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?

A) The demand of bike is less in the month of spring when compared with other seasons.

The demand bike incread in the year 2019 when compared with year 2018

Month Jun to Sep is the period when bike demand is high

The month ~~for~~ Jan is the lowest demand month.

Bike demand is less in holidays in comparison to not being holiday.

~~The bike~~ Bike demand ~~is~~ of bike is almost similar throughout the weekdays.

There is no singnificat change in the bike demand with working day and non working day.

The bike demand is high is high when weather is clear and Few clouds however demand is less in case of Light snow and light rainfall.

We dont have dat for Heavy Rain + Ice pallets + Thunderstorm + mist snow + Fog, so we can not derive ay conclusion. May be company is not operating on those days or there is no demand of bikes.

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

(A) drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation.

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

(A) By looking at the pair plot temp variable has highest (0.63) correlation with target variable 'cnt'.

(4) How did you validate the assumptions of linear Regression after building the model on the Traing set?

(1) Linearity
(2) mean of Residue
(3) check for Homoscedasticity
(4) checke for Normality of error terms/residuals
(5) No autocorrelation of residuals
(6) No perfect multicollinearity
(7) other models for comparison.

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

(A) The top 3 features contributing significantly towards the demands of the shared bikes are:

Weather - light - Snow (negative correlation) ⑨
yr - 2019 (positive correlation).
temp (positive correlation).

General subjective Questions.

① Explain the linear regression algorithm in detail.

Ⓐ Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour 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~~ $y = a + bx$

where "a" and "b" given by the formula:

$$b\,(\text{slope}) = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$$

$$a\,(\text{intercept}) = \frac{n\Sigma y - b(\Sigma x)}{n}$$

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 data set

y = Dependent variable from dataset.

② Explain the Anscombe's quartet - in detail.

Ⓐ Anscombe's quartet comprises four datasets that have nearly identical simple stastical properties, yet appear very different when graphed. Each dataset consists of eleven (x, y) points. They were constructed in 1973 by statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

Single understanding

once franc Then 'Franc' Ansombe who was a statician of great reput found 4 sets of 11 data points in his dreams and requrked the Counsil as his last wish to plot those points. Those 4 sets of 11 - data-points are given below.

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.4	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

After that, the council analyzed them using only descriptive statistics and found the mean, SD, and correlation between x and y.

(3) What is Pearson's R?

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 co-relation between two sets of data. It is covariance of two sets of two variables, divide by the product of their standard deviations; thus it is essentially a normalized measurement of the covariance, such that the result always has 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.

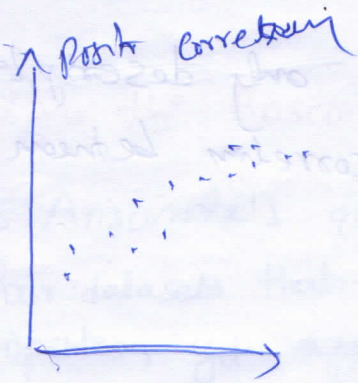
r = -1 means the data is perfectly linear with a negative slope.

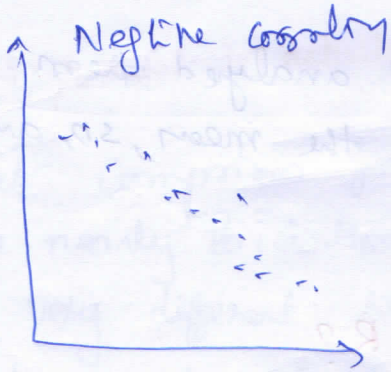
r = 0 means there is no linear association.

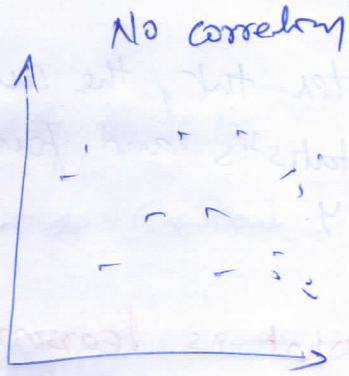
r > 0 < 5 means there is weak association

r > 5 < 8 means there is a moderate association

r > 8 means there is a strong association.







Pearson Formula

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

r = Correlation Coefficient.

x_i = Value of x variable in Sample

$\bar{x}$ = mean of the value of the x-variable.

y_i = value of the y-variable in a sample

$\bar{y}$ = mean of the value of the y-variable.

4) What is the difference between normalized scaling and standardized scaling?

A) Normalized Normalizing typically ~~mean~~ it means ~~that~~ rescale the values into the range of [0,1]

Standardization typically means rescale the data to have a ~~mean~~ mean of "0" and a standard deviation of 1 (unit variance)

S.no	Normalization	Standardisation
(1)	Minimum and maximum value of feature are used for scaling	Mean and Standard deviation is used for scaling.
(2)	It is used when feature are in different scale.	It is used when we want to ensure zero ~~mean~~ mean and unit standard deviation.
(3)	Scale between [0,1] or [-1,1]	It is not bounded a certain Range.
(4)	It is really effected by outliers.	It is much less effected by outliers.
(5)	scikit-learn provides a transform called MinMaxScaler for normalizing	scikit-learn provides a transform called StandardScaler for Standardization.
(6)	This transformation squishes the n-dimensional ~~data~~ data into an n-dimensional unit hypercube.	It translates the ~~data~~ data to the mean vector of original ~~data~~ data to the origin and squishes or expands.

⑦ ~~It is useful useful~~

⑦ It is useful when we don't know about distribution

~~It is useful~~ ~~It is us~~ It is useful when the feature distribution is Normal or Gaussian.

⑧ It is often called as Scaling Normalization

It is often called a Z-score normalization.

⑨ you Might have observed that sometimes the value of VIF is infinite. Why does this happen?

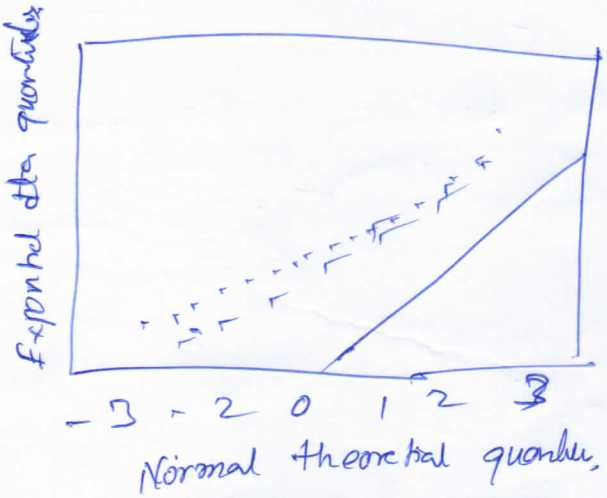
A) If there is a perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get $R_2 = 1$, which lead to $1/(1-R_2)$ Infinity. To solve this problem we need to drop one of the variables from the data set which is 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 shows an infinite VIF as well).

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

A) 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 comes from two data sets come from a common distribution, the points will fall on that reference line.

A Q-Q plot showing the 45° 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 points 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 distributions.

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