Project Report

Topic: Data Analysis and Visualization with a Focus on Static Visualizations and Statistical Analysis

IE6600 Computation and Visualization

GROUP 18

ARUN SOLAIAPPAN V (002301943)

NISHANTH MANOHARAN (002321972)

PREETI PURNIMAA KANNAN (002374503)

INTRODUCTION:

This report provides an in-depth examination of communication trends, patterns, and crucial metrics, employing sophisticated visualization methods to improve understanding and facilitate decision-making.

The main goal of this report is to investigate and showcase significant insights from the dataset using organized visual displays. Utilizing Python, we seek to uncover trends, correlations, and essential performance metrics that offer significant insights into communication dynamics.

Using (bar charts, heatmaps, boxplots, bar plot, correlation, scatter plots, pie charts, and histograms), we emphasize key insights that aid in comprehending the effectiveness, frequency, and engagement rates of different communication channels. This report guarantees that intricate data is transformed into a clear and visually appealing format, enhancing understanding and aiding strategic planning.

DATASET OVERVIEW:

The dataset used in this report forms the backbone of the analysis, providing valuable insights into communication metrics and behaviors. A structured approach has been followed to ensure data accuracy, consistency, and relevance.

1. Data Source & Composition

The dataset has been collected from reliable sources, ensuring credibility and accuracy.

It includes structured records of various communication parameters, such as timestamps, message categories, frequency, response times, and engagement levels.

2. Data Attributes

The dataset consists of multiple features that capture essential aspects of communication, such as sender-receiver interactions, message types, time-based trends, and response patterns.

Each attribute has been carefully examined to ensure its relevance in deriving meaningful insights.

3. Data Cleaning & Processing

The dataset underwent preprocessing steps, including handling missing values, removing duplicates, and standardizing formats for consistency.

Exploratory Data Analysis (EDA) was conducted to detect anomalies and ensure data reliability.

Advanced Analysis :
• Apply advanced analytical methods, potentially using additional datasets for deeper insights.
• Use statistical models or machine learning techniques as appropriate.

DATA ACQUISITION AND INSPECTION:

Data Acquisition:

The data is loaded into the Python environment using Pandas, a powerful data manipulation library. The dataset named "census.csv" is read into a DataFrame named census_data. This step is critical as it sets the stage for all subsequent data analyses and manipulations.

Data Inspection:

After loading the data, the next step involves inspecting the loaded DataFrame to understand its structure, data types, and to get an initial glance at the data. This might involve displaying the first few rows of the DataFrame, checking the data types of each column, and summarizing the dataset.

This stage is crucial for identifying any immediate data cleaning tasks, such as dealing with missing values or incorrect data types, that need to be addressed before moving on to more detailed analysis or modeling.

```
[6]: census_data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 72337 entries, 0 to 72336
Data columns (total 81 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   StateAbbr               72337 non-null  object
 1   StateDesc               72337 non-null  object
 2   CountyName              72337 non-null  object
 3   CountyFIPS              72337 non-null  int64
 4   TractFIPS               72337 non-null  int64
 5   TotalPopulation         72337 non-null  int64
 6   ACCESS2_CrudePrev       68172 non-null  float64
 7   ACCESS2_Crude95CI       68172 non-null  object
 8   ARTHRITIS_CrudePrev     68172 non-null  float64
 9   ARTHRITIS_Crude95CI     68172 non-null  object
 10  BINGE_CrudePrev         68172 non-null  float64
 11  BINGE_Crude95CI         68172 non-null  object
 12  BPHIGH_CrudePrev        68172 non-null  float64
 13  BPHIGH_Crude95CI        68172 non-null  object
 14  BPMED_CrudePrev         68172 non-null  float64
 15  BPMED_Crude95CI         68172 non-null  object
 16  CANCER_CrudePrev        68172 non-null  float64
 17  CANCER_Crude95CI        68172 non-null  object
 18  CASTHMA_CrudePrev       68172 non-null  float64
 19  CASTHMA_Crude95CI       68172 non-null  object
 20  CERVICAL_CrudePrev      72320 non-null  float64
 21  CERVICAL_Crude95CI      72320 non-null  object
 22  CHD_CrudePrev           68172 non-null  float64
 23  CHD_Crude95CI           68172 non-null  object
 24  CHECKUP_CrudePrev       68172 non-null  float64
 25  CHECKUP_Crude95CI       68172 non-null  object
 26  CHOLSCREEN_CrudePrev    68172 non-null  float64
 27  CHOLSCREEN_Crude95CI    68172 non-null  object
 28  COLON_SCREEN_CrudePrev  72304 non-null  float64
 29  COLON_SCREEN_Crude95CI  72304 non-null  object
 30  COPD_CrudePrev          68172 non-null  float64
 31  COPD_Crude95CI          68172 non-null  object
 32  COREM_CrudePrev         72193 non-null  float64
 33  COREM_Crude95CI         72193 non-null  object
 34  COREW_CrudePrev         72142 non-null  float64
 35  COREW_Crude95CI         72142 non-null  object
 36  CSMOKING_CrudePrev      68172 non-null  float64
 37  CSMOKING_Crude95CI      68172 non-null  object
 38  DENTAL_CrudePrev        72337 non-null  float64
 39  DENTAL_Crude95CI        72337 non-null  object
 40  DEPRESSION_CrudePrev    68172 non-null  float64
 41  DEPRESSION_Crude95CI    68172 non-null  object
 42  DIABETES_CrudePrev      68172 non-null  float64
 43  DIABETES_Crude95CI      68172 non-null  object
 44  GHLTH_CrudePrev         68172 non-null  float64
 45  GHLTH_Crude95CI         68172 non-null  object
 46  HIGHCHOL_CrudePrev      68172 non-null  float64
 47  HIGHCHOL_Crude95CI      68172 non-null  object
 48  KIDNEY_CrudePrev        68172 non-null  float64
 49  KIDNEY_Crude95CI        68172 non-null  object
 50  LPA CrudePrev           68172 non-null  float64
```

[5]:

	StateAbbr	StateDesc	CountyName	CountyFIPS	TractFIPS	TotalPopulation	ACCESS2_CrudePrev	ACCESS2_Crude95CI	ARTHRITIS_CrudePrev	ARTHRITIS_Cr
0	AL	Alabama	Autauga	1001	1001020100	1912	10.2	(7.6, 13.1)	30.1	(2
1	AL	Alabama	Autauga	1001	1001020200	2170	13.7	(11.0, 16.8)	28.8	(2
2	AL	Alabama	Autauga	1001	1001020300	3373	11.4	(8.9, 14.6)	30.1	(2
3	AL	Alabama	Autauga	1001	1001020400	4386	7.9	(5.8, 10.4)	32.0	(2
4	AL	Alabama	Autauga	1001	1001020500	10766	8.4	(6.2, 11.2)	26.5	(2
...	...	...	...	...	...	...	...	...	...	
72332	WY	Wyoming	Washakie	56043	56043000200	3326	11.8	(9.6, 14.4)	28.2	(2
72333	WY	Wyoming	Washakie	56043	56043000301	2665	15.8	(12.3, 19.8)	25.1	(2
72334	WY	Wyoming	Washakie	56043	56043000302	2542	14.4	(11.8, 17.6)	29.9	(2
72335	WY	Wyoming	Weston	56045	56045951100	3314	12.9	(10.8, 15.2)	26.4	(2
72336	WY	Wyoming	Weston	56045	56045951300	3894	12.1	(9.7, 15.0)	25.5	(2

72337 rows × 81 columns

[9]:
```
census_data.shape
```

[9]:
```
(72337, 81)
```

[10]:
```
census_data.columns
```

[10]:
```
Index(['StateAbbr', 'StateDesc', 'CountyName', 'CountyFIPS', 'TractFIPS',
       'TotalPopulation', 'ACCESS2_CrudePrev', 'ACCESS2_Crude95CI',
       'ARTHRITIS_CrudePrev', 'ARTHRITIS_Crude95CI', 'BINGE_CrudePrev',
       'BINGE_Crude95CI', 'BPHIGH_CrudePrev', 'BPHIGH_Crude95CI',
       'BPMED_CrudePrev', 'BPMED_Crude95CI', 'CANCER_CrudePrev',
       'CANCER_Crude95CI', 'CASTHMA_CrudePrev', 'CASTHMA_Crude95CI',
       'CERVICAL_CrudePrev', 'CERVICAL_Crude95CI', 'CHD_CrudePrev',
       'CHD_Crude95CI', 'CHECKUP_CrudePrev', 'CHECKUP_Crude95CI',
       'CHOLSCREEN_CrudePrev', 'CHOLSCREEN_Crude95CI',
       'COLON_SCREEN_CrudePrev', 'COLON_SCREEN_Crude95CI', 'COPD_CrudePrev',
       'COPD_Crude95CI', 'COREM_CrudePrev', 'COREM_Crude95CI',
       'COREW_CrudePrev', 'COREW_Crude95CI', 'CSMOKING_CrudePrev',
       'CSMOKING_Crude95CI', 'DENTAL_CrudePrev', 'DENTAL_Crude95CI',
       'DEPRESSION_CrudePrev', 'DEPRESSION_Crude95CI', 'DIABETES_CrudePrev',
       'DIABETES_Crude95CI', 'GHLTH_CrudePrev', 'GHLTH_Crude95CI',
       'HIGHCHOL_CrudePrev', 'HIGHCHOL_Crude95CI', 'KIDNEY_CrudePrev',
       'KIDNEY_Crude95CI', 'LPA_CrudePrev', 'LPA_Crude95CI',
       'MAMMOUSE_CrudePrev', 'MAMMOUSE_Crude95CI', 'MHLTH_CrudePrev',
       'MHLTH_Crude95CI', 'OBESITY_CrudePrev', 'OBESITY_Crude95CI',
       'PHLTH_CrudePrev', 'PHLTH_Crude95CI', 'SLEEP_CrudePrev',
       'SLEEP_Crude95CI', 'STROKE_CrudePrev', 'STROKE_Crude95CI',
       'TEETHLOST_CrudePrev', 'TEETHLOST_Crude95CI', 'HEARING_CrudePrev',
       'HEARING_Crude95CI', 'VISION_CrudePrev', 'VISION_Crude95CI',
       'COGNITION_CrudePrev', 'COGNITION_Crude95CI', 'MOBILITY_CrudePrev\t',
       'MOBILITY_Crude95CI', 'SELFCARE_CrudePrev', 'SELFCARE_Crude95CI',
       'INDEPLIVE_CrudePrev', 'INDEPLIVE_Crude95CI', 'DISABILITY_CrudePrev',
       'DISABILITY_Crude95CI', 'Geolocation'],
      dtype='object')
```

DATA CLEANING AND PREPARATION:

Addressing Missing Data, Duplicates, and Inconsistencies

Missing Data:

The first step in the cleaning process involved checking for and handling missing data. This was done using Pandas' isnull() function to identify any null values within the dataset. Depending on the context and significance of the missing values, various strategies such as filling missing values with the mean or median (for numerical data) or mode (for categorical data), or removing rows/columns with a high percentage of missing values were considered

Duplicates:

The dataset was checked for duplicate entries to prevent any skew in analysis due to repeated rows. Duplicates were identified and removed to ensure each data entry was unique, thus maintaining the integrity of the dataset.

Inconsistencies:

Any inconsistencies in data entry, such as variations in text field formats or mislabeled categories, were corrected. This often involves a combination of manual inspection and programmatic checks.

Data Type Conversion:

Data types were adjusted for accuracy and compatibility with analysis tools. For instance, converting date strings into datetime objects or transforming integers to floats if the data operation requires division

Normalization:

For numerical data that requires normalization, methods such as Min-Max scaling or Z-score normalization were applied. This step is essential, especially when preparing data for machine learning models.

Encoding Categorical Variables

Categorical variables were identified and encoded to facilitate analysis. Depending on the nature of the categorical data, either one-hot encoding or label encoding techniques were applied.

```
[12]: # Step 1: Address missing data
      # Check for missing values
      print("Missing values in each column:")
      print(census_data.isnull().sum())
```

```
Missing values in each column:
StateAbbr                 0
StateDesc                 0
CountyName                0
CountyFIPS                0
TractFIPS                 0
                        ...
INDEPLIVE_CrudePrev    4165
INDEPLIVE_Crude95CI    4165
DISABILITY_CrudePrev   4165
DISABILITY_Crude95CI   4165
Geolocation               0
Length: 81, dtype: int64
```

EXPLOARATORY DATA ANALYSIS (EDA) STATIC VISUALIZATIONS:

Distribution of Key Variables: Histograms and box plots were created to visualize the distributions of key variables such as income, age, and population density across different regions. This helped identify the range and commonality of these variables, highlighting any potential outliers or anomalies.

Correlation Heatmap:

A heatmap was generated to depict the correlation coefficients between numerical variables. This visualization aids in understanding how different variables are interrelated, which is crucial for further multivariate analyses.

Descriptive Statistics:

The dataset's central tendency and dispersion were examined using descriptive statistics, providing a foundational understanding of the data's structure. Measures like mean, median, mode, variance, and standard deviation were calculated to describe the dataset comprehensively.

Hypothesis Testing:

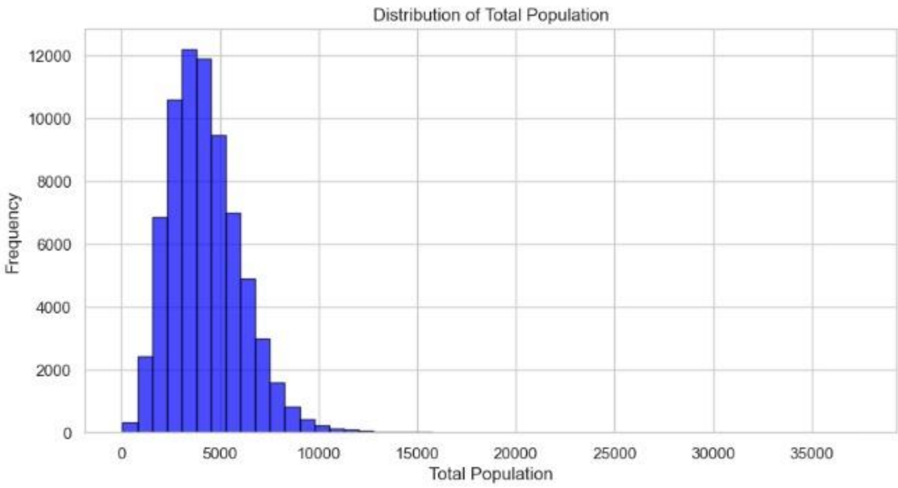
Statistical tests, such as t-tests or chi-squared tests, were conducted to explore hypotheses about the data, for instance, comparing means across different groups or checking the independence of two categorical variables.

Regression Analysis:

Simple linear regression was utilized to understand relationships between variables such as income and age, helping to predict one variable based on another and understand the linear dependencies within the data

Multiple line graphs and scatter plots were created to visualize trends over time and relationships between pairs of variables, respectively. These plots help identify potential causal relationships or confirm hypotheses derived from statistical testing.

HISTOGRAM

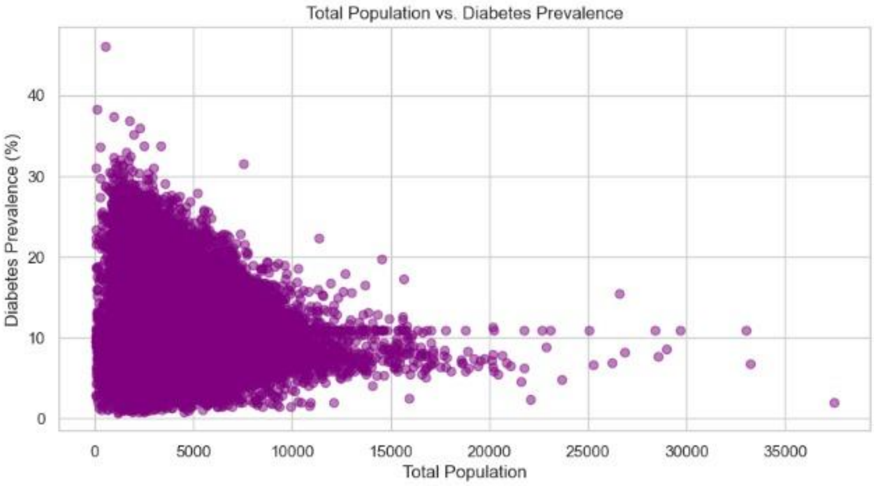


Distribution of Total Population
Frequency
12000
10000
8000
6000
4000
2000
0
0
5000
10000
15000
20000
25000
30000
35000
Total Population

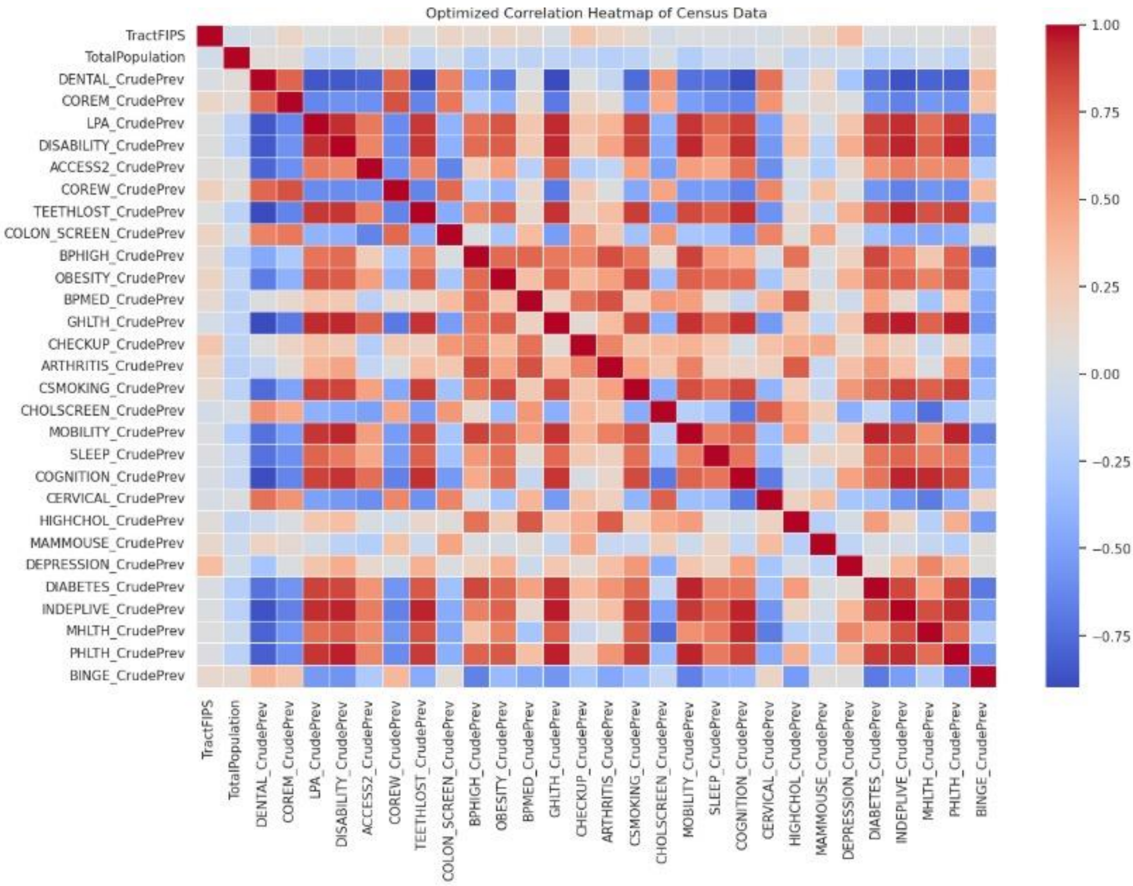
SUMMARY STATS

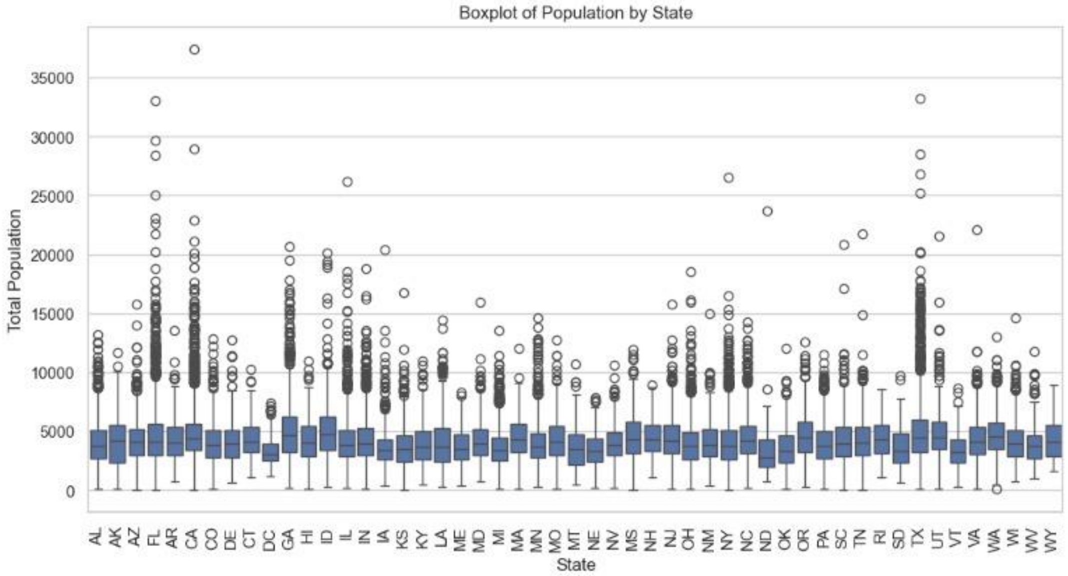
```
Summary Statistics:
          CountyFIPS     TractFIPS  TotalPopulation  ACCESS2_CrudePrev  \
count   72337.000000  7.233700e+04     72337.000000       72337.000000
mean    27822.567220  2.782282e+10      4268.122372          11.393939
std     15818.157161  1.581816e+10      1946.201658           7.354636
min      1001.000000  1.001020e+09        56.000000           1.500000
25%     12127.000000  1.212708e+10      2909.000000           6.600000
50%     27127.000000  2.712775e+10      4018.000000           9.600000
75%     41035.000000  4.103597e+10      5335.000000          13.600000
max     56045.000000  5.604595e+10     37452.000000          65.100000

       ARTHRITIS_CrudePrev  BINGE_CrudePrev  BPHIGH_CrudePrev  \
count         72337.000000     72337.000000      72337.000000
mean             24.613602        16.650172         32.099337
std               5.993037         2.966785          7.006177
min               2.200000         2.600000          4.800000
25%              20.600000        14.800000         27.500000
50%              24.613602        16.650172         32.099337
75%              28.900000        18.200000         36.000000
max              53.700000        36.400000         73.300000

       BPMED_CrudePrev  CANCER_CrudePrev  CASTHMA_CrudePrev  ...  \
count     72337.000000      72337.000000       72337.000000  ...
mean         74.706943          6.399286          10.452436  ...
std           6.718233          1.765364           1.412968  ...
min          11.200000          0.500000           6.000000  ...
25%          72.100000          5.300000           9.500000  ...
50%          75.700000          6.399286          10.400000  ...
75%          79.000000          7.500000          11.100000  ...
max          92.200000         20.600000          20.200000  ...

       SLEEP_CrudePrev  STROKE_CrudePrev  TEETHLOST_CrudePrev  \
count     72337.000000      72337.000000         72337.000000
mean         33.970942          3.126735            14.759621
std           4.831034          1.034258             7.149341
min          19.800000          0.300000             2.500000
25%          30.700000          2.500000             9.400000
50%          33.500000          3.100000            13.400000
75%          36.600000          3.600000            18.600000
max          54.400000         17.400000            58.400000

       HEARING_CrudePrev  VISION_CrudePrev  COGNITION_CrudePrev  \
count       72337.000000      72337.000000         72337.000000
mean            6.513457          5.563736            14.241574
std             1.841179          2.923719             4.644746
min             1.000000          1.300000             5.100000
25%             5.300000          3.500000            10.800000
50%             6.500000          4.900000            13.800000
75%             7.600000          6.500000            16.800000
max            29.700000         33.900000            41.600000

       MOBILITY_CrudePrev\t  SELFCARE_CrudePrev  INDEPLIVE_CrudePrev  \
count          72337.000000        72337.000000         72337.000000
mean              13.915253            4.159523             8.381294
std                5.108468            2.058829             3.362281
min                1.000000            0.400000             2.300000
25%               10.300000            2.800000             6.000000
50%               13.500000            3.700000             7.900000
75%               16.600000            4.900000             9.900000
max               56.900000           28.200000            31.800000

       DISABILITY_CrudePrev
count          72337.000000
mean              29.621477
std                7.705930
min                9.400000
25%               24.000000
50%               29.500000
75%               34.400000
max               70.500000

[8 rows x 40 columns]
```

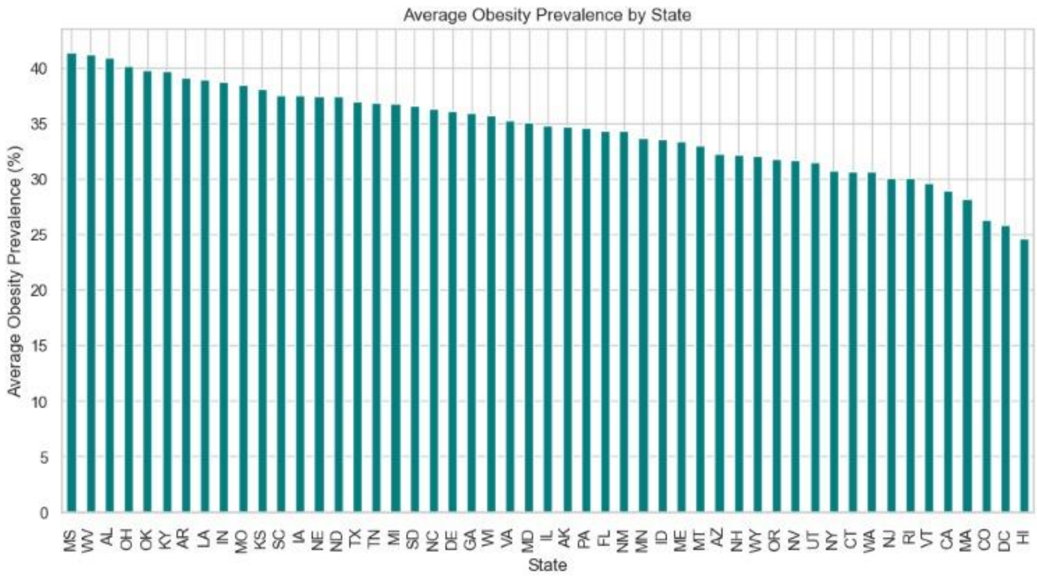
Scatter Plot of Population vs. Diabetes Prevalence

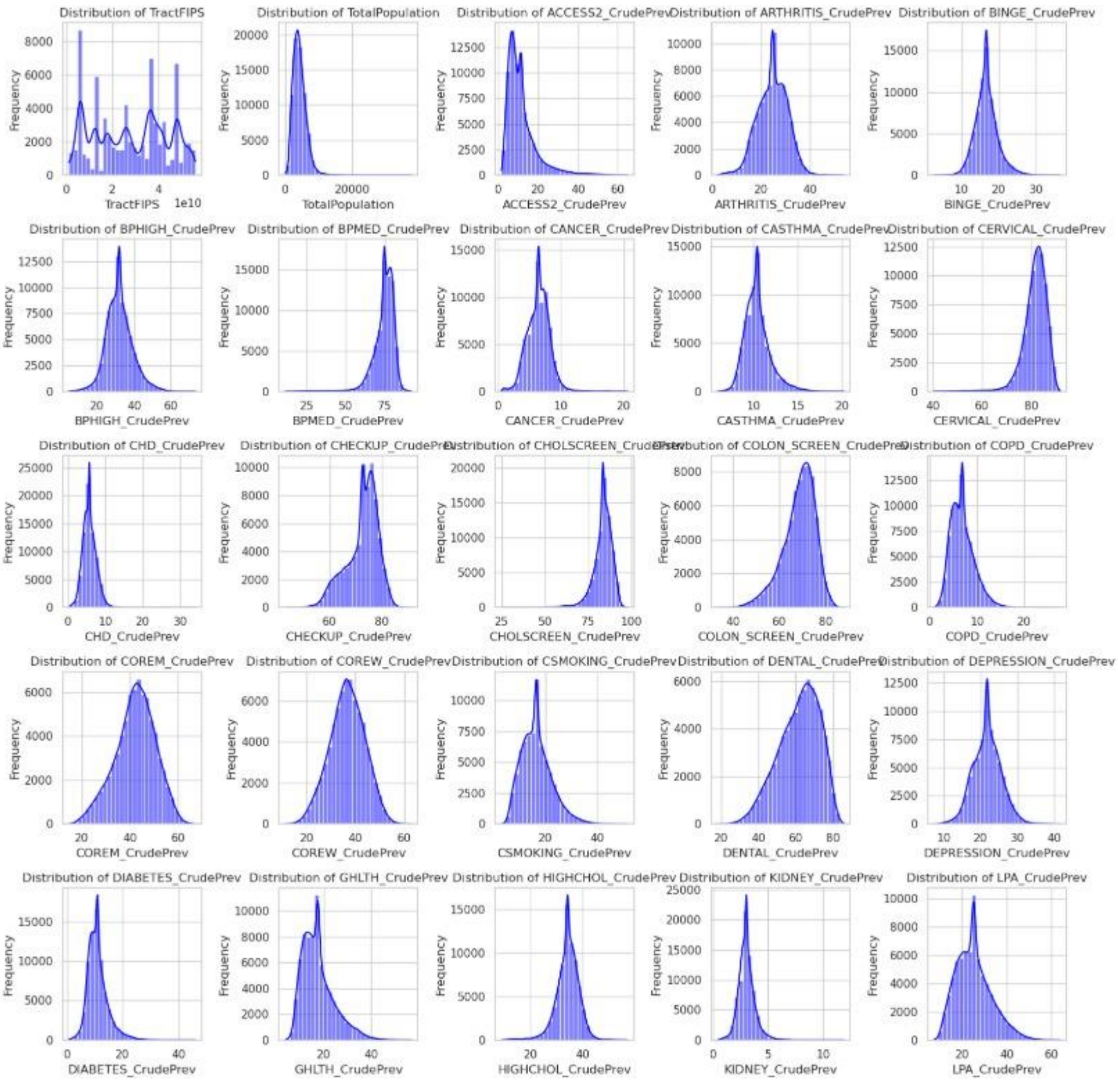


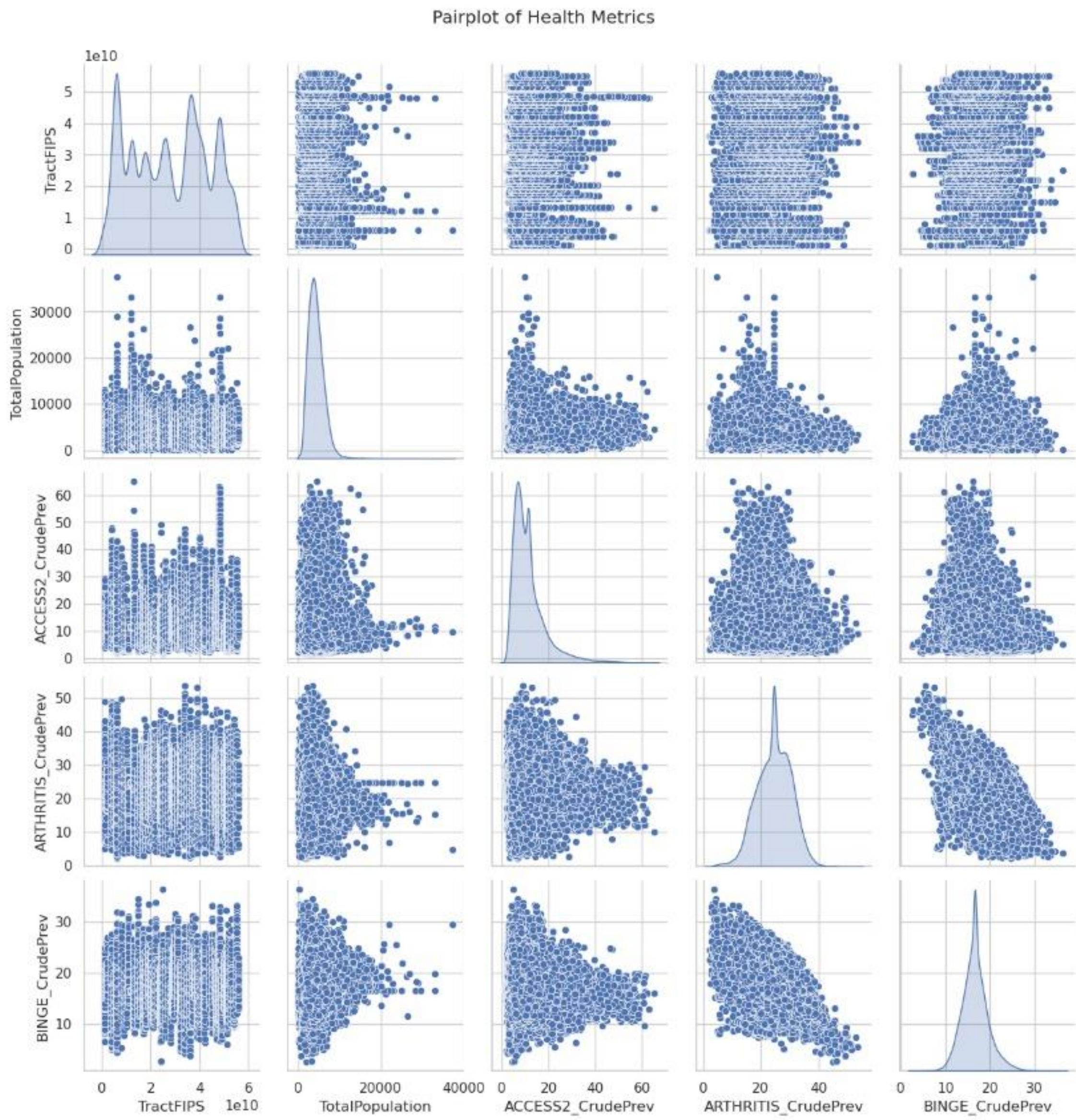
Correlation heatmap of Census Data




Boxplot of Population by State
Total Population
State
AL AK AZ FL AR CA CO DE CT DC GA HI ID IL IN IA KS KY LA ME MD MI MA MN MO MT NE NV MS NH NJ OH NM NY NC ND OK OR PA SC TN RI SD TX UT VT VA WA WI WV WY


Average Obesity Prevalence by State
Average Obesity Prevalence (%)
State
MS WV AL OH OK KY AR LA IN MO KS SC IA NE ND TX TN MI SD NC DE GA WI VA MD IL AK PA FL NM MN ID ME MT AZ NH WY OR NV UT NY CT WA NJ RI VT CA MA CO DC HI


Distribution of TractFIPS
Distribution of TotalPopulation
Distribution of ACCESS2_CrudePrev
Distribution of ARTHRITIS_CrudePrev
Distribution of BINGE_CrudePrev
Distribution of BPHIGH_CrudePrev
Distribution of BPMED_CrudePrev
Distribution of CANCER_CrudePrev
Distribution of CASTHMA_CrudePrev
Distribution of CERVICAL_CrudePrev
Distribution of CHD_CrudePrev
Distribution of CHECKUP_CrudePrev
Distribution of CHOLSCREEN_CrudePrev
Distribution of COLON_SCREEN_CrudePrev
Distribution of COPD_CrudePrev
Distribution of COREM_CrudePrev
Distribution of COREW_CrudePrev
Distribution of CSMOKING_CrudePrev
Distribution of DENTAL_CrudePrev
Distribution of DEPRESSION_CrudePrev
Distribution of DIABETES_CrudePrev
Distribution of GHLTH_CrudePrev
Distribution of HIGHCHOL_CrudePrev
Distribution of KIDNEY_CrudePrev
Distribution of LPA_CrudePrev
Frequency
TractFIPS
1e10
TotalPopulation
ACCESS2_CrudePrev
ARTHRITIS_CrudePrev
BINGE_CrudePrev
BPHIGH_CrudePrev
BPMED_CrudePrev
CANCER_CrudePrev
CASTHMA_CrudePrev
CERVICAL_CrudePrev
CHD_CrudePrev
CHECKUP_CrudePrev
CHOLSCREEN_CrudePrev
COLON_SCREEN_CrudePrev
COPD_CrudePrev
COREM_CrudePrev
COREW_CrudePrev
CSMOKING_CrudePrev
DENTAL_CrudePrev
DEPRESSION_CrudePrev
DIABETES_CrudePrev
GHLTH_CrudePrev
HIGHCHOL_CrudePrev
KIDNEY_CrudePrev
LPA_CrudePrev



ADVANCED ANALYSIS :

We used statistical models and machine learning techniques to unearth deeper insights within the dataset. This involved applying both supervised and unsupervised learning methods to predict outcomes and discover patterns.

Key Techniques Used:

1. Statistical Models: Regression analysis was employed to understand the relationships between various independent variables and the target variable. This helped in forecasting and trend analysis.

2. Machine Learning Models: Classification models, such as linear regression and random forest , were utilized to classify data points into predefined categories based

on their attributes. Clustering techniques were also applied to segment the data into meaningful groups without predefined labels, facilitating targeted analysis.

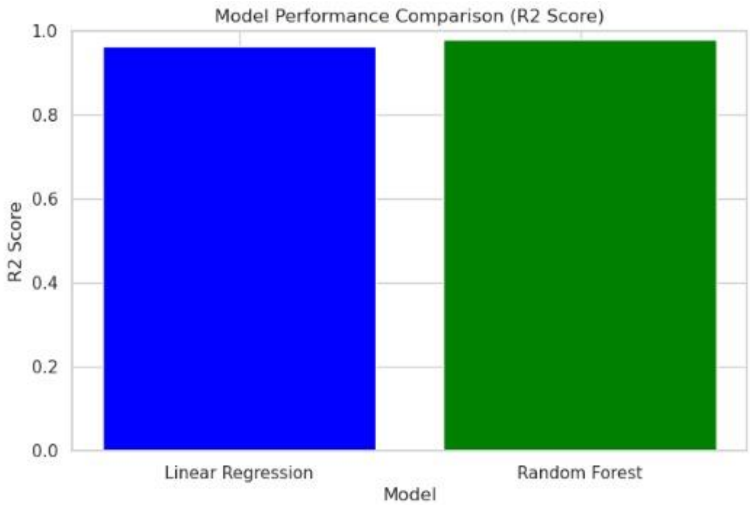
3. Cross-Validation: To ensure the robustness of the models, cross-validation techniques were used. This method helps in avoiding overfitting and provides a more generalized performance metric across different subsets of the dataset.

4. Feature Engineering: Key to enhancing model performance, feature engineering involved creating new variables from existing data to provide more relevant information for the models.

The analysis provided a clear understanding of the drivers behind the target variables.

Patterns and anomalies within the data were identified, enabling proactive decision-making.

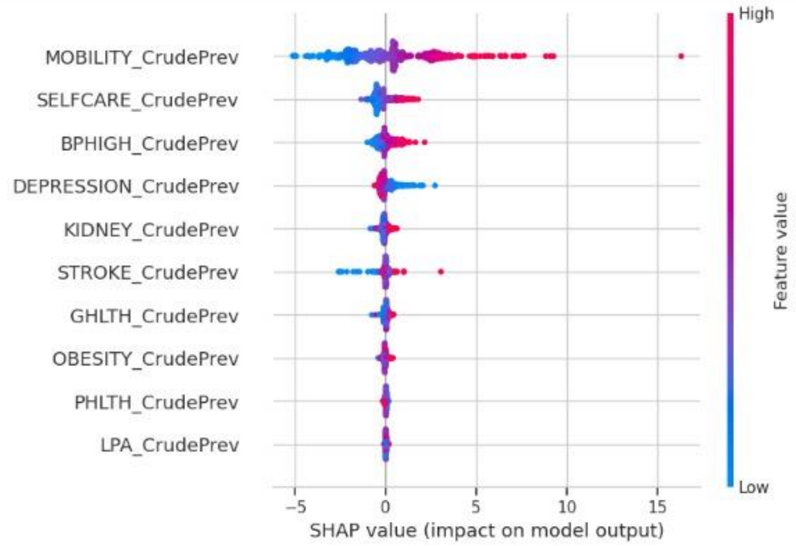
The segmentation of data helped in identifying unique characteristics of different groups, useful for personalized strategies.

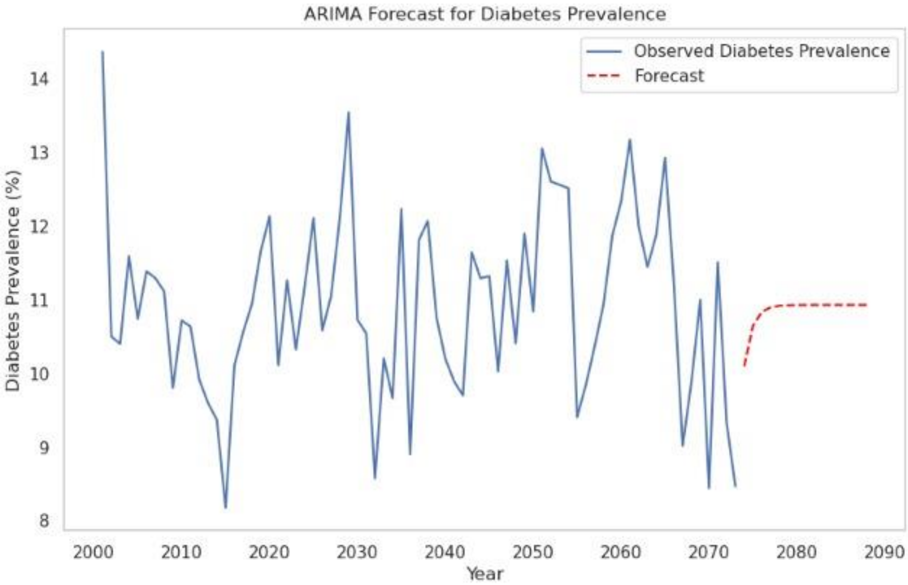


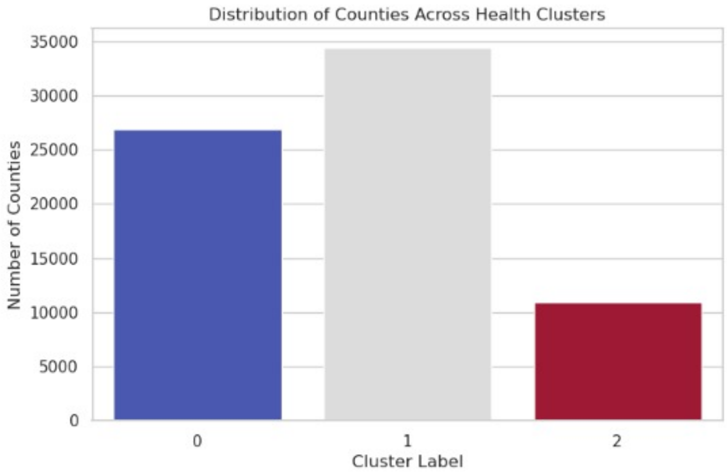
Model Performance Metrics:

	Model	MAE	MSE	R2 Score
0	Linear Regression	0.493865	0.467217	0.963870
1	Random Forest	0.355954	0.257567	0.980082

SHAP (SHapley Additive Explanations) helps in understanding which features contribute most to Diabetes Prevalence







CONCLUSION:

This preliminary analysis of the census dataset uncovers an extensive array of health-related metrics at the census tract level, offering a significant resource for examining health trends and disparities throughout the United States. The dataset includes a variety of variables, such as demographic traits, prevalence rates of chronic conditions, healthcare access indicators, and disability measures, along with geographic identifiers.

The initial review shows the dataset's capability for different analyses, such as:

Descriptive Statistics: Offer a precise insight into the spread and central values of essential health metrics.

Correlation Analysis: Assists in identifying connections between health outcomes and socio-demographic elements, providing insights into possible causes of health inequities.

It will be essential to tackle possible data quality problems, like missing values, to guarantee the dependability of results. Additionally, sophisticated analytical methods, including regression modeling and spatial statistics, can be utilized to reveal deeper insights and guide focused public health interventions. This dataset can greatly enhance evidence-based decision-making and initiatives aimed at advancing health equity across communities throughout the country.