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)

Ans:

- Felling temperature is affecting the number of renting bikes as it is good at good temperature.
- During weathersit is Mist or Light Snow number of bike rented is reduced
- Bikes are demanded well during the whole week.
- Compared to 2018 demand has increased in 2019
- During Spring demand is very low compared to another season

---

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

(2 mark)

Ans:

To avoiding multicollinearity between categorical variables, two dummy columns create the third reference column that can explain the details of 3 dataset as shown below hence n-1 formula is justified

Eg. Summer '00', Winter'01', Rainy'10'

Season	Col1	Col2
Summer	0	0
Winter	0	1
Rainy	1	0

---

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

Ans:

Temp and Atemp as highest correlation

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

Ans:

- Checking R-squared value with target variable getting constant
- Checking Adjusted R-Squared value for target variable constant
- Checking P-value less than 0.05
- Checking VIF less than 5
- Divided data into 2parts, training set and test set and validating both data set to nearby R-squared value

This confirms for the homoscedasticity, linearity of the data. Below is the example tested for Count of rented bikes affected by dependent variable with above points:

CNT = 0.247 **X** Year **+** 0.254 **X** Season Summer **+** 0.314 **X** Season_Fall **+** 0.227 **X** Season Winter **-**0.093 **X** Holiday **-**0.176 **X** WindSpeed **-**0.087 **X** Weathersit 2

- 0.291 **X** Weathersit 3

---

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

Ans:

- During season fall, one day increment may affect count by 0.314 units.
- During summer season, one day increment may affect count by 0.254 units.
- Every Year demand is increasing by 0.247

==========

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**
   - Ans:
   - The line drawn among two points is denoted by Y= m*X +c
   - So to draw a bestfit line from multiple points with the equal distance is denoted by slope βn and the predicted points to get the intercept for straight line β0
   - The Simple linear regression formula becomes y=β0+β1X for one independent and one dependent variable

- The multilinear regression formula becomes y=β0+β1X1+β2X2+.....βnXn for for more independent and one dependent variable with a constant value that rectifies the predicted data on best fit line.
- This dat points are confirmed using R – squared value = (Yi-Ypred)^2
- For R-squared =1 is best fitted.

---

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

Ans:

- In Anscombe's quartet four datasets, having means, variance, R-squared, correlations, and linear regression lines.
- This data set when seen on scattered plot have different representation
- The datasets were created by the statistician Francis Anscombe in 1973
- It is use to show that visualization can addon the correctness of summarize data

Steps to get **Anscombe's quartet regression:**

Find the Descriptive Statistical Properties for the all four Dataset

- Find mean for x and y for all four datasets.
- Find standard deviations for x and y for all four datasets.
- Find correlations with their corresponding pair of each dataset.
- Find slope and intercept for each dataset.
- Find R-square for each dataset.
    - To find R-square first find residual sum of square error and Total sum of square error
- Create a statistical summary by using all these data and print it.

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

Ans:

- Pearson r correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, r which is also called Pearson's r
- Pearson r has the range of value for -1 to 1.
- **Positive r**: Indicates that as one variable increases, the other variable tends to increase as well.

- **Negative r**: Indicates that as one variable increases, the other variable tends to decrease.
- **Magnitude**: The closer r is to −1 or 1, the stronger the linear relationship. A value close to 000 indicates a weak linear relationship.

When to use the Pearson correlation coefficient:

- The Pearson correlation coefficient (r) is one of several correlation coefficients that you need to choose between when you want to measure a correlation. The Pearson correlation coefficient is a good choice when all of the following are true:
- Both variables are quantitative: You will need to use a different method if either of the variables is qualitative.
- The variables are normally distributed: You can create a histogram of each variable to verify whether the distributions are approximately normal. It's not a problem if the variables are a little non-normal.
- The data have no outliers: Outliers are observations that don't follow the same patterns as the rest of the data. A scatterplot is one way to check for outliers—look for points that are far away from the others.
- The relationship is linear: "Linear" means that the relationship between the two variables can be described reasonably well by a straight line. You can use a scatterplot to check whether the relationship between two variables is linear.

---

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

Ans:

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range that helps in speeding up the calculations in an algorithm.

- Scaling helps to comparable scale and have comparable ranges called **normalization**. Larger scale features may dominate the learning process and have an excessive impact on the outcomes.
- Algorithm performance improvement
- Preventing numerical instability:

- Scaling features makes ensuring that each characteristic is given the same consideration during the learning process called as **standardization** of that which have lesser impact for outliers.

---

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

Ans:

VIFi=1/(1−Ri)^2

R-squared value ranges from 0 to 1.

When the value of R- squared =1 that means the model predicts 100% of the relationship with variable, that time VIF value becomes infinite this is the case when one variable is mostly colinear with another one.

This can be rectified by considering only one variable among both.

---

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

Ans:

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

If the points lie approximately along the reference line, the residuals are likely normally distributed.

If the points form an S-shape, it might indicate that the data has heavier or lighter tails than the normal distribution.

A curve or systematic deviation suggests non-normality, such as skewness or kurtosis issues in the residuals.
