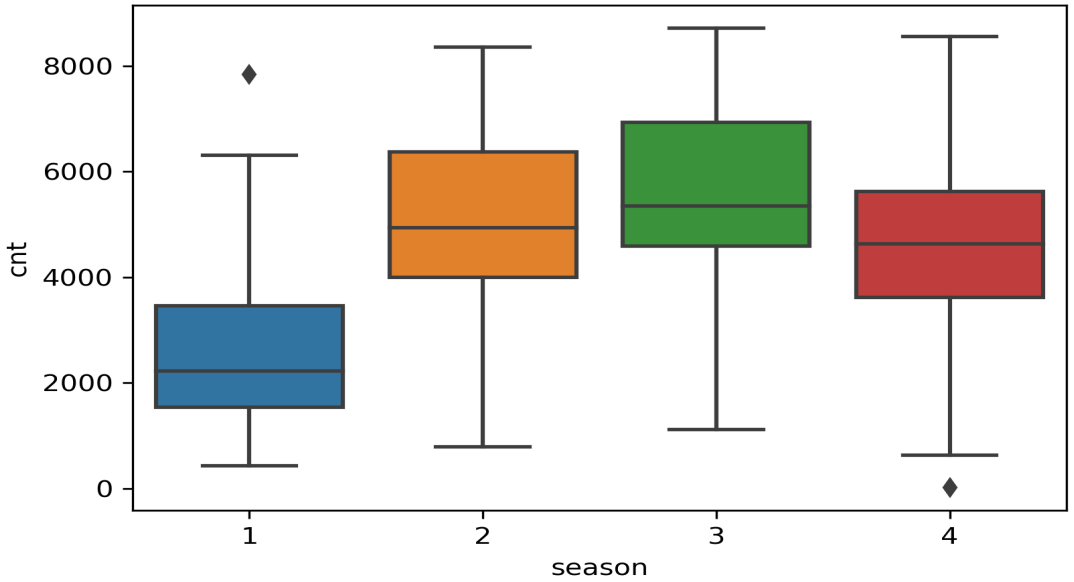
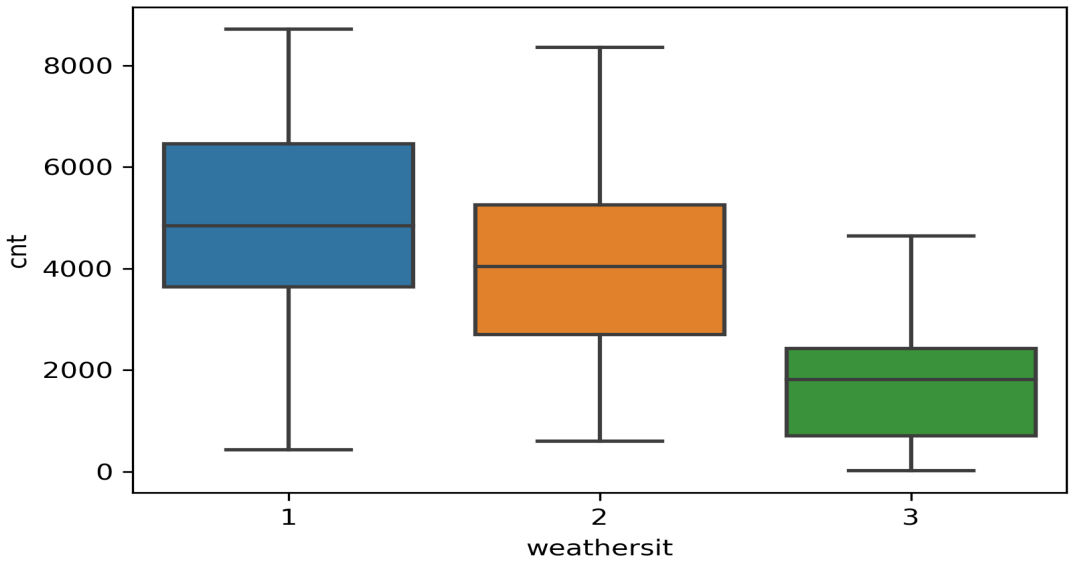
- From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

ANS-

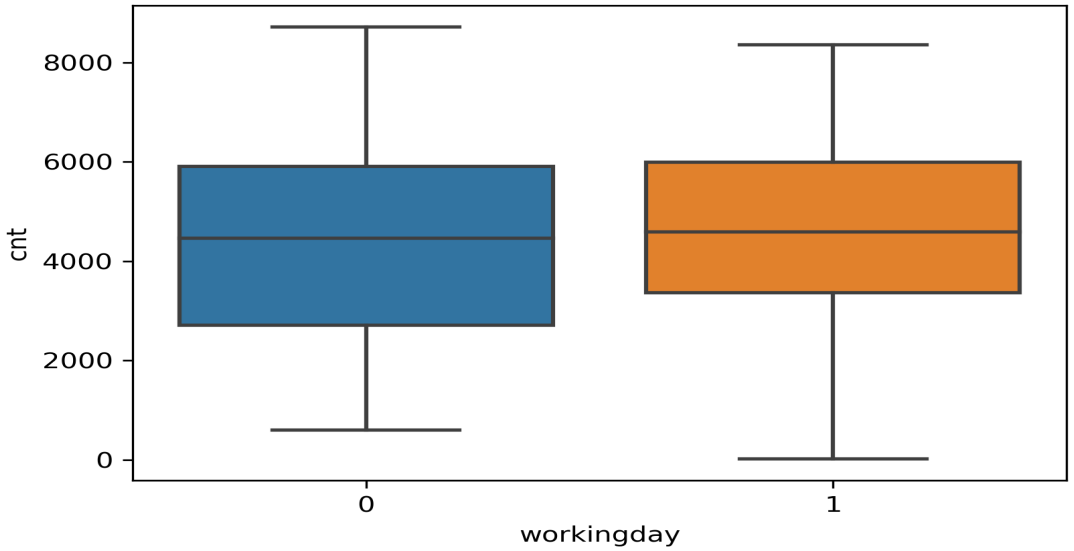


1:spring, 2:summer, 3:fall, 4:winter

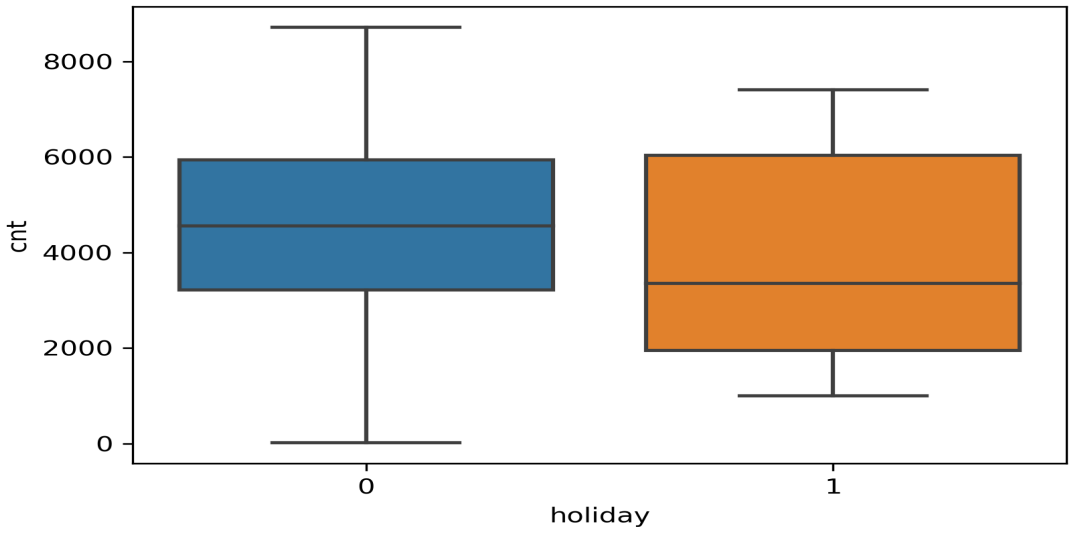
In fall the count of total rental bikes is highest followed by summer , winter and spring.



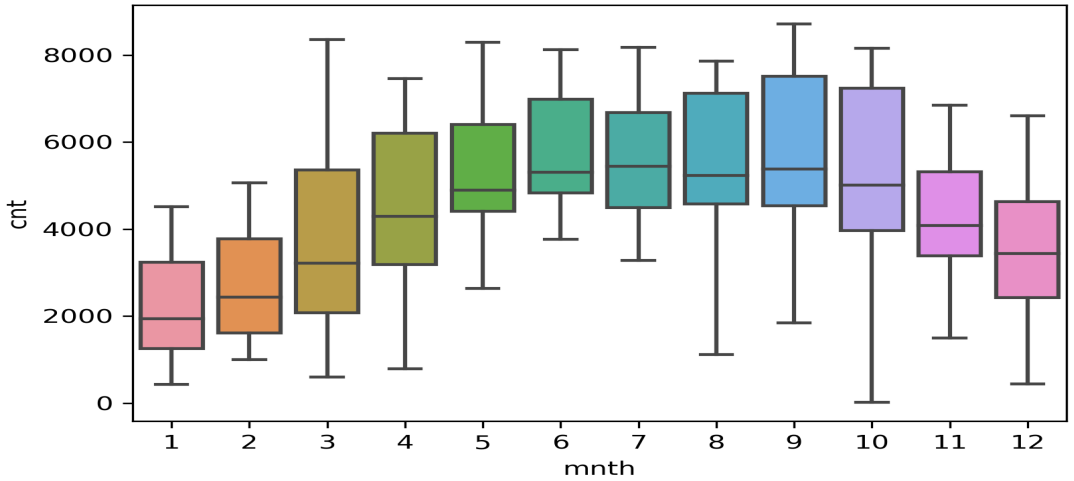
In clear weather CNT is highest followed by cloudy and bad weather.



Working days have slightly higher demand than holidays.

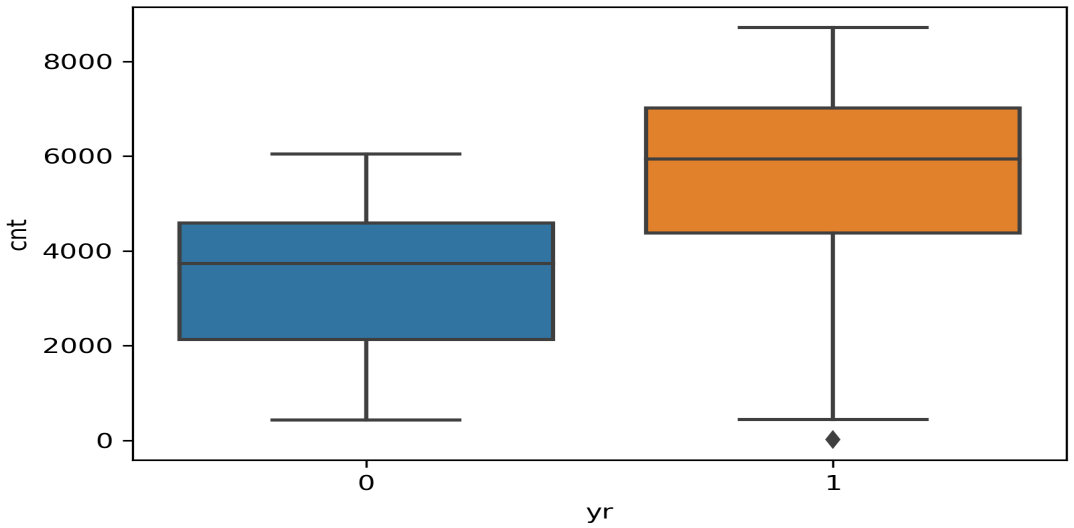


Holidays have slightly lower demand than working days.



month (1 to 12)

The distribution shows that the demand increases by mid year and decreases again.



0: 2018, 1:2019

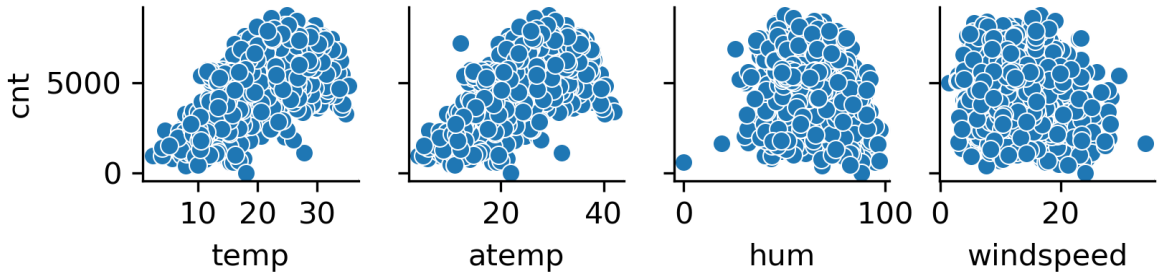
The demand has increased since the previous year.

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

ANS - It is important to use drop_first = True because that avoids creation of extra columns thus reducing the correlation among dummy variables which would be unnecessary.

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

ANS-



Temperature has the highest correlation with the target variable.

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

ANS-

VARIANCE INFLATION FACTOR CODE

```
from statsmodels.stats.outliers_influence import variance_inflation_factor
vif = pd.DataFrame()
vif['Features'] = x_train.columns
vif['VIF'] = [variance_inflation_factor(x_train.values , i) for i in
range(x_train.shape[1])]
vif['VIF'] = round(vif['VIF'],2)
```

P Value CODE

```
x_train_sm = sm.add_constant(x_train)

lr = sm.OLS(y_train, x_train_sm)
lr_model = lr.fit()
lr_model.summary()
```

Significance Value

- High P-value, High VIF(Not Recommended)
- Low P value, Low VIF (Significant)
- High-Low:
  - High P, low VIF (Not Recommended)
  - Low p, High VIF(Significant)

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

ANS- The top 3 features contributing significantly towards demand are:

1. Weathersit
2. Weekdays
3. Temperature

1. Explain the linear regression algorithm in detail.

ANS-

1. **Reading the data:** import pandas and other important libraries for reading the data.
2. **Analysis of data(Exploratory Data Analysis):** Analysis of the correlation between different variables.
3. **Data Preparation:** Converting strings to numeric values using dummy function.
4. **Splitting the Data into Training and Testing Sets:**
5. **Building a linear model**
6. **Residual Analysis of the train data**
7. **Making Predictions Using the Final Model**:Extrapolating the graph.
8. **Model Evaluation:** Plotting the y test and y predicted to find the significant variables.

2. Explain the Anscombe's quartet in detail.

ANS-

Anscombe's Quartet are variables that are similar in statistics but when plotted in scatter plots they behave differently and show some peculiarities. It is a set of four dataset.

3. What is Pearson's R?

ANS- It is the ratio of covariance of two variables and product of their standard deviations.

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

ANS- Scaling is changing the upper and lower bounds of the dataset to a single range.
Normalization:is scaling between 0 and 1
Standardization is scaling between -1 to 1
Scaling is performed so that multiple variables can be viewed on the same graph with different scales.

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

ANS- In the case of perfect correlation , R2 = 1 which leads to 1/(1-R2) to infinite. That implies a perfect correlation between variables.

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

ANS- Quantile -Quantile plots are plots between two Quantiles(ex.25% so that number of points below 25% in a plot).
It compares two different correlation plots in a variable.