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

- Season: 3: 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?**

Ans: During dummy variables the attribute drop_first = True is very important to use because as it helps in reducing the extra column and make the model less complex. And hence it reduces the correlations created among dummy variables.

For the attribute drop_first: bool, the default value is False, which implies whether to get k-1 dummies out of k categorical levels by removing the first level.

For example: Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If value A is 1 then value of B & C is 0, if value B is 1 then value of A & C is 0. Therefore if the value of A & B is 0 then definitely it would be C . So we don't need three variables.

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 with the target variable.

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

Ans: I have validated the assumption of Linear Regression Model based on below 5 assumptions:

- Normality of error terms -> Error terms should be normally distributed
- Multicollinearity check -> There should be insignificant multicollinearity among variables.
- Linear relationship validation -> Linearity should be visible among variables
- Homoscedasticity -> There should be no visible pattern in residual values.
- Independence of residuals -> No auto-correlation

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

Ans: Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes –

- temp
- winter
- sep

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: Linear regression may be defined as the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

Mathematically the relationship can be represented with the help of following equation –

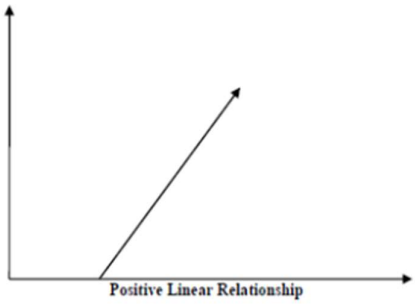
$Y = mX + c$

Here, Y is the dependent variable we are trying to predict. X is the independent variable we are using to make predictions. m is the slope of the regression line which represents the effect X has on Y c is a constant, known as the Y-intercept.

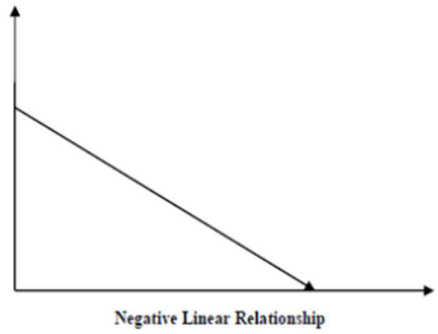
If X = 0, Y would be equal to c.

Furthermore, the linear relationship can be positive or negative in nature as explained below−

o Positive Linear Relationship: ♣ A linear relationship will be called positive if both independent and dependent variable increases. It can be understood with the help of following graph –



o Negative Linear relationship: ♣ A linear relationship will be called positive if independent increases and dependent variable decreases. It can be understood with the help of following graph –



Linear regression is of the following two types –

- Simple Linear Regression
- Multiple Linear Regression

Assumptions –

The following are some assumptions about dataset that is made by Linear Regression model-

- Multi-collinearity –
  o Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.
- Auto-correlation –
  o Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.
- Relationship between variables –
  o Linear regression model assumes that the relationship between response and feature variables must be linear.
- Normality of error terms –
  o Error terms should be normally distributed
- Homoscedasticity –
  o There should be no visible pattern in residual values.

2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

	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

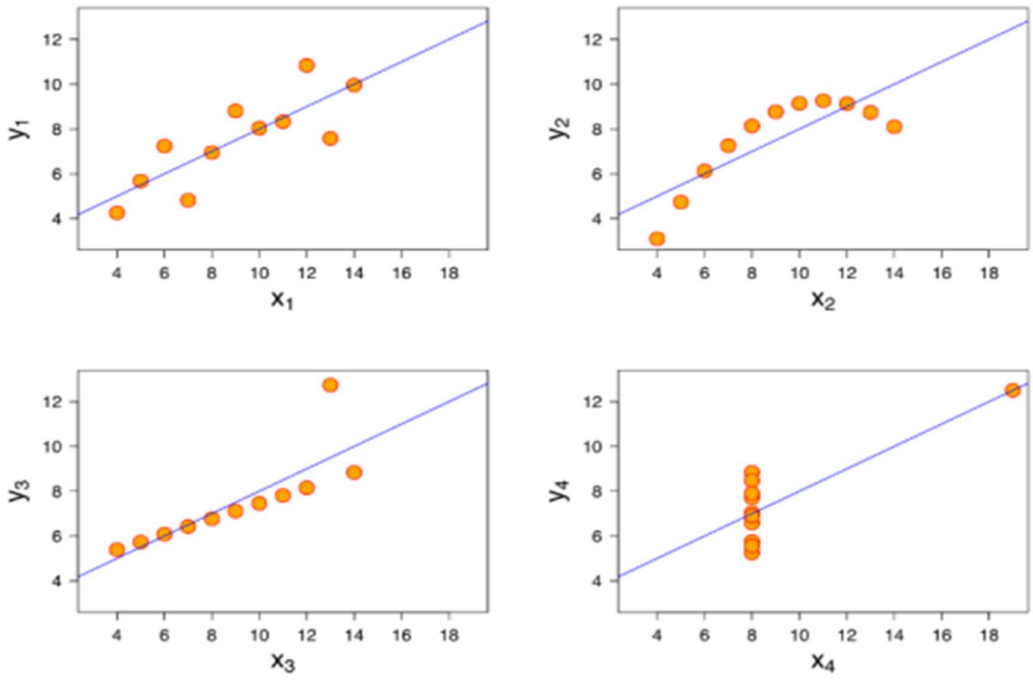
The summary statistics show that the means and the variances were identical for x and y across the groups:

• Mean of x is 9 and mean of y is 7.50 for each dataset.

• Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset

• The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset.

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:



• Dataset I appears to have clean and well-fitting linear models.

• Dataset II is not distributed normally

. • In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.

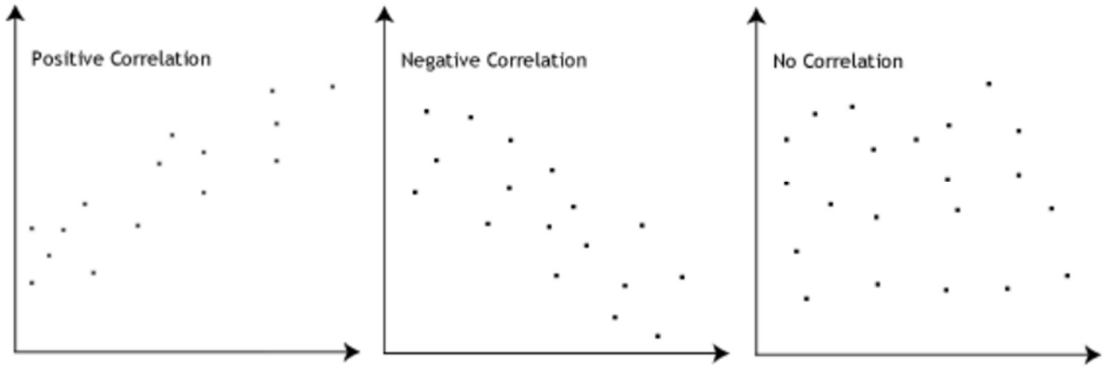
• Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

3. What is Pearson's R?

Ans: Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

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:



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?

Ans: Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Example: If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use

Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

Normalized scaling	Standardized scaling
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
It is really affected by outliers.	It is much less affected by outliers.
Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization

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

Ans: If there is perfect correlation, then VIF = infinity.

A large value of VIF indicates that there is a correlation between the variables.

If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

When the value of VIF is infinite it shows a perfect correlation between two independent variables.

In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Ans: 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

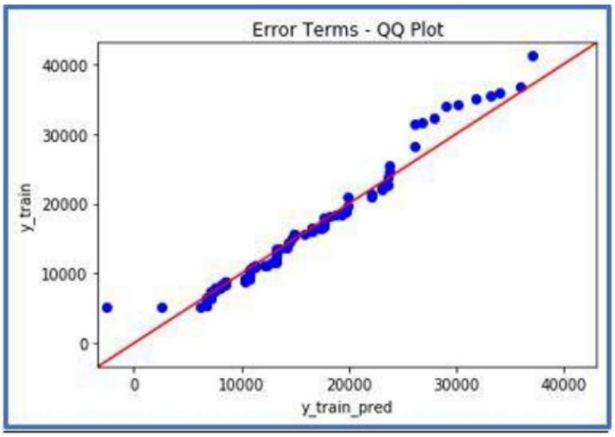
Interpretation:

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 the y-quantiles.



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