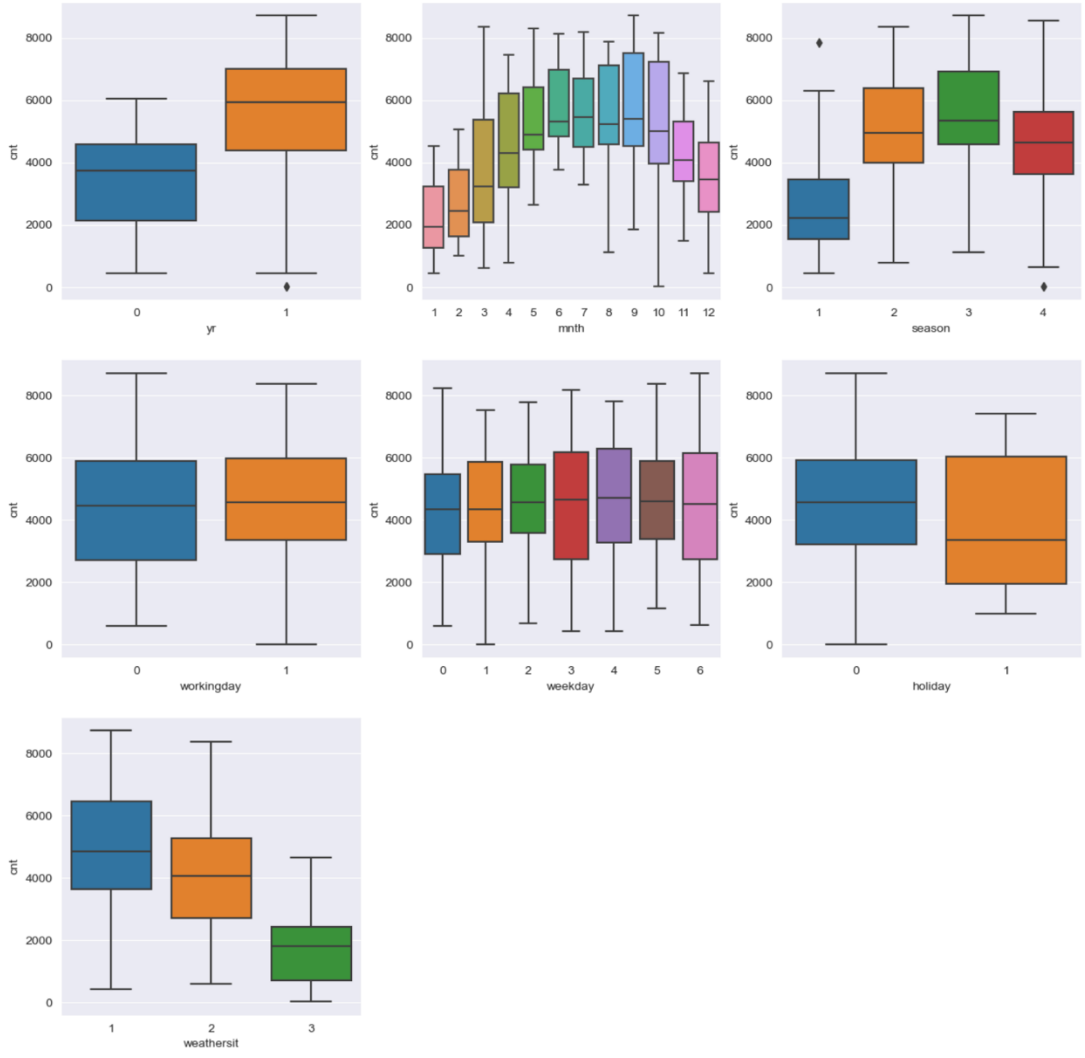
Assignment-based Subjective Questions

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

The categorical variables were - 'yr', 'mnth', 'season', 'workingday', 'weekday', 'holiday', 'weathersit'. The box plot visualisation looks like this (Note – the chart was too big to take screenshot so I had to open the notebook in browser instead of IntelliJ and hence has a light background compared to other screenshots) –

```
In [16]: # "cnt" is the target variable in this case.
         # Let's visualise the categorical variables using box plot to understand the data better

         #plt.rcParams.update({ 'font.size': 14 })

         categorical_vars = ['yr', 'mnth', 'season', 'workingday', 'weekday', 'holiday', 'weathersit']

         plt.figure(figsize=(15, 15))
         for idx, val in enumerate(categorical_vars):
             plt.subplot(3, 3, idx + 1)
             sns.boxplot(data=df1, x=val, y='cnt')
         plt.show()
```



These categorical values had the following effect on the target ("cnt") variable –

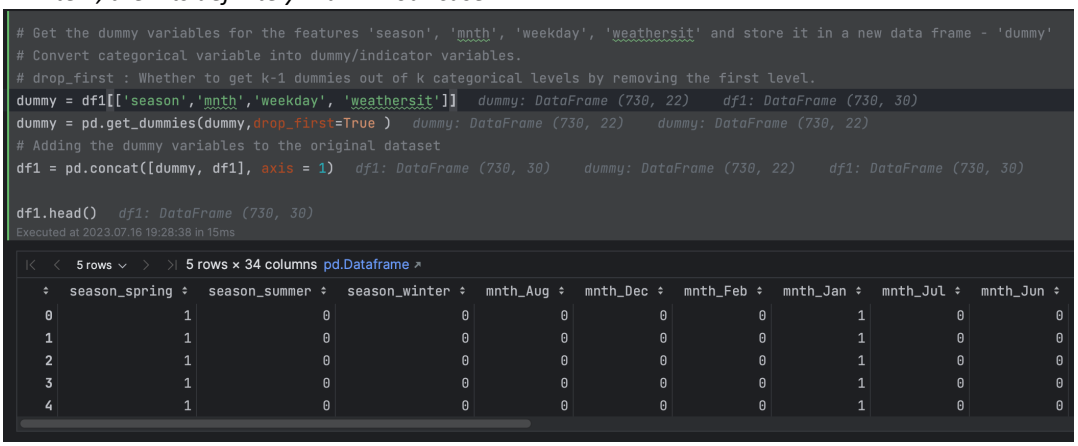
a. *Year 2019 has a higher count of bike rentals/demands as compared to the year 2018.*
b. *The bike rental/demand is highest during the month of June-September.*
    a. *Since its not clear which region the bike-sharing service is operational, we would assume its US (since its a US firm)*

b. *June to September month is a in between Spring and Fall Season i.e. warm months in US*
c. *The bike rentals/demand are pretty much the same throughout out the week. Weekdays or Weekends do not make a significant impact on the demand*
d. *The demand/rental count is between 4000 and 6000 during the week*
e. *The demand is more on working days than holidays (~4500 during working days vs ~3500 during holidays)*
f. *The demand is highest when Weather is "Clear, Few clouds, Partly cloudy, Partly cloudy", followed by when weather is "Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist"*

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

It's 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.

- *E.g. 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 It is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished*
- *From Bike Sharing case study, Season can be "Spring", "Summer", "Fall" , "Winter". By using drop_first, we automatically dropped "Fall" as if the season is not "Spring", "Summer" or "Winter", then its definitely "Fall" in our case.*



```
# Get the dummy variables for the features 'season', 'mnth', 'weekday', 'weathersit' and store it in a new data frame - 'dummy'
# Convert categorical variable into dummy/indicator variables.
# drop_first : Whether to get k-1 dummies out of k categorical levels by removing the first level.
dummy = df1[['season','mnth','weekday', 'weathersit']]   dummy: DataFrame (730, 22)    df1: DataFrame (730, 30)
dummy = pd.get_dummies(dummy,drop_first=True )   dummy: DataFrame (730, 22)    dummy: DataFrame (730, 22)
# Adding the dummy variables to the original dataset
df1 = pd.concat([dummy, df1], axis = 1)   df1: DataFrame (730, 30)    dummy: DataFrame (730, 22)    df1: DataFrame (730, 30)

df1.head()   df1: DataFrame (730, 30)
Executed at 2023.07.16 19:28:38 in 15ms
```

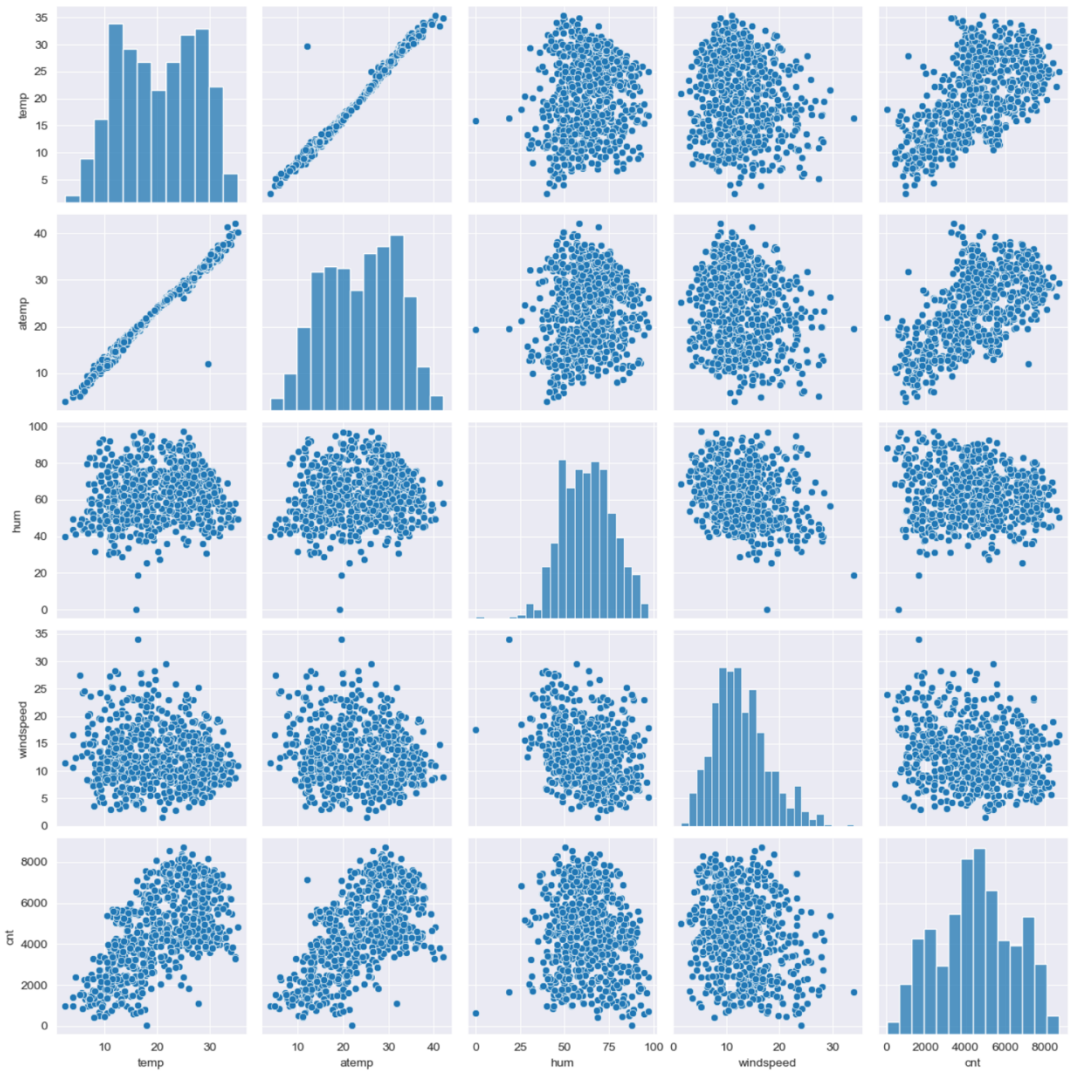
5 rows × 34 columns pd.Dataframe

	season_spring	season_summer	season_winter	mnth_Aug	mnth_Dec	mnth_Feb	mnth_Jan	mnth_Jul	mnth_Jun
0	1	0	0	0	0	0	1	0	0
1	1	0	0	0	0	0	1	0	0
2	1	0	0	0	0	0	1	0	0
3	1	0	0	0	0	0	1	0	0
4	1	0	0	0	0	0	1	0	0

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

Using the below pair plot it can be seen that , "temp" and "atemp" are the two numerical variables which are highly correlated with the target variable (cnt) .
<u>Note</u> – the chart was too big to take screenshot so I had to open the notebook in browser instead of IntelliJ and hence has a light background compared to other screenshots.

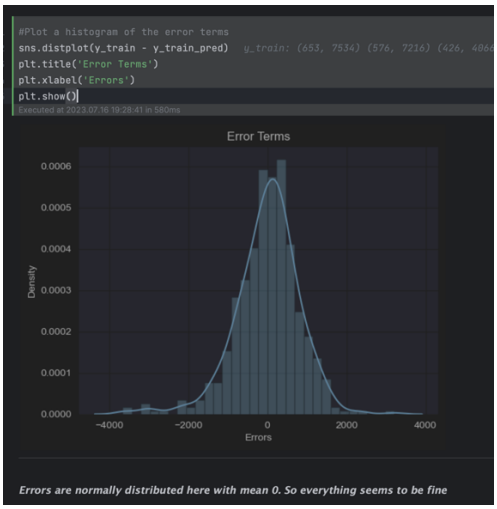
```
In [13]: # using pair-plot to understand numeric values

sns.pairplot(data=df1, vars=['temp','atemp','hum','windspeed','cnt'])
plt.show()
```



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

Following tests were done to validate the assumptions of linear regression, once the model was built on the training set –

a. *Linear regression needs the relationship between the independent and dependent variables to be linear. Pair plot was used to see if any of the variables are linearly related or not. The pair plot is present in the notebook and is also added as a screenshot for Ques. 3 above.*
b. *Residuals distribution should follow normal distribution and centred around 0 (mean = 0). This assumption about residuals was validated by plotting a dist-plot of residuals and it was shown that residuals are following normal distribution or not. The diagram below shows that the residuals are distributed about mean = 0.*

a. 

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

The top 3 features are –

a. ***Temperature*** → *Increase in temp, increases the demand of bike rentals.*
b. ***Year*** → *2019 fared lot better than 2018 i.e. the company should see better rentals/demand once we are out of pandemic.*
c. ***Summer Season*** → *There is more demand during summer season. Particularly month of August, the season of Spring, in which Aug falls, has a better demand than other months.*

General Subjective Questions

1) Explain the linear regression algorithm in detail.

Linear regression is a basic and commonly used type of predictive analysis. It is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features. There are two types of linear regression- Simple and Multiple. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula

$$y = c + b*x$$

where,
y = estimated dependent variable score
c = constant
b = regression coefficient, and
x = score on the independent variable

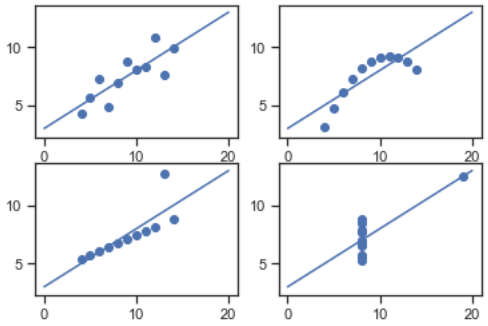
There are two types of linear regression- Simple and Multiple.

1. ***Simple linear regression(SLR)*** *is useful for finding relationship between two continuous variables.*
    a. *Formula →* $Y = \beta_0 + \beta_1 X + \varepsilon$
        i. *Y = Dependent Variable (Target Variable)*
        ii. *X = Independent Variable (predictor Variable)*
        iii. β_0 *= intercept of the line (Gives an additional degree of freedom)*
        iv. β_1 *= Linear regression coefficient (scale factor to each input value).*
        v. ε *= random error*
    b. *One is predictor or independent variable and other is response or dependent variable*
    c. *It looks for statistical relationship but not deterministic relationship. Relationship between two variables is said to be deterministic if one variable can be accurately expressed by the other e.g. Height in Inches can be used to accurately determine height in feet.*
    d. *Following assumptions exists in Linear Regression –*
        i. *There is a linear relationship between X and Y*
        ii. *Error terms are normally distributed with mean zero(not X, Y)*
        iii. *Error terms are independent of each other*
        iv. *Error terms have constant variance (homoscedasticity)*
2. ***Multiple linear regression(MLR)*** is useful when more than one independent variable is used to predict the value of a numerical dependent variable. E.g. Prediction of CO_2 emission based on engine size and number of cylinders in a car. *Key Points about MLR –*
    a. *Formula →* $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_n x_{in} + \epsilon$
        i. y_i *- dependent variable*
        ii. x_i *- explanatory variables*
        iii. β_0 *– y-intercept(constant)*
        iv. β_n *– slope coefficients for each explanatory variable*
        v. ϵ *- Residuals (model's error term)*
    b. *The dependent or target variable(Y) must be the continuous/real, but the predictor or independent variable may be of continuous or categorical form.*
    c. *Each feature variable must model the linear relationship with the dependent variable*
    d. *MLR tries to fit a regression line through a multidimensional space of data-points*

2) Explain the Anscombe's quartet in detail.

Anscombe's quartet is the modal example to demonstrate the importance of data visualization which was developed by the statistician ***Francis Anscombe*** *in 1973 to signify both the importance of plotting data before analysing it with statistical properties. It comprises of four data-set and each data-set*

consists of eleven (x, y) points that shares the same descriptive statistics(mean, variance, standard deviation etc) but different graphical representation. The statistical analysis of these four data-sets are pretty much similar. But when we plot these four data-sets across the x & y coordinate plane, we get the following results & each pictorial view represent the different behaviour.



Graphical Representation of Anscombe's Quartet

a. *Data-set I — consists of a set of (x,y) points that represent a linear relationship with some variance*
b. *Data-set 2 — shows a curve shape but doesn't show a linear relationship (might be quadratic?)*
c. *Data-set 3 — looks like a tight linear relationship between x and y, except for one large outlier*
d. *Data-set 4 — looks like the value of x remains constant, except for one outlier as well.*

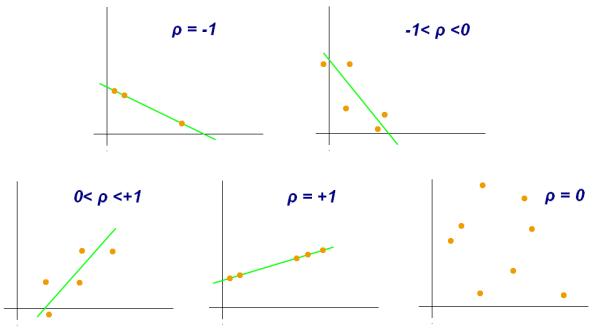
3) What is Pearson's R?

- *Correlation (Pearson) is also called "r" or "Pearson's R". It is a **correlation coefficient** commonly used in Linear Regression and is a measure of linear correlation between two sets of data.*
- *It is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1*
- *E.g. one would expect the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than 0, but less than 1*

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
y_i = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

- As can be seen from the graph below →
  - r = 1, means the data is perfectly linear with a positive slope
  - r = -1, means the data is perfectly linear with a negative slope r = 0 means there is no linear association



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

- *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.*
- *Feature scaling becomes necessary when dealing with datasets containing features that have different ranges, units of measurement, or orders of magnitude. In such cases, the variation in feature values can lead to biased model performance or difficulties during the learning process.*
- *There are several common techniques for feature scaling, including standardization, normalization, and min-max scaling. These methods adjust the feature values while preserving their relative relationships and distributions.*
- ***Normalization*** *is a data pre-processing technique used to adjust the values of features in a dataset to a common scale. This is done to facilitate data analysis and modeling, and to reduce the impact of different scales on the accuracy of machine learning models.*
  - ***Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.***
  - $X' = \frac{X - X_{min}}{X_{max} - X_{min}}$ → *X_{max} and X_{min} are the maximum and the minimum values of the feature*
- ***Standardization*** *is another scaling method where the values are centred around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero, and the resultant distribution has a unit standard deviation.*
  - $X' = \frac{X - \mu}{\sigma}$ → *μ is mean of the feature values & σ is the standard deviation of the feature values*

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

VIF(Variance Inflation Factor) gives how much the variance of the coefficient estimate is being inflated by collinearity. If there is perfect correlation, then VIF = INFINITY.

- It gives a basic quantitative idea about how much the feature variables are correlated with each other. It is an extremely important parameter to test our linear model.
- Formula → $VIF = \frac{1}{1-R^2}$ → R^2 is unadjusted coefficient of determination for regressing the i^{th} independent variable on remaining ones.
- A rule of thumb for interpreting the variance inflation factor:
  - VIF = 1 → not correlated
  - 1 <= VIF <= 5 → moderately correlated.
  - VIF > 5 → highly correlated.

6) What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

- It is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, I mean the fraction (or percent) of points below the given value.
- Advantages of the q-q plot are:
  - The sample sizes do not need to be equal.
  - Many distributional aspects can be simultaneously tested.
    - E.g. 1 → Shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.
    - E.g. 2 → If the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.
- The q-q plot is used to answer the following questions:
  - Do two data sets come from populations with a common distribution?
  - Do two data sets have common location and scale?
  - Do two data sets have similar distributional shapes?
  - Do two data sets have similar tail behavior?
- Importance –
  - 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
  - From the bike sharing assignment, qq-plot looks like this -