Process ETL | ETL pipeline

Extract : Extracting dataset from (LA) ( downloading 2 CSV files) and combining them using python   
  
Transform- data exploration using python + data transformation using AWS Glue

Load- load into redshift

**Data exploration:**   
Dataset description:

1. head, shape, describe, view all columns

Exploration on each column:

1. to get count of null values
2. to get duplicate values – dropna(), fillna()
3. Grouping - ow many times that specific team won at home, on the road, or at a neutral site. (df.groupby('Wteam')['Wscore'].mean().head())
4. identify outliers - min max mean , plot a boxplot, (sns.boxplot(x=boston\_df['DIS']) )
5. count of all distinct values of a col - df['Season'].value\_counts()

Result:

1. Delete 1st row
2. **AREA:   
   - change col name  
   - change datatype  
   - 2 null values**
3. AREA NAME:  
   **- change col name  
   - change datatype**  
   - value “AREA NAME” exists (2 values). It is to be changed to some default value or null value.
4. Rpt Dist No  
   **- change col name  
   - change datatype**  
   **- 2 null values**  
   - Graphical user interface, text, application

   Description automatically generated **count od occurrences**
5. Crm Cd, Crm Cd Desc,Mocodes  
    **- change col name  
   - change datatype**- Table

   Description automatically generated with medium confidenceText, application

   Description automatically generated  
   **- 2 null values(crm cd)**  
   - split column mo-code, 278395 null values
6. Vict Age  
   **- change col name  
   - change datatype**  
   - outliers – negative, >100, 0,1,2 : replace by null  
   **- 2 null values**
7. Vict Sex  
   **- change col name  
   - change datatype**  
   - incorrect values -  
   Can be replaced by not known  
   **- 2 null values**
8. Location-Cross street  
   **- change col name  
   - change datatype**  
   - merge
9. LAT LON  
   - change col name  
   - change datatype  
   - merge or not (need to check visualization queries)
10. Crm Cd 2, Crm Cd 3, Crm Cd 4  
    - need to drop columns as not needed
11. Status, Status Desc, Crm Cd 1  
    -status:2  
    - Crm Cd 1: 13
12. Premis Cd, Premis Desc  
    - Premis Desc:304  
    - Premis Cd: 58
13. Weapon Used Cd, Weapon Desc   
    - Weapon Used Cd:1617408  
    - Weapon Desc: 1617409

**Data transformation:**

Actions to be performed:

* Combining Similar Categories
* Data Reduction.

Process:

1. Uploaded files to S3 bucket  
   Graphical user interface, text, application

   Description automatically generated
2. Created a new database in AWS Glue  
   Graphical user interface, text, application, email

   Description automatically generated
3. Created crawler to load the CSV files into this database run the crawler.

Graphical user interface, text, application

Description automatically generated

1. Run the crawler to load data.  
   Graphical user interface, text, application, email

   Description automatically generated
2. View the database to see uploaded files.  
   Graphical user interface, application

   Description automatically generated
3. Load data to Glue