

2020 Census: Data Reasonableness Review



National Academies of Sciences, Engineering, and Medicine
July 21, 2021

Jason Devine
Assistant Division Chief for Census Programs – Population Division (POP)

Christine Borman
2020 Decennial Census Count Review – Population Division (POP)

The U.S. Census Bureau reviewed this data product for unauthorized disclosure of confidential information and approved the disclosure avoidance practices applied to this release. CBDRB-FY21-POP001-0183

Shape
your future
START HERE >

United States[®]
Census
2020

2020 Census

Presentation Roadmap

This presentation provides an overview of:

1. Objectives for reasonableness review in the 2020 Census
2. The focus of reviews at each stage of file processing
3. Review teams
4. Examples of findings from 2020 review

2020 Census

Purpose of Data Review

Our goals for data review are:

- 1) To **identify data processing errors** and verify that edits and other processing steps have been properly applied.
- 2) To **assess data quality** by looking at item nonresponse/missing rates, population count only responses, proxy responses, and other early indicators of possible data quality issues.
- 3) To **evaluate demographic reasonableness** by looking at census responses and subsequent data files at multiple levels of geography compared to benchmarks, i.e., 2010 Census, American Community Survey data, and Population Estimates.

To do this, we take both a micro approach and a macro approach to the data review:

- At the **micro level**, we're looking to see if processing of individual records has been done correctly.
- At the **macro level**, we're looking to see if the aggregate results appear to be reasonable when compared to benchmark data.

2020 Census

General Categories of Anomalies

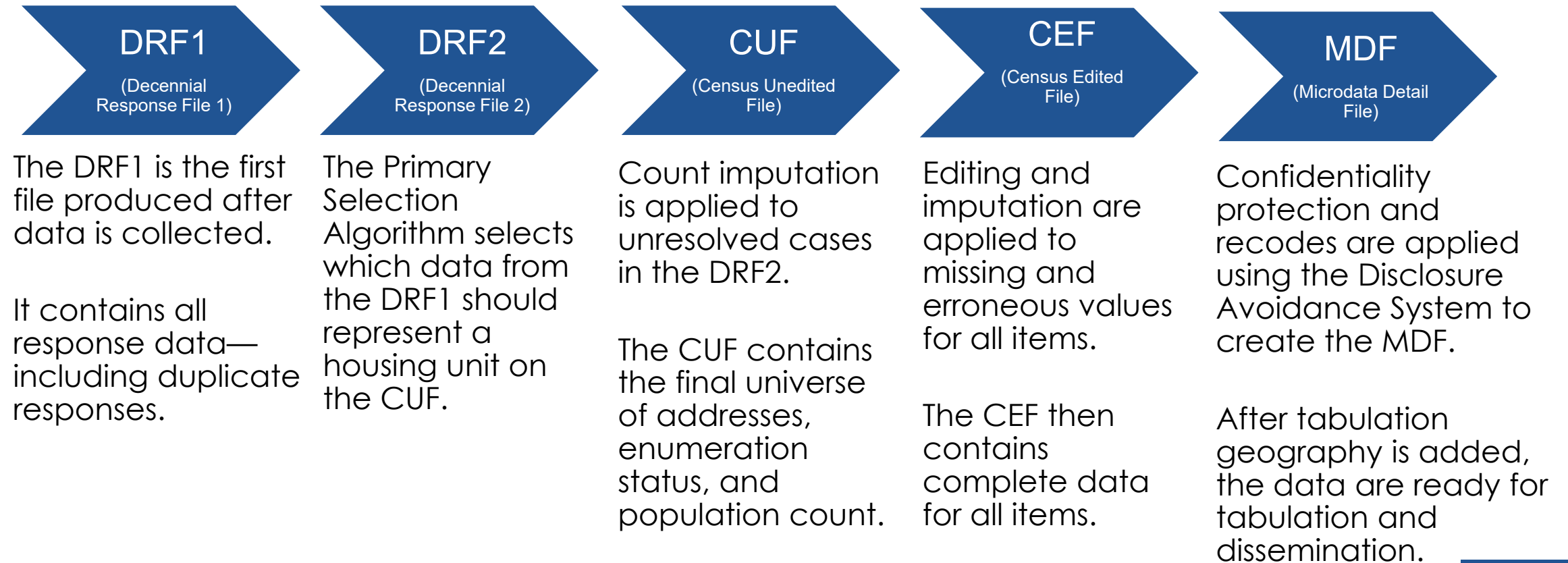
There are two general categories of anomalies – data collection anomalies and data processing anomalies.

- 1) **Data collection anomalies** are less tangible and can lead to unexpected results. While we have developed many processes to address collection-based issues, some anomalies are just outliers or unexpected trends and not “problems” to be fixed. Example of common data collection anomalies that we have built procedures for are:
 - People responding multiple times or in multiple locations (i.e. residence rule confusion)
 - People not responding – either at all or to specific demographic or housing questions
- 2) **Data processing anomalies** are more concrete and can be identified using our review programs where we double program many sections of the specification. Anomalies related to data processing could be because of:
 - A misinterpretation of the specification or a bug in the program itself
 - Situations that were not accounted for in processing specifications

2020 Census

Data Files Under Review

The response data are processed in multiple steps with reasonableness review conducted at each step:

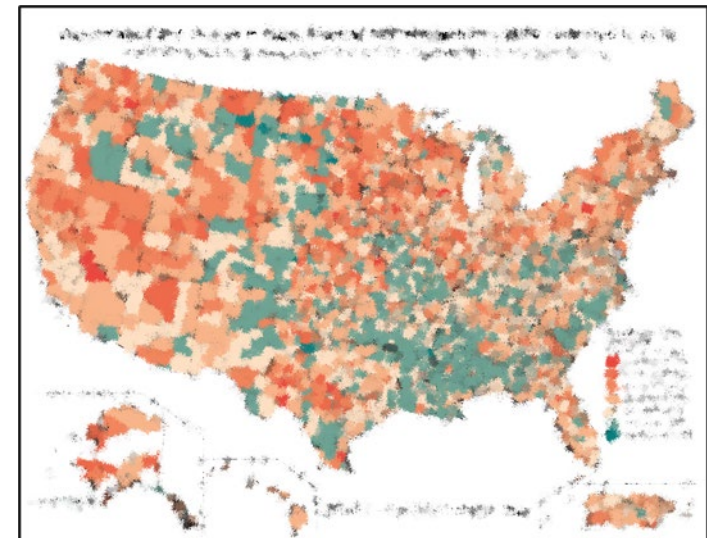
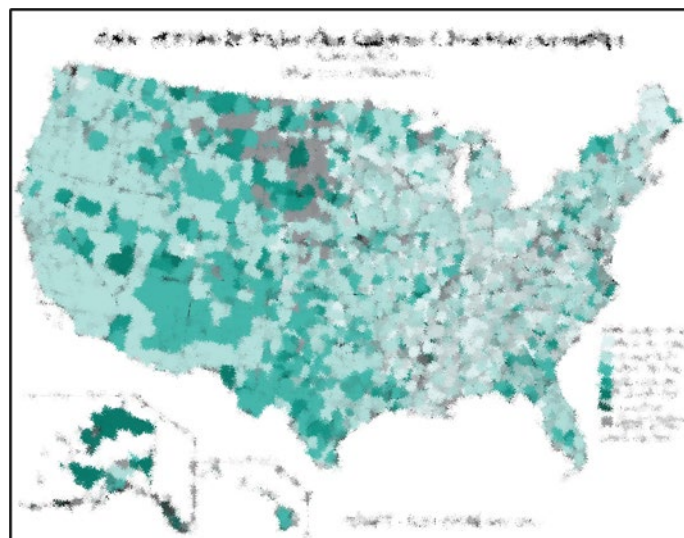
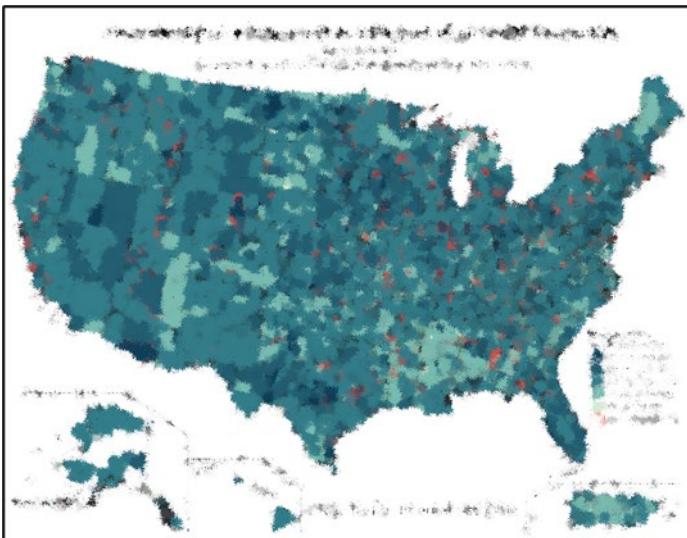


2020 Census

Review of Early 2020 Census Results

Even before production DRF1 review began, experts in the Demographic Directorate (DEMO) reviewed the data as it was collected in order to identify and correct data processing errors and assess potential data quality and reasonableness concerns.

- This early review focused on housing unit responses from self-response and nonresponse followup.
- We used a lot of mapping to allow us to see trends in the data. We know different groups respond at different times, which we saw as the data came and the maps “filled in.”
- For 2030, we have a lot of ideas about how this early review could be expanded to help us identify potential anomalies even earlier – including adding Group Quarters to the mix.



2020 Census

DRF1 Focus of Review

The DRF1 universe is larger than the final universe because it contains multiple responses for some addresses.

- **For population totals**, we focus our reviews on areas with suspected undercounts or extremely large overcounts. This includes looking at the household population and the group quarters population – where potential data collection issues may be more easily evident.
- **For characteristics**, we focus our review on identifying potential errors from data collection and response processing steps that are applied to the DRF1. We also take an early look at data quality and reasonableness.

We get this file for states on a flow basis, so we can't see the full picture for the nation until all states are delivered.

2020 Census

DRF2 Focus of Review

The DRF2 universe is *smaller* than the DRF1 because all duplicate and multiple responses have been resolved.

The DRF2 population counts should be much closer to benchmarks. Accordingly, it is easier to identify areas with populations that are outliers when compared to benchmark data.

For characteristics, the smaller universe allows us to better assess data quality and reasonableness concerns, such as item nonresponse.

We get this file for all states in the nation at the same time – allowing us to do both a national and state-level analysis as soon as we get the data.

2020 Census

CUF Focus of Review

The CUF universe is *slightly larger* than the DRF2 universe because count imputation has been applied.

Because the CUF is the basis for apportionment, we do a critical final look at the population totals compared to benchmarks.

For characteristics, we continue our review for data quality and reasonableness. We also assess the Hispanic origin and race codes from residual coding of write-in responses to ensure there were no coding or processing errors.

We get this file for states on a flow basis, so we can't see the full picture for the nation until all states are delivered.

2020 Census

CEF Focus of Review

The population counts are locked in with the CUF and do not change in the CEF.

At this point, the data in the CEF should closely represent the demographics of the nation.

The focus of this review is on characteristics, including sex, age/date of birth, Hispanic origin, race, relationship, tenure, detailed vacancy status, and group quarters type. All characteristics should now have a valid and consistent response.

We get this file for states on a flow basis, so we can't see the full picture for the nation until all states are delivered.

2020 Census

MDF Focus of Review

At this stage:

- State-level population totals do **not** change.
- The number of housing units and group quarters facilities by type at the block level do **not** change.
- Below-state population counts, person characteristics, and housing unit characteristics **do** change at the lower levels of geography **because of the application of disclosure avoidance**.

For both population totals and characteristics, teams review the MDF with the goal of ensuring that the MDF data are consistent with the expected effects of disclosure avoidance.

We get this file for all states in the nation all at the same time – allowing us to do both a national and state-level analysis as soon as we get the data.

2020 Census Review Teams

There are three DEMO review teams that actively review the census data files – Subject Matter Experts (SMEs), General Experts (GEs), and the Special Total Population team (STP).

DEMO Subject Matter Experts (SME)	DEMO General Experts (GE)	DEMO Special Total Population (STP)
Focus on reasonableness and quality of characteristics using comparisons to benchmarks	Focus on aggregate population, group quarters, and housing units totals	Focus on reasonableness of state population totals for apportionment
Identify data collection and response processing errors	Identify deviations from benchmarks for lower levels of geography	Identify deviations from benchmarks as well as possible demographic trends

2020 Census

Subject Matter Experts (SMEs) - Overview

SMEs come from content-specific branches within DEMO. These analysts focus on specific topics – such as age, sex, race, or Hispanic origin – and have worked on the specifications for both data collections and response processing throughout the decade.

Their primary focuses are the reasonableness of specific characteristics of the population and characteristic-level data quality – particularly looking at item nonresponse, imputation rates, and for specific processing errors that may be impacting the quality of the data. Example of review activities are:

- Ensure all expected variables have values and verify that data collection variables are output as expected
- Examine data reasonableness and item nonresponse rates - by data collection mode - to check for issues with collection and capture of the data
- Verify specific processing steps
- Review percent changes and raw differences in characteristic distributions at multiple levels of geography compared to benchmarks

2020 Census

Examples of SME Review Steps – Mode-Based Variable Checks

The screenshot shows the 2020 Census web interface. At the top, there's a navigation bar with the United States Census 2020 logo, a link to 'Here's how you know', and links for 'FAQ', 'INSTRUCTIONS', and 'ENGLISH'. Below this is a progress bar with four steps: 'Address Verification', 'Household Questions', 'People Questions' (which is highlighted), and 'Final Questions'. The main content area is titled 'Home > Relationship > Sex'. It contains a confirmation prompt: 'Please confirm that your answers are correct. John Doe is recorded as female. Is that correct?' with radio buttons for 'Yes' (selected) and 'No'. Below this is another confirmation prompt: 'Jane Doe is recorded as female. Is that correct?' with radio buttons for 'Yes' and 'No' (selected). Then, it asks 'What is Jane Doe's sex?' with radio buttons for 'Male' and 'Female' (selected). Finally, it asks 'How is John Doe related to Jane Doe?' with a dropdown menu showing 'John Doe is Jane Doe's...' and a list of relationship options: 'Opposite-sex husband/wife/spouse', 'Opposite-sex unmarried partner', 'Same-sex husband/wife/spouse', 'Same-sex unmarried partner', and 'Biological son or daughter'.

Internet Self Response (ISR) Sex-Relationship consistency edit is associated with a specific set of variables and contain the abbreviation “CH” in the variable name.

The sex-relationship consistency edit-specific variables should **never** be filled by responses coming from the paper mode. If it was, then this might indicate a data-transfer or processing error.

The purpose of this review step is to ensure the data itself has not been corrupted or transferred incorrectly during file creation and upstream data collection processes.

2020 Census

Examples of SME Review Steps – Mode-Based Reasonableness Checks

SMEs examine data reasonableness and item nonresponse rates (INR) by data collection mode to check for issues with collection and capture of the data.

Data Collection Mode	N	Missing Sex	Missing Age and Year of Birth	Missing Hispanic Origin	Missing Race	Missing Relationship
2010 Census Total Pop INR from the 2010 CUF						
2020 DRF2 Total Population						
Internet Self Response (ISR)						
Paper Self Response						
NRFU Production						
Census Questionnaire Assistance (CQA)						
Coverage Improvement (CI)						
Update Leave (UL)						
GQ eResponse						
GQ Facility Self-enumeration						
GQ Paper Listing						
NRFU Administrative Records Enumeration						
2020 DRF2 Household Population						
2020 DRF2 Group Quarters Population						

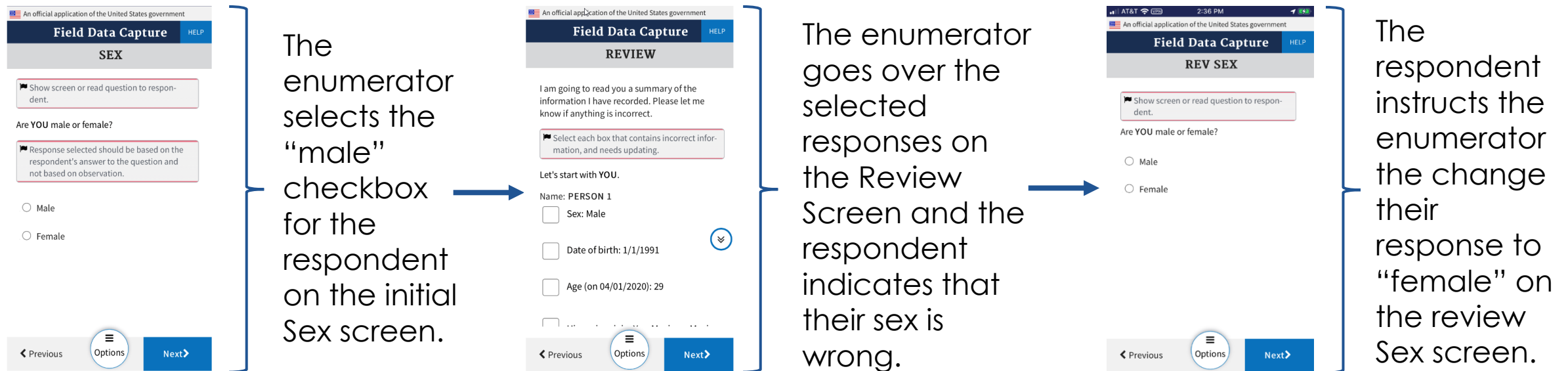
Questions when looking at INR by modes:

- Which mode(s) have the highest INR?
 - Is the INR reasonable based on mode-specific considerations?
- Which person characteristics have the highest INR?
- How does the INR for the total population compare to the 2010 Census?
- How does the INR compare between the household population and the group quarters population?

2020 Census

Examples of SME Review Steps – DRF1 Processing Check

SMEs verify specific processing steps by double programming portions of the DRF1 specification. In one step, DRF1 processing chooses between the responses based on a set of rules created by POP and SEHSD to determine the respondent's final response.



In this example, the respondent's final response to the sex question should be "Female" in the DRF1.

2020 Census

Examples of SME Review Steps – DRF1 Processing Check (Baby Flag)

When people are confirmed to be born after Census day in the Census Questionnaire Assistance (CQA) instrument, they are flagged as confirmed babies born after Census day in DRF1 processing and removed from the Census.

Eligible date of births include:

- any birthdays starting with April 2, 2020
- incomplete date of births, such as May 2020, where the day is missing but the month is after April

SMEs verify that this flag is set correctly through their double programming.

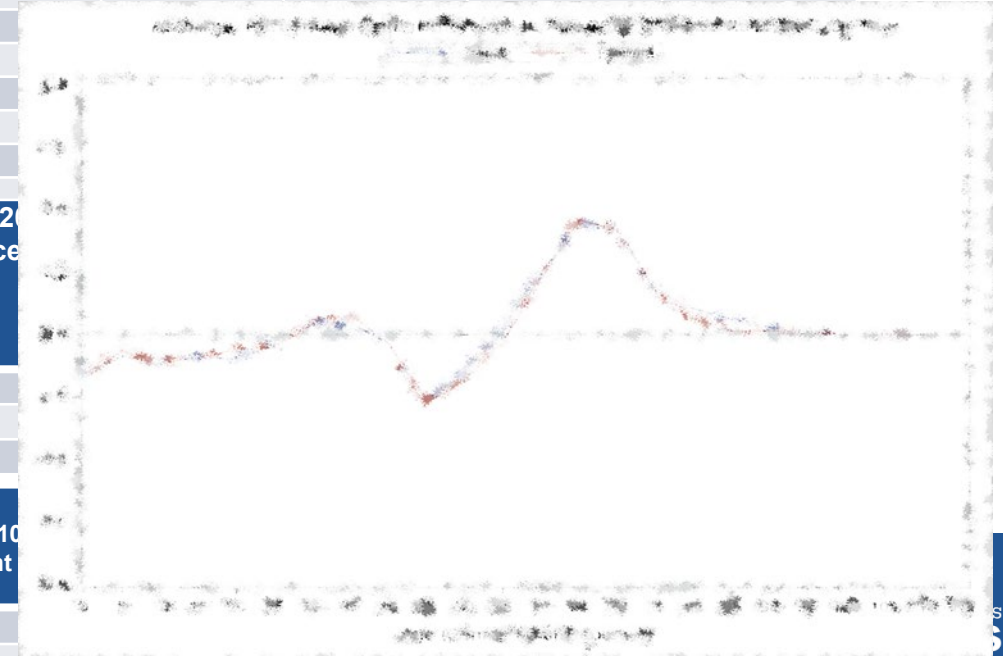
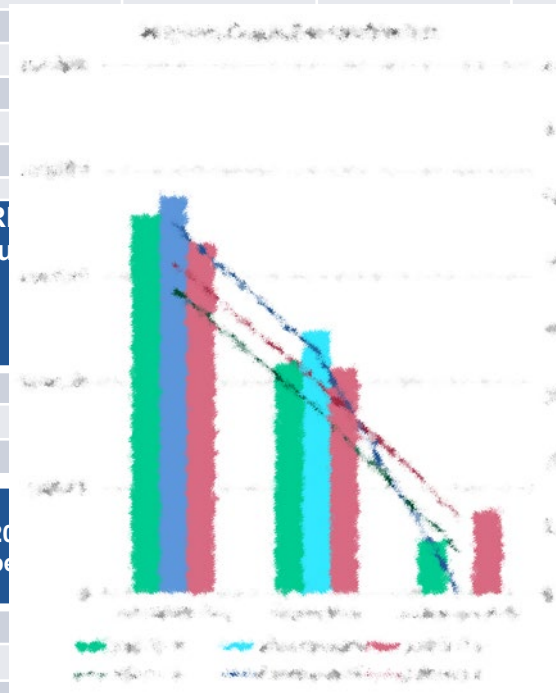
The screenshot shows the '2020 Census - Inbound' interface. At the top, there's a navigation bar with '2020 Census - Inbound', 'FAQ INSTRUCTIONS', and a globe icon for 'ENGLISH'. Below this is a progress bar with four steps: 'Address Verification', 'Household Questions', 'People Questions' (which is highlighted), and 'Final Questions'. The main content area is titled 'BABYFLAG' and contains the text: 'For the 2020 Census, we need to record age as of April 1, 2020. So, just to confirm, Katie was born after April 1, 2020?'. There are three radio button options: 'Yes', 'No', and 'Other Options'. Under 'Other Options', there are two more radio buttons: 'Don't Know' and 'Refused'. A 'Help' link is visible on the right. At the bottom right, there are 'Previous' and 'Next' buttons. The footer of the interface shows a list of languages: ENGLISH, ESPAÑOL, 中文(简体), TIẾNG VIỆT, 한국어, РУССКИЙ, العربية, TAGALOG, POLSKI, FRANÇAIS, KREYÒL AYISYEN, PORTUGUÊS, and 日本語. A small ID number '93c1541d-61b6-47d5-ba1f-3cha6a314f4d' is visible at the very bottom.

2020 Census

Subject Matter Experts (SMEs) – Reasonableness Review

SMEs review percent changes and raw differences in characteristic distributions at multiple levels of geography compared to benchmarks.

Race	CUF 2010 Number	ACS 2019 Number	DRF 2020 Number	Number Difference CUF 2010 vs DRF 2020	Number Difference ACS 2019 vs DRF 2020	CUF 2010 Percent	ACS 2019 Percent	DRF 2020 Percent	Percent Difference CUF 2010 vs DRF 2020	Percent Difference ACS 2019 vs DRF 2020
White alone										
Black alone										
AIAN alone										
Asian alone										
NHPI alone										
Some other race alone										
Two or more races										
Non-response										
Hispanic Origin	CUF 2010 Number	Estimates 2019 Number	DRF 2020 Number							
Not Hispanic										
Hispanic										
No response										
Age (Computed or Reported)	CUF 2010 Number	Estimates 2019 Number	DRF2020 Number							
0-17										
18-64										
65+										



2020 Census

General Experts (GEs) - Overview

GEs are volunteers from within DEMO who have experience with reviewing and analyzing data for reasonableness from their work on other surveys.

- The focus of GE review is the same across files - to identify any systematic issues or anomalies between the census files and benchmark data for the total population, group quarters population, and housing units at various levels of geography.
- GEs use a review tool specifically designed for their use, called the Census Review, Analysis, and Visualization Application (CRAVA).
- In the context of the GE review, an outlier is considered a geography with a 2020 tabulation that appears unreasonable or unexpected when compared to its benchmark tabulation.
- Outliers may exist because of a processing or data collection error, or because of explainable population shifts, such as:
 - changes to the geography's boundaries,
 - natural disasters that require evacuations from a particular area,
 - a new prison or college that increases the group quarters population,
 - economic forces that pull or push people away from a geography.

2020 Census CRAVA Home Page

*Not real data – for illustration only

United States[®]
Census
2020

Census Review, Analysis, and Visualization Application

States Available: 2
States not Available: 50

Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Delaware
District of Columbia
Florida
Georgia
Hawaii
Idaho
Illinois
Indiana
Iowa
Kansas
Kentucky
Louisiana
Maine
Maryland
Massachusetts
Michigan
Minnesota
Mississippi
Missouri

Montana
Nebraska
Nevada
New Hampshire
New Jersey
New Mexico
New York
North Carolina
North Dakota
Ohio
Oklahoma
Oregon
Pennsylvania
Rhode Island
South Carolina
South Dakota
Tennessee
Texas
Utah
Vermont
Virginia
Washington
West Virginia
Wisconsin
Wyoming
Puerto Rico

Welcome
To begin, select a state from the list or map below.

Mapbox

OSM

Keep Only

Exclude

State Name: New York

State ID: 36

Data Available: Yes

New York

Field	Total Population	Population Aged 18+	Households	Avg. Household Size	Housing Units	Vacancy Rate	Group Quarters Population
Total 2020	6,042,710	4,704,671	2,101,000	2.00	2,448,984	5.7	140,400
Total 2018	5,905,717	4,652,457	2,166,309	2.07	2,426,745	5.5	140,400
Total 2010	5,780,642	4,435,694	2,121,047	2.02	2,380,133	5.3	137,930
Numeric Difference 2020-2018	56,001	52,204	14,701	0.01	22,239	0.2	15
Numeric Difference 2020-2010	254,076	268,977	60,046	0.06	68,851	0.4	2,950
Percent Difference 2020-2018	0.94	1.12	0.68	0.37	0.92	3.64	0.01
Percent Difference 2020-2010	4.39	6.06	2.83	2.29	2.89	7.55	1.85

Select a topic to analyze:

General Expert Review

Topic 1: Age

Topic 2: Hispanic Origin

Topic 3: Race

Topic 4: Relationship

Topic 5: Sex

Topic 6: Housing

Topic 7: Group Quarters

From the home page, reviewers can select the state they are assigned to review – in this example, New York – and the topic area that they want to review – 1) General Expert Review, 2) Age, 3) Hispanic Origin, 4) Race, 5) Relationship, 6) Sex, 7) Housing, 8) Group Quarters.

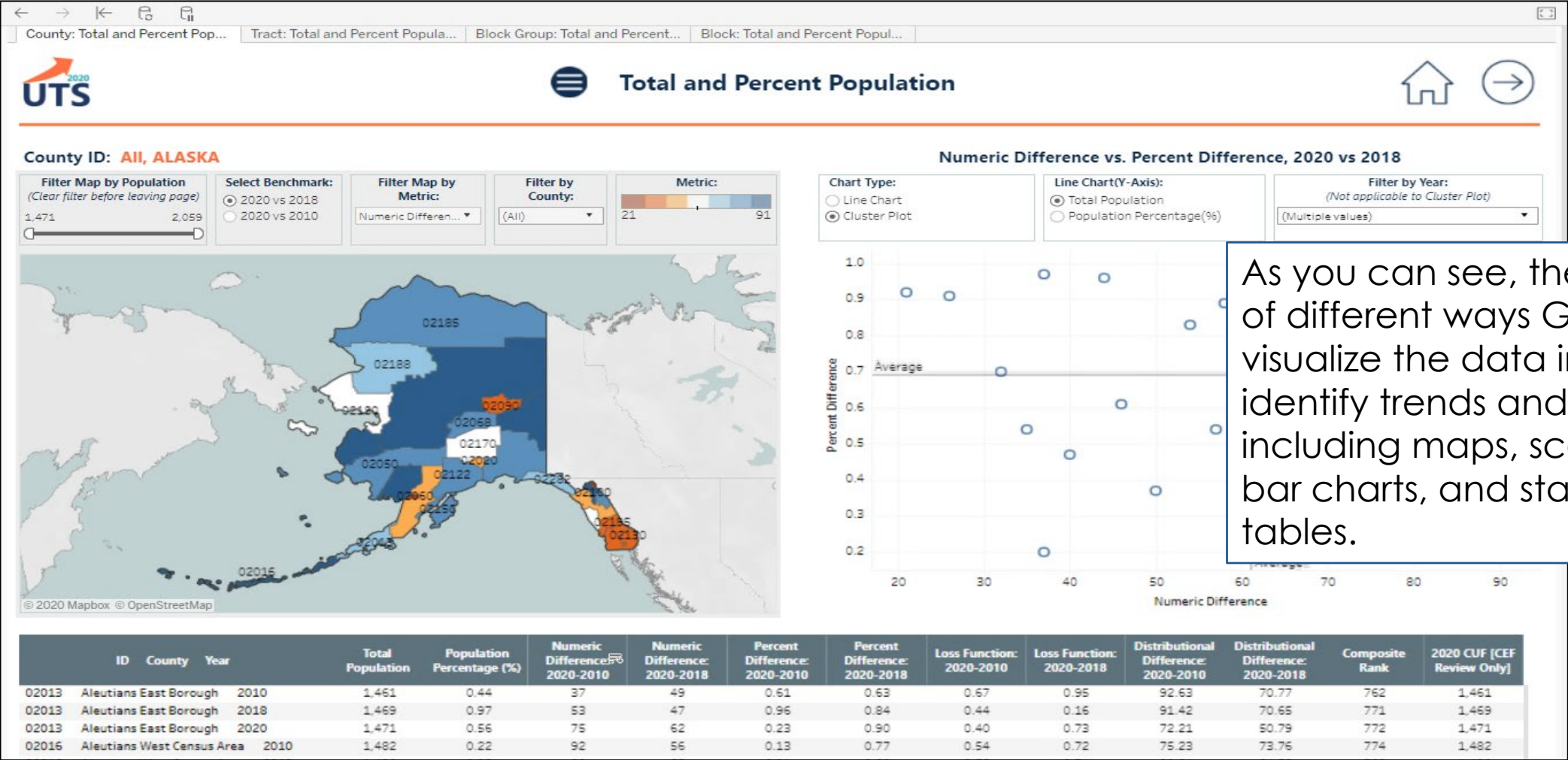
20 2020CENSUS.GOV

Shape
your future
START HERE >

United States[®]
Census
2020

2020 Census CRAVA Review Screen Example 1

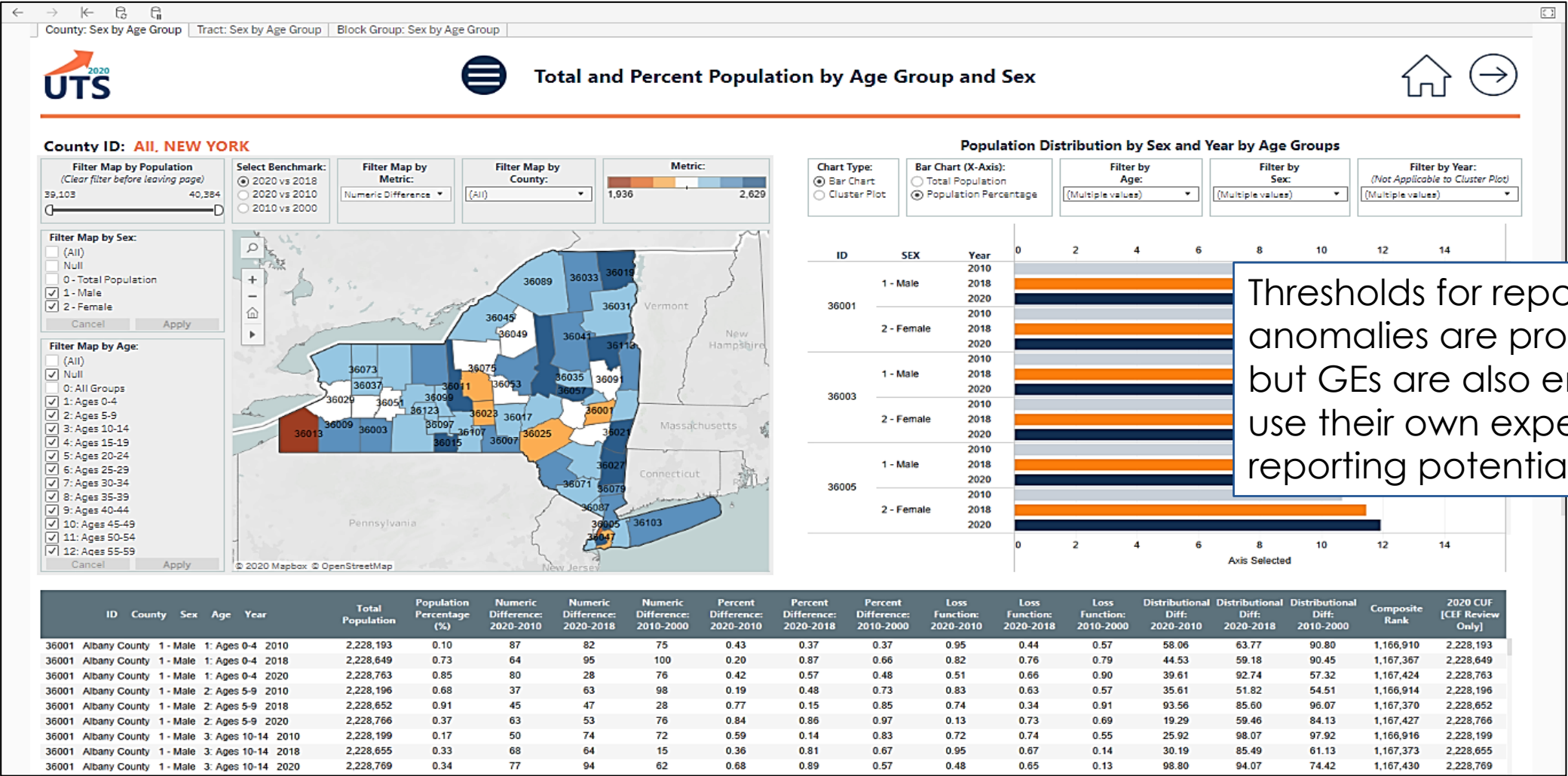
*Not real data – for illustration only



2020 Census

CRAVA Review Screen Example 2

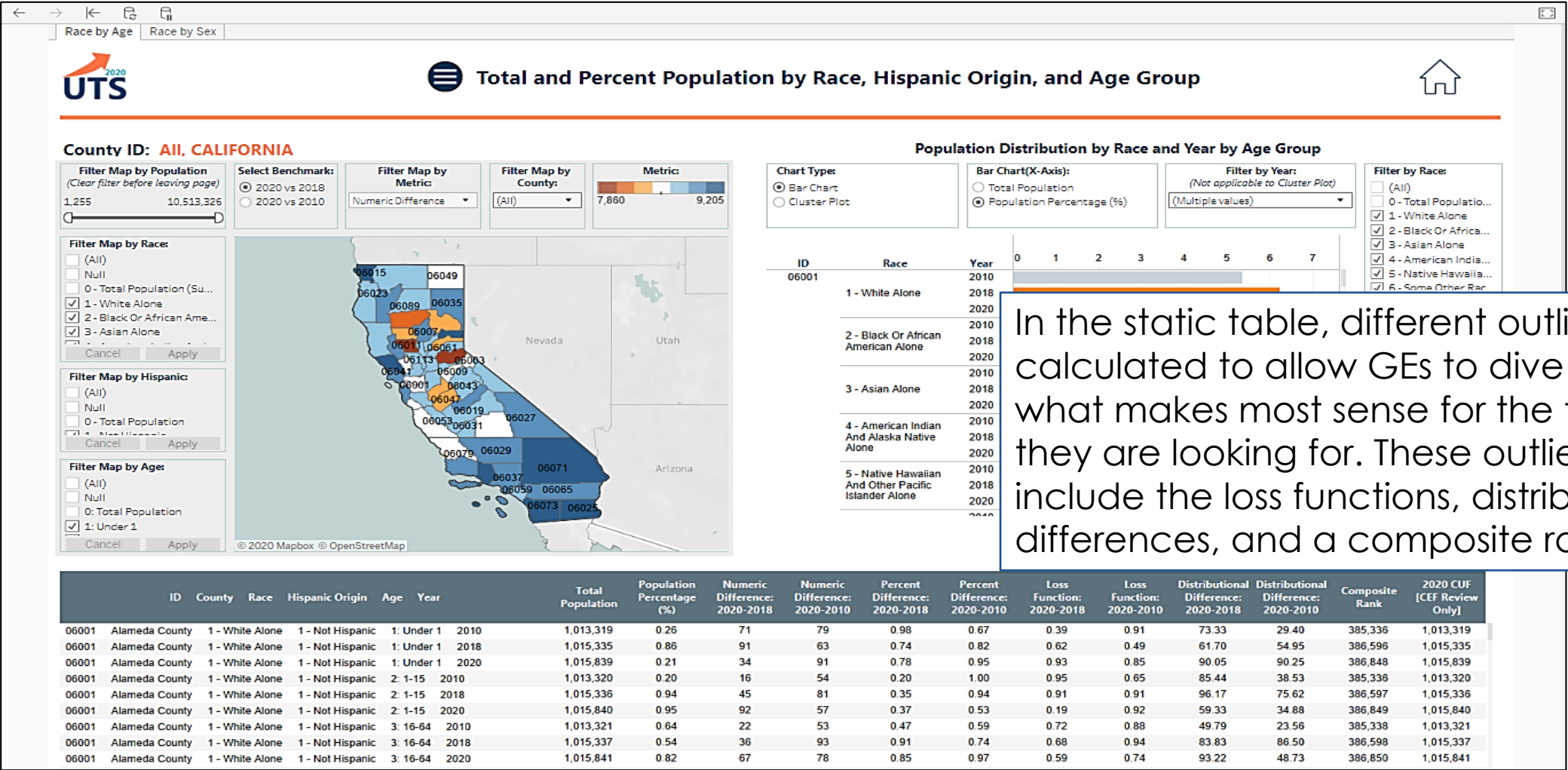
*Not real data – for illustration only



2020 Census

CRAVA Review Screen Example 3

*Not real data – for illustration only



In the static table, different outlier metrics are calculated to allow GEs to dive in and use what makes most sense for the topic and what they are looking for. These outlier metrics include the loss functions, distributional differences, and a composite rank.

2020 Census

Special Total Population Team (STP) – Overview

The STP is made up of experienced analysts and demographers with a variety of backgrounds.

- The STP uses characteristics data to assess the total population counts - looking for potential over and under counts as well as duplication - at the state level and below.
- The team uses a combination of CRAVA and independent tabulations and mapping to identify deviations from benchmarks as well as demographic trends, possible impacts of COVID-19, and data collection operational changes on the census results.
- Beyond nation, state, and county totals, the STP also looks at incorporated places and other levels of geography to identify potential over and undercounts.
- For each stage of review, the STP compares the results to benchmarks based on how we expect the numbers to look at that point in processing.

2020 Census

Examples of Findings from 2020 Review

- Processing errors – both where the code itself was incorrect and where the data led to processing interacting in unexpected ways that needed to be corrected. See Michael Thieme's blog: [Finding 'Anomalies' Illustrates 2020 Census Quality Checks Are Working](#)
- Undercounts and overcounts of group quarters populations. See Deborah Stempowski's blog: [2020 Census Group Quarters](#)
- Anomalies where the population shifted and did not require a fix. See Jason Devine, et al's blog: [2020 Census Data Review](#)

Questions

END