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?

The **dependent variable** in the dataset is the count of total rental bikes, given by **"cnt"**. This variable is a **sum of** the count of casual users, given by **"casual"**, and count of registered users, given by **"registered"**.

The categorical variable in the dataset are:

- season
- year
- month
- holiday
- weekday
- workingday
- weather situation

From the analysis, I could infer the following effect of the categorical variables on the dependent variable:

1. The count of total, casual and registered users are lowest in spring compared to other seasons. The count of total and registered users increases in winter.

2. The count of both types of users were more in 2019.

3. The count of casual users is low in Nov, Dec, Jan, Feb. It increases in June, July, and August. The count of registered users is highest in September. Total users are also highest in September.

4. Although the count of registered users are less during holidays and the count of casual users are more during holidays, the variable had a high VIF, implying that it was correlated with other independent variables and possibly the effect was mediated through those independent variables (i.e., weekday, and weekend).

5. The number of casual users is more during weekends and registered users are less in weekends compared to the other days in a week.

6. I did not find any independent effect of workingday on the count of users. The weekdays had no significant effect on total users (possibly nullified due to opposite effect on casual and registered users).

7. Worsening of weather affects the counts of both types of users. Mist (Type 2) and Snow (Type 3) weather situations reduces registered users count and snow (Type 3) also reduces casual users count. Overall Mist (Type 2) and Snow (Type 3) weather situations reduces user count.

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

n classes of a categorical variable can be represented by (n-1) binary variables. Using drop_first = True, drops the first of the n classes, resulting in (n-1) dummy variables. This approach has two benefits:

1. This obviates the creation of extra columns/ redundant columns;
2. It obviates the correlation created among dummy variables.

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

Feeling temperature (atemp) has highest correlations with the target variable. Temperature (temp) has nearly equal correlation with the target variable.

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

The assumptions that I checked included:

1. error terms are normally distributed.
2. error terms are independent of each other.
3. error terms have constant variance.
4. multicollinearity is absent among independent variables.

Please check plots in code snippets 98 - 103 for the assumption check.

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

Feeling temperature, weather situation, and Year contributed most significantly to the bike selling. The absolute value of the beta coefficients were highest for these features.

General Subjective Questions

---

1. Explain the linear regression algorithm in detail.

Concept: Linear regression analysis is used to explain the relationship between two or more variables and predict the value of one variable using the value of one or more variable(s) using the formula of a straight line or surface. The variable whose value is predicted is called the dependent variable and the variable whose values are used to predict the value of the dependent variable are called the independent variable.

Model: The generic model is an additive one, with the linear equation and the simple linear regression calculator uses least square method to find the best fit line. The standard line equation for SLR is : Y = b0 + b1x. In the least square method, sum of squares of the residual for each data point is measured and minimized to find the line of best fit.

Assessment: The strengthe of linear regression can be determined using co-efficient of determination or R-squared value and residual standard error.

Assumptions: For linear regression, error terms must be normally distributed, independent to each other and have constant variance. For MLR, in addition to the above assumptions, the indep variables should not be correlated with each other (multicolinearity problem).

Execution: Linear regression can be executed using the ScikitLearn and StatsModel libraries of Python. SM provides robust statistical methods. We use recursive feature elimination to reduce the features and variance inflation factor to remove correlated features.

2. Explain the Anscombe’s quartet in detail.

Anscombe's quartet consists of four distinct dataset, which have nearly identical simple decriptive statistics. However, the appear different when plotted. This is important from the perspective that the regression model can be same for all these different sets of data. That's why it is critical to assess the assumptions of linear regression. Linear Regression should be used for data with linear relationships and it is incapable of handling other kind of datasets. In 1973, Ansecomb advanced such datasets.

3. What is Pearson’s R?

Pearson's R is a measure of the strength of correlation between two variables. The variables must be in interval or ratio scale. The value of Pearson's R always lie between -1 and +1. The extreme values indicate perfectly negative or perfectly positive correlations respectively. The values in between denotes the some correlation of two variables. A value of 0 indicates that there is no association between the two variables. It is used to determine the strength of association. A value over .5 or below -.5 indicates considerable correlation. It can also be used to detect multicollinearity.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

When there are multiple features, they can be in different scales. That causes two problems: difficulty in interpretation and difficulty in convergence using gradient descent methods. There are two popular methods of scaling.

In Standardizing features are scaled in a way so that their mean becomes zero and standard deviation becomes 1. This is obtained by substrating the mean of all value of one feature from the corresponding value and then dividing it by the standard deviation of those values of the feature. The formula is:

x_std = (x-mean(x))/ standard_deviation(x)

In Min-Max Scaling features are scaled in such a way that all the values lie between zero and onne. The scaling is done using the following formula:

x_scaled = [x - min(x)]/ [max(x) - min(x)]

Scaling just affects the coefficients and none of the other parameters.

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

VIF provides a measure of the variance of an estimated regression coefficient increasing due to collinearity or correlation between independent variables. If there is no correlation, that means the independent variables are orthogonal, VIF is 1. Similarly, if the variables are perfectly correlated VIF will be infinity. Some methods to address the issues are as follows:

1. The treatment is to remove independent variables step by step, 1 variable in each step.
2. Second approach can be increasing sample size.
3. Third approach can be using Principal Component Analysis.

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

Q-Q or quantile-quantile plots helps in graphically analyzing and comparing two probability distributions by plotting their quantiles against each other. It is a scatterplot created by plotting the two sets of quantiles against one another.

If both sets of quantiles came from the same distribution, we should see the points forming a line that’s roughly straight. The plot can be used to assess the distribution for a random variable. Also, it can be helpful to find out whether both the training and testing data come from the same population.

It checks whether the test and train sets — i. come from populations with a common distribution ii. have common location and scale iii. have similar distributional shapes iv. have similar tail behavior.

If the points at the ends of the curve formed from the points are not very much scattered, the assumption of normal distribution is questioned.