Boston Permitting Project | Team F

Aidan Ruvins, Timothy Evdokimov, Akhil Kokkula, Andre Lesnick, Zachary Gou

December 2023

Emails: aruvins@bu.edu, timevdo@bu.edu, kokkulaa@bu.edu, andrecl@bu.edu, zgou@bu.edu

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. We have answered all the base questions of the project, done modeling of the permit approval process, and further studied the economic factors and influences on the permit approval process.

Our general findings were that the best predictors of a permit being approved or denied were all tied to either wealth or time of year. In short, income, building square footage, number of fireplaces, and tax paid are all good predictors of a permit being approved, all of which are essentially metrics of wealth. Furthermore, the year and month a permit was filed, as well as the hotel occupancy rate at the time the permit was filed, are also good predictors of a permit being approved. While the latter initially struck us as a spurious correlation, we then realized that it was just a proxy for the time of year.

More generally, we also found that the majority of permit applications are for minor projects, almost exclusively for fire alarm installations and electrical work. We also found that most approved permits are in heavily commercialized areas of Boston.

Data Collection:

Our primary data sources were provided by the City of Boston and the United States Census Bureau and are 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 extension project also sourced data from data.boston.gov, pulling data on housing statistics, zoning boundaries, and more general economic indicators.

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. Our data on economic indicators also had several metrics cut off around 2016 or so, presumably when the city stopped collecting that information, so our dataset had to be adjusted to remove that effect.

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 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, house square footage, and other indicators of wealth. We also noted that most permits are for mundane work in commercial

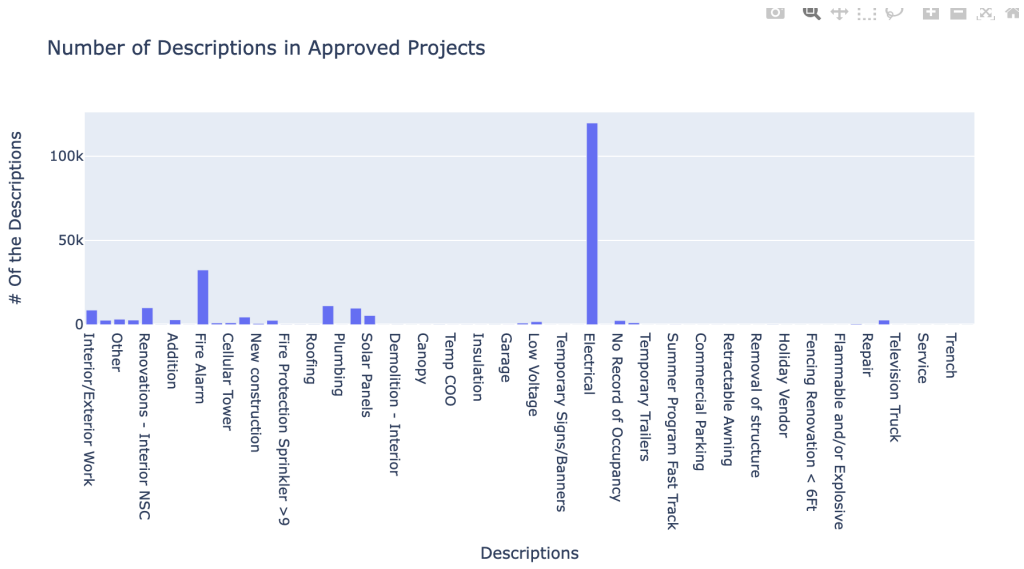
settings. Permit applications, not even approvals, for large-scale projects and residential work are relatively uncommon.

This can be most clearly seen 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 permits. Overall, the general results of the project are that permits are mostly for electrical and fire safety work on commercial properties, and those are most likely to be approved in wealthy areas.

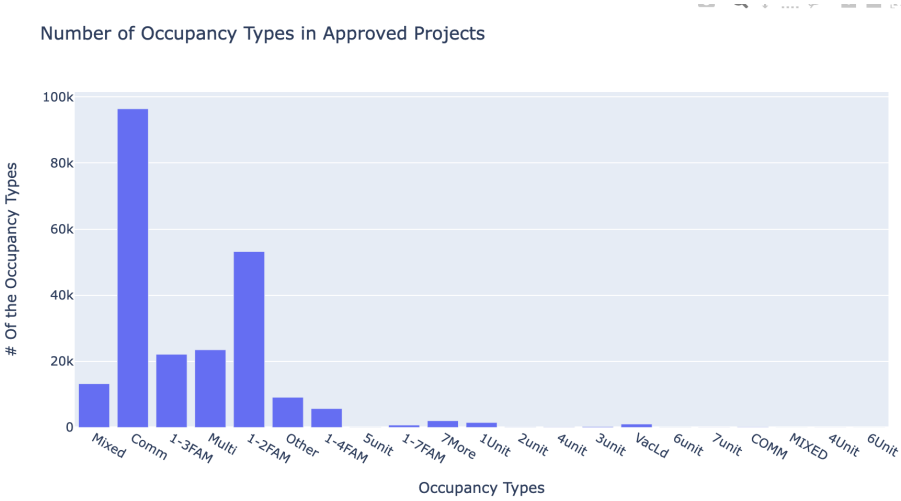
Exploratory Data Analysis:

Question: What type of building permits are approved each year by type (worktype), description, valuation (declared valuation), square footage, occupancy type?

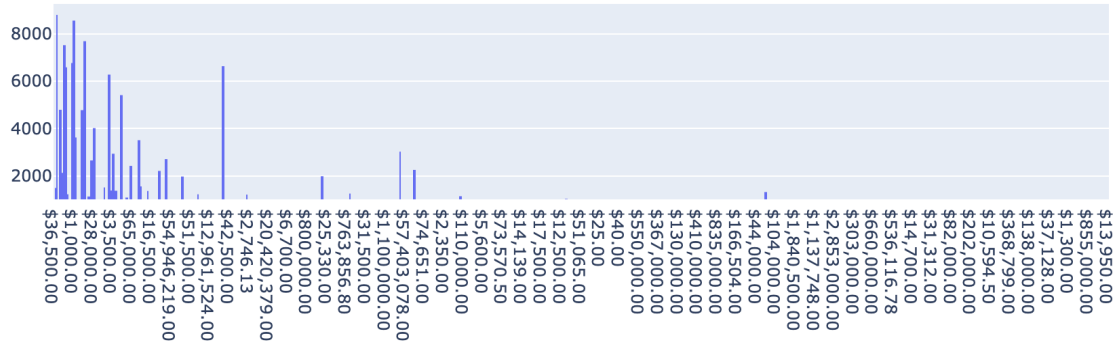
Most of the approved projects are to work on electrical issues. The second most common are fire alarm issues.



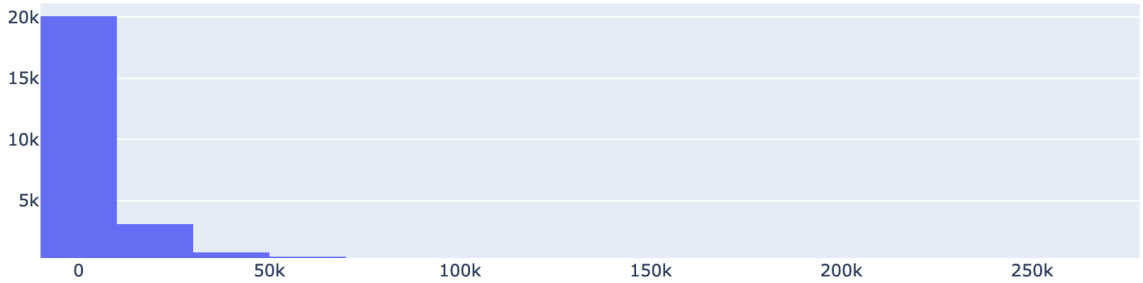
Most of the approved projects are to work on electrical issues. The second most common are fire alarm issues.



The most common project approvals are for commercial properties.

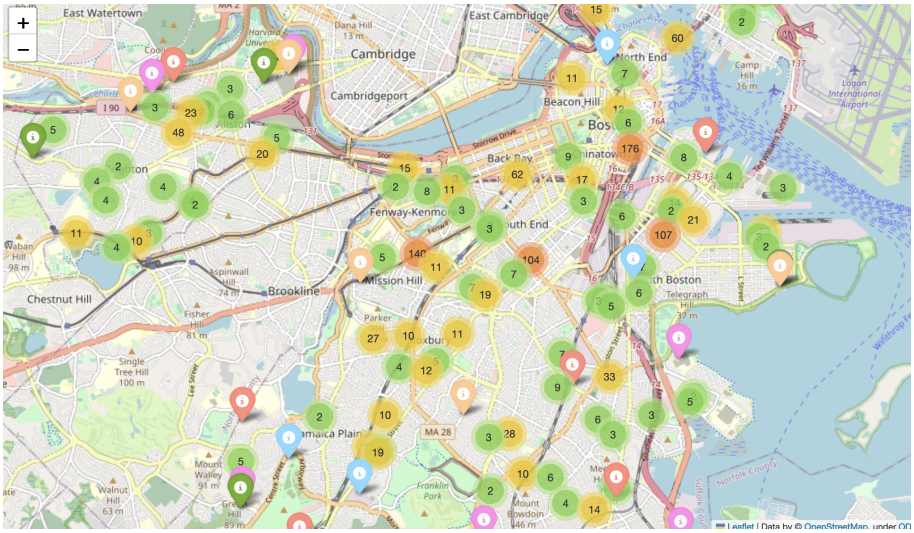


The declared valuation of approved projects tends to fall under $100,000 with a few outliers falling in the millions.



The first bar represents approved projects with square footage from the range of 0-10000 square feet. The second bar represents 10,000 to 30,000 square feet. The third bar represents 30,000 - 50,000 square feet. The fourth bar represents 50,000 to 70,000 square feet. The majority of the approved projects are below 10,000 square feet.

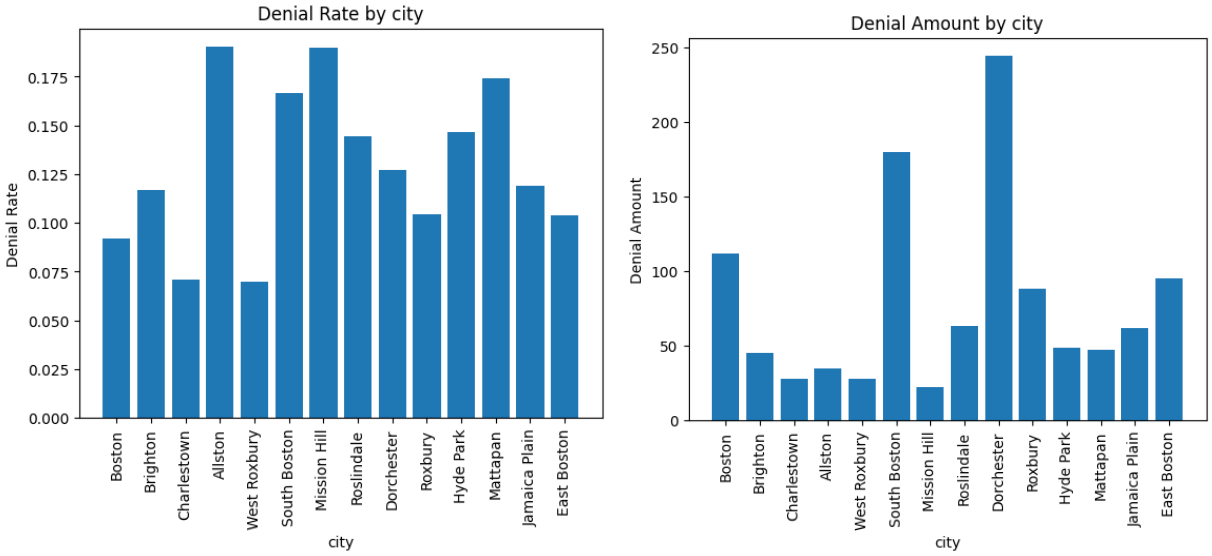
Question: Who is applying for building permits by geography (neighborhood, zip code, zoning district)?



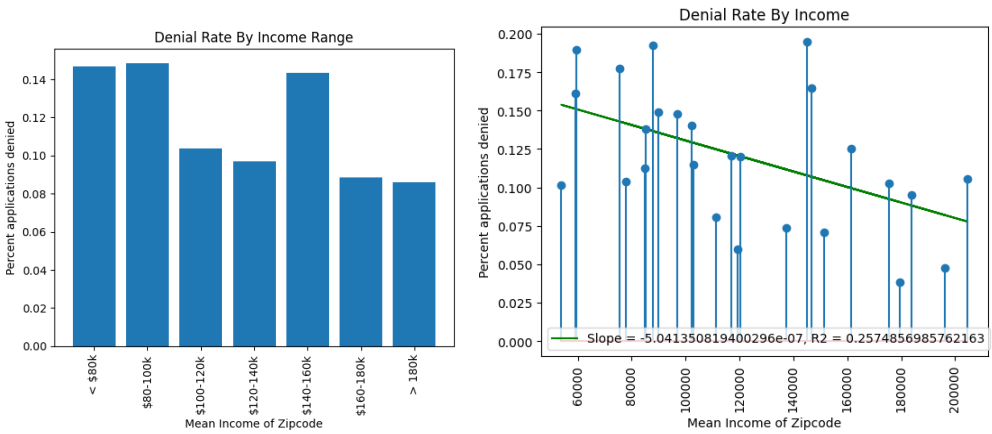
Within the Boston zip code, most approved projects are in the Fenway-Kenmore, South End Mission Hill, Financial District, and Seaport neighborhoods. This is mostly due to commercial properties being located in those areas.

Question: What are the geographic profiles of the census tracts of the addresses for the permits submitted and zoning board approvals and denials (use project address and match to census data)?

There is a clear difference in permit appeal denial rates from one locale to another, Dorchester clearly being the most disadvantaged by a number of denials, but Allston and Mission Hill being worse off in terms of denial *rate*.

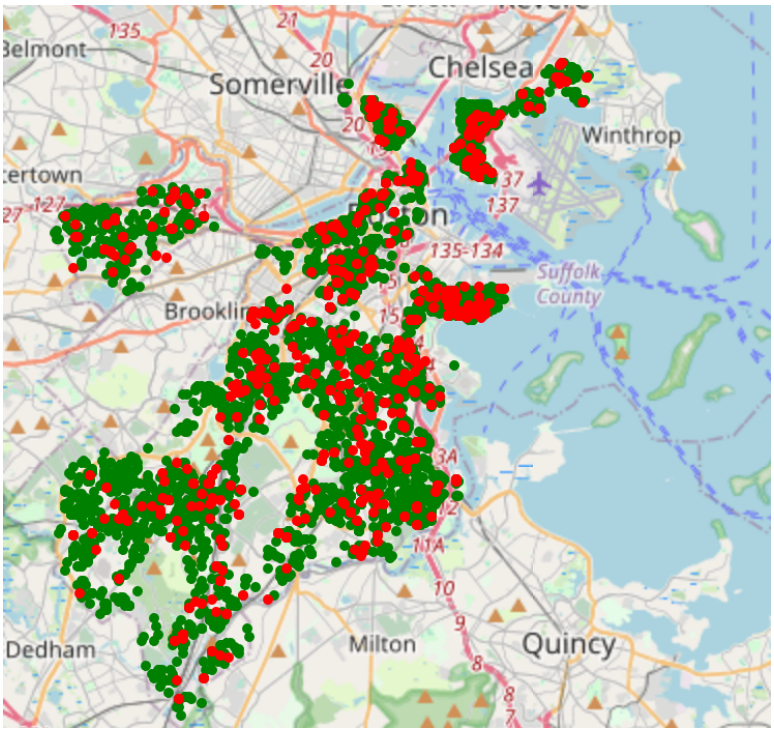


Analyzing this data in terms of per-zip code data found in the MA voter file, our group hypothesized that there would be a strong negative correlation between income and denial rate, i.e. the poorer an area the more likely it would be to be denied a permit. However, the raw data of income vs denial rate showed a fairly weak correlation of R^2 = 0.25, which is a correlation for sure, but a sketchy one. Filtering out the noise by measuring the denial rate vs income *range* proved considerably more useful, and showed a much more visualizable negative correlation between income and denial rate. Note the outlier of the range between 140-160k, which has yet to be adequately explained.

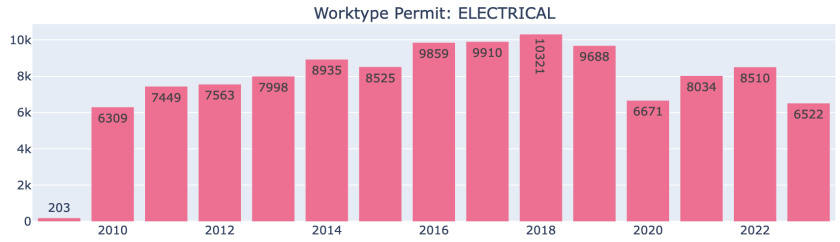


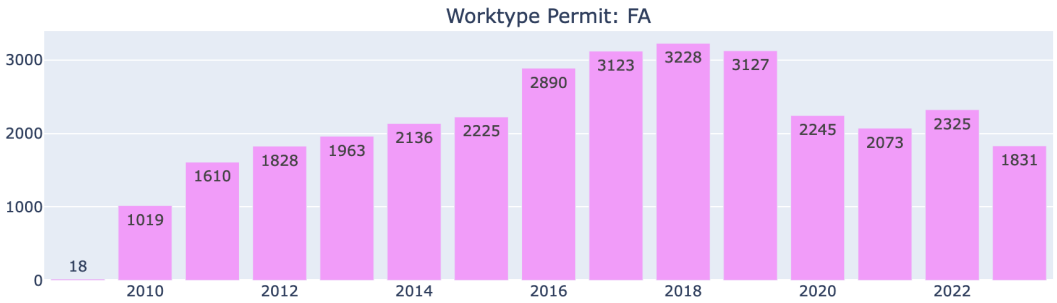
It is worth noting that while some locations have abnormally high denial rates compared to others, the average denial rate across all locations is 12%, and the number of denials, in general, is not very high, with no location exceeding 20% denial rate.

It is also helpful to visualize where the approvals and denials occur on a map, see the chart below. Green dots indicate approvals, while red dots indicate denials. Note the large concentration of denials in South Boston, East Boston, and Jamaica Plain.

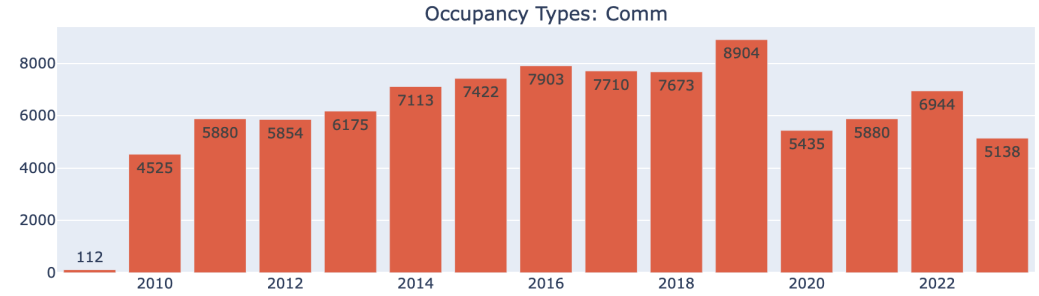


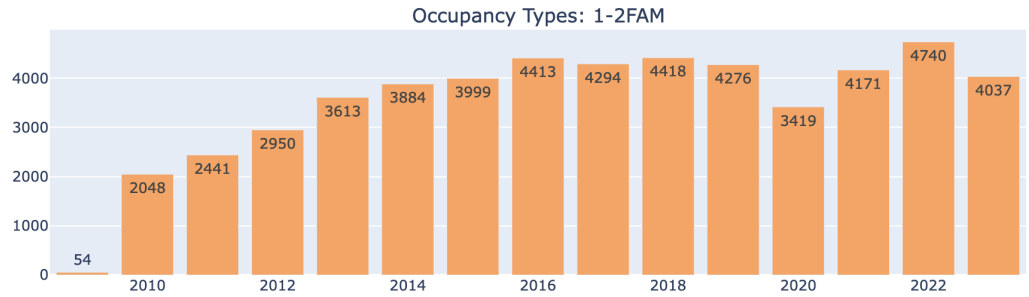
Question: How have work type approvals changed over the past 5 years i.e. a year-over-year analysis?



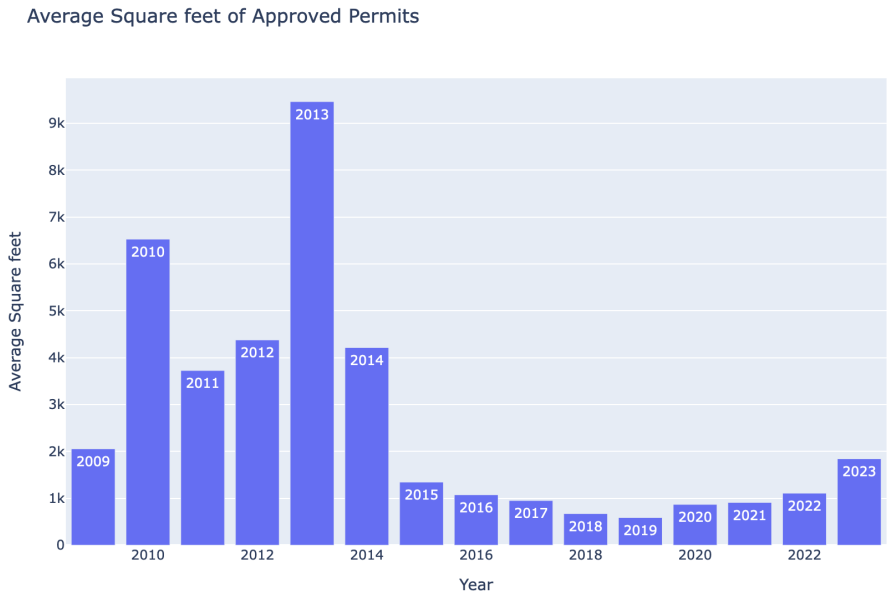


Electrical approvals tended to grow at a steady rate. There is a notable drop in 2020 due to covid. Fire alarm approvals also tended to grow at a steady rate, with a similar drop corresponding to the covid pandemic.



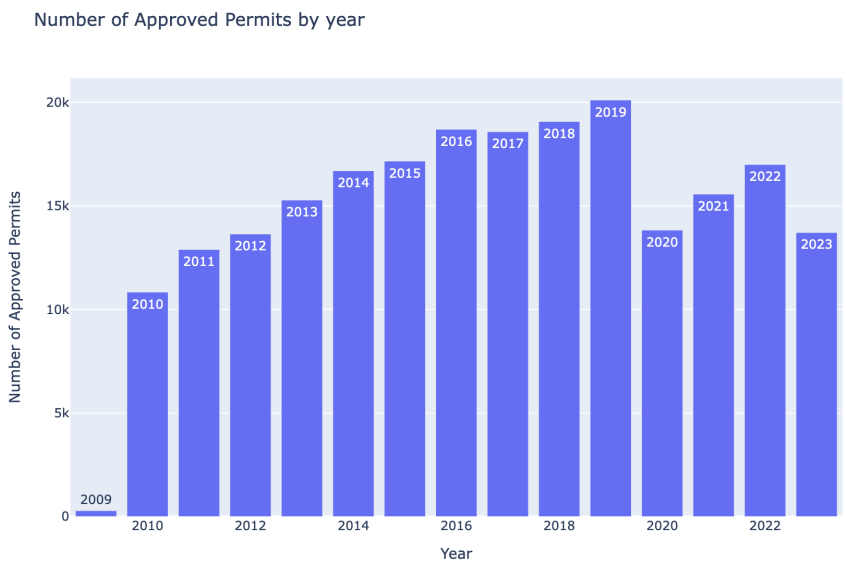


Both commercial property approvals and 1-2 family homes also tended to grow at a steady rate, with a similarly telling drop due to COVID-19. Note that 1-2 family applications have recovered to pre-COVID levels, while commercial applications have not.



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

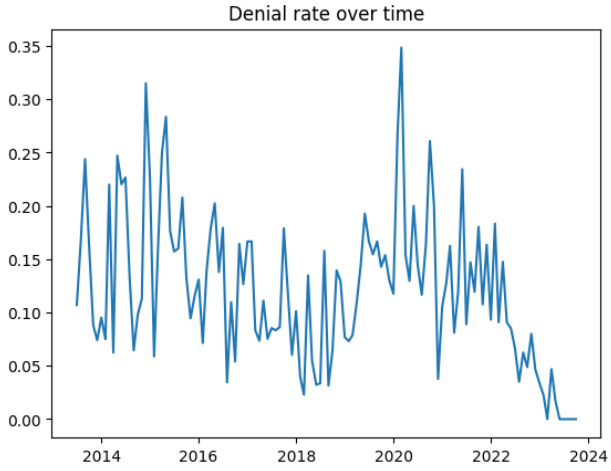
find a conclusive explanation for the sudden drop off after 2013-14, and this phenomenon requires further study.



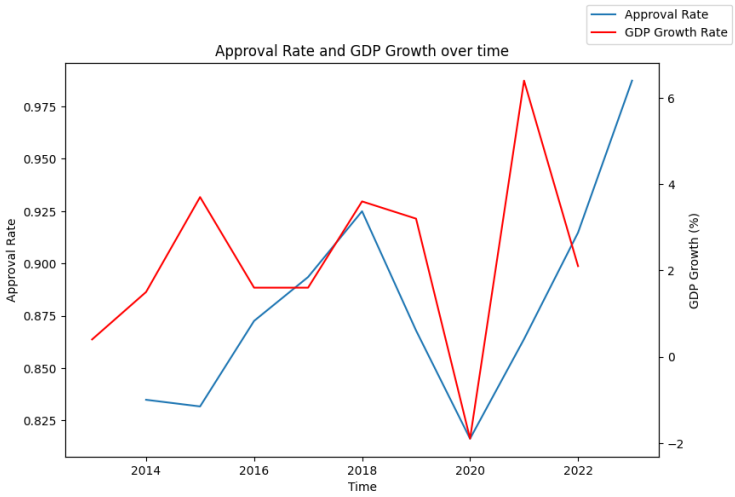
The number of Approved Permits by year steadily increased every year. There was a notable drop in 2020 due to COVID-19 and still has not completely recovered to 2019 levels. This may also be correlated to economic conditions every year.

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 the 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.



There is shown to be a strong correlation between the approval rates of permits and the state of the economy.

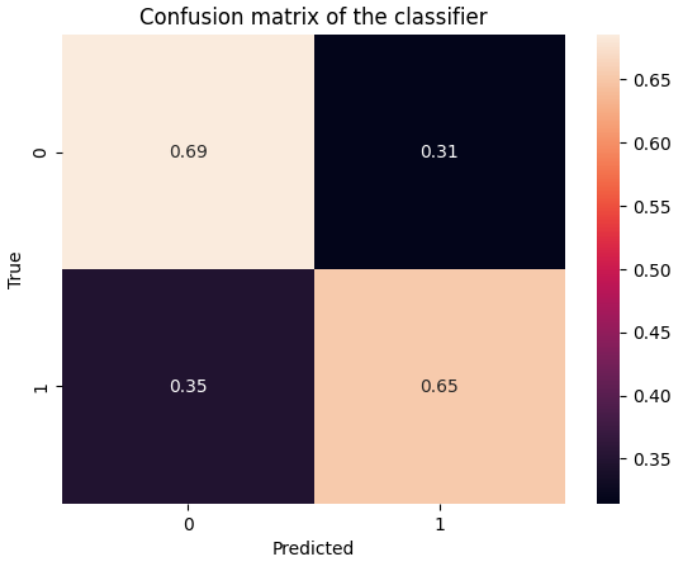


Modeling the Permit Approval Process

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. 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 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 however 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.

Challenges & Limitations

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, filtering data as it is loaded using lambda functions,and clean the data by simply dropping corrupted rows rather than trying to parse them. Then we redid the analysis on as much data as possible once the fast preliminary analysis on a subset of data shows promising results.

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.

Certain problems we were ultimately unable to overcome via limitations inherent in the dataset. For instance, the most salient data on approvals and denials was from the Boston permit appeals dataset, which, at the end of the day, just wasn't very large. After all our filtering, cleaning, merging with extension data, and address-geocoordinate reconciliation, we were left with only about N = 3,541 useful data points out of 9,043 data points in the original dataset, which wasn't all that large to begin with. We strongly suspect that if more appeal data were available, we would be able to construct a more accurate model. Furthermore, other limitations were irresolvable with our extension data, such as the fact that the city of Boston stopped collecting data on house sales, median house price, and foreclosure rate in 2016 — that information would have been really nice to have, since qualitatively it seems like it would have been relevant to understanding housing permits.`

Extension Project Pitch:

During our base project, we observed in our analysis of the base questions that certain areas were heavily skewed in their approval/denial in certain areas, especially those with a reputation for being gentrified. For our extension project, we wanted to look deeper into this trend and understand the economic forces that caused it.

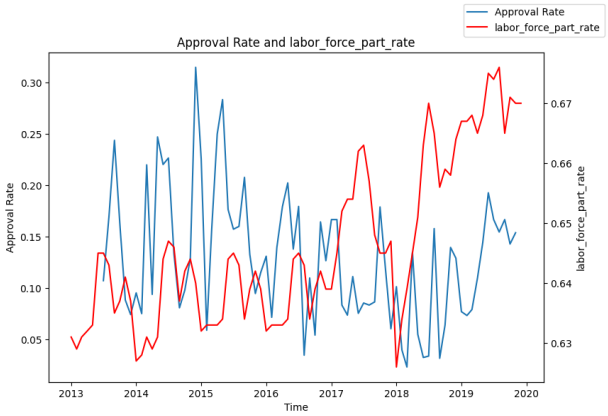
Specifically, we sought out data that would tell us about the economic nature, i.e. overall wealth, broken down by geography as narrowly as possible, to try and analyze permit data by rich/gentrified

areas vs poorer areas. We also wanted to find data on the overall economic state of the city to try and make better sense of our findings on permit approval over time. We then sought to improve our model of the permit process by using all this additional data as model features, refining the model to better predict and understand the permitting process.

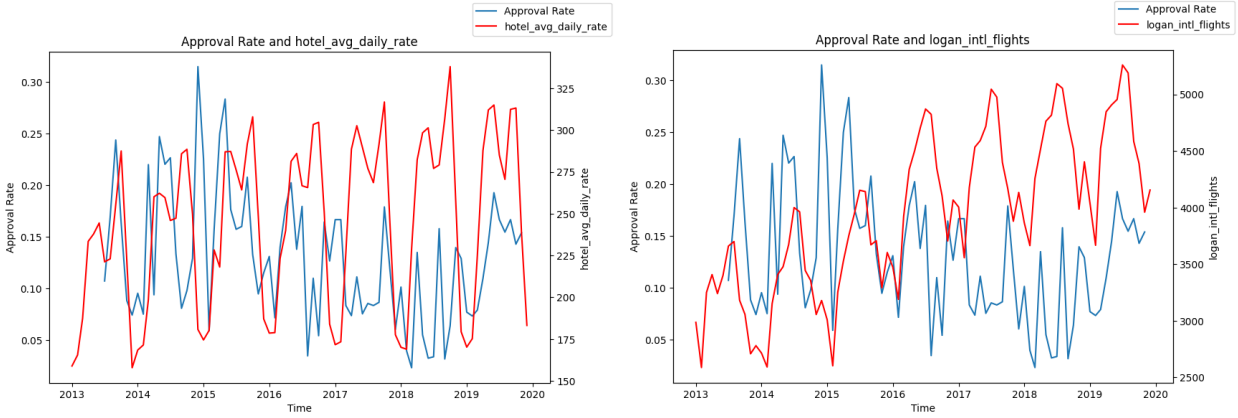
Extension Project Data:

To better understand the economic impact on housing permits, we sought to use several additional datasets, all sourced from https://data.boston.gov. These are the 2023 Boston Property Assessment Data, which is data on a representative sample of Boston properties, the MA Economic Indicators dataset, which tracks several economic indicators over the past decade or so, and the Zoning Subdistrict Data, which gives precise information on where commercial, residential, etc zoned areas are.

We had some notable findings from our exploratory analysis (EDA) of this data, some which seemed legitimate and was backed up by subsequent data modeling, and some are likely spurious correlations. For instance, we found that visually there was a reasonably okay correlation between labor force participation rate and housing permit approval, especially prior to 2016 and after 2019.



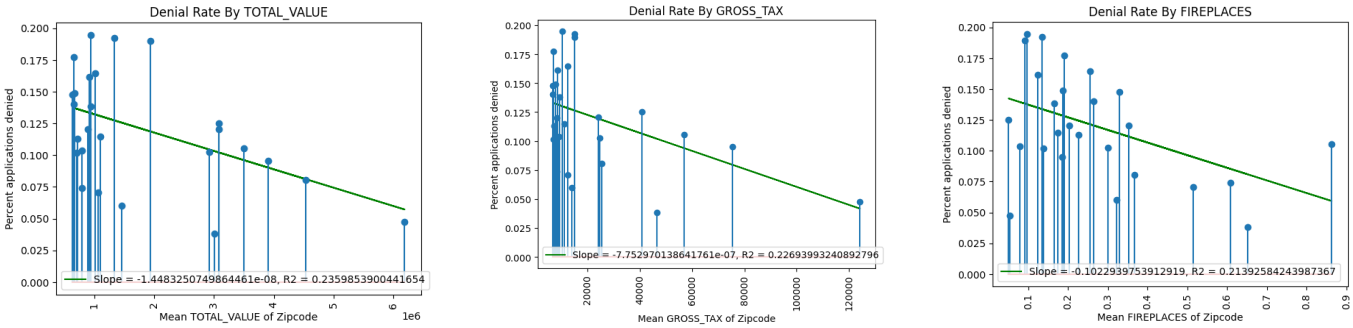
However, the same data also shows that there is a fairly strong correlation between international flights from Boston Logan Airport, as well as hotel occupancy rates.



It could be a totally spurious correlation, except that the hotel's daily occupancy rate finding is also backed up by some of our modeling. We suspect that it may be that both these metrics, which one

would expect to normally oscillate with the seasons and time of year as tourism normally increases and decreases. A possible explanation is that it could be that these metrics are proxies for the time of year.

We also found some interesting results from the housing and property data, by taking the mean property data for a given zip code and comparing it to the permit approval/denial rating of that zip code. We found that the square footage, total land value, number of fireplaces, and gross tax paid on property all correlated with a greater likelihood of approval.

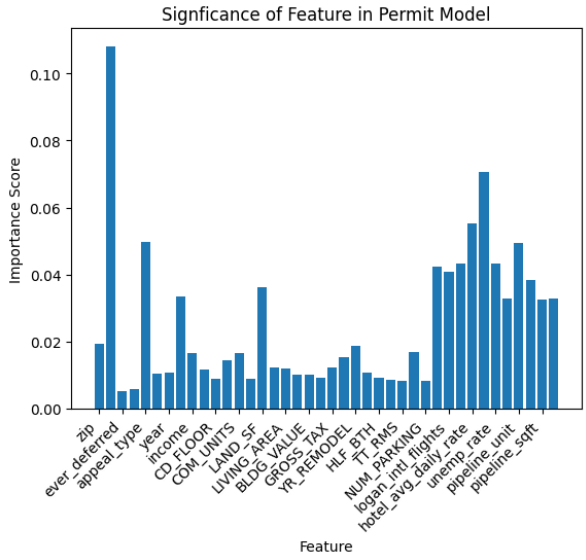


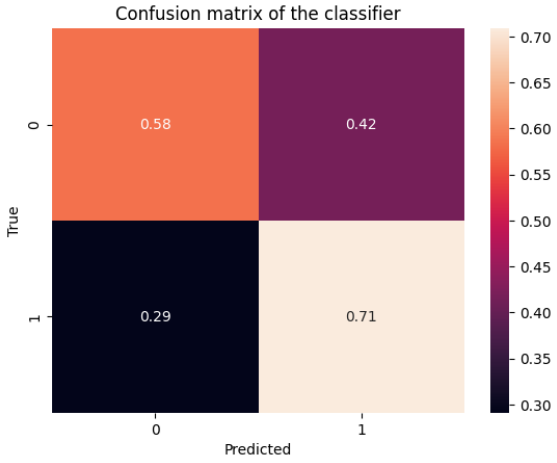
This is most likely due to all of these metrics being a proxy of wealth: wealthier people are likely to have bigger homes, with a greater number of fireplaces, which are worth more money, and in turn, have a larger tax paid on them. This is entirely consistent with our earlier finding that household income is a predictor of being more likely to have a permit approved.

Refining our Model Using Extension Data:

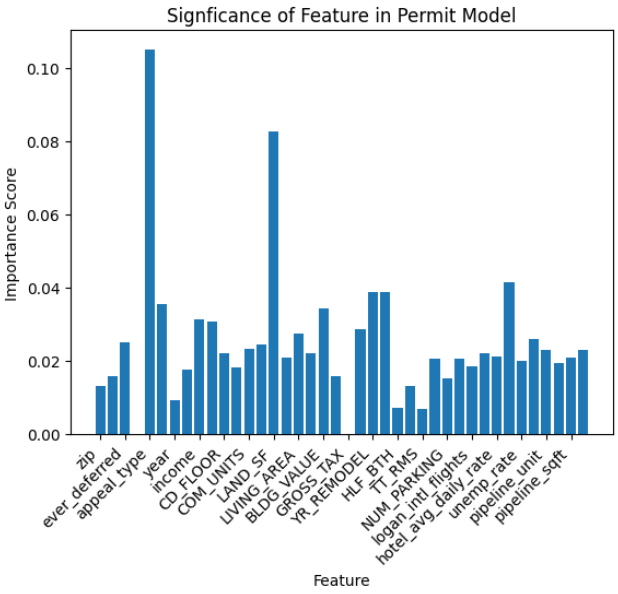
After having cleaned the extension project data, we sought to improve upon our model by incorporating features from our extension data. Some features had to be excluded, as certain economic indicators stopped being collected halfway through the time series, and certain metrics reported in the housing/property survey (such as roofing material) were reported as a string / in natural language and did not always lend itself well to being turned into a numerical classification. That being said, our first approach, which used the same decision tree approach, used undersampling to forcibly balance out the severe imbalance between data points for approved and unapproved class data points.

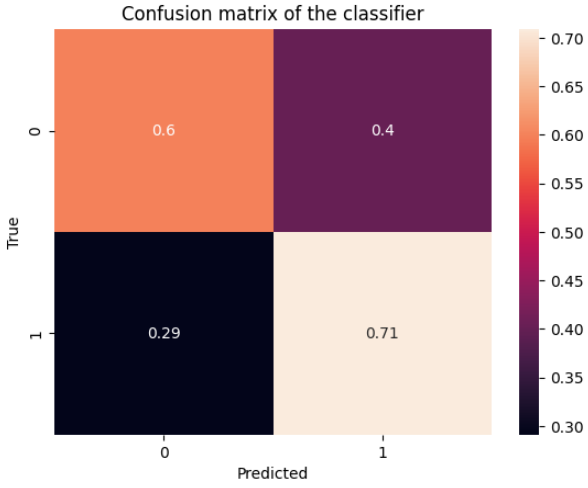
Using the decision tree model, the accuracy was largely unchanged from the original data, except that it became slightly less accurate at predicting approved permits and slightly more accurate at predicting denied ones. However, the most useful feature of this graph is the feature significance chart. It shows that the most predictive thing of a permit being denied is if it were ever deferred in the past, which intuitively makes sense. It also shows the type of appeal, income, house square footage, and the daily hotel occupancy rate have a great effect. The effect of income is consistent with earlier findings, as is house area which is a pretty obvious proxy for wealth/income.





We also tried a more novel modeling approach, using the XGBoost ensemble model. Note that this model is not a part of sklearn but rather can be found here. It is a boosting-type ensemble model developed in 2014, which has proven remarkably effective in many cases. It showed comparable performance to the random forest ensemble, with a marginal improvement for cases of approved permit appeals, correctly classifying 60% of approvals and 71. However, once again the feature significance is the most telling part, with appeal type being the greatest predictor in the XGBoost model, with square footage (still a proxy for wealth) and hotel occupancy again being two leading predictors of permits being approved or denied.





Conclusion

Overall, we found that wealth and time are the biggest predictors of permit approval or denial. This was hardly a surprising result, as we had always hypothesized that wealth is the biggest predictor of permit approval. If anything, the correlation between wealth and approval probability is weaker than we initially thought. However, this correlation, while predictable, is concerning, and could be a reasonably

good metric of broader income inequality, and a demonstration of the knock-on effects of income disparity.

Individual Contributions

Aidan Ruvins: Worked on the content for the README.file, helped with data collection, and created a report for base questions 1,2, and 4.

Timothy Evdokimov: Worked on finding the extension data, modeling the permitting process using the decision tree classifier, and analyzing the data on income versus permit approval in base question 5, and answered base question 3.

Zachary Gou: Worked on the extension project and created visualizations. Created visualizations for base questions 1,2, and 4.

Akhil Kokkula: Worked on the content of the README.md file and also worked on the beginning parts of the final report such as the Introduction, Data Collection, and 2 questions for the Exploratory Data Analysis section

Andre Lesnick: Worked on preliminary analysis and cleaning of the MA voter file data set, and merged the voter file and approved permits data sets to determine latitude and longitude data for each address for use in the extension project.