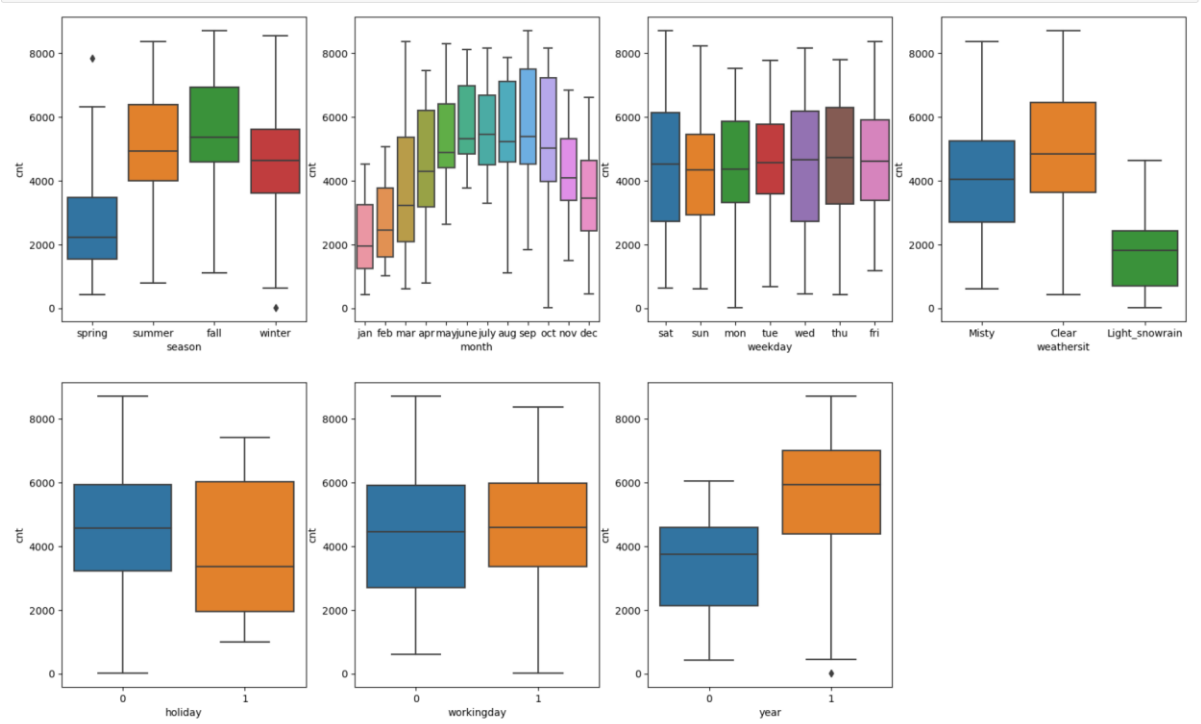
Assignment-based Subjective Questions

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

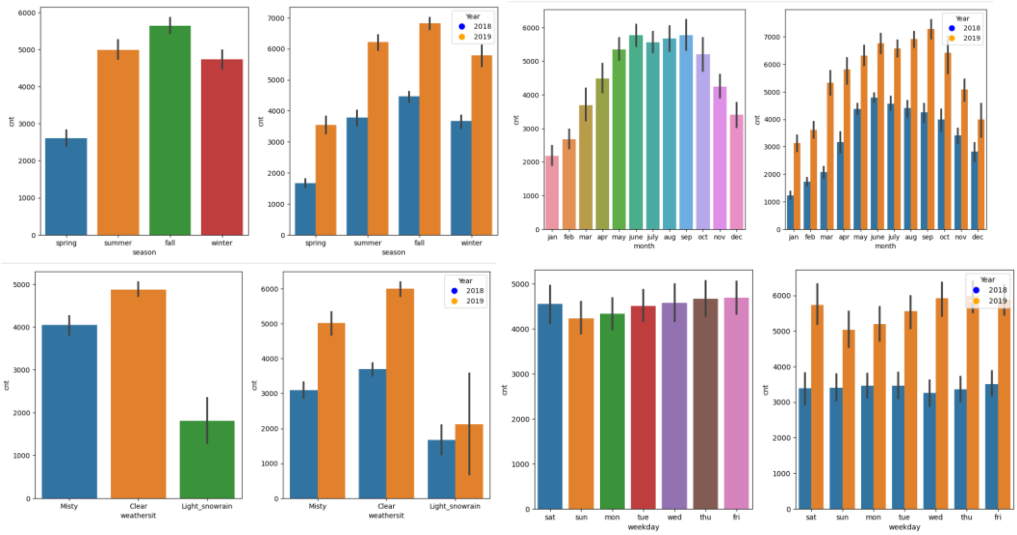
   Answer:

   Analysis on categorical columns using the boxplot and bar plot.

   Boxplot Analysis:



   Barplot Analysis



   Observation:

   a. **Season:** fall (3) season seems to have attracted more booking. And, in each season the booking count has increased drastically from 2018 to 2019.

   b. **Month:** Most of the bookings are done for month duration May to Oct**.** Booking Trend increase from Start of the Year till Mid of the Year and then decline towards the End of the Year.

c. **Weathersit:** Clear Weather (1-Clear, Few clouds, Partly cloudy, Partly cloudy) attracted more bookings.
d. **Weekday:** Though Thu, Fri, Sat seems to have more bookings, but there is no major visible differences in booking for weekdays. It shows very close trend.
e. **Holiday:** During Non-holidays bike hiring is maximum.
f. **Workingday:** Booking seemed to be almost equal either on working day or non-working day.
g. **Year:** There is more booking in year 2019 than year 2018

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

Answer:

a. '**drop_first=True**' is used to drop the first variable from the categorical variable while creating dummy variable from it.
b. It helps in reducing the extra column created whose meaning can be derived anyways from existing variables created. Hence it reduces the correlations created among dummy variables.
c. Let's say we have 3 types of values in Categorical column. We want to create dummy variable for that column. If one variable is not B and C, then It is obvious A. So, we do not need 3rd variable to identify the A. Hence in below case we can drop A as this can be identified as not B and not C

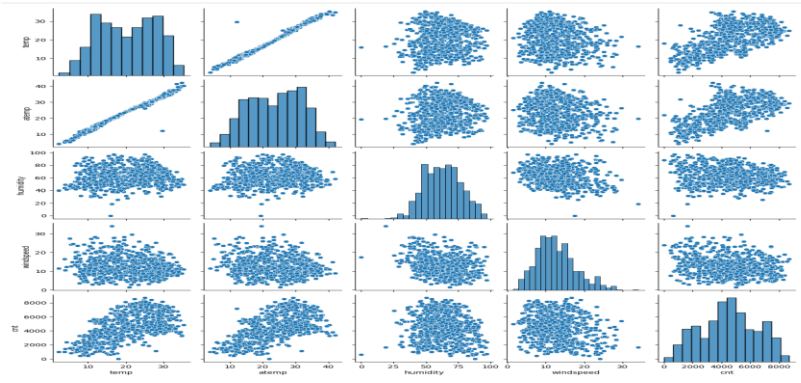
A	B	C
0	0	1
0	1	0
1	0	0

d. Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

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

Answer:

Looking at the pair-plot, 'temp' variable has the highest correlation with the target variable.

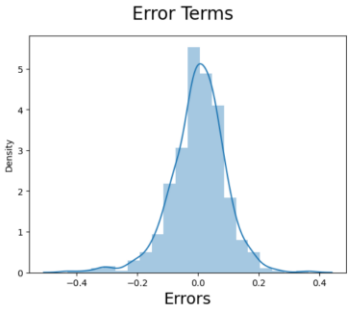


4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

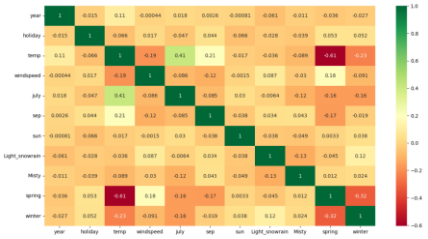
   Answer:

   Validation of the assumptions of Linear Regression Model is done as per below observations:

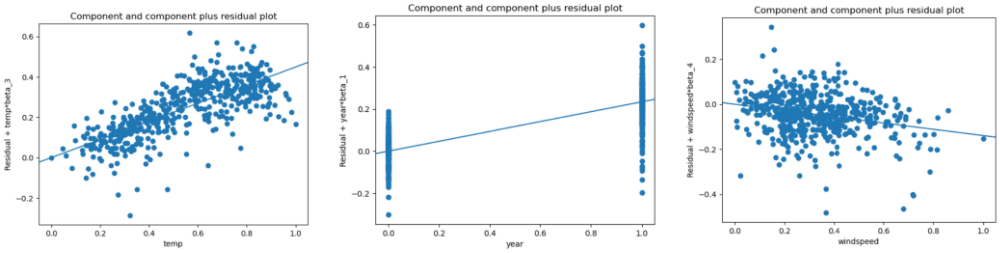
   a. Error terms are normally distributed with mean as 0



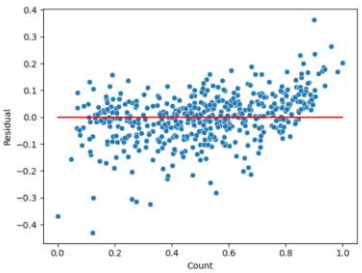
   b. There is insignificant multicollinearity among the predictor variables.



   c. Linearity is well preserved.



   d. Homoscedacity is well preserved.



   e. VIF is < 5 and p-values are 0

	Features	VIF
2	temp	4.67
3	windspeed	4.01
0	year	2.06
9	spring	1.66
8	Misty	1.52
10	winter	1.41
4	july	1.35
5	sep	1.20
6	sun	1.18
7	Light_snowrain	1.08
1	holiday	1.05

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.838
Model:	OLS	Adj. R-squared:	0.834
Method:	Least Squares	F-statistic:	233.6
Date:	Sat, 06 Jan 2024	Prob (F-statistic):	1.32e-188
Time:	23:25:23	Log-Likelihood:	502.54
No. Observations:	510	AIC:	-981.1
Df Residuals:	498	BIC:	-930.3
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2608	0.024	11.027	0.000	0.214	0.307
year	0.2342	0.008	28.644	0.000	0.218	0.250
holiday	-0.1057	0.026	-4.070	0.000	-0.157	-0.055
temp	0.4486	0.030	14.876	0.000	0.389	0.508
windspeed	-0.1393	0.025	-5.617	0.000	-0.188	-0.091
july	-0.0706	0.017	-4.096	0.000	-0.104	-0.037
sep	0.0561	0.016	3.582	0.000	0.025	0.087
sun	-0.0477	0.012	-4.101	0.000	-0.071	-0.025
Light_snowrain	-0.2907	0.025	-11.823	0.000	-0.339	-0.242
Misty	-0.0816	0.009	-9.370	0.000	-0.099	-0.065
spring	-0.1117	0.015	-7.434	0.000	-0.141	-0.082
winter	0.0471	0.012	3.835	0.000	0.023	0.071

Omnibus:	60.439	Durbin-Watson:	2.027
Prob(Omnibus):	0.000	Jarque-Bera (JB):	158.807
Skew:	-0.593	Prob(JB):	3.28e-35
Kurtosis:	5.463	Cond. No.	14.1

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

   Answer:

   Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes.

   a. temp
   b. year (yr as per the Data Dictionary)
   c. Light_snowrain (weathersit: 3 - Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds as per the Data Dictionary)

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

   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).

   Mathematically Equation:

$$Y = \beta_0 + \beta_1 X$$

   Y is the dependent variable we are trying to predict.

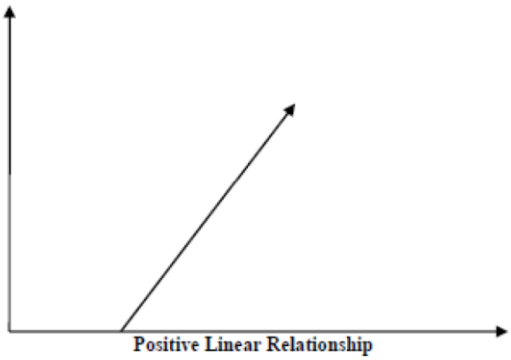
   X is the independent variable we are using to make predictions.

   β_1 is the slope of the regression line which represents the effect X has on Y

   β_0 is a constant, known as the Y-intercept. If X = 0, Y would be equal to β_0.

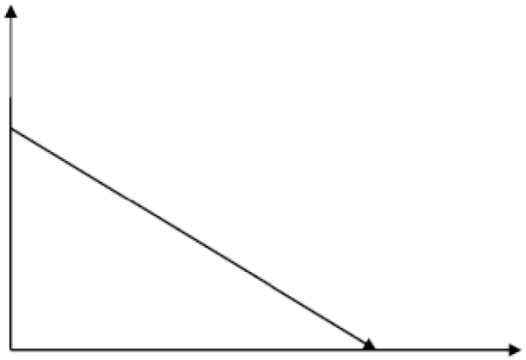
   The linear relationship can be positive or negative in nature.

   a. Positive Linear Relationship:
      - A linear relationship will be called positive if both independent and dependent variable increases.



Positive Linear Relationship

b. Negative Linear relationship:
   - A linear relationship will be called positive if independent increases and dependent variable decreases.



Negative Linear Relationship

Linear regression models can be classified into two types depending upon the number of independent variables:

a. Simple linear regression:
   - When the number of independent variables is 1
b. Multiple linear regression:
   - When the number of independent variables is more than 1.
   - In this case the above formulae is adjusted as below.

Extension of Simple Linear Regression to 'adds' more factors/effects

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

   - Please note the X variable denote more than one independent variable. It is utmost care that independent variables are not corelated else it leads to overfitting.

2. Explain the Anscombe's quartet in detail.

Answer:

a. Anscombe's Quartet is the modal example to demonstrate the importance of data visualization.
b. It signifies the importance of plotting data before analysing it with statistical properties.
c. It comprises of four dataset and each dataset consists of eleven (x, y) points.
d. The basic thing to analyse about these datasets is that they all share the same descriptive statistics (mean, variance, standard deviation etc) but different graphical representation. Each graph plot shows the different behaviour irrespective of statistical analysis.
e. The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyse according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets

Examples of Data Set:

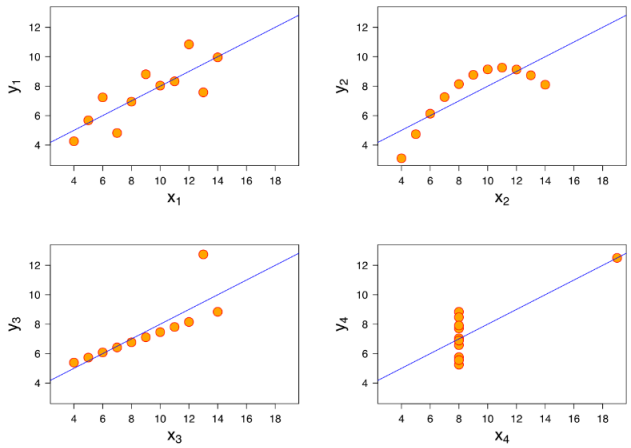
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	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

Mean and Standard Deviation:

Summary

Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817

Data Visualization:



Observation:

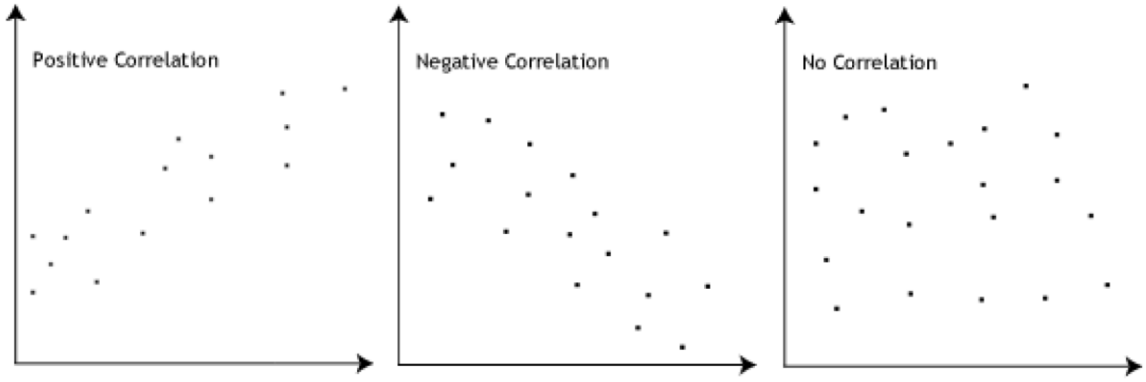
a. In the first one (top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
b. In the second one (top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
c. In the third one (bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
d. Finally, the fourth one (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

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

Answer:

a. Pearson's R is a numerical summary of the strength of the linear association between the variables.
b. 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.
c. The Pearson correlation coefficient, R, can take a range of values from +1 to -1.
d. A value of 0 indicates that there is no association between the two variables.
e. 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.
f. 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?**

Answer:

a. Scaling of a feature is the process of normalizing the range of features in a dataset.
b. It 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.
c. It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Why Scaling performed:

a. Real-world datasets often contain features that are varying in degrees of magnitude, range, and units. Therefore, for machine learning models to interpret these features on the same scale, we need to perform feature scaling.
b. 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.

c. It allows to have faster gradient descent.

Difference between normalized and standardized scaling:

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 really affected by outliers.	It is not bounded to a certain range. 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?

Answer:

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables.

a. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity.
b. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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.

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. 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.
b. 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.
c. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line.

d. 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.