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?**

Answer : The following columns were identified as categorical columns from the given dataset :

season , month , weathersit , weekday

Effect of each of these was observed by plotting graphs against the target using matplotlib in python.

Season Observation: Bikes rented are more during the fall season.
Month Observation: Bikes rented are more in the month of Sep 2019.
Weather Observation: Bikes rented are more in the Clear weather.
Weekday Observation: Bikes rented are more on Saturday and Friday

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

Answer : Multi-collinearity is undesirable, and every time we encode variables, we'll encounter this issue. One way to overcome this issue is by dropping one of the generated columns. So, we can drop either of the the parameter using drop_first which, when set to True, does precisely that.

Example : In our dataset, the variables 'temp' and 'atemp' were identified to have a high correlation value of 0.99 after encoding , ie., they are highly correlated to each other , therefore we needed to drop one of them.

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

Answer : Temp and atemp both indicate a high correlation with target "count".

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

Answer :

1. The linear relationship between predictors and target was verified using scatter plots.
2. The p-values were calculated to check if the variables are significant.

Then, the VIF of the variables was calculated to find if any multicollinearity exists between any variable.

3. Residual analysis was performed to check if he error terms are normally distributed by plotting a histogram of the error terms.

Q5 - Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer : The three most significant variables affecting the demand for shared bikes are :

- **temperature**
- **year**
- **season winter**

as these features are having positive coefficients and an increase in them is going to result into an increase in the demand for shared bikes .

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Answer : Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x).
So, this regression technique finds out a linear relationship between x (input) and y(output).
Simple linear regression is an approach for predicting a response using a single feature**.**
It is assumed that the two variables are linearly related. Hence, we try to find a linear function that predicts the response value(y) as accurately as possible as a function of the feature or independent variable(x).

Example :
If we define :
x as **feature vector**, i.e x = [x_1, x_2, …., x_n],
y as **response vector**, i.e y = [y_1, y_2, …., y_n]
for **n** observations (in above example, n=10).

Aim : to find a **line that fits best in a scatter plot of x and y,** so that we can predict the response for any new feature values.

This line is called a **regression line**.
The equation of regression line is represented as:
Yi = b0 +b1Xi

Here,

- Yi represents the **predicted response value** for i^{th} observation.
- b0 and b1 are regression coefficients and represent **y-intercept** and **slope** of regression line respectively.

our task is to find the value of b0 and b1 for which residual error or cost function is minimum, where cost function is given as :

$$minimize \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

$$J = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

In Multiple linear regression, same algorithm is applied for multiple features.

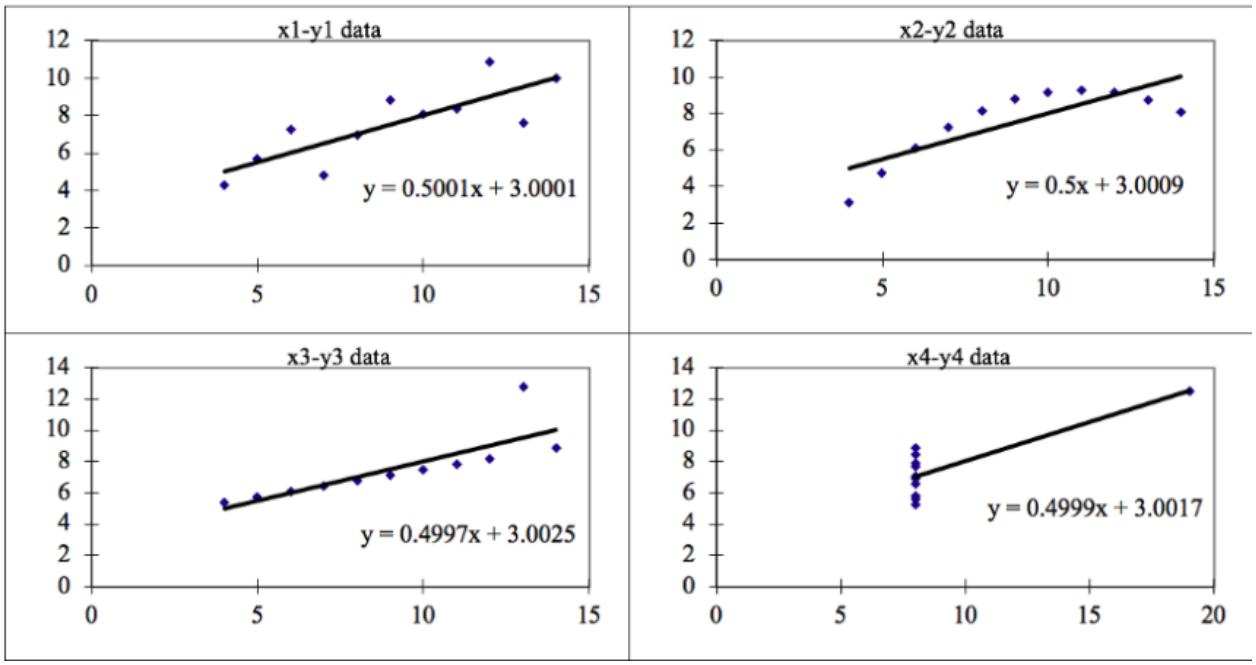
2. **Explain the Anscombe's quartet in detail.**
   Answer: Anscombe's Quartet can be defined as a group of four data sets which **are** nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that **fools the regression model** if built.
   They have very different distributions and **appear differently** when plotted on scatter plots.
   This tells us about the importance of visualizing the data before applying various algorithms 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 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.

   Example of such data set vs their respective scatter plots :

Data:

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

Plots :



3. What is Pearson's R?

Answer : 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; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1. The Pearson's correlation coefficient varies between -1 and +1 where:

- 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

Formula

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

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

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

   Ans : 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.

   Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude. t is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization/Min-Max Scaling:

- It brings all of the data in the range of 0 and
  1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

Standardization Scaling:

- 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 (σ).
- sklearn.preprocessing.scale helps to implement standardization in python.
- One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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.
   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.**
   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.

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.