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?**

Answer:

We visualized the categorical variables with the help of boxplot and these are few of the insights that we could draw from the visualizations –

- Among all the seasons Fall has the highest demand for bikes.
- Between the years 2018 and 2019 we could see a rise in the demand for bikes for the year 2019.
- We also saw that the demand for bikes are lower in the months of January and December due to long holidays and winter. Whereas, the highest demand for bikes is during the month of September.
- During holidays the mean demand for bikes decreases.
- The demand for bikes is more or less the same throughout the week.
- The demand for bikes depends on the weather as during good weathers the demand for bikes is higher compared to the demand for bikes in bad or moderate weather conditions.

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

Answer:

The drop_first = True is important to use as it reduces the extra column created during the creation of dummy variables. It reduces the correlation created among the dummy variables.

Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished nor semi furnished, then It is obviously unfurnished. So we do not need 3rd variable to identify the unfurnished.

Syntax - drop_first: bool, default False, which implies whether to get k-1 dummies out of k categorical levels by removing the first level.

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

Answer:

Looking at the pair plot among the numerical variables, it seems that **'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:

We have validated the assumption of Linear Regression based on the following assumptions-

- Normal distribution of error terms - Error terms should be normally distributed with zeros the mean.
- Checking the Multicollinearity – The multicollinearity should be insignificant among variables.
- Validation of the Linear Relationship – Linearity should be visible among variables.
- Homoscedasticity – No visible pattern in residual values.
- Independence of residuals – No auto-correlation.

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

Answer:

The top 3 features that contribute significantly towards the demand of shared bikes after the final model are –

- temp
- workingday
- spring

General Subjective Questions

1. Explain the linear regression algorithm in detail

Answer:

Linear regression is one of the very basic forms of machine learning where we train a model to predict the behavior of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated. 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 following equation-

$Y = mX + c$

Where,

Y is the dependent variable we are trying to predict

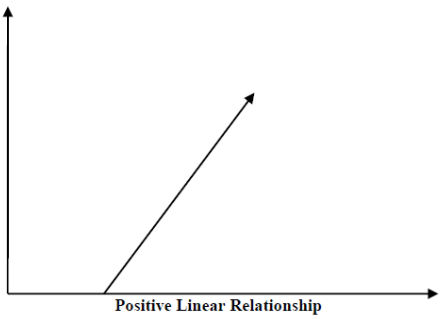
X is the independent variable we are using to make predictions

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.

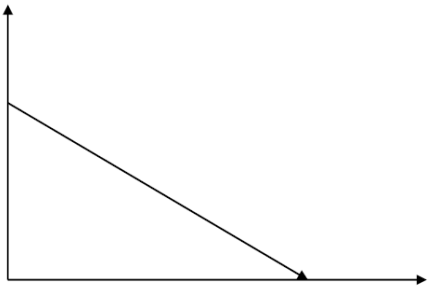
Linear Relationship can either be positive or negative in nature as explained below –

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



Positive Linear Relationship

- Negative Linear Relationship :
  A linear relationship will be called negative if independent increases and dependent variable decreases. It can be understood with the help of following graph –



Negative Linear Relationship

Linear Regression is again of two types –

- Simple Linear Regression
- Multiple Linear Regression

2. . Explain the Anscombe's quartet in detail.

Answer:

Anscombe's Quartet was developed by statistician Francis Anscombe. Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.

	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

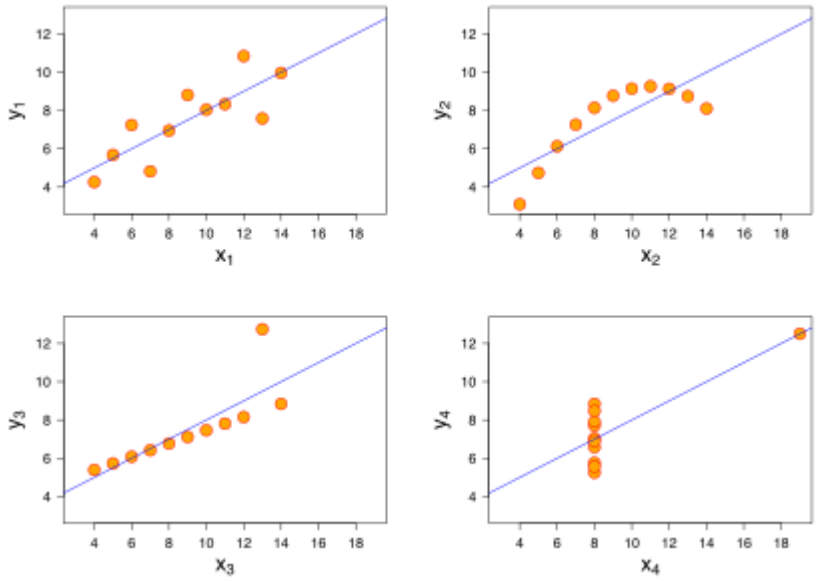
The summary statistics show that the means and the variances were identical for x and y across the groups:

• 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?

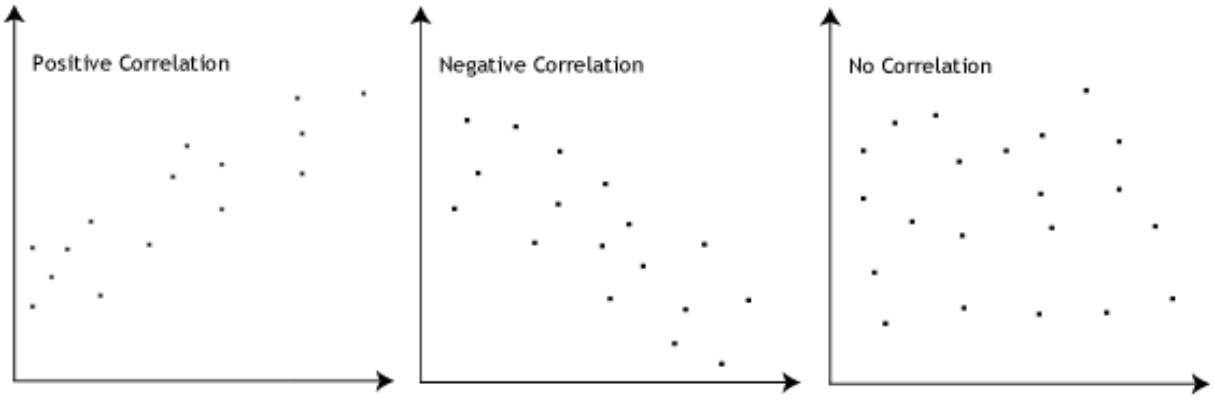
Answer:

A numerical evaluation of the strength of the linear connection between the variables is provided by Pearson's r. 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. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association



Pearson r Formula

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

Where,

- 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

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

Answer:

Feature scaling is a method for uniformly distributing the independent features in the data over a predetermined range. It is done as part of the pre-processing of the data to deal with extremely variable magnitudes, values, or units. In the absence of feature scaling, a machine learning algorithm would often prioritize larger values over smaller ones, regardless of the unit of measurement. It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Serial 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.	It is useful when we don't know about the distribution.	It is useful when the feature distribution is Normal or Gaussian.
6.	It is often called as Scaling Normalization	It is often called as Z-Score Normalization.

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

Answer:

If there is perfect correlation, then VIF = infinity. A high VIF score denotes a strong connection between the variables. The presence of multicollinearity causes the variance of the model coefficient to be exaggerated by a factor of 4 if the VIF is 4. VIF displays a complete correlation between two independent variables when its value is infinite. When the correlation is perfect, we have R-squared (R2) = 1, which results in 1/ (1-R2) infinite. To fix this, we must remove the variable from the dataset that is the exact multicollinearity's cause.

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

Answer:

Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential. 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 is that it is frequently desirable to determine whether the assumption of a common distribution is supported when there are two data samples. If so, location and scale estimators can combine the two sets of data to derive estimates for the shared position and scale. If two samples do differ, it is also helpful to comprehend the variations. Compared to analytical techniques like the chi-square and Kolmogorov-Smirnov 2-sample tests, the q-q plot can shed more light on the nature of the difference.