|  |
| --- |
| **National University of Computer and Emerging Sciences** |
| Lab Manual 6  “ETL with Python” |
|  |
| Data Warehousing and Data Mining |
|  |
|  |
|  |

|  |  |
| --- | --- |
| Course Instructor | Ishaq Raza |
| Lab Instructor(s) | Farhan Azhar  Nouman Ali |
| Section | CS |
| Semester | Fall 2019 |

|  |
| --- |
|  |

Department of Computer Science

FAST-NU, Lahore, Pakistan

**Task 1**

Complete‘Python Data Transformation Tutorial’

**Dataset**: train\_task1.csv

**Task 2**

**Dataset:** train.csv

**Perform the following tasks on the data:**

* Drop columns ‘Ticket’ and ‘Fare’
* Fill the missing values of ‘Cabin’ column. Explore methods of handling missing data with non numerical values.
* Split the ‘Name’ column into three separate columns: ‘FirstName’, ‘MiddleName’ and ‘LastName’.
* Fill the missing values of ‘Age column. Explore methods of handling missing data with numerical values.
* Add a New Column named ‘AgeGroup’ by breaking up the ‘Age’ in 4 categories. The grouping is to be done on the following basis:
* 0 – 13 years: age\_child
* 13 – 25 years: age\_young
* 25 – 50 years: age\_adult
* 50 – 80 years: age\_old
* Add a New Column named ‘SibSpGroup’ by grouping it based on the distribution of the values. The grouping is to be done on the following basis:
* 0 count: sibsp\_none
* 1 or 2: sibsp\_normal
* 3 – 8: sibsp\_high
* Add a New Column named ‘ParchGroup’ by grouping it based on the distribution of the values. The grouping is to be done on the following basis:
* 0 : parch\_none
* 1: parch\_normal
* 2 – 6: parch\_high
* Decode Survived Column: Set values for the ‘Survived’ column from 0 to No and 1 to Yes.

**Summarization**

Each summarization result should be shown separately.

* Find ‘Survived’ wise Sum of Ages of Passengers i.e. The result should have city and sum of ages of customers.
* Find ‘AgeGroup’ Wise number of Passengers. I.e. The result should have AgeGroup and number of Passengers column.
* Find ‘SibSp’ Wise number of Passengers. I.e. The result should have SipSpGroup and number of Passengers column.
* Find ‘ParchGroup’ Wise number of Passengers. I.e. The result should have ParchGroup and number of Passengers column.

**Reference Material**

* **Use the following code to upload the csv file and load it:**

import pandas as pd

from google.colab import files

uploaded = files.upload()

data = pd.read\_csv('train.csv')

* **Intro\_to\_Python\_Helping\_Lecture.ppt**