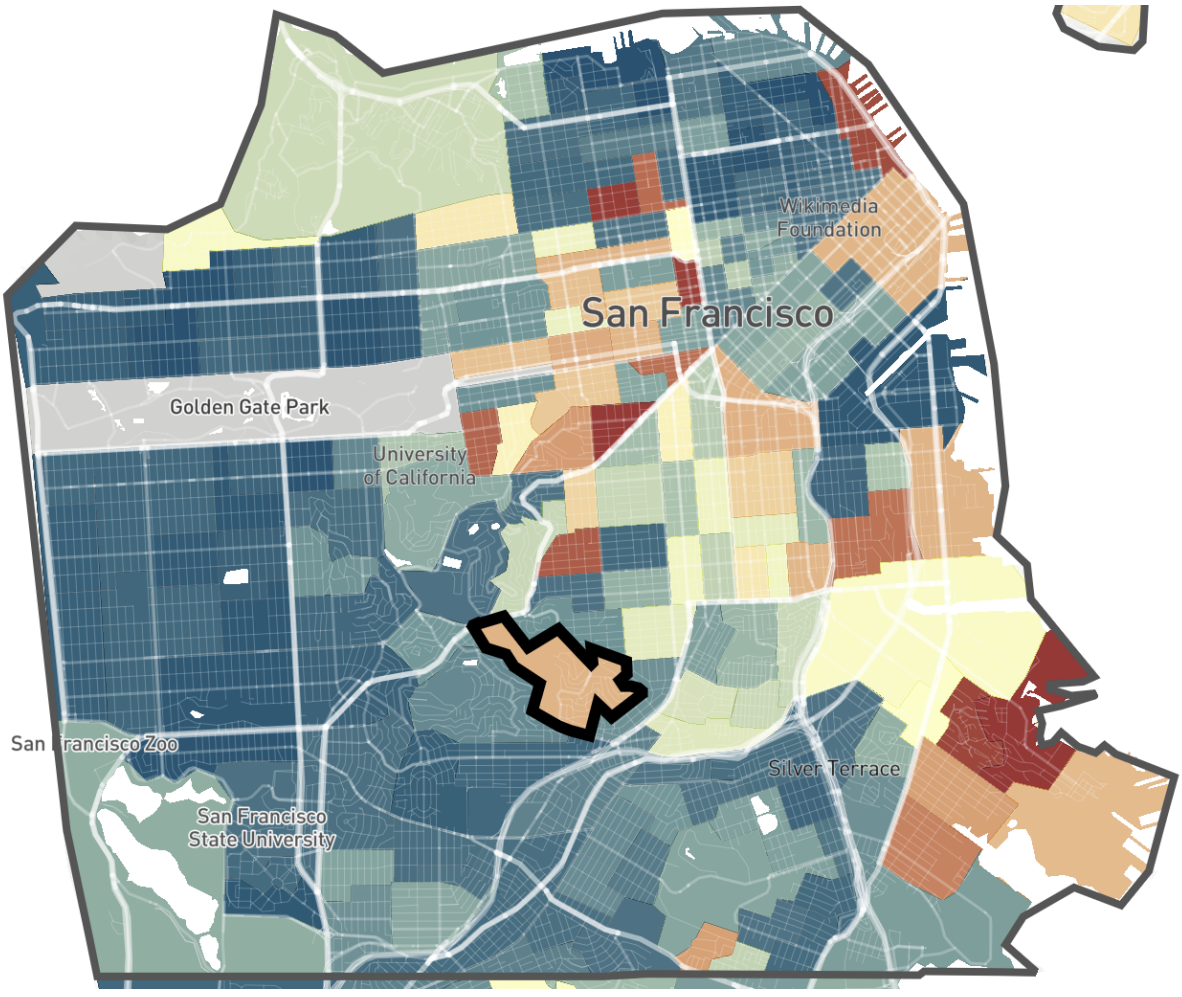
Big data project 1

Moskovets Artem, SMD 2

Map insights

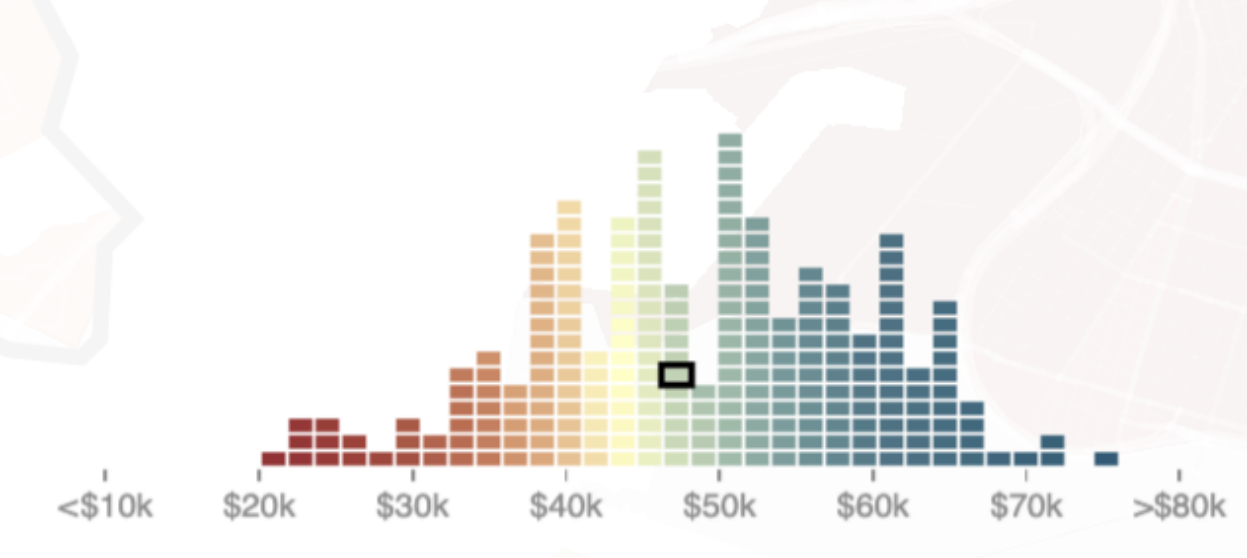
For this project I’ve decided to choose tract **06075021700**, San Francisco, California.



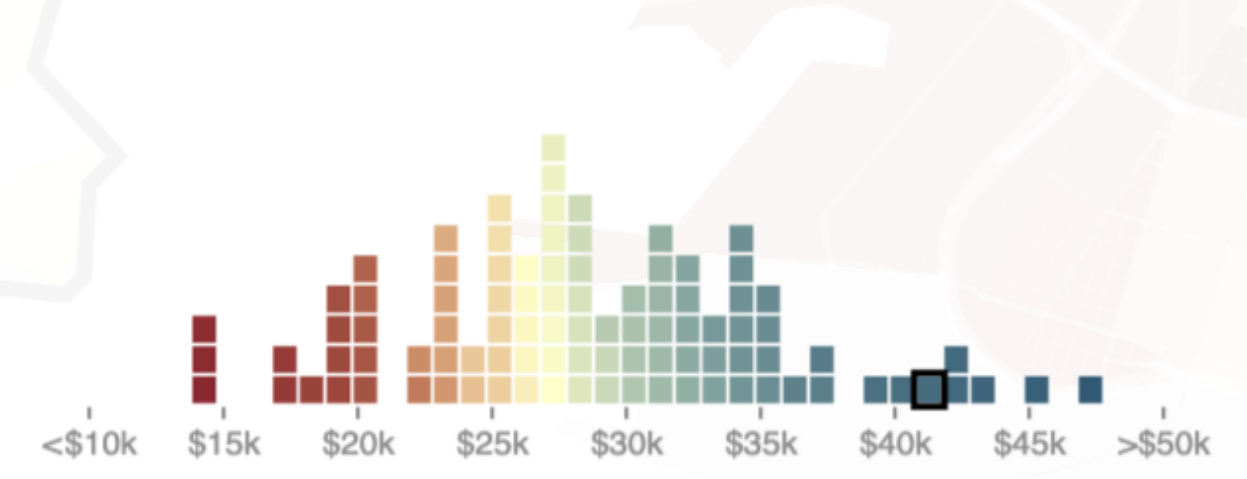
As we can see from the map, household income distributed not equally between the parts of the city. We can see bigger income at the west and north of the city, and lower incomes at the eastern parts of the city.

To understand better this map, let’s look at salary distribution for different race groups:

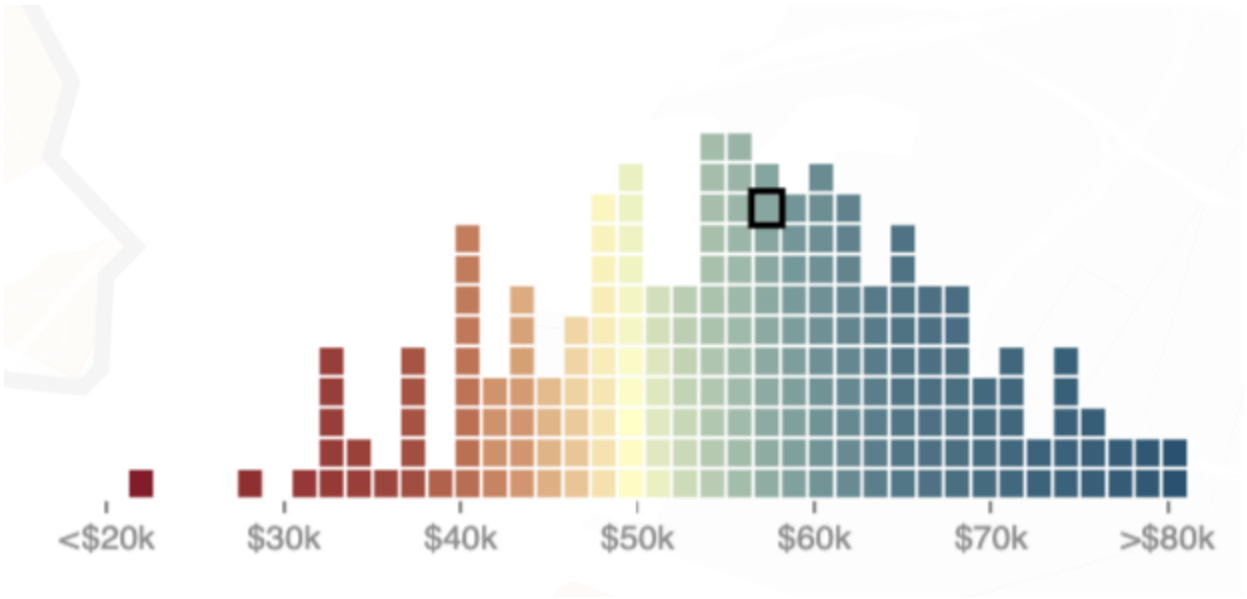
All races



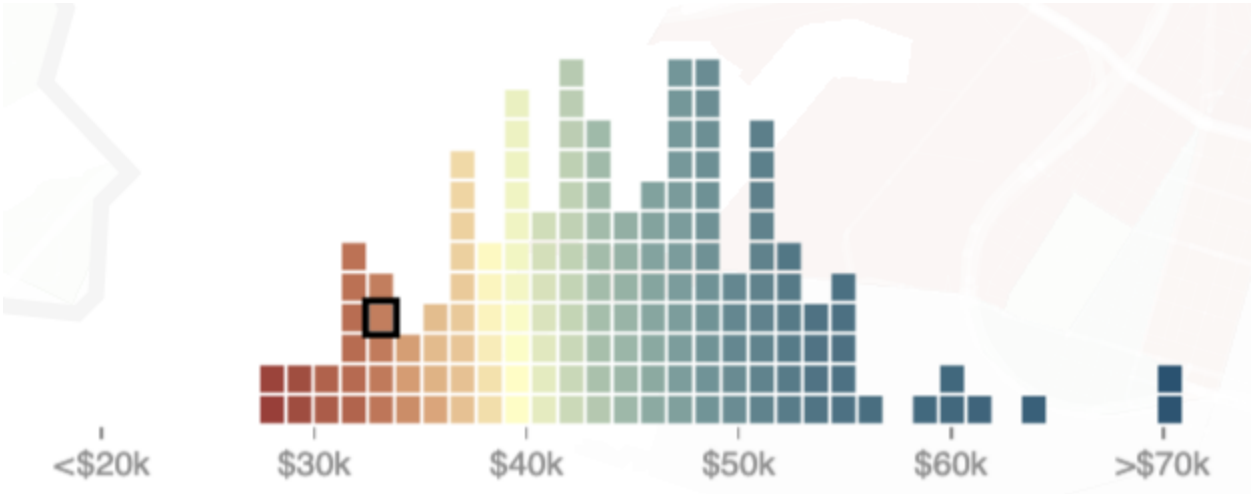
Black



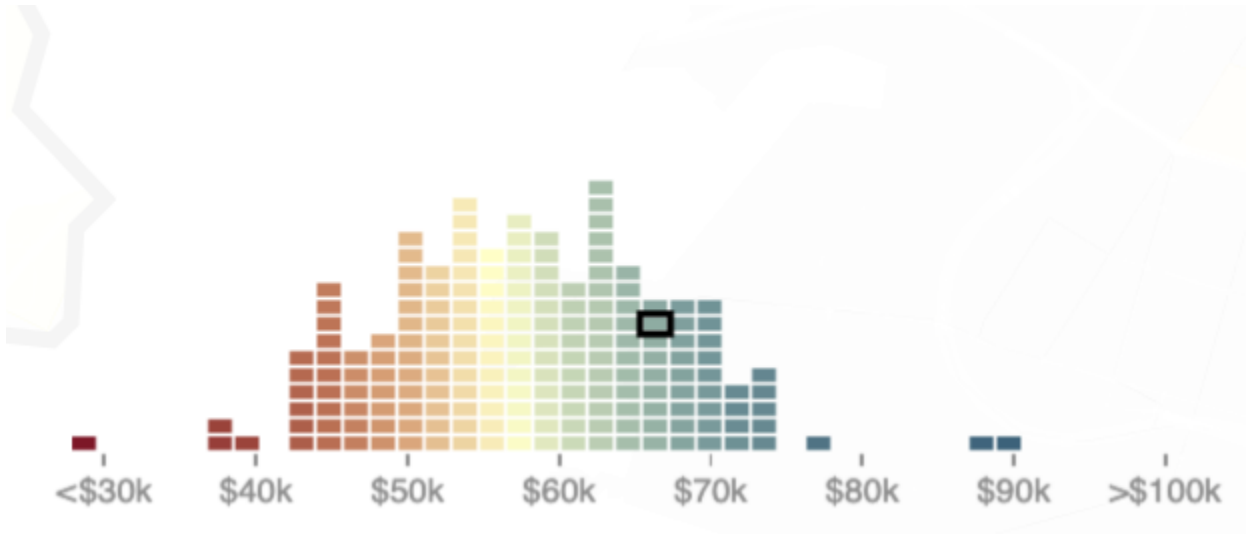
White



Hispanic



Asian



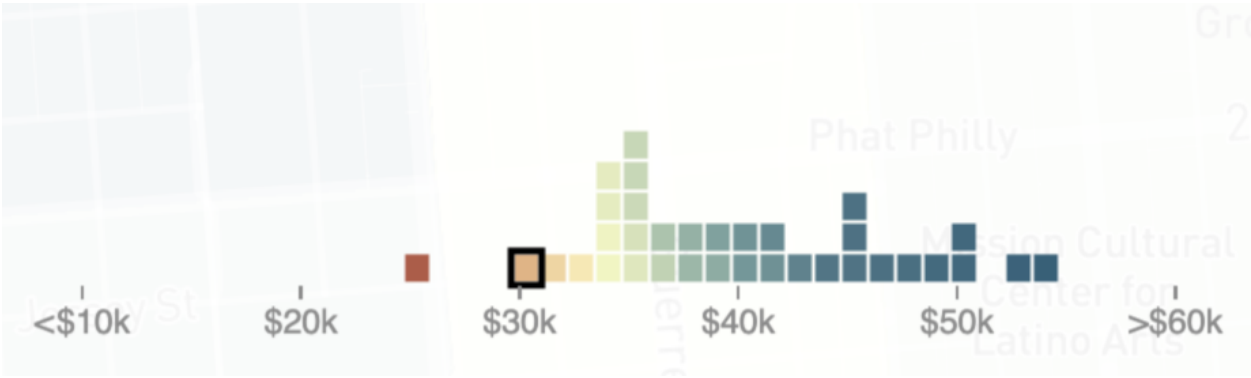
As we can see, the poorest category here is Hispanic. After that Black, White, and Asian. Worth noticing, that for every group we see standard normal data distribution, which is quite similar for all groups (except the absolute values).

Worth to notice, that dataset covers the records from around 50 last years, but doesn't include a new data. The ideas of education and human behavior have changed significantly, and the results for children who'll be 35 in 20-30 years can be significantly different.

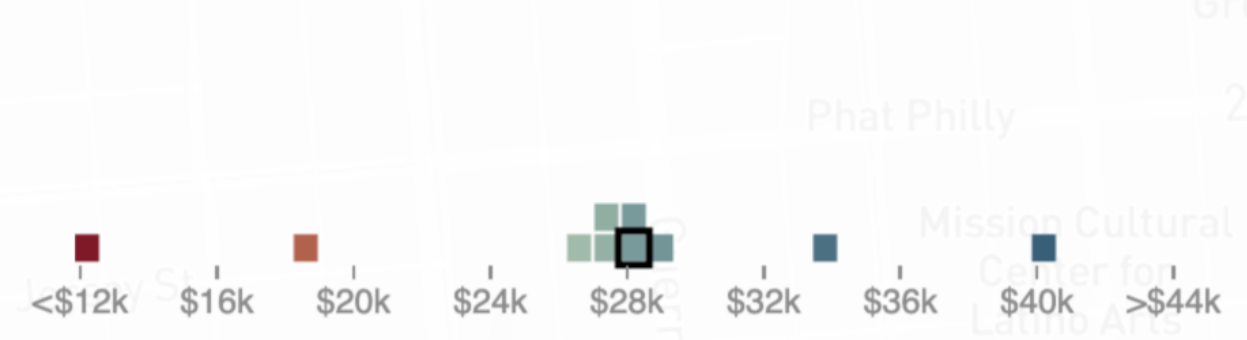
25th percentile analysis

To see the upward mobility for 25th percentile in my tract, let's use the map:

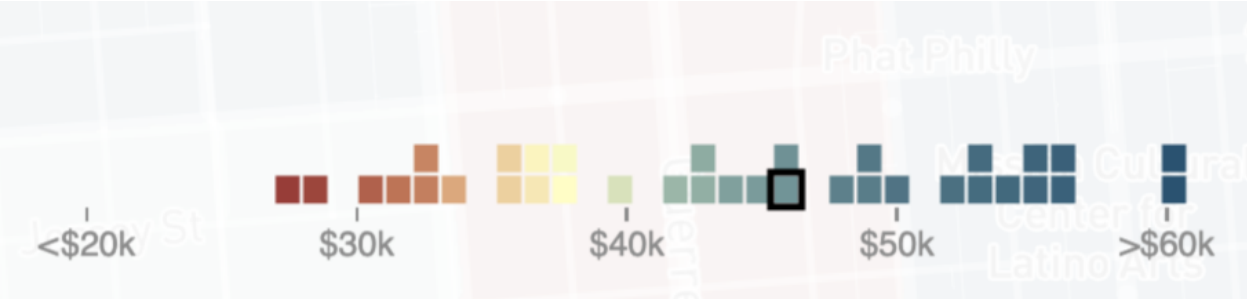
All races



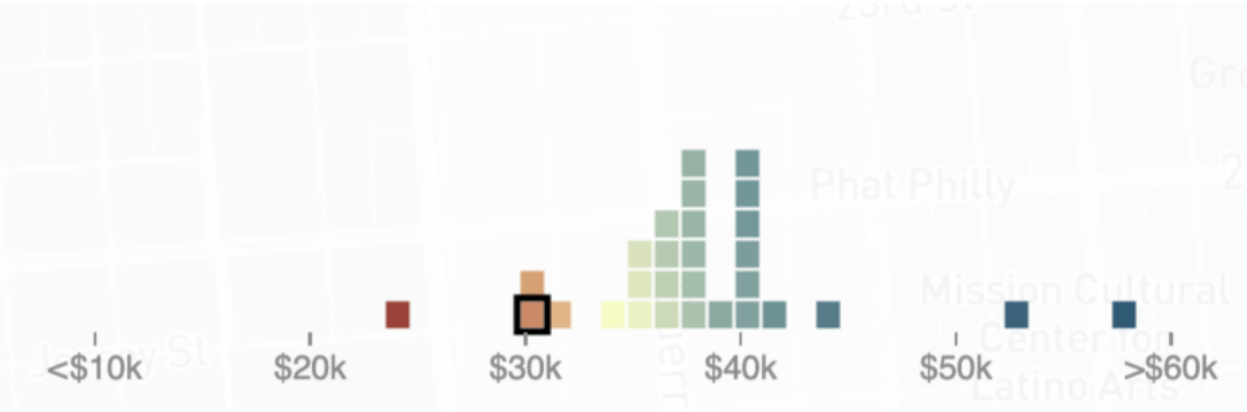
Black



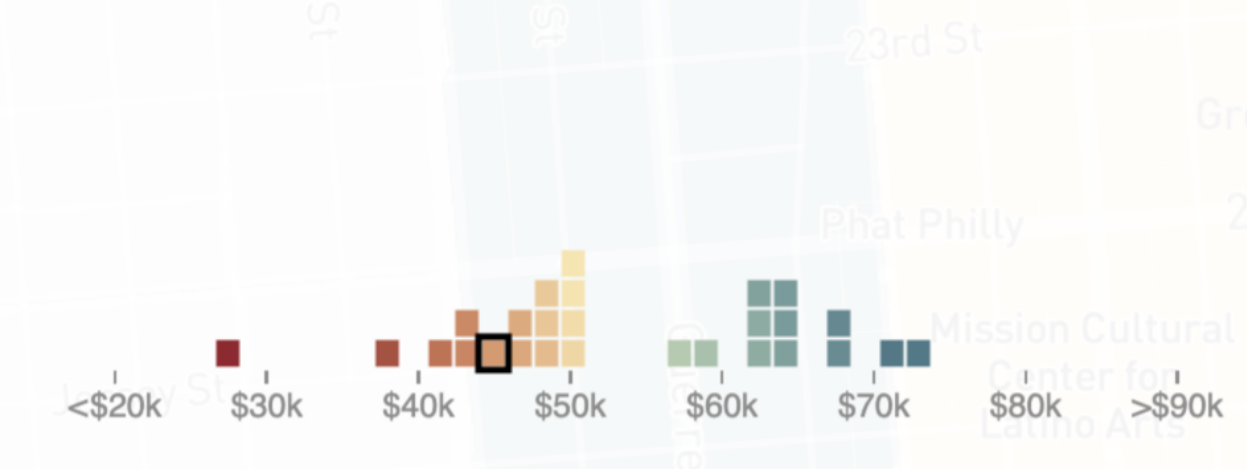
White

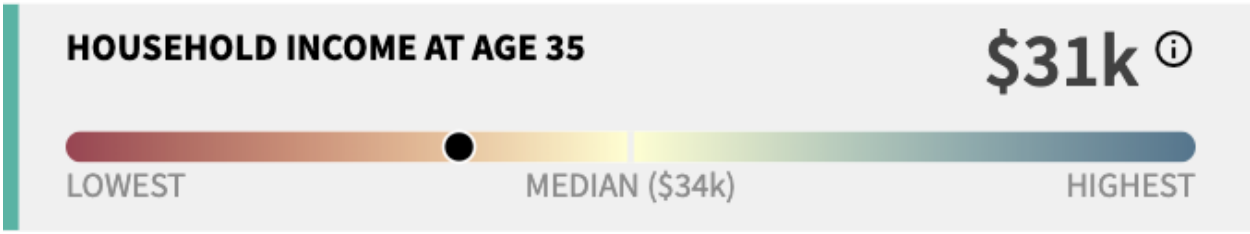


Hispanic

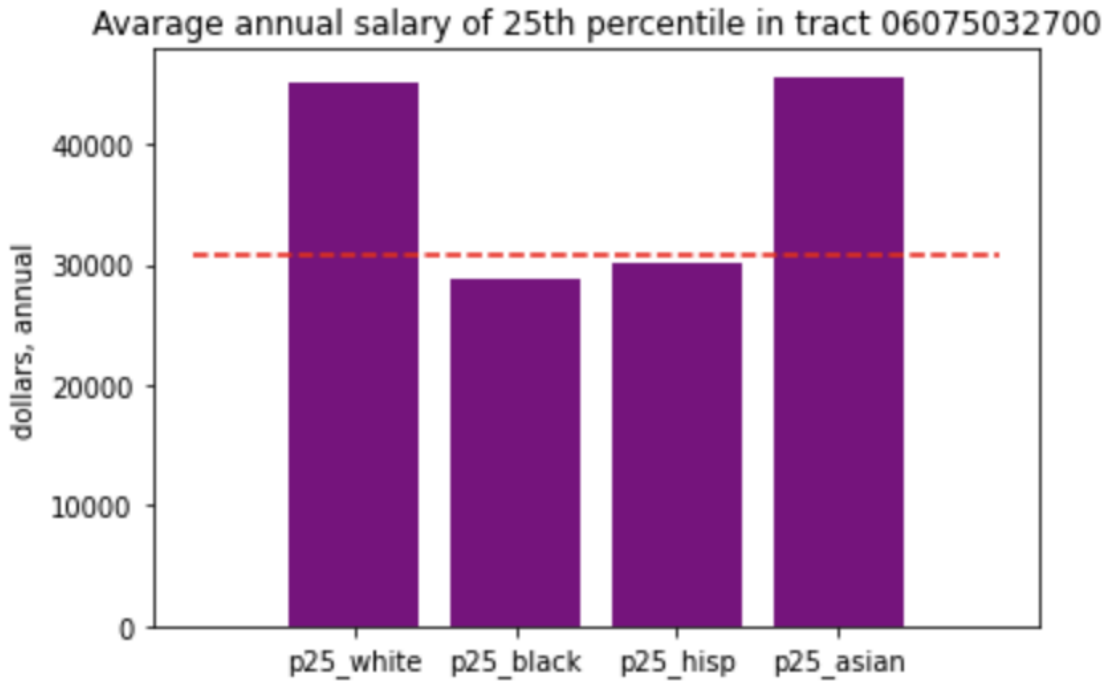


Asian





This plot represents all data together:



So, average income of 31k in my tract is averagely lower, compared to state / the US results.

If we look at the standard deviation of 25th percentile in county, state and the US, we'll see this results:

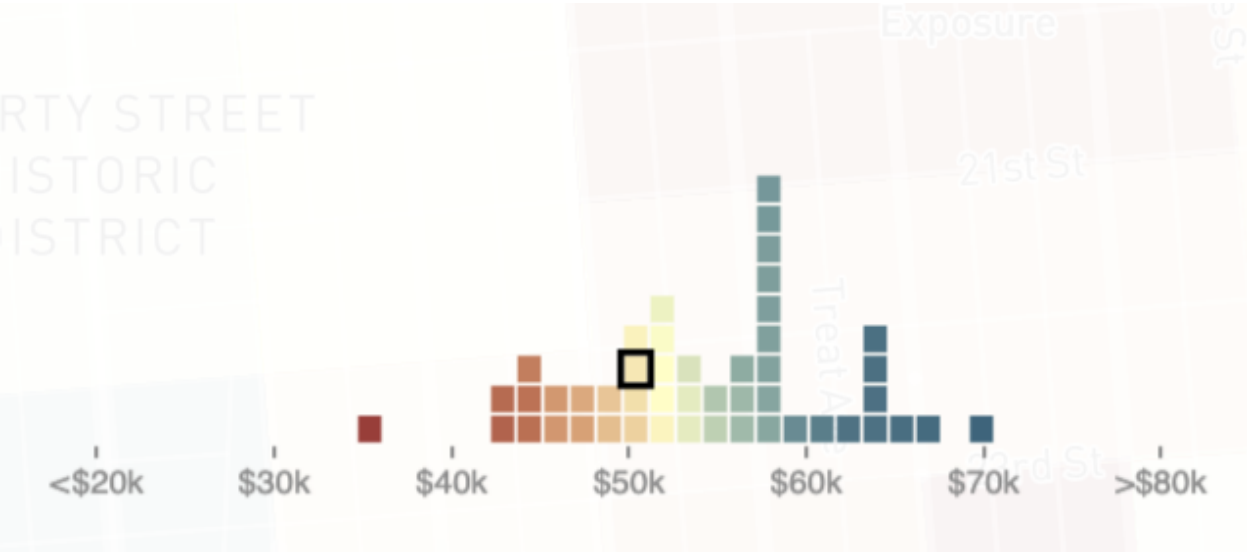
```
average income and real standard deviation in county = 35938.50020, 9161.63003
average income and real standard deviation in California = 35765.82345, 6559.771
average income and real standard deviation in the US = 34311.68270, 7899.53107
```

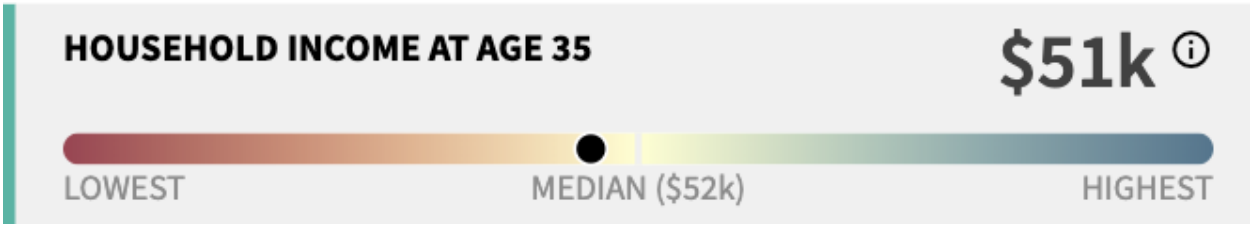
So, average income in San Francisco is bigger than in California state and the US

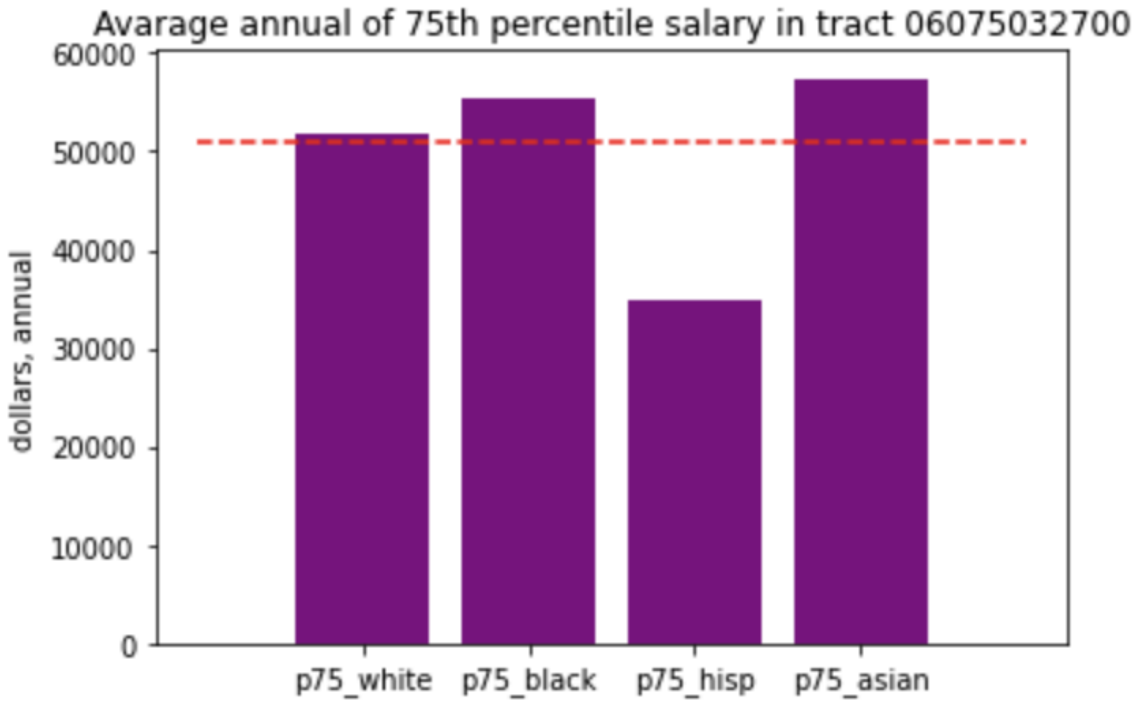
The real standard deviation is much bigger in county, because some of the tracts have much bigger income than others, which we can see on the map. The deviation is bigger for the US, than for California state, because US, obviously includes much more tracts with a huge positive, and negative difference between the neighbourhoods.

Downward mobility for 75th percentile

To see the downward mobility for 75th percentile, let's use the map:



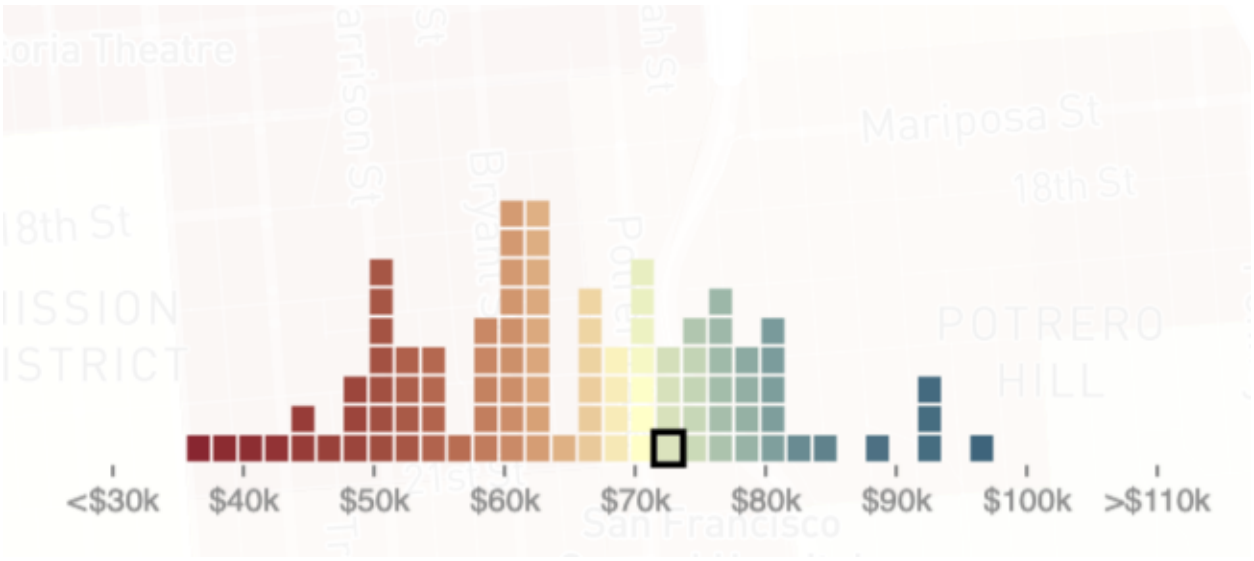


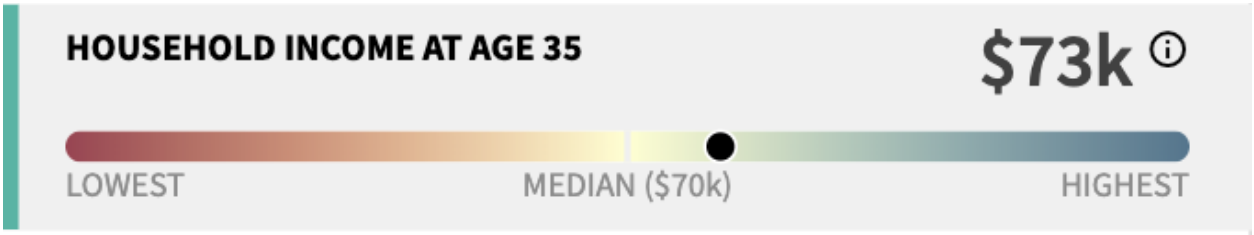


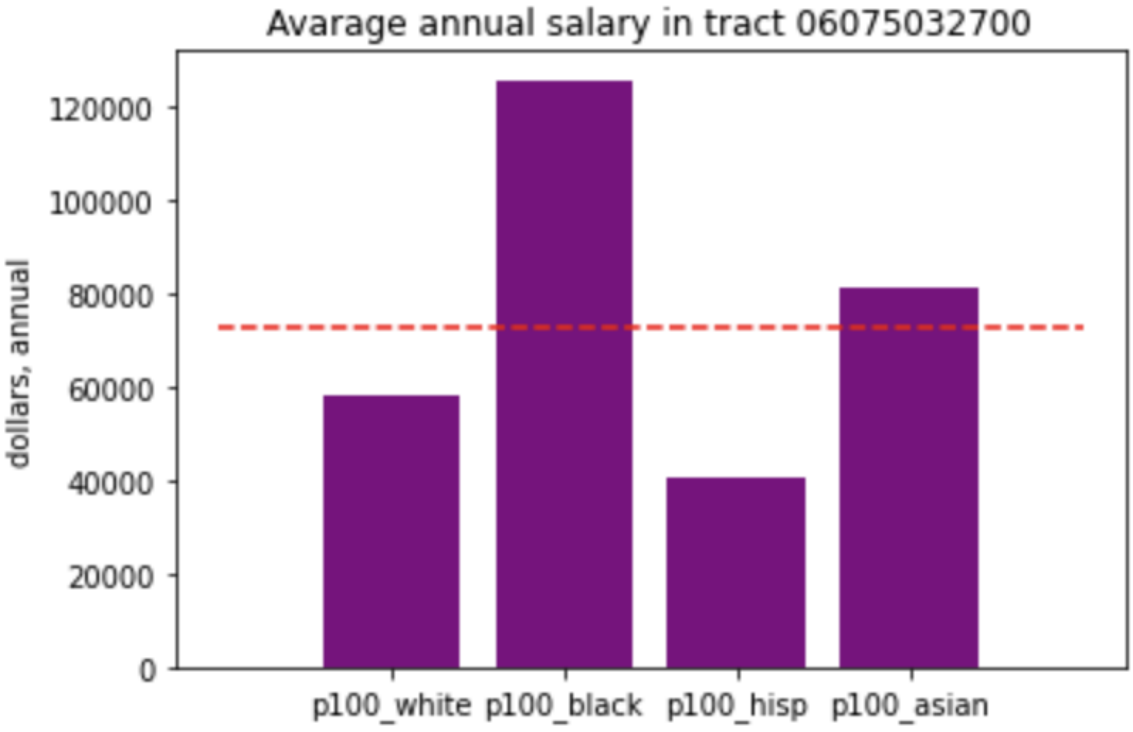

```
average income and real standard deviation in county = 53547.62574, 8903.78771
average income and real standard deviation in California = 48995.07187, 7395.998
average and real standard deviation in the US = 51284.02767, 9326.02872
```

Here we can see quite close income for White, Black and Asian people, but Hispanic's income is much lower. The downward mobility is quite strong for this tract. We can spot the same pattern for std for 75th percentile.

Downward mobility for 100th percentile

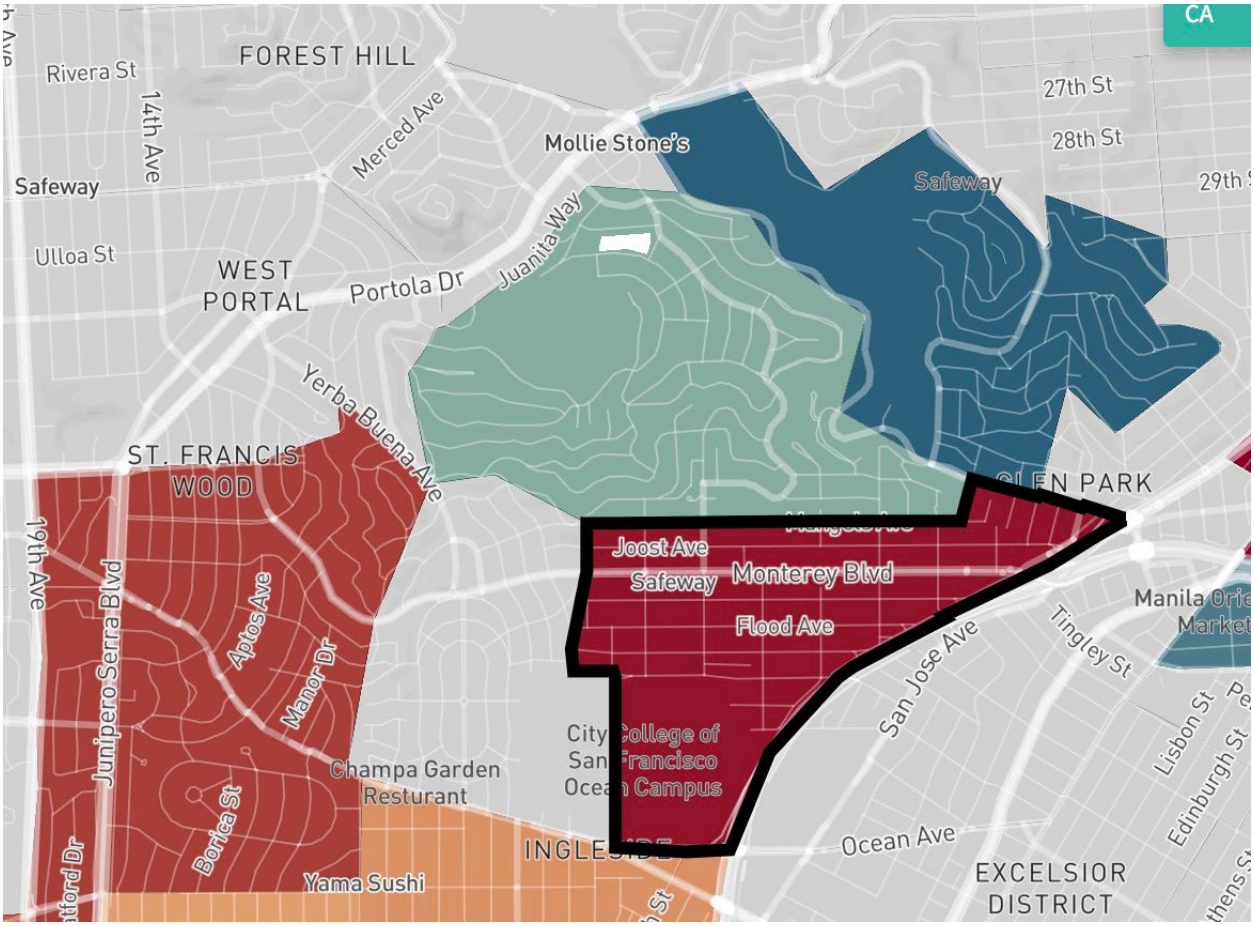

Mariposa St
18th St
POTRERO
HILL
<$30k
$40k
$50k
$60k
$70k
$80k
$90k
$100k
>$110k


HOUSEHOLD INCOME AT AGE 35
$73k
LOWEST
MEDIAN ($70k)
HIGHEST


Avarage annual salary in tract 06075032700
dollars, annual
120000
100000
80000
60000
40000
20000
0
p100_white
p100_black
p100_hisp
p100_asian

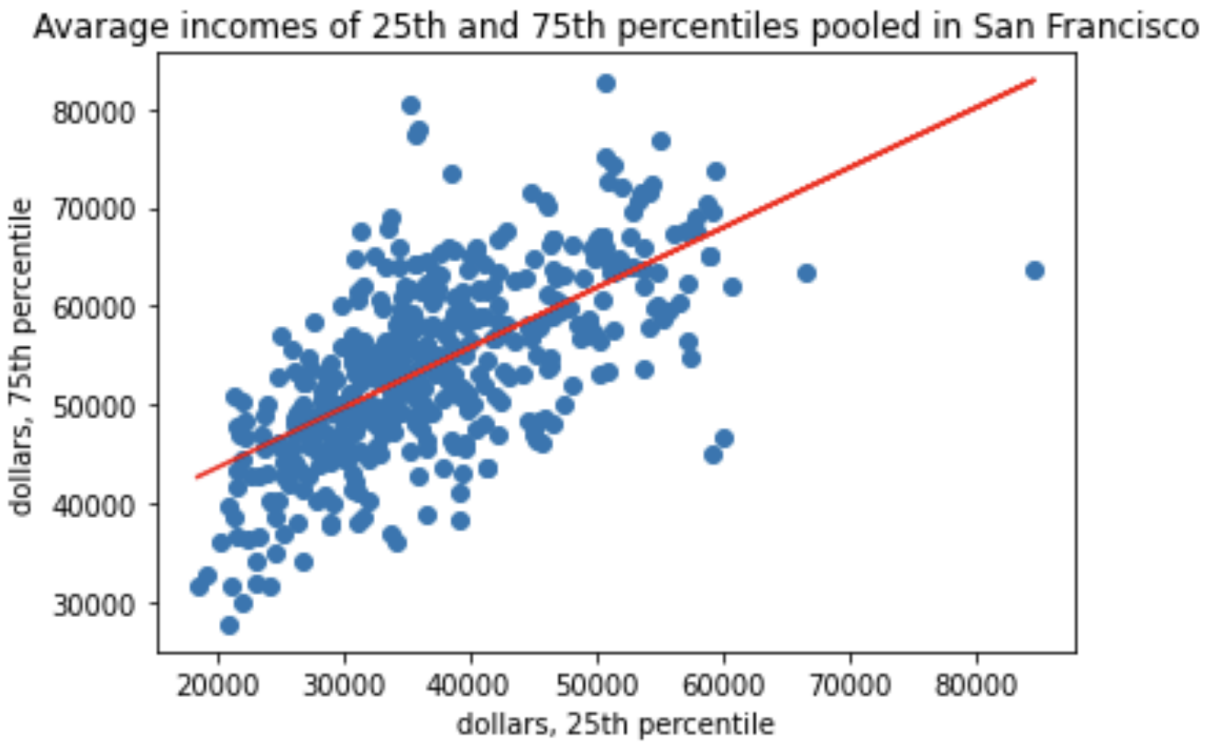
```
average income and real standard deviation in county = 72806.40983, 17267.18964
average income and real standard deviation in California = 62346.79485, 14692.66
average income and real standard deviation in the US = 69218.14595, 16362.69301
```

The downward mobility, predictable, is even higher than for 75th percentile. The interesting thing here - income of Black people. It’s huge, compared to other races, that makes such std for this tract. Btw, if we look at the same 100th percentile of Black people in other tract - we’ll see significant difference - avarage result of 7.3k (in my tract it’s 130k).



Average income of 100th precentile is bigger in San Francisco, than in California / the US.

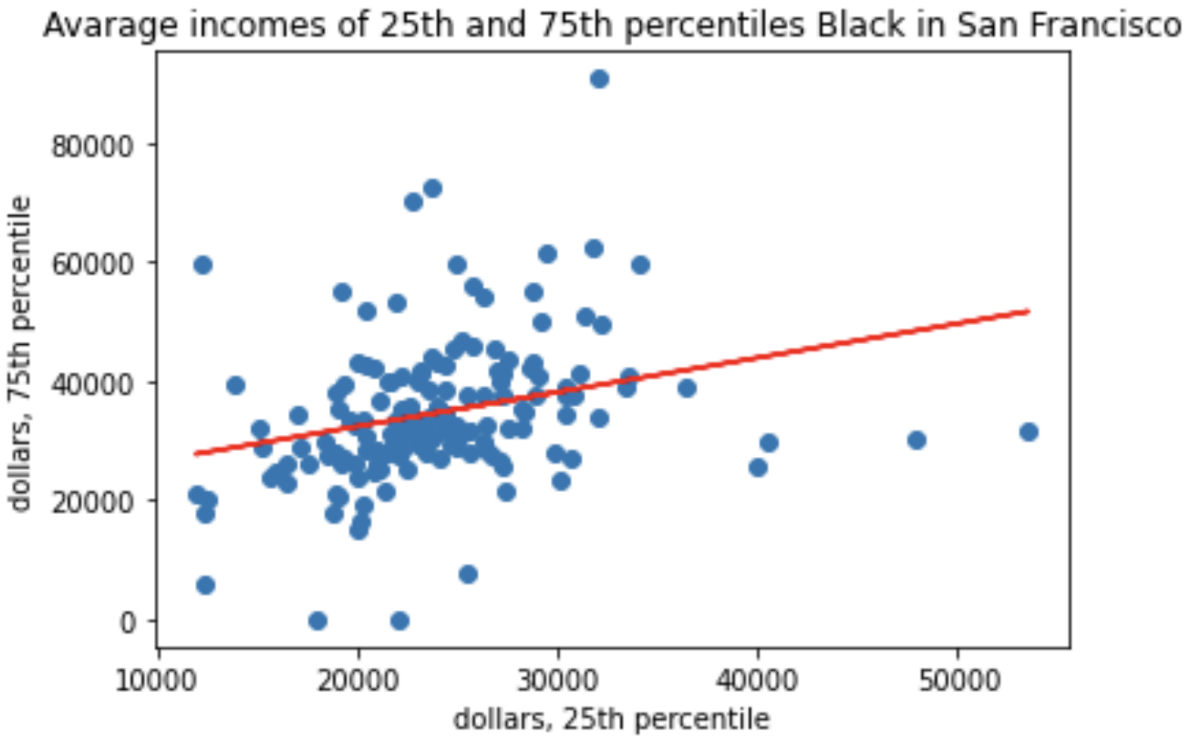
Correlation between income of 25th and 75th percentiles (pooled)




```
r_value = 0.6442171408092052
p_value = 2.3559145552157244e-54
standard error = 0.03405860153282094
```

We can see quite strong correlation, and low p-value and std, analizing relation between 25th and 75th percentile's income in San Francisco. That tells that Linear regression fits well here.

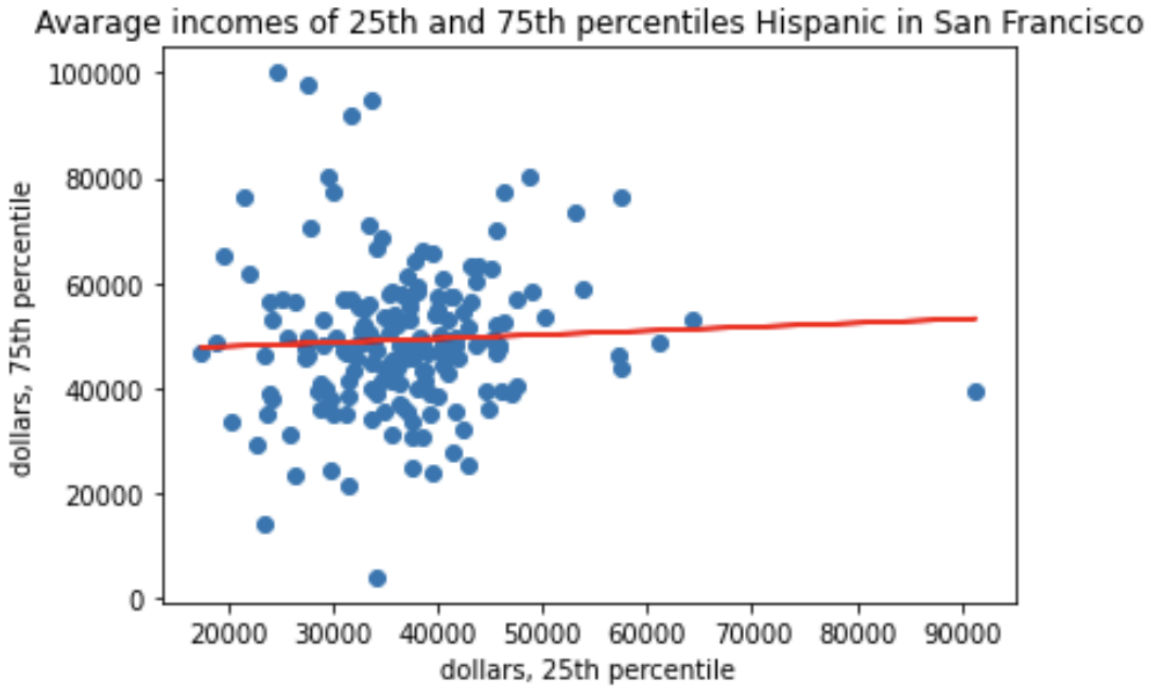
Correlation between income of 25th and 75th percentiles (Black)




```
r_value = 0.27873914073018374
p_value = 0.0004446590654271183
standard error = 0.16068355326204614
```

Here we have very weak correlation, small p-value, but big std. That means that Linear regression doesn't fit well here, and the correlation between 25th and 75th percentile is weak.

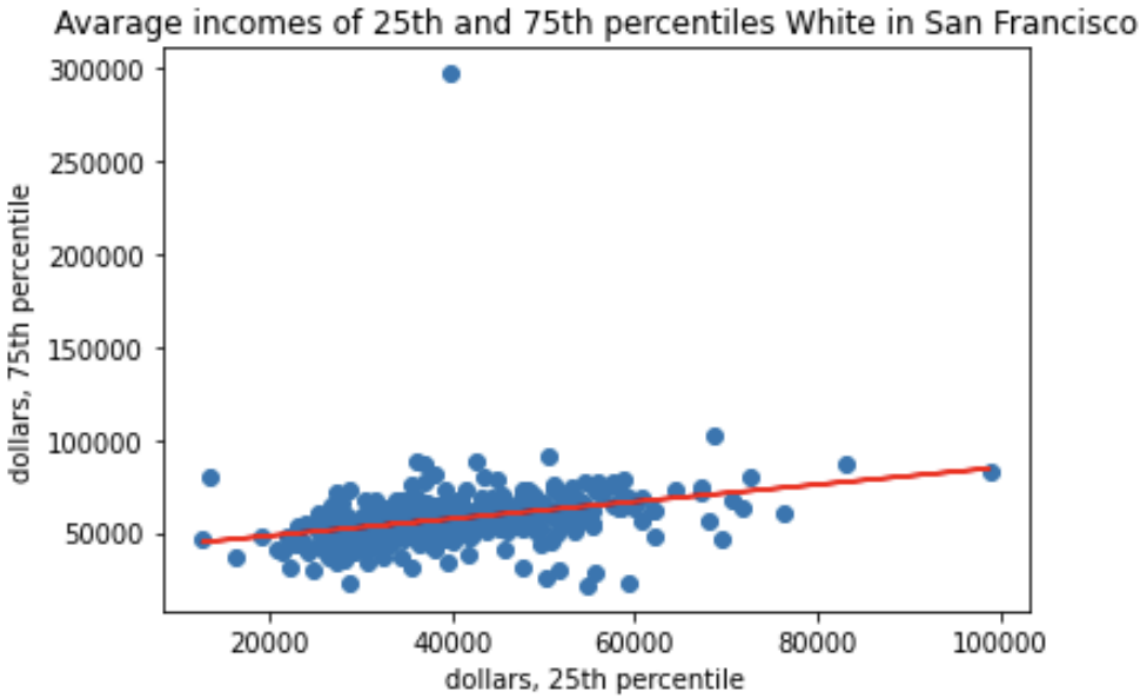
Correlation between income of 25th and 75th percentiles (Hispanic)




```
r_value = 0.0470139545409171
p_value = 0.5366934348383678
standard error = 0.11961191256181797
```

Absolutely no correlation here.

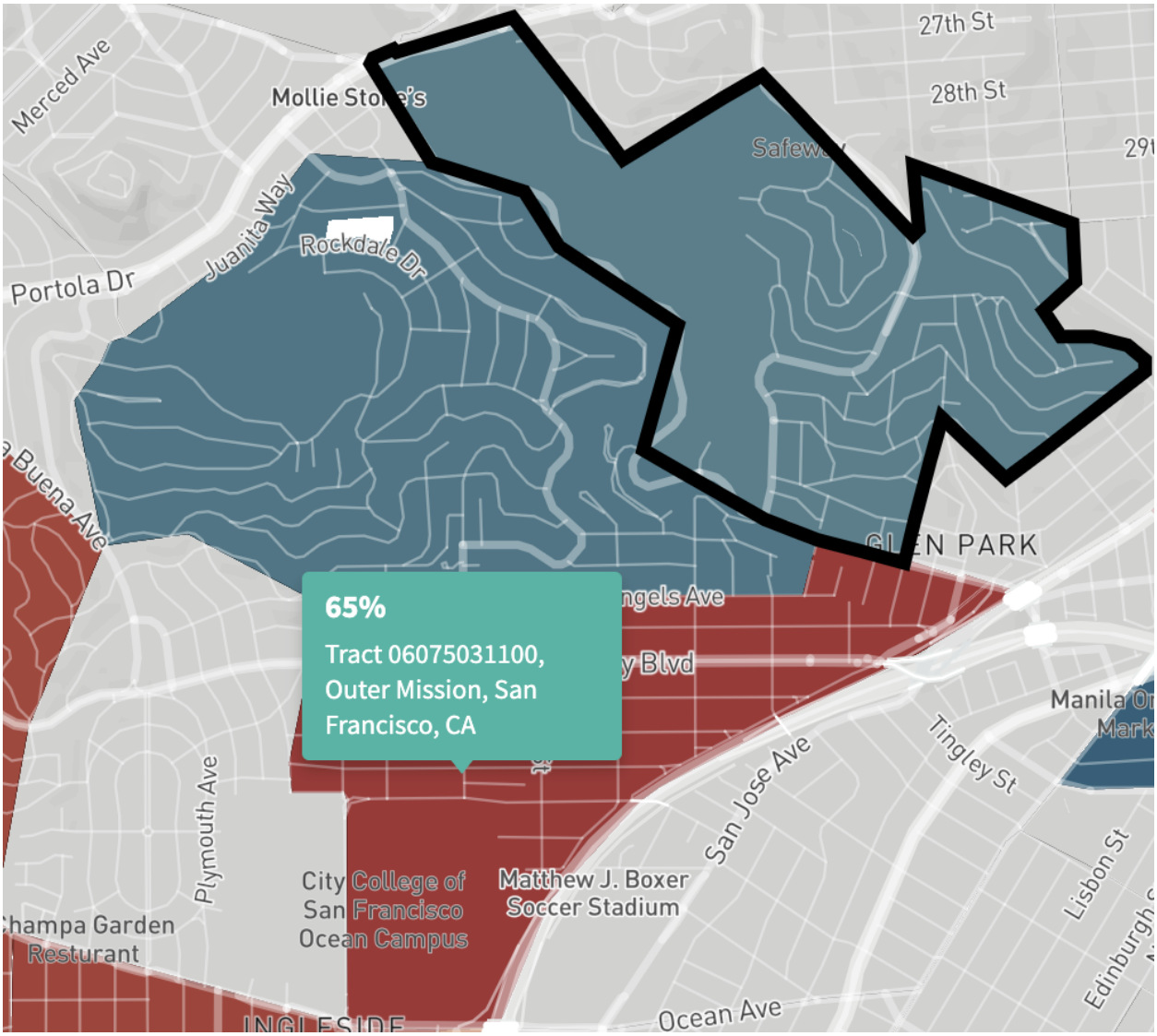
Correlation between income of 25th and 75th percentiles (White)



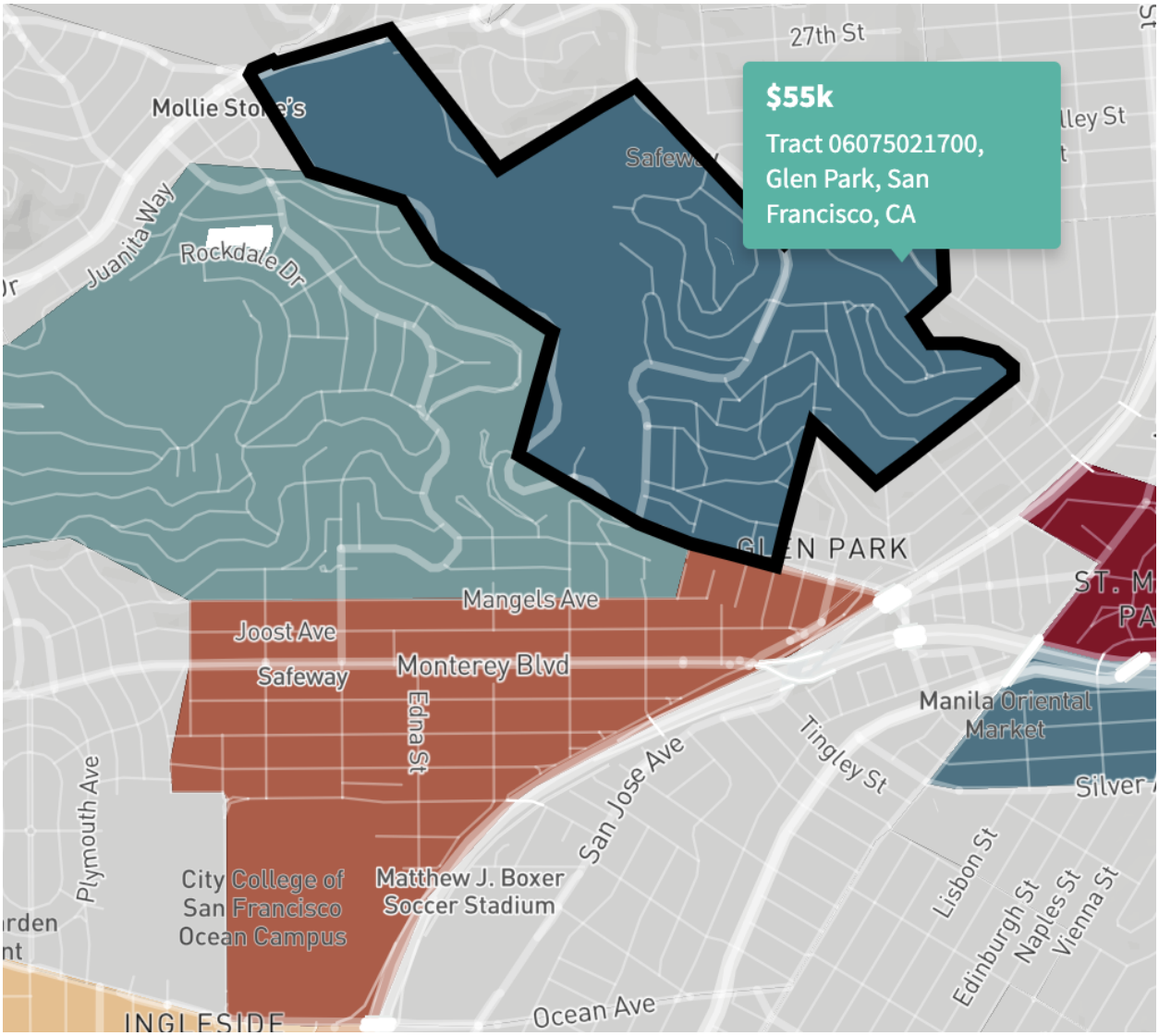

```
r_value = 0.31306902203782033
p_value = 3.290842952819629e-11
standard error = 0.0675957690741579
```

Not really strong correlation, but good p-value and low std. Correlation is not that bad here, and Linear Regression fits quite well here.

Interesting pattern for Black people in 2 close neighbourhoods from 100th percentile is employment rate. In tract with average income 130k employment rate = 93%, and in tract with income 7.3k - 63%. Meanwhile, median household income in red tract is bigger - 110k (in blue it's 100k). That looks really strange - only richest Black people have such a huge difference between these 2 close tracts.

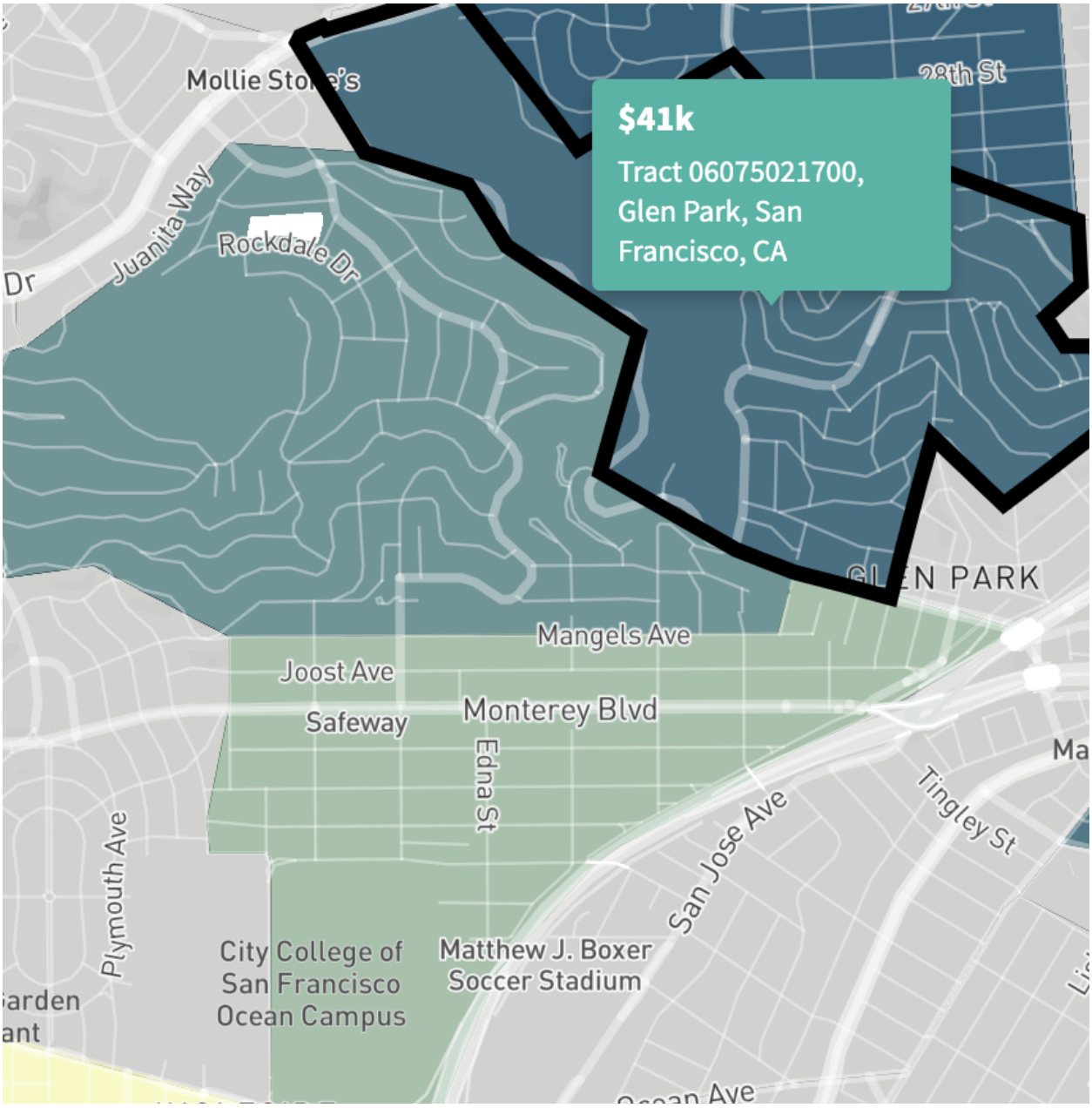


Also, if we look at 75th percentile in these 2 tracts, we'll see, that income of black people from "bad" tract is growing, and in "good" tract - growing.

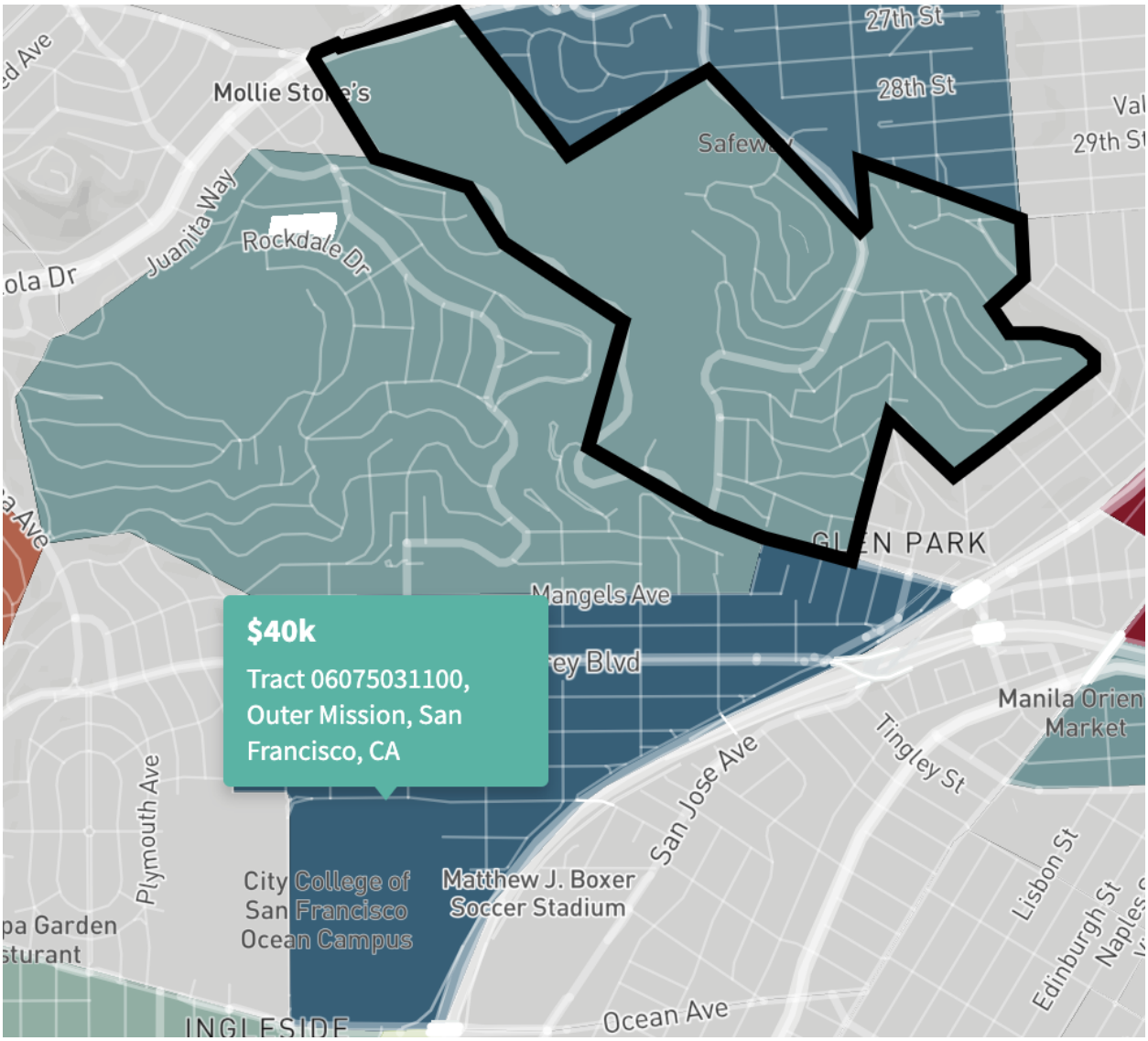


And we can see same tendention for 50th, 25th, and 1th percentiles! Difference between them is getting lower, and for 25th percentile and lower “bad tract” shows better income level!

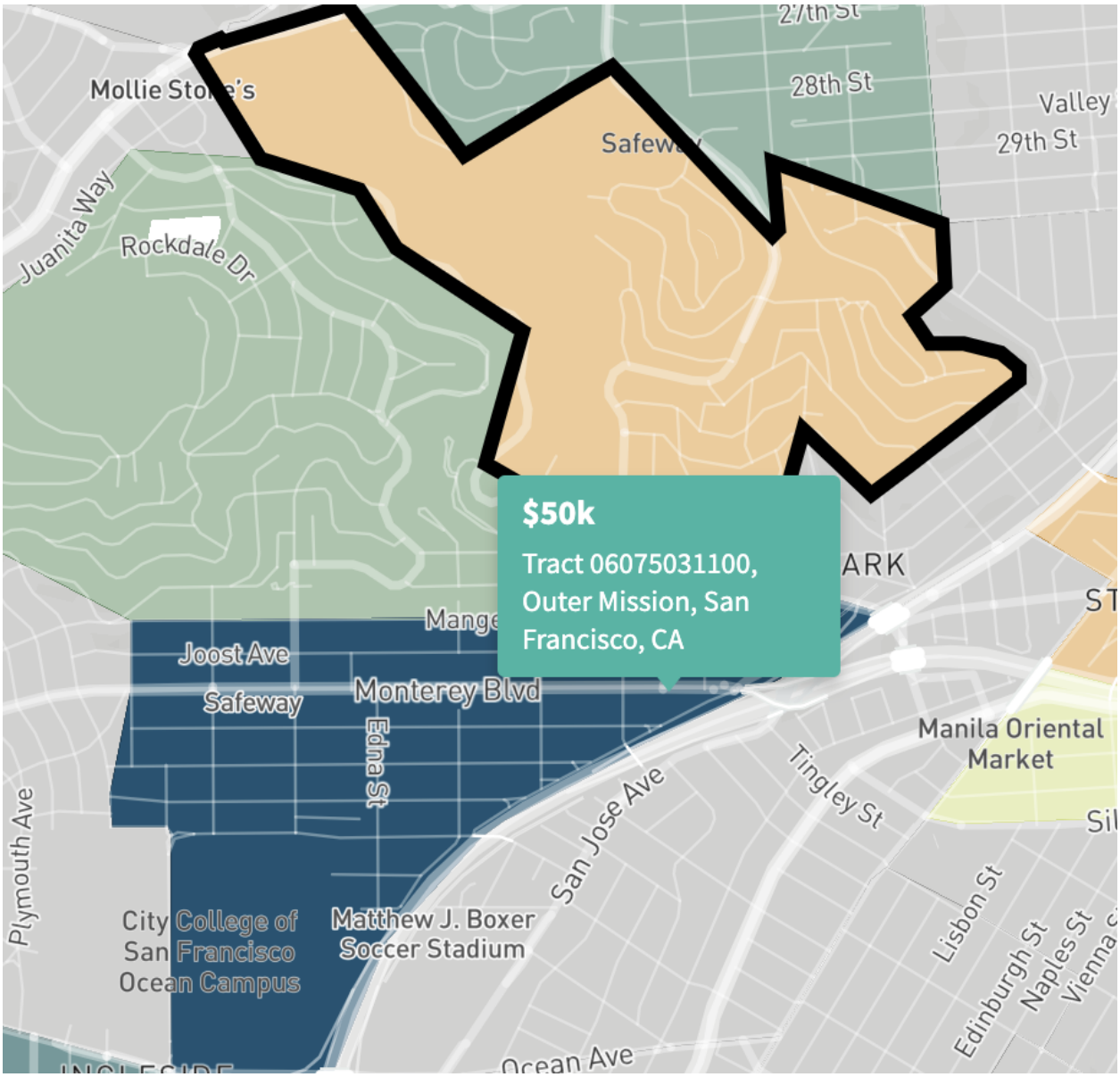
50th percentile



25th percentile



1th percentile



It's easy to notice, that there's a strong correlation between average income and employment rate, as well as for average income and incarceration rate.

I guess that improving these 2 factors is vitally important move to increase the overall income of tracts, especially for low-income families. Other question - that improving these 2 factors is not that easy, and it needs further investigation and analysis. Also, worth mentioning that it's quite hard to understand a difference between such neighborhoods for non-US residents, and I still can't explain some of the patterns I've found.

Moskovets Artem, SMD-2