Assignment-based Subjective Questions and its Answers

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

   Demand for bikes increases in fall which is the highest and drops in spring.
   Demand for bikes in the year 2019 is higher than in the year 2018.
   Demand for bikes goes high during the months May to October.
   Bike demand is high if weather is clear or with mist cloudy while it is low when there is light rain or light snow.
   The demand of bikes is almost similar throughout the weekdays.
   Demand for the bikes doesn't change if day is working day or not.

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

   When we convert a categorical variable with k levels into dummy variables, we end up with k binary columns. So all these k dummy variables in the regression model, will introduce multicollinearity. This is because the sum of all k dummy variables for a given categorical variable will always be 1. This introduces perfect collinearity, making it impossible to estimate the coefficients for these variables accurately.

   Hence by setting drop_first=True, you create k−1 dummy variables instead of k. This eliminates the redundancy and prevents the issue of perfect multicollinearity.

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

   Temperature('temp') and feeling temperature('atemp') in Celsius has highest correlation of 0.62 with the target variable 'cnt'

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

   Linearity assumption is validated using the pair plots with target and predictors.
   Independence of Errors: Seeing the summary statistics of Durbin Watson . This statistical test assesses the presence of autocorrelation in residuals. The value of the Durbin-Watson statistic ranges from 0 to 4. A value close to 2 suggests no autocorrelation.
   Normality of Errors: The residuals were normally distributed, this could be seen in the histplot with residuals(residuals) vs fitted values(y_pred)

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

Top three features contributing significantly towards explaining the demand of the shared bikes are temp, weathersit_3(Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) and month_9(September)

General Subjective Questions

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

Linear regression algorithm is a statistical technique used to model the relationship between the dependent (target) and one or more independent (predictors) variables. It uses the straight line formula .

$y=\beta 0+\beta 1x1+\beta 2x2+\cdots+\beta nxn+\epsilon$

where:

- y is the dependent variable (target).
- x1,x2,…,xn are the independent variables (features).
- β0 is the intercept (constant term).
- β1,β2,…,βn are the coefficients (weights) that represent the impact of each feature on the target variable.
- ϵ is the error term (residual), representing the difference between the observed and predicted values.

The goal of linear regression is to find the best-fitting line (or hyperplane in the case of multiple features) that minimizes the difference between the predicted values and the actual values. This is done by estimating the coefficients β0,β1,…,βn such that the error is minimized.

The most common cost function used in linear regression is the Mean Squared Error (MSE). It measures the average of the squares of the errors, which are the differences between the observed values and the predicted values:

MSE=1/m(∑i=1(yi−y^i)^2)

where:

- m is the number of observations.
- yi is the actual value of the i-th observation.
- y^i is the predicted value for the i-th observation.

After training the model, its performance is evaluated using metrics such as:

- R-squared (R^2): Indicates the proportion of variance in the dependent variable that is predictable from the independent variables. R-squared values range from 0 to 1, where a higher value indicates a better fit.
- Mean Absolute Error (MAE): Measures the average absolute difference between the observed and predicted values.
- Root Mean Squared Error (RMSE): Measures the square root of the average of the squared differences between observed and predicted values.

Linear regression makes several key assumptions:

- Linearity: The relationship between the dependent and independent variables is linear.
- Independence: The residuals (errors) are independent.
- Homoscedasticity: The residuals have constant variance.
- Normality: The residuals are normally distributed.

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

Anscombe's Quartet is designed to show that summary statistics (like mean, variance, and correlation) alone can be misleading and do not capture the underlying structure of data. Visualizing data through plots is crucial for understanding its distribution and potential anomalies.

The quartet consists of four datasets, each with 11 data points. For each dataset, the following statistics are nearly identical:

Mean of x and y
Variance of x and y
Correlation between x and y
Regression line equation for predicting y from x

Despite these similarities in summary statistics, the datasets exhibit different patterns when plotted.

Dataset 1: Linear Relationship
Description: This dataset shows a simple linear relationship between x and y.
Plot Characteristics: Points lie on a straight line with some scattered around it.
Dataset 2: Non-linear Relationship
Description: This dataset features a non-linear relationship between x and y. The data forms a curved shape.
Plot Characteristics: Points form a parabolic curve, demonstrating a clear non-linear pattern.
Dataset 3: Outlier Impact
Description: In this dataset, there is a single outlier that significantly affects the regression line.
Plot Characteristics: Most points align linearly, but one extreme outlier skews the regression line.
Dataset 4: Vertical Line
Description: This dataset shows a vertical distribution of x values, where x is nearly constant but y varies.
Plot Characteristics: Points lie on a vertical line with variability in y, indicating high variability in y but no change in x.

When plotted, the datasets reveal:

Dataset 1: A perfect straight line with minor scatter.
Dataset 2: A curve, indicating a quadratic relationship.
Dataset 3: A linear pattern with an outlier drastically affecting the slope.
Dataset 4: A vertical line showing no apparent relationship between x and y.

Anscombe's Quartet highlights that while summary statistics can be identical, the underlying data can be fundamentally different. Visualizing data helps uncover these differences. It stresses the importance of checking assumptions like linearity before applying linear models. It shows how outliers can significantly impact regression analysis and summary statistics.

3. What is Pearson’s R? (3 marks)

The strength of any linear regression model can be assessed using various metrics. These metrics usually provide a measure of how well the observed outcomes are being replicated by the model, based on the proportion of total variation of outcomes explained by the model. The various metrics are,

1. Coefficient of Determination or R-Squared (R2)
2. Root Mean Squared Error (RSME) and Residual Standard Error (RSE)

Pearson’s R, also known as Pearson's correlation coefficient, is basically the correlation value between the variables. R-Squared is a number which explains what portion of the given data variation is explained by the developed model. It is basically the square of the Pearson’s R correlation value between the variables. It always takes a value between 0 & 1 . Overall, the higher the R-squared, the better the model fits the data. Mathematically it can be represented as

$R^2 = 1-(RSS/TSS)$

- 1 indicates a perfect positive linear relationship,
- -1 indicates a perfect negative linear relationship,
- 0 indicates no linear relationship.

Pearson's R is defined as:

$r=Cov(X,Y)/\sigma X\sigma Yr$

where:

- Cov(X, Y) is the covariance between variables X and Y.
- $\sigma X$ and $\sigma Y$ are the standard deviations of X and Y, respectively.

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

When there are many independent variables in a model, a lot of them might be on very different scales. This leads to a model with very weird coefficients, which are difficult to interpret. Thus, one needs to scale the features for ease of interpretation of the coefficients as well as for the faster convergence of gradient descent methods. One can scale the features using the following methods.

Normalized scaling, also known as min-max scaling, transforms features to a specific range, typically [0, 1]. It rescales the data so that the minimum and maximum values of the feature map to 0 and 1, respectively.

Standardized scaling, also known as z-score normalization, transforms features to have a mean of 0 and a standard deviation of 1. It centers the data around zero and scales based on the standard deviation.

Choosing between normalization and standardization depends on the requirements of machine learning algorithm and the characteristics of the data.

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

Pairwise correlations may not always be useful as it is possible that just one variable might not be able to completely explain some other variable but some of the variables combined might be able to do that. Thus, to check these sorts of relations between variables, one can use VIF. Variance Inflation Factor (VIF) is a measure used to detect multicollinearity among predictor variables in regression analysis. A VIF value of infinity or extremely high indicates a problem with multicollinearity, where predictor variables are highly correlated with each other.

VIF is given by,

VIF = 1/(1-Ri^2)

where, i refers to the i variable which is being represented as a linear combination of the rest of the independent variables.

The common heuristic followed for the VIF values is,

a. V IF > 10 indicates definite removal of the variable.
b. V IF > 5 indicates variable needs inspection.
c. V IF < 5 indicates the variable is good to go.
Once multicollinearity has been detected in the dataset then this can be dealt using the following methods.

1. Highly correlated variables can be dropped.
2. Business Interpretable variables can be picked up.
3. New interpretable features can be derived using the correlated variables.
4. Variable transformations can be done using PCA (Principal Component Analysis) or PLS (Partial Least Squares).

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 plot compares the quantiles of the observed data with the quantiles of a theoretical distribution. If the data follows the theoretical distribution, the points on the Q-Q plot will lie approximately along a straight line.
How It Works:
Quantiles: Divide the data into intervals of equal probability.

Plotting: Plot the quantiles of the sample data on the y-axis against the quantiles of the theoretical distribution on the x-axis.
Straight Line: If the points fall approximately along a straight line (usually the 45-degree line), it indicates that the data follows the theoretical distribution.

Python code for Q-Q plot :

```
import numpy as np
import scipy.stats as stats
 import matplotlib.pyplot as plt
 # Generate a sample data (e.g., normal distribution)
data = np.random.normal(loc=0, scale=1, size=1000)
# Q-Q plot
stats.probplot(data, dist="norm", plot=plt)
plt.title('Q-Q Plot')
plt.show()
```