Census Data Review:

Govt. of India Census, 2001 District-Wise

Jake Clifton, Brandon Damore, Laura Bullard,
Matthew O'Connor, and Kevin Eugene

Questions about the data:

1. Growth Rate - What are the determinants of growth rate? Can it be accurately predicted? What is the relationship between the growth rate of a district and important demographic and infrastructural characteristics?
2. Religion- Which religion is most common in each State? How do the religions differ between high and low populated states?
3. Literacy- Do more literate areas have more access to amenities than less literate ones?
4. Does water availability and quality change in rural areas?
5. How does level of education relate to the quality of housing in each district?

Data Gathering and Cleaning

- We found a great dataset that was pretty clean on Kaggle.com
- Specifically… https://www.kaggle.com/bazuka/census2001

When cleaning data…

- Chose columns we would use
- Created new columns based on the analysis we wanted
- Removed the NAs
- (Refer to our jupyter notebook)

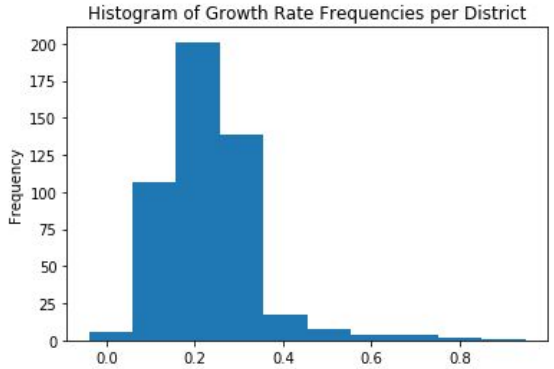
Growth Rate: Data Distribution and Summary

Summary Statistics:

Count	489.0
Mean	0.233
Std. Dev.	0.113
Minimum	-0.04
25th Percentile	0.16
Median	0.23
75th Percentile	0.28
Maximum	0.95

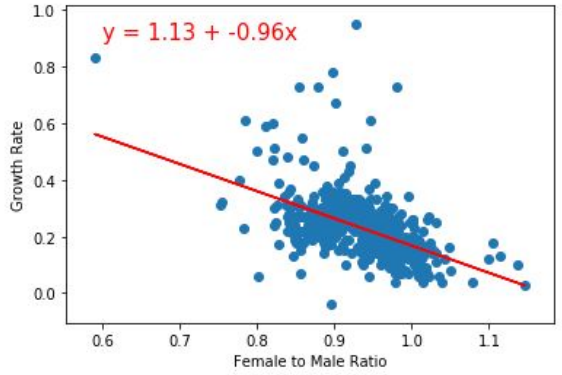
Growth Rate: Data Distribution and Summary

Histogram of Growth Rate frequencies:



Growth Rate (y) vs. Sex Ratio (x)

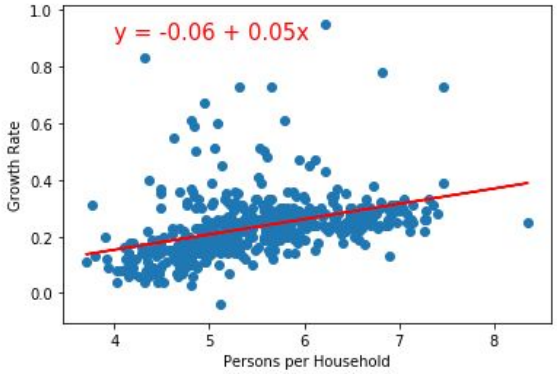
Linear Regression:



- The correlation coefficient between Female to Male Ratio and Growth Rate is -0.51.
- The r-squared of the regression is 0.26.
- The slope of the of our independent variable, Sex Ratio, implies that a 1 unit increase in the Sex Ratio will decrease our dependent variable, Growth Rate, by 0.96.

Growth Rate (y) vs. Persons per Household (x)

Linear Regression:

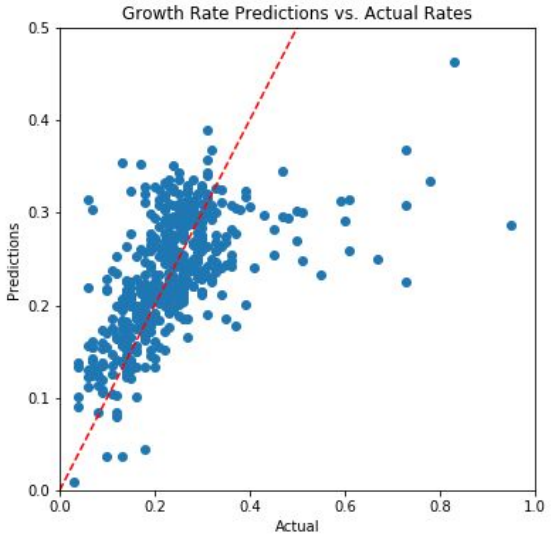


- The correlation coefficient between Persons per Household and Growth Rate is 0.39.
- The r-squared of the regression is 0.15.
- The slope of the of our independent variable, Persons per Household, implies that a 1 unit increase in *x* will increase our dependent variable, Growth Rate, by 0.05.

Predicting Growth Rate using Multiple Covariates

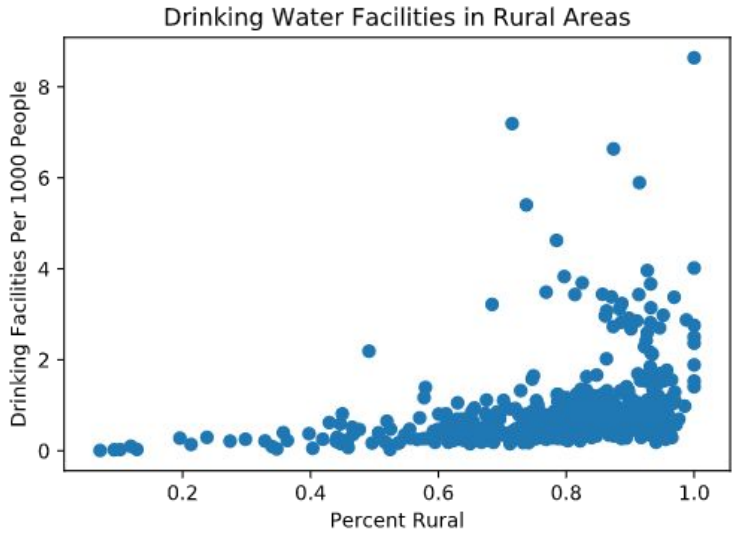
Equation:

- Growth Rate = β0 + β1*Persons + β2*Female to Male Ratio + β3*Persons per Household + β4*Literacy Rate + β5*Permanent Housing Rate



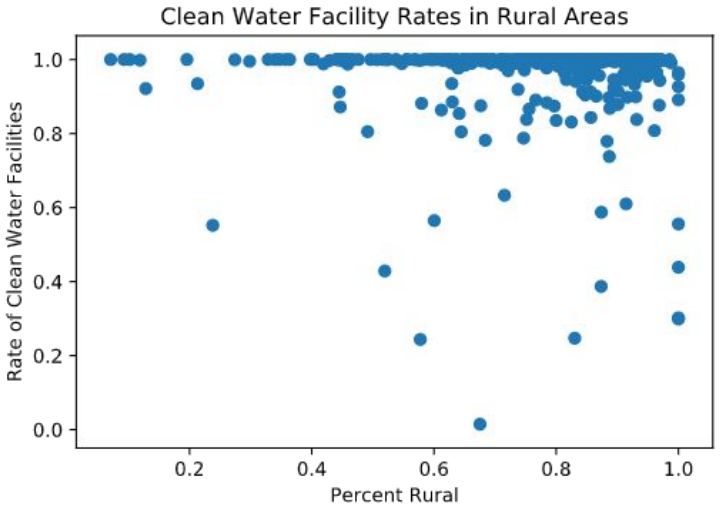
Water Quality and Availability in Rural Areas

- Water facility availability per 1000 people dramatically increases in areas over 75% rural.
- Likely due to households having their own water source, very small towns, etc.



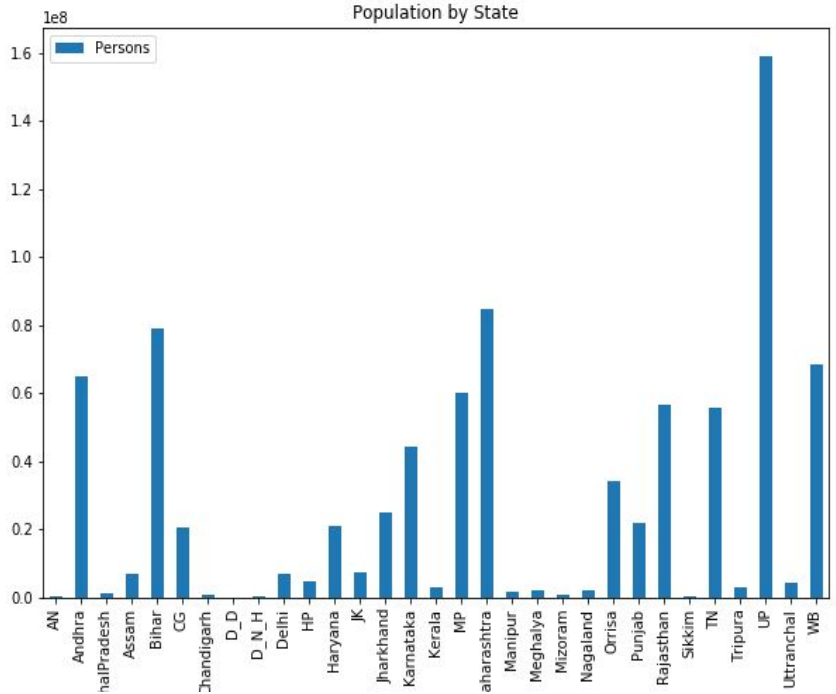
Water Quality and Availability in Rural Areas

- Urban water facilities are more likely to produce safe drinking water than those in rural areas.
- For the most part, regardless of the area, residents are likely to have access to clean drinking water.



Population by State

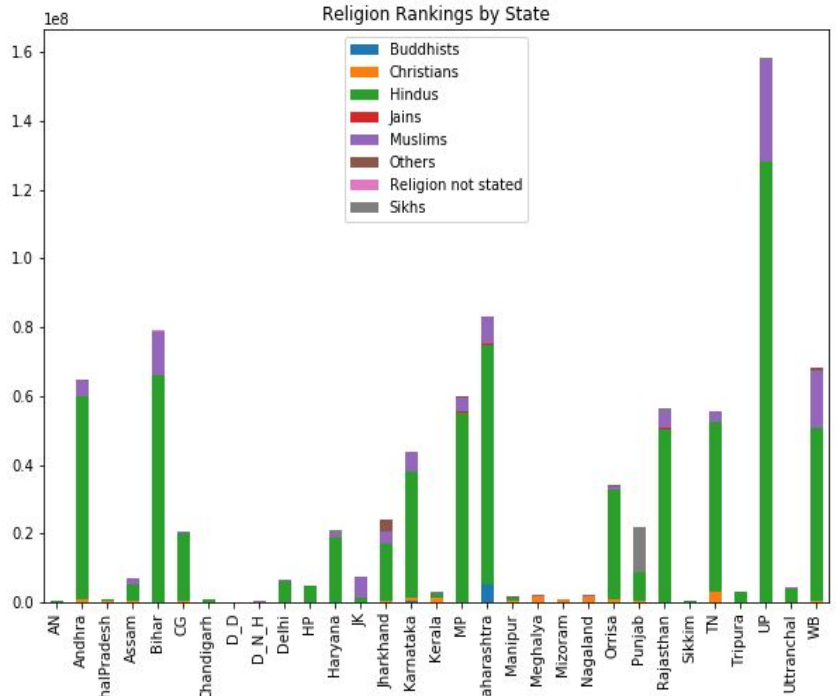
As a country, the population and size of each state vary widely.



Which religions are most common in each state?

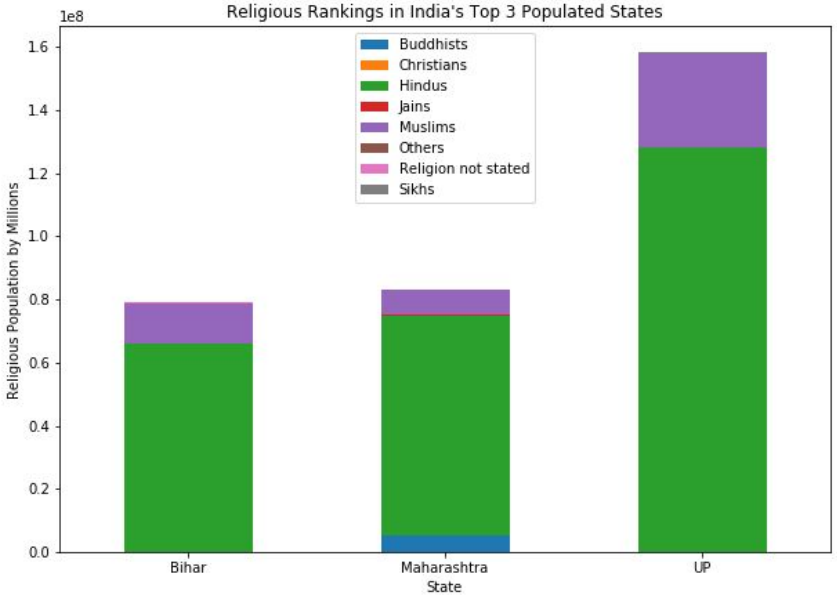
At a glance, it would appear that Hindu & Islam are the most popular religions in most states, no matter population.

As we look at religion by state, we wonder if the most prominent religions are mainly grouped in the highest populated states, or if the less-populated states follow suite.



Religion rankings in the 3 highest populated states

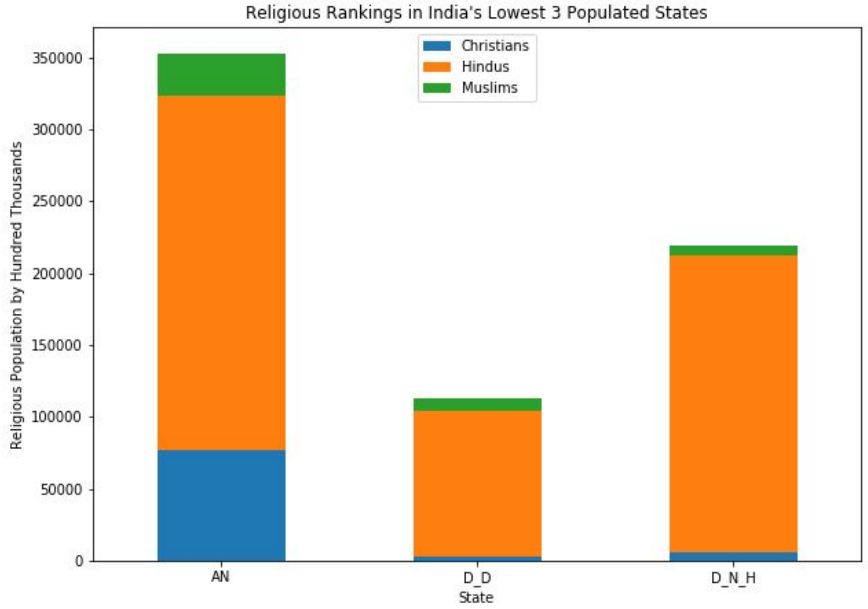
In the three highest populated states, Hindu & Islam fall in line with being the most popular religions, with Buddhism falling in third.



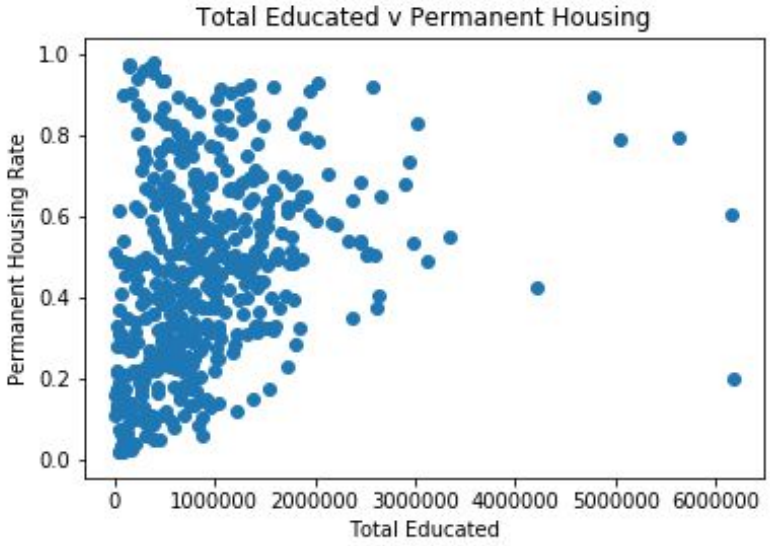
Religion rankings in the 3 lowest populated states

While the two most common religions are Hindu & Islam in the least populated states, we do see Christianity peeking in, making up 13% of the total counted religions. To compare, Christianity only accounted for .02% of religion in the 3 highest populated states.

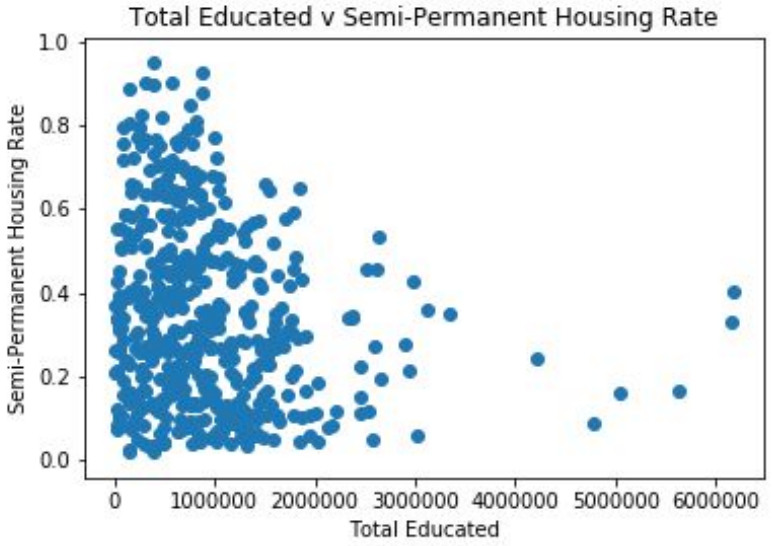
This would indicate that Christians are more likely to either live in small/lower populated areas, or are more likely to openly submit their religion, or both.



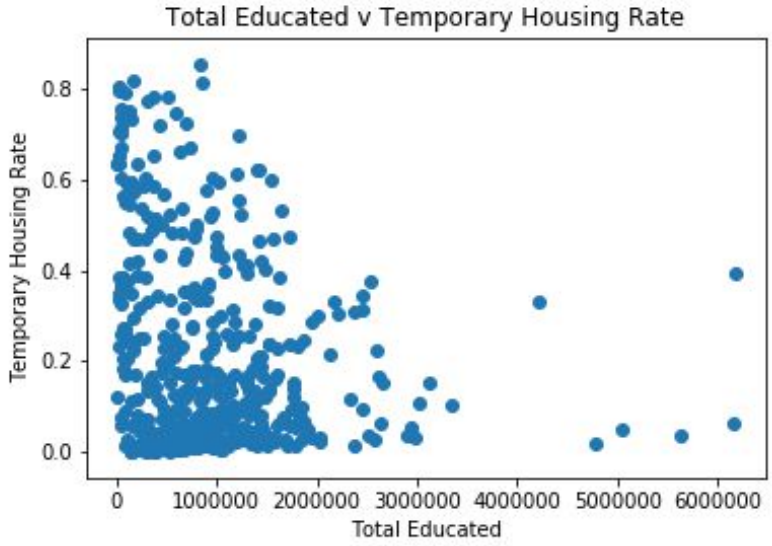
Total Educated vs Permanent Housing



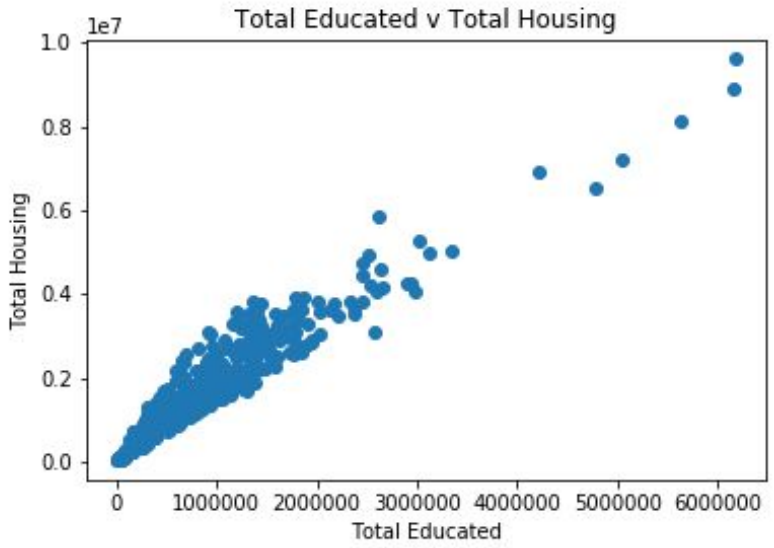
Total Educated vs Semi-Permanent Housing



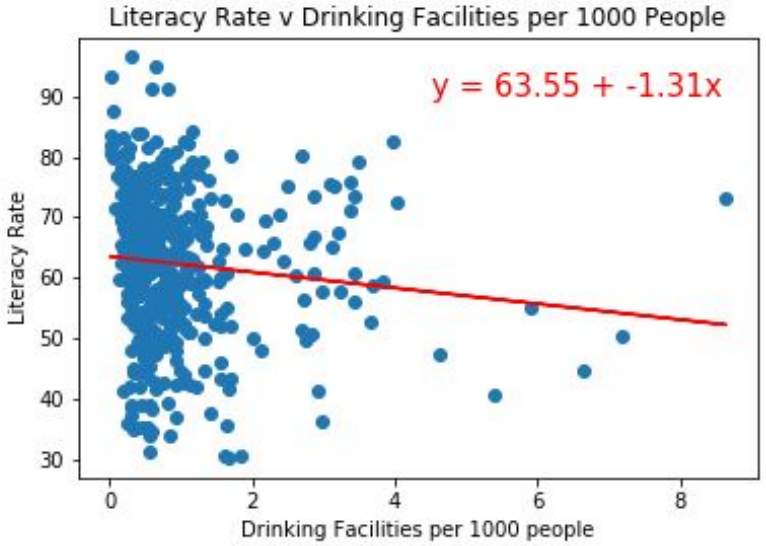
Total Educated vs Temporary Housing



Total Educated vs Total Housing

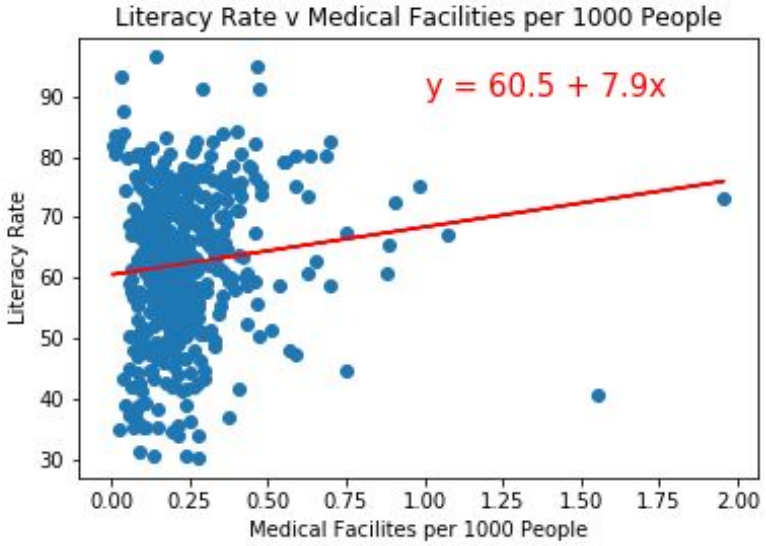


Literacy Rate and Drinking Facilities



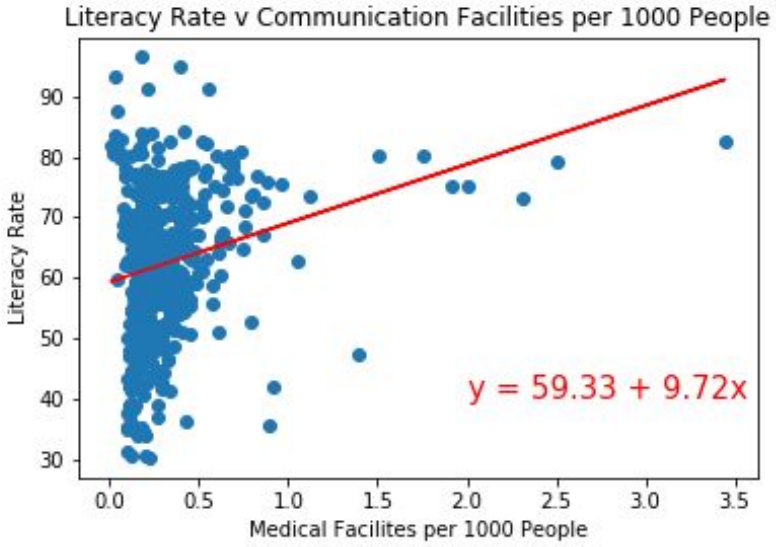
- There is hardly any correlation between the two.
- The r-squared value is -0.1

Literacy Rate and Medical Facilities



- There is not much correlation between the literacy rate and the availability of medical facilities.
- The r-squared value is 0.11

Literacy Rate and Communication Facilities



- While there seems to be more of a correlation between the literacy rate and the number of communication facilities, it is still very low.
- The r-squared value is 0.24

Conclusions

- There is a negative correlation between growth and the ratio of females to males.
- Regardless of population, the top religions are Hindu and Islam.
- The literacy rate of a population has nothing to do with the access to amenities.
- Rural areas are likely to have more water facilities per 1000 people, however urban areas are more likely to have clean water sources. However, citizens are likely to have adequate access to clean drinking water regardless if they live in urban or rural areas.