

DATABASE ANALYTICS AND PROGRAMMING-INDIVIDUAL WORK BREAKDOWN REPORT

Arunkumar Sudhakaran Nair
MSc Data Analytics,
School of Computing,
National College of Ireland,
Dublin, Ireland
x19153317@student.ncirl.ie

Abstract— Documentations have shown that the rate of unemployment is increasing concerning income per capita. The main reason for this is low educated workers whose GDP percapita unemployment levels are rising sharply rather than high educated workers who don't have an income-related unemployment rate. This individual research aims in analyzing the rate of change of unemployment to percapita income and rate of the population [1].

I. INTRODUCTION

Unemployment is one of the biggest social and economic problems nowadays. The economic effects of unemployment are equally serious, a 1% rise in unemployment rise reduces the GDP per capita by 2%. Unemployment has a mixed criminal effect which leads to increased crime rates. Long-lasting effects of unemployment are not only affecting the individuals but also it extends to their families. The longer the unemployment lasts, the greater the health impact, with rising depression and other health conditions declining over time. Besides the obvious income loss, unemployed workers found that they had lost friends and admiration for themselves[2]. In the group research paper, the effect of unemployment and population on per capita income is discussed as a group. The above factors made me choose unemployment as the topic of research.

II. INDIVIDUAL WORK BREAKDOWN

The below-listed works were performed individually during the project report.

A. Installing and setting up MongoDB

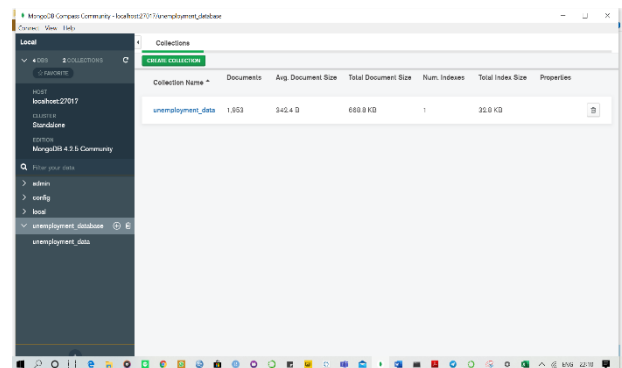
MongoDB was installed for storing unstructured data. MongoDB is a NoSQL database to store semi-structured documents like JSON. MongoDB was installed and setup locally on the same machine for the storage of JSON data. MongoDB was set up by using command prompt by 'mongo' and 'Mongod' commands. Then for ease of use, the mongo compass was installed and connected to the localhost port number 27017.

B. Setting up a connection between MongoDB and python

The connection between MongoDB and python was setup using the mongo client module from the Pymongo package. This package is used for interactions with MongoDB.

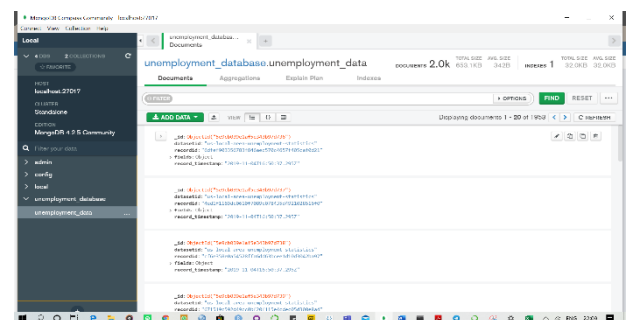
C. Inserting unstructured data into MongoDB

A database named 'unemployment_database' was created and a collection named 'unemployment_data' was created using python. JSON data was inserted to MongoDB collection using 'insert_many' command. Then the connection with the mongo client is closed.



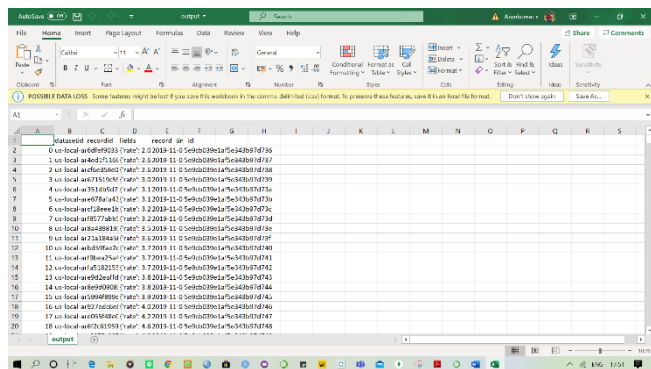
D. Reading the data from MongoDB into an array

The data from MongoDB was read by using an array and keys. The view of the data in MongoDB is shown in the figure below.



E. Converting Array to CSV

The value inside the array is converted into a CSV file of the specified name. The CSV file is named as 'output.csv'.



record_id	unemployment_rate	metropolitan_area	rank
1	0.02039	San Francisco	1
2	0.02039	San Francisco	2
3	0.02039	San Francisco	3
4	0.02039	San Francisco	4
5	0.02039	San Francisco	5
6	0.02039	San Francisco	6
7	0.02039	San Francisco	7
8	0.02039	San Francisco	8
9	0.02039	San Francisco	9
10	0.02039	San Francisco	10
11	0.02039	San Francisco	11
12	0.02039	San Francisco	12
13	0.02039	San Francisco	13
14	0.02039	San Francisco	14
15	0.02039	San Francisco	15
16	0.02039	San Francisco	16
17	0.02039	San Francisco	17
18	0.02039	San Francisco	18
19	0.02039	San Francisco	19
20	0.02039	San Francisco	20

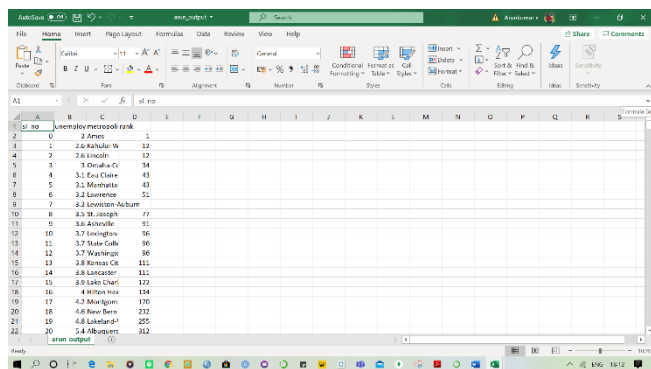
The above figure shows the converted CSV file. This file has multiple fields with data which was not useful for our analysis. So the next step was to clean the data.

F. Cleaning the data

The CSV file was read and stored in a data frame. I have noted that only one column named 'fields' was having the useful data I needed for my future analysis. So I have dropped the other fields. The next steps were to split the 'fields' column to get useful data. For that, I have used the split command. Also, a field named 'Unnamed :0' was present. Dropped all unnecessary fields using drop, split, and rename commands.

G. Convert the cleaned data frame into CSV

The final data comprises only 4 columns namely 'sl_no', 'unemployment_rate', 'metropolitan_area', and 'rank'. The cleaned data frame was then converted to a CSV file.



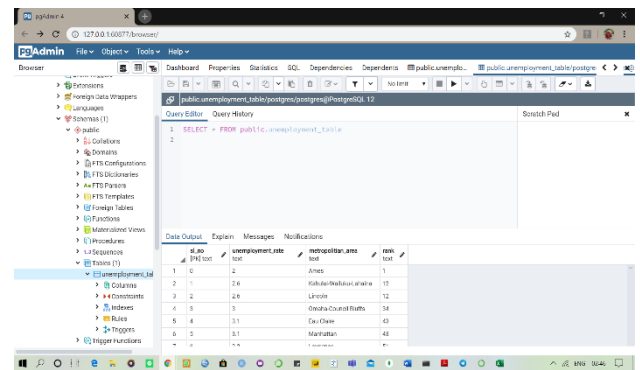
sl_no	unemployment_rate	metropolitan_area	rank
1	0.02039	San Francisco	1
2	0.02039	San Francisco	2
3	0.02039	San Francisco	3
4	0.02039	San Francisco	4
5	0.02039	San Francisco	5
6	0.02039	San Francisco	6
7	0.02039	San Francisco	7
8	0.02039	San Francisco	8
9	0.02039	San Francisco	9
10	0.02039	San Francisco	10
11	0.02039	San Francisco	11
12	0.02039	San Francisco	12
13	0.02039	San Francisco	13
14	0.02039	San Francisco	14
15	0.02039	San Francisco	15
16	0.02039	San Francisco	16
17	0.02039	San Francisco	17
18	0.02039	San Francisco	18
19	0.02039	San Francisco	19
20	0.02039	San Francisco	20
21	0.02039	San Francisco	21
22	0.02039	San Francisco	22

H. Connecting to POSTGRESQL

Postgresql connection was established by importing the psycopg2 package. For better ease of use, I have installed Postgresql locally on windows and using PgAdmin4, I was able to manage the interactions with the database.

I. inserting values into a Table

A table named unemployment_data was created using appropriate SQL queries and data was read from the CSV file and inserted into PostgreSQL. After inserting the data into PostgreSQL, the view is shown in the below figure.

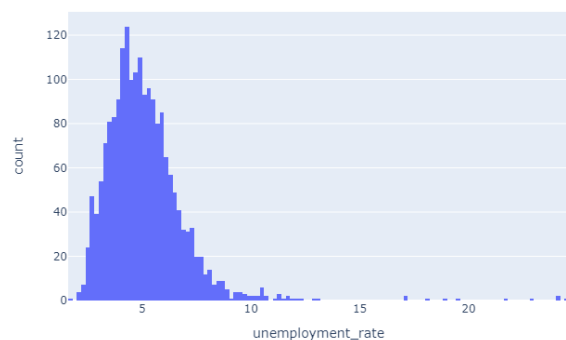


sl_no	unemployment_rate	metropolitan_area	rank
1	0.02039	San Francisco	1
2	0.02039	San Francisco	2
3	0.02039	San Francisco	3
4	0.02039	San Francisco	4
5	0.02039	San Francisco	5
6	0.02039	San Francisco	6
7	0.02039	San Francisco	7
8	0.02039	San Francisco	8
9	0.02039	San Francisco	9
10	0.02039	San Francisco	10
11	0.02039	San Francisco	11
12	0.02039	San Francisco	12
13	0.02039	San Francisco	13
14	0.02039	San Francisco	14
15	0.02039	San Francisco	15
16	0.02039	San Francisco	16
17	0.02039	San Francisco	17
18	0.02039	San Francisco	18
19	0.02039	San Francisco	19
20	0.02039	San Francisco	20
21	0.02039	San Francisco	21
22	0.02039	San Francisco	22

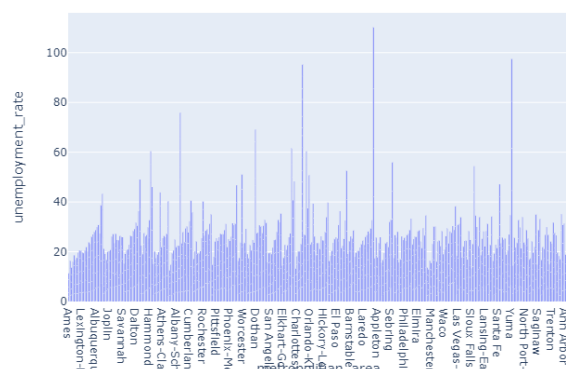
J. Visualization

The package named Plotly was used for the creation of interactive plots.

Histogram representation of the unemployment rate is shown in the plot below. Since the data is correlated with almost similar values, I was not able to gather much inference from the below plot.



The below bar graph representation, it is evident that the Appleton city is having the highest unemployment rate. Yuma city in Arizona ranks second in the unemployment rate.



The below scatterplot shows that even though there are potential outliers in the data, the data is almost normal.

