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

There are below four categorical variables in the dataset:

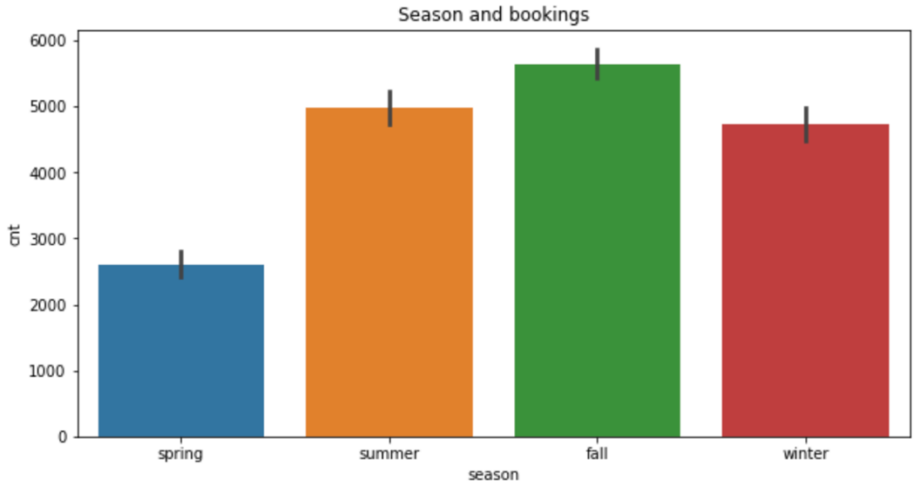
1. season
2. weekday
3. weathersit
4. mnth

Summer:

The season variable explains that most of the bookings were done in fall, followed by summer then winter and season spring saw the least bookings.

```
In [1687]: plt.figure(figsize=(10, 5))

           plt.title('Season and bookings')
           sns.barplot(x='season', y='cnt', data=data)
           plt.show()
```

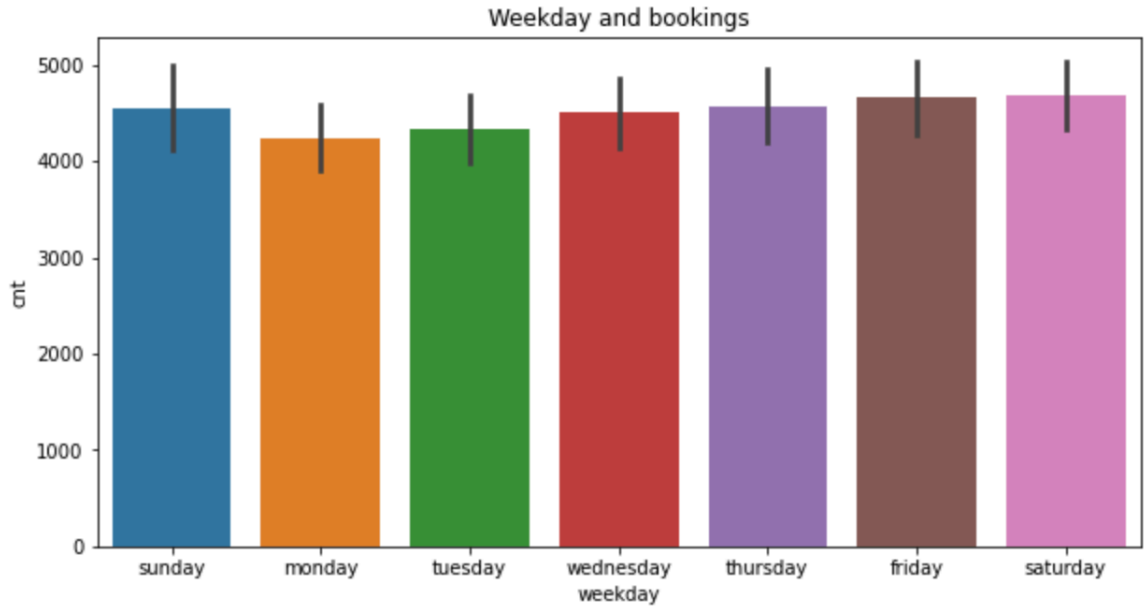


Weekday:

In weekdays variable it looks like bookings for the days - Friday, Saturday and Sunday are pretty close and Monday has the least booking.

We can also see the rise in booking as soon we reach close to weekends

```
plt.figure(figsize=(10, 5))
plt.title('Weekday and bookings')
sns.barplot(x='weekday', y='cnt', data=data)
plt.show()
```

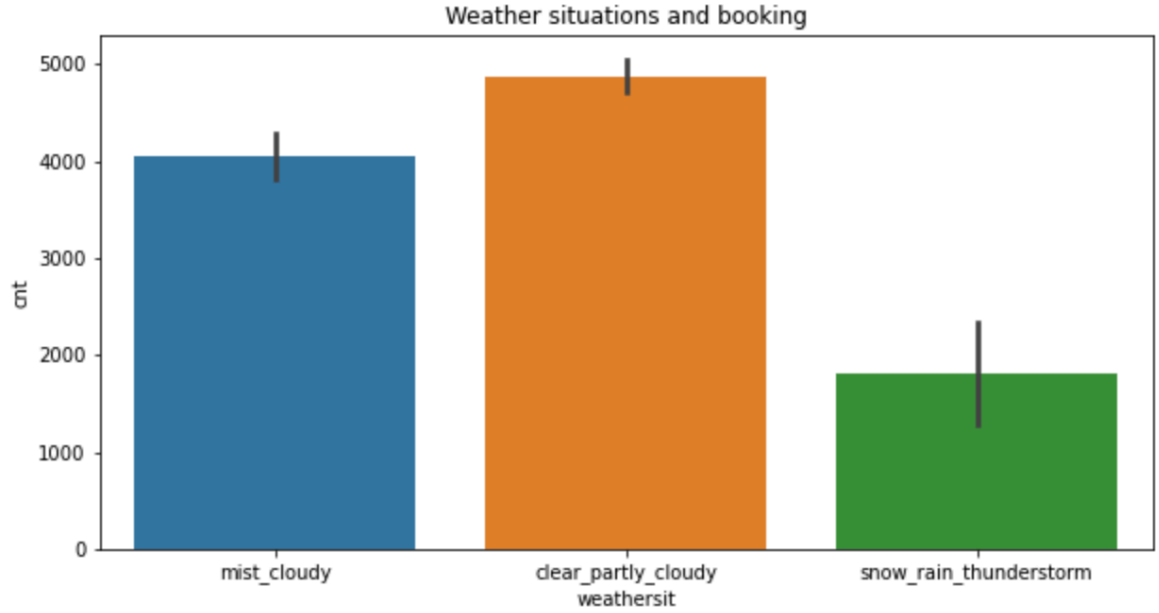


Weathersit:

Looks like customers likes to make booking when the weather is clear or partly cloudy followed by when it is mist and cloudy

During heavy rains and thunderstorms, lowest bookings were made

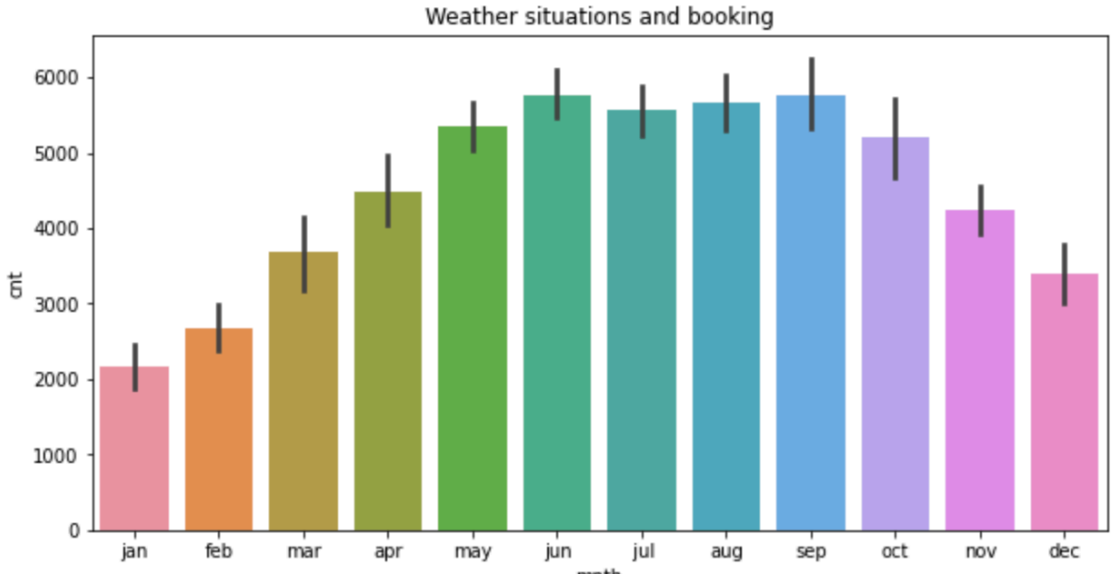
```
plt.figure(figsize=(10, 5))
plt.title('Weather situations and booking')
sns.barplot(x='weathersit', y='cnt', data=data)
plt.show()
```



Mnth:

Bookings were made mostly in the month of June and August and Jan has the least bookings.

```
plt.figure(figsize=(10, 5))
plt.title('Weather situations and booking')
sns.barplot(x='mnth', y='cnt', data=data)
plt.show()
```



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

When dealing with categorical variable, we convert them into numbers so that it can be understood by the program or model, in such cases, the categorical variables are converted in to zeros and ones.

We convert them by adding **k** variables equal to the **k** number of categorical levels and this will create columns with 1s and 0s. This is know as One Hot Encoding or Dummy Variables.

The param, drop_first=True simply converts k number of categories into k-1 as the k categories can be handled with k-1 variables.

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

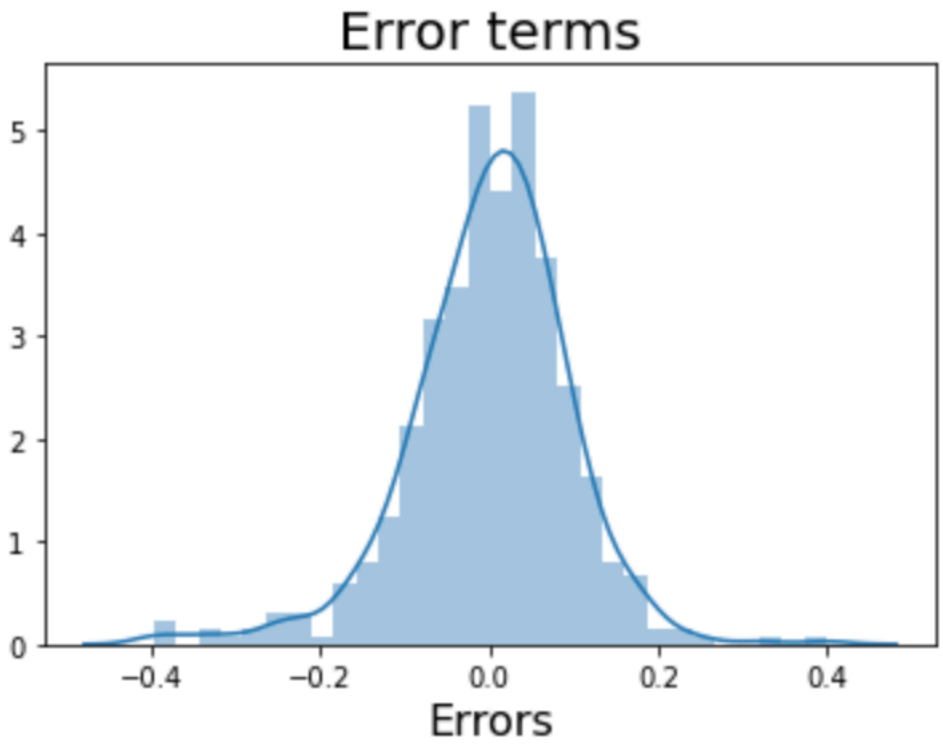
Variable **atemp** has the highest correlation with the target variable with the value of **0.63**

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

Residual analysis of of the prediction using the training data helped in evaluating the error terms:

```
y_train_pred = lm_11.predict(X_train_sm)
```

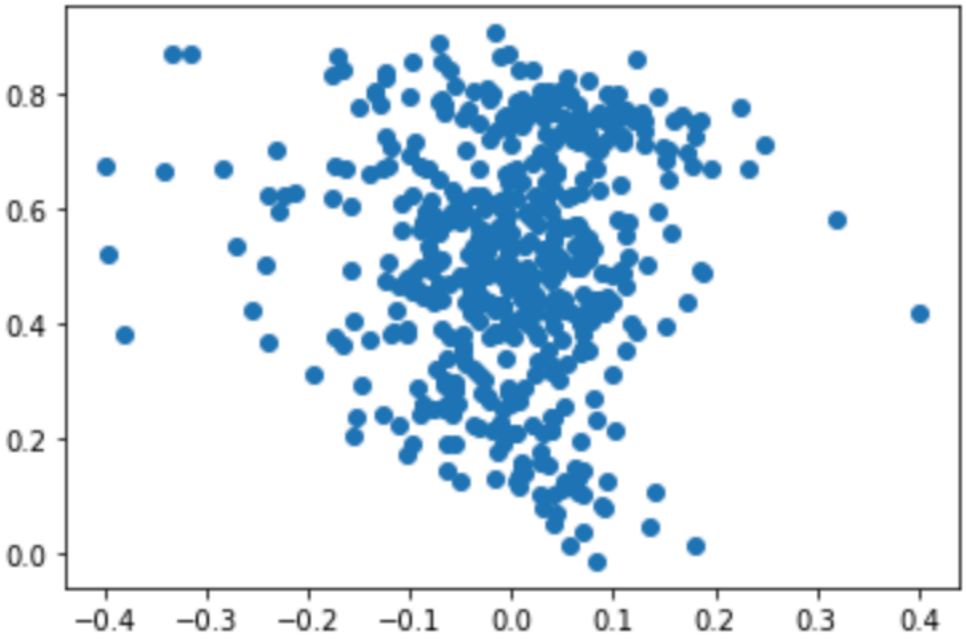
```
plt.figure()
sns.distplot((y_train - y_train_pred))
plt.title('Error terms', fontsize=20)
plt.xlabel('Errors', fontsize=16)
plt.show()
```



The linear relationship between the independent and target variable can be confirmed with the coefficients of the model. In simple, if the coef is positive, that means the target variable will increase in unit change of X and visa versa if the coef is negative.

We also confirmed that the error terms are independent of each other as we do not see any pattern here.

```
plt.scatter((y_train - y_train_pred), y_train_pred)
plt.show()
```



---

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

After the model is built, below are the variables that can be considered as the top 3 feature of the model.

1. atemp
2. yr
3. windspeed

General Subjective Questions.

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

Linear regression models the linear relationship between a target variable (dependent variable) and one or more independent variables.

Overall we examine below things in a regression model.

1. Does a set independent variables are good enough in predicting the dependent variables.
2. What all variables are significant enough to be included in the model for the outcome of the dependent variable.

Above estimates are used to explain the linear relationship between the dependent variable and the independent variable.

A simple linear regression can be explained with below equation.

y = mx + b

where,

m is the slope of the line or regression coef,
b is the constant
x is the independent variable and
y is the dependent variable.

The performance of a regression model is calculated by the params like R-squared, Adj R-Squared, P Values of the column.

MSE and RMSE can also be used to understand the loss in the model and lesser the MSE, better the model is.

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

Anscombe's quartet have four datasets with (x,y) pairs in each of them. An Anscombe's quartet share the same descriptive statistics.

It explains more about the importance of a variable graphically and not simply relying on the basic statistics.

3. **What is Pearson's R?**

   Pearson correlation coefficient (PCC) measures the linear co-relation between two variables.

   The value of Pearson's R lies between -1 to 1, where 1 indicates strong positive relationship with the Y variable and -1 indicates strong negative relationship with Y variable.

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

   Scaling is a mechanism of normalizing the range of the values that have different scales.

   Scaling is performed to make the model easy to understand as the coef of the equations will be within -1 to 1.

   In Standard Scaler will scale the values such that the data will have mean value of zero with a standard deviation of 1.

   In Normalized scaling the data points are scaled between the range of 0 to 1.

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

An infinite VIF tells that the variable maybe explained exactly by a linear combination of some other variable that will also show a VIF of inifinite.

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

QQ plot is used to compare two probability distribution by plotting their quantiles against each other.

QQ plot can help us understand in Linear Regression that the received Train and Test datasets are from populations with same distribution.