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?

Answer:

- Year: Total Demand increased from 2018 to 2019 since it is grown in the company
- Season:
    - In the Spring season demand is at its lowest.
    - In Summer and winter demand is good
    - In fall demand is at highest
- Month:
    - In January demand is at lowest
    - Demand increases from Feb to June
    - From Jun to Oct it remains at a peak
    - Demand decreases from Nov to Dec
- Weekday:
    - On Sunday, Monday there is less demand
    - Demand increases on Tuesday, Wednesday, and Saturday
    - On Thursday, Friday demand is at a peak
- Working days:
    - There is more demand for working days in comparison to non-working days
- Weathers
    - Most demand comes when the weather is Clear, with Few clouds, Partly cloudy, Partly cloudy.
    - Demand are moderate when weather is Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist.
    - Lowest demand when weather is Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds .
    - No demand when weather is Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog.

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

Answer:

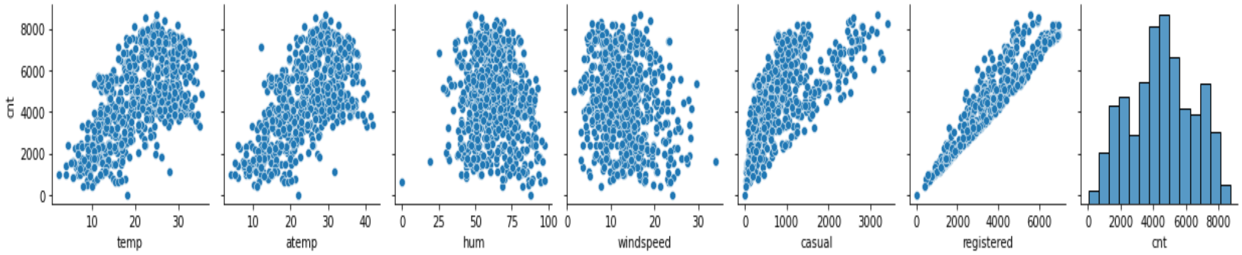
- By setting drop_first=True, the first level of the categorical variable is dropped and not included as a feature. This will remove the multi-collinearity problem, because the predictor variables will no longer be correlated with each other.
- Also it doesn’t affect model accuracy or any feature loss since from remaining features of one hot encoding model can drive values for removed column when all values of one hot encoding is 0.

Ex: column-> grade: A, B, A, C, B, C can be represented as

A	B	C	OR	B	C	We can see When B and C in second table is 0 it will represent that A is 1 even if A is not present in data.
1	0	0		0	0	
0	1	0		1	0	
1	0	0		0	0	
0	0	1		0	1	
0	1	0		1	0	
0	0	1		0	1	

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

Answer:

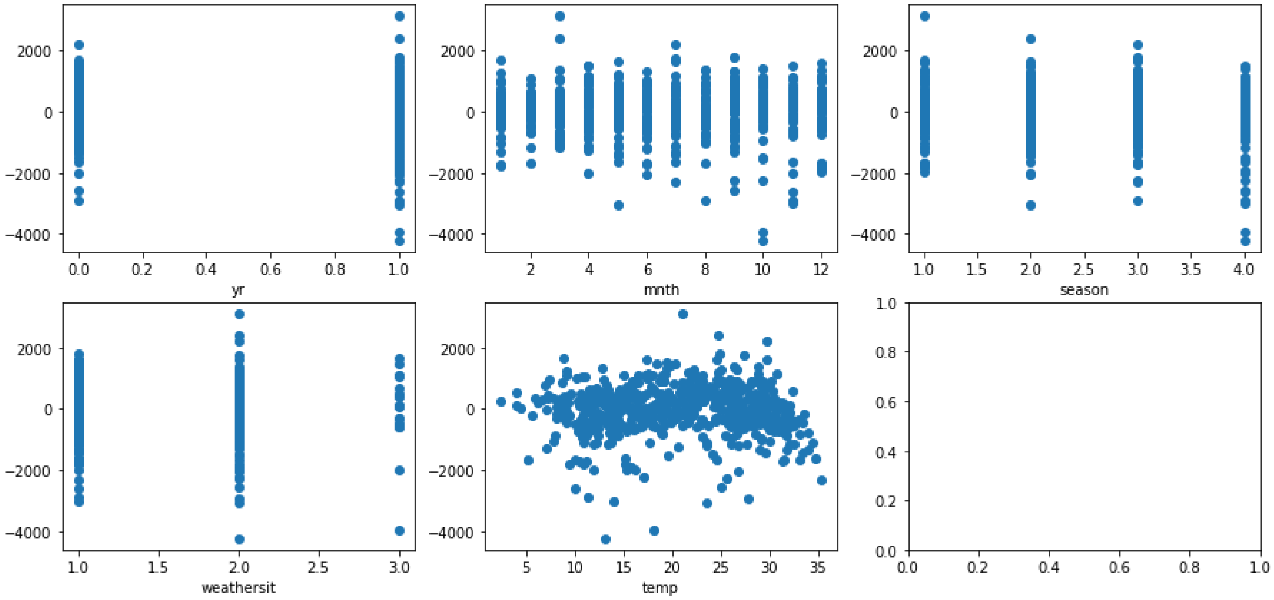


- We can see that registered have highest correlation with target variable followed by casual and it is because target variable is derived from sum of these two columns.
- If we remove registered and casual then temp and atemp have highest correlation with target variable.

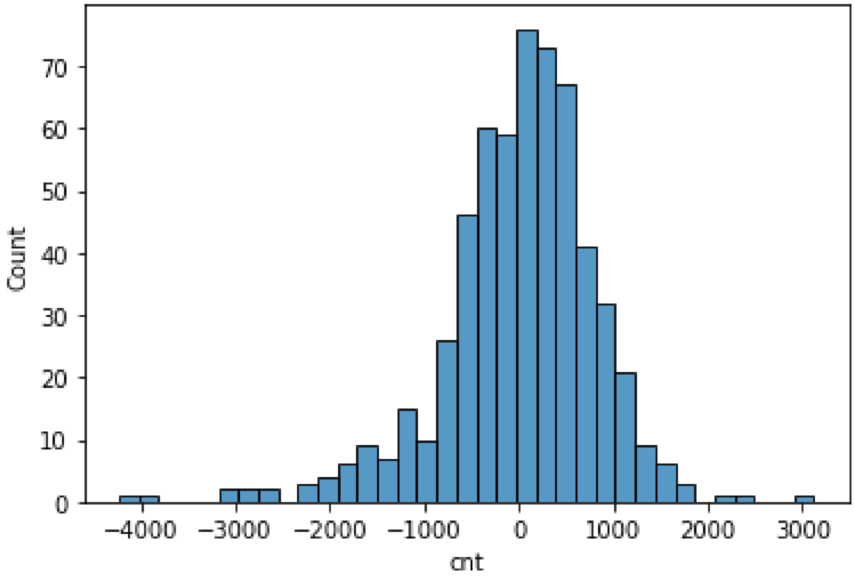
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer:

- Linearity assumption states that the relationship between the independent variables and the dependent variable is linear. This can be checked by plotting the residuals against the independent variables. If a linear relationship exists, the residuals should be randomly dispersed around zero.



- should be normally distributed. We can check this assumption by creating a normal probability plot of the residuals.



- The observations should be independent of each other. We can check this by checking the presence of correlation in the residuals or VIF.
- The independent variables should not be highly correlated with each other.

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

- Year, month, temperature are most significant features in dataset.
- **Year:** Year alone covers/explain 33.5% variance of the data and its correlation with target variable is also high 0.57.
- **Month:** Month column alone covers /explain 37.8% variance and when we add year with month variance covered by just two of them is 70% of the data.
- **Temperature:** Temperature have highest correlation with target variable around .63 and have highest linear coef_. Temperature alone covers/explain 38% variance in data.

General Subjective Questions

1. Explain the linear regression algorithm in detail.
Answer:

- Linear regression is a machine learning model that is used to model the relationship between a dependent variable and independent variables. Linear regression tries to find the best fitting line that help to describe relation between the independent variables and the dependent variable.

*y = b0 + w1x1 + w2x2 + ... + wn*x*

Where y is prediction, b0 is bias/intercept, w1..wn are weights assign to each independent feature.

- In Linear regression model find the best fitting line that minimizes the error between the predicted values and the actual values. It uses least squares, which finds the coefficients of the model that minimize the sum of the squared residuals.

Loss = sqrt(sum(prediction - actual)^2)

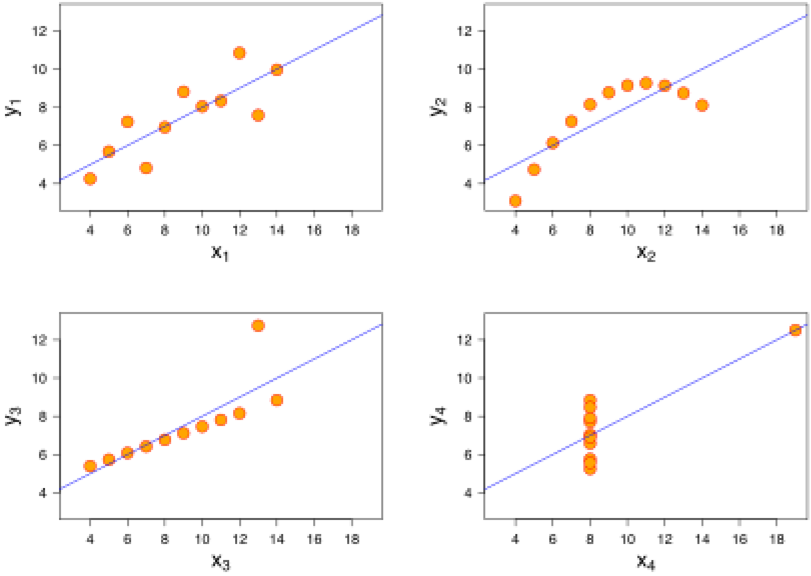
- Linear regression starts with a random set of coefficients and updates them in the direction of the negative gradient of the cost function in order to minimize the error. It is an iterative optimization algorithm that is called gradient descent.

Gradient = d(Loss)/d(w) where w can be w0, w1...wn.

2. Explain the Anscombe's quartet in detail.
Answer:

- Anscombe's quartet is a set of four datasets that were created by the statistician Francis Anscombe in 1973. These datasets were created to demonstrate the importance of visualizing data before analyzing it, and to show how different summary statistics can give very different impressions of the same data. Each of the four datasets in Anscombe's quartet has the following properties:

- The same mean of x and y
- The same variance of x and y
- The same correlation between x and y
- The same linear regression line



- However, when plotted, it is easy to see that each dataset has a very different distribution and shape. The x1 dataset is a roughly linear relationship, the x2 and x3 have an outlier and a non-linear relationship, and the x3has a perfect linear relationship, but with much more noise.

- This example illustrates that just looking at summary statistics such as mean, variance, and correlation can be misleading and that it is important to visualize the data before drawing conclusions. It's important to note that outliers, non-linearity and noise can greatly impact the results and the interpretation of the model in case of using Linear Regression.

3. What is Pearson's R?
Answer:

- The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. A value of 0 denotes no linear relationship between the variables, while a value of 1 denotes a perfect positive linear relationship. As values approach -1 or 1, the link becomes stronger. When one variable changes, the other variable changes in the sign direction.

*R = cov(x, y) / (std(x) * std(y))*

where std(x) and std(y) are the standard deviations of x and y, respectively, and cov(x, y) is the covariance between x and y.

- It's crucial to remember that Pearson R only accounts for linear interactions.

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

Scaling:

- Scaling is the process of transforming the values of a variable to a specific range. It is performed to ensure that the features of a dataset are on a similar scale, which can be important for linear types or distance-based of models.

Why is scaling performed

- It brings data into same range.
- Model can assume that all data have equal importance.
- It helps in speeding up the calculations in an algorithm.
- Curve between weights and loss is more spread in all direction.

Difference between normalized scaling and standardized scaling:

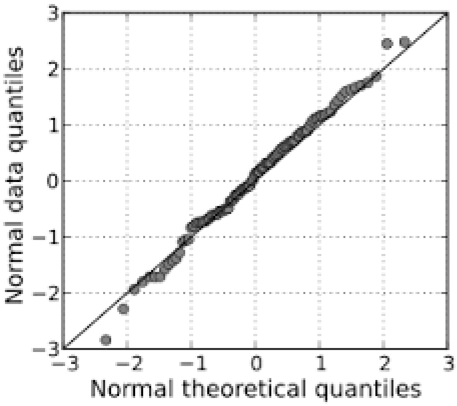
Normalization	Standarization
Normalization is the type of scaling scales the data so that it falls between a specific range, usually 0 and 1	Standardization is the type of scaling scales the data so that it has a mean of 0 and a standard deviation of 1.
x_norm = (x - x.min())/(x.max() - x.min())	x_std = (x - x.mean()) / x.std()
Normalization is useful when the features have to be compared with each other	Standardization is useful when the features are assumed to have a gaussian distribution

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

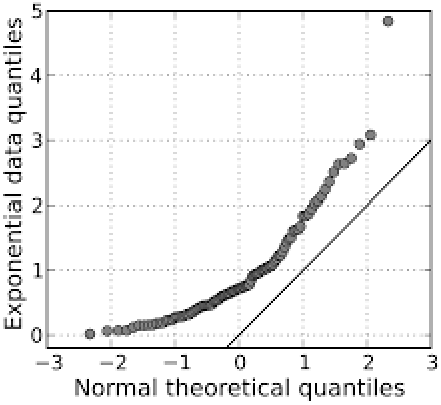
- VIF is a measure of the correlation between multiple independent variables in a multiple regression model. A VIF of 1 indicates that there is no multicollinearity between the independent variable and the other independent variables in the model, while a VIF greater than 1 indicates that there is some multicollinearity and VIF of 5 indicates that there is high multicolinearity.
- A VIF of infinite can occur when there is perfect multicollinearity between an independent variable and the other independent variables in the model. Perfect multicollinearity occurs when an independent variable can be exactly described by a linear combination of the other independent variables. This mostly occurs in One hot encoding when other one hot encoding can perfectly define a particular one hot encoding.

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

- A Q-Q plot, or quantile-quantile plot, is a graphical tool that is used to check the normality assumption of a linear regression model. It compares the distribution of the residuals (prediction errors) of the model with a normal distribution.
- A Q-Q plot is created by plotting the quantiles of the residuals against the quantiles of a normal distribution. The resulting plot shows how closely the residuals of the model match a normal distribution. If the residuals are normally distributed, the points on the Q-Q plot will fall along a straight line. If the residuals deviate from a normal distribution, the points on the Q-Q plot will deviate from a straight line.
- It's important to note that a Q-Q plot is not a formal statistical test, it is a visual tool that can help you to identify if the residuals are following a normal distribution. Additionally, a Q-Q plot can give you an idea of how severe the deviation is, it can help you to decide whether to discard the normality assumption, or to try some transformations.



- This plot indicated that distribution of any data of residual is normally distributed, since all the points are aligned with horizontal line.



- While this plot indicated that distribution of any data of residual is not normally distributed, since all the points are not aligned with horizontal line.