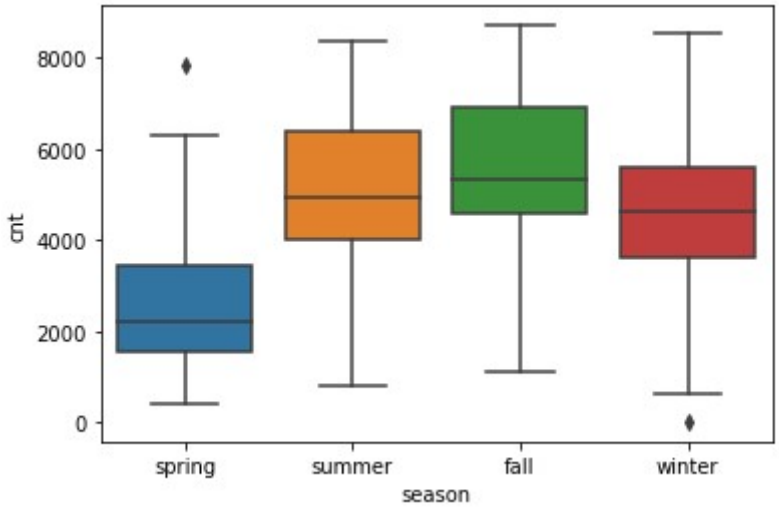
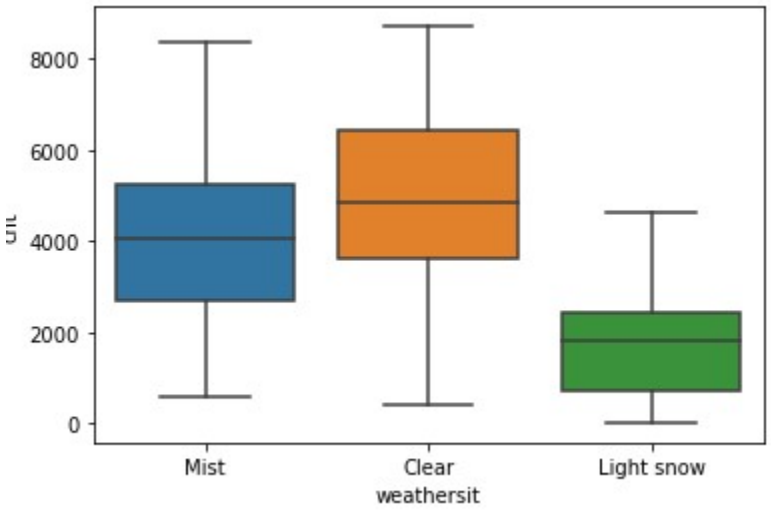
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

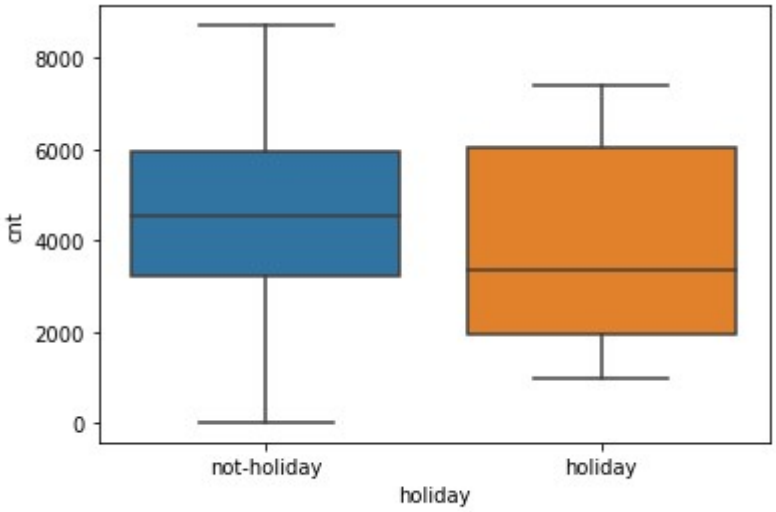
   Categorical variables in bike sharing are "Year", "Month", "Season", "Holiday", "weekday", "working day" and "weathersit"



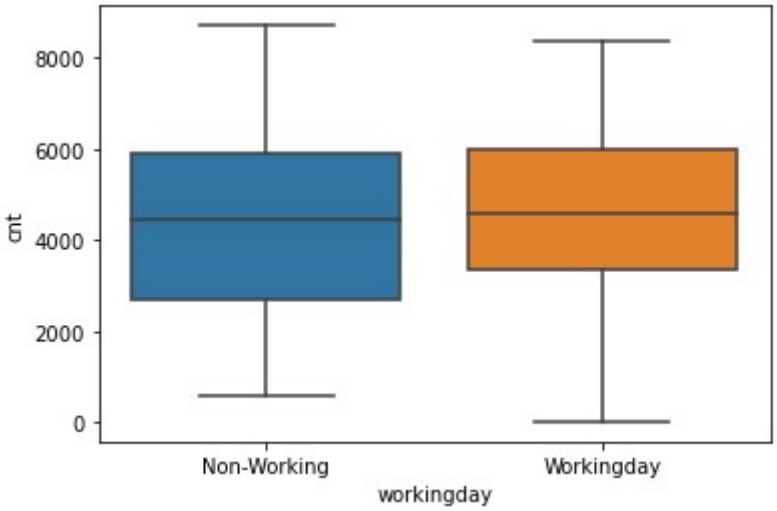
From figure we can say thet Fall has the height bike sharing so we can make more profit fro fall only.



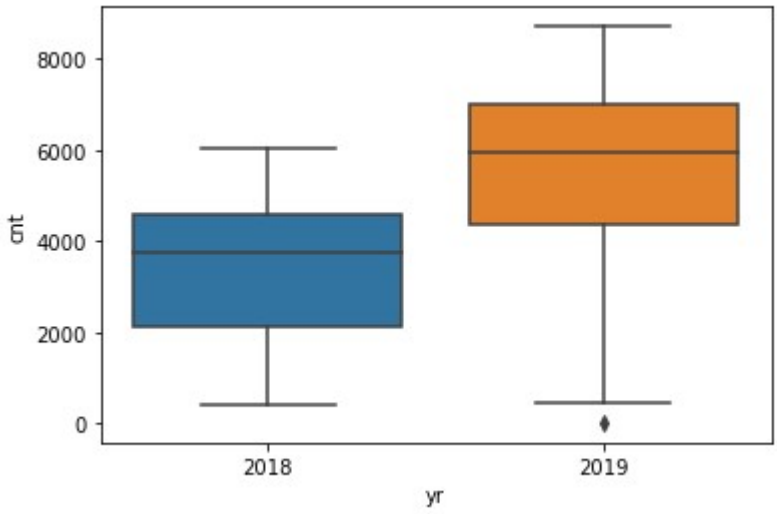
From Visualization we can say that when weather is clear the sharing of bike is more than any other condition



If it is a holiday the number of bike sharing is more than compared to not holiday.



From visualization we can say that holiday and working day has correlation



In 2019 has got the more popular than 2018 for number of shring of bikes

2. **Why is it important to use drop_first=True during dummy variable creation?**
   As the dummy variable creation for categorical variables is done with get_dummies(df[column_name]). Here n types of Categories can be represented with 0 to n-1 features. Here if we don't mention drop_first = True it include n variable for Dataframe.
   When you have applied the statsmodel to Linear regression the VIF of few variable we may get INF(infinity) as denominator R2 get 1.
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

   From the Pair-plot of visualization we can say temperature has heighest correalation with dependent variable. We can see from HeatMap It has 0.63 correlation factor
4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   i. First we can check given R2 value and adjusted R2 value how much variant should be 5%
   ii. Need to check hypothesis testing at least one variable is linearly dependent
   iii. Varience is constant
   iv. Distribution of error is normal
   v. Error is random in nature
5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   "Workingday", 'Saturday' and '2019_year' are having +ve coefficients they are contributing more.

General Questions:

1. Explain the linear regression algorithm in detail?

Linear regression assumes a linear or straight line relationship between the input variables (X) and the single output variable (y).

More specifically, that output (y) can be calculated from a linear combination of the input variables (X). When there is a single input variable, the method is referred to as a simple linear regression.

In simple linear regression we can use statistics on the training data to estimate the coefficients required by the model to make predictions on new data.

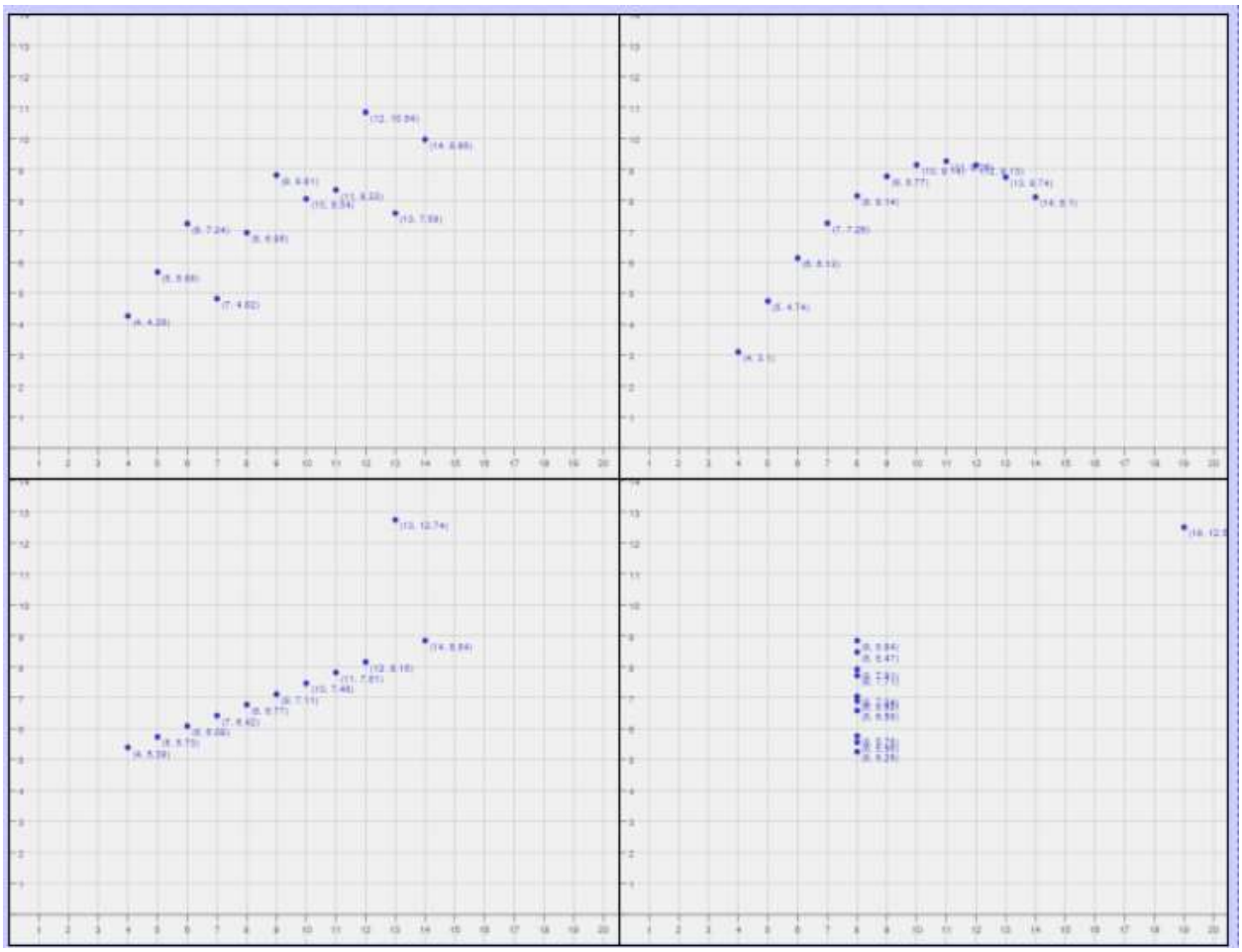
$$Y = B0 + B1X$$

1. Split the data as X_train, y_train, X_test, y_test to model the data
2. Get the linear regression object to make use of linear regression or statsmodel
3. Fit the data to Regression object or Ordinary linear square model
4. Predict the variable by creating constant from available data
5. Analyze error normally distributed and random in nature
6. Check the efficiency of model from predict to test variables

2. Explain the Anscombe's quartet in detail ?

Anscombe's Quartet is a great demonstration of the importance of graphing data to analyze it. Anscombe's Quarter is 4 differet dataset having same mean and variance are of same but the when we plot the data it will shows the difference.

Anscombe's Quartet warns of the dangers of outliers in data sets.



Anscombe's Quartet reminds us that graphing data prior to analysis is good practice,

outliers should be removed when analyzing data,

and statistics about a data set do not fully depict the data set in its entirety.

3. **What is Pearson's R?**
   Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

   Cov is covarience
   Sigmax standard deviation of x
   Sigmay standard deviation of y

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

Simple Linear Regression, scaling doesn't impact your model. So it is extremely important to rescale the variables so that they have a comparable scale. If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients. This might become very annoying at the time of model evaluation. So it is advised to use standardization or normalization so that the units of the coefficients obtained are all on the same scale. As you know, there are two common ways of rescaling:

1. Min-Max scaling
2. Standardisation (mean-0, sigma-1)

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
VIF varience influence factor has the values which explains about the multicolinearity.
As the redundant variables are there in the dataframe the values will be having highest value.
>5 menas having the reduedent variables.
INF value comes due to R2 is one for variables the residual values becomes one. And due to adding all dummies to data set causes the INF values in VIF

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

The “QQ” in QQ plot means quantile-quantile — that is, the QQ plot compares the quantiles of our data against the quantiles of the desired distribution
(defaults to the normal distribution, but it can be other distributions too as long as we supply the proper quantiles).

Quantiles are breakpoints that divide our numerically ordered data into equally proportioned buckets.
For example, you’ve probably heard of percentiles before — percentiles are quantiles that divide our data into 100 buckets (that are ordered by value),
with each bucket containing 1% of observations. Quartiles are quantiles that divide our data into 4 buckets (0–25%, 25–50%, 50–75%, 75–100%).
Even our old friend, the median is a quantile — it divides our data into two buckets where half our observations are lower than the median and half our higher than it.