Team Byte Me:

- Andrew Kuruvilla
- Igor Lucic
- Rayyaan Haamid
- Tales Araujo Leonidas

Professor Patricia McManus

ITAI 1378: Computer Vision

January 29, 2024

Understanding and Integrating GitHub and Jupyter Notebooks in Real-world Applications

Introduction

GitHub and Jupyter Notebooks have emerged as indispensable assets for professionals aiming to enhance collaboration, version control, and interactive computing. This report delves into the core functionalities of GitHub, elucidating its pivotal role in fostering collaborative version control, explores the dynamic capabilities of Jupyter Notebooks in facilitating various tasks through an interactive web-based environment, and how they can integrate into real-world applications.

GitHub

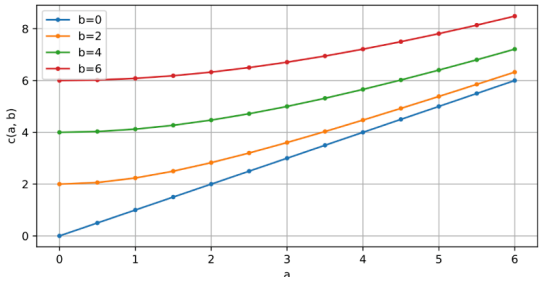
GitHub is a web-based platform that serves as a repository hosting service, fundamentally designed to facilitate version control and collaboration among programmers and developers. At its core, GitHub employs Git, a distributed version control system created by Linus Torvalds, to enable multiple individuals to work on projects simultaneously, efficiently managing changes to documents, programs, and other information types. It offers a suite of collaborative features such as bug tracking, feature requests, task management, and wikis for every project. GitHub's intuitive interface simplifies the process of contributing to open-source projects or managing private repositories for personal or organizational use. By hosting a vast array of projects ranging from simple code scripts to complex software systems, GitHub has become an indispensable tool in the software development process, fostering a community where developers can share, collaborate on, and evolve their software projects across the globe.

Jupyter Notebooks

Jupyter Notebooks are key tools for collaborative computing, offering a unique platform that merges code execution, rich text, mathematical equations, visualizations,

and other media into a single, interactive document. “This single document approach enables users to develop, visualize the results, and add information, charts, and formulas that make work more understandable, repeatable, and shareable” (Wickramasinghe, 6). When integrated with version control systems like GitHub, their collaborative capabilities are further amplified, enabling users to track changes, review code, and merge contributions efficiently. Jupyter Notebooks are extensively utilized for developing and presenting data analysis projects due to their capability of integrating data visualization libraries and machine learning models to assist in dynamic insights.

```
plt.xlabel('a')
plt.ylabel('c(a, b)')
plt.grid()
plt.legend();
```

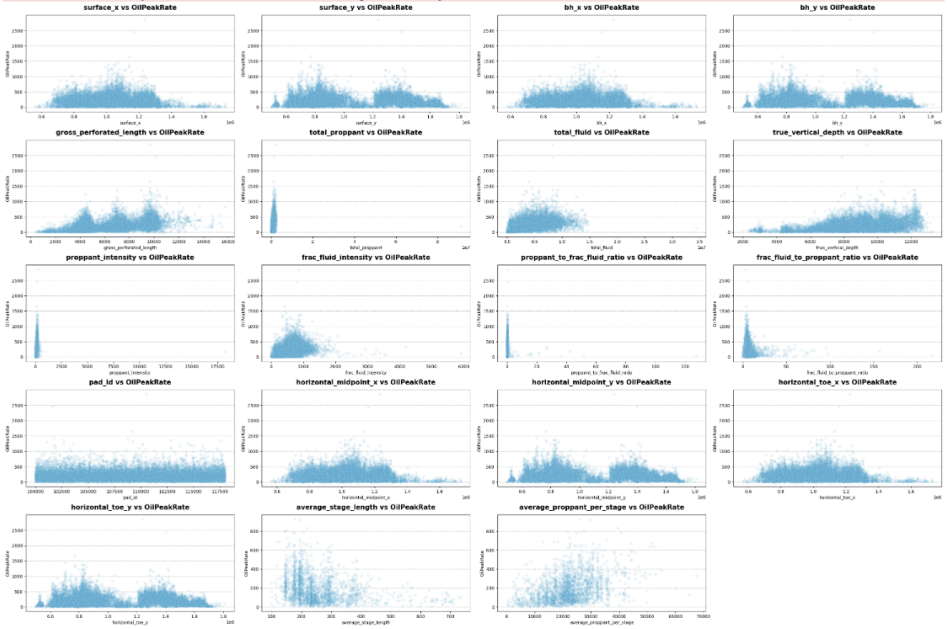


Real-world Applications

Let’s say Chevron needs to find a model to predict the peak oil production of a well. They share the data of 30,000 wells and 21 points of data (location, age of well, depth, etc) so they are giving you 720,000 things to check out! It would be impossible for one, or a team of people to go through that by hand. Let alone be able to use that data to predict anything, so you would put it into Jupyter to visualize the date and get something like this:

```
plt.tight_layout()
plt.show()
```

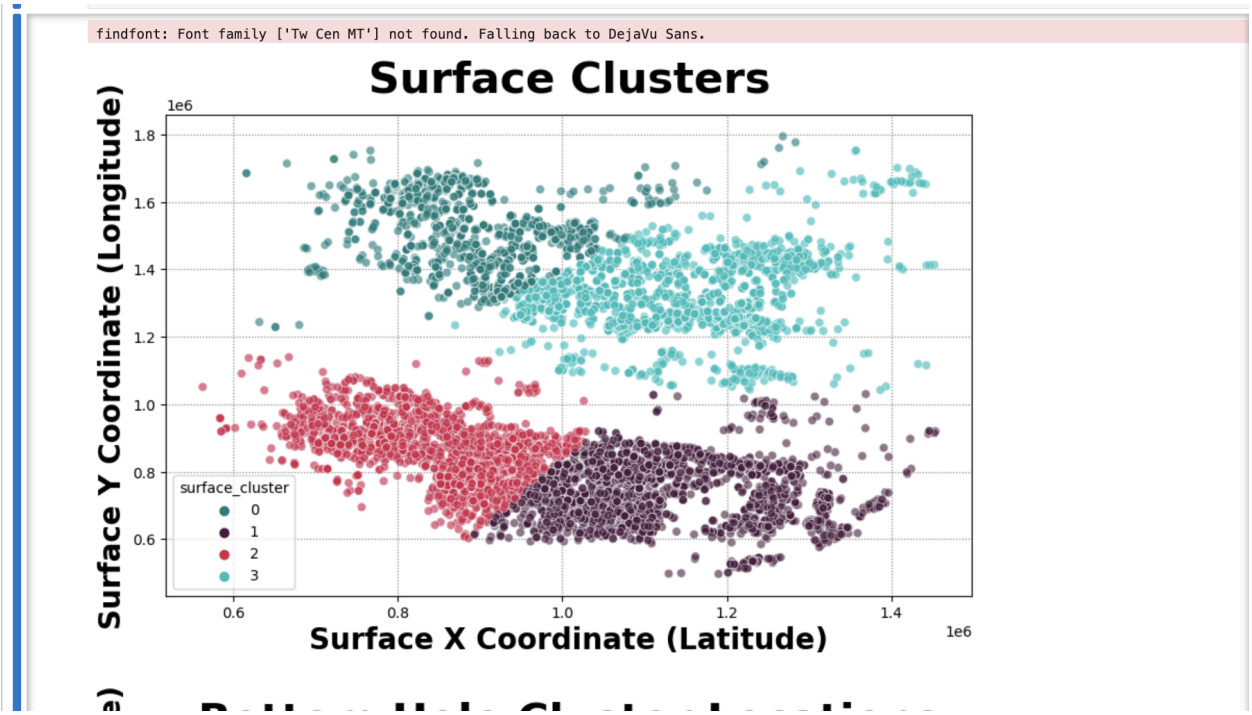
```
findfont: Font family ['Tw Cen MT'] not found. Falling back to DejaVu Sans.
```




```
[7]: data_with_oilpeak_reduced = data_with_oilpeak_reduced.drop(columns=['average_stage_length','average_proppant_per_stage','average_frac_fluid_pe
     test_data1_reduced = test_data1.drop(columns=['average_stage_length','average_proppant_per_stage','average_frac_fluid_per_stage','pad_id'])

     numerical_columns = ['surface_x','surface_y','bh_x','bh_y','gross_perforated_length','total_proppant', 'total_fluid',
```

Now we can get a picture of what is going on!Then you spend hours cleaning up the data and removing outliers and redundancy. You can also utilize a “model” like linear regression. Let’s use that for this case to grab cluster data using random forest:



Conclusion

In conclusion, AI lives off Data. But if we have a messy data points or no organization, we cannot get anywhere. We need to be able to effectively sort important data and keep it as clean as possible. We also need to limit overfitting as much as possible. We want our AI model to learn, not just memorize things. We also need to use GitHub in order to have a “homebase” for our data so our team can grab and update code from a centralized position.

Works Cited

Fangohr, Hans. “Jupyter for Computational Science and Data Science — Computational Science and Data Science.” *Hans Fangohr*, 30 November 2022, https://fangohr.github.io/blog/jupyter-for-computational-science-and-data-science.html. Accessed 29 January 2024.

Wickramasinghe, Shanika. “Jupyter Notebooks for Data Analytics: A Beginner's Guide.” *BMC Software*, 9 August 2021, https://www.bmc.com/blogs/installing-jupyter-for-big-data-and-analytics/. Accessed 29 January 2024.