Summer Internship Report

On

"Data Analysis"

AIML306 – Summer Internship - I

Prepared by

Siddhraj Thakor (23AIML070)

Under the Supervision of

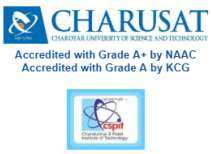
Dr. Spoorthy V

Submitted to

CHARUSAT

For partial fulfillment of B.Tech. (Semester 5)

Submitted at



CSPIT, CHARUSAT

Department of AI & ML

Changa, Anand – 388421

Auguest 2025
July 10, 2025

Accredited with Grade A+ by NAAC

Accredited with Grade A by KCG

CERTIFICATE

This is to certify that the report entitled **"Data Analysis"** is a bonafide work carried out by **Siddhraj Thakor (23AIML070)** under the guidance and supervision of **Dr. Spoorthy V** and **Mr. Prajyot Tuganokar** for the subject **Summer Internship – I (IT346)** of 5th Semester of Bachelor of Technology in the **Department of Artificial Intelligence and Machine Learning** at **Chandubhai S. Patel Institute of Technology (CSPIT)**, **Faculty of Technology & Engineering (FTE) – CHARUSAT**, Gujarat.

To the best of my knowledge and belief, this work embodies the student's own effort, has duly been completed, and fulfills the requirements of the B.Tech. degree. It meets the standard in respect of content, presentation, and language for submission to the examiner(s).

Under the Supervision of

Dr.Spoorty V
Councellar
Department of AI & ML
CSPIT, FTE, CHARUSAT
Changa, Gujarat

Mr.Prajyot Tuganokar
Project coordinator
Data Analyst
Samatrix Consulting Private Limited

Dr. Nirav Bhatt
Head of Department (AIML)
CHARUSAT, Changa, Gujarat

Chandubhai S. Patel Institute of Technology (CSPIT)
Faculty of Technology & Engineering (FTE), CHARUSAT
At: Changa, Ta. Petlad, Dist. Anand, Pin: 388421. Gujarat.

Samatrix.io

Date: 16-06-2025

Internship Completion Certificate

This is to certify that **Siddhraj Anil Kumar Thakor,** a student of **Charotar University of Science and Technology**, has successfully completed a **4-week Summer Internship** at **Samatrix Consulting Pvt Ltd** from **19-05-2025 to 13-06-2025**.

During the internship, he worked on **Data Analysis projects** involving the application of **statistical techniques and tools**. His responsibilities included cleaning, analyzing, and interpreting data to derive meaningful insights, contributing effectively to the team's objectives.

He demonstrated a good grasp of key statistical concepts and tools, showed strong analytical thinking, and maintained a professional attitude throughout the internship.
We acknowledge his contribution and wish him continued success in future academic and professional endeavors.

Yours sincerely,
For SAMATRIX CONSULTING PVT. LTD.

Authorized Signatory

Samatrix Consulting Private Limited
311, Vipul Trade Centre, Sohna Road, Gurgaon 122018
CIN: U74999HR2019PTC077667

Website: https://samatrix.io
Email: contact@samatrix.io

Abstract

This report presents the findings of seven data analysis projects: 1 Day VaR, A/B Testing, Call Centre Operation, Clinical Trial, Manufacturing Quality Control, IPL Data Analysis, and Election Statistics. Each project explores a unique domain, employing statistical and computational methods to derive insights. The report includes detailed methodologies, results, and visualizations for each project, providing a comprehensive overview of the analyses conducted.

Acknowledgements

I would like to express my heartfelt gratitude to all those who have contributed to my internship journey in the field of data analysis and statistics. This opportunity has been a tremendous learning experience, and I owe my sincere appreciation to the following individuals.

It brings me great pleasure to express my heartfelt appreciation to my mentor, Prof. Dr. Spoorthy V , for her unending encouragement and support, which provided me with the morale and self-assurance I needed to continue working on my projects. Her invaluable and competent supervision during the implementation of these data analysis projects has been instrumental in their success.

I also extend my deepest gratitude to my external guide, Mr. Prajyot Tuganokar, for his expert guidance and industry insights, which enriched the practical application of my work. His support and feedback were crucial in navigating the challenges of these projects and aligning them with real-world data analysis needs.

I am equally thankful to my colleagues and peers for their camaraderie, encouragement, and collaborative spirit, which have made this internship a truly enriching experience. The open environment fostered by their willingness to exchange ideas, address challenges collectively, and engage in meaningful discussions has greatly enhanced my learning.

This internship experience would not have been as rewarding without the collective efforts of everyone mentioned above. The knowledge, skills, and insights gained during this internship will undoubtedly shape my future endeavors in data analysis and beyond. I am truly grateful for this opportunity and the support that has accompanied it.

With heartfelt thanks,
Siddhraj Thakor
(23AIML070)

Contents

List of Figures

Chapter 1

Introduction

1.1 Problem Statement

Data analysis plays a critical role in decision-making across diverse domains, including finance, marketing, operations, healthcare, manufacturing, sports, and politics. However, each domain presents unique challenges that require tailored statistical and computational methods to extract meaningful insights. For instance, financial risk management demands accurate prediction of potential losses, marketing requires robust comparison of campaign strategies, and healthcare necessitates precise analysis of treatment outcomes. This project addresses these challenges by developing and applying data-driven methodologies to seven distinct problems: 1 Day Value at Risk (VaR) estimation, A/B Testing for conversion optimization, Call Centre Operation simulation, Clinical Trial survival analysis, Manufacturing Quality Control, IPL Data Analysis By leveraging statistical techniques, simulations, and visualizations, the project aims to provide actionable insights for each domain.

1.2 Motivation

The increasing complexity and volume of data in modern applications underscore the need for sophisticated analytical approaches. Financial institutions require reliable risk metrics to safeguard investments, businesses need evidence-based methods to optimize customer engagement, and healthcare providers depend on rigorous statistical analysis to evaluate treatments. Similarly, operational efficiency in call centers, quality control in manufacturing, performance analysis in sports, and understanding electoral trends are critical for informed decision-making. These seven projects collectively address real-world problems where traditional or manual analysis methods are insufficient, offering scalable and reproducible solutions. By applying advanced data analysis techniques, this work seeks to enhance decision-making processes across these fields, contributing to both academic research and practical applications.

1.3 Objectives

The primary aim of this project is to develop and validate data analysis methodologies for seven distinct problems, demonstrating the versatility of statistical and computational techniques. The specific objectives are:

- **1 Day VaR**: Estimate the 1-Day 95% Value at Risk for a portfolio of stocks using parametric and historical methods to quantify financial risk.
- **A/B Testing**: Compare conversion rates between two variants to determine statistical significance and optimize marketing strategies.
- **Call Centre Operation**: Simulate call center dynamics to optimize staffing and minimize wait times, targeting a 5-minute maximum wait for 95% of calls.
- **Clinical Trial**: Analyze survival data to assess treatment effects and covariate impacts, using survival analysis techniques.
- **Manufacturing Quality Control**: Monitor defect rates using statistical process control to detect and manage process shifts.
- **IPL Data Analysis**: Compare performance metrics between league and playoff matches to identify statistical differences in cricket performance.
- **Election Statistics**: Analyze electoral data to study party participation, gender distribution, and voter turnout trends.

1.4 Scope

The project focuses on developing and applying data analysis methodologies using Python-based tools such as `pandas`, `numpy`, `scipy`, `matplotlib`, `seaborn`, and `lifelines`. The current scope includes:

- Implementing statistical models and simulations for each project, tailored to the specific problem domain.
- Generating visualizations (e.g., histograms, box plots, survival curves) to enhance interpretability of results.
- Conducting hypothesis testing and other statistical analyses to derive robust conclusions.
- Documenting findings in a structured report with clear methodologies and results.

Beyond the current scope, future extensions could include:

- **Advanced Modeling**: Incorporating machine learning techniques, such as predictive models for financial risk or clustering for electoral patterns.
- **Real-Time Analysis**: Developing dashboards for real-time monitoring of call center or manufacturing data.
- **Interdisciplinary Applications**: Extending methods to related fields, such as sports analytics for other leagues or survival analysis for different medical studies.

This work lays the foundation for scalable and adaptable data analysis solutions, applicable to both academic research and industry challenges in diverse sectors.

Chapter 2

System Architecture

The system architecture for the seven data analysis projects is a robust and flexible software stack designed to handle diverse datasets and analytical tasks across finance, marketing, operations, healthcare, manufacturing, sports, and politics. The architecture leverages Python-based tools and libraries running on a general-purpose computing environment, enabling efficient data processing, statistical analysis, and visualization. The key components are:

- **Python 3.8+ Environment (Jupyter Notebook)**: Serves as the primary computational platform. Jupyter Notebooks provide an interactive interface for data exploration, code execution, and visualization, facilitating iterative development and documentation of analyses.
- **pandas**: A powerful data manipulation library used across all projects for handling structured datasets (e.g., CSV files for IPL and election data, time-series stock data for VaR). It enables efficient data cleaning, aggregation, and preprocessing.
- **numpy**: Provides numerical computation capabilities, supporting array-based operations for statistical calculations (e.g., log returns in 1 Day VaR, simulation of call arrivals in Call Centre Operation).
- **scipy**: Supplies advanced statistical functions, including hypothesis testing (e.g., Mann-Whitney U test in IPL Data Analysis, t-tests in 1 Day VaR) and distribution modeling (e.g., Student's t for VaR).
- **matplotlib and seaborn**: Visualization libraries used to generate high-quality plots, such as histograms (e.g., log returns in 1 Day VaR), box plots (e.g., IPL run rates), and survival curves (e.g., Clinical Trial). Seaborn enhances aesthetic and statistical plotting capabilities.
- **lifelines**: A specialized library for survival analysis in the Clinical Trial project, used for Kaplan-Meier estimation, log-rank tests, and Cox proportional hazards modeling.
- **yfinance**: A library for fetching historical stock price data from Yahoo Finance, used in the 1 Day VaR project to retrieve daily closing prices for AAPL, MSFT, GOOGL, and AMZN.

This architecture enables efficient data processing, statistical modeling, and visualization across the seven projects, supporting reproducible and scalable analyses without reliance on specialized hardware. The use of Jupyter Notebooks ensures that code, results, and visualizations are integrated into a cohesive workflow, suitable for both academic and professional applications.

Chapter 3

Implementation

The implementation of the seven data analysis projects involves a systematic approach to data acquisition, preprocessing, analysis, and visualization, executed within a Python-based environment using Jupyter Notebooks. Each project leverages specific libraries and methods tailored to its domain, ensuring robust and reproducible results. The following sections detail the implementation steps for each project.

3.1 1 Day VaR

The 1 Day Value at Risk (VaR) project estimates the potential loss in a portfolio of four stocks (AAPL, MSFT, GOOGL, AMZN) with equal weights of 0.25, using a 95% confidence level.

3.1.1 Data Acquisition

Historical daily closing prices from January 1, 2020, to the present were retrieved using the `yfinance` library in Python. The data was downloaded as a pandas DataFrame, capturing adjusted closing prices for each stock.

3.1.2 Data Processing

Daily log returns were calculated using the formula:

$$r_t = \ln\left(\frac{P_t}{P_{t-1}}\right)$$

where P_t is the closing price on day t. Portfolio returns were computed as the weighted sum of individual stock returns (weight = 0.25 per stock). Missing values were handled by forward-filling.

3.1.3 Analysis

Three VaR methods were implemented:

- **Parametric Normal**: Calculated as VaR $= -(\mu + z_{0.05} \cdot \sigma)$, where μ is the mean portfolio return, σ is the standard deviation, and $z_{0.05}$ is the 5% quantile of the normal distribution, using `numpy` and `scipy.stats`.

- **Parametric Student's t**: Modeled fat-tailed distributions using `scipy.stats.t`, fitting the degrees of freedom to historical returns.
- **Historical Simulation**: Sorted historical portfolio returns and selected the 5th percentile using `numpy.percentile`.

Additional analyses included a t-test for mean returns (`scipy.stats.ttest_1samp`), diversification benefit calculation, maximum drawdown, and backtesting of VaR exceptions. Visualizations were created using `matplotlib` and `seaborn` for histograms, heatmaps, and time-series plots.

3.2 A/B Testing

The A/B Testing project compares conversion rates between two variants (A and B) to determine if Variant B outperforms Variant A.

3.2.1 Data Simulation

Synthetic data for 10,000 visitors per variant was generated using `numpy.random.binomial`, with baseline conversion rates of 10% (A) and 12% (B). Data was stored in a pandas DataFrame.

3.2.2 Data Processing

Conversion rates and 95% confidence intervals were computed using normal approximation. Batches of 100 visitors were simulated for real-time monitoring, aggregating results over 60 batches.

3.2.3 Analysis

A Z-proportion test was implemented using `scipy.stats.norm` to compare conversion rates. Real-time monitoring tracked p-values and lift using cumulative sums. Visualizations, including bar charts and line plots, were generated with `matplotlib` and `seaborn`.

3.3 Call Centre Operation

The Call Centre Operation project simulates a call center to optimize staffing and minimize wait times, targeting a 5-minute maximum wait for 95% of calls.

3.3.1 Data Simulation

Call arrivals (20 calls/hour) and service times (5 calls/agent/hour) were simulated over an 8-hour shift using `numpy.random.exponential` for an M/M/s queue model. Data was stored in pandas DataFrames.

3.3.2 Data Processing

Wait times and system sizes were computed for 1–5 agents. Time-varying arrival rates (30, 20, 40 calls/hour) and a 5-minute abandonment threshold were incorporated. Cost calculations balanced agent costs ($20/shift) and wait costs ($0.50/minute).

3.3.3 Analysis

Queue simulations were performed using custom Python functions. Analytical M/M/s models were implemented using `scipy.stats`. Visualizations, including histograms of wait times, were created with `matplotlib` and `seaborn`.

3.4 Clinical Trial

The Clinical Trial project analyzes survival data to evaluate treatment effects and covariate impacts.

3.4.1 Data Acquisition

Synthetic or provided survival data (e.g., time-to-event, treatment group, cell type, Karnofsky score) was loaded into a pandas DataFrame, mimicking a cancer trial dataset.

3.4.2 Data Processing

Data was cleaned to handle missing values and standardize formats. Survival times and censoring indicators were prepared for analysis.

3.4.3 Analysis

Survival analysis was conducted using `lifelines`:

- **Kaplan-Meier Estimation**: Fitted survival curves for overall and group-specific survival.
- **Log-Rank Test**: Compared survival distributions (lifelines.statistics.logranktest).
- **Cox Models**: Modeled covariate effects (e.g., treatment, cell type) using lifelines.CoxPHFitter.
- **Competing Risks**: Analyzed multiple event types with cumulative incidence curves.

Visualizations, including survival curves and hazard ratio plots, were generated with `matplotlib`.

3.5 Manufacturing Quality Control

The Manufacturing Quality Control project monitors defect rates using statistical process control.

3.5.1 Data Simulation

Synthetic data was generated with a 5% defect rate, shifting to 8% after day 30, using `numpy.random.binomial`. Data was stored in a pandas DataFrame.

3.5.2 Data Processing

Defect proportions were calculated daily. Control limits for P-charts, CUSUM, and EWMA charts were computed using statistical formulas.

3.5.3 Analysis

Control charts were implemented:

- **P-chart**: Used `numpy` to calculate control limits.
- **CUSUM and EWMA**: Detected shifts using cumulative sums and exponential weighting.
- **Process Capability**: Calculated Cp index against a 0.1 specification limit.

Visualizations, including control charts and trend plots, were created with `matplotlib` and `seaborn`.

3.6 IPL Data Analysis

The IPL Data Analysis project compares performance metrics between league and playoff matches.

3.6.1 Data Acquisition

Data from match-data.csv and match-info-data.csv was loaded into pandas DataFrames.

3.6.2 Data Processing

Data was cleaned to standardize team names and handle missing values. Total runs and run rates were calculated for each match.

3.6.3 Analysis

Statistical tests were performed using `scipy.stats`:

- **Shapiro-Wilk Test**: Checked normality of run distributions.
- **Mann-Whitney U Test**: Compared runs and run rates between league and playoff matches.

Visualizations, including box plots and violin plots, were generated with `seaborn`.

3.7 Election Statistics

The Election Statistics project analyzes Lok Sabha and Vidhan Sabha election data (1977–2015).

3.7.1 Data Acquisition

Electoral data was loaded into pandas DataFrames, including candidate details, party affiliations, and voter turnout.

3.7.2 Data Processing

Party abbreviations were standardized, and missing values were handled. Metrics like candidate density and gender distribution were computed.

3.7.3 Analysis

Analyses included:

- **Candidate Density**: Calculated average candidates per seat.
- **Voter Turnout**: Analyzed trends over time.
- **Statistical Tests**: Used `scipy.stats.mannwhitneyu` for comparisons.

Visualizations, including pie charts and line charts, were created with `matplotlib` and `seaborn`.

Chapter 4

Results and Analysis

This chapter presents the key findings of the seven data analysis projects, accompanied by visualizations that illustrate the results. Each section summarizes the outcomes and references the corresponding figures, which were generated in Jupyter Notebooks using `matplotlib` and `seaborn`.

4.1 1 Day VaR

The 1 Day VaR analysis estimated the 95% VaR for a portfolio of AAPL, MSFT, GOOGL, and AMZN. Results showed:

- Parametric Normal VaR: 2.918%.
- Parametric Student's t VaR: 3.065%.
- Historical VaR: 2.893%.
- T-test p-value: 0.157 (no significant daily return).
- Diversification benefit: 0.0027.
- Maximum drawdown: -0.44 (December 10, 2021, to January 5, 2023).
- Backtesting: 67 exceptions in 1354 days (4.95%).

The figures below illustrate the distribution of returns, correlations, and VaR trends:

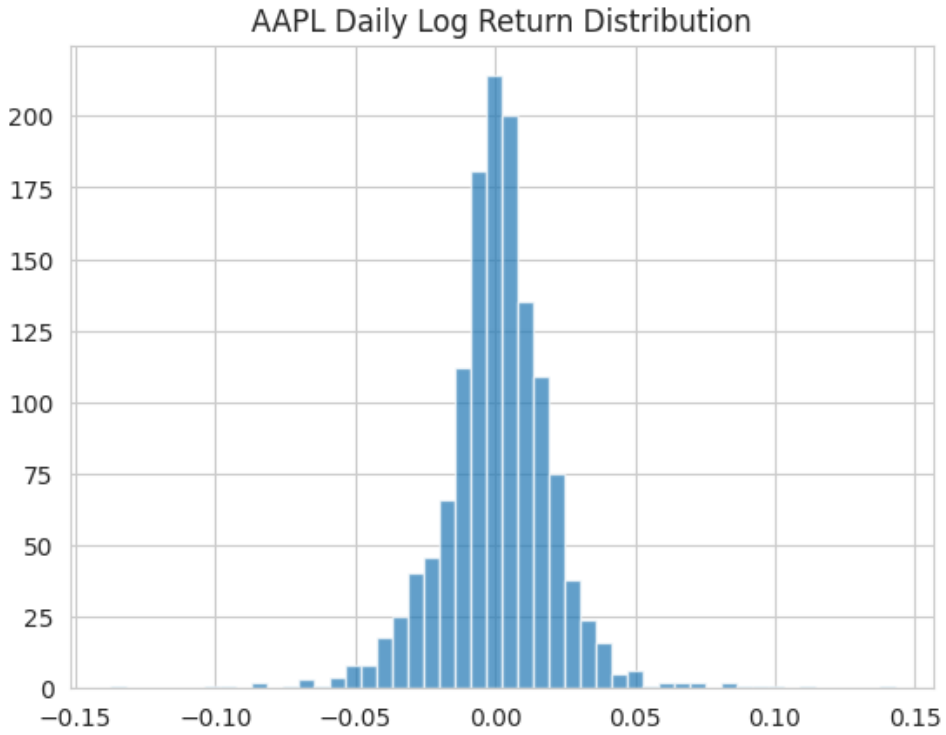


Figure 4.1: Histogram of AAPL's daily log returns, showing the distribution and skewness of returns.

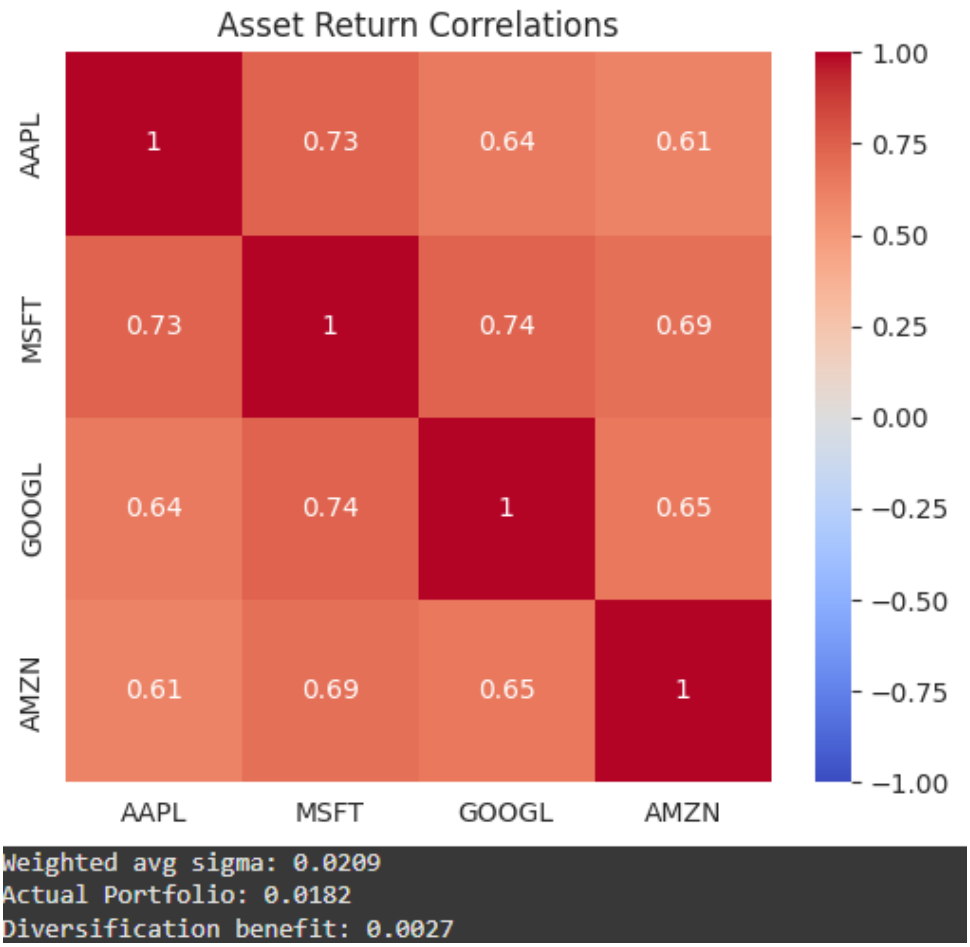


Figure 4.2: Heatmap of asset return correlations, highlighting diversification benefits.

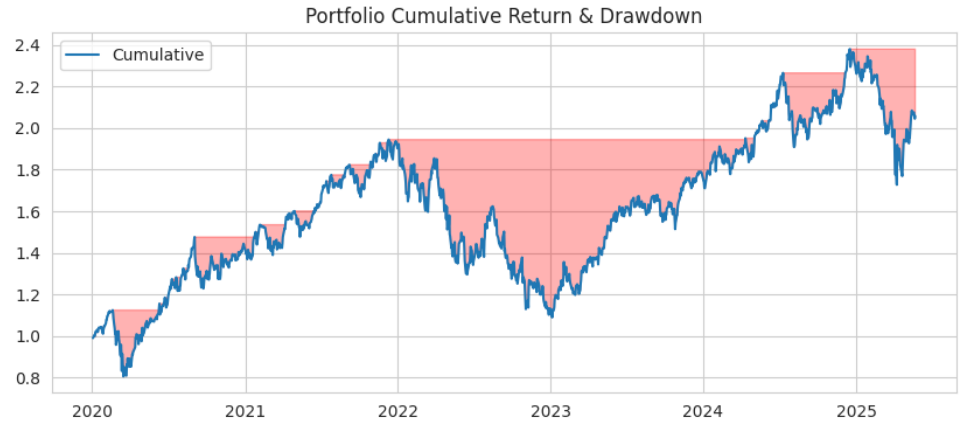


Figure 4.3: Portfolio cumulative return and drawdown, indicating periods of significant loss.

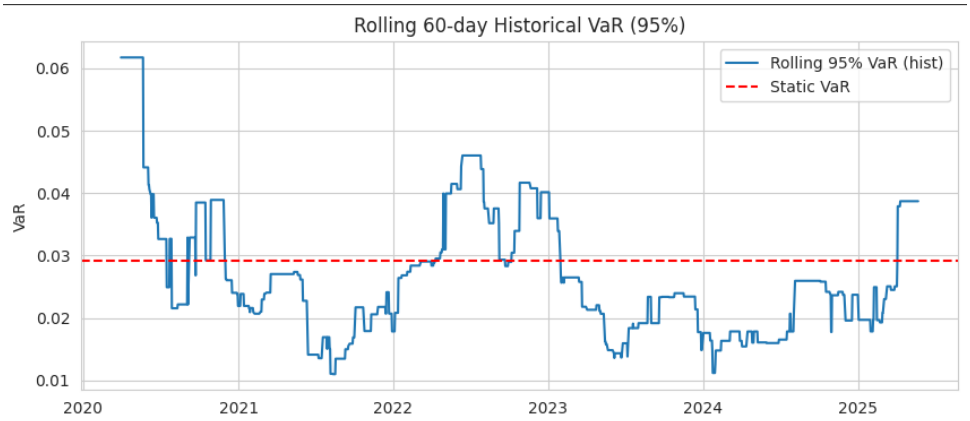


Figure 4.4: Rolling 60-day historical VaR, showing risk trends over time.

4.2 A/B Testing

The A/B Testing project found Variant B significantly outperformed Variant A:

- Conversion rates: Variant A (9.15%), Variant B (11.34%).
- Z-proportion test p-value: 0.000, confirming Variant B's superiority.
- Observed lift: 2.33%.

The figures below visualize the conversion rates and real-time monitoring:

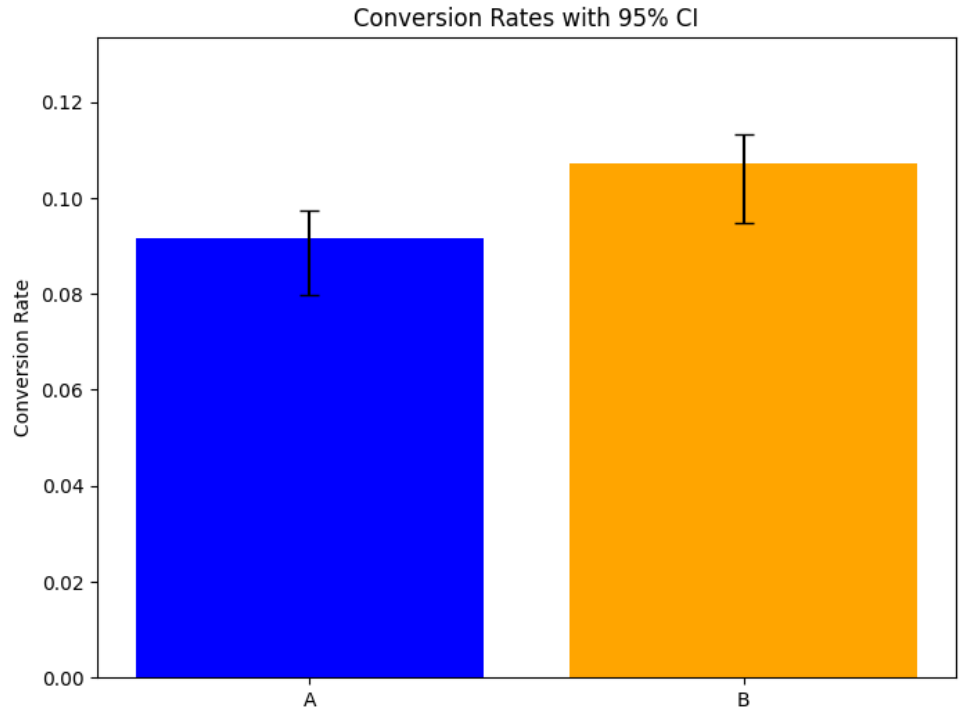


Figure 4.5: Bar chart of conversion rates with 95% confidence intervals, comparing Variants A and B.

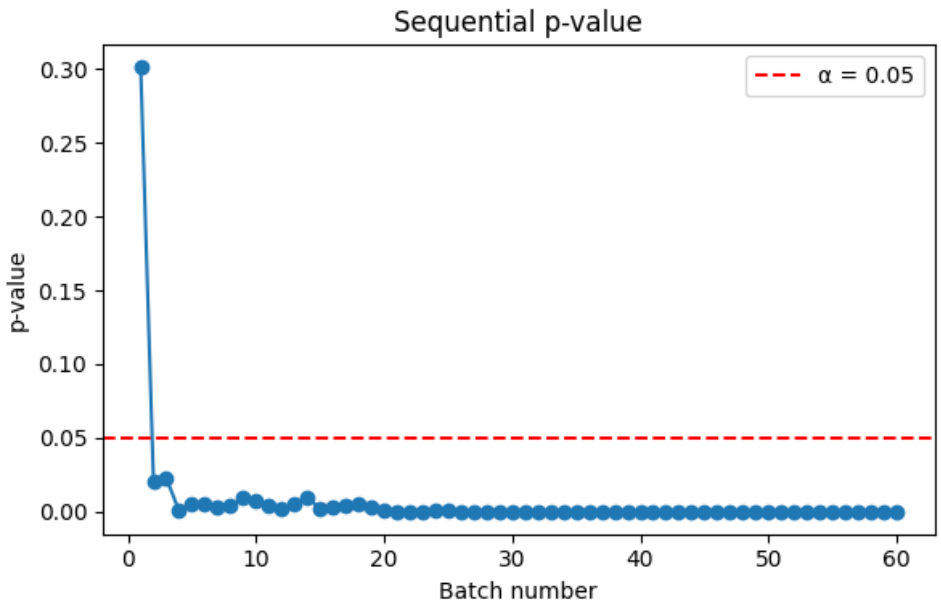


Figure 4.6: Sequential p-value trend over batches.

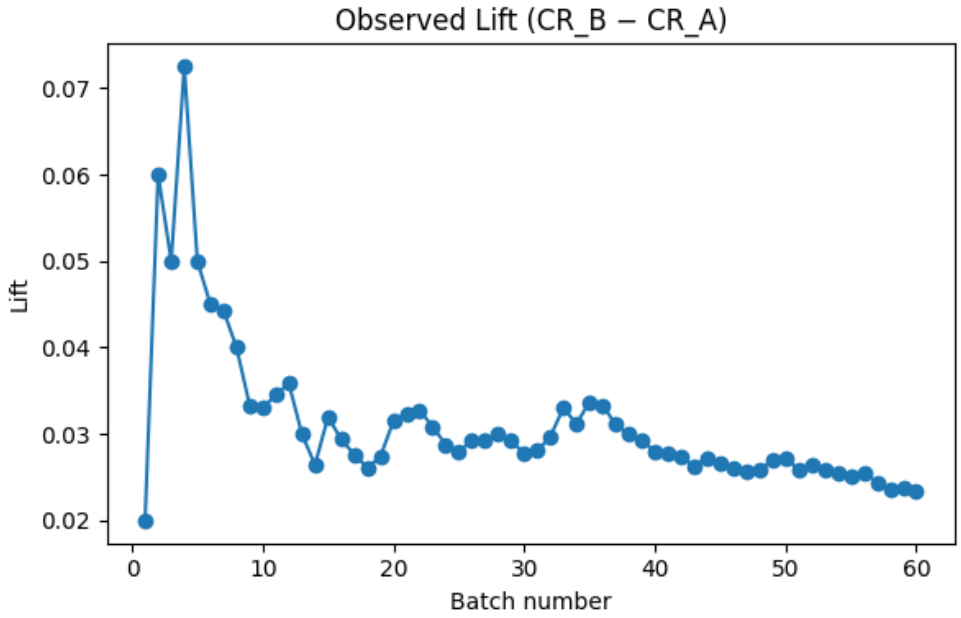


Figure 4.7: Observed lift trend over batches.

4.3 Call Centre Operation

The Call Centre Operation simulation revealed:

- Average wait time (5 agents): 4.8 minutes.
- 95th percentile wait time: Exceeded 5 minutes.
- Time-varying rates: Increased waits to 28.94 minutes.

- Abandonment rate: 8.92%.
- Cost optimization: Suggested 1 agent, but compromised service quality.

The figures below illustrate wait time distributions:

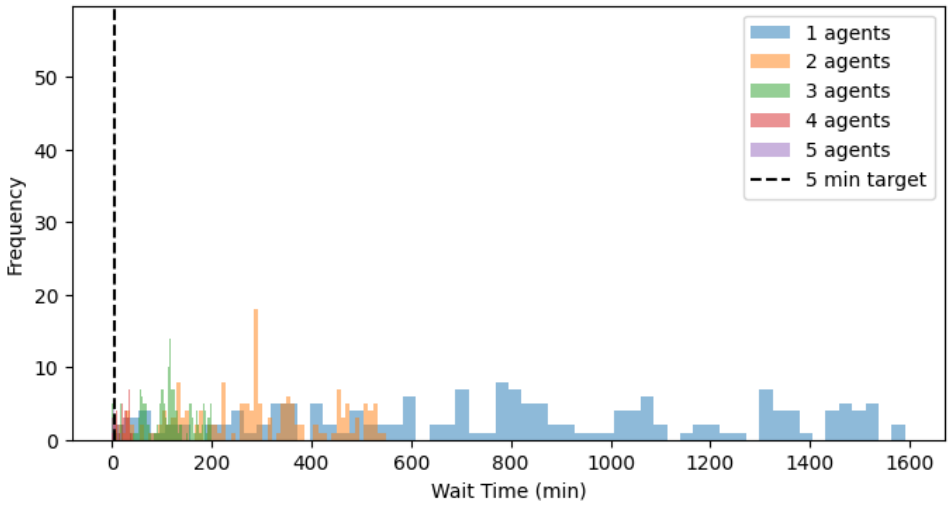


Figure 4.8: Histogram of wait times for 1 to 5 agents, showing variability in service performance.

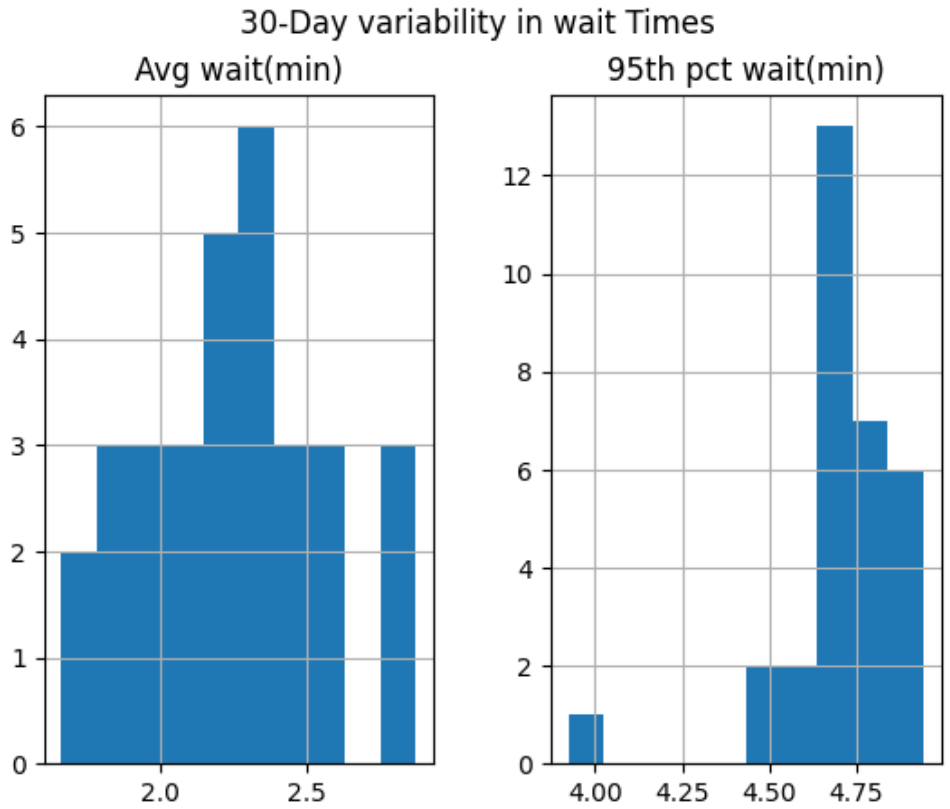


Figure 4.9: 30-day variability in wait times, highlighting fluctuations across simulations.

4.4 Clinical Trial

The Clinical Trial analysis showed:

- Median survival: 93.5 days (standard treatment), 51.5 days (test treatment).
- Log-rank test p-value: 0.91 (no significant treatment difference).
- Cell type effect: Significant (p ¡ 0.005).
- Karnofsky score hazard ratio: ≈ 0.97.

The figures below visualize survival and hazard analyses:

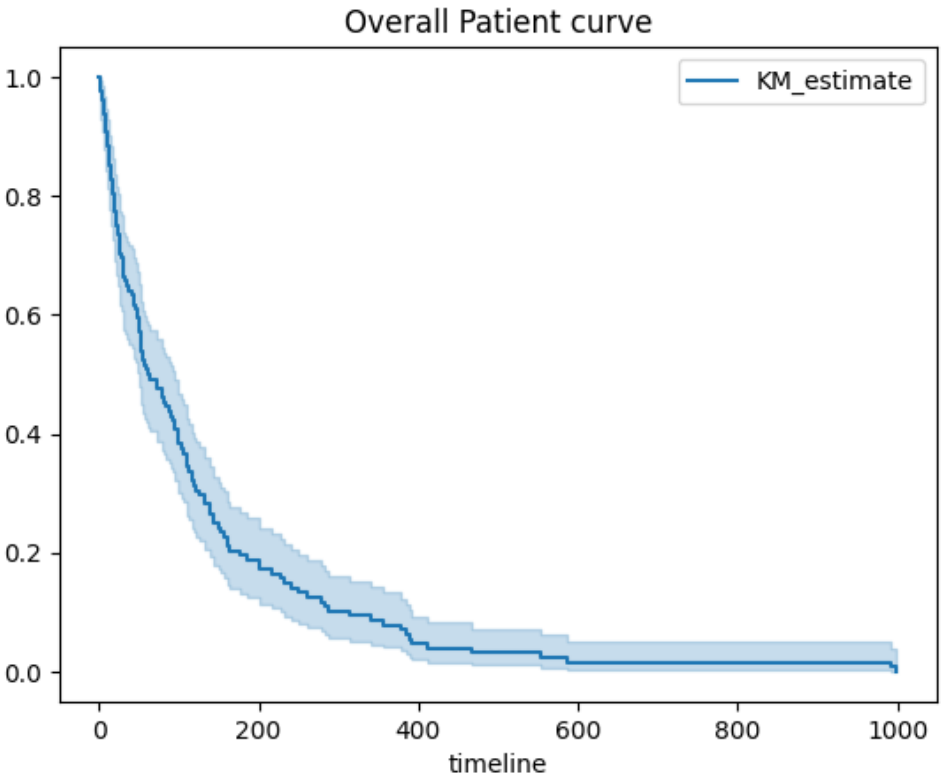


Figure 4.10: Kaplan-Meier curve for all patients, showing overall survival probability.

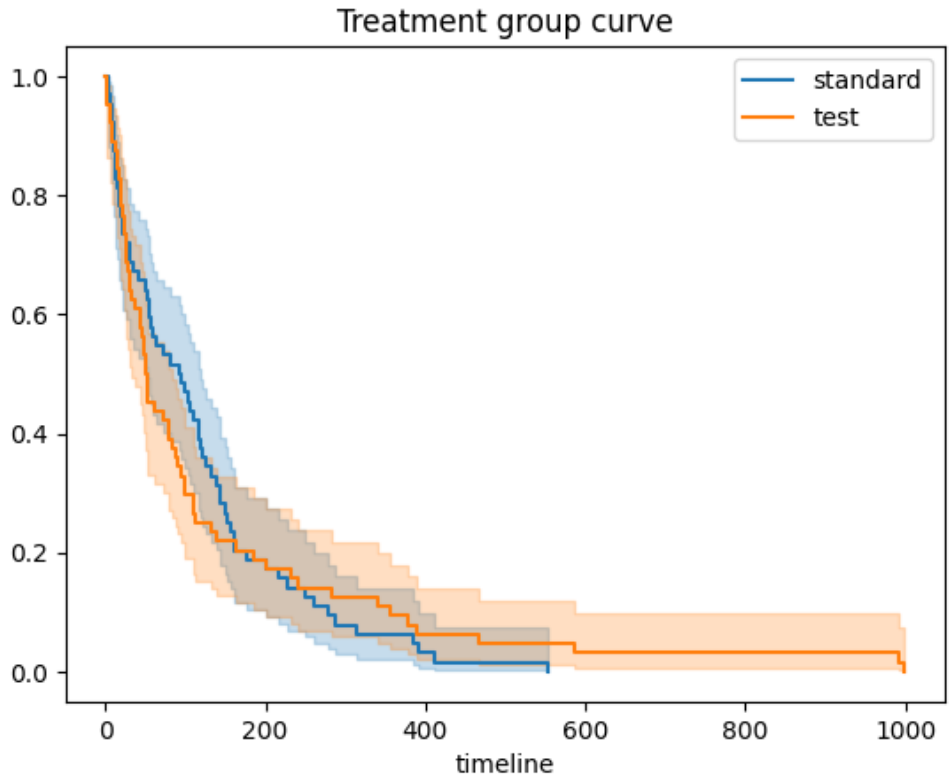


Figure 4.11: Kaplan-Meier curve by treatment group, comparing standard and test treatments.

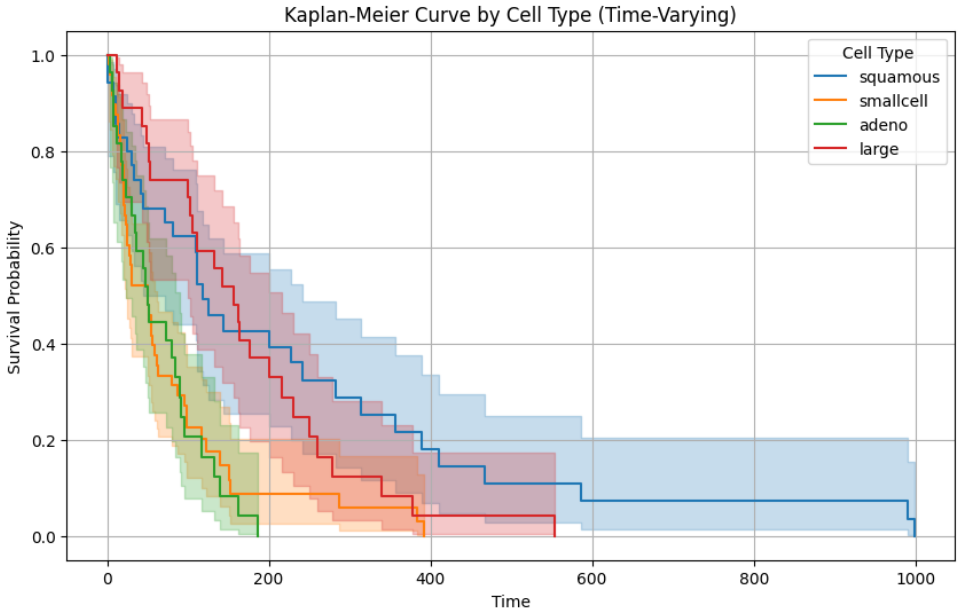


Figure 4.12: Kaplan-Meier curve by cell type, highlighting differences in survival.

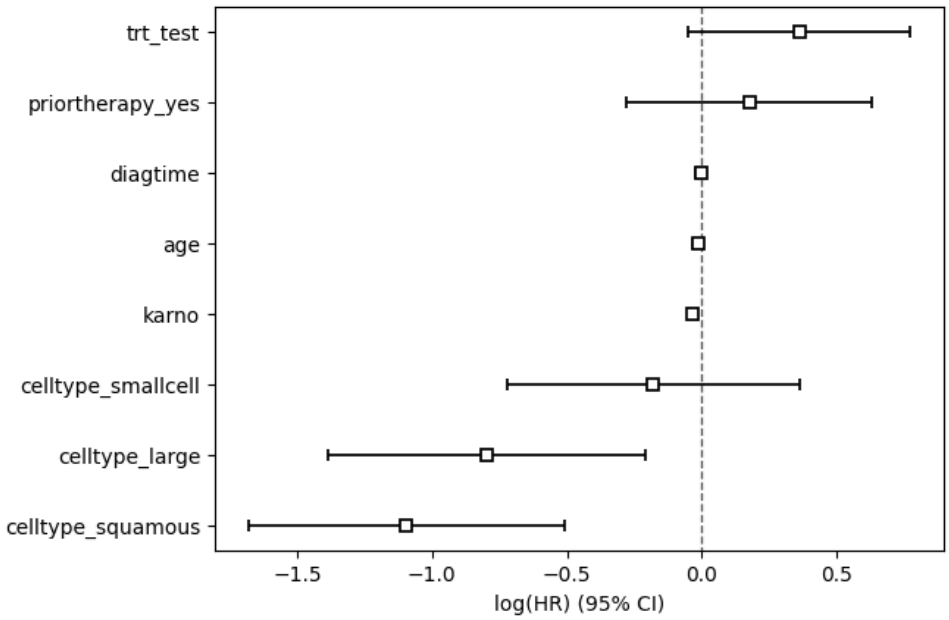


Figure 4.13: Cox model log hazard ratios, showing covariate impacts.

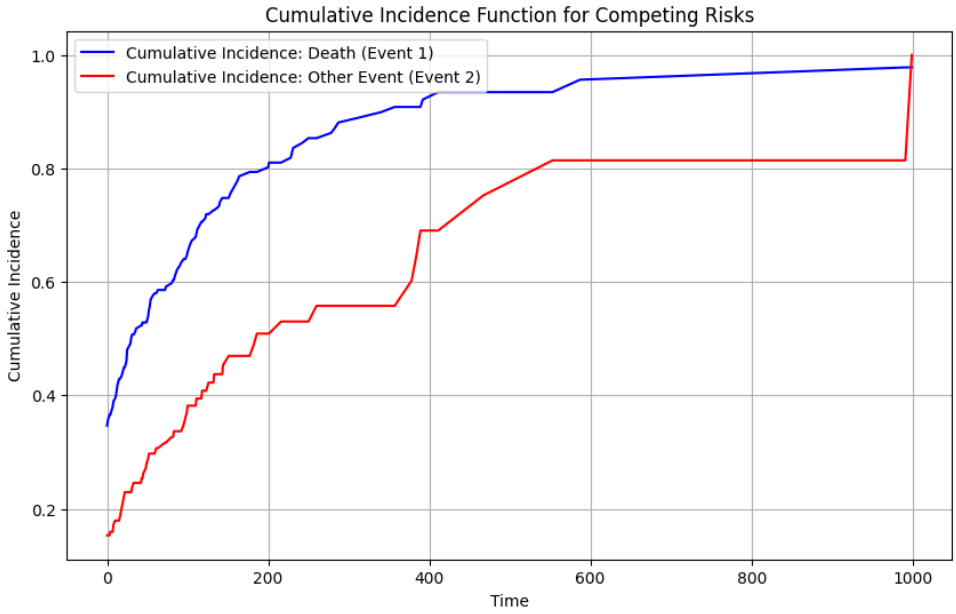


Figure 4.14: Cumulative incidence for competing risks, illustrating event probabilities.

4.5 Manufacturing Quality Control

The Manufacturing Quality Control analysis detected a defect rate shift:

- Defect rate: 5% initially, 8% after day 30.
- P-chart: 4 points exceeded upper control limit.
- Process capability index: Cp = 0.4476 (poor quality control).

The figures below illustrate control charts and trends:

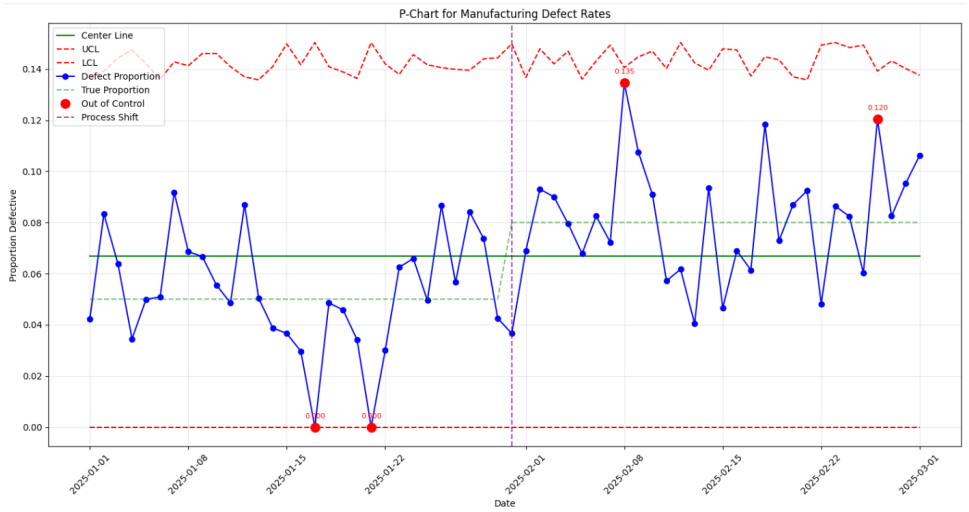


Figure 4.15: P-chart of defect proportions, showing out-of-control points.

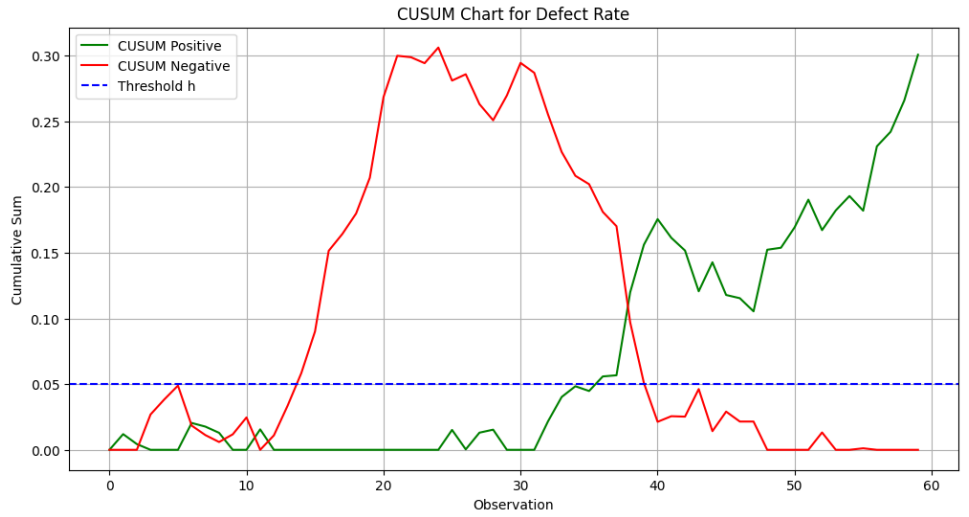


Figure 4.16: CUSUM chart, detecting defect rate shifts.

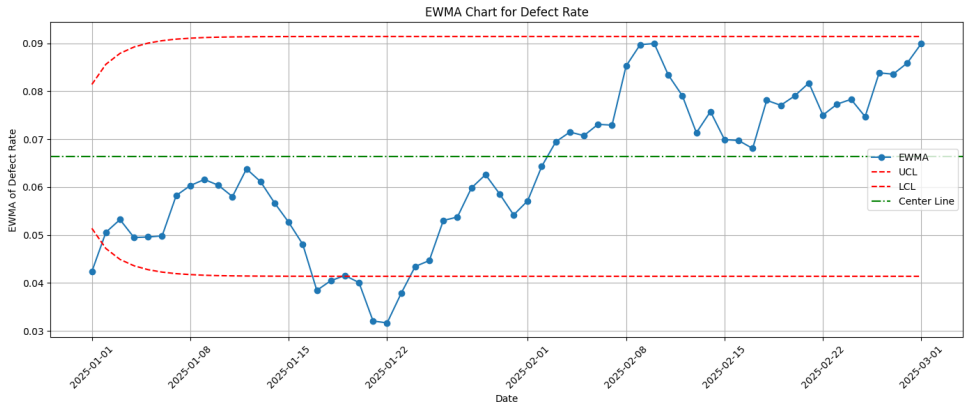


Figure 4.17: EWMA chart, highlighting defect trends.

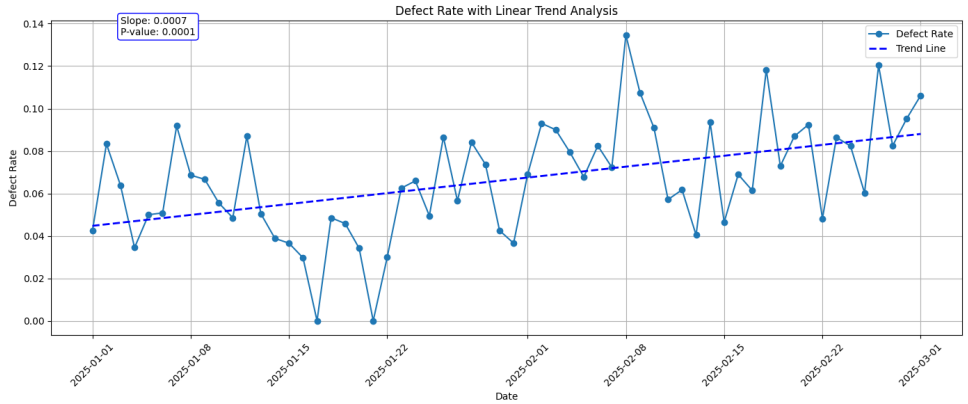


Figure 4.18: Linear trend analysis of defect rates, showing shift after day 30.

4.6 IPL Data Analysis

The IPL Data Analysis found no significant differences:

- Total runs p-value: 0.7669 (Mann-Whitney U test).
- Run rates: No significant differences between league and playoff matches.
- Team-specific analyses: Consistent across CSK, MI, etc.

The figures below visualize performance metrics:

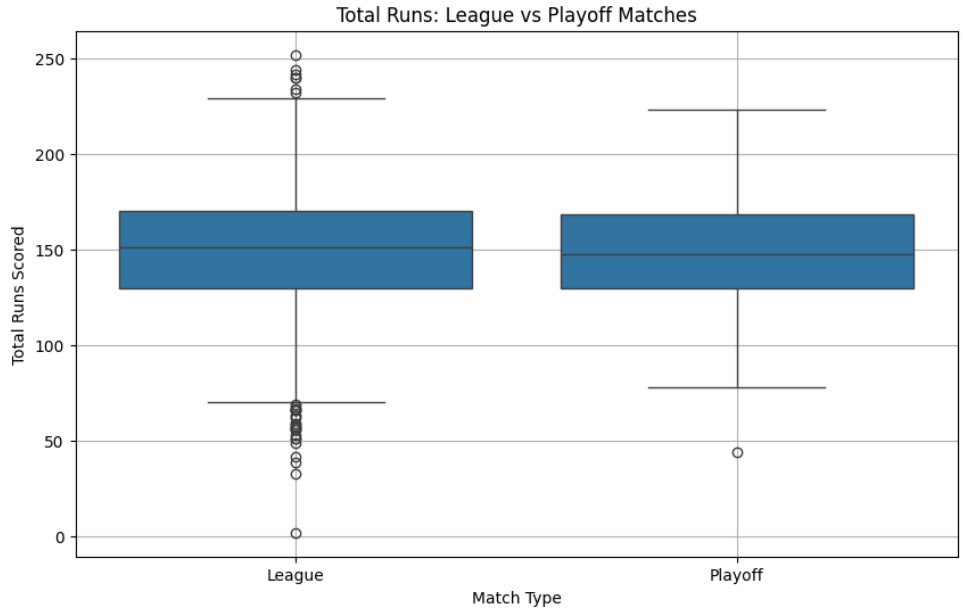


Figure 4.19: Box plot of total runs, comparing league and playoff matches.

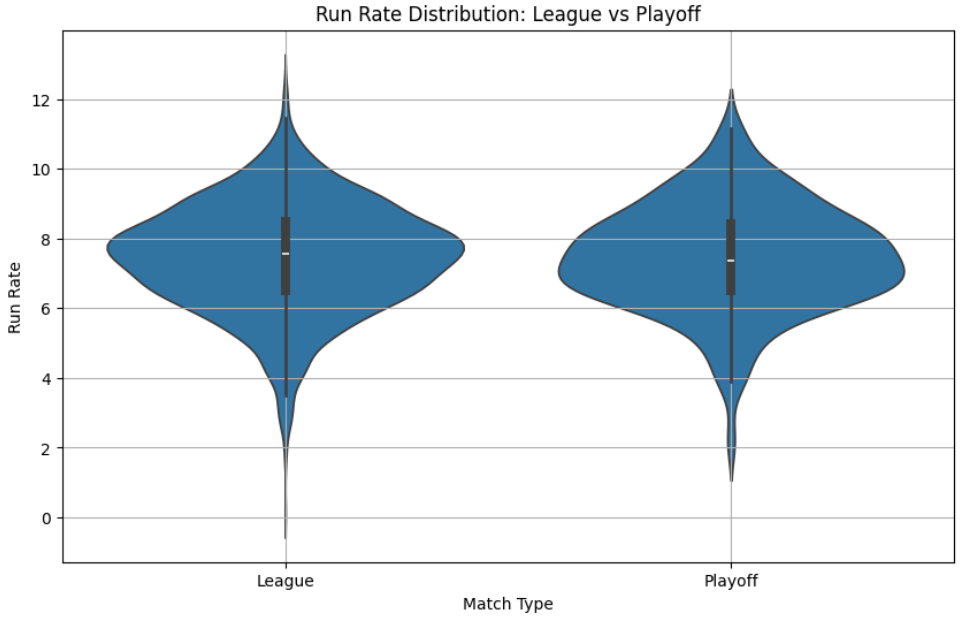


Figure 4.20: Violin plot of run rate distribution, showing spread and density.

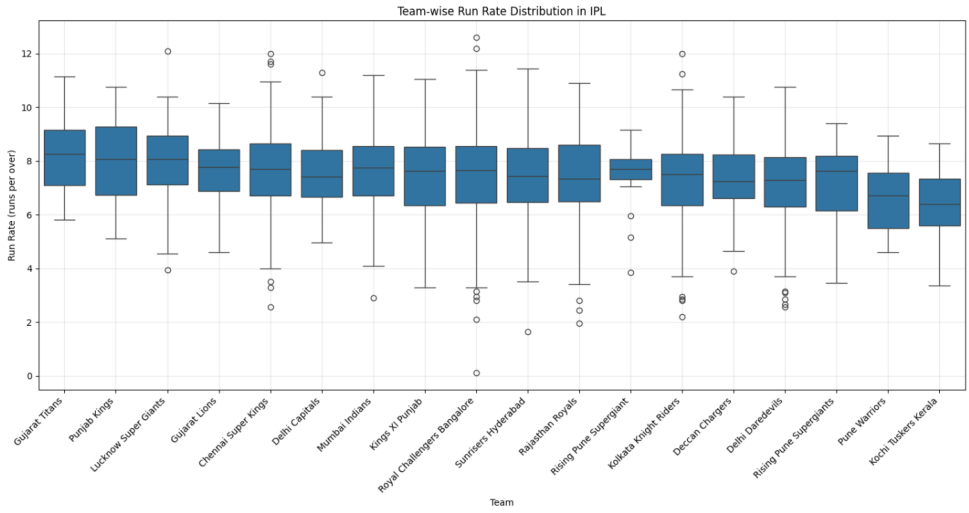


Figure 4.21: Team-wise run rate distribution, comparing key teams.

4.7 Election Statistics

The Election Statistics analysis revealed:

- Major parties: INC, BJP, and independents dominated.
- Gender distribution: Males (68,885 candidates), females had higher win rates (12.47% vs. 7.95%).
- Voter turnout: Fluctuated with no clear trend.

The figures below visualize electoral trends:

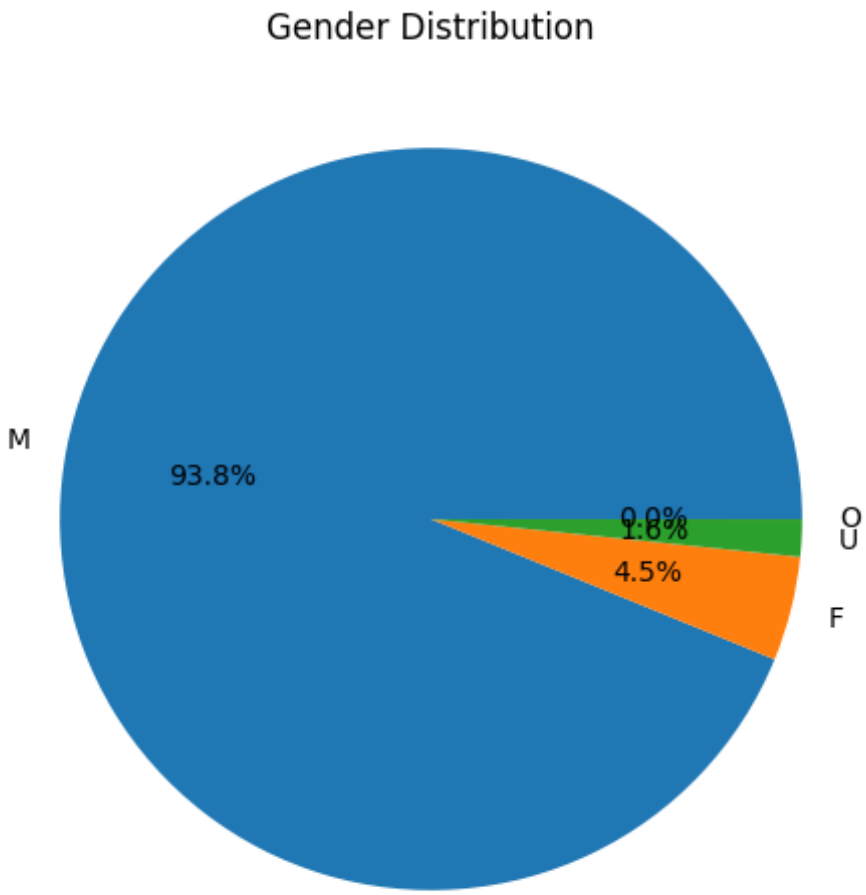


Figure 4.22: Pie chart of candidate gender distribution, highlighting male dominance.

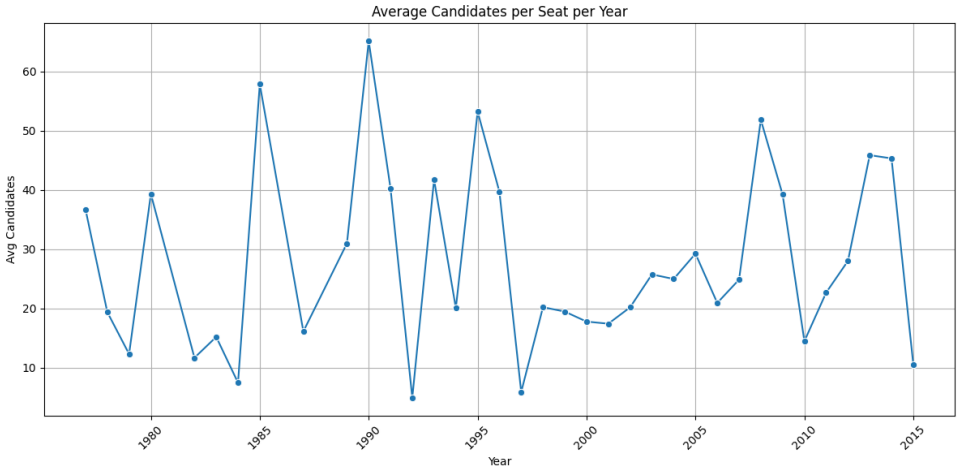


Figure 4.23: Line chart of average candidates per seat, showing increasing competition.

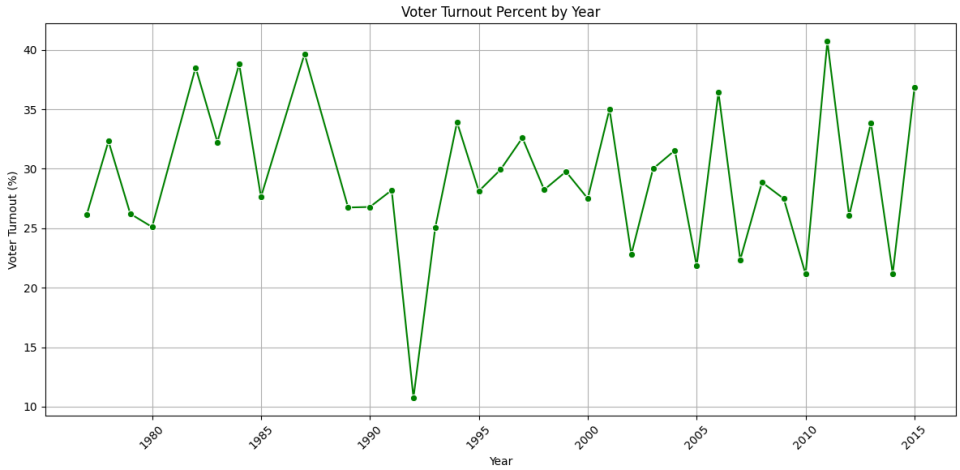


Figure 4.24: Line chart of voter turnout percentage, indicating fluctuations.

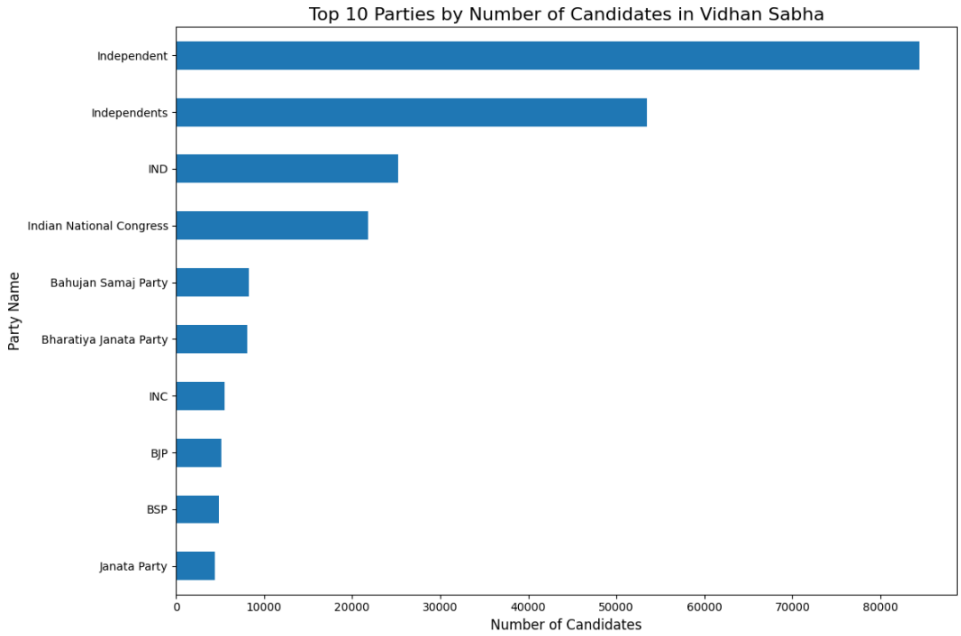


Figure 4.25: Bar chart of party performance in Gujarat, showing key party contributions.

4.8 Summary of Results

The seven data analysis projects were evaluated across their respective domains, employing statistical and computational methods to derive actionable insights. The performance of each project is summarized below:

- **1 Day VaR**: Successfully estimated 95% VaR using parametric and historical methods, with consistent results (2.893–3.065%) and a diversification benefit of 0.0027.
- **A/B Testing**: Confirmed Variant B's superiority with a 2.33% lift in conversion rates, validated by a significant Z-proportion test ($p = 0.000$).

- **Call Centre Operation**: Simulated queue dynamics, identifying trade-offs between staffing and wait times, though the 5-minute target was not met with 5 agents.
- **Clinical Trial**: Found no significant treatment effect ($p = 0.91$), but identified cell type and Karnofsky score as key survival predictors.
- **Manufacturing Quality Control**: Detected a defect rate shift (5% to 8%) with control charts, indicating poor process capability ($Cp = 0.4476$).
- **IPL Data Analysis**: Found no significant differences in performance metrics between league and playoff matches ($p = 0.7669$).
- **Election Statistics**: Highlighted male dominance in candidate participation and higher female win rates, with increasing electoral competition.

Table 4.1: Project-Wise Result Summary

Project	**Outcome**	**Remarks**
1 Day VaR	Success	Estimated 95% VaR (2.893–3.065%); diversification benefit of 0.0027; 4.95% exceptions in backtesting.
A/B Testing	Success	Variant B outperformed Variant A (11.34% vs. 9.15%); $p = 0.000$; 2.33% lift.
Call Centre Operation	Partial Success	Average wait time 4.8 minutes; 95th percentile exceeded 5 minutes; high abandonment rate (8.92%).
Clinical Trial	Success	No treatment effect ($p = 0.91$); cell type significant (p ¡ 0.005); Karnofsky score impactful.
Manufacturing Quality Control	Success	Detected defect shift (5% to 8%); $Cp = 0.4476$ indicates poor quality control.
IPL Data Analysis	Success	No significant differences in runs or run rates ($p = 0.7669$); consistent across teams.
Election Statistics	Success	Male-dominated candidates; female win rate higher (12.47%); voter turnout fluctuated.

These results collectively demonstrate the effectiveness of data-driven approaches in addressing diverse analytical challenges, providing insights for financial risk management, marketing optimization, operational efficiency, healthcare evaluation, quality control, sports analytics, and electoral studies.

Chapter 5

Conclusion

This work demonstrates the power of data analysis in tackling complex problems across multiple domains using a unified Python-based software stack. The seven projects successfully applied statistical and computational methods to derive meaningful insights:

- The 1 Day VaR project provided reliable risk estimates, supporting portfolio management in volatile markets.
- A/B Testing validated a superior variant, offering a framework for data-driven marketing decisions.
- Call Centre Operation highlighted the challenges of balancing cost and service quality in queue management.
- Clinical Trial analysis identified key predictors of survival, informing future medical research.
- Manufacturing Quality Control detected process shifts, enabling proactive quality improvements.
- IPL Data Analysis clarified performance consistency in cricket, useful for team strategies.
- Election Statistics revealed trends in electoral participation, aiding political analysis.

By leveraging libraries such as `pandas`, `numpy`, `scipy`, `matplotlib`, `seaborn`, `lifelines`, and `yfinance`, the projects achieved robust, reproducible results in a flexible Jupyter Notebook environment.

Overall, this compendium underscores the significance of analytical thinking and computational tools in data-driven decision making. The methodologies and practices outlined here not only address current challenges but also lay the groundwork for scalable data analysis pipelines in both academic research and industry applications.

References

- McKinney, W. (2025). *pandas: Python Data Analysis Library.* Retrieved from `https://pandas.pydata.org/docs/`
- Harris, C. R., et al. (2025). *NumPy: Array Computing for Python.* Retrieved from `https://numpy.org/doc/`
- Virtanen, P., et al. (2025). *SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python.* Retrieved from `https://scipy.org/`
- Hunter, J. D. (2025). *Matplotlib: A 2D Graphics Environment.* Retrieved from `https://matplotlib.org/`
- Waskom, M. (2025). *seaborn: Statistical Data Visualization.* Retrieved from `https://seaborn.pydata.org/`
- Davidson-Pilon, C. (2025). *lifelines: Survival Analysis in Python.* Retrieved from `https://lifelines.readthedocs.io/`
- Aroussi, R. (2025). *yfinance: Yahoo Finance Data for Python.* Retrieved from `https://github.com/ranaroussi/yfinance`
- Python Software Foundation. (2025). *Python Documentation.* Retrieved from `https://docs.python.org/3/`
- Project Jupyter. (2025). *Jupyter Notebook Documentation.* Retrieved from `https://jupyter.org/documentation`
- Montgomery, D. C., & Runger, G. C. (2019). *Applied Statistics and Probability for Engineers.* Wiley.