DATA CLEANING AND PROCESSING

NEW YORK CITY TRAFFIC ACCIDENT

Data cleaning, also known as data cleansing or data scrubbing, is the process of identifying and correcting errors, inconsistencies, and inaccuracies in datasets to improve their quality and reliability. It is a crucial step in the data preparation phase before analysis, as the quality of the insights derived from data analysis depends on the cleanliness of the data

Data cleaning ensures that the dataset is reliable, accurate, and suitable for analysis, ultimately leading to more meaningful and trustworthy results. This process is a fundamental step in maintaining data quality and integrity throughout the data lifecycle.

A thorough data cleaning and preparation process was executed on the key columns of the dataset involved in this analysis. This ensures that my analyses are based on a solid foundation, leading to more reliable results and informed decision-making. The following processes were undertaken to ensure high data integrity, accuracy, and to enhance the overall quality of the dataset.

- **Removing Duplicates:** Identified and removed all duplicate entries to avoid redundancy and ensure data integrity.

```sql
--Identify and Remove Duplicate
WITH CTE
AS
(SELECT       Collision_ID,
              [Date],
              [Time],
              Borough,
              Street_Name,
              Cross_Street,
              Latitude,
              Longitude,
              Contributing_Factor,
              Vehicle_Type,
              Persons_Injured,
              Persons_Killed,
              Pedestrians_Injured,
              Pedestrians_Killed,
              Cyclists_Injured,
              Cyclists_Killed,
              Motorists_Injured,
              Motorists_Killed,
              ROW_NUMBER() OVER(PARTITION BY Collision_ID ORDER BY Collision_ID)
AS Row_No
FROM Tbl_NYC_Traffic_Accidents)
```

```sql
DELETE
FROM CTE
WHERE Row_No > 1
```

100 %

Messages

```
(0 rows affected)

Completion time: 2024-01-20T14:06:15.8245106+01:00
```

The generated result indicates that there are no duplicate values in the NYC traffic accident data.

- **Handling Missing Values:** Identified and addressed missing values on the key columns in New York City data.

```sql
--Total Count of Rows Containing Missing Values on the Key Columns
SELECT        COUNT(*) AS Total_Count_of_Rows
FROM          Tbl_NYC_Traffic_Accidents
WHERE         Collision_ID  IS NULL
              OR
              [Date] IS NULL
              OR
              [Time] IS NULL
              OR
              Borough IS NULL
              OR
              Street_Name IS NULL
              OR
              Contributing_Factor IS NULL
              OR
              Vehicle_Type IS NULL
              OR
              Persons_Killed IS NULL
```

100 %

Results | Messages

	Total_Count_of_Rows
1	8783

The generated results reveal that the NYC traffic data includes 8,783 rows with missing values in one or more key columns.

Missing values in the key columns of this dataset indicate incomplete information. Despite the incompleteness, the rows containing these missing values are crucial for this analysis. Therefore, they were addressed separately and were retained rather than deleted.

- **Handling Missing Values in each Key Column on NYC Traffic Data**

```sql
--Identify Missing Values in Collision ID Column
SELECT        COUNT(*) AS Count_of_Missing_Values
FROM          Tbl_NYC_Traffic_Accidents
WHERE         Collision_ID  IS NULL
```

100 %

Results | Messages

	Count_of_Missing_Values
1	0

The generated result indicates that there are no missing values in Collision ID column.

```sql
--Identify Missing Values in Date Column
SELECT        COUNT(*) AS Count_of_Missing_Values
FROM          Tbl_NYC_Traffic_Accidents
WHERE         [Date] IS NULL
```

100 %

Results | Messages

	Count_of_Missing_Values
1	0

The generated result also indicates that there are no missing values in Date column.

```sql
--Identify Missing Values in Time Column
SELECT COUNT(*) AS Count_of_Missing_Values
FROM   Tbl_NYC_Traffic_Accidents
WHERE  [Time] IS NULL
```

100 %

Results | Messages

	Count_of_Missing_Values
1	0

The generated result also indicates that there are no missing values in Time column.

```sql
--Identify Missing Values in Borough Column
SELECT COUNT(*) AS Count_of_Missing_Values
FROM   Tbl_NYC_Traffic_Accidents
WHERE  Borough IS NULL
```

100 %

Results | Messages

	Count_of_Missing_Values
1	7197

The generated result reveals that among the 8,783 rows with missing values in the NYC traffic accident data, 7,197 pertain to the Borough column. These missing values indicate that the borough where the accident occurred is unknown. Therefore, the missing values in the Borough column were replaced with the term 'Unknown'.

```sql
--Update Missing Value in Borough Column to 'Unknown'
UPDATE          Tbl_NYC_Traffic_Accidents
SET             Borough = 'Unknown'
WHERE           Borough IS NULL
```

100 %

Messages

```
(7197 rows affected)

Completion time: 2024-01-20T15:47:06.6905025+01:00
```

```sql
--Identify Missing Values in street Name Column
SELECT          COUNT(*) AS Count_of_Missing_Values
FROM            Tbl_NYC_Traffic_Accidents
WHERE           Street_Name IS NULL
```

100 %

Results | Messages

	Count_of_Missing_Values
1	363

The result obtained indicates that within the NYC traffic accident data, 363 out of 8,783 rows have missing values in the Street Name column. These missing values signify that the street where the accident occurred is unknown. Consequently, the missing values in the street column were substituted with the term 'Unknown.'

```sql
--Update Missing Value in Street Name Column to 'Unknown'
UPDATE          Tbl_NYC_Traffic_Accidents
SET             Street_Name = 'Unknown'
WHERE           Street_Name IS NULL
```

100 %

Messages

```
(363 rows affected)

Completion time: 2024-01-20T16:00:27.7703607+01:00
```

```sql
--Identify Missing Values in Contributing Factor Column
SELECT        COUNT(*) AS Count_of_Missing_Values
FROM          Tbl_NYC_Traffic_Accidents
WHERE         Contributing_Factor IS NULL
```

100 %

Results | Messages

	Count_of_Missing_Values
1	1287

The obtained result reveals that within the NYC traffic accident data, 1,287 out of 8,783 rows have missing values in the Contributing Factor column. These missing values indicate that the factors contributing to the accident for the designated vehicle are unknown.

The Contributing Factor column already includes the term 'Unspecified,' explicitly denoting cases where the contributing factor is known to be unspecified or unknown. As a result, the missing values were substituted with the term 'Unspecified' instead of using 'Unknown.' This choice aims to avoid ambiguity and potential misrepresentation of the nature of the missing values.

```sql
--Update Missing Value in Contributing Factors Column to Unspecified
UPDATE        Tbl_NYC_Traffic_Accidents
SET           Contributing_Factor = 'Unspecified'
WHERE         Contributing_Factor IS NULL
```

100 %

Messages

```
(1287 rows affected)

Completion time: 2024-01-20T16:30:53.1344197+01:00
```

```sql
--Identify Missing Values in Vehicle Type Column
SELECT        COUNT(*) AS Count_of_Missing_Values
FROM          Tbl_NYC_Traffic_Accidents
WHERE         Vehicle_Type IS NULL
```

100 %

Results | Messages

	Count_of_Missing_Values
1	0

The generated result indicates that there are no missing values in Vehicle Type column.

```sql
--Identify Missing Values in Persons Killed Column

SELECT        COUNT(*) AS Count_of_Missing_Values
FROM          Tbl_NYC_Traffic_Accidents
WHERE         Persons_Killed IS NULL
```

100 %

Results | Messages

	Count_of_Missing_Values
1	0

The generated result indicates that there are no missing values in Persons Killed column.

- **Validation of Spelling and Categorization:** A meticulous review was conducted to validate the correctness of spellings and the accuracy of categorization in NYC Traffic Accident data, ensuring consistency and reliability.

```
-- Identify Inconsistencies in Data such as Variation in Spelling or
Categorization in borough Column
SELECT          Borough
FROM            Tbl_NYC_Traffic_Accidents
GROUP BY        Borough
```

100 %

Results | Messages

	Borough
1	Queens
2	Bronx
3	Brooklyn
4	Staten Island
5	Manhattan
6	Unknown

The generated result affirms the accurate spelling and proper categorization of all values in the Borough column.

```
-- Identify Inconsistencies in Data such as Variation in Spelling or
Categorization in Vehicle Type Column
SELECT          Vehicle_Type
FROM            Tbl_NYC_Traffic_Accidents
GROUP BY        Vehicle_Type
```

100 %

Results | Messages

	Vehicle_Type
1	Passenger Vehicle
2	Bus
3	Bicycle
4	Motorcycle
5	Other
6	Transport
7	Other (Open Passenger)
8	Not Reported
9	Fire Services
10	Taxi
11	Scooter
12	Delivery Vehicle
13	Construction
14	Unknown
15	Utility Vehicle
16	Emergency Services
17	Limousine

The generated result affirms the accurate spelling and proper categorization of all values in the Vehicle Type Column.

- **Correcting Data Types:** Ensured that data types are appropriate for each field (e.g., converting date field to date and time field to time) to facilitate accurate analysis

```
--Correcting Datatype on Key Column

ALTER TABLE   Tbl_NYC_Traffic_Accidents
ALTER COLUMN  Collision_ID
NVARCHAR(50)

ALTER TABLE   Tbl_NYC_Traffic_Accidents
ALTER COLUMN  [Date]
DATE

ALTER TABLE   Tbl_NYC_Traffic_Accidents
ALTER COLUMN  [Time]
TIME

ALTER TABLE   Tbl_NYC_Traffic_Accidents
ALTER COLUMN  Borough
NVARCHAR(50)

ALTER TABLE   Tbl_NYC_Traffic_Accidents
ALTER COLUMN  Street_Name
NVARCHAR(100)

ALTER TABLE   Tbl_NYC_Traffic_Accidents
ALTER COLUMN  Contributing_Factor
NVARCHAR(100)

ALTER TABLE   Tbl_NYC_Traffic_Accidents
ALTER COLUMN  Vehicle_Type
NVARCHAR(50)

ALTER TABLE   Tbl_NYC_Traffic_Accidents
ALTER COLUMN  Persons_Killed
INT
```

The following processes were also carried out on New York City Traffic Accident data during data processing.

- **Added New Columns**

New columns were added to NYC Traffic Accident data. These columns are Year, Month Number, Month Name, Week Name, Week Number, Hours and Minutes. These columns are important in this analysis. They helped in exploring how accidents are dispersed across months to identify any seasonal patterns. They also helped to dissect accident frequency by

both the day of the week and the hour of the day to pinpoint when accidents occur most frequently.

```sql
-- Adding New Columns to NYC Traffic Accident Data
ALTER TABLE   Tbl_NYC_Traffic_Accidents
ADD           Week_Name NVARCHAR(50) NULL,
              Week_Number INT NULL,
              Month_Number INT NULL,
              Month_Name NVARCHAR(50) NULL,
              [YEAR] INT NULL,
              [Hours] INT NULL,
              [Minute] INT NULL

-- Updating New Columns Added to NYC Traffic Accident Data
UPDATE Tbl_NYC_Traffic_Accidents
SET    Week_Name = DATENAME(WEEKDAY, [Date]),
       Week_Number = DATEPART(WEEKDAY, [Date]),
       Month_Number = MONTH([Date]),
       Month_Name = DATENAME(MONTH,[Date]),
       [Year] = YEAR ([Date]),
       [Hours] = DATEPART(HOUR,[Time]),
       [Minute] = DATEPART(MINUTE, [Time])
```

- **Added Primary Key to the Collision ID Columns**

A primary key ensures that each record in the table is uniquely identified. This uniqueness prevents the occurrence of duplicate or identical entries in the table, maintaining data integrity. Adding a primary key to a column is essential for maintaining data quality, optimizing query performance, and ensuring the overall integrity of a relational database.

```sql
--Adding a Primary Key on Collision ID
ALTER TABLE    Tbl_NYC_Traffic_Accidents
ADD CONSTRAINT PK_Collision_ID
PRIMARY KEY (Collision_ID)
```

After the execution of data cleaning and processing, the dataset for this analysis is now well-structured, consistent, and free from significant issues that could hinder analysis or interpretation. The data adheres to standardized formats and consistent naming conventions. All necessary data fields are present and well populated, the data values are accurate, data types are appropriately assigned to each column, and there are no duplicate records.

Detailing the specific processes undertaken provides a clearer picture of the thoroughness in data cleaning and preparation efforts. This detailed approach reinforces the reliability and quality of the dataset for subsequent analyses.