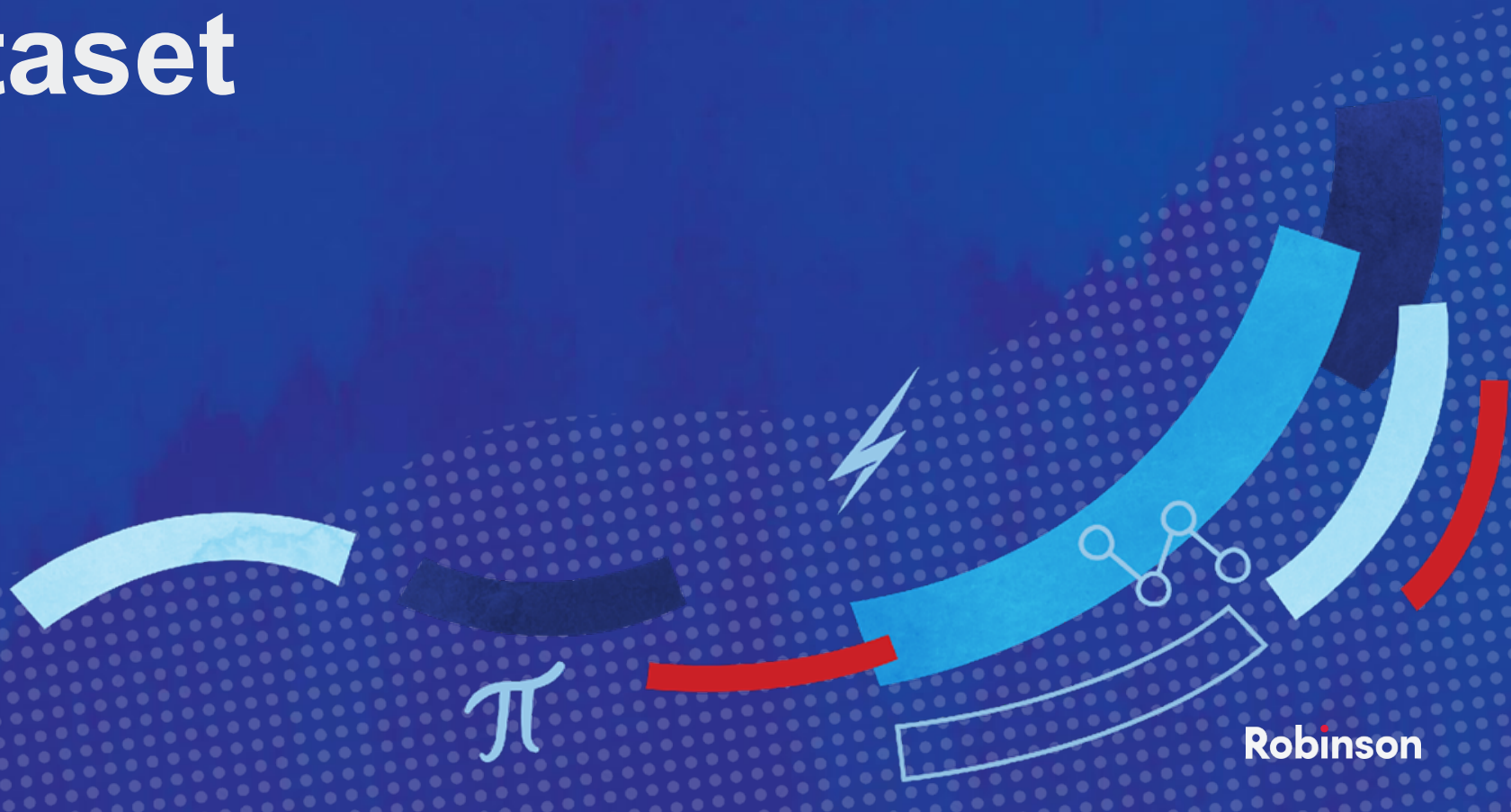
Data Management Project

Estimize Dataset

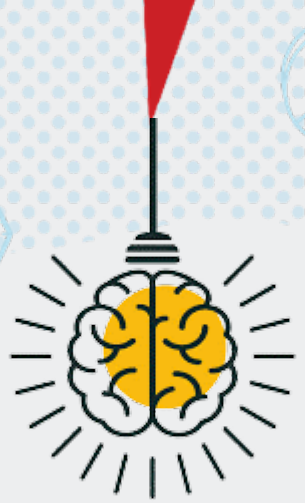
- Group No: 27
- Teja Natireddi
- Yashwant Sai Kishore



Georgia State University | J. Mack Robinson College of Business

Robinson

Problem Statement



- Estimize, an open web-based platform, has become a valuable source for retrieving financial estimates.
- It facilitates the aggregation of financial estimates from a diverse community of individuals.
- To Build an EPS database that stores the analyst's forecasts and accurate EPS for the companies

Contents

Web Scraping

Data Pre-processing

Data- Insertion

ER Model

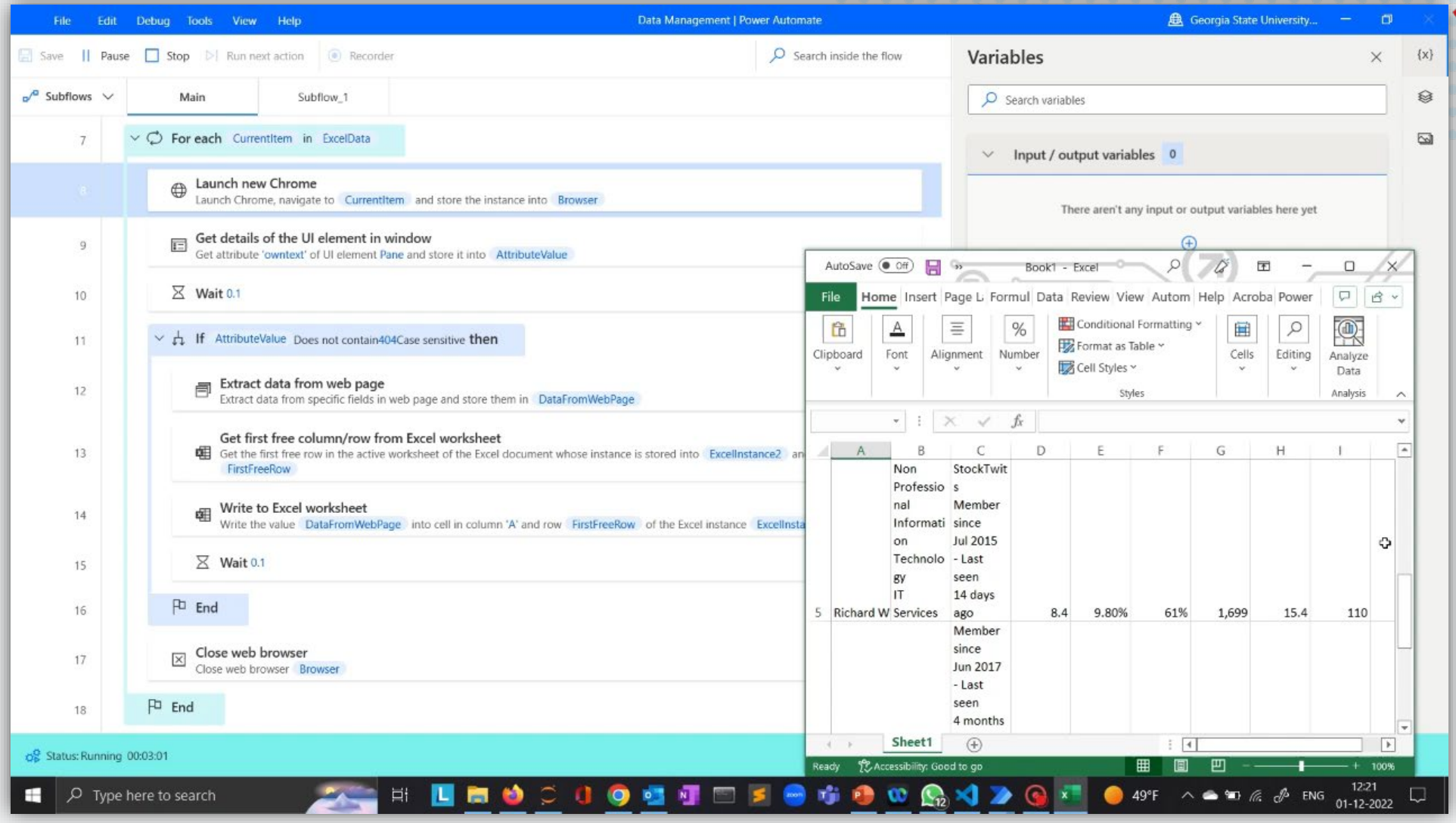
Regression

Data Scraping - Selenium

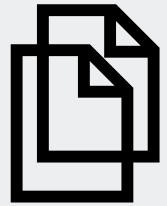
```
tick = ticks

tickers=[]
names=[]
sectors=[]
industries=[]
number_of_followers=[]
number_of_analysts=[]

options = webdriver.ChromeOptions()
options.headless = False
browser = webdriver.Chrome(executable_path = '/Users/arunk/Downloads/chromedriver',chrome_options = options)
link = 'https://www.estimize.com/edge'
browser.get(link)

for i in tick:

    search = browser.find_element_by_name('search') #return an object with value name 'search'
    search.send_keys(i) #feed search bar with text in the search bar
    search.send_keys(Keys.RETURN)

    time.sleep(3)

    ticker = browser. ("//*[@id='releases_show']/div[2]/div[2]/div/div[1]/div/div/div/div[1]/h1/a").text
    tickers.append(ticker)
    name = browser.find_element_by_xpath("//*[@id='releases_show']/div[2]/div[2]/div/div[1]/div/div/div/div[1]/p/a").text
    names.append(name)
    sector =  browser.find_element_by_xpath("//*[@id='releases_show']/div[2]/div[2]/div/div[1]/div/div/div/p/span[1]/a/span").text
    sectors.append(sector)
    ind = browser.find_element_by_xpath('//*[@id="releases_show"]/div[2]/div[2]/div/div[1]/div/div/div/p/span[2]/a/span').text
    industries.append(ind)
    numb_of_followers = browser.find_element_by_xpath("//*[@id='summary-stats']/div/div/div[1]/div[2]").text
    number_of_followers.append(numb_of_followers)
    numb_of_analysts = browser.find_element_by_xpath("//*[@id='summary-stats']/div/div/div[2]/a").text
    number_of_analysts.append(numb_of_analysts)
```

- Used Selenium to extract the basic information of each company, analysts info & their basic details
- Included dynamic sleep to avoid the forbidden error on the webpage
- Stored the information in the form of a pandas data frame for data insertion

Data Scraping - RPA



DATA PREPROCESSING



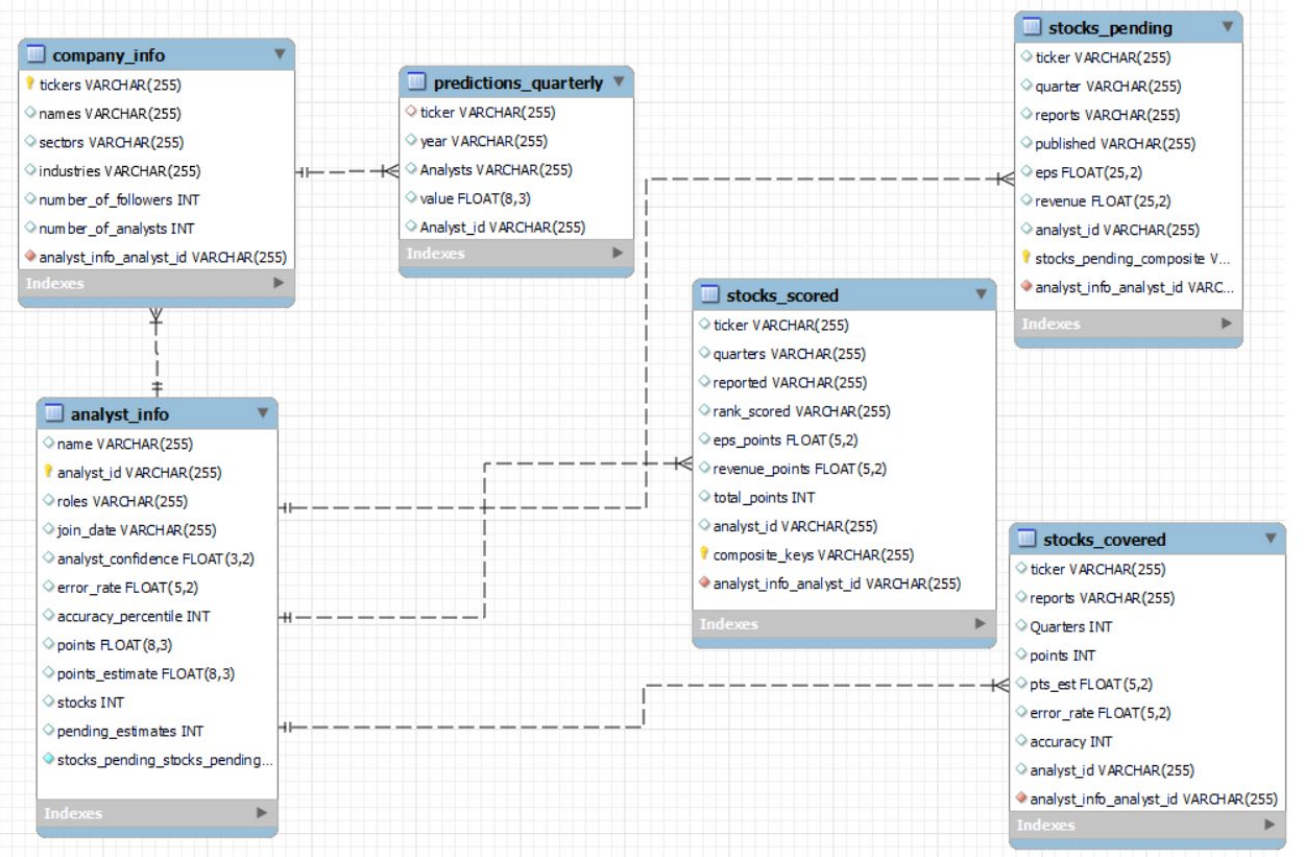
- Merged multiple excel files in which the data was extracted from the webpage
- Dropped the duplicate records of the extracted data frame.
- Changed the case of the text data to lowercase wherever necessary
- Performed string operations such as replacing the special characters with empty spaces
- Changed the data type of attributes such as date field to MySQL Date Time, numeric data to int and float

DATA INSERTION

```
import mysql.connector as msql
from mysql.connector import Error
try:
    conn = msql.connect(host='localhost', database='estimize', user='root', password='******')
    if conn.is_connected():
        cursor = conn.cursor()
        cursor.execute("select database();")
        record = cursor.fetchone()
        print("You're connected to database: ", record)
        cursor.execute('DROP TABLE IF EXISTS analyst_info;')
        print('Creating table....')
        cursor.execute("CREATE TABLE analyst_info(name varchar(255),analyst_id varchar(255),roles varchar(255)\
        ,join_date varchar(255),analyst_confidence float(3,2),error_rate float(5,2)\
        ,accuracy_percentile int, points float(8,3) ,points_estimate float(8,3)\
        , stocks int, pending_estimates int,Primary key (analyst_id))")

        for i,row in analysts_info.iterrows():
            #here %S means string values
            sql = "INSERT INTO estimize.analyst_info VALUES (%s,%s,%s,%s,%s,%s,%s,%s,%s,%s,%s)"
            cursor.execute(sql, tuple(row))

            conn.commit()
except Error as e:
            print("Error while connecting to MySQL", e)
```

- Initiated a connection to MySQL using the MySQL connector library
- Connected to estimate database to Insert the data
- Performed CRUD operations
- Defined the primary keys and referential integrity before loading the data into the tables.

ENTITY RELATIONSHIP MODEL



- The Company info table has information about the no of followers and analysts.
- Prediction info table consists of analysts of each company and their rank.
- The analysts info consists of information about each analyst.
- Regarding Stocks we have three tables , these consists of the data regarding the revenue of each company.

Query Example: Company with max no of followers?

```
2 •  select * from company_info
3    where number_of_followers = (select max(number_of_followers) from company_info);
```

Result Grid | Filter Rows: | Edit: | Export/Import: | Wrap Cell Content:

tickers	names	sectors	industries	number_of_followers	number_of_analysts
AAPL	Apple Inc.	Information Technology	Computers & Peripherals	37719	11974
NULL	NULL	NULL	NULL	NULL	NULL

Query Example: Company with min no of followers?

```
2 •  select * from company_info
3    where number_of_followers = (select min(number_of_followers) from company_info);
```

Result Grid | Filter Rows: | Edit: | Export/Import: | Wrap Cell Content:

tickers	names	sectors	industries	number_of_followers	number_of_analysts
VMW	VMware, Inc.	Information Technology	Software	722	1096
NULL	NULL	NULL	NULL	NULL	NULL

Query Example: Top 5 Analysts

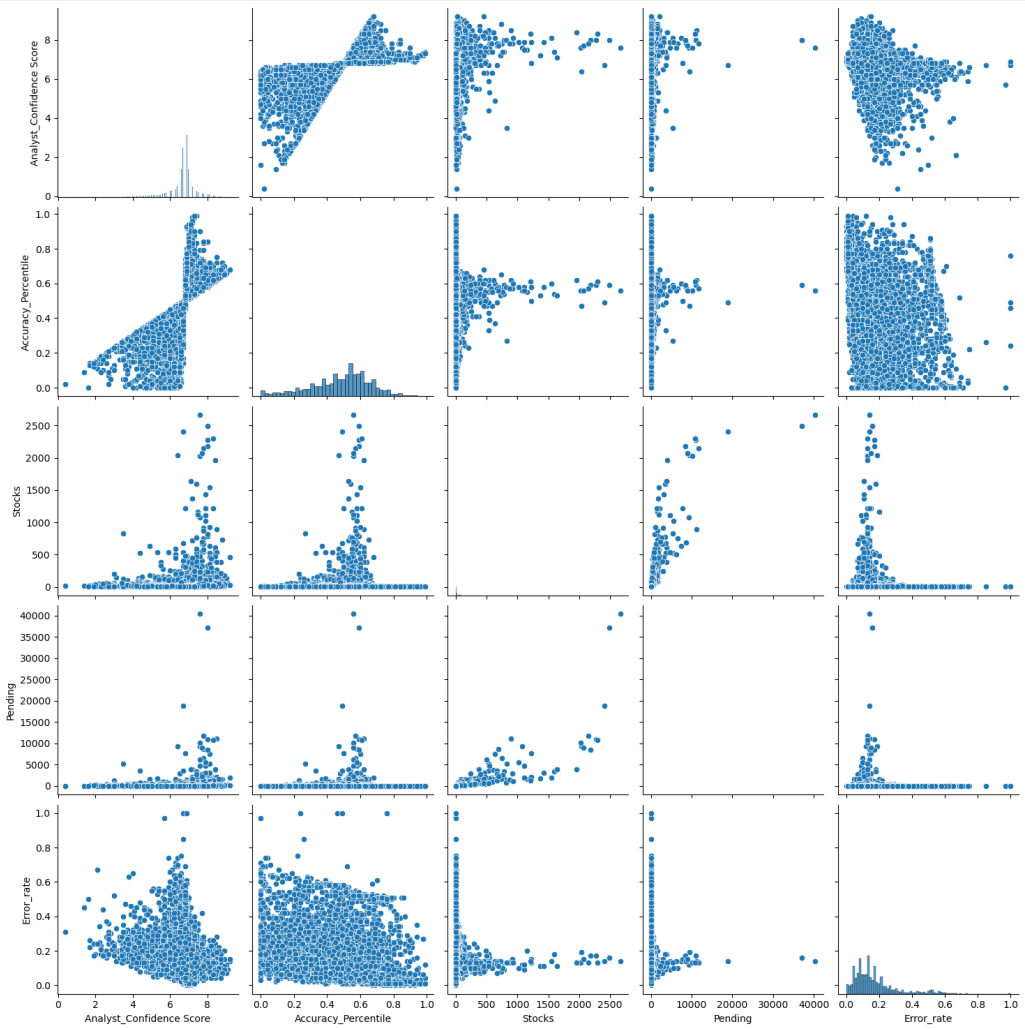
```
1 • use eps;
2 • select * from analyst_info
3   where error_rate = (select min(error_rate) from analyst_info)
4   limit 5;
```

Result Grid | Filter Rows: | Edit: | Export/Import: | Wrap Cell Content:

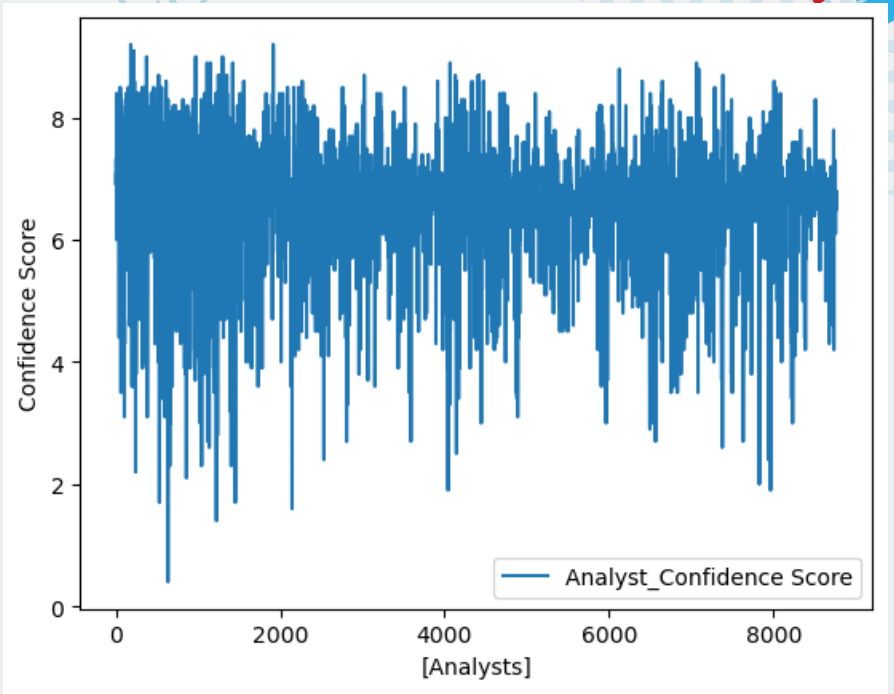
analystid	name	role	join_date	confidence_score	er
analyst_1059793	Analyst_1059793	Non Professional Other Other	Sep-21	6.9	0
analyst_2045619	Analyst_2045619	NULL	Dec-20	6.9	0
analyst_2051311	Analyst_2051311	Non Professional Other Other	May-20	6.9	0
analyst_2425645-4dd8c193-a46e-46ad-88ad-a...	ClaireOkoniewsk	Non Professional Student	Dec-20	6.9	0
analyst_2544793	Analyst_2544793	Non Professional Information Technology IT Ser...	Jun-20	7	0
NULL	NULL	NULL	NULL	NULL	NULL

BONUS - REGRESSION MODEL

Pair Plot



Line Plot



- **Pair Plot – Visualize the numeric data**
- **Line Plot – The average confidence score is – 6.67**

0.03917660043431979
0.03917660043431979
0.03917660043431979

Linear Regression

Accuracy: 60%
RMSE : 0.16683926645091696

```
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score  # For find ac
from sklearn.metrics import mean_squared_error  # F
from math import sqrt
linear_reg = LinearRegression(fit_intercept=False)
linear_reg.fit(X_train, y_train)  # Fit data to the
```

Random Forest Regressor

Accuracy: 91%
RMSE : 0.03917660043431979

```
from sklearn.ensemble import RandomForestRegressor
random_forest_reg = RandomForestRegressor(n_estimators=100,
max_depth=2, random_state=13)
random_forest_reg.fit(X_train, y_train)
```

THANK YOU

Georgia State University | J. MACK ROBINSON COLLEGE OF BUSINESS

Robinson