Hi Folks,

Congratulations on making it this far. Now we are dealing with real-world Data. Now we'll be exploring the Data, Answering the queries using Python and its libraries. So, fasten your seatbelts and get ready for an amazing journey

The concepts you have learned till now will be applied to an exploratory data analysis of a real-world dataset of your choice. You can use this starter notebook as an outline for your project. Include detailed explanations wherever possible in Markdown cells - this Jupyter notebook will also serve as a project report.

Evaluation Criteria

The following criteria will be used to evaluate your submission:

- 10 Projects are mandatory for every student across diverse dataset
- There must be at least three columns and 150 rows in the dataset
- At least four questions must be asked and answered about the dataset
- At least four visualisations (graphs) must be included in your submission
- Apart from the code, your submission must include explanations using markdown cells.
- The work you submit must not be plagiarised, i.e. copied from another source.

Follow this step-by-step guide to work on your project.

Step 1: Select a real-world dataset

- Find an interesting dataset on this page: https://www.kaggle.com/datasets?fileType=csv
- The data should be in CSV format, and should contain at least 3 columns and 150 rows

You can find a list of recommended below in this file

Step 2: Perform data preparation & cleaning

- Load the dataset into a data frame using Pandas
- Explore the number of rows & columns, ranges of values etc.
- Handle missing, incorrect and invalid data
- Perform any additional steps (parsing dates, creating additional columns, merging multiple dataset etc.)

Step 3: Perform exploratory analysis & visualization

Matplotlib - Resources

- Compute the mean, sum, range and other interesting statistics for numeric columns
- Explore distributions of numeric columns using histograms etc.
- Explore relationship between columns using scatter plots, bar charts etc.

- Make a note of interesting insights from the exploratory analysis

Step 4: Ask & answer questions about the data

- Ask at least 4 interesting questions about your dataset
- Answer the questions either by computing the results using Numpy/Pandas or by plotting graphs using Matplotlib/Seaborn
- Create new columns, merge multiple dataset and perform grouping/aggregation wherever necessary
- Wherever you're using a library function from Pandas/Numpy/Matplotlib etc. explain briefly what it does

Step 5: Summarize your inferences & write a conclusion

- Write a summary of what you've learned from the analysis
- Include interesting insights and graphs from previous sections
- Share ideas for future work on the same topic using other relevant datasets
- Share links to resources you found useful during your analysis

Step 6: Make a submission & share your work

- Prepare the whole project over Kaggle
- Share the public link to Rakshit over discord

(Optional) Step 7: Write a blog post

- A blog post is a great way to present and showcase your work.
- Sign up on Medium.com to write a blog post for your project.
- Copy over the explanations from your Jupyter notebook into your blog post, and embed code cells & outputs

Example Projects

Refer to these projects for inspiration:

- Video Games Analytics
- EDA Projects
- Netflix EDA
- University ranking EDA

Some interesting datasets

1. Video Games sales: https://www.kaggle.com/gregorut/videogamesales 465
2. World University Rankings: https://www.kaggle.com/mylesoneill/world-university-rankings 287
3. Netflix Tv shows and Movies: https://www.kaggle.com/shivamb/netflix-shows/notebooks 692
4. StackOverflow Developer Survey: https://www.kaggle.com/stackoverflow/stack-overflow-2018-developer-survey 79
5. Google Play Store Android Apps Data: https://www.kaggle.com/lava18/google-play-store-apps 448
6. Indian Stock Market Data: https://www.kaggle.com/rohanrao/nifty50-stock-market-data 344
7. Indian Air Quality: https://www.kaggle.com/rohanrao/air-quality-data-in-india 475
8. Worldwide Covid-19 Cases: https://www.kaggle.com/imdevskp/corona-virus-report 301
9. USA Covid-19 Cases: https://www.kaggle.com/sudalairajkumar/covid19-in-usa 166
10. US Election Results (2012): https://www.kaggle.com/tunguz/us-elections-dataset 104
11. US Stock Market: https://www.kaggle.com/borismarjanovic/price-volume-data-for-all-us-stocks-etfs/ 104
12. Crop production in India: https://www.kaggle.com/srinivas1/agricuture-crops-production-in-india 314
13. Agricultural raw material prices: https://www.kaggle.com/kianwee/agricultural-raw-material-prices-19902020 128
14. Agricultural land values: https://www.kaggle.com/jmullan/agricultural-land-values-19972017 104
15. Digital payments in India: https://www.kaggle.com/lazycipher/upi-usage-statistics-aug16-to-feb20 366
16. US Unemployment Rate Data: https://www.kaggle.com/jayrav13/unemployment-by-county-us 154
17. India Road accident Data: https://community.data.gov.in/statistics-of-road-accidents-in-india/ 354
18. Data Science Jobs Data:
    a. https://www.kaggle.com/sl6149/data-scientist-job-market-in-the-us 118
    b. https://www.kaggle.com/jonatancr/data-science-jobs-around-the-world 143
    c. https://www.kaggle.com/rkb0023/glassdoor-data-science-jobs 73
19. Youtube Trending Videos: https://www.kaggle.com/datasnaek/youtube-new 334
20. Asteroid Dataset: https://www.kaggle.com/sakhawat18/asteroid-dataset 137

21. Solar flares Data: https://www.kaggle.com/khsamaha/solar-flares-rhessi 83
22. F-1 Race Data: https://www.kaggle.com/cjgdev/formula-1-race-data-19502017 133
23. Automobile Insurance: https://www.kaggle.com/aashishjhamtani/automobile-insurance 98
24. PUBG video game matches: https://www.kaggle.com/skihikingkevin/pubg-match-deaths 198
25. CounterStrike GO (video game)
    a. https://www.kaggle.com/mateusdmachado/csgo-professional-matches 45
    b. https://www.kaggle.com/skihikingkevin/csgo-matchmaking-damage 23
26. Dota 2 (video game): https://www.kaggle.com/devinanzelmo/dota-2-matches 61
27. Cricket One-Day Internationals Data: https://www.kaggle.com/jaykay12/odi-cricket-matches-19712017 161
28. Cricket Indian Premier League Data: https://www.kaggle.com/nowke9/ipldata 329
29. Basketball (NCAA): https://www.kaggle.com/ncaa/ncaa-basketball 71
30. Basketball NBA Players Stats: https://www.kaggle.com/ncaa/ncaa-basketball 71
31. Football datasets:
    a. https://www.kaggle.com/martj42/international-football-results-from-1872-to-2017 90
    b. https://www.kaggle.com/abecklas/fifa-world-cup 99
    c. https://www.kaggle.com/egadharmawan/uefa-champion-league-final-all-season-19552019 87
32. Hotel Booking Demand: https://www.kaggle.com/jessemostipak/hotel-booking-demand 183
33. New York Airbnb listings: https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data 81

Other sources to look for datasets:

- UCI Machine Learning Repository 242
- awesome-public-datasets 428
- Google Dataset Search 301

If you use an external source other than Kaggle, you'll create a new dataset on Kaggle by uploading a CSV file here: https://www.kaggle.com/datasets?new=true 129 (make sure to keep your dataset public, otherwise it will not be downloadable using `open datasets`*)*

Downloading Personal data for EDA

You can also analyse your own personal data for exploratory data analysis, from the following sources:

- Whatsapp Chat data https://jovian.ai/PrajwalPrashanth/whatsapp-chat-data-analysis/v/10#C2 95
- Google Apps data from https://takeout.google.com/ 57
  - Chrome (https://medium.com/free-code-camp/understanding-my-browsing-pattern-using-pandas-and-seaborn-162b97e33e51 43)
  - Contacts
  - Calendar
  - Drive
  - Fit
  - Google Pay
  - Maps
  - …
- Data from Apple's Apps https://appleinsider.com/articles/18/05/23/how-to-request-your-personal-data-using-apples-data-privacy-portal 18
  - Instagram Data https://www.instagram.com/accounts/login/?next=/download/request/ 65
  - Fitbit Data https://help.fitbit.com/articles/en_US/Help_article/1133.htm 18
  - LinkedIn Data https://www.linkedin.com/help/linkedin/answer/50191/downloading-your-account-data?lang=en 41
  - Shopping analysis, Amazon data https://www.amazon.com/gp/help/customer/display.html?nodeId=G5NBVNN2RHXD5BUW 68
  - Spending analysis, check your bank's website and you would be able to export CSV/excel statements for at least a year.