

**SIX WEEKS SUMMER TRAINING**

**REPORT**

on

**HOTEL BOOKING DATA ANALYSIS**

Submitted by

**Garapati Mahendra Siva Sai Satya Venkat**

**Registration No: 12014364**

**Program Name: Introduction to Data Science**

Under Guidance Of

**Pranay Sharma**

**School of Computer Science & Engineering**

**Lovely Professional University, Phagwara**

(June-July, 2022)

**DECLARATION**

I hereby declare that I have completed my six weeks summer training at Allsoft Solutions

from 6th June, 2022 to 10th July, 2022 under the guidance of Pranay Sharma.

I declare that I have worked with full dedication during these six weeks of training and my learning outcomes fulfil the requirements for the award of degree of B. Tech in Computer Science & Engineering, Lovely Professional University, Phagwara.



Name: Garapati Mahendra Siva Sai Satya Venkat

Registration No: 12014364

Date: 16th July, 2022

**ACKNOWLEDGEMENT**

We are highly grateful to the **Dr. Ashok Kumar Mittal, Chancellor** and **Rashmi Mittal, Pro-Chancellor** for providing this opportunity to carry out six-week summer training and project work.The constant guidance and encouragement received from **Pranay Sharma, Allsoft Solutions** has been of great help incarrying our project work and is acknowledged with reverential thanks.We would also like to express a deep sense of gratitude to **HOD,** **……………**who guided during differentphases of project. Without the wise counsel and able guidance, it would have been quite difficult to complete the project in this manner.

**CONTENTS OF REPORT**

1. **Introduction**
2. **Technology Learnt**
3. **Reason for Choosing this Technology**
4. **Profile of Problem**
5. **Existing System**
6. **Problem Analysis**

* **Product Definition**
* **Feasibility Analysis**

1. **Software Requirement Analysis**
2. **Design**

* **Data flow diagrams & ER diagrams**
* **Flow Charts/Pseudo Code**

1. **Implementation**
2. **Learning Outcome from training/technology learnt**
3. **Gantt Chart**
4. **Project Legacy**

* **Technical and Managerial Lessons learnt.**

1. **Bibliography**

**INTRODUCTION**

The title of my project is “**Hotel Booking Data Analysis**”. I have completed my six-weeks summer training course in “Introduction to Data Science” from AllSoft Solutions. In this six-week training course I have revised the basics of python and python modules like NumPy and pandas. I have learned Matplotlib which helps me visualize data. I have learned basics of Data Science like cleaning, sorting and adjusting to our likes. I have also learned various Machine Learning models out of which I used one in my project. The things learnt in the six-week training course are the core to this project and along with these I also used seaborn. “Hotel Booking Data Analysis” is a very fitting project to test out and present my skills tat I have learned in six-week Summer Training course.

For this project I took the dataset from Kaggle.com. It is huge dataset that contains 32 columns and thousands of rows. This is a basic level project in which I answer some significant questions regarding the dataset. We know data is an important part of our lives these days. Every industry is being data dependent as they can learn so much from data and make necessary changes. This project represents how the Hotel Industry can make use of data to know better of there customers and adapt to them. The dataset gives us information about two hotels. We can compare between the two hotels and even tells the difference between their customers.

We use visualization so that everyone could understand and see what the numbers in the dataset actually mean. With visualization the hotel management can better understand the trends of tourism and the needs of customers. And the customers can get to know about the hotels and compare them to know which would be a better pick for them. The hotels can prepare for the spike of customers from just looking at the line plot graphs. They get to know the features they are missing out from others hotels and make changes. They can also know what customers like about their hotel and feature makes them standout from other hotels. Better analysing the data can help the hotels very much improve themselves and have that edge over their competitors and better adapt to the trends and also make predictions using data.

Using Machine Learning models, we can make predictions that help the hotels. Such prediction is made in this project. This prediction lets us know whether a customer cancels his booking or not. We used a Supervised Learning Algorithm called “**Decision Tree**” which predict values of responses by learning decision rules derived from features.

Using the existing datasets extracting the required features from the dataset which may effect the cancellation of the booking, clean the data. And then split the data into training and testing sets. We train our model using training set and then the model makes it predictions. We compare the predictions with the already existing testing set to get the accuracy and precision of the model. An accuracy above 90% or 0.9 out of 1 is considered good. If we don’t achieve this using this algorithm we will have try with another algorithm.

**TECHNOLGY LEARNT**

|  |  |
| --- | --- |
| **Module** | **Contents** |
| **I** | BASIC PYTHON:  ● History of Python  ● Features of python  ● Installation of jupyter notebook  ● Python operators  ● Python Datatypes and its inbuild functions  ● Conditional statement  ● Loops |
| **II** | SCIENTIFIC PYTHON:  ● Pandas  ● Series and its inbuild functions  ● Dataframe and its inbuild functions  ● NumPy  ● 1- D array  ● 2- D array  ● 3- D array |
| **III** | DATA VISULIZATION  ● Using Matplotlib  ● Line Graph  ● Point Graph  ● Bar Graph  ● Pie Chart  ● Histogram |
| **IV** | INTRODUCTION TO MACHINE LEARNING:  ● Application of Machine Learning  ● Supervised vs Unsupervised Learning  ● Python libraries suitable for Machine Learning |
| **V** | DATA PRE-PROCESSING AND DATA  ● Identifying and handling the missing values  ● Encoding the categorical data  ● Normalization  ● Standardization |
| **VI** | SUPERVISED LEARNING REGRESSION  ● Linear Regression  ● Decision Tree Regressor  ● Random Forest Regressor  ● SVR: Support Vector Regressor  ● Model evaluation methods: RMSE value |
| **VII** | SUPERVISED LEARNING CLASSIFICATION:  ● Logistic Regression  ● Decision tree classifier  ● Random Forest classifier  ● SVM: SVR  ● Naïve bayes  ● Gaussian NB  ● Multinomial NB  ● Bernoulli’s NB  ● Model evaluation methods: accuracy score, Precision, recall, F1 -score. |
| **VIII** | Unsupervised Learning:  ● K-means Clustering  ● Hierarchical Clustering  ● PCA |
| **IX** | Reinforcement Learning:  ● Introduction  ● Application Areas RL  ● Basic understanding of Q learning. |
| **X** | Advance demanded projects:  ● Uber-trip NYC data analysis  ● Hotel booking data analysis  ● Covid-19 data analysis  ● Amazon customer data analysis |

**REASON FOR CHOOSING THIS TECHNOLOGY**

Data is considered as the oil of 21st century. In the previous centuries oil is considered very important that the world would stop working if there is no oil and it still does. But as we go further into this century data will replace oil. With the rise of technology and the ease of accessing it every one has mobiles and computers. There is a lot of data being generated which can be used maximize the efficiency in the world. Data Science is the principal study of data. And Machine Learning makes use of data science, they both are correlated.

The ever-evolving nature of data science is part of what makes it such a compelling industry to build your career. You’ll gain a wide array of new skills that will allow you to leverage data to aid companies with their business strategies, and explore exciting new fields developing from within data science—fields like artificial intelligence, machine learning, big data, and more.

You’re probably already familiar with artificial intelligence (AI), at least as a concept. Designed to simulate human intelligence in machines, AI uses multiple algorithms to perform autonomous actions and to understand relationships between different types and different pieces of data. Machine learning is an offshoot discipline of AI focusing on developing machines that will learn from past data automatically without explicit programming.

AI and machine learning are tools used to aid in more efficiently and accurately analysing the data used in data science. As data science continues to evolve, the data that’s available to us becomes more and more complex and it can be a struggle to decipher all on our own. Machine learning lets machines handle a lot of that heavy lifting for us and is used in things we use every day like Google’s search engine and Spotify’s song-choosing algorithms. Though still a developing technology, machine learning is starting to play a role in advances as diverse and influential as self-driving cars and cancer detection. A seemingly elusive concept to some in name (machine learning), it’s actually a discipline that effects individuals on a very practical, daily, and human level.

As data science grows, so will the opportunities it presents. Ten years ago, it would have been difficult to imagine where the fields of data science, AI, and machine learning would take us. It’s equally tough to predict where we’ll be ten years from now, though almost inarguable that it will be a path of increase and growth. If you become a data scientist, those answers will be, in part, up to you and your career contribution influencing in what direction you take the industry in next.

**Python is the programming language of choice for data scientists**. Although it wasn't the first primary programming language, its popularity has grown throughout the years. In 2016, it overtook R on Kaggle, the premier platform for data science competitions.

**NumPy** is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

**Pandas** is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.

**Matplotlib** is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK.

A **decision tree** is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

**PROFILE OF PROBLEM**

This data set contains booking information for a city hotel and a resort hotel and includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things. All personally identifying information has from the data.

We will perform exploratory data analysis with python to get insight from the data.

**We will try to answer the following Questions**

1. How Many Bookings Were Cancelled?
2. What is the booking ratio between Resort Hotel and City Hotel?
3. What is the percentage of booking for each year?
4. Which is the busiest month for hotels?
5. From which country most guests come?
6. How Long People Stay in the hotel?
7. Which was the most booked accommodation type (Single, Couple, Family)?

After that, we will make the predictive model to make predictions in the future whether the booking will be cancelled or not.

**EXISTING SYSTEM**

We have two datasets with hotel demand data. One of the hotels (H1) is a resort hotel and the other is a city hotel (H2). Both datasets share the same structure, with 31 variables describing the 40,060 observations of H1 and 79,330 observations of H2. Each observation represents a hotel booking. Both datasets comprehend bookings due to arrive between the 1st of July of 2015 and the 31st of August 2017, including bookings that effectively arrived and bookings that were cancelled. Since this is hotel real data, all data elements pertaining hotel or costumer identification were deleted. Due to the scarcity of real business data for scientific and educational purposes, these datasets can have an important role for research and education in revenue management, machine learning, or data mining, as well as in other fields.

| **Variable** | **Type** | **Description** |
| --- | --- | --- |
| ***ADR*** | Numeric | Average Daily Rate as defined by [[5]](https://www.sciencedirect.com/science/article/pii/S2352340918315191" \l "bib5) |
| ***Adults*** | Integer | Number of adults |
| ***Agent*** | Categorical | ID of the travel agency that made the booking |
| ***ArrivalDateDayOfMonth*** | Integer | Day of the month of the arrival date |
| ***ArrivalDateMonth*** | Categorical | Month of arrival date with 12 categories: “January” to “December” |
| ***ArrivalDateWeekNumber*** | Integer | Week number of the arrival date |
| ***ArrivalDateYear*** | Integer | Year of arrival date |
| ***AssignedRoomType*** | Categorical | Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due to hotel operation reasons (e.g. overbooking) or by customer request. Code is presented instead of designation for anonymity reasons |
| ***Babies*** | Integer | Number of babies |
| ***BookingChanges*** | Integer | Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation |
| ***Children*** | Integer | Number of children |
| ***Company*** | Categorical | ID of the company/entity that made the booking or responsible for paying the booking. ID is presented instead of designation for anonymity reasons |
| ***Country*** | Categorical | Country of origin. Categories are represented in the ISO 3155–3:2013 format [[6]](https://www.sciencedirect.com/science/article/pii/S2352340918315191" \l "bib6) |
|  |  |  |
| ***CustomerType*** | Categorical | Type of booking, assuming one of four categories: |
| Contract - when the booking has an allotment or other type of contract associated to it; |
| Group – when the booking is associated to a group; |
| Transient – when the booking is not part of a group or contract, and is not associated to other transient booking; |
| Transient-party – when the booking is transient, but is associated to at least other transient booking |
| ***DaysInWaitingList*** | Integer | Number of days the booking was in the waiting list before it was confirmed to the customer |
|  |  |  |
| ***DepositType*** | Categorical | Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories: |
| No Deposit – no deposit was made; |
|
|
| Non-Refund – a deposit was made in the value of the total stay cost; |
|
| Refundable – a deposit was made with a value under the total cost of stay. |
| ***DistributionChannel*** | Categorical | Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators” |
| ***IsCanceled*** | Categorical | Value indicating if the booking was cancelled (1) or not (0) |
| ***IsRepeatedGuest*** | Categorical | Value indicating if the booking name was from a repeated guest (1) or not (0) |
| ***LeadTime*** | Integer | Number of days that elapsed between the entering date of the booking into the PMS and the arrival date |
| ***MarketSegment*** | Categorical | Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators” |
|  |  |  |
| ***Meal*** | Categorical | Type of meal booked. Categories are presented in standard hospitality meal packages: |
| Undefined/SC – no meal package; |
| BB – Bed & Breakfast; |
| HB – Half board (breakfast and one other meal – usually dinner); |
| FB – Full board (breakfast, lunch and dinner) |
| ***PreviousBookingsNotCanceled*** | Integer | Number of previous bookings not cancelled by the customer prior to the current booking |
| ***PreviousCancellations*** | Integer | Number of previous bookings that were cancelled by the customer prior to the current booking |
| ***RequiredCardParkingSpaces*** | Integer | Number of car parking spaces required by the customer |
|  |  |  |
| ***ReservationStatus*** | Categorical | Reservation last status, assuming one of three categories: |
| Cancelled – booking was cancelled by the customer; |
| Check-Out – customer has checked in but already departed; |
| No-Show – customer did not check-in and did inform the hotel of the reason why |
| ***ReservationStatusDate*** | Date | Date at which the last status was set. This variable can be used in conjunction with the *ReservationStatus* to understand when was the booking cancelled or when did the customer checked-out of the hotel |
| ***ReservedRoomType*** | Categorical | Code of room type reserved. Code is presented instead of designation for anonymity reasons |
| ***StaysInWeekendNights*** | Integer | Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel |
| ***StaysInWeekNights*** | Integer | Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel |
| ***TotalOfSpecialRequests*** | Integer | Number of special requests made by the customer (e.g., twin bed or high floor) |

.

**PROBLEM ANALYSIS**

**Product Definition**

* Perform the Feature Engineering to make new features
* Perform the Feature Selection to select only relevant features
* Transform the Data (Categorial to Numerical)
* Split the data (Train Test Split)
* Model the data (Fit the Data)
* And finally, Evaluate our model

Our product will be able to predict the booking cancellations.

**Feasibility Analysis**

Software Requirements:

1. Python & Jupyter Notebook
2. NumPy, Pandas, Matplotlib, Seaborn, SKlearn
3. Windows 7 and above

Hardware Requirements;

1. Intel i5 and above
2. 4gb RAM and above
3. GPU not necessary

If the above requirements are met the project is feasible.

**SOFTWARE REQUIREMENT ANALYSIS**

**PURPOSE:**

The main purpose of this project is to help hotels is to let them know the trends of the hotel industry using the visualisation. To help hotels know more about their customers, know what features their customers like about them and why others are choosing other hotels. To help them know what other hotels are doing better to gain customer and what they can learn from them. Know what gives them the edge over others and they lack that’s taking away customers from them so they make the necessary changes. And with the machine learning model also help them to make prediction about a booking cancellation and what is the main factor that is leading to the cancellation.

**DOCUMENT** **CONVENTIONS:**

This SRS archive keeps up diverse sorts of fonts, sizes, and faces. This record additionally specified Times New Roman: font sort is Bold and Normal.

**PROJECT** **SCOPE:**

In this project, we will try to predict the possibility of a booking for a hotel based on different factors and also try to predict if they need special requests based on different features. The data set contains booking information for a city hotel and a resort hotel, and includes information such as when the booking was made, the number of adults, children, and/or babies, and the number of available parking spaces, among other things. From it, we can understand the customer’s’ behaviour and it might help us make better decisions.

**OBJECTIVES OF THE PROJECT:**

We will calculate and visualize the following

1. How Many Bookings Were Cancelled?
2. What is the booking ratio between Resort Hotel and City Hotel?
3. What is the percentage of booking for each year?
4. Which is the busiest month for hotels?
5. From which country most guests come?
6. How Long People Stay in the hotel?
7. Which was the most booked accommodation type (Single, Couple, Family)?

After that, we will make the predictive model to make predictions in the future whether the booking will be cancelled or not using Decision Tree.

**OVERALL** **DESCRIPTION:**

We have two datasets with hotel demand data. One of the hotels (H1) is a resort hotel and the other is a city hotel (H2). Both datasets share the same structure, with 31 variables describing the 40,060 observations of H1 and 79,330 observations of H2. Each observation represents a hotel booking. Both datasets comprehend bookings due to arrive between the 1st of July of 2015 and the 31st of August 2017, including bookings that effectively arrived and bookings that were cancelled. Since this is hotel real data, all data elements pertaining hotel or costumer identification were deleted. Due to the scarcity of real business data for scientific and educational purposes, these datasets can have an important role for research and education in revenue management, machine learning, or data mining, as well as in other fields.

**PREDICTING BOOKING CANCELLATION:**

Cancellation does not depend on any one feature. It varies from costumer to costumer. From the dataset we drop the “is\_cancelled” column because it indicates the cancellation and it cannot be used in training. We take the remaining features and divide the dataset into two sets. Training set and Testing Test. Training test is 75% of the original dataset and Testing set is the remaining 25% of the original dataset. Using **sklearn** library we import Decision Tree algorithm and train that algorithm. After execution we can make prediction to see whether its correct. We can execute **confusion matrix** to check the correct and wrong predictions. And finally, we get the accuracy.

**PRINCIPLES OF SYSTEM ANALYSIS**

1. Understand the problem before you begin to create the analysis model.

2. Develop prototypes that enable a user to understand how human machine interaction will occur.

3. Record the origin of and the reason for every requirement.

4. Use multiple views of requirements like building data, function and behavioral models.

5. Work to eliminate ambiguity.

**DESIGN**

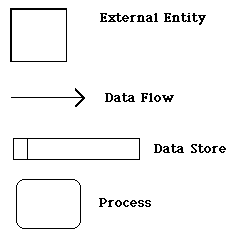
**DATA FLOW DIAGRAMS:**

A data flow diagram (DFD) illustrates how data is processed by a system in terms of inputs and outputs. As its name indicates its focus is on the flow of information, where data comes from, where it goes and how it gets stored.

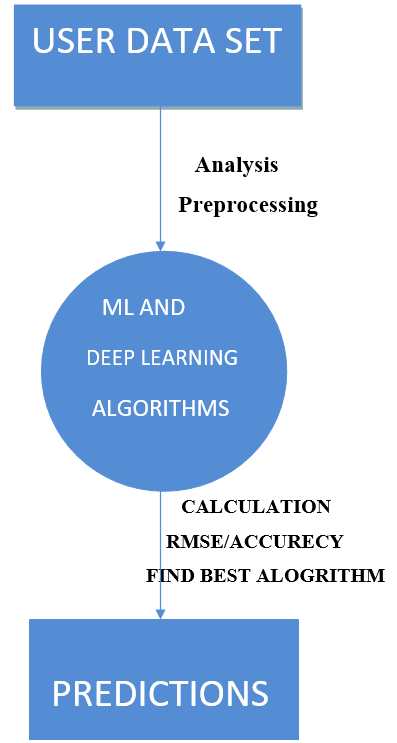
A context diagram is a top level (also known as "Level 0") data flow diagram. It only contains one process node ("Process 0") that generalizes the function of the entire system in relationship to external entities.

The context diagram is the highest level in a data flow diagram and contains only one process, representing the entire system. The process is given the numb external entities are shown on the context diagram, as well as major data flow to and from them. The diagram does not contain any data stores.

The following notations are used for any dfd:



A data flow diagram (DFD) illustrates how data is processed by a system in terms of inputs and outputs. The basic concepts related to smart systems and how machine learning techniques could add smart capabilities to many kinds of systems in almost any domain that you can imagine. Among other things, we learned that a typical workflow for a Machine Learning Project usually looks like the one shown in the image below:

****

**ER DIAGRAMS:**

An Entity Relationship (ER) Diagram is a type of flowchart that illustrates how “entities” such as people, objects or concepts relate to each other within a system. ER Diagrams are most often used to design or debug relational databases in the fields of software engineering, business information systems, education and research. Also known as ERDs or ER Models, they use a defined set of symbols such as rectangles, diamonds, ovals and connecting lines to depict the interconnectedness of entities, relationships and their attributes. They mirror grammatical structure, with entities as nouns and relationships as verbs.

ER diagrams are related to data structure diagrams (DSDs), which focus on the relationships of elements within entities instead of relationships between entities themselves. ER diagrams also are often used in conjunction with data flow diagrams (DFDs), which map out the flow of information for processes or systems.

**Uses** **of** **entity** **relationship** **diagrams:**

* **Database** **design:** ER diagrams are used to model and design relational databases, in terms of logic and business rules (in a logical data model) and in terms of the specific technology to be implemented (in a physical data model.) In software engineering, an ER diagram is often an initial step in determining requirements for an information systems project. It’s also later used to model a particular database or databases. A relational database has an equivalent relational table and can potentially be expressed that way as needed.
* **Database** **troubleshooting:** ER diagrams are used to analyse existing databases to find and resolve problems in logic or deployment. Drawing the diagram should reveal where it’s going wrong.
* **Business** **information** **systems:** The diagrams are used to design or analyse relational databases used in business processes. Any business process that uses fielded data involving entities, actions and interplay can potentially benefit from a relational database. It can streamline processes, uncover information more easily and improve results.
* **Business** **process** **re-engineering** **(BPR):** ER diagrams help in analysing databases used in business process re-engineering and in modelling a new database setup.
* **Education:** Databases are today’s method of storing relational information for educational purposes and later retrieval, so ER Diagrams can be valuable in planning those data structures.
* **Research:** Since so much research focuses on structured data, ER diagrams can play a key role in setting up useful databases to analyse the data.

**The** **components** **and** **features** **of** **an** **ER** **diagram**

ER Diagrams are composed of entities, relationships and attributes. They also depict cardinality, which defines relationships in terms of numbers. Here’s a glossary:

**Entity**

A definable thing—such as a person, object, concept or event—that can have data stored about it. Think of entities as nouns. Examples: a customer, student, car or product. Typically shown as a rectangle.

ENTITY

**Entity** **type:** A group of definable things, such as students or athletes, whereas the entity would be the specific student or athlete. Other examples: customers, cars or products.

**Entity** **set:** Same as an entity type, but defined at a particular point in time, such as students enrolled in a class on the first day. Other examples: Customers who purchased last month, cars currently registered in Florida. A related term is instance, in which the specific person or car would be an instance of the entity set.

**Entity** **categories:** Entities are categorized as strong, weak or associative. A **strong** **entity** can be defined solely by its own attributes, while a **weak** **entity** cannot. An associative entity associates entities (or elements) within an entity set.

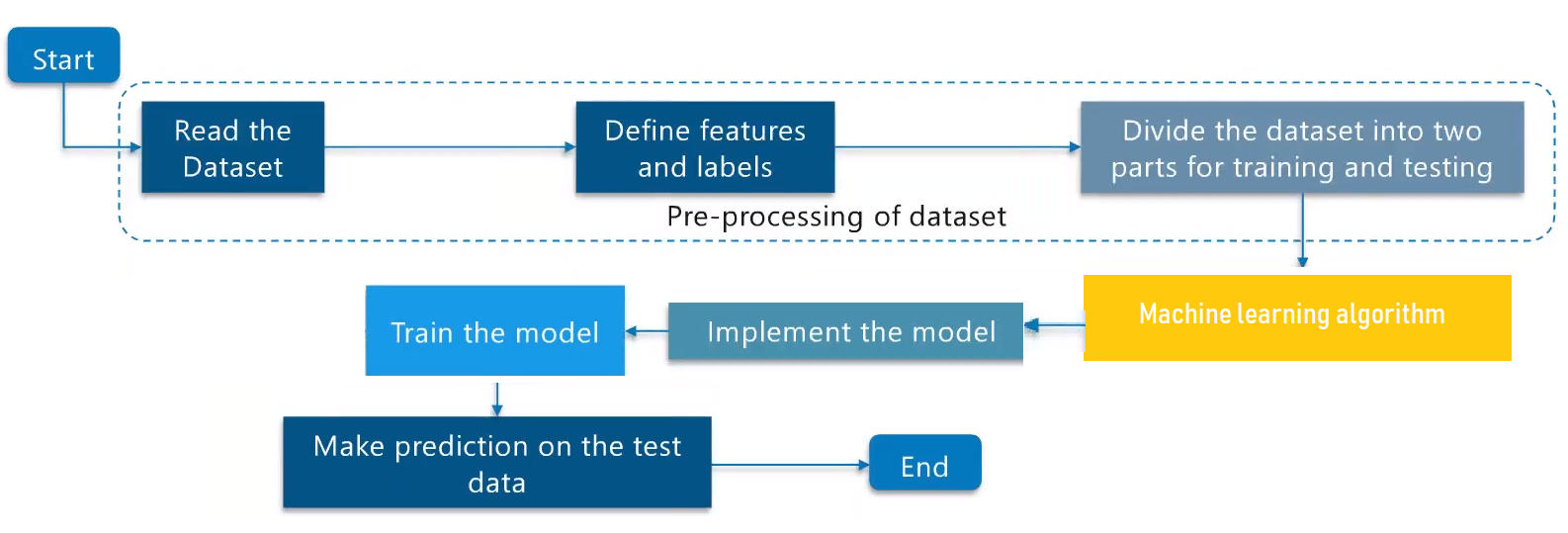
**Entity** **keys:** Refers to an attribute that uniquely defines an entity in an entity set. Entity keys can be super, candidate or primary.

**Super** **key:** A set of attributes (one or more) that together define an entity in an entity set.

**Candidate** **key:** A minimal super key, meaning it has the least possible number of attributes to still be a super key. An entity set may have more than one candidate key. **Primary** **key:** A candidate key chosen by the database designer to uniquely identify the entity set. **Foreign** **key:** Identifies the relationship between entities.

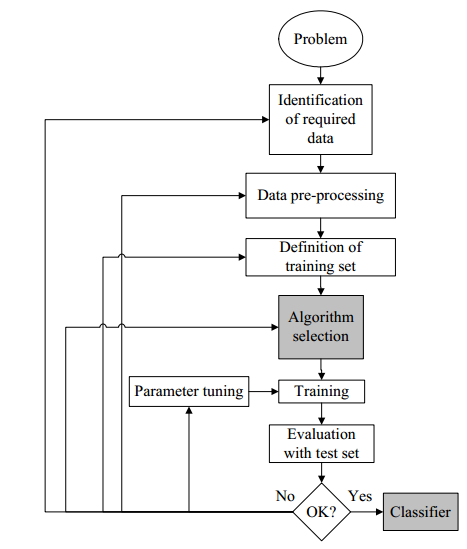
**Relationship:**

How entities act upon each other or are associated with each other. Think of relationships as verbs. For example, the named student might register for a course. The two entities would be the student and the course, and the relationship depicted is the act of enrolling, connecting the two entities in that way. Relationships are typically shown as diamonds or labels directly on the connecting lines.

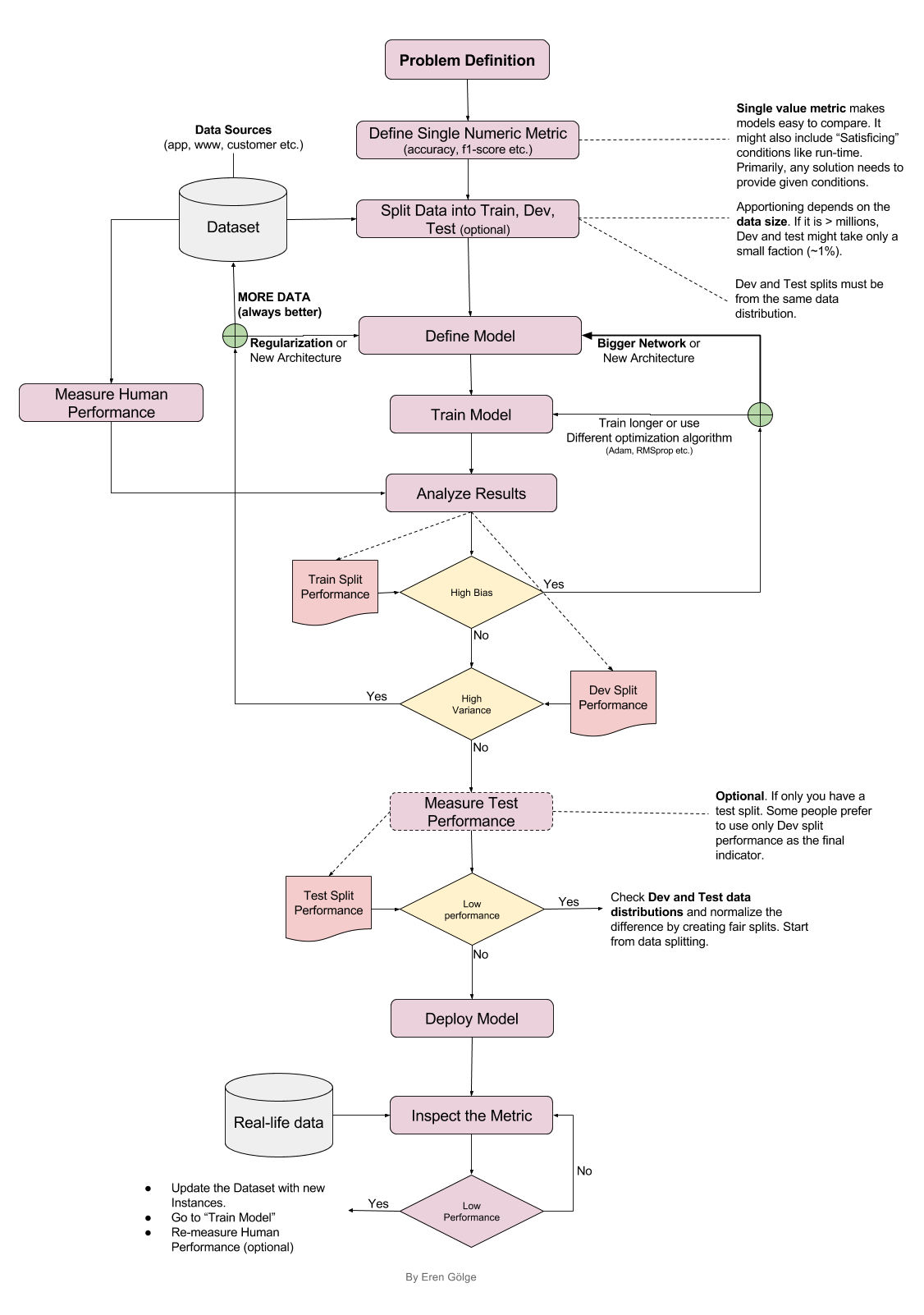


**FLOWCHART**

Flowchart Python is essentially the Python programming language in visual form. You write a program by setting up a flowchart. When you run the flowchart, the software compiles to python byte-code, so that you can easily import modules you write in Flowchart Python into standard Python programs. Flowchart Python does make a few additions to standard Python, which will be outlined on the home page eventually.



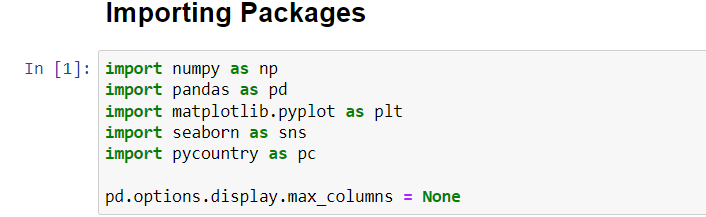
**Fig: FLOW DIAGRAM OF MACHINE LEARNING**

****

**IMPLEMENTATION**

**CODING:**

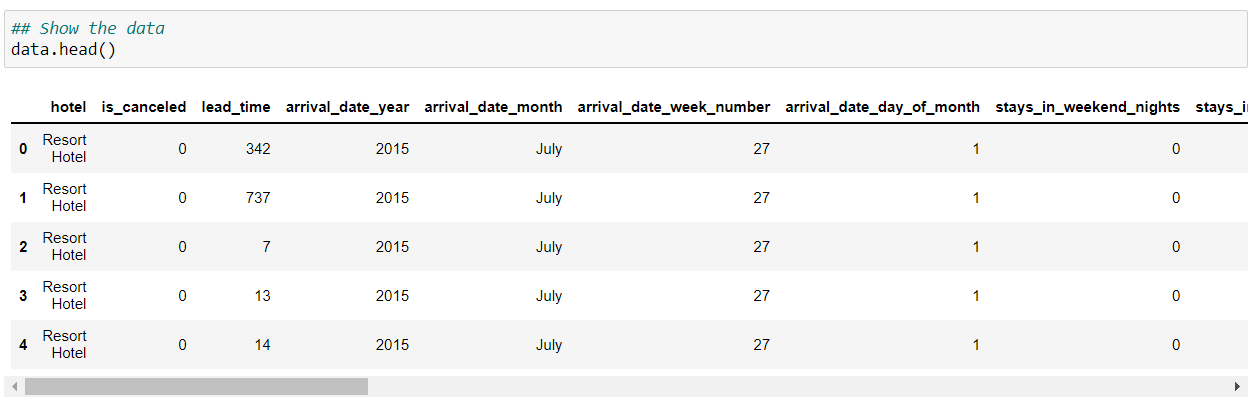
**#** First, we import the libraries required

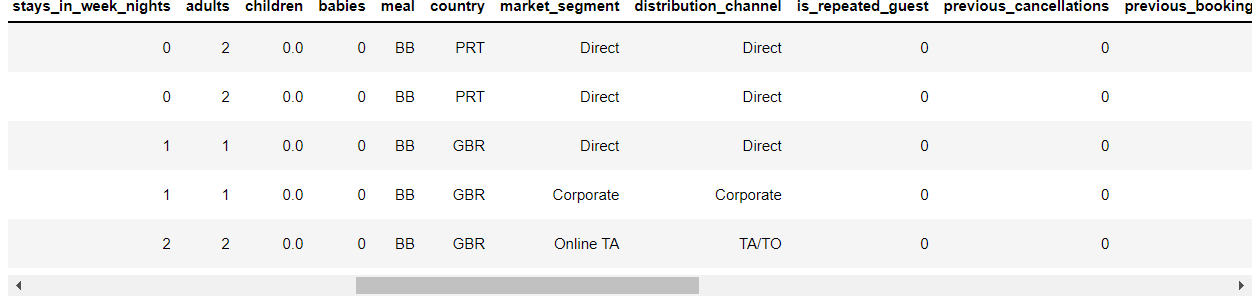
****

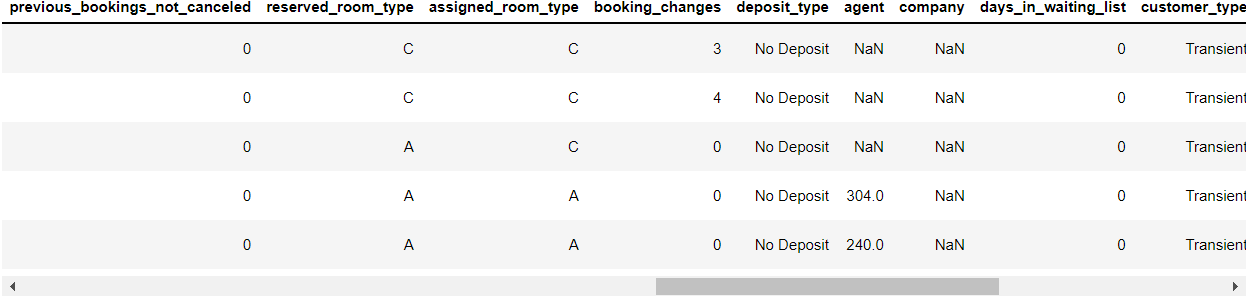
**#** Now we import the dataset

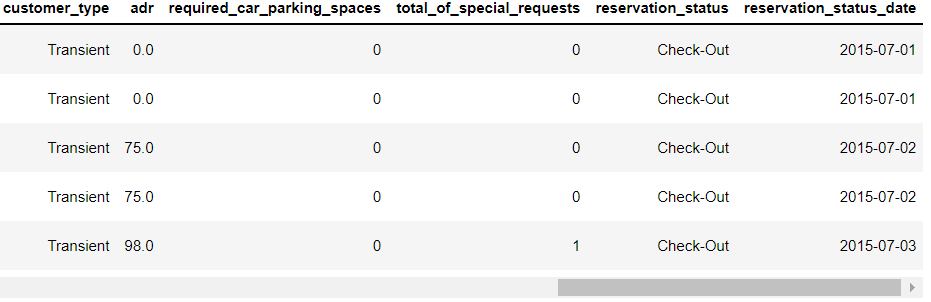


# Let us see the dataset



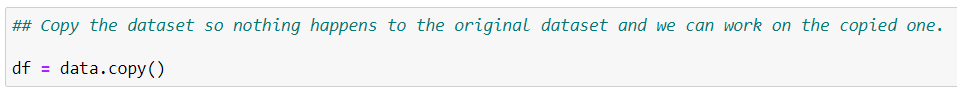




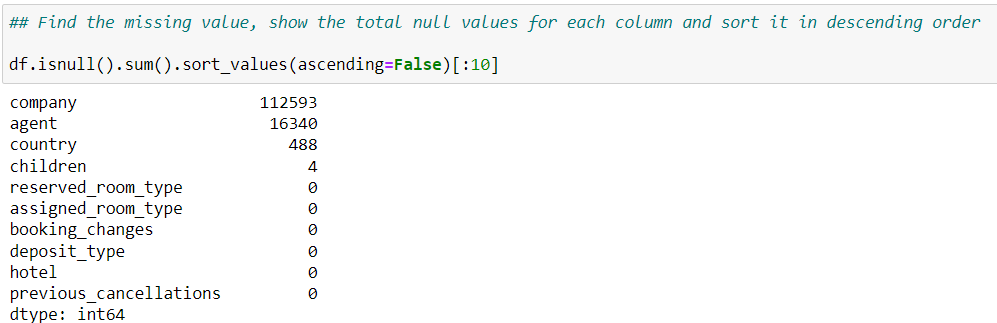


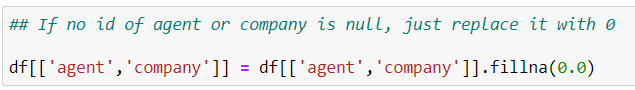
# Now Data Pre-Processing

# We make a copy of the original dataset and work on it so we can make changes to the copied one. If we make any mistakes the original dataset doesn’t get effected.

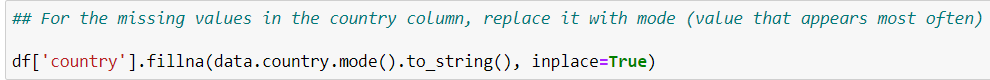


# Now we find the missing values

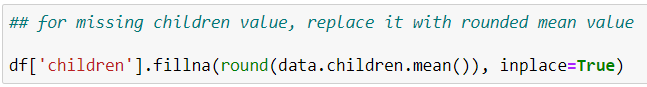


# Now replace null values with 0. 

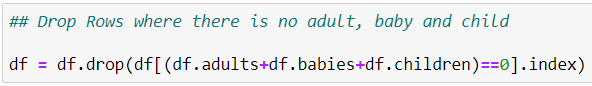
# In country column we replace the missing values with mode of the values.



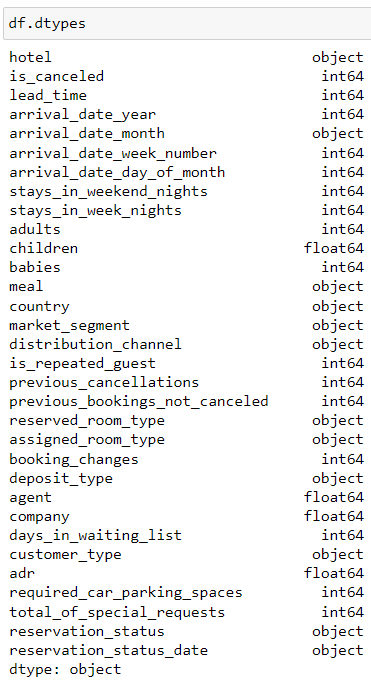
# In children column we replace missing values with rounded mean value.



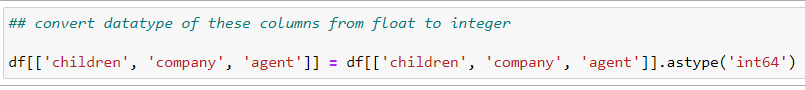
# If there are no adult, children and babies’ values in row that means there are no one in that row. So, we simply drop that row from dataset.



# Let us check the datatypes of the values

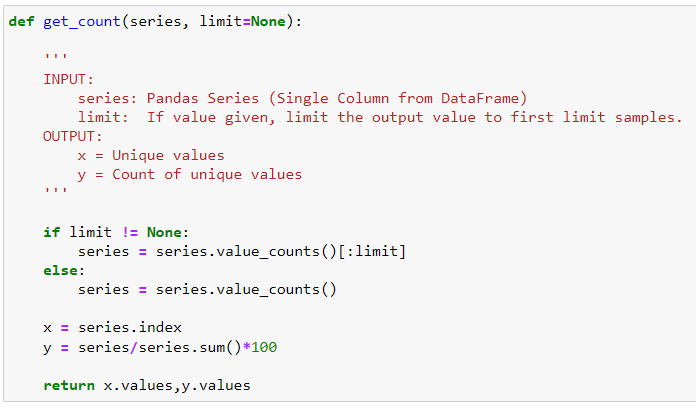


# Now we convert the columns with values of float datatype to integer datatype.



#DATA ANALYSIS

# Let’s write the function to get the percentage of different values.



# This function takes a series or data frame column and returns the two arrays

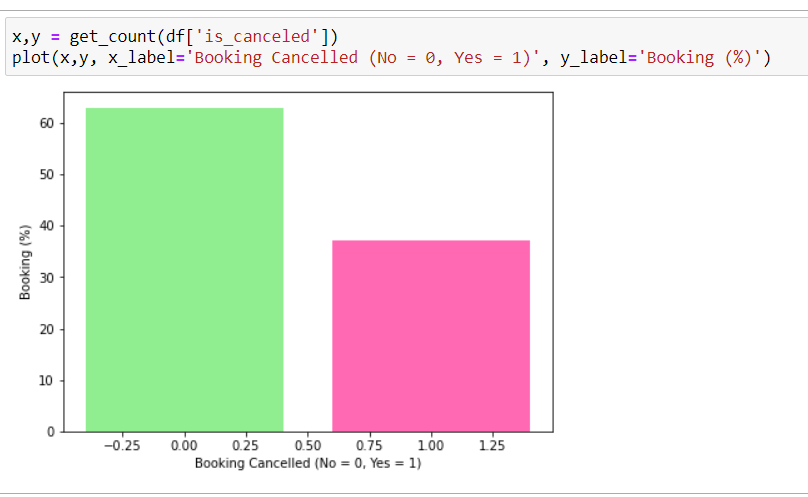
x is our unique values  
 y is the percentage value of each unique value

# I will write another function to plot the diagram. The good thing about writing function is that we can reuse the code again and again.

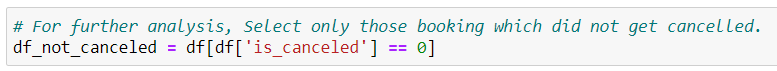
# This function takes two arrays, x, and y and displays the required diagram. The default plot type is a bar plot, but it can also plot the line plot. Optional arguments can be given to display title and labels.



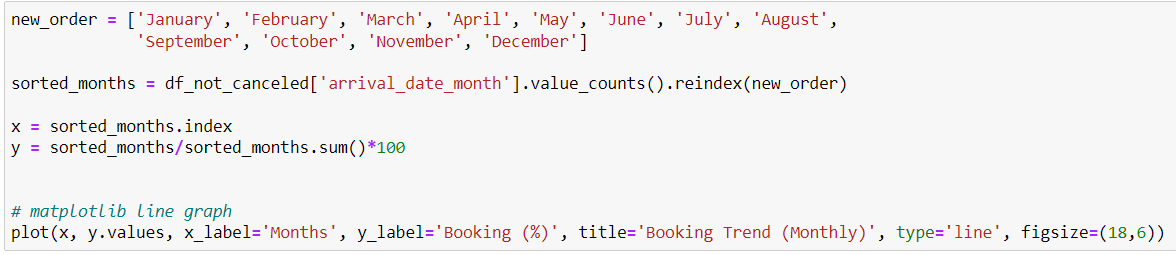
#Let us find out **1. How many Bookings got cancelled?**

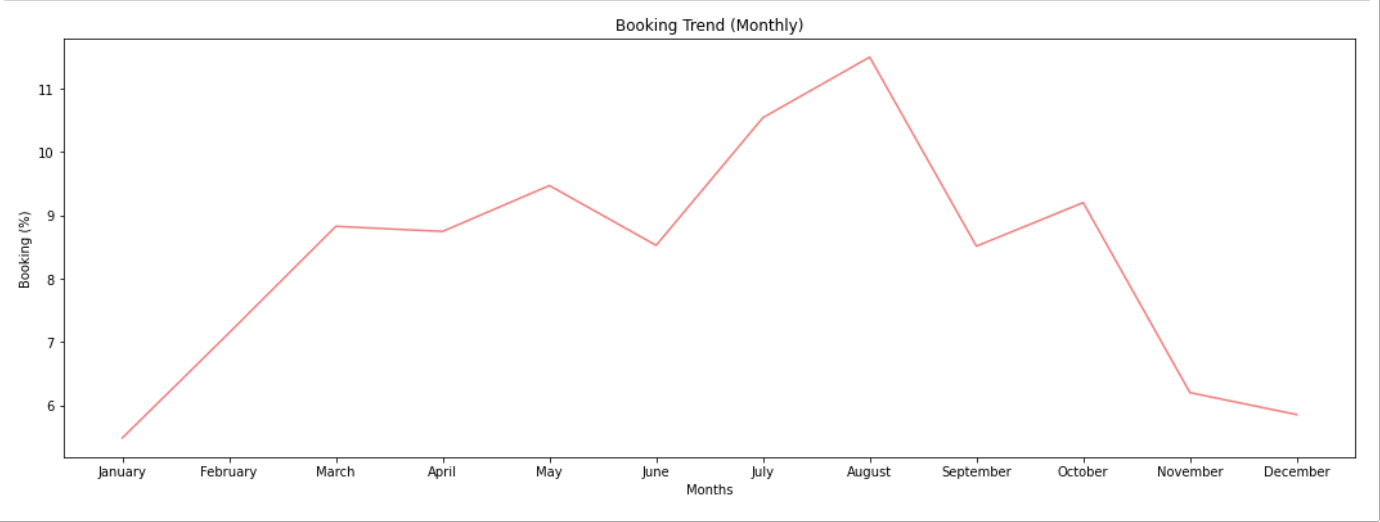
****

**#** Now that we found out cancelled booking these rows won’t be of use to us in further analysis as they won’t have other values required other questions. So we only take rows which booking is not cancelled.



# Now lets us find **2. Which month is most busy for both the hotels combined?**

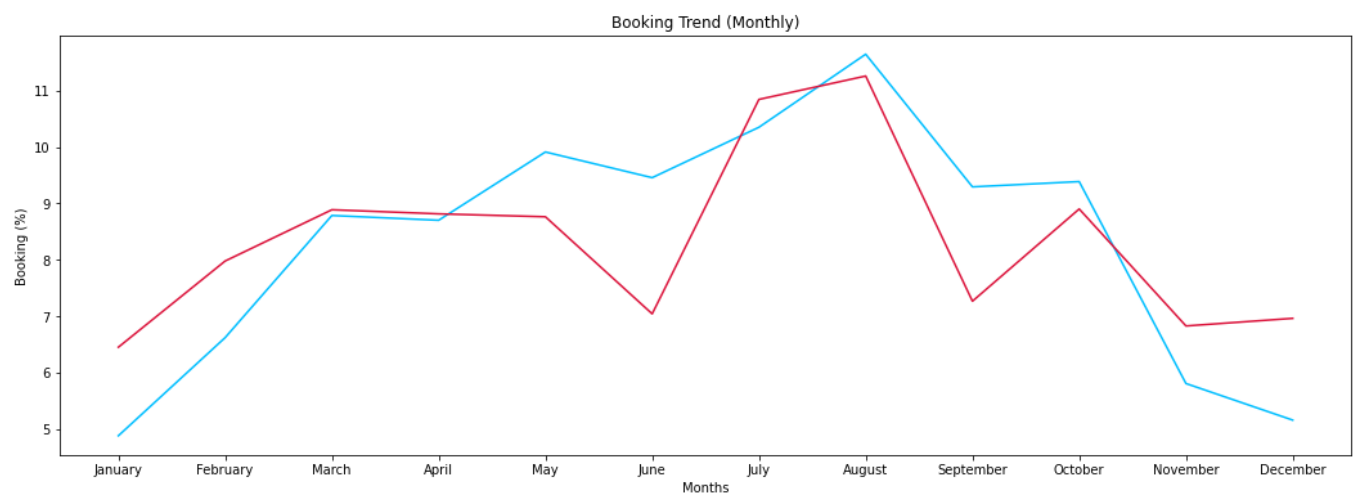




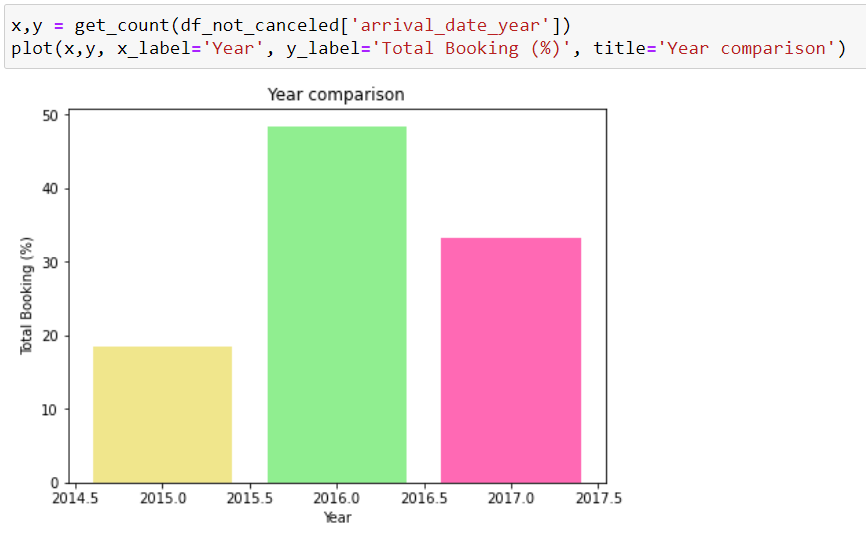
# Now let us find out **2. Which is most busy month for the two hotels separately?**



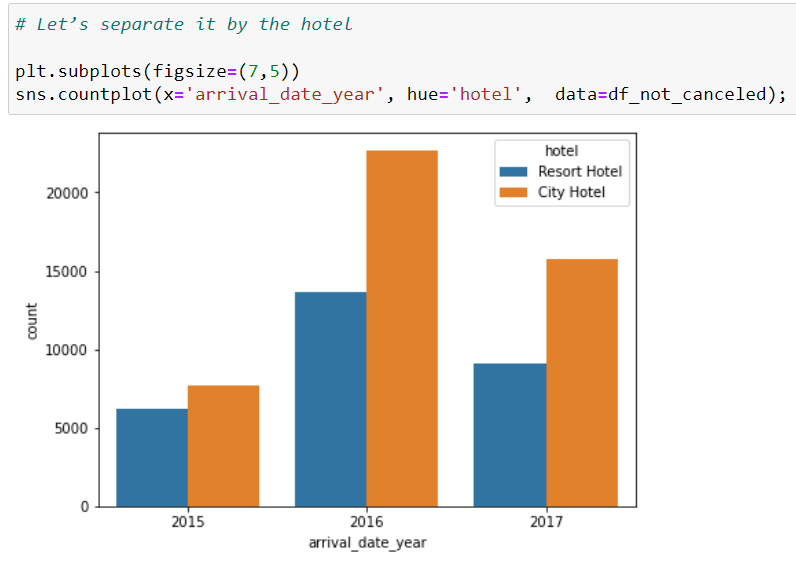




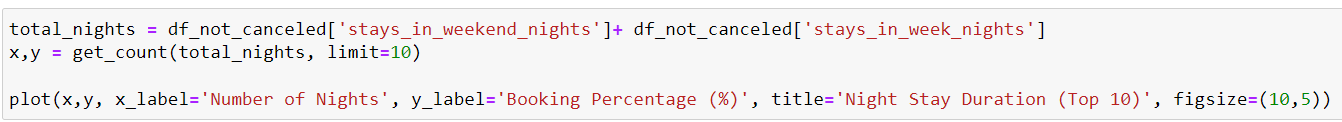
# Now let us find out **3. What is percentage of booking each year?** (combined)

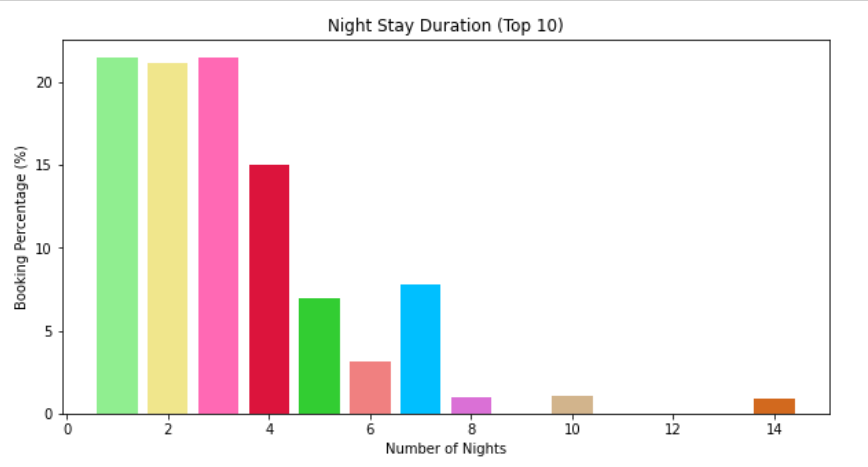


# Now let us find **3. What is percentage of booking each year?** (separately)

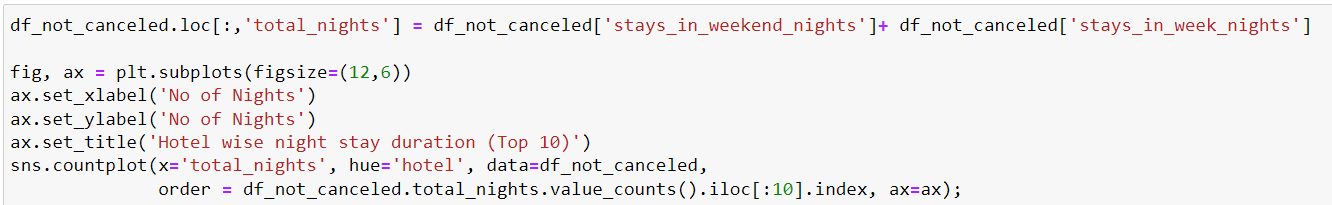


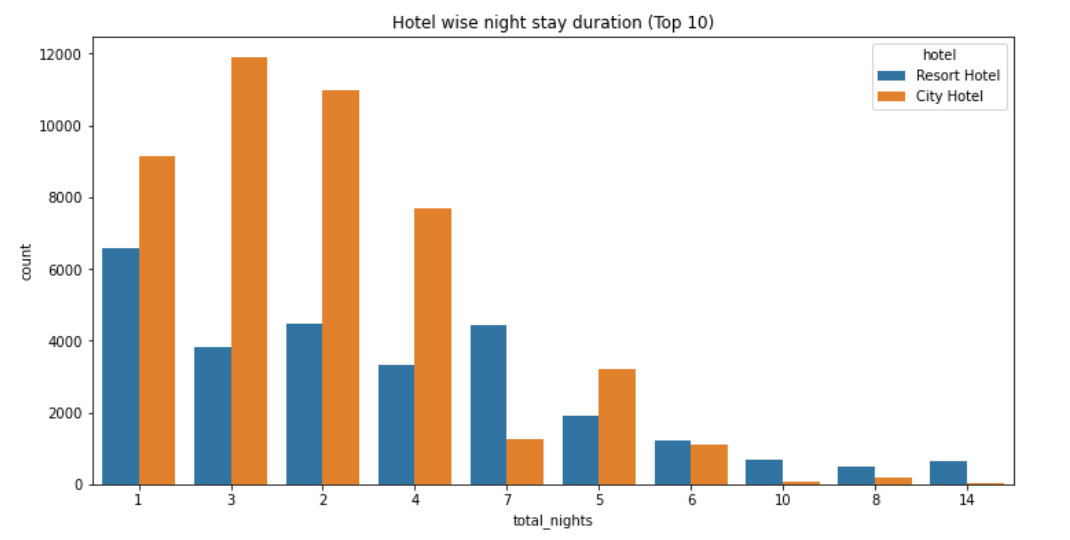
# Now let us find out **4. How long people stay in hotels?** (combined)



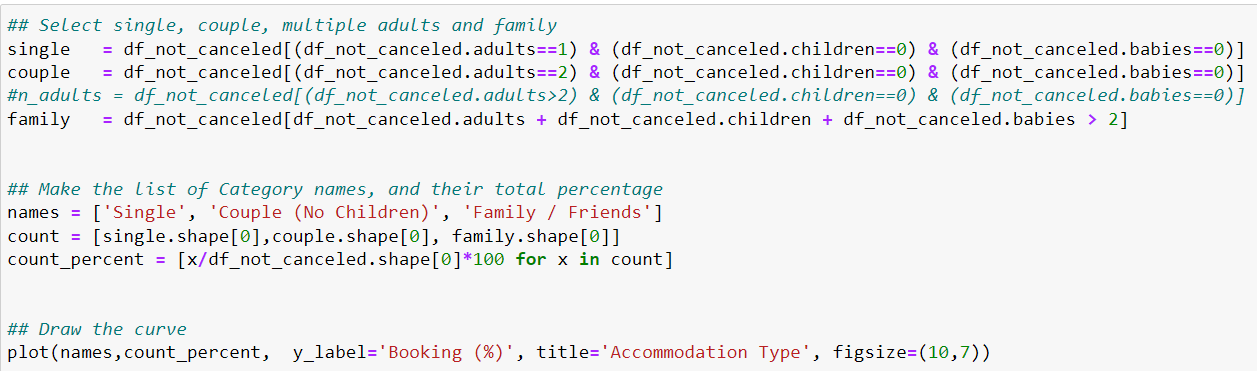


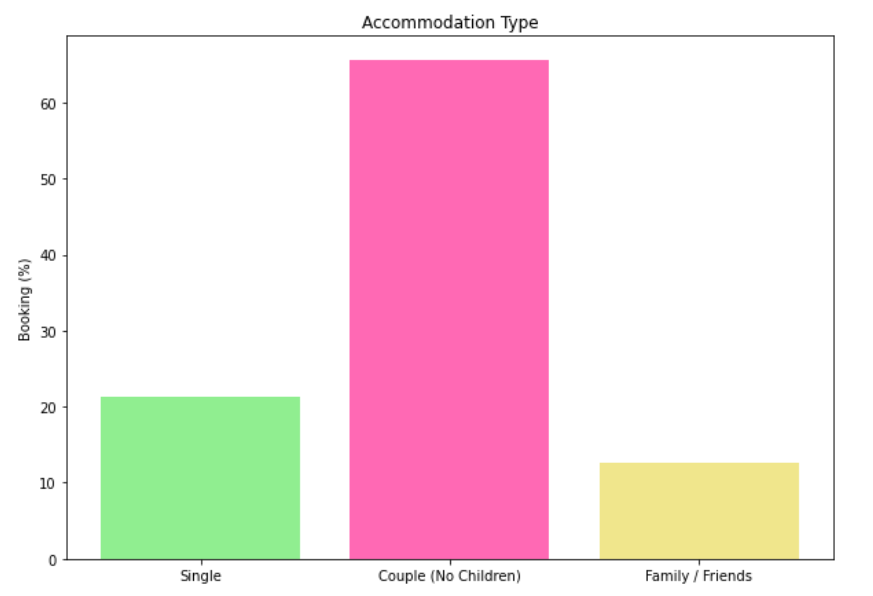
# Now let us find **4. How long people stay in each hotel?**





### # Now let us find **5. Which was the most booked accommodation type (Single, Couple, Family)?**

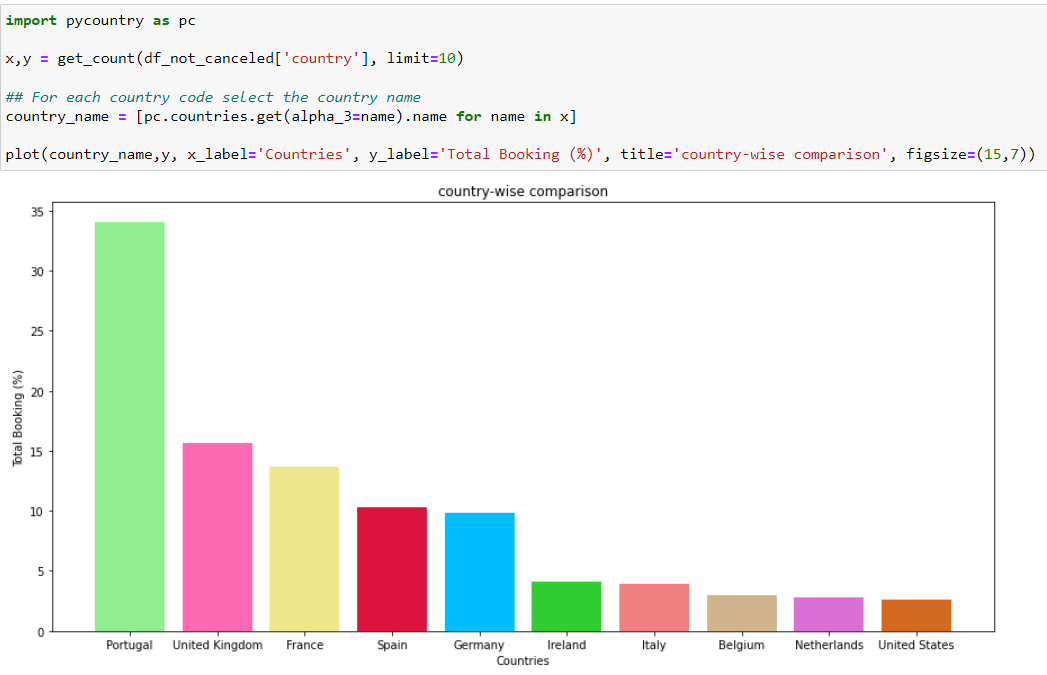




### # Now let us find **6. What is the booking ratio between Resort Hotel and City Hotel?**

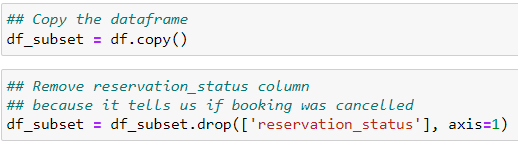


### # Now let us find **7. From which country most guest come?**



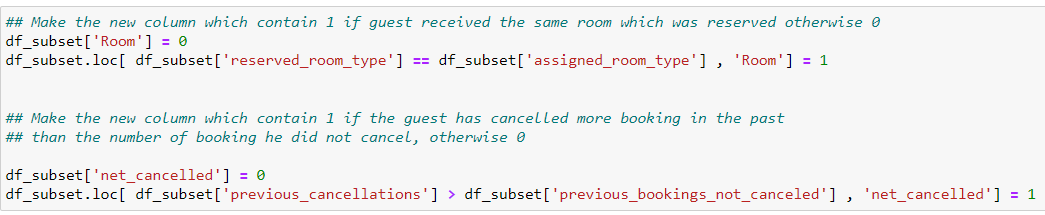
# Feature Selection and Engineering

# Now let us make copy of dataset and remove “reservation\_status” as it tells if the booking is cancelled or not.



# Now we make a column room which has 1 if the guest got same room he reserved and in other cases 0.

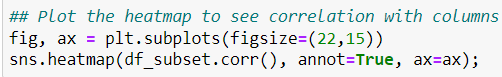
# Let us also make a new column which contain 1 if the guest has cancelled more booking in the past than the number of bookings, he did not cancel, otherwise 0.

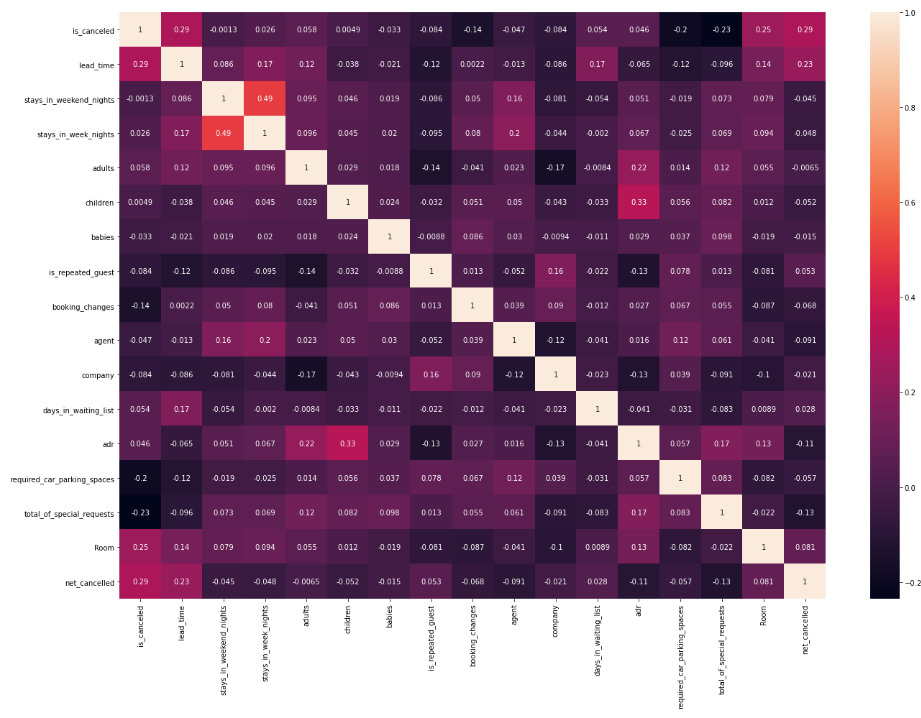


# Now we remove features that won’t affect cancellations.



# Now let us plot heat map to see correlation of columns



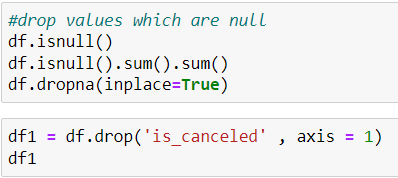


**# MODELING**

### **# 1. Converting Categorical variables to Numerical**

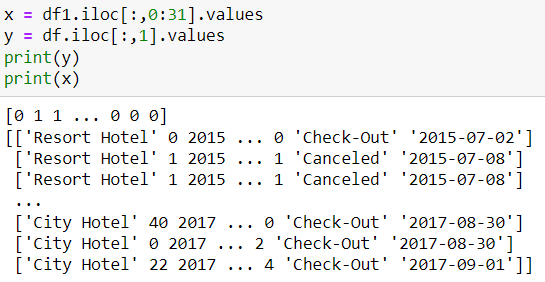
# Drop values that are null

# Make another copy of dataset and drop “is\_cancelled” column

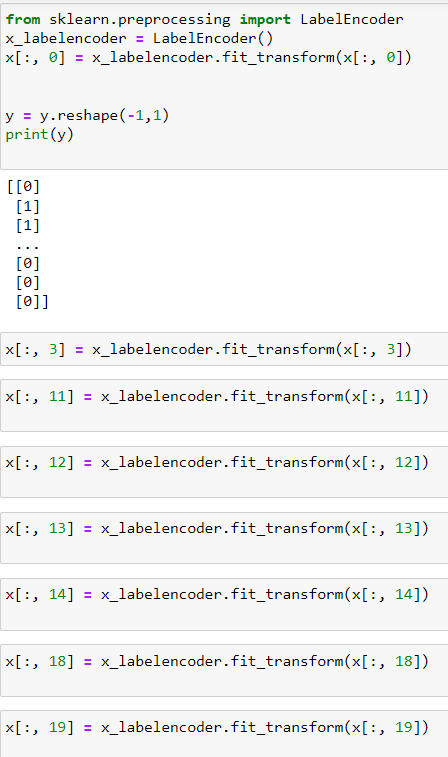


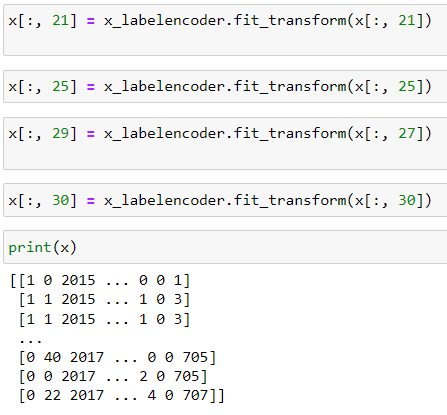
# Assign values x and y

# x has the all values except the “is\_cancelled” column. y consists the “is\_cancelled” column.



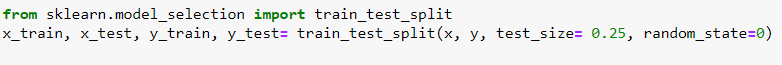
# Import LabelEncoder from sklearn and let us start LabelEncoding those columns that have categorical values to integers.



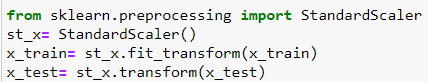
****

### **# 2. Train Test Split**

# Import train\_test\_split from sklearn.model\_selection. Take test size a 0.25(25% of x; training is 75% of x)

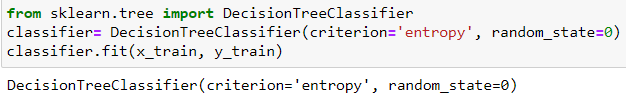


# Import StandardScaling from sklearn.preprocessing and let us scale both testing and training sets.



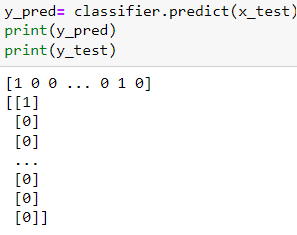
### # **3. Machine Learning Model (Decision Tree)**

# Import DecisionTreeClassifier from sklearn.tree and we choose to execute the classifier using entropy and fit the training set to the model. Now run the code for the model to be trained.

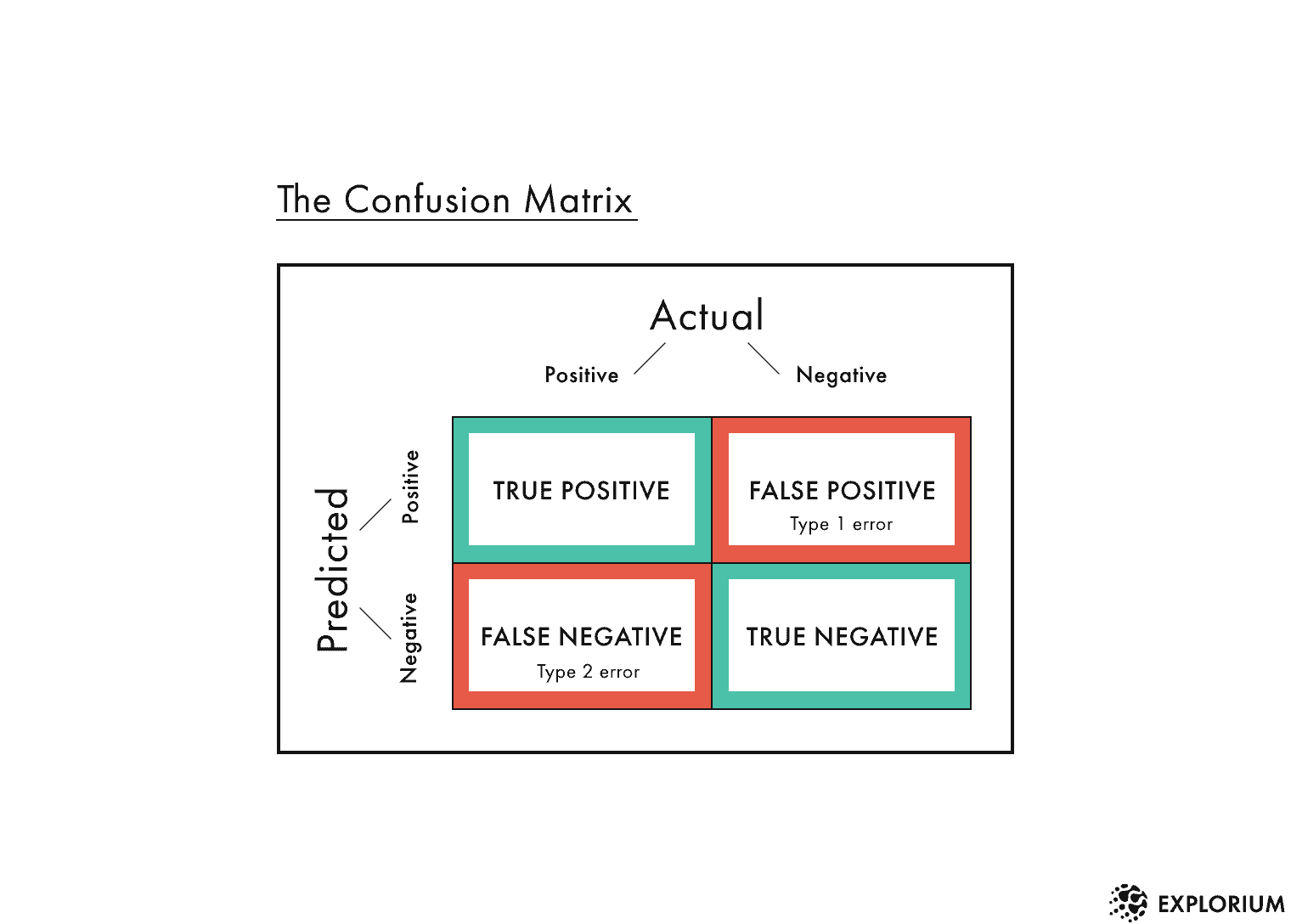


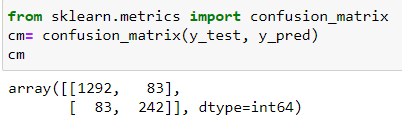
### # **4. Prediction**

# Let us perform prediction.



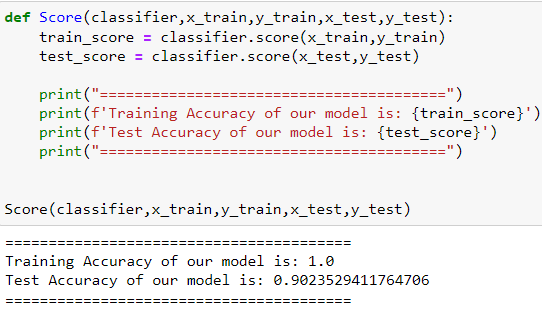
### # **5. Confusion Matrix**

# Import confusion\_matrix from sklearn.metrics and run to know number of correct and wrong predictions.



**# 6. Accuracy**

# Now let us find accuracy of the model.



# Test Accuracy of 0.902 is very good.

**LEARNING OUTCOME FROM THE TRAINING/TECHNOLOGY**

In the six-week summer training course I completed a course on “**Introduction to Data Science**”. The reason to choose this course is I am interested in the field of Artificial Intelligence. My specialization in the coming years is Data Science. So, to have completed this introductory course is a good start. Though I have prior experience in python, NumPy and pandas it was good thing to revise them as I feel clearer with the basics.

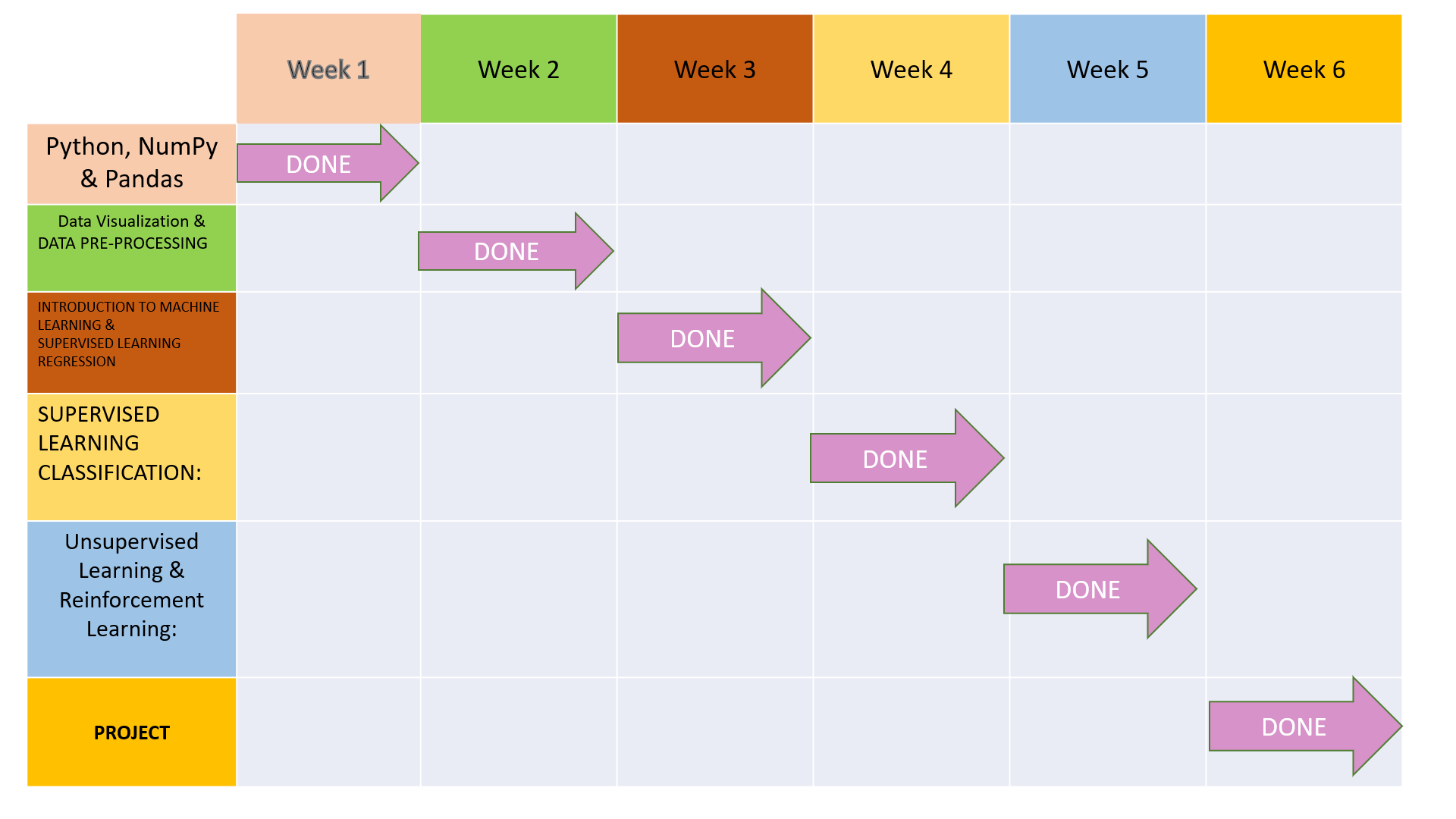
Implementing matplotlib in practical situations everyday made me better at it and also made me clear about by visualization is an important necessity for data science. Now using matplotlib to visualise data in this project was a good experience.

Date Pre-processing is a new topic to me. I did not have any prior experience in it. Learning the basics of data pre-processing gave me a clear idea of how to treat and process data before executing any functions with the data. There is still a lot to work on in the coming years but this is a good start.

I did not have any prior experience with Machine Learning. Learning the basics of Machine Learning was very helpful. Both Data Science and Machine Learning make up a big part of Artificial Intelligence and they correlated. Knowing how both and Machine Learning and Data Science work together to solve the given problem statements gave me a much more precise picture of Artificial Intelligence.

Although six weeks was a short period of time, I am happy I could learn the basics of Data Science not only in theory but also solve a real-life problem statement. Now I have a good understanding what I will be studying in the next few years. But the fact that I should spend more time and gain more experience stays true.

**GANTT CHART**

****

**PROJECT LEGACY**

**TECHNICAL AND MANAGERIAL LESSONS LEARNT**

Even in summer holidays it was good to have classes and learn something. Technically I got the chance to revise python, NumPy and pandas. I got to learn basics of data science and machine learning. During project I got to learn many other things that I could not in class. There were many errors that I faced but I was able to solve them. In summer maintaining a timetable to attend classes and have some own study time is very important. Also, other than studies it is very important to work on extra-curricular activities and explore your hobbies. Having got the time to do all of them I feel I’ve used this summer to best of my capabilities.

**BIBLIOGRAPHY**

1. [**https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand**](https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand)
2. [**https://www.w3schools.com/python/**](https://www.w3schools.com/python/)
3. [**https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm**](https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm)