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?

In the given dataset the categorical variables are yr, weekday, weathersit and month.

After creating dummies for the categorical variables in the dataset and running a regressions model on the same. Here are my inferences:

- Year is having a very positive effect on the variable cnt as the bike shares increased from 2018 to 2019.
- Weekends such as Saturday and Sunday have positive effect on the dependent variable "cnt". That means the demand for Bike shares is more on Saturday and Sunday when compared to other weekdays.
- Months December, July and November are having negative effect on the bike shares. This might be due to the climate in those months people generally do not opt for travelling on a bike, rather they might be going for car.
- Weather conditions like Light Snow and Mist also have a negative effect on the cnt. When the weather is getting cold or rainy the bike shares count will come down.

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

Consider we have a categorical variable called color which is having 3 values red, blue and green. When we create dummy variables for this categorical variable, by default python creates 3 variables i.e. Red, Blue, Green. Each will have values 0 and 1 if the data has red or blue or green respectively. But we can clearly say that if a data item is having value 0 for blue and green that means that its value is red. Hence we can drop the first variable and identify the 3 colors by adding 2 dummy variables as below.

Blue	Green
0	0
0	1
1	0

Here when Blue and Green are having 0 values it can be interpreted that the color is red. Hence drop_first=True is important to use.

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

"yr","temp" and "atemp" have the high correlation with the target variable. Apart from them "casual" and "registered" also have high correlation but they will not be considered for building the model.

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

Below are the assumptions and the checks:

1. Linearity Assumption:

    a. It is assumed that the relationship between dependent variable and independent variables is liner. This can be verified through scatter plots.
2. Assumptions about the residuals:
    a. Normality Distribution:
        i. All the residuals should be normally distributed.
        ii. This can be verified by plotting a distplot(in seaborn).
    b. Zero mean Assumption:
        i. All the residuals are normally distributed around 0 i.e mean of all residuals is equal to 0.
        ii. This can also be verified through the distplot.
    c. Assumption of Homoscedasticity:
        i. The residual terms should have the same variance.
        ii. This can be verified by the scatter plot of the residuals vs the predicted variables. There should be no pattern identified.
3. Multicollinearity:
    a. The independent variables are not linearly correlated to each other.
    b. We can remove the independent variables by checking the vif values regularly for each variable. VIF above 5 is not acceptable.

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

Three features yr, atemp and Light Snow.

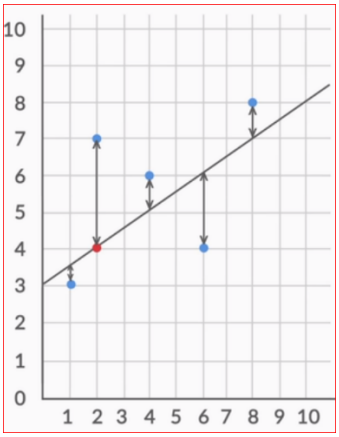
- Yr is contributing positively by having the coefficient of 0.2331.
- Atemp is also contributing positively by having the coefficient of 0.4460.
- Light Snow is contributing negatively with the coefficient value of -0.2413.

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Linear Regression is an algorithm used to find the best linear relationship between independent and dependent variables. It is part of Supervised Machine Learning.
We need to train the model by giving the training data which comprises of independent and dependent variables.
The best fit line is a line which fits the scatter plot in a best way. This is achieved by minimizing the residual sum of squares. [Sum of squares of difference between true value and predicted value]

Regression line is y=mx+c.



RSS(Residual Sum of Squares)

$e_{i} = y_i - y_{pred}$

$RSS = e_1^2 + e_2^2 + e_3^2 + \ldots\ldots\ldots + e_i^2$

To identify the goodness of the fit we use R.squred method which is 1-(RSS/TSS)

TSS = Sum of squares if difference between true value and the mean value of data points.

This R.squared can be used for Simple Linear Regression.

When we move to Multiple Linear Regression the number of independent variables will be more than 1. As adding more and more independent variables may make the model over fit, for Multiple Linear Regression we use Adjusted R.Squared which will penalize the model for having more number of features.

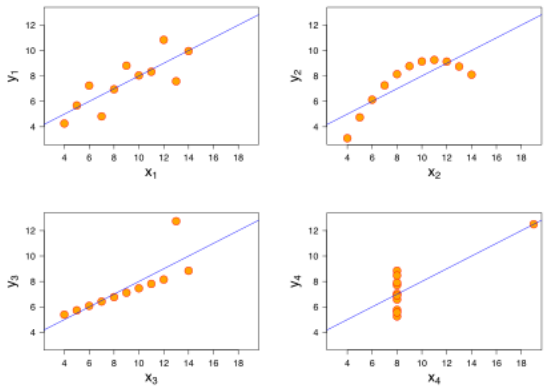
Adj.R-squared = $1-((1-r^2)(n-1)/(n-p-1))$

While Building a Multiple Linear Regression Model there are below assumptions which needs to be taken care of

1. Assumption of Linearity
2. Assumptions about the residuals
    a. Residuals are normally distributed
    b. Residuals have mean at 0
    c. Assumption of homoscedasticity
3. Multi Collinearity.

Q2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is having four data sets for which the regression line is same but the data appear very different when they are graphed as below.



- The first(left top) graph is having a linear relationship between x and y.
- The second(top right) graph has a nonlinear relationship yet the regression line is same.
- The third(bottom left) graph is having a linear distribution but it should have a different line as the datapoints are not equally variated and there is an outlier.
- In the fourth graph there are many data points for a single x variable and there is an outlier. Yet the regression line is same.

This quartet is constructed by statistician Anscombe to demonstrate the important of graphical view of the data before applying any regression model.

Q3. What is Pearson's R?

Pearson's R is the Pearson's Correlation Coefficient which is obtained by dividing the covariance of two variables with the product of their Standard Deviations.

It talks about the linear correlation between 2 variables.

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Cov - covariance
Sigma - Standard Deviation
X,Y - Variables.

Range of the Pearson's R is -1 to 1

-1: -ve correlation. Y decreases as X increases.
0: No Correlation
1: +ve correlation. Y increases as X increases

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling is performed on variables to normalize the data so that the coefficients in the linear regression model will be good enough to compare with each other.

Suppose if a variable Sales is having data in millions and other independent variable Cities is in 10s or 100s, then when the regression model is built, the coefficient of Sales will have a very huge value compared with City.

In order to understand the model well and eliminate the feature based on the impact, scaling is necessary.

There are 2 types of Scaling

- Min Max Scaling:
  - This will bring all the data between 0 and 1.
  - x=(x−min(x))/(max(x)−min(x))
- Standardized Scaling:
  - Brings all the data into a Standardized Normal Distribution with mean 0.
  - x=(x−mean(x))/sd(x)

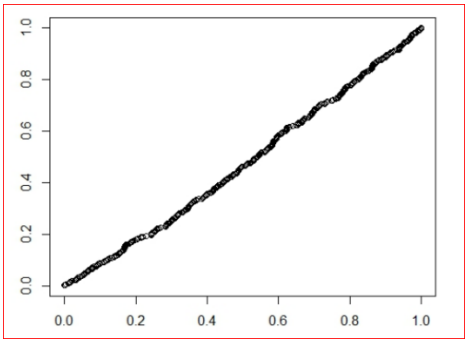
Scaling is always done after the test and train split. The scaler is fit on the train data and then train data is transformed whereas the test data is just transformed.

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

VIF for a variable will be infinity when that variable is having a perfect linear relationship with other variables i.e. the variable can be perfectly represented by a linear combination of other variables. Hence it is good to remove these variables with high VIF in the model.

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

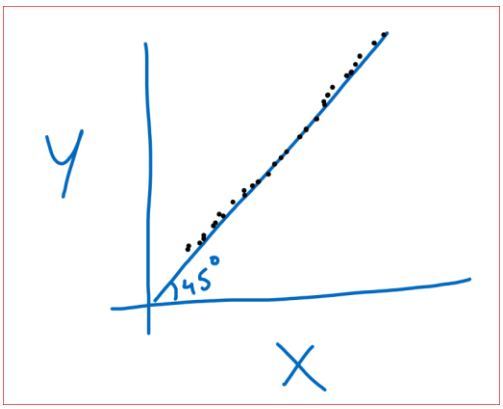
Q-Q plot is a scatter plot created by plotting quantiles of two data against each other. If the quantiles are obtained from the same distribution, then the Q-Q plot will look something like a Straight Line as below.

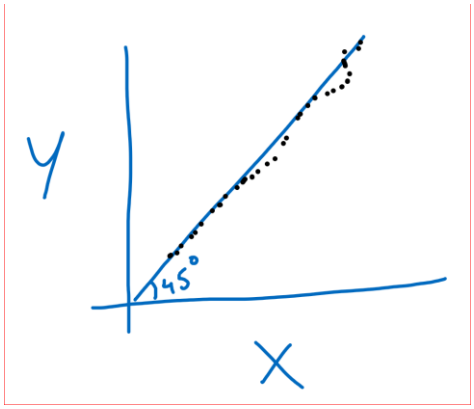


Q-Q plots is mainly used to compare the distributions of 2 variables. If the variables come from similar distribution, then the Q-Q plot looks like a line at 45^0.

The Q-Q plot is used to assess if the data is Normally distributed or not.
In Linear Regression when we receive the training and test data sets separately, then we can use the Q-Q plot to identify if both data sets are from population with same distribution.

We can also plot the predicted vs true target variables to judge the regression model.
Suppose if there are 2 data sets X and Y. If X and Y come from the same distribution the plot will be a line at 45^0.
If X quantiles are less than Y or vice versa, then the respective points will lie below or above the straight line at 45^0.





If both data sets are not from different distributions, then all the data points lie far away from the 45^o line.