# Rishi Guptha Mankala

Stony Brook, NY | P: +1 9342633087 | rishiguptha.mankala@stonybrook.edu | Linkedin

## EDUCATION

**STONY BROOK UNIVERSITY** — Stony Brook, NY

Masters of Science in Data Science GPA: 3.50/4.0 — Jan 2024 - Present

- Coursework: Statistical Computing, Statistical Learning, Data Analysis, Big Data Systems, Data Management

## TECHNICAL SKILLS

**Programming Languages**: Python(Pandas, Numpy), R, SQL (MySQL, PostgreSQL)
**Data Science & ML:** TensorFlow, PyTorch, Scikit-Learn, Pandas, NumPy, LLMs, Deep Learning
**Cloud** : AWS (S3, EC2, Glue, Redshift), Google Cloud (Cloud Storage, BigQuery, VertexAI)
**Big Data Technologies/Frameworks:** Snowflake, Airflow, Apache Spark, DBT(Data Build Tool),Data Modelling
**Data Visualization:** Power BI, Tableau, Looker Studio, AWS Quick Sight, Matplotlib
**Tools**: Collibra, Linux, Docker, Astronomer, Git, GitHub, VS Code, Google Colab, Excel, Google Sheets

## WORK EXPERIENCE

**KIMBERLY CLARK** — Bengaluru, India

Data & Analytics Intern — Jan 2023 – Dec 2023

- **Data Governance & Automation:** Earned 7 Collibra certifications (governance, lineage, cataloging) and automated metadata tagging with Python, accelerating data discovery and boosting team productivity by 25%.
- **Workflow Engineering:** Designed 5 Python/PostgreSQL workflows in Collibra to automate validation processes, reducing manual effort by 30% and enabling real-time analytics for stakeholders.
- **Data Migration & Integrity:** Led end-to-end migration of 10,000+ data assets into Collibra's Physical Data Dictionary via Excel ETL, achieving 100% data integrity and enhancing cross-departmental reporting efficiency.
- **Data Quality Optimization:** Partnered with BI/ML teams to resolve 15% of data inconsistencies in pipelines, improving model accuracy through schema standardization for critical interfaces.
- **Exploratory Analysis:** Conducted EDA on legacy datasets, identifying 20+ anomalies and cleanup strategies that reduced quarterly reporting errors by 18%.
- **Stakeholder Collaboration:** Built interactive Power BI dashboards to visualize Collibra metadata trends, empowering leadership to track governance KPIs and prioritize high-impact domains.

## PROJECTS

**NYC Taxi Insights**

- **ETL Pipeline Development**: Architected a scalable pipeline using Google Cloud Composer and Airflow, processing **10M+ records** with **100% data consistency**, enabling reliable analysis of fare trends and demand patterns.
- **Data Modeling & Visualization**: Optimized BigQuery performance through star schema design and Looker dashboards, **accelerating insights by 35%** for NYC transportation strategy adjustments.
- **Predictive Analytics**: Engineered fare prediction models (LightGBM) with **85% accuracy**, identifying peak pricing periods to inform dynamic pricing recommendations.

**Sales Insights Dashboard**

- **Centralized Reporting System**: Built interactive Power BI dashboards using SQL and DAX, unifying sales data from 5+ sources and **improving forecast accuracy by 30%** through star schema modeling.
- **Stakeholder Engagement**: Designed dynamic visualizations (trend analysis, geospatial mapping) that drove a **20% increase in engagement** during executive reviews.
- **Data Security**: Implemented row-level security protocols to restrict access to sensitive sales metrics, ensuring compliance while maintaining **15% faster forecast cycles**.

**Spotify ETL Pipeline**

- **Automated Data Ingestion**: Developed serverless AWS infrastructure (Lambda, Glue) to process **18k+ records annually**, **reducing manual effort by 90%** and enabling daily trend analysis for marketing teams.
- **Query Optimization**: Leveraged Athena and Glue crawlers to streamline playlist analytics, achieving **40% faster processing** for user behavior insights.

**Statistical Ensemble Learning R Package**

- **Efficiency-Driven Tool:** Engineered an R package integrating 5+ ML models, reducing prediction time by 30% for high-dimensional datasets and enabling faster exploratory analysis.
- **Feature Importance Analysis:** Generated variable importance scores for 1,000+ predictors, helping teams prioritize key drivers in customer segmentation models.

**Political Sentiment Analysis**

- **Unstructured Data Insights:** Analyzed 86k+ tweets using NLP (TextBlob) and logistic regression, forecasting election outcomes with 67% accuracy and identifying swing issues for campaign strategies.
- **Process Optimization:** Streamlined tweet processing to 1,000 tweets/second through efficient data mapping, enabling real-time sentiment tracking during debates.