Extraction, Transformation, and Load Technical Report

Affordable Housing:

Subsidized Housing in Metro ATL

**TABLE OF CONTENTS**

1. [Introduction 3](#_heading=h.2xcytpi)
   1. [Summary 3](#_heading=h.30j0zll)
   2. [Scope 3](#_heading=h.3znysh7)
   3. [Technologies and resource contributions 4](#_heading=h.tyjcwt)
   4. [Definitions, Acronyms and Abbreviations 4](#_heading=h.1t3h5sf)
2. [ETL Details 6](#_heading=h.1ci93xb)
   1. [Data Import/Extract Sources and Method 7](#_heading=h.2s8eyo1)
   2. [Data Acquisition 7](#_heading=h.3rdcrjn)
   3. [Data Transform 8](#_heading=h.lnxbz9)
   4. [Data Integrity 9](#_heading=h.1ksv4uv)
   5. [Data Refresh Frequency 10](#_heading=h.2jxsxqh)
   6. [Data Security 11](#_heading=h.3j2qqm3)
   7. [Data Loading and Availability 13](#_heading=h.4i7ojhp)
3. [Data Quality 13](#_heading=h.3whwml4)

**1. INTRODUCTION**

# Summary

This project builds upon a previous project addressing [Achieving Affordable Housing in Atlanta](https://docs.google.com/document/d/1GLKGeKbesx4VNbkRXtxYxAP07vkFXKUWuxW_3zX30Ds/edit#heading=h.z6ne0og04bp5) (AAHA). This portion of the project will provide the ETL necessary to bring together all the original data sources and two new data sources into a relational database that will be easily accessible and ready to be analyzed.

A question has been asked whether the current allocation of publicly subsidized HUD housing is adequate and if new HUD housing construction is correctly targeted. This project will join data by census tract id to bring demographic information together to allow comparison to the detailed supply of HUD housing. This data should be able to answer these questions and to identify areas of need.

In this report, we provide a logical design of ETL scenarios and provide a generic and customizable framework to assist in understanding the need and adequacy of publicly subsidized housing. First, we present the definitions, acronyms and abbreviations of ETL activities. We then provide explanation of acquisition and transformation steps. We focused our dataset on geographic location to allow uniform comparison of the Atlanta market.

# Scope

We used the ACS 5 (from the Census data set), the HUD Office of Policy Development & Research Picture of Subsidized Households Data Set Income Limits, & National Housing Preservation Database. These data sources will be joined on census tract id in order to provide enough data to form the basis of various analysis.

The components of a data science project used here are:

* Identify appropriate data sources
* Perform cleansing of the data,
* Join the data to other datasets,
* Import these data sources into a database, and
* Provide documentation in the form of an ERD of the database

We will perform these steps in order to enable one to address the initial problem statement given appropriate modeling and analysis.

# Technologies and resource contributions

The final ETL project used the following methods and technologies.

* Data Import/Extract Sources and Method – Alex Novas
* Data Acquisition – Alex Novas, Hudson Heatley
* Data Transform - Alex Novas, Sanaz Paran, Kevin Patterson
* Data Integrity, Data Quality – Melissa Bader
* Data Refresh Frequency and Data Security – Kevin Patterson
* Data Loading and Availability – Alex Novas

The tech stack used included:

* Python
* Jupyter Notebook
* Quick DBD
* DBeaver

# 1.4 Definitions, Acronyms and Abbreviations

* ETL
  + Extract, Transform, Load
  + **Extract** is the process of reading data from a database. In this stage, the data is collected, often from multiple and different types of sources.
  + **Transform** is the process of converting the extracted data from its previous form into the form it needs to be in so that it can be placed into another database. Transformation occurs by using rules or lookup tables or by combining the data with other data.
  + **Load** is the process of writing the data into the target database.
* API
  + Application Programming Interface
  + An application programming interface (API) is an interface or communication protocol between a client and a server intended to simplify the building of client-side software. It has been described as a “contract” between the client and the server, such that if the client makes a request in a specific format, it will always get a response in a specific format or initiate a defined action.[1]
  + An API may be for a web-based system, operating system, database system, computer hardware, or software library
* HUD - US Department of Housing and Urban Development
  + <https://www.hud.gov/>
  + A cabinet level agency, the Department of Housing and Urban Development (HUD) oversees federal programs designed to help Americans meet their housing needs. HUD seeks to increase homeownership, support community development and increase access to affordable housing free from discrimination.
  + HUDUser: Picture of Subsidized Households
  + <https://www.huduser.gov/portal/datasets/assthsg.html#2009-2018_query>
  + Since passage of the U.S. Housing Act of 1937, the federal government has provided housing assistance to low-income renters. Most of these housing subsidies were provided under programs administered by the U.S. Department of Housing and Urban Development (HUD) or predecessor agencies. All programs covered in this report provide subsidies that reduce rents for low-income tenants who meet program eligibility requirements. Generally, households pay rent equal to 30 percent of their incomes, after deductions, while the federal government pays the remainder of rent or rental costs. To qualify for a subsidy, an applicant’s income must initially fall below a certain income limit. These income limits are HUD-determined, location specific, and vary by household size. Applicants for housing assistance are usually placed on a waiting list until a subsidized unit becomes available.
* NHPD - National Housing Preservation Database
  + <https://preservationdatabase.org/>
  + The National Housing Preservation Database (NHPD) was created by the Public and Affordable Housing Research Corporation (PAHRC) and the National Low Income Housing Coalition (NLIHC) in 2011 in an effort to provide communities with the information they need to effectively preserve their stock of public and affordable housing. The database provides over 5,000 users access to de-duplicated information on federally assisted housing inventory across the US at no cost. The NHPD was created thanks to the generous support of the MacArthur Foundation, Fannie Mae, and HAI Group.
  + The NHPD is an address-level inventory of federally assisted rental housing in the US. The agencies and departments that fund these programs have data on the individual programs that they manage, but there is no central location where all these data are integrated. This makes it difficult to get a clear picture of the current stock of public and affordable housing in a community. It also means those who wish to preserve public and affordable housing in their community, cannot easily get the information they need about properties. By creating the NHPD, the PAHRC and NLIHC hope to address these issues.
* ACS - American Community Survey
  + <https://www.census.gov/programs-surveys/acs>
  + The American Community Survey (ACS) is an ongoing survey by the U.S. Census Bureau. It regularly gathers information previously contained only in the long form of the decennial census, such as ancestry, citizenship, educational attainment, income, language proficiency, migration, disability, employment, and housing characteristics. These data are used by many public-sector, private-sector, and not-for-profit stakeholders to allocate funding, track shifting demographics, plan for emergencies, and learn about local communities.[1] Sent to approximately 295,000 addresses monthly (or 3.5 million per year), it is the largest household survey that the Census Bureau administers.[2]
* Neighborhood Nexus
  + <http://www.weaveatlanta.org/weavejs/?file=../weave_all_ct10_mobile.weave>
  + Neighborhood Nexus is a community intelligence system providing over five thousand data variables, from the Census and many other sources, at different levels of geography. Bundled state-of-the-art visualization tools help users to understand and analyze these data. Our goal at Nexus is to support a network of community leaders and residents, government and businesses, advocates and service providers with the information, tools and expertise to make data-driven decisions, help meet challenges, leverage assets, and create new opportunities for policy intervention in community problems.
  + Neighborhood Nexus supports research, analysis, and community engagement. Our Partners use the data and tools to examine past and current socioeconomic and demographic patterns; assess correlations between the equity, health and climate of communities; identify and develop benchmarking metrics; and in so doing make better community decisions.

**2. ETL DETAILS**

# Data Import/Extract Sources and Method

We utilized 3 different datasets for our ETL project.

NHPD Dataset

* Source : <https://nhpd.preservationdatabase.org/Account/Login?ReturnUrl=%2FData>
* Method: CSV download of Georgia Data

HUD Dataset

* Source : <https://www.huduser.gov/portal/datasets/assthsg.html#2009-2018_data>
* Method: API table download with backup to CSV

Census Data (ACS5)

* Source : <https://nhpd.preservationdatabase.org/Account/Login?ReturnUrl=%2FData>
* Method: API table download with backup to CSV

Census Tract Data

* Source : <https://nhpd.preservationdatabase.org/Account/Login?ReturnUrl=%2FData>
* Method: download GEO data for Georgia

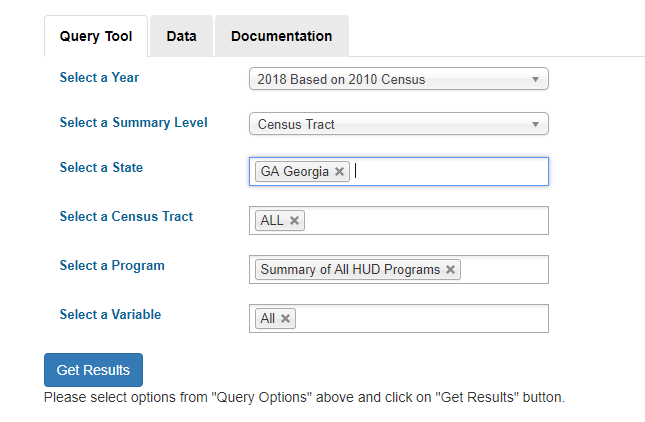
# Data Acquisition

NHPD Dataset

* This dataset is updated yearly. In order to update the dataset, go the source link log-in and download the complete database.
* Update income data variable to new reference CSV.

HUD Dataset

* This dataset is updated yearly. In order to update the dataset, go the source link and filter like below choosing the latest dataset year.



* Update HUD dataset variable to new reference CSV.

Census Data (ACS5)

* This dataset is updated yearly. In order to update this dataset, delete old tables from the data folder, update the yearstopull variable to the new reference year and rerun code.

Census Tract Data

* This dataset is updated yearly. In order to update this dataset, go to the source link, download the Georgia Data. Extract ALL files to the data/tractdata folder.
* Update geofile variable to new shp file.

# Data Transform

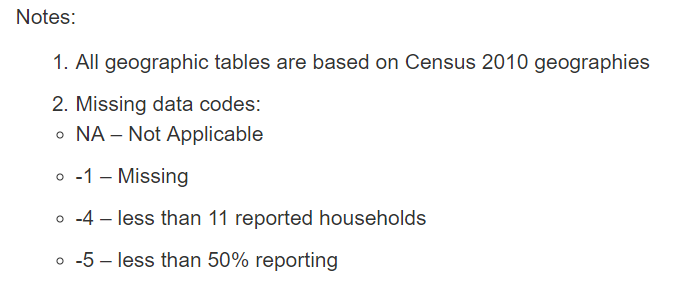
***HUD* Dataset**

HUD dataset has 66 columns with 2075 rows. Based on the needs and definition of the project, several of the columns are chosen as follows:

'Code','Subsidized units available','% Occupied','% Reported','Household income per year','% very low income','% extremely low income ,'% female head','% with disability among Head Spouse Co-head aged 61 years or less','% with disability among Head Spouse Co-head aged 62 years or older', '% in poverty (Census tract)','% minority (Census tract)'

The ‘Code’ column is the census tract ID (GEOID). With the GEOID for the city of Atlanta, the HUD data can be filtered out to the city of Atlanta instead of Georgia.

HUDUser dataset contains codes that correspond to certain data statuses:

**

**NHPD Dataset**

To meet the requirements of the joining dataset, no entries required transformation to another format. Current datatypes included integer, float and object. The original data set had 234 columns with 2177 rows. Using the “City” header column, the data was filtered using the city of “Atlanta”. Columns were further filtered by select columns based on need of the data set. Select column headers includes the following: *'NHPD Property ID','Census Tract','HUD Property ID','Property Address','City','OwnerType', 'PropertyStatus', 'Activesubsidies', 'EarliestStartDate', 'EarliestEndDate', 'TotalUnits', 'S8\_1\_AssistedUnits', 'S202\_1\_AssistedUnits', 'FHA\_1\_AssistedUnits', 'FHA\_2\_AssistedUnits', 'LIHTC\_1\_AssistedUnits', 'LIHTC\_2\_AssistedUnits', 'HOME\_1\_AssistedUnits', 'RHS515\_1\_AssistedUnits'.*

The data in the main Research Tool are displayed at the property level and come from multiple data sources. Each record in the data grid represents one address location. In most instances, one property (or development) is tied to one address.

**Census Dataset (ACS5)**

We pulled the following three tables from the census API. The main conversion we made to the tables as we downloaded them from the Census API was to automatically take coded column headings and convert those to their actual string title for clarity. At this point we also filtered for Metro Atlanta census tracts too. From these extensive tables we decided to only keep summarized information for each census tract.

* TENURE BY HOUSING COSTS AS A PERCENTAGE OF HOUSEHOLD INCOME IN THE PAST 12 MONTHS
* HOUSEHOLD INCOME BY GROSS RENT AS A PERCENTAGE OF HOUSEHOLD INCOME IN THE PAST 12 MONTHS
* MEDIAN HOUSEHOLD INCOME IN THE PAST 12 MONTHS (IN 2016 INFLATION-ADJUSTED DOLLARS) BY HOUSEHOLD SIZE

# Data Integrity

Starting with the Census Data as utilized in the American Community Survey, we found the following characteristics to be present regarding data integrity.

* All data was cleaned as part of creating the ACS data. We found no need to drop nulls, nor did we have any invalid data types, or other data integrity issues.
* The data is published yearly and does not change outside of these windows. It is meant to provide a 5-year slice of data from 2018 and earlier.
* There are no notifications of changes as it does not change as a dynamic data source.

The National Housing Preservation Database had the following data integrity characteristics.

* The data comes from both the US Department of Housing and Urban Development (HUD) and the United States Department of Agriculture (USDA) which is involved in rural housing development.
* It reports on 10 federal subsidy programs to help with housing affordability. There are no state or local subsidy programs in Georgia that are included.
* The NHPD database is refreshed 3 times per year (exact dates not specified). All data has been cleaned and deduplicated. The underlying data is refreshed as follows
  + Section 8 subsidies are reported monthly
  + All other subsidies report annually or less frequently

The following data integrity issues are known and have been reported:

* The subsidy end date was not used for Section 515 loans until September 2017. This may result in an overstatement of inactive subsidies occurring.
* Each record in the NHPD describes a property and includes only one address. A property could consist of several buildings and have several addresses. This limitation will have a negative impact on searches by property addresses.
* An overall data refresh (changes in how subsidies were reported) occurred in June of 2019. The number of changes and the impact to each program was captured in the data documentation and can be found [here.](https://preservationdatabase.org/documentation/data-notes/)

HUD Office of Policy Development and Research Datasets have the following characteristics of data integrity:

* There are many, many datasets that are available and the applicability of each data set to answer a question is assessed in a matrix [here.](https://www.huduser.gov/portal/datasets/HUD_data_matrix.html)
* Each of these data sets have a different refresh schedule so it impossible to qualify a refresh frequency. For a detailed breakdown of refresh schedule, please visit this [link.](https://www.huduser.gov/portal/pdf/Guide-to-HUD-User-Data-Sets.pdf)
* No cleanup of the data was needed beyond ETL activities. The data had been cleaned and deduplicated prior to publication.
* There is no push notification to alert you to dataset changes. You would have to download the data again, based on what data you were interested in, based on its periodicity as described above.

# Data Refresh Frequency

**NHPD Dataset Refresh**

The number of active subsidies reported in the NHPD can change over time as the format of data files change, subsidy programs expand or are phased out, and as the logic applied to determine subsidy status is updated. The NPHD data should be refreshed monthly. The procedure for updating new data and changes in the number of active subsidies and factors that affect the comparability of data over time.

**HUD Dataset Refresh**

HUD provides interested researchers with access to the original data sets generated by Policy Development & Research (PD&R) sponsored data collection efforts. The components of the PD&R data sets are typically updated every one to five years. An annual refresh of the HUD dataset is recommended.

**Census Dataset ACS5 Refresh**

The American Community Survey (ACS) is an ongoing survey that provides data every year. The 5-year estimates from the ACS are "period" estimates that represent data collected over a period. The 5-year estimates are available for all geographies down to the block group level. An annual refresh of the Census Dataset ACS5 is preferred.

# Data Security

**NHPD Database Security**

The NHPD does not collect personally identifiable information from individuals unless it is provided voluntarily. NHPD does not require the user to provide information in order to view parts of their website. Notwithstanding this, The Public and Affordable Housing Research Corporation (**PAHRC**) and The National Low-Income Housing Coalition (NLIHC) may collect and use the following kinds of personal information:

* Information about your use of this website including clicks, downloads, filters, and logins.
* Information that you provide for the purpose of registering with the website, including address, name, e-mail, title, and organization name
* Information about transactions carried out over this website including donations, data license purchases, and sponsorships.
* Information that you provide for the purpose of subscribing to the website services such as submitting user questions or technical support requests and any other information that you send to PAHRC and NLIHC.

NPHD does not share personal information – such as names, mailing addresses, telephone numbers, or email addresses – with anyone, except as necessary to process transactions and provide access to or information about NHPD’s information and services. PAHRC and NLIHC may use your personal information to:

* Administer this website;
* Personalize the website for you;
* Enable your access to and use of the website services;
* Send you products that you purchase;
* Supply you services that you purchase;
* Send to you statements and invoices;
* Collect payments from you;
* Solicit donations and sponsorships and

Send you marketing communications.

Where PAHRC and NLIHC discloses your personal information to its agents or sub-contractors for these purposes, the agent or sub-contractor in question will be obligated to use that personal information in accordance with the terms of this privacy statement.

In addition to the disclosures reasonably necessary for the purposes identified elsewhere above, PAHRC and NLIHC may disclose your personal information to the extent that it is required to do so by law or requested by governmental or law enforcement authorities, subpoena, court order or discovery request, or when we otherwise believe in good faith that such disclosure is appropriate to enforce our Terms of Use or in connection with any activity that may violate the law or rights of others, threaten the safety or security of any person or property (including NHPD) or expose us to liability.

PAHRC and NLIHC will store all the personal information you provide on its servers.

**HUD Database Security**

Personal information requested via the internet to obtain HUD User IDs is protected. The link to any of the REAC systems is a secure connection. Once entered, personal information is automatically encrypted and is only accessible by technical personnel under strictly controlled circumstances.

Personal identifiers are simply a tool to ensure that the designated person or staff using their own identifiers, have the sole authority to verify/adjust data previously submitted to HUD by the assigned PHA. Each person entrusted with such data entry requires their own ID. No other use will be made of personal information entered into the system.

The use of personal identifiers for this purpose has been standard practice by the Federal Government, as well as certain private sector entities, for some time. The Social Security Administration (SSA) guarantees that using your social security number for identification purposes does not provide unauthorized access to your social security records.

**Census Database ACS 5 Security**

All personally identifiable information, such as names, telephone numbers, and addresses, are removed from data files. Various approaches are used to protect user personal information; including computer technologies, statistical methodologies, and security procedures.

Security measures are in place to ensure that only a restricted number of authorized people have access to private information and that access is only granted to conduct the work of census bureau and for no other purposes. Every person who works with census confidential information collected by the Census Bureau is *sworn for life to uphold the law.*

*Violating the confidentiality of a respondent is a federal crime with serious penalties, including a federal prison sentence of up to five years, a fine of up to $250,000, or both.*

# Data Loading and Availability

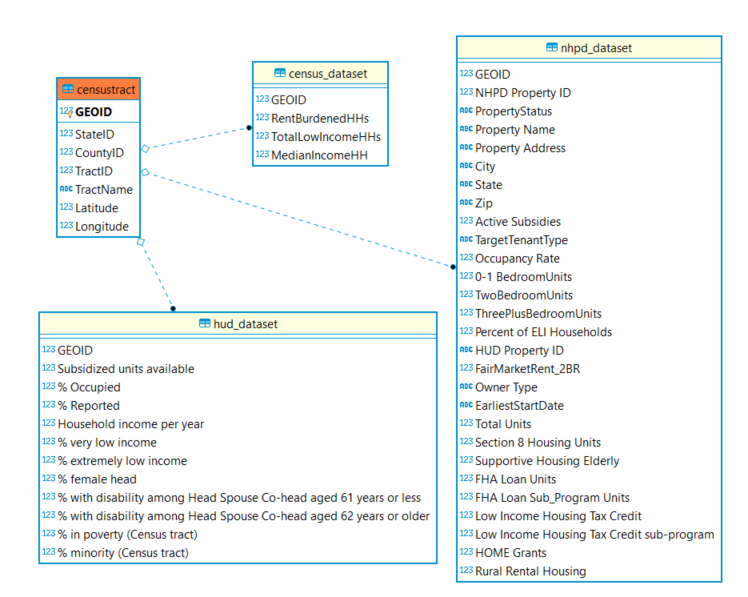
All the transformed datasets have been loaded onto an SQLite database. We recommend using DBeaver to open the SQLite file and use that as an interface to browse the data stored in the database.

census tract PK GEOID

census dataset FK GEOID

nhpd\_dataset FK GEOID

hud\_dataset FK GEOID



**3. DATA QUALITY**

Our team took a more industry standard (dama.org) approach to what data quality is and deviates a bit from the guidance given. Data quality in the simplest understanding is ‘how fit is the data for the intended business purpose’.

Since we do not own these data sets, nor did we do any additional cleaning of the data for data quality, we feel that the ask here is to explain what data quality measures were used by the upstream data owners.

Data quality had been performed in advance of the publication of the datasets. Any data quality impacts have been documented. As we cannot know the specific impact of potential data quality issues, ideally, we could assign a data quality metric to each data set but that is beyond the scope of the knowledge of the data source as exposed to the public.

To share more specifically, let’s look at the data quality metrics given:

For the American Community Survey datasets, we narrowed the data sets to the state of Georgia. There were four quality measurements utilized. We include the metadata about the criteria of data quality employed in each of these dimensions.

* [Coverage rates](https://www.census.gov/acs/www/methodology/sample-size-and-data-quality/coverage-rates/index.php) analyzed to inform the overall data quality of the results
  + [Definition](https://www.census.gov/programs-surveys/acs/methodology/sample-size-and-data-quality/coverage-rates-definitions.html) of what coverage rates is based on.
* [Sample Size](https://www.census.gov/acs/www/methodology/sample-size-and-data-quality/sample-size/index.php) for the state of Georgia information
  + [Definition](https://www.census.gov/programs-surveys/acs/methodology/sample-size-and-data-quality/sample-size-definitions.html) of what measures were used for sample size
* [Response rates](https://www.census.gov/acs/www/methodology/sample-size-and-data-quality/response-rates/index.php) data quality was tracked in this analysis.
  + [Definition](https://www.census.gov/programs-surveys/acs/methodology/sample-size-and-data-quality/response-rates-definitions.html) of what measure were used for response rates
* [Item Allocation Rate](https://www.census.gov/acs/www/methodology/sample-size-and-data-quality/item-allocation-rates/index.php) to inform the breakdown across many categories
  + [Definition](https://www.census.gov/programs-surveys/acs/methodology/sample-size-and-data-quality/item-allocation-rates-definitions.html) of what these categories are, and the data quality dimension used

For the National Housing Preservation Database, we found the following information about the data quality. The issues as noted above in the data integrity session as reiterated here:

* The subsidy end date was not used for Section 515 loans until September 2017. This may result in an overstatement of inactive subsidies occurring.
* Each record in the NHPD describes a property and includes only one address. A property could consist of several buildings and have several addresses. This limitation will have a negative impact on searches by property addresses.
* An overall data refresh (changes in how subsidies were reported) occurred in June of 2019. The number of changes and the impact to each program was captured in the data documentation and can be found [here.](https://preservationdatabase.org/documentation/data-notes/)

Additionally, we found an [updated table of the upstream data sources publication dates](https://preservationdatabase.org/documentation/data-sources/) which impacts the dimension of the timeliness of the data as a data quality measure:

There is a [user guide](https://preservationdatabase.org/documentation/user-guide/) that gives guidance on what the business purpose of the data is, and how to evaluate the data.

In short, if you are evaluating the fitness of the data for the purpose of what property addresses are subsidized, you have a significant data quality issue and should find additional data sources to cross reference the data. These use cases are called out and any analysis of the impacted attributes should be weighed at a lower confidence level than that of other attributes.

For the HUD data sets, it is imperative that one uses the matrix of applicability to help one assess the fit for the intended purpose.

The matrix for how each data set suffices the analysis criteria is published [here.](https://www.huduser.gov/portal/datasets/HUD_data_matrix.html)

The analysis of the data quality rules run on each attribute which tells you exactly what was measured and how is found in the following [documentation](https://www.huduser.gov/portal/datasets/pictures/dictionary_2018.pdf).

The periodicity of the data sources is addressed [here.](https://www.huduser.gov/portal/pdf/Guide-to-HUD-User-Data-Sets.pdf)

In sum, the data sources we performed ETL on had been pre-cleaned, deduplicated and more. Any known data anomalies were pointed out in the data’s documentation. Since these data sources are a gold standard for consumption and analysis by thousands of users, we feel that the data quality is very high.

Ideally in a high-fidelity data quality model, we would have change management notification in place. One would know if any data would be changing and how.

Also, in a high-fidelity data quality model, if there were a slip in data quality, one could expect to be notified, if one were a downstream consumer, of any issues in the data and when to expect it to be resolved.

Our takeaway is one should care about the data quality and ALWAYS read through the data notes and documentation of the data source to determine any potential impacts. One should not just pass data along without making sure these anomalies are documented and stated forthright.