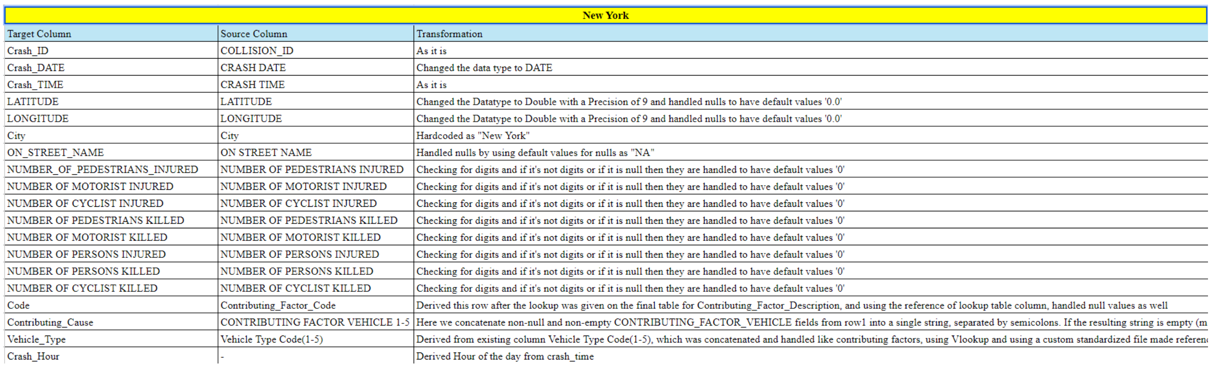
Group 14 - Mapping Document

Prasad Gavas
Shantanu Mahakal
Prarthana Shetty
Soham Shah

Mapping Table:
https://docs.google.com/spreadsheets/d/19v6rCAqg4VMjGnHIaNyz7hUGV-HvxY6nRa61QcVzruo/edit?usp=sharing



New York		
Target Column	Source Column	Transformation
Crash_ID	COLLISION_ID	As it is
Crash_DATE	CRASH DATE	Changed the data type to DATE
Crash_TIME	CRASH TIME	As it is
LATITUDE	LATITUDE	Changed the Datatype to Double with a Precision of 9 and handled nulls to have default values '0.0'
LONGITUDE	LONGITUDE	Changed the Datatype to Double with a Precision of 9 and handled nulls to have default values '0.0'
City	City	Hardcoded as "New York"
ON_STREET_NAME	ON STREET NAME	Handled nulls by using default values for nulls as "NA"
NUMBER_OF_PEDESTRIANS_INJURED	NUMBER OF PEDESTRIANS INJURED	Checking for digits and if it's not digits or if it is null then they are handled to have default values '0'
NUMBER OF MOTORIST INJURED	NUMBER OF MOTORIST INJURED	Checking for digits and if it's not digits or if it is null then they are handled to have default values '0'
NUMBER OF CYCLIST INJURED	NUMBER OF CYCLIST INJURED	Checking for digits and if it's not digits or if it is null then they are handled to have default values '0'
NUMBER OF PEDESTRIANS KILLED	NUMBER OF PEDESTRIANS KILLED	Checking for digits and if it's not digits or if it is null then they are handled to have default values '0'
NUMBER OF MOTORIST KILLED	NUMBER OF MOTORIST KILLED	Checking for digits and if it's not digits or if it is null then they are handled to have default values '0'
NUMBER OF PERSONS INJURED	NUMBER OF PERSONS INJURED	Checking for digits and if it's not digits or if it is null then they are handled to have default values '0'
NUMBER OF PERSONS KILLED	NUMBER OF PERSONS KILLED	Checking for digits and if it's not digits or if it is null then they are handled to have default values '0'
NUMBER OF CYCLIST KILLED	NUMBER OF CYCLIST KILLED	Checking for digits and if it's not digits or if it is null then they are handled to have default values '0'
Code	Contributing_Factor_Code	Derived this row after the lookup was given on the final table for Contributing_Factor_Description, and using the reference of lookup table column, handled null values as well
Contributing_Cause	CONTRIBUTING FACTOR VEHICLE 1-5	Here we concatenate non-null and non-empty CONTRIBUTING_FACTOR_VEHICLE fields from row1 into a single string, separated by semicolons. If the resulting string is empty (m
Vehicle_Type	Vehicle Type Code(1-5)	Derived from existing column Vehicle Type Code(1-5), which was concatenated and handled like contributing factors, using Vlookup and using a custom standardized file made referen
Crash_Hour	-	Derived Hour of the day from crash_time

Mapping Document for Transformation of Columns in NYC Dataset

1. Collision_ID:
  - Transformation: As it is
  - Explanation: - The column retains its original name and data type, maintaining consistency

2. crash date:
  - Transformation: Renamed to CRASH DATE and changed the data type to DATE
  - Explanation: Renaming the column to CRASH DATE enhances its descriptive nature, and parsed the data type from String to DATE to ensure it accurately represents dates

3. crash time:
  - Transformation: No changes, remains as CRASH TIME
  - Explanation: The column retains its original name and data type, maintaining consistency

4. Latitude:
  - Transformation: Renamed to LATITUDE and changed the data type to Double with a Precision of 9
  - Explanation: Renaming the column to LATITUDE improves clarity, and changing the data type accommodates decimal values with increased precision

5. Longitude:
  - Transformation: Renamed to LONGITUDE and changed the data type to Double with a Precision of 9
  - Explanation: Renaming the column to LONGITUDE enhances clarity, and changing the data type ensures accurate representation of geographic coordinates

6. City:
  - Transformation: Hardcoded as "New York".
  - Explanation: The values in this column were replaced with a constant value "New York" to indicate the location of incidents and **source of data**

7. Number of Pedestrians Injured, Number of Motorists Injured, Number of Cyclists injured, No of Pedestrians Killed, Number of Motorists Killed, number of fatalities, Cyclist Killed:

- Transformation: No changes, remain as they are
- Explanation: These columns retain their original names and data types

8. Contributing factor ID:

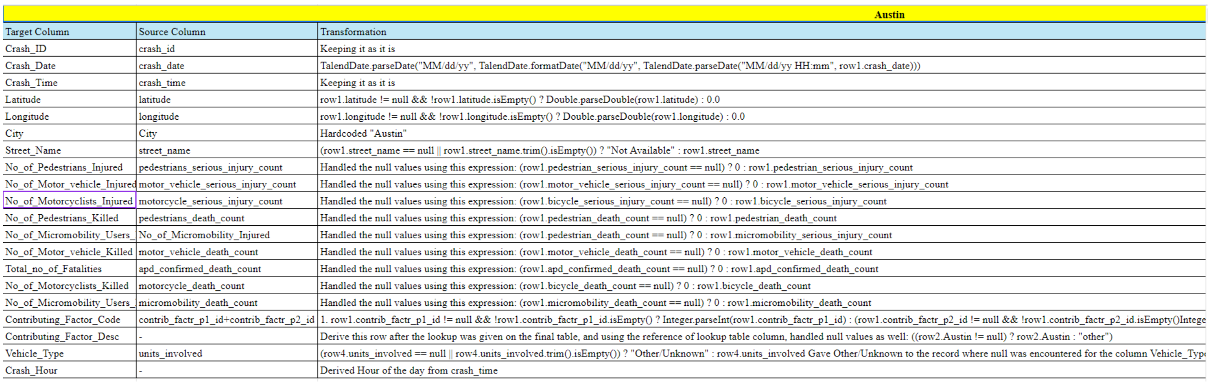
- Transformation: Renamed to Contributing_Factor_Code, derived by carrying out a lookup against a provided file for Contributing Factors.
- Explanation: Renaming the column to Contributing_Factor_Code enhances clarity, and its values were obtained by performing a lookup operation against a custom external file containing Contributing Factors.

9. Contributing Factor Description:

-Here we concatenate non-null and non-empty CONTRIBUTING_FACTOR_VEHICLE1 to 5 fields from row1 into a single string, separated by semicolons. If the resulting string is empty (meaning all contributing factor fields were null or empty), it defaults to "NA". Otherwise, it cleans up any extra semicolons before returning the concatenated string.

10. Vehicle_Type_Codes (1-5):

- Transformation: Combined into one column 'Vehicle_Type_Codes1 to 5' and standardized through a lookup file.
- Explanation: The vehicle type codes were aggregated to streamline the dataset, with null values being handled similarly to the contributing factors, ensuring consistency and clarity in vehicle type reporting. The lookup file was utilized to align disparate vehicle descriptions to a standardized set of categories.



Austin		
Target Column	Source Column	Transformation
Crash_ID	crash_id	Keeping it as it is
Crash_Date	crash_date	TalendDate.parseDate("MM/dd/yy", TalendDate.formatDate("MM/dd/yy", TalendDate.parseDate("MM/dd/yy HH:mm", row1.crash_date)))
Crash_Time	crash_time	Keeping it as it is
Latitude	latitude	row1.latitude != null && !row1.latitude.isEmpty() ? Double.parseDouble(row1.latitude) : 0.0
Longitude	longitude	row1.longitude != null && !row1.longitude.isEmpty() ? Double.parseDouble(row1.longitude) : 0.0
City	City	Hardcoded "Austin"
Street_Name	street_name	(row1.street_name == null || row1.street_name.trim().isEmpty()) ? "Not Available" : row1.street_name
No_of_Pedestrians_Injured	pedestrians_serious_injury_count	Handled the null values using this expression: (row1.pedestrian_serious_injury_count == null) ? 0 : row1.pedestrian_serious_injury_count
No_of_Motor_vehicle_Injured	motor_vehicle_serious_injury_count	Handled the null values using this expression: (row1.motor_vehicle_serious_injury_count == null) ? 0 : row1.motor_vehicle_serious_injury_count
No_of_Motorcyclists_Injured	motorcycle_serious_injury_count	Handled the null values using this expression: (row1.bicycle_serious_injury_count == null) ? 0 : row1.bicycle_serious_injury_count
No_of_Pedestrians_Killed	pedestrians_death_count	Handled the null values using this expression: (row1.pedestrian_death_count == null) ? 0 : row1.pedestrian_death_count
No_of_Micromobility_Users_	No_of_Micromobility_Injured	Handled the null values using this expression: (row1.pedestrian_death_count == null) ? 0 : row1.micromobility_serious_injury_count
No_of_Motor_vehicle_Killed	motor_vehicle_death_count	Handled the null values using this expression: (row1.motor_vehicle_death_count == null) ? 0 : row1.motor_vehicle_death_count
Total_no_of_Fatalities	apd_confirmed_death_count	Handled the null values using this expression: (row1.apd_confirmed_death_count == null) ? 0 : row1.apd_confirmed_death_count
No_of_Motorcyclists_Killed	motorcycle_death_count	Handled the null values using this expression: (row1.bicycle_death_count == null) ? 0 : row1.bicycle_death_count
No_of_Micromobility_Users_	micromobility_death_count	Handled the null values using this expression: (row1.micromobility_death_count == null) ? 0 : row1.micromobility_death_count
Contributing_Factor_Code	contrib_factr_p1_id+contrib_factr_p2_id	1. row1.contrib_factr_p1_id != null && !row1.contrib_factr_p1_id.isEmpty() ? Integer.parseInt(row1.contrib_factr_p1_id) : (row1.contrib_factr_p2_id != null && !row1.contrib_factr_p2_id.isEmpty()Intege
Contributing_Factor_Desc	-	Derive this row after the lookup was given on the final table, and using the reference of lookup table column, handled null values as well: ((row2.Austin != null) ? row2.Austin : "other")
Vehicle_Type	units_involved	(row4.units_involved == null || row4.units_involved.trim().isEmpty()) ? "Other/Unknown" : row4.units_involved Gave Other/Unknown to the record where null was encountered for the column Vehicle_Type
Crash_Hour	-	Derived Hour of the day from crash_time

Mapping Document for Transformation of Columns in Austin Dataset

1.crash_id:

- Transformation: Renamed to CrashID.
- Explanation: The column name was modified to CrashID for clarity and adherence to naming conventions.

2. crash date:

- Transformation: Renamed to CRASH DATE and changed the data type to DATE. TalendDate.parseDate("MM/dd/yy", TalendDate.formatDate("MM/dd/yy", TalendDate.parseDate("MM/dd/yy HH:mm", row1.crash_date)))
- Explanation: Renaming the column to CRASH DATE enhances its descriptive nature, and changing the data type to DATE ensures it accurately represents dates.

3.crash time:

- Transformation: Renamed to Crash Time and changed the data type to String.
- Explanation: Renamed the

4.City:

- Transformation: Hardcoded as "Austin".
- Explanation: The values in this column were replaced with a constant value "Austin" to indicate the location of incidents.

5.Latitude:

- Transformation: Renamed to LATITUDE and changed the data type to Double with a Precision of 9.
- Explanation: Renaming the column to LATITUDE improves clarity, and changing the data type accommodates decimal values with increased precision.

6.Longitude:

- Transformation: Renamed to LONGITUDE and changed the data type to Double with a Precision of 9.
- Explanation: Renaming the column to LONGITUDE enhances clarity, and changing the data type ensures accurate representation of geographic coordinates.

7. Pedestrain_serious_injury_count:

- Transformation: Renamed to Number_of_Pedestrians_Injured and handled null values or checked isEmpty() and passed it a value of 0.
- Explanation: The column Pedestrain_serious_injury_count has been renamed to Number_of_Pedestrians_Injured to improve clarity. Any null or empty values in this column are replaced with zero, ensuring that all records have a valid, numerical entry for the count of seriously injured pedestrians.

8. Pedestrian_death_count:
    - Transformation: Renamed to Number_of_Pedestrains_Killed and handled null values or checked isEmpty() and passed it a value of 0.
    - Explanation: The column Pedestrian_death_count has been renamed to Number_of_Pedestrians_Killed to improve clarity. Any null or empty values in this column are replaced with zero, ensuring that all records have a valid, numerical entry for the count of pedestrians killed.

9. motor_vehicle_serious_injury_count:
    - Transformation: Renamed to Number_of_Motorists_Injured and handled null values or checked isEmpty() and passed it a value of 0.
    - Explanation: The column motor_vehicle_serious_injury_count has been renamed to Number_of_Motorists_injured to improve clarity. Any null or empty values in this column are replaced with zero, ensuring that all records have a valid, numerical entry for the count of motor vehicle killed.

10. Motor_vehicle_death_count:
    - Transformation: Renamed to Number_of Motorists_Killed and handled null values or checked isEmpty() and passed it a value of 0.
    - Explanation: The column Motor_vehicle_death_count has been renamed to Number_of_Motorists_Killed to improve clarity. Any null or empty values in this column are replaced with zero, ensuring that all records have a valid, numerical entry for the count of motorists killed.

11. Motorcycle_serious_injury_count:
    - Transformation: Number_of_Cyclists_Injured
    - Explanation: The column Bicycle_serious_injury_count has been renamed to Number_of_Cyclists_Killed to improve clarity. Any null or empty values in this column are replaced with zero, ensuring that all records have a valid, numerical entry for the count of cyclists killed.

12. Contributing_Factor_Code

Contrib_factor_p1_id

1. row1.contrib_factr_p1_id != null && !row1.contrib_factr_p1_id.isEmpty() ? Integer.parseInt(row1.contrib_factr_p1_id) : (row1.contrib_factr_p2_id != null && !row1.contrib_factr_p2_id.isEmpty()Integer.parseInt(row1.contrib_factr_p2_id) : 101).

contrib_factr_p2_id

2. (row1.contrib_factr_p1_id == null || row1.contrib_factr_p1_id.isEmpty()) && (row1.contrib_factr_p2_id != null && !row1.contrib_factr_p2_id.isEmpty()) ? 8888 :

(row1.contrib_factr_p2_id == null || row1.contrib_factr_p2_id.isEmpty() ? 8888 : Integer.parseInt(row1.contrib_factr_p2_id))

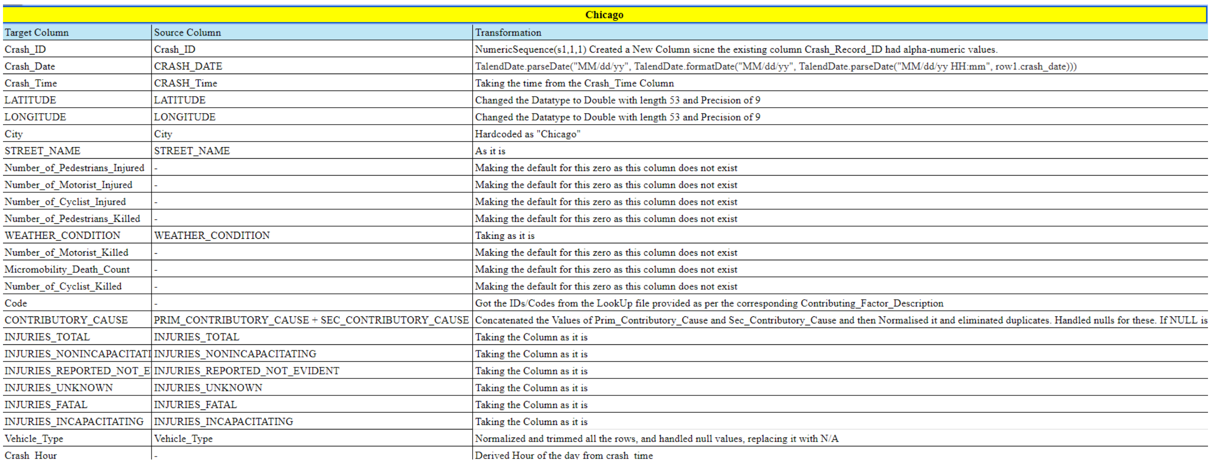
13. Contributing_Factor_Desc
Took a lookup code description from
After lookup this column is added to the final database table : Austin_p_final
Derive this row after the lookup was given on the final table, and using the reference of lookup table column, handled null values as well: ((row2.Austin != null) ? row2.Austin : "other")

14. Vehicle_Type
Normalized all the units_involved in the table
(row4.units_involved == **null** || row4.units_involved.trim().isEmpty()) ? "Other/Unknown" : row4.units_involved
Gave Other/Unknown to the record where null was encountered for the column Vehicle_Type

15. Total_no_of_Fatalities
Handled the null values using this expression: (row1.apd_confirmed_death_count == null) ? 0 : row1.apd_confirmed_death_count

16. Street_name
Handled the null values using this expression: (row1.street_name == null || row1.street_name.trim().isEmpty()) ? "Not Available" : row1.street_name

17. Micromobility_death_count
Handled the null values using this expression: (row1.micromobility_death_count == null) ? 0 : row1.micromobility_death_count



Chicago

Target Column	Source Column	Transformation
Crash_ID	Crash_ID	NumericSequence(s1,1,1) Created a New Column sicne the existing column Crash_Record_ID had alpha-numeric values.
Crash_Date	CRASH_DATE	TalendDate.parseDate("MM/dd/yy", TalendDate.formatDate("MM/dd/yy", TalendDate.parseDate("MM/dd/yy HH:mm", row1.crash_date)))
Crash_Time	CRASH_Time	Taking the time from the Crash_Time Column
LATITUDE	LATITUDE	Changed the Datatype to Double with length 53 and Precision of 9
LONGITUDE	LONGITUDE	Changed the Datatype to Double with length 53 and Precision of 9
City	City	Hardcoded as "Chicago"
STREET_NAME	STREET_NAME	As it is
Number_of_Pedestrians_Injured	-	Making the default for this zero as this column does not exist
Number_of_Motorist_Injured	-	Making the default for this zero as this column does not exist
Number_of_Cyclist_Injured	-	Making the default for this zero as this column does not exist
Number_of_Pedestrians_Killed	-	Making the default for this zero as this column does not exist
WEATHER_CONDITION	WEATHER_CONDITION	Taking as it is
Number_of_Motorist_Killed	-	Making the default for this zero as this column does not exist
Micromobility_Death_Count	-	Making the default for this zero as this column does not exist
Number_of_Cyclist_Killed	-	Making the default for this zero as this column does not exist
Code	-	Got the IDs/Codes from the LookUp file provided as per the corresponding Contributing_Factor_Description
CONTRIBUTORY_CAUSE	PRIM_CONTRIBUTORY_CAUSE + SEC_CONTRIBUTORY_CAUSE	Concatenated the Values of Prim_Contributory_Cause and Sec_Contributory_Cause and then Normalised it and eliminated duplicates. Handled nulls for these. If NULL is
INJURIES_TOTAL	INJURIES_TOTAL	Taking the Column as it is
INJURIES_NONINCAPACITATI	INJURIES_NONINCAPACITATING	Taking the Column as it is
INJURIES_REPORTED_NOT_E	INJURIES_REPORTED_NOT_EVIDENT	Taking the Column as it is
INJURIES_UNKNOWN	INJURIES_UNKNOWN	Taking the Column as it is
INJURIES_FATAL	INJURIES_FATAL	Taking the Column as it is
INJURIES_INCAPACITATING	INJURIES_INCAPACITATING	Taking the Column as it is
Vehicle_Type	Vehicle_Type	Normalized and trimmed all the rows, and handled null values, replacing it with N/A
Crash_Hour	-	Derived Hour of the day from crash_time

Mapping Document for Transformation of Columns in Chicago Dataset

1.crash_id:

- Transformation: Renamed to CrashID.
- Explanation: The column name was modified to CrashID for clarity and adherence to naming conventions.

2. crash date:

- Transformation: Renamed to CRASH DATE and changed the data type to DATE.
  TalendDate.parseDate("MM/dd/yy", TalendDate.formatDate("MM/dd/yy", TalendDate.parseDate("MM/dd/yy HH:mm", row1.crash_date)))
- Explanation: Renaming the column to CRASH DATE enhances its descriptive nature, and changing the data type to DATE ensures it accurately represents dates.

3.crash time:

- Transformation: Renamed to Crash Time and changed the data type to String.
- Explanation: Renamed the

4.City:

- Transformation: Hardcoded as "Chicago".
- Explanation: The values in this column were replaced with a constant value "Austin" to indicate the location of incidents.

5.Latitude:

- Transformation: Renamed to LATITUDE and changed the data type to Double with a Precision of 9.
- Explanation: Renaming the column to LATITUDE improves clarity, and changing the data type accommodates decimal values with increased precision.

6.Longitude:

- Transformation: Renamed to LONGITUDE and changed the data type to Double with a Precision of 9.
- Explanation: Renaming the column to LONGITUDE enhances clarity, and changing the data type ensures accurate representation of geographic coordinates.

7. Number of Pedestrians Injured :

It is Assigned to 0 by default.

8. Number of Motorists Injured :

It is Assigned to 0 by default.

9. Number of Cyclists injured:

It is Assigned to 0 by default.

10. No of Pedestrians Killed :

It is Assigned to 0 by default.

11. Number of Motorists Killed :

It is Assigned to 0 by default.

12. number of fatalities:

It is Assigned to 0 by default.

13. Cyclist Killed :

It is Assigned to 0 by default.

14. Contributing factor ID :

Got the IDs/Codes from the LookUp file provided as per the corresponding Contributing_Factor_Description

15. Contributing Factor Description :

Concatenated the Values of Prim_Contributory_Cause and Sec_Contributory_Cause and then Normalized it and eliminated duplicates. Handled nulls for these. If NULL is encountered then the value is set to NA.

16. Units_involved :

Making the default for this zero as this column doesnot exist