HEALTH INSURANCE CLAIM ANALYSIS

ABSTRACT

Health insurance claim analysis is a critical aspect of modern healthcare and insurance systems, as it enables organizations to improve operational workflows, identify trends, and prevent fraudulent activities. This project undertakes a comprehensive exploration of health insurance claim data to uncover patterns and generate actionable insights. The dataset includes essential variables such as patient demographics age, gender, claim amount, claim type, provider specialty, submission method online, paper, phone, and final claim status approved or denied. Detailed analyses are conducted to examine claim volume by gender and age group, monthly trends in claim submissions, and claim approval rates across different submission methods. Special attention is given to how the mode of submission affects approval outcomes, highlighting the efficiency of digital channels over traditional ones. Claims are also evaluated based on provider specialties and top geographical locations, providing a macro-level understanding of high-activity regions and departments. This helps insurers identify bottlenecks, allocate resources more efficiently, and design customer-centric processes.

The insurance claim analysis segments claim amounts into various monetary ranges to identify high-frequency claim zones and financial outliers. Outliers, such as unusually high or suspiciously low claim amounts, are flagged for further investigation and fraud risk assessment. Advanced cross-analysis explores how combinations of factors such as age group, claim type, and provider affect the final outcome. Patterns emerging from these combinations reveal how certain demographic groups or medical specialties are linked to specific claim behaviors. Insights into the timing of claim submissions, such as peak claim months or seasonal trends, provide insurers with better planning and resource allocation strategies. Visual tools like bar graphs, pie charts, and heatmaps are employed to enhance interpretability and provide quick insights. These visualizations help in comparing approval rates by claim type, monitoring provider performance, and identifying gaps in service delivery. The study also evaluates the consistency of claim decisions across submission methods to identify process inefficiencies. By identifying such trends, insurers can standardize procedures and ensure fairness in claim processing. The findings from this project enable insurance companies to refine their policy frameworks, reduce claim processing times, and improve customer satisfaction. Furthermore, the insights can guide data-driven decision-making and support the creation of predictive models for risk management. Ultimately, this study contributes to building a smarter, faster, and more reliable health insurance claim process.

1. INTRODUCTION

Health insurance claim analysis plays a vital role in the decision-making process of insurance companies and healthcare service providers. In today's data-driven world, organizations in the health insurance sector rely heavily on accurate claim data to evaluate trends, improve service efficiency, detect inconsistencies, and make informed strategic decisions. This project focuses on analyzing historical health insurance claim data to identify significant patterns, behaviors, and insights that can enhance the overall claim management process and customer experience. The dataset used in this project contains detailed information on various attributes, such as Claim ID, Patient Demographics, Claim Amount, Diagnosis and Procedure Codes, Claim Status, and Submission Method. By studying these variables in combination, we aim to understand how different factors influence claim outcomes such as approval, denial, and pending status, as well as identify areas where delays or discrepancies occur most frequently.

The analysis begins with a structured data cleaning and preprocessing phase to ensure accuracy and reliability. Following this, exploratory data analysis (EDA) techniques are applied to uncover hidden patterns and trends. Through the use of visual tools like bar graphs, pie charts, line plots, and histograms, the project highlights key findings such as the age groups filing the most claims, average claim amounts by provider specialty, top reasons for claim rejection, and preferred claim submission methods. Special attention is given to demographic and regional variations, allowing us to identify whether certain patient segments or provider types are more likely to face claim issues or benefit from faster processing.

A key aspect of the analysis involves examining the distribution of claims by gender. This provides insight into which gender submits more claims and whether there is a difference in claim outcomes such as approvals or denials between male and female patients. Closely related are the claim trends by age group, which reveals which age ranges are most active in filing claims and how these age groups differ in terms of claim acceptance, rejections, or delays. This demographic level analysis supports more equitable policy planning and targeted service improvement.

The monthly trend of claims by gender is analyzed to identify seasonal patterns in claim submissions and determine whether there are specific times of the year where claims spike or dip. These findings are valuable for both staffing and operational planning in insurance firms.

One of the central parts of the project is the detailed exploration of submission methods, specifically comparing manual and electronic submissions. The study evaluates how these methods influence claim outcomes in terms of approval speed and accuracy, and highlights that electronic submission tends to result in quicker and more favorable processing.

To understand submission efficiency, the analysis includes a focused review of approved claims by submission method and the overall claim status (approved, denied, pending) across submission types. This reveals which method leads to better claim handling and helps identify gaps in manual processing workflows. The results support the growing shift toward digitization in the health insurance sector. Further, the study examines the distribution of claim status by age group and claim type, showing how different age segments and treatment categories experience varying outcomes. This dual-layered analysis helps stakeholders identify vulnerable groups and optimize policy coverage and communication strategies. The data also provides insight into provider specialties with the highest claim volumes, indicating which healthcare domains have the most billing interactions and might benefit from operational improvements or enhanced billing training.

A key focus of the analysis is the comparison of extreme claim amounts both highest and lowest alongside associated variables such as diagnosis codes, patient age, and healthcare provider information. Examining these outliers can uncover unusual medical events, potential fraudulent activity, or gaps in claim justification. Moreover, analyzing the distribution of claim counts across different amount ranges offers valuable insights into how frequently small, moderate, and high-value claims occur. These patterns are essential for shaping effective insurance pricing strategies, setting appropriate deductible thresholds, and establishing fair reimbursement limits.

Geographical variation is addressed through the analysis of provider locations with the maximum health claims. It also supports the identification of region-specific trends or disparities in healthcare access and insurance utilization. Health insurance claim analyze data to uncover key patterns affecting claim outcomes. By examining variables such as age, gender, claim amount, submission method, and provider details, it provides visual insights that support faster approvals, fraud detection, and operational improvements. Through simplified data visualization, the study emphasizes the importance of data-driven decision-making in building a more efficient, transparent, and customer-oriented health insurance system.

1.1 PROBLEM STATEMENT

Health insurance providers often encounter numerous challenges in managing claims due to the complexity and volume of data involved. Despite having access to detailed historical claim records, many insurers struggle to effectively utilize this data to improve claim processing efficiency, reduce delays, and enhance decision-making. Variations in patient demographics, claim amounts, submission methods, and diagnosis categories make it difficult to identify consistent patterns that influence claim outcomes such as approvals, denials, and pending statuses. These inefficiencies can lead to customer dissatisfaction, increased operational costs, and exposure to fraudulent activities.

Health insurance claim analysis focuses on analyzing comprehensive health insurance claim data to uncover meaningful insights that can drive operational improvements and strategic planning. By exploring trends in patient age, gender, provider specialties, claim submission methods, and geographical distributions, the study aims to highlight key factors affecting claim performance. Additionally, it investigates outliers in claim amounts, compares manual vs. electronic submissions, and evaluates status outcomes across different segments. The goal is to support health insurers in building a more transparent, data-driven, and customer centric claims ecosystem.

- Health insurers lack a clear understanding of which demographic or procedural factors lead to claim approval, denial, or delay.
- Manual submission methods contribute to slower claim processing compared to online submissions.
- Outliers in claim amounts are often not investigated thoroughly, potentially masking fraud or unusual medical events.
- Providers and regions with frequent claims are not always given targeted improvements or operational attention.
- Claim outcome trends by gender and age are not well understood, affecting equitable policy planning.
- Monthly and seasonal fluctuations in claim volume are not effectively used for staffing and operational forecasting.
- Differences in outcomes across provider specialties are often overlooked in training workflow.

1.2 TOOL DESCRIPTION- JUPYTER NOTEBOOK

Jupyter Notebook is an open-source, web-based interactive computing environment that allows users to create and share documents containing live code, equations, visualizations, and narrative text. It supports various programming languages, with Python being the most commonly used in data science. In this project, Jupyter Notebook served as the primary development platform due to its ability to combine code execution, data analysis, and visualization in a single, organized document. It helped streamline the workflow by enabling step-by-step execution, immediate output display, and easy debugging. The notebook format also allowed combining charts, tables, and explanatory notes in a readable and organized way, which helped in better understanding the data and presenting the findings clearly.

PACKAGES USED

To perform the analysis on health insurance claims, the following Python packages were used:

- **Pandas:** Used for loading and managing the dataset. It helped in cleaning the data, handling missing values, filtering rows, grouping data, and performing operations like sorting and summarizing claim records.
- **Numpy:** Provided support for mathematical operations and numerical processing, which were essential for calculating statistics and transforming values also assisted in array-based data transformations and performance optimizations during preprocessing.
- **Matplotlib:** A basic plotting library used to create bar charts, pie charts, and line graphs for visualizing trends and comparisons in the claim data.
- **Seaborn:** Built on top of matplotlib, it made it easier to create attractive and informative visualizations like heatmaps, boxplots, and category-wise comparisons. Enhanced the visual appeal and clarity of complex plots like boxplots, heatmaps, and charts. It was particularly useful in showcasing claim severity by provider type and employment status.
- **Ploty.express and Ploty.graph objects:** These libraries were used for creating interactive visualizations and it helped display dynamic charts such as lollipop plots, horizontal bars, and grouped comparisons that enhanced the user experience. These tools enriched user engagement by allowing on-hover insights and filter-based visual breakdown.

2. DOMAIN-HEALTH INSURANCE AND CLAIM ANALYTICS

The domain of this project is Health Insurance and Claim Analytics, which plays a crucial role in the healthcare and insurance sectors. As the volume of health insurance claims continues to grow, insurance providers face increasing pressure to streamline claim processing, reduce administrative overhead, and provide better service to policyholders. Analyzing claim data can offer critical insights into patterns, behaviors, and inefficiencies within the system.

Health insurance claim analysis focuses on analyzing health insurance claims using demographic, financial, and procedural data. Attributes such as Claim Amount, Gender, Age, Location, Provider Specialty, Claim Type, and Submission Method are examined to uncover meaningful trends. Through this data exploration, insurance companies can better understand how different factors affect claim frequency, amount, and approval status.

The analysis leverages exploratory data analysis (EDA), summary statistics, and data visualization tools to extract insights without applying machine learning techniques. For instance, visualizing claim distribution by gender or age groups can help identify demographic segments that file more claims, while submission methods can reveal process efficiencies or delays.

The importance of this domain lies in its real-world operational and business impact. by identifying patterns in claim submissions, health insurance companies can improve operational efficiency and streamline their processes. Understanding the demographic needs of policyholders helps enhance customer service and satisfaction. Additionally, analyzing the data allows insurers to allocate resources more effectively by identifying high-volume specialties or geographic regions. These insights support data-driven decision-making for policy design and claim handling strategies. Overall, this study empowers health insurance providers to shift from traditional claim processing methods to more insight-driven, strategic approaches and ultimately boosting both profitability and customer experience.

The health insurance claim analysis reveals critical patterns in claim behaviour across age, gender, and region. It helps identify underserved groups and optimize policy offerings accordingly. Frequent and high-cost claim types guide resource and coverage planning. Submission method insights promote faster, tech-driven claim processing. Approval and rejection trends highlight areas for policy and process improvement and collectively the findings enable smarter, more responsive health insurance strategies.

2.1 CASE STUDY – HEALTH INSURANCE CLAIMS ANALYSIS

A major health insurance company faced increasing operational costs and inefficiencies in its claims processing system. With thousands of claims being submitted each month from various regions, specialties, and patient demographics, the company struggled to gain clarity on the root causes of delays, approval disparities, and rising claim costs. Additionally, there was limited understanding of which patient segments or provider types were driving higher volumes and costs. To address these challenges, the company turned to data analytics to analyze historical health insurance claim data. Using Python and relevant data visualization techniques, analysts examined variables such as claim amount, diagnosis and procedure codes, provider specialties, submission methods, and demographic details. The insights revealed trends in patient behaviour, claim statuses, and regional claim concentrations, helping the company to streamline operations, optimize resource allocation, and improve customer service strategies.

OBJECTIVE

The main objective of this project is to analyze health insurance claim data using Python to uncover key trends and inefficiencies through data visualization.

- Identify the most frequent diagnosis and procedure codes contributing to high claim amounts.
- Analyze historical claim data to uncover trends and patterns related to claim volume, status, and associated costs.
- Examine the impact of patient demographics and provider specialties on claim approval and rejection rates.
- Enable the insurance firm to detect and address operational delays in claims processing workflows.
- Support strategic planning by visualizing regional claim distributions and submission method effectiveness.
- Provide actionable insights to improve policy design, resource allocation, and customer service strategies.

This analysis helps the company make better decisions using data, so the claim process becomes quicker, more correct, and easier for customer.

3. DATA MODELLING

The health insurance claim records dataset contains claim amount, patient demographics, diagnosis type, and provider information. These features are used to analyze trends, detect outliers, and segment claims by amount, frequency, and risk category. The system visualizes key insights such as extreme claim distributions, common medical categories, and demographic patterns, enabling insurers to improve fraud detection, pricing strategies, and resource allocation.

3.1 PROCESS FLOW

The process flow diagram shows how raw health insurance claim data is transformed into insights through data collection, cleaning, encoding, EDA, and visualizations based on age, gender, location, and submission method, leading to informed decision-making.

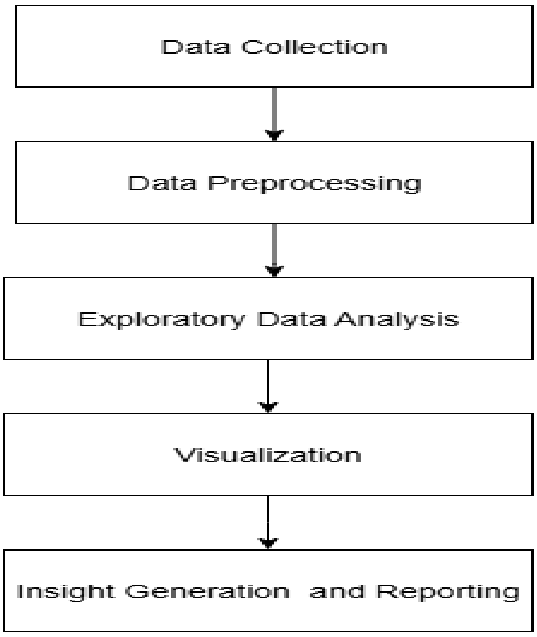


Figure 3.1-Process flow

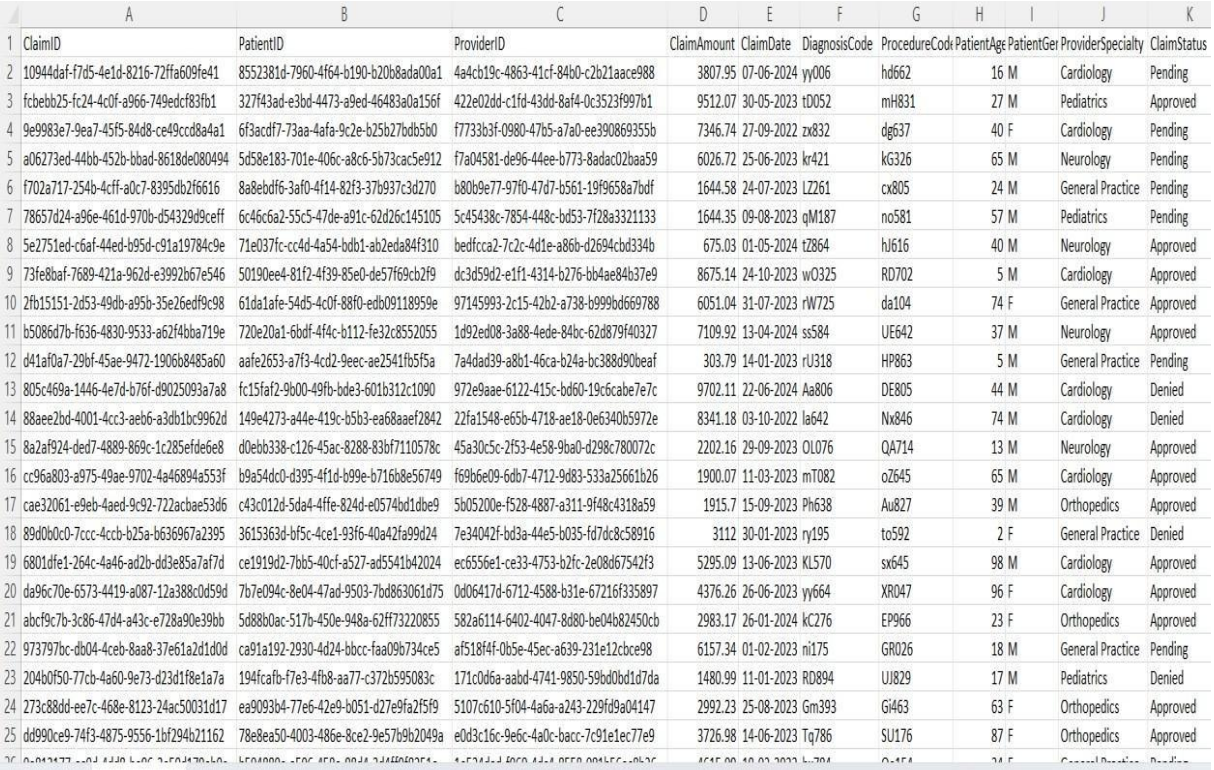
Figure 3.1 represents the processflow begins with data collection and preprocessing to clean and prepare the health insurance claim dataset. It then proceeds through exploratory data analysis to uncover patterns in claim amounts, submission methods, patient demographics, and provider specialties. The process concludes with the generation of meaningful business insights to support better decision-making in health insurance operations.

3.2 DATA COLLECTION

The dataset was sourced from a health insurance provider’s internal claim management system. It contains 4,501 claim records selected to study claim behaviour and approval patterns. key attributes include patient demographics, provider details, diagnosis and procedure codes. Additional fields cover claim status, type, amount, and submission methods. The data was extracted in format and cleaned to ensure accuracy and consistency. Personally identifiable information was anonymized to maintain data privacy and security.

3.3 SAMPLE DATASET

The dataset includes essential attributes like ClaimID, Patient, ProviderID, Claim Amount, Claim Date, Diagnosis Code, and Procedure Code, enabling detailed analysis of claim patterns. Demographic fields such as Patient Age, Gender, Income, Marital Status, and Employment Status help explore how personal factors influence claims. Provider-related details like Specialty and Location reveal trends across regions and services. Claim Status and Claim Type offer insights into claim outcomes, while Claim Submission Method highlights processing efficiency. Overall, the dataset supports comprehensive analysis of health insurance claims to uncover patterns, trends, and operational improvements.



	A	B	C	D	E	F	G	H	I	J	K
1	ClaimID	PatientID	ProviderID	ClaimAmount	ClaimDate	DiagnosisCode	ProcedureCod	PatientAg	PatientGer	ProviderSpecialty	ClaimStatus
2	10944daf-f7d5-4e1d-8216-72ffa609fe41	8552381d-7960-4f64-b190-b20b8ada00a1	4a4cb19c-4863-41cf-84b0-c2b21aace988	3807.95	07-06-2024	yy006	hd662	16	M	Cardiology	Pending
3	fcbebb25-fc24-4c0f-a966-749edcf83fb1	327f43ad-e3bd-4473-a9ed-46483a0a156f	422e02dd-c1fd-43dd-8af4-0c3523f997b1	9512.07	30-05-2023	tD052	mH831	27	M	Pediatrics	Approved
4	9e9983e7-9ea7-45f5-84d8-ce49ccd8a4a1	6f3acdf7-73aa-4afa-9c2e-b25b27bdb5b0	f7733b3f-0980-47b5-a7a0-ee390869355b	7346.74	27-09-2022	zx832	dg637	40	F	Cardiology	Pending
5	a06273ed-44bb-452b-bbad-8618de080494	5d58e183-701e-406c-a8c6-5b73cac5e912	f7a04581-de96-44ee-b773-8adac02baa59	6026.72	25-06-2023	kr421	kG326	65	M	Neurology	Pending
6	f702a717-254b-4cff-a0c7-8395db2f6616	8a8ebdf6-3af0-4f14-82f3-37b937c3d270	b80b9e77-97f0-47d7-b561-19f9658a7bdf	1644.58	24-07-2023	LZ261	cx805	24	M	General Practice	Pending
7	78657d24-a96e-461d-970b-d54329d9ceff	6c46c6a2-55c5-47de-a91c-62d26c145105	5c45438c-7854-448c-bd53-7f28a3321133	1644.35	09-08-2023	qM187	no581	57	M	Pediatrics	Pending
8	5e2751ed-c6af-44ed-b95d-c91a19784c9e	71e037fc-cc4d-4a54-bdb1-ab2eda84f310	bedfcca2-7c2c-4d1e-a86b-d2694cbd334b	675.03	01-05-2024	tZ864	hJ616	40	M	Neurology	Approved
9	73fe8baf-7689-421a-962d-e3992b67e546	50190ee4-81f2-4f39-85e0-de57f69cb2f9	dc3d59d2-e1f1-4314-b276-bb4ae84b37e9	8675.14	24-10-2023	wO325	RD702	5	M	Cardiology	Approved
10	2fb15151-2d53-49db-a95b-35e26edf9c98	61da1afe-54d5-4c0f-88f0-edb09118959e	97145993-2c15-42b2-a738-b999bd669788	6051.04	31-07-2023	rW725	da104	74	F	General Practice	Approved
11	b5086d7b-f636-4830-9533-a62f4bba719e	720e20a1-6bdf-4f4c-b112-fe32c8552055	1d92ed08-3a88-4ede-84bc-62d879f40327	7109.92	13-04-2024	ss584	UE642	37	M	Neurology	Approved
12	d41af0a7-29bf-45ae-9472-1906b8485a60	aafe2653-a7f3-4cd2-9eec-ae2541fb5f5a	7a4dad39-a8b1-46ca-b24a-bc388d90beaf	303.79	14-01-2023	rU318	HP863	5	M	General Practice	Pending
13	805c469a-1446-4e7d-b76f-d9025093a7a8	fc15faf2-9b00-49fb-bde3-601b312c1090	972e9aae-6122-415c-bd60-19c6cabe7e7c	9702.11	22-06-2024	Aa806	DE805	44	M	Cardiology	Denied
14	88aee2bd-4001-4cc3-aeb6-a3db1bc9962d	149e4273-a44e-419c-b5b3-ea68aaef2842	22fa1548-e65b-4718-ae18-0e6340b5972e	8341.18	03-10-2022	la642	Nx846	74	M	Cardiology	Denied
15	8a2af924-ded7-4889-869c-1c285efde6e8	d0ebb338-c126-45ac-8288-83bf7110578c	45a30c5c-2f53-4e58-9ba0-d298c780072c	2202.16	29-09-2023	OL076	QA714	13	M	Neurology	Approved
16	cc96a803-a975-49ae-9702-4a46894a553f	b9a54dc0-d395-4f1d-b99e-b716b8e56749	f69b6e09-6db7-4712-9d83-533a25661b26	1900.07	11-03-2023	mT082	oZ645	65	M	Cardiology	Approved
17	cae32061-e9eb-4aed-9c92-722acbae53d6	c43c012d-5da4-4ffe-824d-e0574bd1dbe9	5b05200e-f528-4887-a311-9f48c4318a59	1915.7	15-09-2023	Ph638	Au827	39	M	Orthopedics	Approved
18	89d0b0c0-7ccc-4ccb-b25a-b636967a2395	3615363d-bf5c-4ce1-93f6-40a42fa99d24	7e34042f-bd3a-44e5-b035-fd7dc8c58916	3112	30-01-2023	ry195	to592	2	F	General Practice	Denied
19	6801dfe1-264c-4a46-ad2b-dd3e85a7af7d	ce1919d2-7bb5-40cf-a527-ad5541b42024	ec6556e1-ce33-4753-b2fc-2e08d67542f3	5295.09	13-06-2023	KL570	sx645	98	M	Cardiology	Approved
20	da96c70e-6573-4419-a087-12a388c0d59d	7b7e094c-8e04-47ad-9503-7bd863061d75	0d06417d-6712-4588-b31e-67216f335897	4376.26	26-06-2023	yy664	XR047	96	F	Cardiology	Approved
21	abcf9c7b-3c86-47d4-a43c-e728a90e39bb	5d88b0ac-517b-450e-948a-62ff73220855	582a6114-6402-4047-8d80-be04b82450cb	2983.17	26-01-2024	kC276	EP966	23	F	Orthopedics	Approved
22	973797bc-db04-4ceb-8aa8-37e61a2d1d0d	ca91a192-2930-4d24-bbcc-faa09b734ce5	af518f4f-0b5e-45ec-a639-231e12cbce98	6157.34	01-02-2023	ni175	GR026	18	M	General Practice	Pending
23	204b0f50-77cb-4a60-9e73-d23d1f8e1a7a	194fcafb-f7e3-4fb8-aa77-c372b595083c	171c0d6a-aabd-4741-9850-59bd0bd1d7da	1480.99	11-01-2023	RD894	UJ829	17	M	Pediatrics	Denied
24	273c88dd-ee7c-468e-8123-24ac50031d17	ea9093b4-77e6-42e9-b051-d27e9fa2f5f9	5107c610-5f04-4a6a-a243-229fd9a04147	2992.23	25-08-2023	Gm393	Gi463	63	F	Orthopedics	Approved
25	dd990ce9-74f3-4875-9556-1bf294b21162	78e8ea50-4003-486e-8ce2-9e57b9b2049a	e0d3c16c-9e6c-4a0c-bacc-7c91e1ec77e9	3726.98	14-06-2023	Tq786	SU176	87	F	Orthopedics	Approved
26	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	General Practice	Pending

Figure 3.3 - Sample Dataset

Figure 3.3 represents the sample dataset displays a sample view of the health insurance claim dataset used for analysis. It includes various attributes such as ClaimID, PatientID, ProviderID, ClaimAmount, ClaimDate, DiagnosisCode, ProcedureCode, PatientAge, PatientGender, ProviderSpecialty, ClaimStatus, and more. These features capture demographic, clinical, and transactional data, which help in identifying claim patterns, patient behaviour, and potential approval likelihood.

3.4 DATASET DESCRIPTION

Table 3.4 -Dataset Description

FEATURES	DESCRIPTION
ClaimID	Unique identifier assigned to each health insurance claim.
PatientID	Unique identifier for each patient who submitted a claim.
ProviderID	Unique identifier representing the healthcare provider.
ClaimAmount	Amount claimed by patient for reimbursement.
ClaimDate	Date on which the claim was submitted or processed.
DiagnosisCode	Medical code for the patient diagnosed illness.
ProcedureCode	Code indicating the specific medical procedure provided.
PatientAge	Age of the patient at the time of the claim.
PatientGender	Gender of the patient (Male or Female).
ProviderSpecialty	Medical specialty of the healthcare provider
PatientIncome	Income range of the patient used for socio-economic analysis.
ClaimStatus	Indicates Current claim status Approved, Denied, or Pending.
PatientIncome	Income range of the patient for socio-economic analysis.
PatientMaritalStatus	Indicates whether the patient is Single, Married, Divorced
PatientEmploymentStatus	Patient’s employment status Employed, Unemployed, Retired.
ProviderLocation	Geographical location of the healthcare provider.
ClaimType	Patients type of claim Inpatient, Outpatient or Emergency.
ClaimSubmissionMethod	Claim Submission mode Online, Phone, or Paper-based.

Table 3.4 shows that dataset contains demographic, medical, and transactional attributes of health insurance claims, including details like patient age, gender, income, diagnosis and procedure codes, claim amount, and claim status. These variables are used to analyze claim patterns, provider behaviour, and predict outcomes such as claim approval.

3.5 DATA PREPROCESSING

Data preprocessing was carried out to clean and prepare the health insurance claim dataset for analysis. Missing values were handled, and categorical columns like gender and claim status were encoded. Dates were converted to datetime format to extract features like claim month and year. Patient age was grouped into categories for demographic analysis. Outliers in claim amounts were identified and highlighted. Numeric columns were scaled where necessary, and unnecessary or duplicate records were removed to ensure the dataset was ready for accurate visualizations.

- **Handling Missing Values**: Null values were identified in critical fields like ClaimAmount, DiagnosisCode, and ClaimStatus. Missing entries in non-essential columns were dropped, while missing numeric values were filled using statistical measures (mean/median) to avoid data distortion in claim amount or status distribution analysis.
- **Data Type Conversion:** The ClaimDate column was converted to datetime format. This allowed the extraction of new temporal features such as Claim Month and ClaimYear, which were later used for trend-based visualizations and seasonal claim pattern analysis.
- **Encoding Categorical Variables:** Encoding Categorical Variables: Categorical variables such as PatientGender, ClaimType, ClaimStatus, ClaimSubmission Method, and ProviderSpecialty were encoded for consistency and to support grouped visualizations. Label encoding was applied to binary features, while one-hot encoding was used where needed for claim types or methods
- **Binning and Grouping:** The PatientAge column was binned into age groups to support demographic breakdowns like “Claims by Age Group”. Similar grouping was applied for income categories and claim amounts for easier interpretation.
- **Feature standardization**: It ensures applied to numerical fields like ClaimAmount and PatientAge. This improves data consistency and ensures reliable analytical results.

4. ANALYSIS AND REPORT

The health insurance claim dataset was carefully analyzed to understand patterns and issues in claim processing. The analysis focused on important areas like patient details, claim amounts, submission methods, diagnosis and procedure codes, and provider information. This helped identify trends, delays, and claim approvals or rejections, making it easier to improve the overall claim process and customer experience.

4.1 GENDER AND AGE-BASED CLAIM ANALYSIS

Gender and age-wise distribution of health insurance claims is examined to identify usage patterns across demographics. Visualizations reveal which age groups or genders contribute to higher claim volumes. Such insights help in optimizing healthcare plans for specific populations**.**

4.1.1 GENDER AND CLAIM DISTRIBUTION

Gender and Claim Distribution analysis examines how health insurance claims vary by gender, revealing which group male or female is more likely to submit claims, and offering insights into gender-based healthcare trends.

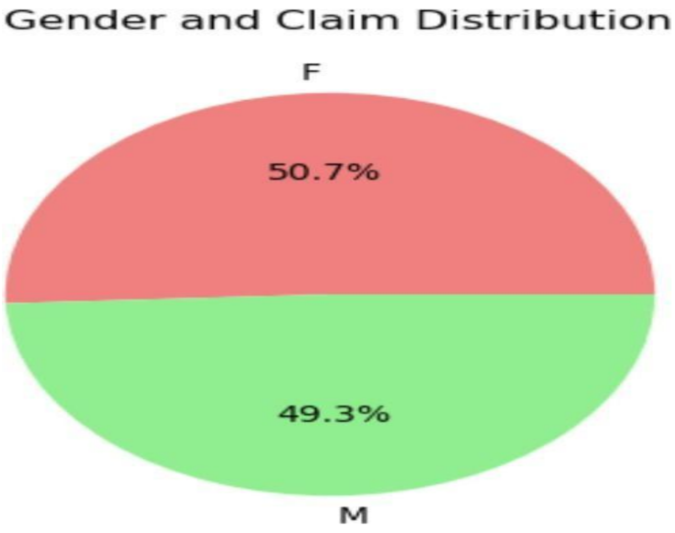


Figure 4.1.1 Gender and Claim Distribution

Figure 4.1.1 represents the pie chart shows that female patients filed 50.7% of the claims, while male patients filed 49.3%. This indicates that both genders claim health insurance almost equally, with a slightly higher proportion from females.

4.1.2 CLAIM TRENDS BY AGE GROUP

Age segment claim analysis highlights the health insurance claims are distributed across various age groups based on the claim data and helps in identifying the age groups with the highest number of claims.

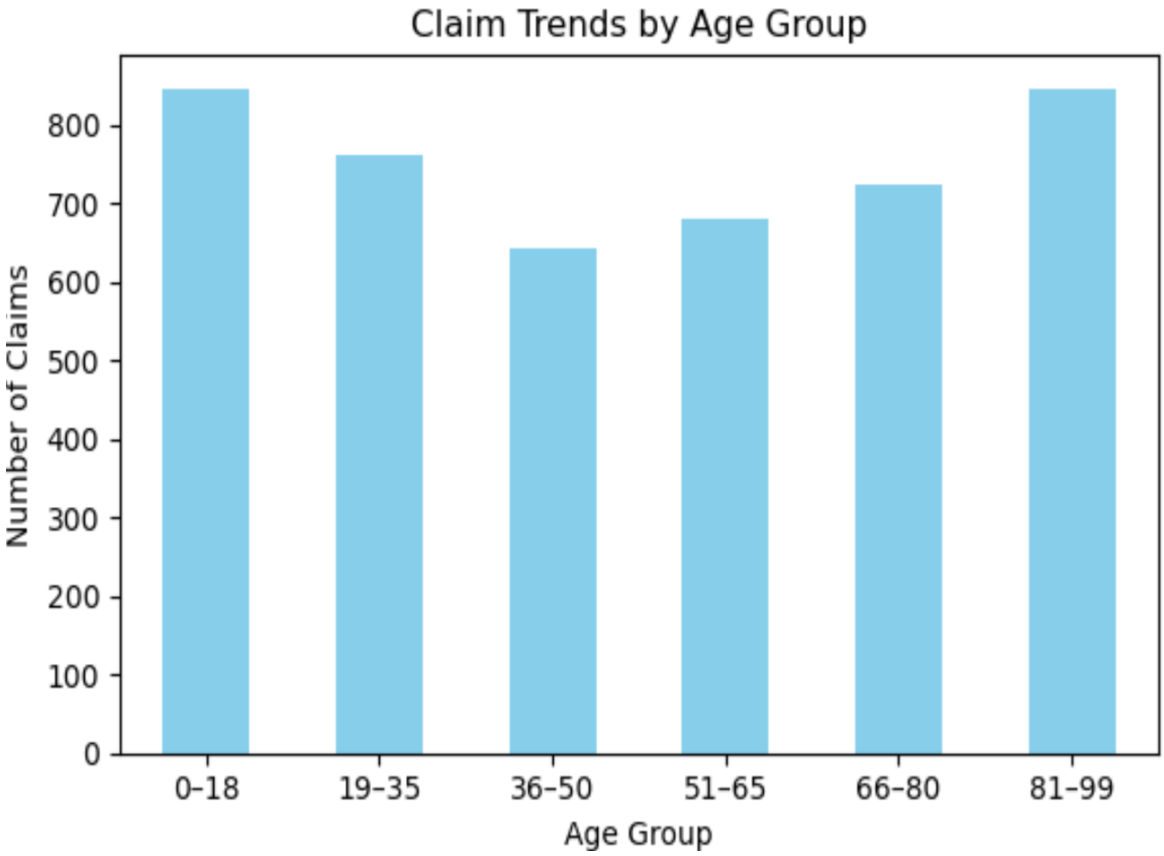


Figure 4.1.2 Claim Trends by Age Group

Figure 4.1.2 This bar chart visually represents how claims are distributed across different age groups. It highlights that the 0–18 and 81–99 age groups have the highest number of claims, showing greater usage of health insurance among the youngest and oldest patients.

4.1.3 MONTHLY HEALTH INSURANCE CLAIMS BY GENDER

The monthly trend of health insurance claims illustrates how male and female patients make claims across different months. By analyzing and comparing these gender-wise patterns, we can identify seasonal variations and determine if one gender tends to claim more frequently during specific periods. This insight, generated using gender-based bar charts for each month, helps healthcare providers and insurers plan resources, design targeted services, and address gender-specific healthcare needs more effectively.

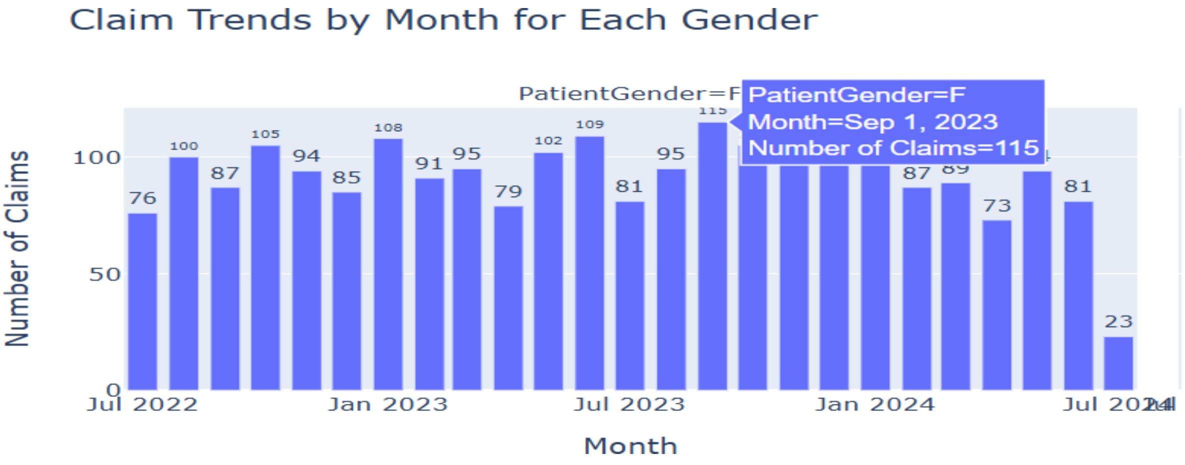


Figure 4.1.3 Monthly Claim Trend – Female Patients

Figure 4.1.3 This bar chart visually represents Monthly Claim Data for Female Patients September 2023.In September 2023, a total of 115 claims were made by female patients, highlighting their active participation in health insurance during this period.

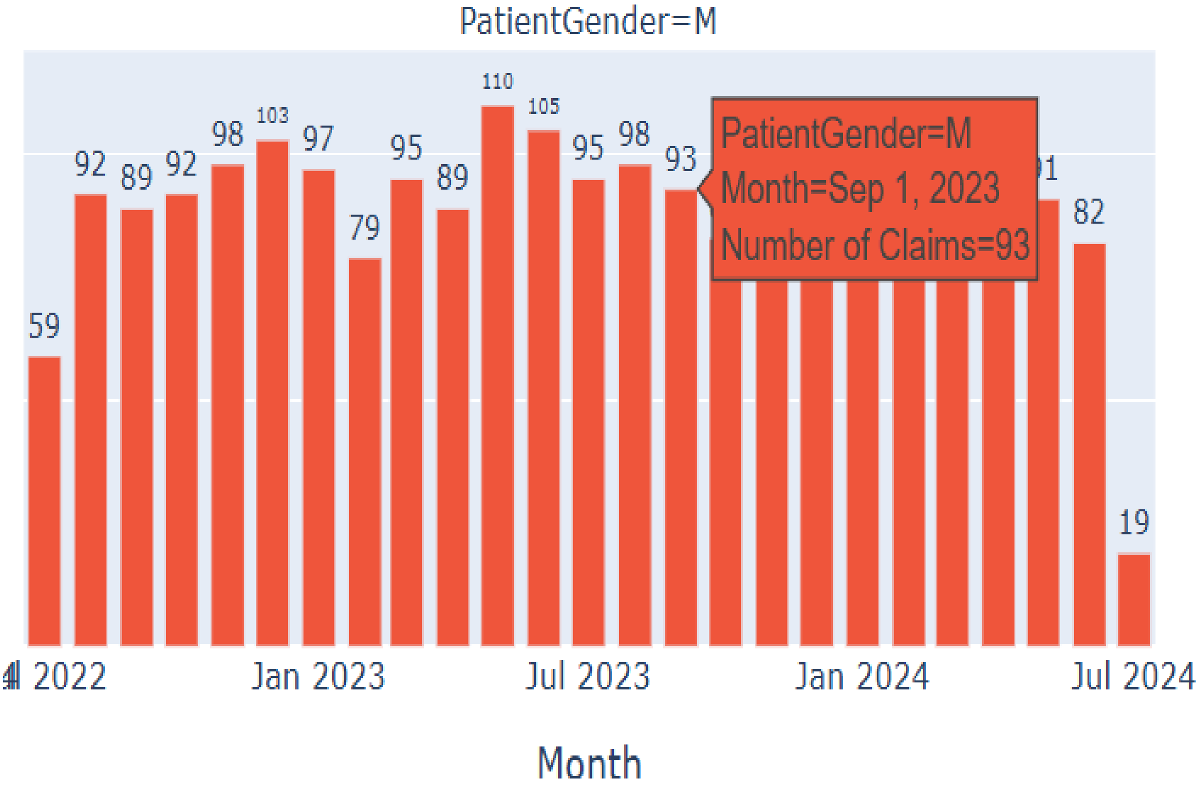


Figure 4.1.4 Monthly Claim Trend – Male Patients

Figure 4.1.4 This bar chart visually represents Monthly Claim Data for Male Patients September 2023. In September 2023, a total of 93 claims were made by male patients, reflecting their significant involvement in health insurance activities during this period.

4.2 SUBMISSION METHODS AND CLAIM STATUS ANALYSIS

Analyzes the relationship between claim submission methods and their approval or rejection outcomes. Reveals trends in processing efficiency across different channels and identifies methods associated with higher success rates.

4.2.1 APPROVED CLAIMS BY SUBMISSION METHOD

Approved Claims by Submission Method shows how many approved claims were submitted using different methods. It helps to see which submission method like online or offline was used most often. This can help insurance companies improve the way people submit their claims.

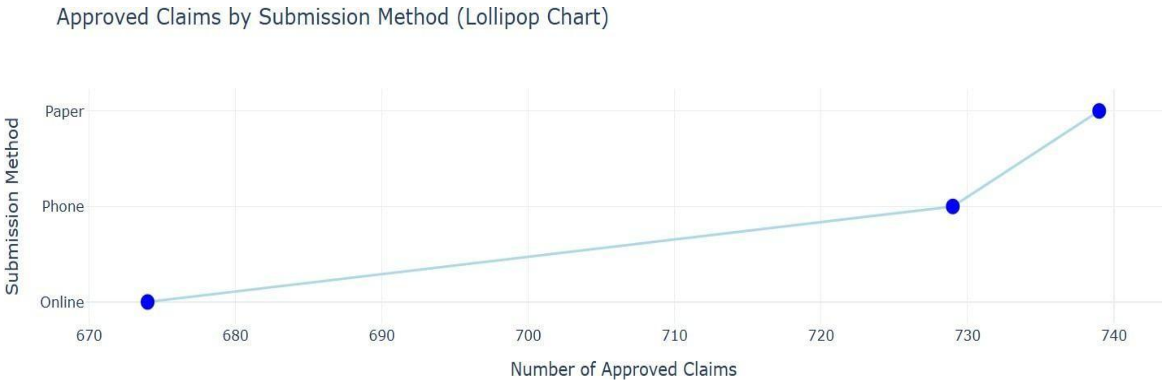


Figure 4.2.1 Approved Claims by Submission Method

Figure 4.2.1This Lollipop Chart displays the number of approved health insurance claims based on how they were submitted. In this case, most claims were approved through paper submission 739, followed by phone 727 and then online 674. This information helps understand which methods are preferred by patients and can guide improvements in the claim submission process**.**

4.2.2 CLAIM SUBMISSION METHODS

Claim Submission Methods were submitted by patients using three different methods Paper, Online, and Phone. It helps us understand which submission method is most commonly used.

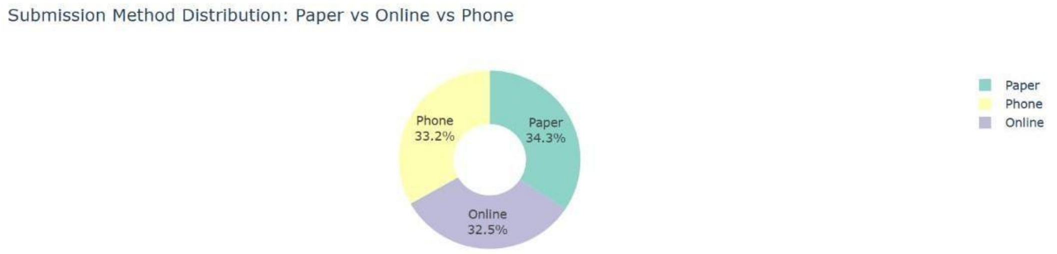


Figure4.2.2 Claims Submission Method

Figure 4.2.2 This donut chart shows health insurance claims were submitted by patients using different methods by Paper 34.3%, Online 32.5%, and Phone 33.2%. It helps to understand which submission method is most commonly used. From the chart, we can easily see the percentage share of each method.

4.2.3 CLAIM STATUS BY SUBMISSION METHOD

Claim Status by Submission Method provides a clear visual breakdown of claim submission methods like Paper, Online, and Phone along with their respective claim statuses Approved, Rejected, Pending. It helps identify which method is most used and how successful each method is in getting claims approved**.**

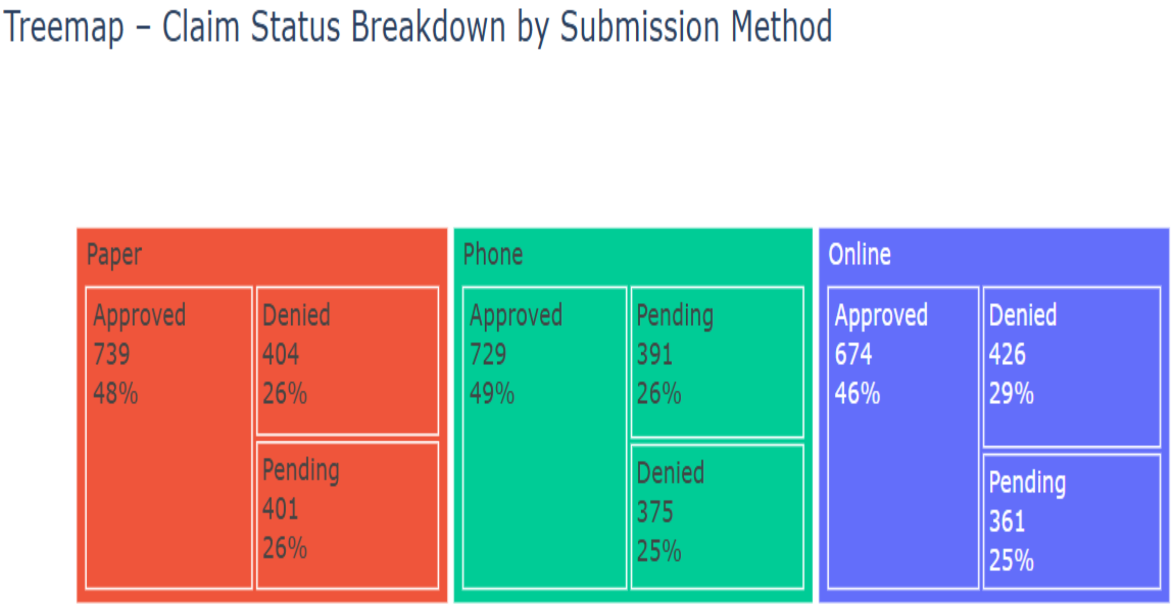


Figure 4.2.3 Claim Status by Submission Method.

Figure 4.2.3 This treemap visually represents the breakdown of claim statuses Approved, Rejected, Pending for each submission method through Paper, Online, Phone. Each

box size indicates the number of claims, making it easy to compare which method is most used and how successful each method is. It helps identify which submission method leads to better approval outcomes.

4.2.4 CLAIM STATUS DISTRIBUTION BY AGE GROUP AND CLAIM TYPE

Claim Status Distribution by Age Group and Claim type to know how many claims were approved, denied, and pending for different age groups under each claim type Inpatient, Outpatient, and Routine. Using a heatmap, it clearly shows the count of each status across age groups without comparing which is highest or lowest.

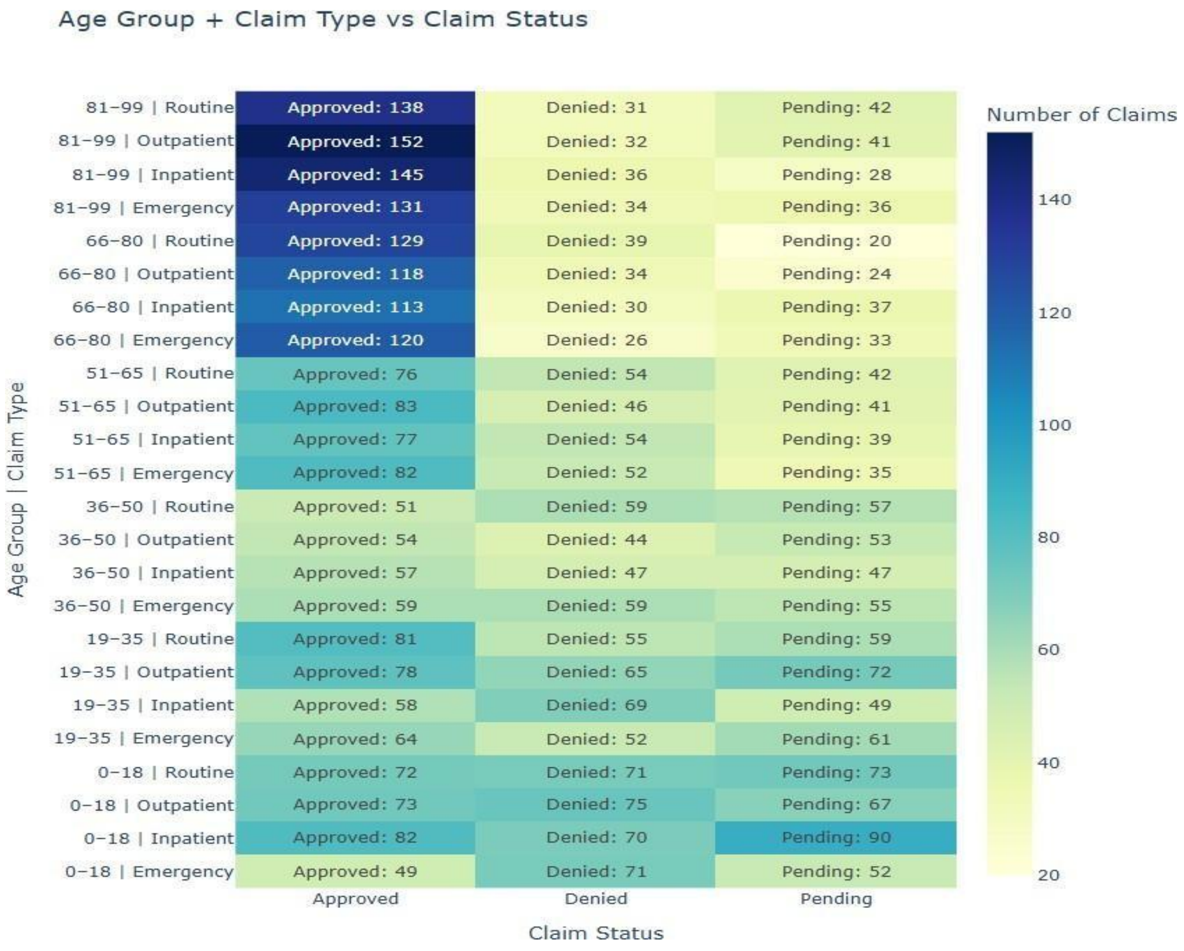


Figure 4.2.4 Claim Status Distribution by Age Group and Claim Type

Figure 4.2.4 This heatmap shows the distribution of claim statuses Approved, Denied, and Pending across different age groups for each claim type: Inpatient, Outpatient, and Routine. It is used to understand how many claims fall into each status category within each age group

and claim type. The heatmap provides a clear view of claim counts without focusing on which is highest or lowest**.**

4.3 PROVIDER AND SPECIALTY-BASED CLAIM ANALYSIS

Analyzes claim patterns across various healthcare providers and specialties. Highlights the specialties with the highest claim volumes and identifies provider locations contributing significantly to overall claims.

4.3.1 SPECIALTIES BY CLAIM VOLUME

Visualized the proportion of health insurance claims across various medical specialties and highlighting to identify which specialties contribute most to the total claim volume.

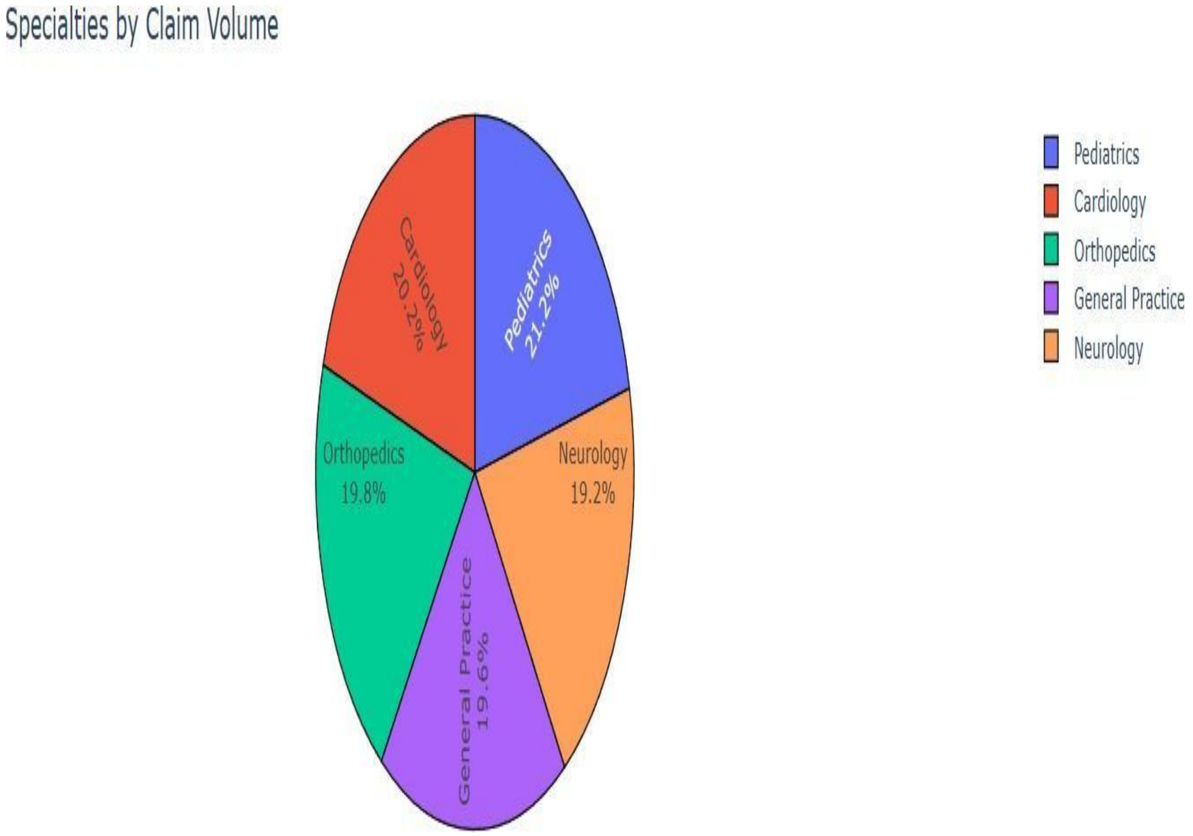


Figure 4.3.1 Specialties by Claim Volume

Figure 4.3.1 The pie chart illustrates the distribution of health insurance claims across key medical specialties. Pediatrics 21.2% and Cardiology 20.2% account for the highest claim volumes, followed by Orthopedics 19.8%, General Practice 19.6%, and Neurology 19.2%. indicating a distribution of claims among these specialties**.**

4.3.2 PROVIDER LOCATIONS WITH MAXIMUM HEALTH CLAIMS

This horizontal bar chart presents the top 10 provider locations that recorded the highest number of health insurance claims. Each bar represents a location and the total claims submitted from that area. The visualization helps identify regions with high claim activity, offering useful insights into geographic trends and healthcare service utilization patterns.

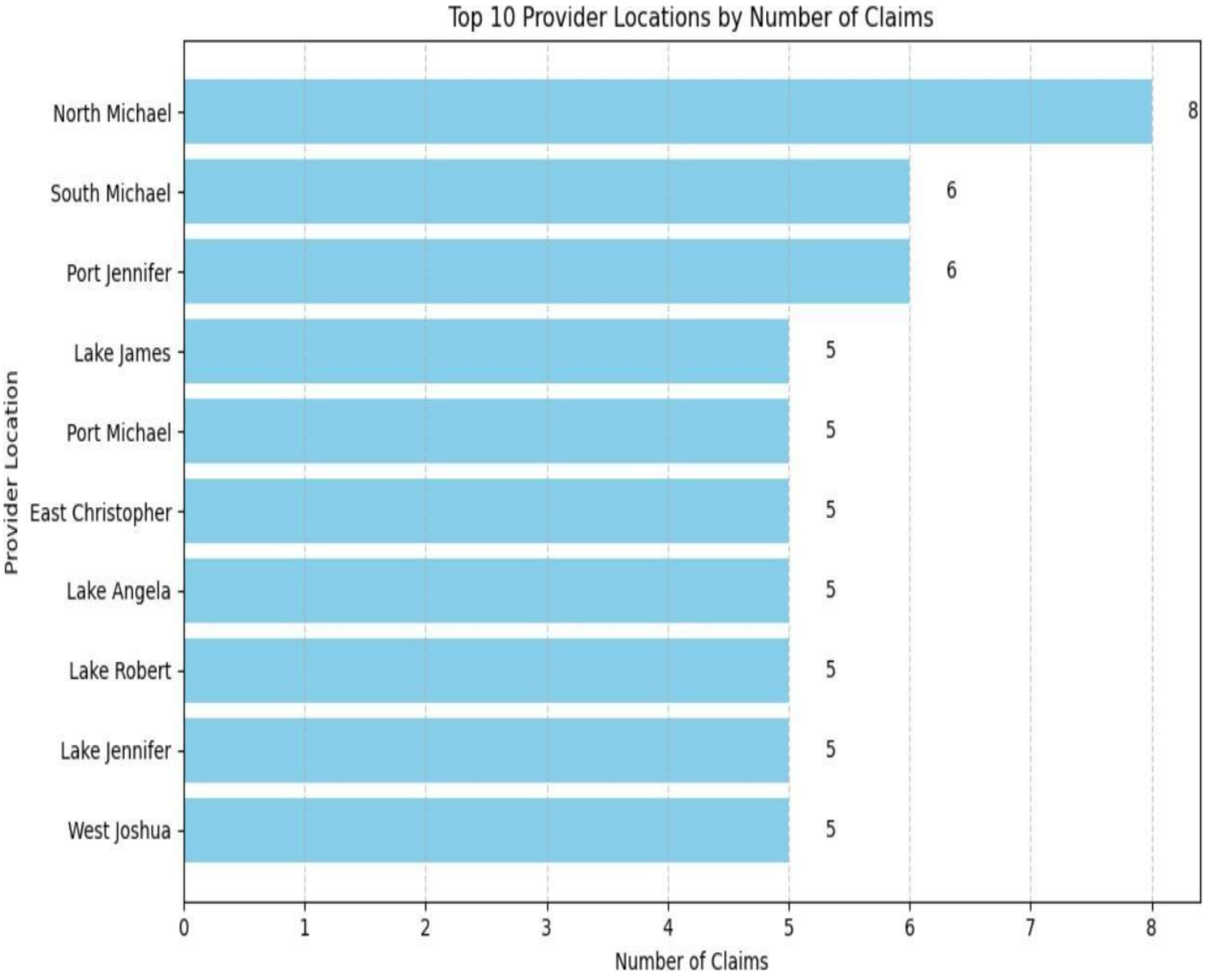


Figure 4.3.2 Provider Locations with Maximum Health Claims

Figure 4.3.2 This horizontal bar chart displays the top 10 provider locations with the highest number of health insurance claims. North Michael tops the list with 8 claims, followed by South Michael and Port Jennifer with 6 claims each. The remaining locations, including Lake James, Port Michael, and others, each recorded 5 claims. This visualization highlights regional trends in claim activity and helps identify locations with greater demand for healthcare services.

4.4 ANALYSIS OF CLAIM AMOUNTS AND VOLUME PATTERNS

Claim value and frequency analysis focuses on how claim amounts vary across individual cases and how frequently different claim ranges appear within the dataset. It provides deeper insights into the overall distribution of financial claims, highlighting zones with high-volume, low-cost claims versus low-frequency, high-value claims. Understanding these patterns helps identify common cost brackets, outliers, and financial impact zones that are critical for optimizing policy design and resource allocation.

4.4.1 COMPARISON OF EXTREME CLAIM AMOUNTS WITH DETAILS

This visualization is used to compare the highest and lowest health insurance claims, helping to understand the characteristics of extreme cases. It aids in identifying patterns related to large and small claims, which can be useful for risk assessment, policy design.

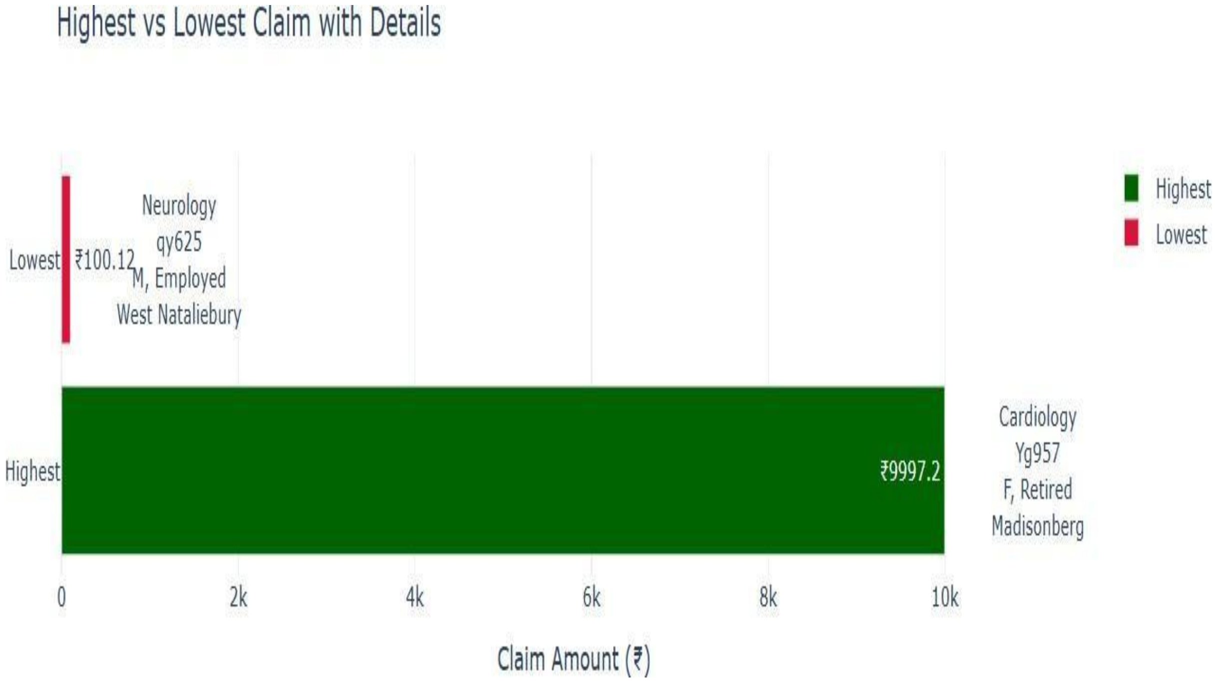


Figure 4.4.1 Comparison Of Extreme Claim Amounts With Details

Figure 4.4.1The horizontal bar chart shows the highest and lowest claim amounts with key details. The highest claim is ₹9,997.20 for a retired female patient in Cardiology the procedurecode YG957 from Madisonberg. The lowest claim is ₹1,000.12 for an employed male patient in Neurology procedurecode QY625 from West Nataliebury. It highlights claim amount extremes along with patient and provider information.

4.4.2 CLAIM COUNT DISTRIBUTION BY CLAIM AMOUNT RANGE

Analyzing the frequency of claims across different claim amount ranges to understand common claim values, detect cost patterns, and support decision-making in health insurance policy design and risk assessment.

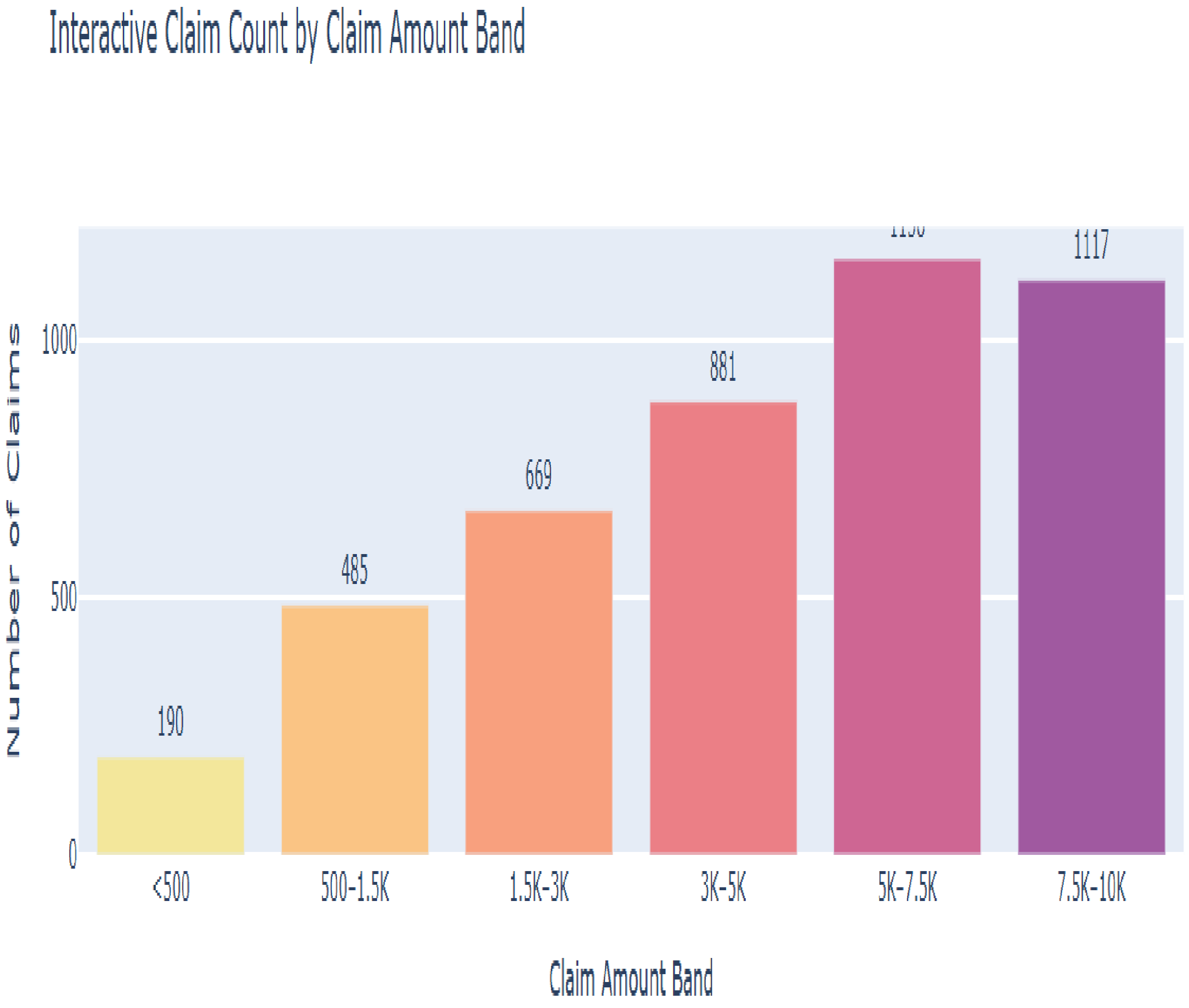


Figure 4.4.2 Claim Count Distribution By Claim Amount Range

Figure 4.4.2 This bar chart shows the distribution of health insurance claims across various claim amount bands. The number of claims in each range is as follows: less than ₹500 190 claims, ₹500–1.5K – 485 claims, ₹1.5K–3K – 669 claims, ₹5K–7.5K – 1,158 claims, and ₹7.5K–10K – 1,117 claims. From the chart, it is evident that higher claim amounts, especially between ₹5K and ₹10K, are more frequent. This analysis helps identify common claim value ranges and supports better understanding of claim cost patterns within the dataset.

5. CONCLUSION

The health insurance claim analysis project provided valuable insights into the behaviour and distribution of claims using data visualization techniques. By examining key attributes such as claim amount, provider location, patient age group, claim type, and claim status, the study offered a clear understanding of how different factors influence claim outcomes. A significant portion of claims was found in the ₹5,000–₹10,000 range, indicating that many policyholders file claims for moderate to serious medical treatments. This insight is crucial for insurers when designing coverage plans and allocating financial resources effectively.

Geographical analysis revealed that certain provider locations consistently handled higher volumes of claims, highlighting regional trends in healthcare access and utilization. Age group analysis showed that elderly patients, particularly those aged 81–99, submitted a larger number of approved claims especially for routine and outpatient care—emphasizing the growing medical needs of aging populations.

Claim behaviour also varied based on demographic factors like gender, income, and marital status, helping insurers tailor services more effectively. The submission method analysis showed that online claims were generally more efficient, with higher approval rates. Provider specialties with frequent or high-cost claims were identified, guiding insurers in resource allocation and partnership strategies.

To summarize, this health insurance claim analysis demonstrated how data-driven approaches can enhance decision-making, increase transparency, and improve operational efficiency in the health insurance sector. These findings are not only useful for insurers but also for healthcare providers and policymakers aiming to deliver better, more equitable services.

To further enhancements may include applying machine learning models to predict claim approval or detect anomalies for fraud prevention. Real-time claim data integration and interactive dashboards can provide dynamic, user-friendly insights to support quicker decisions. Predictive modeling could help forecast healthcare demand and optimize premium pricing. Additionally, incorporating more granular features such as diagnosis severity, treatment duration, or follow-up visits can enable a deeper and more accurate analysis of claim trends.

REFERENCES

1. Hitesh S.M. and Yukthi A., Sales Prediction Using Machine Learning Techniques, published in the International Journal of Novel Research and Development (IJNRD), Volume 9, Issue 1, January 2023.
2. Giordani M., Polese M., Roy A., Castor D., and Zorzi M., Toward 6G Networks: Use Cases and Technologies, published in IEEE Communications Magazine, Volume 58, Issue 3, March 2020.
3. Kumar P. and Singh R., Quality of Service Management in 6G Networks, published in ResearchGate, August 2022.
4. Xu J., Fang J., and Lin W., AI-Driven Monitoring in Healthcare Networks: A Real-Time Approach, published in the National Center for Biotechnology Information (NCBI), February 2021.
5. Sharma R. and Mehta V., Real-Time Network Performance Analysis in Healthcare IoT Systems, published in the International Journal of Computer Applications (IJCA), Volume 184, Issue 42, April 2022.
6. Sharma R. and Mehta V., Real-Time Network Performance Analysis in Healthcare IoT Systems, published in the International Journal of Computer Applications (IJCA), Volume 184, Issue 42, April 2022.
7. Xu J., Fang J., and Lin W., AI-Driven Monitoring in Healthcare Networks: A Real-Time Approach, published in the National Center for Biotechnology Information (NCBI), February 2021.
8. Pandas Documentation, Python Data Analysis Library, available at https://pandas.pydata.org/, accessed July 2025.
9. Hunter J. D., Matplotlib: A 2D Graphics Environment, published in Computing in Science & Engineering, Volume 9, Issue 3, May 2007.
10. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., et al Scikit-learn Machine Learning in Python, published in the Journal of Machine Learning Research, Volume 12, 2011.