Insights from Four Decades of US Natality Data

Sanket Dalvi
Incedo Inc
sanket.dalvi@incedoinc.com

Nishitha Chidipothu
Incedo Inc
nishitha.c@incedoinc.com

ABSTRACT

This study explores the extensive United States natality dataset provided by the Centers for Disease Control and Prevention (CDC), covering birth records from 1969 to 2008. Our research focuses on the crucial preprocessing and data cleaning steps necessary for effective analysis, as well as the application of machine learning models to extract meaningful insights from this rich demographic resource. We discuss the challenges encountered in handling such a large-scale, longitudinal dataset and present our methodologies for overcoming these obstacles. Our analysis reveals significant trends in birth outcomes, maternal health, and demographic shifts over the four-decade span. Furthermore, we demonstrate the potential of machine learning algorithms in predicting various birth-related outcomes and identifying key factors influencing maternal and infant health. This research contributes to the fields of public health, demography, and data science by showcasing the power of big data analytics in understanding long-term population health trends.

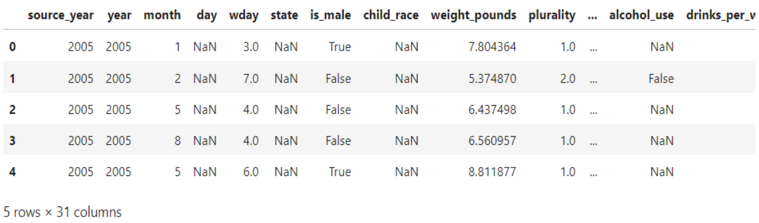
1 INTRODUCTION

The United States natality dataset, compiled by the Centers for Disease Control and Prevention (CDC), represents one of the most comprehensive resources for understanding birth trends and outcomes in the United States. Spanning from 1969 to 2008, this dataset offers a unique opportunity to analyze four decades of demographic and health information related to births across the nation. The sheer volume and complexity of this data present both exciting opportunities and significant challenges for researchers and data scientists.

Maternal and infant health continues to remain a critical global challenge with disparities in access to healthcare contributing significantly to adverse birth outcomes. Effective allocation of resources for care of mothers in mitigating these disparities is very important. This study aims to contribute to this effort by analyzing historical natality data to identify key factors influencing birth outcomes.

In recent years, the advent of advanced data processing techniques and machine learning algorithms has opened new avenues for extracting insights from large-scale datasets. However, the effective utilization of such extensive historical data requires careful preprocessing and cleaning to ensure the validity and reliability of subsequent analyses. This study aims to address these challenges and leverage the potential of the CDC natality dataset to uncover long-term trends and patterns in US birth statistics.

2 DATASET



	source_year	year	month	day	wday	state	is_male	child_race	weight_pounds	plurality	...	alcohol_use	drinks_per_v
0	2005	2005	1	NaN	3.0	NaN	True	NaN	7.804364	1.0	...	NaN	
1	2005	2005	2	NaN	7.0	NaN	False	NaN	5.374870	2.0	...	False	
2	2005	2005	5	NaN	4.0	NaN	False	NaN	6.437498	1.0	...	NaN	
3	2005	2005	8	NaN	4.0	NaN	False	NaN	6.560957	1.0	...	NaN	
4	2005	2005	5	NaN	6.0	NaN	True	NaN	8.811877	1.0	...	NaN	

5 rows × 31 columns

Figure 1: Dataset

Dataset Description: US CDC Natality Data (1969-2008). The dataset used in this project is the United States Natality dataset, compiled by the Centers for Disease Control and Prevention (CDC).

Time span: 40 years, from 1969 to 2008

Key variables: Maternal characteristics (age, race, education) Paternal information Birth outcomes (birth weight, gestational age) Geographic data (state)

Size: 1,40,000 rows with 31 columns

3 PROJECT ARCHITECTURE

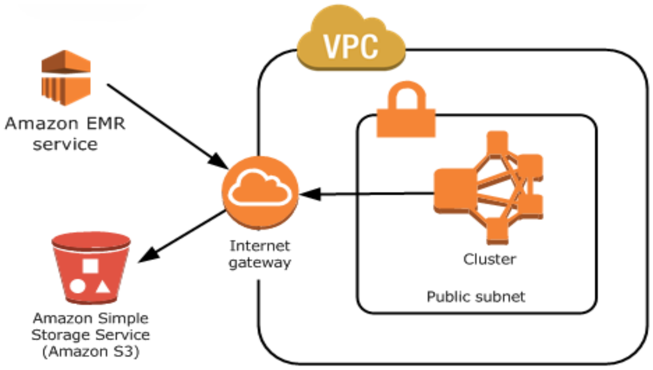


Figure 2: Architectural Diagram

VPC (Virtual Private Cloud): This is the overarching container for the cloud resources, providing a logically isolated section of the AWS cloud.

Conference'17, Washington, DC, USA
© 2016 ACM. 978-x-xxxx-xxxx-x/YY/MM…$15.00
DOI: 10.1145/nnnnnnn.nnnnnnn

Cluster: An Amazon EMR (Elastic MapReduce) cluster for data processing.

Internet Gateway: This component allows communication between the VPC and the internet, enabling the cluster to access external resources and services.

Amazon EMR Service: Managed service provided by AWS for big data processing.

Amazon Simple Storage Service (Amazon S3): Object storage service for input data and output results.

This architecture allows for secure, scalable big data processing. The EMR cluster in the public subnet can process data, potentially sourced from S3, while being managed by the EMR service. The VPC and security measures ensure that the data processing occurs in a controlled environment, while still allowing necessary internet access for data transfer and service management.

4 SERVICES USED

To effecitvely handle the massive CDC natality dataset, we leveraged a suite of Amazon Web Services (AWS) to streamline data storage, processing and analysis,

AWS provides a comphrehensive cloud computing platform offering a wide range of services to build and deploy applications. For our project we primarily utilized-

4.1 Data Storage and Access

Amazon S3- Used to store the massive CDV natality dataset in a highly durable and scalable object storage. The dataset's size and longevity make s3 an ideal choice for long term data rentention.

Amazon VPC- Provides a secure, isolated network environment for processing the sensitive birth data. By creating a VPC, we ensure that only authorized personnel and systems can access the data.

4.2 Data Processing and Transformation

Within the EMR cluster, PySpark was the primary tool for data cleaning and preprocessing. This involved tasks such as : Data Ingestion: Loading the natality data from S3 into a Spark DataFrame. Data quality assessment- Identifying and handling missing values, inconsistencies, and outliers in the dataset. Data Cleaning: Correcting errors, standardizing data formats, and creating derived varaibles. Data Transformation: Agggregating data, creating summary statistic, and preparing the data for machine learning modeling.

4.3 Security

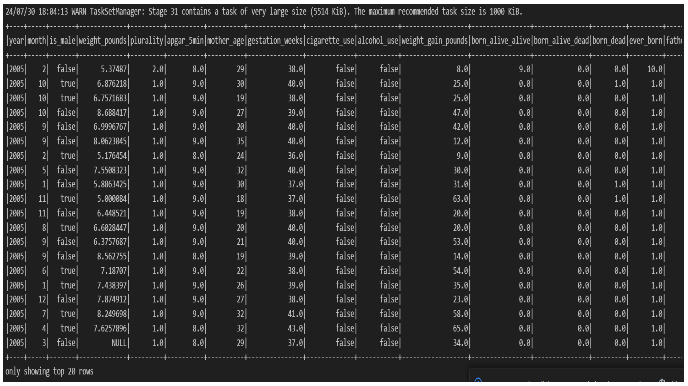
VPC provides a logically isolated section of the AWS Cloud where we launch our AWS resources. Security Groups act as a virtual firewall for the instances to control inbound and outbound traffic.

5 CODE

We have used different packages - boto3, pandas, ydata-profiling, numpy, pyspark

Nishitha Link

Sanket Link.



```
24/07/30 18:04:13 WARN TaskSetManager: Stage 31 contains a task of very large size (5514 KiB). The maximum recommended task size is 1000 KiB.
+----+-----+-------+-------------+---------+----------+----------+---------------+-------------+-----------+------------------+----------------+---------------+---------+---------+----
|year|month|is_male|weight_pounds|plurality|apgar_5min|mother_age|gestation_weeks|cigarette_use|alcohol_use|weight_gain_pounds|born_alive_alive|born_alive_dead|born_dead|ever_born|fathe
+----+-----+-------+-------------+---------+----------+----------+---------------+-------------+-----------+------------------+----------------+---------------+---------+---------+----
|2005|    2|  false|      5.37487|      2.0|       8.0|        29|           38.0|        false|      false|               8.0|             9.0|            0.0|      0.0|     10.0|
|2005|   10|   true|     6.876218|      1.0|       9.0|        30|           40.0|        false|      false|              25.0|             0.0|            0.0|      1.0|      1.0|
|2005|   10|   true|    6.7571683|      1.0|       9.0|        19|           38.0|        false|      false|              25.0|             0.0|            0.0|      0.0|      1.0|
|2005|   10|  false|     8.688417|      1.0|       9.0|        27|           39.0|        false|      false|              47.0|             0.0|            0.0|      0.0|      1.0|
|2005|    9|  false|    6.9996767|      1.0|       9.0|        20|           40.0|        false|      false|              42.0|             0.0|            0.0|      0.0|      1.0|
|2005|    9|  false|    8.0623045|      1.0|       9.0|        35|           40.0|        false|      false|              12.0|             0.0|            0.0|      0.0|      1.0|
|2005|    2|   true|     5.176454|      1.0|       8.0|        24|           36.0|        false|      false|               9.0|             0.0|            0.0|      0.0|      1.0|
|2005|    5|  false|    7.5508323|      1.0|       9.0|        32|           40.0|        false|      false|              30.0|             0.0|            0.0|      0.0|      1.0|
|2005|    1|  false|    5.8863425|      1.0|       9.0|        30|           37.0|        false|      false|              31.0|             0.0|            0.0|      1.0|      1.0|
|2005|   11|   true|     5.000084|      1.0|       9.0|        18|           37.0|        false|      false|              63.0|             0.0|            0.0|      1.0|      1.0|
|2005|   11|  false|     6.448521|      1.0|       9.0|        19|           38.0|        false|      false|              20.0|             0.0|            0.0|      0.0|      1.0|
|2005|    8|   true|    6.6028447|      1.0|       9.0|        20|           40.0|        false|      false|              20.0|             0.0|            0.0|      0.0|      1.0|
|2005|    9|  false|    6.3757687|      1.0|       9.0|        21|           40.0|        false|      false|              53.0|             0.0|            0.0|      0.0|      1.0|
|2005|    9|  false|     8.562755|      1.0|       8.0|        19|           39.0|        false|      false|              14.0|             0.0|            0.0|      0.0|      1.0|
|2005|    6|   true|      7.18707|      1.0|       9.0|        22|           38.0|        false|      false|              54.0|             0.0|            0.0|      0.0|      1.0|
|2005|    1|   true|     7.438397|      1.0|       9.0|        26|           39.0|        false|      false|              35.0|             0.0|            0.0|      0.0|      1.0|
|2005|   12|  false|     7.874912|      1.0|       9.0|        27|           38.0|        false|      false|              23.0|             0.0|            0.0|      0.0|      1.0|
|2005|    7|   true|     8.249698|      1.0|       9.0|        32|           41.0|        false|      false|              58.0|             0.0|            0.0|      0.0|      1.0|
|2005|    4|   true|    7.6257896|      1.0|       8.0|        32|           43.0|        false|      false|              65.0|             0.0|            0.0|      0.0|      1.0|
|2005|    3|  false|         NULL|      1.0|       8.0|        29|           37.0|        false|      false|              34.0|             0.0|            0.0|      0.0|      1.0|
+----+-----+-------+-------------+---------+----------+----------+---------------+-------------+-----------+------------------+----------------+---------------+---------+---------+----
only showing top 20 rows
```

Figure 3: clean data

6 CONCLUSION

This study has demonstrated the value of applying data preprocessing techniques and machine learning models to the US CDC Natality dataset (1969-2008). Our analysis has uncovered significant trends in birth outcomes, maternal health shifts over four decades. Our preprocessing methodologies have successfully addressed challenges such as missing data, coding inconsistencies, and evolving data collection practices. The application of machine learning models has enabled us to predict birth outcomes with improved accuracy and identify key factors influencing maternal and infant health. This research contributes to public health, demography, and data science by showcasing the power of big data analytics in understanding long-term population health trends. The insights gained have potential implications for public health policy, healthcare resource allocation, and targeted interventions to improve maternal and infant health outcomes.

7 FUTURE SCOPE

Extending the time series analysis to incorporate post-2008 data would allow examination of recent trends and potential forecasting. Integration with other datasets, such as socioeconomic indicators or environmental factors, could provide a more comprehensive understanding of influences on birth outcomes. Exploring advanced machine learning techniques, including deep learning may uncover more nuanced patterns in the data. Developing real-time data processing systems could enable more timely interventions and policy adjustments. Ethical considerations, including bias mitigation in AI models, should be a priority as these tools become more widely adopted in healthcare settings. Finally, fostering interdisciplinary collaborations between data scientists, public health experts, policymakers, and healthcare providers will be crucial to translate insights into actionable strategies.