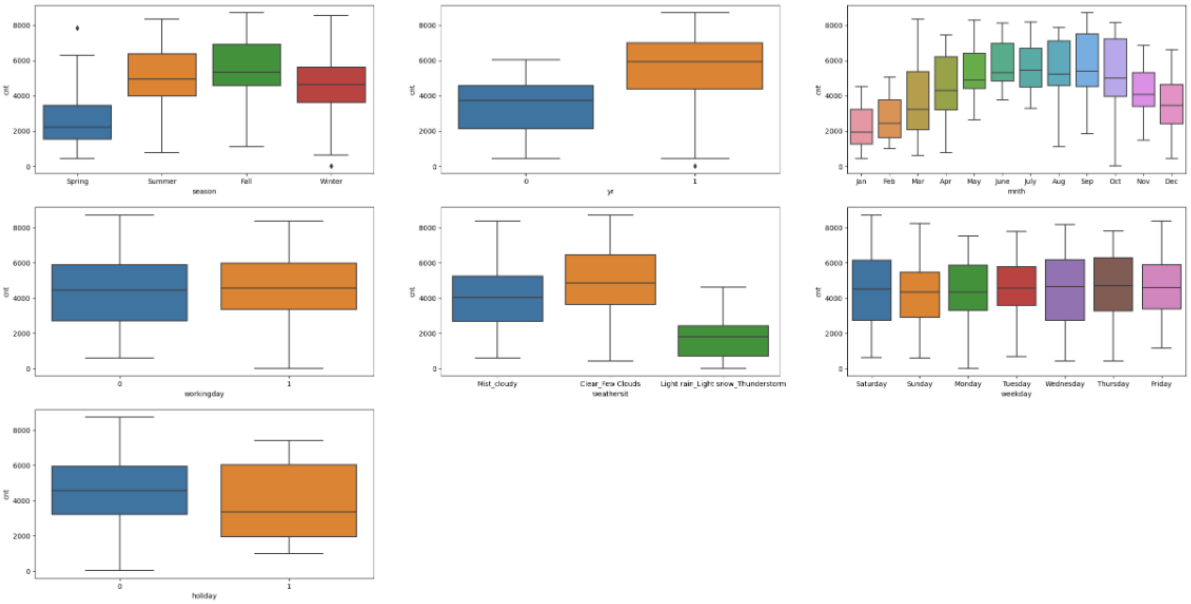
Assignment-based Subjective Questions:

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

Ans >>

We have 7 categorical `variables ['season', 'yr', 'mnth', 'weathersit', 'holiday', 'weekday', 'workingday']`

To examine the impact of the categorical variables on target variable 'cnt' used the boxplots:



The following conclusions can be drawn from the data:

a) The dependent variable ‘cnt’ is highly correlated/dependant on categorical variable such as season, weather, workingday etc.

b) season: Season 3/Fall had the highest proportion of bike bookings, with 32% of the total and a median of over 5000 bookings for the two-year period. Season 2/Summer and season 4/Winter followed with 27% and 25% respectively. This suggests that season is a good predictor for the dependent variable.

c) mnth: The months May, June, July, Aug, and Sept had the most bike bookings, with 10% each and a median of over4000bookings per month. This suggests that mnth has some trend for bookings and can be a good predictor for the dependent variable.

d) weathersit: Weathersit 1 had the most bike bookings, with 67% of the total and a median of close to 5000 bookings for the two-year period. Weathersit 2 followed with 30% of the total. This suggests that weathersit has some influence on the bike bookings and can be a good predictor for the dependent variable.

e) holiday: Only 2.4% of the bike bookings occurred on holidays, which means this data is clearly biased. This suggests that holiday cannot be a good predictor for the dependent variable.
f) weekday: Weekday variable showed a similar trend (between13.5%-14.8%of the total bookings on all days of the week) with their independent medians between 4000 to 5000 bookings. This variable may have some or no effect on the predictor. I will let the model decide if this needs to be added or not.
g) workingday: Workingday had 69% of the bike bookings, with a median of close to 5000 bookings for the two-year period. This suggests that workingday can be a good predictor for the dependent variable.
h) yr: Year also had a strong correlation with cnt, which can be seen from the boxplot diagram.

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

Ans >>

Using drop_first=True during dummy variable creation is important because it helps to avoid the dummy variable trap, which is a situation where one or more of the dummy variables are redundant and can be predicted by the others.

This can cause multicollinearity, which is a problem for some regression models that assume the predictor variables are independent of each other.

By dropping the first column, we reduce the number of dummy variables by one and ensure that they are not perfectly correlated. This way, we avoid the dummy variable trap and multicollinearity.

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

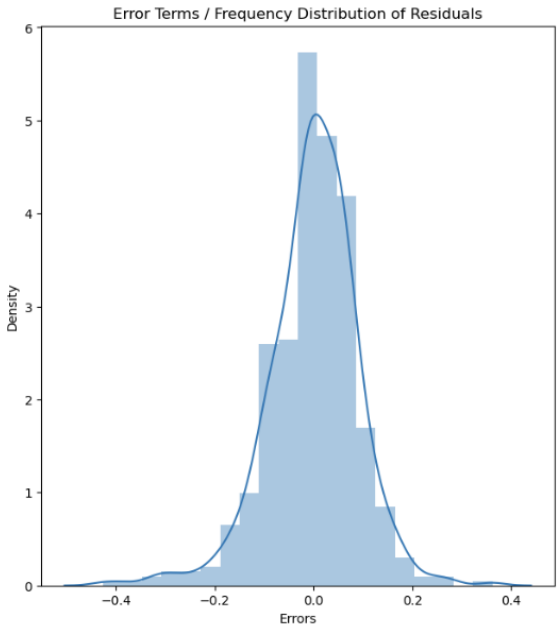
Ans >>

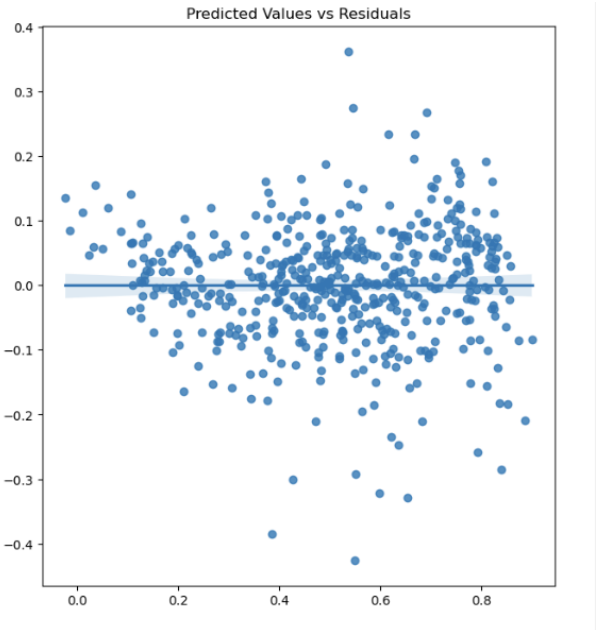
The pairplot shows a linear relationship between temp, atemp, and the target variable 'cnt'. This means that as the temperature (either actual or felt) increases, so does the number of bike bookings.

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

Ans >>

a) The errors have a normal distribution.

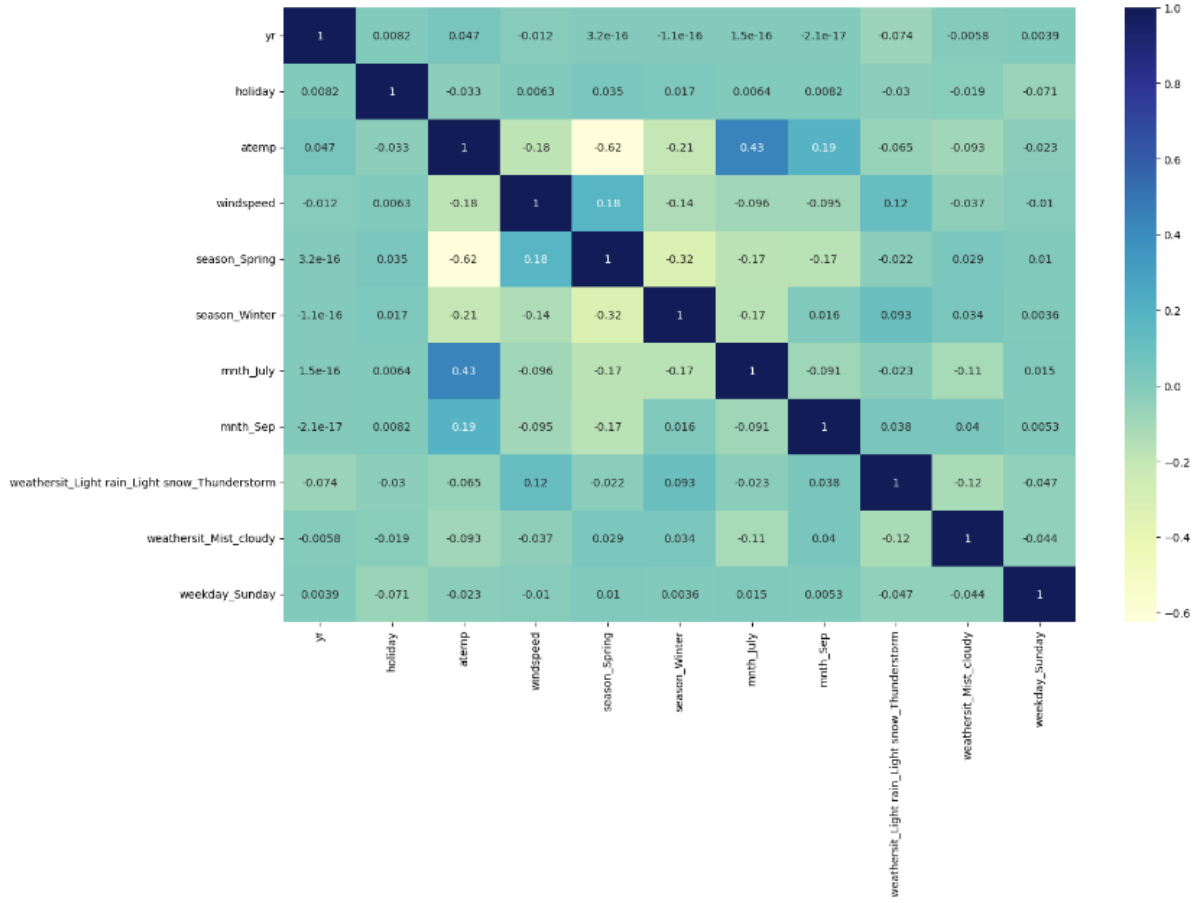




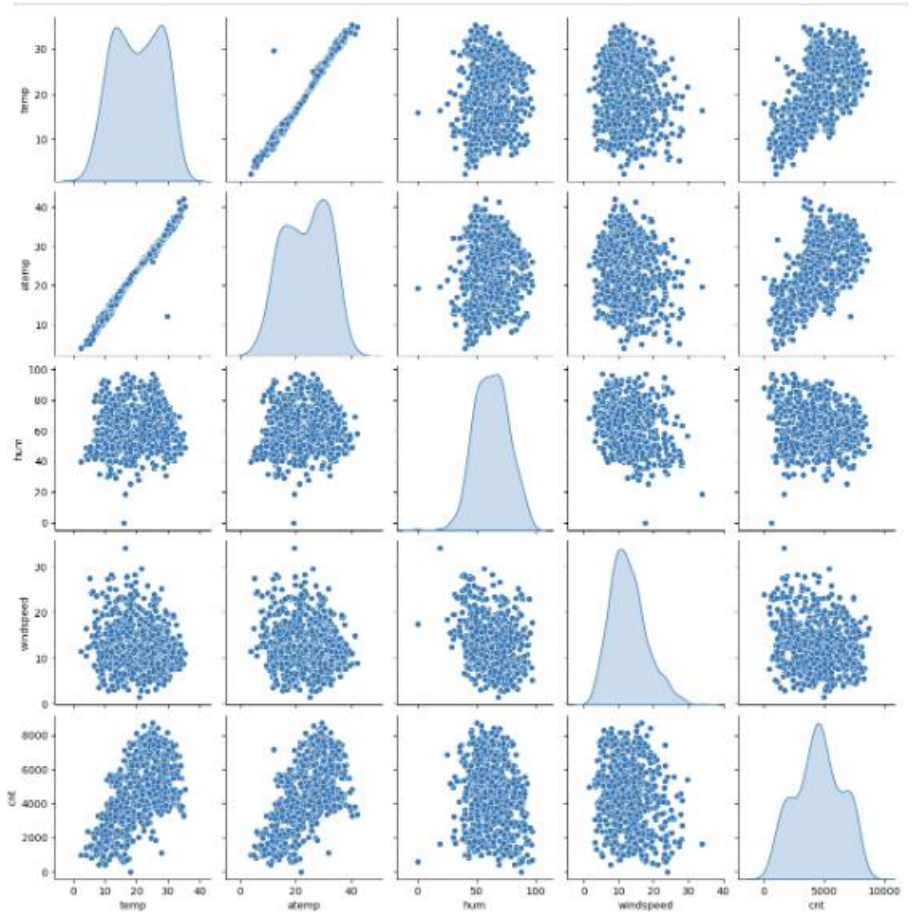
b) The predictor variables are not correlated with each other. (No Multicollinearity).

Checking the correlations using final pred variables

```
plt.figure(figsize = (16, 10))
sns.heatmap(df_subset[cols].corr(), annot=True, cmap="YlGnBu")
plt.show()
```



c) There is a linear relationship between temp, atemp and cnt or in other words the number of bike bookings increases with the temperature (both actual and felt).



The Pair-Plot graphs show that there is a linear relationship between temp, atemp and cnt (target var).

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

And >>

The features that have the highest positive correlation with the target variable are atemp/temp, yr, season_winter and mnth_Sep, with correlation values of 0.4632, 0.2350, 0.0412 and 0.0587 respectively.

The following factors have a significant impact on the bike bookings: (considering positive coef only):

- temp: The users prefer to ride bikes when the temperature is moderate and comfortable.
- yr: Demand increases(as coefficient is positive) in case of yr
- season: The company should target seasons, when the demand is higher based on positive coef's.
- weather: The users prefer to ride bikes when the weather is pleasant and clear.

We can see that the equation for best fitted line is: `cnt = 0.2620 + 0.2350 X yr - 0.1028 X holiday + 0.4632 X atemp - 0.1254 X windspeed -0.1167 X season_Spring + 0.0412 X season_Winter - 0.0657 X mnth_July + 0.0587 X mnth_Sep - 0.2872 X`

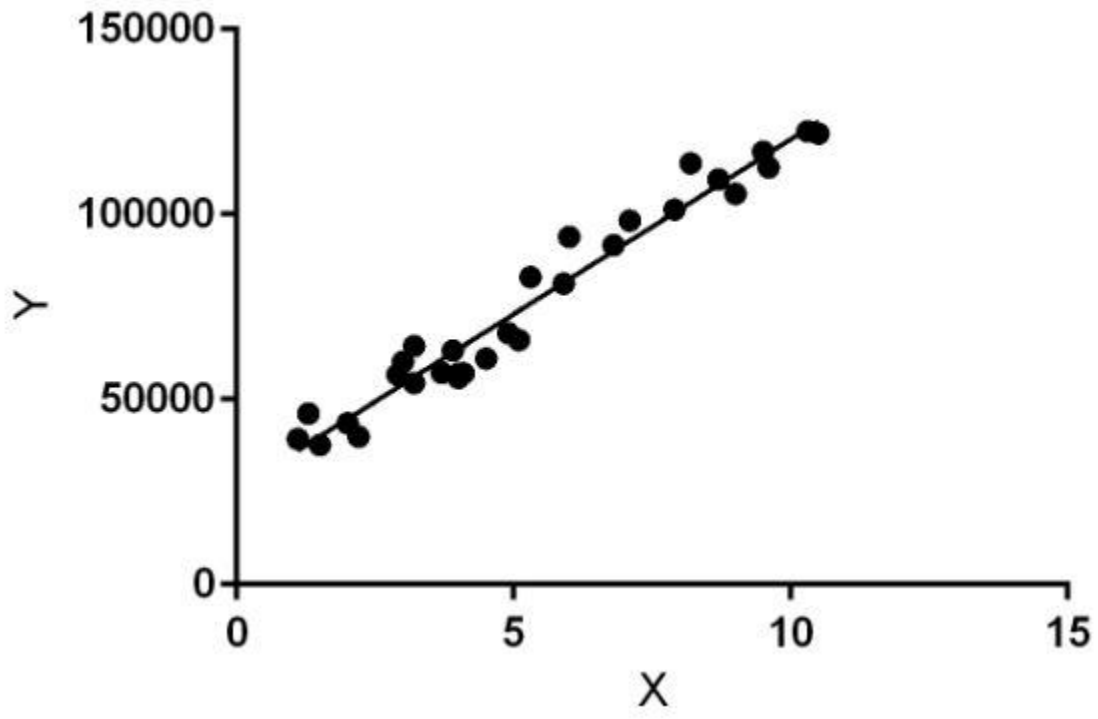
```
weathersit_Light rain_Light snow_Thunderstorm - 0.0837 X
weathersit_Mist_cloudy -0.0484 X weekday_Sunday
```

General Subjective Questions:

1. Explain the linear regression algorithm in detail.

Ans >>

In simple terms, linear regression is a method of finding the best straight line fitting to the given data,i.e. finding the best linear relationship between independent and dependent variables. In technical terms,linear regression is a machine learning algorithm that finds the best linear-fit relationship on any given data,between independent and dependent variables. It is mostly done by Sum Of Squared Residuals Method.



The line in the above graph is referred as the best fit straight line. Based on the given data points we try to plot a line that models the points the best.

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

Y = mX + b

Here, Y is the dependent variable we are trying to predict
X is the dependent variable we are using to make predictions.
m is the slop of the regression line which represents the effect X has on Y
b is a constant, known as the Y-intercept. If X = 0, Y would be equal to b

Assumption for Linear Regression Model

- Linear regression is a powerful tool for understanding and predicting the behavior of a variable, however, it needs to meet a few conditions in order to be accurate and dependable solutions.
- Linearity: The independent and dependent variables have a linear relationship with one another. This implies that changes in the dependent variable follow those in the independent variable(s) in a linear fashion.
- Independence: The observations in the dataset are independent of each other. This means that the value of the dependent variable for one observation does not depend on the value of the dependent variable for another observation.
- Homoscedasticity: Across all levels of the independent variable(s), the variance of the errors is constant. This indicates that the amount of the independent variable(s) has no impact on the variance of the errors.
- Normality: The errors in the model are normally distributed.
- No multicollinearity: There is no high correlation between the independent variables. This indicates that there is little or no correlation between the independent variables.

2. Explain the Anscombe's quartet in detail

Ans >>

Anscombe's Quartet is a set of four data sets that have similar simple descriptive statistics, but they are very different in their shapes and patterns when plotted on scatter plots. They show that regression models can be misled by the data if they do not consider the underlying distribution and outliers.

3. What is Pearson's R?

Ans >>

The Pearson correlation coefficient, r, measures the strength and direction of the linear relationship between two continuous variables. The value of r ranges from -1 to 1. The closer r is to 1 or -1, the stronger the linear relationship is. The sign of r indicates the direction of the relationship. A positive r means that the variables tend to increase together. A negative r means that the variables tend to decrease together. A perfect linear relationship (r=-1 or r=1) means that one variable can be perfectly predicted by a linear function of the other variable.

The assumptions for a Pearson correlation are: level of measurement, related pairs, absence of outliers, and linearity. Level of measurement refers to the type of data for each variable. For a Pearson correlation, both variables should be continuous.

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

Ans >>

Scaling is a data preprocessing technique that transforms the values of features or variables in a dataset to a similar scale. The purpose of scaling is to ensure that all features contribute equally to the model and to avoid the domination of features with larger values. Scaling can also improve the performance and convergence of some machine learning algorithms that rely on distance calculations or gradient descent optimization.

There are different methods of scaling, such as normalization and standardization :

- Normalization, also known as min-max scaling, rescales the values of a feature to a range between 0 and 1, or sometimes -1 and 1. This is done by subtracting the minimum value and dividing by the maximum minus the minimum value of each feature Normalization is useful when the features have different scales or units of measurement, and when there are no outliers in the data.
- Standardization, also known as z-score normalization, transforms the values of a feature such that they have a mean of 0 and a standard deviation of 1. This is done by subtracting the mean and dividing by the standard deviation of each feature. Standardization is useful when the features have a normal or Gaussian distribution, or when there are outliers in the data. Standardization does not have a fixed range for the transformed values, unlike normalization.

The difference between normalization and standardization is that normalization rescales the values of a feature to a fixed range, while standardization changes the mean and variance of a feature to 0 and 1 respectively. Normalization is more sensitive to outliers than standardization, and standardization preserves the shape of the distribution better than normalization. The choice of scaling method depends on the type of data and the algorithm used for modelling.

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

Ans >>

The value of VIF is infinite when there is perfect correlation between a given independent variable and other variables in the model. This means that the given variable can be perfectly predicted by a linear combination of the other variables. In other words, the given variable is redundant and does not add any new information to the model. This situation can cause problems for some regression methods that rely on the inverse of the covariance matrix, which becomes singular when there is perfect correlation.

To avoid infinite VIF values, one should check for multicollinearity among the independent variables and remove any variables that are highly correlated with others. Alternatively, one can use regularization methods such as ridge or lasso regression that can handle multicollinearity by shrinking the coefficients of correlated variables.

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

Ans >>

A Q-Q plot, or quantile-quantile plot, is a graphical tool that helps us assess if a set of data plausibly came from some theoretical distribution, such as a normal, exponential, or uniform distribution. It can also help us compare if two data sets come from populations with a common distribution.

A Q-Q plot is created by plotting the quantiles of the first data set against the quantiles of the second data set. If the points lie approximately on a straight line, it means that the two data sets have similar distributions. The slope and intercept of the line indicate the relative location and scale of the two data sets.

In linear regression, a Q-Q plot is often used to check the normality assumption of the error terms or residuals. By plotting the standardized residuals against the theoretical quantiles of a standard normal distribution, we can see if the residuals are normally distributed. If the residuals are not normally distributed, it implies that the standard confidence intervals and significance tests for the regression coefficients may be invalid.

A Q-Q plot is important in linear regression because it can help us diagnose potential problems with our model, such as outliers, skewness, heteroscedasticity, or non-linearity. It can also help us decide if we need to transform our data or use a different regression method to improve our model fit. Source/More Info : Q–Q plot - Wikipedia