**COIMBATORE INSTITUTE OF TECHNOLOGY**

****

**PROJECT REPORT**

**IN**

**COMPUTING LAB (DATA SCIENCE)**

**(2020 - 2021)**

**SUBMITTED BY**

**S.S.NANDHAKUMAR MDS 2032022**

**S.NAVIN MDS 2032026**

**ACKNOWLEDGEMENT**

At the outset, I would like to take this opportunity to express my gratitude to the almighty God, for keeping me in good health all through this grueling project work.

I place on record my sincere gratitude and appreciation to my project guide Mrs. Rathika and Mrs. Chandina my Computing Lab professor for her kind cooperation and guidance which enabled me to complete this project in time.

I take this opportunity to dedicate my project to all the faculty members who were a constant source of motivation and I express my deep gratitude to their never ending support and encouragement during this project.

Finally, I thank each and every one who helped me to complete

the task.

(S.S.NANDHAKUMAR)

MDS 2032022

**PROJECT TITLE:**

**PREDICTING SATISFACTION OF AIRLINE PASSENGERS**

**ABSTRACT**

In the aviation industry, high-grade customer satisfaction is a key factor to run the business, as the airline industry is very competitive and customer loyalty varies with small changes in the services. Therefore, companies need to understand the customers’ need to deliver unparalleled experiences to retain customers. Using the customer’s satisfaction dataset, we here to analysis the reasons for customer experience being satisfied or not. Based on that, improvements will be made to provide better service by the airline company. Also, as part of the analysis, we will be able to understand several factors which improve customer satisfaction level.

**INTRODUCTION**

Due to fierce competition in the airline industry, the airline company needs to focus on the passenger’s experience and satisfaction. Customer feedback, in particular, is critical since it is an outcome measurement for business performance. According to the international air transport association (IATA), numbers of airline passengers were increasing by about 7% every year since 2015. However, the net profit per airline passenger was decreasing by $10 for 2015, $9 for 2016 and 2017, and it was estimated at only $7.4 for 2018. This is mainly due to intense competition, and also airline costs have been rising recently. The major expenses that affect companies in the airline industry are labor, fuel and other maintenance costs. The airline industry continues to be competitive, even though many people are traveling by aircraft. The Internet has also created greater price transparency, reducing margins.

Many studies have employed survey methods to measure service quality in the airline industry. However, a few recent studies have highlighted the advantages of analyzing online review data for studying customers’ satisfaction or their experience of the airline. Online reviews are critical since it is a significant source for business growth, performance and improvement of customer experience, and allow airline companies to conduct two-way communication with airline passengers. Moreover, electronic word of mouth (eWOM) shared by other airline passengers are considered trustworthy, fast and widespread.

**OBJECTIVE**

The main aim of this project is to determine the relative importance of each parameter with regards to their contribution to passenger satisfaction using both analytical techniques and predictive algorithm.

Specific objectives are:

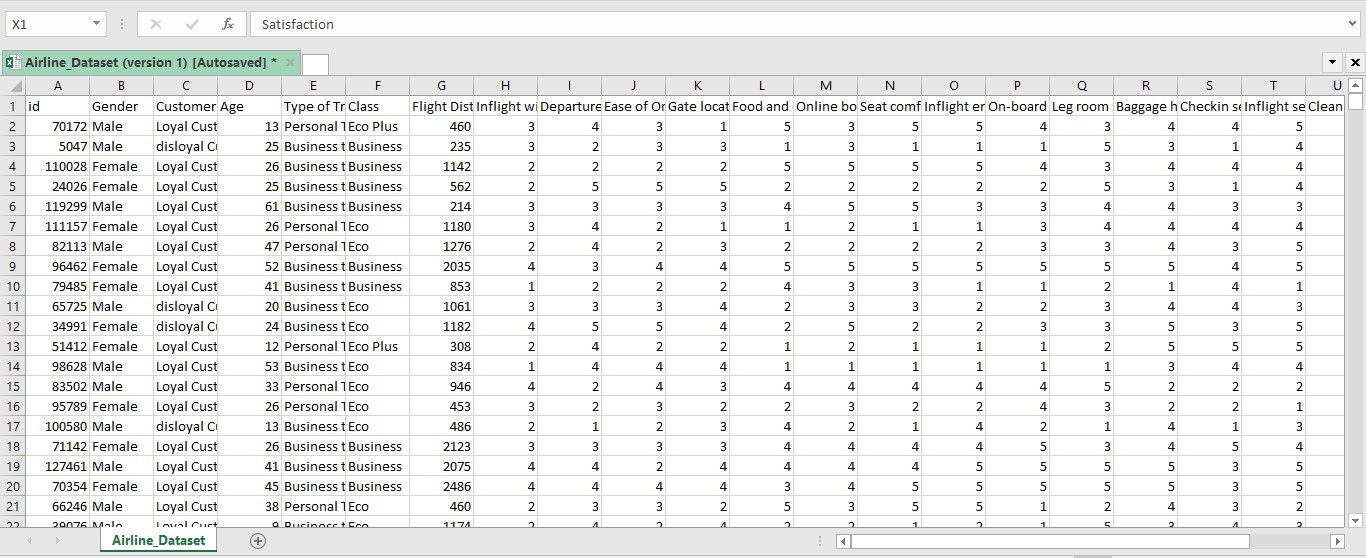
* To analysis the dataset and understand which variable plays an important role in passengers satisfaction.
* And to predict their satisfaction on the basis of their feedback.
* And to check the accuracy of the prediction.

**DATASET**

* Our dataset of passengers feedback and rating was downloaded from Kaggle and imported into the program.
* The dataset contains over 1,29,880 entries.
* It contains passengers personal information such as ID, gender, purpose of travel and age.
* It also contains their details of their flight class (Economy, Business, Eco-Plus), type of customer(first time or regular customer) and flight distance.

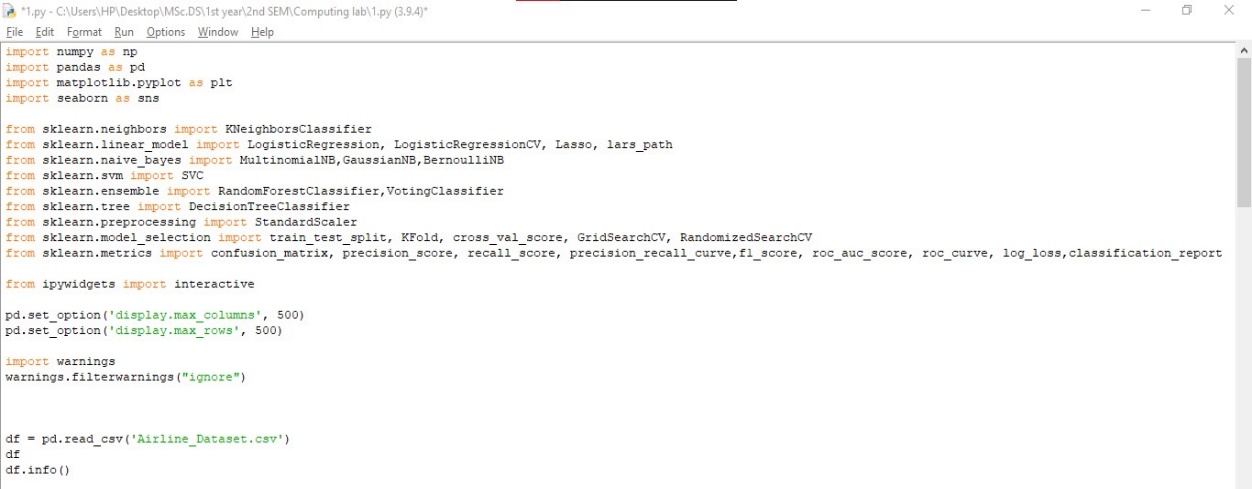
And it also contains passengers ratings and feedback on different categories and features of the flight(Wi-Fi, Leg room, Food, etc.).

A sample part of our dataset(Airline\_dataset).

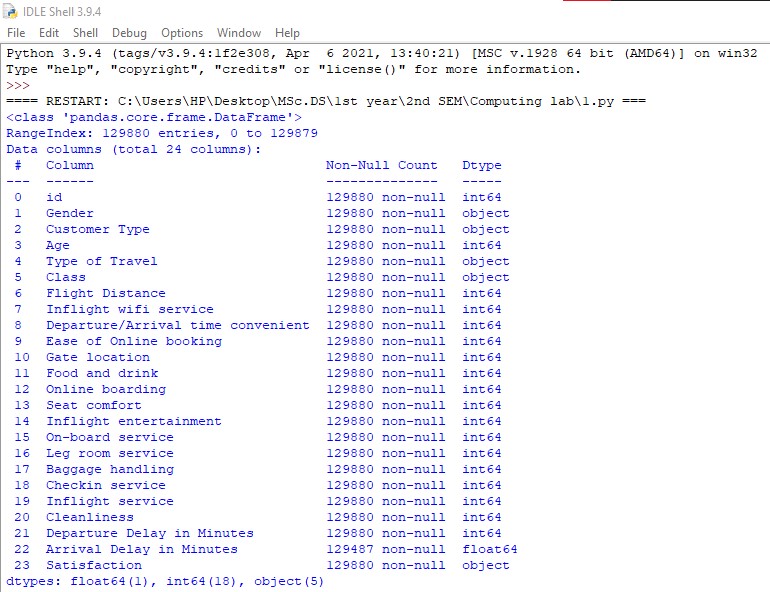


**IMPORTING DATA**

All the important python library and the DATASET are imported into python



Here is the information of DATASET



**EXPLORATORY DATA ANALYSIS AND FEATURE SELECTION**

**Create visualizations to first understand business problem, and also identify important features for model building:**

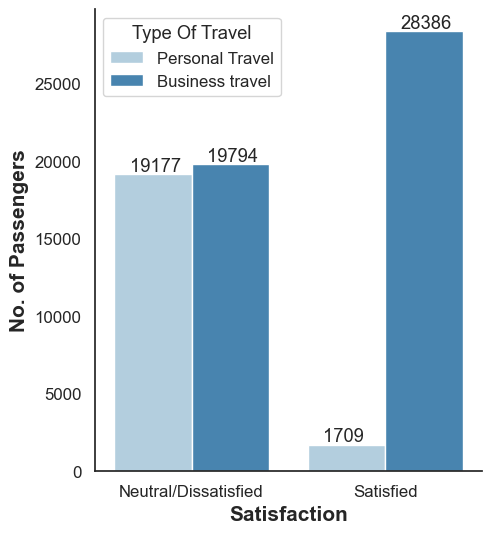
* Find out proportion of classes in target, and split them by Type of Travel and Type of Customers (To understand trend of satisfaction - useful later in model evaluation)
* Identify feature significance for model through visualizing Graph plots, LASSO path and correlation heatmap.
* After evaluation and discreet selection, We have decided to drop 'Gender, 'Total Delay','Flight Distance','Age','Gate Location' and 'Departure/Arrival Time Convenience'

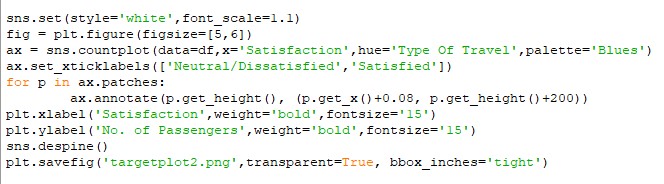
**GRAPHS**

Graphs play an important  role in displaying and analysing quantitative data.

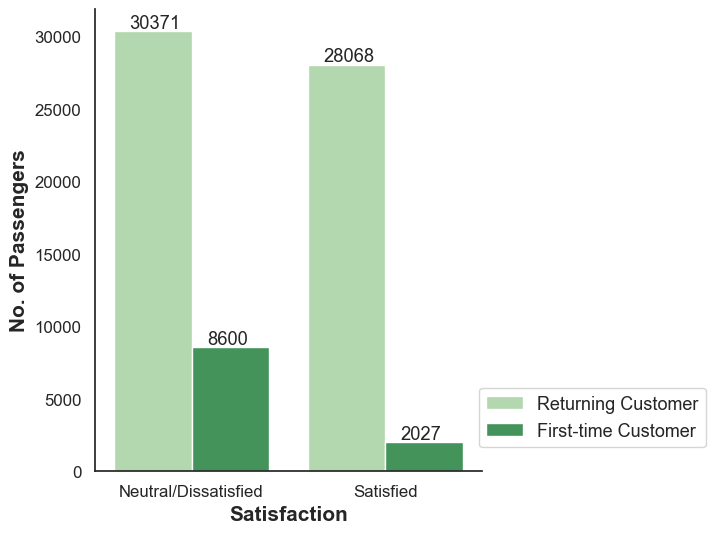
A graph data structure consists of a finite (and possibly mutable) set of *vertices* (also called *nodes* or *points*), together with a set of unordered pairs of these vertices for an undirected graph or a set of ordered pairs for a directed graph. These pairs are known as *edges* (also called *links* or *lines*), and for a directed graph are also known as *edges* but also sometimes *arrows* or *arcs*. The vertices may be part of the graph structure, or may be external entities represented by integer indices or references.

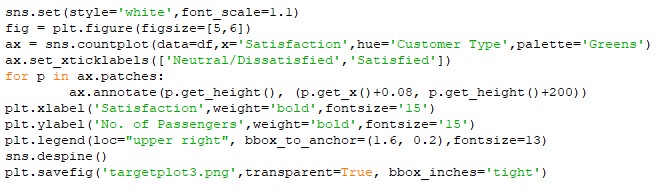
* Here is a customized graph showing passenger satisfaction based on the purpose of travel:



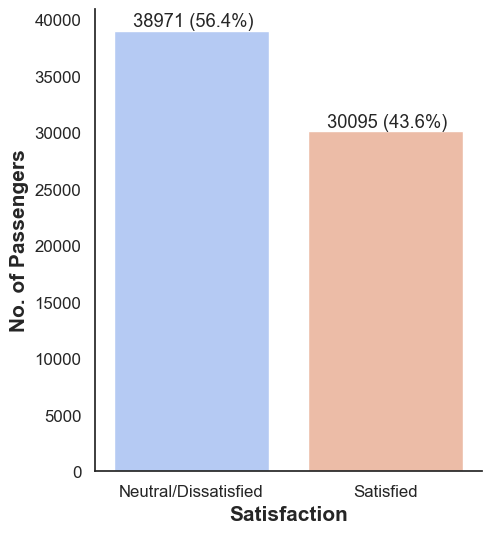


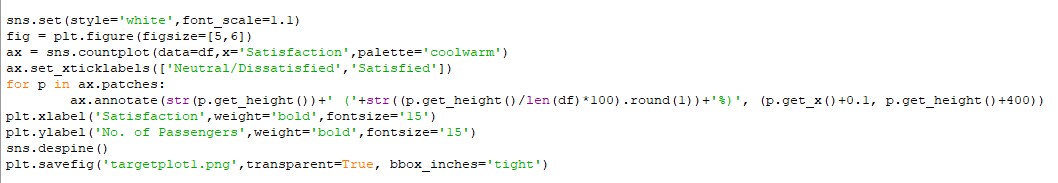
* This graph shows passenger satisfaction based on type of customer(whether first time or regular customer):





* Here this graph shows overall passengers satisfaction irrespective of the categories and feature:





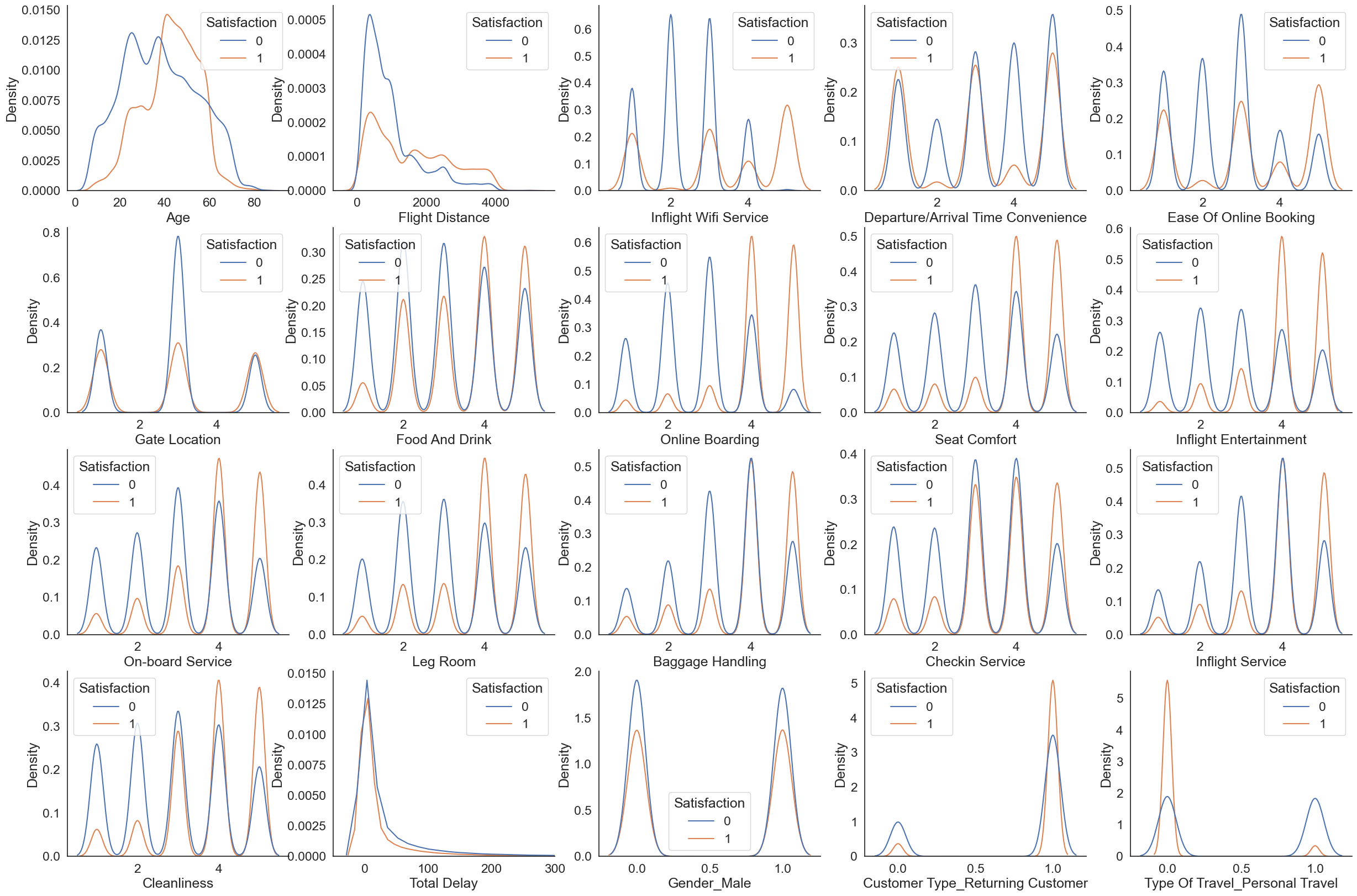
**FEATURE SELECTION**

**KERNEL DENSITY ESTIMATION:**

Now with the help of **KDE**, a technique that let's you create a smooth curve given a set of data. This can be useful if you want to visualize just the “shape” of some data, as a kind of continuous replacement for the discrete histogram.

This helps us in determining the important features(variables) that matters in passengers satisfaction. We can remove all the other features which does not play any roll in customers satisfaction prediction.

KED plots:



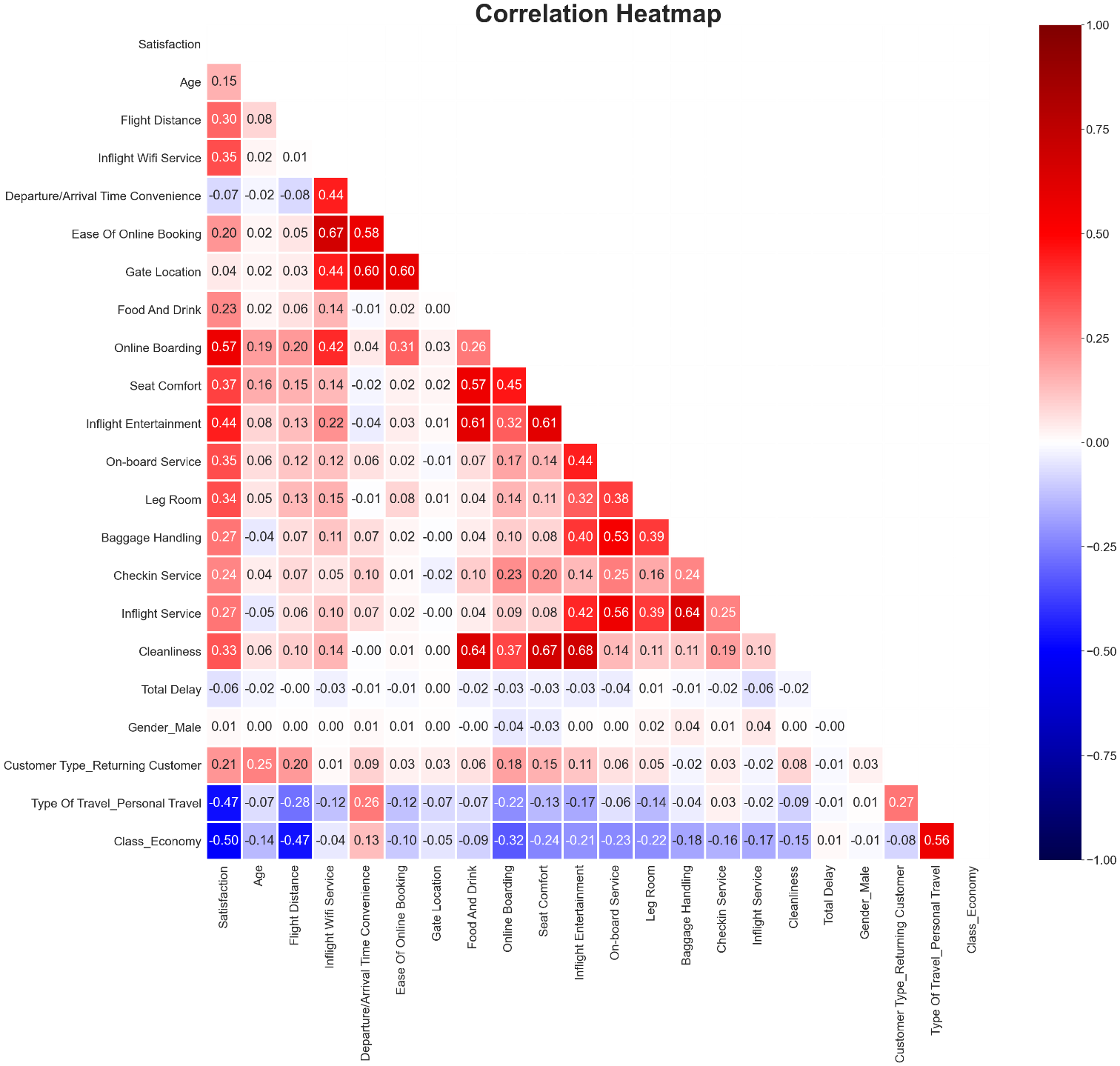
**CORELLATION HEATMAP:**

In statistics, **correlation**or **dependence**is any statistical relationship, whether causal or not, between two random variables or bivariate data. In the broadest sense **correlation** is any statistical association, though it commonly refers to the degree to which a pair of variables are linearly related.

Correlations are useful because they can indicate a predictive relationship that can be exploited in practice.

A **correlation heatmap** uses coloured cells, typically in a monochromatic scale, to show a 2D **correlation** matrix (table) between two discrete dimensions or event types. **Correlation heatmaps** are ideal for comparing the measurement for each pair of dimension values.

Correlation heatmap:

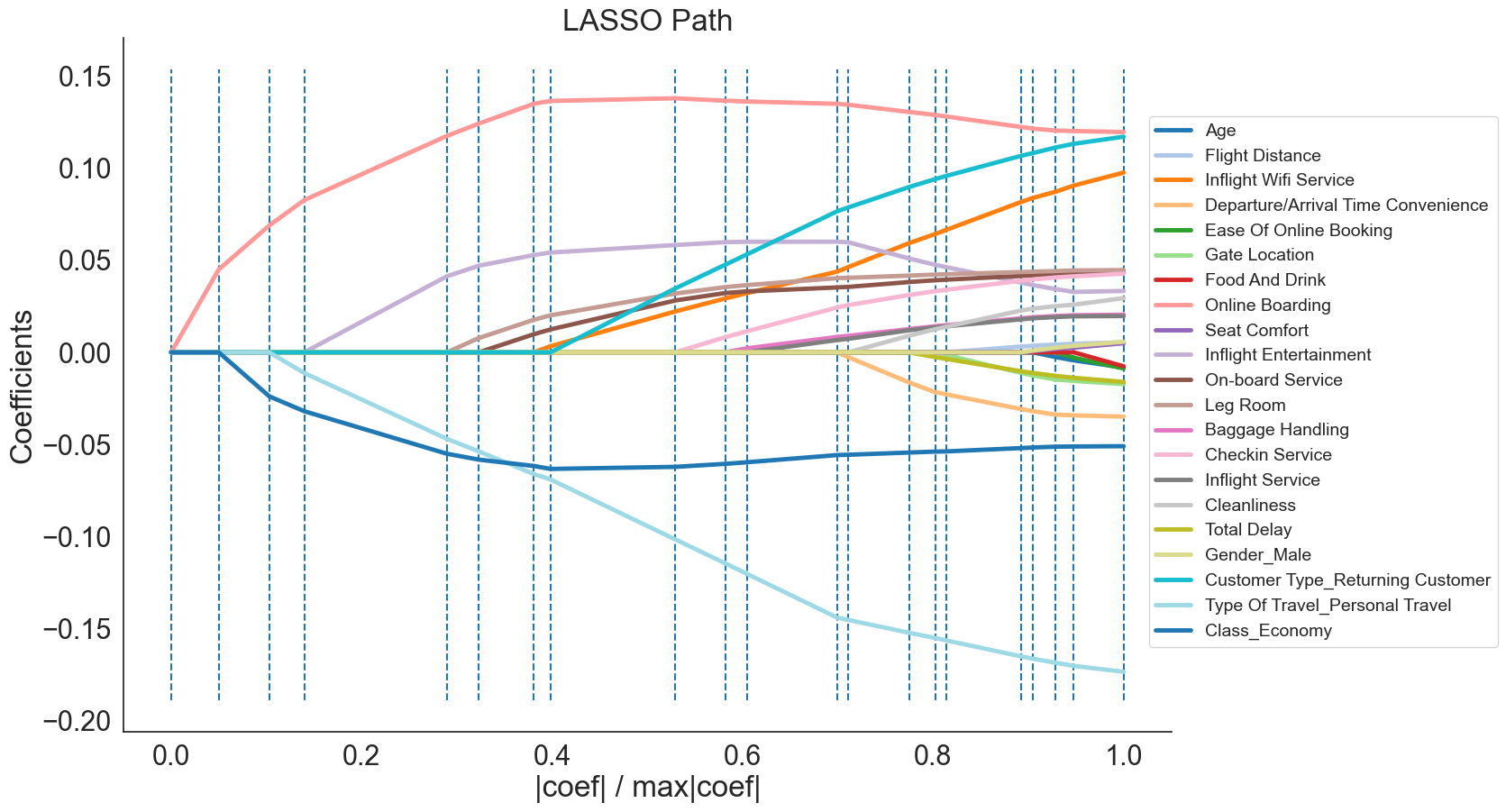


**LASSO PATH:**

In statistics and machine learning, **lasso** (**least absolute shrinkage and selection operator**; also **Lasso** or **LASSO**) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model.

Lasso was originally formulated for linear regression models. This simple case reveals a substantial amount about the estimator. These include its relationship to ridge regression and best subset selection and the connections between lasso coefficient estimates and so-called soft thresholding. It also reveals that (like standard linear regression) the coefficient estimates do not need to be unique if covariates are collinear.

LASSO path:



**FEATURES TO DROP**

After analysing the dataset thoroughly we understand which features play important role in passengers satisfaction and which does not.

Features to drop:

* Gender
* Age
* Gate Location
* Total Delay
* Flight Distance
* Departure/Arrival Time Convenience

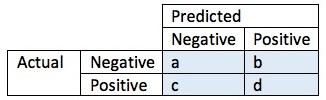
**AFTER ANALYSING THE DATASET THERE ARE FEW CONCLUSIONS**

* Airlines should highly focus on **inflight wi-fi experience**.
* **Ease of online booking** is important for business customers.
* The airlines should provide better on-board service.
* Need to concentrate more on economy class passengers.
* The customers are satisfied with the food and services.
* The customers like the seat comfort.

**PREDICTIVE ANALYSIS USING CONFUSION MATRIX**

A **confusion matrix**, in predictive analytics, is a two-by-two table that tells us the rate of false positives, false negatives, true positives and true negatives for a test or predictor.

Suppose your confusion matrix is a simple 2 by 2 table, given by:

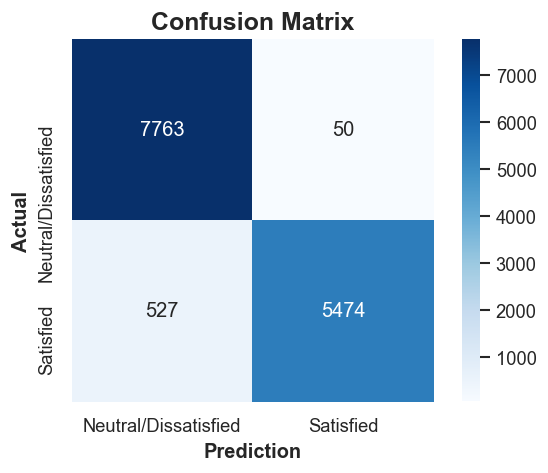


Here **a** is the number of true negatives, and **d** the number of true positives. **b** is the number of false positives, and **c** is the number of false negatives.

The **accuracy** of the prediction or test is defined as (a + d)/(a + c + d + e).

Therefore according to our confusion matrix:

The **accuracy** of the prediction or test is defined as (a + d)/(a + c + d + e).



**PRECISION = 99.1%**

**FUTURE DEVELOPMENTS**

* Can add more features for checking passengers satisfaction
* Collect more data and check satisfaction
* Improve the precision value
* Drop all unwanted features
* Personal review and feedback can also be added.
* Flight booking services can also be added

**CONCLUSION**

Take a passenger-first approach. Stop focusing on yourself, and focus on your customers instead. In order to shape a positive experience from the customer's point of view, your airline brand can follow these five steps:

Interact in a friendly, authentic, and personalized way.

Listen to the customer's unique situation and acknowledge their needs.

Give real-time, accurate information that empowers the customer to make decisions.

Go beyond the customer's service expectations so they will remember and share their great experience.

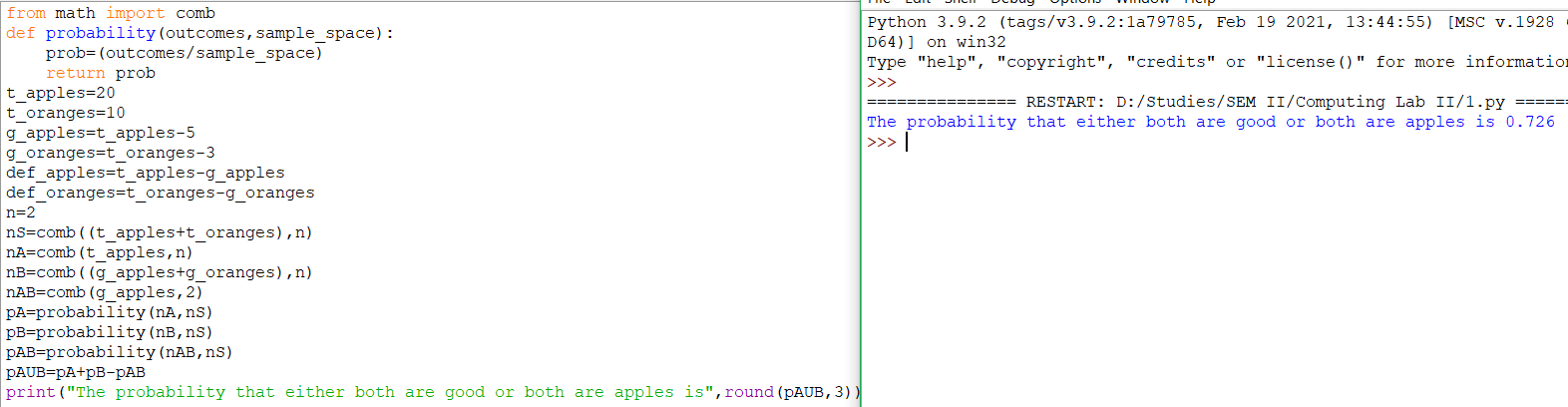
Remember the customer's preferences and anticipate their future needs.

Don’t make customers repeat themselves.

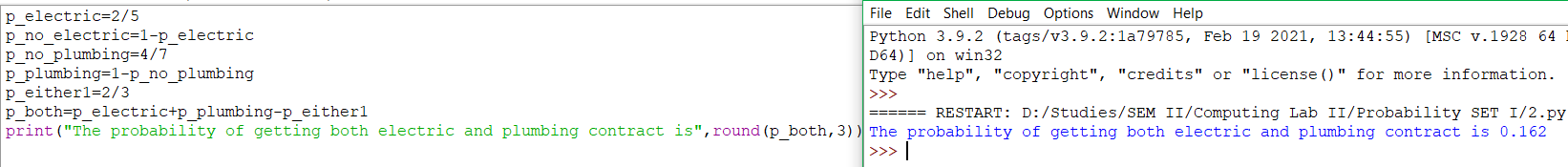
**ASSIGNMENT DOCUMENTATION**

**Date:16-03-2021 SET 1 (Probability)**

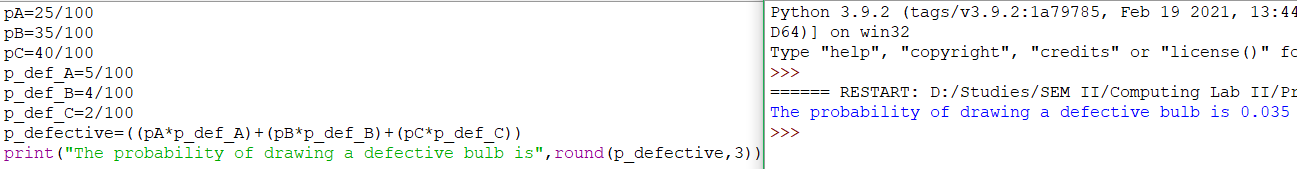
A bag contains 10 oranges and 20 apples out of which 5 apples and 3 oranges are defective .If a person takes out two at random, what is the probability that either both are good or both are apples ?



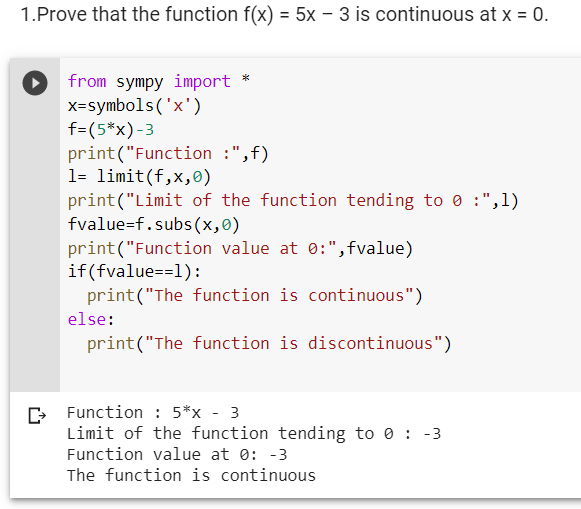
The probability that a person will get an electric contract is 2/5 and probability that he will not get plumbing contract is 4/7 . If the probability of getting at least one contact is 2/3, what is the probability of getting both ?

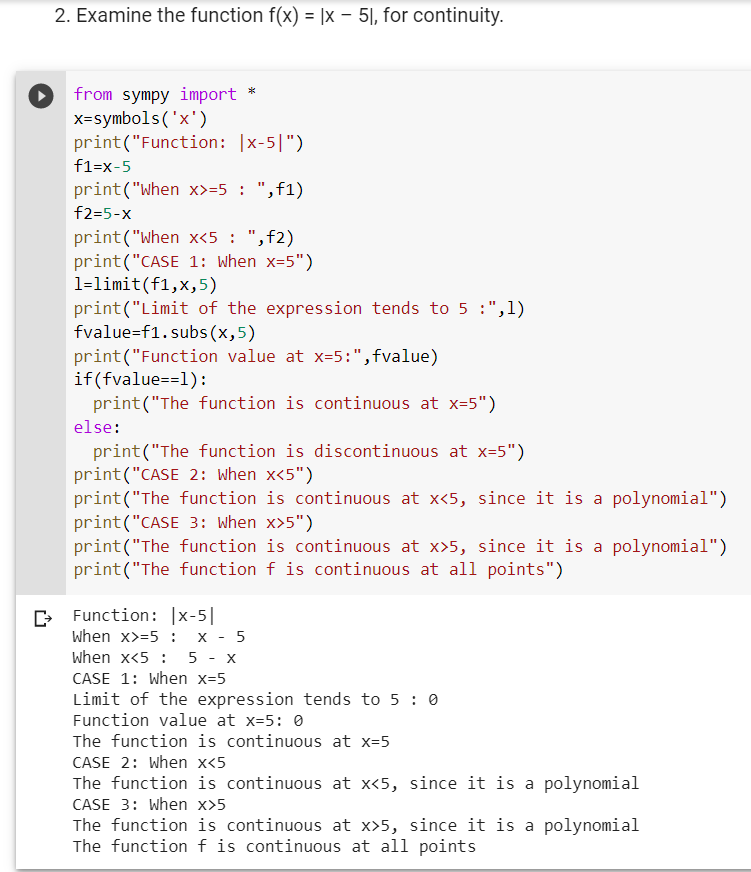


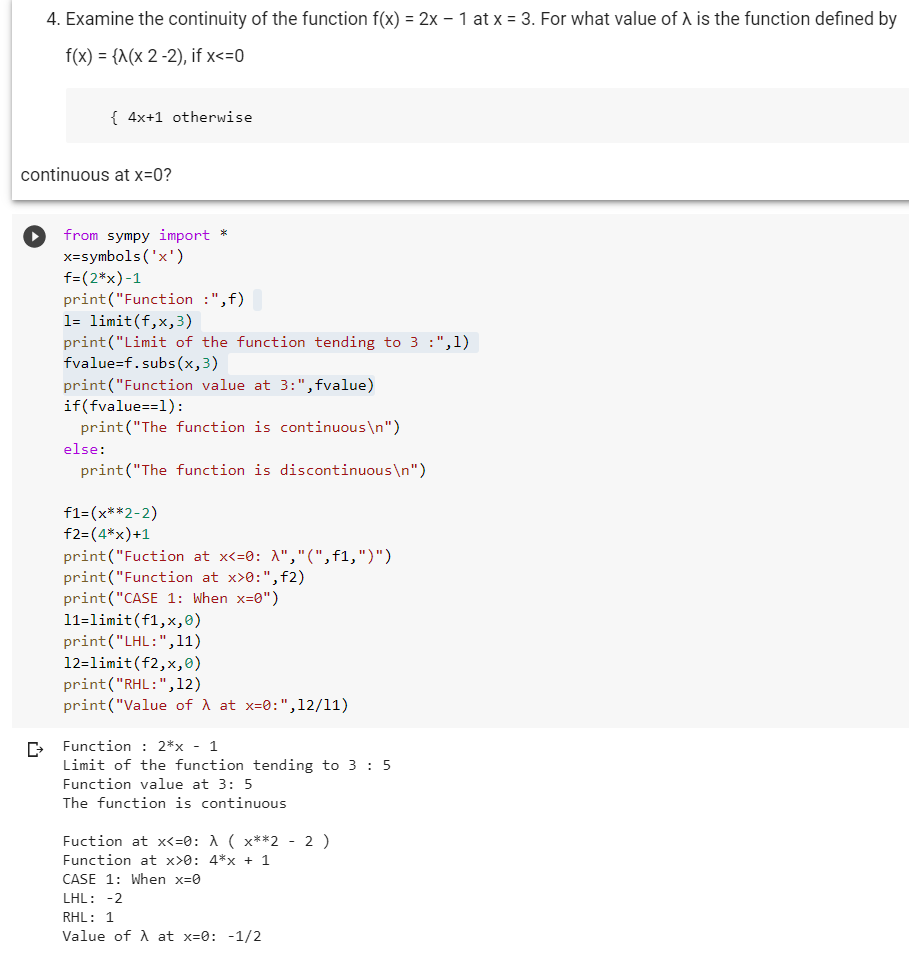
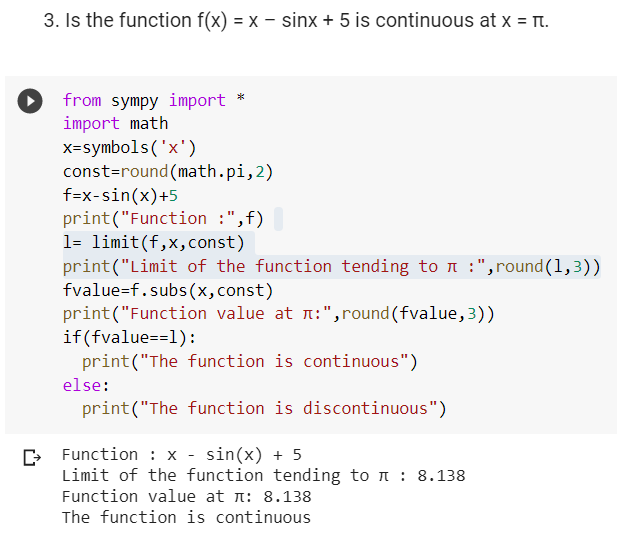
In a bulb factory, three machines namely A, B, C produces 25%, 35% and 40% of the total bulbs respectively . Of their output, 5, 4 and 2 percent are defective bulbs respectively . A bulb is drawn is drawn at random from products . What is the probability that bulb drawn is defective ?

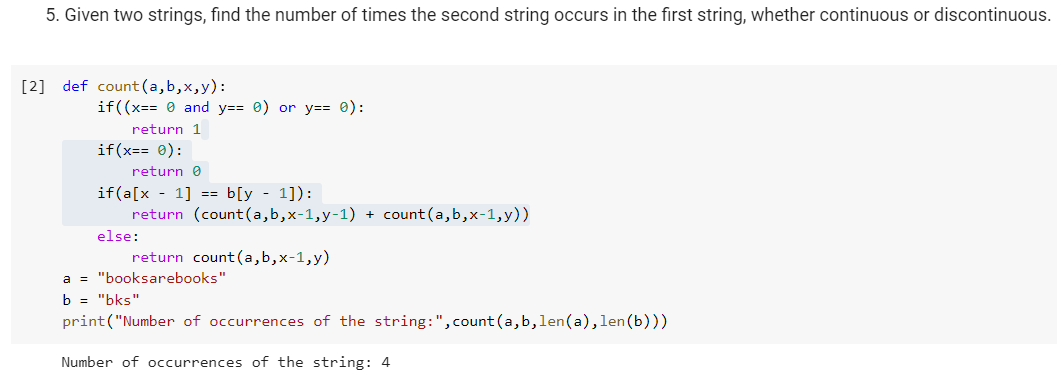


**Date:27-04-2021 Continuous and Discontinuous**









**Date:04-05-2021 PROBABILITY DISTRIBUTIONS**

Geometric Distributions:

1. A fair die is thrown until I get the number 4.

a. Find the probability distribution of the number of throws required.

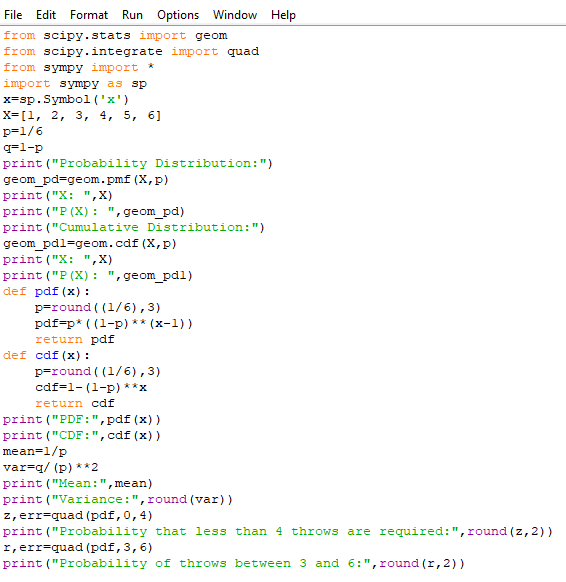
b. Find also the cumulative distribution function.

c. What are the mean and variance of the number of throws required.

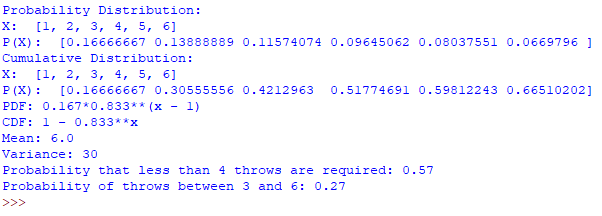
d. Find the probability that less than 4 throws are required.

e. Find the probability that the number of throws required lies between 3 and 6.

CODING:



OUTPUT:

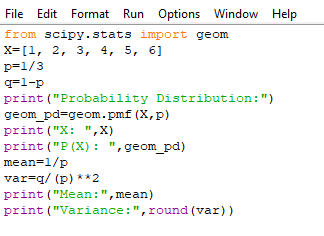


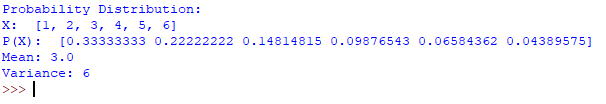
2. A fair die is thrown until I get the number 2 or 5.

a. Find the probability distribution of the number of throws required.

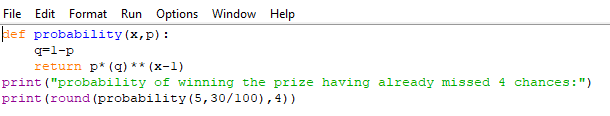
b. What are the mean and variance?

CODING:



OUTPUT: 

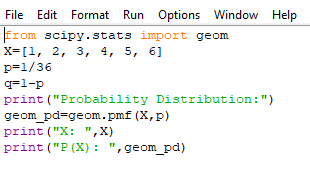
3. In an amusement fair, a competitor is entitled for a prize if he throws a ring on a peg from a certain distance. It is observed that only 30% of the competitors are able to do this. If someone is given 5 chances, what is the probability of his winning the prize when he has already missed 4 chances?

CODING: 

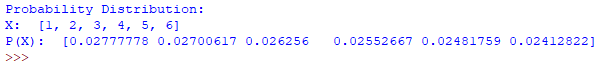
OUTPUT: 

4. Two dies are thrown until both dies show aces. Find the probability distribution of the number of throws required.

CODING:



OUTPUT:



Uniform Distribution:

1. State the pdf and cdf of the uniformly distributed random variable defined in the interval

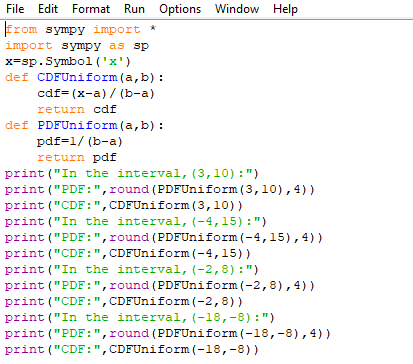
a. (3,10)

b. (-4,15)

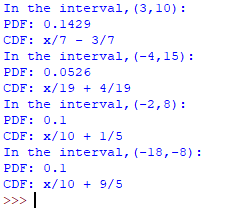
c. (-2,8)

d. (-18, -8)

CODING:



OUTPUT:



2. A random variable X is uniformly distributed in the interval (4,14). Find

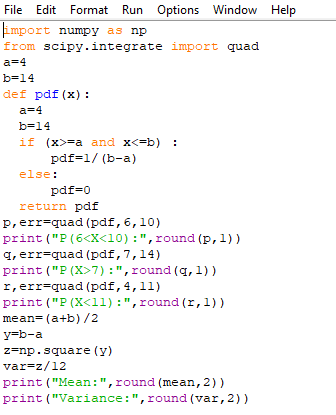
a. Pr(6 < X <10)

b. Pr(X > 7)

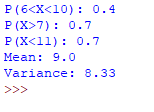
c. Pr(X < 11)

d. The mean and variance

CODING:



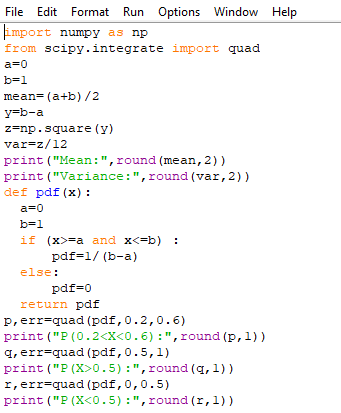
OUTPUT:



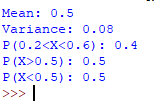
3. Derive the mean and variance of standard uniform distribution defined in the interval (0, 1).

Also find Pr(0.2 < X < 0.6) ,  Pr(X > 0.5) and Pr(X < 0.5)

CODING:



OUTPUT:



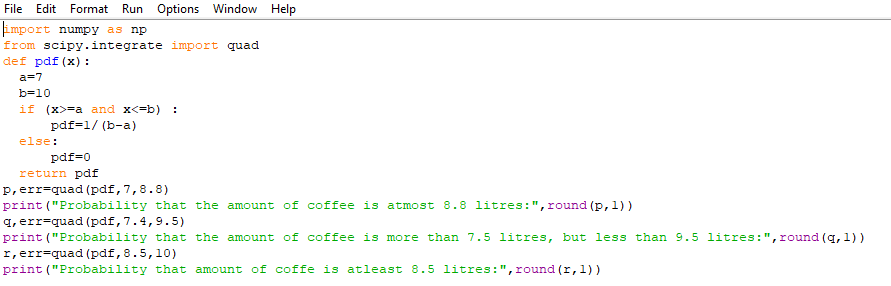
4. The daily amount of coffee, in litres, dispensed by a machine located in an airport lobby is a random variable *X* having a continuous uniform distribution with *A* = 7 and *B* = 10. Find the probability that on a given day the amount of coffee dispensed by this machine will be

(a) at most 8.8 litres;

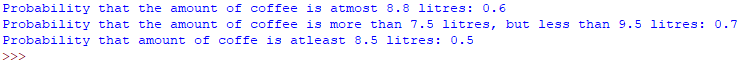
(b) more than 7.4 litres but less than 9.5 litres;

(c) at least 8.5 litres.

CODING:

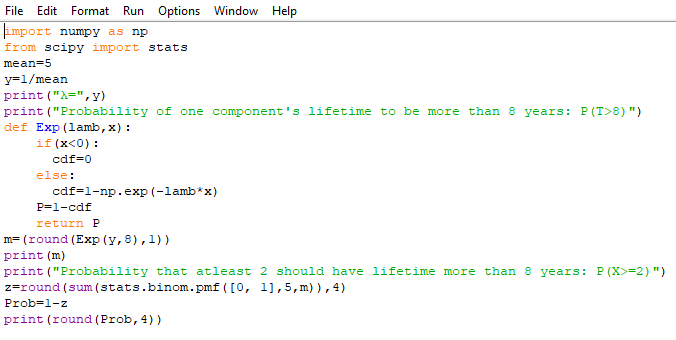


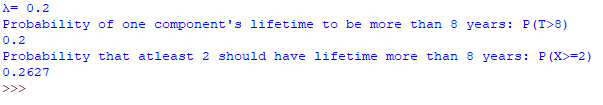
OUTPUT:



Exponential Distribution:

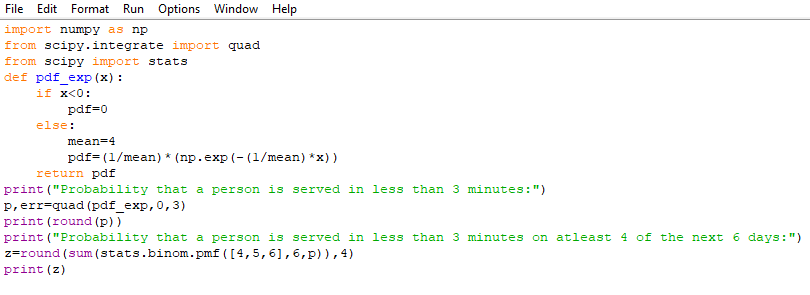
1. Suppose a system contains certain type of component whose lifetime is T , which is modelled by exponential distribution, its mean time to failure is 5. If 5 of these components are installed in different systems, what is the probability that at least 2 are still functioning at the end of 8 years?

CODING: 

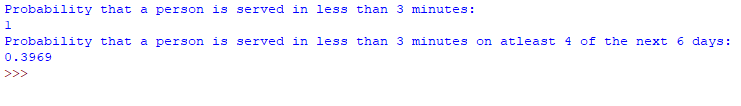
OUTPUT: 

2. The length of time for one individual to be served at a cafeteria is a random variable having an exponential distribution with a mean of 4 minutes. What is the probability that a person is served in less than 3 minutes on at least 4 of the next 6 days?

CODING:



OUTPUT:



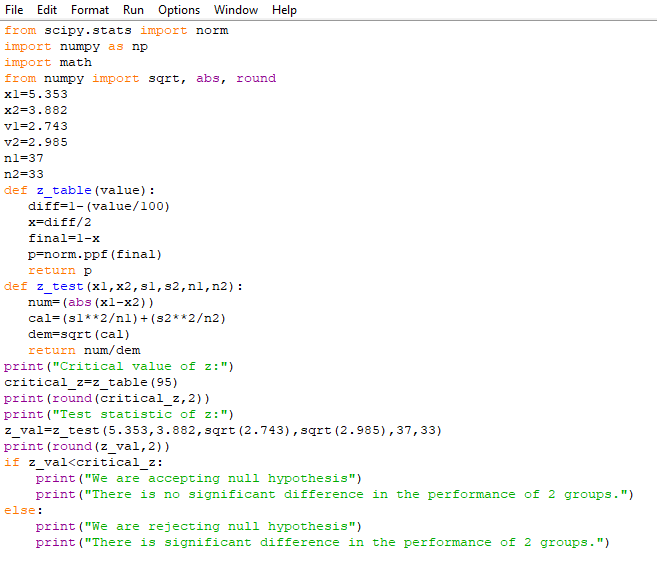
**Date:16-06-21 HYPOTHESIS TESTING**

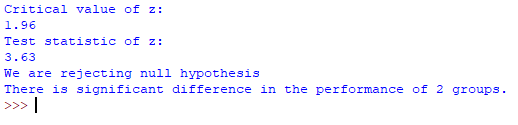
**1**.Consider  two groups A and B whose means and variances are as given below:

|  |  |  |  |
| --- | --- | --- | --- |
| **Group** | **n** | **Mean** | **Variance** |
| A | 37 | 5.353 | 2.743 |
| B | 33 | 3.882 | 2.985 |

Use Z test to test the significance of the difference between two means assuming population variances are unequal and not known.

**CODING:**

****

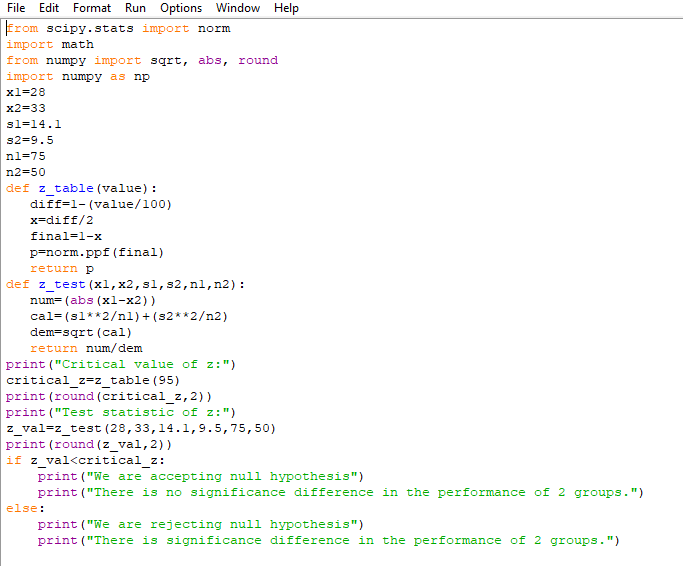
**OUTPUT: **

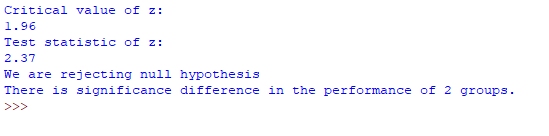
**2.** Test the significance of the difference between two means given:

Mean of group 1 = 28; standard deviation of group 1=14.1; size of sample =75

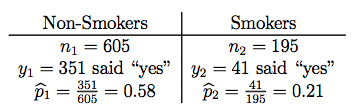
Mean of group 2 = 33; standard deviation of group 2=9.5; size of sample =50

**CODING:**

****

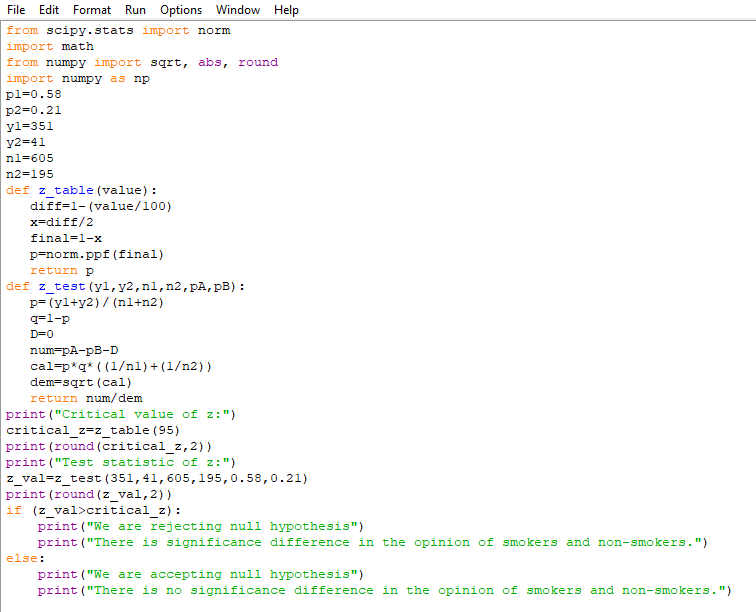
**OUTPUT: **

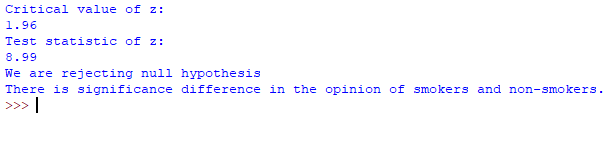
**3.** Time magazine reported the result of a telephone poll of 800 adult Americans. The question posed of the Americans who were surveyed was: "Should the federal tax on cigarettes be raised to pay for health care reform?" The results of the survey were:



Is there sufficient evidence at the *α* = 0.05 level, say, to conclude that the two populations — smokers and non-smokers — differ significantly with respect to their opinions?

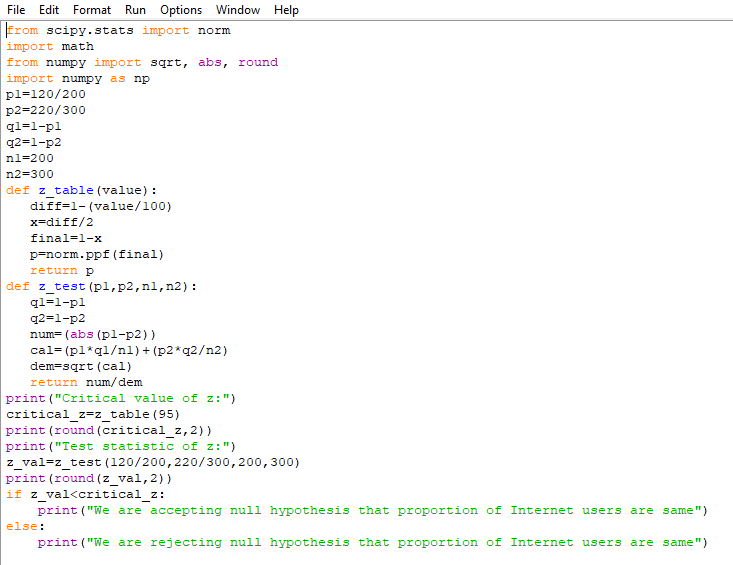
**CODING:**

****

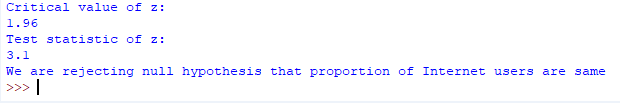
**OUTPUT: **

**4.** In a random sample of 200 households, 120 of them use Internet. In another random sample of 300 households, 220 of them use Internet. Is the hypothesis “proportion of Internet users in two groups same” is tenable?

**CODING:**

****

**OUTPUT:**

****