ASSIGNMENT BASED SUBJECTIVE QUESTONS

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

From the final model of my analysis, most of the categorical variables have a negative impact on the target variable. For example, (mist + cloudy), (heavy rain + ice pallets) have negative impact on bike usage. Also, some months like November, December, January have negative impact. Also, spring and summer have negative impact on dependent variable.

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

If there are three categories in a categorical variable, we can explain which category it belongs to using two dummy variables. Let's say we have a **house** variable having 3 values furnished, semi-furnished, unfurnished.

Furnished	Semi-furnished
1	0
0	1
0	0

From the above table, if both are zero, we can say that house is unfurnished. So, it is clear that we can explain three categories with 2 dummy variables. So, using **drop_first** , it drops the first category and creates 1 less dummy variable.

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

After looking at pair plot, target variable **cnt** is highly correlated with **registered** with correlation value (0.95), followed by **casual** with value 0.67. **Temp** and **Atemp** are third highest correlated variable with value 0.63

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

Assumptions of Linear Regression are:

a. There is a Linear relationship between X and Y
b. Error terms are normally distributed with mean zero
c. Error terms are independent of each other

d. Error terms have constant variance

After building the model, plotting the histogram of the residuals shows us whether error terms are normally distributed or not. By plotting a scatter plot of the residuals, we can identify any patterns in the error terms. Also, looking at the cap of error terms, we get an idea whether error terms have a constant variance or not. All these are plotted in the Jupyter notebook

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

From the final model submitted, **yr**, **season_spring, weathersit_C** variables have big coefficients. yr has a significant role in explaining the bike sharing count. Weathersit_C (heavy rain + ice pallets) negative impact on bike usage. Season_spring also has negative impact on bike usage.

General Subjective Questions

1. **Explain Linear Regression algorithm in detail**.

Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things:

a. does a set of predictor variables do a good job in predicting an outcome (dependent) variable?

b. Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable?

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.

A linear regression line has an equation of the form $Y = a + bX$, where X is the explanatory variable and Y is the dependent variable. The slope of the line is b, and a is the intercept (the value of y when $x = 0$).

The most common method for fitting a regression line is the method of least-squares. This method calculates the best-fitting line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to

the line (if a point lies on the fitted line exactly, then its vertical deviation is 0). Because the deviations are first squared, then summed, there are no cancellations between positive and negative values.

General steps we follow in Simple Linear Regression:

1. **Estimating the coefficients**
   Let's assume we only have one variable and one target. Then, linear regression is expressed as: Y = a + bX. In the equation above, the betas are the coefficients. These coefficients are what we need in order to make predictions with our model.
   So how do we find these parameters?
   To find the parameters, we need to minimize the least squares or the sum of squared errors. Of course, the linear model is not perfect and it will not predict all the data accurately, meaning that there is a difference between the actual value and the prediction.
   The error is easily calculated with: $e_i = y_i - \hat{y_i}$
2. **Estimate the relevancy of the coefficients**

   Now that you have coefficients, how can you tell if they are relevant to predict your target?

   The best way is to find the p-value. The p-value is used to quantify statistical significance; it allows to tell whether the null hypothesis is to be rejected or not.For any modelling task, the hypothesis is that there is some correlation between the features and the target. The null hypothesis is therefore the opposite: there is no correlation between the features and the target.

   So, finding the p-value for each coefficient will tell if the variable is statistically significant to predict the target. As a general rule of thumb, if the p-value is less than 0.05: there is a strong relationship between the variable and the target.

3. **Assess the accuracy of the model**

   You found out that your variable was statistically significant by finding its p-value. Now, how do you know if your linear model is any good?

   To assess that, we usually use the RSE (residual standard error) and the R^2 statistic.

2. **What is Anscombe's Quartet in detail**

Summary statistics allow us to describe a vast, complex dataset using just a few key numbers. This gives us something easy to optimize against and use as a barometer for our business. But there's a danger in relying only on summary statistics and ignoring the overall distribution.

the most elegant demonstration of the dangers of summary statistics is Anscombe's Quartet. It's a group of four datasets that appear to be similar when using typical summary statistics, yet tell four different stories when graphed. Each dataset consists of eleven (x,y) pairs as follows:

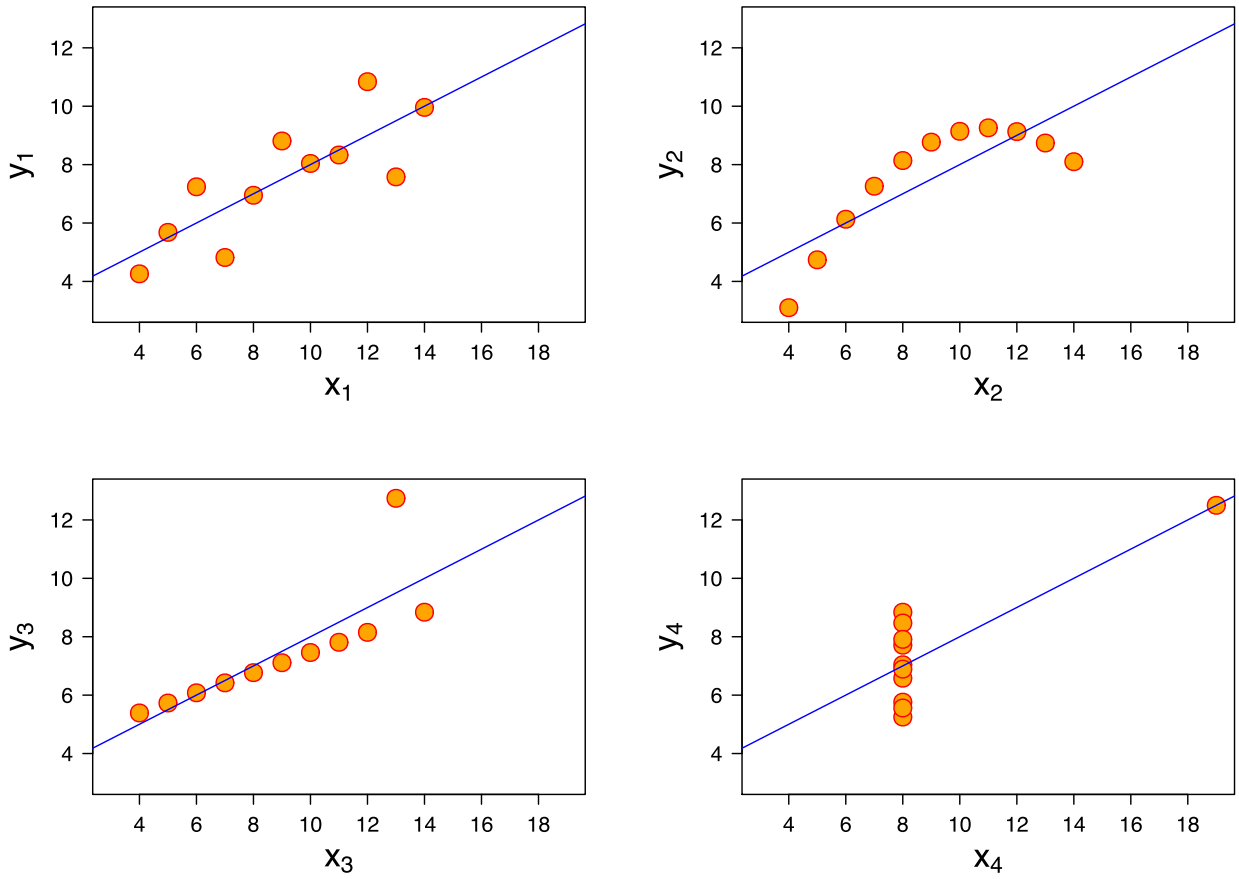
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

All the summary statistics you'd think to compute are close to identical:

- The average x value is 9 for each dataset
- The average y value is 7.50 for each dataset
- The variance for x is 11 and the variance for y is 4.12
- The correlation between x and y is 0.816 for each dataset

A linear regression (line of best fit) for each dataset follows the equation $y = 0.5x + 3$.

So far these four datasets appear to be pretty similar. But when we plot these four data sets on an x/y coordinate plane, we get the following results:



Now we see the real relationships in the datasets start to emerge. Dataset I consists of a set of points that appear to follow a rough linear relationship with some variance. Dataset II fits a neat curve but doesn't follow a linear relationship (maybe it's quadratic?). Dataset III looks like a tight linear relationship between x and y, except for one large outlier. Dataset IV looks like x remains constant, except for one outlier as well.

Computing summary statistics or staring at the data wouldn't have told us any of these stories. Instead, it's important to visualize the data to get a clear picture of what's going on.

3. **What is Peason's R?**

Correlation coefficients are used in statistics to measure how strong a relationship is between two variables. There are several types of correlation coefficient, but the most popular is Pearson's. Pearson's correlation (also called Pearson's R) is a correlation coefficient commonly used in linear regression

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation

coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

1. For the Pearson r correlation, both variables should be normally distributed. i.e the normal distribution describes how the values of a variable are distributed. This is sometimes called the 'Bell Curve' or the 'Gaussian Curve'. A simple way to do this is to determine the normality of each variable separately using the Shapiro-Wilk Test.
2. There should be no significant outliers. We all know what outliers are but we don't know the effect of outliers on Pearson's correlation coefficient, r. Pearson's correlation coefficient, r, is very sensitive to outliers, which can have a very large effect on the line of best fit and the Pearson correlation coefficient. This means — including outliers in your analysis can lead to misleading results.
3. Each variable should be continuous i.e. interval or ratios for example weight, time, height, age etc. If one or both of the variables are ordinal in measurement, then a Spearman correlation could be conducted instead.
4. The two variables have a linear relationship. Scatter plots will help you tell whether the variables have a linear relationship. If the data points have a straight line (and not a curve), then the data satisfies the linearity assumption. If the data you have is not linearly related you might have to run a non-parametric.
5. The observations are paired observations. That is, for every observation of the independent variable, there must be a corresponding observation of the dependent variable. For example if you're calculating the correlation between age and weight. If there are 12 observations of weight, you should have 12 observations of age. i.e. no blanks.
6. Homoscedascity. I've saved best for last. The hard is hard to pronounce but the concept is simple. Homoscedascity simply refers to 'equal variances'. A scatter-plot makes it easy to check for this. If the points lie equally on both sides of the line of best fit, then the data is homoscedastic. As a bonus — the opposite of homoscedascity is heteroscedascity which refers to refers to the circumstance in which the variability of a variable is unequal across the range of values of a second variable that predicts it.

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

Scaling 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.

$$x' = \frac{x - \bar{x}}{\sigma}$$

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.

$$\text{VIF} = \frac{1}{1 - R^2(x_1)}$$

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).

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

VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables. For example, we would fit the following models to estimate the coefficient of determination R1 and use this value to estimate the VIF.

As we see VIF s calculated based on R-squared value. If we have an R-squared value of 1(100% prediction) , VIF will become zero. Which tells us that the feature are highly correlated.

If there is perfect correlation, then VIF = infinity. We get perfect correlation when we compare a variable with itself. When we try to predict a variable with itself, we'll get an infinite VIF value.

$$X_{sc} = \frac{X - X_{min}}{X_{max} - X_{min}}.$$

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

Quantile-Quantile (Q-Q) plot, 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.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

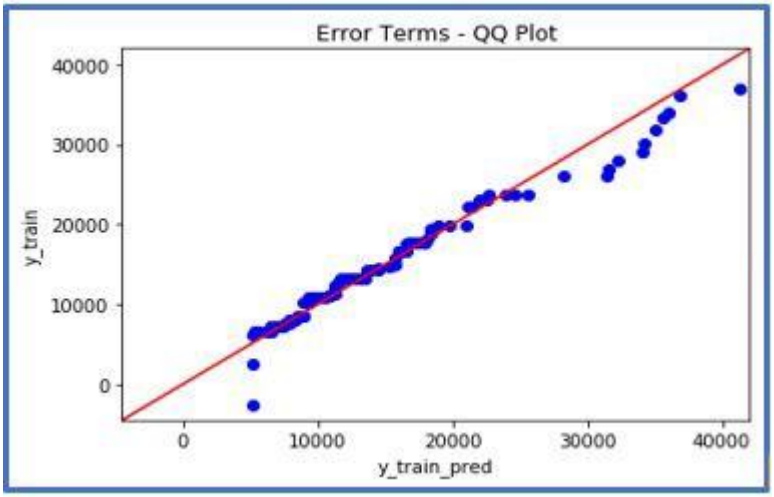
nterpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

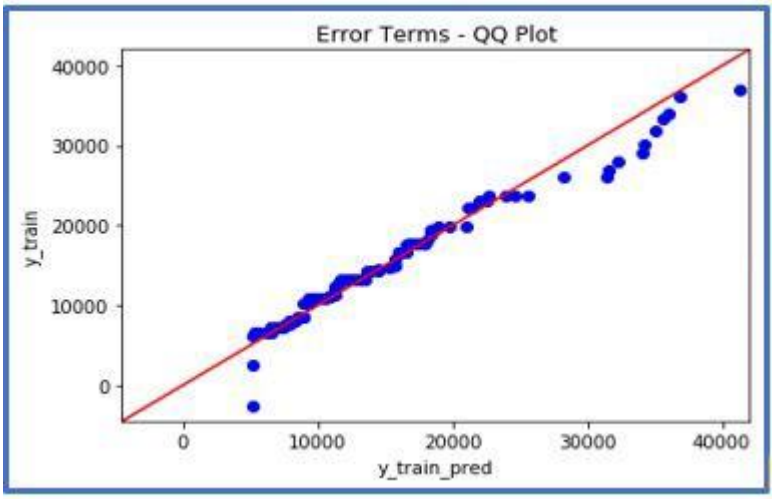
Below are the possible interpretations for two data sets.

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.



c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.



d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis