Linear Regression - Subjective Questions

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?

(Plots on the next page.)

From the analysis of the categorical variables from the dataset, we can infer that,

1. *season:* The demand is lowest in the *spring* and highest in the *fall*, as we can see. The following is the order: *fall* > *summer* > *winter* > *spring*
2. *yr*: The box plot clearly shows that demand in *2019* was higher than in *2018*.
3. *mnth*: We can infer that the demand is the lowest in the month of *Jan* and the highest in the month of *Sep*. *mnth* shows a curve of increasing demand from *Jan-Sep* and then again decreasing demand from *Sep-Jan*.
4. holiday: We cannot infer a strong relationship between the dependent variable and the independent variable.
5. weekday: We cannot infer a strong relationship between the independent variable and the dependent variable, but we can see a slight curve with *Mon* at the bottom and *Thu*/*Fri* at the top.
6. workingday: We cannot infer a strong relationship between the dependent variable and the independent variable.
7. weathersit: *weathersit* has a strong relationship, with demand increasing from *LRS* to *MST* to *CLD*, with *CLD* being the highest. The better the weather, the higher the demand.

As a result, we can conclude that the variables *yr*, *season*, *mnth*, and *weathersit* have a positive influence on the dependent variable, whereas the others do not.

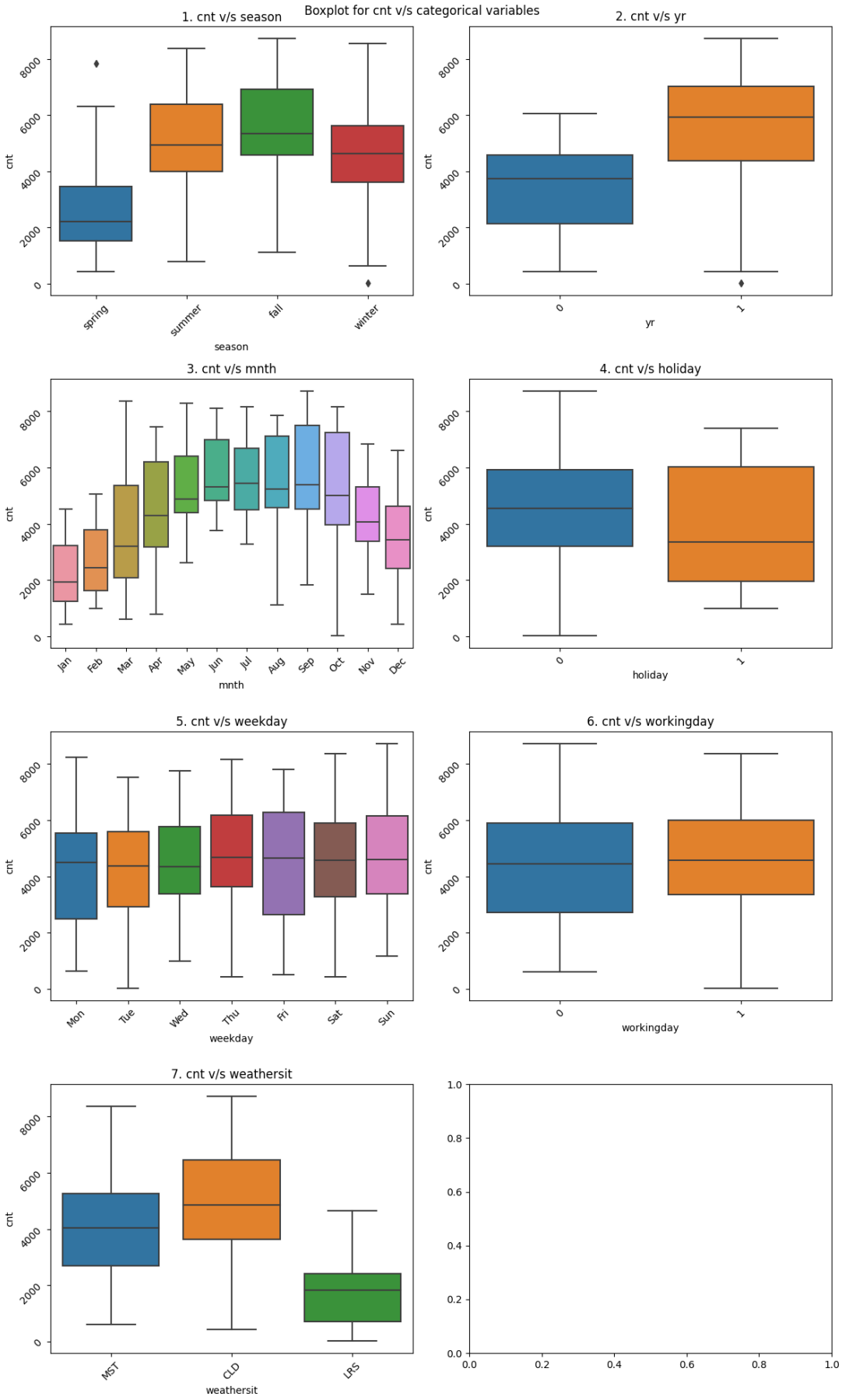

Fig. 1.a. Box Plots - Target variable vs Categorical variable

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

We will need to build *n-1* dummy variables for any categorical variable with *n* unique values. The Pandas library contains a method get_dummies that creates *n* variables for us and *drop_first=True* that drops one of the variables, resulting in *n-1* variables being produced for us. For example, we have a weathersit column with the unique values *CLD*, *MST*, *LRS*, and *HRT*. The table below depicts the options in the left column and their representation as dummy variables in the *MST*, *LRS*, and *HRT* columns.

	MST	**LRS**	**HRT**
CLD	0	0	0
MST	1	0	0
LRS	0	1	0
HRT	0	0	1

Table. 2.a. Dummy variables for weathersit

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

The variables *temp* and *atemp* have the highest correlation with the target variable. Furthermore, *temp* and *atemp* are highly connected. In comparison to *temp* and *atemp*, *hum* and *windspeed* do not appear to have a significant impact on the target variable.

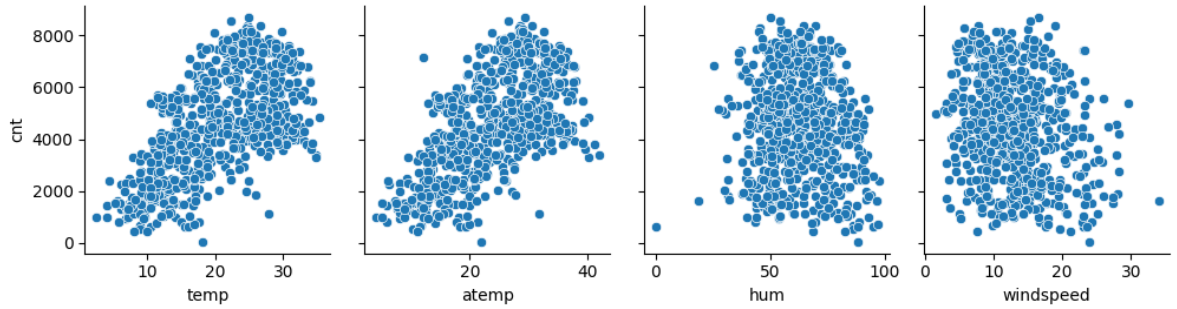

Fig. 3.a. Pair Plot - Target variable vs Numerical variables.

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

To validate the assumptions of Linear Regression after building the model on the training set,

1. Firstly, I halved the number of features using *Recursive Feature Elimination (RFE)* from *scikit-learn*.
2. Then, using *p_value* and *VIF* to select the features. I dropped all the features that in the following order,
    a. Features with *p_value* of more than *0.05* and *VIF* more than *5*. There were no features fitting in the criteria as we had dropped many using RFE.
    b. Features with *p_value* more than *0.05*.
    c. Features with *VIF* more than *5*.
3. While dropping the features, it was seen that after dropping *temp*, the *p_value* of *Jul* crossed the set threshold limit i.e. *p_value* more than *0.05*. Hence, after dropping *Jul* but keeping *temp* in the features overall good results were seen.
4. The R-Squared and Adj. R-squared did not have a high difference which is a good measure to say the quality of the model.
5. After making the selection of the features, a density plot for the residuals (error terms) helped to validate the model performance. The residuals lie near the 0. Hence, this says that the model is good enough.

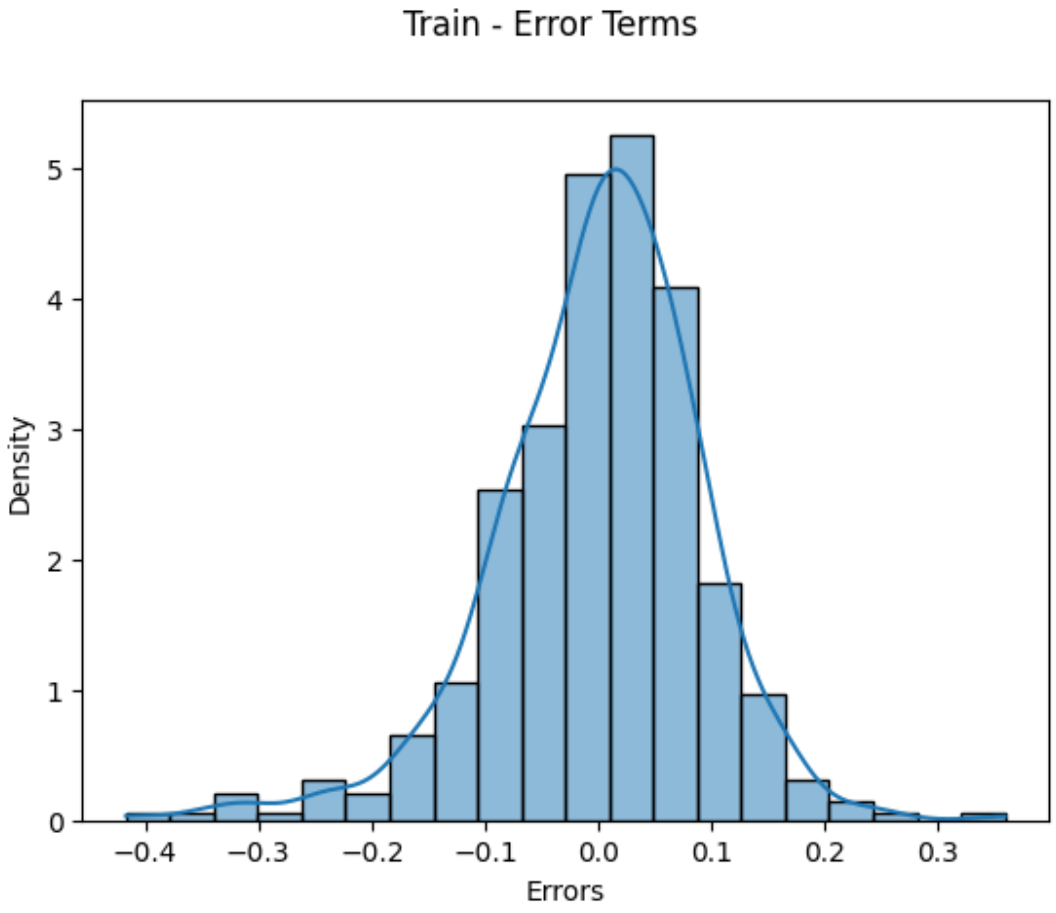

Fig. 4.a. Density Plot - Train - Error Terms

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

According to the OLS regression results, the top three factors that contribute considerably to understanding the demand for shared bikes are:

1. *Temperature (temp)* - with a coefficient of 0.4777 and a t-value of 14.423, this variable has a strong positive connection with the dependent variable and contributes significantly to explaining the variance in demand for shared bikes.
2. *Year (2019)* - with a coefficient of 0.2341 and a t-value of 28.237, indicating a strong positive association with the dependent variable and a considerable contribution to explaining the variance in shared bike demand.
3. *Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds (LRS)* - with a coefficient of -0.2850 and a t-value of -11.444, indicating a strong negative connection with the dependent variable and a considerable contribution to explaining variance in shared bike demand.

Windspeed, holiday, and *season (spring, summer, winter)* all have significant coefficients and help to explain the variation in demand for shared bikes.

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Linear regression is a statistical technique to determine a link between one or more independent variables and a dependent variable. The technique works by fitting a straight line to a group of data points that best captures the relationship between the variables.

The linear regression process begins by choosing a dependent variable, also known as the response variable, to be predicted. The independent variables, also known as predictor variables, are chosen based on their ability to influence the dependent variable. The process then generates a linear equation with the formula $y = mX + b$, where y is the dependent variable, X is the independent variable, m is the line's slope, and b is the y-intercept.

The linear regression algorithm's goal is to identify the values of m and b that best fit the data points. This is accomplished by minimising the sum of the squared differences between predicted and actual values. This is referred to as least-squares regression.

The basic algorithm for linear regression is as follows,

1. Calculate the mean of X and y.
2. Calculate the variance of X and the covariance between X and y.
3. Calculate the slope of the regression line, m: $m = covariance(X, y) / variance(X)$
4. Calculate the y-intercept of the regression line, b: $b = mean(y) - m * mean(X)$
5. Use the equation of the regression line to make predictions for new values of X: $y\_predicted = m * X + b$
6. Return m, b, and $y\_predicted$ as outputs.

For multiple linear regression, the algorithm is similar but involves calculating the coefficients of each independent variable instead of just a single slope value.

Q2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a collection of four datasets with identical statistical properties but strikingly different graph patterns. The quartet, which was introduced in 1973 by statistician *Francis Anscombe*, serves as a powerful reminder of the importance of data visualisation in statistical analysis.

Each dataset in the quartet contains 11 paired observations of two variables, and the summary statistics such as *mean*, *standard deviation*, *correlation coefficient*, and *regression line* are the same for all four datasets. When the datasets are graphed, however, they show starkly different patterns, emphasising the limitations of relying solely on summary statistics. For example, a dataset has a linear relationship between the two variables, whereas another has a non-linear relationship that is better modelled by a quadratic function. The quartet emphasises the importance of visualising data in order to gain a deeper understanding of its patterns and relationships, as well as the importance of employing multiple techniques in statistical analysis, such as summary statistics, visualisation, and statistical models.

Q3. What is Pearson’s R?

Pearson's R is a statistical metric that measures the degree of linear association between two variables. It is also known as the Pearson correlation coefficient or Pearson's product-moment correlation coefficient. It is used to assess the strength and direction of a link between two variables and to forecast one variable based on the other.

It is denoted by *r* and has a range of *-1* to *+1*. A positive value of *r* shows a positive correlation, which means that if one variable increases, so does the other. A negative value of *r* implies a negative correlation, which means that if one variable rises, the other tends to fall. A value of *r* near to zero suggests that the variables have little or no association.

Pearson's R has several important qualities in addition to quantifying the strength and direction of a linear relationship between two variables. For example, it is a standardised measure, which means that discrepancies in the scales or units of the two variables have no effect on it. It is also a symmetrical measure, which means that the values of *r* for variables *X* and *Y* are the same. Finally, Pearson's R can be used to test hypotheses to see if the observed connection is statistically significant.

Q4. What is scaling? Why is scaling performed? What is the difference between normalised scaling and standardised scaling?

Scaling is the process of transforming numerical data to fit within a specific range or scale. It involves converting raw data to a standardised format that machine learning algorithms can more easily interpret. Scaling is a common pre-processing step in data analysis and machine learning that ensures features or variables are all on the same scale, making them more comparable and easier to interpret.

Scaling is performed to,

- Prevent higher-valued features from dominating lower-valued features. This is important when using distance-based algorithms like K-Nearest Neighbours, where larger features can have a disproportionate impact on the output.
- Improve machine learning models' accuracy and performance. Scaling can help reduce the impact of outliers and improve algorithm convergence.
- Improve the data's interpretability. Scaling can aid in the identification of trends and patterns that may not be visible in raw data.

Normalised scaling, also known as min-max scaling, implies reducing the data to a range of 0 to 1. This is done by subtracting the lowest value from each observation and dividing the result by the range (the difference between the maximum and minimum values). The formula for normalised scaling is: $x\_norm = (x - min(x)) / (max(x) - min(x))$

Standardised scaling, also known as z-score scaling, implies adjusting the data so that it has a mean of 0 and a standard deviation of 1. Subtracting the mean from each observation and dividing by the standard deviation accomplishes this. The formula for standardised scaling is: $x\_std = (x - mean(x)) / std(x)$

Normalised scaling preserves the shape of the distribution and the relative distance between observations, whereas standardised scaling transforms the data to have a mean of 0 and a standard deviation of 1, making it easier to compare across variables.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

The *Variance Inflation Factor (VIF)* indicates how much the variance of the predicted regression coefficients is inflated as a result of multicollinearity in the independent variables. *Multicollinearity* is indicated by VIF values larger than one, with values greater than 5 or 10 indicating substantial multicollinearity.

VIF can have an infinite value in some instances. When one or more independent variables are exactly linearly connected to each other, the regression procedure produces a singular matrix. When a matrix is singular, it signifies that its inverse does not exist, resulting in an error in the calculation of the VIF.

When the same variable is included several times in the regression model or when a variable is a function of other variables in the model, perfect multicollinearity can arise. The VIF will be undefined or infinite in such instances.

For instance, consider a dataset with two independent variables, *X1* and *X2*, where *X1* is a linear combination of *X2*. This means that *X1* can be predicted precisely from *X2*. Because the variance of the estimated regression coefficient for *X1* cannot be computed due to perfect multicollinearity with *X2*, the VIF for *X1* will be infinite in this case.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A Q-Q plot, also known as a *quantile-quantile* plot, is a graphical tool used to determine whether a dataset is roughly normally distributed. The quantiles of the dataset are plotted against the theoretical quantiles of a normal distribution in a Q-Q plot. The data is considered normally distributed if the points on the plot fall roughly on a straight line. The data is not normally distributed if the points deviate significantly from a straight line.

In linear regression, Q-Q plots are widely used to evaluate the assumption of normality of the residuals, which are the discrepancies between the actual values of the dependent variable and the projected values from the regression model. For the linear regression model to be valid, the residuals should preferably be normally distributed with a mean of zero and a constant variance.

The use of a Q-Q plot in linear regression is important because it can visually detect deviations from normality in the residuals. If the residuals are not normally distributed, it might result in biassed regression coefficient estimations, erroneous standard errors, and unreliable p-values. A Q-Q plot can help uncover non-normality in residuals, allowing the researcher to take corrective actions such as data transformation or using an alternative model.

To generate a Q-Q plot for a linear regression model, first compute the residuals by subtracting the predicted values from the actual dependent variable values. A scatter plot is then used to compare the residuals to the theoretical quantiles of a normal distribution. If the residuals are regularly distributed, the plot points should resemble a straight line. If the points depart from a straight line, it indicates that the residuals are not normally distributed and that more research is needed.