Assignment-based Subjective Questions

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

A. We could explain 81% of variance on train data and 79 % of variance on test data based on 8 prominent features

a. Temperature: It has positive co-efficient of 0.489. It means as the temperature (predictor variable) increases the count (response / target) variable also increases. It says that the count parameter increases by 0.489 for every 1 unit change in the temperature.

b. Year: It has a positive co-efficient of 0.2322. It tells that the 2019 performed better than 2018.

c. Weathersit:

i. Misty: Represented as 2 in the data dictionary, has a negative co-efficient of -0.073. The performance of the target variable decreases as weather becomes mist & cloudy or mist with broken clouds or mist with few clouds or its all misty

ii. Light snow: Represented as 3 in the data dictionary, has a negative co-efficient of -0.296. The performance of the target variables decreases Light snow or Light rain with thunderstorm and scattered clouds or light rain with scattered clouds

d. Season:

i. Summer: It has a positive co-efficient of 0.03, indicating the number of people renting the bike increases in summer season

ii. Winter: It has a positive co-efficient of 0.08, indicating the number of people renting the bike increases in winter season

iii. Spring: It has a negative co-efficient of -0.093, indicating that the number of persons renting the bike decreases in spring season

e. Holiday: It has a negative co-efficient of -0.109, indicating that the number of people renting the bike decreases in holidays.

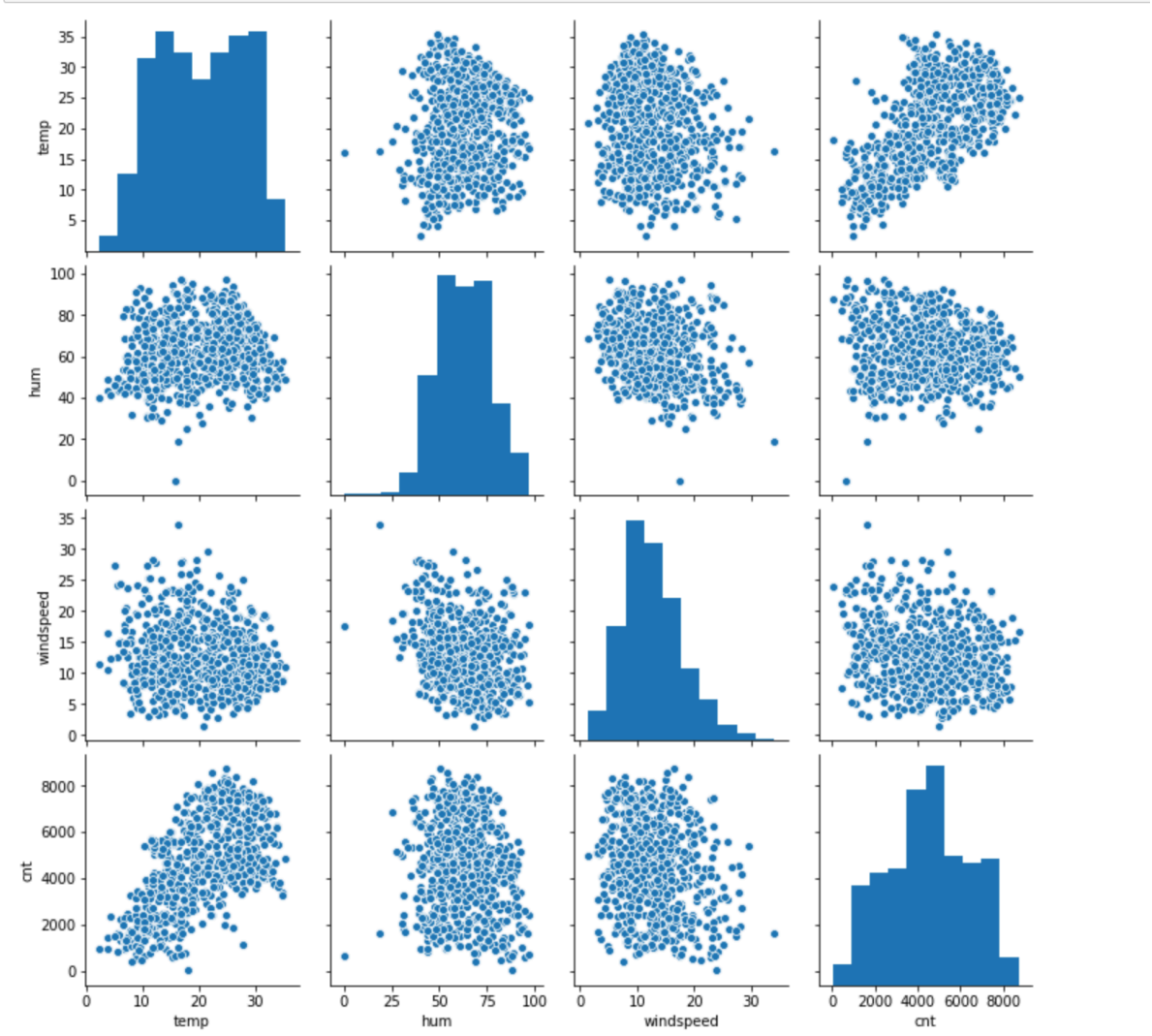
Conclusion: We are seeing better renting with increase in temperature and when the season is Summer or Winter. We see an increase in rentals in 2019. We see that when the weather is Misty, cloudy or snowy or when the season is spring, we see a decreased performance in renting.

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

A. The drop_first = true tells, weather to get K-1 dummy variables out of K categorical variables by removing the first level. This is needed as it reduces the number of extra columns and reduces the number of correlations created against those dummy variables. As we could explain the K categorical variables with K-1 categorical variables it is redundant.

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

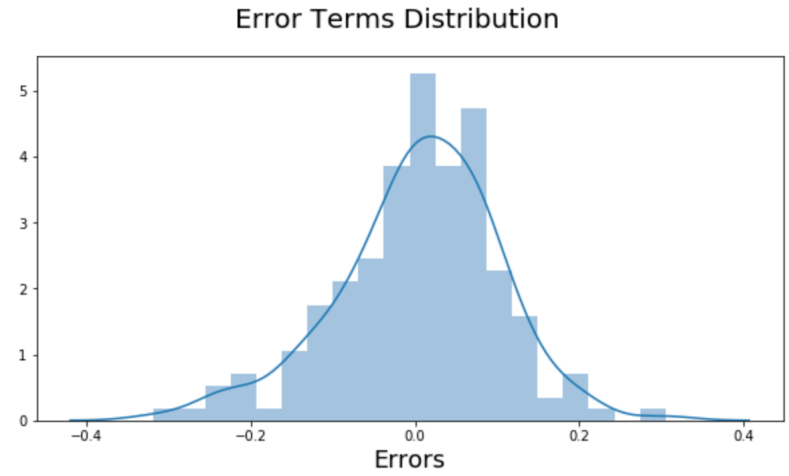
A. Looking the pair plot of the numerical variables on the pair – plot below, we could see that a model for linear representation can be made with the temperature (temp) variable and it has the highest correlation which is represented in the heat map.



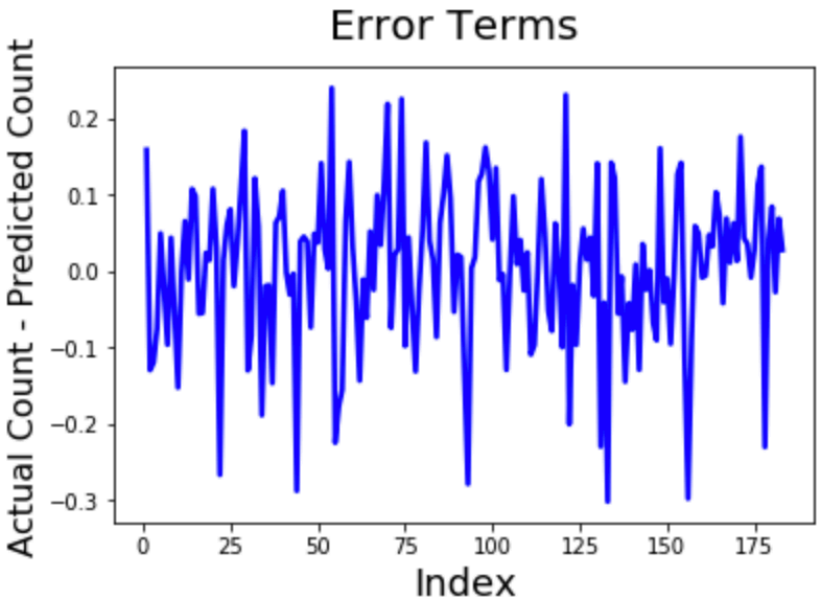
Q. How did you validate the assumptions of Linear Regression after building the model on the training set?

A. We can validate the assumptions of linear regression the following way

1. If you look at the below graph, we could observe that the error terms are normally distributed with mean = 0



2. We could see in the below graph that the error terms are independent of one and other



3. We can see from the above graph that the error terms have a constant value at a given point
4. The model fits a hyper plane as we have 8 variables defining the target variable

count = 0.164 + (0.2322) * yr + (-0.1098) * Holiday + (0.4891)* temp * (-0.2962) * light_snow + (-0.0732) * misty + (-0.0923) * misty + (0.03) * summer + (0.0841) * winter

5. As the algorithm uses the sum of squared errors, the co-efficient(s) are obtained using the model

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

A. We can derive the 3 top features based on the co-efficient of predictor variables. Because the count parameter increases / decrease by co-efficient for every, one unit change in the predictor variables. As we observe from the stats below the top 3 would be Temperature, Light Snow and the Year variables.

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.815
Model:	OLS	Adj. R-squared:	0.812
Method:	Least Squares	F-statistic:	295.5
Date:	Mon, 13 Dec 2021	Prob (F-statistic):	2.34e-191
Time:	13:42:32	Log-Likelihood:	501.19
No. Observations:	547	AIC:	-984.4
Df Residuals:	538	BIC:	-945.6
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1640	0.028	5.819	0.000	0.109	0.219
yr	0.2322	0.008	27.596	0.000	0.216	0.249
holiday	-0.1098	0.026	-4.266	0.000	-0.160	-0.059
temp	0.4891	0.033	14.660	0.000	0.424	0.555
light_snow	-0.2962	0.026	-11.389	0.000	-0.347	-0.245
misty	-0.0732	0.009	-8.240	0.000	-0.091	-0.056
spring	-0.0923	0.020	-4.525	0.000	-0.132	-0.052
summer	0.0300	0.014	2.196	0.029	0.003	0.057
winter	0.0841	0.017	5.093	0.000	0.052	0.117

Omnibus:	73.418	Durbin-Watson:	2.034
Prob(Omnibus):	0.000	Jarque-Bera (JB):	185.264
Skew:	-0.690	Prob(JB):	5.90e-41
Kurtosis:	5.495	Cond. No.	16.3

General Subjective Questions

Q. Explain the linear regression algorithm in detail

A. There are 2 types of linear regression

1. Simple Linear Regression
    a. The most elementary type of regression model is the simple linear regression which explains the relationship between a dependent variable and one independent variable using a straight line
    b. The standard equation of the regression line is given by the following expression: $Y = \beta 0 + \beta 1.X$
    c. The strength of the linear regression model can be assessed using 2 metrics:
        i. R^2 or Coefficient of Determination: It is a number which explains what portion of the given data variation is explained by the developed model. It always takes a value between 0 & 1

$$R^2 = 1 - (RSS / TSS)$$

            a. RSS (Residual Sum of squares): It is the measure of the difference between the expected and the actual output. A small RSS indicates a tight fit of the model to the data
            b. TSS (Total sum of squares): It is the sum of errors of the data points from mean of response variable

        ii. Residual Standard Error (RSE) : The Residual Standard Error is the average amount that the response (dist) will deviate from the true regression line.
    d. Assumptions of Linear Regression:
        i. Linear relationship between X and Y
        ii. Error Terms are normally distributed
        iii. Error Terms are independent of each other
        iv. Error terms have constant variance(homoscedasticity)

2. Multiple Linear Regression
    a. Statistical technique to understand the relationship between one dependent variable and several independent variables
    b. Expressed with equation

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

    c. Model fits a hyperplane instead of a line
    d. Coefficients still obtained by minimizing sum of squared error
    e. The assumption of linear regression still hold

f. Overfitting: When you add more and more variables, there could be a problem with generalization, as it fits all of the training data and when it runs on the test data the accuracy drops significantly
g. Multicollinearity: Some of the variables might completely explain some other independent variable in the model due to which the presence of that variable in the model is redundant.
    i. Affects of multicollinearity are
        1. Co-efficient(s) swing wildly, and can invert
        2. P-values may not be reliable
    ii. Doesn't effect multicollinearity are
        1. Precision of the predictions
        2. R-squared value
    iii. 2 ways to deal with multicollinearity are
        1. Correlations: Highly correlated variables are multi-collinear
        2. Variance Inflation factor (VIF): VIF basically helps explaining the relationship of one independent variable with all the other independent variables
        3. VIF greater than 10 is definitely high, a VIF of greater than 5 should also not be ignored and inspected appropriate
    iv. The few methods to deal with multicollinearity are:
        1. Dropping variables
        2. Create new variable with older variables
    v. Feature Scaling:
        1. When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret
        2. Helps with ease of interpretation and faster gradient descent
        3. We can use standardization or MixMax Scaling

Handling categorical Variables: But when you have multiple variables, there might be some categorical variables that might turn out to be useful for the model. So it is essential to handle these variables appropriately in order to get a good model. One way to deal with them is creating dummy variables

Hypothesis testing: The Hypothesis says that the co-efficient(s) should be zero

$$H_0 : \beta_1 = 0$$

$$H_A : \beta_1 \neq 0$$

Once the model runs we can determined that the coefficient is significant, using p-values.

Params to determine the model:

1. **T-statistic:** Used to determine the p-value and hence, helps in determining whether the coefficient is significant or not
2. **F statistic:** Used to assess whether the overall model fit is significant or not. Generally, the higher the value of F statistic, the more significant a model turns out to be
3. **R-squared:** After it has been concluded that the model fit is significant, the R-squared the extent of the fit, i.e. how well the straight line describes the variance in the data. Its value ranges from 0 to 1, with the value 1 being the best fit and the value 0 showcasing the worst

Q. Explain the Anscombe's quartet in detail.

A. It tells us about the importance of visualizing the data before applying various, which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc

It can be defined as a group of four data sets which are **nearly identical in simple descriptive statistics (Same Mean, Standard Deviation and Correlation)**, but there are some peculiarities in the dataset

The four data sets can be described as:

1. Fits the linear regression model
2. As the data is non-linear the linear regression model could not fit the data quite well
3. The outliers involved in the dataset which cannot be handled by linear regression model
4. one high-leverage point is enough to produce a high correlation coefficient.

Q. What is Pearson's R?

A. Pearson's R, the Pearson product-moment correlation coefficient (PPMCC), and bivariate correlation are all names for the Pearson correlation coefficient. It is a statistic with a numerical value between -1.0 and +1.0 that measures the linear correlation between two variables. It is incapable of capturing nonlinear relationships between two variables and of distinguishing between dependent and independent variables.

Pearson's correlation coefficient is calculated by dividing the covariance of the two variables by the product of their standard deviations. The definition takes the form of a "product moment," which is the mean (the first moment about the origin) of the product of the mean-adjusted random variables; thus, the name includes the modifier product-moment.

$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

Where,

N = the number of pairs of scores

Σxy = the sum of the products of paired scores

Σx = the sum of x scores

Σy = the sum of y scores

Σx2 = the sum of squared x scores

Σy2 = the sum of squared y scores

The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association

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

A. It is a data Pre-Processing step that is applied to independent variables in order to normalize the data within a specific range. It also aids in the speeding up of algorithm calculations.

Most of the time, the collected data set contains features with widely disparate magnitudes, units, and ranges. If scaling is not performed, the algorithm only considers magnitude rather than units, resulting in incorrect modeling. To solve this problem, we must scale all of the variables to the same magnitude level. It is important to note that scaling has no effect on the other parameters such as t-statistic, F-statistic, p-values, R-squared, and so on.

Normalization / Mix Max Scaling:

- Normalization is a scaling technique that shifts and rescales values to make them range between 0 and 1.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

- Xmax and Xmin are the maximum and the minimum values of the feature respectively. When the value of X is the minimum value in the column, the numerator will be 0, and hence X' is 0. On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator and thus the value of X' is 1. If the value of X is between the minimum and the maximum value, then the value of X' is between 0 and 1
- When you know that your data does not follow a Gaussian distribution, normalization is a suitable option.

Standardization:

- Another scaling strategy is standardization, in which the values are centered around the mean with a unit standard deviation. This results in the attribute's mean becoming zero, and the resulting distribution having a unit standard deviation.

$$X' = \frac{X - \mu}{\sigma}$$

- In circumstances where the data follows a Gaussian distribution, on the other hand, standardization can be beneficial. This, however, does not have to be the case. Standardization, unlike normalization, does not have a bounding range. As a result, even if your data contains outliers, normalization will have no effect on them.

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

- VIF = infinity indicates perfect correlation. This demonstrates that two independent variables have a perfect correlation.
- We get R2 = 1 in the event of perfect correlation, which leads to 1/(1-R2) infinite.

- To remedy this issue, we must remove one of the factors that is creating this perfect multicollinearity from the dataset.

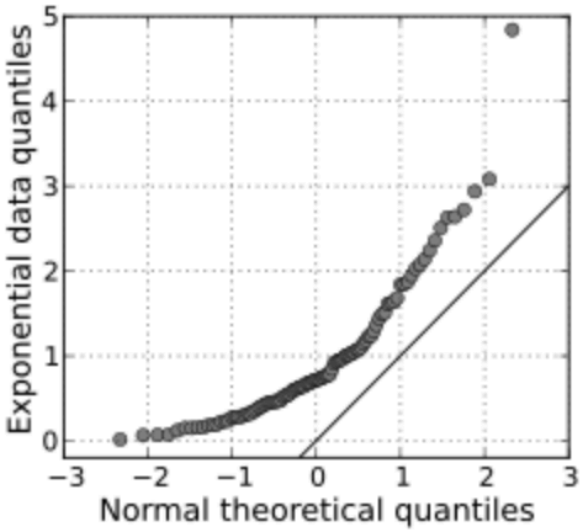
Q. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
A. It helps to determine if two data sets come from populations with a common distribution. This is useful in a linear regression scenario where the training and test data sets are obtained separately, and the Q-Q plot is used to demonstrate that both data sets are from populations with similar distributions.

It's used to see if the following scenarios are true and if the 2 data sets have:

- from populations having a similar distribution
- share a common scale and location
- have distributional shapes that are similar
- have a tail that behaves similarly

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.). A quantile is a percentage of the population in which specific values fall below it. The median, for example, is a quantile where 50% of the data falls below it and 50% of the data falls above it.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line $y = x$. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line $y = x$. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.



References:

- https://towardsdatascience.com/
- https://www.geeksforgeeks.org/anscombes-quartet/
- https://www.analyticssteps.com/blogs/pearsons-correlation-coefficient-r-in-statistics
- http://www.people.vcu.edu/~pdattalo/706SuppRead/Pearson's%20r.html
- https://medium.com/@premal.matalia/what-is-scaling-why-is-scaling-performed-normalized-vs-standardized-scaling-5113c86688f8
- https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/
- https://medium.com/@premal.matalia/q-q-plot-in-linear-regression-explained-ab040567d86f
- https://stackoverflow.com/questions/60163405/vif-function-returns-all-inf-values
- https://www.statisticshowto.com/q-q-plots/
- https://en.wikipedia.org/wiki/Q%E2%80%93Q_plot