Vibhor Malik - Assignment-based Subjective Questions

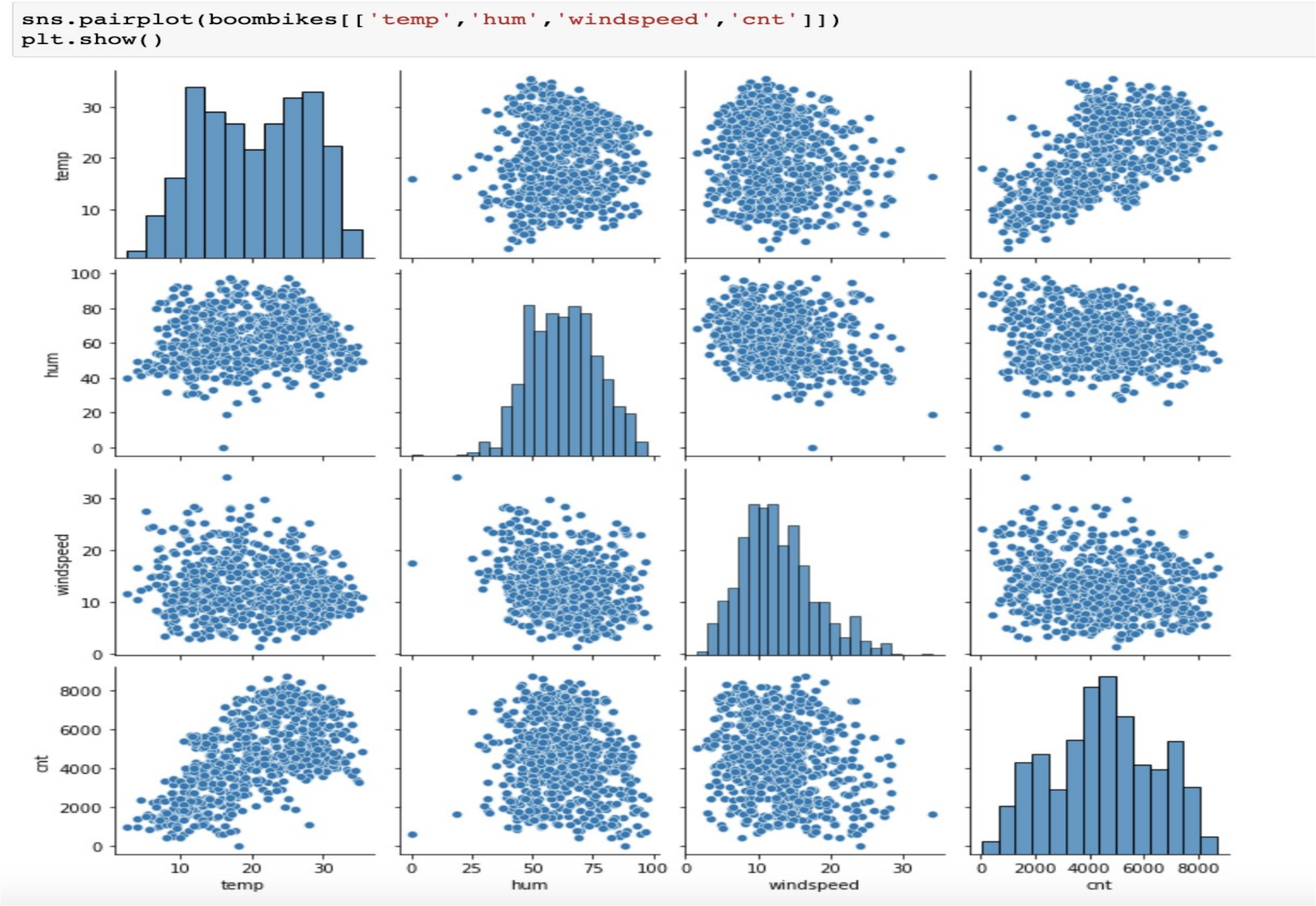
Qus 1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
Ans 1: According to the observations from bar charts for categorical variables :

- In Year 2019 more bikes are rented as compared to 2018
- Clear weather attracts more customers
- Bike renting is seen during working days and non holidays.
- More bikes are rented during the 'summer' season.

➔ More bikes are rented during the May-Oct month.

Qus 2: Why is it important to use drop_first=True during dummy variable creation?
Ans 2: While creating dummy variables, we can always find out which possible combination would be the last column, thus to save execution time.
e.g. if there are three values in a categorical columns –
we can use 000, 100, 010, 001
in this one if non is selected we can always know that which column it is

Qus 3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Ans 3: `Temp` and 'atemp' variables have highest correlation with Target Variable `cnt` but we only use one of these to avoid multicollinearity

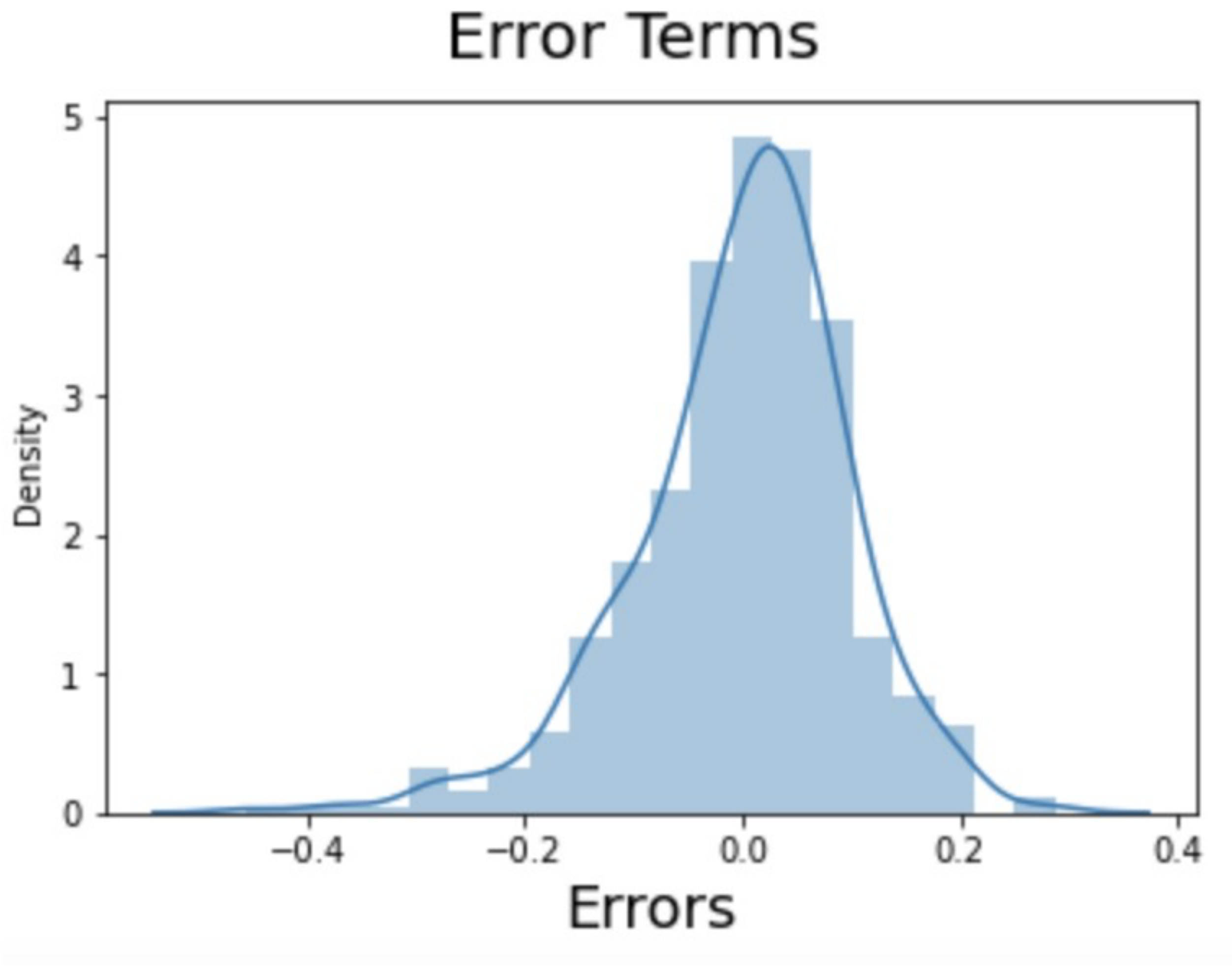


```
sns.pairplot(boombikes[['temp','hum','windspeed','cnt']])
plt.show()
```

<u>Qus 4:</u> How did you validate the assumptions of Linear Regression after building the model on the training set?

<u>Ans 4:</u>

We created a PDF of errors to check if its is normally distributed or not.
If the error terms don’t follow a normaldistribution, confidence intervals may become too wide or narrow.



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

Ans 5:

1. Temperature (+ ve Correlation)
2. Year 2019 (+ ve Correlation)
3. Clear weather (- ve Correlation)

General Subjective Questions

Qus 1: Explain the linear regression algorithm in detail.

Ans 1:

In we talk dictionary meaning – linear regression “a measure of the relation between the mean value of one variable (e.g. output) and corresponding values of other variables (e.g. time and cost).”

We can find the best first line which can pass through all the data points and have minimal residue. Use that line to find intercept and slope of the line. Using that slope and intercept predict data for new data points.

By math we can say a linear regression equation as: $y = a + bx$

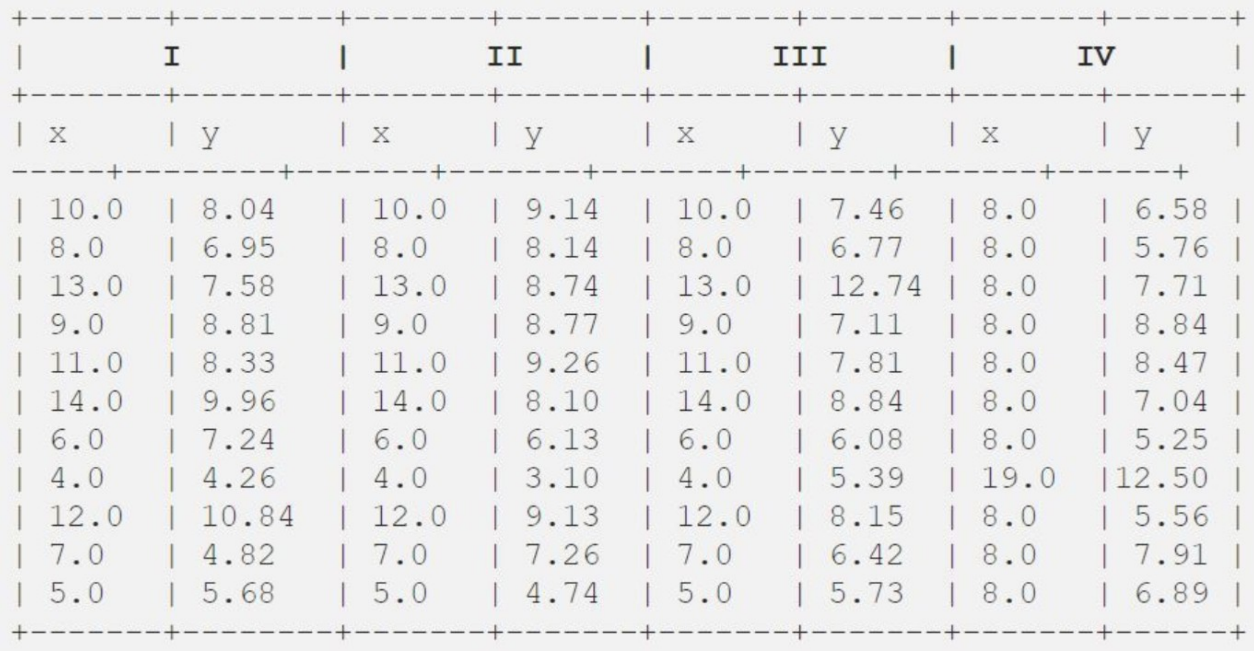
Here, x and y are two variables on the regression line.

b = Slope of the line
a = y-intercept of the line
y = Dependent variable from dataset
x = Independent variable from dataset

Qus 2: Explain the Anscombe's quartet in detail.
Ans 2:

Comprises 4 datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points.

Those 4 sets of 11 data-points are given below.



I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Qus 3: What is Pearson’s R?

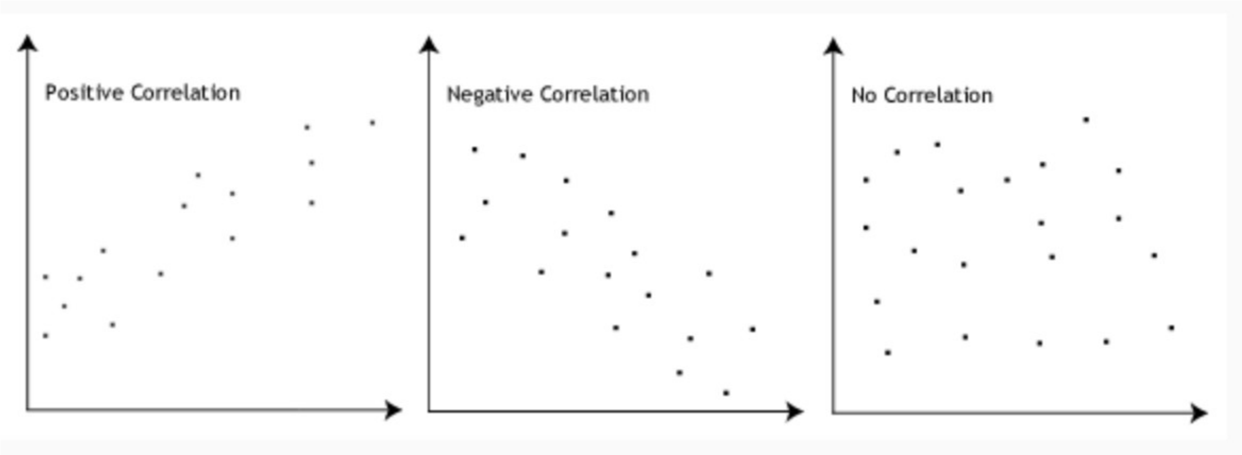
Ans 3:

The Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. 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.

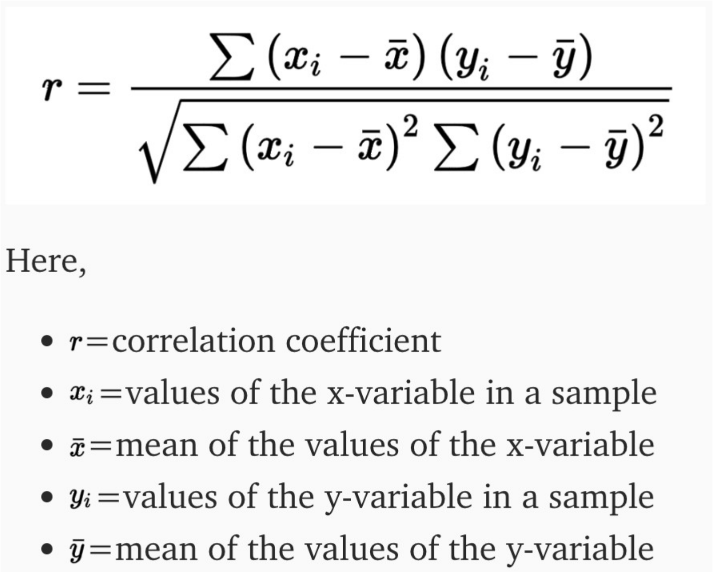
The Pearson's correlation coefficient varies between -1 and +1 where:

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:



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

Here,

- 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

<u>Qus 4:</u> What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans 4: Scaling is to convert few columns to a different scale so all columns are similar in scale before giving those columns to train in the model

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.

Normalization typically rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

S.NO.	Normalisation	Standardisation
1.	Min and maxi 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.

6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7.	It is useful when we don’t know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is a often called as Scaling Normalization	It is often called Z-Score Normalization.

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

Ans 5:

In perfect correlation, then VIF equal to infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which leads to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Ans 6: A Q–Q plot is used to compare the shapes of distributions, providing a graphicalview of how properties such as location, scale, and skewness are similar or different in the two distributions.

Quantile-Quantile plots are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. Let’s say, the median is a quantile where 50% of the data fall below that point and 50% lie above it.