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 :- Inference from the categorical variables is as follows:-

- The demand of bike is less in the month of spring when compared with other seasons
- The demand bike increased in the year 2019 when compared with year 2018.
- Month Jun to Sep is the period when bike demand is high. The Month Jan is the lowest demand month.
- Bike demand is less in holidays in comparison to not being holiday.
- The demand of bike is almost similar throughout the weekdays.
- There is no significant change in bike demand with working day and non working day.
- The bike demand is high when weather is clear and Few clouds however demand is less in case of Lights now and light rainfall. We do not have any data for Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog , so we cannot derive any conclusion. May be the company is not operating on those days or there is no demand of bike.

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

Ans :- drop_first=True helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. known as multicollinearity and helps in avoiding dummy Variable Trap.

Ex:- If we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi furnished, then it is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished.

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

Ans :- temp and atemp has the highest correlation with cnt the target variable 0.64,0.65 respectively and A Negative correlation observed with cnt vs hum and cnt vs windspeed (-0.06 and -0.25)

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

Ans :- Linear Regression makes certain assumptions about the data and provides predictions based on that. Naturally, if we don't take care of those assumptions Linear Regression will penalise us with a bad model. We will check following assumption are met or not

1. *The Dependent variable and Independent variable must have a linear relationship.*
   A simple pairplot of the dataframe can help us see if the Independent variables exhibit linear relationship with the Dependent Variable.

2. Error terms or residuals are normally distributed
   Use Distribution plot on the residuals and see if it is normally distributed.
3. No perfect multicollinearity
   Check VIF (Variance Inflation Factor) if:-
   VIF=1, Very Less Multicollinearity
   VIF<5, Moderate Multicollinearity
   VIF>5 , Extreme Multicollinearity (This is what we have to avoid)
4. Error terms have constant variance (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.
   Residuals are nothing but the difference between actual and fitted values
5. Error terms are independent of each other :-Check this assumption by examining a scatterplot of "residuals versus fits"; the correlation should be approximately 0.

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

Ans :- temp with coefficient of 0.59 i.e., 59%

weathersit_Light_snow_rain_thunderstrom:- Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds with coefficient of -0.24 i.e., -24%

yr_2019 with coefficient of 0.22 i.e., 22%

so temp, weathersit_Light_snow_rain_thunderstrom and yr_2019 are top 3 features contributing significantly towards explaining the demand of the shared bikes

General Subjective Questions

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

Ans :- Linear regression is one of the very basic forms of machine learning where we train a model to predict the behavior of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

Mathematically, we can write a linear regression equation as:

$$y = a + bx$$

Where a and b given by the formulas:

$$b(slobe) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a(intercept) = \frac{n\sum y - b(\sum x)}{n}$$

Here, x and y are two variables on the regression line.

b = Slope of the line

a = y-intercept of the line

x = Independent variable from dataset

y = Dependent variable from dataset

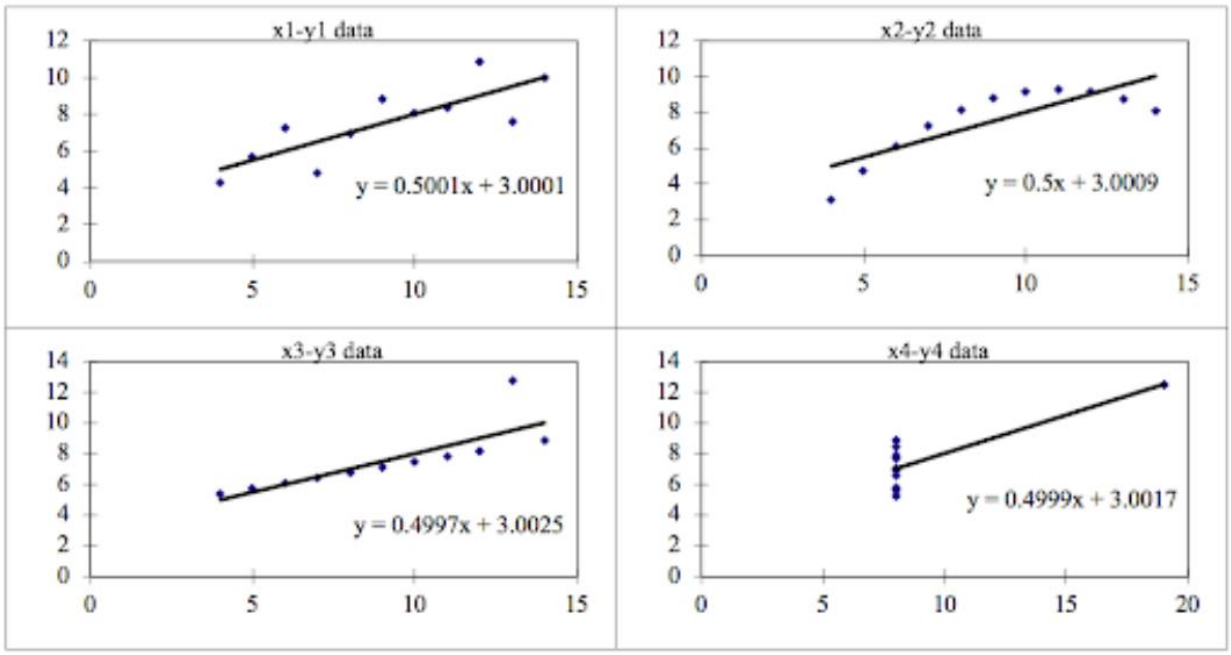
2. Explain the Anscombe's quartet in detail. (3 marks)

Ans :- Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

Ex:- The statistical information for four data sets are approximately similar. We can compute them as follows

Anscombe's Data								
Observation	x1	y1	x2	y2	x3	y3	x4	y4
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.1	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.1	4	5.39	19	12.5
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89
			Summary Statistics					
N	11	11	11	11	11	11	11	11
mean	9.00	7.50	9.00	7.500909	9.00	7.50	9.00	7.50
SD	3.16	1.94	3.16	1.94	3.16	1.94	3.16	1.94
r	0.82		0.82		0.82		0.82	

However, when these models are plotted on a scatter plot, each data set generates a different kind of plot that isn't interpretable by any regression algorithm, as you can see below:



ANSCOMBE'S QUARTET FOUR DATASETS

- **Data Set 1:** fits the linear regression model pretty well.
- **Data Set 2:** cannot fit the linear regression model because the data is non-linear.
- **Data Set 3:** shows the outliers involved in the data set, which cannot be handled by the linear regression model.
- **Data Set 4:** shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

As seen above, Anscombe's quartet helps us to understand the importance of data visualization and how easy it is to fool a regression algorithm. So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set in order to help build a well-fit model.

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

Ans :- Pearson's R is defined as measure of strength of relationship between the two variables

and their association with each other, In other words Pearson's correlation coefficient calculates the effect of change in one variable when the other variable changes.

It seeks to draw a line through the data of two variables to show their relationship. The relationship of the variables is measured with the help Pearson correlation coefficient calculator. This linear relationship can be positive or negative.

For example:

- **Positive linear relationship:** In most cases, universally, the income of a person increases as his/her age increases.
- **Negative linear relationship:** If the vehicle increases its speed, the time taken to travel decreases, and vice versa.

From the example above, it is evident that the Pearson correlation coefficient, r, tries to find out two things – the strength and the direction of the relationship from the given sample sizes.

Pearson correlation coefficient formula and calculation

The correlation coefficient formula finds out the relation between the variables. It returns the values between -1 and 1. Use the below Pearson coefficient correlation calculator to measure the strength of two variables.

Pearson correlation coefficient formula:

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

Where:
N = the number of pairs of scores
Σxy = the sum of the products of paired scores
Σx = the sum of x scores
Σy = the sum of y scores
Σx2 = the sum of squared x scores
Σy2 = the sum of squared y scores

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

Ans :- It is a step in data preprocessing which is applied to the independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in the algorithm.

Scaling is performed because most of the times collected dataset contains features highly varying in magnitude,units and range, if scaling is not done then algorithm takes only magnitude in account and not units hence incorrect modelling, To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

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.

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? (3 marks)

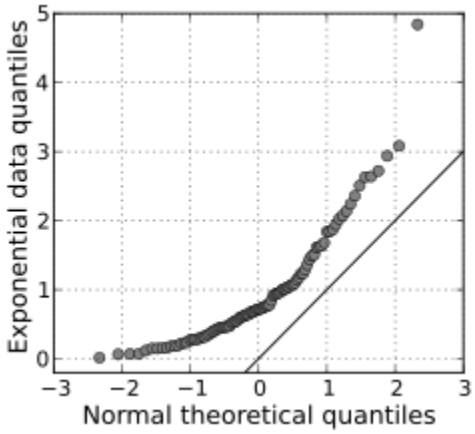
Ans :- If there is a perfect correlation, then VIF = infinity. This shows perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Ans:- Q-Q plots are plot of two quantiles against each other, A quantile is a fraction where certain values fall below that quantile and certain values fall above that quantile , for example the median is a quantile where 50% lie above it. The purpose of QQ plot is to find out if the two datasets come from same distribution. A 45 degree angle is plotted on the QQ plot; if the two datasets come from a common distribution, the points will fall on that reference line.

A Q Q plot showing the 45 degree reference line:



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.