Harsh Bajpai

+1(669) 204-2897 • San Francisco, CA

bajpaih03@gmail.com • LinkedIn • GitHub • Tableau • Portfolio

Professional Summary

Analytical and detail-oriented aspiring Data Analyst/Business Intelligence Analyst with a strong foundation in statistics, data visualization, and machine learning. Currently pursuing an M.S. in Statistical Data Science (Graduating May 2026). Skilled in uncovering insights from complex datasets through storytelling, dashboards, and predictive models. Strong communicator with a passion for simplifying data for non-technical audiences.

Education

Master of Science in Statistical Data Science — San Francisco State University *Expected July 2026*
GPA: 3.71/4.0

Relevant Coursework: Statistical Learning, Data Mining, Computational Statistics, Pattern Analysis, Machine Learning

Bachelor of Technology in Electronics and Communication — RGPV University *June 2024*
GPA: 8.98/10.0 (US Equivalent: 3.7/4.0)

Relevant Coursework: Data Structures and Algorithms, Embedded Systems, Signal Processing, IoT Development

Technical Skills

- **Programming:** Python, SQL(Structured Query Language), R, C, MATLAB
- **Data Analytics & Visualization:** Tableau, Excel (Advanced Excel, VLOOKUP, PivotTables, charts, Power Query, DAX, Macros, VBA), Matplotlib, Seaborn
- **Libraries & Machine Learning :** Pandas, Numpy, Scikit-learn, Hypothesis testing, A/B Testing, Regression Analysis, Time Series Analysis, Clustering
- **Tools & Platforms:** Jira, Jupyter Notebook, VS Code, Google Colab, R Studio, Scrum, Kanban
- **API's:** Flask API, Google Cloud APIs

Projects

- **NYC Yellow Taxi Fare & Total Amount Prediction [Live Demo]**
  - Analyzed 6.4M+ NYC Yellow Taxi records (18 features); performed data cleaning, feature extraction, and exploratory data analysis (EDA) to uncover key fare patterns
  - Trained and evaluated multiple regression models (Linear, Ridge, Lasso, Decision Tree, Random Forest) using MAE and R^2 metrics for model selection
  - Built a hybrid prediction system combining Machine Learning (ML)-based metered fare estimation and rule-based JFK flat-rate logic, including dynamic surcharges
  - Deployed a full-stack web app using Flask, Geocoding API, Directions API, Places API, Time Zone API, Google Maps API for real-time predictions with map-based trip input
- **Netflix Data-Driven Analysis [GitHub]**
  - Analyzed 9,800+ Netflix-style movie records to uncover genre trends, audience ratings, and content patterns using structured and text data.
  - Performed extensive feature engineering (release decade, genre count, overview word count) and EDA to identify rating-popularity dynamics.
  - Conducted clustering using KMeans on numeric and TF-IDF-transformed text data to group movies by performance and narrative themes
  - Applied ANOVA and Tukey's HSD tests to assess statistically significant genre effects on ratings and popularity, hightlighting content-performance relationships.
- **Hotel Booking Cancellation Analysis [GitHub]**
  - Explored 119K+ bookings across 36 features to identify patterns in cancellations for City vs Resort hotels
  - Found that cancellations (37%) were higher for OTA bookings, during January, & when ADR was elevated.
  - Recommended pricing adjustments, direct booking incentives, and country-specific strategies to reduce cancellation rates.
- **Spotify SQL Analysis [GitHub]**
  - Answered business questions using PostgreSQL across 20K+ track records. Investigated artist popularity, track features, and YouTube-Spotify cross-platform engagement.Created summary reports with CTEs, joins, aggregates, and window functions.