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?

1. **YR :** 2019 year have a higher average rental bike per day than overall average rental
2. **MNTH :** 6,9,8,7,5 and 10 months have a higher average rental bike per day than overall average rental
3. 6,9,8,7,5 and 10 have a higher average rental bike per day than 4,11,3,12,2,1
4. **September month** has maximum demand of bikes and January month has minimum demands of bikes.
5. **Demand** of bikes has increased in year **2019.**
6. **HOLIDAY :** Non Holidays have a higher average rental bike per day than overall average rental
7. Non Holidays have a higher average rental bike per day than Holidays
8. **WEEKDAY** : Weekday 2,3,4,5,6 have a higher average rental bike per day than overall average rental
9. Weekday 2,3,4,5,6 have a higher average rental bike per day than Weekday 1,2
10. **WORKINGDAY** : Working day have a higher average rental bike per day than overall average rental
11. Working day have a higher average rental bike per day than Non Working Day
12. **WEATHERSIT :** Clear, Few clouds, Partly cloudy, Partly cloudy Weather have a higher average rental bike per day than overall average rental
13. Clear, Few clouds, Partly cloudy, Partly cloudy Weather have a higher average rental bike per day than Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist, Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

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

It is important to use drop first as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. By dropping one of the one-hot encoded columns from each categorical feature, we make sure that there are no reference columns and the remaining columns become linearly independent. If there are n categorical values then n-1 dummy columns have to be created. In the Booming Bikes Sharing Assignment the following are the categorical columns for which we have to create the dummy variables.

a. SEASON

b. MNTH

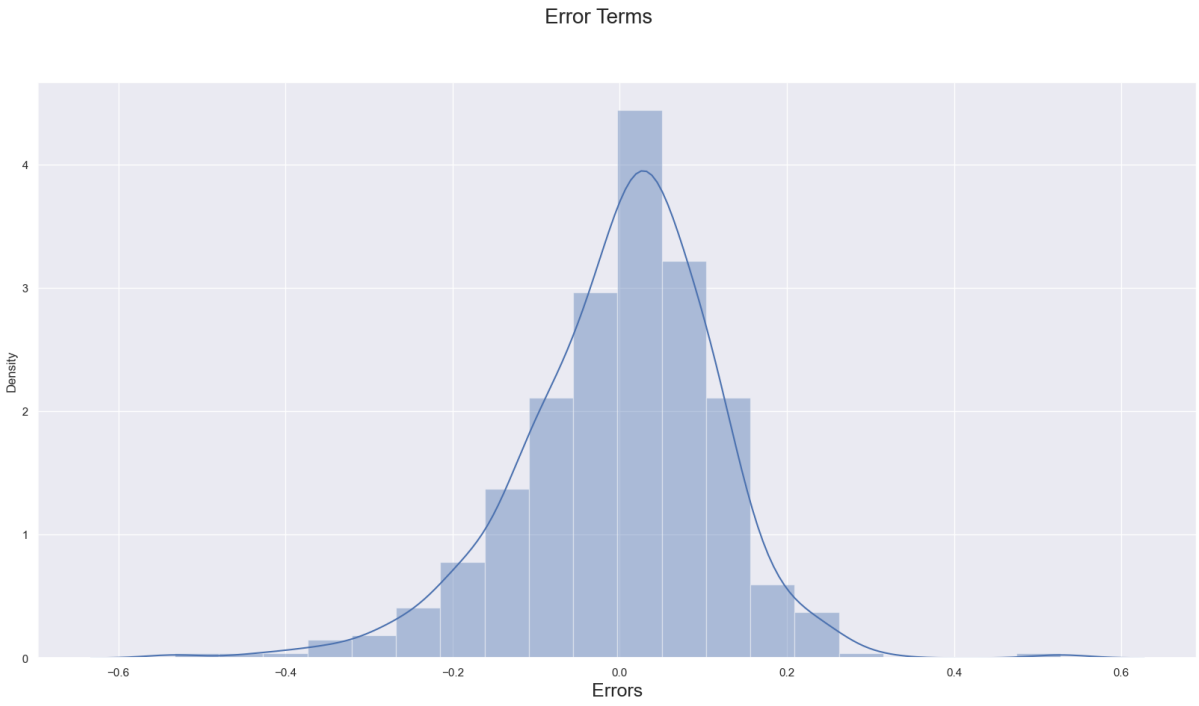
c. WEEKDAY

d. WEATHERSIT

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

Registered variable has the highest correlation with the target variable (**cnt**).

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



After building the model on training set, Residual Analysis validates the assumptions of Linear Regression. Dist. plot shows that the error terms are normally distributed with mean equal to zero.

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

Based on the final model, top 3 features contributing significantly towards explaining the demand of the shared bikes are:

1. **weathersit - Light Snow and Light Rain**
2. **Season - spring**
3. **Year - yr**

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the **dependent variable**. The variable you are using to predict the other variable's value is called the **independent variable**.

This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a "**least squares**" method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).

Naming the Variables. There are many names for a regression's dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regressand. The independent variables can be called exogenous variables, predictor variables, or regressors.

Types of Regression

Simple linear regression

1 dependent variable (interval or ratio), 1 independent variable

(Interval or ratio or dichotomous)

Multiple linear regression

1 dependent variable (interval or ratio), 2+ independent variables (interval or ratio or dichotomous)

Logistic regression

1 dependent variable (dichotomous), 2+ independent variable(s) (interval or ratio or dichotomous)

Ordinal regression

1 dependent variable (ordinal), 1+ independent variable(s) (nominal or dichotomous)

Multinomial regression

1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio or dichotomous)

Discriminant analysis

1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio)

Application of model for Linear Regression:

(1) Determining the strength of predictors

(2) Forecasting an effect

(3) Trend forecasting

Formula and Calculation of Linear Regression:

y = c + b*x (Simple linear regression)

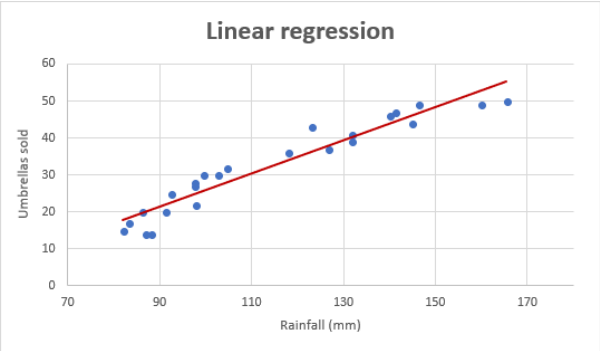
Where y = estimated dependent variable score,

c = constant,

b = regression coefficient, and

x = score on the independent variable.

Graph for Linear Regression:



Significance (P - value use to predict):

A p-value is **a statistical measurement used to validate a hypothesis against observed data**. A p-value measures the probability of obtaining the observed results, assuming that the null hypothesis is true. The lower the p-value, the greater the statistical significance of the observed difference.

(Less than 0.05 is significance for Hypothesis testing.)

Multiple Linear Regression (MLR)

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression is to model the linear relationship between the explanatory (independent) variables and response (dependent) variables. In essence, multiple regression is the extension of ordinary least-squares (OLS) regression because it involves more than one explanatory variable.

KEY POINTS:

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable.

Multiple regression is an extension of linear (OLS) regression that uses just one explanatory variable.

Application of model: (MLR is used extensively in)

5. Econometrics
6. Financial inference.

Formula and Calculation of Multiple Linear Regression

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$$

Where, for i=n observations:

y = dependent variable

x = explanatory variables

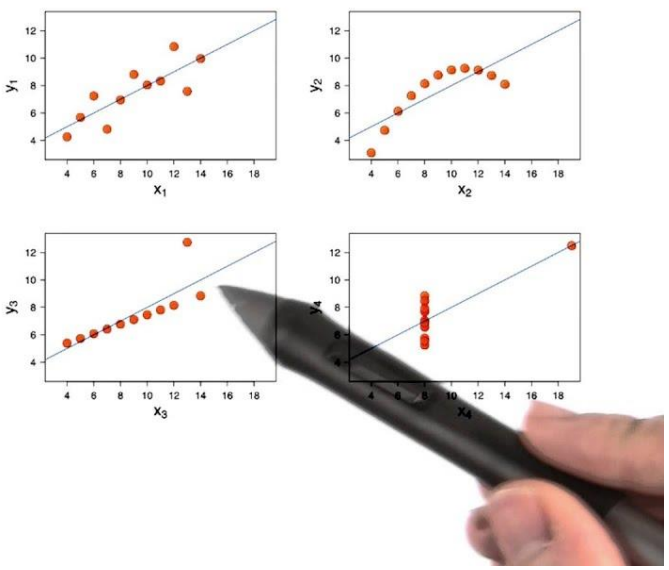
β0 = y-intercept (constant term)

β1 =slope coefficients for each explanatory variable

ϵ=the model's error term (also known as the residuals)

2. Explain the Anscombe's quartet in detail.

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



Anscombe's quartet consist of four data sets which have nearly identical simple descriptive statistics. But still they have very different distributions and they appear very different when they are graphed. Each data set comprises of 11 (x, y) points. They were constructed in 1973 by statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers & other influential observations on statistical properties.

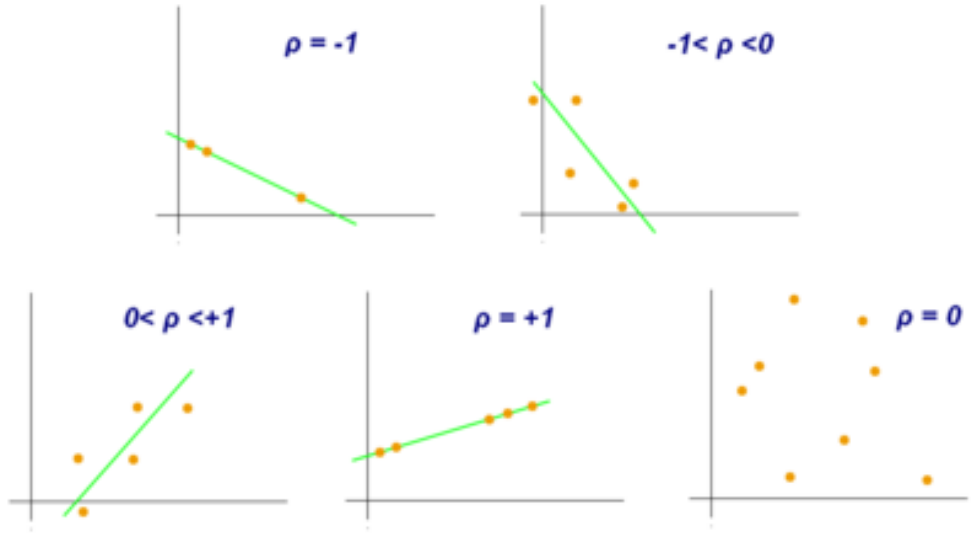
3. What is Pearson's R?

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that **measures the strength and direction of the relationship between two variables**.

Mathematically, it is the ratio between the covariance of two variables and the product of their standard deviations. When it is applied to population then it is denoted by Greek letter ρ (rho) Graphs for different values of ρ are

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

1. r = Pearson Coefficient.
2. n= number of the pairs of the stock.
3. ∑xy = sum of products of the paired stocks.
4. ∑x = sum of the x scores.
5. ∑y= sum of the y scores.
6. $\sum x^2$ = sum of the squared x scores.
7. $\sum y^2$ = sum of the squared y scores.



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

Scaling means that you're transforming your data so that it **fits within a specific scale**, like 0-100 or 0-1. You want to scale data when you're using methods based on measures of how far apart data points, like support vector machines, or SVM or k-nearest neighbours, or KNN.

When you have lots of independent variables which are on different scale in a model then scaling is performed for the ease of interpretation and for faster convergence for gradient descent method.

If the data in any conditions has data points far from each other, scaling is a technique to make them closer to each other.

In Normalization is used when the **data doesn't have Gaussian distribution**.

In Standardization is used on **data having Gaussian distribution**.

In normalized scaling, variables are scaled in such a way that all values lie between zero and one **using maximum and minimum values in the data**.

In standardized scaling, variables are scaled in such a way that their mean is zero and **standard deviation is one**.

Standardisation (Z-score Normalization)	Max-Min Normalization
$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation } (x)}$	$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$

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

An infinite VIF value indicates that the variable with infinite VIF value is highly correlated with other variables of the model and R square of this variable is equal to 1.

If all the independent variables are orthogonal to each other, then VIF = 1.0. **If there is perfect correlation**, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.

Values of VIF that exceed 10 are often regarded as indicating multicollinearity, but in weaker models values above 2.5 may be a cause for concern.

$$VIF_i = \frac{1}{1-R_i^2}$$

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.