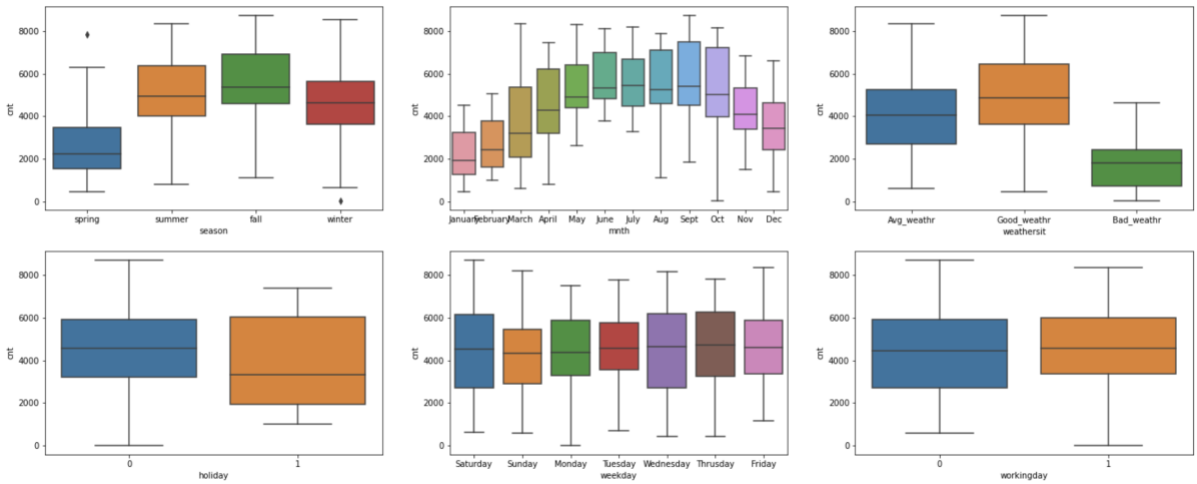
Assignment-based Subjective Questions

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

- **season**: Maximum bike booking were happening in season3[fall] . This was followed by season2[summer] & season4[winter]. A good predictor for the dependent variable. Spring season people don't hire much bikes
- **mnth**: More bookings happen from May – Sept . This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.
- **weathersit**: Bad weather like Light Snow, Light Rain with Scattered clouds " leads to significant decrease in bike hiring . More bike booking were happening during 'weathersit1 or good weather .It can be a good predictor for the dependent variable.
- **holiday**: During Non-holidays bike hiring is maximum .
- **weekday**: weekday variable shows very close trend. We could only understand once we built the model .
- **workingday**: Working day has more hiring vs non-working day



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

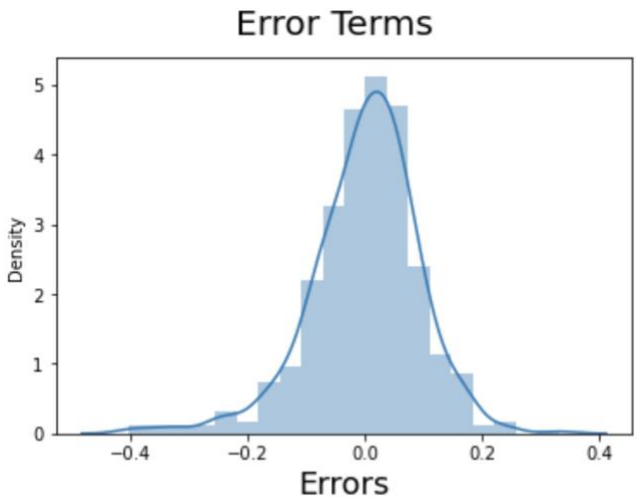
- We use drop_first=True to drop the first variable from the categorical variable while creating dummy variable from it . **It helps in reducing the extra column created whose meaning can be derived anyways from existing variables created.Hence it reduces the correlations created among dummy variables.**
- Ex : 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 furnished and semi_furnished, then we know automatically its unfurnished. So we do not need 3rd variable to identify the unfurnished.
- 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? (1 mark)

- Temp has highest correlation

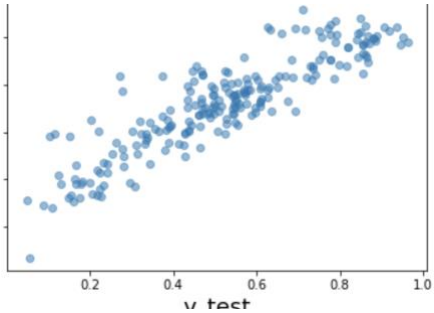
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

- We saw Error terms are normally distributed with mean zero (not X, Y)



- **Linear relationship between X and Y .**
  We plotted graphs and found linear relationship between X and Y .
  We found temp and atemp and cnt are linearly related in pair plots

- **Error terms have constant variance and are homoscedastic in nature .**



- **Error terms are independent of each other .**
- **After buildlng the final model we see from p-value and VIF value there is no multicollinearity . P-Values less than 0.05 and VIF less than 5 .**

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0442	0.017	2.601	0.010	0.011	0.078
yr	0.2332	0.008	27.645	0.000	0.217	0.250
temp	0.5527	0.020	27.295	0.000	0.513	0.592
windspeed	-0.1552	0.026	-6.041	0.000	-0.206	-0.105
season_summer	0.0894	0.011	8.460	0.000	0.069	0.110
season_winter	0.1281	0.011	12.051	0.000	0.107	0.149
mnth_Sept	0.0978	0.016	6.052	0.000	0.066	0.130
weathersit_Bad_weathr	-0.2019	0.026	-7.839	0.000	-0.252	-0.151
weathersit_Good_weathr	0.0767	0.009	8.553	0.000	0.059	0.094

	Features	VIF
1	temp	4.37
2	windspeed	3.16
7	weathersit_Good_weathr	2.66
0	yr	2.00
3	season_summer	1.55
4	season_winter	1.35
5	mnth_Sept	1.20
6	weathersit_Bad_weathr	1.11

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Top Three Predictor variables are :

- **temp [0.5527] .** An increase of temp will increase the bike request.
- **yr[0.2332] .** AN increase in years will increase the bike request . This might be due to increase in popularity .
- **weathersit_Bad_weathr [-0.2019]** Bad weather would lead to lead to decrease in bike request . This is referred as " Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds " .

General Subjective Questions

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

- Linear regression - basic forms of machine learning where we train a model to predict the behavior of dependent variable based on independent variables.
- Please note the dependent and independent variables which are on the x-axis and y-axis should be linearly correlated.
- An example, a promotion for a selling sim cards and expecting a certain number of count of customers to buy the sim .
    - Now what we do here is look the previous promotions and plot a chart to see whether there is an increment into the number of customers whenever you rate the promotions . With the help of the previous historical data we try to figure it out and try to estimate what will be the count or what will be the estimated count for the current promotion .
    - This allows to know how many numbers of stalls maybe you need or how many increase number of employees you need to serve the customer.

    - Here the idea is to estimate the future value based on the historical data by learning the behavior or patterns from the historical data.
- In some cases, the value will be linearly upward that means whenever X is increasing Y is also increasing or vice versa that means they have a correlation or there can be a linear downward relationship too which denotes a decrease
- Mathematically, we can write a linear regression equation as:

$$y = a + bx$$

Here, x and y are two variables on the regression line.

b = Slope of the line., a = y-intercept of the line, x = Independent variable from dataset , y = Dependent variable from dataset

- b slope also called as coefficient inform how the variable or feature will affect the value of y . It can be prositie to affect y positivitely or negative to affect negatively . In case its 0 , it does not affect .

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

    - Simple linear regression: When the number of independent variables is 1
    - Multiple linear regression: When the number of independent variables is more than 1 . In this case the above formulae is adjusted as below . Please note the X variable denote more than one independent variable . It is utmost care that independent variable are not corelated else it leads to overfitting .

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

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

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

- Anscombe's Quartet is the modal example to demonstrate the importance of data visualization. It signifies the importance of plotting data before analyzing it with statistical properties.
- It comprises of four data-set and each data-set consists of eleven (x,y) points.
- The basic thing to analyze about these data-sets is that they all share the same descriptive statistics(mean, variance, standard deviation etc) but different graphical representation.
- Each graph plot shows the different behavior irrespective of statistical analysis.
- Example to explain this :

The data-points :

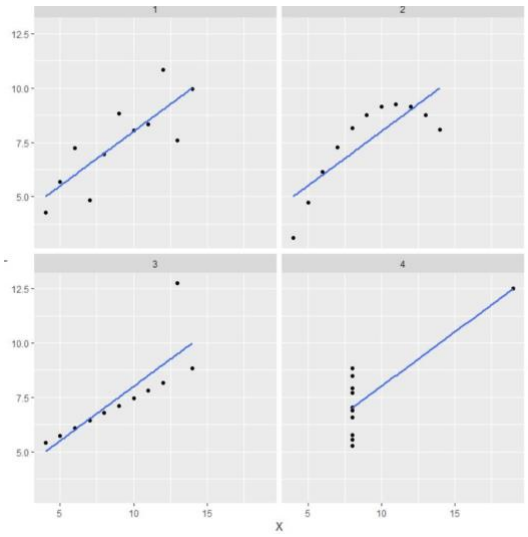
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0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

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

Explanation using graph :



Conclusion of this output:

- 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.
- 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.
- 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.
- Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

Application:

- The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

3. What is Pearson's R? (3 marks)

- Pearson's Correlation Coefficient is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.
- 1 means that they are highly correlated and 0 means no correlation. -1 means that there is a negative correlation.
- However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.
- The t-test is a correlation coefficient testing for any correlation between two values.

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

Scaling Meaning :

- Scaling of a feature is the process of normalizing the range of features in a dataset.
- 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.
- 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 its used :

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

Difference :

- There are two major methods to scale the variables, i.e. standardisation and MinMax scaling.
- Standardisation basically brings all of the data into a standard normal distribution with mean zero and standard deviation one.
- MinMax scaling, on the other hand, brings all of the data in the range of 0 and 1.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

- If there is perfect correlation, then VIF = infinity. This 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.
- 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. (3 marks)

- Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile.
    - For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution.
- A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.
- If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

Use and Importance :

- A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.

- In linear regression , Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential. This can allow to follow the assumptions of linear regression .