**FDS Lab**

**Instructions for Practical Journal:**

1. Add at least two examples for each concept with two different datasets.
2. After completing your practical, please show and check your notebook.
3. Once finalized, you can upload your notebook on your Git repository.
4. Complete your task on or before 4th March 10 pm only.
5. Once the link of all notebooks are added to the file, you are expected to complete the remaining practicals and upload the notebook on your own Git Repository.
6. Your own repository will be considered as your lab work for FDS.
7. For practical dates refer <https://docs.google.com/document/d/1GfQLuMe2p9A2tmYF1lAWKXWeDybvlmyhSPSdieVinwY/edit>
8. Use following Link to paste your Github accounts link: <https://docs.google.com/document/d/1iN7QPwP6MTEzsa20EB2IN0PIK66v5KMR98yVFQAGSVs/edit?usp=sharingk>
9. Prepare at least two examples for each concept for each practical/concept.

**Instructions for Internals:**

1. Study the topic thoroughly with theory and practical for the Internal exam.
2. Prepare at least five different examples for each concept for the internal exam.
3. The schedule for Internals will be displayed into the classroom.
4. Prepare your notebook with the following:
   1. Importance/role of topic in statistical modelling and machine learning.
   2. Solve any file examples related to the topic.

**Index**

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Title** | **Date** |
| 1 | Python Basics   1. Jupyter Notebook, Google Colab & Github 2. Python 3. Numpy 4. Pandas 5. Matplotlib and Seaborn |  |
| 1. Jupyter Notebook, Google Colab and Github    1. What is Jupyter Notebook, Google Colab and Github.    2. Create a Github Account and repository for FDS    3. Create a simple Jupyter notebook with a simple python example and save it into the Github account. 2. Python   Perform Python basic concepts using Jupyter Notebook and upload the notebook on your FDS Git repository.  <https://www.w3schools.com/python/>   1. Numpy   Perform Numpy basic concepts using Jupyter Notebook and upload the notebook on your FDS Git repository.  <https://www.w3schools.com/python/numpy_intro.asp>   1. Pandas   Perform Pandas basic concepts using Jupyter Notebook and upload the notebook on your FDS Git repository.  <https://www.w3schools.com/python/pandas/default.asp>   1. Matplotlib & Seaborn   Demonstrate the Matplotlib & Seaborn basic concepts with some examples using Jupyter Notebook and upload the notebook on your FDS Git repository.  <https://www.w3schools.com/python/matplotlib_intro.asp> | |
| 2 | Data Collection and Datasets   1. From .csv Files 2. From Excel Files 3. From SQL Files |
| Create the csv ,excel and sql database and table for storing the following data. Also load and display the data from all these sources into Google colab notebook.  **EmpData**   |  |  |  |  |  | | --- | --- | --- | --- | --- | | EmpId | FName | LName | Designation | Basic Salary | | E101 | Shweta | Subnis | ProjectEngineer | 40000 | | E102 | Manisha | Perdesi | Tester | 50000 | | E103 | Madhu | Mali | SrSoftwareEngineer | 60000 | | E104 | Anita | Rai | WebDeveloper | 70000 | | |
| 3 | Data Preprocessing (Data Preparation and Data Transformations)   1. Missing Values 2. Group By 3. Merge 4. Transform 5. Apply |  |
| Examples from...   1. <https://sergilehkyi.com/extracting-and-transforming-data-in-python/> 2. <https://www.analyticsvidhya.com/blog/2020/03/understanding-transform-function-python/> | |
| 4 | Descriptive Statistics   1. Measures of Central Tendency    1. Mean    2. Median    3. Mode 2. Measures of Variability (    1. Standard Deviation    2. Variance    3. Min and Max Variables    4. Skewness and Kurtosis    5. Range    6. Percentiles/Quartiles    7. Box Plots |  |
| **Measures of Central Tendency**   1. <https://github.com/PacktPublishing/Principles-of-Data-Science-Second-Edition/blob/master/Chapter02/Chapter_2%20.ipynb> 2. <https://github.com/gedeck/practical-statistics-for-data-scientists/blob/master/python/notebooks/Chapter%201%20-%20Exploratory%20Data%20Analysis.ipynb>   **Measures of Variability**   1. [https://colab.research.google.com/github/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/Index.ipynb#scrollTo=t-1jH lyFE](https://colab.research.google.com/github/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/Index.ipynb#scrollTo=t-Ca1jHZlyFE) 2. <https://colab.research.google.com/github/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/02.04-Computation-on-arrays-aggregates.ipynb#scrollTo=L3xu2NEmlag5> 3. <https://github.com/gedeck/practical-statistics-for-data-scientists/blob/master/python/notebooks/Chapter%201%20-%20Exploratory%20Data%20Analysis.ipynb> 4. <https://www.dataquest.io/blog/basic-statistics-with-python-descriptive-statistics/> | |
| 5 | Exploratory Data Analysis   1. Rectangular Data 2. Non-Rectangular Data |  |
| 1. <https://github.com/gedeck/practical-statistics-for-data-scientists/blob/master/python/notebooks/Chapter%201%20-%20Exploratory%20Data%20Analysis.ipynb> 2. For time series data<https://medium.com/analytics-vidhya/how-to-guide-on-exploratory-data-analysis-for-time-series-data-34250ff1d04f> | |
| 6 | Visualization   1. Frequency Table 2. Scatter Plot 3. Line Chart 4. Bar Chart 5. Histogram 6. Density Curve |  |
| **Frequency Plot**   1. <https://github.com/gedeck/practical-statistics-for-data-scientists/blob/master/python/notebooks/Chapter%201%20-%20Exploratory%20Data%20Analysis.ipynb> 2. <https://github.com/PacktPublishing/Principles-of-Data-Science-Second-Edition/blob/master/Chapter01/Chapter_1.ipynb> 3. <https://colab.research.google.com/github/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/04.00-Introduction-To-Matplotlib.ipynb> | |
| 7 | Exploratory Data Analysis of Binary and Categorical Variables |  |
| 1. <https://github.com/gedeck/practical-statistics-for-data-scientists/blob/master/python/notebooks/Chapter%201%20-%20Exploratory%20Data%20Analysis.ipynb> | |
| 8 | Generation of Random Variable   1. Discrete Random Variable 2. Continuous Random Variable |  |
| 1. <https://github.com/PacktPublishing/Principles-of-Data-Science-Second-Edition> 2. W3School | |
| 9 | Probability Distributions   1. Discrete Random Distributions    1. Bernoulli Distribution    2. Uniform Distribution    3. Binomial Distribution 2. Continuous Random Distributions    1. Normal Distribution    2. Exponential Distribution    3. Weibull Distribution    4. Standard Normal and QQ Plots    5. Long Tailed Distribution |  |
| 1. <https://github.com/gedeck/practical-statistics-for-data-scientists/tree/master/python/notebooks> | |
| 10 | Data & Sampling Distributions   1. Sampling Distribution 2. Bootstrap 3. Central Limit Theorem 4. Law of Large Numbers 5. Plotting Estimates (Mean and Standard Deviation) 6. Defining Confidence Interval |  |
| 1. <https://github.com/gedeck/practical-statistics-for-data-scientists/tree/master/python/notebooks> 2. <https://github.com/PacktPublishing/Principles-of-Data-Science-Second-Edition> | |
| 11 | Inferential Statistics   1. Calculating Expected Parameters (Mean and Standard Deviation) 2. Estimating population parameters | Divya |
| 1. <https://github.com/PacktPublishing/Principles-of-Data-Science-Second-Edition> | |
| 12 | Probability   1. Types of Events 2. Rules of Probability 3. Permutations and Combinations 4. Probability Likelihood 5. Probability Expectation 6. Probability Estimation |  |
| 1. <https://github.com/PacktPublishing/Principles-of-Data-Science-Second-Edition> | |
| 13 | Significance and Hypothesis Testing   1. Significance Testing 2. Hypothesis Testing 3. t-Test 4. A/B Testing 5. Type I and Type II Errors 6. One-Sample Test 7. Two-Sample Test |  |
| 1. <https://github.com/gedeck/practical-statistics-for-data-scientists/tree/master/python/notebooks> | |
| 14 | Supervised Learning   1. Univariate Linear Regression 2. Bivariate Linear Regression 3. Multivariate Linear Regression 4. Logistic Regression 5. Classification 6. Decision Tree |  |
| 1. <https://colab.research.google.com/github/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/05.00-Machine-Learning.ipynb> 2. <https://github.com/PacktPublishing/Principles-of-Data-Science-Second-Edition> | |
| 15 | Unsupervised Learning   1. K-means Clustering 2. Association Rules |  |
| 1. <https://colab.research.google.com/github/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/05.00-Machine-Learning.ipynb> 2. <https://github.com/gedeck/practical-statistics-for-data-scientists/tree/master/python/notebooks> 3. <https://github.com/PacktPublishing/Principles-of-Data-Science-Second-Edition> | |
| 16 | Time Series Analysis |  |
| 1. <https://colab.research.google.com/github/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/03.11-Working-with-Time-Series.ipynb> 2. Chapter 11 from Python for Data Analysis: Data Wrangling with Pandas, Numpy, and IPython, Wes McKinney. | |
| **References**   1. <https://www.oreilly.com/library/view/python-data-science/9781491912126/> 2. <https://colab.research.google.com/github/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/Index.ipynb#scrollTo=t-Ca1jHZlyFE> 3. <https://github.com/PacktPublishing/Principles-of-Data-Science-Second-Edition> 4. <https://github.com/gedeck/practical-statistics-for-data-scientists/tree/master/python/notebooks> | |

**Practical 1 - Python, Numpy and Google Colab Basics**

**Aim:** To Learn Python, Numpy and Google Colab Basics

**Problem Statement**

Q. Create a notebook in Google Colab and execute all basic statements of Python and Numpy.

**Home Work**

Revise and Practice Python and Numpy Concepts.

**Reference** : https://www.w3schools.com/python/

**Practical 2- Data Collection**

**Aim**: To learn how to load/store the data from/into Google Colab from cav, excel and SQL files.

**Problem Statement**

Q. Create a notebook in Google Colab to load/store the data from/into Google Colab from cav, excel and SQL files.

**Home Assignment**

Execute the following example in Google Colab

<https://datatofish.com/pandas-dataframe-to-sql/>

**Code for Pract 2**

1. Load data from csv file and display data without headers

import pandas as pd

Location = "datasets/smallgradesh.csv"

df = pd.read\_csv(Location, header=None)

df.head()

2. Load data from csv file and display data with headers

import pandas as pd

Location = "datasets/gradedata.csv"

df = pd.read\_csv(Location)

df.head()

3. Loading Data from CSV File and Adding Headers

import pandas as pd

Location = "datasets/smallgrades.csv"

# To add headers as we load the data...

df = pd.read\_csv(Location, names=['Names','Grades'])

# To add headers to a dataframe

df.columns = ['Names','Grades']

df.head()

4. Exporting a Dataset to CSV

import pandas as pd

names = ['Bob','Jessica','Mary','John','Mel']

grades = [76,95,77,78,99]

GradeList = zip(names,grades)

df = pd.DataFrame(data = GradeList, columns=['Names','Grades'])

df.to\_csv('studentgrades.csv',index=False,header=False)

5. Creating a Dataset for the Exercise

import pandas as pd

names = ['Bob','Jessica','Mary','John','Mel']

grades = [76,95,77,78,99]

bsdegrees = [1,1,0,0,1]

msdegrees = [2,1,0,0,0]

phddegrees = [0,1,0,0,0]

Degrees = zip(names,grades,bsdegrees,msdegrees,phddegrees)

columns = ['Names','Grades','BS','MS','PhD']

df = pd.DataFrame(data = Degrees, columns=column)

df

6. Loading Data from Excel File and changing column names

import pandas as pd

Location = "datasets/gradedata.xlsx"

df = pd.read\_excel(Location)

df.columns = ['first','last','sex','age','exer','hrs','grd','addr']

df.head()

7. Exporting a Dataframe to Excel

import pandas as pd

names = ['Bob','Jessica','Mary','John','Mel']

grades = [76,95,77,78,99]

GradeList = zip(names,grades)

df = pd.DataFrame(data = GradeList,

columns=['Names','Grades'])

writer = pd.ExcelWriter('dataframe.xlsx', engine='xlsxwriter')

df.to\_excel(writer, sheet\_name='Sheet1')

writer.save()

8. Load Data from sqlite

import pandas as pd

from sqlalchemy import create\_engine

# Connect to sqlite db

db\_file = r'datasets/gradedata.db'

engine = create\_engine(r"sqlite:///{}" .format(db\_file))

sql = 'SELECT \* from test'

'where Grades in (76,77,78)'

sales\_data\_df = pd.read\_sql(sql, engine)

sales\_data\_df

9. Finding the Table Names

sql = "select name from sqlite\_master"

"where type = 'table';"

10. Create Dataset to Save

import pandas as pd

names = ['Bob','Jessica','Mary','John','Mel']

grades = [76,95,77,78,99]

GradeList = zip(names,grades)

df = pd.DataFrame(data = GradeList,

columns=['Names', 'Grades'])

df

11. To export it to SQL

import os

import sqlite3 as lite

db\_filename = r'mydb.db'

con = lite.connect(db\_filename)

df.to\_sql('mytable',

con,

schema=None,

if\_exists='replace',

index=True,

index\_label=None,

chunksize=None,

dtype=None)

con.close()

**Practice Questions**

Q. Create the csv,excel and sql database and table for storing the following data. Also load and display the data from all these sources into Google colab notebook.

EmpData

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| EmpId | FName | LName | Designation | Basic Salary |
| E101 | Shweta | Subnis | ProjectEngineer | 40000 |
| E102 | Manisha | Perdesi | Tester | 50000 |
| E103 | Madhu | Mali | SrSoftwareEngineer | 60000 |
| E104 | Anita | Rai | WebDeveloper | 70000 |

**Pract 3**