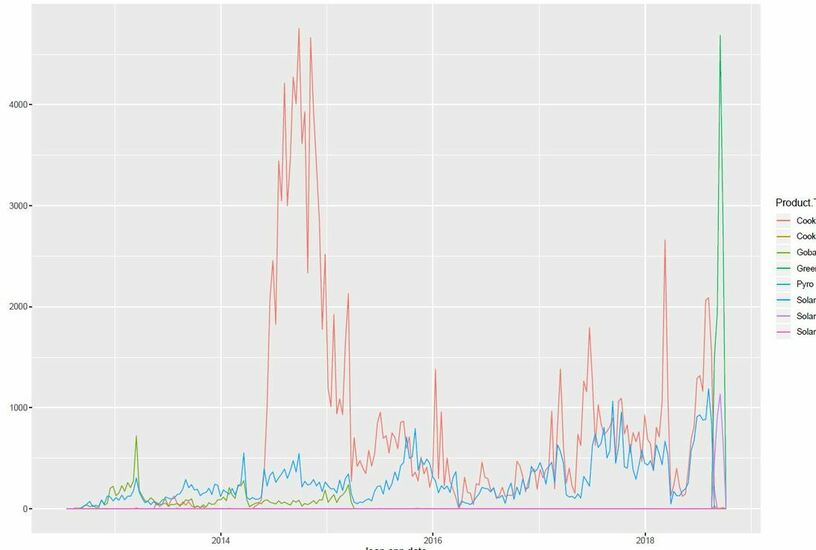
**Hypothetical Study w/ Simulated Data, a walk-through**

am writing this post primarily to describe the phases and processes a researcher has to undergo when working with data.  
  
Step 1: Getting the data -- this is an integral part of the process. Often it can involve registering with a site to obtain data, documentation, and code books, through a legitimate request submission. At other times, it can be more 'scrappy', literally involving scraping websites. There are ethical considerations to scraping that are to be kept in mind. And then, there is data handed to you by clients. In these instances, you must absolutely abide by client confidentiality, revealing no identifiable records, and treading carefully when sharing analyses.   
  
Step 2: Sizing up a data set: My most recent study with the hypothetical data in question addresses a population sample in India. The sample is sold some products- subsidized household items - from 2012 to 2018. Across 31 districts, we strive to tabulate demand against various key indicators. What these key indicators are is up to you to determine.   
​  
We figured regional variation is typically a good measure of an area's demand. And so, we begin by outlining our districts in the graph below.

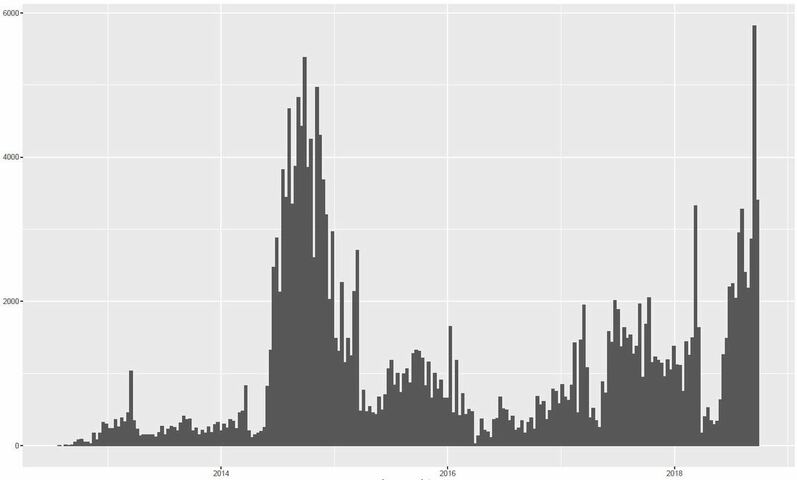


Continuing with our commentary, we next wanted to see variation in products sold by region, and over time. However, because the whole point of research is to narrow down the questions we're asking, and to narrow down the explanatory mechanisms, we will now proceed with a few key assumptions. These assumptions include what is the best way of aggregating and slicing data.   
  
Step 3: Assumptions. The original data has been aggregated from individual level to the district level. One thing to keep in mind is the fact that sheer numbers do not account for sales that accurately. The question to ask is what is driving these numbers? Is it the fact there is just a greater population in this area? Is it that people are being given larger loans to purchase products? Or, are they simply better positioned economically, endowed with greater material well being?  
  
One thing that was not previously mentioned about this rich set of mock data, is that loans are made for the individuals to purchase products. This is a standard model followed by many NGOs in the developing world.  
  
Step 4: Focusing In: We've run a summary of key indicators, probably plotted a few haphazard graphs by now. But which graphs can we use to justify the next decisions we'll make? This is the goal of this step. To share a few graphical tabulations that justify the next leap in conclusions, and so, I shall start from the top.   
  
A slightly tricky part of the process is determining which tools to use for your work. Different people have different preferences. Two people may prefer different tools for the same process, all else equal. And so, here is my takeaway:  
On this project I found it most convenient to do my reshaping using R. I found it convenient to merge data using Stata. And I found it easiest to chug out a quick plot in Tableau.  
  
Step 5: Going back and cleaning records of files and folders: For the sake of consistency, I will not post any of my earlier plots made in R. For the sake of consistency, I will also discard intermediate excel and csv files I may have outputted at the time, but which now seem like a completely useless step. In other words, whereas outputting versions of data is a useful practice in the flow, there is a point when you want to take stock.  
  
Step 6: Preliminary finding: Now that all these checks are put in place, we are ready to start collecting some preliminary findings. 



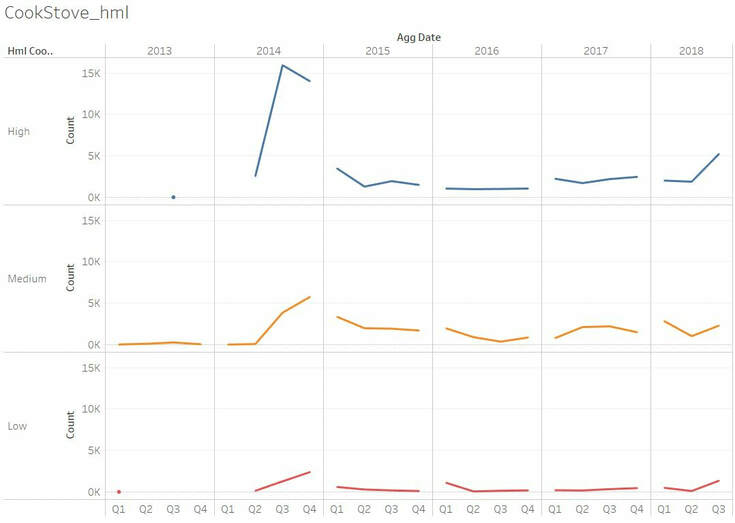
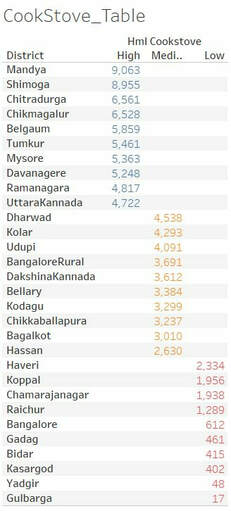
Probably the most telling of our preliminary graphs, this shows us that of our products, cook stoves (red) and solar lights (blue) produce the most consistent sales over time, with cook stoves far outstripping solar lights. Closely following are gobar gas stoves, and jumbo stoves.

Below is a most elementary tabulation of total sales over time.​ Here we clearly see a spike in sale roughly 2014-15. But this we know from the previous graph is explained by the high sales of Cook Stoves during that time period. The question remains, what drove the high sales in cook stoves from 2014-15? How did it vary by region? Was it particularly highly correlated with any other metric in the data set?  
  
Next we turn our attention to the spike in sales from 2018 - onward. Here we find that the dis-aggregated plot indicates high numbers in every major product type. Why might that be?

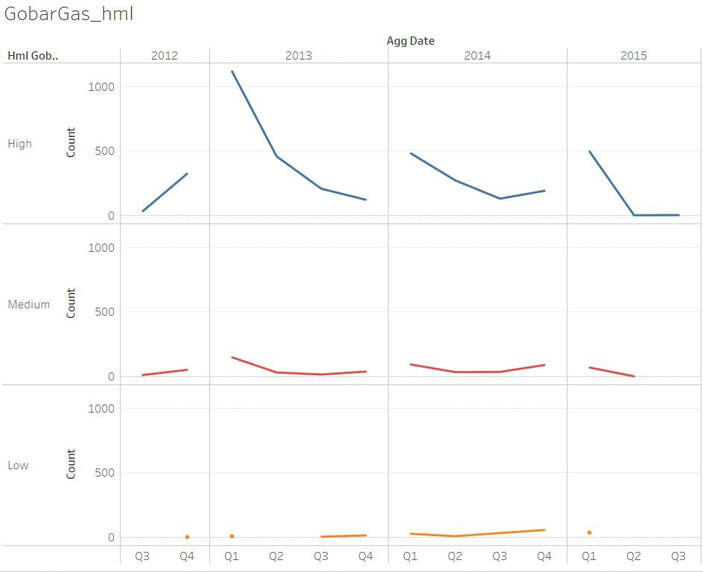
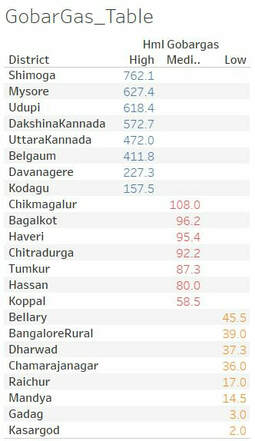


Step 7: Checkpoints: At this stage, once I am fairly satisfied with a few preliminary findings at the individual level, I deem it appropriate to aggregate some of my data. At this stage, I produce an aggregated excel file that will serve as a basis of much of my analysis moving forward, practically replacing the original data file. It is important therefore, not only to save the original file safely, but to save a copy or two of this new data someplace. Importantly, we do not plan on going back to re-run this section too often. Although it must be maintained for the sake of record-keeping.  
  
Step 8: Looking beyond the data for Answers: This step involves looking beyond the data set at hand for answers. It is an optional step, particularly if you already have a rich source of data. If, however, you doubt some of the key metrics therein, you might want to bring in additional sources of data, and link it to the present data. This involves identifying a variable by which to match our records.   
  
However, beware. While linking records by aggregate measures is tempting, it must be noted that due to differences in sampling, this is not a fully representative method of comparing aggregate measures. For instance, if linking records for districts between two data sets, one may have sampled a region within the district, whereas the other might have a random sampling. Nonetheless, when comparing across regions, you can capture some of the variation contained in the aggregate measures.

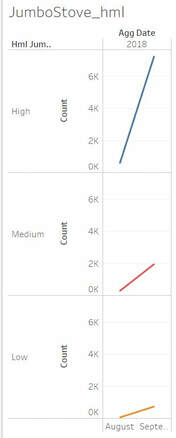
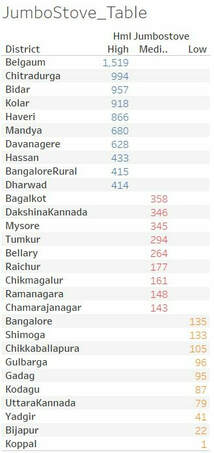
Step 9: Secondary Data Linkages: At this point my supervisor and I cannot agree on whether this secondary data carries sufficient explanatory power or not. And so, I use my judgement to append any available data, carefully documenting the nature of the combined data.   
In this case, I bring in a point estimate of various factors, particularly household material goods, and use this as an indicator of varying wealth by region.   
  
Additionally, I choose to tack on geographic indicators, simply to create powerful visuals for stakeholders.   
  
Now, equipped with a fuller data set, I find it useful to discard all the various excel files I'd outputted along the way, that will not serve much purpose moving forward. Moreover, I now opt to switch my mode for creating visuals to Tableau. I decide it might be worthwhile compiling previous R files into 1, by task/section/step of the process.  
  
Step 10: My foray into appending secondary data is not very fruitful. In terms of merging additional data, I prefer to use Stata. However, I prefer R for manipulating strings. Aha.   
  
And so, I do a quick formatting of strings, merge my files and remind myself where I'm headed.  
  
Step 10: Taking stock: Ask yourself, what's the scope of this study again?   
Answer: To chart seasonal trends in sales data.   
  
That's the simple answer. To decompose these trends by region and product, is the next layer of complexity. Region can comprise agglomerates of districts.   
  
Alternatively, decompose seasonal trends by high and low performers.   
  
An idea hits me while I write this goal out. Maybe aggregating numbers wasn't the answer. The whole idea is to leave the original data set untouched. Population is already there as an aggregate measure for weights.  
  
Step 11: Instead of hindering progress, stop, appreciate your thought process. Technically, adding coordinates was frills. And goes beyond the immediate scope. And so, instead of rewriting history at this point, I choose to forge ahead with fewer blanks to fill in.  
  
Here's the thing, while Anahit mentioned separating regions by low medium and high, she did not account for the fact that one region can be a low performer in one time period for one product, and be a high performer for another time period and product. Ultimately, we want to categorize every district/branch using an intersection of time period and product. Using this, setting a criteria, we can mark cutoffs.   
  
Step 12: Cutoffs produced, you have a setup for running some preliminary regressions to assess probabilities of being a high - medium - low purchaser of one of the four products. Now, being as this is the third day of writing this blog post, the writing process gets a bit tiresome. And drawn out. And so, I will try to deliver some brief concluding remarks. Thereafter, cleaning this post for a first brief to be submitted to my supervisor.   
The next post will detail any  
  
Step 13: Preliminary Findings of Market Research, with segments:

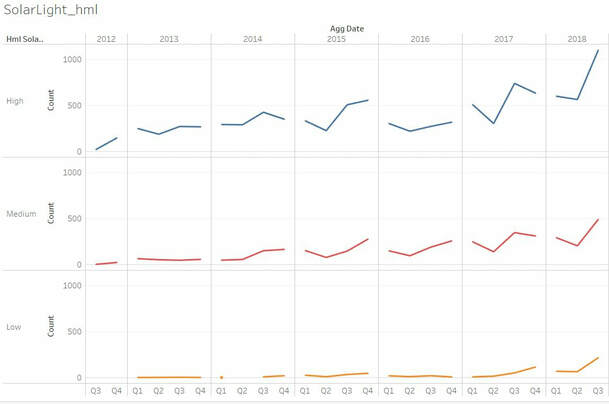
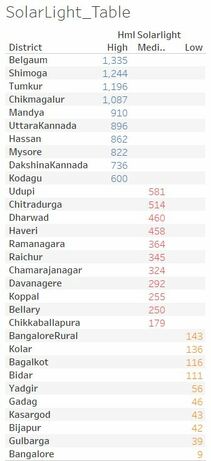
As you can see here, the basic trend suggests an uptick in sales at the beginning of each year, and a gradual tapering off through the course of the year. There is also a general overall decline in sales after the spike in 2014.

Once again, spending on gobar gas plants is higher towards the start of the year, in all categories. Here we even see a slight increase towards the end of the year. Absolute numbers are far lower than those of cookstoves.

This time round, we find that jumbo stove sales enter the market rather late in the game. However, judging by their absolute numbers, they are an instant hit, relative to gobar gas plants at least.

A bit more of a trend is discernible in the above graph. Sales for solar light increase over time. The yearly cycle is a bit difficult to untangle, although there is evidence of a rather sharp decline in the 2nd quarter for multiple years. This makes sense given longer days during the spring and summer.

Step 14: CheckPoint: Let this be a check point. At this time, I look to consolidate my preliminary findings, into a 2 page executive summary with minimal bullet points. My next goal is to produce a map for each of these 4 products, plotting low - medium - high purchasers by districts. This can be done within a day. Time to take down any sensitive information.