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

   Answer:

   I have done analysis on categorical columns using the boxplot and bar plot. Below are the few points we can infer from the visualization

   - Fall season seems to have attracted more booking. And, in each season the booking count has increased drastically from 2018 to 2019.
   - Most of the bookings has been done during the month of may, june, july, aug, sep and oct. Trend increased starting of the year till mid of the year and then it started decreasing as we approached the end of year.
   - Clear weather attracted more booking which seems obvious.
   - Thu, Fir, Sat and Sun have more number of bookings as compared to the start of the week.
   - When it's not holiday, booking seems to be less in number which seems reasonable as on holidays, people may want to spend time at home and enjoy with family.
   - Booking seemed to be almost equal either on working day or non-working day.
   - 2019 attracted more number of booking from the previous year, which shows good progress in terms of business.

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

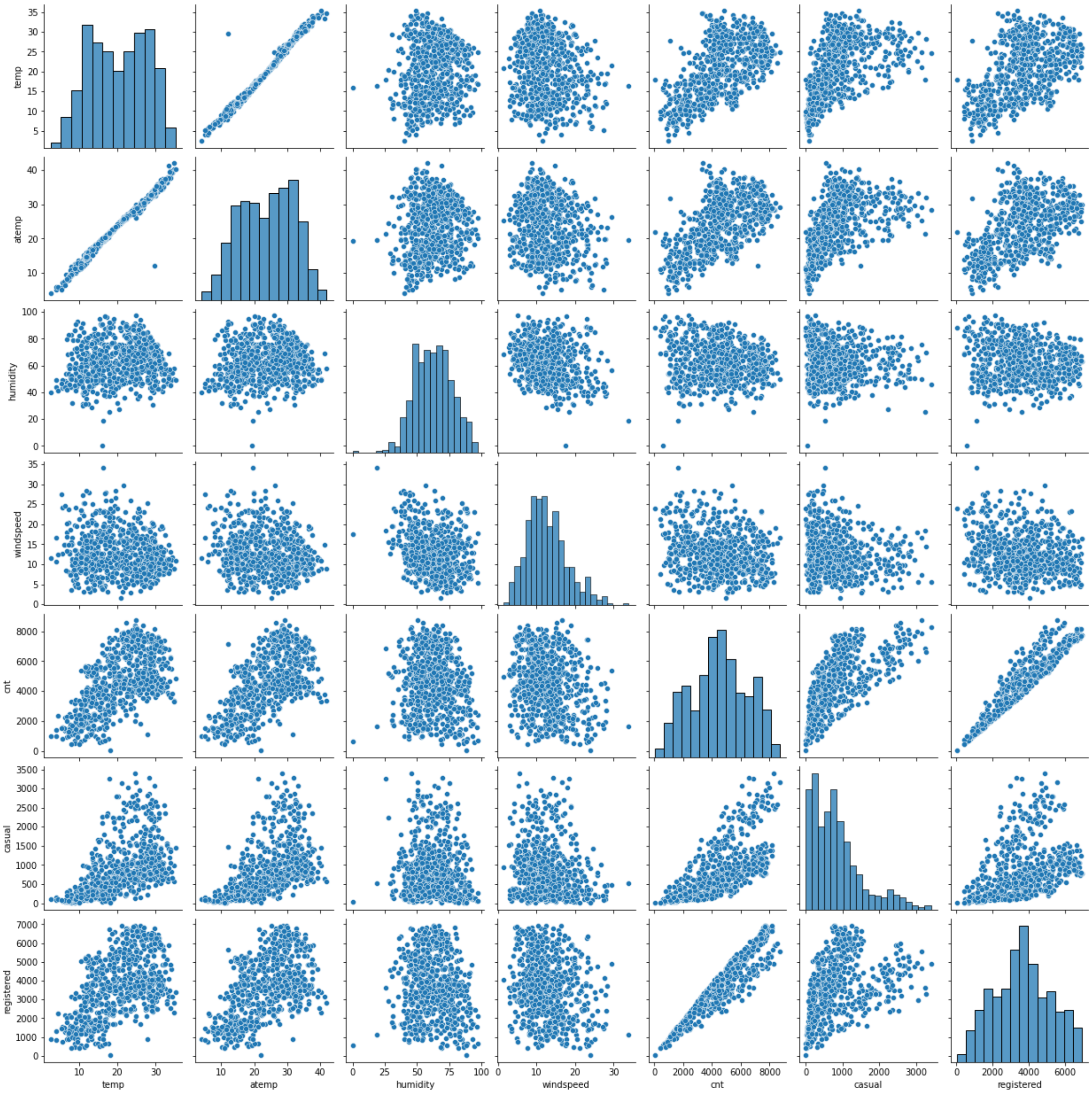
   Answer:

   drop_first = True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

   Syntax –

   > drop_first: bool, default False, which implies whether to get k-1 dummies out of k categorical levels by removing the first level.
   >
   > Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not A and B, then It is obvious C. So we do not need 3rd variable to identify the C.

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



⟶ Looking at the pairplot, it is quite evident that there is a clear trend between the temp, atemp variables and the target variable, count.

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

Answer :

I have validated the assumption of Linear Regression Model based on below 5 assumptions

- Normality of error terms
    - o Error terms should be normally distributed
- Multicollinearity check
    - o There should be insignificant multicollinearity among variables.

- Linear relationship validation
    - o Linearity should be visible among variables
- Homoscedasticity
    - o There should be no visible pattern in residual values.
- Independence of residuals
    - o No auto-correlation

5. **. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**

Answer :

Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes –

- temp
- winter
- sep

General Subjective Questions :

1. **Explain the linear regression algorithm in detail**

Answer:

Linear regression may be defined as the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

Mathematically the relationship can be represented with the help of following equation

Y = mX + c

Here, Y is the dependent variable we are trying to predict.

X is the independent variable we are using to make predictions

. m is the slope of the regression line which represents the effect X has on Y

c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

Furthermore, the linear relationship can be positive or negative in nature as explained below−

❖ Positive Linear Relationship:

- A linear relationship will be called positive if both independent and dependent variable increases. It can be understood with the help of following graph

❖ Negative Linear relationship:

- A linear relationship will be called positive if independent increases and dependent variable decreases. It can be understood with the help of following graph –

Linear regression is of the following two types –

¬ Simple Linear Regression

¬ Multiple Linear Regression

Assumptions –

The following are some assumptions about dataset that is made by Linear Regression model –

- Multi-collinearity –

o Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.

- Auto-correlation –

o Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.

- Relationship between variables –

o Linear regression model assumes that the relationship between response and feature variables must be linear.

- Normality of error terms –

o Error terms should be normally distributed

- Homoscedasticity –

o There should be no visible pattern in residual values.

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

Answer :

⟶ Anscombe's Quartet is a concept in Statistics which explains the importance of visualizing data in addition to looking at the Summary Statistics.

Summary Statistics of any data gives us the big picture rather than showing the value of each datapoint and allowing us to intuitively estimate the ranges and values of the data.

Summary Statistics consists of calculating values like the Average (mean), Median, Mode (if categorical), and the spread of data — minimum value, 25th percentile, 50th percentile (same as median), 75th percentile and maximum value — if it is a numerical column.

While it does give us a good idea about the data, it doesn't really help us look at the shape or the distribution of data.

Anscombe's Quartet is a collection of 4 different datasets with different individual data points have the same average values for the data yet widely different distributions (shape).

In the above picture, the summary statistics of the 4 distinct dataset have identical values - 7.50 for mean of the data, 1.94 for Standard Deviation and 0.82 for the Correlation Coefficient - yet they have vastly different distributions.

This is the crux of the concept of Anscombe's Quartet - How summary statistics don't give the full picture, literally!

3. What is Pearson’s R?

Answer :

⟶ Pearson's R is a statistical measure that is used to determine the measure of the strength of association between two numerical and Linear Variables. Pearson's correlation coefficient is usually calculated by plotting the values of the independent variable of a sample on the x-axis and the corresponding values of the dependent variable of the sample on the y-axis. Note that, the variables strictly don't have to be dependent on one and another. After plotting the values on the graph, the covariance values are calculated by the formula

$Cov(x, y) = \Sigma(xi - xbar)(yi - ybar)/N - 1$

where xi - X value of an individual data point

yi - y value of an individual data point

x bar - Mean of X

y bar - Mean of y

Once the covariance value is calculated, the Correlation coefficient is calculated by dividing the value of covariance with the standard deviation of X and Y.

σx - Standard Deviation of X

σy - Standard Deviation of y

The Correlation coefficient tells us if there is a strong positive relationship, strong negative relationship, weak positive relationship, weak negative relationship or no relationship between the 2 variables. The Correlation Coefficient value would only range in between 1 to1

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

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.

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.

S.NO.	Normalized scaling	Standardized scaling
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

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

Answer:

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) 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.**
   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.