# Milestone One

### Subject area

For the project required for DSC540 Data Preparation, I have decided to explore real estate data. At this stage in my life, my wife and I are ready to leave the state that we’ve grown up in and find a new place to call home.

### Data Sources

The data sources I have chosen for this project are as follows:

* Flat File:
  + This data source comes from Kaggle.com and consists of 58 columns. The file size is large compared to most files I’ve used for this Data Science program.
  + [Link to Kaggle.com](https://www.kaggle.com/datasets/vincentvaseghi/us-cities-housing-market-data)
* API:
  + The API I found that was free and seemed to fit the use case I wanted is the “Population Estimates API” from the US government census site. It includes data about population estimates as well as national monthly population estimates. The main variables I would like to utilize are “state”, “date\_code”, “age\_group”, and “pop”.
  + [Population Estimates API](https://www.census.gov/data/developers/data-sets/popest-popproj/popest.html)
* Website:
  + The website I’ve chosen is Weather Underground. Since this project will focus on a potential move, the weather will be a key factor. The website has a table of daily observations in weather for varying periods of time.
  + [Weather Underground Link](https://www.wunderground.com/history/monthly/us/nc/charlotte/KCLT/date/2023-4)

### Relationships

For the selection of data, I’ve tried to make an obvious connection between all three sources. It happens that there can be at least two different connections, State and/or City. Our intention in moving from Florida is to move to North Carolina, but perhaps more specifically Charlotte, NC. These three different data sets should all have this in common.

My approach for this term project will evolve over time as the insights from the data sets become clearer. I’ve picked three data sets that should work well together and be able to be joined based on location variables such as “state” or “city”. My overall plan is to clean and transform the data in a manner that will help me make a more informed decision about moving out of state.

There are several concerns that I have in regard to the data sets. One of these concerns is that perhaps the data sets will not work well together. I’ve reviewed the data sets but there is potential that I’ve missed something. I would hate to be midway through the project and have an issue with compatibility between data sets.

Another concern that I have is specific to a particular data set. The flat data set that I have chosen from Kaggle is the largest data set I have ever worked with. It is a .tsv file that is about 1GB in size when downloaded. There will be 59 columns to filter through with many more rows. The process of working with this file should not differ much from any other file but I hope to not be overwhelmed by the size and amount of data.

In regard to the ethical implications of the project topic, I do not believe there are any. I will be analyzing readily available data for my own research and analysis. If there is an ethical argument to be made, perhaps it could be that this research would be better suited for a realtor or realty group to perform, and by attempting to do this myself I am taking business from another group. However, this really will be a high level overview for my decision, and I would require the expertise of a realtor to purchase my next home.