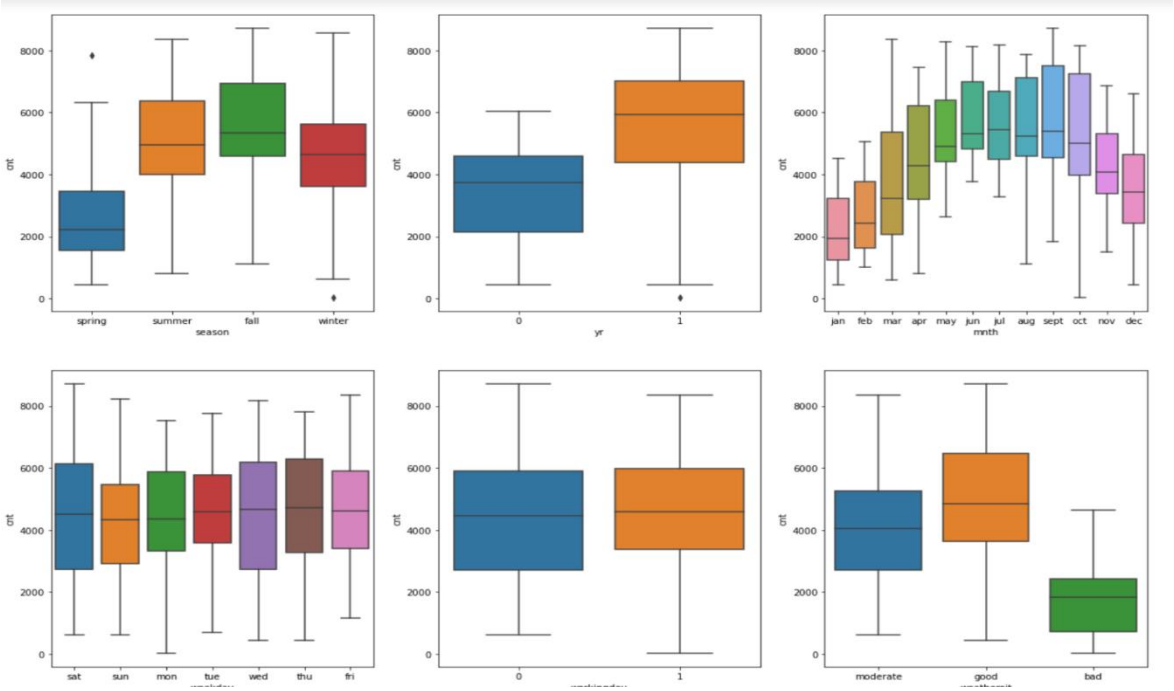
Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the data set, what could you infer about their effect on the dependent variable? (3 marks)**

Answer:

Here are a couple of categorical variables namely season,mnth,yr,weekday, working day and weathersit. These categorical variables have a major effect on the dependent variable 'cnt'. The below fig shows the correlation among the same

These variables are visualized using bar plot and Box plot both.



I have done analysis on categorical columns using the box plot 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 more 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 more number of booking 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 -

```
pd.get_dummies(data['Category'], drop_first=True)
```

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.

Benefits:

1. Avoids Multicollinearity: By creating k−1 dummy variables instead of k, we prevent perfect multicollinearity among the dummy variables.
2. Simplifies Interpretation: The coefficients of the dummy variables in a regression model will be interpreted relative to the reference category, making it easier to understand the effects of each category.

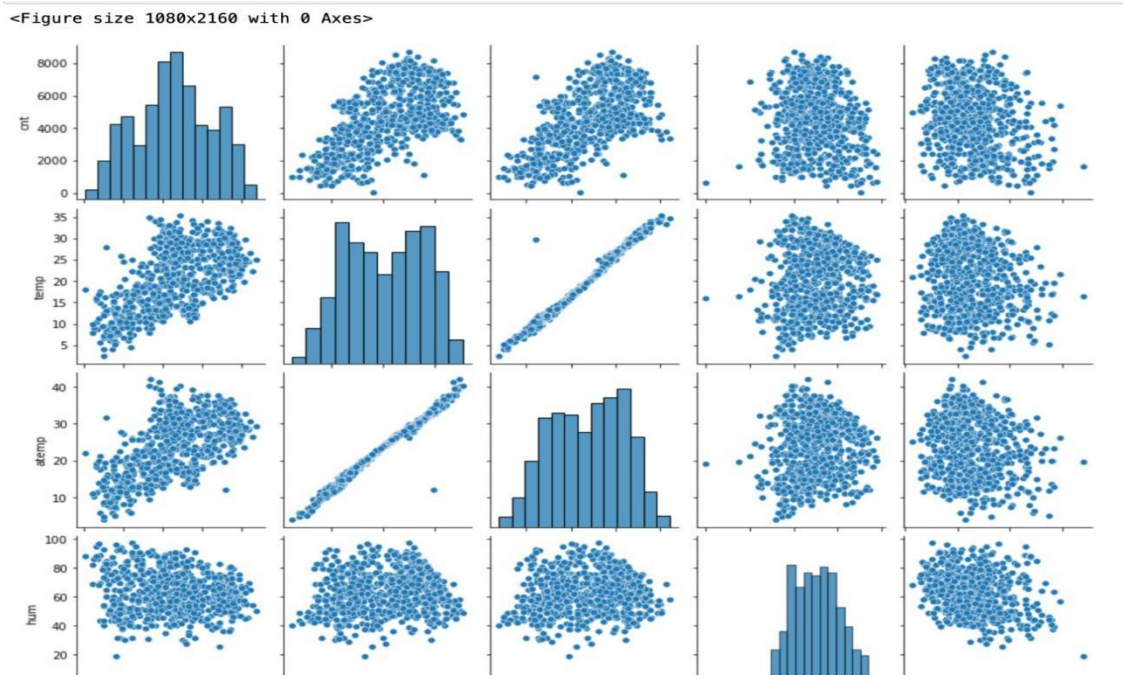
The intention behind the dummy variable is that for a categorical variable with 'n' levels, you create 'n-1' new columns each indicating whether that level exists or not using a zero or one.Hence drop_first=True is used so that the resultant can match up n-1 levels. Hence it reduces the correlation among the dummy variables.

Eg: If there are 3 levels, the drop_first will drop the first column.

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

Answer:

The **'**temp' and 'atemp' variables have highest correlation when compared to the rest with target variable as 'cnt'.



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 -

   A. Normality of error terms : Error terms should be normally distributed
   B. Multicollinearity check : There should be insignificant multicollinearity among variables.
   C. Linear relationship validation : Linearity should be visible among variables
   D. Homoscedasticity : There should be no visible pattern in residual values.
   E. 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 –

   - Temperature
   - Season : winter
   - Year : sep

General Subjective Questions

1. **Explain the linear regression algorithm in detail.** **(4 marks)**

Answer:

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).

Linear regression is a fundamental statistical technique used in machine learning and data analysis to model the relationship between a dependent variable (also known as the target) and one or more independent variables (features). The objective is to determine the best-fitting line or hyperplane that describes this relationship.

Key Concepts

Dependent Variable (Y): The outcome or target variable that the model aims to predict.

Independent Variables (X): The input features used to predict the dependent variable

.Mathematically the relationship can be represented with the help of following equation −

1Here, Y is the dependent variable we are trying to predict.

X is the independent variable we are using to make predictions.

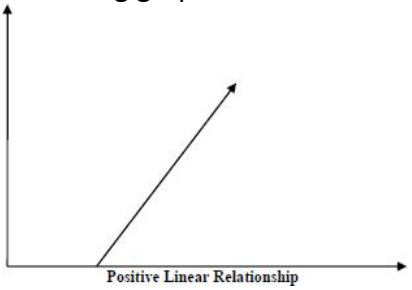
m is the slope of the regression line which represents the effect X

has on Y c is a constant, known as the Y-intercept. If X = 0, Y would

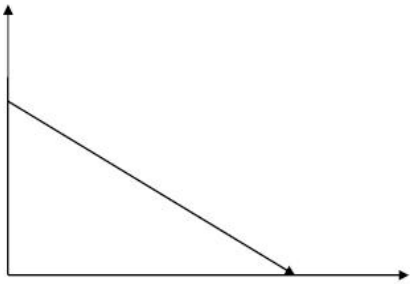
be equal to c.

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 −



- 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 auto-correlation 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:

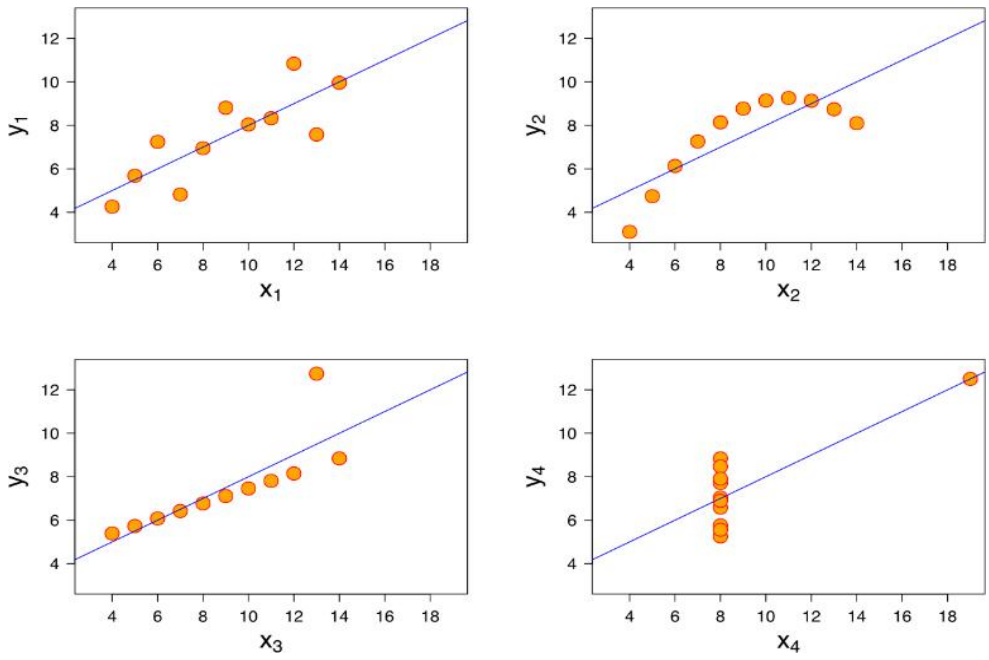
Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when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

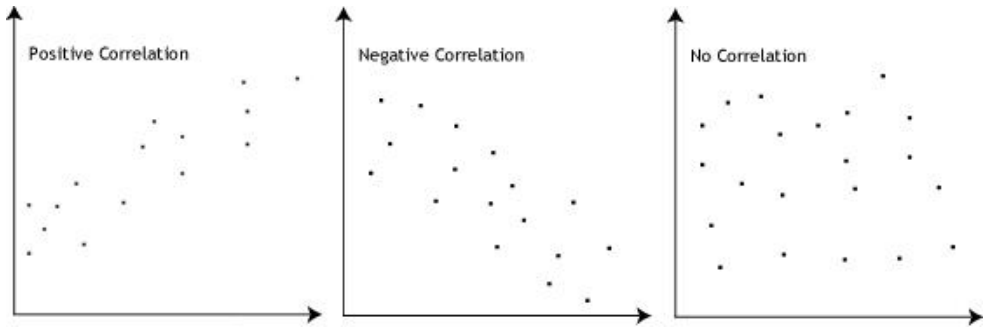
This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

3. **What is Pearson's R?** **(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:



In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables.

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

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?** **(3 marks)**

Answer:

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.

S.NO.	Normalized scaling	Standardized scaling
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

Difference between Normalizing Scaling and Standardize Scaling:

1. In normalized scaling minimum and maximum value of features being used whereas inStandardize scaling mean and standard deviation is used for scaling.
2. Normalized scaling is used when features are of different scales whereas standardized scaling is used to ensure zero mean and unit standard deviation.
3. Normalized scaling scales values between (0,1) or (-1,1) whereas standardized scaling is not having or is not bounded in a certain range.
4. Normalized scaling is affected by outliers whereas standardized scaling is not having any effect by outliers.
5. Normalized scaling is used when we don't know about the distribution whereasstandardized scaling is used when distribution is normal.
6. Normalized scaling is called as scaling normalization whereas standardized scaling is called as Z Score Normalization.

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.

VIF(VarianceInflationFactor) basically helps explain the relationship of one independent variable with all the other independent variables. The formulation of VIF is given below:A VIF value of greater than 10 is definitely high, a VIF of greater than 5 should also notbe ignored and inspected appropriately.
A very high VIF value shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem 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 for the conclusion that the two data sets have come from populations with different distributions.

Importance of Q-Q plot:
When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.

QQ plot can also be used to determine whether or not two distributions are similar or not. If they are quite similar you can expect the QQ plot to be more linear. The linearity assumption can best be tested with scatter plots. Secondly, the linear regression analysis requires all variables to be multivariate normal. This assumption can best be checked with a histogram or a Q-Q-Plot.

Importance of QQ Plot in Linear Regression :
In Linear Regression when we have a train and test dataset then we can create Q-Q plot by

which we can confirm that both the data train and test data set are from the population with the same distribution or not.

Advantages:

- It can be used with sample size also
- Many distributional aspects like shifts in location, shifts in scale, changes insymmetry, and the presence of outliers can all be detected from this plot

Q-Q plot use on two datasets to check

- If both datasets came from population with common distribution
- If both datasets have common location and common scale
- If both datasets have similar type of distribution shape
- If both datasets have tail behavior