Daivanshu Gandhi

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)

Answer: I have plotted the categorical variables with the target variables on boxplot and has inferred following effect on target:

- ✓ Season: fall has highest demand for rental bikes
- ✓ I see that demand for next year has grown
- ✓ Demand is continuously growing each month till June. September month has highest demand. After September, demand is decreasing
- ✓ When there is a holiday, demand has decreased.
- ✓ Weekday is not giving clear picture about demand.
- ✓ The clear weathershit has highest demand

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

Answer: drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.
If we do not drop one of the dummy variables created from a categorical variable then it becomes redundant with dataset as we will have constant variable(intercept) which will create multicollinearity issue.

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

Answer: The feature "temp" has highest correlation. It is very well linearly related with target "cnt"

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

Answer: I have checked the following assumptions:

- ✓ Error terms are normally distributed with mean 0.
- ✓ Error Terms do not follow any pattern.
- ✓ Multicollinearity check using VIF(s).
- ✓ Linearity Check.
- ✓ Ensured the overfitting by looking the R2 value and Adjusted R2.

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

Answer: Features “yr”, “temp” and season “summer” are highly related with target column, so these are top contributing features in model building.

General Subjective Questions:

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

Answer: Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. There are two types of regression–

1. Simple linear regression: Model with one independent variable.

2. Multiple linear regression: Model with more than one independent variable.

Steps that we take while building a model:
1. Reading and understanding the data.
2. Visualising the data: If there is some obvious multicollinearity going on, this is the first place to catch it. This is where you'll also identify if some predictors directly have a strong association with the outcome variable.
3. Data preparation:
• You can see if your dataset has columns with values as 'Yes' or 'No'. To fit a regression line, we would need numerical values and not string. Hence, we need to convert them to 1s and 0s, where 1 is a 'Yes' and 0 is a 'No'.
• Also convert the categorical variables into numerical using dummy variables.
• Treating the outliers if observed.
• Treating the missing values if observed.
4. Splitting the data into train and test set.
5. Rescaling the data, if required, using the minmax scaling or standardisation.
6. Dividing the train data into X and y sets for model building.
7. Building a linear model: Fit a regression line through the training data using statsmodels if statistics is of importance or else sklearn can also be used.
8. Add/remove variables unless the model has all variables with p values, VIF, r-square and prob(F-statistics) in acceptable range.
9. Residual analysis of the train data: check if the error terms are also normally distributed and also other assumptions of linear regression.
10. Making Predictions using the final model.
11. Evaluating the model

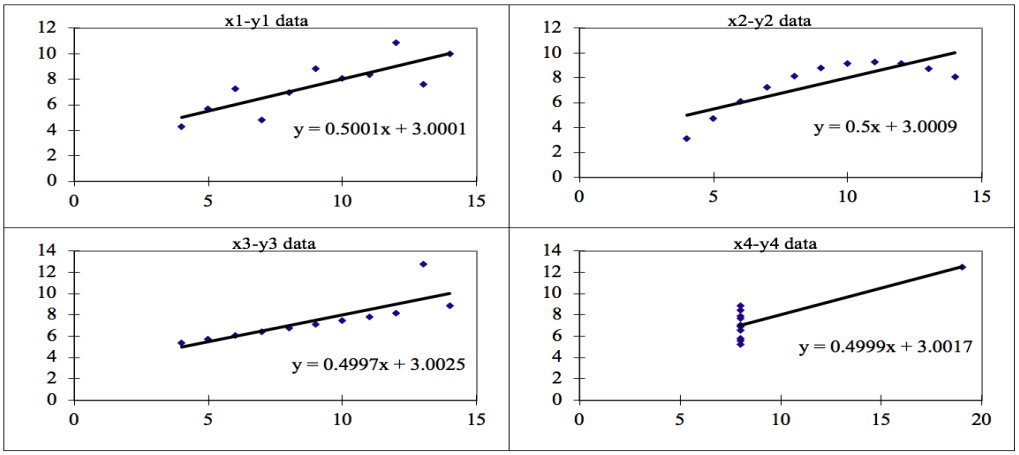
2. Explain the Anscombe’s quartet in detail. (3 marks)

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.
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

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

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:

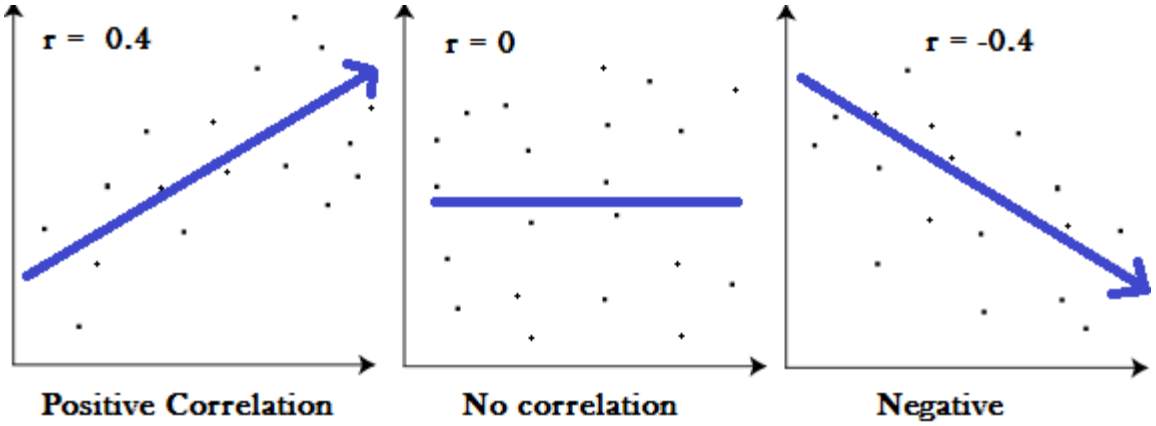


Now we see the real relationships in the datasets start to emerge. Dataset I consists of a set of points that appear to follow a rough linear relationship with some variance. Dataset II fits a neat curve but does not follow a linear relationship. Dataset III looks like a tight linear relationship between x and y, except for one large outlier. Dataset IV looks like x remains constant, except for one outlier as well

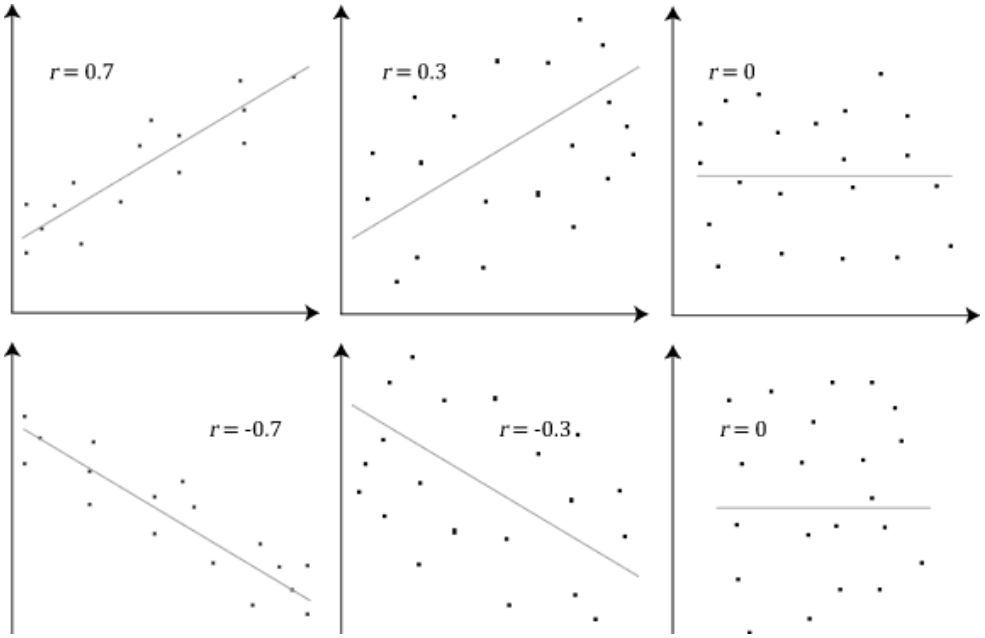
Conclusion:
We have described the four datasets that were intentionally created to describe the importance of data visualisation and how any regression algorithm can be fooled by the same. Hence, all the important features in the dataset must be visualised before implementing any machine learning algorithm on them which will help to make a good fit model.

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

Answer: The Pearson product-moment correlation coefficient (or Pearson correlation coefficient, for short) is a measure of the strength of a linear association between two variables and is denoted by r. Basically, a Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, r, indicates how far away all these data points are to this line of best fit. The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



The stronger the association of the two variables, the closer the Pearson correlation coefficient, r, will be to either +1 or -1 depending on whether the relationship is positive or negative, respectively. Achieving a value of +1 or -1 means that all your data points are included on the line of best fit – there are no data points that show any variation away from this line. Values for r between +1 and -1 (for example, r = 0.8 or -0.4) indicate that there is variation around the line of best fit. The closer the value of r to 0 the greater the variation around the line of best fit. Different relationships and their correlation coefficients are shown in the diagram below:



Pearson's r measures degree of correlation or correlation coefficient between 2 numerical variables.

Its value varies between -1 and 1.

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

0<r<5 means there is a weak association
5< r< 8 means there is a moderate association
8 < r means there is a strong association

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

Answer: Scaling is a method to normalize the range of independent variables. It is performed to bring all the independent variables on a same scale in regression. If Scaling is not done, then regression algorithm will consider greater values as higher and smaller values as lower values.

It 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.

Example Weight of a device = 500 grams, and weight of another device is 5 kg. In this example machine learning algorithm will consider 500 as greater value which is not the case. And it will do wrong prediction.

Machine Learning algorithm works on numbers not units. So, before regression on a dataset it is a necessary step to perform.

Scaling can be performed in two ways: Normalization: It scale a variable in range 0 and 1. Standardization: It transforms data to have a mean of 0 and standard deviation of 1.

Normalization vs. standardization

Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.

Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

However, at the end of the day, the choice of using normalization or standardization will depend on your problem and the machine learning algorithm you are using. There is no hard and fast rule to tell you when to normalize or standardize your data. You can always start by fitting your model to raw, normalized and standardized data and compare the performance for best results

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

Answer: VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables.

If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.

The standard error of the coefficient determines the confidence interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity. A general rule of thumb is that if VIF > 10 then there is multicollinearity.

An infinite VIF value indicates that the dependent variable may be expressed exactly by a linear combination of other variables. VIF = 1/ (1-R2), when R2=1 then VIF = Infinity Example: In our Assignment, Registered Users + Casual Users = Total no. of Users If we fit the model including these 2 variables then VIF will be infinity because of this

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

Answer: **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

It is used for determining if two data sets come from populations with a common distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. Whether the Distributions is Gaussian, Uniform, Exponential or even Pareto distribution, it can be found out.

Few advantages:

a) It can be used with sample sizes also
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.

It is used to check following scenarios:

If two data sets —
i. come from populations with a common distribution
ii. have common location and scale
iii. have similar distributional shapes
iv. have similar tail behavior

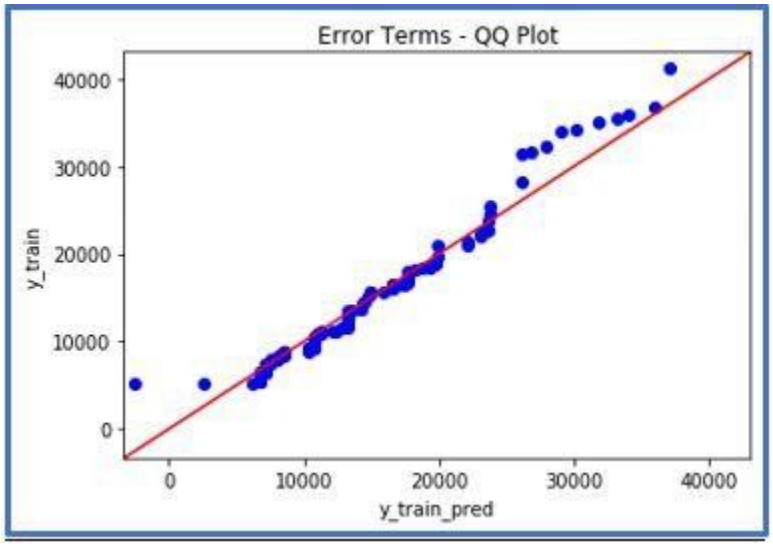
A q-q plot is a plot of the

quantiles of the first data set

against the quantiles of the

second data set.

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 they-quantiles.



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