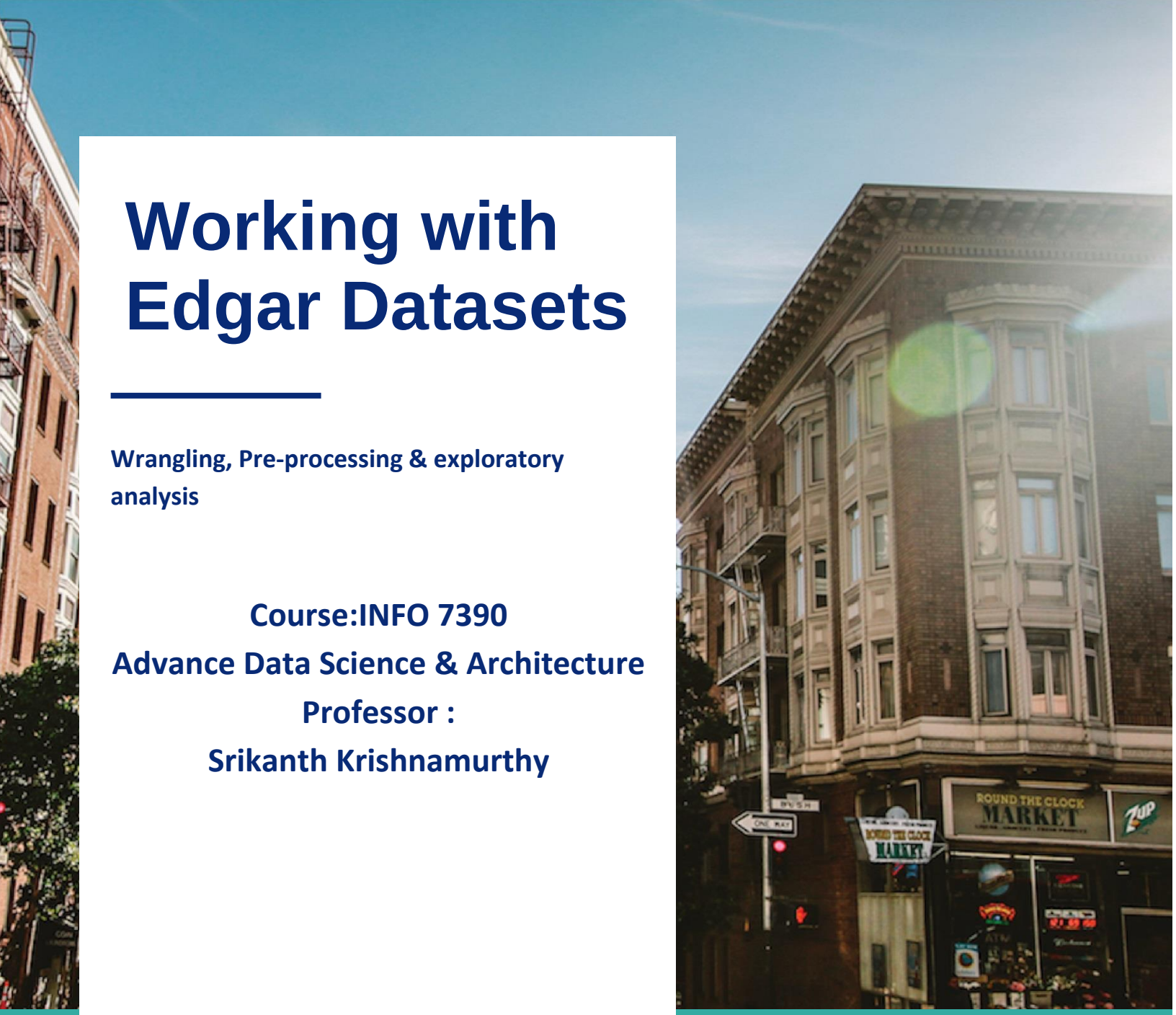

Working with Edgar Datasets

Wrangling, Pre-processing & exploratory analysis

Course:INFO 7390

Advance Data Science & Architecture

Professor :

Srikanth Krishnamurthy

FEBRUARY 16

COMPANY NAME

Authored by:1] Aditya Pawar

2]Dhairya Jaiswal

3]Ranga Chari Vinjamuri

Logo Name

Table of Content

Programming Language & Libraries used

We are using Python to solve the problem. Following is the list of python libraries used in the code:

Beautifulsoup	**Web scrapping and prettifying**
urllib	**Web scrapping**
urllib.request	**To check the response of the URL**
csv	**To write into csv format**
logging	**To track log details**
os	**To create and delete directories and files**
zipfile	**Creating and Handling zipfiles**
boto3	**For handling of AWS S3**
pandas	**Data manipulation**
shutil	**For removing files**
Glob	**To read the files**

Logging

Logging is a means of tracking events that happen when some software runs. The software's developer adds logging calls to their code to indicate that certain events have occurred. An event is described by a descriptive message which can optionally contain variable data.
We are generating log files at every stage of the code, we are generating timestamps, level names and customized error message.
We have use 4 levels of Logs:

1. DEBUG: It is used to return information to the log file that is created. Report events that occur during normal operation of a program
2. Error: Due to a more serious problem, the software has not been able to perform some function.
3. Warning: Issue a warning regarding a particular runtime event.
4. Info: Confirmation that things are working as expected.

We are also outputting the same log message on the console using the Sys method of the logging library.

Data Wrangling Edgar data from text files

The objective of this assignment is to extract tables from 10Q fillings using Python.

1. We need to give 5 arguments to access the code they are as follows:
   A) Cik
   B) Account_Number
   C)Amazon Access_key
   D)Amazon Secret Access_key
   E)S3 Location
2. We are using urllib and urllib.request libraries to navigate and generate page specific URL.
3. We are using BeautifulSoup libraries to handle html tags in Python and also to prettify the data content.
4. Once URL is opened using the urllib.request.get() we have a foreach loop to run through all the <a> tags and <href> tags of the html page. If there is no 10-Q file the program exists.
5. Once we get and open the 10-Q url link we extract the data by looking for table content which are part of the tag.
6. We clean the fetched data as it contains special characters like '$' and '%' values inside the table data. So we have to iterate through the tables and look for patterns.
7. We created a refined table variable were we stored the clean table data which is inside <td> tags. We remove unwanted characters like '\n' and '\xa0' characters.
8. After data is cleaned every table is stored as individual csv file which is inside extracted_csv folder.
9. The program zips the folder and puts it inside Problem_1.zip.
10. We create a bucket and upload a zip file .The bucket name is always unique.
11. Exception Handling
    - If amazon access keys are not provided or are invalid, the program would log an error and exit.
    - If location is not provided it creates a bucket in uswest2 location (default).

Missing Data Analysis

1. Import all the required libraries for the code and also initialize the log file as done in the previous problem.
2. If the directories are not present create one to put the zip files and then clean the required directories.
3. The input year must be between 2003 to 2018. If the year is not between this range the program will exit.There is no data for the first quarter of 2003.
4. To the extract the csv ,unzip the downloaded log files.
5. Load all the csv files into individual dataframe.
6. For every individual log file:
   a) Count the null values for every variable.
   b) Check if the variables *idx*, *norefer* and *noagent* have any other values except 0 and 1.
7. For each log file we have handled the missing value as the following:
   a. The variables cik, ip, accession, data or time are most important for the log file in Edgar and we cannot dp correct analysis without these values.So skip the row if they consist null value.
   b. Replacement for NaN according to the properties of variables.

VARIABLES	REPLACE	REPLACE WITH
noagent	NaN	1
code	NaN	Most used code
browser	NaN	Most used browser
idx	NaN	Most used idx
norefer	NaN	1
find	NaN	Most used find
crawler	NaN	0(assume it as empty)
extension	NaN	Most used extension
zone	NaN	Most used zone
size	NaN	Mean of the size

8. Summary Metrics are as follows:
   a. Compute mean size
   b. Compute maximum used browser
   c. Compute distinct count of ip per month i.e. per log file
   d. Compute distinct count of cik per month i.e per log file
9. Provide logging.info to all the variables and check for anomalies in the data and remove the row if found.
10. Combine all dataframes and compute overall summary metric. Now export it to a csv.

11. Create a bucket on S3 through the code and upload both the files.

Exception handling:

- Made the bucket name unique by making it as a concatenation of ACCESS_KEY + Current_Timestamp.
- If the keys are invalid, the program will log an error and exit.
- If amazon keys are not provided, then program exits after displaying the appropriate message on console.