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?

A: The categorical variables present in the model are :

- Yr
- Season_2
- Season_4
- Mnth_8
- Mnth_9
- Weathersit_2
- Weathersit_3
- Holiday

❖ Yr : yr has a coefficient of '0.232'. The dependent variable 'cnt' changes linearly. With unit increase in yr variable,the count of bike rentals significantly increases by a factor of '0.232'.

❖ Season_2 : season_2 has a coefficient of '0.0999. The dependent variable 'cnt' changes linearly. With unit increase in season_2 variable,the count of bike rentals increases by a factor of '0.0999'.

❖ Season_4 : season_4 has a coefficient of '0.138' . The dependent variable 'cnt' changes linearly. With unit increase in season_4 variable,the count of bike rentals significantly increases by a factor of '0.138'.

❖ Mnth_8 : mnth_8 has a coefficient of '0.054' . The dependent variable 'cnt' changes linearly. With unit increase in mnth_8 variable,the count of bike rentals increases by a factor of '0.054'.

❖ Mnth_9 : mnth_9 has a coefficient of '0.116' . The dependent variable 'cnt' changes linearly. With unit increase in mnth_9 variable,the count of bike rentals increases by a factor of '0.116'.

❖ Weathersit_2 : weathersit_2 has a coefficient of '-0.082' . With unit increase in weathersit_2 variable,the count of bike rentals decreases by a factor of '0.082'.

❖ Weathersit_3 : weathersit_3 has a coefficient of '-0.282' . With unit increase in weathersit_3 variable,the count of bike rentals decreases by a factor of '0.282', which is a significant decrease.

❖ Holiday : holiday has a coefficient of '-0.097' . With unit increase in holiday variable,the count of bike rentals decreases by a factor of '0.097'.

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

A: Dummy variable creation is a process of converting categorical variables into numerical variables by assigning 0 or 1 values to each category. If we don't specify drop_first = True, It creates a problem of multicollinearity, which means that some of the dummy variables are linearly dependent on others. For example, if there are 3 dummy variables D_A,D_B,D_C

We can always infer the value of D_C from the values of D_A,D_B using the formula:

D_C=1-D_A-D_B. This can cause issues in regression analysis,such as **inflated standard errors,unreliable estimates, and poor model performance.**

To avoid this problem, we can use drop_first = True option in pandas.get_dummies() function,which drops the first dummy variable for each categorical variable.

This **reduces the number of dummy variables by one for each categorical variable and eliminates multicollinearity issue**.

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

A: atemp(with correlation coefficient of 0.63) and yr (with correlation coefficient of 0.57) have highest correlation coefficient with the target variable.

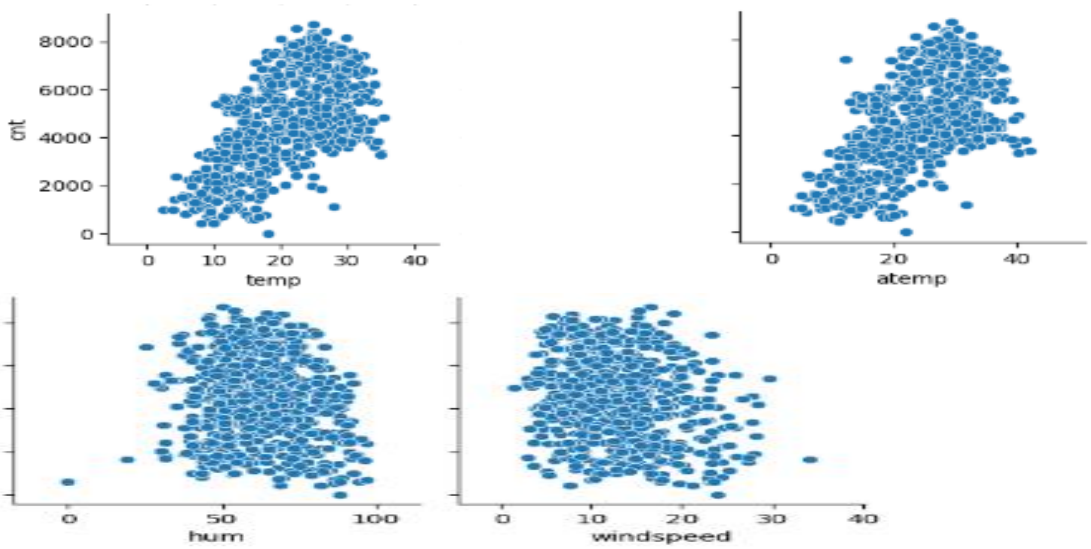
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

A: Linear relationship: There exists a linear relationship between the independent variables, and the dependent variable, 'cnt'. We can use a scatter plot of x vs. 'cnt' to visually inspect the linearity.

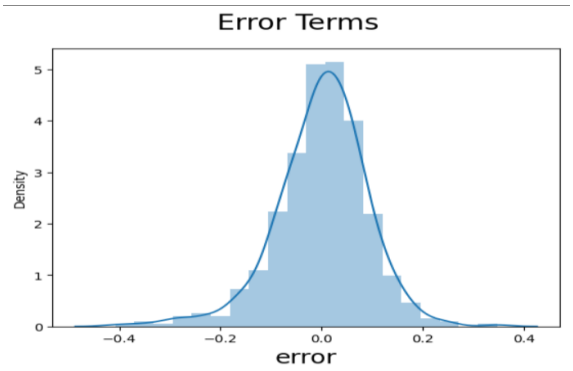
Homoscedasticity: The residuals have constant variance at every level of x. We can use a scatter plot of residuals vs. Fitted values.

Normality: The residuals of the model are normally distributed. We can use a Q-Q plot, which is a plot of the quantiles of the residuals vs. The quantiles of a normal distribution,to check for any deviations from normality,such as skewness or kurtosis.

The linearity of various numerical variables:



ERROR is in the form of Normal curve with mean 0.



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

A: The coefficients of three independent variables (temp, weathersit_3, and yr) that predict the dependent variable (bike-sharing count).

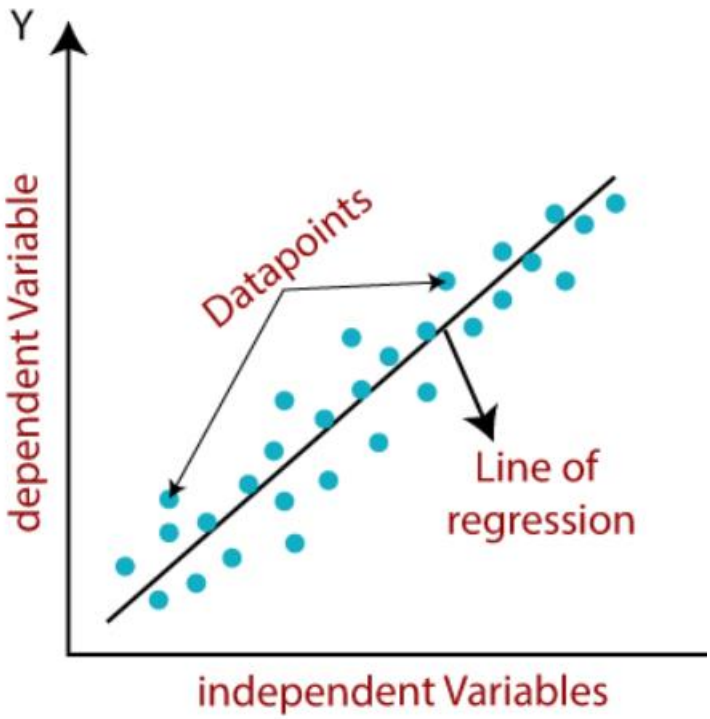
- The coefficients indicate the direction and magnitude of the effect of each independent variable on the dependent variable, holding the other variables constant.
- The coefficient of temp is positive and large, which means that higher temperatures are associated with higher bike-sharing counts.
- The coefficient of weathersit_3 is negative and moderate, which means that worse weather conditions are associated with lower bike-sharing counts.
- The coefficient of yr is positive and small, which means that later years are associated with higher bike-sharing counts, but the effect is not very strong.

General Subjective Questions

1) Explain the linear regression algorithm in detail.

A: It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as **sales, salary, age, product price,** etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.



$$y= a_0+a_1x+ \varepsilon$$

Y= Dependent Variable (Target Variable)
X= Independent Variable (predictor Variable)
a0= intercept of the line (Gives an additional degree of freedom)
a1 = Linear regression coefficient (scale factor to each input value).
ε = random error

The values for x and y variables are training datasets for Linear Regression model representation.

Types:

- **Simple_Linear_Regression(SLR):**
  If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.
- **Multiple_Linear_regression(MLR):**
  If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

Assumptions of Linear Regression model are:

Assumptions of Linear Regression

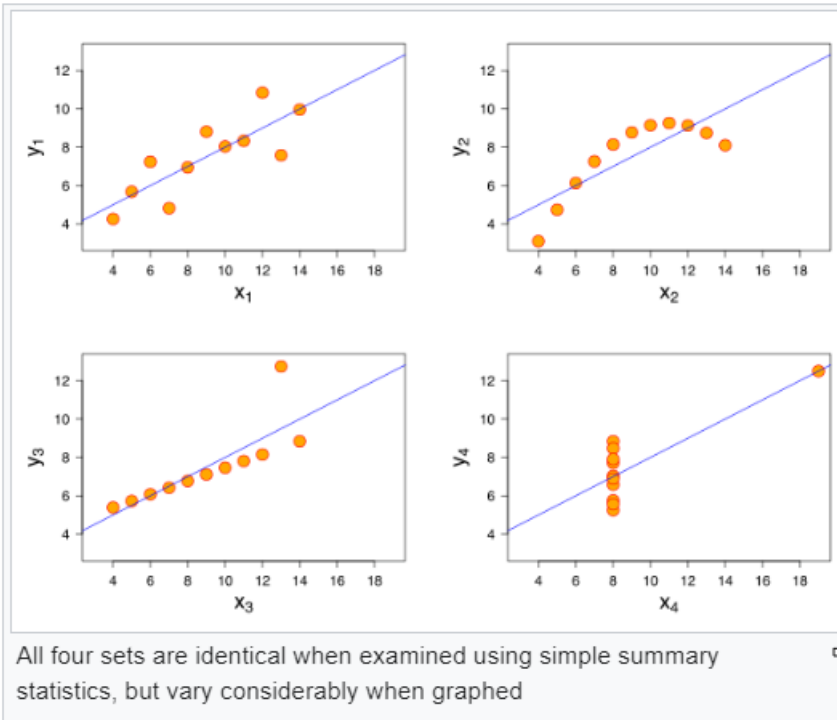
Below are some important assumptions of Linear Regression. These are some formal checks while building a Linear Regression model, which ensures to get the best possible result from the given dataset.

- **Linear relationship between the features and target:**
  Linear regression assumes the linear relationship between the dependent and independent variables.
- **Small or no multicollinearity between the features:**
  Multicollinearity means high-correlation between the independent variables. Due to multicollinearity, it may difficult to find the true relationship between the predictors and target variables. Or we can say, it is difficult to determine which predictor variable is affecting the target variable and which is not. So, the model assumes either little or no multicollinearity between the features or independent variables.
- **Homoscedasticity:**
  Homoscedasticity is a situation when the error term is the same for all the values of independent variables. With homoscedasticity, there should be no clear pattern distribution of data in the scatter plot.
- **Normal distribution of error terms:**
  Linear regression assumes that the error term should follow the normal distribution pattern. If error terms are not normally distributed, then confidence intervals will become either too wide or too narrow, which may cause difficulties in finding coefficients.
  It can be checked using the **q-q plot**. If the plot shows a straight line without any deviation, which means the error is normally distributed.

- **autocorrelations:**
  The linear regression model assumes no autocorrelation in error terms. If there will be any correlation in the error term, then it will drastically reduce the accuracy of the model. Autocorrelation usually occurs if there is a dependency between residual errors.

2) Explain the Anscombe's quartet in detail.

A: **Anscombe's quartet** comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (*x*, *y*) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough".



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

For all four datasets:

Property	Value	Accuracy

Mean of x	9	exact
Sample variance of x: s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y: s_y^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression:	0.67	to 2 decimal places

- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two correlated variables, where y could be modelled as gaussian with mean linearly dependent on x.
- For the second graph (top right), while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier, which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

The datasets are as follows. The x values are the same for the first three datasets.[1]

Anscombe's quartet			
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

It is not known how Anscombe created his datasets. Since its publication, several methods to generate similar data sets with identical statistics and dissimilar graphics have been developed.One of these, the *Datasaurus Dozen*, consists of points tracing out the outline of a dinosaur, plus twelve other data sets that have the same summary statistics

3) . What is Pearson's R?

A: Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier *product-moment* in the name.

For a population

Pearson's correlation coefficient, when applied to a population, is commonly represented by the Greek letter ρ (rho) and may be referred to as the *population correlation coefficient* or the *population Pearson correlation coefficient*. Given a pair of random variables (for example, Height and Weight), the formula for ρ is

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Statistical inference based on Pearson's correlation coefficient often focuses on one of the following two aims:

- One aim is to test the null hypothesis that the true correlation coefficient ρ is equal to 0, based on the value of the sample correlation coefficient *r*.
- The other aim is to derive a confidence interval that, on repeated sampling, has a given probability of containing ρ.

where

- n is sample size
- x_i, y_i are the individual sample points indexed with *i*
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$.

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

A: Scaling is the process of transforming data to a common range of values, such as between 0 and 1, or between -1 and 1. Scaling is performed to make the data compatible with most machine learning algorithms, which assume that the data is normally distributed and has similar scales. Scaling can also improve the performance and stability of the algorithms, and prevent numerical errors.

There are two common methods of scaling: normalized scaling and standardized scaling. Normalized scaling, also known as min-max scaling, rescales the data to the range of 0 to 1, by subtracting the minimum value and dividing by the maximum value. Standardized scaling, also known as z-score scaling, rescales the data to have zero mean and unit variance, by subtracting the mean and dividing by the standard deviation.

The difference between normalized scaling and standardized scaling is that normalized scaling preserves the original distribution of the data, while standardized scaling makes the data follow a standard normal distribution. Normalized scaling is sensitive to outliers, while standardized scaling is robust to outliers. Normalized scaling is suitable for data that has a fixed range, while standardized scaling is suitable for data that has an unknown or infinite range.

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

A: The formula for calculating the variance inflation factor (VIF) for a predictor variable X in a multiple linear regression model is:

VIF(X)=1/(1−R^2(X))

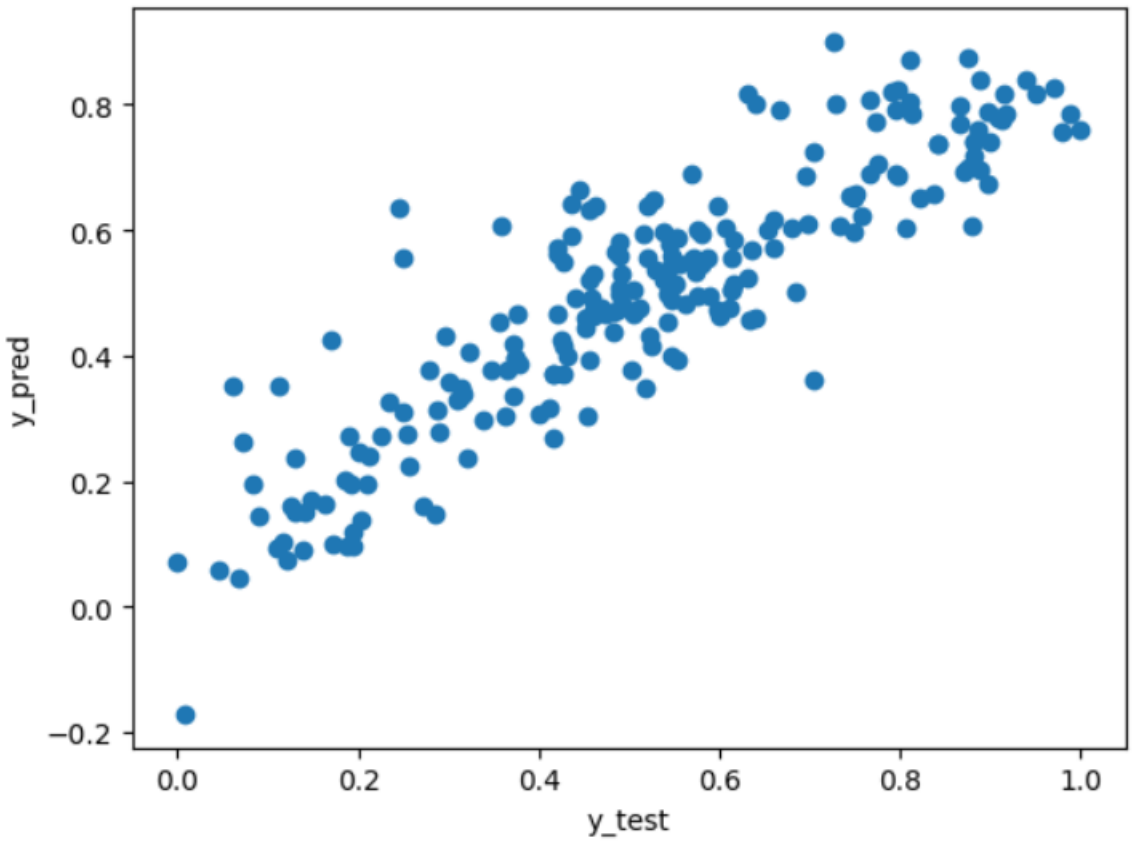
where R^2(X) is the coefficient of determination from a linear regression model where X is the dependent variable and all other predictor variables are used to predict X1. The VIF measures how much the variance of the estimated regression coefficient is increased because of collinearity. You can find more information and examples of how to use and interpret VIF in the web search results that I have provided

The value of VIF (variance inflation factor) is infinite when there is perfect multicollinearity between the independent variables in a regression model. This means that one or more independent variables can be exactly predicted by a linear combination of the other independent variables. This causes the denominator of the VIF formula, which is 1 minus the R-squared of the regression of the independent variable on the other independent variables, to be zero. Therefore, the VIF becomes infinite. To avoid this problem, we need to remove the redundant variables from the model or use a different method of estimation.

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

A: A Q-Q plot, short for quantile-quantile plot, is a graphical tool that compares the distribution of the residuals of a linear regression model to a normal distribution. It plots the quantiles of the standardized residuals against the quantiles of a standard normal distribution. If the points on the plot roughly form a straight diagonal line, then the normality assumption of the residuals is met. This is important because many inference methods and tests for linear regression rely on the normality assumption.

A Q-Q plot can also help to detect and characterize the deviations from normality, such as skewness, kurtosis, outliers, and heteroscedasticity. These can affect the accuracy and validity of the linear regression model and may require some transformations or corrections. You can find more information and examples of how to use and interpret Q-Q plots in linear regression in the web search results that I have provided.


y_pred
0.8
0.6
0.4
0.2
0.0
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
y_test