Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

From the analysis of categorical variables, the following are their effects on the dependent variable:

- There is no significant difference in the bike rental in various seasons, but the company can increase their revenue by having more bikes for hire during SUmmer, fall and Winter.
- The sales is highest during the month of September
- The revenue has increased from previous year to the current year hence this is an indicator that the sales will increase in the further years
- There is no demand during worst weather condition and less in bad weather
- There is no significant rise in demand during holidays

Q2.Why is it important to use drop_first=True during dummy variable creation?

If we do not use drop_first = True, then multicollinearity occurs which causes Dummy Variable Trap. That is there would be redundancy of variables we need drop_first=True.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

From the pair plot among numerical variables, the variables temp and atemp have the highest correlation with the target variable.

Q4.How did you validate the assumptions of Linear Regression after building the model on the training set?

We validate the assumptions of Linear Regression using residual Analysis,
By plotting a distplot, we check if the plot is following normal distribution.

Q5.Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

From the final model, the top 3 features that contribute to the demand of shared bikes are year, weather and temp.

General Subjective Questions

Q1.Explain the linear regression algorithm in detail.

It is a supervised machine learning model in which the predicted outcome is continuous with a constant slope. It shows the linear relationship between the independent and dependent variable which is called linear regression. It is of two types :

1. Simple Linear Regression
2. Multiple Linear Regression

Simple Linear Regression : This uses the equation of straight line (slope intercept form)

$$Y = mx + c$$

Where x - is the input and y is the prediction

m - intercept and c - coefficient

Using this method, the model tries to find a best fit line which shows the relationship between x & Y with minimal error.

Multiple Linear Regression: This uses a complex equation since there are more than one inputs to detect the output., This uses the equation:

$$Y = B0 + B1X1 + B2X2 + ... + BnXn$$

Where B0 is the constant
B1,B2,...Bn are model parameters
X1,X2,...Xn are feature values.

There are certain assumptions to keep in mind :

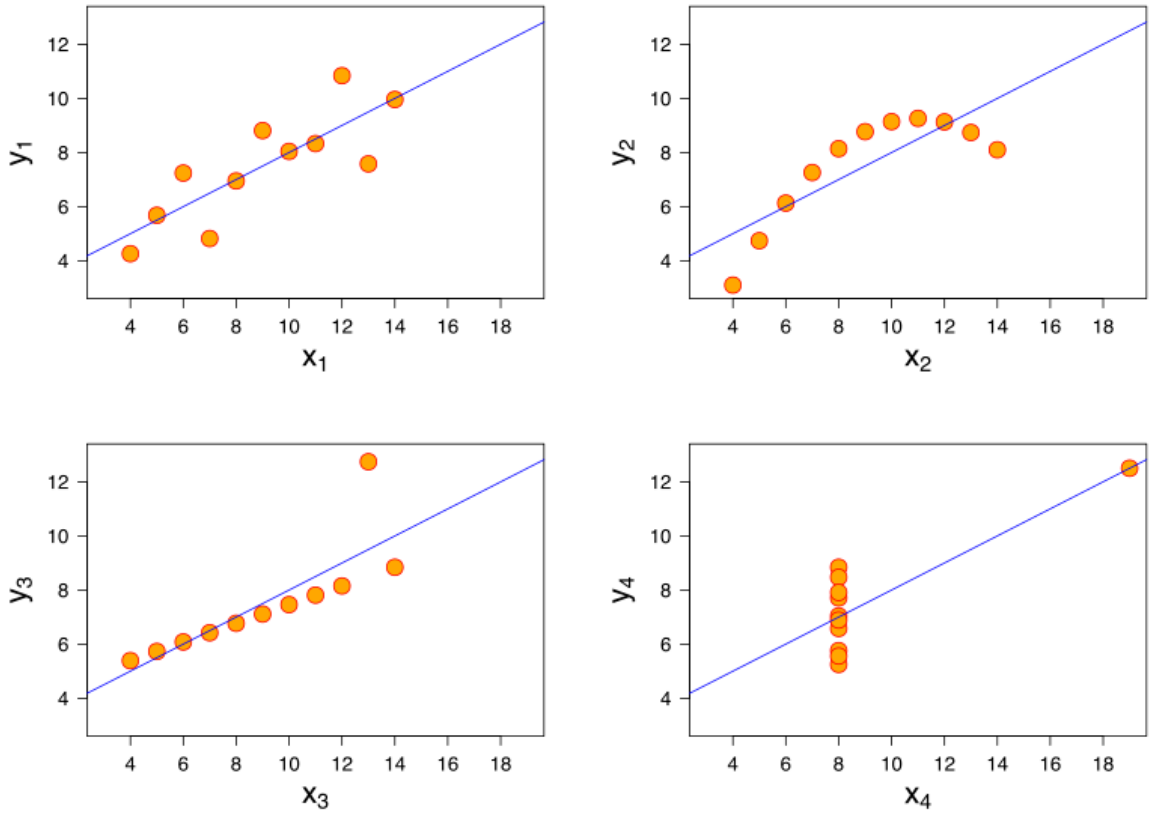
1. Linear RelationShip : There is a linear relationship between predictor and response variable.
2. No Multicollinearity : No predictor variables are highly collinear
3. Independence : All observations are independent
4. Homoscedasticity : The residuals have constant variance.
5. Multivariate Normality: The residuals follow normal distribution

Q2.Explain the Anscombe's quartet in detail.

Anscombe's quartet is a collection of four datasets that, although they share virtually identical statistical characteristics, appear to be very distinct from one another when graphed. Francis Anscombe, a statistician, developed the quartet in 1973 to show how important it is to visualise data in addition to computing statistical measures.

The four datasets in Anscombe's quartet all have the same descriptive statistics for their x and y variables, which include:

Mean of x: 9.0
Variance of x: 10.0
Mean of y: 7.5
Variance of y: 3.75
Correlation between x and y: 0.816



However, when plotted, the datasets appear quite different:

- The first dataset shows a linear relationship between x and y with a clear positive slope.
- The second dataset appears to follow a non-linear relationship with an outlier in the dataset.
- The third dataset is a perfect example of why you should always visualize your data. The dataset has a relationship between x and y but is influenced by one outlier that is skewing the line of best fit.
- The fourth dataset appears to have no correlation between x and y, but upon closer inspection, it is revealed that it is actually a perfect example of a y being dependent on x, but only when x is one of two specific values.

The quartet shows that it can be misleading to comprehend a dataset merely by looking at summary statistics.When the data is plotted, significant trends or outliers that would not be apparent from a simple number analysis are shown.As a result, data visualisation is a crucial component of data analysis and should be utilised in addition to, not in substitute of, statistical methodologies.

The table with the values are as shown below:

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

Q3.What is Pearson's R?

Pearson's Correlation is the commonly used method for numerical variables, where it assigns a value from -1 to +1 , where -1 means that the variables are highly correlated but with a negative slope, and +1 means that the variables are highly correlated with positive slope, 0 means that the variables are not correlated.

The formula for it is as follows:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where,

- n is sample size
- x i , y i are the individual sample points indexed with *i.*
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$.

Q4 What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling is the way to normalise the data to fit into a particular range. This is done when pre-processing the data for Machine Learning or other data analytic purposes.

scaling is performed for making data points generalized so that the distance between them will be lower.

In case of normalised Scaling we use the minimum value of the attributes and the maximum value of attributes to generalise a data point.

The formula for it as below:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

This brings the datapoint between the range of 0-1.

In case of standardised scaling, we use the standard deviation to generalize a value.

The formula for that is as below:

$$X' = \frac{X - \mu}{\sigma}$$

Q5.You might have observed that sometimes the value of VIF is infinite. Why does this happen?

VIF or Variance Inflation Factor is the measure of amount of multicollinearity in Linear Regression. It is a tool to identify the degree of multicollinearity.

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}$$

where:

R_i^2 = Unadjusted coefficient of determination for regressing the ith independent variable on the remaining ones

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

Q6.What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Quantile-Quantile Plot is a graphical tool which helps in plotting quantiles of a sample distribution against quantiles of a theoretical Distribution. This helps us in determining if a dataset follows which kind of distribution

QQ plots is very useful to determine

- If two populations are of the same distribution
- If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it's met using this.
- Skewness of distribution