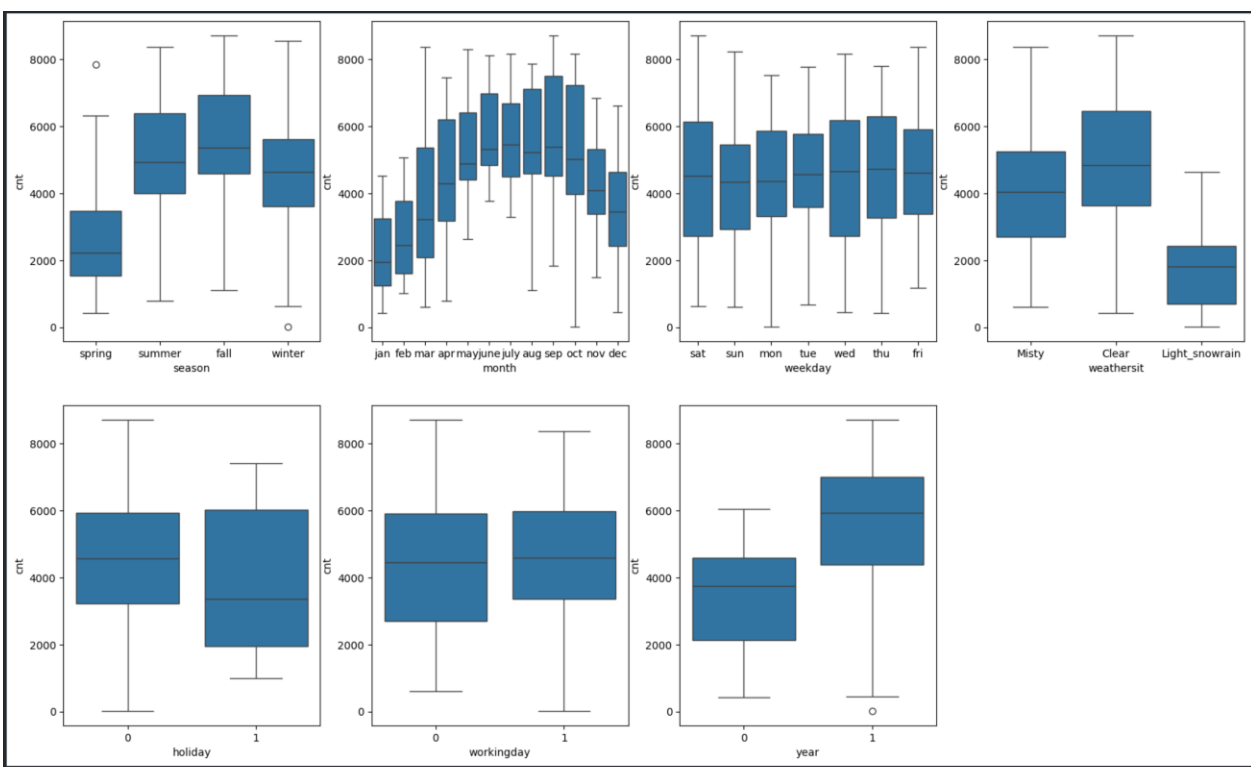
Assignment-based Subjective Questions

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:

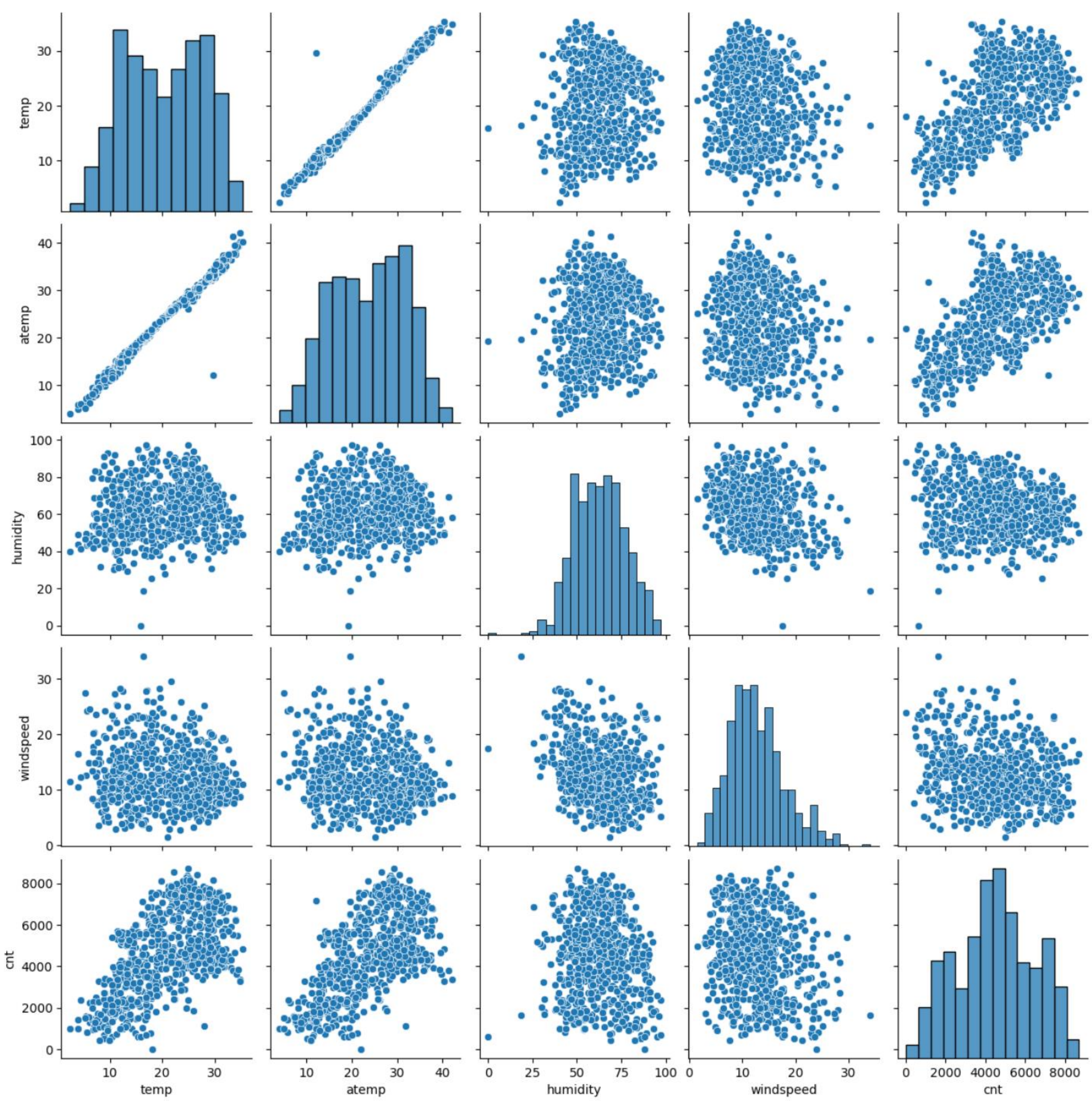
I have done analysis on categorical columns using the boxplot and bar plot. Below are the few points we can infer from the visualization –

- Fall season seems to have attracted more booking. And, in each season the booking count has increased drastically from 2018 to 2019.
- Most of the bookings has been done during the month of May, June, July, August, September, and October. Trend increased starting of the year till mid of the year and then it started decreasing as we approached the end of year.
- Clear weather attracted more booking which seems obvious.
- Thu, Fir, Sat, and Sun have a greater number of bookings as compared to the start of the week.
- When it's not holiday, booking seems to be less in number which seems reasonable as on holidays, people may want to spend time at home and enjoy with family.
- Booking seemed to be almost equal either on working day or non-working day.
- 2019 attracted a greater number of bookings from the previous year, which shows good progress in terms of business.



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.
Syntax -
drop_first: bool, default False, which implies whether to get k-1 dummies out of k categorical levels by removing the first level.
Let's say we have 3 types of values in Categorical column, and we want to create dummy variable for that column. If one variable is not A and B, then It is obvious C. So, we do not need 3rd variable to identify the C.

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



‘temp’ & ‘atemp’ 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? (3 marks)
Answer:

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? (2 marks)
Answer:
Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes –

- temp
- year
- season_winter

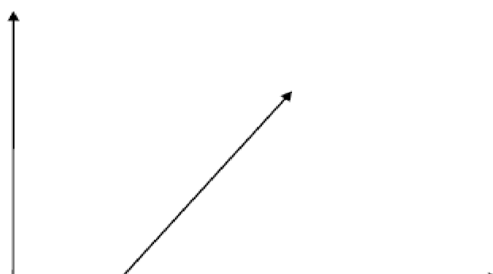
General Subjective Questions

1. Explain the linear regression algorithm in detail?

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.

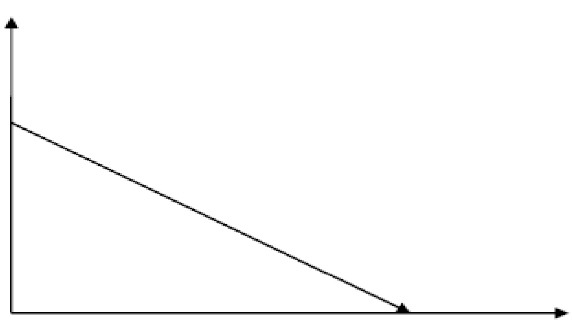
Furthermore, the linear relationship can be positive or negative in nature as explained below−

- 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 −



Positive Linear Relationship

- 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 –



Negative Linear Relationship

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 –

- 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 –

- Another assumption Linear regression model assumes is that there is very little or no autocorrelation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.

Relationship between variables –

- Linear regression model assumes that the relationship between response and feature variables must be linear.

Normality of error terms –

- Error terms should be normally distributed.

Homoscedasticity –

- There should be no visible pattern in residual values.

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

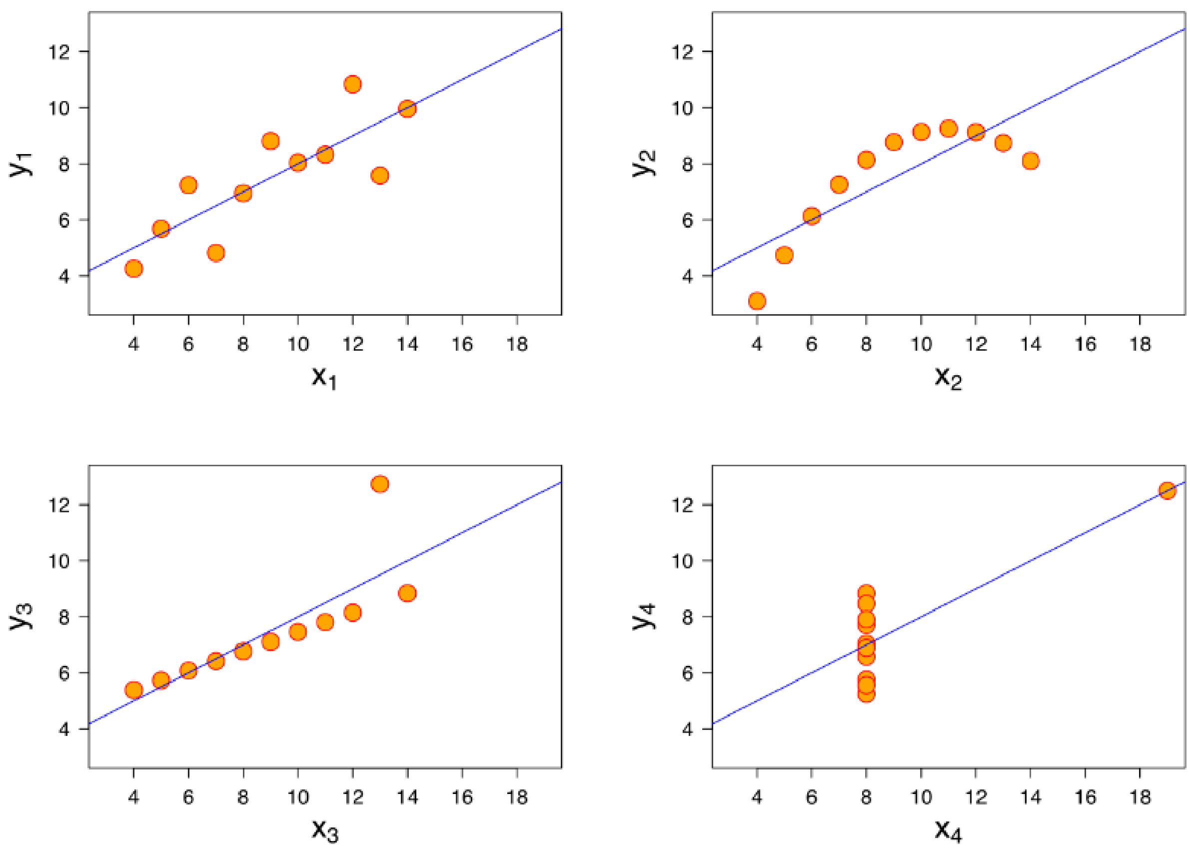
Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when 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

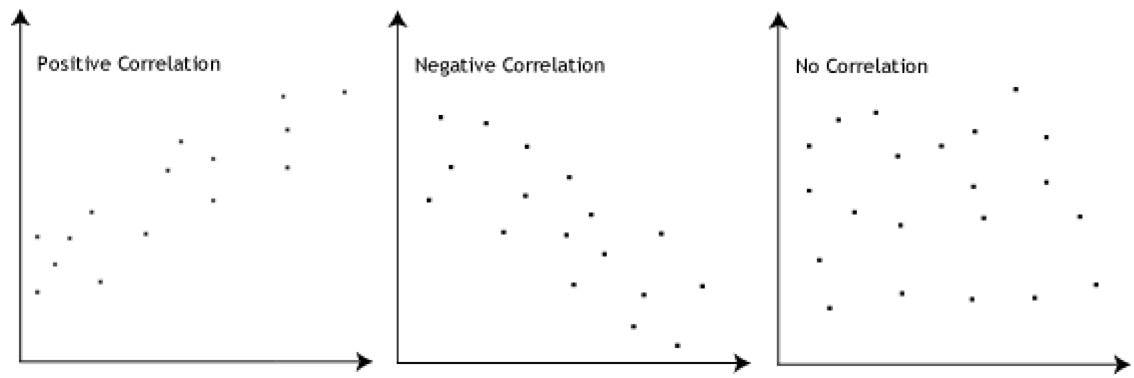
When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:



- Dataset I appear 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? (3 marks)**
Answer:
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:



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

Scaling:

The process of bringing all the independent features present in the given dataset to a same scale for making the data interpretation easier is called **SCALING.**

Need For Scaling:

In Multiple Linear Regression, if the variables are at different scale the interpretation of the coefficients of the model after we train the model becomes little difficult. For example, I have column A & column B in my dataset which stores integer values.
Column A values goes like this: 1,4,2,3,4,1,2,2 etc.…

Column B values goes like this: 4580,4400,5800,6878,2245,6565 etc.

Now, after we train the model if we get the values of coefficients like

ColumnA: 580, Column B: 0.684. with the values like this we can't really say that one column is significant than other. If we want to interpret the coefficients in multiple linear regression, it's important to have the values of the variables on a similar scale. Therefore, scaling is required.

There are two types of scaling techniques that are present.

1. Min-Max Scaling (Normalization)
2. Standardization

Min-Max Scaling:

In this technique, the data is converted to a value between 0 and 1.

If X is our data vector, then the Normalized version would be $(X-X_{min})/(X_{max}-X_{min})$.

Standardization:

In this technique, the data is converted in such a way that the mean of the data is 0 and the standard deviation(σ) of the data is 1.
If X is our data vector, the standardized version of X would be $(X-\mu)/\sigma$.

Normalization	Standardization
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.
Can alter the shape of the original distribution, particularly if outliers are present.	Does not change the shape of the original distribution; it only standardizes the scale.
More sensitive to outliers because the scale is heavily influenced by the extreme maximum and minimum values.	Less sensitive to outliers since it focuses on standard deviation.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**
Answer:
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.**
(3 marks)
Answer:
The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.
Use of Q-Q plot:
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence