**BA820 - Project Milestone 2**

Cover Page

* **Project Title:** San Francisco Building Complaints Analysis
* **Section and Team Number:** B1 Team 7
* **Members:** Yu-Hsiang (Rick) Wang, Ching-Hsuan (Shawn) Lin, Kuang-Ching (Amanda) Ting, Ming-Hua (Jasmine) Tsai
* **Date:** Mar 3, 2025

# A. Proposal

## Problem statement

Building inspections are crucial for urban safety. A well-managed system can reduce risks like fires and building collapses while boosting a city’s image and economy, but many cities struggle with slow complaint resolution due to misclassification and inefficient processing. In San Francisco, the Department of Building Inspection allocated an average of $105 million annually [(2018–2022),](#_heading=h.3rdcrjn) but about 42% of cases take over three months to resolve, with 19% being reassigned multiple times. We aim to use unsupervised machine learning to help the government improve efficiency by identifying division assignment patterns and exploring communities prone to specific types of complaints to minimize the need for manual reassignment further and enable systematic regional issue resolution.

# B. Exploratory Data Analysis & Preprocessing

**Data Source:**

San Francisco Government's Open Data (<https://reurl.cc/eGKE5K>)

**Summary Statistics, Outliers and Patterns:**

Based on the mean, median, and standard deviation, the Street Number data is highly dispersed, while the ZIP Code values are more concentrated. The Supervisor District distribution appears relatively balanced across different areas. Since most of our data consists of categorical variables, we have not yet isolated outliers. [*(Appendix Figure 1)*](#_heading=h.17dp8vu)

**Preprocessing:**

Time-related fields such as data\_loaded\_at and data\_as\_of were deemed unnecessary and could introduce noise. For the Closed Date field, we dropped the column for further analysis because unresolved cases made it difficult to determine whether an empty value was due to missing data or an unresolved issue. Additionally, we removed excessive location details to retain only the ZIP Code and the Analysis Neighborhood. Regarding the complaint description column, we employed text mining techniques to transform sentences into topic labels, which were then incorporated into our subsequent analysis. A detailed explanation of the text mining steps will be provided later.

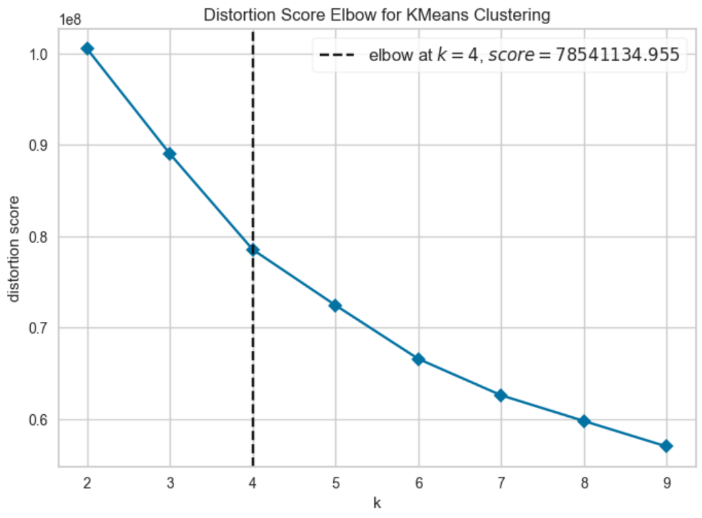
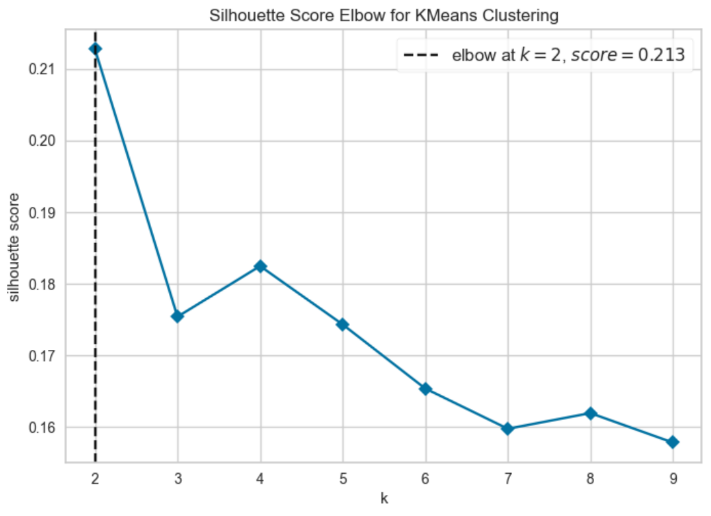
**Observation:**

We began by comparing the counts of complaints received and assigned by different divisions. Our analysis revealed that most complaints were assigned to the Housing Inspection Services Division and the Building Inspection Division.[*(Appendix Figure 2)*](#_heading=h.17dp8vu)

We then created a bar chart [*(Appendix Figure 3)*](#_heading=h.17dp8vu) to gain insight into the mismatch rates across departments. We found that although the Housing Inspection Services Division received the highest number of complaints, its mismatch rate was only about 1.38%, indicating a minimal difference between the cases received and those assigned. In contrast, the Plumbing Inspection Division and the Building Inspection Division, despite handling a high volume of complaints, exhibited relatively high mismatch rates, suggesting significant room for improvement. Lastly, and importantly, the Disabled Access Division, Help Desk/Technical Services, and Central Permit Bureau all had mismatch rates of 100%, meaning that every case received by these departments was ultimately reassigned rather than handled in-house.

We also created a bar chart to display the top ten areas with the highest number of complaints *(Appendix Figure 4)*. The Mission District stands out with a significantly higher volume, far surpassing Tenderloin, which ranks second. Both Nob Hill and Sunset/Parkside report considerable numbers of complaints. We will later analyze the content of complaints from each area to address key issues.

# C. Analysis & Experiments

****

**K-means**: We implemented a k-means model to address misclassification errors in the current system, thereby reducing the need for manual reassignment and enhancing overall efficiency. We chose K-means as the start point due to its straightforward implementation, faster computation speed, and better suitability for our large dataset. We tested K-means with 2 to 10 clusters, examining both silhouette scores and a distortion elbow plot. The silhouette analysis shows that two clusters (k=2) perform best. In contrast, the elbow plot suggests k=4, with a distinct bend in the curve. We plan to use k=4 for further analysis because the richer cluster breakdown will help us explore the many variables present in each feature.

**Hierarchical Clustering**: We then applied hierarchical clustering to compare its performance with k-means. However, due to the large size of our dataset, directly running the hierarchical clustering kernel caused it to crash, so we had to sample the data before applying the method. We experimented with different sample sizes and found that a sampling rate of 0.2 was the critical threshold at which the kernel would crash. The attached figure illustrates these results. Additionally, we evaluated the impact of various linkage methods on the model. The results showed that the silhouette scores for the single, complete, and average linkage methods were all identical at 0.4136, while the Ward method produced a lower score of 0.215. Although the silhouette score obtained from hierarchical clustering was higher than that observed with basic k-means, it is not directly comparable since it was derived from only a subset of the data.

**Text mining:** Since we have a complaint description field, we first need to apply text mining techniques to convert the raw strings into a format the model can recognize, which will be used for subsequent topic modeling. Initially, we preprocess the text by converting it to lowercase, removing punctuation, and eliminating stop words. We also apply lemmatization to restore words to their base forms—instead of stemming—since stemming does not consider semantics or part-of-speech and offers lower accuracy. For vectorization, we experimented with both Bag-of-Words (BOW) and TF-IDF, ultimately choosing TF-IDF because it weights words to diminish the impact of standard terms while highlighting keywords.

**Topic modeling:** After conducting text mining, we decided to perform topic modeling because, after vectorization, our data had around 20,000 dimensions. We might lose too much information if we applied dimensionality reduction at that stage. Instead, topic modeling provides a more effective way to reduce the feature space to 10 topics. Due to computational cost considerations, we chose 10 topics because the original system only allowed residents to select from 5 issue categories*(Appendix Figure 5)* ; as a result, we experimented with twice as many topics to see whether this would better categorize the issues. We then utilized ChatGPT to identify the top 10 significant words for each topic and used our subject knowledge to name them .*(Appendix Figure 6)* Finally, we reassigned these labels to the original dataset for subsequent association rule analysis and clustering.

**Back to K-means:** To evaluate whether incorporating features from the complaint descriptions improved our clustering, we reran K-means. After performing topic modeling, we integrated the resulting topic labels with the original variables and executed K-means again. The results showed a slight improvement in the average silhouette score, increasing from 0.169 to 0.178. However, the improvement was minimal, indicating that while adding description-derived features was somewhat beneficial, its overall impact was limited.

**Association rules**: We focus on Division, Neighborhood, and Topic. From the division aspect, by comparing the receiving and assigned division graphs, we identify unusual mismatch rates to pinpoint divisions that often receive cases outside their responsibility. For neighborhoods, after placing the top 10 neighborhoods by complaint volume, we dive deeper into those areas that are more prone to specific complaint types, aiming to reduce manual reassignments and accelerate resolution. Finally, at the topic level, we examine which labels coincide with mismatches, signaling potential misclassification. We refine our categorization process by highlighting topics with higher mismatch odds, minimizing errors, and improving complaint-handling efficiency. More details about this analysis can be found in the Findings section and[*Appendix 5 – Appendix 7.*](#_heading=h.17dp8vu)

# 

# D. Challenges, Dead Ends & Adjustments

**Challenges in Dataset Processing**

The raw data from the San Francisco Government's Open Data is quite messy and requires thorough preprocessing to ensure accuracy. This includes filling in missing values, standardizing data types, encoding categories, handling outliers, and scaling numerical features. Removing duplicates also helps maintain data integrity and improve model performance.

The dataset's size makes feature selection, model training, and tuning computationally expensive. Sampling risks losing key groups, so we use label encoding to reduce dimensionality, which requires extra preprocessing. To optimize resources, we refine hyperparameter tuning by halving grid and random search while prioritizing key features via forward/backward selection and SelectKBest. Cross-validation ensures generalization, balancing efficiency, and performance.

**Dead Ends in Data Analysis & Adjustments**

When we ran the hierarchical clustering, it crashed on large datasets due to its rapidly increasing computational and memory demands. Therefore, we applied sampling (0.4) for processing, but it failed to represent the entire dataset fully. As a result, we replaced it with K-Means.

# E. Findings and Interpretations

## Division Mismatch Rates and Clustering Patterns in Complaints

As our EDA shows, the Building Inspection Division and Plumbing Inspection Division receive a high volume of complaints, ranking second and third overall. Further analysis [*(Appendix 9)*](#_heading=h.17dp8vu) reveals that when the Plumbing Inspection Division receives a complaint, around 14% are ultimately assigned to the Code Enforcement Section. Similarly, when the Building Inspection Division receives a complaint, about 10% end up in the Code Enforcement Section. This suggests a significant number of Code Enforcement related complaints are initially routed to Plumbing and Building Inspection by mistake. In addition, three divisions, Department of Bldg Inspection, Help Desk/Technical Services, and Central Permit Bureau, exhibit a 100% mismatch rate, despite having relatively few complaints. Notably, when the Department of Bldg Inspection receives a complaint, 80% of these are rerouted to Housing Inspection Services.

By mapping the KMeans cluster labels back to the original dataset, we observe that cluster label 3 and Building Inspection Division are positively correlated, each showing a lift value of about 1.4 [*(Appendix 10)*](#_heading=h.17dp8vu) Additionally, cluster label 0 is positively correlated with both Code Enforcement Section and Building Inspection Division, with lift values around 1.39 in both directions. Finally, cluster label 2 shows a positive correlation with Housing Inspection Services. These findings suggest distinct patterns in how complaint cases are grouped by cluster labels and the divisions involved, pointing to specific divisions commonly receiving similar complaints. This underscores the need for improved complaint-routing procedures to ensure each complaint is directed to the correct division and handled efficiently.

### **Complaint Topics Patterns in Neighborhoods**

After applying association rules between the topic labels from our topic modeling and neighborhood data, the results revealed that Tenderloin and Nob Hill[*(Appendix 11)*](#_heading=h.17dp8vu), the neighborhoods with the second and third highest complaint volumes, exhibited the strongest association with issues related to illegal garage/deck construction, indicating that unauthorized construction is particularly severe in these areas. Additionally, the Mission District, which recorded the highest number of complaints, was primarily linked to topics such as construction noise & complaints, illegal garage/deck construction, and property exterior & neighborhood issues. We recommend enhancing monitoring and management efforts through measures like intensified fire safety inspections, stricter control over construction noise, and prompt intervention in cases of illegal construction.

Our analysis provided more profound insight to help government agencies and community organizations develop more targeted regulatory and improvement measures. By bolstering inspections, raising public awareness, and prioritizing the resolution of common or high-risk violations, unauthorized construction and fire safety hazards can be effectively reduced, ultimately improving the overall quality of life.

### **Specific Complaint Types Are Prone to Mismatch**

Our analysis [(Appendix 12)](#_heading=h.17dp8vu) highlights specific complaint types that are more prone to misclassification, leading to inefficiencies in case handling. Complaints related to "Illegal Basement & Water Damage" strongly correlate with a mismatch. Similarly, "Interior Plumbing & Mold" and "Commercial/Health Violations" also exhibit a notable lift above 1.4, suggesting a meaningful relationship between these complaints and mismatch occurrences.

These findings suggest that specific complaint categories may require refined classification standards, improved training for case handlers, or enhanced data validation processes to reduce errors and improve case resolution efficiency.

**F. Appendix**

|  |  |
| --- | --- |
|  |  |
| Figure 1 | Figure 2 |
|  |  |
| Figure 3 | Figure 4 |
| 一張含有 文字, 圖表, 螢幕擷取畫面, Rectangle 的圖片  AI 產生的內容可能不正確。 | 一張含有 文字, 字型, 螢幕擷取畫面, 數字 的圖片  AI 產生的內容可能不正確。 |
| Figure 5 | Figure 6 |
|  |  |
| Figure 7 | Figure 8 |
|  | |
| Figure 9 | |
|  | |
| Figure 10 | |
|  | |
| Figure 11 | |

|  |
| --- |
|  |
| Figure 12 |

**2. Contribution**

|  |  |  |  |
| --- | --- | --- | --- |
| Team Member | Milestone 1 | Milestone 2 | Overall  Contribution |
| Yu-Hsiang  (Rick) Wang | Data Preprocessing  Model Construction  Deliverable Drafting | Text Mining  K-Means Clustering  Storyline Integration  Deliverable Drafting | 25% |
| Ching-Hsuan  (Shawn) Lin | EDA  Preliminary Analysis  Deliverable Drafting | Text Mining  Topic Modeling  Storyline Integration  Deliverable Drafting | 25% |
| Kuang-Ching  (Amanda) Ting | EDA  Model Construction  Deliverable Drafting | Association Rule  Hierarchical Clustering  Storyline Integration  Deliverable Drafting | 25% |
| Ming-Hua  (Jasmine) Tsai | Data Preprocessing  Preliminary Analysis  Deliverable Drafting | Topic Modeling  Association Rule  Storyline Integration  Deliverable Drafting | 25% |

**3. GitHub Project:**

<https://github.com/shanelin0107/BA820_Group_Project_Team7.git>

**4. Reference**

1. ChatGPT: <https://chatgpt.com/share/67b6a8ac-4e4c-8009-8670-3f927ce830c9>
2. City and County of San Francisco. *Annual Budget Reports, Department of Building Inspection (DBI), 2018–2022.* <https://sf.gov>.
3. San Francisco Open Data. *Building Complaints and Violations Dataset, 2017–2022*.<https://data.sfgov.org>.
4. San Francisco Controller’s Office. *Audit Report on Building Complaint Classification Efficiency, 2021*.<https://sfcontroller.org>.

**5. Timeline**

**2/20~ 2/26**

* Implement Hierarchical Clustering and compare its grouping with the K-Means results to assess interpretability and actionability.
* **Text Preprocessing**: Reintroduce complaint descriptions. Perform text cleaning: remove stop words, punctuation, and special characters; convert text to lowercase; apply lemmatization.
* **Text Vectorization**: Convert the cleaned text data into numerical representations (TF-IDF, Bag-of-Words).

**2/27~3/5**

* **Topic Modeling**: Apply LDA (or an alternative method) to extract latent topics from the text.
* **Dimensionality Reduction & Visualization**: Use PCA or t-SNE to visualize the clustering results from both structured data and text-derived features.
* **Integration & Analysis**: Integrate insights from both clustering and topic modeling, and assign severity levels to the clusters.
* **Review & Reporting**: Refine the models and visualizations. Prepare and finalize the milestone report.