INTRODUCTION

- Challenges in Insurance Complaints: Frequent issues include claim denials, premium disputes, and policy misunderstandings, leading to dissatisfaction and loss of trust.
- Technologies Used: Leverage PySpark, Spark MLlib, and NoSQL tools like Databricks DBFS.
- Machine Learning Integration: Build models to predict resolution times and streamline complaint processes.
- Key Outcomes: Deliver insights into complaint trends, improve resolution efficiency, and enhance customer satisfaction and trust.

PROBLEM STATEMENT

The inefficient handling of insurance complaints, including delays, claim denials, and policy misunderstandings, leads to customer dissatisfaction and mistrust. This project addresses these challenges by leveraging Big Data technologies and machine learning to analyze complaint patterns, improve resolution processes, and enhance overall customer satisfaction and trust in the insurance industry.

OBJECTIVES

- Identify Complaint Patterns: Discover frequent causes of complaints (e.g., claims, premiums, policy terms) and **highlight areas needing attention** for different insurance products.
- Predictive Modeling: Use machine learning to **predict complaint resolution times**, **aiding resource planning and efficiency**.
- Coverage-Specific Trends: Analyze complaint trends across health, accident, and group insurance to **uncover unique challenges** and **opportunities**.
- Develop Actionable Recommendations: Provide insights to improve resolution processes, enhance **customer satisfaction**, and **build trust**.

PROJECT TIMELINE

Phase	Tasks	Weeks
Phase 1:Data Understanding	Dataset structure review, column definition, and metadata Handle missing values, remove duplicates, standardize data formats, and clean anomalies.	Week 1-2
Phase 2: Data Cleaning	Handle missing values, remove duplicates, standardize data formats, and clean anomalies.	Week 1-2
Phase 3: Exploratory Data Analysis	Generate descriptive statistics, visualize trends and identify key variables	Week 2- 3
Phase 4: Model Development	Build predictive models-e.g., bridge lifespan, perform parameter tuning, and validation of results.	Week 3-4
Phase 5: Documentation and Reporting	Prepare project report, create visualizations, finalize results for presentation	Week 4 - 5

SELECTED DATASET

- **Dataset Overview**: Contains insurance complaints, resolutions, and recovery amounts, enabling pattern and trend analysis of customer grievances.
- The dataset provides detailed information on customer complaints and their resolution status, including columns for the insurance company, unique case identifier (File No.), complaint filing and resolution dates (Opened/Closed), policy types (Coverage/SubCoverage), causes of dissatisfaction (Reason/SubReason), final decision (Disposition), recovery amount (Recovery), and complaint status (Status). It also includes a calculated feature, Resolution_Time_Days, which measures the efficiency of the complaint resolution process.
- The dataset contains over 63,000 records of data and has 12 columns in total.

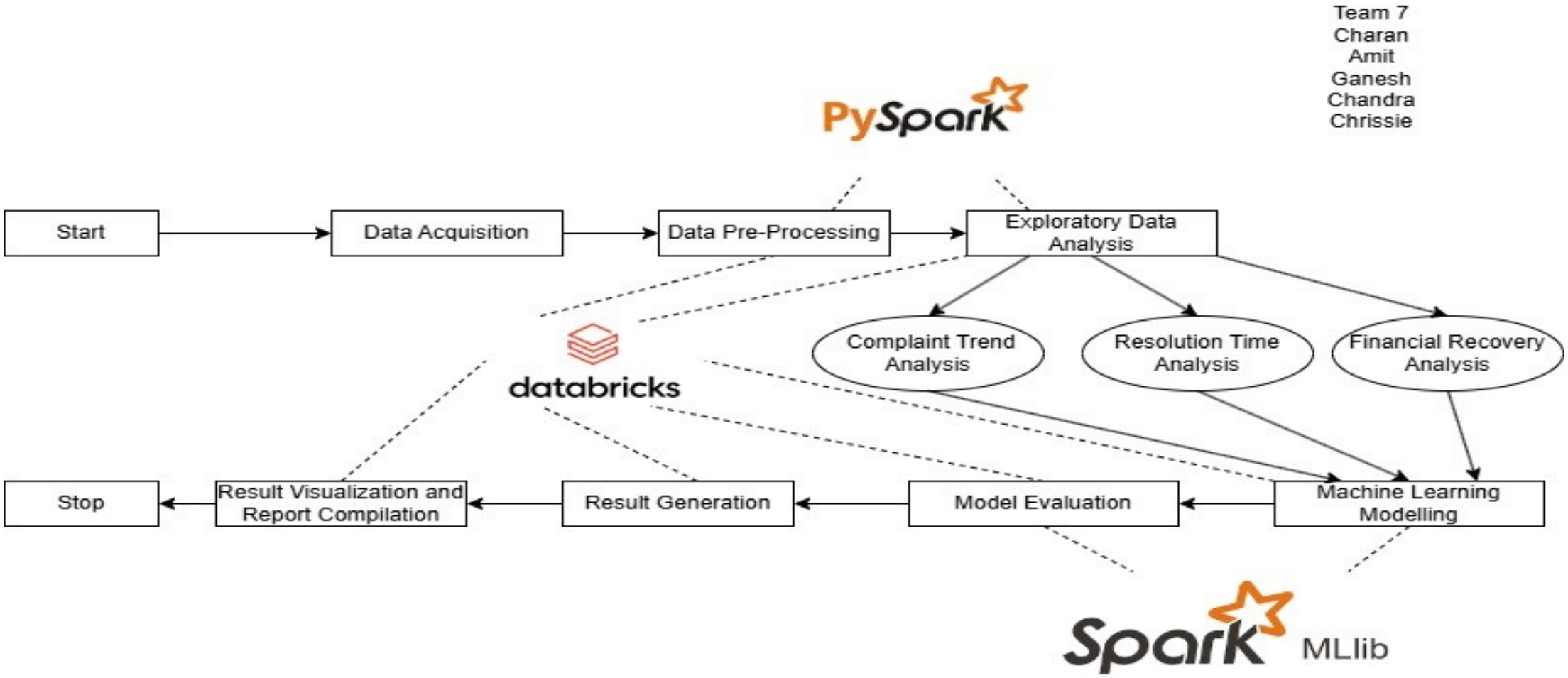
- **Purpose**: Identify common complaint causes, resolution times, and improvement areas to enhance customer satisfaction and operational efficiency.

DATASET

	Company	File No.	Opened	Closed	Coverage	SubCoverage	Reason	SubReason	Disposition	Conclusion	Recovery	Status
2	Anthem Health Plan	7045593	05/31/2022	06-02-2022	Group	Health Only	Claim Handling	Medically Necessar	Company Position S	Company Position U	0	Closed
3	Anthem Health Plan	7043381	02/28/2022	06-02-2022	Group	Health Only	Claim Handling	Provider Contract Is	Claim Settled	Satisfied	6467.3	Closed
4	Anthem Health Plan	7044860	05-03-2022	06-02-2022	A & H	Health Only	Claim Handling	Denial	Claim Settled	Claim Paid	147.58	Closed
5	Anthem Health Plan	7043381	02/28/2022	06-02-2022	Group	A & H	Claim Handling	Provider Contract Is	Claim Settled	Satisfied	6467.3	Closed
6	Anthem Health Plan	7052007	02/23/2023	03/17/2023	A & H	A & H	Marketing & Sales	Duplicate Coverage	Compromised Settle	Premium Refund	2179.32	Closed
7	Anthem Health Plan	7056820	08/16/2023	09/15/2023	A & H	Self Funded/ERISA	Claim Handling	Medical Necessity Denial			0	Closed
8	Metropolitan Life Ins	7053748	04/26/2023	05-02-2023	Individual	Long Term Care	PolicyHolder Servic	Premium/Notice		Rate Increase Expla	0	Closed
9	Anthem Health Plan	7044860	05-03-2022	06-02-2022	A & H	A & H	Claim Handling	Denial	Claim Settled	Claim Paid	147.58	Closed
10	Anthem Health Plan	7048359	10-03-2022	03-08-2023	A & H	Health Only	Claim Handling	Provider Contract Is	Company Position	Overturned	51094.84	Closed
11	GEICO General Insu	7054688	05/30/2023	06/15/2023	Individual Private Pa	Collision	Claim Handling	Claim Denial	Company Position	Substantiated	0	Closed
12	ConnectiCare Inc	7041052	11/22/2021	06-02-2022							0	Closed
13	Vigilant Insurance C	7005924	08/18/2021	04-11-2022	Homeowners	Homeowners	Claim Handling	Unsatisfactory Settlement/Offer		Company Position U	0	Closed
14	UnitedHealthcare In	7052741	03/17/2023	05-02-2023	A & H	A & H	Claim Handling	Provider Contract Is	Claim Settled	Provider Issue	1550.78	Closed
15	Anthem Health Plan	7044266	04-07-2022	06-02-2022	A & H	Health Exchange	Claim Handling	Surprise Billing	Claim Settled	Furnished Informati	0	Closed
16	Western World Insu	7042656	02-03-2022	06-02-2022	Commercial Multi-P	Fire - Real Property	Claim Handling	Claim Delays	Claim Settled	Corrective Action	311596.73	Closed
17	Anthem Health Plan	7053777	04/26/2023	06/15/2023	Group	A & H	Claim Handling	External Review			0	Closed
18	Bankers Life and Ca	7000032	01-10-2020	01/30/2020	Individual Life		Claim Handling	Other		Justified	0	Closed
19	Anthem Health Plan	7044266	04-07-2022	06-02-2022	A & H	Health Only	Claim Handling	Surprise Billing	Claim Settled	Furnished Informati	0	Closed
20	Aetna Health Inc	7007715	09/15/2021	06-02-2022							0	Closed
21	Continental Casualt	7053553	04/19/2023	05-02-2023	Individual	A & H	Claim Handling	Medical Necessity	Question of Fact/Co	Coverage Denied	0	Closed
22	Nationwide Mutual I	7048101	09/21/2022	03/17/2023	Travel		Claim Handling	Unsatisfactory Settl	Claim Settled	Additional Money Re	35.52	Closed
23	Anthem Health Plan	7043381	02/28/2022	06-02-2022	A & H	A & H	Claim Handling	Provider Contract Is	Claim Settled	Satisfied	6467.3	Closed
24	Cigna Health and Lif	7013857	01/23/2019	01/23/2019	Group	Health Only	Claim Handling	UR Case Management		Furnished Informati	0	Closed
25	Continental Casualt	7053553	04/19/2023	05-02-2023	Individual	Long Term Care	Claim Handling	Medical Necessity D	Question of Fact/Co	Coverage Denied	0	Closed
26	Cigna Health and Lif	7013857	01/23/2019	01/23/2019	Group	Health Only	Claim Handling	UR Case Management		Refer-Judicial/Attor	0	Closed
27	ConnectiCare Inc	7052648	03/15/2023	05-02-2023	Group	Health Only	Claim Handling	Claim Denial		Furnished Informati	0	Closed

SYSTEM ARCHITECTURE

Team 7
Charan
Amit
Ganesh
Chandra
Chrissie



SYSTEM AND ARCHITECTURE

Data Cleaning and Preprocessing:

- Handle missing values and resolve inconsistencies to ensure high-quality data.
- Standardize categorical variables for uniformity across the dataset.

Feature Engineering:

- Convert categorical variables into machine-readable formats using techniques like Label Encoding.
- Calculate the Resolution_Time_Days feature, which represents the time difference between the complaint creation and resolution dates.

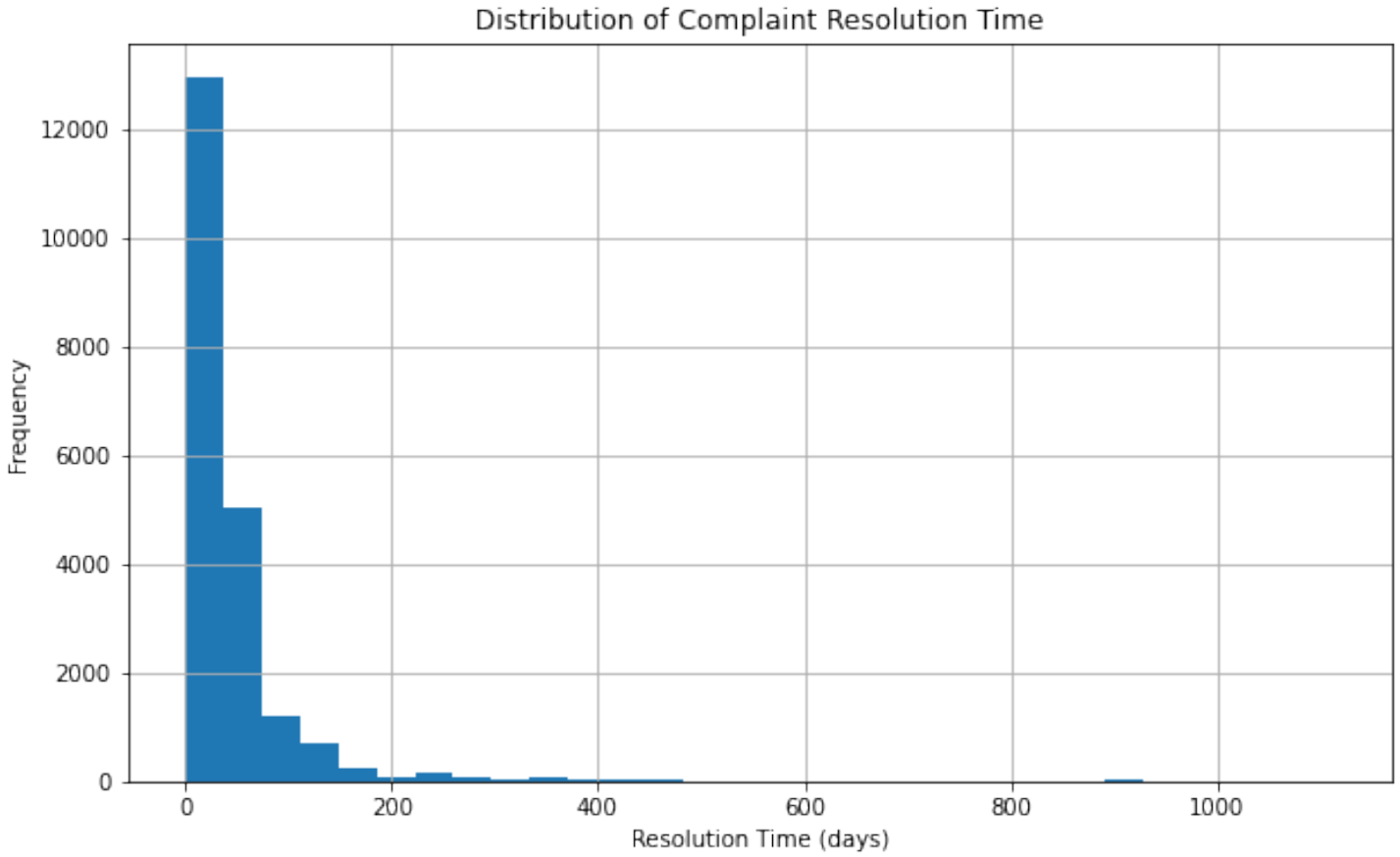
Model Training:

- Apply a variety of machine learning models, including Random Forest, Decision Tree, Logistic Regression, Support Vector Classifier, and Gradient Boosting.
- Train the models to predict complaint dispositions and identify key variables influencing resolution outcomes.

Model Evaluation:

- Assess model performance using metrics like accuracy and classification reports.
- Validate results using cross-validation techniques to ensure robustness and reliability.

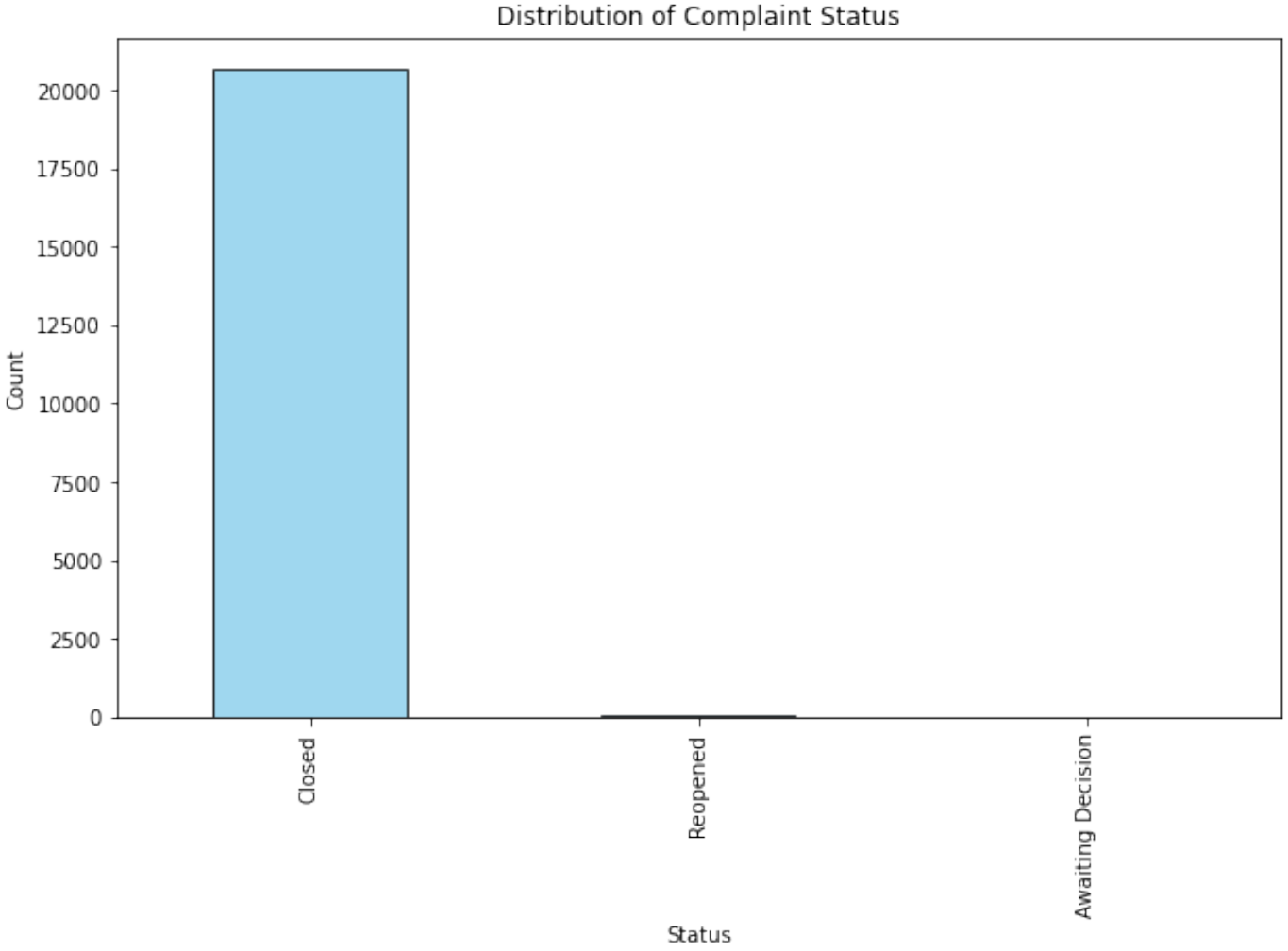
KEY INSIGHTS FROM VISUALIZATIONS



Title: Distribution of Complaint Resolution Time

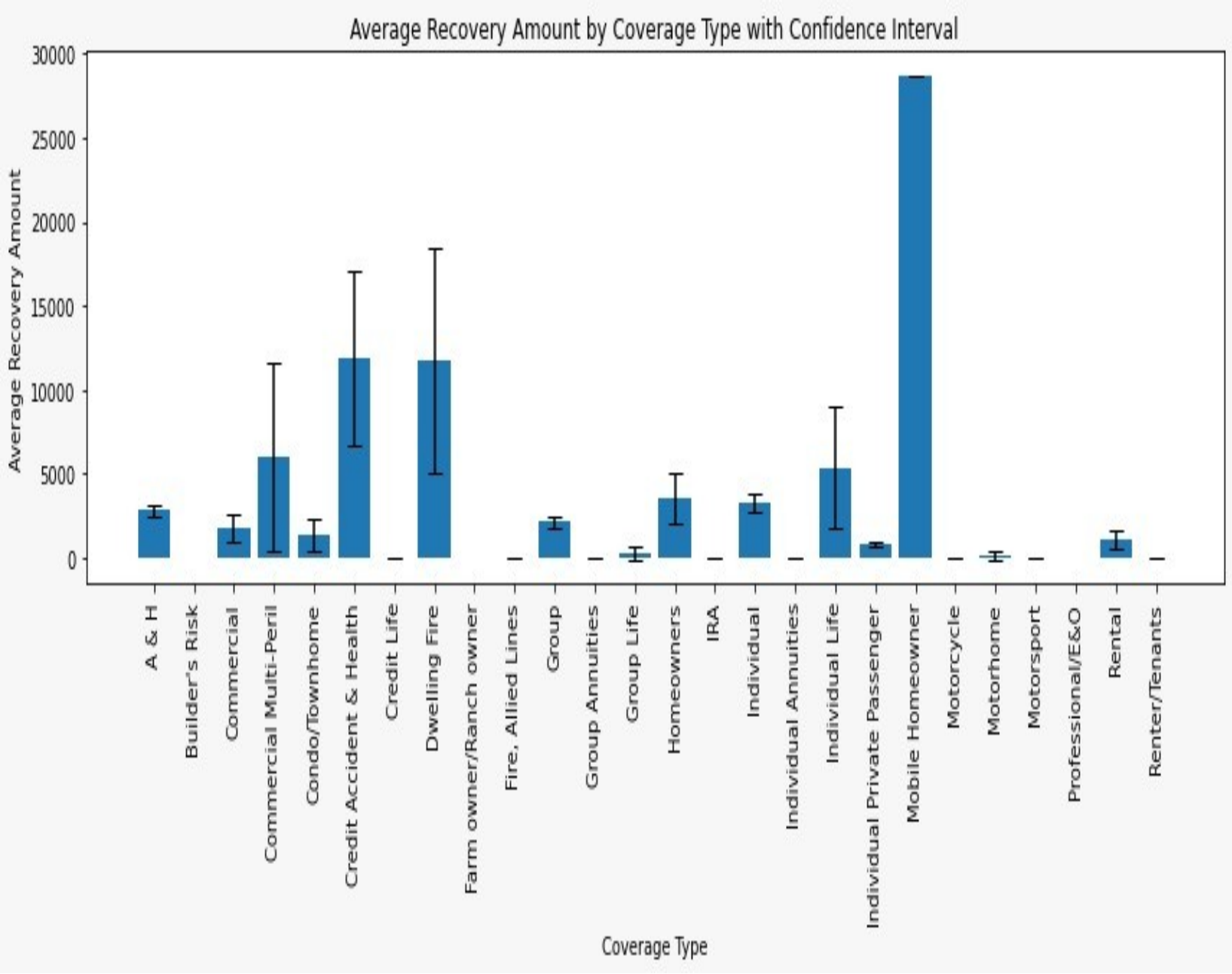
Description:

The histogram illustrates that most complaints are resolved within 50 days, reflecting efficient handling in most cases. However, outliers with resolution times beyond 200 days indicate potential inefficiencies or complexities in specific situations. These anomalies highlight the importance of identifying and addressing delays to enhance overall complaint resolution processes.

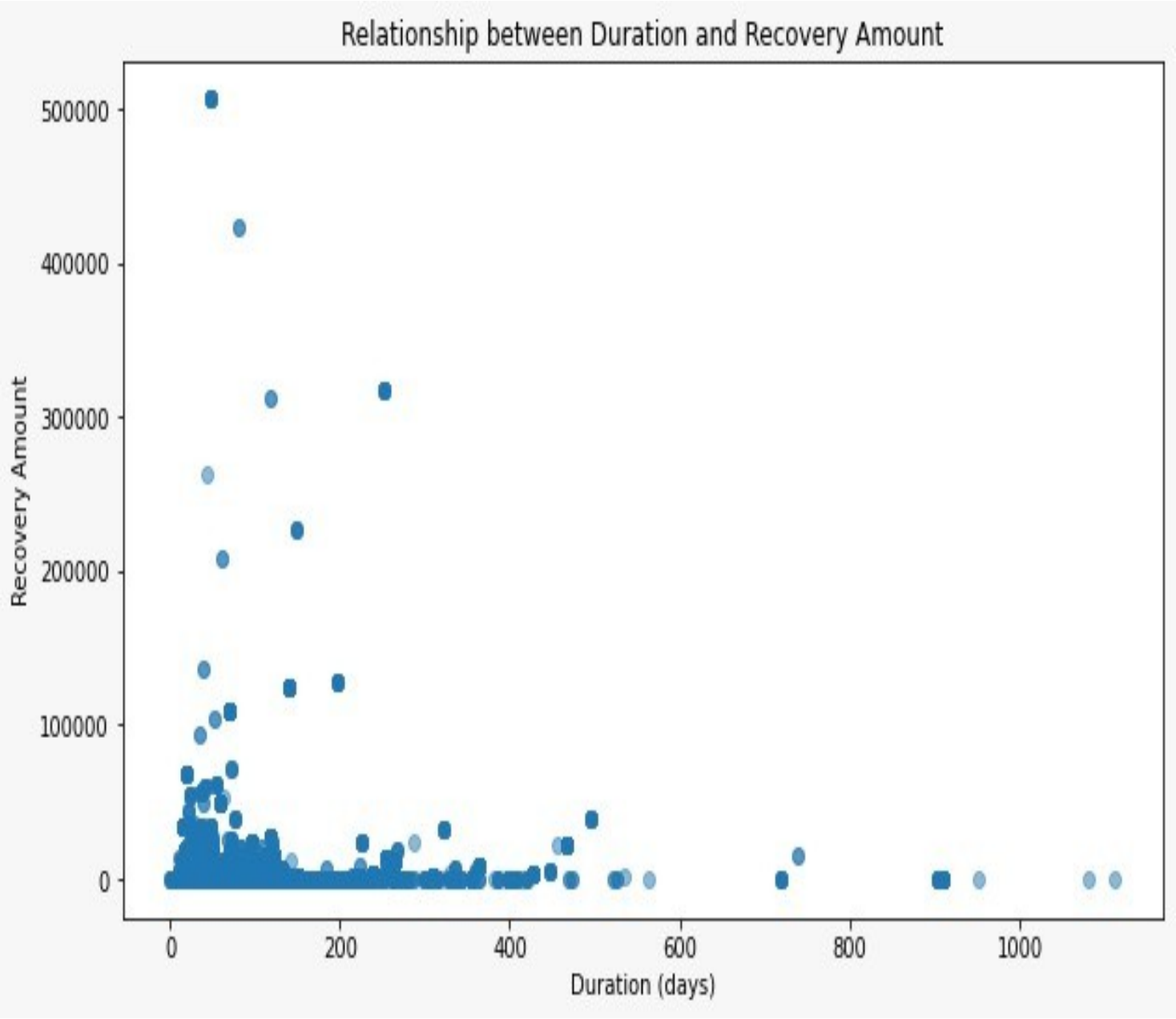


Title: Distribution of Complaint Status

Description:
This bar chart shows most complaints are categorized as "Closed," reflecting strong resolution efficiency. Minimal cases in "Reopened" and "Awaiting Decision" suggest the need for further investigation to ensure these exceptions are addressed effectively. The analysis highlights a high-resolution rate but points to areas for improving unresolved case management.

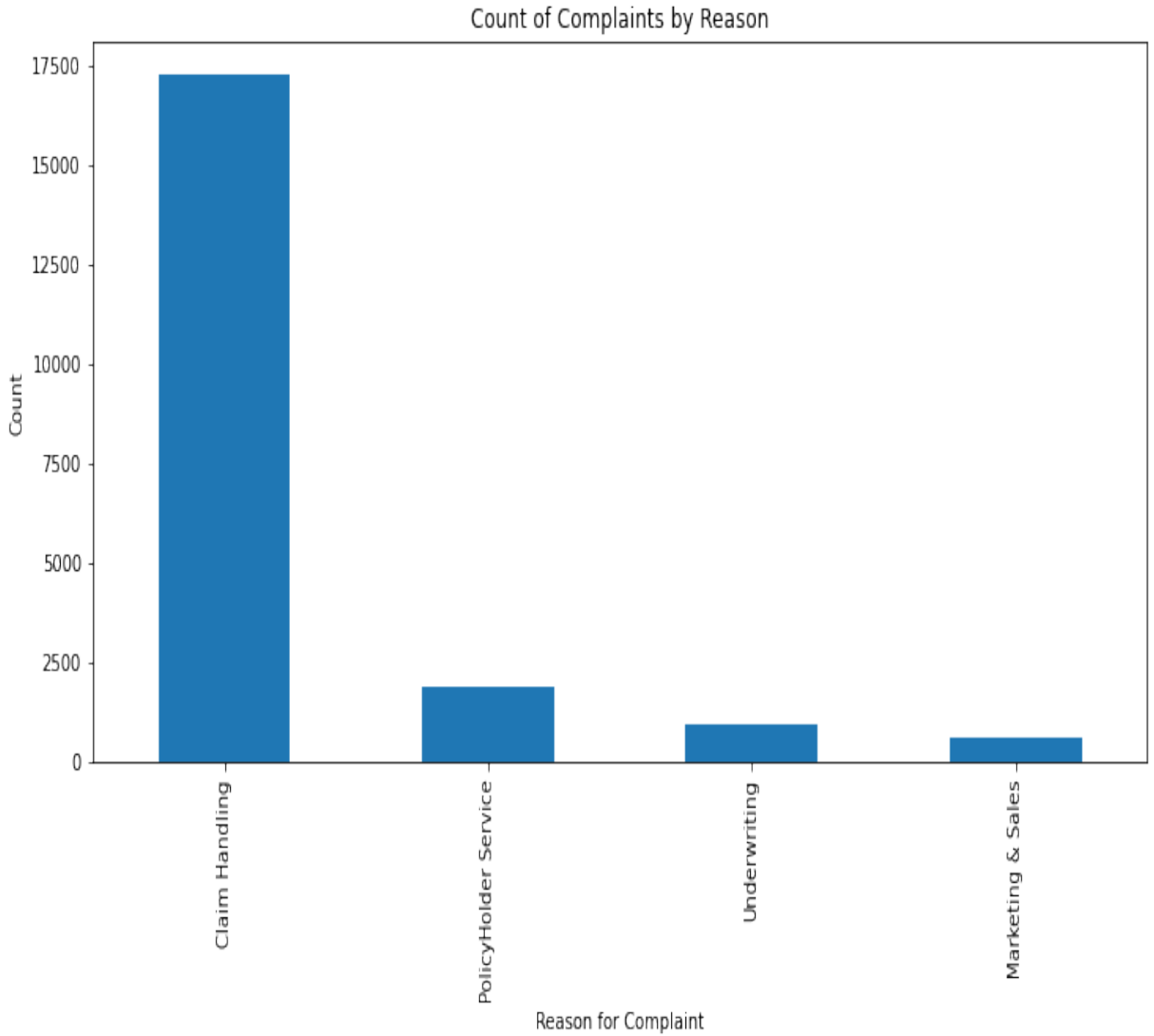


- **Title:** Average Recovery Amount by Coverage Type
- **Description:**
  This bar chart illustrates the average recovery amounts for various insurance coverage types, with confidence intervals representing variability. Categories like Life and Fire show significantly higher average recoveries, suggesting complex or high-value claims. In contrast, types like Renters/Tenants and Commercial Multi-Peril exhibit smaller, more consistent recovery amounts. Variability is evident in coverage types such as Fire, where larger intervals indicate diverse claim values.



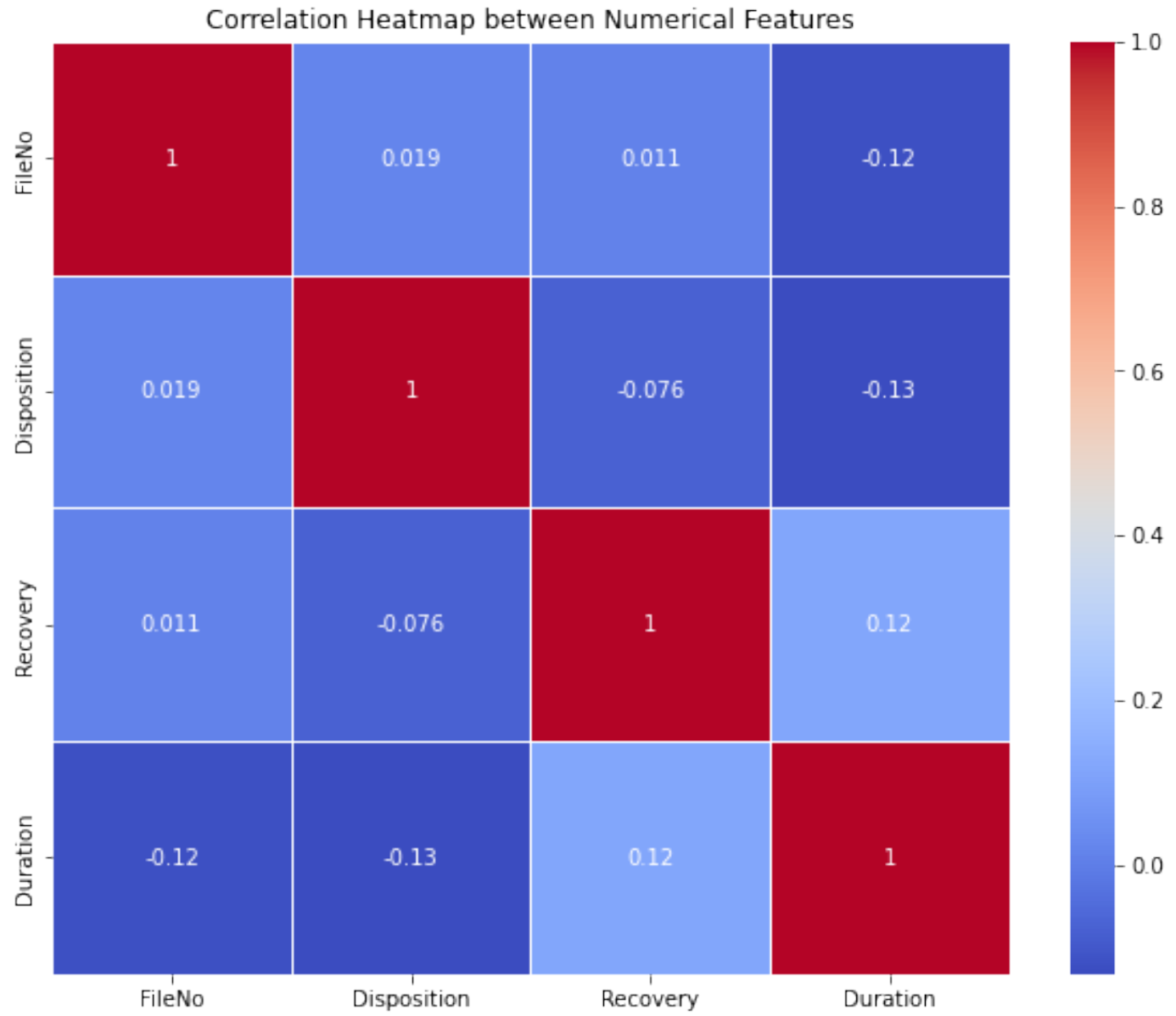
Title: Relationship Between Complaint Duration and Recovery Amount

Description:
This scatter plot shows the relationship between complaint duration and recovery amounts. No clear linear pattern emerges, as short-duration complaints exhibit both low and high recovery amounts, while longer durations show a wider recovery range. This suggests resolution time may not directly affect recovery but is worth examining for cases involving larger recoveries, often reflecting complex or contested claims.



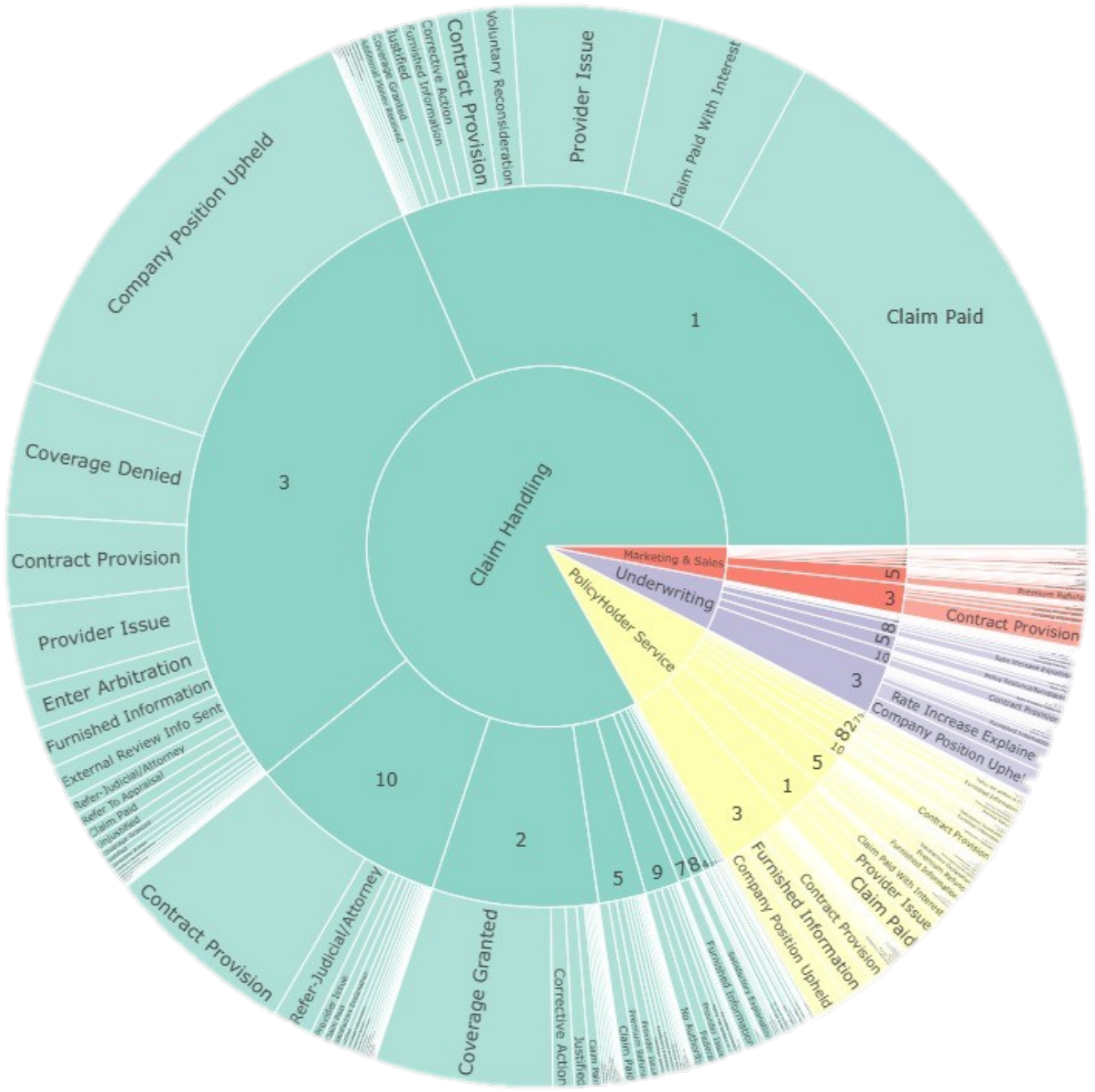
Title: Count of Complaints by Reason

Description:
This bar chart shows the frequency of complaints categorized by specific reasons. The largest bar represents the majority of complaints, while the smaller bars show significantly fewer complaints for other reasons. It provides a clear visual of the disparity between the most common and less common reasons for complaints.



Title: Heatmap of Correlation between Numerical Features

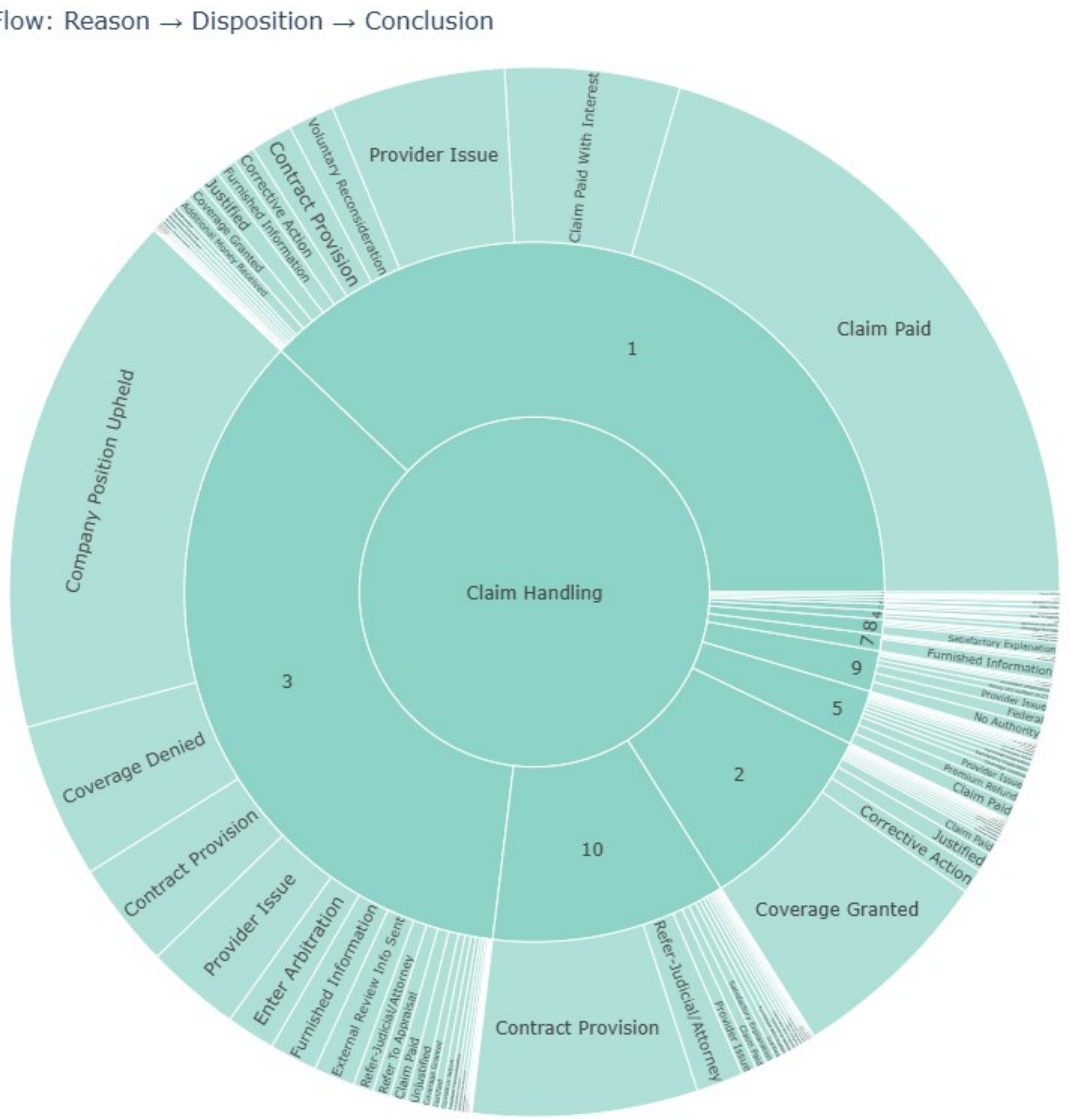
Description:
This heatmap shows the correlations between numerical features, with dark red indicating strong positive correlations and dark blue for strong negative correlations. Key insights, like the link between Resolution_Time_Days and Recovery, suggest that longer resolutions may result in higher recovery amounts. This helps guide future analysis and feature engineering.



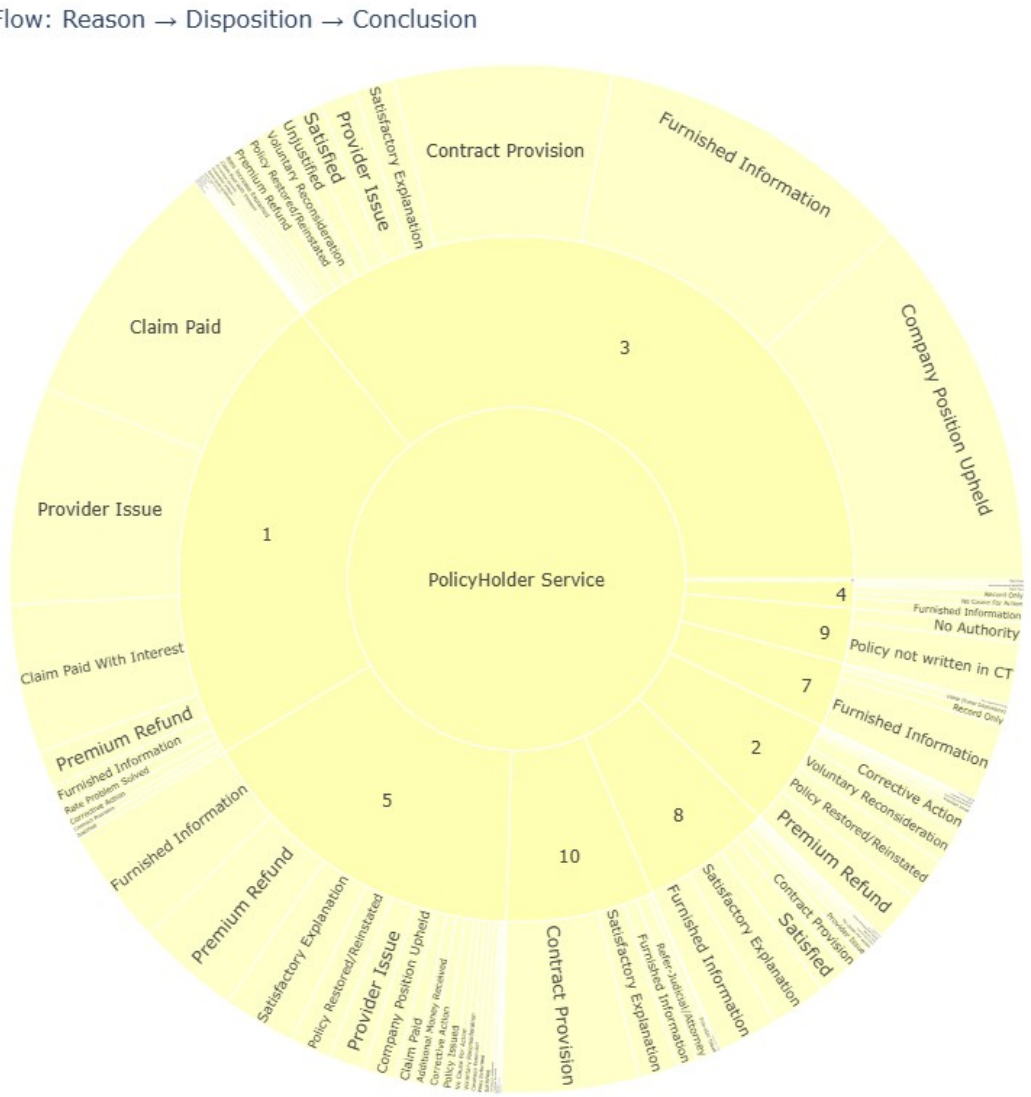
Title: Sunburst Chart: Complaint Flow Analysis

Description:
This Sunburst Chart visualizes the complaint flow from Reason to Disposition to Conclusion. The inner ring shows the primary causes, like Claim Handling or Policyholder Service. The middle ring highlights how complaints are addressed (e.g., Claim Paid, Coverage Denied). The outer ring reveals the final resolution, such as whether the issue was resolved or escalated. Key insights include Claim Handling as the most common complaint reason, with many claims resolved by "Claim Paid," though some escalate to "Coverage Denied" or further disputes. This breakdown identifies areas for improvement, focusing on the most frequent complaint paths.

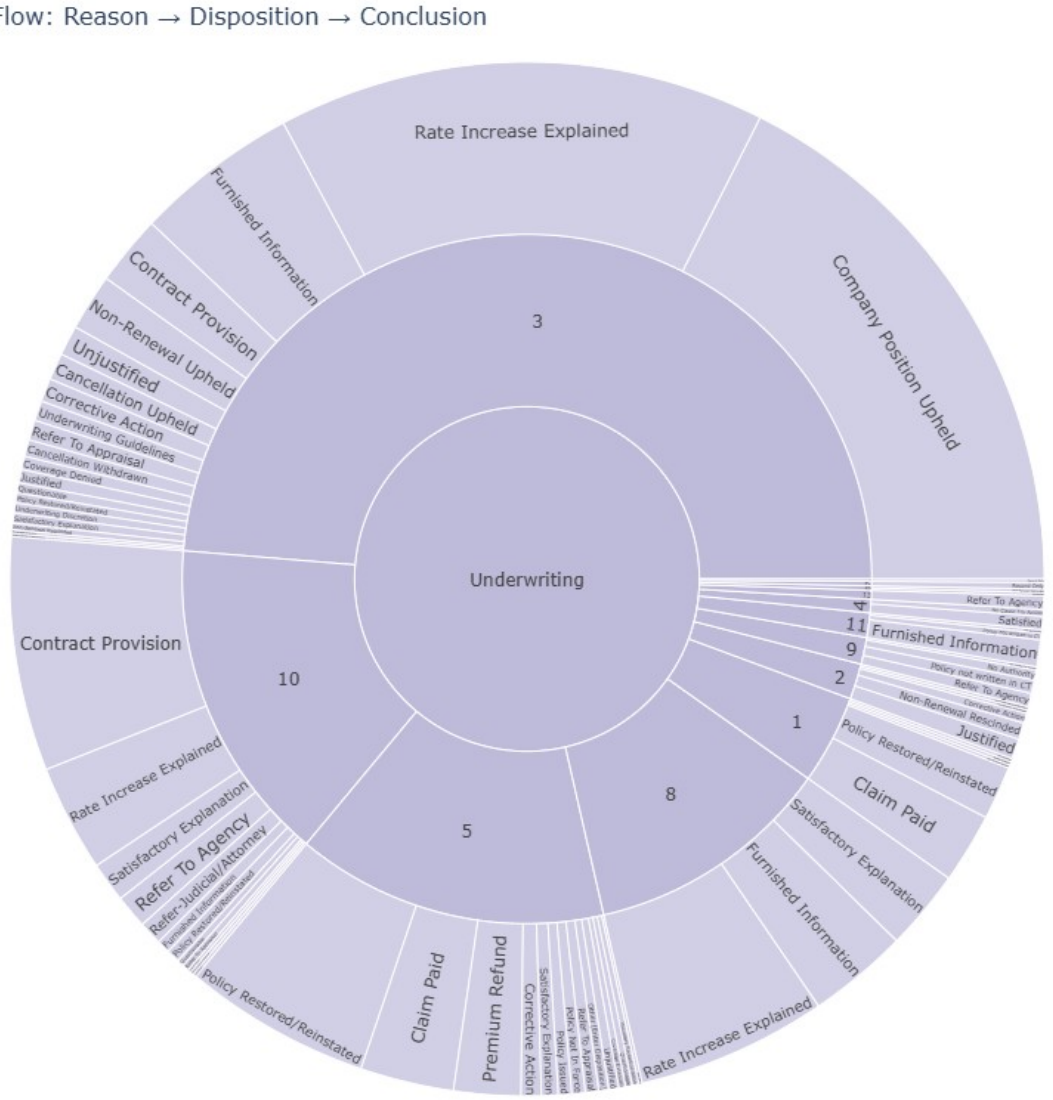
By expanding the data, we generate sunburst charts for key complaint categories such as marketing and sales, policyholder service, underwriting, and claim handling, providing detailed insights into their resolution paths.



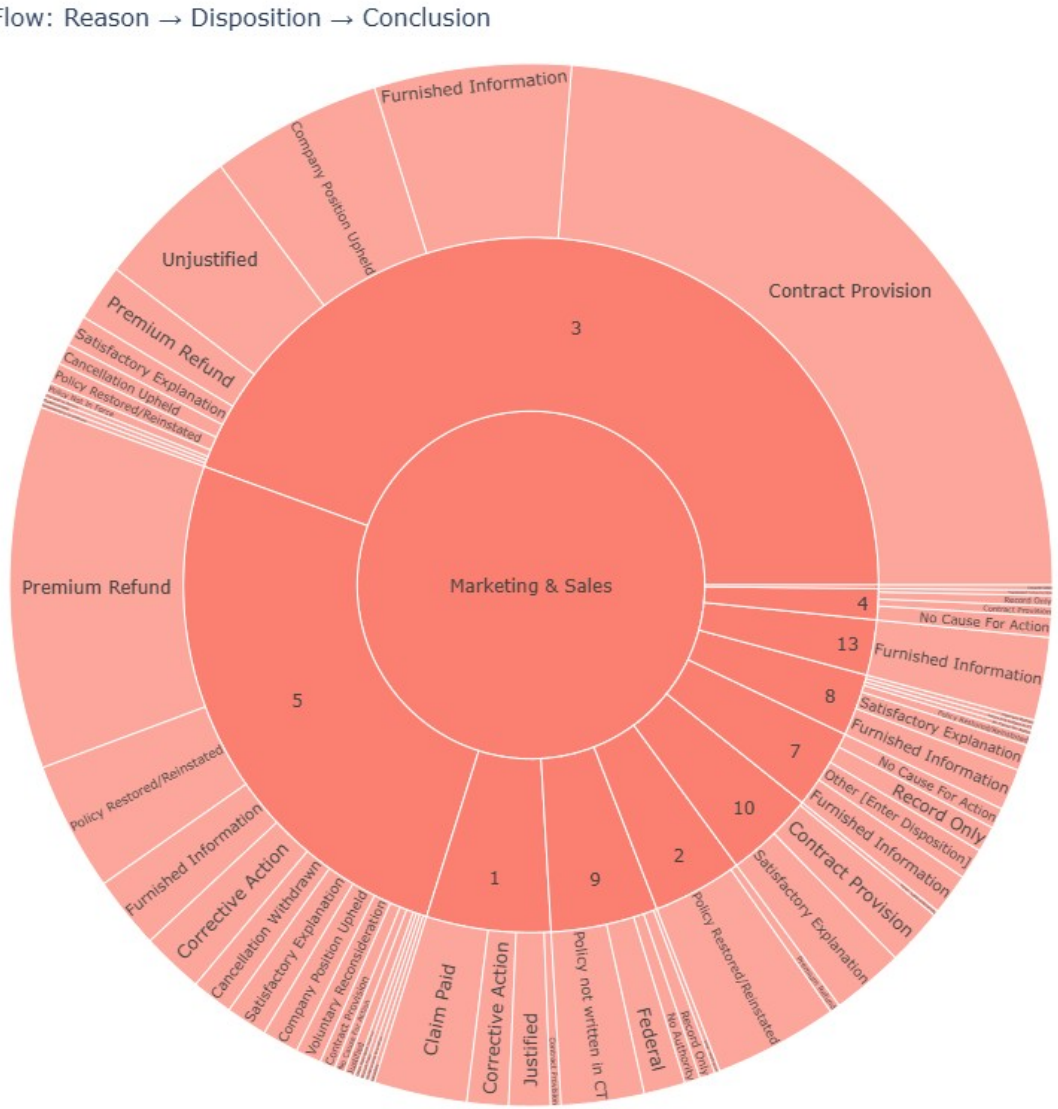
This Sunburst Chart highlights the complaint flow for 'Policyholder Service,' showing how complaints progress from reasons to dispositions and final outcomes. It provides a clear visualization of trends and resolution pathways within this category.



This Sunburst Chart illustrates the complaint flow for Underwriting, detailing how complaints transition from specific reasons to dispositions and final conclusions. It helps visualize resolution patterns and identify areas for process optimization within this category.



This Sunburst Chart showcases the complaint flow for 'Marketing & Sales,' tracking how complaints progress from initial reasons to dispositions and their final conclusions. It provides insights into resolution patterns and areas requiring process improvements in this category.

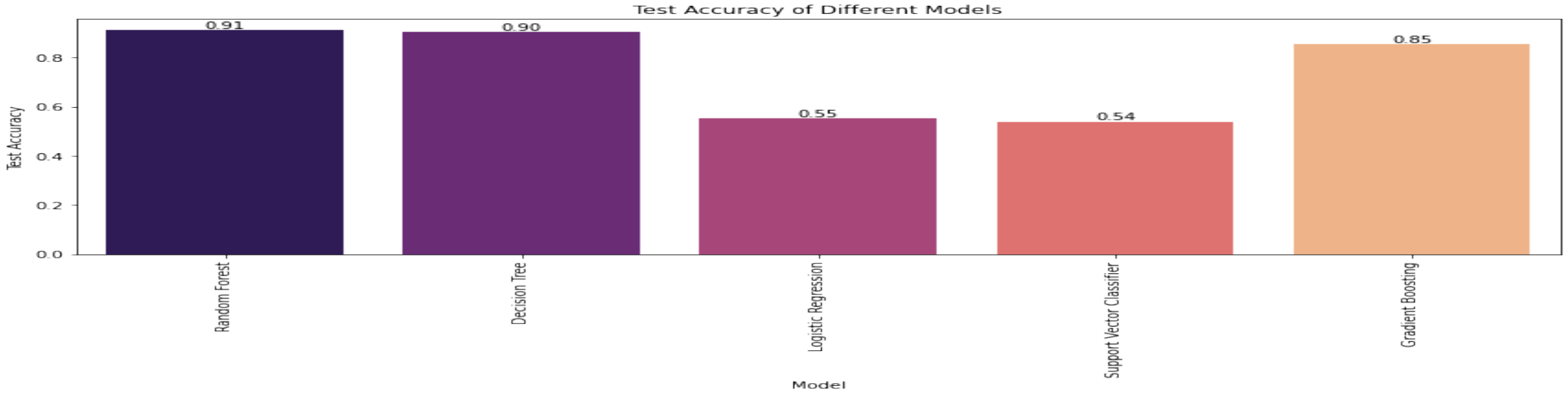


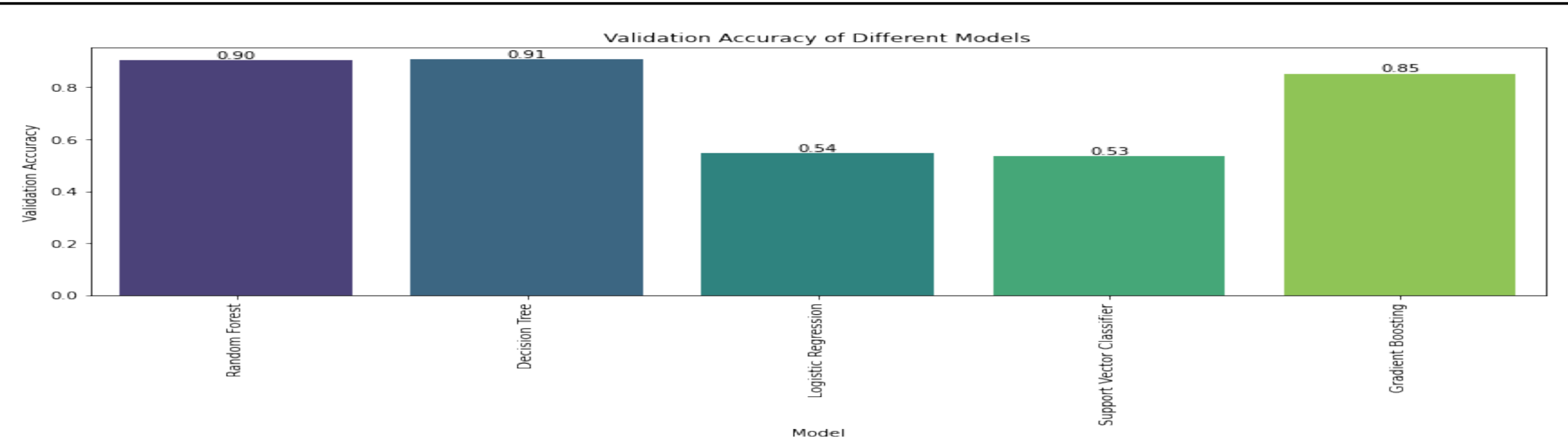
Validation and Test Accuracy of Different Models

1. **Models Compared:**
    1. Random Forest
    2. Decision Tree
    3. Logistic Regression
    4. Support Vector Classifier
    5. Gradient Boosting

2. **Insights:**
    1. **Random Forest** and **Decision Tree** achieved the highest accuracy for both validation and test datasets, demonstrating their effectiveness in handling the dataset's characteristics.
    2. **Gradient Boosting** also performed well, making it a strong alternative.
    3. Logistic Regression and Support Vector Classifier showed significantly lower accuracy, suggesting they are less suitable for this task due to potential underfitting or data complexity.

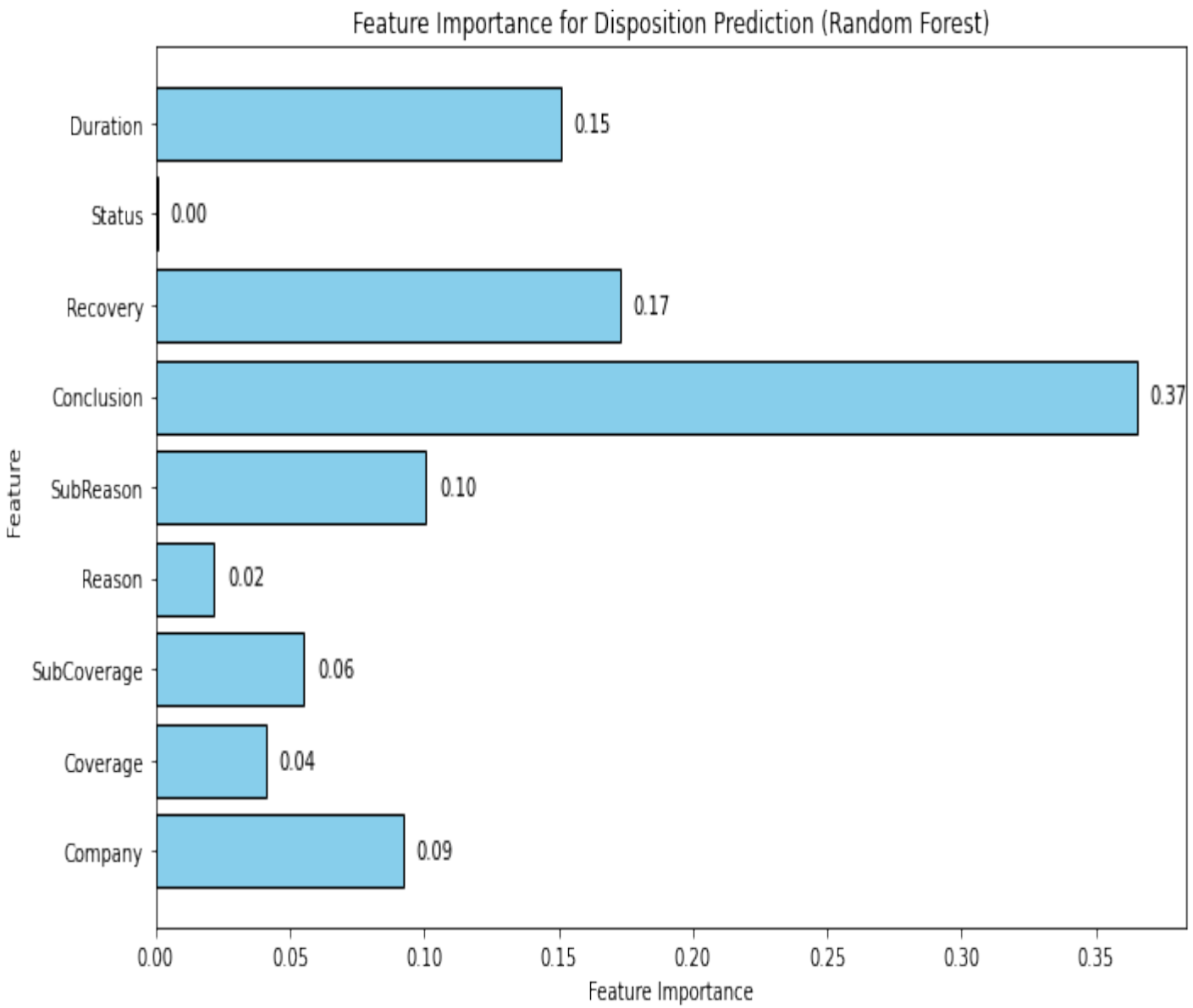
3. **Key Takeaway:** The evaluation suggests that Random Forest and Decision Tree models are the best performers, making them the most reliable choices for predictive tasks in this project.


Test Accuracy of Different Models
Test Accuracy
0.0
0.2
0.4
0.6
0.8
0.91
0.90
0.55
0.54
0.85
Random Forest
Decision Tree
Logistic Regression
Support Vector Classifier
Gradient Boosting
Model

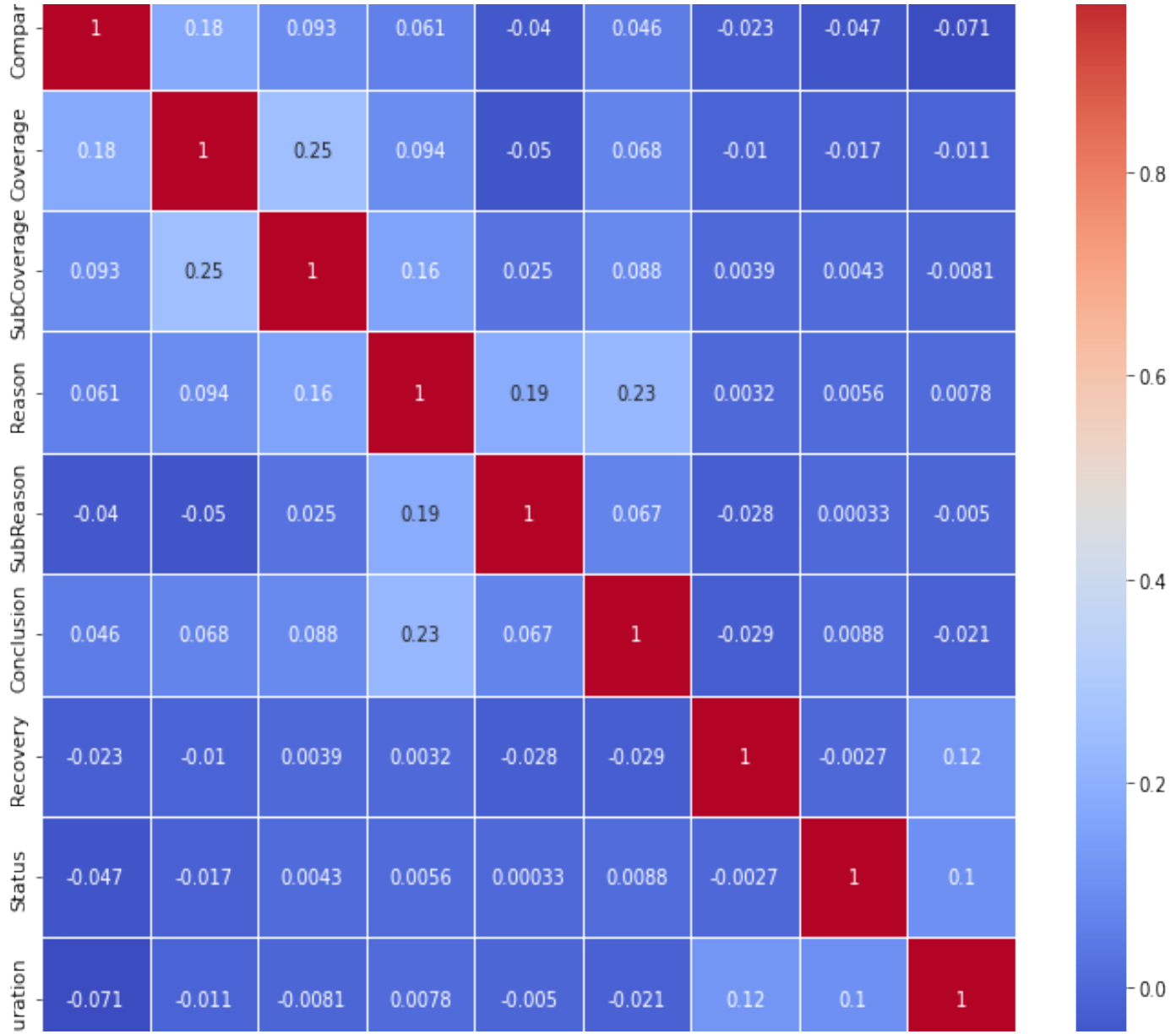

Validation Accuracy of Different Models
Validation Accuracy
0.0
0.2
0.4
0.6
0.8
0.90
0.91
0.54
0.53
0.85
Random Forest
Decision Tree
Logistic Regression
Support Vector Classifier
Gradient Boosting
Model

ACCURACY VALUES

Model	Validation Accuracy	Test Accuracy
Random Forest	90%	91%
Decision Tree	91%	90%
Gradient Boosting	85%	85%
Logistic Regression	54%	55%
Support Vector Classifier	53%	54%

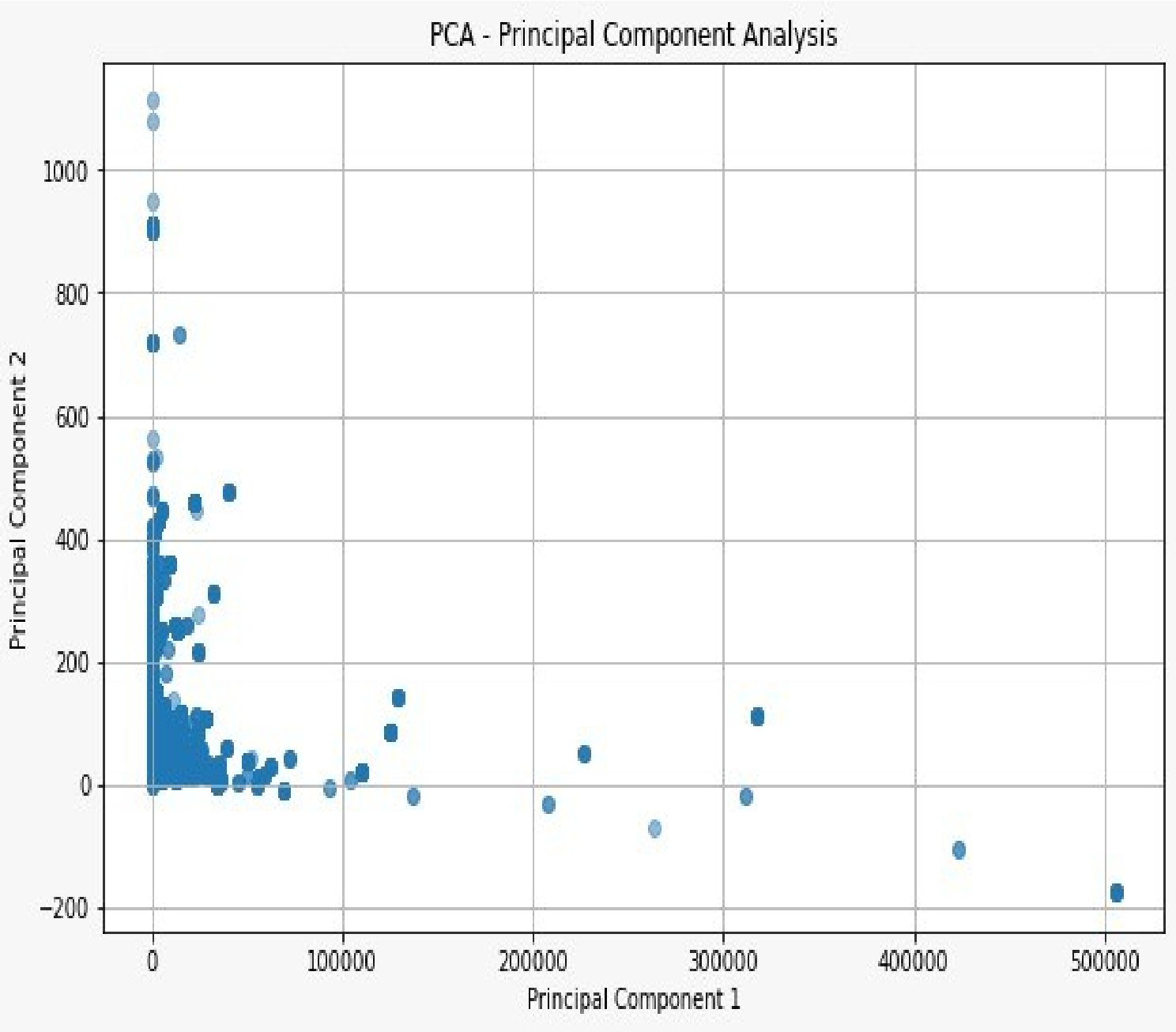


- **Title:** Feature Importance for Disposition Prediction (Random Forest)
- **Description:**
  This horizontal bar chart displays the importance of each feature in predicting the disposition of complaints using a Random Forest model. Key features like Recovery, Conclusion, and Duration may have higher importance, indicating they strongly influence the final disposition. Features such as Status may have lower importance, suggesting limited impact. Insights show that Recovery Amount is likely crucial, as larger recoveries are linked to specific dispositions, while Conclusion reflects the final outcome. Understanding feature importance helps optimize prediction models by focusing on the most critical variables for improved process efficiency.



Title: Feature Correlation Heatmap

Description:
This heatmap shows the correlation between features in the predictive model, with values ranging from -1 to +1. Positive correlations (Red) indicate that as one feature increases, the other tends to increase. Negative correlations (Blue) suggest that as one feature increases, the other decreases. Zero or near-zero correlations imply no significant relationship. Insights reveal that strongly correlated features may introduce redundancy, while weakly correlated features, like Recovery and Disposition, might offer independent contributions that enhance model performance. This visualization guides feature engineering and selection for optimizing model predictions.



- **Title:** PCA - Principal Component Analysis
- **Description:**
  This scatter plot visualizes the results of Principal Component Analysis (PCA), a technique that reduces dimensionality while preserving variability. Axes: Principal Component 1 (PC1): Captures the maximum variance. Principal Component 2 (PC2): Captures the next highest variance, orthogonal to PC1.Insights:The dense cluster near the origin suggests similar patterns in features like Duration and Recovery. Points spread along PC1 and PC2 indicate data variability, with outliers farther from the origin. Applications: PCA simplifies complex data for visualization and modeling. Helps identify patterns, clusters, and outliers. Key Takeaway: PCA reduces dimensionality while preserving important variance, aiding exploratory data analysis and machine learning model preparation.

CONCLUSION

Summary of Findings:

1. The analysis highlighted:
    1. **Key Complaint Reasons** such as claim handling and policy issues.
    2. **Trends in Resolution Times**, with the majority resolved within 30 days but some taking significantly longer.
    3. **Predictive Modeling Insights**, demonstrating the capability of models to predict outcomes with over 90% accuracy.

Future Scope:

2. **Integrate External Datasets:** Combine customer feedback, market data, or operational metrics for more robust predictive models.
3. **Implement Predictive Systems:** Utilize predictive tools in customer service workflows to proactively resolve complaints and enhance customer satisfaction.

In the end, we wanted to utilize machine learning models to predict complaint outcomes and enable better resource allocation and faster resolution times, reducing operational bottlenecks and associated costs.

ACKNOWLEDGEMENT AND REFERENCES

- We extend our gratitude to everyone who contributed to the success of this project. Special thanks to the data providers for the insurance complaints dataset, which formed the foundation of our analysis. We are deeply grateful to our mentors for their invaluable guidance and to our team members for their collaboration and expertise. Finally, we express our sincere appreciation to end-users whose feedback played a vital role in shaping the system's development and practical application.

- Data.gov. (2024, October 19). State of Connecticut - Insurance Company complaints, resolutions, status, and recoveries. https://catalog.data.gov/dataset/insurance-company-complaints-resolutions-status-and-recoveries
- Ross, H. L. (1975). Insurance Claims Complaints: a private appeals procedure. Law & Society Review, 9(2), 275. https://doi.org/10.2307/3052977
- Alamir, E., Urgessa, T., Hunegnaw, A., & Gopikrishna, T. (2021). Motor Insurance Claim Status Prediction using Machine Learning Techniques. International Journal of Advanced Computer Science and Applications, 12(3). https://doi.org/10.14569/ijacsa.2021.0120354