FHB Home Work Report

# Summary

1. I used python3 to do data evaluation and correction. I found 1802 ERRORs (refer to Tables 1 for more details about errors) and then tried to use Open Street Map API to correct information of street, city, state, and zip code.
2. I did some basic statistical analysis and found that **March and September** are two peak-months for opening accounts over a year.
3. I used Open Street Maps to evaluate and correct address information; then I got the longitude and latitude data for each address to find out some Geospatial Relations. K-means algorithm was also applied to cluster the data and get some hotspots, which would give some clues for building new branches in the hotspots.
4. I used the data of number of accounts from the past N years to predict for the next year. After comparing 12 algorithms, I chose ARDRegression to build a model and the results indicated that there will be about 218 people opening an account in 2018.
5. In the future, we can initially use Google Map API or USPS API for address correction. Open Street Maps API also can do it but not powerful enough. Furthermore, after Geospatial Clustering, we can project the clustering results onto the map to better visualize our results. In addition, the size of the test data is too small, larger data size can improve the accuracy of Regression Model.

# Environments

* **Language**: Python3
* **IDE:** Jupyter
* **Models:** numpy, pandas, phonenumbers, matplotlib, scipy, sklearn
* **The name of scripts:** Homework.ipynb

# Data Pre-processing

I found totally 1802 ERRORs (refer to Tables 1 for more details about errors)

Table1. Error distribution by columns

|  |  |
| --- | --- |
| Column Name | Count |
| address | 81 |
| zip | 514 |
| phone | 270 |
| work address | 79 |
| work zipcode | 569 |
| work phone | 247 |
| work email | 4 |
| account created on | 56 |
| Total | 1820 |

**Issue justification:**

* I used “phonenumbers” package to evaluate phone number information transfer them to International format.
* I used python’s datetime tool to transfer all the date to the format of “%m/%d/%Y %H:%M (11/26/12 16:47)”.
* I used Regex to evaluate Email info and marked all wrong email addresses.
* I used Open Street Map API to correct the wrong info of the street, city, state, and zip code. For example:

When I run following code:

address\_validation('3029 lowrey ave', 'Bentonville','Hawaii','97655'))

#wrong city and wrong zip code

I will get :

('3029 lowrey ave', 'UPDATE(F(Bentonville),T(Honolulu))', 'Hawaii', 'UPDATE(F(97655),T(96822))', '21.308664251275', '-157.807035707976')

This function (address\_validation) successfully corrects the city and zip code.

But when I run the code for all data sets, I got an error as follow:

Bandwidth limit exceeded

You have been temporarily blocked because you have been overusing OSM's geocoding service or because you have not provided sufficient identification of your application. This block will be automatically lifted after a while. Please take the time and adapt your scripts to reduce the number of requests and make sure that you send a valid UserAgent or Referer.

For more information, consult the [usage policy](https://operations.osmfoundation.org/policies/nominatim/) for the OSM Nominatim server.

This was caused by we call the API too frequently.

# Basic Data Analysis

I did some basic statistical analysis and found that **March and September** are two peak-months for opening accounts over a year (Fig. 1). I also clustered the data by State (Fig. 2) and by Year (Fig. 3 and Fig.4, normalized to percentages) and found that the data are evenly distributed.

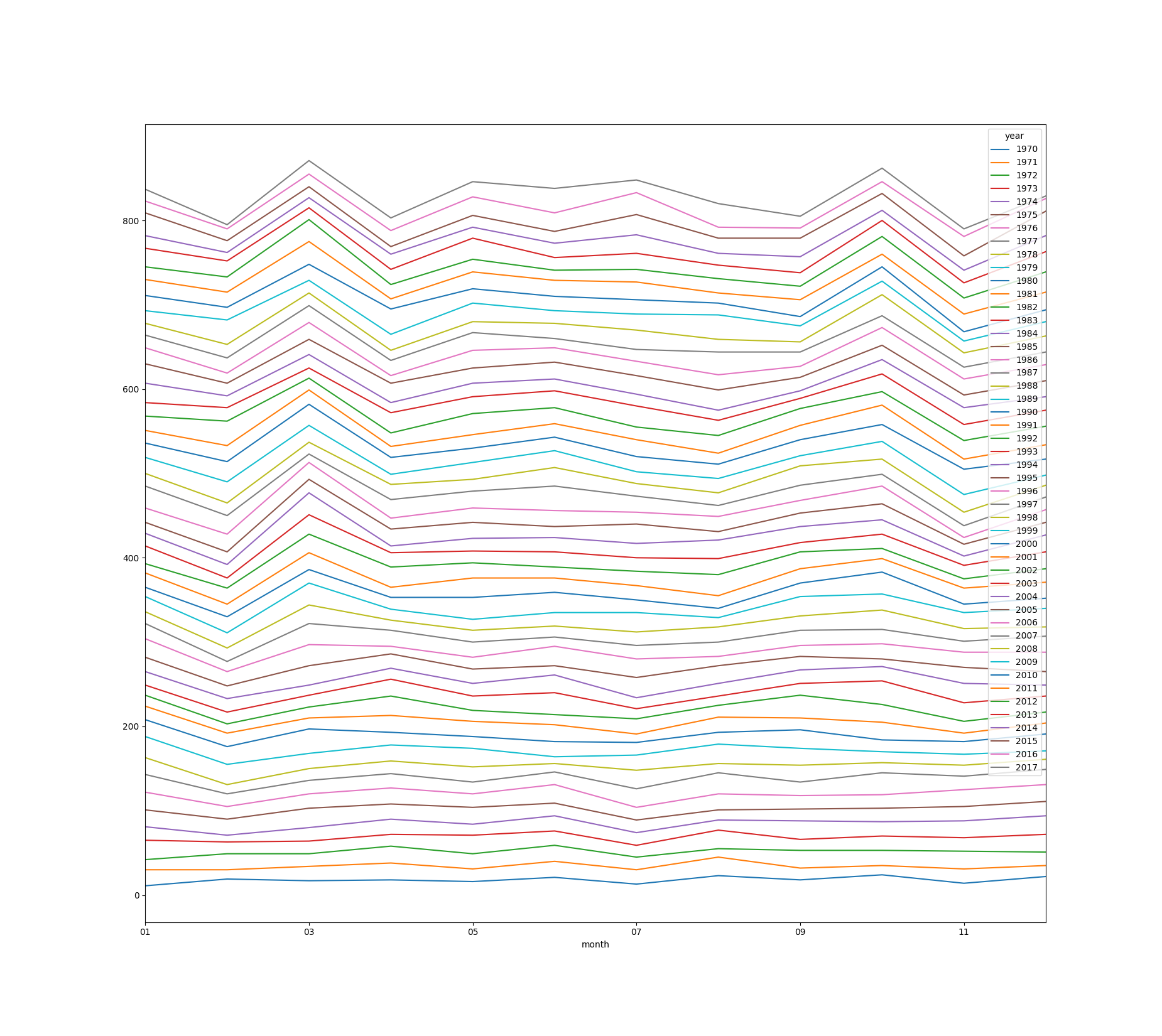


Fig. 1. Data grouping by Month and Year

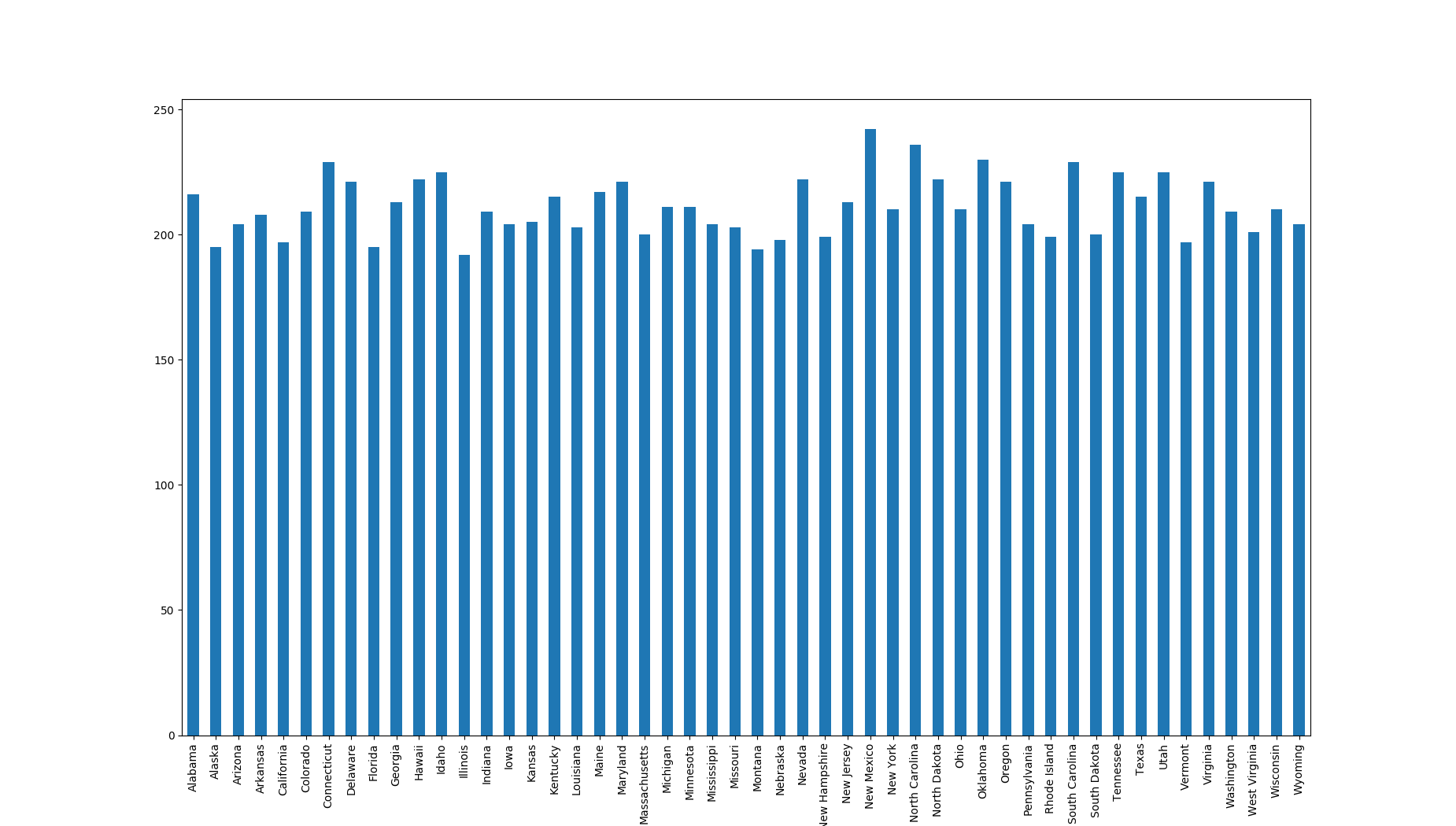


Fig. 2. Data grouping by State

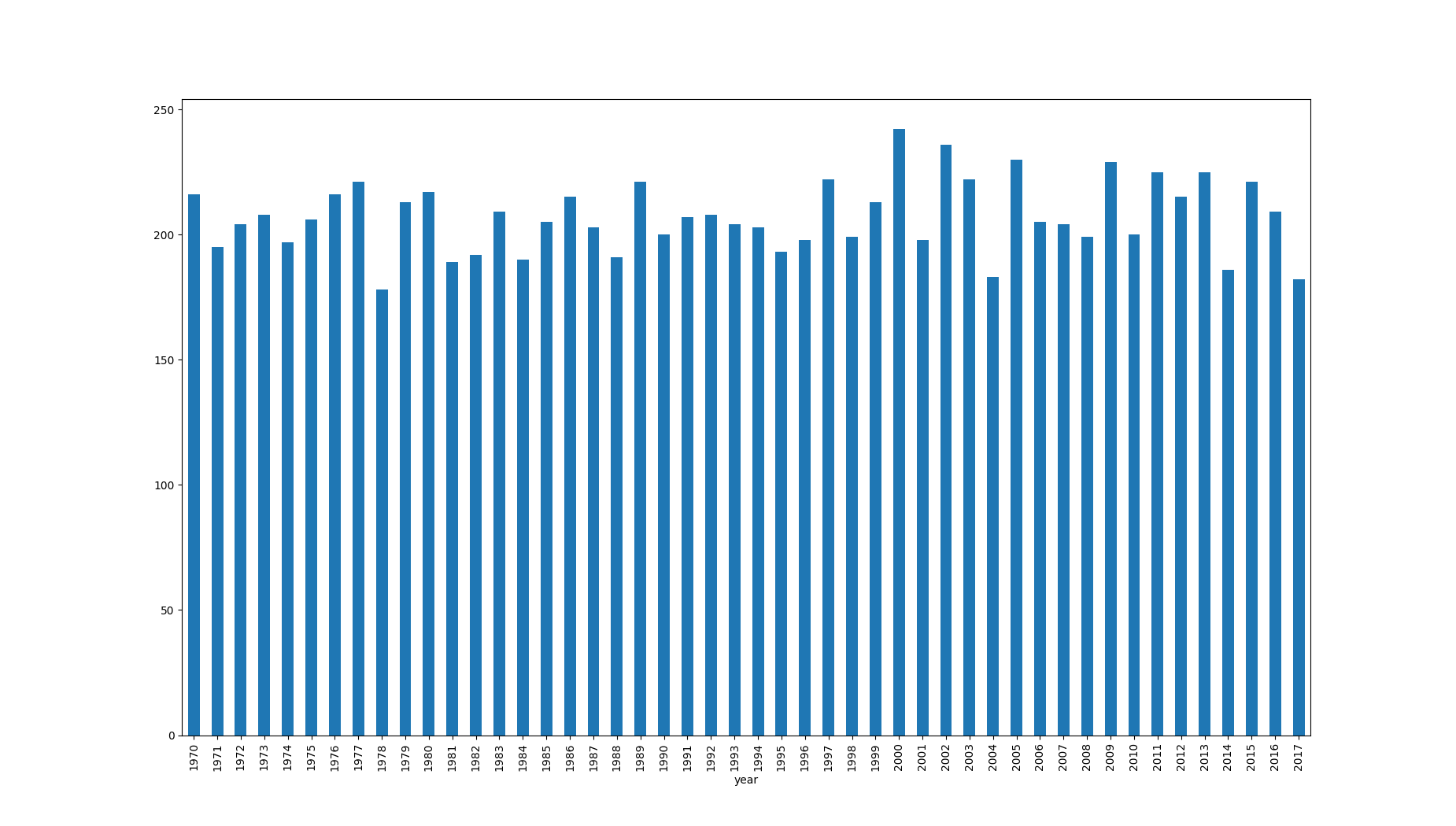


Fig. 3. Data grouping by Year

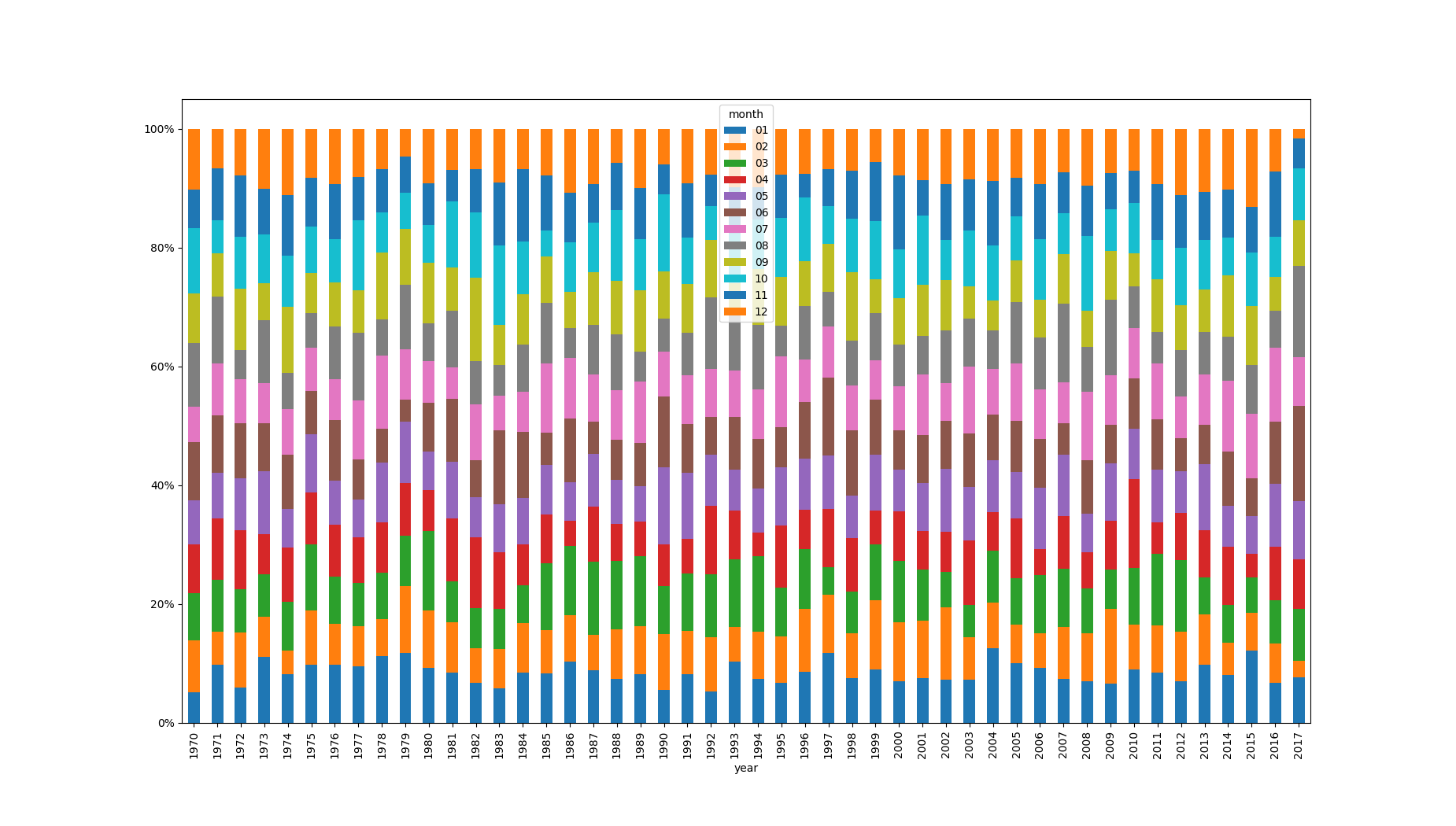


Fig. 4. Data grouping by Year and Month (Normalized to percentages)

# Geospatial Clustering

I used Open Street Maps to evaluate and correct address information; then I got the longitude and latitude data for each address to find out some Geospatial Relations. K-means algorithm was also applied to cluster 100 records and eventually I got 3 clusters that are geospatially close each other (Fig. 5).

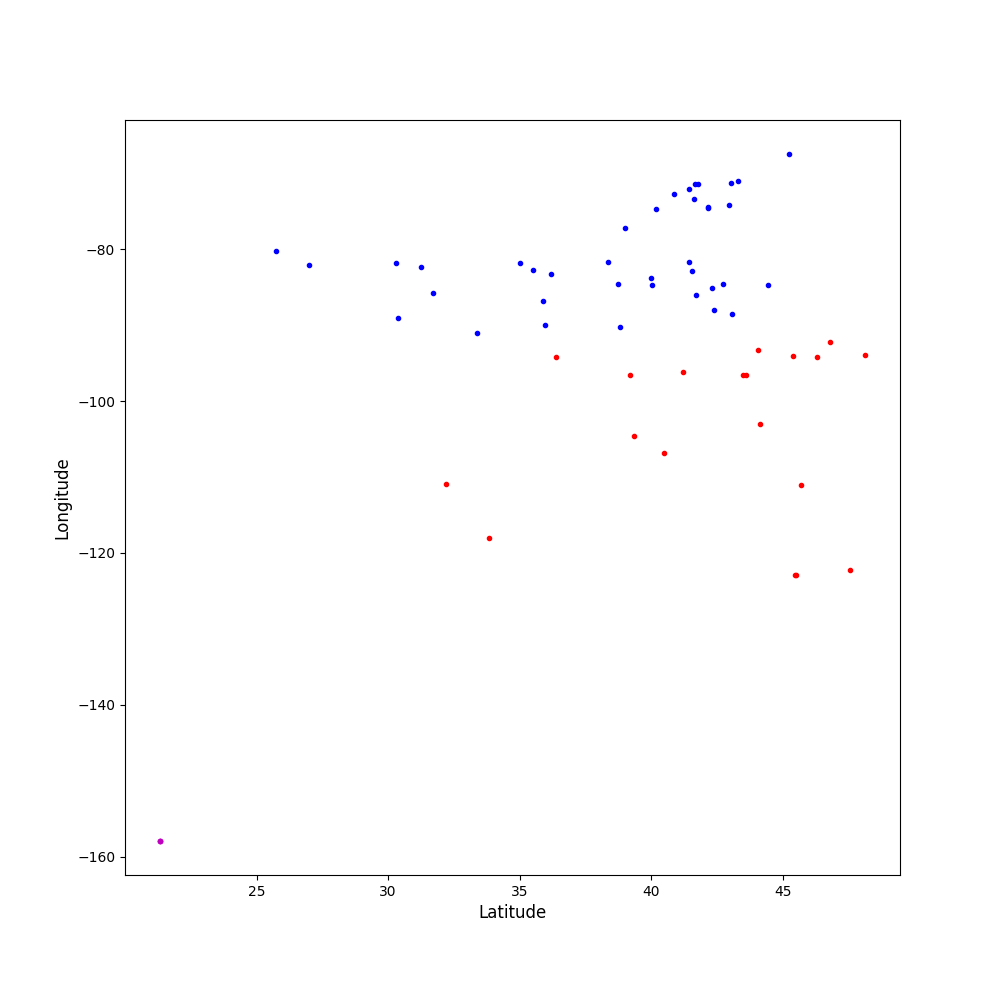


Fig. 5. K-means clustering result

# Account Prediction

I used the data of number of accounts from the past N years to predict for the next year. After comparing 12 algorithms (Fig. 6 and Fig. 7), I chose **ARDRegression** to build a model and the results indicated that there will be about 218 people opening an account in 2018 (Fig. 8).

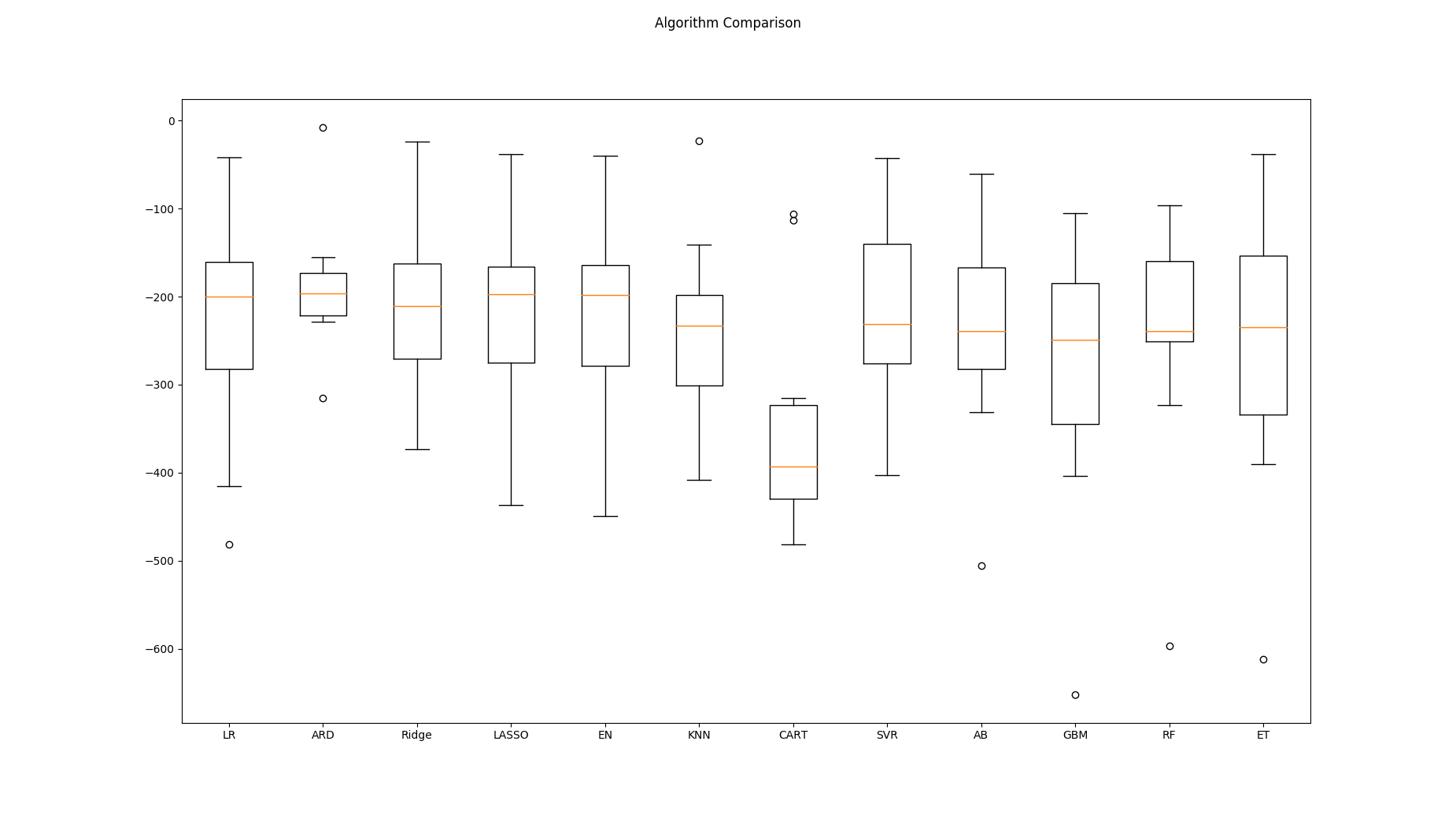


Fig. 6. Mean Squared Error of 12 Algorithms

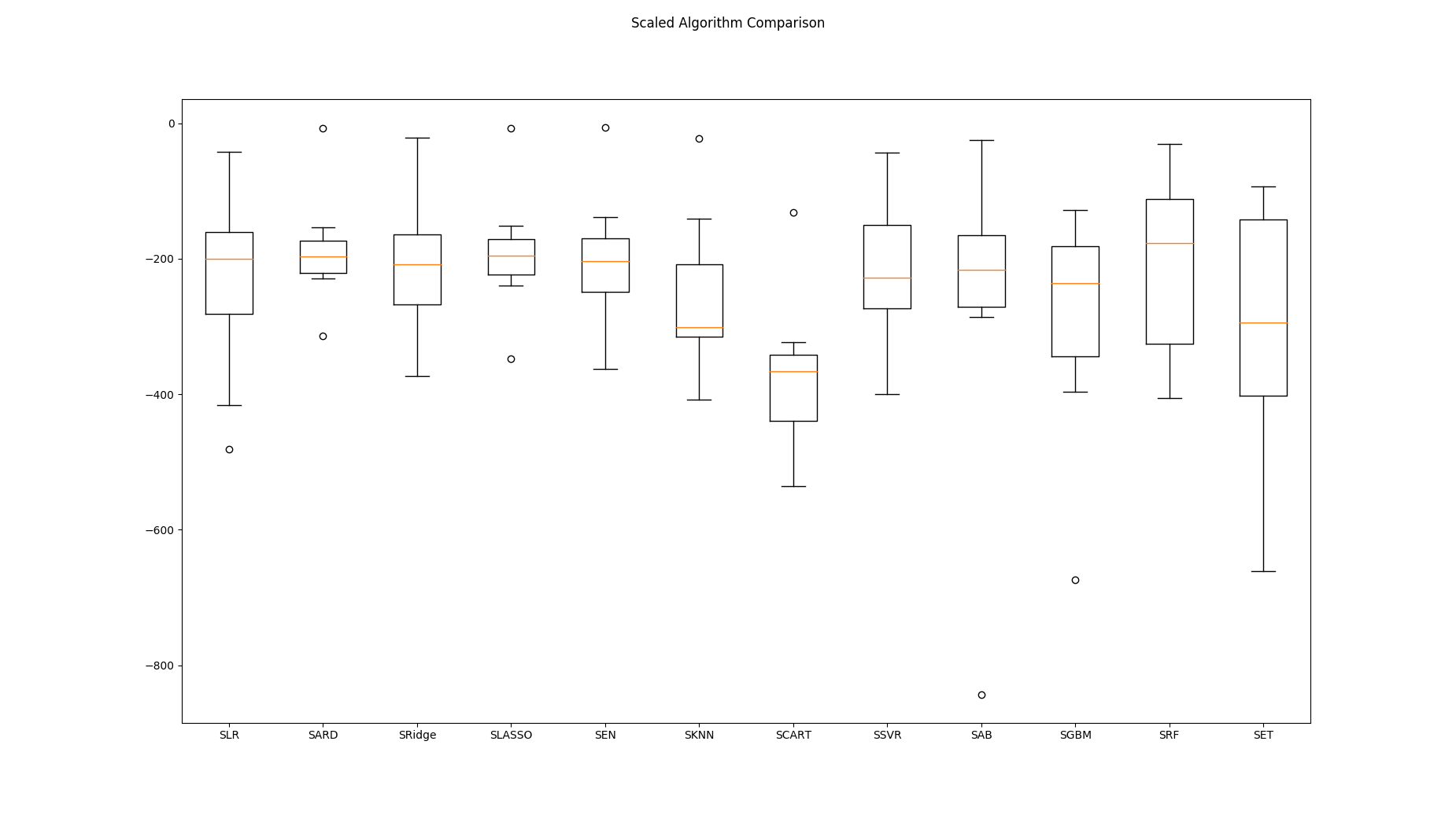


Fig. 7. Mean Squared Error after standardization for each algorithm

Model Information:

ARDRegression(alpha\_1=1e-06, alpha\_2=1e-06, compute\_score=False, copy\_X=True, fit\_intercept=True, lambda\_1=1e-06, lambda\_2=1e-06, n\_iter=300, normalize=False, threshold\_lambda=10000.0, tol=0.001, verbose=False)

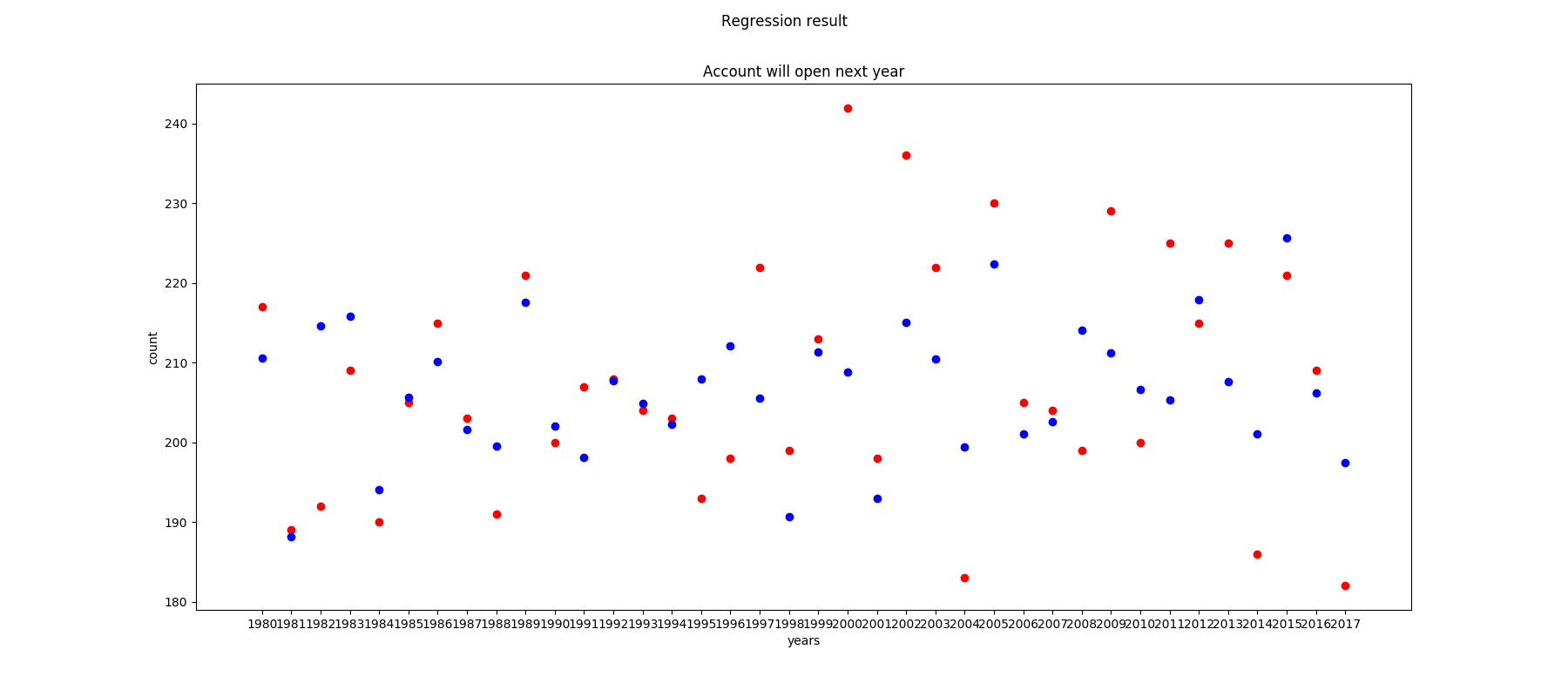


Fig. 8. The Regression Result (Red is the real data, and Blue is the predicted data)