**Ids Project Report**

**Intro-**

Our data set is based on Population densities of 300 cities across USA based on Census from 2010 and 2016.The aim of project is to find the interconnection and correlation between various factors such as overall population, % change in population and population density and also to find how each one of them depends on several other factors.

**Data cleaning-**

1)In the % change in population column, all the cities for which the % change was negative was represented by \_(underscore)symbol rather then -ve sign.So by iterating through that particular column, we checked if \_(underscore) was present in that row and replaced it with -ve sign if that was the case.

2)In the Land area column in the data set, each numeric value was followed by units in square miles. We are only interested in relative values of items wrt to each other in the data set and the not the absolute values, the units don’t play a major role and therefore not required.

3)We have also converted the strings in Census, RecentCensus and Percentage\_change columns into integers since some of data visualization methods cannot be directly applied on to strings data type.

**Questions,Inferences and Data Visualistions -**

1)Corelation between % change in population and total population in the recent census-

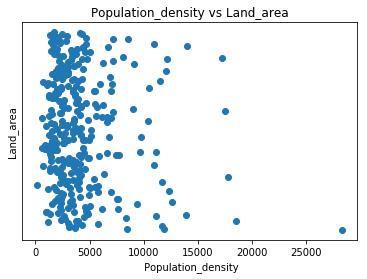
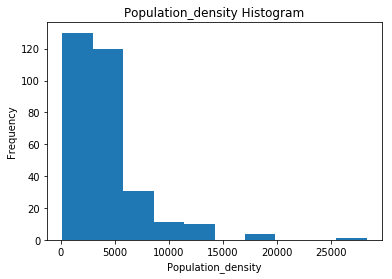
|  |  |  |  |
| --- | --- | --- | --- |
| Rank | City | RecentCensus | Change |
| 1 | New York | 8,537,673 | 4.43% |
| 2 | Los Angeles | 3,976,322 | 4.84% |
| 3 | Chicago | 2,704,958 | 0.35% |
| 4 | Houston | 2,303,482 | 9.68% |
| 5 | Pheonix | 1,615,017 | 11.72% |
| 6 | Philadelphia | 1,567,872 | 2.74% |
| 7 | San Antonio | 1,492,510 | 12.44% |
| 8 | San Diego | 1,406,630 | 7.59% |
| 9 | Dallas | 1,317,929 | 10.03% |
| 10 | San Jose | 1,025,350 | 8.39% |
| 11 | Austin | 947,890 | 19.93% |
| 12 | Jacksonville | 880,619 | 7.16% |
| 13 | San Francisco | 870,887 | 8.15% |
| 14 | Columbus | 860,090 | 9.28% |
| 15 | Indianapolis | 855,164 | 4.23% |
| 16 | Fort Worth | 854,113 | 15.23% |
| 17 | Charlotte | 842,051 | 15.12% |
| 18 | Seattle | 704,352 | 15.72% |
| 19 | Denver | 693,060 | 15.48% |
| 20 | El Paso | 683,080 | 5.23% |

|  |  |  |  |
| --- | --- | --- | --- |
| Rank | City | RecentCensus | Change |
| 157 | Frisco | 163,656 | 39.89% |
| 168 | Macon | 152,555 | 67.00% |

Using these 2 tables we can see that among the top 20 populous cities not a single city has -ve % change in population. All these cities have % change less than 20% while there are 2 cities (in table 2) which have % change in population 39.89% and 67% respectively. Using the previous points we can conclude that % change in population can be really high only when the total population itself is relatively low.

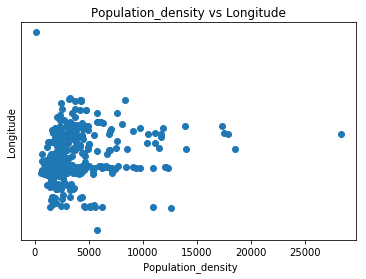
2)Factors that make New York the most populated city in USA

Population is basically the product of population density and land area. As we can observe from scatter plot, New York is the city with highest Population density (>25000 units), even though land area for New York city is low the population density is so high that it almost makes land area irrelevant.



3)Why Anchorage’s population density so low?

Anchorage is a city in Alaska whose Longitude is around 60(outlier as we can see from the Longitude Histogram), As we can see from the scatter plot Anchorage has very high Longitude but very low Population density. Since its Longitude is 60, it is much closer to the arctic circle than the other cities. So because of extreme cold climate Anchorage’s population density is really low as compared to other cities.



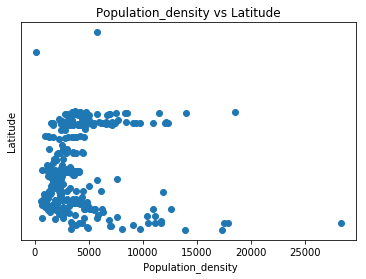
4)Reasons for range of values of % Change in population being so wide

In the whole data set a very high % of cities have +ve rate of change in population and very few have -ve rate, so we conclude that overall the population of all this cities combined has increased. However if we consider each city as a separate entity, Range of values of % change in population is vast. As already pointed out, some cities have very high increase in the rate of population (example- 67% in Macon). This high increase is only when only the total population of that city is relatively small as compared to other cities. On other hand of the spectrum there are some cities which have -ve rate of change in population. The reasons for -ve values can be natural calamities like Floods, Earthquakes, increase in crime rates, extreme climate and also Emigration.

5)From East to west how does population density vary?

As we can observe from the scatter plot there are 3 outliers ie-1 city in the east with high population density (NY) and 2 other cities with low population density in the west.

If we consider the main body of the scatter plot excluding those 3 outliers it almost a C type shape. So basically cities having high or low value of latitude have relatively higher population density. We can conclude that cities which either extreme east or west have higher population density, ie - cities with high population density lie in the coastal areas.



**Conclusions/inferences-**

1)% change in Population can only be really high only when total population is relatively low. All the top cities in terms of overall population have +ve % change in population.

2)New York is the most populated city in the states due to its really high population density which compensates for the low land area.

3)Anchorage’s population density is really low because of high longitude it has extremely cold weather.

4)Range of values of % change of population is really high/wide.

5)US cities lying in the coastal areas generally have higher population density than those lying in the central region.