Designing Advance Data Architecture for BI
Individual Project

Name: Yashasvi Nagar
Email: nagar.ya@northeastern.edu
NUID: 002056347

Dataset Cleaning Plan and Problem Documentation

Introduction: This document outlines the identified data quality issues in the dataset containing 5,000 records across 14 columns and provides a comprehensive cleaning strategy to prepare the data for analysis.

1. Identified Data Quality Problems

- High Percentage of Missing Values
  - Category2 column: 58.58% null values (2,929 missing records)
  - Category3 column: 68.62% null values (3,431 missing records)
  - Impact: These high missing value rates severely limit analysis capabilities and may introduce bias if not handled properly
- Data Completeness Concerns
  - With only 10% of records being fully populated (based on the "percent of records: 10" notation), the dataset suffers from significant incompleteness
  - This suggests potential systematic data collection issues
- Potential Issues

Since multiple visualizations were generated for field summaries, the following issues are likely to present:

- Outliers and anomalies in numerical fields
- Inconsistent data formats across categorical variables
- Distribution irregularities that may affect statistical analysis
- Potential duplicate records that could skew results
- Data type inconsistencies between similar fields

2. Data Cleaning Strategy

Generate a comprehensive data quality report including:

- Complete missing value analysis for all 14 columns
- Data type verification for each column
- Basic statistical summaries (mean, median, mode, standard deviation)
- Unique value counts for categorical variables

SQL Queries:

(1) SERVICE REQUESTS OVER TIME

Query 1a – Yearly Trend (2018–2021)

```
SELECT
   YEAR(CAST(creationdate AS date)) AS Year,
   COUNT(*) AS Requests
FROM dbo.servicerequest
WHERE YEAR(CAST(creationdate AS date)) BETWEEN 2018 AND 2021
GROUP BY YEAR(CAST(creationdate AS date))
ORDER BY Year;
```

- Shows total service requests per year.
- Helps identify whether the overall request volume is increasing, decreasing, or stable.
- Example: A rising trend may indicate higher citizen engagement or operational strain.

Query 1b – Monthly Trend (2018–2021)

```
SELECT
   FORMAT(CAST(creationdate AS date), 'yyyy-MM') AS YearMonth,
   COUNT(*) AS Requests
FROM dbo.servicerequest
WHERE YEAR(CAST(creationdate AS date)) BETWEEN 2018 AND 2021
GROUP BY FORMAT(CAST(creationdate AS date), 'yyyy-MM')
ORDER BY YearMonth;
```

- Detects seasonal patterns or spikes in service requests.
- Example: Certain months consistently show higher requests, indicating seasonal demand or recurring maintenance cycles.

(2) VOLUME OF SERVICE REQUESTS BY SOURCE

```
SELECT
   source,
   COUNT(*) AS Requests
FROM dbo.servicerequest
GROUP BY source
ORDER BY Requests DESC;
```

- Identifies which channels (phone, email, web, etc.) citizens use most.
- Example: High online requests suggest web platform popularity; low usage of some channels may indicate underutilization.

(3) VOLUME OF SERVICE REQUESTS BY DEPARTMENT

```
SELECT
   department,
   COUNT(*) AS Requests
FROM dbo.servicerequest
GROUP BY department
ORDER BY Requests DESC;
```

- Shows which departments handle the most requests.
- Example: High-volume departments may require additional staffing or process optimization.

(4) TOP 10 FASTEST RESPONSE TIMES (BY CATEGORY1 & TYPE)

```
SELECT TOP (10)
   caseid,
   category1,
   type,
   department,
   source,
   creationdate,
   closedate,
   daystoclose
FROM dbo.servicerequest
WHERE closedate IS NOT NULL
ORDER BY daystoclose ASC;
```

- Highlights cases with the quickest closure times.
- Example: Efficient departments/workgroups can serve as benchmarks for improving overall service performance.

(5) GEOGRAPHICAL VISUALIZATION (TOP 10 AREAS)

Query 5a – By Street Address

```
SELECT TOP (10)
   streetaddress,
   COUNT(*) AS Requests
FROM dbo.servicerequest
GROUP BY streetaddress
ORDER BY Requests DESC;
```

Query 5b – By Zip Code

```
SELECT TOP (10)
   zipcode,
   COUNT(*) AS Requests
```

```
FROM dbo.servicerequest
GROUP BY zipcode
ORDER BY Requests DESC;
```

Query 5c – By Latitude/Longitude

```
SELECT TOP (10)
   latitude,
   longitude,
   COUNT(*) AS Requests
FROM dbo.servicerequest
GROUP BY latitude, longitude
ORDER BY Requests DESC;
```

- Identifies geographic hotspots of service requests.
- Example: High-demand areas may need proactive planning, such as targeted maintenance or staff allocation.

(6) DEPARTMENTAL WORKLOAD VS WORKGROUP

```
SELECT
   department,
   workgroup,
   COUNT(*) AS Requests
FROM dbo.servicerequest
GROUP BY department, workgroup
ORDER BY department, Requests DESC;
```

- Shows workload distribution among workgroups within departments.
- Example: Uneven distribution may indicate overloaded teams that need task reallocation.

(7) RESPONSE TIME ANALYSIS PER DEPARTMENT

```
SELECT
   department,
   COUNT(*) AS ClosedCount,
   AVG(daystoclose) AS AvgDays,
   MIN(daystoclose) AS MinDays,
   MAX(daystoclose) AS MaxDays
FROM dbo.servicerequest
WHERE daystoclose IS NOT NULL
GROUP BY department
ORDER BY AvgDays DESC;
```

- Assesses department efficiency based on response times.
- Example: Departments with high average closure days may face bottlenecks; those with low averages indicate high efficiency.

(8) SERVICE REQUEST STATUS COMPOSITION (2018–2021)

```
SELECT
   YEAR(CAST(creationdate AS date)) AS Year,
   status,
   COUNT(*) AS Requests
FROM dbo.servicerequest
WHERE YEAR(CAST(creationdate AS date)) BETWEEN 2018 AND 2021
GROUP BY YEAR(CAST(creationdate AS date)), status
ORDER BY Year, Requests DESC;
```

- Shows request status trends over the years (e.g., closed, pending, overdue).
- Example: Increasing closed requests indicate efficiency; growing open/overdue requests highlight backlog issues.

(9) TIME TO CLOSURE BY CATEGORY1 (TOP 10 LONGEST)

```
SELECT TOP (10)
   category1,
   COUNT(*) AS ClosedCount,
   AVG(daystoclose) AS AvgDaysToClose
FROM dbo.servicerequest
WHERE daystoclose IS NOT NULL
GROUP BY category1
ORDER BY AvgDaysToClose DESC;
```

- Highlights categories with the slowest resolution times.
- Example: Long closure times in certain categories indicate areas for process improvement or additional resources.

(10) WORKLOAD VS EFFICIENCY BY DEPARTMENT

```
SELECT
   department,
   COUNT(*) AS TotalRequests,
   AVG(daystoclose) AS AvgDaysToClose
FROM dbo.servicerequest
GROUP BY department
ORDER BY TotalRequests DESC;
```

- Compares workload (total requests) and efficiency (average closure time).
- Example: Departments with high requests and slow closures need more resources, while low-volume, fast-closing departments may be underutilized.

	Year	Requests
1	2018	124280
2	2019	166021
3	2020	125906
4	2021	19683

	YearMonth	Requests
1	2018-01	10870
2	2018-02	7728
3	2018-03	9625
4	2018-04	9657
5	2018-05	12295
6	2018-06	11644
7	2018-07	12329
8	2018-08	10960

	source	Requests
1	PHONE	1204236
2	WEB	211721
3	EMAIL	80585
4	SYS	19226
5	INSPE	14690
6	BOT	13396
7	TWIR	8311
8	VOICE	6021

	department	Requests
1	NHS	783094
2	Public Works	353787
3	Water Serv...	216852
4	Parks and ...	87954
5	Health	39543
6	KCPD	36369
7	City Manag...	13098
8	City Planni...	12575

	caseid	category1	type	department	source	creationdate	closedate	daystoclose
1	2020086740	Water	MFS Referral (Meter Field Services)	Water Services	PHONE	06/23/2020	06/30/2020	NULL
2	2020160258	Trash	Trash Collection	NHS	WEB	12/31/2020	01/03/2021	NULL
3	2020103086	Water	No Water / Pressure	Water Services	PHONE	07/28/2020	07/30/2020	NULL
4	2020119506	Property	Property Maintenance	South	PHONE	09/03/2020	09/04/2020	NULL
5	2019175412	Water	MFS Referral (Meter Field Services)	Water Services	PHONE	10/03/2019	10/08/2019	NULL
6	2019186289	Water	MFS Referral (Meter Field Services)	Water Services	PHONE	10/29/2019	11/05/2019	NULL
7	2020060588	Water	Leak	Water Services	PHONE	05/01/2020	05/06/2020	NULL
8	2020104964	Water	Leak	Water Services	PHONE	07/31/2020	08/04/2020	NULL

	streetaddress	Requests
1	414 E 12TH ST	5773
2	4800 E 63RD ST	2406
3	415 E 12th st	1816
4	400 W 31ST ST	509
5	5100 Wornall Rd	381
6	4900 Raytown ...	291
7	2400 Troost Ave	290
8	2700 MAIN ST	278

	zipcode	Requests
1	64130	133154
2	64127	90867
3	64114	83166
4	64134	77097
5	64131	76437
6	64132	76361
7	64128	73073
8	64110	64771

	latitude	longitude	Requests
1	-103.890579	35.797519	40610
2	-94.577919	39.100387	4940
3	-94.529933	39.014493	2468
4	-94.577983	39.099123	1246
5	-94.591904	39.074934	468
6	-94.472954	39.285963	378
7	-94.576776	39.100349	371
8	-94.59448	39.031378	368

	department	workgroup	Requests
2	City Clerks Office	City Clerks Office--	2
3	City Council	City Council--	3
4	City Managers ...	City Managers O...	8106
5	City Managers ...	City Managers O...	2131
6	City Managers ...	City Managers O...	1235
7	City Managers ...	City Managers O...	579
8	City Managers ...	City Managers O...	397
9	City Managers ...	City Managers O...	187

	department	ClosedCount	AvgDays	MinDays	MaxDays
3	City Plannin...	12023	105.12525991849	0	3595
4	NHS	762712	77.7277530706217	0	4525
5	City Council	3	77.6666666666667	69	83
6	General Ser...	513	57.5438596491228	0	621
7	IT	1	47	47	47
8	Northeast	33	45.7878787878788	1	296
9	Parks & Re...	11	45.5454545454545	4	117
10	South	304	44.5986842105263	0	386

	Year	status	Requests
5	2018	FAIL	1
6	2019	RESOL	163187
7	2019	OPEN	2229
8	2019	CANC	432
9	2019	DUP	171
10	2019	ASSIG	1
11	2019	FAIL	1
12	2020	RESOL	119480

	category1	ClosedCount	AvgDaysToClose
2	Weeds	12398	420.305371834167
3	Property & Nuisa...	20447	187.382354379616
4	Property Violations	133543	180.22246766959
5	Water Main Break	99	172.50505050505
6	Property	69038	171.236130826501
7	Mowing	33733	162.946076542258
8	Information Req...	4	103.5
9	Water Services	5942	99.108044429485

	department	TotalRequests	AvgDaysToClose
1	NHS	783094	77.7277530706217
2	Public Works	353787	10.9746796706317
3	Water Serv...	216852	32.4879225011505
4	Parks and ...	87954	19.9493849658314
5	Health	39543	11.1741284683403
6	KCPD	36369	3.62778924198652
7	City Manag...	13098	22.1688514357054
8	City Planni...	12575	105.12525991849