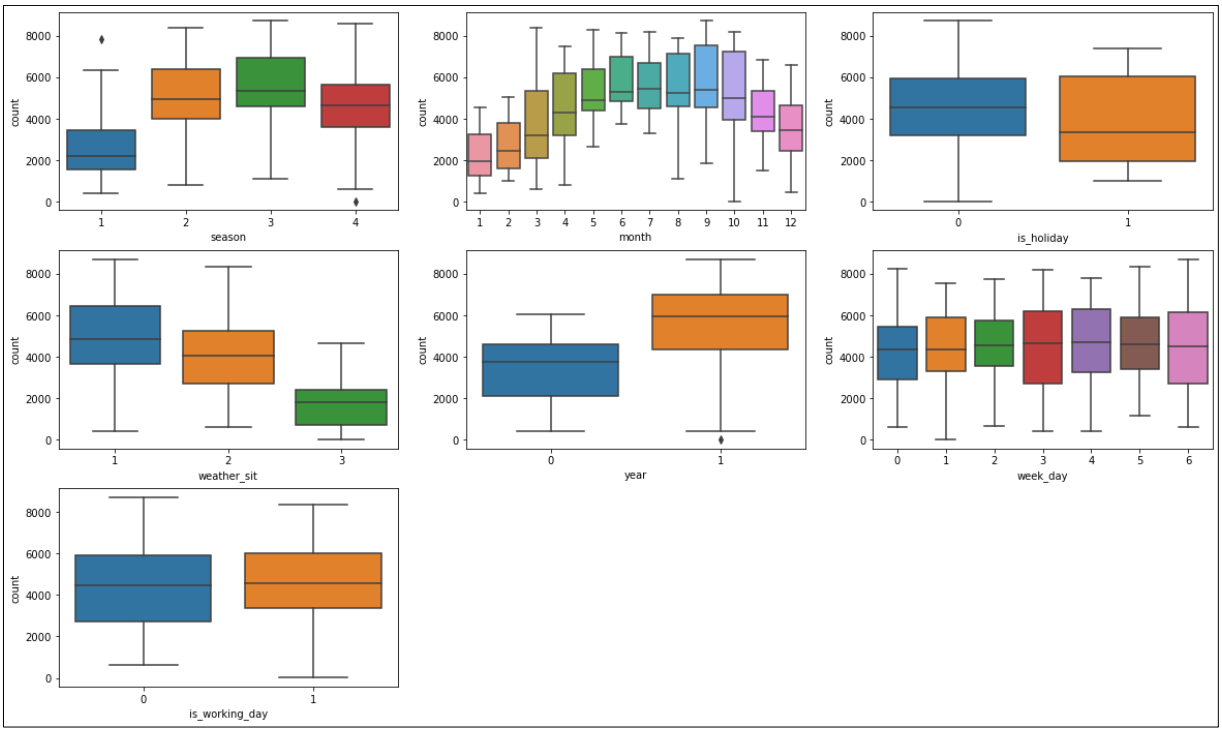
Assignment-based Subjective Questions

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:



The categorical variables in the dataset are

1. Season
2. Weathersit
3. Holiday
4. Month
5. Year
6. Weekday.
7. Workingday

These variables are analyzed using a boxplot and have the following effect on the dependent variable: -

- The number of bikers is the **highest** in **Fall season, followed by Summer**. **Spring season** has the **lowest** number of bikers.
- The demand for bikes is **higher in July and September** comparatively. The demand increased constantly from January to September and dropped after October.
- The usage of bikes on **weekday is higher than on holida**y. Also, the bike usage throughout the week days is almost the same.
- The demand of bikes has increased in 2019 as compared to 2018.
- People prefer to ride bikes on a **clear or partly cloudy day**. There is **no demand** of bikes on during **heavy rain or thunderstorm.**

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

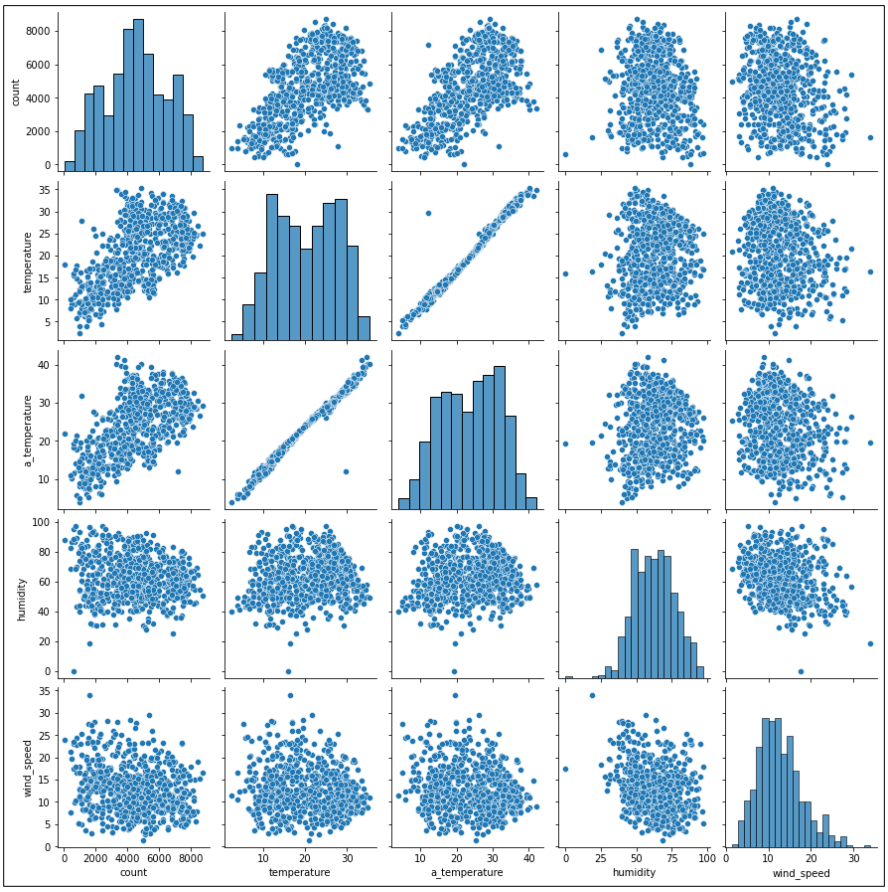
Answer:

If the first column is not dropped then your dummy variables will be correlated. This may affect some models adversely and the effect is stronger when the cardinality is smaller. For example, iterative models may have trouble converging and lists of variable importance may be distorted. Another reason is, keeping all dummy variables leads to multicollinearity between the dummy variables.

To avoid this, it is important to use drop_first=True, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

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

Answer:



'temp' numerical variable has the highest correlation with the target variable, followed by "a_temp".

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

Answer:

- Linear Relationship between the features and target: This is validated by plotting pairplot to check the linearity between features and target variable.
- Little or No Multicollinearity between the features: This is validated by calculating the VIF of the predictor variables. Accepted VIF is less than 5.
- Homoscedasticity Assumption: This is validated by plotting scatter plot to ensure no visible pattern in error terms.
- Normal distribution of error terms: This is validated by plotting distpplot for residuals. The mean is centered at 0.
- Little or No autocorrelation in the residuals: This is validated by plotting scatter plot to ensure no correlation in error terms.

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:

The top 3 features based on the final model are

1. a_temperature (4268.69)

2.year (2037.19)

3. month_Sep (548.94)

General Subjective Questions

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

Answer:

Linear Regression is a type of supervised Machine Learning algorithm that is used for the prediction of numeric values. Linear Regression is the most basic form of regression analysis.

It assumes that there is a linear relationship between the dependent variable(y) and the predictor(s)/independent variable(x). In regression, we calculate the best fit line which describes the relationship between the independent and dependent variable.

Regression is performed when the dependent variable is of continuous data type and Predictors or independent variables could be of any data type like continuous, nominal/categorical etc. Regression method tries to find the best fit line which shows the relationship between the dependent variable and

predictors with least error. In regression, the output/dependent variable is the function of an independent variable and the coefficient and the error term.

Regression is broadly divided into simple linear regression and multiple linear regression.

1. **Simple Linear Regression**: SLR is used when the dependent variable is predicted using only one independent variable. The equation for SLR is given by

Y = β0 + β1X.

2. **Multiple Linear Regression**: MLR is used when the dependent variable is predicted using multiple independent variables. The equation for MLR will be:

Y = β0 + β1X1 + β2X2 + β3X3 + …….

β1 = coefficient for X1 variable

β2 = coefficient for X2 variable

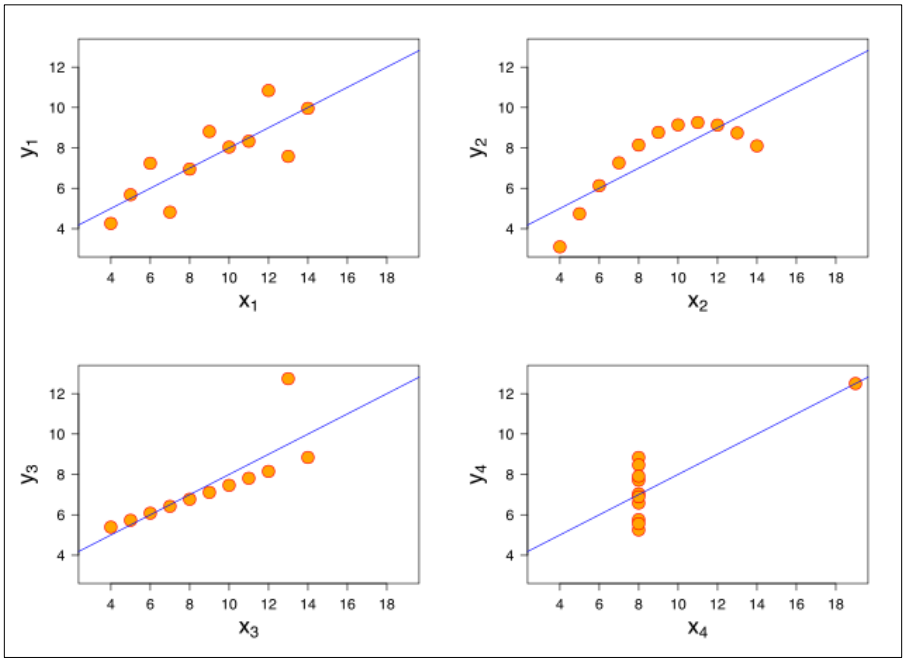
β3 = coefficient for X3 variable and so on…

β0 is the intercept (constant term).

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

Answer:

Anscombe's Quartet was developed by statistician Francis Anscombe. It includes four data sets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph. It was developed to emphasize both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties.



● The first scatter plot (top left) appears to be a simple linear relationship.

● The second graph (top right) is not distributed normally; while there is a relation between them, it's not linear.

● In the third graph (bottom left), the distribution is linear, but should have a different regression line The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.

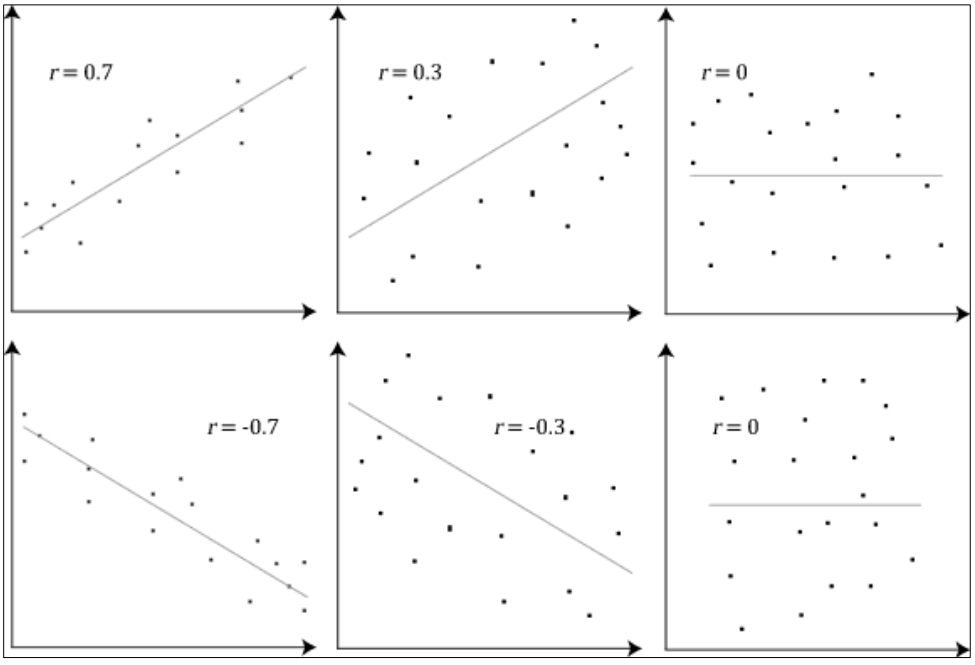
● Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset

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

Answer:

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.



The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A

value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below

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

Answer:

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

For Example: if an algorithm is not using feature scaling method, then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and, in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

Normalization/Min-Max Scaling:

- It brings all of the data in the range of 0 and 1.
- sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

MinMax Scaling : x = x – min(x) / max(x) - min(x)

Standardization Scaling:

- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).
- sklearn.preprocessing.scale helps to implement standardization in python.

Standardization: x = x – mean(x) / sd(x)

One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

Answer:

VIF - the variance inflation factor -The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity.

(VIF) =1/ (1-R_1^2).

If there is perfect correlation, then VIF = infinity, where R-1 is the R-square value of that independent variable which we want to check how well this independent variable is explained well by other independent variables. If that independent variable can be explained perfectly 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”.

To solve this, we need to drop one of the variables from the dataset which is causing this

perfect multicollinearity.

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

Answer:

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

Use of Q-Q plot:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Importance of Q-Q plot:

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.