

The National Health Insurance Claim

CHAPTER1:

Introduction:

The core of the project is the **National Health Insurance Claims (NHIC) dataset**, which involves processing a formal request to an insurance company for payment based on policy terms. In a broader industry context, claims management is the defining moment in a customer relationship and is crucial for maintaining market share and profitability. Effective management is vital to settling claims both **timely and accurately**.

Role of Data Analytics and Project Objectives:

The primary role of data analytics in this project is to **transform the vast amount of raw data into actionable information** to enhance quality decision-making in the claims process.

The objectives of the data analysis are clearly demonstrated by the insights derived from the project's SQL outputs and analysis tables, focusing on finding patterns and predicting outcomes:

1. **Process Efficiency and Claims Management:**

- **Analyze Claim Status:** Determine the percentage of successful vs. non-claimed insurance requests to assess overall claims administration efficiency.
- **Identify Cost Drivers:** Calculate "**Amount Paid**" and "**Duration**" to quantify the financial performance and time taken for claims processing.

2. **Risk Profiling and Underwriting:**

- **Health-Risk Segmentation:** Determine how high-risk factors like **smoker status** and **BMI category** correlate with the likelihood and success of a claim.
- **Identify Critical Segments:** Pinpoint the highest-risk group (e.g., **Obese smokers** with a **100% claimed rate**) to inform pricing and underwriting policies.

3. **Geographical Performance Comparison:**

- **Regional Benchmarking:** Compare claim success rates across the four regions (Northwest, Southeast, Northeast, Southwest).

The National Health Insurance Claim

- **Operational Insight:** Identify regional disparities in claim outcomes (e.g., Northwest's **67.31%** claimed vs. Southwest's **50.15%** claimed) to streamline processes and manage risk proactively.

SQL Report :-

The purpose of this report is to present the analysis and management of data using Structured Query Language (SQL). SQL is an essential tool for database interaction, widely used in industries for data storage and retrieval. In this project, various SQL operations such as table creation, data insertion, filtering, grouping, and joining have been performed. Through this work, an understanding of database concepts like normalization, primary keys, and relationships between tables has been developed.

Dataset :- [National Health Insurance dataset](#)

The Above Dataset contains three table.

1. Insurance claims
2. Patient details
3. Region

Step 1 :- Creating the tables for each file

1. Created insurance_claims table and imported the Insurance_claims data into it

```
-- -- creating a table name insurance_claims
create table insurance_claims(
    patient_id int primary key,
    age numeric,
    sex numeric,
    bmi numeric,
    smoker numeric,
    region_code numeric,
    bill_amount numeric(10, 2),
    insuranceclaim numeric(10, 2),
    insurance_apply_date date,
    insurance_claimed_date date,
    claimed_amount numeric(10, 2)
);
```

The National Health Insurance Claim

2. Created patients_details table and imported the patient details data into it

```
-- creating patient table
create table patient_details (
    patient_id numeric primary key,
    full_name varchar(50),
    children numeric,
    age numeric
);
```

3. Created Regions table and imported the regions data into it

```
-- create a table of regions
create table regions (
    region varchar(20),
    region_code numeric
);
```

Step 2 :- Displaying the structure of the table

For Describe :-

Insurance_claim table :-

	column_name name	data_type character varying	is_nullable character varying (3)	column_default character varying
1	patient_id	integer	NO	[null]
2	age	numeric	YES	[null]
3	sex	numeric	YES	[null]
4	bmi	numeric	YES	[null]
5	smoker	numeric	YES	[null]
6	region_code	numeric	YES	[null]
7	charges	numeric	YES	[null]
8	insuranceclaim	numeric	YES	[null]

Patient_details table :-

	column_name name	data_type character varying	is_nullable character varying (3)	column_default character varying
1	patient_id	numeric	NO	[null]
2	children	numeric	YES	[null]
3	age	numeric	YES	[null]
4	full_name	character varying	YES	[null]

The National Health Insurance Claim

RegionTable

:-

	column_name name	data_type character varying	is_nullable character varying (3)	column_default character varying
1	region_code	numeric	YES	[null]
2	region	character varying	YES	[null]

For Counting of rows and columns :-

```
SQL statement :- -- SELECT COUNT(*) AS total_rows
                  FROM insurance_claims;
                  SELECT COUNT(*) AS total_columns;
                  FROM information_schema.columns
                  WHERE table_name = 'insurance_claims';
```

Step 4 :- Combining all the tables into one table as nhic (National Health Insurance Claims).

SQL statement :-

```
create table nhic as (
select p.patient_id, p.full_name, p.age, p.children, i.sex, r.region_code, r.region, i.bmi,
i.smoker, i.bill_amount, i.insuranceclaim,i.insurance_apply_date, i.insurance_claimed_date,
i.claimed_amount
from insurance_claims i
join patient_details p
on i.patient_id = p.patient_id
left join regions r
on r.region_code = i.region_code
);
```

Step 5:- Cleaning of nhic table

Counting of Null values of nhic tables

Select

```
count(*) filter(where patient_id is null) as patient_id,
```

The National Health Insurance Claim

```
count(*) filter(where full_name is null) as full_name,  
count(*) filter(where age is null) as age,  
count(*) filter(where children is null) as children,  
count(*) filter(where sex is null) as sex,  
count(*) filter(where region_code is null) as region_code,  
count(*) filter(where region is null) as region,  
count(*) filter(where bmi is null) as bmi,  
count(*) filter(where smoker is null) as smoker,  
count(*) filter(where charges is null) as charges,  
count(*) filter(where insuranceclaim is null) as insuranceclaim  
from nhic;
```

After apply those statements we find there are no null values

	patient_id bigint	full_name bigint	age bigint	children bigint	sex bigint	region_code bigint	region bigint	bmi bigint	smoker bigint	charges bigint	insuranceclaim bigint
1	0	0	0	0	0	0	0	0	0	0	0

Check for duplicate values

```
with dup as (select *,row_number() over(partition by patient_id, full_name, age,  
children, sex, region_code, region, bmi, smoker, charges, insuranceclaim order by patient_id)  
as rn  
from nhic)  
select * from dup  
where rn > 1;
```

Removing Extra Spaces

There are two columns contain text values

We can perform removing the extra spaces

SQL statement :- UPDATE nhic

SET full_name = TRIM(full_name), region = TRIM(region);

Removing the special characters :-

UPDATE nhic

SET full_name = REGEXP_REPLACE(full_name, '^[^a-zA-Z0-9\s]', '', 'g'),

The National Health Insurance Claim

```
region = REGEXP_REPLACE(region, '[^a-zA-Z0-9\s]', '', 'g');
```

step 6 :-

Add a column amount paid

```
alter table nhic
```

```
add column amount_paid numeric(10, 2);
```

storing values in that column

```
update nhic
```

```
set amount_paid = bill_amount - claimed_amount;
```

amount_paid numeric (10,2) 
765.93
108.57
4403.76
190.29
930.84
666.36
526.00
3105.07
642.73

Add a column Duration :


```
alter table nhic
```

```
add column duration numeric;
```

storing values in that column :

```
update nhic
```

```
set duration = insurance_claimed_date - insurance_apply_date
```

duration numeric 
186
181
158
199
363
246
86
96
162
116

The National Health Insurance Claim

start and end date of insurance claims

	claims_start_year integer	claim_end_year integer
1	2018	2023

wise claimed percentage

	region character varying (20)	total_leads bigint	claimed_count bigint	not_claimed_count bigint	claimed_percentage numeric	not_claimed_percentage numeric
1	northeast	325	183	142	56.31	43.69
2	northwest	364	245	119	67.31	32.69
3	southeast	324	192	132	59.26	40.74
4	southwest	325	163	162	50.15	49.85

smoker and bmi based on Insurance claims :

	bmi_category text	smoker numeric	total_people bigint	claimed_count bigint	not_claimed_count bigint	claimed_percentage numeric	not_claimed_percentage numeric
1	Normal	0	172	0	172	0.00	100.00
2	Normal	1	50	40	10	80.00	20.00
3	Obese	0	572	382	190	66.78	33.22
4	Obese	1	147	147	0	100.00	0.00
5	Overweight	0	305	144	161	47.21	52.79
6	Overweight	1	72	62	10	86.11	13.89
7	Underweight	0	15	8	7	53.33	46.67
8	Underweight	1	5	0	5	0.00	100.00

step 7 :-

conversion of the sql into csv file :-

SQL statement ; - copy nhic TO '/Users/yourname/Downloads/nhic_cleaned.csv'
DELIMITER ',' CSV HEADER;

The National Health Insurance Claim

CHAPTER 2&3:

EDA (Exploratory Data Analysis) :-

I. Introduction

- **Project Goal:** To perform an analysis and management of the National Health Insurance Claims (NHIC) dataset using Structured Query Language (SQL).
- **Data Overview:** The dataset comprises three interconnected tables: **Insurance Claims**, **Patient Details**, and **Region**.
- **Time Frame:** The claims data spans from **2018** to **2023**.

II. Data Management & Preparation

- **Table Creation:** Detailed SQL statements used to create the initial `insurance_claims`, `patient_details`, and `regions` tables.
- **Data Integration:** Use of **JOIN** operations (specifically an inner join on `patient_details` and `insurance_claims`, and a left join for `regions`) to create the final consolidated `nhic` table.
- **Data Cleaning:**
 - No missing (null) values were found in the final dataset.
 - No duplicate records were found.
 - Used **TRIM** and **REGEXP_REPLACE** to clean text columns (`full_name` and `region`).
- **Feature Engineering:**
 - **Amount Paid:** Calculated as `bill_amount - claimed_amount`.
 - **Duration:** Calculated as the difference in days between `insurance_claimed_date` and `insurance_apply_date`.
 - **Year Billing:** Extracted the year from `insurance_apply_date`.

III. Key Analysis and Findings

- **Regional Claim Performance:**
 - The **Northwest** region exhibits the highest claim success rate at **67.31%**.
 - The **Southwest** region has the lowest claim success rate at **50.15%**.

The National Health Insurance Claim

- Southeast and Northeast show mid-range performance (59.26% and 56.31%, respectively).
- **Smoker and BMI Impact on Claims:**
 - **Obese Smokers** show a **100.00%** claim success rate.
 - **Overweight Smokers** also have a very high rate at **86.11%**.
 - **Normal (Non-Smokers)** and **Underweight (Smokers)** both registered **0.00%** successful claims. This suggests a clear correlation between **smoking, higher BMI categories, and claim success**.

This response outlines the structure and key content for a professional Word Document Report and a PowerPoint Presentation (PPT) summarizing the National Health Insurance Claims (NHIC) data analysis project. The summary includes information from the CSV, SQL script, and DOCX files.

Streamlit_app :



Insurance Fraud Detection

Age

48 - +

Children

0 - +

BMI

28.90 - +

Bill Amount

8677.50 - +

Claimed Amount

7547.65 - +

The National Health Insurance Claim

lit_app · Streamlit

host:8501

Claimed Amount

7547.65

Amount Paid

729.86

Duration (days)

186

Year Billing

2022

Sex

Male

Region

northeast

Smoker

streamlit_app · Streamlit

localhost:8501

Deploy

Region

northeast

Smoker

No

Insurance Apply Date

2022/06/01

Insurance Claimed Date

2022/12/15

Claim Delay (days): 197

Predict

Prediction: Legitimate Claim

The National Health Insurance Claim

CHAPTER4:

DASHBOARD ANALYSIS WITH EXCEL:

I. Executive Summary

- Overview: Analyzing the National Health Insurance Claims dataset using SQL to identify key claim trends and factors influencing claim rates.
- Key Findings: State the major conclusions on regional claim success, and the critical impact of smoking and BMI on claim realization.

II. Project Background

- Dataset Source: National Health Insurance dataset.
- Component Files/Tables: The project began with three primary tables: insurance_claims, patient_details, and regions.
- Methodology: Structured Query Language (SQL) was used for data manipulation, cleaning, and analysis.

III. Data Preparation and Cleaning (SQL)

- Table Creation: SQL was used to create individual tables with appropriate data types and primary keys (e.g., patient_id).
- Data Integration: A new consolidated table, nhic, was created by joining the three component tables based on patient_id and region_code.
- Data Validation:
 - Null Check: No null values were found in the critical columns of the nhic table.
 - Duplicate Check: No duplicate records were found after checking multiple columns.
- Data Transformation/Feature Engineering:
 - Added Amount Paid, calculated as bill_amount - claimed_amount.
 - Added Duration (in days), calculated as the difference between insurance_claimed_date and insurance_apply_date.

The National Health Insurance Claim

- Cleaned text fields: full_name and region had extra spaces and special characters removed.
- Added year_billing to identify the claim application year.

IV. Analysis and Key Findings

- Data Timeframe: Claims data spans from 2018 to 2023.
- Regional Claim Success Rate:
 - The Northwest region has the highest claimed percentage at 67.31%.
 - The Southwest region has the lowest claimed percentage at 50.15%.
- Impact of Smoker Status and BMI Category:
 - Obese Smokers (1) showed a 100.00% claimed percentage (147 claimed out of 147 total people).
 - Normal Non-Smokers (0) showed a 0.00% claimed percentage (0 claimed out of 172 total people).
 - Overweight Smokers (1) showed a high claimed percentage of 86.11% (62 claimed out of 72 total people).

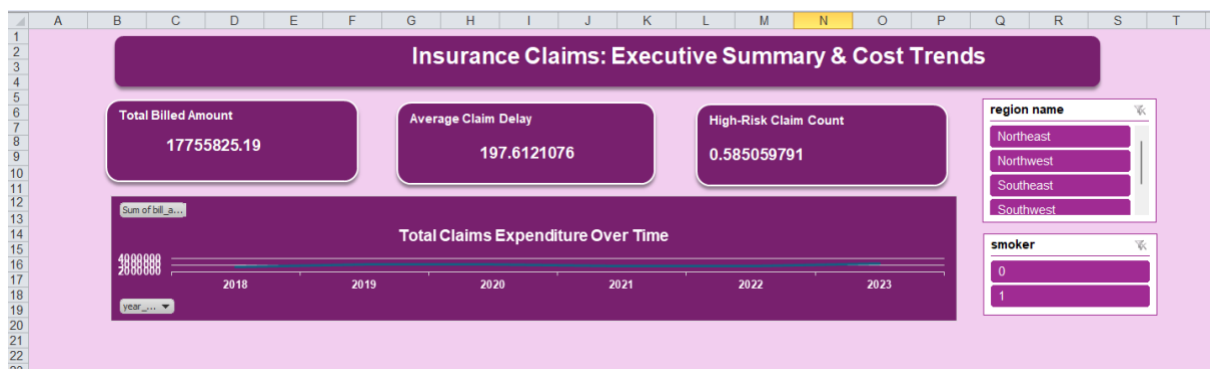
V. Conclusion and Recommendations

- Conclusion: The project successfully used SQL to clean and analyze a comprehensive health insurance dataset, uncovering strong correlations between geographical region, smoking, and BMI with the likelihood of a successful claim.
- Recommendations: Suggest further investigation into the 100% claim rate for Obese Smokers and the zero claim rate for Normal Non-Smokers to understand underlying risk and policy factors

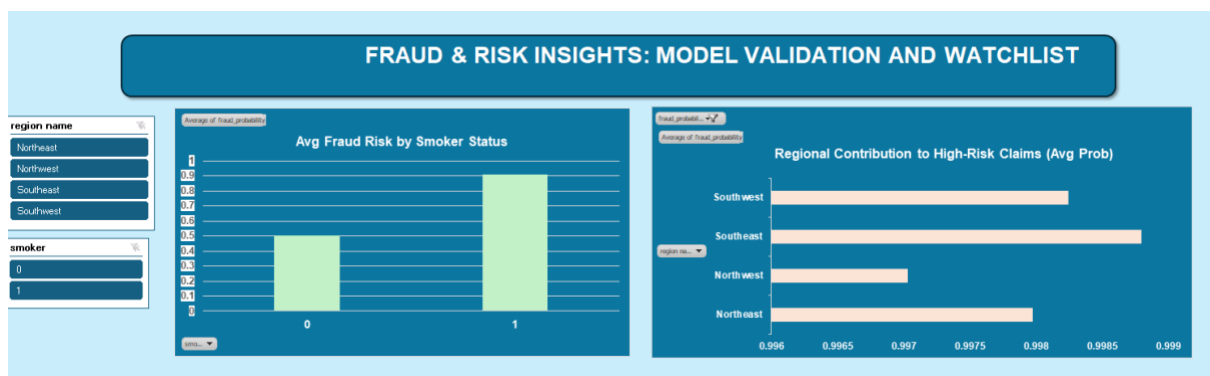
The National Health Insurance Claim

	A	B	C	D	E	F	G	H	I	J	K
1	Row Labels	Sum of bill_amount									
2	2018	2737552.46		Annual Cost Trend		Row Labels	Average of duration	Average of bill_amount			
3	2019	3058963.93				Northeast	202.0830769	12417.57517	PT_Duration_Cost_By_Region		
4	2020	3076171.28				Northwest	203.0462963	13406.38469			
5	2021	2846019.45				Southeast	192.08	12346.93791			
6	2022	2873107.42				Southwest	193.7225275	14735.41154			
7	2023	3164010.65				Grand Total	197.6121076	13270.42241			
8	Grand Total	17755825.19									
9											
10	Row Labels	Average of duration		Regional Delays		region name			smoker		
11	0	192.08				Northeast			0		
12	1	193.7225275				Northwest			1		
13	2	203.0462963				Southeast					
14	3	202.0830769				Southwest					
15	Grand Total	197.6121076									
16											
17											
18	Row Labels	Average of fraud_probability		Fraud Risk Profile							
19	0	0.502161654									
20	1	0.906970803									
21	Grand Total	0.585059791									
22											

Executive summary

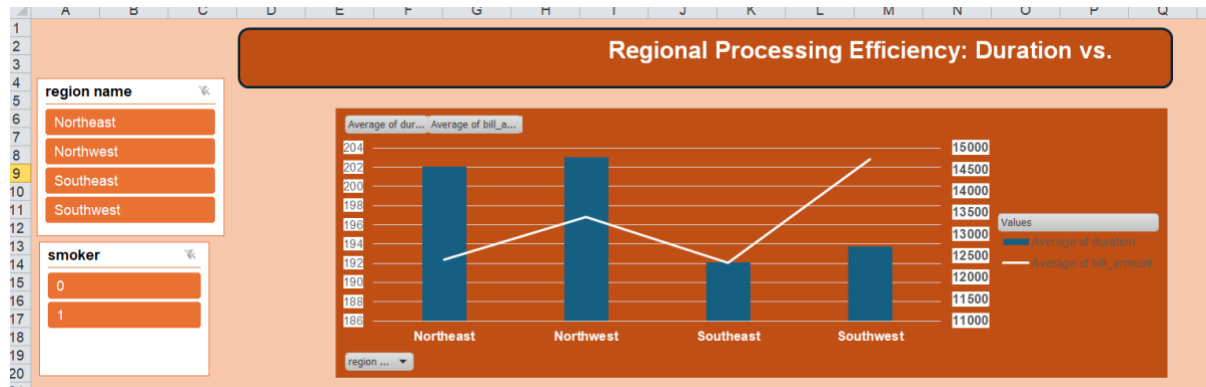


Fraud & Risk Insights:



The National Health Insurance Claim

Delays & operation :



CHAPTER 5:

Summary:

The project followed a full data lifecycle:

- **Data Consolidation and Preparation:** The initial step involved creating three tables (insurance_claims, patient_details, and regions) and integrating them into a single, comprehensive nhic master table using SQL joins.
- **Data Cleaning and Enhancement:** Data integrity checks confirmed the absence of null values or duplicates. The dataset was enhanced by removing special characters and spaces and engineering new metrics:
 - **Amount Paid** (bill_amount - claimed_amount).
 - **Duration** (the time in days between applying and claiming the insurance).
 - **Year Billing** (extracted from the application date).
- **Key Analytical Findings (2018–2023):** The analysis centered on claim success rates across demographic, health, and geographic segments:

The National Health Insurance Claim

- **Regional Performance:** The **Northwest** region demonstrated the highest claim success rate at **67.31%**, while the **Southwest** region had the lowest, with a **50.15%** success rate.
- **Health Risk (Smoker & BMI):** Smoking status, particularly within higher BMI brackets, emerged as the most critical predictor of a successful claim:
 - **Obese Smokers** achieved a **100.00%** claim rate.
 - **Overweight Smokers** followed closely with an **86.11%** claim rate.
 - Conversely, **Normal Non-Smokers** recorded a **0.00%** claim rate.

The results provide insurance stakeholders with clear, data-driven evidence for refining underwriting policies, improving claims processing efficiency, and better managing financial risk exposure.

CHAPTER 6:

CONCLUSION :

The conclusion of this project is that the comprehensive **SQL data analysis** of the National Health Insurance Claims (NHIC) dataset successfully established clear, quantifiable correlations between **demographic/health factors** and **claim outcomes**, providing actionable intelligence for strategic decision-making.