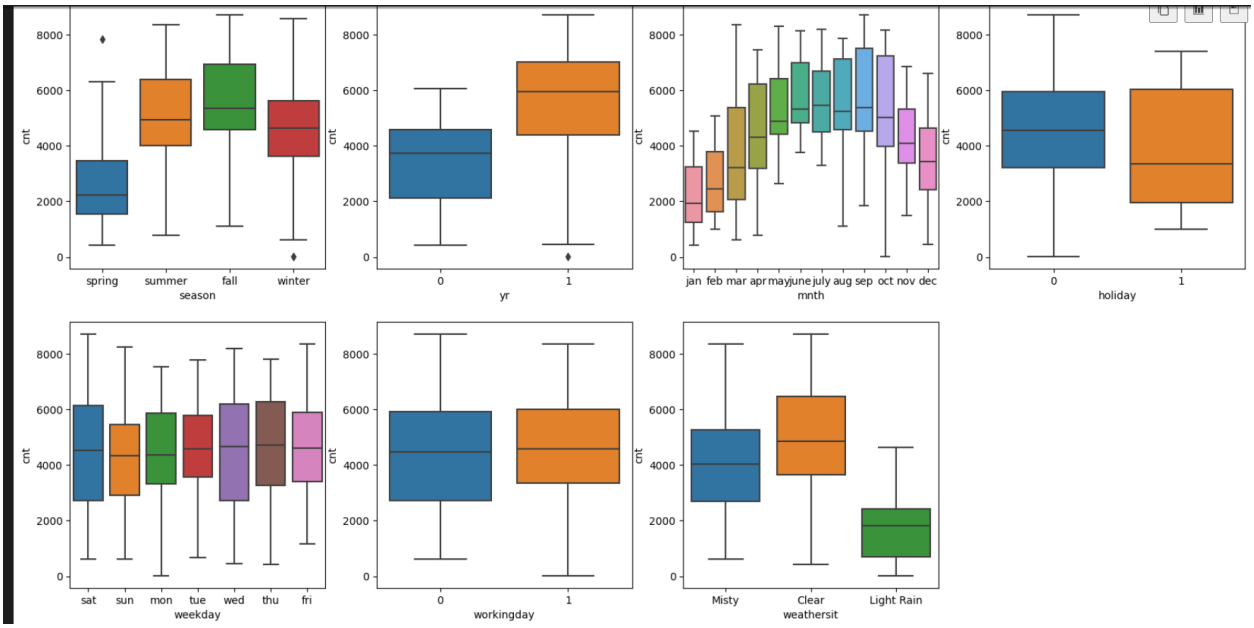
Assignment-based Subjective Questions

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

Answer:

Based on the EDA of Bivariate analysis on Categorical varibles,

- Season (Fall seem to have more bookings)
- The next year the booking numbers is grown
- Booking in Jun, Jul, Aug, Sep and Oct are high. Sep has high no of booking among them
- Year-end and Year-Beginning have less amount of bookings
- When its not holiday booking has less numbers
- Fri, Sat and Sunday has more no of bookings
- In Clear weather set, The booking numbers are increasing.



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

Answer:

It is mandatory to use drop_first-True, while creating dummy variables in order to not having the extra columns

Syntax:

```
season_d = pd.get_dummies (data_bike['season'], drop_first=True).astype (int)
mnth_d = pd.get_dummies (data_bike['mnth'], drop_first=True).astype (int)
```

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

Answer:

Target Variable is "cnt", And it has a high correlation with "temp, atemp" variable.

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

Answer:

List of validations for the assumptions that handled in the analysis:

- Error terms (Should be normally distributed)
- Linear Relationship (Visibility of linearity should be there among the variables)
- Multicollinearity (multicollinearity among variables is insignificantly)
- Homoscedasticity (variance of the residual values is constant)
- Autocorrelation (Residuals should not be independent of each other)

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

Answer:

- temp
- sep (month)
- fall (season)

==========

General Subjective Questions

1. Explain the linear regression algorithm in detail?

Answer:

Linear regression algorithm is an analysis used to predict the value of a variable based on the value of another variable. The dependent variable is the one you want to be able to forecast. The independent variable is the one you're using to make a prediction about the value of the other variable.

When modeling the relationship between a scalar answer and one or more explanatory factors (also known as dependent and independent variables) in statistics, linear regression is a linear approach. Simple linear regression is used when there is only one explanatory variable, and multiple linear regression is used when there are numerous variables.

The best-fit line for a set of paired data can be found using straightforward linear regression calculators that employ the "least squares" technique.

Mathematically the relationship can be defined by "y = mX+c"

y - Dependent variable

X - Independent variable

m - regression line slope X on y

c - Constant, y-intercept.

Types:

1. Simple linear Regression

2. Multiple linear Regression

1. Positive linearity: Increase in both Dependent variables.
2. Negative linearity: Independent variables increases, Dependent variable value decreases.

Assumption:

- Error terms
- Linearity
- Multicollinearity
- Homoscedasticity
- Autocorrelation

2. Explain the Anscombe’s quartet in detail.

Answer:

Anscombe's quartet is a collection of four data sets that, when plotted separately, have idiosyncrasies that trick the regression model despite being almost similar in terms of simple descriptive statistics. It demonstrates both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties.

It's often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

It is made up of four data sets, each of which has eleven (x,y) points. The fundamental characteristic of these data sets that needs to be analyzed is that they all have distinct graphical representations but the same descriptive statistics (mean, variance, standard deviation, etc.)

Statistical Formula:

mean of x = 9, mean of y = 7.50

Variance of x = 11, Variance of y =4.12

Correlation Coefficient = 0.816

Linear Regression Equation : y = 0.5 x + 3

3. What is Pearson’s R?

Answer:

The most popular method for determining a linear connection is the Pearson correlation coefficient (r). The intensity and direction of the link between two variables is expressed as a number between -1 and 1.

Formula:

$p(X, Y) = cov(X, Y)/ \sigma X * \sigma Y$
cov -> Coveriance
σX -> Standard deviation of X
σY -> Standard deviation of Y

$cov(X, Y) = E[(X-\mu X)/ (Y-\mu Y)]$

μX -> mean of X
μY -> mean of Y
E -> Expectation

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

Answer:

Scaling:

It is a data pre-processing procedure used to normalize data within a specific range by applying it to independent variables. Additionally, it aids in accelerating algorithmic calculations.

Why Scaling is performed:

Most of the time, the obtained data set includes characteristics that vary greatly in magnitudes, units, and range. If scaling is not done, the algorithm will only consider magnitude and not units, which will result in inaccurate modeling. We must scale all the variables to the same degree of magnitude in order to resolve this problem.

Normalized Scaling	Standardized Scaling
Scales min & max values	Scales mean & standard deviation
Scale falls between -1 to 1	Values on scale not constrained to a defined range
Also called as Scaling normalization	Also called as Z-Score normalization
Feature distribution is unclear	Feature distribution is clean
Python pkg: sklearn.preprocessing.MinMaxScaler	Python pkg: sklearn.preprocessing.scale
Formula: x = [x - min(x)]/ max(x)-min(x)	Formula: x = [x-mean(x)]/ sd(x)

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

Answer:

If every independent variable is perpendicular to every other independent variable, then VIF = 1.0. In the case of perfect correlation, VIF equals infinite. There is a correlation between the variables if VIF has a high value.

In the exact correlation, we will get R2 as 1.0, this leads to 1/ (1-R2) as infinity.

To overcome this, we need to drop the one of the variables from the dataset causing this correlation.

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

Answer:
Quantile-Quantile (Q-Q) plot, A graphical method for assessing if two data sets originate from populations with a common distribution. The quantiles of the first data set are plotted against the quantiles of the second data set in a q-q graphic.

Use of Q-Q plot:

Knowing whether the distribution is normal or not is important so that we can apply different statistical techniques to the data and interpret it in a way that makes sense to people. This is where the Q-Q plot comes into picture.

Showing in a graphic the similarities and differences between the two distributions' position, scale, and skewness. Theoretical distributions or sets of data can be compared using Q-Q graphs.

Importance of Q-Q plot:

TO say whether it is a Pareto distribution, Gaussian distribution, uniform distribution, exponential distribution, etc. Simply by glancing at the plot, you may determine the type of distribution using the Q-Q plot's power.