Registered Email ID – antonyphilly@gmail.com , **Name**: Antony John Sundar

Assignment Path- https://github.com/AntonyJohnSundar/BikeSharingDemandPrediction.git

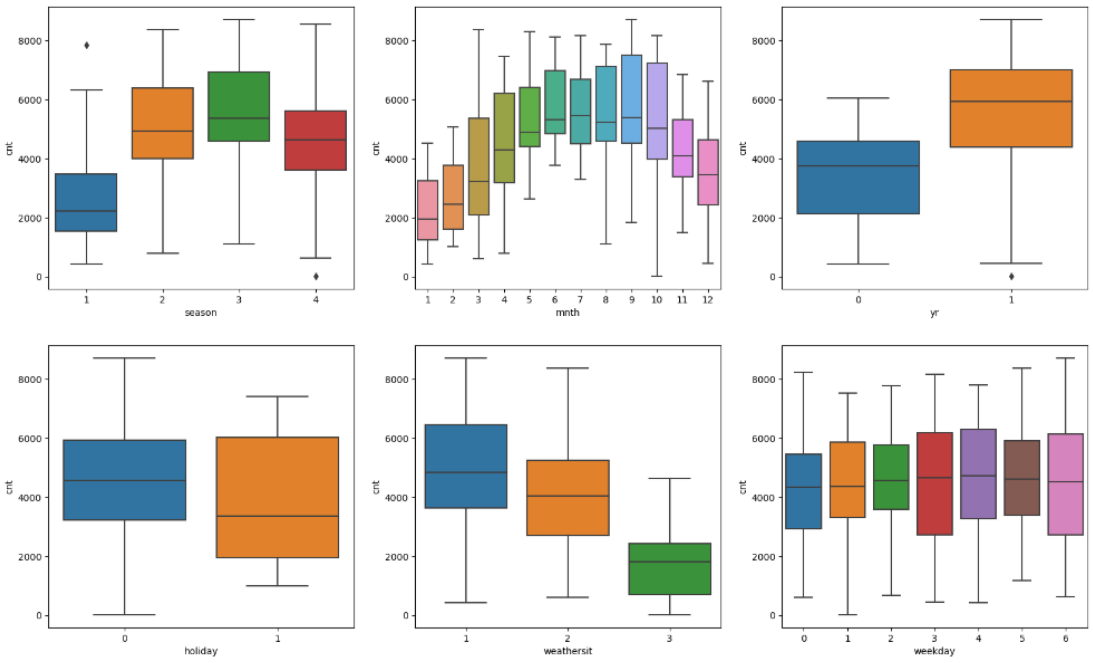
Assignment-based Subjective Questions:

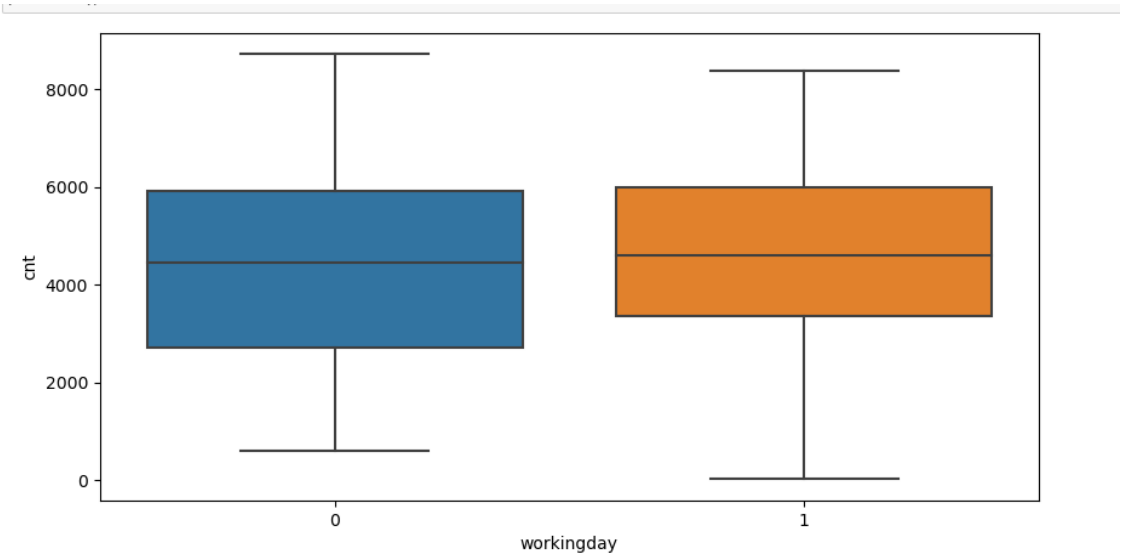
Questions 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer:

The identified categorical variables from the given data set were 'season', ' weathersit', 'holiday', 'mnth' , 'yr' , 'weekday', and 'workingday'.

These variables were visualized using a box plot on the target variable cnt





Inference on their effect on the dependent variables are as follows:

1. **season** – Bike demand was high on Fall and Summer, and was low on Spring and Winter
2. **mnth** – The Bike demand increases from March and peaking at September, the demand is low between November and February (This could be the fact associated with season as in winter the demand is low, most parts of the USA experience winter between Nov and Feb)
3. **yr**- The bike demand in the year 2019 is higher than 2018
4. **holiday**- The bike demand was lower on the holidays
5. **weathersit**- The bike demand was higher when the weather is pleasant with clear sky and low on light rain or snow. The bike demand was zero when the weather is harsh with heavy rain or heavy snow as per the data

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

Answer:

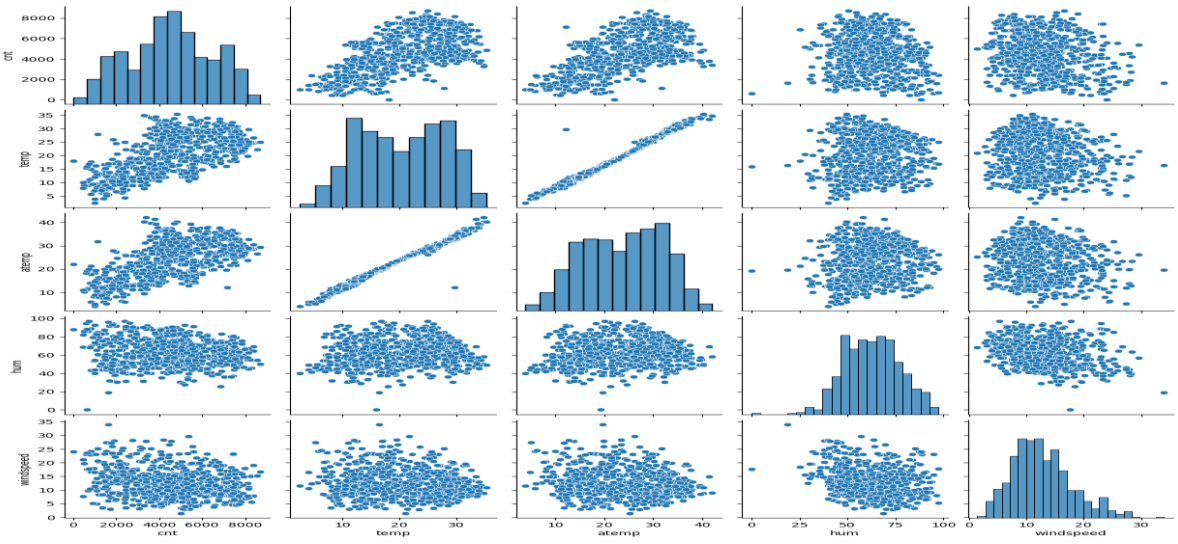
Dummy variables from pandas.get_dummies() function, essentially creates new binary (0 or 1) columns for each category of a feature. For 'n' category variables 'n' new columns are created.

This can be represented by n-1 columns because if all n-1 columns have a value of 0, the nth category must be True.

By using drop_first=True, we avoid creating an extra column, reducing the dimensionality of the data and preventing multicollinearity. This is critical on models like linear regression, which assumes that the predictor variables are not highly correlated.

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

Answer:

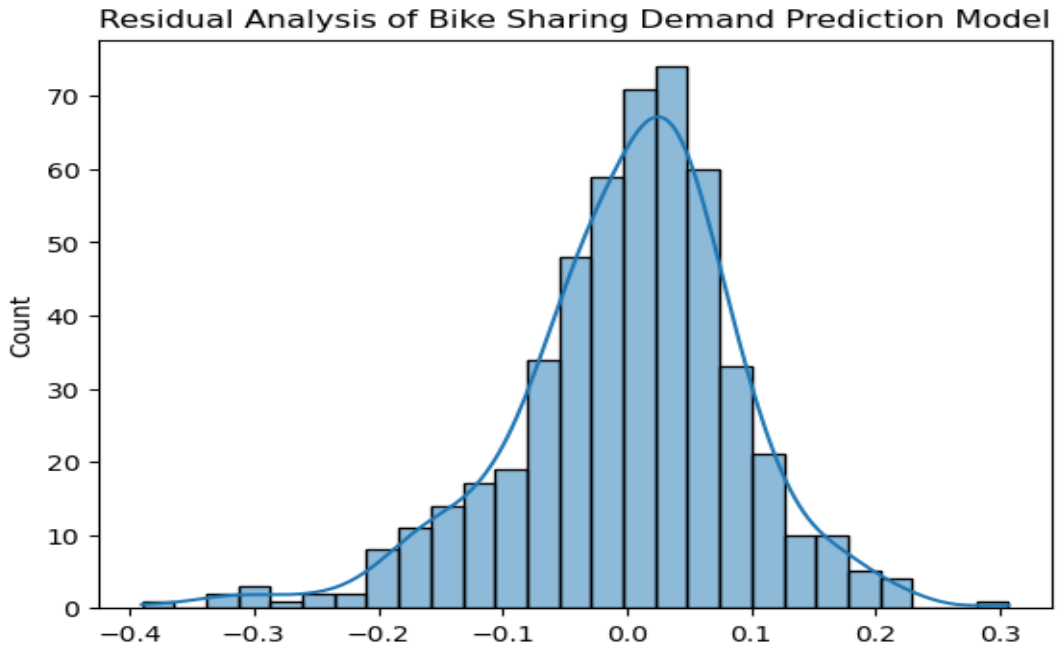


The pair plot on the numerical variables indicates that variable '**temp'** and '**atemp'** has high correlation with the target variable.

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

Answer:

By using Residual Analysis method, Plotting a graph for the difference between the observed value of the dependent variable (y) and the predicted value (y_pred)



Residuals distribution should follow normal distribution and centred around 0 (mean = 0). We validate this assumption about residuals by plotting a histplot of residuals and see if residuals are following normal distribution or not. The above diagram shows that the residuals are distributed about mean = 0. Thus, indicating that the linear regression approach is appropriate for this model building.

Question 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:

Best fit regression line for the final model is expressed as below:

cnt = 0.268203 **+** 0.619981*temp + 0.225886*yr **+** 0.138953*season_Winter **+** 0.094378*mnth_Sep

+ 0.079967*season_ Summer **-** 0.048206*mnth_Jul **-** 0.191358*weathersit_Light Snow & Rain

- 0.093018*holiday **-** 0.205869*windspeed **-** 0.286785*hum

temp, hum and yr has a significant impact on the bike demand compared to other predictor variables

General Subjective Questions:

Question 1. Explain the linear regression algorithm in detail.

Answer:

Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features. It's one of the simplest types of regression, and it's used for predicting a continuous output variable based on one or more input variables. The input variables can be of categorical or numerical

The linear regression algorithm:

Model Specification: The first step in linear regression is to specify the model. In simple linear regression, the model is a straight line (only 2 features), and in multiple linear regression (3 or more features), it's a hyperplane.

In general, MLR can be expressed as

$$y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \epsilon$$

- y = the predicted value of the dependent variable
- B_0 = the y-intercept (value of y when all other parameters are set to 0)
- $B_1 X_1$ = the regression coefficient (B_1) of the first independent variable (X_1) (a.k.a. the effect that increasing the value of the independent variable has on the predicted y value)
- … = do the same for however many independent variables you are testing
- $B_n X_n$ = the regression coefficient of the last independent variable
- ϵ = model error (a.k.a. how much variation there is in our estimate of y)

Best Fit Line: The algorithm calculates the best fit line (or hyperplane) that minimizes the sum of the squared differences between the actual and predicted values. This is also known as the least squares method.

Cost Function: The cost function for linear regression is the Mean Squared Error (MSE), which measures the average squared difference between the actual and predicted values. The goal of the algorithm is to minimize this cost function.

Gradient Descent or Normal Equation: The algorithm uses either the gradient descent method or the normal equation to find the parameters (coefficients) that minimize the cost function.

Prediction: Once the model is trained (i.e., the parameters are learned), it can be used to predict the output variable for new input data.

Evaluation: The performance of the model is evaluated using various metrics, such as R-squared, Mean Squared Error (MSE), or Root Mean Squared Error (RMSE).

The model's equation provides clear coefficients that elucidate the impact of each independent variable on the dependent variable, facilitating a deeper understanding of the underlying dynamics.

It's important to note that linear regression makes several assumptions, including linearity, independence, homoscedasticity, and normality. If these assumptions are violated, the results of the regression may not be reliable.

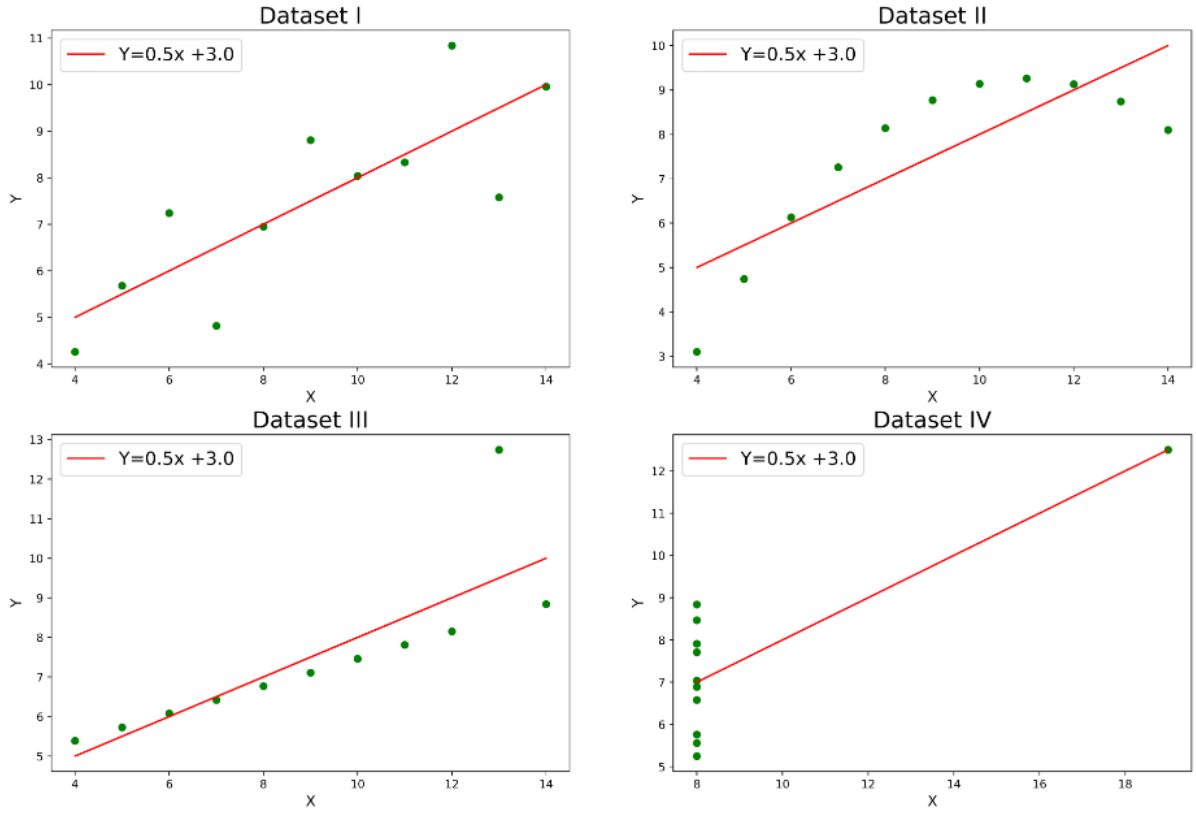
Question 2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet is a set of four datasets that were created by the statistician Francis Anscombe in 1973. Each dataset consists of 11 x-y pairs. The quartet is known for its interesting properties despite having nearly identical descriptive statistical properties (means, variances, correlations, and linear regression lines), the datasets look very different when visualized on a scatter plot.

Data set:

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Visualization the above data set



The key points from Anscombe's Quartet

Statistical Properties: Each dataset in the quartet has the same mean and variance for both x and y variables. They also have the same correlation between x and y, and the same linear regression line.

Visual Differences: Despite the similar statistical properties, each dataset appears very different when visualized. Dataset 1 has perfect linear relationship, Dataset 2 is a curve, Dataset 3 is cloud of points, and Dataset 4 might have outliers.

Purpose: Anscombe quartet to demonstrates the importance of visualizing data before analysing it. The quartet shows that datasets with identical statistical properties can be very different, underscoring the fact that summary statistics don't tell the whole story.

Implications: The quartet reminds the importance of exploratory data analysis and data visualization. Summary statistics are not reliable when analysing a dataset.

In conclusion, Anscombe's Quartet is a classic demonstration of why visual data exploration is an essential step in data analysis. It reminds that summary statistics, while useful, cannot capture all the nuances and patterns in a dataset.

Question 3. What is Pearson's R?

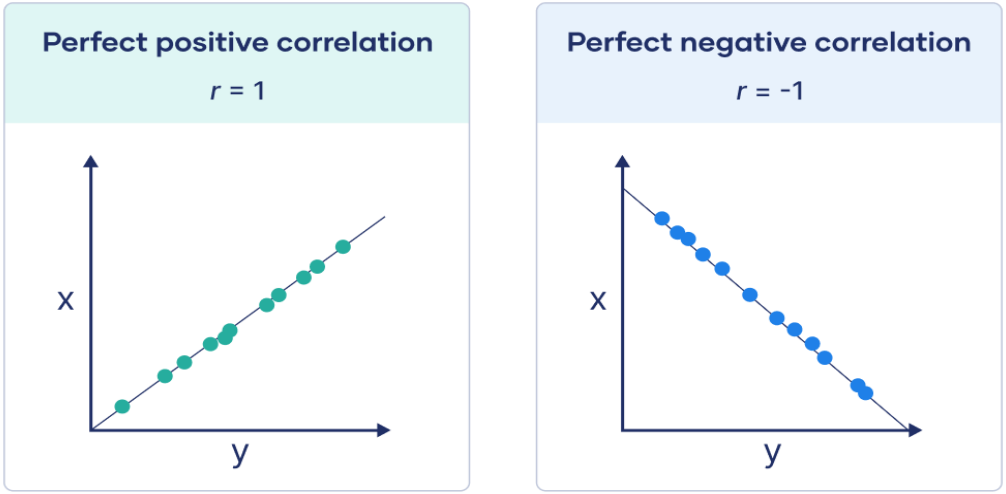
Answer:

Pearson's r is a numerical summary of the strength and direction of the linear association between two variables. Its value ranges between -1 to +1. It shows the linear relationship between two sets of data. In simple terms, it tells us can we draw a line graph to represent the data

r = 1 means the data is perfectly linear with a positive slope

r = -1 means the data is perfectly linear with a negative slope

r = 0 means there is no linear association



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

Answer:

Scaling: Scaling is a step of data preprocessing applied to independent variables to normalize the data within a particular range. It's a technique to standardize the independent variables of the dataset in a specific range. It's generally performed during the data preprocessing stage and helps to normalize the data within a particular range.

Why Scaling is Performed: Scaling is performed because most of the times, collected datasets contain features highly varying in magnitudes, units, and range. If scaling is not done, then a machine learning algorithm tends to weigh greater values higher and consider smaller values as lower values, regardless of the unit of the values. This can also speed up the convergence of stochastic gradient descent, as features would be on a similar scale.

Normalization and **Standardization** are two different techniques used in feature scaling that prepare the data for machine learning algorithms.

Normalization (Min-Max Scaling): This method scales the features to be within a specific range, usually 0 to 1. The formula for normalization is: **Xnew = (X−Xmin) / (Xmax−Xmin)**

Normalization is useful when your data has varied scales and the algorithm you are using does not make assumptions about the distribution of your data1. However, normalization is sensitive to outliers12.

Standardization (Z-Score Normalization): This method standardizes features by subtracting the mean and scaling to unit variance12. The result is a distribution with a mean of 0 and a standard deviation of 1. The formula for standardization is: **Xnew = (X−μ)/σ**

Where μ is the mean and σ is the standard deviation12. Standardization can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true12. Unlike normalization, standardization does not have a bounding range, which makes it less affected by outliers.

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

Answer:

VIF - the Variance Inflation Factor -The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity. If there is perfect correlation, then VIF = infinity.

VIF =1/ (1-R^2).

Where R^2 is the co-efficient of determination for regressing one independent with other independent variables.

If that independent variable can be perfectly predicted by other independent variables, then it will have perfect correlation and its R-squared value will be equal to 1.

So, VIF = 1/ (1-1) which gives VIF = 1/0 which results in "infinity" mathematically.

Question 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, short for quantile-quantile plot, is a graphical tool used in statistics to assess if a set of data plausibly came from some theoretical distribution such as a Normal, Exponential, or Uniform distribution. It is a plot of the quantiles of two distributions against each other

A Q-Q plot provides a comprehensive, visual means for assessing the validity of the normality assumption in linear regression, which is crucial for the reliability of the regression analysis. If the normality assumption is violated, it may be necessary to apply data transformations or use different statistical methods that do not assume normality

The importance of a Q-Q plot in linear regression analysis lies in its ability to:

- Determine if two data sets come from populations with a common distribution.
- Check if the two data sets have common location and scale.
- Assess if the two data sets have similar distributional shapes.
- Evaluate if the two data sets have similar tail behaviour.
- Detect many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers