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?

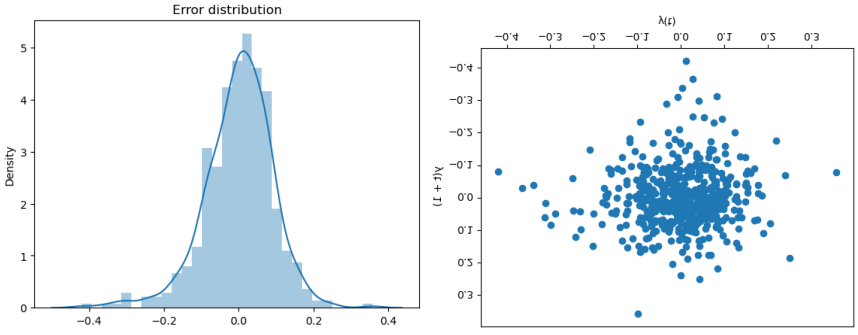
holiday,windspeed,jul,spring,mist+cloud,light rain/snow are negative coefficient if these incr ease then it has inverse effect in hiring bike.
yr,temp,sep,winter have positive coefficient which means if 1 unit of this var increae then co unt of rental bike will also increase.
The equation of our best fitted line is:

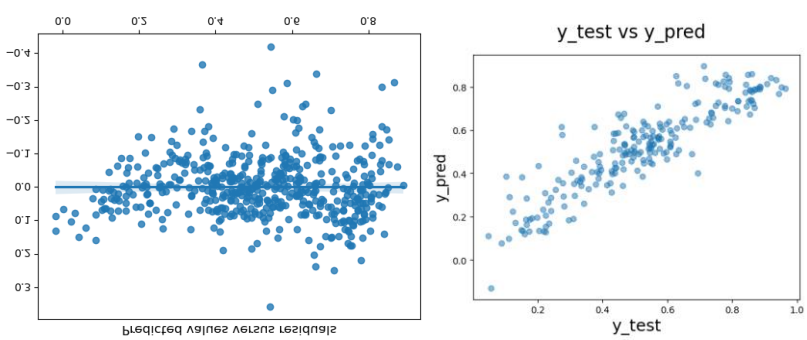
cnt=0.253+(0.253×yr)-(0.098×holiday)+(0.450×temp)−(0.140×windspeed)-(0.073×Jul)+(0.0 57×sep)-(0.112×spring)+(0.045×winter)−(0.080×Mist + Cloudy)-(0.285*lightrain/snow

Why is it important to use drop_first=True during dummy variable creation?
When Dummy variables created for the independent variables categorical once.if not droppe d then multicollinearity happens between dummy var and independent var .To avoid redan dants in dataset we have to use drop_first.

3. Looking at the pair-plot among the numerical variables, which one has the highest correla tion with the target variable?
Ans:Temp and atemp

4.How did you validate the assumptions of Linear Regression after building the model on the training set?
Ans:Vif is <5% and p<0.05 which rejects the null hypothesis
Verifying multicolinearity





- Residuals are normally distributed
- Mean Residual errors are independent of each other since the Predicted values vs Residuals plot doesn't show any trend.
- Lagplot of residuals shows no trend. Hence the error terms have constant variance
- Predicted vs observed value plots shows that the model is reasonably accurate.
- Hence, assumptions of Linear Regression are satisfied by this model

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

Ans:Temp,yr (positive coeff) and light rain/snow(negative coeffient)are the three which has h ighest coefficients among rest in this model

General Subjective Questions

I. 1.Explain the linear regression algorithm in detail.

Ans: The mathematical equation of Linear regression:$Y = \beta0 + \beta1*x$ for simple Linear regression

Y-Target variable,B0 is constant, β1 is independent variable ,x is the coefficient.

Positive coeff->y increases(coefficient of x) with increase in 1 unit x

Neg coeff->y decreases(coefficient of x) with increase in 1 unit x

For multi $Y = \beta0 + \beta1*x + \beta2*2\ldots.. \beta i*xi$

This equation that best describes the correlation of the explanatory variables with the dependent variable. This is achieved by fitting a line to the data using least squares

In Linear Regression, Mean Squared Error (MSE) cost function is used, which helps to figure out the best possible values for a0 and a1, which provides the best fit line for the data points

Steps followed while performing Linear regression:

I. Draw the scatterplot of the variables. Check for linearity or non-linearity of the data and any deviations. If the pattern is non-linear, perform a transformation. If there are outliers, remove them only if there is a non-statistical reason.

II. Fit the least-squares regression line to the data and check the assumptions of the model by looking at the residual plot and normal probability plot. If the assumptions of the model are not met, a transformation may be needed.
III. If necessary, transform the data and re-fit the least-squares regression line using the transformed data.
IV. If a transformation was performed, then go back to step 1. Otherwise, move to step 5.
V. Once a "good-fitting" model is obtained, write the equation of the least-squares regression line. Consider the standard errors of the estimates, the estimate of standard deviation, and R-squared.
VI. Find if the explanatory variable is a significant predictor of the target variable by performing a t-test or F-test.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a group of datasets (x, y) that have the same mean, standard deviation, and regression line, but which are qualitatively different. It is often used to illustrate the importance of looking at a set of data graphically and not only relying on basic statistic properties.

3.What is Pearson's R?

Pearson's product moment correlation coefficient (r) is given as a measure of linear association between the two variables: r^2 is the proportion of the total variance (s^2) of Y that can be explained by the linear regression of Y on x.

For example:

- Positive linear relationship: In most cases, universally, the income of a person increases as his/her age increases.
- Negative linear relationship: If the vehicle increases its speed, the time taken to travel decreases, and vice versa.

From the example above, it is evident that the Pearson correlation coefficient, r, tries to find out two things – the strength and the direction of the relationship from the sample sizes.

	Coefficient, r	
Strength of Association	Positive	Negative
Small	.1 to .3	-0.1 to -0.3
Medium	.3 to .5	-0.3 to -0.5
Large	.5 to 1.0	-0.5 to 1.0

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

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm. Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization/Min-Max Scaling:

It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

sklearn.preprocessing.scale helps to implement standardization in python.

One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.

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

Q-Q plots are also known as Quantile-Quantile plots , 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 to show if two data sets come from the same distribution. may also be used to determine outliers.