Wiki for Leveraging External Data Sources to Enhance Official Statistics and Products

Sallie Keller (Project Lead),
Director Social and Decision Analytics Division
Distinguished Professor in Biocomplexity,
Professor of Public Health Sciences, School of Medicine
https://orcid.org/0000-0001-7303-7267
sak9tr@virginia.edu

Stephanie Shipp https://orcid.org/0000-0002-2142-2136
Mark Orr, https://orcid.org/0000-0001-7950-8752
David Higdon
Gizem Korkmaz 0000-0002-4947-6320
Aaron Schroeder https://orcid.org/0000-0003-4372-2241
Emily Molfino https://orcid.org/0000-0002-6575-030X
Bianica Pires, https://orcid.org/0000-0002-4710-4849
Kathryn Schaefer Ziemer,
Daniel H. Weinberg https://orcid.org/0000-0002-2799-5015

Social and Decision Analytics Division
Biocomplexity Institute & Initiative
University of Virginia

January 5, 2016

Funding: This evaluation was conducted for the U.S. Census Bureau under contract to the MITRE Corporation

Citation for wiki: Keller, S., Schroeder, A., Shipp, S., Orr, M., Higdon, D., Korkmaz, G., Molfino, E., Pires, B., Ziemer, K., Weinberg, D. "Leveraging External Data Sources to Enhance Official Statistics and Products." Proceedings of the Biocomplexity Institute, **Wiki for Technical Report**. **TR# 2021-065.** University of Virginia. The dynamic wiki is at this link - https://uva-bi-sdad.github.io/census2016_wiki/ (https://doi.org/10.18130/kn89-xm38)

Citation for related technical report: Keller, S., Shipp, S., Orr, M., Higdon, D., Korkmaz, G., Schroeder A., Molfino, E., Pires, B., Ziemer, K., Weinberg, D. "Leveraging External Data Sources to Enhance Official Statistics and Products." Proceedings of the Biocomplexity Institute, Technical Report. TR# 2021-060. University of Virginia**.** https://doi.org/10.18130/fyyx-0t81.

Abstract

The Census Bureau wants to understand how to leverage external data sources with traditional survey data and understand the effects on statistical data quality and standards of use resulting from incorporating external data. The Census Bureau tasked the Social and Decision Analytics Division, part of the Biocomplexity Institute of the University of Virginia, to address the question: how can we know if external data are useful for federal statistical needs? Today, the broad expansion of data collection across state and local governments, nonprofits, and commercial entities has created many opportunities to leverage external data sources to complement and even improve federal statistics. In the context of this report, external data are those collected by organizations outside of the federal statistical system, including city, county, and state governments, and the commercial sector.

The SDAD team developed an initial data framework that encompasses the theory and methods capable of capturing, repurposing and integrating sources of data. This framework cannot be developed in isolation; rather it must be deeply grounded in real problems. Two specific case studies were chosen to ground the data framework development. The first case study considered measuring housing information directly based on locally available data. The second case study explored the use of state longitudinal education and workforce data. For both case studies, analyses were conducted to determine statistical properties, quality, accuracy, availability, and timeliness for each data source. Alternative estimates of housing and education information were created and recommendations were developed for external data sources, methodologies to produce estimates for the 2009-2013 American Community Survey (ACS) and selected uses of the ACS data.

The documentation of research to conduct is enclosed in this document. The dynamic version of the wiki is at this link: https://uva-bi-sdad.github.io/census2016_wiki/

Table of Contents

VT Census Case Studies : Data Inventory Guidelines 4
VT Census Case Studies : Housing Data Inventory 5
VT Census Case Studies : Housing Data Profiling, Preparation Process, and Benchmarking 7
Areas that Need to be Addressed 9
VT Census Case Studies : Housing Data Profiling, Preparation Process, and Benchmarking 21
Areas that Need to be Addressed 23
VT Census Case Studies: Education Data Inventory 33
VT Census Case Studies : Education Data Profiling, Preparation Process, and Benchmarking 36
Definitions 37
Issues in Benchmarking 37
Areas that Need to be Addressed 38
Documentation of Assumptions 45
ACS Education Tables and External Data 48
Benchmarking Rules 50

VT Census Case Studies : Data Inventory Guidelines

Created by molfe, last modified by steph19 on 22 Dec 2015

As stated in our project description, we will develop a comprehensive list of data sources for housing and education. This page describes the steps taken to complete this task.

Data Dictionary

Key throughout this process is agreement on definitions of key terms found and used throughout the inventory process. Thus, a data dictionary was created. This list is a working document to be returned to and updated throughout the inventory process. The document can be found **here.**

Screening

In this stage, each data source will have a brief description.

Given the large number of data sources, we will triage the sources and complete inventories on those that meet our screening criteria. These criteria were based on the purpose of this data inventory and subsequent case studies.

- Are the data collected opinion-based (e.g., people's attitudes, preferences, etc.)?
- Are the data collection recurring (must be collected at least annually)?
- Are there data available for 2013?
- Geographic granularity
    - For Education: Is the data collected at school level?
    - For Housing: Is the data collected at the property or housing unit level?
- Can we access the data by August 15th?

Short Inventory

During this screen phase, a smaller set of additional information was gathered to insure an adequate understanding of the data set, accurate responses to the screening questions, and to aid with the data description. These included questions:

- Purpose
- Method
- Description
- Selectivity
- Accessibility

The Shortened Inventory sheet can be found **here.**

Full Inventory

If a data source passes the above triage, the complete data inventory form is completed, which can be found **here.** Just like the dictionary, this is a working document that is updated throughout the data inventory process.

This long inventory form was created with the ultimate use of this project in mind ands contain questions pertaining to:

- Purpose
- Method
- Description
- Metadata
- Selectivity
- Stability/coherence
- Accuracy
- Accessibility

- Privacy and security,
- Research
- Gaps and concerns.

Attachments:

Census Data Inventory Definitions.docx (application/vnd.openxmlformats-officedocument.wordprocessingml.document)
AAPOR BigDataTaskForceReport_FINAL_2_12_15.pdf (application/pdf)
Data Quality Assessment Tool for Administrative Data.pdf (application/pdf)
O'Hara AR_Quality_Dimensions_draft.pdf (application/pdf)
NAS Frontiers in Massive Data Analysis.pdf (application/pdf)
NAS Training Students to Extract Value from Big Data.pdf (application/pdf)
Berard Quality_Task_Team_StatCan_Toronto.v3.pdf (application/pdf)
UN 2015 summary.pdf (application/pdf)
UN Big Data Quality Framework - final- Jan08-2015.pdf (application/pdf)
Census Data Inventory Full Template.docx (application/vnd.openxmlformats-officedocument.wordprocessingml.document)
Census Data Inventory Shortened Template.docx (application/vnd.openxmlformats-officedocument.wordprocessingml.document)

VT Census Case Studies : Housing Data Inventory

Created by molfe, last modified by sallie41 on 15 Nov 2015

This is a list of housing related data sources that have been inventoried. Clicking each source will take you to a separate site that includes information on that source's triage or triage and data inventory. The list of data sources reviewed, links to sources, and status of review can be found **here.**

Commercial:

- Black Knight Financial Services (BKFS)
- MPF Research
- National Association of REALTORS
- Real Capital Analytics
- Zillow
- Mortgage Bankers Association
- Redfin
- CoreLogic
- National Association of Home Builders and "Housing Economies" (Data/Forecasting Companion Site to NAHB)
- MLS Data
- MRIS (metropolitan regional information system)
- RealtyTrac
- WegoWise
- Williamsburg Local MLS Data
- Equifax Credit Scores
- TransUnion Credit Data
- Experian
- Foot Traffic - SentriLock
- Axiometrics, Inc
- Planet Labs
- Blackbridge
- CoStar

Local:

- Arlington County: CPHD data
- Arlington County: Permitting
- Arlington County: Real Estate Assessments
- Arlington Economic Development
- Arlington County: Mapping Center
- Arlington County: Crime data
- Arlington County: Building Energy Report Cards
- Arlington County: Bicycle & Pedestrian Counters
- Arlington County Affordable Housing Study: Resident Poll Results
- James City County: Real Estate Assessments
- James City County: GIS/Mapping
- James City County: Crime Data
- James City County Citizen Survey
- James City County Housing and Community Development Programs

State:

- Housing Virginia
- Northern Virginia Association of Realtors
- VHDA Housing Analysis
- Virginia Housing Coalition

Other:

- National Change Database (NCDB)
- Community Commons Maps
- Crime Reports
- IPUMS-USA
- National Council on Real Estate Investment and Fiduciaries
- Location Inc (Neighborhoodscout)
- USDA Forest
- Maponics
- Center for Regional Analysis
- Urban Tree Canopy Analysis of Virginia Localities
- Urban Institute
- Yelp
- Walk Score
- RS Metrics
- AirBnB
- TripAdvisor
- InfoUSA Mailing List
- ARLnow
- Factual

Federal:

- Comprehensive Housing Affordability Strategy
- Home Mortgage Disclosure Act
- Panel Study of Income Dynamics (PSID)
- Uniform Crime Reporting Statistics

- Location Affordability Portal
- Longitudinal Employer Household Dynamics
- National Household Travel Survey
- U.S. Postal Service, Vacancies
- Federal Reserve Economic Data (FRED)
- Economic Census
- Construction Statistics
- Freddie Mac
- Fannie Mae, National Housing Survey (NHS)
- United Nations Statistics Division- Housing
- American Housing Survey
- Decennial Censuses of Housing
- Survey of Market Absorption of Apartments (SOMA)
- U.S. Bureau of Labor Statistics, Consumer Price Index (CPI)
- National Low-Income Housing Tax Credit (LIHTC) Database
- New Residential Construction
- Residential Energy Consumption Survey (RECS)
- Department of Commerce Economic Indicators
- Rental Housing Finance Survey (RHFS)
- Assisted Housing (HUD)
- Homelessness Data Exchange (HUD)
- Property Owners and Managers Survey (POMS)
- Manufactured Homes Survey (MHS)
- Survey of Residential Alterations and Repairs (SORAR)
- Survey of Consumer Finances (SCF)
- Residential Finance Survey (RFS)
- Housing Vacancy Survey (HVS)
- Federal Housing Finance Agency
- Community Reinvestment Act (CRA)

Attachments:

Housing Data List.xlsx (application/vnd.openxmlformats-officedocument.spreadsheetml.sheet)

VT Census Case Studies : Housing Data Profiling, Preparation Process, and Benchmarking

Created and last modified by molfe on 31 Dec 2015

The list of data that needs to be profiled and clean are listed below. Click the link to see more detail on where we are on the process.

** priorities

Data	Profiling	Cleaning	Transformation	Restructuring
BlackKnight **	Completed (Arlington County)	In progress (Arlington County)	In progress (Arlington County)	In progress (Arlington County)

Data	Profiling	Cleaning	Transformation	Restructuring
	Completed (James City County)	In progress (James City County)	In progress (James City County)	In progress (James City County)
CoreLogic **	Completed (Arlington County) Completed (James City County)	Completed (Arlington County)) Completed (James City County)	Completed (Arlington County) Completed (James City County)	Completed (Arlington County) Completed (James City County)
WMLS	Completed	Completed	Completed	Completed
MRIS-MLS	Completed (2010-2014) Completed (2009)	Completed (2010-2014) Completed (2009)	Completed (2010-2014)	Completed (2010-2014) Completed (2009)
Location, Inc.	Completed	Completed	Completed	Completed
Arlington County Real Estate Assessments	Completed	Completed	Completed	Completed
Arlington County Permits	In progress			
Arlington County Housing/ATRACK **	Completed			
Arlington County Crime	Completed	Completed	Completed	Completed
Arlington County GIS **	Completed	Completed	Completed	Not Needed
Arlington County Economic Development				
James City County GIS **	Completed	Completed	Completed	Not Needed

Data	Profiling	Cleaning	Transformation	Restructuring
James City County Parcel **	Completed	Completed	Completed	Completed
James City County Permits				
HMDA	Completed			
USDA Tree Canopy				

Areas that Need to be Addressed

This section lists the areas of the profiling or cleaning process that requires further attention (e.g. definition clarification) in order to proceed.

Addressed (Y/N)?	Issues
Y	General: Need to come up with common format of Addresses **(Will base location on lat/lon)**
Y	General: What year to base constant dollars? Should I also multiply by adjustment factor found in ACS data? **2013; no need to adjustment factor**
Y	WMLS: Need to get parcel IDs for units. This will allow us then make unique identifiers for housing units and then restructure data where appropriate. **Decided this step not needed (see profiling on what it would entail)**
Y	Real Estate Assessment: Figure out what provalLrsnId" means. **Proval land record serial number: identification for all properties Factor**
Y	Real Estate Assessment: Confirm what "Improvement.Value.Amount" means. **An "improvement' usually means a home/house/structure of some type.**
Y	Real Estate Assessment: Difference between "realEstatePropertyCode" and "masterRealEstatePropertyCode" **realEstatePropertyCode is an identification for all properties and masterRealEstatePropertyCode" is specific for condo buildings only its really used for GIS display.**
Y	Real Estate Assessment: Confirm what "reasPropertyStatusCode" means (Levels: A, T) **A: Active T: Inactive**

Y	Real Estate Dwelling: What to do with 66 properties that appear more than once but have different information. No time stamp. **Emailed both real estate and API people: All data in these tables is current. There are no duplicated Property-Dwelling properties**
Y	Real Estate Interior: What to do with 34,365 (32% of data) properties that appear more than once but have different information. No time stamp. **Emailed both real estate and API people: All data in these tables is current. There are no duplicated Property-Floor observations.**
Y	Real Estate Property: some property list as having 1 unit count are also listed as being in high right apartments (NumberofUnits AND Apartment highrise/midrise is invalid) **Made these unit counts NA**
Y	Real Estate: multiple dwelling properties... how to handle when restructuring? **Minimal effects on final estimates and doesn't effect all variables**
Y	Real Estate: Is time-stamped data available? **Yes, got the data from the department of Real Estate Assessment**
	JCC Parcel: What years are current and past assessment data from? Current year is 2015
Y	Real Estate: Check assessments that around low (~100) **Removed empty land**
Y	Real Estate: Why are some payments and levies missing? Set as 0 (non payment)? Mark those that have have some sort of reduction in price **Included deferred/adjusted/relief category which explains most of the difference**
Y	Real Estate: Remove certain types of sales? **See Description for which categories removed (those that are primarily non market value sales)**
Y	Real Estate: Why do 58 have mismatched years? **Multiple dwelling properties. Rename the Year Built Variable to Reflect this**
Y	Real Estate: 2013 unit counts are off (some greater than 2,000), which makes the weighted values in the millions. **Consistency check across the years shows 2013 errors. Emailed county. Switched systems which might have lead to error. Took previous year's value.**
Y	Real Estate: Why are there still assessments that are low (~1000)? **Parcels that later became inactive are still included so are some parking spots (that are not clearly marked) and new construction with low improvement value and 0 land value**

ACS Housing Tables and External Data

Arlington County

ACS Table		BKFS	CoreLogic	AC Real Estate	MRIS MLS
B25001 Housing Units	5 Year		X	X	
	1 Year	X	X	X	
B25003 Occupancy Status	5 Year		X		
	1 Year	X	X		
B25024 Units in Structure	5 Year		X	X	
	1 Year	X	X	X	
B25034 Year Structure Built	5 Year		X	X	
	1 Year	X	X	X	X
B25035 Median Year Structure Built	5 Year		X	X	X
	1 Year	X	X	X	X
B25041 Bedrooms	5 Year		X	X	X
	1 Year	X	X	X	X
B25075 Value	5 Year		X	X	X
	1 Year	X	X	X	X

ACS Table		BKFS	CoreLogic	AC Real Estate	MRIS MLS
B25077 Median Value (Dollars)	5 Year		X	X	X
	1 Year	X	X	X	X
B25102 Real Estate Taxes Paid	5 Year		X	X	X
	1 Year	X	X	X	X

James City County

ACS Table		BKFS	CoreLogic	JCC Parcel	WMLS
B25001 Housing Units	5 Year		X	X	
	1 Year		X	X	
B25003 Occupancy Status	5 Year		X		
	1 Year		X		
B25024 Units in Structure	5 Year		X		
	1 Year		X		
B25034 Year Structure Built	5 Year		X	X	X
	1 Year		X	X	X
B25035 Median Year Structure Built	5 Year		X	X	X
	1 Year		X	X	X

ACS Table		BKFS	CoreLogic	JCC Parcel	WMLS
B25041 Bedrooms	5 Year		X	X	X
	1 Year		X	X	X
B25075 Value	5 Year		X	X	X
	1 Year		X	X	X
B25077 Median Value (Dollars)	5 Year		X	X	X
	1 Year		X	X	X
B25102 real Estate Taxes Paid	5 Year		X		
	1 Year		X		

NOTE: Highlighted tables are those of primary focus.

ACS Table	Black Knight Variable Name (Field No.)	Core Logic	WMLS	MRIS-MLS	Location, Inc.	Arlington County Real Estate Assessments	Arlington County Permits	Arlington County Housing/ ATRACK	James City County Parcel
B25001 Housing Units		X				X			
B25002 Occupancy Status	Owner-Occupied (36)	X							

ACS Table	Black Knight Variable Name (Field No.)	Core Logic	WMLS	MRIS-MLS	Location, Inc.	Arlington County Real Estate Assessments	Arlington County Permits	Arlington County Housing/ ATRACK	James City County Parcel
B25024 Units In Structure	No of Units (81) No of Buildings (79)	X				X		X	X
B25034 Year Structure Built	Year Built (78) Effective Year Built (149)	X	X	X		X		X	X
B25035 Median Year Structure Built	X	X	X	X		X		X	X
B25036 Tenure By Year Structure Built	X	X						X	
B25037 Median Year Structure	X	X							

ACS Table	Black Knight Variable Name (Field No.)	Core Logic	WMLS	MRIS-MLS	Location, Inc.	Arlington County Real Estate Assessments	Arlington County Permits	Arlington County Housing/ ATRACK	James City County Parcel
Built By Tenure									
B25038 Tenure By Year Householder Moved Into Unit									
B25040 Housing Heating Fuel	Heating (104) Heating Fuel Type (150)	X		X		X			
B25041 Bedrooms	No of Bedrooms (83) Total # Rooms (82) Other Rooms (176)	X	X	X		X		X	X

ACS Table	Black Knight Variable Name (Field No.)	Core Logic	WMLS	MRIS-MLS	Location, Inc.	Arlington County Real Estate Assessments	Arlington County Permits	Arlington County Housing/ ATRACK	James City County Parcel
B25042 Tenure By Bedrooms	X	X						X	
B25047 Plumbing Facilities For All Housing Units	# of Plumbing Fixtures (134)	X	X	X		X			X
B25048 Plumbing Facilities For Occupied Housing Units									
B25049 Tenure By Plumbing Facilities	X	X							
B25051 Kitchen Facilities For All Housing Units									
B25063 Gross Rent									
B25064 Median								X	

ACS Table	Black Knight Variable Name (Field No.)	Core Logic	WMLS	MRIS-MLS	Location, Inc.	Arlington County Real Estate Assessments	Arlington County Permits	Arlington County Housing/ ATRACK	James City County Parcel
Gross Rent (Dollars)									
B25066 Aggregate Gross Rent (Dollars) By Units In Structure									
B25068 Bedrooms By Gross Rent									
B25075 Value	Total Assessed Value (39) = Assessed Land Value (37) + Assessed Improvement Value (38) Total Market	Calculated Assessments Market Appraisal	Final Selling Price	Original Listing Final Selling Price		Assessments Selling Price			Selling Price Assessments (no year)

ACS Table	Black Knight Variable Name (Field No.)	Core Logic	WMLS	MRIS-MLS	Location, Inc.	Arlington County Real Estate Assessments	Arlington County Permits	Arlington County Housing/ ATRACK	James City County Parcel
	Value (96) = Market Value: Land (94) + Market Value Improvement (95) Sale Price (52) Prior Sale Price (55) Tax Amount (44)								
B25077 Median Value (Dollars)	X	X	X	X		X			X

ACS Table	Black Knight Variable Name (Field No.)	Core Logic	WMLS	MRIS-MLS	Location, Inc.	Arlington County Real Estate Assessments	Arlington County Permits	Arlington County Housing/ ATRACK	James City County Parcel
B25081 Mortgage Status	Mortgage Lender Name? (153)								
B25082 Aggregate Value (Dollars) By Mortgage Status									
B25096 Mortgage Status By Value									
B25102 Mortgage Status By Real Estate Taxes Paid									
B25103 Mortgage Status By Median Real Estate Taxes Paid (Dollars)									

ACS Table	Black Knight Variable Name (Field No.)	Core Logic	WMLS	MRIS-MLS	Location, Inc.	Arlington County Real Estate Assessments	Arlington County Permits	Arlington County Housing/ ATRACK	James City County Parcel
B25107 Median Value By Year Structure	X	X	X	X		X			X
B25082 Aggregate Value (Dollars) by Units in Structure	X	X				X			
B25111 Median Gross Rent (Dollars) By Year Structure Built									X
B25114 Aggregate Gross Rent (Dollars) By Year Householder Moved Into Unit									
B25117 Tenure By House Heating Fuel	X	X							

ACS Table	Black Knight Variable Name (Field No.)	Core Logic	WMLS	MRIS-MLS	Location, Inc.	Arlington County Real Estate Assessments	Arlington County Permits	Arlington County Housing/ ATRACK	James City County Parcel
B25127 Tenure By Year Structure Built By Units In Structure	X	X						X	

VT Census Case Studies : Housing Data Profiling, Preparation Process, and Benchmarking

Created and last modified by molfe on 31 Dec 2015

The list of data that needs to be profiled and clean are listed below. Click the link to see more detail on where we are on the process.

** priorities

Data	Profiling	Cleaning	Transformation	Restructuring
BlackKnight **	Completed (Arlington County) Completed (James City County)	In progress (Arlington County) In progress (James City County)	In progress (Arlington County) In progress (James City County)	In progress (Arlington County) In progress (James City County)
CoreLogic **	Completed (Arlington County) Completed (James City County)	Completed (Arlington County)) Completed (James City County)	Completed (Arlington County) Completed (James City County)	Completed (Arlington County) Completed (James City County)
WMLS	Completed	Completed	Completed	Completed

Data	Profiling	Cleaning	Transformation	Restructuring
MRIS-MLS	Completed (2010-2014) Completed (2009)	Completed (2010-2014) Completed (2009)	Completed (2010-2014)	Completed (2010-2014) Completed (2009)
Location, Inc.	Completed	Completed	Completed	Completed
Arlington County Real Estate Assessments	Completed	Completed	Completed	Completed
Arlington County Permits	In progress			
Arlington County Housing/ATRACK **	Completed			
Arlington County Crime	Completed	Completed	Completed	Completed
Arlington County GIS **	Completed	Completed	Completed	Not Needed
Arlington County Economic Development				
James City County GIS **	Completed	Completed	Completed	Not Needed
James City County Parcel **	Completed	Completed	Completed	Completed
James City County Permits				
HMDA	Completed			
USDA Tree Canopy				

Areas that Need to be Addressed

This section lists the areas of the profiling or cleaning process that requires further attention (e.g. definition clarification) in order to proceed.

Addressed (Y/N)?	Issues
Y	General: Need to come up with common format of Addresses **(Will base location on lat/lon)**
Y	General: What year to base constant dollars? Should I also multiply by adjustment factor found in ACS data? **2013; no need to adjustment factor**
Y	WMLS: Need to get parcel IDs for units. This will allow us then make unique identifiers for housing units and then restructure data where appropriate. **Decided this step not needed (see profiling on what it would entail)**
Y	Real Estate Assessment: Figure out what provalLrsnId" means. **Proval land record serial number: identification for all properties Factor**
Y	Real Estate Assessment: Confirm what "Improvement.Value.Amount" means. **An "improvement' usually means a home/house/structure of some type.**
Y	Real Estate Assessment: Difference between "realEstatePropertyCode" and "masterRealEstatePropertyCode" **realEstatePropertyCode is an identification for all properties and masterRealEstatePropertyCode" is specific for condo buildings only its really used for GIS display.**
Y	Real Estate Assessment: Confirm what "reasPropertyStatusCode" means (Levels: A, T) **A: Active T: Inactive**
Y	Real Estate Dwelling: What to do with 66 properties that appear more than once but have different information. No time stamp. **Emailed both real estate and API people: All data in these tables is current. There are no duplicated Property-Dwelling properties**
Y	Real Estate Interior: What to do with 34,365 (32% of data) properties that appear more than once but have different information. No time stamp. **Emailed both real estate and API people: All data in these tables is current. There are no duplicated Property-Floor observations.**
Y	Real Estate Property: some property list as having 1 unit count are also listed as being in high right apartments (NumberofUnits AND Apartment highrise/midrise is invalid) **Made these unit counts NA**

Addressed (Y/N)?	Issues
Y	Real Estate: multiple dwelling properties... how to handle when restructuring? **Minimal effects on final estimates and doesn't effect all variables**
Y	Real Estate: Is time-stamped data available? **Yes, got the data from the department of Real Estate Assessment**
	JCC Parcel: What years are current and past assessment data from? Current year is 2015
Y	Real Estate: Check assessments that around low (~100) **Removed empty land**
Y	Real Estate: Why are some payments and levies missing? Set as 0 (non payment)? Mark those that have have some sort of reduction in price **Included deferred/adjusted/relief category which explains most of the difference**
Y	Real Estate: Remove certain types of sales? **See Description for which categories removed (those that are primarily non market value sales)**
Y	Real Estate: Why do 58 have mismatched years? **Multiple dwelling properties. Rename the Year Built Variable to Reflect this**
Y	Real Estate: 2013 unit counts are off (some greater than 2,000), which makes the weighted values in the millions. **Consistency check across the years shows 2013 errors. Emailed county. Switched systems which might have lead to error. Took previous year's value.**
Y	Real Estate: Why are there still assessments that are low (~1000)? **Parcels that later became inactive are still included so are some parking spots (that are not clearly marked) and new construction with low improvement value and 0 land value**

ACS Housing Tables and External Data

Arlington County

ACS Table		BKFS	CoreLogic	AC Real Estate	MRIS MLS
B25001 Housing Units	5 Year		X	X	
	1 Year	X	X	X	
B25003 Occupancy Status	5 Year		X		
	1 Year	X	X		
B25024 Units in Structure	5 Year		X	X	
	1 Year	X	X	X	
B25034 Year Structure Built	5 Year		X	X	
	1 Year	X	X	X	X
B25035 Median Year Structure Built	5 Year		X	X	X
	1 Year	X	X	X	X
B25041 Bedrooms	5 Year		X	X	X
	1 Year	X	X	X	X
B25075 Value	5 Year		X	X	X
	1 Year	X	X	X	X

ACS Table		BKFS	CoreLogic	AC Real Estate	MRIS MLS
B25077 Median Value (Dollars)	5 Year		X	X	X
	1 Year	X	X	X	X
B25102 Real Estate Taxes Paid	5 Year		X	X	X
	1 Year	X	X	X	X

James City County

ACS Table		BKFS	CoreLogic	JCC Parcel	WMLS
B25001 Housing Units	5 Year		X	X	
	1 Year		X	X	
B25003 Occupancy Status	5 Year		X		
	1 Year		X		
B25024 Units in Structure	5 Year		X		
	1 Year		X		
B25034 Year Structure Built	5 Year		X	X	X
	1 Year		X	X	X
B25035 Median Year Structure Built	5 Year		X	X	X
	1 Year		X	X	X

B25041 Bedrooms	**5 Year**		X	X	X
	1 Year		X	X	X
B25075 Value	**5 Year**		X	X	X
	1 Year		X	X	X
B25077 Median Value (Dollars)	**5 Year**		X	X	X
	1 Year		X	X	X
B25102 real Estate Taxes Paid	**5 Year**		X		
	1 Year		X		

NOTE: Highlighted tables are those of primary focus.

ACS Table	Black Knight Variable Name (Field No.)	Core Logic	WMLS	MRIS-MLS	Location, Inc.	Arlington County Real Estate Assessments	Arlington County Permits	Arlington County Housing/ ATRACK	James City County Parcel
B25001 Housing Units		X				X			
B25002 Occupancy Status	Owner-Occupied (36)	X							

B25024 Units In Structure	No of Units (81) No of Buildings (79)	X				X		X	X
B25034 Year Structure Built	Year Built (78) Effective Year Built (149)	X	X	X		X		X	X
B25035 Median Year Structure Built	X	X	X	X		X		X	X
B25036 Tenure By Year Structure Built	X	X						X	
B25037 Median Year Structure Built By Tenure	X	X							
B25038 Tenure By Year Householder Moved Into Unit									
B25040 Housing	Heating (104)	X		X		X			

Heating Fuel	Heating Fuel Type (150)								
B25041 Bedrooms	No of Bedrooms (83) Total # Rooms (82) Other Rooms (176)	X	X	X		X		X	X
B25042 Tenure By Bedrooms	X	X						X	
B25047 Plumbing Facilities For All Housing Units	# of Plumbing Fixtures (134)	X	X	X		X			X
B25048 Plumbing Facilities For Occupied Housing Units									
B25049 Tenure By Plumbing Facilities	X	X							
B25051 Kitchen									

Facilities For All Housing Units									
B25063 Gross Rent									
B25064 Median Gross Rent (Dollars)								X	
B25066 Aggregate Gross Rent (Dollars) By Units In Structure									
B25068 Bedrooms By Gross Rent									
B25075 Value	Total Assessed Value (39) = Assessed Land Value (37) + Assessed Improvement Value (38) Total Market Value	Calculated Assessments Market Appraisal	Final Selling Price	Original Listing Final Selling Price		Assessments Selling Price			Selling Price Assessments (no year)

	(96) = Market Value: Land (94) + Market Value Improvement (95) Sale Price (52) Prior Sale Price (55) Tax Amount (44)								
B25077 Median Value (Dollars)	X	X	X	X		X			X
B25081 Mortgage Status	Mortgage Lender Name? (153)								
B25082 Aggregate Value (Dollars) By									

Mortgage Status									
B25096 Mortgage Status By Value									
B25102 Mortgage Status By Real Estate Taxes Paid									
B25103 Mortgage Status By Median Real Estate Taxes Paid (Dollars)									
B25107 Median Value By Year Structure	X	X	X	X		X			X
B25082 Aggregate Value (Dollars) by Units in Structure	X	X				X			
B25111 Median Gross Rent (Dollars) By Year Structure Built									X
B25114 Aggregate Gross Rent (Dollars)									

By Year Householder Moved Into Unit									
B25117 Tenure By House Heating Fuel	X	X							
B25127 Tenure By Year Structure Built By Units In Structure	X	X						X	

VT Census Case Studies: Education Data Inventory

Created by molfe, last modified by ziemer on 29 Dec 2015

This is a list of education related data sources that have been inventoried. Clicking each source will take you to a separate site that includes information on that source's triage or triage and data inventory.

State Level Data

- Alabama K-12 Education
- Alaska K-12 Education
- Arizona K-12 Education
- Arkansas Longitudinal Education Data
- California K-12 Education
- Colorado K-12 Education
- Connecticut K-12 Education
- DC Longitudinal Education Data
- Delaware K-12 Education
- Florida Longitudinal Education Data
- Georgia K-12 Education
- Hawaii K-12 Education
- Idaho Longitudinal Education Data
- Illinois K-12 Education and Workforce
- Illinois Higher Education
- Indiana K-12 Education
- Indiana Higher Education/Workforce
- Iowa K-12 Education
- Kansas K-12 Education
- Kentucky Longitudinal Data System
- Louisiana K-12 Education
- Maine K-12 Education

- Maryland Longitudinal Education Data
- Massachusetts Longitudinal Education Data
- Michigan Longitudinal Education Data
- Minnesota Longitudinal Education Data
- Mississippi K-12 Education
- Missouri K-12 Education
- Montana K-12 Education
- Nebraska K-12 Education
- Nevada K-12 Education
- New Hampshire Longitudinal Education Data
- New Jersey K-12 Education
- New Mexico K-12 Education
- New York Longitudinal Education Data
- North Carolina K-12 Education
- North Carolina Higher Education/Workforce
- North Dakota K-12 Education
- Ohio Longitudinal Education Data
- Oklahoma K-12 Education
- Oklahoma Higher Education/Workforce
- Oregon K-12 Education
- Pennsylvania Longitudinal Education Data
- Pennsylvania Workforce
- Rhode Island Longitudinal Education Data
- South Carolina K-12 Education
- South Carolina Higher Education
- South Dakota K-12 Education
- Tennessee K-12 Education
- Texas K-12 Education
- Texas Higher Education
- Utah Longitudinal Education Data
- Vermont K-12 Education Data
- Virginia Longitudinal Education Data
- Washington Longitudinal Education Data
- West Virginia K-12 Education
- West Virginia Higher Education
- West Virginia Workforce
- Wisconsin K-12 Education
- Wyoming K-12 Education

Commercial Data

- College Board
- Donors Choose
- eSparks
- Glassdoor
- Great School Ratings
- LinkedIn
- Location Inc (Neighborhoodscout)
- Maponics - School Boundaries
- Monster Resume Database
- National Student Clearinghouse
- School Attendance Boundary Information System (SABINS)
- School Digger

- US News and World Report Rankings

Federal Data

- Bureau of Labor Statistics - Business Employment Dynamics
- Bureau of Labor Statistics - Current Population Survey
- Bureau of Labor Statistics - Employment Projections
- Bureau of Labor Statistics - Job Openings and Labor Turnover Survey (JOLTS)
- Bureau of Labor Statistics - Occupational Employment Statistics
- Bureau of Labor Statistics - Quarterly Census of Employment and Wages (QCEW)
- Bureau of Labor Statistics - State and Metro Area Employment
- Bureau of Labor Statistics - State and Local Area Unemployment Statistics
- Department of Education - Academic Library Survey
- Department of Education - Adult Literacy and Lifeskills Survey
- Department of Education - Common Core of Data
- Department of Education - Consolidated State Performance Report
- Department of Education - EDFacts
- Department of Education - Fast Response Survey System
- Department of Education - Federal Student Loan Program
- Department of Education - Free Application for Federal Student Aid (FAFSA)
- Department of Education - High School and Beyond Survey
- Department of Education - National Assessment of Educational Progress
- Department of Education - National Household Education Surveys Program
- Department of Education - National Postsecondary Student Aid Study
- Department of Education - National Study of Postsecondary Faculty
- Department of Education - Private School Universe Survey
- National Longitudinal Surveys
- OECD - Programme for International Student Assessment (PISA)
- OECD - Teaching and Learning International Survey (TALIS)
- Quarterly Wage Indicators (QWI) Explorer

VT Census Case Studies : Education Data Profiling, Preparation Process, and Benchmarking

Created and last modified by ziemer on 30 Dec 2015

The list of data that needs to be profiled and clean are listed below. Click the state link in the table to see more detail on where we are on the process.

Data	Profiling	Cleaning	Transformation	Restructuring	Univariate Descriptives	Direct Estimates/Benchmarks
North Carolina	North Carolina - 1. Data Profiling	North Carolina - 2. Data Cleaning	North Carolina - 3. Data Transformation	NA	North Carolina - 4. Univariate Descriptions	North Carolina - 5. Benchmark ACS Estimates
Texas	Texas - 1. Data Profiling	Texas - 2. Data Cleaning	Texas - 3. Data Transformation	NA	Texas - 4. Univariate Descriptions	Texas - 5. Benchmark ACS Estimates
Virginia	Virginia K-12 - 1. Data Profiling Virginia Higher Ed - 1. Data Profiling	NA	NA	Virginia K-12 - 2. Data Restructuring	Virginia K-12 - 3. Univariate Descriptions	Virginia K-12 - 4. Benchmark ACS Estimates
Kentucky	Kentucky - 1. Data Profiling	Kentucky - 2. Data Cleaning	Kentucky - 3. Data Transformation	NA		Kentucky - 4. Benchmark ACS Estimates

Data	Profiling	Cleaning	Transformation	Restructuring	Univariate Descriptives	Direct Estimates/Benchmarks
Washington	Washington - 1. Data Profiling					

Definitions

Benchmark Estimate: Estimate from an ACS table, or from a tabulation of the ACS PUMS.

Direct Estimate: Estimate from an alternate source that tries to replicate an ACS estimate.

Issues in Benchmarking

Issues that Affect the Estimates:

Masked data due to FERPA: Some of the state data (e.g., Texas) required cell counts under a certain number (e.g., 10) to be masked due to FERPA laws. Therefore, we did not receive comprehensive data from Texas and Virginia K-12. Because the counts were much lower for Texas (i.e., 1/5th of the total counts), we used weights in order to approximate student enrollment counts.

County/District Areas: In Texas, there are many cases where county and school district boundaries do not align. When a school district overlaps two or more counties, the county ID assigned to that school district is based on the geographical location of the district office. Therefore the county estimates may not be as accurate as the district estimates when comparing to ACS.

Issues that Could be Addressed with Future Research

Public vs. Private schools: The ACS includes both public and private schools, whereas the state K-12 data include only public schools. Some ACS tables (e.g., B14002) break down counts for public and private schools, however, most do not.

PUMA vs. County/District Areas: PUMS data could be used to identify public versus private school enrollment, however, this information would be geographically aligned with the PUMA areas. The state data is geographically aligned with school district and county areas rather than PUMA areas. PUMA areas often to not directly correspond to county or school district boundaries. One county/district can equal multiple PUMAs, multiple counties/districts can equal one PUMA, and multiple PUMAs can equal multiple counties/districts.

Charter schools: The state data includes charter schools, however, these schools are not geographically aligned with the ACS school district boundaries. The ACS uses Local Education Agency (LEA) numbers to identify school districts and charter schools do not have LEA numbers. Therefore, charter schools were not included in benchmarking.

School to work transitions: Data on workforce from the states did not become available in time. However, this could be used in future research.

Areas that Need to be Addressed

This section lists the areas of the profiling or cleaning process that requires further attention (e.g. definition clarification) in order to proceed.

Addressed (Y/N)?	Issues
Y	Why are there 5 more District IDs than District Names for Texas (this is consistent across all 5 years)? **Resolution: The District IDs are unique whereas the District Names are not. There are different school districts with the same name**
Y	Why do the number of Districts change so much from year to year (788-851)? **Resolution: it's due to suppressed information that is not available due to FERPA (records are removed when there are counts less than 5)**
Y	The ACS 2013 5 year estimates 1022 unified school districts for Texas. This is several hundred greater than the number of districts in the Texas data. **Resolution: it's due to suppressed information that is not available due to FERPA (records are removed when there are counts less than 5)**
Y	What does EE means for Grade in Texas? **Resolution: Early Education**
Y	For student-level, should duplication only be assessed for IDs or should it be based on an entire dataset (e.g., duplicates on all variables except for IDs)? **Resolution: Duplication will only assessed for IDs**
Y	In Texas, only approximately 30% of the students can be linked from year to year. **Resolution (response from TX): This is likely due to the FERPA removal process where students were removed separately for each year. Therefore, the same student could be in the 2009 dataset, not in the 2010 dataset**
Y	How to differentiate between blank rows and missing? Exit Reasons that are blank indicate that a student is still enrolled and none are actually coded as "missing". **Resolution: Anyone blank row is considered "still enrolled".**
Y	School districts change every year (e.g., some are combined, some become obsolete). Texas school district directories are available for 2009-2013 only in pdf or web format. Need to scrape this information in order to compare to the datasets for Validity Step. Have already compared District IDs and District Names against the District codebooks provided by the Texas Education Agency (which list the school districts, district IDs, and matching counties and county IDs). **Resolution: Only compare the year-by-year District codebooks against the dataset.**

Addressed (Y/N)?	Issues
Y	North Carolina's "school code" variable isn't a unique value for each school. That is school number, but the mbuild dataset doesn't include schools. **Resolution: Since the focus is at the district level rather than the school level, the school code variable does not need to be cleaned for now.**
Y	For some reason the North Carolina mbuild dataset, which is suppose to include every person in North Carolina schools, does not have anyone listed in Pre-K, Kindergarten and First Grade. It also says only three kids are in second grade statewide. **Resolution: The MBuild datasets only include students in grades 3-12.**
Y	For North Carolina data, some people are listed as being in grade 13 even though the codebook says K-12 are the only valid values. Is grade 13 representing a repeated senior year? Adult education? **Resolution (response from NC): Grade 13 is a 5th year of high school and I believe most if not all of these are Early College High Schools.**
Y	What do 21, NU, and SS mean for grades? **Resolution (response from NC): 21 is likely a data entry error and NU is NULL.**
Y	What do 1 and U mean in terms of Limited English Proficiency for North Carolina Students? **Resolution (response from NC): 1=1st year exempt; U=Exited LEP**
Y	Why are almost 50% of students in North Carolina listed as economically disadvantaged? Is there a different threshold for free lunch in NC different than food stamps? **Resolution: This is likely correct as the North Carolina Department of Public Instruction lists the percentage of students enrolled in free and reduced lunch ranging from 48 - 56%. http://datacenter.kidscount.org/data/tables/2239-percent-of-students-enrolled-in-free-and-reduced-lunch?loc=35#detailed/2/any/false/1021,909,857,105,118/any/4682**
Y	Several LEA codes are not in the code book. **Resolution (response from NC): LEA 209 is Cherokee Central School, 997 is NC Health and Human Services, and 998 is NC Department of Juvenile Justice (http://www.ncpublicschools.org/docs/fbs/resources/leacharterlist.pdf).**
Y	Why does Mbuild 2013 not have a year variable? **Resolution (response from NC): Variables change from year to year. In the case of the year variable, you can add one as all of those records are for 2012-13.**
Y	For NC, why isn't there a public school universe file for 2013 that would allow us to validate the LEAs? **Resolution: variable not being used.**

Addressed (Y/N)?	Issues
Y	When there are people in the NC datasets listed as in 2nd Grade is this a mistake or do these represent second graders who are taking third grade tests? **Resolution (response from NC): while there are 2nd graders in the masterbuild, they should not have test scores.**
Y	How do we address duplicates when they represent students who move to different school districts during the school year? **Resolution: The first entry for a student in the dataset should be retained and the duplicates should be removed.**
Y	How should Race/Ethnicity be cleaned when there are changes in the codes across years? (i.e., what should the final race code be?) **Resolution: The most recent race/ethnicity should be used for all years.**
Y	Should Economic Status be cleaned to be consistent across years? **Resolution: No, this may be inconsistent across years since students' economic status may change.**
Y	For NC, in 2011 Course Membership Ethnicity column what is N and why does it represent 15% of the population (also appears in 2012 and 2013)? **Resolution (response from NC): We recommend you use ethnicity and grade from the demographics file rather than course membership.**
Y	For NC, in 2013 Course Membership Ethnicity column do 3 and 5 represent a mistaken use of the pre-2011 codes? **Resolution (response from NC): We recommend you use ethnicity and grade from the demographics file rather than course membership.**
Y	For NC, in 2013 Course Membership grade column what does SS mean? **Resolution: variable not being used.**
Y	Should grade be modified in the cleaned dataset if there are inconsistencies? **Resolution: Not at this time. Inconsistencies were flagged - see data cleaning page.**
Y	Should we create weights for the Texas Administrative data? **Resolution: Use Texas aggregated data to create weights.**
Y	For NC, why are there so many duplicates in the demographics dataset? Do many students take tests both in the fall and spring? **Resolution (response from NC): Yes, demographic information is collected on students in both the fall and the spring.**
Y	For NC, does a NULL in diploma type just mean that the student is not yet in high school, and therefore does not have a diploma type? **Resolution: Diploma**

Addressed (Y/N)?	Issues
	type does not indicate graduation (it's a track within HS coursework) therefore it does not need to be profiled at this point.
Y	For NC, is there any documentation on how accurate the Pre-K - 2 data collection is (i.e. which schools reported them in the demographics tables and which didn't)? Would Course Membership be a better way to measure Pre-K - 2? Main concern is for 2009, 2010 looks accurate, will update as more years are profiled. The demographics codebook states, "Note that some students in grades PK through 2 are included in this file, but the inclusion of these students is not required and therefore may not reflect all students in these grades in North Carolina public schools." **Resolution (response from NC): In recent years, inclusion of PK-2 students appears to be complete. You can compare these counts to other grades to get a sense of how comprehensive each file/year might be.**
Y	For NC, why do some schools (typically charter schools) appear in public school universe but never have any students associated with them? **Resolution: Variable not being used.**
Y	For NC, what does FRC with no number following it mean in Demographics 2010 Diploma Type? **Resolution: Diploma type does not indicate graduation (it's a track within HS coursework) therefore it does not need to be profiled at this point.**
	What withdrawal codes does the demographics file use? Are they the same as dropout?
Y	What is the CERT diploma type? **Resolution: Diploma type does not indicate graduation (it's a track within HS coursework) therefore it does not need to be profiled at this point.**
Y	For North Carolina, when were demographics collected during the spring semester for the Demographics table? **(response from NC): They were collected on the first day of spring testing which is during the End of Grade testing which takes place within the last 3 weeks of school.**
Y	In the ACS, there are several Elementary School Districts that are not in the Unified School Districts for Virginia (1) and Texas (~7). Should these be viewed as separate tables when doing direct estimates or combined? **Resolution: These should be combined into one table with the unified school districts.**
Y	Charter schools do not have LEAs. How should we account for them when comparing SLDS data to ACS estimates at the school district level? (note. ACS

Addressed (Y/N)?	Issues
	school districts are based on where people live, not necessarily where they go to school. Also, ACS does not ask about charter schools - they only differentiate public and private schools). **Resolution: We will leave out the charter schools for now and see how much of a difference there are in the benchmark estimates.**
Y	Only one ACS table (B14002) differentiates between public and private school. We only have public school data from the states, therefore the counts will not be accurate comparisons to the ACS. What is the best way to account for this? **Resolution: We can use the PUMS data which differentiates the public and private schools, however, the PUMS areas do not always align with county or school district boundaries which would make it difficult to match to the Texas data. Another alternative is to use private school enrollment estimates from the Department of Education.**
Y	For KY, why are birth years off (e.g., 1900-1950 range for the 2009-2014 K-12 file) in the master demographics dataset? **(response from KY): Data entry errors.**
Y	For KY, why does the preschool dataset have grade variables in the KG-9th grade range? **(response from KY): Data entry errors.**
Y	For KY, in the KPEDS Cohort file, what do the codes GRS and FTU stand for in the Cohort variable? **(response from KY):** **FTU = First-time undergraduate = An undergraduate student who has not previously attended any postsecondary institution or who attended postsecondary level courses as a high school student and is currently enrolled for the first time since high school graduation. Includes students who were first-time in the summer semester and returned in the fall.** **GRS = For the fall enrollment report only, indicates 'Y' if the student is full-time and eligible to be included in the Integrated Postsecondary Education Data System (IPEDS) Graduation Rate Survey (GRS) Cohort, P if the student meets the criteria and is enrolled part-time, or N if the student is not eligible. The following criteria determine whether an undergraduate student is included in the cohort:** **Full-time First-time: Include students who attended college for the first time (either part-time or full-time) in the prior summer term whether at the same college, another college in Kentucky, in another state, or another country. Summer semester will be counted whether the student enrolled as degree-seeking or not. Also include students who have entered with advanced standing (college credits earned before graduation from high school).**

Addressed (Y/N)?	Issues
	Degree/certificate seeking: Students must be enrolled in courses creditable toward a degree, diploma, certificate, or other formal award. Include students in occupational and vocational programs. Students at four-year institutions whose intent is not known are to be reported in the bachelor's cohort.
Y	For KY, In the KPEDS Readiness file, how is college readiness (Y/N) calculated? **(response from KY): Here are Kentucky's college readiness indicators:** http://cpe.ky.gov/NR/rdonlyres/78B3510A-CECD-4157-8F20-3E3499707DAA/0/CollegeReadinessIndicators.pdf
Y	For KY, In the TEDS Enrollment file, in the student objective variable, what is the difference between prepatory and exploring? **(response from KY):** **From this publication: http://education.ky.gov/CTE/teds/Documents/TEDS_User_Guide.pdf** **Enrollment Status: Identifies the enrollment status of each student in a CTE program/pathway. Students can be identified as EXPLORING (secondary or post-secondary) or PREPARATORY (secondary or post-secondary).** **POST-SECONDARY: Information must be entered for any student who has been enrolled in a technical education course(s) long enough to be required to pay tuition. Enter "exploring" in the student objective field for any student who has 12 or fewer credits in the program/CIP code. Once the student has completed 12 credits in their program AND is enrolled for the next course (resulting in 13 or more credits in the program), change the status to "preparatory".**
Y	For KY, In the KPEDS Financial Aid file, how is the Total Income variable calculated? Is it based on family income? student and parent income? **(response from KY): This indicates family income and reflects the most updated amount that was used for calculating student aid during the academic year.**
Y	For KY, how and when are data collected? **(response from KY): The K-12 Person Enrollment records are submitted to KDE by the individual districts -- basically in real-time. Kentucky utilizes a statewide software called Infinite Campus and the individual school districts are responsible for creating new Person Enrollment records whenever they have a new enrollment. About six months after the end of a school year (Oct/Nov),**

Addressed (Y/N)?	Issues
	KDE takes all of these enrollments and sends them over to us so that we can add them to the KLDS. The K-12 Annual Person records are created from the individual Person Enrollment records at the end of the school year and are based on each student's "end status".
Y	For KY, how are dropouts classified and is this code mutually exclusive from other codes (e.g., withdraw, move, etc.)? **(response from KY): IC_PersonEnrollments are created by the individual school districts using a statewide reporting software called Infinite Campus. The individual districts then send their enrollment data to KDE. Students get a new person enrollment record every time they change schools during the year, so an individual student could have 8 or 10 IC_PersonEnrollment records in the same school year if they were especially transient. Schools/districts are supposed to use the dropout code when there has been a confirmed dropout, but we suspect that some districts report dropouts as "transferred to home school" or other codes in order to avoid reporting high dropout counts.**
Y	For KY, what would be the best indicator to use for dropouts? **(response from KY): An easier way to calculate dropouts might be to use the IC_Dropout_Code/IC_Dropout_Reason in the IC_AnnualPerson table. The annual person records are created from the person enrollment records and are based on the final status of each student at the end of the school year.**
Y	For KY, why are the 2011-2014 dropout codes all NULL, whereas there are Dropout Reasons for these years? **(response from KY): KDE used to supply both a dropout code and dropout reason, but stopped doing that after 2010.**
Y	For KY, what are the corresponding Dropout Reasons for the Dropout Code? **(response from KY):** **IC_Dropout_Code IC_Dropout_Reason** **01 Boredom** **02 Course Selection** **03 Employment** **04 Expulsion** **05 Failing Classes** **06 Family Problems** **07 Illness** **08 Marriage** **09 Pregnancy** **10 Student/Teacher Conflict** **99 No Substantiated Enrollment**

Addressed (Y/N)?	Issues
Y	For KY, for the IC_AnnualPerson, is the school district for each student the last school district they attended that year? **(response from KY): Usually, but not necessarily. It is possible for a student to have a different school/district listed in their annual person record than the school/district at which they finished the school year if they attended 100 or more days at another school/district during the year, but then finished the year at another school. In that case, the school/district at which they attended more than 100 days would be listed.**
Y	For KY, why are there 511 NULL for grade for 2014 for IC_AnnualPerson? **(response from KY): "Crappy data from KDE"**
Y	For KY, for the IC_Annual Person why is there a separate grade category for over 17yrs old by Oct 1st? Are these students also included in the 12th Grade category? **(response from KY): I believe that category is typically reserved for students with an IEP on file (special education), which allows them extra time to graduate and still be included in their original cohort graduation rates.**
Y	For KY, for the IC_AnnualPerson are the 1-4 year olds included because they're enrolled in Pre-K? **(response from KY): Yes**
	For KY, for the IC_AnnualPerson, why are there missing District Names for 2011-2014?
Y	For KY, why are there slightly different counts in IC_AnnualPerson between "Dropout Code" and "Dropout Reason" for Boredom, Failing Classes, Family Problems, No Substantial Enrollment, Student/Teacher Conflict, and NULL? Which one is more accurate? **(response from KY): I think you're okay to use whichever field is populated. Districts are supposed to enter both a dropout code and dropout reason, but we've seen that that's not always the case.**
Y	For KY, why are there duplicates in the IC_AnnualPerson? Some seem to be because they switched school districts, some seem to be due to errors in the data entries (e.g., one entry has inaccurate information), and some seem to be duplicates on all other variables except for # of days enrolled, unexcused days, etc. **(response from KY): The cause of this could be crappy data from KDE or matching issues when we import that data into our system.**

Documentation of Assumptions

This section lists the assumptions that we follow for different data sources (e.g., The ACS defines Hispanic/Not Hispanic as ethnicity which is separate from race).

Assumption	Documentation	Resolution
The Texas data is documented as "2009" for the 2008-2009 school year. The most recent file is the 2012-2013 school year. Therefore we will be comparing the 2012-2013 school year with 2013 ACS data.		
Texas county codes	http://comptroller.texas.gov/taxinfo/tx_county_codes.html	
Texas county and school district match	http://mansfield.tea.state.tx.us/TEA.AskTED.Web/Forms/DownloadFile.aspx	
We are getting a representative sample of students from each state	The digest of educational statistics is really useful for getting state-level enrollment counts: https://nces.ed.gov/programs/digest/d13/tables/dt13_203.20.asp	This table can be used to compare to the state data we're receiving.
For North Carolina,		

don't include 2009 Pre-K through 2nd grade estimates in the 5 year estimates because reporting was optional for these grades in this year and estimates are below that of the other years as a result.		
The higher education institutions in Virginia include both private and public.		
NCES LEA IDs are based on official documentation	https://www2.ed.gov/admins/lead/account/consolidated/sy09-10part1/kys.pdf	

ACS Education Tables and External Data

ACS Table	Texas	North Carolina	Virginia	Kentucky	Washington
B14002: SEX BY SCHOOL ENROLLMENT BY LEVEL OF SCHOOL BY TYPE OF SCHOOL FOR THE POPULATION 3 YEARS AND OVER	X	X	X	X	
B14003: SEX BY SCHOOL ENROLLMENT BY TYPE OF SCHOOL BY AGE FOR THE POPULATION 3 YEARS AND OVER		X	X (for higher ed only)	X	
B14004: SEX BY COLLEGE OR GRADUATE SCHOOL ENROLLMENT BY TYPE OF SCHOOL BY AGE FOR THE POPULATION 15 YEARS AND OVER			X (for higher ed only)	X (for higher ed only)	
B14007B: SCHOOL ENROLLMENT BY DETAILED LEVEL OF SCHOOL FOR THE POPULATION 3 YEARS AND OVER (BLACK ALONE)	X	X	X	X	
B14007C: SCHOOL ENROLLMENT BY DETAILED LEVEL OF SCHOOL FOR THE POPULATION 3 YEARS AND OVER (AMERICAN INDIAN AND ALASKA NATIVE ALONE)	X	X	X	X	

ACS Table	Texas	North Carolina	Virginia	Kentucky	Washington
B14007D: SCHOOL ENROLLMENT BY DETAILED LEVEL OF SCHOOL FOR THE POPULATION 3 YEARS AND OVER (ASIAN ALONE)	X	X	X	X	
B14007E: SCHOOL ENROLLMENT BY DETAILED LEVEL OF SCHOOL FOR THE POPULATION 3 YEARS AND OVER (NATIVE HAWAIIAN AND PACIFIC ISLANDER ALONE)	X	X	X	X	
B14007G: SCHOOL ENROLLMENT BY DETAILED LEVEL OF SCHOOL FOR THE POPULATION 3 YEARS AND OVER (TWO OR MORE RACES)	X	X	X	X	
B14007H: SCHOOL ENROLLMENT BY DETAILED LEVEL OF SCHOOL FOR THE POPULATION 3 YEARS AND OVER (WHITE ALONE, NOT HISPANIC OR LATINO)	X	X	X	X	
B14007I: SCHOOL ENROLLMENT BY DETAILED LEVEL OF SCHOOL FOR THE POPULATION 3 YEARS AND OVER (HISPANIC OR LATINO)	X	X	X	X	

ACS Table	Texas	North Carolina	Virginia	Kentucky	Washington
B15001: SEX BY AGE BY EDUCATIONAL ATTAINMENT FOR THE POPULATION 18 YEARS AND OVER ***Note. Only for the age category of 18-24.**			X (for higher ed only)	X (for higher ed only)	
B15011: SEX BY AGE BY FIELD OF BACHELOR'S DEGREE FOR FIRST MAJOR FOR THE POPULATION 25 YEARS AND OVER ***Note. Only for the age category of 25-39.**			X (for higher ed only)	X (for higher ed only)	

Benchmarking Rules

Ratio Plots

- Put the number of districts/counties represented in the plots (this can be found in the County and District tabs of the excel spreadsheet).
- Y-Axis range
  - County and District 1-year estimates: -7.5 to 7.5
  - County and District 5-year estimates: -15 to 15
  - **Exception:** B14002, county 1-year estimate: -10 to 7.5
- Saving the plots: export using 1000 for width and 1000 for length
- Outliers not included in this range should go in a table. For B14007, can put all race outliers in one table. See Benchmarking Examples - Virginia (bottom of page).