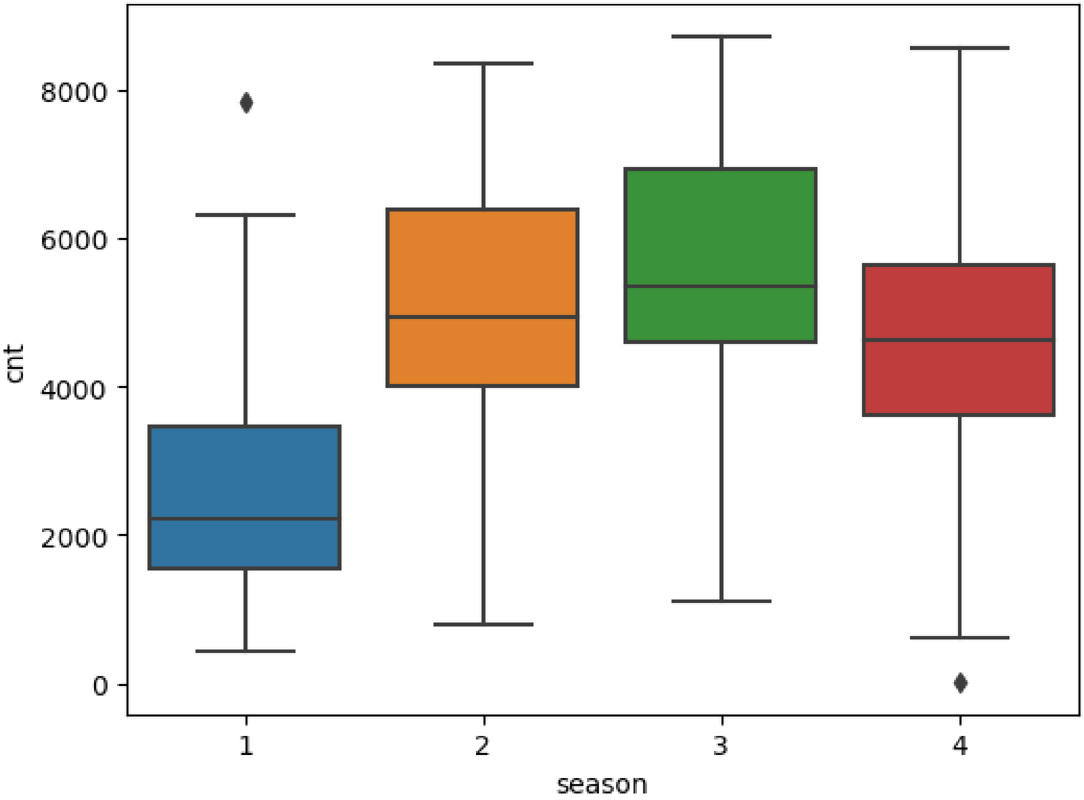
Assignment-based Subjective Questions

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

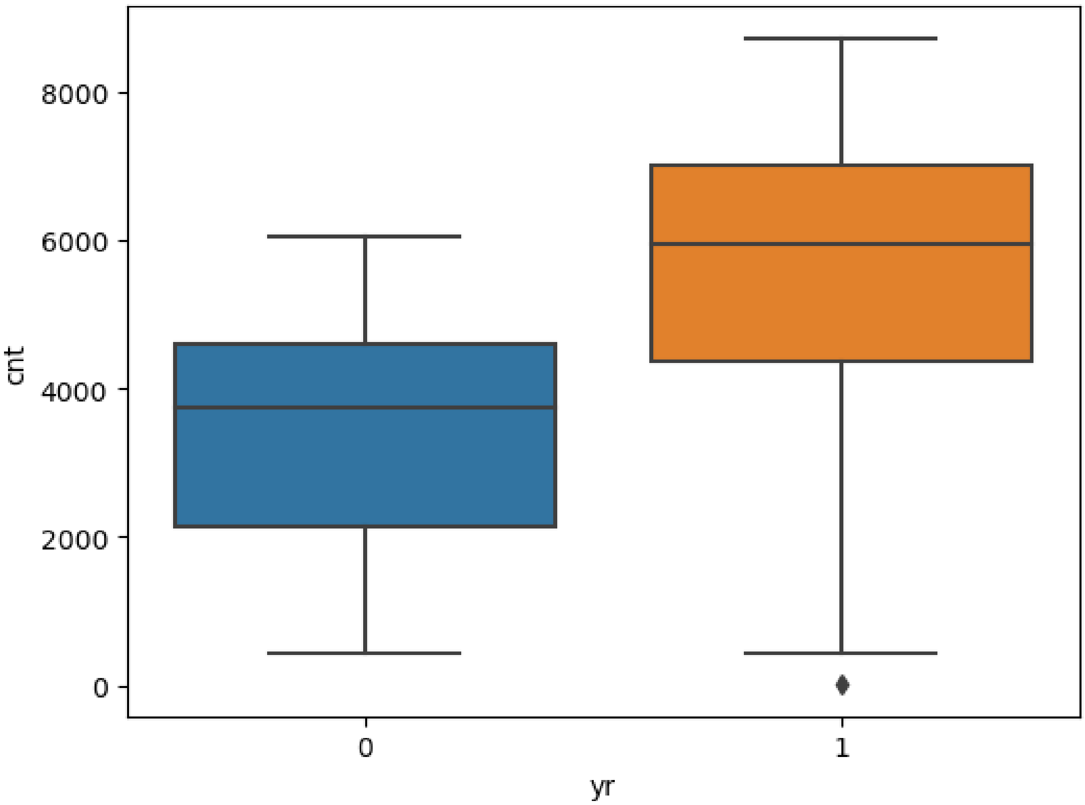
Ans: Categorical variables and their effect on dependent variables

a) Season



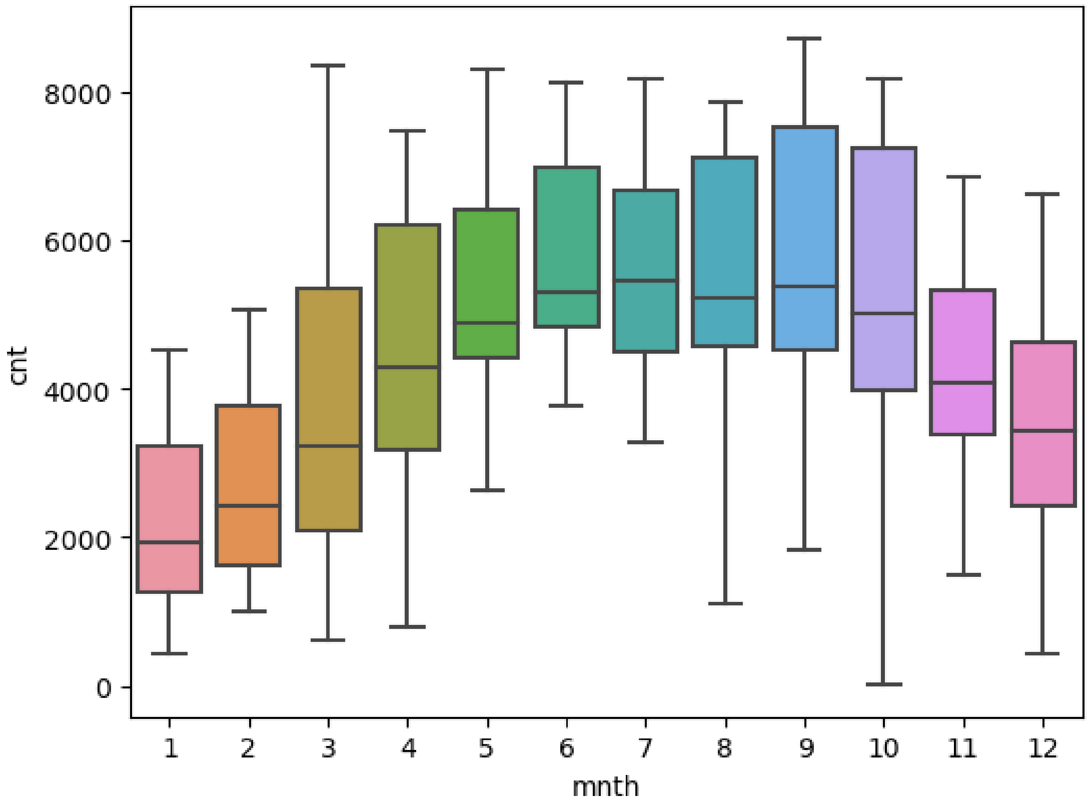
We could see that bike rentals increase up to season 3 and drop for season 4. Hence we say that ‘Season’ is good predictor of bike rentals.

b) Year:



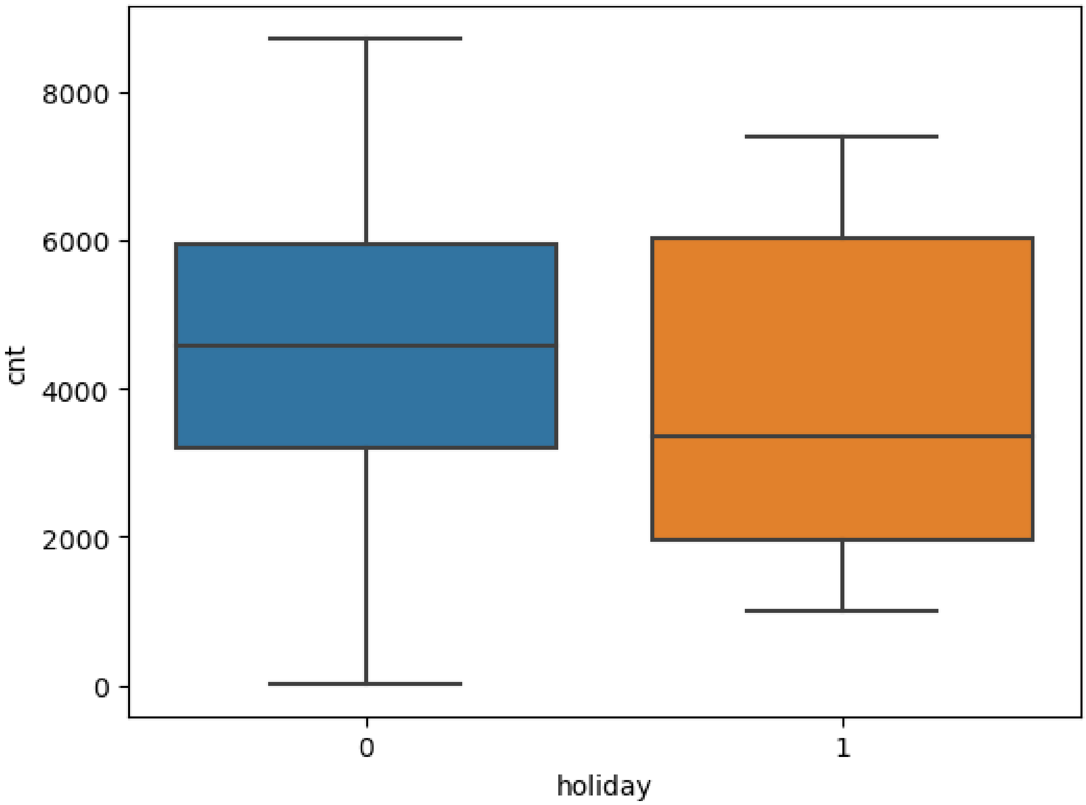
We could see that rentals were significantly high in year 2019 than 2018. So ‘Year’ is good predictor of bike rentals.

c) **Months:**



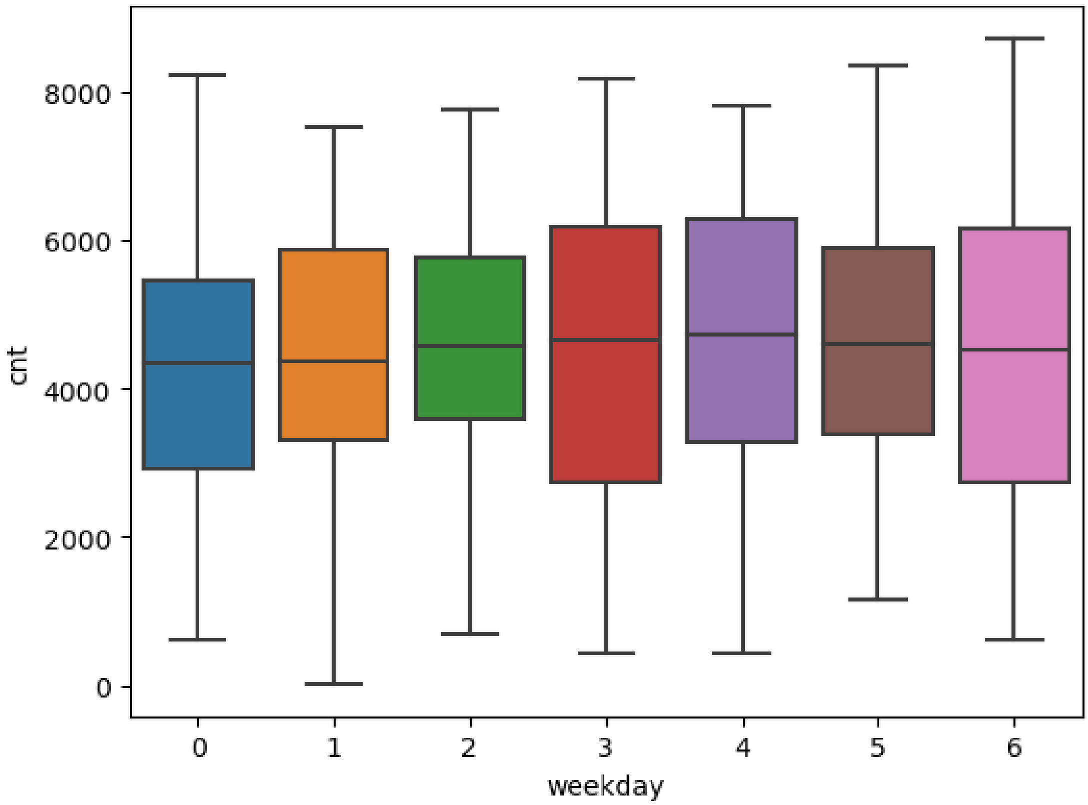
We could see that bike rentals increase upto month 9 than decreases. So we could say 'Month' is a good predictor of bike rentals

d) Holiday :



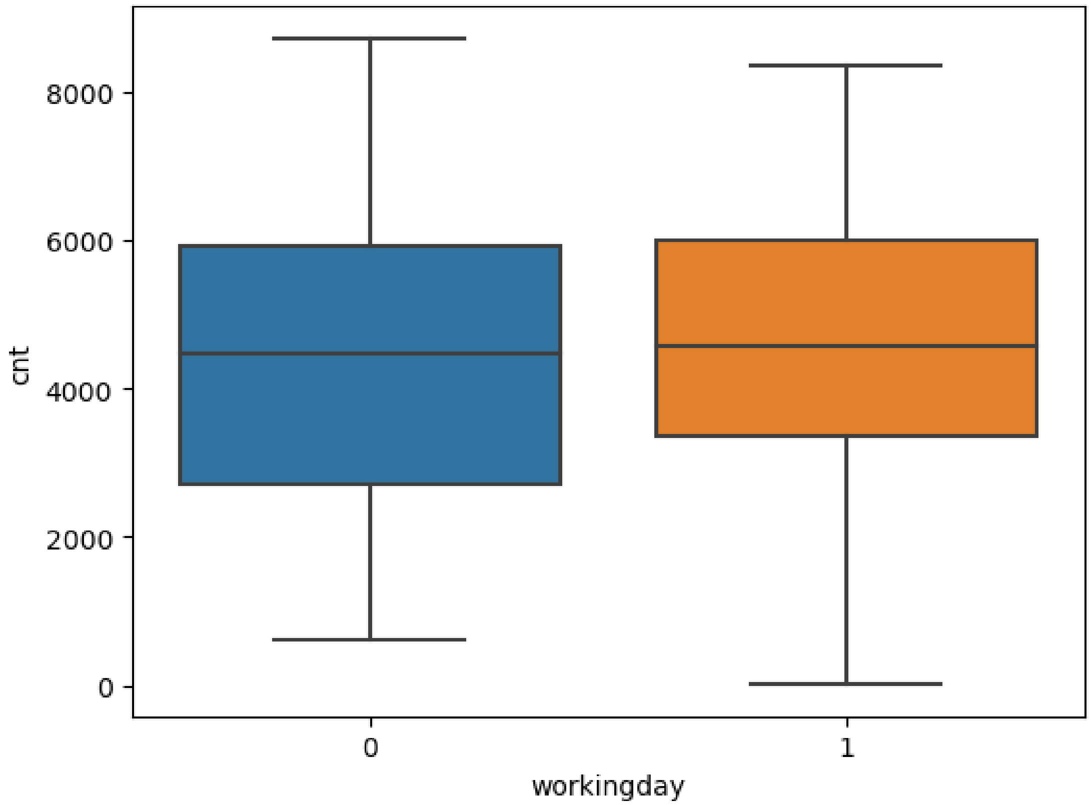
About 97% of the bookings are done when it is not a holiday. This is extremely biased data. So holiday is not good predictor of bike rentals.

e) Weekday:



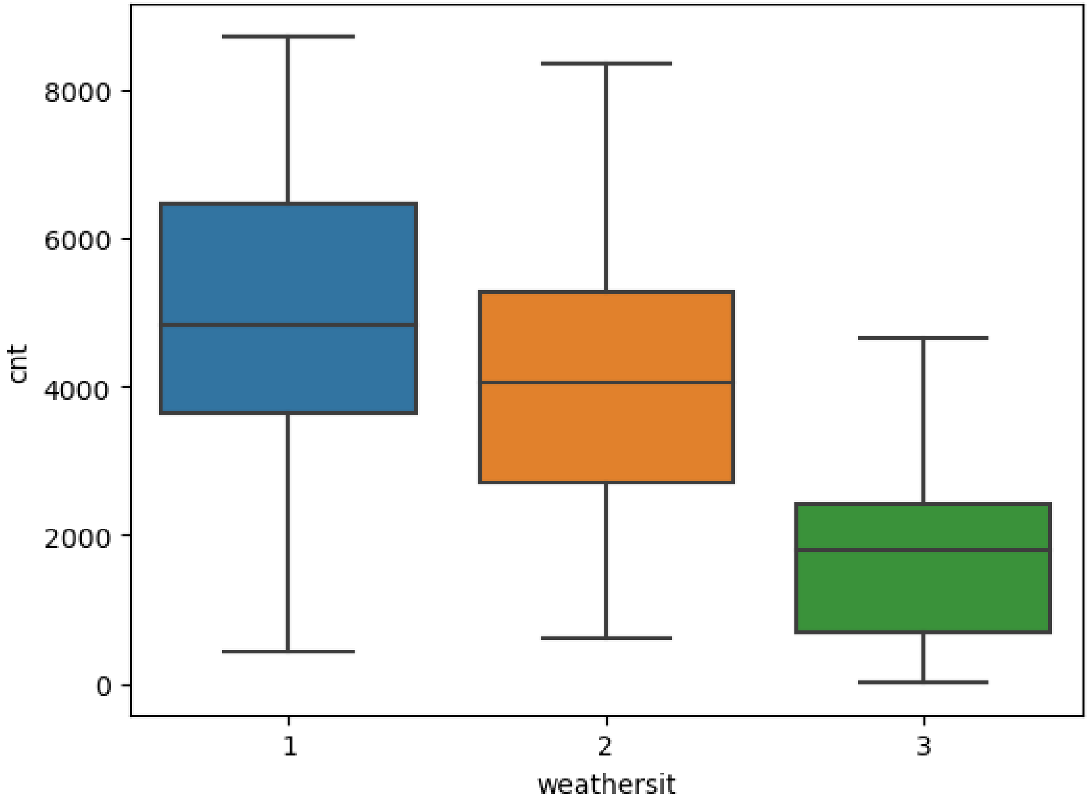
Since all the weekdays have approximately same median of bike rentals therefore weekday may or may not be a good predictor of sales.

f) **Working day**



Since 69% of the bookings happened on working day . Therefore it's a good predictor of bike rentals.

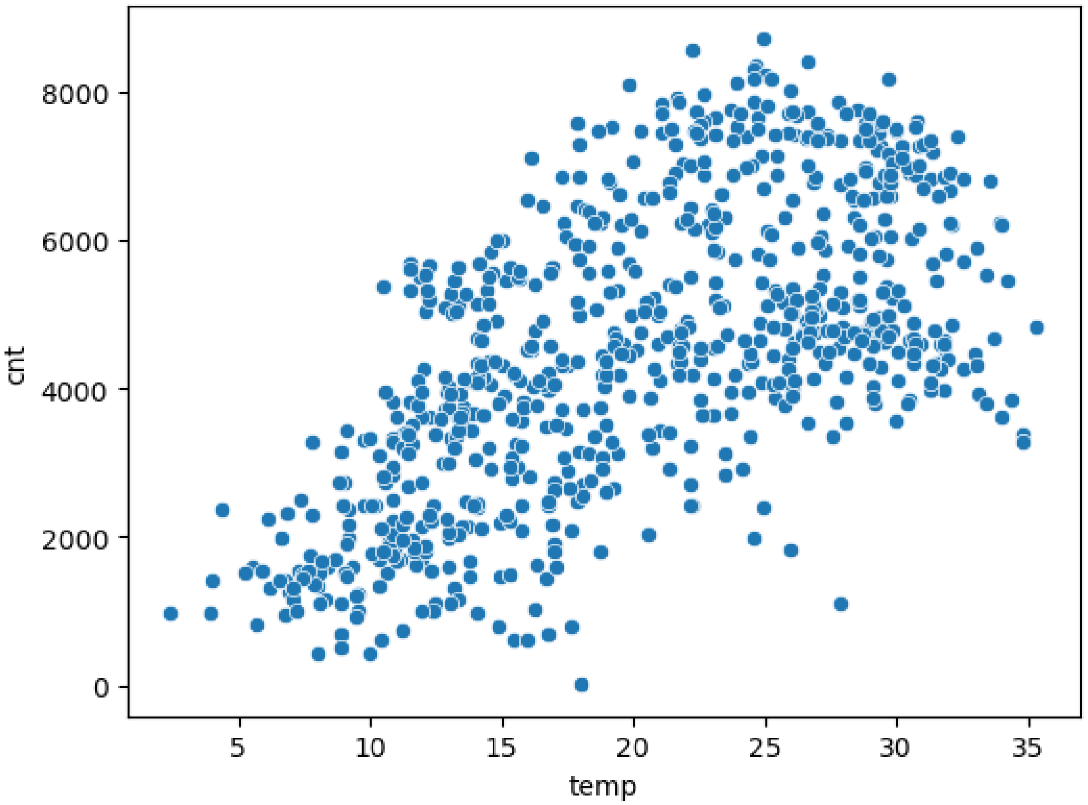
g) Weather situation:

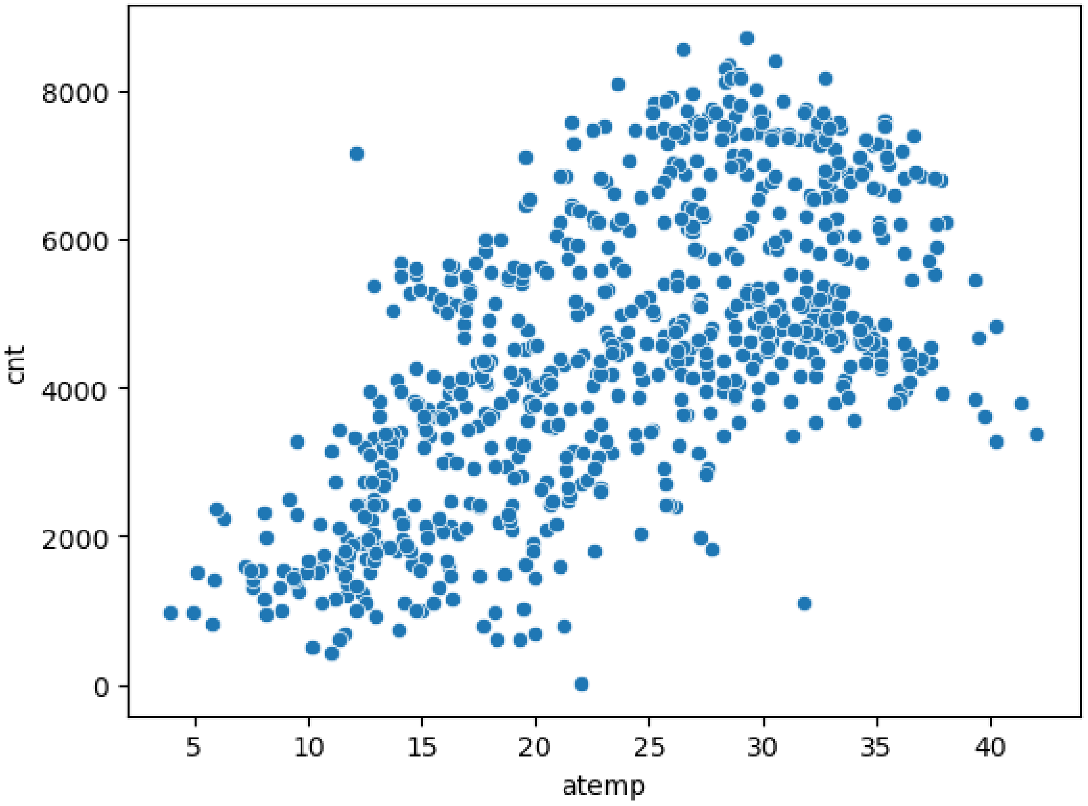


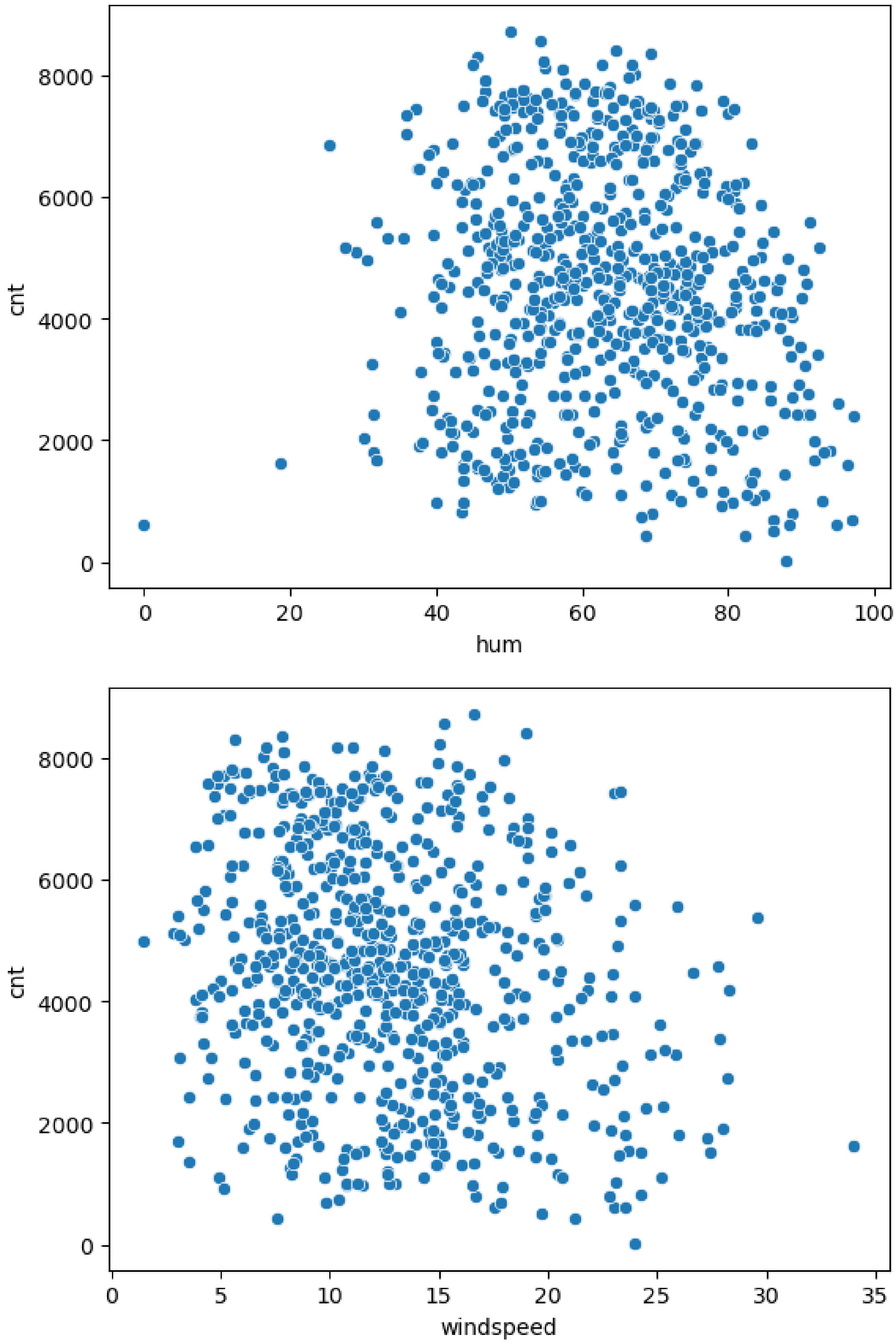
As we can see more number of bikes were rented on clear weather. So weather situation is good predictor of bike rentals.

2. **Why is it important to use drop_first=True during dummy variable creation?**
   Ans: Because if there are n levels of a categorical variable then only n-1 dummy variables are actually needed . drop_first = True will drop the first dummy variable, leaving us with n-1 dummy variables.

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


8000
6000
4000
2000
0
cnt
5
10
15
20
25
30
35
temp


8000
6000
4000
2000
0
cnt
5
10
15
20
25
30
35
40
atemp

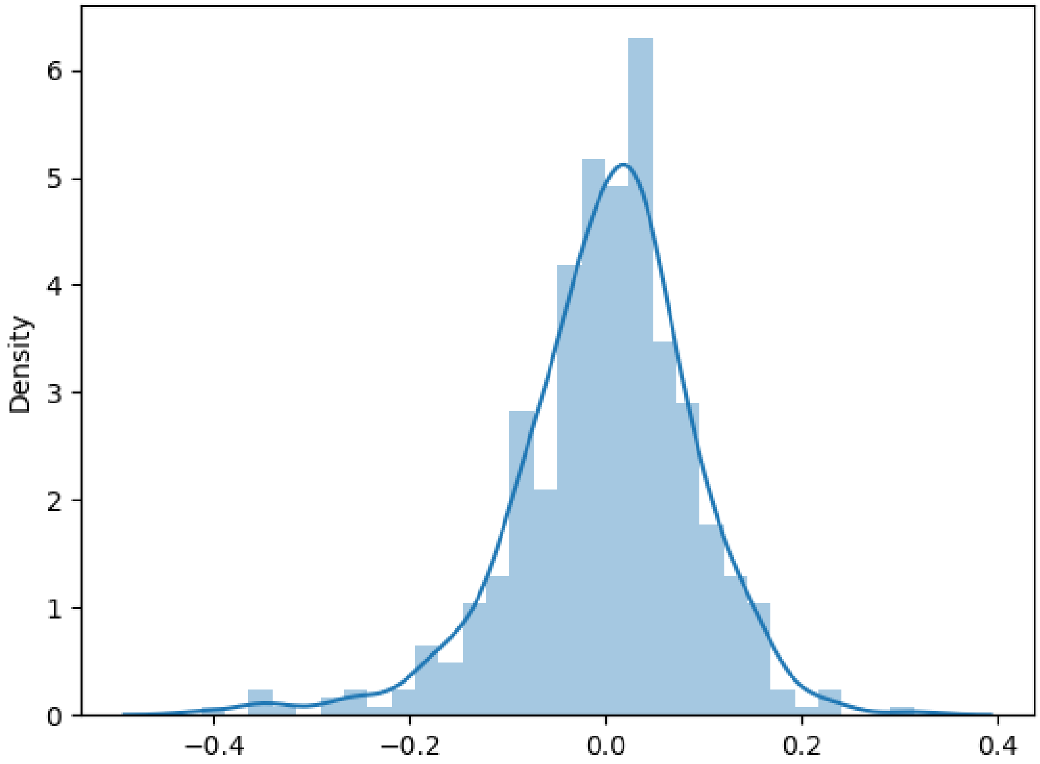

8000
6000
4000
2000
0
cnt
0
20
40
60
80
100
hum
8000
6000
4000
2000
0
cnt
0
5
10
15
20
25
30
35
windspeed

As we can see in above plots that temp and atemp show good linear relationship with target variable. So temp and atemp have highest correlation with target variable.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   Ans:
   1) X and Y should have linear relationship. We could see that temp variable had linear relationship with bike rentals.

   We did residual analysis for validating assumptions of linear regression.

   2) Residual terms should show normal distribution with mean centered around zero. As we could see in below plot of residuals that it is normal distribution with mean centered around zero.



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

const	**0.074417**
yr	**0.233643**
workingday	**0.056634**
temp	**0.550076**

```
windspeed   -0.155035
season_2     0.088584
season_4     0.131824
month_9      0.097117
weekday_6    0.067612
weather_2   -0.080480
weather_3   -0.287809
dtype: float64
```

The above coefficients are obtained in the final model. The top 3 contributors to sales are

- **temp(Temperature) :** For a unit increase in temp the sales increases by 0.55
- **weather_3:** For a unit increase in extreme weather situations the sales decreases by 0.28
- **yr(Year):** Sales increase by 0.23 if year is 2019

General Subjective Questions

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

Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features. When the number of the independent feature, is 1 then it is known as Univariate Linear regression, and in the case of more than one feature, it is known as multivariate linear regression. The goal of the algorithm is to find the best linear equation that can predict the value of the dependent variable based on the independent variables. The equation provides a straight line that represents the relationship between the dependent and independent variables. The slope of the line indicates how much the dependent variable changes for a unit change in the independent variables.

Linear regression is used in many different fields, including finance, economics, and psychology, to understand and predict the behavior of a particular variable. For example, in finance, linear regression might be used to understand the relationship between a company's stock price and its earnings or to predict the future value of a currency based on its past performance.

One of the most important supervised learning tasks is regression. In regression set of records are present with X and Y values and these values are used to learn a function so if you want to predict Y from an unknown X this learned function can be used. In regression we have to find the value of Y, So, a function is required that predicts continuous Y in the case of regression given X as independent features.

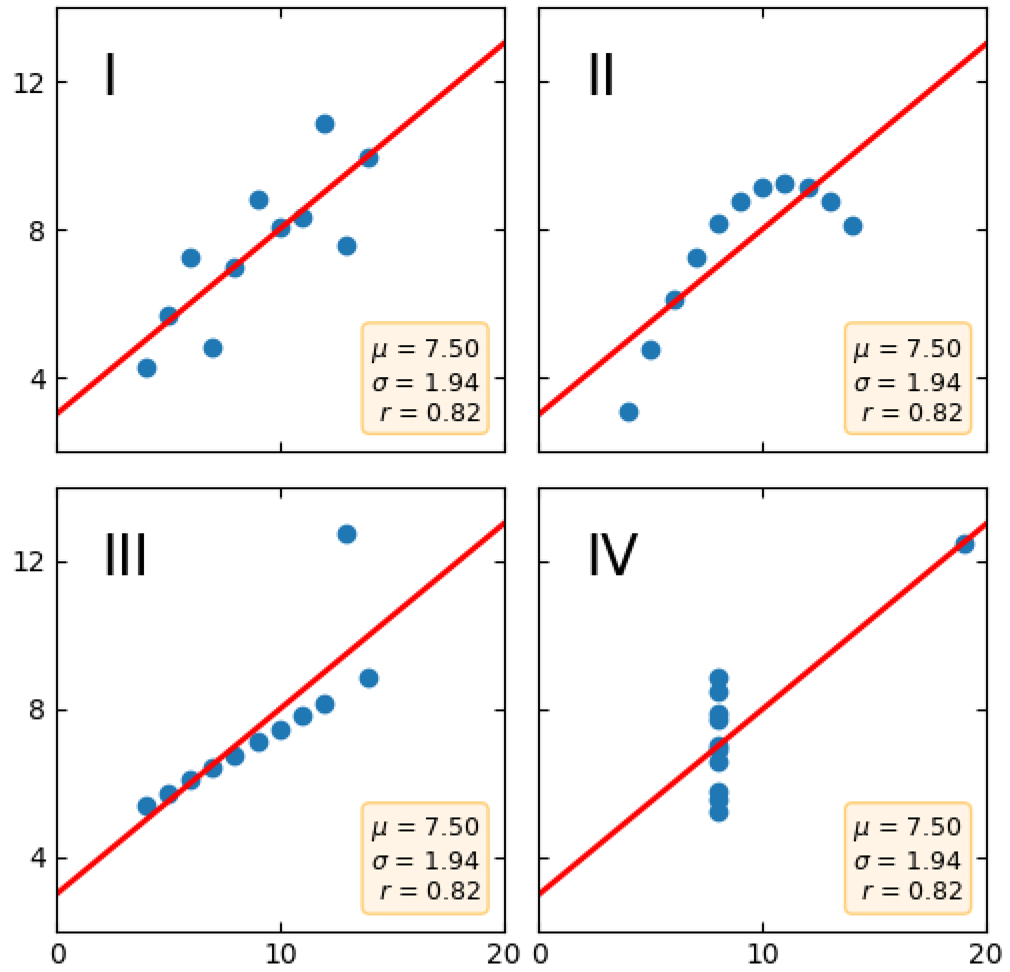
Here Y is called a dependent or target variable and X is called an independent variable also known as the predictor of Y. There are many types of functions or modules that can be used for regression. A linear function is the simplest type of function. Here, X may be a single feature or multiple features representing the problem.

2. **Explain the Anscombe's quartet in detail.**
   Ans: Anscombe's quartet is a group of datasets (x, y) that have the same mean, standard deviation, and regression line, but which are qualitatively different.

   It is often used to illustrate the importance of looking at a set of data graphically and not only relying on basic statistic properties.

   Consider following datasets illustrated below.

```
x = [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5]
y1 = [8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68]
y2 = [9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74]
y3 = [7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73]
x4 = [8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8]
y4 = [6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91, 6.89]

datasets = {

  'I': (x, y1),

  'II': (x, y2),

  'III': (x, y3),

  'IV': (x4, y4)

}
```



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

   The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

   The Pearson correlation coefficient (r) is the most widely used correlation coefficient and is known by many names:

   - Pearson's r
   - Bivariate correlation

- Pearson product-moment correlation coefficient (PPMCC)
- The correlation coefficient

The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

The Pearson correlation coefficient is also an inferential statistic, meaning that it can be used to test statistical hypotheses. Specifically, we can test whether there is a significant relationship between two variables.

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

Ans:

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization/Min-Max Scaling:

It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

Standardization Scaling:

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

sklearn.preprocessing.scale helps to implement standardization in python.
One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

Ans:

This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect correlation.

5. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**
   Ans:
   Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

   This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

   Few advantages:
   a) It can be used with sample sizes also

   b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

   It is used to check following scenarios:

   If two data sets —

   i. come from populations with a common distribution

   ii. have common location and scale

   iii. have similar distributional shapes

   iv. have similar tail behavior

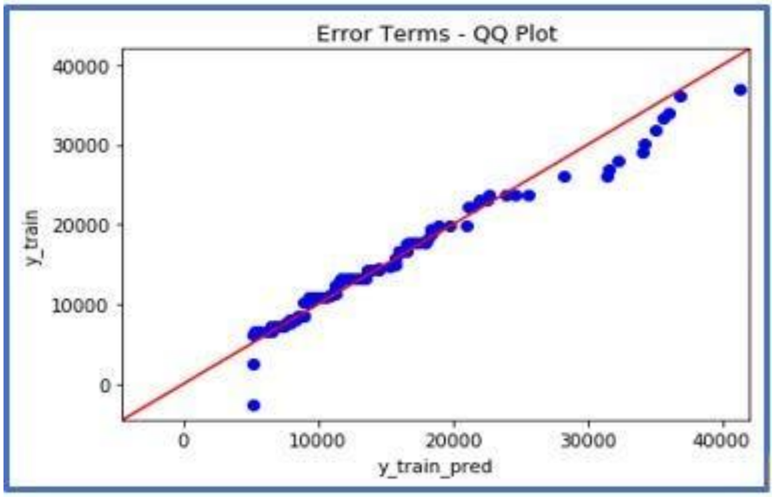
Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

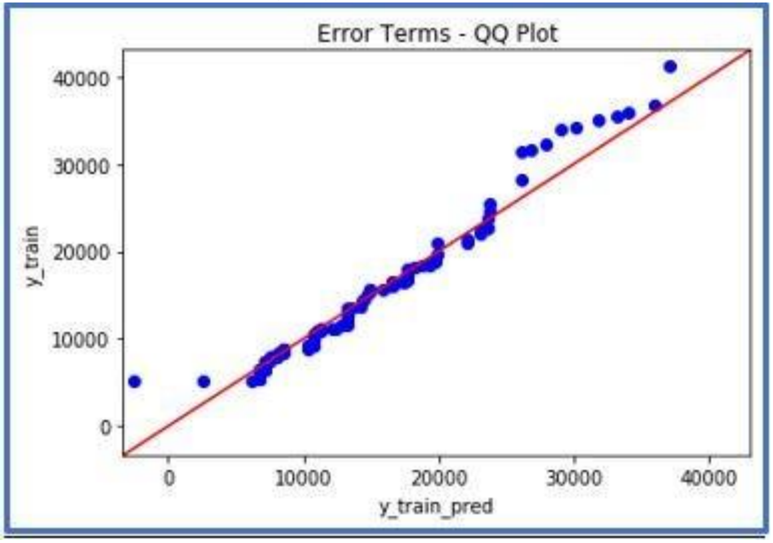
Below are the possible interpretations for two data sets.

a) ***Similar distribution****: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis*

b) ***Y-values < X-values:*** *If y-quantiles are lower than the x-quantiles.*



c) X-values < Y-values: *If x-quantiles are lower than the y-quantiles.*


Error Terms - QQ Plot
y_train
y_train_pred