A Case Study on Using Databricks for COVID-19 Dataset Analysis

Team Members

Siddharth Dudugu ([sd3627@drexel.edu](mailto:sd3627@drexel.edu))

Harsh Ramesh ([hr439@drexel.edu](mailto:hr439@drexel.edu))

Surrender Pothuri ([sp3796@drexel.edu](mailto:sp3796@drexel.edu))

This projects presents an overview of the applications of Databricks and PySpark for ana lyzing COVID-19 data. The project includes exploratory data analysis and the application of a machine learning model to predict mortality rates due to COVID-19. The documentation covers various aspects of the project, including data cleaning and preprocessing, data visualization, predictive modeling, collaborative data analysis, data governance and security, challenges, and limitations.

**1. What is DataBricks?**

Databricks is designed to provide a unified platform for big data analytics, leveraging the capabilities of Apache Spark. By combining Spark's distributed computing framework with a user-friendly interface, Databricks makes it easier for data scientists and analysts to work with large datasets.

The scalable architecture of Databricks allows organizations to handle the processing and analysis of big data efficiently. With the ability to distribute computations across clusters, it enables parallel processing, reducing the time required for data transformations and analytics tasks.

Databricks provides an interactive workspace that facilitates data exploration and analysis. Data scientists can leverage its interactive notebooks, which support code execution, data visualization, and documentation. This interactive environment helps in iterative analysis and quick experimentation.

One of the key advantages of Databricks is its built-in collaboration tools. These tools enable teams to work together effectively, share notebooks, code, and visualizations, and collaborate on data analysis projects. This collaborative aspect fosters knowledge sharing and improves the overall productivity of the team.

Furthermore, Databricks supports multiple programming languages, including Python, which is widely used in data analysis and machine learning. This flexibility allows data scientists to leverage their existing skills and libraries in Python for COVID-19 data analysis within the Databricks platform.

**2. Dataset**

The dataset used in this project was provided by the Mexican government and contains anonymized patient-related information, including pre-conditions. It consists of 21 unique features and 1,048,576 unique patients, such as sex, age, COVID-19 test findings, patient type, and various medical conditions. The dataset also provides details about the medical unit, intubation, ICU admission, and the date of death if applicable. The objective of the project is to build a machine learning model that predicts high-risk COVID-19 patients based on their symptoms and medical history, aiding in resource allocation, and providing appropriate care.

This dataset contains an enormous number of anonymized patient-related information including pre-conditions. **In the Boolean features, 1 means "yes" and 2 means "no". values as 97 and 99 are missing data**.

* sex: 1 for female and 2 for male.
* age: of the patient.
* classification: covid test findings. Values 1-3 mean that the patient was diagnosed with covid in different  
   degrees. 4 or higher means that the patient is not a carrier of covid or that the test is inconclusive.
* patient type: type of care the patient received in the unit. 1 for returned home and 2 for hospitalization.
* pneumonia: whether the patient already have air sacs inflammation or not.
* pregnancy: whether the patient is pregnant or not.
* diabetes: whether the patient has diabetes or not.
* copd: Indicates whether the patient has Chronic obstructive pulmonary disease or not.
* asthma: whether the patient has asthma or not.
* inmsupr: whether the patient is immunosuppressed or not.
* hypertension: whether the patient has hypertension or not.
* cardiovascular: whether the patient has heart or blood vessels related disease.
* renal chronic: whether the patient has chronic renal disease or not.
* other disease: whether the patient has other disease or not.
* obesity: whether the patient is obese or not.
* tobacco: whether the patient is a tobacco user.
* usmr: Indicates whether the patient treated medical units of the first, second or third level.
* medical unit: type of institution of the National Health System that provided the care.
* intubed: whether the patient was connected to the ventilator.
* icu: Indicates whether the patient had been admitted to an Intensive Care Unit.
* date died: If the patient died indicate the date of death, and 9999-99-99 otherwise.

3. Databricks & PySpark for Data Analysis

Databricks and PySpark are crucial tools for us when it comes to processing and analyzing massive datasets in big data analytics. They provide us with the necessary capabilities to handle large volumes of data and perform complex computations on a scalable architecture.

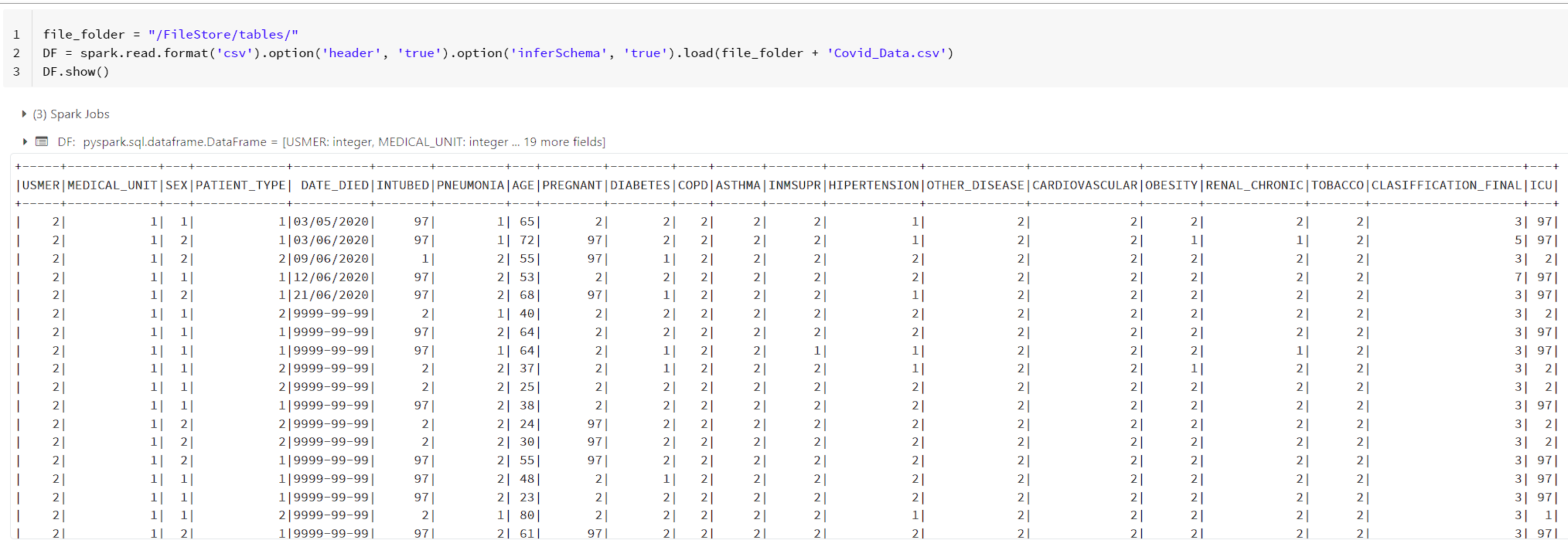
One of the key advantages of using Databricks and PySpark is their ability to leverage distributed computing. By harnessing the power of clusters, we can process data in parallel, significantly reducing the time required for data transformations and analytics tasks. This parallel processing capability is essential for us to efficiently analyze the vast amount of COVID-19 data we encounter.

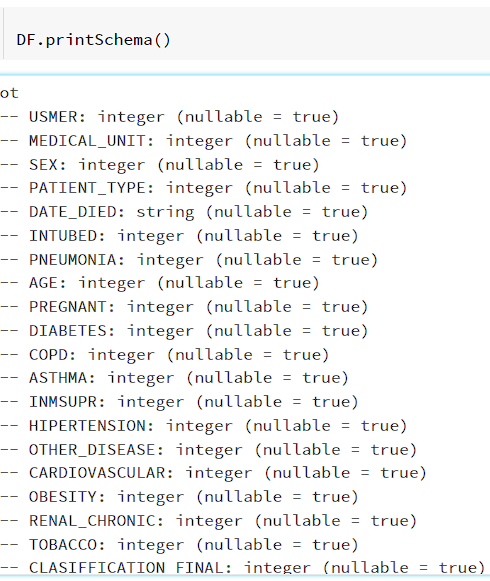
Moreover, the advanced analytics libraries and APIs available in Databricks and PySpark greatly enhance our data analysis capabilities. These libraries, including MLlib and Spark SQL, provide us with a wide range of tools to conduct complex analyses and build predictive models. We can apply sophisticated algorithms and techniques to gain valuable insights and make informed decisions based on the COVID-19 data.

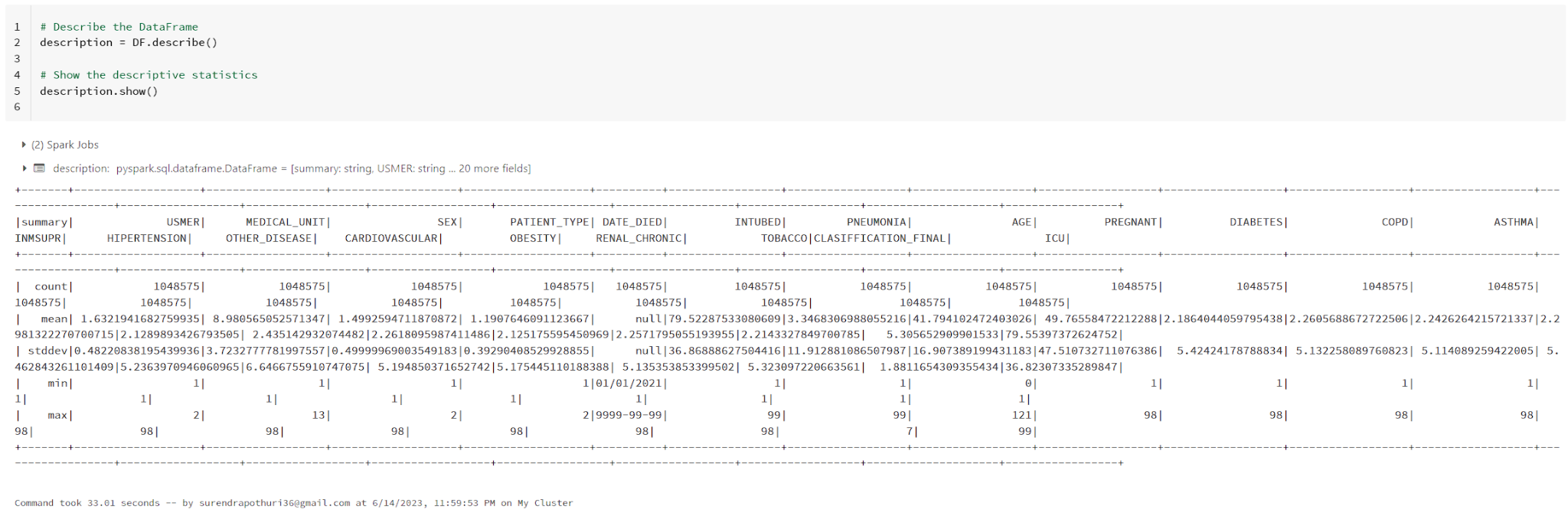
The flexibility of PySpark, particularly its support for Python, is another advantage for our organization. Many of our data scientists and analysts are already proficient in Python and have access to a wealth of existing libraries and resources. With PySpark, we can seamlessly integrate our Python skills and leverage these libraries such as NumPy, pandas, and scikit-learn. This compatibility allows us to work with COVID-19 data using familiar tools, making our analysis more efficient and effective.

Additionally, the scalability and parallel processing capabilities of Databricks and PySpark are essential for us. We can handle the ever-increasing volume of COVID-19 data without compromising on performance. The ability to process large datasets efficiently enables us to derive meaningful insights in a timely manner, contributing to our decision-making processes and the fight against the pandemic.

Data Exploration:





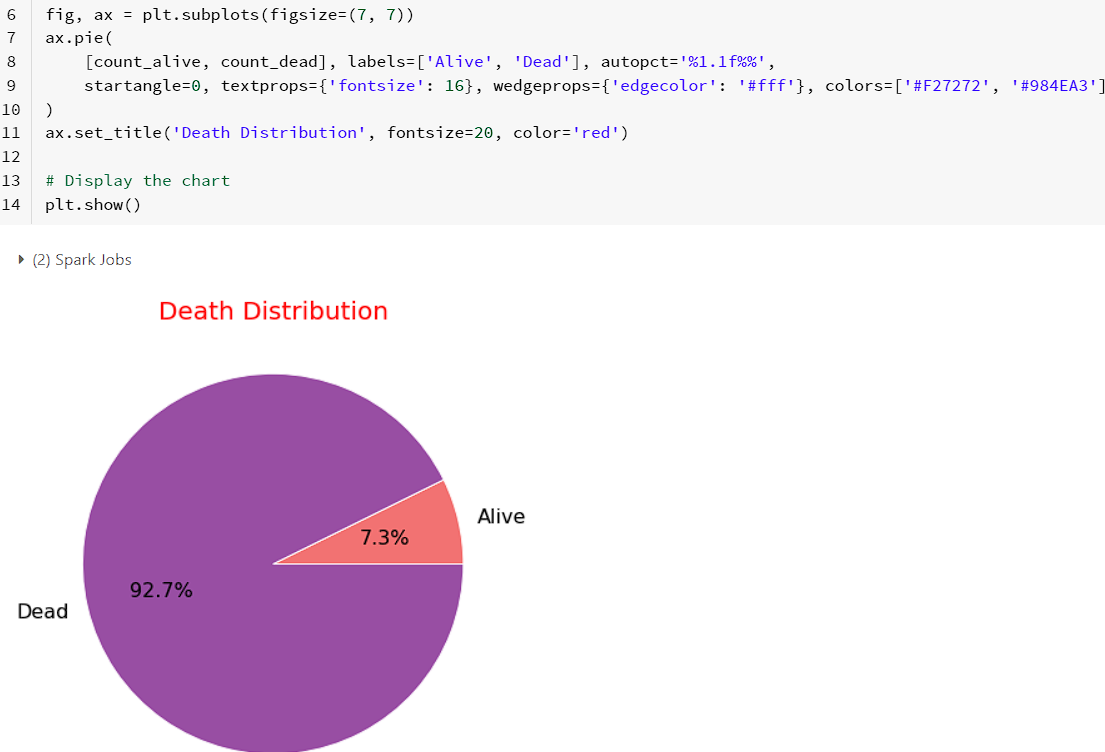


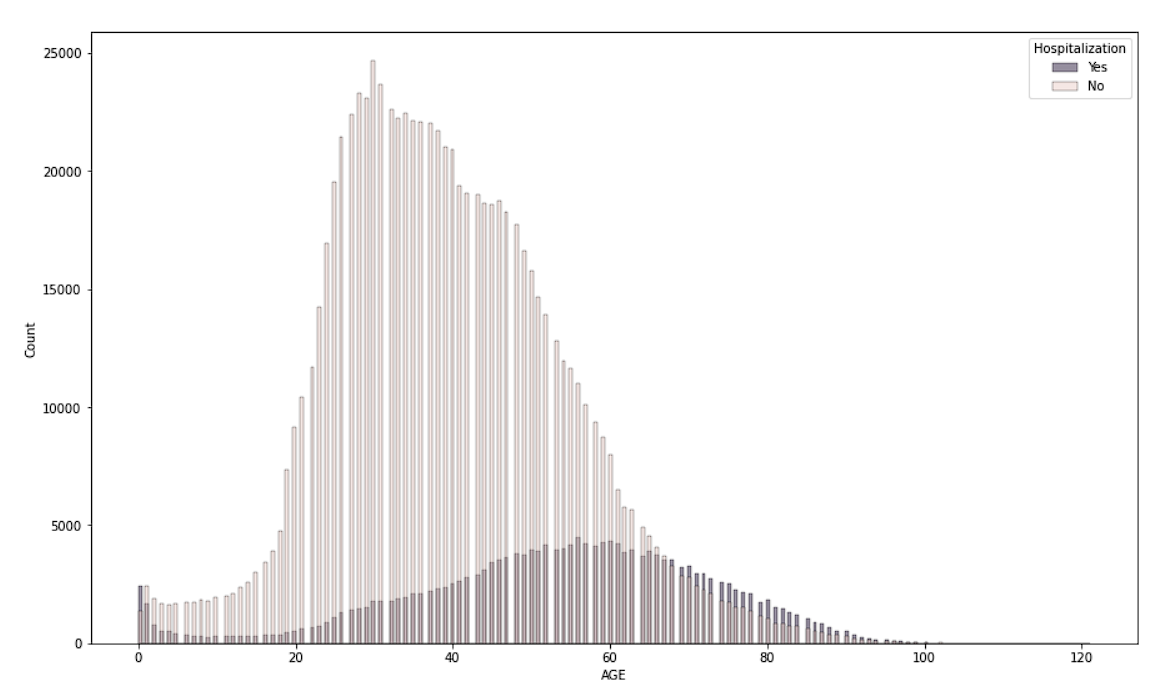
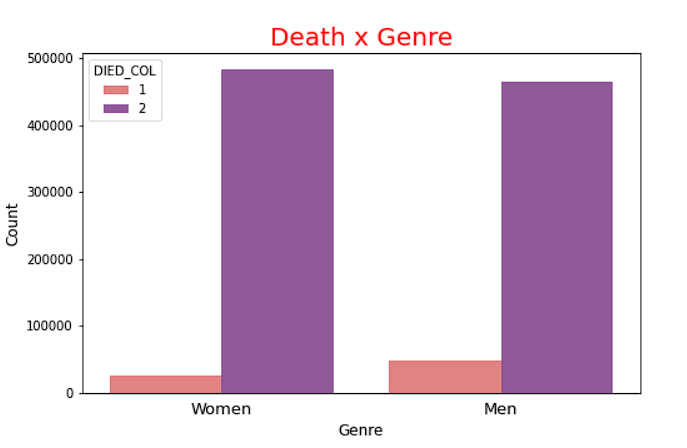
4. Data Cleaning & Pre-processing

During the data exploration phase, it was discovered that all the categorical columns were represented as numerical values. The dataset's author used integers '97', '98', and '99' to indicate missing values, which affected attributes such as INTUBED (whether the patient was intubed or not), PREGNANT (whether the patient was pregnant or not), and ICU (whether the patient was admitted to the ICU or not). Furthermore, the 'DATE\_DIED' attribute had values representing patients who were still alive, indicated by date values '9999-99-99'. To handle missing information, the following steps were taken:

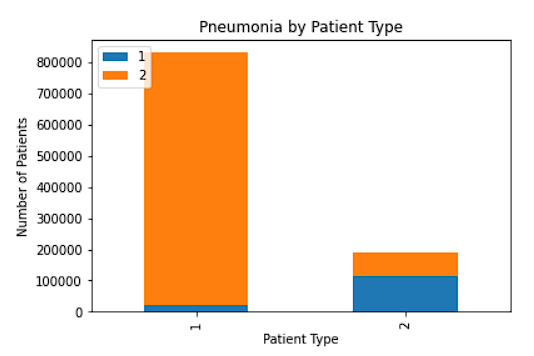
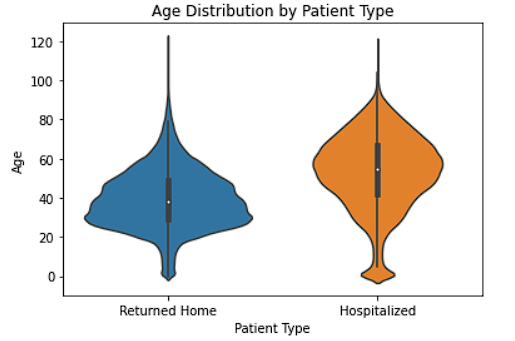
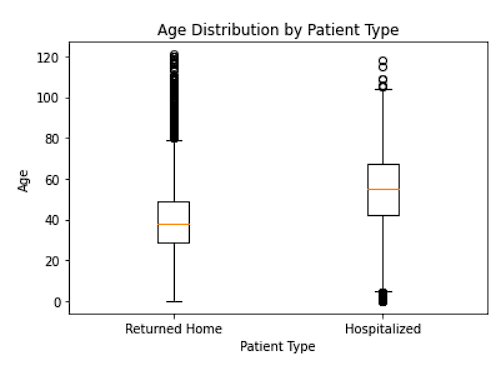
* Replaced missing values in 'PREGNANT' attribute with '2', representing 'no', when the patient's sex was male.
* Dropped columns 'ICU' and 'INTUBED'.
* Dropped remaining missing values in other attributes except for the 'DATE\_DIED' attribute.
* Created a new attribute representing whether the patient is alive or not, with values '2' and '1', respectively.

5. Data Visualisation & Exploratory Data Analysis using Pyspark:

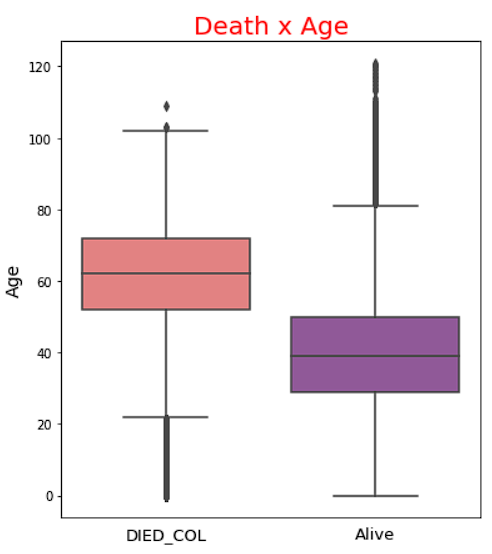
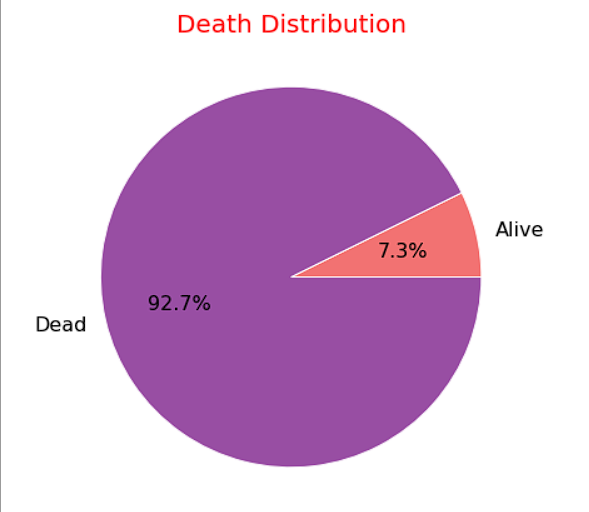
PySpark offers powerful visualization libraries such as Matplotlib and Seaborn. Matplotlib provides diverse plot types, including line charts, bar graphs, and scatter plots, while Seaborn enhances plots aesthetically and provides additional functionalities. Data visualization plays a crucial role in presenting COVID-19 data effectively for decision-making. Visualizations simplify complex information, reveal patterns and correlations, and aid in data exploration 

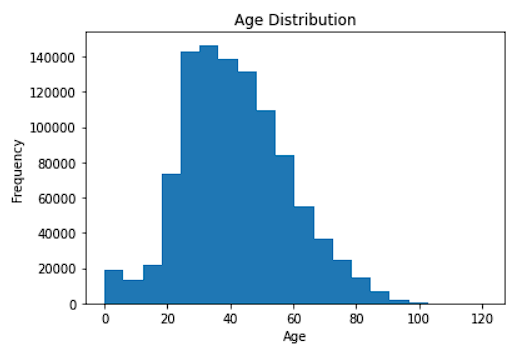
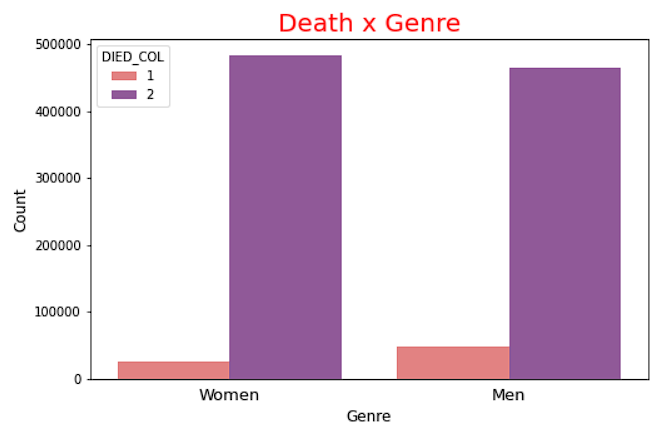


Disease Prevalence Analysis:

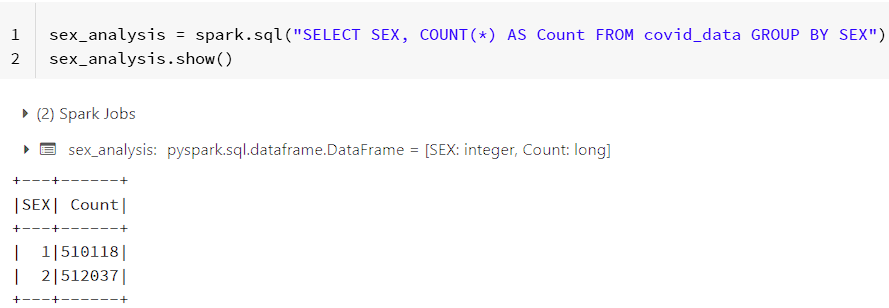


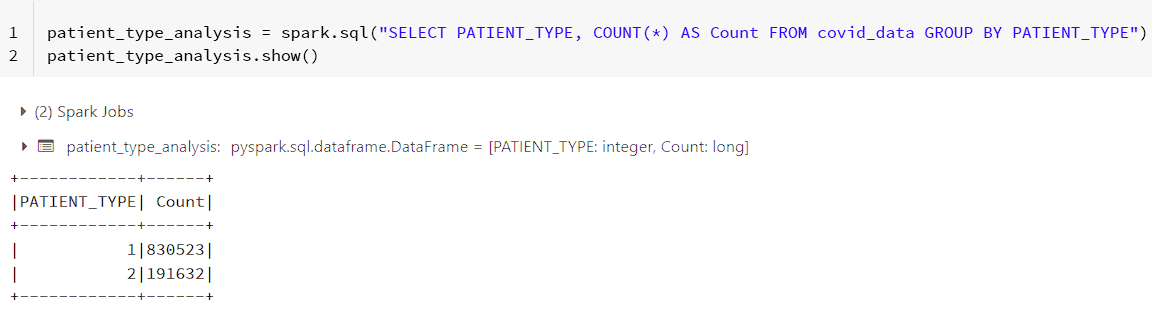
Mortality Analysis:

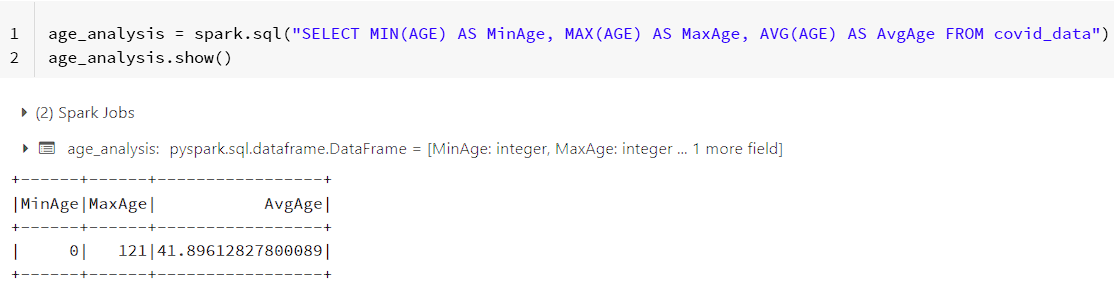


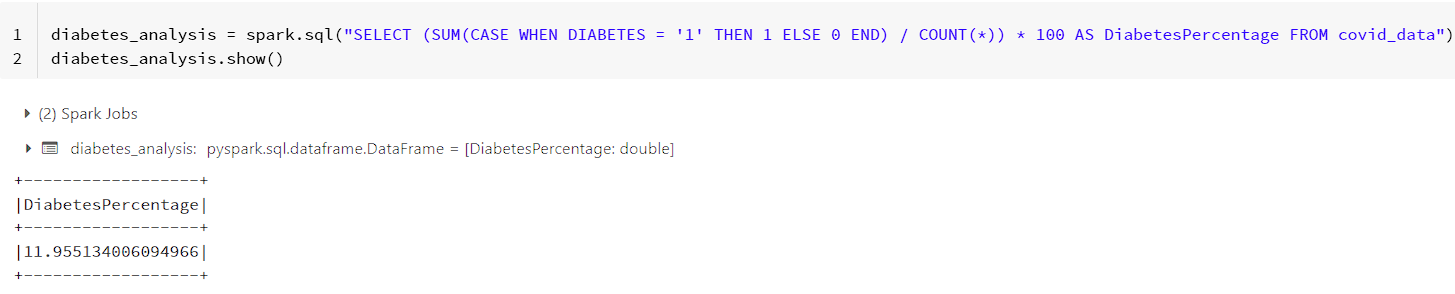


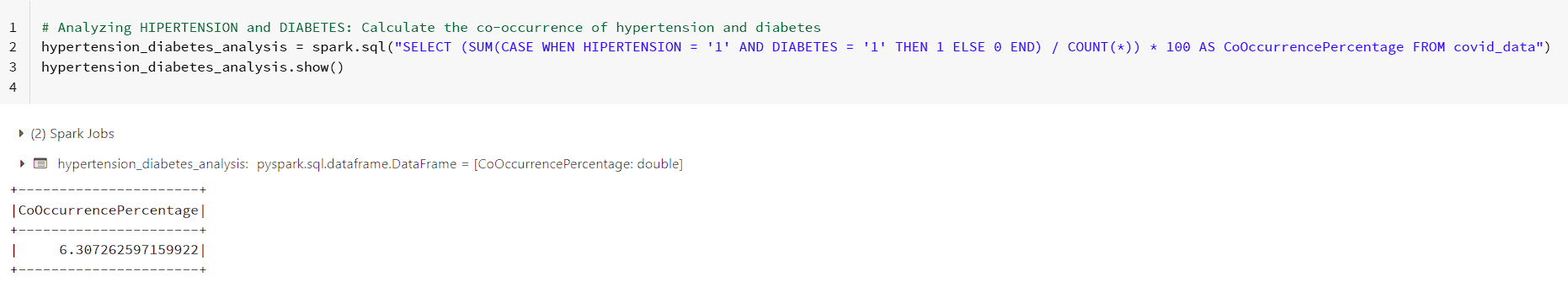
6. EDA using Pyspark:



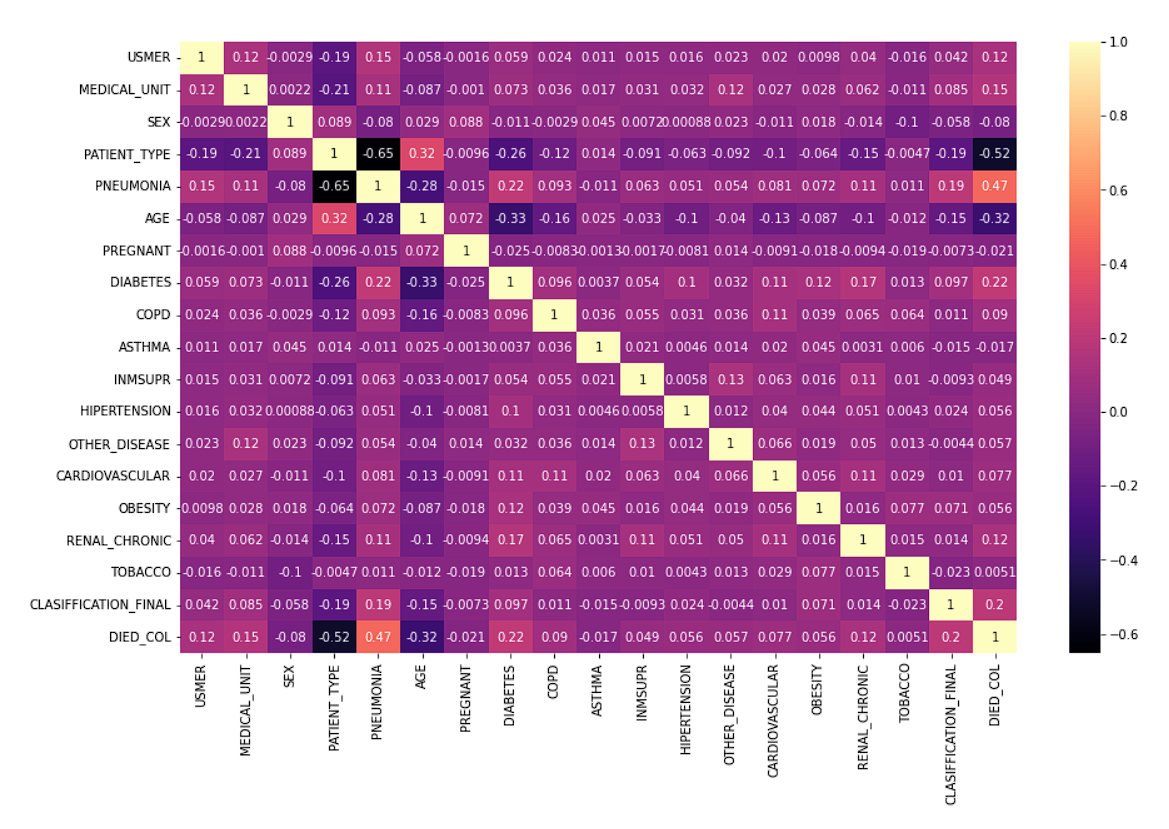








Correlation Analyis:



7. COVID-19 Predictive Modeling using PySpark

In the predictive modeling task, we followed a systematic approach to build a model that could accurately predict COVID-19 mortality rates. The steps involved in this process are as follows:

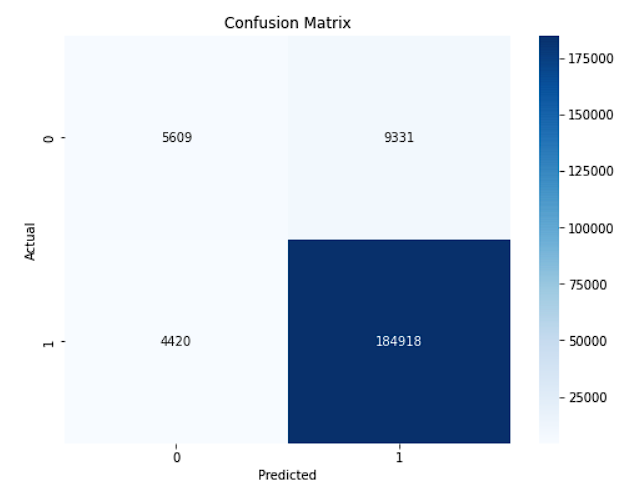
* Feature Selection: We carefully selected a set of features that we believed would be informative in predicting COVID-19 mortality rates. These features included 'AGE', 'PATIENT\_TYPE', 'PNEUMONIA', 'MEDICAL\_UNIT', and 'HIPERTENSION'. By considering these specific attributes, we aimed to capture important patient characteristics and medical conditions that could contribute to the prediction.
* Data Transformation: To prepare the data for modeling, we utilized a VectorAssembler to combine the selected features into a single vector column. This transformation allowed us to create a unified representation of the features, enabling the model to efficiently process and analyze the data.
* Data Splitting: In order to evaluate the performance of the model, we divided the data into training and test sets. We allocated 80% of the data for training the model and kept the remaining 20% for testing its accuracy. By using a random seed of 42, we ensured that the split was consistent for reproducibility.
* Model Training: We trained a Logistic Regression model on the training set. Logistic Regression is a popular classification algorithm that is well-suited for predicting binary outcomes, making it suitable for our task of predicting COVID-19 mortality. The model learned from the training data and identified patterns and relationships between the selected features and the target variable.
* Model Evaluation: To assess the accuracy of our model, we evaluated its performance on the test set. We compared the model's predictions against the actual outcomes in the 'DIED\_COL' column. The accuracy of the model on the test set was found to be 93%, indicating that it was able to accurately classify high-risk COVID-19 patients.
* Prediction Display: To gain insights into the model's predictions, we displayed the target column 'DIED\_COL' and the corresponding 'prediction' column for the top 20 rows. This allowed us to visually compare the predicted outcomes with the actual outcomes, providing a clear understanding of how well the model performed on individual instances.

By following these steps, we were able to build and evaluate a predictive model for COVID-19 mortality rates. The model demonstrated a high accuracy level, suggesting its potential usefulness in identifying high-risk patients. These findings can aid healthcare authorities in efficiently allocating resources and providing appropriate care to those in need.

8. Model Evaluation

To evaluate the performance of the predictive model, several techniques were employed, including:

* To evaluate the performance of our predictive model, we employed several techniques that allowed us to gain insights into its effectiveness in predicting COVID-19 mortality rates. These techniques provided us with a comprehensive assessment of the model's performance and helped us make informed decisions based on the results.
* First and foremost, we calculated the accuracy of the model. Accuracy represents the overall correctness of the predictions made by the model. By comparing the predicted outcomes with the actual outcomes, we were able to determine the percentage of correct predictions. This metric provided us with a general measure of the model's performance and gave us an idea of its reliability.
* In addition to accuracy, we constructed a confusion matrix to further analyze the model's predictions. The confusion matrix allowed us to examine the true positives, true negatives, false positives, and false negatives generated by the model. This analysis provided us with a more detailed understanding of the model's performance, highlighting its strengths and weaknesses in terms of correctly identifying COVID-19 mortality cases.
* Furthermore, we utilized evaluation metrics such as precision, recall, and F1 score to assess the model's performance. Precision measures the proportion of correctly predicted positive cases among all predicted positive cases, while recall measures the proportion of correctly predicted positive cases among all actual positive cases. The F1 score combines precision and recall into a single metric, providing a balanced evaluation of the model's performance. By considering these metrics, we were able to gain insights into the model's ability to correctly classify high-risk COVID-19 patients.
* To enhance the visualization of the confusion matrix, we employed a heatmap. The heatmap visually represented the distribution of correct and incorrect predictions, allowing us to identify patterns and discrepancies in the model's performance. This visualization technique enabled us to quickly identify areas where the model excelled and areas that required improvement, facilitating our decision-making process and future model iterations.



By employing these evaluation techniques, including accuracy, confusion matrix, evaluation metrics, and heatmap visualization, we obtained a comprehensive understanding of the predictive model's performance. This information empowered us to make informed decisions, assess the model's effectiveness, and identify areas for improvement. Ultimately, these evaluation techniques played a vital role in ensuring the reliability and usefulness of our predictive model in predicting COVID-19 mortality rates.

9. Data Governance & Security

Data governance and security are crucial considerations in COVID-19 data analysis. The importance of data governance lies in ensuring data accuracy, integrity, privacy, and compliance with regulations. When dealing with sensitive health data, organizations must protect it from unauthorized access and ensure data confidentiality. Databricks addresses these concerns by implementing robust security measures, including encryption, access controls, and audit logs. It also provides data protection mechanisms to safeguard COVID-19 data throughout its lifecycle, helping organizations adhere to data protection regulations such as HIPAA and GDPR.

10. Challenges & Limitations

Throughout our project, we encountered several challenges and limitations that are inherent to COVID-19 data analysis. These challenges have shaped our understanding of the complexities involved and influenced the outcomes of our analyses. We would like to highlight the key challenges and limitations we faced from our perspective:

Data Quality: Ensuring the quality of the data has been a significant challenge. COVID-19 datasets can vary in terms of accuracy, completeness, and consistency. We dedicated considerable effort to cleaning and preprocessing the data, addressing issues such as missing values, outliers, and inconsistencies. However, we acknowledge that there may still be residual data quality issues that could impact the reliability of our findings. Maintaining data integrity and accuracy is an ongoing effort that requires continuous vigilance.

Data Privacy: Handling sensitive health data and maintaining data privacy has been a top priority for us. COVID-19 data often contains personally identifiable information that must be protected. We have taken stringent measures to ensure data anonymization and comply with relevant data protection regulations. Safeguarding individual privacy while conducting meaningful analyses has been a constant consideration throughout the project.

Computational Resources: Analyzing large-scale COVID-19 datasets demands substantial computational resources. Processing and analyzing massive volumes of data can be computationally intensive and require scalable infrastructure. We have worked diligently to optimize our workflows, leverage distributed computing capabilities, and make efficient use of the resources provided by Databricks and PySpark. However, the availability and allocation of computational resources remain an ongoing challenge that influences the scale and scope of our analyses.

Data Integration: Integrating diverse datasets from multiple sources has been essential for comprehensive COVID-19 data analysis. However, data integration poses challenges due to variations in data formats, structures, and semantics across different sources. We have made concerted efforts to harmonize and consolidate the data, but we acknowledge that some level of heterogeneity and data integration complexity may persist. Overcoming these challenges requires careful data mapping and transformation techniques.

Data Interpretation & Context: Interpreting COVID-19 data within the appropriate context has been crucial for meaningful analysis. Understanding the limitations, biases, and uncertainties associated with the data is essential for accurate interpretation. Factors such as testing strategies, reporting discrepancies, and evolving understanding of the virus have influenced our interpretation of results. We have strived to provide insightful interpretations while remaining mindful of the inherent limitations and potential biases present in the data.

Despite these challenges and limitations, we have remained committed to conducting rigorous analyses and generating valuable insights. By employing best practices, leveraging advanced analytics techniques, and utilizing the capabilities of Databricks and PySpark, we have aimed to contribute to the collective understanding of the COVID-19 pandemic. Our focus has been on providing evidence-based insights while being transparent about the challenges we have encountered along the way.

11. Databricks VS Other Analysis Tools:

Our experience with Databricks in this project was highly beneficial. We encountered several situations where Databricks proved to be advantageous:

Big Data Processing: Handling a large amount of data with various abnormalities was made easy by Databricks. Its integration with Apache Spark provided the scalability and distributed computing capabilities necessary to efficiently process the workload. Databricks allowed us to perform complex data transformations, aggregations, and machine learning tasks on massive datasets.

Machine Learning Tasks: Databricks simplified the implementation of end-to-end machine learning models. We found it convenient to use Databricks notebooks to write code for data preprocessing, feature engineering, model training, and evaluation. The distributed computing capabilities of Databricks ensured fast execution of these tasks, particularly when dealing with large datasets.

Collaborative Data Analysis: Databricks provided a collaborative environment that enabled multiple data scientists and analysts to work together on the project. Since we worked on the project online, we could easily share notebooks, collaborate on code, and provide feedback to each other. This collaborative aspect of Databricks was particularly useful for our team.

Overall, Databricks streamlined our data processing, machine learning, and collaborative analysis tasks, making the project more efficient and productive.

12. Conclusion

As students, we have recognized the significant impact of data analysis in addressing the challenges posed by the COVID-19 pandemic. In our project, we have leveraged the powerful tools and scalable architecture of Databricks and PySpark to analyze large-scale COVID-19 datasets and gain valuable insights.

Working collaboratively within the Databricks environment, we have utilized features such as notebooks and version control to enhance our teamwork and knowledge sharing. This has allowed us to tap into the collective expertise of our team members, fostering a dynamic and collaborative learning environment.

During our project, we encountered various challenges and limitations that are common in COVID-19 data analysis. These include ensuring data quality, addressing privacy concerns, managing computational resources, integrating diverse datasets, and interpreting the data within the appropriate context. While navigating these challenges, we have learned valuable lessons about the complexities involved in real-world data analysis projects.

To optimize our analyses, we have followed best practices such as efficient coding, performance optimization, and reproducibility. These practices have enabled us to streamline our workflows, improve efficiency, and enhance the transparency and reproducibility of our results.

In summary, this project has allowed us to contribute to the field of COVID-19 data analysis using Databricks and PySpark. Through this project, we have gained practical experience in handling large-scale datasets, collaborating effectively, and addressing real-world challenges. We are proud to have made meaningful contributions to our understanding of the COVID-19 pandemic and the application of data analysis in addressing global health crises.

References:

Databricks glossary:<https://www.databricks.com/glossary/big-data-analytics>

Databricks documentation on exploratory data analysis:<https://docs.databricks.com/exploratory-data-analysis/index.html>

Article on exploratory data analysis with PySpark on Databricks:<https://towardsdatascience.com/exploratory-data-analysis-eda-with-pyspark-on-databricks-e8d6529626b1>

Dataset Link:<https://www.kaggle.com/datasets/meirnizri/covid19-dataset>