# Shared Bikes Demand Prediction

- Submitter:
  - Aditya Singh

# Use Case

- A US bike-sharing provider ’BoomBikes’ has recently suffered considerable dips in their revenues due to the ongoing Corona pandemic. The company is finding it very difficult to sustain in the current market scenario.
- BoomBikes aspires to understand the demand for shared bikes among the people after this ongoing quarantine situation ends across the nation due to Covid-19.
- They have contracted a consulting company to understand the factors on which the demand for these shared bikes depends, specifically which variables are significant in predicting the demand for shared bikes and how well those variables describe the bike demands.
- We are required to model the demand for shared bikes with the available independent variables.

# Data Understanding

1) The loan dataset had 39717 rows and 111 columns.

2) Their are 74 columns of float type, 13 of integer type

   and 24 of textual/object type.

3) Cnt representing Bicycle Count is the dependent or target column.

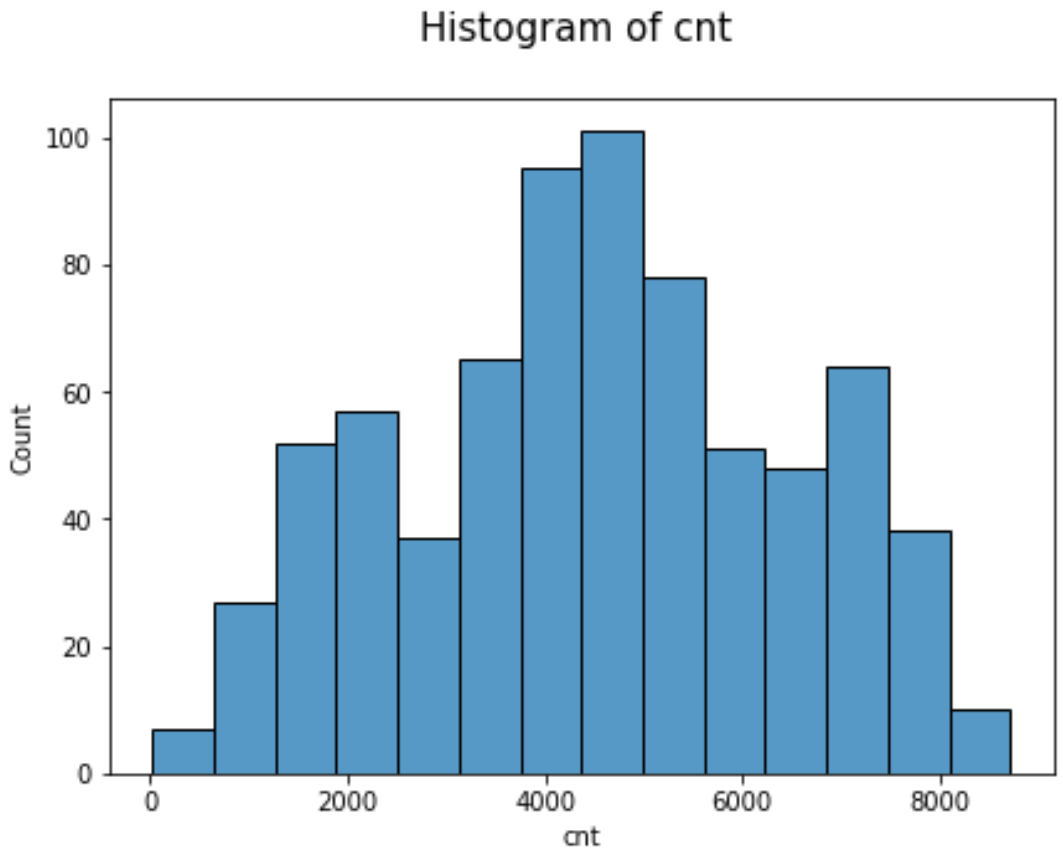
# Removing Irrelevant Columns

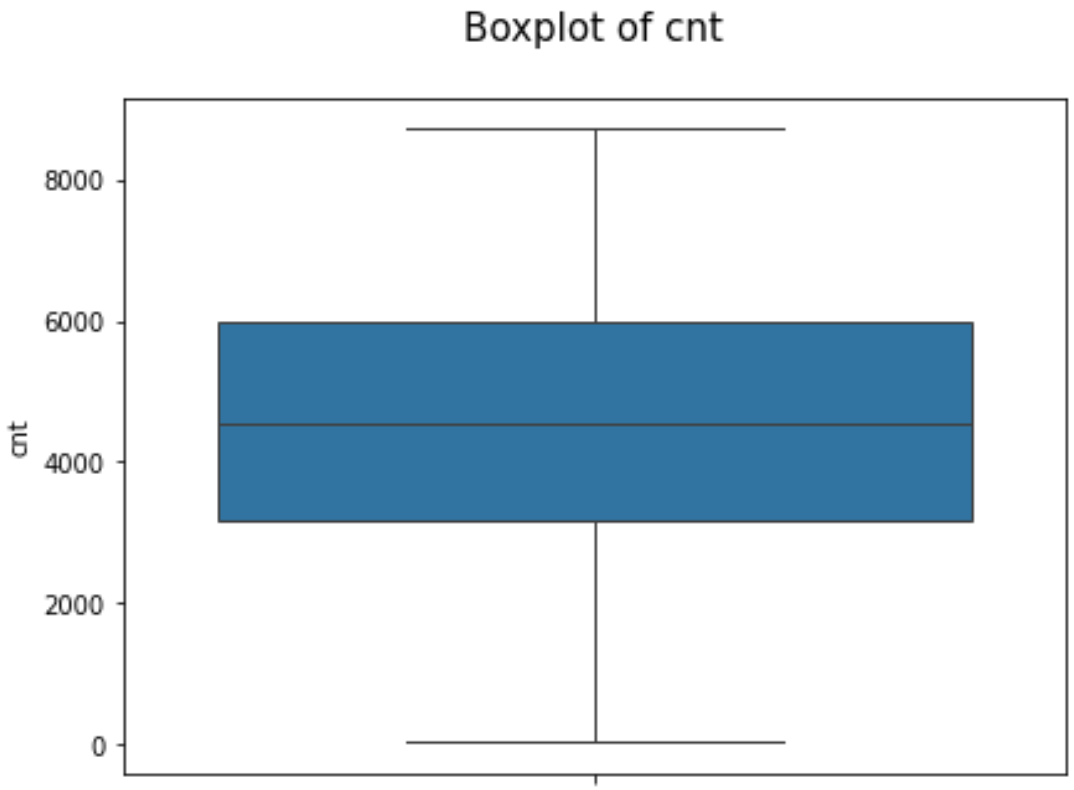
- Instant,  dteday,  atemp, casual, registered columns are removed from analysis as they are correlated with other predictors
- After removing these columns we had 11 columns for analysis

# Segregrating the remaining features into continuous and numeric discrete type

- continuous = ['temp’,  'hum’,  'windspeed']
- categorical = ['season’,  'yr’,  'mnth’,  'holiday’,  'weekday', 'workingday’,  'weathersit']
- target = 'cnt’
- We had 3 continuous and 7 numerical discrete features respectively.

# Bicycle Count



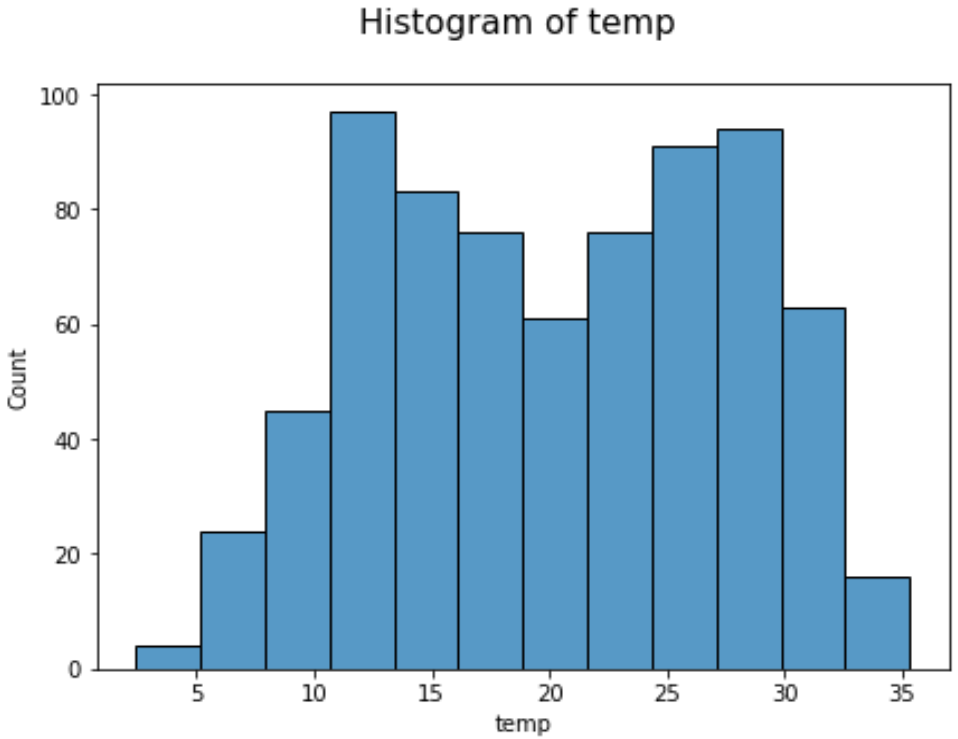



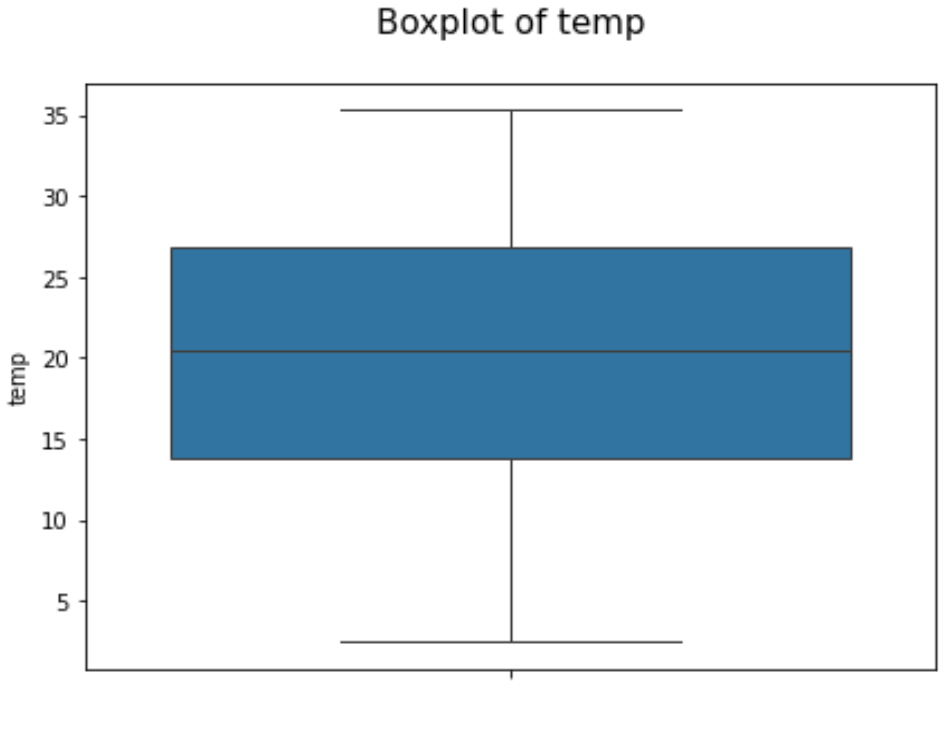

Inference:

1) The minimum and maximum value of bicycle count is 22 and 8714.
2) The mean value of bicycle count is 4508
3) Their are no outliers in bicycle count variable

# Univariate Analysis( Continuous Features)

## 1) Temperature in Celsius (temp)

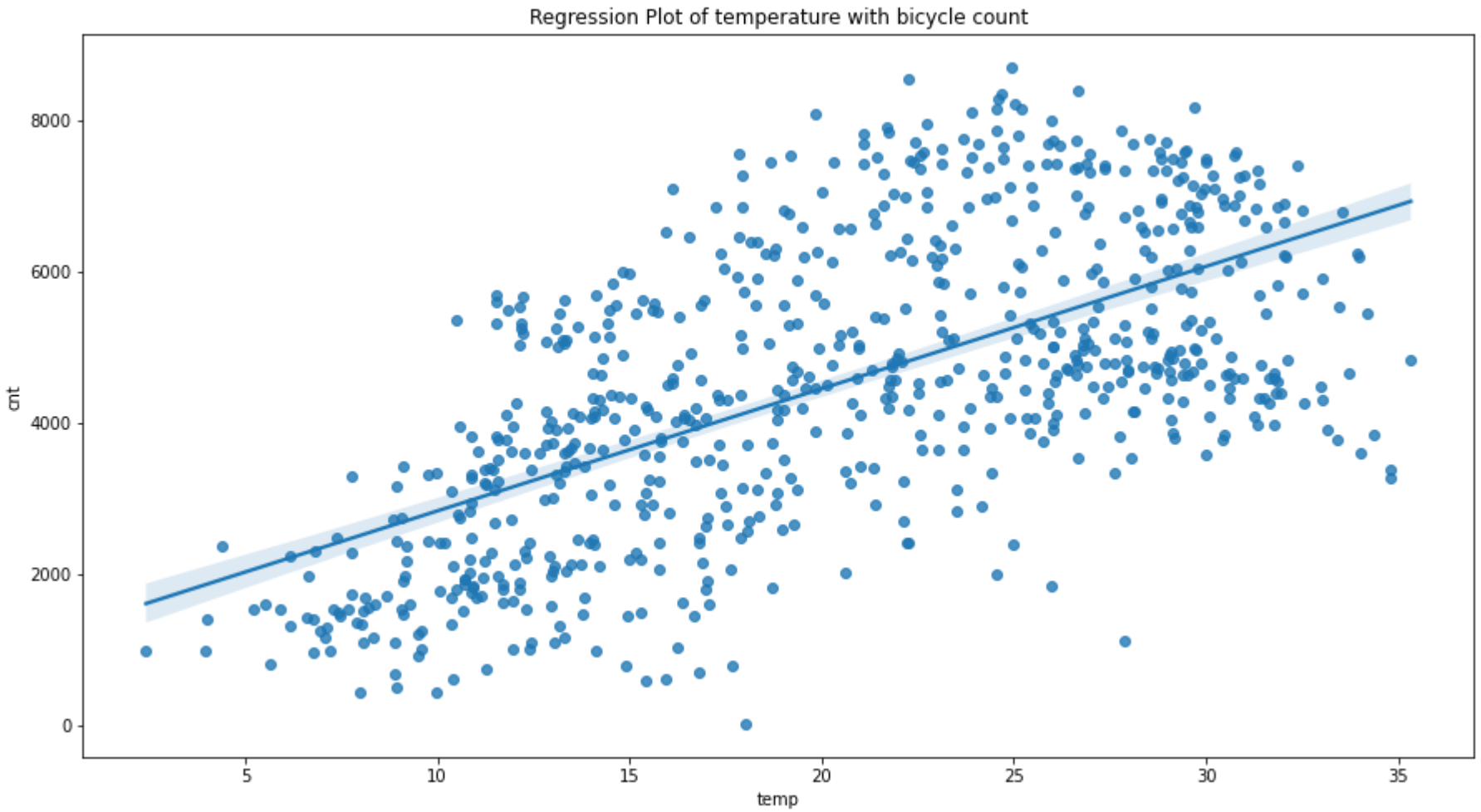



Inference:

1) The minimum and maximum value of temperature is 2.42 and 35.3.
2) The mean value of temperature is 20.32
3) Their are no outliers in temperature

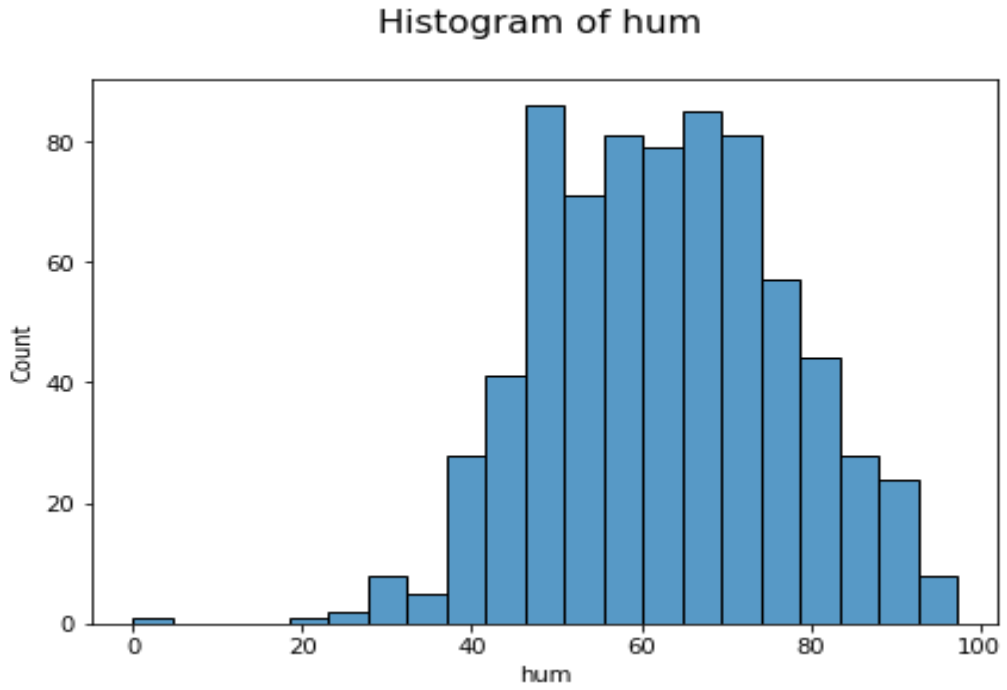
# Relation of Temp with bicycle count

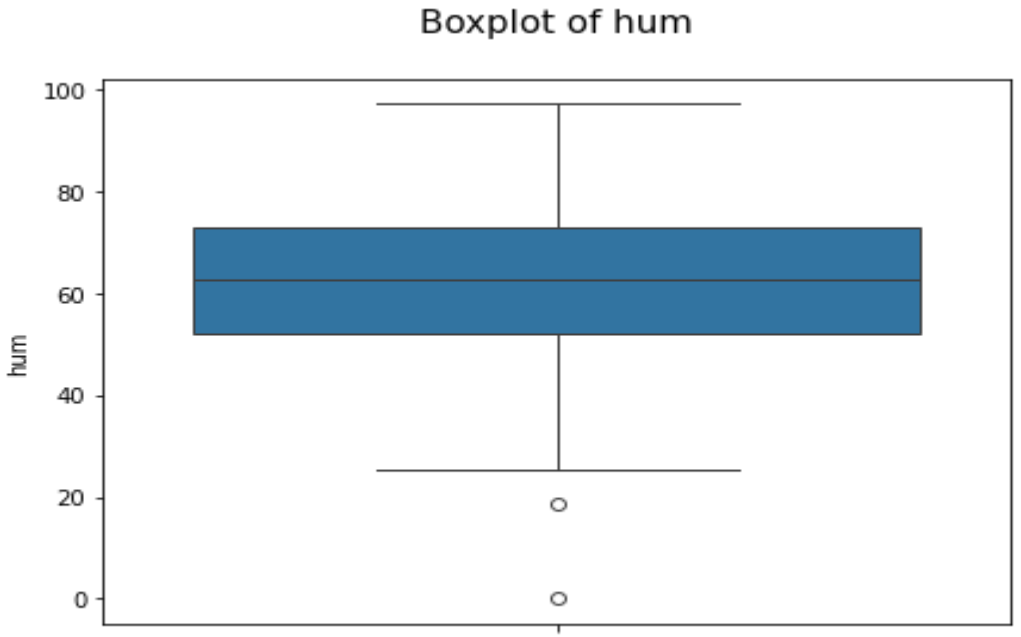



Inference:

1) In general, with increasing temperture the demand for bicycles increases
2) Their is a weak linear correlation(~63%) of temperture with bicycle count

# 2) Humidity




Inference:

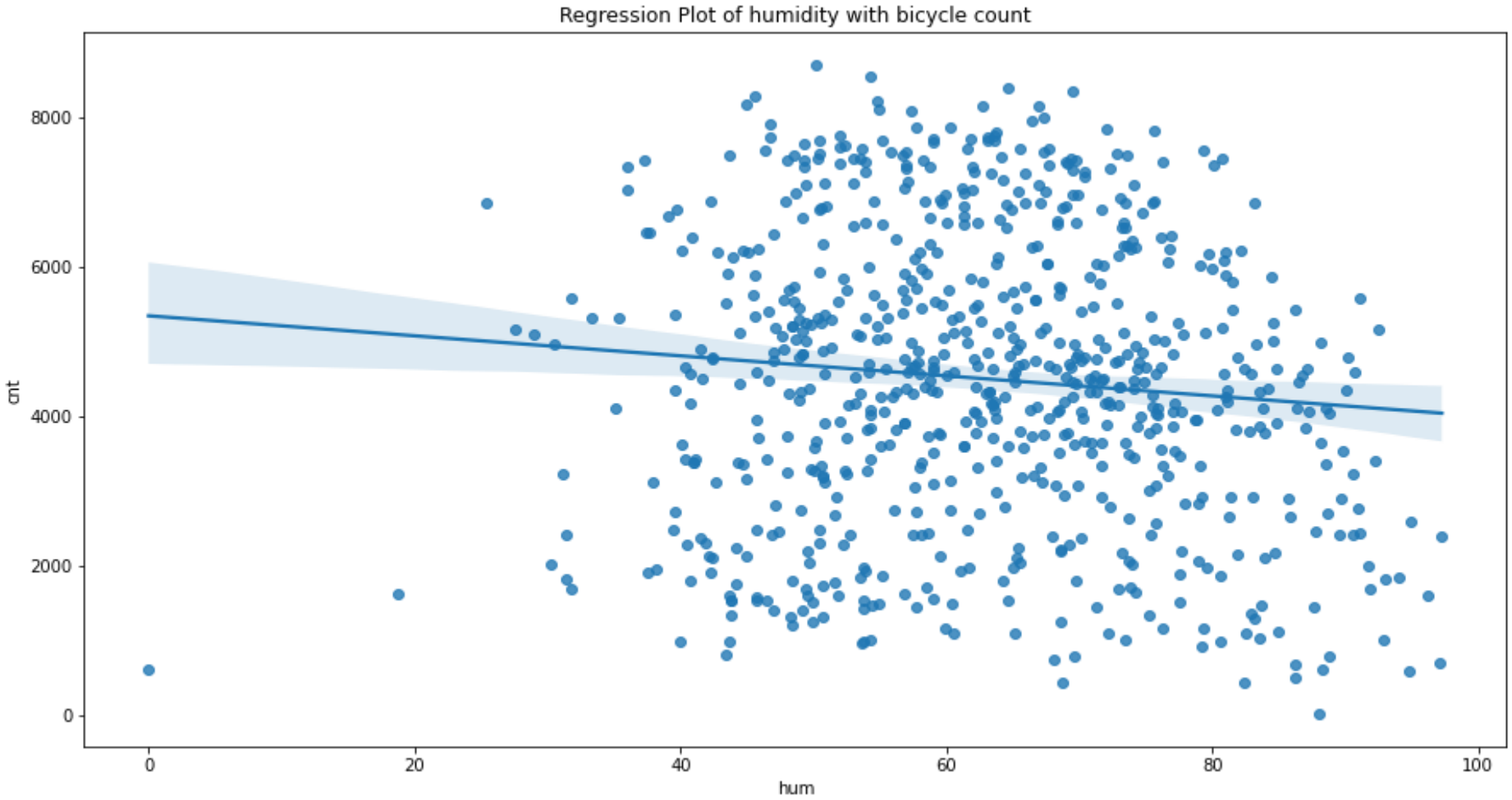
1) The minimum and maximum value of humidity is 0 and 97.25.

2) The mean value of temperature is 62.76

3) Their are very few outliers in below lower whiskers of humidity as shown in the boxplot

4) Since these are very few in number, so they will imputed by the median

# Relation of Humidity with bicycle count

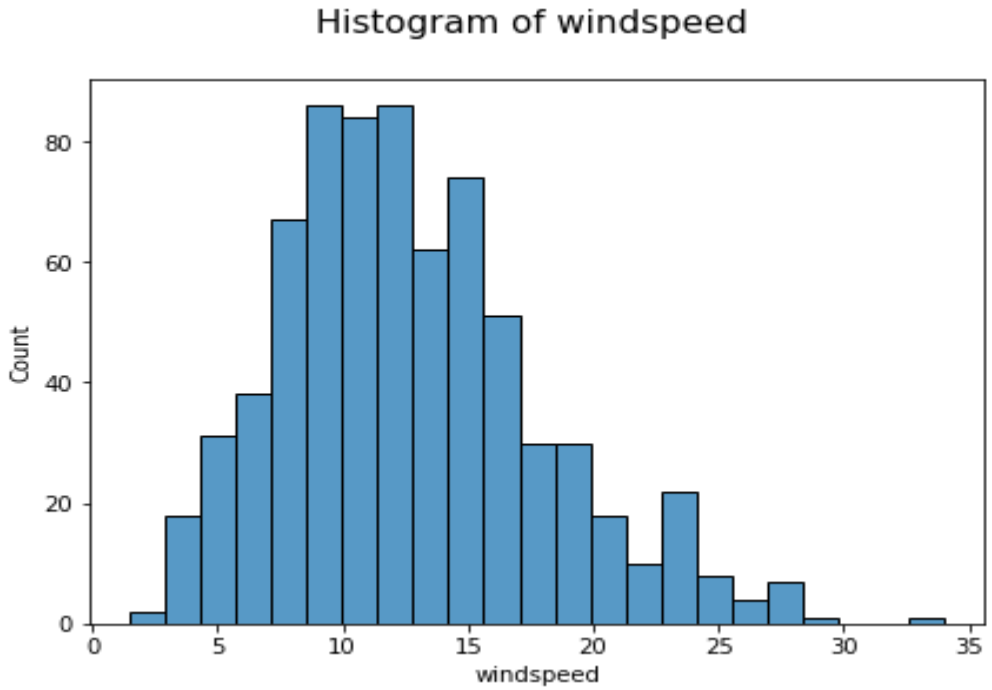


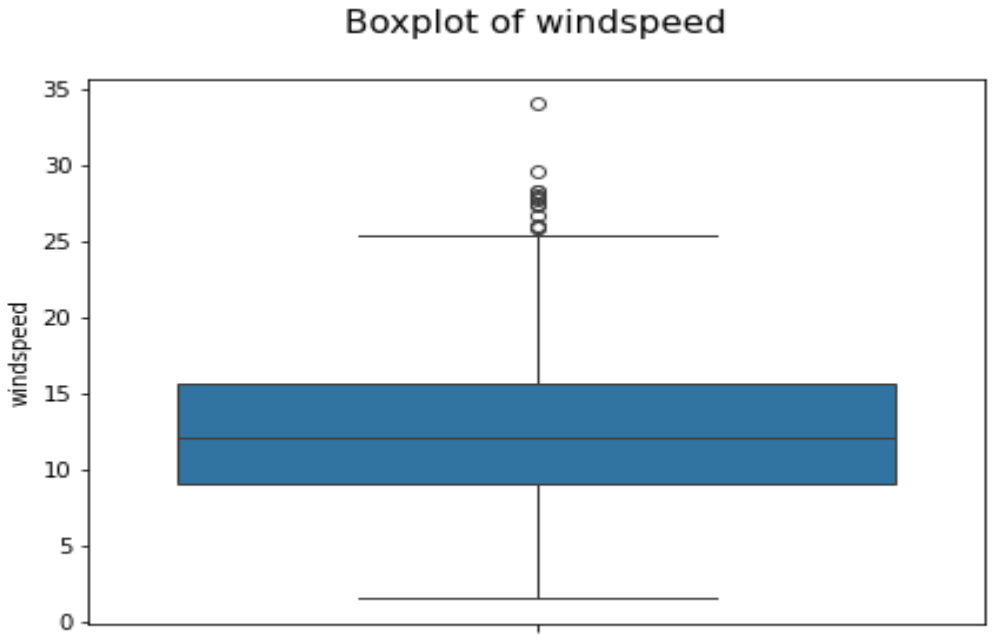

Inference:

1) Their is almost no bicycle demand below humidity value of 30

2) Their is very very weak negative correlation(~10%) of humidity with bicycle count

# 3) Windspeed



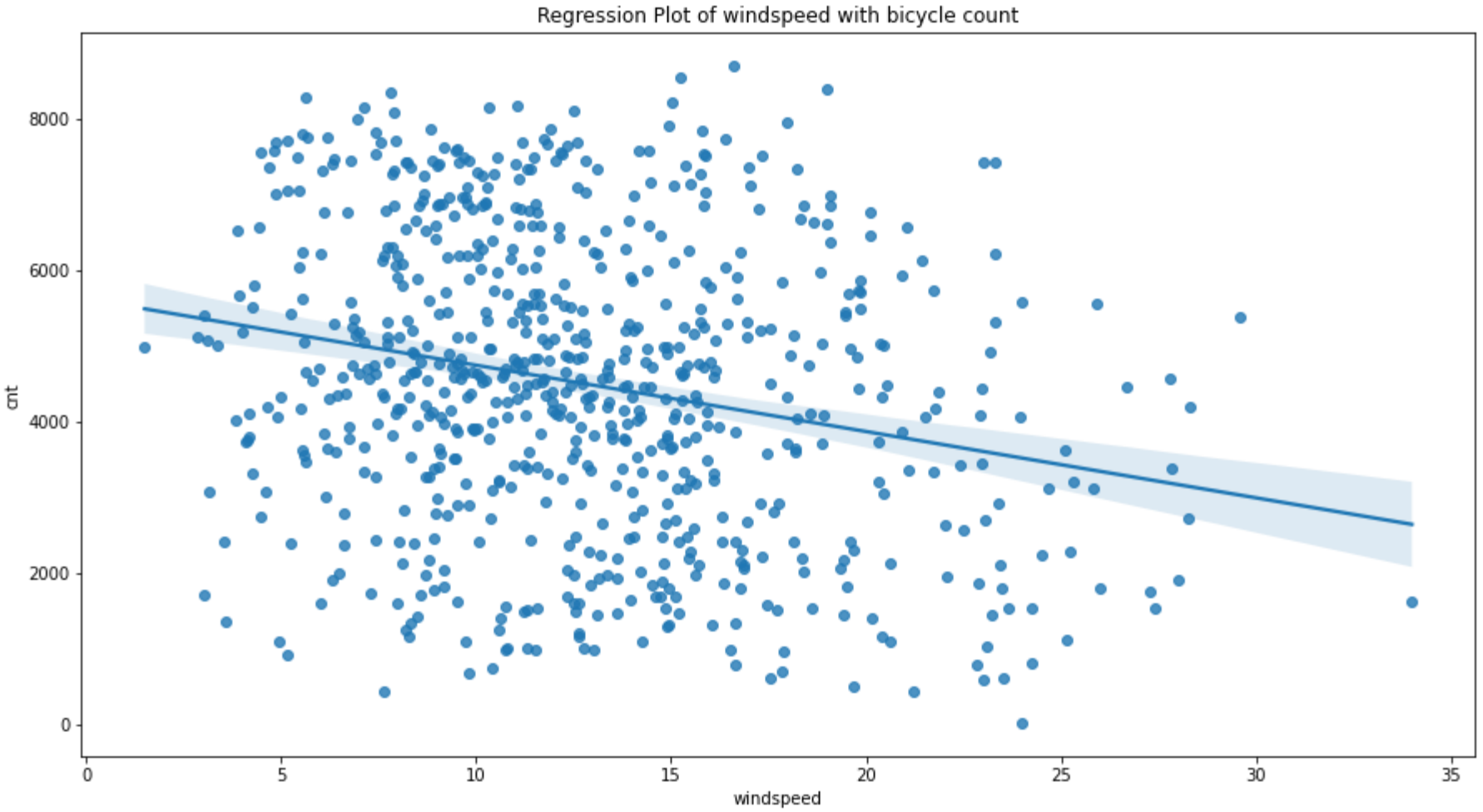

Inference:

1) The minimum and maximum value of windspeed is 1.5 and 34.

2) The mean value of windspeed is 12.76

3) Their are outliers in windspeed after a value of around 26 as shown in boxplot which will be removed by upper capping using the interquartile range method

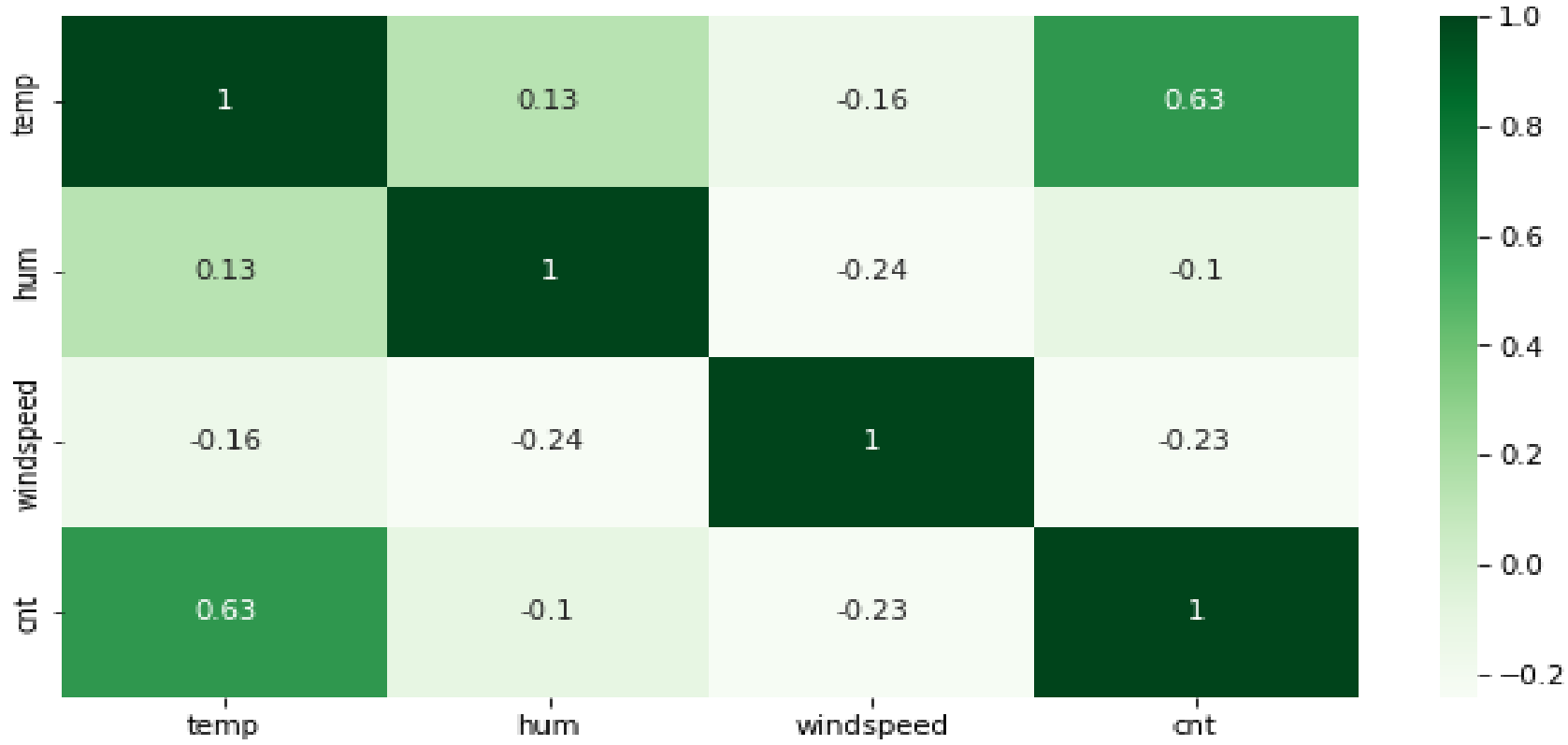
# Relation of windspeed with bicycle count




Inference:

1) Their is very weak negative correlation(~23%) of windspeed with bicycle count

# Correlation Analysis




Inference:

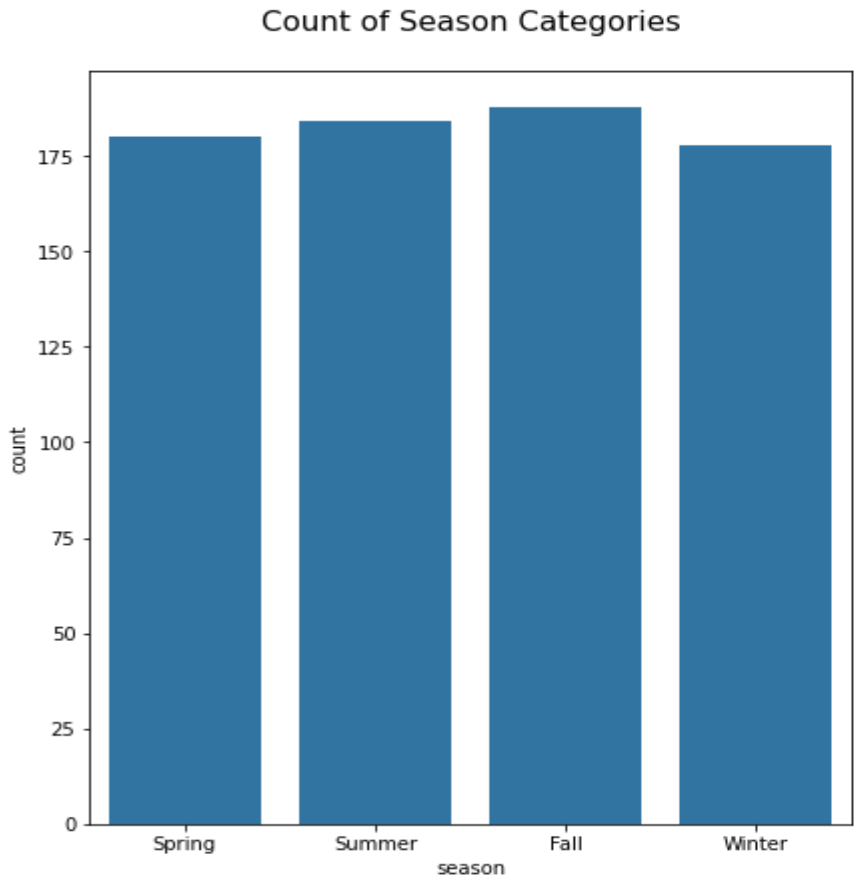
1) None of the independent continuous variables have a very strong correlation with each other

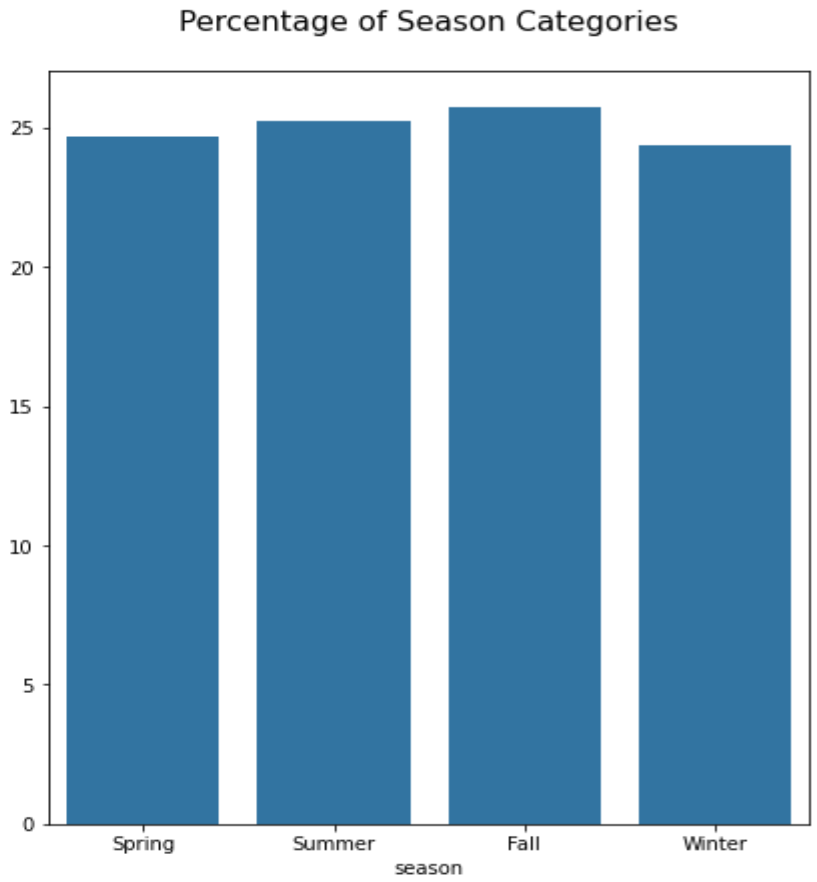
2) None of the continuous variables have a very strong correlation with bicycle count

3) Temperature has a correlation of ~63% whereas humidity and windspeed have a very weak negative correlation of 10% and 23% respectively with bicycle count

# Categorical Variables

## 1) Season

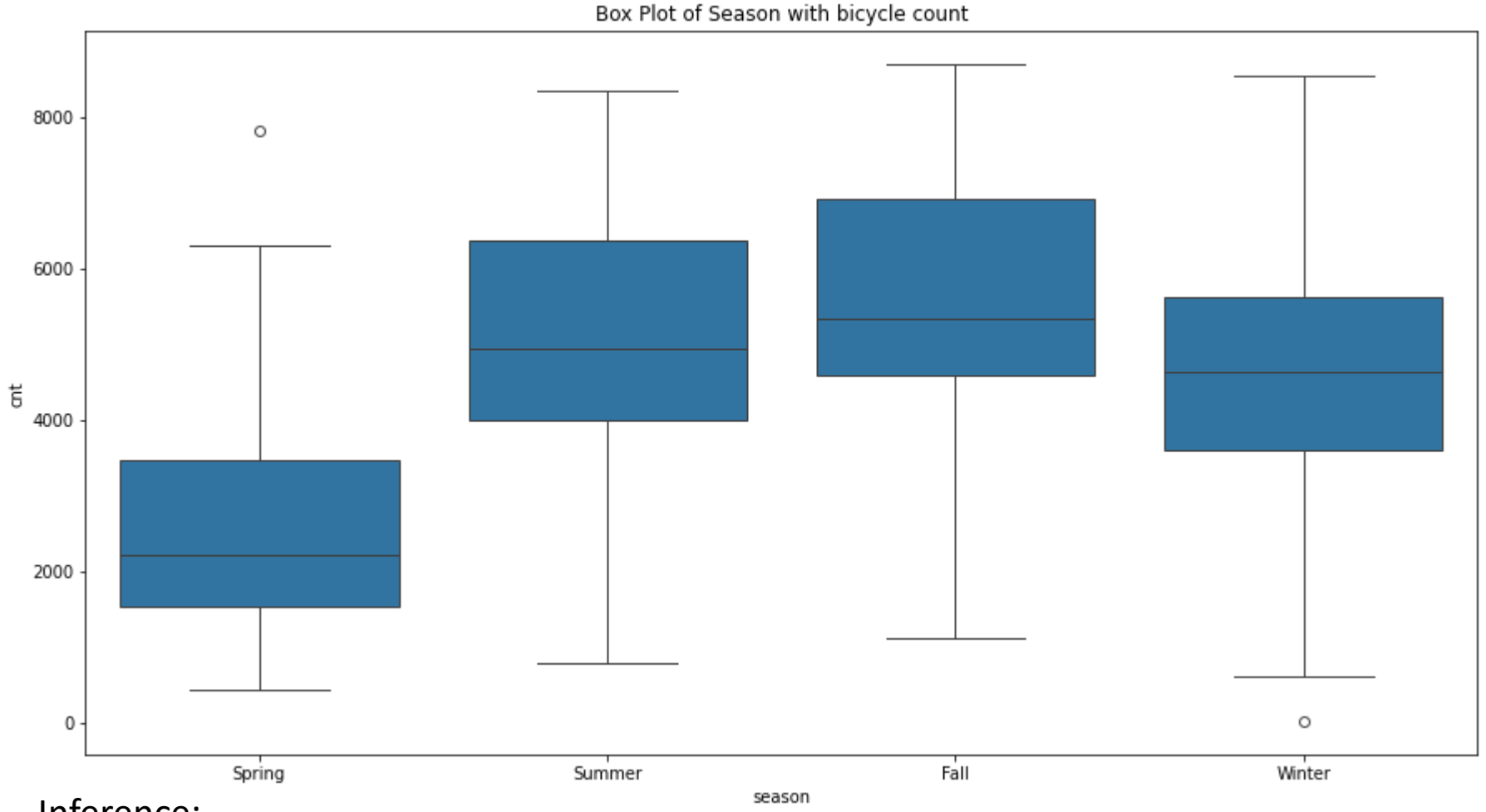



Inference:

1) The counts of each season category is almost equal

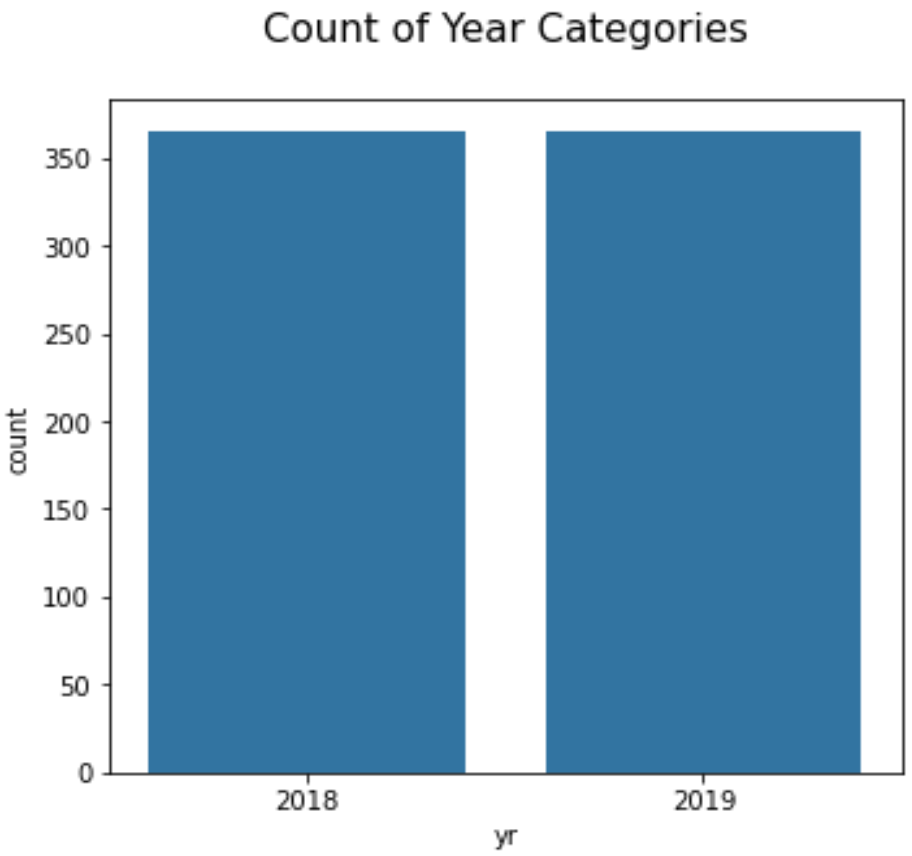
# Relation of Season with Count

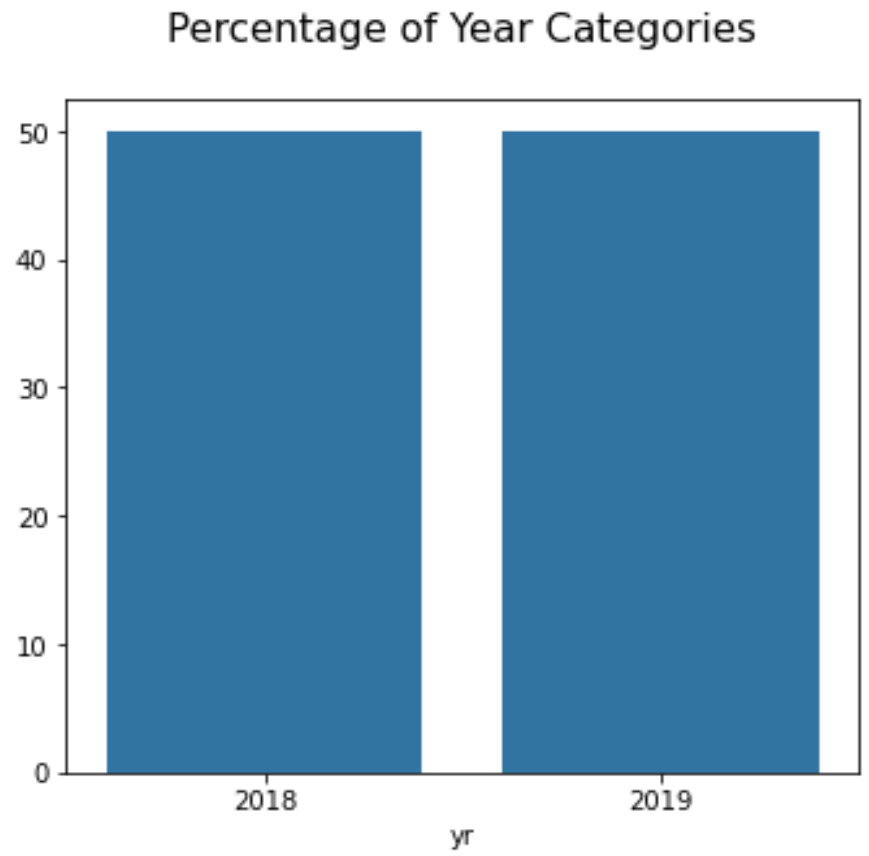



Inference:

1) The demand for bicycles is highest in fall followed by summer and least in spring
2) The total count and median count of bicycle in fall season is 1061129 and 5353
3) The total count and median count of bicycle in summer season is 918589 and 4941
4) The total count and median count of bicycle in spring season is 469514 and 2222

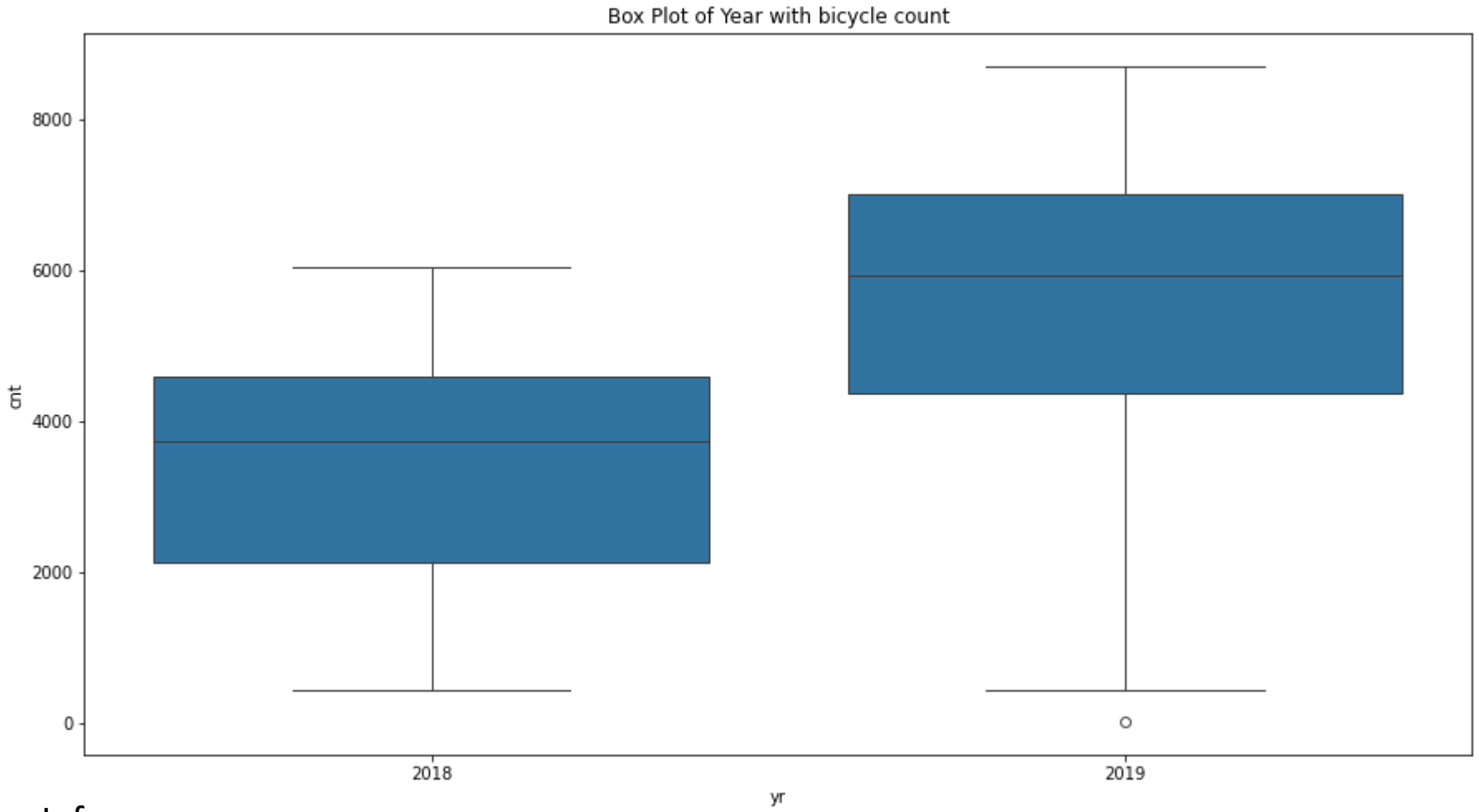
# 2) Year




Inference: The counts of each year is almost equal

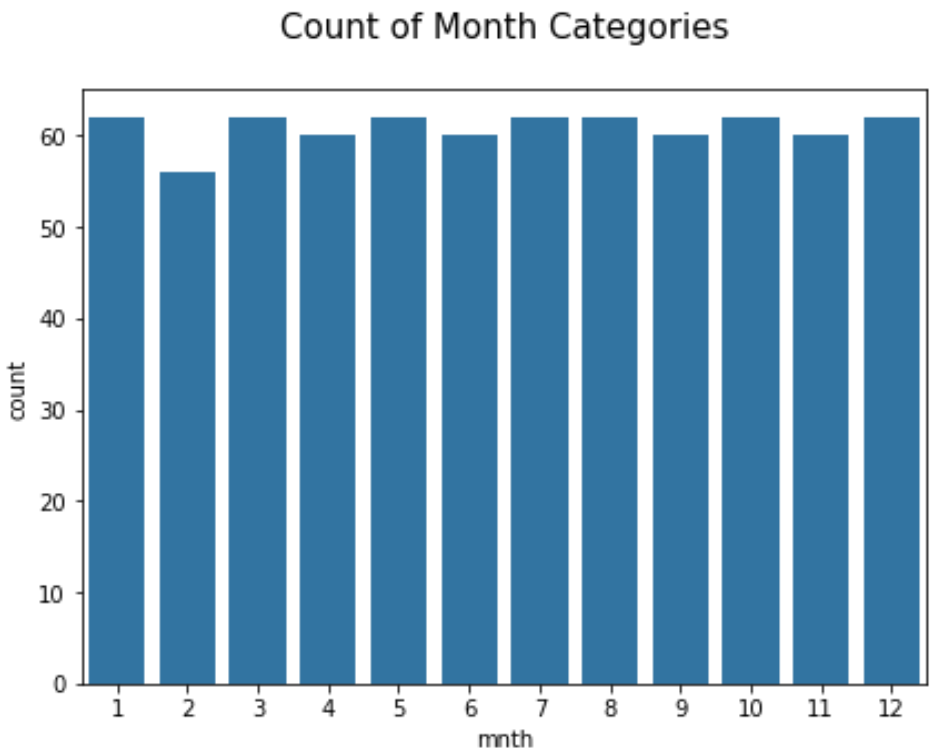
# Relation of year with bicycle count

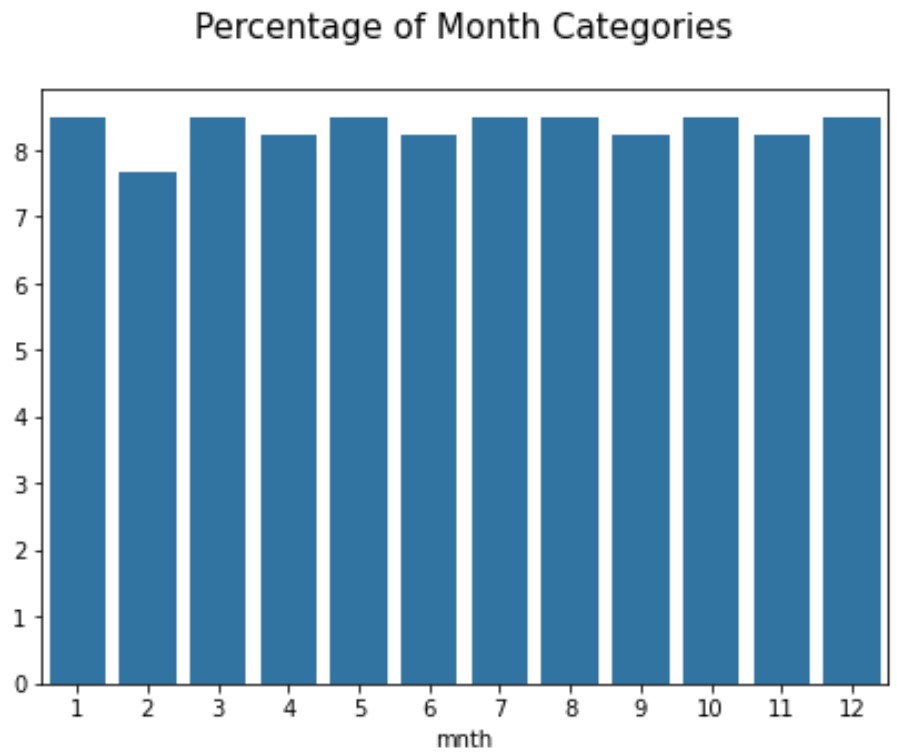



Inference:

1) The demand for bicycles is higher in 2019 than 2018
2) The total count and median count of bicycle in 2018 is 1243103 and 3740
3) The total count and median count of bicycle in 2018 is 2047742 and 5936
4) Bicycle demand has increased by almost 65% in 2019 as compared to 2018

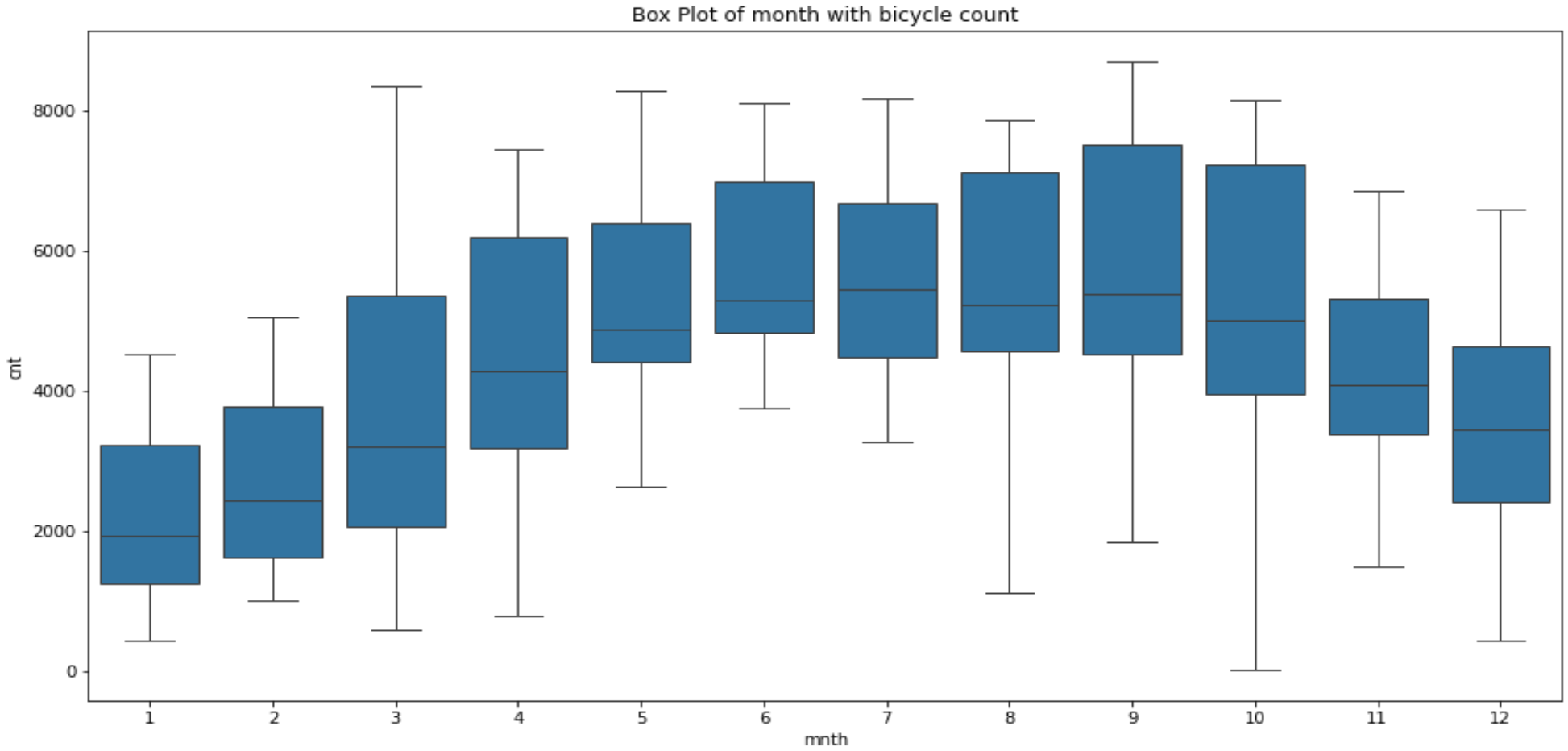
# 3) Month




Inference: Month variable has a almost uniform distribution

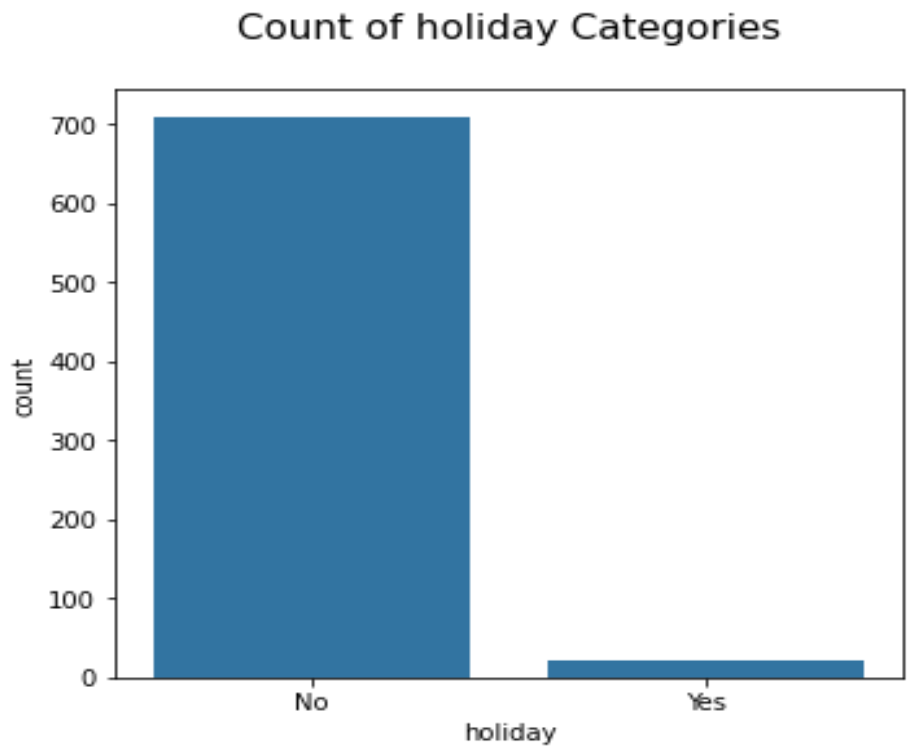
# Relation of month with bicycle count

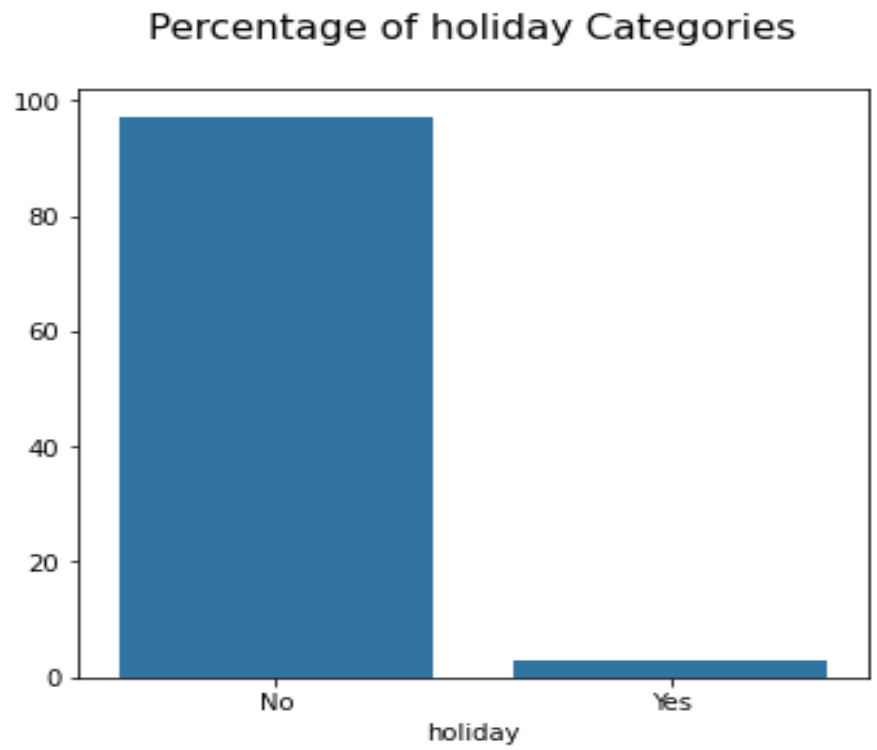



Inference:

1) The demand for bicycles increases constantly from January till September and then gradually decreases till December
2) The bicycle demand is more in April till December as compared to months from January to March.
3) The demand is highest in months of June till September

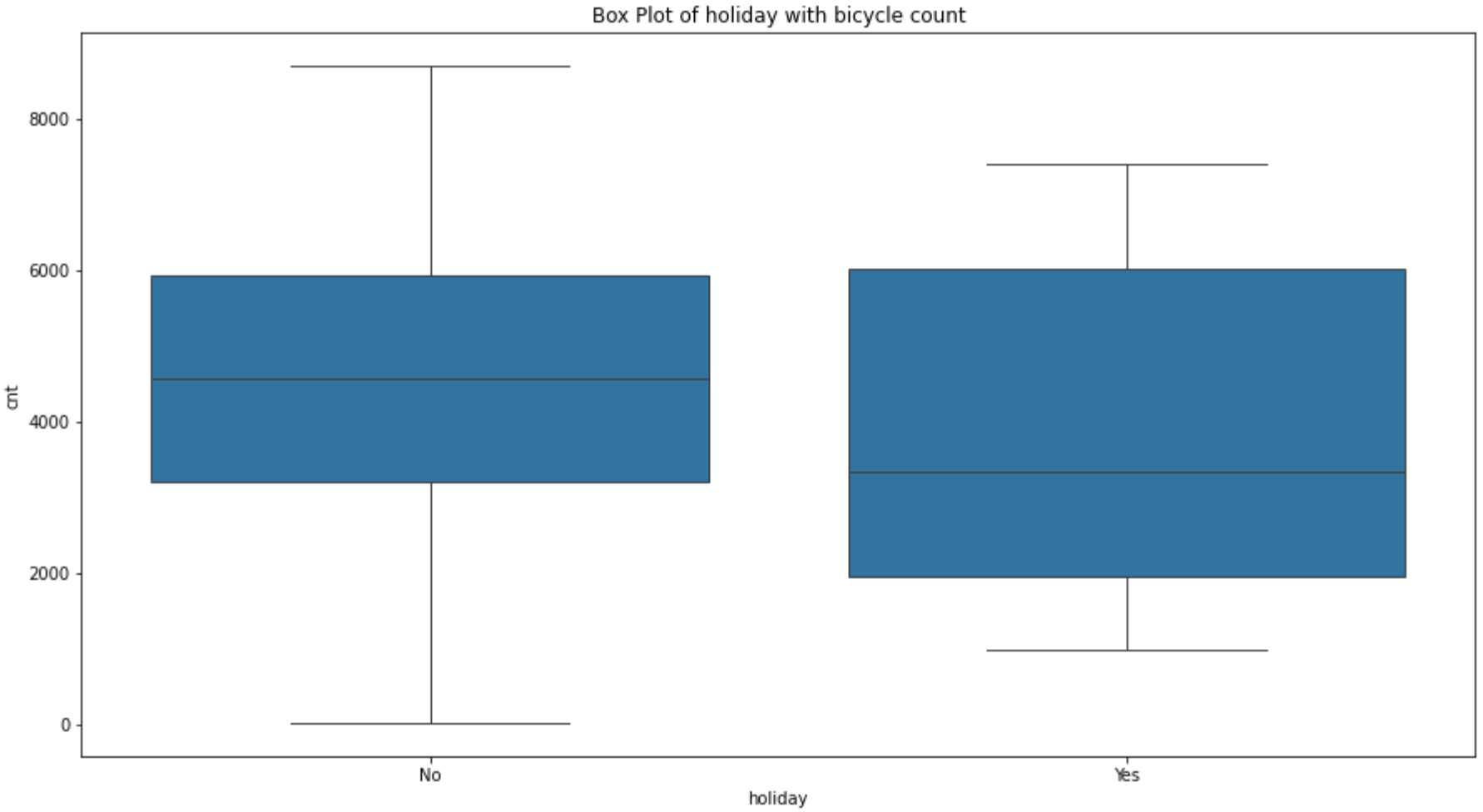
# 4) Holiday



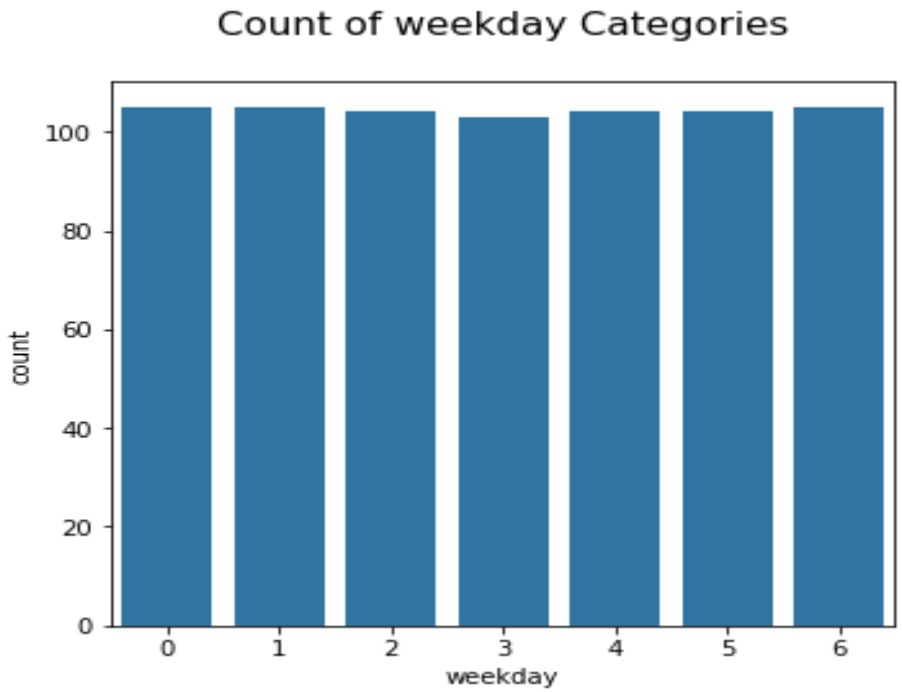

Inference:  The counts of no holiday days are much more than yes ones

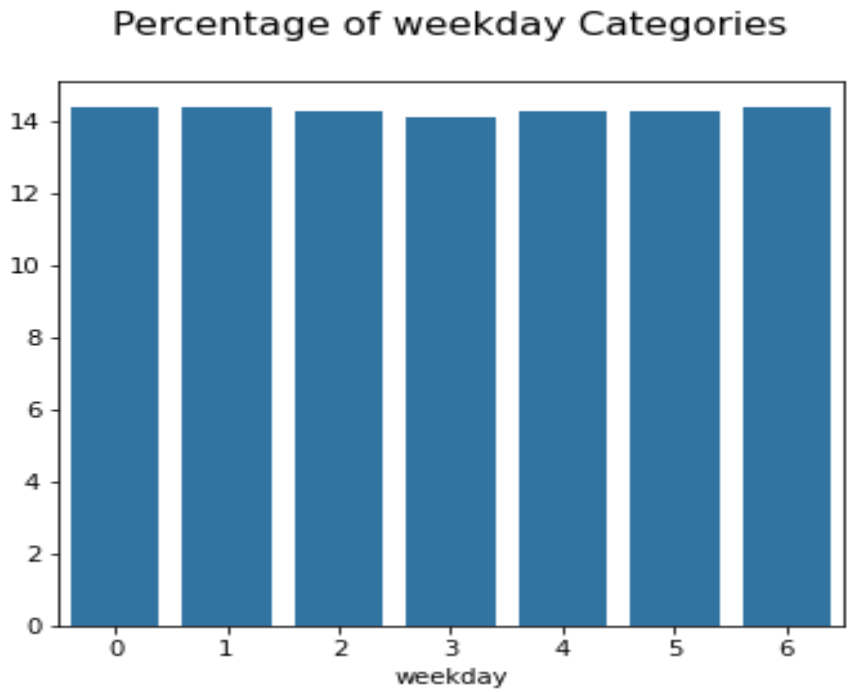
# Relation of holiday with bicycle count




Inference: The median demand is much higher on no holiday days(4563) as compared to holiday ones(3351).

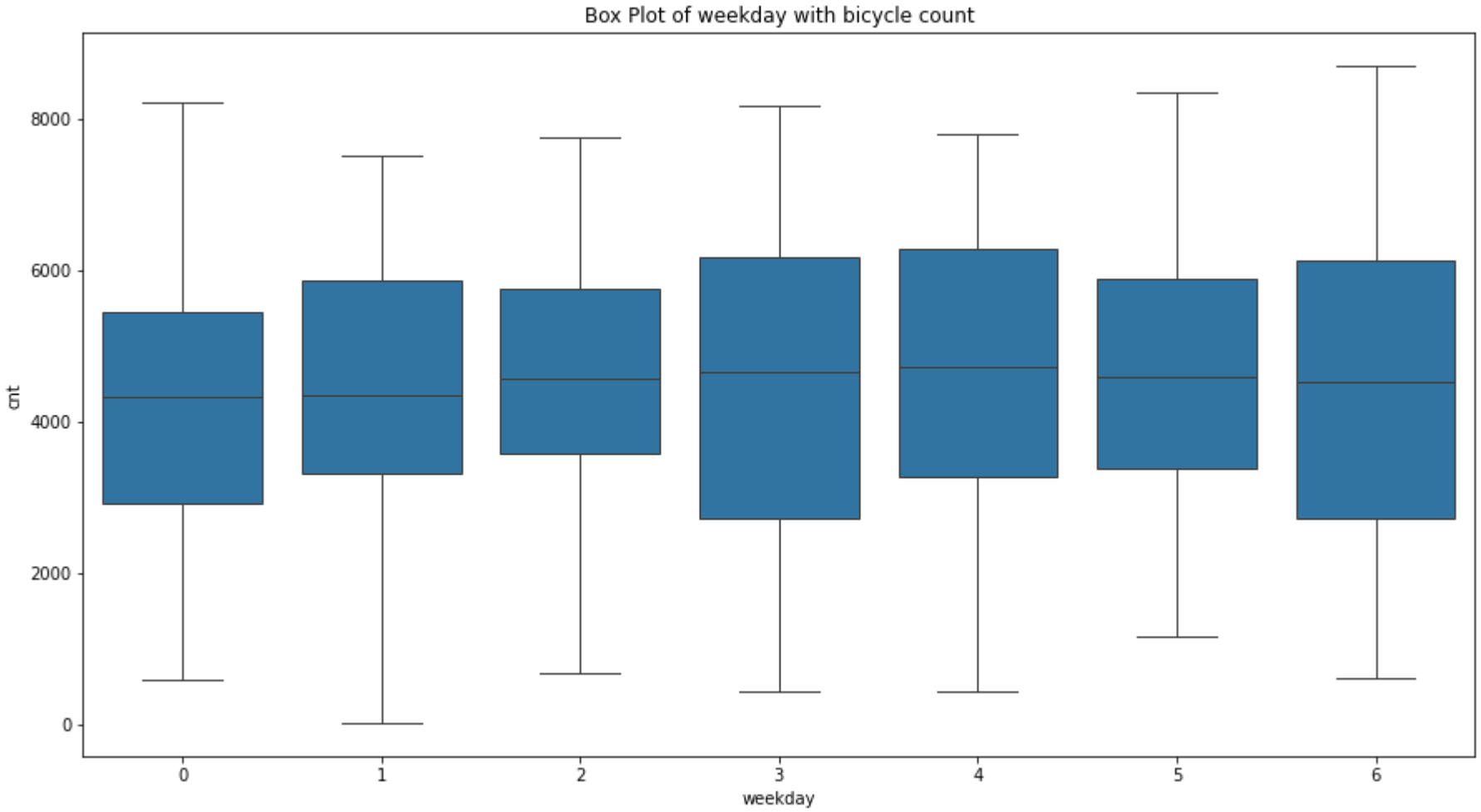
# 5) Weekday




Inference: Weekday variable has a almost uniform distribution

# Relation of Weekday with bicycle count

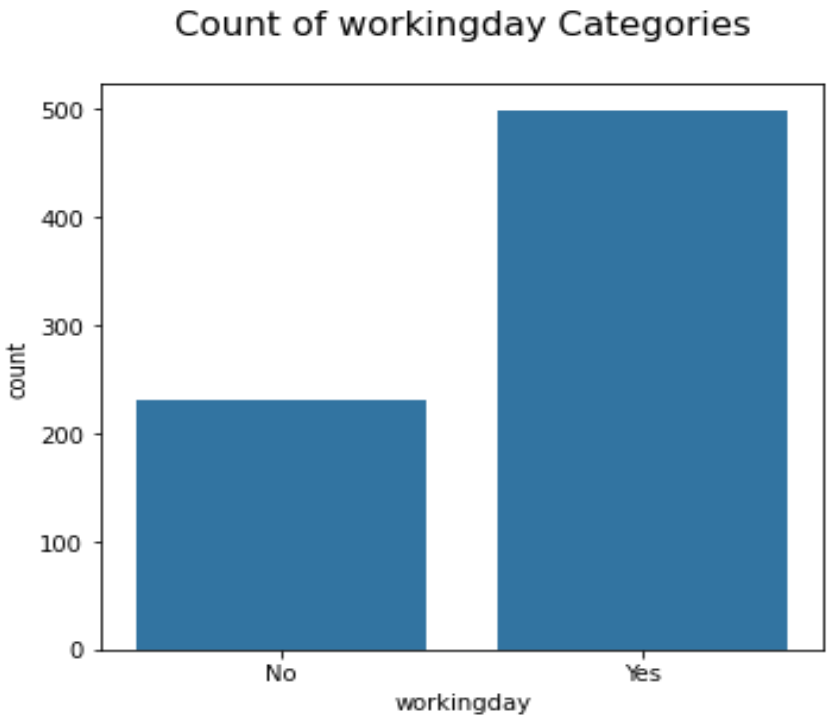


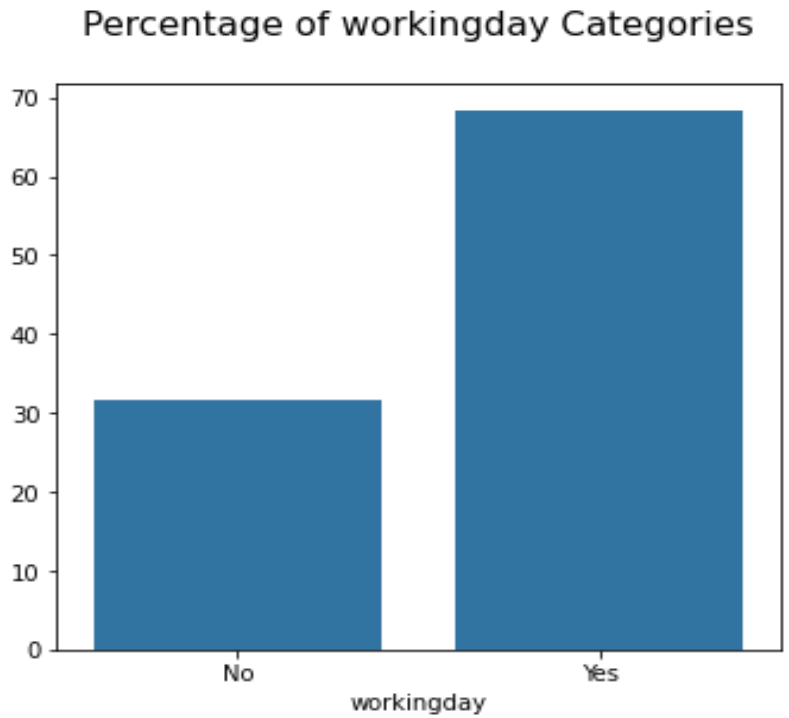

Inference:

1) All the weekday's have almost same amount of median demand of bicycles

2) 25% of bicycle demand on each day very gradually increases as week progresses

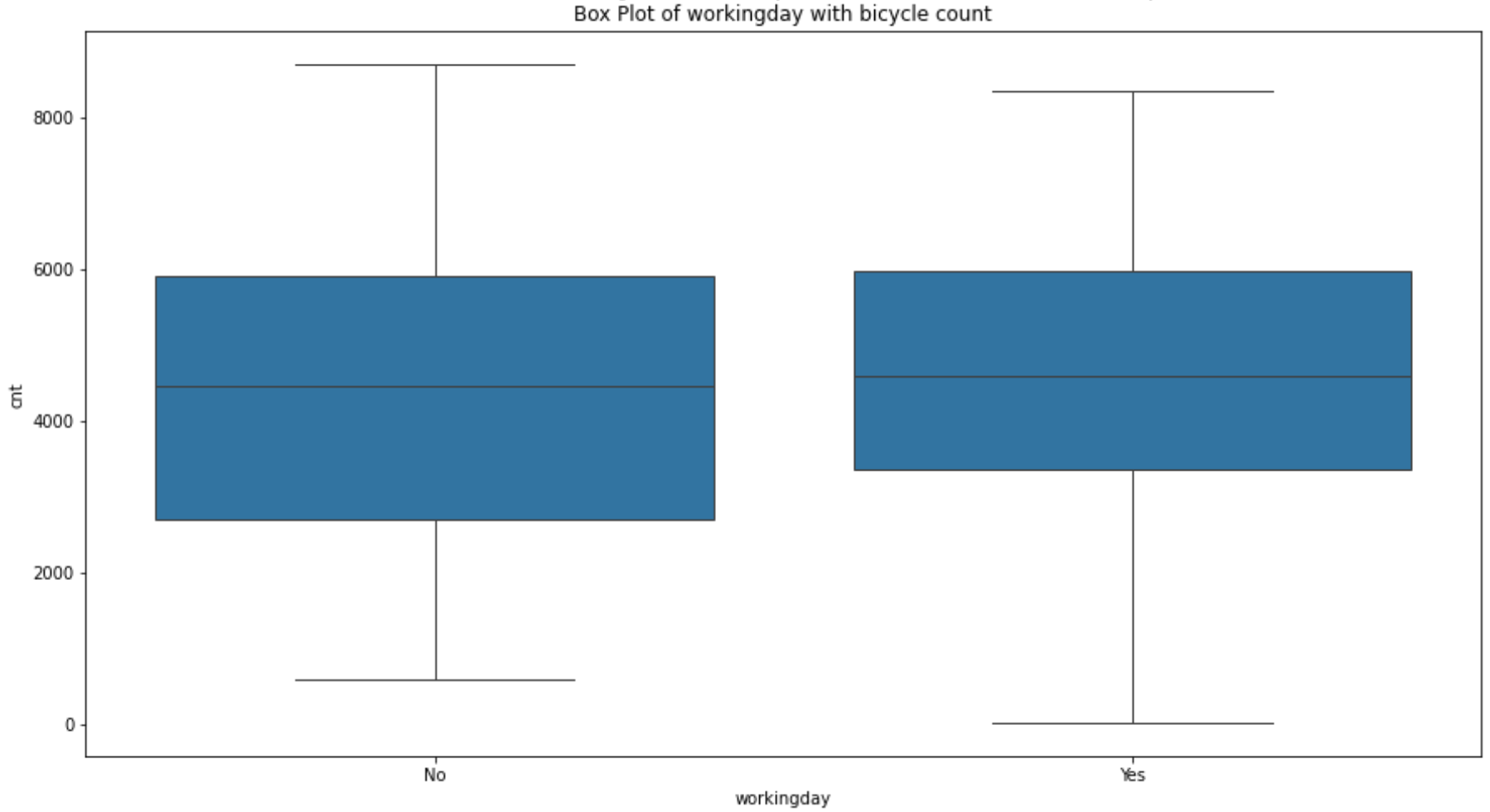
# 6) Working Day




Inference: Working day datapoints(499) are more in number than non working day datapoints(231)

# Relation of working day with bicycle count

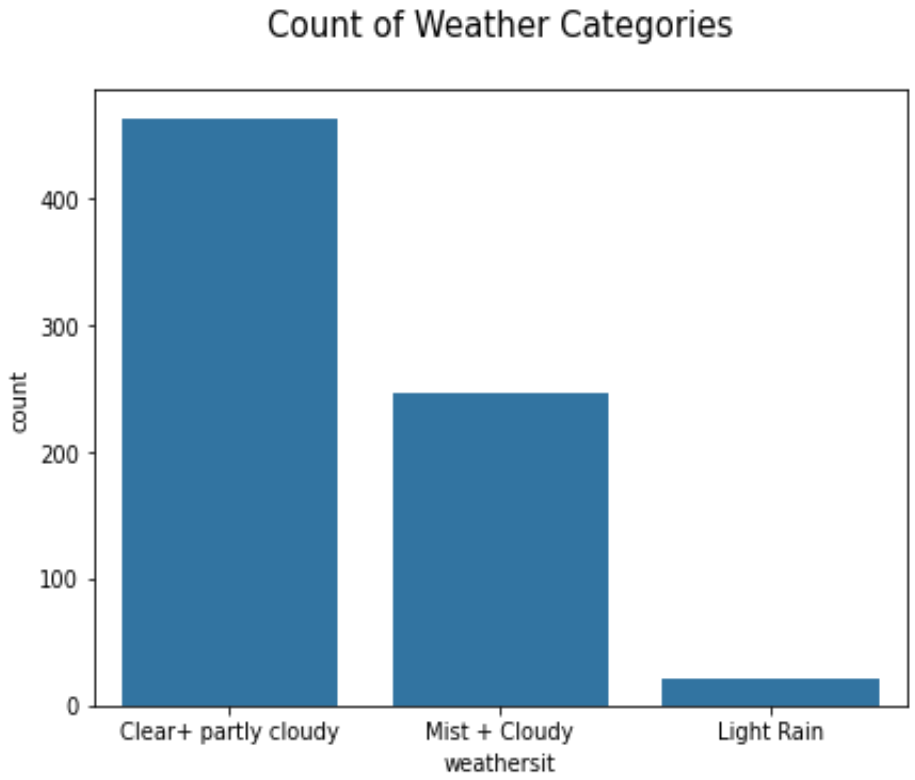


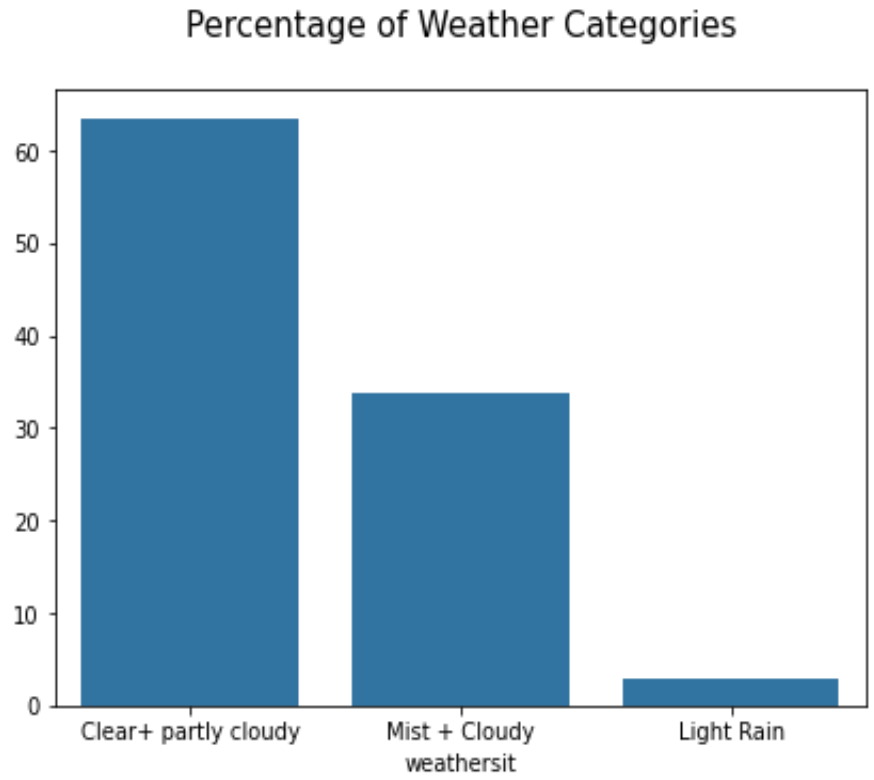

Inference:

1) The demand for bicycles is more on working days

2) 25% of the bicycle demand on working days is more than the 25% on non-working days.

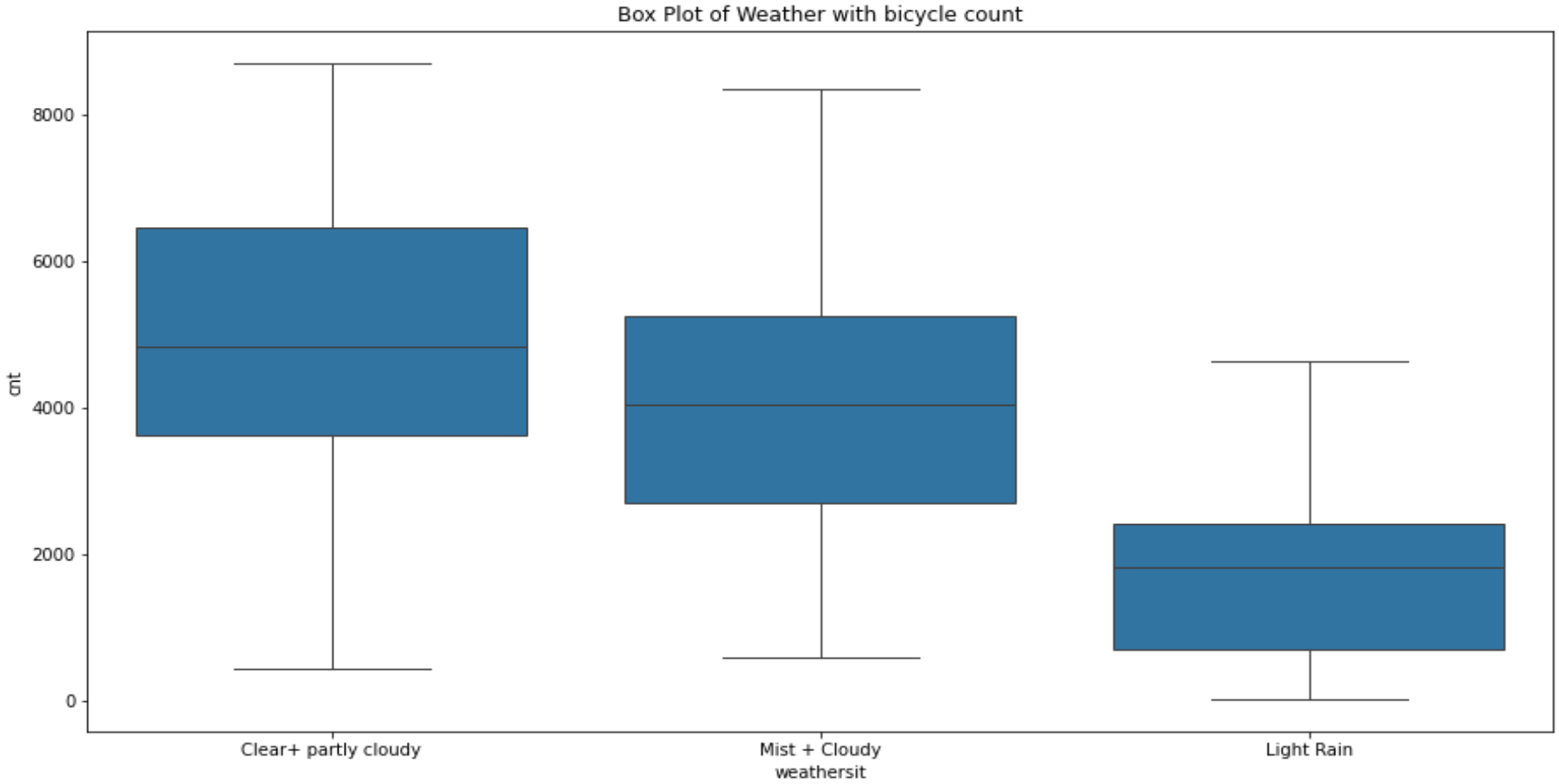
# 7) Weather




Inference: The counts of clear/partly cloudy and mist/cloudy data points is much more than light rain datapoints.

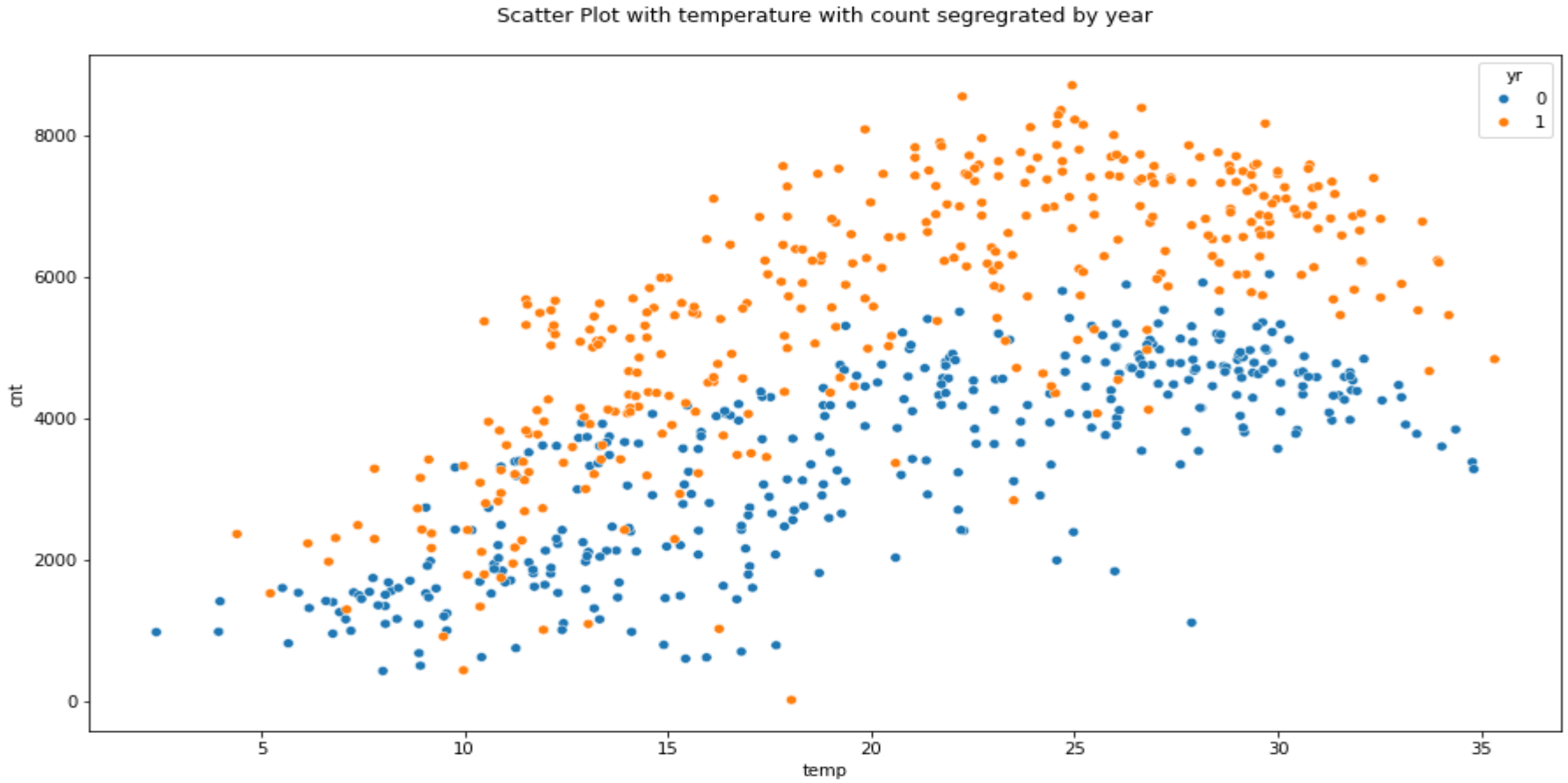
# Relation of weather with bicycle count




Inference: The demand for bicycle is maximum on clear/partly cloudy days followerd by mist/cloudy days and least on light rain days.

# Multivariate Analysis

## 1) Temperature + year with Bicycle Count

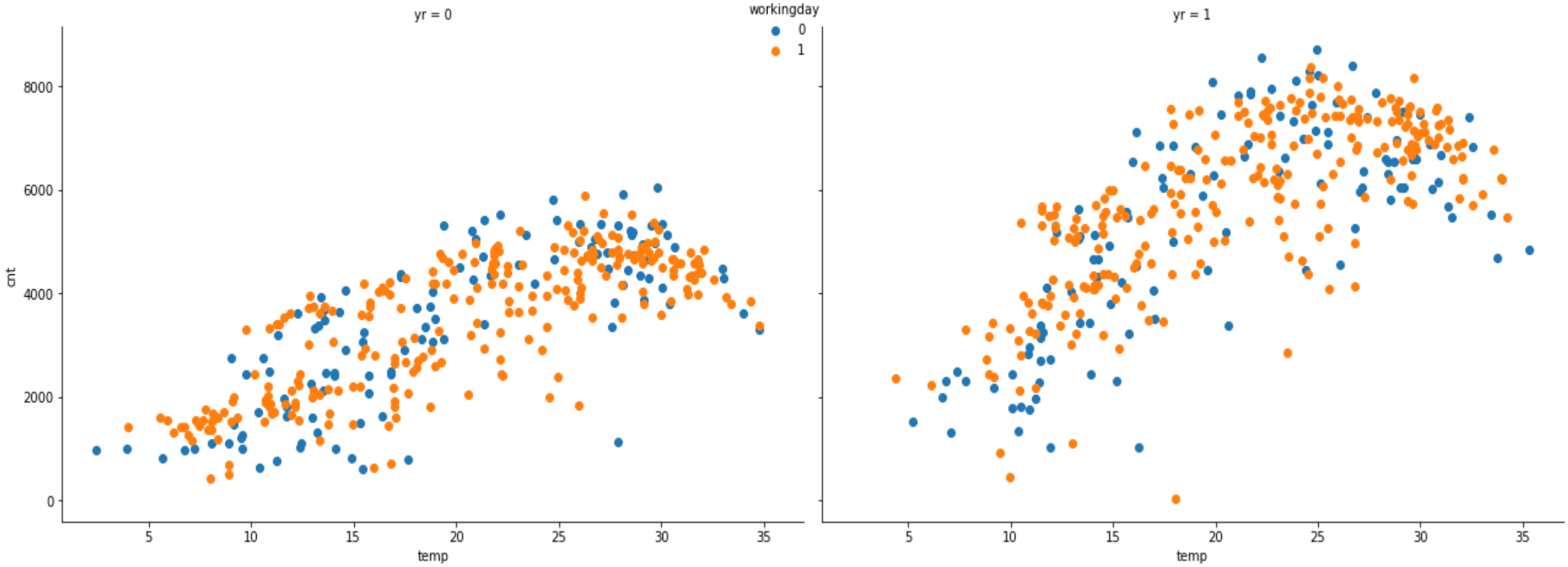



Inference:

1) The demand for bicycle is increasing with year and with increasing temperature

2) In general, the demand at the same temperature is more in 2019 than 2018.

# 2) Year + Temperature + Working Day with Bicycle count

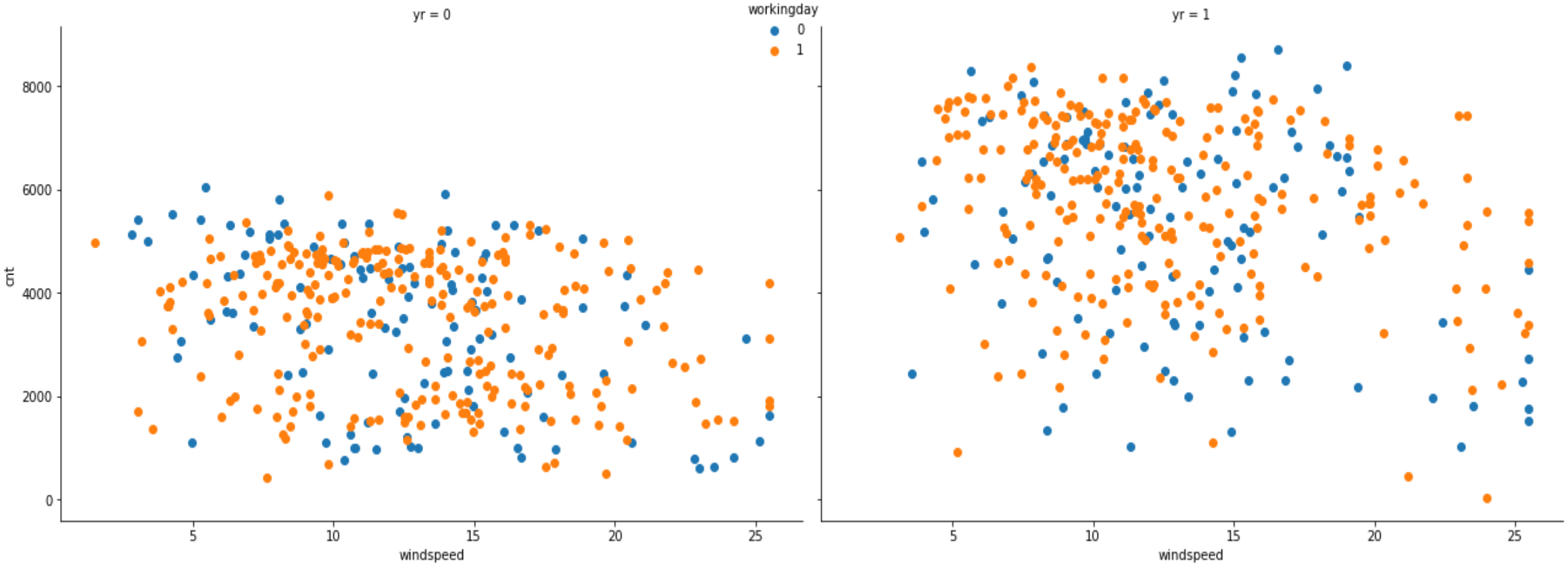



Inference:

1) The demand is low when the temperature is low irrespective of whether the day is working or not.

2) The demand is high when the temperature is low irrespective of whether the day is working or not.

# 3) Year +  Windspeed + Working Day with Bicycle Count



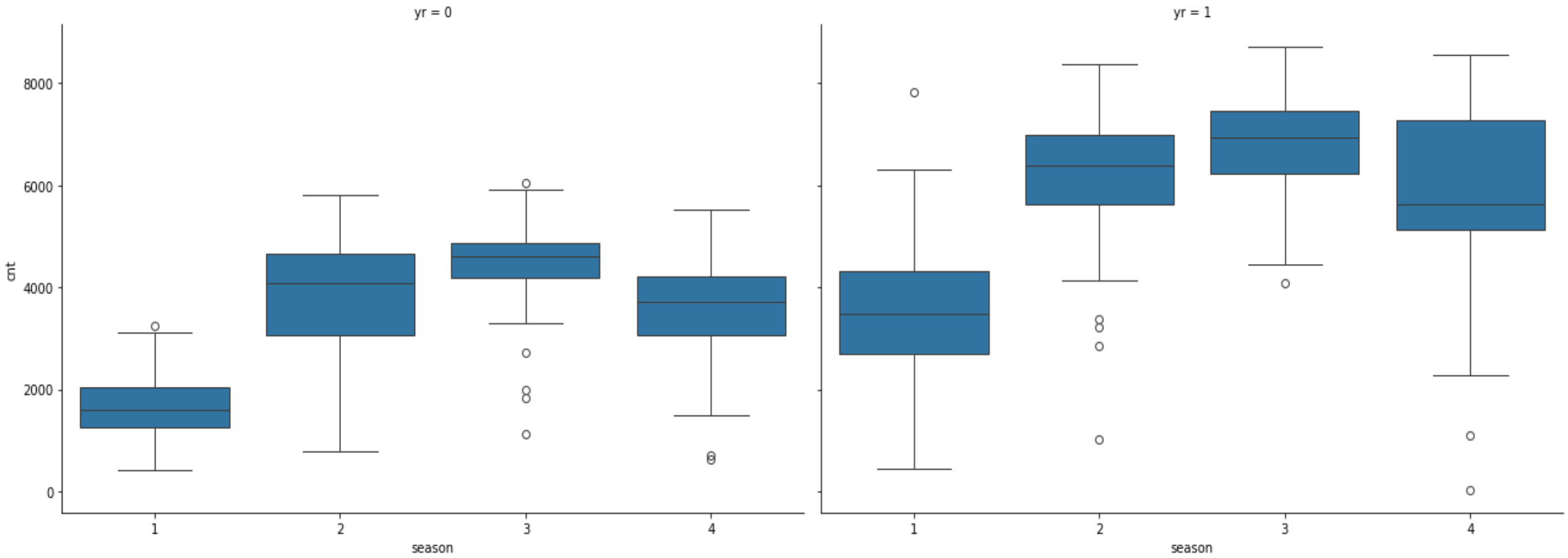

Inference:

1) When the windspeed is low(>18) the demand is good irrespective of working day or no working day.

2) The demand generally decreases when windspeed is greater than 18

3) When the windspeed is high(>18) and when it is a working day the demand is still high than at the same speed on non- working days

# 4) Year + Season with Bicycle Count




Inference:

1) The demand for bicycle is increasing with year and with increasing temperature

2) In general, the demand in the same season is more in 2019 than 2018.

# Statistical Analysis

1) Multicollinearity : The final list of features with VIF < 5 is

| | |
|---|---|
| temp | 3.88793 |
| mnth_7 | 2.53216 |
| mnth_8 | 2.28102 |
| weathersit_2 | 2.16194 |
| mnth_6 | 2.15102 |
| yr | 1.95714 |
| weekday_1 | 1.90206 |
| weekday_6 | 1.80422 |
| weekday_5 | 1.80385 |
| mnth_9 | 1.77822 |
| weekday_4 | 1.76180 |
| weekday_2 | 1.75313 |
| weekday_3 | 1.74774 |
| hum | 1.73047 |
| mnth_5 | 1.72327 |
| mnth_12 | 1.48606 |
| mnth_10 | 1.46511 |
| mnth_2 | 1.45249 |
| mnth_4 | 1.44928 |
| mnth_3 | 1.42727 |
| mnth_11 | 1.39955 |
| windspeed | 1.14197 |
| holiday | 1.13690 |

# 2) Welch t-test

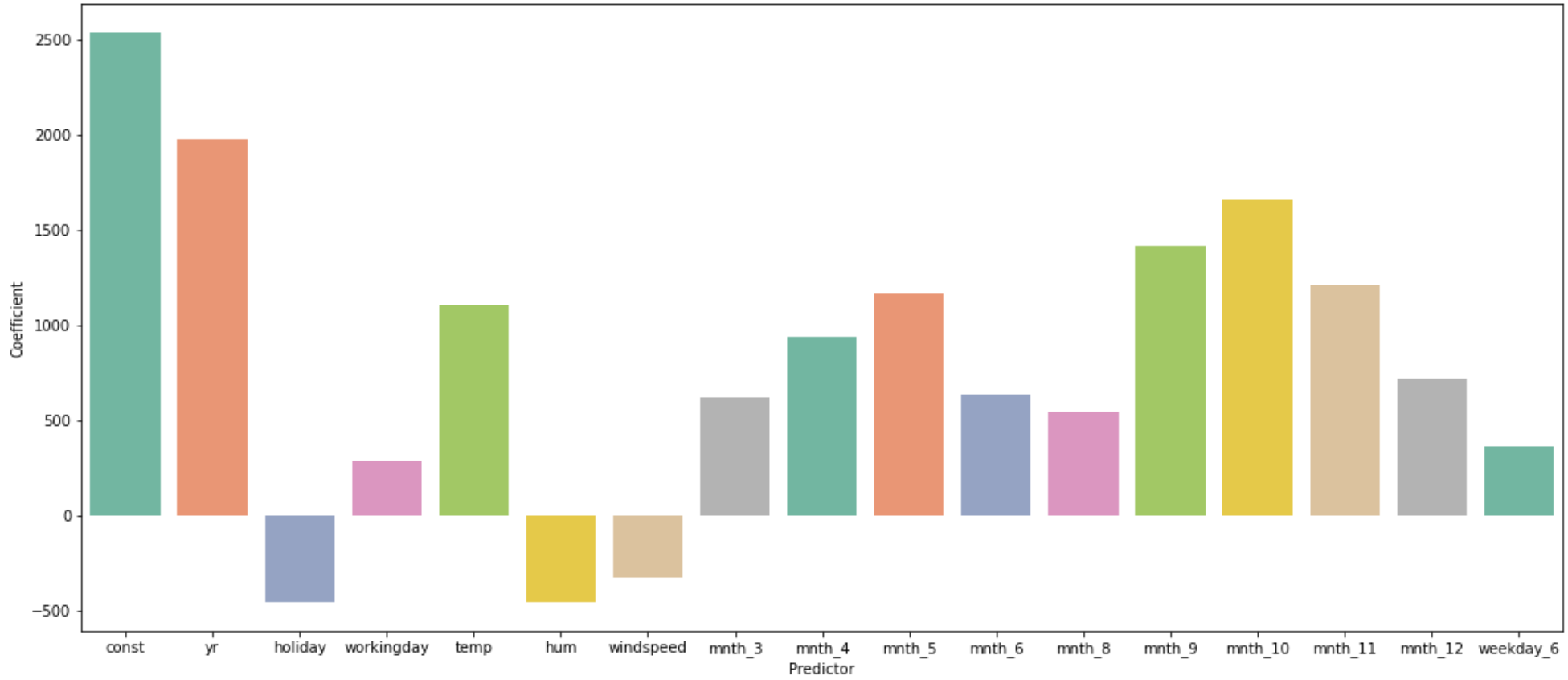
- For variable year and holiday we rejected the null hypothesis as p-value obtained was < 0.05. It means there is a significant difference between the two groups of year on relationship with bicycle count
- For variables workday and holiday we failed to reject the null hypothesis as p-value obtained was > 0.05. It means there is no significant difference between the two groups of holiday or two groups of working data on relationship with bicycle count

# 3) Welch anova test

- For variables season, month and weather situation we rejected the null hypothesis as p-value obtained was < 0.05. It means there is a significant difference between the respective groups of season, month and weather situation with respect to bicycle count.
- For variable weekday we failed to reject the null hypothesis as p-value obtained was > 0.05. It means there is no significant difference between the respective groups of weekday with respect to bicycle count.

# Model Training

- We divided the dataset into 90% training and 10% testing datasets respectively.
- After the few iterations, we were able to get a statistically significant multiple linear regression model with statistically significant variables for bicycle prediction. These variables are:



Cont...

| Predictor | Coefficient |
|---|---|
| const | 2539.75617 |
| yr | 1977.31669 |
| holiday | -454.29925 |
| workingday | 285.90375 |
| temp | 1106.63833 |
| hum | -456.31968 |
| windspeed | -327.26445 |
| mnth_3 | 623.34261 |
| mnth_4 | 938.51983 |
| mnth_5 | 1163.85114 |
| mnth_6 | 633.14090 |
| mnth_8 | 542.06614 |
| mnth_9 | 1418.34258 |
| mnth_10 | 1658.77665 |
| mnth_11 | 1215.09678 |
| mnth_12 | 720.10296 |

Inference:

1) The variables holiday, humidity, windspeed have negative coefficients, i.e when they increase the demand of bicycle decreases
2) Rest all variables have positive coefficients.

# Final Multiple Linear Equation

(2539.76 * const) + (1977.32 * yr) + (-454.3 * holiday) + (285.9 * workingday) + (1106.64 * temp) + (-456.32 * hum) + (-327.26 * windspeed) + (623.34 * mnth_3) + (938.52 * mnth_4) + (1163.85 * mnth_5) + (633.14 * mnth_6) + (542.07 * mnth_8) + (1418.34 * mnth_9) + (1658.78 * mnth_10) + (1215.1 * mnth_11) + (720.1 * mnth_12) + (361.88 * weekday_6)

# Model training metrics

- R2 score: 0.804
- Adjusted R2 score : 0.80
- P-value for F-test : <0.05
- Mean absolute error : ~0
- Durbin- Watson: 1.985

# Linear Regression Assumptions Check

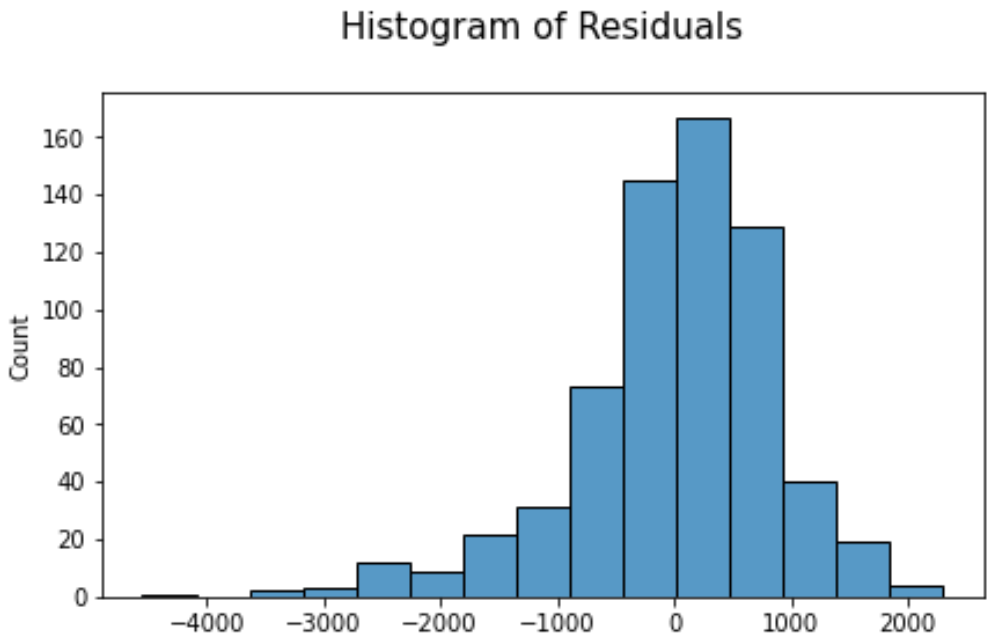
1) Predictors should have a linear relationship: Verified this in EDA section of continuous variables

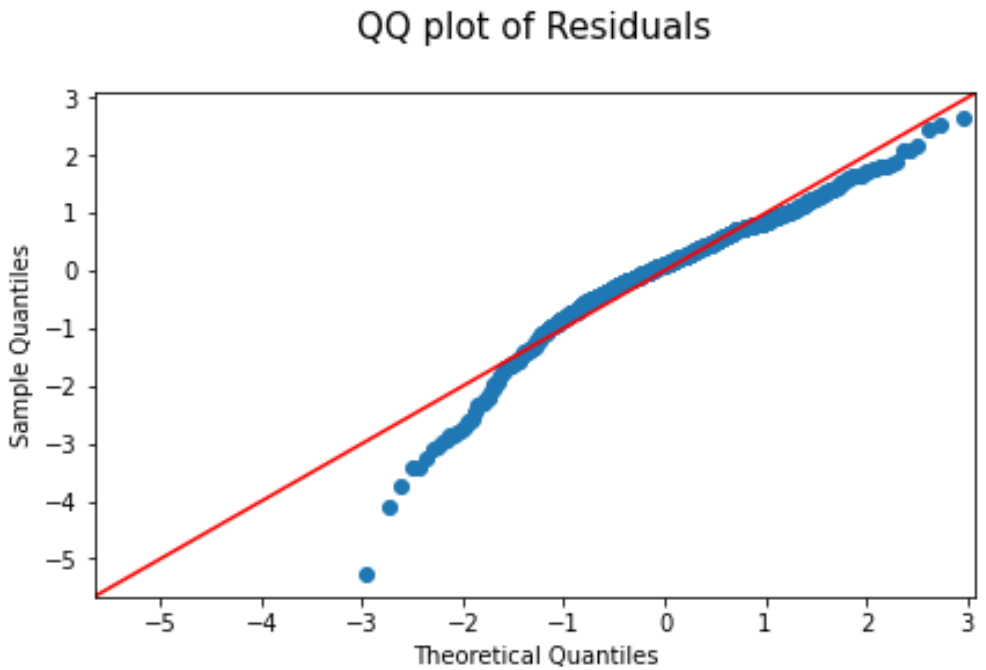
2) No Multicollinearity among predictors: Removed multicollinearity in Statistical Analysis section

3) Observations must be independent: All observations are independent or no autocorrelation is present as shown from Durbin-watson test. The value of the test statistic is ~2 which shows no autocorrelation

# Cont....

## 4) Errors should have a normal distribution



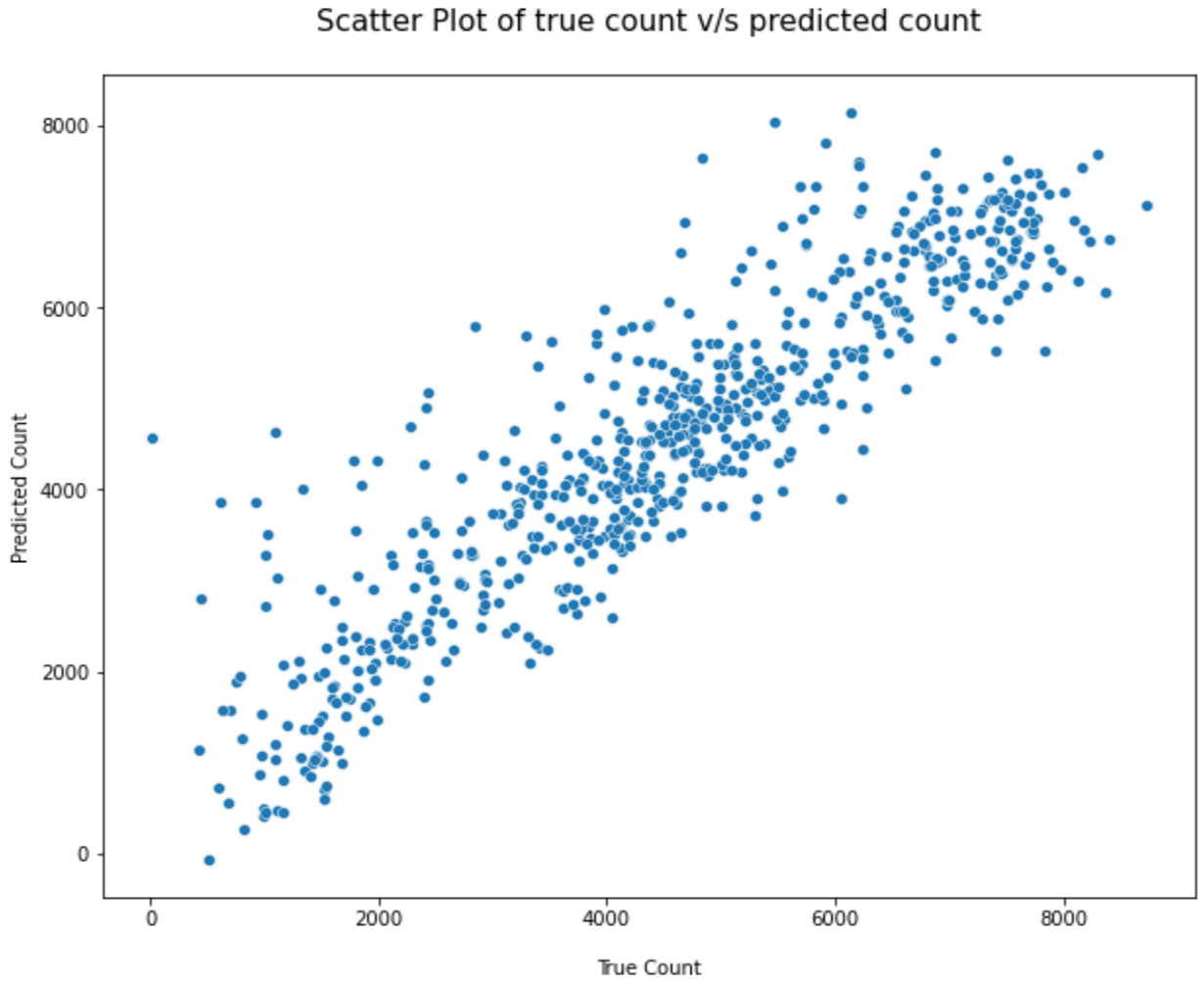

Inference:

1) Residuals have a mean of zero as shown in histogram

2) Residuals follow an approximate normal distribution as shown in QQ-plot

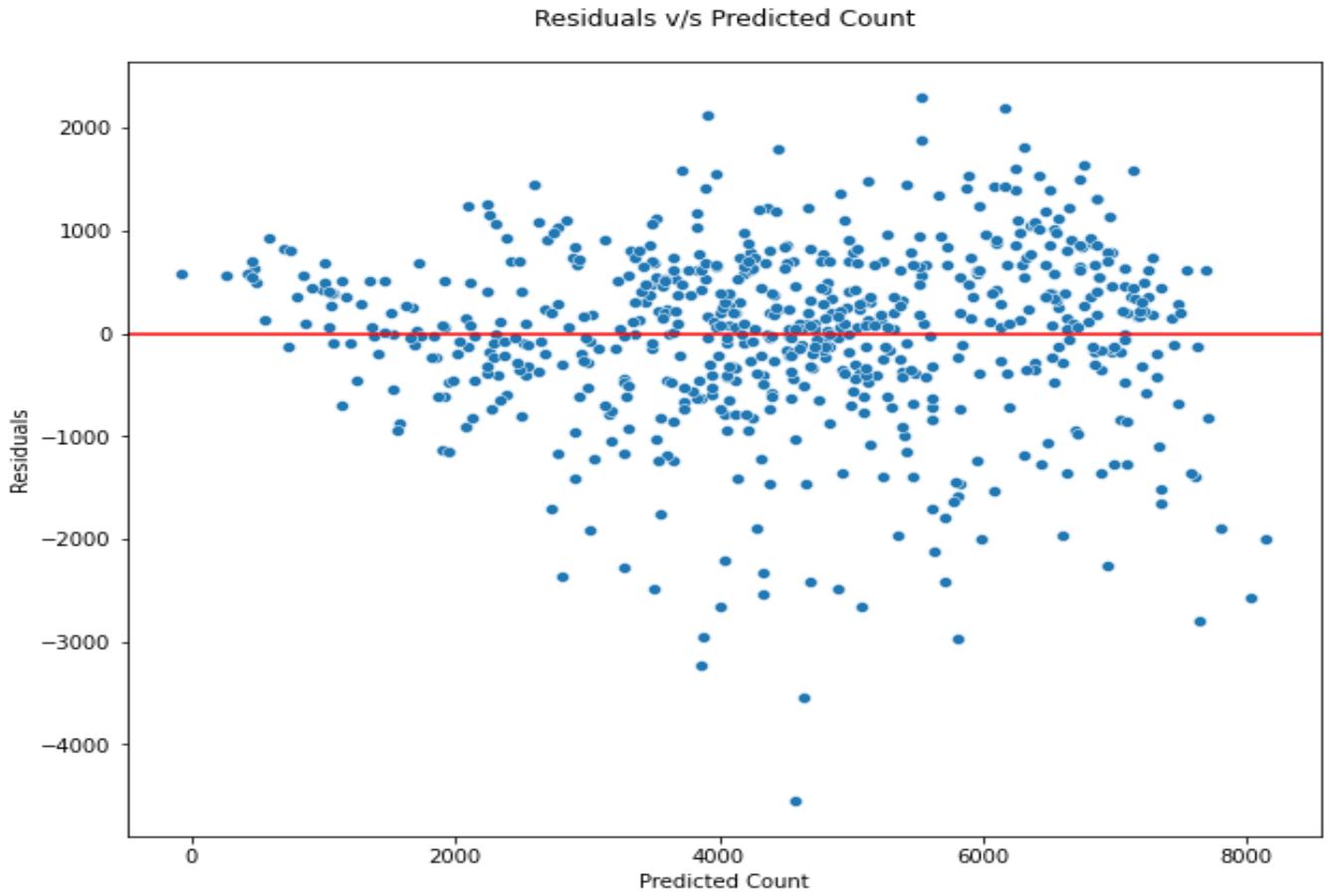
# Cont...

## 5) Residuals should have a constant variance (Homoscedasticity Check)




Inference: Errors have a constant variance

# Cont…




Inference: The spread of Residuals has no pattern

# Inference about Model:

1) All the assumptions of linear regression, i.e linear relationship of predictors with target, no multicollinearity among predictors, residuals are normally distributed and residuals having constant variance are all satisfied

2) The model is statiscally significant as seen from the p-value of F-test. It means atleast one of the predictor variables is statiscally significant in predicting bicycle count

3) The final model has all predictors with p-value less than 0.05, thus all predictors of the final model are statiscally significant in predicting bicycle count

# Test Dataset Prediction

- R2 score: 0.82
- Mean absolute error: 79.78
- Inference: The Training R2 score was 80% and of test is 82%, so the model is not overfitting and generalizing well.

# Final Recommendations

- Top 5 variables which increase bicycle demand, higher the value of coefficient more they increase the demand

| Predictor | Coefficient |
|---|---|
| yr | 1977.31669 |
| mnth_10 | 1658.77665 |
| mnth_9 | 1418.34258 |
| mnth_11 | 1215.09678 |
| mnth_5 | 1163.85114 |

# Cont…

- Top 3 variables which decrease bicycle demand, lower the value of coefficient more they decrease the demand

| Predictor | Coefficient |
|---|---|
| windspeed | -327.26445 |
| holiday | -454.29925 |
| hum | -456.31968 |