SHANU THAKUR -20070243037

SUBJECT: Programming for Spatial Sciences (Mini-Project)

REPORT

INTRODUCTION

The Education is a key ingredient for the growth of any sector. The process of learning and the process of learning how we shall learn is continuous in nature. With the help of right teachers they can drive the mass youth population towards a better generation of ideologist, entrepreneurs, growth, and livelihood. Post script education is like a dough in pizza, if you want your pizza to be tasty and complete you necessarily need education in today's world. And in order to see that If we're going in a right direction, I'll have to go through the data what I can possibly get and interpret optimum valuable information. So I went on and selected the the domain of Education to study the situation we poses and chose my area of interest as to be the whole nation.

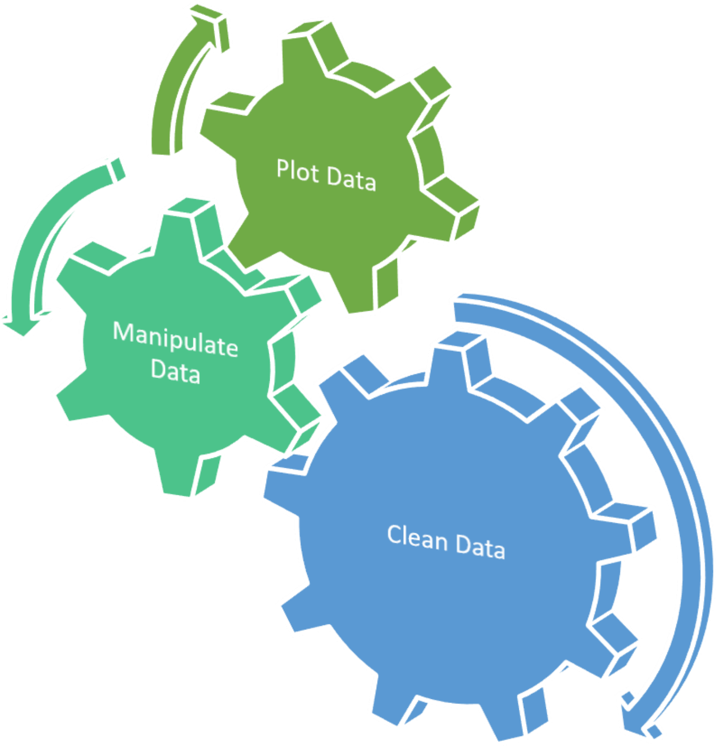
Note: There are places where it's mentioned as "we" , that we totally in refernce with me and me only used that word just for effective reading.

OBJECTIVE

- To study and know about the adequate availability of teachers , schools, classrooms with respect to the population and thereby also analyze literacy proportion.
- To study the Literacy proportion with respect to the Quality of Schools we have.
- To study the Literacy rate against total population using a scatterplot for better interpretation.
- To study the statewise total number of schools and rank them in terms of number.
- To study the total number of teachers and statewise ranking in terms of State Size using stacked bar graph.
- To study some analysis using interactive maps using Leaflet package.

FLOW OF PROJECT

After selecting the topic it was usually the process what we often read about data scientists/ analyst. For reference see the diagram. below

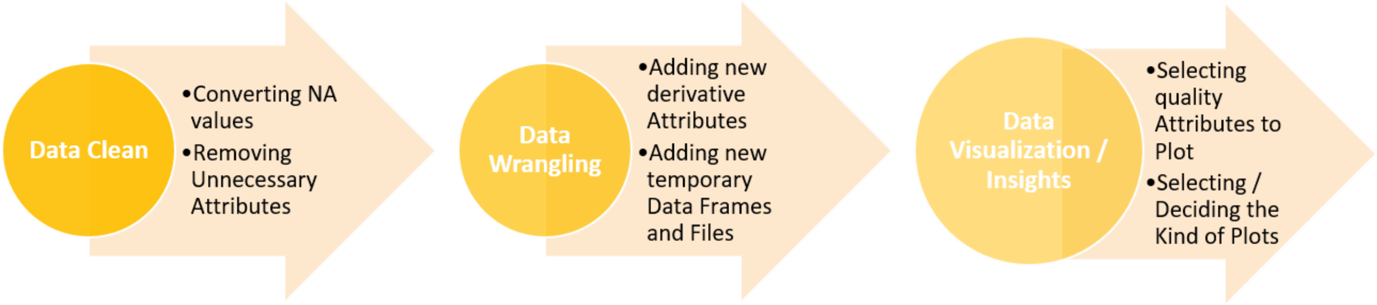


So yes we first went through the cleaning of the datasets , the data we collected was from an open organization i.e hub.arcGIS along with the data from the WorldBank. We tried going throughout out the national data center the NIC and the data.gov but we found that our governments have failed to provide with any such data. So had to work with very low resources and they me also seem incomplete or missing information, largely affecting the analytics. But we still had to go through this so we wrangled and cleaned , sorted and filyter as much as possible. Since the appropriate data wasn't available at any one source except for the hub.arcGIS, only they were providing with such a effective data but with a drawback of having recorded information only of 2013.

ACTUAL EXPERIENCES AND THINGS LEARNED

It started with the taking overview of data we saw each and every column (the attributes of the datasets), and the theorrotically we decided what elements to keep and what not, etc.

Later on we went to clean data following with the wrangling of data to get some columns ready for plotting it later on. Our approach is depicted in below given diagram.


Data Clean
•Converting NA values
•Removing Unnecessary Attributes
Data Wrangling
•Adding new derivative Attributes
•Adding new temporary Data Frames and Files
Data Visualization / Insights
•Selecting quality Attributes to Plot
•Selecting / Deciding the Kind of Plots

Data Cleaning included removing NA values and deleting certain waste columns . Following the data wrangling part included adding of data and datasets like Ranks, Rank2 , temp, etc

It was often tough choices to select what attribiute shall we select to plot against what. Because total population vs anything was a good analytic But we got through it and made some amazing plots. That too via notebook HTML.!

Many times it was some minor understanding that took so long that we had to stay for like 3/ 4 and sometimes 6 hours straigt to rectify a minute error. Like sometimes it was about the merging the shape file and the data frame. In my case I was continuously merging in a way that in output it did not give me a merged shape file instead it gave a data frame with class of list! But after some long R&D it was later rectified, another was that about the stacked bar graph it was not possible to plot more then one numeric attribute at once but some how I tricked it into data frame(Rank2) and I was then able to complete that too.