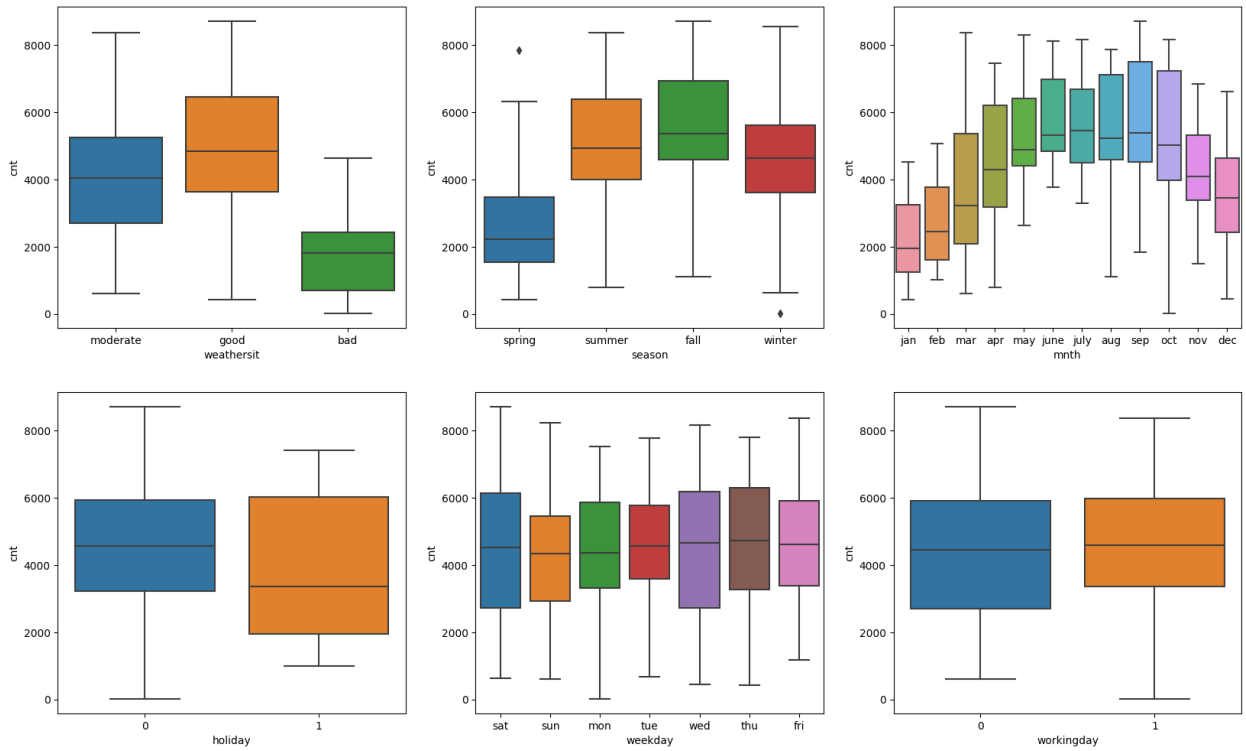
Assignment-based Subjective Questions

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

   After my analysis, we can witness the effect of categorical variables on the dependent variable



   - **weathersit**: The demand for shared bikes dropped drastically during bad weather, when it was high in good and moderate weather, indicates that weather could be a good predictor
   - **season**: The demand for shared bikes was highest in fall, and lowest in spring
   - **month**: The demand reached its peak during the middle of year, from June to September
   - **holiday**: The median of the demand is lower in holiday compared to non-holiday day
   - **weekday**: We can see little difference in demand for shared bikes between days in week
   - **workingday**: Booking trends seemed to be almost equal between working and non-working day

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

   drop_first=True is very important to use, it will remove the extra column when creating dummy variable, therefore it can reduce the correlations created among dummy variables.

   For example, for **season** column, we have four values: spring, summer, fall, winter. The dummy variables created will be four columns: spring, summer, fall, winter.

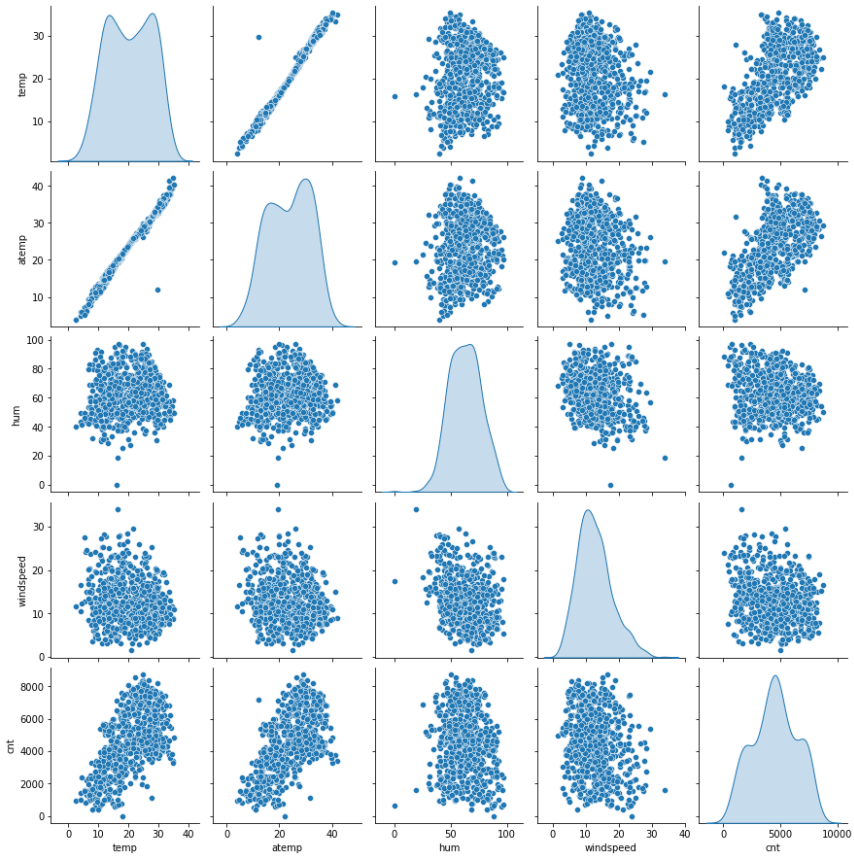
So, a row will be mapping to those values:

- **Spring**: spring 1 – summer 0 – fall 0 – winter 0
- **Summer:** spring 0 – summer 1 – fall 0 – winter 0
- **Fall**: spring 0 – summer 0 – fall 1 – winter 0
- **Winter:** spring 0 – summer 0 – fall 0 – winter 1

However, we can see the redundant case can be one of the above cases. If the season is **Spring**, so summer 0 – fall 0 – winter 0 will be enough, hence **drop_first=True** will remove the **spring** column, and the result will be:

- **Spring**: summer 0 – fall 0 – winter 0
- **Summer:** summer 1 – fall 0 – winter 0
- **Fall**: summer 0 – fall 1 – winter 0
- **Winter:** summer 0 – fall 0 – winter 1

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



According to the pair plot above, we can see that the **temp** and **atemp** show a strong correlation with the target variable, and we can see a linear regression pattern on the graph between temp and target variable, atemp and target variable.

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

   We can validate the assumptions of Linear Regression Model on the training set by:

   - **Normal Distribution of Error terms:** The error terms should be normally distributed

   - **Multicollinearity:** There should not be no significant multicollinearity between variables (The independent variables should not show high correlations)

   - **Linearity**: The relationship between independent variable and target value must be linear.

   - **Homoscedasticity:** The error should not be constant along the values of the dependent variables. We can check by drawing a scatterplot with the residuals, or using Breusch Pagan Test.

   - **No autocorrelation of errors:** Error terms should be independent. We can detect by Durbin Watson test, a value between 1.8 and 2.2 indicates no autocorrelation.

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

   After our final model, we can see the top 3 predictor variables:

   - **Temperature** (temp) has a coefficient value of 0.5038, which indicates that it is a strong predictor variable influence the shared bike demand

   - **Year** (yr) has a coefficient value of 0.2390, that means the demand for shared bike will increase as year increase

   - **Windspeed** (windspeed) has a coefficient value of -0.1777, indicates that strong wind will affect the booking of shared bikes

General Subjective Questions

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

   Linear regression is a basic form of machine learning in which we train a model to predict the behavior of the target variable base on some variables. Linear regression algorithm must show a linear relationship between a dependent (y) and one or more independent (x) variables. The linear regression model should provide a sloped straight line that represents the relationship between the variables.

   Mathematically, the linear regression equation can be written as:

   $$y = a + bx$$

   In which the formula of a and b can be:

   $$b = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

   $$a = \frac{n\sum y - b\sum x}{n}$$

   y – Dependent variable (Target variable)

   x – Independent variable (Predictor variable)

   a – Interceptor of the line

   b – Slope of the line

   The goal of the linear regression algorithm is to get the best value for a and b, or known as finding the best fit line, the best fit line should have the least error.

   Some possible use cases of linear regression

   - *Price prediction*: Predict the change in price of stock or product based on some predictors. For example: Housing price, …

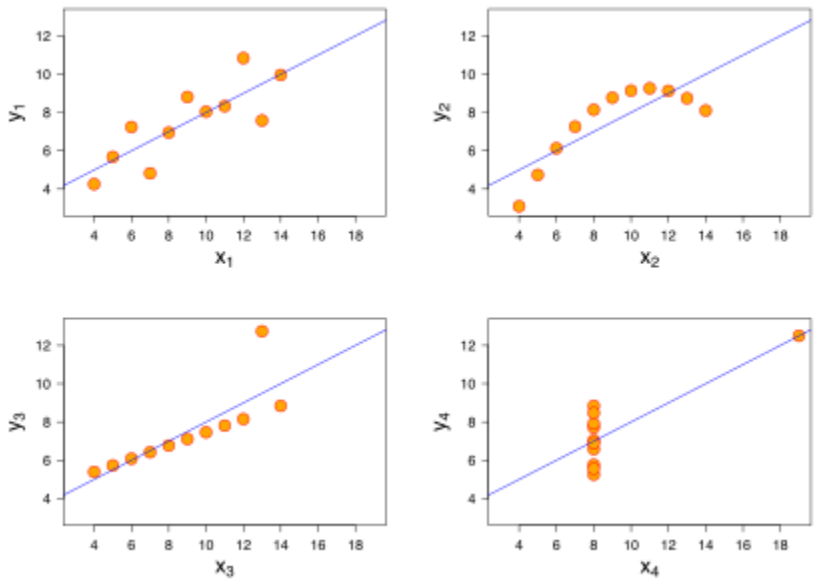
   - *Trends and sales target prediction*: Predict how many sales target industry may be able to achieve in the future.

2. **Explain the Anscombe's quartet in detail**

   Anscombe's quartet is a group of four datasets that are nearly identical in simple descriptive statistics, but they have very different distributions and appear very different when visualized. Let's look at Anscombe's Data:

Dataset / Observe	I		II		III		IV	
	x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
1	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
2	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
3	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
4	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
5	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
6	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
7	6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
8	4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
9	12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
10	7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
11	5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89
Statistics Summary								
N	11	11	11	11	11	11	11	11
Mean	9.0	7.50	9.0	7.50	9.0	7.5	7.0	7.5
SD	3.16	1.94	3.16	1.94	3.16	1.94	3.16	1.94
r	0.82		0.82		0.82		0.82	

We can see 4 datasets share the same statistics summary (mean, standard deviation, r). However, after visualization, the data will look like below:



We can see that:

- 1st dataset (top left) appears to be a simple linear relationship between X and Y
- 2nd dataset (top right) shows a visible relationship between X and Y, but it is not linear.
- 3rd dataset (bottom left), the relationship is linear, but should have different line because of there is one outlier which can strongly affect the correlation coefficient.
- 4th dataset (bottom right) has a high leverage point that can produce a high correlation coefficient, even though the other data points do not demonstrate any relationship between the variables.

In conclusion, **Anscombe's quartet** helps us understand the important of data visualization before building a machine learning model, as it can easily fool a regression algorithm.

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

In statistics, the **Pearson correlation coefficient**, also known as **Pearson's R,** the **Pearson product-moment correlation coefficient** (**PPMCC**), the **bivariate correlation**, or simply **the correlation coefficient,** is a measure of linear correlation between two sets of data. It is the most common way of measuring a linear correlation. It is a number between -1 and 1 indicates that it can measure the strength and the direction of the relationship between two variables.

Pearson's R	Interpretation	Example
$0 < r < 1$	When one variable changes, the other variable should change i**n the same direction** (both increase or decrease)	Height and weight. The higher a person, the heavier his/her weight.
$r = 0$	There is no relationship between two variables.	Fuel prices and adopting pets – These two variables demonstrate a weak / no correlation
$-1 < r < 0$	When one variable changes, the other variable should change i**n the opposite direction** (one increases, the other decreases and vice versa)	Housing price vs distance from city center – The greater distance from city center, the lower the housing price

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

Feature scaling is a technique to standardize the predictor variables in a fixed range. It should be performed during the data pre-processing step to handle highly varying magnitudes or

values or units. If it is not performed, machine learning algorithm tends to weigh greater values higher and consider small values as the lower values regardless of the unit of the values.

Example: In housing dataset, the number of bedrooms is ranged from 1 to 5, while the area feature has a wider range, can be up to 500 square meters. Machine learning algorithm can consider 500 to be greater than 1 or 5 but it is not true, so the prediction could be wrong.

Two most used techniques to perform Feature Scaling

- **Min-Max Normalization:** Rescale a feature or observation value with distribution between 0 and 1

$$X_{new} = \frac{X_i - min(X)}{\max(X) - min(X)}$$

- **Standardization:** Rescale a feature value so that it has distribution with the mean of 0 and variance of 1

$$X_{new} = \frac{X_i - \overline{X}}{\sigma_x}$$

Differences between Min-Max Normalization and Standardization:

Min-Max Normalization	Standardization
Use minimum and maximum value of features for scaling	Use mean and standard deviation for scaling
Used when features are of different scales	Used when we want to ensure zero mean and unit standard deviation
Scale value between $[0,1]$ or $[-1,1]$	Not bound to certain range
Affected by outliers	Much less affected by outliers

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

VIF (Variant Inflation Factor) helps explain the relationship of one independent variable with all the other independent variables. The formular of VIF is described as below:

$$VIF_i = \frac{1}{1 - R_i^2}$$

When the VIF is infinity, that's mean R_i^2 is nearly or equal to 1, which shows a perfect correlation between two independent variables. In order to solve this, we need to drop one of the variables from dataset to prevent them from causing this perfect multicollinearity.

Example of perfect multicollinearity:

- One predictor is a multiple of another

Suppose we have a dataset that has two columns **height in meter** and **height in centimeter**

Weight	Height in meter	Height in centimeter
400	1.3	130
460	1.4	140
470	1.5	150
475	1.2	120
490	0.9	90

We can see that **height in centimeter** is simply the product of **height in meter** and 100. This is the case of perfect multicollinearity.

- One predictor is a transformation of another

This case usually happens when we perform feature scaling and forget to remove the original column. For example: **areas** and **transform_areas** in housing dataset, they will produce perfect multicollinearity.

- The dummy variable trap

This scenario will occur when we want to use a categorical variable in a regression model, then we convert it without **drop_first=True.** This can lead to the happening of perfect multicollinearity.

For example, in the bike dataset, the **season** column. If we mistakenly omit **drop_first=True,** those dummy variables will be created.

Season		Spring	Summer	Fall	Winter
Spring		1	0	0	0
Summer	⟶	0	1	0	0
Fall		0	0	1	0
Winter		0	0	0	1

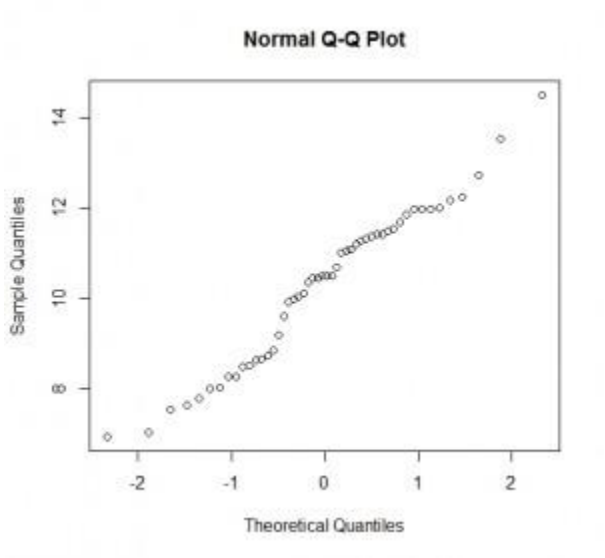
In this scenario, the column **Spring** is redundant because a combination of **summer 0 – fall 0 – winter 0** will be enough to indicate that the season is Spring. We can say that the variable **Spring** is a perfect linear combination of **Summer, Fall** and **Winter** variables. This is an example of perfect multicollinearity.

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

In statistics, **Q-Q plot (quantile-quantile plot)** is a probability plot, also a graphical technique for comparing two probability distributions by plotting these quantiles against each other. If

both sets of quantiles came from same distribution, we could see the points forming a line that's roughly straight.

This is an example of a Normal Q-Q plot when both sets of quantiles came from Normal distributions:



A 45-degree angle will be plotted on the Q-Q plot if the two datasets came from a common distribution, then the points will fall on that reference line.

In linear regression, Q-Q plot can help in the scenario when we have training and test dataset received separately and then we use Q-Q plot to ensure they are from populations with the same distribution. This can explain the importance of Q-Q plot in linear regression, it can detect many distributional aspects like shifts in location, shifts in scale, changes in symmetry, or the presence of outliers.