

HOTEL BOOKING CANCELATION PREDICTION USING ML ALGORITHMS

at

Submitted in partial fulfillment of the requirements for the award of
Bachelor of Engineering Degree in
Computer Science and Engineering

By

MANDALA VENKAT RAKESH

(Reg. No. 38110293)

SUNKARA PRASANNA KUMAR

(Reg. No. 38110576)



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SCHOOL OF COMPUTING

SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with Grade “A” by NAAC

JEPPIAAR NAGAR, RAJIV GANDHI SALAI,

CHENNAI – 600119, TAMILNADU

MARCH 2022



SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with Grade “A” by NAAC

(Established under Section 3 of UGC Act,1956)

JEPPIAAR NAGAR, RAJIV GANDHI SALAI, Chennai -600119

www.sathyabama.ac.in



DEPARTMENT OF COMPUTER SCIENCE AND
ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Professional Training Report is the bonafide work of **Mandala Venkat Rakesh (Reg No. 38110293)**, **Sunkara Prasanna Kumar (Reg No. 38110576)** who have done the project work as a team who carried out the project entitled “**HOTEL BOOKING CANCELTION PREDICTION USING ML ALGORITHMS**” under my supervision from nov 2021 to MAR 2022.

Internal Guide

MS. YOGITHA., M.E., PH.D.,

Head of the Department

Submitted for Viva voce Examination held on_____

Internal Examiner

External Examiner

DECLARATION

I, **MANDALA VENKAT RAKESH (RegNo.38110293), SUNKARA PRASANNA KUMAR (Reg No. 38110576)** hereby declare that the professional Training Report on “**HOTEL BOOKING CANCELATION PREDICTION USING ML ALGORITHMS**” done by me under the guidance of **MS. YOGITHA., M.E., PH.D.**, at Sathyabama Institute of Science and technology is submitted in partial fulfillment of the requirements of the award of Bachelor of Engineering degree in Computer Science and Engineering.

DATE:

PLACE:

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

We are pleased to acknowledge our sincere thanks to Board of management of **SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

We convey our thanks to **Dr. T. SASIKALA M.E., Ph.D., Dean of the Department, Department of Computer Science and Engineering** for providing us the necessary support and details at the right time during the progressive reviews.

We convey our thanks to **Dr. L. LAKSHMANAN., M.E., Ph.D., Head of the Department, Department of Computer Science and Engineering** for providing us the necessary support and details at the right time during the progressive reviews.

We would like to express our sincere and deep sense of gratitude to our Project Guide **MS. YOGITHA., M.E., Ph.D.**, for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

We wish to express our thanks to all Teaching and Non-teaching staff members of the Department of **COMPUTER SCIENCE AND ENGINEERING** who were helpful in many ways for the completion of the project.

ABSTARCT

Cancellations in bookings show a negative impact in the hospitality industry field while making management decisions. To prevent the negative impact of the cancellations, they implement a lot of policies along with few overbooking techniques, this, in turn, can severely damage the income and reputation of that particular hotel. To prevent this situation, they developed machine learning models. These models use previous data from the hotel and then it gets trained and predicts the particular booking is going to get cancelled or not. We have taken two hotels namely Resort hotel and City hotel and then we are going to make our ML models to predict how the particular actions taken by the hotel management showing the impact on the hotel revenue and cancellations. This, in turn, makes management rethink about policies and their decisions. Our ML models will help management to predict number of cancellations are going to happen.

LIST OF FIGURES

FIGURE NO.	FIGURE NAME	PAGE NO
3.3.1	SYSTEM ARCHITECTURE	05
3.6.1	BLOCK DIAGRAM	07
3.9.4	WORKFLOW DIAGRAM	11
3.11.1	BOOTSTRAPPING DIAGRAM	19
4.2	OUTPUTS	22

TABLE OF CONTENTS

CHAPTER NO.	TITLE ABSTRACT.	PAGE NO V
	LIST OF FIGURES.	VI
1.	INTRODUCTION	01
	1.1 OUTLINE OF PROJECT	01
	1.2 OBJECTIVES OF PROJECT	01
	1.3 DOMAIN INTRODUCTION	02
	1.3.1 BASICS OF PYTHON	02
	1.3.2 PYTHON FEATURES	02
2.	AIM AND SCOPE	03
	2.1 AIM OF PROJECT	03
	2.2 SCOPE OF PROJECT	03
	2.3 SOFTWARE REQUIREMENTS.	03
	2.4 HARDWARE REQUIREMENTS.	03
3.	METHODS AND ALGORITHMS USED	04
	3.1 DATA COLLECTION.	04
	3.2 DATA CLEANING.	04
	3.3 DATA PREPROCESSING	04

3.3.1	SYSTEM ARCHITECTURE	05
3.4	FEATURE EXTRACTION	05
3.5	DATA PREPERATION	06
3.6	SECOND STAGE FILTERING	06
3.6.1	BLOCK DIAGRAM	07
3.7	TESTING MODEL.	07
3.8	PERFORMANCE EVALUATION.	08
3.9	PREDICTION.	09
3.9.1	NLTK.	10
3.9.2	TENSORFLOW.	10
3.9.3	KERAS.	10
3.9.4	WORKFLOW DIAGRAM.	11
3.10	SUPPORT VECTOR MACHINE	11
3.10.1	WHAT IS SVM.	12
3.10.2	HOW DOES IT WORK	13
3.11	RANDOM FOREST ALGORITHM	18
3.11.1	WORKING	19
4.	RESULTS	22
4.1	RESULTS	22
4.2	SCREEN SHORTS.	23

5.	CONCLUSION AND FUTURE WORK.	24
	5.1 CONCLUSION	24
	5.2 FUTURE SCOPE.	24
6.	REFERENCE	25
7.	APPENDIX	26
	A. SOURCE CODE	26

CHAPTER 1

INTRODUCTION

1.1 OUTLINE OF THE PROJECT

The ultimate aim of this project is to predict hotel booking cancellations, and develop some Machine Learning models with some best accuracy. After Machine Learning models, use the same data to develop some interactive reports to ease user's task whenever they need any information, they can get it from Power Bi reports and also, they can get future predictions using Machine Learning models and to check the country-wise bookings and cancellations and also to view which month has more bookings. So, our ML models will predict the future bookings based on this experience

1.2 OBJECTIVES OF THE PROJECT

- To ease user's task whenever they need information, they can get from our power bi report directly by just one click. And they can export the data in the form tables if needed.
- The hotel management has to build such a model that it should send all the details regarding a hotel to the right customer at right time depending upon his spending score.
- Even though there is a bit of difference between cancellation and no-show, we are treating both as cancellations for our convenience. Cancellations mean the customer initially booked a room in a hotel but canceled again.
- No-show means the customer failed to check-in. Some of the cancellations occur because of comprehensive reasons like changes in business vacations, change in timings, flight cancellations. But majority cancellations occur because of customers deal-seeking shows in figure1.

Home country of guests

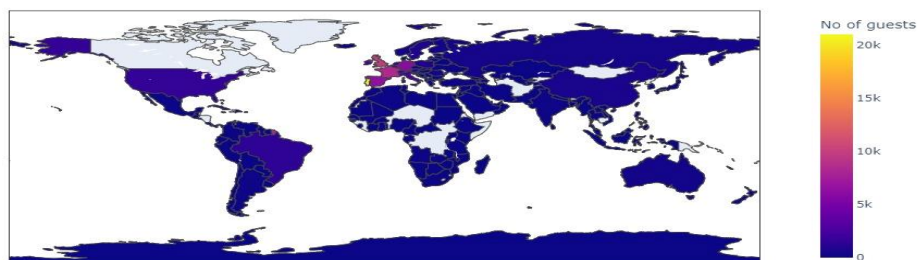


Fig.1 Customer Deal-seeking

1.3 DOMAIN INTRODUCTION:

Python is a widely used general-purpose, high level programming language. It was created by Guido van Rossum in 1991 and further developed by the Python Software Foundation. It was designed with an emphasis on code readability, and its syntax allows programmers to express their concepts in fewer lines of code.

1.3.1 Basics of Python

Python has a simple syntax similar to the English language. Python has syntax that allows developers to write programs with fewer lines than some other programming languages. Python runs on an interpreter system, meaning that code can be executed as soon as it is written.

1.3.2 Python features

Python is a dynamic, high level, free open source and interpreted programming language. It supports object-oriented programming as well as procedural oriented programming. In Python, we don't need to declare the type of variable because it is a dynamically typed language.

CHAPTER 2

AIM AND SCOPE

2.1 AIM OF THE PROJECT

- The purpose of developing this HOTEL BOOKING CANCELLATION PREDICTION is to develop some Machine Learning models and to predict the future bookings and cancellations.
- And to check the country wise bookings and cancellations and also to view which month has more bookings. So, our ML models will predict the future bookings based on this experience.

2.2 SCOPE OF THE PROJECT

To ease user's task whenever they need information, they can get from our power bi report directly by just one click. And they can export the data in the form tables if needed

2.3 SOFTWARE REQUIREMENTS

- Operating system: Windows xp,10,8,8.1,7.
- Coding Language: python 3.7
- Data Base: SQL lite
- IDE: python IDLE 3.7

2.4 HARDWARE REQUIREMENTS

- System: Pentium i5 Processor and above or equivalent
- Processor speed: Minimum 500 MHZ
- Ram: Minimum 256 MB

CHAPTER – 3

METHODS AND ALGORITHMS USED

3.1 LINEAR REGRESSION:

Today in many fields Machine learning algorithms are widely used it solves complex tasks that cannot be solved by humans or by traditional machine approaches one of the simple and mostly used machine learning approach is linear regression. For exploratory data analysis Linear Regression is one of the algorithms that many people use. This is implemented in a place where we need to predict the output or forecast the information

By using one or more independent variables. This will find the relation between the input and independent variables with output or dependent variables by using a gradient descent algorithm. Regression is used in two cases, the first case is to make predictions based on the previous experience, and the second case is to find the relations between different variables. Experience means data in Machine learning.

3.2 K- NEAREST NEIGHBOR ALGORITHM:

A super simple way of classifying data, it is also a supervised learning technique. And it is a parametric less Algorithm, which is without assumptions it takes up the data. Since it is not learning from data available for training this can also be called as lazy or idle learner but during classification the data will

be stored and then it carryout some actions on our data. This algorithm will search for similarities at the time of classification.

3.2.1 Why KNN?

For example, if we have two categories, A and B, if a new data arrives between A and B, using KNN it was easy to classify that new data belongs to category A or B.

3.2.2 Working:

Step1: Number of neighbors should be collected at first.

Step2: Euclidean distance will be calculated from every data point to N estimators.

Step3: According to the distance we had calculated nearest neighbor should be taken

Step4: Now allocate the data points to the group for which the number of neighbors is highest or greater in number

Step6: Our model is ready.

Based on the EUCLIDEAN DISTANCE also it works.

EUCLIDEAN DISTANCE b/w points A and B given below (1)

$$d = \sqrt{[(x_2 - x_1)^2 + (y_2 - y_1)^2]} \quad (1)$$

K- Value as a no particular way to determine, will try different ways and find the best output. Most preferred K=5.

If we go with very less value of K, it will lead to outliers in the model and also it gives noisy in output

3.2.2.1 Advantages:

Implementation of KNN is quite simple

No discriminative function will be derived from data

If training data is more, our KNN will be more effective

3.2.2.2 Disadvantages:

We need to complete Standardization and normalization prior to the implementation of KNN.

KNN algorithm is computationally expensive because it needs to calculate distance for each data point of our dataset.

3.3 DECISION TREE:

It comes under supervised learning, and it has also been used for Regression as well as classification problems. Mostly performable for classification.

- It maintains a tree structure.
- Internal nodes, data set, branches, decision rules are applicable.
- It takes the features of data sets in the form of nodes.
- Decision rules are taken as arms.
- One output will be given by every leaf node.

- In a decision tree there are two types of nodes: 1) Decision node. 2) Leaf node.
- Decision nodes are useful for taking up decisions, branches are there.
- Leaf node is an output for decisions we made, no branches are there.
- In a decision tree, it tries to give a solution in all possible ways, and it's been represented in a graphical way, it is called a decision tree.
- We use a CART-algorithm to build a decision tree.
- CART is a classification, regression algorithm.
- Initially the decision tree starts with a DECISION NODE called as root node, then branches, if no branches it will be leaf node.

3.3.1 Why Decision Tree?

Reason 1: It thinks like a human and gives the decision.

Reason 2: Logic is easily understandable.

3.3.2 Working:

Step1: The starting of the tree should be a root which contains all the variables or features in our dataset

Step2: To find the best attributes of the dataset we have to go for attribute selection or Principal component analysis (PCA)

Step3: Later, we have to divide the root to lot of sub-trees using best attributes or features

Step4: Now we will pick the decision tree which has the best feature

Step5: We have to repeat the process of making new decision trees using step3 until we reach a stage where no further trees can be built from the features in dataset

3.3.3 Rules for feature Selection:

1. After dividing the dataset based on attribute, we will measure the change in entropy; the change is also called as information gain. Information gain calculates how much details will be provided by a feature in our dataset about a class. We will use this information gain value to divide the node to build decision tree. Usually, the Decision tree algorithm tries to boost the IG or information gain value, and it splits the node which is having the highest Value of Information gain. Information gain can be calculate using below (4)

$IG = E(S) - [(WA) * E(I)]$	(4)
-----------------------------	-----

Where,

E= entropy

WA= weighted average

I= each feature

IG= Information Gain

Entropy: The impurity present in a feature of our dataset is called entropy. It describes the randomness in data. Entropy can be calculated below (5)

$s_e - (16/40 * s1_e + 24/40 * s2_e)$	(5)
--	-----

2. Gini Index:

This is the measurement of impurity or cleanness which is used in making decision tree by CART (Classification and regression Tree) algorithm.

The attribute having Low Gini Index should be given more preference when we compared to High Gini Index

The Gini Index will be used by CART algorithm to prepare binary splits

Gini Index can be calculated using the below formula (2) and (3):

$$GI = 1 - \sum_j P_j^2 \quad (2)$$

$$Gi = \text{Gini Index} \quad (3)$$

3.3.3.1 Advantages of the decision Tree:

- For decision-related problems it is very useful.
- For a problem it helps to think about all possible ways.
- It won't require much data wrangling compared to other models.

3.3.3.2 Decision Tree Disadvantages:

- Since its having many number of layers it will become complex
- Using Random Forest algorithm over fitting issue can be resolved.
- It requires more time for model training or we can say it as computational complexity

3.4 NAIVE BAYES CLASSIFIER:

This belongs to the probabilistic classifier family and it is based on Bayes' Theorem. This is also based on kernel density estimation along with simple

Bayesian network. Naive Bayes is a type of classifier that is highly scalable and it also requires a lot of variables that are linear in the problem. For the construction of classifiers, we use a simple technique

Like Naive Bayes. There is no other particular algorithm for training some classifiers, but here we assume a particular feature as independent over any other feature.

For example, if a fruit like mango which is yellow, round, and has a diameter of 14cm. Now our naive Bayes' algorithm consider that all these features should contribute equally or independently for finding whether the fruit is mango or not.

Naive Bayes' classifier is used very efficiently for some supervised learning probability models. Naive Bayes uses conditional probability.

$$k = (c * i) / I \quad \text{Where,}$$

Posterior=K, Prior=c, Likelihood=I, Evidence=I

Interestingly the model or algorithm accuracy is not correlated with the number of independencies

CHAPTER – 4

LITERATURE SURVEY

- A) Predicting cancellations in hotel bookings using machine learning algorithms is simple and common thing in current scenario. These cancellations might bring lot of economic losses to management; hence it is important to predict these cancellations.
- B) This base paper is helpful in explaining how machine learning is applied in this situation to identify which booking is going to get cancelled and this in term prevent losses. For accuracy it should be evaluated in real environment.
- C) Hotel industry is also get affected by the IT industry growth. But the change and impact are quite slow. Many people are doing research in testing and implementing new AI technology along with learning environment in hotels Machine learning is quite trending these days.
- D) The reviews given by customers or Visitors play a major role in reputation as well as revenue system on the hotel.
- E) But most of the people won't read all the comments or reviews given by previous customers of the hotel. These overlooked comments will be grouped and taken necessary steps and emotional analysis is done. Then improvement in the service is done

CHAPTER – 5

PROPOSED WORK

The main aim of the hospitality industry is to provide the quality services to its customers which in turn leave a positive impact in their minds. This will create some positive reputation and reviews. Customer satisfaction is the main aim of this hospitality industry.

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number
2224	Resort Hotel	0	1	2015	October	41
2409	Resort Hotel	0	0	2015	October	42
3181	Resort Hotel	0	36	2015	November	47
3684	Resort Hotel	0	165	2015	December	53
3708	Resort Hotel	0	165	2015	December	53
...	...	...	...	...	...	...
115029	City Hotel	0	107	2017	June	26

Fig 2: Monitored data

5.1 Power bi:

Is a Microsoft data analytics tool. This will help any user to prepare interactive reports using power view. It also allows for data cleaning using power query. Using power bi service, we can share our dashboard to our team or with client, it can periodically load the newly added data and also it will send report to the end user at frequent intervals of time

5.2 Components in Power BI

- Power Bi has big collection of applications which we can use in mobile devices or on Power Bi desktop application as SaaS product
- Power BI Desktop is a software that we can use for free and it is on-premises
- Power Query will be used for data cleaning
- . Power Pivot is for data modeling
- Power View is for developing interactive charts

The different components of Power BI are meant to let users create and share business insights in a way that fits with their role shows in figure 3.

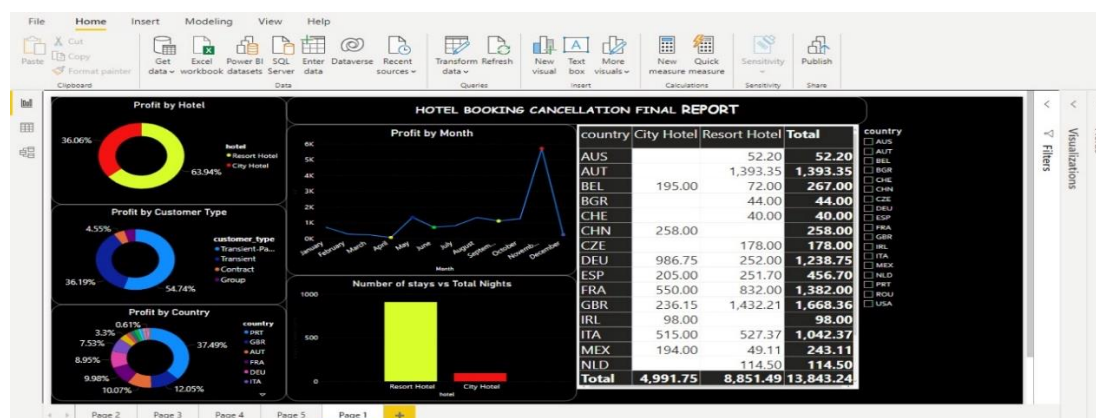


Fig 3: Components of Power BI (Business Insight)

CHAPTER - 6

SYSTEM ARCHITECTURE AND METHODOLOGY

The below is the system Architecture of the project shows in figure 4.

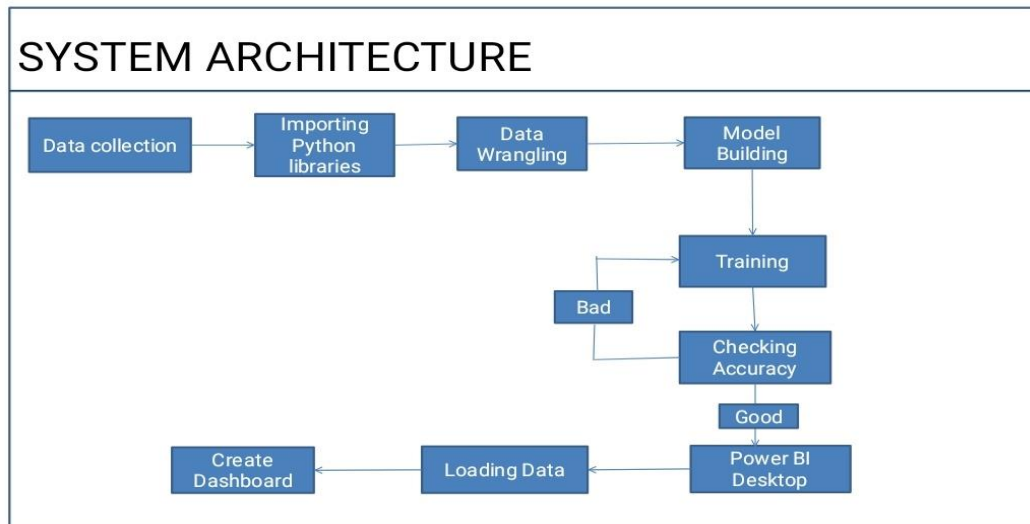


Fig 4: System Architecture

- 1) First the data is assemblage and then stored in the working directory
- 2) By using python libraries, we do data wrangling and data analysis
- 3) Then we go for modeling the data depending on our requirement
- 4) After all these steps we will build our machine learning models
- 5) Then we will make the model to interact with the real world and accuracy will be checked. If the accuracy is good, we will go for building power bi dashboards. If not the training of the model will be done again, we go for modeling the data depending on our requirement
- 6) After all these steps we will build our machine learning models
- 7) Then we will make the model to interact with the real world and accuracy will be checked. If the accuracy is good, we will go for building power bi dashboards. If not the training of the model will be done again.

CHAPTER – 7

RESULTS AND DISCUSSIONS

At this stage it allows us to know various predictions for the data shows in figure 5.

```
LogisticRegression
[[15068 4985]
 [ 3571 6179]]
0.7129148072341711

Naive Bayes
[[ 8547 1253]
 [10092 9911]]
0.6193336241317988

RandomForest
[[18494 1243]
 [  145 9921]]
0.953427507297923

Decision Tree
[[17788 821]
 [  851 10343]]
0.9438982652753078

KNN
[[18484 1440]
 [  155 9724]]
0.9464818977955239
```

Fig 5: Predicted data for Hotel Cancellation

CHAPTER – 8

CONCLUSION AND FUTURE ENHANCEMENT

8.1 CONCLUSION:

The machine learning model is built and get trained using the past data of the hotel. In this way it can understand the patterns of cancellations. So, our ML model can now be able to predict which booking is going to get cancelled and prevent the loss to hotel management field. And interactive dashboards are built using powerbi. So, the end client from hotel side can now be able to see and filter the data and can make on time decisions on their own without human interference.

8.2 FUTURE ENHANCEMENT:

Once after completing all the Machine Learning part, we will move to POWER BI which contains all required charts, slicers, cleaning the data, visualization and exporting the details after filtering. While adding any sales record to our sales excel sheet, they must add all the required details i.e., order-ID, order date, ship date, ship mode, customer id, segment, country, region..... Etc.. The user can also add new measure mean new calculation whenever he or she thinks it will be useful by just clicking on NEW MEASURE option on the head ribbon of the POWERI tool. Data visualization is the concept of presenting data through visuals, such as infographics, chart, Sparkline, and geographic maps etc. Data visualization provides an efficient and effective way to communicate concepts in general because the human brain processes visualized information

CHAPTER – 9

REFERENCES

- [1] Antonio, N., De Almeida, A., & Nunes, L. (2017). Predicting hotel booking cancellations to decrease uncertainty and increase revenue. *Tourism & Management Studies*, 13(2), 25-39.
- [2] Sánchez-Medina, A. J., & Eleazar, C. (2020). Using machine learning and big data for efficient forecasting of hotel booking cancellations. *International Journal of Hospitality Management*, 89, 102546.
- [3] Satu, M. S., Ahammed, K., & Abedin, M. Z. (2020, December). Performance Analysis of Machine Learning Techniques to Predict Hotel booking Cancellations in Hospitality Industry. In *2020 23rd International Conference on Computer and Information Technology (ICCIT)* (pp. 1-6). IEEE.
- [4] Falk, M., & Vieru, M. (2018). Modelling the cancellation behaviour of hotel guests. *International Journal of Contemporary Hospitality Management*.
- [5] Adil, M., Ansari, M. F., Alahmadi, A., Wu, J. Z., & Chakraborty, R. K. (2021). Solving the problem of class imbalance in the prediction of hotel cancelations: A hybridized machine learning approach. *Processes*, 9(10), 1713.

- [6] Antonio, N., de Almeida, A., & Nunes, L. (2017, December). Predicting hotel bookings cancellation with a machine learning classification model. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 1049-1054).
- [7] Saputro, P. H., & Nanang, H. (2021). Exploratory Data Analysis & Booking Cancellation Prediction on Hotel Booking Demands Datasets. *Journal of Applied Data Sciences*, 2(1), 40-56.
- [8] Chen, C. C., Schwartz, Z., & Vargas, P. (2011). The search for the best deal: How hotel cancellation policies affect the search and booking decisions of deal-seeking customers. *International Journal of Hospitality Management*, 30(1), 129-135.
- [9] Lee, H., Yang, S. B., & Chung, N. (2021). Out of sight, out of cancellation: The impact of psychological distance on the cancellation behavior of tourists. *Journal of Air Transport Management*, 90, 101942.
- [10] Novakovic, J., & Turina, S. Hotel reservation cancellations: analysis and prediction using machine learning algorithms. *ACADEMIC JOURNAL*, 4.
- [11] Morrison, A. M., Jing, S., O'Leary, J. T., & Cai, L. A. (2001). Predicting usage of the Internet for travel bookings: An exploratory study. *Information Technology & Tourism*, 4(1), 15-30.
- [12] Alotaibi, E. (2020). Application of Machine Learning in the Hotel Industry: A Critical Review. *Journal of Association of Arab Universities for Tourism and Hospitality*, 18(3), 78-96.
- [13] Wang, H. Y., & Wang, S. H. (2010). Predicting mobile hotel reservation adoption: Insight from a perceived value standpoint. *International Journal of Hospitality Management*, 29(4), 598-608.
- [14] Ansari, A., Shaikh, A., Mapkar, S., & Khan, M. (2019, April). Cancellation Prediction for Flight Data Using Machine Learning. In *2nd International Conference on Advances in Science & Technology (ICAST)*.
- [15] Subramanian, J., Stidham Jr, S., & Lautenbacher, C. J. (1999). Airline yield management with overbooking, cancellations, and no-shows. *Transportation science*, 33(2), 147-167.
- [16] Antonio, N., de Almeida, A., & Nunes, L. (2019). Big data in hotel revenue management: Exploring cancellation drivers to gain insights into

- booking cancellation behavior. *Cornell Hospitality Quarterly*, 60(4), 298-319.
- [17] Pereira, L. N., & Cerqueira, V. (2021). Forecasting hotel demand for revenue management using machine learning regression methods. *Current Issues in Tourism*, 1-18
- [18] Tsai, T. H. (2011). A temporal case-based procedure for cancellation forecasting: a case study. *Current Politics and Economics of South, Southeastern, and Central Asia*, 20(2), 159.
- [19] Putro, N. A., Septian, R., Widiastuti, W., Maulidah, M., & Pardede, H. F. (2021). PREDICTION OF HOTEL BOOKING CANCELLATION USING DEEP NEURAL NETWORK AND LOGISTIC REGRESSION ALGORITHM. *Jurnal Techno Nusa Mandiri*, 18(1), 1-8.
- [20] Song, K. S. (2021). Simultaneous statistical modelling of excess zeros, over/underdispersion, and multimodality with applications in hotel industry. *Journal of Applied Statistics*, 48(9), 1603-1627.
- [21] Lambert, C. U., Lambert, J. M., & Cullen, T. P. (1989). The overbooking question: A simulation. *Cornell Hotel and Restaurant Administration Quarterly*, 30(2), 14-20.
- [22] Putra, M. S. T., & Azhar, Y. (2021). Perbandingan Model Logistic Regression dan Artificial Neural Network pada Prediksi Pembatalan Hotel. *JISKA (Jurnal Informatika Sunan Kalijaga)*, 6(1), 29-37.
- [23] Pinheiro, A. B., Pinto, A. S., Abreu, A., Costa, E., & Borges, I. (2020, October). The Impact of Artificial Intelligence on the Tourism Industry: A Systematic Review. In *International Conference on Tourism, Technology and Systems* (pp. 458-469). Springer, Singapore.
- [24] Abreu, A., Costa, E., & Borges, I. (2020). The Impact of Artificial Intelligence on the Tourism Industry: A Systematic Review. *Advances in Tourism, Technology and Systems: Selected Papers from ICOTTS20, Volume 1*, 208, 458.
- [25] Petricek, M., Chalupa, S., & Melas, D. (2021). Model of Price Optimization as a Part of Hotel Revenue Management—Stochastic Approach. *Mathematics*, 9(13), 1552.
- [26] Antonio, N., de Almeida, A., & Nunes, L. (2019). An automated machine learning based decision support system to predict hotel booking

- cancellations. *An automated machine learning based decision support system to predict hotel booking cancellations*, (1), 1-20.
- [27] Vitek, T., Pachner, D., & Stecha, J. (2007, October). Mathematical Models for Hotel Yield Management. In *2007 IEEE International Conference on Control Applications* (pp. 807-812). IEEE.
- [28] Basson, M. D., Butler, T. W., & Verma, H. (2006). Predicting patient nonappearance for surgery as a scheduling strategy to optimize operating room utilization in a veterans' administration hospital. *The Journal of the American Society of Anesthesiologists*, 104(4), 826-834.
- [29] Barris, L. J. (2006). Hotels and Hurricanes: Legal Duties and Liabilities When Natural Disasters Strike. *Int'l Travel LJ*, 212.
- [30] Schwartz, Z. (2008). Time, price, and advanced booking of hotel rooms. *International Journal of Hospitality & Tourism Administration*, 9(2), 128-146.
- [31] Höpken, W., Fuchs, M., Keil, D., & Lexhagen, M. (2015). Business intelligence for cross-process knowledge extraction at tourism destinations. *Information Technology & Tourism*, 15(2), 101-130.
- [32] Upchurch, R. S., Ellis, T., & Seo, J. (2002). Revenue management underpinnings: an exploratory review. *International Journal of Hospitality Management*, 21(1), 67-83
- [33] Azhar, Y., Mahesa, G. A., & Mustaqim, M. C. (2021). Prediksi pembatalan pemesanan hotel menggunakan optimalisasi hiperparameter pada algoritme Random Forest. *Jurnal Teknologi dan Sistem Komputer*, 9(1), 15-21.
- [34] Putranto, Y., Sartono, B., & Djuraidah, A. (2021). Topic modelling and hotel rating prediction based on customer review in Indonesia. *International Journal of Management and Decision Making*, 20(3), 282-307.
- [35] Benítez-Aurioles, B. (2018). Why are flexible booking policies priced negatively?. *Tourism Management*, 67, 312-325.
- [36] Schwartz, Z., Uysal, M., Webb, T., & Altin, M. (2016). Hotel daily occupancy forecasting with competitive sets: a recursive algorithm. *International Journal of Contemporary Hospitality Management*.

CHAPTER - 10

APPENDIX

A. SOURCE CODE &

B. OUTPUTS :

This is our data in the CSV file which contains

hotel
is_canceled
lead_time
arrival_date_year
arrival_date_month
arrival_date_week_number
arrival_date_day_of_month
stays_in_weekend_nights
stays_in_week_nights
adults
children
babies
meal
country
market_segment
distribution_channel
is_repeated_guest
previous_cancellations
previous_bookings_not_canceled
reserved_room_type
assigned_room_type
booking_changes
deposit_type
agent

company
days_in_waiting_list
customer_type
adr
required_car_parking_spaces
total_of_special_requests
reservation_status
reservation_status_date

hotel	is_cancelled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week	arrival_date_day_of_month	stays_in_weekend	stays_in_week	ni_adults	children
Resort Hotel	0	342	2015	July	27	1	0	0	2	2
Resort Hotel	0	737	2015	July	27	1	0	0	2	2
Resort Hotel	0	7	2015	July	27	1	0	1	1	1
Resort Hotel	0	13	2015	July	27	1	0	1	2	2
Resort Hotel	0	14	2015	July	27	1	0	2	2	2
Resort Hotel	0	14	2015	July	27	1	0	2	2	2
Resort Hotel	0	0	2015	July	27	1	0	2	2	2
Resort Hotel	0	9	2015	July	27	1	0	2	2	2
Resort Hotel	1	85	2015	July	27	1	0	3	2	2
Resort Hotel	1	75	2015	July	27	1	0	3	2	2
Resort Hotel	1	23	2015	July	27	1	0	4	2	2
Resort Hotel	0	35	2015	July	27	1	0	4	2	2
Resort Hotel	0	68	2015	July	27	1	0	4	2	2
Resort Hotel	0	18	2015	July	27	1	0	4	2	2
Resort Hotel	0	37	2015	July	27	1	0	4	2	2
Resort Hotel	0	68	2015	July	27	1	0	4	2	2
Resort Hotel	0	37	2015	July	27	1	0	4	2	2
Resort Hotel	0	12	2015	July	27	1	0	1	2	2
Resort Hotel	0	0	2015	July	27	1	0	1	2	2
Resort Hotel	0	7	2015	July	27	1	0	4	2	2
Resort Hotel	0	37	2015	July	27	1	1	4	1	1
Resort Hotel	0	72	2015	July	27	1	2	4	2	2
Resort Hotel	0	72	2015	July	27	1	2	4	2	2

Now we have to import all the necessary libraries

```
In [ ]:

In [1]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

In [ ]:
```

Now we have to read the csv file in pandas. And check the top few records

```
df=pd.read_csv('hotel_bookings.csv')

df.head()
```



```

children has unique values as [ 0.  1.  2. 10.  3.]
adults has unique values as [ 2  1  3  4 40 26 50 27 55  0
20  6  5 10]
babies has unique values as [ 0  1  2 10  9]

```

There is some abnormalities in our data,like at the same time children, adults, babies should not be zero

```

filter=(df['children']==0) & (df['adults']==0) & (df['babies']==0)
df[filter]

```

Visualise Entire Dataframe where adult,children & babies are 0

Home Page - Select or create a notebook

Hotel Bookings Cancellation Final Output

localhost:8888/notebooks/Hotel%20Bookings%20Cancellation%20Final%20Output.ipynb

jupyter

Hotel Bookings Cancellation Final Output Last Checkpoint: 07/23/2021 (autosaved)

Logout

File Edit View Insert Cell Kernel Help

Trusted Python 3

↑ ↓ ↺ ↻ ↵

Run

Code

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	st
2224	Resort Hotel	0	1	2015	October	41	6	0	
2409	Resort Hotel	0	0	2015	October	42	12	0	
3181	Resort Hotel	0	36	2015	November	47	20	1	
3684	Resort Hotel	0	165	2015	December	53	30	1	
3708	Resort Hotel	0	165	2015	December	53	30	2	
...	...	...	...	...	...	...	...	...	
115029	City Hotel	0	107	2017	June	26	27	0	
115091	City Hotel	0	1	2017	June	26	30	0	
116251	City Hotel	0	44	2017	July	28	15	1	
116534	City Hotel	0	2	2017	July	28	15	2	
117087	City Hotel	0	170	2017	July	30	27	0	

180 rows x 32 columns

11:40 AM

24-Sep-21

And now we will apply complement to that filter,so all the above abnormalities will be get deleted.

```
data=df[~filter]
```

```
data
```

now we can see all the abnormalities will be get deleted.

```

In [17]: data.shape
Out[17]: (119210, 32)

In [18]: data.head()
Out[18]:
  hotel  is_canceled  lead_time  arrival_date_year  arrival_date_month  arrival_date_week_number  arrival_date_day_of_month  stays_in_weekend_nights  stays_in_week_not_nights
0  Resort Hotel      0       342             2015              July                27                    1                        0
1  Resort Hotel      0       737             2015              July                27                    1                        0
2  Resort Hotel      0        7             2015              July                27                    1                        0
3  Resort Hotel      0       13             2015              July                27                    1                        0
4  Resort Hotel      0       14             2015              July                27                    1                        0

Type Markdown and LaTeX: α²

Where do the guests come from

In [19]: country_wise_data=data[data['is_canceled']==0]['country'].value_counts().reset_index()

```

Where do the guests come from¶

```
country_wise_data=data[data['is_canceled']==0]['country'].value_counts().reset_index()
```

```
country_wise_data.columns=['country','No of guests']
```

```
country_wise_data
```

o/p:-

```

country_wise_data

Out[19]:
  country  No of guests
0  PRT      20977
1  GBR      9668
2  FRA      8468
3  ESP      6383
4  DEU      6067
...
161 MMR      1
162 MLI      1
163 ZMB      1
164 NCL      1
165 BFA      1

166 rows x 2 columns

```

```

filt=(df['is_canceled']==0)country_=df[filt]['country'].value_counts().reset_index()country_

```

```

import folium
from folium.plugins import HeatMap

# show on map
map_guest = px.choropleth(country_wise_data,

                           locations=country_wise_data['country'],

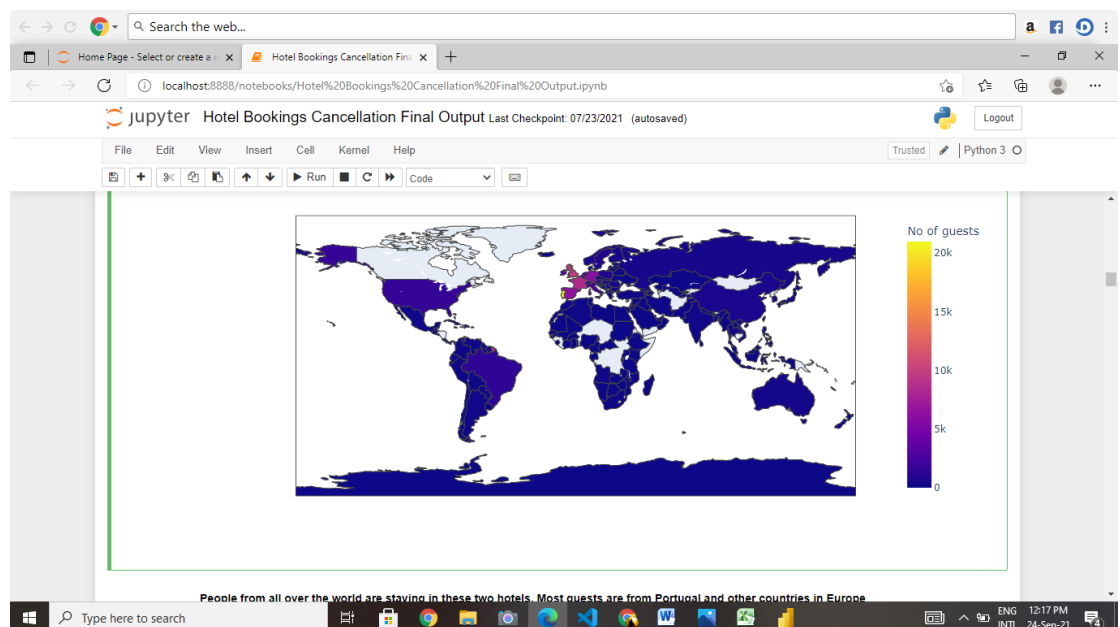
                           color=country_wise_data['No of guests'],

                           hover_name=country_wise_data['country'],

                           title="Home country of guests")

map_guest.show()

```



Conclusion:- People from all over the world are staying in these two hotels.
Most guests are from Portugal and other countries in Europe

How much do guests pay for a room per night?

```
In [26]: data.head()
```

Out[26]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights
0	Resort Hotel	0	342	2015	July	27	1	0	
1	Resort Hotel	0	737	2015	July	27	1	0	
2	Resort Hotel	0	7	2015	July	27	1	0	
3	Resort Hotel	0	13	2015	July	27	1	0	
4	Resort Hotel	0	14	2015	July	27	1	0	

Both hotels have different room types and different meal arrangements. Seasonal factors are also important. So the prices vary a lot. Since no currency information is given, but Portugal is part of the European Monetary Union, I assume that all prices are in EUR.

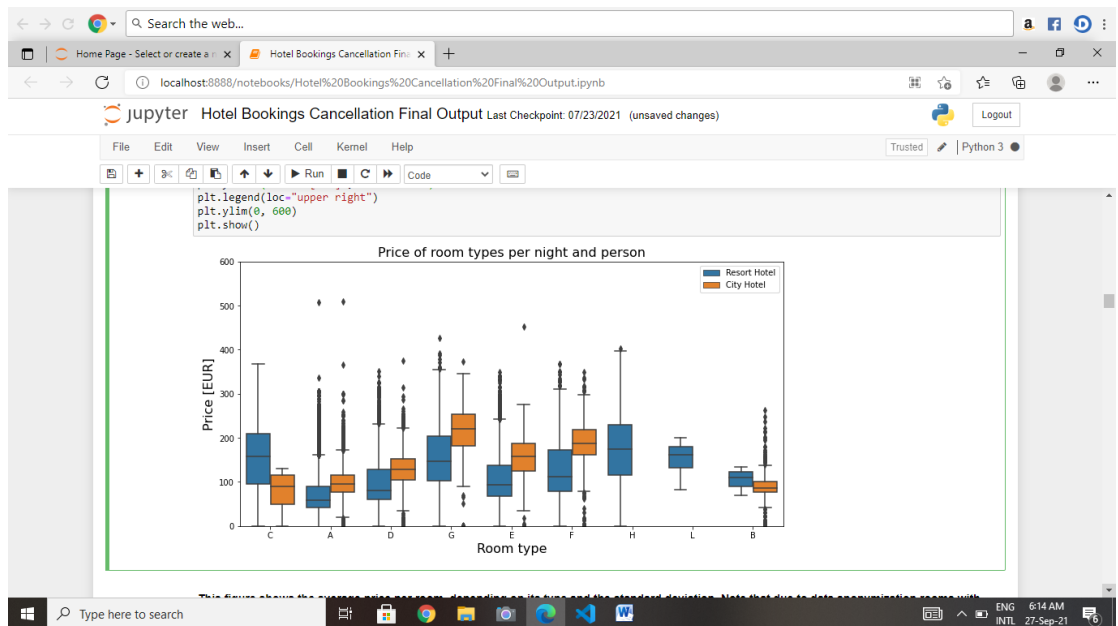
```
In [27]: data2=data[data['is_canceled']==0]
```

How much do guests pay for a room per night?

Both hotels have different room types and different meal arrangements. Seasonal factors are also important. So the prices vary a lot. Since no currency information is given, but Portugal is part of the European Monetary Union, I assume that all prices are in EUR.

```
data2=data[data['is_canceled']==0]
plt.figure(figsize=(12, 8))
sns.boxplot(x="reserved_room_type",
            y="adr",
            hue="hotel",
            data=data2)

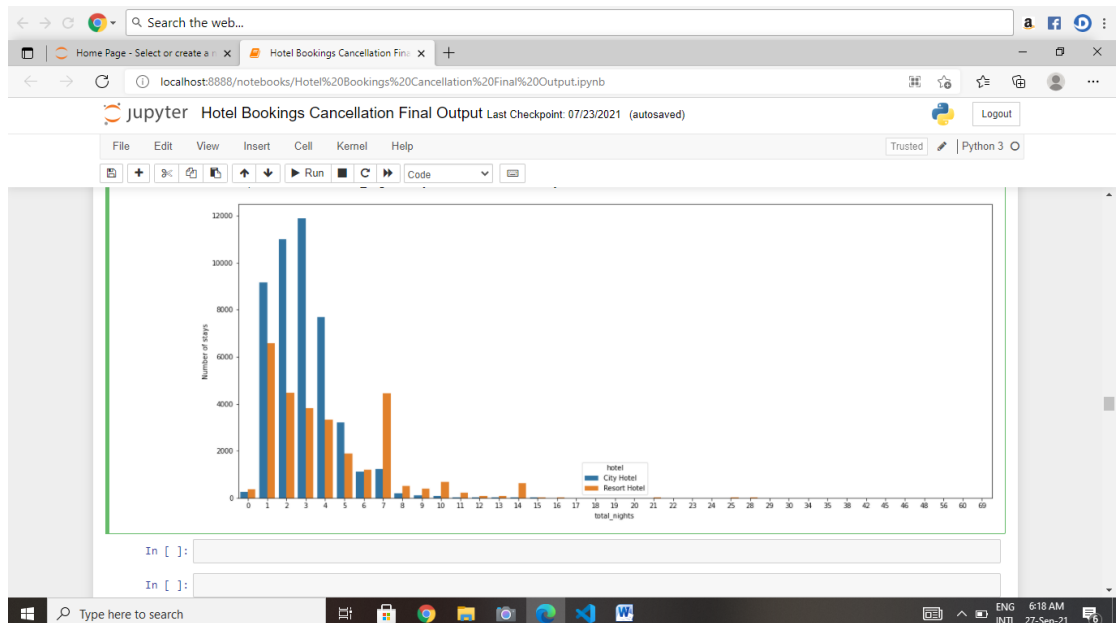
plt.title("Price of room types per night and person", fontsize=16)
plt.xlabel("Room type", fontsize=16)
plt.ylabel("Price [EUR]", fontsize=16)
plt.legend(loc="upper right")
plt.ylim(0, 600)
plt.show()
```



This figure shows the average price per room, depending on its type and the standard deviation. Note that due to data anonymization rooms with the same type letter may not necessarily be the same across hotels.

```
plt.figure(figsize=(20, 8))
```

```
sns.barplot(x = "total_nights", y = "Number of stays", hue="hotel",  
            hue_order = ["City Hotel", "Resort Hotel"], data=stay)
```



Conclusion- Both hotels have different room types and different meal arrangements. Seasonal factors are also important. So the prices vary a lot. Since no currency information is given, but Portugal is part of the European Monetary Union, I assume that all prices are in EUR.

splitting dataset & model Building

```
from sklearn.model_selection import train_test_split

x_train,x_test,y_train,y_test=train_test_split(x,y,train_size=0.75,random
_state=0)

from sklearn.linear_model import LogisticRegression

logreg=LogisticRegression()

logreg.fit(x_train,y_train)

y_pred=logreg.predict(x_test)

from sklearn.metrics import confusion_matrix

cm=confusion_matrix(y_test,y_pred)

cm

from sklearn.metrics import accuracy_score

score=accuracy_score(y_test,y_pred)

score
```

```

from sklearn.model_selection import cross_val_score

score=cross_val_score(logreg,x,y,cv=10)

o/p:- array([0.67796326,  0.71990605,  0.70715544,  0.73081117,
0.75438302,
          0.7411291 , 0.57746833, 0.79162822, 0.72653301, 0.73557047])
#fit naive bayes

from sklearn.naive_bayes import GaussianNB

from sklearn.linear_model import LogisticRegression

from sklearn.neighbors import KNeighborsClassifier

from sklearn.ensemble import RandomForestClassifier

from sklearn.tree import DecisionTreeClassifier

models = []

models.append(('LogisticRegression', LogisticRegression()))

models.append(('Naive Bayes',GaussianNB()))

models.append(('RandomForest', RandomForestClassifier()))

models.append(('Decision Tree', DecisionTreeClassifier()))

models.append(('KNN', KNeighborsClassifier(n_neighbors = 5)))

for name, model in models:

    print(name)

```

```
model.fit(x_train, y_train)
```

```
# Make predictions.
```

```
predictions = model.predict(x_test)
```

```
# Compute the error.
```

```
from sklearn.metrics import confusion_matrix
```

```
print(confusion_matrix(predictions, y_test))
```

```
from sklearn.metrics import accuracy_score
```

```
print(accuracy_score(predictions,y_test))
```

```
print('\n')
```

o/p:- LogisticRegression

```
[[15878 5274]
```

```
[ 2761 5890]]
```

```
0.7303962688319968
```

Naive Bayes

```
[[ 8547 1253]
```

```
[10092 9911]]
```

0.6193336241317988

RandomForest

[[18521 1252]

[118 9912]]

0.9540314733416099

Decision Tree

[[17799 825]

[840 10339]]

0.9441331409589638

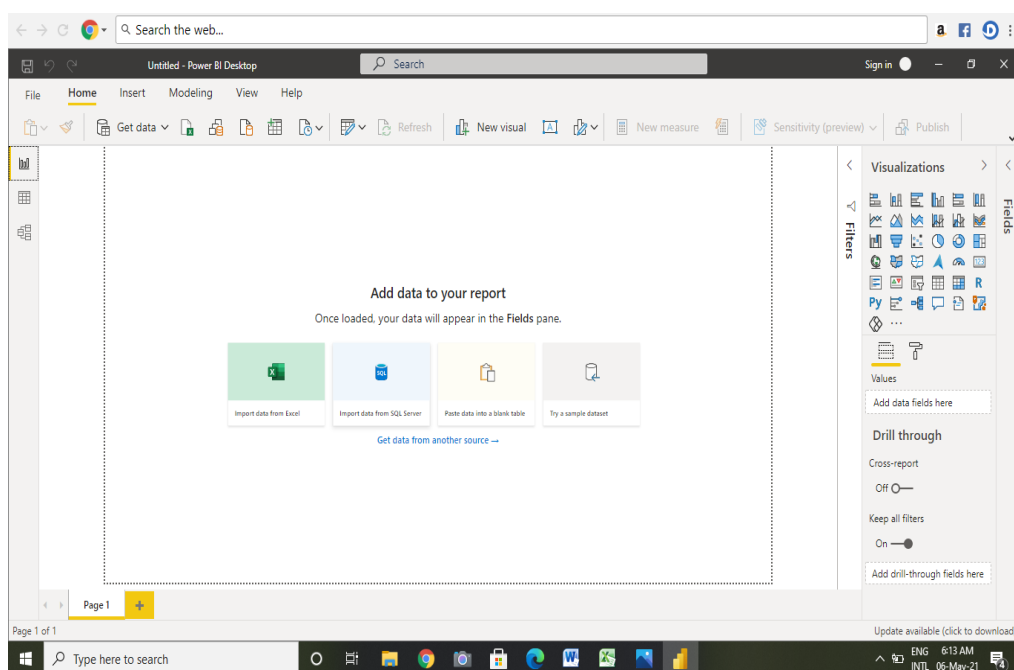
KNN

[[18484 1440]

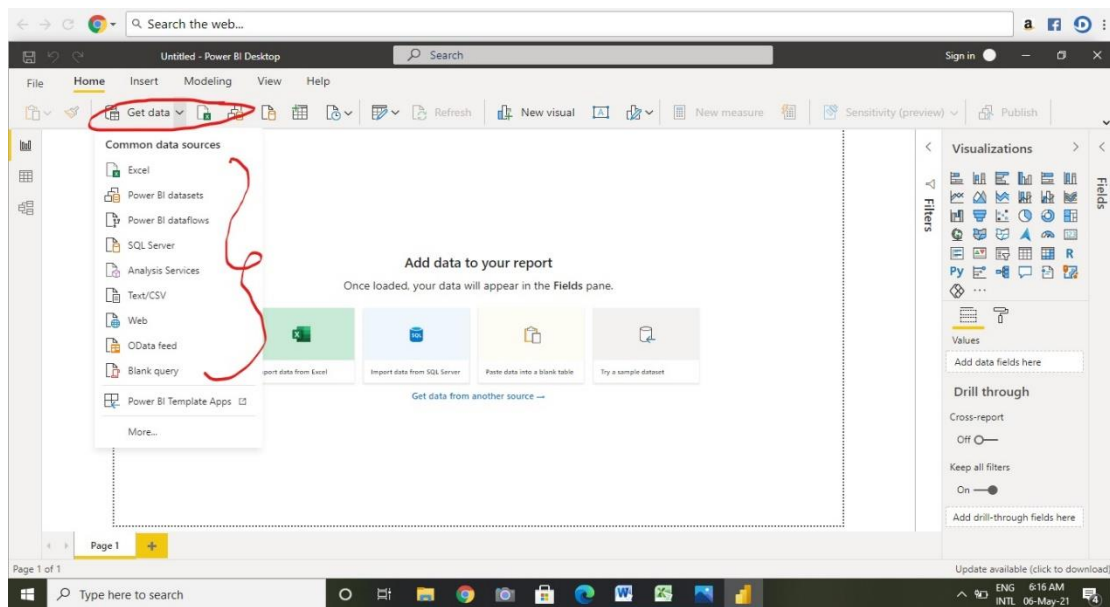
[155 9724]]

0.9464818977955239

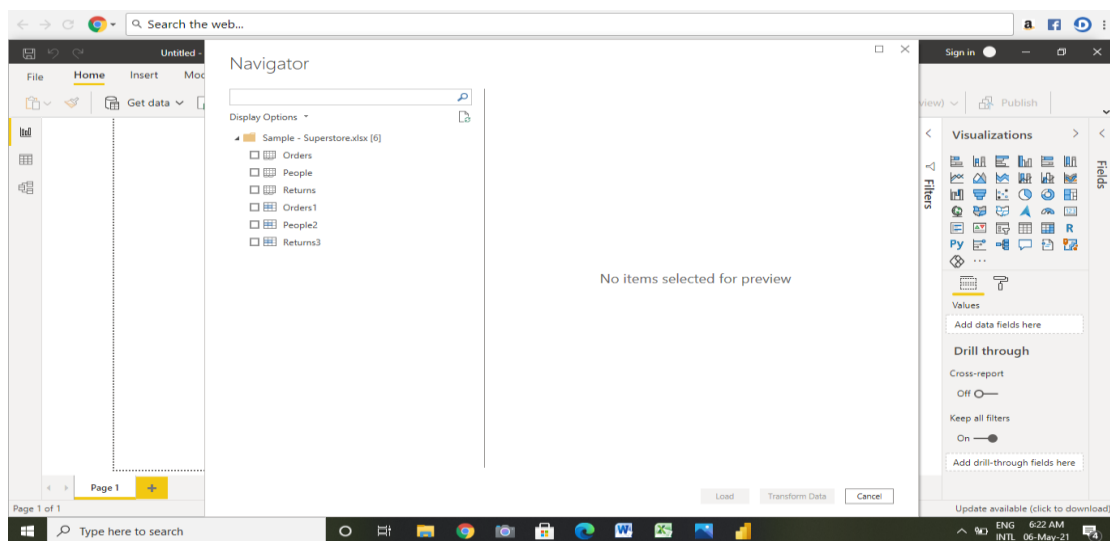
This the user interface of POWERBI tool



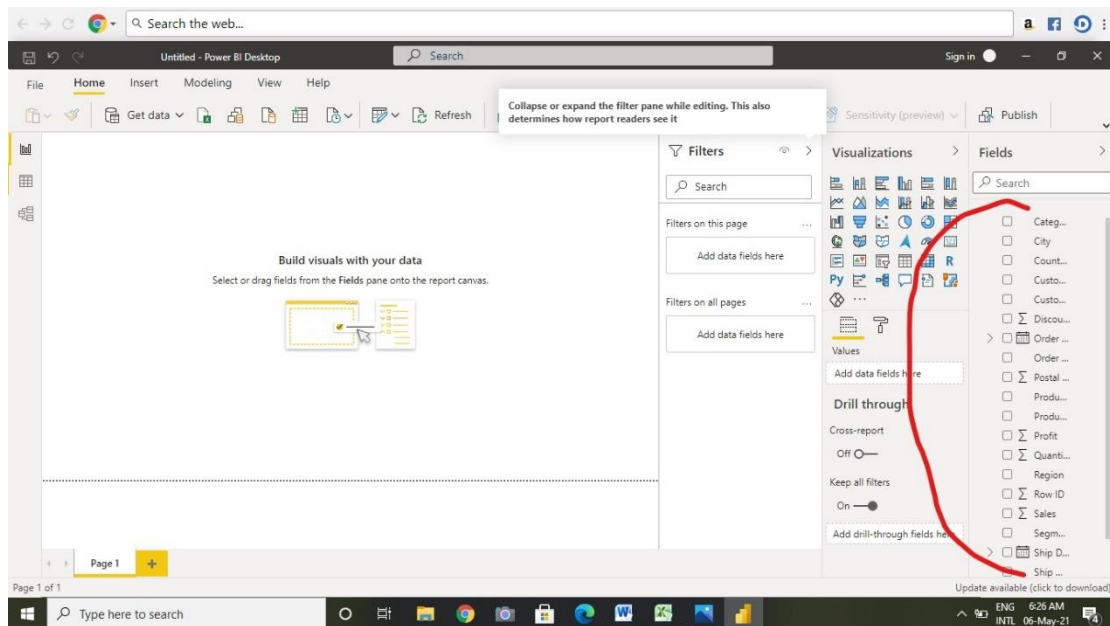
Powerbi supports almost 100plus data sources:-



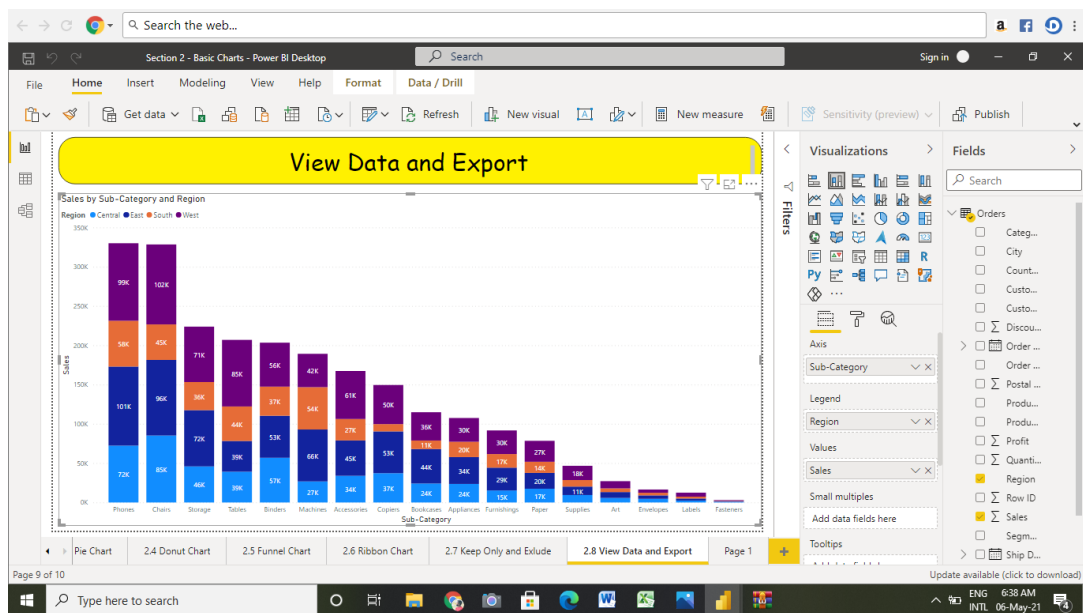
After loading the particular data sheet we have to select the name we want to work on

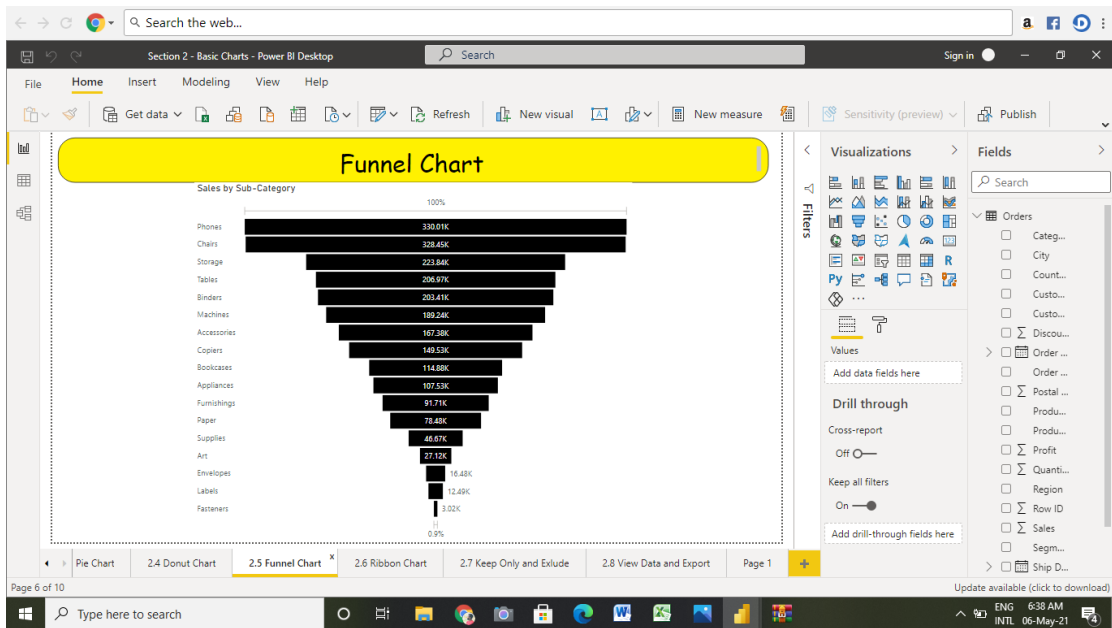
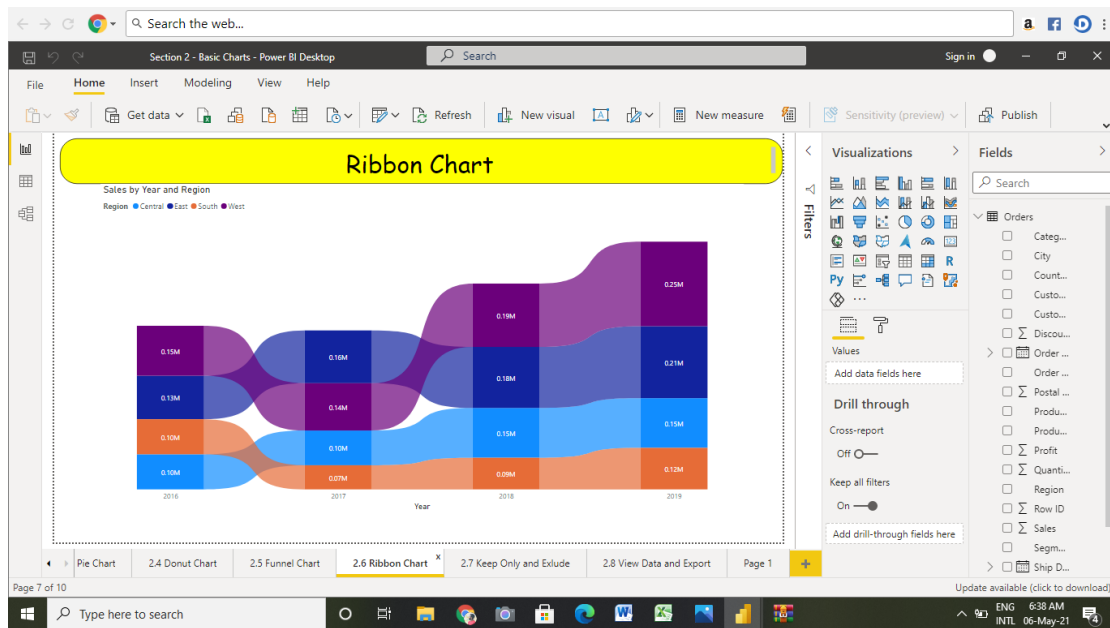


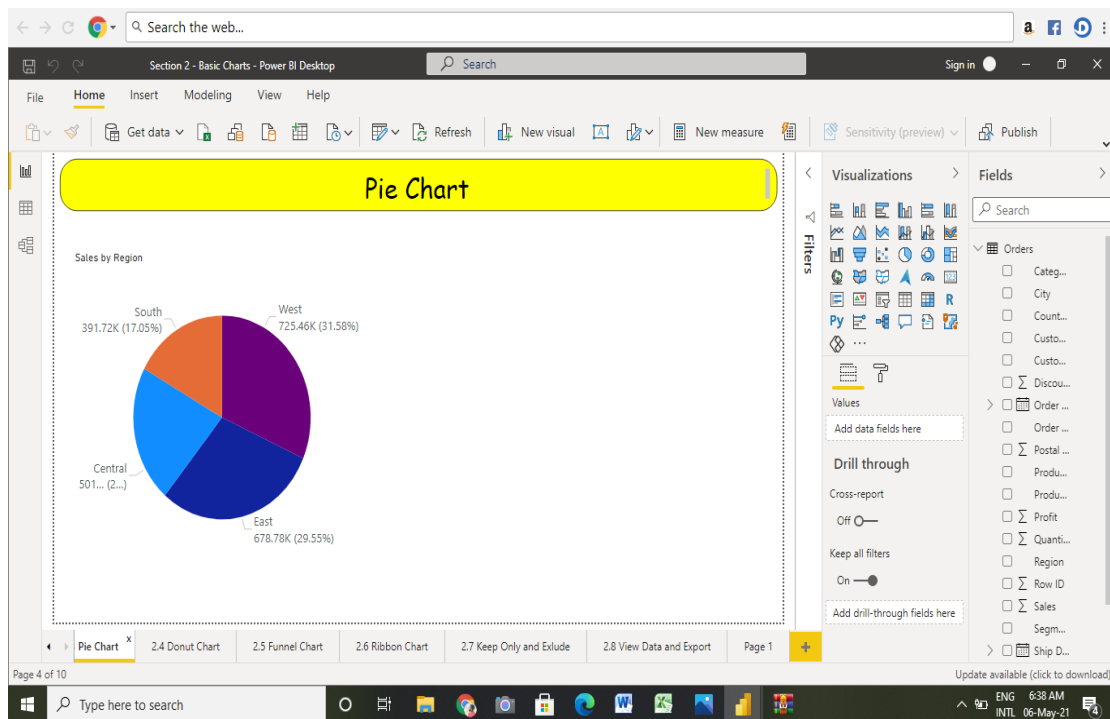
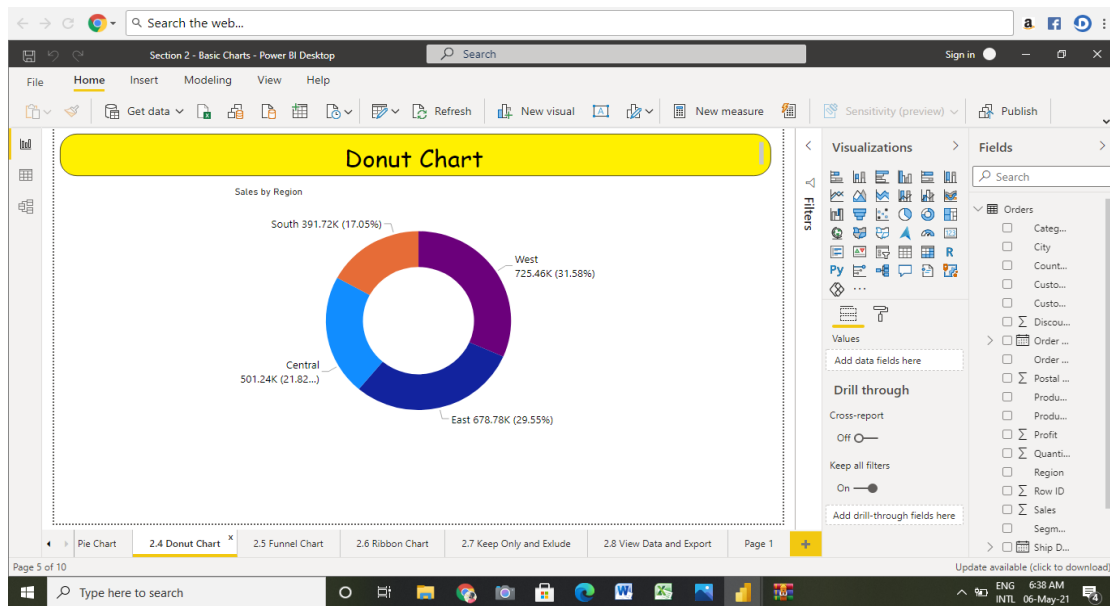
After loading all the data, it will be available for visualizations, and this can be seen as



The basic charts in PowerBi

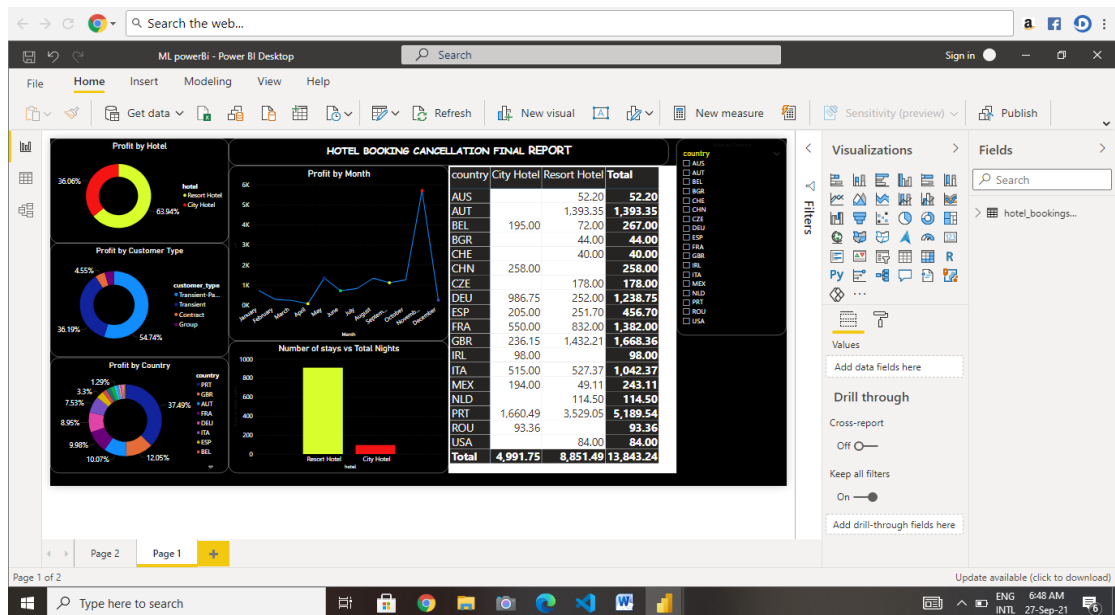






The above charts are the basic charts given by Power BI, from this now we are going to develop a dashboard

- This is the dashboard that having the overall results of our data



And if we want to filter out the data, we can do this with just by one click on slicer which is available at left side of the screenshot i.e I want to study and analysis about Indiana, Iowa and cansas.

