YOHAN M MARKOSE
Boston, MA | markose.y@northeastern.edu | (857) 867-5489 | LinkedIn | Git Hub | Website

EDUCATION

Northeastern University | Boston, MA | Sep 2024 – May 2026 (Expected)

- ***Master of Science in Information Systems (GPA: 4.0/4.0)***
  Relevant Courses*: Data Science Engineering Methods and Tools****,*** *Big-Data Systems and Intelligence Analytics****,*** *Neural Modeling Methods and Tools****,*** *Programming Structures and Algorithm (PSA)*

Mar Athanasius College of Engineering (MACE) | Kerala, India | Aug 2016 – Jun 2020

- ***Bachelor of Technology in Mechanical Engineering***
  Relevant Courses: *Calculus, Linear Algebra and Complex Analysis, Probability Distributions*

SKILLS & CERTIFICATION

- **Languages:** Python, SQL, Java, C++, HTML5, CSS3
- **Frameworks & Tools:** Apache Airflow, Docker, MCP, Git, GitHub, Microsoft Excel, Power BI, Pinecone, ChromaDB, Mistral OCR
- **Databases:** Snowflake, MySQL, dbt, Redis Streams
- **Cloude Technologies:** GCP, Google Cloud Run, AWS S3, GitHub Actions, Google Compute Engine
- **Python Libraries:** NumPy, Pandas, Matplotlib, Scikit-learn, TensorFlow, SQLAlchemy ,Beautiful Soup, Selenium, LangChain, FastAPI, Streamlit, LiteLLM, LangGraph, spaCy, PyMuPDF, snowflake-connector-python
- **Certifications:** Microsoft Certified: Power BI Data Analyst Associate

PROFESSIONAL EXPERIENCE

IQVIA | Kochi, Kerala, India
Software Developer **| Oct 2020 – Apr 2023**

- Built **LangChain** and vector database **internal chatbot** using **FastAPI** to accelerate pharmaceutical domain knowledge of drug sales behavior, reducing manual research time for analysis from hours to minuets
- Extracted, cleaned, and transformed large datasets (500K+ records) from multiple pharmaceutical data sources (API, Dashboards, Flat Files) using Python (Pandas, NumPy) to create training datasets for quarterly **sales predictive models**, ensuring data quality and **feature engineering** for optimal model performance
- Engineered end-to-end automated **ETL pipelines in Python** and SQL (Snowflake) with **scikit-learn statistical validation** and anomaly detection, reducing data preparation and analysis time from 10 hours to 10 minutes per cycle
- Built **time series forecasting models using** Python (**scikit-learn**, Pandas) to predict quarterly sales performance and identify potential revenue decline patterns, achieving 85% forecast accuracy
- **Led the automation team** (20 members) in my department, streamlining processes and reducing manual data processing and analysis tasks by 250+ hours/month
- **Collaborated with internal stakeholders** to understand data needs and deliver insights with 100% on-time delivery
- Delivered 15+ monthly reports through data analysis and validation using **Excel, Power BI** (integrating multiple data sources with Power Query) and **SQL**, ensuring 100% on-time delivery and action points to stakeholders

Software Developer- Intern **| Jan 2020 – Apr 2020**

- Developed Python scripts to integrate data from multiple environments and **analyze anomalies using NumPy, Pandas and Matplotlib libraries** reducing data processing time by 80%
- Transformed raw data into actionable business insights using Python and Excel, enabling data-driven decision making for pharmaceutical client projects

PROJECTS

Electric Cars - Impact Analysis: ***Bayesian Statistics, NumPy, Pandas, Matplotlib, Data Visualization*****|** **GitHub Link**

- Developed **Bayesian statistical models** using Gamma and Student-T distributions with change-point detection to analyze EV impact on China's transportation emissions, identifying key transition periods (2005-2009, 2022) with 94% confidence
- Built predictive models forecasting emissions based on EV adoption targets (till 2030), revealing potential 188+ million ton reductions while uncovering that lithium production surge could offset EV benefits, providing actionable policy insights

Venture-Scope (Multi Agent - Agentic RAG): MCP, ***LangGraph, Pincecone , CI/CD Pipeline, LLM, AWS S3*****|** **GitHub Link**

- Architected **AI Ops platform** with **multi-agent** orchestration for automated business intelligence, implementing **CI/CD deployment** pipelines and real-time monitoring workflows helping entrepreneurs make data-driven decisions
- Deployed scalable cloud infrastructure integrating automated workflows with AWS S3 storage, containerized architecture, and comprehensive system monitoring for nationwide business accelerators

Snowflake Pipeline - FRED: ***Snowflake, Snowpark, CI/CD Pipeline, Tasks(DAGs), Github Actions, AWS S3*** **|** **GitHub Link**

- Engineered **end-end orchestrated pipeline** tracking U.S. Treasury yield curves using Federal Reserve data with real-time extraction, scheduled processing, and cloud storage integration
- Created **interactive dashboard** displaying yield curve inversions and economic indicators, enabling financial analysts to monitor market conditions and recession predictors effectively