Dissertation Submitted for the partial fulfilment of the **M.Sc. (Integrated) Five Years Program AIML/Data Science** degree to the Department of AIML & Data Science.

**M.Sc. Project Dissertation**

**MedRec: Medical Report Data Extraction**

**& Maintenance Application**

submitted to

****

By

**Agarwal Shruti Hemant (DS - 01)**

**Sandhu Prableen Kaur Charanjitsingh (DS - 12)**

**Sharma Radhika Shivkumar (DS - 15)**

**Esha Mishra (AIML - 09)**

**Semester- VII**

**M.Sc. (Integrated) Five Years Program Data Science/AIML**

Department of AIML & Data Science

School of Emerging Science and Technology

Gujarat University

**December 2022**

**DECLARATION**

This is to certify that the research work reported in this dissertation entitled “**MedRec: Medical Report Data Extraction & Maintenance Application**” for the partial fulfilment of Master of Science in Data Science degree is the result of investigation done by myself.

|  |  |
| --- | --- |
|  |  |
|  |  |

Place: Ahmedabad Agarwal Shruti Hemant Date: 19th December 2022

**DECLARATION**

This is to certify that the research work reported in this dissertation entitled “**MedRec: Medical Report Data Extraction & Maintenance Application**” for the partial fulfilment of Master of Science in Data Science degree is the result of investigation done by myself.

Place: Ahmedabad Sandhu Prableen Kaur Charanjitsingh

Date: 19th December 2022

**DECLARATION**

This is to certify that the research work reported in this dissertation entitled “**MedRec: Medical Report Data Extraction & Maintenance Application**” for the partial fulfilment of Master of Science in Data Science degree is the result of investigation done by myself.

Place: Ahmedabad Sharma Radhika Shivkumar

Date: 19th December 2022

**DECLARATION**

This is to certify that the research work reported in this dissertation entitled “**MedRec: Medical Report Data Extraction & Maintenance Application**” for the partial fulfilment of Master of Science in Artificial Intelligence and Machine Learning degree is the result of investigation done by myself.

|  |  |
| --- | --- |
| Place: Ahmedabad | Esha Mishra |
| Date: 19th December 2022 |  |

Table of Contents

[Chapter 1 8](#_Toc122288846)

[Abstract & Key Words 8](#_Toc122288847)

[Abstract 9](#_Toc122288848)

[Key words 9](#_Toc122288850)

[Chapter 2 10](#_Toc122288852)

[Introduction 10](#_Toc122288853)

[Background 11](#_Toc122288854)

[Problem Statement 11](#_Toc122288855)

[Objective 11](#_Toc122288856)

[Introduction 11](#_Toc122288857)

[Chapter 3 12](#_Toc122288858)

[Literature review 12](#_Toc122288859)

[Benefits of Data Archiving 13](#_Toc122288860)

[Storage Requirements 13](#_Toc122288861)

[Unified medical database system launched by government 14](#_Toc122288862)

[What is Unified Health Interface? 14](#_Toc122288863)

[Medical Record Management 15](#_Toc122288864)

[What is a personal health record? 15](#_Toc122288865)

[Chapter 4 17](#_Toc122288866)

[Basic Terminology 17](#_Toc122288867)

[Chapter 5 19](#_Toc122288868)

[Methodology 19](#_Toc122288869)

[Selection of OS 20](#_Toc122288870)

[Microsoft Windows was used for this project because it is user friendly & it's robust. The code works on all OS. 20](#_Toc122288871)

[Selection of Coding Language 20](#_Toc122288872)

[Libraries Used 22](#_Toc122288874)

[Project Design 29](#_Toc122288876)

[Project GUI 29](#_Toc122288877)

[Chapter 6 31](#_Toc122288878)

[Data Analysis 31](#_Toc122288879)

[Data Collection 32](#_Toc122288880)

[About Data 32](#_Toc122288882)

[Training & Testing 32](#_Toc122288884)

[Result & Discussion 40](#_Toc122288885)

[Conclusion 45](#_Toc122288886)

[Future work 47](#_Toc122288887)

[Bibliography 49](#_Toc122288888)

# 

# Chapter 1

# Abstract & Key Words

## Abstract

## For any significant diagnosis or prescription, a patient's medical report is a must. On an average a person gets at least 5 tests done every year. India is the diabetes capital of the world with 77 million formally diagnosed patients. These people need to get their blood glucose levels regularly tested. Patients with chronic illnesses like thyroid disorder, PCOS/ PCOD get tested every 3 months. Physical file maintenance of these medical reports is laborious and they are easy to misplace or missing when one needs them, especially in case of emergency. Maintaining these reports becomes more challenging with time. Therefore to overcome this problem an end-to-end solution for storing and maintaining medical reports using deep learning algorithms is provided. where The user will give an input image of their medical report and then on the backend the machine will extract the date of report using a OCR from the report's image. This data will be maintained in the Firebase database and will be easily accessible to the user. This will help users to store data in the app itself and thus reduce the storage cost and burden of maintaining data physically.

## Key words

## Medical, Maintenance, Patients, Report, Data, Extraction.

# Chapter 2

# Introduction

## Background

80% of all healthcare data is unstructured and inaccessible for further processing. This limits the quantity of usable data and also limits a healthcare organization’s decision-making capabilities. According to a study, 30% of healthcare costs are associated with administrative tasks. AI can automate some of these tasks. The global healthcare AI market size is expected to grow from USD 3.64 billion in 2019 to USD 33.42 billion by 2026, at a Compound Annual Growth Rate (CAGR) of 46.21% during the forecast period. Medical report storage is a practical solution to free up space while keeping records safe, ensuring they are in structured form and not lost, stolen, or damaged.

## Problem Statement

With more and more physical medical reports of a patient, it becomes a cumbersome task to store and maintain them as it can get misplaced or lost.To overcome this problem we need to organize it digitally and effectively.So that report can be easily accessible when needed.

## Objective

Our objective is to build a practical end-to-end solution that helps in keeping records safe along with easy access using deep learning and image processing. The goal of medical record keeping is to help ensure that your patients receive quality care at the right time.

## Introduction

Archiving is the process of securely storing inactive information in any format that you no longer use regularly for long-term retention. Such information is still important to organizations and must be retained for future reference or regulatory compliance.

For life-saving documents like medical records, there is not solution available that allows easy access to patient medical history. In case of a medical emergency, easy and fast access to medical history reduces diagnosis time and enables doctors to get the treatment quickly. Big hospitals have their own softwares which will not be accessible to other medical institutions outside the hospital network. Thus, there is a need for a solution that's accessible to all.

# Chapter 3

# Literature review

## Benefits of Data Archiving

The primary benefits of archiving data are:

* Reduced cost━data is typically stored on low performance, high capacity media with lower associated maintenance and operation costs.
* Better backup and restore performance━archiving removes data from backups, reducing their size and eliminating restoration of unnecessary files.
* Prevention of data loss━archiving reduces the ability to modify data, preventing data loss
* Increased security━archiving removes documents from circulation, limiting the chance of cyberattack or malware infection.
* Regulatory compliance━built-in policies ensure records are kept for an appropriate amount of time and indexing makes data more retrievable.

## Storage Requirements

The type of storage you choose plays a big role in how accessible your data is, how much your archive costs to create and store, and how safe your data is once it’s archived. An archive is only useful if you can retrieve data when you need it, so it’s important to periodically verify that the storage you select continues to be functional.

* Online storage storing your archive online allows you to easily access it from multiple locations and ensures that you can retrieve the data quickly. It also makes it easier to manage efficiently and add more data to it. The downside of online storage is that it increases opportunities for theft or tampering and is only accessible when you have a network connection. Private clouds can reduce your security risks but have high upfront and operating costs whereas public clouds are cheaper upfront and include built-in support and encryption but require ongoing fees for use.
* Offline storage storing archives offline, such as with disk or tape drives, reduces the risk of theft or modification as well as maintenance and storage costs. Offline storage often has a better capacity to cost ratio but means longer retrieval times and greater barriers to managing or transferring data.

## Unified medical database system launched by government

The Prime Minister of India announced the launch of the National Digital Health Mission (NDHM) in 2021. Under NDHM, the Unified Health Interface launch (UHI) was also announced in India. This mission aims toward the digital development of healthcare across the country.

The goal of the Unified Health Interface is to make it as common as UPI in the country. This is why it is important to understand what UHI, Unified Health Interface, uses, and benefits. Read on to know more and how to register for Unified Health Interface.

## What is Unified Health Interface?

Unified Health Interface is an open IT network to provide a digital platform for all health services. It is considered a part of the foundational layer of ABDM (Ayushman Bharat Digital Mission). Under NDHM, UHI aims to enable the following things: [1]

For health service providers (doctors, pharmacists, hospitals):

* List of their services (appointment, teleconsultations)
* Immediate access to user demand generated in UHI
* Ongoing connectivity with existing customers
* Access to health records in a single place

For Patients:

* Facility to contact doctors immediately via the UHI platform
* Easy digital health access to all Indians
* Option to share information with healthcare service providers
* Features to receive doctors’ prescriptions and lab reports digitally on your device
* Availability of a wide variety of health services with the provision of full transparency

## Medical Record Management

Medical records management is organizing and handling patients’ health information in a way convenient for medical professionals to access and carry out healthcare workflows. Patient records are created when people walk into a clinic or healthcare centre for the first time. Health records management involves rules, regulations, procedures, and protocols to manage health records throughout their existence. This involves updating, modifying, communicating with other providers, and ensuring privacy and integrity. Managing health records is important since the data it contains leads to treatment decisions by doctors.

A medical record management system enables physicians to access patients' medical information. This contains details like treatment history, immunizations, allergies, previous medications, pre-existing conditions, etc. This is vital for doctors who need to prescribe tests, scans, or medications for patients. A person’s health history can impact how he or she responds to certain treatments, so health records are important for making informed decisions.

An [electronic health record](https://www.healthit.gov/providers-professionals/learn-ehr-basics) (EHR) is a digital version of a patient’s paper chart. EHRs are real-time, patient-centred records that make information available instantly and securely to authorized users. While an EHR does contain the medical and treatment histories of patients, an EHR system is built to go beyond standard clinical data collected in a provider’s office and can be inclusive of a broader view of a patient’s care. EHRs are a vital part of health IT and can:

* Contain a patient’s medical history, diagnoses, medications, treatment plans, immunization dates, allergies, radiology images, and laboratory and test results
* Allow access to evidence-based tools that providers can use to make decisions about a patient’s care
* Automate and streamline provider workflow

## What is a personal health record?

A personal health record is simply a collection of information about your health. If you have a shot record or a folder of medical papers, you already have a basic personal health record.

And you've probably encountered the big drawback of paper records: You rarely have them with you when you need them.

Electronic personal health records (PHRs) remedy that problem by making your information accessible to you anytime via web-enabled devices, such as computers, smartphones and tablets.

# Chapter 4

# Basic Terminology

**Date Extraction -** Date extraction refers to the process of identifying and extracting dates from a given piece of text. This can be useful in a variety of applications, such as information retrieval, natural language processing, and data analysis. There are a few different approaches that can be used to extract dates from text, depending on the specific requirements and context of the task. One common approach is to use regular expressions to identify patterns in the text that match the format of a date. Another approach is to use natural language processing techniques to identify dates in the text. This involves understanding the context and meaning of the text, it can be more accurate and flexible, as it can handle a wider range of date formats and variations.

**OCR -** OCR (Optical Character Recognition) is a technology that allows for the conversion of scanned documents or images containing text into editable and searchable digital text. OCR software works by analyzing the text in an image or scanned document and recognizing the individual characters, then converting them into a digital format that can be edited or searched.

**Cloud Storage -** Firebase is a cloud-based platform that provides a variety of tools and services for mobile and web application development. One of the services offered by Firebase is a real-time database, which allows developers to store and synchronize data across multiple devices in real-time.

**GUI -** A GUI, or graphical user interface, is a type of user interface that allows users to interact with electronic devices through visual elements such as windows, icons, and buttons. GUIs can be found on a wide variety of electronic devices, including computers, smartphones, and tablets.

**Augmentation -** Data augmentation is a technique used to increase the size and diversity of a dataset by generating new data samples from existing ones. The goal of data augmentation is to improve the performance of a model by providing it with more diverse and representative training data.

# Chapter 5

# Methodology

## 

## 

## Selection of OS

## Microsoft Windows was used for this project because it is user friendly & it's robust. The code works on all OS.

## Selection of Coding Language

## Python:



Python is a high-level, general-purpose programming language that is widely used for web development, data analysis, scientific computing, and many other purposes. It is known for its simplicity, readability, and flexibility, as well as its large and active developer community. Some of the key features of Python include: A large standard library that supports many common programming tasks, such as connecting to web servers, reading and writing files, and working with data, An interactive interpreter, which allows you to try out code snippets and experiment with the language in an interactive environment, Support for object-oriented, imperative, and functional programming styles, Dynamically-typed, which means that you don't have to specify the data type of a variable when you declare it, Cross-platform compatibility, which means that Python programs can run on multiple operating systems.

* **HTML:**



HTML (HyperText Markup Language) is a markup language used to structure and format content on the web. It consists of elements that are represented by tags, which are used to define the structure and layout of content on a web page. It was used to design web applications.

## Libraries Used

* **OS:**

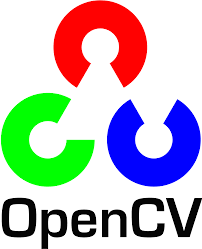
The os library is a module in the Python standard library that provides functions and utilities for interacting with the operating system. It allows you to perform tasks such as reading and writing files, creating and deleting directories, and interacting with the file system and environment variables.

# AugLy : Facebook’s New Augmentation Library



AugLy is a novel open source data augmentation library that combines multiple modalities: audio, image, video, and text, which is increasingly important in many AI research fields.

* **OpenCV (cv2):**



cv2 (OpenCV) is a computer vision library that is widely used in a variety of applications, including image and video processing, machine learning, and robotics. It is a free and open-source library that provides functions and utilities for performing tasks such as image and video manipulation, object detection, and feature extraction.

* **Numpy:**



NumPy is a library for the Python programming language that is used for scientific computing and data analysis. It provides functions and utilities for working with large, multi-dimensional arrays and matrices of numerical data, as well as for performing mathematical operations on these data. One of the main advantages of NumPy is that it is very efficient for performing operations on large arrays and matrices, as it is implemented in C and can make use of the performance of compiled code.

* **PIL:**



Pillow (PIL) is a library for the Python programming language that is used for handling and manipulating image data. It provides functions and utilities for reading and writing image files, as well as for performing tasks such as resizing, cropping, and applying image filters.

* **Matplotlib:**



Matplotlib is a library for the Python programming language that is used for data visualization. It provides functions and utilities for creating a wide variety of plots and charts, including line plots, bar plots, scatter plots, and pie charts.

* **Pandas:**



Pandas is a library for the Python programming language that is used for data manipulation and analysis. It provides functions and utilities for working with tabular data, such as data stored in a spreadsheet or a database table. One of the main advantages of Pandas is that it provides a high-level interface for working with data, making it easy to perform tasks such as filtering, aggregating, and transforming data. It also integrates well with other libraries for data analysis, such as NumPy and Matplotlib.

* **TensorFlow:**



TensorFlow is an open-source software library for machine learning and artificial intelligence. It was developed by Google and is widely used in a variety of applications, including natural language processing, image and video analysis, and predictive modeling. It is designed to be flexible and scalable, and it can be used for tasks such as training and deploying machine learning models, performing data preprocessing and feature engineering, and optimizing model performance. It can be used on a variety of platforms, including desktop computers, servers, and mobile devices.

* **Glob:**



glob is a module in the Python standard library that provides functions for working with file paths and patterns. It allows you to search for files in a directory based on a pattern, and it returns a list of matching file paths. The glob module also supports more advanced pattern matching, using wildcards.

* **Keras:**



Keras is an open-source software library for machine learning that is built on top of other libraries, such as TensorFlow, Theano, and CNTK. It provides a high-level interface for building and training neural network models, making it easy to experiment with different architectures and hyperparameters. Keras is designed to be user-friendly and easy to use, and it is particularly well-suited for prototyping and developing machine learning models quickly. It supports a wide range of tasks, including classification, regression, and generation, and it can be used with a variety of different types of neural networks, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs).

* **Pytesseract:**



Pytesseract is a Python wrapper for the Tesseract OCR engine. Tesseract is an open-source OCR (Optical Character Recognition) engine that can be used to extract text from images and convert them into digital text.Pytesseract provides various options for customizing the OCR process, such as specifying the language of the text in the image, or disabling image preprocessing.

* **Dateparser:**



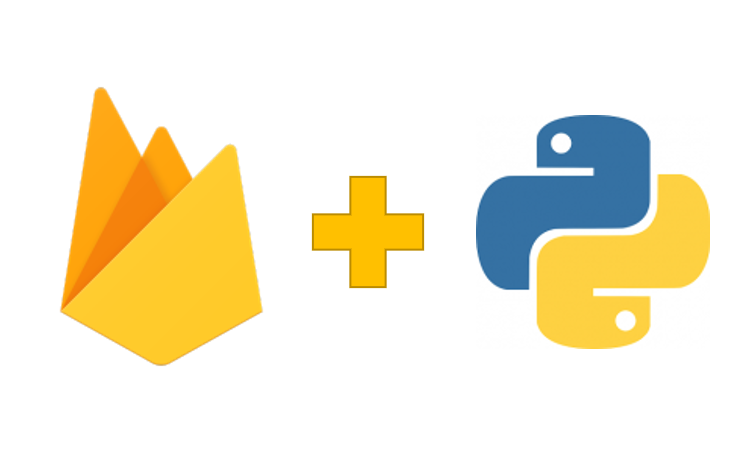
dateparser is a Python library that provides functions for parsing and manipulating dates and times. It is designed to be flexible and handle a wide range of date and time formats, including complex and ambiguous input. It can also handle timezone information and handle dates in different languages.

* **Datetime:**



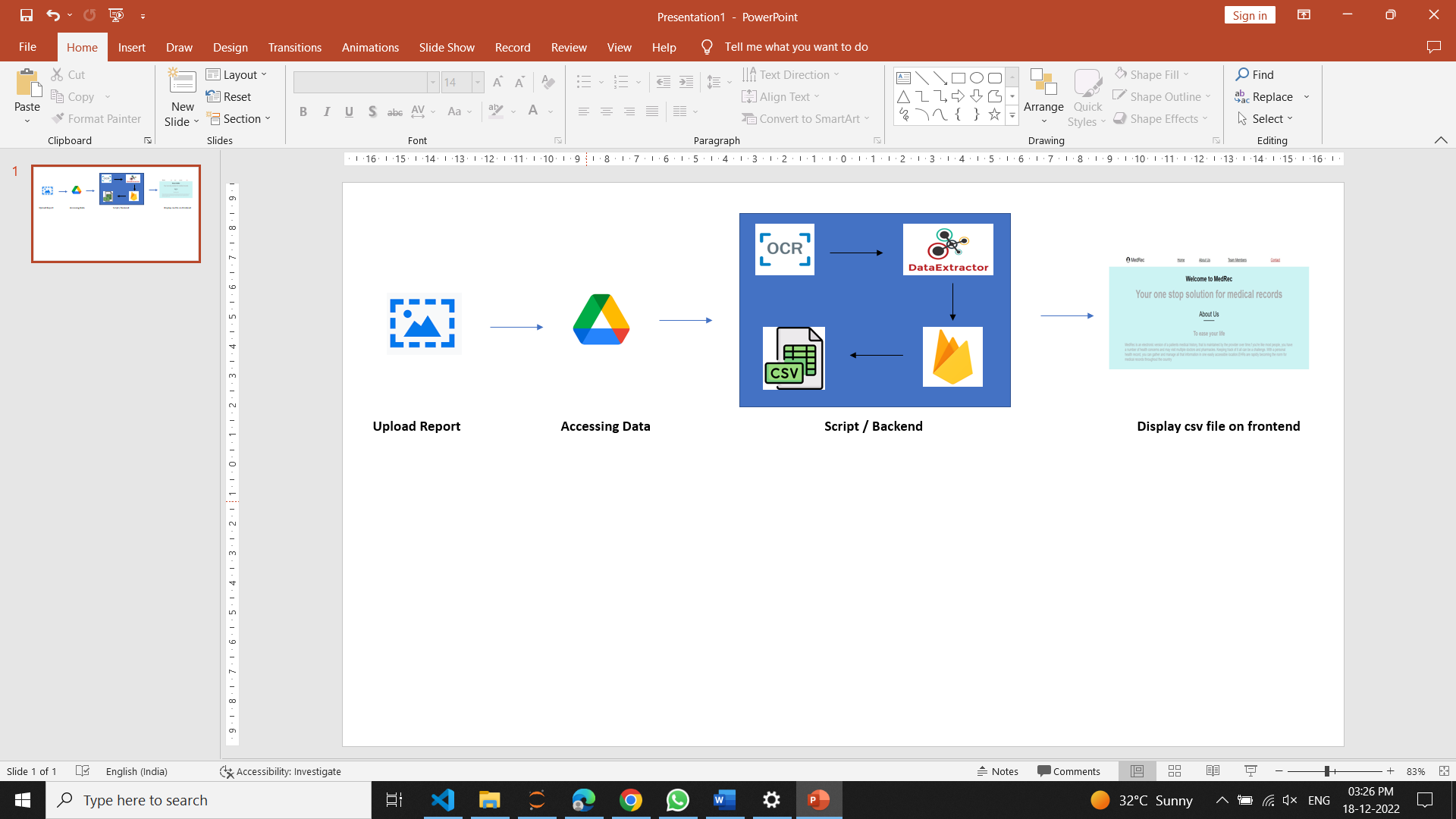
datetime is a module in the Python standard library that provides classes for working with dates and times. It allows you to store and manipulate dates and times in a variety of formats, and to perform tasks such as formatting dates and times for display, calculating differences between dates, and parsing dates and times from strings.

* **Pyrebase:**



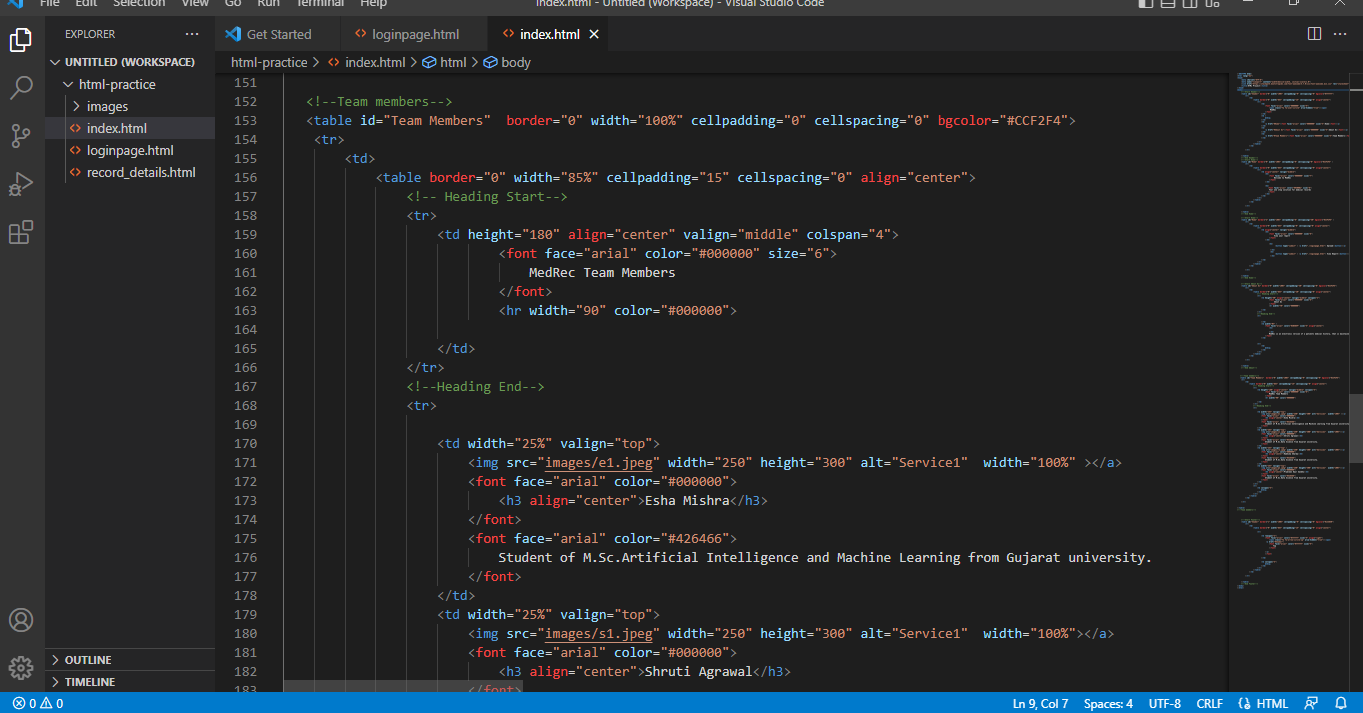
Pyrebase is a Python library that provides a simple interface for interacting with the Firebase Realtime Database. Firebase is a cloud-based database service that allows you to store and synchronize data in real-time across multiple devices and users.

## Project Design



## Project GUI

A graphical user interface (GUI) is an interface through which a user interacts with electronic devices such as computers and smartphones through the use of icons, menus and other visual indicators or representations (graphics). A website is proposed for the same in this project. The code of the website is written in html language.



A screenshot of a computer

Description automatically generated

# Chapter 6

# Data Analysis

## 

## 

## Data Collection

## Capturing digital images of medical records. The images should be clear and taken in good lighting. These images were uploaded to google drive.

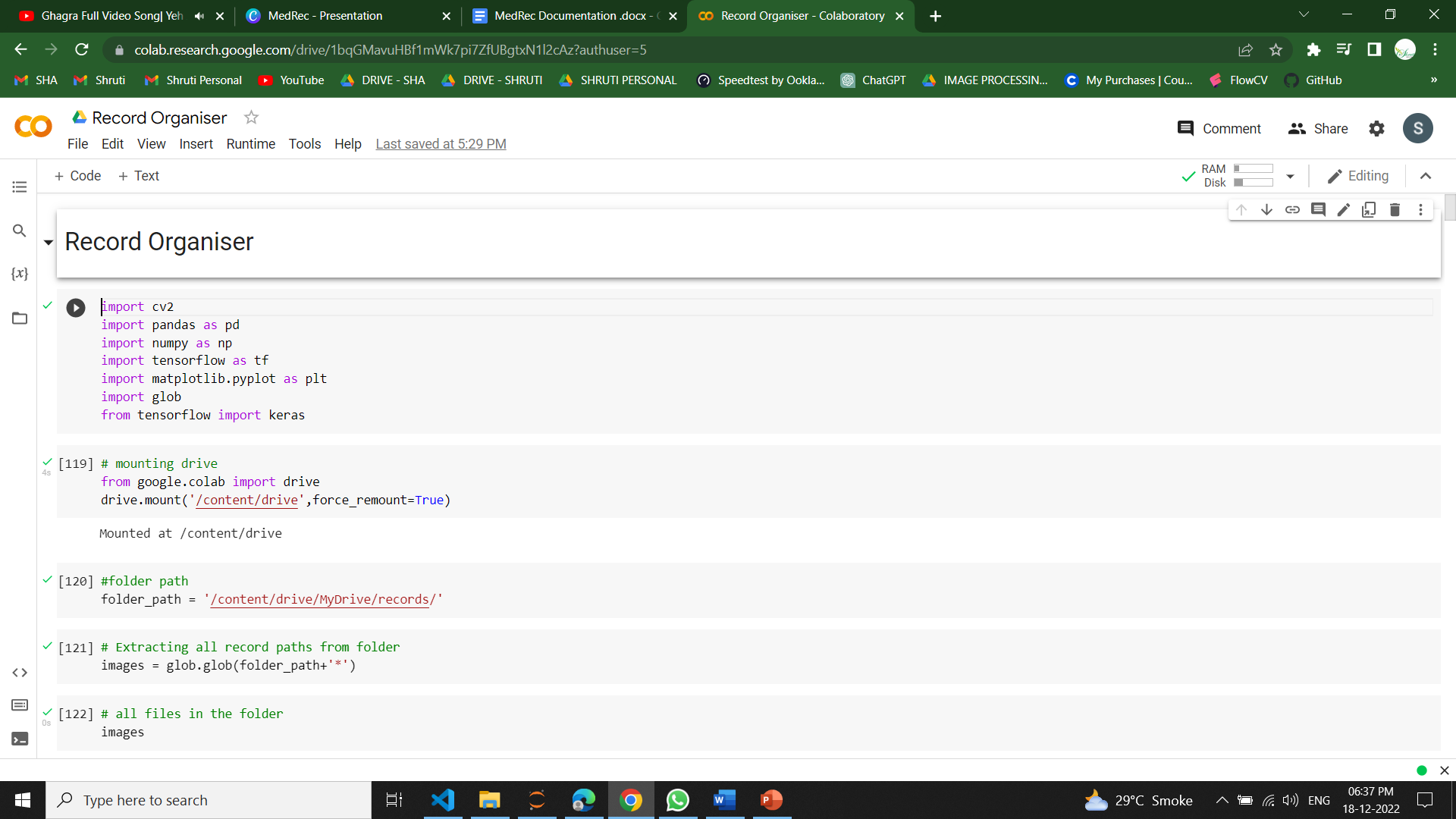
## 

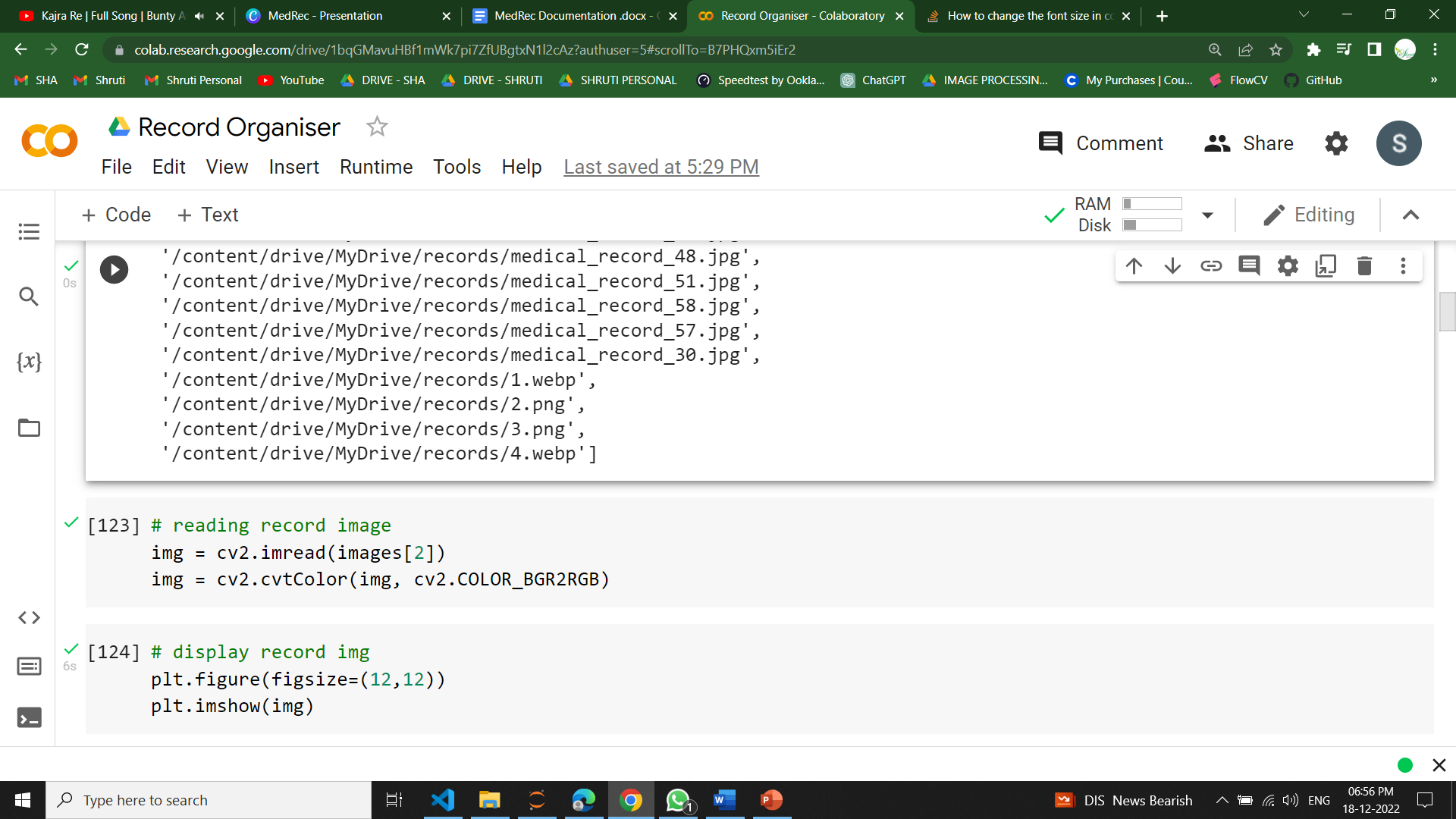
## About Data

## The data was collected by using medical reports of team acquaintances with consent.

## Training & Testing

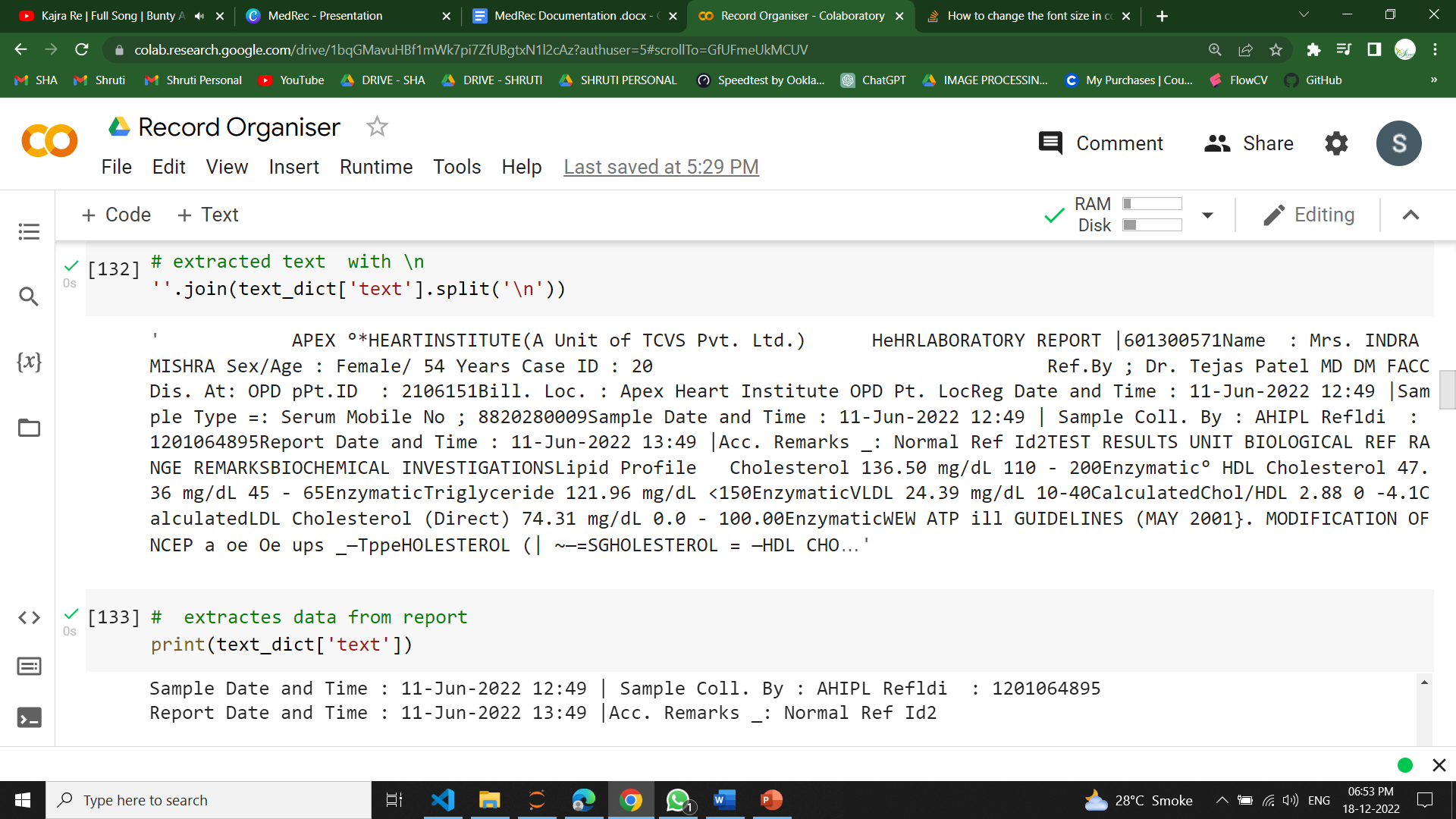
* The first step is to read all the record names from the given folder using glob, These record names were later used to read record images from the drive using cv2.



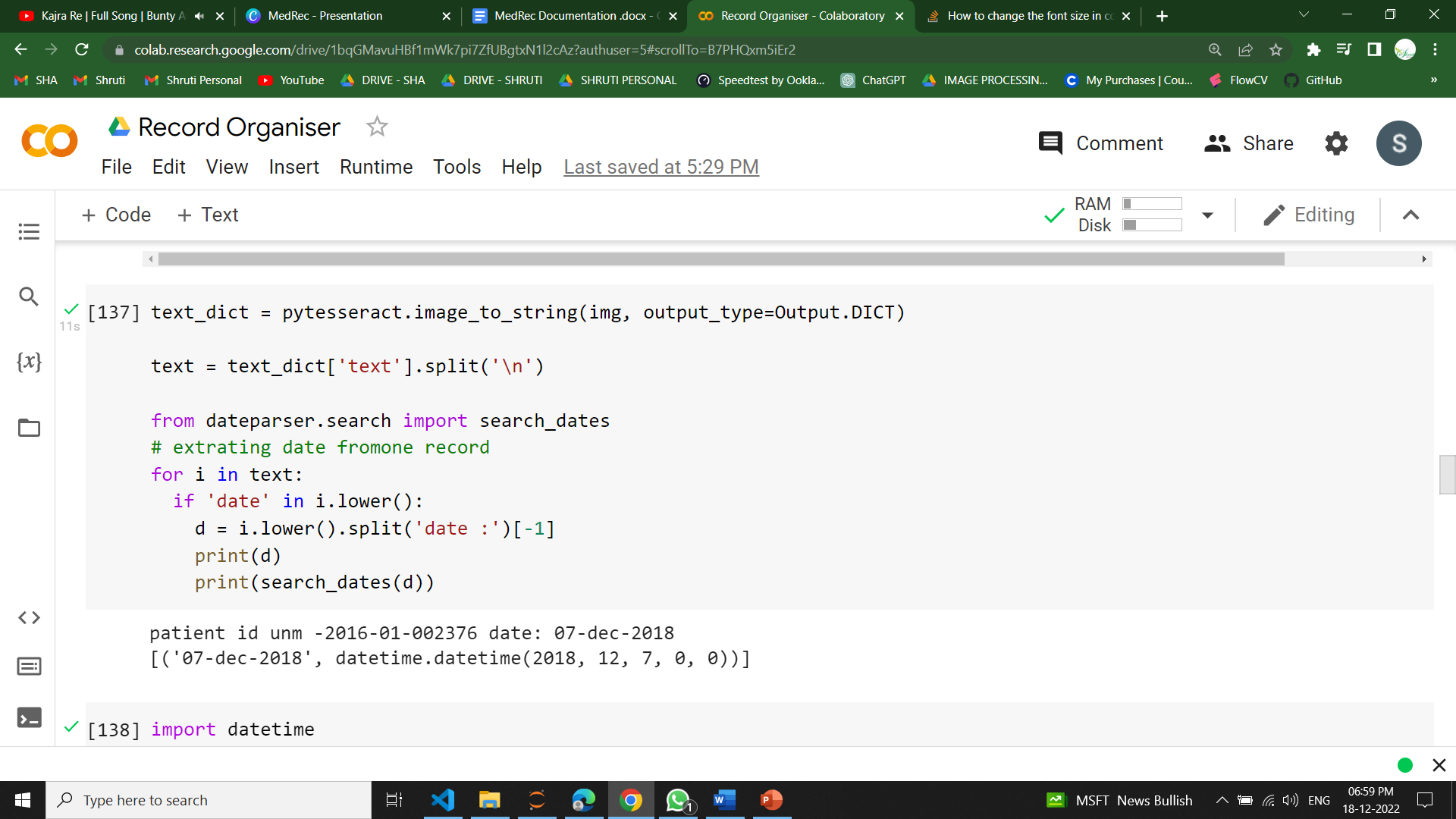


* Pre-trained model from pytesseract was used to perform OCR on images.

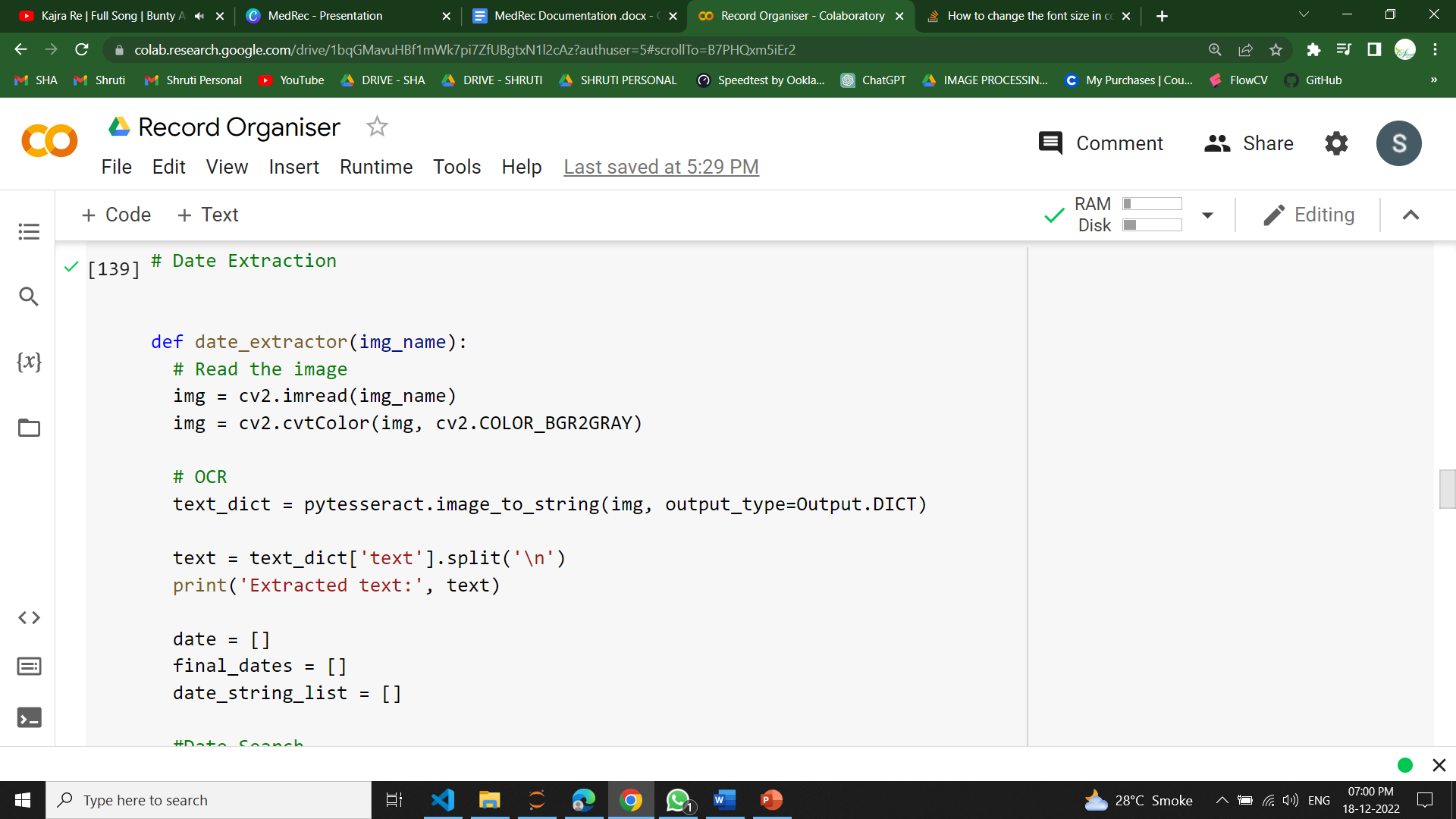




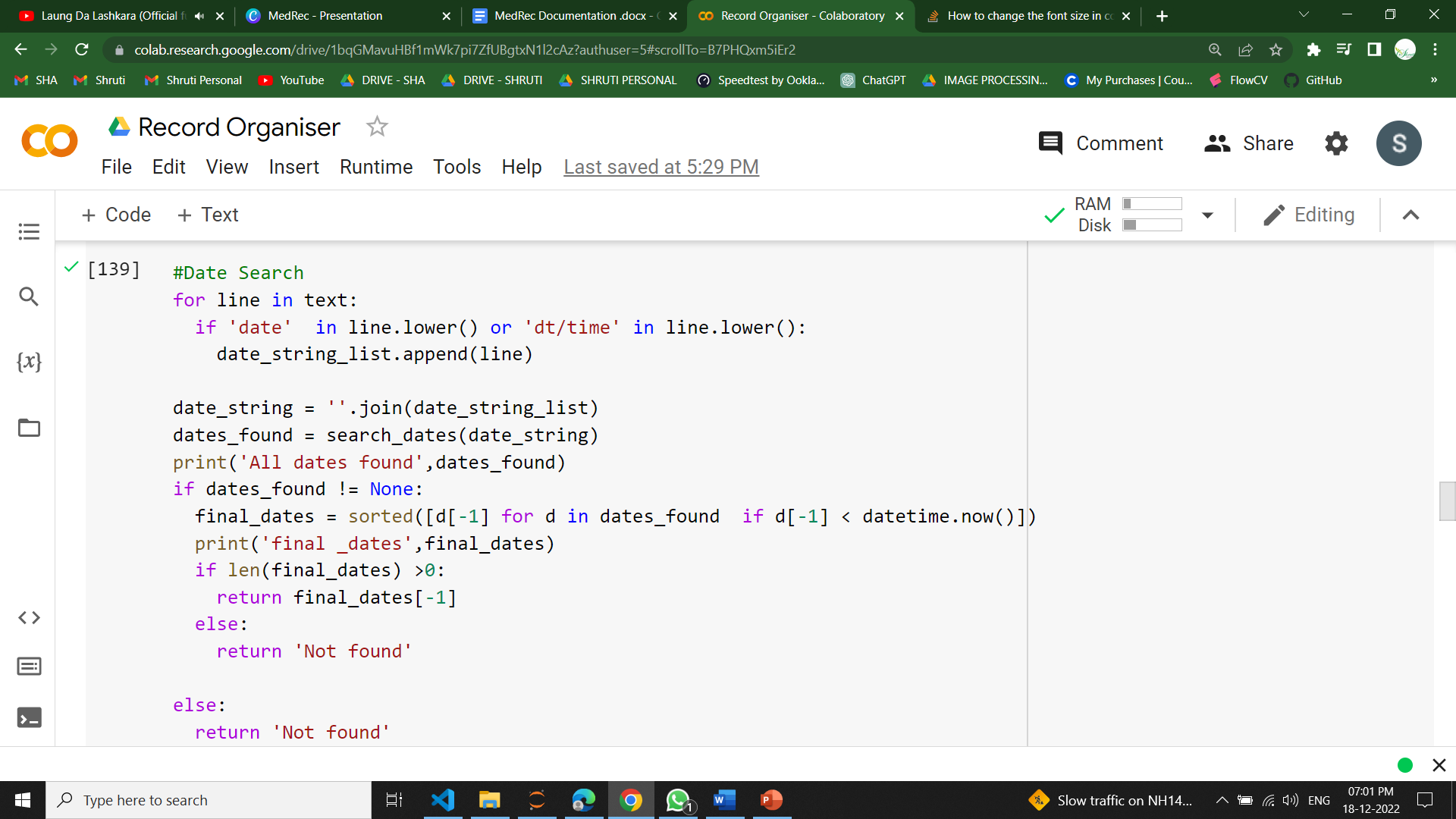
* ‘If’ statement was used to identify lines with the ‘date’ word from extracted text. These lines were used to extract all possible dates using search\_dates() from the dateparser module.



* This was followed by filtering out invalid dates. All dates greater than the present date were removed.



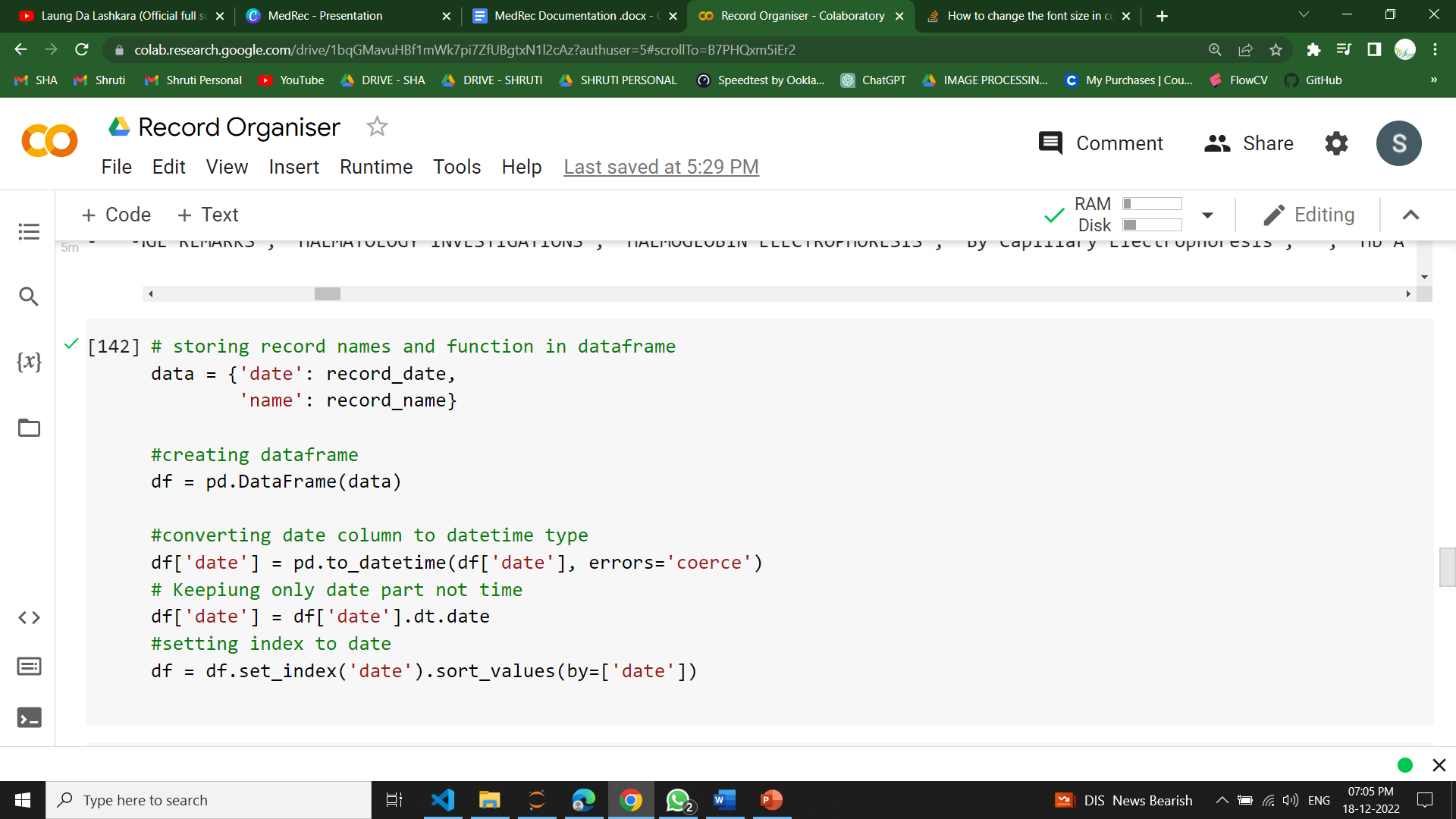
* Final list of dates was sorted and the most recent one was chosen.

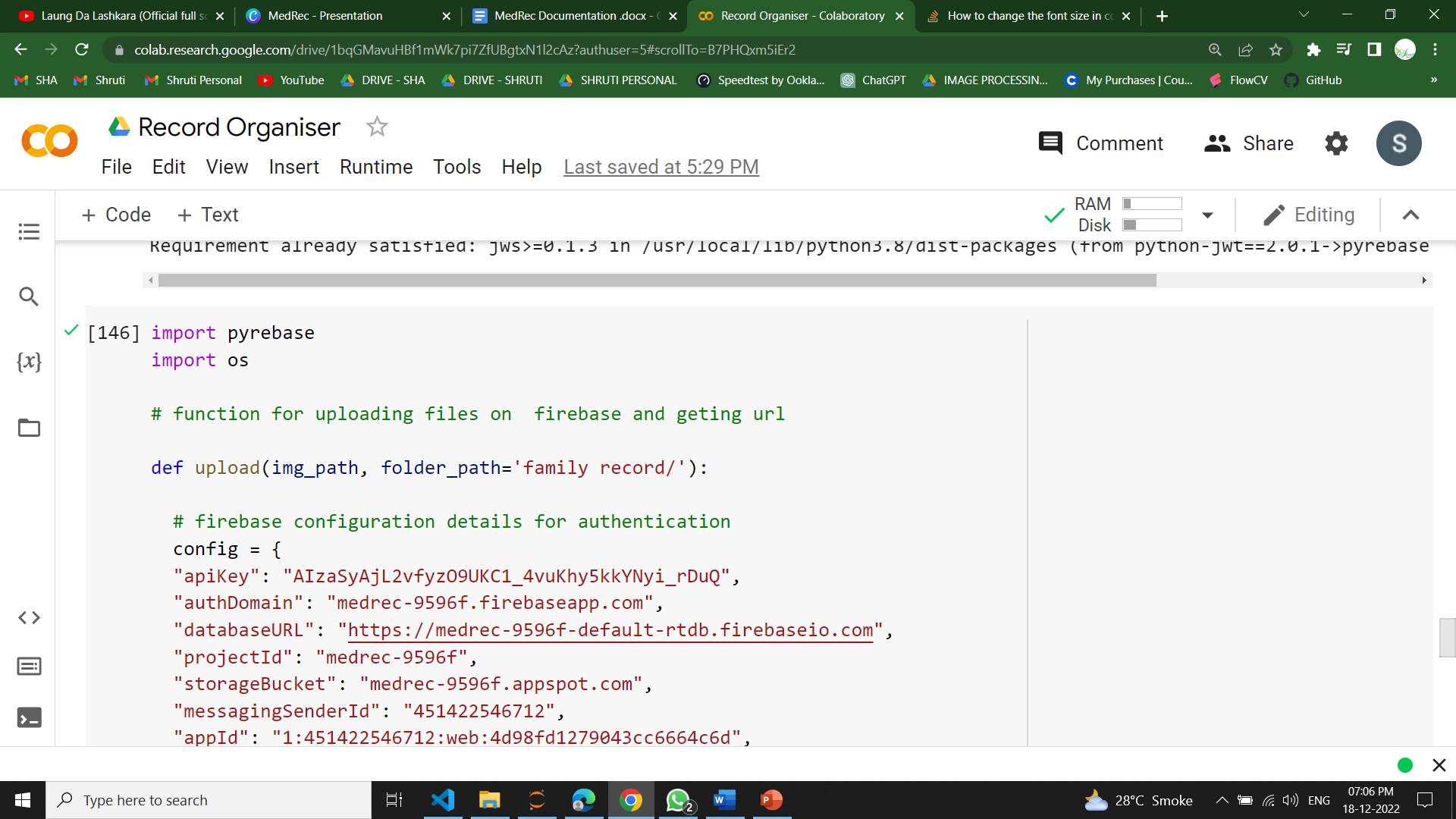


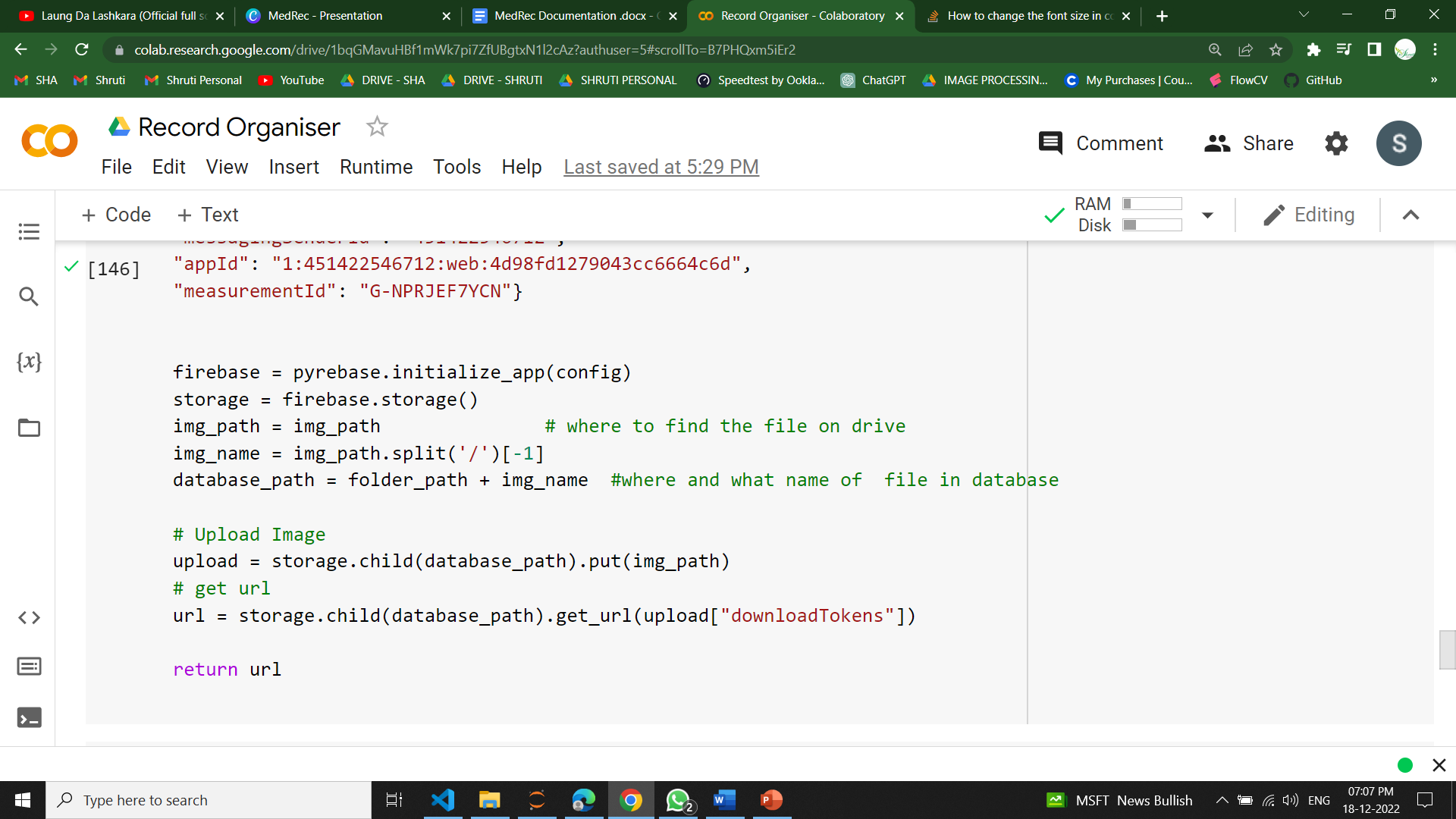
* The record names and their extracted dates were added in a pandas dataframe. The date column is converted to a datetime object and only dates are saved, not time.

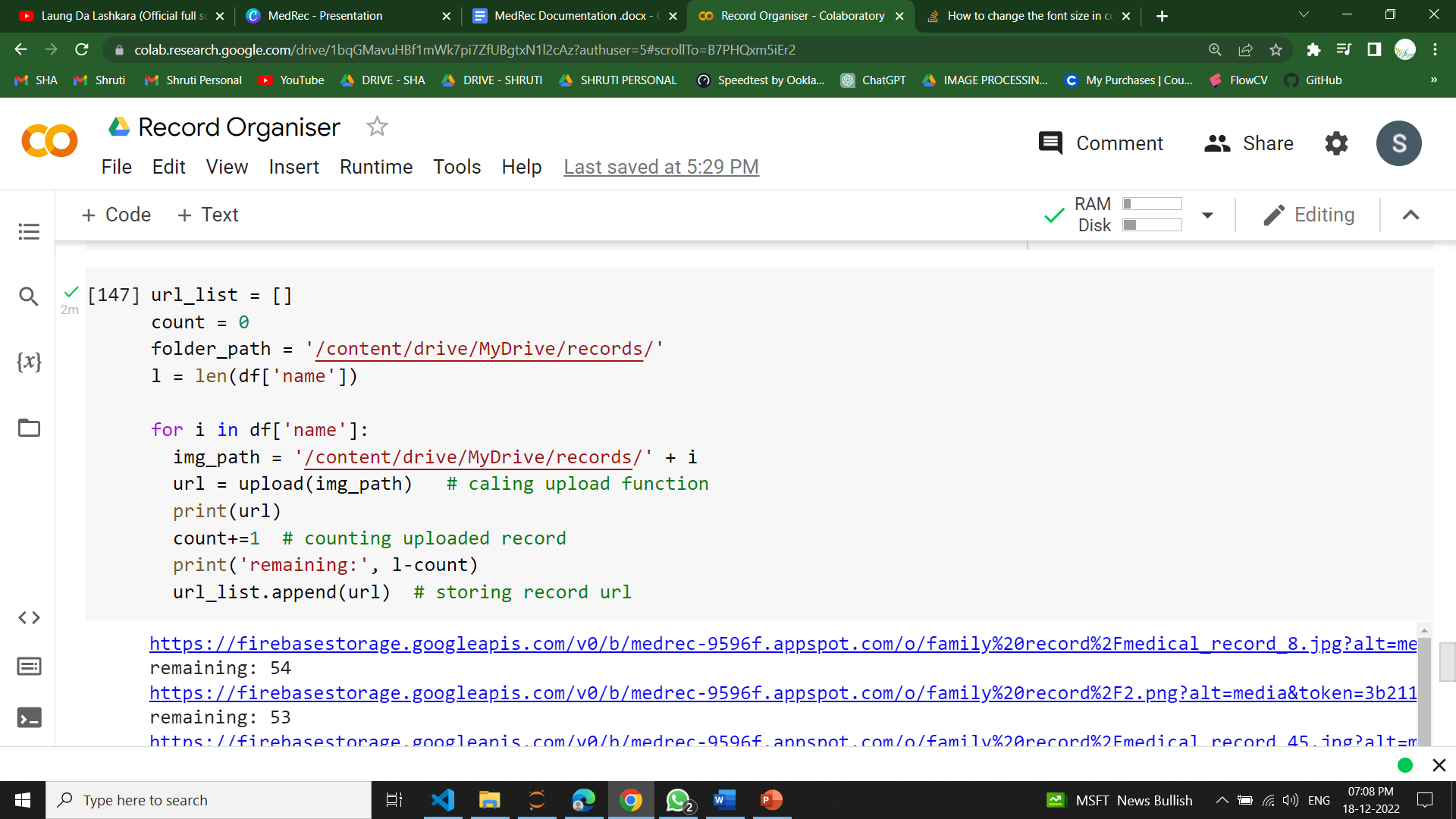


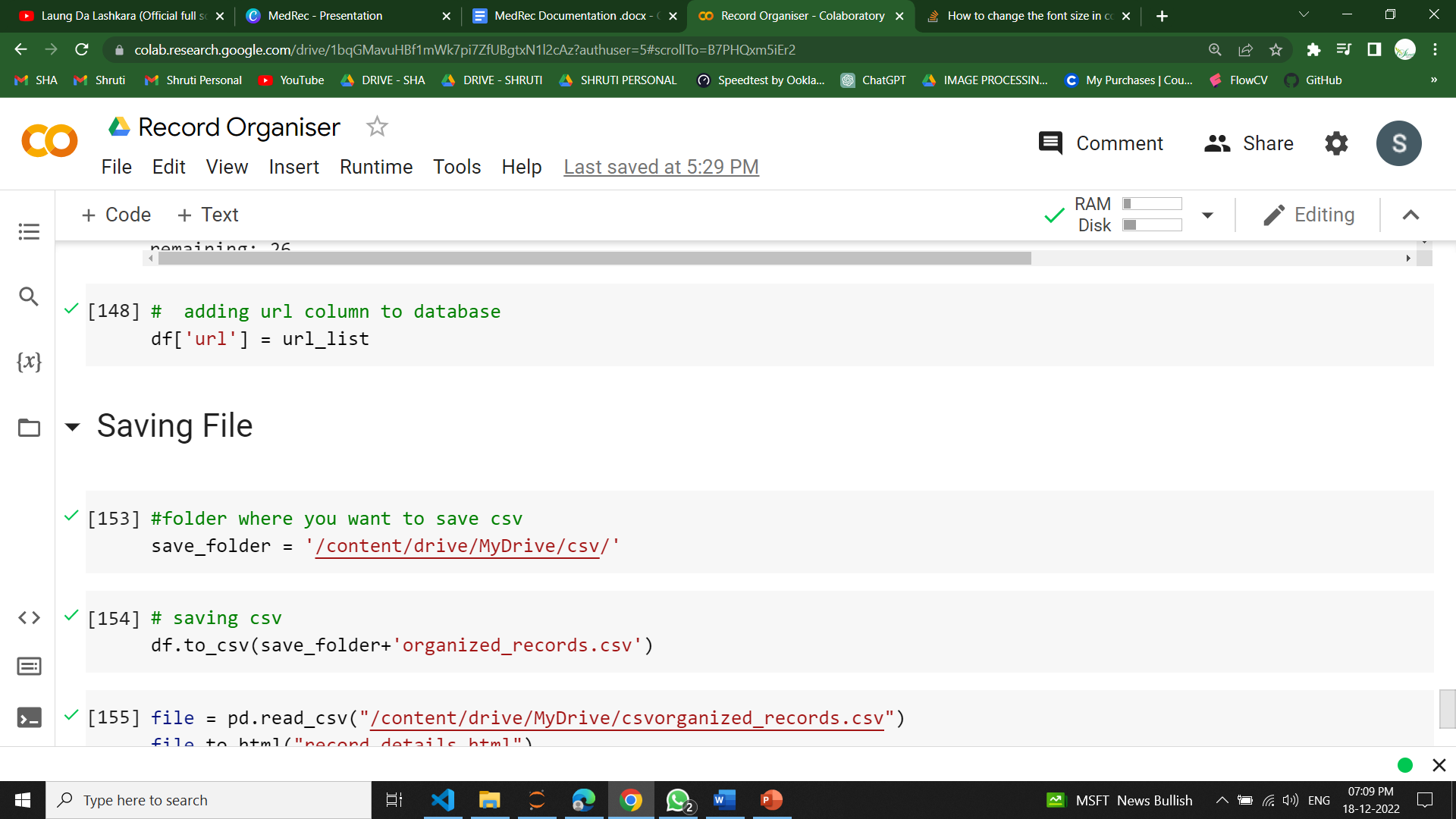
* All these files were uploaded on firebase and their urls were added in the dataframe. This date frame was saved as a csv file.











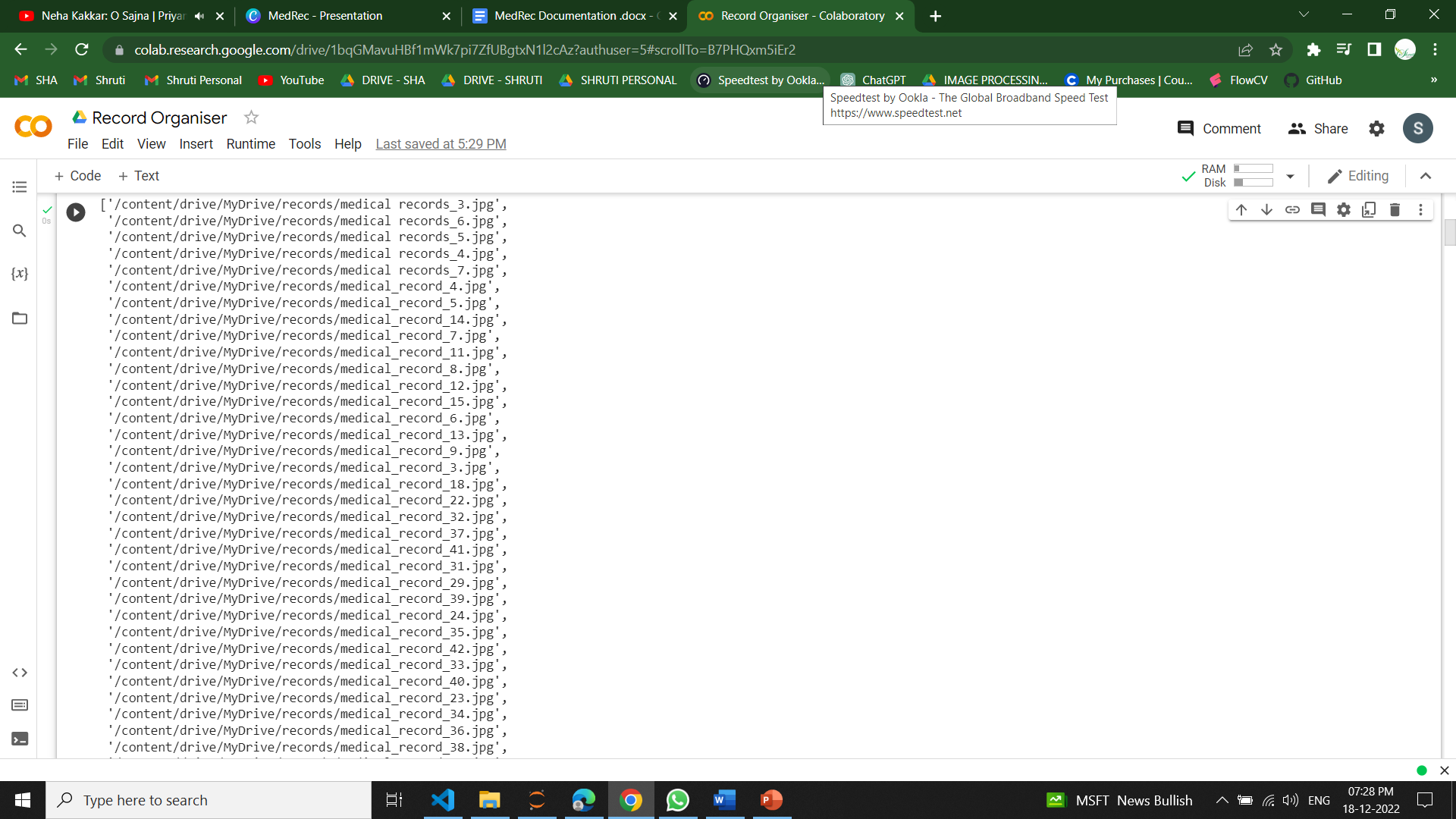
* One can open the csv file to access any record of a given date and can use the url to view the original report.
* This csv file was then converted to an html file and then displayed on the html website.

# 

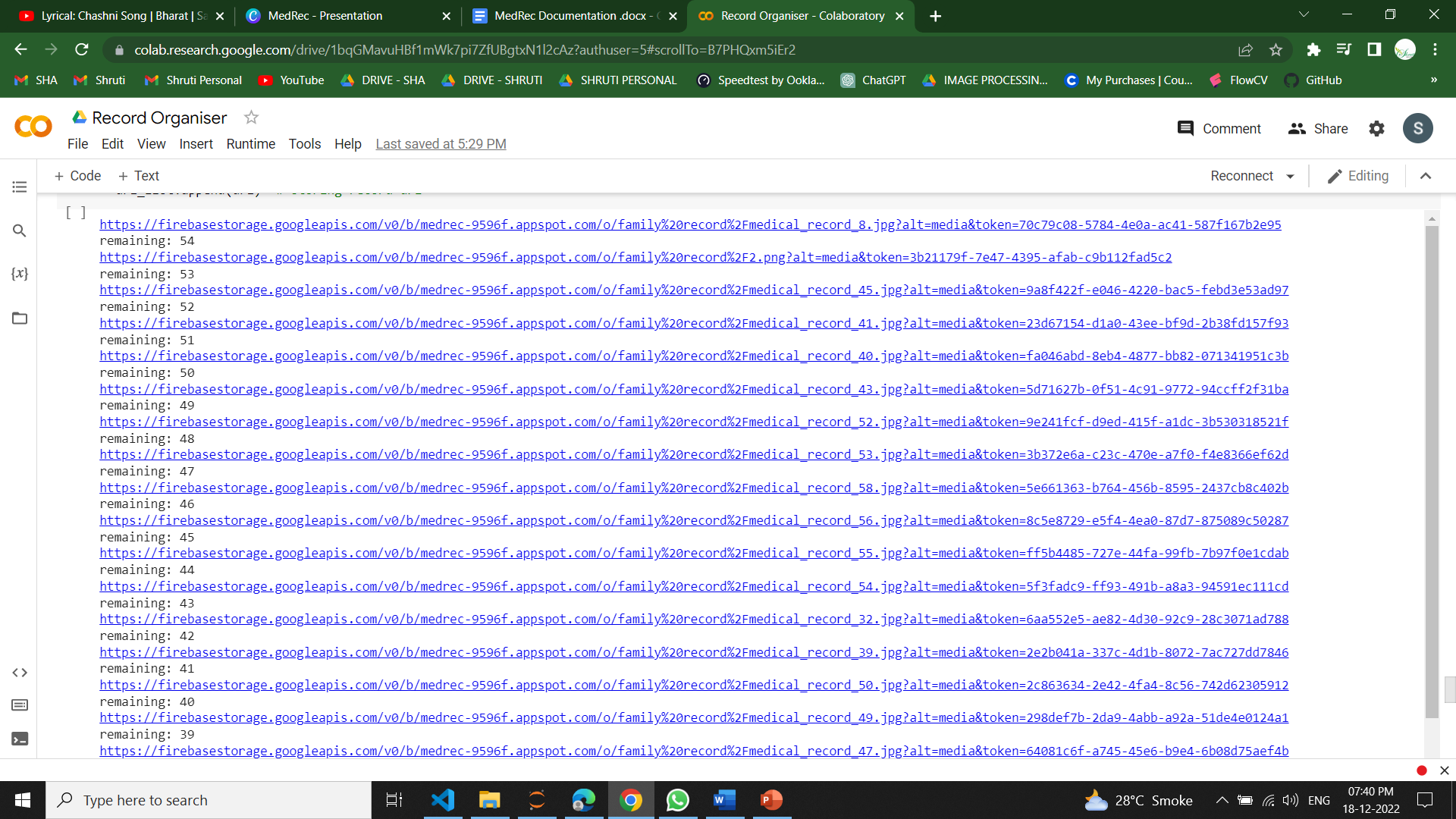
# Result & Discussion

* The output is a csv file with record date, name and url.
* Date helps in organising the documents and url provides easy access to the record image.
* Date extracted is not always accurate and can be improved. For best results the digital image of the document should be clear.
* Here is the glimpse of result:

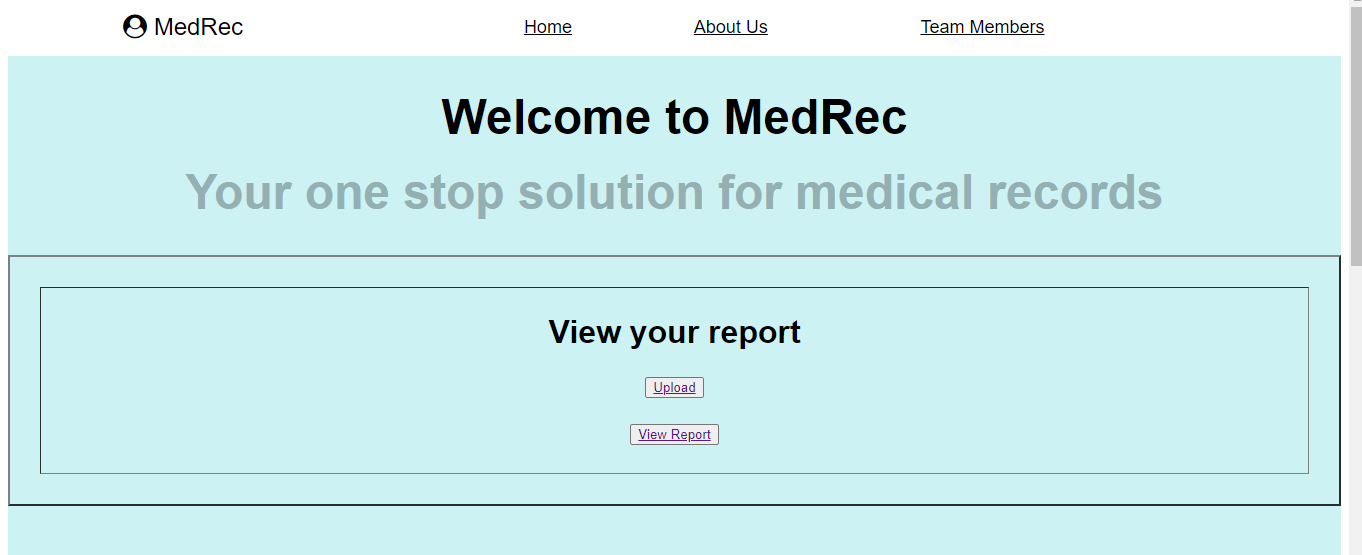
Reading the data



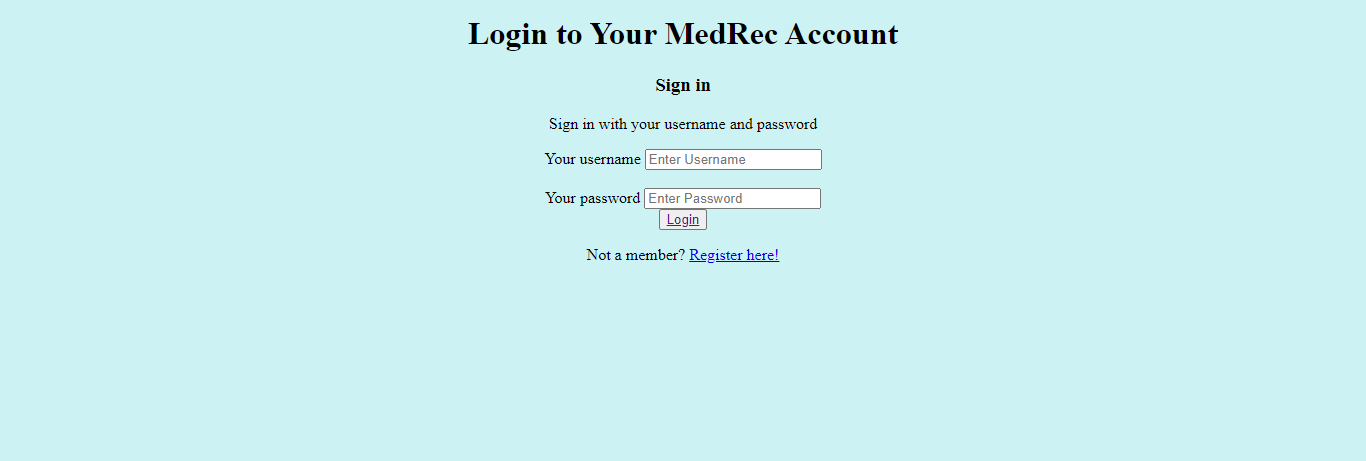
Extracted URL of Uploaded images



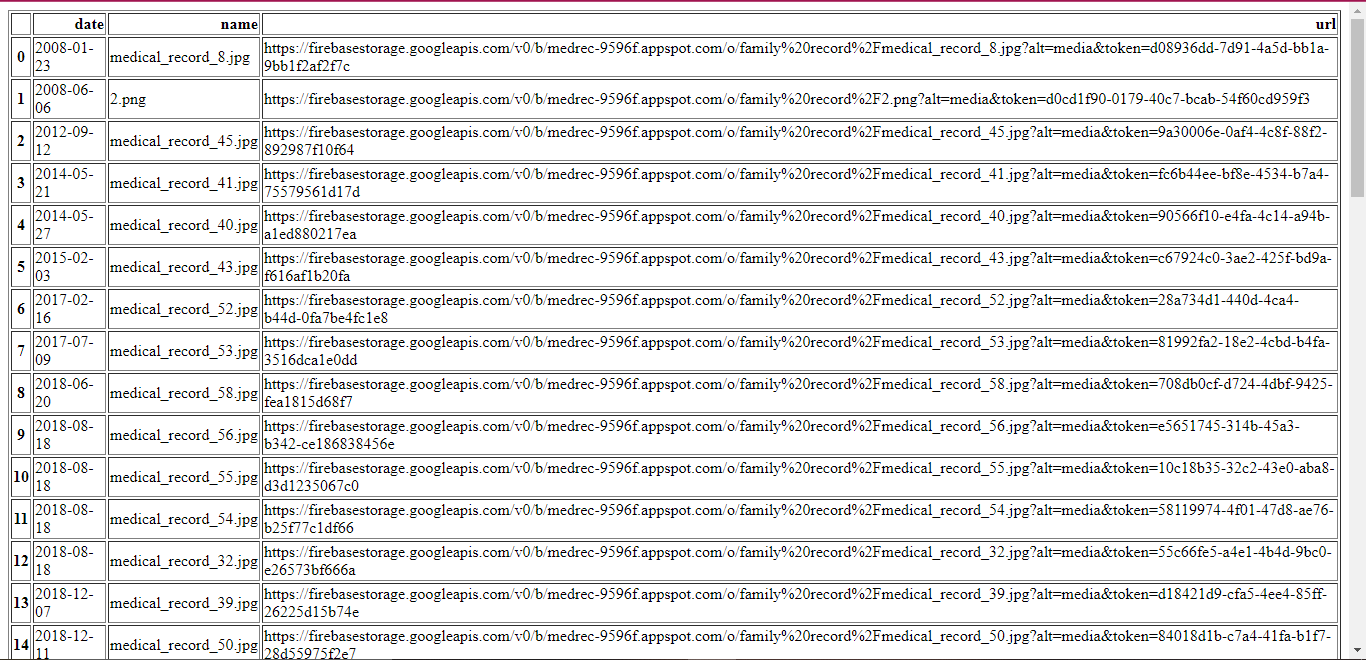
Frontend



Authentication Page



Displayed CSV File



* This is a static solution.

# Conclusion

Medical records are very important lifesaving documents. Maintaining them is a laborious task. Archiving is used for long term information storage. It has two parts to it. 1.Storing the information so that it is never lost. 2. Easy access to information. Our solution uses firebase by google for storage. The extracted date from medical record images is used to organize the records and save them in a csv file where you can access the record using the given URL. This will help users to store data in the app itself and thus reduce the storage cost and burden of maintaining data physically.

We aim to develop an application based solution for this problem. The user will

# Future work

* Improve accuracy of date extraction function.
* A dynamic system that works in real time.
* Additional authentication for privacy and security

# Bibliography

Bajaj Finserv Health. (2022, March 09). *Unified Health Interface: Services, Benefits & Registration*. Bajaj Finserv Health. Retrieved December 18, 2022, from https://www.bajajfinservhealth.in/articles/unified-health-interface

*Data Archives and Why You Need Them*. (n.d.). Cloudian. Retrieved December 18, 2022, from https://cloudian.com/guides/data-backup/data-archive/

*Improve patient experience with custom medical records management*. (n.d.). OSP Labs. Retrieved December 18, 2022, from https://www.osplabs.com/medical-records-management/

*MySirG-official/html-project: HTML only Page*. (2020, November 25). GitHub. Retrieved December 18, 2022, from https://github.com/MySirG-official/html-project

*Personal health records and patient portals*. (n.d.). Mayo Clinic. Retrieved December 18, 2022, from https://www.mayoclinic.org/healthy-lifestyle/consumer-health/in-depth/personal-health-record/art-20047273

*thisbe jim/Pyrebase: A simple python wrapper for the Firebase API.* (n.d.). GitHub. Retrieved December 18, 2022, from https://github.com/thisbejim/Pyrebase