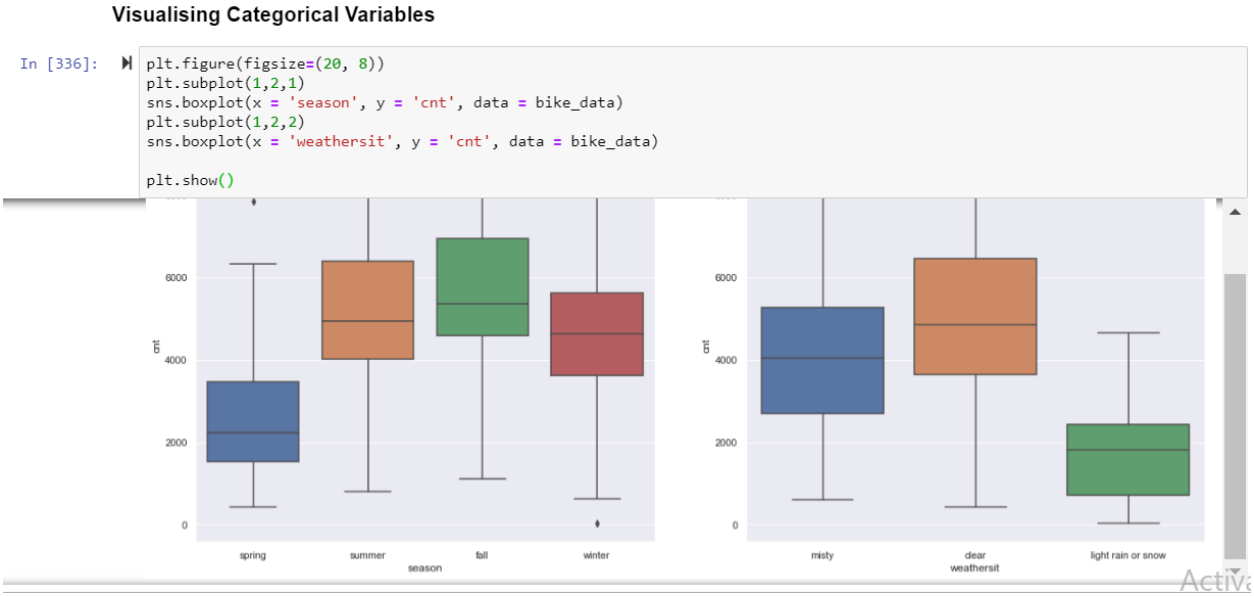
Assignment-based Subjective Questions

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 two categorical variables 'Season' and 'Weathersit'. When these two variables are visualized using box plot, it gives fair idea about their impact on the dependent variable which is 'cnt'



It tells 'fall' season has the largest footfall with mean of ~ 5500 while spring season would generally have low footfall compared to all seasons with mean of ~ 2200

Similarly, when weathers comparison with cnt is seen, it tells people would like to hire bike on a clear sky day while when it is raining or having snow footfall with mean going down to below 2000 footfall.

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

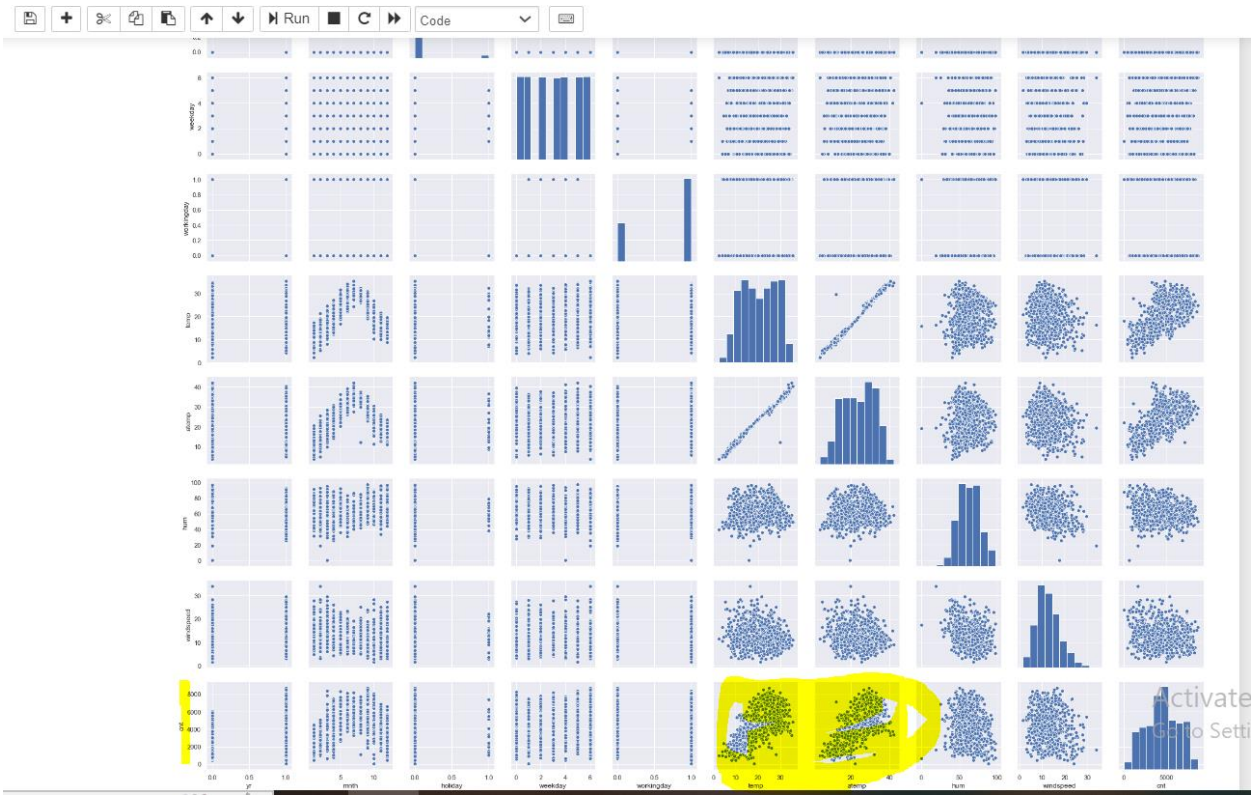
Answer: When categorical columns are converted to integer column. It creates N number of columns where N = number of categories. We do not need all N column to determine the categorical values as it can be done using N-1 columns. If we do not use drop_first=True, then all N columns gets used while building the model increasing the model complexity. That's why it is important to use drop_first=True

Let's take an example of a categorical column Is_Furnished having 3 values. When it is converted to numerical you get 3 columns. Now, you don't need three columns. You can drop the `one` column, as the type of furnishing can be identified with just the last two columns where —

- `00` will correspond to `furnished`
- `01` will correspond to `unfurnished`
- `10` will correspond to `semi-furnished`

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

Answer: Looking at the pair-plot among the numerical variables, Temp / Atemp has the 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? (3 marks)

Answer: We run the model on the test data. The outcome of the model is then compared with the test data target variable. R-squared score is then derived to calculate the accuracy of the model.

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: Based on final model, below features are having higher weightage to get predicted demand of the shared bike as they are having the high coefficient values. :

1) Yr : high positive coef
2) Spring : High negative coef
3) Light rain or snow : High negative coef.

General Subjective Questions

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

Answer:

Linear Regression is one of the most widely used and simple machine learning algorithms. It is a supervised learning technique that tries to find the best-fit line for a given set of data points.

Linear regression assumes that there is a linear relationship between the input variables (X) and the output variable (y). That is, we can model y as a linear combination of X, plus some error term:

$$y = b0 + b1*X1 + b2*X2 + ... + bn*Xn + e$$

where b0, b1, ..., bn are the coefficients or weights of the model, and e is the error term. The goal of linear regression is to find the optimal values of the coefficients that minimize the error term, or equivalently, the sum of squared errors (SSE):

$$SSE = sum((y - y\_pred)^2)$$

where y_pred is the predicted value of y for a given X. To find the optimal coefficients, we can use different methods, such as gradient descent, normal equation, or least squares.

One of the advantages of linear regression is that it is easy to interpret and explain. We can see how each input variable affects the output variable by looking at the corresponding coefficient. For example, if b1 is positive, it means that X1 has a positive effect on y, and vice versa. We can also measure the goodness of fit of the model by using metrics such as R-squared, adjusted R-squared, or root mean squared error (RMSE).

Step 1: Import the necessary libraries and load the data. We use the sklearn library for linear regression and RFE, pandas for data manipulation, and matplotlib for visualization. We can take bike sharing dataset as an example.

Step 2: Split the data into features (X) and target (y). The target variable is the count of bikes shared on a given day. The features are the various characteristics, such as month, weather, season, temp , windspeed , humidity etc.

Step 3: Split the data into training and test sets. We will use 80% of the data for training and 20% for testing. We will also set a random state for reproducibility.

Step 4: Create a linear regression model and fit it to the training data. We use the LinearRegression class from sklearn.linear_model.

Step 5: Evaluate the model performance on the test data. We use the mean squared error (MSE) and the coefficient of determination (R-squared) as the metrics.

Step 6: Create an RFE object and fit it to the training data. We use the RFE class from sklearn.feature_selection. RFE is a method that selects the best features by recursively eliminating the least important ones based on the model coefficients or feature importances. We will specify the number of features we want to select as a parameter.

Step 7: Get the selected features from the RFE object. We can use the support_ attribute to get a boolean mask of the selected features, and then use it to filter the feature names. We calculate the VIF for selected features and eliminated those which are having high VIF values, generally more than 5,

Step 8: Create a new linear regression model using only the selected features and fit it to the training data.

Step 9: Evaluate the new model performance on the test data using the same metrics as before.

Step 10: Compare the results of the original model and the RFE model. We can use a bar plot to visualize the MSE and R-squared values of both models.

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

Answer: Anscombe's quartet is a group of four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points.

The quartet was constructed demonstrate the importance of plotting data before analyzing it and the effect of outliers and other influential observations on statistical properties.

The four datasets have the same mean, variance, correlation, and linear regression line for both x and y variables, as shown in the table below:

However, when the datasets are plotted on a scatter plot, they look very different from each other, as shown in the figure below:

We can describe the four datasets as follows:

- Dataset I: fits the linear regression model well and has no outliers.

- Dataset II: has a non-linear relationship between x and y and cannot be modeled by a linear regression.

- Dataset III: has a linear relationship between x and y but is influenced by an outlier that lowers the correlation coefficient and changes the slope of the regression line.

- Dataset IV: has no relationship between x and y except for one high-leverage point that produces a high correlation coefficient and a misleading regression line.

Anscombe's quartet shows us why we need to visualize our data before applying any statistical methods or algorithms to it. It also shows us that simple summary statistics are not enough to capture the complexity and diversity of real-world data.

3. What is Pearson’s R? (3 marks)

Answer:

Pearson's R is a measure of how well a linear model fits the data. Here are few points to understand it:

- It is also called the correlation coefficient or the coefficient of determination.

- It ranges from -1 to 1, where -1 means a perfect negative linear relationship, 0 means no linear relationship, and 1 means a perfect positive linear relationship.

- It indicates how much of the variation in the dependent variable (the outcome) is explained by the variation in the independent variable (the predictor).

- It can be calculated by dividing the covariance of the two variables by the product of their standard deviations.

- It can be used to assess the strength and direction of a linear relationship, but it does not imply causation or account for other factors that may affect the outcome.

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

Answer :

Scaling is a data preprocessing technique that transforms the values of numerical features to a common scale. Scaling is performed to avoid bias or distortion in machine learning models due to features having different ranges or units. Scaling can also improve the performance and convergence of some algorithms.

There are two common types of scaling: normalized scaling and standardized scaling. Normalized scaling rescales the values of a feature to the range [0, 1] or [-1, 1], depending on the minimum and maximum values of the feature. Standardized scaling rescales the values of a feature to have zero mean and unit variance, depending on the mean and standard deviation of the feature.

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

VIF stands for variance inflation factor, which measures how much the variance of a regression coefficient is inflated due to multicollinearity.

Multicollinearity occurs when there is a high correlation between two or more predictor variables in a regression model. When multicollinearity is perfect, meaning that one predictor variable can be expressed as a linear combination of other predictor variables, the VIF becomes infinite.

This means that the regression coefficient cannot be estimated uniquely or reliably. To avoid this problem, we should check the VIF values for each predictor variable and remove or combine the variables that have very high VIF values. Some possible points to explain this are:

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

(3 marks)

Answer :

`

A Q-Q plot, or quantile-quantile plot, is a graphical tool that compares two probability distributions by plotting their quantiles against each other. A quantile is a value that divides a distribution into equal proportions. For example, the median is the 50th percentile, or the 0.5 quantile, of a distribution.

A Q-Q plot can be used to assess how well a sample of data fits a theoretical distribution, such as the normal distribution. If the sample data follows the theoretical distribution, the points on the Q-Q plot will lie on a straight line. A Q-Q plot can also be used to compare two samples of data and see if they come from the same distribution.

A Q-Q plot is useful in linear regression for several reasons:

- It can help check the assumption of normality of the error terms. If the errors are normally distributed, the Q-Q plot of the residuals will show a straight line.

- It can help identify outliers and influential observations. If there are points that deviate significantly from the straight line, they may indicate outliers or influential observations that affect the fit of the regression model.

- It can help assess the goodness-of-fit of the regression model. If the model fits the data well, the Q-Q plot of the residuals will show a straight line with no systematic patterns.

- It can help compare different regression models and choose the best one. If there are multiple models that fit the data, the Q-Q plot can help compare their residuals and see which one has the smallest deviation from normality.