Climate and US Migration

By:
Sanjana Chintalapudi
Sai Rohitha Challa
Minal Pawar

Table of Contents:

Abstract:

Climate change has emerged as one of the most formidable challenges facing humanity, with its impacts resonating across the world. Annual global temperature has increased by about 1.8°F (1.0°C) according to a linear trend from 1901 to 2016. 2023 was the hottest year on record. With the rise of global temperatures and increasing frequency of climate disasters, there is a profound consequence on shifting migration and settlement patterns. Individuals and communities are grappling with the short- and long-term ramifications of environmental changes and in turn rethinking where and how they live. We are going to dive deeper into understanding how changing weather patterns and natural disasters in the United States is impacting climate migration. We are going to look at both climate disasters that could force people to migrate and the slower shifting weather patterns that changing where people may want to live. By analysing the different dimensions of migration in the United States and comparing it to national climate patterns, we will shed light on the impacts of climate change on US migration. Our goal is to comprehend the subtle effects of climate change that go beyond the obvious by looking at how shifting weather patterns affect migration to states.

Introduction:

Climate change

Climate change has emerged as one of the most formidable challenges facing humanity, with its impacts resonating across the world. Annual global temperature has increased by about 1.8°F (1.0°C) according to a linear trend from 1901 to 2016. [1] 2023 was the hottest year on record. [2] According to National Academy of Sciences, the earth will see a greater increase in temperature in the next 50 years than compared to the whole last 6,000 years combined. Extremely hot zones like Sahara could cover fifth of the land surface by 2070 [3]. By the next century, around 3 to 6 billion people could be trapped in places facing extreme heat and food scarcity. South-Asia, which houses one-fourth of the global population, will be most affected in the near future given the current rates of global warming. In places like India and China, even a few hours outside would lead to death, with the greenhouse gas emissions unabated. [4] As the earth heats up, it has led to volatile weather patterns and more frequent climate disasters, impacting people everywhere significantly. From 2000 to 2019, there were 7,348 major recorded disaster events claiming 1.23 million lives, affecting 4.2 billion people, resulting in approximately US$2.97 trillion in global economic losses. [5] That is more than half the global population that has been affected. In 2022 alone, there were 33 million natural disaster-related displacements. [6]

Climate Migration

With the rise of global temperatures and increasing frequency of climate disasters, there is a profound consequence on shifting migration and settlement patterns. Individuals and communities are grappling with the short- and long-term ramifications of environmental changes and in turn rethinking where and how they live. This type of migration is called climate migration. Climate migration occurs when people leave their homes due to climate disasters, such as floods, droughts, and wildfires, as well as slower-moving climate challenges such as rising seas and increasing water scarcity. [9] People, regardless of immediate threat, are increasingly proactive in reevaluating their habitats. Across the United States, nearly 1 in 2 people will experience a decline in the quality of their environment, namely more heat and less water. [15] The term "billion-dollar disasters" has been coined to name the natural disasters that cause over a billion dollars in losses. These billion-dollar disasters cause about $60.5 billion in losses every year. [11] 14.5 million homes were impacted by natural disasters in 2021. That is about 1 in 10 homes in the US. [12] Droughts in the west are changing agricultural landscapes, forcing people to move due to their livelihoods.

Data Collection and Cleaning:

The data for migration within states was collected from the US Census website, and the weather related data was collected from the National Oceanic and Atmospheric Administration (NOAA) website. The collected migration data includes the estimated number of people who migrated within the states in United States (internal migration) during 2005 to 2022 with the exception of data of year 2020, and for weather data, parameters like average temperature (F), average snow (in), average precipitation (in), maximum temperature (F) and minimum temperature (F) were gathered for 372 stations spread over all the states of US, except for Delaware, on a monthly basis from 2005 to 2022. The list of stations were obtained from the Global Summary of the Month dataset of NOAA. API link

Weather data:
After requesting the data from NOAA, using an API, the data was cleaned, reduced and transformed into seasonal and yearly data. First, the data types of the obtained data are converted to help with further operations. All the stations didn't have the parameters requested, resulting in a lot of blanks. To handle the NaNs and blanks, statistics like average, maximum and minimum values for the data were calculated by grouping for states and years. For data exploration, the calculated monthly statistics were further calculated for summer and winter months, where April to October are considered summer months and the rest of the months considered as winter months. The data-frames were sorted for ease of use and to help with merging the dataframes.

Initial weather data obtained:

	DATE	STATION	SNOW	TMAX	TAVG	TMIN	PRCP
0	2005-01	USW00024029	NaN	35.5	23.2	11.0	0.19
1	2005-02	USW00024029	NaN	44.9	31.2	17.5	0.19
2	2005-03	USW00024029	NaN	52.4	38.9	25.5	0.46
3	2005-04	USW00024029	NaN	58.3	44.0	29.6	1.83
4	2005-05	USW00024029	NaN	63.7	51.0	38.2	6.18
...	...	...	...	...	...	...	...
80050	2022-08	USW00093806	0.0	89.2	80.8	72.3	4.53
80051	2022-09	USW00093806	0.0	86.5	75.2	63.8	2.71
80052	2022-10	USW00093806	0.0	76.8	62.2	47.5	2.86
80053	2022-11	USW00093806	0.0	67.4	55.3	43.2	5.02
80054	2022-12	USW00093806	0.0	59.5	50.4	41.2	4.95

	NAME
0	SHERIDAN AIRPORT, WY US
1	SHERIDAN AIRPORT, WY US
2	SHERIDAN AIRPORT, WY US
3	SHERIDAN AIRPORT, WY US
4	SHERIDAN AIRPORT, WY US
...	...
80050	TUSCALOOSA AIRPORT ASOS, AL US
80051	TUSCALOOSA AIRPORT ASOS, AL US
80052	TUSCALOOSA AIRPORT ASOS, AL US
80053	TUSCALOOSA AIRPORT ASOS, AL US
80054	TUSCALOOSA AIRPORT ASOS, AL US

Weather Dataframe after data-cleaning:

	DATE	State	avg_snow	max_temp	min_temp	avg_temp	avg_prcp
0	2005-01-01	WY US	10.90	43.4	7.8	26.34	0.46
1	2005-02-01	WY US	2.43	45.6	11.7	29.44	0.16
2	2005-03-01	WY US	11.13	52.6	22.5	35.96	0.60
3	2005-04-01	WY US	16.80	60.3	25.8	43.07	1.35
4	2005-05-01	WY US	0.20	66.8	35.0	51.31	2.93

Migration data:
For migration data, the preliminary data cleaning was done using MS Excel, as the table formats in the files changed over the years. The given data was a cross table for all the states that includes the in-state population, the population that moved from one state to another and the margin of error of census data. The datasets on the website were given per year, so combined 17 years of data files into one large dataframe. Since the data was given in segments, we transformed the dataset and calculated variables that can be leveraged in our analysis, Total_Pop, Percentage, Top1, Top1 State.

Total_Pop : We calculated the total population for each state per year, by adding the population currently living in that state and summing the populations that moved from other states to that current state.
Percentage: The percentage was calculated by dividing the people that moved to the state (Total Population - In-state population) by the total population.
Top1 : We found the top state that people moved from into the current state and the population moved by using the max function.
Top1_State : The top1 population variable identified the top state.

The excel sheets with these parameters data were read into separate data-frames. The data-frames were concatenated and cleaned. Unnecessary columns were dropped and the percentage column was transformed to show the values in percentages. After dropping the data for Delaware, District of Columbia, Puerto Rico (District of Columbia and Puerto Rico data was dropped as they weren't states), the weather and migration data-frames were merged on state and date columns.

Initial migration data obtained:

	Residence 1 year ago in --									
	Alabama		Alaska		Arizona		Arkansas		California	
Current residence in --	Estimate	MOE	Estimate	MOE	Estimate	MOE	Estimate	MOE	Estimate	MOE
Alabama	580,353	+/- 22,382	608	+/- 432	574	+/- 542	3,168	+/- 1,655	3,025	+/- 1,149
Alaska	152	+/- 176	84,150	+/- 6,528	883	+/- 530	54	+/- 83	3,898	+/- 1,708
Arizona	1,427	+/- 784	1,836	+/- 1,443	917,186	+/- 26,917	1,845	+/- 948	94,296	+/- 7,715
Arkansas	1,894	+/- 1,228	660	+/- 504	1,266	+/- 617	390,364	+/- 17,198	6,167	+/- 2,263
California	4,375	+/- 3,603	5,274	+/- 1,769	28,156	+/- 4,098	2,481	+/- 1,232	4,904,586	+/- 75,490

Migration Data-frame after data-cleaning:

	State	Total Pop	In-state Pop	Percentage	Top1	Top1 State	Year
0	Alabama	4716343	4607620	2.305239	21644	Georgia	2010
1	Alaska	697235	660909	5.210008	4123	Texas	2010
2	Arizona	6293718	6070993	3.538846	47164	California	2010
3	Arkansas	2879930	2800803	2.747532	13707	Texas	2010
4	California	36648257	36203508	1.213561	36582	Texas	2010

Both the Weather and the Migration data-frames didn't have any missing values, NaN values or blanks after the data cleaning. There were data-points with average snow as zero, this data is to be expected for some years and states.

We merged the weather and migration data frames so that each state and its corresponding year had both the migration data and the weather data. We did some additional cleanup by using the state's full name instead of its abbreviation and changing the date column to a year column.

	state	year	yearly_avg_snow	yearly_max_temp	yearly_min_temp	yearly_avg_temp	yearly_avg_prcp	yearly_min_prcp	yearly_max_prcp
0	Alaska	2005	6.237500	73.5	-22.7	35.382500	3.295833	1.65	6.27
1	Alaska	2006	7.075833	70.1	-29.7	32.372500	2.904167	0.99	5.89
2	Alaska	2007	6.545833	75.1	-25.8	33.729167	2.703333	0.80	5.92
3	Alaska	2008	8.650000	70.2	-23.0	31.442500	3.033333	2.09	5.57
4	Alaska	2009	8.765833	78.6	-24.4	33.072500	2.826667	1.32	4.66

The merged data frame has 24 columns representing the data across movement between states and the state's weather patterns.

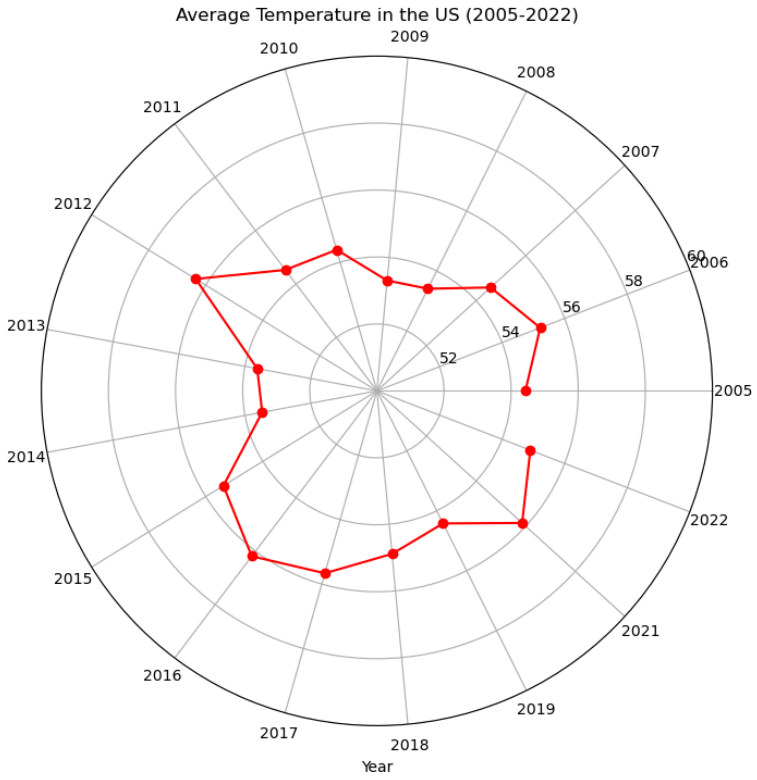
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 833 entries, 0 to 832
Data columns (total 24 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   state            833 non-null    object
 1   year             833 non-null    object
 2   yearly_avg_snow  817 non-null    float64
 3   yearly_max_temp  833 non-null    float64
 4   yearly_min_temp  833 non-null    float64
 5   yearly_avg_temp  833 non-null    float64
 6   yearly_avg_prcp  833 non-null    float64
 7   yearly_min_prcp  833 non-null    float64
 8   yearly_max_prcp  833 non-null    float64
 9   avg_snow_summer  815 non-null    float64
 10  avg_snow_winter  813 non-null    float64
 11  max_temp_summer  833 non-null    float64
 12  max_temp_winter  833 non-null    float64
 13  min_temp_summer  833 non-null    float64
 14  min_temp_winter  833 non-null    float64
 15  avg_temp_summer  833 non-null    float64
 16  avg_temp_winter  833 non-null    float64
 17  avg_prcp_summer  833 non-null    float64
 18  avg_prcp_winter  833 non-null    float64
 19  total pop        833 non-null    int64
 20  in-state pop     833 non-null    int64
 21  percentage       833 non-null    float64
 22  top1             833 non-null    int64
 23  top1 state       833 non-null    object
dtypes: float64(18), int64(3), object(3)
```

We analysed the distribution of the data for the merged data-frame using qq plots. From the qq-plots of the merged distribution, we could see that in the weather related columns, temperature appears to be relatively normally distributed, while the migration related columns are non-normally distributed.

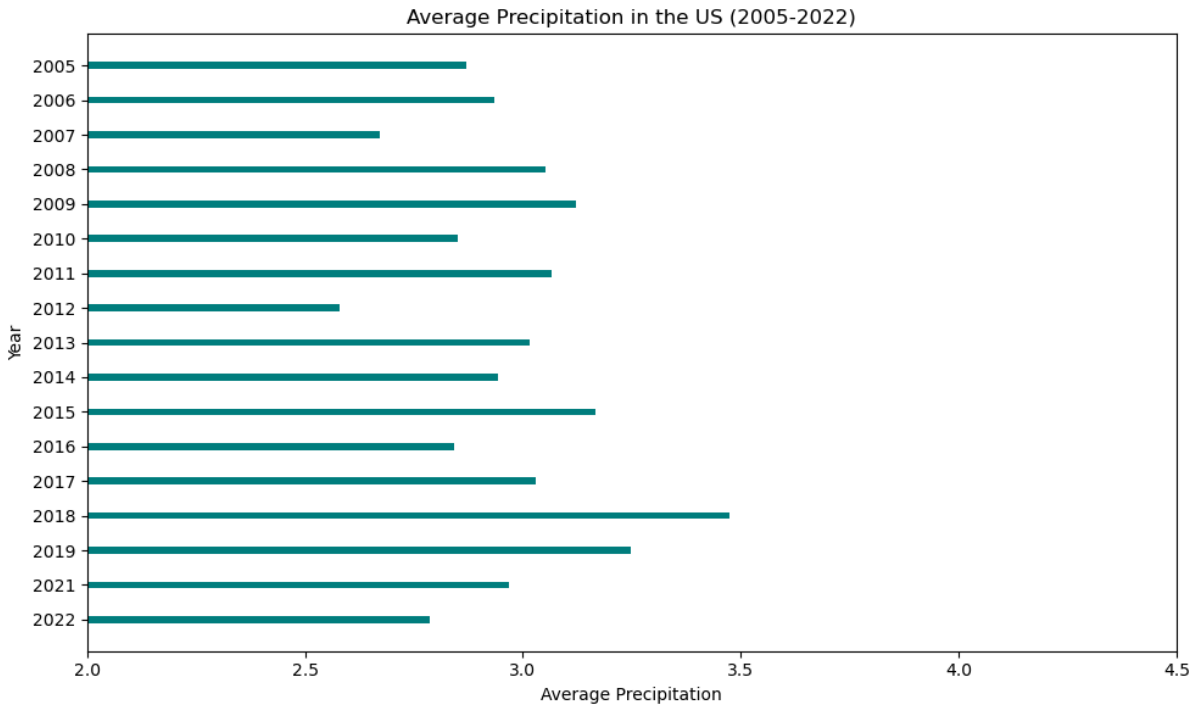
Visualisations:

To explore the weather data and migration data for the United States, we used visualisations so that we can uncover initial patterns. Link of website.

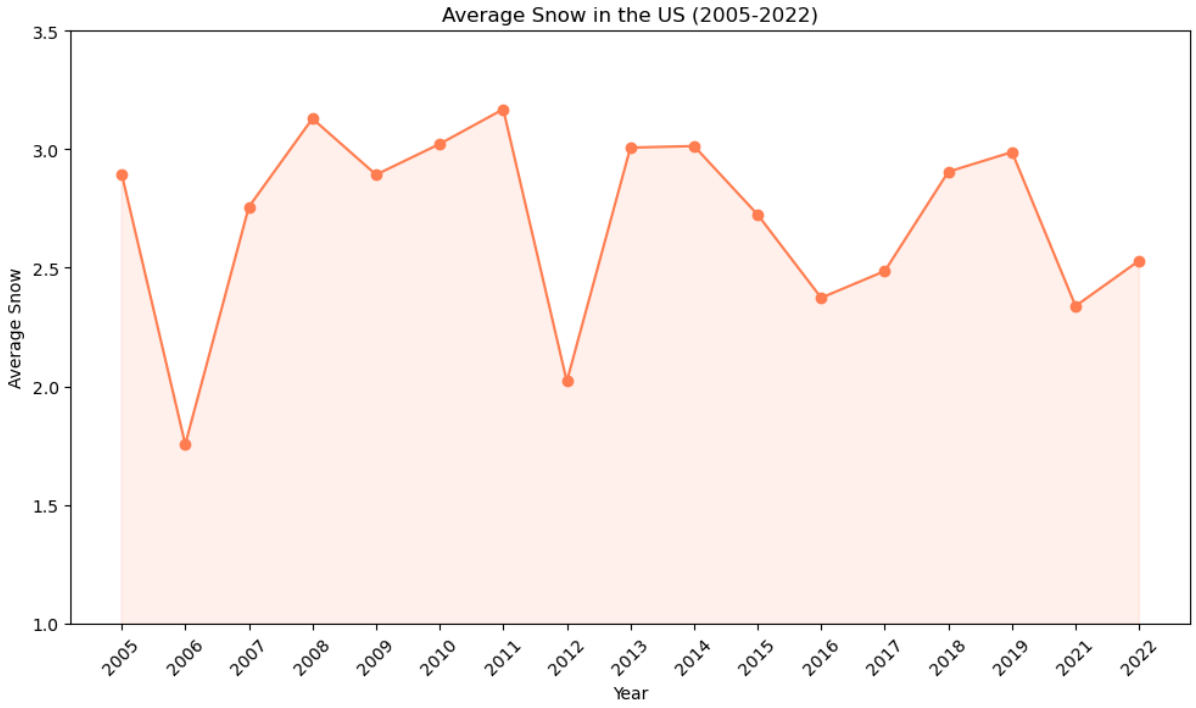
First we started by looking at the weather patterns across the years and the states.



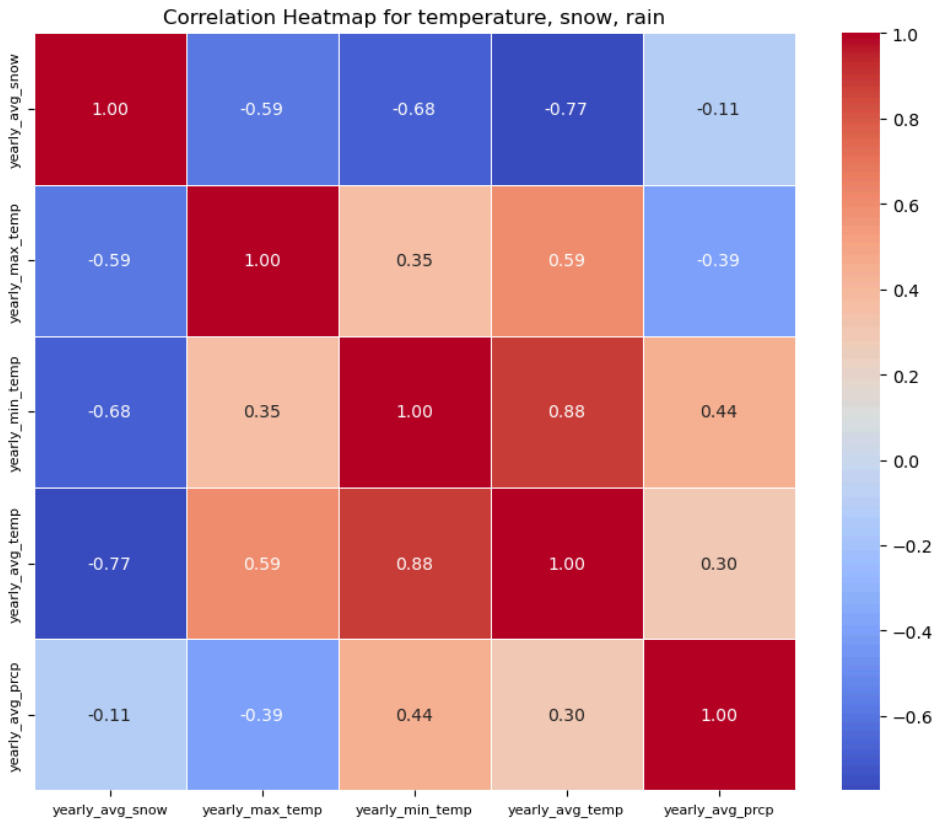
In the above plot we can see that the average temperatures in the United States are showing a cyclical behaviour with consecutive high and low dips in the temperature with the mean average temperature 55F and the range being 53F to 56.5F.



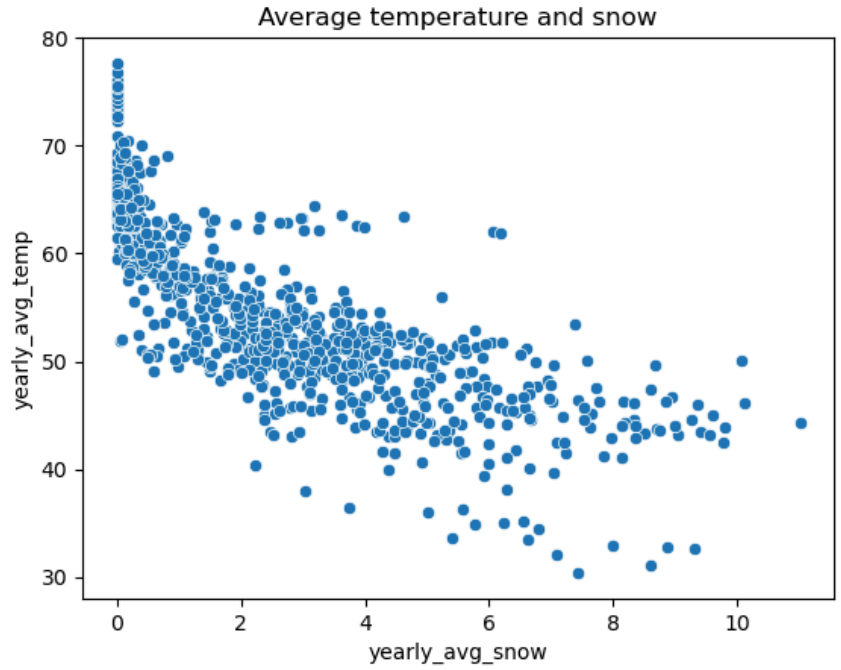
The above plot shows the average precipitation (in) in the United States over the years, with the mean precipitation being 3 inches. Concerningly, we can see that after 2018, average precipitation is declining.



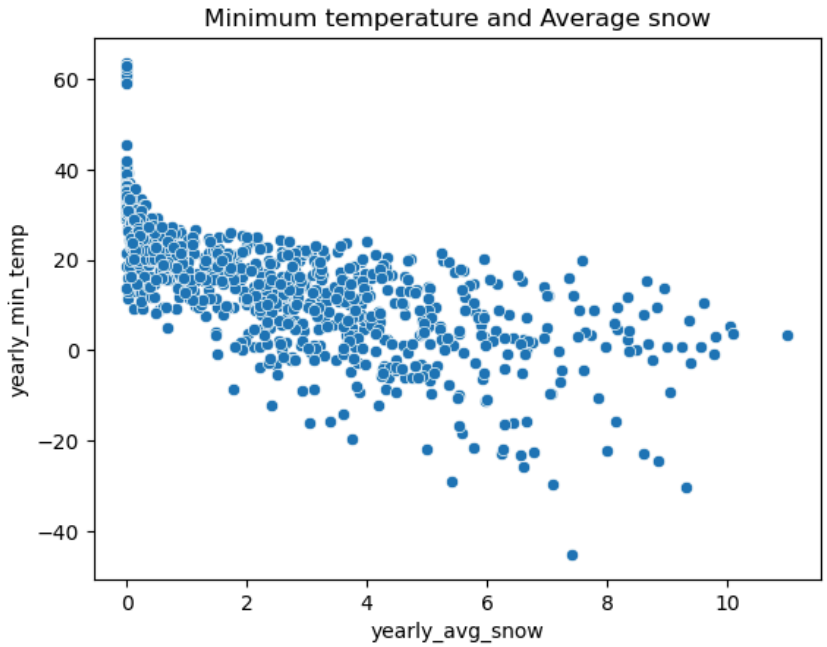
The above plot shows us the average snow(in) in the United States from 2005 to 2022. There is a sharp drop in the average snow in 2006 and 2012, after that the average snow ranges between 2 to 3 inches a year.



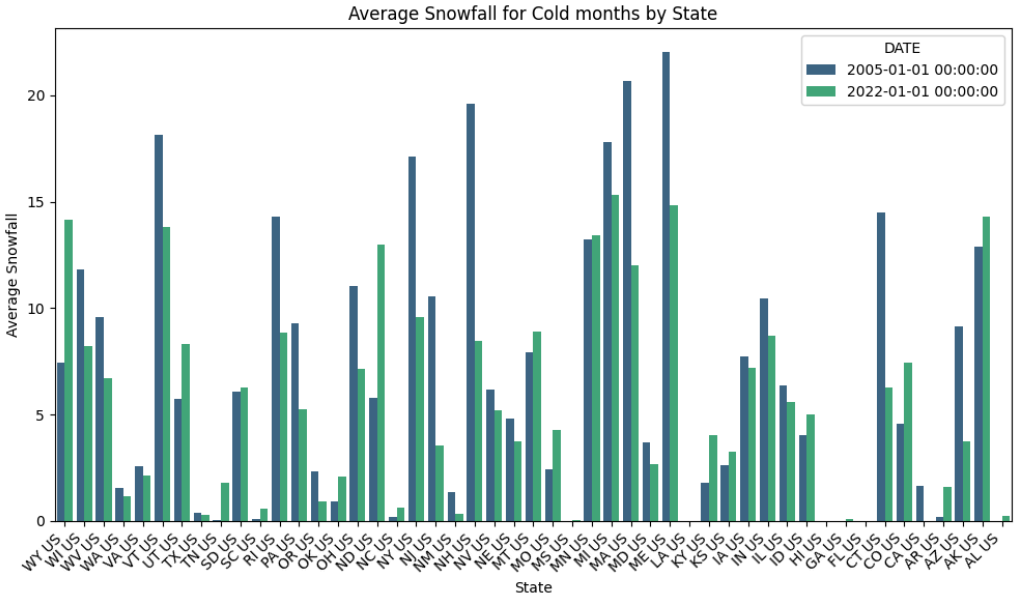
The above heatmap shows the correlation between the temperatures, snow and rain. We can see that the average snow has the highest correlation with average temperature and then with minimum temperature. Lets see how exactly the data is distributed with the help of scatter plots.



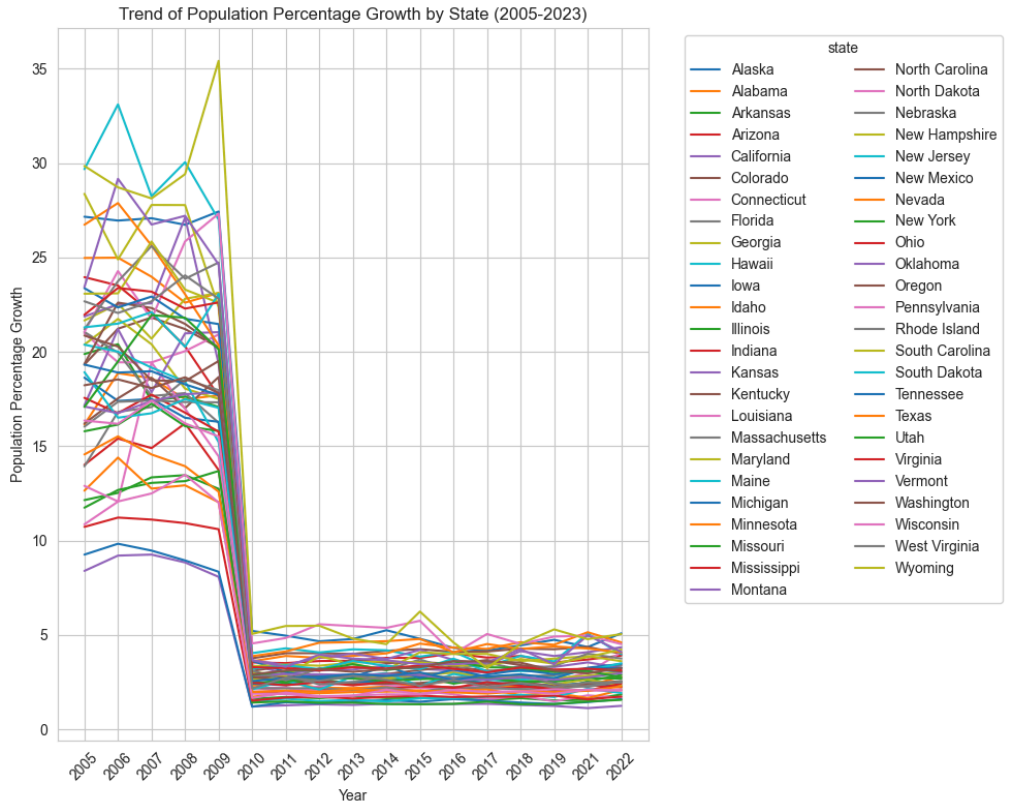
The above scatterplot shows that the yearly average snow and yearly average temperature tend to move together.



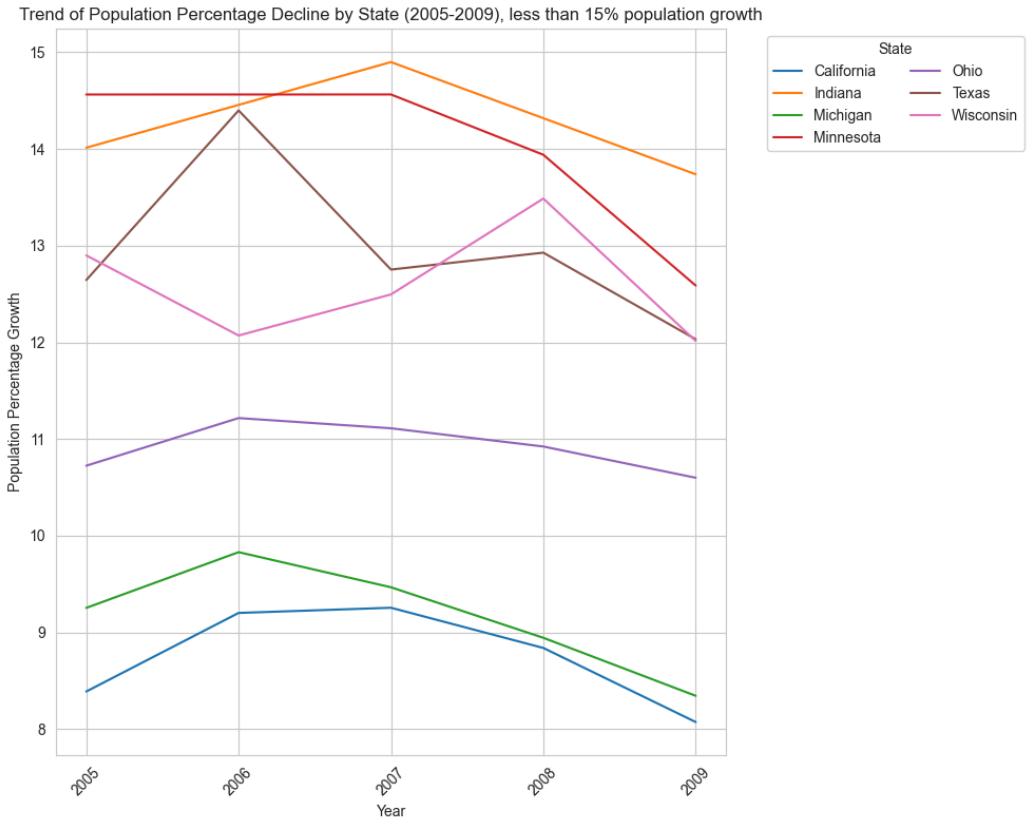
The above scatterplot shows that the yearly average snow and yearly minimum temperature tend to move together. We did not find any worthwhile trends when looking at temperature and precipitation.



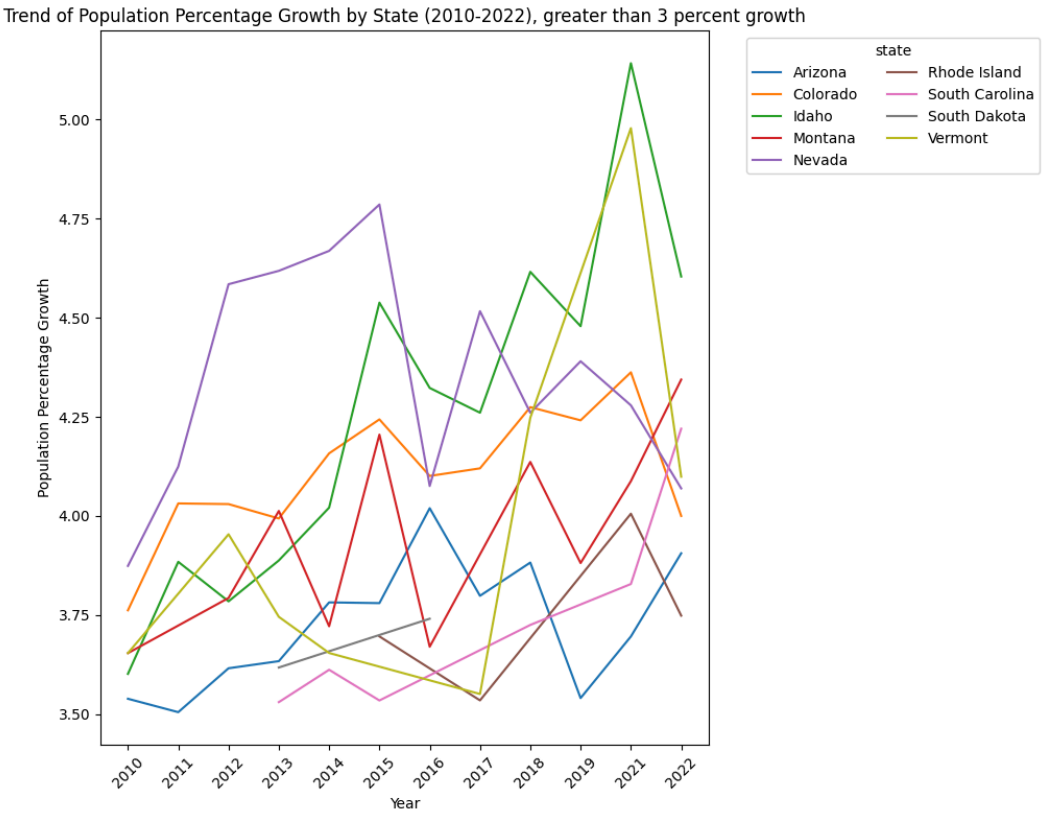
The above graph compares snowfall in the winter by state in 2005 and 2022. In most states, it appears that snowfall was more prevalent in 2005 compared to 2022. This aligns with the broader trends of climate change, including rising global temperatures resulting in lower snow patterns.



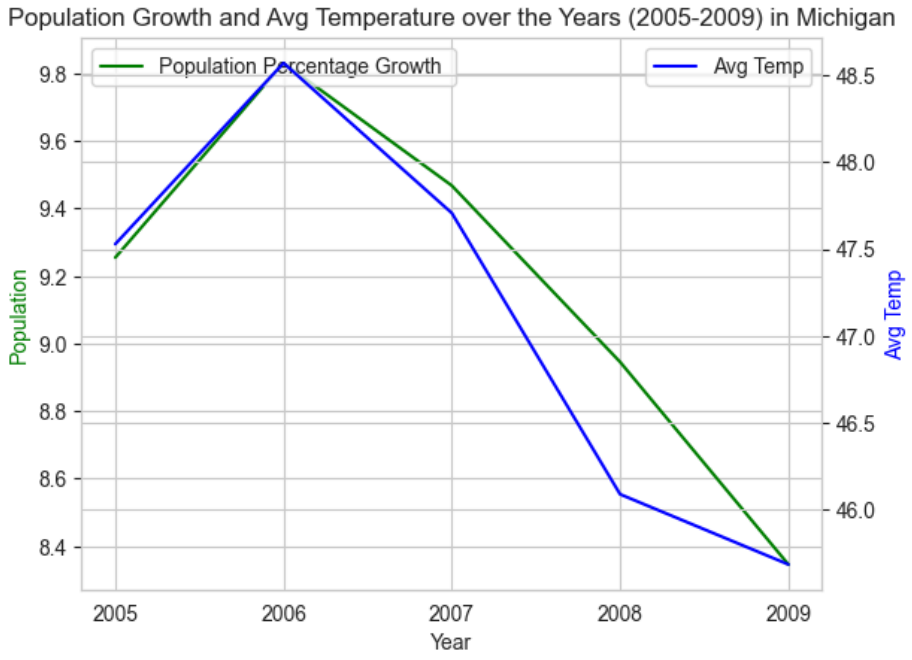
To better understand the migration patterns for each of the states, we plotted migration trends over the years for each state. The line plot shows a significant common trend across all the states; between 2009 and 2010, every state's population growth percentage dropped from over 5% to under 5%. One can speculate that the drop could be associated by the major event in the United States at that time, the Great Recession. To analyse the states' migration data better, we split the data to look at 2005-2009 together and the data between 2010-2022 together.



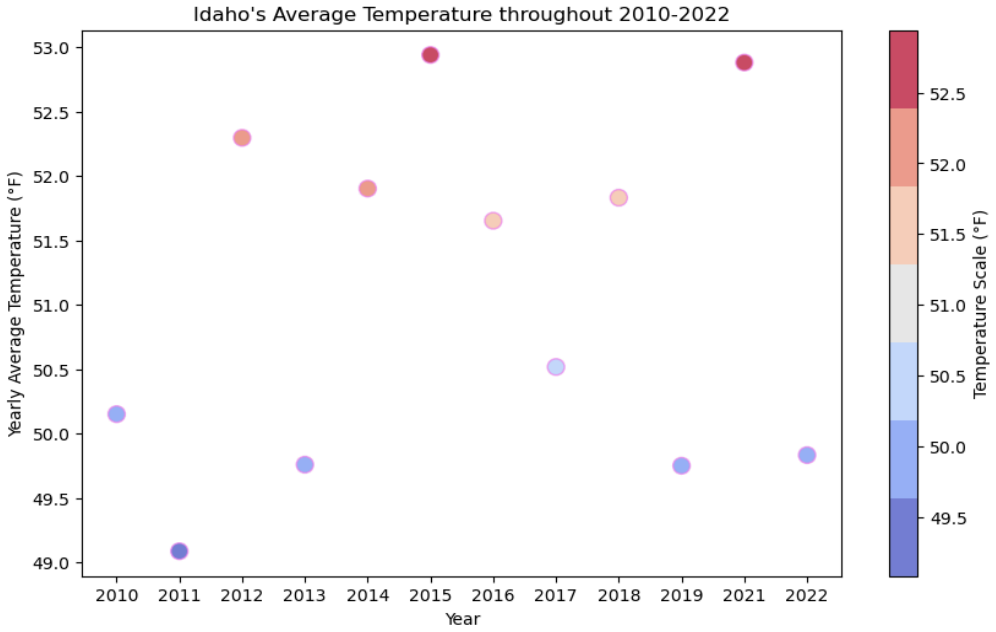
We first focused on exploring migration data from 2005 to 2009. We found that there were specific states that had decreasing population growth and significantly less population growth than most other states: less than 15%. California and Michigan had lower population growth than most states, less than 10% and declining.



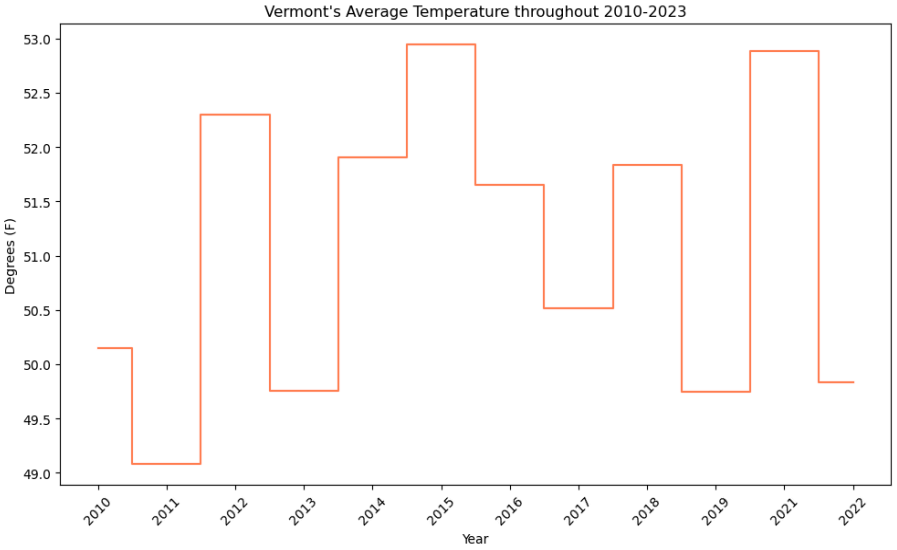
Between 2010 and 2022, most states did not have more than 10% population growth. There were specific states that were increasing greater than 3 percent and exponentially over most other states. In Vermont, there was a significant increase in population between 2017 and 2021. Keep in mind that there is no data for 2020, due to Covid so there is a gap for that year. Idaho had the most significant population growth percentage between 2010 and 2022 with over 5% in 2021.



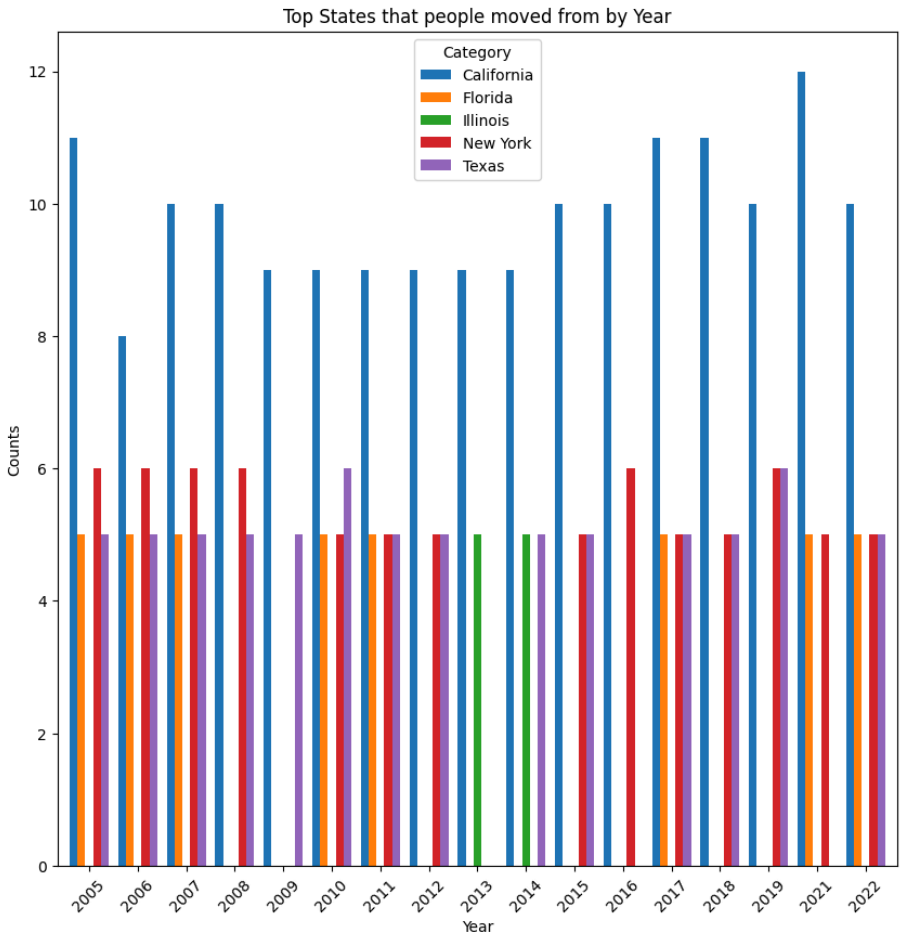
In Michigan, average temperature and population growth percentage had similar trend lines throughout 2005 and 2009. Between 2005 and 2006, population growth and average temperature increased. Between 2006 and 2009, the population percentage decreased and so did the average temperature. Both population growth and average temperature had similar slopes throughout the increase and the decrease.



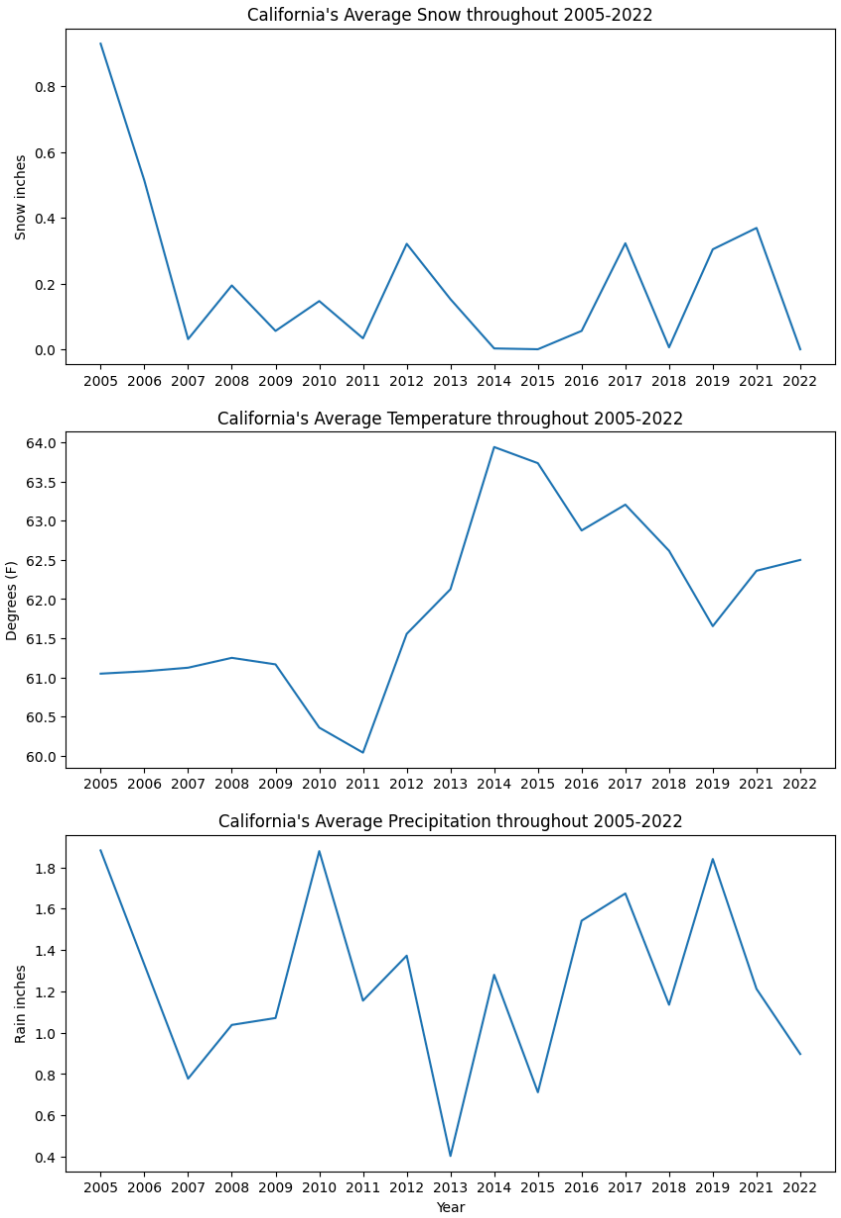
Idaho’s average temperature throughout 2010 to 2022 stayed consistent - around the 50s.



Vermont’s average temperature throughout 2010 to 2022 stayed consistent - around the high 40s, between 44 and 48 degrees Fahrenheit.



The above graph shows the top states that people move from to another state by year. The top states are California, Florida, Illinois, New York and Texas. California is the top state that people move from throughout 2005 and 2022, surpassing other states throughout those years.



California’s average temperature increased by 3 points between 2011 and 2015. California’s snowfall dropped by a quarter since 2005. California’s rainfall is also volatile with it fluctuating between less than 1in to no more than 2 inches throughout the last 17 years.

Models Implemented:

Data Cleaning: Our data is a combined dataset of weather and migration datasets. When we looked closely at the migration data, there was a sharp drop in the metrics calculated,which was caused due to the change in the data collection methods by the census department of the USA during 2009 and 2010. Therefore, the data before 2010 was removed from further inferences and analysis. We performed preliminary data analysis in R programming (the code and result snippets of R are at the end of the report) to understand the data and evidence of multicollinearity was found due to the variables that were derived from other variables, such as temperatures that correlated to winter and summer. These were originally leveraged to see if there were any prominent trends between winter and summer weather patterns that needed to be taken into account for migration. However during our analysis stage, we found that they weren't significant. Therefore, these low-significant features were dropped.

For one feature in the dataset, named yearly_average_snow, NAs were replaced with 0s. For modeling purposes, the state names were transformed into numbers using scikit-library functions. The features or X variables of the models are the weather data and the response or y variable of the model is the percentage of population migrated to a state in a given year.

Since the dataset comprises of time-series and numerical data, the models were chosen to accommodate for these factors, specifically that these factors would not affect the models' performances. Originally, we attempted to split the data by years for our train and test data, with train being from 2010 to 2018 and test data being from 2019 to 2022. However, when tested against different machine learning models, the models performed poorly. We found that a random split worked better for most of the models as indicated below. We did leverage splitting the data by years for train and test in the ARIMA model.

We implemented 5 models: Random Forest, Support Vector Machine (SVM), Gradient Boosting Machines (GBM), Linear Regression, and Autoregressive integrated moving average (ARIMA).

To see the implementation of the models check the link.

Data before Transformation for models:

```
state               object
year                 int64
yearly_avg_snow    float64
yearly_max_temp    float64
yearly_min_temp    float64
yearly_avg_temp    float64
yearly_avg_prcp    float64
yearly_min_prcp    float64
yearly_max_prcp    float64
avg_snow_summer    float64
avg_snow_winter    float64
max_temp_summer    float64
max_temp_winter    float64
min_temp_summer    float64
min_temp_winter    float64
avg_temp_summer    float64
avg_temp_winter    float64
avg_prcp_summer    float64
avg_prcp_winter    float64
total pop            int64
in-state pop         int64
percentage         float64
top1                 int64
top1 state          object
dtype: object
```

Dataset after transforming for all the models implementation:

```
year                int64
yearly_avg_snow    float64
yearly_max_temp    float64
yearly_min_temp    float64
yearly_avg_temp    float64
yearly_avg_prcp    float64
yearly_min_prcp    float64
yearly_max_prcp    float64
total pop            int64
in-state pop         int64
state_encoded        int64
dtype: object
```

```
Name: percentage, Length: 567, dtype: float64
```

X (features) Y (response)

Random Forest Regression:

The Random Forest Regression was used as a model since it builds multiple decision trees during training and gives the average of their predictions. Since our dataset has multiple weather features that don't exactly have a linear relationship with the percentage of migration, this model was suitable for predicting migration percentage. The X(feature) and y(response) were randomly split into 80% and 20%, named as train and test datasets. Hyper-parameters **max_depth** and **max_features** were used to create a more accurate model and different parameters were estimated. The hyper-parameter max_depth controls the depth of the tree, with a deeper tree risking overfitting and a shallow tree risking underfitting. Max_features defines the number of features to look at per decision tree in the random forest.

Mean Squared Error (MSE) and **Mean Absolute Errors (MAE)** were leveraged as metrics to test the accuracy and reliability of the model. The MSE indicates the average squared difference between the estimated values and the actual value and its purpose is to test the quality of the model. Without hyper parameters, the **MSE is 10.8%.** With hyper-parameters of max_depth = 20 and max_features = “log2”, the **Mean Squared Error (MSE)** was **10.6%**. Although the difference between having hyper-parameters and not is quite small, 0.002, it could be of some significance considering that the percentages of migration fall between 0-10%.

The **Mean Absolute Error (MAE)** of the model without hyper parameters is **0.234.** With hyper-parameters, the MAE is **0.232** indicating that the model is about 0.23 percentage points away from the actual values.

MSE and the MAE are relatively small indicating that the Random Forest model is accurate and reliable for predicting migration patterns based on weather patterns.

Support Vector Machine (SVM):

The support vector regression model is used to predict continuous outcomes by fitting as many data points as possible within a specified error margin while reducing the model complexity to prevent overfitting. Similar to Random Forest the X(feature) and y(response) were randomly split into 80% and 20% train and test data. Without hyper-parameters, the MSE for the SVM was 18% and the MAE is 0.32.

The hyper-parameters used for SVM were **C, gamma and epsilon**. The C parameter controls the complexity and accuracy of the model; gamma defines how much influence individual data points have on the prediction model; epsilon controls how sensitive the model is to the “noise” in the dataset. The hyper parameters chosen that allowed for a decrease in MSE and MAE were **C = 5, gamma = 0.1, and epsilon = 0.2.**

With hyper-parameters, the MSE is 16% and the MAE is 0.30. **The MSE difference of 0.03 amounts to a 11.11% error reduction** which is a significant reduction for migration patterns as the migration percentages are small. With hyper-parameters the model was more accurate and reliable, albeit not as accurate as random forest.

Gradient Boosting Machines (GBM):

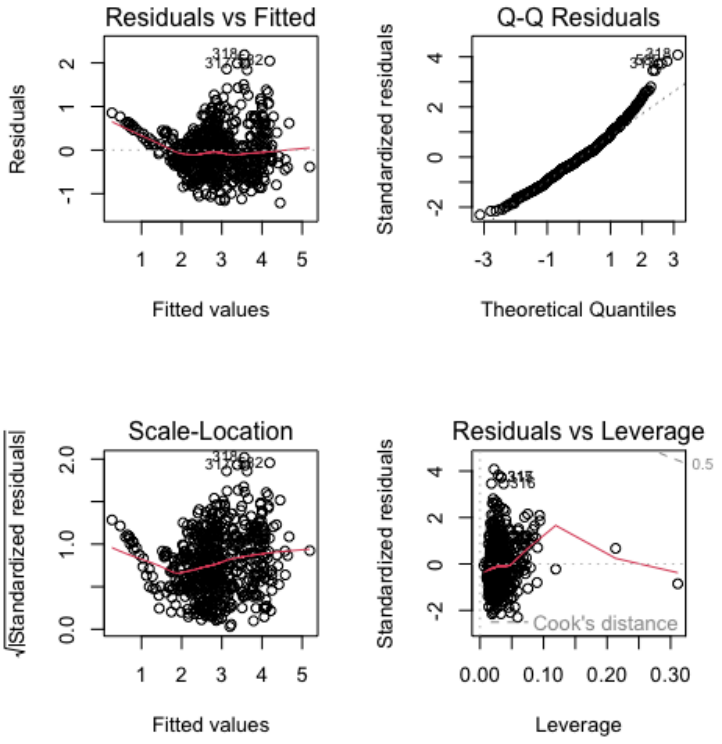
In Gradient Boosting Machine (GBM) model, the "gradient" helps the model get better by fixing its errors step by step. Boosting means each new iteration tries to learn from the previous ones to perform better. GBM builds trees, with each new tree fixing the mistakes of the ones before it.

Our dataset was split into two parts: one for training the model and the other for testing it. We set **80%** of our data for **training** (X_train, y_train) and kept **20% for testing** (X_test, y_test).

Mean Squared Error (MSE) was used to measure how close the model's predictions were to the actual values. The MSE of our model was approximately **14.94%**. This number indicates that on average, the GBM model's predictions are about 0.1494 away from the real values. This was not the best MSE of the models that have been implemented so far. To make the model better, we can adjust its settings, understand which factors are most important, or test it with different data to see if it works well in general.

Multiple Linear Regression Model (MLR):

The Multiple linear regression model is one of the standard statistical models to evaluate the relation between the feature and response variables. Its assumptions make it easier to interpret and infer about the data. The regression models assume **homoscedasticity, normality and linearity** from the data for the model to be unbiased and as efficient as possible. When checked for the assumptions using **residual vs fitted plot** and **Q-Q plot** below, we find that the assumptions are satisfied but not in its entirety. The X(feature) and y(response) were randomly split into 80% and 20%, named as train and test datasets. The model is trained on the train dataset and tested for accuracy and errors on test data. After fitting the linear regression model, the MSE (Mean-squared-error) metric was used to check the models' performance.



Below we find the fitted MLR model summary, and we can see that the p-values for many variables are very low. When the p-values are lower than 0.05 (default significance level), that means that those feature variables are significant in the model and that those features contribute to explaining the patterns in the response variable.

```
Call:
lm(formula = percentage ~ ., data = weather)

Residuals:
     Min       1Q   Median       3Q      Max
-1.38625 -0.42438 -0.06277  0.28979  2.35256

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -2.477e+01  1.415e+01  -1.751  0.08044 .
year             1.531e-02  7.003e-03   2.186  0.02922 *
yearly_avg_snow -9.446e-03  1.850e-02  -0.511  0.60987
yearly_max_temp -1.429e-02  7.417e-03  -1.927  0.05446 .
yearly_min_temp  1.023e-02  4.583e-03   2.232  0.02599 *
yearly_avg_temp -1.493e-02  1.046e-02  -1.427  0.15417
yearly_avg_prcp -4.921e-01  5.511e-02  -8.929  < 2e-16 ***
yearly_min_prcp  7.291e-02  5.859e-02   1.245  0.21380
yearly_max_prcp  5.971e-02  1.936e-02   3.084  0.00214 **
total.pop        6.275e-06  4.887e-07  12.840  < 2e-16 ***
in.state.pop    -6.441e-06  4.959e-07 -12.989  < 2e-16 ***
state_encoded   -2.400e-03  1.919e-03  -1.251  0.21156
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6202 on 576 degrees of freedom
Multiple R-squared:  0.5877,    Adjusted R-squared:  0.5798
F-statistic: 74.63 on 11 and 576 DF,  p-value: < 2.2e-16
```

The **MSE** turned out to be more than **30%** on any run, which indicates the model is performing poorly. There are a number of factors for the result we obtained, like the small size of the dataset, the data being time-series data while the MLR assumes the data-points to be independent, the features not capturing the total changes in the response variable and so on. We next perform a model that is used mainly for time-series data.

Autoregressive Integrated Moving Average Model (ARIMA):

The ARIMA model is often used to forecast time-series data. The model takes the response-variable as the input, accounts for the seasonality of the variable, identifies the trend of the data and forecasts the next value for the future time-point. Given the methodology, we fitted an ARIMA model for each of the 50 states in the United States and calculated the MSE for each state separately. The response variable for each state consists of 11 data-points which were divided into train data(2010-2018) and test data (2019-2022) of 80% and 20% split.

The **mean MSE** for all the states is **13%** with standard deviation of MSE being 22%**,** the median of the MSEs calculated for all the states is 4% and the model for Virginia state leads to a MSE of 0.4%, the best performing model out of all states. The median gives us a better idea of the model statistics, as it is robust to outliers and since there is huge difference between the mean and median we can say that there is a presence of outliers, meaning some of the ARIMA models performed very poorly.

As we know lower MSE means that the predicted values are close to the actual test data values and that the model is performing good, looking at the obtained MSEs above, we can say that ARIMA models are one of the best performing models for time-series data, but with more data-points, this would lead to better results overall and not just for one or two states.

Research Questions and Objectives

We are going to dive deeper into understanding how changing weather patterns and natural disasters in the United States is impacting climate migration. We are going to look at both climate disasters that could force people to migrate and the slower shifting weather patterns that changing where people may want to live. By analyzing the difference dimensions of migration in the United States and comparing it to national climate patterns, we will shed light on the impacts of climate change on US migration. Our goal is to comprehend the subtle effects of climate change that go beyond the obvious by looking at how shifting weather patterns affect migration on states. In addition, we want to determine the socioeconomic and demographic variables that influence migration. Our research aims to evaluate the following questions:

- How do changing weather patterns influence migration patterns in the United States?
- What are the socioeconomic factors driving climate-induced migration?
- What are the variations in migratory patterns caused by climate change among distinct demographic categories?
- How are households migration affected by climate change?
- How are people migrating based on age demographics?
- What states are going to be more populated in the next 5 years?
- What regions are going to be more affected by weather patterns and natural disasters and what is the population there going to be?
- How are regions with seasonal changes affecting migration patterns versus regions that are mostly consistent throughout the year?
- What type of natural disasters cause people to migrate?
- How has changes in natural disasters in an area impacted migration?

Data sources:

[1]https://www.census.gov/data/tables/time-series/demo/geographic-mobility/state-to-state-migration.html
[2] https://www.ncei.noaa.gov/

R-code:

```
weather <- read.csv('data/merged_dataframe.csv') #reading file
weather <- weather[!weather$year %in% c(2005, 2006, 2007, 2008, 2009), ] #dropping those year rows
rownames(weather) <- NULL #reindexing
weather$state <- factor(weather$state) #converting the state to numerical
weather$state <- as.numeric(weather$state)
colSums(is.na(weather)) #checking nas
weather$yearly_avg_snow[is.na(weather$yearly_avg_snow)] <- 0 #replacing nas with 0s
weather <- weather[, !(names(weather) %in% c("top1.state"))] #dropping top1 state as it is a string variable
weatherlm <- lm(percentage ~ ., data=weather)
summary(weatherlm)
```

Summary

```
Call:
lm(formula = percentage ~ ., data = weather)

Residuals:
     Min       1Q   Median       3Q      Max
-1.21039 -0.34909 -0.03536  0.29193  2.18758

Coefficients: (5 not defined because of singularities)
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -2.550e+01  1.315e+01  -1.939 0.052979 .
state            1.463e-03  1.768e-03   0.828 0.408289
year             1.601e-02  6.523e-03   2.454 0.014431 *
yearly_avg_snow -3.954e-02  1.816e-02  -2.177 0.029904 *
yearly_max_temp  8.350e-03  8.638e-03   0.967 0.334125
yearly_min_temp -2.519e-04  5.865e-03  -0.043 0.965757
yearly_avg_temp  7.251e-02  3.192e-02   2.271 0.023521 *
yearly_avg_prcp -2.953e-01  6.875e-02  -4.295 2.07e-05 ***
yearly_min_prcp  1.145e-01  5.376e-02   2.129 0.033707 *
yearly_max_prcp  6.781e-02  1.764e-02   3.845 0.000135 ***
avg_snow_summer  1.733e-01  3.890e-02   4.455 1.02e-05 ***
avg_snow_winter         NA         NA      NA       NA
max_temp_summer         NA         NA      NA       NA
max_temp_winter  7.535e-03  8.181e-03   0.921 0.357455
min_temp_summer -4.253e-03  6.676e-03  -0.637 0.524380
min_temp_winter         NA         NA      NA       NA
avg_temp_summer -1.230e-01  2.494e-02  -4.931 1.09e-06 ***
avg_temp_winter         NA         NA      NA       NA
avg_prcp_summer -6.642e-02  5.155e-02  -1.289 0.198079
avg_prcp_winter         NA         NA      NA       NA
total.pop        7.525e-06  6.180e-07  12.176  < 2e-16 ***
in.state.pop    -7.705e-06  6.248e-07 -12.332  < 2e-16 ***
top1            -7.963e-06  2.394e-06  -3.327 0.000937 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5423 on 549 degrees of freedom
  (21 observations deleted due to missingness)
Multiple R-squared:  0.6834,    Adjusted R-squared:  0.6736
F-statistic:  69.7 on 17 and 549 DF,  p-value: < 2.2e-16
```

```
options(repr.plot.width = 9, repr.plot.height = 9)
par(mfrow = c(2,2))
plot(weatherlm) #plots of the fitted regression model
#(plot in the model section)
```

References:

[1]https://nca2018.globalchange.gov/chapter/2/
[2
https://www.climate.gov/news-features/understanding-climate/climate-change-global-temperature#:~:text=2023%20was%20the%20warmest%20year,1850%20by%20a%20wide%20margin.
[3]
https://features.propublica.org/climate-migration/model-how-climate-refugees-move-across-continents/
[4]https://www.propublica.org/article/climate-crisis-niche-migration-environment-population
[5]https://www.undrr.org/publication/human-cost-disasters-overview-last-20-years-2000-2019
[6]
https://www.migrationpolicy.org/article/climate-migration-101-explainer#origins
[7]https://openknowledge.worldbank.org/entities/publication/2be91c76-d023-5809-9c94-d41b71c25635
[8] https://news.un.org/en/story/2021/09/1098662
[9]
https://www.cfr.org/in-brief/climate-change-fueling-migration-do-climate-migrants-have-legal-protections#:~:text=Climate%20migration%20occurs%20when%20people,seas%20and%20intensifying%20water%20stress
[10] https://link.springer.com/article/10.1007/s11111-017-0290-2
[11]https://www.ncei.noaa.gov/access/billions/
[12]https://www.forbes.com/home-improvement/features/americans-moving-climate-change/
[13]
https://www.policygenius.com/homeowners-insurance/home-insurance-pricing-report-2023/
[14]
https://www.cnn.com/2023/11/07/homes/homeowners-insurance-climate-real-estate/index.html

[15]
https://www.propublica.org/article/climate-change-will-force-a-new-american-migration