The Great Divide:

An Analysis of the Disparities between Public and Private Health Care Coverage in the United States

PREPARED BY

Lateef Babatunde
Brandi Bugg
Maia Clemons
Jessica Lawrence
Ebitimere Ogobri
Teara Peeples

Introduction	**3**
Overview of the Problem	3
Specific Problem to be Solved	4
Importance in American Society	4
Data Analysis and Computation	**5**
Datasets	5
Data Cleaning and Wrangling	5
Exploratory Data Analysis	7
Statistical Analysis	12
Dashboard	**15**
Use Case	15
Data Engineering	18
Conclusion	**19**
Future Work	**20**
References	**21**

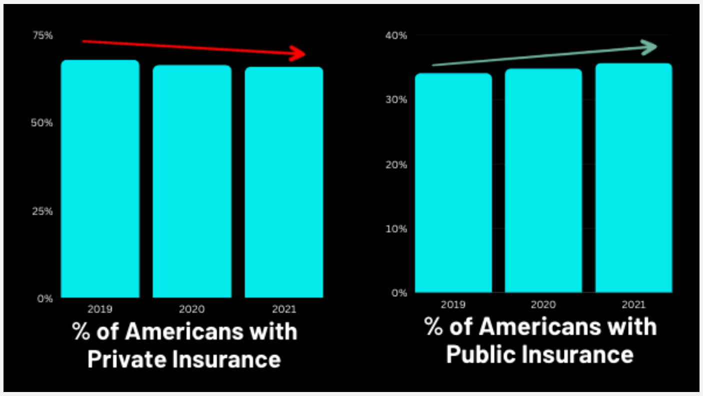
Introduction

Overview of the Problem

Health insurance coverage is a trending topic in the United States of America. It can directly impact an individual's access to care and is therefore nearly essential. The American health insurance system is also highly influenced by the fact that it is pluralistic as it is divided by a private sector and a public sector.

In analyzing the "great divide" between the two sectors, private insurance and public insurance, the following factors are important to note:

- ➔ Inadequate healthcare coverage is one of the largest barriers to healthcare access.
- ➔ In the United States unlike other higher income countries, we currently do not have universal healthcare programs. The type of health insurance you have, among other factors, has a major influence on recipients' health outcomes regardless of initial health status.
- ➔ The average life expectancy in the United States is on the decline again and it is headed for the sharpest drop in 100 years. Studies suggest that lack of health insurance negatively impacts health and overall lifespan.
- ➔ Health coverage plays a major role in enabling people to access health care and protecting families from high medical costs.
- ➔ People of color have faced longstanding disparities in health coverage that contribute to disparities in health.
- ➔ Individuals on public insurance have less access to healthcare resources versus those with private insurance; due to public insurance having restrictions on services and certain facilities only accepting private insurance.



Graphical depiction of recent trends in the U.S. Health Insurance Coverage.

Specific Problem to be Solved

The specific problems that we aim to solve in our project are centered around the disparities that having private insurance over public insurance, or vice versa, can have in your overall health status in America.
We will further analyze the data on individuals on public versus private insurance using the following factors that include but are not limited to:

- Socioeconomic Factors
    - Race and/or Ethnicity
    - Income
    - Education Level
- Access to Care
    - Most Recent Doctor Visit
    - Wellness Visit
    - Time Since Last Visit
- Prevalence of Chronic Health Diseases or Conditions
    - Chronic Obstructive Pulmonary Disease (COPD)
    - Coronary Heart Disease
    - Cancer
    - Diabetes
    - Hypertension

Our analysis will seek to highlight the disparities in the coverage of public and private insurance, especially in relation to socioeconomic factors, health outcomes, and access to care.

Importance in American Society

Over the years, the number of insured Americans has increased significantly with now 92% of Americans being insured. As we continue to progress towards the measure of all citizens having coverage, we must examine the divide in coverage for those publicly and privately insured. In recent years, the number of Americans with private insurance has continued to decrease while the number of Americans with public insurance have steadily increased. This trend highlights the importance of addressing any disparities that exist between those privately and publicly insured. Health insurance coverage is a factor that can affect every American.

Affordability of health insurance affects an individual's ability to reduce risks of preventable conditions. The motivation for our group would be to use data science to identify the gap between private and public insurance to propose more affordable and effective insurance.
It appears as though the type of health insurance you have, among other factors, has a major influence on recipients’ health outcomes (whether or not they’re in good health).

“Health coverage plays a major role in enabling people to access health care and protecting families from high medical costs. People of color have faced longstanding disparities in health coverage that contribute to disparities in health.”
Individuals on public insurance have less access to healthcare resources versus those with private insurance; due to public insurance having restrictions on services and certain facilities only accepting private insurance. Individuals with private insurance have less restrictions, but higher rates; the insurance is typically available through their employer, and one could lose the insurance in the event of job change or termination.

Data Analysis and Computation

Datasets

We used the National Health Interview Survey (NHIS) dataset from the Centers for Disease Control and Prevention (CDC) for our analysis. The NHIS is a self-reported survey with the purpose of monitoring the health of U.S. citizens. This survey is only available to the non-institutionalized population. We used the 2019, 2020, and 2021 version of the datasets. The datasets included information about insurance type, general health status, income type, and disease just to name a few. We were able to sift through all of the columns and pick out the columns that would allow us to compare health insurance types to different socioeconomic factors and diseases. These datasets are basically identical and were cleaned and analyzed in the exact same manner. Any images included are from the 2021 dataset as it was the most recent data available to us.

Data Cleaning and Wrangling

The CSV file with over 600 columns was loaded into Google Colab. We needed to figure out which columns we wanted to analyze first. We decided to focus on the columns that described insurance types, socioeconomic factors like education level and race, diseases, overall health status, overall satisfaction, prescription drugs, and any other variables that might tell us about their access to care like if they have a primary doctor they visit regularly.

```
[ ] df=pd.DataFrame(adult21,columns=['AGEP_A','AGE65','SEX_A','EDUCP_A','HISP_A','PAYBLL12M_A','RACEALLP_A','LASTDI
```

This figure depicts the creation of a new, smaller dataframe with some of our target variables.

```
[ ] adult21.columns

    Index(['URBRRL', 'RATCAT_A', 'IMPINCFLG_A', 'CVDVAC2YR_A', 'CVDVAC2MR_A',
           'CVDVAC1YR_A', 'CVDVAC1MR_A', 'SHTCVD19AV_A', 'SHTCVD19NM_A',
           'SHTCVD19_A',
           ...
           'PROXYREL_A', 'PROXY_A', 'AVAIL_A', 'HHSTAT_A', 'INTV_MON', 'RECTYPE',
           'IMPNUM_A', 'WTFA_A', 'HHX', 'POVRATTC_A'],
          dtype='object', length=622)
```

622 columns. Oh my. We went through the columns and decided to pick out the ones that made the most sense for what we are trying to analyze.

This figure depicts us examining the columns in the dataset. It shows the amount of columns and shows some of the original column names.

The dataset included column names that were not intuitive or easy to recognize so we decided to modify them to make it easier to understand the data without having to constantly reference the NHIS codebook. Numeric codes were used to represent values inside the columns so we opted to modify those as well for our convenience. These steps greatly improved the readability of our database and were essential for our analysis.

```
[ ] #renaming every column
    df = df.rename(columns={'AGEP_A': 'AGE', 'AGE65': 'AGE (DOB not verified)','SEX_A':'SEX','EDUCP_A':'HIGHEST EDU
```

This figure depicts the renaming of the columns included in the new dataframe.

```
[ ] #renaming values in columns. Some columsn have been grouped because they have the same variables.
    insurancecols = ['HAS PRIVATE INSURANCE', 'MEDICARE','MEDICARE SUPPLEMENT(MEDIGAP)', 'MEDICAID','CHIP', 'MILITARY(TRIC
    newdf[insurancecols]=newdf[insurancecols].replace({1:'mentioned',2:'not mentioned',7:'refused',8:'not accertained',9:'
    yesnocols =['DELAYED MEDICAL CARE DUE TO COST','HOSPITALIZED OVERNIGHT LAST 12 MNTHS','HISPANIC','TIME SINCE LAST VISI
    newdf[yesnocols] = newdf[yesnocols].replace({1:'yes',2:'no',7:'refused',8:'not accertained',9:'dont know'})
    newdf[yesnocols] = newdf[yesnocols].replace({1:'yes',2:'no',7:'refused',8:'not accertained',9:'dont know'})
    newdf=newdf.replace({'SEX' : { 1 : 'male', 2 : 'female', 7 : 'refused',8:'not accertained',9 :'dont know'}})
    newdf=newdf.replace({'HIGHEST EDUCATION' : { 1 : 'Grade 1-11 ', 2 : '12th grade, no diploma', 3 : 'GED or equivalent',

[ ] newdf=newdf.replace({'PAYBILL' : { 1 : 'yes', 2 : 'no', 7 : 'refused',8:'not accertained',9 :'dont know'}})
    newdf=newdf.replace({'SEX' : { 1 : 'male', 2 : 'female', 7 : 'refused',8:'not accertained',9 :'dont know'}})

    #renaming columns continued
    newdf.replace({'SEX' : { 1 : 'male', 2 : 'female', 7 : 'refused',8:'not accertained',9 :'dont know'}})
    newdf.replace({'GENERAL HEALTH STATUS' : { 1 : 'excellent', 2 : 'very good', 3:'good',4:'fair',5:'poor', 7 : 'refused'
    timescols =['FREQ URGENT CARE VISITS LAST 12 MNTHS','FREQ ER VISITS LAST 12 MNTHS']
    newdf[timescols] = newdf[timescols].replace({0:'0 times',1:'1 time',2:'2 times',3:'3 times',4:'4 times',5:'5+ times',7

    newdf=newdf.replace({'RACE' : { 1 : 'white', 2 : 'Black/AFAM', 3:'Asian',4:'AI/AN',5:'AIAN and any other group', 6:'Ot

[ ] newdf=newdf.replace({'COULDNT AFFORD BALANCED MEALS' : { 1 : 'often true', 2 : 'sometimes true',3:'never true', 7 : 'r
    newdf=newdf.replace({'GENERAL HEALTH STATUS' : { 1 : 'excellent', 2 : 'very good', 3:'good',4:'fair',5:'poor', 7 : 're
    timescols =['FREQ URGENT CARE VISITS LAST 12 MNTHS','FREQ ER VISITS LAST 12 MNTHS']
    newdf[timescols] = newdf[timescols].replace({0:'0 times',1:'1 time',2:'2 times',3:'3 times',4:'4 times',5:'5+ times',7

    newdf=newdf.replace({'AGE (DOB not verified)' : { 1 : 'less than 65',2:'65 and older',7:'refused',8:'not accertained',
    newdf=newdf.replace({'TYPE OF PLACE FOR USUAL CARE' : { 1 : 'doctors office or health center', 2 : 'Urgent care center
```

This figure depicts the assigning of variable names to the numerical code variables given by the NHIS.

It was determined that there were almost 400,000 null values in our dataset. Our dataset is unique in the way that the data is scattered throughout different columns. For example, there is a different column for each insurance type (Medicaid, Medicare, private, etc) and in each column, there is a “yes” meaning that they have that insurance type and the rest would be null values. So even with this large number of null values in the dataset, a decision was made to rename the null values as “Not Available” rather than to completely drop them so that the integrity of the data could be maintained. This left us with 64 columns and 31,568 rows.

```
[ ] # Count total NaN in a DataFrame
    print(" \nCount total NaN in a DataFrame : \n\n",
           df.isnull().sum().sum())

    Count total NaN in a DataFrame :

     382604
```

```
#creating a new variable for NaN values.
df = df.fillna('Not Answered')
```

The figures above depict the total number of nulls in the dataset and changing the variable to “Not Answered”.

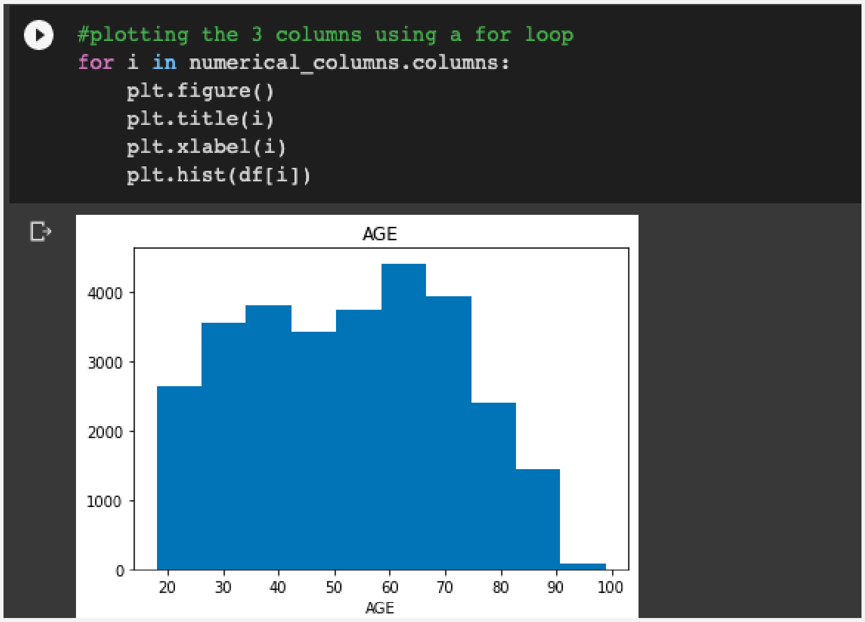
Exploratory Data Analysis

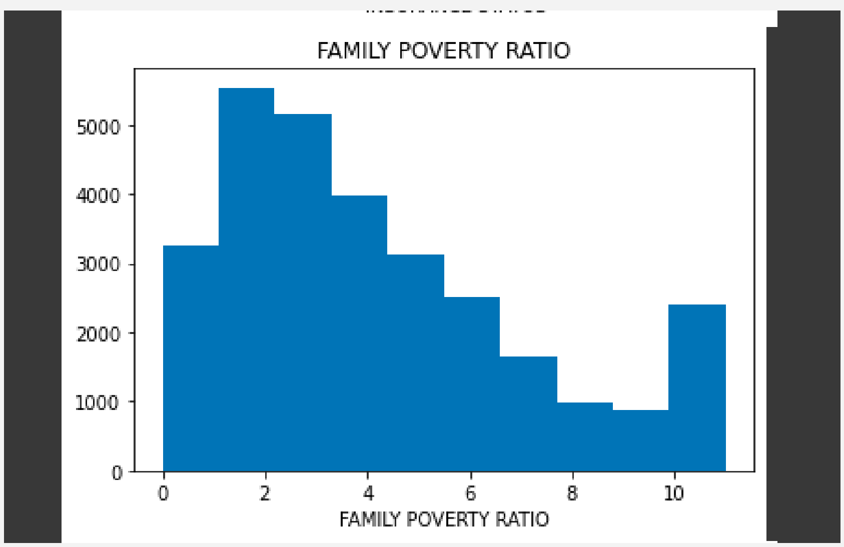
The goal of our exploratory data analysis (EDA) of the NHIS data was to identify any potential patterns or relationships in the data. First, it was important for us to examine the number of categorical and numerical columns we were working with. There were only three numerical columns out of 64. This is common in epidemiological survey data. This information was important for us in order to determine the models we would use to analyze the data since we had to compare a large amount of categorical data.

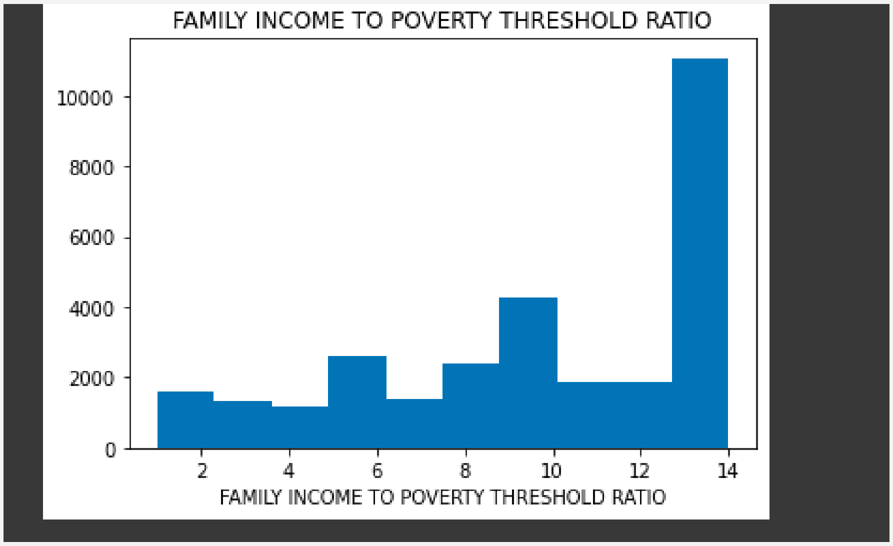
```
[ ] #examining how many columns are numerical
    newdf.info()

    <class 'pandas.core.frame.DataFrame'>
    RangeIndex: 29482 entries, 0 to 29481
    Data columns (total 64 columns):
     #   Column                              Non-Null Count  Dtype
    ---  ------                              --------------  -----
     0   AGE                                 29482 non-null  int64
     1   AGE (DOB not verified)              29482 non-null  object
     2   SEX                                 29482 non-null  object
     3   HIGHEST EDUCATION                   29482 non-null  object
     4   HISPANIC                            29482 non-null  object
     5   PAYBILL                             29482 non-null  object
     6   RACE                                29482 non-null  object
     7   MOST RECENT DOCTOR VISIT            29482 non-null  object
     8   WELLNESS VISIT                      29482 non-null  object
     9   TIME SINCE LAST VISIT               29482 non-null  object
     10  REGULAR DOCTOR ACCESS               29482 non-null  object
     11  TYPE OF PLACE FOR USUAL CARE        29482 non-null  object
     12  FREQ URGENT CARE VISITS LAST 12 MNTHS  29482 non-null  object
     13  FREQ ER VISITS LAST 12 MNTHS        29482 non-null  object
     14  HOSPITALIZED OVERNIGHT LAST 12 MNTHS  29482 non-null  object
     15  DELAYED MEDICAL CARE DUE TO COST    29482 non-null  object
     16  GENERAL HEALTH STATUS               29482 non-null  object
```

This figure depicts the examination of the columns in the new dataframe. It includes the Non-Null count and data type along with the column name and index.

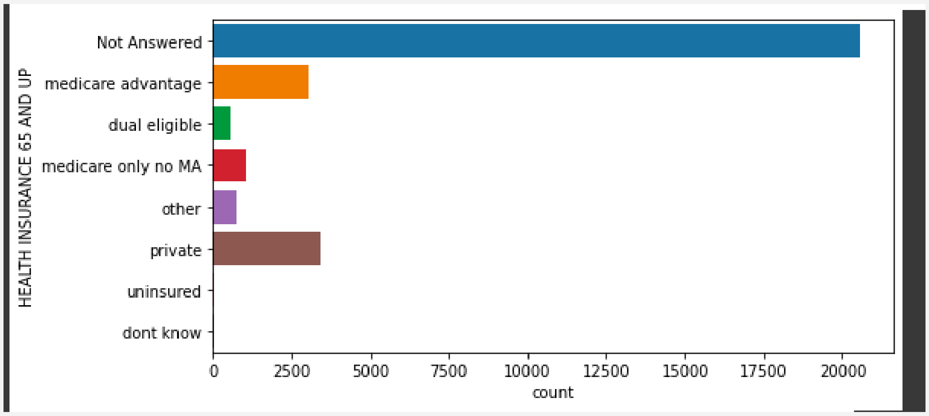
The figures below represent the three numerical columns that were plotted as histograms to examine the distribution.



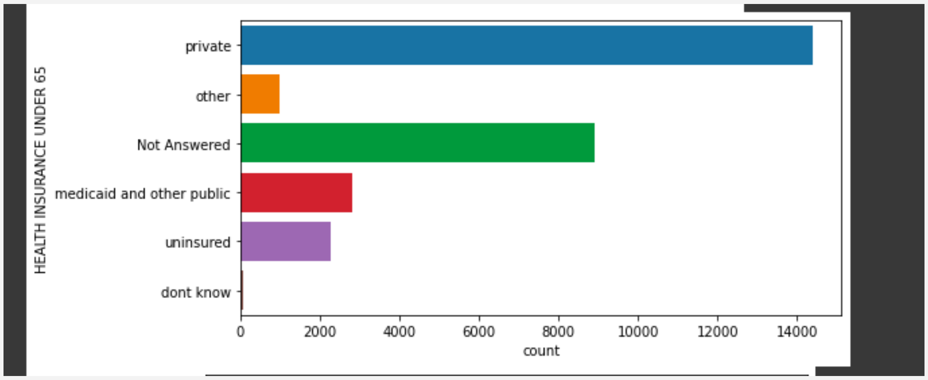




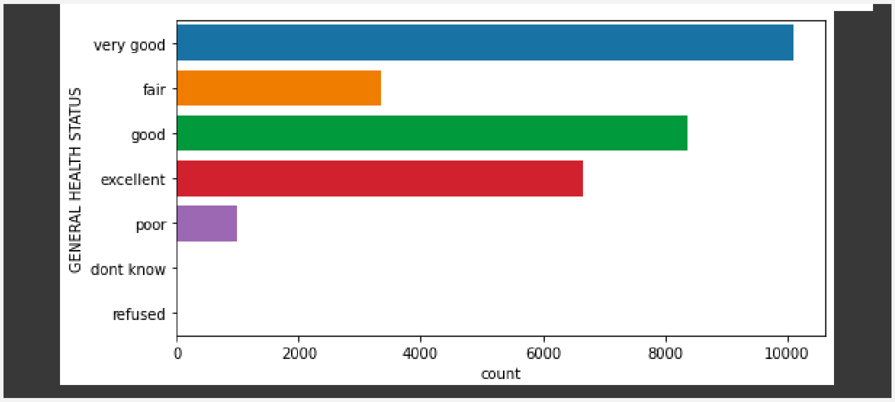
The figures below represent the categorical variables that were plotted as bar plots to examine the distribution and magnitude of each value.



Health Insurance 65 and up by Coverage Type



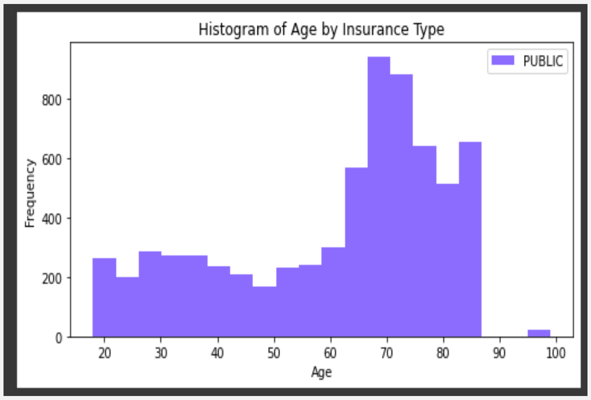
Health Insurance Under 65 by Coverage Type

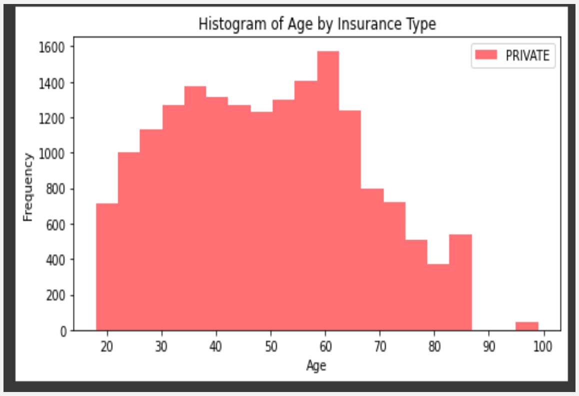


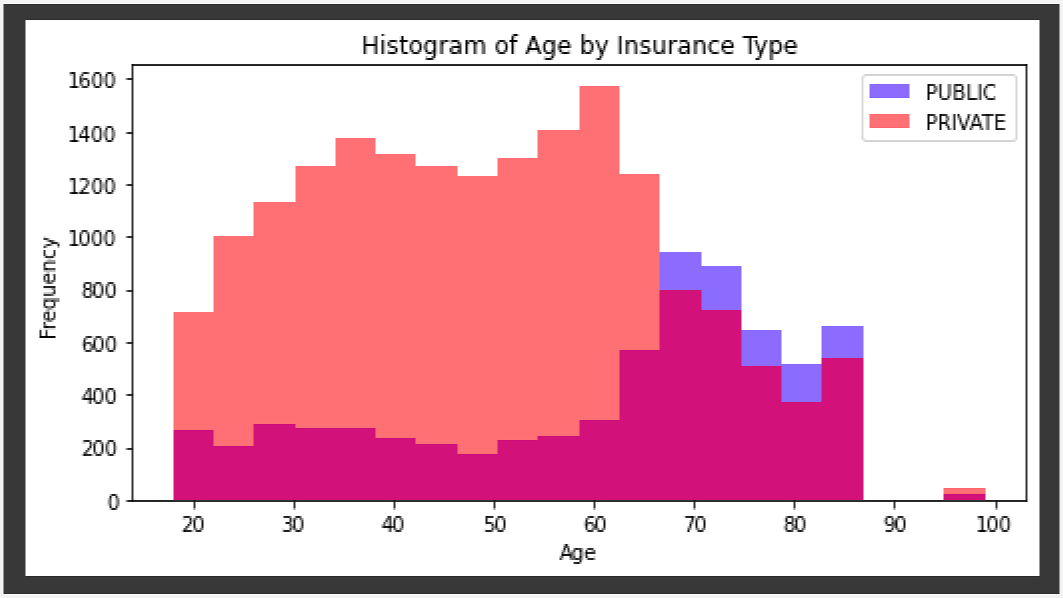
In order to be able to compare the insurance types to all of the other variables when doing our statistical analysis, they needed to be condensed into one column. We did this using a for loop so that we could have the values “PRIVATE”, “PUBLIC”, “OTHER”, “NO COVERAGE” AND “NOT AVAILABLE”.

```
coverage_all = []

for i in range(len(newdf)):
    if newdf.loc[i, 'HEALTH INSURANCE UNDER 65'] == 'private':
        coverage_all.append('PRIVATE')
    elif newdf.loc[i, 'HEALTH INSURANCE UNDER 65'] == 'medicaid and other public':
        coverage_all.append('PUBLIC')
    elif newdf.loc[i, 'HEALTH INSURANCE UNDER 65'] == 'other':
        coverage_all.append('OTHER')
    elif newdf.loc[i, 'HEALTH INSURANCE UNDER 65'] == 'uninsured' :
        coverage_all.append('NO COVERAGE')
    elif newdf.loc[i, 'HEALTH INSURANCE UNDER 65'] == 'Not Answered':
        coverage_all.append('NOT AVAILABLE')
    elif newdf.loc[i, 'HEALTH INSURANCE UNDER 65'] == 'dont know':
        coverage_all.append('NOT AVAILABLE')
    elif newdf.loc[i, 'HEALTH INSURANCE 65 AND UP'] == 'private':
        coverage_all.append('PRIVATE')
    elif newdf.loc[i, 'HEALTH INSURANCE 65 AND UP'] == 'dual eligible':
        coverage_all.append('OTHER')
    elif newdf.loc[i, 'HEALTH INSURANCE 65 AND UP'] == 'medicare advantage':
        coverage_all.append('PUBLIC')
    elif newdf.loc[i, 'HEALTH INSURANCE 65 AND UP'] == 'medicare only no MA':
        coverage_all.append('PUBLIC')
    elif newdf.loc[i, 'HEALTH INSURANCE 65 AND UP'] == 'other' :
        coverage_all.append('OTHER')
    elif newdf.loc[i, 'HEALTH INSURANCE 65 AND UP'] == 'uninsured':
        coverage_all.append('NO COVERAGE')
    elif newdf.loc[i, 'HEALTH INSURANCE 65 AND UP'] == 'dont know ':
        coverage_all.append('NOT AVAILABLE')
    else:
        coverage_all.append('NOT AVAILABLE')

newdf['ALL COVERAGE'] = coverage_all
```

The public and private insurance values were made into their own columns in order to create histograms to examine the distribution in public and private insurance types with the numerical variables.







The biggest takeaways from the EDA were that there were some significant differences between the public and private population in terms of the numerical values in the dataset. We were able to see that the 65+ population relies on public insurance through the histograms plotted. We were also able to see that the family income to poverty ratio for people with private insurance was skewed left (not pictured).

Statistical Analysis

We conducted our analysis in Python. The most informative statistical methods used were linear regression and chi-square test. We used a simple linear regression to analyze the variable "ALL_COVERAGE" (which is the same column as "ALL COVERAGE" with no spaces in the name) against the numerical variables. When conducting these tests, "ALL_COVERAGE", the predictor value, was used as the independent variable because it was the main focus of the study. In every case, the p-values were very small (below the predictor) and the t-values were very large. This indicated to us that the predictor value had a significant effect on the dependent or response values which are our three numerical columns.

```
#Linear Regression - AGE
formula1='AGE ~ ALL_COVERAGE'
model = sm.ols(formula=formula1, data=spaceless)
fitted = model.fit()
print(fitted.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    AGE   R-squared:                       0.124
Model:                            OLS   Adj. R-squared:                  0.124
Method:                 Least Squares   F-statistic:                     1040.
Date:                Fri, 10 Feb 2023   Prob (F-statistic):               0.00
Time:                        05:34:03   Log-Likelihood:            -1.2588e+05
No. Observations:               29482   AIC:                         2.518e+05
Df Residuals:                   29477   BIC:                         2.518e+05
Df Model:                           4
Covariance Type:            nonrobust
=================================================================================================
                                    coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
Intercept                        41.3022      0.360    114.788      0.000      40.597      42.007
ALL_COVERAGE[T.NOT AVAILABLE]     7.7939      1.735      4.493      0.000       4.394      11.194
ALL_COVERAGE[T.OTHER]            22.1853      0.508     43.660      0.000      21.189      23.181
ALL_COVERAGE[T.PRIVATE]           8.1246      0.382     21.243      0.000       7.375       8.874
ALL_COVERAGE[T.PUBLIC]           19.7428      0.416     47.508      0.000      18.928      20.557
==============================================================================
Omnibus:                     1099.215   Durbin-Watson:                   1.820
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              573.901
Skew:                          -0.160   Prob(JB):                    2.39e-125
Kurtosis:                       2.396   Cond. No.                         21.0
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

This figure depicts the linear regression model of Age and "ALL_COVERGAGE".

Performing the chi-square tests provided us with an abundance of new information. The contingency tables allowed us to see the data in a completely new way. We were able to see the relationships between our many categorical variables and our main predictor variable, "ALL COVERAGE". The contingency tables gave us our observed frequency. They allowed us to see how many individuals had cancer for example that also had public health insurance. We took these contingency tables and conducted chi-square tests. The tables were modified to exclude the "NOT AVAILABLE" values as they were unnecessary and didn't provide us any important information. The contingency tables (observed values) were divided by expected frequency from the chi-square test in order to provide us with the standardized residual which measures how much the observed frequency deviated from the expected frequency if the values were independent. We found that in several of the tested variables the residual value was greater than 1 and this indicated that there was a strong association between the variables and that the particular combination was more likely to occur. When the residual value was below one, that was an indication of a weak correlation and that combination was less likely to occur than expected.

	General Health Status							
Coverage Type	Excellent	Very Good	Good	Fair	Poor	Refused	Don't Know	Grand Total
NO COVERAGE	26.46%	31.30%	28.84%	10.59%	2.77%	0.04%		100.00%
NOT AVAILABLE	25.00%	26.92%	31.73%	12.50%	2.88%	0.96%		100.00%
OTHER	12.25%	23.30%	32.20%	23.39%	8.77%	0.04%	0.04%	100.00%
PRIVATE	25.80%	38.43%	26.34%	7.81%	1.59%	0.01%	0.02%	100.00%
PUBLIC	16.45%	28.37%	31.89%	16.75%	6.50%	0.01%	0.03%	100.00%
Grand Total	22.58%	34.28%	28.32%	11.37%	3.41%	0.02%	0.02%	100.00%

The three figures below depict the process of using the chi-square test and contingency table to find the standardized residual.

```
#contingency table
edu_crosstab = pd.crosstab(df['ALL COVERAGE'],
                           df['HIGHEST EDUCATION'],
                             margins = True)
print(edu_crosstab)
```

```
HIGHEST EDUCATION     1    2    3     4     5     6     7     8     9    10  \
ALL COVERAGE
NO COVERAGE         420   77   94   718   340    96   165   277    65    32
NOT AVAILABLE        11    3    1    40    17     2     5    15     6     1
OTHER               312   38   67   590   401    98   228   336   185    52
PRIVATE             509  185  268  3352  2521   641  1618  5220  2554   883
PUBLIC              817  161  215  1906  1174   289   550  1120   474   181
All                2069  464  645  6606  4453  1126  2566  6968  3284  1149

HIGHEST EDUCATION  97   99    All
ALL COVERAGE
NO COVERAGE         3   26   2313
NOT AVAILABLE       3    0    104
OTHER               5   14   2326
PRIVATE            29   33  17813
PUBLIC             11   28   6926
All                51  101  29482
```

```
[111] edu_crosstab.index
```

```
Index(['NO COVERAGE', 'NOT AVAILABLE', 'OTHER', 'PRIVATE', 'PUBLIC', 'All'], dtype='object', name='ALL COVERAGE')
```

```
[112] edu_chi=edu_crosstab.loc[['NO COVERAGE','OTHER', 'PRIVATE','PUBLIC']]
```

```
[113] print(edu_chi)

    HIGHEST EDUCATION    1    2    3     4     5    6     7     8     9   10  97  \
    ALL COVERAGE
    NO COVERAGE        420   77   94   718   340   96   165   277    65   32   3
    OTHER              312   38   67   590   401   98   228   336   185   52   5
    PRIVATE            509  185  268  3352  2521  641  1618  5220  2554  883  29
    PUBLIC             817  161  215  1906  1174  289   550  1120   474  181  11

    HIGHEST EDUCATION  99    All
    ALL COVERAGE
    NO COVERAGE        26   2313
    OTHER              14   2326
    PRIVATE            33  17813
    PUBLIC             28   6926
```

```
[114] chi_squared, p_value, dof, expected_freq = stats.chi2_contingency(edu_chi)
      print("Chi-squared test statistic: ", chi_squared)
      print("p-value: ", p_value)
      print("Degrees of freedom: ", dof)
      print("Expected frequencies: \n", expected_freq)

    Chi-squared test statistic:  3005.602590380369
    p-value:  0.0
    Degrees of freedom:  36
    Expected frequencies:
     [[1.62031248e+02 3.62956294e+01 5.07036558e+01 5.16956838e+02
       3.49256859e+02 8.84952005e+01 2.01633637e+02 5.47426271e+02
       2.58084757e+02 9.03847777e+01 3.77915447e+00 7.95197086e+00
       2.31300000e+03]
      [1.62941929e+02 3.64996256e+01 5.09886309e+01 5.19862346e+02
       3.51219824e+02 8.89925795e+01 2.02766900e+02 5.50503029e+02
```

```
edu_chi/expected_freq
```

HIGHEST EDUCATION / ALL COVERAGE	1	2	3	4	5	6	7	8	9	10	97	99	All
NO COVERAGE	2.592093	2.121468	1.853910	1.388897	0.973496	1.084805	0.818316	0.506004	0.251855	0.354042	0.793828	3.269630	1.0
OTHER	1.914793	1.041107	1.314018	1.134916	1.141735	1.101215	1.124444	0.610351	0.712812	0.572103	1.315653	1.750730	1.0
PRIVATE	0.407904	0.661844	0.686332	0.841954	0.937274	0.940539	1.041967	1.238179	1.284982	1.268539	0.996419	0.538862	1.0
PUBLIC	1.683900	1.481374	1.416095	1.231294	1.122577	1.090615	0.910946	0.683259	0.613351	0.668769	0.972056	1.175916	1.0

We also performed t-tests with our main predictor variable and our numerical variables. This was especially informative when examining the “FAMILY INCOME TO POVERTY THRESHOLD RATIO”. The t-statistic was negative with a strong magnitude. This indicated that the mean family income to poverty threshold ratio for families with public health insurance coverage was significantly lower and the difference was strong. This was confirmed by calculating the means separately.

```
t_stat, p_value = stats.ttest_ind(public_df['FAMILY INCOME TO POVERTY THRESHOLD RATIO'], private_df['FAMILY INC
print("t-statistic: ", t_stat)
print("p-value: ", p_value)
```

```
t-statistic:  -71.46525014315785
p-value:  0.0
```

+ Code + Text

```
[100] public_mean = public_df['FAMILY INCOME TO POVERTY THRESHOLD RATIO'].mean()
      private_mean = private_df['FAMILY INCOME TO POVERTY THRESHOLD RATIO'].mean()
      print(public_mean)
      print( private_mean)
```

```
7.6962171527577246
11.265536405995622
```

This figure depicts the t-test of "FAMILY INCOME TO POVERTY THRESHOLD RATIO" for public and private insurance types.

Dashboard

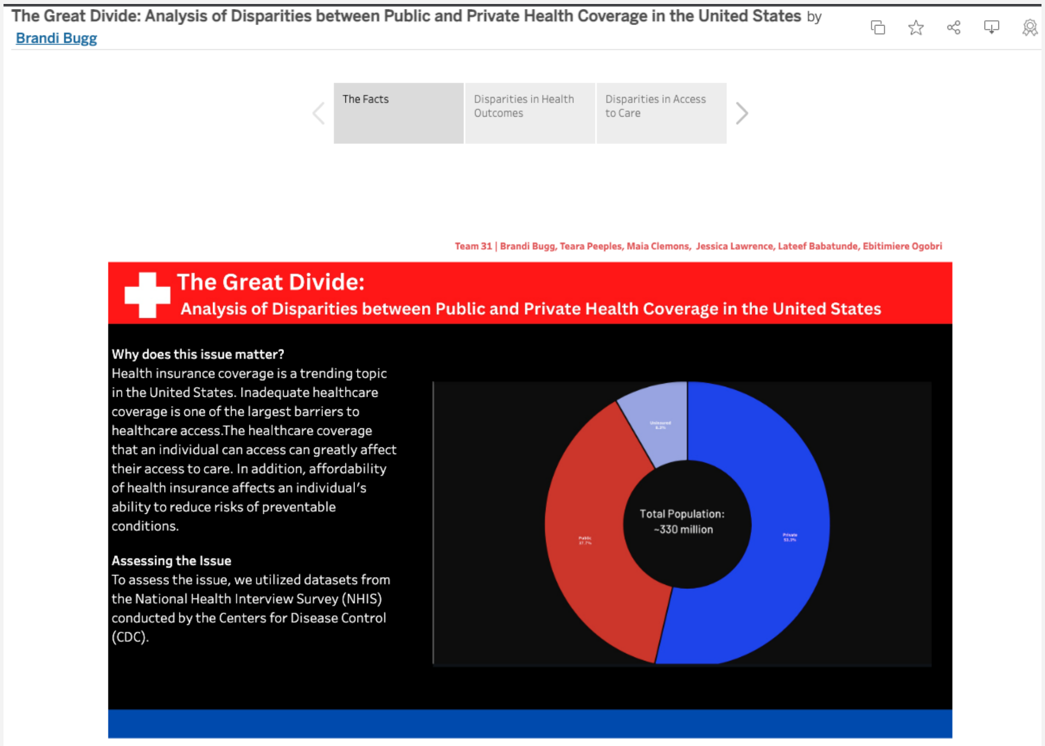
Use Case

Our dashboard was designed to showcase our most insightful analyses discovered from our dataset. The dashboard features 3 pages featuring visualizations that will allow end users to explore the impact that the type of coverage has on factors such as health outcomes and access to care.

The dashboard developed can be accessed at the following location
[The Great Divide: Analysis of Disparities between Public and Private Health Coverage in the United States]

Please note that for the purposes of all dashboards the insurance types focused on private and public insurance only.

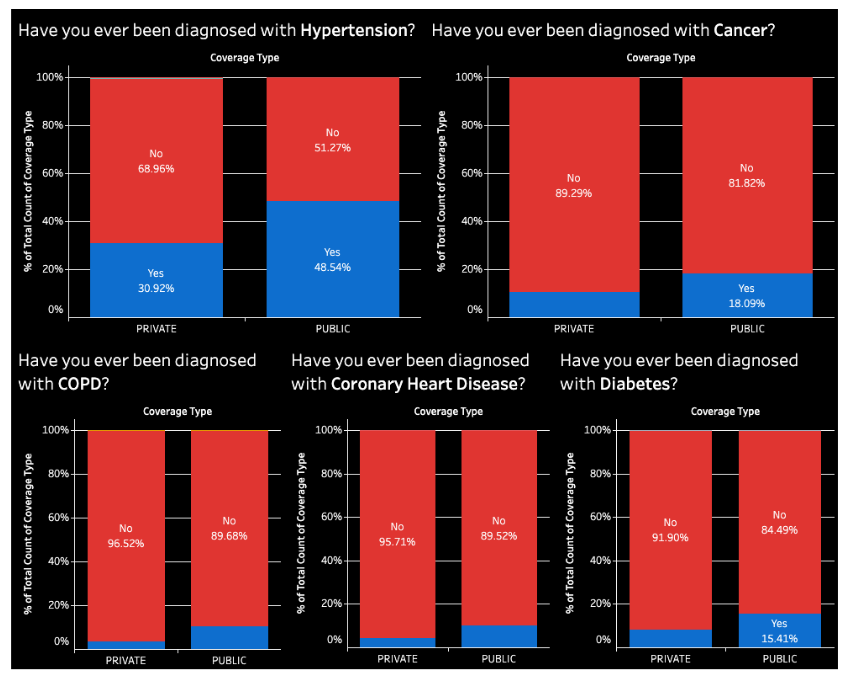
The Facts page provides an overview of the problem that we were analyzing and the purpose of the dashboard. This page introduced the issue discussed in this dashboard, the assessment of the issue, and a pie chart breaking down the total population of America and the associated percentages for the coverage types of private, public, and uninsured.



Dashboard Introductory Page

The second page of our dashboard provided an interactive page to explore Disparities in Health Outcomes. On this page of the dashboard, we spotlighted five health outcomes that are some of the most prevalent chronic diseases or chronic medical conditions in American society. Those five health outcomes featured were: Hypertension, Cancer, Chronic Obstructive Pulmonary Disease, Coronary Heart Disease, and Diabetes.

The blue area in the charts denoted that an individual stated yes while taking the National Health Interview Survey (2021) that they had a particular aforementioned disease or condition. The red area in the charts denotes that the survey taker stated that they had not been diagnosed with a certain condition.For these visualizations,we provided details such as percentages, so that it was apparent to any end user of our dashboard, the difference in percentages for public versus private. We also focused and filtered on only yes and no responses.



Disparities in Health Outcomes dashboard page

The last page of our dashboard provided an interactive page to explore Disparities in Access to Care. This dashboard looked at following in relation to disparities to access to care:

- Health Status by Coverage Type
- Challenges in Accessing Prescription Medicines
- Visits to the Hospital Emergency Room
- Challenges in Paying Medical Bills

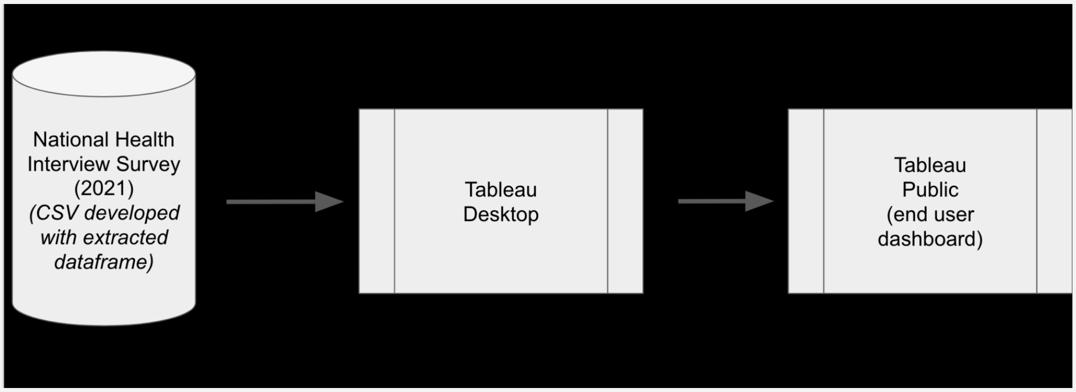
All visualizations provided here show a significant difference in the access to care, health visits, and health status in those with public insurance versus those on private insurance.

Disparities in Access to Care Dashboard Page

Data Engineering

Our dashboard was developed using the features available in Tableau. For this specific dashboard and most of our extensive analysis, we utilized one specific dataframe developed from the datasource which as mentioned above was adapted from the 2021 National Health Interview Survey. Due to the additional insights that we displayed using this dashboard, the information included was succinct in nature.

IIn developing the tool, we used a CSV file of a dataframe that we wanted to present visualizations for. This CSV file was extracted from a dataframe within our data analysis notebook. The CSV file will not need to be updated at any point, but it was extracted to make the Tableau dashboard data available for end users viewing the dashboard online.



Dashboard Data Engineering Process

Conclusion

Our data analysis revealed that there were disparities between public and private health insurance coverage in the United States. The residual value obtained from the chi-square tests guided us on choosing the variables that were significant. We then used information from the contingency tables to draw conclusions. We found that individuals with public insurance were three times more likely to be diagnosed with Chronic Obstructive Pulmonary Disease (COPD) than people with private insurance. It was also found that individuals with public insurance were two times more likely to be diagnosed with cancer. Fifty percent of people with public insurance were diagnosed with hypertension while only thirty percent of people with private insurance were diagnosed. People with private insurance were less likely to ever visit the Emergency Room(ER) in a year while people with public insurance were twice as likely to visit the ER four or more times in a year. We also found that people with public insurance were four times more likely to report "poor" health status. We understand that this does not indicate a cause-and-effect relationship but these things were shown to be correlated in our data.

	Copd				
Coverage Type	No	Yes	Refused	Not Ascertaine	Grand Total
NO COVERAGE	97.67%	2.29%	0.04%		**100.00%**
NOT AVAILABLE	93.27%	5.77%	0.96%		**100.00%**
OTHER	85.98%	13.63%	0.04%	0.34%	**100.00%**
PRIVATE	96.52%	3.37%	0.07%	0.04%	**100.00%**
PUBLIC	89.68%	10.21%	0.04%	0.07%	**100.00%**
Grand Total	**94.16%**	**5.71%**	**0.06%**	**0.07%**	**100.00%**

This figure depicts a contingency table for the variable "COPD" that represents the question "Have you ever been diagnosed with COPD?".

Future Work

We used data that was collected over three years: 2019, 2020, and 2021. COVID-19 impacted the 2020 dataset tremendously. According to NHIS, the surveys were shifted to telephone interviews which resulted in a decline in the number of surveys in comparison to previous years. This resulted in a limited sample size and some of the data being estimated. We still thought it was important to use this data because the questions had changed around this time to include some information that we found pertinent to the project. In the future, we would like to analyze different years to be sure that the population was truly represented.

The lack of data in our dataset hindered our ability to examine the relationship between insurance type and mortality rate. In future studies, it may also be beneficial to have access to a dataset where a group was followed for a long period of time. This would provide a better understanding of how the diseases and the insurance types are impacting mortality and influencing any barriers to access over a long period of time. Maybe a long term study would include more numerical data and there would be the opportunity to analyze the data in additional ways. We also did not have information on the locations of the individuals which would have given us an idea about the disparities and the differences in access to care in different parts of the country. Overall, having access to more comprehensive data would allow us to examine these relationships in a more holistic way.

References

American Progress. (Jan. 2023). Policies to Improve Health Insurance Coverage in America to Recover from COVID-19. Retrieved from https://www.americanprogress.org/article/policies-improve-health-insurance-coverage-america-recover-covid-19/

Center for American Progress. (2022). Health Disparities by Race and Ethnicity. Retrieved February 11, 2023, from https://www.americanprogress.org/article/health-disparities-race-ethnicity/

Centers for Disease Control and Prevention. (Dec. 2022). National Health Interview Survey. Retrieved from https://www.cdc.gov/visionhealth/vehss/data/national-surveys/national-health-interview-survey.html

U.S. Census Bureau. (2021). Health Insurance Coverage in the United States: 2020. Retrieved from https://www.census.gov/library/publications/2021/demo/p60-274.html

U.S. Census Bureau. (2022). Health Insurance Coverage in the United States: 2021. Retrieved from https://www.census.gov/library/publications/2022/demo/p60-278.html