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)

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

a. Season: 3: fall has highest demand for rental bikes
b. I see that demand for next year has grown
c. Demand is continuously growing each month till June. September month has highest demand. After September, demand is decreasing
d. When there is a holiday, demand has decreased.
e. Weekday is not giving clear picture about demand.
f. The clear weathershit has highest demand

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

Answer: 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. That 'why it is important to use drop_first=True during dummy variable creation.

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:

a. Error terms are normally distributed with mean 0

b. Error Terms do not follow any pattern

c. Multicollinearity check using VIF(s). Whether a feature is being explained by other features.

d. Linearity Check

e. 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)

Answers: Features “yr”, “temp” and season “spring” 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)

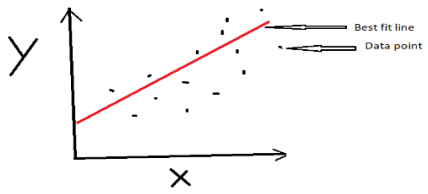
Answer: Linear regression algorithm establishes linear relationship between independent variable and dependent variable. It is used in prediction and projection. Simple linear regression where there is only one independent variable and in multiple linear regression, there would be more than one independent variables.

Independent variables are also called predictors and dependent variable is output variable. We select type of linear regression based on number of dependent variables. The equation of simple linear regression as follow

$$y = mx + c$$

Where y is the predicted variable (dependent variable), m is slope of the line, x is independent variable, c is intercept(constant).

It is cost function which helps to find the best possible value for m and c which in turn provide the best fit line for the data points.



$e1 = y1 - y_{pred}$, here e1 is called residual

To find the best fit line, we need to minimize sum of squares of residuals

Cost Function

$$J(m,c) = \sum(y1-mx1-c)^2$$

Here, it is Gradient Descent concept which help us reduce the cost function. It changes value for slope and intercept to reduce cost

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

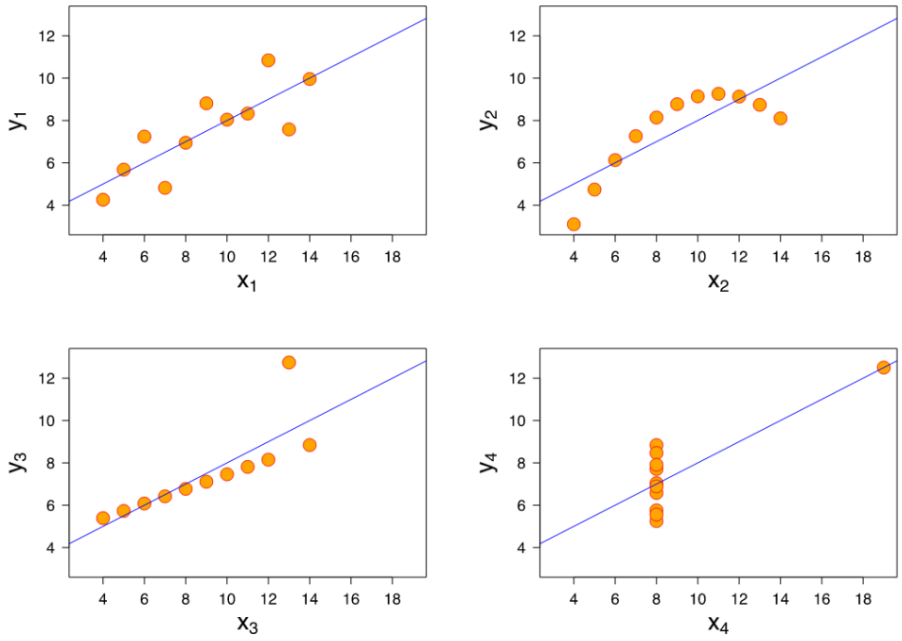
Answer: Anscombe's quartet prove that only relying on summary statistics can be dangerous. It is groups of four datasets that appear very similar statistically means all the four data set have identical statistics such as mean, variance, average, correlation and even it follow the same linear equation but it tells different stories when it plotted on graph. Example: See below the four datasets below which consists eleven *(x, y)* pairs.

I		**II**		**III**		**IV**	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

The above data set have following statistics:

a. Average value for x variable in each dataset is 9
b. Average value for y variable in each dataset is 7.50
c. Variance for x and y are 11 and 4.12 respectively for each data set.
d. Correlation between x any variable is 0.816 for each dataset
e. Even Linear regression for each dataset is as y = 0.5x + 3

Despite so much similarities from statistical point when we do scatter plot it looks like as



So, it is very important to visualize data.

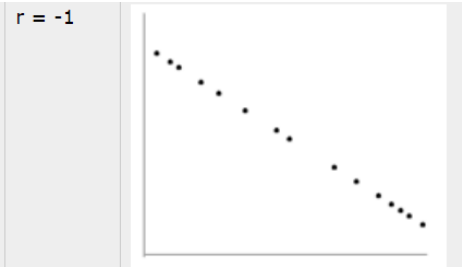
3. What is Pearson's R? (3 marks)

Answer: Pearson's R indicates the relationship between two continuous and quantitative variables. Example:- Salary and Experience, Age and IQ, Number of days and Covid-19 infected persons etc.

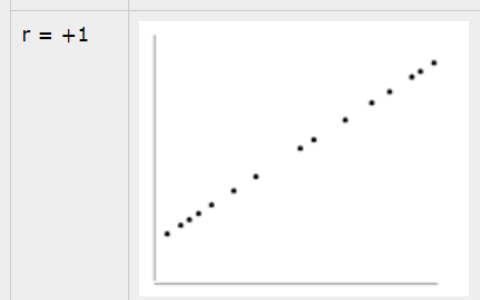
How strongly the two variables are associated can be drawn by pair plot or scatter plot. There is linear relationship between two variables. If the two variables are not linearly correlated, then it is not Pearson's R.

The Pearson's correlation coefficient for a continuous variable can range from -1 to + 1

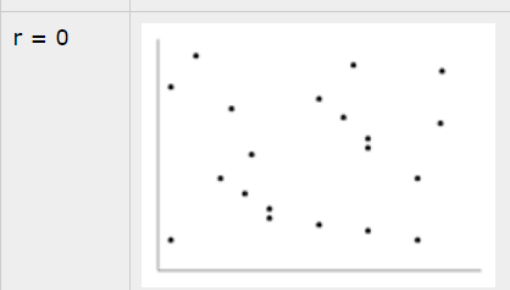
If data lie on a perfect straight line with negative slope, then r = -1



If data lie on a perfect straight line with positive slope, then r = +1



If there is no linear relationship between variables , then r = 0



Positive correlation indicates the both the variable increase and decrease together.
Negative correlation indicates the one the variable increase and the other variable decrease and vice versa.

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. 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

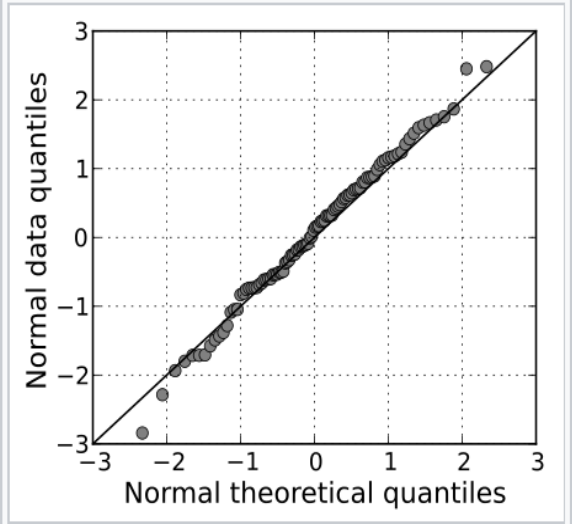
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer: When there is a perfect relationship then VIF = Infinity whereas if all the independent variables are orthogonal then to each other then VIF = 1.0. Means if a variable is expressed exactly by a linear combination of other variable then it is said that VIF is infinite.

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

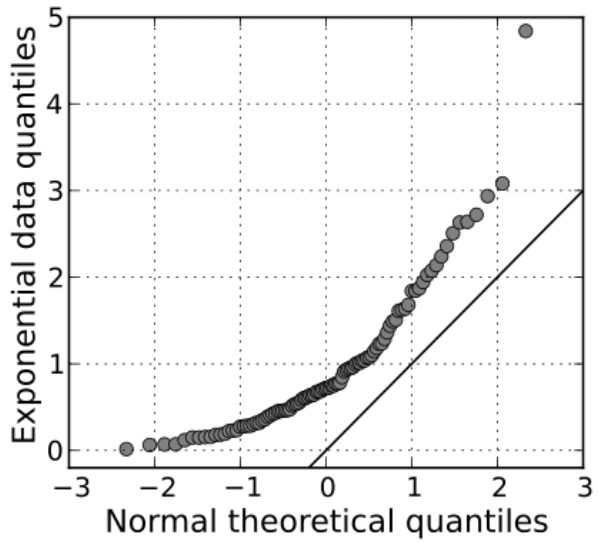
Answer: Q-Q plot is graphical probability plot. It is used for comparing two probability distributions by plotting their quantile against each other. Is the Curve on a graph is normally distributed? Q-Q plot can answer this. Whether the Distributions is Gaussian, Uniform, Exponential or even Pareto distribution, it can be found out.

If all the points plotted on the graph perfectly lies on a straight line, then we can clearly say that this distribution is Normally distribution.



The linearity of the points suggests that the data are normally distributed.

This Q–Q plot compares a sample of data on the vertical axis to a statistical population on the horizontal axis. The points below follow a strongly nonlinear pattern



A normal Q–Q plot of randomly generated, independent standard exponential data, (X ~ Exp(1))