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)

   The categorical variables that are part of the dataset are: “season”, “workingday”, “weathersit”, “weekday”, “yr”, “holiday”, and “mnth”.

- Season:
  - Analysing the data, we can come to the inference that the most favourable seasons for biking are summer and fall. Hence, the recommendation to the company would be to plan higher numbers for these seasons.
  - On the other hand, Sprint has a significantly low demand for bikes. Hence, this could be the season where maintenance work could be understaken by the company.
- Workingday:
  - Analysing the data, it is clear that the registered users prefer renting the bikes on working days, whereas the casual users prefer non-working days. But when we consider the average, the overall impact is linear.
  - Registered and casual users’ identity and relevant strategy for working and not working days shall help to increase the numbers.
- Weathersit:
  - From the data, it is evident the most preferred weather for renting bikes is clean/few cloud days.
  - Registered users have a comparatively higher number even on light rainy days and this could be inferred as they are consuming the service for their daily routine.
  - Most favourable weather condition is the clean/few clouds days.
- Weekday:
  - Weekday has no significant impact on cnt. The demans is consistent across weekdays.
- yr:
  - Based on the data we can see that the demand of bikes is increasing annually.
- Holiday:
  - Consumption of bikes on holidays is more for causal users than registered users.
- Mnth:
  - The demand is high for June, July, August, September and October months.

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

   drop_first=True helps in reducing the extra column created during the dummy variable creation and hence avoid redundancy of any kind. This helps in avoiding the multi-collinearity among the variables present in the model resulting from the dummy varaibles being included. The reference category can be easily deduced where 0 is present in a single row for all the other dummy variables of a particular category.

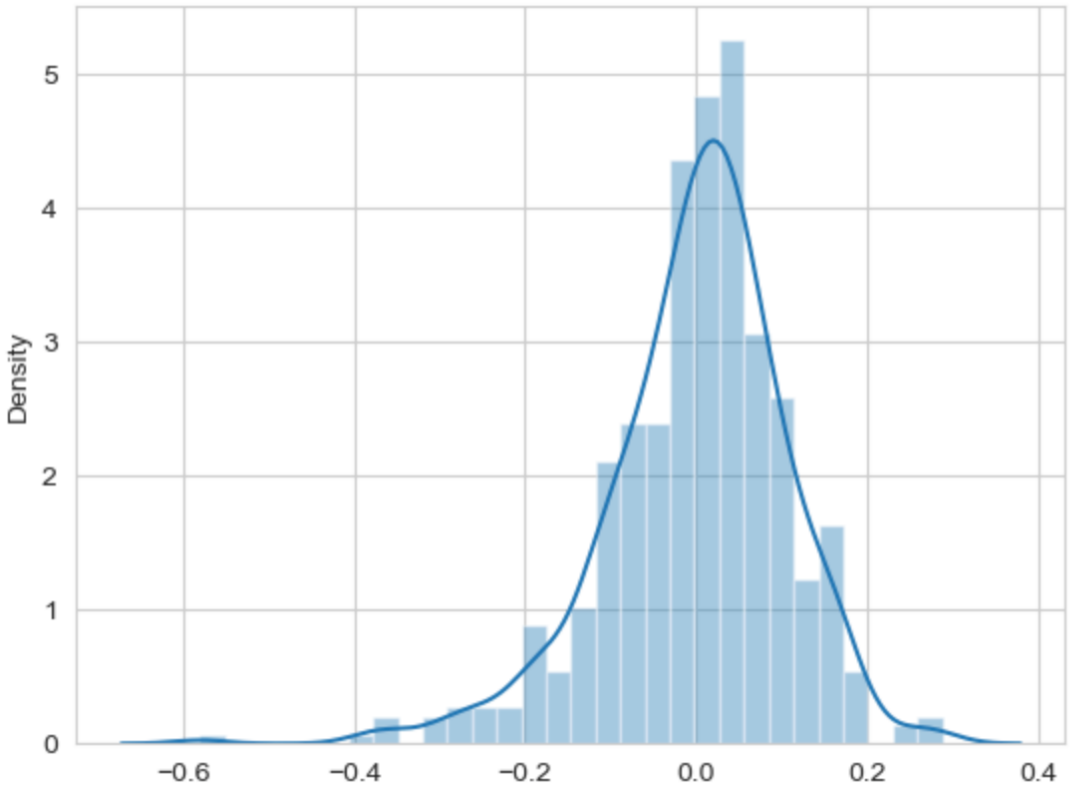
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

   The temp and atemp variable has the highest correlation with the target variable (0.64). Additionally, these variables also have an inter corelation of 0.99. This is also the reason why we decided to remove atemp and keep only temp, so that the assumptions for Linear Regression can hold valid.

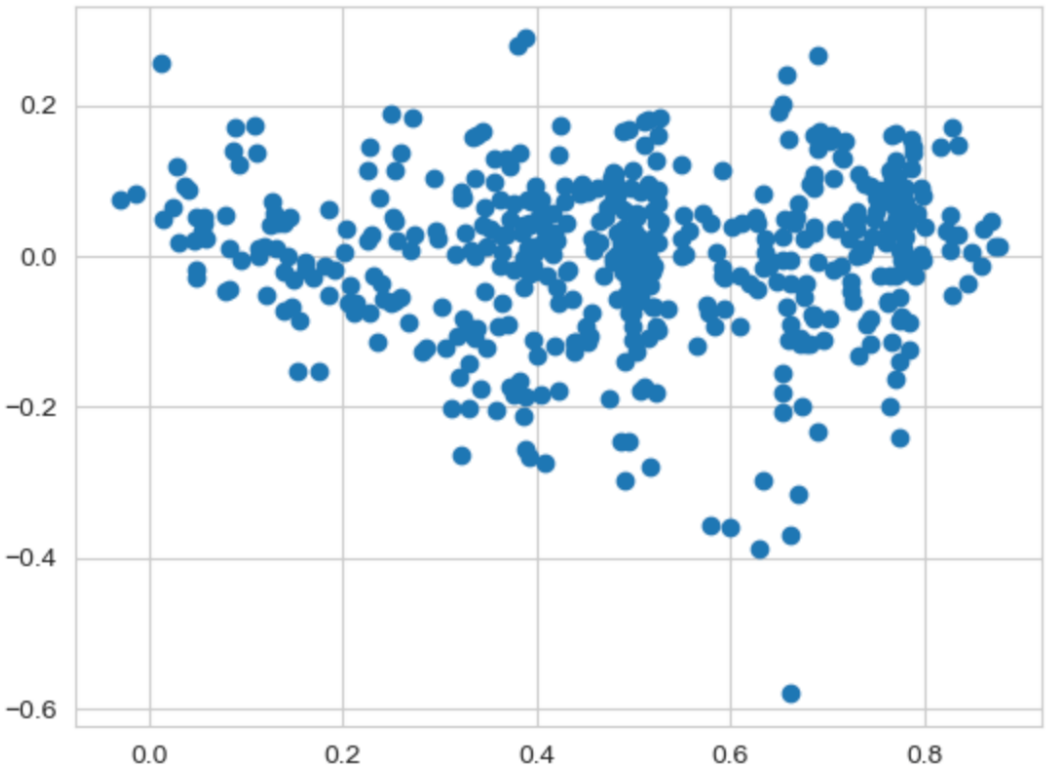
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

   Validated the assumptions of linear regression by checking the VIF, error distribution of residuals and linear relationship between the dependent variable and a feature variable.

   Error terms are normally distributed:



   Error terms are independent of each other:



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

   According to my model, the top 3 features contributing significantly towards the demand of the shared bikes are the year, workingday and the windspeed variables.

General Subjective Questions

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

   Linear Regression is a supervised learning algorithm. The primary purpose of using Linear Regression is to predict a dependent variable based on the given data of one or more independent variables. When only independent variable is present, we refer to it as Simple Linear Regression and when more are there it is referred to as Multiple Linear Regression. A positive linear relationship is when the dependent variable on the Y-axis increases along with the independent variable in the X-axis. However, if dependent variables value decreases with increase in independent variable value increase in X-axis, it is a negative linear relationship.

   - Simple Linear Regression – Single independent variable is used.
     - $Y = \beta_0 + \beta_1 X$ is the line equation used for SLR.
   - Multiple Linear Regression – Multiple independent variables are used.
     - $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \in$ is the line equation for MLR.
   - $\beta_0 = value\ of\ the\ Y\ when\ X = 0\ (Y\ intercept)$
   - $\beta_1, \beta_2, \ldots, \beta_p = Slope\ or\ the\ gradient.$

   Linear Regression also has certain assumptions:

1. Linear relationship: There exists a linear relationship between the independent variable, x, and the dependent variable, y.
2. Independence: The residuals are independent. In particular, there is no correlation between consecutive residuals in time series data.
3. Homoscedasticity: The residuals have constant variance at every level of x.
4. Normality: The residuals of the model are normally distributed.

If one or more of these assumptions are violated, then the results of our linear regression may be unreliable or even misleading.

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

Statistics like variance and standard deviation are usually considered good enough parameters to understand the variation of some data without actually looking at every data point. The statistics are great to for describing the general trends and aspects of the data.

Francis Anscombe realized in 1973 that only statistical measures are not good enough to depict the data sets. He created several data sets all with several identical statistical properties to illustrate the fact.

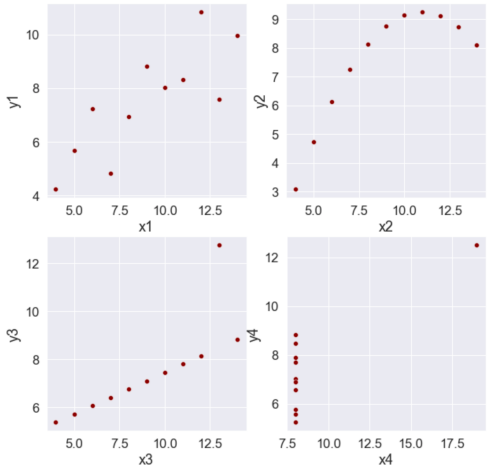
- Illustrations
  - One of the data set is as follows:

	x1	x2	x3	x4	y1	y2	y3	y4
0	10	10	10	8	8.040000	9.140000	7.460000	6.580000
1	8	8	8	8	6.950000	8.140000	6.770000	5.760000
2	13	13	13	8	7.580000	8.740000	12.740000	7.710000
3	9	9	9	8	8.810000	8.770000	7.110000	8.840000
4	11	11	11	8	8.330000	9.260000	7.810000	8.470000
5	14	14	14	8	9.960000	8.100000	8.840000	7.040000
6	6	6	6	8	7.240000	6.130000	6.080000	5.250000
7	4	4	4	19	4.260000	3.100000	5.390000	12.500000
8	12	12	12	8	10.840000	9.130000	8.150000	5.560000
9	7	7	7	8	4.820000	7.260000	6.420000	7.910000
10	5	5	5	8	5.680000	4.740000	5.730000	6.890000

  - If the descriptive statistics are checked for above data set then all looks same:

	x1	x2	x3	x4	y1	y2	y3	y4
count	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000
mean	9.000000	9.000000	9.000000	9.000000	7.500909	7.500909	7.500000	7.500909
std	3.316625	3.316625	3.316625	3.316625	2.031568	2.031657	2.030424	2.030579
min	4.000000	4.000000	4.000000	8.000000	4.260000	3.100000	5.390000	5.250000
25%	6.500000	6.500000	6.500000	8.000000	6.315000	6.695000	6.250000	6.170000
50%	9.000000	9.000000	9.000000	8.000000	7.580000	8.140000	7.110000	7.040000
75%	11.500000	11.500000	11.500000	8.000000	8.570000	8.950000	7.980000	8.190000
max	14.000000	14.000000	14.000000	19.000000	10.840000	9.260000	12.740000	12.500000

  - However, when plotted these points, the relation looks completely different as depicted below.



- Anscombe's Quartet signifies that multiple data sets with many similar statistical properties could still be different from one another when plotted.
- The dangers of outliers in data sets are warned by the quartet. Check the bottom 2 graphs. If those outliers would have not been there the descriptive stats would have been completely different in that case.
- Important points
    - Plotting the data is very important and a good practice before analysing the data.
    - Outliers should be removed while analysing the data.
    - Descriptive statistics do not fully depict the data set in its entirety.

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

Pearson's Correlation Coefficient is used to establish a linear relationship between two quantities. It gives an indication of the measure of strength between two variables and the value of the coefficient can be between -1 and +1.

- -1 coefficient indicates strong inversely proportional relationship.
- 0 coefficient indicates no relationship.
- 1 coefficient indicates strong proportional relationship.

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

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

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

   Scaling is a technique performed in pre-processing during building a machine learning model to standardize the independent feature variables in the dataset in a fixed range.

   The dataset could have several features which are highly ranging between high magnitudes and units. If there is no scaling performed on this data, it leads to incorrect modelling as there will be some mismatch in the units of all the features involved in the model.

   There are two types of scaling:

$$MinMaxScaling: x = \frac{x - \min(x)}{\max(x) - \min(x)}$$

$$Standardization: x = \frac{x - mean(x)}{sd(x)}$$

   The difference between normalization and standardization is that while normalization brings all the data points in a range between 0 and 1, standardization replaces the values with their Z scores.

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

   The value of VIF is infinite when there is a perfect correlation between the two independent variables. The R-squared value is 1 in this case. This leads to VIF infinity as VIF equals to 1/(1-R2). This concept suggests that is there is a problem of multi-collinearity and one of these variables need to be dropped in order to define a working model for regression.

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

   The VIF formula clearly signifies when the VIF will be infinite. If the R2 is 1 then the VIF is infinite. The reason for R2 to be 1 is that there is a perfect correlation between 2 independent variables.

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 are the quantile-quantile plots. It is a graphical tool to assess the 2 data sets are from common distribution. The theoretical distributions could be of type normal, exponential, or uniform. The Q-Q plots are useful in the linear regression to identify the train data set and test data set are from the populations with same distributions. This is another method to check the normal distribution of the data sets in a straight line with patterns explained below

   - Interpretations
     - Similar distribution: If all the data points of quantile are lying around the straight line at an angle of 45 degree from x-axis.
     - Y values < X values: If y-values quantiles are lower than x-values quantiles.

    - X values < Y values: If x-values quantiles are lower than y-values quantiles.
    - Different distributions – If all the data points are lying away from the straight line.
- Advantages
    - Distribution aspects like loc, scale shifts, symmetry changes and the outliers all can be daintified from the single plot.
    - The plot has a provision to mention the sample size as well.

To summarise, the quantile-quantile (Q-Q) plot are used to plot quantiles of a sample distribution with a theoretical distribution to determine if any dataset concerned follows any distribution such as normal, uniform or exponential distribution. It helps us determine if two datasets follow the same kind of distribution. It also helps to find out if the errors in dataset are normal in nature or not.