DAV Technical Team Assignment

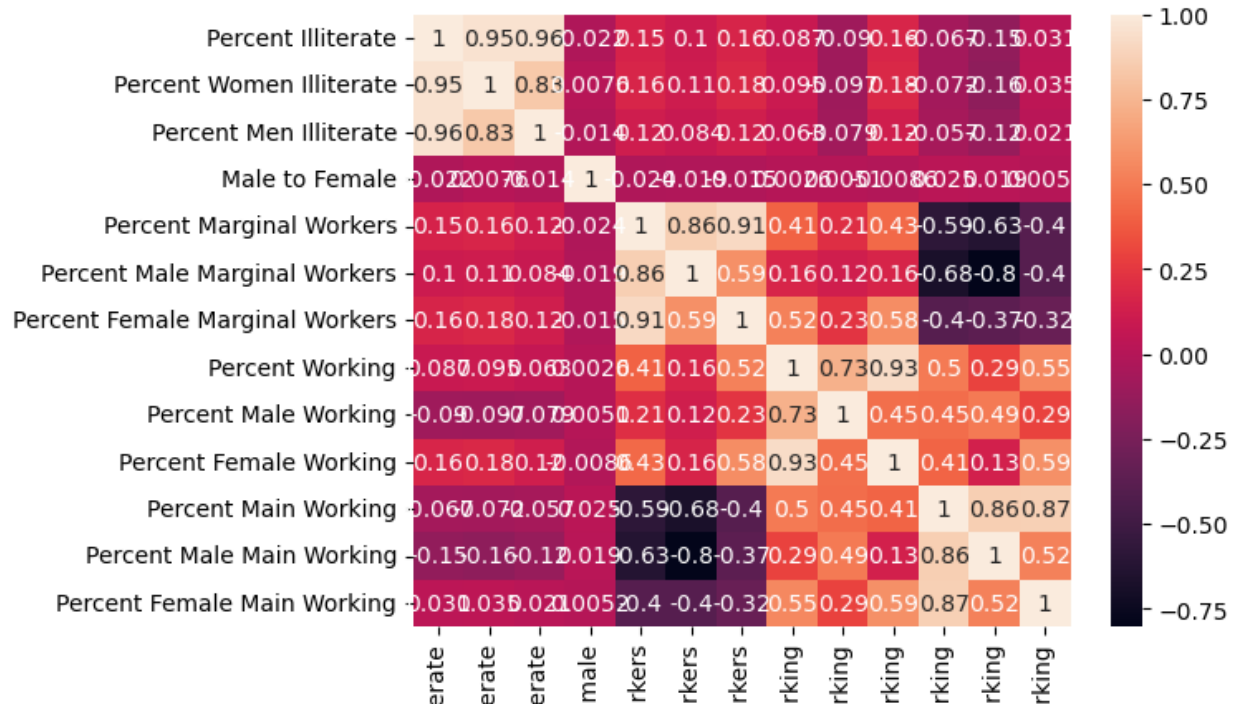
Pratik Sahoo
(22B0968)

Q1]

1.

First and foremost, we must understand that these data points are at a single time frame, so we cannot comment on time-based trends for anything. A very obvious start is to calculate the correlation between different features and see their relations. On calculating correlations and plotting a heatmap, however, we see that a huge amount of real-life variables seem to have a large correlation. Anyone with even a few days experience of data analysis knows that it is incredibly rare for a pair of real-world variables (that are not directly consequences of each other) to have correlations greater than 0.9, yet here we have it for many pairs of variables.

On some thought, I realized that this high correlation comes from a simple fact: big towns have a lot of everything, and small towns have a little of everything. To overcome this, we must deal with ratios or percentages instead of raw values. Thus, I created a new dataframe (correlation heatmap below) and a function to populate the new dataframe with ratios of columns of the original.



Question 1

Do literacy rates have any correlation with the gender ratio? Other factors?

Answer: From the correlation data (displayed in the Q1.ipynb file) , it appears that there is a significant correlation (not necessarily causation) between illiteracy and percentage of marginal workers in a society (especially female).
Also, illiteracy has a strong correlation with the percentage of females working, but not the percentage of females working main jobs.

Finally, illiteracy has a significant *negative* correlation with the percentage of men working main jobs.

Also, gender ratio has negligible correlation with literacy.

What this indicates is that communities with high literacy tend to have a higher percentage of men working full-time jobs, and a lower percentage of people working marginal jobs, especially women (and vice-versa for low literacy).

Question 2

What factors differentiate between male and female literacy rates?

I found the difference between the correlations of various features with male and female literacy rates. I understand that while this is not a rigorous method (A rigorous approach could include linear regression with the feature being studied as the target and male and female literacy amongst the features, because the optimal solution of a linear regression problem yields the eigenvectors of the linear transformation matrix, and different eigenspaces are orthogonal to each other, thus reducing the common effect of both variables. Comparing the coefficients of the solutions would better indicate the difference in how strongly they are related.), it serves as a rough measure of how much stronger a certain factor moves with one of these rather than the other.

Clearly, of the factors explored, the one differentiating male and female literacy the most is percentage of female marginal workers.
The high negative difference shows that the portion of female marginal workers tends to increase with female illiteracy far more than male illiteracy. This may be an indicating factor that females work marginal jobs due to their illiteracy more often than illiteracy prevalent in society; however, we must be careful in our assumptions since correlation does not imply causation.

Once again, the male-to-female ratio is not a large contributing factor.

An interesting thing to notice is that the percentage of men working main jobs has a significantly higher correlation with the difference we are observing than percent female main workers, or percent male marginal workers. This may suggest that men working full-time jobs encourage female literacy more than their counterparts in marginal jobs, and may hold more ability to enforce it than women working full-time jobs. While this seems like a very sound and plausible argument, since we have not been given any time-dependent data, we can only draw correlations between different factors; we cannot, in any way, conclude a causal theory.

Question 3

Does the presence of SC population have any correlation with literacy rates? Other observations?

There does not seem to be a significant correlation of SC population with most factors we have already enlisted.

There seems to be a significant *negative* correlation between percent SC population and percent female working. One can draw many causal theories from this data, however, like stated above, one cannot state any of these with any surety based on the given data.

We must also note that percent SC population has a low negative correlation with illiteracy (total, male and female). While this cannot be neglected and can be used as evidence to support other theories, the correlation is too low to formulate a theory based solely on it.

Question 4

Does the presence of ST population have any correlation with working or literacy rates?

Unlike the correlations with the SC population we say above, we see that the ST population of an area has high correlations with a lot of employment and literacy parameters. This may be a result of ST people being alienated by society more, or them living a different lifestyle from the ordinary person, or neither. It might also be a statistical anomaly, born out of how the population is distributed, with no real-life basis. All we can say for sure is that an area having a large portion of ST population is likely to have a different demographic status than the average. Below, I will try to summarize the strongest takeaways.

Firstly, ST population has a surprisingly strong positive correlation with illiteracy (across gender). Again, based on the given data, we cannot comment on whether they have been denied access to education, or do not treat it as a priority themselves, or any other theory.

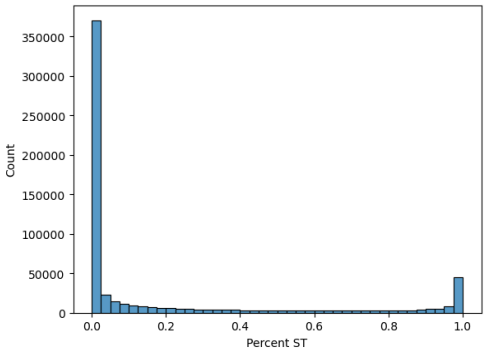
Secondly, there is a strong correlation between percent of ST population and percent of people working marginal jobs (all genders), and percent women working full-time jobs. Across all jobs, the percent of ST population tends to increase alongside female employment quite strongly, whereas there less to no correlation with male employment. While we cannot comment on what exactly the impact is, we can say that heavy ST populations impact women's employment far more than men's.

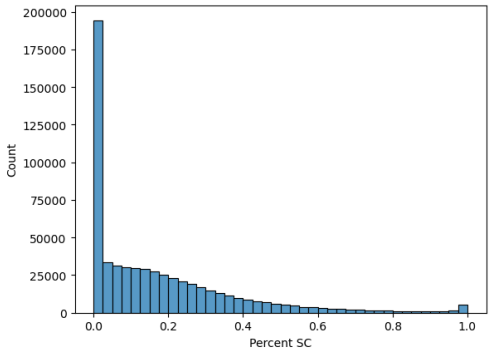
Lastly, and quite unexpectedly, there is a very strong negative correlation (strongest of all factors considered, with percent ST) between SC and ST populations. I will address this next.

Question 5

Is there anything more to be learnt regarding the negative correlation between percent SC and percent ST populations?

After some more correlation analysis, I got the idea of plotting the SC and ST percentages as a histogram.





As you can see, both histograms look somewhat similar, in such that the vast majority of villages have ST population percentages very close to 0%, or very close to 100%, with very few villages having intermediate values. One can say the same statement for the SC population distribution, with slightly more villages having intermediate values.

This distribution of SC and ST populations shows that most villages have a negligible percentage of ST population, hence putting its mean value much closer to 0 than 1. The sizable deviations from its mean occur when it takes values close to 100% for a village (since intermediate values are very rare), and since almost the entire village belongs to ST population, there is close to 0% of SC population in said village. Hence, they usually go in opposite directions from their means (~0.18 for both) together, and thus, they have a high negative correlation (when ST increases, SC below its mean). One could argue the same logic in reverse, but since the ST population is much heavier distributed at 0% and 100%, the logic is better understood this way.

We conclude this analysis by saying that the high negative correlation between percent SC and percent ST is due to the way their individual populations are distributed, and is a statistical anomaly, not a representation of some real-world interaction/conflict between SC and ST populations.

I personally found this question quite intellectually challenging, since it tested my understanding of statistics, required use of good data visualization, and concluded in a statistical anomaly that requires in-depth study to not be misinterpreted.

Question 6

What relation does the percentage of cultivators have with other factors?

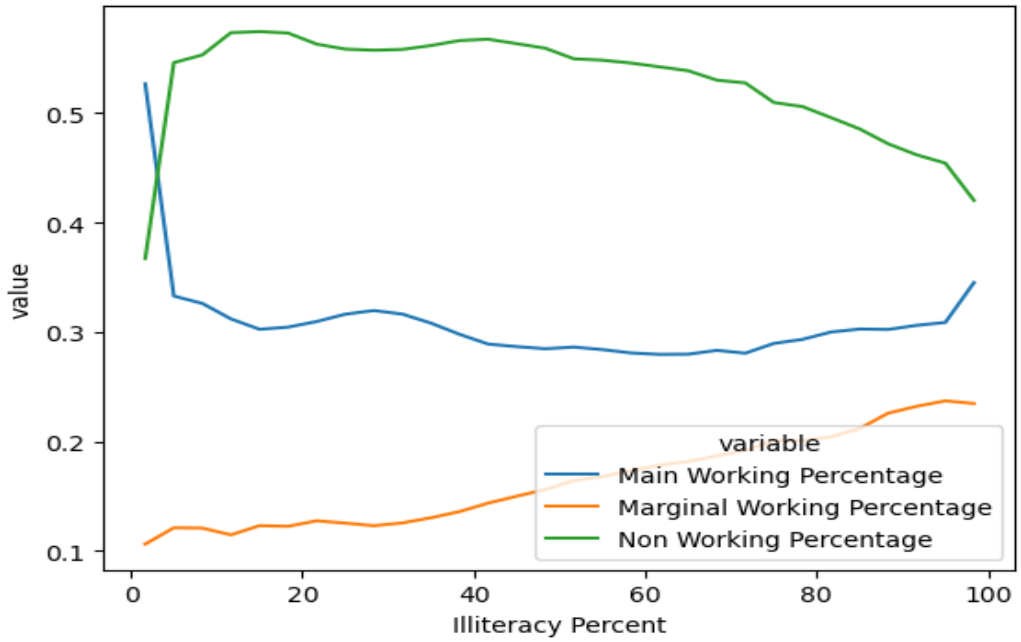
We proceed to add relevant data to our dataframe and calculate correlation values.
As can be seen from the data, there is no strong correlation between illiteracy and percent cultivators. Most of the correlation values with employment fields should be ignored, as cultivation is a form of working, and hence the fields between which correlation is being taken are in a set-subset relationship, influencing the data. However, it is interesting to note that the correlation with percent female marginal workers is significantly higher than that with percent male marginal workers, indicating that a lot of the female marginal workers are likely involved in cultivation, more so than their male counterparts.

Also, we can notice the not-negligible-but-not-very-strong correlations, positive of ST with cultivation and negative of SC with cultivation. This suggests that many of those in the ST community are likely involved with cultivation, while the opposite may be true of the SC community. As a further inference, this trend in correlations is less prominent in men than women, indicating the above scenarios are more applicable to the women involved than the men.

Question 7

How does literacy influence percentages of full-time, marginal and non-workers?

For this question, I believed there would not be a simple linear correlation but rather the presence of maximas and minimas, which would not be reflected in the correlation data. Hence, I decided to plot the following graph:



Clearly, from the above graph, we can understand the trends present. Also, note that we have taken the average across all villages, not the average across the entire population in those villages. This is to ensure that certain villages with large populations do not skew the average, and we get a better representation of the employment situations in all villages.

Firstly, the percentage of people with full-time jobs is very high for the most educated societies, but is practically constant for all others. The fact that full-time employment does not increase continuously with literacy suggests that literacy (and presence of more trained and educated workers) does not directly cause the high full-time employment in highly educated societies, but that they are correlated due to some other factors, say, through societal mindsets.

Secondly, the non-working percentage is quite low in highly literate societies, suddenly increases and reaches a maxima around 20% illiteracy, and then steadily decreases with increasing illiteracy. This may indicate that very high illiteracy is correlated with higher employment; one plausible reason is that poverty might be a common cause to both (denial to education, and more hands required to feed families). However, that is just a conjecture that requires more data to conclude.

Thirdly, the percentage of marginal workers steadily increases with illiteracy. This reflects the fact that marginal workers have a lower literacy rate than full-time workers; however, we are not sure whether illiteracy causes people to take up marginal jobs or there are other societal factors at play.

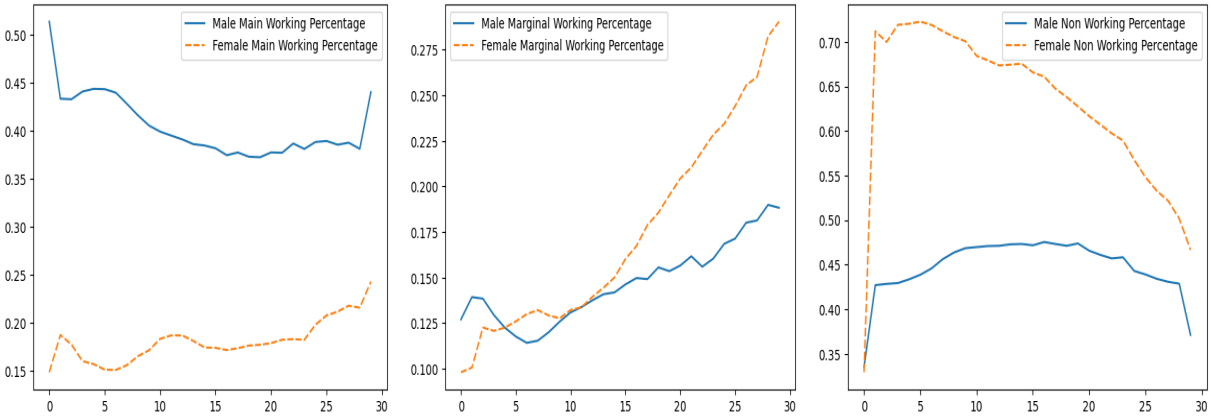
Note: From now on, all the graphs have X-axes of 0 to 30. This is because, to plot the graphs, I distributed villages into 30 buckets based on (male/female/total respectively) illiteracy rates (0-3.33, 3.33-6.66, etc.). Hence, one can regard the X-axis roughly as increasing illiteracy

---

Question 8

Does literacy affect the nature of men's or women's employment more?

In accordance with the method previously followed, I plotted the following graphs:

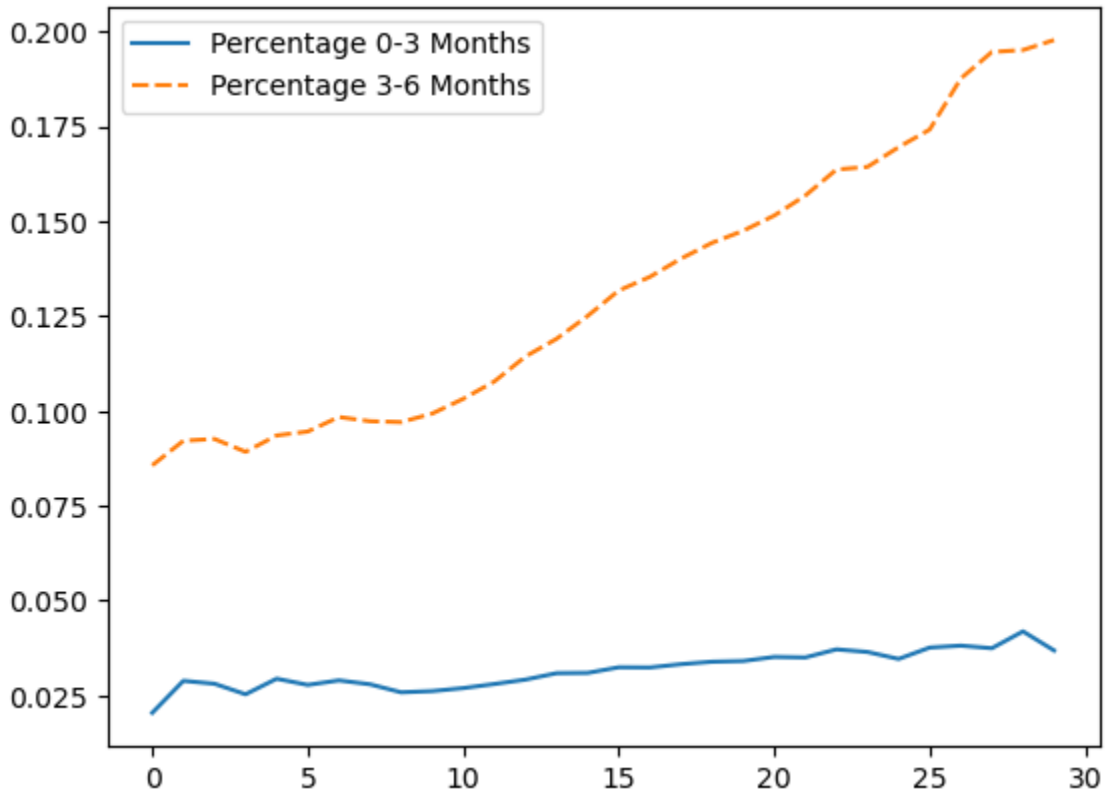


As can be seen very clearly from the graphs, the main working percentage of both men and women is relatively unaffected by the literacy rate (except in highly literate societies, which see a jump in male main occupations).

However, regarding unemployment and marginal employment rates, we can see that the female rates are affected far more than the male rates by literacy.

Question 9

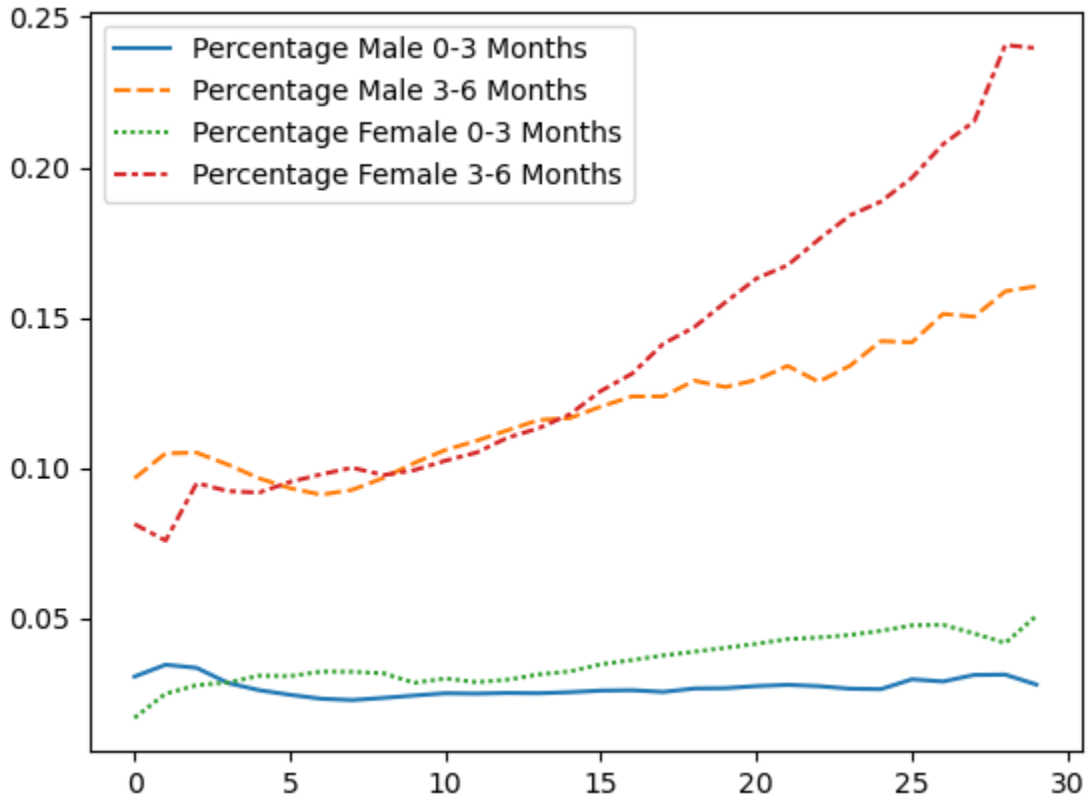
Is there any difference in trends (with literacy) of marginal workers who worked 3-6 months and those who worked 0-3 months a year?



Clearly, the percentage of marginal workers working 3-6 months a year is affected much more by illiteracy than those working 0-3 months a year. Any explanation of this phenomena undoubtedly has multiple societal and logistical components, which is beyond our score right now.

Question 10

Is the relation between literacy and the amount of time worked by marginal workers same across genders?



As can be clearly seen above, in both 0-3 and 3-6 months workers, women are less (in proportion) for literate societies but more for illiterate societies. Thus, in all categories of marginal workers, the proportion of women is more correlated with illiteracy.

Also, notice that in females, the difference in the dependance of 0-3 and 3-6 month workers with illiteracy is more than that in men.

2.

I would look into more detailed literacy data, such as level of literacy, type of school, etc. I would also consider educational opportunities available, such as schools and colleges in the area. I would also, if possible, like to have data regarding the literacy rate of families having workers in different industries/kinds of occupations.

https://censusindia.gov.in/census.website/data/census-tables
Has a lot of census tables, including some that list literacy by socio-economic background.

https://www.kaggle.com/datasets
Is also a very standard and wealthy source of data, and I found a dataset there that might be relevant:
https://www.kaggle.com/datasets/lazer999/indianschools-dataset
While difficult, it is possible to match the addresses of the schools with the districts to gain more knowledge of the available opportunities for education by district.

3.

While the given dataset does provide adequate population demographic data, there are a few more features I can think of that might be significant in said analysis.
I feel that data of the type of housing will be significant to addressing this question, as will be detailed age demographics.

https://censusindia.gov.in/census.website/data/census-tables
As mentioned before, the Indian census gives us a wealth of data from which we can deduce the type of housing (for example, from a dataset on the material used to build the walls/roof). There also exist datasets on the same that can give us detailed age demographics, district wise.

https://data.covid19india.org/

The above webpage contains the best district-wise COVID-19 data, collected as part of a crowd-sourced open source project.