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)

Yes, because that dependent variable sometime can have a complete control over the other variable which is depend. For instance if the season is summer, mostly the temperature is will be high, likewise, month is dependent to season (like fall, winter, spring, summer).

So VIF will result a correlation of variable among all variables in the data frame. So VIF can give us the most dependent variable to us.

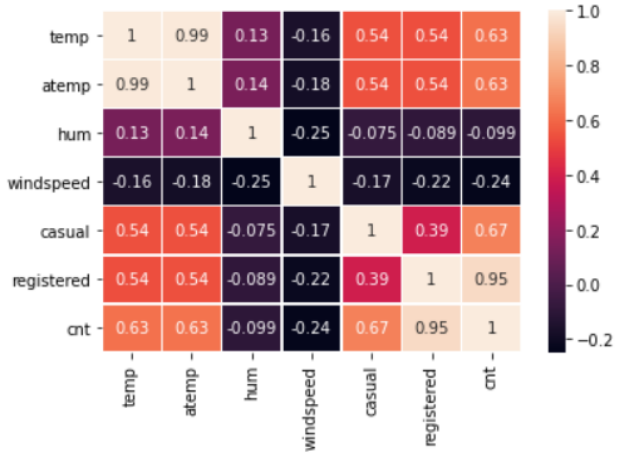
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Because then only we can derive n-1 columns as per the data efficiency standard. If not we have to manually drop any other column to achieve the same.

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

atemp(Atmosphere Temperature) column have the highest correlation on the cnt variable.

This result derived after dropping registered variable due to multicollinearity.



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

1. Fitted regression line is linear.
2. Where the R square value is >=80% the model is better.
3. P value should not be greater than 0.05
4. Adjusted R square should be +/- 5% from the R square.

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

1. Yr (Year) Positive

2. weathersit (weather situation) Negatively contributing

3. Sep(September month) Positive

	coef
const	3411.2033
yr	2138.3540
workingday	463.0695
weathersit	-1957.7191
spring	-1045.8557
winter	745.8585
Aug	1058.8777
Dec	-977.8591
Feb	-689.4066
Jan	-1196.2775
July	1018.7556
June	1090.8084
May	847.0134
Nov	-1135.2818
Sep	1264.7184
Sat	417.6773

General Subjective Questions

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

1. Reading and understanding the data
2. Visualizing the data (Exploratory Data Analysis)
3. Data Preparation
4. Splitting the data into training and test sets
5. Building a linear model
6. Residual analysis of the train data:
7. Making predictions using the final model and evaluation

Divide the test sets into X_test and y_test and calculate r2_score of test set. The train and test set should have similar r2_score. A difference of +/- 5% between r2_score of train and test score is acceptable as per the standards.

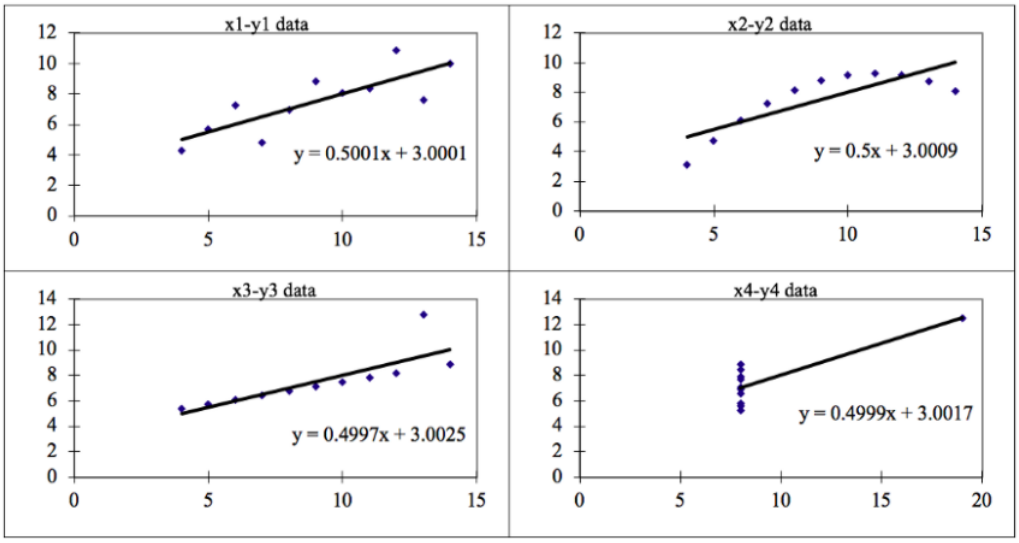
2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's Quartet can be defined a a group of four data sets which are nearly identical in simple descriptive statistics, 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.

It illustrate the importance of plotting the graphs before analyzing and model building an the effect of other observations on statistical properties. There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance and mean of all x, y points in all four datasets.

The statistical information for all these four datasets approximately similar. 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:



1. Dataset 1: This fits the linear regression model pretty well.
2. Dataset 2: This could not fit linear regression model on the data quite well as the data is non-linear.
3. Dataset 3: Shows the outliers involved in the dataset which cannot be handled by linear regression model.
4. Dataset 4: Shows the outliers involved in the dataset which cannot be handled by linear regression mode.

Conclusion:

All the important features in the dataset must be visualized before implementing any machine learning algorithm on them which will help to make a good fit model.

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

In statistics, the Pearson correlation coefficient, also known as Pearson's r, the Pearson product-moment correlation coefficient **(PPMCC)** or the bivariate correlation, or colloquially simply as the correlation coefficient — is a measure of linear correlation between two sets of data.

Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

we can calculate a **linear relationship** between the two given variables. For example, a child's height increases with his increasing age (different factors affect this biological change). So, we can calculate the relationship between these two variables by obtaining the value of Pearson's Correlation Coefficient r. There are certain requirements for Pearson's Correlation Coefficient:

- Scale of measurement should be interval or ratio
- Variables should be approximately normally distributed
- The association should be linear
- There should be no outliers in the data

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

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

The positive value of Pearson's correlation coefficient implies that if we change either of these variables, there will be a **positive effect** on the other. For example, if we increase the age there will be an increase in the income.

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

Scaling a method to minimize the variable values without changing the actual values of it.

It basically means modify all the values in a variable so that the actual value will effect the same.

Ex: values are like, 25, 39, 95, 22, 39 are the age this we can divide by 10 to get lesser value (0-10) it will result. So that plotting it will give crisp insight.

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Here's the formula for normalization:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

Here's the formula for standardization:

$$X' = \frac{X - \mu}{\sigma}$$

μ is the mean of the feature values and σ

- Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.
- Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So even if you have outliers in your data, they will not be affected by standardization.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

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).

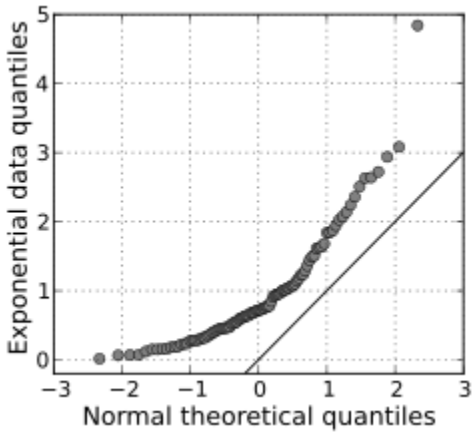
If all the independent variables are orthogonal to each other, then VIF= 1.0. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. This would mean that that standard error of this coefficient is inflated by a factor of 2 (square root of variance is the standard deviation).

If VIF is large and multicollinearity affects your analysis results, then you need to take some corrective actions before you can use multiple regression.

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

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.