

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?

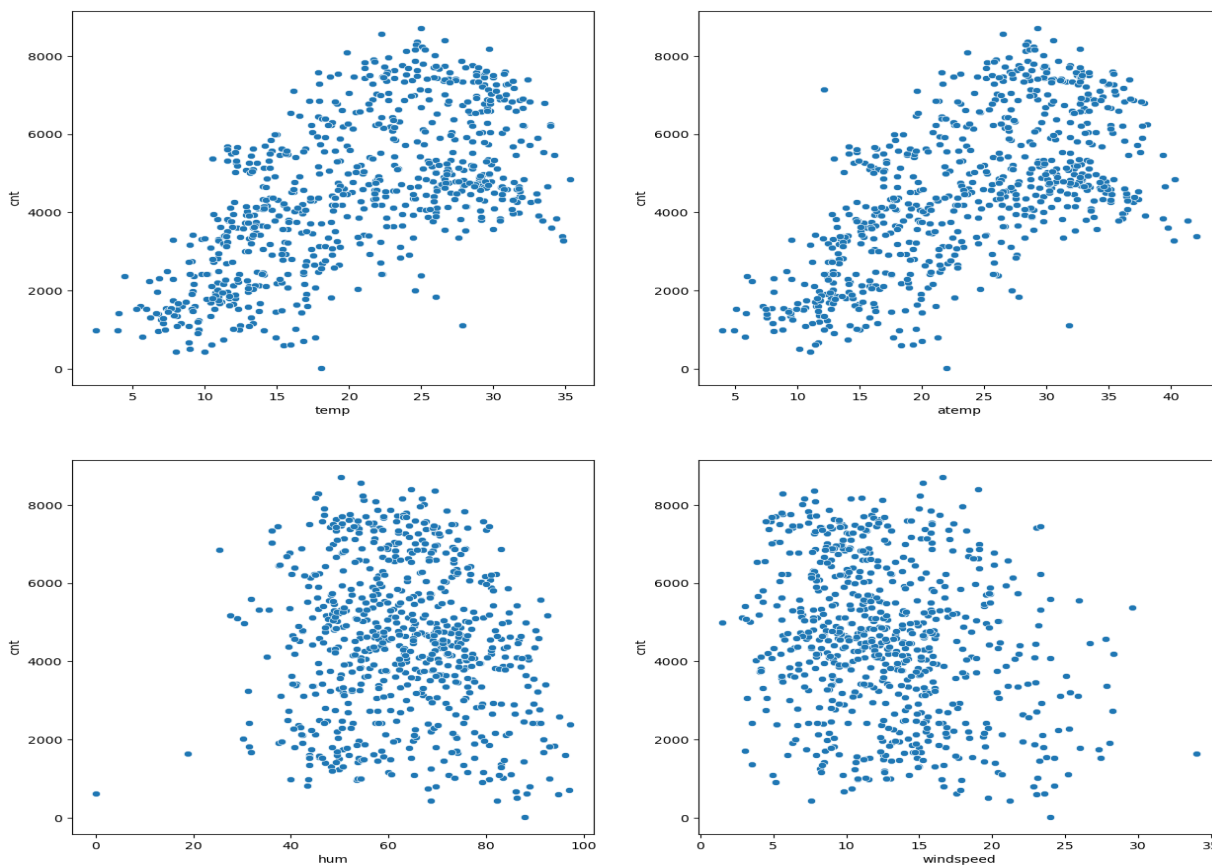
Answer : There are some categorical variable like season, weekdays, holiday, weathershit, month this categorical variable have the most impact on dependent variable **cnt**

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

Answer : Because when we create dummy variable system generated the variable equal to the count of category in the variable let say n to define the correlation , but the same correlation can be achieved by n-1 dummy variables as well ,that's the reason why we use **drop_first=True** .

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 below we can say that the temp and atemp variable have 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 assumption of Linear Regression Model based

1: Error terms should be normally distributed

2: There should be insignificant multicollinearity among variables.

3: variables should be linear

4: There should be no visible pattern in residual values.

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

Answer:

Temp, Year, and Season are the top 3 features which contributing significantly towards explaining the demand of the shared bikes

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Answer:

Linear regression is a statistical algorithm that uses a linear relationship between an independent variable and a dependent variable to predict future events. It's a data science and machine learning method used for predictive analysis.

The goal of linear regression is to find the best values for the coefficients in a model to minimize the difference between the predicted values and the actual values. This difference is known as the residuals. The best fit line should minimize the error between the predicted and actual values.

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

$$Y = mX + c$$

Here, 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 regression is of the following two types –

1: **Simple Linear Regression** – Dependent variable predicted using Single Independent variable

2: **Multiple Linear Regression** - Regression – Dependent variable predicted using Single Independent variable.

Assumptions -

The following are some assumptions about dataset that is made by Linear Regression model –

- Multi-collinearity –
 - o 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.
- Auto-correlation –
 - o 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.
- Relationship between variables –
 - o Linear regression model assumes that the relationship between response and feature variables must be linear.
- Normality of error terms –
 - o Error terms should be normally distributed
- Homoscedasticity –
 - o 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	

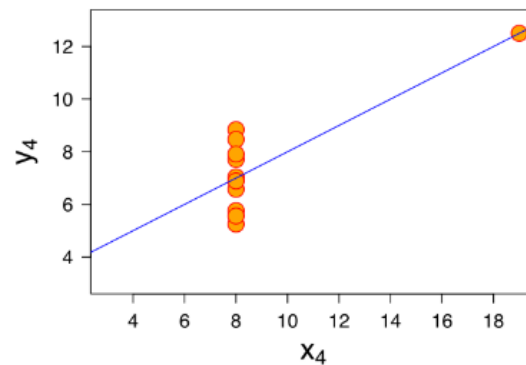
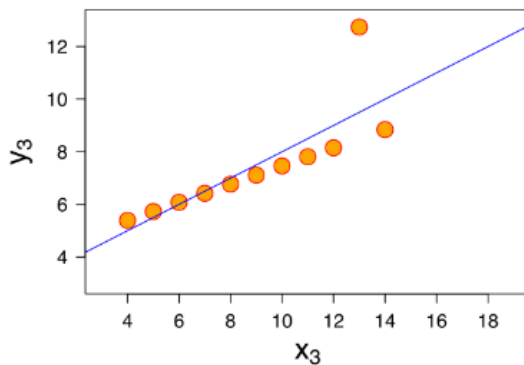
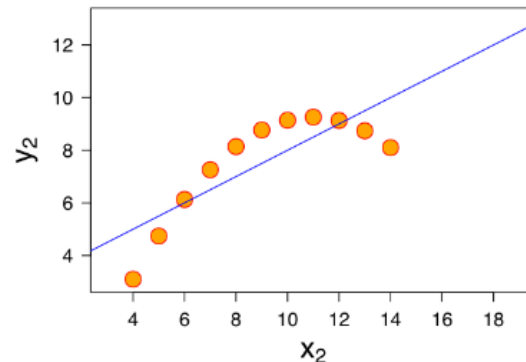
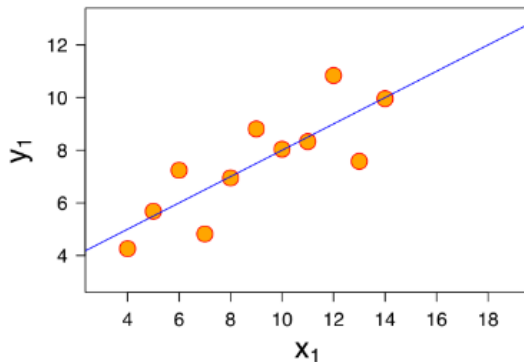
The summary statistics show that the means and the variances were identical for x and y across the groups:

- 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.

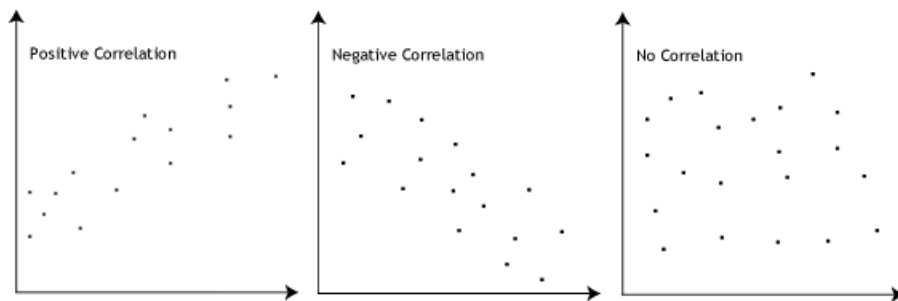
3. What is Pearson's R?

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:



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

Answer:

When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. So we need to scale features because of two reasons:

1. Ease of interpretation

2. Faster convergence for gradient descent methods You can scale the features using two very popular method:

1. Standardizing: The variables are scaled in such a way that their mean is zero and standard deviation is one.

2. MinMax Scaling: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

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-square, etc.

Normalized scaling	Standardized scaling
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.

scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
It is really affected by outliers.	It is much less affected by outliers.
ikit-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:

If there is perfect correlation, then $VIF = \infty$. 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 $R\text{-squared } (R^2) = 1$, which lead to $1 / (1 - R^2) = \infty$. 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:

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

