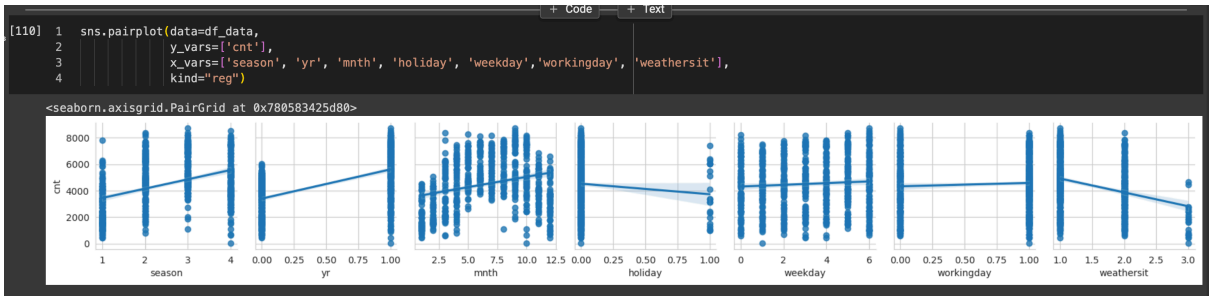
Linear Regression Subjective Questions & Answers

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)



Upon pair plotting categorical variables against target dependent variable, one can observe that,

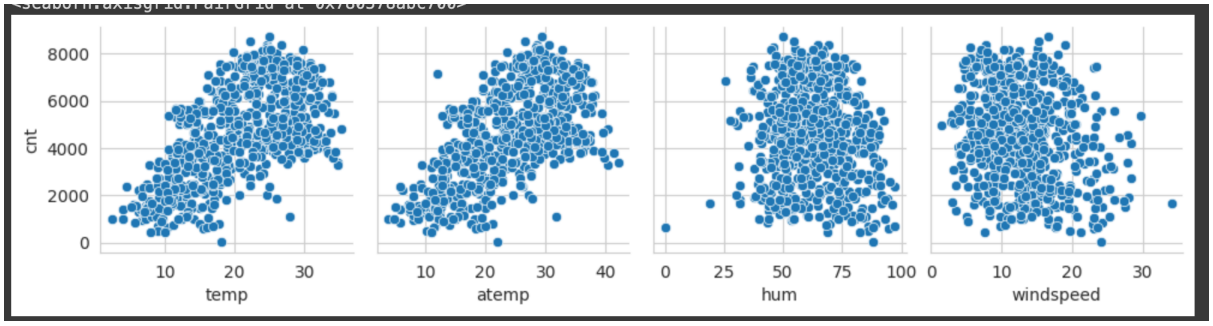
1. Bike hires are uniform across all season of an year.
2. There is rise in bike hires between 2018 and 2019
3. ***Bike hires count is more on holidays compared to non-holidays.***
4. Bike hires are uniform across all week days.
5. Weather situation impacts bike hires. Bad weather has less hires.

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

Categorical data (text) cannot be fed into ML model. ML models needs information in numeric format. Pd.get_dummies() encodes categorical data into numerical representation by introducing additional columns. Dummy encoding uses “N-1” features to represent N labels/categories.

Once, categorical data is encoded. The original categorical feature is removed as it is redundant (the information is transformed into new “N-1” features).

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



Based on above pair plots of numerical variables with target variable suggest,

1. "temp" feature and "atemp" feature are highly correlated with target "cnt"
2. More bike hires during warm weather over cold conditions.
3. "temp" and "atemp" are highly correlated. Correlation heatmap indicates the same.

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

The original dataset has been split into two datasets called training and validation/testing. The training and test data split was don 80:20 ratio.

Scikit-Learn's "train_test_split" function has been used to split the data.

Linear Regression model is then trained with training dataset. Later is validated with test dataset. R2 Score is used for evaluation.

5. Based on the final model, which are the top 3 features contributing significantly towards

Based on coefficients of Linear Regression model, following are top 3 features,

1. Year
2. Temp
3. Weather situation

6. explaining the demand of the shared bikes? (2 marks)

The objective in the assignment is to find demand of shared bikes. In given dataset 'cnt' indicated demand for bikes.

'cnt' attribute indicates demand.

The demand for bikes is highly dependent on year, temperature and weather situation.

General Subjective Questions

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

The function of a regression model is to determine a linear function between the X and Y variables that best describes the relationship between the two variables.

The relationship between the input variables (X) and the target variables (Y) can be portrayed by drawing a line through the points in the graph. The line represents the function that best describes the relationship between X and Y. The goal is to find an optimal "regression line", or the line/function that best fits the data.

Lines are typically represented by the equation: Y = m*X + b. X refers to the dependent variable while Y is the independent variable. In ML lingo it is represented as,

y(x) = w0 + w1 * x

In the case of "multiple linear regression", the equation is,
y(x) = w0 + w1x1 + w2x2 + … + w(n)x(n)

After establishing the formula for linear regression, the machine learning model will use different values for the weights, drawing different lines of fit. Remember that the goal is to find the line that best fits the data.

A cost function is used to measure how close the assumed Y values are to the actual Y values when given a particular weight value. The cost function for linear regression is mean squared error, which just takes the average (squared) error between the predicted value and the true value for all of the various data points in the dataset.

$$Cost function = \frac{1}{2n}\sum_{i=1}^{n}(y_{predicted} - y_{actual})^2$$

By using Gradient descent process, machine learning models tune parameters and reduce cost function.

The weights optimization means the linear regressor founds a line that best fits the data. Later this line is used to predict y value by plugging in independent variable x.

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

Anscombe's quartet is a group of four data sets that are nearly identical in simple descriptive statistics (involving variance and mean), but there are peculiarities that fool the regression model once you plot each data set.

However, when we plot these data sets, they look very different from one another.

Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.).

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

Pearson's R is a correlation coefficient for assessing linear relations. It describes what is happening in the scatterplot. The Pearson correlation coefficient is used to measure the strength of a linear association between two variables, where the value *r* = 1 means a perfect positive correlation and the value *r* = -1 means a perfect negative correlation.
The formula for r is,

$$r = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2}\sqrt{\sum_i (y_i - \overline{y})^2}}$$

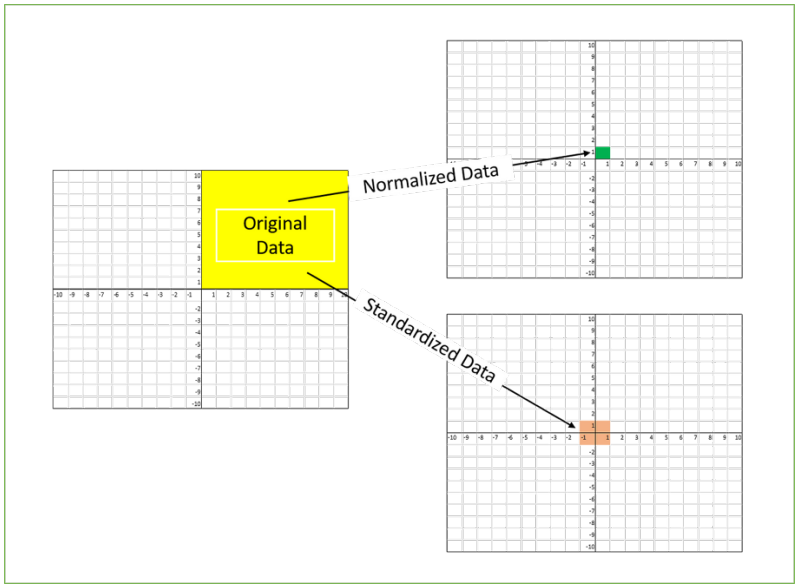
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Machine learning algorithms just sees numbers, if there is a vast difference in the range say few ranging in thousands and few ranging in the tens, and it makes the underlying assumption that higher ranging numbers have superiority of some sort. So these more significant number starts playing a more decisive role while training the model
Scaling brings all features in the same standing,

The most common techniques of feature scaling are Normalization and Standardization.

- Normalization is used when we want to bound our values between two numbers, typically, between [0,1] or [-1,1].

- Standardization transforms the data to have zero mean and a variance of 1, they make our data unitless.



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

Variance Inflation Factor (VIF) in order to determine if we have a multicollinearity problem.

Consider the following linear regression model:
$Y = \beta 0 + \beta 1 \times X1 + \beta 2 \times X2 + \beta 3 \times X3 + \varepsilon$

To calculate VIF for independent variable X1 is,

$$VIF_1 = \frac{1}{1 - R^2}$$

R^2 in this formula is the coefficient of determination from the linear regression model which has:

- X1 as dependent variable
- X2 and X3 as independent variables

R^2 comes from the following linear regression model:

$X_1 = \beta_0 + \beta_1 \times X_2 + \beta_2 \times X_3 + \varepsilon$

And because R^2 is a number between 0 and 1:

- When R^2 is close to 1 (i.e. X2 and X3 are highly predictive of X1): the VIF will be very large
- When R^2 is close to 0 (i.e. X2 and X3 are not related to X1): the VIF will be close to 1

Therefore the range of VIF is between 1 and infinity.

An infinite value of VIF for a given independent variable indicates that it can be perfectly predicted by other variables in the model.

VIF of 1 for a given independent variable (say for X_1 from the model above) indicates the total absence of collinearity between this variable and other predictors in the model

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Quantile-Quantile (Q-Q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

QQ plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.

Q-Q plots are useful for checking whether a dataset follows a certain theoretical distribution, such as a normal distribution or a log-normal distribution. If the points on the Q-Q plot fall on a straight line, it indicates that the two datasets have the same distribution.