# Capstone Project I: Milestone Report

---

## Project Objective

**Predict the Bike Rental volume from the dataset given by Capital Bike Sharing System. Also determine the factors or features that influence Bike Rental Count most.**

The client is Capital Bikeshare System and this research to predict  bike rental count will be useful to them in knowing:

- What features in the dataset influence the bike rental count
- When is the demand for bike share program maximum during the day, season, quarter or year.
- Does weather conditions like temperature, humidity, windspeed have any impact on the demand? If yes, then is it to advantage or adverse.
- Are bike users whether Registered or Casual drive Bike Rental Count? If yes, do they have similar influence on Bike Rental Count distribution?

Capital Bikeshare posts quarterly data reports of bike trip times, start and end locations, and type of user (registered or casual). Each trip is on one line of data.  These data are readily and publicly available at https://www.capitalbikeshare.com/system-data. The data look like this:

| | instant | dteday | season | yr | mnth | hr | holiday | weekday | workingda | weathersi | temp | atemp | hum | windspee | casual | registered | cnt |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 1 | 1/1/11 | 1 | 0 | 1 | 0 | 0 | 6 | 0 | 1 | 0.24 | 0.2879 | 0.81 | 0 | 3 | 13 | 16 |
| 3 | 2 | 1/1/11 | 1 | 0 | 1 | 1 | 0 | 6 | 0 | 1 | 0.22 | 0.2727 | 0.8 | 0 | 8 | 32 | 40 |
| 4 | 3 | 1/1/11 | 1 | 0 | 1 | 2 | 0 | 6 | 0 | 1 | 0.22 | 0.2727 | 0.8 | 0 | 5 | 27 | 32 |
| 5 | 4 | 1/1/11 | 1 | 0 | 1 | 3 | 0 | 6 | 0 | 1 | 0.24 | 0.2879 | 0.75 | 0 | 3 | 10 | 13 |
| 6 | 5 | 1/1/11 | 1 | 0 | 1 | 4 | 0 | 6 | 0 | 1 | 0.24 | 0.2879 | 0.75 | 0 | 0 | 1 | 1 |
| 7 | 6 | 1/1/11 | 1 | 0 | 1 | 5 | 0 | 6 | 0 | 2 | 0.24 | 0.2576 | 0.75 | 0.0896 | 0 | 1 | 1 |
| 8 | 7 | 1/1/11 | 1 | 0 | 1 | 6 | 0 | 6 | 0 | 1 | 0.22 | 0.2727 | 0.8 | 0 | 2 | 0 | 2 |

# Data Wrangling

The Capital bikeshare datasets required little data wrangling other than renaming a few columns based on preference, formatting the date and time columns to match with the weather data.

1. **Renaming the columns as below:**

    'instant': 'rec_id',

    'dteday':'datetime',

    'holiday': 'is_holiday',

    'workingday': 'is_workingday',

    'weathersit': 'weather_condition',

    'hum': 'humidity',

    'mnth': 'month',

    'cnt': 'total_count',

    'hr': 'hour',

    'yr': 'year'

2. **There were not any missing values to drop or replace. Type casting the attributes as ‘datetime’ or ‘category’ shown below**

    stats['datetime'] = pd.to_datetime(stats.datetime)#dae time conversion

```
# categorical variables

stats['season'] = stats.season.astype('category')

stats['is_holiday'] = stats.is_holiday.astype('category')

stats['weekday'] = stats.weekday.astype('category')

stats['weather_condition'] = stats.weather_condition.astype('category')

stats['is_workingday'] = stats.is_workingday.astype('category')

stats['month'] = stats.month.astype('category')

stats['year'] = stats.year.astype('category')

stats['hour'] = stats.hour.astype('category')
```

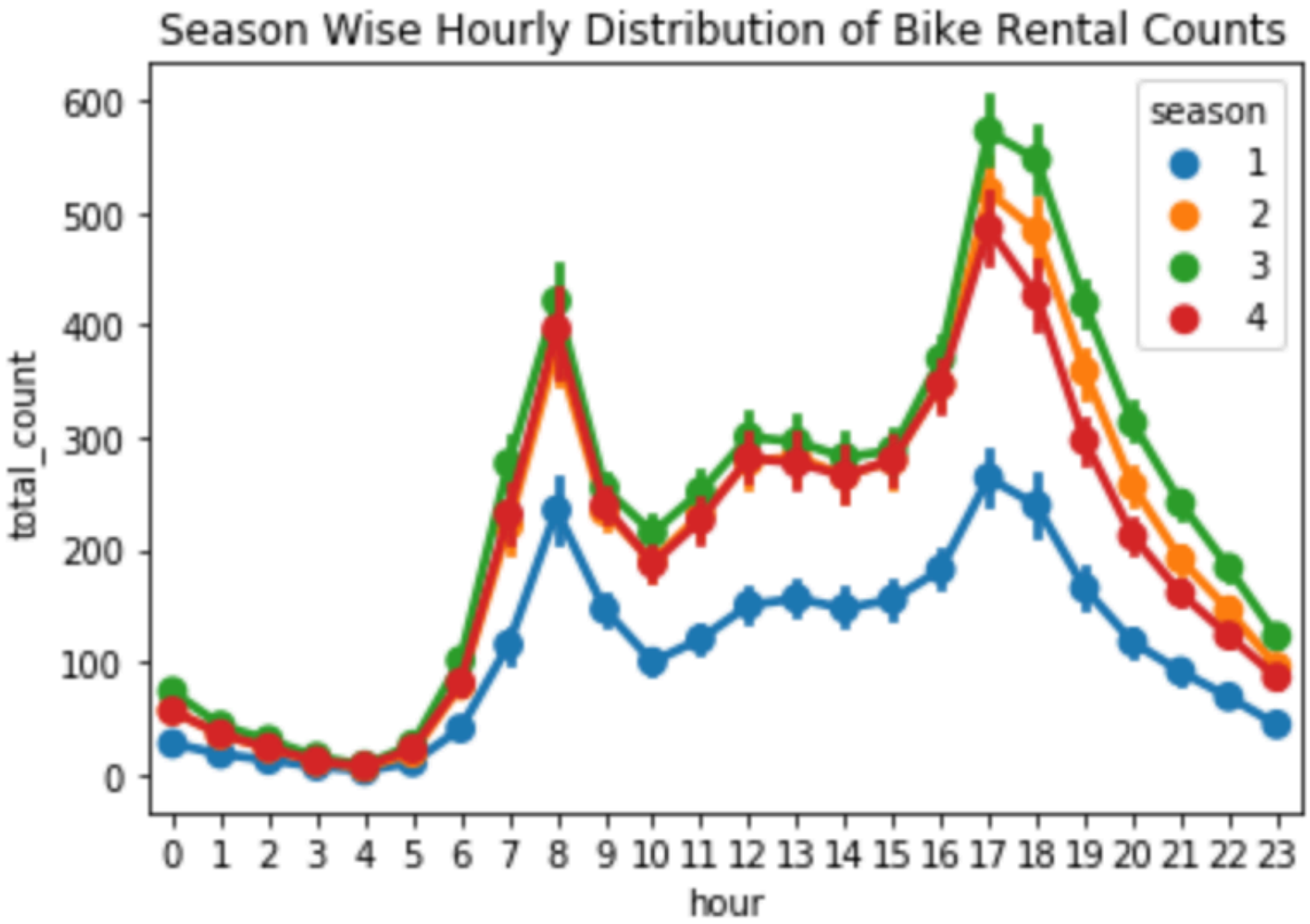
## Exploratory Data Analysis

Of the 16 features, following features showed **correlation** with the target variable ‘cnt’ i.e. Bike Rental Count:

- Season
- Month
- Temperature (temp)
- Humidity (hum)
- Windspeed
- Casual
- Registered
- Hour (hr)

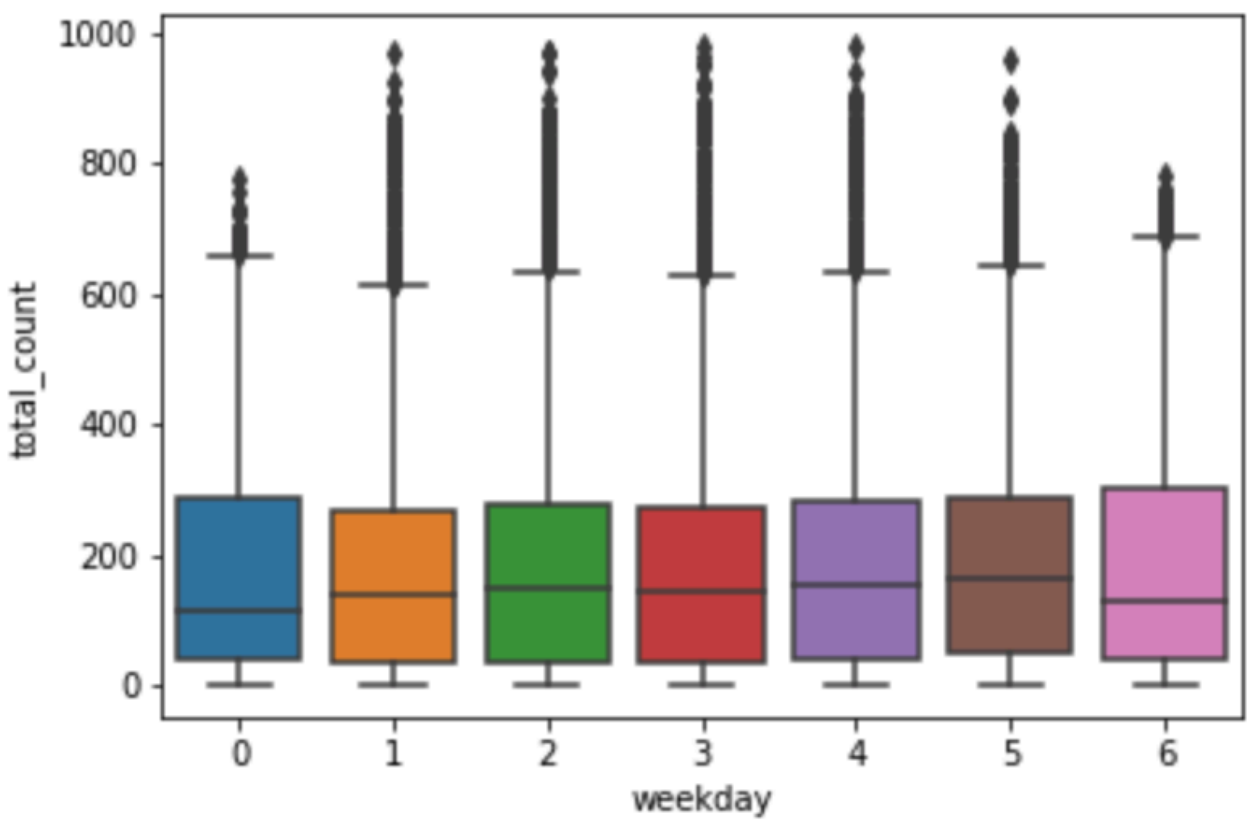
Above dependencies can be verified with following visualization graph plots and inferential statistics (code book attached too):

1. Line Chart between Bike Rental Count vs. Hour across Seasons




**Season 1 = Spring, 2 = Summer, 3= Fall, 4=Winter. Above graph shows similar trends for all seasons with counts peaking in the morning between 7 -9 AM and in the evening between 4-6 PM for the reason those are business hours. The counts are lowest for spring season (Legend 1) while highest for Fall (Legend 3) across 24 hours**

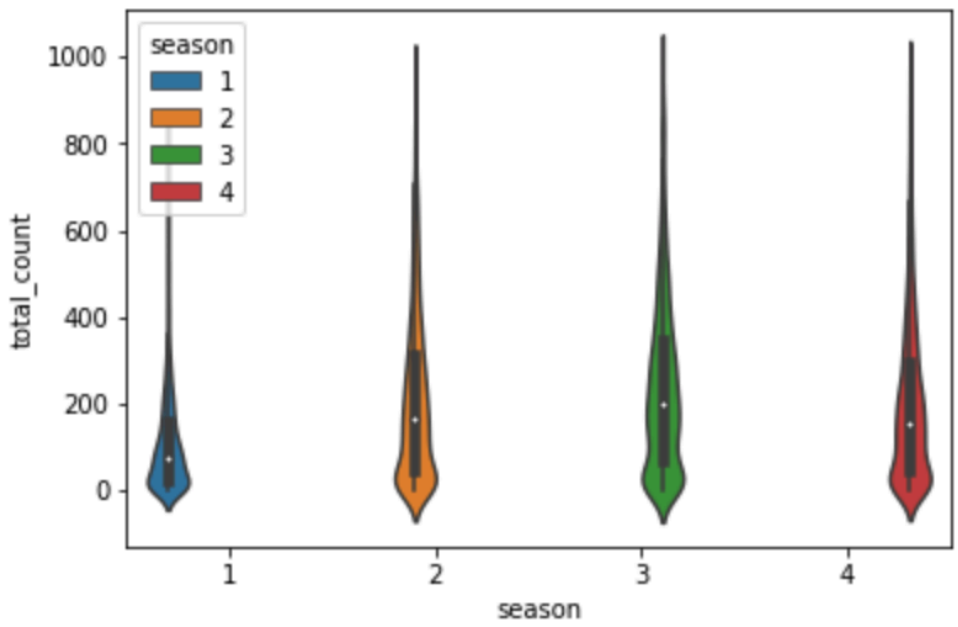
2. Box Plot between Bike Rental Count vs Weekday



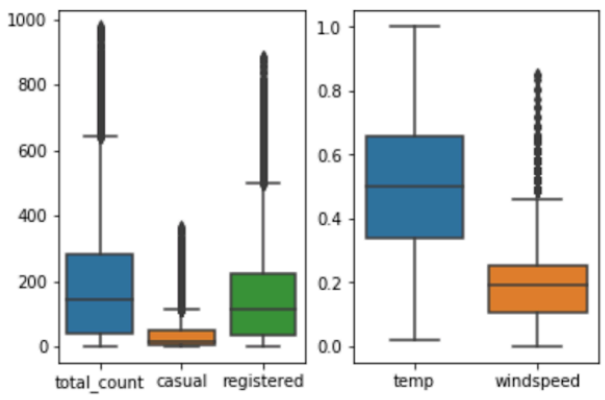
**Weekday 0= Sunday, 1= Monday, 2= Tuesday & so on. During weekdays Mon -Fri, I see median of Bike REntal count is similar as opposed to weekends.**

3. Violin Plot between Bike Rental Count vs Seasons

```
z = sn.violinplot(data=stats, x='season', y= 'total_count', hue = 'season')
```



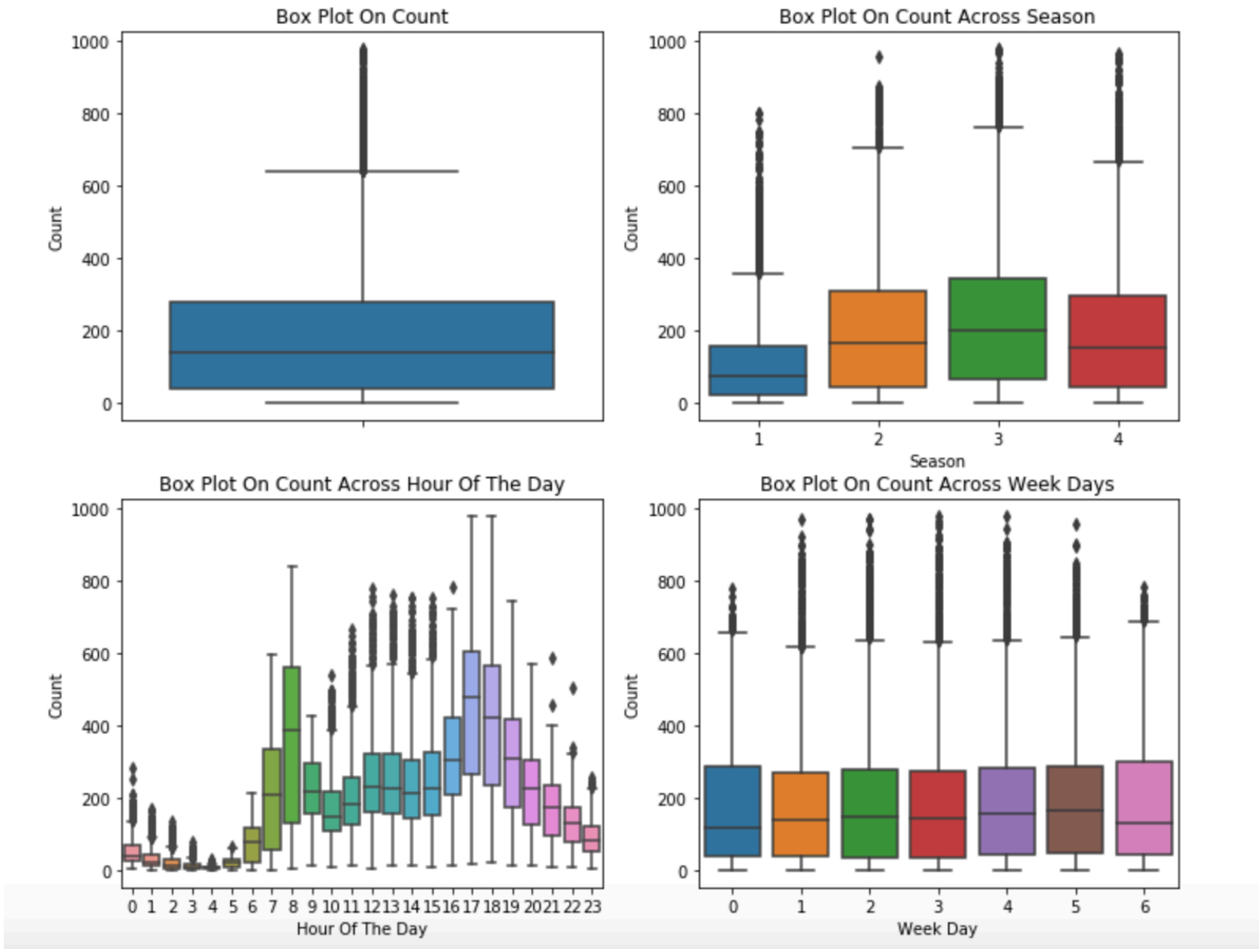

4. Box Plot between Bike Rental Count vs Casual, Registered Users




**The total, casual & registered type users show sizeable number of outlier values, however casual show lower numbers though. For weather attributes of temperature and wind speed, we see outliers only in the case of windspeed.**

**The total, casual & registered type users show sizeable number of outlier values, however casual show lower numbers though. For weather attributes of temperature and wind speed, we see outliers only in the case of windspeed.**

5. Outlier Analysis




**Outliers Analysis** At first look, "count" variable contains lot of outlier data points which skews the distribution towards right (as there are more data points beyond Outer Quartile Limit).But in addition to that, following inferences can also been made from the simple boxplots given below.

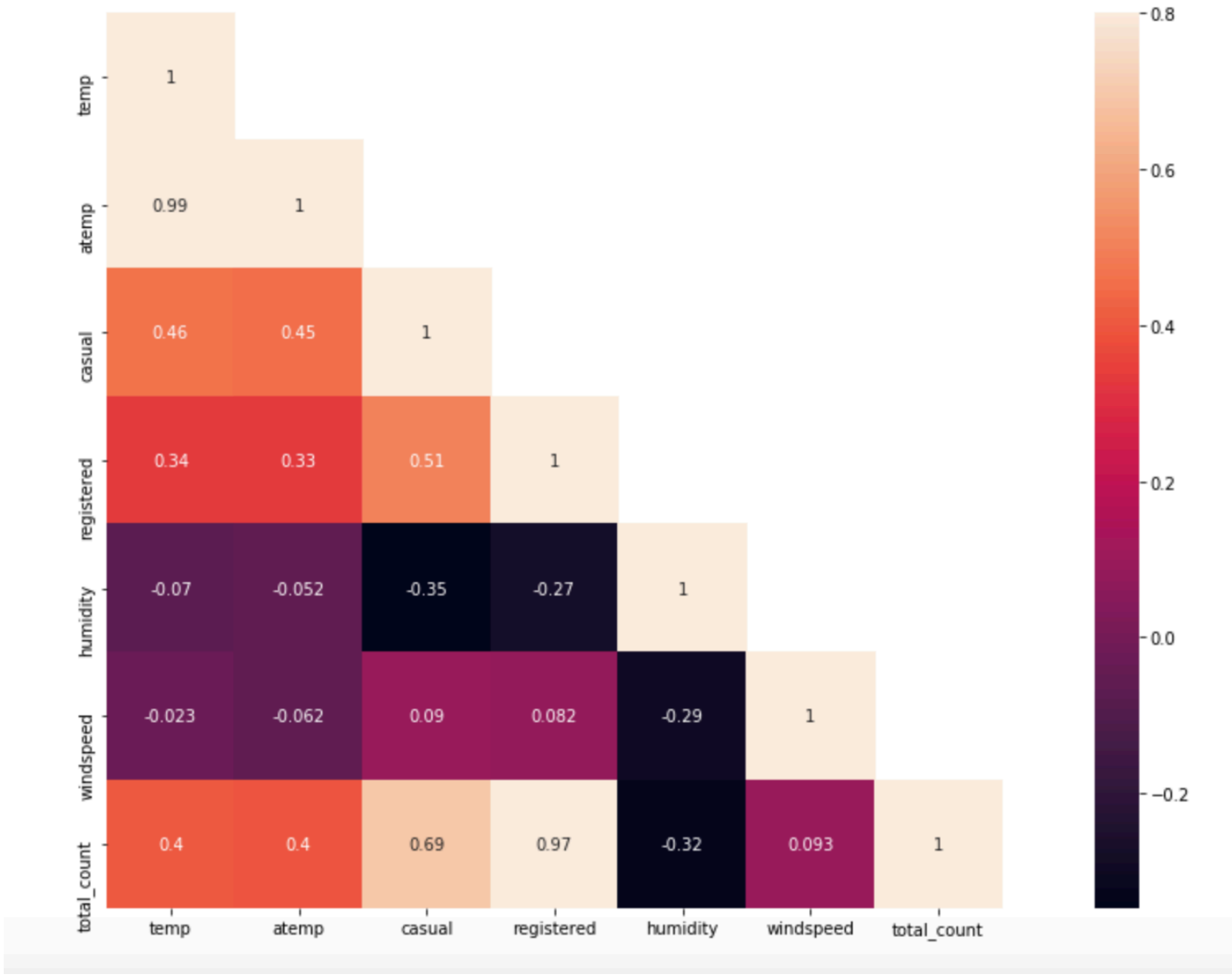
Spring season has got relatively lower count.The dip in median value in boxplot gives evidence for it. The boxplot with "Hour Of The Day" is quiet interesting.The median value are relatively higher at 7AM - 8AM and 5PM - 6PM. It can be attributed to regular school and office users at that time. Most of the outlier points are mainly contributed from "Working Day" than "Non Working Day". It is quiet visible from from figure 4.

6. Correlation plot between "count" and ["temp","atemp","humidity","windspeed"]

One common to understand how a dependent variable is influenced by features (numerical) is to find a correlation matrix between them. Lets plot a correlation plot between "count" and ["temp","atemp","humidity","windspeed"].

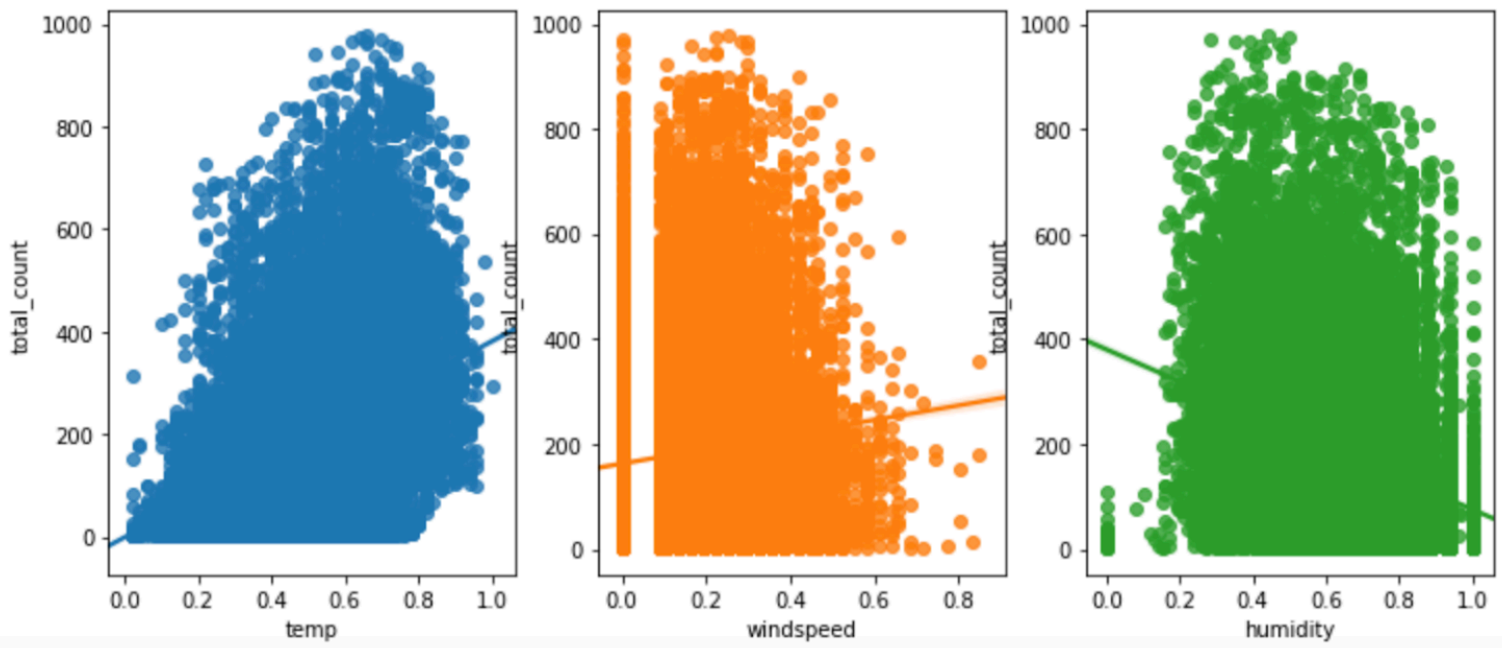
temp and humidity features has got positive and negative correlation with count respectively.Although the correlation between them are not very prominent still the count variable has got little dependency on "temp" and "humidity". windspeed is not gonna be really useful numerical feature and it is visible from it correlation value with "count" "atemp" is variable is not

taken into since "atemp" and "temp" has got strong correlation with each other. During model building any one of the variable has to be dropped since they will exhibit multicollinearity in the data. "Casual" and "Registered" are also not taken into account since they are leakage variables in nature and need to dropped during model building. Regression plot in seaborn is one useful way to depict the relationship between two features. Here we consider "count" vs "temp", "humidity", "windspeed".



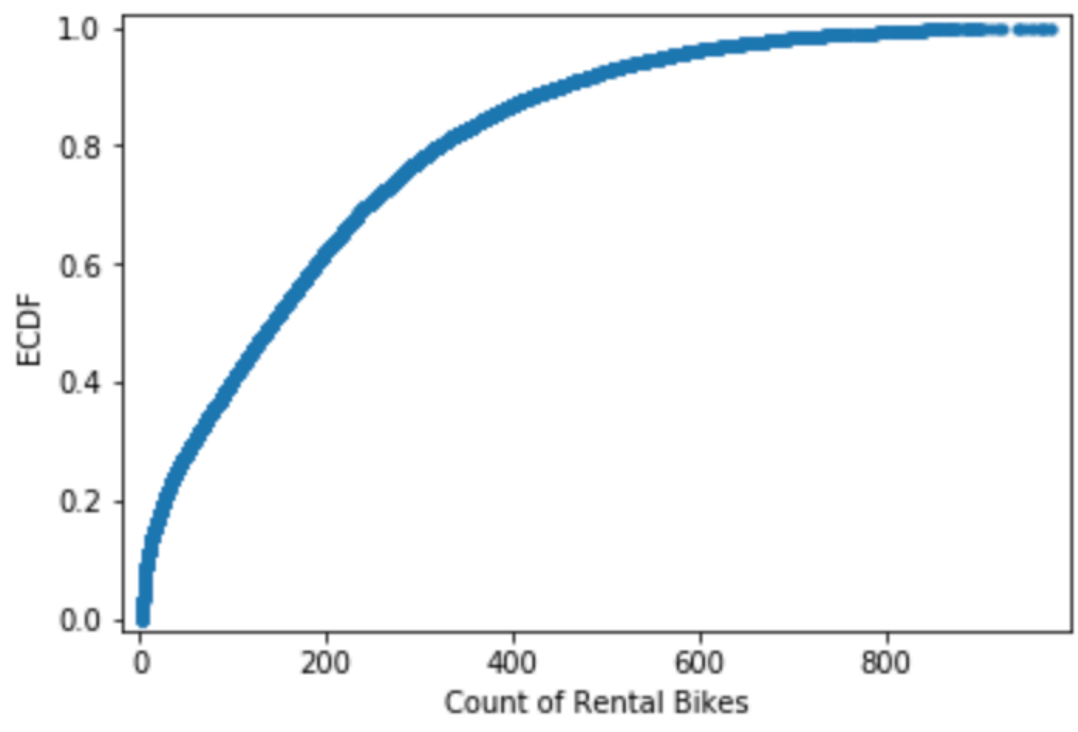

**Correlation between Bike Rental Volume (total_count) and 'registered' user type is the highest. Followed by 'casual' user type. I will explore this dependency of Bike rental volume by User Type in Null Hypothesis under Inferential statistics coming later in EDA. There is moderate collinearity between 'total_count' and 'temp'(temperature)too.**

7. Linear Regression plot between Bike Rental Count vs Temp, humidity, Windspeed




**There is direct positive relation between Bike Rental volume(total_count) vs 'temp' while negative relation with 'windspeed'**

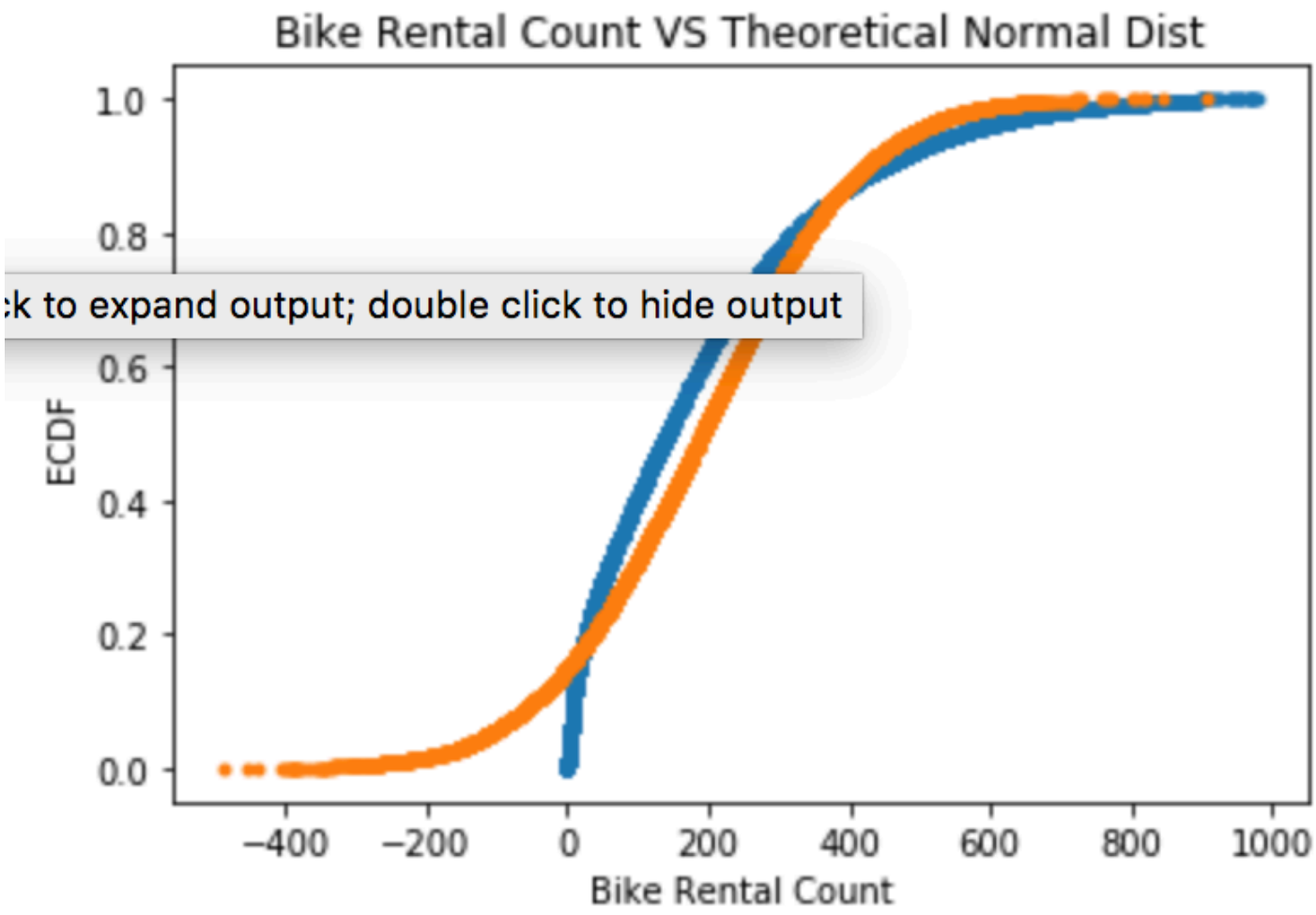
8. Empirical Continuous Distribution ECDF plot for Bike Rental Count




```
[33]: np.percentile(stats['total_count'], [25, 50, 75, 90, 98, 100])
```

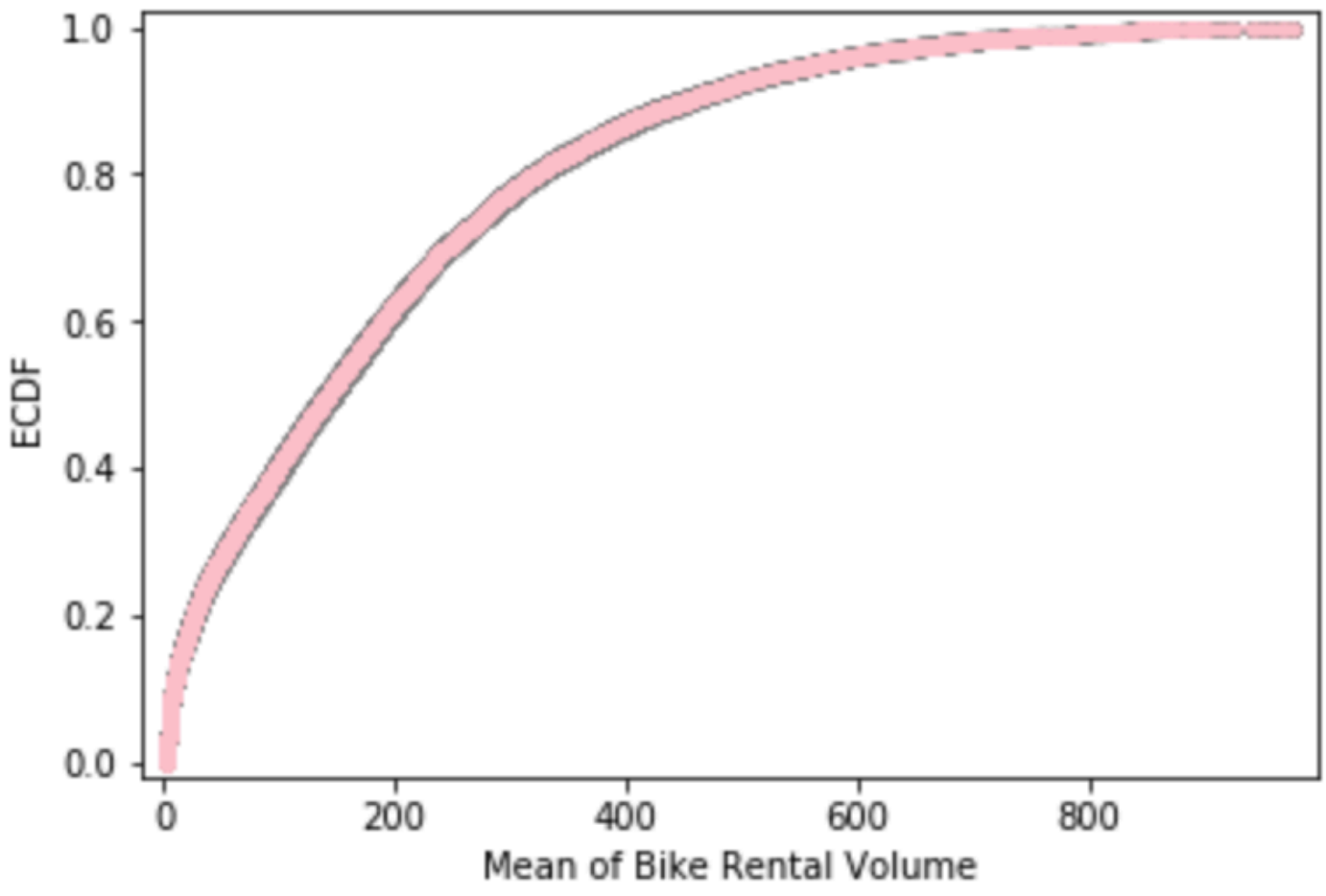
```
[33]: array([ 40. , 142. , 281. , 451.2, 690. , 977. ])
```

9. Checking ECDF Distribution of Bike Rental Count across two years(2011 & 2012) and theoretical samples of data




**Compare the distribution of the data to the theoretical distribution of the data. This is done by comparing the ecdf First define a function for computing the ecdf from a data set. Next use np.random.normal to sample the theoretical normal distribution and overlay the ecdf of both data sets to compare distribution. We see how closely the real data set follows the theoretical normal distribution curve.**

## 10. **Visualizing ECDF using bootstrap samples**



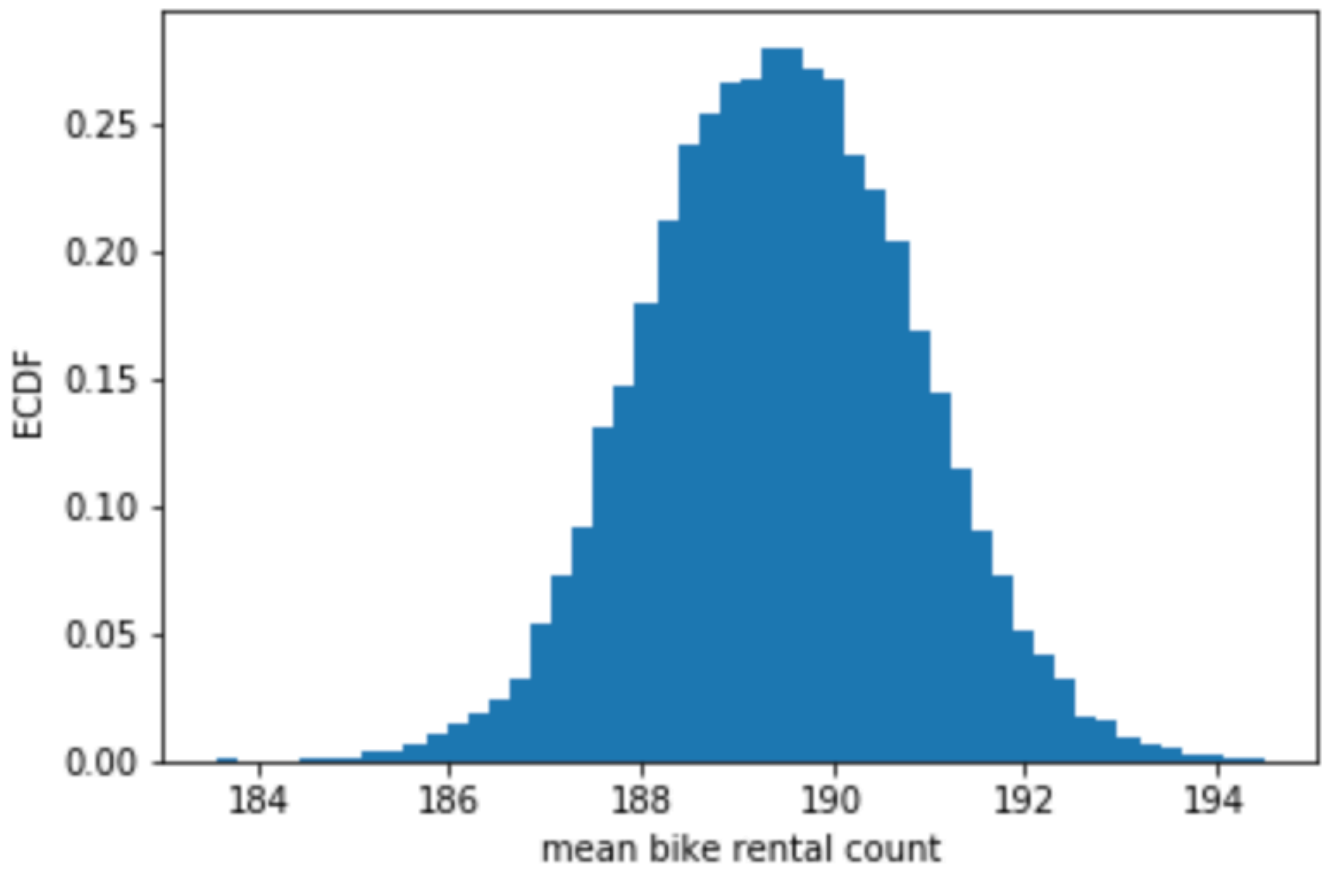

By graphically displaying the bootstrap samples with an ECDF, I see how bootstrap sampling allows probabilistic distribution of data.

# 11. **Confidence Interval**

**Assuming 95% Confidence interval i.e. give the 2.5th and 97.5th percentile of bootstrap replicates is stored as bs_replicates**

**np.percentile(bs_replicates, [2.5, 97.5])**
**O/p: array([186.82940186, 192.19181915])**

**Verifying it with histogram for bootstrap replicates**

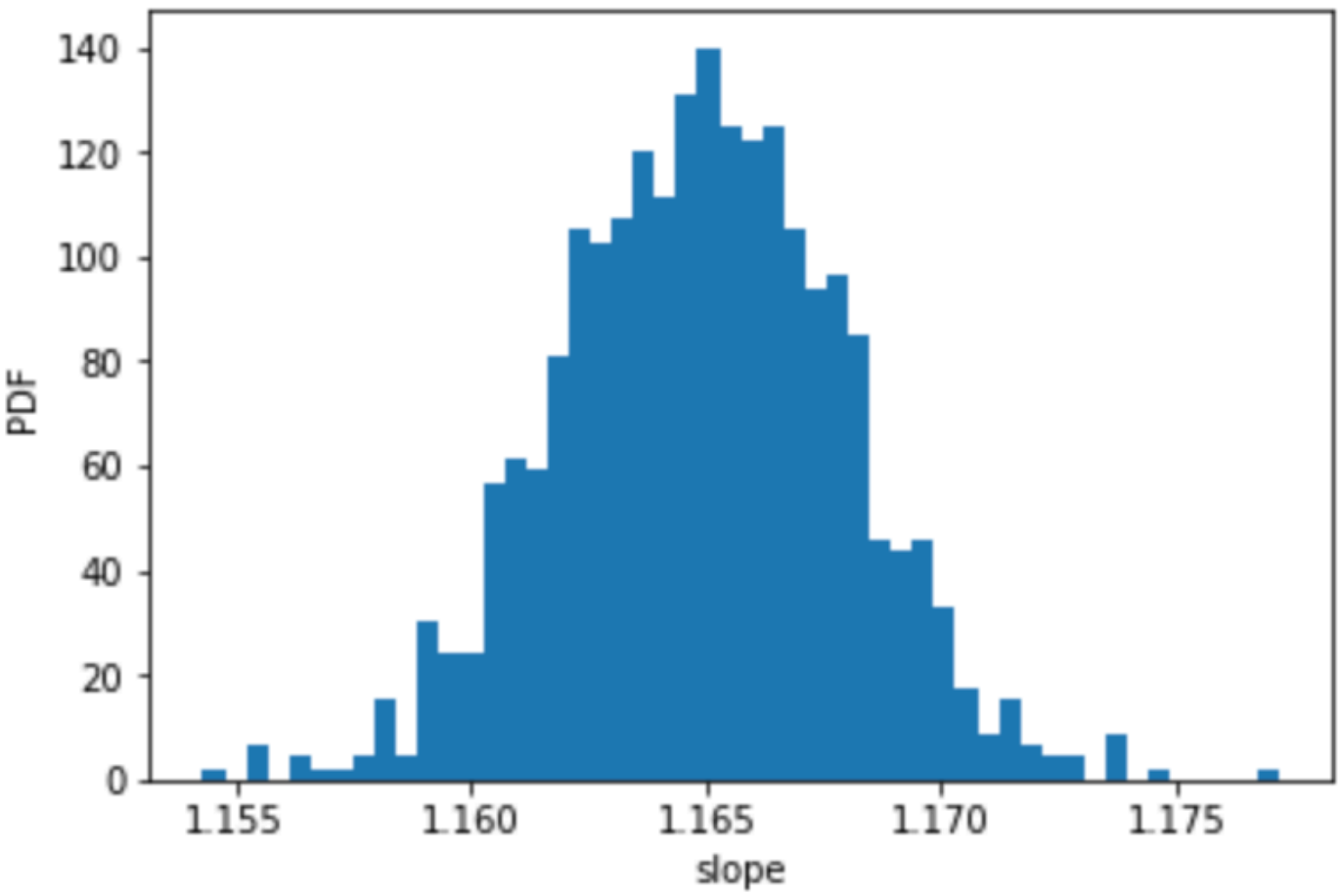


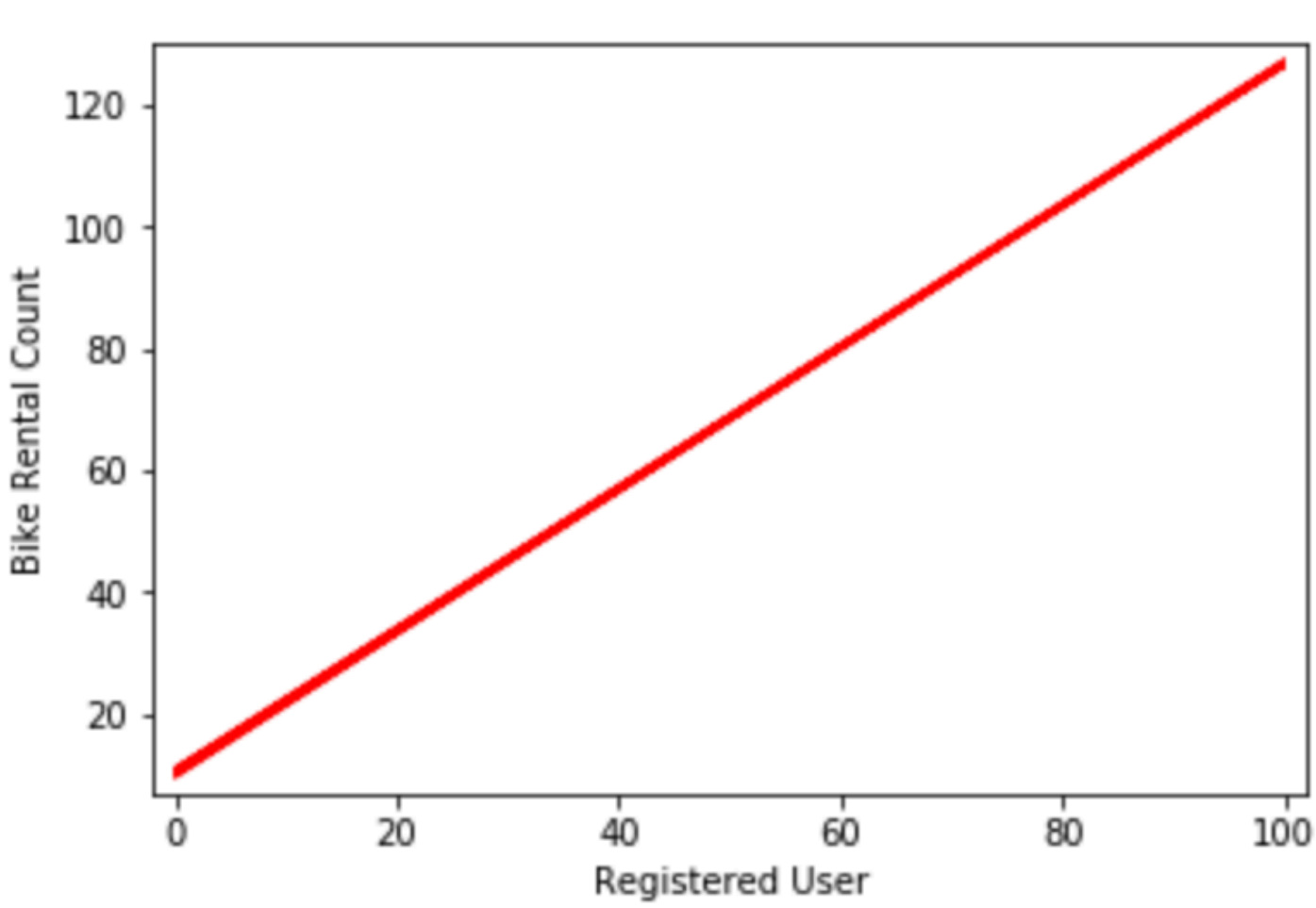

**This is bootstrap estimate of the probability distribution function of the mean Bike Rental Count at the Capital Bikeshare System. Remember, we are estimating the mean Bike Rental Count we would get if the Capital Bikeshare System could repeat all of the measurements from 2011 to 2012 over and over again. This is a probabilistic estimate of the mean. I plot the PDF as a histogram, and I see that it is not Normal as it has slightly longer left tail.**

**In fact, it can be shown theoretically that under not-too-restrictive conditions, the value of the mean will always be Normally distributed. (This does not hold in general, just for the mean and a few other statistics.) The standard deviation of this distribution, called the standard error of the mean, or SEM, is given by the standard deviation of the data divided by the square root of the number of data points. I.e., for a data set. Notice that the SEM we got from the known expression and the bootstrap replicates is the same and the distribution of the bootstrap replicates of the mean is Normal.**

12. **Extending Confidence Interval Concept to Pairs Bootstrap**

**Finding pairs bootstrap for slope & intercept of a linear function between Bike REntal Count and Registered User Type**




## 13. **Hypothesis Testing**

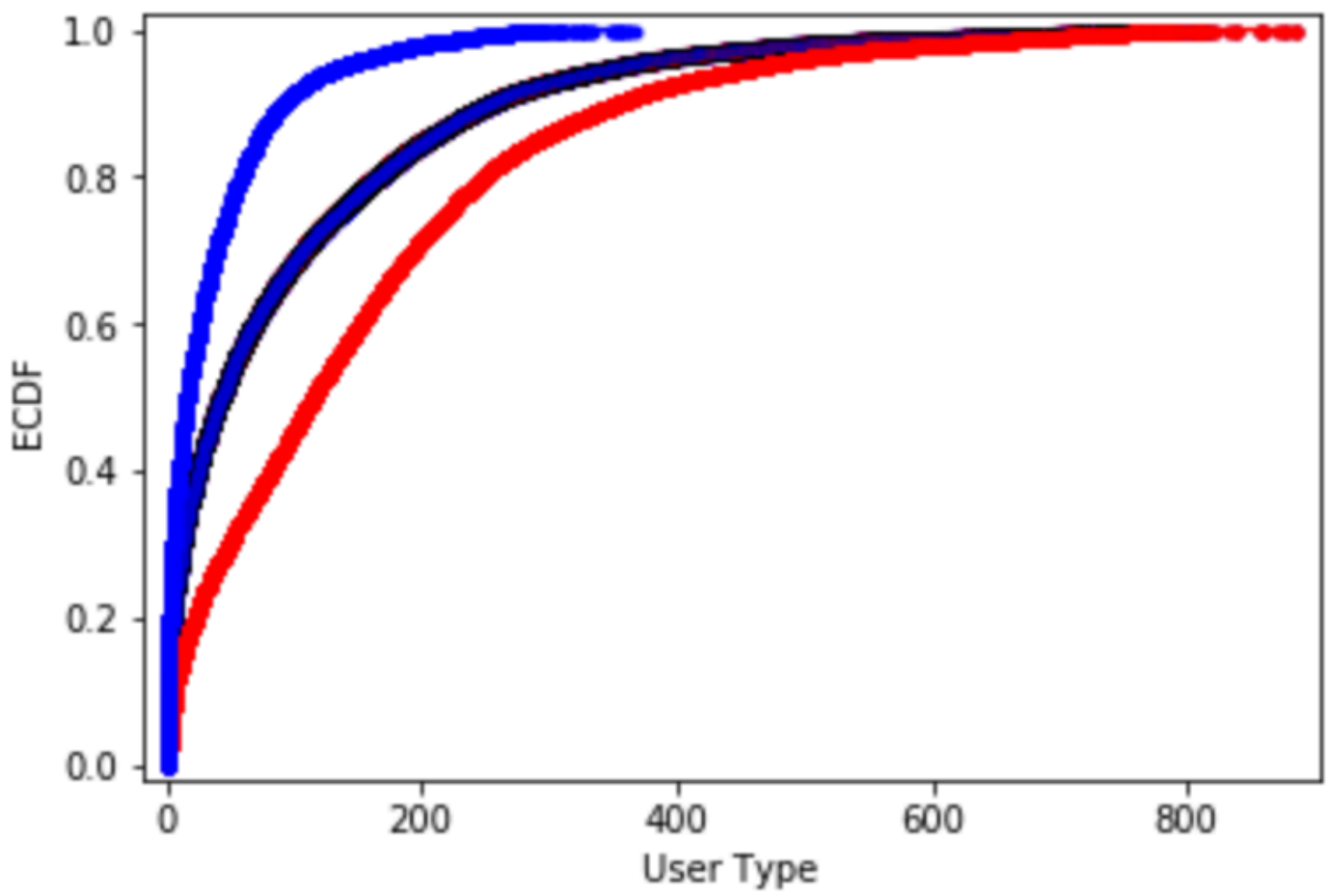
**Null Hypothesis- There is no significant difference between registered and casual user type mean on Bike Rental Count.**

H0:μregistered−μcasual=0

Significance Level: 95% Confidence α=0.05

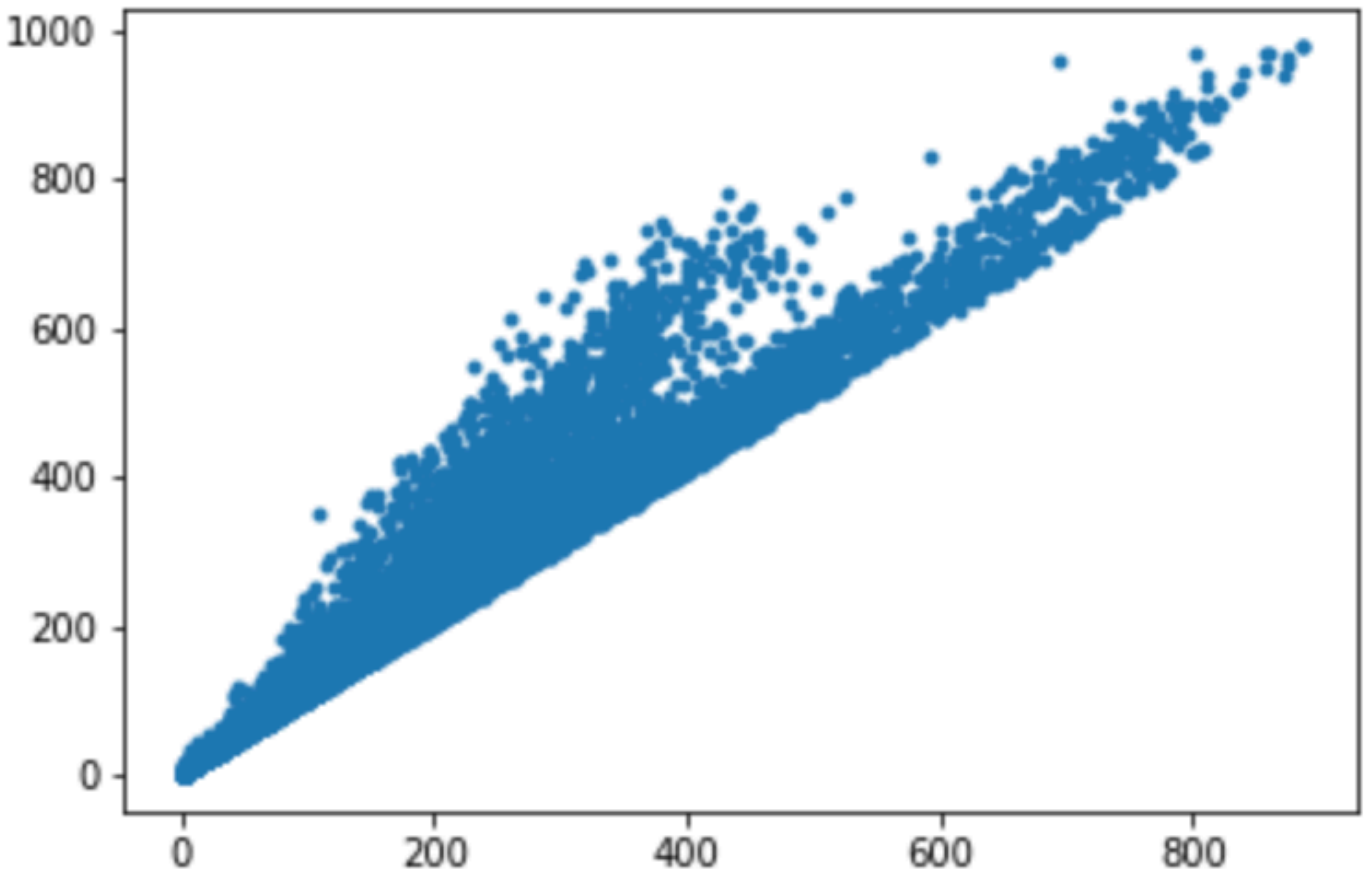
**Alternate Hypothesis -**

**There is significant difference between registered and casual user type mean on Bike Rental Count HA : μregistered−μcasual != 0**



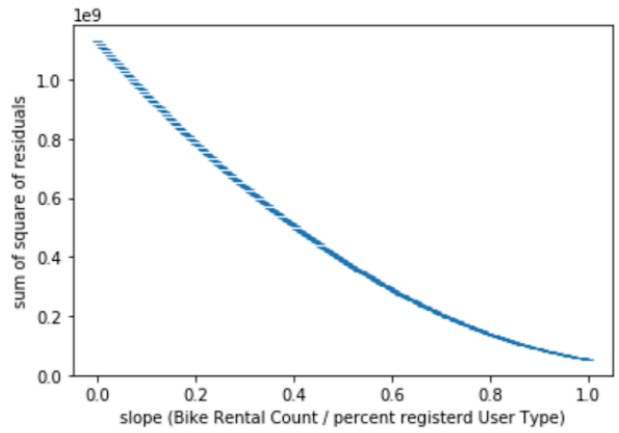

**Permutation samples ECDFs overlap and give a purple haze. Few of the ECDFs from the permutation samples overlap with the observed Registered User type data towards right of the graph & even fewer overlap towards left, suggesting that the hypothesis is not commensurate with the data. Registered & Casual User Type are not identically distributed and do not influence data in similar way. So Null Hypothesis is not correct.**

14. If ECDF is not the right estimated mean another approach to find optimal parameters and residual sum of squares is adopted.
    Took the approach to establish relation between ‘Bike Rental Count’ and ‘Registered’ user type by:
    - Finding Pearson correlation coefficient
    - Scatter Plot




- Optimal parameter (slope, intercept) finding to find best fit linear function

- Comparing above derived slope with slope of minimum RSS & found similar, thus confirming validity of optimal parameters




**minimum on the plot, that is the value of the slope (~1.16) that gives the minimum sum of the square of the residua performing the regression above using np.polyfit()**

**minimum on the plot, that is the value of the slope (~1.16) that gives the minimum sum of the square of the residuals, is the same value I got when performing the regression above using np.polyfit(). Hence Bike Rental Count vs. User Type is a linear continuous function and so ECDF graphs to show Bike Rental Count distribution by User Types is correct and Null hypothesis i.e. User Type Casual & Registered carry similar influence on Bike Rental Count may be rejected.**

**15. ('Registered Sample Size:', 17379, '\nRegistered User Type Mean:', 153.78686920996606) ('\nCasual Sample Size:', 17379, '\nCasual User Type Mean:', 35.67621842453536)**

**There is a difference between the mean of registered and casual User Type in the sample data, but a statistical analysis will help determine if the difference is significant. Null Hypothesis: There is no significant difference between registered and casual user type mean on Bike Rental Count.**

$$H_0: \mu_{registered} - \mu_{casual} = 0$$

**Significance Level: 95% Confidence**

$$\alpha = 0.05$$

**16. Difference of Means by Permutation Samples**

Using np.concatenate, np.random.permutation methods is
('Difference of Means', 118.11065078543069)
('p-value =', 0)

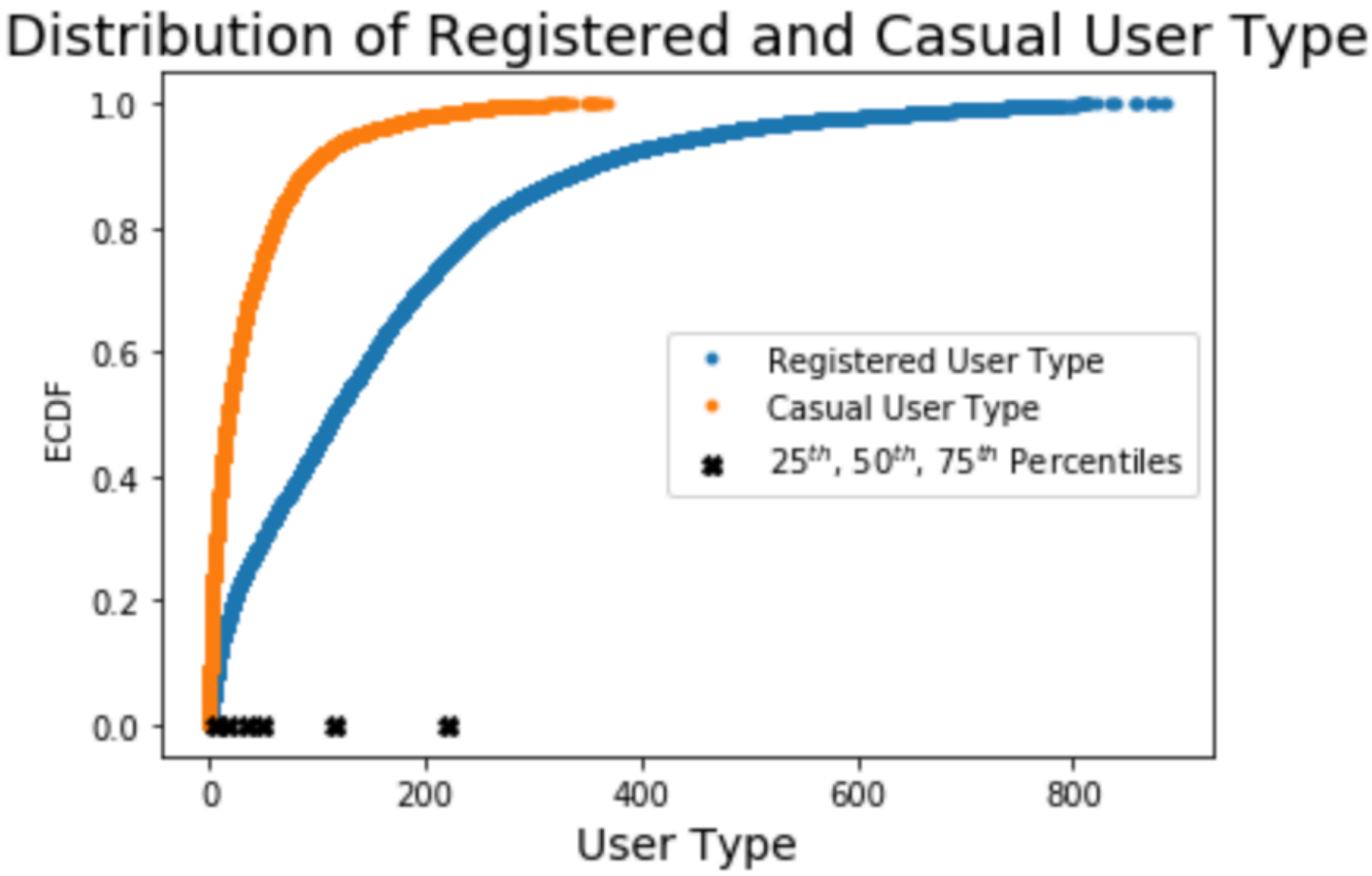
**17. T-Test and P-Value**

Using scipy import stats method performed the T-Test and P-value

('t-statistic:', 97.81332643791566)
('p-value:', 0.0)

This confirms Null Hypothesis may be rejected as means of Bike Rental Count by User Type is not same.

**18. ECDF Plots for Bike Rental Count Distribution by Registered and casual User Types**




**Hence by above plot, we may reject the Null hypothesis since there is significant difference between Bike Rental Count distribution of Registered and Casual User Type.**

Based on above EDA & Inferential analysis, I next propose to do in-depth analysis on the given dataset using one of the Supervised Machine Learning algorithms of type Regression since output datasets are provided and I can use this to predict future outcomes of target variable. Additionally this is Regression problem as dependent variable ‘Bike Rental Count’ is continuous values and can be used to predict the output value using training data.

Here is the link to iPython Notebook:
https://github.com/rashi-n/Machine-Learning-Projects/blob/master/Capstone%20Projects/CS%20I%20-%20EDA%20%26%20Inferential%20Statistics.ipynb