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)

Answer: **Cnt** is the dependent variable and we have categorical as '**weathersit'** which is **negatively** correlated with the Cnt by **0.3**, '**season'** which is **positively** correlated with **0.4** correlation, '**mnth'** this is also **positively** correlated with **0.28** correlation and '**weekday'** this **positively** correlated with **0.068** correlation.

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

Answer: The drop_first parameter specifies whether or not you want to drop the first category of the categorical variable you're encoding. By default, this is set to **drop_first = False**. This will cause get_dummies to create one dummy variable for every level of the input categorical variable.

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

Answer: There is high correlation between '**temp'** and '**atemp'** with the target variable '**cnt'**. The correlation value is '**0.63'**.

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

Answer: Hypothesis testing states that: H0:B1=B2=...=Bn=0 H1: at least one Bi!=0. We can see that all of our variables are not equal to 0, which means we can reject the null hypothesis.

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

Answer: As per our final Model, the top 3 predictor variables that influences the bike booking are:

- Temperature (temp) - A coefficient value of '0.575942' indicated that a unit increase in temp variable increases the bike hire numbers by 0.575942 units.
- Weather Situation 3 (weathersit_3) - A coefficient value of '-0.253650' indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire numbers by 0.253650 units.
- Year (yr) - A coefficient value of '0.234668' indicated that a unit increase in yr variable increases the bike hire numbers by 0.234668 units.

General Subjective Questions

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

Answer: Linear Regression algorithm is a **Machine Learning** algorithm. It comes **under supervised learning** algorithms. The model is able to predict the value of dependent/ target variable, using the independent variable(s). It is mostly used for finding out the relationship between variables and forecasting. When one independent variable is used to build the model, it is called **Simple Linear Regression** model and when more than one independent variable is used to build the model it is called **Multiple Linear Regression** model.

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

Answer: Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (X,Y) points. They were constructed in 1973 by the statistician **Francis Anscombe** to demonstrate both the **importance of graphing** data before analyzing it and the **effect of outliers** on statistical properties.

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

Answer: The Pearson correlation coefficient (r) is the most common way of measuring **a linear correlation**. It is a number between **–1 and 1** that measures the **strength and direction of the relationship** between two variables. When one variable changes, the other variable changes in the same direction.

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

Answer: Scaling is a process in **data-preprocessing**, which is applied to numerical(continuous) variables to normalize the data in a particular range.

It also helps in faster convergence of **gradient decent** algorithm, which gives us the minimum cost function and improving the model.

When parameters present in our model tend to be of different units, there is a difference in magnitude for the given, which might lead to bias in our model and in turn lead to model being incorrect.

Normalized scaling - scales the values from the range **0 and 1**. We use **MinMaxScaler** from **sklearn.preprocessing** to implement this in Python. This is basically used in deep learning applications such as image processing where the model is based on the pixel levels (non-negative).

Formula – $(x-x_{min})/(X_{max}-X_{min})$

Standardized scaling – also called as **Z-Score Normalization** is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as Z-score.

It brings all of the data in standard normal distribution, which has a **mean of 0(μ)** and **standard deviation of 1(σ)**. This is used in machine learning models such as weather forecasting and pricing prediction, where the model can have both positive and negative coefficients.

Formula – $(x- \mu(x))/ \sigma(x)$

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

Answer: The formula for **VIF is $1/(1-R_i^2)$**. VIF value can be infinite only when R_i^2 values becomes **1** (Denominator value becomes 0). It implies that the corresponding variable can be perfectly expressed as linear combination of other variables.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**

(3 marks)

Answer: The quantile-quantile plot is a graphical method for determining whether two samples of data came from the same population or not. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set