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. I have carried out the analysis of categorical variables using boxplot and these are the inferences that I got,

- In season fall has highest demand for rental bikes.
- The demand of bikes for next year has grown.
- Demand is continuously growing each month till June. September month has highest demand, after that the demand is decreasing.
- When there is a holiday, demand has decreased.
- In weekday Sunday has the highest demand while tuesday has the lowest.
- The clear weathershit has highest demand.
- During September, bike sharing is more. During the year end and beginning, it is less, could be due to extereme weather conditions.

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

Ans. It is important to remove the unnecessary column created during dummy variable creation. So that the correlations created among dummy variables is minimised. **drop_first=True** removes the first column created while the dummy variable is created.

For example,

Lets say in our assignment, pd.get_dummies(data=bikeshare,columns=["season"],drop_first=True) creates 4 dummy variables season_fall,season_spring, seaso_ summer, season_winter. Here season_fall gets created first so it it removed after dummy variable creation.

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

Ans. Temp variable has the highest correlation of 0.99.

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

Ans. I have verified the Linear regression model on training set by :

i. Checking the normality of error terms: We found out that the error terms were normally distributed
ii. Checking the linearity: There was a linear relationship between the variables.
iii. Checking for independence of residuals: We saw that the error terms are independent of each other.
iv. Homoscedastic check: We observed no relationship in residual values.
v. Multicollinearity check: We observed that the VIF value of features got below five which indicates that there is insignificant multicollinearity between variables.

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

Ans. The top 3 features that contribute significantly in the final model are:

- temp
- yr
- weathersit_good

General Subjective Questions

6.Explain the linear regression algorithm in detail

Ans. Linear regression is defined as a statistical model that can be used to describe the linear relationship between a dependent variable (y variable) and one or more independent variables (x variable). According to the linear relationship, the change in the dependent variable(increase or decrease) is proportionally reflected in the independent variable(increase or decrease) also.

Mathematically the relationship is defined as,

$$y=mx+c$$

where,

y is the dependent variable,

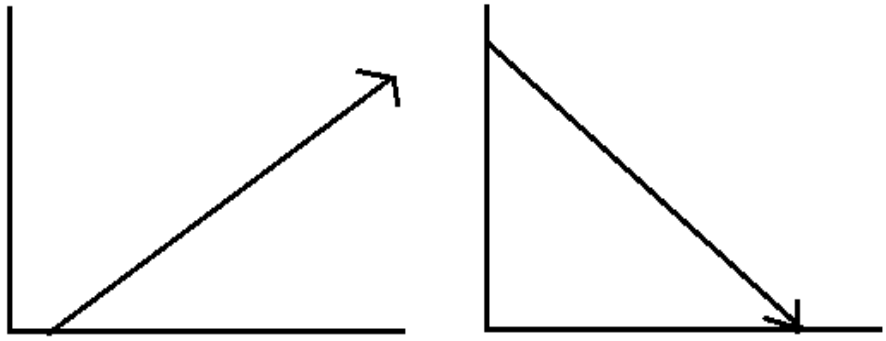
X is the independent variable,

M is the slope of the regression line

and c is the intercept of y or constant.

Furthermore, the linear relationship can be classified into two−

o Positive Linear Relationship: If the dependent variable increases with increase in dependent variable the relationship between these variable is known as positive linear relationship.



Positive Linear Relationship Negative Linear Relationship

o Negative Linear relationship: If independent increases and dependent variable decreases it can be termed as a negative linear relationship.

Linear Regression can be classified into two:

- Simple linear regression: It is used when there is only 1 independent variable.
- Multiple linear regression: It is used when there is more than one independent variables.

There are several assumptions that are to be made in linear regression for the provided dataset:

- The relationship between the dependent and independent variable should be linear.
- Error terms are normally distributed.
- There is constant variance of errors across i.e. no visible pattern in residual values. This is known as homoscedasticity.
- There is no dependency between the independent and independent variable.ie little to no multicollinearity.
- Also it assumes that there is no dependency between the residual error .i.e no autocorrelation.

7.Explain the Anscombe's quartet in detail.

Ans. It is a set of four datasets that was created by statistician Francis Anscombe in 1973. The specialty of these 4 datasets is that they share the same descriptive statistics (mean, variance, R-Squared, correlations, and linear regression lines) and each of them contain eleven (x, y) pairs. But things change entirely when they are graphed. Each graph displays a different story irrespective of their similar

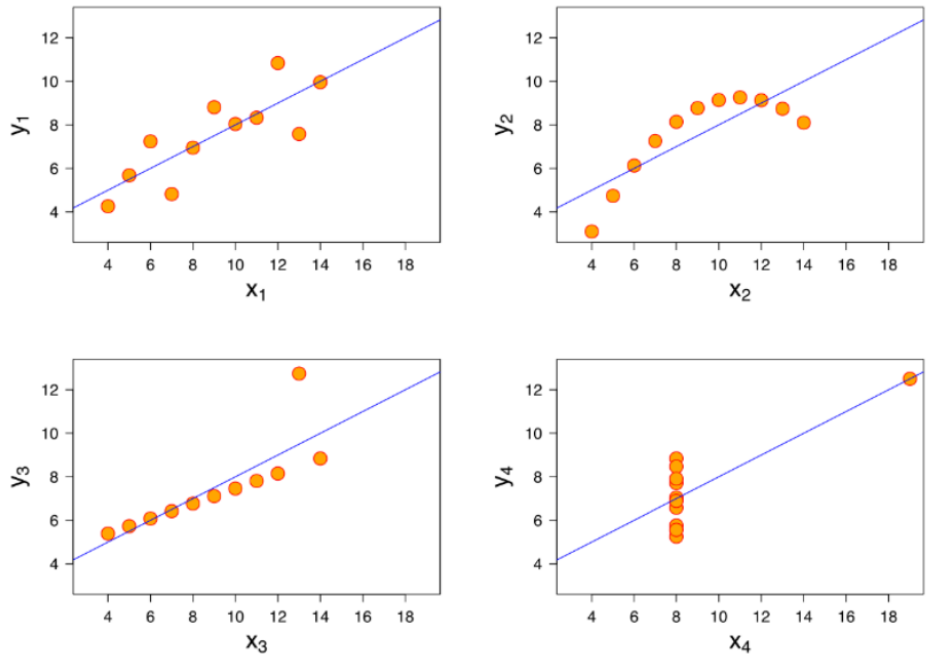
summary statistics. These datasets were created by Anscombe to signify the value of data visualization in understanding and interpreting statistical analyses.

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	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
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

By analysing the summary statistics it is apparent that the mean and variance are identical across the four datasets,

- Mean of x is 9 and mean of y is 7.50 across 4 datasets.
- For the four datasets, the variance of x is 11 and variance of y is 4.13
- Correlation coefficient is 0.816.

While plotting the four datasets along the x-y coordinate , we get different regression lines.

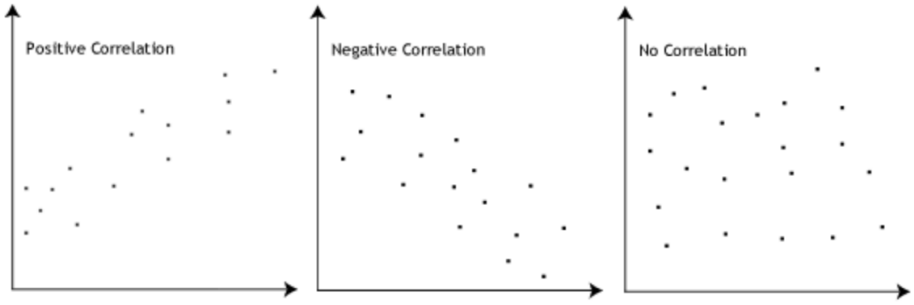


- Dataset 1 looks good and seems to have well fitted models.
- While dataset 2 isn't normally distributed.
- Dataset 3 displays linear distribution with an outlier present.
- Dataset 4 has high correlation coefficient due the presence of a single outlier.

8.What is Pearson's R?

Ans. The strength of the linear relationship between two variables is determined by Pearson's R. Its denoted by r, and it is also known as Pearson's correlation coefficient. The range of Pearson's correlation coefficient is from -1 to +1.

When the Pearson's correlation coefficient is zero, it indicates that there's no association between the two variables. If the coefficient value is greater than 0, it emphasizes that there is a positive linear relationship between the variables i.e., when one variable increase or decrease, there is a proportional change in the other variable. If the value is less than 0 indicates a negative linear relationship between the variables i.e., when one variable increase or decrease, there is an inversely proportional change in the other variable.



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

Ans. Scaling is done to standardise the independent variables present in the data within a fixed range. It is carried out during the preprocessing of data to deal with highly varying magnitudes, values or units. Feature scaling ensures that no feature is dominated by another variable in raw data.

Normalized Scaling	**Standardized Scaling**
1. Mininum and maximum values used for scaling .	1. Mean and standard deviation used for scaling.
2. Used when distribution of data is unknown.	2.Used when we want zero mean and unit standard deviation.
3.Gets highly affected by outliers.	3.Gets least affected by outliers.
4.There's a transformer called MixMaxScaler in Scikit Learn.	4. There's a tranformer called StandardScaler in Scikit Learn.

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

Ans. VIF stands for variance inflation factor, it quantifies the amount of multicollinearity present in the regression analysis. In a regression model, multicollinearity happens when two or more independent variables have a strong correlation with one another. VIF indicated how much variance of regression model coefficient is inflated by presence of multicollinearity.

The value of VIF becoming infinite states that there is a strong correlation between the independent variables. If the correlation is perfect, R-squared value(denoted by R) will be equal to 0.

VIF= 1/(1-R^2)

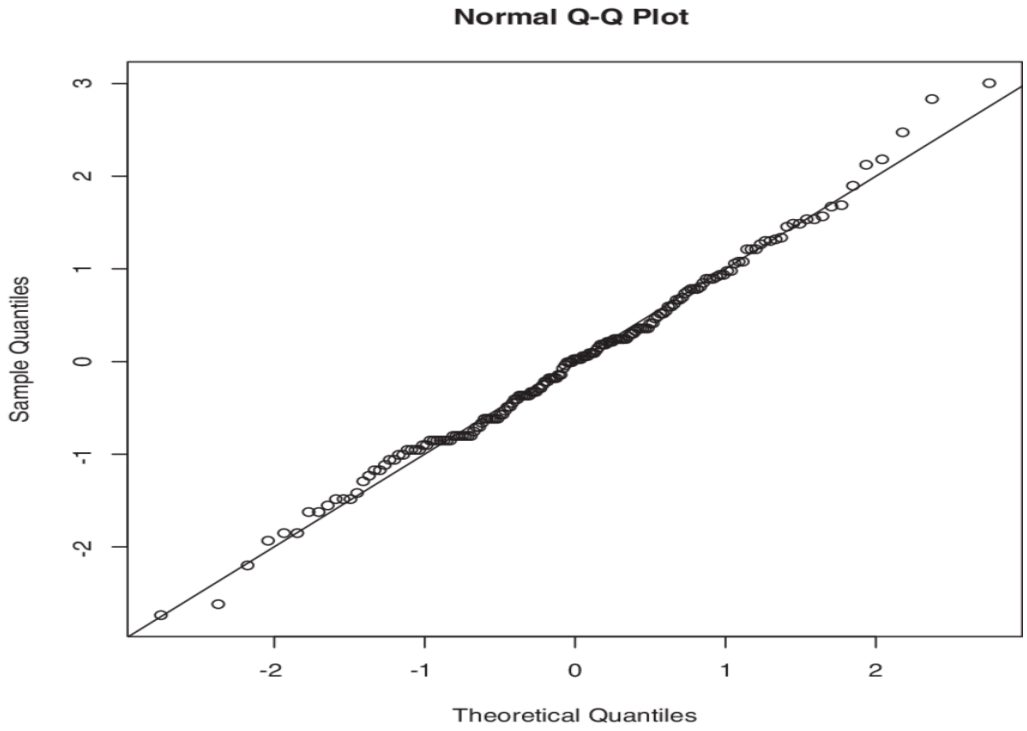
=1/(1-1) =1/0

= infinity

So VIF becomes equal to infinity. In order to solve this problem, we will have one variable that causes this multicollinearity among the two variables.

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

Ans. The graphical method used to find out if two datasets come from populations with a common distribution is called as quantile-quantile plot.



It is primarily used to find out the distribution similarity between the datasets. In linear regression, Q-Q plot is used to find the difference between the observed values and predicted values which is known as residuals. We can check the assumption of normality of residuals by plotting the quantiles of residuals against the quantiles of theoretical distribution. Q-Q plots can also be used for the outlier detection. It can also to check whether the linearity assumption between the dependent and independent variables.