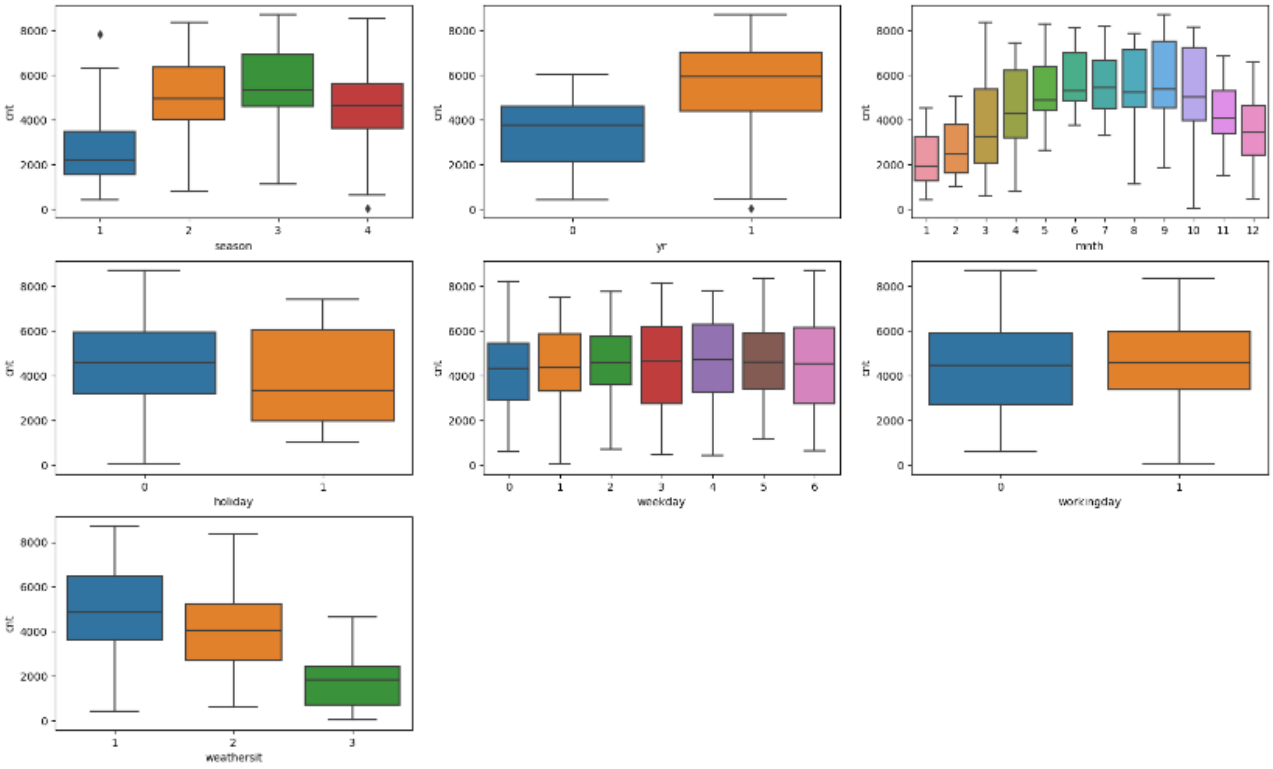
Assignment-based Subjective Questions

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

Answer)

There are several categorical variables in the dataset such as season, yr, mnth, holiday, weekday, workingday, weathersit.

These variables are visualized using Box plot with target variable 'cnt':



There are four types of weathersit viz., 1. Clear, 2. Mist, 3. Light Snow, 4. Heavy Rain

Weathersit – After looking at box plot we can observe that clear weather situations are good, while mist weather situations mean is lower than clear weathersit.

Post creating linear regression equation we can observe following:

Light Snow (3) has major negative impact with coefficient of -0.28. Which means whenever there is light snow, it has negative impact on bike sharing cnt.

Mist cloud (2) also has negative impact with coefficient of -0.07. This has lower impact than light snow, but still significant and should be considered.

Clear (1) does not have significant impact in predicting dependent variable i.e., cnt

Heavy Rain (4) was not observed in the dataset

Season – After looking at box plot mean is highest for Fall (3), followed by Summer (2), then Winter (4) and then Spring (1).

Post creating linear regression equation we can observe following:

Spring (1) has negative impact with coefficient of -0.11. Which means customers do not prefer bike sharing in spring season

Winter (4) has positive impact with coefficient of 0.04. Which means customers prefer bike sharing more in winter season compared to others

Summer (2) is neutral and hence not taken in linear regression model

Fall (3) is neutral and hence not taken in linear regression model

yr – After looking at box plot mean is high for yr (1), followed by yr (0).

Post creating linear regression equation we can observe following:

yr has positive impact with coefficient of 0.23. Which means bike sharing has gained popularity from 2018 -> 2019.

mnth – After looking at box plot mean increases to mid year and then drops again.

Post creating linear regression equation we can observe following:

mnth does not have any significance in predicting cnt.

holiday – After looking at box plot mean is higher for 0 and lower for 1.

Post creating linear regression equation we can observe following:

Holiday has negative impact in predicting cnt with coefficient of -0.0892.

Workingday, weekday – not much insights coming out of box plot or regression.

Overall inference for categorical variables:

1. It is observed that popularity has increased year over year. 2018 -> 2019. Hence, once situation stabilizes, company can expect increase in bike sharing business
2. On holidays, there is less demand
3. Weathersit light snow has negative impact on business
4. Weathersit mistcloud has negative impact on business
5. Season spring has negative impact on business
6. Season winter has positive impact on business and this time can be used to promote business

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

Answer)

It is important to use Drop_first = True during dummy variable creation as n-1 variables can explain n variables. For instance, gender has two values male and female. When we create dummy variables, it will create two dummy variables male and female containing 0,1 as values. However, male 0 already indicates that the person is female. Hence, this will only increase complexity of the model and not be value adding. It will also increase processing of extra variable which is not optimal. Hence, it is advisable to use drop first = true during dummy variable creation.

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

Answer)

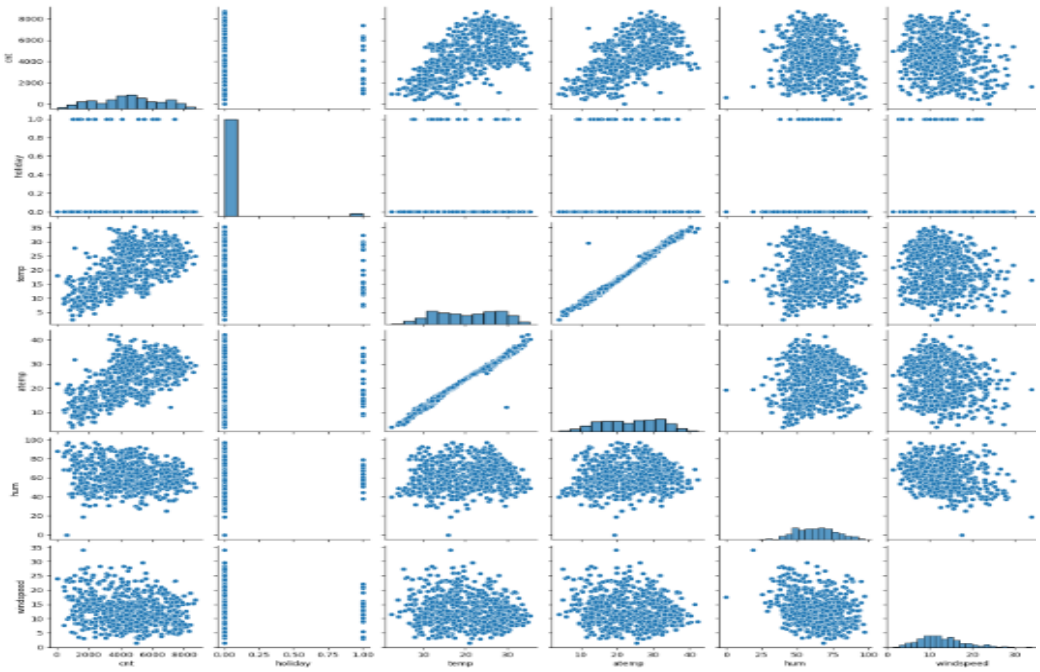
Atemp has the highest correlation with the target variable cnt.

Correlation value is **0.6464**.

While **temp** is also very close with correlation value of 0.6435.

This is explainable as temp vs atemp correlation value is 0.98. Which means both these numerical variables are similar in nature.

Diagram for reference:

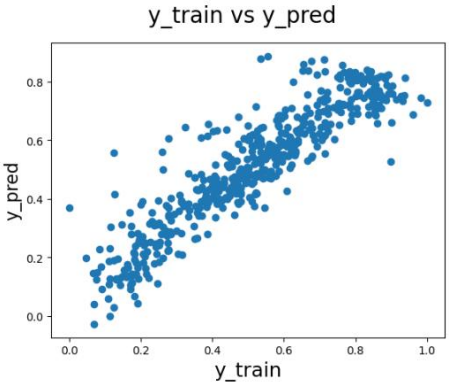


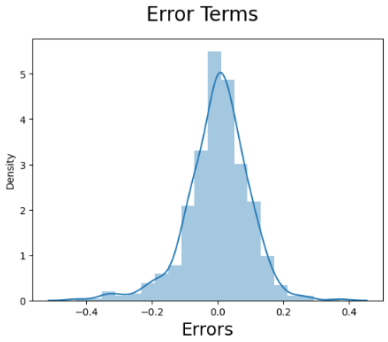
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

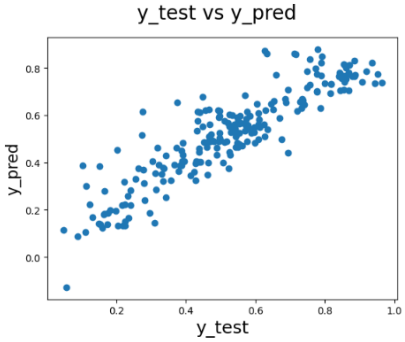
Answer)

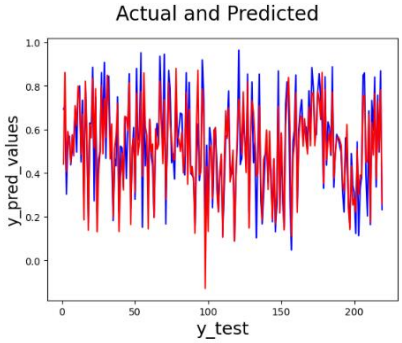
I validated the assumptions of linear regression after building the model on the training set by doing following steps:

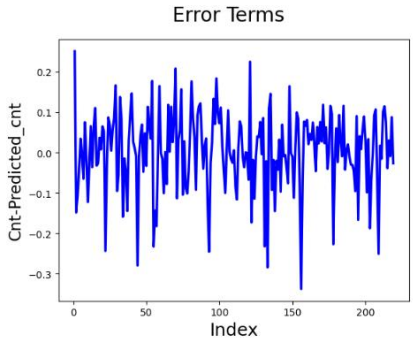
a. Validated on the test set and checked R square and adjusted R square value
b. Validated Error term (y_train – y_train_price) distribution to be normal distribution
c. Validated error term mean to be at 0
d. Plotted y_test vs y_pred to understand the spread
e. Checked R2_score of y_test, y_pred. It is 0.801. Showing high correlation.
f. Validated based on **linearity, No auto-correlation, Normality of error, Homoscedasticity, Multicollinearity (VIF value < 5), p value (less than 0.05) for significance**











Observations from screenshots attached:

1. Residual distribution should follow normal distribution and centred around 0 (mean =0). Errors are normally distributed – normality of error
2. Error mean is at 0
3. Y test vs y pred value plotted to check how close they are. Whether highs and lows are being explained properly.
4. Spread of y test vs y pred (also y train vs y pred)

5. No error term pattern observed

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

Answer)

Temp (0.42): This shows that favourable higher temperatures days have high demand of the shared bikes

Weathersit (3) Light Snow (-0.28): This shows that whenever there is snow, rain, thunderstorms, scattered cloud, customers do not prefer bike sharing. Which means unfavourable weather situation has negative impact on demand of the shared bikes

Yr (0.23): Positive 0.23 indicates that year on year, popularity and demand of the shared bike has increased

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

General Subjective Questions

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

Answer) Linear regression is a form of predictive modelling technique which tells us the relationship between the dependent (target variable) and independent variables (predictors). Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable. If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression. The linear regression model gives a sloped straight line describing the relationship within the variables.

Simple linear regression equation looks like: y= a0 + a1x
where,
y = dependent variable
x = independent variable
m = intercept of the line
b = linear regression constant

This helps in predicting change in dependent variable with every unit change in independent variable. Relationship between x and y is explained.

Assumptions of simple linear regression:

1. Linearity: the relationship between x and y is linear
2. Homoscedasticity: the variance of residual is same for any value of X
3. Independence: observations are independent of each other

4. Normality: for any fixed value of x, y is normally distributed

A regression line can be a Positive Linear Relationship or a Negative Linear Relationship. The goal of the linear regression algorithm is to get the best values for ‘a0’ and ‘a1’ to find the best fit line and the best fit line should have the least error. In Linear Regression, RFE or Mean Squared Error (MSE) or cost function is used, which helps to figure out the best possible values for ‘a0’ and ‘a1’, which provides the best fit line for the data points.

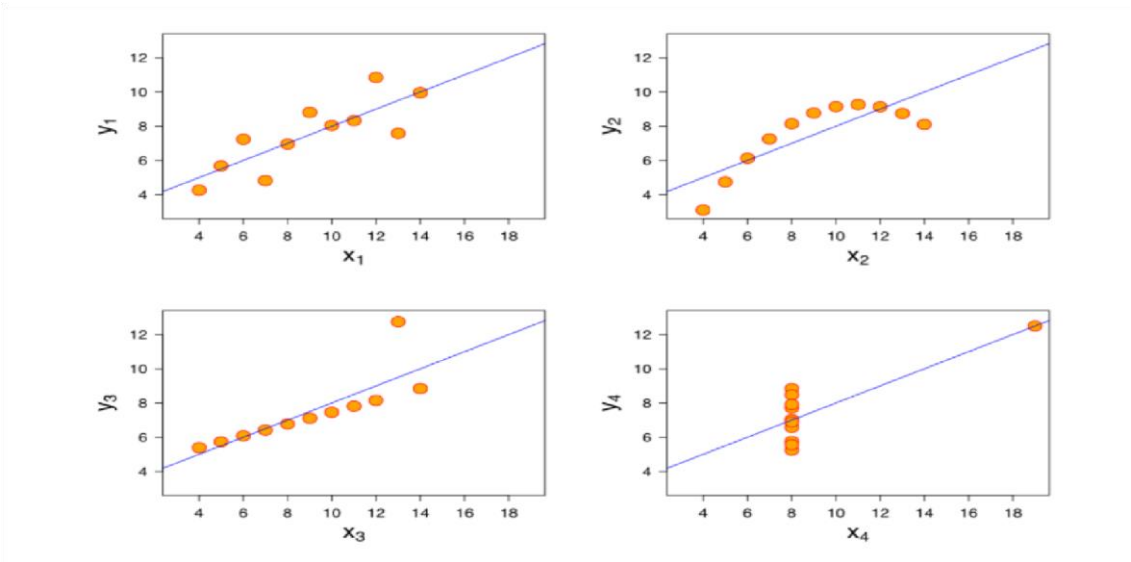
Equation for multiple linear equation looks like this:
y= a0 + a1x1 + a2x2 + a3x3

where a represents the weights, our model will try to learn.

Assumptions of multiple linear regression:
1. Linear relationship: there exist a linear relationship between predictor variable and response variable
2. No multicollinearity: none of the predicted variables are highly correlated to each other
3. Independence: the observations are independent
4. Homoscedasticity: the residuals have constant variance at every point in the linear model
5. Multivariate normality: the residuals of the model are normally distributed

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

Answer) Anscombe’s Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots. It was constructed to illustrate the importance of plotting the graphs before analysing and model building, and the effect of other observations on statistical properties. There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.



- 1st data set fits linear regression model as it seems to be linear relationship between X and y
- 2nd data set does not show a linear relationship between X and Y, which means it does not fit the linear regression model.
- 3rd data set shows some outliers present in the dataset which can't be handled by a linear regression model.
- 4th data set has a high leverage point means it produces a high correlation coeff.

Its conclusion is that regression algorithms can be fooled so, it's important to data visualization before building machine learning model.

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

Answer) In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

1. r = correlation coefficient
2. x_i = values of the x-variable in a sample
3. $\bar{x}$ = mean of the values of the x-variable
4. y_i = values of the y-variable in a sample
5. $\bar{y}$ = mean of the values of the y-variable

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

Answer) Scaling means you're transforming your data so that it fits within a specific scale. It is one type of data pre-processing step where we will fit data in specific scale and speed up the calculations in an algorithm. Collected data contains features varying in magnitudes, units and range. If scaling is not performed than algorithm tends to weigh high values magnitudes and ignore other parameters which will result in incorrect modelling. Difference between Normalizing Scaling and Standardize Scaling:

1. In normalized scaling minimum and maximum value of features being used whereas in Standardize scaling mean and standard deviation is used for scaling.
2. Normalized scaling is used when features are of different scales whereas standardized scaling is used to ensure zero mean and unit standard deviation.
3. Normalized scaling scales values between (0,1) or (-1,1) whereas standardized scaling is not having or is not bounded in a certain range.
4. Normalized scaling is affected by outliers whereas standardized scaling is not having any effect by outliers.
5. Normalized scaling is used when we don't know about the distribution whereas standardized scaling is used when distribution is normal.
6. Normalized scaling is called as scaling normalization whereas standardized scaling is called as Z Score Normalization.

Formula for normalized scaling / min max scaling:

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Formula for standardized scaling:

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

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

Answer) VIF(VarianceInflationFactor) basically helps explain the relationship of one independent variable with all the other independent variables. The formulation of VIF is given below: A VIF value of greater than 10 is definitely high, a VIF of greater than 5 should also not be ignored and inspected appropriately.

A very high VIF value 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. 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)

Answer) Q–Q plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data possibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution.

QQ plot can also be used to determine whether or not two distributions are similar or not. If they are quite similar you can expect the QQ plot to be more linear. The linearity assumption can best be tested with scatter plots. Secondly, the linear regression analysis requires all variables to be multivariate normal. This assumption can best be checked with a histogram or a Q-Q-Plot.

Importance of QQ Plot in Linear Regression :

In Linear Regression when we have a train and test dataset then we can create Q-Q plot by which we can confirm that both the data train and test data set are from the population with the same distribution or not. Advantages:

- It can be used with sample size also
- Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot Q-Q plot use on two datasets to check
- If both datasets came from population with common distribution
- If both datasets have common location and common scale
- If both datasets have similar type of distribution shape
- If both datasets have tail behavior

Advantages :

1. The sample size do not need to be equal
2. Many distributional aspects can be simultaneously tested

Q-Q plot is very useful to determine:

1. If two populations have same distribution
2. If residuals follow normal distribution. Having a normal error term is assumption of regression. We can verify if its met using this.
3. Skewness of distribution