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)

There are two main categorial variables in dataset –

1. Season – A season can be one of the 4 seasons from spring, summer, fall, winter. The season can have effect on the bike sharing as people will not prefer to use bike in bad weather. We expect to find overall higher or lower trend for a season.
2. Weathersit – This variable shows the status of the weather conditions on given day. As season it would have effect on the number of users using the service as people would not prefer bike share in bad weather. As season gives us seasonal reading on the weather, which would stay consistent for weeks/months, this variable can provide us with the variation seen in the weather on daily basis.

Above variables are not ordered and hence would need to be categorized and then we would need to create the dummy variables for the same.

Apart from these below additional variables can be treated as categorial variables.

1. Yr – 2018 or 2019
2. Mnth – Month of year for the reading (ordered 1-12)
3. Weekday – Day of week (ordered 0-6)
4. Holiday – if day is holiday or not (unordered binary)
5. Workingday – weather day is working day or not (unordered binary)

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

When we have three categorial values for the column we can represent those values using two columns because 00 can be used as third value.

For example – For season we get dummies like this –

	fall	spring	summer	winter
0	0	1	0	0
1	0	1	0	0
2	0	1	0	0
3	0	1	0	0
4	0	1	0	0

Here we don't need four columns. We can drop the `fall` column, as the season can be identified with just the last three columns where —

- `000` will correspond to `fall`
- `100` will correspond to `spring`
- `010` will correspond to `summer`
- `001` will correspond to `winter`

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

The column ‘registered’ has the highest correlation with the target variable.

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

We need to plot the errors and find if their distribution is normal distribution. If the errors are distributed normally, then the Problem is good fit for the Linear Regression problem.

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

Based on the final model built, below are the three factors significantly contributing towards explaining the demand of the shared bikes.

1. registered
2. casual
3. temp

General Subjective Questions

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

Linear regression is used to find how the value of the dependent variable is changing according to the value of the independent variable and make predictions for continuous/real or numeric variables. Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (x) variables.

The linear regression model provides a sloped straight line representing the relationship between the variables.

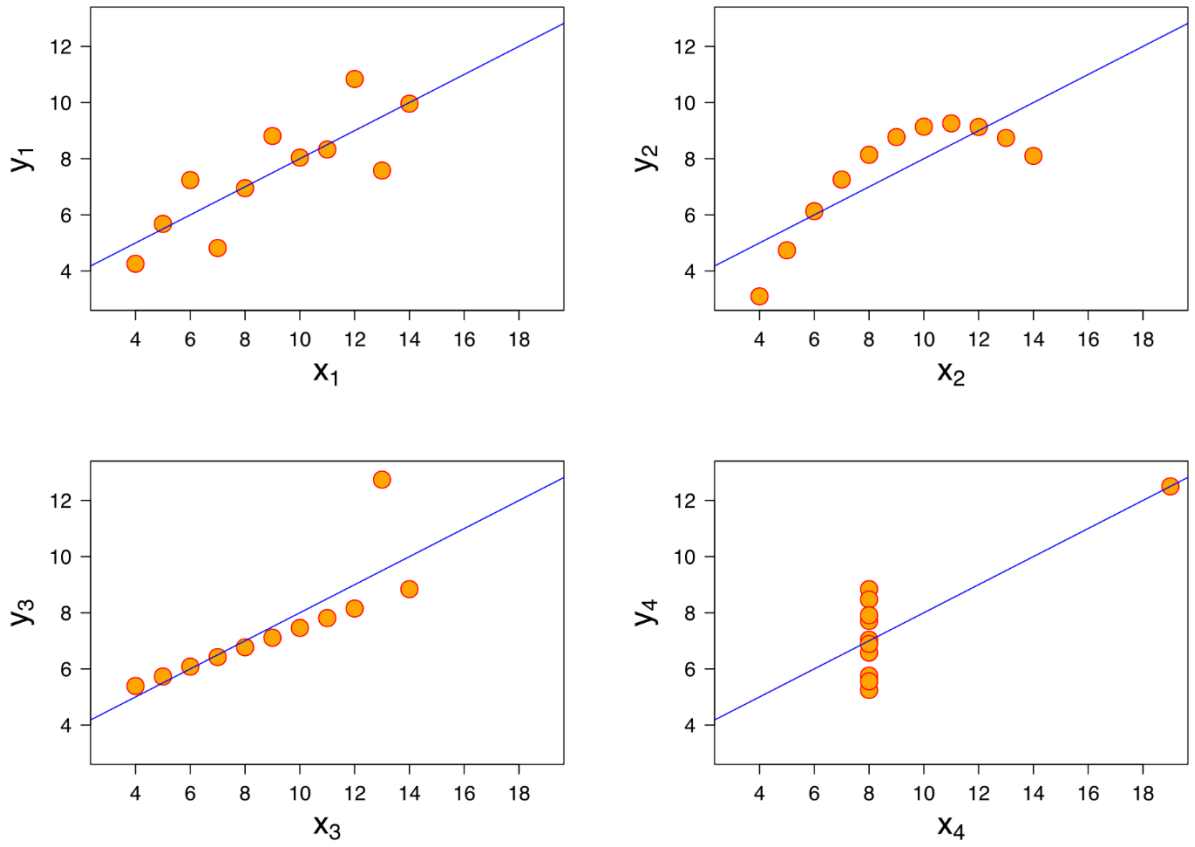
There are two types of linear regression algorithms –

- Simple Linear Regression:
  A single independent variable is used to predict the value of a numerical dependent variable.
- Multiple Linear regression:
  More than one independent variable is used to predict the value of a numerical dependent variable.

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

Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x,y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

For example, if we consider below graphs,



- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
- The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets

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

The Pearson's R is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and 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. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation. As a simple example, one would expect the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than 0, but less than 1 (as 1 would represent an unrealistically perfect correlation)

The absolute values of both the sample and population Pearson correlation coefficients are on or between -1 and 1. Correlations equal to +1 or −1 correspond to data points lying exactly on a line (in the case of the sample correlation), or to a bivariate distribution entirely supported on a line (in the case of the population correlation). The Pearson correlation coefficient is symmetric: corr(X,Y) = corr(Y,X).

A key mathematical property of the Pearson correlation coefficient is that it is invariant under separate changes in location and scale in the two variables. That is, we may transform X to a + bX and transform Y to c + dY, where a, b, c, and d are constants with b, d > 0, without changing the correlation coefficient. (This holds for both the population and sample Pearson correlation coefficients.) Note that more general linear transformations do change the correlation: see § Decorrelation of n random variables for an application of this.

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

What is scaling?

Feature scaling is a method used to normalize the range of independent variables or features of data.

Why is scaling performed?

Since the range of values of raw data varies widely objective functions will not work properly without normalization. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.

Another reason why feature scaling is applied is that gradient descent converges much faster with feature scaling than without it.

It's also important to apply feature scaling if regularization is used as part of the loss function (so that coefficients are penalized appropriately).

What is the difference between normalized scaling and standardized scaling?

In Normalized scaling consists in rescaling the range of features to scale the range in [0, 1] or [−1, 1]. In standardized scaling method of calculation is to determine the distribution mean and standard deviation for each feature. Next we subtract the mean from each feature. Then we divide the values (mean is already subtracted) of each feature by its standard deviation.

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

If there is perfect correlation between the dependent variables, then the value of VIF is infinite.

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.

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.