**Web Exercise 1: Introduction of GitHub and Online Data Science Resources**

1. ***What are the key functions of GitHub? Who are the users? Please use your own words to describe them.***

**GitHub** is one of the most widely used web-based platforms for version control and project collaboration. **Git** is an open-source software that is used for version control. GitHub leverages Git to manage version control for project collaborations in the cloud, allowing users worldwide to access and contribute to private or public projects seamlessly.

The key functions of GitHub are (but not limited to following):

1. **Manage Codebases:** Store and manage code in cloud repositories, making it accessible over the internet for team collaboration on development tasks.
2. **Version Control:** Compare, push, and merge local code changes to centralized cloud repositories.
3. **Sync with Latest Code:** Fetch the latest code from the repository to stay up to date with the current source code of any development project.

**GitHub** offers both free and paid licensing options for different use cases. The target audiences for GitHub include:

1. **Multinational Organizations:** Companies working on various software development projects can set up private cloud workspaces (paid license), separate from the public GitHub domain.
2. **Small-Scale Development Projects:** Open-source developers or startups in the early stages of development might use the free version, without the need for a paid private cloud workspace.
3. **Individual Developers:** Developers looking to host personal projects online.
4. **General Users:** Anyone needing a secure public cloud workspace to showcase their projects.
5. ***What are the differences between “clone” and “pull request” in GitHub?***

Cloning, as the term suggests, involves creating a copy of a repository from a remote source, such as GitHub, onto your local machine. This process includes downloading the entire codebase along with all branches and historical metadata. We typically use the **git clone** command to set up a local workspace from an existing remote repository. This cloned repository becomes a new node in the version control network, allowing the user to collaborate with the main (or master) codebase.

On the other hand, the **git pull** command is used to fetch the latest changes from the remote repository and synchronize them with your local copy. After cloning a repository, you usually use the **git pull** command to keep your local codebase up to date with the latest changes from the remote repository.

1. ***Introduce ONE selected “Dataset” from the [awesome-datascience] GitHub with its URL and describe its potential applications and values.***

***Dataset*** ***-*** [data.gov](https://catalog.data.gov/dataset) (The home of the U.S. Government's open data)

This website hosts various datasets collected and maintained by the U.S. Government over the years, all available in the public domain. I am particularly interested in working with financial datasets. Instead of focusing on a single dataset, I plan to utilize multiple related datasets from the website that are associated with financial fraud.

For example, the “**21st Century Corporate Financial Fraud, United States, 2005-2010**” dataset can serve as the foundation for my fraud detection pipeline, while other databases, such as the “**Consumer Complaint Database**” and “**Debt to the Penny**” can act as supplementary datasets for predictive analysis in fraud detection and prevention. These supporting datasets may not be directly linked to the core fraud datasets, but they could provide crucial insights into the factors that contribute to financial fraud.

Although this is an initial concept for applying these datasets, we can enhance the fraud detection pipeline by incorporating additional datasets during the detailed Knowledge Discovery in Databases (KDD) phase.

1. ***Introduce ONE selected “free-data-science-book” with its URL and author/institute information. Explain briefly why you are interested in this book.***

**Data Science Book -** <https://otexts.com/fpp2/> (Forecasting: Principles and Practice)

This online textbook provides a thorough introduction to forecasting methods, emphasizing practical applications over theoretical discussions. It highlights the use of R for forecasting, offering practical examples with real data through the **fpp2** and **forecast** R packages. Additionally, the book focuses on graphical methods, primarily using **ggplot2**, to explore data and present forecasting outcomes.

These topics will help me gain a strong understanding of forecasting methodologies, which I can apply to various data pipelines. Currently, I can use these methods to develop a pipeline for financial fraud detection utilizing the datasets available on [data.gov](https://catalog.data.gov/dataset)

1. ***Print out the README.md (in a regular display mode inside GitHub) as a PDF. Add this to your submission as Attachment #1.***

Link to the attachment #1 – [Link](https://github.com/pTidke/BDA594-ptidke/blob/master/Attachments/Attachment%20%231.pdf)

1. ***Print out your GitHub website from a web browser (such as Chrome) as a PDF. Add this to your submission as Attachment #2. (Make sure you introduced your website URL, your data set example, and data science book on your GitHub website).***

Link to the attachment #2 – [Link](https://github.com/pTidke/BDA594-ptidke/blob/master/Attachments/Attachment%20%232.pdf)

1. ***GitHub Details for the assignment –*** 
   * GitHub repository - [BDA594-ptidke](https://github.com/pTidke/BDA594-ptidke.git)
   * GitHub personal website - [Link](https://ptidke.github.io/)