Data science notes

Categories of Data Science

1. Data management
2. Data integration and transformation
3. Data Visualization
4. Model Building – train data and analyze patterns using suitable machine learning algorithms.
5. Model Deployment – Process of integrating developed model into production.
6. Model Monitoring and Assessment – Continuous check for accuracy and fairness

Github is an example of code asset management

Open Source Tools for Data Science

Data management tools – MySQL, PostgresSQL

Data integration and transformation – Apache airflow, apache spark sql, alteryx

Data visualization – kibana, PixieDust, Tableau

Model Deployment – Prediction IO, Seldom

Model Monitoring – ModelDB

Jupyter

Libraries for data science

Scientific computing libraries

1. Pandas
2. numpy

Visualization libraries

1. matplotlib
2. seaborn

Machine learning and deep learning libraries.

1. Scikit-learn
2. Keras
3. Tensorflow
4. pytorch

Other libraries

1. Apache spark

Libraries are a collection of function and method that allow you to perform many actions without writing code

**Application Programming Interfaces (APIs)**

An API allows communication between two pieces of software

API Library

REST – Representational State transfer API – comuunicate thourhg the internet

Takes advantage fo storage. Data and artificially intelligent algorithm

**Data sets powering data science**

Data is set is a collection of data

Data structures

* Tabular data
* Hierarchical data
* Raw files

Where to find data sets – datacatalogs.org, Kaggle.com, data.setsearch.research.google.com

Machine learning Model – uses algortihms and models to learns data patterns

Afte a model is trained it can make predictions

**Introduction to Jupyter Notebooks**

* Define jupyter note-book
* Explain how to use jupyterlab
* Describe how to use the notebooks in jupyter lab

JUpyter notebook is a web browser application that allows you to create and share documents containing code, equations, visuallizations and more.

Allows ti combine text, code block and code output in a sigle file

Jupyter Lab allows access to multiple jupyter notebook files, other codes and data files

Enables working in an integrated manner

Is compatible with several file formats and is open source

**Getting started with Jupyter**

* Describe how to run, insert and delte a cell in a notebook
* Work with multiple notebooks
* Work with multiple notebooks
* Close a notebook after done with session

**Jupyter Kernels**

* Define kernel
* Describe how to work with kernels

Kernle is a computational engine that executes the code contained in a notebook file

**Jupyter Architecture**

* Describe basic jupyter architecture
* Explain jupyter architecture for conversion to a file format

2 process model – kernel and client

The notebook server is responsible for saving and loading the notebooks

The kernel executes the cells of code contained in the notebook

Jupter architecture used the NB convert tool to convert files to other formats

**AdditionlAL Anaconda Jupyter Environment**

* Describe anaconda and its dagta science features
* Describe anaconda jupyter environments
* Identify tools in anaconda juptyer environment

Combine coded cimput

Jupyterlab – an open source web based application , enables the cration of code and interactive visualizations and included pre installed python libraries (munpy, pands, matplotlib)

Anaconda is a free and open source distributor for python and R

Jupyter is a poplar computational notebook tool because it supports dozens of programming languages

The anaconda navigator GUI can launch multiple appl;ications oon a local device

Jupyter environment in the anaconda navigator include jupyterlab and VS code

**Cloud Based Jupyter Environment**

* JupyterLite – executes in brower
* GoogleCollab

**Introduction to R and R Studio**

Explain what R is

List R capabilities

Describe RStudio

List the R libraries for data science

R is a statistical processing language

Used for data processing and manipulation, statistical analysis, and machine learning

R is used by most academics, health care amd government

Supports importing data from different data sources; flat file,.databases, software, web(STSS Stata)

R Capabilities

It is easy to use compared to other data science tools

Great tool for visualization

Basic data analysis doesn’t require installing packages

RStudio is an integrated development environment (IDE)

IT INCREASE PRODUCTIVITY IN RUNNING THE Programming language

Includes; code editor, console, workspace, history tab, files, plots, packages

Packages - dplyr - data manipulation, stringr - string manipulation, ggplot - data visualization, caret - machine learning

**Overview of GIT/GITHUB**

Version Control – allows you to keep track of changes to your documents. Makes collaboration a lot easier

**Git**

Free and open-source software

Distributed version control system

Accessible anywhere in the world

One of the most common version control systems available

Can also version control images, documents

SSH protocol – a method for secure remote login from one computer to another

Repository – the folders of your project that are set up for version control

Fork – a copy of a repository

Pull request – the process you use to request that someone reviews and approves your changes before they become final

Working directory = a directory on your file system including its files and subdirectories that is associated with a git repository

Basic git commands

Get init

Get add

Get status

Get commit

Get rest

Get log

Git branch

Git checkout

Git merge

**Introduction to GitHub**

Describe the purpose of source repositories

Explain how GitHub satisfies the needs for a source repository

**Git Repository Model.**

Distributed version control system

Tracks source code

Coordinates among programmers (supports non linear workflows)

Created in 2005

**What is Git**

Git is a distributed version control system (tracks changes to content, provides a central point for collaboration)

Git allows for centralized administration (teams have controlled access scope, the main branch should always correspond to deployable code)

**GitHUB**

Github is an online hosting service for git repositories (hosted by a subsidiary of Microsoft, offers free and professional accounts)

Repository is a data structure for storing documents including application source code (can track and maintain version control)

**GitLab – enables developers to collaborate, work from a local copy, branch and merge code, steamline testing and delivery with CI/CD (continuous integration/continuous delivery)**

Devops platform delivered as a single application

Provides access to git repositories

Provides source code management