DATA MANAGEMENT: DATA ACQUISITION AND DATA CLEANING

HANDBOOK

Topic of Dataset - Alcohol Related Disease Impact

Source - Founded more than 70 years ago, the Centres for Disease Control and Prevention (CDC) is the United States of America's leading data-driven, science-based public health organisation. The CDC tackles a range of health issues by using evidence-based procedures and thorough scientific research, all while maintaining a strong commitment to safeguarding public health. Its diverse goal includes responding to public health emergencies, protecting children's health, preventing and controlling diseases, and assisting families, companies, and communities. The CDC is a vital tool in the country's continuous quest of public health excellence because it constantly applies science to practice, helping to shape and carry out policies that advance people's health and well-being.

Link:healthhttps://data.cdc.gov/browse?category=AlcoholRelated+Disease+Impact&q=alcohol&sortBy=relevance

Size of the Dataset: Rows :30,000

Columns: 36

Purpose of the topic

The topic "Alcohol-Related Disease Impact" was selected due to its intriguing and uncommon prospect for investigation, as well as its relative lack of comprehensive coverage. This emphasis promises fresh perspectives on the complex effects of alcohol use on health, illuminating previously unexplored areas and advancing our knowledge of the complex interplay between alcohol and illness. This research has the potential to close knowledge gaps and provide new insights into an important yet understudied area of public health. There is an urgent need to increase public knowledge of alcohol's negative effects and associated illnesses given the growing influence of alcohol intake among young people through emphasising the negative impacts on one's physical and mental health, along with the increased vulnerability to alcohol-related illnesses youth need to make wise decisions. Overall, the project's emphasis on the effects of alcohol on disease is in line with the larger objective of advancing public health and emphasises the importance of using evidence-based approaches to address this pressing social issue.

Uses cases

The "Alcohol-Related Disease Impact" dataset can be valuable in addressing public health issues and providing guidance for evidence-based policy. Here are a few possible applications for this dataset:

- **Epidemiological Research**: By examining the dataset, researchers can gain insight into the trends and prevalence of diseases linked to alcohol consumption in various geographic areas, socioeconomic categories, and demographic groupings. To identify at-risk people and modify interventions appropriately, this knowledge is essential.
- **Healthcare Resource Allocation:** By using the dataset, hospitals and other healthcare institutions may plan ahead and distribute resources efficiently. Planning for treatment centres, specialised care units, and preventive programmes can be made easier by having a better understanding of how alcohol-related disorders affect healthcare systems.
- **Educational Initiatives:** By using the dataset, educational institutions can improve their alcohol education initiatives. Incorporating empirical data regarding the health implications of alcohol use allows instructors to provide students a more thorough awareness of the risks related to binge drinking.
- **Long-Term Health Planning**: The dataset can be used by policymakers and public health experts to guide long-term health planning. Through the projection of future alcohol-related illness burden, policymakers can create long-term plans for treatment, prevention, and the expansion of healthcare infrastructure.
- **Research on the Effectiveness of Interventions**: Scientists can assess how well different programmes and laws work to lower the incidence of disorders linked to alcohol use. This feedback loop makes sure that tactics can be improved in light of actual results.

DATA PROFILING

Data profiling is a crucial step in assessing the quality of a dataset, and it involves several key checks to ensure data integrity and reliability. The first step in data quality checking is completeness, where the presence of missing values in each column is determined. This helps identify areas that may require further investigation or imputation. Consistency checks follow, addressing potential inconsistencies in data formats or unexpected values within categorical columns. Validity checks assess whether the data in each column adhere to expected formats or ranges, ensuring data integrity. Conformity verification ensures that the data align with specified formats or standards, maintaining consistency across the dataset. Accuracy, although challenging to measure without external benchmarks, is evaluated through indicators like plausible values in numerical columns. Lastly, uniqueness checks for duplicate rows or values within columns expected to be unique, such as identifiers. Together, these steps form a comprehensive approach to evaluating and enhancing data quality.

UNCLEANED	Completeness	Consistency	Validity	Conformity	Accuracy	Uniqueness
Location_New	O	X	O	X	O	X
Phone Number	O	X	O	X	O	X
YearStartNew	O	X	O	X	O	X
YearEndNew	X	X	O	X	O	X
Random	O	X	O	X	O	O
Source Row Number	O	O	O	O	O	O
YearEnd	O	O	O	O	O	O
LocationAbbr	O	X	O	X	O	O
LocationDesc	O	X	O	X	O	O
DataSource	O	O	O	O	O	O
ConditionType	O	O	O	O	O	O
ConditionType-1	O	O	O	O	O	O
Category	O	O	O	O	O	O
Cause_of_Death	O	O	O	O	O	O
Data_Value_Unit	O	O	O	O	O	O
Data_Value_Type	O	O	O	O	O	O
Data_Value	O	O	X	O	X	X
Data_Value_Alt	O	O	X	O	X	X
Data_Value_Footnote_Symbol	X	X	X	X	X	X
Data_Value_Footnote	X	X	X	X	X	X
Effect	O	O	O	O	O	O
ConsumptionPattern	O	O	O	O	O	O
Sex	O	O	O	O	O	O
AgeCategory	O	O	O	O	O	O
AgeGroup	O	O	O	O	O	O
LocationID	O	O	O	X	X	X
ConditionTypeID	O	O	O	O	O	O
CategoryID	O	O	O	O	O	O
Cause_of_Death_ID	O	O	O	O	O	O
EffectID	O	O	O	O	O	O
ConsumptionID	O	O	O	O	O	O
SexID	O	O	O	O	O	O
AgeCategoryID	O	O	O	O	O	O
AgeGroupID	O	O	O	O	O	O
DataValueUnitID	O	O	O	O	O	O
DataValueTypeID	O	O	O	O	O	O

UNPROCESSED DATA

Profile report – Alochol Related Disease.xlsx/Sheet1

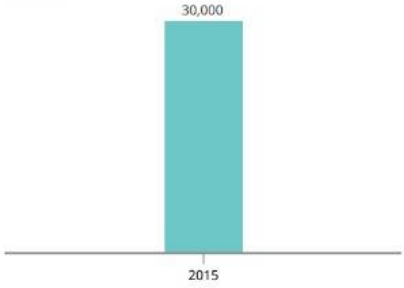
Untitled Flow – Job ID: 24243548

All Data — 30 columns 30,000 rows 4 data types

93% valid values · 0.1% mismatching values · 7% missing values

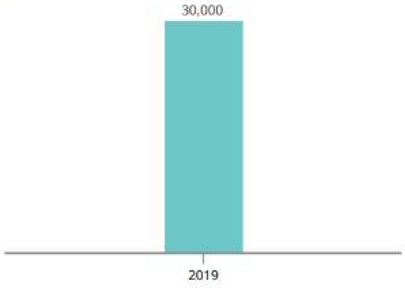
YearStart

Type	Date / Time (yyyy)
Valid	30,000
Mismatched	0
Empty	0



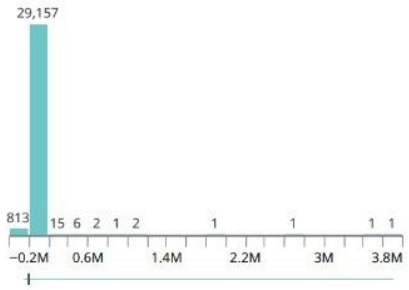
YearEnd

Type	Date / Time (yyyy)
Valid	30,000
Mismatched	0
Empty	0



Data_Value

Type	Integer
Valid	30,000
Mismatched	0
Empty	0



Minimum	-41,647
Lower quartile	0
Median	13
Upper quartile	167
Maximum	3,713,177

UNCLEAN DATA

Profile report – Alcohol Diseases Dataset-Unclean-30k.

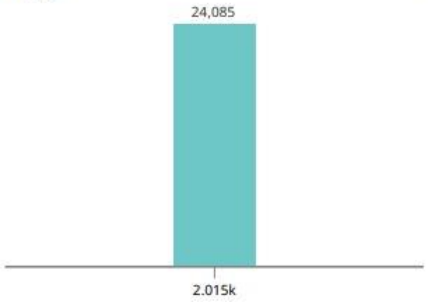
Untitled Flow – 2 – Job ID: 24244679

All Data — 36 columns 30,000 rows 6 data types

● 93% valid values ● 1% mismatching values ● 6% missing values

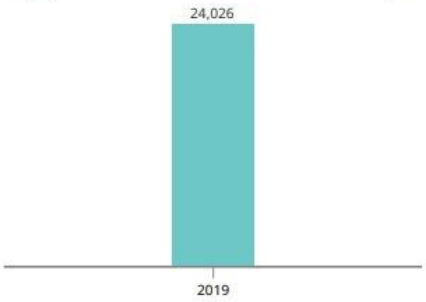
YearStartNew

Type	Integer
Valid	24,085
Mismatched	5,915
Empty	0



YearEndNew

Type	Date / Time (yyyy)
Valid	24,026
Mismatched	0
Empty	5,974



ConditionType

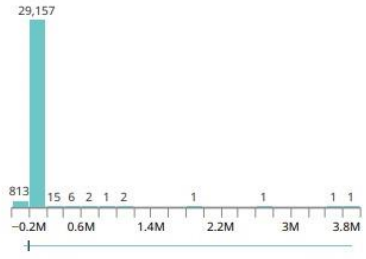
Type	String
Valid	30,000
Mismatched	0
Empty	0
Top 3 values	
Chronic	20,393
Acute	9,116
Total	491

ConditionType-1

Type	String
Valid	30,000
Mismatched	0
Empty	0
Top 3 values	
Chronic	20,393
Acute	9,116
Total	491

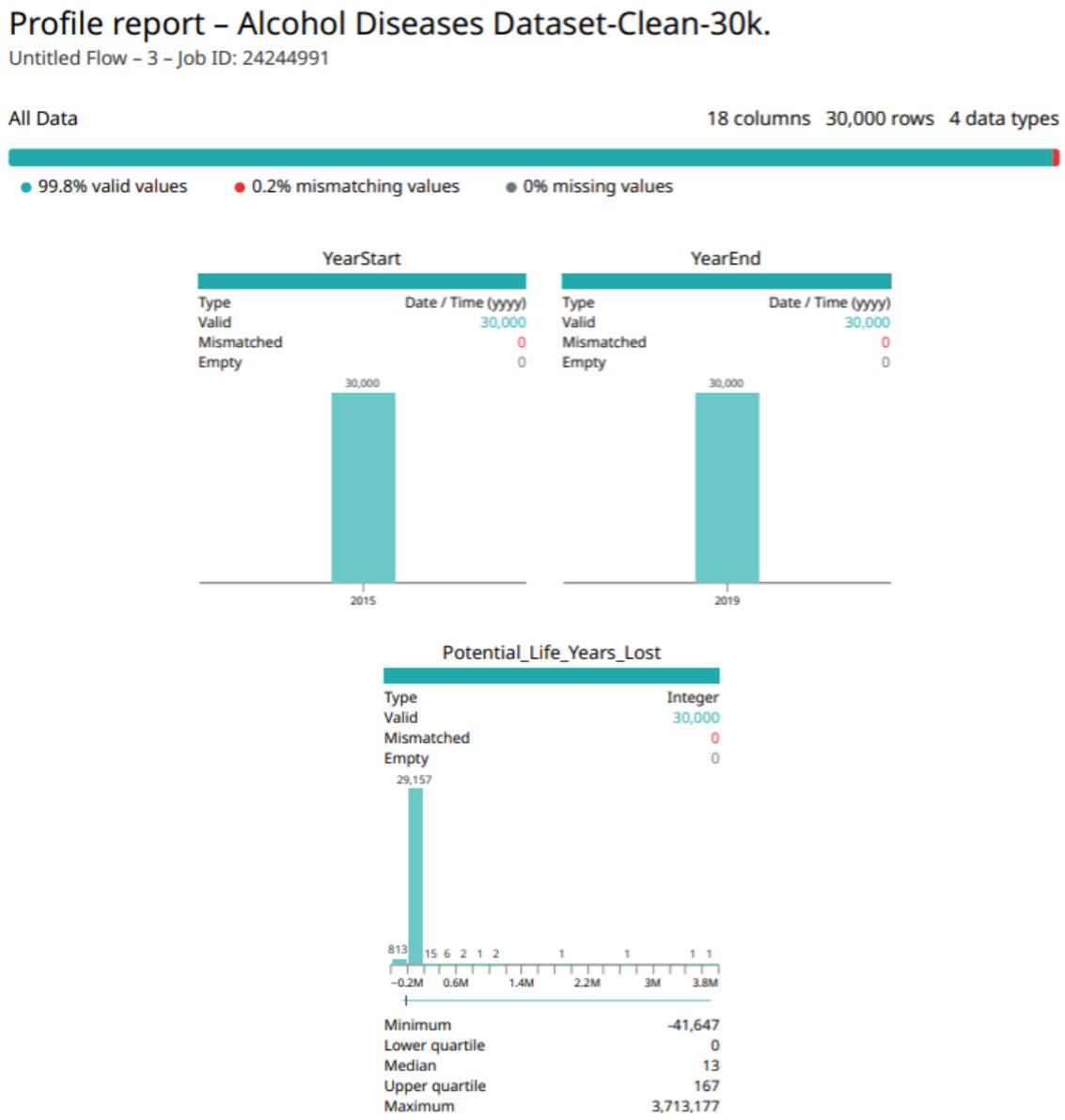
Data_Value

Type	Integer
Valid	30,000
Mismatched	0
Empty	0



Minimum	-41,647
Lower quartile	0
Median	13
Upper quartile	167
Maximum	3,713,177

CLEAN DATA



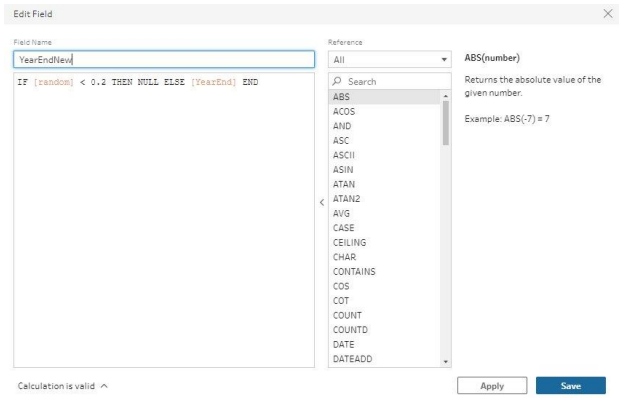
DATA UNCLEANING STEPS

Step1: Missing Values

To assess how well the data analysis techniques can handle scenarios involving missing values, we can purposefully exclude some observations or add missing values to particular variables. We can learn more about the limitations and dependability of the analytical procedures by purposefully "uncleaning" or adding missing variables. This enables us to make more educated decisions about the robustness of their methods in real-world applications. The procedures used on our dataset are as follows:

1-Create RANDOM () field using New Calculated field
2-Create a new calculated field using IF [Random] < 0.2 THEN NULL ELSE [Your Field] END (if we want threshold as 20%)

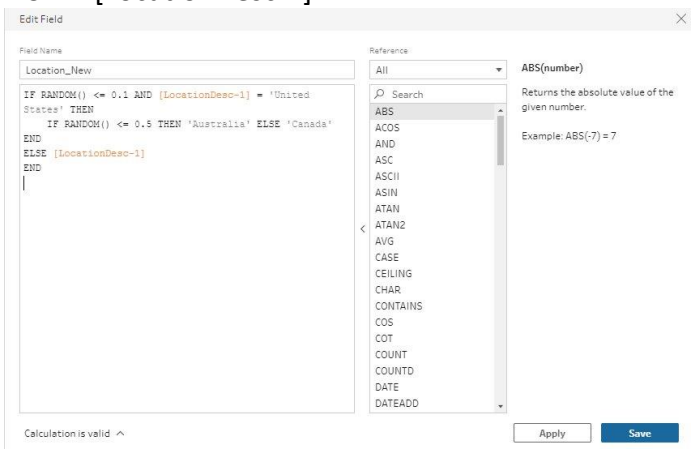
3-Delete the old column
4-Rename the new column with the old column's name
5-Delete the RANDOM field



Step2: Incorrect values

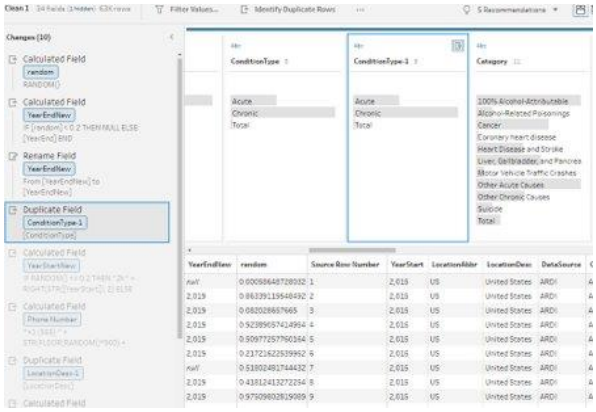
We should model situations in which wrong values occur. For the purpose of assessing how well their statistical models or machine learning algorithms can recognise and manage such disparities, we may include outliers, anomalies, or numbers with known mistakes. This method aids in evaluating how robust data analysis techniques are to anomalies and obstacles encountered in the actual world.

IF RANDOM () <= 0.1 AND [LocationDesc-1] = 'United States' THEN IF RANDOM () <= 0.5 THEN 'Australia' ELSE 'Canada' END
ELSE [LocationDesc-1] END



Step3: Duplicate Entries

Intentionally duplicating current records and adding them back into the dataset is what this method entails, which can lead to redundancy and complicate data processing and analysis. To assess the robustness of the data handling methods, we include duplicate entries, particularly in situations where failures in data integration or collection procedures frequently result in duplicate data.
Select the column click on menu and select duplicate (Condition type).

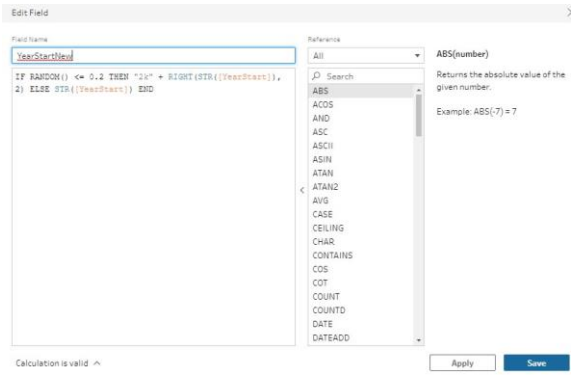


Step4: Inconsistent Formats like date-time

purposely changing date-time representations within a dataset, introducing formats like "YYYY-MM-DD," "MM/DD/YYYY," or other variations. This intentional uncleaning helps assess the robustness of data processing and analysis tools, especially those sensitive to date-time formats. Analysing data with diverse date-time representations is particularly relevant in scenarios where information comes from different systems or sources with varying conventions.

Column changed YearStart

-IF RANDOM () <= 0.2 THEN "2k" + RIGHT(STR([YearStart]), 2) ELSE STR([YearStart]) END

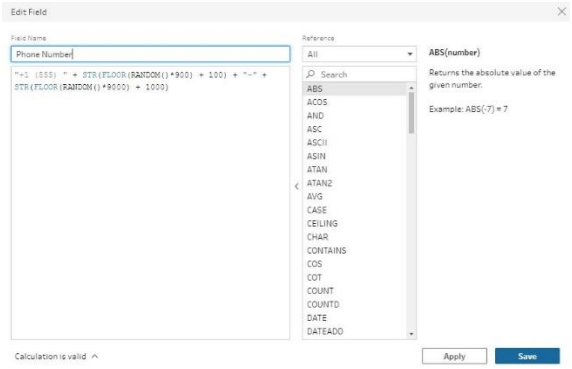


Step5: Privacy Violation - sensitive data in the dataset

Intentionally including personally identifiable information (PII) or other private information about people might be considered a privacy breach. As they violate people's right to privacy and may even violate legal requirements like data protection laws, such actions give rise to grave ethical questions. Prioritising data privacy and confidentiality is crucial for conducting ethical research and data analysis.

Added Phone Numbers

"+1 (555)
" + STR(FLOOR(RANDOM ()*900) + 100) + "-" + STR(FLOOR(RANDOM()*9000) + 1000)

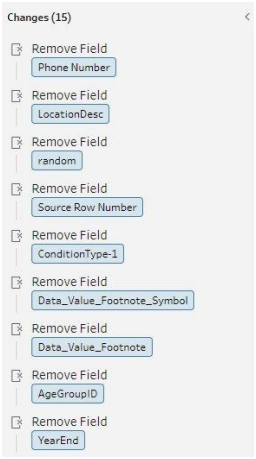


DATA CLEANING STEPS

1. Removed Phone numbers since it is personal info and violates privacy.
2. Removed Location Desc Duplication of column.
3. Removed Random – Irrelevant Column
4. Removed Source row number – duplication.
5. Removed Condition type1 – duplication.
6. Removed Data Value Footnote Symbol- It has Null value.
7. Removed Data Value Footnote – It has Null value
8. Removed Age group ID- Irrelevant data.(Age0, Age1 etc)
9. Year Start – making data consistent.
10. Year End- Filling null values
11. Location mapping – mapping location with correct location ID
12. Removed Data Source – One Value(ARDI)
13. Removed Data_Value_Unit- One Value(Years of Potential life lost)
14. Data_Value_Type- One Value (5 years average)
15. Data Value Unit Id- One Value (YPLL)
16. Data Value UnitType Id- One Value (5YEARSAVG)
17. Removed (Effect ID ,Consumption data, SexId, AgeCategoryID) – Redundant Data
18. Rename Data Value Field to Potential_life_Years_Lost

Step1: Eliminating unwanted columns

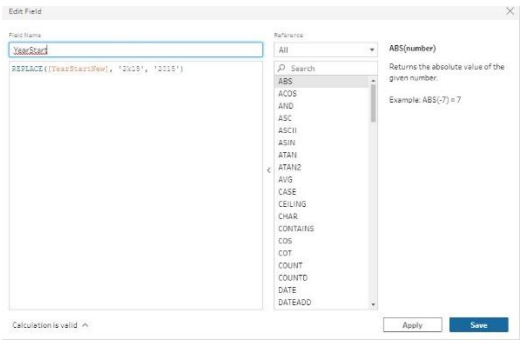
Eliminating unnecessary columns, which are frequently characterised by their repetition, irrelevance, or lack of relevant information, is a step in data cleaning process. This is a crucial step since it reduces dimensionality, makes data more interpretable, boosts computational performance, fixes collinearity problems, and improves data quality overall.



Step2: Standardize

Standardization is a crucial data cleaning step that involves transforming the numerical values of different variables in a dataset to a common scale. This process ensures that the data adhere to a consistent and comparable metric, preventing issues arising from differing units or scales.

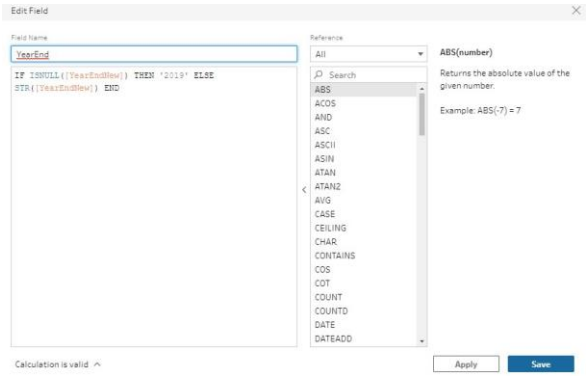
YearStart -REPLACE([YearStartNew], '2k15', '2015')



Step3: Replacing null values

Imputation, or the replacement of null values in data, is a stage in data cleansing that keeps a dataset entire and accurate. It is imperative to address null values resulting from errors or gaps in data gathering to guarantee the validity of the analyses that follow. Common approaches include forward or backward fill for time-series data, mean, median, or mode imputation for numerical variables, and more sophisticated techniques like predictive modelling with regression or machine learning algorithms.

IF ISNULL([YearEndNew]) THEN '2019' ELSE STR([YearEndNew]) END

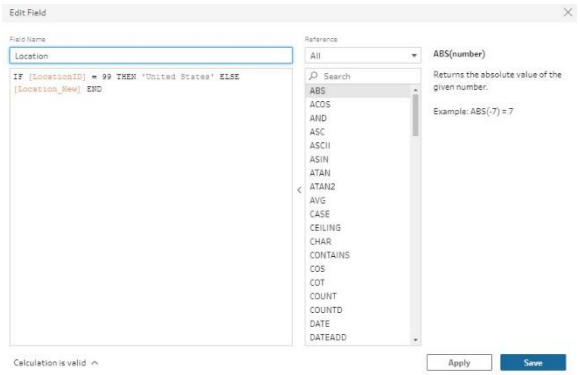


Step4: Replacing text

This process involves correcting misspellings, standardizing formats, removing special characters, and handling case sensitivity to ensure uniformity and accuracy. Addressing synonyms, abbreviations, and inconsistent language contributes to a more coherent dataset, reducing ambiguity and enhancing the quality of subsequent analyses.

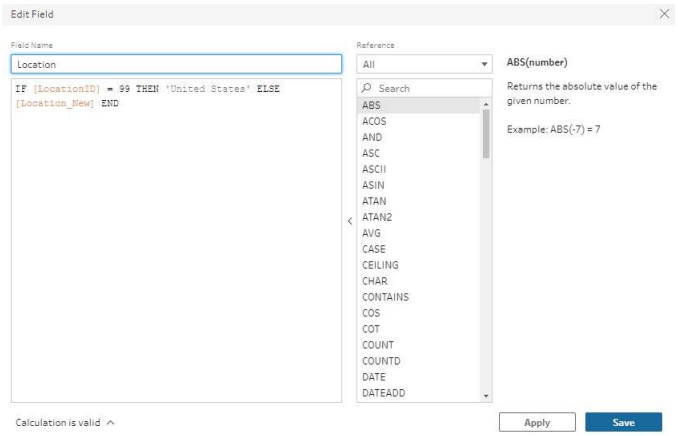
Replacing location name with correct name using location id -

IF [LocationID] = 99 THEN 'United States' ELSE [Location_New] END



Step5: Renaming Column Name

Certain columns were found to have irrelevant or ambiguous column names that did not accurately express the information they contained throughout the data cleansing process. Renaming these columns was an essential step in improving the dataset's readability and clarity. In addition to improving the dataset's readability for upcoming analyses, this calculated renaming made sure that the column names appropriately reflected the type of data they contained.



MEETING TIMELINES

Selecting the dataset topic (22-02-2024):

Minutes of Meeting

The process of selecting a topic often involves collaborative discussions and considerations within a group. In our case, we engaged in conversations with our peers to explore potential database topics. Sanath was tasked with gathering the dataset, and he produced three or four datasets. After consulting with our colleagues and taking into account every dataset, we decided on "Alcohol-Related Disease Impact" as the subject of our study. This choice was motivated by several factors, primarily the inherent research potential that the topic offers. Exploring the intersection of alcohol consumption and its health consequences opens avenues for understanding the complexities of public health. We agreed to talk about which tool could be used to complete our project's remaining steps at our next meeting.

Selecting data processing tool and data profiling (27-02-2024)

Minutes of Meeting

We as a group decided the technology to use for data cleaning and uncleaning on the same day that we discussed how to profile the data. For preparing the data Tableau Prep was used and to profile the data Trifacta by GCP was used. Tableau Prep's seamless integration with Tableau for visualization, its user-friendly interface, and interactive data exploration make it an attractive choice for preparing data within the Tableau ecosystem. On the other hand, Trifacta is renowned for its advanced data profiling features, utilizing machine learning for

automated detection of data patterns and anomalies. Our teammate Spoorti took up the task to profile the data, each column of our dataset was profiled in GCP Trifacta and later the values were added to the data profiling table.

Data quality (*28-02-2024)*

Minutes of Meeting

One of our group's members, Aniket, has taken up the duty of assessing the quality of the data. This entails a thorough analysis that includes Completeness to make sure all necessary information is included; Accuracy, certifying data precision; Conformity, assuring conformity to standards; Consistency, confirming uniformity across datasets; Validity, confirming adherence to stated rules; Currency measures the information's timeliness; Timeliness measures the data's relevancy; and Uniqueness measures the absence of duplicate entries. Aniket's attention to these important details shows that he takes great care to keep accurate data, which is essential for well-informed decision-making and overall project operating efficiency.

Uncleaned the data (01-03-2024)

Minutes of Meeting

Data uncleaning task was taken by Kshema. Kshema worked in our experiment by intentionally making our data messy to see how well our tools can handle challenges. Added missing values, wrong information, repeated entries, different ways of writing dates, and even private details. This helped test if the analysis methods can deal with real-world data issues. However, it's crucial to remember that messing with private information is not okay. Our goal was to learn how to handle tricky data situations better and improve our analysis methods. We need to be careful and ethical when dealing with people's private details in any research.

Cleaned the data (03-02-2024)

Minutes of Meeting

In the data cleaning process, streamlining of our dataset by removing unnecessary columns, reducing complexity, and focusing on relevant information took place. Standardization ensured consistent scales for numerical data, enhancing analysis accuracy. Addressing null values involved replacing missing entries through methods like mean or mode imputation, ensuring completeness. Text data underwent

meticulous cleaning, correcting errors, and standardizing formats for improved uniformity. These steps collectively were performed by Vandit and aimed to refine our dataset, promoting reliability and facilitating more accurate and meaningful analyses.

Discussions on presentation and project report (05-03-2024)

We discussed the presentation and project report together, going over important results, improving interpretations of the analyses, and working through any issues. Clarity was key when discussing our findings and observations. The project's goals and results were successfully communicated in the final presentation and report thanks to open communication that encouraged honest feedback.

INDIVIDUAL ROLES IN PROJECT

Task	Deadline	Member responsible
Dataset Collection	23-02-2024	Sanath Haritsa
Data Profiling	27-02-2024	Spoorti Basarkod Math
Data Quality	29-02-2024	Aniket Ghetla
Data Uncleaning	02-03-2024	Kshema Iliger
Data Cleaning	04-02-2024	Vandit Bhalla
Presentation	05-03-2024	All Members