to see whether the error terms are independent of each other.

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 –

   - Temperature
   - Year
   - weathersit_light snow & rain

General Subjective Questions

1) Explain the linear regression algorithm in detail.

   Linear Regression is a type of supervised Machine Learning algorithm that is used for the prediction of numeric values. Linear Regression is the most basic form of regression analysis. Regression is the most used predictive analysis model.

   Linear regression is based on the popular equation “y = mx + c”.

   It assumes that there is a linear relationship between the dependent variable(y) and the predictor(s)/independent variable(x).

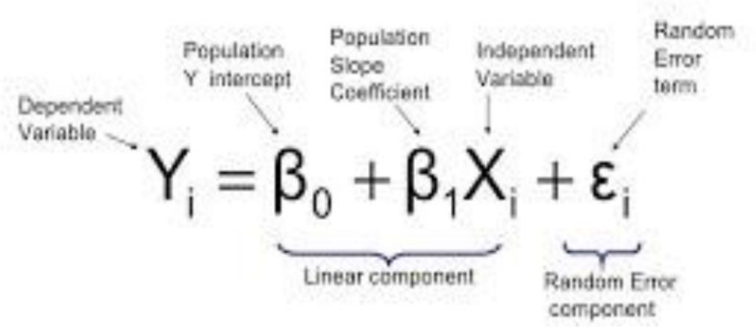
   In regression, we calculate the best fit line which describes the relationship between the independent and dependent variable. Regression is performed when the dependent variable is of continuous data type and Predictors or independent variables could be of any data type like continuous, nominal/categorical etc. Regression method tries to find the best fit line which shows the relationship between the dependent variable and predictors with least error.

In regression, the output/dependent variable is the function of an independent variable and the coefficient and the error term.

Regression is broadly divided into simple linear regression and multiple linear regression.

1. Simple Linear Regression: SLR is used when the dependent variable is predicted using only one independent variable.

The equation for SLR will be:



2. Multiple Linear Regression: MLR is used when the dependent variable is predicted using multiple independent variables.

The equation for MLR will be:

$$\textit{observed data} \rightarrow \quad y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p + \varepsilon$$

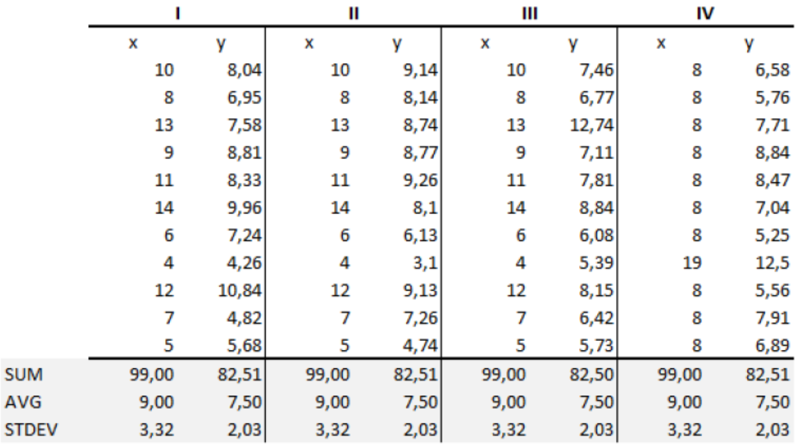
$$\textit{predicted data} \rightarrow \quad y' = b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p$$

$$\textit{error} \rightarrow \quad \varepsilon = y - y'$$

B1 = coefficient for X1 variable
B2 = coefficient for X2 variable
B3 = coefficient for X3 variable and so on…
B0 is the intercept (constant term)

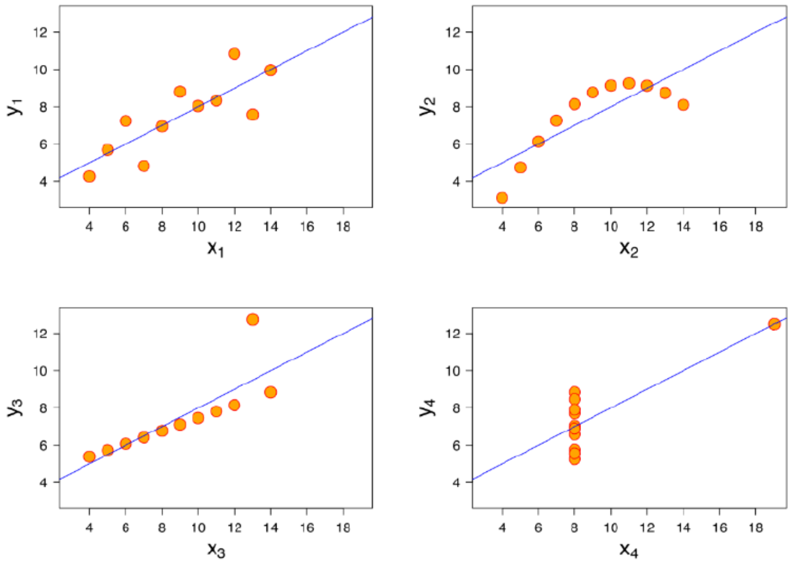
2) Explain the Anscombe’s quartet in detail.

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:

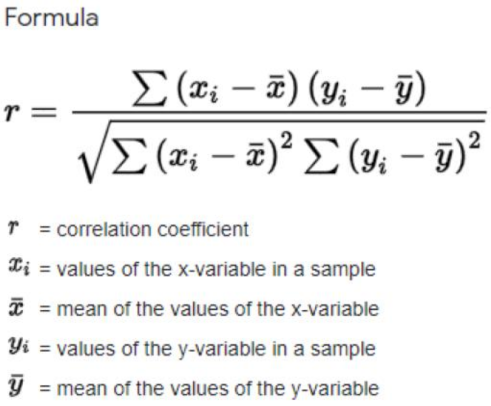
- The first scatter plot (top left) appears to be a simple linear relationship.
- The second graph (top right) is not distributed normally; while there is a relation between them, it’s not linear.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.

- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

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?

Pearson's r is a numerical summary of the strength of the linear association between the variables.It value ranges between -1 to +1. It shows the linear relationship between two sets of data. In simple terms, it tells us "can we draw a line graph to represent the data? "



Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

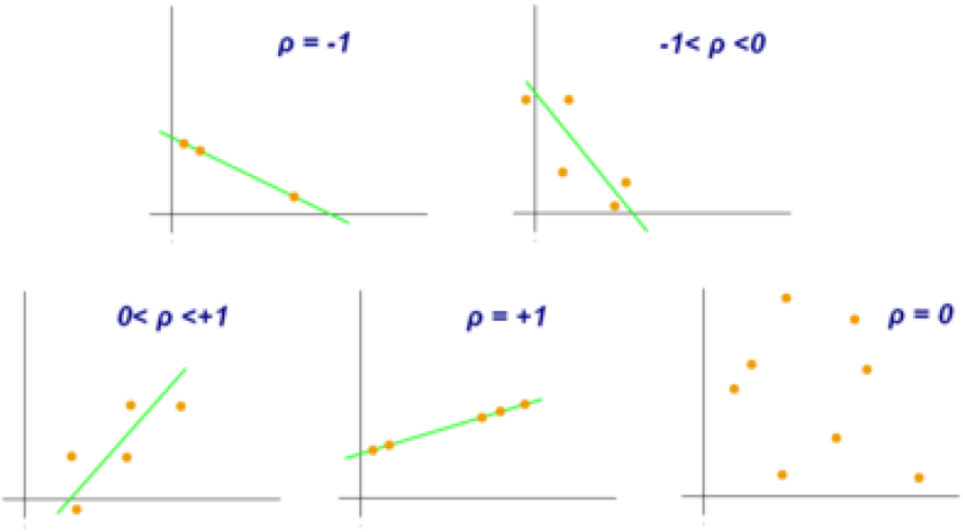
x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

As can be seen from the graph below,
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



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 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?

The VIF(Variance Inflation Factor) gives how much the variance of the coefficient estimate is being inflated by collinearity. If there is perfect correlation, then VIF = infinity. It gives a basic quantitative idea about how much the feature variables are correlated with each other. It is an extremely important parameter to test our linear model.

$$VIF = \frac{1}{1 - R^2}$$

Where R-1 is the R-square value of that independent variable which we want to check how well this independent variable is explained well by other independent variables. If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and its R-squared value will be equal to 1. So, VIF = 1/ (1-1) which gives VIF = 1/0 which results in “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.