Introduction:

The Main project we are assigned is a spark project named "South Boston Neighborhood Development" the goal of which is to identify buildings that may not even be on the market yet and then finding contact information of the person/people who own it. Since the initial project doesn't meet the requirements for class project we decided to define some interesting questions within the scope of our main project which could be potentially answered using the data sets given to us by spark.

The questions to be answered are the following:

1- Is there any correlation between the crime rate in a neighborhood and the average property value in that neighborhood? We answer this question by finding information about the buildings/properties in the most dangerous neighborhood of Boston area. These information include the value of these properties, type of the crimes happening in the neighborhood, permit number of buildings, owners of the properties,...

2- Is there any correlation between the number of food establishments in Boston areas and the average value of properties in that neighborhood? Can we model the relationship using a linear regression?We answer this question by Clustering food establishments in Boston area using their location and then finding information about properties in the each cluster.

3- Is there any correlation between the age and number of properties owned by a person?Can we model the relationship using a linear regression?

The detailed description of data transformation and analysis and results for each part is given in the following sections. First an overview of the data sets used is given.

Datasets:

Permit Database:

https://data.boston.gov/api/3/action/datastore_search?resource_id=6ddcd912-32a0-43df-9908-63574f8c7e77&limit=125650

Crime IncidentDatabase:

https://data.boston.gov/api/3/action/datastore_search?resource_id=12cb3883-56f5-47de-afa5-3b1cf61b257b&limit=366640

Active Food Establishments: https://data.boston.gov/api/3/action/datastore_search?resource_id=f1e13724-284d-478c-b8bc-ef042aa5b70b&limit=3010

Street Names:

https://data.boston.gov/api/3/action/datastore_search?resource_id=a07cc1c6-aa78-4eb3-a005-dcf7a949249f&limit=18992

Voter File:

Spark has been able to get the voter file from city hall which contains information (including residency address, phone, age, …) about people in south Boston who voted. This data set is really important since it contains personal information of people including their age and occupation and phone number. We have cleaned up this data set and converted it into json format.

now this data set is available at: http://datamechanics.io/data/asadeg02/Voter-File.json

Additional resources:

Zillow Search API:

https://www.zillow.com/howto/api/GetSearchResults.htm

We are using this API to create a database of the property addresses in south Boston that are on market.

Assesssors (Assessing online - City Of Boston):

https://www.cityofboston.gov/assessing/search

We are scraping this website to find the information we are interested in about the properties in City Of Boston including the value of properties, the owners and finding exact addresses on a street name.

Googlemaps Geocoding API:

https://www.cityofboston.gov/assessing/search

we are using this Geocoding API to find the latitude and longitude of the street addresses. We need to find latitude and longitude for being able to render the data on in an interactive map.

Algorithms And Transformations For Project:

In order to find the properties that are not on the market in South Boston we first need to somehow build a database of all the property “**addresses**” in south Boston. It is worth mentioning that we already have a data set of all the street **names** in all Boston neighborhoods but we are interested in finding the owner of the properties and in order to be able to find the owners we need to have addresses. To get the addresses we first filter the streets names in Boston street segment data-set and and then we scrape assessors website using these names to find all the properties information on these streets(including their address, owner, value, parcel id) . Then we get latitude and longitude for these addresses using Googlemaps Geocoding API.

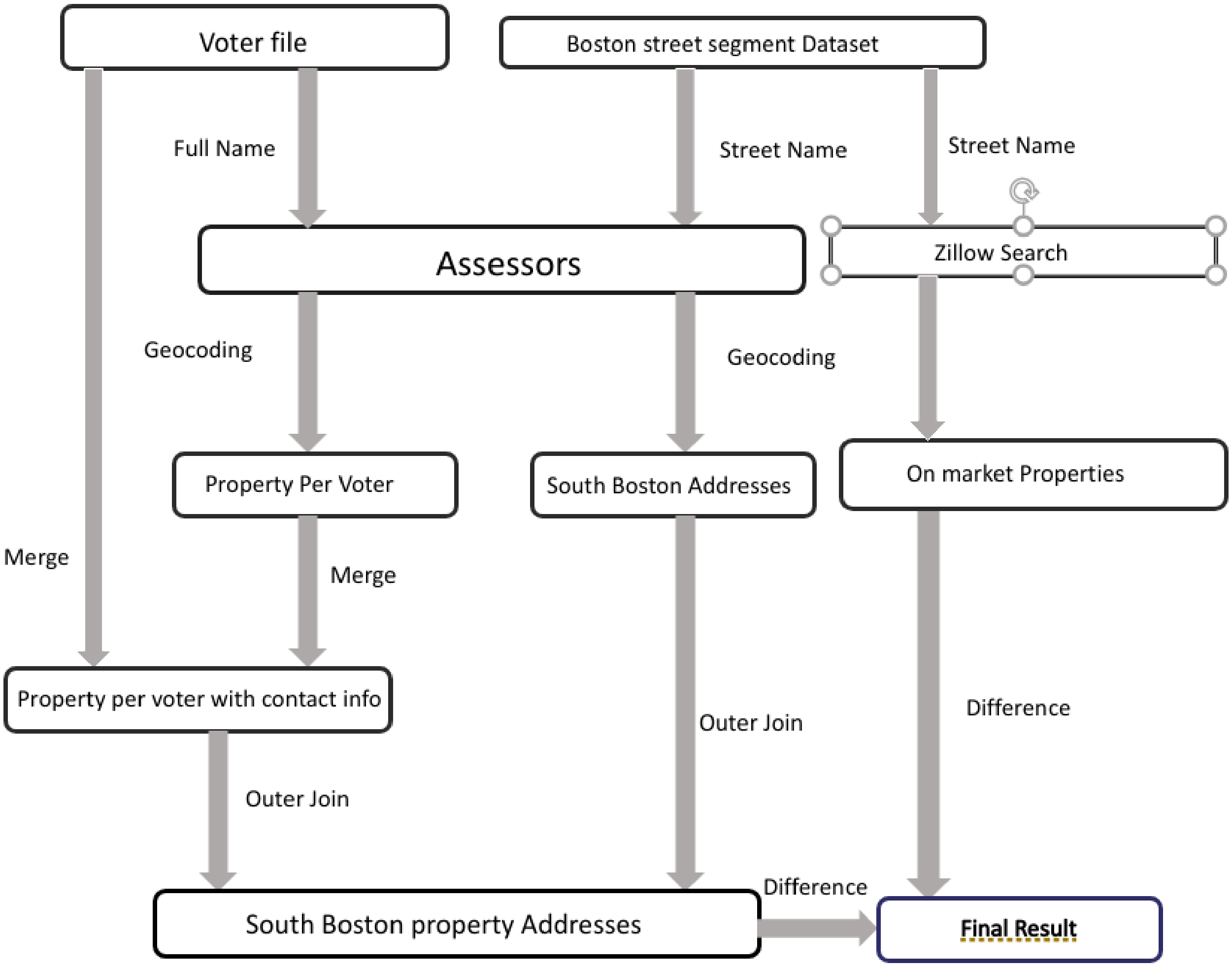
Since our ultimate goal is to find the contact information of people and the only people whose contact information we have are the ones in the voter file we need to find the properties owned by theses people.(even in the voter file the phone attribute of some of the entries is nan).

Please keep in mind addresses in voter file are the voters residency addresses but we are interested in finding first all the properties (Potentially in Boston) they own. In order to do so, we extracted all the first names and last names from voter file and we used last name + first name the search key to scrap assessors which gives us the exact addresses of all buildings/properties in Boston owned people. This data set is now available at: http://datamechanics.io/data/asadeg02/Property-Per-Voter.jso

then we get latitude and longitude of these properties.

Then we do a left inner join on Address and latitude and longitude between these 2 data sets to filter only the properties in south Boston owned by people in voter file(these properties are the only ones we contact information of their owners) then we do a right outer join between property per voter data set and south Boston properties (to get the union)

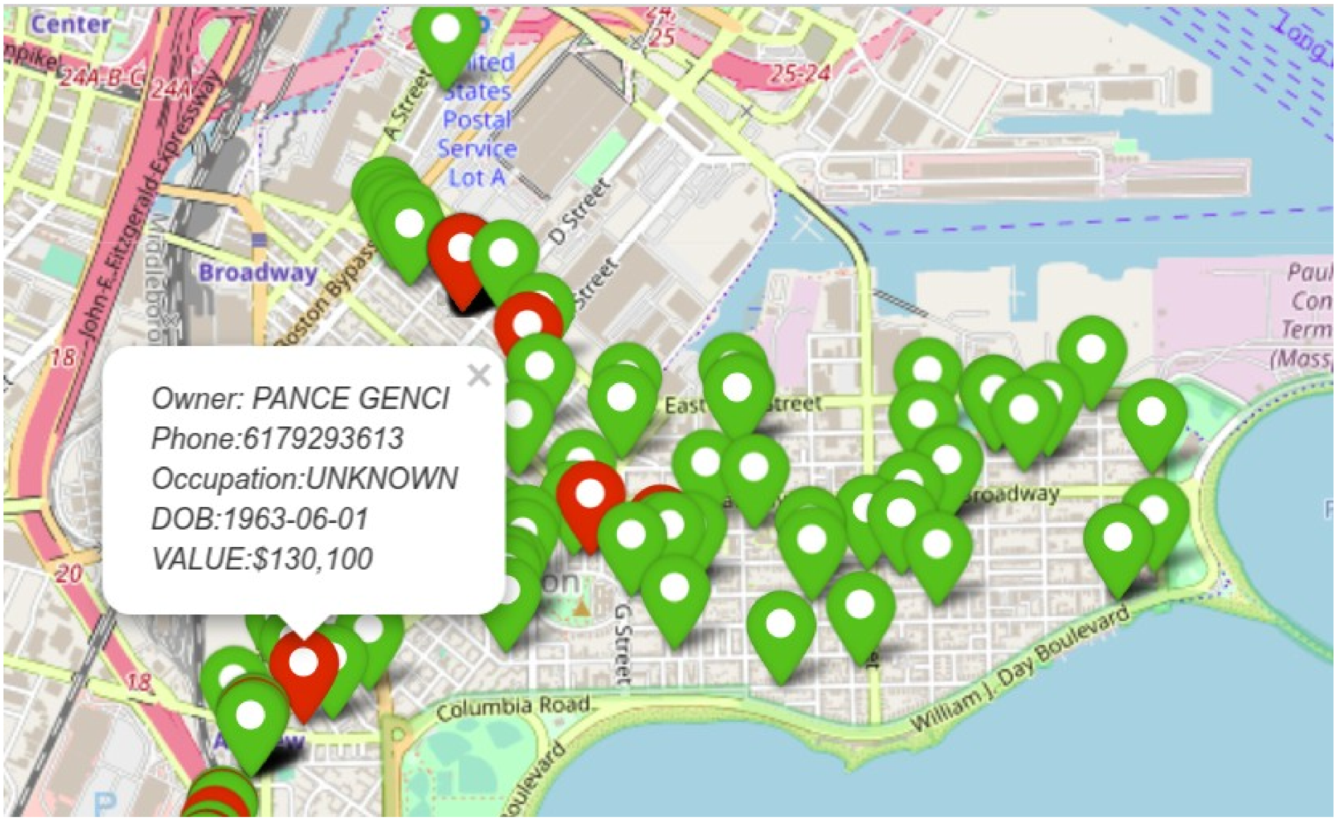
then merge this data set with the data set obtained from Zillow search API (address of properties in south Boston which are on market) to get their intersection and then remove this part since we are only interested in the properties which are not on market.



Result: we have built a java script interactive web-base simple app which shows the results in two different formats (data table with search and pagination functionality) and and an interactive map. Please note that this map has been created using folium and the HTM file generated by folium has been embedded into the app.

Red markers on the map are the properties that we have their owners' contact information. As you can see there aren't many red markers on the map since

a lot of people in the voter file either didn't own any properties at all or the properties they own is not in south Boston.



Owner: PANCE GENCI
Phone:6179293613
Occupation:UNKNOWN
DOB:1963-06-01
VALUE:$130,100

Challenges:

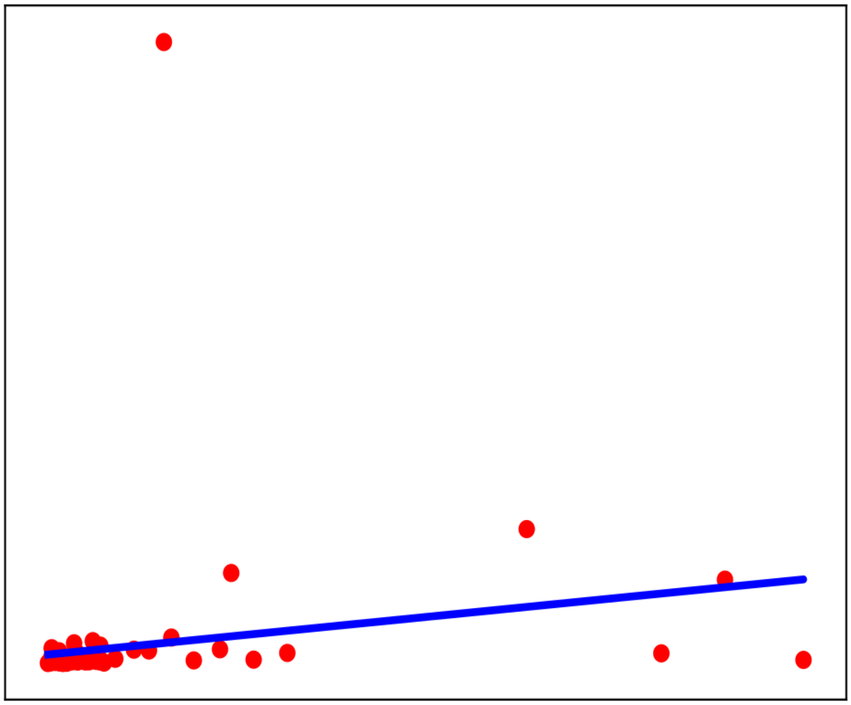
Finding all the addresses in South Boston without having false positives is a challenging task since scraping assessors using street names gives you a all addresses from around Boston with the same street names as in South Boston neighborhood. Filtering the final result by latitudes and longitudes require user-specified constraints.

Correlation between safety of an area and the average value of properties in that area:

In order to find the correlations between the safety of a street and the value of the properties in that street and see if there is a good linear model for modeling the relationship between these two, we have divided the Boston streets into two categories: safe street and dangerous street. safe streets are the ones which don't have any record in "address_crime_rate" which is a data set derived from “Boston crime incident” data set and dangerous streets which appear in address_crime_rate" data set. obviously crime rate for safe streets is zero. for modeling the regression and correlation we only sample a few streets from safe streets since the number of street with crime rate value equal to one is a lot more than dangerous streets. In order to get good regression results we have constrained the number of safe street we include in our analysis.

In order to computer the average property value in a street, we have scrapped assessors and stored the result data set in "crime_rate_mean_value" repo.

Result:

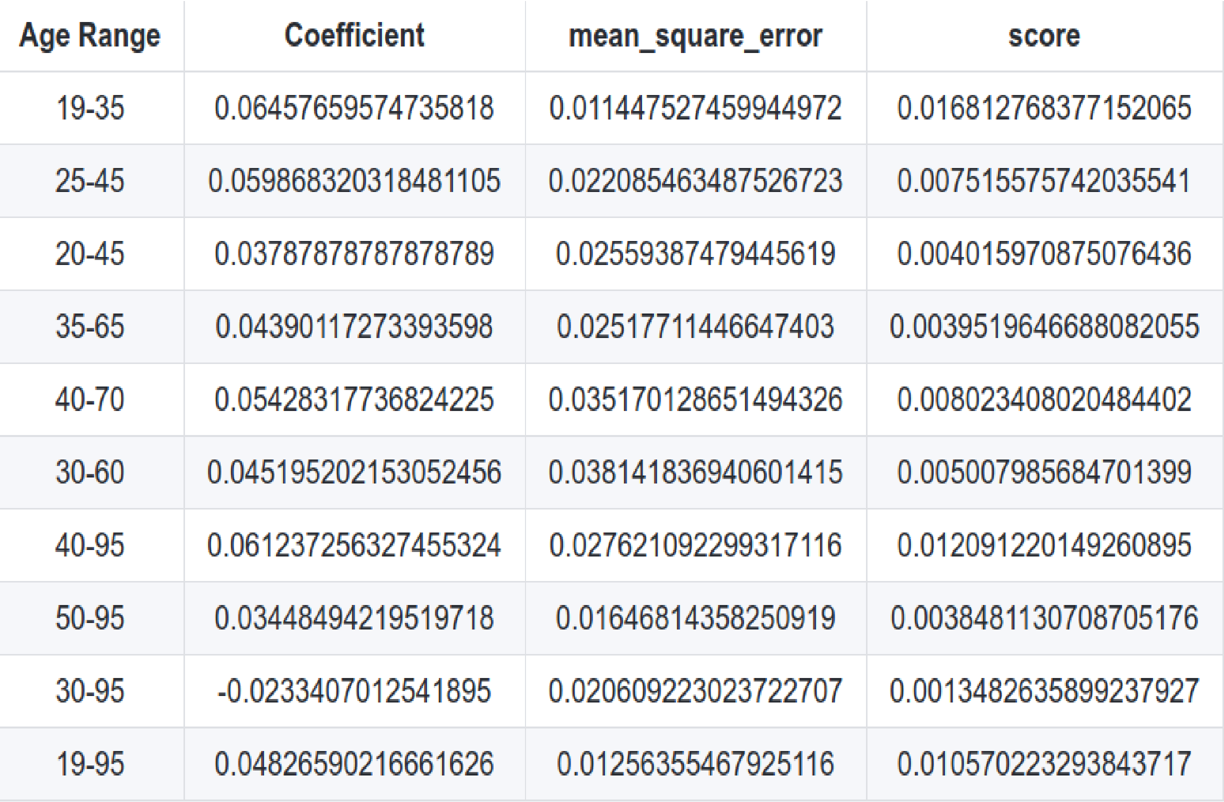


Correlation between the age and number of properties owned by a person

In order to find the correlation between age of a person and the number of properties that person owns. we have merged "voters_info" repo and "properties_per_voter" repo to be able to get a data set whose documents include "age", "numProperties", "phone", "occupation" "addr" attribute. This address is the address of the buildings they own and not (the address where the live included in voter file) that's why we we had to merge these two data sets to be able to access personal information of owners. the merge has been done on both first name and last name. and finally since we aggregate the results in by owner's complete names to count the number of properties each person owns and get their age.

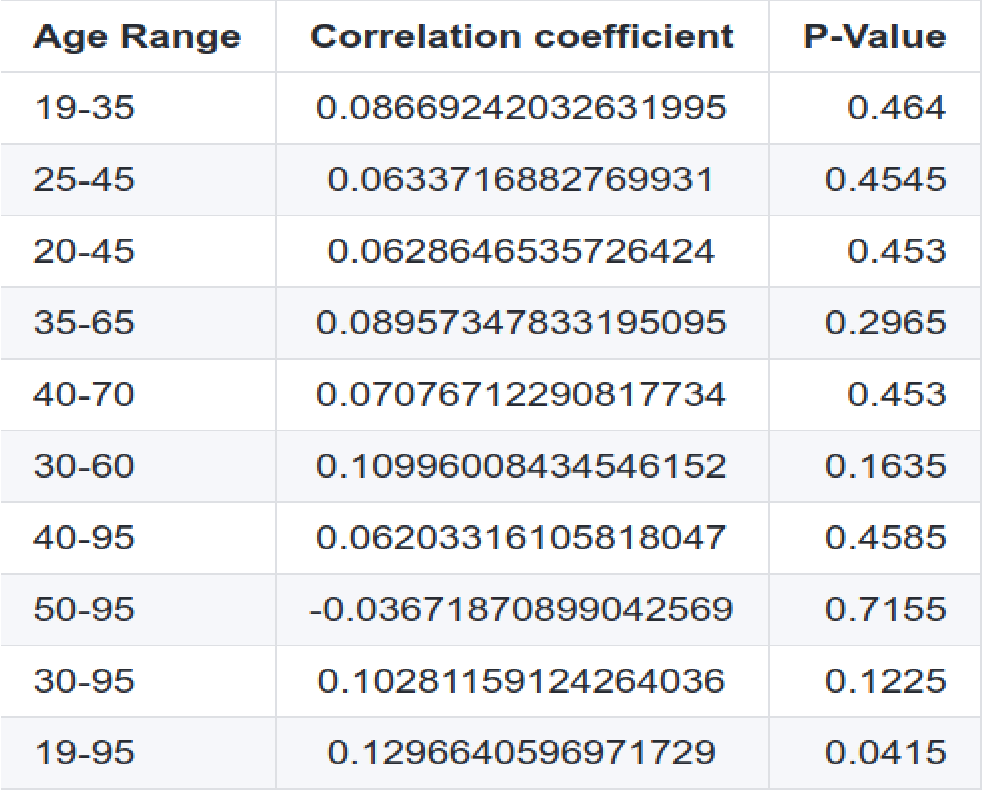
Results:

Table1



Age Range	Coefficient	mean_square_error	score
19-35	0.06457659574735818	0.011447527459944972	0.016812768377152065
25-45	0.059868320318481105	0.022085463487526723	0.00751557574203541
20-45	0.03787878787878789	0.02559387479445619	0.004015970875076436
35-65	0.04390117273393598	0.02517711446647403	0.0039519646688082055
40-70	0.05428317736824225	0.035170128651494326	0.008023408020484402
30-60	0.045195202153052456	0.038141836940601415	0.005007985684701399
40-95	0.061237256327455324	0.027621092299317116	0.012091220149260895
50-95	0.03448494219519718	0.01646814358250919	0.0038481130708705176
30-95	-0.0233407012541895	0.020609223023722707	0.0013482635899237927
19-95	0.04826590216661626	0.01256355467925116	0.010570223293843717

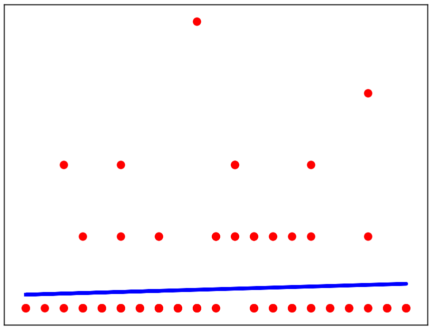
Table2



Age Range	Correlation coefficient	P-Value
19-35	0.08669242032631995	0.464
25-45	0.0633716882769931	0.4545
20-45	0.0628646535726424	0.453
35-65	0.08957347833195095	0.2965
40-70	0.07076712290817734	0.453
30-60	0.10996008434546152	0.1635
40-95	0.06203316105818047	0.4585
50-95	-0.03671870899042569	0.7155
30-95	0.10281159124264036	0.1225
19-95	0.1296640596971729	0.0415

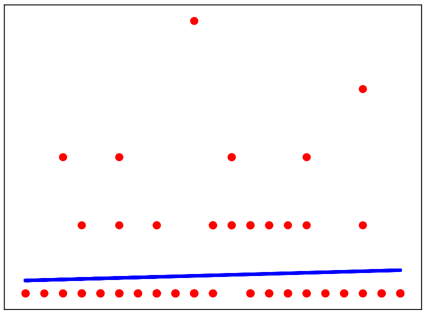
As you can see when we the do the analysis on the entire data set (19-95 age range) the model seems to be a better predictor in the sense that he average distance of the red circles from the blue line is less. Looking at the table1 the mean square error for the two has its minimum value for age ranges (19-35) and (19-95) and the mean square error has its maximum value for age ranges (30-60) and (40-70).

Age range: 20-40

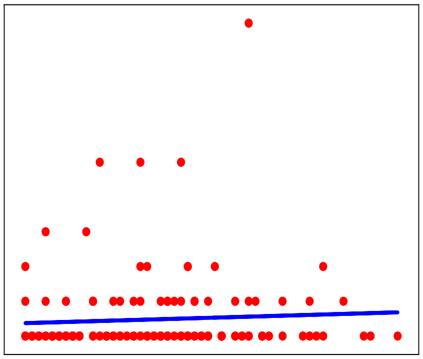


\

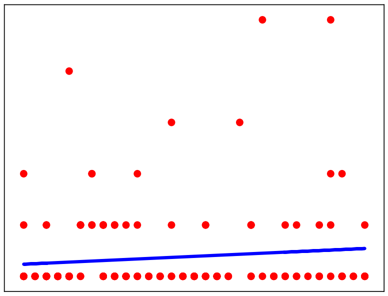
Age- range:25 -45



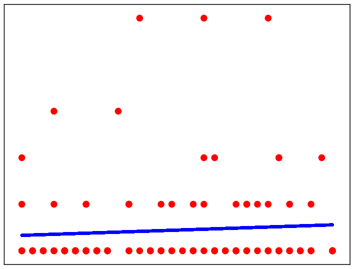
Age range: 19-95



Age range 30-60



Age range: 35-65



We have created a web app with flask which lets you choose the age range interactively and display the results.