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:

   By analyzing, categorical columns using the box plots and bar plot, following points can be inferred following effects on the dependent variable:

   - Bookings are higher in near weekdays like Friday, Saturday and Sunday.
   - Summer and Fall season attracted more booking.
   - Booking is increased from 2018 to 2019 which shows progress in business.
   - Booking seems to be higher when temperature was neither too high nor too low.
   - Bookings were higher on the month May, June, July, August and September.
   - When weather was clear, bookings were higher.
   - Bookings was higher on non-holiday.

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

   Answer:

   drop_first = true is important to use, as it helps to decrease the number of extra columns created during the creation of dummy variable. So, it helps to decrease correlation created among dummy variables.

   Example:

   We have 3 types of value in column, and we want to create dummy variable for that column, if one variable is not X and Y, then for sure it is Z. So, we need third variable to identify Z.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

Answer:

By looking at the pair-plot among the numerical variables, temp variable is the one with the highest correlation with the target variable and that target variable is cnt.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

Answer:

Once the building model on the training set was built, following are assumption of Linear Regression Model:

- **Normality of error terms**: Normally distributed.
- **Homoscedasticity**: Visible pattern was not found.
- **Linear Relationship Diagram**: Linearity was visible.
- **Multicollinearity Test**: As VIF value was below 5, approved.
- **Independence of Residuals**: Ir_6 was 2.111, which confirms there is no auto-correction.

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:

These are top 3 features which contributed significantly towards explaining the demands of the shared bikes:

- Temp
- Windspeed
- Winter

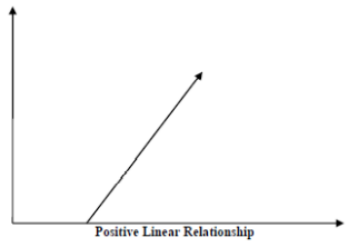
General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Answer:

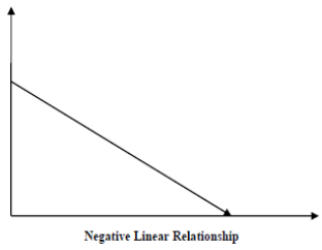
Linear Regression can be defined as the statistical model which analyses the linear relationship between a dependent variable with given set of independent variables. It is supervised leaning-based machine learning algorithm which is used to carryout regression task. When value of one or more independent variable changes, either by increasing or decreasing then the value of dependent variable also changes. Mathematically, this is represented by Y = mX+c where Y is dependent variable that needs to be predicted, X is the independent variable which we need to predict, m is the slop of regression line which shows effect of X on Y and c is a constant, which is also known as the Y-intercept. If X = 0 then Y = c. Linear regression can be either positive or negative in nature,

- Positive Linear Relationship



Positive Linear Relationship

- Negative Linear Relationship



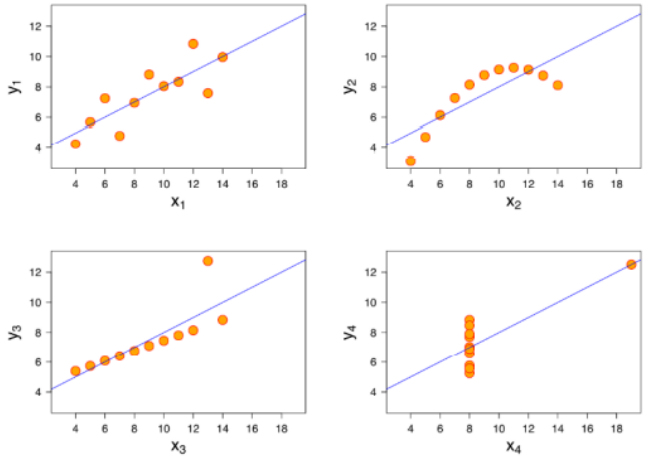
Negative Linear Relationship

2. Explain the Anscombe's quartet in detail. (3 marks)

Answer:

Anscombe's quarter comprises four datasets, each containing eleven (x, y) pairs. Important thing about these datasets is that they all share same descriptive statistics. Anscombe's quarter helps to demonstrate importance of graphical visualization, effect of outliers and other observations and statistical properties.

The graph visualization looks like



Example:

	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 = 9, mean of y = 7.5 for each dataset.**
- **Variance of x = 11, Variance of y = 4.13 for each dataset.**
- **Correlation Coefficient between x and y = 0.816 for each dataset.**

3. What is Pearson's R? (3 marks)

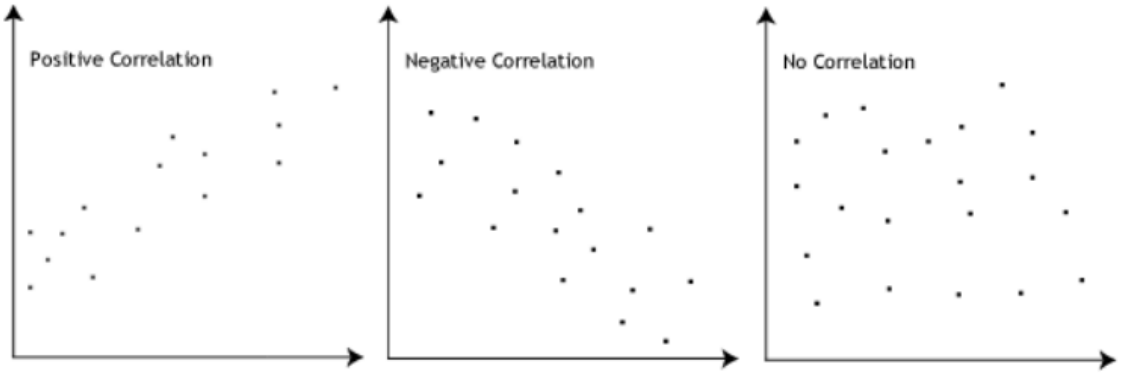
Answer:

Pearson's R which is also known as Correlation Coefficient numerical summary of the strength of the linear association between the variables. It is a measurement of linear correlation between two sets of data. The Pearson Correlation Coefficient, r can take a range of values from -1 to 1. If the value is 0, then it indicates that there is no association between the two variables. If the value is greater than 0 then, it indicates that there is a positive association. If then value is less than 0 than, it indicates that there is a negative association.

The formula for Pearson's R is:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

When the value of one variable increase then value of other variable decreases and vice versa.



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

Answer:

Scaling which is also known as 'Feature Scaling' is a technique to standardizing the feature which is present in the data in a fixed range. Feature Scaling is performed while data pre-processing to handle the highly varying magnitude of values/units. Machine Learning algorithm also weighs some greater value, higher including some small values as the values, regardless of the unit.

Example:

If an algorithm dies not use scaling, it will consider few until values 7000 meter to be greater than 10KM but that's not true and in such cases, algorithm will give wrong predictions. This is the main reason to use Feature Scaling to bring all the values to same magnitudes or units.

Normalization	Standardization
Minimum and maximum value of features are used for scaling.	Mean and S.D is used for scaling.
Normalization is done when two features are of different units.	Standardization is done to ensure mean = 0 and S.D = 1.
Normalization is affected by outliers.	Standardization is much less effected by outliers.
Scales values between [0, 1] or [-1, 1].	Standardization is not bounded by any range.
Scikit-Learn provides a transformer which is called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer which is called StandardScaler for Standardization.

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

Answer:

Value of VIF can be observed infinite only when there is perfect correlation or r^2 = 1, which indicates that there is high correlation between the variables.

Example:

VIF of column is 2 which means that the model coefficient variance is inflated by 2 factor due to multicollinearity. VIF value become infinity only when there is perfect correlation between two independent variables. As VIF = 1/(1-r^2), where r^2 = 1 then the equation becomes 1/ (1-1) =1/0 whose value is infinity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

(3 marks)

Answer:

Q-Q plot which is also known as the quantile-quantile plot is probability plot. Q-Q plot is a graphical representation of the two datasets which comes from the populations with a common distribution. The patterns which are there in the plots is used to compare the two distributions.

Q-Q plot known as quantile – quantile plot is used to determine the following:

- To find whether two populations of the same distribution or not.
- To find the skewness of distribution.
- To know whether residuals follow normal distribution.