Introduction to Data Science - Project

There are two kinds of project you can choose from:

A. Data Preprocessing and Preliminary Analysis (Data should be numerical) and get inferences from the data. You should use the statical analysis that you had learned in the first two chapters.
B. Applying ML Classification algorithms on the data set and getting inferences from the data. You may use the appropriate ML algorithm packages available in R or Python. But know which algorithm you use and know the concept behind it.

For both the above options, these are the following steps that needs to be addressed as part of your project:

1. Collect a dataset from data.gov.in or https://archive.ics.uci.edu/ml/datasets.php
2. Understand the various features (columns) of the dataset.
3. Do preprocessing on the dataset.
4. Do preliminary analysis on the dataset as much as needed and get a summarisation of your inferences with proper validation
5. You may use ether R or Python for this project
6. For ML classification you may use any appropriate ML classification algorithm (you need to justify why you are using the same)
7. If the dataset is huge start doing your analysis on a random sample dataset from the original huge dataset and slowly scale it up.
8. Write a complete report as a Google doc with needed plots, charts, bar graphs, pie charts and so on and submit it. Report should also contain the description about the data set, preprocessing done - why, how and what, analysis done / ML classification done, for analysis what is the purpose of doing, for ML why specific algorithms / packages were used, what is the output, what are you inferences from the output, what are your questions about the data? what are questions addressed by your analysis and so on.
9. Report should also contain all the details, like, from where you collected the dataset, how many rows and columns are there, what are the features, why we do such and such preprocessing and after preprocessing what is changed in the dataset, what are your basic questions about your data, how you are going to address those questions through an analysis (give this with substantiation about why you use such analysis).
10. Report should contain all the codes written by you and in addition to that it should list out the inferences you made through the analysis and how this answers your questions about the dataset.
11. The report should be submitted by Nov 30, 2020 (11.59pm) through Moodle (use the Main server). Late submission will have penalties and if it is submitted too late (more than a day) then it will not be evaluated.
12. Do not copy paste the codes from any Internet source or from your friends: If we come to know about this then your project report will be totally rejected.
13. In general, marks will be awarded mainly based on how much you were able to dig into the data and get inferences out of it.
14. The report should be comprehensive and it should be formatted nicely like a project report. Report should clearly specify the team member names and their roll numbers.
15. You should form a team of 4 members (not less and not more) for this project. Members can be across batches.
16. If you have any doubt regarding the project raise the doubts on Piazza. DO NOT send emails.

All the best!