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?
   - **Thursday, Friday and Saturday (485395,487790 and 477807 respectively)** have a greater number of bookings than other days. **Sunday** has the least number of rentals.
   - Most Bookings are done during the month of **May, Jun, July, Aug and September**
   - Clear Weather attracted the most rentals.
   - Bike rentals are low during **Light snow**
   - The months during **Fall and Summer attracted** the most rentals.
   - 2019 had **a greater number** of bike rentals than previous year. Which shows good improvement in Rentals
   - Rental counts are almost the same during working or non-working days.
2. Why is it important to use **drop_first=True** during dummy variable creation?
   Answer:

   It is important to use **drop_first=True** when we create dummy variables using pd.get_dummies().

   When we don't use drop_first, it causes multicollinearity among given features.

   In my assignment I wanted to look into the impact of **season_fall** in bike rentals. My VIF had a value of Infinity which shows very high multicollinearity. This should have been removed using drop_first=True while creating dummies.

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

   Answer:

   '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?
   Answer:
   I have validated the assumptions of Linear regression based on 5 checks on my model. They are:

   - Normality of errors terms.
     - Errors terms are normally distributed in my model
   - Linear Relationship validation between actual and predicted .
     - My model has linearity visible among variables.
   - Multicollinearity Check
     - There shouldn't be significant correlation among independent variables.
   - Homoscedasticity
     - There should be no visible patterns in residual values.
   - Independence of Residuals
     - No auto correlation.
   - All the above validation are done after model building.

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

   The top 3 features contributing significantly to the model are:

   (a) Temp
   (b) Winter
   (c) September month

General Subjective Questions:

1. Explain the linear regression algorithm in detail.

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

The representation is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). As such, both the input values (x) and the output value are numeric.

The linear equation assigns one scale factor to each input value or column, called a coefficient and represented by the capital Greek letter Beta (B). One additional coefficient is also added, giving the line an additional degree of freedom (e.g. moving up and down on a two-dimensional plot) and is often called the intercept or the bias coefficient.

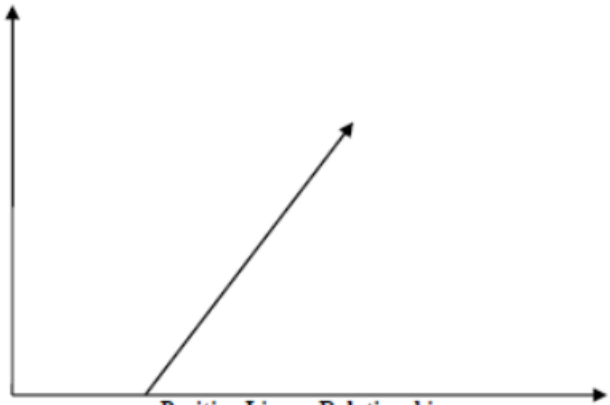
For example, in a simple regression problem (a single x and a single y), the form of the model would be:

$$y = B0 + B1*x$$

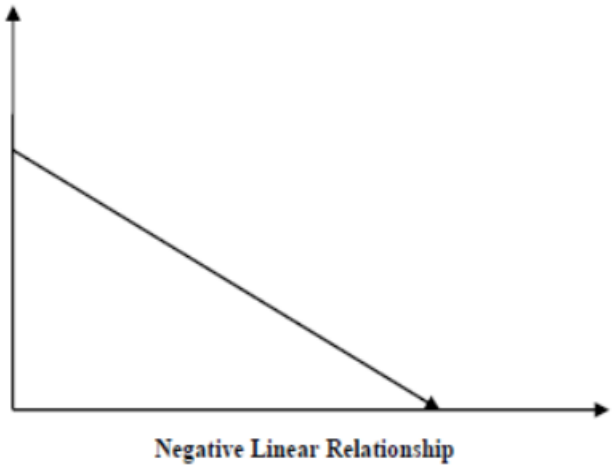
In higher dimensions when we have more than one input (x), the line is called a plane or a hyper-plane. The representation therefore is the form of the equation and the specific values used for the coefficients (e.g. B0 and B1 in the above example).

Furthermore, the linear relationship can be positive or negative in nature as explained below:-

1) Positive Linear Relationship:
   a) A linear relationship will be called positive if both independent and dependent variable increases. It can be understood with the help of following graph –



2) Negative Linear relationship:
   a) A linear relationship will be called positive if independent increases and dependent variable decreases. It can be understood with the help of following graph –



1 Linear regression is of the following two types –
   a. Simple Linear Regression
   b. Multiple Linear Regression

The following are some assumptions about dataset that is made by Linear Regression model -
1. 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.

2. 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.

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

4. Normality of error terms –
   - Error terms should be normally distributed

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

2. Explain the Anscombe's quartet in detail.

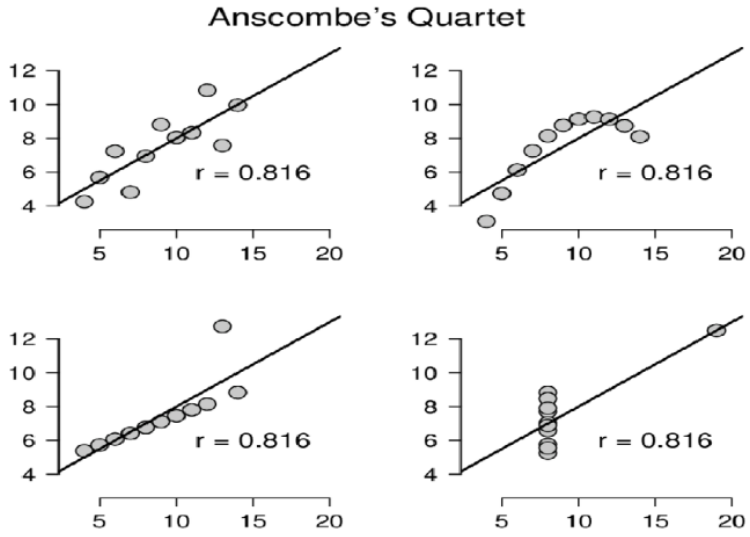
Answer:

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, 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:

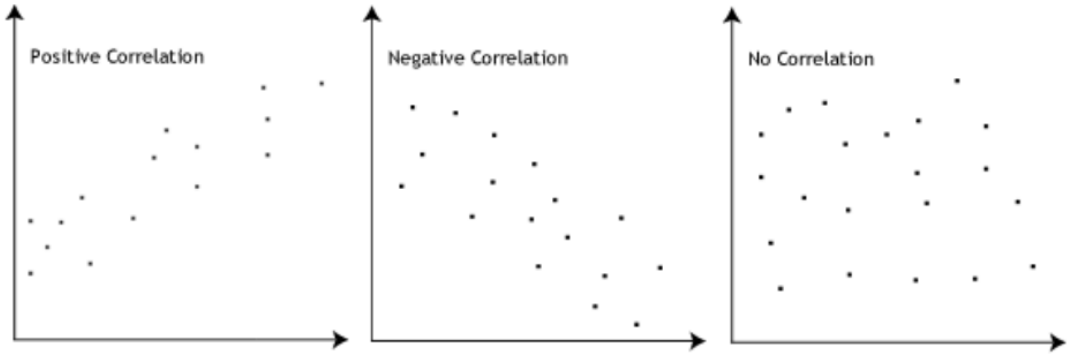


- Dataset I appears to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.
- 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?

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 with low values of one variable associated with 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. This is shown in the diagram below:



Pearson Correlation coefficient Formula

$$r = \Sigma (xi- \bar{x})(yi- \bar{y}) / \sqrt{\Sigma(xi- \bar{x})^2 \Sigma (yi- \bar{y})^2}$$

Where,

r = correlation coefficient
xi = x-values in the sample
$\bar{x}$ = mean of x variable
yi = x-values in the sample
$\bar{y}$ = mean of y variable

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

Answer:

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?

Answer:

Formula to calculate VIF:

$$VIF = 1 / (1- R^2)$$

I had similar case while building my own model for bike sharing assignment. It happened when I wanted to keep season fall to see the impact it had for bike sharing. This happens when there is a perfect correlation between season fall and other features.

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

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 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.

Advantages of Q-Q plot

- To find whether two populations are of same distribution.
- Skewness of distribution
- It can be used with sample sizes
- Presence of outliers, distributional aspects like shifts in location, shifts in scale, changes in symmetry.

Interpretation of two data sets

- Similar Distribution : If all point of quantiles lies on or close to straight line at an angle of
45 degree from x-axis
- Y-value < X-value: If y quantiles are lower than x quantiles.
- X-value < y-value: If x quantiles are lower than y quantiles.
- Different Distribution: If all point of quantiles lies away from the straight line at an angle
of 45 degree from x-axis