PES University, Bangalore
(Established under Karnataka Act No. 16 of 2013)

UE17CS203

B.Tech, Sem III
Session : Aug-Dec, 2018

UE17CS203 – INTRODUCTION TO DATA SCIENCE

REPORT

EXPLORATORY ANALYSIS ON
Karnataka State Education

DATA SET LINK :	**https://www.kaggle.com/pavansanagapati/karnataka-state-education**
TEAM MEMBERS	**NAME : Shashidhar R** **SRN :PES1201701417** **EMAIL ID :shashir14399@gmail.com** **CONTACT NO. :9071039351**
	NAME : Anish Sekhar **SRN :PES1201700242** **EMAIL ID :pogchampoo24@gmail.com** **CONTACT NO. :9900206446**

ABSTRACT

This is the IDS assignment for picking up a random data set of suitable size from the internet and cleaning and analyzing the data. The kind of analysis performed is to plot graphs such as bar charts for various fields and concluding answers from it.

DATA SET

Our data set is the number of literate people in different districts/areas in Karnataka, with their being specifications such as literacy numbers for different age groups (such as 0-6,13,14,15 etc.) as well as for different levels of literacy(such as literate without education, diploma graduates.etc). It is from https://www.kaggle.com/pavansanagapati/karnataka-state-education .There are 812 rows and 46 columns (we do not use all the columns her, except to clean them). This is a large dataset to me, but it is probably tiny compared to the datasets companies and corporates analyze.

PROCESSING:

The only processing we have done is to impute mean values in columns where there are zeros instead of proper values. Mean imputation is a fair enough metric to replace the zeros, as we do not want to delete these rows entirely. Filling it in as the previous or next value is a fair enough method too, as all the adjacent values are from the same area. However, they may not be listed by proximity , and a really rural part of an area may be next to a more developed one, so we avoid this. We could have used regression as well, but coding mean values is just simpler. Outliers in the data set, we ignore. This is because our data set

is from a census and thus represents the population. It is not a sample, so we will not be removing outliers from this dataset.

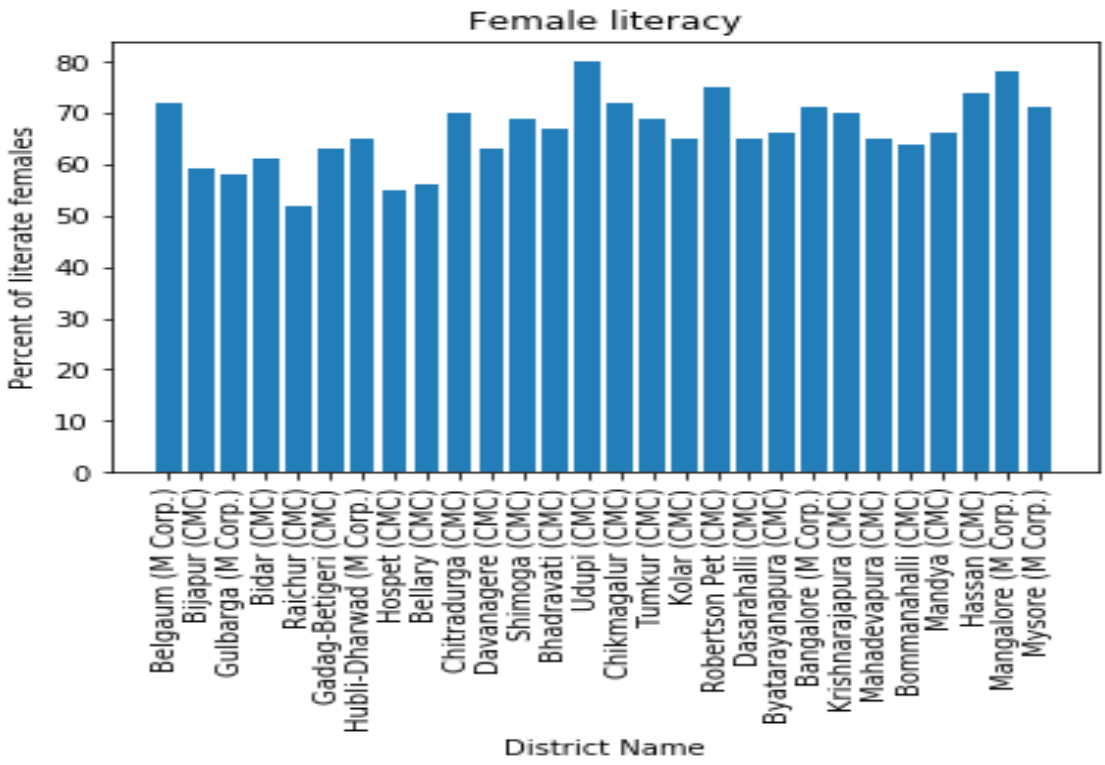
EXPLORATORY ANALYSIS:

We primarily use bar charts because:

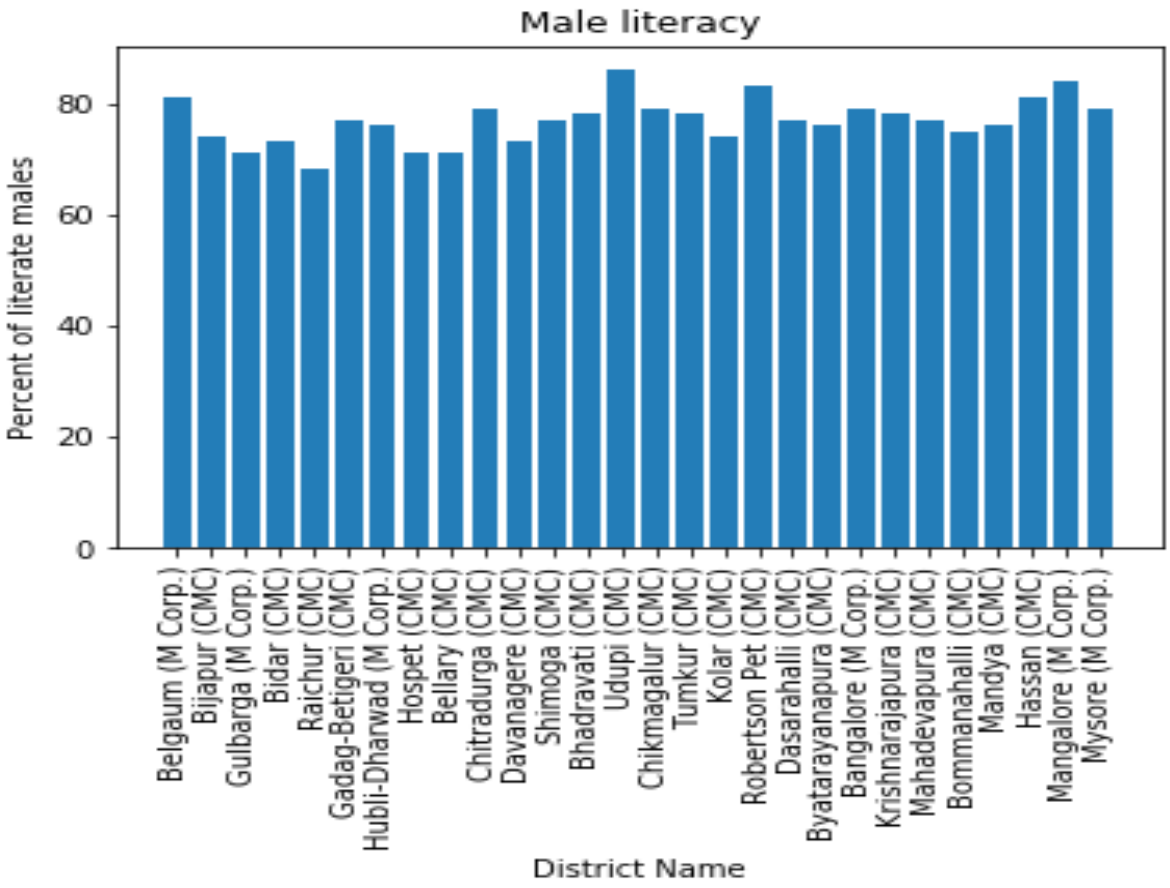
1. Not time based, no sequential events, so line graphs are useless.

2. Always more than 6 categories, so pi charts are also quite ineffective.

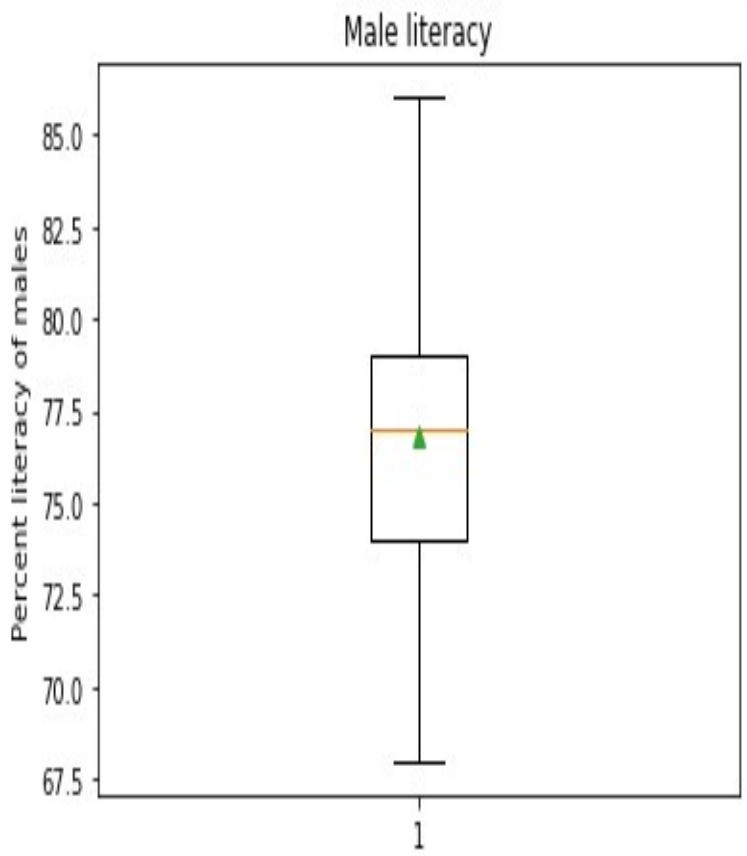
3. Scatter plots are also not useful, as there are no dependent variables, just the same variables categorized and for different regions.

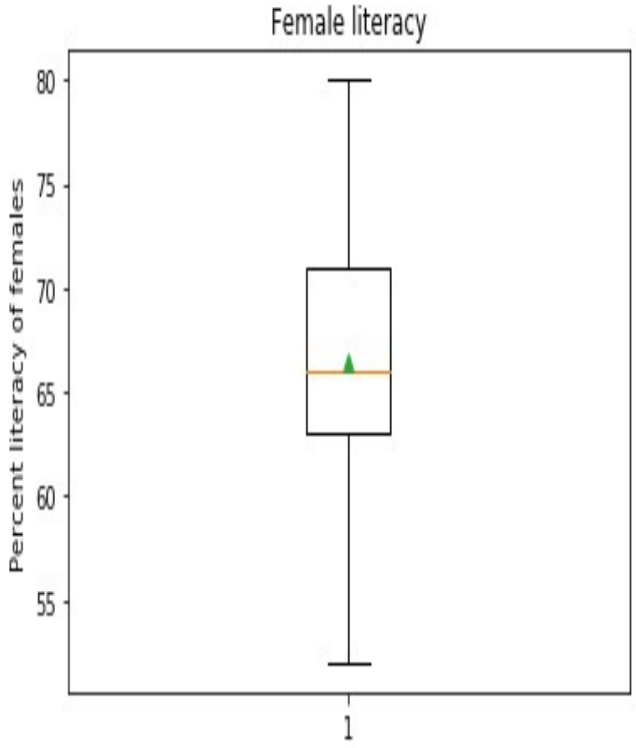


This bar chart shows the literacy rate of females of all age groups in various districts of Karnataka . From the graph it is evident that Udupi district has the highest female literacy,80% where as Raichur has the least with 52%. The average female literacy rate of Karnataka is 66%.So by looking at the graph we can conclude that most of the northern districts of Karnataka have a lesser female literacy than the mean.

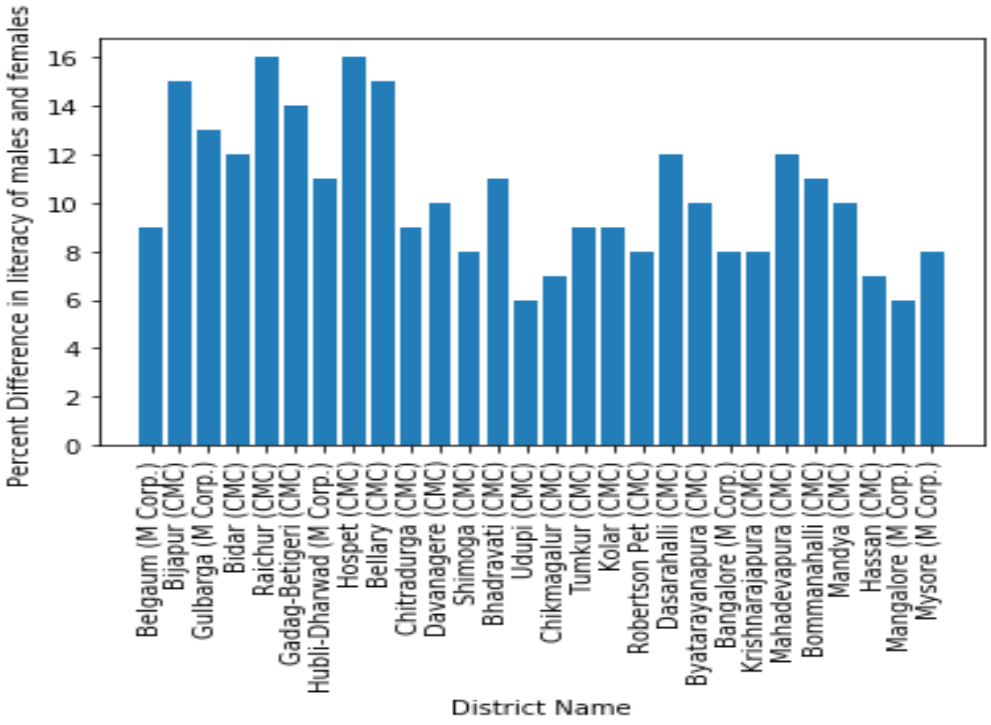


This bar chart shows the literacy rate of males of all age groups in various districts of Karnataka . From the graph it is evident that Udupi district has the highest male literacy,86% where as Raichur has the least with 68%. The average male literacy rate of Karnataka is 77%.So by looking at the graph we can conclude that most of the northern districts of Karnataka have a lesser male literacy than the mean.


Male literacy
Percent literacy of males
85.0
82.5
80.0
77.5
75.0
72.5
70.0
67.5
1

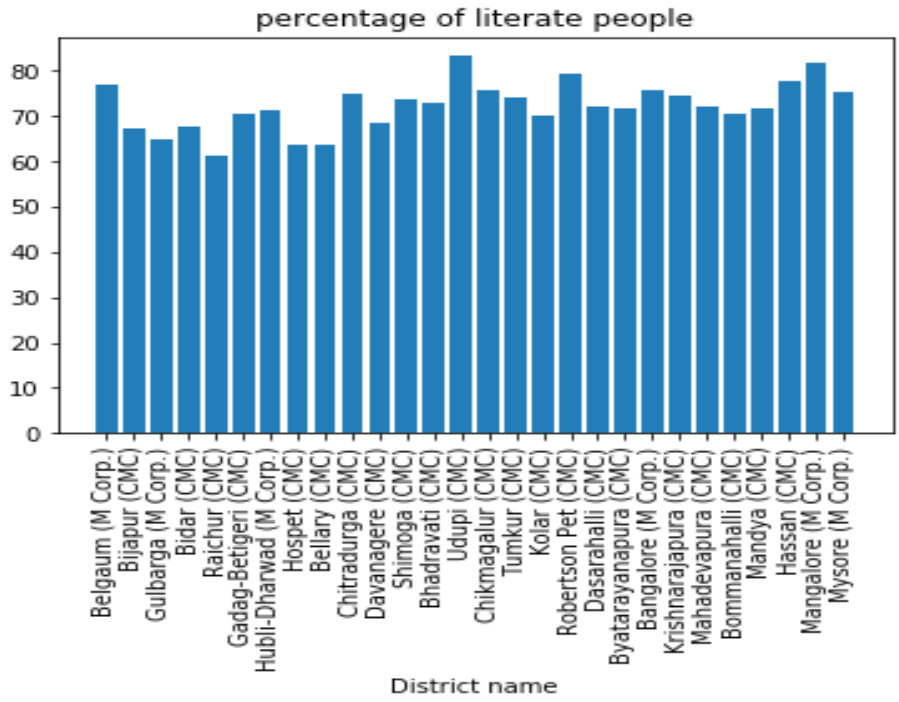


These are the box plots for percent literacy of males and females .By observing these two boxplots we can see that the mean and median of the male literacy of all the districts of Karnataka is greater than that of the female literacy. Also the upper whisker and lower whisker of the male literacy is greater than that of the female literacy. So we can conclude that males are more literate than females in Karnataka



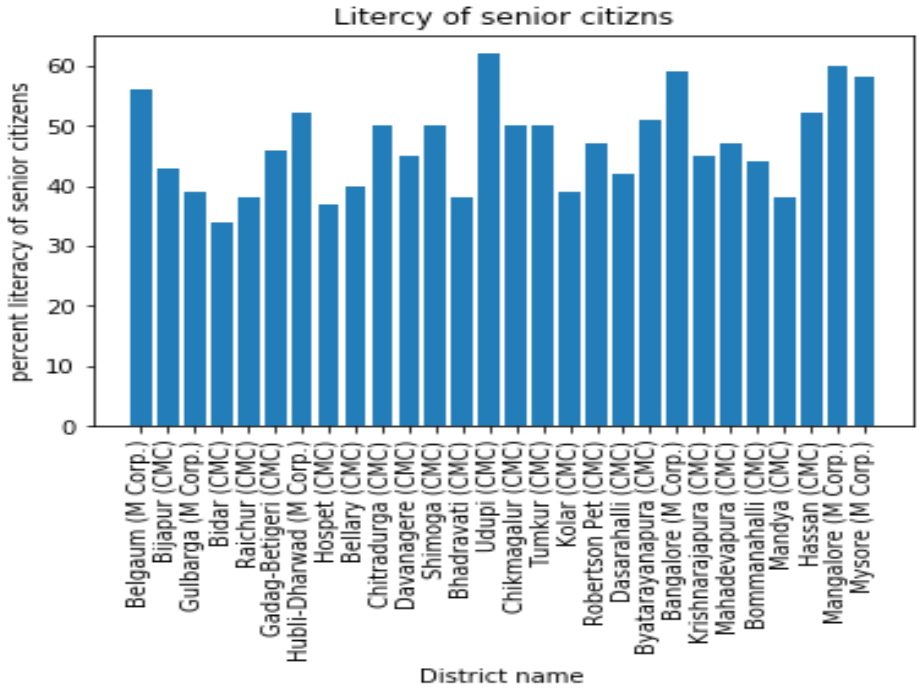
This bar chart shows the difference in the literacy rate of females and males in various districts of Karnataka . In all the districts the

percentage of male literates are more than that of the female literates . From the bar chart it is evident that Udupi has the least difference in male and female literacy rate whereas Raichur and Hospet have the highest difference



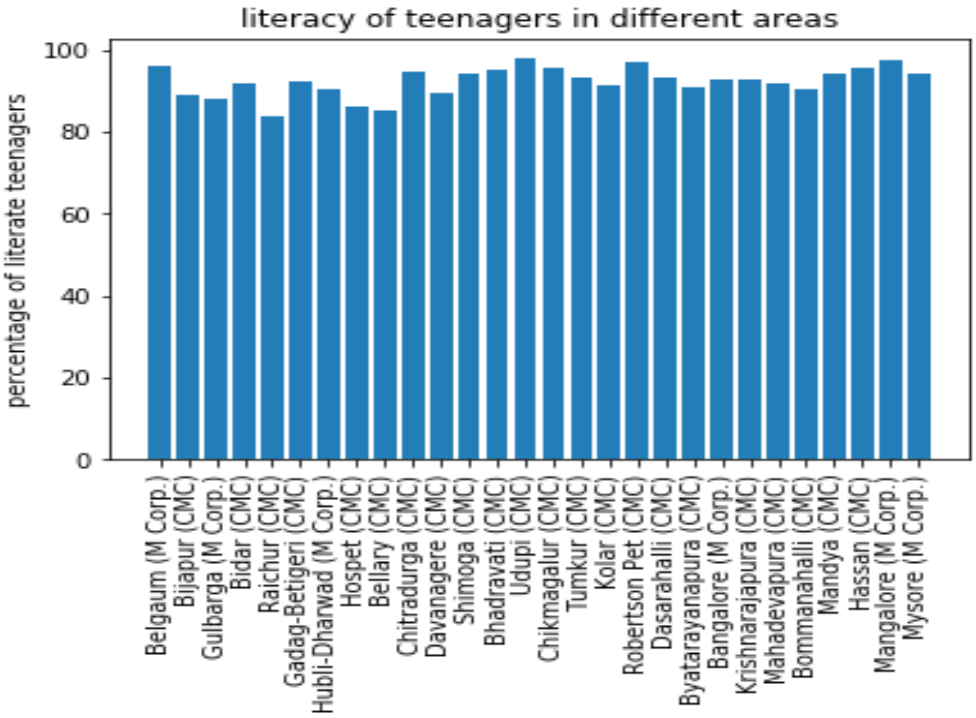
Literacy rates across different areas in Karnataka. As expected, more developed areas like Udupi, Bangalore,Mangalore,Mysore have

higher literacy rates as they are more developed. Bangalore is not the highest , despite being the only metropolitan city. This may be because of a lot of movement of unskilled labor form rural areas looking for work and going to Bangalore. Raichur is the lowest, being the least developed of all these areas.



As seen in the other graphs , bangalore has a pretty high literacy rate, being a metropolitan city and more advanced . However, it is not the

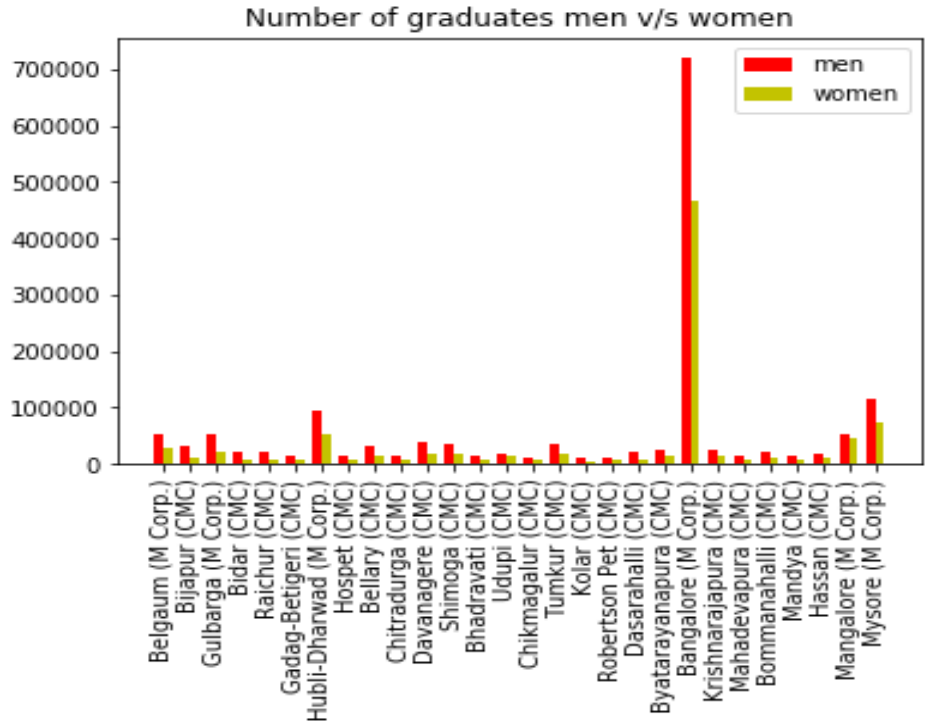
highest. This maybe because number of senior citizens in bangalore are higher than in other places. perhaps they are from rural areas , gone to bangalore to make a living in the city.
since many of these senior citizens are not necessarily from bangalore, this data indicates how progressive these areas were in terms of educating children many years ago, as well as how well developing industries in these areas has made these people choose to become literate



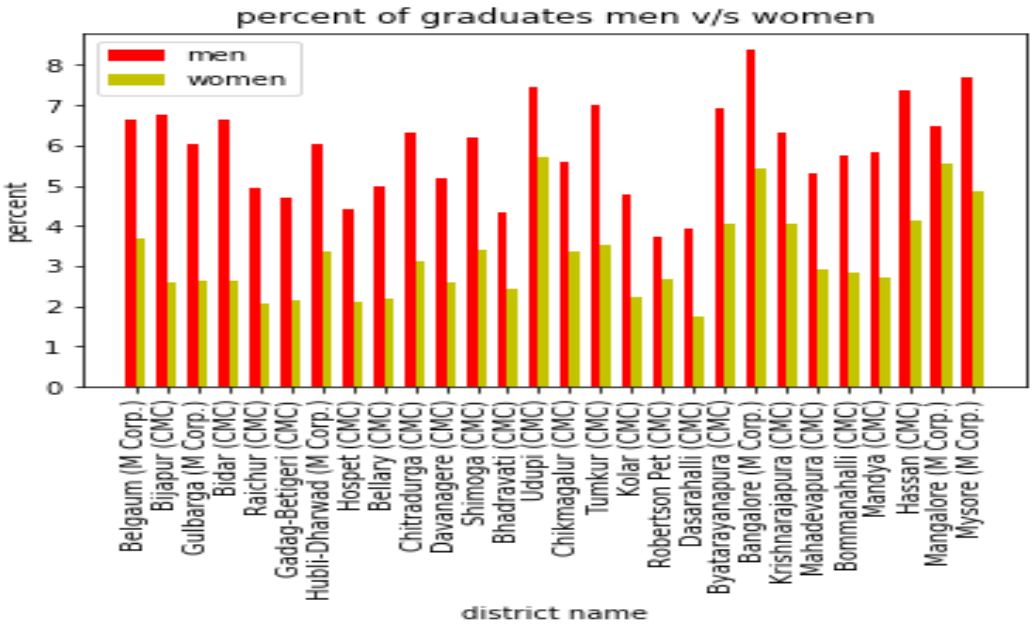
Our dataset consists only of urban areas, so as expected, most of the rates are high. Consistent with other graphs, Udupi is again the

highest in this.

Raichur is also the lowest, as is the case in most of the other graphs. This fits the general trend of toppers coming from Udupi, and Raichur is the most underdeveloped of these areas.



Here we see total number of graduates (NOT PERCENTAGE) and we see how difference in population gives us a false sense of inequality. Actually, this graph is like this because Bangalore's population is more.
In the next graph, which is a percentage, it shows the true difference between areas.



As expected, when taking percentage of total population, the number of graduate are pretty low. This is the highest level of education in our

dataset. From the previous slide, we see a huge difference. Now, the areas are much more close in values (though bangalore is still the highest).
We do see that in all the states, male graduates are more in number. Udupi again shows it is a great place, with high percentages AND a relatively lower difference in percentages between males and females.

CONCLUSION:

We can conclude that as expected, the more developed and advanced areas such as Bangalore not only score high in total numbers(because of more population) but also in percentages, suggesting that education is valued more highly here, or maybe just that the standard of living of people in such areas is such that families can afford to educate.

Takeaways from this exercise:

1. More familiarity with pandas, and how to use it to extract data from csvs

2. Ability to work with csv files greatly increased. Can extract information from such files easily

3. Matplotlib familiarity, where we now know how to plot various kinds of graphs given lists of data. (though in our case there are only box plots and bar charts)