Student Name: Hua Zhang

Panther ID: 6355065

Course : CAP4770 U01 1241

Professor: Dr Kaoutar Ben Ahmed

Data Understanding & Preprocessing Report

Description

This dataset provides a broad understanding of the financial health of software developers in major U.S. cities and metropolitan areas. We explored differences between states and cities in terms of average software developer salary, median home price, average cost of living, average rent, cost of living plus average rent, and average local purchasing power. With this dataset, we can gain insights into how to better understand which fields are more financially viable than others when looking for employment opportunities in software development. This data allows us to discover patterns among certain geographies in order to identify other compelling financial opportunities that software developers may benefit from.

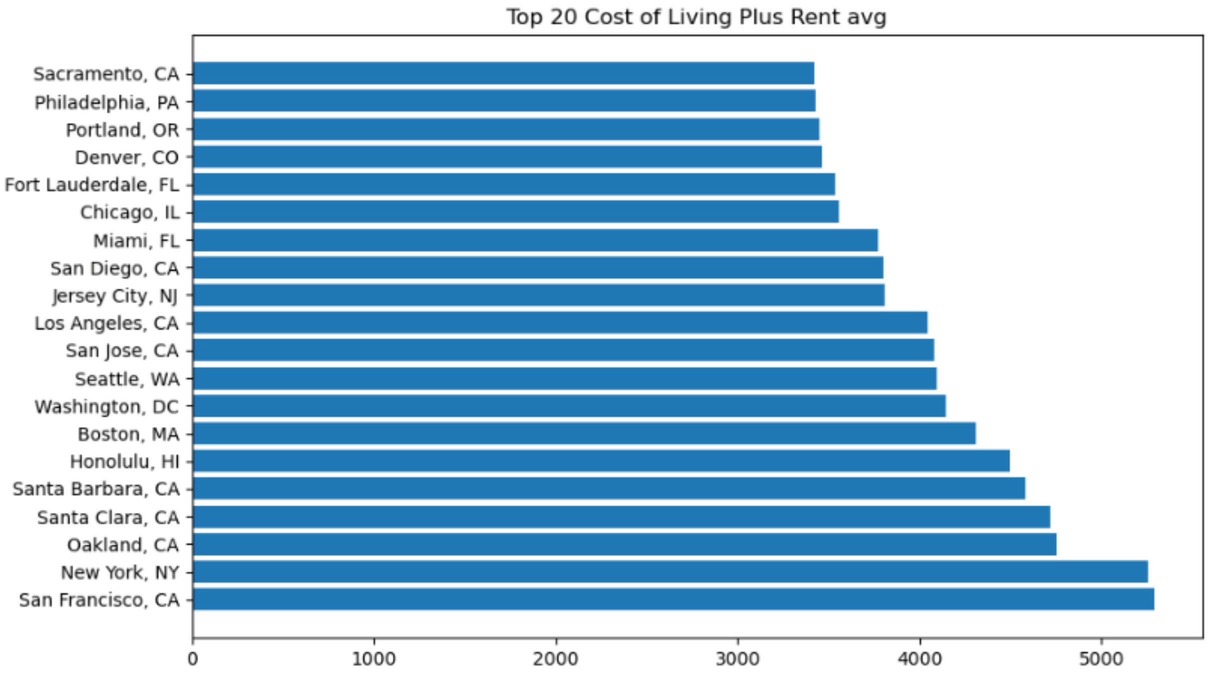
Observations

The most Cost of Living plus rent is **New York**.

The mean salary and numbers of jobs to software developer, they **don't** have association.

Where is the **most numbers of Jobs** for software developer.

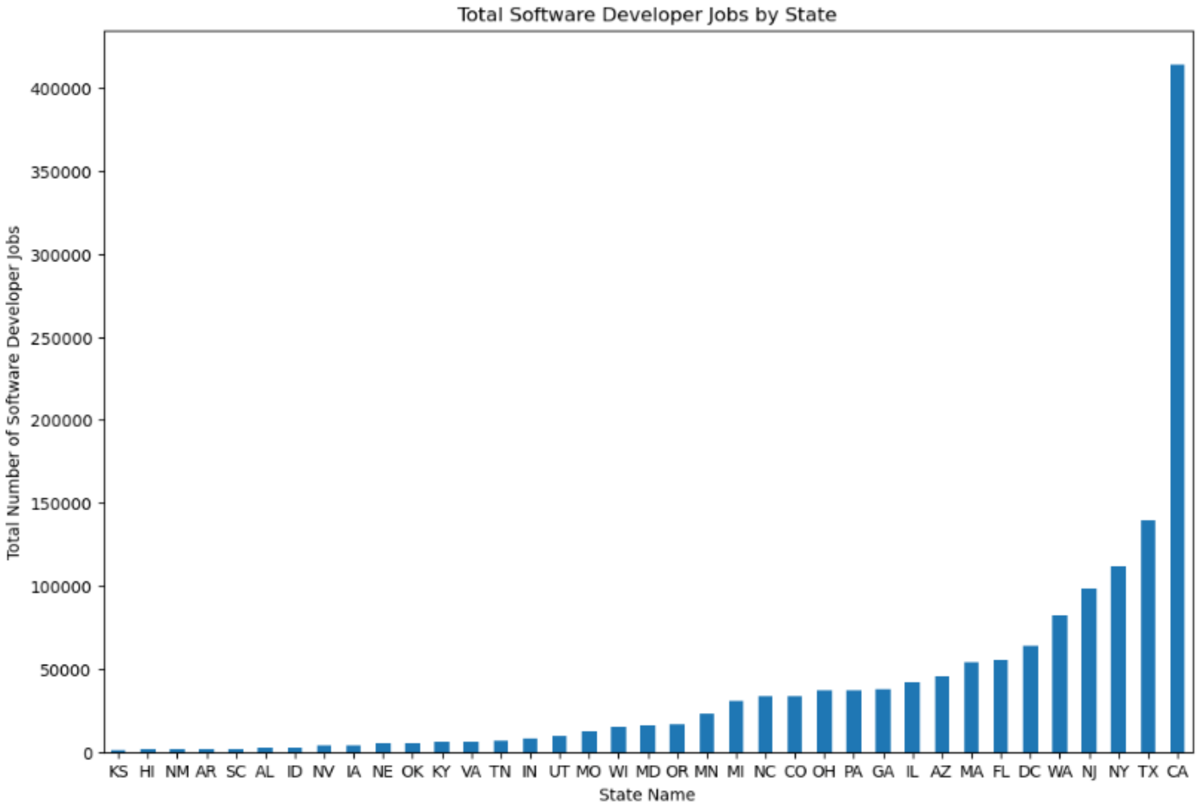
Visualizations



Explanation figure (Top 20 Cost of Living Plus Rent avg)

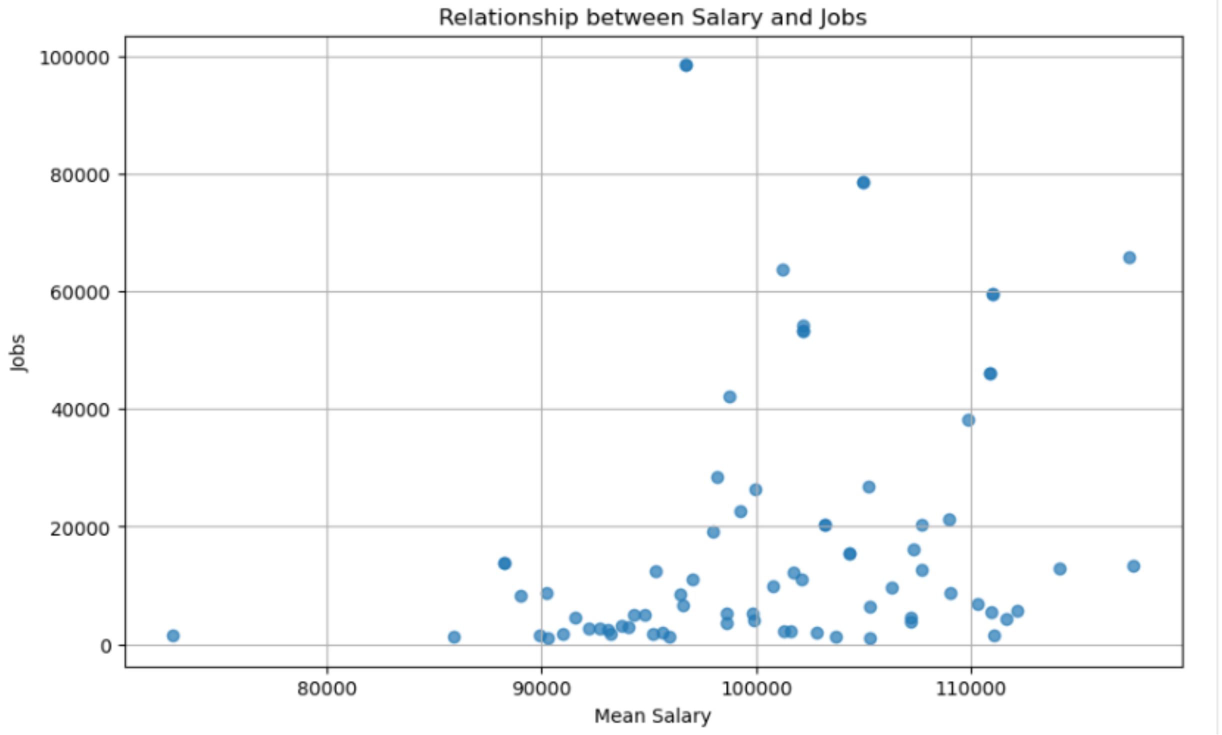
This graphic provides a visual representation of the cost of living in various locations. Each bar on the graph represents a city, and the height of the bar corresponds to the combined cost of living and rent in that city.

The information is valuable as it allows for quick comparison and identification of cities with higher living expenses. For individuals considering relocation or businesses exploring new locations, this graph offers insights into the financial aspects of residing in different areas, aiding in decision-making and financial planning. It serves as a concise and informative snapshot of the relative affordability of these cities.



Explanation figure (Total Software Developer jobs by State)

The bar chart illustrating the total number of software developer jobs by state provides a clear overview of the demand for developers in different regions. Each bar represents a state, with the height indicating the total number of software developer jobs. This information is crucial for job seekers, employers, and policymakers as it highlights the distribution of opportunities across states.



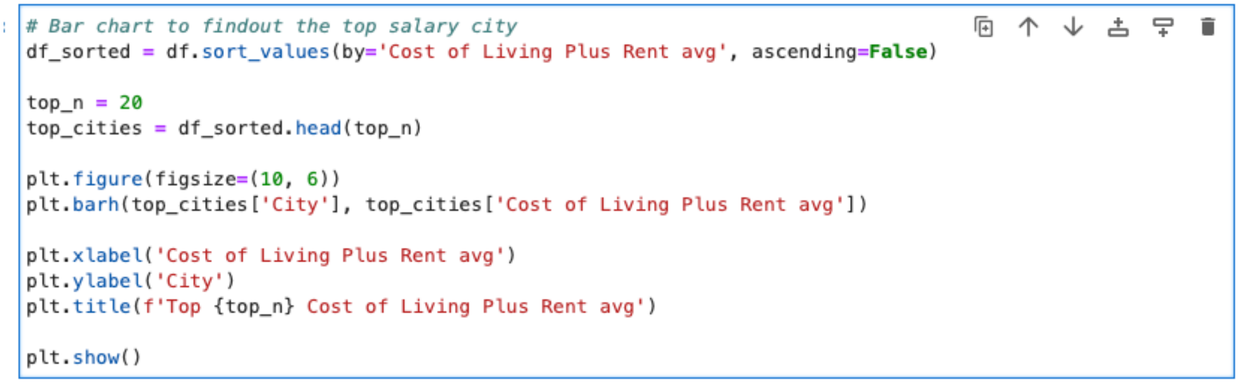
Explanation figure (Relationship between Salary and jobs)

A graphical representation of the relationship between average software developer salary and number of jobs provides insight into the employment outlook. Each point on the scatter plot corresponds to a city, with the x-axis representing average salary and the y-axis representing the number of software developer jobs. This visualization can identify potential correlations or patterns. For job seekers and employers, it can quickly assess the relationship between salary levels and job opportunities. It is obvious in the sample data. There is no relationship between them, and there is no relationship between your salary level and the number of job opportunities in the area.

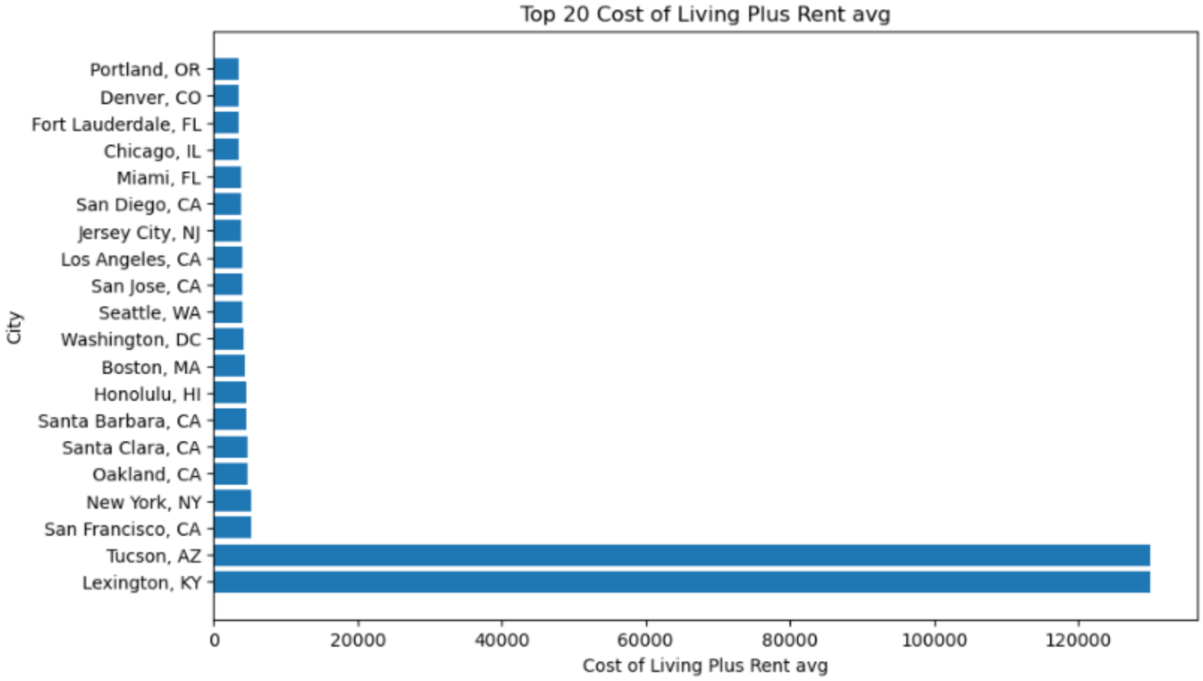
Data cleaning and preprocessing

- Data cleaning

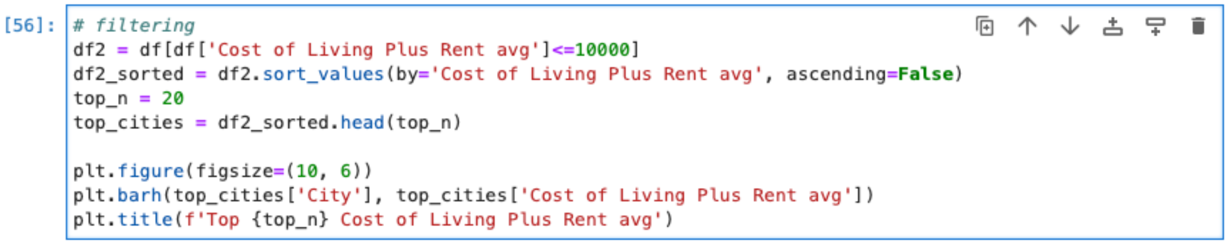
  Find out outliers then remove. When I making graph to check which city have highest cost of living. The result shows two outliers (Tucson AZ and Lexington KY), because it much far from other instances.



```
# Bar chart to findout the top salary city
df_sorted = df.sort_values(by='Cost of Living Plus Rent avg', ascending=False)

top_n = 20
top_cities = df_sorted.head(top_n)

plt.figure(figsize=(10, 6))
plt.barh(top_cities['City'], top_cities['Cost of Living Plus Rent avg'])

plt.xlabel('Cost of Living Plus Rent avg')
plt.ylabel('City')
plt.title(f'Top {top_n} Cost of Living Plus Rent avg')

plt.show()
```



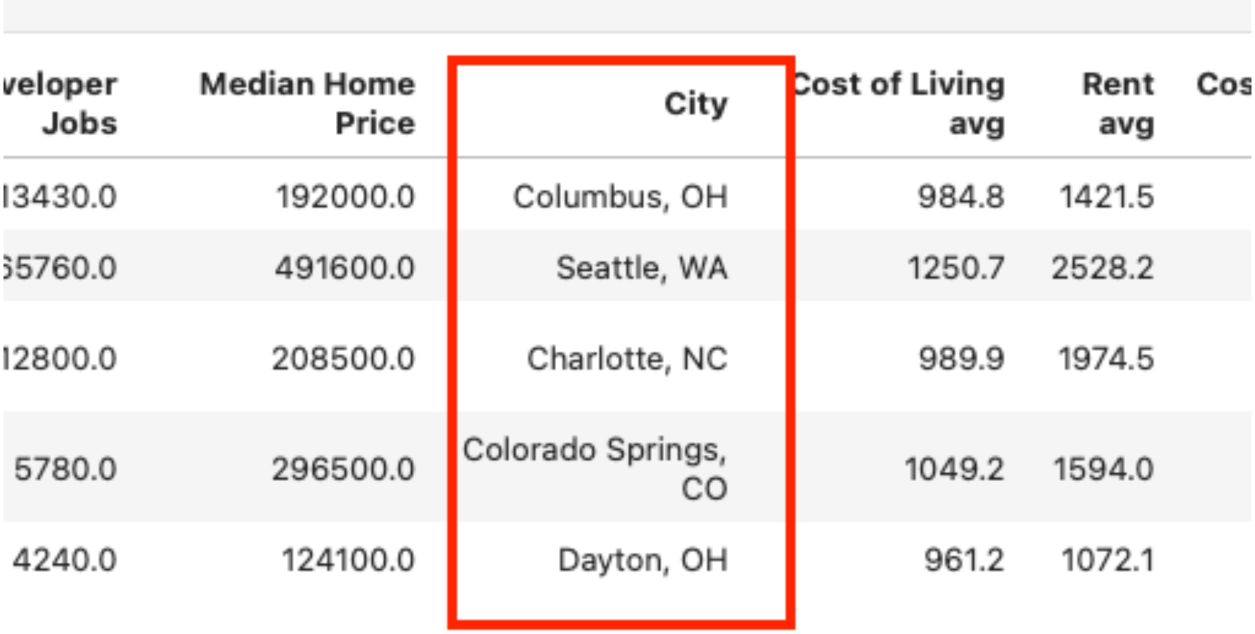
- To avoid this to impact our result I removed them.



```
[56]: # filtering
df2 = df[df['Cost of Living Plus Rent avg']<=10000]
df2_sorted = df2.sort_values(by='Cost of Living Plus Rent avg', ascending=False)
top_n = 20
top_cities = df2_sorted.head(top_n)

plt.figure(figsize=(10, 6))
plt.barh(top_cities['City'], top_cities['Cost of Living Plus Rent avg'])
plt.title(f'Top {top_n} Cost of Living Plus Rent avg')
```

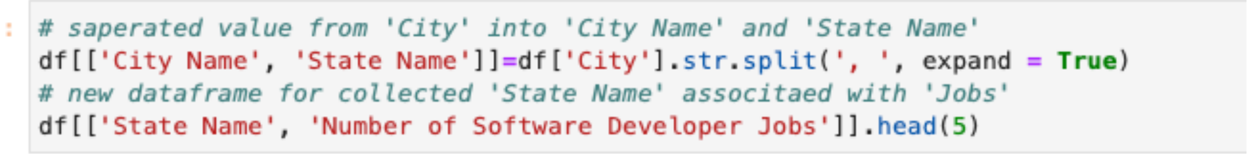
- Preprocessing

  I want to know the sum of the Jobs for each state. But it’s don’t give us state feature directly. So, I am using feature extraction to achieve it.

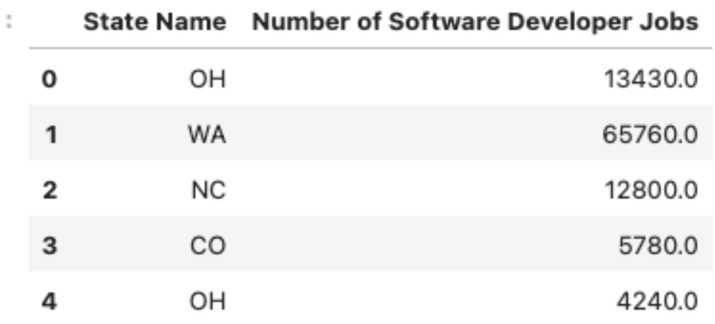


veloper Jobs	Median Home Price	City	Cost of Living avg	Rent avg	Cos
13430.0	192000.0	Columbus, OH	984.8	1421.5	
65760.0	491600.0	Seattle, WA	1250.7	2528.2	
12800.0	208500.0	Charlotte, NC	989.9	1974.5	
5780.0	296500.0	Colorado Springs, CO	1049.2	1594.0	
4240.0	124100.0	Dayton, OH	961.2	1072.1	

  - We find out the feature what we need in the city column.



```
# saperated value from 'City' into 'City Name' and 'State Name'
df[['City Name', 'State Name']]=df['City'].str.split(', ', expand = True)
# new dataframe for collected 'State Name' associtaed with 'Jobs'
df[['State Name', 'Number of Software Developer Jobs']].head(5)
```



	State Name	Number of Software Developer Jobs
0	OH	13430.0
1	WA	65760.0
2	NC	12800.0
3	CO	5780.0
4	OH	4240.0

  - Then using ‘split’ command to extract state value.

Techniques

- Library

  pandas

  numpy

  matplotlib

- Method

  sort_values(): *sorts values in a DataFrame along the selected axis and returns a DataFrame with sorted values or None.*

  str.split(): splits a string into a list. You can specify the separator, default separator is any whitespace.