MALL FOOTFALL PREDICTION

CAPSTONE PROJECT

Submitted in partial fulfilment of the requirements of the

Post Graduate Certification Program in

Artificial Intelligence and Machine Learning

By

Ajay Singaraju

2019AIML060

Sudhakar Yeelarthi

2019AIML070

Aji ED

2019AIML090

Sagar Sangamnerkar

2019AIML104

Rakesh Rudra

2019AIML152

Satish Kadagala

2019AIML176

Under the supervision of

Mr. Satyaki Dasgupta

Project work carried out at

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE

Pilani (Rajasthan) INDIA

September, 2020

PCAM ZC321 CAPSTONE PROJECT

MALL FOOTFALL PREDICTION

Submitted in partial fulfilment of the requirements of the

PGP - Artificial Intelligence and Machine Learning

By

Ajay Singaraju

2019AIML060

Sudhakar Yeelarthi

2019AIML070

Aji ED

2019AIML090

Sagar Sangamnerkar

2019AIML104

Rakesh Rudra

2019AIML152

Satish Kadagala

2019AIML176

Under the supervision of

Mr. Satyaki Dasgupta

Project work carried out at

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE

PILANI (RAJASTHAN)

(Month, Year)

**ACKNOWLEDGEMENTS**

**BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI**

**CERTIFICATE**

This is to certify that the Capstone Project entitled\_\_\_\_\_\_\_Mall Footfall Prediction \_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

and submitted by Mr./Ms.\_\_\_First name last name \_\_\_\_\_\_\_ ID No. 2019AIMLnnn\_\_\_\_\_\_

in partial fulfilment of the requirements of PCAM ZC321 Capstone Project, embodies the work

done by him/her under my supervision.

Place : Hyderabad Signature of the Mentor

Date : DD/MMM/2020 Name : Mr. Satyaki Dasgupta

**BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI** **SECOND SEMESTER 2019-20**

**PCAM ZC321 CAPSTONE PROJECT**

|  |  |
| --- | --- |
|  |  |
| Project Title | : MALL FOOTFALL PREDICTION |
| Name of Mentor | : Mr. SATYAKI DASGUPTA |
| Name of Student | : First name Last Name |
| ID No. of Student | : 2019AIMLnnn |
| **Abstract**  Footfall — also known as People Counting, Shopper Counting or traffic is the measurement of the number of people entering a shop or shopping mall.  Why should anyone count and predict the footfall?   * Calculating footfall is a key metric or measurement for retail businesses wanting to increase sales and drive profits up. By counting footfall, retailers can identify at any time of the day, the purchasing opportunities they have in-store and crucially, by collating that with sales data, identify the ones missed out. * Knowing the hourly, daily and weekly retail footfall can give retailers that extra edge to target peak hours, convert browsers into customers and ultimately optimize store performance. * Measuring footfall lets businesses and retailers can reap multitude of benefits like valuable analytics and key performance indicators, improve the visitor experience, understand customer behaviour, optimise operations, popular store areas etc   How can we predict the footfall?   * Collect the past footfall data and other data related to the impacting factors like holidays, weather conditions each day etc * Apply Pre-Processing and feature engineering techniques   + Understanding Data, fields involved, Visualize and Impute Missing Values   + Data Merger from Various sources - In Count Data, Holidays, Weather   + Data Normalization and Scaling, Data Reshape, Data Encoding   + Outlier Analysis and PCA   + Derive Additional Data - Lags * Decompose time series into trend and seasonal component, Detrend Time series * Apply Augmented Dickey-Fuller Test to Evaluate the stationarity * Visualization the data in various plots - Kernel Density, SNS/ggplot/panda, strip, Swarm, Box, Violen, Boxen, Point, Bar, Count, Joinplot, Displot, Pairplot, rugplot, Line plot, Numerical scatterplot, Multiple relationships with facets * Check Auto Correlation * Build and apply appropriate machine learning models to predict the data | |
| **Key Words:** |  |

(​***Note:*** *The Abstract should briefly describe the work done with respect to the goals, in about 500 words. The Abstract in the above format should be included in the bound Report, after the Acknowledgements and immediately before the Table of Contents.).*

# List of Symbols and Abbreviations used

|  |  |
| --- | --- |
| Symbol | Meaning/Usage |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

Table :List of Symbols

|  |  |
| --- | --- |
| Abbreviation | Meaning/Usage |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

Table :List of Abbreviations

# List of Tables

[Table 1:List of Symbols 6](#_Toc53864339)

[Table 2:List of Abbreviations 7](#_Toc53864340)

[Table 3: Completion Checklist 32](#_Toc53864341)

# List of Figures

[Figure 1: Process flow 11](#_Toc53864347)

[Figure 2: Classes and functions 12](#_Toc53864348)

[Figure 3:Execution Plan 15](#_Toc53864349)

[Figure 4 Pre-processing and Feature Engineering flow 16](#_Toc53864350)

[Figure 5: Original data snapshot 16](#_Toc53864351)

[Figure 6: Reshaped Data snapshot 17](#_Toc53864352)

[Figure 7: Aggregated data snapshot 17](#_Toc53864353)

[Figure 8: Merged data snapshot 17](#_Toc53864354)

Table of Contents

[List of Symbols and Abbreviations used 10](#_Toc52657139)

[List of Tables 11](#_Toc52657140)

[List of Figures 12](#_Toc52657141)

[1. Problem statement (what is the problem being addressed) 14](#_Toc52657142)

[2. Objective of the project 14](#_Toc52657143)

[3. Background of previous work done in the chosen area (Literature Review) 14](#_Toc52657144)

[4. Machine Leaning process flow (Consolidated Approach / Solution Architecture) 14](#_Toc52657145)

[5. Resources needed for the project, including people, hardware, software, etc. 14](#_Toc52657146)

[6. Potential data challenges & risks in doing the project 15](#_Toc52657147)

[7. Detailed Plan of Work 16](#_Toc52657148)

[8. Pre-Processing Steps 17](#_Toc52657149)

[Data processing/Feature pre-processing 17](#_Toc52657150)

[Outlier detection 17](#_Toc52657151)

[Visualization for Summarization 17](#_Toc52657152)

[9. Machine Learning Modelling & Techniques Applied 17](#_Toc52657153)

[10. Code & Screenshots (Full screenshot without cropping the system date-time) 17](#_Toc52657154)

[11. Interpretation 17](#_Toc52657155)

[Justification of Measures / Metrics used 17](#_Toc52657156)

[Project output in terms of above measures/metrics 17](#_Toc52657157)

[Visualization Plots 17](#_Toc52657158)

[12. Future Work & Extension or Scope of improvements 17](#_Toc52657159)

[13. Conclusions / Recommendations 17](#_Toc52657160)

[14. Directions for future work 17](#_Toc52657161)

[15. Bibliography / References 17](#_Toc52657162)

[16. Appendices 18](#_Toc52657163)

[17. List of Publications/Conference Presentations, if any. 18](#_Toc52657164)

[18. Duly Completed Checklist 19](#_Toc52657165)

1. Problem statement (what is the problem being addressed)

Footfall is the measurement of the number of people entering a shop or shopping mall. Job at hand is to predict

* Peak time in each day of week
* Footfall in peak time of each day of week

1. Objective of the project

The primary objective of this project is to predict the possible in flow at any given future point in time, making use of the data available – 1) The in count at each hour at various locations (Also referred to as Gates) of the mall. 2) The supplementary data like weather and holiday data.

The objective is also to thoroughly understand the data in hand and derive business insights that would help predict the required information in an efficient manner.

1. Background of previous work done in the chosen area (Literature Review)

aaaaaaaaaaaaaaa

1. Machine Leaning process flow (Consolidated Approach / Solution Architecture)

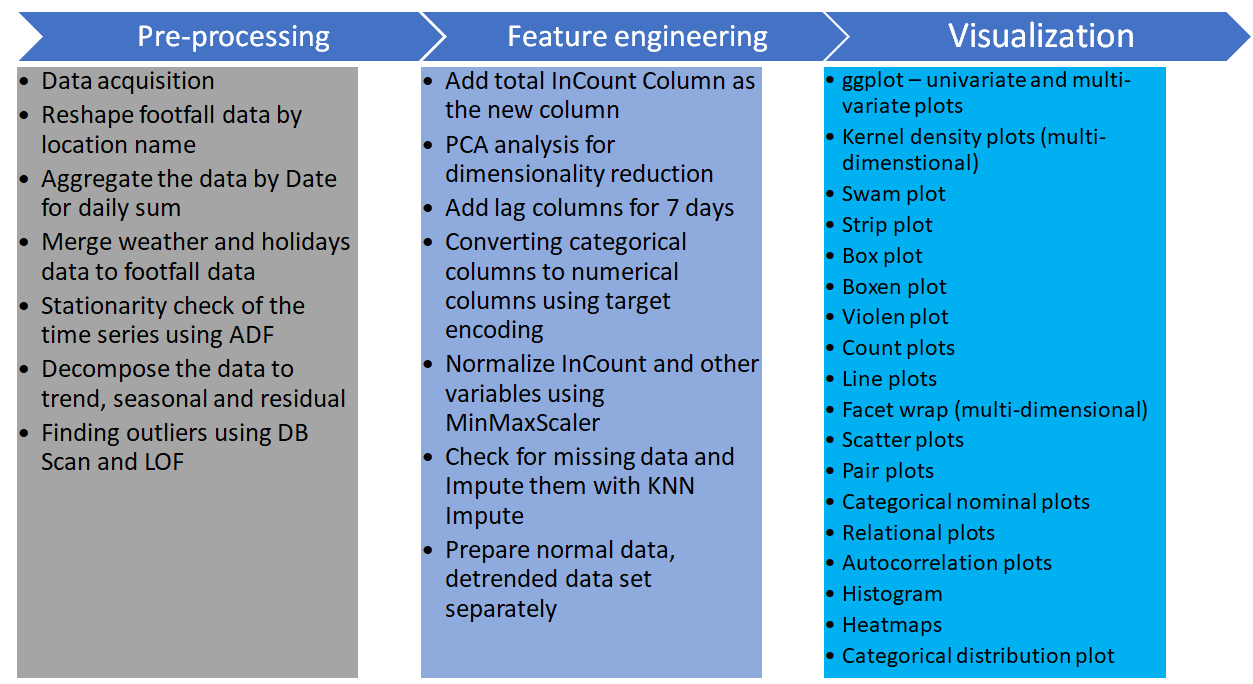


Figure : Process flow

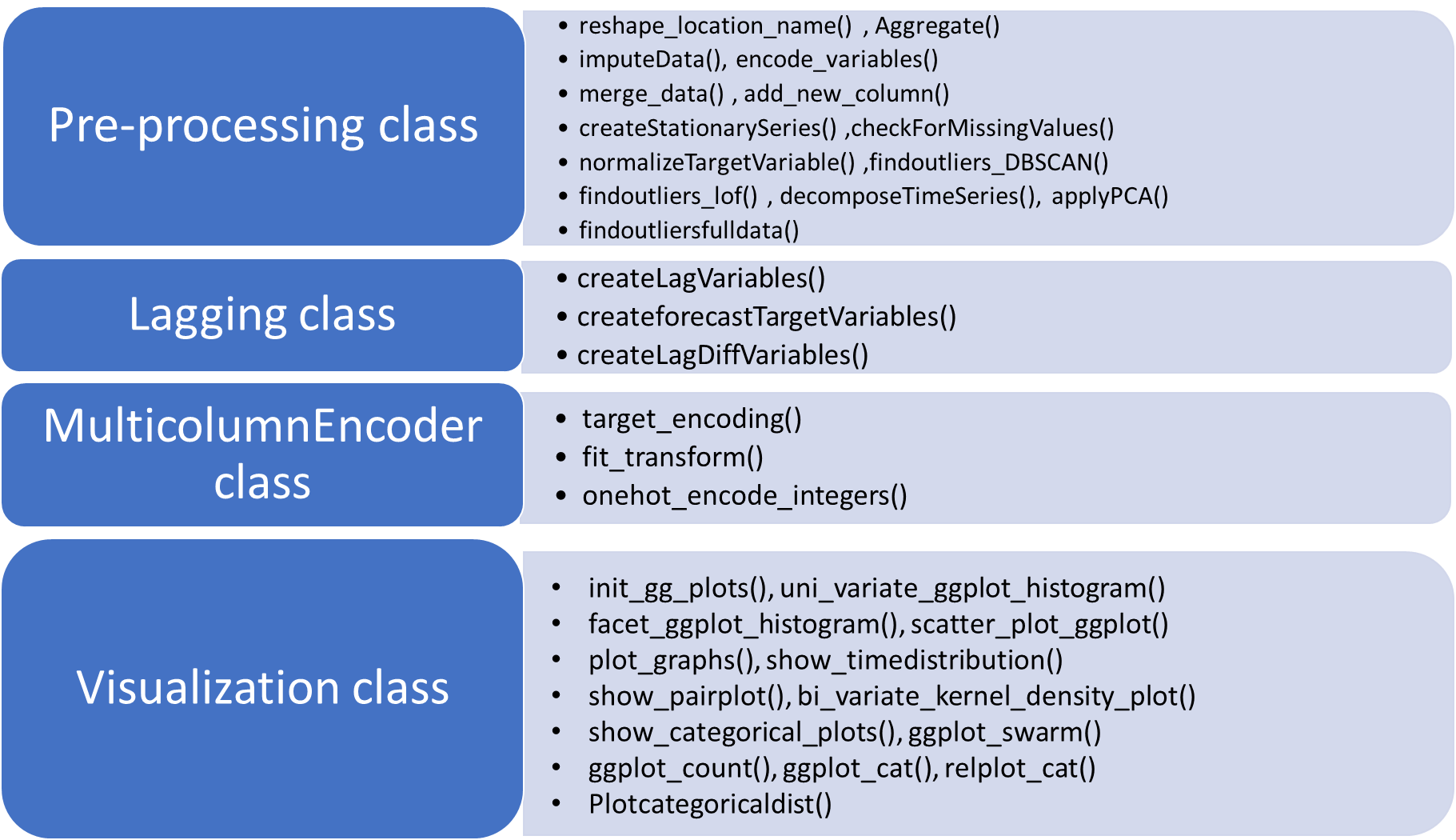


Figure : Classes and functions

1. Resources needed for the project, including people, hardware, software, etc.

The resources needed for the project can be classified in to four broad categories – Information, Technology, Process and People

* Information
  + The past data that can be used as an input for various machine learning models, after thorough pre-processing, inference and visualization of the underlying information
  + Business knowledge of what all could impact the in count of people in to the Mall at various times and the degree of impact.
  + Knowledge of various tools and machine learning techniques that can be applied to be able to understand the data available better, zero down on the right machine learning models that can be applied and how to tune them to be able to predict the required parameters with the at most accuracy
* Technology
  + Data storage at a common location accessible to all the team members
  + Python run time environments – Jupyter/Spyder/Google Colab
* Process
  + Business Understanding
  + Project Planning and Execution
  + Resource management
    - Version Management of the code
    - Work distribution and collaboration
  + Data Collection
  + Data Pre-processing
    - Data Merger from various sources
    - Data Normalization
    - Outlier Analysis
    - Missing Values Imputation
  + Data Inference
    - Uni Variate Analysis
    - Multi Variate Analysis
    - Trend Analysis
    - Data Correlation Analysis
* People
  + A guide to direct the team in right direction so that learnt knowledge can be applied in an efficient
  + 5-6 People with good understanding of Machine learning concepts and Python hands-on experience – preferably with Object Oriented Programming experience

1. Potential data challenges & risks in doing the project

* Data Challenges
  + Possible missing critical data or Noise Ex: An Unplanned Special Event in the Mall or a Local holiday (Un Common) not captured there by resulting in outliers/noisy data
* Risks

1. Detailed Plan of Work



Figure :Execution Plan

1. Pre-Processing and Feature engineering

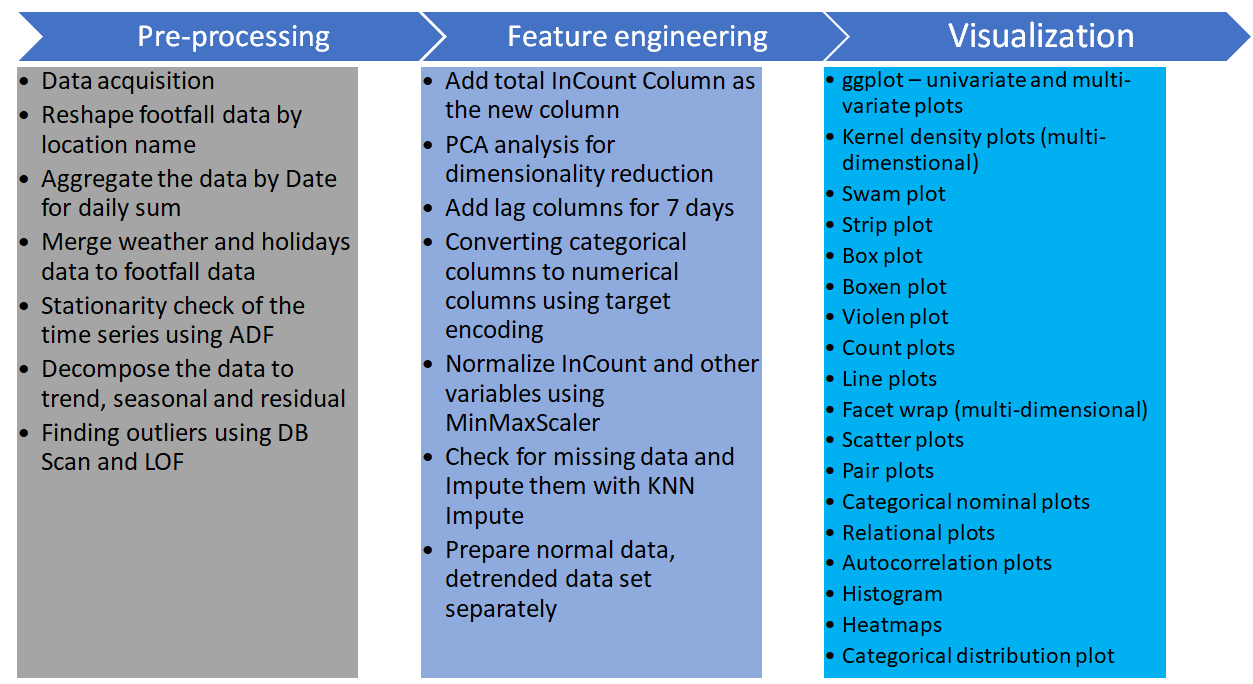


Figure Pre-processing and Feature Engineering flow

# Data Acquisition

The given csv file datasets are read into multiple Panda dataframes.

*df\_foot\_fall =pd.read\_csv (‘footfall.csv')*

*df\_weather\_data= pd.read\_csv('overall\_weather.csv')*

*df\_easter\_sunday\_holiday\_data= pd.read\_csv('Easter-Sundays.csv')*

*df\_university\_long\_holidays\_data= pd.read\_csv('uni\_hols\_long.csv')*

*df\_UK\_bank\_holidays\_data= pd.read\_csv('UKbankholidays.csv')*

*df\_school\_long\_holidays\_data= pd.read\_csv('school\_hols\_long.csv')*

* 1. Reshape footfall data by location name

Original foot fall data:

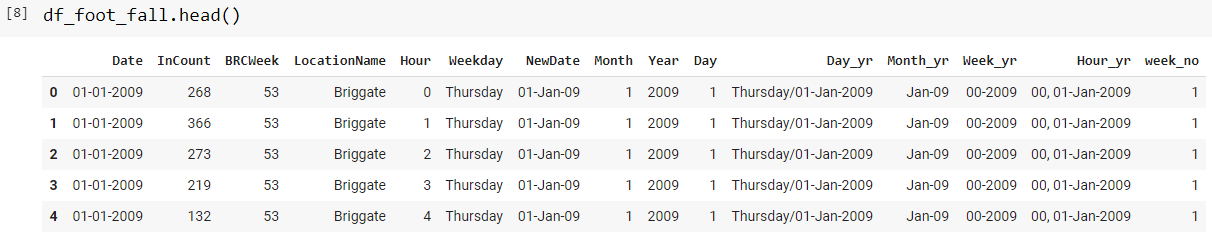


Figure : Original data snapshot

This is reshaped based on location name and multiple columns are created for each location.

After reshaping :

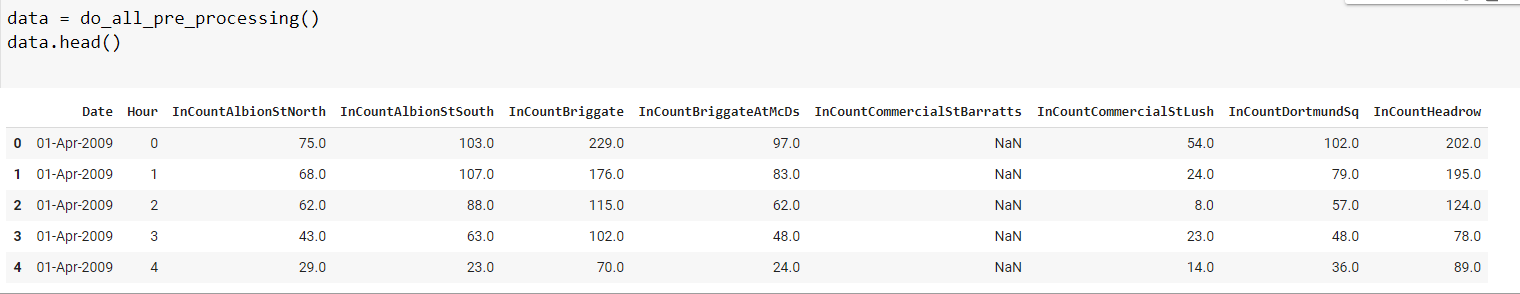


Figure : Reshaped Data snapshot

* 1. Aggregation of Data

Footfall data is aggregated based on Date and Hour to create a daily data.

After aggregation, daily data looks like this:

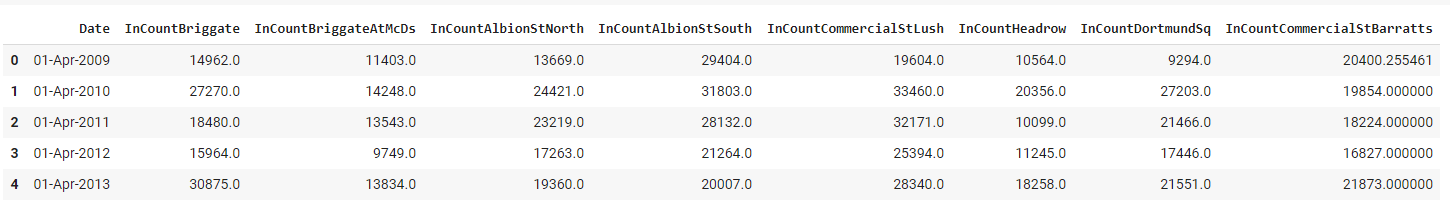
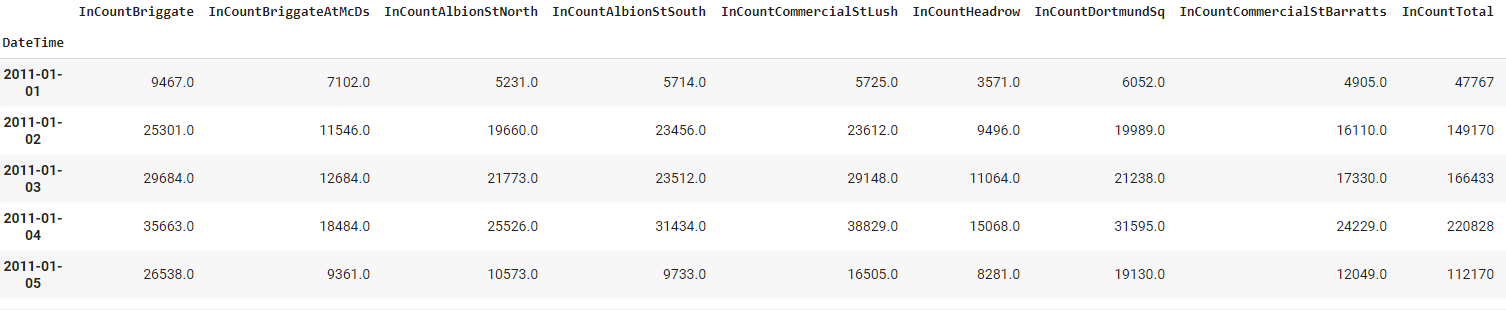


Figure : Aggregated data snapshot

* 1. Merge of Weather and Holidays data to footfall data

Weather and holiday’s data are merged to footfall data. The full dataset after merge :



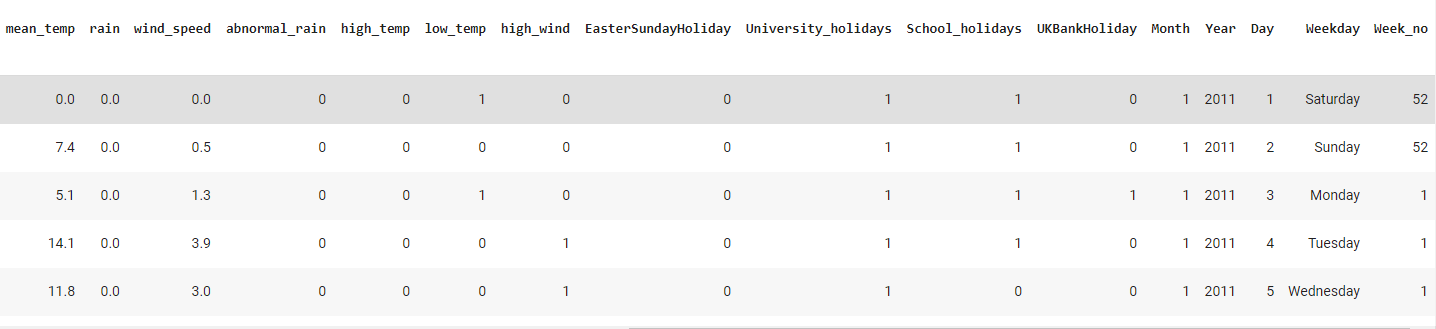
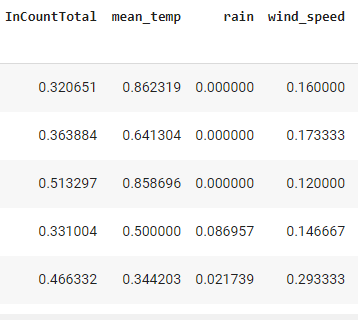


Figure : Merged data snapshot

* 1. Normalize the data

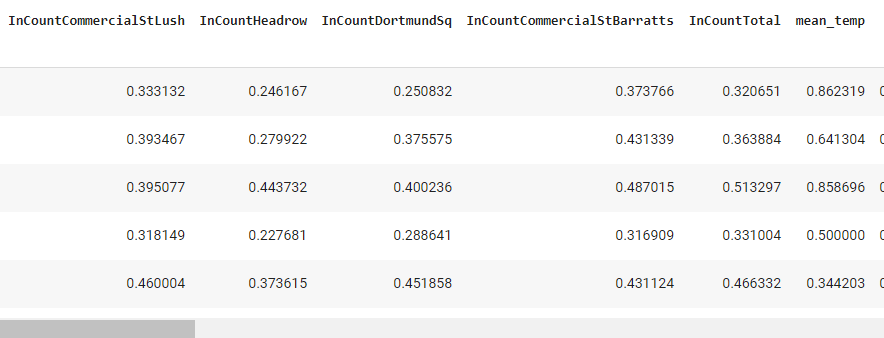
MinMaxScaler() is used for normalizing the numerical columns



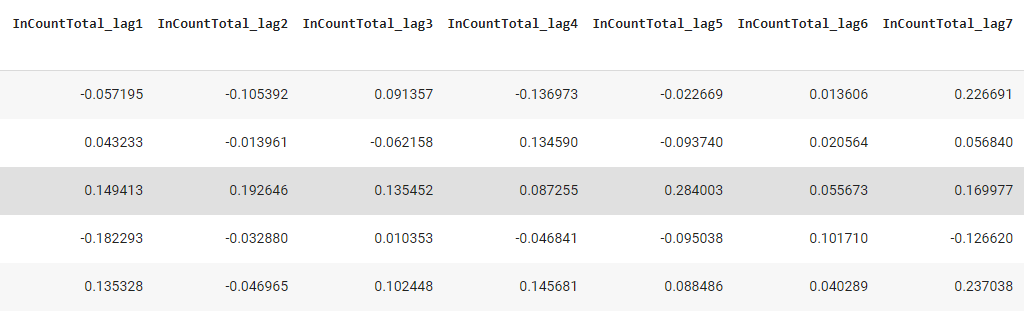
* 1. Derive new columns

The following new columns are created :

1. InCountTotal is derived as the sum of InCount of all locations.



1. Lag columns : 7 lag columns are created to use for supervised learning – previous 7 days of InCount is used as the Incount of these lag columns.



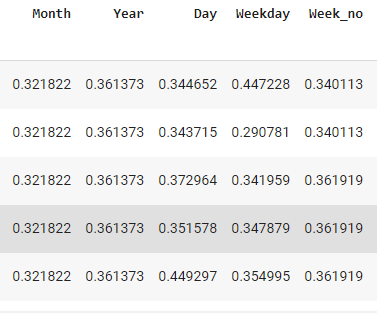
1. Target column : The target column is shifted by 1 day.
   1. Conversion of Categorical columns to numerical columns

TargetEncoding is used to convert multi-class categorical values to numerical values. Target encoding encodes the categorical values based on their distribution against the target variable.

Before target encoding :



After target encoding :



* 1. Stationarity check

ADF (Augmented Dicky-Fuller) test is used to determine if the given the dataset is stationary.

ADF test results for InCountTotal

ADF Statistic: -7.638551

p-value: 1.9231439534300742e-11

full result (-7.638551478845224, 1.9231439534300742e-11, 28, 2869, {'1%': -3.4326313386090552, '5%': -2.8625479402718623, '10%': -2.5673065559730883}, 70040.73650505886)

Inference : Since p-value is << 0.05, we can cleary deduce that the given series is stationary.

* 1. Decomposition of data

Decomposition of data is done using seasonal\_decompose from statsmodels.

|  |  |
| --- | --- |
| C:\Users\ed\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\5FB383A1.tmp | **Technique:** decomposition using seasonal\_decompose  **Stage Used**: After Normalising the data, applied on InCountTotal.  **Purpose**: To understand the trend and seasonality in the dataset  **Data**: InCount data  **Inference**: 1) there is no clear trend visible 2) seasonality is visible  **Action Taken**: created a separate dataset after detrending. |

|  |  |
| --- | --- |
| C:\Users\ed\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\F3212757.tmp | **Technique:** detrending using seasonal\_decompose  **Stage Used**: after decomposition of the data.  **Purpose**: To remove trend and seasonality  **Data**: InCount data  **Inference**: after detrending , the data becomes zero mean centered.  **Action Taken**: created a separate dataset for detrending. |

* 1. Outlier detection
     1. DB Scan

|  |  |
| --- | --- |
| C:\Users\ed\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\E35ED5B1.tmp | **Technique:** DBSCAN  **Stage Used**: After Normalising the data, applied on the entire dataset.  **Purpose**: To understand outliers in the data  **Data**: all the data variables  **Inference**: only 2 points are seen as outliers. They also don’t seem to real far outlies  **Action Taken**: Retain those points in the dataset |

* + 1. Using LOF ( Local Outlier Factor)

|  |  |
| --- | --- |
| C:\Users\ed\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\4DB1A7C8.tmp | **Technique:** LOF  **Stage Used**: After Normalising the data, applied on the entire dataset.  **Purpose**: To understand outliers in the data  **Data**: all the data variables  **Inference**: many points are shown as outliers, but all are based on local reachability density. In our specific business context, the LOF doesn’t seem applicable  **Action Taken**: Retain those points in the dataset |

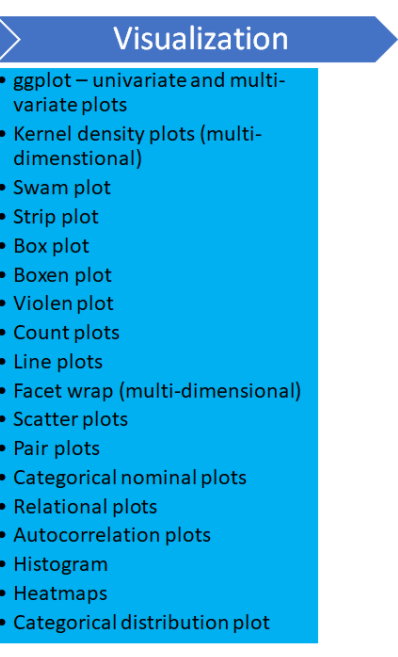
* 1. PCA analysis for dimensionality reduction

|  |  |
| --- | --- |
| C:\Users\ed\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\E009693.tmp | **Technique:** PCA  **Stage Used**: PCA was done towards the end of pre-processing step.  **Purpose**: To understand how many principal components give maximum variance  **Data**: all the data variables  **Inference**: 85% variance can be preserved with 6 principal components and 90% can be preserved with 11 principal components  **Action Taken**: Since we don’t have too many number of variables, we don’t intend to use PCA to reduce the dimensions |

* 1. Missing value detection and imputation

|  |  |
| --- | --- |
| C:\Users\ed\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\CE6D0916.tmp | **Technique:** missingno  **Stage Used**: After reshaping and merging the data, missing value check was done.  **Purpose**: To understand missing values in the data  **Data**: all the data variables  **Inference**: only 2 missing values found for CommercialSt.Barrats gate  **Action Taken**: Missing values are imputed using KNNImputer |

## **Visualization for Summarization**



|  |  |
| --- | --- |
|  | **Technique:** ggplot  **Stage Used**: After Pre processing data  **Purpose**: To Understand the histogram of Univariate variables IncountTotal, Rain, mean\_temp and wind speed  **Data**: Visualization data  **Inference**:   * mean of InCount : 180000 * mean of rain < 1 * mean of temp : 10.5 * mean of wind\_speed : ~1.75   **Action Taken**: |
|  | **Technique:** facet ggplot  **Stage Used**: After Pre processing data  **Purpose**: To understand the incount distribution based on Weekday, high temperature and abnormal rain  **Data**: Visualization data  **Inference**:   * The Incount is significanty low during the days with Abnormal rain and High temperature * Abnormal rain seems to have more impact on the Incount than the High temperature. Days with only Abnormal rain has slightly low footfall than the days with only High Temperature   **Action Taken**: |
|  | **Technique:** seaborn swarm and strip plots  **Stage Used**: After Pre processing data  **Purpose**: To understand the scatter plot for incount based on the weekday and month  **Data**: Visualization data  **Inference**:   * Weekdays: There is a gradual increase in incount from Sunday to Saturday. Saturday being highest incount * Months : we can see a seasonal trend in months - peaks in the months of Dec, Nov and March. Lower in the summer months(Jul and August) and also in Jan   **Action Taken**: |
|  | **Technique:** seaborn box, violin and boxen plot for data distribution  **Stage Used**: After Pre processing data  **Purpose**:To understand the Incount Total Distribution based on weekday.  **Data**: Visualization data  **Inference**:   * Distribution of data in more on Saturday which is between 100000 and 300000 * Monday to Friday the distribution is similar * On some of the week days the distribution is also minimum except monday and saturdays. * Saturday we can see that there are some data which has count of ,400000 count compared to other days where data is normally at 350000   **Action Taken**: |
|  | **Technique:**  **Stage Used**:  **Purpose**:  **Data**:  **Inference**:  **Action Taken**: |

1. Machine Learning Modelling & Techniques Applied

aaaaaaaaaaaaaaa

1. Code & Screenshots (Full screenshot without cropping the system date-time)

aaaaaaaaaaaaaaa

1. Interpretation

aaaaaaaaaaaaaaa

## **Justification of Measures / Metrics used**

aaaaaaaaaaaaaaa

## **Project output in terms of above measures/metrics**

aaaaaaaaaaaaaaa

## **Visualization Plots**

aaaaaaaaaaaaaaa

1. Future Work & Extension or Scope of improvements

aaaaaaaaaaaaaaa

1. Conclusions / Recommendations

aaaaaaaaaaaaaaa

1. Directions for future work

aaaaaaaaaaaaaaa

1. Bibliography / References

aaaaaaaaaaaaaaa

1. Appendices

aaaaaaaaaaaaaaa

1. List of Publications/Conference Presentations, if any.

aaaaaaaaaaaaaaa

1. Duly Completed Checklist

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

Table : Completion Checklist

**Note: A copy of this checklist should be included as the last page of the Final report. This checklist, duly completed and signed by the student, should also be verified and signed by the evaluators. Evaluators are requested to ensure that the students have submitted their reports properly.**