Deliverable 2 | Boston Permitting Project | Team F
Aidan Ruvins, Timothy Evdokimov, Akhil Kokkula, Andre Lesnick, Zachary Gou
November 2023

Introduction:
We are studying the City of Boston Permitting Project, the goal of which is to analyze trends in the approval and denial of the City of Boston building permits. Thus far we answered all the base questions of the project, done some preliminary modeling of the permit approval process, and are now proposing to further study the economic factors and influences on the permit approval process.

Our general findings were that the vast majority of applications are there for electrical or fire alarm-related works and that most approved permits are in heavily commercialized areas of Boston. Furthermore, we found that there is a negative correlation between area income and denial rate of permits, but is considerably weaker than we hypothesized and requires further study.

Data Collection:
Our primary data sources were provided by the City of Boston and the United States Census Bureau and is comprised of the following information

- Zoning Board Appeals Data
- Article 80 Development Projects Data
- Approved Building Permit Data
- Income and other demographic data pulled from the MA voter roll and census data
- GDP data pulled from the federal government

Our data cleaning steps mainly included removing incomplete data points by filtering out various “N/A”, “NaN”, etc values in the data, and filtering out obvious outliers from the datasets to prevent them from skewing means and other measurements of central tendency.

In addition to filtering out unhelpful data, “reconciling” data from the different datasets to make sure that information from one dataset, e.g. income data from the MA voter registry, was correctly match to any given datum about a permit application proved somewhat challenging, the main issue was with data being different resolutions: i.e. some data was accurate to the address, while other data was only accurate to the nearest township or zip code.

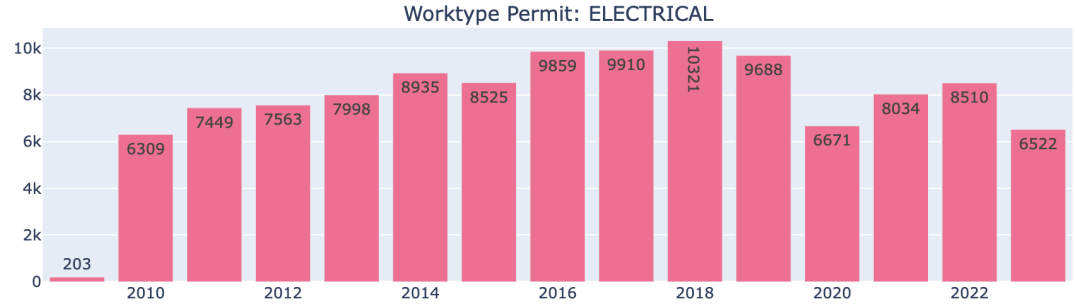
Summary of Early Findings:
The overall conclusion of our findings is that geographic and economic data is currently the best predictor we have of whether a permit will be approved or denied. Certain zip codes have vastly different approval rates than others, which generally correlates with income. We also noted that most permits are for mundane work in commercial settings. Permit applications, not even approvals, for large scale projects and residential work are relatively uncommon.

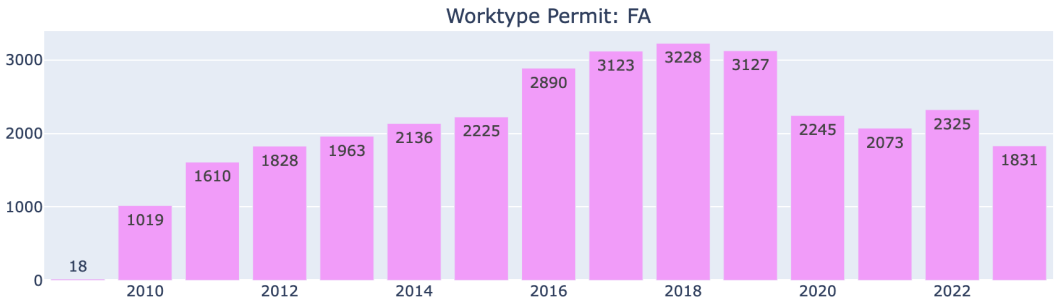
This can be most clearly seen in by the effect of location on approval rate (see question 5 in deliverable 1), by the sharp changes in approval rate over time, and by the vastly different approval statistics for different categories of permit.

Exploratory Data Analysis:

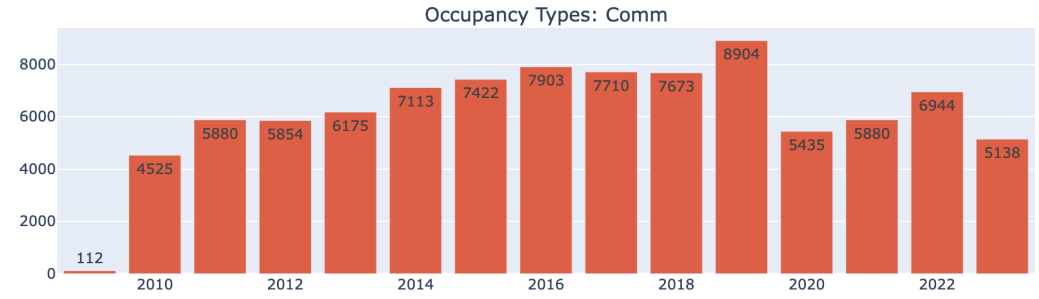
Note: Questions 1, 3 and 5 were already answered in Deliverable 1 for Boston Permitting Team F.

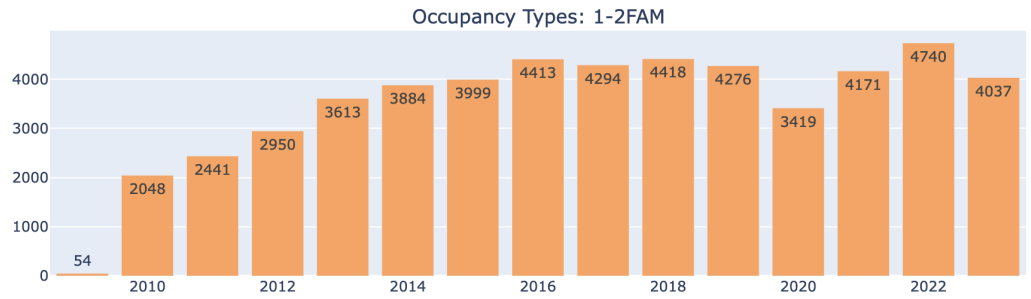
Question 2: How have work type approvals changed over the past 5 years i.e. a year-over-year analysis?



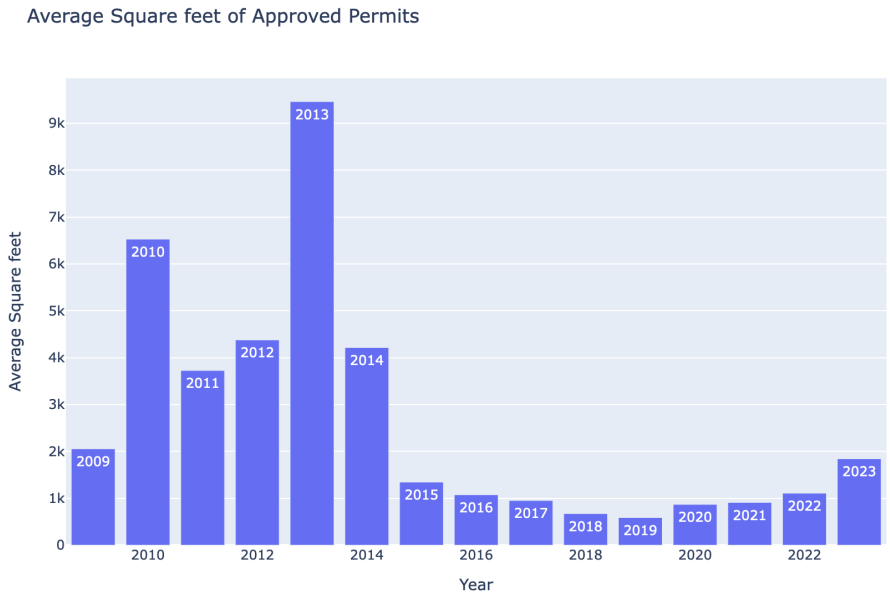


Electrical approvals tended to grow at a steady rate. There is a notable drop in 2020 due to covid. Fire alarm approvals also tended to grow at a steady rate, with a similar drop corresponding to the covid pandemic.

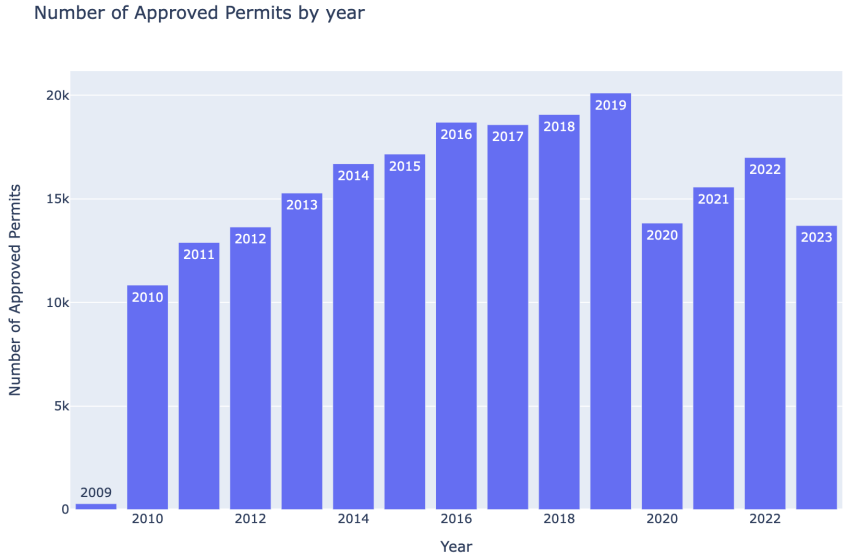




Both commercial property approvals and 1-2 family homes also tended to grow at a steady rate, with a similarly telling drop due to covid. Note that 1-2 family applications have recovered to pre-covid levels, while commercial applications have not.



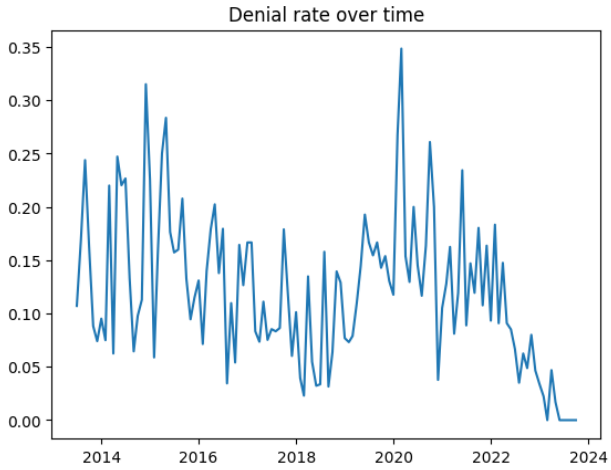
The average square footage per year shows a clear spike in 2013, and a sharp decline starting in 2015 that continued until 2019, when the average square footage began increasing again. We have yet to find a conclusive explanation for the sudden drop off after 2013-14, and this phenomenon requires further study.



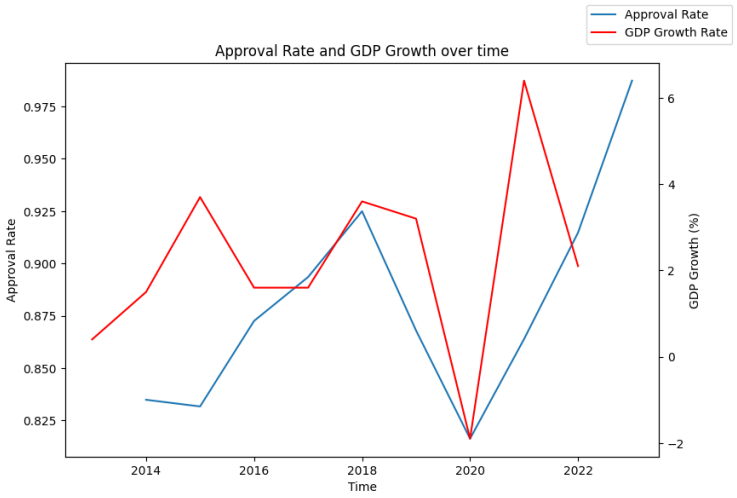
The number of Approved Permits by year steadily increased every year. There is a notable drop in 2020 due to covid, and still has not completely recovered to 2019 levels. This may also be correlated to economic conditions every year.

Question: What are the year over year trends visible in the zoning board of appeal approvals and denials by geography (neighborhood - listed as city, zip code, zoning district)? You'll want to normalize the data, perhaps ratio of permits to approvals or denials, etc.

If we look at denial rate over time as broken down in one-month intervals, we get a very noisy trend, suggesting that in many ways approval/denial trends are down to random chance, however certain trends do nonetheless emerge. For instance, looking at the denial rate, we can see a giant spike to 35% denial rate in early 2020, which is a very significant result. The cause of this is very likely to be the covid pandemic, a cause of many statistical anomalies and likely to be the source asterisks in data tables for years to come.



We further hypothesized that the approval rate of permits is tied to the state economy, a hypothesis that we were able to start exploring by comparing *approval* rate to percent change in GDP. The results are preliminary, as there is more to the ever-monolithic metric of "*the* economy", but initial results show a strong correlation.



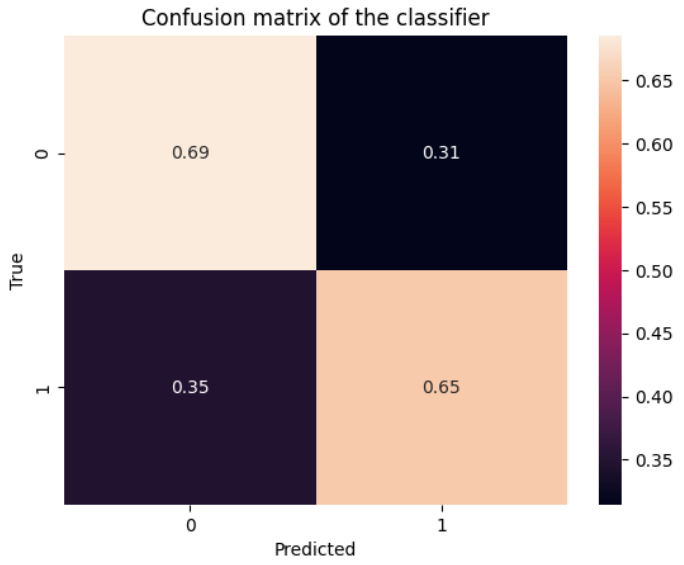
It is still unclear where the causation lies here though, i.e. if covid caused both the economic downturn and spike in denials, or perhaps vice versa. More work is needed to eliminate the possibility of confounding variables.

Modeling the Permit Approval Process

When trying to analyze what factors most affect the approval or denial of a permit, a natural approach is to try to model the permit approval process and see which features are most important in predicting the outcome of a permit approval process.

The first step was determining what features to base the model off of: apart from readily available and numerical data, such as ward and zip code numbers, we chose to include the number of deferrals a permit previously had, the income of the neighborhood the permit application is in, and the date the permit application took place (expressed as two separate features, year and month). Many features originally encoded textually (e.g. permit status) were turned into numerical classes. Also note that the classification problem itself was reduced to a permit being either denied (class 1) or not denied (class 0), as simplifying the problem to get the best model possible is more important than getting a little extra resolution by differentiating “denied” and “denied with prejudice”, which ultimately have the same result as far as the permit is concerned.

The initial naive attempt to use K-nearest-neighbors initially looked incredibly successful, with a suspiciously high accuracy of 83% on the first attempt. However, upon closer inspection, this model proved to be 96% accurate for approved permits but only 13% accurate for denied ones, meaning it was a very heavily skewed model and therefore not very useful in distinguishing approved permits from denied ones.



To address the data being heavily skewed in favor of approved data points, undersampling was used to forcefully even out the data by dropping a large fraction of the approved permit data at random in order to make a training set with approximately equal numbers of points for each class. This provided a much more even result, with a random forest model using N=1,000 decision tree estimators being able to distinguish between approved and unapproved permits approximately 67% of the time, but doing so evenly.

One thing to note is that the model did not suffer at all from overfitting, as the accuracy of the model did not ever decrease by limiting the maximum number of features applied to the random forest. Also, note that these are preliminary models, and further analysis more heavily based on economic data as model features is needed to get a more accurate model.

Additional Challenges

Cleaning the data was challenging since there were many columns in the large dataset and we needed to fix or remove incorrectly formatted, duplicate, and incomplete data. For instance, the Massachusetts voter data was very messy, with many missing pieces of information, and a sheer volume of data that made it difficult to work with. The solution to this was to iterate with a small subset of the data, clean

the data by simply dropping corrupted rows rather than trying to parse them, and then redo the analysis on as much data as possible once preliminary analysis shows promising results.

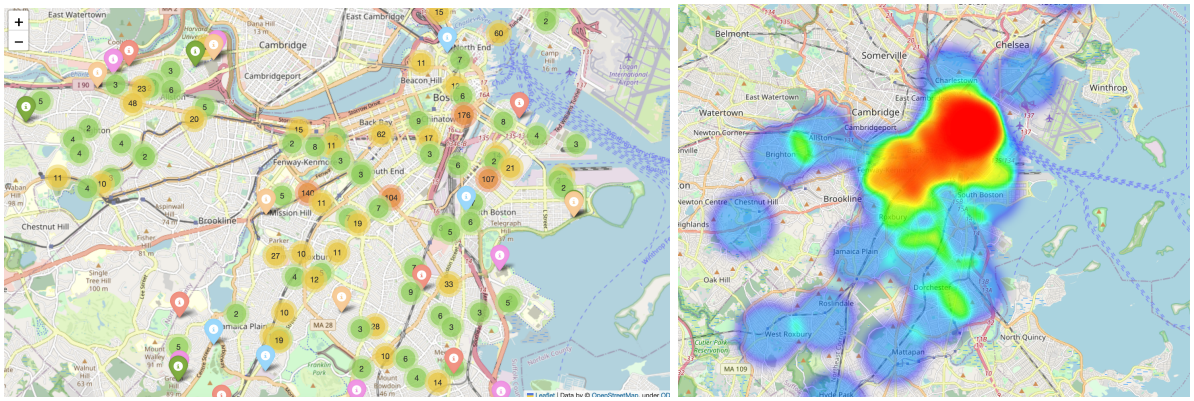
Limitations

Important variables were scattered around different datasets, so reconciling them was a bit challenging, and some data was missing or didn't go back far enough. For instance, the voter appeal data did not list geographic coordinates, merely zip codes and addresses, if that, making mapping the data extremely difficult as most geocoding services (e.g. the geopy library used access Google and other geocoding APIs, a lookup table of zip codes to latitude and longitude proved to be a suitable alternative, although did not provide the address-level resolution a more costly approach might've) are either rate-limited or cost money. Another example would be correlating approval rate over time to the economy — our data only went back to 2014 or so, which is almost 10 years, but at the same time, not that long on the timescales of economics.

Extension Proposal

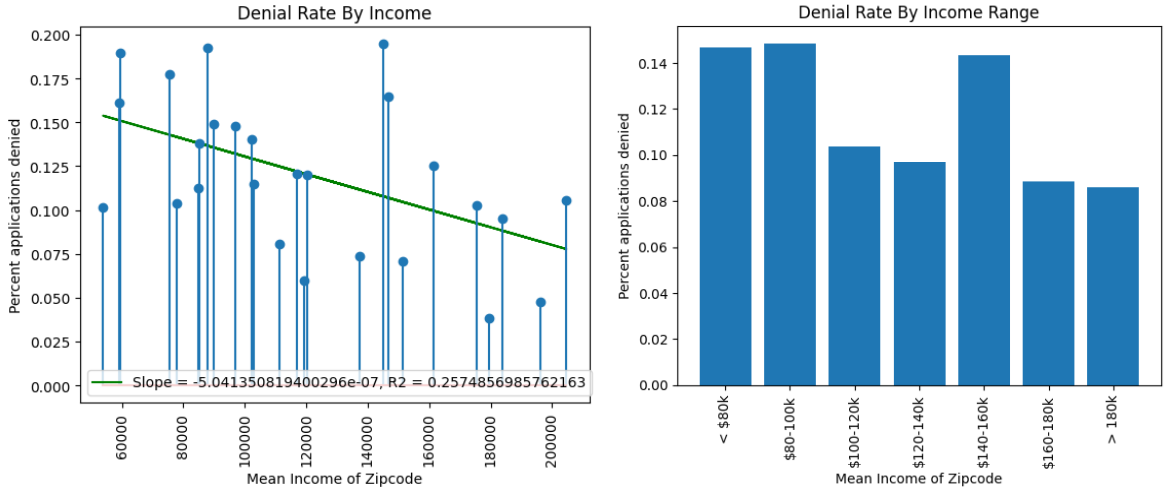
We are planning on analyzing why certain areas have a lower or higher permit approval rate, based on demographics and economic data (i.e. commercial vs residential zoning), specifically looking at heavily gentrified areas. This extension is important because it will show which areas tend to have more approvals. This is interesting because it can reveal which areas receive more approvals and why. We hypothesize that gentrified areas and wealthier areas will have more approval rates. We think this because more commercial properties will be located there and will find it necessary to go through the approval process. This problem came about because we noticed in Deliverable 1 that certain neighborhoods have more approvals than others. Wealth data across geographic locations will also be helpful for analysis.

The questions we mainly seek to answer with this extension project are to pinpoint which specific economic indicators (income? Zoning information? GDP? Tax data? Something else?) have the best predictive value for permit approval, on a neighborhood-by-neighborhood basis.

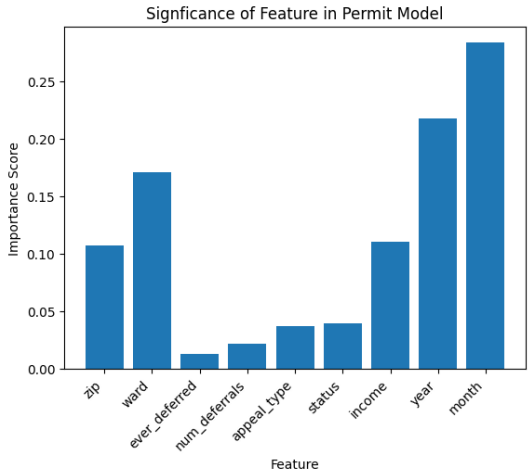


Consider these maps, with the one on the left mapping out locations of approved permits, and the one on the right being a heatmap of income level in the Boston area, with the two highlighting largely similar areas. This is one of the main reasons we hypothesize that income, and other economic factors are a large factor in permitting outcome. For the extension project, we hope to further analyze this correlation. We also noted that the areas with large numbers of approvals, as seen above, have a reputation of being "gentrified" — one other goal of the extension project should be to develop a more quantitative metric of how "gentrified" an area is, and see if it has any effect on permit outcome

Note that we already have data indicating a correlation between income and approval rating from deliverable 1, question 5, and which supports our initial hypothesis. However, we should investigate the outlier of the spike in approval for incomes of 120k to 140k as part of the extension.



Furthermore, when looking at our random forest decision tree model, analysis of the features that most heavily impacted the model, we found that location, date, and income were overwhelmingly the most deciding factors in the model's classification.



This supports our extension project focusing on economic factors, as this result tells us the most important factor is location, which is strongly tied to economic factors and supports the approach of analyzing our data on a location-by-location basis. The fact that the time of submission has such a strong impact tells us that a very impactful feature is one that changed significantly over the last few years but was omitted from the features we included, which is likely an economic feature, although it's possible it's simply a direct impact of the COVID-19 pandemic.

The data concerning around the individual permit, such as whether or not it is an appeal, whether it has previously been denied before, and other data concerning the permit itself is notably much less significant than factors concerning where and when the permit was filed, further suggesting larger external factors being key in understanding the disparity in Boston city permits.

Next Steps

To continue with this project, our group will shift focus to working on the extension project. We will collect further data on economic factors concerning permits, with the city of boston (https://data.boston.gov/) being an incredibly useful source of economic data, as well as other data from the state and federal government on state-wide economic trends that may prove useful.

We will also continue to refine our model of boston city permitting, with a main goal being to create a model that can predict the outcome of a permit based on economic data more accurately than our current best model. We also hope to better visualize the economic data we analyze to provide a better intuitive understanding of how economic factors affect permits, with a focus on mapping and time-series data.

Note that a significant limitation is that our permitting only goes to 2014, so trends caused by major events prior to that, such as the effects of the Big Dig of 1982-2008. This will also make it harder to model the effects of larger economic events that happened prior to 2014, such as the great recession of 2008 and 2009.

Individual Contributions

Aidan Ruvins: assisted exploratory data analysis for question 2, worked on deliverable report/presentation

Timothy Evdokimov: Modeling of permitting data as KNN model and random decision tree, addressing skewed classification data, visualization of income data as heatmap, answering base question 4, and sourcing preliminary economic data on GDP growth trends.

Akhil Kokkula: Cleaned up data and helped with the visualizations for base question 2, found helpful resources for the extension project, and assisted with deliverable report/presentation.

Andre Lesnick: Cleaned data and assisted with deliverable report/presentation

Zachary Gou: Exploratory data analysis for question 2