DSC 550

OpenFDA Tobacco Reports Analysis

By Adithya Prahasith

Agenda

- Introduction
- Problems Addressed
- Impact / Influence
- Project Approach
- Data Overview
- Tools & Technologies
- Insights Walkthrough
- Forecasting
- RAG End User Application
- Conclusion & Future Improvements

Introduction

> Tobacco remains the leading cause of preventable death and disease in America, killing 490,000 people each year, including 50,000 Black adults, 15,000 Hispanic adults and 400,000 white adults.
>
> **American Lung Association**

- Tobacco consumption is a major global health threat, contributing to diseases like lung cancer, heart disease, and chronic respiratory issues.
- Despite awareness campaigns, the rates of consumption remain high, posing significant risks to public health.
- This project aims to address this issue by using data-driven insights to forecast and classify health problems related to tobacco use.
- By providing early warnings and easy access to complaints data, this project helps regulators, healthcare providers, and the public take proactive steps to reduce tobacco consumption and its associated health risks.

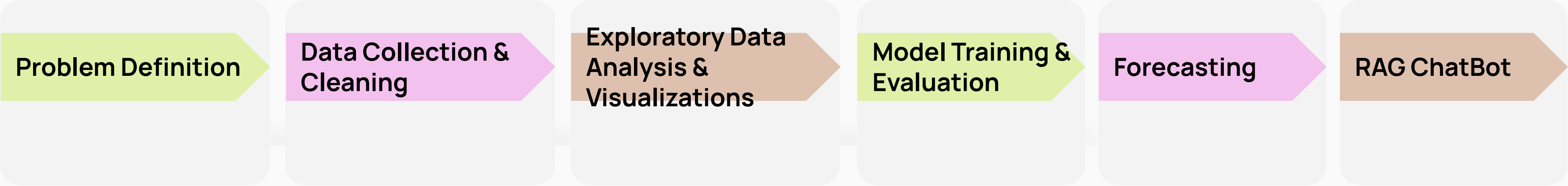
Problems Addressed

- Despite the known health risks of tobacco consumption, such as respiratory diseases, heart conditions, and cancer, there remains a gap in actively monitoring and addressing emerging public health threats linked to tobacco use.
- The FDA collects tobacco-related complaints, but the data is static and underutilized, making it difficult to predict future risks or efficiently identify trends. Current methods rely heavily on manual analysis, which is time-consuming and prone to oversight.
- This project aims to solve these issues by building predictive models to forecast health problems and classify complaint types, helping to detect rising health threats early.
- Additionally, we develop a Retrieval-Augmented Generation (RAG) system to make it easier for stakeholders to access relevant information from the data, empowering faster, data-driven responses to tobacco-related health concerns.

Impact / Influence

- **FDA & Regulators:** Early detection of health risks from tobacco products, aiding faster regulatory interventions.
- **Healthcare Providers:** Better understanding of tobacco-associated illnesses (especially respiratory and cardiovascular issues).
- **Manufacturers:** Insight into frequent product defects or adverse health effects to improve product design and safety.
- **Consumers & Public:** Easy access to complaint trends through RAG applications, enabling informed personal choices about tobacco product use.
- **Researchers:** Enhanced access to structured complaint datasets for public health studies and policy recommendations.

Project Approach



Problem Definition

Data Collection & Cleaning

Exploratory Data Analysis & Visualizations

Model Training & Evaluation

Forecasting

RAG ChatBot

Classification Models

Predict whether a complaint is likely related to respiratory or cardiovascular problems.

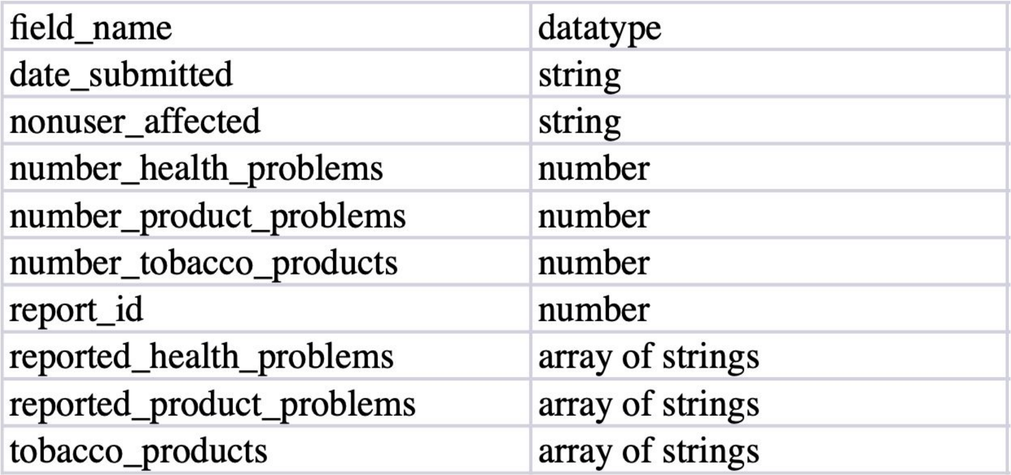
Forecasting the Health Problems

Forecast the number of health and product complaints for the next 2 years using historical data.

ChatBot

Build an AI assistant that answers specific questions from the complaint dataset only.

Data Overview



field_name	datatype
date_submitted	string
nonuser_affected	string
number_health_problems	number
number_product_problems	number
number_tobacco_products	number
report_id	number
reported_health_problems	array of strings
reported_product_problems	array of strings
tobacco_products	array of strings

OpenFDA API by FDA

Data Source

JSON

Data Format

1250

Tobacco Problem Reports

9

Key Features

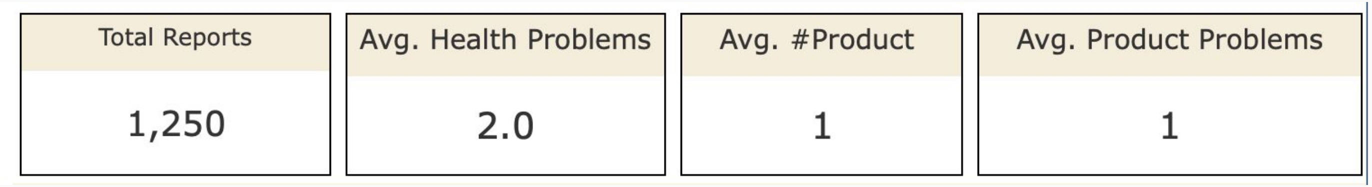
2017-2024

Reported Years

Tools & Technologies

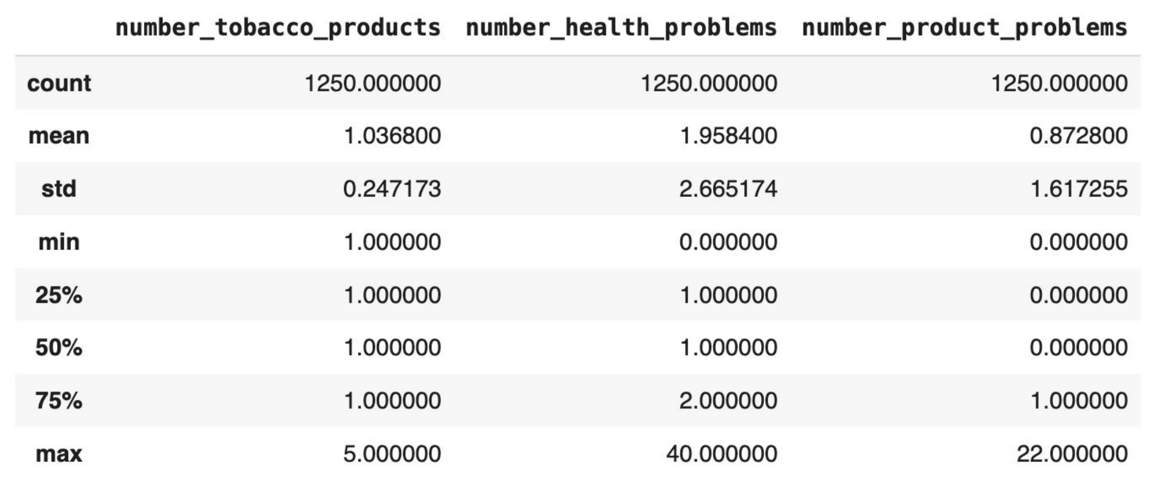
Excel	Data Cleaning, Validation checks
Tableau	Exploratory Data Analysis & Visualizations
Python	Data Preprocessing, Correlation Analysis, Models Training & Evaluation, Forecasting
LangChain , Openai, Llama_Index	Python Libraries for LLM Applications

Overall Insights



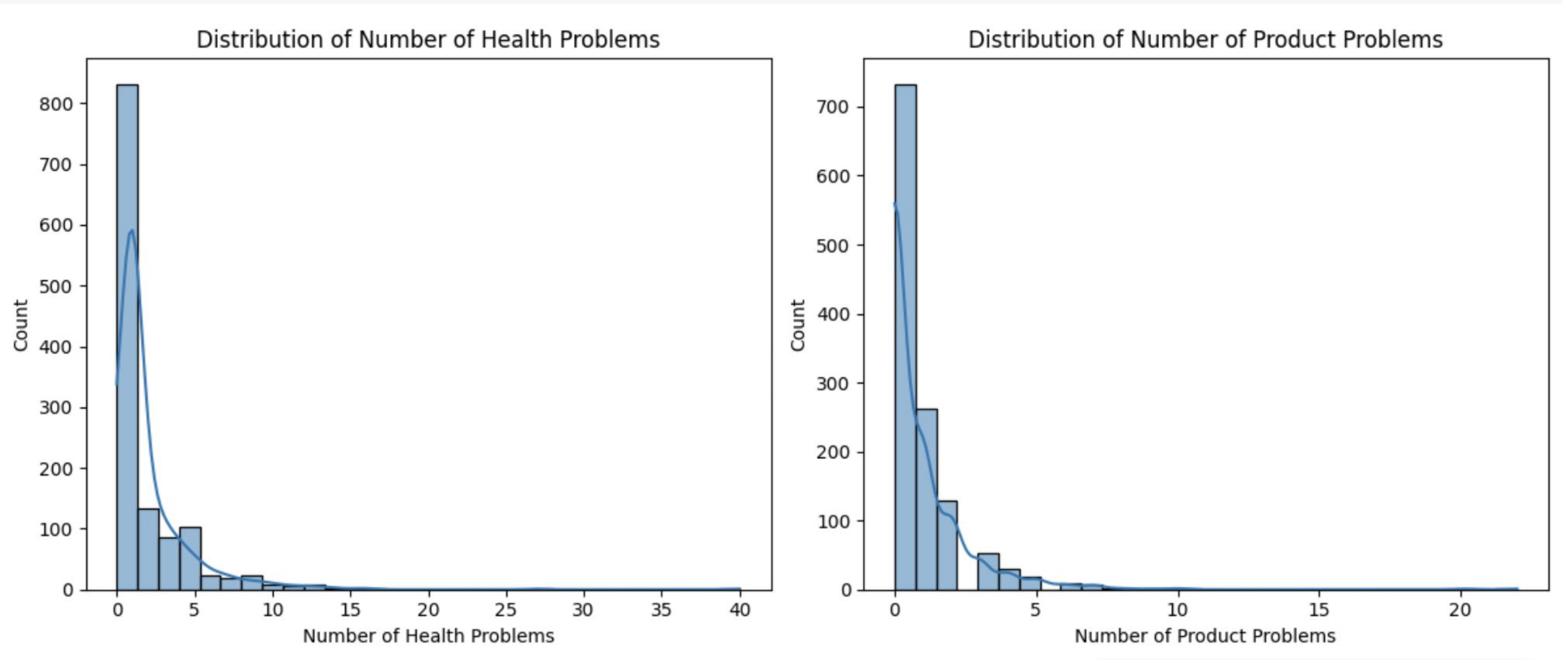
Total Reports	Avg. Health Problems	Avg. #Product	Avg. Product Problems
1,250	2.0	1	1

Descriptive Stats



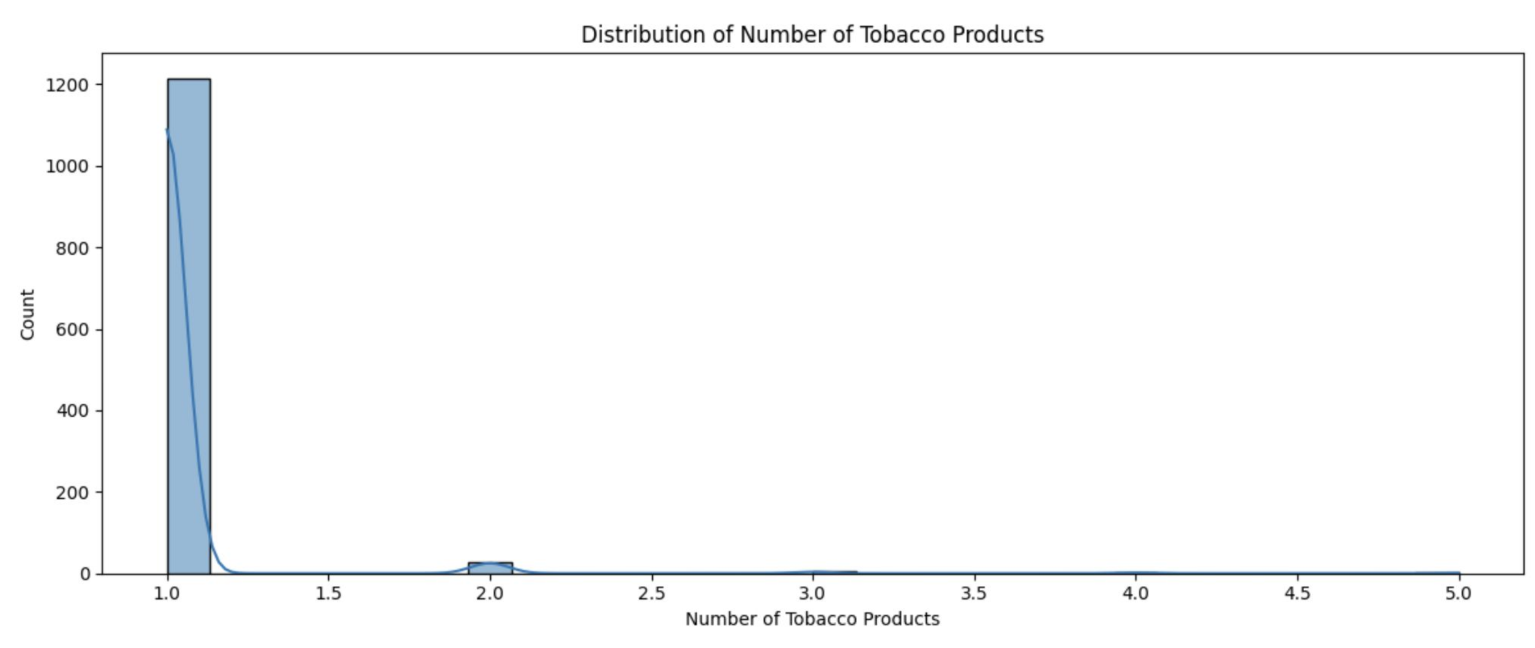
	number_tobacco_products	number_health_problems	number_product_problems
count	1250.000000	1250.000000	1250.000000
mean	1.036800	1.958400	0.872800
std	0.247173	2.665174	1.617255
min	1.000000	0.000000	0.000000
25%	1.000000	1.000000	0.000000
50%	1.000000	1.000000	0.000000
75%	1.000000	2.000000	1.000000
max	5.000000	40.000000	22.000000

Data Distributions



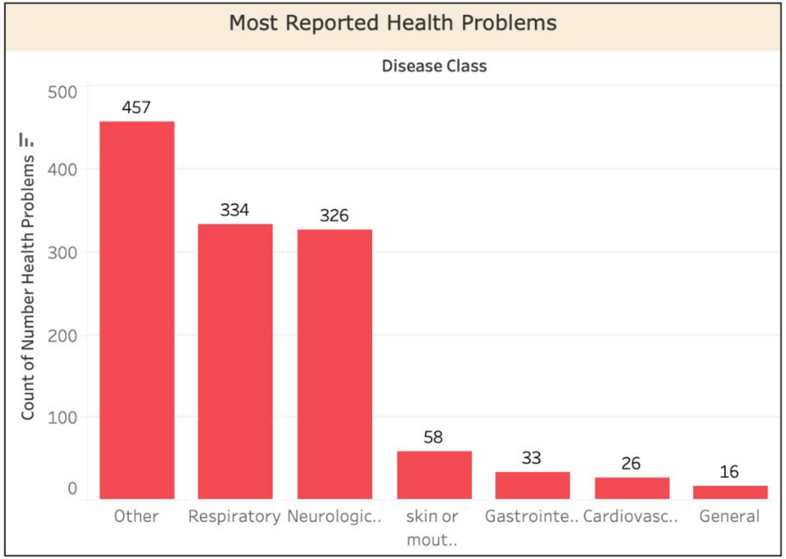
Right Skewed

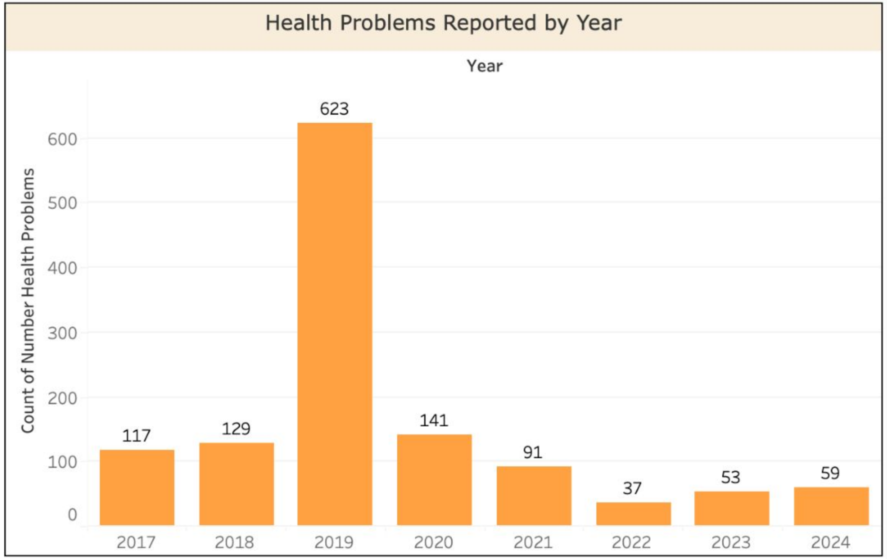
Products Distributions



Right Skewed

Health Insights



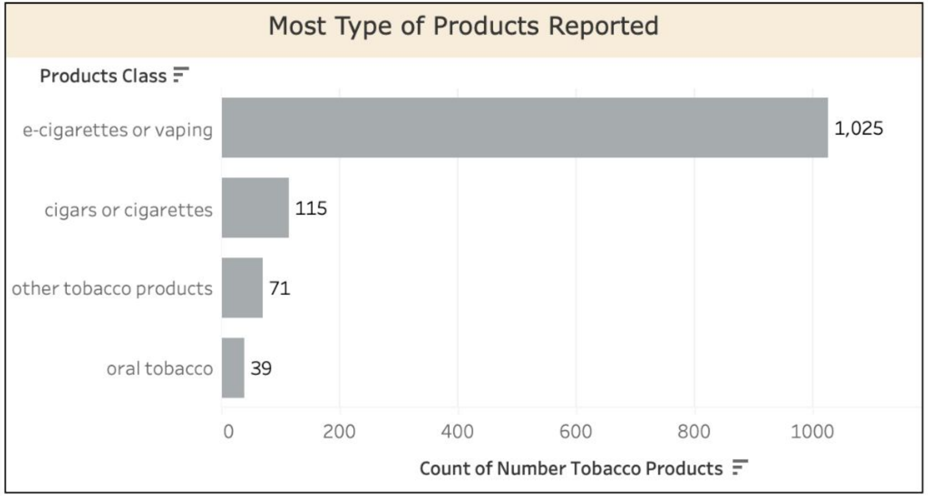


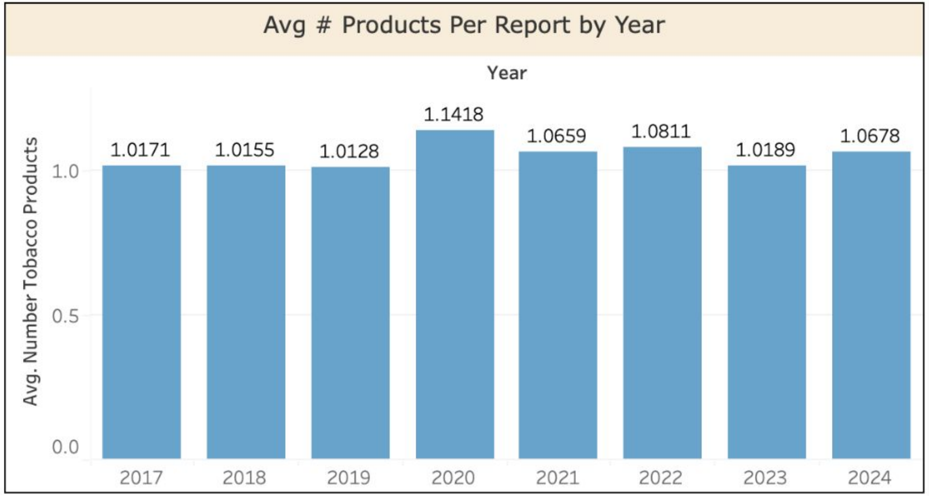
['breath', 'cough', 'lung', 'wheez', 'asthma','throat','dyspnea','respiratory','pulmonary','Nasal','Pneumonia']-Respiratory

['seizure', 'headache', 'dizz', 'brain','migraine','Unconsciousness','anger','confusion','fatigue']-Neurological

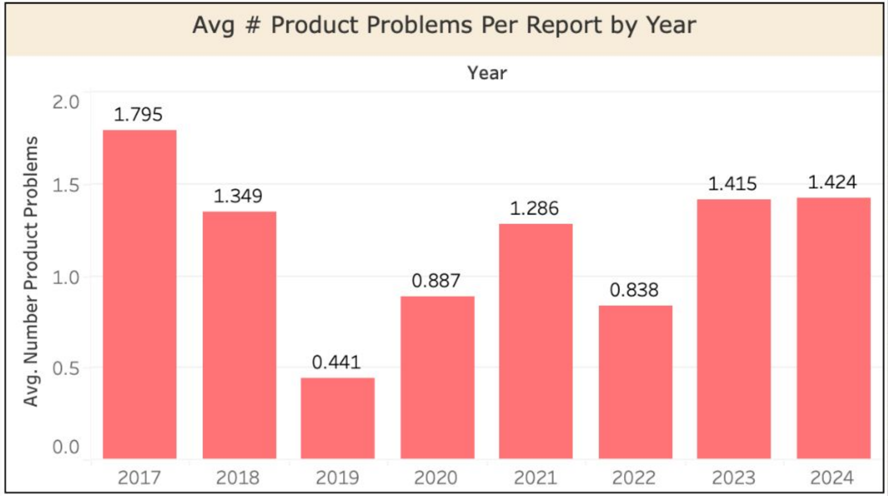
['heart', 'cardiac', 'chest','stroke','blood pressure','pulse']-Cardiac

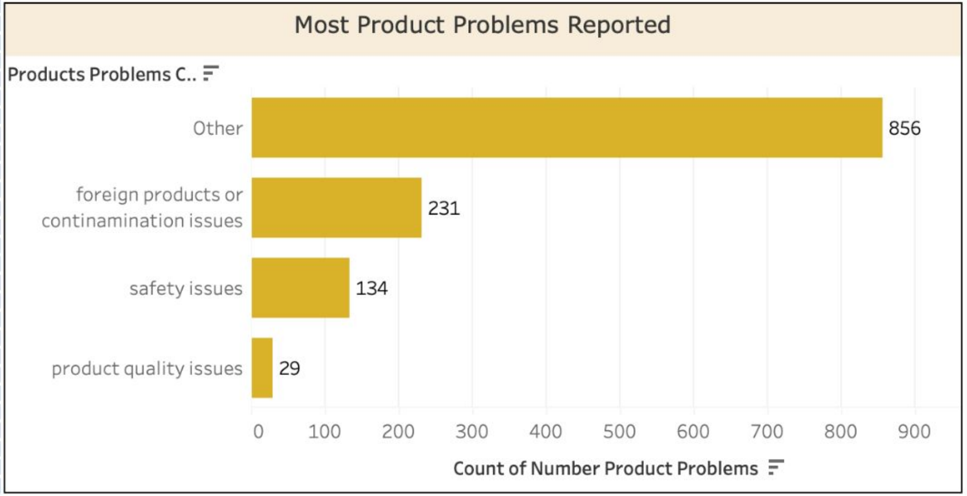
Product Usage Insights



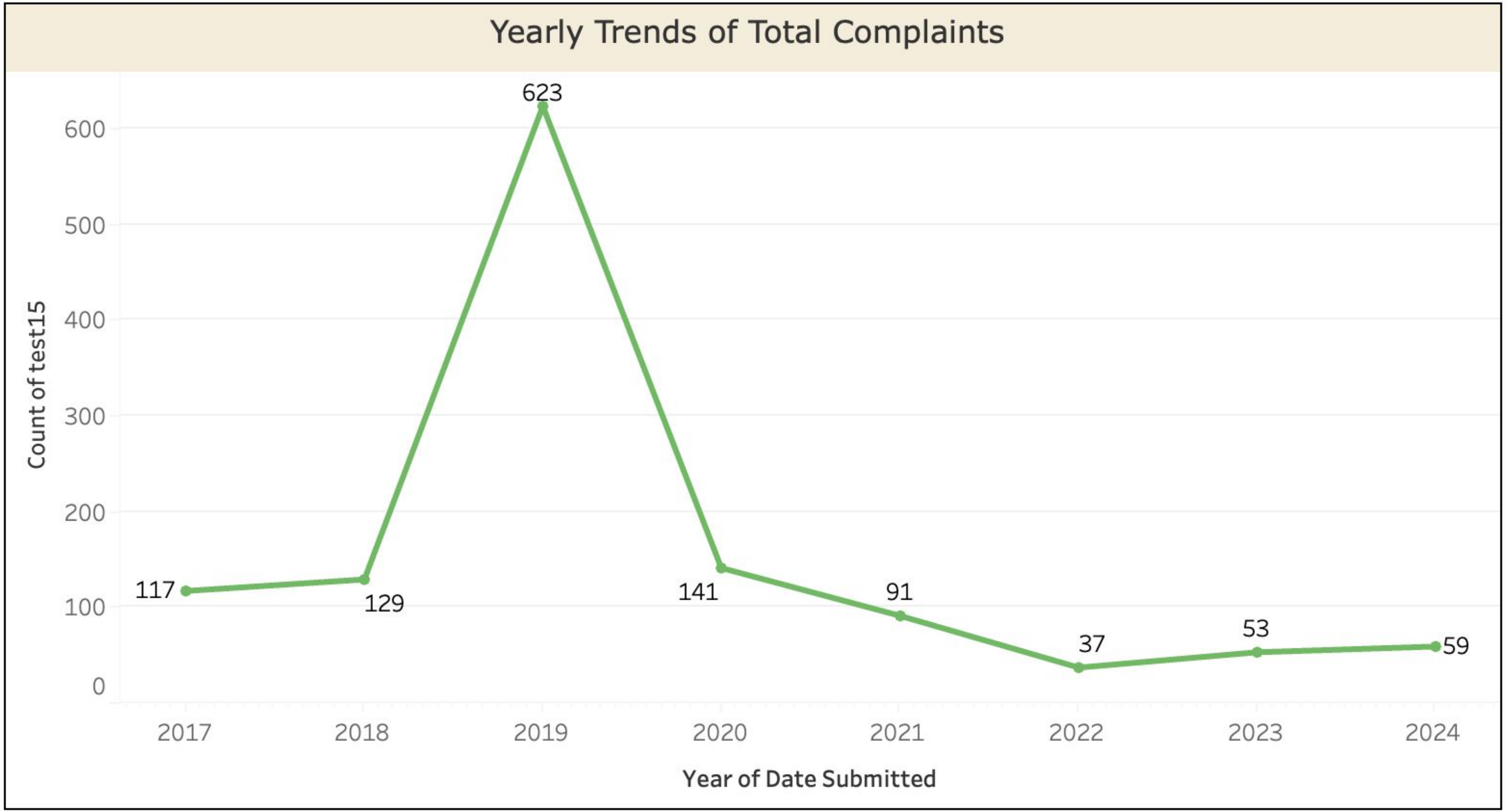


Product Problems Insights

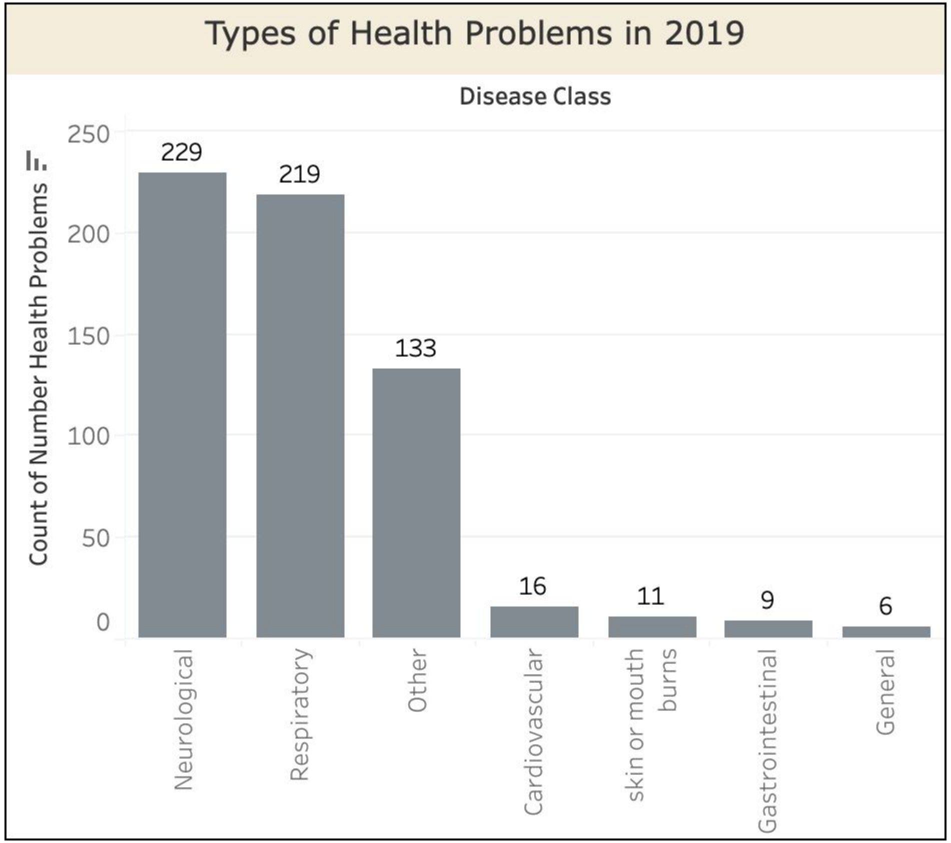




Yearly Trends of Complaints



Deep Dive Analysis of Year 2019



EVALI Outbreak: A national health crisis tied to vaping caused a surge in respiratory and neurological complaints.

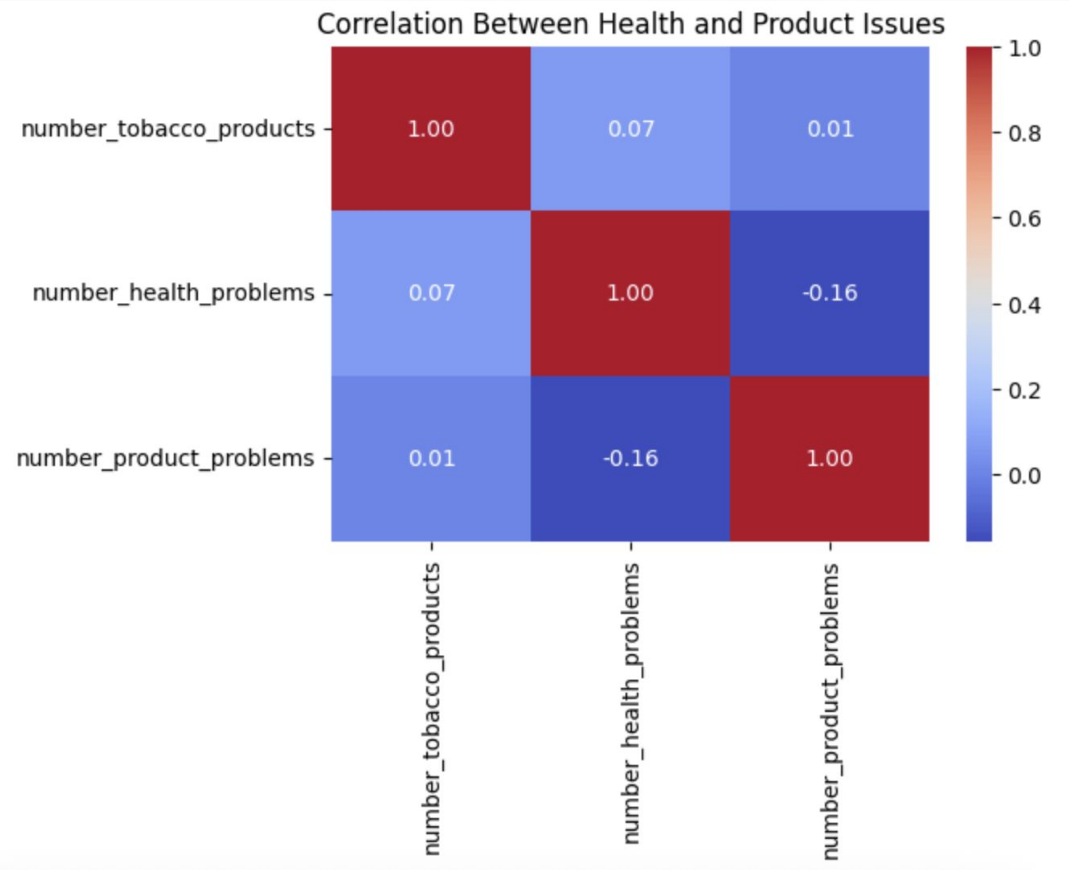
Data Confirmation: Dataset shows a dramatic rise in health issues in 2019, especially respiratory (219) and neurological (229) complaints.

Media & Awareness: Extensive media coverage and public concern led to increased complaint reporting to the FDA.

Not Product Failures: Statistical tests show 2019 complaints were health-driven, not due to more product defects.

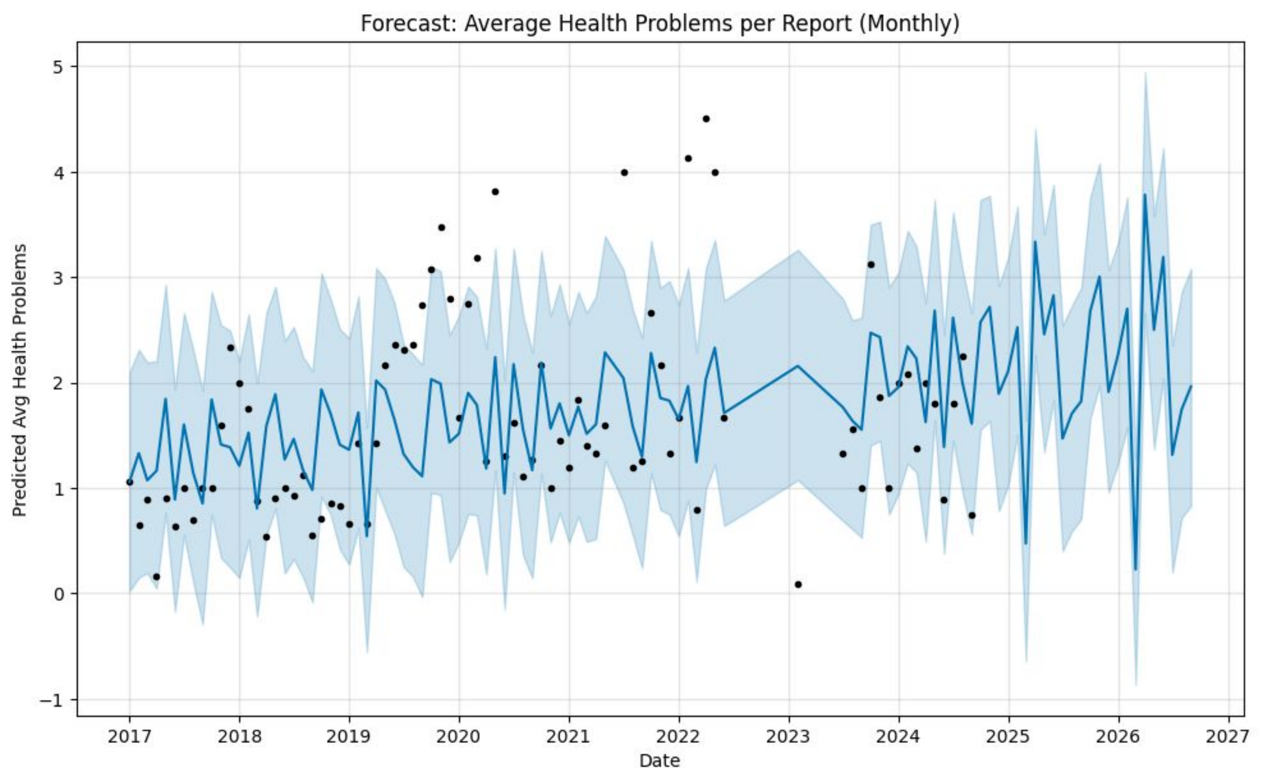
Regulatory Lag: The lack of early regulation allowed risky products to spread before stronger oversight reduced complaints post-2019.

Correlation Analysis



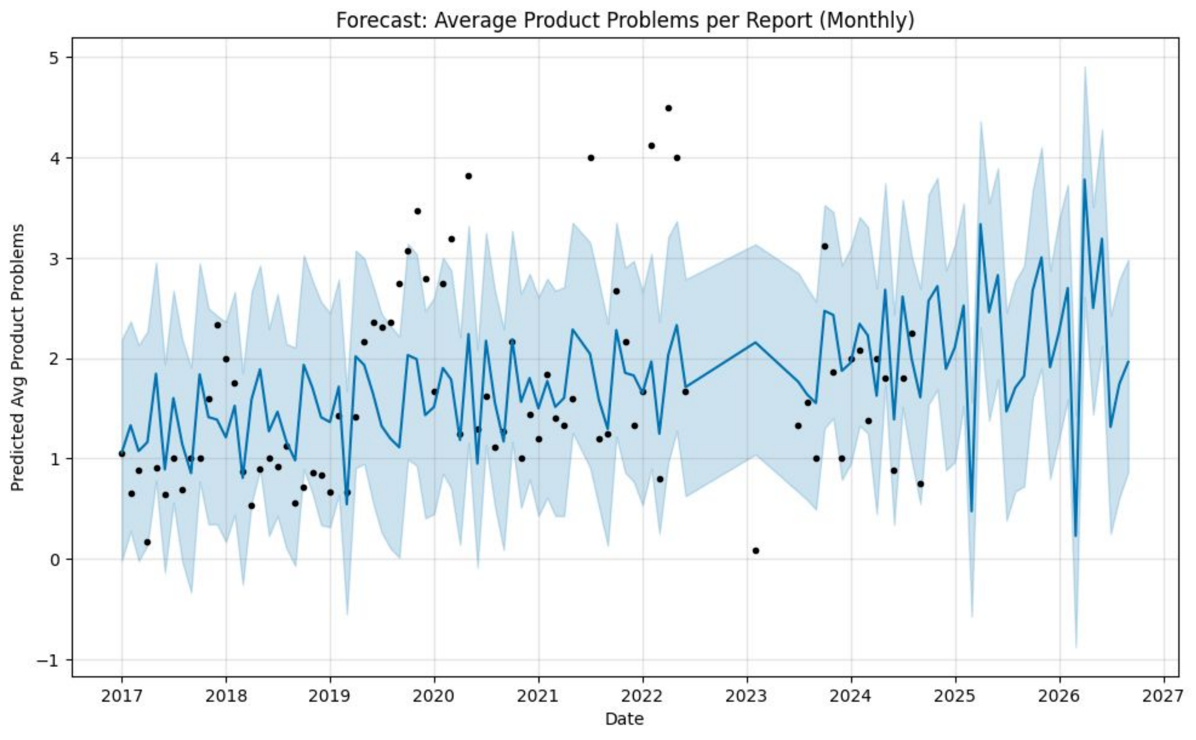
- Health problems and product issues are weakly negatively correlated (−0.16), suggesting they occur independently.
- More tobacco products per report does not strongly increase health complaints (correlation = 0.07).
- Product problems are not linked to the number of tobacco products (correlation = 0.01), indicating no relationship.

Forecasting # Health Problems



- The overall trend shows an increase in the average number of health problems reported per tobacco incident, rising from approximately 1.0 in 2017 to forecasted values of around 2.0-3.0 by 2025-2026.
- A major peak reaching approximately 4.5 health problems per report in 2022
- Projected extreme peaks approaching 5.0 in the 2026 forecast

Forecasting # Product Problems



Parallel Trend with Health Problems: The exact match between this graph and the health problems graph suggests a strong correlation between reported product problems and health problems.
This indicates that when users report tobacco product issues, they typically report corresponding health issues in similar numbers.
Increasing Complexity: The upward trend from 2017 to 2026 suggests reports are becoming more detailed over time, with users identifying more specific product problems per report.
Similar Peak Periods: The significant spikes around 2020 and 2022 that approached 4-4.5 problems per report appear in both datasets, suggesting specific incidents or product issues that generated multiple problems and health effects simultaneously.

Random Forest Classifier Model

50%

Classification Accuracy

Predict All Diseases

98%

Binary Classification Accuracy

Predict Respiratory Disease

90%

Multi- Class Accuracy

Predict All Diseases

Before Label Classification

After Label Classification

DSC 550

RAG ChatBot

What?

- A question-answering AI assistant built using Retrieval-Augmented Generation (RAG)
- Allows users to ask natural-language questions about the complaint data (e.g., *“What were the top health issues in 2019?”*)
- The system retrieves relevant complaint records and generates accurate, context-aware answers

Why?

- The original dataset is large and unstructured, making it hard to explore manually
- Enables researchers, health professionals, and the public to access insights without technical skills

How?

- Combines a vector store (embeddings) for document retrieval with a language model (LLM) for response generation
- Uses tools like LangChain + OpenAI or HuggingFace to build the retrieval and generation pipeline
- Data is chunked, embedded, and indexed — then queried in real time based on user input

Conclusion & Future Improvements

- Examined 1,250 tobacco product reports capturing health problems, product issues, and impact patterns.
- 2019 health complaint spike was driven by serious health effects, not product failures
- Identified rising trend in reported health and product problems (1.0 → ~2.5 per report from 2017-2025)
- Successfully implemented Random Forest classification model to predict respiratory conditions
- Forecasting indicates increasing complexity of tobacco-related health impacts

Recommendations

- Implement an early warning system for product problem spikes.
- Expand the dataset to include more recent and broader complaint records
- Enhance the classification model with deeper disease subcategories
- Develop a full-scale RAG chatbot for public and professional use
- Integrate geographic analysis to identify regional patterns

Thank you