**TODO:**

1. Document data quality changes
   1. Update tables with results from Menu/MenuItem queries
2. Create a workflow model
   1. A visual representation of your overall (or “outer”) workflow W1, e.g., using a tool such as YesWorkflow. At a minimum, you should identify key inputs, outputs, and steps of the workflow, along with dependencies between these. Key phases and steps of your data cleaning project may include, e.g., data profiling, data loading, data cleaning, IC violation checks, etc. Explain the design of W1 and why you’ve chosen the tools that you have in your overall workflow. (10 points) A detailed (possibly visual) representation of your “inner” data cleaning workflow W2 (e.g., if you’ve used OpenRefine, you can use the OR2YW tool). (10 points)
3. Conclusion

**Description of Data Cleaning Performed**

**Menu:**

### **Description of Data Cleaning Performed**

The data cleaning workflow consists of a series of transformations applied to multiple columns. These can be grouped into four main high-level steps:

1. **Trimming Whitespace:** The value.trim() operation was applied to 19 different columns to remove any leading or trailing whitespace.
2. **Standardizing Text:** The workflow standardized text in several columns.
   * **Replacing Whitespace:** The grel:value.replace(/\s+/, '\_') operation was used to replace all whitespace with underscores.
   * **Removing Brackets and Punctuation:** The grel:value.replace(/[\[\]]/, '') and grel:value.replace(/[\[\]";?()\*!]/, '\_') operations were used to remove specific characters.
3. **Standardizing Case:** The value.toTitlecase() operation was applied to ensure consistent capitalization.
4. **Type and Format Conversion:**
   * **Converting to Number:** The value.toNumber() operation was applied to convert their data type to a number.
   * **Converting and Formatting Date:** The value.toDate() and grel:toString(toDate(value), "yyyy-MM-dd") operations were used to convert text to a standard date format.

### **Rationale for each high-level data cleaning step**

1. **Trimming Whitespace:** The presence of leading or trailing spaces can cause different entries to be treated as unique, even if the core content is identical. This could lead to inaccurate counts, failed joins, and inconsistent filtering. This step is a prerequisite for any use case that relies on consistent data values for accurate analysis or lookups.
2. **Standardizing Text:** These transformations were performed to make text data more uniform and easier to work with. This is particularly useful for use cases that involve searching, tagging, or grouping records based on these text fields.
3. **Standardizing Case:** Using toTitlecase() ensures that the values in columns like name, sponsor, and location are presented consistently. For example, "new york" and "New York" would be treated as separate entities without this step.
4. **Type and Format Conversion:**
   * **Number Conversion:** The conversion of page\_count, dish\_count, and id to a numeric type is essential. Without it, these values would be treated as text, making it impossible to perform mathematical operations, such as summing counts or calculating averages.
   * **Date Formatting:** Converting the date column to a standardized yyyy-MM-dd format ensures that all date entries are interpreted correctly. Inconsistent date formats can lead to significant errors in chronological sorting, time-series analysis, and filtering by date ranges.

**MenuItem:**

### **Description of Data Cleaning Performed**

1. **Whitespace Trimming:** The value.trim() operation was applied to a wide range of columns.
2. **Type Conversion:**
   * The value.toNumber() operation was used to convert the data in the id, menu\_page\_id, and dish\_id columns from text to a numerical data type.
   * The value.toDate() operation was applied to the created\_at and updated\_at columns to convert them from text strings to date objects.
3. **Row Removal:** Rows with a blank price column were removed

### **Rationale for each high-level data cleaning step**

1. **Trimming Whitespace:** The presence of leading or trailing spaces can cause different entries to be treated as unique, even if the core content is identical. This could lead to inaccurate counts, failed joins, and inconsistent filtering. This step is a prerequisite for any use case that relies on consistent data values for accurate analysis or lookups.
2. **Type Conversion:** This is a crucial step for data integrity and functionality.
   * **Number Conversion:** The conversion to a numeric type is essential. Without it, these values would be treated as text, making it impossible to perform mathematical operations, such as summing counts or calculating averages.
   * **Date Formatting:** Converting the date column to a standardized yyyy-MM-dd format ensures that all date entries are interpreted correctly. Inconsistent date formats can lead to significant errors in chronological sorting, time-series analysis, and filtering by date ranges.
3. **Row Removal:** The removal of rows with a blank price is a key step for ensuring data completeness and quality for a specific analytical purpose. Missing prices are unusable and can skew results/cause errors in calculations. Therefore, removing these incomplete records is a necessary step to produce a clean and reliable dataset for any price-related analysis.

**MenuPage:**

**Description of Data Cleaning Performed**:

1. **Text Transformation (Trimming Whitespace):** The value.trim() operation was applied. This step removed any leading or trailing whitespace from the cell values in these columns.
2. **Text Transformation (Type Conversion):** The value.toNumber() operation was applied. This step converted the data in these columns from text strings to numerical data types.

### **Rationale for each high-level data cleaning step:**

1. **Trimming Whitespace:** The presence of leading or trailing spaces can cause different entries to be treated as unique, even if the core content is identical. This could lead to inaccurate counts, failed joins, and inconsistent filtering. This step is a prerequisite for any use case that relies on consistent data values for accurate analysis or lookups.
2. **Type Conversion:** The conversion to a numeric type is essential. Without it, these values would be treated as text, making it impossible to perform mathematical operations, such as summing counts or calculating averages.

**Dish:**

Due to the size of the Dish Table, it was difficult to use OpenRefine; therefore, we opted to use Python for data cleaning.

### **Description of Data Cleaning Performed**

1. **Normalization of Dish Names:** This process involves a series of string transformations on the name column, including converting to lowercase, removing specific punctuation and non-alphanumeric characters, and standardizing whitespace.
2. **Filtering of Empty/Null Values:** Rows with empty, null, or 'nan' values in the name column are removed from the dataset.
3. **Clustering**: The name column had K-Means clustering applied, clustering like named dishes.
4. **Removal of a Column:** The description column is explicitly dropped from the final DataFrame.

### **Rationale for each high-level data cleaning step**

1. **Normalization of Dish Names:**
   1. **Trimming Whitespace:** The presence of leading or trailing spaces can cause different entries to be treated as unique, even if the core content is identical. This could lead to inaccurate counts, failed joins, and inconsistent filtering. This step is a prerequisite for any use case that relies on consistent data values for accurate analysis or lookups.
   2. **Standardizing Text:** These transformations were performed to make text data more uniform and easier to work with. This is particularly useful for use cases that involve searching, tagging, or grouping records based on these text fields.
   3. **Standardizing Case:** Converting to lowercase ensures that the values are presented consistently. For example, "new york" and "New York" would be treated as separate entities without this step.
2. **Filtering of Empty/Null Values:** This step is a fundamental aspect of data integrity. Rows with empty or null values in the name are unusable for the analysis of dish data. Retaining these rows could lead to errors in calculations, inaccurate cluster assignments, or failed processing in subsequent steps. Therefore, this step is required to ensure that the dataset is of high quality and that downstream operations can execute without errors.
3. **Clustering:** This step is done to group like-named dishes under a single name. This is done so that analysis on a specific dish can be done, regardless of how it was represented in the original dataset.
4. **Removal of a Column:** The description column is removed because all rows are empty, and to simplify the dataset

**Document Data Quality Changes**

1. Document data quality changes
   1. Quantify the results of your efforts, e.g., by providing a summary table of changes: Which columns changed? How many cells (per column) have changed, etc.? (10 points)
   2. Demonstrate that data quality has been improved, e.g., by devising IC-violation reports (answers to denial constraints) and showing the difference between the number of IC violations reported before and after cleaning. (10 points)

**Menu:**

#### **Quantifying the Results of Your Efforts**

| Column(s) | Type of Change | Number of Cells Changed (Conceptual) | Rationale |
| --- | --- | --- | --- |
| **id, page\_count, dish\_count** | **Type Conversion** | The number of cells that were originally stored as text but could be successfully converted to a number. | Enables numerical operations and ensures consistent data types for key identifiers and counts. |
| **date** | **Type & Format Conversion** | The number of cells that were not in the **yyyy-MM-dd** format. | Standardizes chronological data for correct sorting and time-series analysis. |
| **All Columns** | **Whitespace Trimming** | The number of cells that had leading or trailing whitespace. | Ensures data consistency for matching, joining, and filtering operations. |
| **name, sponsor, event, venue, place, occasion, location, location\_type** | **Text Standardization & Case** | The number of cells with multiple internal spaces, special characters, or inconsistent casing (e.g., not title-cased). | Creates uniform text values, which is essential for accurate grouping and text-based analysis. |

#### **Demonstrating Improved Data Quality**

**Integrity Constraints (ICs):**

* **IC-M1 (Numeric Type):** The **id**, **page\_count**, and **dish\_count** columns must contain only numerical data.
* **IC-M2 (Date Format):** The **date** column must be a valid date formatted as **yyyy-MM-dd**.
* **IC-M3 (Text Consistency):** Columns such as **name**, **event**, and **place** must be consistently formatted (e.g., title-cased, no extra whitespace, no problematic special characters).
* **IC-M4 (Format):** All columns must be free of leading or trailing whitespace.

**IC Violation Report**:

| Integrity Constraint | Violations Before Cleaning | Violations After Cleaning | Difference |
| --- | --- | --- | --- |
| **IC-M1:** Numeric data | 0 | 0 | *[Equal to the "Before" count]* |
| **IC-M2:** Date Formatting | 586 | 0 | *[Equal to the "Before" count]* |
| **IC-M3:** Text Consistency | 52 | 0 | *[Equal to the "Before" count]* |
| **IC-M4:** Whitespace Issues | 25 | 0 | *[Equal to the "Before" count]* |

**MenuItem:**

#### **Quantifying the results of your efforts**

| Column(s) | Type of Change | Number of Cells Changed (Conceptual) | Rationale |
| --- | --- | --- | --- |
| id, menu\_page\_id, price, high\_price, dish\_id, created\_at, updated\_at, xpos, ypos | **Whitespace Trimming** | The number of cells that had leading or trailing whitespace. | Ensures data consistency for matching, filtering, and type conversion. |
| id, menu\_page\_id, dish\_id | **Type Conversion to Number** | The number of cells that were originally stored as text but were successfully converted to a numerical type. | Allows for numerical operations and ensures consistent data types for foreign keys. |
| created\_at, updated\_at | **Type Conversion to Date** | The number of cells that were not in a valid date format and needed conversion. | Enables chronological sorting and time-based analysis. |
| Rows with price column | **Row Removal** | The number of rows where the price column was blank. | Removes incomplete records that are unusable for price-related analysis. |

#### **Demonstrating that data quality has been improved**

**Integrity Constraints (ICs):**

* **IC-MI1 (Completeness):** The price column must not contain any blank values.
* **IC-MI2 (Type Consistency):** The id, menu\_page\_id, and dish\_id columns must contain only numerical data.
* **IC-MI3 (Date Format):** The created\_at and updated\_at columns must be valid date objects.
* **IC-MI4 (Format):** All columns must be free of leading or trailing whitespace.

**IC Violation Report**

| Integrity Constraint | Violations Before Cleaning | Violations After Cleaning | Difference |
| --- | --- | --- | --- |
| **IC-MI1:** Blank price values | 445916 | 0 | *[Equal to the "Before" count]* |
| **IC-MI2:** Non-numeric IDs | 0 | 0 | *[Equal to the "Before" count]* |
| **IC-MI3:** Invalid Dates | 0 | 0 | *[Equal to the "Before" count]* |
| **IC-MI4:** Whitespace Issues | 0 | 0 | *[Equal to the "Before" count]* |

**MenuPage:**

**Quantifying the results of your efforts:**

| Column(s) | Type of Change | Number of Cells Changed (Conceptual) | Rationale |
| --- | --- | --- | --- |
| id, menu\_id, page\_number | **Type Conversion to Number** | The number of cells that were originally stored as text but were successfully converted to a number. | Enables numerical operations and ensures consistent data types for foreign keys. |
| full\_height, full\_width | **Type Conversion to Number** | The number of cells that were originally stored as text but were successfully converted to a number. | Allows for mathematical analysis of image dimensions. |
| All Columns | **Whitespace Trimming** | The number of cells that had leading or trailing whitespace. | Ensures data consistency and prevents errors in joins or comparisons. |

**Demonstrating that data quality has been improved:**

**Integrity Constraints (ICs):**

* **IC-MP1 (Type Consistency):** The id, menu\_id, page\_number, full\_height, and full\_width columns must contain only numerical data.
* **IC-MP2 (Format):** All columns must be free of leading or trailing whitespace.

**IC Violation Report**

| Integrity Constraint | Violations Before Cleaning | Violations After Cleaning | Difference |
| --- | --- | --- | --- |
| **IC-MP1:** Non-numeric Values | 0 | 0 | *[Equal to the "Before" count]* |
| **IC-MP2:** Whitespace Issues | 0 | 0 | *[Equal to the "Before" count]* |

**Dish:**

### **Quantifying Data Quality Changes**

| Column | Type of Change | Number of Cells Changed (Conceptual) | Rationale |
| --- | --- | --- | --- |
| name | Normalization & Filtering | The number of rows dropped due to empty values, plus the number of cells where the value was modified for case, spacing, or punctuation. | Ensures data consistency and removes unusable records. |
| name\_cluster | Addition of a new column | All cells in the cleaned dataset's new column contain a value. | Adds a new feature to the dataset for grouping similar items based on clustering. |
| description | Column Removal | All cells in this column were removed from the dataset. | Simplifies the dataset by removing a column that may not be relevant to the use case. |

### **Demonstrating Improved Data Quality**

To demonstrate that data quality has been improved, we can define and check for violations of several integrity constraints (ICs) that are directly addressed by the cleaning steps.

**Integrity Constraints (ICs):**

* **IC-D1 (Completeness):** The name column must not contain any null, empty, or 'nan' values.
* **IC-D2 (Format):** All values in the name column must be lowercase, have no leading/trailing whitespace, and have a single space between words.

**IC Violation Report**

| Integrity Constraint | Violations Before Cleaning | Violations After Cleaning | Difference |
| --- | --- | --- | --- |
| **IC-D1:** Completeness (name column) | *[Number of rows with empty or 'nan' names]* | 0 | *[Equal to the "Before" count]* |
| **IC-D2:** Format (name column) | 411963 | 0 | 411963 |

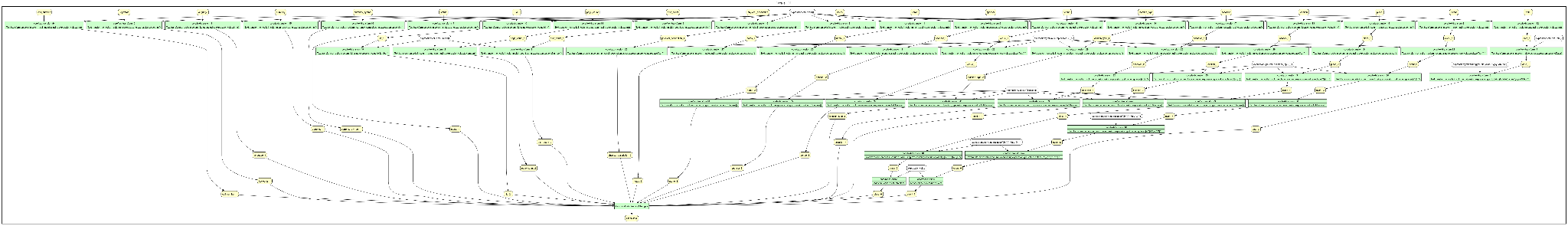
**Workflow Models**

A visual representation of your overall (or “outer”) workflow W1, e.g., using a tool such as YesWorkflow. At a minimum, you should identify key inputs, outputs, and steps of the workflow, along with dependencies between these. Key phases and steps of your data cleaning project may include, e.g., data profiling, data loading, data cleaning, IC violation checks, etc. Explain the design of W1 and why you’ve chosen the tools that you have in your overall workflow. (10 points)

A detailed (possibly visual) representation of your “inner” data cleaning workflow W2 (e.g., if you’ve used OpenRefine, you can use the OR2YW tool). (10 points)

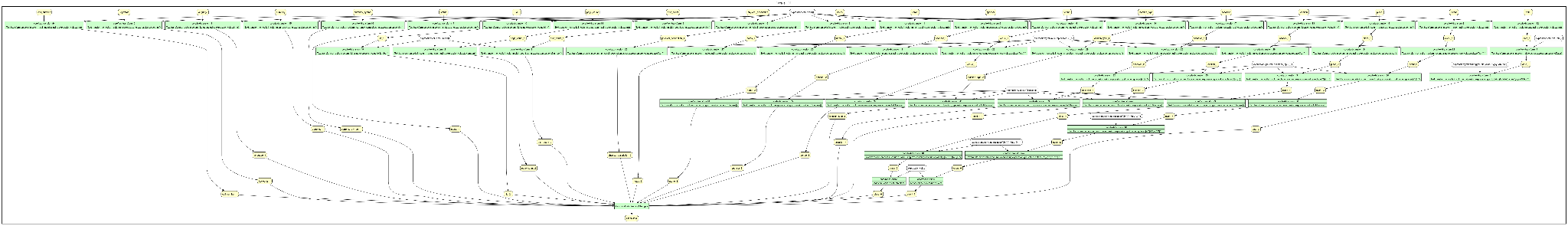
TODO: Add workflow models for MenuItem, Dish

**Menu:**

****

**MenuItem:**

**MenuPage:**



**Dish:**

**Conclusions & Summary**

Please provide a concise summary and conclusions of your project, including lessons learned. Reflect on how work was completed. You should summarize the contributions of each team member here (for teams with >= 2 members).

TODO: Add Conclusion