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)**
   a) Among four seasons, summer and fall shows high demand for bikes. This is because the season suitable and pleasant to ride a bike.
   b) Considering the year variable, the demand for bikes has increased in year 2019 by 2X compared to 2018. This might be because of increased popularity.
   c) The month variable is closely related to weather, from graph it is visible that, the demand increases from January and reduces in December. Month of July and September has recorded highest count of rented bikes.
   d) People preferred to rent a bike on Friday. However, the difference is very small between Friday and other weekdays.
   e) The demand has decreased on holidays. This might be due to people go for picnic or choose to rest at home.
   f) When the weather is good and clear, people prefer to ride.

---

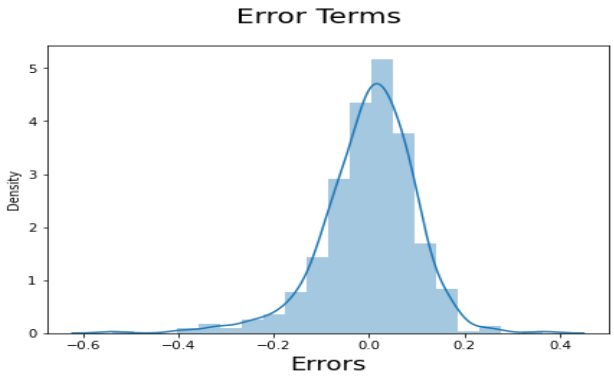
2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**
   It is important to use, as it helps in reducing the extra column created during dummy variable creation. Which inturn, reduces the correlations created among dummy variables. If we don't use drop_first = True, we will end up with an extra column with high collinearity. In short, it helps avoid redundant features.
   For example, if we consider year column, two columns will be created as year_2018 and year_2019 which are binary. With the help of either of the column the other column can be explained. If there is 0 in year_2018 that means the year was 2019. Hence, if there are n columns of certain category we will need only n-1 column to explain all related data points.

---

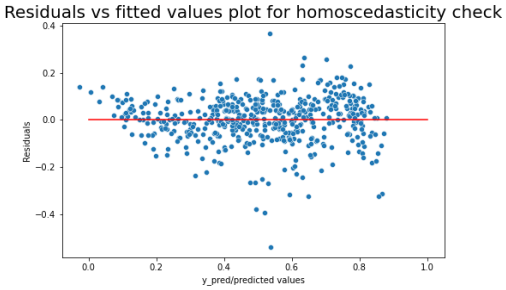
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**
   By looking at the pair-plot it is clear that, target variable i.e. count has highest correlation with temp (temperature) and atemp.

---

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**
   In order to validate assumptions I have followed following steps on the training dataset:
   a) To check if the errors are normally distributed or not by doing residual analysis with the help of histogram.



b) Check for homoscedasticity by plotting predicted values and residuals on scatter plot.



---

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, the top 3 features that contribute significantly towards explaining demand of shared bikes are: Temp, Weather, Year.

General Subjective Questions

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

- Linear Regression is one of the most fundamental algorithms in the Machine Learning world which comes under supervised learning. Basically it performs a regression task. Regression models predict a dependent (target) value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between the dependent and independent variables they are considering and the number of independent variables being used.

- Linear Regression may further divided into two types as mentioned below:

a) Simple Linear Regression –

- The most elementary type of regression model is the simple linear regression which explains the relationship between one dependent variable and one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points.
- The standard equation of the regression line is given by the following expression: $Y = \beta_o + \beta_1 X$

b) Multiple Linear Regression-

- Multiple linear regression is a statistical technique to understand the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X.
- The formulation for multiple linear regression is also similar to simple linear linear regression with the small change that instead of having beta for just one variable, you will now have betas for all the variables used. The formula now can be simply given as: $Y=\beta 0+\beta 1X1+\beta 2X2+...+\beta pXp+\epsilon$

- The best-fit line is found by minimizing the expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot. Residuals for any data point are found by subtracting predicted value of dependent variable from actual value of dependent variable.
- The strength of the linear regression model can be assessed using 2 metrics: R^2 or Coefficient of Determination and Residual Standard Error (RSE)
- Assumptions of linear regression are:
  - Linear relationship between X and Y
  - Error terms are normally distributed (not X,Y)
  - Error terms are independent of each other
  - Error terms have constant variance (homoscedasticity)
- The normal distribution of the residual terms is a very crucial assumption when it comes to making inferences from a linear regression model. Hence, it is very important that to analyze these residual terms before moving forward. The simplest method to check for the normality is to plot a histogram of the error terms and check whether the error terms are normal.
- The next step is hypothesis testing on beta co-efficient in order to support the significance of the model. The null and alternate hypothesis in case of linear regression are

H_0: $\beta_i = 0$, $\beta_j = 0$, $j \neq i$	The regression coefficient β_i equals 0, while all other coefficient β_j are set at zero.
H_A: $\beta_i \neq 0$, $\beta_j = 0$	The regression coefficient β_i unequal 0, while all other coefficient β_j are set at zero.

- Next step is to check p-value of all variables. The considered standard limit of p-value is 0.05. The variables having p-values less than 0.05 can be considered as significant for model fit.
- Another parameters to to assess the model are t-statistic, F-statistic and R square. The value of R square ranges from 0 to 1 with 1 showing best fit and 0 shows worst fit.
- For multiple linear regaression, where we are considering more than one independent variable, apart from p-value it is necessary to check other important parameters. There can be two new considerations of overfitting and multicolinearity because of which just p-value can not be a reliable parameter.
- In order to check multicolinearity of variables the heat map is best solution.

- VIF is yet another important factor to check. VIF helps explaining the relationship of one independent variable with all other independent variables. The formulation of VIF is as given below:

$$VIF_i = \frac{1}{1 - R_i^2}$$

- Next step is dropping variables based on p-value and VIF. VIF > 10 is considered as worst and VIF>5 shouldn't be ignored. Based on this criteria and simultaneously checking correlation heat map, we can drop variables one by one.
- After this, assumptions need to be checked by doing residual analysis.
- Once it is done, we can go for test dataset and check R-square and adjusted R-square percentages for both train and test datasets. The difference should not be more than 5%.

---

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

Anscombe's Quartet can be defined as a group of four data sets which are almost 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 was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and 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.

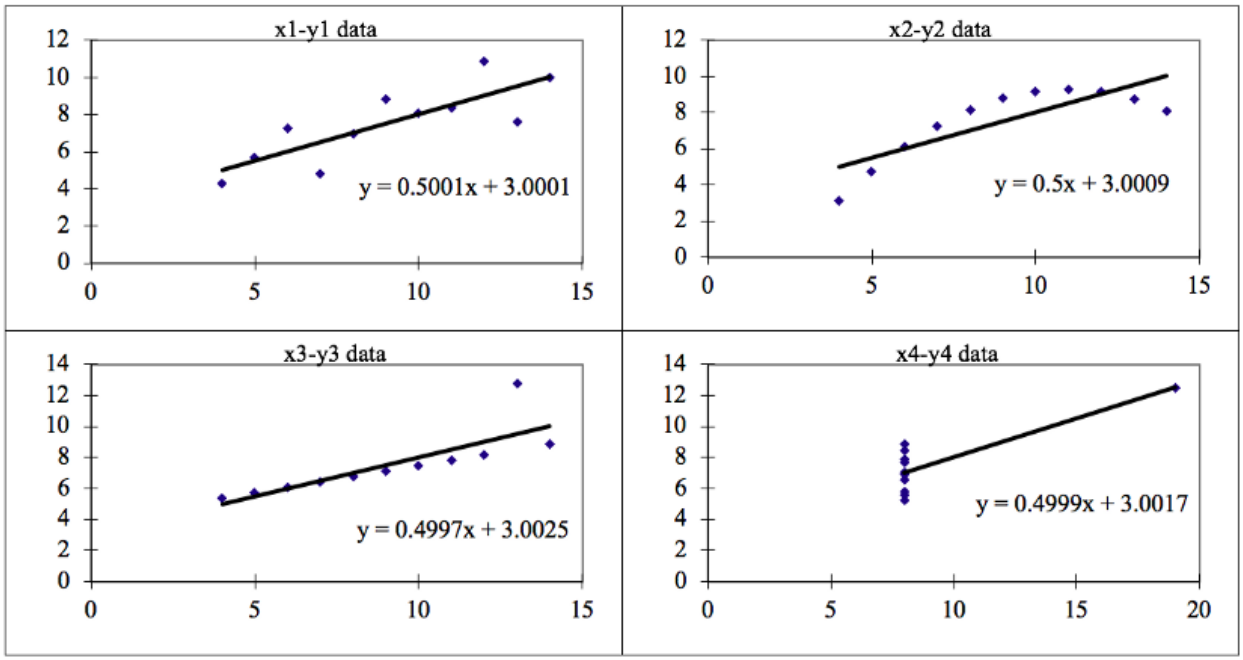
This tells us about the importance of visualizing the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets. These four plots can be defined as follows:

Anscombe's Data										
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89

The statistical information for all these four datasets are approximately similar and can be computed as follows:

Anscombe's Data										
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

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:



The four datasets can be described as:

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

---

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

Pearson's R is defined in statistics as the measurement of the strength of the relationship between two variables and their association with each other. In simple words, Pearson's correlation coefficient calculates the effect of change in one variable when the other variable changes.

For example: Up till a certain age, (in most cases) a child's height will keep increasing as his/her age increases. Of course, his/her growth depends upon various factors like genes, location, diet, lifestyle, etc.

This approach is based on covariance and thus is the best method to measure the relationship between two variables.

The Pearson coefficient correlation has a high statistical significance. It looks at the relationship between two variables. It seeks to draw a line through the data of two variables to show their relationship. The relationship of the variables is measured with the help Pearson correlation coefficient calculator. This linear relationship can be positive or negative.

For example:

- **Positive linear relationship:** In most cases, universally, the income of a person increases as his/her age increases.
- **Negative linear relationship:** If the vehicle increases its speed, the time taken to travel decreases, and vice versa.

From the example above, it is evident that the Pearson correlation coefficient, R, tries to find out two things – the strength and the direction of the relationship from the given sample sizes.

Pearson correlation coefficient formula

The correlation coefficient formula finds out the relation between the variables. It returns the values between -1 and 1. Use the below Pearson coefficient correlation calculator to measure the strength of two variables.

Pearson correlation coefficient formula:

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

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

---

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

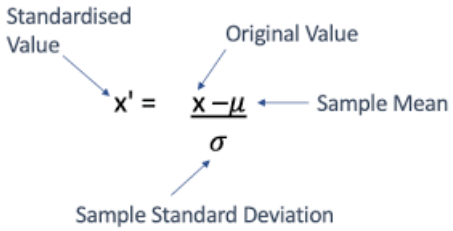
Feature scaling is the process of normalizing the range of features in a dataset. Real-world datasets often contain features that are varying in degrees of magnitude, range and units. Therefore, in order for machine learning models to interpret these features on the same scale, we need to perform feature scaling.

Normalization:

It is also known as min-max scaling, is a scaling technique whereby the values in a column are shifted so that they are bounded between a fixed range of 0 and 1.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Standardization:
On the other hand, standardization or Z-score normalization is another scaling technique whereby the values in a column are rescaled so that they demonstrate the properties of a standard Gaussian distribution, that is mean = 0 and variance = 1.



Normalization Vs Standardization Scaling:
Standardization is generally preferred over normalization in most machine learning context as it is especially important when comparing the similarities between features based on certain distance measures. This is most prominent in Principal Component Analysis (PCA), a dimensionality reduction algorithm, where we are interested in the components that maximize the variance in the data.

Normalization, on the other hand, also offers many practical applications particularly in computer vision and image processing where pixel intensities have to be normalized in order to fit within the RGB color range between 0 and 255. Moreover, neural network algorithms typically require data to be normalized to a 0 to 1 scale before model training.
At the end of the day, there is no definitive answer as to whether you should normalize or standardize your data. One can always apply both techniques and compare the model performance under each approach for the best result.

---

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**
   The extent to which a predictor is correlated with the other predictor variables in linear regression can be quantified as the R-squared statistic of the regression where the predictor of interest is predicted by all the other predictor variables. The variance inflation for a variable is then computed as:

$$VIF_i = \frac{1}{1 - R_i^2}$$

When R-square becomes 1, VIF will be infinity. When R-squared reaches 1 then it means multicolinearity exists. If the variables are highly correlated with each other then this can happens that VIF value becomes infinity.
To solve this problem we need to drop one of the variables from the dataset 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 show an infinite VIF as well).

---

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

In Statistics, Q-Q(quantile-quantile) plots play a very vital role to graphically analyze and compare two probability distributions by plotting their quantiles against each other.
Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.
This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Advantages of Q-Q Plot:

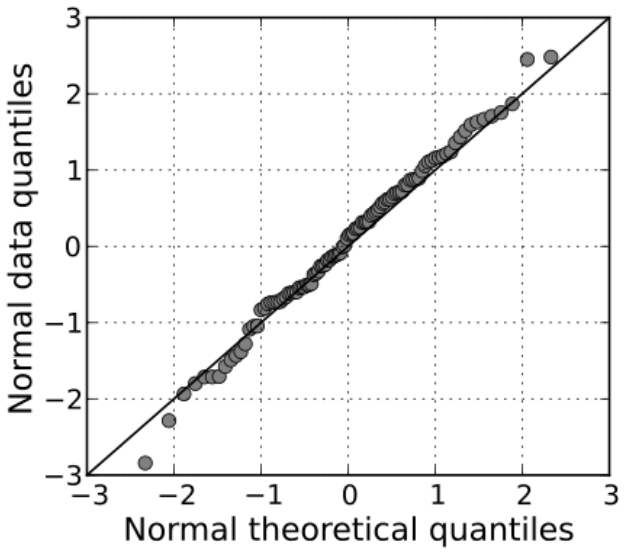
a) It can be used with sample sizes
b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.
c) It is used to check following scenarios:
If two data sets :

- come from populations with a common distribution
- have common location and scale
- have similar distributional shapes
- have similar tail behavior

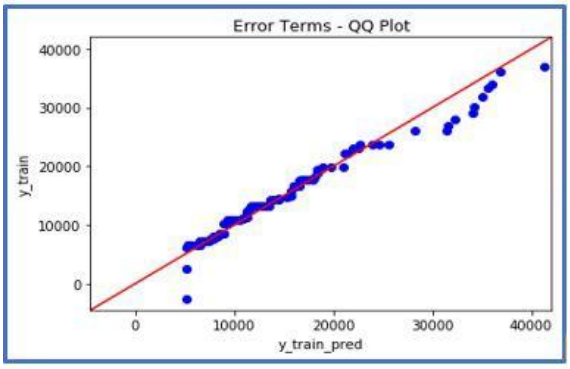
Interpretation:

Below are the possible interpretations for two data sets.

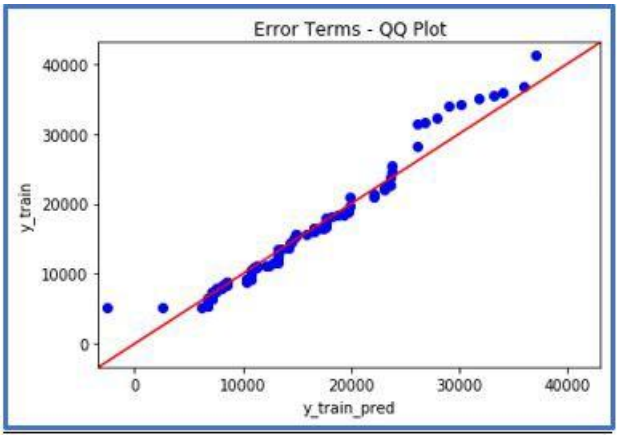
a) **Similar distribution**: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x –axis.



b) **Y-values < X-values:** If y-quantiles are lower than the x-quantiles.



c) **X-values < Y-values:** If x-quantiles are lower than the y-quantiles.



d) **Different distribution:** If all point of quantiles lies away from the straight line at an angle of 45 degree from x –axis.