Subjective Test Answers

Chaitanya Namburu

**Assignment-based Subjective Questions**

***Q1.*** *From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?*

Answer:

Season: The count is higher in Summer and Fall, compared to Winter and Spring. It seems like hot seasons are preferred by the customers.

Holiday: The average count is higher on a non-holiday, compared to a holiday.

Weather Situation: The bikes were taken more when the weather was calm and clear, compared to misty conditions. It was followed by light snow and the bikes were literally not taken during heavy rains and thunderstorms.

Month: The count gradually increased from the start of the year to around Sep/Oct and decreased towards the end of the year. It coincided with our understanding of seasons. The busiest months were Jul-Oct.

Weekday: Thursday, Friday, Saturday & Sunday seemed to have higher count compared to the beginning of the week.

**Q2.** *Why is it important to use drop\_first=True during dummy variable creation?*

Answer:

“drop\_first” option is important to avoid multicollinearity among the variables. This is used for categorical variables.

We need k-1 dummy variables to represent a categorical variable with k levels.

Example:

If there are 4 levels in a categorical variable (e.g., seasons = summer, winter, autumn, rainy) we need only 3 variables to represent the mentioned categorical variable. If we don’t drop the first column, it leads to multicollinearity among the variables.

**Q3.** *Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?*

Answer:

It can be observed from the pair-plot that Count has the highest correlation with “temp” (Temperature) variable.

**Q4.** *How did you validate the assumptions of Linear Regression after building the model on the training set?*

Answer:

I have validated the assumptions of the Linear Regression model by ensuring the following:

1. The residuals (error terms) should have a normal distribution. I have graphed out the residuals and observed that they are approximately in a normal distribution.
2. Homoscedasticity – I have ensured that there is no pattern among the residuals.
3. There was no autocorrelation among the residuals.
4. There was no multicollinearity among the variables. The VIF of the variables in the final model was less as per the acceptable standards.
5. Linear relationship – there was linear relationship between the variables and the dependent variable.

**Q5.** *Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?*

Answer:

The top features contributing significantly explaining the demand of bikes are:

* Temperature
* Weather Situation 3 (Light snow)
* Year
* Windspeed

**General Subjective Questions**

**Q1. *Explain the linear regression algorithm in detail.***

**Answer:**

**Linear regression is a machine learning technique to predict unknown values of a target variable by studying known values of the target variable with respect to a few other independent variables. The technique involves the development of a model by using a training set of data which establishes a linear relationship the independent variables and the dependent variable. Then we use the model to determine the values of the target variable for other values of independent variables.**

**It is classified as Simple Linear Regression and Multiple Linear Regression based on the number of independent variables.**

**The overall goal is to mathematically establish a linear relation using the formula:**

**Y = mx + c**

**Where Y = target variable and x = independent variable and c = constant**

**Using the training data properly, we try to find the coefficients and constant value to best fit the available data. We use various techniques like gradient descent to find the optimal coefficients.**

**The generated model should satisfy some assumptions to be an efficient model:**

* **Linearity**
* **Absence of multicollinearity**
* **Absence of Autocorrelation**
* **Normality of residuals**
* **Homoscedasticity**

**Q2.** Explain the Anscombe’s quartet in detail.

Answer:

Developed by the French statistician Anscombe, the quartet consists of four datasets which have similar descriptive statistics but are different in the nature of the datasets. The 4 datasets appear very different in