Regresja liniowa – predykcja ceny majątku (z punktu widzenia urządu skarbowego)

**Ładowanie potrzebnych danych**

Z początku musimy wybrać z wybranego datasetu potrzebne pola potrzebne do analizy wartości domu, zostały wybrane następne pola:

|  |  |
| --- | --- |
| **Year Property Built** | Rok budowy majątku |
| **Number of Bathrooms** | Liczba łazienek |
| **Number of Bedrooms** | Liczba sypialń |
| **Number of Rooms** | Liczba pokojów |
| **Number of Stories** | Liczba piętra |
| **Number of Units** | Liczba budynków na terenie |
| **Property Area** | Pole budynku |
| **Basement Area** | Pole piwnicy |
| **Lot Area** | Pole terytorium |

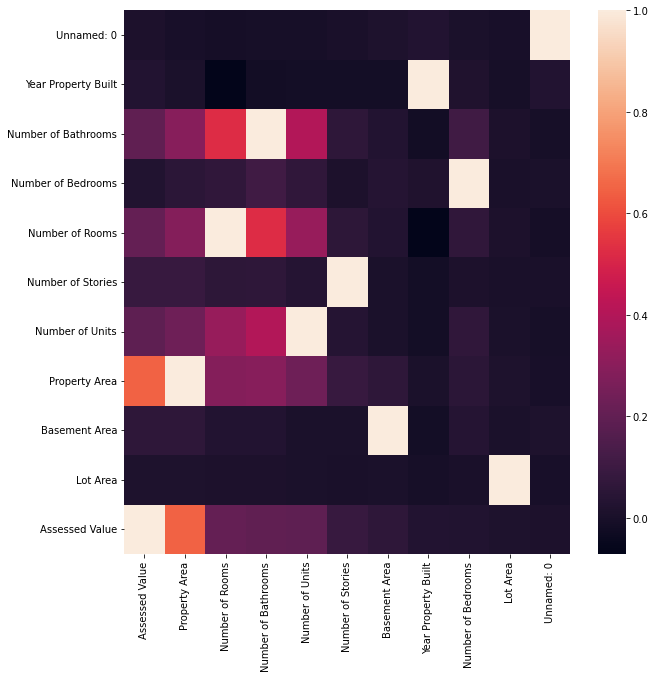
Następnie musimy pobrać sam koszt majątku, ale nie jest ono podane wprost, wartość jest rozdzielona wśród pol: *Assessed Fixtures Value, Assessed Improvement Value, Assessed Land Value, Assessed Personal Property Value*. „Assesed value” to pewna cześć kosztu rynkowej, która jest wykorzystywana dla liczenia podatku, zwykle jest równa 80-90%. Niestety nie zdążyłem znaleźć tej wartości dla San Francisco.

Formuła dla całkowitego kosztu majątku będzie:

W taki sposób przekształcamy dane, wyrzucała blednę dane, w danym przypadku wszystkie wartości mniejsze od zera i wartości równe zero dla kosztu.

**Analiza korelacji danych**

Następnie zbadamy jak wybrane pola korelują z wartością, wykorzystamy dla tego pandas.corr dla wyliczenia i sns.heatmap dla reprezentacji. Dodatkowo korelacji zostały posortowane dla Assessed Value. To daje następny rezultat:



Możemy zobaczyć, ze samym związanym z ceną parametrem jest pole majątku, potem idzie liczba pokojów, łazienek i budynków. Liczba pięter i pole piwnicy mają małe powiazanie z ceną, a pozostałe wartości prawie niezwiązane. Dlatego do dalszej pracy zostaną wybrane wymienione atrybuty.

**Normalizacja danych**

Dla bardziej efektywnego nauczania modelu musimy znormalizować dane według wzoru

Gdzie to średnia arytmetyczna, a to odchylenie standardowe.

**Budowa modelu**