

Data Analytics Tools

Dr. Ghodrat Moghadampour

mg@vamk.fi

Vaasa University of Applied Sciences



Outline

- Data Analytics
- Data Analytics Tools
- Main Categories
 - Programming Languages
 - Visualization Platforms
 - Big Data and Cloud Tools
 - Database Technologies

Data Analytics

- Data Analytics: the process of examining raw data to find useful information, trends, and patterns that help in decision making.
- Data analytics typically involves:
 - **Collecting data:** databases, sensors, logs, social media, etc.
 - **Cleaning and preparing data:** removing errors, formatting
 - **Analyzing data:** using statistics, queries, or machine learning
 - **Visualizing data:** charts, dashboards, reports
 - **Making decisions:** business strategies, process improvements, predictions

Data Analytics Tools

- Software and platforms used to **collect, process, analyze, and visualize data**
- Help transform **raw data into insights and decisions**
- Range from **simple spreadsheets** to advanced **machine learning frameworks**
- The backbone of modern analytics: **reduce complexity** and make **data-driven decisions** possible

Categories of Analytics Tools

- **Programming and Scripting:**
 - Python, R and SAS
- **Visualization and BI:**
 - Tableau, Power BI and Grafana
- **Big Data and Cloud:**
 - Spark, Hadoop, BigQuery and AWS Redshift
- **Databases:**
 - SQL, NoSQL like MongoDB and Cassandra
- **Specialized and Emerging:**
 - RapidMiner, KNIME and AutoML tools

Programming Languages

- **Python**, dominates due to ease and versatility:
 - **Pandas, NumPy, Scikit-learn and Matplotlib**
- **R**, popular in academia and statistics-heavy fields :
 - **ggplot2: an R package for producing statistical, or data, graphics**
 - **Caret R package for predictive modeling in R, streamlining the entire machine learning workflow from data preparation to model training, evaluation, and tuning**
- **SAS**, strong in enterprises :
 - **advanced statistical modeling for enterprise use**

Visualization and BI Tools

- Visualization tools make data accessible to non-technical users.
- Very useful for demonstration and decision-making:
 - **Tableau:**
 - Interactive dashboards, drag-and-drop, enterprise-ready
 - **Power BI:**
 - Microsoft ecosystem, easy Excel integration
 - **Grafana:**
 - Open-source, real-time monitoring and dashboards
 - **Google Data Studio:**
 - Cloud-native, integrates with Google products

Big Data and Cloud Tools

- Big data platforms are needed when dealing with millions or billions of records
- Cloud tools make it easier and cheaper to scale:
 - **Apache Hadoop:**
 - Distributed storage and processing
 - **Apache Spark:**
 - Fast big data processing, ML support
 - **Google BigQuery:**
 - Serverless, cloud data warehouse
 - **AWS Redshift:**
 - Scalable, fast queries in Amazon cloud
 - **Azure Synapse:**
 - Microsoft's analytics cloud platform

Databases

- The choice of database depends on the data type and scalability needs:
 - **Relational (SQL):**
 - for structured data:
 - MySQL
 - PostgreSQL
 - Oracle DB
 - **NoSQL:**
 - for flexible, large-scale, unstructured data:
 - MongoDB (document)
 - Cassandra (wide-column)
 - Neo4j (graph)

Specialized and Emerging Tools

- **RapidMiner:**
 - no-code analytics platform
- **KNIME:**
 - open-source workflow-based analytics
- **Google AutoML / Azure ML:**
 - Automated machine learning
- **Databricks:**
 - Unified platform for data, analytics, and ML

Data Analytics with Python

- We can use Python together with **Pandas** (for data manipulation) and **Matplotlib** and **Seaborn** for visualization) to analyze data.
- Prerequisites:
 - **Python**
 - **PIP (Python 3's alias pip3)**: a **package manager** (package management system) written in Python used to install and manage software packages.
 - **Pandas**: a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language.
 - **Matplotlib**: a comprehensive library **for creating static, animated, and interactive visualizations** in Python. Matplotlib makes easy things easy and hard things possible.
 - **Seaborn**: a Python **data visualization library** based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

Data Analytics with Python

- Prerequisites:

- **NumPy**: an open-source Python library **for numerical computing**, known for its powerful support for large, multi-dimensional arrays and matrices, and a collection of high-level mathematical functions to operate on them.
- **Scikit-learn**: provides an array of built-in metrics **for both classification and regression problems**, thereby aiding in the decision-making process regarding model optimization or model selection.
- **SciPy**: provides algorithms **for optimization, integration, interpolation, eigenvalue problems, algebraic equations, differential equations, statistics and many other classes of problems.**

Data Analytics with Python

- Prerequisites:
 - **Plotly**: an open-source **graphing library** and a technical computing company that provides tools for interactive data visualization and analysis.
 - **JupyterLab**: a next-generation, open-source, web-based platform from Project Jupyter that **acts as a flexible, extensible, and integrated development environment** for interactive computing.
 - **OpenPyXL** : a popular Python library used **for reading from and writing to Excel 2010 (.xlsx) files**, making it an essential tool for data analysis, automation, and reporting.

Data Analytics with Python

- Setting up the environment:
 - Download Python from python.org
 - `python --version`
 - Make sure to have **pip**:
 - `python -m ensurepip --upgrade`
 - `pip install package_name`
 - Install data analytics libraries
 - `pip install pandas numpy matplotlib seaborn scikit-learn scipy plotly jupyterlab openpyxl`
 - `pip list`
 - **Start JupyterLab**: great for exploration:
 - `jupyter lab`

Data Analytics with Python

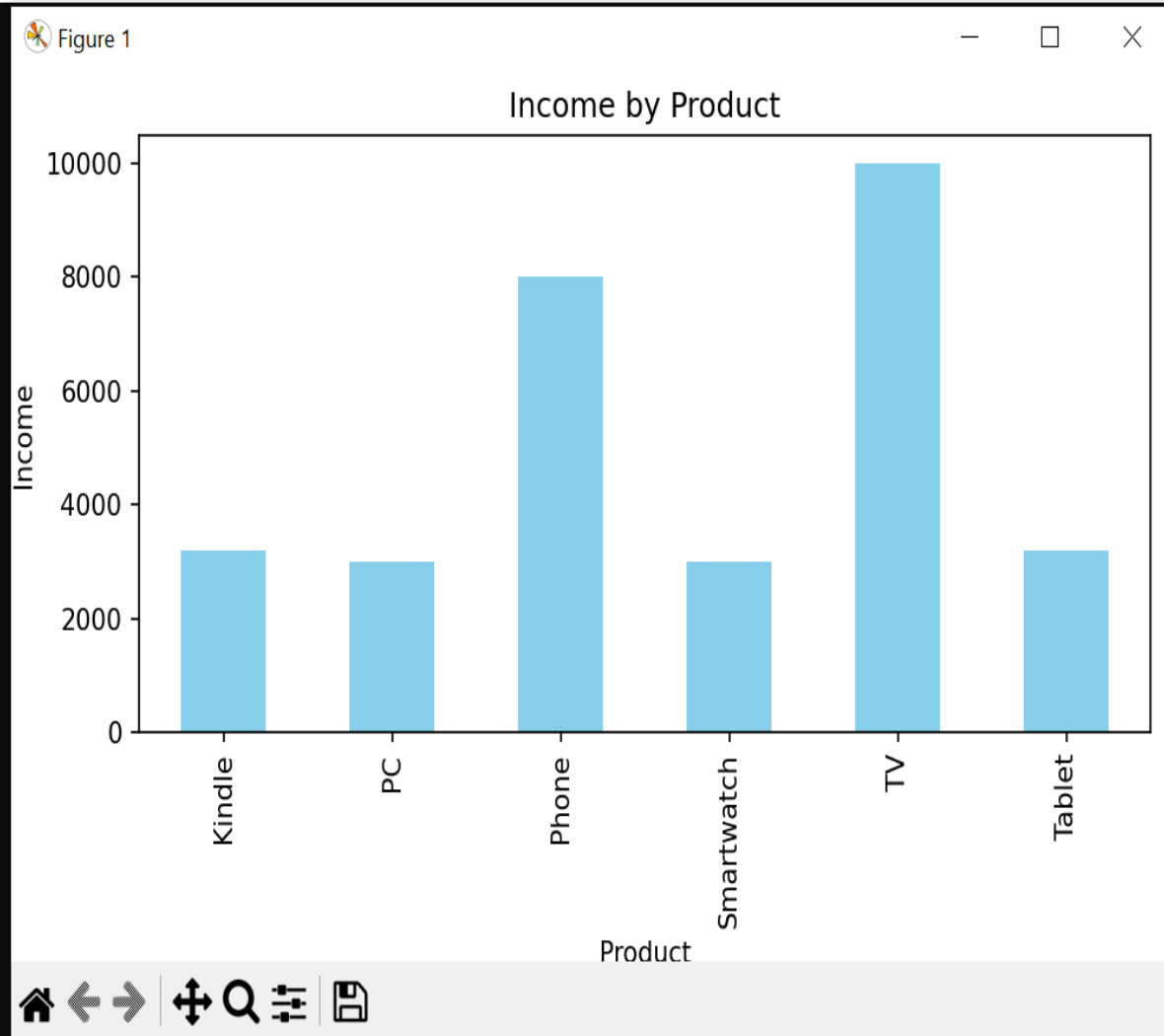
- Example 1:
 - This example shows how **pandas** and **pyplot** of **matplotlib** libraries can be used to draw statistics of sales data in CSV format, which consists of the following columns: **Date**, **Product**, **Sales** and **Income**.

Data Analytics with Python

Summary statistics:

	Sales	Income
count	8.000000	8.000000
mean	14.000000	3800.000000
std	11.006492	841.766849
min	2.000000	3000.000000
25%	4.750000	3150.000000
50%	11.500000	3600.000000
75%	23.000000	4250.000000
max	32.000000	5000.000000

Total Income: 30400



Data Analytics with Python

- Example 2:
 - This example shows how **pandas**, **numpy**, **pyplot** of **matplotlib**, **seaborn**, **ticker** from **matplotlib** and **FuncFormatter** libraries can be used to draw extensive statistics and diagrams from sales data in CSV format, which consists of the following columns: **Date**, **Product**, **Quantity**, **UnitPrice** and **TotalSale**.

Data Analytics with Python

```
C:\Users\mg\2025-2026\AmiES2025\code\ex2>python sales_analysis.py --input
sales.csv --outdir output_figures
```

```
----Data BASIC SUMMARY----
```

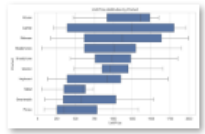
```
Rows: 100
```

```
Date range: 2025-01-01 to 2025-03-31
```

```
Products: 10
```

	Date	Product	...	UnitPrice	TotalSale
count	100	100	...	100.000000	100.000000
unique	NaN	10	...	NaN	NaN
top	NaN	Headphones	...	NaN	NaN
freq	NaN	19	...	NaN	NaN
mean	2025-02-15 01:26:24	NaN	...	910.900000	2751.420000
min	2025-01-01 00:00:00	NaN	...	51.000000	174.000000
25%	2025-01-24 00:00:00	NaN	...	462.250000	965.250000
50%	2025-02-15 12:00:00	NaN	...	851.000000	1636.000000
75%	2025-03-12 12:00:00	NaN	...	1287.250000	3873.000000

Data Analytics with Python



boxplot_unitprice_by_product.png



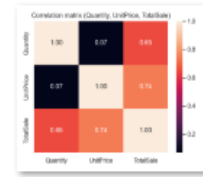
by_month.csv



by_product.csv



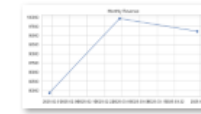
correlation_matrix.csv



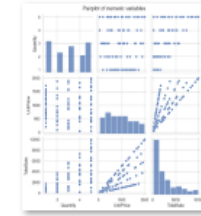
correlation_matrix_x.png



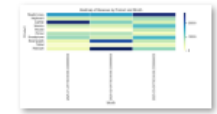
daily_revenue_time_series.png



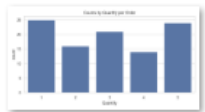
monthly_revenue.png



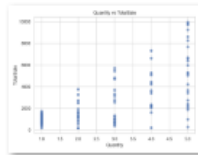
pairplot_numeric.png



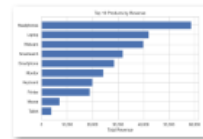
product_month_heatmap.png



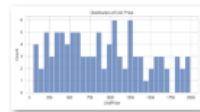
quantity_counts.png



quantity_vs_total_sale.png



top_10_products_revenue.png



unitprice_distribution.png



weekly_revenue_Headphones.png



weekly_revenue_Keyboard.png



weekly_revenue_Laptop.png



weekly_revenue_Monitor.png



weekly_revenue_Mouse.png



weekly_revenue_Printer.png



weekly_revenue_Smartphone.png



weekly_revenue_Smartwatch.png



weekly_revenue_Tablet.png



weekly_revenue_Webcam.png

Thanks for your attention 😊

