

# Task 4.1

## Why Python is so popular among data analysts?

Python has gained widespread popularity among data analysts due to its versatility, ease of use, and extensive support for data manipulation, statistical analysis, and machine learning. There are some reasons:

- It is an open-source programming language, meaning it is free to use, and anyone can contribute to its development.
- It has vast libraries, that simplify data analysis tasks. These libraries offer pre-built functions for data cleaning, transformation, and visualization, streamlining the analysis process.
- Python is designed to handle large datasets beyond Excel's row limits, making it ideal for large-scale data processing
- Good compatibility with databases, APIs, and cloud platforms makes it a valuable tool in data engineering and data science pipelines.
- Large Community Support.

## Top 5 Companies Using Python

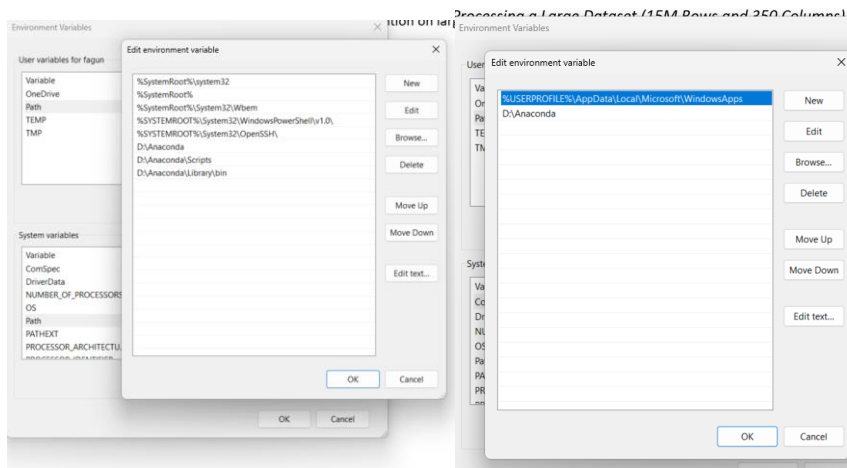
I could find five of the top companies globally recognized for their extensive use of Python:

1. **Google:** Uses Python for multiple purposes, including back-end web development, data analysis, and machine learning.
2. **Facebook (Meta):** Utilizes Python for its data infrastructure, especially for data manipulation, engineering, and machine learning models.
3. **Amazon:** Leverages Python for AI-based recommendations, data analytics, and data management processes within AWS.
4. **Netflix:** Employs Python extensively for data analysis, especially for A/B testing, algorithm personalization, and infrastructure monitoring.
5. **Spotify:** Uses Python for backend services and data analysis, particularly for its recommendation engine and analytics.


## Tool Recommendations for Data Scenarios

1. **Quick Tweaks and Minor Analysis with a Small Dataset**
  - **Tool:** Microsoft Excel.
  - **Reason:** For quick data manipulation and visualization, Excel or Google Sheets offer a straightforward, user-friendly interface. They support essential functions like filtering, sorting, and creating charts without the need to code.
2. **Scenario 2: Retrieve Data from a Very Large Database**
  - **Tool:** SQL – PostgreSQL.
  - **Reason:** SQL is ideal for retrieving, aggregating, and manipulating data directly within a database. Its structured querying capabilities are optimized for large datasets, allowing efficient extraction without needing to load the entire dataset.
3. **Scenario 3: Processing a Large Dataset (15M Rows and 350 Columns) for Advanced Analysis**
  - **Tool:** Python (using Pandas and Dask).
  - **Reason:** Python, with libraries like Pandas and Dask, is equipped for handling large datasets that exceed typical memory constraints. Dask, in particular, is designed for parallel computing and can process large datasets by splitting them into manageable parts. This setup allows for efficient data preparation and manipulation on large datasets.

## Environment Variables



# Jupyter

 jupyter

FileViewSettingsHelp

FilesRunning

Select items to perform actions on them.

NewUpload

/

<input type="checkbox"/> Name	Last Modified	File Size
<input type="checkbox"/> anaconda3	9 minutes ago	
<input type="checkbox"/> Contacts	10 months ago	
<input type="checkbox"/> Documents	10 months ago	
<input type="checkbox"/> Downloads	13 minutes ago	
<input type="checkbox"/> Favorites	10 months ago	
<input type="checkbox"/> iCloudDrive	10 hours ago	
<input type="checkbox"/> iCloudPhotos	10 hours ago	
<input type="checkbox"/> Links	10 months ago	
<input type="checkbox"/> Music	10 months ago	
<input type="checkbox"/> OneDrive	10 hours ago	
<input type="checkbox"/> Saved Games	10 months ago	
<input type="checkbox"/> Searches	10 months ago	
<input type="checkbox"/> Tracing	3 months ago	
<input type="checkbox"/> Videos	9 months ago	
<input type="checkbox"/> Untitled.ipynb	1 minute ago	72 B