Analyzing U.S. Cost of Living, Socioeconomic Disparities and Quality of Life Trends

Bill Joshua Swamidoss
MSc. Data Analytics
National College of Ireland
x23178981@student.ncirl.ie

Balasubramanian C
Senior Digital Engineer,
Sonata Software, India
balavicky85@gmail.com

Abstract—**The "US Cost of Living Dataset - 3,171 Counties" gathers a wealth of information on living costs in a variety of U.S. counties, covering demographics like population size, income levels, and unemployment rates as well as economic indicators like housing, transportation, food, and healthcare. Its main goal is to make cost of living comparisons possible, which will help companies, governments, and scholars better understand regional economic differences. This dataset assists in evaluating affordability and economic trends, and it provides information for budgetary and migration decisions. While policymakers can use data to guide economic strategies, researchers can examine relationships between living expenditures and socioeconomic characteristics. Similar to this, the U.S. Bureau of Labor Statistics' "Unemployment in America per US State" dataset offers comprehensive state-level unemployment data. It facilitates understanding of local economic circumstances, supporting economics.**

Keywords— ***US Cost of Living, Employment and Unemployment, Quality of Life***

I. Introduction

Raising costs, tax rises, rising energy prices, reductions in social security, stagnant salary levels, when combined, create a very challenging atmosphere, particularly for families who are already having a hard time moving forward. True to its right, the cost-of-living problem has garnered a lot of attention since the beginning of 2022 and will put millions of households under genuine, ongoing financial strain. Unfortunately, though, the current state of our social security system renders it useless. Even before to the start of the epidemic, this was evident due to the ongoing reductions in assistance and the inability to raise benefits in proportion to inflation, resulting in insufficient benefit levels that frequently fall short of meeting necessities [1].

Although inflation is declining and unemployment is low, consumer confidence is still low. Economists have been perplexed by this as they have always relied on these two factors to determine how consumers feel about the state of the economy. We suggest that a major contributing factor to this discrepancy is the rise in borrowing costs, which have not decreased in decades. Since traditional price indexes do not now incorporate the cost of money, there is a discrepancy between the metrics that economists like and the actual prices that consumers bear. We demonstrate that there is a robust correlation between borrowing prices and the availability of consumer credit and the low points in US consumer mood that cannot be attributed to unemployment or official inflation. There is a peak in worries about borrowing expenses, which traditionally correspond with the cost of money [2].

The idea of quality of life in the United States encompasses several aspects and is indicative of the general fulfilment and well-being of its citizens. It includes a wide range of elements, including social support networks, healthcare accessibility, high-quality education, economic opportunity, safety, and environmental circumstances. The United States of America, one of the biggest and most varied countries in the world, provides a variety of experiences and lifestyles, from thriving urban areas to tranquil rural villages. In their efforts to build conditions where people may thrive physically, psychologically, and emotionally, legislators, urban planners, and communities all place a high focus on understanding and improving quality of life [3].

Our project's goal is to determine the different elements that make up the US country, such as state-by-state unemployment, Quality of Life, and cost of living. The cost of living must be understood by anyone who wishes to assess regional disparities, plan a relocation, or make financially sensible decisions. This includes individuals, organizations, lawmakers, and scientists. This dataset contains a wide range of economic and demographic variables that taken together give an overall picture of the cost of living in different counties. By examining socioeconomic trends, identifying regions with greater or lower living expenses, and developing solutions to communities' affordability problems, we may analyze data on the cost of living in the United States. This dataset is necessary to make knowledgeable judgments about economics.

II. METHODOLOGY

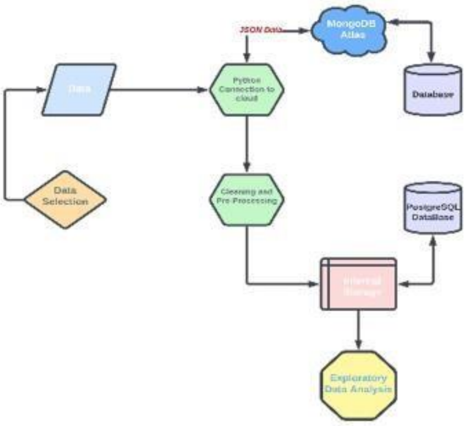


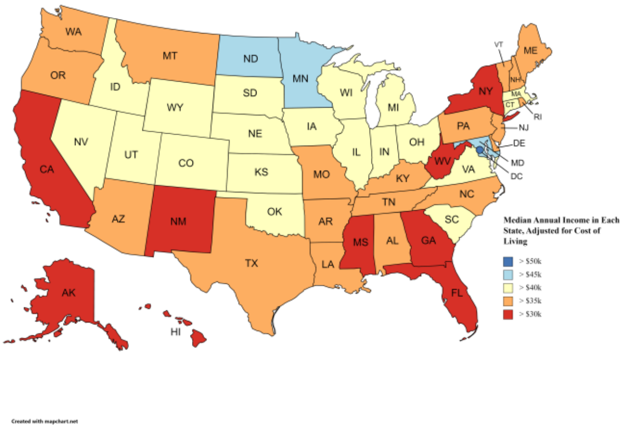
Fig: 1 Workflow

A. *Platforms Utilized:*

Our datasets are divided into two separate categories: semi-structured and structured formats. For semi-structured datasets, we used MongoDB as the storage platform and the Dagster library for ETL work to turn the information into a structured format. We then used pandas to aggregate all three data-frames into a single data frame, which we then merged into Postgresql using the join function.

B. *Process:*

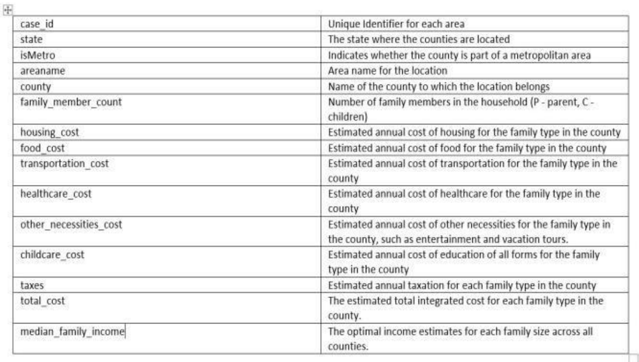
1) **Data Selection:** In this study, we used 3 distinct datasets, the specification of the same provided below.



a) Cost of Living: The dataset includes the cost of living in each state of the United States, as well as information for each county, with metro and non-metro segmentation. For additional analytical value, the data includes distinct family income data as well as specific feasible family size.

This dataset provides community-specific estimates for ten family types, including one or two parents and zero to four children, in all 1877 counties and metropolitan areas in the United States as shown in *fig [2]*.

Data Description:

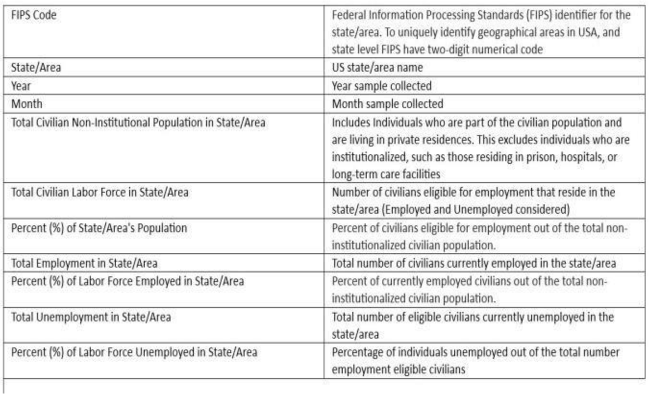


case_id	Unique Identifier for each area
state	The state where the counties are located
isMetro	Indicates whether the county is part of a metropolitan area
areaname	Area name for the location
county	Name of the county to which the location belongs
family_member_count	Number of family members in the household (P - parent, C - children)
housing_cost	Estimated annual cost of housing for the family type in the county
food_cost	Estimated annual cost of food for the family type in the county
transportation_cost	Estimated annual cost of transportation for the family type in the county
healthcare_cost	Estimated annual cost of healthcare for the family type in the county
other_necessities_cost	Estimated annual cost of other necessities for the family type in the county, such as entertainment and vacation tours.
childcare_cost	Estimated annual cost of education of all forms for the family type in the county
taxes	Estimated annual taxation for each family type in the county
total_cost	The estimated total integrated cost for each family type in the county.
median_family_income	The optimal income estimates for each family size across all counties.

Fig: 2 Data Description of US Cost of Living

b) Unemployment: This data gives us the view of relevant population statistics and employment rate across all US states from 1976 to 2022.The data is majorly sourced from Bureau of Labor Statistics which published employment data pertaining to America.

Data Description:



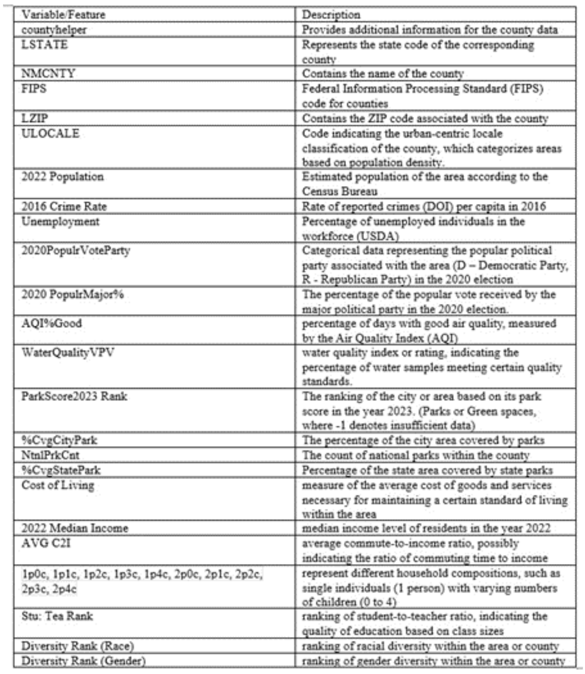
FIPS Code	Federal Information Processing Standards (FIPS) identifier for the state/area. To uniquely identify geographical areas in USA, and state level FIPS have two-digit numerical code
State/Area	US state/area name
Year	Year sample collected
Month	Month sample collected
Total Civilian Non-Institutional Population in State/Area	Includes Individuals who are part of the civilian population and are living in private residences. This excludes individuals who are institutionalized, such as those residing in prison, hospitals, or long-term care facilities
Total Civilian Labor Force in State/Area	Number of civilians eligible for employment that reside in the state/area (Employed and Unemployed considered)
Percent (%) of State/Area's Population	Percent of civilians eligible for employment out of the total non-institutionalized civilian population.
Total Employment in State/Area	Total number of civilians currently employed in the state/area
Percent (%) of Labor Force Employed in State/Area	Percent of currently employed civilians out of the total non-institutionalized civilian population.
Total Unemployment in State/Area	Total number of eligible civilians currently unemployed in the state/area
Percent (%) of Labor Force Unemployed in State/Area	Percentage of individuals unemployed out of the total number employment eligible civilians

Fig: 3 Data Description for Unemployment

c) Quality of Life: This data set provides the details at granular level, to analyze the quality of life in USA for each state and county level. The data collected here covers air and water quality, weather and climatic conditions, population analysis that classifies employed and unemployed rate along with the crime rate at multiple levels and specific details such as green spaces in the area. And all these attributes are segregated for different family sizes.

The major source of the data set is EPA (United States Environmental Protection Agency) that has the mission to protect human health and the environment.

Data Description:

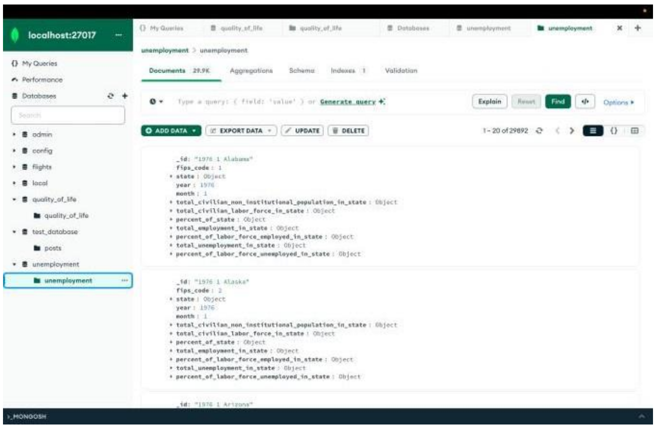


Variable/Feature	Description
countyhelper	Provides additional information for the county data
LSTATE	Represents the state code of the corresponding county
NMCNTY	Contains the name of the county
FIPS	Federal Information Processing Standard (FIPS) code for counties
LZIP	Contains the ZIP code associated with the county
ULOCALE	Code indicating the urban-centric locale classification of the county, which categorizes areas based on population density.
2022 Population	Estimated population of the area according to the Census Bureau
2016 Crime Rate	Rate of reported crimes (DOI) per capita in 2016
Unemployment	Percentage of unemployed individuals in the workforce (USDA)
2020PopulrVoteParty	Categorical data representing the popular political party associated with the area (D – Democratic Party, R - Republican Party) in the 2020 election
2020 PopulrMajor%	The percentage of the popular vote received by the major political party in the 2020 election.
AQI%Good	percentage of days with good air quality, measured by the Air Quality Index (AQI)
WaterQualityVPV	water quality index or rating, indicating the percentage of water samples meeting certain quality standards.
ParkScore2023 Rank	The ranking of the city or area based on its park score in the year 2023. (Parks or Green spaces, where -1 denotes insufficient data)
%CvgCityPark	The percentage of the city area covered by parks
NtnlPrkCnt	The count of national parks within the county
%CvgStatePark	Percentage of the state area covered by state parks
Cost of Living	measure of the average cost of goods and services necessary for maintaining a certain standard of living within the area
2022 Median Income	median income level of residents in the year 2022
AVG C2I	average commute-to-income ratio, possibly indicating the ratio of commuting time to income
1p0c, 1p1c, 1p2c, 1p3c, 1p4c, 2p0c, 2p1c, 2p2c, 2p3c, 2p4c	represent different household compositions, such as single individuals (1 person) with varying numbers of children (0 to 4)
Stu: Tea Rank	ranking of student-to-teacher ratio, indicating the quality of education based on class sizes
Diversity Rank (Race)	ranking of racial diversity within the area or county
Diversity Rank (Gender)	ranking of gender diversity within the area or county

Fig: 4 Data Description of US Quality of Life

C) *Data Warehousing and Retrieval:*

The MongoDB Atlas was used to hold the data once it was sourced using Python Programming. Before the data was uploaded to the MongoDB collections, it was made sure that it was all in JSON format.



```
_id: "1976 1 Alabama"
fips_code : 1
state : Object
year : 1976
month : 1
total_civilian_non_institutional_population_in_state : Object
total_civilian_labor_force_in_state : Object
percent_of_state : Object
total_employment_in_state : Object
percent_of_labor_force_employed_in_state : Object
total_unemployment_in_state : Object
percent_of_labor_force_unemployed_in_state : Object

_id: "1976 1 Alaska"
fips_code : 2
state : Object
year : 1976
month : 1
total_civilian_non_institutional_population_in_state : Object
total_civilian_labor_force_in_state : Object
percent_of_state : Object
total_employment_in_state : Object
percent_of_labor_force_employed_in_state : Object
total_unemployment_in_state : Object
percent_of_labor_force_unemployed_in_state : Object

_id: "1976 1 Arizona"
```

Fig: 5 Reference for MongoDB Atlas

D) *Data Cleaning and Preprocessing:*

During this process, each teammate thoroughly analyzed all three datasets to ensure that all null values, missing values, duplicates and unnecessary data/columns were handled correctly. The cleansed, pre-processed data was then put in a PostgreSQL database.

E) *Exploratory Data Analysis:*

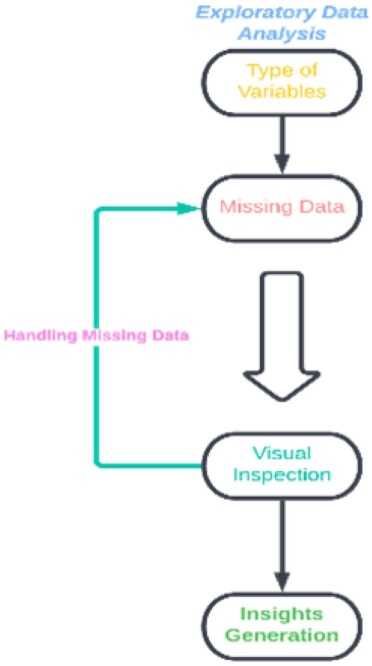


Fig: 6 Exploratory Data Analysis Process

The processed and combined data was fetched from PostgreSQL and to be accurate and precise each of us further studied the data, to understand the core competence of the features available.

Once we made sure that the data quality is top notch, each of us started visualizing and extracting effective insights from the respective datasets.

III. Results

Through the data exploration done, we can clearly see the major source of the collected data is EPI (Economic Policy Institute) which provides the family budget calculations for United States.

- ***Dataset 1: Unemployment in US***

In this research paper, we carry out the necessary exploratory data analytics to conclude the most inhabitable and adaptable state or area in the whole of USA by including all the 50 states along with three more areas (Columbia, New York City, Los Angeles-Long Beach- Glendale) to increase the analytical potential of the dataset. And to begin with, USA which has one of the world’s most vast and diverse economy and has a strong culture of innovation and entrepreneurship since its home to many startups and tech giants and to compliment all these, US has one of the largest consumer markets globally.

From our selective data, we have the employment and unemployment data ranging from the year 1976 to 2022, and the percentage of average employment all over USA has increased gradually as the total working population of the country increases along the range of years, which is a positive sign that the country has been growing ever since 1976 to present. While carefully analyzing we can see a significant decline in employment during the COVID-19 Pandemic and there is no other major impact for the job market of the USA.

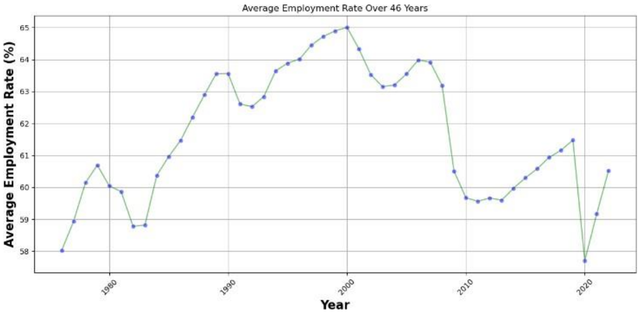


Fig: 7 Average Employment Rate Over 46 Years

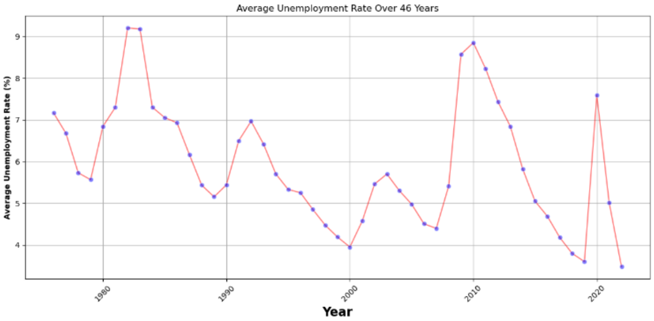


Fig: 8 Average Unemployment Rate Over 46 Years

Fig 8 shows the unemployment rate over the period of 1980-2020. After 1982, the unemployment rate peaked in the year 2010 and gradually came to a lowest percent. The rate again spiked in the year 2020 and it has been decreasing ever since.

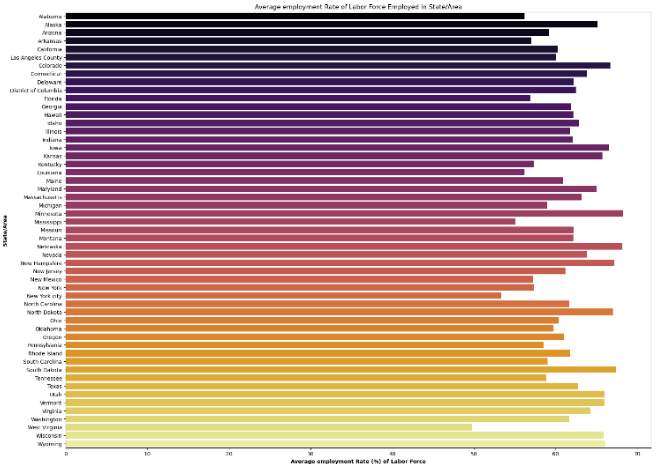


Fig: 9 Average Employment Rate of Labor Force

Fig 9 represents the rate of labor force in each state, where Minnesota state has the highest, followed by Nebraska and West Virginia has the lowest amount of labor force.

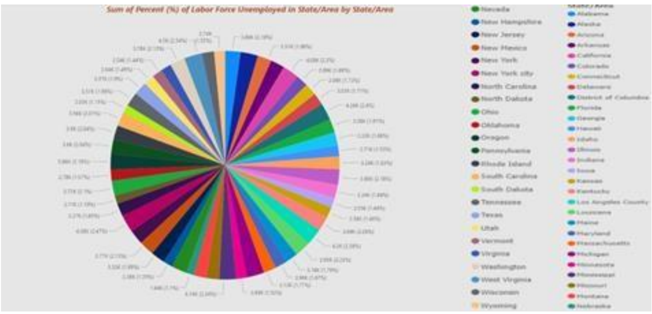


Fig: 10 Sum of Percent of Labor Force in State

Above image shows the unemployment rate in each state from which it can be seen that West Virginia has the highest and Nebraska has the lowest percent of unemployment.

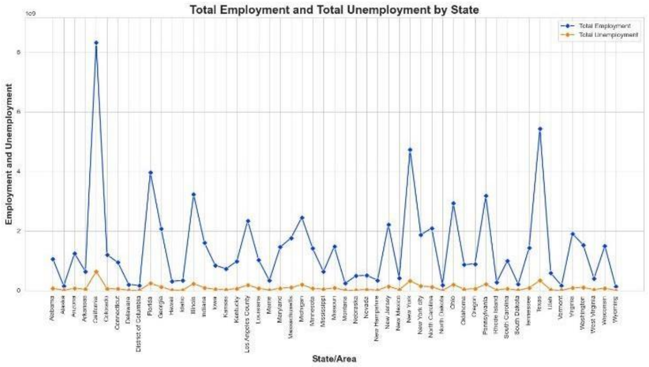


Fig: 11 Total Employment and Total Unemployment by State

Here as we can see, the two major features are employment and unemployment rate in each of the states in the US. The plot perfectly represents the states that have leading edge opportunities and provide high employment such as California, Texas and New York which are all leading IT hubs and high-tech industrialized regions. For career growth and entrepreneurial opportunities, we can clearly understand the locations to go for which maximum confidence.

- ***Dataset – 2: Cost of Living in US***

The dataset was acquired from the Economic Policy Institute (EPI), which provides family budget projections for the United States. It displays the cost of living for each state, considering all the important elements that contribute to the overall cost of living.

Factors such as transportation costs, healthcare costs, housing costs, childcare costs, food costs, and other necessities are estimated. As the above few factors, covers almost all kinds of spending by an average household.

Here from the plot, we can tell the difference between the total cost of living of each state in the USA. This would give us a general view that which of the states is the more expensive and which ones are more cost effective and economically friendly. However, as we all know, cost alone will not indicate which area is best suited to live or migrate to; however, it is one of the critical variables that allows people to rank or even choose a location that is better than others.

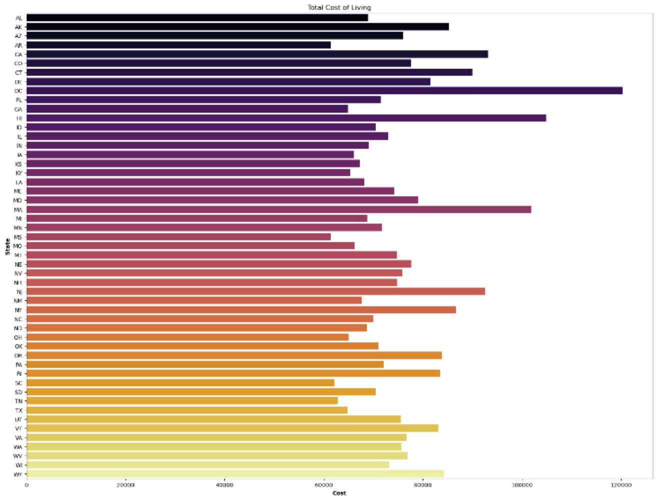


Fig: 12 Total Cost of Living

And to be more precise we looked even further into the data for classified cost values as it enables us to focus on the most critical factor that affects the total cost of an area or a state.

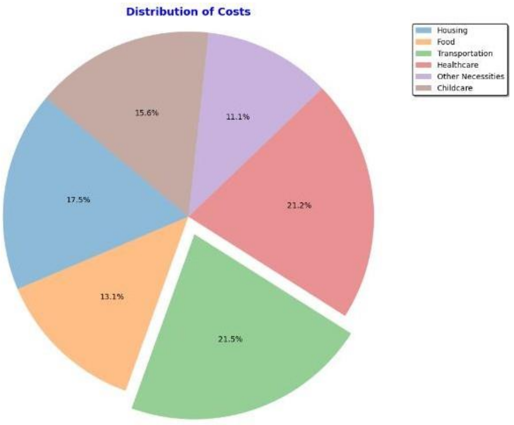


Fig: 13 Distribution of Costs

This pie representation makes it easy to determine which factors are most crucial and which are least critical. Basically, transportation and healthcare expenditures are the most critical, and more attention is needed to address these issues.

So far, we have seen the cost and its many components, which finish the cost of the US state. It is now equally crucial for us to assess and estimate the survival rate of the same state based on the expenses we have seen. Like the cost of each state, we make the chart for median family income.

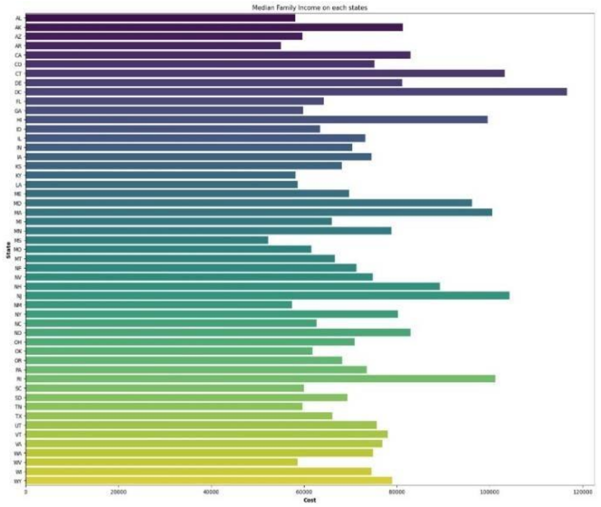


Fig: 14 Median of Family Income on each States

Here let’s take some samples, to reason with choosing the best state and the most challenging one. Best state in the sense that might even have high living expenses or lower, but either way the state should give people the opportunity to utilize their skills and with that enables them to survive comfortably even in those high expectations with ease.

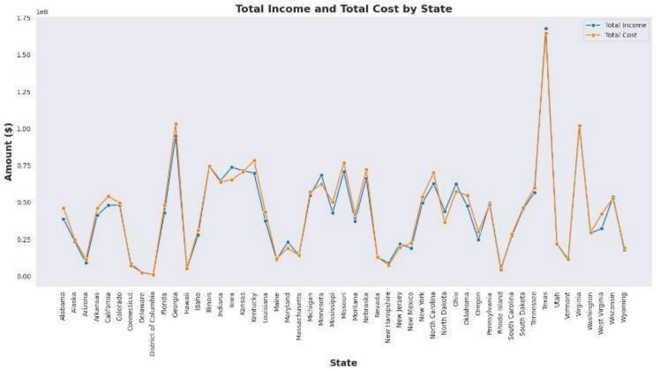


Fig: 15 Total Income and Total Cost by State

One such example is Washington, DC, which has the highest household expense level but also provides the most opportunities for workable people to shine and live with ease and comfort, which also sprouts happiness in the family, along with the negative byproduct of a busy schedule for parents, but it is acceptable to a certain extent. And we may see for ourselves what other locations fit the above statements.

Now we'd shift our focus to more accurate analytics, looking at costs and expenses from a different angle. We'd break it down into metro areas versus non-metro regions, and the pictorial illustration below makes it clear to see how the two aspects differ.

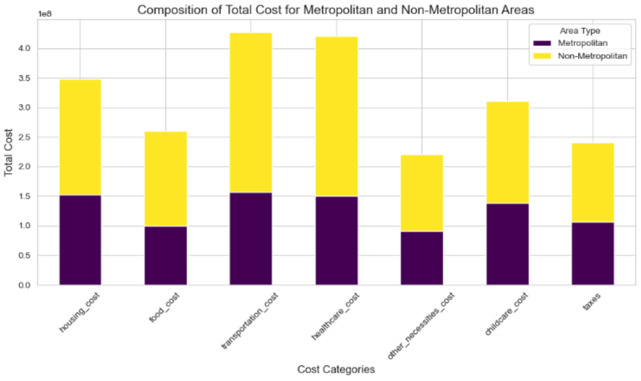


Fig: 16 Composition of Total Cost for Metropolitan and Non-Metropolitan Areas

To have a better understanding of this truth, break it down even further. We captured the cost classification effect across the USA, with metro and non-metro, respectively.

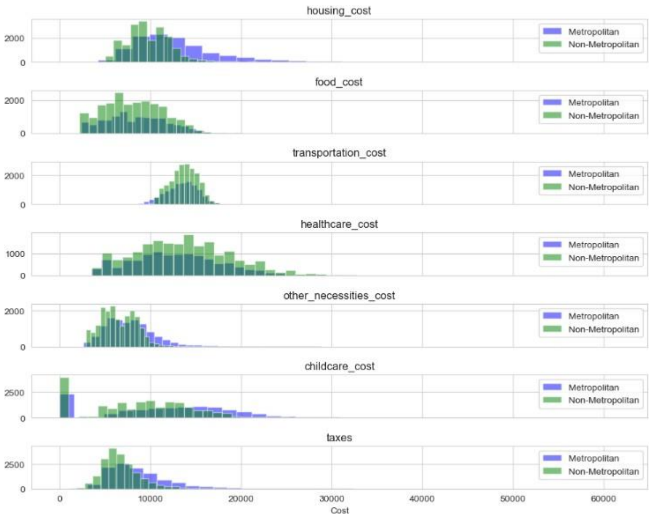


Fig: 17 Cost of each sector by Metropolitan and Non-Metropolitan States

The findings here may contradict our expectations, given the cost of living is higher in non-metro regions than in metropolitan areas. This helps us understand and gain a lot of insights from all the above depictions; metro regions, despite being hectic and crowded, provide several benefits for cost-effective solutions for each situation, such as feasible and frequent transportation and housing allowances, education scholarships, and support for entrepreneurship.

In contrast, these benefits are not reasonable in small and medium-sized non-metropolitan areas. To end with the total study, portrayals, and collective insights, we can say the best region ideal to establish a business, enhance our job abilities, and relocate from a foreign. For someone looking to settle down in a quiet and calm location non-metro is suitable, this would be compatible for US native civilians.

- ***Dataset – 3: Quality of Life in US***

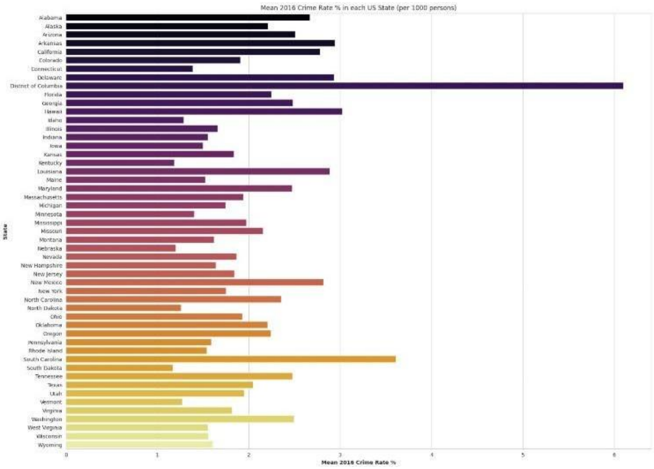


Fig: 18 Mean of Crime Rate in the year 2016

The above figure shows the mean crime rate per 1000 persons in the year 2016. It can be observed that the District of Columbia has the highest crime rate. South Dakota records the lowest crime rate.

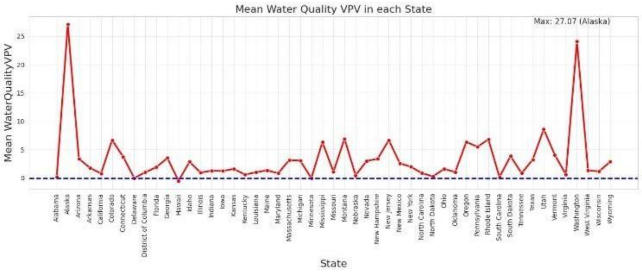


Fig: 19 Mean of Water Quality

Water quality can be observed from the above figure where the quality of water is better in Alaska and Washington, compared to the other states.

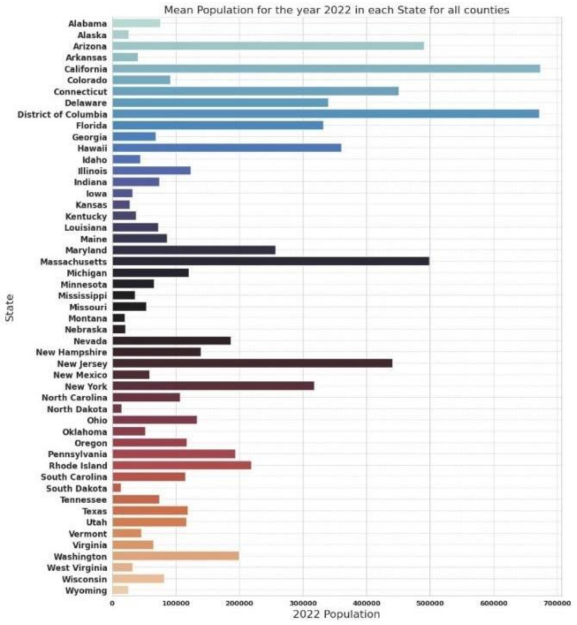


Fig: 20 Mean of Population for the year 2022

From the above image, the mean of population in the year 2022 is visualized where California has the highest overall population.

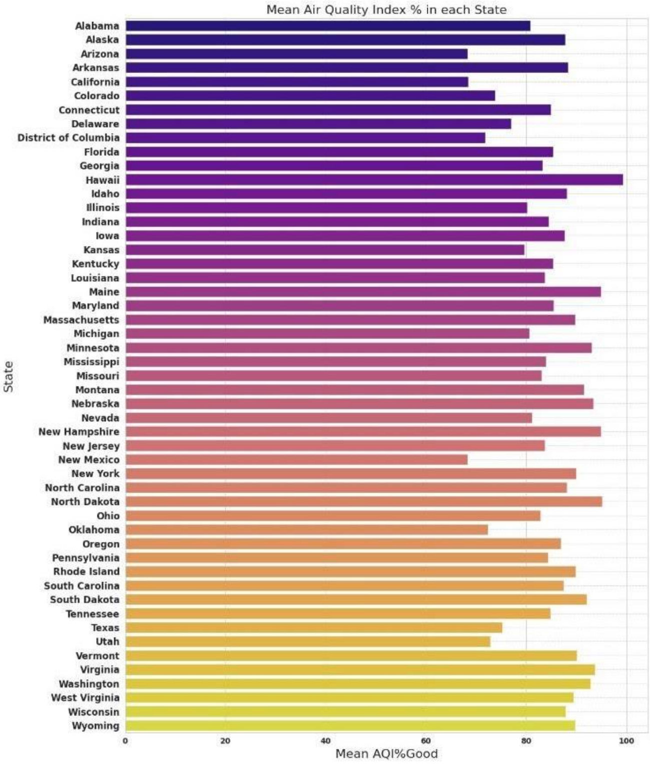
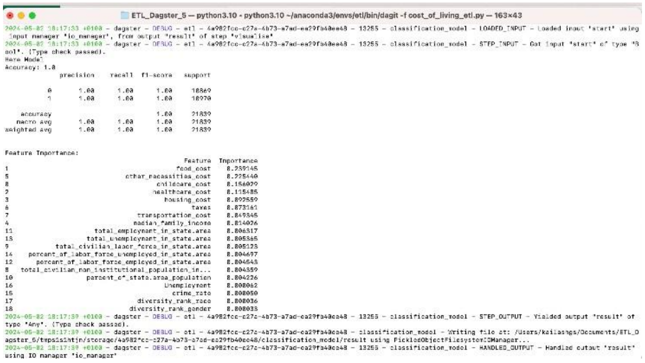


Fig: 21 Mean of Air Quality Index

The image above shows the mean of air quality index in each state. Hawaii has better air quality than other states, whereas Arizona and California have the worst quality of air.


ETL_Dagster_5 — python3.10 · python3.10 ~/anaconda3/envs/etl/bin/dagit -f cost_of_living_etl.py — 163×43

```
2024-05-02 18:17:33 +0100 - dagster - DEBUG - etl - 4a982fcc-c27a-4b73-a7ad-ea29fb40ee48 - 13255 - classification_model - LOADED_INPUT - Loaded input 'start' using
 input manager 'io_manager', from output 'result' of step 'visualise'
2024-05-02 18:17:33 +0100 - dagster - DEBUG - etl - 4a982fcc-c27a-4b73-a7ad-ea29fb40ee48 - 13255 - classification_model - STEP_INPUT - Got input 'start' of type "B
ool". (Type check passed).
Here Model
Accuracy: 1.0
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     10869
           1       1.00      1.00      1.00     10970

    accuracy                           1.00     21839
   macro avg       1.00      1.00      1.00     21839
weighted avg       1.00      1.00      1.00     21839


Feature Importance:
                                             Feature  Importance
1                                          food_cost    0.239145
5                              other_necessities_cost    0.225440
0                                      childcare_cost    0.156029
2                                     healthcare_cost    0.115485
3                                        housing_cost    0.092559
6                                               taxes    0.073161
7                                 transportation_cost    0.049345
4                                median_family_income    0.014026
11                  total_employment_in_state.area    0.006317
13                total_unemployment_in_state.area    0.005365
9          total_civilian_labor_force_in_state.area    0.005123
14  percent_of_labor_force_unemployed_in_state.area    0.004697
12    percent_of_labor_force_employed_in_state.area    0.004543
8   total_civilian_non_institutional_population_in...    0.004359
10                percent_of_state.area_population    0.004226
16                                      Unemployment    0.000062
15                                         crime_rate    0.000050
17                               diversity_rank_race    0.000036
18                             diversity_rank_gender    0.000033
2024-05-02 18:17:39 +0100 - dagster - DEBUG - etl - 4a982fcc-c27a-4b73-a7ad-ea29fb40ee48 - 13255 - classification_model - STEP_OUTPUT - Yielded output 'result' of
type "Any". (Type check passed).
2024-05-02 18:17:39 +0100 - dagster - DEBUG - etl - 4a982fcc-c27a-4b73-a7ad-ea29fb40ee48 - classification_model - Writing file at: /Users/kailashgs/Documents/ETL_D
agster_5/tmps1s1htjn/storage/4a982fcc-c27a-4b73-a7ad-ea29fb40ee48/classification_model/result using PickledObjectFilesystemIOManager...
2024-05-02 18:17:39 +0100 - dagster - DEBUG - etl - 4a982fcc-c27a-4b73-a7ad-ea29fb40ee48 - 13255 - classification_model - HANDLED_OUTPUT - Handled output 'result'
using IO manager 'io_manager'
```

Fig: 22 Accuracy of Random Forest Model

Comparing the output from the provided console values with the code snippet and analysis, it's apparent that the Random Forest Classifier achieved a perfect accuracy of 1.0. This indicates that the model perfectly predicted the total_cost for all instances in the test set. The precision, recall, and F1-score values for both classes (0 and 1) are also perfect, indicating that the model performed exceptionally well in classifying both categories.

However, it's essential to interpret these results cautiously. While the model achieved perfect accuracy on the test set, there's a possibility of overfitting, especially given the complexity of the model and the high dimensionality of the feature space. Additionally, further analysis is warranted to validate the model's generalizability and robustness, such as assessing its performance on unseen data and exploring potential limitations or biases in the dataset. Nonetheless, the feature importance analysis provides valuable insights for understanding the underlying factors driving the total_cost and can support decision-making processes related to cost estimation and resource allocation.

IV. Conclusion

Have completed all potential visualizations that support our goal for this project. After careful consideration, we can confidently state that the United States will continue to be a beacon of hope for our world in terms of opportunities, inventions, and career growth for international society. While it may be difficult due to high living costs, it is always willing to provide appropriate and sustainable income.

Future Work:

- **Comparative Analysis:** To find patterns and differences, do a thorough comparative study of the cost of living, unemployment rate and standard of living in several US states.
- **Demographic Analysis:** Determine how age, gender, ethnicity and educational attainment affect living expenses, unemployment and living standards. This can help identify suitable areas for focused intervention or policy modifications.
- **Improved Modelling:** Create predictive models using historical data and external variables like economic metrics and legislative modifications to predict future trends in living expenses, unemployment and life’s quality.