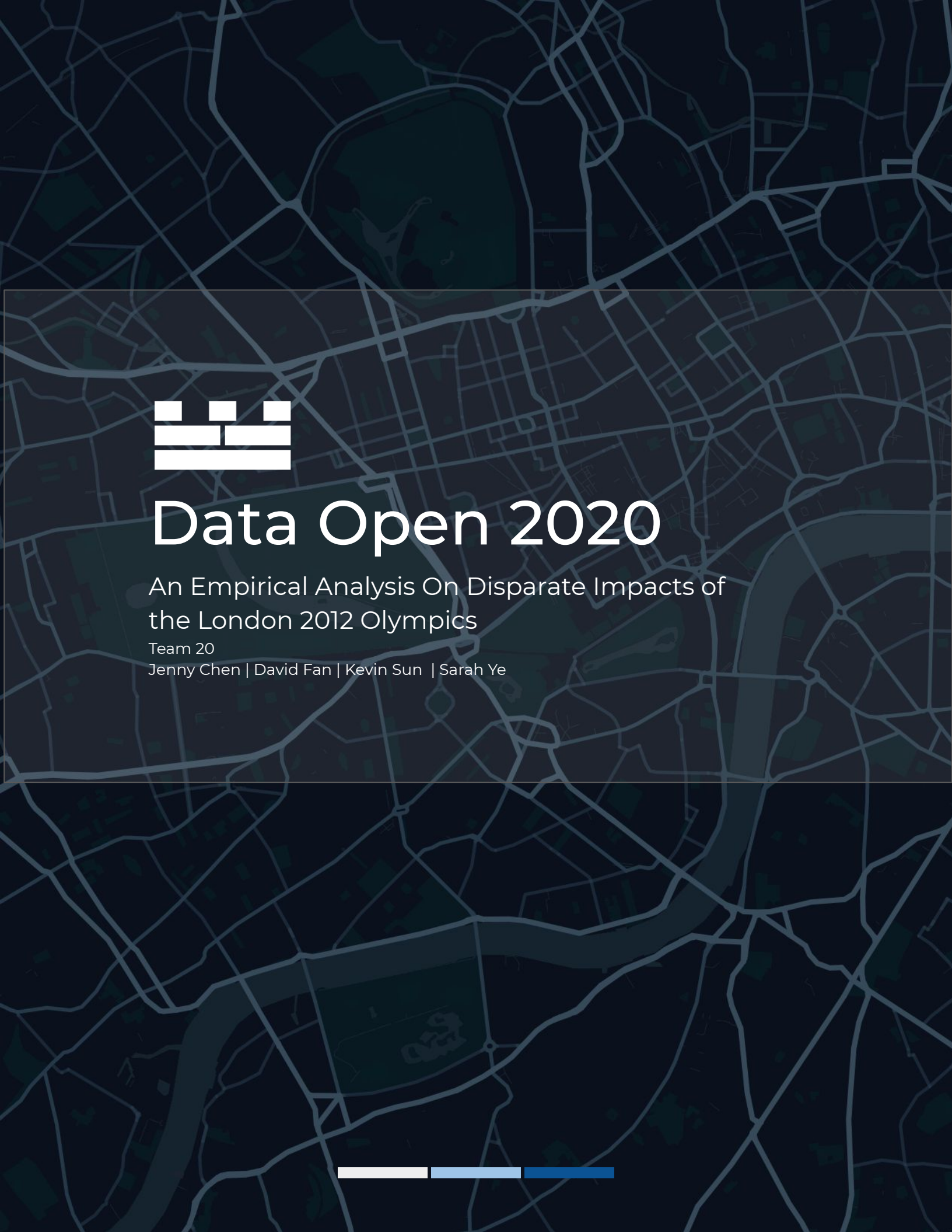

Data Open 2020

An Empirical Analysis On Disparate Impacts of the London 2012 Olympics

Team 20
Jenny Chen | David Fan | Kevin Sun | Sarah Ye

Executive Summary | Key Summary

The Data Open

General Background

The London Olympics was certainly a mega event, and while studies have been done extensively on the economic impact of the London Olympics overall, there currently lack literature specifically targeting how has the London Olympics affected the more **intangible aspects** of its various boroughs.

Knowing that the London Olympics targeted specific so-called 'growth boroughs', it is particularly important for us to **examine whether or not it has lived up to the promises of bridging the gap between the wealthier and less well-off neighborhoods**. Furthermore, we are interested in understanding whether or not there are unintended consequences of hosting the Olympics in terms of affecting various minority populations.

Key Findings

We employed causal methodology, geographical analysis, and ML attribution extensively for our analysis and found that:

Observation 1: London boroughs see **varying effects of hosting the Olympics**. As intended by the organizing committee - There appears to be a short term impact on the less economically well-off and more leisure-focused cities in the year 2012. However, the trends soon reverse to benefit their counterparts more in the longer term.

Observation 2: Upon a deep dive into the difference between immediate effect and overall aggregate effect, we found that the major host cities (particularly Newham), being the less well-off cities from the get-go, did, in fact, have a splendid initial gain during 2012, however the aggregate results clearly show a positive trend for the western London boroughs, that are known to be economically well-off.

Observation 3: We further found that while the economically well-off neighborhoods started their economic gains from 2012, on aggregate, there actually contains a spillover effect from these economically well-off neighborhoods to the neighborhoods near them.

Observation 4: There also seems to exist **unintended disparity between the development of boroughs where specific minorities reside**. In particular, boroughs with a high Asian and Black population saw a roughly negative impact from the Olympics while those with a high mixed-race population saw a sustainable long term gain.

Our Story

While there may be short term effects from the tourist crowd - the direct establishment of Olympic stadiums and other sports facilities and venues are unlikely to be standalone sustainable tourist attractions due to the plethora of stadiums in the UK and a bigger emphasis on league based sports such as soccer. Similar trends can also be seen in Rio, as the Olympics stadium now simply sits idly.

The more indirect benefit of the Olympics is likely present. These include **increased awareness and interest for the host city**, which potentially translates to an increasing flow of tourism and possible foreign investment but also general infrastructure revamp such as road construction and the like. However, we pose it that these benefits exist as a **multiplier effect** of the already well-off neighborhood since awareness at a city-level does not translate directly to host boroughs benefiting from the Olympics.

Tangentially, however, while the host boroughs did not benefit in the long term, as the better of boroughs grew, nearby boroughs experience an economic ripple effect in gaining positive impacts in the long run. This benefit however is more market dependent and hence speculative and something that the state can't directly control or alter.

Recommendation

While the Olympics have long been employed as a strategy to rejuvenate a specific neighborhood, we believe that **it alone is unlikely to result in the effect that governments generally expect**.

In line with that, we would recommend that **host cities should more actively try to ensure that the areas which they'd want to rejuvenate have the capabilities to sustain the growth** obtained from the Olympics prior to hosting such an event. These measures should include less direct investment in areas such as leisure but more on **general wellness infrastructures such as Education**, revitalizing the business scene, and ensuring neighborhood safety and quality of life.

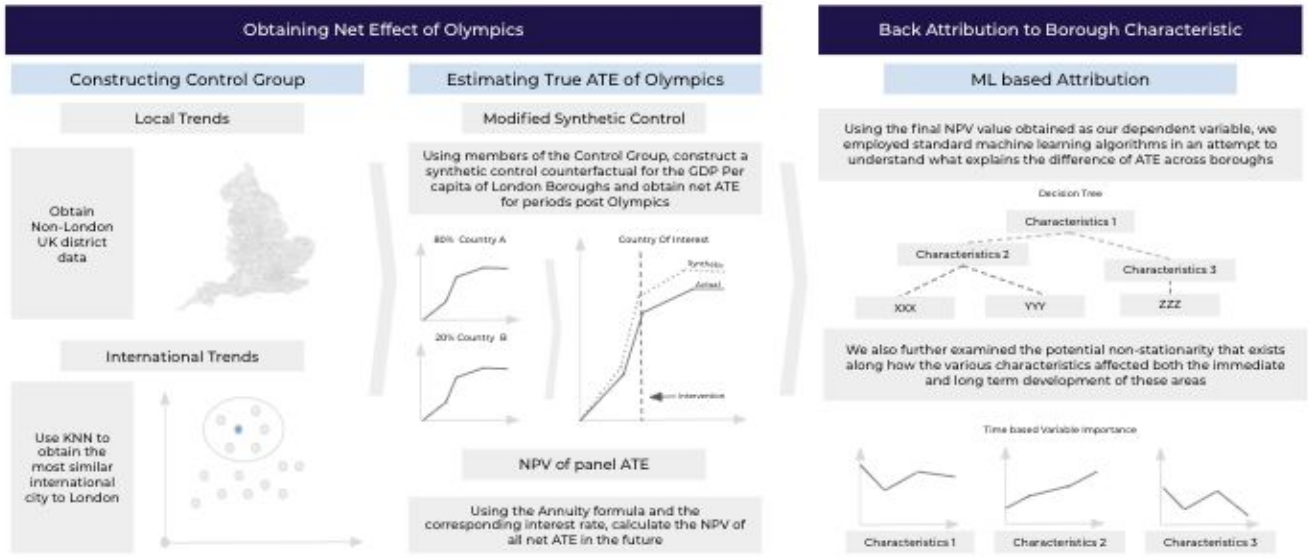

Figure 1: Analytical Approach Summary

Background

Mega sporting events like the Olympics are about competition and boosting national pride. But they are also big businesses and often serve as **key factors in both local and national development strategies**. Hosting cities invest massive sums to create the necessary infrastructure with the hope of reaping substantial gains for the local and national economy. The claimed benefits include increased tourist spending during the Games; the long-run "Olympic legacy" which might include improvements in infrastructure and increased trade, foreign investment, or tourism after the Games; and intangible benefits such as civic pride.

While governments justify their decision to host Olympics on the basis that GDP will increase as a result of hosting the event, previous studies have shown conflicting results. And we think it's worth researching into the real impact of the Olympics on the host city and country. Furthermore, we believe that there's a **disparate impact on different boroughs of London** and that prompted us to analyze the data on a granular level. Greater London is made up of 32 boroughs and the Olympics venues are located in the boroughs of Barking & Dagenham, Greenwich, Hackney, Newham, Tower Hamlets, and Waltham Forest. The host boroughs are referenced as 'growth boroughs' and they are specifically chosen because they were underdeveloped as compared to their neighbor boroughs and one of the Olympics' objectives was to **close the economic and social gap that existed between them and the average of London**.

Analytical Approach Summary

While past articles and studies have been made to investigate the effect of the Olympics on its host boroughs, several gaps remain within this space that we aim to investigate further.

First of all, we believe that London boroughs, as a whole, ought to be affected by the London Olympics. This is hypothesized to both exhibit a short term effect (i.e. a direct result of Olympic tourists traveling beyond the constraint of the Olympic village), but also long term effects driven by the increased awareness of the city around the world. In particular, we believe that the **long term effect across various boroughs** is especially worth examining further since the branding of the London Olympics brings together attention to the city of London as opposed to specific neighborhoods. We further posed that, while the London Olympics may have been intended to boost the status of its host boroughs, it may have not worked as intended, and may have simply driven more consumption and tourism for other more established areas in London.

Secondly, current literature focuses mainly on the observed change of certain metrics, such as median income, and aims to provide a causal relationship between the Olympics and these metrics directly. We believe that further work is needed in particular to estimate the **TRUE causal effects of Olympics** on the various cities, especially since key metrics like median income or GDP per capita are constantly affected by nationwide policies in the UK and also international trends.

Thirdly, we want to go further from simply pinpointing disparate treatment across boroughs, but create a meaningful analysis that **generalizes why certain boroughs may have benefited from the Olympics while the others have not**. In particular, we are curious to not only estimate the causal effects of Olympics in improving economic statuses of these boroughs but also to assess whether the benefits/harm has been disparate in areas that embody characteristics concerning minorities and protected classes.

Exploratory Data Analysis

We gathered a handful of external data and performed a number of preliminary graphical analysis.

Datasets

To perform our borough-level analysis, we needed time-series data on various socio-economic indicators for different boroughs in the United Kingdom. We started with the provided London taxpayer income data, which we hypothesized contained relevant data revealing the disproportionate impact of the 2012 London Olympics on various boroughs. However, we chose not to use many of the provided borough-level London datasets (e.g. earnings and underground activity) because we also needed that data for other non-London boroughs and international cities for comparison. Instead, we scoured the internet and the provided raw data sources for time-series (around the years 2000-2018), borough-level data on selected socioeconomic features—GDP per capita, unemployment, number of bars to reflect social activity, car traffic in kilometers, number of productivity jobs, number of productivity hours worked per week, distribution of enterprise category (e.g. Construction, Finance, Retail), and distribution of ethnic groups. The social indicators were primarily used for back attribution later on to explain economic differences between boroughs. Additionally, we used the UK inflation data using 2018 as the base year to quantify the financial impact in our modeling.

Other than traditional data cleaning methods such as removing or imputing missing values, we also paid special attention to the borough names, since we could not find a standard naming convention or a map from borough to area code. Specifically, if a dataset combined two boroughs together, we split them out into the two separate boroughs with values either copied (for features such as GDP per capita) or halved (for features such as the number of bars). It is worth noting that while such manipulation provides us with the complete set of data among all UK boroughs, it risks the danger of having one treatment group having the exact set of data as a control group, resulting in erroneous perfect model fittings. To counter such risks, all data duplicates of the region of interest are dropped while we fit our model on any specific borough. In the end, we were able to identify a list of London boroughs, UK boroughs outside of London, and other international cities to use for our control group.

Income and GDP per capita

First, we sought to visualize the London taxpayer income and GDP per capita data for various UK boroughs over time to motivate any future modeling. For GDP per capita, we found **a large disparity between boroughs**, with the minimum GDP per capita for years during the 2012 Olympic Games coming from Barking and Dagenham (~£16,000) and the maximum GDP per capita coming from Westminster (~£300,000). Figure 2 notes some summary statistics for London boroughs:

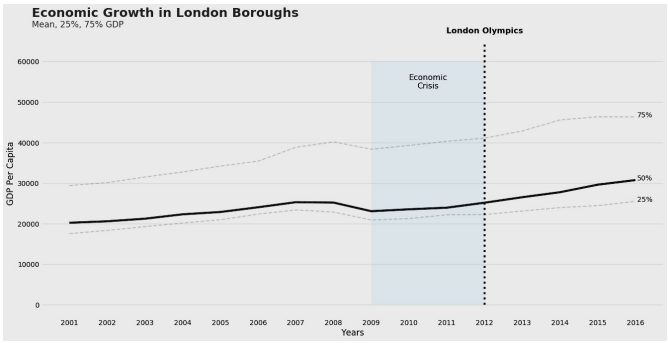

Figure 2: GDP per Capita Time Series

We can immediately note the dip in GDP per capita during the Great Recession, but just looking at the region around the 2012 Olympic year **does not reveal any significant conclusions about the impact of the London 2012 Olympics**.

Next, we wanted to see where the UK boroughs were geographically located and if boroughs with similar levels of income or GDP per capita were clustered together. Utilizing external GeoJSON data from [borough_boundaries.csv12]—containing the latitude, longitude, and polygon shape data for all UK boroughs—we decided to visualize income and GDP per capita for 2011 using Uber's Kepler.gl mapping tool.

We chose the year 2011 to establish the economic landscape before the London 2012 Olympics. For boroughs that have missing data, we median imputed them (with the middle shade of blue) because we discovered that the mean was dragged up by outliers such as York. From the heatmaps of income and GDP per capita, we concluded that they provide similar economic information on various UK boroughs, so we chose to narrow in on GDP per capita. Out of all the socio-economic indicators we explored, **GDP per capita acts as a proxy of each borough's overall economic well-being**, taking into account all industries that might have been impacted by the Olympics. From the heatmap, we can see that **Central and Western London have higher GDP per capita as represented by darker colors while poorer regions lie on the outskirts of London and the outer UK**.

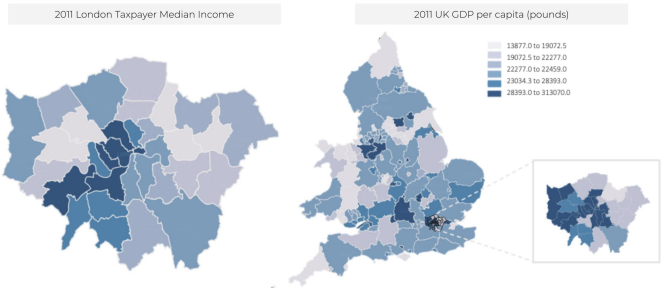

Figure 3: Geo Visualization of Income and GDP per cap

Selection of Target Variable

We have shown with our EDA that there exists a large degree of heterogeneity among different London boroughs prior to the Game, which makes nominal GDP per capita increase incomparable across different boroughs. For example, a £100 GDP increase in a borough with a £10,000 baseline and a different borough with a £100 baseline should not be treated equally. To account for such differences in baselines, we will instead mainly examine the **percent change of GDP per capita** from the expected rate. We consider a borough to be "better off" as a result of the Olympics if the net percent change from its expected GDP per capita is positive, and "worse" otherwise.

Building Control Unit & Donor Pool

To effectively isolate the impact of the Olympics, a set of control regions are needed for comparisons. That is, we need to find regions that are comparable to London boroughs while not being directly impacted by the London 2012 Olympics.

K-Nearest Neighbors (KNN)

Our control group for the analysis lacked international areas outside of our UK boroughs. We chose to add international cities due to the lack of borough-level data on many international cities. To identify cities similar to our target city, London, we used the **traditional KNN algorithm and Euclidean distance to compare two cities**.

The feature columns included socio-economic data on population, revenues of large corporations, city area, GDP per capita, and the number of higher education institutions from our [city.csv1]. The year of this data was not relevant to our analysis of "similar" cities, but we found data around the Olympic year, 2012. After normalizing these feature values and performing various imputation techniques (e.g. mean imputation) to fill out missing values, our trained model identified cities including **Paris, Chicago, San Francisco, Madrid, and Osaka**.

Although Beijing and Hong Kong were clustered near London, we left them out due to the lack of economic data on these cities. For the included cities, we found time-series GDP per capita data to complete our [gdp_with_international_cities.csv3] dataset. Below, we visualized the Principle Component Analysis (PCA) representation of our KNN:

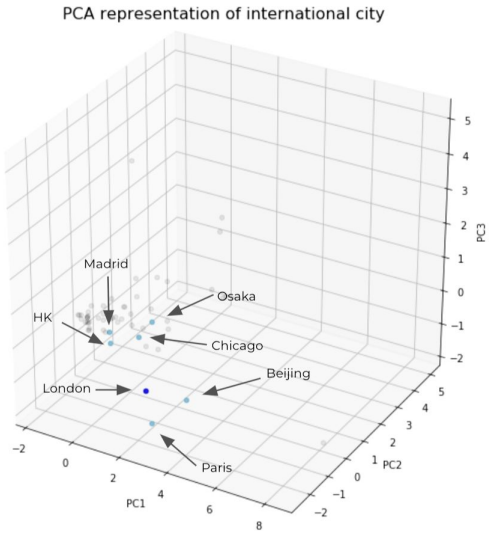

Figure 4 PCA Visualization of International City

Shallow men believe in luck or in circumstance. Strong men believe in cause and effect." — Ralph Waldo Emerson

Modeling Approaches

After identifying the control groups and the treatment groups, we now need to employ comparative case study methods in order to infer **causality between the impact of the London 2012 Olympics and the prosperity of different boroughs within London**.

Synthetic Control Method

Among existing methodologies, our research supports the use of the **synthetic control method (SCM) approach**. Unlike the traditional approach where one treatment region is matched with a separate control region, the synthetic control method leverages a combination of a limitless number of control units, which often provides a better comparison for the treatment unit than any single control unit alone. The algorithm aims to reproduce the GDP per capita of each treatment unit pre-intervention and to take this trend into post-intervention time for comparisons with the actual trend. In other words, we want to use the time-series GDP per capita data from all non-foreign boroughs in the UK as well as the foreign cities we have previously identified with our clustering algorithm to **reproduce the GDP per capita trend of each London borough pre-Olympics, and to project the expected GDP per capita for these boroughs without the intervention of the games**. To further infer causality, a placebo process takes place in which the same algorithm is repeatedly applied to every control region. The hope is that we observe a higher marginal difference in our treatment group in comparison to the placebo group, which we can then contribute as an impact from the Olympics.

This methodology improved upon existing methodologies such as Difference-in-Difference by using an algorithmic approach towards producing the best estimate of the counterfactual. With the placebo process, the model further contributes the marginal difference as an impact rather than a random process. We believe this will lead the projection line to be more precise, encapsulating more trends but also less subjected to arbitrary biases that we may have.

However, further research reveals that some constraints with the existing framework do not support our intuition behind the Olympic impact. First and foremost, under the SCM model, all weights

are constrained to be positive. This is likely not true with our Olympic scenario. Since resources are limited within the United Kingdom, we should expect some resources shifting from certain areas of the country to increase productivity within London in preparation for the Olympics. Such constraints on the weights do not give us the flexibility of assigning negative weights to those places with a likely-temporary negatively correlated GDP against those of London boroughs. Second, the algorithm uses a two-step optimization, which greatly reduces the search space in exchange for a possible improvement in running time. Since we don't have a large amount of data to optimize over, we do not wish to sacrifice our accuracy with processing time. Lastly, SCM weights are constrained to sum to one. While this might allow for better extrapolation, it largely increases the difficulty of the training process since some London boroughs, e.g. Westminster, has one of the highest GDP per capita by boroughs in the world.

Our Model - Customized Lasso

Due to the above limitations, we decided to employ our own model while taking inspiration from the SCM framework. In particular, our model similarly leverages past GDP per capita data of our control group to approximate the counterfactual for those in our treatment group. Except, in this case, our model base uses a **lasso regression** which uses cross-validation to select the best lambda penalty to determine the coefficients. The process to obtain placebo results is generally similar, wherein we simply ran our algorithm for all members of our control group. This serves 4 main purposes:

Purpose 1: Only one set of optimization is now necessary to obtain the coefficient for counterfactual which speeds up our modeling speed significantly

Figure 5: Table Structure Synthetic Control

Purpose 2: We now accommodate for possibly negative weightings to encapsulate **potential cannibalizing relationships for the counterfactual**

Purpose 3: Removing the weighting constraint to have all coefficients sum up to 1, we can now offer appropriate counterfactuals for boroughs with higher GDP than those in other UK boroughs and international cities

Purpose 4: Lasso regression allows us to account for both parsimony but also model fitting, and the additional cross-validation step helps us make the **optimal trade-off**. This is particularly important in this case since our predictors (members of our control group) are typically larger than our observations. Without a penalty like Lasso, this will inevitably lead to overfitting as a result of the curse of dimensionality.

Methodology Outline

- For every i-th borough of our treatment group
  - Obtain the pre-treatment (2011 and before) GDP per capita data of the i-th borough
  - Obtain the pre-treatment (2011 and before) GDP per capita data of members of our control group (other UK boroughs and International cities)
  - Run a cross-validation model to select the best lambda penalty that produces a model that predicts the out of sample GDP per capita of the ith borough using the GDP per capita data of the members of our control group.
  - Run the lasso model on the full pre-treatment dataset with the selected lambda penalty
  - Using the control group features in the post-treatment period, project out the synthetic control GDP per capita or our ith borough as our counterfactual
  - Compare and contrast the counterfactual with the real GDP of the one in the ith borough and obtain the difference per year to construct the average treatment effect of the Olympics every year
- Repeat the step above for all boroughs in our treatment group to obtain ATE
- Repeat the above steps by pretending the control group is a treatment group to generate placebo effect mentioned earlier

We believe this is a proper causal setup for a number of reasons. Firstly our control group very clearly encapsulates possible local trends within the UK that may be reflected upon London along with the inclusion of international cities to possibly encapsulate global macro-trends that may affect London as an international city. We believe this part is essential to control for the natural growth rate of real GDP as a result of increasing productivity and various economic policies.

Furthermore, we believe that we satisfy the crucial assumption of any causal modeling and that the only real major events that differentiate members of our treatment group (London boroughs) from those of our control group are the existence of Olympics, **our research did not return any significant results in terms of external events that may cause significant differences between the GDP per capita**. We believe this effect ought to be unique and large on our treatment variable. We understand that by 2016

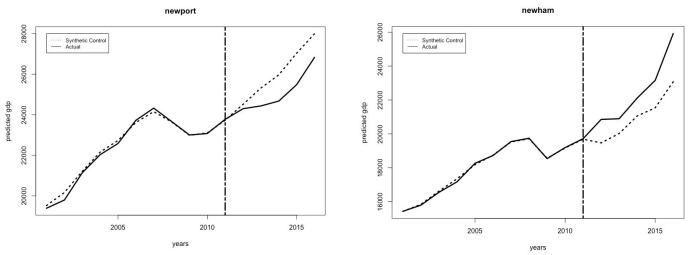

Figure 6: Output Comparison (L: Treatment, R: Control)

Brexit may have affected and introduced further potential GDP per capita movements that are harder to control, which is why we limited our holdout only until 2016.

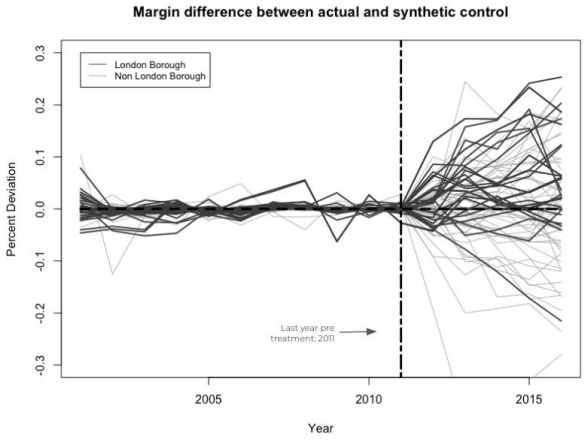

Figure 7: Margin (Percent diff between actual and predicted GDP per cap visualization

Interpreting Results

Figure 6 shows a set of sample model outputs. We obtain Figure 7 by plotting the **net percent deviation between the actual vs predicted GDP per capita**. Note that while we included all treatment units (in black) in our final result, we only kept the control units (in grey) for which the model had a maximum training deviation of 0.5 percent to ensure model accuracy. This leaves us with all 33 London boroughs as treatment regions, and 63 non-London boroughs as control regions. At first glance, there seems to be **more treatment units than control units with wider margins**. What is needed is an approach to quantify such observations, which we will be discussing in the next section.

Revamping the idea of p-value using Bayesian statistics

As we know, the statistical p-value is often used to determine the level of significance of model outputs. Empirically for us, we can compute the percent of control units with a more extreme margin than that observed of each of the London boroughs as a proxy for p-value. While taking the average across all p-values per borough can possibly resolve the year-to-year difference, we believe that there should be a coherent story to show how p-value progresses over time, which prompts us to use a **Bayesian update on the p-values**.

Specifically, we believe that p-values should follow a distribution rather than a one-point estimate. Given a specific treatment borough and a fixed year, we propose its significance to follow a **binomial distribution**, with the number of trials being the number of control units (63), and the number of successes being the number of control units for which our unit has a higher absolute margin. Using Jeffreys prior of Beta(½, ½), we can use the posterior distribution to then update our guess for the distribution of the parameter every period. Coining it as the **significance index**, we interpret this parameter as the probability that our treatment unit has a higher absolute margin than a randomly chosen control unit. Our results show that 7 out of 33 London boroughs have an estimated significance index greater than 85% by 2015, with Kensington-Chelsea having the highest significance index at 96.8%.

Borough	Index
Hammersmith and Fulham	0.968
Kensington and Chelsea	0.968
Hounslow	0.933
Richmond upon Thames	0.933
Barnet	0.893

Table 1: Bayesian P-value results

Quantifying financial impact

Using the GDP per capita margins we computed earlier, we can also quantify financial impact by computing the net GDP per capita gain/loss after hosting the Olympics. Leveraging the UK inflation dataset with 2018 as our base year, we computed the aggregate net gain for each London borough. From the table (below), we see that the median margin across London boroughs is consistently higher than that of the UK, which supports our claim of the overall positive impact of the Olympics.

Year	Median (London)	Median (UK)
2012	296.9297	25.83255
2013	1412.2424	83.96637
2014	2068.1216	98.13020
2015	2248.6445	290.20417
2016	2924.6604	-16.7900

Table 2: Median of GDP per capita impact

positive_uk	positive_london	year
0.6060606	0.5328095	2012
0.7878788	0.4920635	2013
0.8181818	0.5079365	2014
0.7575758	0.4920635	2015
07575758	0.5555556	2016

Table 3: Percent with positive margins

Index construction

In order to perform our back attribution and identify the main drivers of GDP margin and growth, we selected a number of features from external data sources. Through these indexes we seek to answer the following:

Question 1: How does the initial economic wealth and status of boroughs impact the economic gains the borough will have from the Olympics?
Answer 1: An economic index is constructed by extracting the principal component of a borough's unemployment rate, GDP, and median income. This index seeks to summarize the general economic wellness of a borough

Question 2: Do we observe disparate impacts across various minority groups? Is it uniform across minority groups or is it heterogeneous?
Answer 2: We obtained the ethnic breakdown of all London boroughs including percent of white, black, Asian, and mixed-race population. These values will be used to assess how the Olympics may have impacted these top minority groups differently.

Question 3: Do we observe an increase in particular for cities with a rich leisure supply that may be heavily utilized by the tourists? Is that increase long-lasting?
Answer 3: A leisure index is created by extracting the principal component of a borough's bar count and active enterprises in retail and accommodations. While the variable list is non-exhaustive, and there may be other forms of leisure supply, we believe that cities with a high value across these features are likely also high in terms of hospitality, making it a viable candidate to examine how Olympics affects boroughs with these tourism-relevant characteristics.

Decision Tree on Aggregated Margin

We hope to understand the determinants of the different clusters of ATE based on borough attributions using a **decision tree**. We observe that boroughs with a large mixed-race population and low economic indexes appear to have the lowest GDP per capita growth margin in aggregate and happen to be the only cluster with a negative NPV from the Olympics. In general, it appears that the growth of GDP per capita is generally lower for areas without a diverse population, which can be a trait of the more suburban boroughs. Similarly, it also seems that on the margin, **boroughs that are already more developed tend to experience a faster acceleration in terms of economic growth**. Furthermore, a large mixed-race population but a low proportion of the Asian population seem to be experiencing the largest GDP per capita NPV boost, indicating potential disparate outcomes for specific minority populations.

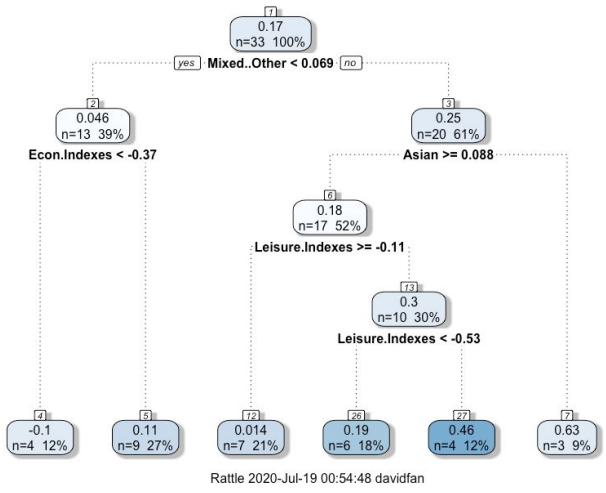

Figure 8: Decision Tree Visualization

Non-stationarity

It is worth noting, however, that our hypothesis already indicates potential nonstationarities for all our effects. Thus as opposed to simply using a

decision tree back attribution, we further investigated an attribution on a per-year basis. The process of this attribution was done by **regressing the yearly margin on the key variables and assessing how the coefficient changed across time**.

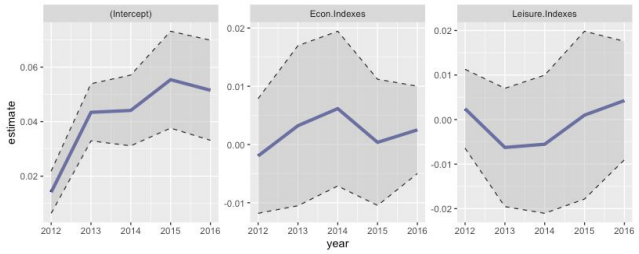

Figure 9: Time based regression graph [No Race]

The regression coefficient clearly showed a nonlinear trend. From the economic index, we could see that **during the year 2012, boroughs with lower economic status seem to have benefited more from the Olympics**. However, after 2012 the boroughs with higher economic index seem to be the ones that ended receiving a more positive impact.

Looking at the Leisure index, on the other hand, it seems like **in 2012 the Leisure index had a net positive effect but the effect is evidently not sustained in the long-term in the following years**. This makes logical sense from our point of view since the London Olympics was hosted around areas with lower economic wellness, and the surge of visitors likely has boosted the economic status of cities with a lot of leisure activities that people enjoy. However, these are simply momentary impacts from the time of the Olympics but do not translate to sustainable future development.

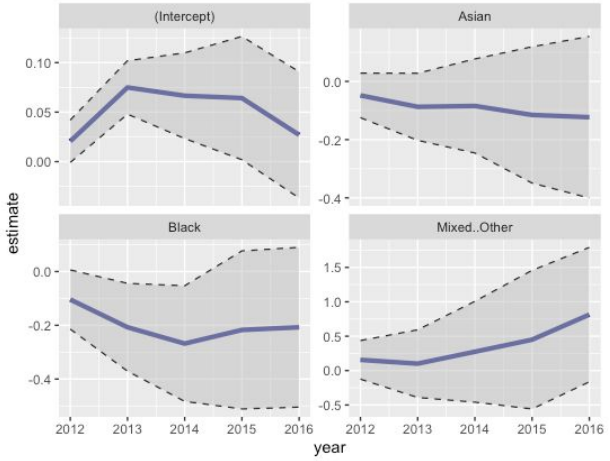

Figure 10: Time based regression graph on Race

It is also worth noting that there exists an interesting **disparity between boroughs with different groups of minorities**. While unlikely to be a direct causal link, we observed that for boroughs that tended to have benefited more from the Olympics, there are mixed-race individuals and less Asian and Black people.

Deep Dive

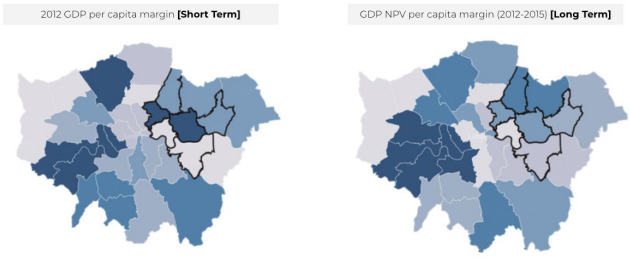

Figure 11: Short Term Long term Impact Comparison

In addition to exploring the different borough-level factors affecting GDP through back attribution modeling, we were curious to know the **difference between long-term and short-term impacts of the Olympics on different boroughs**. Figure 11 demonstrates the degree of GDP per capita impact among different boroughs in both 2012 alone as well as an aggregate from 2012 to 2015, with darker color representing a more positive impact. We observe two major points.

Point 1: Geo-proximity plays a significant role in the long run. While certain boroughs might not see the positive impact of the Games within the first year, those with nearby boroughs that do demonstrate a positive impact during 2012 are likely to demonstrate similar trends in the long run.

We contribute the first observation to a natural economic ripple effect of well-developed regions. Since GDP per capita is highly correlated with median income, in addition to contributing to their own borough's GDP, well off residents are likely to go to neighboring boroughs for food and entertainment, hence increasing their level of GDP.

Point 2: While the majority of the host boroughs (outlined in black) experienced a positive impact during 2012, such gains are not sustained in the long run.

For our second observation, our EDA research has

demonstrated that the Olympic Games were situated in a list of boroughs that were relatively underdeveloped compared to the rest. In accordance with our previous claim, the Olympics renovations and tourism boosted the GDP per capita in these regions in 2012. **However, It is important to note that the Olympic Park and subsequent stadium alone is not really a strong stand-alone tourist site**. While the London Olympics have certainly brought more awareness to London and have helped improve infrastructure at hand through various road construction and projects alike, these improvements and the long term growth stemming from them thereof are contingent ultimately around the kind of soft characteristics that the borough espouses.

For instance, when London hosted the Olympics and has further crystallized London's reputation internationally, it will potentially cause more tourists to visit London and bring in more foreign investment. **Yet, such values will most likely go to areas that are already more developed not only because it offers a better attraction spot but also because it has a mature market (more activities for tourists, more opportunities for investors).** Similarly with infrastructure, having better infrastructure within London, can potentially make transportation more efficient, but such an effect is unlikely to transcend directly to an improvement of life for individuals from a less well-off borough. **This is mainly because transportation serves as an enabler to allow people to seek better opportunities (through jobs, attracting tourists)**, however by just having transportation infrastructure, it does not drive to improve the employability of the people nor the attraction of the place directly; **While it allows people to reach that area better, it does not provide people a reason to visit those areas.**

It is worth noting that the most central areas in London such as Camden and City of London don't seem to reap as much accelerated growth. This is likely due to the fact that these cities already boast an extremely high GDP per capita such that the percentage increase appears less significant. **Nonetheless, they did still reap of a huge amount of monetary reward from the Olympics (around 30k GDP per capita)**

Concluding Thoughts

As hosting a successful Olympics takes years of preparation and millions of dollars of investment, it is sometimes dubbed as the "single most important event" for the host city. While existing research has explored the impacts of Olympics to the host city, **few have studied such impacts across different regions within the host city.** To address some of the limitations with existing analytical frameworks, we took inspiration from the synthetic control method and developed our own lasso-based model. Our results aligned with our hypothesis that while hosting the Olympic Games has a positive impact on the host city in the long run, such impacts are disproportionately reflected across different regions of the city. In particular, boroughs with lower economic indexes tend to be more well-off in the short term but not as well in the long term. We further noticed two interesting effects of hosting the Olympics at the borough level. **First, while the host boroughs receive short term gain due to spikes in tourism and entertainment, such gains do not last across the years.** Based on external research, we contributed this observation to the education and income gap that persists between the relatively underdeveloped host boroughs and the rest of the city. Second, there exists a spillover effect of borough impacts, where boroughs situated next to other boroughs who have demonstrated high positive impacts from the Games are likely to follow the same trend in the long term. We contribute this observation to the geographical ripple effect of economic gains.

The Olympics draws much public attention to the host city, particularly to the host borough within the city. **Rather than focusing on developing short term gains within the host borough, the city should utilize such attention and resources as an opportunity to narrow down the gaps that will impact the Olympic boroughs in the long run**. For starters, they can develop Olympics related educational programs to increase literacy across underdeveloped boroughs, designing Olympic infrastructures with the goals to be revamped into public properties after the Games, and increase security enforcements to permanently improve safety in the nearby regions. Only by looking past the Olympics Games, investing in changes that will simultaneously make future tourism safer and make local residency more desirable and accessible, the host borough will reap the most benefits from hosting the Olympic Games.

Caveats, challenges, and future research areas / unsuccessful routes

The main challenge we encountered during the course of this analysis is the data availability in terms of panel data for the host city. In particular, the back attribution could be further improved if we had more indexes to reflect other areas that are currently uncaptured. **However, we believe this methodology is a principally grounded way to understand the impact of the Olympics and we would recommend extending a similar methodology to other Olympics host cities with a focus on host cities that have actually invested in education programs and aimed at sustainable growth for comparison.**

For our submission, we included this report, our datafolio, and a zip containing code and a data folder for our external datasets. The files are aptly named towards their purpose in the analysis.

References

1. Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. Journal of the American Statistical Association, 105(490), 493–505. https://doi.org/10.1198/jasa.2009.ap08746

Data Sources

Provided Data Sources:

1. **UK_inflation**: UK inflation data using 2018 as a base year, covering back to 1751.
2. **london_taxpayer_income**: median annual income of taxpayers by borough and year.

Additional Data Sources:

1. **city.csv**: Used for our k-nearest-neighbors model, this dataset contains company financial data (https://www.lboro.ac.uk/gawc/datasets/da26.html), population data (https://www.kaggle.com/max-mind/world-cities-database), and other indicators such as city area, GDP per capita, and number of educational institutions (https://data.london.gov.uk/dataset/global-city-data) for several international cities.
2. **gdp.csv**: Time-series GDP per capita data for UK boroughs (https://www.ons.gov.uk/economy/grossdomesticproductgdp/datasets/regionalgrossdomesticproductallnutslevelregions)
3. **gdp_with_international_cities**: Gdp.csv augmented with time-series GDP per capita data for additional international cities, retrieved from (https://stats.oecd.org/Index.aspx?DataSetCode=CITIES)
4. **unemployment.csv**: TIme-series data for UK boroughs on unemployment rate (https://www.ons.gov.uk/employmentandlabourmarket/peoplenotinwork/unemployment/datasets/modelledunemploymentforlocalandunitaryauthoritiesm01/current)
5. **bars.csv**: Time-series data for UK Boroughs on number of bars (https://www.ons.gov.uk/businessindustryandtrade/changestobusiness/businessbirthsdeathsandsurvivalrates/adhocs/11806enterpriseactivesanddeathsbydistrict2009to2013)
6. **housing_price.csv**: Time-series data for UK boroughs on medium house prices (https://www.ons.gov.uk/peoplepopulationandcommunity/housing/datasets/medianhousepricefornationalandsubnationalgeographiesquarterlyrollingyearhpssadataset09)
7. **car_traffic_kilometers.csv**: Time-series data for UK boroughs on the vehicle miles traveled in kilometers (https://www.gov.uk/government/collections/road-traffic-statistics)
8. **productivity_filled_jobs.csv**: Time-series data for UK boroughs on the number of productivity jobs (https://www.ons.gov.uk/economy/economicoutputandproductivity/productivitymeasures/datasets/subregionalproductivitylabourproductivitygvaperhourworkedandgvaperfilledjobindicesbycityregion)
9. **productivity_per_hour.csv**: Time-series data for UK boroughs on number of productivity hours worked per week (https://www.ons.gov.uk/economy/economicoutputandproductivity/productivitymeasures/datasets/subregionalproductivitylabourproductivitygvaperhourworkedandgvaperfilledjobindicesbycityregion)
10. **enterprise.csv**: Distribution of enterprise category for UK boroughs in 2011 (https://www.ons.gov.uk/businessindustryandtrade/changestobusiness/businessbirthsdeathsandsurvivalrates/adhocs/11806enterpriseactivesanddeathsbydistrict2009to2013)
11. **ethnic.csv**: Distribution of ethnic groups for UK boroughs in 2012 (https://data.london.gov.uk/dataset/ethnic-groups-borough)
12. **borough_boundaries.geojson**: GeoJSON data on UK boroughs containing latitude, longitude, and polygon shape data (https://data.gov.uk/dataset/cd97a8df-e2fe-4f3d-a60f-1f871a317d31/counties-and-unitary-authorities-december-2016-full-extent-boundaries-in-england-and-wales)

Code

For our analysis, we used the following programming languages, tools, and packages:

- **Python** and **Jupyter Notebook** were primarily used for data wrangling, EDA, and preliminary visualization. We used standard libraries such as **pandas**, **numpy**, **matplotlib**, **seaborn**, **scipy**, etc.
- **R** and **RStudio** were used for Lasso, decision tree, and synthetic control modeling
- **Kepler.gl** by Uber was an external mapping tool we used for geographical visualizations: https://kepler.gl/demo/map?mapUrl=https://dl.dropboxusercontent.com/s/at4glqjcqf4w451/keplergl_w7381yi.json