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 the EDA inferences of categorical variables.

- more demand for bikes in summer and fall seasons
- Demand very less in rainy weather
- Demand is less on holiday
- Weekday and working day does not impact the bikes demand
- The bike demand increases gradually from jan till july and then it starts decreasing from july till dec.

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

Ans: We use **drop_first=True** to remove the first variable of the dummy columns.

- If there are n distinct categories for the given categorical variable, we can represent all the n values using binary data (0,1) of n-1 dummy variables.
- If we keep the n^{th} variable then the dummy variables have high correlation among them and it is not useful.
- Removing one variable also decreases the number of dummy columns that are added

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

Ans: temp, atemp has positive correlation with target variable cnt.

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

Ans: The Multiple Linear Regression model is validated based on the following:

- Plotted the error terms histogram. Its close to normal distribution, Mean value of error terms is almost 0. Hence this model best fit.
- The R-Squared(84.3%) and Adjusted R-Squared(83.9%) values of training data set are above 80 %
- The R-Squared value of test dataset is 80.5 %, the difference between training and test dataset R-Squared is 3.8% which is less than 5%

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: Following are the top 3 features that explain the demand of shared bikes

- Temp (coefficient: 0.4728)
- Rain (coefficient: -0.2917)
- Year (coefficient: 0.2344)

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Ans: Linear Regression is an algorithm that belongs to supervised Machine Learning. It tries to apply relations that will predict the outcome of an event based on the independent variable data points. The relation is usually a straight line that best fits the different data points as close as possible. The output is of a continuous form,

Eg: Revenue or Sales in currency,

Number of products sold, etc.

Linear regression can be expressed mathematically as: **$y = \beta 0 + \beta 1 x$**

Y= Dependent Variable

X= Independent Variable

β 0= intercept of the line

β1 = Linear regression coefficient (slope of the line)

Best-fit line — the line which fits the given scatter-plot in the best way.

We use Ordinary least squares method to find Best fit line

RSS (Residual Sum of Squares): In statistics, it is defined as the total sum of error across the whole sample. It is the measure of the difference between the expected and the actual output. A small RSS indicates a tight fit of the model to the data. It is also defined as follows:

$$RSS = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

The strength of the linear regression model can be assessed using **R^2 or Coefficient of Determination**

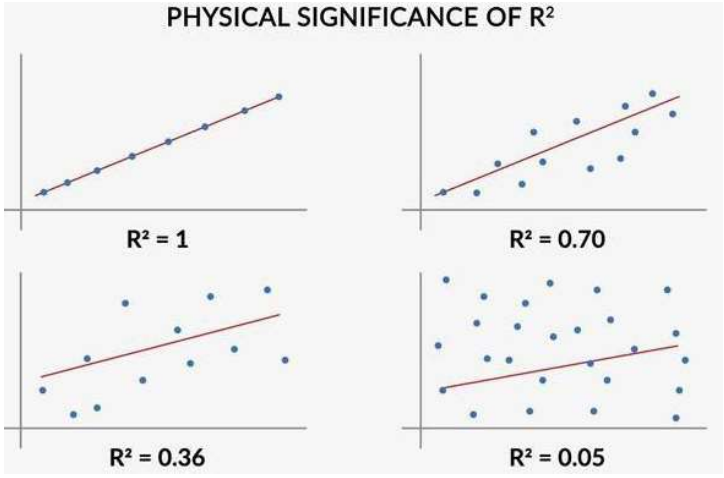
R2 is a number which explains what portion of the given data variation is explained by the developed model. It always takes a value between 0 & 1. The higher the R-squared, the better the model fits your data.

$R^2 = 1 - (RSS / TSS)$

RSS (Residual Sum of Squares) – explained above

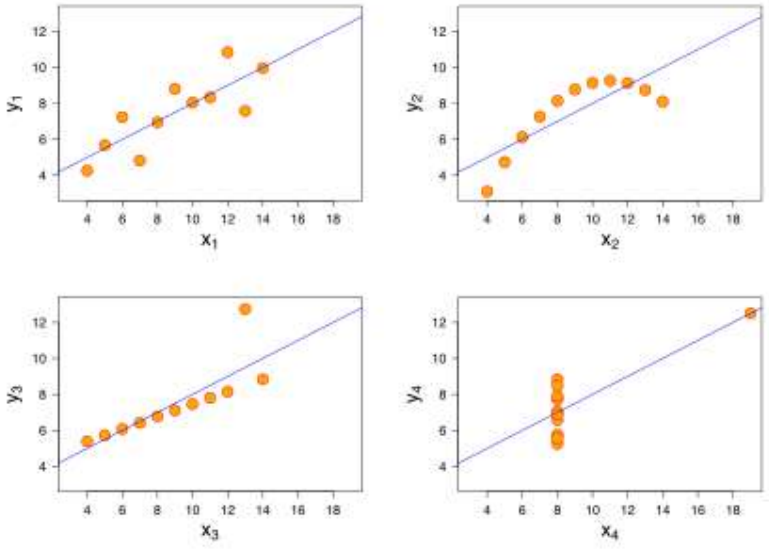
TSS (Total sum of squares): It is the sum of errors of the data points from mean of response variable. Mathematically, TSS is:

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$



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

Ans: Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots. It was constructed in 1973 by statistician Francis Anscombe to **illustrate the importance of plotting the graphs before analyzing and model building**



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

Below is the statistical summay for all the above 4 sets

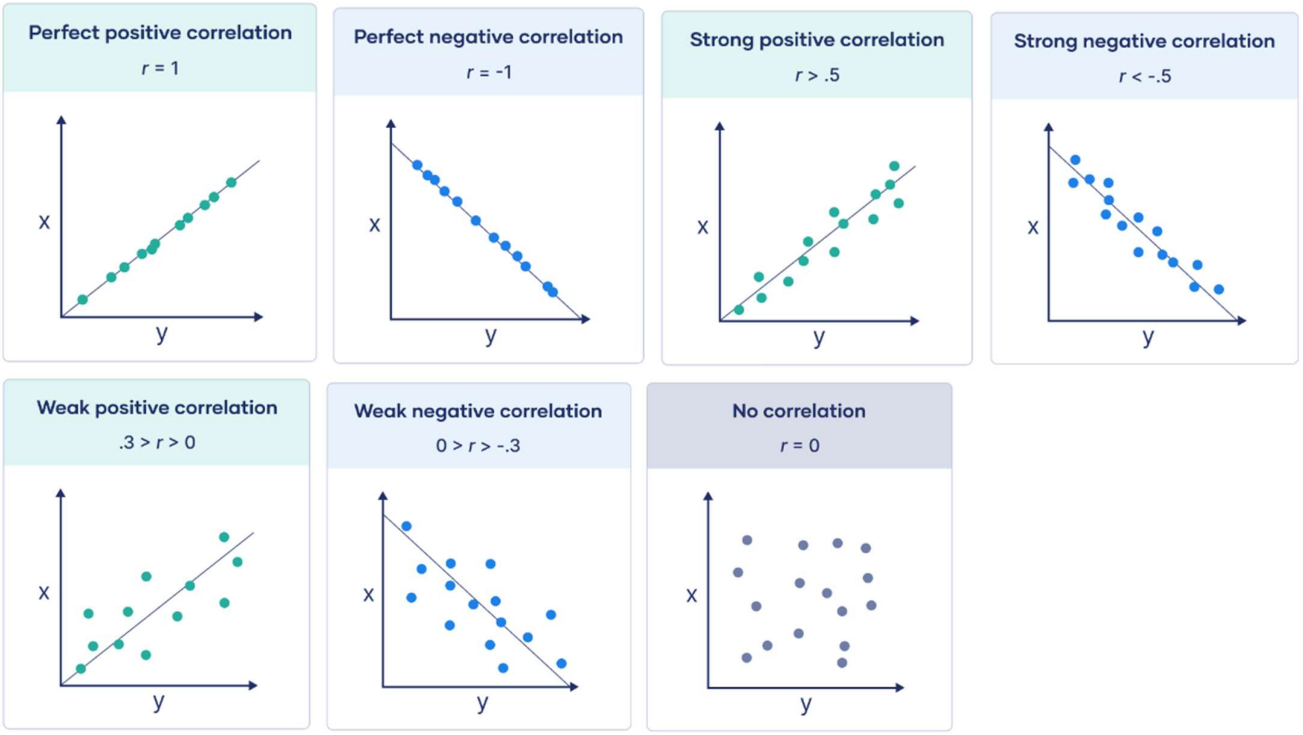
Property	Value
Mean of x	9
Sample variance of x : s_x^2	11
Mean of y	7.50
Sample variance of y : s_y^2	4.125
Correlation between x and y	0.816
Linear regression line	$y = 3.00 + 0.500x$
Coefficient of determination of the linear regression : R^2	0.67

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

Ans: The Pearson correlation coefficient also known as Pearson's 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.

R between 0 and 1 – positive correlation

R between 0 and -1 – negative correlation



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

Ans:

Scaling - In machine learning the feature scaling means bringing all the feature values into the same range.

Why Scaling - Scaling is important because scaling enables the coefficients of all the independent variables comparable.

There are 2 methods of scaling

a. Standardisation: Standardisation basically brings all of the data into a standard normal distribution with mean zero

   x=(x−mean(x)) / sd(x)

b. Min Max Scaling: MinMax scaling, on the other hand, brings all of the data in the range of 0 and 1

$x=(x-\min(x)) / (\max(x)-\min(x))$

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.

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

Ans: VIF = $1/(1-R^2)$

In the case of perfect positive($R=1$) or negative($R=-1$) correlation, we get $R^2 = 1$, which makes VIF infinity as per the above formula.

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

Ans: In Statistics, Q-Q(quantile-quantile) plots play a very vital role to graphically analyse and compare two probability distributions by plotting their quantiles against each other. If the two distributions which we are comparing are exactly equal then the points on the Q-Q plot will perfectly lie on a straight-line y = x.