Part 1

The environment that our team chose to use for this assignment is the docker environment. We use docker containers and dropbox, provided by the course, to set our environment to perform all tasks necessary to sufficiently complete this course to the best of our abilities. In this google cloud environment, we set our operating system to Ubuntu Linux. Then once we were using the Ubuntu operating system, we installed the course docker that provided us with all necessary packages to complete this course.

We felt that our chosen approach was a very capable environment but did see rather a few advantages and disadvantages compared to other environments. Compared to Kaggle Kernels, the other three environments: Docker, Collab, and Conda, provide isolated, fully configurable environments allowing developers to avoid dependency issues (Misal, D. 2019b). Docker and Conda are relatively similar, except that Docker deals in Dockerfiles rather than the package manager concept that Conda utilizes (Lodge, M. 2018). Simply put, Conda is a package manager, where Docker is a container platform that lets you package your environments in an isolated container. On the other hand, Google Collab already has frequently used packages installed, which simplifies the process and reduces customization abilities (Aaryan, Y. 2020). Additionally, Collab offers a collaborative environment for sharing code and working in groups which is helpful (Aaryan, Y. 2020). Besides, Kaggle Kernels is good when developing simple tasks as it saves you the time to set up and it allows user to use Jupyter notebook in the browser (Misal, D. 2019a).

For this course, we felt Docker is more suitable because there are various dependencies conflicts and software issues if we were to set the environment for each task by ourselves. Containers provided by the course are easy to install thanks to the thorough instruction you have provided. One of the issues we faced when creating our own environment is that the Java version, we installed for another course was not supported by H2O. Another issue is also with pickle object that if the object was created by a different version of pandas, it could not be called from another version of pandas. Also, working on a similar Linux virtual environment allows us to share code seamlessly as our group has both Mac and Windows users.

Overall, our choice of using Docker was based on the simplicity of installing the complex environment and the reproducibility associated with the environment where everyone on the team could be on the exact same page without having to face multiple technical issues.

Reference list

Aaryan, Y. (2020). *Colab vs Kaggle - Which is better? | Data Science and Machine Learning*. [online] a. Available at: https://www.kaggle.com/product-feedback/147587 [Accessed 16 Jul. 2021].

Lodge, M. (2018). *Conda, Docker, and Kubernetes: The cloud-native future of data science (sponsored by Anaconda): Big data conference & machine learning training | Strata Data*. [online] 71478.html. Available at: https://conferences.oreilly.com/strata/strata-ny-2018/public/schedule/detail/71478.html [Accessed 16 Jul. 2021].

Misal, D. (2019a). *5 Alternatives To Google Colab For Data Scientists*. [online] Analytics India Magazine. Available at: https://analyticsindiamag.com/5-alternatives-to-google-colab-for-data-scientists/ [Accessed 16 Jul. 2021].

Misal, D. (2019b). *Google Colab Vs Kaggle Kernels: Which Of The Two Platforms Should You Go For?* [online] Analytics India Magazine. Available at: https://analyticsindiamag.com/google-colab-vs-kaggle-kernels-which-of-the-two-platforms-should-you-go-for/ [Accessed 16 Jul. 2021].