Citizen Friendly Report of diversitydatakids.org

Anil Gubbala, Ali Hussain Ladiwala, Urja Naik and Priyanka Cornelius
Department of Software Engineering, San Jose State University´
San Jose, CA´
anil.gubbala@sjsu.edu, alihussain.ladiwala@sjsu.edu, urjamitulkumar.naik@sjsu.edu,
priyanka.cornelius@sjsu.edu

Abstract—Using Divrsitykids datasets, create a dashboard and use NLG to generate a citizen friendly report.About Diversitydatakids.org: It is a research project that examines who our children are, whether they have what they need to grow up healthy and achieve their full potential, whether social policies are well designed to improve children's lives and how to make them better to improve equity.Data collected by diversitydatakids is in tabular format, which makes it difficult for a general citizen to understand the content.Our goal is to convert few topics of this tabular content to a citizen friendly report so that it becomes more readable & helps increase awareness among citizens.

I. Introduction

Early childhood is a period of rapid development for children. Early childhood experiences at home, in school and in their neighborhoods shape their ability to thrive in childhood and beyond. diversitydatakids.org examines equity (or inequity) in children's early experiences and in policies that support their wellbeing in early childhood, such as early care and education so that policymakers and practitioners can build on the work as they consider expansions of existing early childhood programs and to inform new federal, state and local policies. However users from non-engineering backgrounds might find it cumbersome to interpret these extensive datasets as numerical or tabular data is not intuitive to the brain. Our attempt is to obtain the data in citizen friendly textual format so as to augment its impact.

II. Architecture

There are 5 major components:
Dataset on diversitydatakids.org which was fetched using CKAN's API , Processing & filtering this data in node.js & react.js, Natural Language Generation library RosaeNLG based on pug templates, Display tabular data and generated text in frontend, Deploy application on AWS server.

CKAN API is used to fetch datasets that are used in the diversitydatakids.org website and then do data filtering & data extraction. The filtered data is sent to the Node backend server which pushes the data to RosaeNLG that uses Pug templates for NLG structuring.

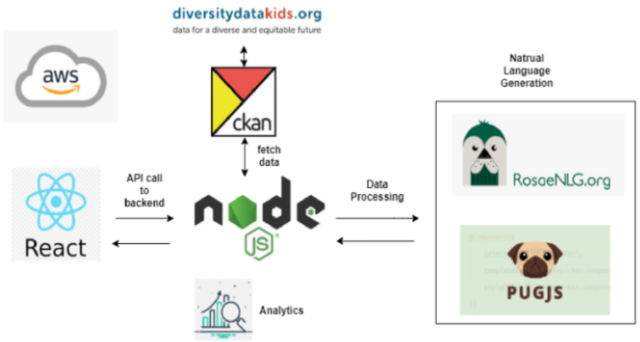


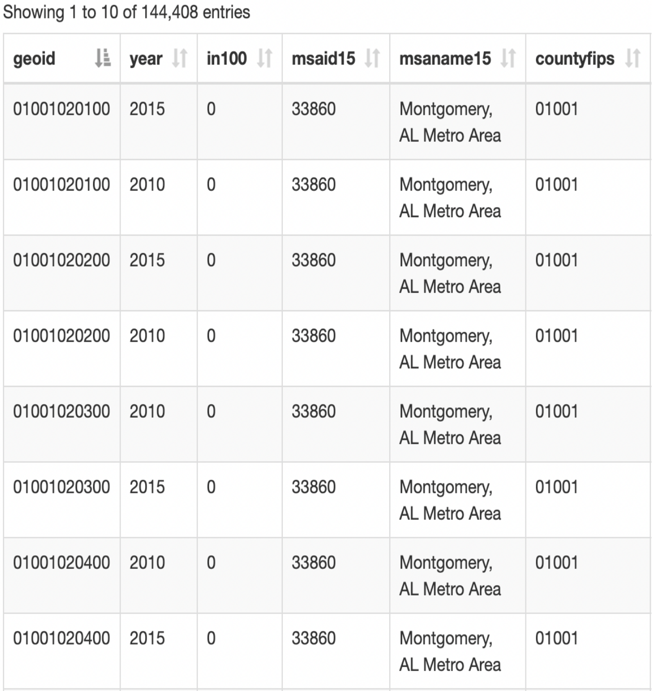
Fig. 1. Project Architectures

III. Data Analysis

Established in 2014 with support from the W.K. Kellogg Foundation and the Robert Wood Johnson Foundation, diversitydatakids.org has 323 datasets, subdivided into Topics such as Child Opportunity Index, Demographics, Income, Education, Family, Health, Housing, Early Childhood, Neighborhood and Work. The datasets can further be filtered on the basis of 'Availability by Race and Ethnicity', 'Nativity', 'Age Group', 'Gender', Geography' and 'Time Scale'. The dataset is publicly accessible from the author's website [1].

A. Data Distribution

The dataset available is in the form of csv files & JSON, categorised into several topics and subtopics like 'Early Childhood', 'Income, Work and Family' and 'Neighbourhoods' etc. In total there are over 6000+ datasets available in the database and can be fetched using CKAN data API. Majority of the datasets contain data examining how racial/ethnic and income segregation in neighborhoods and early childhood programs contribute to inequities in children's early experiences and opportunities. Column titles are tags that have been described in a separate table and include geographic ID, region and numerical values for opportunity indices termed as raw indicator values. Sample Data for Child Opportunity Index is shown in Fig 2.



Showing 1 to 10 of 144,408 entries

geoid	year	in100	msaid15	msaname15	countyfips
01001020100	2015	0	33860	Montgomery, AL Metro Area	01001
01001020100	2010	0	33860	Montgomery, AL Metro Area	01001
01001020200	2015	0	33860	Montgomery, AL Metro Area	01001
01001020200	2010	0	33860	Montgomery, AL Metro Area	01001
01001020300	2010	0	33860	Montgomery, AL Metro Area	01001
01001020300	2015	0	33860	Montgomery, AL Metro Area	01001
01001020400	2010	0	33860	Montgomery, AL Metro Area	01001
01001020400	2015	0	33860	Montgomery, AL Metro Area	01001

.

Fig.2. COI 2.0 Index Data

B. Fetching Dataset

Fetching a dataset is not a straightforward process. Inorder to fetch the dataset first we have to select a subtopic and then selection geographic region to get the dataset. In datastore at present there are over 320 subtopics available. Only alternative is to enter the exact resource ID and package ID of the dataset to access that particular dataset and view the data. Our implementation includes both these approaches.

C. Dataset Column titles

The column titles in the datasets on diversitydatakids.org are column tags or labels that have been separately described in a table. In order to use these column headers for natural language generation, they have been converted into a more descriptive format that clearly explains the contents of the column.

D. Dataset statistics

The CKAN data API supports several forms of querying including Javascript , python and URI. In addition, it also supports inclusion of direct SQL query & Apache Solr querying for a few API calls. So we make use of these features to extract summary of the datasets using aggregate functions in SQL.

E. Text Decoding

Even though there are over 6000 datasets available in the datastore, the structural format of all the datasets is the same for all the datasets. Every dataset also maintains some keywords specifying to which topics, subtopics or categories the current dataset represents. There are around 10 major categories which are further categorized to several other sub categories. This structural uniformity made it easy to extract data from the dataset required for NLG for almost all the datasets available in the datastore.

F. Keywords in NLG

The filtered and extracted data is sent to RosaeNLG where the actual text generation is implemented. The keywords & the metadata information in each dataset is used here to generate the text in RosaeNLG. This data is used in the major 5 stages of Natural language generation to generate the text[3]. Few of the text used in the NLG is static but most of the text is dynamically imported. So the text generation supports almost all the datasets in datastore.

IV. ARCHITECTURAL COMPONENTS

Natural Language Generation tools and libraries require structured data in order to convert it to textual format. We obtain the required structured data in the form of JSON using CKAN data API. We fetch the data from the website using CKAN API employing SQL, Apache Solr queries and then format the JSON output in a Pug template that RosaeNLG uses to generate natural language text.

A. CKAN data API

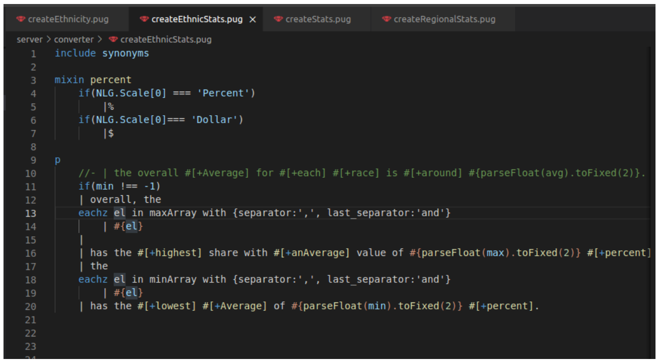
CKAN is an open-source DMS (data management system) for storage and distribution of open data. Instead of a database, CKAN API has been used to fetch data from the datastore. We performed API calls from both react as well as node.js based on functionality being implemented. [4][6]

B. PostgreSQL & Apache Solr

CKAN maintains information in PostgreSQL databases since PostgreSQL is open source and widely available. Ckan has an Apache Solr component that enables search implementation with results in JSON format. Once we get the JSON content from CKAN we filtered the data and updated the content based on the filters. [5]

C. Pug

Pug is a templating engine designed for node or browser execution. It is fast, robust and feature-rich. RosaeNLG templates are basically pug templates where you use RosaeNLG structures and mixins to complete the standard pug syntax.[2][8]



```
include synonyms

mixin percent
    if(NLG.Scale[0] === 'Percent')
        |%
    if(NLG.Scale[0]=== 'Dollar')
        |$

p
    //- | the overall #[+Average] for #[+each] #[+race] is #[+around] #{parseFloat(avg).toFixed(2)}.
    if(min !== -1)
    | overall, the
    eachz el in maxArray with {separator:',', last_separator:'and'}
        | #{el}
    |
    | has the #[+highest] share with #[+anAverage] value of #{parseFloat(max).toFixed(2)} #[+percent]
    | the
    eachz el in minArray with {separator:',', last_separator:'and'}
        | #{el}
    | has the #[+lowest] #[+Average] of #{parseFloat(min).toFixed(2)} #[+percent].
```

Fig3.Pug template

D. RosaeNLG

RosaeNLG is a Natural Language Generation library for node.js or client side (browser) execution, based on the Pug template engine[8]. The filtered data obtained from CKAN is sent to the Node backend server which in turn sends the data as an input to the RosaeNLG library in order to perform NLG conversion.[7]

E. React.js & Node.js

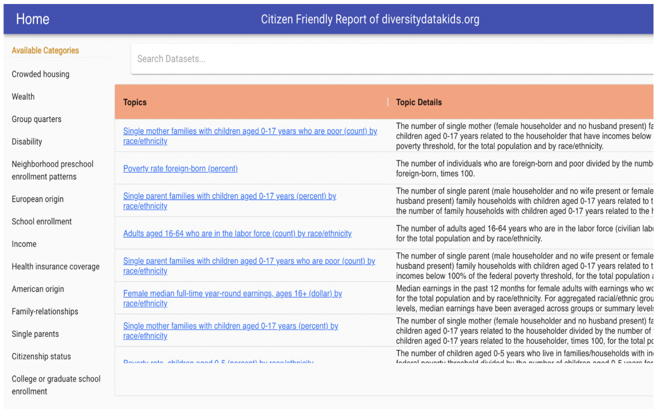
In our application we used react.js for frontend development and Node.js for backend development. Node.js is used to connect to RosaeNLG whereas both react.js & node.js are used to connect to CKAN API to fetch required data.

V. IMPLEMENTATION

The Citizen Friendly Report has been presented on a dashboard that lets the user select datasets from a wide range of topics, search for datasets of their choice using relevant search keywords or choose from a list of available categories and further select datasets according to the region and tenure for which the data was recorded. The selected dataset can then be viewed in the dashboard and the textual report is generated summarising the range of the entire dataset, range of the region wise filtered data and the selected rows. Each selected row has also been drawn as a bar chart for visualization and comparison.

A. Find relevant SubTopic using Keyword Search

Considering the huge count of datasets on the website and the vast research topics they cover, a search keyword filtering mechanism has been implemented that allows the user to enter keywords and search datasets relevant to their choice. The user can also select relevant search keys from a list of 'Available Categories' that are found to be the most common in the dataset descriptions. The search results appear as a list of clickable links to the dataset's region-wise categorization. A description of each dataset has been provided alongside the topic, fetched from the topic descriptions on diversitydatakids.org .

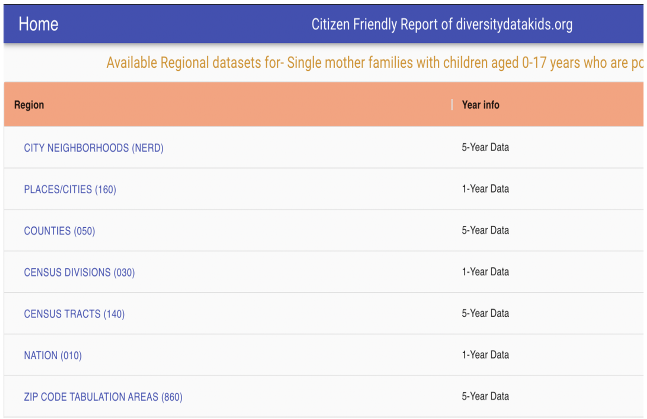


Topics	Topic Details
Single mother families with children aged 0-17 years who are poor (count) by race/ethnicity	The number of single mother (female householder and no husband present) fa children aged 0-17 years related to the householder that have incomes below poverty threshold, for the total population and by race/ethnicity.
Poverty rate foreign-born (percent)	The number of individuals who are foreign-born and poor divided by the numb foreign-born, times 100.
Single parent families with children aged 0-17 years (percent) by race/ethnicity	The number of single parent (male householder and no wife present or female husband present) family households with children aged 0-17 years related to t the number of family households with children aged 0-17 years related to the h
Adults aged 16-64 who are in the labor force (count) by race/ethnicity	The number of adults aged 16-64 years who are in the labor force (civilian lab for the total population and by race/ethnicity.
Single parent families with children aged 0-17 years who are poor (count) by race/ethnicity	The number of single parent (male householder and no wife present or female husband present) family households with children aged 0-17 years related to t incomes below 100% of the federal poverty threshold, for the total population a
Female median full-time year-round earnings, ages 16+ (dollar) by race/ethnicity	Median earnings in the past 12 months for female adults with earnings who wo for the total population and by race/ethnicity. For aggregated racial/ethnic grou levels, median earnings have been averaged across groups or summary levels
Single mother families with children aged 0-17 years (percent) by race/ethnicity	The number of single mother (female householder and no husband present) fa children aged 0-17 years related to the householder divided by the number of children aged 0-17 years related to the householder, times 100, for the total po
Poverty rate, children aged 0-5 (percent) by race/ethnicity	The number of children aged 0-5 years who live in families/households with in federal poverty threshold divided by the number of children aged 0-5 years for

Fig. 4.Find relevant datasets using keyword search

B. Region wise filtering of data

Each Subtopic listed as a search result in the Search Page can be further segregated according to region and tenure represented by the dataset. This segregated list of datasets can further enhance the search for relevant datasets. The user may choose a dataset belonging to a particular City, State, Nation, Zip Code, Census Division, Unified School Districts, Metro Areas, Counties etc, and further pick a 1 year dataset or a 5 year dataset in most cases.



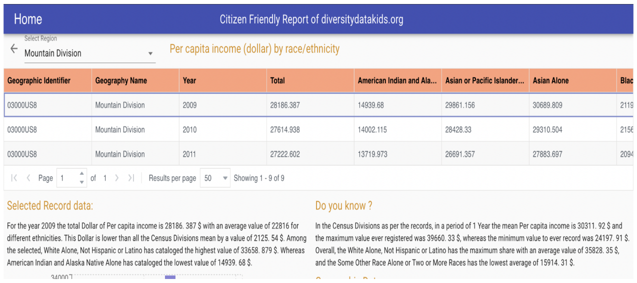
Available Regional datasets for- Single mother families with children aged 0-17 years who are pc

Region	Year info
CITY NEIGHBORHOODS (NERD)	5-Year Data
PLACES/CITIES (160)	1-Year Data
COUNTIES (050)	5-Year Data
CENSUS DIVISIONS (030)	1-Year Data
CENSUS TRACTS (140)	5-Year Data
NATION (010)	1-Year Data
ZIP CODE TABULATION AREAS (860)	5-Year Data

Fig. 5. Region wise filtering of data

C. Additional Filters

Once a user accesses a particular dataset , by default a region is selected and the data corresponding to that region alone is displayed to the user. Users can search the region manually or select from the available dropdown which gets lazily updated on input. Without this filter there can be thousands of records per dataset which makes it difficult for users to process the information.



Mountain Division

Per capita income (dollar) by race/ethnicity

Geographic Identifier	Geography Name	Year	Total	American Indian and Ala...	Asian or Pacific Islander...	Asian Alone	Blac
03000US8	Mountain Division	2009	28186.387	14939.68	29861.156	30689.809	2119
03000US8	Mountain Division	2010	27614.938	14002.115	28428.33	29310.504	2156
03000US8	Mountain Division	2011	27222.602	13719.973	26691.357	27883.697	2094

Page 1 of 1 Results per page 50 Showing 1 - 9 of 9

Selected Record data:

For the year 2009 the total Dollar of Per capita income is 28186. 387 $ with an average value of 22816 for different ethnicities. This Dollar is lower than all the Census Divisions mean by a value of 2125. 54 $. Among the selected, White Alone, Not Hispanic or Latino has cataloged the highest value of 33658. 879 $. Whereas American Indian and Alaska Native Alone has cataloged the lowest value of 14939. 68 $.

Do you know ?

In the Census Divisions as per the records, in a period of 1 Year the mean Per capita income is 30311. 92 $ and the maximum value ever registered was 39660. 33 $, whereas the minimum value to ever record was 24197. 91 $. Overall, the White Alone, Not Hispanic or Latino has the maximum share with an average value of 35828. 35 $, and the Some Other Race Alone or Two or More Races has the lowest average of 15914. 31 $.

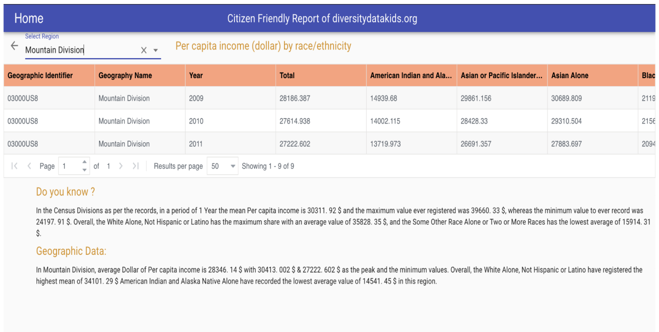
Fig.7. Default region selected at top left

C. Citizen Friendly Report

The Citizen friendly report generated by the rosaenlg library has been presented under three subsections that can provide a holistic view of the dataset as well as represent each row of the data in textual format.

Firstly, running an analysis on the selected dataset, a textual representation of the range of data is represented which includes observations such as the mean, topmost and the least values cataloged in the dataset for the selected region and tenure.

Following this, a textual report has been generated that presents an analysis of the data belonging to the selected city/state. A dropdown list of each distinct region recorded in the dataset has been provided so as to enable the user to generalise region specific data. This report shows mean and the highest and lowest values recorded in the dataset for the selected region and the average distribution of factors for ethnicities, races, income groups etc.



Mountain Division

Per capita income (dollar) by race/ethnicity

Geographic Identifier	Geography Name	Year	Total	American Indian and Ala...	Asian or Pacific Islander...	Asian Alone	Blac
03000US8	Mountain Division	2009	28186.387	14939.68	29861.156	30689.809	2119
03000US8	Mountain Division	2010	27614.938	14002.115	28428.33	29310.504	2156
03000US8	Mountain Division	2011	27222.602	13719.973	26691.357	27883.697	2094

Page 1 of 1 Results per page 50 Showing 1 - 9 of 9

Do you know ?

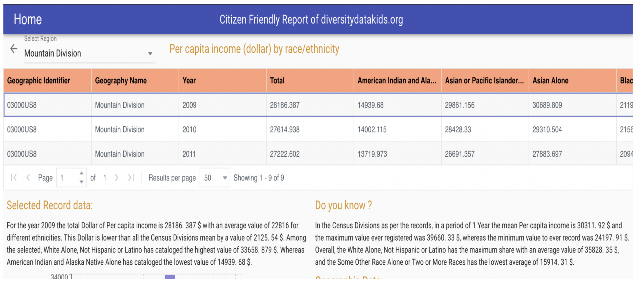
In the Census Divisions as per the records, in a period of 1 Year the mean Per capita income is 30311. 92 $ and the maximum value ever registered was 39660. 33 $, whereas the minimum value to ever record was 24197. 91 $. Overall, the White Alone, Not Hispanic or Latino has the maximum share with an average value of 35828. 35 $, and the Some Other Race Alone or Two or More Races has the lowest average of 15914. 31 $.

Geographic Data:

In Mountain Division, average Dollar of Per capita income is 28346. 14 $ with 30413. 002 $ & 27222. 602 $ as the peak and the minimum values. Overall, the White Alone, Not Hispanic or Latino have registered the highest mean of 34101. 29 $ American Indian and Alaska Native Alone have recorded the lowest average value of 14541. 45 $ in this region.

Fig. 6. Range of cataloged data in textual format

Finally, upon selecting each row in the dataset, a textual representation of the selected row is presented which gives the stats and observations recorded in the row in a structured format that is more intuitive for users.



Mountain Division

Per capita income (dollar) by race/ethnicity

Geographic Identifier	Geography Name	Year	Total	American Indian and Ala...	Asian or Pacific Islander...	Asian Alone	Blac
03000US8	Mountain Division	2009	28186.387	14939.68	29861.156	30689.809	2119
03000US8	Mountain Division	2010	27614.938	14002.115	28428.33	29310.504	2156
03000US8	Mountain Division	2011	27222.602	13719.973	26691.357	27883.697	2094

Page 1 of 1 Results per page 50 Showing 1 - 9 of 9

Selected Record data:

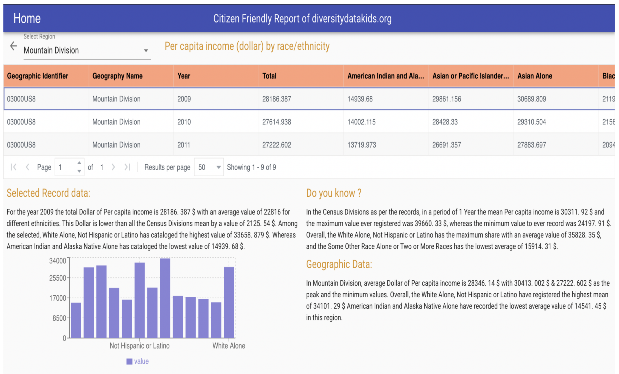
For the year 2009 the total Dollar of Per capita income is 28186. 387 $ with an average value of 22816 for different ethnicities. This Dollar is lower than all the Census Divisions mean by a value of 2125. 54 $. Among the selected, White Alone, Not Hispanic or Latino has cataloged the highest value of 33658. 879 $. Whereas American Indian and Alaska Native Alone has cataloged the lowest value of 14939. 68 $.

Do you know ?

In the Census Divisions as per the records, in a period of 1 Year the mean Per capita income is 30311. 92 $ and the maximum value ever registered was 39660. 33 $, whereas the minimum value to ever record was 24197. 91 $. Overall, the White Alone, Not Hispanic or Latino has the maximum share with an average value of 35828. 35 $, and the Some Other Race Alone or Two or More Races has the lowest average of 15914. 31 $.

Fi. 8. Selected row data in textual format

Since each row in the dataset contains numerical stats for various races, ethnicities or such factors specific to the dataset topic, a graphical representation has been included that aids users in visual comparison. In most cases a bar chart has been used to represent the data values for each ethnicity/racial group.



Mountain Division

Per capita income (dollar) by race/ethnicity

Geographic Identifier	Geography Name	Year	Total	American Indian and Ala...	Asian or Pacific Islander...	Asian Alone	Blac
03000US8	Mountain Division	2009	28186.387	14939.68	29861.156	30689.809	2119
03000US8	Mountain Division	2010	27614.938	14002.115	28428.33	29310.504	2156
03000US8	Mountain Division	2011	27222.602	13719.973	26691.357	27883.697	2094

Page 1 of 1 Results per page 50 Showing 1 - 9 of 9

Selected Record data:

For the year 2009 the total Dollar of Per capita income is 28186. 387 $ with an average value of 22816 for different ethnicities. This Dollar is lower than all the Census Divisions mean by a value of 2125. 54 $. Among the selected, White Alone, Not Hispanic or Latino has cataloged the highest value of 33658. 879 $. Whereas American Indian and Alaska Native Alone has cataloged the lowest value of 14939. 68 $.

Do you know ?

In the Census Divisions as per the records, in a period of 1 Year the mean Per capita income is 30311. 92 $ and the maximum value ever registered was 39660. 33 $, whereas the minimum value to ever record was 24197. 91 $. Overall, the White Alone, Not Hispanic or Latino has the maximum share with an average value of 35828. 35 $, and the Some Other Race Alone or Two or More Races has the lowest average of 15914. 31 $.

Geographic Data:

In Mountain Division, average Dollar of Per capita income is 28346. 14 $ with 30413. 002 $ & 27222. 602 $ as the peak and the minimum values. Overall, the White Alone, Not Hispanic or Latino have registered the highest mean of 34101. 29 $ American Indian and Alaska Native Alone have recorded the lowest average value of 14541. 45 $ in this region.

Fig. 9. Graphical representation of selected row

X. Conclusion

We were able to develop a citizen friendly version of diversitydatakids.org website with the help of nlg. This will help US citizens to get a proper knowledge about the data and help in making the rules beneficial for the children and helps in research of the needs for kids to get proper development.

XI. Future Improvements

While we are proud that we were able to develop a site which is citizen friendly and can create a greater knowledge for the society in one semester, we know the idea still has room to grow.

A. Adding more graphical Data to the website for better visualization of the data

B. Add more analytics in NLG like what's the trend in past few years or trend of a particular ethnic group.

C. Data visualization for the overview of the whole dataset.

XII. Deliverables

A. GitHub Repository

https://github.com/SJSUSpring21/Citizen-Friendly-Report-of-diversitydatakids.org

B. Presentation

https://github.com/SJSUSpring21/Citizen-Friendly-Report-of-diversitydatakids.org/blob/main/Citizen%20Friendly%20Report%20of%20DiversityDataKids.org.pdf

C. Link of Running Application (Deployed in AWS)

http://3.16.112.6:3000

References

[1] https://data.diversitydatakids.org/dataset

[2] https://rosaenlg.org/rosaenlg/3.0.0/advanced/pug.html

[3] https://en.wikipedia.org/wiki/Natural-language_generation

[4] https://docs.ckan.org/en/latest/api/index.html

[5] https://solr.apache.org/guide/6_6/common-query-parameters.html

[6] https://docs.ckan.org/en/2.8/maintaining/datastore.html

[7] https://rosaenlg.org/rosaenlg/3.0.0/tutorials/tutorial_en_US.html

[8] https://pugjs.org/api/getting-started.html