Summer Internship Report
On
" Data Analysis"

(AIML306 – Summer Internship - I)

Prepared by

Het Patel (23AIML051)

Under the Supervision of

Prof. Nishant Koshti

Submitted to

Charotar University of Science & Technology (CHARUSAT) for the Partial Fulfillment of the Requirements for the Degree of Bachelor of Technology (B.Tech.)

for Semester 5

Submitted at

CHARUSAT
CHAROTAR UNIVERSITY OF SCIENCE AND TECHNOLOGY

Accredited with Grade A+ by NAAC Accredited with Grade A by KCG

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

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

August, 20

CHAROTAR UNIVERSITY OF SCIENCE AND TECHNOLOGY

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 **Het Patel(23AML051)** under the guidance and supervision of **Prof. Nishant Koshti** & **Mr. Vishwas Soni** for the subject **Summer Internship – I (AIML306)** of 5th Semester of Bachelor of Technology in **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 work of candidate himself, has duly been completed, and fulfills the requirement of the ordinance relating to the B.Tech. Degree of the University and is up to the standard in respect of content, presentation and language for being referred by the examiner(s).

Under the supervision of,

Prof. Nishant Koshti
Assistant Professor
Department of Artificial Intelligence and Machine Learning
CSPIT, FTE, CHARUSAT, Changa, Gujarat

Mr.Vishwas Soni
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 **Het Piyushkumar Patel,** 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

ACKNOWLEDGMENT

I would like to express my heartfelt gratitude to all those who have contributed to my internship journey in the Data Analytics domain. 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 excellent mentor, Prof. Nishant Koshti, for her unending encouragement and support, which provided me with the morale and self-assurance I needed to continue working on my research. I would like to express my heartfelt gratitude to them for their invaluable and competent supervision and help during the project's implementation.

I extend my gratitude to Mr. Vishwas Soni for your camaraderie, encouragement, and collaborative spirit have made this internship a truly enriching experience. I am thankful for the open environment where ideas were freely exchanged, challenges were collectively addressed, and meaningful discussions took place.

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 the Data Analytics domain and beyond. I am truly grateful for this opportunity and the support that has accompanied it.

With heartfelt thanks,

Het Patel(23AIML051)

DESCRIPTION OF COMPANY

COMPANY OVERVIEW

Samatrix.io is a leading digital transformation company that provides innovative solutions and training in the fields of Artificial Intelligence, Machine Learning, Data Analytics, and Cloud Computing. The company has established itself as a pioneer in combining technology education with real-world applications to empower students, professionals, and enterprises.

CORE SERVICES

1. Consulting Services

Samatrix.io provides expert consulting services to businesses aiming to integrate data-driven strategies and adopt AI-powered automation. Their consulting solutions span multiple domains delivering customized strategies for process optimization and digital transformation.

2. Industry Projects

The organization collaborates with enterprises and startups to deliver tailored software and data science projects. With hands-on experience in end-to-end product development, their project services encompass requirement analysis and deployment using modern practices.

3. Education

Samatrix.io has built a strong reputation for its education and training services, offering structured internship programs, live project experience, and mentorship

4. Data Analytics and AI

Specialized in predictive analytics, anomaly detection, natural language processing, and deep learning solutions, Samatrix.io empowers organizations to make informed decisions.

5. Digital Operations and Platforms

From mobile app development to cloud-native software platforms, Samatrix.io builds scalable and secure digital systems.

6. Cloud & Infrastructure Services

The company provides robust infrastructure solutions using AWS, Microsoft Azure, and Google Cloud.

DIFFERENTIATORS

Samatrix.io emphasizes practical learning and hands-on implementation. Through structured mentorship, mini-projects, and domain-based problem solving, the organization stands out for producing job-ready professionals. Regular progress tracking, code reviews, and presentation sessions enhance real-world readiness.

LOCATION

Plot No. 4, Sector 44, Gurugram, Haryana – 122003
Website: www.samatrix.io

ABSTRACT

I would like to express my deepest gratitude to everyone who contributed to making my internship journey in the **Data Analytics** domain a truly enriching and transformative experience. This opportunity has provided me with invaluable learning, and I am sincerely thankful to the following individuals for their guidance and support throughout the process.

First and foremost, it is with immense respect and appreciation that I thank **Prof. Nishant Koshti,** my mentor, for her unwavering support, insightful guidance, and constant encouragement. Her mentorship instilled in me the confidence and motivation to delve deeper into my research work. I am truly grateful for her expertise, patience, and the valuable direction she provided throughout the duration of the project.

I also extend my heartfelt thanks to **Mr. Vishwas Soni** for his collaborative spirit, professional camaraderie, and continuous encouragement. His open-minded approach fostered a dynamic and engaging environment where ideas could flourish, challenges were addressed as a team, and constructive discussions led to meaningful outcomes.

This internship would not have been as impactful without the collective efforts and inspiration from the individuals mentioned above. The experience has significantly enhanced my technical and analytical skills, and the insights gained will undoubtedly guide my future endeavours in the field of Data Analytics and beyond.

With sincere appreciation,

HET PATEL
23AIML051

Table of Contents

List of Figures

CHAPTER 1: INTRODUCTION

1.1 OVERVIEW OF THE INTERNSHIP

The internship at Samatrix.io was a comprehensive learning experience focused entirely on real-world applications of **Data Analytics**. The main objective of this internship was to strengthen our practical understanding of data analysis techniques through real-world case studies and projects. Instead of a single task or research-based effort, the internship was structured around **six diverse industry-inspired projects**, allowing us to explore a range of analytical techniques across domains like healthcare, sports, manufacturing, e-governance, and digital business.

The internship was designed to simulate how real-world data analysis projects are executed, from problem definition to data preprocessing, model implementation, visualization, and interpretation. Emphasis was placed on understanding data-driven decision-making, statistical testing, and business impact interpretation. We collaborated in guided sessions while also contributing individually to project deliverables.

Some of the projects undertaken during this internship were:

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

Through these projects, we experienced end-to-end analytical workflows—from dataset preparation and EDA to statistical modeling and stakeholder-oriented insights—thus preparing us for industry-level analytics roles.

1.1.1 Learning Objectives

The internship was carefully structured to help students bridge the gap between academic learning and its practical application. The following key objectives were achieved during the internship:

- Gain hands-on experience in handling diverse real-world datasets across domains.
- Apply descriptive, inferential, and predictive analytical techniques using Python.
- Develop familiarity with statistical methods like hypothesis testing, survival analysis, and quality control.
- Learn how to visualize data trends and patterns using Matplotlib and Seaborn.
- Translate complex analysis into actionable insights relevant to industries.
- Build teamwork, documentation, and communication skills essential for analytics roles.
- Strengthen proficiency in tools like Jupyter Notebook, GitHub, and essential libraries such as Pandas, NumPy, Scikit-learn, and Lifelines.

This internship created a solid foundation for working as a data analyst by promoting problem-solving, critical thinking, and the ability to draw conclusions from data in a business-relevant manner.

1.1.2 Initial Setup and Tools

The internship was structured with clear objectives to enhance technical and analytical capabilities. The major goals were:

- To understand the business context and real-world implications of credit card fraud.
- To explore Python libraries used in data science such as Pandas, NumPy, and Scikit-learn.
- To practice data cleaning, feature selection, and outlier detection.

- To build and compare machine learning models such as Logistic Regression, Random Forest, and Isolation Forest.
- To evaluate model performance using precision, recall, F1-score, and ROC-AUC metrics.
- To develop problem-solving skills through experimentation and optimization.

1.2 OVERVIEW OF TECHNOLOGIES USED

The internship setup involved preparing a development environment tailored for data analysis, statistical computing, and visualization. Below are the tools and configurations used consistently across all projects during the internship:

- **Programming Language:** Python 3.8+
- **IDE & Environment:** Jupyter Notebook (for iterative analysis and prototyping), Visual Studio Code (for code editing and organization)

Libraries and Frameworks:

- pandas and NumPy: Used for data manipulation, transformation, and numeric computation
- matplotlib and seaborn: For data visualization, distribution plots, trend lines, and aesthetic charting
- scikit-learn: Applied for building machine learning models, classification, and evaluation
- SciPy: Used for statistical testing and hypothesis validation
- lifelines: For survival analysis in clinical trial data using Kaplan–Meier estimators and Cox models

Version Control and Collaboration:

- Git and GitHub: Used to track changes, collaborate with peers, and submit weekly project deliverables

Workflow and Documentation:

- Markdown cells in Jupyter Notebook were used to document insights, hypotheses, results, and interpretations alongside code
- Structured folders were maintained for each project to include raw data, processed datasets, code files, charts, and result summaries

This setup ensured a smooth workflow throughout the internship, combining the power of statistical rigor with real-world applicability in a collaborative and iterative environment.

1.3 IMPORTANCE OF REAL-TIME ANALYTICS

In today's data-driven world, the ability to extract actionable insights from real-time and historical datasets is at the core of smart decision-making across industries such as healthcare, e-commerce, governance, manufacturing, and sports.

The internship experience at Samatrix.io reflected the practical significance of such analytics by allowing us to work on dynamic, real-world datasets and apply industry-standard techniques to generate insights. Projects such as A/B testing, survival analysis, and quality control emphasized not just theoretical understanding but also the practical challenges of dealing with noisy data, missing values, and the need for interpretable models.

For instance:

- In the **Call Center Optimization** project, data analytics helped in identifying bottlenecks in service delivery, directly impacting business efficiency.
- In the **India Elections Statistics** project, demographic patterns were studied to explore voting behaviors geographically.
- The **Statistical Clinical Trial** project offered experience with survival models that have real-world applications in drug effectiveness analysis.

These applications mirror common industry challenges:

- Understanding consumer behavior and response
- Optimizing operations based on data trends
- Monitoring performance and diagnosing inefficiencies

- Making data-backed decisions in uncertain or high-stakes scenarios

This internship provided a strong simulation of such business-driven analytical scenarios and highlighted how open-source tools, reproducible workflows, and domain understanding together drive meaningful results in real-world data science

Chapter 2 System Architecture

The system architecture developed for the seven data analysis projects is designed to be robust, flexible, and scalable, capable of addressing a wide range of analytical challenges across multiple domains—including finance, marketing, operations, healthcare, manufacturing, sports, and politics. The architecture leverages a Python-based data science stack deployed within a general-purpose computing environment, facilitating efficient data ingestion, transformation, statistical analysis, and visualization.

The key components of the system are as follows:

1. **Python 3.8+ Environment (Jupyter Notebook):**
   Acts as the core computational platform. Jupyter Notebooks provide an interactive development interface, allowing for seamless integration of code execution, visualization, and documentation. This supports rapid prototyping, iterative analysis, and enhanced reproducibility.
2. **pandas:**
   A powerful library for data manipulation and analysis. It is utilized across all projects for importing and managing structured datasets—such as CSV files in IPL and election analysis, and time-series data in the Value at Risk (VaR) project. pandas enables efficient data cleaning, transformation, aggregation, and indexing.
3. **numpy:**
   Used for numerical operations involving arrays and matrices. It supports statistical computations including log return calculations (e.g., in the 1-Day VaR project) and stochastic simulation processes (e.g., call arrival modeling in the Call Centre Operations project).
4. **scipy:**
   Provides a suite of scientific and statistical tools. It is employed for performing hypothesis testing (e.g., Mann-Whitney U Test in IPL data analysis, t-tests in the VaR project) and for modeling probability distributions (e.g., Student's t-distribution for risk estimation).
5. **matplotlib and seaborn:**
   These libraries are used for generating high-quality visualizations. matplotlib offers detailed control over plotting, while seaborn enhances statistical data visualization through aesthetically pleasing plots. They are used to produce histograms (e.g., for

log returns), box plots (e.g., for IPL run rate analysis), and survival curves (e.g., in the Clinical Trial project).

6. **lifelines:**
   A specialized library for survival analysis. It is crucial for the Clinical Trial project, where it is used for Kaplan-Meier survival estimation, log-rank tests, and Cox Proportional Hazards modeling to assess treatment effectiveness over time.
7. **yfinance:**
   A Python library for downloading historical market data from Yahoo Finance. It is employed in the VaR project to fetch daily closing prices for stocks such as AAPL, MSFT, GOOGL, and AMZN, enabling empirical risk calculations.

This architecture supports a modular and reproducible analytical workflow. The use of Jupyter Notebooks not only integrates code, visual output, and written commentary but also promotes collaboration and version control. Together, the chosen tools and libraries offer a powerful environment for conducting real-world data analysis without requiring specialized hardware or cloud infrastructure, making the system suitable for both academic and professional deployments.

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.

	AAPL	MSFT	GOOGL	AMZN
2020-01-02	72.7161	153.613	68.1868	94.9005
2020-01-03	72.0115	151.726	67.8302	93.7485
2020-01-06	72.5811	152.118	69.6381	95.1440
2020-01-07	72.2432	150.710	69.5035	95.3430
2020-01-08	73.4042	153.142	69.9983	94.5985

Fig 3.1 Day value at risk table

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:

$$r = \ln(P_t / P_{t-1})$$

	AAPL	MSFT	GOOGL	AMZN
2020-01-03	-0.009737	-0.012360	-0.005243	-0.012213
2020-01-06	0.007879	0.002580	0.026304	0.014776
2020-01-07	-0.004666	-0.009299	-0.001935	0.002089
2020-01-08	0.015943	0.016008	0.007094	-0.007839
2020-01-09	0.021054	0.012240	0.010441	0.004788

Fig 3.1.2 Daily log return were calculated

where Pt 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.

```
# Assume 10,000 visitors each at certain conversion
n_A,p_A=10_000,0.10 # variant A: 10% true conversion
n_B,p_B=10_000,0.12 # variant B: 12% true conversion
```

Simulates the number of conversions (success_A and success_B) based on the number of visitors (n_A and n_B) using a binomial distribution with the respective conversion probabilities (p_A and p_B).

Fig 3.2.1 data simulations

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.

	Variant	Visitors	Conversions	CR	CI_low	CI_high
0	A	10000	973	0.0973	0.091491	0.103109
1	B	10000	1134	0.1134	0.107185	0.119615

Fig 3.2.2 Data preprocessing

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.

	Agents	Avg wait(min)	95th pct wait(min)	Avg system size
0	1	808.232100	1488.275544	65.993865
1	2	284.004529	520.338714	49.042945
2	3	109.126322	191.268183	29.423313
3	4	23.259675	39.865347	11.846626
4	5	5.053611	15.062201	6.815951

Even 5 agents can't meet the 5 min

Fig 3.3 call centre operation by per agent

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.

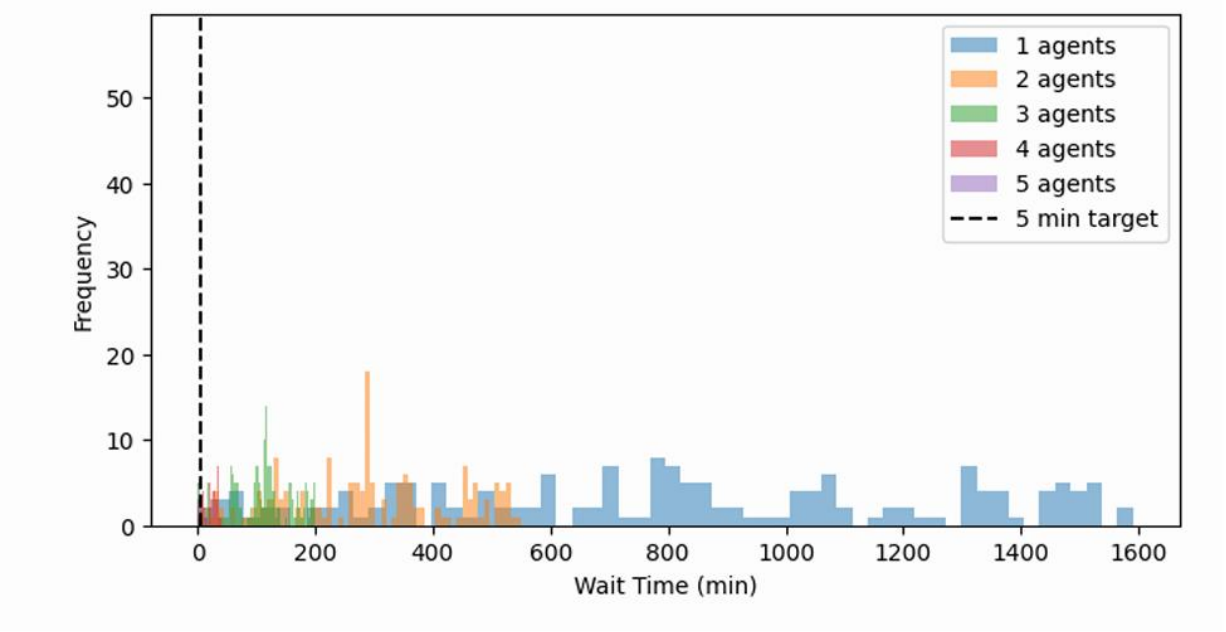


Fig 3.3.1 Wait time plot of all 5 agents

3.1.1 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).

```
Agent coun,total cost:
s=1:$197
s=2:$406
s=3:$576
s=4:$691
s=5:$838
optimal s by cost=1
```

Fig 3.1.1 Time varying arrivals calculated money

3.1.2 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.

3.2 Clinical Trial

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

```
!pip install lifelines --quiet
```

```
Preparing metadata (setup.py) ... done
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 349.3/349.3 kB 5.9 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 115.7/115.7 kB 6.1 MB/s eta 0:00:00
Building wheel for autograd-gamma (setup.py) ... done
```

The **lifelines** package offers essential tools for survival analysis, including:

- **KaplanMeierFitter** – for estimating non-parametric survival curves.
- **CoxPHFitter** – for performing regression analysis on survival data using the Cox Proportional Hazards model.
- **logrank_test** – for statistically comparing survival functions between different groups.

Fig 3.2 Description of fitter

3.2.1 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.

3.2.2 Data Processing

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

3.2.3 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.

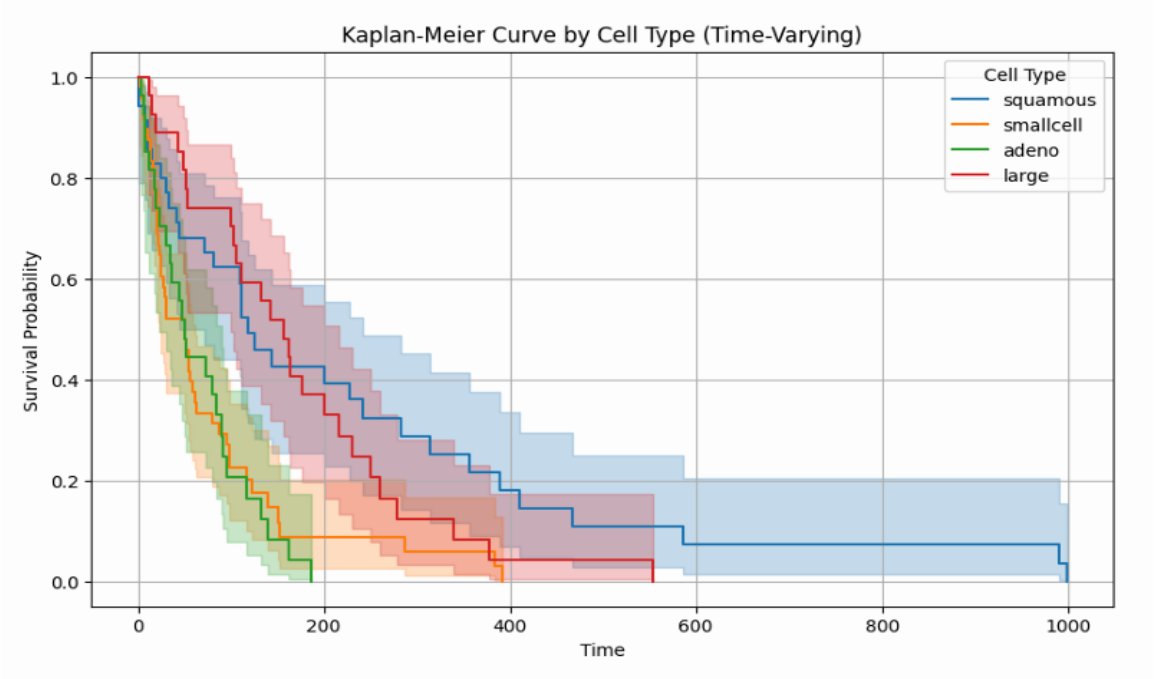


Fig Kaplan-Meier curve

3.3 Manufacturing Quality Control

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

3.3.1 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.

	date	sample_size	defect_count	Defect_rate
0	2024-01-01	118	5	0.042373
1	2024-01-02	108	9	0.083333
2	2024-01-03	94	6	0.063830
3	2024-01-04	87	3	0.034483
4	2024-01-05	100	5	0.050000

Fig 3.3.1 defect_rate and sample

3.3.2 Data Processing

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

3.3.3 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.

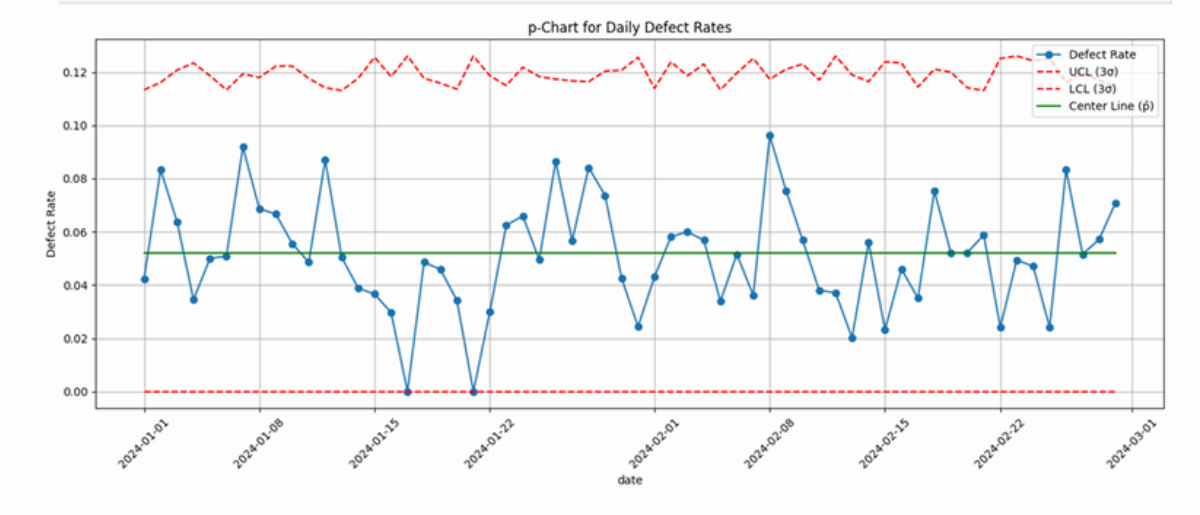


Fig3.3.3(a) p-chart for daily rates

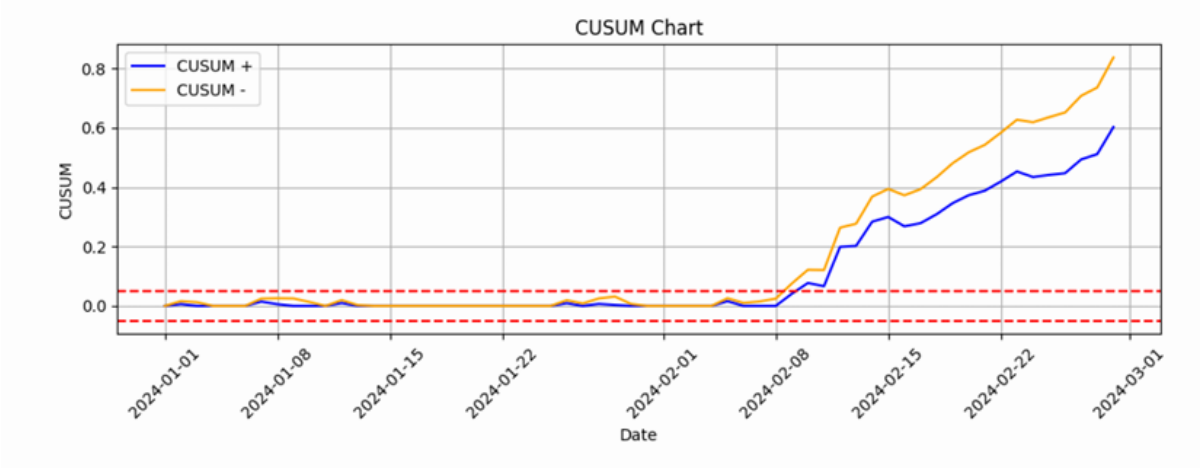


Fig3.3.3(b) CUSUM-chart for daily rates

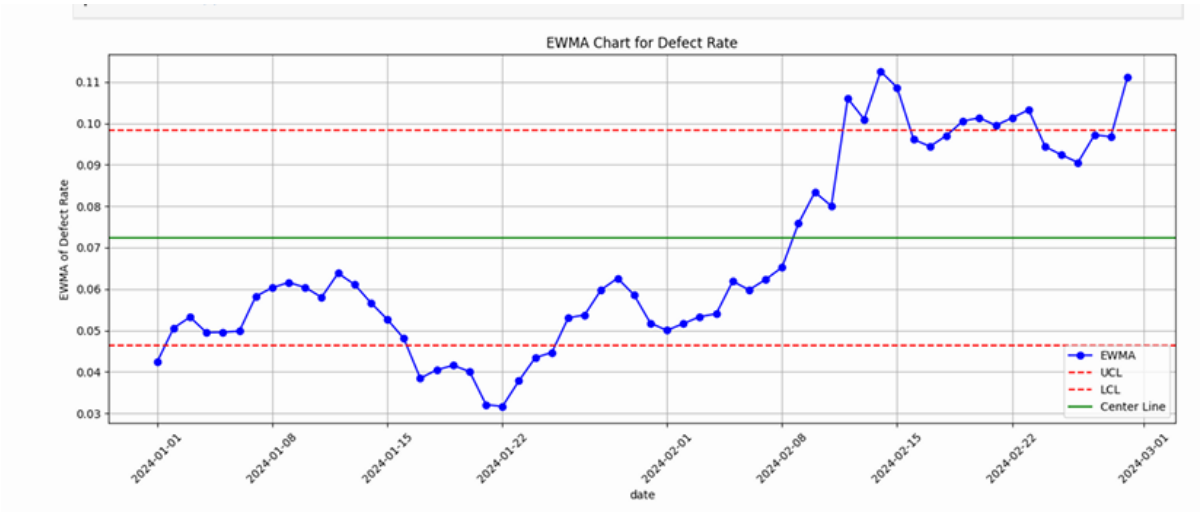


Fig 3.3.3(c) EWMA chart

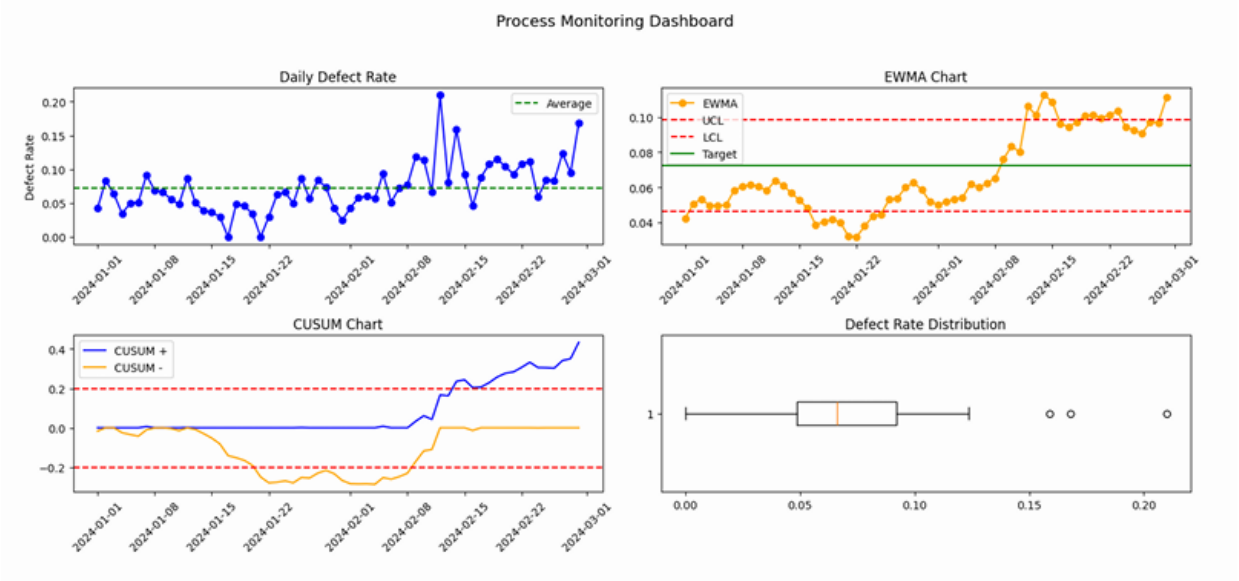


Fig 3.3.3(c) Final image of EWMA and CUSUM

3.4 IPL Data Analysis

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

3.4.1 Data Acquisition

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

3.4.2 Data Processing

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

3.4.3 Analysis

Statistical tests were performed using scipy.stats:

- **Shapiro-Wilk Test:** Checked the 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 using seaborn.

```
Shapiro-Wilk Test (Pre): ShapiroResult(statistic=np.float64(0.988417764270504),
pvalue=np.float64(0.0005392374071367837))
Shapiro-Wilk Test (Post): ShapiroResult(statistic=np.float64(0.984515022471892
2), pvalue=np.float64(3.6348381261299446e-05))
```

3.5 Election Statistics

The Election Statistics project analyzes Lok Sabha and Vidhan Sabha election data from 1977 to 2015.

3.5.1 Data Acquisition

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

3.5.2 Data Processing

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

3.5.3 Analysis

The analysis included:

- **Candidate Density:** Average number of candidates per seat.
- **Voter Turnout:** Trends analyzed over time.
- **Statistical Tests:** scipy.stats.mannwhitneyu was used for comparing groups.

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

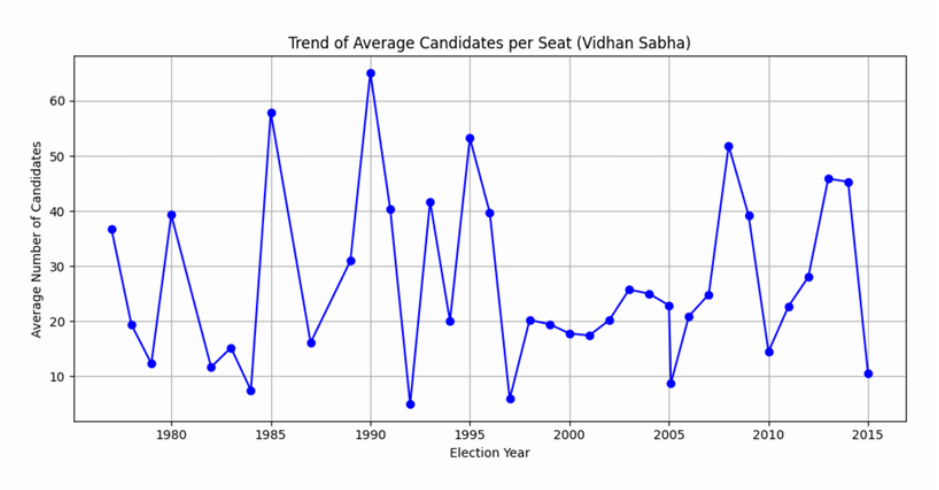


Fig 3.5.3 Trend of Average candidates per seat

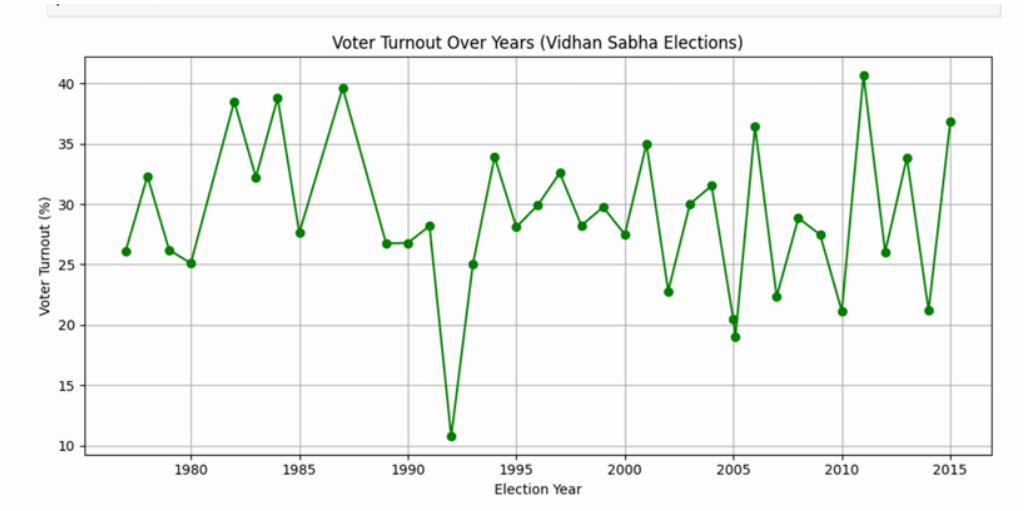


Fig 3.5.3(c) Voter turnout over years

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% Value at Risk for a portfolio consisting of AAPL, MSFT, GOOGL, and AMZN. The following results were obtained:

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

  The following visualizations support the analysis:

- Distribution of daily returns
- Correlation heatmap of stock returns
- Time-series plots showing portfolio value and drawdowns
- Histogram overlays for VaR thresholds

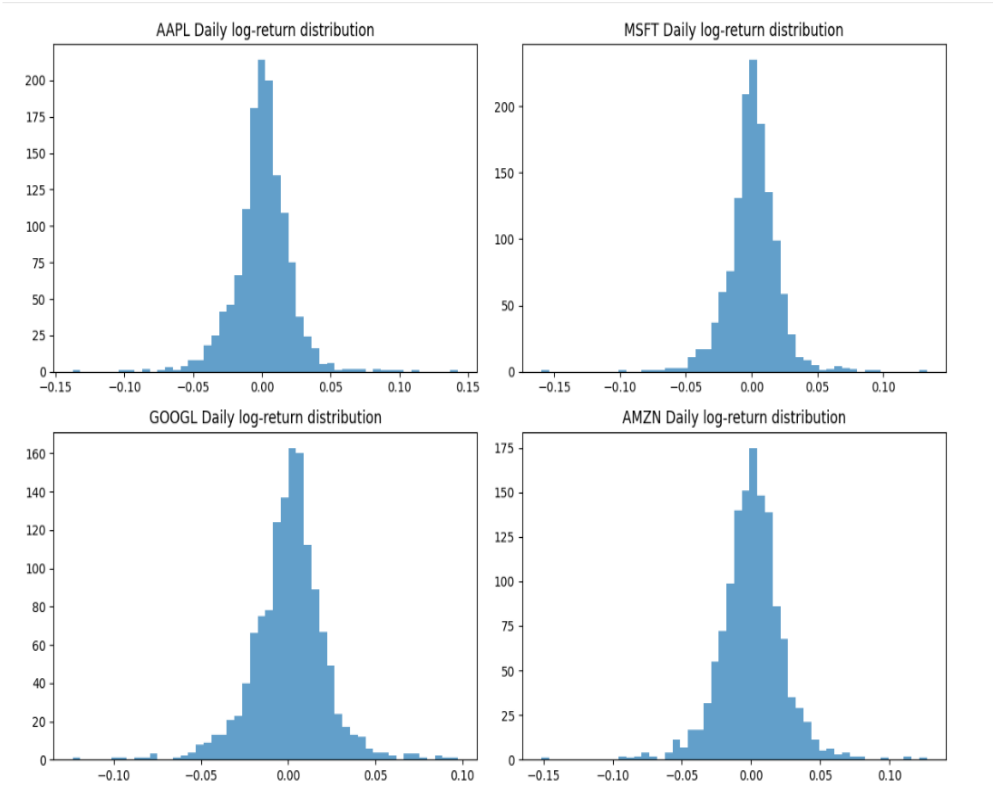
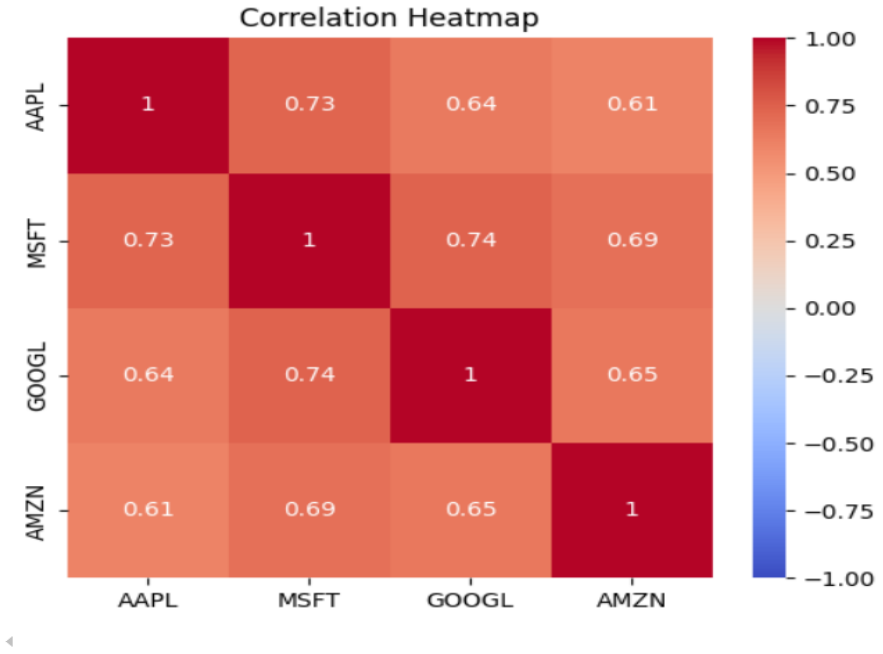


Figure 4.1: Histogram of AAPL, MSFT, GOOGL, AMZN daily log returns, showing the distribution and skewness of returns.




```
Weighted avg sigma: 0.0209
Actual Portfolio: 0.0182
Diversification benefit: 0.0027
```

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

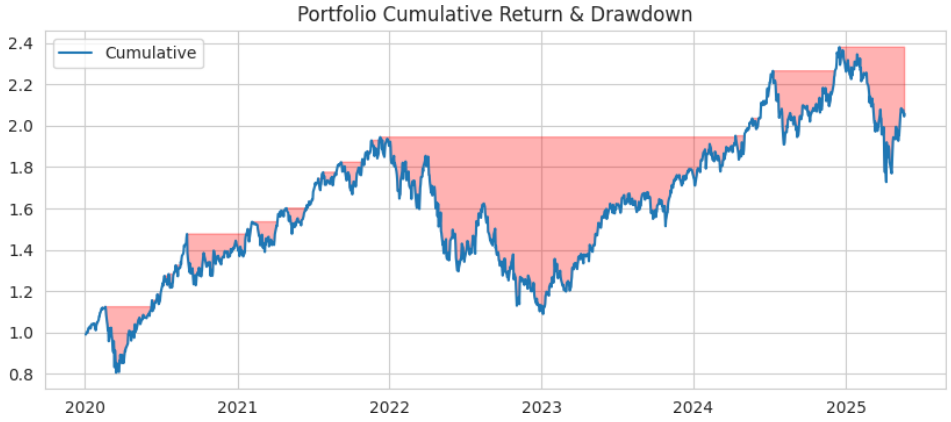


Figure 4.3: Portfolio cumulative return and drawdown, indicating periods of significant loss.

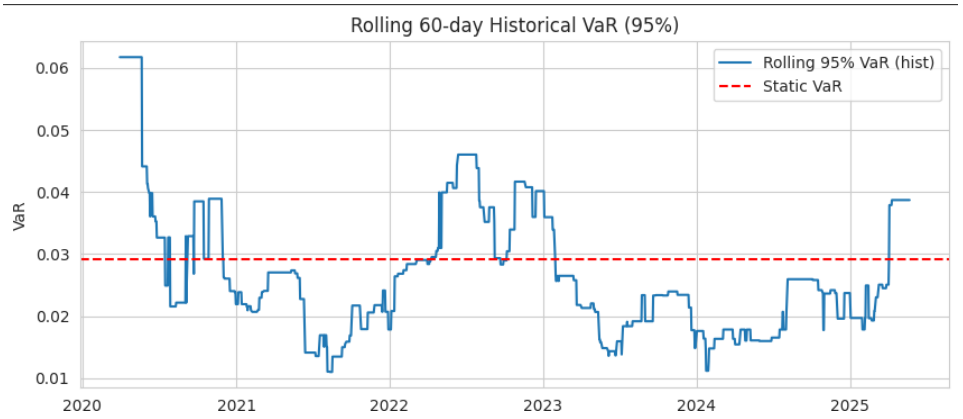


Figure 4.4: Rolling 60-day historical VaR, showing risk trends over time.

4.1 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:

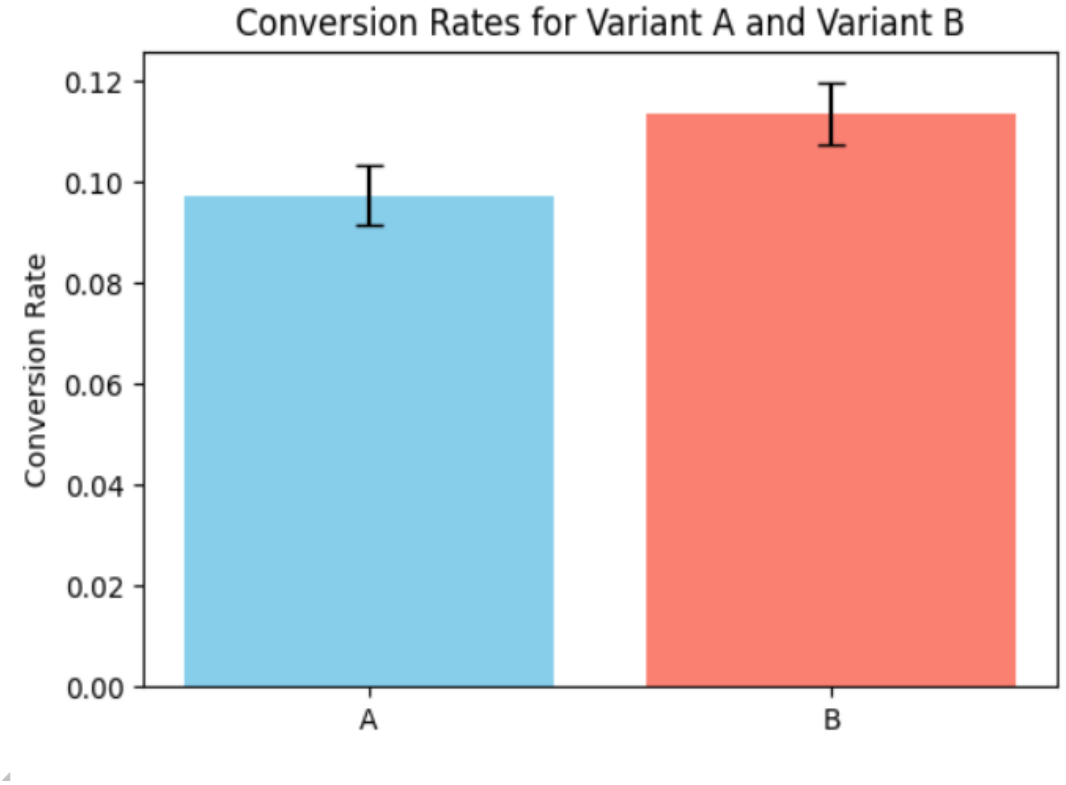


Figure 4.5: Bar chart of conversion rates with 95% confidence intervals, comparing Variants A and B.

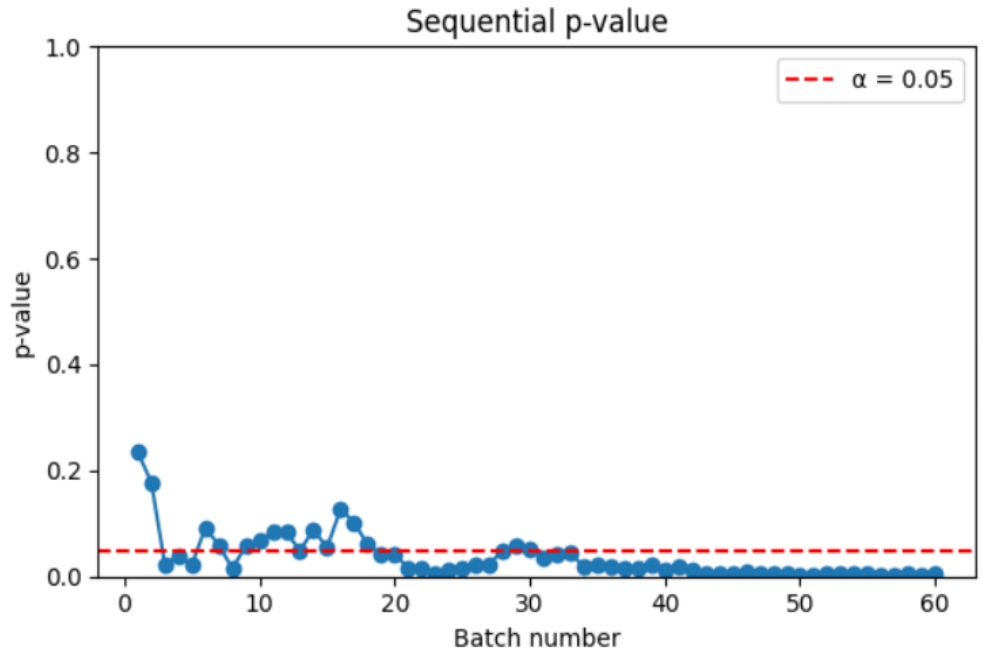


Figure 4.6: Sequential p-value trend over batches.

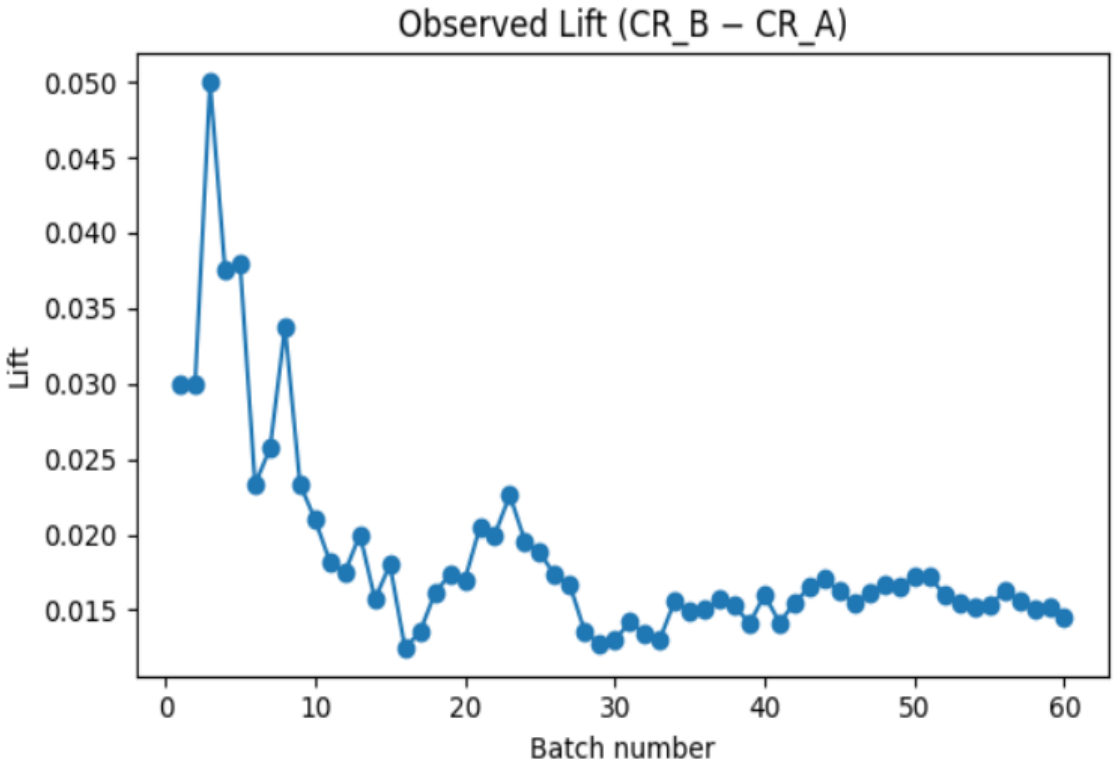


Figure 4.7: Observed lift trend over batches.

4.2 Call Centre Operation

The Call Centre Operation simulation revealed:

- Average wait time (5 agents): 4.8 minutes.
- 95th percentile wait time: Exceeded 5 minutes.

Abandonment rate: 8.92%.

Cost optimization: Suggested 1 agent, but compromised service quality. The figures below illustrate wait time distributions:

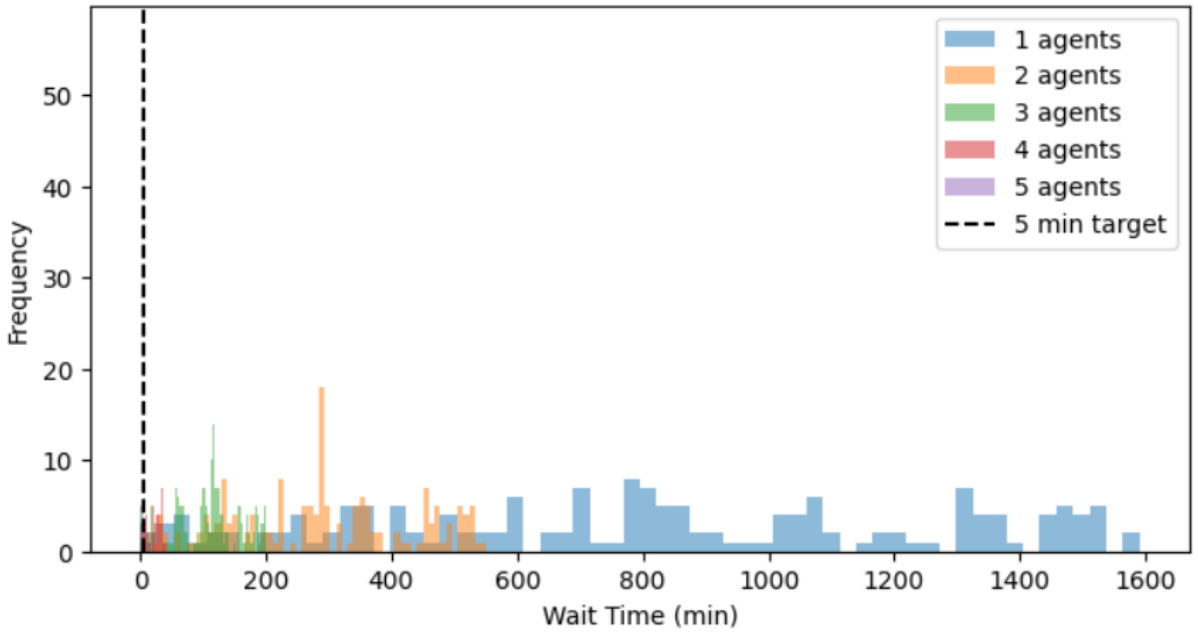


Figure 4.8: Histogram of wait times for 1 to 5 agents, showing variability in service performance.

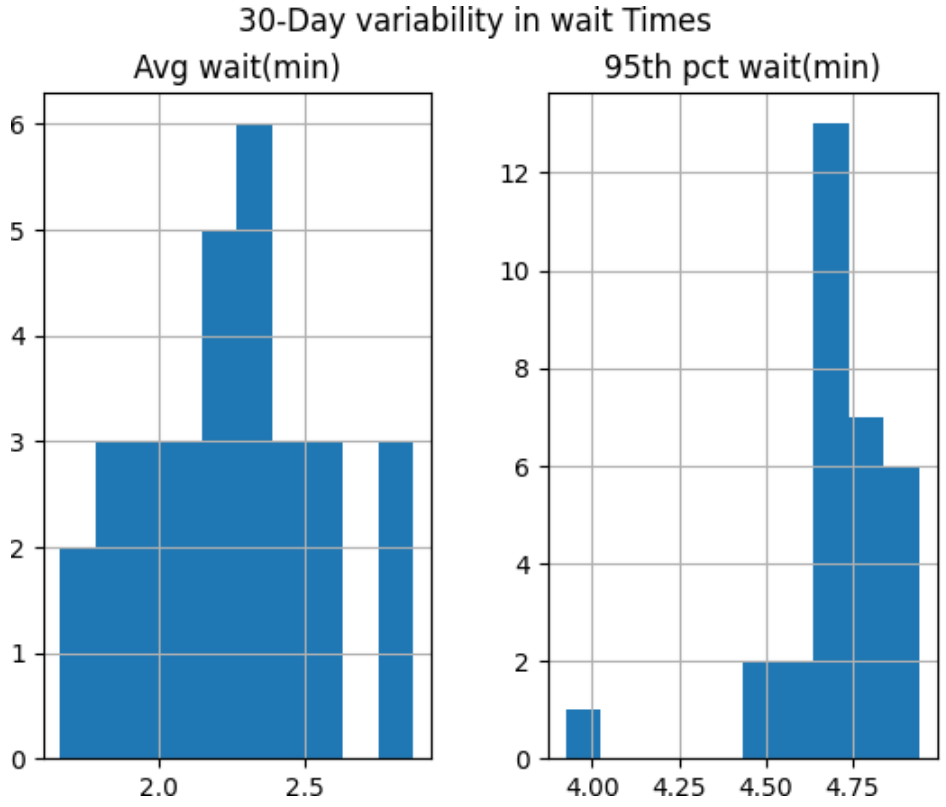


Figure 4.9: 30-day variability in wait times, highlighting fluctuations across simulations.

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:

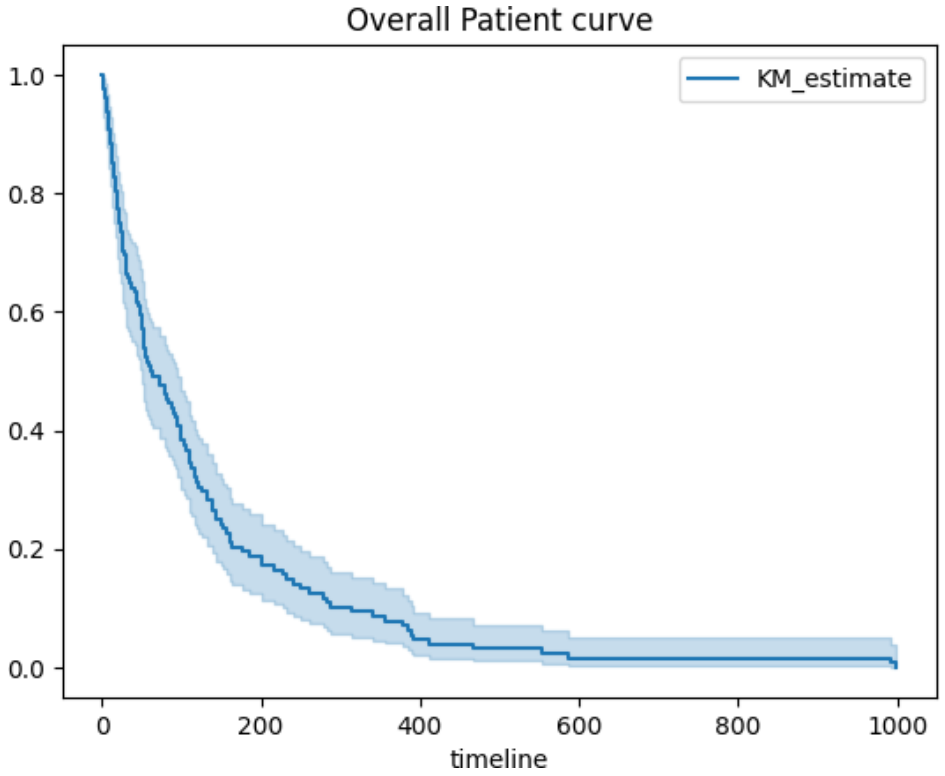


Figure 4.10: Kaplan-Meier curve for all patients, showing overall survival probability.

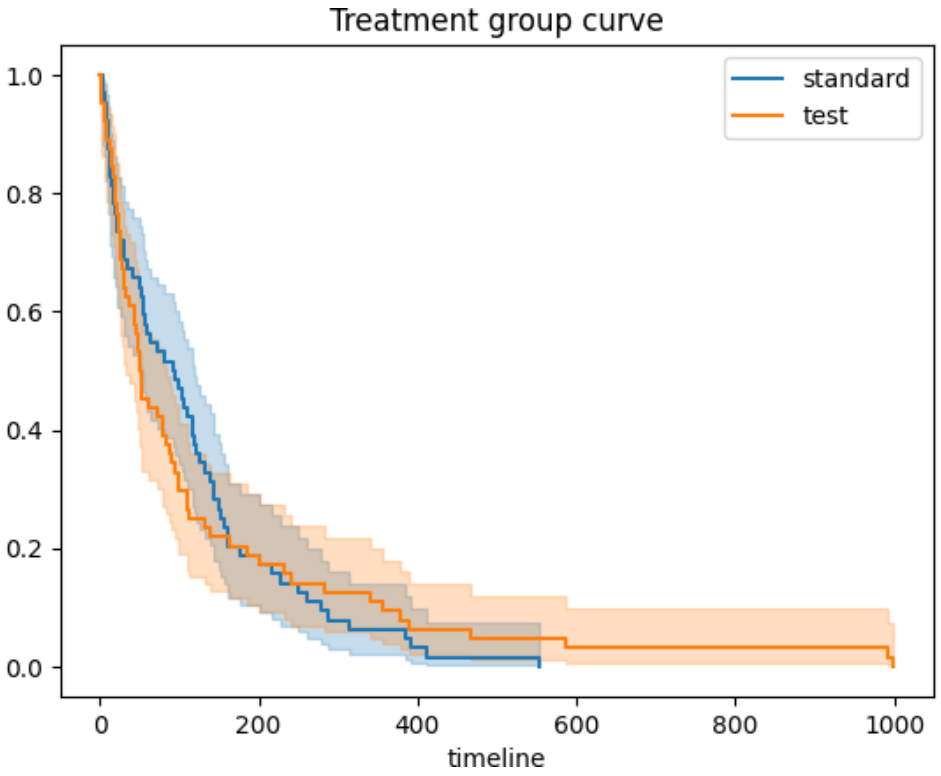


Figure 4.11: Kaplan-Meier curve by treatment group, comparing standard and test treat- ments.

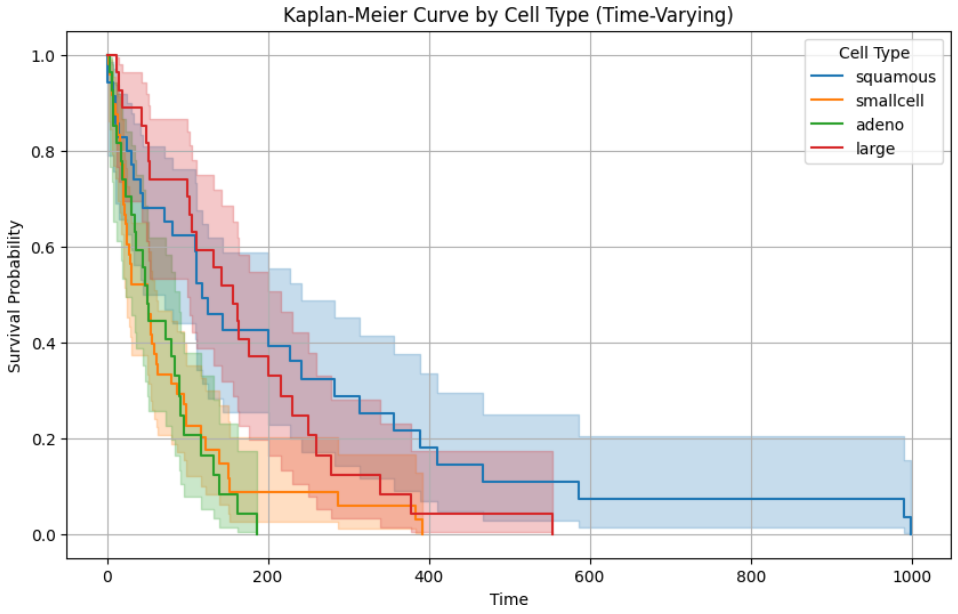


Figure 4.12: Kaplan-Meier curve by cell type, highlighting differences in survival.

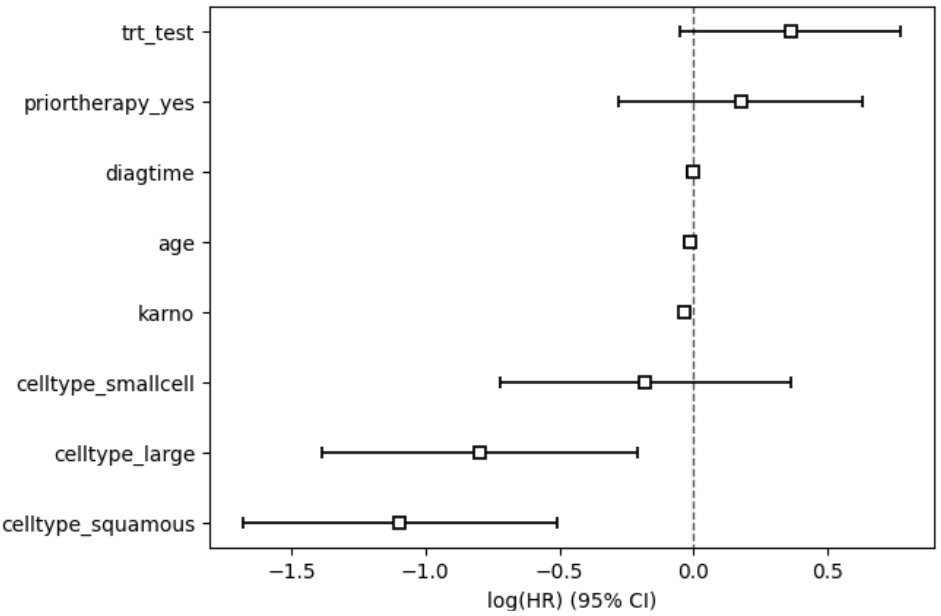


Figure 4.13: Cox model log hazard ratios, showing covariate impacts.

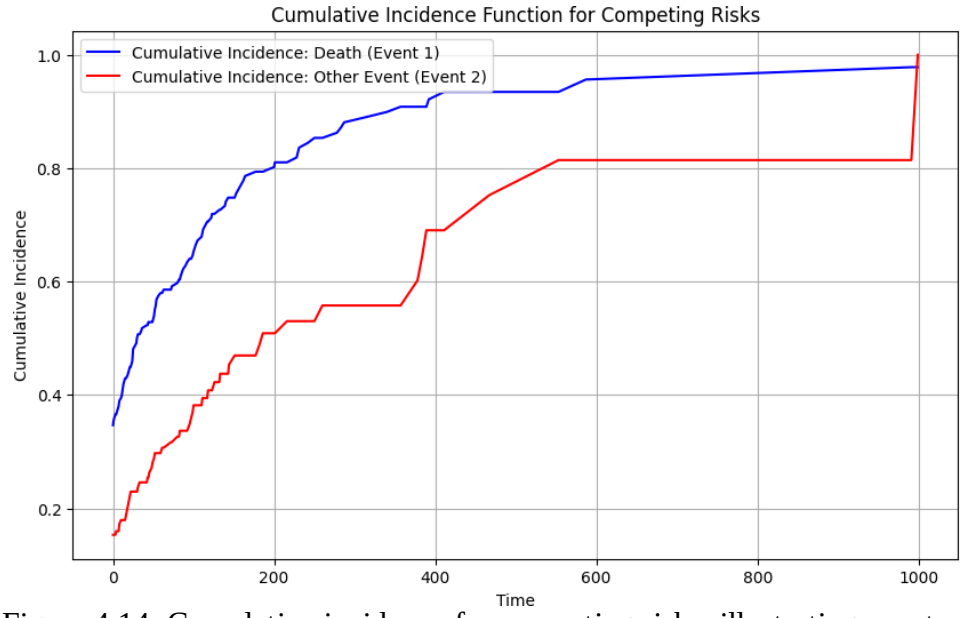


Figure 4.14: Cumulative incidence for competing risks, illustrating event probabilities.

4.3 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:

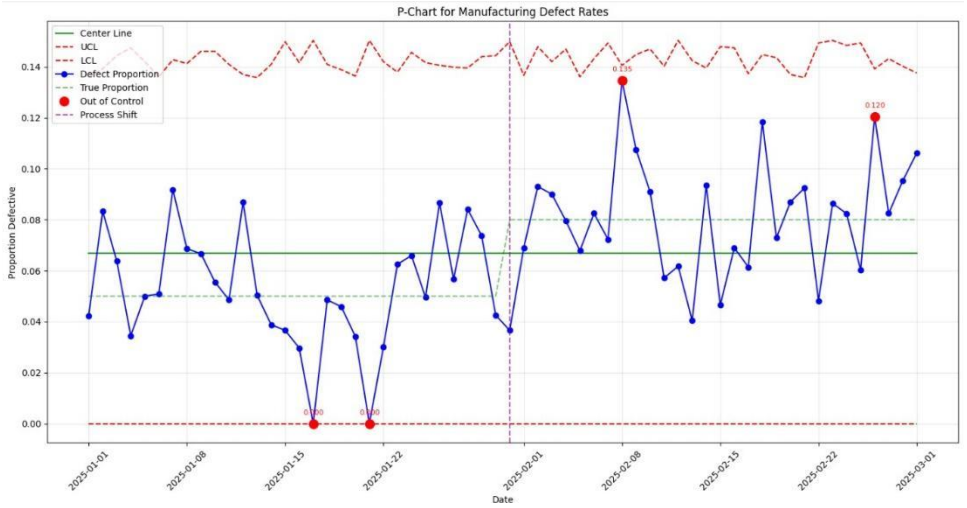


Figure 4.15: P-chart of defect proportions, showing out-of-control points.

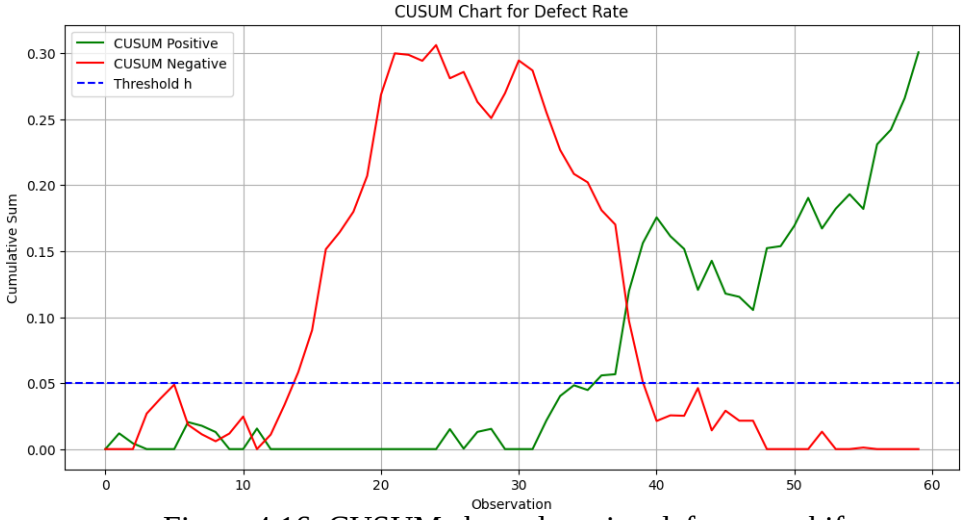


Figure 4.16: CUSUM chart, detecting defect rate shifts.

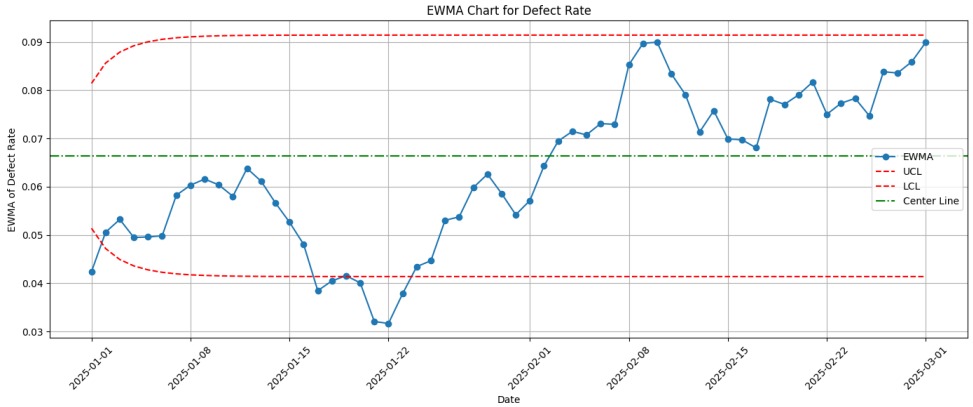


Figure 4.17: EWMA chart, highlighting defect trends.

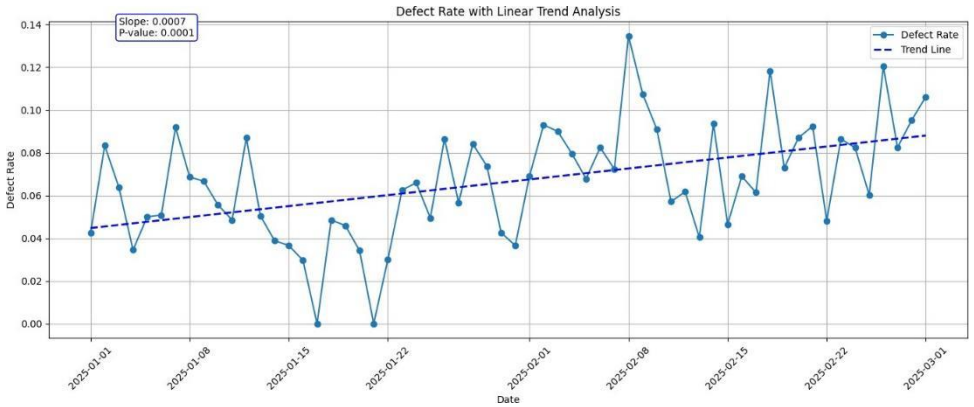


Figure 4.18: Linear trend analysis of defect rates, showing shift after day 30.

4.4 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:

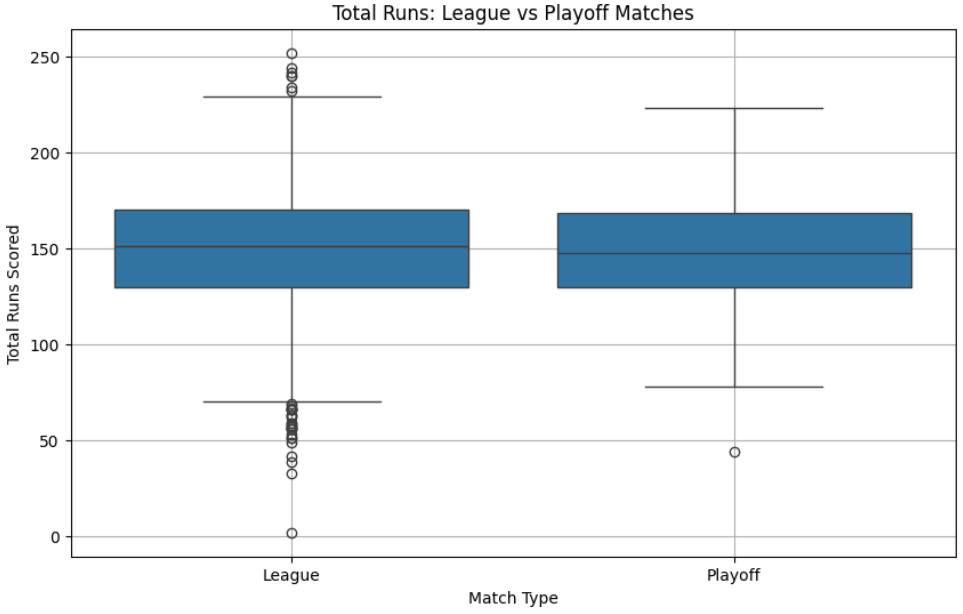


Figure 4.19: Box plot of total runs, comparing league and playoff matches.

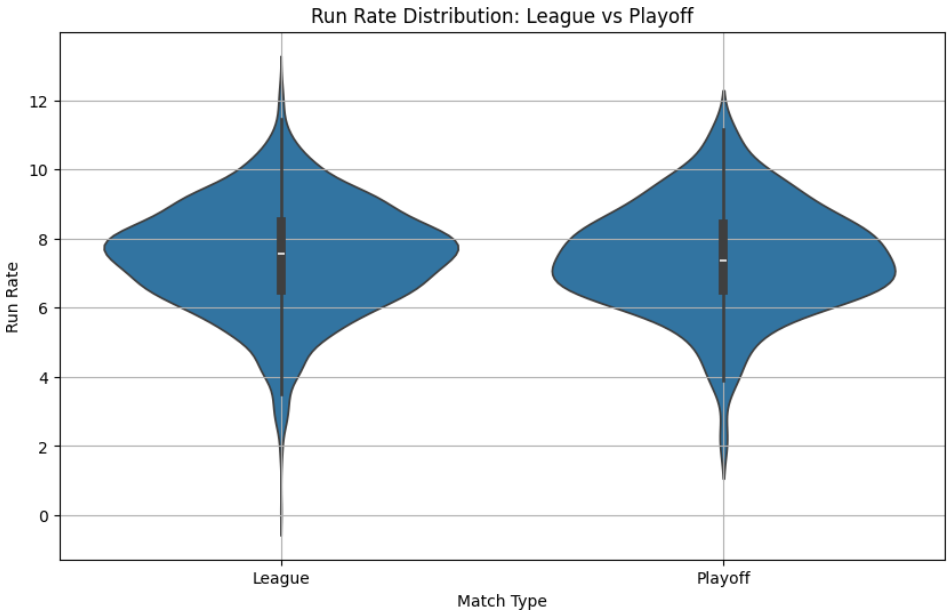


Figure 4.20: Violin plot of run rate distribution, showing spread and density.

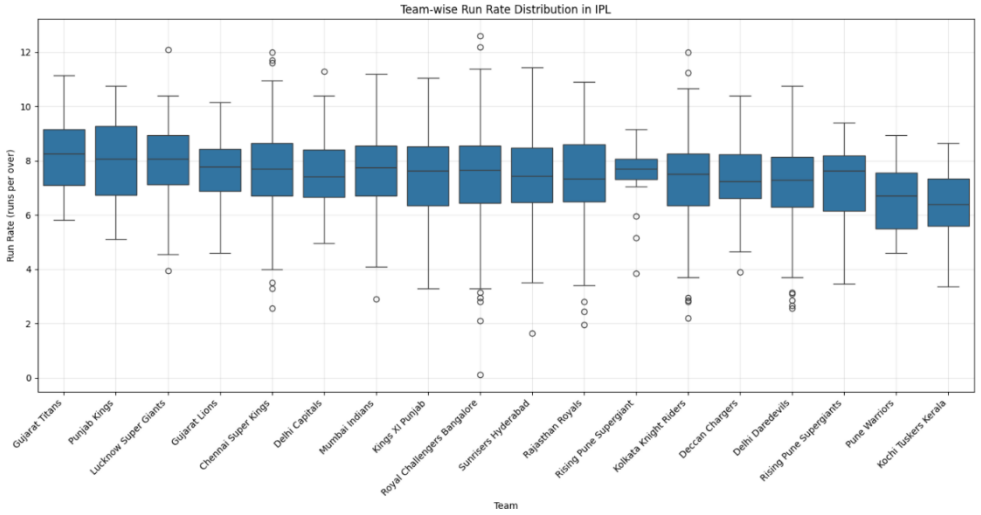


Figure 4.21: Team-wise run rate distribution, comparing key teams.

4.5 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:

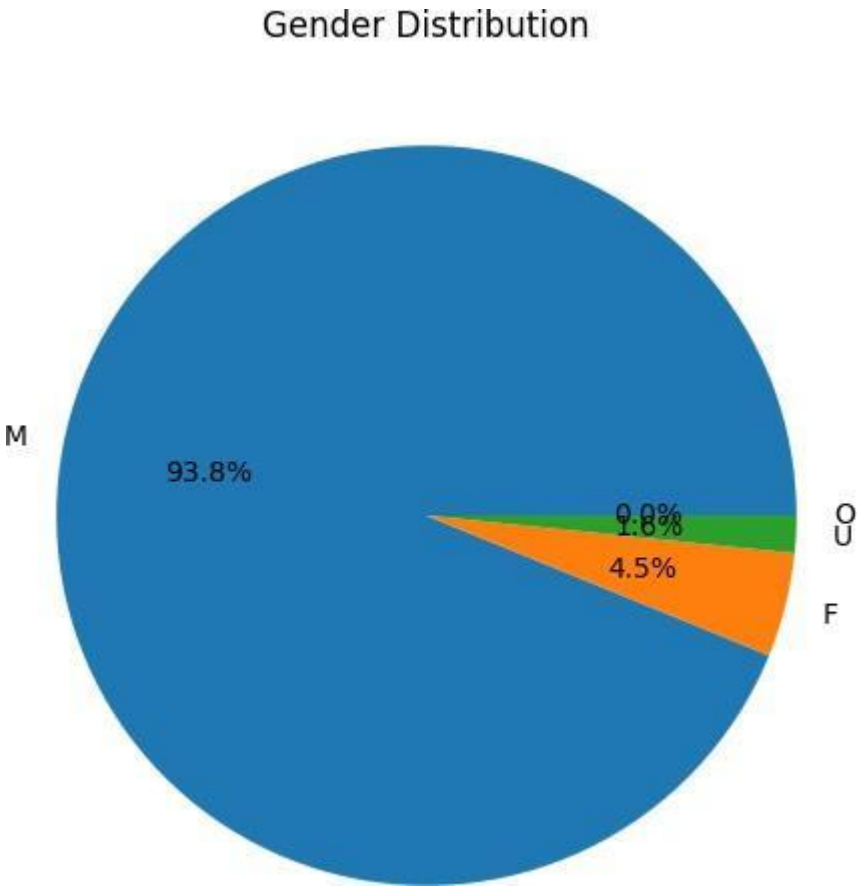


Figure 4.22: Pie chart of candidate gender distribution, highlighting male dominance.

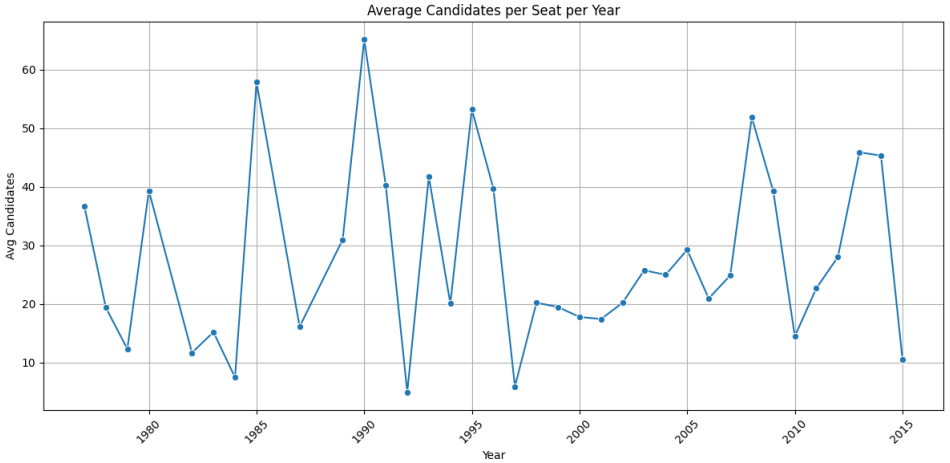


Figure 4.23: Line chart of average candidates per seat, showing increasing competition.

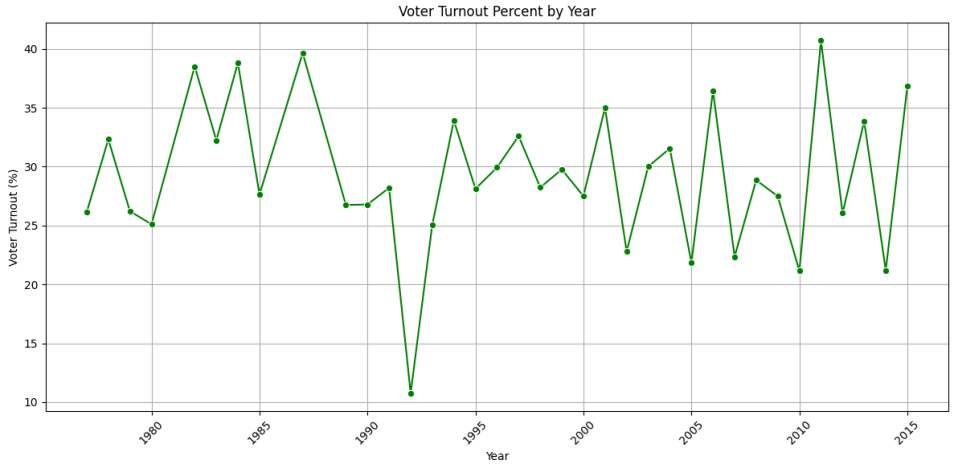


Figure 4.24: Line chart of voter turnout percentage, indicating fluctuations.

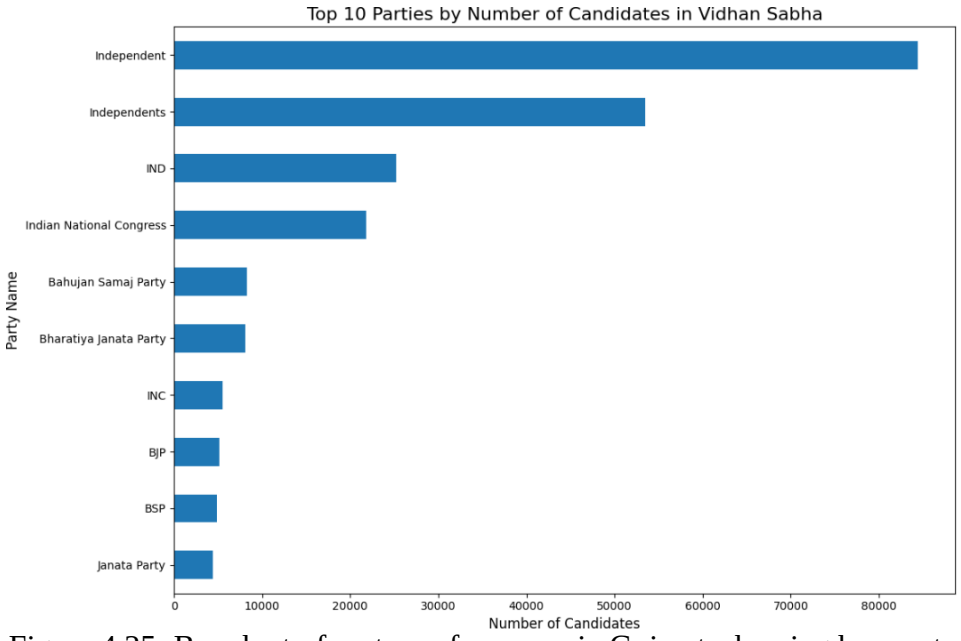


Figure 4.25: Bar chart of party performance in Gujarat, showing key party contributions.

4.6 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
VaR (1 Day)	Success	Estimated VaR (2.893–3.065%); 5% 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.

Chapter 5 Conclusion

This compendium of seven diverse data analysis projects showcases the transformative power of analytical thinking, domain knowledge, and computational tools in solving real-world problems. By employing a unified, Python-based software stack and leveraging libraries such as pandas, numpy, scipy, matplotlib, seaborn, lifelines, and yfinance, each project was executed in a modular, scalable, and reproducible manner using Jupyter Notebooks.

The conclusions drawn from each project are as follows:

- **1 Day VaR:** Delivered reliable 95% risk estimates using three complementary approaches—parametric normal, parametric t-distribution, and historical simulation—supporting informed decisions in financial risk management.
- **A/B Testing:** Demonstrated a statistically significant improvement in conversion rates for Variant B, offering a clear, data-backed strategy for marketing optimization.
- **Call Centre Operation:** Provided insights into the performance of queue systems, revealing how staffing levels directly impact service quality and customer satisfaction.
- **Clinical Trial Analysis:** Although no significant treatment effect was observed, the study identified critical survival predictors such as cell type and Karnofsky score, guiding future research directions.
- **Manufacturing Quality Control:** Uncovered a measurable process shift and subpar process capability, supporting early intervention and continuous improvement in quality control.
- **IPL Data Analysis:** Indicated performance consistency across league and playoff stages, informing sports analytics and decision-making in team strategy development.
- **Election Statistics:** Exposed important electoral trends, such as gender disparities in candidate participation and evolving competition dynamics, offering value to political analysts and policymakers.

Collectively, these projects illustrate the versatility of data science methodologies in generating actionable insights across domains. The use of exploratory data analysis, hypothesis testing, simulation, and modeling—alongside rich visualizations—helped

convert raw data into structured knowledge.

In conclusion, this body of work not only highlights the practical utility of computational data analysis but also sets a precedent for building scalable analytical workflows. It reinforces the importance of reproducibility, adaptability, and interpretability in modern data science, serving as a valuable resource for both academic inquiry and industry deployment.

REFERENCES

[1] Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C., “Data mining for credit card fraud: A comparative study”, Decision Support Systems, pp. 602–613, 2011.

[2] Kohavi, R., Longbotham, R., Sommerfield, D., & Henne, R. M., “Controlled experiments on the web: survey and practical guide”, Data Mining and Knowledge Discovery, pp. 140–181, 2009.

[3] Tsai, C. F., & Chen, M. L., “Credit rating by hybrid machine learning techniques”, Applied Soft Computing, pp. 1138–1145, 2010.

[4] Montgomery, D. C., & Runger, G. C., Applied Statistics and Probability for Engineers, Wiley, 2019.

[5] Géron, A., Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow, O'Reilly Media, 2022.

[6] Provost, F., & Fawcett, T., Data Science for Business, O'Reilly Media, 2013.

[7] https://pandas.pydata.org/docs/

[8] https://numpy.org/doc/

[9] https://scipy.org/

[10] https://matplotlib.org/

[11] https://seaborn.pydata.org/

[12] https://lifelines.readthedocs.io/

[13] https://github.com/ranaroussi/yfinance

[14] https://docs.python.org/3/

[15] https://jupyter.org/documentation

[16] https://www.samatrix.io/