Linear Regression Subjective Questions

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)

Bike rent increases on

1. temperature between 25-35 degrees.
2. 45-80 humidity.
3. 7-17 wind speed
4. Non holidays
5. working days
6. Clear weather
7. Fall season

Renting on bikes heavily depends on the upper factors but few of the factors are correlated.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
To make use of categorical variables in Linear Regression, we generally create dummies and if the categorical variable has more than 2 unique values then we remove the first using drop_first=True as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.
Example: Let's support season has 4 categorical values

1. Summer
2. Winter
3. Fall
4. Monson

Now if we create dummies for those without drop_first=True then 4 new columns will be generated

Summer	Winter	Fall	Monsoon	Indication
1	0	0	0	Summer season
0	1	0	0	Winter season
0	0	1	0	Fall season

0	0	0	1	Monsoon season

But to be very specific we do not need 4 columns to materialize the data.
Example:

Summer	Winter	Fall	Indication
1	0	0	Summer season
0	1	0	Winter season
0	0	1	Fall season
0	0	0	Monsoon season

So if all the columns have 0 value then it means Monsoon only so we required actually n-1 columns for a categorical column having n unique values. That is why **drop_first=True** is important

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

temp/atemp has the highest correlation with cnt.

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

Linear regression assumptions are:

1. Linear relationship between X and Y : Once I finalized the model, I checked the correlation between the cnt(dependent variable) and other features(independent variables) and found out there exists some kind of correlation between them.
2. Error terms are normally distributed (not X, Y) : Plotted the residuals/errors and verified these are normally distributed.
3. Error terms are independent of each other : Plotted a line graph of y_train_pred vs residuals and tried to find out any correlation between them present or not. I do not see any correlation.
4. Error terms have constant variance (homoscedasticity): Plotted y_train_pred vs residuals and find out the regression line to verify the constant variance.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

1. Atemp
2. yr
3. Snow weather

General Subjective Questions

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

Linear Regression algorithm is a machine learning algorithm which is used to define the relationship between one dependent variable and one or more independent variables.
Dependent Variables: The variables which we try to predict
Independent Variables: The variables that we believe to have influence on dependent variables.

Key concept:
It assumes that there is a linear relation between dependent and independent variables.
Equation: y=ax1+bx2+.....+mxn+z
Where
z= y or the dependent variable value when all the independent variables are 0.
a= slope of x1
b= slope of x2…….

Goal:
The goal is to find out the best fit line to minimize the overall error between the actual value and the predicted value. This is such a line where residual sum of squares(RSS) or error terms are minimal. There are different approaches present to achieve RSS close to 0.

1. Differential approach
2. Gradient Descent(Preferable)

Assumptions:

1. Linear relationship between X and Y
2. Error terms are normally distributed (not X, Y)
3. Error terms are independent of each other

4. Error terms have constant variance (homoscedasticity)

Usage:

1. Economics (predicting sales or demand)
2. Finance (predicting stock prices or interest rates)
3. Healthcare (predicting patient outcomes)
4. Many other fields where predicting a numerical outcome is important.

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

Anscombe's quartet is a group of datasets (x, y) that have the same mean, standard deviation, and regression line, but which are qualitatively different.These datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.
The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

Importance:
Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

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

Pearson correlation coefficient or Pearson's R, is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables i.e. it is used to measure how strong a relationship is between two variables.

R Value and corresponding relationship:
The r value ranges between -1 and 1, where:
1 indicates a perfect positive linear relationship,
-1 indicates a perfect negative linear relationship, and
0 indicates no linear relationship between the variables.

Formula:
$r = n(\sum xy) - (\sum x)(\sum y) / \sqrt{[n\sum x^2-(\sum x)^2][n\sum y^2-(\sum y)^2}$

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

Feature scaling a data preprocessing technique, a step often taken before applying machine learning algorithms. It involves transforming the features (independent variables) in your dataset to bring them into a similar range or scale.

Why used:
Equalizes Feature Importance: Without scaling, features with larger magnitudes (e.g., salary in dollars) can dominate features with smaller magnitudes (e.g., age in years) in the model.

Improves Algorithm Performance: Many machine learning algorithms rely on distance calculations (e.g., k-nearest neighbors, support vector machines). Scaling prevents features with larger ranges from disproportionately affecting these distances, leading to better model performance.

Faster Convergence: Algorithms like gradient descent can converge faster when features are scaled.

Difference between normalized scaling and standardized scaling

Normalized scaling(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.
$X\_scaled = (X - X\_min) / (X\_max - X\_min)$

Standardized scaling: The variables are scaled in such a way that their mean is zero and standard deviation is one.
$X\_scaled = (X - \mu) / \sigma$

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

VIF is used to detect and measure multicollinearity, which is a situation where two or more predictor variables (independent variables) are highly correlated with each other in the model.

A VIF value of infinity indicates **perfect multicollinearity**. This happens because one predictor variable can be perfectly predicted from the other predictor variables. In other words, the information contained in that variable is entirely redundant and doesn't add any new information to the model.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

A Q-Q (Quantile-Quantile) plot is a visual tool used in statistics to assess whether a dataset follows a particular theoretical distribution (like a normal distribution) or to compare the distributions of two datasets.

How Q-Q Plots Work

Quantiles: Q-Q plots work by plotting the quantiles of one dataset against the quantiles of another. Quantiles are values that divide a dataset into equal portions. For example, the median is the 0.5 quantile (50th percentile), splitting the data in half.
Comparison: The Q-Q plot then compares these quantiles:
If the two datasets come from the same distribution, the points in the Q-Q plot will roughly form a straight line.
If the distributions are different, the points will deviate from a straight line, with the shape of the deviation providing clues about how the distributions differ.

Use and Importance:

1. QQ plots are very useful to assess whether the assumptions underlying the linear regression model are met, specifically the assumption of normality of residuals.
2. Compare Model Fits: Q-Q plots can be used to compare the goodness of fit of different linear regression models by comparing the straightness of their respective Q-Q plots.