Name: Tapas Ranjan Nayak. Batch : C34

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? (3 marks)

Ans:

Following are few observations:

- Number of bikes rented is high when the weather is good.
- Number of bikes rented is high in fall season.
- Number of bikes rented goes high during mid of the week.
- Number of bikes rented goes high during mid of the year.

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

Ans:

drop_first=True is used to reduce the extra columns created during dummy variable creation. Hence it reduces the correlations created among dummy variables. This is important to use. By default, it is False.

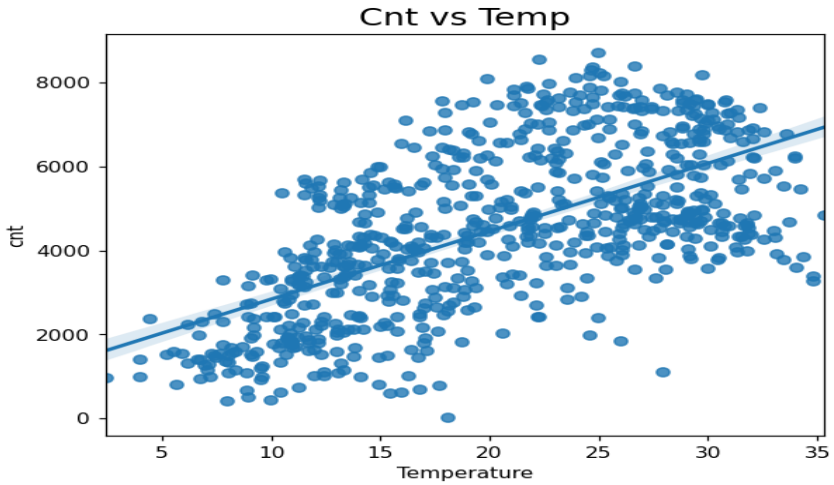
Syntax Example:

```
dummies = pd.get_dummies(df[['column_1']], drop_first=True)
```

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans:

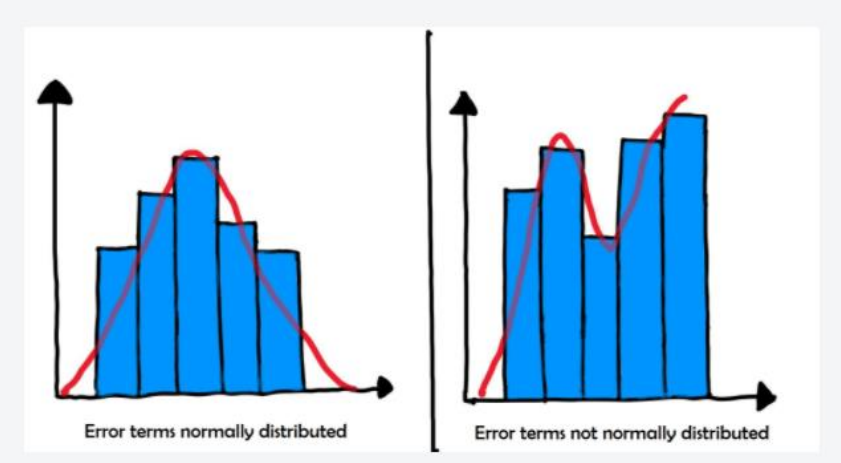
Temp



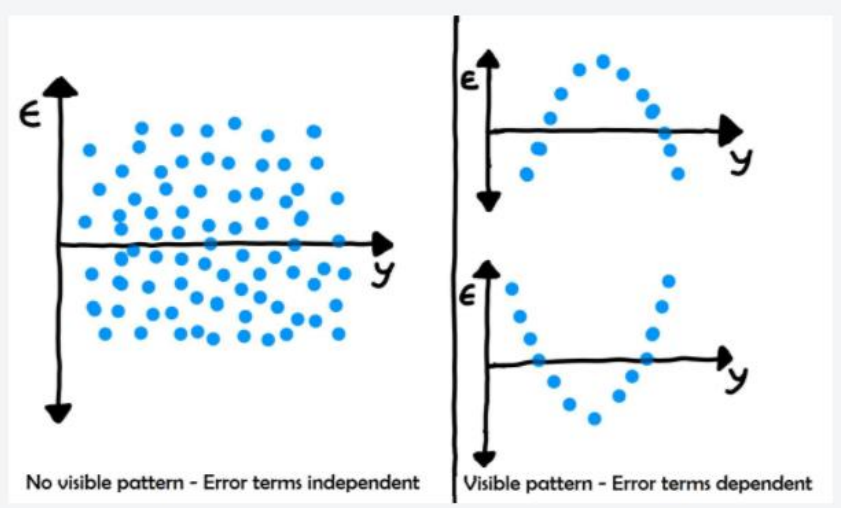
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans:

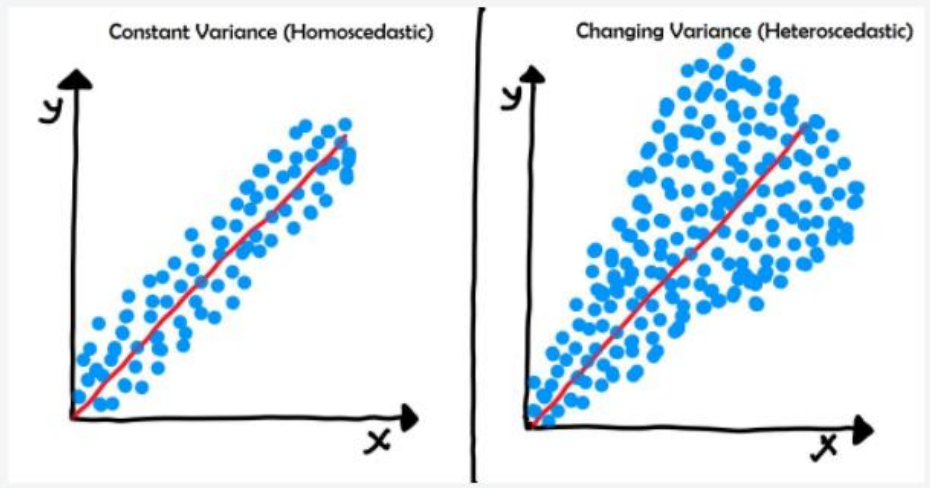
1. A simple pairplot of the dataframe can help us see if the independent variables exhibit linear relationship with the Dependent Variable.
2. Use Distribution plot on the residuals and see if error terms normally distributed like below.



3. **Error terms are *independent* of each other like below plot.**



4. **No** Heteroskedasticity**.** Residual vs Fitted values plot can tell if Heteroskedasticity is present or not. If the plot shows a funnel shape pattern, then we say that Heteroskedasticity is present. Below plot explains.



5. No Perfect Multicollinearity, this can be checked with use of heatmap in case of small set of variables, but for large variables, the calculation of VIF. If VIF=1 very less multicollinearity. VIF>5 extreme collinearity and VIF <5 moderate multicollinearity.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
Ans:

Temperature, Holiday and humidity.

General Subjective Questions:

1. Explain the linear regression algorithm in detail. (4 marks)
Ans:
Linear regression algorithm is a very simple machine learning approach for supervised learning. It performs a regression task. It performs the task to predict a dependent variable(target) based on the given independent variable(s). So, this regression technique finds out a linear relationship between a dependent variable and the other given independent variables.
Linear regression is one of the most commonly used ML algorithms for predictive

analysis. The overall idea of regression is to examine two things:

- Does a set of predictor variables do a good job in predicting an outcome (dependent) Variable?
- 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?

These regression estimates are in turn used to explain the relationship between one dependent variable with one or more independent variables.

The simplest form of the regression equation with one dependent and one independent variable is defined by the formula **y = b0 + b1*x**, where y = estimated dependent variable score, **b0 = constant, b1 = regression coefficient, and x = score on the independent variable**.

Three major uses for regression analysis are:

- determining the strength of predictors
- forecasting an effect, and
- trend forecasting.

First, the regression might be used to identify the strength of the effect that the independent variable(s) have on a dependent variable. Typical questions are what is the strength of relationship between dose and effect, sales and marketing spending, or age and income.

Second, it can be used to forecast effects or impact of changes. That is, the regression analysis helps us to understand how much the dependent variable changes with a change in one or more independent variables. A typical question is, “how much additional sales income do I get for each additional $1000 spent on marketing?”

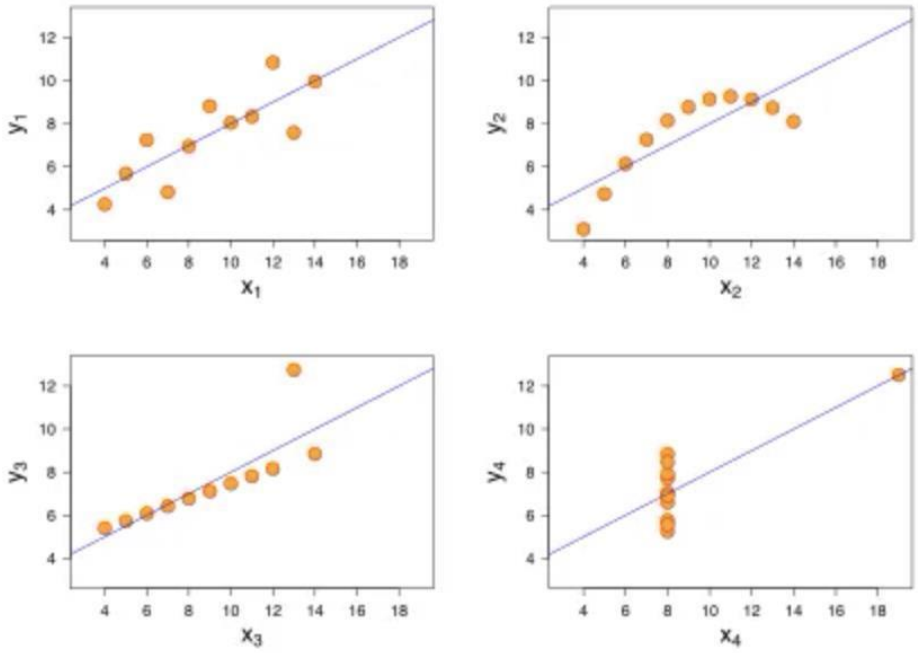
Third, regression analysis predicts trends and future values. The regression analysis can be used to get point estimates. A typical question is, “what will the price of gold be in 6 months?”

Linear regression models are often fitted using the least squares approach, but they may also be fitted in other ways, such as by minimizing the "lack of fit" in some other norm (as with least absolute deviations regression), or by minimizing a penalized version of the least squares cost function as in ridge regression (L2-norm penalty) and lasso (L1-norm penalty). Conversely, the least squares approach can be used to fit models that are not linear models. Thus, although the terms "least squares" and "linear model" are closely linked, they are not synonymous.

2. Explain the Anscombe’s quartet in detail. (3 marks)

Ans:

Anscombe’s quartet comprises four datasets that have nearly identical simple statistical properties, yet 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 before analyzing it and the effect of outliers on statistical properties.



The first scatter plot (top left) appears to be a simple linear regression, corresponding to two variables correlated and following the assumption of normality. The second graph (top right) is not distributed normally; while an obvious relationship between the two variables can be observed, it is not linear, and the Pearson’s Correlation Coefficient is not relevant. In the third graph (bottom left), the distribution is linear, but with a different regression line, which is offset by the one outlier, which exerts enough influence to alter the regression line and lower the correlation coefficient from 1 to 0.816. Finally, the fourth graph (bottom right) shows another example when one outlier is enough to produce a high correlation coefficient, even though the relationship between the two variables is not linear.

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyse 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.

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

3. What is Pearson's R? (3 marks)

Ans:

It is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

In a sample it is denoted by r and is by design constrained as follows: Furthermore,

1. Positive values denote positive linear correlation.

2. Negative values denote negative linear correlation.

3. 0 denotes no linear correlation.

4. The closer the value is to 1 or -1, the stronger the linear correlation. It Is given

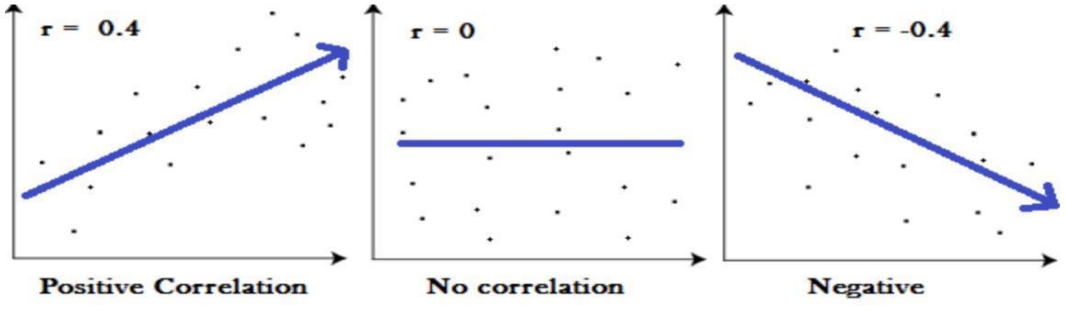
by, There are several types of correlation coefficient formulas.

One of the most commonly used formulas in stats is Pearson's correlation coefficient formula. If you're taking a basic stats class, this is the one you'll probably use:

$$\rho_{X,Y} = \frac{\mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y]}{\sqrt{\mathrm{E}[X^2] - [\mathrm{E}[X]]^2}\,\sqrt{\mathrm{E}[Y^2] - [\mathrm{E}[Y]]^2}}.$$

Correlation coefficient formulas are used to find how strong a relationship is between data.
The formulas return a value between -1 and 1, where:

- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all.



Graphs showing a correlation of -1, 0 and +1

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans:

Scaling (also known as **data normalization**) is the method used to standardize the range of features of data. Since, the range of values of data may vary widely, it becomes a necessary step in data pre-processing while using machine learning algorithms.

In scaling (also called min-max scaling), you transform the data such that the features are within a specific range e.g. [0, 1].

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

where x' is the normalized value.

It is mainly performed because most of the times, your dataset will contain features highly varying in magnitudes, units and range. But since, most of the machine learning algorithms use Euclidian distance between two data points in their computations, this is a problem. If left alone, these algorithms only take in the magnitude of features neglecting the units. The results would vary greatly between different units, 5kg and 5000gms. The features with high magnitudes will weigh in a lot more in the distance calculations than features with low magnitudes. To suppress this effect, we need to bring all features to the same level of magnitudes. This can be achieved by scaling.

Normalized Scaling:

Also known as min-max scaling or min-max normalization, is the simplest method and consists in rescaling the range of features to scale the range in [0, 1] or [−1, 1]. Selecting the target range depends on the nature of the data. The general formula for a min-max of [0, 1] is given as:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

where x is an original value, x' is the normalized value. For example, suppose that we have the students' weight data, and the students' weights span [160 pounds, 200 pounds]. To rescale this data, we first

subtract 160 from each student's weight and divide the result by 40 (the difference between the maximum and minimum weights).

To rescale a range between an arbitrary set of values [a, b], the formula becomes:

$$x' = a + \frac{(x - \min(x))(b - a)}{\max(x) - \min(x)}$$

where a, b are the min-max values. Mean normalization

$$x' = \frac{x - \text{average}(x)}{\max(x) - \min(x)}$$

where x is an original value, x' is the normalized value. There is another form of the mean normalization which is when we divide by the standard deviation which is also called standardization.

Standardized Scaling:

In machine learning, we can handle various types of data, e.g. audio signals and pixel values for image data, and this data can include multiple dimensions. Feature standardization makes the values of each feature in the data have zero-mean (when subtracting the mean in the numerator) and unit-variance. This method is widely used for normalization in many machine learning algorithms (e.g., support vector machines, logistic regression, and artificial neural networks). The general method of calculation is to determine the distribution mean and standard deviation for each feature. Next we subtract the mean from each feature. Then we divide the values (mean is already subtracted) of each feature by its standard deviation.

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

Where x is the original feature vector, x = average (x) is the mean of that feature vector, and σ is its standard deviation. Scaling to unit length

Another option that is widely used in machine-learning is to scale the components of a feature vector such that the complete vector has length one. This usually means dividing each component by the Euclidean length of the

vector:

$$x' = \frac{x}{\|x\|}$$

In some applications (e.g. Histogram features) it can be more practical to use the L1 norm (i.e. Manhattan Distance, City-Block Length or Taxicab Geometry) of the feature vector. This is especially important if in the following learning steps the Scalar Metric is used as a distance measure.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Ans:

The extent to which a predictor is correlated with the other predictor variables in a linear regression can be quantified as the R-squared statistic of the regression where the predictor of interest is predicted by all the other predictor variables. The variance inflation for a variable is then computed as:

$$VIF = \frac{1}{1 - R^2}$$

When R-squared reaches 1, VIF reaches infinity. When R-squared reaches 1 then it means multicollinearity exists. Different variables are highly correlated with each other.

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

Ans:

The quantile-quantile plot is a graphical method for determining whether two samples of data came from the same population or not. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value.

The Quantile-Quantile plot is used for the following purpose:

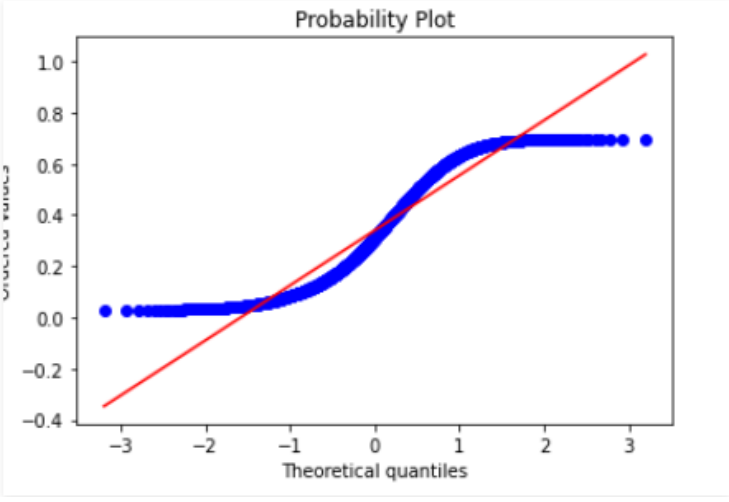
- Determine whether two samples are from the same population.
- Whether two samples have the same tail
- Whether two samples have the same distribution shape.
- Whether two samples have common location behavior.

Advantages of Q-Q plot

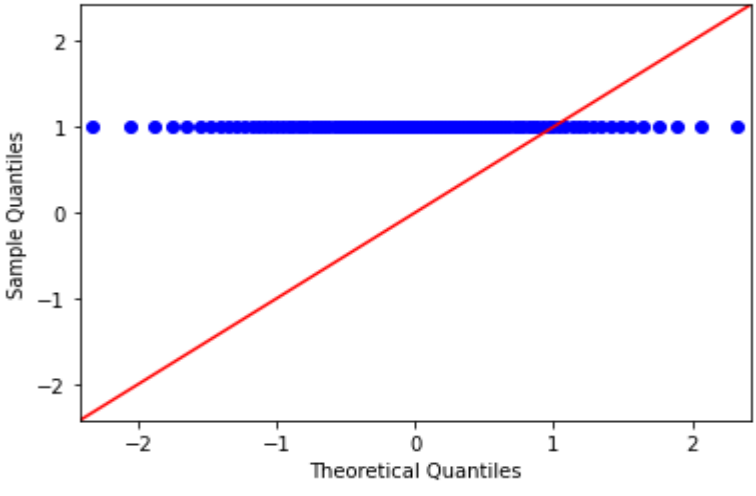
- Since Q-Q plot is like probability plot. So, while comparing two datasets the sample size need not to be equal.
- Since we need to normalize the dataset, so we don't need to care about the dimensions of values.

Types of Q-Q plots

- For Left-tailed distribution: Below is the



- For the uniform distribution: Below is the q-q plot distribution for uniform distribution:



uniform distribution Q-Q plot