**Data Science Training**

**Day 1**

**Outline**

1. Non-technical
   * Excel
   * Power BI
2. Technical
   * SQL – RDBMS, MySQL
   * Python – EDA, packages, stats, linear regression, logistic regression, classification (SVM, decision tree), text mining, clustering, dimension reduction

**Software**

1. Microsoft Excel
2. Power BI
3. VSCode
4. Python

set path=%path%;address;

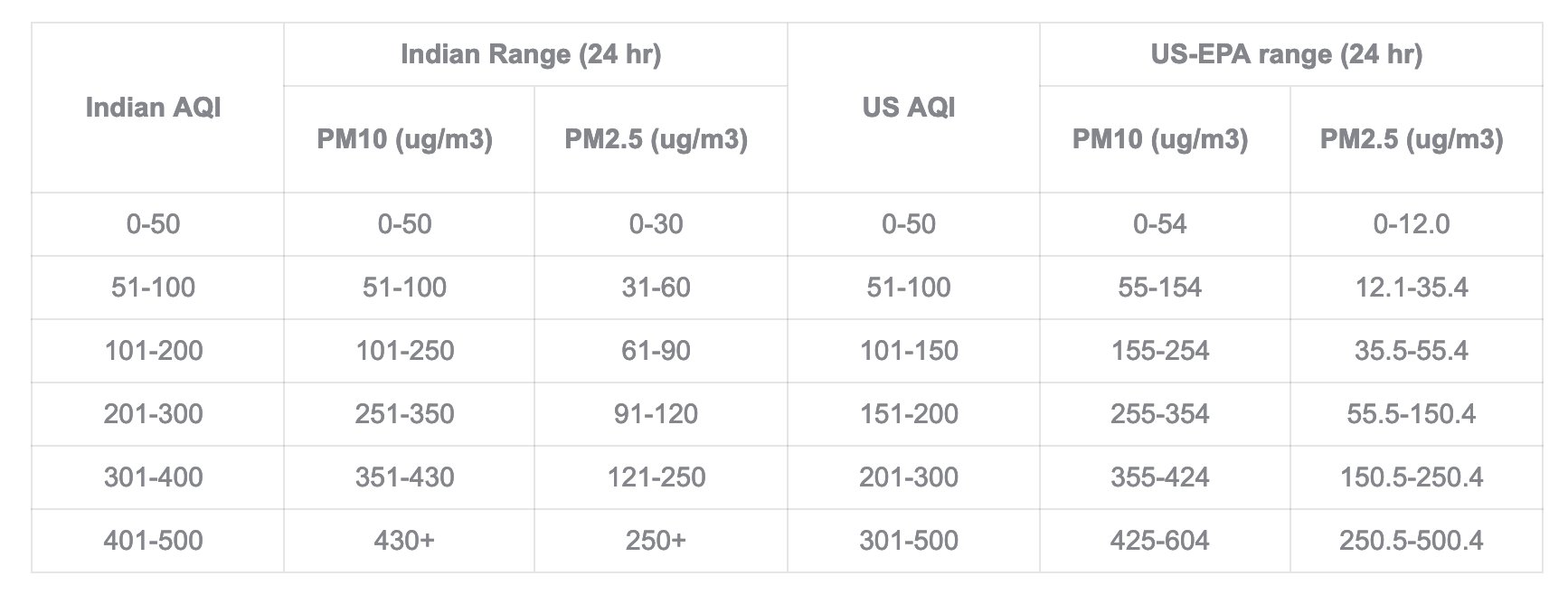
1. ANSI SQL – MySQL
   * Server – cloud (Azure/AWS)
   * Client – DBeaver (connect to Database)
2. Git
   * Server – GitHub
   * Client – Git SCM (software configuration management)

**Excel**

AQI – air quality index

* + Govt CPCB central pollution control board (NAQI)
  + 0 (good) – 500 (severe)
  + PM = particulate matter
  + CO, O3, NO2, SO2, Amonia NH3, Lead Pb

<https://www.pranaair.com/in/blog/what-is-air-quality-index-aqi-and-its-calculation/>



**ML Project**

1. Problem definition
2. Data collection
   * Structured – SQL
   * Unstructured – image, audio, video
   * Semi structured – JSON, CSV
3. Data Cleaning
   * Missing values
   * Outliers
   * Format inconsistencies
4. EDA
   * Analyze
   * Visualization
   * Understand patterns
   * Correlation
   * Potential biases
5. Feature Engineering
6. Feature Selection
7. Data Splitting – training, validation, testing
8. Model Selection
   * Choose algorithm
9. Training the Model
10. Hyperparameter Tunning
11. Model Selection for deployment
12. Model Testing
13. Deployment – application programming interface
14. Monitoring and Maintenance
15. Documentation

Power BI

* + Visualization
  + Analyze
  + Dashboard
  + Report

**VCS**

* + Repository
    1. src
    2. Output – artifact – s3, Jfrog
    3. Containers – hub, docker
    4. Dependencies – maven gradle
    5. System
    6. Update
  + Central – SVN
  + Distributed – Git/Bajaar/Marcial

|  |  |  |  |
| --- | --- | --- | --- |
| Local Repo | | | Remote |
| WD | Stage | LR | Branch |
| add |  | commit | push |
|  |  |  |  |

1. Origin: branch to local

**Day 2**

System

* + Application (process) – Python
  + Data (experience) – SQL

Server – GitHub

Client – GUI based, command based (init, remote, add, commit, push)

Authentication: key vault, PAT token

GitHub – settings – developers setting – PAT – generate

PAT token: ghp\_pOUKca2onTICf5vlTRfzgEElk30ban2D82rA

<https://github.com/yh2784/training.git>

https://oauth:ghp\_pOUKca2onTICf5vlTRfzgEElk30ban2D82rA@github.com/yh2784/training.git