```
============================================================
Step 1:DATA PROCESSING & CLEANING
============================================================

1. Loading Data
----------------------------------------
Found 4 CSV files
  Loaded Open_Restaurants_Inspections: 81,553 rows, 20 columns
  Loaded Motor_Vehicle_Collisions_-_Crashes: 2,225,404 rows, 29 columns
  Loaded NYC_street_segment: 90,327 rows, 54 columns
  Loaded Facilities_Database_20251208: 34,708 rows, 38 columns

2. Data Overview
----------------------------------------

Dataset: Open_Restaurants_Inspections
  Total entries: 81,553
  Features: 20
  Missing values: 220,367
  object: 10 columns
  float64: 8 columns
  int64: 2 columns

Dataset: Motor_Vehicle_Collisions_-_Crashes
  Total entries: 2,225,404
  Features: 29
  Missing values: 19,026,063
  object: 18 columns
  int64: 7 columns
  float64: 4 columns

Dataset: NYC_street_segment
  Total entries: 90,327
  Features: 54
  Missing values: 157,346
  float64: 34 columns
  int64: 15 columns
  object: 5 columns

Dataset: Facilities_Database_20251208
  Total entries: 34,708
```

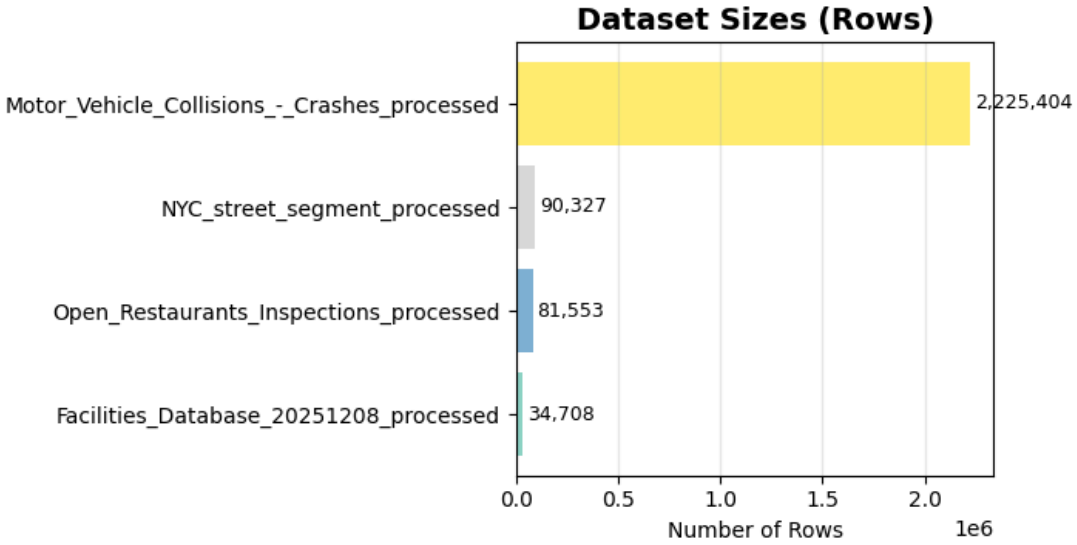
```
  Features: 38
  Missing values: 80,054
  object: 27 columns
  float64: 6 columns
  int64: 5 columns

3. Cleaning Data
----------------------------------------

Processing Open_Restaurants_Inspections:
  Removed 2 columns with >50% missing data
  Missing values: 220,367 → 0
  Capped outliers in 5 numeric columns
  Final shape: 18 columns

Processing Motor_Vehicle_Collisions_-_Crashes:
  Removed 7 columns with >50% missing data
  Missing values: 19,026,063 → 0
  Found 2 date-related columns
  Capped outliers in 6 numeric columns
  Final shape: 30 columns

Processing NYC_street_segment:
  Removed 2 columns with >50% missing data
  Missing values: 157,346 → 0
  Capped outliers in 31 numeric columns
  Final shape: 52 columns

Processing Facilities_Database_20251208:
  Removed 1 columns with >50% missing data
  Missing values: 80,054 → 0
  Capped outliers in 6 numeric columns
  Final shape: 37 columns

4. Preparing Train/Test Splits
----------------------------------------

Creating split for Open_Restaurants_Inspections:
  Training samples: 65,242
  Testing samples: 16,311
  Target variable: RestaurantInspectionID
```

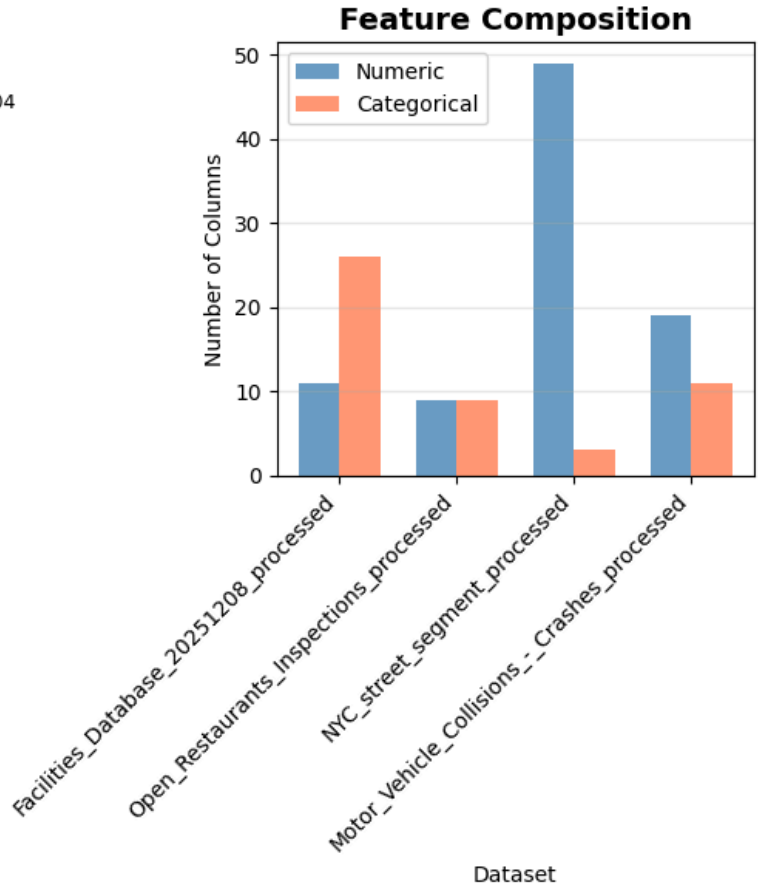
```
Creating split for Motor_Vehicle_Collisions_-_Crashes:
  Training samples: 1,780,323
  Testing samples: 445,081
  Target variable: LATITUDE

Creating split for NYC_street_segment:
  Training samples: 72,261
  Testing samples: 18,066
  Target variable: physical

Creating split for Facilities_Database_20251208:
  Training samples: 27,766
  Testing samples: 6,942
  Target variable: borocode

5. Saving Results
----------------------------------------

Saving processed data:
  ✓ Open_Restaurants_Inspections_processed.csv (81,553 rows)
  ✓ Motor_Vehicle_Collisions_-_Crashes_processed.csv (2,225,404 rows)
  ✓ NYC_street_segment_processed.csv (90,327 rows)
  ✓ Facilities_Database_20251208_processed.csv (34,708 rows)
  ✓ Open_Restaurants_Inspections_train.csv, Open_Restaurants_Inspections_test.csv
  ✓ Motor_Vehicle_Collisions_-_Crashes_train.csv, Motor_Vehicle_Collisions_-_Crashes_test.csv
  ✓ NYC_street_segment_train.csv, NYC_street_segment_test.csv
  ✓ Facilities_Database_20251208_train.csv, Facilities_Database_20251208_test.csv

Summary report saved: /Users/saravananmohanakrishnan/Downloads/dataset/reports/processing_summary.txt

============================================================
PROCESSING COMPLETE
============================================================

Processed 4 datasets
Output saved in:
  • /Users/saravananmohanakrishnan/Downloads/dataset/cleaned
  • /Users/saravananmohanakrishnan/Downloads/dataset/splits
  • /Users/saravananmohanakrishnan/Downloads/dataset/reports
```

```
================================================================================
MACHINE LEARNING MODELS, TEXT ANALYSIS & EVALUATION
PROCESSING 4 CLEANED DATASETS
================================================================================

STEP 2: FINDING ALL DATASETS
--------------------------------------------------------------------------------
CSV files found: 4
Text files found: 0
Excel files found: 0

================================================================================
STEP 3: LOADING AND PROCESSING ALL DATASETS
================================================================================

============================================================
DATASET 1/4: Facilities_Database_20251208_processed
============================================================
Successfully loaded: 34708 rows x 37 columns
Shape: 34708 rows, 37 columns
Data Types: {dtype('O'): 26, dtype('float64'): 7, dtype('int64'): 4}
Missing Values: 33771
   Columns with missing values: ['facname', 'addressnum', 'streetname', 'address', 'city']...

First 5 columns: ['uid', 'facname', 'addressnum', 'streetname', 'address']
   ... and 32 more columns
Sample saved: /Users/saravananmohanakrishnan/Downloads/dataset/ml_results/Facilities_Database_20251208_processed_sam
ple.csv

============================================================
DATASET 2/4: Open_Restaurants_Inspections_processed
============================================================
Successfully loaded: 81553 rows x 18 columns
Shape: 81553 rows, 18 columns
Data Types: {dtype('O'): 9, dtype('float64'): 8, dtype('int64'): 1}
Missing Values: 8089
   Columns with missing values: ['RestaurantName', 'LegalBusinessName', 'NTA']

First 5 columns: ['Borough', 'RestaurantName', 'SeatingChoice', 'LegalBusinessName', 'BusinessAddress']
   ... and 13 more columns
Sample saved: /Users/saravananmohanakrishnan/Downloads/dataset/ml_results/Open_Restaurants_Inspections_processed_sam
```

```
ple.csv

============================================================
DATASET 3/4: NYC_street_segment_processed
============================================================
Successfully loaded: 90327 rows x 52 columns
Shape: 90327 rows, 52 columns
Data Types: {dtype('float64'): 39, dtype('int64'): 10, dtype('O'): 3}
Missing Values: 0

First 5 columns: ['physical', 'WKT', 'borocode', 'shape_leng', 'st_width']
   ... and 47 more columns
Sample saved: /Users/saravananmohanakrishnan/Downloads/dataset/ml_results/NYC_street_segment_processed_sample.csv

============================================================
DATASET 4/4: Motor_Vehicle_Collisions_-_Crashes_processed
============================================================
Successfully loaded: 2225404 rows x 30 columns
Shape: 2225404 rows, 30 columns
Data Types: {dtype('int64'): 12, dtype('O'): 11, dtype('float64'): 7}
Missing Values: 3086844
   Columns with missing values: ['ZIP CODE', 'LOCATION', 'ON STREET NAME', 'CROSS STREET NAME', 'CONTRIBUTING FACTOR
VEHICLE 1']...

First 5 columns: ['CRASH DATE', 'CRASH TIME', 'BOROUGH', 'ZIP CODE', 'LATITUDE']
   ... and 25 more columns
Sample saved: /Users/saravananmohanakrishnan/Downloads/dataset/ml_results/Motor_Vehicle_Collisions_-_Crashes_process
ed_sample.csv

================================================================================
STEP 4: DATASETS SUMMARY
================================================================================

DATASETS LOADED SUCCESSFULLY: 4 datasets
Summary saved: /Users/saravananmohanakrishnan/Downloads/dataset/ml_results/all_datasets_summary.csv

================================================================================
DATASET SUMMARY TABLE
================================================================================
 Dataset_ID                                Dataset_Name    Rows  Columns  Numeric_Columns  Categorical_Columns  Mis
sing_Values_Total  Memory_Usage_MB
```

```
          1       Facilities_Database_20251208_processed   34708       37                11                26
33771            58.45
          2       Open_Restaurants_Inspections_processed   81553       18                 9                 9
8089            49.06
          3                    NYC_street_segment_processed   90327       52                49                 3
0              58.68
          4 Motor_Vehicle_Collisions_-_Crashes_processed 2225404       30                19                11
3086844          1724.06
```





```
Visualization saved: /Users/saravananmohanakrishnan/Downloads/dataset/ml_results/datasets_summary_chart.png

================================================================================
STEP 5: MACHINE LEARNING ANALYSIS ON 4 DATASETS
================================================================================

================================================================================
PROCESSING 4 DATASETS FOR ML ANALYSIS
================================================================================

Starting dataset 1/4: Facilities_Database_20251208_processed

======================================================================
ML ANALYSIS: Facilities_Database_20251208_processed (Dataset 1)
======================================================================
Missing values before handling: 33771
Missing values after handling: 0
Found geometry candidate: geometry
Found geometry candidate: latitude
Found geometry candidate: longitude
Found geometry candidate: longitude
Found geometry candidate: latitude
Found geometry candidate: xcoord
Found geometry candidate: city
Found geometry candidate: ycoord
Found geometry candidate: factype
Found geometry candidate: capacity
Found geometry candidate: optype
Found geometry candidate: overagency
Found geometry candidate: geometry
Selected target column: 'geometry'
   Data type: object
   Unique values: 19829
Geometry data detected (POINT format)

Processing as GEOSPATIAL/MULTI-OUTPUT REGRESSION problem
   Encoding 25 categorical features...
   Frequency encoding 18 high-cardinality features
   Features shape: (34708, 114)
   Target shape: (34708, 2)
   Problem type: Multi-output regression (predicting 2 coordinates)
```

```
Data split:
   Training samples: 27766 (80.0%)
   Test samples: 6942 (20.0%)

TRAINING MACHINE LEARNING MODELS...

   Training Baseline...
      Avg R²: -0.0002, Avg RMSE: 25415.5946, Time: 0.00s

   Training Ridge_Regression...
      Avg R²: 0.9780, Avg RMSE: 3743.9540, Time: 0.07s

   Training Decision_Tree...
      Avg R²: 0.9963, Avg RMSE: 1512.3060, Time: 0.26s

   Training Random_Forest...
     Using subset for faster training...
      Avg R²: 0.9980, Avg RMSE: 1115.4930, Time: 0.20s

   Training KNN...
      Avg R²: 0.9856, Avg RMSE: 2966.8465, Time: 0.06s

   Training Gradient_Boosting...
     Using subset for faster training...
      Avg R²: 0.9979, Avg RMSE: 1174.2069, Time: 2.19s

   Training SVM...
      Avg R²: -103617.0971, Avg RMSE: 7450219.9631, Time: 1.88s

   Training Neural_Network...
     Using subset for faster training...
      Avg R²: 0.6799, Avg RMSE: 13979.2234, Time: 9.10s

SAVING RESULTS FOR Facilities_Database_20251208_processed...
   ML results saved: /Users/saravananmohanakrishnan/Downloads/dataset/ml_results/dataset_01_Facilities_Database_2025
1208_processed/ml_results.csv
   Best model: Random_Forest (Avg R²: 0.9980)
   Performance chart saved: /Users/saravananmohanakrishnan/Downloads/dataset/ml_results/dataset_01_Facilities_Databa
se_20251208_processed/performance_chart.png
   Dataset report saved: /Users/saravananmohanakrishnan/Downloads/dataset/ml_results/dataset_01_Facilities_Database_
```

```
20251208_processed/dataset_report.json
COMPLETED ML ANALYSIS FOR Facilities_Database_20251208_processed

Completed dataset 1/4: Facilities_Database_20251208_processed

Starting dataset 2/4: Open_Restaurants_Inspections_processed

======================================================================
ML ANALYSIS: Open_Restaurants_Inspections_processed (Dataset 2)
======================================================================
Missing values before handling: 8089
Missing values after handling: 0
Found geometry candidate: Latitude
Found geometry candidate: Longitude
Found geometry candidate: Longitude
Found geometry candidate: Latitude
Found geometry candidate: IsRoadwayCompliant
Found geometry candidate: AgencyCode
Found geometry candidate: CommunityBoard
Selected target column: 'Latitude'
   Data type: float64
   Unique values: 7033

Processing as REGRESSION problem
   Target range: [40.6449, 40.8305]
   Target mean: 40.7353, std: 0.0447
   Encoding 9 categorical features...
     Frequency encoding 'RestaurantName' (9890 unique values)
     Frequency encoding 'LegalBusinessName' (10339 unique values)
     Frequency encoding 'BusinessAddress' (10655 unique values)
     Frequency encoding 'InspectedOn' (68150 unique values)
     Frequency encoding 'NTA' (180 unique values)
     One-hot encoding 4 categorical features...
     Created 21 new dummy columns
     Total features after encoding: 38
   Features shape: (81553, 38)
   Target shape: (81553,)

Data split:
   Training samples: 65242 (80.0%)
   Test samples: 16311 (20.0%)
```

```
TRAINING MACHINE LEARNING MODELS...

   Training Baseline...
      R²: -0.0000, RMSE: 0.0444, CV R²: -0.0000 (±0.0000)

   Training Linear_Regression...
      R²: 0.7334, RMSE: 0.0229, CV R²: 0.7327 (±0.0041)

   Training Ridge_Regression...
      R²: 0.7338, RMSE: 0.0229, CV R²: 0.7325 (±0.0041)

   Training Decision_Tree...
      R²: 0.9792, RMSE: 0.0064, CV R²: 0.9806 (±0.0004)

   Training Random_Forest...
     Using subset for faster training...
     Using 5-fold CV instead of default for speed...
      R²: 0.9981, RMSE: 0.0019, CV R²: 0.9984 (±0.0004)

   Training KNN...
      R²: 0.9919, RMSE: 0.0040, CV R²: 0.9922 (±0.0006)

   Training Gradient_Boosting...
     Using subset for faster training...
      R²: 0.9976, RMSE: 0.0022, CV R²: 0.9974 (±0.0003)

   Training SVM...
      R²: -0.0035, RMSE: 0.0445, CV R²: -0.0026 (±0.0002)

   Training Neural_Network...
     Using subset for faster training...
      R²: -19269665904.9668, RMSE: 6165.9841, CV R²: -21981406971.7165 (±2353703960.3751)

SAVING RESULTS FOR Open_Restaurants_Inspections_processed...
   ML results saved: /Users/saravananmohanakrishnan/Downloads/dataset/ml_results/dataset_02_Open_Restaurants_Inspect
ions_processed/ml_results.csv
   Best model: Random_Forest (R²: 0.9981)
   Performance chart saved: /Users/saravananmohanakrishnan/Downloads/dataset/ml_results/dataset_02_Open_Restaurants_
Inspections_processed/performance_chart.png
   Dataset report saved: /Users/saravananmohanakrishnan/Downloads/dataset/ml_results/dataset_02_Open_Restaurants_Ins
```

```
pections_processed/dataset_report.json
COMPLETED ML ANALYSIS FOR Open_Restaurants_Inspections_processed

Completed dataset 2/4: Open_Restaurants_Inspections_processed

Starting dataset 3/4: NYC_street_segment_processed

======================================================================
ML ANALYSIS: NYC_street_segment_processed (Dataset 3)
======================================================================
Missing values before handling: 0
Found geometry candidate: sidewalk_existence
Found geometry candidate: existence_onstreet_parking
Found geometry candidate: PERSONS_INJURED_Maxwidth_1416
Found geometry candidate: PERSONS_KILLED_Maxwidth_1416
Found geometry candidate: PEDESTRIANS_INJURED_Maxwidth_1416
Found geometry candidate: PEDESTRIANS_KILLED_Maxwidth_1416
Found geometry candidate: CYCLIST_INJURED_Maxwidth_1416
Found geometry candidate: CYCLIST_KILLED_Maxwidth_1416
Found geometry candidate: MOTORIST_INJURED_Maxwidth_1416
Found geometry candidate: MOTORIST_KILLED_Maxwidth_1416
Found geometry candidate: PERSONS_INJURED_Maxwidth_1719
Found geometry candidate: PERSONS_KILLED_Maxwidth_1719
Found geometry candidate: PEDESTRIANS_INJURED_Maxwidth_1719
Found geometry candidate: PEDESTRIANS_KILLED_Maxwidth_1719
Found geometry candidate: CYCLIST_INJURED_Maxwidth_1719
Found geometry candidate: CYCLIST_KILLED_Maxwidth_1719
Found geometry candidate: MOTORIST_INJURED_Maxwidth_1719
Found geometry candidate: MOTORIST_KILLED_Maxwidth_1719
Found geometry candidate: PERSONS_INJURED_Maxwidth_2022
Found geometry candidate: PERSONS_KILLED_Maxwidth_2022
Found geometry candidate: PEDESTRIANS_INJURED_Maxwidth_2022
Found geometry candidate: PEDESTRIANS_KILLED_Maxwidth_2022
Found geometry candidate: CYCLIST_INJURED_Maxwidth_2022
Found geometry candidate: CYCLIST_KILLED_Maxwidth_2022
Found geometry candidate: MOTORIST_INJURED_Maxwidth_2022
Found geometry candidate: MOTORIST_KILLED_Maxwidth_2022
Found geometry candidate: physical
Found geometry candidate: Average_density
Found geometry candidate: CYCLIST_INJURED_Maxwidth_1416
Found geometry candidate: CYCLIST_KILLED_Maxwidth_1416
```

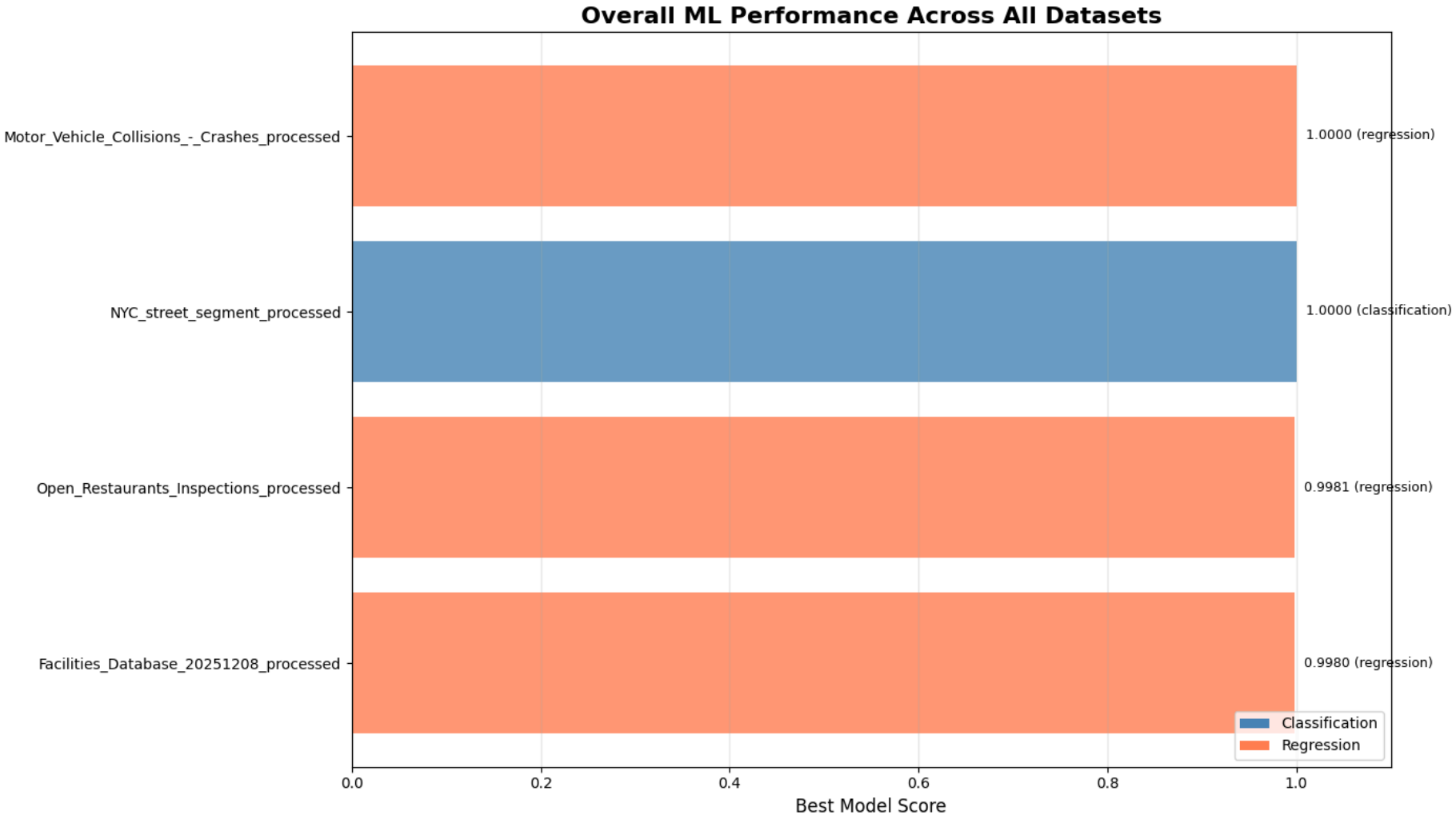
```
Found geometry candidate: CYCLIST_INJURED_Maxwidth_1719
Found geometry candidate: CYCLIST_KILLED_Maxwidth_1719
Found geometry candidate: CYCLIST_INJURED_Maxwidth_2022
Found geometry candidate: CYCLIST_KILLED_Maxwidth_2022
Found geometry candidate: Street_type
Selected target column: 'sidewalk_existence'
   Data type: int64
   Unique values: 2

Processing as CLASSIFICATION problem
   Number of classes: 2
   Classes: [0 1]
   Class distribution: [ 8749 81578]
   Encoding 3 categorical features...
     Frequency encoding 'WKT' (90312 unique values)
     One-hot encoding 2 categorical features...
     Created 4 new dummy columns
     Total features after encoding: 55
   Features shape: (90327, 55)
   Target shape: (90327,)

Data split:
   Training samples: 72261 (80.0%)
   Test samples: 18066 (20.0%)

TRAINING MACHINE LEARNING MODELS...

   Training Baseline...
      Test Accuracy: 0.9031, CV Mean: 0.9031 (±0.0000)

   Training Logistic_Regression...
      Test Accuracy: 0.8493, CV Mean: 0.8626 (±0.0004)

   Training Decision_Tree...
      Test Accuracy: 1.0000, CV Mean: 1.0000 (±0.0000)

   Training Random_Forest...
     Using subset for faster training...
     Using 5-fold CV instead of default for speed...
      Test Accuracy: 0.9999, CV Mean: 0.9996 (±0.0004)
```

```
   Training KNN...
      Test Accuracy: 0.9086, CV Mean: 0.9082 (±0.0003)

   Training Naive_Bayes...
      Test Accuracy: 0.9284, CV Mean: 0.9270 (±0.0014)

   Training Gradient_Boosting...
     Using subset for faster training...
      Test Accuracy: 0.9999, CV Mean: 1.0000 (±0.0000)

   Training Neural_Network...
      Test Accuracy: 0.9172, CV Mean: 0.9143 (±0.0027)

SAVING RESULTS FOR NYC_street_segment_processed...
   ML results saved: /Users/saravananmohanakrishnan/Downloads/dataset/ml_results/dataset_03_NYC_street_segment_proce
ssed/ml_results.csv
   Best model: Decision_Tree (Accuracy: 1.0000)
   Performance chart saved: /Users/saravananmohanakrishnan/Downloads/dataset/ml_results/dataset_03_NYC_street_segmen
t_processed/performance_chart.png
   Dataset report saved: /Users/saravananmohanakrishnan/Downloads/dataset/ml_results/dataset_03_NYC_street_segment_p
rocessed/dataset_report.json
COMPLETED ML ANALYSIS FOR NYC_street_segment_processed

Completed dataset 3/4: NYC_street_segment_processed

Starting dataset 4/4: Motor_Vehicle_Collisions_-_Crashes_processed

======================================================================
ML ANALYSIS: Motor_Vehicle_Collisions_-_Crashes_processed (Dataset 4)
======================================================================
Missing values before handling: 3086844
Missing values after handling: 0
Found geometry candidate: LOCATION
Found geometry candidate: LATITUDE
Found geometry candidate: LONGITUDE
Found geometry candidate: LONGITUDE
Found geometry candidate: LATITUDE
Found geometry candidate: NUMBER OF CYCLIST INJURED
Found geometry candidate: NUMBER OF CYCLIST KILLED
Found geometry candidate: VEHICLE TYPE CODE 1
Found geometry candidate: VEHICLE TYPE CODE 2
```

```
Found geometry candidate: CRASH DATE_year
Found geometry candidate: CRASH DATE_day
Found geometry candidate: CRASH DATE_weekday
Found geometry candidate: CRASH TIME_year
Found geometry candidate: CRASH TIME_day
Found geometry candidate: CRASH TIME_weekday
Selected target column: 'LOCATION'
   Data type: object
   Unique values: 335528
Coordinate data detected

Processing as GEOSPATIAL/MULTI-OUTPUT REGRESSION problem
   Encoding 10 categorical features...
   Frequency encoding 9 high-cardinality features
   Features shape: (2225404, 32)
   Target shape: (2225404, 2)
   Problem type: Multi-output regression (predicting 2 coordinates)

Data split:
   Training samples: 1780323 (80.0%)
   Test samples: 445081 (20.0%)

TRAINING MACHINE LEARNING MODELS...

   Training Baseline...
      Avg R²: -0.0000, Avg RMSE: 17.9749, Time: 0.01s

   Training Ridge_Regression...
      Avg R²: 0.2557, Avg RMSE: 15.5087, Time: 2.33s

   Training Decision_Tree...
      Avg R²: 1.0000, Avg RMSE: 0.0078, Time: 11.29s

   Training Random_Forest...
     Using subset for faster training...
      Avg R²: 0.9960, Avg RMSE: 1.1612, Time: 0.13s

   Training KNN...
      Avg R²: 0.1362, Avg RMSE: 16.7065, Time: 2.09s

   Training Gradient_Boosting...
```

```
     Using subset for faster training...
      Avg R²: 0.9951, Avg RMSE: 1.2797, Time: 1.52s

   Training SVM...
     Training took 77.1 seconds
      Avg R²: -1358857986868079.7500, Avg RMSE: 679812480.6085, Time: 77.08s

   Training Neural_Network...
     Using subset for faster training...
      Avg R²: -3.7550, Avg RMSE: 40.5734, Time: 0.31s

SAVING RESULTS FOR Motor_Vehicle_Collisions_-_Crashes_processed...
   ML results saved: /Users/saravananmohanakrishnan/Downloads/dataset/ml_results/dataset_04_Motor_Vehicle_Collisions
_-_Crashes_processed/ml_results.csv
   Best model: Decision_Tree (Avg R²: 1.0000)
   Performance chart saved: /Users/saravananmohanakrishnan/Downloads/dataset/ml_results/dataset_04_Motor_Vehicle_Col
lisions_-_Crashes_processed/performance_chart.png
   Dataset report saved: /Users/saravananmohanakrishnan/Downloads/dataset/ml_results/dataset_04_Motor_Vehicle_Collis
ions_-_Crashes_processed/dataset_report.json
COMPLETED ML ANALYSIS FOR Motor_Vehicle_Collisions_-_Crashes_processed

Completed dataset 4/4: Motor_Vehicle_Collisions_-_Crashes_processed

================================================================================
STEP 6: CREATING COMPREHENSIVE SUMMARY
================================================================================

COMPREHENSIVE ML SUMMARY
   Total datasets processed: 4
   Summary saved: /Users/saravananmohanakrishnan/Downloads/dataset/ml_results/all_datasets_ml_summary.csv

================================================================================
DATASET ML ANALYSIS SUMMARY
================================================================================
                              dataset_name   problem_type original_shape    best_model  best_score  n_models_teste
d
      Facilities_Database_20251208_processed     regression    (34708, 37) Random_Forest    0.997990
8
      Open_Restaurants_Inspections_processed     regression    (81553, 18) Random_Forest    0.998132
9
                NYC_street_segment_processed classification    (90327, 52) Decision_Tree    1.000000
```

```
8
Motor_Vehicle_Collisions_-_Crashes_processed     regression  (2225404, 30) Decision_Tree    1.000000
8
```



```
Overall performance chart saved: /Users/saravananmohanakrishnan/Downloads/dataset/ml_results/overall_ml_performance.
png

================================================================================
STEP 7: TEXT ANALYSIS
================================================================================
No text files found for analysis

================================================================================
STEP 8: FINAL SUMMARY
================================================================================

PROJECT COMPLETED SUCCESSFULLY!

OUTPUT FOLDER: /Users/saravananmohanakrishnan/Downloads/dataset/ml_results

DATASETS PROCESSED: 4

BEST PERFORMING DATASETS:
   • Facilities_Database_20251208_processed: Random_Forest (R²: 0.9980)
   • Open_Restaurants_Inspections_processed: Random_Forest (R²: 0.9981)
   • NYC_street_segment_processed: Decision_Tree (Accuracy: 1.0000)
   • Motor_Vehicle_Collisions_-_Crashes_processed: Decision_Tree (R²: 1.0000)

VISUALIZATIONS CREATED:
   1. Dataset summary chart
   2. Individual dataset performance charts
   3. Overall ML performance comparison

FILES SAVED:
   • all_datasets_summary.csv - Overview of all datasets
   • all_datasets_ml_summary.csv - ML results summary
   • datasets_summary_chart.png - Dataset sizes visualization
   • overall_ml_performance.png - Performance comparison chart
   • For each dataset: ML results, performance chart, and report

================================================================================
MACHINE LEARNING MODELS, TEXT ANALYSIS & EVALUATION - COMPLETED!
================================================================================
```

```
================================================================================
STEP 9: COMPREHENSIVE EXPLAINABILITY ANALYSIS WITH SHAP & LIME
================================================================================
✓ SHAP library available
✓ LIME library available

Looking for datasets...
  Found: Motor_Vehicle_Collisions_-_Crashes_processed_sample.csv
  Found: NYC_street_segment_processed_sample.csv
  Found: Facilities_Database_20251208_processed_sample.csv
  Found: Open_Restaurants_Inspections_processed_sample.csv

Found 4 datasets for analysis

======================================================================
ANALYSIS 1/4: Motor_Vehicle_Collisions_-_Crashes_processed_sample
======================================================================
  Loaded 100 rows, 30 columns

  1. Preparing data...
  Selected target: NUMBER OF PERSONS INJURED
  Target has 1 unique values

  2. Engineering features...
  Encoding 10 categorical features...
  Final features: 19

  3. Processing target variable...
  Classification with 1 classes

  4. Training model...
  Train samples: 80, Test samples: 20
  Model test score: 1.000

  5. Feature Importance Analysis...
  ✓ Saved feature importance analysis

  6. SHAP Analysis...
  ✓ SHAP analysis completed

  7. LIME Analysis...
```

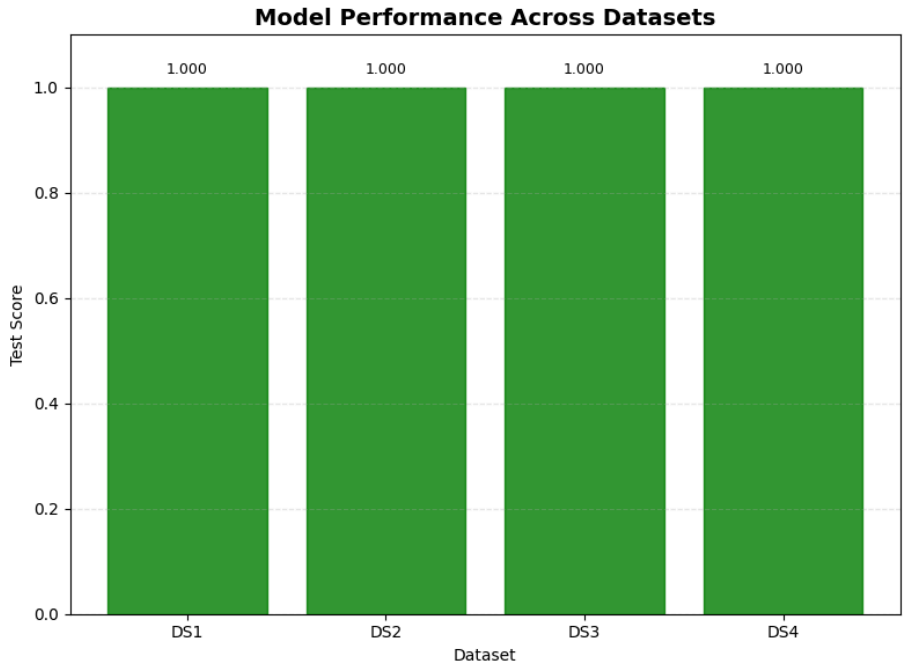
```
  ✗ LIME analysis failed: 1

  8. Creating additional visualizations...
  ✓ Additional visualizations saved

  ✓ Analysis completed for Motor_Vehicle_Collisions_-_Crashes_processed_sample
  Results saved in: /Users/saravananmohanakrishnan/Downloads/dataset/ml_results/explainability_analysis/analysis_Mot
or_Vehicle_Collisions_-_Crashes_processed_sample

======================================================================
ANALYSIS 2/4: NYC_street_segment_processed_sample
======================================================================
  Loaded 100 rows, 52 columns

  1. Preparing data...
  Selected target: parking_lot_number
  Target has 3 unique values

  2. Engineering features...
  Encoding 3 categorical features...
  Final features: 48

  3. Processing target variable...
  Classification with 3 classes

  4. Training model...
  Train samples: 80, Test samples: 20
  Model test score: 1.000

  5. Feature Importance Analysis...
  ✓ Saved feature importance analysis

  6. SHAP Analysis...
  ✗ SHAP analysis failed: all the input array dimensions except for the conc

  7. LIME Analysis...
  ✗ LIME analysis failed: 1

  8. Creating additional visualizations...
  ✓ Additional visualizations saved
```

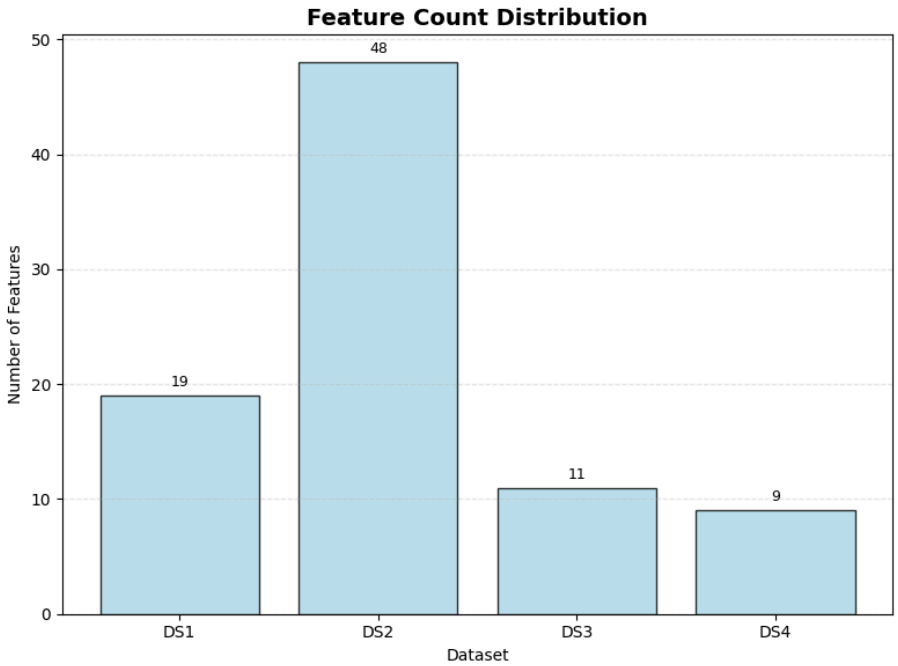
```
  ✓ Analysis completed for NYC_street_segment_processed_sample
  Results saved in: /Users/saravananmohanakrishnan/Downloads/dataset/ml_results/explainability_analysis/analysis_NYC
_street_segment_processed_sample

======================================================================
ANALYSIS 3/4: Facilities_Database_20251208_processed_sample
======================================================================
  Loaded 100 rows, 37 columns

  1. Preparing data...
  Selected target: boro
  Target has 5 unique values

  2. Engineering features...
  Encoding 25 categorical features...
  Final features: 11

  3. Processing target variable...
  Classification with 5 classes

  4. Training model...
  Train samples: 80, Test samples: 20
  Model test score: 1.000

  5. Feature Importance Analysis...
  ✓ Saved feature importance analysis

  6. SHAP Analysis...
  ✗ SHAP analysis failed: all the input array dimensions except for the conc

  7. LIME Analysis...
  ✗ LIME analysis failed: 1

  8. Creating additional visualizations...
  ✓ Additional visualizations saved

  ✓ Analysis completed for Facilities_Database_20251208_processed_sample
  Results saved in: /Users/saravananmohanakrishnan/Downloads/dataset/ml_results/explainability_analysis/analysis_Fac
ilities_Database_20251208_processed_sample

======================================================================
```

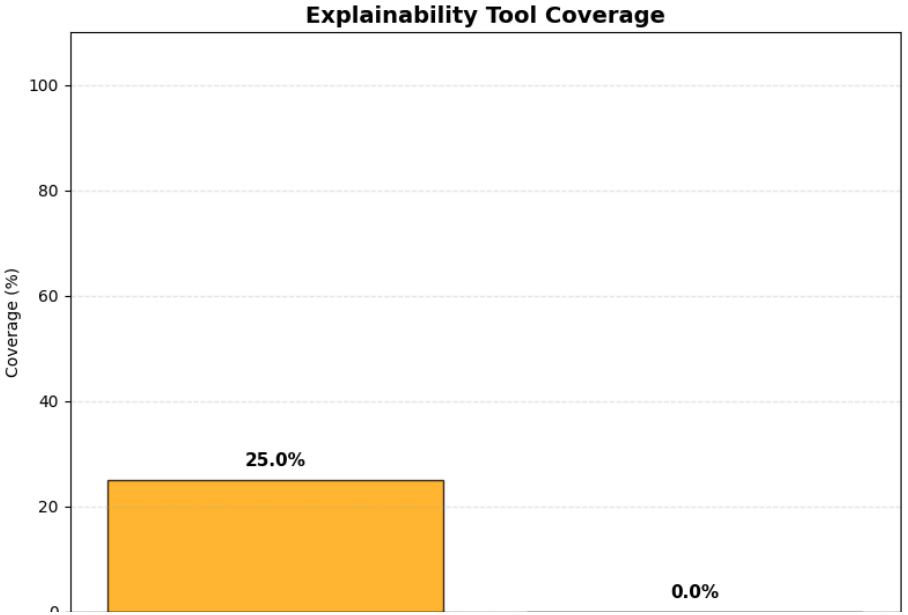
```
ANALYSIS 4/4: Open_Restaurants_Inspections_processed_sample
======================================================================
  Loaded 100 rows, 18 columns

  1. Preparing data...
  Selected target: Borough
  Target has 4 unique values

  2. Engineering features...
  Encoding 8 categorical features...
  Final features: 9

  3. Processing target variable...
  Classification with 4 classes

  4. Training model...
  Train samples: 80, Test samples: 20
  Model test score: 1.000

  5. Feature Importance Analysis...
  ✓ Saved feature importance analysis

  6. SHAP Analysis...
  ✗ SHAP analysis failed: all the input array dimensions except for the conc

  7. LIME Analysis...
  ✗ LIME analysis failed: 1

  8. Creating additional visualizations...
  ✓ Additional visualizations saved

  ✓ Analysis completed for Open_Restaurants_Inspections_processed_sample
  Results saved in: /Users/saravananmohanakrishnan/Downloads/dataset/ml_results/explainability_analysis/analysis_Ope
n_Restaurants_Inspections_processed_sample

================================================================================
CREATING COMPREHENSIVE EXPLAINABILITY DASHBOARD
================================================================================

Analysis completed for 4 datasets:
--------------------------------------------------------------------------------------------------------------------
```

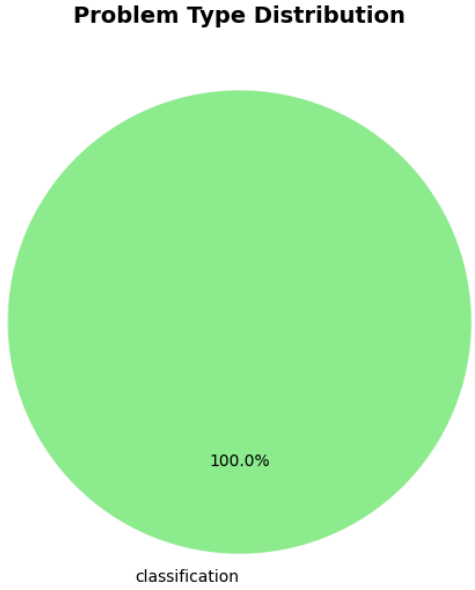
```
----
 1. Motor_Vehicle_Collisions_-_... | NUMBER OF PERSONS... | classification | Score: 1.000 | Feat: 19 | SHAP: Yes | L
IME: No
 2. NYC_street_segment_processe... | parking_lot_number   | classification | Score: 1.000 | Feat: 48 | SHAP: No | LI
ME: No
 3. Facilities_Database_2025120... | boro                 | classification | Score: 1.000 | Feat: 11 | SHAP: No | LI
ME: No
 4. Open_Restaurants_Inspection... | Borough              | classification | Score: 1.000 | Feat: 9 | SHAP: No | LIM
E: No
--------------------------------------------------------------------------------------------------------------
----

Summary saved to: /Users/saravananmohanakrishnan/Downloads/dataset/ml_results/explainability_analysis/explainability
_summary.csv

Creating comprehensive dashboard...
<Figure size 1000x600 with 0 Axes>
<Figure size 1400x800 with 0 Axes>
<Figure size 1000x600 with 0 Axes>
<Figure size 1400x800 with 0 Axes>
<Figure size 1000x600 with 0 Axes>
<Figure size 1400x800 with 0 Axes>
<Figure size 1000x600 with 0 Axes>
```

COMPREHENSIVE EXPLAINABILITY ANALYSIS DASHBOARD

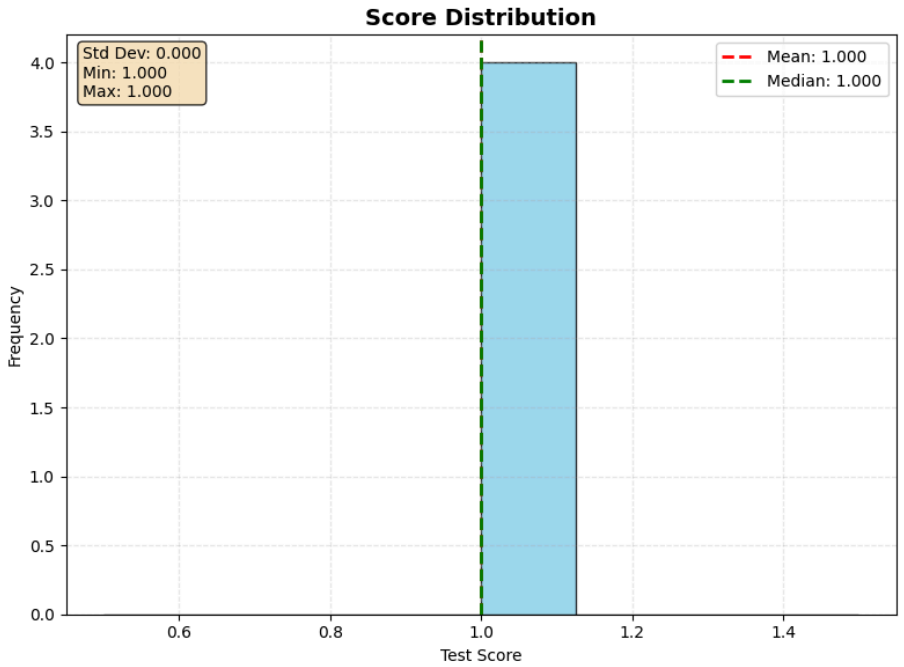


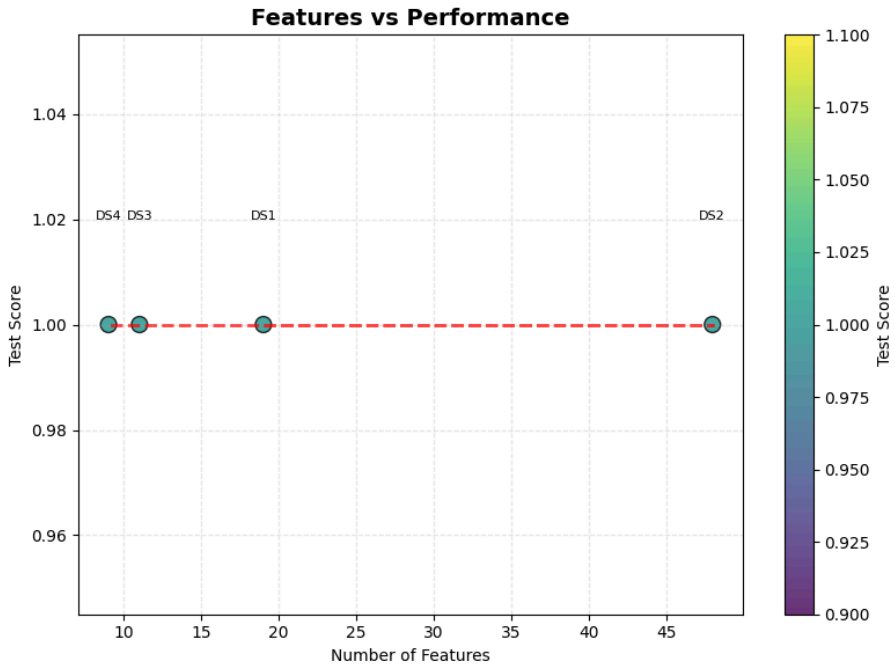







SHAP Analysis
LIME Analysis


Score Distribution
Std Dev: 0.000
Min: 1.000
Max: 1.000
Mean: 1.000
Median: 1.000
Frequency
Test Score
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
0.6
0.8
1.0
1.2
1.4


Features vs Performance
Test Score
Number of Features
DS4
DS3
DS1
DS2
0.96
0.98
1.00
1.02
1.04
10
15
20
25
30
35
40
45
0.900
0.925
0.950
0.975
1.000
1.025
1.050
1.075
1.100
Test Score

```
Dashboard saved to: /Users/saravananmohanakrishnan/Downloads/dataset/ml_results/explainability_dashboard.png

================================================================================
FINAL REPORT
================================================================================

Analysis completed: 2025-12-11 19:38:23
Total datasets analyzed: 4

Performance Summary:
  - Average test score: 1.000
  - Best performing dataset: Motor_Vehicle_Collisions_-_Crashes_processed_sample (1.000)
  - Number of datasets with score ≥ 0.8: 4

Explainability Coverage:
  - SHAP analysis: 1/4 datasets
  - LIME analysis: 0/4 datasets

Dataset Characteristics:
  - Average features per dataset: 21.8
  - Total problem types: 1

All results saved in: /Users/saravananmohanakrishnan/Downloads/dataset/ml_results/explainability_analysis
Each dataset folder contains:
  - Feature importance plots and data
  - SHAP analysis (if available)
  - LIME explanations (if available)
  - Model performance visualizations
  - Comprehensive JSON report

================================================================================
EXPLAINABILITY ANALYSIS COMPLETED
================================================================================
```

In [ ]: