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)

These are the categorical variables (year, month, season, weather situation) among the categorical variables year the most effect on the demand

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

If a categorical column has n values, we can represent that variable using n-1 dummy variables with out any loss of information pd. get_dummies returns n columns, if we set drop_first to True then it rremoves the first columns and returns n-1 columns

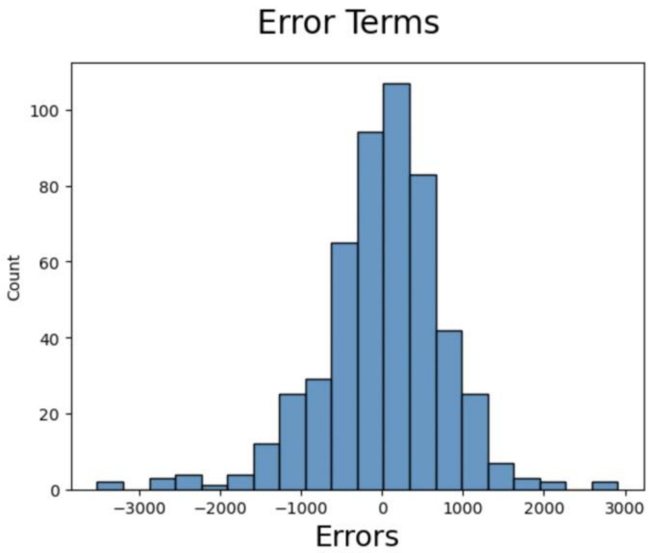
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Temp and atemp has the highest correlation with target variable cnt

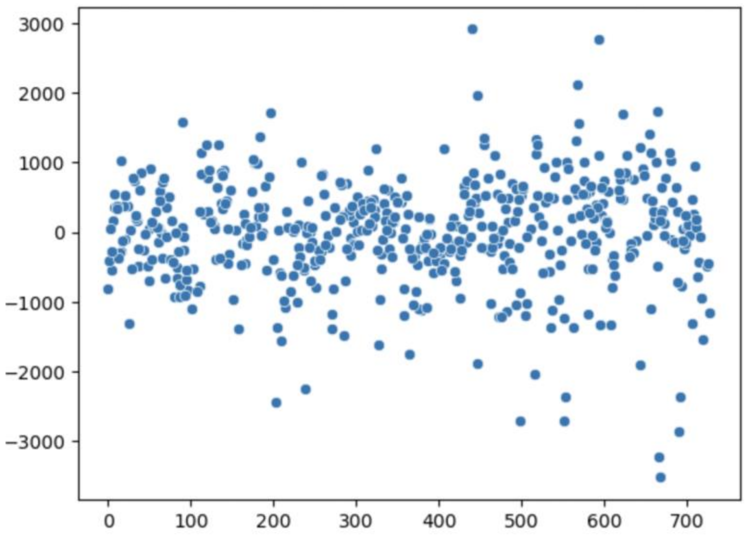
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

These are the linear regression assumptions.

1. there is some form of linear relationship exists among X and Y
   a. From the pair plot among the independent variables and target variable y we can see there is a linear relationship among the variables
2. from the residual analysis on the training data we can see the error(residual) follows normal distribution



3. scattar plot of residual shows that the errors are independent



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

- Atemp
- Yr
- Humidity
- windspeed

General Subjective Questions

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

linear regression is a statistical model which estimates the linear relationship between a dependent variable and one or more independent variables. If the there is only one independent variable it is called simple linear regression, if there are more than multiple independent variables it is called multiple linear regression.

It comes under supervised learning, where data is given with labels(output is known), and the model is built by learning from the data. The model is run in multiple iterations where in each iteration, the model learns from the mistakes done in the prediction until the model produces satisfactory results(the error between the actual and predicted values is minimal). The built model is used to predict the label for the data that is not seen.

Linear regression is used to predict a continuous variable, if the target is a categorical variable logistic regression is used.

Simple linear regression:

Lets say if we have a single independent variable (area of the house) and dependent variable (cost of the house), and given a set of data with area of houses and the corresponding cost of the house, simple linear regression can be used to predict the values of a house given area of the house.

It is similar to fitting a straight line that passes through X-Y plane. It can be given like this,

Y = mx + b. -> house = m . area + b

Where b is intercept and m is slope of the line. What we need to find out is m and b. Once m and b are found, we can predict cost of any house given area.

The above equation can be rewritten into y = ß0 + ß1 . area

If cost of the house dependes not only one the area but no of bed rooms, locality, no of bath rooms then we need to use multiple linear regression and it can be written as

Y = ß0 + ß1 . area + ß1 . no of bedrooms + ß2 . no of bath rooms + ß3 . locality

And we need to find the ß coefficiants.

The way to find the the coefficiants are to minimize the difference called as residual between the actual value and the predicted value, and it is called as cost function and we need to minimize this cost function.

There are several approaches to minimize the cost function.

- Using linear equations and solving them, it is easy to solve the equations for small order but as the number of featues increases the complexity increases
- Using gradiant discent algorithm – it is closed form solution and is an iteration solution.
    - We start from a random point,
    - From that point, we look which way is minimum and we make small strides towards that point
    - We do this iteratively until we reach minimum
    - The rate at which we make small strides is called learning rate it control
    - The smaller the learning rate it takes long time to reach the minimum(converge to minimum), if the learning rate is big we will never converge to minimum.

2. Explain the Anscombe's quartet in detail

It stresses the importance to visualize the data and not rely on the descriptive statistics to understand the data. Anscombe's created 4 datasets with 11 points(x,y) which have the same descritptive statistics such as mean, std, variance and correlation yet when we visualize the data points on graph they seems entirely different. One data set has linear relationship, other data set has non linear relationship, third dataset has linear relationship but there is an outlier.

3. What is Pearson's R?

It is called Pearson's correlation coefficient (R), it is used to define the correlation between the two variables. It can take a value from [-1, 1],

- If the value is between (0 ,1], then we say the two variables are positively correlated meaning if one increases the other also increases
- If the value is 0 then we say the two variables are not correlated

- If the value is between [-1, 0] then we say they are negatively correlated meaning if one increases the other decreases

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

Generally we have many numerical variables with different values and different ranges for example salary can range take anything from 0 to 6 digit or more, and where as age can have range of 0-100, etc.. when we have these kind of variable/features exist in the data which can take different ranges of data, scaling is performed for these reasons

- in linear regression the coefficiants can be smaller for high range of variables and small for lower range features, in such scenario we will not be able to tell which feature is significant in prediction.
- The gradiant discent will not reach to minimum quickly

Scaling is a measure to bring all features with different ranges into fixed range so that

- The magnitude of coefficiants is relavnat and by magnitude we can tell which feature is more important
- The Gradient dissent coverges to minimum quickly

There are two types of normalization techiques used

1. Standardized scaling

   x – mean(x) / std(x)

2. Normalized scaling or min- max scaling which scales the data into [0, 1] range

   X – min(x) / max(x) – min (x)

Scaling is done for the numerical variables

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

- VIF is used to identify if there is any multi collinearity exists among the independent variables using which we try to predict the dependent variable, general rule of thumb is if VIF > 5 then that variable can be expressed as a combination of other independent variables i.e is this variable has a multicolinerariy with the combination of other independent variable and this variable can be removed
- If the VIF is infinite means the variables are highly correlated, the R2 is almost one. Formula to compute the VIF is

$$VIF_{i}=.\ 1/1-R_{i2}$$

As we can see if the Rsquare value is 1 then the denominator is 0 resulting in VIF being infinite.

- So, If there are highly correlated variables present in the data using which we try to predict a dependent variable, computing VIF among these highly correlated variables would result VIF being infinite, solution is to remove one of the variable
- In this assignment, these are highly correlated variables
  - Temp and atemp (0.99)
  - Weather situation clear and weather situation mist(-0.95)

- Computing VIF for these will result in VIF being negative

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

Quartile-Quartile plot is also known as Q-Q plot, as name suggests it is a plot against Quantile of one variable against the quantines of another variable. It is used to check whether the samples are taken from the same population or not and also to understand the distribution.