**PROJECT REPORT ON**

**Predicting Airbnb Host Revenue**



**Submitted in partial fulfillment for the award of PG Diploma in Big Data Analytics from C-DAC ACTS (Pune)**

**Presented by:**

#### Kshitij Jagtap PRN 220340125020

#### Malvika Sharma PRN 220340125024

#### Pratik Nikalje PRN 220340125027

#### Sarwesh Karande PRN 220340125039

#### Shital Kore PRN 220340125045

**Guided by:**

**Dr. Krishnanjan Bhattacharjee**

CERTIFICATE

### TO WHOMSOEVER IT MAY CONCERN

**This is to certify that Mr. Kshitij Jagtap Ms. Malvika Sharma Mr. Pratik Nikalje**

### Mr. Sarwesh Karande

### Ms. Shital Kore

**have successfully completed their project on**

**Predicting Airbnb Host Revenue**

## under the guidance of Dr. Krishnanjan B. Bhattacharjee

**Project Guide Project Supervisor**

#### HOD ACTS

**Mr. Sundar Gaur**

**ACKNOWLEDGEMENT**

This project ‘**Predicting Airbnb Host Revenue**’ was a great learning experience for us and we are submitting this work to Advanced Computing Training School (**CDAC ACTS**).

We all are very glad to mention the name of Dr. Krishnanjan B. Bhattacharjee for his valuable guidance to work on this project. The guidance and support from him helped us to overcome various obstacles and intricacies during the course of project work.

Our most heartfelt gratitude goes to Mr. Sunder Gaur (HOD, ACTS), Ms. Risha P.R. (Program Head, ACTS) and Ms. Priyanka Ranade (Course Coordinator, PGDBDA) who gave us all the required support and coordinated with us to provide all the necessities we needed to complete the project and throughout the course up to the last day here in C-DAC ACTS, Pune.

**From**

**Mr. Kshitij Jagtap Ms. Malvika Sharma Mr. Pratik Nikalje**

### Mr. Sarwesh Karande

### Ms. Shital Kore

## Table of Contents

**Predicting Airbnb Host Revenue**

**0**

1. [Abstract 4](#_TOC_250035)
2. [Introduction and Overview of Project 5](#_TOC_250034)
   1. [Objectives of the project: 6](#_TOC_250033)
3. [Data Description and Technologies Implemented 7](#_TOC_250032)
   1. [Data Description 7](#_TOC_250031)
   2. [Information on the technologies being used: 8](#_TOC_250029)
      1. [Python 8](#_TOC_250028)
      2. [Pandas 8](#_TOC_250027)
      3. [Streamlit 9](#_TOC_250026)
      4. [Scikit-learn 9](#_TOC_250025)
      5. [Apache Airflow 9](#_TOC_250024)

[3.2.6. Data visualization libraries - Seaborn, Matplotlib 9](#_TOC_250023)

[3.2..7.Apache Hadoop 10](#_TOC_250023)

1. [Initial Steps 11](#_TOC_250022)
2. [Workflow 12](#_TOC_250021)
   1. [Research 13](#_TOC_250020)
      1. [Deep analysis on the price distribution, based on locations and apartment properties](https://github.com/ruchigupta19/Boston-Airbnb-data-analysis) 13
      2. Revenue Vs Price 13
   2. [Problem Statement 13](#_TOC_250018)
   3. [Project](#_TOC_250017) Structure 14
   4. Environments and Applications 15
      1. [Environments 15](#_TOC_250015)

5.4.2.Applications 15

5.5. [Data Science Process 16](#_TOC_250013)

5.5.1. Data Collection 16

5.5.2. Data Cleaning and Preprocessing 17

5.5.3. Exploratory Data Analysis (EDA) 19

5.5.4. Feature Engineering and Selection 22

5.5.5. Modelling 23

5.5.5.1. Comparison of Model Performance 25

5.5.5.2. Findings 25

**6. Generalizing Data Science Lifecycle 26**

6.1.[Pipeline 26](#_TOC_250011)

6.1.1. ETL 27

6.1.2. DataInterface 29

6.1.3. Modelling 30

6.1.3.1. Feature Engineering and Selection 30

6.1.3.2. Model Building 30

6.1.4. Insights 31

6.1.5. Streamlit 32

6.1.5.1. Internal dashboard 34

1. [Conclusion 36](#_TOC_250002)
2. [Future Scope 37](#_TOC_250001)
3. [Bibliography 38](#_TOC_250000)

Table of Figures

Figure Name

1. Project Structure
2. Insideairbnb Website
3. Code Snippet 1
4. Code Snippet 2
5. Code Snippet 3
6. Code Snippet 4
7. Code Snippet 5
8. Code Snippet 6
9. Code Snippet 7
10. Code Snippet 8
11. Modelling
12. Metrics Comparison
13. ETL Block Diagram
14. Table of ETL Class
15. Table of DataInterface Class
16. Modelling Block Diagram
17. Table of Model Class
18. Table of Insights Class
19. Table of InsightEngine Class
20. Streamlit Block Diagram
21. Streamlit Prediction Page
22. Streamlit Explore Data Page
23. Streamlit Statistics Page
24. Streamlit Pipeline Page
25. Streamlit Metrics Page

Page no.

14

9

16

16

17

18

18

19

20

21

22

23

25

26

26

28 29

29

31

31

32

33

33

34

34

35

# Abstract

Big data analytics and machine learning are the technologies primarily used in analysis and prediction of large datasets. These allow us to gain valuable insights by processing large amounts of data. This paper included Airbnb host revenue and booking rate prediction. Data collection is done from <http://insideairbnb.com/>. It provides datasets for various cities.

Apache Airflow helps to build project workflow using pipelines. Python being majorly used for Data Science projects provides easy functions and APIs to manipulate the data. The powerful visualization libraries Seaborn and Matplotlib help in visualizing and exploring the data.

Machine learning is the technology helping in predictions of the host revenue and booking rates. Streamlit is the backbone to provide User Interfaces for trained machine learning models. Ths project uses Apache Airflow to manage the workflow with which project tasks are automated using pipeline. Pipeline outputs the processed dataset which is used by machine learning model for revenue and booking rate predictions. Streamlit uses the model for real-time revenue and booking predictions.

# Introduction and Overview of Project

Big data refers to vast amounts of data that traditional storage methods cannot handle. Machine learning is the ability of computer systems to learn to make predictions from observations and data. Machine learning can use the information provided. Big data includes collecting data from various sources and making effective storage of data. Web scraping can scrape data from websites using libraries such as BeautifulSoup. Machine learning is used by various sectors along with the huge amount of data to make predictions and to get insights from the data.

Human mind cannot process such large amounts of data easily and make any meaningful sense of it. New developments in machine learning and big data have driven analytics to turn these vast quantities of data into actionable information. E-commerce uses it to give product recommendations. Shopping malls can use it to predict demand for the products in future to avoid shortage or excessive storage. Hotel chains can use it to predict their revenue and booking rates depending on the historical data of the customers.

The project uses such historical data for the Airbnb hosts from various cities to predict revenue and booking rate. Using more generalized model from various models we tend to make accurate predictions to increase the business. Streamlit helps host to make real-time predictions of their presently available data.

## 2.1. Objectives of the project:

* Optimizing booking rate and revenue.
* Create a platform for Airbnb to process, transform, store data of different cities.
* Analyse and model the data to get useful insights with a user interface to display these insights.

# Data Description and Technologies Implemented

## Data Description

Data source used for data collection is

<http://insideairbnb.com/get-the-data/>

The source has dataset for hosts of different cities.

## Information on the technologies being used:

* + 1. **Python** –

[Python](https://www.geeksforgeeks.org/python-programming-language/) is open source, interpreted, high level language and provides great approach for object-oriented programming. It is one of the best language used by data scientist for various data science projects/application. Due to various APIs in Python it is easy to use for data handling operations. Project uses libraries such as pandas, numpy, matplotlib and seaborn to handle data frames and visualizations.

* + 1. **Pandas** -

Pandas is an open-source library that is made mainly for working with relational or labeled data both easily and intuitively. It provides various data structures and operations for manipulating numerical data and time series. This library is built on top of the NumPy library. Pandas is fast and it has high performance & productivity for users. Project fetches datasets from insideairbnb.com which Pandas can best manipulate with its available functionalities.

**Features:**

* Fast and efficient for manipulating and analyzing data.
* Easy handling of missing data (represented as NaN) in floating point as well as non-floating point data
* Data set merging and joining.
* Provides time-series functionality.

* + 1. **Streamlit-**

Streamlit is an open source app framework in Python language. It helps us create web apps for data science and machine learning in a short time. It is compatible with major Python libraries such as scikit-learn, NumPy, pandas, Matplotlib etc.Data caching simplifies and speeds up computation pipelines.

Project uses Streamlit to show User Interface where trained model makes predictions for unlabelled data.

* + 1. **Scikit-learn-**

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering, etc in Python. Revenue and booking prediction being a regression problem statement, we use different supervised models such as KNN, Xgboost, Random forest, regularization for regression.

* + 1. **Apache Airflow-**

Apache Airflow is a workflow engine that will easily schedule and run complex data pipelines. It makes sure that each task of the data pipeline gets executed in the correct order and each task gets the required resources. We can monitor pipelines easily which is a tedious task in normal use but Airflow User Interface makes it easier. It is compatible with python.

* + 1. **Data visualization libraries - Seaborn, Matplotlib-**

Data Visualization is an extremely important part of Data Analysis. It helps to find the hidden patterns and layers in the data in a visual format. Human brain interprets visuals than textual data. Data visualization charts like **bar charts, scatterplots, line charts, geographical maps,** etc. are extremely important.

Matplotlib is a python bases 2-D data visualization library. Seaborn is a Python data visualization library that is based on Matplotlib and closely integrated with the NumPy and pandas data structures.

* + 1. **Apache Hadoop –**

Hadoop is an Apache open source framework written in java that allows distributed storing and processing of large datasets across clusters of computers using simple programming models. Hadoop is designed to scale up from single server to thousands of machines, each offering local computation and storage. Hadoop comes with a distributed file system called HDFS. In HDFS data is distributed over several machines and replicated to ensure their durability to failure and high availability to parallel application.It is cost effective as it uses commodity hardware.

# Initial Steps

* Initially gain the knowledge about Airbnb and different research papers on Airbnb predictions.
* After getting sufficient knowledge, we define our problem statement and accordingly we decide what type of data required solving our problem statement.
* Then we began to search different sources from where we could get historical data of Airbnb.
* After sufficient research we finalize one resource for our data.

<http://insideairbnb.com/get-the-data>

# Workflow

1. Research
2. Problem Statement
3. Deciding on program structure
4. Setting & running of required Environments & Applications
5. Data Science process:
   * Data Collection
   * Data cleaning and preprocessing
   * EDA
   * Feature Engineering
   * Model Building
6. Generalized Pipeline
   * ETL
   * Modeling
   * Insights
     + Graphs
     + Statistics
7. Building Streamlit UI

## Research

**5.1.1.** [**Deep analysis on the price distribution, based on locations and apartment properties**](https://github.com/ruchigupta19/Boston-Airbnb-data-analysis)

The article explores Boston dataset to find various business questions with the help of three different datasets -

listings csv consists of details of all the listings in Boston

calendar.csv consists of details of listings and its availability and its price.

reviews.csv consists of reviews for each listing in Boston.

Business questions -

* What causes difference in prices of listings?
* Where to invest a property in boston to get maximum returns from airbnb?
* Seasonal pattern of prices
* Sentiment analysis of reviews & its relation with price
* Host analysis & recommendation system for prices
  + 1. **Revenue vs. Price**

Revenue is the total monitory value of services sold while Price is the amount of money set as consideration for the sale of specified service.

## Problem Statement

1. Exploratory analysis of Airbnb data of a few cities to get insights.
2. Building model for revenue and booking rate predictions.
3. Building pipeline by generalizing data science lifecycle to fit any city.
4. Web application interface for host for revenue and booking predictions and insights.

## Project structure

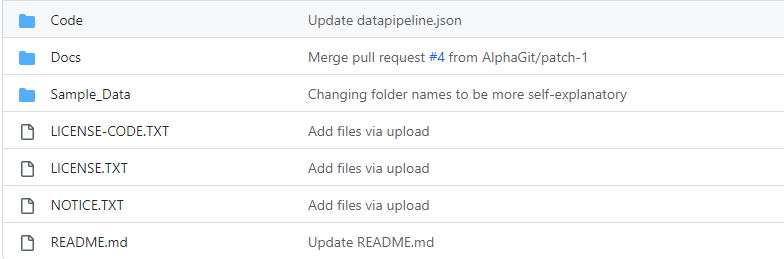


Fig 1: Project Structure

## 5.4. Environments and Applications

### Environments:

#### Python

#### [Python](https://www.geeksforgeeks.org/python-programming-language/) is open source, interpreted, high level language and provides great approach for object-oriented programming.

#### HDFS:

Setting up of Hadoop distributed file system as data warehouse for project.

#### Apache Airflow:

[Apache Airflow](https://airflow.apache.org/docs/stable/) is a workflow engine that will easily schedule and run complex data pipelines. It is one of the most robust platforms used by Data Engineers for orchestrating workflows or pipelines.

#### Apache Spark:

Setting up spark compatible to our Hadoop environment and accordingly installing python interface of Apache Spark(Pyspark).

#### Hive:

Setting up of hive for storing the tables onto Metastore db as externally managed tables.

* + - 1. **Streamlit:**

Open source framework to create data applications.

### Applications:

* + - 1. IDE : Pycharm, Spyder, Jupyter Notebook
      2. Apache Airflow
      3. VMWare workstation
      4. Github

## Data Science process:

## 5.5.1. Data collection

Data collection is process of gathering data from various sources of data.

Data collection is done from <http://insideairbnb.com/get-the-data/>.

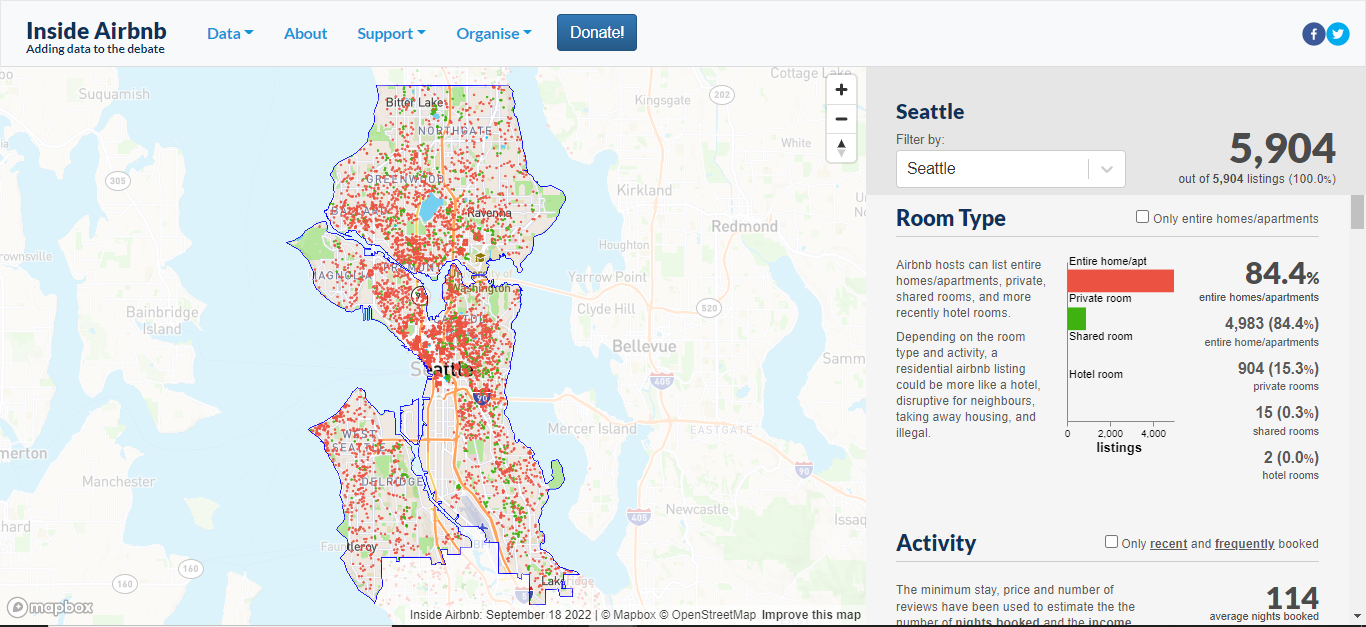


Fig 3: Insideairbnb Website

Data scraping is done using BeautifulSoup python library to scrape links from the site. The dataframe is created with links metadata which includes city and its repsective CSV file links.



Fig 4: Code Snippet 1

## Data Cleaning and preprocessing

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format. Any machine learning model needs cleaned data to be able to predict the labels. Data cleaning and preprocessing is important step which includes various tasks such as Imputation, handling missing values.



Fig 5: Code Snippet 2



Fig 6: Code Snippet 3

Fig 7: Code Snippet 4

## Exploratory Data Analysis (EDA)

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations. This step took a longer time to explore data as it is better to understand data as much as possible. This answers initial business questions of Airbnb Hosts.

**Few examples of EDA –**

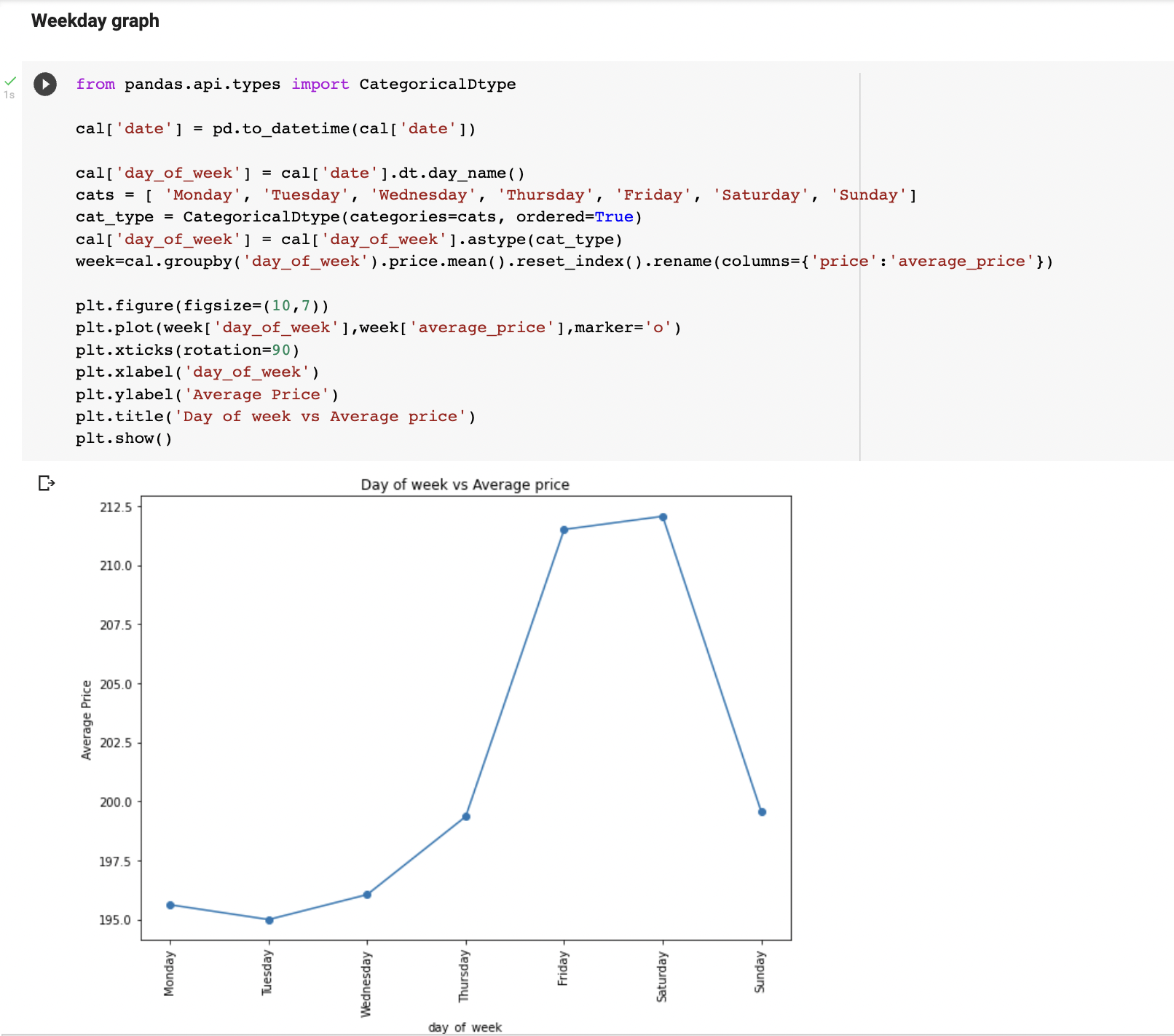


Fig 8: Code Snippet 5



Fig 9: Code Snippet 6

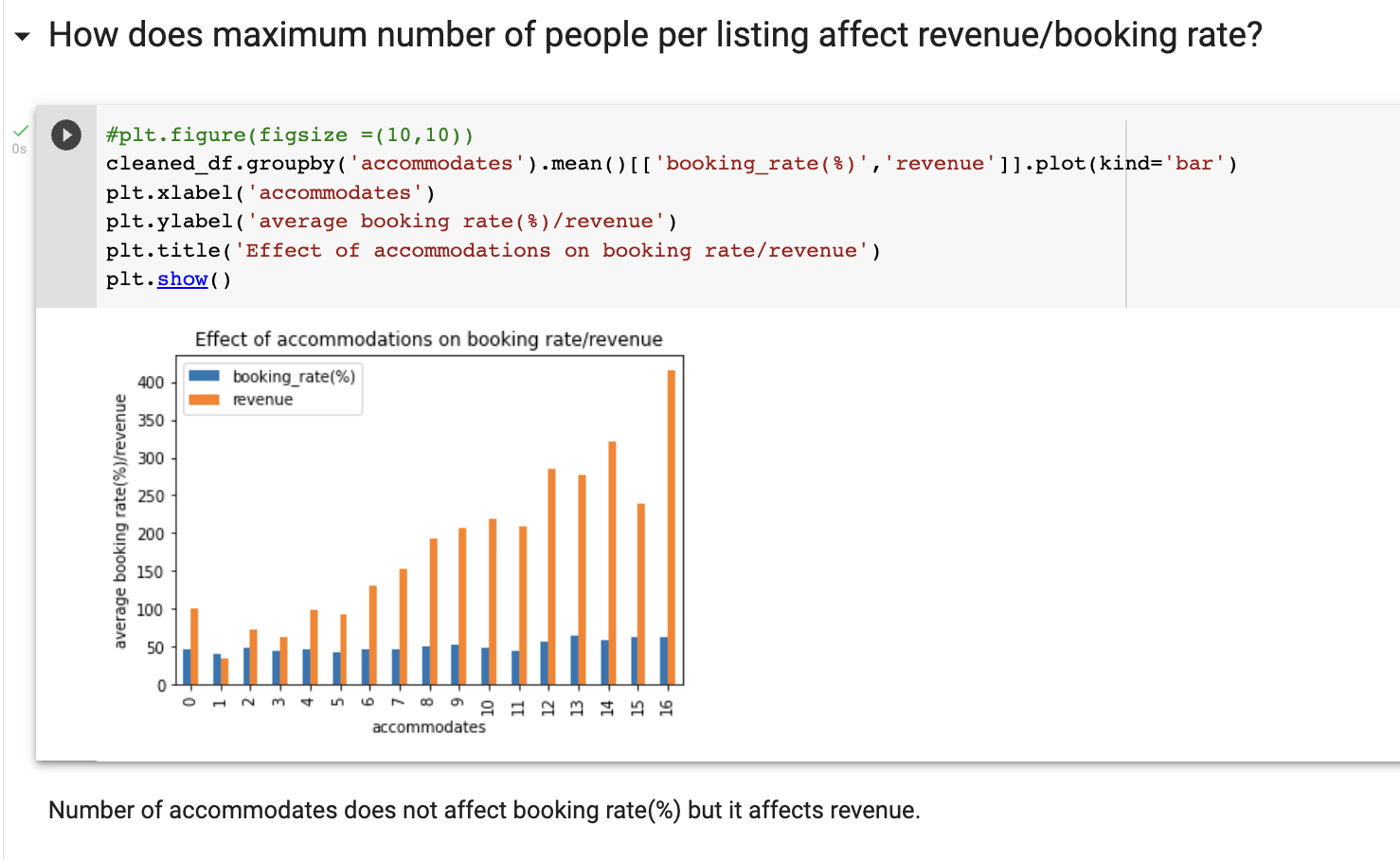


Fig 10: Code Snippet 7

## Feature Engineering and Selection

Feature engineering is the pre-processing step of machine learning, which is used to transform raw data into features that can be used for creating a predictive model using Machine learning or statistical Modelling. Our preprocessed dataset contains 44 features. Keeping large number of features causes bad model performance and increases training time. Hence it is necessary to select most important features which have maximum influence of revenue and booking rate.

Also it is important to create new features from existing features which influence revenue and booking rate. This is called as Feature Construction.

We constructed following features:

1. Accommodation per bed
2. Price surge percent - To check abrupt changes in price during peak season
3. Price standard percentage

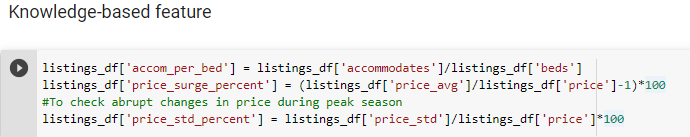


Fig 11: Code Snippet 8

**Using various feature selection algorithms we selected top important features:**

1. price – daily price of room in $
2. availability\_365 - availability of room for X number of days of 365 days
3. beds – number of beds
4. accommodates - The maximum capacity of the listing
5. availability\_60 - availability of room for X number of days of 60 days
6. review\_scores\_rating – Ratings given to host
7. reviews\_per\_month - The number of reviews the listing has over the lifetime of the listing from 1 to 5
8. availability\_30 - availability of room for X number of days of 30 days

## Modelling

A machine learning model is a mathematical structure that is trained to recognize certain types of patterns. Model is trained over a set of data, providing it an algorithm to learn from those data.

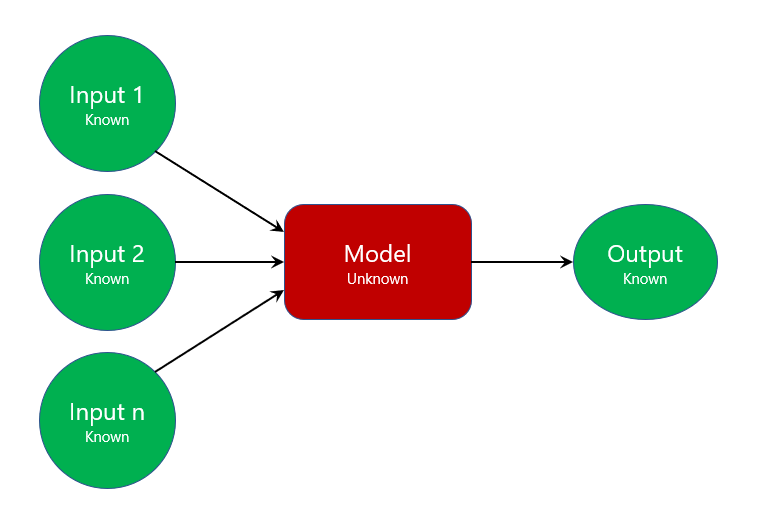


Fig 12: Modelling

A model in machine learning is the output of a machine learning algorithm run on data. We train machine learning algorithm on the data available and create a model. This model helps in prediction of unlabelled data. We used different models and compared there metrics to find most generalized model to predict revenue and booking rate.

1. **Linear Regression**

It is a supervised machine learning algorithm used for regression. It predicts dependent variable based on independent variables by finding linear relationship between them using equation



It is simple to implement but it assumes linear relationship. Linear model can be ovefitted but it can be avoided using Regularization techniques.

1. **Regularization**
   1. **Ridge (L2 Regularization)**

We avoid overfitting of Linear model by adding penalty which is square of coefficient to Linear regression, called as L2 Regularization or Ridge regression. But Ridge model also gives importance to those features not having significant impact on response variable. To avoid this we use L2.

* 1. **Lasso (L2 Regularization)**

Lasso helps in feature selection by giving importance to those features having maximum impact on response variables and also prevents overfitting.

1. **Polynomial + Regularization**

Polynomial comes with different degrees with each degree, linearity of data decreases. It can give best performace with Regularization.

1. **K Nearest Neighbors**

This is a supervised machine learning algorithm used to predict classification as well as regression. For revenue and booking rate prediction we used K Nearest Neighbors regressor. But if dataset is unbalanced then it gives biased output. Also it is affected by outliers

1. **SVR**

In simple regression we try to minimise the error rate. While in SVR we try to fit the error within a certain threshold.

1. **Random Forest**

This is bagging technique which internally uses multiple decision tress to reduce overall variance of all the weak learning models. But we need to limit the height else it tends to overfit.

1. **Xgboost**

It is a boosting technique based on Gradient Boosting. It uses multiple decision tress arranged sequentially. Wrongly predicted results are fed to next tree which tries to correct prediction of previous tree.

.

* + - 1. **Comparison of model performace –**

|  |  |
| --- | --- |
| Model | R2 score |
| K Nearest Neighbors | 0.8450578330169121 |
| SVR – Linear | 0.7852993518404301 |
| SVR – Radial | 0.4949102276874262 |
| XGBRegressor | 0.9533370165073544 |
| Random Forest | 0.8810420834518464 |
| Linear Regression | 0.8068319637301892 |
| Ridge Regularization | 0.8067676164771325 |
| Lasso Regularization | 0.8069592093371641 |
| Polynomial with Ridge | 0.8067676164771325 |
| Polynomial with Lasso | 0.9802121530953398 |

Fig 13: Metrics Comparison

* + - 1. **Findings -**

From the table, linear models tend to perform well with the given data. But polynomial with Lasso gives best result which may suggest the relationship between independent and dependent variable is not perfectly linear, suggesting a curvilinear relationship.

Xgboost and random forest seem to perform well on almost every dataset with minimum bias and variance tradeoff.

# Generalizing Data Science Lifecycle

* 1. **Pipeline**

**ETL**

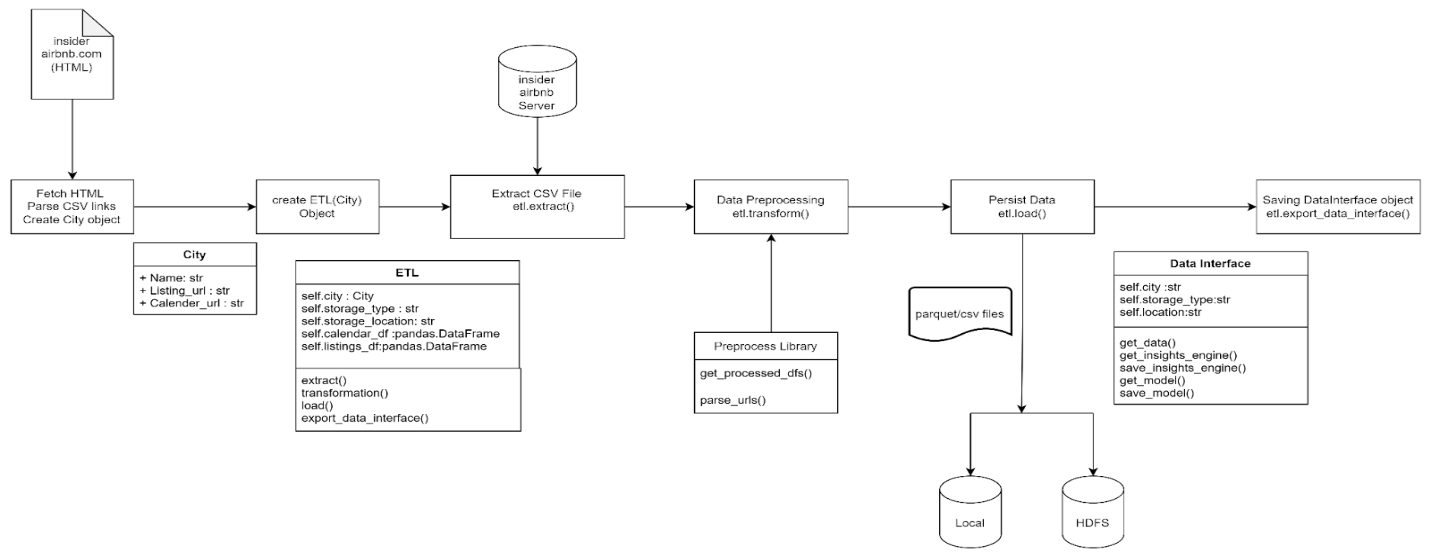


Fig 14: ETL Block Diagram

**Table of ETL Class**

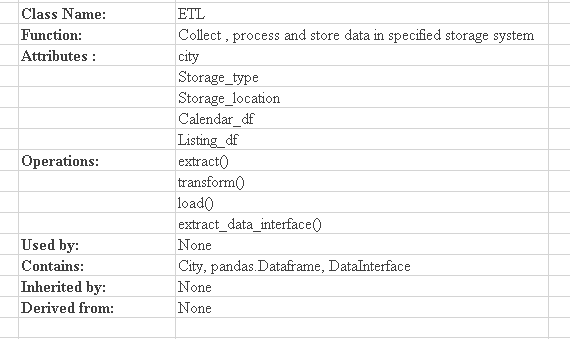


Fig 15: Table of ETL Class

**6.1.1. ETL**

1. **Extraction:**

Data Extraction can be defined as the process of collecting data from various sources for the purpose of storing that data, transforming it, and feeding it to another system for subsequent analysis.

We initiated the process of web scrapping with the help of beautiful soup library and fetched required data of different cities from InsideAirbnb.com which is in HTML format and then we passed only csv links from these HTMLs for different cities and created a dataframe from csv links that includes links, name of city, name of country, and type of dataset for all cities respectively. Then we created a city object to fetch url for a particular city we want to use in various sections of the project and ultimately converted the URL into pandas dataframe.

After the completion of extraction process we got two dataframes i.e calendar.csv and listings.csv for particular city that we specified.

1. **Transformation:**

Data transformation is an essential data preprocessing technique that must be performed on the data before data mining to provide patterns that are easier to understand. Data transformation changes the format, structure, or values of the data and converts them into [clean, usable data](https://www.zuar.com/blog/data-cleaning-the-benefits-and-steps-to-creating-and-using-clean-data/).

After getting the extracted data, we start processing the data by passing the raw dataframe to Preprocess class for following purposes:

* cleaning the data
* converting raw data in required data format
* categorical data converted to numerical data for calculation purpose
* handled null values
* Merge two dataframes(listings.csv and calendar.csv)

After Transformation process we get completely processed cleaned dataframe

1. **Load:**

Loading the data into the data warehouse is the last step of the ETL process. The vast volume of data needs to load into the data warehouse for a concise time. For increasing the performance, loading should be optimized. In this step, the extracted data and the transformed data are loaded into the target storage location.

In this project we provided two storage loactions i.e local and HDFS to load the cleaned dataframe. The dataframe is loaded in respective city’s folder which is automatically created at time of loading dataframe.

We defined a function to create and save datainterface object which will be later used in the program.

**Table of DataInterface Class**

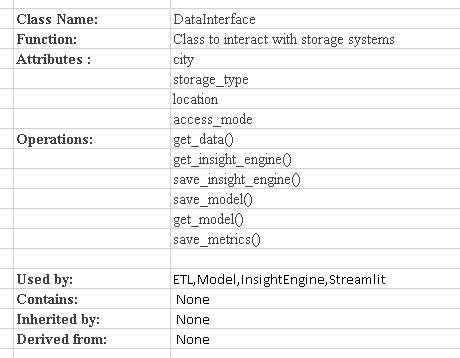


Fig 16: Table of DataInterface Class

**6.1.2. DataInterface:**

A gateway to independent and persistent storage. The main aim of implementing this class is to achieve Functional Independence. It is providing most classes with single point access to all stored data.

**Modelling**

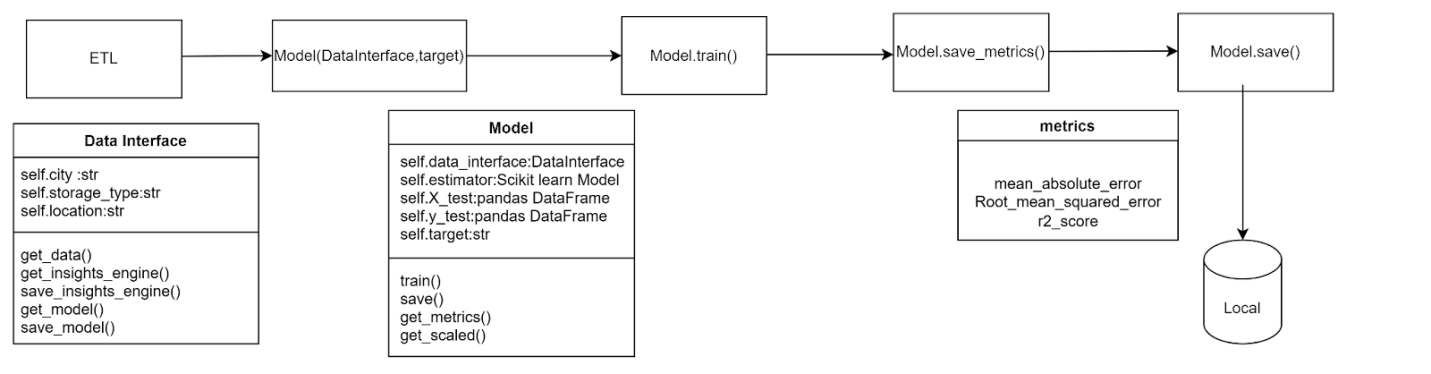


Fig 17:Modelling Block Diagram

**Table of Model Class**

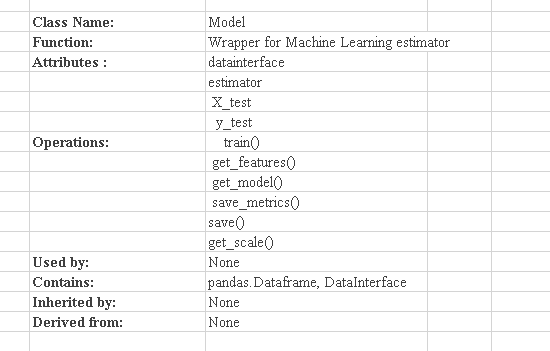


Fig 18: Table of Model Class

**6.1.3. Modelling:**

**6.1.3.1. Feature Engineering and selection**:

**Feature engineering is the pre-processing step of machine learning, which is used to transform raw data into features that can be used for creating a predictive model using Machine learning or statistical Modelling.** Feature engineering in machine learning aims to improve the performance of models.

To fetch cleaned and processed data we use get\_data() method of DataInterface Class. This data is then passed to get\_features() method in Model class for selecting the required features for model building. In the dataframe there were large number of fetaures having multiple categories, this causes the curse of dimensionality which means that the error increases with the increase in the number of features. So to avoid it we chose such features which covered 90-95% information of respective dataframe.

Ultimately, we selected top 10 features using various techniques that would be used to predict target variable.

* + - 1. **Model Building:**

In the train method() two tasks are being performed one for selecting features and the other to build models. The selected features dataframe obtained is used to train the model. In the process of building model various Supervised Machine Learning Algorithms were used such as XGBoost, Random Forest ,Linear Regression with L1 Regularization with poly degree 3.

In the project two models have been built , one for Revenue and the other for Booking Rate. Each model has their respective metrics and model object saved in the specified location . The saved csv metrics file provides the information about Mean Absolute Error, Root Mean Squared Error and R2\_score for Revenue and Booking rate model per city.

**Table of Insights Class**

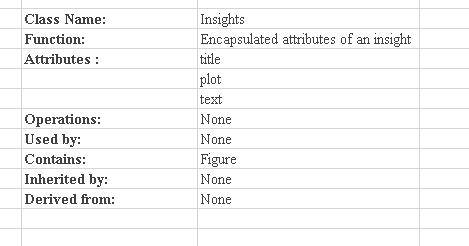


Fig 19: Table of Insights Class

**Table of InsightEngine Class**

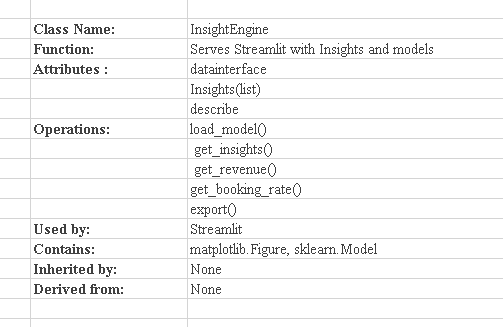


Fig 20: Table of InsighEngine Class

* + 1. **Insights:**

Insights are the knowledge gained through analyzing data, generating conclusions from the data that can benefit your business. Data are the input. Insights are the output.

create\_insight() method with the help of DataInterface is required to create useful insights. In this method many plots have been created that are stored in a list object and passed to InsightEngine Class. InsightEngine Class saves the insights object in the persistent storage.

**Streamlit**

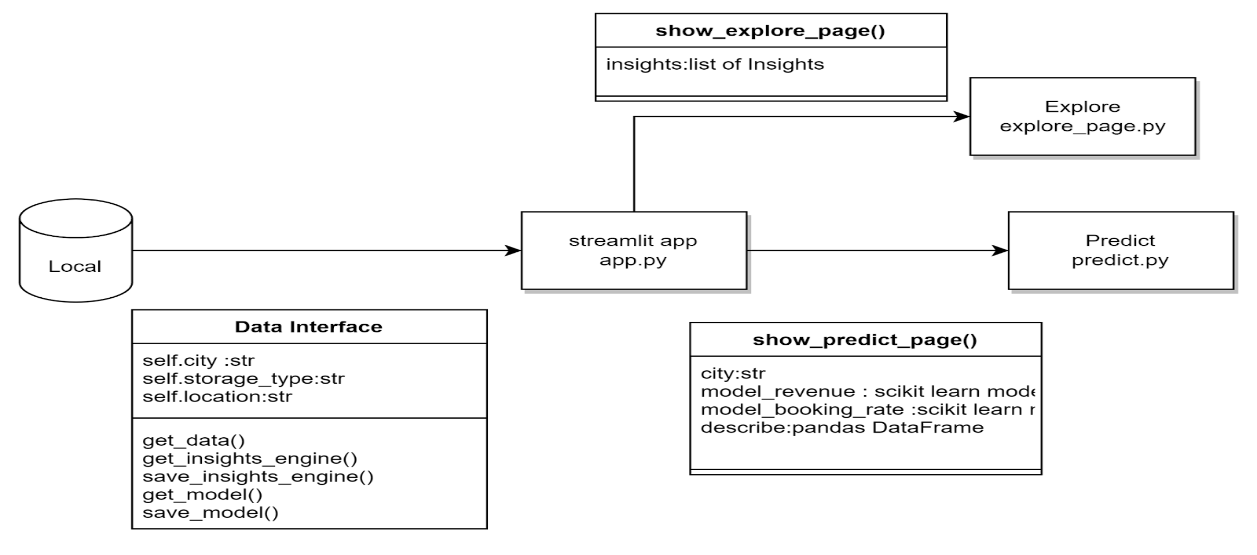


Fig 21: Streamlit Block Diagram

* + 1. **Streamlit:**

Streamlit UI contains 3 windows having their own functionalities

In Streamlit we have provided Airbnb host with list of cities for which he can take some useful decisions based on predictions and insights.

* **Predict**: This page shows predicted Revenue and Booking Rate by adjusting provided sidebar features. The Calculation is carried out using Revenue and Booking Rate model object which are already present in their city directory.

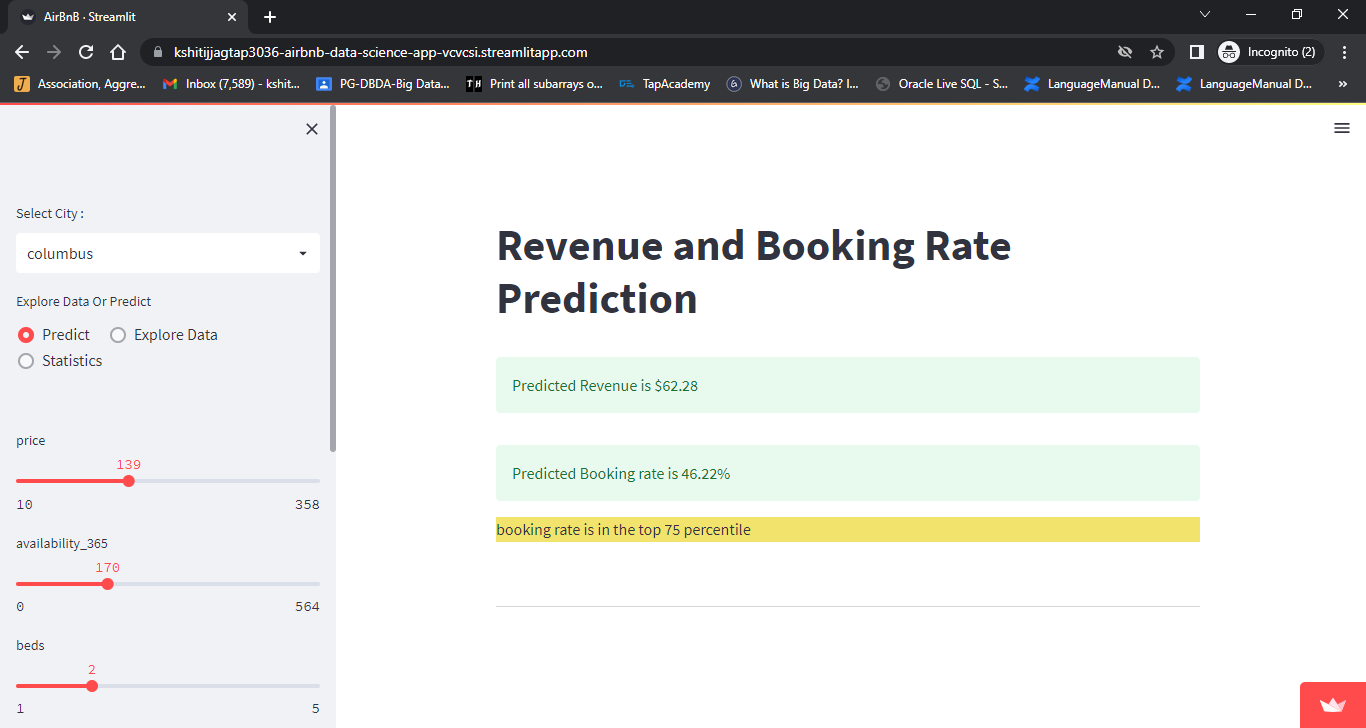


Fig 22: Streamlit Prediction Page

* **Explore Data:** This page shows various graphs that were created using insight object.

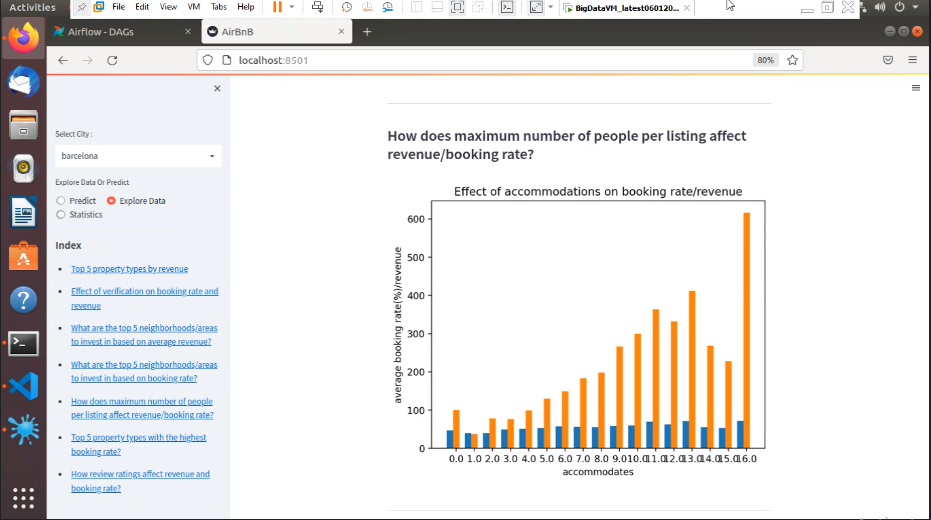


Fig 23: Streamlit Explore Data Page

* **Statistics**: This page shows the statistics for selected features that are provided in the first predict window.

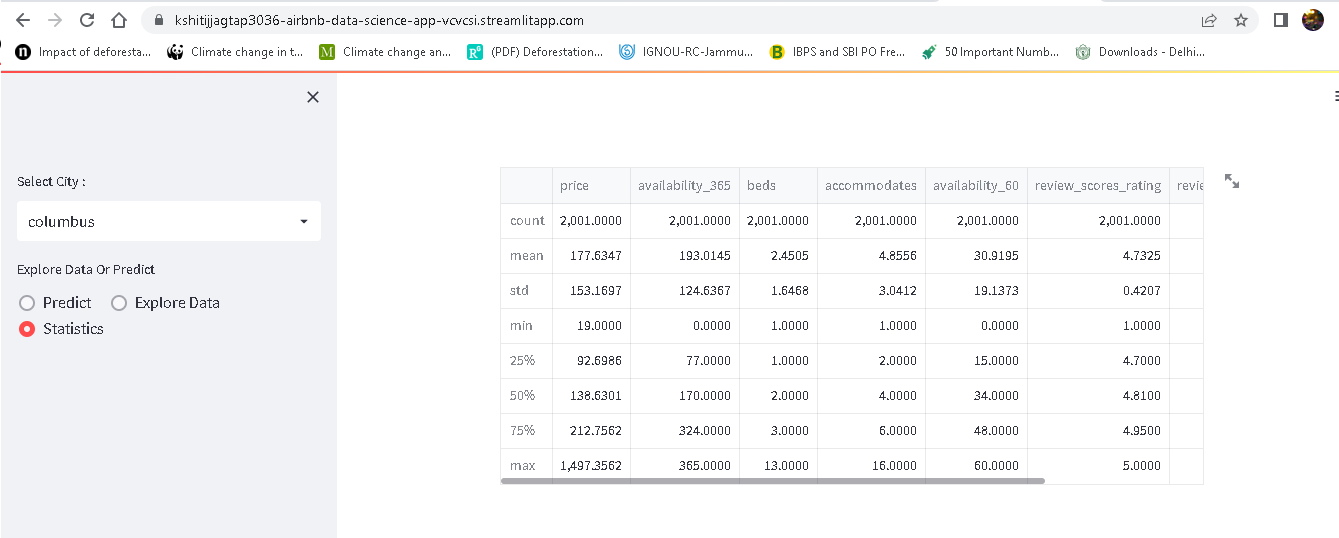


Fig 24: Streamlit Statistics Page

**6.1.5.1. Internal Dashboard**

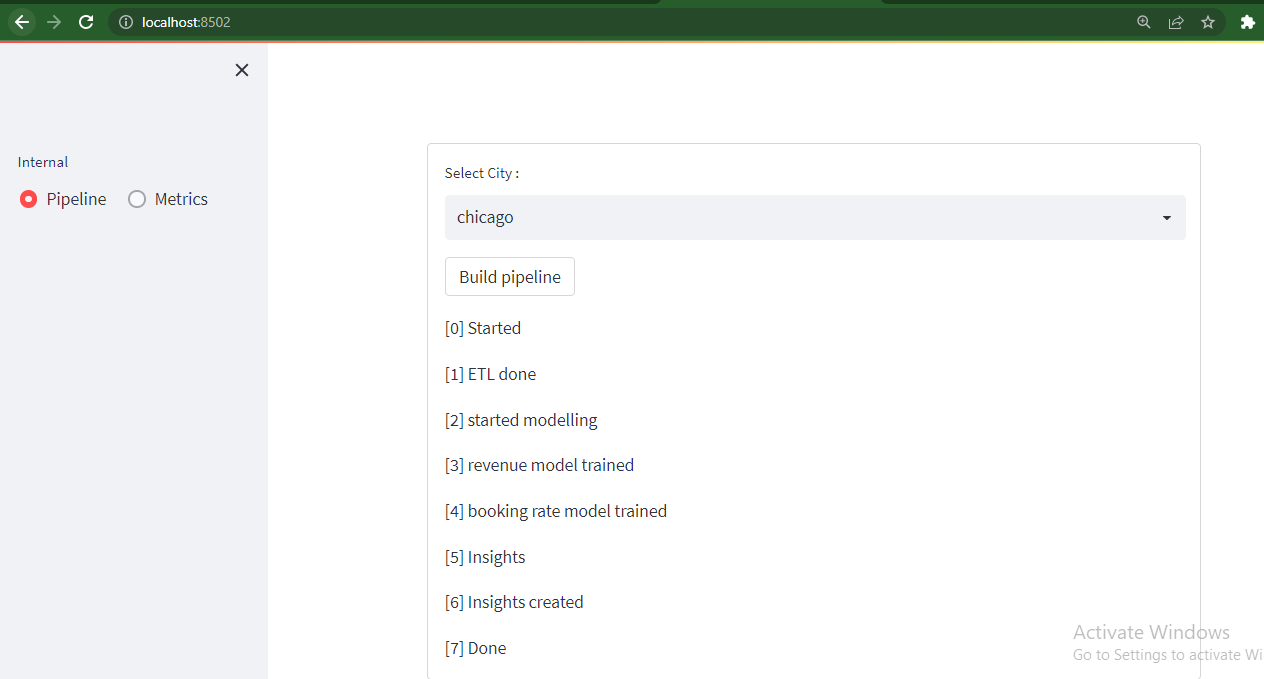


Fig 25: Streamlit Pipeline page

Above page is created to build pipeline for unprocessed cities.

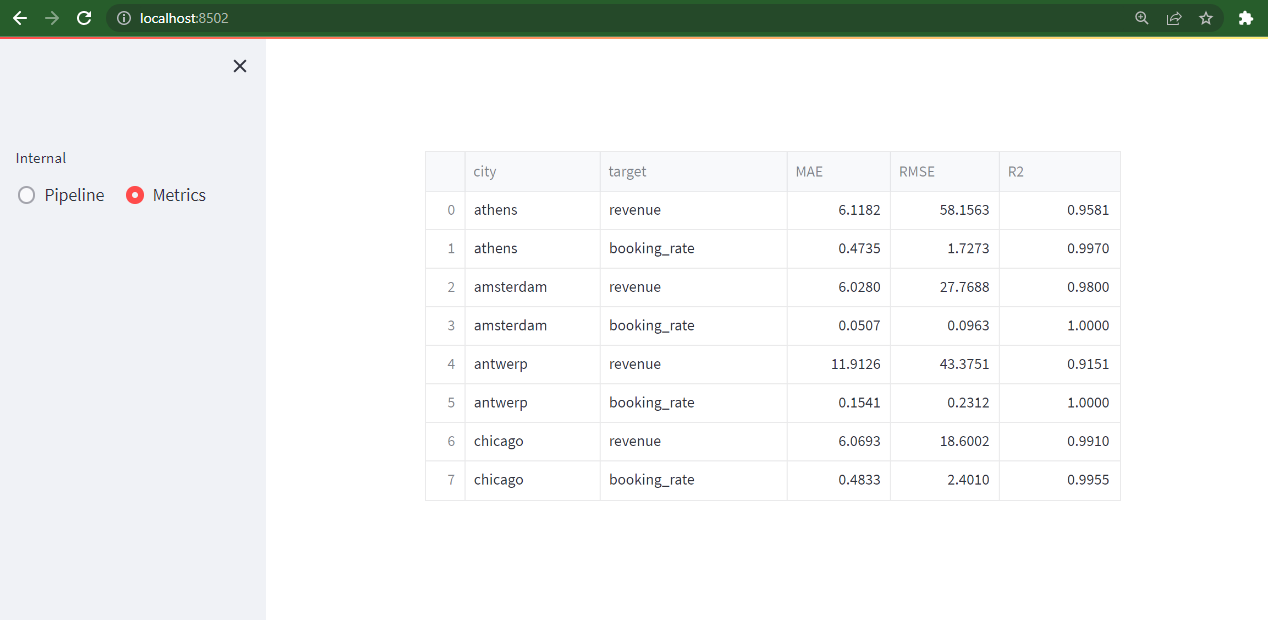


Fig 26: Streamlit Metrics page

Above page is created to compare metrics of different machine learning models for processed cities.

# Conclusion.

Through our analysis of Airbnb data of a handful of cities, we came up with insights which help hosts to improve business. We trained different models on datasets and compared them to find suitable models. We observed a curvilinear relationship and accordingly tried finding a more generalized model.

Inferring patterns from the previous step we created a generalized whole data science process. Object oriented programming helped to develop modularized and configurable code. This code can be easily reused with other use cases such as Uber with a few changes. This contributed to developing an automated workflow using Apache Airflow. Data warehouses created through the option of storing data to HDFS can be leveraged by other data science and business processes.

Usefulness of this process is demonstrated through Streamlit user interface which provides meaningful insights and predictions to Airbnb host to make business decisions.

# Future Scope.

* Reviews can be used with Natural Language Processing to get further insights on hosts.
* Cloud storage can be used as a data store depending upon requirement.
* With more time we can explore every feature in detail.

# Bibliography.

**Github Project Link :** [**https://github.com/saturn279/Decision-Analytics-Platform-Airbnb**](https://github.com/saturn279/Decision-Analytics-Platform-Airbnb)

**Airbnb datasets:** [**http://insideairbnb.com/get-the-data/**](http://insideairbnb.com/get-the-data/)

**Apache Airflow -** [**https://airflow.apache.org/**](https://airflow.apache.org/)

**Apache Hadoop -** [**https://hadoop.apache.org/**](https://hadoop.apache.org/)

**Streamlit -** [**https://streamlit.io/**](https://streamlit.io/)

**Python -** [**https://www.python.org/downloads/release/python-3810/**](https://www.python.org/downloads/release/python-3810/)