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?**

Ans: From the analysis of categorical variables on dependent variable, following points are analysed:

- Rental bike counts start rising from summer and it maximum demand happen in fall season.
- The rental bike counts start rising from month of March and start dropping from month of August. But maximum demand happens in August, September and October.
- The average of rental bike is almost same for every day of week.
- The average of rental bike is almost same for working and non-working day.
- The average of rental bike is more on non-holiday than on holiday, which means people may use rental bike for their office commute.
- People prefer more bikes renting during clear and mist weather.
- From the trend we can clearly visualise that counts of rental bikes is more in year 2019 than 2018.

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

Ans: When we have categorical variable with, say 'n' levels, pd.get_dummies() helps us to create dummy variables but this categorical variable can be easily get explained by (n-1) dummy variables as one of them will give redundancy, so **drop_first=True** helps us to drop the excess one variable.

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

Ans: 'temp' (i.e., temperature in Celsius) and 'atemp' (i.e., feeling temperature in Celsius) have 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?**

Ans: On analysing the R-squared and Adjusted R-squared values of our final model which came 0.829, 0.825 respectively, which are approximately same, which indicate that our model is stable and by analysing the VIF table we also found that all predictors value are less than 5 which means there is no multicollinearity among the predictors. By this, I validated my model.

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

Ans: Top 3 features are:
a. Temperature
b. Year
c. Winter season
These 3 features are contributing most towards explaining the demand of the shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression model is used to predict the target/dependent variable based on independent variables. It is mostly used for finding out the relationship between variables and forecasting and the relationship must be linear.
Linear regression can be characterised into two types:

a) Simple Linear Regression: It has only one independent variable. Equation for Simple Linear Regression is as below:

$$\mathbf{Y = \beta_0 + \beta_1 X + E}$$

Where,
Y is dependent variable
X is independent variable
β_1 is population slope coefficient
β_0 is population Y intercept
E is the random error for X

b) Multiple Linear Regression: It has more than one independent variables. Equation for Multiple Linear Regression is as below:

$$\mathbf{Y = \beta_0 + \beta_1 X_1 + \beta_1 X_2 + \beta_1 X_3 + \ldots\ldots + \beta_1 X_n + E_i}$$

Random/Residual error E or E_i is the error or difference between actual Y or Y_a value(s) and the predicted Y_{pred} values which we got from the model.

$$\mathbf{E_i = Y_a - Y_{pred}}$$

To minimize this error we generally tend to use **Ordinary Least square** method in which we square all the residual error of X_i and then sum up to get the **Residual Sum of Square (RSS)** and then we try to minimize the RSS using **Gradient Descent method**.

$$\text{Therefore, } \mathbf{RSS} = \sum_{i=0}^{N} (\mathbf{Yi - \beta 0 - \beta 1 Xi})\mathbf{2}$$

As X and Y are known to us, we need to find the optimum value of slope (β_1) and intercept (β_0) in order to find the best fit line. To find the optimum value of β_1 and β_0 using **Gradient Descent method,** we first initialise the β_1 and β_0 as β_1=0 and β_0=0. Here a new term introduced as **Cost**

function, which is a mathematical function (error function) which is used as a criteria to find the best fit line. It is also termed as **Mean Square Error (MSE).**

$$\textbf{Cost function or MSE = RSS/N}$$

$$\textbf{Cost function, J (m, C)} = \frac{1}{N}\sum_{i=0}^{N}(\textbf{Yi} - \boldsymbol{\beta}\textbf{0} - \boldsymbol{\beta}\textbf{1Xi})\textbf{2}$$

After initiating $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_0}$, now update $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_0}$ as

$$\boldsymbol{\beta_1} = \boldsymbol{\beta_1} - \alpha\frac{\partial J}{\partial \boldsymbol{\beta 1}}$$

$$\boldsymbol{\beta_0} = \boldsymbol{\beta_0} - \alpha\frac{\partial J}{\partial \boldsymbol{\beta 0}}$$

Update $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_0}$ till we get their optimum values. Here, $\boldsymbol{\alpha}$ is the learning rate.

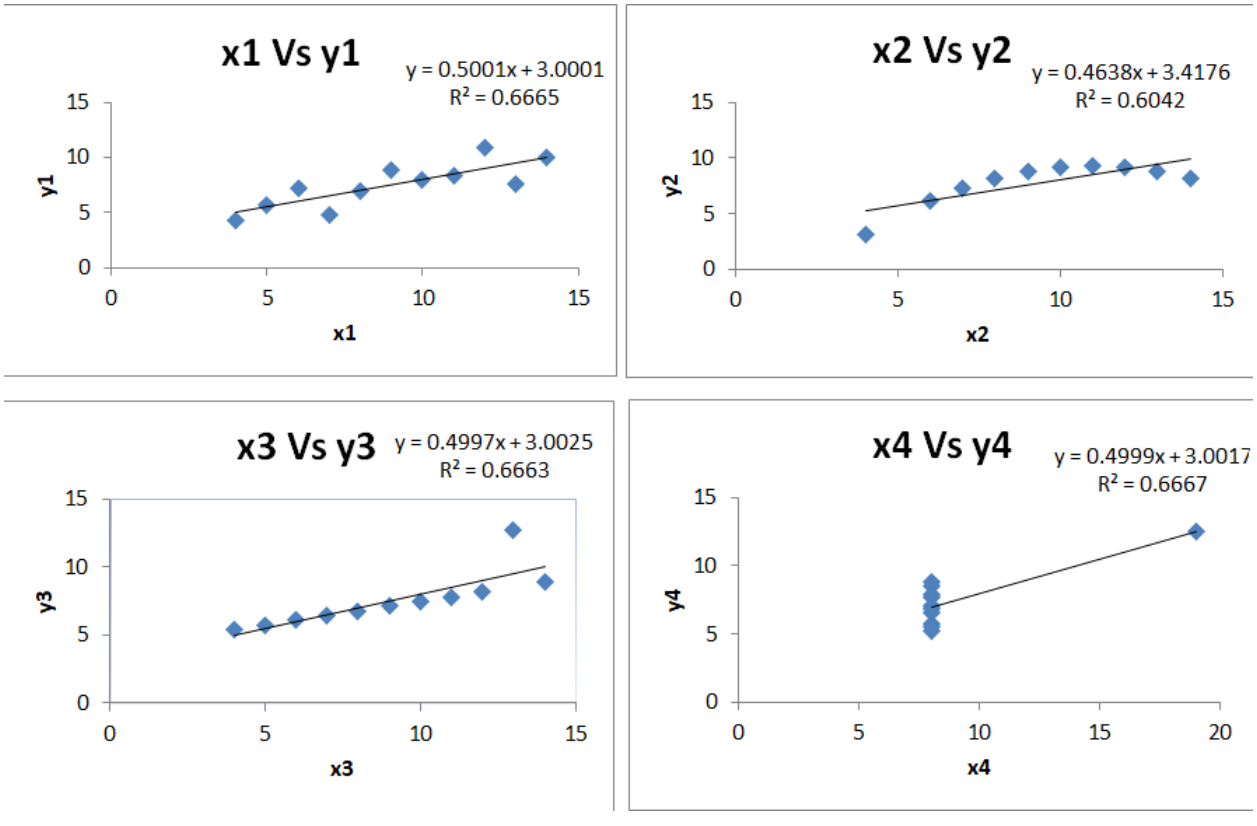
2. Explain the Anscombe's quartet in detail.

Ans: 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.

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analysing 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 provide same statistical information that involves variance, and mean of all x, y points in all four datasets.

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
N	**11**	**11**	**11**	**11**	**11**	**11**	**11**	**11**
mean	**9**	**7.5**	**9**	**7.5**	**9**	**7.5**	**9**	**7.5**
SD	**3.32**	**2.03**	**3.32**	**2.03**	**3.32**	**2.03**	**3.32**	**2.03**
R2	**0.67**		**0.60**		**0.67**		**0.67**	
Corr.	**0.82**		**0.82**		**0.82**		**0.82**	

This tells us about the importance of visualising 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 etc. Also, the Linear Regression can be only being 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:



When we see the dataset, we will find all that all the statistical parameters are almost same but when we plot scatter plot of all the four dataset, we can see various anomalies and got to know that some plot not could not even handle linear regression model. So, it is important that all the important data set features or variables must be visualized first before implementing machine learning algorithm on them which will help to develop a good fit model.

3. What is Pearson's R?

Ans: The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is the specific measure that quantifies the strength of the linear relationship between two variables in a correlation analysis. The magnitude of the correlation coefficient indicates the strength of the association. For example, a correlation of r = 0.9 suggest a strong, positive association between two variables, whereas a correlation of r = -0.2 suggest a weak and negative association.

Population Correlation Coefficient between x and y,

$$P_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum(x_i - \bar{x})^2)(\sum(y_i - \bar{y})^2)}}$$

Where,

σ_x, σ_y = Population standard deviations

σ_{xy} = Population covariance

$\bar{x}, \bar{y}$ = Population mean

Sample Correlation coefficient between x and y,

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum(x_i - \bar{x})^2)(\sum(y_i - \bar{y})^2)}}$$

Where,

S_x, S_y = Sample standard deviations

S_{xy} = Sample covariance

The value of the correlation coefficient always lies between -1 and +1, where a negative value implies negative correlation and a positive value implies positive correlation and a zero value shows no correlation.

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

Ans: Scaling is a step of data pre-processing which is applied to independent variables to normalize data within a particular range.
Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling not done then algorithm only take magnitude into account and not units and which may lead to wrong modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude. Scaling also helps in speeding up the calculations in an algorithm. It is important to note that scaling just affects the coefficients of variables and none other parameters like p-value, R-squared, F-statistic, t-statistic, etc.

Normalized Scaling:

- It is also known as MinMax Scaling.
- It brings all data in the range of 0 and 1.
- Sklearn.preprocessing.MinMaxScaler helps to implement normalization in Python.

MinMax Scaling: $X = \frac{x - \min(x)}{\max(x) - \min(x)}$

Standardized 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 (σ) one.

Standardization: $X = \frac{x - \mu}{\sigma}$

- Sklearn**.**preprocessing.scale helps to implement standardization in python.
- One disadvantage of normalization over standardization is that it loses some information in data, especially about outliers.

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

Ans: This actually happen when there is perfect correlation between two independent variables. In such case our R^2 value will become 1 and as we know that:

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

So, in such case $(1-R^2)$ will become zero and VIF becomes infinite.

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

Ans: Q-Q plot or Quantile-Quantile plot 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 plot is to find out if two data sets come from the same distribution. A 45 degree angle plotted on Q-Q plot; if the two data sets come from a common distribution, the plots 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 distribution.
A Q-Q plot is used to compare the shapes of distributions, proving a graphical view of how properties such as location, scale and skewness are similar or different in the two distributions.