**513 Project**

Menu.csv: <https://uofi.app.box.com/s/66vze2odpr78hgyrvz2ilvtmd23xp5bv/file/802768115355>

Reference: <https://campuswire.com/c/G6023D093/feed/374> (CampusWire 374)

<https://campuswire.com/c/G6023D093/feed/432> (CampusWire #432)

<https://campuswire.com/c/G6023D093/feed/326> (CampusWire #326)

<https://campuswire.com/c/G6023D093/feed/493> (CampusWire #493)

For testing: <https://github.com/pyangyu/5l3-Project>

**1. Description of Dataset (25 points)** a. Here you will provide an ER diagram, an ontology, or a detailed database schema (10 points), and b. a narrative description of the dataset covering structure and content (15 points)

* Hanhsun2:
  + Narrative & ER diagram
    - The dataset is the record of the menus provided for various events.
    - Record
      * (PK, FK) id: A unique identifier for each menu which can be used for referencing the record.
      * notes: addition information for the menu or the event
      * call\_number: a unique identifier used in libraries or archives to locate the physical record
    - Event
      * id (PK, FK)
      * event: the type of the event, such as breakfast, supper…
      * sponsor: the entity or organization that sponsored the events
      * name:
      * date: the date when the event took place
      * status: the status of the event (menu making?)
      * occasion: the reason for the event
    - Venue
      * id (PK, FK)
      * venue: the type of venue where the event took place
      * place: the geographical location of the event
      * location: the specific location of the event
    - menu
      * id (PK, FK)
      * physical\_description: the physical description of the menus
      * currency: the currency specified in the menu
      * currency\_symbol: the symbol for the currency used in the menu
      * page\_count: the number of pages for the menu
      * dish\_count: the number of dishes listed in the menu
* Jw138:

Final Version\_Submitted

**Narrative:**

The menu.csv file is used as the dataset. The ER diagram has been used as the conceptual model of the project. The ER diagram consists of five entities: menu, period, sponsor, venue, and location. Below is a detailed explanation of each entity and its relationships.

* + Menu Entity:
    - Primary key: id (dtype : int64)
    - Attributes:
      * Physical\_description(dtype :object): Material and size information of the menus.
      * Page\_count(dtype : int64): Number of pages in the menu.
      * Notes(dtype :object ): Additional information about the menu or event (e.g., decoration, language).
      * Currency(dtype: object): Currency used in the menu pricing.
      * Call\_numbe(dtype: object)r: Unique identifier for locating the physical record.
      * Dish\_count(dtype:int64 ): Number of dishes listed in the menu.
  + Period Entity:
    - Primary key: id
    - Attributes:
      * Date(dtype :object ): Date of the event.
      * Event(dtype : object): Type of the event.
      * Occasion(dtype :object ): Frequency of the event.
  + Sponsor Entity:
    - Primary key: id
    - Attributes:
      * Location(dtype :object ): Geographical location of the sponsor.
      * Sponsor(dtype :object ): Company sponsoring the event.
  + Venue Entity:
    - Primary key: id
    - Attributes:
      * Venue(dtype : object): Field of the event.
  + Location Entity:
    - Primary key: id
    - Attributes:
      * Place(dtype : object): Geographical location of the event.

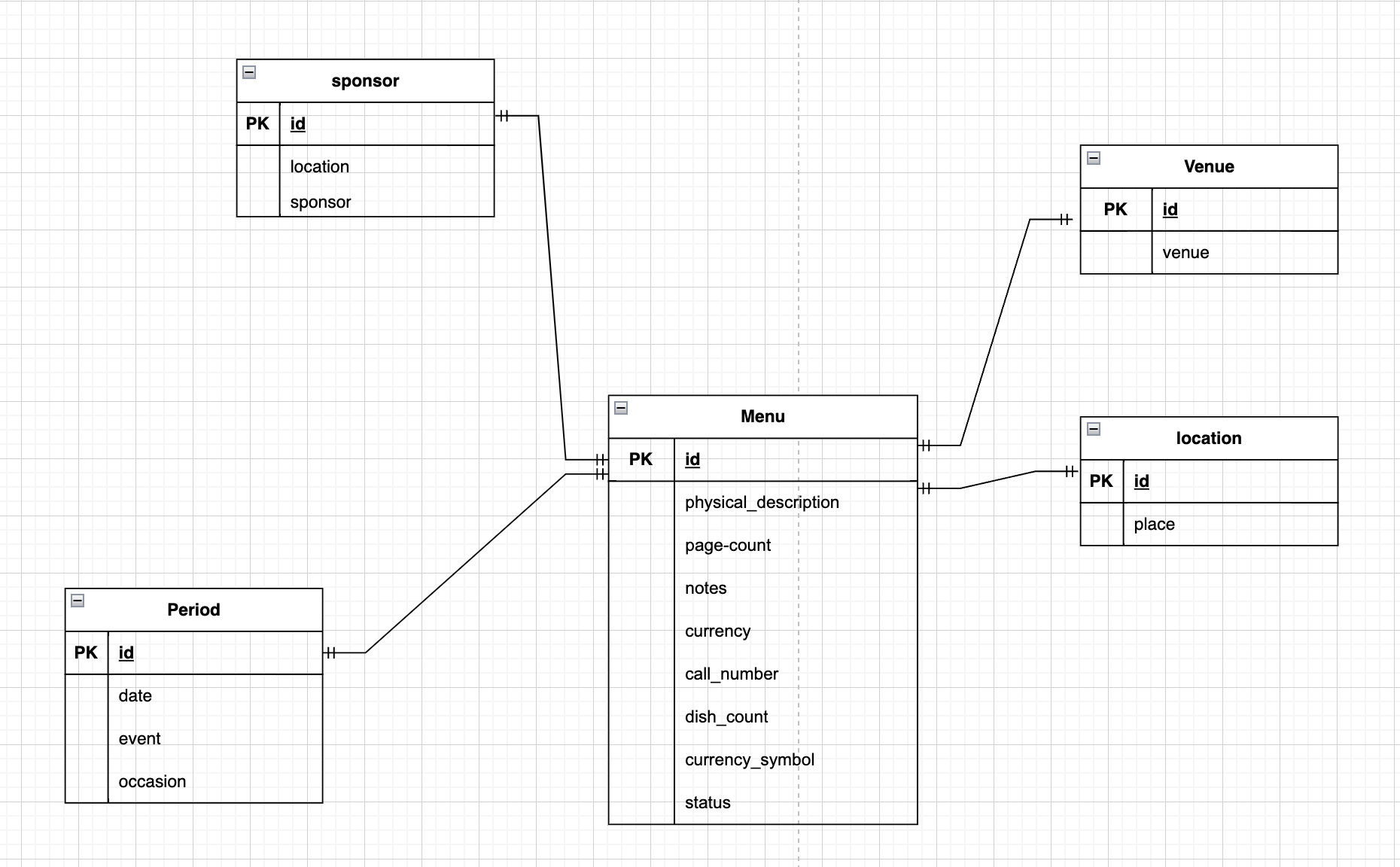
**Relationship Types:**

* + Menu - Sponsor
    - Type: One-to-One
    - Description: One sponsor can be associated with one menu.
  + Menu - Period
    - Type: One-to-One
    - Description: Each period can be associated with one menu.
  + Menu - Venue
    - Type: One-to-One
    - Description: Each venue can be associated one menu.
  + Menu - Location
    - Type: One-to-One
    - Description: Each location can be associated with one menu.

**Temporal Extent:** What is the time span between the earliest and latest dates in the date attribute?

**Spatial Extent:** How many unique places are listed in the place attribute of the dataset?

**ER Diagram**



* Pyang33:

Group into three types:

* Menu
  + id
  + Name: name on the menu
  + physical\_description (e.g., size of the menu)
  + Page\_ count
  + Currency
  + Currency\_symbol
  + status
* Sponsor
  + sponsor
  + call\_number
* Venue
  + venue
  + Place
* Date
  + Date
  + Occasion

Narrative:

**2. Use Cases (30 points)** a. Target (Main) use case U1: data cleaning is necessary and sufficient (20 points) b. “Zero data cleaning” use case U0: data cleaning is not necessary (5 points) c. “Never enough” use case U2 : data cleaning is not sufficient (5 points)

* hanhsun2
* jw138  
  **Use Case U0 (Data cleaning is not necessary):**

**Use Case U1 (Data cleaning is necessary and sufficient):** For use case U1, the objective is to determine how many menus have been provided in the dinner event from the sponsor of Canadian Pacific Railway Company. During the data cleaning process, several issues were identified that need to be addressed to achieve the objective:

* Spelling Errors: Occurrences of incorrect spellings (e.g., "DINNE" instead of "DINNER").
* Unnecessary Punctuation: Presence of extraneous punctuation marks such as question marks, semicolons, and brackets (e.g., "SUPPER(?)", "[DINNER]", "DINNER;").

**Use Case U2 (Data cleaning is not sufficient):**

* Pyang33:

**Use Case U0 (Data cleaning is not necessary):** In Use Case U0, I analyze the data distribution of values in the columns "page count" and "dish count" to determine the typical range for these values. This use case will help us to understand what constitutes a normal (or proper) page and dish count in the menus, helping sponsors in providing appropriate quantities. In this use case U0, data cleaning is not unnecessary because:

* Both columns contain no blank entries.
* The data types are correct; all values are integers.
* From our analysis, there are no data points that are significantly high that could be considered outliers, hence no need for data cleaning in these two columns.

**Use Case U1 (Data cleaning is necessary and sufficient):** For Use Case U1, the use case is trying to determine the most popular sponsor across the menus. Data cleaning is essential because of the following issues (data quality problems) in this column:

* Some entries have missing values NULL.
* Incorrect spellings, unnecessary punctuation (e.g., question marks and periods), and variations in showing the same name (e.g., with the same sponsor, there may be "restaurant" after the sponsor name).
* Some sponsors are within quote marks.

**Use Case U2 (Data cleaning is not sufficient):** Use Case U2 will determine the most common language used in the menus. This case shows that data cleaning is not sufficient because:

* All entries in the "language" column are null, which is a total absence of all useful values.
* With no available values to analyze, it is impossible to impute missing values (although we can conduct from the address places but data cleaning on the language column will never generate useful inputs). Therefore, data cleaning is not sufficient in this use case.

Final Version\_Submitted:  
**Use Case U0 (Data cleaning is not necessary):** In Use Case U0, I analyze the data distribution of values in the columns "page count" and "dish count" to determine the typical range for these values. This use case will help us to understand what constitutes a normal (or proper) page and dish count in the menus, helping sponsors in providing appropriate quantities. In this use case U0, data cleaning is not unnecessary because:

* Both columns contain no blank entries.
* The data types are correct; all values are integers.
* From our analysis, there are no data points that are significantly high that could be considered outliers, hence no need for data cleaning in these two columns.

**Use Case U1 (Data cleaning is necessary and sufficient):** For Use Case U1, the use case is trying to determine the most popular sponsor across the menus. Data cleaning is essential because of the following issues (data quality problems) in this column:

* Some entries have missing values NULL.
* Incorrect spellings, unnecessary punctuation (e.g., question marks and periods), and variations in showing the same name (e.g., with the same sponsor, there may be "restaurant" after the sponsor name).
* Some sponsors are within quote marks.

The other objective is to determine how many menus have been provided in the dinner event by the sponsor of Canadian Pacific Railway Company. During the data cleaning process, several issues were identified that need to be addressed to achieve the objective:

* Spelling Errors: Occurrences of incorrect spellings (e.g., "DINNE" instead of "DINNER").
* Unnecessary Punctuation: Presence of extraneous punctuation marks such as question marks, semicolons, and brackets (e.g., "SUPPER(?)", "[DINNER]", "DINNER;").

**Use Case U2 (Data cleaning is not sufficient):** Use Case U2 will determine the most common language used in the menus. This case shows that data cleaning is not sufficient because:

* All entries in the "language" column are null, which is a total absence of all useful values.
* With no available values to analyze, it is impossible to impute missing values (although we can conduct from the address places but data cleaning on the language column will never generate useful inputs). Therefore, data cleaning is not sufficient in this use case.

**3. Data Quality Problems (30 points)** a. List obvious data quality problems with evidence (examples and/or screenshots) (20 points) b. Explain why / how data cleaning is necessary to support the main use case U1 (10 points)

* hanhsun2
* Jw138

Some obvious data quality problems of dataset have been identified as follows:

* + sponsor:
    - Spelling: OCCIDENTAL & ORIENTAL, OCCIDENTAL & ORIENTAL STEAMSHIP COMPANY, OCCIDENTAL & ORIENTAL STEAMSHIP CO.
    - Spelling: PACIFIC MAIL STEAMSHIP CO., PACIFIC MAIL STEAMSHIP COMPANY
    - Spelling: CANADIAN PACIFIC RAILWAY, CANADIAN PACIFIC RAILWAY COMPANY
    - Spelling?: Red star line, RED STAR LINE - S.S.SOUTHWARK
  + Event:
    - Breakfast: BREAKFAST, FRUHSTUCK/BREAKFAST
      * split: FRUHSTUCK/BREAKFAST
    - Lunch:LUNCH;, LUNCH, CAFE LUNCHEON, LUNCHEON, LUNCH & DINNER
      * Split: LUNCH & DINNER
      * extraneous punctuation marks: LUNCH;,
    - Dinner: [DINNER], DINNER;, DINNER, DINNE, TWENTY-NINTH ANNUAL DINNER, SUNDAY DINNER, SECOND ANNUAL DINNER, SEVENTH ANNUAL DINNER, LUNCH & DINNER, ANNUAL DINNER, DINNER TO THE BOARD OF OFFICERS OF THE CATHOLIC CLUB OF THE CITY OF NEW YORK, SEMI-ANNUAL DINNER, SUPPER, DINNER TO THE BOARD OF OFFICERS OF THE CATHOLIC CLUB OF THE CITY OF NEW YORK, SOUPER, SUPPER(?),
      * Spelling:DINNER, DINNE; SUPPER, SOUPER, SUPPER(?)
      * Extraneous punctuation marks: [DINNER], DINNER;,
      * Type: WENTY-NINTH ANNUAL DINNER, SUNDAY DINNER, SECOND ANNUAL DINNER, SEVENTH ANNUAL DINNER, ANNUAL DINNER, SEMI-ANNUAL DINNER,DINNER TO THE BOARD OF OFFICERS OF THE CATHOLIC CLUB OF THE CITY OF NEW YORK
      * Split: LUNCH & DINNER
    - Banquet: ANNUAL BANQUET, 26TH ANNIVERSARY BANQUET
    - Daily Menu: DAILY MENU, DAILY MENU - 11;30 TO 3;00, MENU
    - Others:CARTE DU JOUR, ANNUAL MEETING, BANQUET OF THE FIFTY NINTH ANNUAL CONVENTION, FIFTH ANNUAL REUNION, SHAKEPEARE COMMENMORATION, COMPLIMENTARY BANQUET GIVEN BY THE CITY GOVERNMENT OF BOSTON TO THE BOARDS OF TRADE OF THE WESTERN CITIES, FEST-BANKETT, ANNUAL OUTING OF EDWARD THOMPSON CO., BANQUE

1. Explain why / how data cleaning is necessary to support the main use case U1 (10 points)

* Cleaning sponsor values is necessary to find the most popular sponsor across menus. There are some entries with NULL values, which we need to impute the missing value by predicting the sponsors from related columns like place or physical\_description, etc. It is important because NULL values can lead to misrepresentation; commonly, a menu should have an associated sponsor. We might also delete rows with NULL sponsors where a NULL sponsor may indicate anonymous sponsorship. So, the NULL values should be handled properly before doing data distribution.
* Misspellings and variations in formatting will lead to wrong results in finding popular sponsors because the same sponsor might be treated differently. To make sure accuracy in identifying popular sponsors, it's necessary to clean the sponsor column before summarizing the data.
* Errors such as "DINNE" instead of "DINNER" need to be corrected, and extraneous punctuation marks like [DINNER], DINNER;, SUPPER(?), etc., should be removed to standardize event names. Correct spelling and standardize event names to ensure that all relevant dinner events are included in the count, preventing underreporting due to misspellings and unstandardized format.
* Data cleaning ensures that variations like "OCCIDENTAL & ORIENTAL," "OCCIDENTAL & ORIENTAL STEAMSHIP COMPANY," and "OCCIDENTAL & ORIENTAL STEAMSHIP CO." are standardized to a single format (e.g., "OCCIDENTAL & ORIENTAL"). Standardization format helps to accurately count specific sponsors.
* Pyang33
  + Data quality problems: (17546 line)

0 id 17545 non-null int64

1 name 3197 non-null object

2 sponsor 15984 non-null object

3 event 8154 non-null object

4 venue 8119 non-null object

5 place 8123 non-null object

6 physical\_description 14763 non-null object

7 occasion 3791 non-null object

8 notes 10613 non-null object

9 call\_number 15983 non-null object

10 keywords 0 non-null float64

11 language 0 non-null float64

12 date 16959 non-null object

13 location 17545 non-null object

14 location\_type 0 non-null float64

15 currency 6456 non-null object

16 currency\_symbol 6456 non-null object

17 status 17545 non-null object

18 page\_count 17545 non-null int64

19 dish\_count 17545 non-null int64

**4. Initial Plan for Phase-II (15 points)** a. Below is a possible plan, listing typical data cleaning workflow steps. In your Plan for Phase-II, fill in additional details for the project steps as needed. In particular, include who of your team members will be responsible for which steps, and list the timeline that you are setting yourselves! ■ S1: Review (and update if necessary) your use case description and dataset description ■ S2: Profile D to identify DQ problems: How do you plan to do it? What tools are you going to use? ■ S3: Perform DC “proper”: How are you going to do it? What tools do you plan to use? Who does what? ■ S4: Data quality checking: is D’ really “cleaner” than D? ● Develop test examples / demos ■ S5: Document and quantify change (e.g. columns and cells changed, IC violations detected: before vs after, etc.)

**S1:**

**Description of the dataset and user case briefly:**

The menu.csv file is used as the dataset for our team; the data types of the dataset include int64, object, and float64. The dataset comprises five entities: menu, period, sponsor, venue, and location. Use Case U1 aims to determine the most popular sponsor across the menus and quantify the number of menus provided for dinner events sponsored by the Canadian Pacific Railway Company. To achieve this objective, data cleaning is necessary. Several issues have been identified and need to be further addressed in Phase 2 of the project.

**S2 & S3:**

**Plan:** planning to use Python for this project. In this use case, we want to find the trends in the popularity of the sponsors over different time periods (e.g., determining the most popular sponsor in 1990 compared to other years. Like sponsor A may be the most popular in 1889, but sponsor B is more popular in 1990).

**7/8/2024 ~ 7/14/2024:**

**Step 1: Prepare Data by data type and simple revising (hanhsun2)**

* Convert date values to a uniform ISO format in column date and trim leading and trailing whitespace from the sponsor column.

**Step 2: Standardizing, using the same formats for values (hanhsun2, payng33)**

* Correct spelling errors and remove unnecessary punctuation in the sponsor column. Adjust the case of letters to either uppercase or lowercase if it is needed.
* Standardize the formatting by removing any keywords following the sponsor names. Like, removing extra unnecessary information behind the sponsors' names, and making the column values more uniform. Using value\_counts will help identify and moderate variations in sponsor names.

**7/12/2024 ~ 7/20/2024:**

**Step 3: Handling Missing Values (jw138 )**

* Impute missing sponsor values by considering related information from the place column, for example, certain sponsors may only host events or provide menus at specific places.
* We may also use the name columns to analyze the sponsor column because they should be the same.

**Step 4: Integrity Checks (pyang33, jw138)**

* Use pandas to identify and fix integrity constraint violations, such as duplicate records or inconsistencies within the data.
* Remove duplicate entries with the same IDs to maintain data quality or something similar like that. ? (probably need to do IC checks before or after the changing values. Before data cleaning, the IC checks would probably help to identify some structural problems or relationship problems within the ER diagram we created. Detecting duplicates early in this project will prevent us from introducing further errors to the datasets. After cells change, we can do an IC check to ensure there is no issue introduced while we clean the data.)

**7/20/2024 ~ 7/28/2024:**

**Step 5: Summarization and Verification (pyang33, jw138, hanhsun2)**

* Transition from different versions and verify that the cleaning methodologies are implemented correctly.
* Compare results and conclude test cases by using datasets D and D'. Estimate the performance of the data cleaning methods to check how cleaning data from D to D' can improve accuracy.
* Write reports that document the results and conclusions from the data cleaning projects and document quality changes in public platforms, e.g. Github, Tableau.

refined version hanhsun2

**Step 1: Use Case and Dataset Review**

1. Use case description:

* For Use Case U1, data cleaning is necessary to find the most popular sponsor and determine the number of dinner menus provided by Canadian Pacific Railway Company. To achieve this objective, data cleaning is necessary. Several issues have been identified, including but not limited to extra punctuation marks, missing values, and spelling errors. Correcting these issues ensures accurate and reliable results.

1. Dataset description:

* The menu.csv file is used as the dataset for our team; the data types of the dataset include int64, object, and float64. The dataset comprises five entities: menu, period, sponsor, venue, and location. Each entity has one-to-one relationships connecting to menu. This structure helps identify the most popular sponsor and analyze menu details accurately.

**7/8/2024 ~ 7/14/2024:**

**Step 2: Standardizing Data Formats to Identify Data Quality (DQ) Problems (hanhsun2, payng33)**

To identify and address data quality problems, we will standardize the data formats and ensure consistency across the dataset.

1. Correct spelling errors:

* Scan the sponsor column for spelling errors and correct them to maintain consistency.

1. Remove unnecessary punctuation:

* Identify and remove any unnecessary punctuation marks in the sponsor names to ensure uniformity.

1. Adjust letter case:

* Standardize the case of letters in the sponsor names to either all uppercase, all lowercase, or title case if needed.

1. Remove extraneous keywords:

* Remove any additional keywords or unnecessary information following the sponsor names to make the column values more uniform.

1. Standardize formatting:

* Use value\_counts() to identify variations in sponsor names. This will help spot inconsistencies and standardize them, ensuring that variations like " PENNSYLVANIA R.R." and "P.R.R" are treated as a single entity.

By following these steps, we will enhance data quality by ensuring that all sponsor names are consistent, properly formatted, and free of errors. This standardization process will make the data more reliable and easier to analyze.

**7/12/2024 ~ 7/20/2024:  
Step 3: Handling Missing Values in Data Cleaning (DC) Proper (jw138)**

To handle missing values effectively and ensure the integrity of the dataset, we will adopt the following steps.

1. Impute missing sponsor values through related information:

* Analyze the 'place' column to infer possible sponsors. For instance, if certain sponsors are known to host events or provide menus at specific places, use this relationship to fill in the missing sponsor values.

1. Impose consistency for name and sponsor columns:

* Cross-reference the sponsor column with name column. Since these two columns should match with each other. We can use this consistency to fill in missing values.

By following these steps, we can systematically handle missing values, ensuring that the data remains reliable for further analysis.

**Step 4: Data Quality Checking (pyang33, jw138)**

To verify that the cleaned dataset (D') is indeed cleaner and more reliable than the original dataset (D), we will perform integrity checks and develop test examples/demos.

1. Define data quality metrics:

* Establish clear metrics for data quality, such as the number of duplicate records, percentage of missing values, consistency of categorical data, and adherence to integrity constraints. With these metrics, we can compare (D’) and (D) to check if (D’) is cleaner.

1. Integrity checks for (D’) and (D):

* Use pandas to identify if there are still any integrity constraint violations such as duplicate records, missing values, and inconsistencies within the data.

1. Comparison and validation:

* Compare key metrics between D and D' to validate the improvements. Visualize the results to demonstrate the changes.

**7/21/2024 ~ 7/27/2024:**

**Step 5: Documenting and Quantifying Changes**

To document and quantify changes in the dataset, we will **track key items** before and after data processing.

1. Columns and cells changed:

* Record the cells and columns that were modified during cleaning.

1. Integrity checks result:

* Measure and record the data quality metrics before and after the data cleaning.

1. Methodology:

* Record the methodology applied for the data cleaning including tools and techniques.

Documenting process analysis:

* Start by recording the analysis histories. There are mechanisms to record the history of changes, and it's important to keep useful information. Some tools and platforms, like openRefine, can provide traces of the data investigation and cleaning process.
  + Document the tools we will use to investigate the data. **(pyang33)**
  + Draw a plan of how we will perform the data cleaning**. (jw138)**
  + Decide the use cases and columns we use to modify. We will have more than one use case as candidates, and we will have a discussion of deciding the final use case. **(jw138)**
* Document processes and findings at each step of data cleaning. Share this information with teammates and other developers for collaboration and continuous improvement. Sharing ideas with others and analyzing the data cleaning results and step-by-step conclusions throughout the process is useful. GitHub is one of the platforms where we can share our data-cleaning processes and results.
  + Keep a record of all changes made to columns and cells from the beginning (this would be documented in a version. That is, we may change our methodologies and document different approaches in other views). **(pyang33，jw138)**
  + Note all the integrity constraint (IC) violations and compare the data before and after handling IC violations. **(hanhsun2)**
  + Use platforms like GitHub to share the data and record the feedback between our team members. **(pyang33)**
* Guide other developers or explorers through the data cleaning processes with analysis, so we can communicate our findings and understanding with others. Tools like YesWorkflow can be used to show the provenance of the data and show how data cleaning can refine the data, where it can provide a way of storytelling with sequential flow charts.
  + Use YesWorkFlow to document data’s provenance and formalize the data cleaning steps **(jw138, hanhsun2)**
  + Story-telling with the data with more detailed analysis. **(hanhsun2)**