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: The Categorical variables are season, month, holiday, weather situation, year, weekday and working day. Boxplot was used to analyze. Below are the inferences:

a. Season -The demand is least for spring season.
b. Month -The number of bike shares gradually increase until September and then starts to decrease.
c. Holiday -The cnt values are less during holidays.
d. Weather situation -We do not find any values for weathersit - Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog. However, there is high number of bike-sharing during Clear, Few clouds, partly cloudy weathersit.
e. Year -The cnt values have increased in 2019 compared to 2018.

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

Answer: drop_first =True is important, to reduce the number of columns used. When the data can be explained using two columns instead of three, it would help in the readability and reduces the correlation created among the dummy variables.

Example: Say we have three values (A, B, C) for a particular variable.

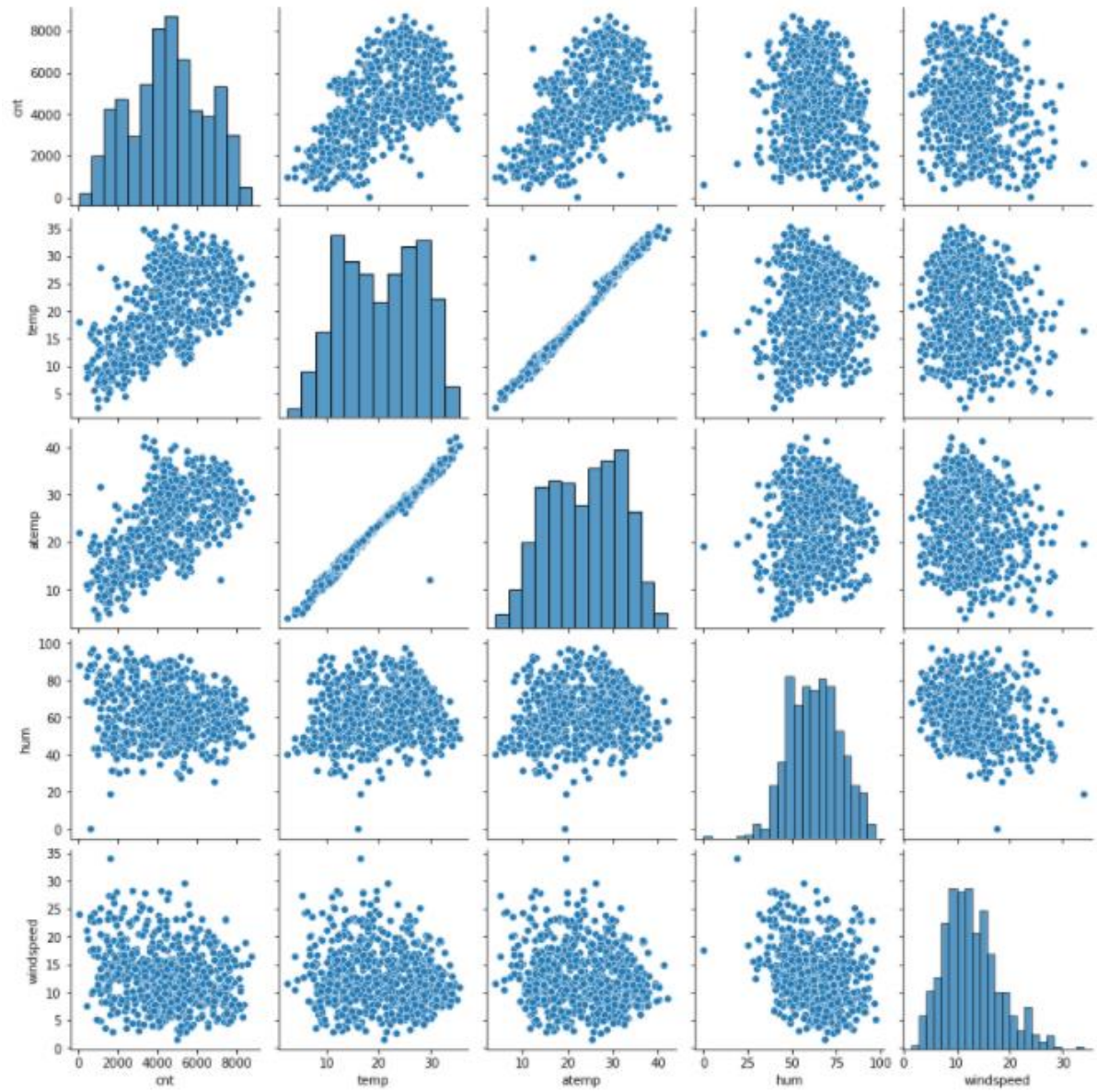
We can interpret using 10 -A, 01 -B, 00 -C, so only two columns are enough to explain three values.

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

with the target variable? (1 mark)

Answer: Temp and Atemp have the highest correlation with the target variable cnt.

```
# Numeric variables:
sns.pairplot(data=bike_new,vars=['cnt', 'temp', 'atemp', 'hum','windspeed'])
plt.show()
```

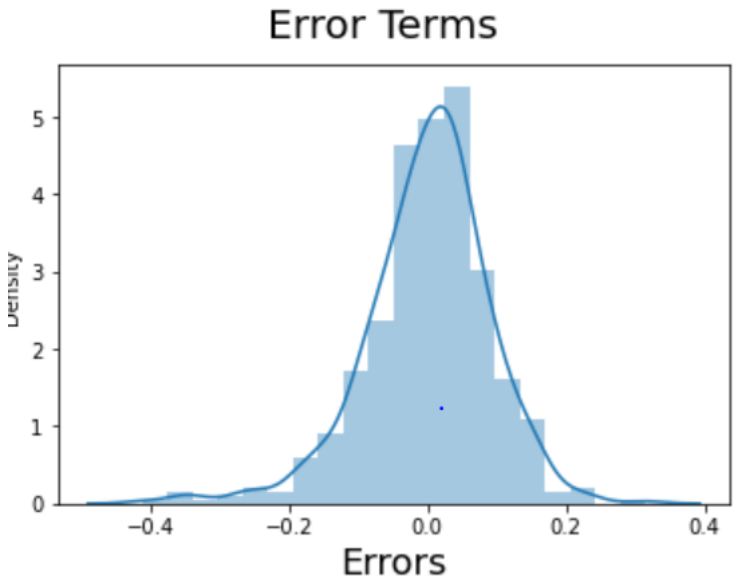


---

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

Answer: Based on

1. Residual Analysis: Error terms are normally distributed centered around zero.



2. Linearity between target and input variables [straight line].

3. Homoscedasticity: Cone shape should not be present; no visible pattern should be observed.

4. Errors should be independent; this is checked using the DW value. [in our case it was: 2.089].

5. VIF values and p-values are optimal.

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:

1. Temperature - Temp is the most significant with the largest coefficient of 0.5499
2. Year -Followed by year with coefficient: 0.2331
3. Season – with coefficient 0.1307

General Subjective Questions

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

Answer: Linear Regression is a supervised ML model in which the model finds the best fit line which uses independent variables to predict the dependent variable, provided the output

variables to be predicted are continuous. The model finds the linear relation between predictor and target variables.

We will use straight line plot to predict the possible outcomes for a particular value of independent variable. We use ordinary least square method to identify the best fit line.

Equation of straight line is y = mx+c, where m is the slope and c is the interpret.

Linear regression can be classified into Simple LR and Multiple LR.

Assumptions on LR:

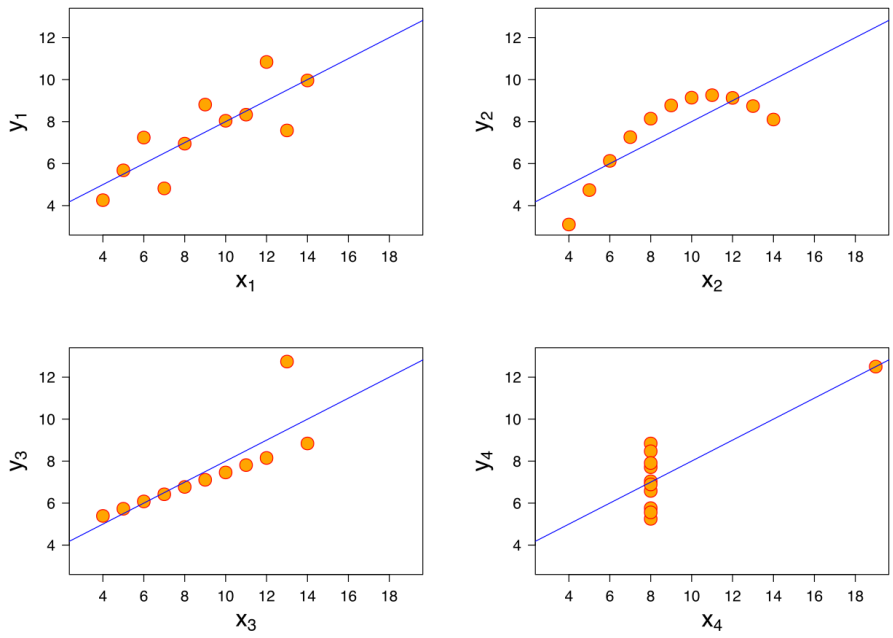
a. We assume that the target variable y and input variable x are linearly dependent
   Y =beta0 + beta1 *X + error
b. Error terms are normally distributed.
c. Error terms are independent of each other.
d. Error terms have the same variance. – Homoscedascity.

These assumptions allow us to make inferences on the sample data to the population.

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

Answer: Anscombe's quartet has four data sets that are identical in simple descriptive statistics, however when plotted they have very different distributions.

Each dataset consists of eleven (x,y) points.



It was built by Francis Anscombe to demonstrate the importance of visualizing the data using graphs before model building.

The statistical information such as mean and std dev for these four datasets are approximately similar.

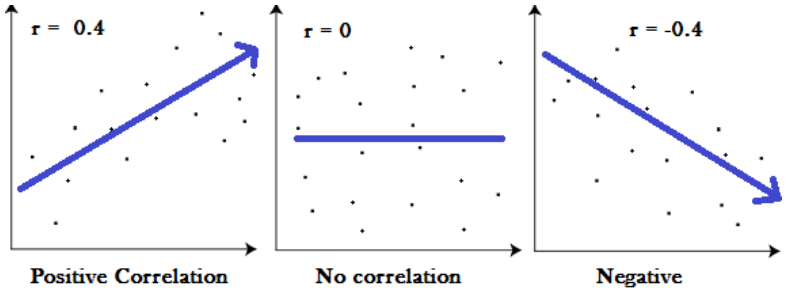
The four datasets can be explained as:

a. The first dataset X1, fits the linear regression model well.
b. X2, could not fit linear regression model on the data as the data is non-linear.
c. X3 shows the distribution is linear, however the outliers involved in the dataset are not handled by linear regression model
d. X4 shows that one outlier is enough to produce a high correlation coefficient.

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

Answer: Pearson correlation coefficient known as Pearson's r, is a measure of linear correlation between two variables. It is the ration between their covariance and product of their std dev. The outcome is always in the range 1 and -1, where:

- 1 indicates positive relationship.
- -1 indicates negative relationship.
- zero indicates no relationship.



Pearson's Correlation coefficient is given as:

Rho(x,y) =Cov(x,y)/sigma(x)* sigma(y)

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

Answer: It is a data processing step which helps to normalize/standardize the independent variables within a particular range. This step is performed to maintain all variables in the same

range for better results. Scaling only affects the coefficients and not the T-statistics, F-statistics, P-statistics, or the R-squared values.

Scaling Methods:

1. Normalization – min-max scaling that uses the minimum and maximum values and scales them between 0 and 1.
   Minmax scaling = X -min(x)/max(X)-min(X)
2. Standardization –uses the mean and std dev for scaling, it doesn't have any range, is used when we want to have zero mean and unit std dev.
   Standard scaling = X-mean(X)/std(X)

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

(3 marks)

Answer: The Variance Inflation Factor (VIF) is a measure of collinearity among predictor variables within a multiple regression. It is the ratio of the variance of all betas in the model and the variance of a single beta if it were fit alone.

VIF = 1/(1-Rsqu)

A large value of VIF indicates that there is high correlation between the variables, therefore, if there is perfect correlation, then VIF = infinity. To fix this issue, we need to drop one of the variables from the dataset which is causing this perfect correlation.

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 is a probability plot, which compares two probability distributions by plotting their quantiles against each other. It is a scatterplot.

Use: to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions. it helps to determine if two data sets

It is very important in linear regression when we have training and test data set received separately, we can confirm using Q-Q plot that both the data sets are from populations with same distributions.