DATA_T1

BUILDFLAKE

THE ALL IN ONE BUILDING DATA SOLUTION

The Team: Samuel Park, Quinn Thompson, Cecilia Wu, Kevin Yin

MEET OUR TEAM

Samuel Park

Business Value Analyst

Kevin Yin

Data Analyst

Quinn Thompson

Cloud Engineer/BI Analyst

Cecilia Wu

Data Architect

WHAT IS THE PROBLEM AT HAND?

(EXECUTIVE SUMMARY)

SMART GOVERNMENT - GOVERNMENT AS A BUSINESS

1. Modernizing Data Management Systems

- 1) **Information accessibility and quality**: Enhance public access to accurate and timely information
- 2) **Data management efficiency**: Improve data management system (ETL)
- Improve overall government operation

2. Building a Data-Driven Budget Optimization Cycle

- A sustainable financial structure through transparent, data-driven budgeting.
- 1) **Recover unpaid fees** to secure additional funding.
- 2) **Optimize workforce productivity** through data-informed decision

BEFORE DIVING INTO EDA GRAPHS.....

SOME KEYWORDS NEED TO KNOW

BUILDING PERMIT ? VIOLATION?

BUILDING VIOLATION:

OCCURS WHEN A PROPERTY **DOES NOT COMPLY** WITH LOCAL BUILDING CODES. THE DEPARTMENT OF BUILDINGS (DOB) IS RESPONSIBLE FOR INVESTIGATING BUILDING VIOLATIONS. E.G.**ILLEGAL CONSTRUCTION**, **BLOCKED EMERGENCY EXITS**

BUILDING PERMIT:

AN OFFICIAL DOCUMENT ISSUED BY A LOCAL GOVERNMENT AUTHORITY THAT **GRANTS PERMISSION** TO **CONSTRUCT, ALTER, REPAIR, OR DEMOLISH** A BUILDING OR STRUCTURE.

311?

311 IS A NON-EMERGENCY SERVICE REQUEST SYSTEM HELPS TRACK AND ADDRESS COMMUNITY CONCERNS E.G. TRASH OR RECYCLING ISSUES, NOISE COMPLAINTS

WHAT DO WE KNOW ABOUT THE DATA PRIOR TO BUILDFLAKE?

(EXPLORATORY DATA ANALYSIS)

EXPLORATORY DATA ANALYSIS

SOURCE OF DATA

CHICAGO DATA PORTAL Chicago Data Portal

Developed by the City of Chicago, by the Department of Innovation and Technology (DoIT), Free and Open data platform, Diverse Datasets

TABLE WE SELECT

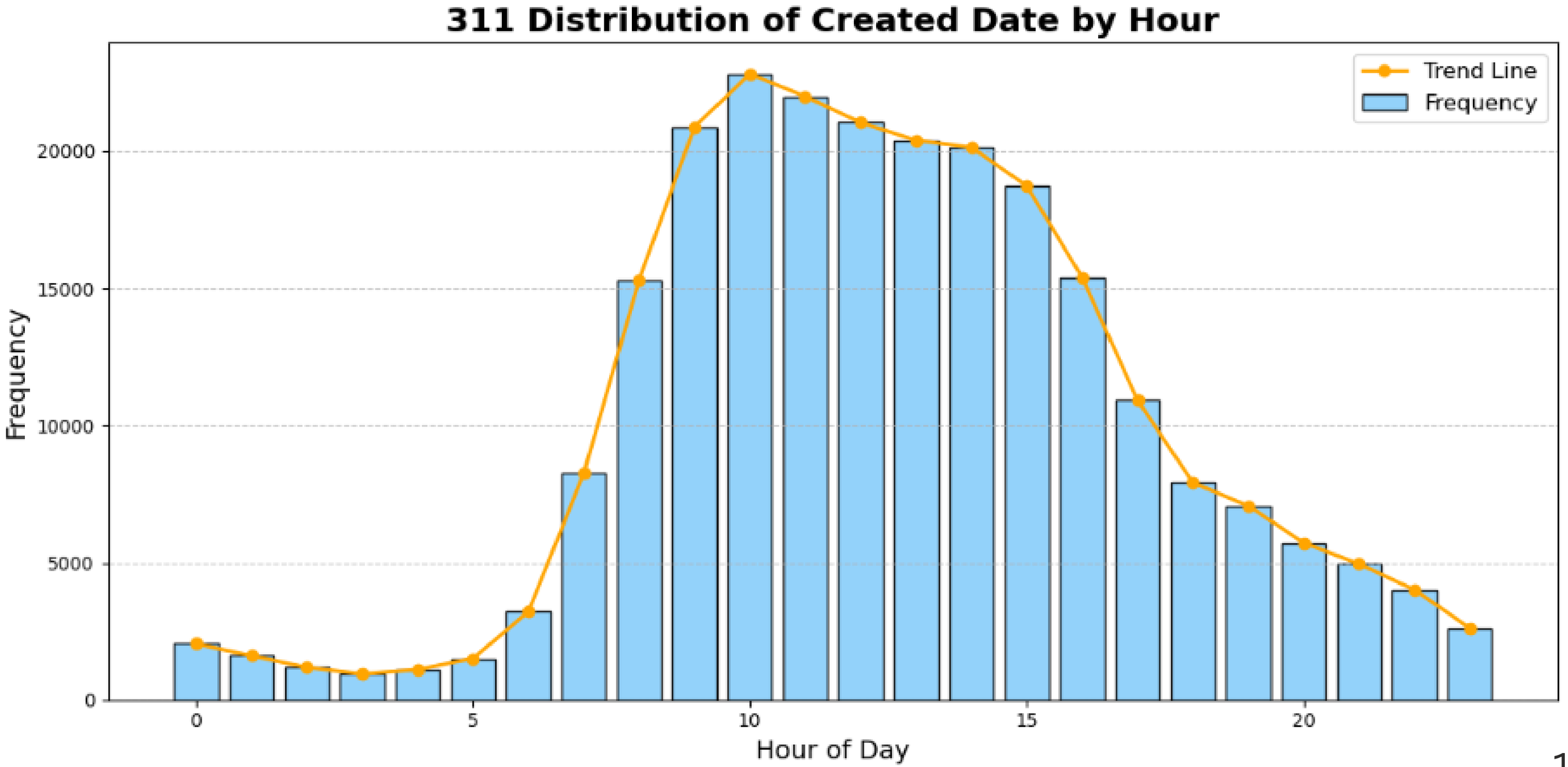
- Building Permits — Buildings
- Building Violations — Buildings
- buildings — Buildings
- 311 Service Requests — Service Requests
- Building Footprints (current) — Buildings

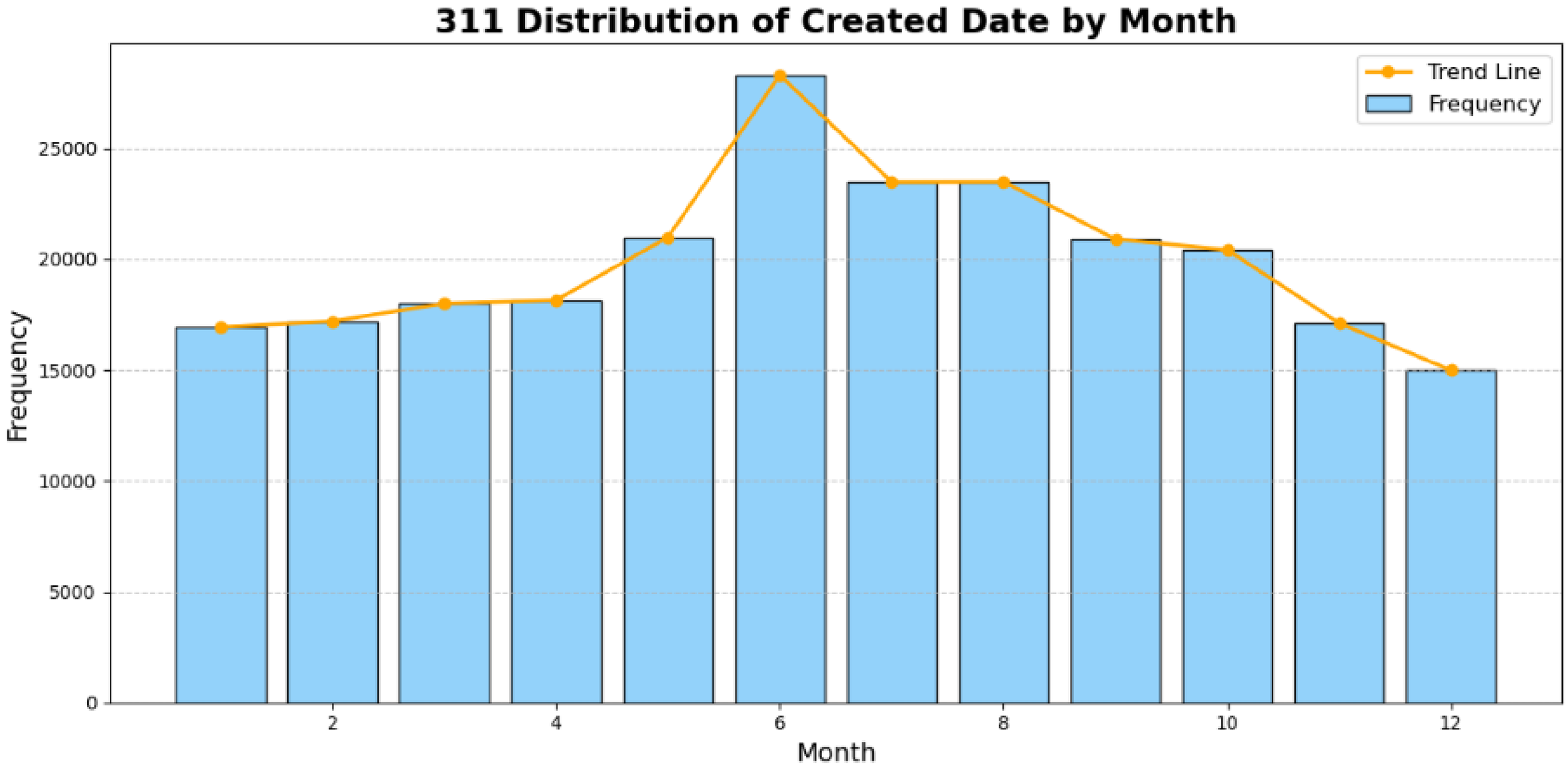
DATA SIZE COMPARISON

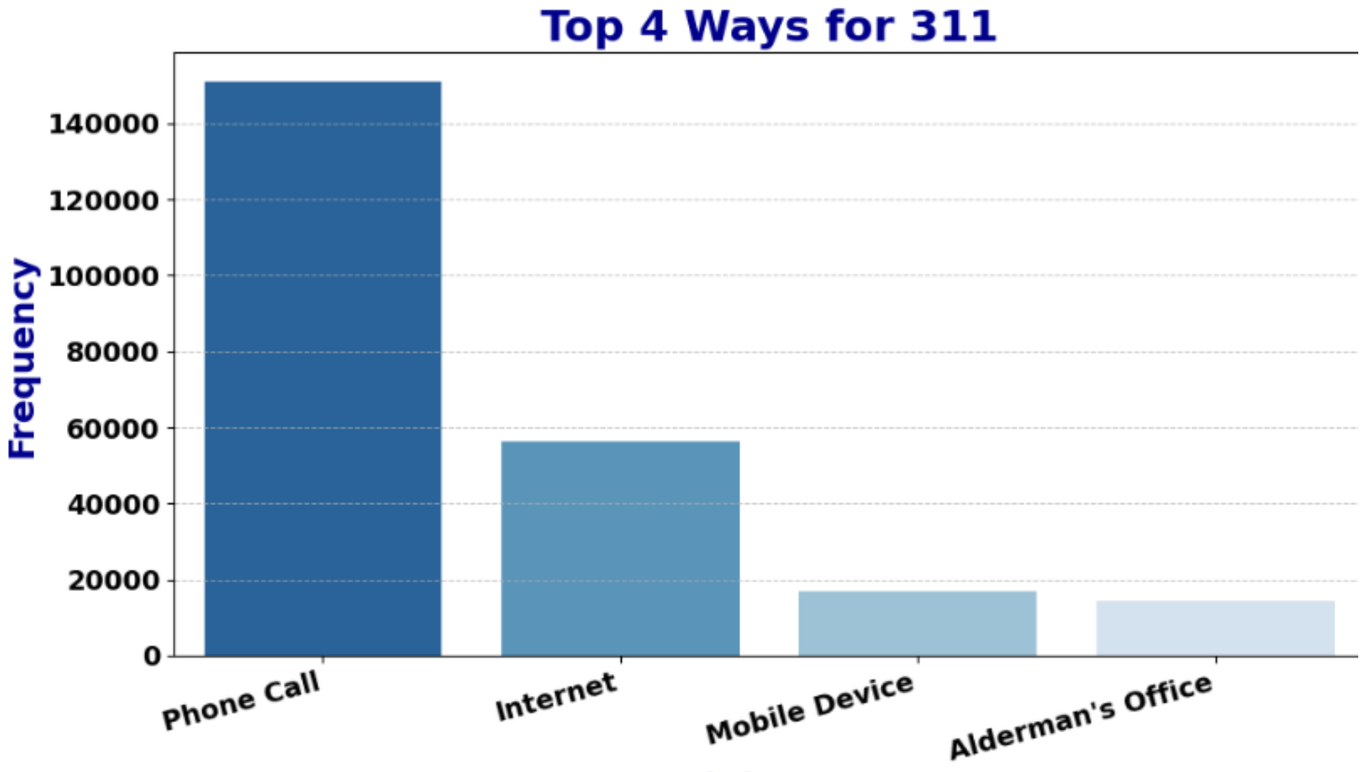
Data	Rows and Columns
Building Data	(821K, 43)
Ordinance Violation Data	(797K, 22)
Violation Data	(1.94M, 26)
Permit Data	(792K, 115)

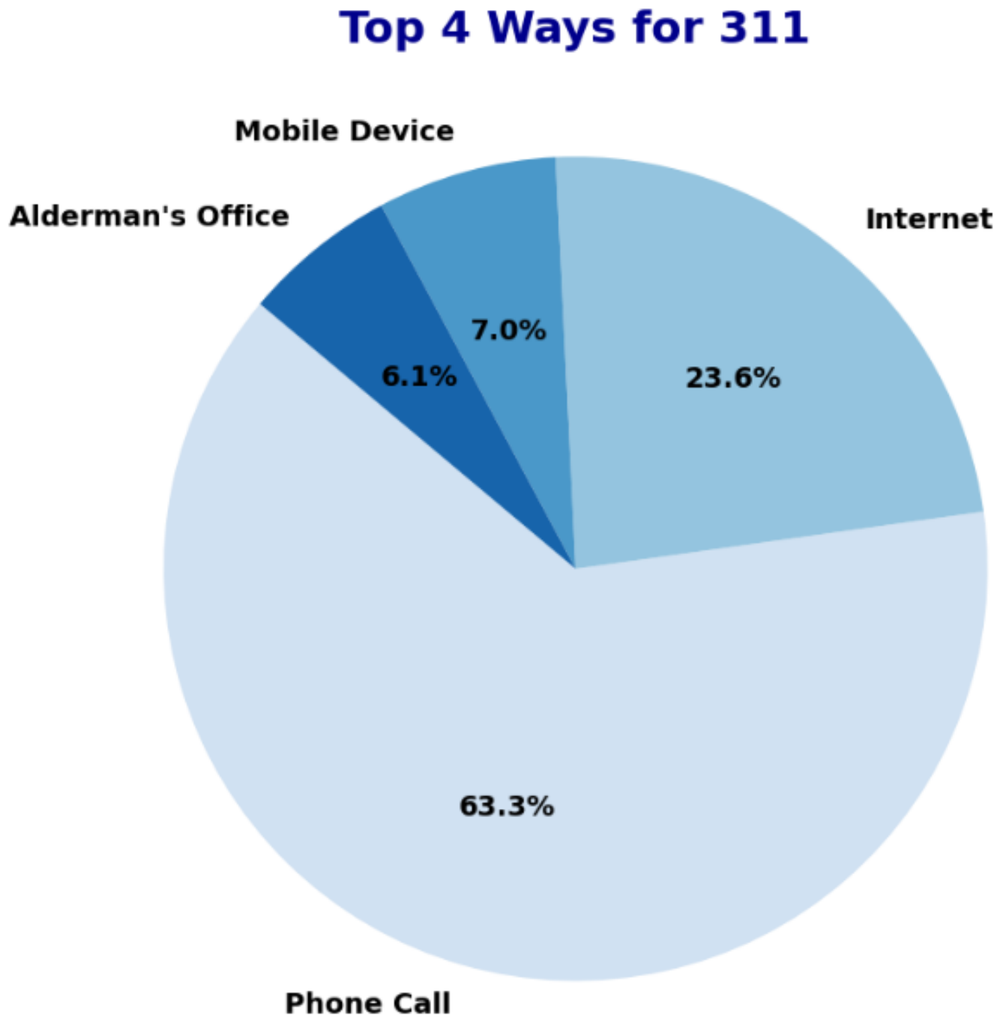
EXAMPLE: PERMIT DATA OVERVIEW

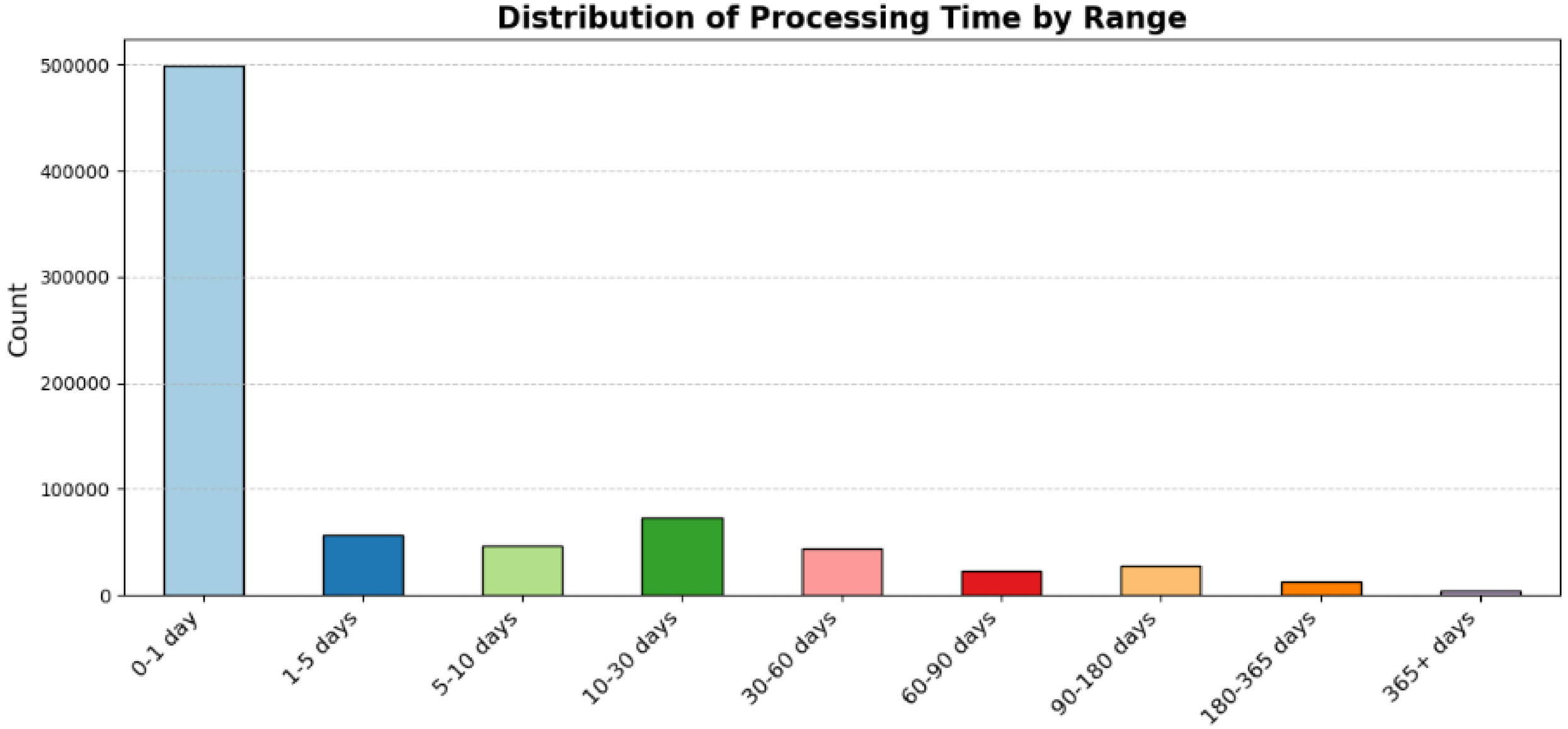
	building_fee_paid	zoning_fee_paid	other_fee_paid	subtotal_paid	building_fee_unpaid	zoning_fee_unpaid	other_fee_unpaid	subtotal_unpaid	building_fee_waived	total_fee
count	95,038.00	95,038.00	95,038.00	95,038.00	95,038.00	95,038.00	95,038.00	95,038.00	95,038.00	95,038.00
mean	824.74	44.10	64.28	933.12	4.38	0.08	0.51	4.96	247.21	1,200.91
std	6,618.40	250.45	1,871.88	7,715.12	201.60	6.05	33.62	207.81	6,820.63	10,685.99
min	0.00	0.00	0.00	0.00	-1,000.00	0.00	0.00	-1,000.00	0.00	0.00
25%	75.00	0.00	0.00	75.00	0.00	0.00	0.00	0.00	0.00	75.00
50%	150.00	0.00	0.00	200.00	0.00	0.00	0.00	0.00	0.00	225.00
75%	475.00	75.00	0.00	525.00	0.00	0.00	0.00	0.00	0.00	550.00
max	722,911.00	46,426.00	243,514.00	799,744.00	51,637.00	1,188.00	8,306.00	51,637.00	1,133,182.00	1,134,257.00

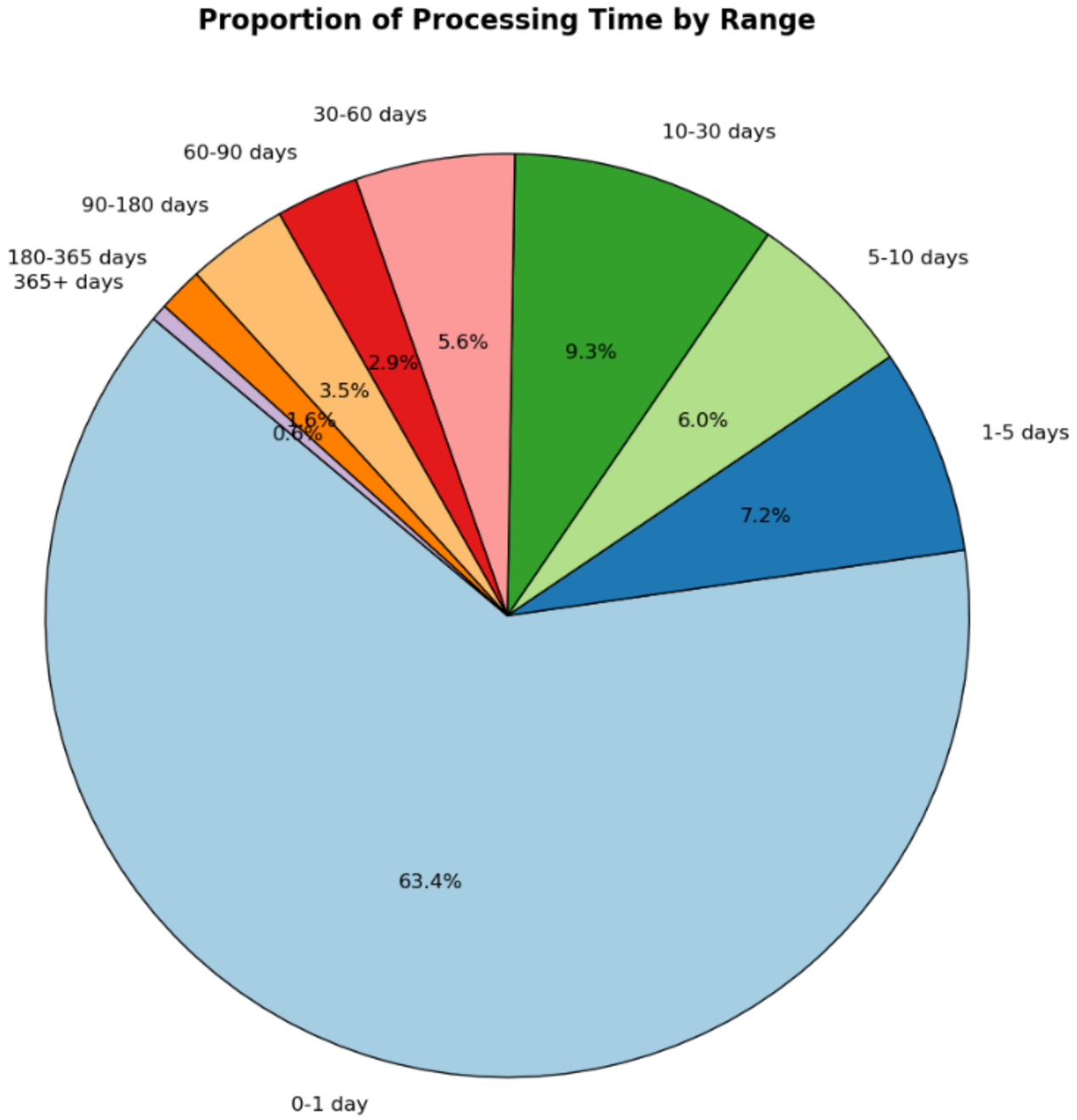

311 Distribution of Created Date by Hour
Trend Line
Frequency
Frequency
20000
15000
10000
5000
0
0
5
10
15
20
Hour of Day


311 Distribution of Created Date by Month
Trend Line
Frequency
25000
20000
15000
10000
5000
0
Frequency
2
4
6
8
10
12
Month


Top 4 Ways for 311
Frequency
140000
120000
100000
80000
60000
40000
20000
0
Phone Call
Internet
Mobile Device
Alderman's Office


Top 4 Ways for 311
Mobile Device
Alderman's Office
Internet
7.0%
6.1%
23.6%
63.3%
Phone Call


Distribution of Processing Time by Range
Count
500000
400000
300000
200000
100000
0
0-1 day
1-5 days
5-10 days
10-30 days
30-60 days
60-90 days
90-180 days
180-365 days
365+ days

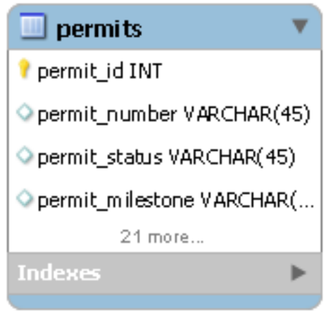

Proportion of Processing Time by Range
30-60 days
10-30 days
60-90 days
90-180 days
5-10 days
180-365 days
365+ days
1-5 days
5.6%
9.3%
2.9%
6.0%
3.5%
1.6%
0.7%
7.2%
63.4%
0-1 day

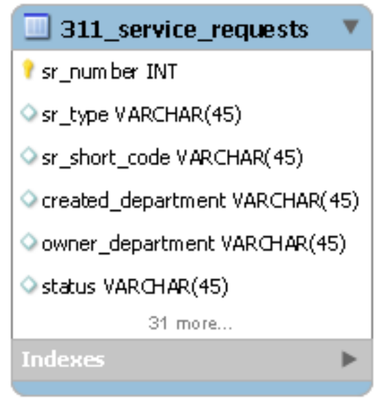
WHAT DOES BUILDFLAKE LOOK LIKE?

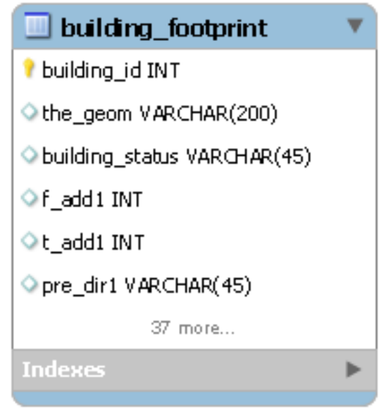
(BUILDFLAKE DATA MODEL CREATION OVERVIEW)

WHAT DOES THE BUSINESS DO?

Original Business Data Model (Conceptual)







SOURCE OF DATA

CHICAGO DATA PORTAL

Chicago Data Portal

Data Portal System

- Data separated to multiple tables
- Individual ID with no linkage between tables
- Single Source (Data Portal)
- “Generally” linked by address, address separated into multiple columns
- Data lookup times through API or otherwise can take a while

DATA_T1

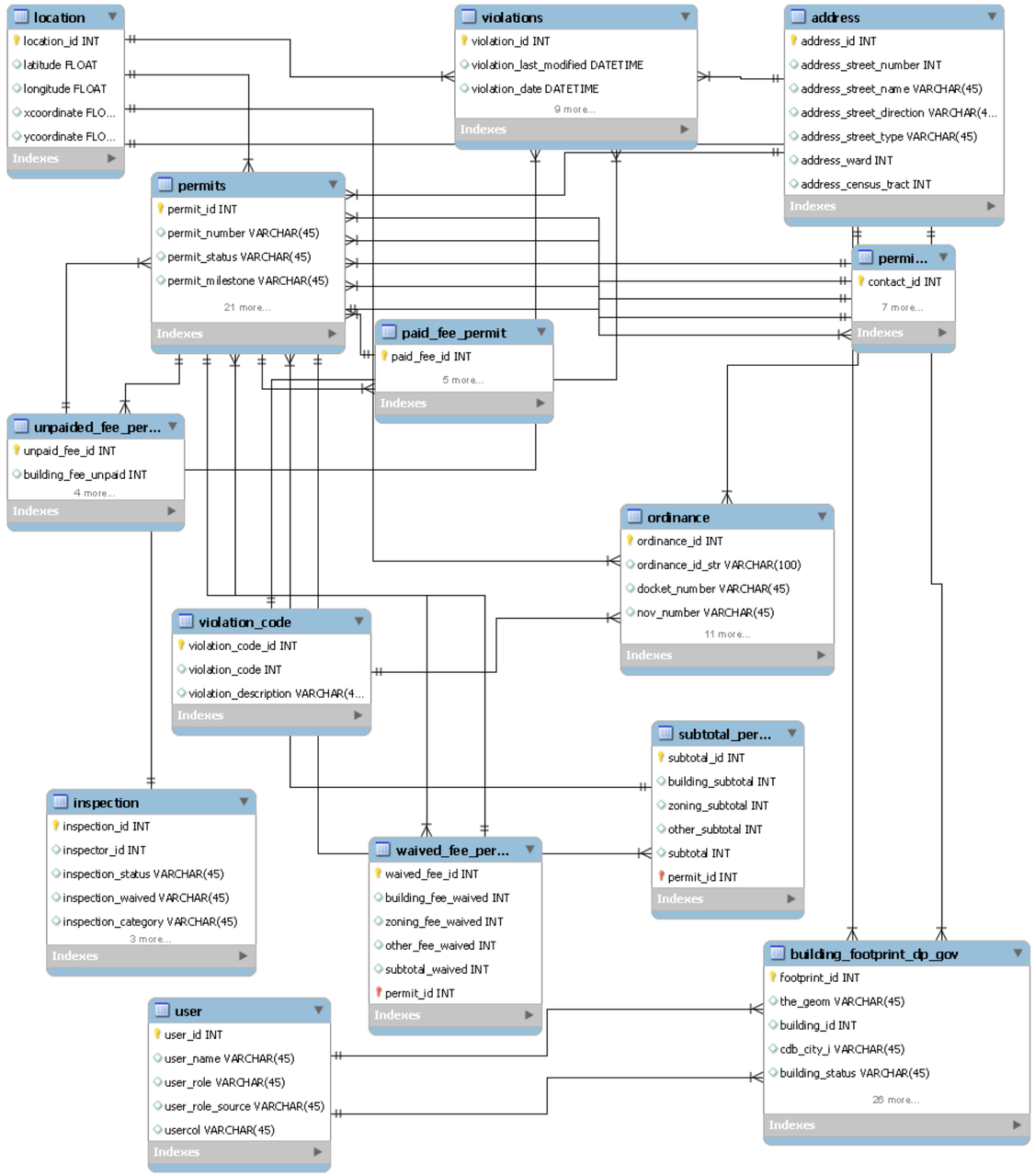
WHERE DID BUILDFLAKE START?

DATA MODEL (OLTP, LOGICAL MODEL)

SOURCES OF DATA:

CHICAGO DATA PORTAL

CHICAGO koordinates

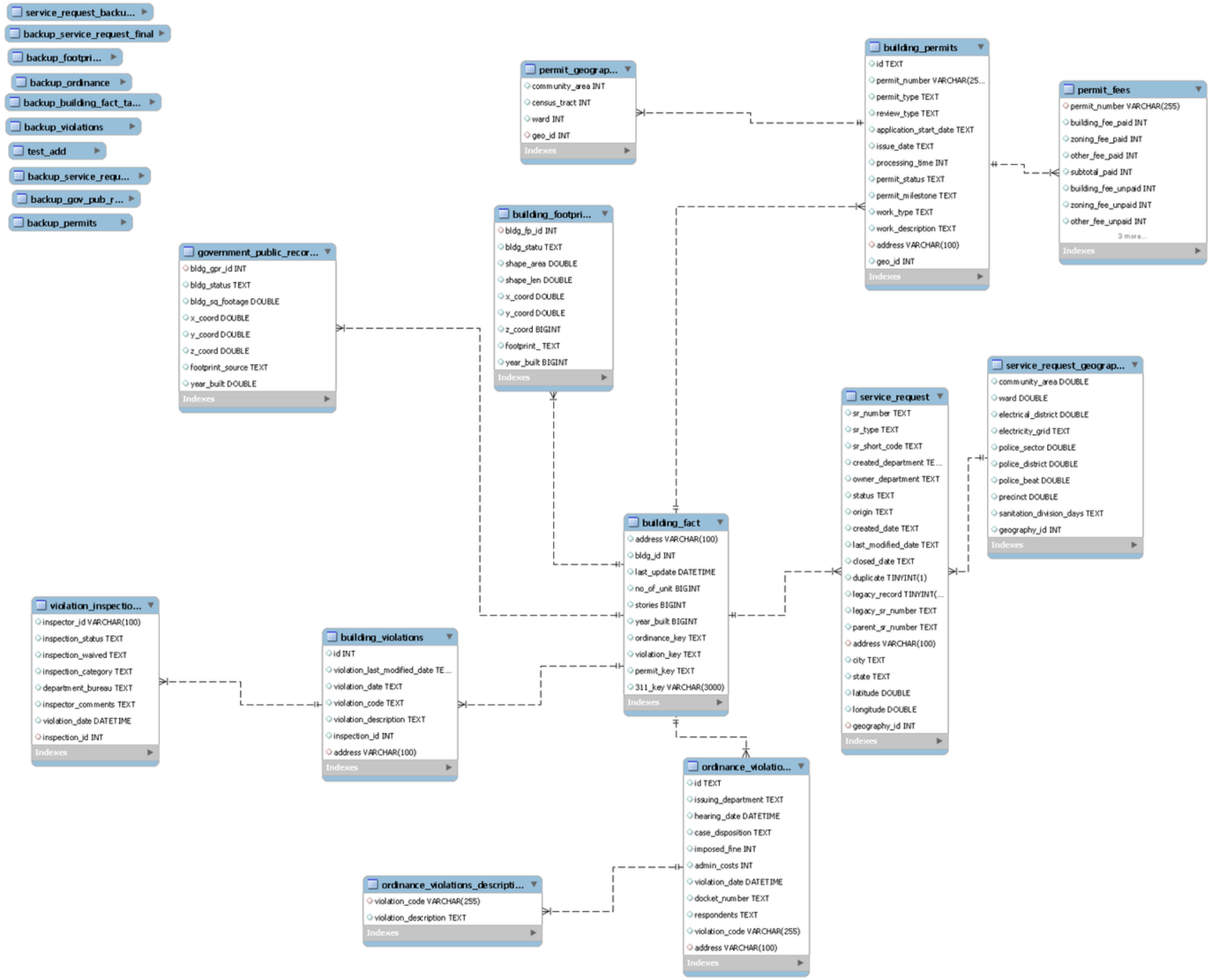


Our Preliminary System Design

- Data in multiple tables linked via address and location id table
- Normalized
- Sources data from the data portal and outside resources (Chicago Building Department Records, Koordinates)

BUILDFLAKE'S FINAL FORM

DATA MODEL: BUILDFLAKE (OLAP, PHYSICAL MODEL)



Our Final System Design

- Data in multiple tables linked via address
- Dimension tables built around building records (building_fact face table) linked by address
- Denormalized for faster analysis and queries
- Over 13,000 unique addresses

NORMALIZATION AND DENORMALIZATION

Normalization (OLTP)

1NF
- Checked for and removed duplicate rows

2NF
- Aspects of permits (permit fees, permit contacts)
- Footprint data users are separated into new table
- Inspection information for a violation is separated into a new table

3NF
- Location and Address are given IDs and placed into separate tables

Denormalization

- The address itself becomes the primary link between all tables
- Tables such as permit contacts and users are removed
- Locational information removed or attached to the dimension table

BUILDFLAKE DESIGN

Table Name	Table Type	Details
building_fact	Fact Table	This table contains several building specific ids linking to other tables. The table itself is indexed by address and contains some base information about buildings such as the number of units and year built.
building_permits	Dimensional Table	This table contains the permits listed for a given address as well as information such as status, type, and work description.
service_request	Dimensional Table	This table contains records of 311 service requests made to addresses. The table contains information on type, origin, and description of the request.
odinance_violations	Dimensional Table	This table contains records of ordinance violations by address. The table contains information such as the fine, case disposition, and violation code.
build_footprints	Dimensional Table	This table contains the governmental building footprint data for addresses in Chicago, providing information on the building statuses and coordinates.
government_public_records	Dimensional Table	This table contains public building footprint data for addresses in Chicago, providing information on building statuses and coordinates.

BUILDFLAKE DESIGN

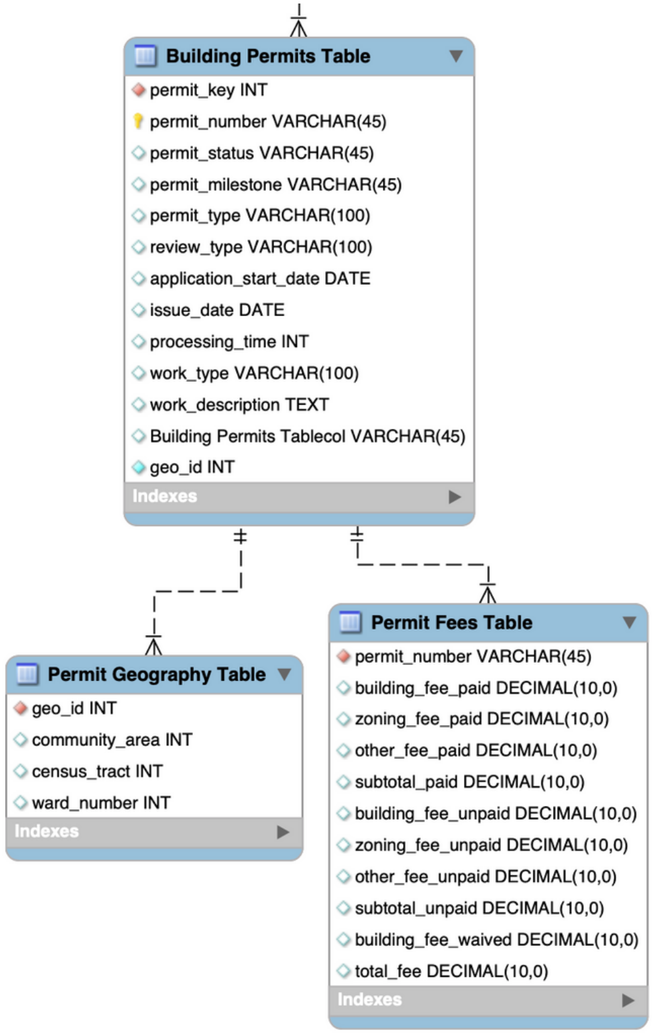
Table Name	Table Type	Details
permit_fees	Sub-Dimensional Table	This table contains the fee information for specific permits based on their permit number.
permit_geography	Sub-Dimensional Table	This table contains geographical information on permits as specified by a geography id.
violation_inspections	Sub-Dimensional Table	This table contains the inspection information for a violations, identified by the inspection id.
ordinance_violation_description	Sub-Dimensional Table	This table contains the text descriptions of ordinance violation codes.
service_request_geography	Sub-Dimensional Table	This table contains geographical information on service requests as specified by a geography id.

BUILDING BUILDFLAKE'S DATA

(DATA PROFILING AND CLEANING OVERVIEW)

DATA PROFILING

Permits Table as an Example



Normalization

- **Create Fact and Demention Tables**
- **Create Artificial Keys**
- **Implement Relationships**
- **Add Foreign Key Constraints**

Created permit geography id and permit number as artificial keys to ensure unique and consistent identifiers across permit table, permit geography table, and permit fees table.

Denormalization

Denormalized street number, street direction, street name and type to one column address, using it to identify unique buildings among tables

Column Names Standardization

Rename and Categorize Columns

Renamed all columns to consistent format
Renamed permit number, geo_id to make sure they are consistent in every tables

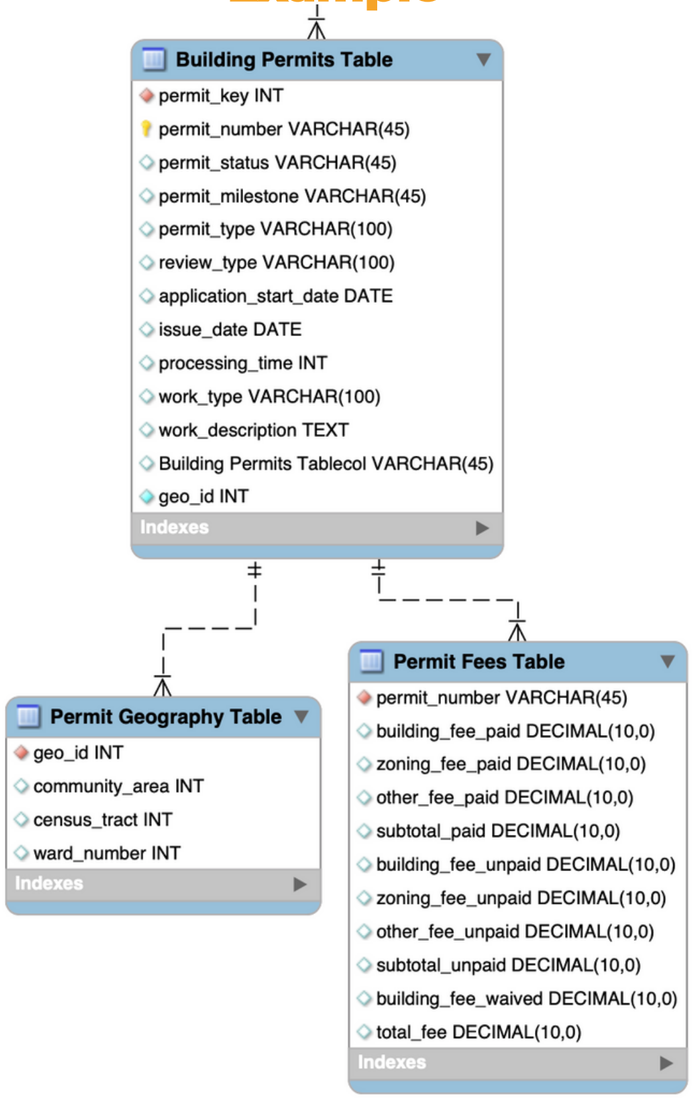
Data Type Vailidation

Assign Correct Data Type

Converted geo_id from VARCHAR to INT where applicable
Adjusted mismatched data types

DATA CLEANING

Permits Table as an Example



After conducting our EDA, we identified inconsistencies, missing values, and outliers in the building datasets. We thoroughly investigated these issues and applied data cleaning techniques to ensure data quality and reliability.

	Observation	Data Cleaning Actions
Missing Values	Missing value in rows of permit geography table	Drop missing rows before adding artificial keys (geo_id)
Outliers	Outliers in processing_time, such as -1 days	Drop negative value in processing_time
Data Duplication	Raw dataset has location, x &y coordinates, and address information	Drop location, x and y coordinates column

HOW WE BUILT BUILDFLAKE
(ETL AND TOOLS OVERVIEW)

TOOLS USED THROUGHOUT THE PROCESS

Original Data Sources

CHICAGO DATA PORTAL

CHICAGO

koordinates

EDA and ETL

Data Warehouse

Google Cloud Platform

Data Visualization and Reporting

+ableau

TOOLS USED THROUGHOUT THE PROCESS

- Google Cloud Platform
  - Group has the most experience with GCP compared to other cloud platforms
  - UChicago Organization allows for easy storage and lookup of our project
  - SQL functionality
  - IAM Specification
  - Scheduled maintenance and auto-scaling

TOOLS USED THROUGHOUT THE PROCESS

- MySQL Workbench
  - Create and work with ER Diagrams
  - Quick and easy query execution

TOOLS USED THROUGHOUT THE PROCESS

- Python
  - Helpful in aggregating, cleaning, preparing, and eventually sending data to the cloud
  - Jupyter Notebook allows for effective organization of work
  - Integral to EDA
  - Notable packages: Socrata, Pandas, Matplotlib

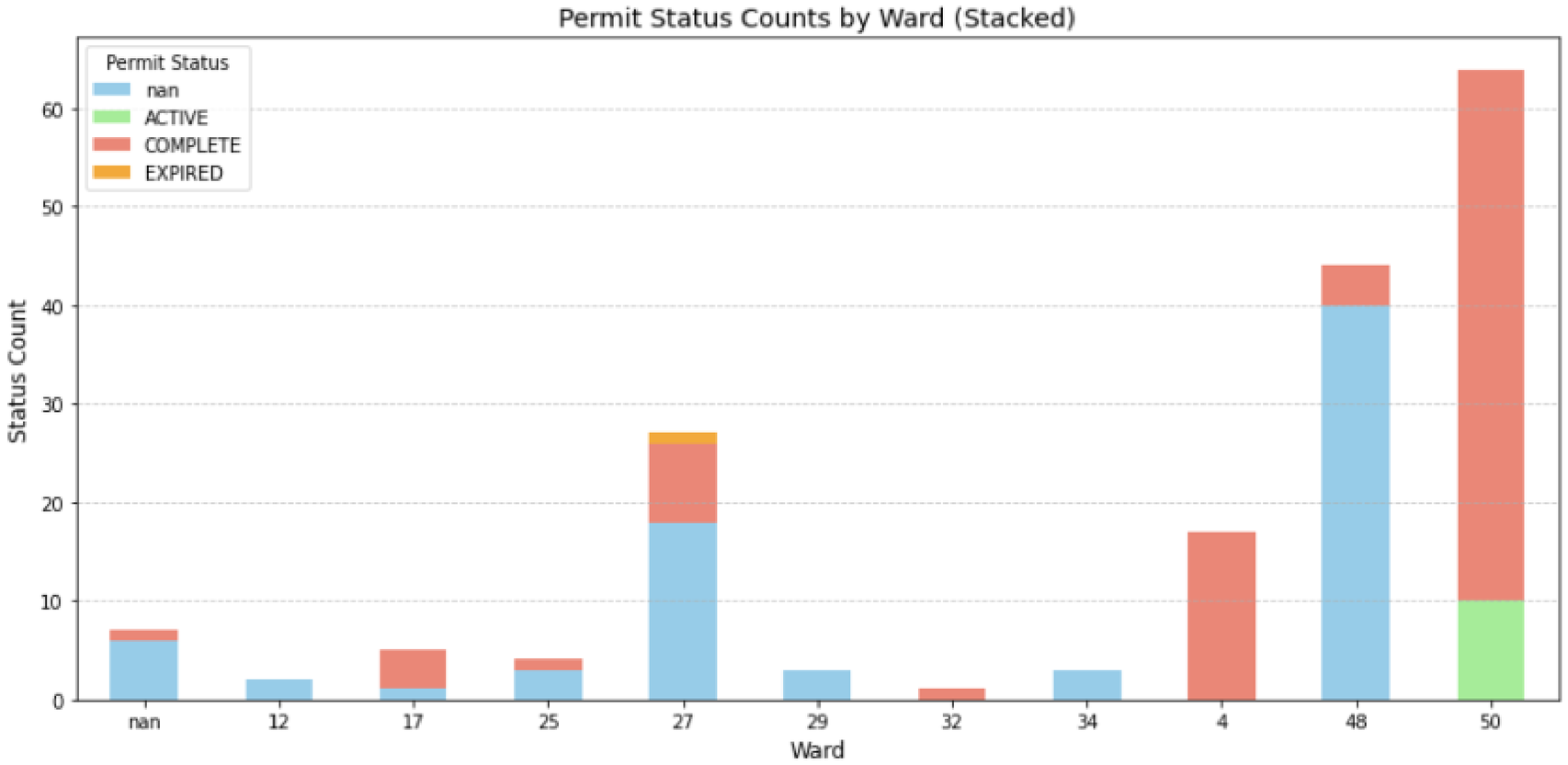
TOOLS USED THROUGHOUT THE PROCESS

- Tableau
- Useful for creating BI reports
- Connects directly to our SQL instance
- Dashboards help put together business information

WHAT DO WE KNOW ABOUT BUILDINGS IN CHICAGO

(BUSINESS INTELLIGENCE INSIGHTS FROM BUILDFLAKE)

PERMIT STATUS: WHERE TO WORK ON



Suspended Permits

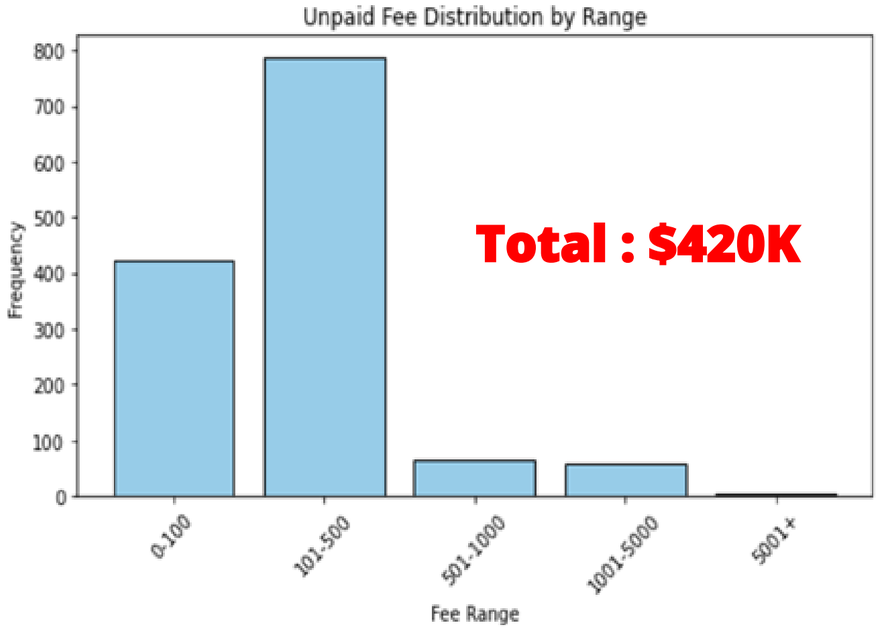
- Highlights delays or compliance issues.

Action Plan

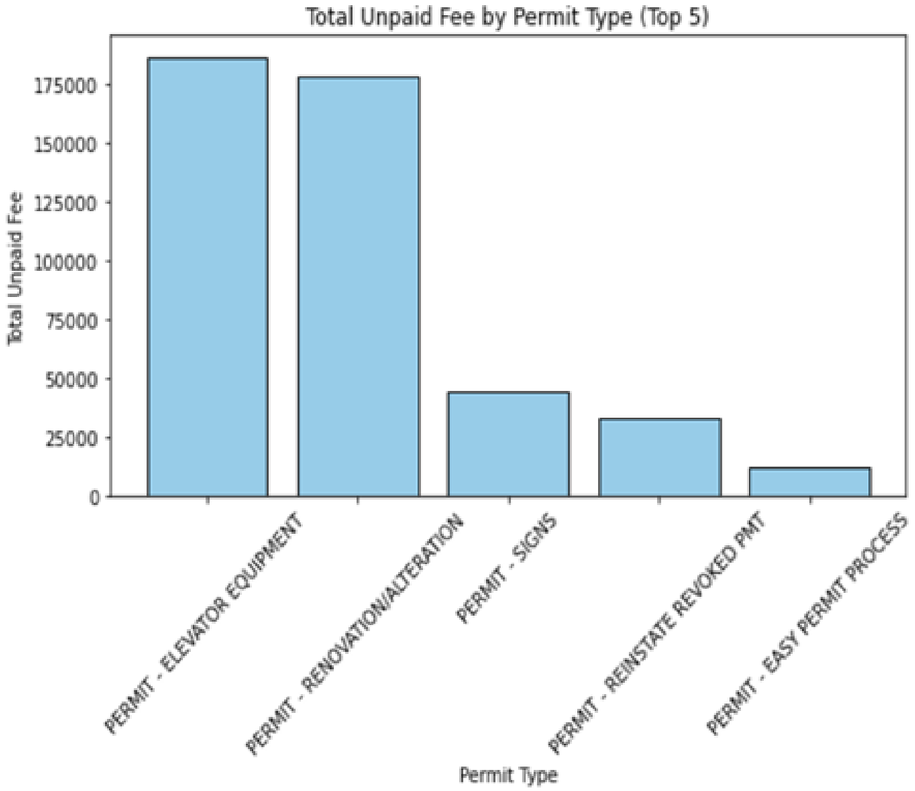
- Investigate and prioritize wards with the highest delayed permits.
- Resolve to optimize processing and reduce delays.

```
query = """
SELECT
    ward,
    permit_status,
    COUNT(*) AS status_count
FROM
    building_permits bp
JOIN
    permit_geography pg
ON
    bp.geo_id = pg.geo_id
GROUP BY
    ward, permit_status
ORDER BY
    ward ASC, status_count DESC;
"""

# Execute the query and load results into a DataFrame
df = pd.read_sql(query, engine)

df = pd.DataFrame(df)
df.columns
```

UNPAID FEE: EXTRA BUDGET FOR GOVERNMENT



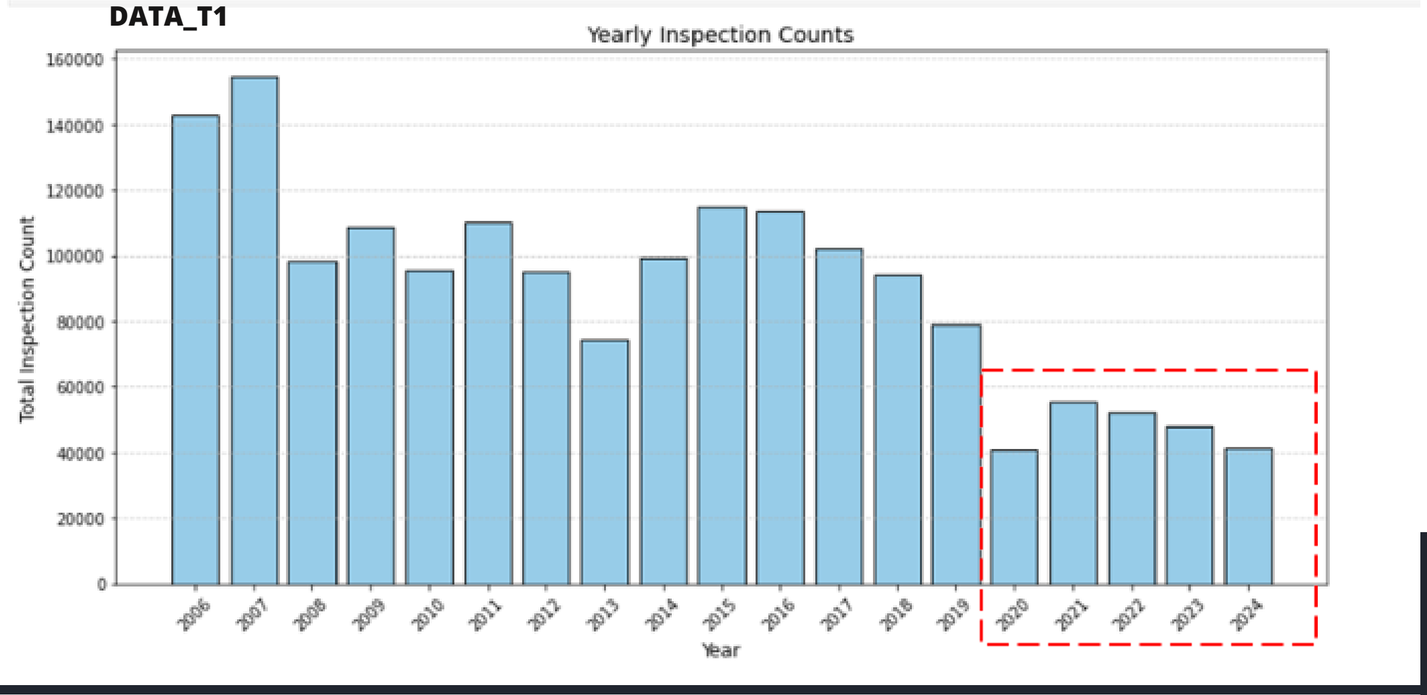
Total : $420K

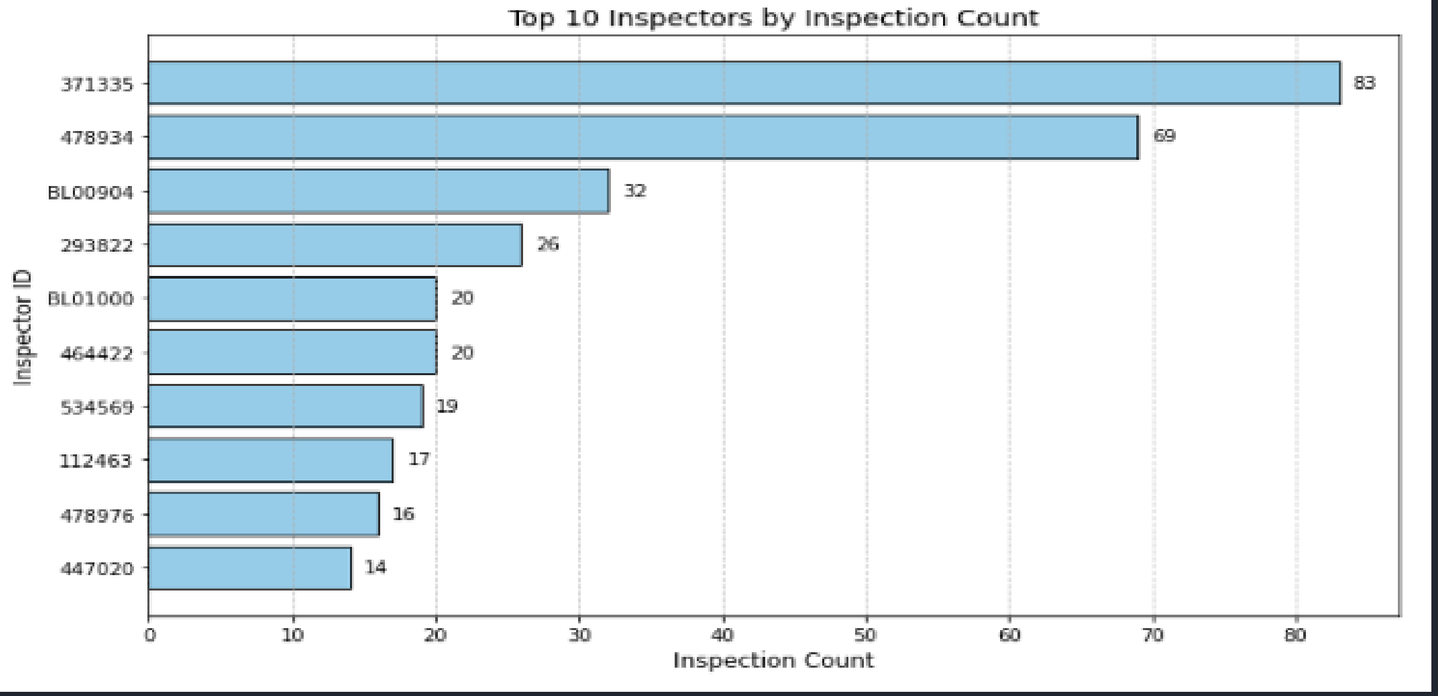


1. Total Unpaid Fees: $420K
2. Major Fee Types:
   a. Elevator Equipment
   b. Renovation/Alteration
3. Next Steps:
   a. Identify high-impact areas.
   b. Focus on recovery strategies.

```
#1.Unpaid permit fees (by type and description)
```

```
query = """
SELECT
    a.subtotal_unpaid,
    b.permit_type,
    b.work_description
FROM
    permit_fees a
LEFT JOIN
    building_permits b
ON
    a.permit_number = b.permit_number
WHERE
    a.subtotal_unpaid > 0;
"""

# Execute the query and load results into a DataFrame
df = pd.read_sql(query, engine)

# Display the result
print("\nQuery Results:")
print(df)
```





LINKING INSPECTION DATA TO WORKFORCE PRODUCTIVITY

The decline in inspection counts since 2020 is likely due to the COVID-19 pandemic, resource constraints, or policy changes. This highlights the need to reassess workforce utilization.

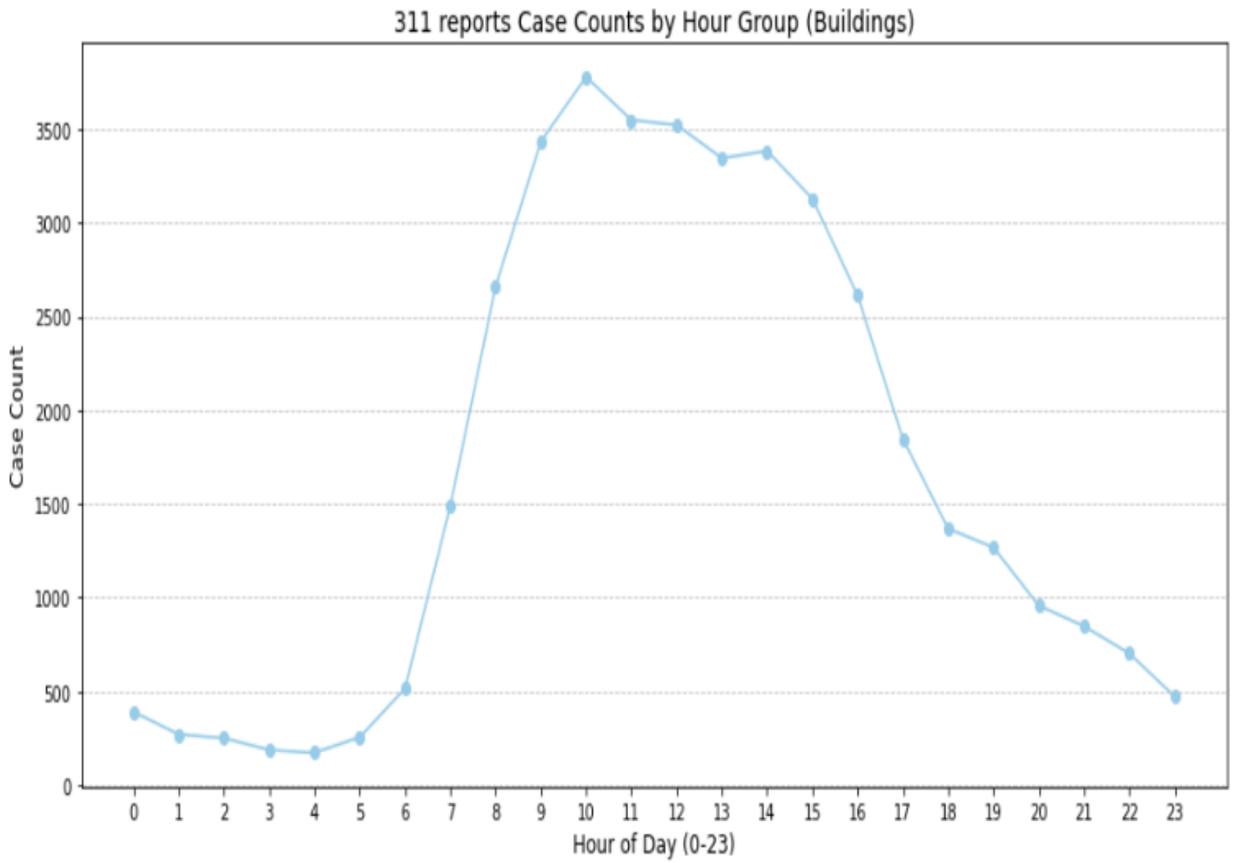
1. **Performance Analysis**
   - Optimizing task assignments can improve overall team productivity.
2. **Strategic Planning**
   - Insights from workload trends enable better forecasting and workforce planning.
   - Ensures consistent inspection capacity to meet demand.
3. **Resource Allocation**
   - Uneven workload distribution could lead to inefficiencies and burnout.

```
query = """
SELECT
    YEAR(violation_date) AS year,
    MONTH(violation_date) AS month,
    COUNT(*) AS inspection_count
FROM
    violation_inspections
GROUP BY
    YEAR(violation_date), MONTH(violation_date)
ORDER BY
    year DESC, month DESC;

"""
```

```
query='''
SELECT
    inspector_id,
    COUNT(*) AS inspection_count,
    min(YEAR(violation_date)) as 'month',
    min(MONTH(violation_date)) as 'year'
FROM
    violation_inspections
WHERE
    YEAR(violation_date) = YEAR(CURDATE()) AND
    MONTH(violation_date) = MONTH(CURDATE())
GROUP BY
    inspector_id
ORDER BY
    inspection_count DESC
LIMIT 10;
```

DATA_T1

311 REQUESTS – OPTIMIZING STAFFING BY TIME OF DAY



Key Insights

- **Peak Hours (10 AM – 3 PM):** High request volume indicates the need for adaptable staffing during this period.
- **Low Activity (12 AM – 6 AM):** Minimal requests suggest the opportunity to adjust staffing levels and allocate resources efficiently.

Actionable Steps

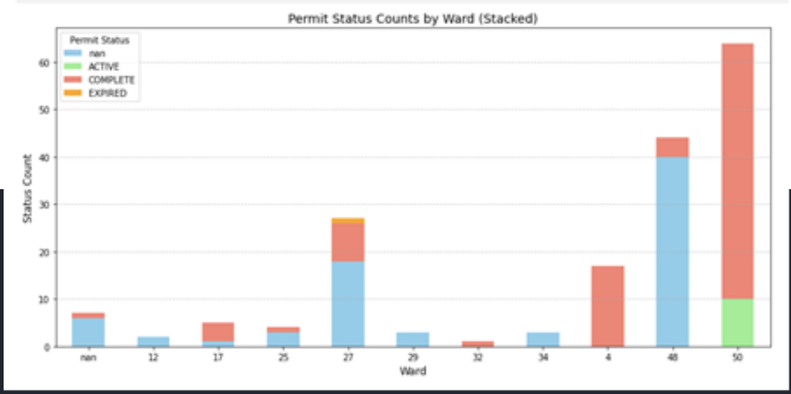
1. **Flexible Staffing Plan:** Align workforce availability with request patterns to enhance efficiency.
2. **Resource Allocation:** Minimize overstaffing during quiet hours to reduce costs while maintaining response quality.
3. **Data-Informed Decisions:** Use real-time monitoring and request patterns to continuously optimize staffing and ensure responsiveness.

```
query = """
SELECT
    HOUR(STR_TO_DATE(created_date, '%%m/%%d/%%Y %%r')) AS hour_group,
    COUNT(*) AS case_count
FROM
    service_request
WHERE
    created_date IS NOT NULL
GROUP BY
    hour_group
ORDER BY
    hour_group ASC;
"""

# Execute the query and load results into a DataFrame
df = pd.read_sql(query, engine)

# Display the result
print("\nQuery Results:")
print(df)
```

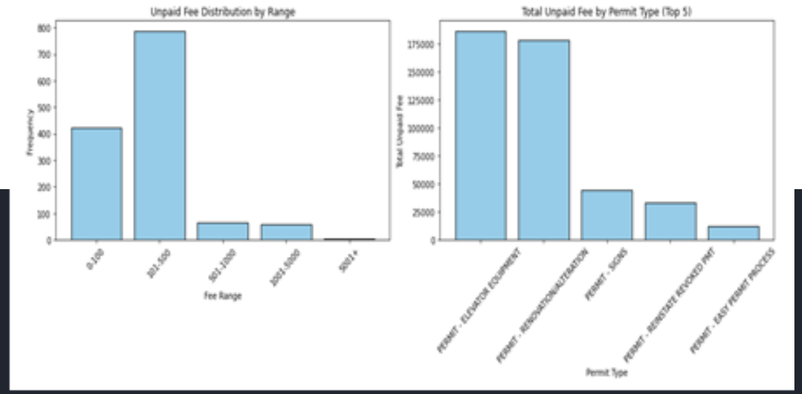
DATA_T1

SUSTAINABLE BUDGET OPTIMIZATION CYCLE

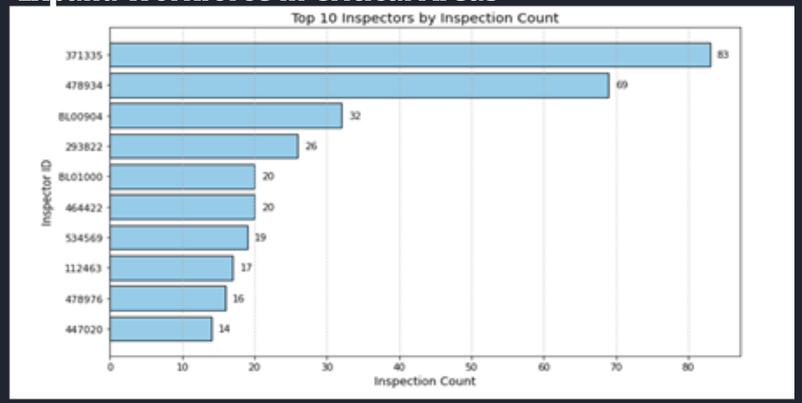
Identify Budget Expansion Opportunities



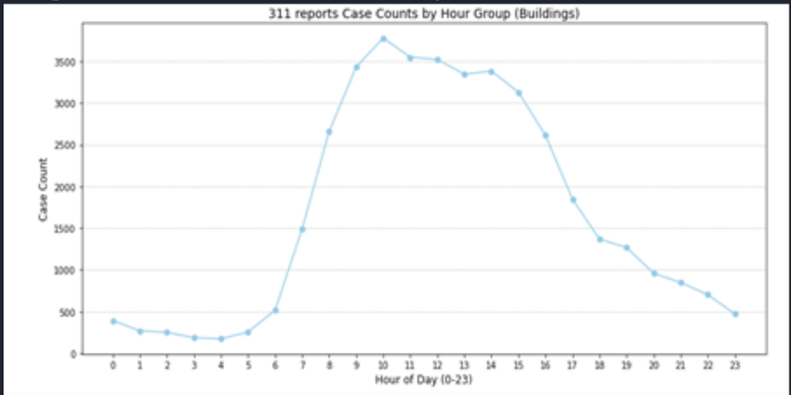
Secure Additional Budget



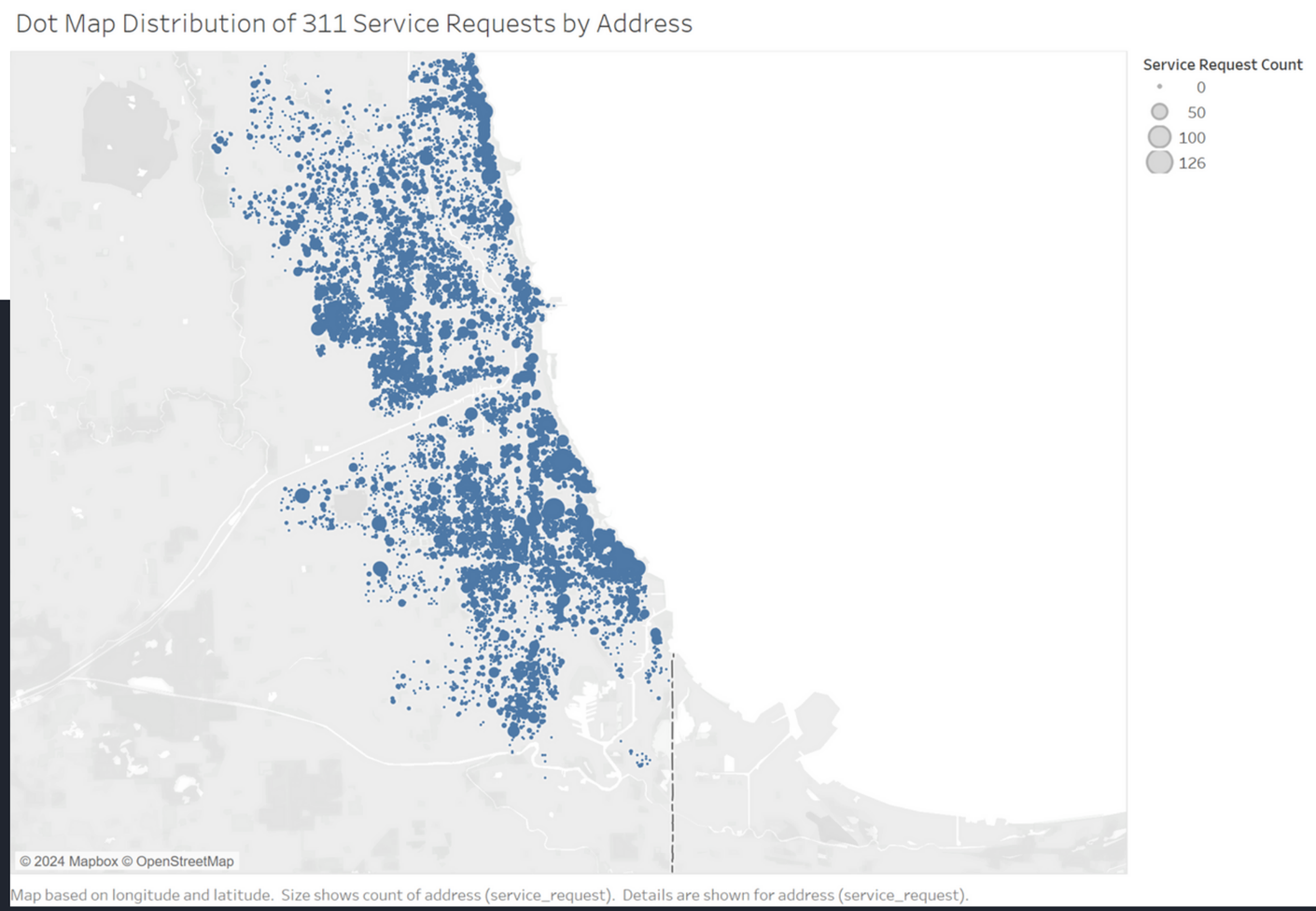
Expand Workforce in Critical Areas



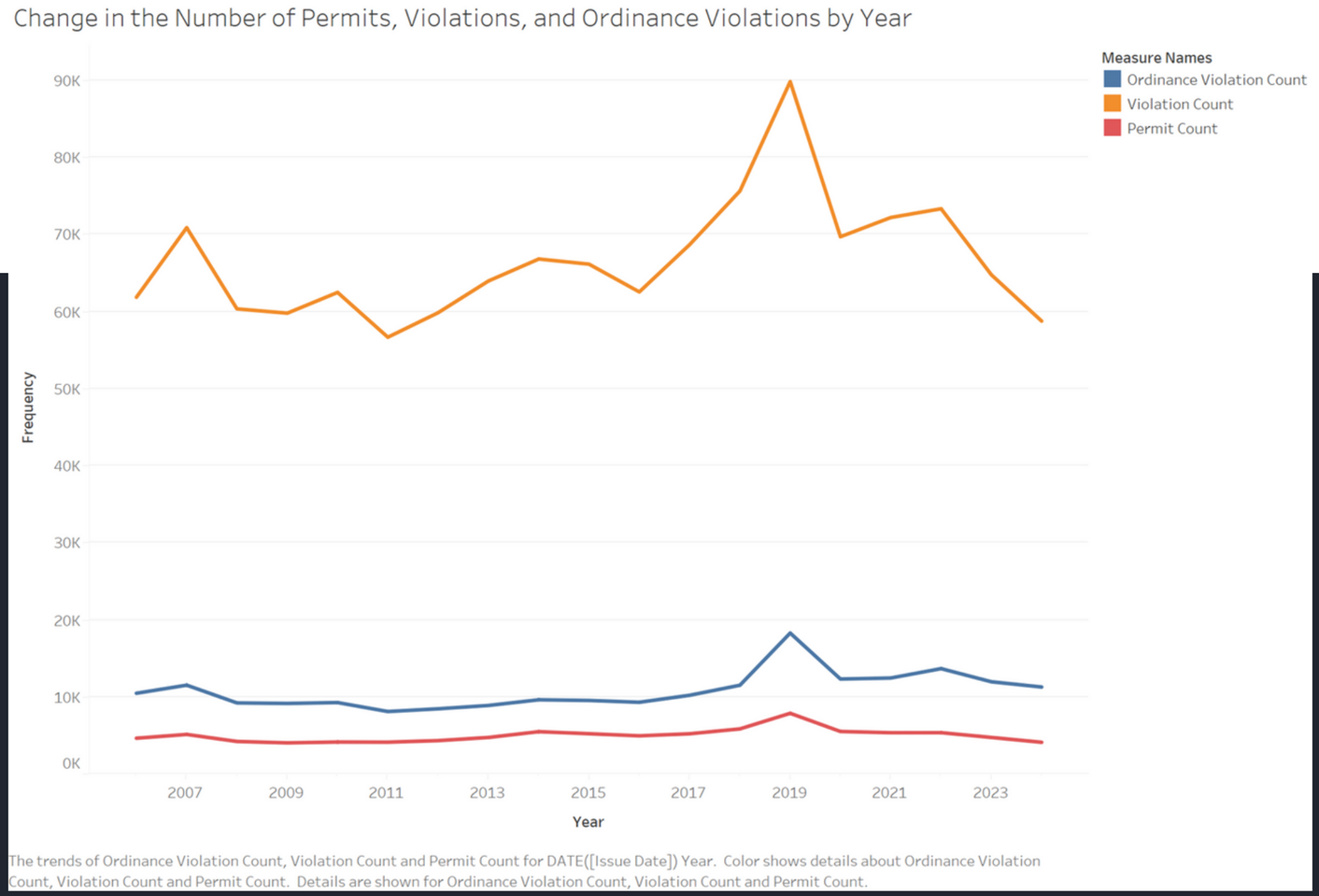
Optimize Workforce Efficiency



DISTRIBUTION OF 311 SERVICE REQUESTS



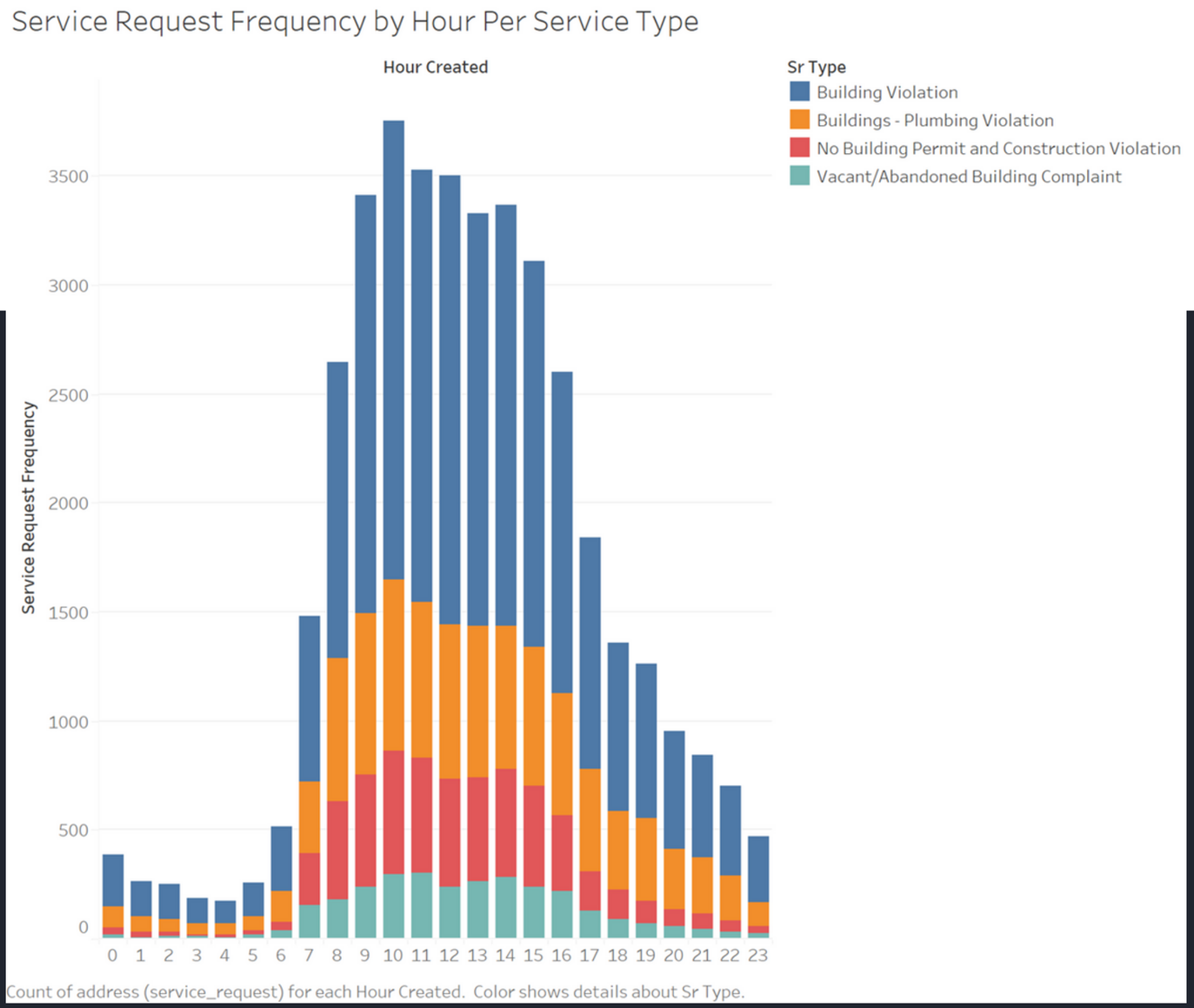
- More on the South and West parts of Chicago
- Most service requests: 4624 S Ellis Ave (apartment complex)
- Average Number of Service Requests: 4.25



CHANGE IN PERMITS, VIOLATION, AND ORDINANCES OVER TIME

- Violations, permits, and ordinances peak in 2019
- Tractable to real life changes
  - Comprehensive modernization of building code
  - New standard, leading to more write-ups

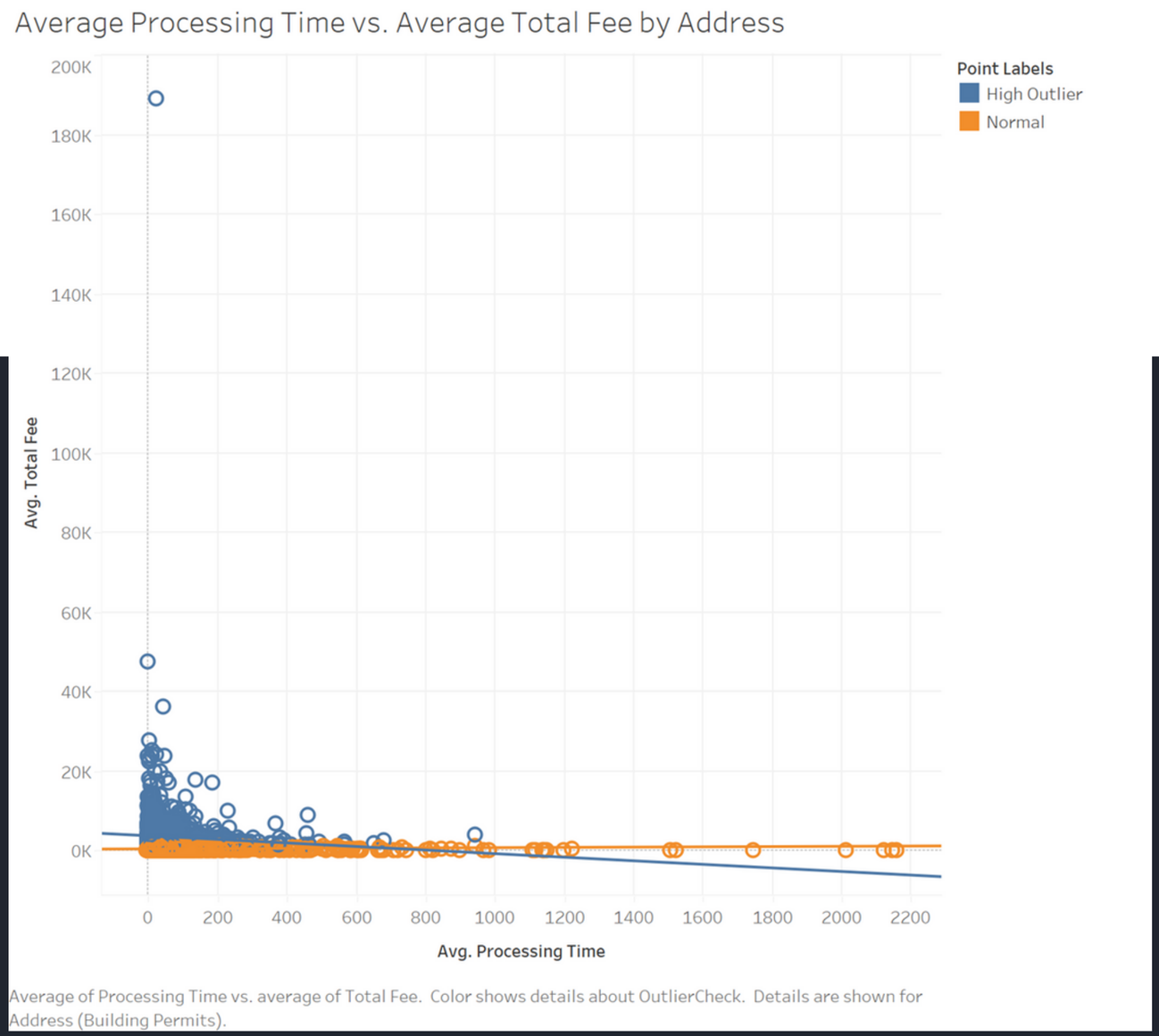
GRADY, GRAHAM C. "BUILDING CODE MODERNIZATION ORDINANCE: MY TAKE." BOMA/CHICAGO, APRIL 2, 2019. HTTPS://WWW.BOMACHICAGO.ORG/2019/04/02/BUILDING-CODE-MODERNIZATION-ORDINANCE-MY-TAKE/
CITY OF CHICAGO. "PROCEDURES FOR OPTIONAL USE OF 2019 CHICAGO CONSTRUCTION CODES." CODE MEMORANDUM 2019-003, DECEMBER 15, 2019. HTTPS://WWW.CHICAGO.GOV/CONTENT/DAM/CITY/DEPTS/BLDGS/GENERAL/CODEMODERNIZATION/IMPLEMENTATION%20MEMO%2012-15-2019.PDF
BRADFORD MILLER LAW. "FIVE COMMON BUILDING CODE VIOLATIONS IN CHICAGO." BRADFORD MILLER LAW, JUNE 14, 2024. HTTPS://WWW.BRADFORDMILLERLAW.COM/FIVE-COMMON-BUILDING-CODE-VIOLATIONS-IN-CHICAGO/



DISTRIBUTION OF SERVICE REQUEST FREQUENCY BY TYPE

- Most fell under general building violations
- Least common was vacant/abandoned buildings
- Most calls are in the middle of the day

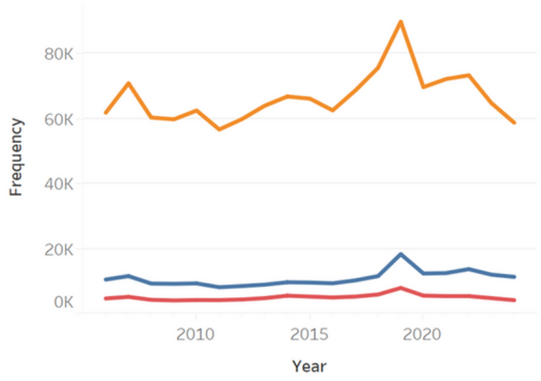
DATA_T1



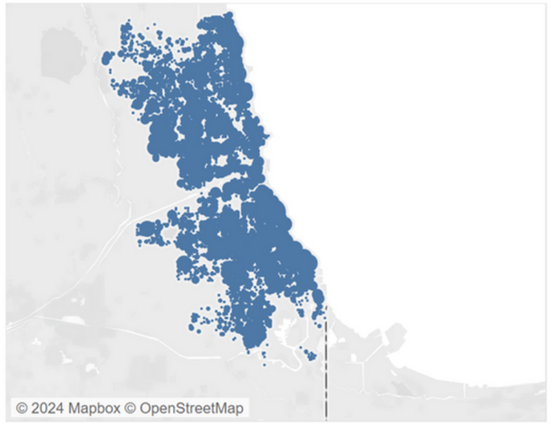
DISTRIBUTION OF SERVICE REQUEST FREQUENCY BY TYPE

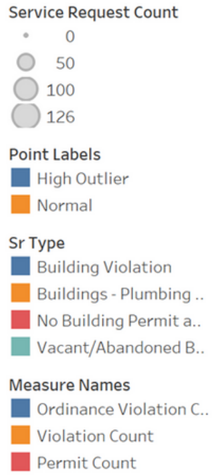
- Tons of high outliers lead to slightly different fit lines
- In general, the higher the average fee, the lower the processing time is for that fee

Change in the Number of Permits, Violations, and Ordinance Violations by Year

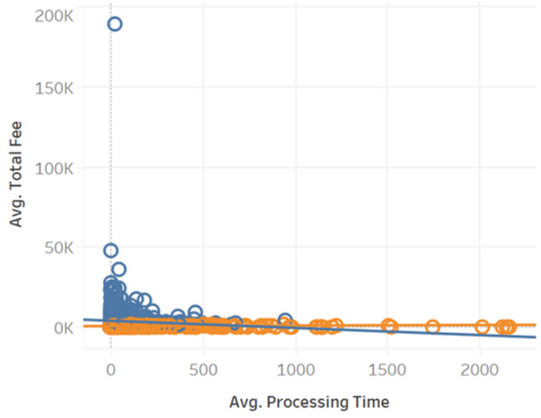


Dot Map Distribution of 311 Service Requests by Address

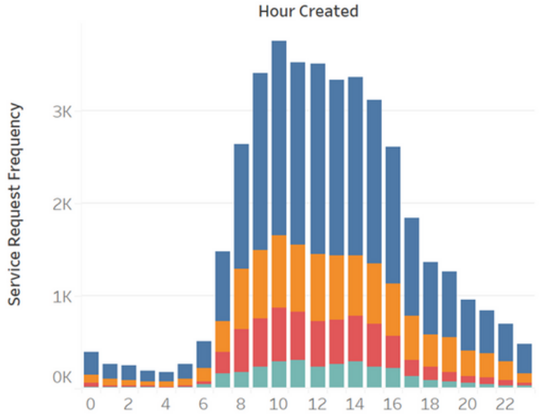




Average Processing Time vs. Average Total Fee by Address



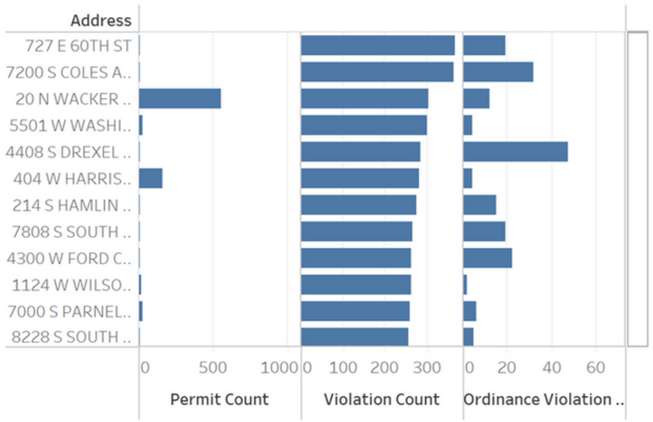
Service Request Frequency by Hour Per Service Type



DASHBOARD 1: VISUAL STATISTICS

- Easier to see general trends
- Color-coded legends make tracking specific aspects of this building information more feasible
- Takes 2-3 seconds to execute

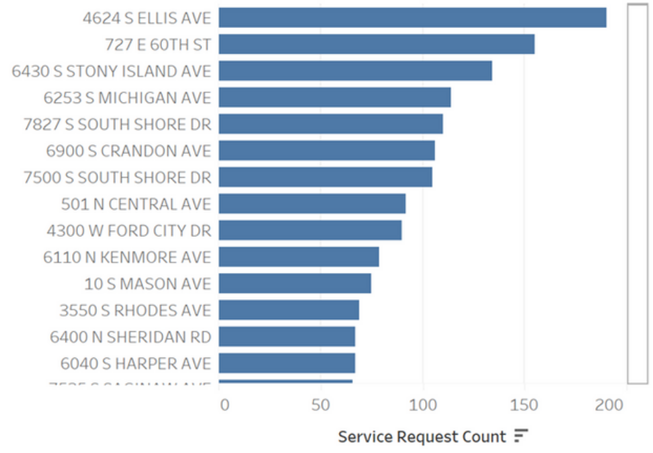
Count of Permits, Building Violations, and Ordinance Violations by Address



Count of Permits for Addresses by Permit Status

	Permit Status		
Address	ACTIVE	CANCELLED	COMPLETE
20 N WACK..	304		304
1124 W WIL..			262
200 E RAND..	185		185
505 W BEL..	178		178
7827 S SOU..			165
500 W MAD..	160	160	160
130 E RAND..	160		160
141 W JACK..	159		159
1000 W WA..	153	153	153
151 E WAC..	148		148
2501 W AD..	146		146
1 S WACKE..	139		139
4520 S DRE..	136		136
900 N MICH..	134	134	134

Count of Service Requests by Address



Subtotal Paid, Unpaid, and Total Fee for Permits by Address

	Sum of Subtotal Paid	Sum of Subtotal Unpaid	Total Fe
1 E 8TH ST	167,870	695	
1 E ERIE ST	57,050	150	
1 E JACKSO..	31,107	500	
1 E SUPERI..	4,305	2,475	
1 E WACKE..	360,825	0	
1 N HALSTE..	59,187	0	
1 N LA SALL..	242,090	1,298	
1 N STATE ST	350,499	150	
1 N WACKE..	1,130,511	100	
1 S FRANKL..	256,613	600	
1 S HALSTE..	2,230	0	
1 S STATE ST	229,744	315	
1 S WACKE..	817,865	560	
1 W GRAND..	14,926	0	

DASHBOARD 2: GRANULAR, ADDRESS LEVEL STATISTICS

- Allows you to check on individual property stats
- Execution takes about 2-3 seconds to complete
- Clear labeling and sizing

HOW GOOD IS BUILDFLAKE?
(SOLUTION OVERVIEW AND REVIEW)

Speed/Performance Comparison

Chicago Data Portal

Data	API Call Time
Building Data	76.49 secs
Permits Data	205.49 secs
Violation Data	702.94 secs

BuildFlake

- Query Execution Time: <1 second
- Business Intelligence Graphs and Statistics: 2-3 seconds (Max 10 seconds)

BUSINESS MONETARY VALUE

- Business Monetary Value
- Department of Technology budget: ~$4 million Average cost for data portal system (see Los Angeles, New York): $200-$400 thousand Can reduce these overhead costs with our optimizations

RECOMMENDATIONS AND LESSONS LEARNED

Recommendations

- Re-input building data from addresses not in our original building data
- Weekly/monthly maintenance checks on the cloud instance
- Automate data

Lessons Learned

- Start with OLTP, move into OLAP
- Keep an eye on cloud costs
- When gathering and modeling data, start early!

APPENDIX

REFERENCES

- City of Chicago Data Portal. "City of Chicago Data Portal." Accessed December 10, 2024. https://data.cityofchicago.org/.
- City of Chicago Data Portal. "Vacant and Abandoned Buildings Violations." Last modified December 10, 2024. https://data.cityofchicago.org/Buildings/Vacant-and-Abandoned-Buildings-Violations/kc9i-wq85/about_data.
- City of Chicago Data Portal. "Building Violations (Circa 2006 to Present)." Last modified December 10, 2024. https://data.cityofchicago.org/Buildings/Building-Violations/22u3-xenr/about_data.
- City of Chicago Data Portal. "Ordinance Violations (Filled During Department of Administrative Hearings)." Last modified December 10, 2024. https://data.cityofchicago.org/Administration-Finance/Ordinance-Violations-Buildings-/awqx-tuwv/data_preview.
- City of Chicago Data Portal. "Building Permits (Circa 2006 to Present)." Last modified December 10, 2024. https://data.cityofchicago.org/Buildings/Building-Permits/ydr8-5enu/about_data.
- City of Chicago Data Portal. "Buildings (Building Footprints Data)." Last modified December 10, 2024. https://data.cityofchicago.org/Buildings/buildings/syp8-uezg/about_data.
- Koordinates. "Chicago Buildings." Accessed December 10, 2024. https://koordinates.com/layer/97859-chicago-buildings/.
- City of Chicago. "Building Violations." Accessed December 10, 2024. https://www.chicago.gov/city/en/depts/bldgs/provdrs/inspect/svcs/building_violationsonline.html.
- U.S. Census Bureau. "Chicago City, Illinois." Last modified December 10, 2024. https://data.census.gov/profile/Chicago_city,_Illinois?g=160XX00US1714000.
- City of Chicago. Departmental Budget Hearings Book. Accessed December 10, 2024. https://www.chicago.gov/content/dam/city/depts/COFA/COFA%20Departmental%20Budget%20Hearings%20Book.pdf.
- Governing. "The Cost and Problems of Open Data." Last modified December 10, 2024. https://www.governing.com/archive/gov-open-data-cost-problems.html.
- City of New York. "What Responsibilities Does Socrata Have with the NYC Open Data Portal, and What Is the Duration and Cost of Their Contract?" February 27, 2024. https://citymeetings.nyc/city-council/2024-02-27-1000-am-committee-on-technology/chapter/what-responsibilities-does-socrata-have-with-the-nyc-open-data-portal-and-what-is-the-duration-and-cost-of-their-contract?utm_source=chatgpt.com.