## Exploratory Data Analysis - Inferential Staistics

After you've obtained, cleaned, and wrangled your dataset into a form that's ready for analysis, you'll perform preliminary exploration. This exploratory data analysis (EDA) uses a combination of inferential statistics and data visualization to find interesting trends and identify significant features in the dataset. For example:

- Are there significant variables that help explain the answer to your project question?
- Are there strong correlations between pairs of independent variables or between an

independent and a dependent variable?

**Learning Objective**

- Identify variables tin the data to answer to a project question.
- Identify strong correlations between pairs of independent variables or between an independent and a dependent variable.
- Practice identifying the most appropriate tests to use to analyse relationships between variables.

| Criteria | Meets Expectations |
|---|---|
| Completion | ❏ A 1-2 page report on the steps and findings from inferential statistical analysis, uploaded to GitHub. |
| Process and understanding | ❏ The submission shows that the student applied applied inferential statistics techniques to the data for their capstone project. |
| Presentation | ❏ The submission is complete and uploaded in full. |

Identify variables tin the data to answer to a project question.
Project Question: Predict the Bike rental usage by the features given in the dataset from Capital Bike Sharing System
Dataset given has 17 features with 17K +rows

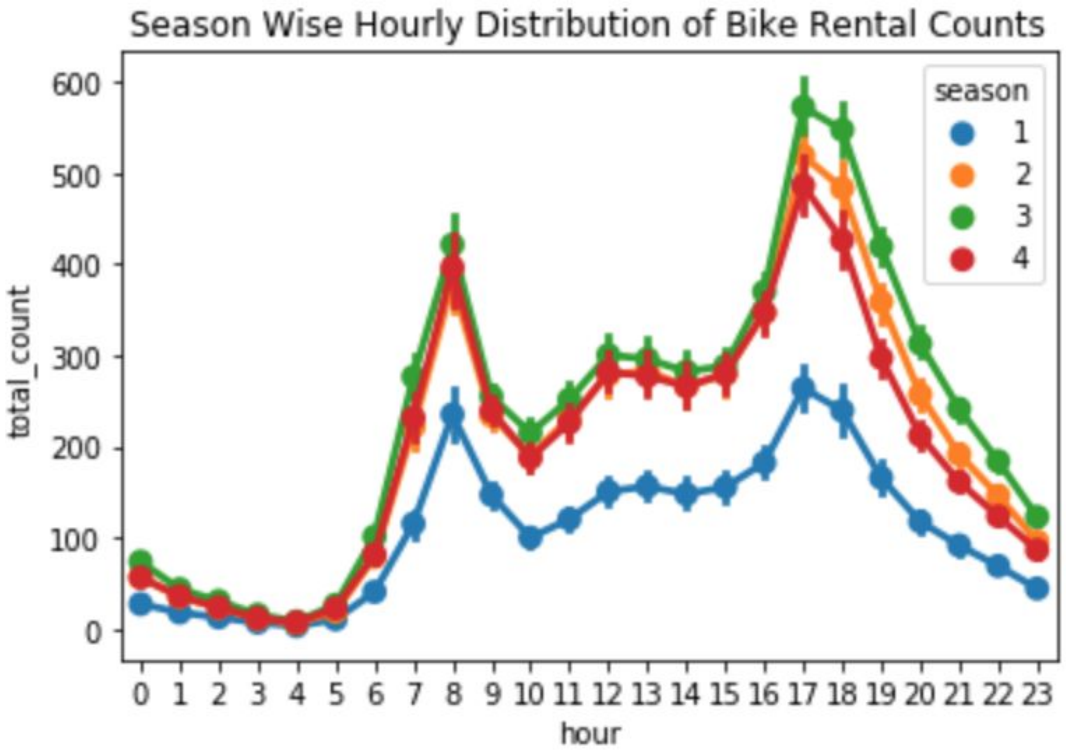
| instant | dteday | season | yr | mnth | hr | holiday | weekday | workingday | weathersit | temp | atemp | hum | windspeed | casual | registered | cnt |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1/1/11 | 1 | 0 | 1 | 0 | 0 | 6 | 0 | 1 | 0.24 | 0.2879 | 0.81 | 0.0 | 3 | 13 | 16 |
| 2 | 1/1/11 | 1 | 0 | 1 | 1 | 0 | 6 | 0 | 1 | 0.22 | 0.2727 | 0.80 | 0.0 | 8 | 32 | 40 |
| 3 | 1/1/11 | 1 | 0 | 1 | 2 | 0 | 6 | 0 | 1 | 0.22 | 0.2727 | 0.80 | 0.0 | 5 | 27 | 32 |

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

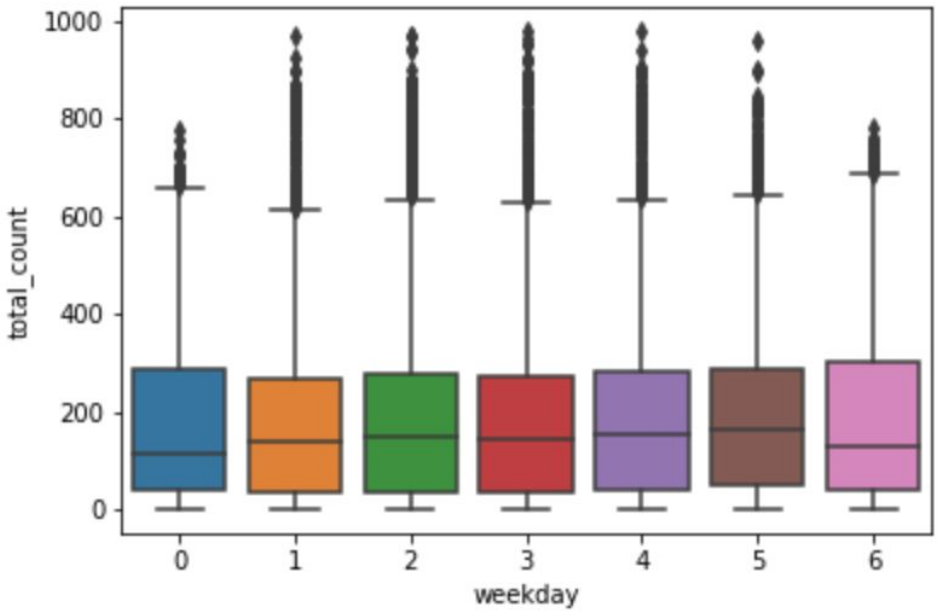
- Season
- Month
- Temperature (temp)
- Humidity (hum)
- Windspeed
- Casual
- Registered
- Hour (hr)

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

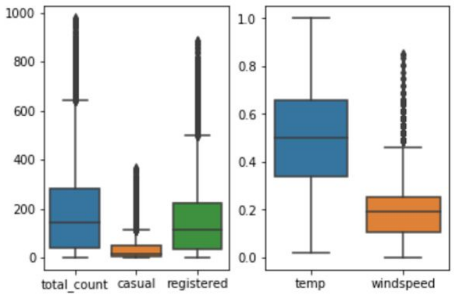
1. Line Chart between Bike Rental Count vs. Hour across Seasons




2. Box Plot between Bike Rental Count vs Weekday

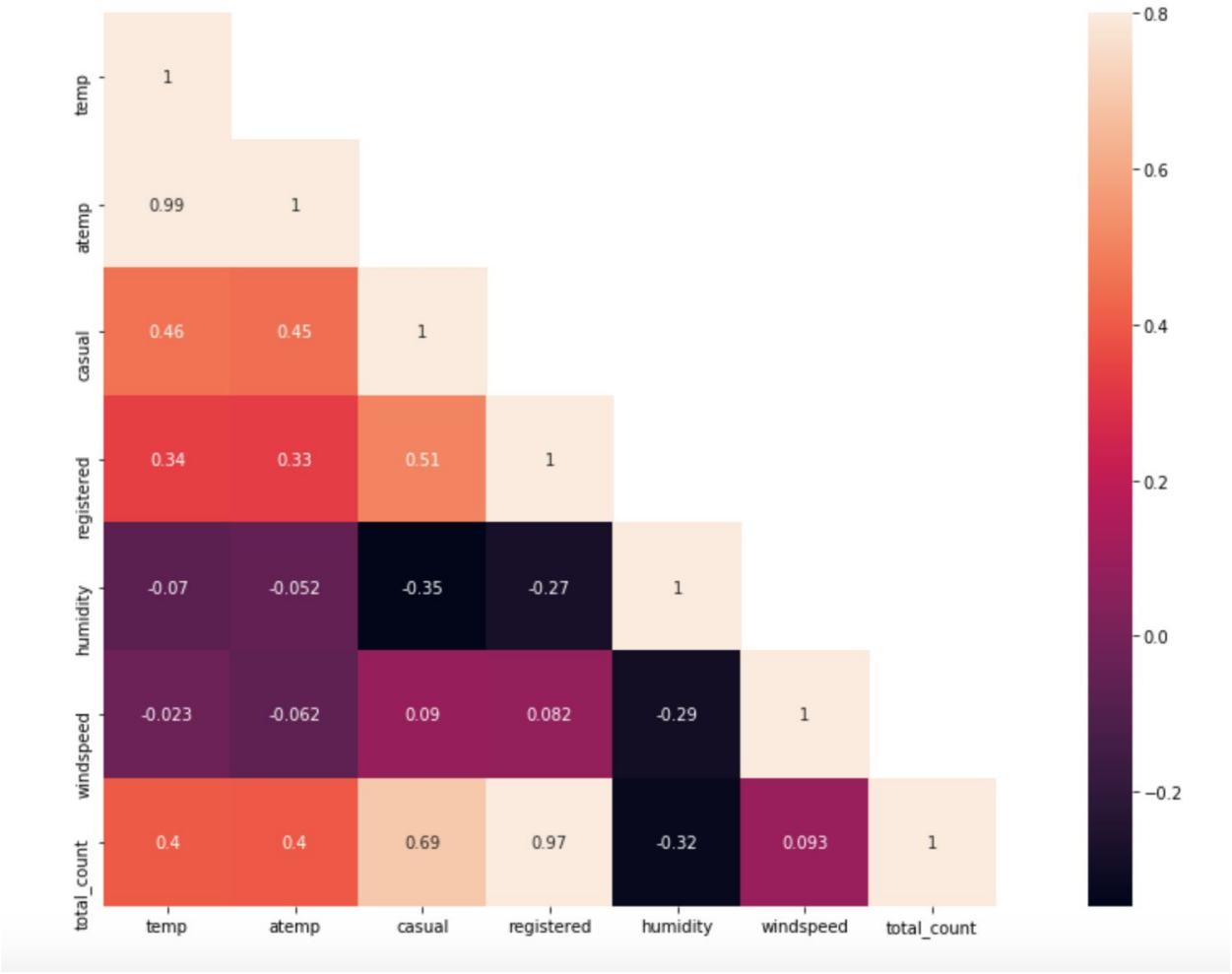



3. Violin Plot between Bike Rental Count vs Seasons
4. Box Plot between Bike Rental Count vs Casual, Registered Users

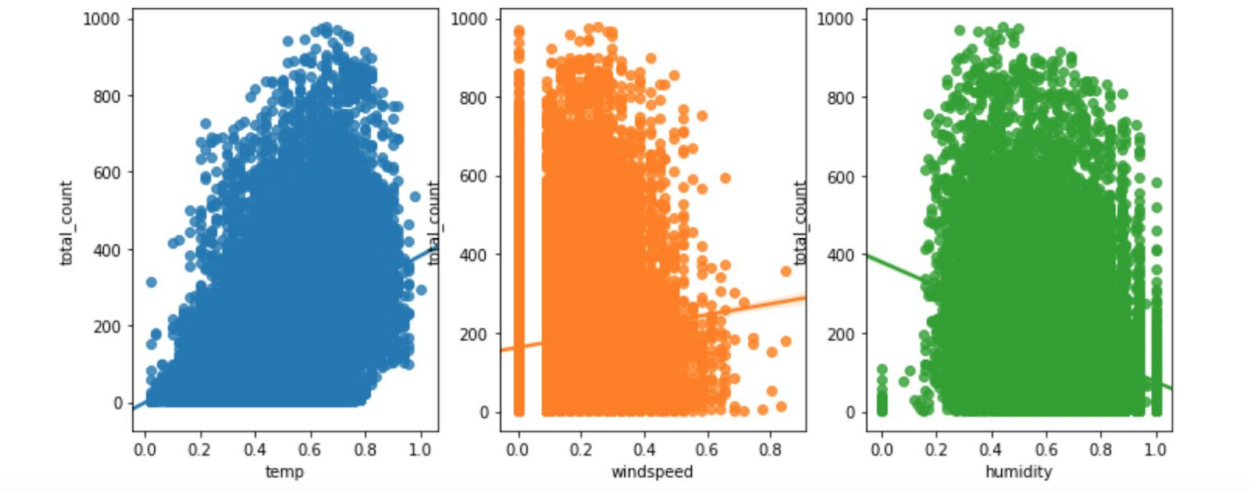



**The total, casual & registered type users show sizeable number of outlier values, however casual show lower numbers though. For weather attributes of temperature and wind speed, we see outliers only in the case of windspeed.**

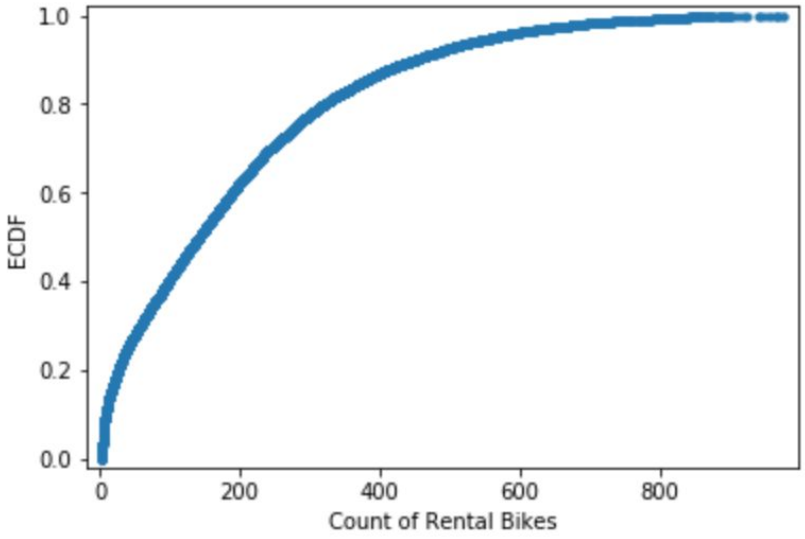
5. Correlation plot between "count" and ["temp","atemp","humidity","windspeed"]




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



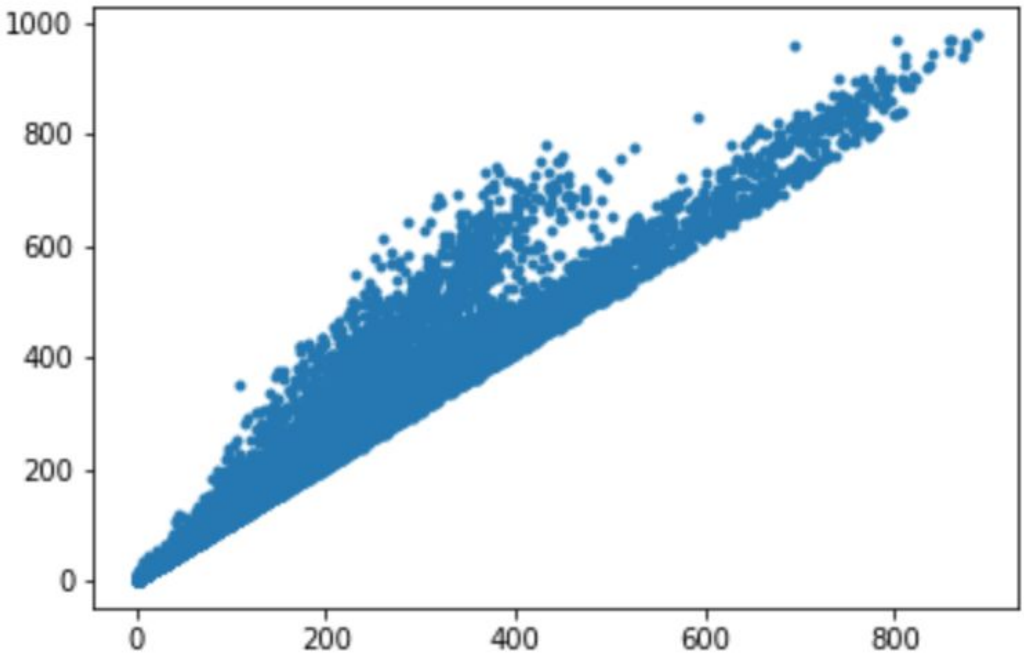

7. Empirical Continuous Distribution ECDF plot for Bike Rental Count




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

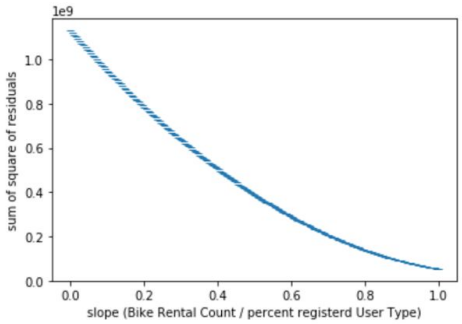
```
[33]: array([ 40. , 142. , 281. , 451.2, 690. , 977. ])
```

8. 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



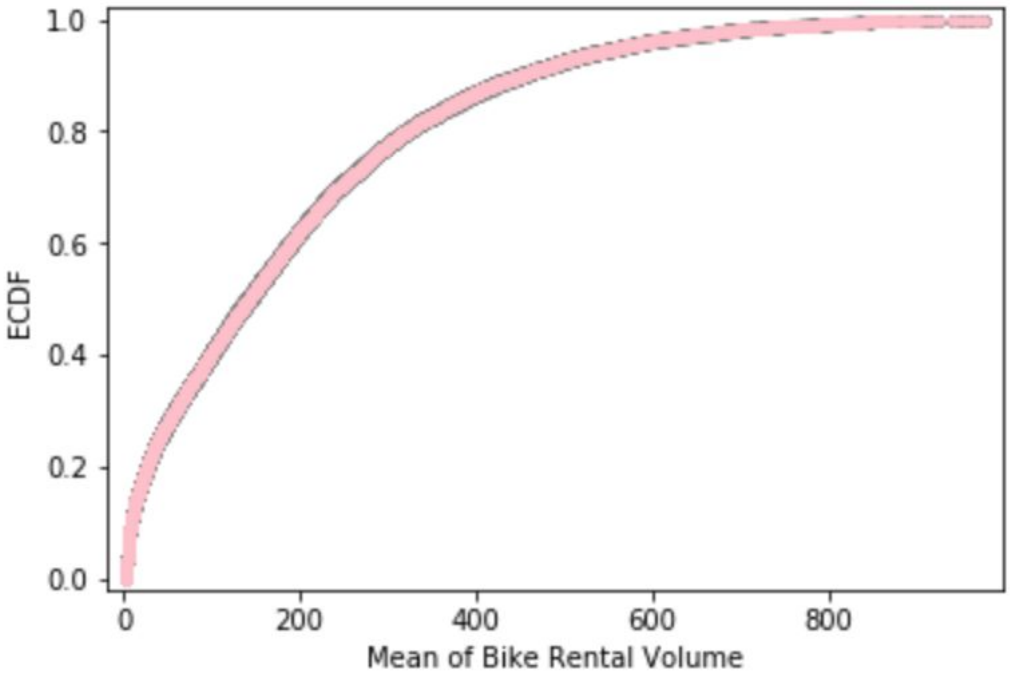

- Optimal parameter (slope, intercept) finding to find best fit linear function

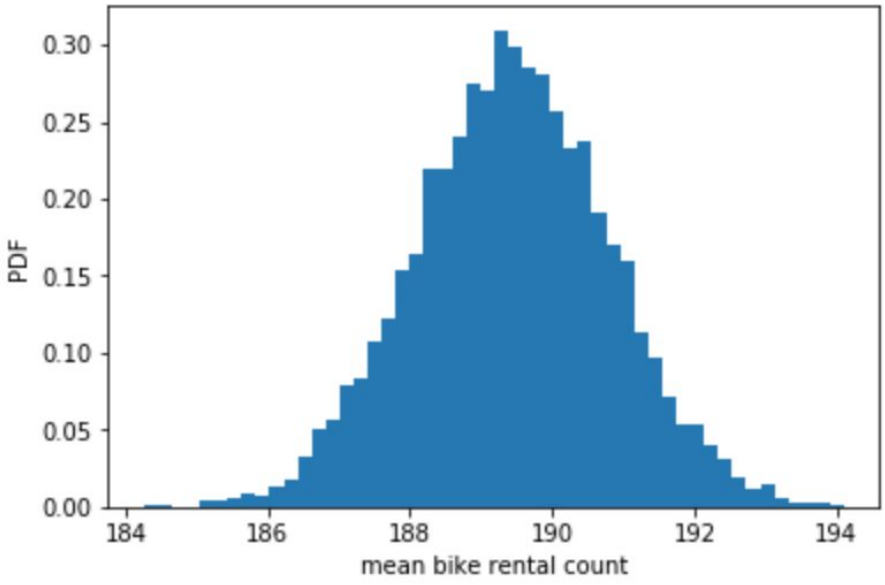
- Comparing above derived slope with slope of minimum RSS & found similar, thus confirming validity of optimal parameters




**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()**

9. Mean of Bike Rental Counts was resampled using Bootstrap replicate function to plot ECDF and histogram as shown below with assumption confidence interval 95% proving true:


1.0
0.8
0.6
0.4
0.2
0.0
ECDF
0
200
400
600
800
Mean of Bike Rental Volume


0.30
0.25
0.20
0.15
0.10
0.05
0.00
PDF
184
186
188
190
192
194
mean bike rental count