

SUMANTEJA

Data Analyst

✉ goudsumanteja@gmail.com



ABOUT ME

Data Analyst with **5+** years of experience in extracting insights from complex data to support business decisions. Skilled in statistical analysis, data visualization, and process optimization to drive efficiency and growth.

EDUCATION

Masters in Business Analytics,
University of Hartford

PROFILE SUMMARY

- Over **5+ years** of experience in data analysis, specializing in extracting insights from structured and unstructured data using **SQL, Python, and R**, with a focus on automation and business intelligence.
- Expert in crafting dynamic dashboards and visual reports using **Tableau, Power BI, and Matplotlib**, enabling informed decision-making through clear and interactive data presentations.
- Proficient in end-to-end ETL processes using tools like **Alteryx, Talend, and Apache NiFi** for cleansing, transforming, and loading large volumes of data across multiple systems.
- Skilled in applying statistical modelling, **machine learning** algorithms, and predictive analytics using **Scikit-learn, TensorFlow, and Keras**, and processing large datasets in **Hadoop, Hive, and Spark** environments.
- Experienced with cloud platforms such as **AWS, Azure, and Google Cloud**, implementing scalable data solutions and working with services like **Redshift, BigQuery, Glue, and S3**.
- Hands-on expertise in **Snowflake, Amazon Redshift, and Google BigQuery** for data warehousing, combined with **Git** for version control and **Apache Airflow** for workflow orchestration.
- Developed and maintained data pipelines using **Apache NiFi**, and automated reporting and transformation tasks with **PowerShell** and **Python** scripting.
- Strong background in data wrangling and visualization with **Pandas, NumPy, Seaborn, and SciPy**, used extensively for statistical analysis and performance reporting.
- Worked with **Jupyter Notebooks, Looker, and Google Data Studio** to build and share analytics insights across cross-functional teams and executive stakeholders.
- Familiar with real-time data analytics using **Elasticsearch** and **Kibana**, and experienced in using **Presto/Trino** for querying distributed data sources efficiently.
- Built and monitored CI/CD pipelines with **Jenkins**, containerized analytics environments using **Docker**, and implemented predictive workflows with **KNIME**.
- Experienced in working with **Databricks** and **MLflow** for collaborative data science, including model development, tracking, and deployment in production environments.

TECHNICAL SKILLS

- **Data Analysis & Visualization:** Tableau, Power BI, Looker, Google Data Studio, Matplotlib, Seaborn, Plotly, D3.js, Chart.js, Excel (Advanced), Google Sheets
- **Programming Languages:** Python (Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn).
- **SQL** (MySQL, PostgreSQL, SQL Server, Oracle, SQLite, JavaScript)
- **Database Management:** MySQL, PostgreSQL, SQL Server, Oracle, NoSQL databases (MongoDB, Cassandra)
- **Data Warehousing:** Snowflake, Amazon Redshift, Google Big Query
- **ETL Tools:** Informatica, Talend, SSIS, Apache NiFi, Pentaho
- **Big Data & Cloud Technologies:** Hadoop, Apache Spark, Kafka, AWS (S3, Redshift, Athena, Lambda), Azure (Azure Data Factory, Synapse Analytics, Databricks), Google Cloud Platform (GCP) (Big Query, Dataflow, Pub/Sub)
- **Machine Learning & Statistical Analysis:** Scikit-learn, TensorFlow, Keras, PyTorch, XGBoost
- **Statistical Analysis:** Hypothesis testing, A/B testing, Regression analysis
- **Data Engineering & Automation:** Apache Airflow, Jenkins, Docker, Kubernetes (CI/CD), DataOps, Git, GitHub
- **Data Integration & APIs:** RESTful APIs, GraphQL, Postman, Swagger
- **Web Scraping:** BeautifulSoup, Scrapy, Selenium
- **Other Tools & Technologies:** Jupyter Notebooks, Google Collab, Elasticsearch, Kibana (ELK Stack), JDBC/ODBC.

WORK EXPERIENCE

W. R. Berkley Corporation, Greenwich, Connecticut

Senior Data Analyst

Jan 2025 – Present

Description: W. R. Berkley Corporation is an American commercial lines property and casualty insurance holding company. Built and managed interactive dashboards using Tableau and Power BI to visualize KPIs related to cost, efficiency, and performance. Delivered clear, data-driven insights to support strategic decisions at W. R. Berkley Corporation.

Responsibilities

- Designed and optimized relational data models and **SQL-based pipelines**, implementing scalable **ETL workflows** using **Informatica** and **Apache NiFi** to process streaming and transactional financial data.
- Architected big data frameworks with **Apache Spark** and **Hadoop**, enabling high-performance computing for large-scale credit risk assessment and customer segmentation projects.
- Centralized data storage by developing cloud-based warehousing on **Snowflake** and **Amazon Redshift**, integrating mortgage, loan, and investment datasets for consolidated reporting and analytics.
- Created real-time data streaming infrastructures using **Apache Kafka** and **AWS Kinesis**, enabling live monitoring and alerting for transactions, fraud prevention, and account behavior.
- Automated batch and real-time data workflows with **Apache Airflow**, orchestrating jobs to support financial forecasting, reporting, and regulatory compliance processes.
- Enhanced system interoperability by developing integrations with **REST APIs** and **GraphQL**, enabling seamless communication between mobile, web, and core banking systems.
- Built and deployed predictive models using **TensorFlow** and **Scikit-learn**, while applying **Python**, **Pandas**, **NumPy**, and **SciPy** for statistical analytics that support customer retention and cross-sell initiatives.
- Implemented containerized deployment with **Docker**, managed orchestration using **Kubernetes**, ensuring high availability across development, testing, and production environments.
- Established **CI/CD pipelines** via **Jenkins** to streamline deployment of analytics applications and fraud detection models, reducing time-to-production.
- Developed dynamic dashboards in **Power BI**, **Tableau**, **Looker**, and **Excel**, tracking KPIs such as loan quality, acquisition cost, and transaction health for executive stakeholders.
- Leveraged **Java Spring Boot** microservices with **Apache Kafka** to support event-driven processing, and used **AWS Lambda** and **Apache Camel** to enable near-real-time synchronization with cloud analytics systems.
- Managed high-speed unstructured data ingestion using **MongoDB** and **Cassandra**, and powered search analytics and anomaly detection with **Elasticsearch** for user activity and log data.
- Streamlined finance and operational reporting via **Excel Macros** and **VBA scripting**, significantly reducing manual effort in data consolidation and ensuring high accuracy in repeated report generation.

Environment: SQL, Informatica, Apache NiFi, Apache Spark, Hadoop, Snowflake, Amazon Redshift, Apache Kafka, AWS Kinesis, Apache Airflow, REST APIs, Python, TensorFlow, Docker, Kubernetes.

FactSet Research Systems, Norwalk, Connecticut

Senior Data Analyst

Jul 2024 - Dec 2024

Description: FactSet Research Systems Inc., is an American financial data and software company. Collected and analysed data from diverse sources using statistical techniques to identify trends and anomalies. Investigated root causes of issues and provided data-driven recommendations for effective solutions.

Responsibilities

- Automated and managed **ETL pipelines** using **Informatica**, **Apache NiFi**, and **Talend** to integrate CRM, actuarial, and policy data efficiently for analytics and reporting.
- Optimized large-scale claims and risk processing workflows leveraging **Hadoop**, **Hive**, and **Apache Spark** to enhance system performance and scalability.
- Managed structured and semi-structured policyholder data using **MongoDB**, improving accessibility for underwriting, compliance, and marketing analysis.
- Built real-time financial monitoring solutions with **Apache Kafka**, supported by **SQL** alerts and **Jenkins** automation, ensuring compliance in annuity transaction tracking.

- Developed cloud-based data warehousing on **AWS Glue**, **Amazon Redshift**, and **S3**, enabling scalable analytics for financial and actuarial teams, and created predictive models with **Prophet** and **Dask** to forecast claims, customer churn, and annuity fund performance, enhancing strategic planning.
- Applied machine learning using **Scikit-learn**, **TensorFlow**, and **Keras** for fraud detection, customer segmentation, and annuity surrender predictions.
- Delivered interactive dashboards and self-service analytics via **Tableau**, **Power BI**, and **Looker**, empowering stakeholders with real-time insights.
- Ensured data governance and streamlined deployment by integrating **Collibra**, **Terraform**, **Jenkins**, and **GitLab**, alongside containerizing applications with **Docker** and orchestrating with **Kubernetes**.

Environment: Informatica, Apache NiFi, Talend, Hadoop, Hive, Apache Spark, MongoDB, Apache Kafka, AWS Redshift, Docker.

Sundaram Finance Limited(MerDevelop), Hyderabad, India

Data Analyst

Aug 2021 – Jul 2023

Description: Sundaram Finance Limited is an Indian financial and investment service provider. Managed databases for efficient data access, while analysing financial data to detect risks and improve fraud detection. Supported product development, pricing, and marketing through data-driven insights.

Responsibilities

- Extracted and consolidated data from various insurance systems using **SQL**, **Python**, and **APIs** to support pricing analysis, claims tracking, and policy performance reviews.
- Conducted through data cleaning and transformation using **Pandas**, **NumPy**, **R**, and **SQL**, ensuring accurate datasets for underwriting and actuarial projections.
- Developed insurance risk and forecasting models using **R**, **Python**, and **SAS**, enabling advanced insights into claims behaviour and portfolio performance.
- Built interactive dashboards with **Tableau**, **Power BI**, and **Excel** to visualize KPIs such as claim turnaround time, premium collections, and policy renewals.
- Managed scalable data storage with **Amazon Redshift**, **Google BigQuery**, and **AWS S3**, and streamlined data integration using **Apache Airflow**, **Talend**, and **AWS Glue**.
- Applied machine learning algorithms via **Scikit-learn**, **TensorFlow**, and **Keras** to detect potential fraud and predict healthcare claim costs for better underwriting outcomes.
- Automated regulatory reporting and analytics with **Python**, **Matplotlib**, and **Looker**, ensuring IRDAI compliance and enabling self-service insights for claims, sales, and actuarial teams.

Environment: Python, SQL, R, Tableau, Power BI, Amazon Redshift, Google BigQuery, AWS S3, Apache Airflow, Talend

Care Health Insurance (Cognizant), Hyderabad, India

Programmer Analyst

May 2019 – Jul 2021

Description: Care Health Insurance is a specialized health insurance provider. Designed and managed reliable databases to ensure accurate, accessible health insurance data. Built predictive models to identify future trends and developed interactive dashboards and visualizations to present actionable insights to business teams and stakeholders.

Responsibilities

- Extracted and combined policyholder, claims, and provider data using **SQL**, **Python**, and **REST APIs**, supporting analytics for pricing models, fraud detection, and performance monitoring.
- Cleaned and transformed raw datasets using **Pandas**, **NumPy**, and **R**, improving data accuracy for actuarial risk models and underwriting insights across multiple health insurance products.
- Built and maintained interactive dashboards with **Tableau**, **Power BI**, and **Excel**, visualizing KPIs such as claim turnaround time, premium trends, and customer satisfaction metrics.
- Developed automated ETL workflows using **Apache Airflow**, **AWS Glue**, and **Talend**, integrating data from hospital networks, CRM systems, and internal claim engines into **Amazon Redshift** and **Google BigQuery**.
- Supported regulatory and internal reporting by creating scalable models in **Python**, using libraries like **Scikit-learn** and **Matplotlib**, and enabled teams with self-service dashboards via **Looker** and **Snowflake**.

Environment: SQL Server, Python, R, Tableau, Power BI, Apache Airflow, AWS Glue, Amazon Redshift, Google BigQuery, Snowflake.