**Batch: B2 Experiment Number:1**

**Roll Number: 16010421073 Name: Keyur Patel**

**Aim of the Experiment: Data pre-processing by applying data normalization and data discretization**

**Output/Result:**

# Program/ Steps: 1) Importing modules

import pandas as pd

import numpy as np

# 2) Importing the excel file

import io

import pandas as pd

from google.colab import files

uploaded=files.upload()

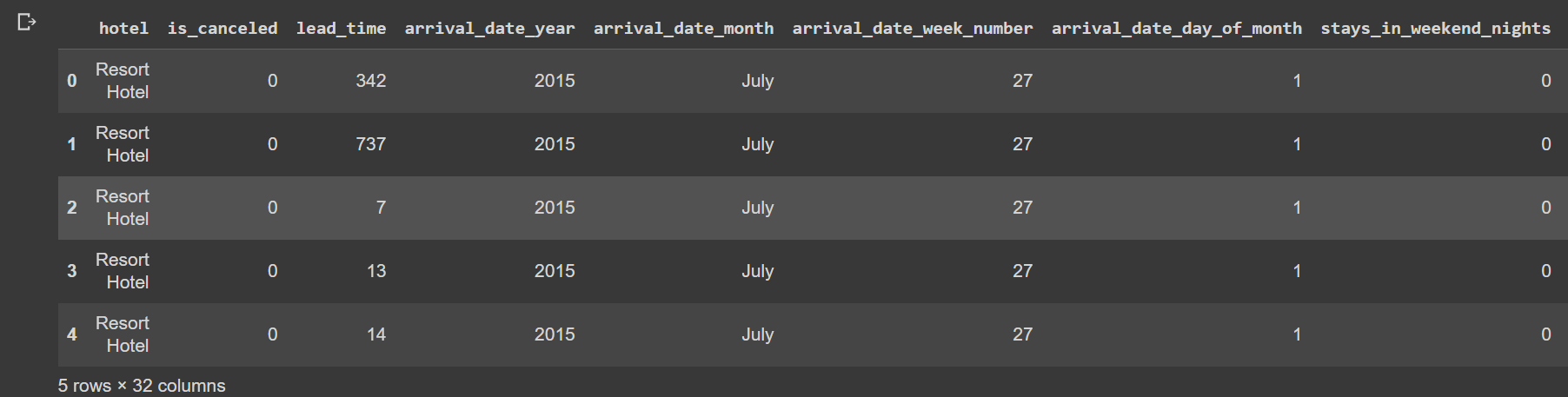
# 

**3) Storing the dataset in a variable df**

df=pd.read\_excel(io.BytesIO(uploaded['hotel\_booking.xlsx']))

**4) First 5 rows of the dataset**

df.head()

****

**5) Choosing an attribute for normalization**

**Attribute chosen – Lead time**

**Reason for choosing this attribute for normalization :**

Lead time is likely to have a wide range of values and can vary significantly. Normalizing this attribute will scale it to a common range, typically between 0 and 1, making it easier for machine learning algorithms to handle. Normalization ensures that each data point is relative to the entire range of values in the "lead\_time" attribute.

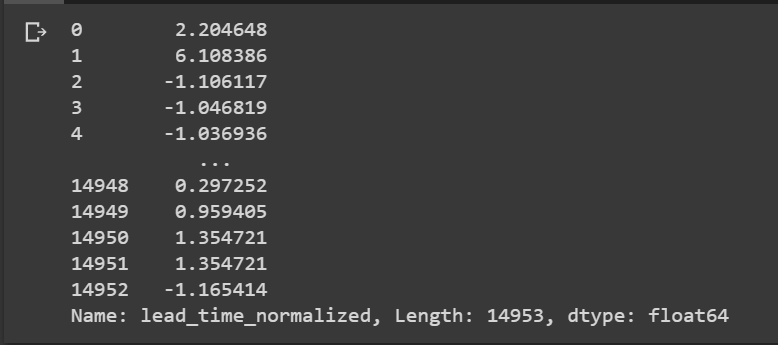
lead\_time = df['lead\_time']

mean\_lead\_time = lead\_time.mean()

std\_lead\_time = lead\_time.std()

df['lead\_time\_normalized'] = (lead\_time - mean\_lead\_time) / std\_lead\_time

print(df['lead\_time\_normalized'])



**6) Choosing an attribute for discretization**

**Attribute chosen – adr (average daily rate)**

**Reason for choosing this attribute for discretization :**

Discretizing the average daily rate can be beneficial if you want to group the values into specific categories or bins. For example, you can create bins for different price ranges like "low," "medium," and "high" to represent different levels of hotel room rates. Discretization can help reduce the impact of outliers and make the data more manageable and interpretable for certain types of machine learning algorithms.

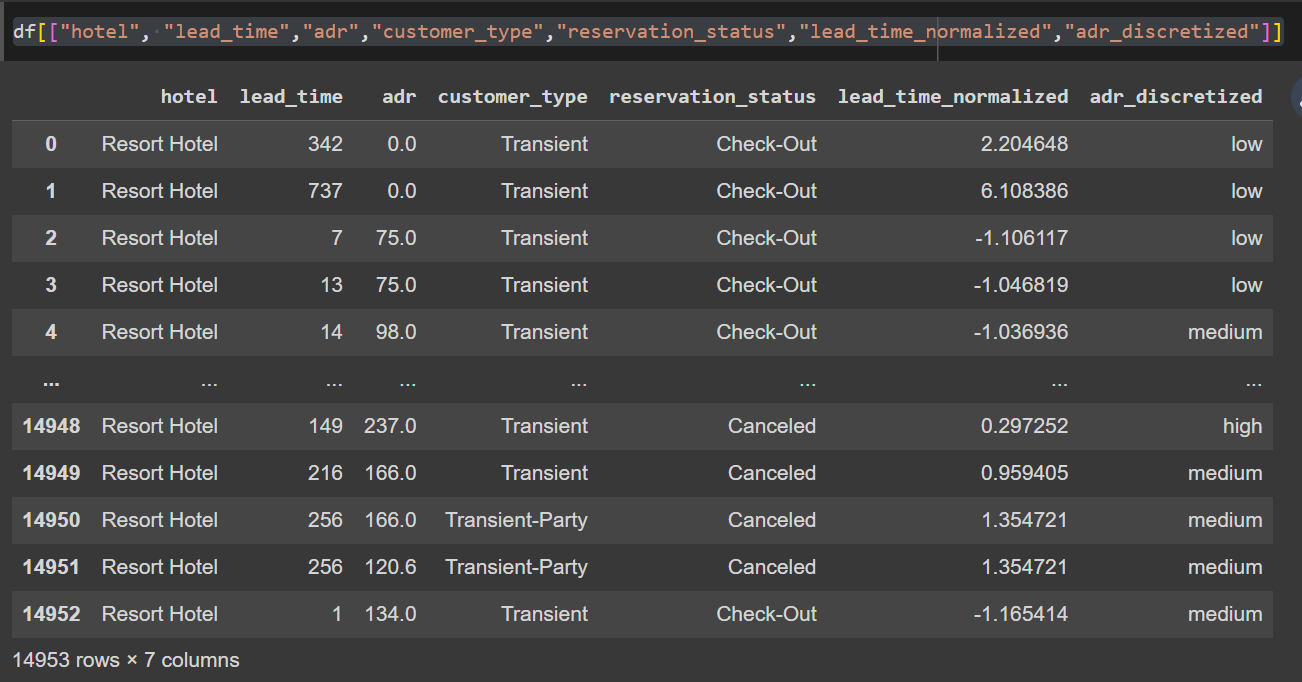
Creating number of bins: num\_bins = 5

labels = ['low', 'medium', 'high', 'very high', 'extremely high']

df['adr\_discretized'] = pd.cut(df['adr'], bins=num\_bins, labels=labels)

df[["hotel","lead\_time","adr","customer\_type","reservation\_status","lead\_time\_normalized","adr\_discretized"]]

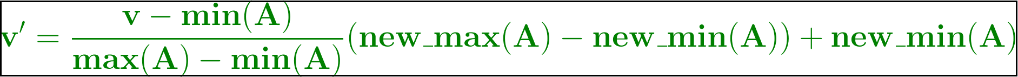
**Both lead\_time\_normalized and adr\_discretized columns were added to the dataset**



**Questions:**

1. **Explain with example Min-Max normalization technique.**

**Ans:** Min-Max normalization, also known as feature scaling, is a data normalization technique used to transform numerical data into a specific range, typically between 0 and 1. It linearly scales the original values to fit within this range based on the minimum and maximum values of the attribute. This normalization is given by the formula:



To understand the formula, here is an example. Suppose a company wants to decide on a promotion based on the years of work experience of its employees. So, it needs to analyze a database that looks like this:

|  |  |
| --- | --- |
| **Employee Name** | **Years of Experience** |
| ABC | 8 |
| XYZ | 20 |
| PQR | 10 |
| MNO | 15 |

* The minimum value is 8
* The maximum value is 20

As this formula scales the data between 0 and 1,

* The new min is 0
* The new max is 1

Here, V stands for the respective value of the attribute, i.e., 8, 10, 15, 20

After applying the min-max normalization formula, the following are the **V’**values for the attributes:

* For 8 years of experience: **v’= 0**
* For 10 years of experience: **v’ = 0.16**
* For 15 years of experience: **v’ = 0.58**
* For 20 years of experience:**v’ = 1**

So, the min-max normalization can reduce big numbers to much smaller values.  This makes it extremely easy to read the difference between the ranging numbers.

# Outcomes: CO1: Comprehend basics of machine learning

**Conclusion (based on the Results and outcomes achieved):**

Through this experiment we learned the concepts of data preprocessing by by applying data normalization and data discretization.

**References:**

Books/ Journals/ Websites:

* 1. Han, Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann 3nd Edition