Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

Categorical variables in the given dataset are ‘**season’**, ‘**year’**, ‘**month’**, ‘**weekday’**, ‘**workingday’**,**’holiday’** and ‘**weathersit’**

Season

- Fall: Highest bike demand
- Spring : Lowest bike demands
- Summer : Demand is intermediate but higher than winter
- Winter : Demand is lesser than Summer but higher than Spring

Year : The demand for bike has increased from 2018 to 2019 which suggests that the bile rentals gained popularity gradually

Month:

- June, July, August, September and October have seen highest demand for rental bikes
- Jan, Feb, November and December have faced the lowest demand
- This suggests that September is very much suitable for the customers for bike rides

Weekday: The demand is not significantly impacted by this feature. The demand is almost same on all the weekdays and average demand is slightly higher on Saturdays

Workingday: This feature doesn’t seem to have any impact on the demand for rental bikes

Weathersit: The demand for rental bikes is highly impacted by weather. It is seen that the demand is hight when the weather is clear and partially cloudy and the demand is lowest when there is light precipitation

---

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit)

When you create dummy variables for a categorical feature, each unique category gets its own column with binary (0 or 1) encoding. If we include all dummy variables, one column becomes redundant because its value can be perfectly predicted by the others. This redundancy leads to multicollinearity, which can negatively affect regression models or statistical inferences.

Multicollinearity Problems: Some models, like linear regression, don’t work well if the inputs are too closely related (multicollinear). Keeping all dummy variables can cause this issue, making the model results less reliable.

Easier to Understand: Dropping one dummy column makes the model simpler to read. The remaining dummy variables are compared to the one you dropped, which becomes the baseline.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

Total Marks: 1 mark (Do not edit)

Answer: <Your answer for Question 3 goes below this line> (Do not edit)

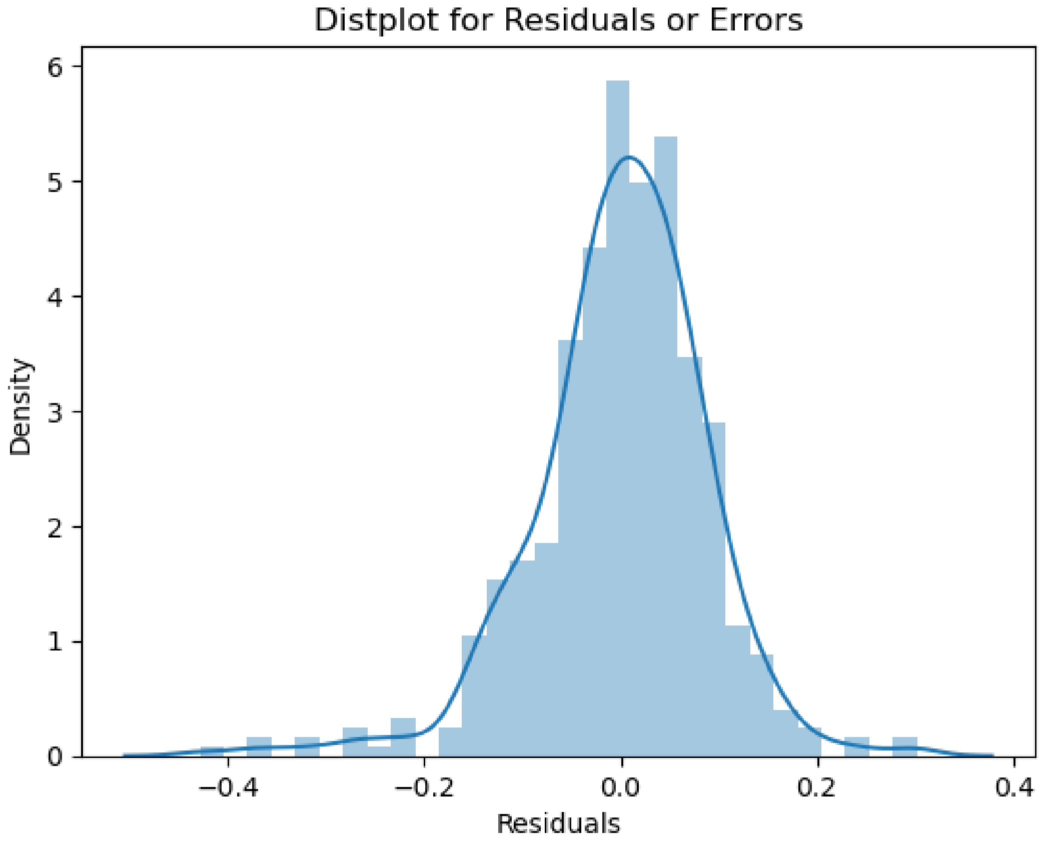
The highest correlation is seen between 'casual', 'registered' and 'cnt'. But since 'cnt' is the sum of both 'casual' and 'registered', we cannot consider this.

Otherwise, 'temp' and 'atemp' are seen to be highly correlated with the target variable 'cnt'

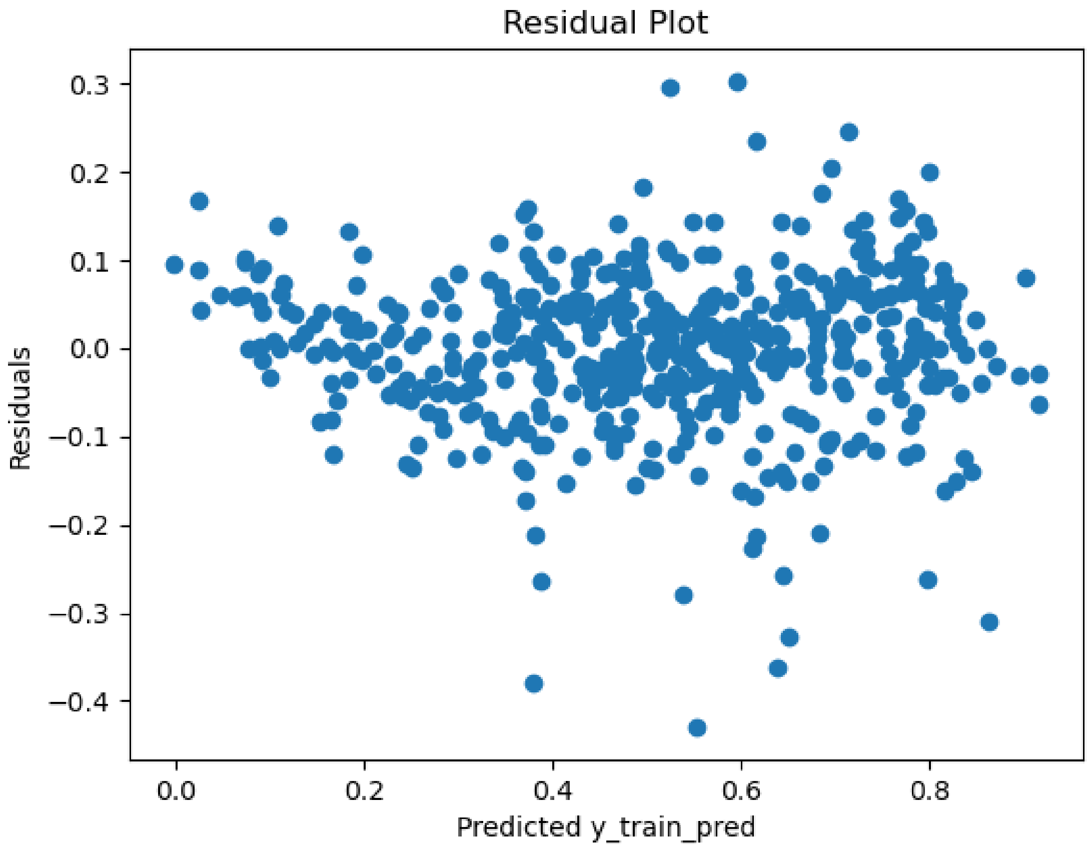
Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

The assumptions of Linear Regression are validated by calculating the residuals and plotting various graphs for the same.

1. Distribution plot of residuals is as follows. This shows that the errors are mostly centered around zero and they are normally distributed



2. Scatter plot of residuals is as follows which shows that the spread of the errors are same across the model which means that the model is consistent and reliable



Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

Based on the model summary, the equation for the target variable cnt can be written as follows

cnt = 0.2426
+ 0.2317 * yr
+ 0.0511 * workingday
+ 0.4763 * atemp
- 0.1518 * hum
- 0.1769 * windspeed
+ 0.0575 * Mnth_Aug
- 0.0525 * Mnth_Dec
- 0.0500 * Mnth_Feb
- 0.0792 * Mnth_Jan
- 0.0463 * Mnth_Nov
+ 0.1068 * Mnth_Sep
+ 0.0780 * Season_Summer
+ 0.1357 * Season_Winter
+ 0.0593 * weekday_Saturday
- 0.2487 * weathersit_Light Precipitation

- 0.0589 * weathersit_Mist + Cloud

So, we can conclude that the top three contributing features are 'year', 'atemp' and 'windspeed'

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 6 goes here>
Linear Regression: It is one of the simplest and most widely used algorithm in machine learning and statistics. It is used to find the relationship between the **Target Variable** and **one or more independent variables.** In simple terms, it tries to draw a straight line that best fits the data points

The simple equation for Linear Regression is **Y = mX+b**
Where Y is target variable, m is the slope of the line (Coefficient) and X is independent variable and b is the constant or intercept (The value of Y when x is 0)

When we have multiple independent variables/features, the equation is

$Y = b + m_1X_1+m_2X_2+......+m_nX_n$

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 7 goes here>

Anscombe's Four Data Sets are a set of four data sets that statistician Francis Anscombe created in 1973. These data sets have nearly the same statistical properties (such as mean, variance, and correlation), but they look different. Very different when plotted as a graph...

The idea behind all four is to emphasize the importance of data visualization. and do not rely on summary statistics alone to understand the data.

Each dataset contains 11 pairs (rows) of x,y values. Though each dataset have different values for x, y, They share the below statistic properties for all the datasets

1. Mean of x : 9
2. Mean of y : 7.5
3. Variance of x : 11
4. Variance of y : ~ 4.1
5. Correlation between x and y : 0.816
6. Linear Regression equation : y = 3+0.5x

Though all the above statistic properties are same for all the datasets, When we visualize them using the charts, it is clearly seen that, each dataset follows a unique pattern.

Dataset 1 : follows nearly linear pattern
Dataset 2: forms a curved pattern.
Dataset 3: only one point is an extreme outlier, and this influences the regression line
Dataset 4 : all points are vertically aligned except one outlier but this outlier decides the regression line

Significance of Anscombe's Quartet is:

1. It is always recommended to visualize the data rather than just analyzing the statistical data as they do not reveal the underlying pattern and structure.
2. Always keep the context in mind while performing the linear regression
3. Avoid relying completely on statistics as each dataset requires different interpretation despite having the same summary statistics.

Example:
Imagine four students who receive the same average score on a test. On paper they seem to perform equally well, but if you examine the scores in detail:

One student will be able to achieve average scores consistently.
Another person may be good at one thing and fail at another.
Third, it can be gradually getting better over time.
The last one can work erratically without pattern.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 8 goes here>
Pearson's R is also known as Pearson's correlation coefficient. It is a statistical measure used to evaluate the strength and direction of a linear relationship between two variables. They are represented by values from -1 to +1:

+1 indicates a perfect positive relationship. (When one variable increases the other one increased.)
-1 indicates perfect negative correlation. (When one variable increases Another variable decreased)
0 indicates no relationship. (There is no linear relationship between the variables.)

Formula for Pearson's R is as below

$$\rho_{X,Y} = \frac{\mathbb{E}[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

Where **µX** – mean of X,
Where **µY** – mean of Y
Where **σX** – Standard Deviation of X
Where **σY** – Standard Deviation of Y

Example of Pearson’s Correlation can be shown as follows.

No. of hours studied	Marks in the exam
20	45
30	52
50	60
60	68
72	78
80	85
85	95

Pearson’s R value for the above data **0.98** shows a strong correlation between the two variables.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 9 goes here>
Scaling: Scaling is a data processing technique used in machine learning to adjust the range of a data feature so that the features are proportionate to each other.

Scaling is performed to achieve the below :
Consistency in features: Features with larger values can dominate the optimization process and affect the accuracy of the model. Scaling ensures that all features are treated equally.

Improved model performance: Many machine learning algorithms (e.g., support vector machines) are sensitive to the size of the input feature
.
Faster Convergence: Gradient-based algorithms converge faster when features are scaled. Because it avoids fluctuations caused by measuring different properties.

Compatibility with distance-based algorithms: Algorithms such as KNN and K-means rely on distance measures. This is greatly influenced by the size of the feature.

	Normalized Scaling	Standardized Scaling
Definition	Rescales data to a fixed range, typically in [0, 1] range	Centers the data around the mean and scales it to have a standard deviation of 1.
Usage	When the data needs to be bound to a specific range (e.g., image pixel values)	When the algorithm assumes a Gaussian distribution or is sensitive to variance

		(e.g., logistic regression).
Sensitive to Outliers	Yes, because the range depends on the min and max values.	Less sensitive, as it uses the mean and standard deviation.
Example Algorithms	KNN, K Means	Logistic Regression, SVM, PCA
Output Range	Fixed range (e.g., [0, 1]).	Not fixed; depends on the distribution of the data.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 10 goes here>
When there is perfect multicollinearity between the independent variables in a regression model. The variance inflation rate (VIF) can be infinite.

When one independent variable can be expressed as exact linear combination of other independent variables, R^2_J value will be 1.
When $R^2_J = 1$,

$VIF_J = 1/ (1- R^2_J) = 1/(1-1) = 1/0 = \infty$

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 11 goes here>
A Q-Q plot (quantile-quantile plot) is a graphical tool used to compare the distribution of a data set with the theoretical distribution. It plots the quantiles of the data set against the quantiles of the theoretical normal distribution.

Components of a Q-Q plot:
The X-axis is the scale of the theoretical distribution (such as the standard normal distribution).
Y axis: Volume of sample data.
Diagonal line (45-degree line): Represents where the drop would be if the data followed the theoretical distribution exactly.

Using the QQ plot in linear regression:

The main assumption in linear regression is that the residuals (errors) are normally distributed. A Q-Q plot helps to evaluate this assumption.

Assessment of normality of residues:

If the remainder were normally distributed, the points in the Q-Q diagram would be aligned along the 45-degree line.
Deviations from this line indicate deviations from normality, such as skewness or heavy tailing.
Identifying outliers or non-normative patterns:

Points further away from the line indicate possible outliers or specific patterns (for example, a curved pattern may indicate skewness). while an S-shaped deviation indicates a heavy tail).
Checking model assumptions:

in linear regression the assumption of normality of the residuals is important for valid hypothesis testing and confidence intervals. Q-Q plots help detect violations of this assumption.
Importance of Q-Q Plot:
Imaging Diagnostic Tools: Provides a visual and easy-to-use method of checking for normality. without relying on statistical tests alone, such as the Shapiro-Wilk test
Identify required transformations: If normality is violated, a Q-Q plot can suggest the type of transformation (such as logarithmic or square root) to be applied to the data or residuals.