DATABASE ANALYTICS PRAGRAMMING A Comprehensive Data Analysis of Economic Indicators in the United States

SHREYANSH GUPTA
x21239347
MSc in Data analytics

SAIF RABBANI
x21223149
MSc in Data analytics

KARAN SONAWANE
x22167005
MSc in Data analytics

***Abstract*—Data analytics is used to ensure innovative data exploitation as the data economy becomes essential to a country's progress and future success. Economic data analytics will provide insights into the performance of the economy, it also helps identify trends and patterns, and informs policymakers on how to make informed decisions that can impact people's lives and businesses. In this study, we analyze the economic data of the United States of America, considering the minimum wage, presidential voting, and unemployment rate in different counties. We gathered data from multiple sources, including the Federal Reserve Economic Data and the Election Commission, spanning different time periods. We utilized various statistical tools to analyze the data. Our results indicate a significant positive correlation between minimum wage and presidential voting in favor of the Republican Party, and a negative correlation between minimum wage and unemployment rate. Additionally, we observed notable variations in these correlations among different counties in the United States.**

I. Introduction

The volume and variety of data have significantly increased, resulting in valuable economic and practical applications of data analysis. The most important objective of data analysis work is to analyse and utilize data efficiently and effectively. Economic data provides valuable insights into economic performance, identifying trends and patterns, evaluating policy effectiveness, improving resource allocation, understanding economic inequality, supporting international trade and investment. It informs decision-making, promotes growth, and improves well-being. The federal government accumulates enormous volumes of detailed administrative data through the management of taxes, social programs, and rules, including micro-level statistics from organizations like the Social Security Administration, Internal Revenue Service, and Centres for Medicare and Medicaid. Significant administrative statistics are also produced by state and local governments, particularly in sectors like education, social insurance, and local government spending. Despite this, limited access to administrative data from the government undoubtedly prevents researchers and private data sellers from making new discoveries. [1]

In our research study, for economic data analysis of the United States of America, considering factors such as presidential voting, unemployment rate, and minimum wage, is essential as it provides insights into the economic and political landscape of the country. By studying the relationship between these factors, policymakers, researchers, and other stakeholders can understand the impact of government policies, economic conditions, and social factors on the economy. The objective of this project is to extract data, store and retrieve it from appropriate databases and identify the correlation between minimum wage, presidential voting, and unemployment rate in different counties of the United States, using analytical tools. By analysing these factors, the study aims to provide insights into the economic landscape of different regions in the country, and how it may affect political preferences and unemployment rates. The study also aims to identify variations in the correlations between these factors among different counties in the United States.

The findings of this study are relevant as they can inform policymakers, researchers, and stakeholders on the effects of minimum wage policies on political outcomes and employment rates in different regions of the United States. It can also help policymakers to design policies that aim to improve economic conditions in areas that are experiencing higher unemployment rates or lower political participation. The study's insights could potentially lead to a more equitable and inclusive economic environment for the United States, benefiting both citizens and businesses alike. The proposed methodology will help data scientists and data analysts working under the government to successfully build models to analyse economic data.

II. RELATED WORK

PROBLEM STATEMENT: To analyze the economic data indicators and extract the insightful factors from it and then apply data visualization techniques on it to understand the trend.

The study by H. Kang titled "Analysis of the Impact of COVID-19 on the U.S. Economy Based on Big Data Processing and Experimental Analyses" has certain limitations. Firstly, the study heavily relies on big data processing techniques, which may not always produce accurate results. The paper does not account for potential biases in the data or address the issue of missing data. The study does not consider

the impact of the pandemic on specific industries or regions, which can affect the overall analysis. Most importantly the paper does not explore the long-term effects of the pandemic on the economy and only focuses on the short-term impact. [2]

The study by A. Kyung and S. Nam focuses on studying the unemployment rate in the USA using computational and statistical methods. The authors used a dataset of monthly unemployment rates and applied various statistical and machine learning methods, including linear regression and decision tree algorithms, to predict the unemployment rate. However, the paper did not address the issue of data bias, and the dataset used was limited to only one economic indicator, which could affect the accuracy and reliability of the analysis. Additionally, the paper did not consider other economic indicators, such as GDP or inflation, which could provide a more comprehensive understanding of the economic landscape. [3]

The research paper titled "Statistical analysis and data processing: A case study of employment effects of minimum wages" by Qiong Wang was presented at the 12th International Conference on Fuzzy Systems and Knowledge Discovery in 2015. The objective of the study is to examine the relationship between minimum wages and employment by utilizing statistical techniques. The data was collected from various sources and used econometric models to identify the impact of minimum wages on employment. The results suggest that minimum wage increases negatively affect employment, especially for low-skilled workers. The study has some limitations, the study only focuses on the employment effects of minimum wages and does not consider other factors that may impact employment. The study relies on aggregate data and does not account for differences in the impact of minimum wages across various industries and regions. This study does not consider the long-term effects of minimum wage increases on the economy, which could potentially limit the scope of the analysis. [4]

The study "Scientific Research in Relation to Gross Domestic Product (GDP) A Comparative Study of China and India" has its limits. It only looks at the amount of research papers and patents filed, as well as funding for research and development, but leaves out other factors that affect research output, like how researchers work together and the quality of their work. The study also doesn't take a detailed look at the economic and political factors that impact research output in China and India. This study is mostly quantitative, so it doesn't provide insights into what researchers and other people involved in research think about their work. [5]

Our analysis includes data from FRED (Federal Reserve Economic Data) which aims to address some of these limitations. By including multiple datasets, we can create a more comprehensive understanding of the economic landscape. we will use advanced statistical methods, including machine learning algorithms, to analyse complex relationships and patterns in the data. In this approach we will also address potential biases and inaccuracies in the data by using multiple sources. By utilizing a diverse set of data sources and analysis methods, we can achieve a more robust and accurate economic analysis.

III. METHODOLOGY

The dataset we are analysing includes information on the unemployment rate [6], the minimum wage, and presidential election results in different counties over time. We scrapped data on the postal codes of different counties through an API and used it to link the datasets together.

The reason for choosing these datasets is that they provide key economic indicators that can help us better understand trends in the market and the impact of political events on the economy. The unemployment rate is a widely used measure of economic activity, and tracking changes in minimum wage can help us understand the impact of policies on different regions of the country. Presidential elections can also have a significant impact on economic policies and investor sentiment, making them an important factor to consider when analysing economic data. By linking these datasets together using postal codes, we can perform more nuanced analyses of economic trends across different regions of the country. We used FRED to perform statistical analyses on the dataset and generate insights into the relationships between these economic indicators.

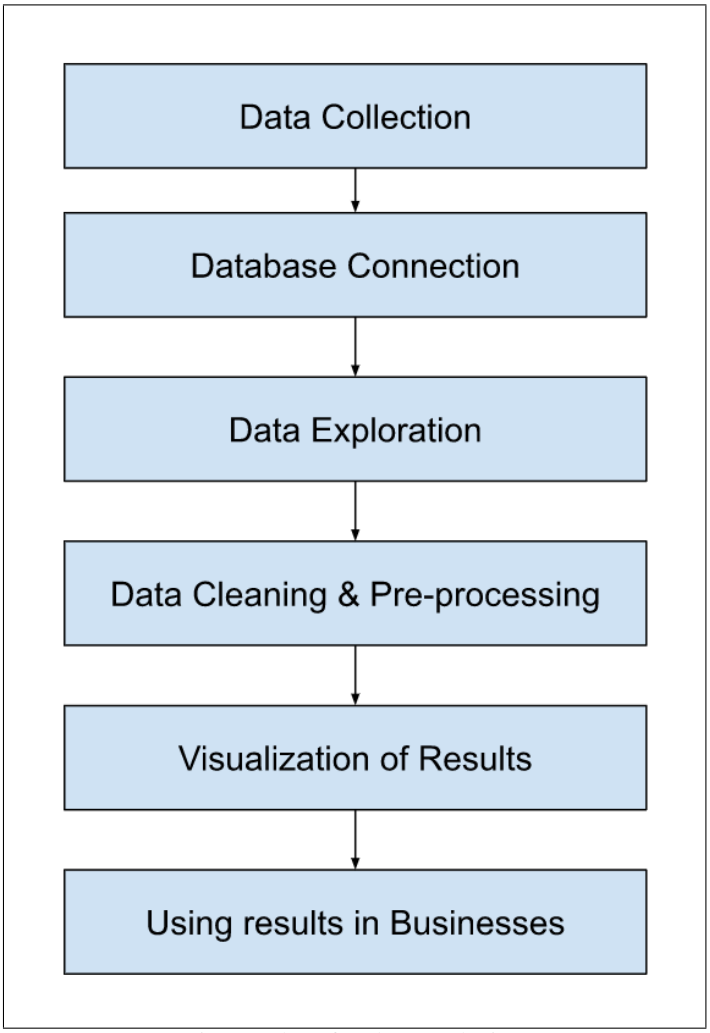

Fig. 1: Flow for data analysis

The data analytics model follows a specific flow, starting with the collection of data, followed by processing the data using techniques such as transformation, cleaning, and preprocessing. We then derive visualisation from the data using plots or graphs, which can then be used to extract meaningful insights. The entire project is cloud-based, including coding files, SQL, and NoSQL database tables. [7] There are no dependencies on local machines, resulting in increased portability and compatibility. We imported all necessary libraries, such as 'pandas' for performing DataFrame-related operations, 'json' for working with JSON files, 'requests' to handle HTTP requests, 'time' for working with database connections, 'pprint' for reading data structures in a more readable way, 'pymongo' and 'mongoClient' for connecting to the NoSQL MongoDB database, 'matplotlib' and 'seaborn' for data exploration and visualization, and 'postgresql' for connecting to a SQL server. [8]

Structuring data storage can be a challenging task, especially as data becomes more dispersed across different economic locations. Failing to have a clear understanding of where data is stored can result in severe consequences, including regulatory penalties, damage to a brand's reputation, and loss of customer trust. To gain insight into handling data with a NoSQL database, we incorporated MongoDB into our project. MongoDB [9] can be deployed on-premises, in the cloud, or as a fully managed global cloud database. With MongoDB, data is highly available, as multiple copies are made and distributed across different servers. This redundancy ensures that data can be recovered without delay in the event of a server failure, we imported semi-structured datasets into the mongodb server and for structured dataset, we used Postgresql server.

We established mongodb connection using mongo client function where json file namely 'unemployment data' is stored. This json file contains parameter such as : years, months, states, county and most importantly unemployment rate. We explored this file using data exploration functions like head(), info(), shape(), describe() functions. This file contains 8,85,548 records. After exploring the data, we moved forward for data cleaning process where we checked the null values, using 'isna()', 'isnan()' functions.

	Year	Month	State	County	Rate
0	2015	February	Mississippi	Newton County	6.1
1	2015	February	Mississippi	Panola County	9.4
2	2015	February	Mississippi	Monroe County	7.9
3	2015	February	Mississippi	Hinds County	6.1
4	2015	February	Mississippi	Kemper County	10.6

Fig. 2: Unemployment rate dataset

The unemployment rate is a significant economic indicator as it shows the percentage of unemployed people actively looking for work. It reflects the economic growth and labour market health. Similarly, minimum wage is crucial as it affects business labour costs and consumer purchasing power, which, in turn, impacts employment levels, consumer spending, and economic growth.

To better understand the American economy, the unemployment rate dataset [6] was mapped to another dataset, which is also in JSON file format, related to minimum wage. This helps explore the relationship between these two variables and the regional differences in the economy, as minimum wages vary by county in many states. This helps in determining appropriate minimum wage levels in different regions or targeting interventions to reduce unemployment in specific areas. The figure below shows the snippet of the minimum wage dataset. [10]

Year	State	State.Minimum.Wage	State.Minimum.Wage.2020.Dollars	Federal.Minimum.Wage	Federal.Minimum.Wage.2020.Dollars	Effective.Minimum.Wage
1968	Alabama	0.00000	0.00	1.15	8.55	1.15
1968	Alaska	2.10000	15.61	1.15	8.55	2.10
1968	Arizona	0.46800	3.48	1.15	8.55	1.15
1968	Arkansas	0.15625	1.16	1.15	8.55	1.15
1968	California	1.65000	12.26	1.15	8.55	1.65
...	...	...	...	...	...	...
2020	Virginia	7.25000	7.25	7.25	7.25	7.25

Fig. 3: Minimum wage dataset

The dataset 'minimum wage' [10] contains parameters such as years, state, state-wise minimum wage, state-wise minimum wage in the year 2020, federal minimum wage, effective minimum wage. We explored this dataset in the similar manner as the first dataset. We applied groupby function on states parameter which selects all states indexed by year and with columns labelled with each state's name.

	Alaska	Arkansas	California	Colorado	Connecticut	Delaware	District of Columbia	Guam	Hawaii	Idaho
Alaska	1.000000	-0.332127	0.622245	-0.078678	0.646909	0.260265	0.271810	0.787515	0.630212	0.371745
Arkansas	-0.332127	1.000000	0.030265	0.333597	0.183662	0.107639	0.433775	0.037006	0.324870	0.008276
California	0.622245	0.030265	1.000000	0.649967	0.881831	0.518481	0.727373	0.411729	0.610784	0.542006
Colorado	-0.078678	0.333597	0.649967	1.000000	0.520931	0.581162	0.798145	-0.152291	0.291995	0.436981
Connecticut	0.646909	0.183662	0.881831	0.520931	1.000000	0.586472	0.700081	0.587469	0.674581	0.536894

Fig. 4: Result of grouby function.

We created a function to get the minimum wage of a particular county. By doing this we are trying to find out if there is any relationship between unemployment rate and the minimum wage, to identify the relationship we found a way to apply correlation and covariance on some DataFrame. We checked row by row, state by state to that of unemployment dataset 'map' function.

The function get_min_wage was taken from 'minimum wage' dataset which is mapped to the year and state from 'unemployment rate' dataset, here we merged the two datasets into one.

```
get_min_wage(2012, "Colorado")

8.61
```

Fig. 5: Extracting minimum wage of a particular county.

	Year	Month	State	County	Rate	min_wage
885543	2009	November	Maine	Somerset County	10.5	8.46
885544	2009	November	Maine	Oxford County	10.5	8.46
885545	2009	November	Maine	Knox County	7.5	8.46
885546	2009	November	Maine	Piscataquis County	11.3	8.46
885547	2009	November	Maine	Aroostook County	9.0	8.46

Fig. 6: Merged dataset

The values of correlation and covariance in the merged dataset can be seen in the images below, which describes relationship between unemployment rate and minimum wage

	Rate	min_wage
Rate	1.000000	0.153047
min_wage	0.153047	1.000000

Fig. 7: Correlation between the unemployment rate and minimum wage

The above image proves that unemployment rate and minimum wage are co-related, but it looks like there is only a slight positive relationship.

	Rate	min_wage
Rate	9.687873	0.651586
min_wage	0.651586	1.874228

Fig. 8: Covariance

The Fig. 8 projects that there is a strong covariance which means these two vary together, even after both factors varying together, the impact of one on another is not significant.

The figure 9 shows data distribution of a variable against density distribution using univariate data analysis.

The figure 10 shows how the data is diverse.

A president who supports minimum wage increases may enact policies to raise it, while a president who prioritizes job

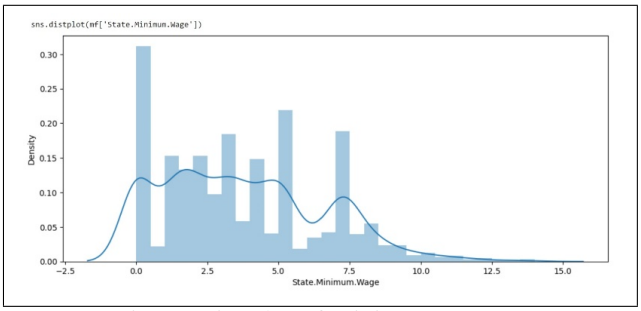

Fig. 9: Dist plot of minimum wage.

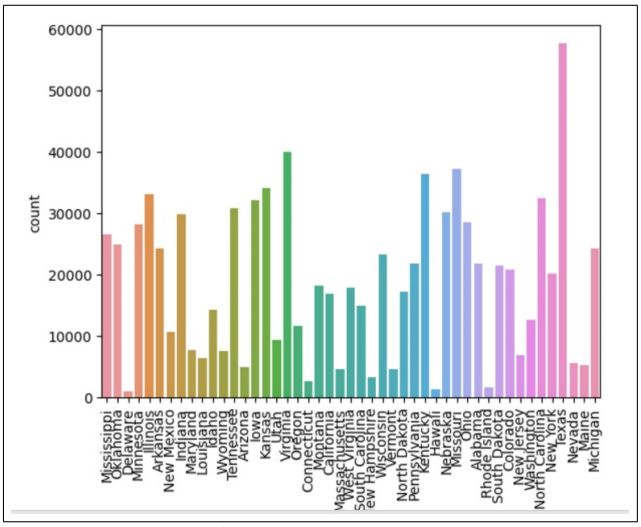

Fig. 10: Count Plot

creation may implement policies to lower unemployment rates, hence presidential voting is closely related to unemployment rate and minimum wage. The dataset [11] which we considered includes presidential voting of United states of America of the year 2016. The below figure shows a snippet of presidential voting dataset which is in CSV format.

fips	cand	st	pct_report	votes	total_votes	pct	lead
US	Donald Trump	US	0.9951	60350241.0	127592176.0	0.472993	Donald Trump
US	Hillary Clinton	US	0.9951	60981118.0	127592176.0	0.477938	Donald Trump
US	Gary Johnson	US	0.9951	4164589.0	127592176.0	0.032640	Donald Trump
US	Jill Stein	US	0.9951	1255968.0	127592176.0	0.009844	Donald Trump
US	Evan McMullin	US	0.9951	451636.0	127592176.0	0.003540	Donald Trump
US	Darrell Castle	US	0.9951	180877.0	127592176.0	0.001418	Donald Trump

Fig. 11: Dataset after using Map Function

The aggregate statistic for the entire US is the starting point for this data, which is then broken down by state, county, and candidate which displays the top ten candidates of the year 2016. The dataset where we merged unemployed rate and minimum wage has some values which says 'nan', to remove this we scrapped the data through an API, which helped us in getting the postal codes of different states. We mapped these postal codes on the same dataset using dictionary.

State Name/District	
Alabama	AL
Alaska	AK
Arizona	AZ
Arkansas	AR
California	CA

Fig. 12: Postal code mapped in state name.

The figure 12 shows depiction of merged columns in the dataset.

State	pct
US	0.472993
CA	0.330641
FL	0.490640
TX	0.525830

Fig. 13: State wise PCT value

```
['Donald Trump' 'Hillary Clinton' 'Gary Johnson' 'Jill Stein'
 'Evan McMullin' 'Darrell Castle' 'Gloria La Riva' 'Rocky De La Fuente'
 ' None of these candidates' 'Richard Duncan']
```

Fig. 14: Top ten candidates in presidential voting.

County	State	Month	Rate	min_wage	pct
Major County	OK	February	2.6	2.11	0.864960
Pottawatomie County	OK	February	4.5	2.11	0.701342
Johnston County	OK	February	6.5	2.11	0.770057
Jefferson County	OK	February	5.0	2.11	0.812367
Beaver County	OK	February	2.8	2.11	0.888243

Fig. 15: Merged Dataset2

	Rate	min_wage	pct
Rate	1.000000	0.186689	-0.085985
min_wage	0.186689	1.000000	-0.325036
pct	-0.085985	-0.325036	1.000000

Fig. 16: Correlation of the dataset

The above figure 16 depicts the Correlation of the altogether dataset. [12]

	Rate	min_wage	pct
Rate	5.743199	0.683870	-0.031771
min_wage	0.683870	2.336451	-0.076602
pct	-0.031771	-0.076602	0.023772

Fig. 17: Covariance of the dataset

The above figure 17 depicts the Covariance of the altogether dataset. [12]

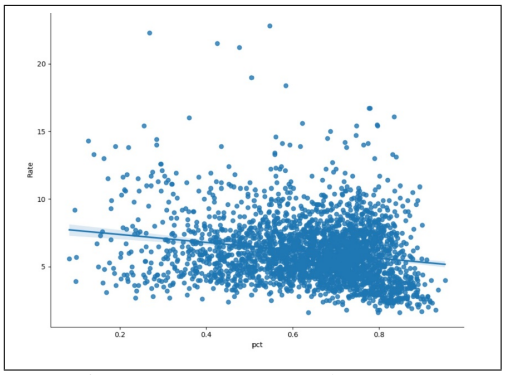

Fig. 18: PCT vs unemployment rate

The above figure 18 shows a plot using Implot(), this method is used to plot data and draw regression model fits

across grids where multiple plots can be plotted. This plot shows us the dense data and gives us the best fit line.

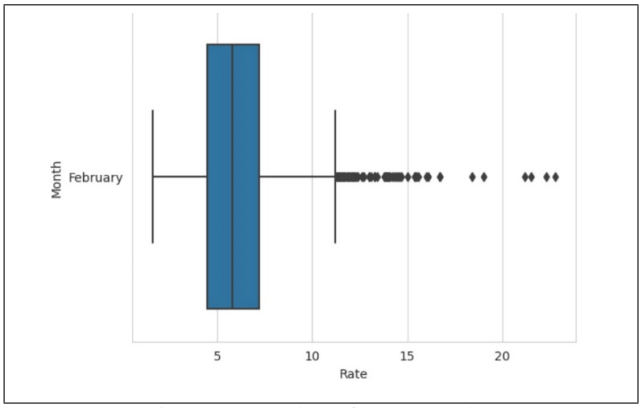

Fig. 19: Boxplot of one month

We used categorical variable to plot the boxplot of 'altogether' data, where the month selected was 'February'.

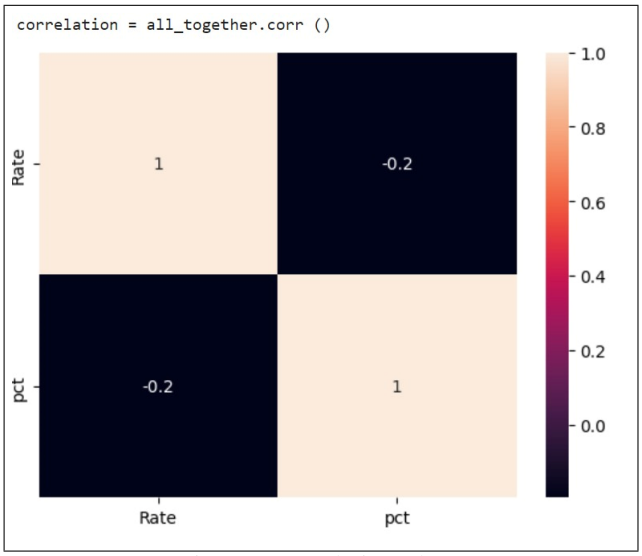

Fig. 20: Correlation plot

The above shows the Correlation between PCT and Unemployment rate after merging all the datasets. The target database 'Alltogether' was then finally stored into cloud-PostgreSQL server.

For further analysis, we decided to explore more about the American Economy and compare the results with our analysis. We found the website called 'Federal Reserve Economic Data' where we generated an API-key, this API key was then used in our notebook to scrap the data. The figure below shows the dataset which we scrapped from the website.

After analysing the dataset further, we plotted visualisations which gives us a better insight into the unemployment rate of the United States of America from the year 1980 to 2020, here the graph shows peak of unemployment rate in the year 2020

series id	id	realtime_start	realtime_end	title	observation_start	observation_end	frequency	frequency_short	units	units_short	seasonal_adjustment
BAMLH0A0HYM2	BAMLH0A0HYM2	2023-04-24	2023-04-24	ICE BofA US High Yield Index Option-Adjusted S...	1996-12-31	2023-04-20	Daily, Close	D	Percent	%	Not Seasonally Adjusted
CSUSHPINSA	CSUSHPINSA	2023-04-24	2023-04-24	S&P/Case-Shiller U.S. National Home Price Index	1987-01-01	2023-01-01	Monthly	M	Index Jan 2000=100	Index Jan 2000=100	Not Seasonally Adjusted
SP500	SP500	2023-04-24	2023-04-24	S&P 500	2013-04-22	2023-04-21	Daily, Close	D	Index	Index	Not Seasonally Adjusted
BAMLH0A0HYM2EY	BAMLH0A0HYM2EY	2023-04-24	2023-04-24	ICE BofA US High Yield Index Effective Yield	1996-12-31	2023-04-20	Daily, Close	D	Percent	%	Not Seasonally Adjusted
CSUSHPISA	CSUSHPISA	2023-04-24	2023-04-24	S&P/Case-Shiller U.S. National Home Price Index	1987-01-01	2023-01-01	Monthly	M	Index Jan 2000=100	Index Jan 2000=100	Seasonally Adjusted

Fig. 21: fred dataset

which depicts the effects of Pandemic. The visualisations of the same are depicted here.

Fig. 22: series based data

This above series uses the fredapi library to fetch S&P 500 data from FRED and plots it as a line graph with a specified size, title and line width, which is ordered by Popularity.

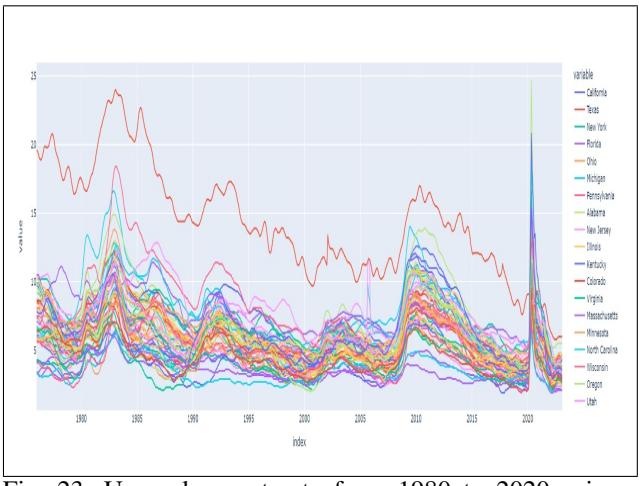

Fig. 23: Unemployment rate from 1980 to 2020 using FRED.

The above figure 23 contains all the counties of America ranging from California to Utah which are depicted in different

colours, which depicts the unemployment rate over 4 decades.

IV. RESULT & CONCLUSION

In this report, we analysed the economic data of different counties in the United States by merging two datasets, which were unemployment rate and minimum wage. After merging, we found a correlation and covariance between these two factors. Further, we included the presidential voting dataset to investigate the influence of democratic and republican states on the economy. We also stored the preprocessed data into appropriate databases. We also established a PostgreSQL connection on the cloud to store the dataset and merged it based on the county.

We obtained data from the Federal Reserve Economic Data (FRED) to forecast unemployment rates, which helped us to understand the American economy better. We drew several conclusions from all the datasets, including that the minimum wage has a negative correlation with the percentage of votes for the respective president. As minimum wage rises, people are less likely to vote for Trump. This is consistent with higher minimum wages being typically part of the Democratic agenda. Even if they are correlated, they do not attempt to vary together. It is important to note that there are complex interrelationships between economic factors and political outcomes. It highlights the importance of understanding these relationships to make informed decisions about policy and politics.

V. FUTURE WORK

The following could be the follow-ups that can be made in order to delve deeper into this work based on the findings of this study.

I. It would be interesting to look at how other economic variables like inflation, GDP growth, and interest rates affect these two variables in addition to the link between the unemployment rate and minimum wage. The cost of living is significantly impacted by inflation, which can have an influence on the minimum wage and total employment rates. Employment and salary patterns can also be influenced by GDP growth and interest rates since they show how the economy is doing overall. A more thorough picture of the American economy and assistance in guiding policy decisions might result from analyzing these elements along with the unemployment rate and minimum wage. Additionally, users may acquire a better understanding of these aspects by putting them into the web application and BI dashboards indicated in the section on future projects.

II. The COVID-19 pandemic significantly affected the world economy, including that of the United States. Understanding the financial effects of the pandemic can be done by looking at how COVID-19 has affected various US counties and industries. By identifying the sectors and regions that are most affected, this analysis can help guide the development of policies and interventions that aim to lessen the adverse effects on the economy. Additionally, it can serve as a foundation for assessing the success of economic aid programs and locating any gaps that require attention. Furthermore, being aware of how COVID-19 has affected the economy can help with long-term strategy and choice-making, especially in terms of crisis preparedness.

III. Building predictive models with the ability to analyze historical data and spot patterns and trends would be necessary to use methods of machine learning to predict the rate of unemployment and minimum wage for various areas in the United States. These models are able to utilize the forecast future, minimum wages and unemployment rates. The datasets can be used to apply a variety of machine learning algorithms, including regression, time series analysis, and neural networks. Additionally, relevant features that are likely to have an impact on the outcomes of interest can be extracted from the data using feature engineering techniques. Insights from this analysis could help businesses and policymakers make sound choices regarding the economy, such as finding regions where action is required to lower unemployment rates or raising the minimum wage to promote economic growth.

A potential area for further investigation, in addition to the work mentioned above, is the creation of an online application that would enable the general public to access and engage alongside economic information that we have examined. Users of such an application might benefit from an easy-to-use interface that enables simpler and interactive data visualization and exploration. The creation of business intelligence (BI) dashboards that compile the most important findings from our analysis into one view may be one feature of this web application. Line charts, scatter diagrams, and heat maps are just a few examples of the visualizations that could be included in these dashboards to help users quickly understand the connections between various economic variables.

One dashboard, for instance, might display county-level data on the correlation between minimum wage and unemployment rates, with the option to filter the information by state or year. Another dashboard could show how presidential voting trends and economic metrics like job openings or median income relate to one another. As fresh data becomes available, these visualizations could be instantly updated to give users the most recent information.

In conclusion, the creation of a web-based application via BI dashboards could increase the usability and applicability of our analysis for a larger audience. It would give decision-makers, analysts, journalists, and interested citizens a potent tool for examining and comprehending the intricate connections between various economic variables.

REFERENCES

[1] R. Elmasri, S. B. Navathe, and N. Shamkant, *Fundamentals of database systems*. Pearson Education, 2021.

[2] A. Chawla *et al.*, "Post-pandemic economic transformations in the united states of america," in *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, Orlando, FL, USA, 2022, pp. 1–5.

[3] A. Kyung and S. Nam, "Study on unemployment rate in usa using computational and statistical methods," in *2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, New York, NY, USA, 2019, pp. 0987–0993.

[4] Q. Wang, "Statistical analysis and data processing: A case study of employment effects of minimum wages," in *2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, Zhangjiajie, 2015, pp. 1085–1090.

[5] T. Zainab, Z. A. Wani, and M. A. Bhat, "Scientific research in relation to gross domestic product (gdp) a comparative study of china and india," in *2018 5th International Symposium on Emerging Trends and Technologies in Libraries and Information Services (ETTLIS)*, Noida, India, 2018, pp. 328–330.

[6] "Bureau of labor statistics," https://www.bls.gov/, accessed: May 10, 2023.

[7] J. Han, E. Haihong, G. Le, and J. Du, "Survey on nosql database," in *2011 6th International Conference on Pervasive Computing and Applications*. IEEE, 2011, pp. 363–366.

[8] "Postgresql," https://www.postgresql.org/, accessed: 2023-05-07.

[9] MongoDB, "Mongodb documentation," https://docs.mongodb.com/, accessed 2023.

[10] "History of federal minimum wage rates under the fair labor standards act, 1938-2009," https://www.dol.gov/agencies/whd/state/minimum-wage/history, accessed May 10, 2023.

[11] The New York Times, "2016 election results: President live map by state, real-time voting updates," https://www.nytimes.com/elections/2016/results/president, 2016, accessed on May 7, 2023.

[12] Wikipedia contributors, "Covariance and correlation — Wikipedia, the free encyclopedia," 2023, [Online; accessed 7-May-2023]. [Online]. Available: https://en.wikipedia.org/wiki/Covariance_and_correlation