

**Tech Saksham**

**Capstone Project Report**

**ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING FUNDAMENTALS**

**“Heart Disease Prediction using Logistic Regression”**

**“UNIVERSITY COLLEGE OF ENGINEERING (BIT CAMPUS) TIRUCHIRAPALLI”**

|  |  |
| --- | --- |
| **NM ID** | **NAME** |
| au810021214011 | CHITRA C |

|  |  |
| --- | --- |
|  |  |
|  | Trainer Name  P.Raja |
|  | Master Trainer |

**ABSTRACT**

This machine learning Logistic Regression model endeavors to create a predictive framework for identifying individuals prone to heart disease. The primary objective of this research is to determine the percentage of patients at elevated risk of developing cardiovascular disease (CVD). Leveraging a dataset containing pertinent medical and demographic data, Logistic Regression, a sophisticated statistical method adept at binary classification tasks, will be utilized to gauge the likelihood of heart disease occurrence in an individual. Key stages of the project encompass data preprocessing, feature selection, model training, and assessment. Thorough preprocessing of the dataset will be conducted, entailing the removal of missing values, encoding categorical variables, and scaling numerical features as necessary. The resultant model holds promise in aiding healthcare providers with timely intervention and personalized risk evaluation, ultimately leading to reduced rates of CVD-related morbidity and mortality.

**INDEX**

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Table of Contents** | **Page No.** |
| 1 | Chapter 1: Introduction | 1 |
| 2 | Chapter 2: Services and Tools Required | 3 |
| 3 | Chapter 3: Project Architecture | 4 |
| 4 | Chapter 4: Modeling and Project Outcome | 6 |
| 5 | Chapter 5: Project Result | 21 |
| 6 | Conclusion | 22 |
| 7 | Future Scope | 23 |
| 8 | References | 24 |
| 9 | Links | 25 |

**CHAPTER 1**

**INTRODUCTION**

* 1. **Problem Statement**

Heart disease remains a leading cause of mortality globally, necessitating effective predictive models to identify at-risk individuals for timely intervention. While various risk factors are known, there is a need for a robust predictive framework to accurately assess the likelihood of heart disease occurrence. The goal of this project is to develop a logistic regression model leveraging clinical and demographic features to predict the probability of heart disease presence in individuals.

**1.2. Proposed Solution**

The proposed logistic regression model for heart disease prediction, implemented in Google Colab, encompasses data preprocessing, feature selection, and model building. By leveraging logistic regression, it becomes feasible to evaluate an individual's likelihood of developing cardiovascular disease. Such a model serves as a valuable tool for healthcare professionals, aiding in the identification of high-risk patients. With this insight, tailored interventions and lifestyle modifications can be initiated to mitigate the risk of cardiovascular disease-related outcomes.

**1.3. Feature**

1. **Data Preprocessing:** After data collection, preprocess the data. This involves handling missing values, scaling numerical features, and encoding categorical variables
2. **Model Training:** Train a logistic regression model using the training data. During training, the model learns the coefficients for each feature, which determine their influence on the likelihood of having heart disease**.**
3. **Model Evaluation**: Evaluate the performance of the trained model using the testing data. Common evaluation metrics for binary classification tasks include accuracy, precision, recall, F1 score, and area under the ROC curve (AUC-ROC).
4. **Feature Selection:** Relevant features are selected based on their significance in predicting CVD risk, using techniques such as correlation analysis and feature importance ranking**.**
5. **Model Deployment**: Once satisfied with the model's performance, deploy it for making predictions on new data. This could involve integrating the model into a software application or system where it can be used to predict heart disease for new patients

**1.4. Advantages**

* One of the primary benefits of utilizing Logistic Regression for heart disease prediction is its interpretability. Unlike more complicated machine learning models, Logistic Regression explains the association between input data and the chance of heart disease.
* Logistic Regression is a computationally efficient and simple algorithm, making it ideal for situations with limited computational resources or rapid deployment, like clinical settings.
* Logistic Regression is a model that can effectively use diverse patient data, including demographics, medical history, and diagnostic test results, to improve prediction accuracy.
* Logistic Regression offers probabilistic predictions for heart disease risk stratification, enabling healthcare providers to prioritize interventions for patients with the highest predicted risk.

**1.5. Scope**

The outlook for heart disease prediction using Logistic Regression looks optimistic. By integrating additional risk factors and leveraging advanced machine learning methods, there's potential for the development of more sophisticated prediction models. This could lead to improved accuracy and reliability in identifying individuals at risk of cardiovascular disease (CVD). Moreover, the emergence of personalized medicine is poised to revolutionize the management of heart disease. Telemedicine and remote patient monitoring technologies allow Logistic Regression models to estimate heart disease risk, notify healthcare professionals early, permit prompt interventions, and reduce hospital readmissions.

**CHAPTER 2**

**SERVICES AND TOOLS REQUIRED**

**2.1 Services Used**

**Data Collection and Storage Services**: Access and use healthcare databases like MIMIC-III, NHANES, or the Framingham Heart Study dataset, which contain anonymised patient data related to heart disease or may also use cloud storage platforms like Github or Kaggle.

**Data Processing Services**: Python packages like Pandas, NumPy, and scikit-learn are used for data cleaning, preprocessing, and feature engineering and Matplotlib with Seaborn to visualize data distributions, correlations, and trends.

**Machine Learning Services**: Machine learning frameworks such as google collab, jupyter, scikit-learn, or PyTorch to implement and train Logistic Regression models on a dataset.

**2.2 Tools and Software used**

* **Data collection and storage:**

**Databases:** Github and Kaggle are used and the heart disease prediction dataset are downloaded.

* **Data Processing:**

Python libraries like Pandas, NumPy, scikit-learn are used and Matplotlib for

visualization used

* **Model Evaluation and interpretation**:

Google collab software is used to run the program ad the results are predicted.

**CHAPTER 3**

**PROJECT ARCHITECTURE**

**3.1 Architecture**

 **Data Collection > Data Preprocessing >**

** Feature Engineering >** **Model Development >**

** Model Evaluation >**

Here’s a high-level architecture for the project:

**Data Collection**: Collect a broad dataset that includes demographic information, lifestyle factors (such as smoking, diet, and exercise habits), clinical parameters (such as blood pressure, cholesterol levels, and CVD family history), and outcomes (CVD presence or absence).   
**Data preprocessing**: Clean up the dataset by removing missing values, encoding category variables, and normalizing numerical characteristics.   
Exploratory data analysis (EDA): Examine the dataset for trends, correlations, and outliers that could influence CVD risk.   
**Model Development**: Use logistic regression to create a predictive model that assesses the likelihood of CVD occurrence depending on input features.   
**Model Evaluation**: Evaluate the performance of the logistic regression model using appropriate evaluation measures, such as accuracy and precision.

**Interpretation and Insights**: Analyse the logistic regression model coefficients to determine the impact of various risk factors on CVD risk and provide practical conclusions for preventive healthcare.

This architecture provides vital insights into Logistic Regression's prediction skills in identifying individuals at risk of heart disease. Furthermore, the created model may assist healthcare providers in early intervention and tailored risk assessment, ultimately contributing to a reduction in CVD-related morbidity and mortality rates.

**CHAPTER 4**

**MODELING AND PROJECT OUTCOME**

**Datasets**

**1.** The datasets are retrieved from the GitHub platform and analyzed

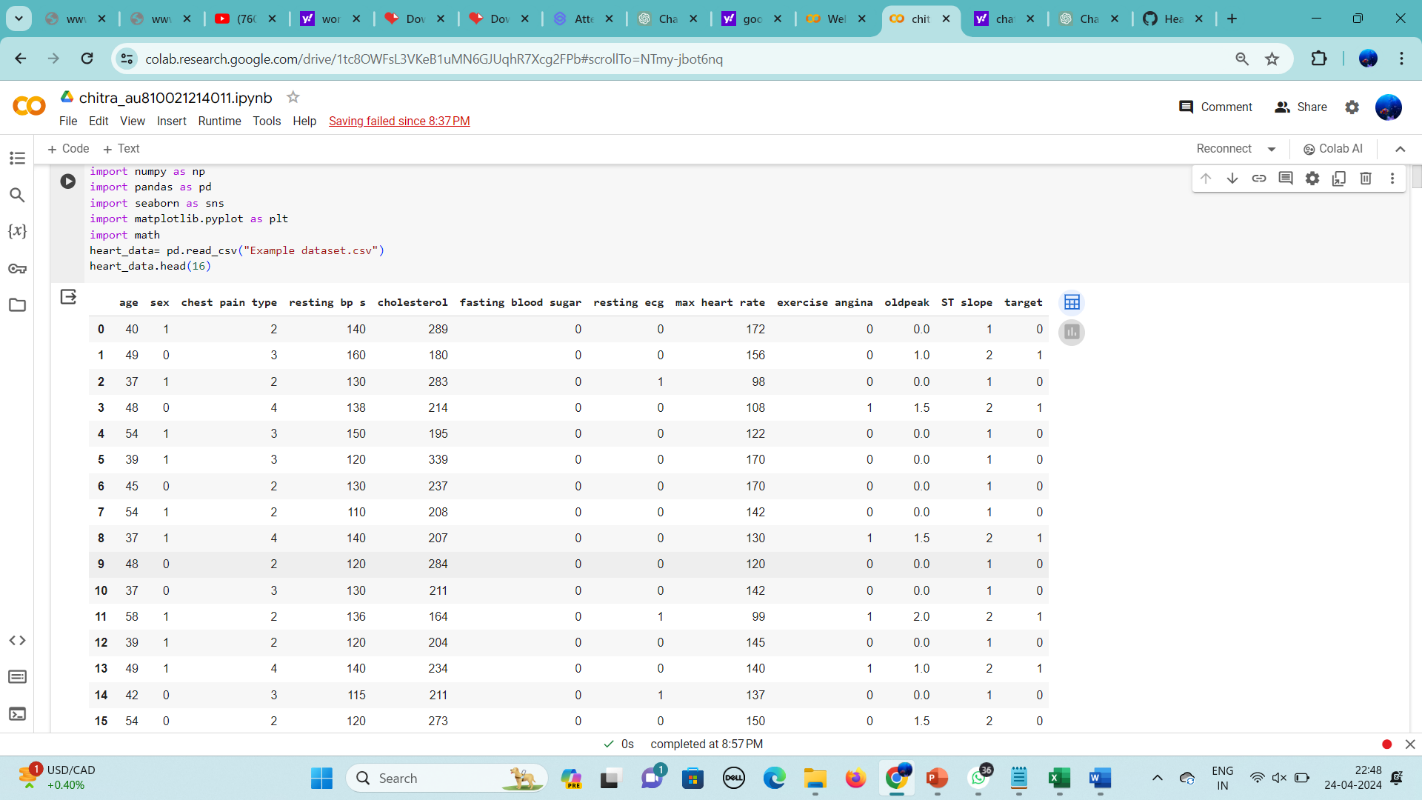


Google collab software is used to run the program and predict the heart disease in the people community

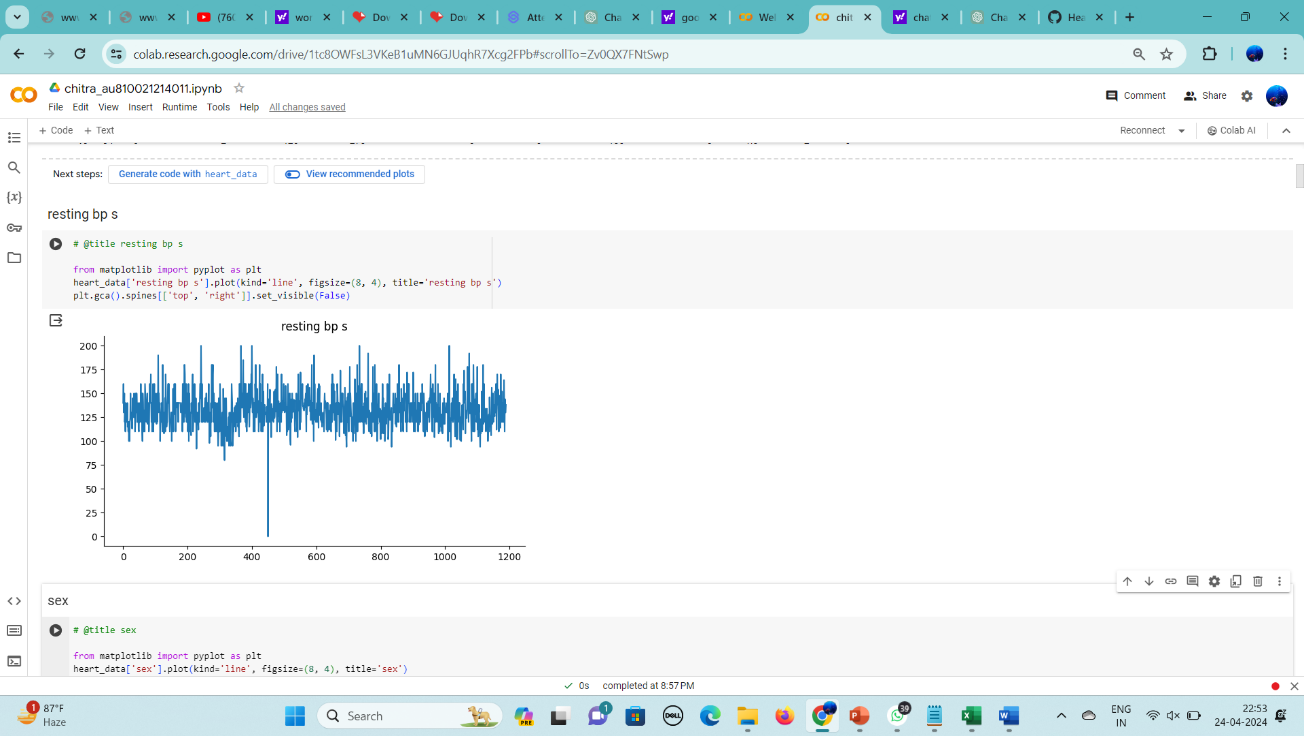
Code:

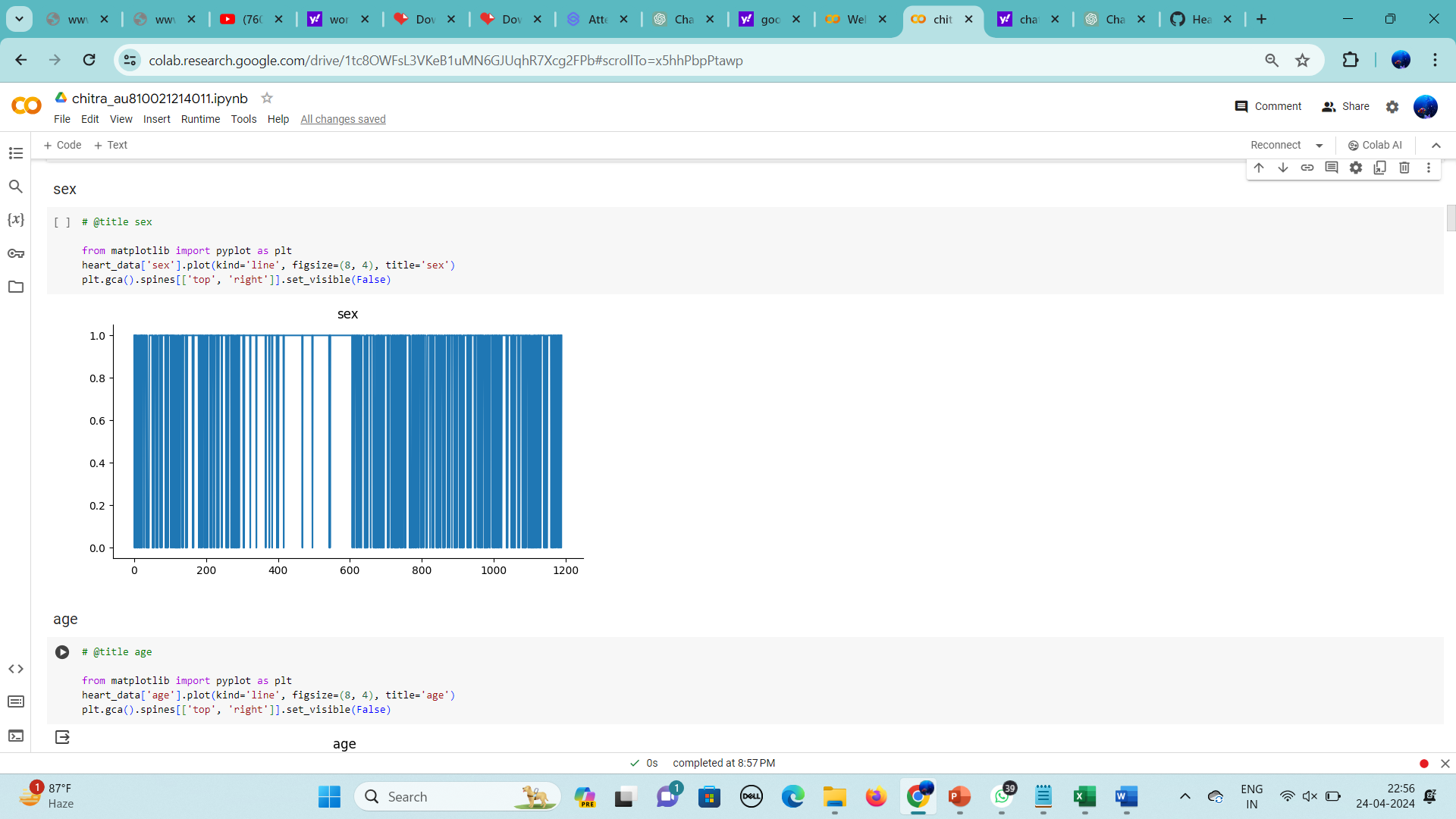
* Data Collection
* Data Preprocessing

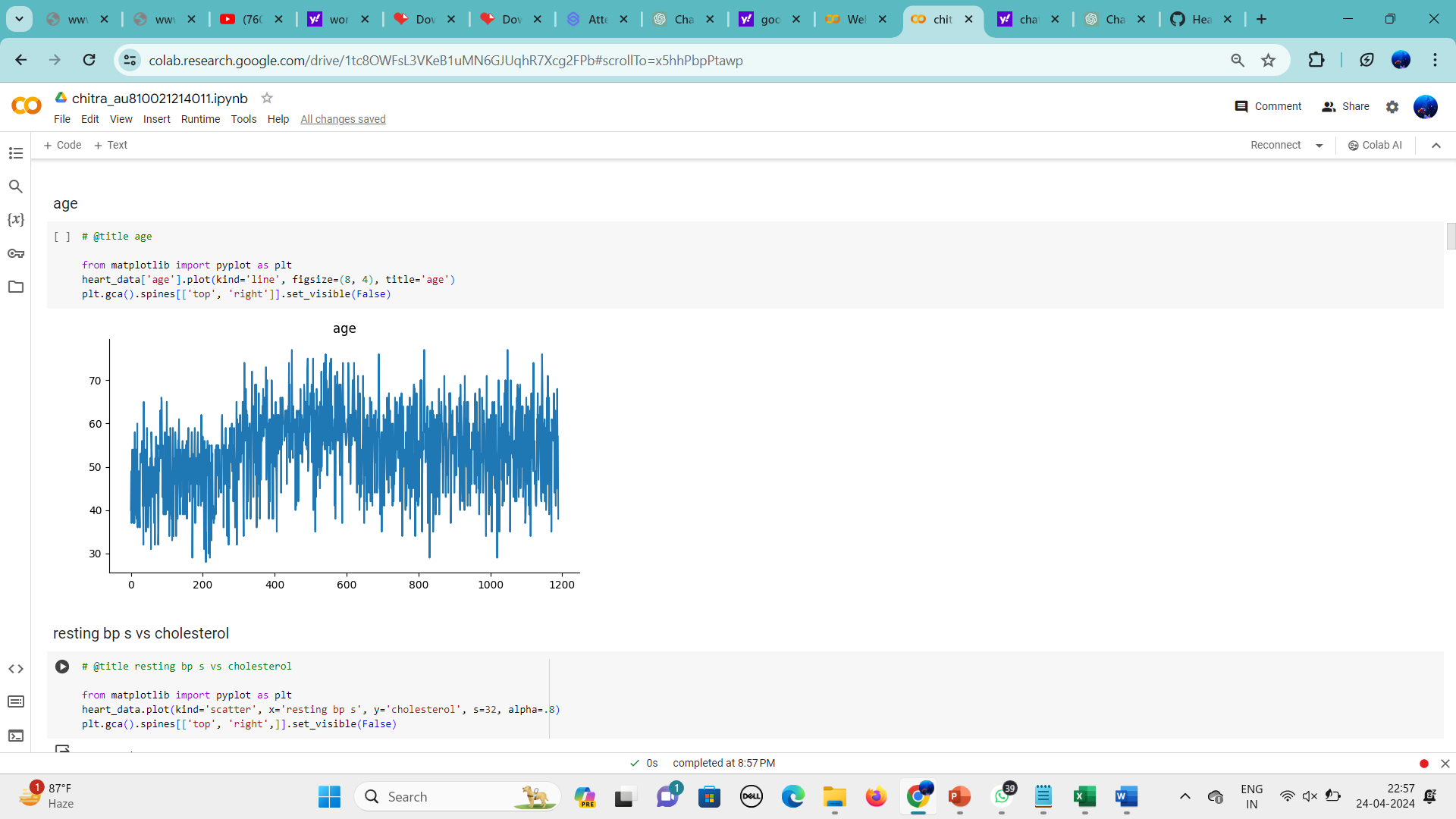
2. Import packages (Numpy, Pandas) for processing the dataset from GitHub.

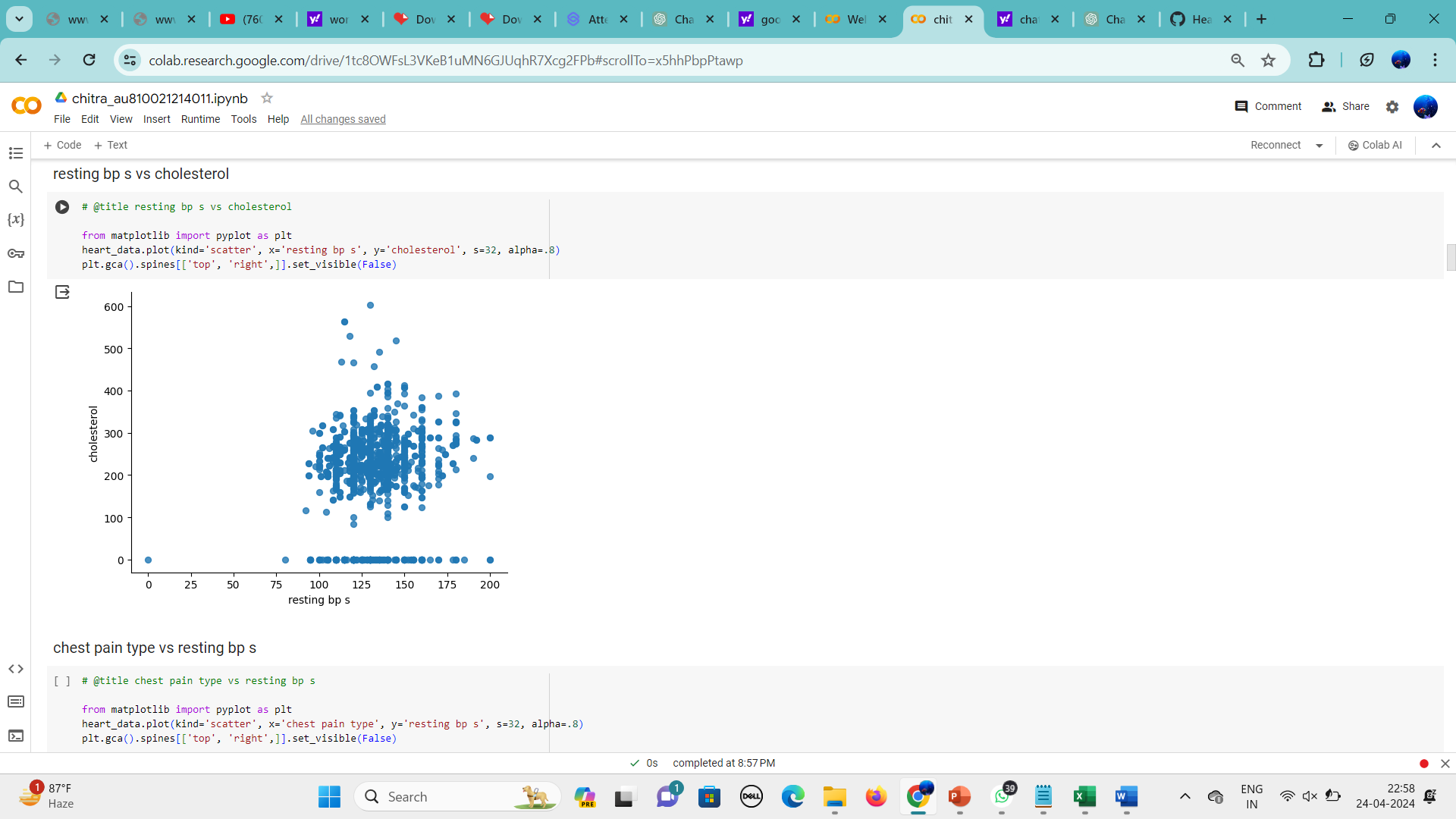


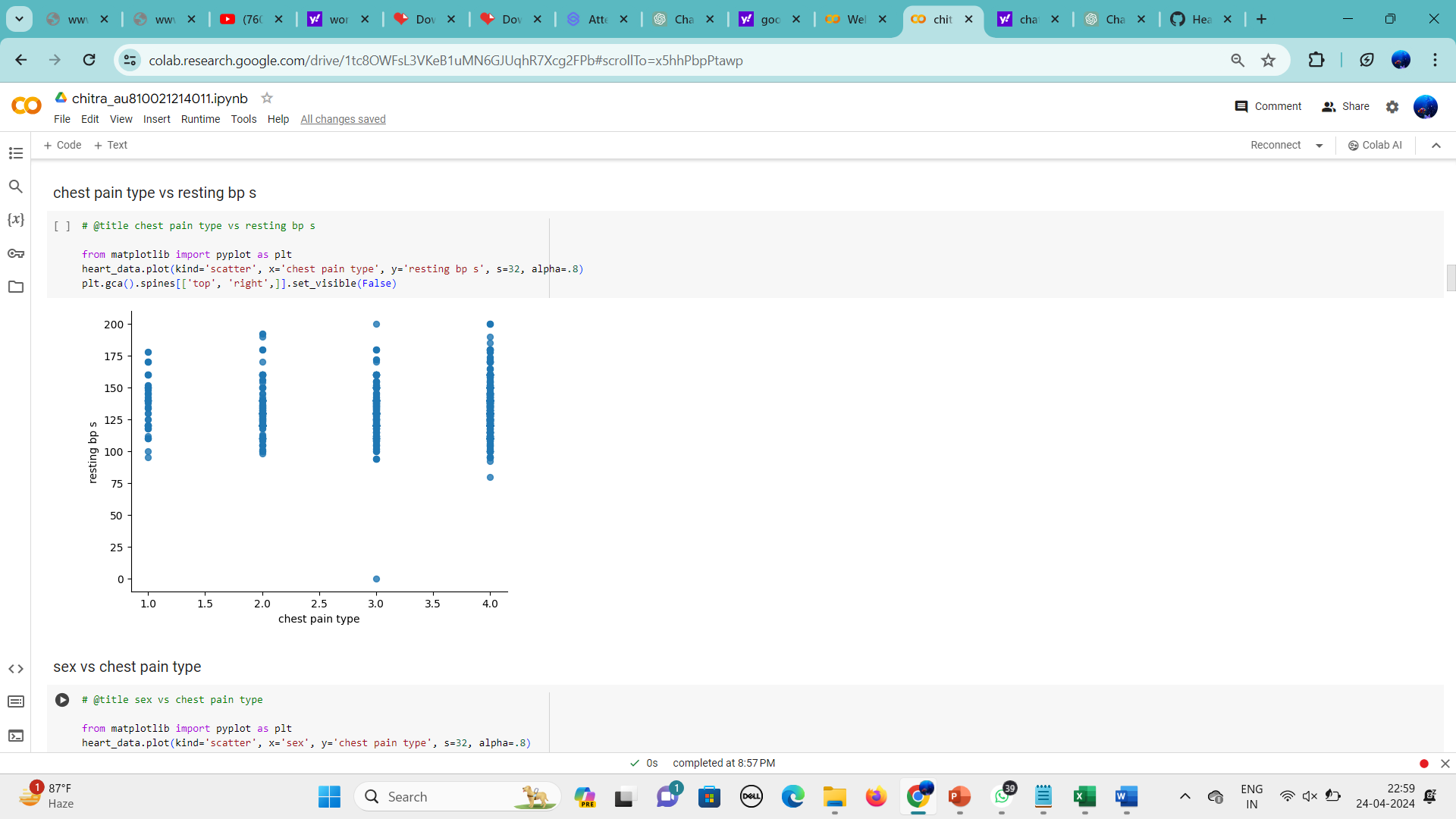
3.Visualizing data

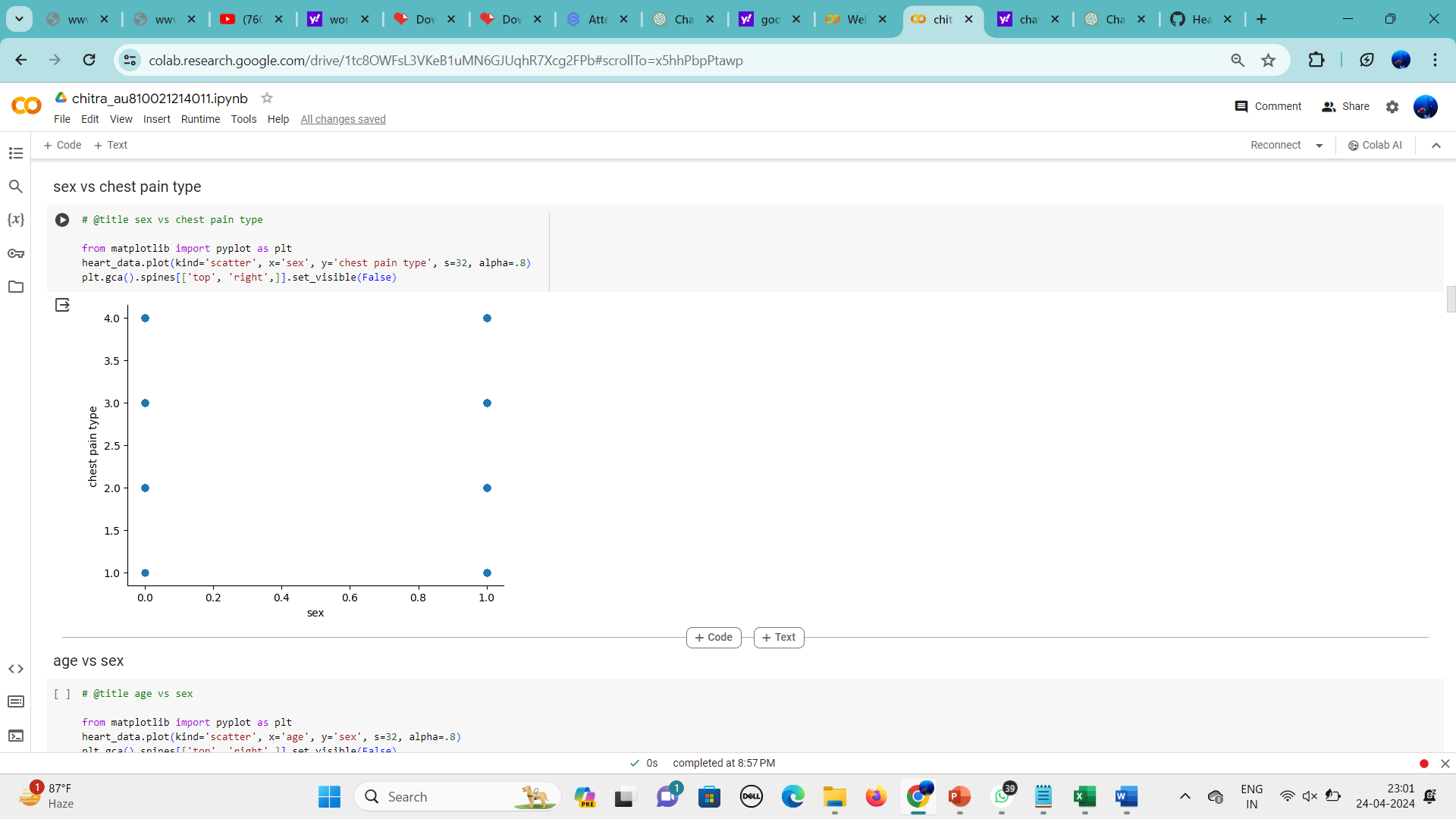


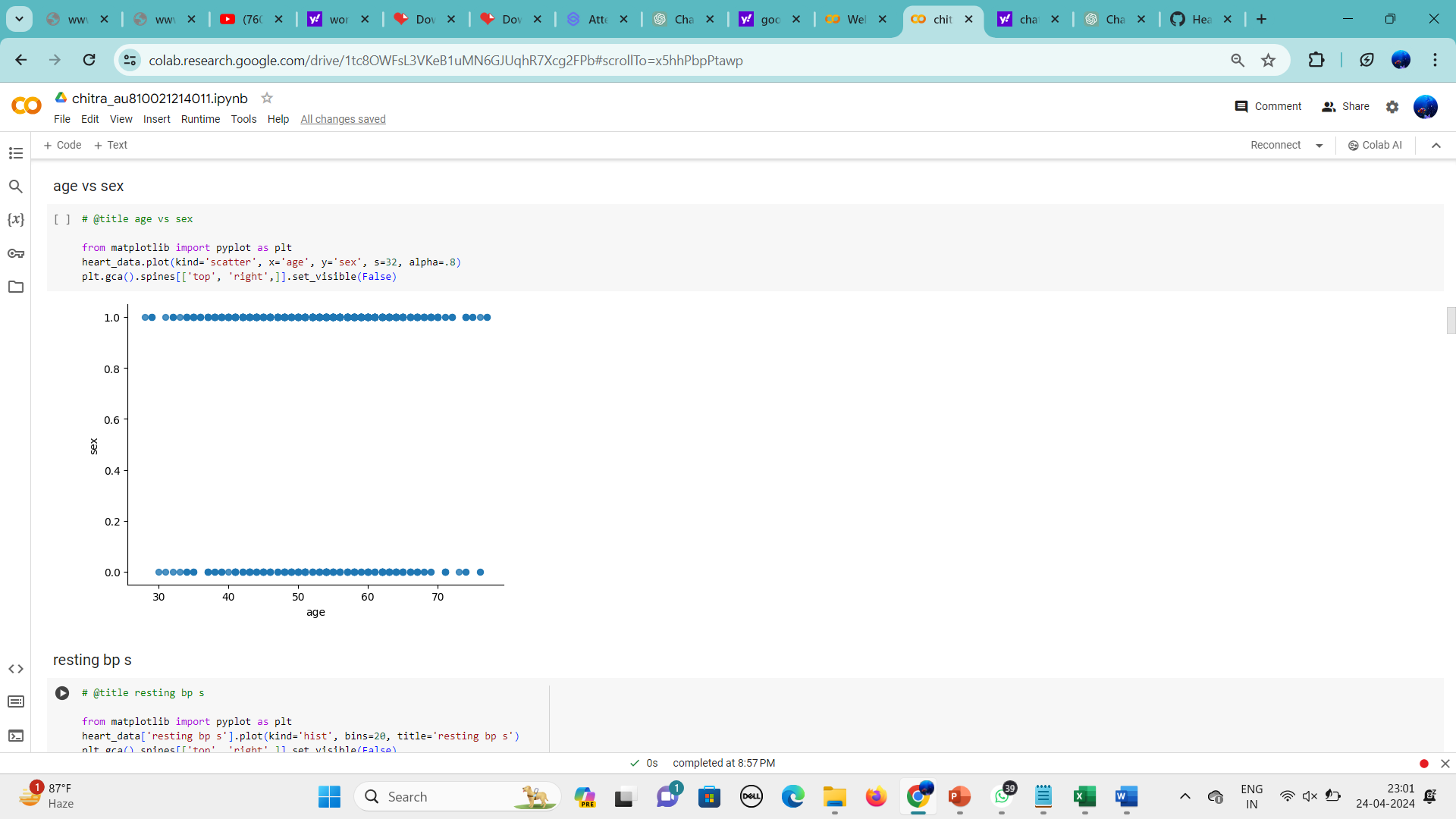




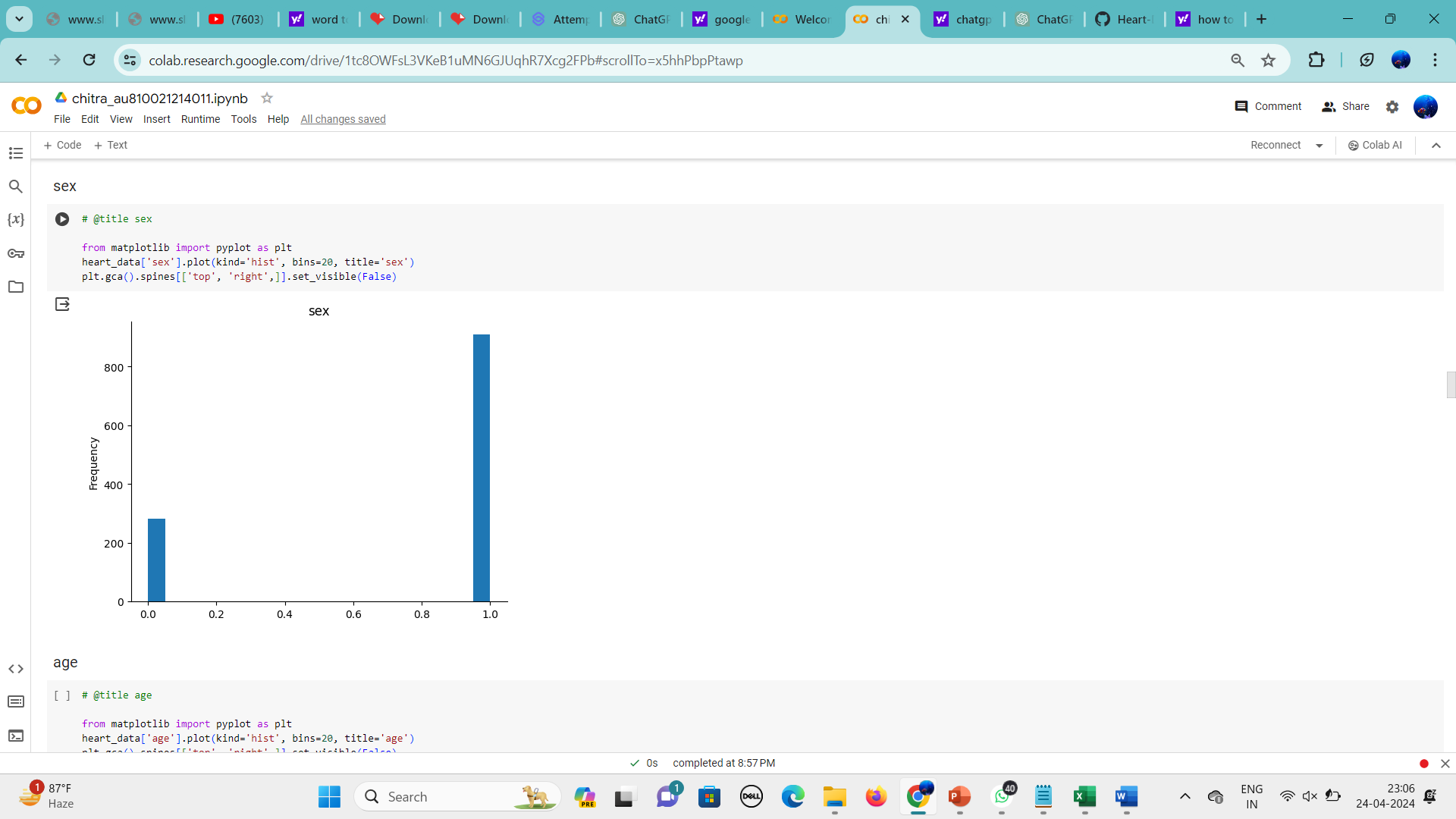


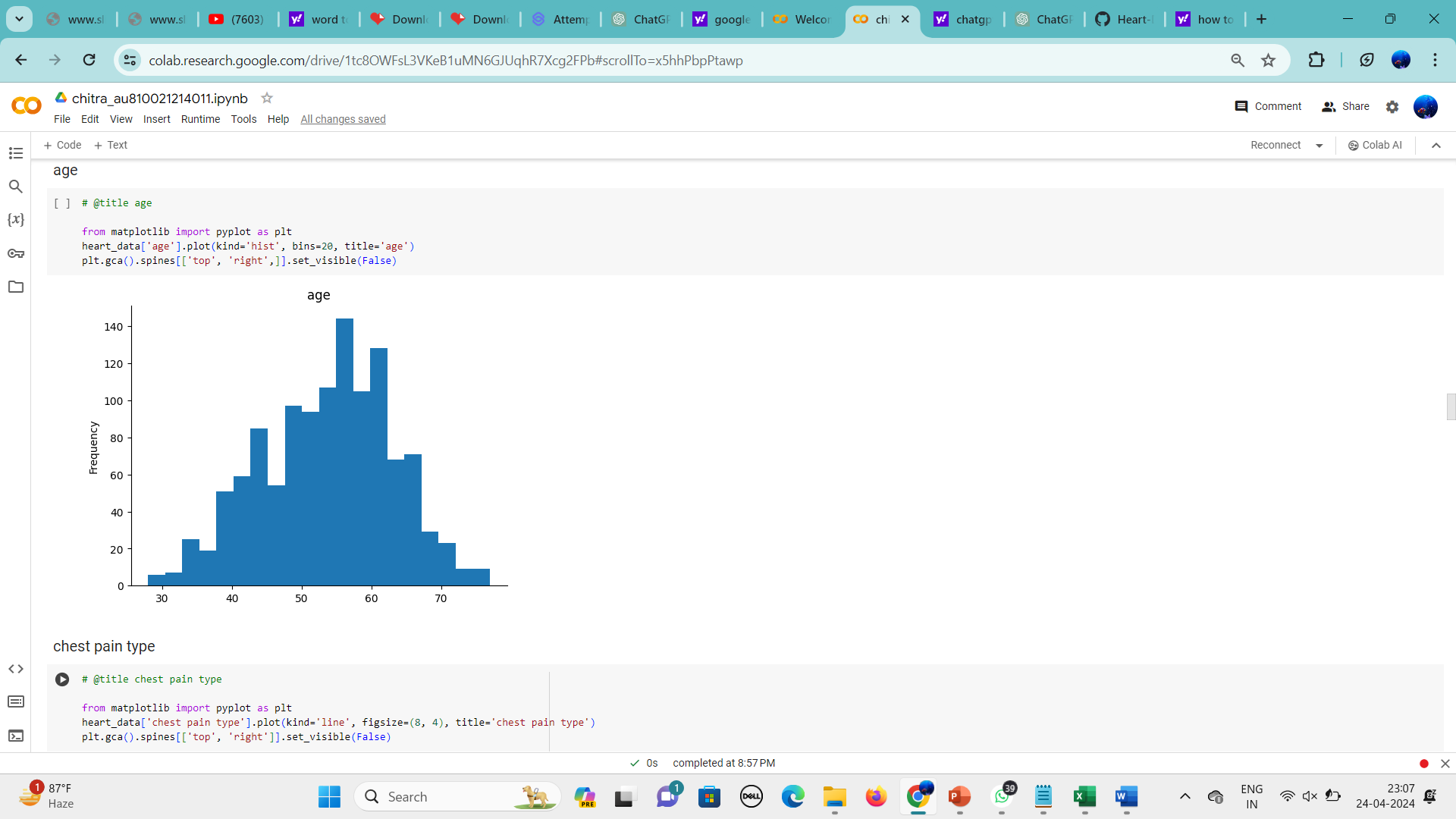


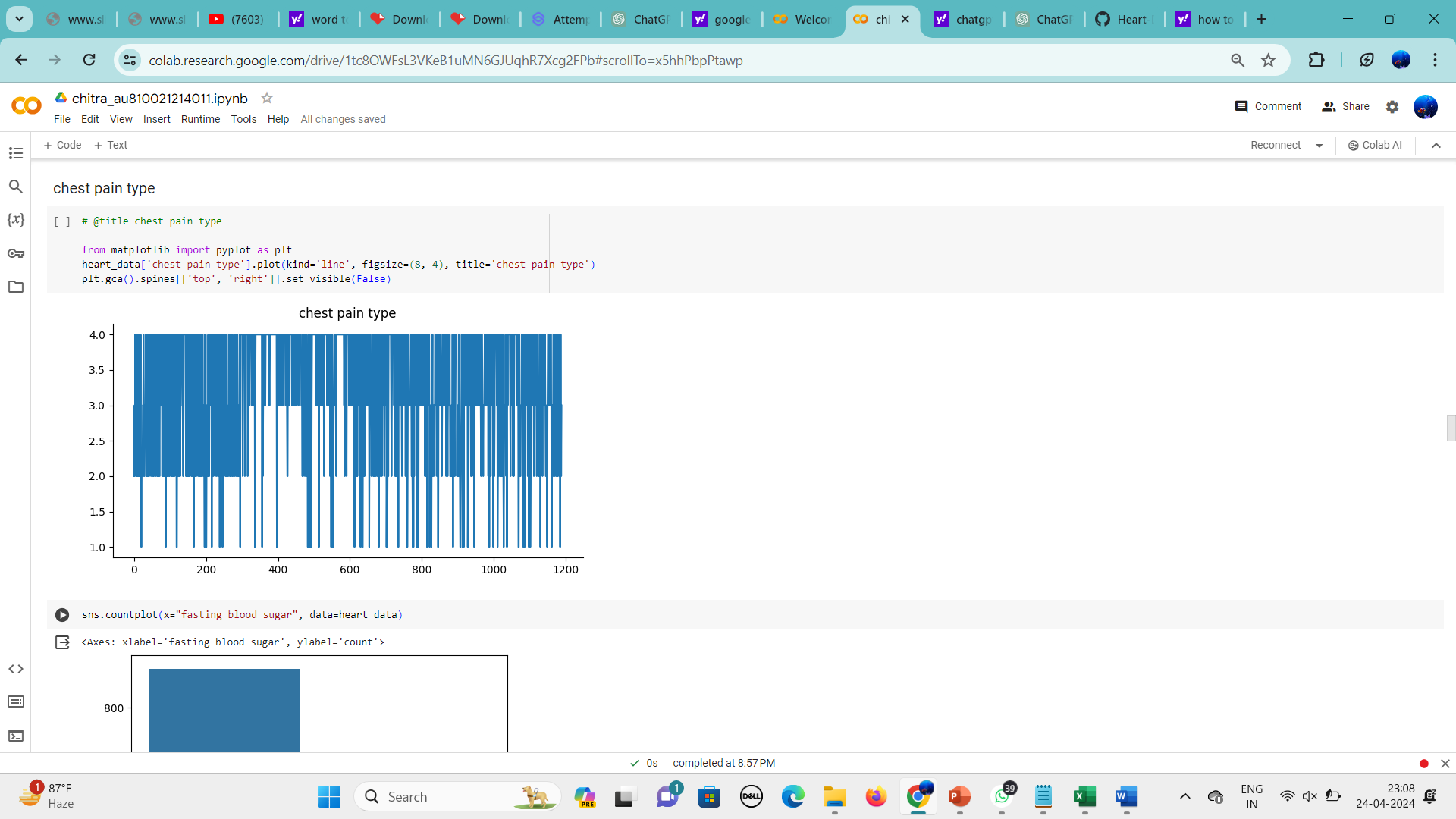


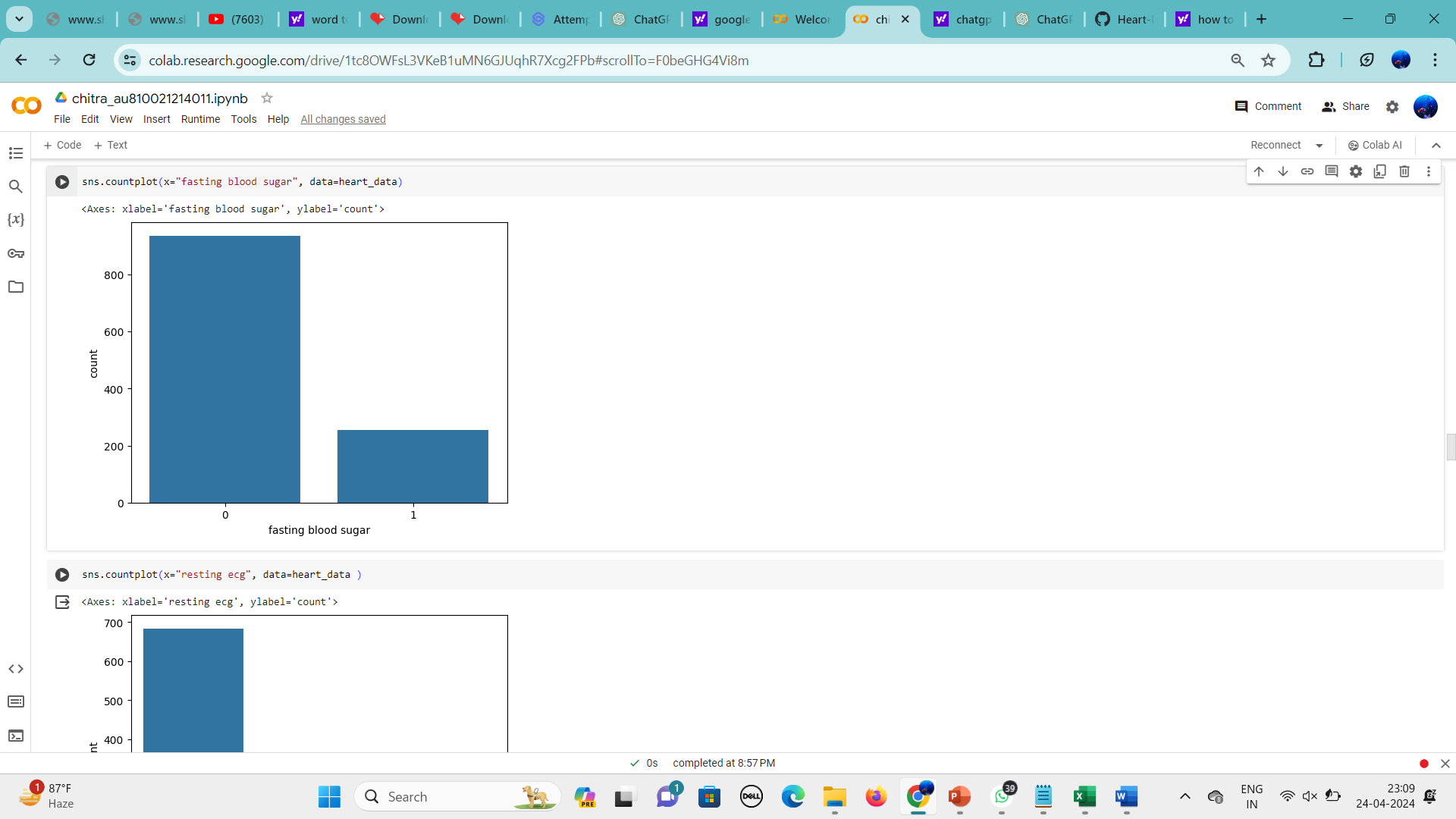
****

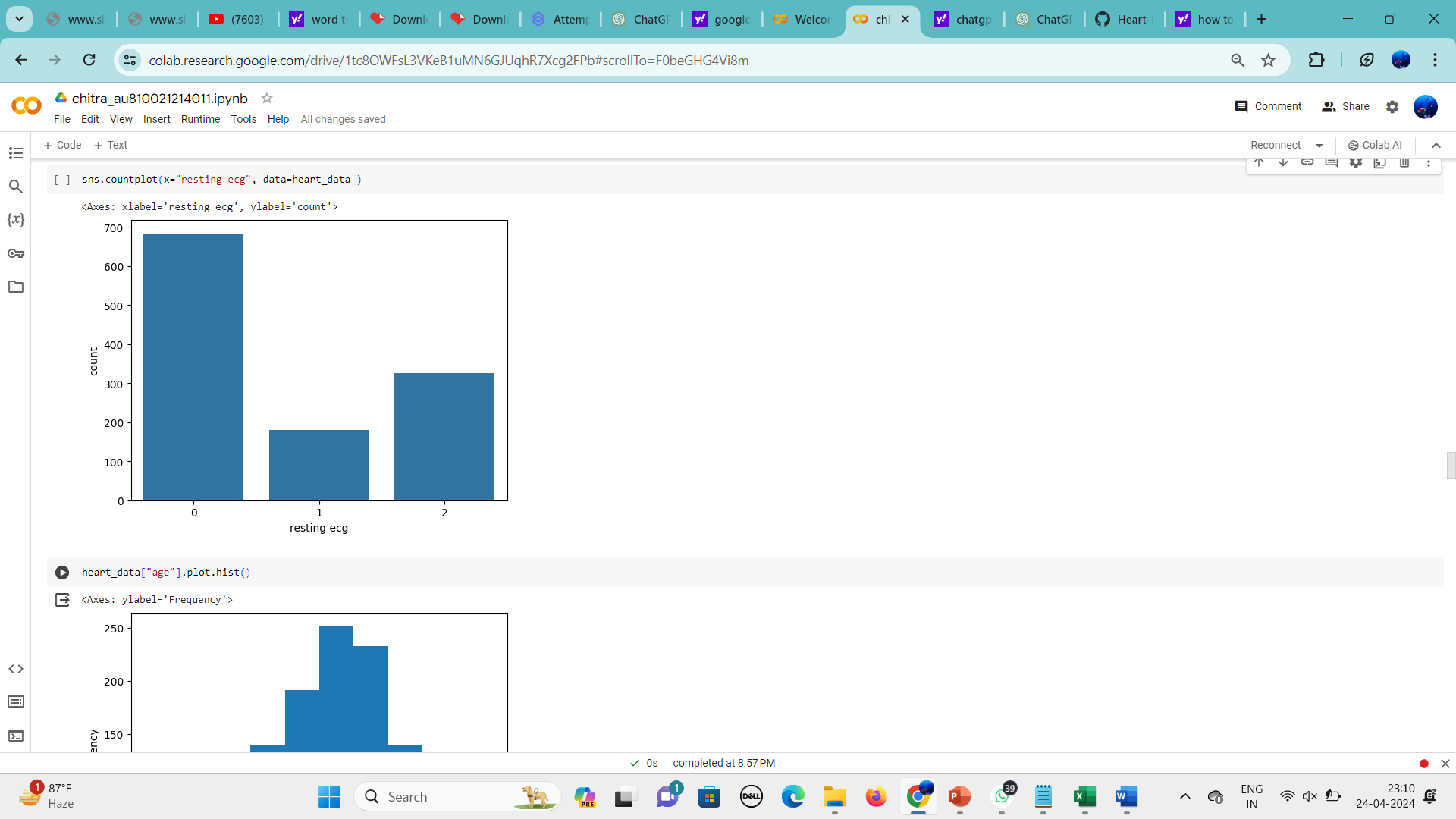
****

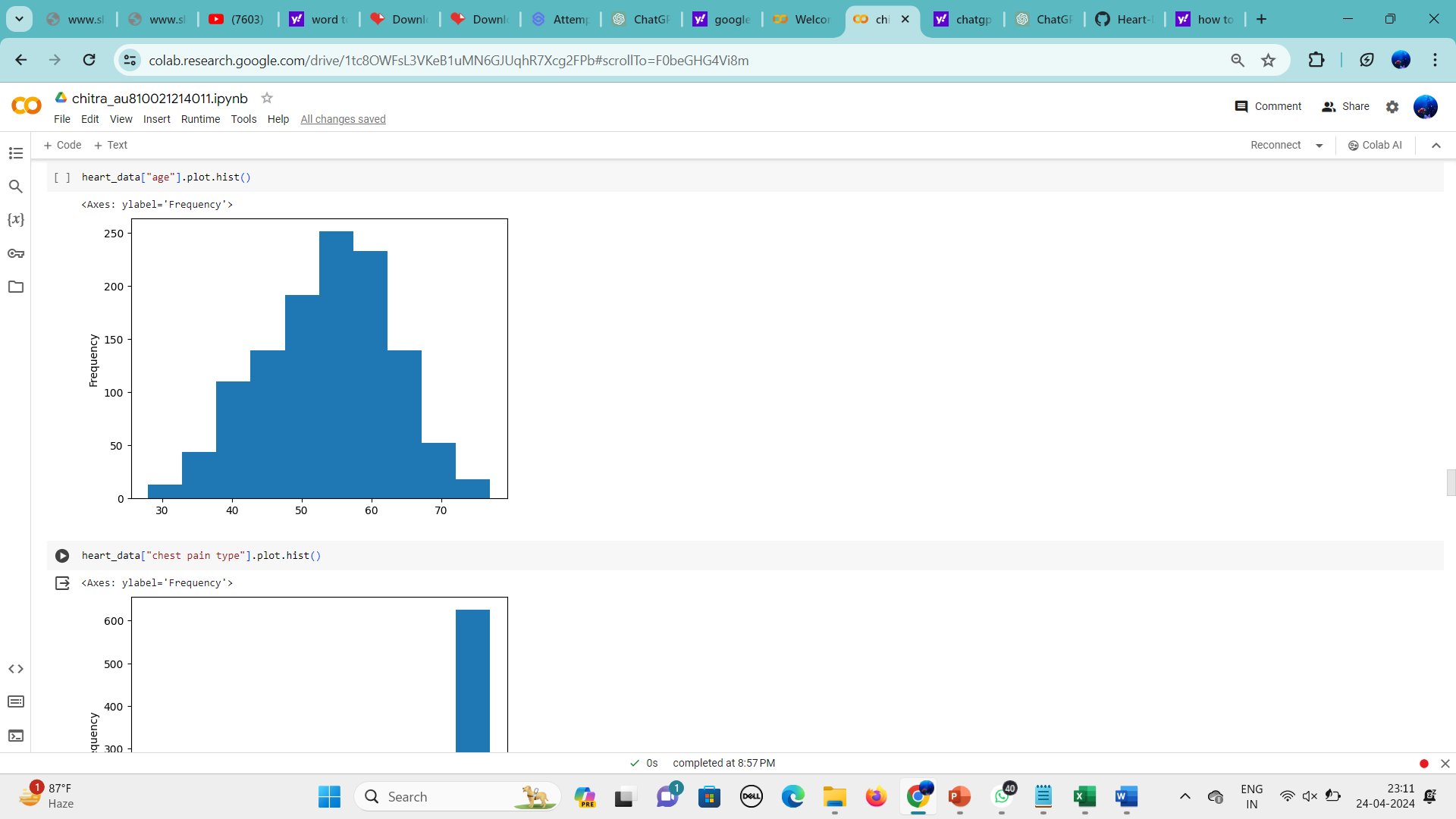
****

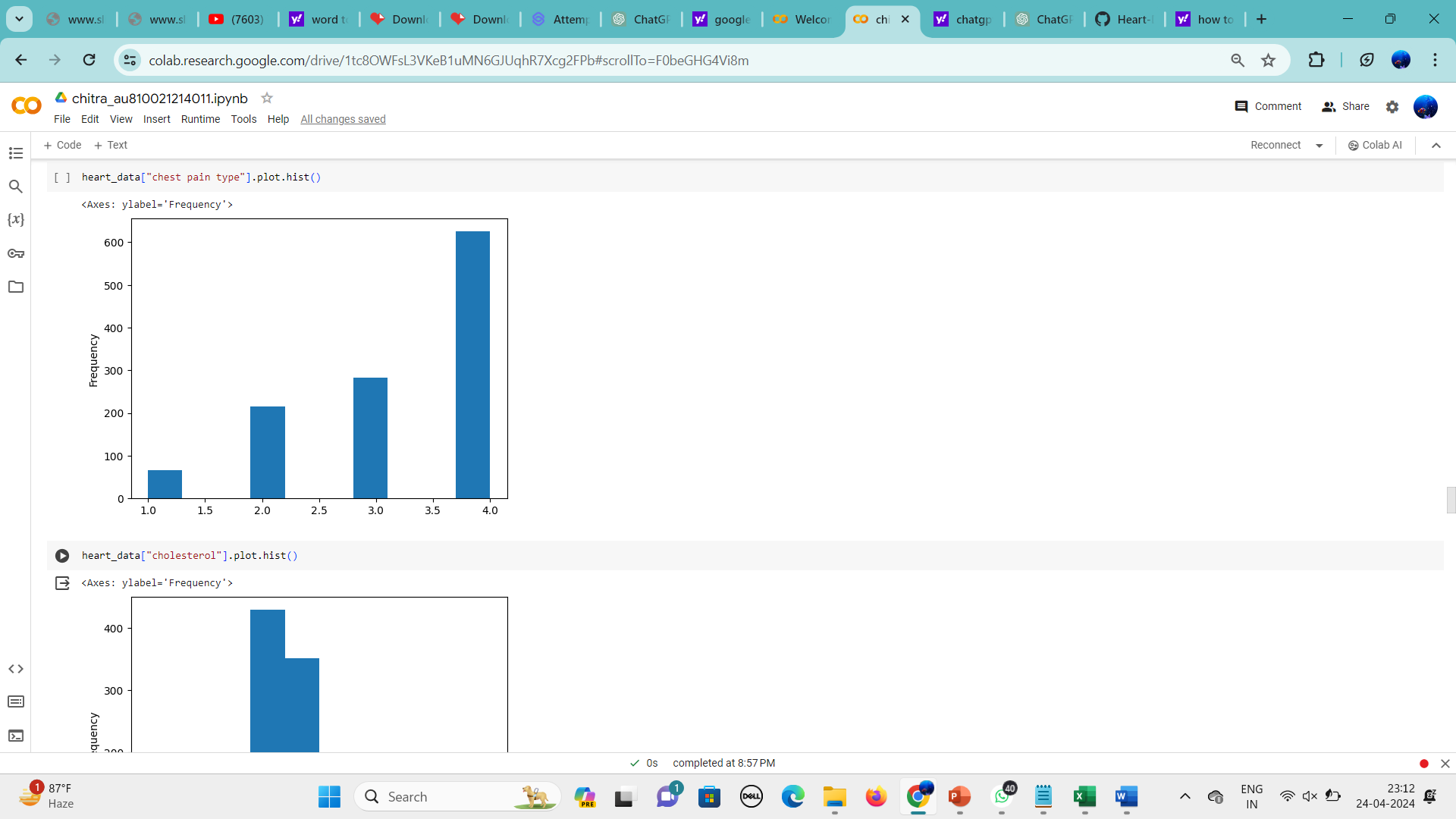
****

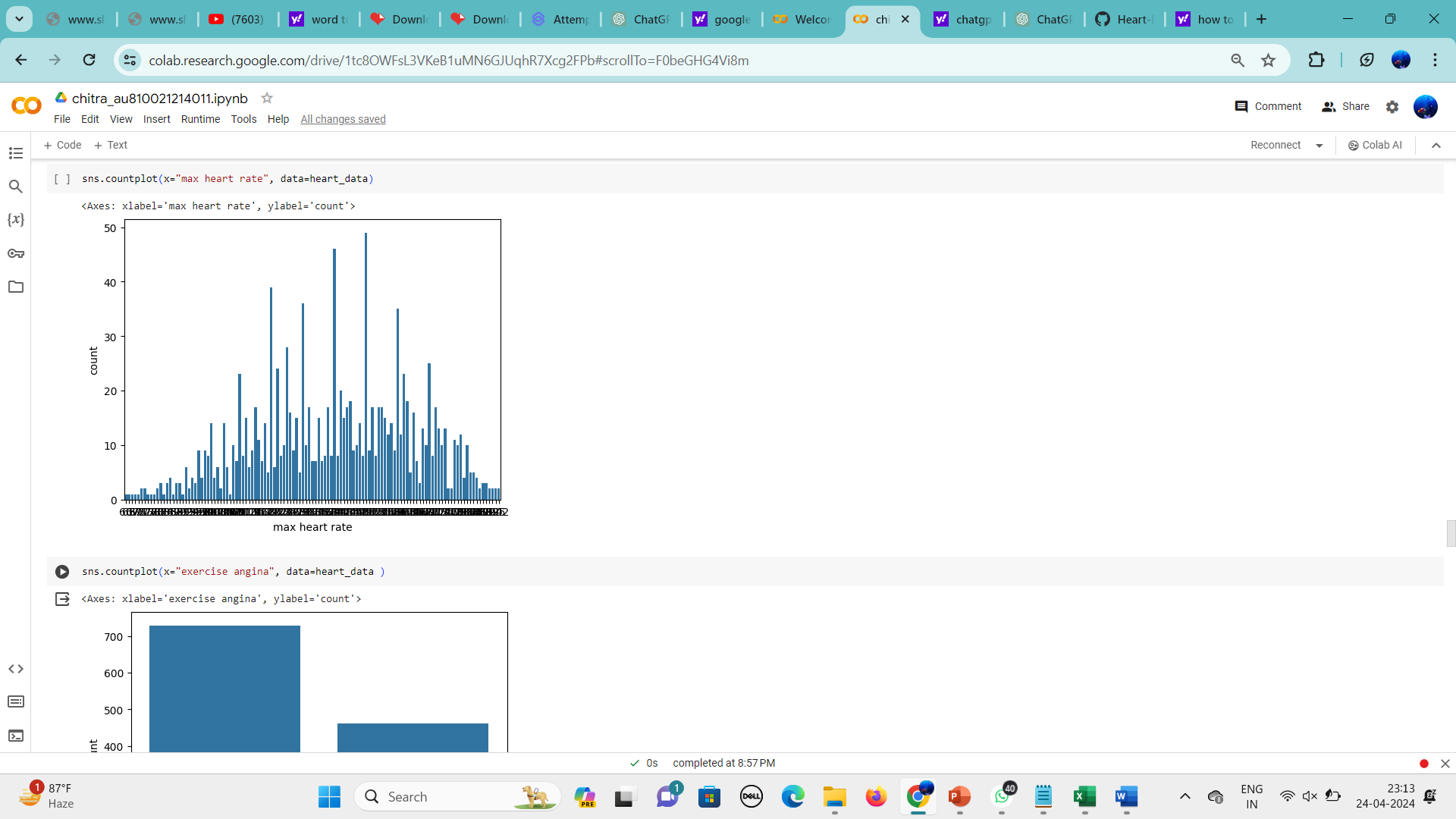
****

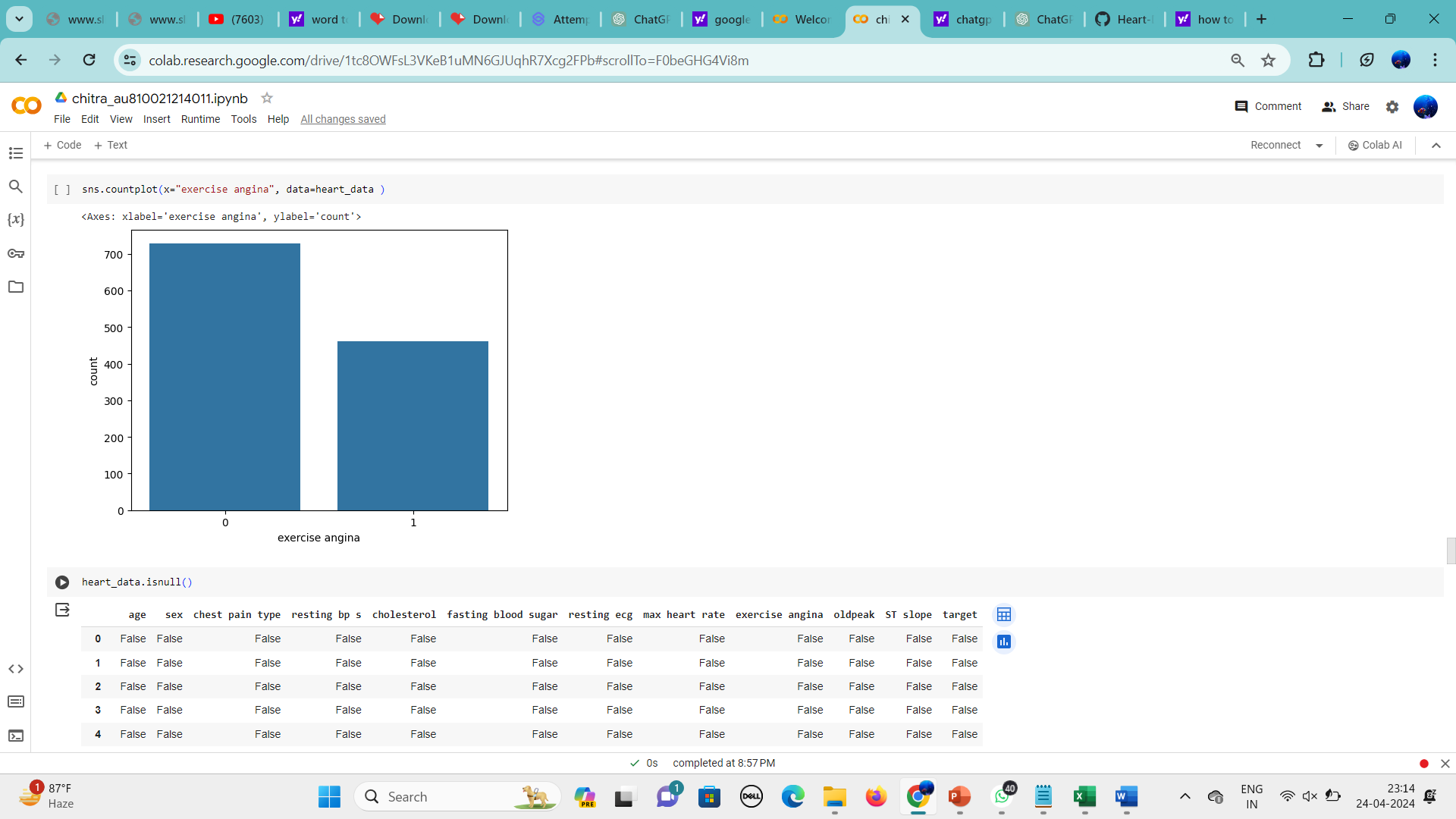
****

****

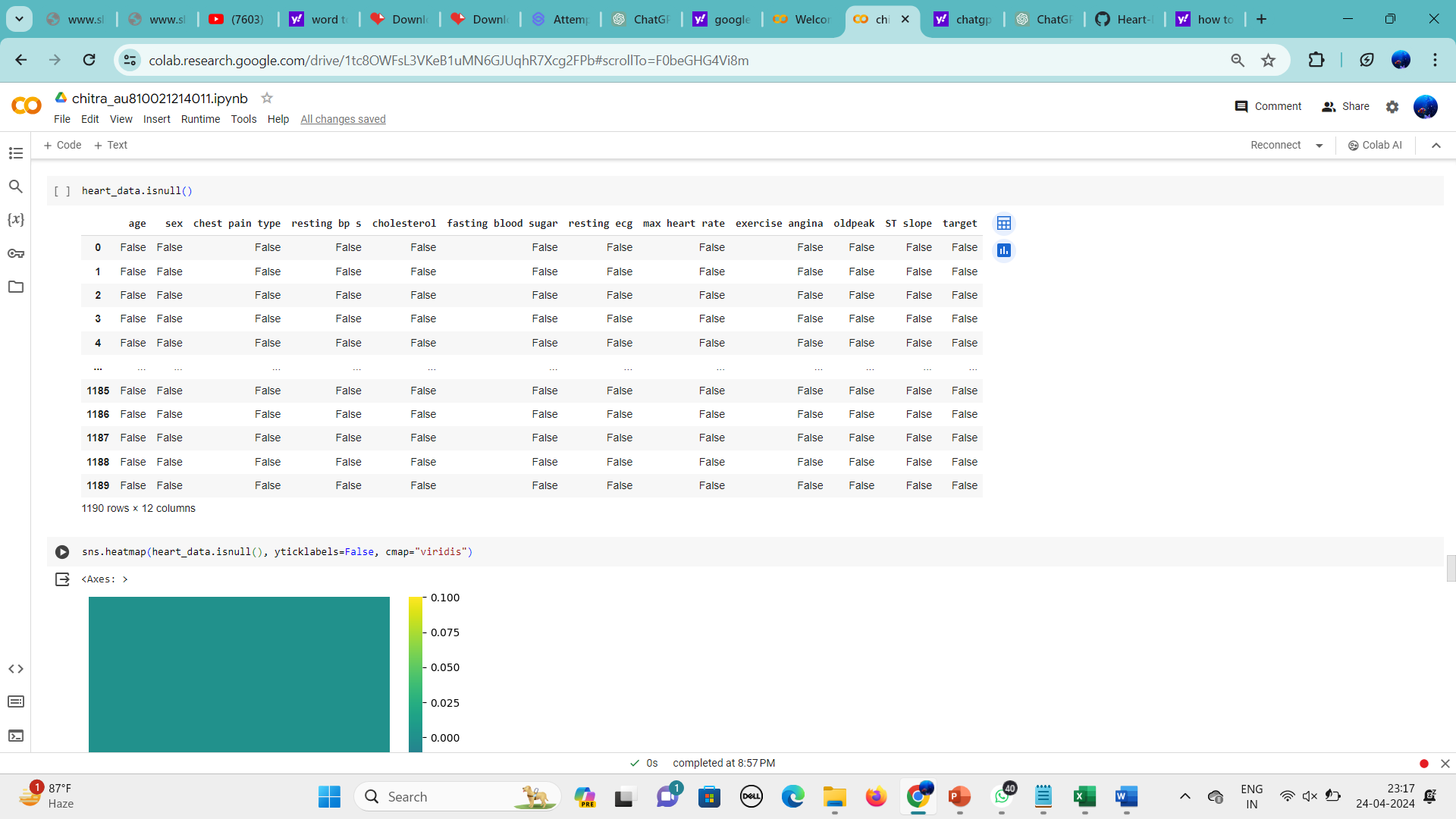
****

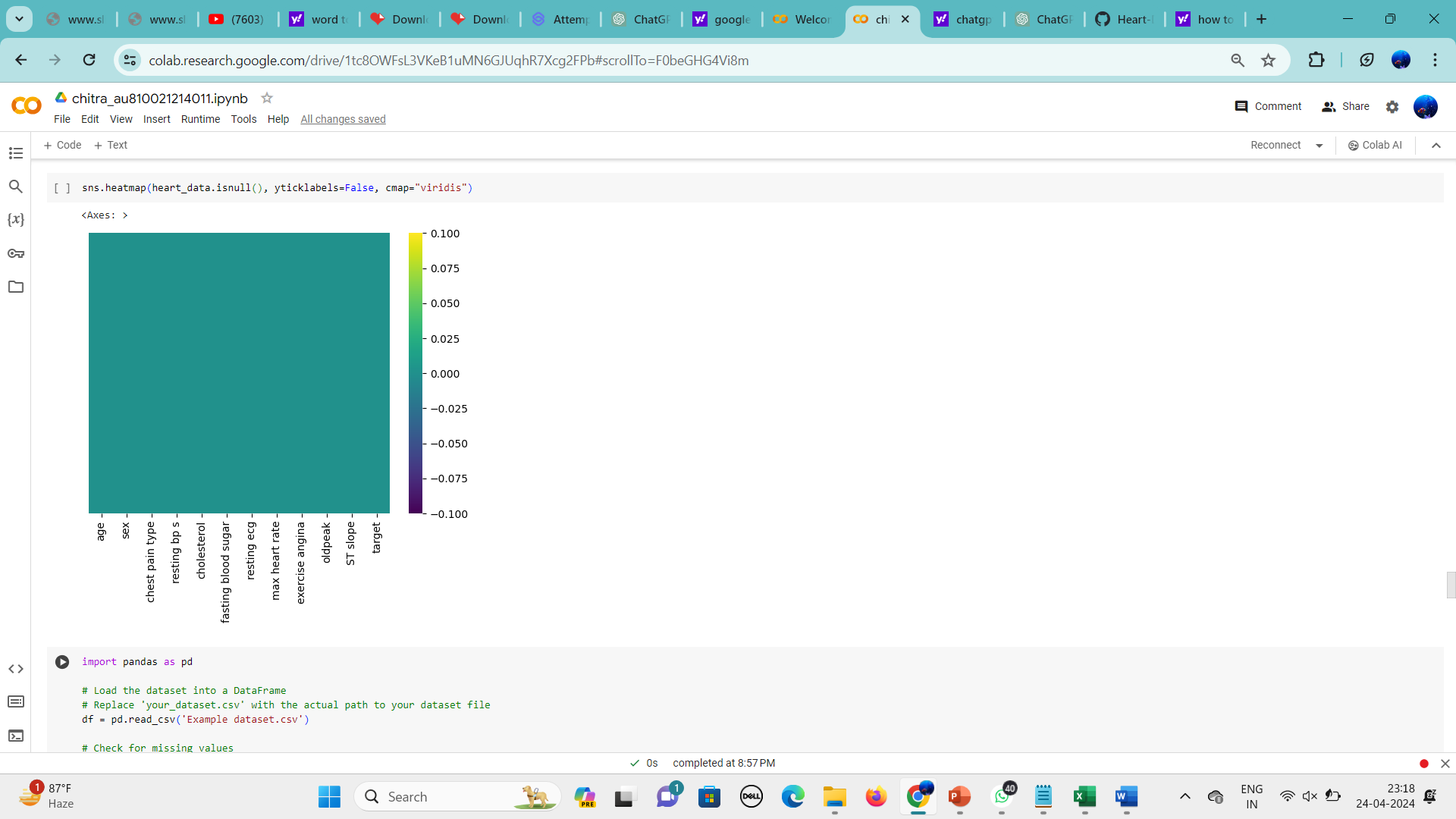
****

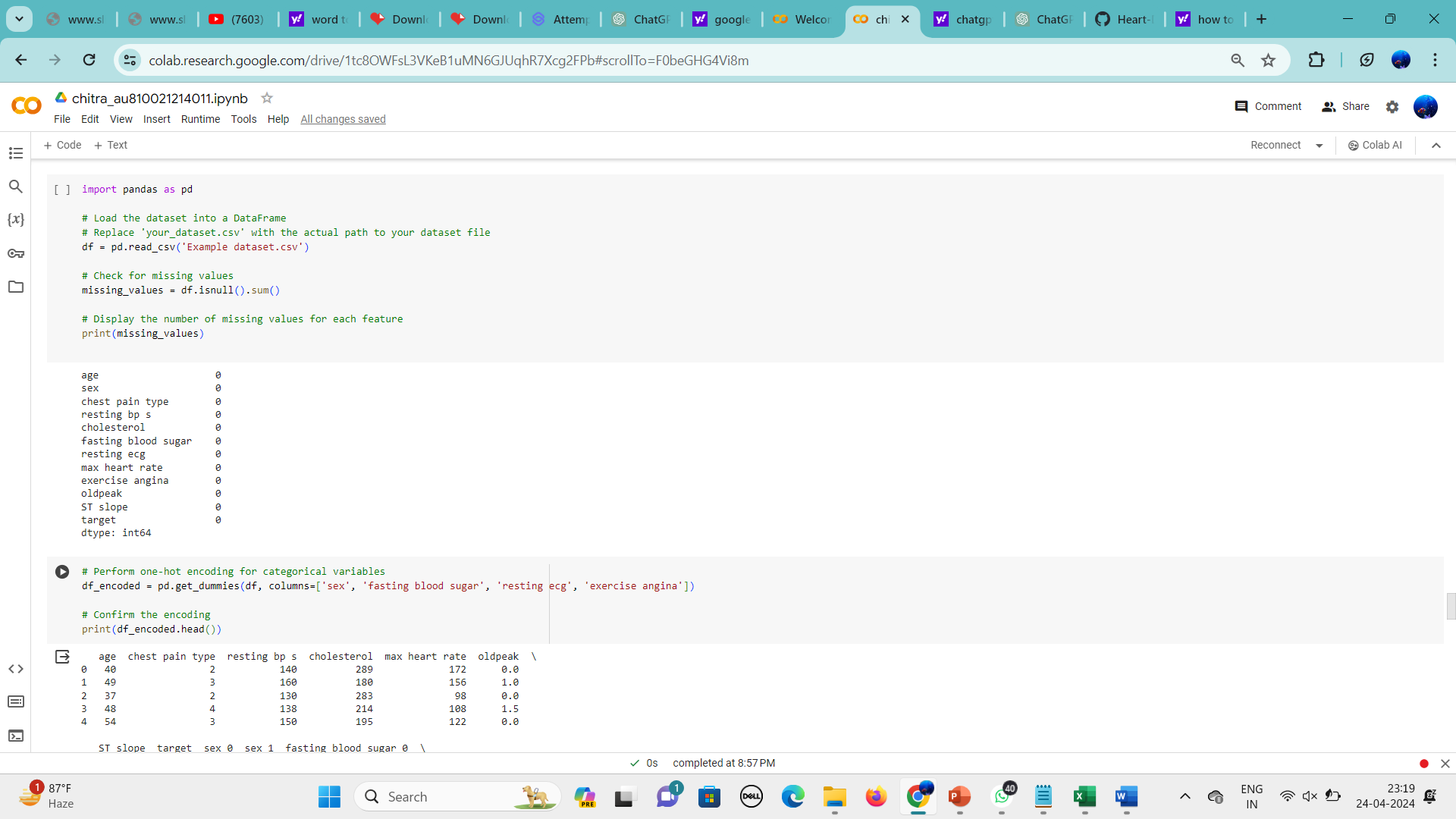
****

****

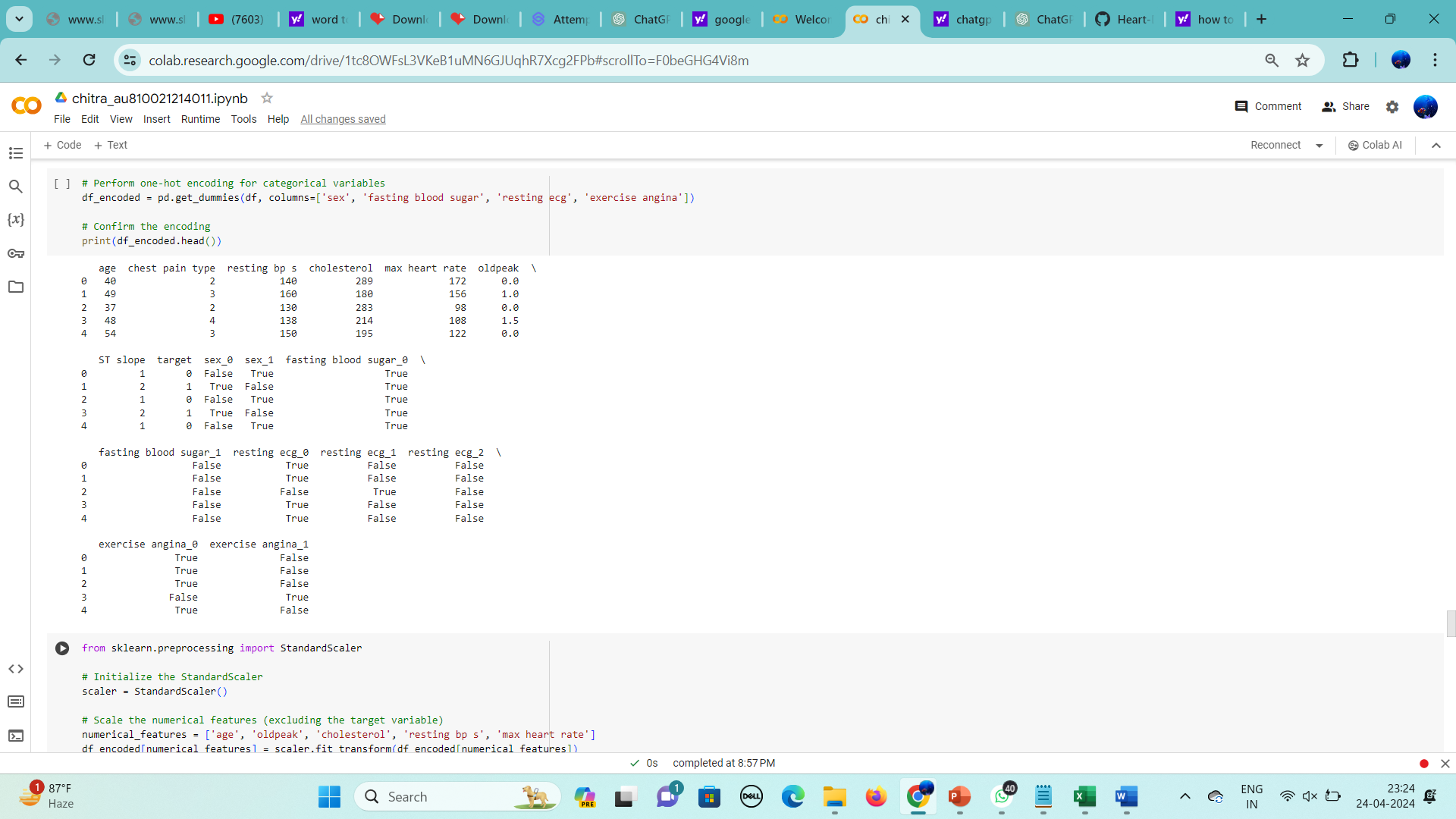
4. Finding null values



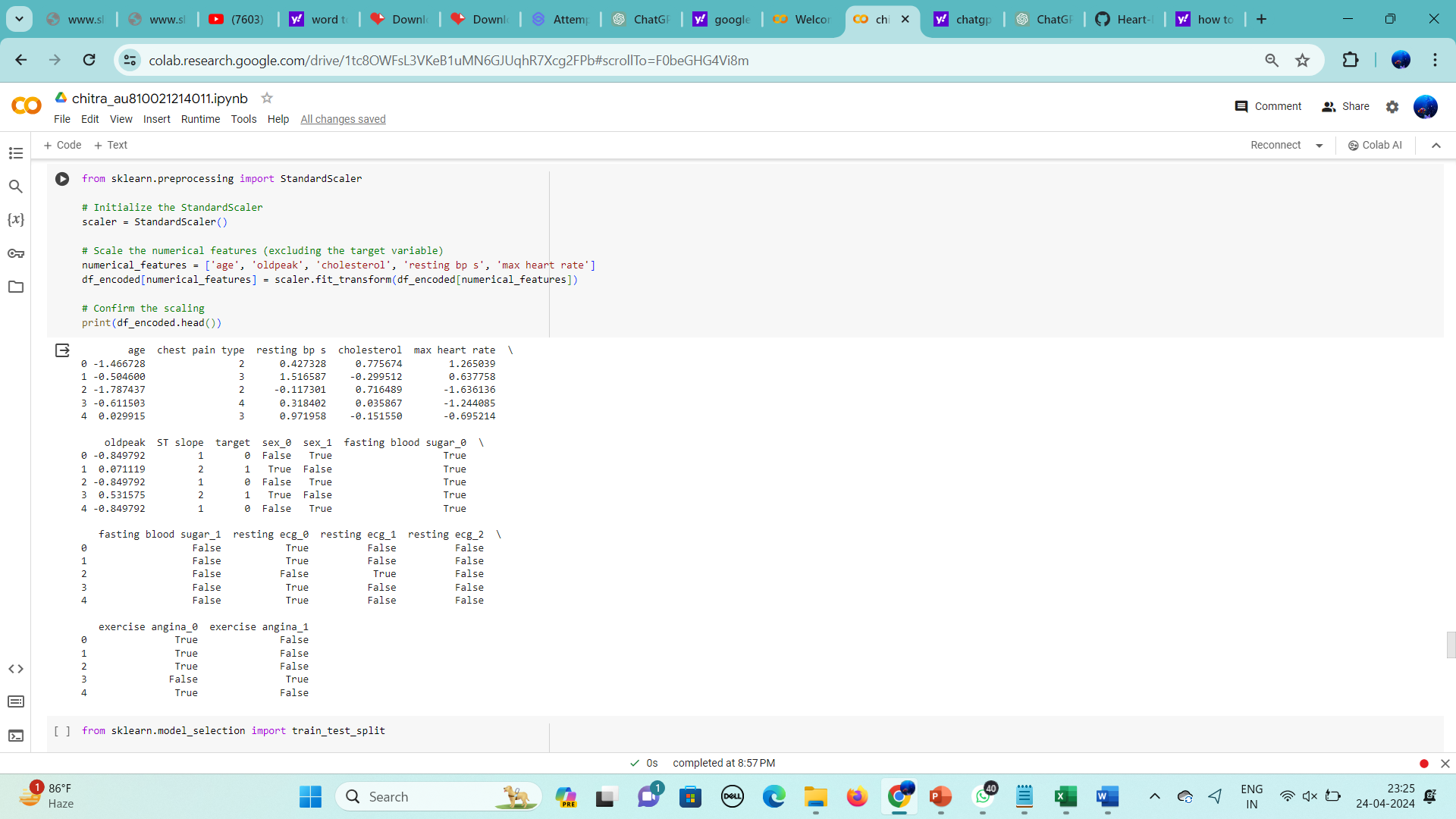




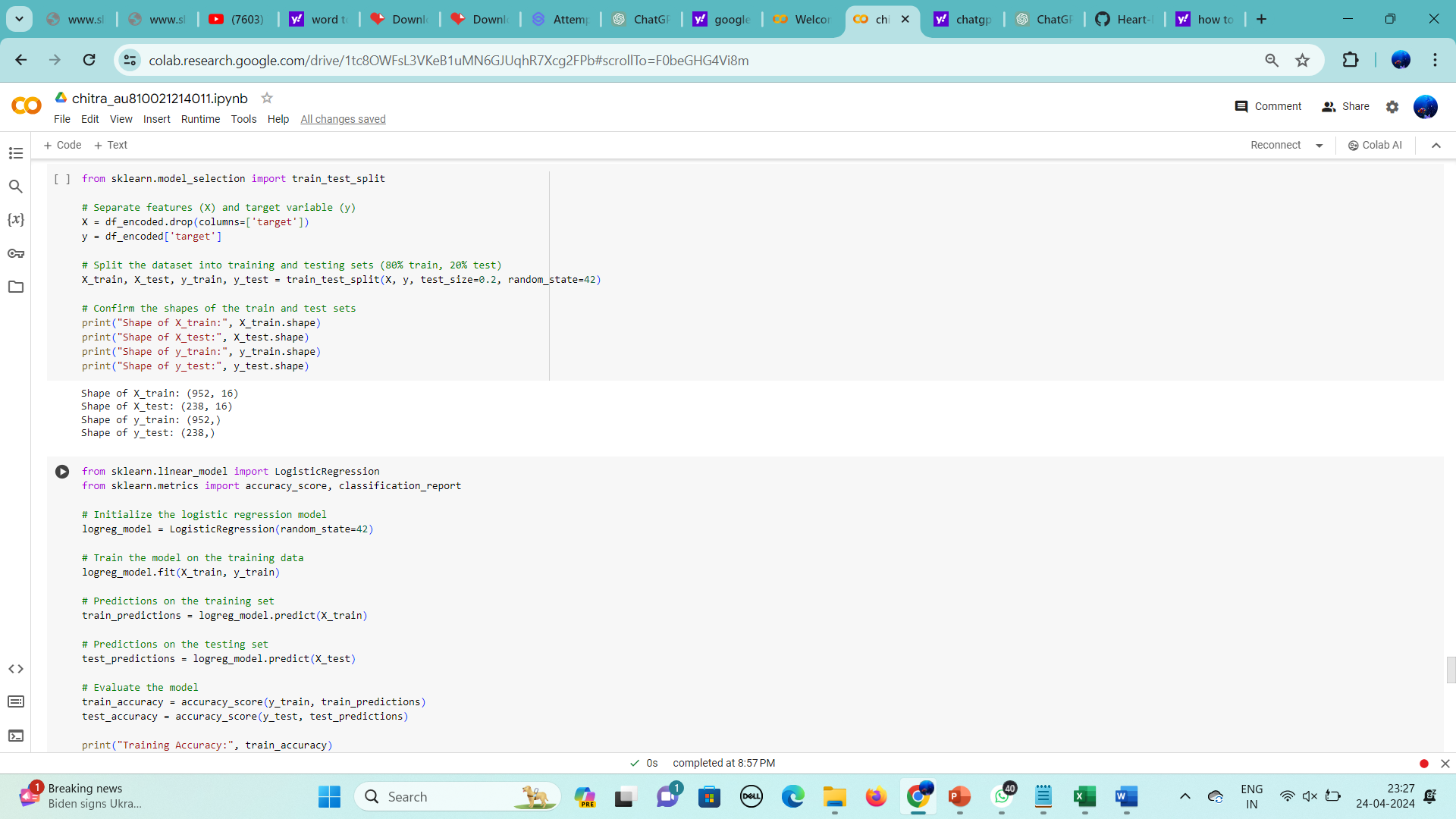
5.This code snippet performs one-hot encoding for categorical variables in the DataFrame df, where can visually inspect how the categorical variables have been transformed into binary features.

****

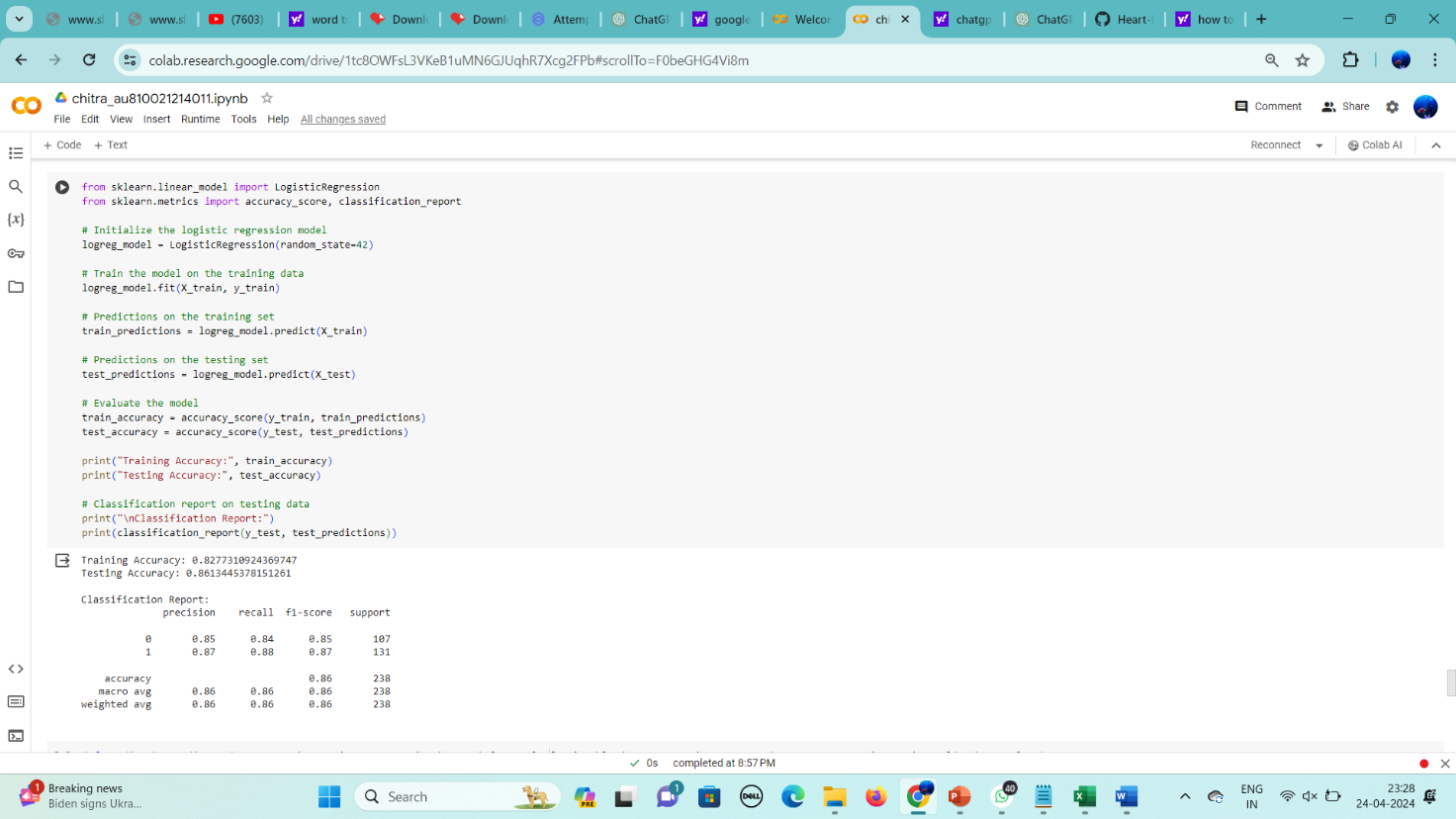
6**.** To standardize the numerical features in the DataFrame

****

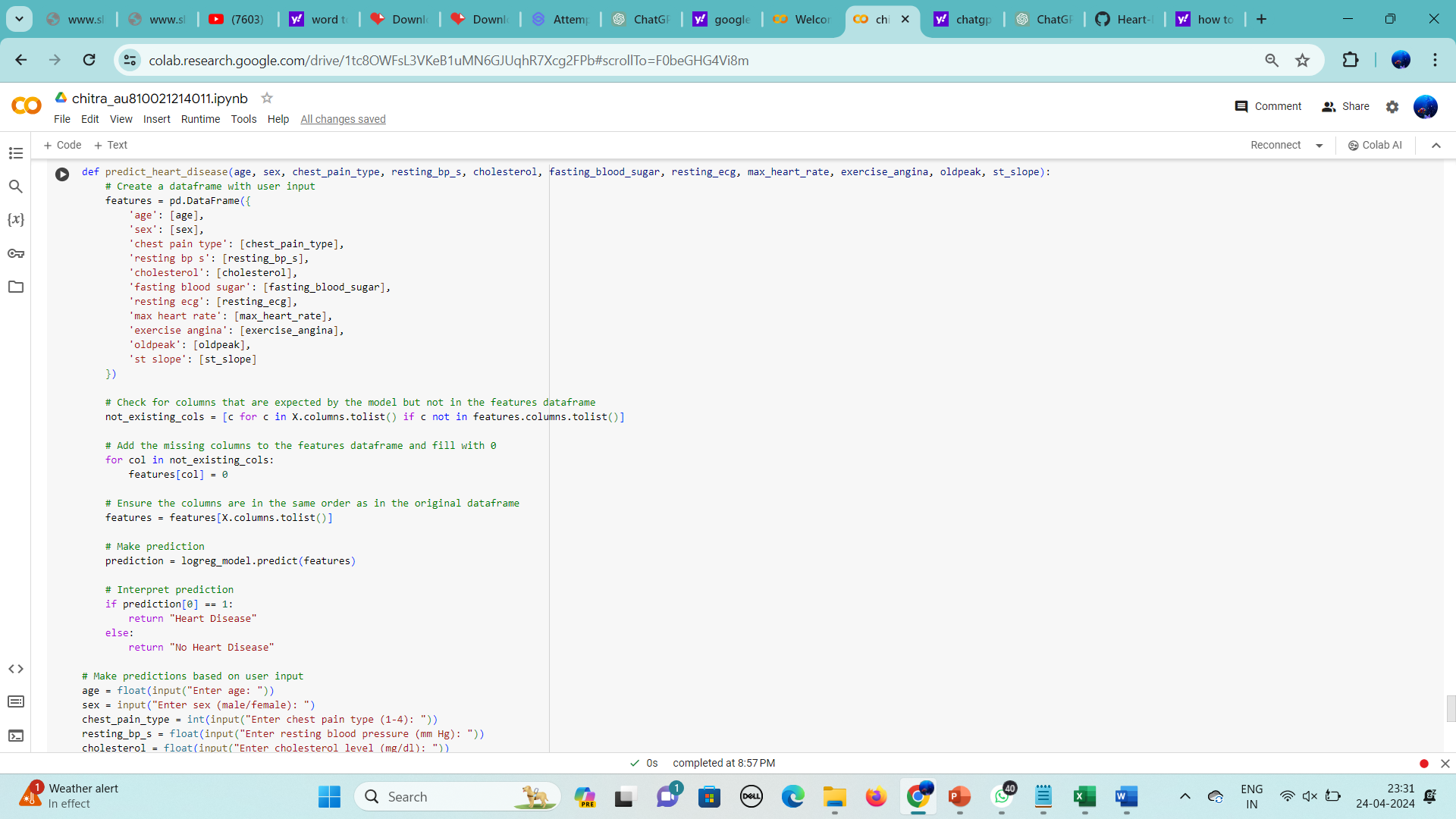
**7.** Prepares the dataset for model training and evaluation by splitting it into separate sets for training and testing, allowing for unbiased assessment of the model's performance on unseen data.

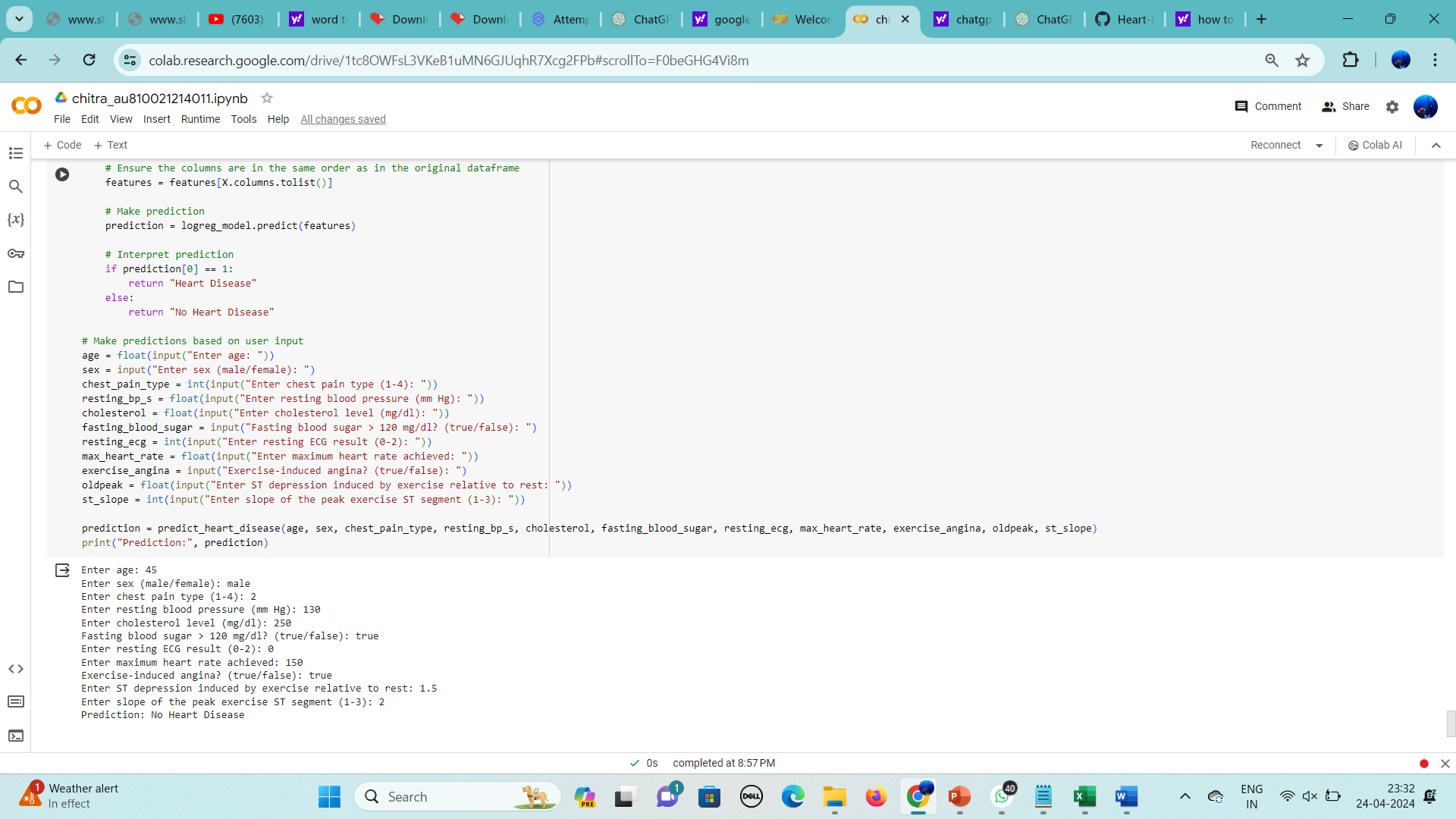
****

8**.** A Logistic Regression model on the training data, making predictions on both the training and testing sets, and evaluating the model's performance using accuracy score and classification report



9. This takes user input for various attributes related to heart health, creates a DataFrame with the input, and uses a pre-trained Logistic Regression model to predict whether the user is likely to have heart disease or not.

****

****

**CHAPTER 5**

**Project result**

The Logistic Regression model underwent training and achieved a testing accuracy when evaluated on the testing set. The classification report furnishes precise performance metrics for the model, including precision, recall, F1-score, and support for each class. This project's Logistic Regression model demonstrates promise in heart disease prediction utilizing demographic and clinical data. However, further adjustments and optimization of the model, along with validation on additional datasets, may be necessary to enhance its performance and applicability. The attained accuracy scores suggest that the model effectively predicts heart disease based on the provided features. Furthermore, the classification report offers deeper insights into the model's performance, including its capacity to accurately classify positive and negative instances, and its balance between precision and recall.

**CONCLUSION**

Utilizing Logistic Regression to forecast heart disease risk yields valuable insights for identifying individuals more prone to developing cardiovascular disease (CVD) and potentially averting adverse health consequences. Logistic Regression effectively models the likelihood of CVD occurrence by considering various risk factors, enabling early intervention and targeted healthcare strategies. The findings underscore the significance of proactive healthcare measures, such as lifestyle adjustments, regular screenings, and personalized therapies, in mitigating CVD risk factors and enhancing patient outcomes. The Logistic Regression model developed in this project exhibits promise in predicting heart disease based on demographic and clinical factors. Further refinement and optimization of the model, alongside validation on additional datasets, may be necessary to bolster its performance and applicability.

**FUTURE SCOPE**

The future of Logistic Regression-based heart disease prediction appears promising. When additional risk factors and powerful machine learning approaches are combined, advanced prediction models are likely to develop, resulting in better accuracy and reliability in identifying people at risk of cardiovascular disease (CVD). Furthermore, the concept of customized medicine is set to change heart disease management. Investigating personalized medicine approaches that tailor predictive models to individual patient characteristics, such as genetic predispositions, lifestyle factors, and medical history, could result in more accurate and targeted risk assessments, allowing for personalized prevention and treatment strategies.

**REFERENCES**

1. Project Github link, Ramar Bose , 2024
2. Project video recorded link (youtube/github), Ramar Bose , 2024
3. Project PPT & Report github link, Ramar Bose , 2024

# **GIT Hub Link of Project Code:**

<https://github.com/ChitraChellappa/Heart-Disease_au011>

YouTube link :

<https://youtu.be/F5T8bge390E>

Presentation link :

[Chitra Heart report.pptx](https://1drv.ms/p/c/68e6c1801a67dec6/ETXPaD28xXtBtBSrDLxcThABK1nEymKp6WvIVT415ZHbRw?e=d3j8Gx)