https://github.com/sfp3000/numpy\_pandas\_Pyspark\_tutorials.git

>git clone https://github.com/sfp3000/numpy\_pandas\_Pyspark\_tutorials.git

run the following to get the latest code. I'll be constantly adding code.

>git pull

Intro: my name is Saad Fazelpour and I'm going to be your teacher/tutor/instructor/study partner

sfp300@gmail.com

Saad Fazelpour

https://www.linkedin.com/in/saad-f-b947a411/

https://github.com/sfp3000/numpy\_pandas\_Pyspark\_tutorials.git

>git clone https://github.com/sfp3000/numpy\_pandas\_Pyspark\_tutorials.git

>git add --all

>git status

>git commit -m "Description of commit"

>git push -u origin main

**005**

**1-install python Done**

**2- install git Done**

**3-configure env variables Done**

**4-pip install numpy Done**

**5-pip install pandas Done**

**6-pip install pyspark Done**

**7-pip install findspark Done**

**8-pip install jupyter Done**

**9-pip install matplotlib Done**

**10-install MySQL later in the pandas section**

**11-Install mysql workbench**

**Navigate to your source folder and type:**

**> jupyter notebook**

# Sources of Data

Kaggle.com

<https://www.kaggle.com/datasets>

**Comprehensive List of Data Sources with URLs**

**🌐 Government & Official Data Portals**

**United States Government**

1. [**Data.gov**](https://data.gov/) - US government's open data
   * [https://data.gov](https://data.gov/)
   * 200,000+ datasets across all agencies
2. **US Census Bureau**
   * <https://www.census.gov/data.html>
   * Demographic, economic, and geographic data
3. **Bureau of Labor Statistics**
   * <https://www.bls.gov/data/>
   * Employment, pricing, productivity data
4. **CDC Data & Statistics**
   * <https://www.cdc.gov/datastatistics/>
   * Public health and disease data
5. **NASA Open Data**
   * <https://data.nasa.gov/>
   * Space, earth science, aeronautics data
6. **NOAA Data**
   * <https://www.noaa.gov/weather-climate-data>
   * Weather, climate, ocean data
7. **FDA Open Data**
   * <https://open.fda.gov/>
   * Drug, device, food data
8. **Energy Information Administration**
   * <https://www.eia.gov/opendata/>
   * Energy statistics and data

**International Government Data**

1. **European Union Open Data Portal**
   * <https://data.europa.eu/euodp/en/data/>
   * EU institutions and bodies data
2. **UK Government Data**
   * <https://data.gov.uk/>
   * UK public sector data
3. **World Bank Open Data**
   * <https://data.worldbank.org/>
   * Global development data
4. **UN Data**
   * <http://data.un.org/>
   * United Nations datasets
5. **IMF Data**
   * <https://www.imf.org/en/Data>
   * International financial statistics
6. **OECD Data**
   * <https://data.oecd.org/>
   * Economic and social data from developed countries

**📊 Data Science & ML Platforms**

1. **Kaggle Datasets**
   * <https://www.kaggle.com/datasets>
   * 50,000+ community-shared datasets
2. **UCI Machine Learning Repository**
   * <https://archive.ics.uci.edu/ml/index.php>
   * 600+ datasets for machine learning
3. **Google Dataset Search**
   * <https://datasetsearch.research.google.com/>
   * Search engine for datasets
4. **AWS Open Data Registry**
   * <https://registry.opendata.aws/>
   * Cloud-optimized datasets
5. **Microsoft Research Open Data**
   * <https://msropendata.com/>
   * Datasets from Microsoft research
6. **IBM Data Asset Exchange**
   * <https://developer.ibm.com/exchanges/data/>
   * Curated datasets for AI development

**💹 Financial & Economic Data**

1. **Yahoo Finance**
   * <https://finance.yahoo.com/>
   * Stock market data and news
2. **Quandl**
   * <https://www.quandl.com/>
   * Financial and economic data (now part of NASDAQ)
3. **Alpha Vantage**
   * <https://www.alphavantage.co/>
   * Free stock APIs and historical data
4. **FRED (Federal Reserve Economic Data)**
   * <https://fred.stlouisfed.org/>
   * 800,000+ economic time series
5. **Intrinio**
   * <https://intrinio.com/>
   * Financial data and APIs
6. **Tiingo**
   * <https://api.tiingo.com/>
   * Financial data platform

**📱 Social Media & Web Data**

1. **Twitter API**
   * <https://developer.twitter.com/en/docs/twitter-api>
   * Social media data access
2. **Reddit API**
   * <https://www.reddit.com/dev/api/>
   * Forum and discussion data
3. **Facebook Graph API**
   * <https://developers.facebook.com/docs/graph-api/>
   * Limited social data access
4. **Google Trends**
   * <https://trends.google.com/trends/>
   * Search trend data
5. **Common Crawl**
   * <https://commoncrawl.org/>
   * Open web crawl data
6. **Wikipedia API**
   * <https://www.mediawiki.org/wiki/API:Main_page>
   * Encyclopedia data access

**🗺️ Geospatial & Location Data**

1. **OpenStreetMap**
   * <https://www.openstreetmap.org/>
   * Crowdsourced map data
2. **Google Maps Platform**
   * <https://developers.google.com/maps>
   * Mapping and location APIs
3. **USGS EarthExplorer**
   * <https://earthexplorer.usgs.gov/>
   * Satellite imagery and data
4. **Sentinel Hub**
   * <https://www.sentinel-hub.com/>
   * Satellite data access
5. **Natural Earth Data**
   * <https://www.naturalearthdata.com/>
   * Cultural and physical vector data
6. **OpenAddresses**
   * <https://openaddresses.io/>
   * Global address data

**🏢 Business & Commerce Data**

1. **Crunchbase**
   * <https://www.crunchbase.com/>
   * Company and startup database
2. **Glassdoor API**
   * <https://www.glassdoor.com/developer/index.htm>
   * Company reviews and salary data
3. **Amazon Product Advertising API**
   * <https://affiliate-program.amazon.com/>
   * E-commerce product data
4. **Yelp Fusion API**
   * <https://www.yelp.com/developers>
   * Business reviews and ratings
5. **LinkedIn**
   * <https://developer.linkedin.com/>
   * Professional network data (limited access)

**🏥 Healthcare & Biomedical Data**

1. **NCBI Datasets**
   * <https://www.ncbi.nlm.nih.gov/datasets/>
   * Biological and genetic data
2. [**ClinicalTrials.gov**](https://clinicaltrials.gov/)
   * <https://clinicaltrials.gov/ct2/resources/download>
   * Clinical research data
3. **MIMIC Critical Care Database**
   * <https://mimic.mit.edu/>
   * Medical ICU data
4. **WHO Data**
   * <https://www.who.int/data>
   * World Health Organization data
5. **OpenNeuro**
   * <https://openneuro.org/>
   * Brain imaging data
6. **Cancer Imaging Archive**
   * <https://www.cancerimagingarchive.net/>
   * Cancer-related images and data

**🎓 Education & Research Data**

1. **Google Scholar**
   * <https://scholar.google.com/>
   * Academic paper metadata
2. **arXiv API**
   * <https://arxiv.org/help/api>
   * Scientific pre-print papers
3. [**Data.world**](https://data.world/)
   * <https://data.world/>
   * Collaborative data platform
4. **Figshare**
   * <https://figshare.com/>
   * Research data repository
5. **Zenodo**
   * <https://zenodo.org/>
   * Research data repository
6. **Harvard Dataverse**
   * <https://dataverse.harvard.edu/>
   * Research data repository

**⚡ Real-time & API Data**

1. **RapidAPI**
   * <https://rapidapi.com/>
   * API marketplace with 35,000+ APIs
2. **APILayer**
   * <https://apilayer.com/>
   * Various API services
3. **JSONPlaceholder**
   * <https://jsonplaceholder.typicode.com/>
   * Fake REST API for testing
4. **OpenWeatherMap**
   * <https://openweathermap.org/api>
   * Weather data API
5. **NewsAPI**
   * <https://newsapi.org/>
   * News articles API
6. **Twilio**
   * <https://www.twilio.com/docs/usage/api>
   * Communication APIs

**🎯 Specialized & Niche Data**

1. **SportsReference**
   * <https://www.sports-reference.com/>
   * Sports statistics
2. **IMDb Datasets**
   * <https://www.imdb.com/interfaces/>
   * Movie and TV data
3. **MusicBrainz**
   * <https://musicbrainz.org/doc/MusicBrainz_Database>
   * Music metadata
4. **Lichess Database**
   * <https://database.lichess.org/>
   * Chess game data
5. **GitHub Archive**
   * <https://www.gharchive.org/>
   * GitHub activity data
6. **DBPedia**
   * <http://wiki.dbpedia.org/>
   * Structured Wikipedia content
7. **FiveThirtyEight Data**
   * <https://data.fivethirtyeight.com/>
   * Politics, sports, science data

**📈 Market Research & Consumer Data**

1. **Pew Research Center**
   * <https://www.pewresearch.org/download-datasets/>
   * Social science research data
2. **Nielsen**
   * <https://www.nielsen.com/us/en/insights/>
   * Consumer behavior data
3. **Statista**
   * <https://www.statista.com/>
   * Market and consumer data
4. **Gallup**
   * <https://www.gallup.com/analytics/354374/gallup-analytics.aspx>
   * Public opinion data

**🔬 Scientific & Research Data**

1. **Dryad Digital Repository**
   * <https://datadryad.org/stash>
   * Scientific research data
2. **PANGAEA**
   * <https://www.pangaea.de/>
   * Earth & environmental science data
3. **Gene Expression Omnibus**
   * <https://www.ncbi.nlm.nih.gov/geo/>
   * Gene expression data
4. **Protein Data Bank**
   * <https://www.rcsb.org/>
   * 3D protein structure data

**🌍 Climate & Environmental Data**

1. **Copernicus Open Access Hub**
   * <https://scihub.copernicus.eu/>
   * Satellite observation data
2. **WorldClim**
   * <https://www.worldclim.org/>
   * Global climate data
3. **Global Biodiversity Information Facility**
   * <https://www.gbif.org/>
   * Biodiversity data
4. **Open Topography**
   * <https://opentopography.org/>
   * Topographic data

**💡 Key Categories Summary:**

* **Government**: Official statistics, economic indicators
* **Financial**: Markets, economic data, company information
* **Social**: Social media, web content, user-generated data
* **Geospatial**: Maps, satellite imagery, location data
* **Healthcare**: Medical, genetic, clinical trial data
* **Academic**: Research papers, scientific data, educational resources
* **Business**: Company data, market research, e-commerce
* **Real-time**: APIs, live data feeds, streaming data
* **Specialized**: Sports, entertainment, niche domains

# Data Formats/types/shapes

**1. Structured Data**

* **Tabular Data (rows & columns, fixed schema) → Example: SQL Tables, Excel Sheets →** [**https://en.wikipedia.org/wiki/Tabular\_data**](https://en.wikipedia.org/wiki/Tabular_data)
* **Relational Databases → Example: MySQL, PostgreSQL, Oracle DB →** [**https://en.wikipedia.org/wiki/Relational\_database**](https://en.wikipedia.org/wiki/Relational_database)
* **Spreadsheets → Example: Microsoft Excel, Google Sheets →** [**https://en.wikipedia.org/wiki/Spreadsheet**](https://en.wikipedia.org/wiki/Spreadsheet)

**2. Semi-Structured Data**

* **JSON (hierarchical, flexible schema) → Example: MongoDB stores JSON-like documents →** [**https://en.wikipedia.org/wiki/JSON**](https://en.wikipedia.org/wiki/JSON)
* **XML (tree-like structure for data exchange) → Example: RSS Feeds →** [**https://en.wikipedia.org/wiki/XML**](https://en.wikipedia.org/wiki/XML)
* **YAML (human-readable configuration format) → Example: Kubernetes Config Files →** [**https://en.wikipedia.org/wiki/YAML**](https://en.wikipedia.org/wiki/YAML)
* **CSV (flat files, comma-separated) → Example: Dataset exports →** [**https://en.wikipedia.org/wiki/Comma-separated\_values**](https://en.wikipedia.org/wiki/Comma-separated_values)
* **NoSQL Databases → Example: MongoDB, Cassandra, CouchDB →** [**https://en.wikipedia.org/wiki/NoSQL**](https://en.wikipedia.org/wiki/NoSQL)

**3. Unstructured Data**

* **Text Documents → Example: Word Docs (.docx), PDFs →** [**https://en.wikipedia.org/wiki/Text\_file**](https://en.wikipedia.org/wiki/Text_file)
* **Logs → Example: Web Server Logs, Application Logs →** [**https://en.wikipedia.org/wiki/Logfile**](https://en.wikipedia.org/wiki/Logfile)
* **Images → Example: JPG, PNG, TIFF →** [**https://en.wikipedia.org/wiki/Digital\_image**](https://en.wikipedia.org/wiki/Digital_image)
* **Audio → Example: MP3, WAV, FLAC →** [**https://en.wikipedia.org/wiki/Digital\_audio**](https://en.wikipedia.org/wiki/Digital_audio)
* **Video → Example: MP4, AVI, MKV →** [**https://en.wikipedia.org/wiki/Digital\_video**](https://en.wikipedia.org/wiki/Digital_video)

**4. Time-Oriented Data**

* **Time Series (data indexed by time) → Example: Stock Prices, Weather Data →** [**https://en.wikipedia.org/wiki/Time\_series**](https://en.wikipedia.org/wiki/Time_series)
* **Event Data (timestamped logs of events) → Example: Clickstream Data →** [**https://en.wikipedia.org/wiki/Event\_log**](https://en.wikipedia.org/wiki/Event_log)
* **Temporal Databases → Example: Oracle Temporal Data, SQL:2011 standard →** [**https://en.wikipedia.org/wiki/Temporal\_database**](https://en.wikipedia.org/wiki/Temporal_database)

**5. Graph & Network Data**

* **Graphs (nodes & edges) → Example: Neo4j Database →** [**https://en.wikipedia.org/wiki/Graph\_(abstract\_data\_type)**](https://en.wikipedia.org/wiki/Graph_(abstract_data_type))
* **Networks → Example: Computer Networks, Transportation Networks →** [**https://en.wikipedia.org/wiki/Network\_theory**](https://en.wikipedia.org/wiki/Network_theory)
* **Social Graphs → Example: Facebook Graph API →** [**https://en.wikipedia.org/wiki/Social\_graph**](https://en.wikipedia.org/wiki/Social_graph)

**6. Geospatial Data**

* **Raster Data (gridded pixels) → Example: Satellite Images, GeoTIFF →** [**https://en.wikipedia.org/wiki/Raster\_graphics**](https://en.wikipedia.org/wiki/Raster_graphics)
* **Vector Data (points, lines, polygons) → Example: Shapefiles in GIS →** [**https://en.wikipedia.org/wiki/Vector\_graphics**](https://en.wikipedia.org/wiki/Vector_graphics?utm_source=chatgpt.com)
* **GIS Data → Example: ArcGIS, QGIS Datasets →** [**https://en.wikipedia.org/wiki/Geographic\_information\_system**](https://en.wikipedia.org/wiki/Geographic_information_system)

**7. Scientific & Numeric Data**

* **Matrices → Example: NumPy Arrays →** [**https://en.wikipedia.org/wiki/Matrix\_(mathematics)**](https://en.wikipedia.org/wiki/Matrix_(mathematics))
* **Tensors → Example: TensorFlow, PyTorch Models →** [**https://en.wikipedia.org/wiki/Tensor**](https://en.wikipedia.org/wiki/Tensor)
* **Multidimensional Arrays → Example: HDF5 Data Storage →** [**https://en.wikipedia.org/wiki/N-dimensional\_array**](https://en.wikipedia.org/wiki/N-dimensional_array)

**8. Streaming & Real-Time Data**

* **Data Streams → Example: Apache Kafka, Flink →** [**https://en.wikipedia.org/wiki/Data\_stream**](https://en.wikipedia.org/wiki/Data_stream)
* **Sensor Data → Example: IoT Temperature Sensor Readings →** [**https://en.wikipedia.org/wiki/Sensor**](https://en.wikipedia.org/wiki/Sensor)
* **IoT Data → Example: Smart Home Devices (Nest, Alexa) →** [**https://en.wikipedia.org/wiki/Internet\_of\_things**](https://en.wikipedia.org/wiki/Internet_of_things)

**9. Big Data Formats**

* **Apache Parquet (columnar storage) → Example: Spark & Hadoop Datasets →** [**https://en.wikipedia.org/wiki/Apache\_Parquet**](https://en.wikipedia.org/wiki/Apache_Parquet)
* **Apache Avro (row-oriented serialization) → Example: Kafka Event Streaming →** [**https://en.wikipedia.org/wiki/Apache\_Avro**](https://en.wikipedia.org/wiki/Apache_Avro)
* **ORC (optimized row-columnar format) → Example: Hive Datasets →** [**https://en.wikipedia.org/wiki/Optimized\_Row\_Columnar**](https://en.wikipedia.org/wiki/Optimized_Row_Columnar)

**10. Specialized Data**

* **Genomic Data → Example: DNA Sequences (FASTA format) →** [**https://en.wikipedia.org/wiki/Genomics**](https://en.wikipedia.org/wiki/Genomics)
* **Medical Imaging → Example: MRI (DICOM Format) →** [**https://en.wikipedia.org/wiki/Medical\_imaging**](https://en.wikipedia.org/wiki/Medical_imaging)
* **Financial Data → Example: Bloomberg Feeds, SEC Filings →** [**https://en.wikipedia.org/wiki/Financial\_data**](https://en.wikipedia.org/wiki/Financial_data)

# Data jobs and salaries

| **Role** | **Typical Salary Range (USA)** | **Notes / Factors Driving Salary** |
| --- | --- | --- |
| **Data Analyst (Entry / Junior)** | **~$60,000 - $90,000** | **Basic analytics, dashboards, SQL; depends on domain + experience. (**[**Syntax Technologies**](https://www.syntaxtechs.com/blog/data-analytics-job-salaries/?utm_source=chatgpt.com)**)** |
| **Business Intelligence (BI) Analyst** | **~$80,000 - $130,000** | **Building reports, PowerBI/Tableau, database querying. (**[**Syntax Technologies**](https://www.syntaxtechs.com/blog/data-analytics-job-salaries/?utm_source=chatgpt.com)**)** |
| **Data Scientist (Mid-level)** | **~$110,000 - $180,000** | **Machine learning models, statistical analysis; depends on ML/AI exposure. (**[**Syntax Technologies**](https://www.syntaxtechs.com/blog/data-analytics-job-salaries/?utm_source=chatgpt.com)**)** |
| **Senior Data Scientist / Staff Scientist** | **~$160,000 - $220,000+** | **Leadership, specialized ML, large scale. Often includes bonus / stock. ([Refonte Learning](https://www.refontelearning.com/salary-guide/data-science-salaries-unveiled?utm_source=chatgpt.com" \o "Refonte Learning : Data Science Salary Guide 2025: Average Pay, Top Roles & Career Advice))** |
| **Data Engineer (Junior → Senior)** | **~$90,000 - $180,000+** | **Building pipelines, ETL, big data tools; senior roles with real-time/data scale pay more. (**[**Syntax Technologies**](https://www.syntaxtechs.com/blog/data-analytics-job-salaries/?utm_source=chatgpt.com)**)** |
| **Big Data Engineer / Distributed Systems Engineer** | **~$140,000 - $220,000+** | **Handling massive datasets, distributed computing (Spark, Hadoop, etc.). (**[**Syntax Technologies**](https://www.syntaxtechs.com/blog/data-analytics-job-salaries/?utm_source=chatgpt.com)**)** |
| **Machine Learning Engineer** | **~$120,000 - $200,000+** | **Building/deploying ML models, software engineering mix. ([Refonte Learning](https://www.refontelearning.com/salary-guide/data-science-salaries-unveiled?utm_source=chatgpt.com" \o "Refonte Learning : Data Science Salary Guide 2025: Average Pay, Top Roles & Career Advice))** |
| **AI / Deep Learning Engineer** | **Similar or higher than ML Engineer, often $150,000–250,000+ for experienced in niche (CV, NLP, etc.)** | **Specialized skills like DL, large models, etc. ([Refonte Learning](https://www.refontelearning.com/salary-guide/data-science-salaries-unveiled?utm_source=chatgpt.com" \o "Refonte Learning : Data Science Salary Guide 2025: Average Pay, Top Roles & Career Advice))** |
| **Quantitative Analyst (“Quant”)** | **~$130,000 - $200,000+** | **Finance / trading roles; math/stats heavy. Top quants make much more. (**[**Investopedia**](https://www.investopedia.com/articles/professionals/121615/quantitative-analyst-job-description-average-salary.asp?utm_source=chatgpt.com)**)** |
| **Data Architect** | **~$120,000 - $180,000+** | **Designing data infrastructure, schemas, storage. Senior architects command $$$. (**[**Salary.com**](https://www.salary.com/research/salary/listing/data-architect-salary?utm_source=chatgpt.com)**)** |
| **Business Intelligence Architect** | **~$130,000 - $200,000+** | **BI infrastructure, reporting strategy, dashboards at scale. ([Harnham](https://www.harnham.com/wp-content/uploads/2023/02/Harnham-Data-Analytics-Recruitment-US-Salary-Guide-2021.pdf?utm_source=chatgpt.com" \o "-))** |
| **Head of Data / Director of Data Science** | **~$180,000 - $300,000+** | **Leading teams, setting vision, managing high stakes/data budgets. ([Harnham](https://www.harnham.com/wp-content/uploads/2023/02/Harnham-Data-Analytics-Recruitment-US-Salary-Guide-2021.pdf?utm_source=chatgpt.com" \o "-))** |
| **Data Product Manager** | **~$115,000 - $200,000+** | **Bridging product & data science; defining data products. ([Refonte Learning](https://www.refontelearning.com/salary-guide/data-science-salaries-unveiled?utm_source=chatgpt.com" \o "Refonte Learning : Data Science Salary Guide 2025: Average Pay, Top Roles & Career Advice))** |
| **Data Governance / Privacy / Compliance Specialist** | **~$100,000 - $180,000+** | **Ensuring data policies, privacy regulation, ethics. Function critical in sensitive industries. ([Harnham](https://www.harnham.com/wp-content/uploads/2023/02/Harnham-Data-Analytics-Recruitment-US-Salary-Guide-2021.pdf?utm_source=chatgpt.com" \o "-))** |