Sarah Liang
08/13/2024

The why behind this project

As a Californian resident for 20+ years, the mix of rising housing prices and the increasingly competitive and impacted job market is making out-of-state opportunities more and more appealing. The prospect of leaving California to further my professional career and life, is difficult to grasp, especially when all my family lives in the San Francisco Bay Area.

Furthermore, the job search for data analyst and/or business analyst positions often require and actively use Microsoft Excel proficiency skills. This is my opportunity to do a deep dive on the living situations in different U.S. states, as well as practice and hone my Excel and Tableau skills.

Motivating research questions

1. What is the situation in California right now; How are California residents faring in terms of how their income matches up to living expenses?
2. Would Californian residents benefit from leaving California and taking job opportunities out of state? Will returning to California after a period of time be plausible?
3. How does the pay for jobs in the tech industry compare across states?
4. Would relocation prove beneficial to a professional career in tech? If so, which states would provide the most financial/retirement savings?
5. Is it financially plausible to have children in a single income household? If not, double income?

Datasets

The final dataset is a combination of information pulling from the following publically available datasets:

- U.S. Cost of Living Dataset (1877 Counties) from Asaniczka on Kaggle
- May 2023 National Occupational Employment and Wage Estimates from the U.S. Bureau of Labor Statistics
- U.S. Counties Database (data pulled from authoritative sources) from Simple Maps

Excel Workbook files

See the Github Readme for this project for a full description (including variables) of all datasets in the Excel workbook.

File Name	Description
`us-cost-of-living-2023`	The main Excel dataset comprising of U.S. cost of living data for 1877 counties.
`us-occupation-salary-may2023`	Dataset with occupational income information, corresponding to U.S. cities and counties.

File Name	Description
`uscounties`	Dataset with U.S. counties information, including latitude and longitude geographical information.
`us-states`	Lookup table of U.S. abbreviation and full name data.
`us-occ-swdev`	Dataset/lookup table of filtered software developer income information from `us-occupation-salary-may2023`; includes variables `area_title`, `OCC_CODE`, and `A_MEAN`
`us-occ-ds`	Dataset/lookup table of filtered data scientist income information from `us-occupation-salary-may2023`; includes variables `area_title`, `OCC_CODE`, and `A_MEAN`
`us-occ-dataentry`	Dataset/lookup table of filtered data entry income information from `us-occupation-salary-may2023`; includes variables `area_title`, `OCC_CODE`, and `A_MEAN`
`us-occ-statassistant`	Dataset/lookup table of filtered statistical assistant income information from `us-occupation-salary-may2023`; includes variables `area_title`, `OCC_CODE`, and `A_MEAN`
`us-occ-acutary`	Dataset/lookup table of filtered actuary income information from `us-occupation-salary-may2023`; includes variables `area_title`, `OCC_CODE`, and `A_MEAN`

Data Cleaning and Manipulation in Microsoft Excel

A combination of techniques were used to combine necessary data into a single dataset for further visualization purposes. The U.S. Cost of Living Dataset had the most useful information, so supplemental data from other datasets was pulled and appended to the former. Here is a general workflow of steps achieved:

1. `CONCATENATE()` and `VLOOKUP()` were used to pull geographical latitude and longitude data from the U.S. Counties Database into a single variable `geo_point`, formatted as `latitude, longitude`.

   For example the geo_point cells were filled using the formula:

   ```
   CONCATENATE(
       VLOOKUP(county, us_counties, 5, FALSE)
       &", "&
       VLOOKUP(county, us_counties, 6, FALSE)
       )
   ```

   where columns 5 and 6 in the us_counties Excel sheet were latitude and longitude values respectively.

2. `IF()`, `ISNA()`, `VLOOKUP()`, `CONCATENATE()`, and filtering functions were used to include `median_[occupation]_salary` for different occupations in the tech industry. First, annual mean income for different occupations was filtered from the May 2023 National Occupational Employment and Wage Estimates dataset and deposited into their respective Excel sheets named `us-occ-actuary`, `us-occ-swdev`, and such.

An example formula like this was used to add occupational mean income for actuaries, software developers, and data scientists as variables `median_actuary_salary`, `median_swdev_salary`, and `median_ds_salary` to the main U.S. Cost of Living Dataset.

```
IF(
    ISNA(VLOOKUP(CONCATENATE(area_name&", "&state),
    us-occ-actuary, 4, FALSE)),
    IF(
        ISNA(VLOOKUP(state_name, us-occ-actuary,
        4, FALSE)),
        VLOOKUP("U.S.", us-occ-actuary, 4, FALSE),
        VLOOKUP(state_name, us-occ-actuary,
        4, FALSE)),
    VLOOKUP(CONCATENATE(area_name&", "&state),
    us-occ-actuary, 4, FALSE)
    )
```

In terms of pseudo-code and/or logic, the above formula is searching the occupational income data in priority of city-level mean income, then state-level mean income, then the U.S. national mean income for that particular occupation.

  - It is important to note that all the occupational income data include the **state level** for these three occupations: actuaries, software developers, and data scientists. I.e. the U.S. national mean income is not included in the accumulated dataset.

3. The variables `median_savings`, `median_savings_actuary`, `median_savings_swdev`, and `median_savings_ds` were calculated as the difference between the median annual income or respective occupational salaries and the total estimated annual cost of living.

```
For example, median_savings_actuary = median_actuary_salary - total_cost.
```

4. The `family_member_count` variable was split using delimiters into two columns of numerical values: `parent` and `child`. Similarly, the variable `areaname` was split into `areaname` and `state`.

5. A variable `state_name` was created for convenience using `VLOOKUP()` and a lookup table/Excel sheet `us-states` with two columns of state abbreviations and their full names.

Interactive data visualization

Due to the data size and for aesthetic reasons, I want to leverage Tableau to create the final interactive data visualization. MySQL is also being experimented with at the moment.

Through data visualization, I would like to convey the financial living situation in each U.S. state (how to maximize savings, how savings/cost can be impacted by other factors, etc.). There should be a factor of comparing income levels in different occupations across states, leading viewers to determine whether or not other states may provide better financial opportunities than their original state. The goal here is to display the following plots:

- **interactive geographical U.S. map** `(states and counties)` highlighting the annual `income` levels, `cost of living`, and amount of potential `savings`, dependent on the `family type` and `occupation`
  - *This plot aims to be the main plot. It will convey the most information about the U.S. 2023 financial situation, and provide some insight into tech occupations.*
- **bar plot** highlighting the `cost of having children` and the effect of having children on `total cost` and/or `median savings` dependent on a `single or double income` household
  - *This plot will help answer questions about the percentage of income that goes into raising children, etc.*
- **geographical plot** highlighting the `cost of housing` in `metropolitan` area/counties vs. not
- **scatter or bar plot** highlighting the `annual savings` as other plots and settings are toggled
  - *explores how other factors directly impact savings; also highlight maximized savings/retirement potential: savings that are 15% of pretax savings*

Results and meaningful insights