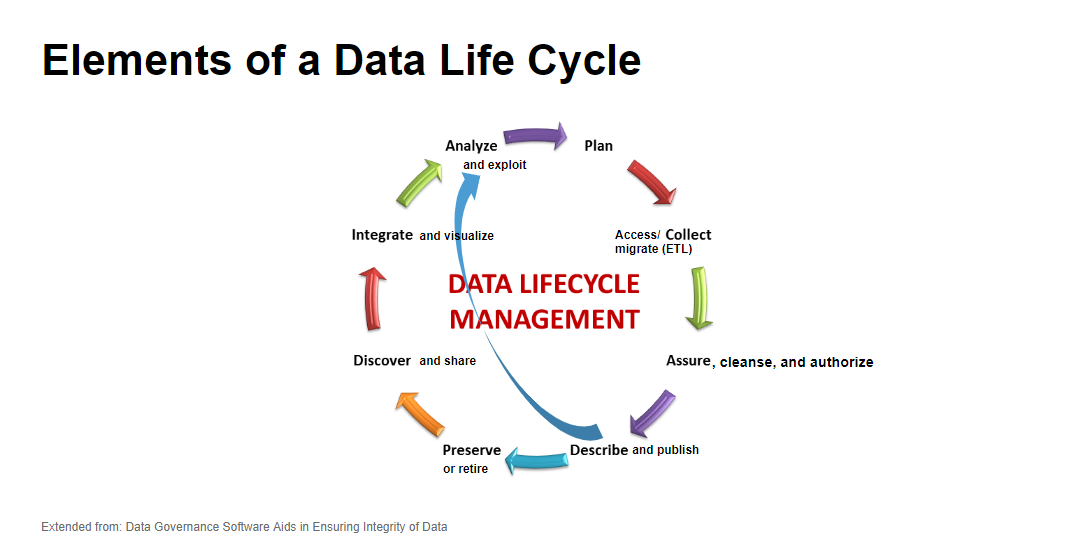
**Introduction to Applied Data Science**

**Homework 1**

**Provide answers to the following questions as a PDF file or Word document file.**

**Q1. The website at:** [**https://www2.census.gov/cac/sac/meetings/2015-11/2020-census-lifecycle-chart.pdf**](https://www2.census.gov/cac/sac/meetings/2015-11/2020-census-lifecycle-chart.pdf) **provides the lifecycle for operationalizing the US census data acquisition. Look at the various aspects of the census lifecycle and relate them to the various steps in the data lifecycle that we saw in class.**

**As your answer give an overlaid lifecycle figure with your insight about the census lifecycle on top of the data lifecycle.**

****

|  |  |
| --- | --- |
| Plan | * Begin Formal Planning for 2020 Census (**11/11**) * Develop High-Level Conceptual Design (**01/13**) * Deliver Initial 2020 Project Requirements (**12/14**) * Preliminary Operational Plan with Cost Estimates (**10/15**) * Begin Advance Notice Mail Package for LUCA Program (01/17) * Local Update of Census Addresses (LUCA) * Finalize Field Office Scope/Locations (**03/16–08/17**) * Finalize Staffing Structure (**09/17**) * Deliver 2020 Operational Plan (**09/18**) * Begin Recruiting for Address Canvassing (**06/19**) |
| Collect | * Census Day (**2015** Census and Optimizing Self-Response Tests) * Begin in-Office Canvasing (**9/15**) * Begin Initial Basic Collection Unit (BCUs) Delineation (**11/15**) * Conduct Address Canvassing (**08/19–12/19**) * Internet Self-Response Operation (**04/20**) * Non-Response Follow-up Operation (**04/20**) * Group Quarters Enumeration (**02/20**) * Conduct Response Processing Operation (Data Collection, **01/20–09/20**) * In-Field Data Collection (**Remote Alaska, Enumeration**) |
| Assure, Cleanse | * Administrative Records Usage, LUCA Program Improvements (***ongoing***) * Begin Address Validation Test (AVT) (**09/14**) * Resolve LUCAs OMB Clearance (**12/16**) * Finalize Residence Rules (**12/17**) * **2016–2018** Census Day Tests (Data Collection Validation) -> in Test and Evaluation pg. 2 * Complete LUCA Program (**09/21**) |
| Describe, Publish | * Deliver 2020 Census Topics to Congress (**04/17**) * Deliver Final 2020 Census Architecture and IT Roadmap (**09/18**) * Deliver 2020 Census Initial Universe (**1/20**) * Deliver Final Tabulation Geographic Products (**09/20**) * Deliver Apportionment Counts to President (**12/20**) * Deliver Redistricting Counts to States (**03/21**) * Complete LUCA Program (**09/21**) |
| Preserve, Retire | * Close Field Offices/Regional Centers (**12/20–06/21**) * Complete System Deployment/Retirement (**06/19–12/20**) * Complete the 2020 Census (**9/23**) |
| Discover, Share | * Launch 2020 Census Website (**01/15**) * Begin Recruiting Campaign/Partnership Program (**06/17**) * Deploy Recruitment Website (**06/17**) * Launch Advertising Campaign (**11/19**) * Complete Delivery of Redistricting Counts to the States (**03/21**) |
| Integrate, Visualize | * Develop Non-GDCoP Systems (**05/16–12/17**) * Award Contract for 2020 Census Questionnaire Assistance (**06/16**) * Finalize IT Roadmap (**09/18**) * Deploy Non-GDCoP/CEDCaP Systems (**06/19–12/20**) * Conduct Nonresponse Follow-up (**04/20–08/20**) |
| Analyze, Exploit | * Census Day (**2015** Census and **Optimizing Self-Response Tests**) * Conduct Housing Unit Count Review (**02/20**) * Conduct Group Quarters Count Review (**08/20**) * Tabulate Census Counts (**11/20**) * Use Data for Redistricting/Apportionment (**12/20–03/21**) |



The left-hand column is the steps in the lifecycle figure above with the right-hand column being processes from the census lifecycle. The processes in the census lifecycle contain dates which are color coded by the same legend that appears in the pdf given.

We can see from placing the tasks into bins that fit the data management lifecycle there is a general trend where tasks belong. **Research & Testing Milestones** are mostly categorized as **Planning** while having some outliers in creating systems/tests to report and validate later down the pipeline. **Program Level Milestones** are mostly **Publishing** the collected and analyzed data while having high level milestones for distribution of data and dissolvement of temporary offices. **Decision Points** and **Execution Milestones** fill in steps across the lifecycle where you would normally expect entry/exit points for data, while also controlling the inner loop where data is refined before being **Retired** -> **Integrated**. As an example, the process for the census follows both loops over the ten-year life span like so:

**Plan** (2011–2019) -> **Collect** (2015 tests, 2020 data) -> **Assure** (validation) -> **Describe** (publish counts, report data) -> **Preserve** (close offices) -> **Discover** (share results with states) -> **Integrate** (IT systems, publish websites, award contracts) -> **Analyze** (exploit data, redistrict) -> **Plan** (future cycles)

**Plan** (address canvassing strategy) -> **Collect** (address data) -> **Assure** (cleanse errors) -> **Describe** (publish LUCA results) -> **Analyze** (address challenges) -> **Plan** adjustments (2021 governmental unit challenges CQR)

**Q2. The website at:** [**https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/data-quality.html**](https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/data-quality.html) **provides an overview of data quality measures taken by the US Census Bureau. More information on some of these measures are given in** [**https://unece.org/sites/default/files/2022-07/ECE\_CES\_GE.41\_2022\_13-2210837E.pdf**](https://unece.org/sites/default/files/2022-07/ECE_CES_GE.41_2022_13-2210837E.pdf)

**What we want from you for this question is to look at the six causes for poor data quality and relate them to how these are mitigated by the Census bureau when collecting data.**

The six main data quality issues (Data Entry Errors, Incomplete Data, Duplicate Data, Outdated Data, Lack of Data Standards, Technical Issues) mapped to census bureau problems and mitigation techniques they performed:

|  |  |  |
| --- | --- | --- |
| Incomplete Data (Missing Data) | Failing to collect data from households and individuals due to lack of participation | Seven Contact Attempts, In-Person Follow-Up, Administrative Records, Real Time Analysis of Data (RTAD) |
| Data Entry Errors (Inaccurate Data) | Errors due to misunderstood questions, language barriers, or respondent mistakes. Errors introduced during data transcription, coding, or tabulation | Built-In Prompts, Cognitive Research (pre-census focus groups), 13 Languages for Forms, Decennial Field Quality Monitoring (prevent proxy bias), Online forms and inconsistency detection (automation), Benchmark Comparisons against previous census and other datasets (American Community Survey) |
| Duplicate Data | Missing or double-counting individuals due to incomplete address lists or mobility | Address Canvassing, Post-Enumeration Survey (Independent survey match results to estimate), Demographic Analysis, Community Partnerships |
| Outdated Data | Data that is irrelevant due to the date it was collected (Incomplete or obsolete address lists) | American Community Survey is conducted annually and used sparingly to supplement data for the census. Only the most recent ACS is used. IRS, USPS, and state records to update address lists. Monitored address resolution to identify geographic gaps |
| Lack of Data Standards | Systemic or human biases leading to misclassification (Lack of consistent data formatting for addresses, race, etc.) | Disaggregated Categories, Independent Reviews (CNSTAT, ASA, JASON), forms with built-in prompts, cross-system data integration (weather alerts, IRS, etc.) |
| Technical Issues | Data collection disruptions (pandemic shutdowns), system errors, and cybersecurity risks | Real-Time Analysis of Data, Multimode Response Options, encryption and differential privacy, paper questionnaires |

**Q3. The office of the national coordinator for health Information Technology, has set out a framework for patient demographics data quality (see:** [**https://www.healthit.gov/playbook/pddq-framework/data-quality/data-quality-assessment/**](https://www.healthit.gov/playbook/pddq-framework/data-quality/data-quality-assessment/) **). They have identified a number of quality dimensions that they expect a health organization to comply. In another document they also denote methods to check data quality through data profiling (see:** [**https://www.healthit.gov/playbook/pddq-framework/data-quality/data-profiling/**](https://www.healthit.gov/playbook/pddq-framework/data-quality/data-profiling/) **).**

**What we want from you for this question is an assessment of their quality dimensions with respect to data quality issues that we saw in class. Also compare their data profiling tasks to the data quality mitigation tasks that we identified in class. We are looking at your insight in data quality assessment and mitigation applied to this type of data.**

The main data quality issues (Data Entry Errors, Incomplete Data, Duplicate Data, Outdated Data, Lack of Data Standards, Technical Issues) that we discussed normally deal with V-I-P-R-T, as such, I will map the quality criterion to each issue it is meant to address.

|  |  |  |
| --- | --- | --- |
| Accessibility | the data is available when needed | **Timeliness** – Accessibility is vital for timely decisions***Technical Issues*** |
| Accuracy | affinity with original intent, veracity as compared to an authoritative source, correlation of data elements, and measurement precision | **Validity** – Data must represent reality *Data* ***Entry Errors*** |
| Completeness | availability of required data attributes | **Precision** – Incomplete data lacks usability. ***Incomplete Data*** |
| Coverage | availability of required data records | **Reliability/Precision** – Data must be readily available and not missing ***Technical Issues*** |
| Conformity | alignment of content with required standards | **Validity** – Data must comply with formats for usability and trustworthiness ***Lack of Data Standards*** |
| Consistency | compliance with required patterns and uniformity rules, supported by data entry standards, workflow management, and technical design standards | **Reliability** – Maintains uniformity across sources and time ***Lack of Data Standards*** |
| Integrity | accuracy of data relationships | **Integrity** – Ensures security and accuracy ***Incomplete Data*** |
| Timeliness | the currency of content | **Timeliness** – Accessibility is essential for timely decisions ***Outdated Data*** |
| Uniqueness | each record can be unambiguously identified, also includes checks for redundancy of records | **Integrity** – Ensures one-to-one correspondence between data and entities ***Duplicate Data*** |

To address them more directly, these assessment standards prevent:

1. A poorly designed system that does not scale properly (**Accessibility**, **Timeliness**, **Conformity**)
   1. Inefficient or potentially malicious programming
   2. Poor database design
2. Potentially erroneous data, incomplete records, and duplicate records (**Uniqueness**, **Coverage**, **Consistency**, **Completeness**)
3. Possibly misuse or misrepresentation of the data (**Integrity**, **Accuracy**)

As such, theOffice of the National Coordinator for Health Information Technology has also created a profiling process prior to data assessment to reach conclusions about the data set itself. This in itself is a sort of ‘mitigation’ as it’s questions to ask during the data cleanse process to feed into the quality assessment. One of the examples they give is to profile the new source of data before migrating it (in the context of healthcare) to ensure that values are not missing and data types match the target destination (‘YYYYMMDD’ vs ‘DD-MM-YY’). This can be seen in the assessment under data profiling tasks, the questions are copied below in a table:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | | 1.1 Has the organization profiled patient demographic data? |  |  | | --- | |  | | |  | | --- | | Checks that initial steps were taken to analyze and assess the quality of its demographic data |  |  | | --- | |  | | |  |  |  | | --- | --- | --- | | |  | | --- | | Applies to **initial audit** and **gap analysis** -> profiling data to identify missing values, outliers, or duplicates |  |  | | --- | |  | |  |  | | --- | |  | |
| |  | | --- | | 2.1 Has the organization defined an approach and method for profiling a dataset? |  |  | | --- | |  | | |  | | --- | | Checks whether a structured, repeatable process for profiling is in place (tools, techniques, and workflows) |  |  | | --- | |  | | |  |  |  | | --- | --- | --- | | |  | | --- | | Applies to **schema and process standardization** -> Structuring profiling ensures repeatability (**reliability**), helps avoid unexpected side effects and ensures consistency when identifying anomalies across datasets |  |  | | --- | |  | |  |  | | --- | |  | |
| |  | | --- | | 2.2 Are defects and anomalies identified through data profiling, and are resulting recommendations for remediation reported to stakeholders? |  |  | | --- | |  | | |  | | --- | | Checks if profiling identifies issues and translates findings into actionable remediation steps shared with decision-makers |  |  | | --- | |  | | |  |  |  | | --- | --- | --- | | |  | | --- | | Applies to **incident tracking** and **scenario simulations** -> Mitigation includes identifying issues (anomalies, duplicates) and reporting them. This ensures corrective actions like deduplication or normalization. Simulation techniques such as ablation studies could support profiling decisions by testing different scenarios for data correction |  |  | | --- | |  | |  |  | | --- | |  | |
| |  | | --- | | 3.1 Has data governance approved the set of patient demographic data attributes and sources for regular profiling and monitoring? |  |  | | --- | |  | | |  | | --- | | Checks whether governance bodies have validated the scope and rules for profiling (key attributes, systems, and frequency) |  |  | | --- | |  | | |  |  |  | | --- | --- | --- | | |  | | --- | | Applies to **governance-driven decision-making** -> oversight ensures schema changes, deduplication, or error correction are documented and aligned with organizational priorities. For example, governance might dictate whether duplicate values should be merged or maintained as separate entities due to policy requirements |  |  | | --- | |  | |  |  | | --- | |  | |
| |  | | --- | | 3.2 Has the organization developed standard data profiling tools, metrics, and result report templates? |  |  | | --- | |  | | |  | | --- | | Checks that tools for profiling are standardized (templates for profiling reports and consistent metrics) |  |  | | --- | |  | | Applies to **standardized tools and query design** -> Standard tools and templates support regular profiling, are pre-tested, audited, and standardized to. Reduces the risk of introducing errors when fixing missing values or reconciling duplicates |

**Q4. USAID/DRC provides a checklist of data quality assessment for projects that are funded by them. They have established data quality standards of Validity, Integrity, Reliability, Precision, and Timeliness (V-I-P-R-T) to assess the quality of data and information an implementing partner (IP) submits by analyzing the process used to collect, store, and transmit data to USAID/DRC. The website:** [**https://pdf.usaid.gov/pdf\_docs/pnaec151.pdf**](https://pdf.usaid.gov/pdf_docs/pnaec151.pdf) **provides information about this process.**

**What we want from you is look at the checklist (pages 13-15) and for each item in the checklist find/deduce which of the 5 (V-I-P-R-T) data quality issues that item is assessing.**

The five data quality issues (V-I-R-P-T) are **Validity**, **Integrity**, **Reliability**, **Precision**, and **Timeliness**.

**GENERAL INFORMATION**

1. Name(s) of USAID Assessors -> **Timeliness/Integrity** (prevents falsified information, new and accurate record entry)
2. Date of Interview -> **Timeliness/Integrity** (data is relevant to the time frame of the activities)
3. Reason for Assessment -> **Validity** (purpose of the assessment aligns with data needs)

**PARTNER INFORMATION**

1. Name(s) -> **Integrity** (names provided are accurate and not falsified)
2. Name of Organization -> **Integrity** (same reason as above)
3. Name of USAID Funded Activity -> **Validity** (data must correspond to the correct funded activity)

**DATA COLLECTION AND REPORTING**

1. Does your organization systematically monitor progress of its activities? -> **Validity** (ensures monitoring aligns with objectives)
2. If yes, describe the data collected: level (i.e. household, individual, individual, etc.); types i.e. nutrition, opinion, etc, frequency (i.e. annually) -> **Precision** (details of data granularity and relevance to objectives)
3. Describe collection process: (i.e. Trip reports, surveys, internal reports, technical records, sub ‐ grantee reports) -> **Reliability/Integrity** (to be able to repeat methods for data collection / consistency, no manipulation of data, predefined protocols of legal/ethical guidelines)
4. How is data received by the partner main office? -> **Timeliness/Integrity** (data flow occurs timely, protection from possible tampering)
5. Do they have baseline data? If so, how was the data collected, and what is the baseline for the USAID‐ funded activity? -> **Validity** (baseline data establishes credible benchmarks)
6. Do they have targets for USAID-funded activities? -> **Validity** (ensures alignment with program goals)
7. If so, how were the targets set, and what are they? -> **Validity** (accuracy and relevance of set targets)
8. Describe how the partner organization monitors or plans to monitor USAID funded activities -> **Reliability** (consistency in monitoring processes)
9. How does the partner organization evaluate progress and impact? -> **Precision** (detailed and measurable evaluation criteria)

**INTERNAL DATA MANAGEMENT ISSUES**

1. Are there controls on access to database systems? -> **Integrity/Validity** (protects against unauthorized data manipulation)
2. Do they cross-check data input? -> **Reliability** (ensures consistency and correctness in data entry)
3. Do they use procedures to ensure the quality of financial information? -> **Integrity** (prevents errors or fraud in financial reporting)
4. Do they train their data collectors? -> **Reliability** (consistency through well-trained staff)
5. Do they have a staff position responsible for data monitoring? -> **Precision** (clear roles enhance accountability and detailed oversight)

**ROLES AND RESPONSIBILITIES**

1. Data and name of USAID Staff to follow ‐ up any recommended actions. -> **Timeliness** (relevant updates on responsible parties)
2. Date and name of next USAID staff spot check of partner data files-> **Timeliness** (keeps data monitoring schedules up-to-date)
3. Date and name of next USAID staff field visit to observe activities -> **Timeliness** (relevance of upcoming observations)
4. Date and name of independent organization expected to conduct survey or evaluation to ensure accuracy of data reporting (if needed) -> **Validity** (ensures evaluation accuracy through credible sources)

**Q5. Find all the anomalies and errors in the following table pointed by the arrows.**

**A close-up of a phone number

Description automatically generated**

1. Duplicate entries for Jenny Roberts (R3-4 C2)
   1. Improperly spelt name for one entry (R4C2)
2. First name of Smith missing (R1C3)
3. Incomplete stress address for Jeff Jones (R2C4)
4. Inconsistent numerical convention for phone and fax (R3-4 C8-9)
5. Incorrect entry email for Jeff Jones (R2C10)

**Q6. Find all the anomalies and errors in the following table (color added to point errors)**

**A table with multiple colored text

Description automatically generated**

1. Duplicate customer entries for Thomas Bond with differing GlobalID and CustomerStatus (**Red**)
2. Missing entries for multiple records across the table (**Yellow**)
3. Duplicate GlobalID (**Orange**) and CustomerID for completely different customers
4. Random text entry of ‘Sales’ for CustomerNAME (**Green**)
5. Incomplete data entry for PostalCode (**Purple**)
6. Random text entry ‘4’ and ‘a3’ in City (**Blue**)
7. Data entered into the wrong column ‘Long Road 23’, ‘GB’, IM-M’, ‘A’ (**Blue**)
8. Inconsistent formatting for PostalCode? *uncolored*
9. Inconsistent formatting for City (NY/London)? *uncolored*

**Q7. Find all the anomalies and errors you can find in the following table.**

**A screenshot of a computer

Description automatically generated**

Basic assumptions to be made about the given table given no context:

* Organizations *seem to be* charitable (or non-profit) judging by the nomenclature
* Some First/Last-Name pairs are married couples, but they *don’t have to be*
* All Addresses are on Main st. however the city and zip differ for some entries

1. Missing Zip values for multiple records, given an address these should exist
   1. One zip value is shorter than the rest (H7)
2. Missing names for the Authors Association (A13-D13)
3. Abbreviated first name for record 20 (A20)
4. Improper data entry for record 11 ‘Guest of Lewis Carroll’ (C11)
5. Improper data entry for record 10 First Name 1 and Last Name 1 (A10, B10)
6. Inconsistent data entry for record 6, Last Name 1 ‘Barrett Browning’ (*assuming they’re married)* (B6)
7. Inconsistent data entry for record 14, First Name 2 and Last Name 2 (*assuming they’re married)* (B14, D14)
8. Incomplete data entry for multiple records (*assuming there has to be an organization*) (E2-4, E6-7, E9, E11-12, E14-17, E19-20)