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)

Answer:
1 .We can see **fall season** seems more booking of bike among other season.
2. We can see the in each **season** demand of bike is drastically increasing from 2018 to 2019.
3 .The demand of bike is high in **may**, **june**, **july**, **august**, **sep**, **oct** and less demand in last two month.
4. The demand of bike booking is high in good weather situation.
5. Also demand of bike booking is drastically increase in 2019 when **weather situation is good**.
6. Booking Demand is high in **Thursday**, **Friday** and **Saturday**
7. The demand of bike booking is high in **Holiday**.
8. The demand of bike booking is low in **holiday** of **2019 Year**.

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

Answer: This has something to do with Mutli-colinearity in case if Multiple Linear Regression. Beacause, keeping k dummies for k levels of a categorical variable is good idea, but there is a redundancy of one level, which is here in separate column. This is not needed since one of the combination will be uniquely representing this redundant column. Hence, it's better to drop one of the column and just have k-1 dummies (columns) to represent k levels.

This Overall approach reduces Multi-colinearity in the dataset, which is one of the prime Assumption of Multiple Linear Regression.

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

Answer: From pair plot it can be observed that **temp** and **atemp** is highly correlated.

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

Answer: We can validate the assumption of Linear Regression after the Model on training set by following way:
1. Residual Analysis of the train data and validation
2. Normality of error terms
3. Residual Errors Have a Mean Value of Zero
4. Residual Errors Have Constant Variance
5. Residual Errors Are Independent from Each Other and Predictors (x)

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

Answer:
1. Temperature (temp) - A coefficient value of '0.630' indicated that a unit increase in temp variable increases the bike hire numbers by 0.630 units.

2 .Humidity - A coefficient value of '-0.345' indicated that, a unit increase in humidity variable decreases the bike hire numbers by 0.345 units.

3 .Year - A coefficient value of '0.228' indicated that a unit increase in Year variable increases the bike hire numbers by 0.228 units.

General Subjective Questions

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

Answer: Linear Regression is a type of supervised learning model which is used for forecasting. In the Supervised Learning model, we use training data to build the model and then use test data to test its accuracy.
Linear Regression shows the relationships between a set of independent variables to that of the dependent variable.

Linear Regression is the plotting of a straight line of a form y=mx+c such that it predicts the data points. In other words, if our model is well trained using Linear Regression then, in that case, the predicted point will lie on the regression line.

Let's suppose that we have 2 axis x & y where the x-axis has independent variables and the y-axis has dependent variables and here our aim is to draw the Regression Line. If we have a data point on the x-axis which increases and is independent in nature, similarly we have a data point on the y-axis which is also increasing and is dependent in nature thus we will get a positive regression line.

Suppose that the data point of the y-axis is decreasing then we will get the negative regression line.
As we know the equation of a line is y=mx+c. Here

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

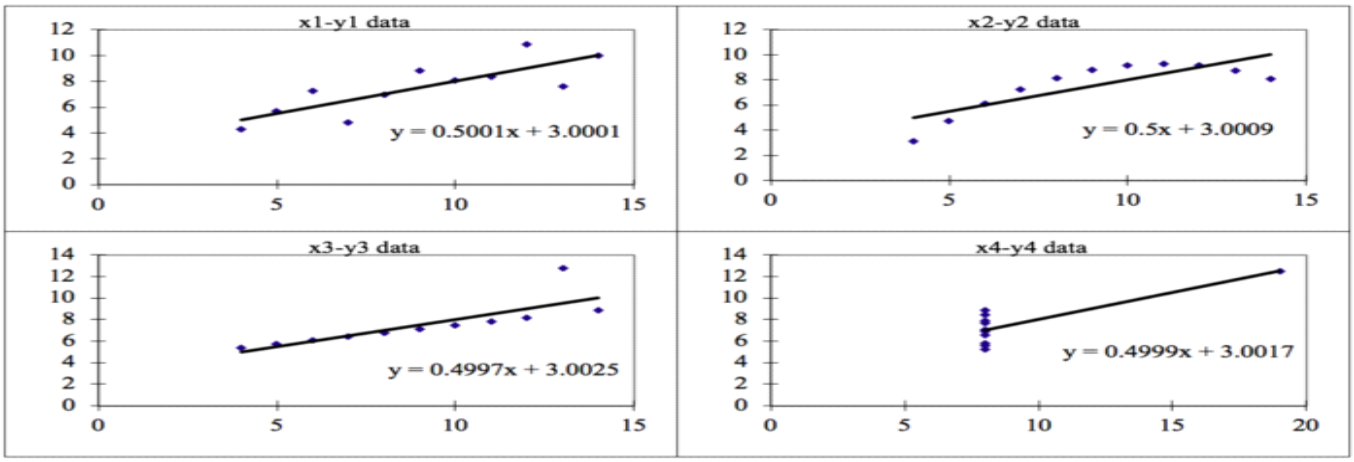
2. Explain the Anscombe's quartet in detail. (3 marks)

Answer: 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.

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.

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	

When these models are plotted on a scatter plot, all datasets generates a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities and can be seen as follows:



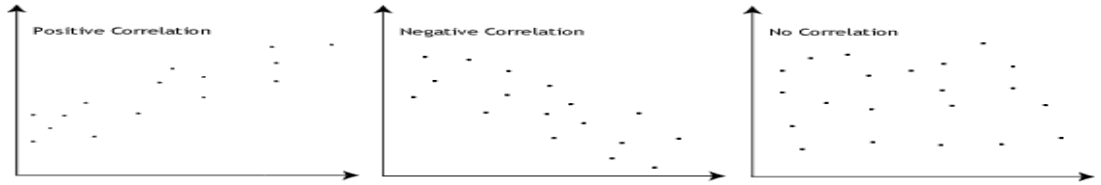
- **Dataset 1:** this **fits** the linear regression model pretty well.

- **Dataset 2:** this **could not fit** linear regression model on the data quite well as the data is non-linear.
- **Dataset 3:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model
- **Dataset 4:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model

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

Answer: In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.
The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong associatio
-



$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Here,

- =correlation coefficient
- =values of the x-variable in a sample
- =mean of the values of the x-variable
- =values of the y-variable in a sample
- =mean of the values of the y-variable

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

Answer: (Scaling) It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes 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.

Difference between normalized scaling and standardized scaling

Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

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)

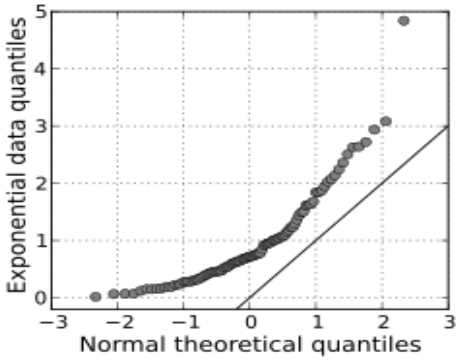
Answer: If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing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

Answer: Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets 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.