Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

1. "**weathersit**":

A value of 1 (Clear, Few clouds, Partly cloudy, Partly cloudy) sees maximum number of shared bike riders followed by value 2(Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist). Value 3 (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) sees very few bike rides and there are no bike rides reported for value 4(Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog)

2. "**holiday**":

Non holiday days(value=0) have a higher number of shared bike rides as compared to the holidays.

3. "**season**":

Summer and fall season see a higher number of shared bike rides as compared to spring and winters.

4. "**month**":

Months June to October see a higher number of shared bike rides as compared to the rest of the year.

5. "**workingday" and "weekday**":

These columns do not seem to have much impact on the shared bike rides.

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit)

get_dummies if a pandas function that we use to convert the categorical variables into numeric variables. Each distinct value of the given categorical variable gets divided into a separate binary column which can take either a value of 1 or 0.

Now, while creating these columns we can say that if there are 6 distinct values of the variable, we can do with just 5 columns because in case all those 5 columns have a value of 0, we can assume that the 6th is one.

This is where “drop_first=True” comes into picture. Here we tell the function to create one less column of the categorical variable. It also helps to avoid the issue of “multi-collinearity”

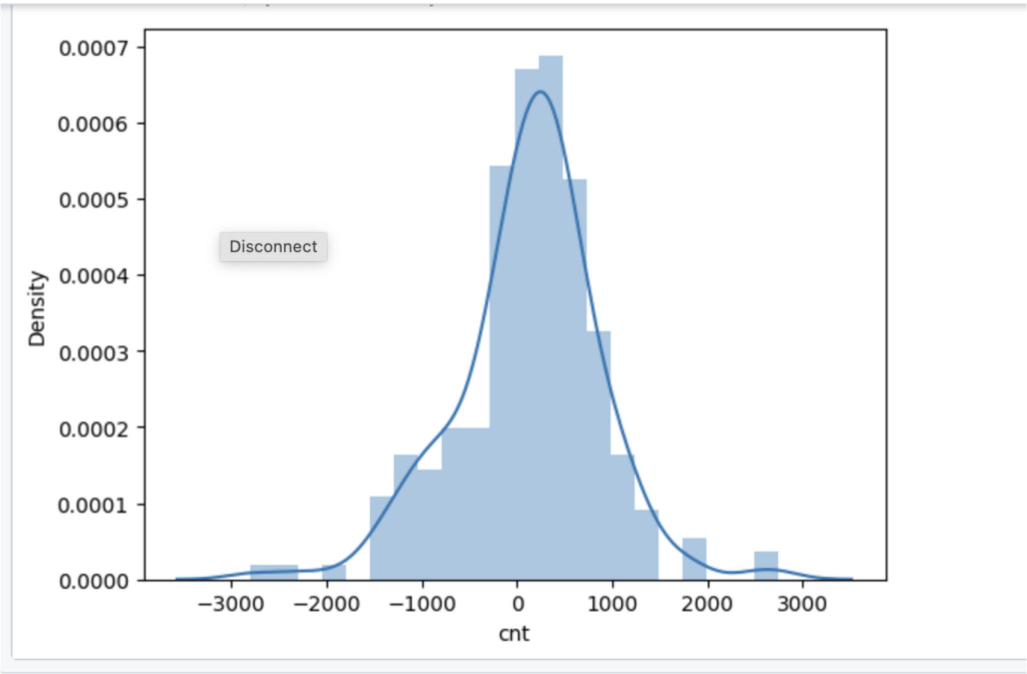
Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
Total Marks: 1 mark (Do not edit)
Answer: <Your answer for Question 3 goes below this line> (Do not edit)

"**atemp**" seems to have the highest correlation of the value 0.630685.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

After building the model, I created a graph of the residuals. The residuals i.e. the graph of the difference between the actual and the expected values of the target variable must be a normal distribution graph centered around 0.

Below is the output we got for the residual plot.



Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

1. **"atemp" and "temp"** (We removed "temp" from the final model because the two were highly correlated and could lead to multi-collinearity)
"atemp" has a very high positive correlation value [0.630685348953104] with the number of shared bike rides.
Higher the temperature, higher is the demand

2. "**yr**"
The year 2019 seems to have seen much higher number of shared bike rides as compared to the year 2018.
Positive correlation value being [0.5697284652110435]

3. "**windspeed**"
It seems to be highly negatively correlated with the demand for shared bike rides.
More the wind, lesser the demand and vice-versa.
Negative correlation value being [-0.2351324951410363]

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 6 goes here>

Linear regression is a supervised machine learning algorithm that computes the linear relationship between the dependent variable and one or more independent features. it does so by fitting a linear equation to observed data of the form

y=a+bx1+cx2

y : target dependent variable
x1 : independent input variable 1
x2 : independent input variable 2
a : intercept of the linear regression model
b : slope of the variable x1. It signifies the amount of change in the target y when input x1 changes by 1 unit.
c : slope of the variable x2. It signifies the amount of change in the target y when input x2 changes by 1 unit.

Linear Regression is of two types:
Simple Linear Regression:
When there is only one independent feature, it is known as Simple Linear Regression
The equation for simple linear regression is:
y=a+bx1

Multiple Linear Regression:
When there are more than one features, it is known as Multiple Linear Regression.
The equation for multiple linear regression is:
y=a+bx1+cx2

Assumption of the Linear Regression model:
Linearity:
The relationship between the independent and dependent variables is linear. Scatter plots can be used to test this assumption.

Homoscedasticity:
The error term has a constant variance across all levels of the independent variable. This means that the spread of the residuals is consistent along the regression line.

Normality of errors:
The error term is normally distributed. This assumption is important for hypothesis testing and constructing confidence intervals. Histograms or QQ plots can be used to assess the normality of errors.

Independence of observations:

Each observation is independent of the others. This means that the significance of one observation does not depend on the significance of another.

Residual errors have a mean value of zero:

The residual errors have a mean value of zero.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 7 goes here>

Anscombe's quartet is a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

Purpose of Anscombe's Quartet

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 8 goes here>

Pearson's R is a statistical tool that is being used to measure the correlation of the independent variables with the dependent target variable. It not only measures the strength of the correlation, but it also measures the direction of the correlation. It tells us of the correlation is positive or negative and the extent of the correlation.

Value Range: -1 to 1

-1: Highly negative correlation

+1: Highly positive correlation

How to calculate in Python?
It is calculated using the scipy.stats.pearsonr(x, y) function from the scipy library

Uses of Pearson's R
Feature Importance:
Understand the correlation between features to identify the most important ones for your model
Avoid overfitting:
Identify features that are correlated with the target variable but aren't informative

---

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 9 goes here>

What is Scaling?
Scaling is one of the preprocessing techniques that we use before training a machine learning model. Scaling is done on the data to make scales of the different independent variables comparable. This helps to avoid the dominance of one of the features as compared to others.

Why is scaling performed?
Scaling is performed to make the input variables comparable to each other. This is specially useful for the algorithms that use distance between the data in their functionality. The distances like Euclidian Distance are impacted by the scales of the features.
Also, the gradient descend performs better and faster when the data is scaled.

What is the difference between normalized scaling and standardized scaling?

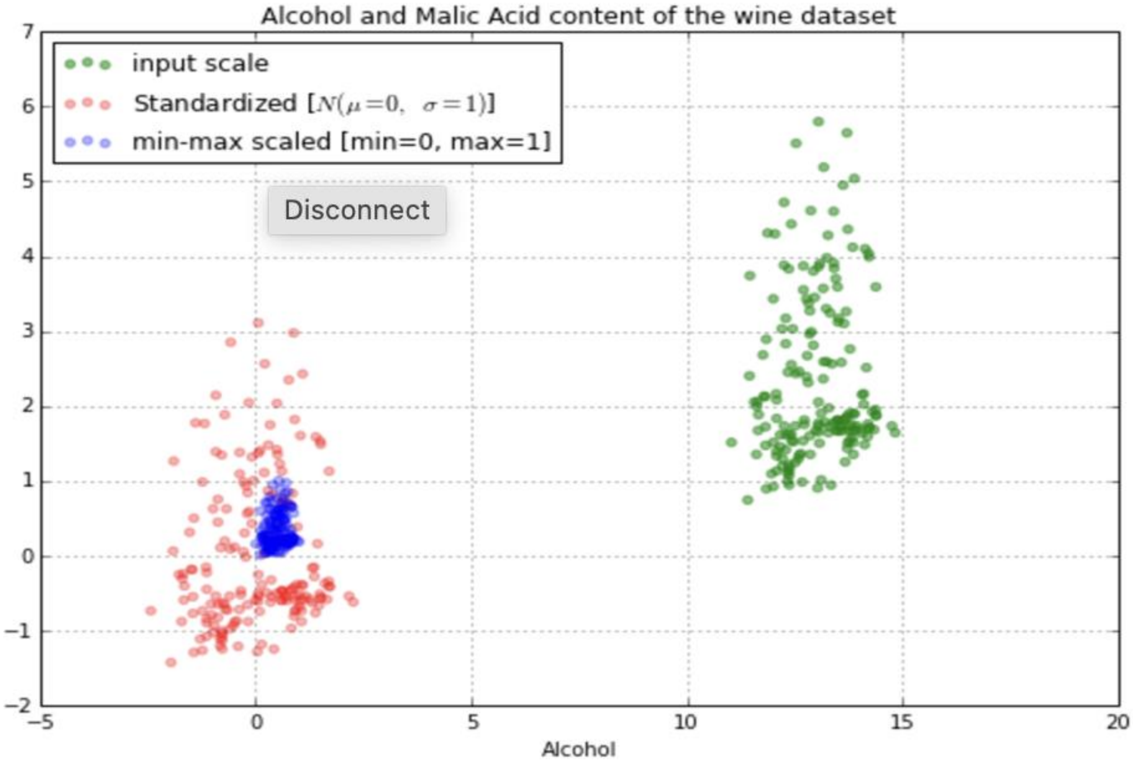
Normalized Scaling (Min-max scaling)
It converts the entire data between 0-1 while retaining the shape of the original data
Formula - (x-xmin)/(xmax-xmin)
Max value is mapped to 1
Min value is mapped to 0

Entire data is compressed between 0 and 1.
It also takes care of the outliers and maps them to the max or the min value as required

Standardisation
Standardisation refers to converting/scaling the data such that it has zero mean unit variance.
Formula - (X - Xmean)/standard_deviation

Shape of the data is retained. There is bound for the min or the max value.

The diagram below shows the difference between the two techniques.



Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 10 goes here>

VIF – Variance Inflation Factor
VIF is generally performed on the independent variables to identify the variables that can be explained well by the other independent variables.

If a variable can be explained well by others, we can ideally remove that variable from the model training because its effect on the dependent variable is being covered well by the other variables.

A value of inf (infinite) means that some variables are able to create perfect multiple regressions on that variable. Such a variable can be removed from model training.
VIF Formula:
VIF = 1/(1-R2)

R2 : It is the r-squared value that comes up when that variables is taken as target and all the other variables are taken as input and a regression is run. If R2 comes out to be 1, it means there is perfect regression
VIF = 1/(1-1) = 1/0 = infinite

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
(Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 11 goes here>

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.
This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Few advantages:
a) It can be used with sample sizes also
b) 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.

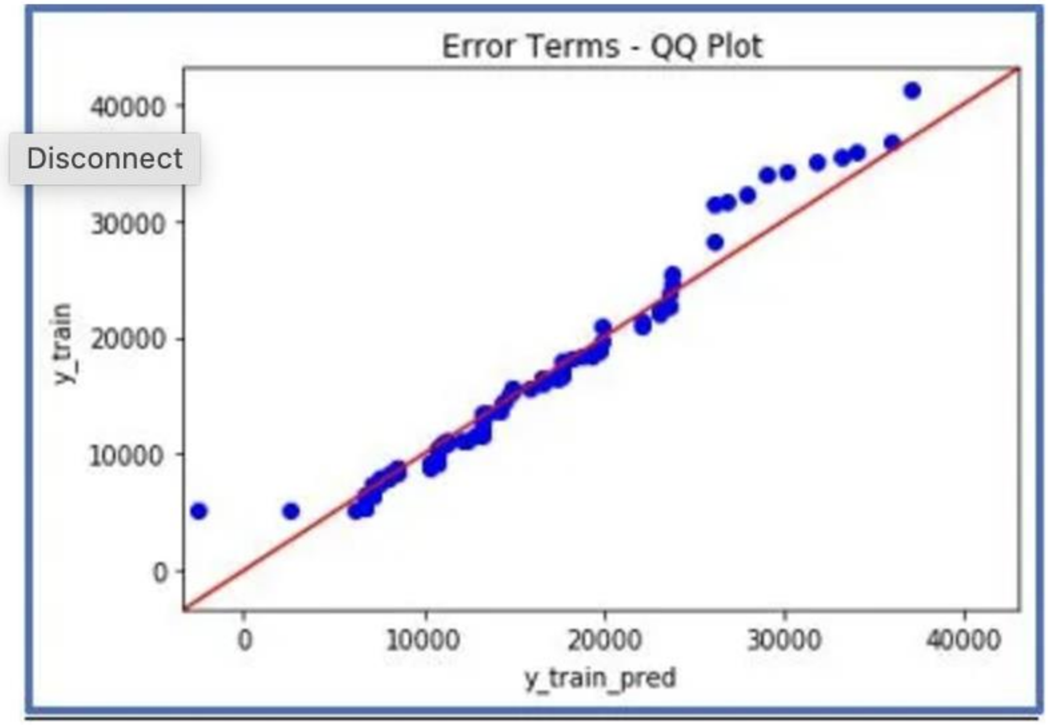
It is used to check following scenarios:

If two data sets —

i. come from populations with a common distribution
ii. have common location and scale
iii. have similar distributional shapes
iv. have similar tail behaviour

Interpretation:
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. Below are the possible interpretations for two data sets.

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis
b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.
c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.
d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis


Error Terms - QQ Plot
Disconnect
y_train
y_train_pred