Analysis of Healthcare-Stroke Data

Project\_Stroke\_Spring\_2023

Dharani Kurra Veera Poorna Prakash Reddy Chekkera Rajashekar Reddy Devireddy Krupa Sagar Reddy

16334850

ABSTRACT

Today, data analysis is important in business, research, metrology, and many other sectors. The retrieved information from the databases aids in making meaningful decisions, presenting research articles, forecasting weather, and many other tasks. We can collect a large amount of data, which helps us to derive relevant findings and make sound business decisions. However, as the volume of data grows, it becomes more difficult to interpret and study the data. Visualizations, when utilized properly and appropriately, may be powerful techniques in exploratory data analysis. Visualizations may also be used to convey the information to the target audience or to educate them about the results. This Stroke Data Analysis will show us how to do exploratory data analysis on stroke-related datasets using Pyspark. Due to its strength in managing big data, Pyspark is a particularly wonderful choice as our team is intending to work on a project with a huge dataset. The data set that we have taken consists of the following data **ID, bmi, region,marital\_status,hypertension,Stroke\_chances,Occupation,avg\_glucose\_level, Smoking\_status**. The considered dataset is of 200 mb size and consists of nearly 50 thousand rows with 9 Columns.

**Introduction**

In this project the dataset contains the patient data across various features. The entire data was collected from Kaggle. This is its entire collection since January 1, 2017. The dimensions of the dataset are 53000\* 9. The Region is from where the patient belongs to. The bmi is the patients’ body mass index. The smoking\_status refers to the smoking habit of the patient. The hypertension and stroke\_chances features refer to history of the patient. The avg\_glucose\_levels refer to the glucose level of the patient recorded. The occupation column refers to the occupation of the patient. The gender column refers to the gender of the patient. And finally using these variables or predictors we perform the analysis like main contributors to the chances of stroke in the patients.

Pyspark on data bricks when combined with a cloud platform like Azure enables us to store and create data pipelines and analytics that are more scalable. This lesson begins by exposing you to PySpark's capability for handling massive dataset analysis. You'll discover how to connect to Spark on Windows as a local computer and communicate with Spark from Python. In this project, we are going to attempt an investigation of the Stroke dataset, which is available on Kaggle. Furthermore, Stroke analysis gives us a clear picture of dominant features causing the stroke. We are attempting to analyze patients’ data using this dataset to acquire some beneficial insights about the features causing the stroke.

**Motivation/Purpose**

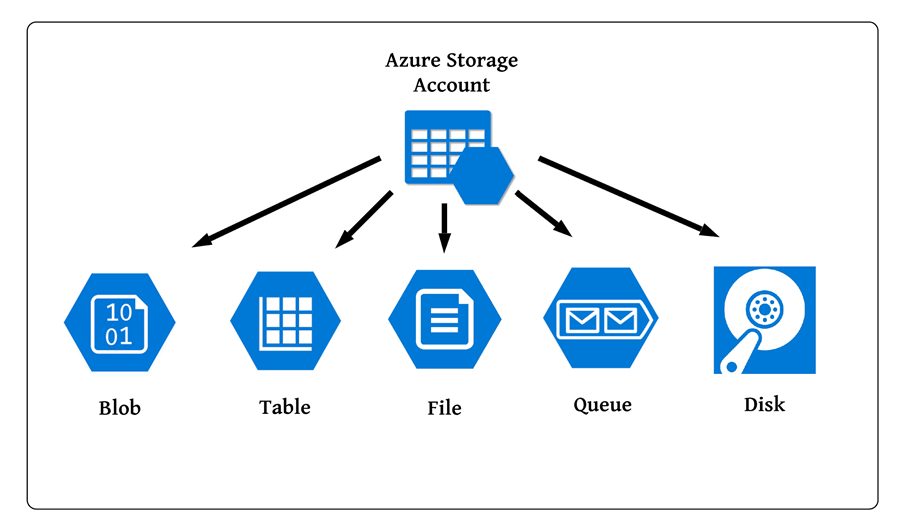
The main purpose why we chose to do this project is our interest in data analysis of some useful data. The data analysis on the healthcare data or mainly the stroke related data gives us the insight of the dominant features or causes for the heart stroke in the people. We tried our best to point out the major factors for the stroke in the people. The analysis we made can be further extended as the data comes in and may finally lead to general patterns in the patients and may be of great help to the people know of the reasons for the stroke. We hope this analysis can be put to great use to mitigate the chances of stroke by oppressing the factors causing it.

**System Architecture**

The main components involved in this project and their brief introduction are

**Azure Storage Account**

An Azure storage account contains all Azure Storage data objects, blobs, file shares, queues, tables, and disks. It provides a unique namespace and is accessible from anywhere in the world over HTTP or HTTPS. It is Durable, highly available, secure, and massively scalable.



**Azure Blob Storage/Azure Data Lake Gen2:** Blob storage is a part of Cloud Storage which is Optimized for storing massive amounts of Unstructured Data.

**Azure Data Bricks**

It is Apache Spark-based analytics platform optimized for Azure which is used for creating your notebooks and integrate them the cluster of you desired.



It has several features like optimized spark engine, machine learning run time, creation and Autoscaling the clusters, choosing the language of your choice like Python, R, Scala, SQL.

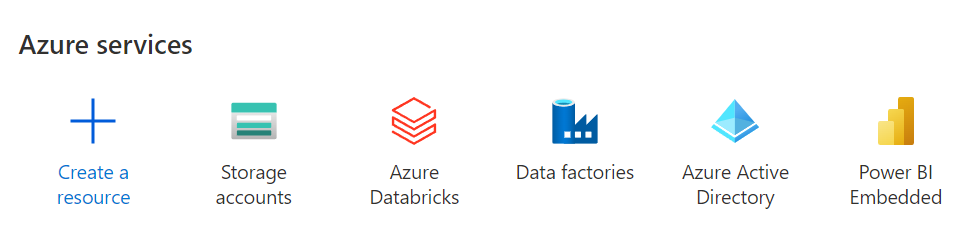
It has interactive workspace which lets you collaborate with other users like data scientists, data engineers and business analysts.

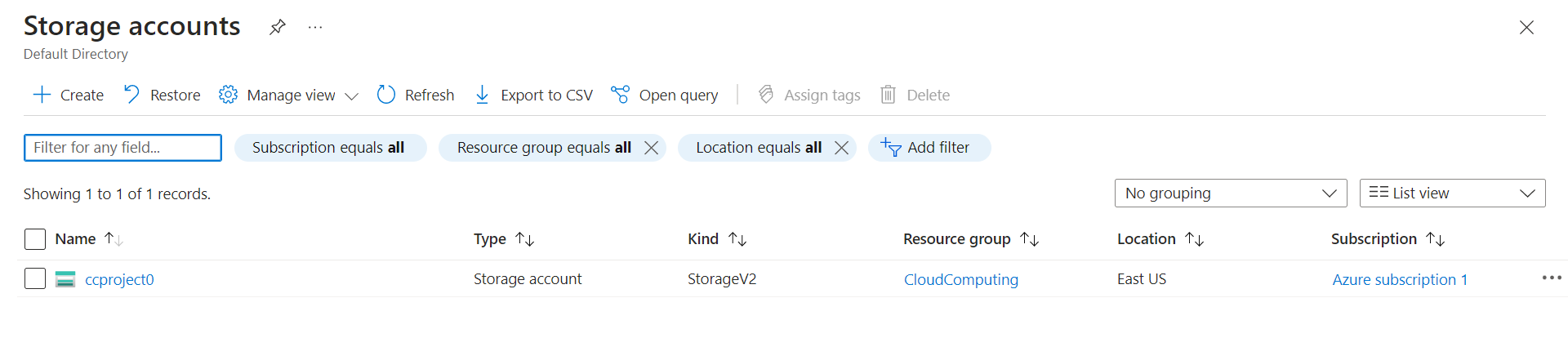
**Analysis and Related Work/ Features/Steps involved**

For the Stroke data analysis, we extracted data from Kaggle. We ensured that we get a large dataset with wide variations so as to get a clear picture about the general trend from the data. For the storage of the huge, extracted dataset from the Kaggle, we have selected to go with a cloud platform for one stop solution of storage and analysis purposes. The overall project comprises of the following steps

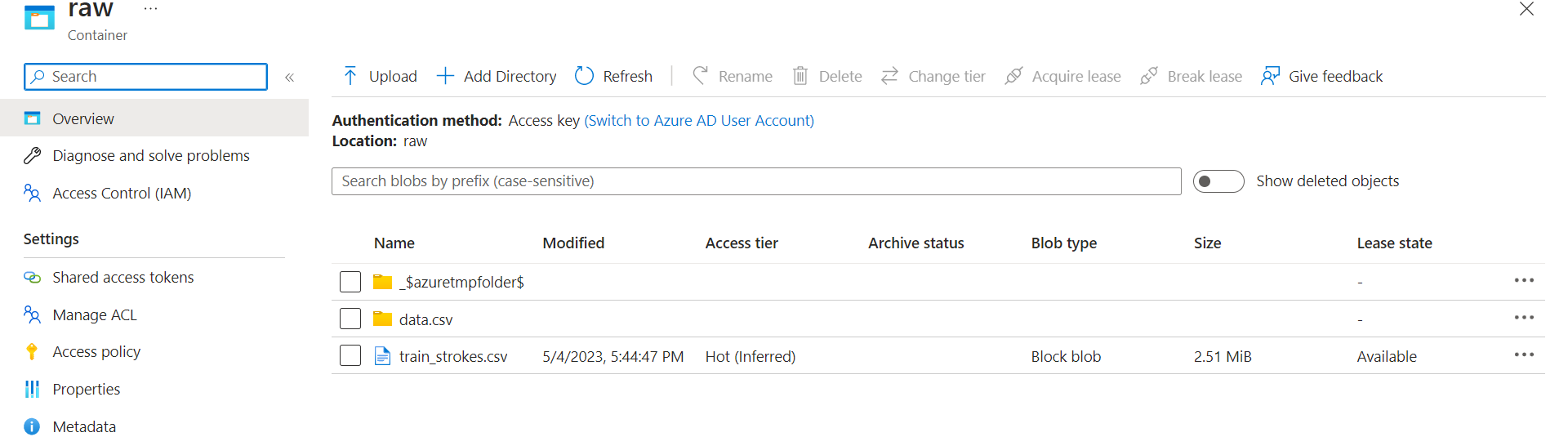
1. Loaded the complete dataset into Azure Blob Storage.
2. Integrating Azure Storage with Azure Databricks.
3. Creating of Notebooks and workspaces.
4. Reading the data from Blob Storage and cleaning and transforming the data.
5. Creation of Hops for optimised data retrieval and data processing.
6. Automated the process of data loading into azure using a pipeline built using Azure data Factory. We considered ADF as data needs to be continuously migrated. Also, ADF helps in migrating data from on-premises to the cloud easily.

The folder structure in Azure blob storage looks like

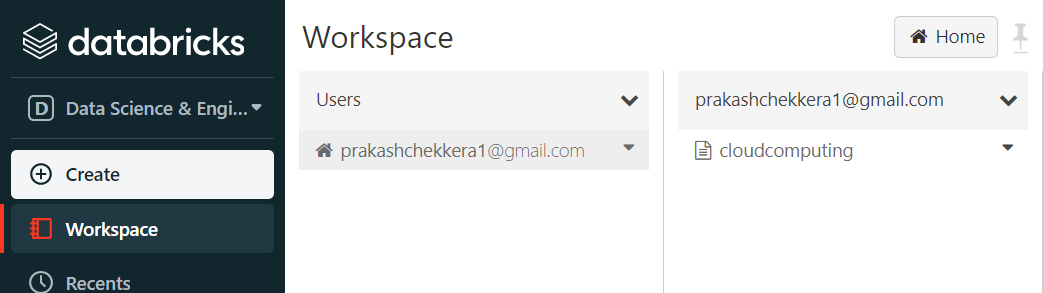


****

We made use of an Azure storage account to store the data. The account mentioned above serves as a storage location for our data files, data tables, and data objects. Using an HTTP connection, we may access the Azure storage account from anywhere. We utilize Azure Blog storage for cloud storage. Unstructured data can be kept in blog storage.

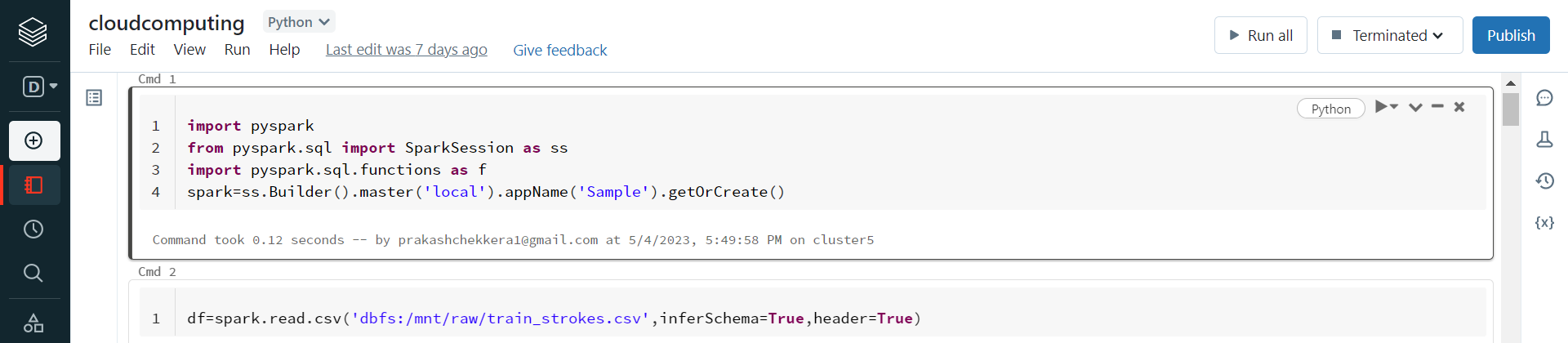
****

We are using Azure containers to carry out a variety of tasks in this without environment management or dependencies. advanced data analysis.



We have developed advanced ETL procedures that convert data using visual data flows or compute services like Azure HDInsight Hadoop, Azure Databricks, and Azure SQL Database.

Furthermore, we leverage modified data by disseminating it to data repositories like Azure Synapse Analytics so that business intelligence (BI) applications may consume it. In the end, Azure Data Factory may enable the organization of unstructured data into practical data warehouses and data lakes for more informed business choices.

****

1. Regarding the analysis, we created a cluster and a notebook in the Databricks platform provided by Azure. We mounted the Blob storage in Databricks to import later for the analysis part and to store back the cleaned dataset so as to make way for visualizations using PowerBI and we also captured some visualizations using Matplotlib of Python Package as projected below.
2. In the cleaning process of the data, we made these changes to the dataset.
3. We removed the null and inconsistent values in the data and we even made manual corrections to the data.

Proposed Techniques

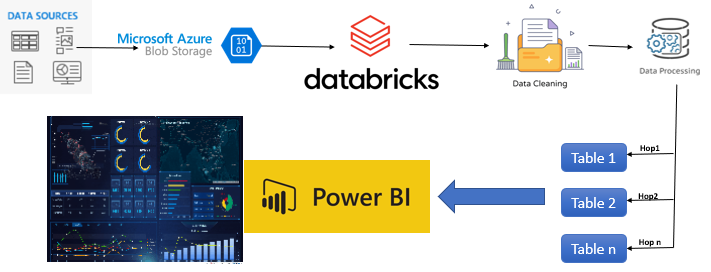
Big data demands a service that can coordinate and operationalize processes in order to turn these enormous amounts of raw data into actionable business insights. The managed cloud platform Azure Data Factory is ideal for these difficult projects that include extract-transform-load (ETL), extract-load-transform (ELT), and data integration. A Spark cluster in the cloud (Azure HDInsight) will analyze the linked data to extract insights, and the transformed data will be delivered into a cloud data warehouse like Azure Synapse Analytics so that a report can easily be constructed on top of it. They want to automate, control, and continuously monitor this workflow. Additionally, they need it to launch each time a file is deposited into a blob storage container.

The platform that handles various data streams is Azure Data Factory. You may develop data-driven processes for coordinating data movement and converting data at scale using the cloud-based ETL and data integration service. You may design and plan data-driven processes (also known as pipelines) that can ingest data from various data repositories using Azure Data Factory. You may build complex ETL processes that convert data using visual data flows or compute services like Azure HDInsight Hadoop, Azure Databricks, and Azure SQL Database.

Additionally, you can publish your converted data to data repositories like Azure Synapse Analytics so that business intelligence (BI) applications may use it. In the end, Azure Data Factory may arrange raw data into practical data warehouses and data lakes for better business choices.

For data storage, we can use an Azure storage account. In this storage account, we can access our data objects, files, and data tables. We can access the Azure storage account anywhere with HTTP link. For cloud storage we can use Azure Blog storage. We can store unstructured data in blog storage. In this, we are using Azure containers to run a bunch of jobs without any environment management and without any dependencies. For advanced data analysis,

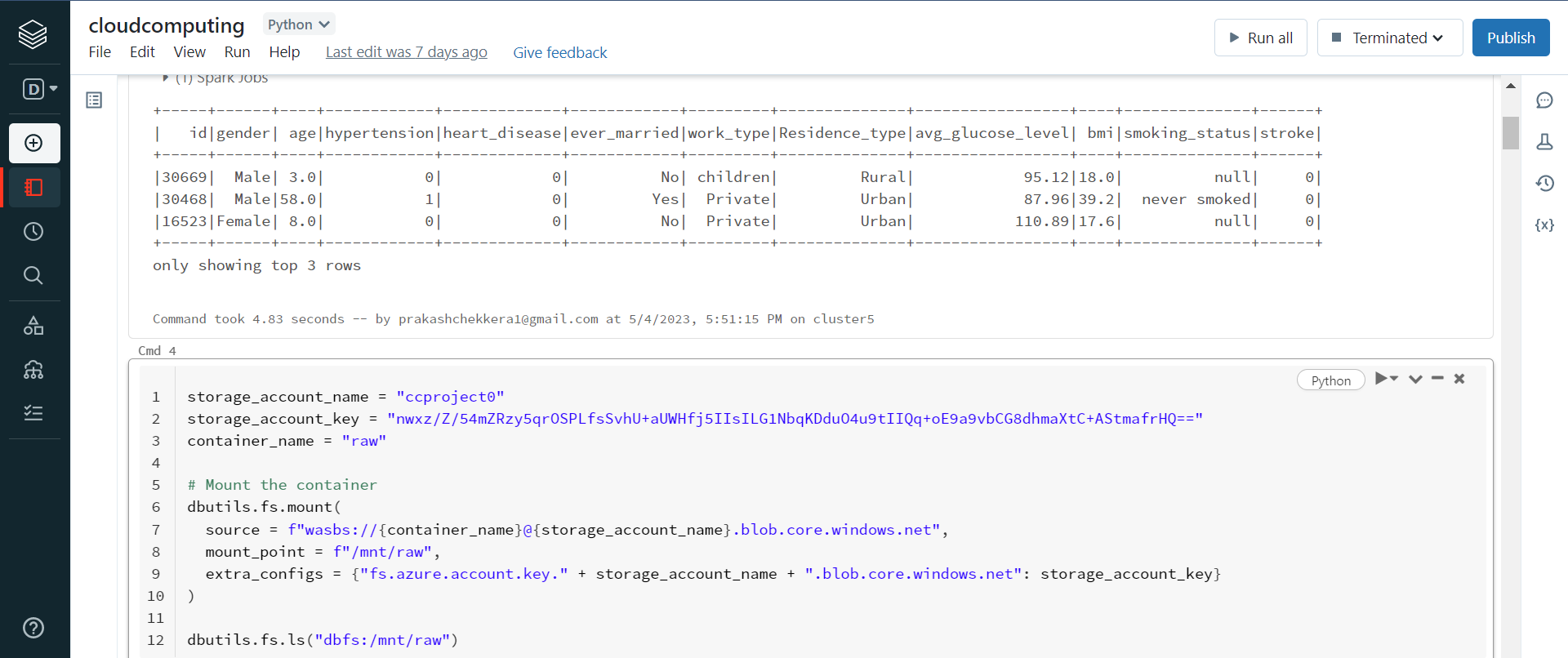
we are using Power BI (a tool for data analytics). We can establish the connection between Azure data bricks to Power BI using some connectors and we can create some advanced visualizations.



BIG DATA and CLOUD TECHNIQUES USED

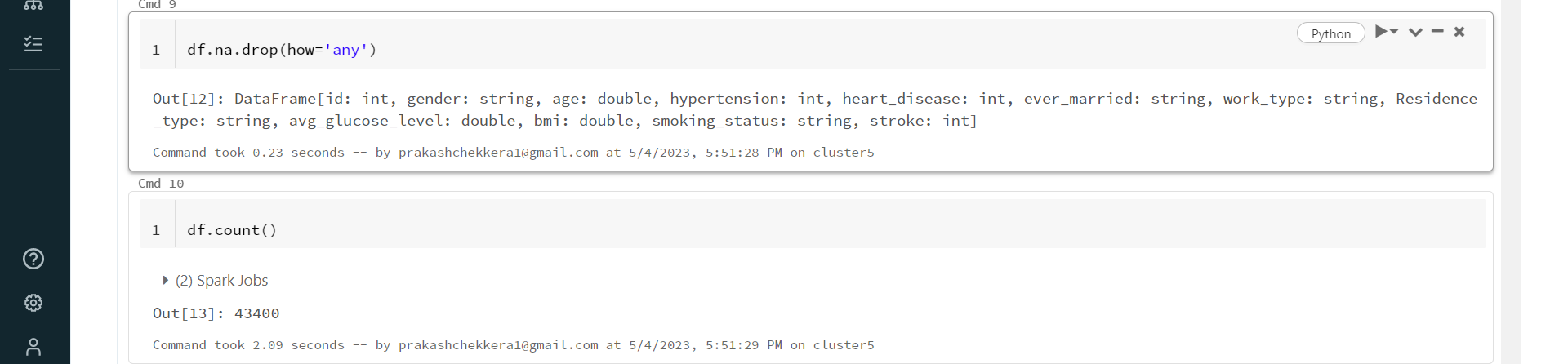
DATA PRE-PROCESSING

Fig 1 shows how dataset is mounted in Databricks from the resources created in Microsoft azure.



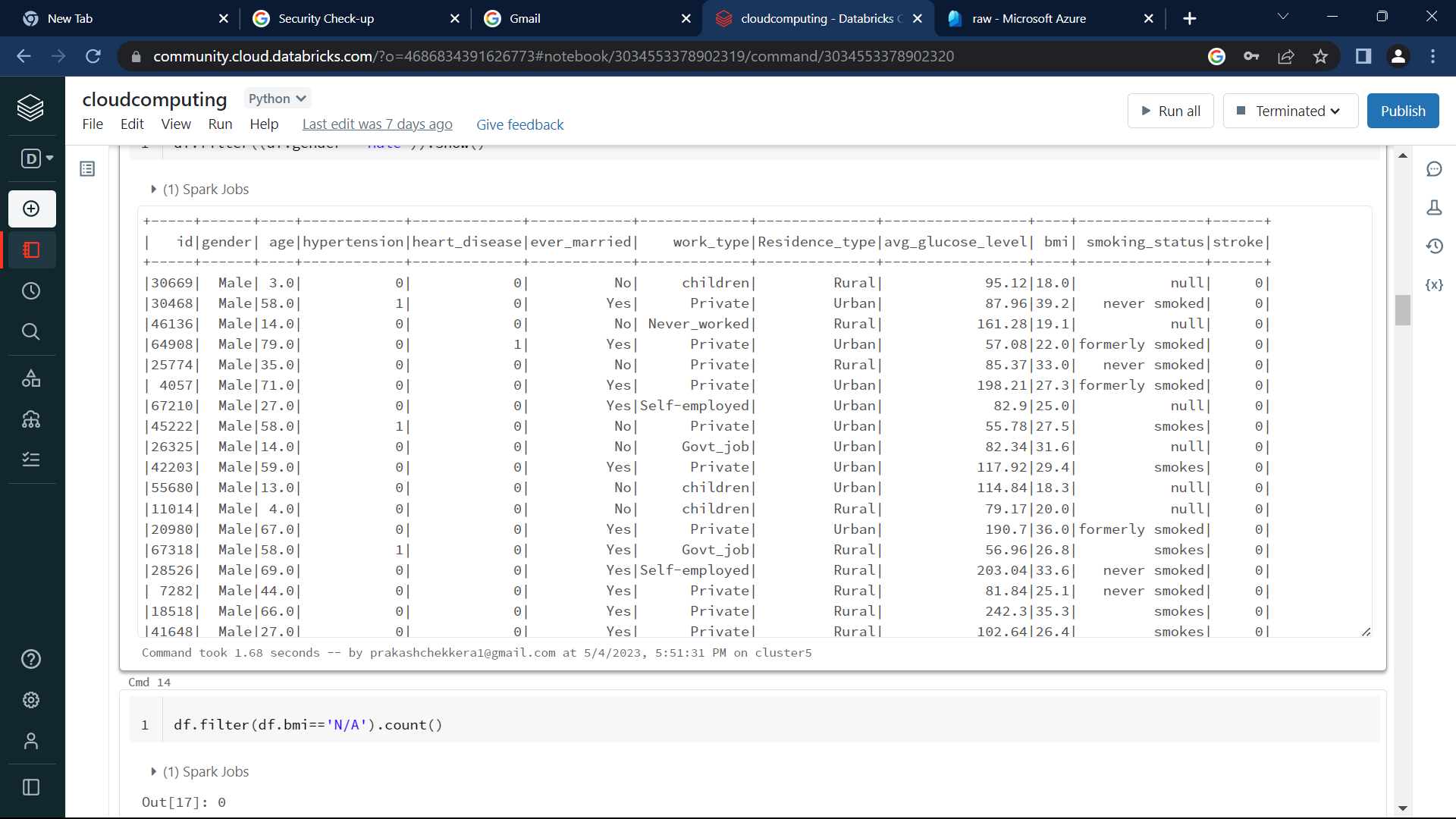
**Fig 1**

Fig 2 shows the datasets that has its null values replaced.



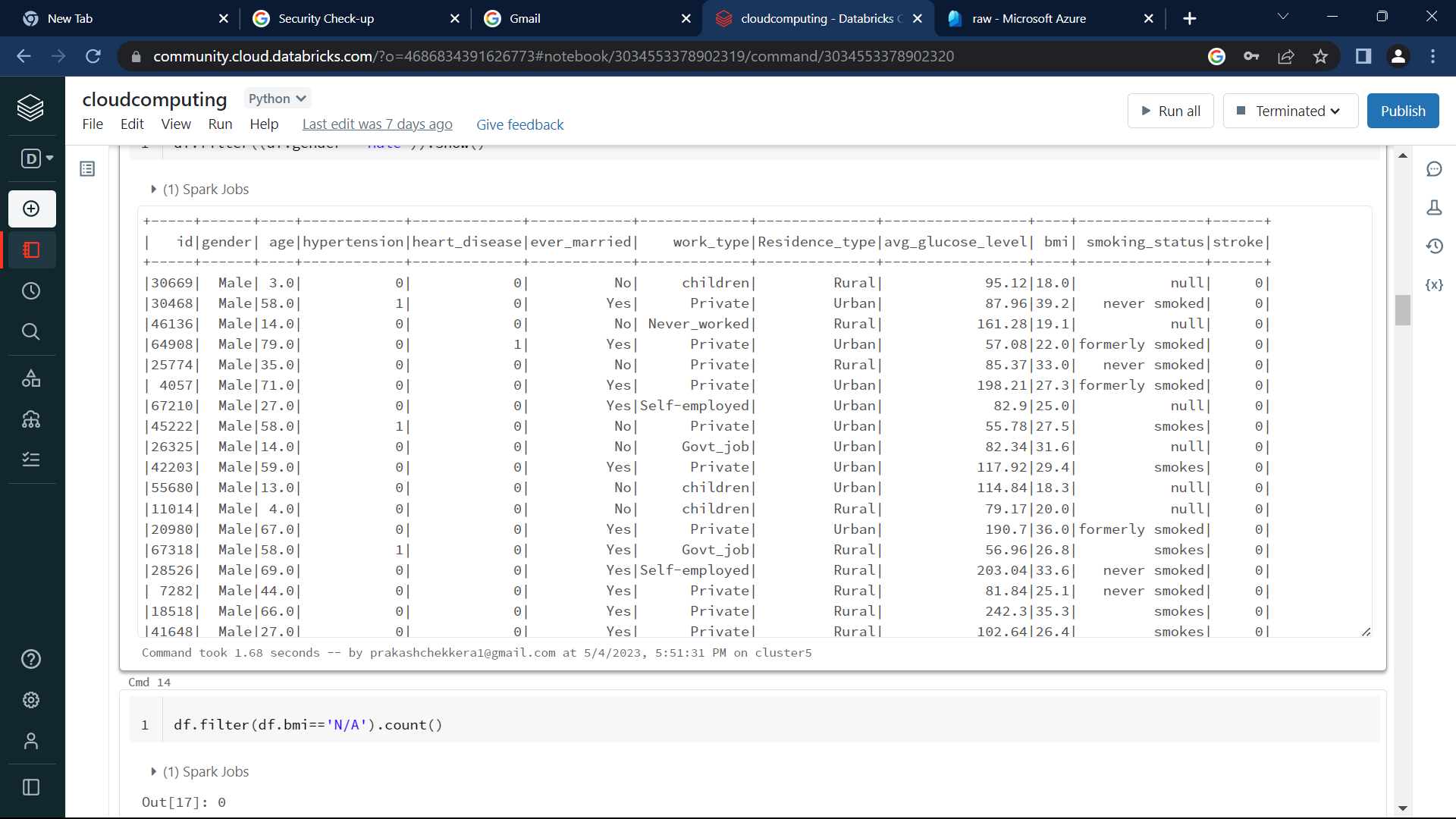
**Fig 2**

After completing data cleaning, data is processed. Fig 3 shows the processed data that is sorted accordingly.



**Fig 3**

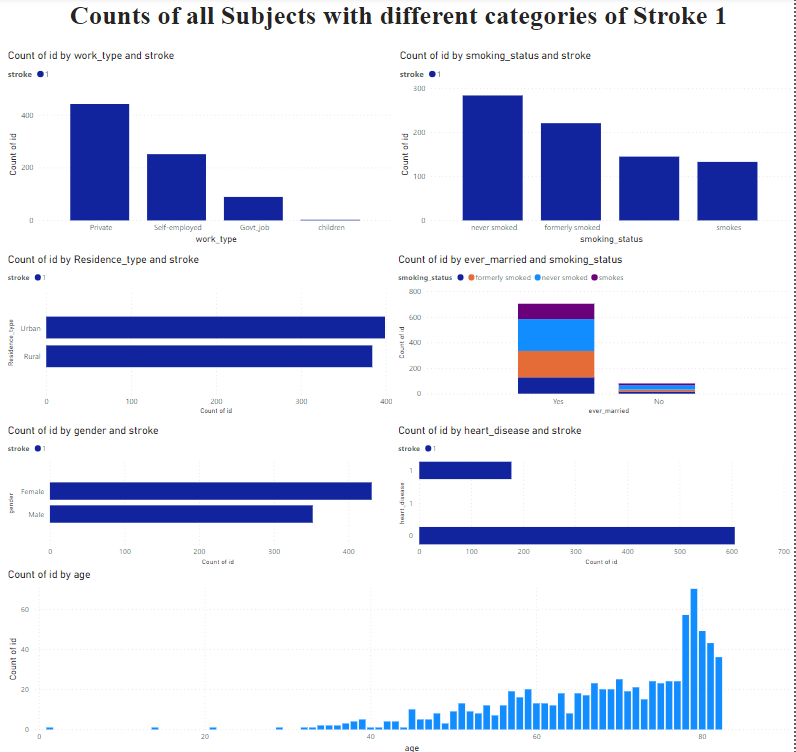
Fig 4 shows the processed data that shows how cleaned data looked like



**Fig 4**

The following Fig 5 and Fig 6 shows the initial visualizations are plotted using matplotlib and seaborn packages.

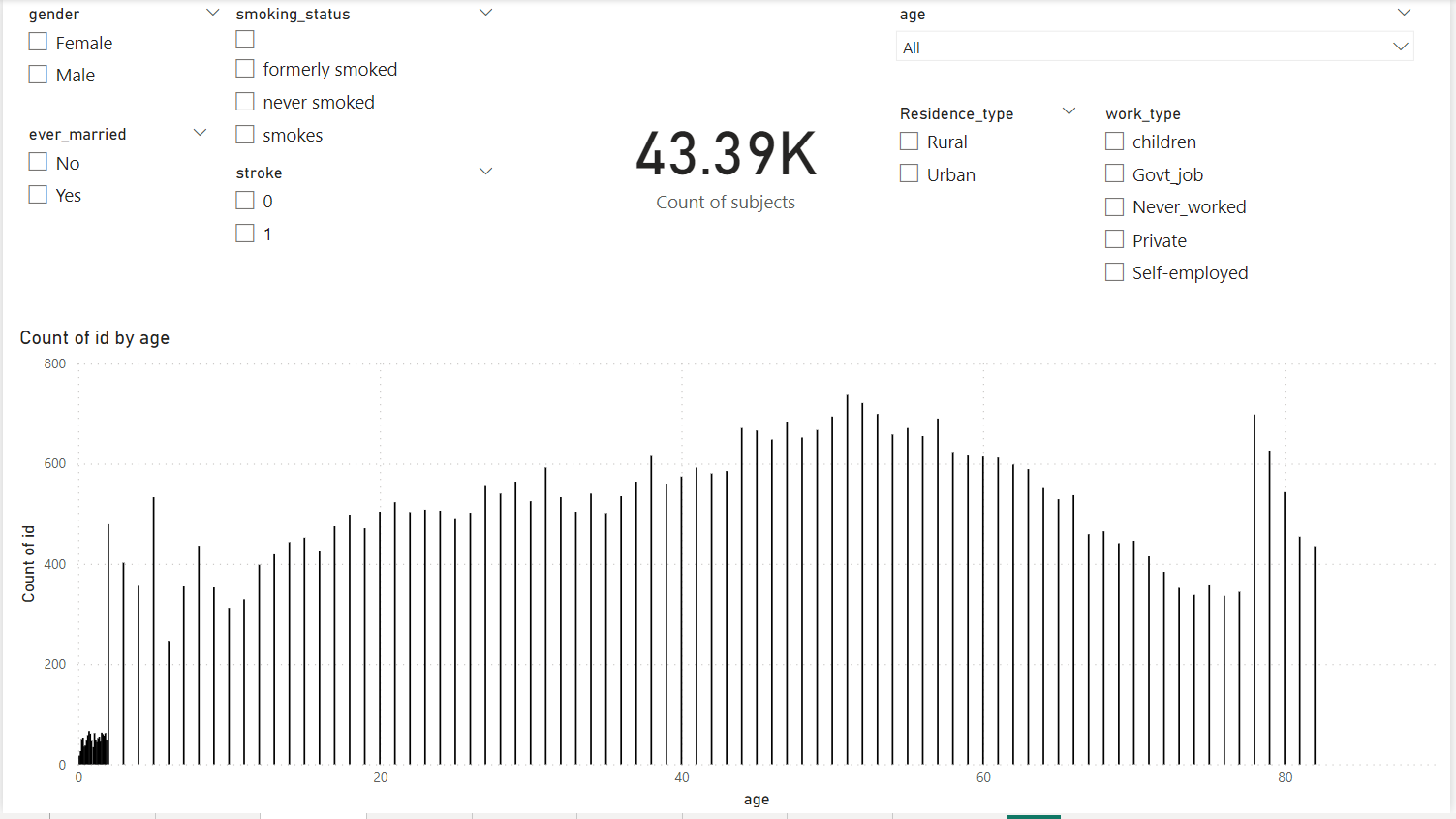
* The count of all subjects shares among all the patient’s data is plotted as

****

**Fig 5**

**Fig 6**

The famous dynamic dashboard and their drilldowns are plotted as

****

**Fig 6**

* Here, the dashboard shows the entire data and can be filtered as needed.

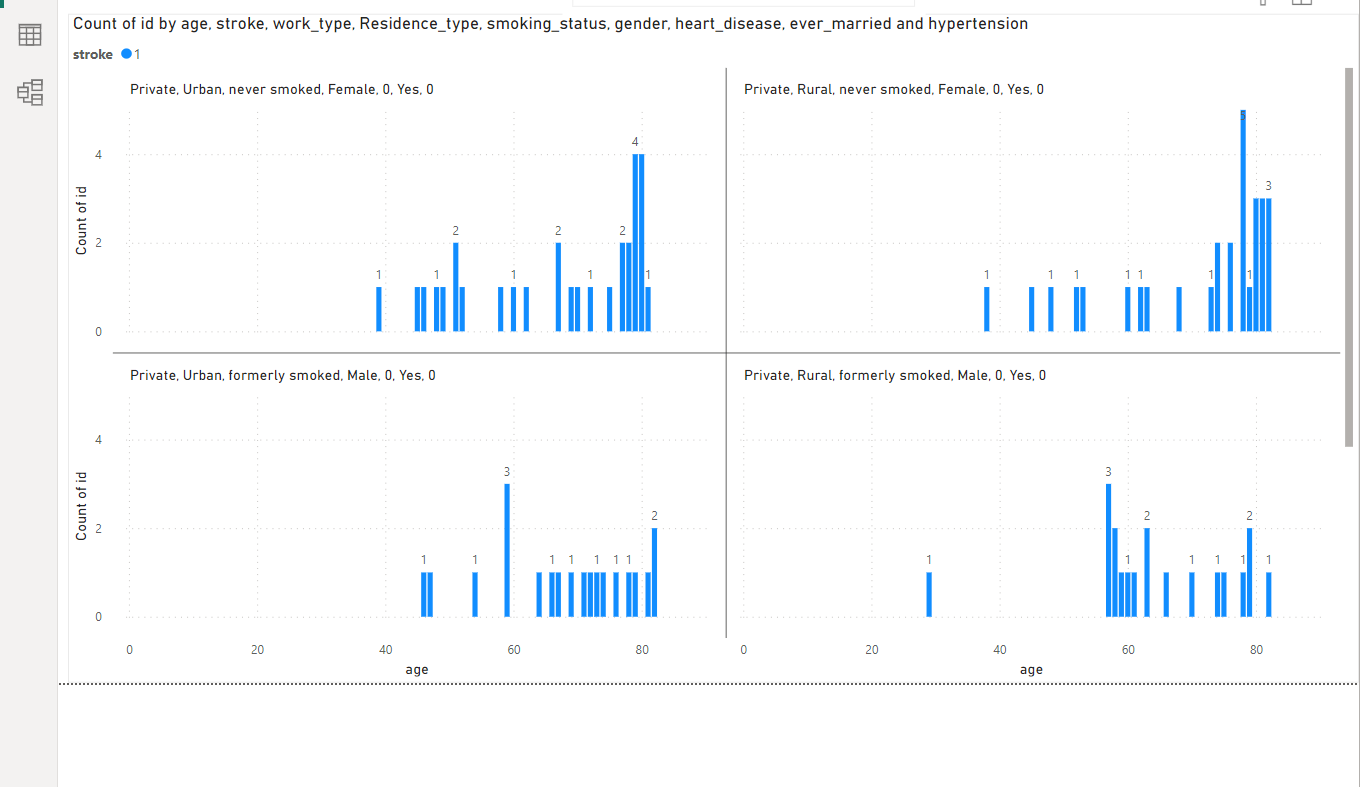
**Results**

The following are the results we concluded from the dataset.

* **It seemed like both BMI and Age were positively correlated, though the association was not strong.**
* **Older patient was more likely to suffer a stroke than a younger patient.**
* **Higher BMI does not increase the stroke risk.**
* **Higher proportion of patients who suffered from hypertension or heart disease experienced a stroke, all else being equal.**
* **Regardless of patient’s gender, and where they stayed, they have the same likelihood to experience stroke.**
* **Work type variable was highly associated with age.**
* **Marital status variable was highly associated with age.**
* **Smoking\_status has very less influence on the stroke.**

The complete analysis can be found in the pyspark notebook we made use of for analysis.

[**https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/4686834391626773/3034553378902319/4011599185336554/latest.html**](https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/4686834391626773/3034553378902319/4011599185336554/latest.html)



This is an attracting PowerBI live visualization graphic that allows us to view data as new data gets added to it. Mostly all the powerbi pages we created are dynamic and changes as new data gets added to it.

**Conclusion**

We have analyzed the data and created various powerful visualizations using Pyspark and PowerBI. The future scope of this project can include various works related to machine learning and advanced deep learning algorithms to predict the chances of a stroke by collecting the user related data. The scope of this work can also be extended by creating live interactive dashboards that enable users to view the advanced charts and other streams.

Author Contributions

Equal contribution has been done by all the group members. But more specifically **Prakash** has worked on Data Cleaning and Analysis, creating resources and hops in Azure Blob Storage and Databricks. **Dharani** has worked on collecting datasets and data visualization. **Krupa** has worked on platform architecture and. **Rajashekar** worked on Deployment in Azure and worked on PowerBI for creating an interactive dashboard.

**References**

1. <https://github.com/prakash961/Project_Stroke/tree/main>
2. <https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/4686834391626773/3034553378902319/4011599185336554/latest.html>
3. https://portal.azure.com/#@prakashchekkera1gmail.onmicrosoft.com/resource/subscriptions/609f3d36-1f28-423a-9af6-acf5b39f7340/resourceGroups/CloudComputing/providers/Microsoft.Storage/storageAccounts/ccproject0/overview

Conference Name:ACM Woodstock conference

Conference Short Name:WOODSTOCK’18

Conference Location:El Paso, Texas USA

ISBN:978-1-4503-0000-0/18/06

Year:2018

Date:June

Copyright Year:2018

Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00