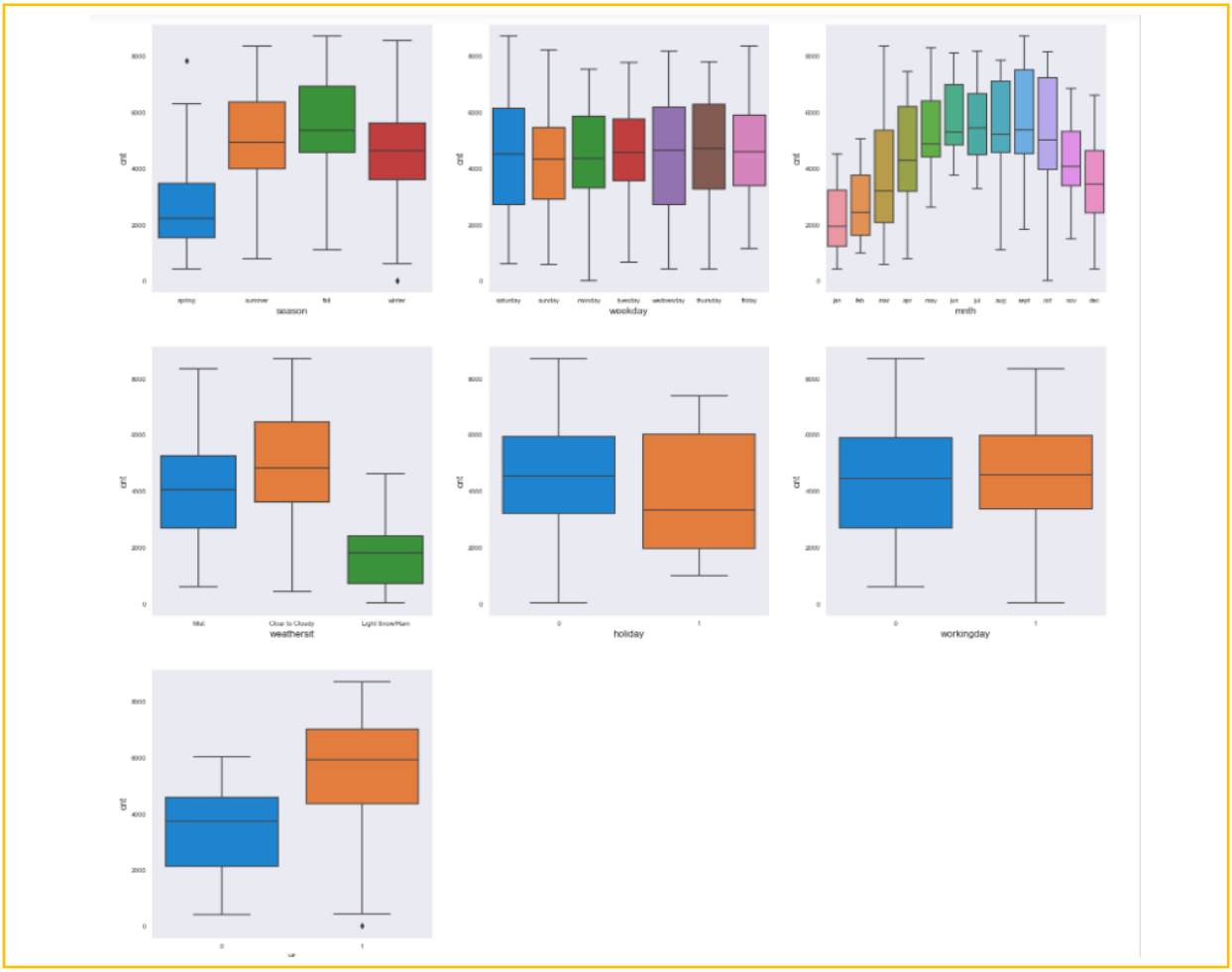
Sheetal

ML C48 Batch

Assignment based subjective questions

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



From Box plot we can infer the following:

- The Usage of bikes increase in **fall** followed by **summer**. There is significant reduction in usage of bike in **spring**.
- - The increase in usage during **september** followed by august and **october** confirms the season boxplot observation above.
- There is dependency on weather too as we see people tend to use bike during **Clear to Cloudy** followed by during **Misty** condition. They tend to use less bike during Light Rain/Snow. There is 0 usage of bike during Heavy Rain/Snow.
- Based on analysis of boxplot of mean usage on weekday we see they are used in same proportion.

- The mean of usage during **not** a holiday is higher.
- The usage of bikes during **working** and **non-working** day does not have a huge impact as bikes are used equally on most days.
- The usage of bikes in **2019** has increased from **2018** which is a good indicator.

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

   drop_first=True helps in reducing the extra column created during dummy variable creation. Hence it aids in reducing multicollinearity and cyclic dependency.

   For example: In code below notice for **season** dummies are created for season_summer, season_winter, season_spring. **Fall** is assumed when none of the other seasons are true. Hence reducing the need of dummy variable.

   Thus it reduces the number of dummy variables by 1, thus aiding in deletion of extra column while creating dummy variables.

```
# Subset all categorical variables
 bs_day_categorical=bs_day.select_dtypes(include=['object'])
# Let's drop the first column from resultant bs_day_categorical column's  using 'drop_first = True'
  bs_day_dummies = pd.get_dummies(bs_day_categorical, drop_first=True)
  bs_day_dummies.head()
```

	season_spring	season_summer	season_winter	weekday_monday	weekday_saturday	weekday_sunday	weekday_thursday	weekday_tuesday	weekday_wedn
0	1	0	0	0	1	0	0	0	
1	1	0	0	0	0	1	0	0	
2	1	0	0	1	0	0	0	0	
3	1	0	0	0	0	0	0	1	
4	1	0	0	0	0	0	0	0	

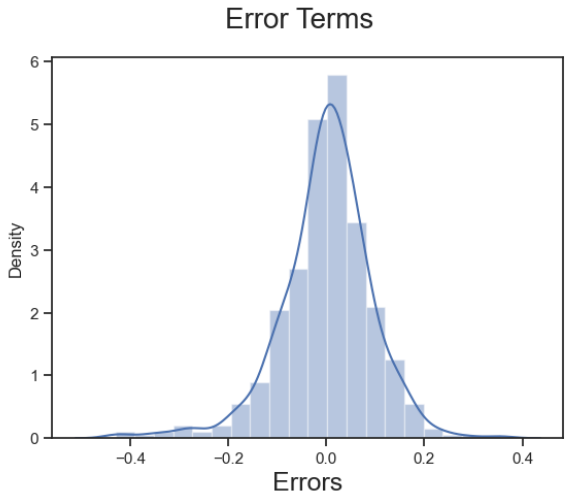
3. **Looking at the pair plot among the numerical variable, which one has the highest correlation with target variable?**

   The field's atemp and temp have high correlation with cnt variable. It is 0.63. The field registered and casual can be eliminated as their sum leads to cnt and thus are multicollinearity.

4. **How did you validate the assumptions of linear regression after building the model on training set**
   Firstly we developed the model using recursive feature elimination and calculated the VIF.
   Linear relationship between independent variable and target variable should be validated using pair plots. This was created and train model was developed using p|t| < 0.05 and VIF < 5
   Summary of find from Jupyter notebook:

   > All the VIF values and p-values seem to be in the permissible range now. Also the `Adjusted R-squared` value has dropped from `82.5%` with **10 variables** to just `82%` using **9 variables**. This model is explaining most of the variance without being too complex. **This was obtained after using RFE and removing variable based on high VIF value.**

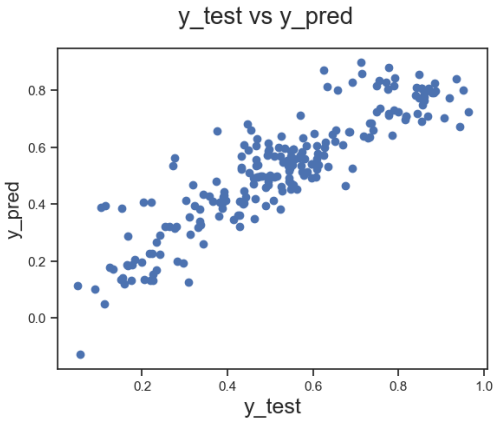
   Based on residual analysis:
   Error Terms should be normally distributed
   The normal distribution of the residual terms is a very crucial assumption when it comes to making inferences from a linear regression model. Hence, it is very important that you analyse these residual terms before you can move forward. The simplest method to check for the normality is to plot a histogram of the error terms and check whether the error terms are normal.
   As you can see the error terms has **mean at 0**



Now checking the correlation between actual **y** and predicted **y** we see that model is doing pretty good with predicted versus test data set.

Error Terms should be independent of each other

Error variance should be constant or shouldn't follow any pattern



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

   From the linear regression summary we see

   temp with co-efficient of 0.4689,

   weathersit_Light Snow/Rain with co-efficient of (-0.2809)

   yr with co-efficient of 0.2342 are the features which can have impact on bike sharing cnt.

$$cnt = 0.2342 \times yr + 0.4689 \times temp + (-0.1559) \times windspeed + (-0.0824) \times season_spring + 0.0388 \times season_summer + 0.0766 \times season_winter + (-0.0480) \times weekday_sunday + (-0.2809) \times weathersit_LightSnow/Rain + (-0.0770) \times weathersit_Mist$$

```
In [1050]: lr_rfe.summary()
```

Out[1050]:

OLS Regression Results

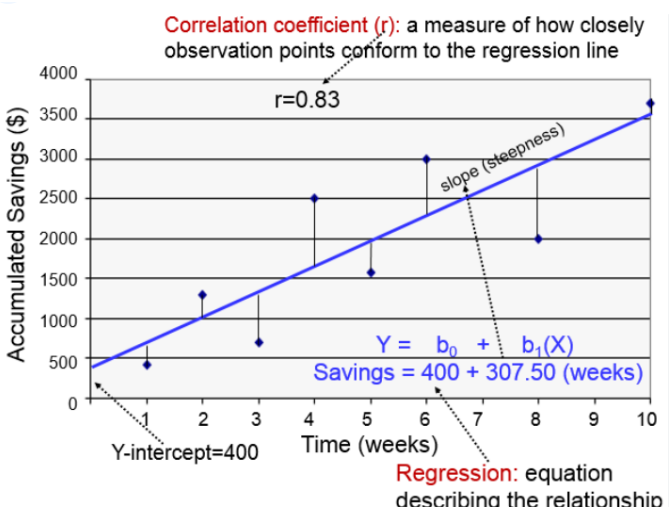
Dep. Variable:	cnt	**R-squared:**	0.825
Model:	OLS	**Adj. R-squared:**	0.821
Method:	Least Squares	**F-statistic:**	261.2
Date:	Tue, 14 Mar 2023	**Prob (F-statistic):**	1.08e-182
Time:	22:45:32	**Log-Likelihood:**	482.44
No. Observations:	510	**AIC:**	-944.9
Df Residuals:	500	**BIC:**	-902.5
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2257	0.030	7.580	0.000	0.167	0.284
yr	0.2342	0.008	27.602	0.000	0.218	0.251
temp	0.4689	0.034	13.861	0.000	0.402	0.535
windspeed	-0.1559	0.026	-6.027	0.000	-0.207	-0.105
season_spring	-0.0824	0.021	-3.968	0.000	-0.123	-0.042
season_summer	0.0388	0.014	2.784	0.006	0.011	0.066
season_winter	0.0766	0.017	4.574	0.000	0.044	0.110
weekday_sunday	-0.0480	0.012	-3.986	0.000	-0.072	-0.024
weathersit_Light Snow/Rain	-0.2809	0.025	-11.029	0.000	-0.331	-0.231
weathersit_Mist	-0.0770	0.009	-8.547	0.000	-0.095	-0.059

General Subjective Question

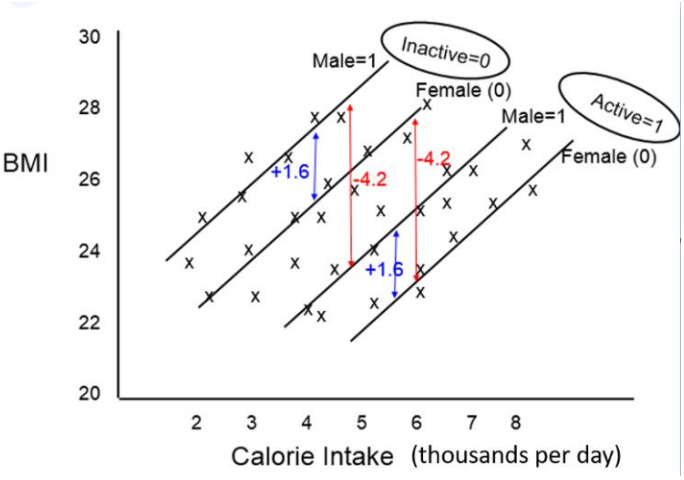
1. **Explain Linear Regression Algorithm in detail**

Linear regression is a supervised machine learning method that is used by the Train Using AutoML tool and finds a linear equation that best describes the correlation of the dependent (target variable) and independent variables (predictors). This is achieved by fitting a line to the data using least squares. The line **tries** to minimize the sum of the squares of the residuals. The residual is the distance between the line and the actual value of the explanatory variable. Finding the line of best fit is an iterative process.



The following is an example of a resulting linear regression equation:
Y=β0+β1X1+β2X2+...

In the example above, y is the dependent variable, and X1, X2, and so on, are the explanatory variables. The coefficients (β 1, β 2, and so on) explain the correlation of the explanatory variables with the dependent variable. The sign of the coefficients (+/-) designates whether the variable is positively or negatively correlated. β 0 is the intercept that indicates the value of the dependent variable assuming all explanatory variables are 0.

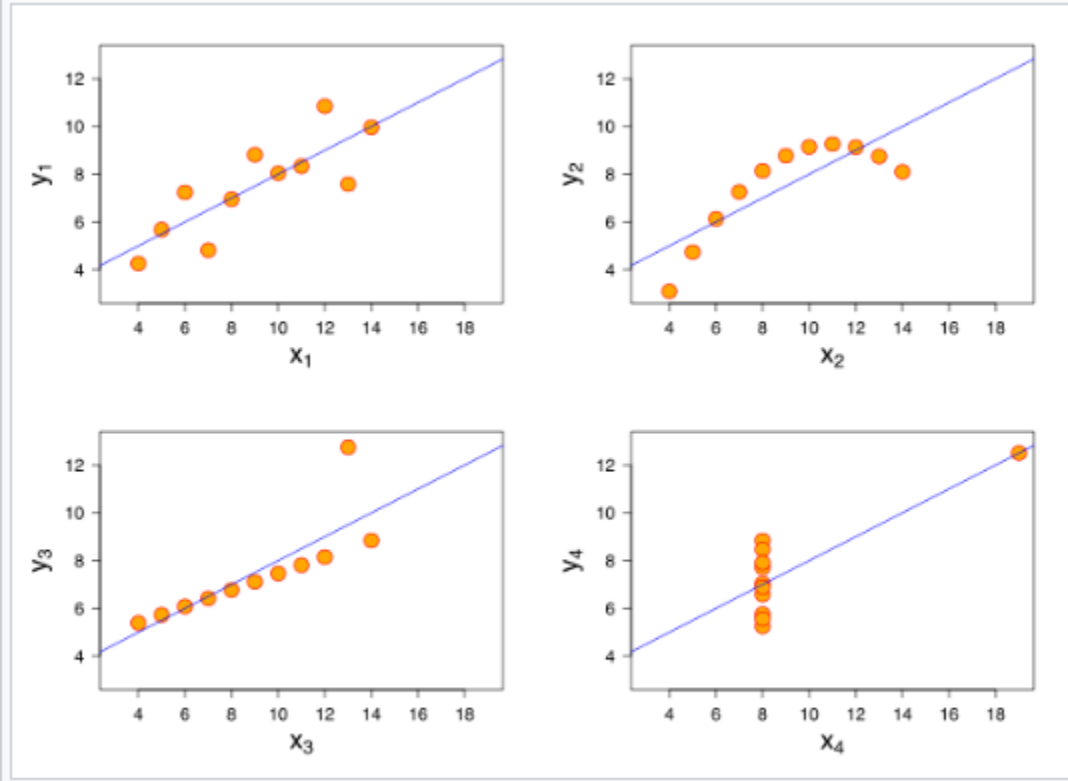


In case of multiple linear regression the resulting equation is
$Y=\beta 0+\beta 1X1+\beta 2X2+...+\beta pXp+\epsilon$

The model now fits a hyperplane instead of a line
Coefficients are still obtained by minimizing the sum of squared errors, the least squares criteria
For inference, the assumptions from simple linear regression still hold - zero-mean, independent and normally distributed error terms with constant variance.

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

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (*x*,*y*) points.



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on *x*.
- The second graph (top right); while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line (a robust regression would have been called for). The calculated

regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.

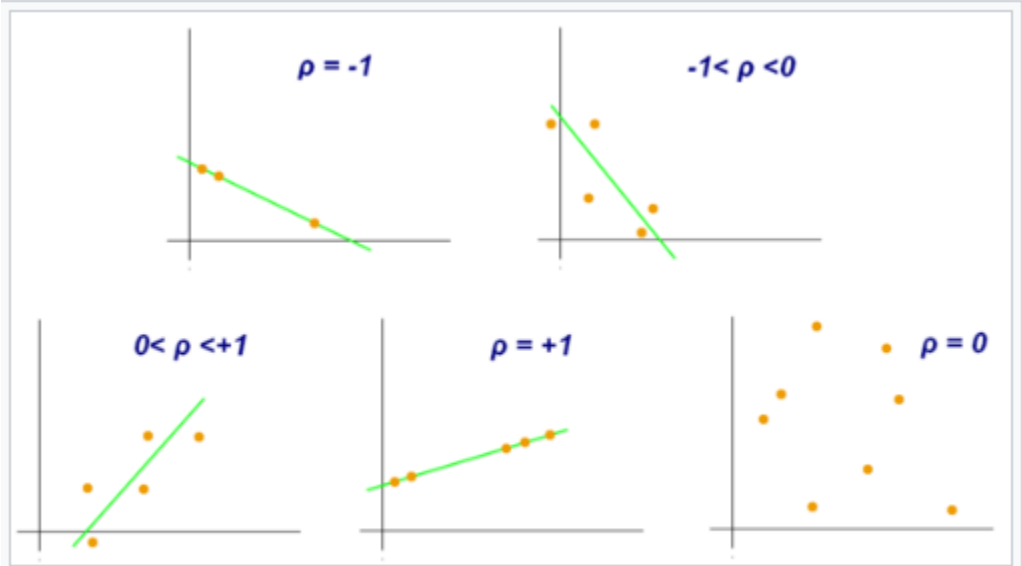
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

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**

In statistics, the **Pearson correlation coefficient** also known as **Pearson's *R***, the **Pearson product-moment correlation coefficient** (**PPMCC**), the **bivariate correlation**, or colloquially simply as **the correlation coefficient** — is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationships or correlations. As a simple example, 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 (as 1 would represent an unrealistically perfect correlation).

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier *product-moment* in the name.



Examples of scatter diagrams with different values of correlation coefficient (ρ)

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

Feature Scaling or scaling as it is termed is used when you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. So we need to scale features because of two reasons:

1. Ease of interpretation
2. Faster convergence for gradient descent methods you can scale the features using.

Two very popular method:

1. **Standardizing**: The variables are scaled in such a way that their mean is zero and standard deviation is one.

$$x' = \frac{x - \bar{x}}{\sigma}$$

2. **MinMax Scaling**: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

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-square, etc.

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

Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. 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 R^2 =1,
Which lead to 1/(1-R^2) infinity. Where R^2 is the coefficient of the determination
To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.
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).

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

Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential
The power of Q-Q plots lies in their ability to summarize any distribution visually.
QQ plots is very useful to determine

- If two populations are of the same distribution
- If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it's met using this.

- Skewness of distribution

In Q-Q plots, we plot the theoretical Quantile values with the sample Quantile values. Quantiles are obtained by sorting the data. It determines how many values in a distribution are above or below a certain limit.

If the datasets we are comparing are of the same type of distribution type, we would get a roughly straight line.