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?

Ans:

We can infer the following points from the analysis of categorical variable:

1. From the season and mnth column, we can predict that shared bike demand is increasing in mid-year such as in summer and fall seasons.
2. People not prefer bikes when there is snow.
3. In 2019 demand of bikes got increase as compare 2018, it could be that craze of shared bikes is increasing day by day

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

Ans:

drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

Ans:

Looking at the pair plot, “temp” column has highest correlation with target variable. On first look we can see that casual and registered columns has highest correlation with target but it doesn’t make sense because target variable is addition of casual and registered column values.

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

Ans:

By validating error terms are normally distributed or not:

Steps:

- Find the predicted value of target variable
- Find the difference between actual and predicted value of target variable
- Plot the distplot of error term
- Make sure that it is normally distributed

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

Ans:

Top 3 Features are as follows:

1. Temp
2. windspeed
3. weathersit

General Subjective Questions

1. Explain the linear regression algorithm in detail?

Ans:

Linear Regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. It's used to predict values within a continuous range, (e.g. sales, price) rather than trying to classify them into categories (e.g. cat, dog). There are two main types:

1.Simple regression:

Simple linear regression uses traditional slope-intercept form, where m and b are the variables our algorithm will try to "learn" to produce the most accurate predictions. x represents our input data and y represents our prediction.

$y=mx+b$

2.Multivariable regression:

A more complex, multi-variable linear equation might look like this, where w represents the coefficients, or weights, our model will try to learn.

$f(x,y,z)=w1x+w2y+w3z$

The variables x,y,z represent the attributes, or distinct pieces of information, we have about each observation. For sales predictions, these attributes might include a company's advertising spend on radio, TV, and newspapers.

Sales=w1Radio+w2TV+w3News

2. Explain the Anscombe's quartet in detail.

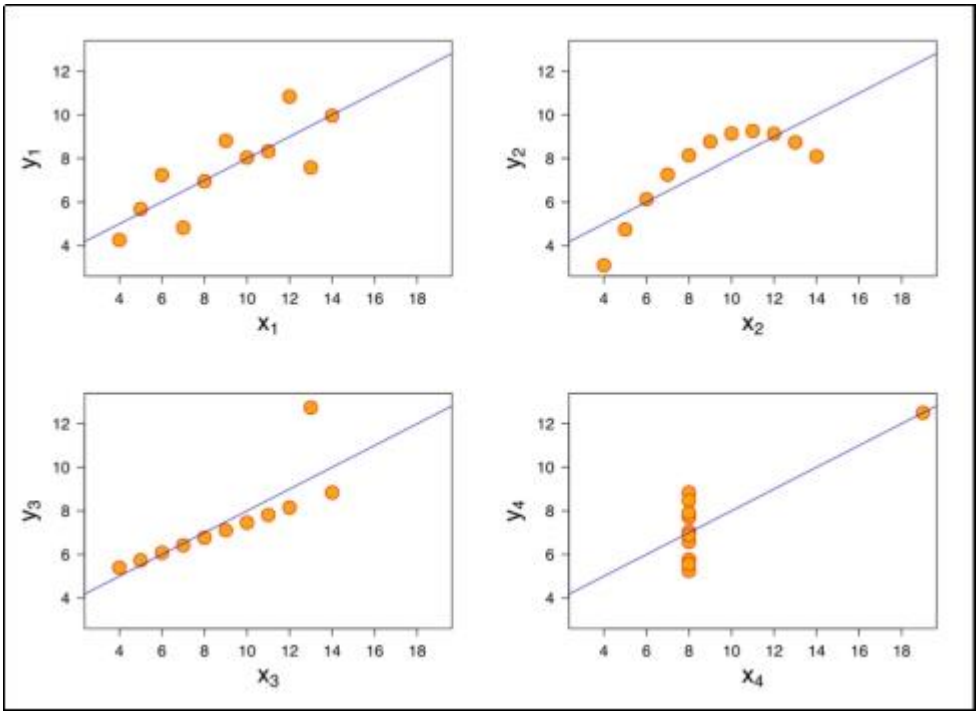
Ans:

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 before analysing it and the effect of outliers and other influential observations on statistical properties,

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : S_x	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : S_y	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 : R^2	0.67	to 2 decimal places

From the graphical visualisation, we see that all the four datasets can be expressed by the same linear regression model : y = 3.00 + 0.500x



- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
- The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear and the Pearson correlation coefficient is not relevant.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line.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.

This is the dataset used,

Anscombe's quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

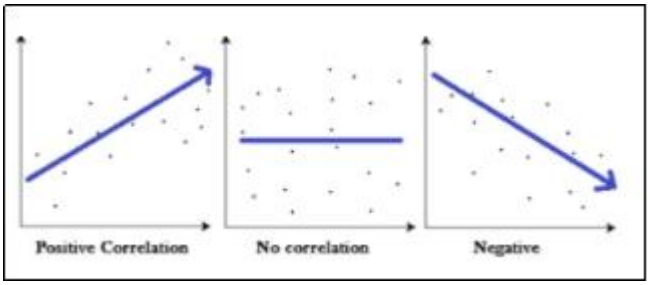
The quartet is used to illustrate the importance of looking at a set of data graphically before starting to analyse particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets

3. What is Pearson's R?

Ans:

In statistics, the Pearson correlation coefficient (**PCC**), also referred to as **Pearson's r**, or he Pearson product-moment correlation coefficient (**PPMCC**) is a measure of the strength of a linear association between two variables and is denoted by *r*. Basically, it attempts to draw a line of best fit through the data of two variables, and the *r*, indicates how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit).

The Pearson correlation coefficient, r, can take a range of values from +1 to -1 A value of 1 implies that a linear equation describes the relationship between X and Y perfectly, with all data points lying on a line for which Y increases as X increases. A value of −1 implies that all data points lie on a line for which Y decreases as X increases. A value of 0 implies that there is no linear correlation between the variables



The Formula for Calculating the Pearson's R is the ratio of Covariance of X and Y divided by the product of the Standard deviations on X and Y.

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

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

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

Ans:

When there are a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients making it difficult to interpret. Thus we need to scale features because of two reasons for

1. Ease of interpretation
2. Faster convergence for gradient descent methods

Difference:

S.NO.	Normalisation	Standardisation
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called `MinMaxScaler` for Normalization.	Scikit-Learn provides a transformer called `StandardScaler` for standardization.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

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

Ans:

For a Regression Model like - $Y = \beta0 + \beta1 \times X1 + \beta2 \times X2 + \beta3 \times X3 + \varepsilon$,

VIF is given by, $$VIF_1 = \frac{1}{1 - R^2}$$

Since R^2 which is the coefficient of determination from the linear regression model, lies between 0 and 1,

When $R^2 = 1$ the VIF tends to infinity. While determining VIF for a predictive variable X, $R^2 = 1$ implies that all the other predictor variables are completely (100 %) able to explain the variance in Xi and that this variable can be perfectly predicted by other variables in the model.

This signifies a perfect correlation.

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

Ans:

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data 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.

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

The advantages of the q-q plot are:

1. The sample sizes do not need to be equal.
2. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.
3. To check *If two data sets —*
    i. come from populations with a common distribution
    ii. have common location and scale
    iii. have similar distributional shapes
    iv. have similar tail behaviour

The q-q plot is formed by:
Vertical axis: Estimated quantiles from data set 1
Horizontal axis: Estimated quantiles from data set 2

Both axes are in units of their respective data sets. That is, the actual quantile level is not plotted. For a given point on the q-q plot, we know that the quantile level is the same for both points, but not what that quantile level actually is.