1. Introduction

1.1 Background

In the modern era with a globally connected economy, young professionals are increasingly willing to relocate for the right opportunity. Often these opportunities come from among many active applications for the applicant, with 30 days being a typical timeframe for their relocation. 30 days is a crunch for any individual to assess a living situation and accurately anticipate the actual annualized cost of living that relies on more than simply the monthly rental fee associated with living. How can a young professional quickly and easily determine what areas are within their budget and what their commute will be like? This capstone will aim to provide a young professional high-level information about areas surrounding a potential job location so that they can make an informed subjective decision with a more complete dataset than would otherwise likely be achievable in such a short timeframe for relocation.

1.2 Problem

This idea is focused on the state of Massachusetts with a job based in the center of Boston. Selfishly, the idea is derived from my own struggles to find an appropriate living situation when a job opportunity presented itself in a similar way in another city. Knowing the stress of the scramble for living accommodations, this tool could be of great use to anyone seeking deeper knowledge about cost of living on a tight timeline. There is plenty of relevant data but much of it scattered or focused on individual property prices rather than area cost and the true cost of living that is often centered in the nearest metropolitan area and lowered as one moves away. By taking available property data and improving the context in which it is

presented, a clearer comparison will be available for the target audience to make the most informed decision.

1.3 Interest

While the scope has been intentionally limited to Massachusetts with a job in the Boston area, the extension of the application would be quite simple to execute with an exponential increase in data and added user input functions for job location. Any young professional or professional looking to quickly relocate for a new opportunity would be glad for concise information on what monthly living and commuting expenses might be. In turn this could benefit the companies looking to hire those individuals by having a method to provide objective and contextualized data to prospective hires, making the decision to accept that much easier.

2. Data Sources and Cleaning

2.1 Data Approach

For the given problem, the number of cost factors is quite significant and so the scope will be limited to giving a holistic view of recurring monetary costs associated with accommodations and transportation, namely median rent in all surrounding areas and the commuting distance to the job location to be extrapolated to cost. For this the following data is required:

- A list of cities/towns and counties in a state
- Geolocation data for cities/towns and counties for map display
- Home pricing data from the active housing market
- Reported apartment rental prices

For this data to be manageable and appropriately complex for this project, the scope is limited by the following assumptions:

- The target audience is searching for a job based in the heart of Boston, MA
- The target audience has excellent credit
- The target audience is a potential first time home buyer
- The target audience has a compact car for commuting purposes
- Property inflation is not a predictable factor and will not be accounted for
- Any prices for the previous year (2020) are identical in this year so that the dataset will accurately represent a whole year of price fluctuation
- Annual averages will be used as seasonal price fluctuation is highly volatile

To complete the analysis of this data and provide the target audience with a more holistic assessment of living options, several unique formulas will be used to add depth to the gathered data:

- The Haversine formula, used to calculate the great-circle distance between two coordinates, will be used to approximate commuting costs using the job location coordinates and all towns/cities. Once calculated as a distance in miles, an approximate monthly commuting cost can be approximated based on average mpg and an average gas price
- A monthly mortgage payment calculation. Using the principal cost of the home and assuming a 3% down payment (first time home buyer option) and 3% interest rate for a 30 year-fixed mortgage, a monthly cost can be approximated and compared to rental costs in a less intimidating way than the total home price.

Buying is often not as expensive in some areas especially when compared to renting within large cities, this contextualized number will allow the target audience to make a more informed decision

2.2 Data Sources and Extraction

Mapsofworld.com (https://www.mapsofworld.com/usa/states/massachusetts/lat-long.html) provides a list of all cities and towns in with their corresponding latitudes and longitudes in a neat table. This data can be scraped into a data frame through the html processing contained in pandas.

Wikipedia (https://en.wikipedia.org/wiki/List_of_villages_in_Massachusetts) contains a table of towns and their counties in Massachusetts in a formatted table. Again, using pandas html read functionality, this data can be scraped into a data frame and the geolocation data joined with the county data through a city/town match case.

A summary site of monthly home sale data is scraped for median sales prices of homes and the corresponding city and an excel spreadsheet download from a popular rentals site is concerted to a csv for two bedroom monthly rental costs by town/city.

Geolocation data for the coordinates of the job location can be obtained through the provided geolocation package and all other points of consideration for the user will be achieved through rigorous data wrangling, calculation, and organized visualization.

Although machine learning might suggest what could be considered ideal through methods such as k-nearest neighbor, the weights of these variables will vary wildly between potential users within the target audience and therefore there is no truly optimized values to

identify. To avoid over-processing and allowing the user to apply weight as they see fit, machine learning will not be used to provide a recommendation.

2.3 Data Cleaning

From the sites noted above, four separate tables were created, one of all towns/cities and their global coordinates, a second table of each town/city and its parent county, a third of median single family home prices by city, and a final table of two-bedroom rental prices by city.

There are issues with the four sources inclusion of smaller towns in the reported data and therefore there is missing data for some smaller town in the state. Rather than trying to dig deeper for more granular supporting data, this is taken care of in the same step as combining the tables together. In order to make ensure that there is no ‘small-to-big’ data mismatch, the table joins are performed from the highest level of data down (County>City, City>Prices), data not included in any step in the matching location columns is automatically removed from the process, resulting in data that is not 100% complete in its coverage of pricing but has two benefits. One is that there is still good area data since the smaller towns that aren’t included are often lumped in with other nearby areas for pricing depending on the source so the area pricing is still captured, the second is that the data did not have to be thoroughly sifted through for mismatch and filling in missing data which can be time consuming. The resulting table consists of the features shown below and is the starting point for deeping the analysis for the target audience’s benefit

	Town/City	County	Median Home Price	Median 2BR Rent	Latitude	Longitude
0	Abington	Plymouth	$530,000	$1,365	42.12	-70.96
1	Adams	Berkshire	$191,700	$1,000	42.63	-73.12
2	Andover	Essex	$930,000	$1,187	42.66	-71.14
3	Arlington	Middlesex	$687,000	$1,740	42.42	-71.16
4	Athol	Worcester	$190,150	$922	42.59	-72.23

2.4 Feature Selection

Due to the specific data source selection, there are only several features that needed to be removed from the original four tables. All of them come from the property pricing data where each scarped table contained a problem for the price change from previous year which is not of critical importance in determining affordability. These rows were removed before joining the tables together

3. Methodology and Analysis

3.1 Calculation of Secondary Metrics

The generic pricing data ordered by city lacks two key features for young professionals, the cost of commuting from any given location, and the monthly cost of buying a home rather than renting, which is often much less of a monthly cost than the daunting milestone may seem. The assumptions for the commuting metric is that the target audience will use a car for commuting, that vehicle achieves 16.2mpg, and that the price of gas will hold at $3.00/gallon; for the monthly mortgage payment cost, it will be assumed that the audience is a first time home

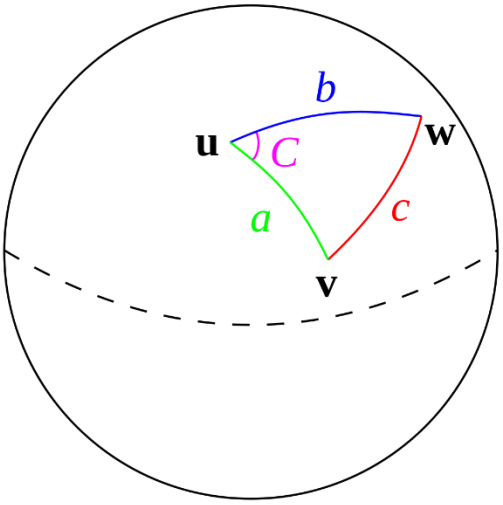
buyer with good credit, eligible for first-time home buyer down payment with a 3% interest 30-year-fixed home loan. Through these provided variables, calculation can begin.

For commuting, the coordinates of all towns and the target job location are already in the table, the issue is approximating the commuting distance and extrapolating that to a monthly cost to be factored in with the cost of accommodations. The concept of ‘great-circle distance’ is a critical one in any navigation and is a form of third dimension trigonometry, allowing the calculation of distance between two coordinates, “as the crow flies”, as if they were points of a large triangle laid on a sphere, more commonly known in mathematics as the haversine formula (which they are in a macro sense!).

$a = \sin^2(\varphi B - \varphi A/2) + \cos \varphi A * \cos \varphi B * \sin^2(\lambda B - \lambda A/2)$

$b = 2 * \text{atan2}(\sqrt{a}, \sqrt{(1-a)})$

$c = R \cdot b$ (where R is the radius of the earth)



Creating a function from this preexisting formula, this calculation in miles within our notebook we can then get the most direct distance between two global coordinate points. From here it is quite simple to perform the algebraic steps to calculate the commutes taken in a 30-day month and the resulting cost from gas prices and fuel efficiency. This can be added as a new column in our database and generated using the defined formula below.

```
def haversine_distance(lat1, lon1, lat2, lon2):
   r = 6371
   phi1 = np.radians(lat1)
   phi2 = np.radians(lat2)
   delta_phi = np.radians(lat2 - lat1)
   delta_lambda = np.radians(lon2 - lon1)
   a = np.sin(delta_phi / 2)**2 + np.cos(phi1) * np.cos(phi2) *   np.sin(delta_lambda / 2)**2
   res = r * (2 * np.arctan2(np.sqrt(a), np.sqrt(1 - a)))
   res_mi=res*0.621371
   return np.round(res, 2)
```

The other additional metric is to reduce the daunting cost of a family home down into its monthly value for an unbiased assessment of monthly comparative costs for the target audience, any bias can cost the audience financial consternation and regret, diminishing their potential satisfaction with their decision and this is required to eliminate the potential unconscious avoidance of large price tags. Using a common formula for mortgage payments, from the total price another column can be calculated with a integrated function as before using the formula and function shown below.

$$M = P [I (1 + I)^N] / [(1 + I)^N – 1]$$

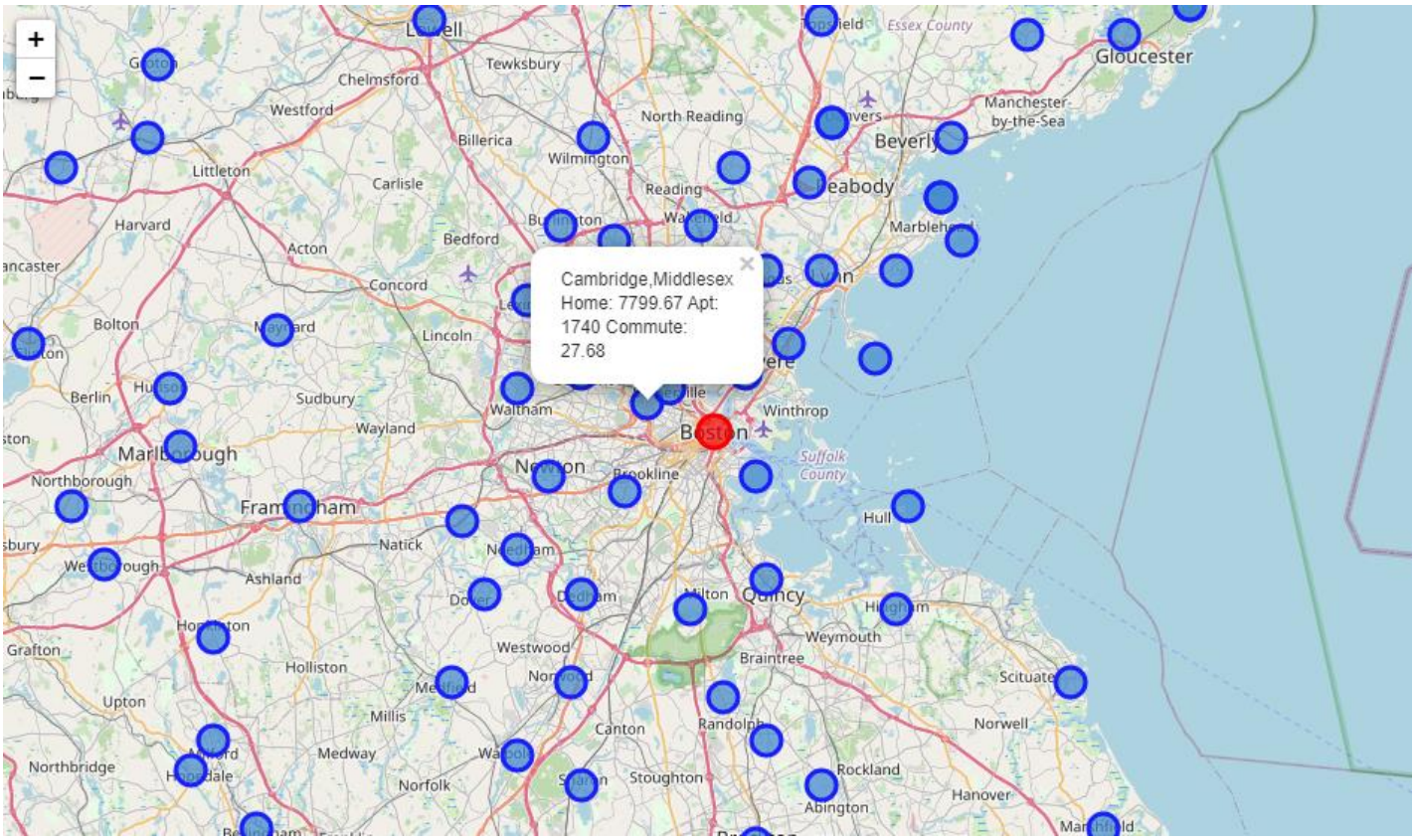
```
def mon_mortgage(P,r,n):
    M = P*(r*(1+r)**n)/((1+r)**n-1)
    return np.round(M,2)
```

The final data frame as shown below holds enough secondary metrics to give an overview of expected monthly cost of living and transportation in any given area.

	Town/City	County	Median Home Price	Median 2BR Rent	Latitude	Longitude	Monthly Commute Cost	Monthly Home Payment
0	Abington	Plymouth	530000	1365	42.12	-70.96	139.95	2234.5
1	Adams	Berkshire	191700	1000	42.63	-73.12	860.93	808.21
2	Andover	Essex	930000	1187	42.66	-71.14	170.46	3920.92
3	Arlington	Middlesex	687000	1740	42.42	-71.16	53.51	2896.42
4	Athol	Worcester	190150	922	42.59	-72.23	498.71	801.68
5	Attleboro	Bristol	430000	1014	41.93	-71.29	258.35	1812.9
6	Ayer	Middlesex	415000	1740	42.56	-71.58	241.81	1749.66
7	Barre	Worcester	225000	1192	42.42	-72.11	434.46	948.61

4. Results and Conclusion

Since the result of this project is providing a user with the means to make more informed choices the results and conclusion are the same in this case. The final product allows a user to browse a map of their new potential living location, clicking any bubble to glean information about the home and apartment costs in that area as well as monthly expected commuting costs. An image of the display is below to show the user interface and resulting information where the job location is in red and all cities/towns are in blue and can all be clicked to display information.



While I would have been happy to have this tool for my own relocation, it is apparent to me that there are not only lacking factors to paint an even more detailed picture of cost, but better data sources and data wrangling as well. Median prices are certainly not the best representative of data for a young professional looking in a metropolitan area. Future analyses should look at the total distribution of homes and selectively remove the right tail before calculating the mean rather than median, this will eliminate high cost homes driving up relative prices using the current method, giving less disparaging price differences in a more realistic fashion. Additionally, user input fields vie tkinter or html would make this tool even more powerful and user friendly if proper data sources could be found for the US (or globally!) for input job locations and desired travel distances.