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.

atemp (feeling temperature) and temp: (note, we have considered atemp in our model)

With increase in temperature, the count of boombike users increases.

With decrease ins temperature, count of boombike users decreases.

There is strong positive correlation of count with temperature

Season:

Fall season has the maximum boom bike usage followed by summer and then winter

Month:

June and September months have maximum boom bike usage followed by summer and then winter

May,June,July,August,September and October are 6 months which sees lot of boombike usage.

This further confirms the number users of boom bikes are higher in Fall and Summer

Weather Situation:

As per the case study these are the 4 weather situations

- 1: Clear, Few clouds, Partly cloudy, Partly cloudy

  This seems to be great weather and boombike usage increases under such weather conditions

- 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

  This seems to be OK weather if not great weather and boombike usage decreases slightly under such weather conditions in comparison to Great Weather conditions.But still the count is good

- 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

  This seems to be bad weather and boombike usage decreases drastically under such weather conditions

Conclusion as the Weather Conditions deteriorate, the boombike usage decreases

Year (age of Boombike):

Year variable tells us the age of the company boombike.It gives a clear indication with passage of time, boombike has gained popularity and that reflect in the count of boombike usage.

2019 shows more users in comparison to 2018

Holiday: It means non-weekend holiday. It shows a negative correlation with boombike usage count.

Weekend: On weekends there is slight increase in count

IsLongWeekend: On LongWeekends, there is slight increase in count

Windspeed: higher the windspeed lower the count, lower the windspeed higher the count.There is a negative correlation

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

Answer:

Drop_first = True eliminates the unnecessary dummy variable which can be easily derived from other dummy variables

Example: Season has 4 levels. A-Spring, B-Summer, C-Fall, D-Winter

On creating Dummy variables it looks something like this:

	Spring	Summer	Fall	Winter
Sprint	1	0	0	0
Summer	0	1	0	0
Fall	0	0	1	0
Winter	0	0	0	1

The 4^{th} column is not necessary as from the values of Spring, Summer and Fall, we can know whether it is Winter or not. When the other columns have values as 0, then we know it is winter or any of the first 3 columns have a value as 1, then we know it is not winter.

So, we have n levels in a categorical variable, we just need (n-1) dummy variables

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

Answer: temp and atemp have the highest correlation

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

Answer:

- Error Terms are normally distributed

- Error terms have zero mean
- Error Terms are independent
- By checking Error Terms have constant variance

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

- yr (Age of Boombike – 0 means it 2018, 1 means 2019)
- atemp (feeling temperature)
- windspeed

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
Answer:
Linear Regression is a kind of Supervised Learning where the output variable to be predicted is numeric or a continuous variable and prediction is done based on past data.

Linear regression follows the equation of a straight line:
Simple Linear Regression: $y = mx + c$ or $y = B0 + B1X1$
Multiple Linear Regression: $y = B0 + B1X1 + B2X2 \ldots. + BnXn$

Where, y is the response, c or B0 is the intercept, B1 is the coefficient of X1 and similarly B2 is coefficient of X2 and so on.

Coefficients: the per unit increase in the variable when other predictors are held constant, there is increase in Y by the coefficient of the changed variable
Below are the Linear Regression Assumptions:

- Linearity: Getting the correct functional form written as a linear equation is extremely important In case of a simple linear regression model fits a line where a in case of multiple linear regression model fits a hyperplane.
- Homoscedasticity: means constant variance of standard error (Standard Error = Y_Predicted - Y_Actual)
- Independent Error terms: no co-relationship should exist between Standard errors
- Normal distribution of error terms
- For multiple linear regression, there should not be multi-collinearity. It happens when independent or predictor variables are collinear. It affects the coefficients and the standard errors

2. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's quartet involves 4 data sets having similar statistical observations which provide the same information (i.e. variance and mean) for each x and y point in each of the four data sets. But, when these data sets are represented graphically, they appear completely different from each other.

Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. Through visualization, one can identify the data anomalies like outliers.

Key here is before creating a model, always visualize the data.

3. What is Pearson's R?

Answer:

The **Pearson correlation coefficient (*r*)** is a method to measure a linear correlation. measures the strength and direction of the relationship between two variables. It varies between -1 to 1.

If the value is less than 0 it means it is a negative correlation (increase in x sees decrease in y).

If the value is greater than 0 it means positive correlation (increase in x sees decrease in y).

Values closer to -1 and 1 are considered strong correlation.

It also helps in identifying how close the observations are to a line of best fit.

It tells whether the slope of the line of best fit is negative or positive. When the slope is negative, *r* is negative. When the slope is positive, *r* is positive.

The Pearson correlation coefficient is a good choice when:

- The variables are quantitative
- The variables are normally distributed
- The data have no outliers
- The relationship is linear

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

Answer:

What is scaling: Here, Scaling refers to Feature Scaling. It is process of putting the Feature values within the same range.

Why is scaling performed: For faster and smoother gradient descent**.** With different features having different scales or range of values, the algorithm will follow different step sizes. This not only makes the process (like gradient descent) unsmooth but also increases the execution time. Therefore, to ensure

that gradient descent converges more smoothly and quickly, we need to scale our features so that they share a similar scale.

Difference between normalized scaling and standardized scaling:

Normalization	Standardization
Minimum and maximum value is used to do the scaling	Mean and standard deviation is used
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 -1 to 1	No specific range is defined
It is used when there are no outliers	less affected by outliers.
It is used when we don't know about the distribution	It is used when the distribution is Normal or Gaussian.

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

VIF stands for Variance Inflation Factor.

VIF = 1/(1-R^2) where R^2 = Residual Squared Error

When the VIF is infinite, it means the R^2 value is 1. This indicates that the independent variables are highly collinear with each other. This is case of **multi-collinearity**

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 also called Quantile-Quantile plot.

It plots Quantiles of one dataset against quantiles of another dataset.

The graph might result in datapoints on a straight line or datapoints on either side of a straight line.

If the more datapoints are on a straight line, it means the two datasets have a similar distribution else they do not have a similar distribution.

Example: