Assignment-based Subjective Questions

Question 1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer : Followings are the categorical variables and their effect on the independent variable : -

- Season : As compare to any other season, Fall has high demand while Spring season low demand of bike.
- Year : The year 2019 was highly demanding as compared to the year 2018.
- Month : Demand is high in month of August, September, October and low in January.
- Holiday : In holidays there is less bike demand as compare to not holiday.
- Weekday : Demand almost similar on all the weekdays.
- Weathersit :
    - High demand with - "Clear, Few clouds".
    - Low demand with "Light Snow, Light Rain"

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

Answer : As per the concept, while creating dummy for categorical column with n unique values, we only need n-1 column to represent with **drop_first=True** and removes the first column. We can infer the same information without extra column.

e.g., let's we have a column furnishing status with 3 categories as

- furnished,
- semi-furnished
- unfurnished.

Without dropping the dummy variables looks like

furnished	1	0	0
semi-furnished	0	1	0
unfurnished	0	0	1

Now, we don't need three columns. we can drop the "furnished" column, as the type of furnishing can be identified with just the last two columns where —

- 00 will correspond to "furnished"
- 01 will correspond to "unfurnished"
- 10 will correspond to "semi-furnished"

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

Answer : registered (Resisted Users Count) both has the highest correlation with cnt (Target variable).

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

Answer : With the help of following point, we can validate the assumption:

- Error Terms: The Error term should be normally distributed.
- Linear relationship between x and y.
- Multi-collinearity: There should be no or minimal collinearity in the independent variables.
- Homoscedasticity: There should be no visible pattern in the distribution of the residual/error term.

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

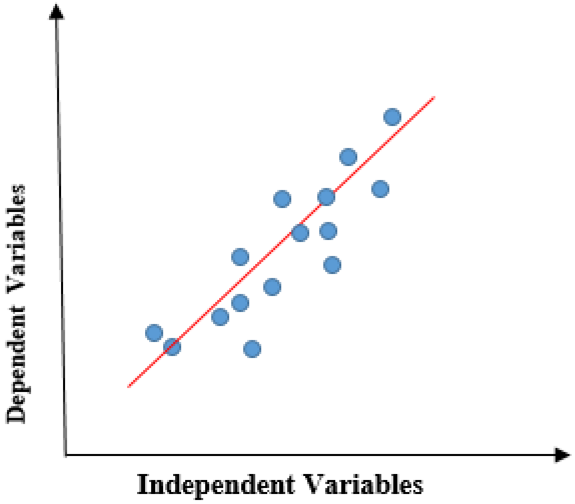
Answer : Following are the top 3 features contributing significantly toward explaining the demand of the shared bikes: -

- Year (yr) - A unit increased into the year increases the bike demand by 0.246 unit.
- Weekday (Saturday) - A unit increased in this the bike demand by 0.067 unit.
- WindSpeed – A unit increased in the windspeed the bike demand by 0.057 unit.

General Subjective Questions

Question 1: Explain the linear regression algorithm in detail.

Answer : Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression. If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression. The linear regression model gives a sloped straight line describing the relationship within the variables..



In the above figure,
X-axis = Independent variable
Y-axis = Output / dependent variable
Line of regression = Best fit line for a model

Here, a line is plotted for the given data points that suitably fit all the issues. Hence, it is called the 'best fit line.' The goal of the linear regression algorithm is to find this best fit line seen in the above figure.

Mathematically the relationship can be represented with the help of following equation −

y = a + bx

Where a and b given by the formulas:

$$b\,(slope) = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\sum x^2 - \left(\sum x\right)^2}$$

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

Here, y is the dependent variable we are trying to predict.

x -> is the independent variable we are using to make predictions.
b -> is the slop of the regression line which represents the effect x has on y
a -> is a constant, known as the y-intercept.

The following are some assumptions about dataset that is made by Linear Regression model –

Multi-collinearity − Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.
Auto-correlation − Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.
Relationship between variables − Linear regression model assumes that the relationship between response and feature variables must be linear.

Question 2: Explain the Anscombe's quartet in detail.

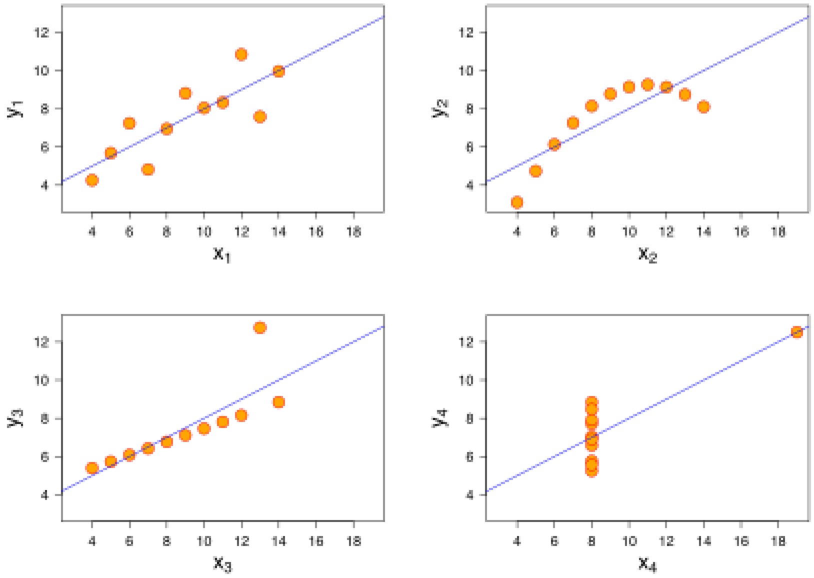
Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing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

Simple understanding:

Once Francis John "Frank" Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council 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	9.14	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 finding the mean, standard deviation, and the correlation between x and y and plotting them we will see the following output –



Conclusion of the above output:

- In the first one (top left) if we look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the second one (top right) if we look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the third one (bottom left) you we say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- The fourth one (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

Question 3: What is Pearson's R?

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.

It is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

- 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 association

Calculating the Calculating pearson correlation in python –

```
from scipy.stats import pearsonr
list1 = [1,2,3,4], list2 = [1,6,3,7]
corr, _ = pearsonr(list1, list2)
```

Pearson r Formula -

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

Here,

- r = correlation coefficient
- x_{i} = values of the x-variable in a sample
- bar{x} = mean of the values of the x-variable
- y_{i} = values of the y-variable in a sample
- bar{y} = mean of the values of the y-variable

Question 4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer: 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.

Why?

Algorithms that compute the distance between the features are biased towards numerically larger or different unit of values if the data is not scaled. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling and here scaling comes to the picture to solve the issue.

Tree-based algorithms are fairly insensitive to the scale of the features. Also, feature scaling helps machine learning, and deep learning algorithms train and converge faster.

Scaling only affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization or Min-Max Scaling: It brings all of the data in the range of 0 and 1. The new point is calculated as:

X_new = (X - X_min)/(X_max - X_min)

Standardization or Z-Score Normalization: Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one.

X_new = (X - mean)/Std

Normalisation	**Standardisation**
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
It is really affected by outliers.	It is much less affected by outliers.
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.
It is useful when we don’t know about the distribution.	It is useful when the feature distribution is Normal or Gaussian.

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

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.

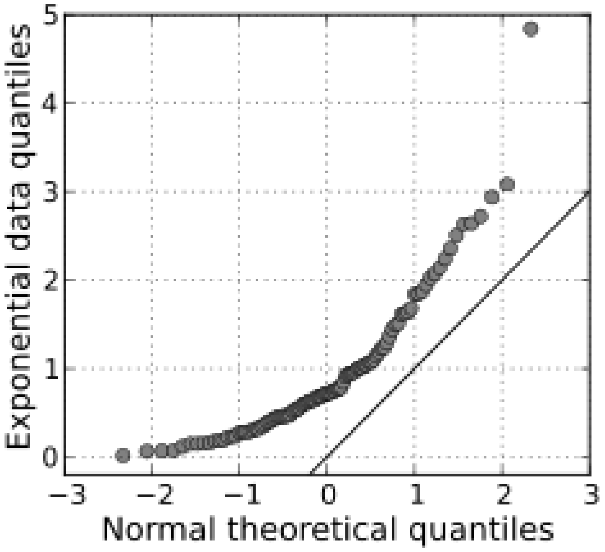
$$\text{VIF} = \frac{1}{1 - R^2}$$

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).

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

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.