



Analysing and Predicting the Spatial Penetration of Airbnb in the U.S.

Finding the Distinguishing Features of the Sharing Economy

Submitted by

Andrew Greateorex

Supervised by Licia Capra and Giovanni Quattrone

Master of Engineering

Department of Computer Science, UCL

2016

This report is submitted as part requirement for the MEng Degree in Computer Science at UCL. It is substantially the result of my own work except where explicitly indicated in the text. The report may be freely copied and distributed provided the source is explicitly acknowledged.

Abstract

Platforms exhibiting peer-to-peer characteristics akin to the sharing economy have altered the landscape of markets ranging from transportation to manual labour. Airbnb, the poster child and archetype of the sharing economy, is one of the most prominent examples that utilise this business model. Despite its decentralised, collaborative platform yielding an often cheaper alternative to traditional businesses, such as hotels, detractors argue that its success has been accelerated by a regulatory environment that is severely lacking. Unfortunately, there is little numeric evidence describing Airbnb's spatial penetration in metropolitan areas upon which to design such legislation. In this study, we apply several explanatory and predictive models to discern Airbnb's spatial-distribution in eight disparate urban areas in the U.S. with respect to a discrete set of geographic and socio-demographic factors. We find that areas of increased penetration are positively correlated with a creative class of young, talented and creative individuals, often in gentrified and touristic areas.

Keywords— Sharing economy, socio-demographic, classification, creative class

Contents

1	Introduction	3
1.1	Motivation	3
1.2	Report Structure	5
2	Related Work	6
3	Datasets	9
3.1	Research Questions	9
3.2	Spatial Unit	10
3.3	Cities	11
3.4	Airbnb Data	14
3.5	Explanatory Variables	15
3.5.1	Socio-Economic Indexes	16
3.5.2	People	18
3.5.3	City Geography	19
3.5.4	Other	21
3.6	Dataset Summary	22
4	Method	23
4.1	Data Normalisation	23
4.2	Preliminary Analysis	24
4.3	Multiple Linear Regression	26
4.3.1	Aims and Overview	26

4.3.2	Moran Test	27
4.4	Classification	28
4.4.1	Aims and Overview	28
4.4.2	Dataset Manipulation and K-Folds	28
4.4.3	Benchmark	30
4.4.4	Classification Algorithms	30
5	Results	35
5.1	Regression Results	35
5.1.1	Model Fit for Listings Density	36
5.1.2	Model Fit for Median Price	40
5.1.3	Model Fit for Review Density	42
5.2	Prediction Results	44
5.2.1	Quartile Classification	44
5.2.2	Extreme Classification	45
6	Conclusion	47
6.1	Summary	47
6.2	Critical Evaluation	48
6.3	Implications	49
6.4	Future Work	49
A	Data Collection	54
B	Chloroplasts and Other Maps	63
C	Regression Analysis	64
D	Classification Analysis	66

Chapter 1

Introduction

1.1 Motivation

The digital age has witnessed a surge in the number of platforms considered part of the 'sharing economy', a peer-to-peer hybrid market model that has defined a new way of doing business online. The sharing economy builds communities by sweating underutilised economic resources, such as assets and skills, and has flipped the traditional business model by transforming consumers into providers. Stapled by the idiom of collaborative consumption, it has seeped into a broad and diverse range of marketplaces, disrupting the previously considered (and saturated) status quo in business.

In the years that followed the global economic downturn in 2008, a plethora of peer-to-peer businesses emerged around the world. The business model offered an attractive post-crisis solution to overconsumption and exposed the truth that often sharing, or perhaps more accurately *renting*, is cheaper than ownership. More than this, growth was driven by an increased environmental consciousness which captured the imagination of the current generation of tech savvy individuals [1].

The Internet is the ecosystem that makes such inter-peer collaboration possible, providing a platform upon which to build communities of common motivations and interests. Companies that utilise user collaboration like this are becoming more prolific as they expose flaws in the classical

business model. eBay¹ and Craigslist² are considered the pioneers of collaborative consumption, and have paved the way for companies such as Uber³, TaskRabbit⁴ and Fon⁵, all operating near exclusively in specific, and comparatively disparate marketplaces. Uber’s aggressive expansion has transformed the transportation market in over 58 countries. Fon operates a bi-directional network of Wi-Fi hotspots, and Taskrabbit matches freelance labour with local demand.

The poster child of the modern collaborative economy, Airbnb⁶, allows users to rent out their spare rooms or properties. Founded in 2008, the company now has over 1,500,000 listings in 34,000 cities across the globe⁷. Unlike traditional hotel services that expand by undertaking lengthy and expensive property acquisitions for lodging conversions, Airbnb scales by increasing the number of hosts and matching them with guests.

The rapid growth experienced by Airbnb and other sharing economy companies has been accelerated by an exceedingly lacking regulatory environment that has been ineffective in forming policies to govern them. This has given rise to political and regulatory debates around the world about how best to compile legislation for businesses utilising collaborative consumption. However, few pieces of empirical evidence exist that elucidate the current climate of the sharing economy nor offer quantitative backing for effective legislation.

In this study, we use Airbnb as a case study into the sharing economy to build a numerically backed analysis of the service’s spatial penetration across the U.S. The results will provide regulators with a quantitative model upon which to design efficacious legislation. We present both explanatory and predictive models to explicate the socio-economic and geographic characteristics that have influenced the rapid expansion of Airbnb in large metropolitan areas.

Despite previous explanatory research into Airbnb’s growth in London, this is the first of its kind to attempt to explain Airbnb’s spatial context on a nationwide scale. The findings will likely fall

¹ebay.co.uk

²craigslist.com

³uber.com

⁴taskrabbit.co.uk

⁵fon.com

⁶airbnb.com

⁷<https://www.airbnb.com/about/about-us>

into one of two categories; either Airbnb’s penetration will follow local, city-wide patterns, or they will be indicative of universal phenomena. Any regulations that follow may lean on the results as a suggestion that legislation should be designed on a municipal or national granularity. Note, however, that this study provides numeric evidence to aid the design of such legislation, not regulatory recommendations.

1.2 Report Structure

This section outlines the structure and summarises the content of the report.

Related Work: This section provides an overview of current studies into Airbnb, the sharing economy and its regulatory concerns by looking at a mixture of qualitative and quantitative studies.

Datasets: This section focuses on the data that will be used to refine and build a model to conduct a spatial analysis into Airbnb. We begin by posing research questions, before defining the spatial unit to use and a critical selection of U.S. cities. Then, which dependent Airbnb variables to measure against a choice of explanatory variables aiming to capture an area’s characteristics. The section ends by providing a tabular summary of the dataset to be used in the analysis.

Method: This section specifies how the dataset is manipulated to match regression requirements, followed by a preliminary, non-quantitative analysis. We then move on to a multiple linear regression analysis, covering its process and spatial autocorrelation considerations. Finally, the steps taken towards computing an applicable and contextually correct classification model.

Results: This section describes the results obtained through both the regression and classification approaches for answering the posed hypotheses. We also provide an explanatory indication of the findings, comparing them to previous studies of a similar vain.

Conclusion: Finally, the paper concludes by summarising the results, critically discussing the method’s limitations, it’s contextual implications, and recommendations for potential future work.

Chapter 2

Related Work

This study contributes to a burgeoning range of literature on Airbnb, the sharing economy and its regulatory concerns. Related work is generally qualitative in nature, and only a handful of studies have empirically evaluated Airbnb’s spatial penetration in urban areas.

Despite little academic work into understanding its spatial penetration, other efforts provide insight into the company’s temporal evolution towards its market value of over \$25bn (July 2015)¹. The company’s rise is well documented; Schneiderman (2014) [2] found that short term rentals rose by 13,831 between 2010 and mid 2014 in New York City. In London, a city that saw 17.4 million unique visitors in 2015 [3], listings grew at a staggering 87% in 2014². PwC estimate that there are currently over 31,000 listings in London [4]. The blog maintained by Airbnb recently released figures that showed the total number of guests that utilised their platform to lodge grew over 353 times between 2010 and 2014³.

Whereas hotels have extensive upfront costs to grow, Airbnb may efficiently scale simply by amassing hosts and pairing them to guests. Airbnb’s rapid expansion may be attributed to the near zero marginal cost for additional users of the platform (Rifkin 2014) [5]. Two-sided platforms increas-

¹<http://www.wsj.com/articles/airbnb-raises-1-5-billion-in-one-of-largest-private-placements-1435363506?mod=LS1>

²<http://www.telegraph.co.uk/finance/newsbysector/retailandconsumer/11702399/Airbnb-boss-calls-UK-the-centre-of-the-sharing-economy.html>

³<http://blog.airbnb.com/wp-content/uploads/2015/09/Airbnb-Summer-Travel-Report-1.pdf>

ingly have cost-advantages over consumer services offered by traditional, incumbent firms, such as hotels. Bakos (1997) finds that electronic marketplaces reduce inefficiencies caused by buyer search and increase the ability of markets to optimally allocate resources [6].

Further explanations for Airbnb's popularity are that firstly, they are more personal than other hospitality services [7]; whereas hotel bookings online tend to be more robotic, Airbnb have profiles for each host and guest, establishing a more personal experience for users. Indeed, it may explain the reason as to why nearly 95% of Airbnb properties have a rating of either 4.5 or 5 stars. Comparatively, cross-listed properties on both Airbnb and TripAdvisor have a proportionally higher rating on Airbnb (Zervas, Proserpio and Byers 2015) [8]. Interestingly, this also gives rise to discrimination. Edelman and Luca (2014) found that in New York, non-black hosts charge roughly 12% more than black hosts for equivalent rental [9].

Airbnb's rapid growth has been nurtured by a severe lack of regulations in the sharing economy. While proponents make the case that peer-to-peer markets create wealth, stimulate optimal resource utilisation, increases independence and self-reliance by decentralisation and in some instances reduce environmental impacts (2013) [10, 11], detractors argue that the sharing economy's lack of regulations is predatory in nature [12]. For example, tourists have long been an important source of tax revenue and income for governments and hotels alike. As the sharing economy, and Airbnb, continue to aggressively expand in unregulated locales, industries and governments alike may suffer (Malhotra, Van Alstyne 2014) [13].

Zervas et al. focus on Airbnb's impact on the hotel industry in Austin, Texas (2013) [14]. The authors investigate whether Airbnb's spatial emergence is consistent with the locality of Austin's hotels; they estimate that the casual impact on hotel revenue is in the 8-10% range. This is in line with a separate analysis carried out by Credit Suisse, suggesting Airbnb placed downward pricing pressure on hotels in New York⁴.

The question remains, however, where in urban areas is Airbnb experiencing the most growth? Pioneering work by Richard Florida, an urban studies theorist, sheds light on where one may find

⁴<https://www.tnooz.com/article/airbnb-responsible-softening-new-york-revpar/>

higher concentrations of a 'creative class' in the U.S., a group of talented, bohemian individuals that foster a milieu that, as an implication, entices early tech adoption [15,16]. Such creative individuals would likely be early adopters of a service such as Airbnb. Clifton (2008) notes similar findings in the UK. [17]. The results of this study will provide backing for or against the work by Florida, and test whether his findings are still relevant in 2016.

Quantitative research into the geographic distribution of Airbnb is scarce. As well as Zervas et al.'s aforementioned spatial analysis in Austin, some examples in London offer additional insight [18,19]. Hughes [18] conducts an explanatory analysis on the growth of Airbnb in London. Her main findings show that touristic areas have a higher density of Airbnb's, and also tend to charge a higher price. A second, temporal, analysis conducted by Capra, Quattrone et al. [19] confirms Hughes' findings. Moreover, their results indicate that hosts of Airbnb listings are young people who do not own a house. However, the demographic makeup of American cities, where racial segregation⁵ and demographic divides is often high, is vastly different to London. Our work will serve as an illustration of the demographic differences between London and cities across the U.S.

⁵<http://fivethirtyeight.com/features/the-most-diverse-cities-are-often-the-most-segregated/>

Chapter 3

Datasets

This chapter explains in detail the process that was undertaken to create the sets of data necessary to carry out an unbiased analysis. We begin by defining a list of hypothesis to test. Next we select cities to analyse, make spatial considerations and define a list of dependent and explanatory features that will make up the final dataset on a per-city basis. The tools used to obtain, clean and collate this data are included in this section. The chapter ends with a tabular summary of the dataset.

3.1 Research Questions

This study is concerned with the adoption of Airbnb in cities across the U.S. A thorough analysis in this vain is reliant on three types of hypotheses. Firstly, *understanding* Airbnb’s core demographics on a spatial basis. Then, *comparing* the socio-demographic differences from city to city; does penetration change according to different cities? Finally, is it possible to *predict* what level of spatial-penetration Airbnb will have in a given area of a city? This presents three main research questions:

- What are the characteristics of urban areas that also contain a high or low density of Airbnb listings? How does this change for the offer, price and demand of Airbnb listings?
- Do these characteristics differ between American cities? If so, how are they different?

- In American cities where Airbnb has matured, is it possible to predict the listing-penetration in a given area?

3.2 Spatial Unit

Before compiling, processing and normalising final datasets, an important consideration was the spatial scope at which data would be collected. The aim here is to justify the granularity at which information can be collected and compared on a city-wide basis. Such data aggregation is a source of statistical bias and may heavily affect the results of aforementioned hypothesis test. The phenomena is known as the Modifiable Area Unit Problem (Gehlke et al. 1934 [20]). Since the results of this study rely on a spatial analysis of Airbnb, adopting a resolution that minimises bias and avoids spurious conclusions is paramount.

The smallest granularity at which the US Census Bureau collate data is at a 'tract' level. Although tract areas are not totally homogeneous, they provide three characteristics that will be important to this analysis. Firstly, since each unit has roughly the same population (average size 4000¹), statistical comparisons may be made between tracts. Secondly, census tracts cover a contiguous area. If this were not the case, it would be difficult to measure spatial autocorrelation by analyzing clusters and dispersion of data. Thirdly, census tracts represent a unit of measurement that captures a statistically significant number of data points in both city and self.

Note that all data captured with different spatial granularities will be converted to a tract level, otherwise they will be omitted. For example, information whose location is represented as coordinates should undergo a conversion such that its information can be understood on a tract level.

¹https://www.census.gov/geo/reference/gtc/gtc_ct.html

3.3 Cities

Previous work [18, 19] in London identified correlations between Airbnb and area characteristics. However, an analysis of a single city does not allow inference's to be made nationwide, due to the lack of sampling size significance. One may not assume that the demographics of one city reflects those of the whole country. To ascertain conclusive findings that may apply to all (major) metropolitan areas, a sample of cities should be chosen that represent the broad range of demographics, economics and traits that exist throughout urban areas in a single nation. We choose to analyse cities located within the U.S., as it has a rich cultural spectrum, covers a large region of land, and is also the most mature Airbnb marketplace.

Before identifying metropolitan areas to include in the study, conditions for sample size and city characteristics are required. First, the sample should be practical in terms of magnitude, but large enough to ensure any relationships that are found are statistically significant and avoid sampling error. To establish that identified correlations found in a city are valid, we follow a rule of thumb such that only cities with greater than 100 tracts may be considered as part of the analysis, and must have a population greater than 250,000.

As well as these constraints, an agenda to collate a list comprising of cities with highly contrasting socio-demographic data was adopted. Quantitative data from the most recent US Census Index was used to help make an informed decision. This was combined with other urban-index rankings such as a city's art and technology score as well as industrial capability. Due to the available Airbnb data, the number of cities to choose from was limited, however still reflected the varied makeup of cities in the US.

Since the study concerns Airbnb, the obvious first choice was San Francisco, where the company was founded in 2008 and is currently headquartered. As Airbnb's hometown, it offers insight into the most developed Airbnb marketplace. Furthermore, San Francisco, the second most densely populated U.S. city, is home to many budding technology entrepreneurs who work in the nearby heart of the U.S. technology scene; Silicon Valley. It is a very ethnically diverse city ², has a very high average

²<http://priceconomics.com/the-most-and-least-diverse-cities-in-america/>

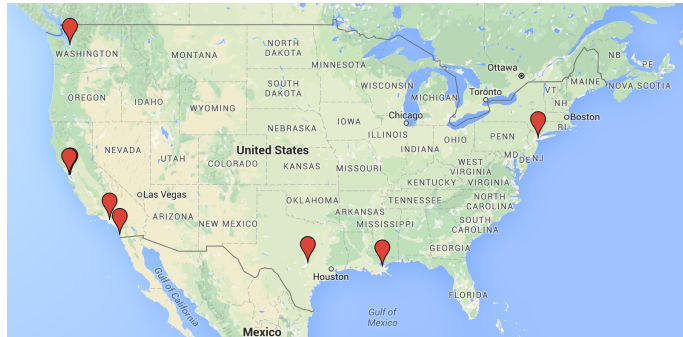


Figure 3.1: Map of Chosen Cities

age and, and despite having high median income, has a large disparity between the rich and poor.³

On the West Coast, we find New York City, which is the most dense, highest populated city in the U.S. It is a metropolis. It has many attributes that make it akin to London; it is a focus for immigration, it is expensive and is similarly very diverse. The city is second on the Global Power Index list⁴ and is the second largest city worldwide by GDP⁵. New York consists of five counties; New York County (Borough of Manhattan), Kings County (Borough of Brooklyn), Bronx County (Borough of the Bronx), Richmond County (Borough of Straten Island) and Queens County (Borough of Queens). However, the distribution of Airbnb in New York is highly concentrated in Manhattan, and is sparse across all other counties (see chloropaths in chapter 4.2). The lack of data points across other counties that make up New York City forces the analysis to focus only on Manhattan.

Across the bay from San Francisco lies the city of Oakland. Unlike San Francisco it serves as a center for trade and is the busiest port in California. Despite its close proximity to San Francisco, the characteristics of Oakland's demographic makeup differ considerably and median pay is roughly two thirds that of San Francisco's. Oakland is the next city to be added to the dataset.

The fourth identified city was New Orleans, which resides in Louisiana in the South. In stark contrast to Manhattan, New Orleans is the smallest of the chosen cities, with a population of 378,000.

³<http://www.fastcoexist.com/3050320/solving-inequality/the-10-us-cities-with-the-biggest-gap-between-rich-and-poor>

⁴<http://www.forbes.com/forbes/welcome/#7bbf56cb7370>

⁵<http://www.brookings.edu/research/reports2/2015/01/22-global-metro-monitor>

The city has seen a decline in population in recent times, where in 2010 it was 76% of what it was in 2005. As further proof of contrast to Manhattan, the median income of the city is \$26,900 (2010 US Census), to Manhattan's \$72,200, almost three times greater. It has one of the highest income diversity rates in the country, as well as ranking as the city of the highest average per annual homicide rates.

Having added two wealthy cities and a poor city, the next city that was added was Austin. Which, as the unofficial slogan 'Keep Austin Weird' suggests, differs vastly to both the metropolis of New York and the quaint New Orleans. Austin is the fastest growing city of the top 50 largest US cities⁶ and is not so ethnically diverse. Where New Orleans is predominantly black (60.2%), the majority of Austin's populous is white (66.8%). Unlike New Orleans and New York, Austin is also one of the safest cities in the US; the FBI ranked it as the second safest major city in 2012. It is also the youngest city in the dataset, with a median age of 31.1, contrasting to the US median age of 36.8.

Back in California, Los Angeles was the next city to be added. Los Angeles is a global center of commerce and has a diverse economy in business, technology, culture and sport. It has the highest educational diversity in the country and ranks highly on the diversification of its economy business-sectors⁷. Despite its size and economic power, it has a low median income and a disproportionately high cost of living. Despite having similar median income to New Orleans, the cost of living in Los Angeles is 43% higher.

The Pacific Northwest city of Seattle, in Washington, is the seventh city to be added to the dataset. A city home to Amazon⁸, Microsoft⁹ and Boeing¹⁰, Seattle is an important center for technology and is a major gateway for trade with Asia. Like Austin, it is a predominantly white city. However, it is far older, has a much higher median income and a greater cost of living.

Finally, San Diego was added as the third major city (population greater than 1,000,000) to the

⁶http://www.slate.com/blogs/moneybox/2015/05/21/population_growth_in_u_s_cities_austin_is_blowing_away_the_competition.html

⁷<https://wallethub.com/edu/most-diverse-cities/12690/#detailed>

⁸amazon.com

⁹microsoft.com

¹⁰boeing.com

dataset. The city, which has an immediate proximity to the Mexican border, is not a technology hub like Seattle or New York. Its main economic engines are the military and tourism. Due to its closeness to Mexico, it has a large Hispanic population and a low proportion of blacks (6.7%). This is in stark contrast to New Orleans, which has a black proportion of 60.2%.

Table 3.1 summaries the socio-demographic information of the final set of chosen cites.

City	Pop	Median Age	Median Income	% White	Cost of Living
Austin,TX	885,000	31.1	32,300	68.3	95.5
Los Angeles, CA	3.8m	34.4	27,800	49.8	136.6
Manhattan, NY	1.63m	35.5	72,200	48.0	216.7
New Orleans, LS	379,000	34.6	26,900	33.0	95.0
Oakland, CA	406,000	35.1	32,000	34.5	139.1
San Diego, CA	1.36m	35.6	33,200	58.9	132.3
San Francisco, CA	837,000	38.5	48,500	48.5	164.0
Seattle, WA	652,000	36.1	43,200	68.3	121.4

Table 3.1: Summary of Demographics in Chosen Cities

The final dataset of cities provide a sound representation of the size, demographics and geography of cities in the United States. Conclusions gathered from an analysis into the offer, demand and price of Airbnb in these locations will give reason to believe that any relationships will hold across the country.

Having decided on which cities to include in the study, the next step is to take an in depth look at the Airbnb features we will aim to predict.

3.4 Airbnb Data

Broadly, this study is concerned with the *offer*, *price* and *demand* of Airbnb listings, as embodied by the metrics listing density, median price and review density. That is, the number of Airbnb listings that are available in a unit area (tract), the median price paid for a listing in a unit area,

and the number of reviews that are left in a unit area¹¹.

A comprehensive dataset that was compiled by Murray Cox as part of his Inside Airbnb project¹² was used to gather consumer-facing Airbnb data. Given a set of coordinate, price and number of reviews per Airbnb listing per city, data underwent a collective conversion to be represented at a tract base level.

For each city, there were three datasets given; listings, neighbourhoods and reviews. The raw listings file was the only dataset that was of interest. In said file, where each row represented a single Airbnb listing, the columns of significance were the *latitude*, *longitude*, *price* and *number_of_reviews*. Then followed a process of extraction and normalisation. For each listing, convert the latitude and longitude to its corresponding tract ID and then enumerate how many listings were in each tract. This gives the number of listings in an area. For price and reviews, again we find how many data points existed in each area, then calculated the median price in each area and the number of reviews in an area. The number of listings and the number of reviews are then normalised to square kilometers. Table 3.2 summarises the cities by Airbnb data.

Previous research on Airbnb in London suggests that it is more difficult to explain the price and review density in an area than it is to explain the listing density. That being said, London has a vastly different geo-social-economic makeup than many American cities.

3.5 Explanatory Variables

A extensive set of explanatory variables that capture the social, economic and geographic characteristics is necessary to best understand the aforementioned Airbnb attributes. To make an informed choice of variables to use, we refer to several pieces of literature that have analysed the general effects of specific metrics on other city characteristics such as technology concentration. Explanatory variables are split into socio-economic indexes, people, city geography and other.

¹¹Review density used for demand since high proportion of guests (80%) leave reviews

¹²InsideAirbnb.com - Murray Cox

City	Number of Tracts	Number of Airbnb Listings	Number of Airbnb Reviews	Median Price Per Night
Austin, TX	220	5,193	46,019	180
Los Angeles, CA	1,011	17,044	236,175	115
Manhattan, NY	288	16,041	200,618	155
New Orleans, LS	177	2,646	53,519	135
Oakland, CA	114	1,155	16,302	95
San Diego, CA	314	3,530	43,455	149
San Francisco, CA	197	6,361	126,132	169
Seattle, WA	132	2,711	57,871	119

Table 3.2: Summary of Airbnb Data

The majority of data was collected from the US Census bureau, which gathers decennial population data. The data is compiled on a tract granularity and grouped by state. This study focuses only on tracts bounded by the aforementioned cities. Thus, matching tracts were identified and extracted, while all other tracts in the state were removed. The typical process followed was to identify the correct county code of a city and select the corresponding tracts. However, this was not the case for Austin, Texas, which resides completely in one county and partially across two neighbouring counties. To find Austin’s tracts, we used the *FME Workbench*¹³ to extract information from Austin’s shapefile.

After collecting the following data, all variables were combined into their respective city tracts.

3.5.1 Socio-Economic Indexes

Indexes are used to measure types of diversity in cities. Since there are many different forms of social diversity that may affect the characteristics of Airbnb in an area, we choose several metrics that have previously been shown to have high impact on technology uptake and urban consumption.

Race Diversity Index; (US Census Bureau) The Race Diversity Index [21] is a measure of how much racial diversity exists in an area. First coined by Meyer and Macintosh [22], it is formulated as a Gini-Simpson Index [23] and acts as a probability measure. It measures the likelihood that

¹³safe.com

two people selected at random from a given area represent the same type. In this case it is a measure of whether the race of the chosen person is the same. We formulate the problem with seven distinct racial categories: white, black or African American, Hispanic or Latino, American Indian or Alaska native, Asian, native Hawaiian or Pacific Islander and finally two or more races. The index is expressed as a percentage, with 0 representing a completely homogeneous area. The greater the race diversity index, the greater the probability that two people selected from random will be from different races. The formula is given below.

$$1 - \sum p^2$$

Income Diversity Index; (US Census Bureau) The income diversity index [21] shows how diverse an area is in terms of average household income for the population of that area. It is calculated using the Gini-Simpson index using three distinct wage bands: low income (annual incomes less than \$35,000), middle band income (annual incomes between \$35,000 and \$100,000) and high income (annual incomes greater than \$100,000).

Bohemian Index; (US Census Bureau) A bohemian is a socially unconventional person with interests in art or literacy. Richard Florida’s 2001 paper *Bohemia and Economic Geography* [15] examines the relationship between geographic concentrations of bohemia and a strong technology presence by directly measuring the bohemian population at an MSA (Metropolitan Statistical Area) level. Though there are other variations of the Bohemian Index (notably Clifton, 2008 [17]), we use the Florida’s definition, which is defined as the proportion of the number of bohemians to the number of residents in an area as compared to the national proportion of bohemians to the number to the total population

Talent Index; (US Census Bureau) The basic talent index [24] is a measure of highly educated people, defined as those with a bachelor’s degree or above. The index is normalised per thousand people and based on 1990 decennial census Public-Use Microdata Sample (Florida, 2002). Florida hypothesises that a high talent index is correlated with a larger concentration of bohemians. Given this, we may infer that areas with a strong technology presence, such as those areas with high

Airbnb uptake, will have a higher index for talent.

3.5.2 People

Unemployment Proportion; (US Census Bureau) The unemployment proportion is calculated as the number of people aged 16 and over currently out of work (unemployed), against the total number of people in an area. Unemployment rates often provide a strong indication of the economic health of an area. Florida's work on the Creative Class (2004) suggests that areas of lower unemployment (amongst other factors) are symbolic of a creative class, and transitively may lead to greater technology concentration. However, the Wall Street Journal ¹⁴ found a large percentage of renters are offering up living spaces due to unemployment. In Paris, only one third of hosts have full time jobs. If the relationship holds in across the US, we may see a correlation between unemployment and Airbnb dependent variables.

Poverty By Income Percentage; (US Census Bureau) Michael Zweig [25] defines poverty as "a state of deprivation, or a lack of the usual or socially acceptable amount of money or material possessions". In the US, the most common poverty measurements are the "poverty thresholds", as defined by the US Census Bureau ¹⁵. This variable is calculated, in a given area, as the percentage of households in poverty (as defined by their income) against the total number of households in that area. The logical hypothesis is that Airbnb's penetration will fall in areas of increased poverty.

Median Household Income Estimate; (US Census Bureau) For each area, the US Census Bureau measures the median household income for the local populous. A temporal study on Airbnb in London [19] showed that income became an increasingly more negative and statistically significant variable, signaling that Airbnb hosts are using the extra income generated from Airbnb to support themselves.

Median Household Value Estimate; (US Census Bureau) The US Census Bureau also provides a measure of median household value for each area. We predict that, for the dependent variable

¹⁴<http://www.wsj.com/news/articles/SB10001424052702304007504579348781019477384>

¹⁵<https://www.census.gov/hhes/www/poverty/data/threshld/>

Airbnb Price, there should exist some positive correlation to household value. Variables such as median household value may be used as a strong indicator of socio-economic makeup of a city, and may be useful in identifying clusters of cities later.

Proportion of Young People; (US Census Bureau) This was calculated as the proportion of people aged between 20 and 34 years old in a given area against the population of that area. Florida (2001) [15] suggests that, as well as the bohemian index, areas with higher concentrations of young people is often a driver of the technology uptake in that area. Florida argues that areas made up of a 'creative class' attract business and technology. Indeed, the Technology Adoption Lifecycle model [26] predicts that the 'Early Adopters' of technology are generally younger, more educated individuals. In our case, we anticipate that younger areas will represent larger density of Airbnbs.

3.5.3 City Geography

Distance to Center - Before computing the distance from each area to the city center, it is in our interest to first choose the type of city center a city has and whether or not the city is poly-centric. London, is an example of a poly-centric city, as shown by Roth et al. (2010) [27]. In the US, it is a fair generalisation to use the 'downtown district' or CBD (central business district) as the center of the city. Some cities, such as Philadelphia, use the City Hall as its center. The metric is measured in meters from the center of a given tract.

Points of Interest; (Open Street Map) A point of interest (PoI) is a feature on a map that someone may find useful or interesting [26]. Examples of PoI's include pubs, town halls and post offices and (by OpenStreetMap's definition) exclude linear features such as roads and landfills. A study on the geography of Airbnb in London [18] found that the 'tourism factor' of an area, as shown by PoI density, had the greatest positive significance on the number of Airbnb offerings in that area. We expect that the relationship will hold for American cities, such that areas of higher PoI concentration, indicating greater tourist appeal, will also have increased a Airbnb density.

OpenStreetMap, a collaborate project to create a free and editable world map, was used to collate thousands of coordinate pairs of Points of Interest for each city. Using the online tool Overpass-

Turbo, where one may search for, a set of types of data points. The collated PoI's were in line with Hughes's method to indicate tourist appeal and area attractiveness. only including PoI's that fell under one of the following categories: Accommodation, Eating and Drinking; Attractions; Sports and Entertainment; Retail. A script was written to work out the density of PoIs in each area given the raw coordinate list from OpenStreetMap.

Number of Hotels; (Google) Despite a previous analysis showing that, in London, there is little relationship between hotels and Airbnb adoption, we do not know whether the conclusion holds in US cities. Airbnb's economic blog, which reports and measures Airbnb's effect on city economies, states that 72% of Airbnb properties in San Francisco are outside the central hotel district. However, little other work has been covered on the spatial patterns between Airbnbs and hotels. Intuitively, the density of hotels in an area should give a sound proxy for the level of tourism of that area. Furthermore, results highlighting where Airbnbs appear in a city relative to hotels will provide regulators with a source of quantitative information to make more informed decisions.

Since there is no publicly available dataset for the number of hotels in all cities, hotel data had to be crawled from Google. Searching Google 'city_name' + 'hotels' returned a list of hotels, spread across a number of pages, which had basic information such as address, star rating and basic features. Using the Chrome app kimonolabs, a list of all city hotel addresses was returned. A script was then written to, firstly, convert the raw addresses to latitude-longitude pairs and, secondly, to collate all coordinates into hotel densities per area.

Bus Stops; (Open Street Map, official city websites) The strength of an area's infrastructure and transport links have historically been a key component in the performance of property prices, due to the ease of connection to major areas of that city. For tourists visiting a city, although they may spend time and money in tourist centers, their choice of where they stay is likely influenced by the connectivity of an area. The density of bus stops in an area provides a proxy to the strength of said area's transport links. Thus we expect to see a relationship between Airbnb offerings and the number of bus stops. This metric is normalized to the size of the area it resides in, giving the density of bus stops.

Despite the prevalence of OpenStreetMap nowadays, some cities naturally have gaps in their data. Though cities such as New York and Seattle have extremely rich datasets on the map, New Orleans has been found to be lacking in particular data types (such as bus stops). PoIs, which are also sourced from OpenStreetMap, do not have this issue since the sheer number of data points dictates that it will still be distributed much in the same way. However, for bus stops, a single type of data point, this causes an issue. Thus, some data had to be crawled from Google and official bus timetables.

3.5.4 Other

Proportion of Owner Occupied Residences; (US Census Bureau) [19] found that the proportion of owner occupied residences became negative and very significant as Airbnb found its place in the London market, suggesting that Airbnb hosts are typically not the owners of the property they rent. Airbnb is a mature player in the North American marketplace, just as it is in London. A continuation of the proportion of owner occupied residences having a strong and opposite relationship with Airbnb adoption is expected, and at the very least will allow for comparisons to be made between the characteristics of Airbnb in London and American cities.

3.6 Dataset Summary

Each city has a corresponding dataset with the following features.

Metric	Source	Description
Airbnb Count	InsideAirbnb	Normalised to square kilometre, the number of Airbnb listings in an area
Airbnb Price	InsideAirbnb	The average price per night of all listing in an area
Airbnb Reviews	InsideAirbnb	Normalised to square kilometer, the number of Airbnb reviews in an area
Hotel Count	Google Maps	Normalised to square kilometer, the number of hotels in an area
POI Count	OpenStreetMap	Normalised to square kilometer, the number of points of interest (POI) in an area
Bus Stop Count	OpenStreetMap	Normalised to square kilometer, the number of bust stops in an area
Talent Index	OpenStreetMap, Official Timetables	The number of people in an area with degrees higher than an associate degree per thousand people
Income Diversity Index	Census	The variance between three levels of income in an area
Bohemian Index	Census	For ages 16 and over, the proportion of people employed in arts, entertainment and media in an area to the same proportion nationwide.
Race Diversity Index	Census	The variance between the total counts of people of differing races in an area.
Unemployment Ratio	Census	For ages 16 and over, the proportion of unemployed to the total populous
Proportion of Poverty by Income	Census	The proportion of the number of residents with income in poverty to the number of residents with income in an area
Proportion of Young Persons	Census	The proportion of people aged between 20 and 34 to the total populous of the area.
Proportion of Owner Occupied Residences	Census	The proportion of dwellings that are owned, to those that are occupied
Median Household Income	Census	The median estimate of household income in an area
Median Household Value	Census	The median value of a household in an area
Median Age	Census	The median age of an area
Distance to Center	Routing Machine	The distance from the center of a given area to the center of a city

Table 3.3: Dataset Summary

Chapter 4

Method

The main aim of this project is to analyse offer, demand and price in terms of spatiality, to supply regulators and policy makers the quantitative evidence they require to construct informed legislation. Intuitively, the aforestated demographic and geographic urban attributes will offer insight into the disparities between Airbnb penetration for these topics. This section focuses on the design and implementation of mathematical models that will both explain Airbnb penetration and also predict penetration in new cities. The *R* programming language and a stated collection of core libraries were used to normalise and analyse the data. Relevant code is shown in the appendices and in the attached folder.

4.1 Data Normalisation

The assumption of normality may tend to lead to Type I and Type II errors, code for the rejection and acceptance, respectfully, of the null hypothesis. Micceri (1989) [28] points out that real world data rarely normally distributed, which we quickly discovered when visualizing our dataset.

Since this is a multivariate analysis using regression, whereby we aim to compare phenomena in multiple metropolitan areas, the data must be scaled and transformed to remove skewness. All the cities in our dataset have different characteristics, and as such is not possible to compare them

linearly. For example, prior to scaling, the influence of median household income in Manhattan is not comparable to that of New Orleans. First, we check for skewness and convert each explanatory and dependent metric to fit a normal distribution. Over eight cities, this constituted of manually checking 144 metrics, and testing multiple functions to ensure normality.

On inspection, some variables contained noticeably more skew than others. The most severe cases of skew were the hotel densities, the POI densities and the bus stop densities, which were heavily positively skewed in multiple cities. This is because the hotels, POIs and bus stops are more concentrated and clustered in specific areas of cities, hotels the most so. These variables required a $\log(x + 1)$ transformation. Common transformations were $\log(x + 1)$, $\log^2(x + 1)$, $1/\log(x + 1)$ and $\exp(x)$. To reduce the effect of adding 1 to data points, data in the format of a proportion (i.e. $0 \leq x \leq 1$), was first converted to a percentage (i.e. $0 \leq x \leq 100$). Adding 1 to a percentage makes little difference to the relative value of the data point, unlike adding it to a proportion. Some raw distributions were already normal, and were thus left as such. Data was then converted to zscores so that multiple variable distributions could be compared.

4.2 Preliminary Analysis

After datasets were compiled and normalised for each city, a preliminary analysis was conducted to build a basic understanding of relationships between variables. The final regression and prediction models may thereby be cross checked with this preliminary analysis for a sanity check.

To begin, variables were visually inspected on city chloropleths, created using the compiled dataset and official city-shapefiles; figures 4.1 to 4.4 are examples of such inspections. Since one may be created for each variable for each city, only a small selection are presented. Darker areas represent greater concentration.

As expected, Airbnb listings are more concentrated towards city centers, supporting findings in London where distance from center had an inverse relationship with Airbnb penetration. Note that the most concentrated areas of Airbnbs in New York City are virtually all in Manhattan county, and much of the other counties are zero-valued (white).

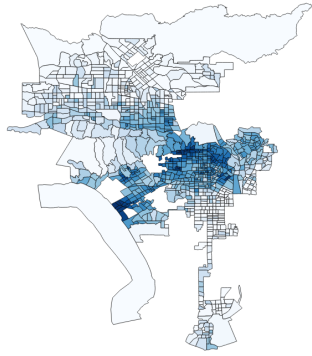


Figure 4.1: Los Angeles Airbnb Adoption

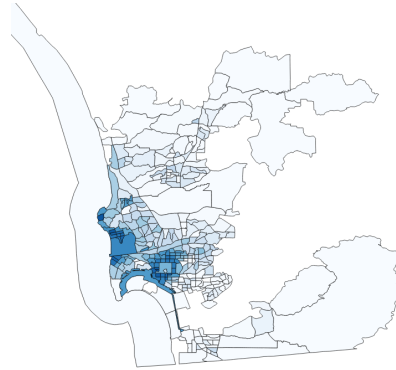


Figure 4.2: San Diego Airbnb Adoption.

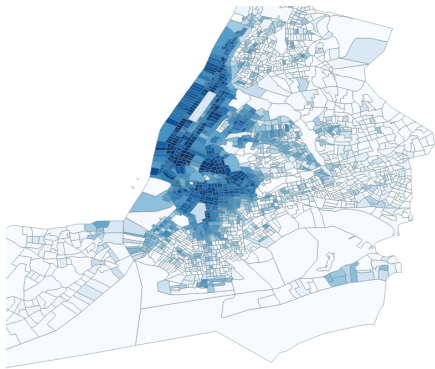


Figure 4.3: New York Airbnb Adoption

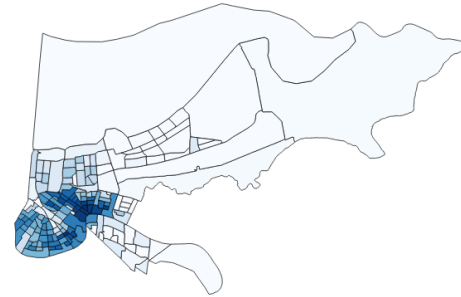


Figure 4.4: New Orleans Airbnb Adoption.

Following a visual inspection, correlation matrices were computed using the *corrplot* library in *R*, which allowed us to compare the variables used in the study and to have a stronger idea behind what results we may expect.

There are further chloroplaths and correlation matrices in the appendix section.

4.3 Multiple Linear Regression

An initial, simple least squares regression model was used to validate and recognise which variables were important in explaining Airbnb metrics. It helps to understand how these Airbnb metrics vary relative to the geo-demographic characteristics defined above. Moreover, using regression as an initial model provides results that are simple to interpret and allows us to then progress onto more advanced predictive models based on these results. The formula for least squares regression is given below:

$$y_i = \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \dots, n,$$

Where y_i represents the dependent variable (an Airbnb metric), x_i represents the regressors (the aforementioned explanatory variables), ε_i the error term and β the beta coefficient.

4.3.1 Aims and Overview

Given the final sets of data for each city, we construct a least squares regression model. The results of which produce an adjusted R^2 and β values for each model (listing density, median price, review density). The β value shows the importance that a variable has in the final model and the adjusted R^2 shows how well data fits a statistical model. An R^2 equal to 1 indicates a perfect fit of the line to the data, in our case that means that one of the Airbnb metrics is perfectly explained by its complementary model. Furthermore, each term has an associated p -value. The p -value tests the null hypothesis that the coefficient is equal to zero, and hence has no effect. A low p -value (generally < 0.05) indicates that one may reject the null-hypothesis, this means that terms with low, associated p -value's are a meaningful addition to the model.

Following a full regression on all variables, we run stepwise regression on the full model to reduce the model to fewer variables of more importance. The stepwise process works by building a parsimonious model from the initial least squares regression, which successively adds and removes variables based on the t-statistics of their estimated coefficients. Despite the drawbacks of stepwise regression, it will find the minimal, most simple model that explains the most variance in the output data. A smaller model is easier to interpret and the identification of a smaller subset of useful

features makes the next step of prediction more simple and powerful, whereby a smaller number of predictor variables will reduce the effects of overfitting by disposing the *curse of dimensionality*, though this is an additional consideration as oppose to an expected issue.

4.3.2 Moran Test

"Everything is related to everything else, but near things are more related than distant things." - Waldo R. Tobler.

The regression model is based on an assumption of independent observations, which in the case of geo-spatial data-points, is not always the case. This means that any results we obtain may be misleading if observations are not independent from one another. In the context of this study, an example is asking whether young people are clustered in specific areas that cover multiple nearby areas.

Spatial autocorrelation is the degree to which one object is similar to nearby objects. We use *Moran's I* (Index) to measure spatial autocorrelation. The result of Moran's I is a number between -1 and 1. 1 indicates perfect positive autocorrelation, in our example this would imply young people are extremely clustered on a map. A result of -1 indicates perfect negative autocorrelation, dissimilar values neighbour one another. A good example of this is a chessboard. A value of 0 implies no spatial autocorrelation.

Moran's I is calculated as follows:

$$I = \frac{N}{\sum_i \sum_j w_{i,j}} \frac{\sum_i \sum_j w_{i,j} (X_i - \bar{X})(X_j - \bar{X})}{\sum_i (X_i - \bar{X})^2}$$

where the number of spatial units is denoted by N , and is indexed by i and j . X is the variable of interest and \bar{X} is the mean of X . Finally, w is a matrix of spatial weights, indexed by w_{ij} .

4.4 Classification

The second step of our analysis is to attempt to predict the penetration of Airbnb in American cities. If classification yields sound results it will provide regulators and policy makers with quantitative evidence upon which to design new regulations. If it is not possible to accurately classify Airbnb penetration in unseen cities, we may infer that the Airbnb’s urban penetration cannot be explained universally, but on a more local scale.

4.4.1 Aims and Overview

The classification analysis will take a supervised approach, such that we may infer a function from a set of labelled data. Here, the labelled data refers to the penetration of Airbnb in a city and is the natural progression for this analysis. Since the dataset has been setup for regression analysis, the next step is to alter the dataset to prepare it for a supervised learning classifier. Next, benchmark methods are discussed to compare the results of the classifier to. Following this, we discuss the various classification algorithms that were considered and make an informed choice for which to use.

4.4.2 Dataset Manipulation and K-Folds

The datasets for this step should have a reduced dimensionality to cut both the chance of overfitting and the time taken to run these algorithms, many of which are based on non-linear objective functions. The previous regression analysis identified the most important variables for each model, which may be used to trim the new datasets. The chosen features are discussed in the results section.

Furthermore, since the target variables are continuous, the datasets are in an incompatible form for classification. They must be stratified into a categorical hierarchy, to form a discrete set of labels. The target variable was converted into quartiles, resulting in the following labels for listing density:

VLP: Very Low Penetration

LP: Low Penetration

HP: High Penetration

VHP: Very High Penetration

In layman’s terms, Airbnb listings with a density less than the lower quartile were labelled as VLP, listings with a density between the lower quartile and the median were classed as LP, and so on. This is an example of ordinal classification.

While some supervised approaches for classification are inherently multiclass, many lack the ability to handle multiclassification problems, but only binary classification problems. A multiclassification problem is one where each training point belongs to one of N distinct classes. Here, the aim is to construct a function which can class a new data point as one of four classes. The *one-vs-all* strategy is used, resulting in maximisation problem over four binary classifiers, whereby the highest confidence score is taken to suggest the correctly classified label. The maximisation problem for the *one-vs-all* strategy is provided below:

$$\hat{y} = \arg \max_{k \in 1 \dots K} f_k(x)$$

Where \hat{y} represents one of k (four) categories: VLP, LP, HP, VHP. $f_k(x)$ is the classification function, which takes in a vector of datapoints for each tract, and returns a confidence score for each four categories. The model’s performance may be measured in multiple ways, in this study we focus on the accuracy, precision and recall relative to the results of a benchmark, as defined in the next section.

A robust predictive model is defined by its ability to perform well on previously unseen, test data. Thus, we first split the dataset into both a training set and a test set. The model is created using the training set and is followed by predicting the test set. This method is known as the hold out method. However, to increase the amount of testing possible on unseen data, we use a technique called *k-folds cross validation*. In k-folds cross validation, the initial dataset is split into k equal sized samples. Of these k samples, k-1 are combined and used as the training set, and the final one used as a test set. In this, every data point is considered exactly once in the test set, giving a stronger case to the argue the performance of the model and its ability to generalise.

In this model, k represents eight cities. For each of the eight iterations of the algorithm, seven

cities will make up the training set, and the final city as the testing set (the 'unseen' data sample). This testing strategy will provide insight into Airbnb's penetration over the spectrum of U.S. cities. If the classifier convincingly exceeds the benchmark in all cities, we may conclude that the given model presents valid, quantitative proof that explains where Airbnb will grow in a city.

4.4.3 Benchmark

Raw values from the model are alone inconsequential without a metric to contrast them to. In this section we define the benchmark/ baseline that we will use to compare our model to.

Multiple methods exist for computing a benchmark. We present two benchmark approaches for comparison. The first is a simple random predictor. Since the classification dataset has four categories of equal quartile length, the first benchmark classifies simply at 25%. However, this is based on the assumption, and mistaken belief, that Airbnb's growth in a city occurs as a random event. Given all indications of previous research, the area that a new Airbnb will appear in is not random. We expect that it is not an independent trial of a random event, but an event that is dependent on the area's given spatial-characteristics. Benchmarking of this random nature is not dissimilar to the Monte Carlo fallacy.

The benchmark used in this study is based on the variable of most significance. We define the variable of most significance to be the largest, absolute standardised regression coefficient of the prior regression analysis, averaged over all eight cities. The size of the beta coefficient yields a sound indication of the given variable's influence on the predicted variable.

4.4.4 Classification Algorithms

There are many types of classification algorithms that could be considered for use in this study. Important considerations to take when choosing classifiers are ones of dimensionality, sample size, linear separability of data, feature independence, speed and performance and overfitting considerations. In statistics, one should generally follow Occam's razor, a heuristic technique that states

that "*Among competing hypotheses, the one with the fewest assumptions should be selected*". In the case of choosing a classifier, we choose algorithms that the needs of the problem scope, and will only pursue more complex solutions where strictly necessary.

Moreover, the results of a single classifier may not be indicative of Airbnb's true penetration. Even if an algorithm seems well-suited to this particular problem, in practice it may fail searching through its hypothesis space. Thus, and as the *Netflix prize* has shown, it is often wise to use an ensemble method, using multiple machine learning classifiers to obtain a more indicative set of results. Next, we present arguments for and against a selection of machine learning classifiers and kernels to make a more informed choice.

Some of the most widely used supervised classifiers are Support Vector Machines (SVM), Logistic Regression, Naive Bayes and Random Forests (tree ensemble). Alternatively, one may adopt a non-supervised or semi-supervised approach using Deep Learning, however, this is not a general purpose solution and adds far more complexity than is strictly necessary. As a baseline, the supervised approaches are more relevant to the current problem scope. For future work, if it is felt supervised methods do not reflect the whole picture, a deep learning approach may be considered.

The relationship between rows and features is an important characteristic to consider. The final dataset (all cities) will contain 2456 inputs, and less than 18 features (cut based on significance). Here, the input count dominates the feature space. Clearly, given the number of training samples and possible number of features that the final dataset will contain, time complexity is not a major concern.

The following subsections provide a brief description of the algorithm and note the contextual use, benefits and flaws of each.

Support Vector Machines

Support Vector Machines (SVM) are a class of supervised algorithms used for regression and classification. SVM's work by maximising the margin around a separating hyperplane and may linearly

or non-linearly separate a dataset. SVM's are time-expensive, however as mentioned, the size of the dataset indicates this is a null-issue. SVM's are extremely versatile as they may utilise different kernel's for the decision function in different context. Common SVM kernels are the **linear kernel**, the **radial basis kernel**, the **polynomial kernel** and the **sigmoid kernel**.

Linear kernels are the most simple kernel function, and may be derived from the logistic regression algorithm. The main reason to use SVM's over logistic regression is because the problem may not be linearly separable (logistic regression can be used non-linearly, however for practical reasons a more accepted strategy is to use SVM's). We do not know whether the problem scope is linearly separable or not, and so should use both a linear and non-linear algorithm. Andrew Ng, associate professor at Stanford and a renowned machine learning researcher, recommends that linear kernels should be used when the number of features is greater than the number of training examples. If not, we may run the risk of overfitting. Thus, we should not make use of the SVM linear kernel.

The radial basis function kernel (RBF) is a Gaussian, non-linear kernel that is based on the belief that a higher dimensional space is more likely to be separable in a lower dimensional space, utilising the *kernel trick*. It works at least as well as the linear kernel if it is tuned properly. Andrew Ng recommends that it should be used when the number of training examples exceeds the number of features, which matches the dimensions of this study's classification dataset. However, what of the polynomial and sigmoid kernel? Currently, the RBF kernel is far more widely used than the polynomial or sigmoid kernel. Under similar training and testing costs, the polynomial kernel may not give higher accuracy than the RBF kernel ¹ (Chang, Hsieh, Chang 2010). The sigmoid kernel can do anything a two layer neural network can do, and may work as well as the RBF kernel, however this is adding further complexity to the problem. We use an RBF kernel as it fits the problem scope and also may test a non-linear SVM solution.

Logistic Regression

The logistic regression classifier works by measuring the relationship between two categorical dependent variables by using a logistic function that estimates probabilities of both classes. Logistic

¹<http://www.jmlr.org/papers/volume11/chang10a/chang10a.pdf>

regression is a special case of linear regression, but is based on differing assumptions; mainly that the data is Bernoulli distributed (i.e. case 1 and case 0) and not Gaussian. It analyses dichotomous (binary) dependents, thus we use multinomial logistic regression, an extension of logistic regression that works on multiple categories. With multinomial logistic we can handle the multiclass (VLP, LP, HP, VHP) problem without manually rerunning the classifier using the one-vs-all method, stated above.

Logistic regression uses many ways to regularise its model, and is less concerned about collinearity, as in Naive Bayes. It is a commonly used and generally well behaved algorithm that fits many problem scopes. Thus, logistic regression is the second addition to the ensemble set over the linear-kernel SVM for its alternative implementation.

Naive Bayes

Naive Bayes applies Bayes' theorem to classify data. It is a relatively simplistic technique that assumes conditional independence. Even if this assumption does not hold, Naive Bayes has been found to perform well in practice. The main advantage of Naive Bayes is that it is a fast and simple algorithm and performs well in many contexts. Its main drawback is that it is unable to learn feature interaction, whereby there may exist some between our chosen features which measure Airbnb penetration. Intuitively, for example, one would expect the median household value to have some collinearity to median household income. However, we use Naive Bayes as the third ensemble classifier as in practice it often works well, in spite of its assumption.

Tree Ensemble

Tree Ensembles are an ensemble method in themselves, such that they construct a multitude of decision trees for the training set and output the mode of the classes for classification. In this context, the mode will correspond to the class of penetration that is has the highest output frequency among the decision trees. One of the main advantages over Logistic Regression is that they do not assume linearity, nor collinearity. Tree ensembles handle both high dimensional spaces and examples such as ours whereby the training count is far higher than the feature space. Further-

more, because they handle interactions well, they handle outliers well. Examples of tree ensemble algorithms include **Random Forests** and **Gradient Boosting**. Gradient Boosted Decision Trees have more hyper-parameters to tune and are more prone to overfitting than Random Forests, thus, for ease of implementation we will add the Random Forest classifier to our ensemble.

The final list of algorithms that will be used are 1) SVM with a Radial Basis Function (RBF) Kernel, Multinomial Logistic Regression, Naive Bayes and Random Forests. The accuracy, precision and recall of each algorithm will be computed for each city and compared to measure their performance and find cross city patterns.

The next section focuses on the results of the regression and classification techniques for analysing Airbnb's spatial penetration.

Chapter 5

Results

This section outlines the findings of methods described in the previous section to answer the research questions posed in section 3.1. Firstly, what are the characteristics of urban areas that also contain a high or low density of Airbnb listings? How does this change for the offer, price and demand of Airbnb listings? Secondly, do these characteristics differ between American cities? If so, how are they different? Finally, can one predict Airbnb’s spatial penetration for unseen data?

5.1 Regression Results

We begin by presenting the results of the regression models. Since the models exist in higher dimensional space, we represent the numeric results in a tabular format. Tables 5.1-5.6 show both the measured p -values and beta (β) coefficients for each variable, and the adjusted R^2 and Moran’s I test values for each model. For each of the three Airbnb variables, there are two tables.

The tables shown below list the findings of the full analysis, containing *all* explanatory variables, and not the outcomes of the stepwise regression, which display a subset of variables. The stepwise results are discussed after the full models, but results of the remaining variables tended to be similar in size to the full results. It is more informative to show the findings for each variable, given that there is little difference between the results for the variables left in stepwise and the same variables

in the full regression. The quality of the model is determined by the given adjusted R^2 value.

For clarity, the p -values are presented as their categorical significance. '***' is associated with terms that have a p -value less than 0.001, and thus very important to the model. '**' is a p -value between 0.001 and 0.01. '*' is in the range of 0.01 and 0.05. '.' between 0.05 and 0.1. Finally, p -value's above 0.1 are presented as null space. Values that are '.' or ' ' indicate that we cannot reject the null hypothesis. Furthermore, the β values are accompanied by green and red bars, representing the size and sign of the coefficient; green bars represent positive coefficients and red bars represent negative coefficients. The most important variables in each model are given by larger absolute beta values.

First and foremost, the results mean little unless they pass the Moran's I test. We find that all models pass. Consequently, we may reject the hypothesis that results are based on spatial autocorrelative factors.

Now, we look at the models given by listings density, followed by price per night and finally review density.

5.1.1 Model Fit for Listings Density

The listing density results are shown in tables 5.1-5.2. It is interesting to note that, despite clear differences between some features across different cities. The findings are indicative of cross-city patterns. Such patterns imply that it will be possible to accurately predict the adoption of Airbnb on a national scale in urban areas. Moreover, the results of adjusted R^2 are consistent and high. They range between 0.66 and 0.81, with a median value of 0.725.

To appreciate the existence of numeric patterns between cities, we compute the absolute sum and median of beta (β) values, the median number of 'stars' for each independent variable and sum of 'signs', which is then made absolute. These metrics will help identify the variables of greatest importance prior to conducting a predictive analysis.

Interestingly, the results are not dissimilar of the findings of Hughes’ [18] and Capra et al. [19] in London. Firstly, distance from its center has a strong, negative relationship with the density of Airbnbs in that area across all cities in our dataset. That is, the further from the location of the center, the fewer the number of Airbnbs. Only in Oakland is distance from center not considered one of the most important variables, at least by measurement of its p -value (two stars). Additionally, the attractiveness of an area, characterised by the density of points of interest (PoI) is positively correlated with listings. This also mirrors the findings of Airbnb listings in London and is indicative that Airbnbs existing in more touristic areas.

The *bohemian index* and *talent index* both exhibit positive correlation against the listings density. The bohemian index shows positive correlation across the board, and the talent index is only negative in Austin, Texas. This finding is in agreement with Florida’s research [16] from 2002. Perhaps the idiom that a creative class attracts early technology adoption still holds in 2016. This line of thinking is further substantiated by the betas of the number of young persons, which similarly follows a cross-city pattern of positive correlation. Perhaps related to results exhibited by proportion of young people, the median income of an area is inversely correlated. We hypothesize that this is because young people are generally less well off than later generations.

Finally, and despite playing a lesser role in each model, the median household value is positively correlated across the board. This represents an interesting phenomena, since the median income is generally inversely correlated. A possible explanation is that young persons are renting out rooms in houses they do not own. Findings in London back up this theory, whereby the writers hypothesise that hosts are young persons who do not own a house. Here, the proportion of owner occupied residences is negatively correlated in six of the eight cities. However, the generally small beta values are not telling of the true story, thus we take this finding with a grain of salt.

Given the results, we can hypothesise that the socio-demographics of areas featuring high levels of Airbnb are young, talented and creative people. Since the listings density across all cities is explained well by each model, the next step is to take the most important variables and use them in a predictive model. The variables that we will use in the predictive model are as follows:

- Bohemian Index

- Distance to Center
- Median Household Value
- PoI Density
- Talent Index
- Proportion of Young Persons

Before then, however, we look at the results of the model fit for the median price and review density of Airbnb listings.

Model Fit for Airbnb Listings Density

Tables 4 and 5 show the least squares multiple linear regression results for all eight cities.














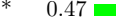













































	Austin		Los Angeles		Manhattan		New Orleans	
Indip. var	<i>p</i> -val	β	<i>p</i> -val	β	<i>p</i> -val	β	<i>p</i> -val	β
Hotel Count	*	-0.08 		-0.02 		-0.06 		-0.05 
POI Count		0.03 	.	0.03 	***	0.21 	***	0.26 
Bus Stop Count	*	0.12 	***	-0.07 		-0.04 		0.04 
Talent Index		-0.13 	***	0.47 		0.22 	**	0.29 
Income Diversity Index		-0.09 	*	0.05 	**	0.22 	*	0.17 
Bohemian Index	***	0.17 	***	0.29 	***	0.47 	***	0.23 
Race Diversity Index	***	-0.16 		0.01 	**	-0.15 	.	0.10 
Unemployment Ratio	.	0.07 		0.03 		0.00		0.05 
Poverty by Income		-0.04 	*	0.09 		0.04 	*	0.20 
Young Persons Prop		0.11 	***	0.11 	**	0.20 		0.04 
Owner Occupied		0.02 	***	-0.27 		-0.03 		-0.08 
Median Income		-0.11 	*	-0.08 	***	-0.35 	*	-0.25 
Median Value	*	0.07 	*	0.05 	.	0.11 	.	0.14 
Median Age		0.26 		-0.03 	***	-0.19 		0.04 
Distance to Center	***	-0.72 	***	-0.33 	***	-0.25 	***	-0.26 
<i>Adjusted R-squared</i>		0.81		0.71		0.66		0.70
<i>Moran's Test</i>		0.06		0.05		0.03		0.04

Table 5.1: Regression Analysis - Listing Density I




























































	Oakland		San Diego		San Francisco		Seattle	
Indip. var	<i>p</i> -val	β	<i>p</i> -val	β	<i>p</i> -val	β	<i>p</i> -val	β
Hotel Count	***	-0.21 		-0.04 		0.02 	***	0.19 
POI Count	**	0.05 		0.03 	***	0.22 	***	0.29 
Bus Stop Count		0.20 	***	0.18 	*	0.15 		-0.04 
Talent Index	**	0.49 	***	0.41 		0.14 		0.02 
Income Diversity Index		0.05 	*	0.09 		0.07 		0.05 
Bohemian Index	**	0.22 		0.04 	***	0.27 	.	0.09 
Race Diversity Index		0.04 	***	-0.16 	*	-0.14 		-0.07 
Unemployment Ratio		0.08 	**	0.13 	*	-0.12 		-0.06 
Poverty by Income		-0.10 	**	0.17 		0.04 		0.05 
Young Persons Prop	*	0.20 	***	0.38 		0.07 	*	0.26 
Owner Occupied		-0.03 	*	-0.09 	*	0.21 		-0.06 
Median Income	*	-0.32 	*	-0.17 	*	-0.22 		0.06 
Median Value		0.03 	*	0.12 		0.04 		0.00
Median Age		0.14 	***	-0.23 	*	-0.17 		0.03 
Distance to Center	**	-0.26 	***	-0.33 	***	-0.33 	***	-0.49 
<i>Adjusted R-squared</i>		0.79		0.76		0.66		0.74
<i>Moran's Test</i>		0.04		0.03		0.06		0.02

Table 5.2: Regression Analysis - Listing Density II

5.1.2 Model Fit for Median Price

The foremost observation is that models across the board are far less confident in their findings than explaining Airbnb adoption by listing density. Whereas the median coefficient of determination (adjusted R^2) for listing density was 0.725, the median for price was 0.47. There is a clear imbalance in the range of results; they lie between 0.2 and 0.58. This suggests that the asking prices for Airbnb's are difficult to explain with the given choice of variables, and thereby will be more difficult to predict.

Notably, we discover two variables that tend to maintain their script for most models. They are the talent index and the distance from city center. The talent index was the most important characteristic in determining price per night. In all instances (cities) this was positively correlated, and often the most significant variable in the whole model. In Oakland, for example, the talent index stood at 0.73, over twice as large as the coefficient of the next highest variable. Hughes' [18] results are similar in London, and suggests that this is indicative of higher prices existing in more gentrified areas of the city. This is a tempting argument to follow, given that the bohemian index is, six times out of eight, positively correlated with price. However, we must approach this suggestion with caution, given that the overall quality of each model was often lacking.

The distance from city center was, in all cases but one, significantly anti-correlated with price. This is intuitive, as properties towards the suburbs are naturally more inexpensive than the more popular areas in the center of the city.

Surprisingly, the price per night was rarely influenced by either median household income nor household value. This is an unexpected result, as intuitively, a household worth more would charge higher rents. Unlike London, where areas have high disparity between property values, metropolitan areas in America are considered economically and demographically segregated. One would expect, given a narrower range of property values, that listing prices would naturally correlate.

Given an overall lack of statistical significance, and general poor performance of the regression model on median price per night, we will not attempt a predictive analysis on this variable. Please refer to tables 6 and 7 to see the results of this section.

Model Fit for Airbnb Price per Night

Tables 6 and 7 show the least squares multiple linear regression results for all eight cities.

	Austin		Los Angeles		Manhattan		New Orleans	
Indip. var	<i>p</i> -val	β	<i>p</i> -val	β	<i>p</i> -val	β	<i>p</i> -val	β
Hotel Count		0.00	*	0.06	***	0.15		-0.03
POI Count	***	-0.32		-0.01		0.06		-0.02
Bus Stop Count	*	0.17	*	-0.06		0.03		0.01
Talent Index		0.22	***	0.49	***	0.41	***	0.58
Income Diversity Index		0.01		0.00		0.09	*	-0.24
Bohemian Index	*	0.18	*	0.10	***	-0.20	**	0.18
Race Diversity Index	.	-0.13		-0.01	***	-0.23		0.07
Unemployment Ratio		0.01		0.02		0.00		0.01
Poverty by Income	.	0.18		0.05		-0.01		-0.16
Young Persons Prop		-0.10		-0.02	.	-0.13		-0.07
Owner Occupied		-0.04	***	-0.18	.	0.11		-0.02
Median Income		0.08	**	-0.12		0.07	**	-0.43
Median Value		0.06		0.04		-0.05	*	0.22
Median Age		-0.01	***	0.18		-0.07		0.07
Distance to Center	***	-0.43	***	-0.21	***	-0.42	***	-0.33
<i>Adjusted R-squared</i>		0.37		0.45		0.58		0.49
<i>Moran's Test</i>		0.02		0.03		0.01		0.02

Table 5.3: Regression Analysis - Median Price I

	Oakland		San Diego		San Francisco		Seattle	
Indip. var	<i>p</i> -val	β	<i>p</i> -val	β	<i>p</i> -val	β	<i>p</i> -val	β
Hotel Count		-0.10		-0.06		0.05		0.04
POI Count		-0.09	*	-0.18		-0.10	.	-0.23
Bus Stop Count		-0.07		-0.05		0.12		0.01
Talent Index	***	0.73	***	0.42		0.09	**	0.47
Income Diversity Index		-0.06	**	0.26		0.07		-0.12
Bohemian Index	*	0.20		0.03		0.09	.	-0.14
Race Diversity Index		0.12	*	0.16	.	-0.14		0.03
Unemployment Ratio		0.12		-0.01		-0.02		0.07
Poverty by Income		0.11		0.15	***	0.33		0.06
Young Persons Prop		-0.17	.	0.16		0.11	*	-0.36
Owner Occupied		0.05	*	0.20		0.03		0.08
Median Income		-0.05	*	-0.28	**	0.39		-0.09
Median Value		-0.07		-0.10		0.08		-0.20
Median Age		0.02		0.05	*	0.18		-0.04
Distance to Center	*	-0.35	*	-0.18		-0.12	**	0.38
<i>Adjusted R-squared</i>		0.59		0.20		0.50		0.40
<i>Moran's Test</i>		0.03		0.03		0.00		0.04

Table 5.4: Regression Analysis - Median Price II

5.1.3 Model Fit for Review Density

Finally, we look at the results for review density. The first observation is that they seem to follow the findings of listings density. Though similar in substance, the significance of the results is on average 0.11 less than Airbnb's spatial density, with a median adjusted R^2 of 0.61.

The most significant variables in both reviews and listings density were extremely similar. That is, the distance to center and medium household income are negatively correlated with the density of reviews, and the talent index, bohemian index and young person proportion also follow a positive correlation.

This relationship is better understood when it is noted that, on average, 80% of Airbnb guests leave reviews. Intuitively, if most guests who lodge with Airbnb leave reviews, then the relationships should scale accordingly, with only a 20% margin of results unaccounted for.

Since the findings are so similar, we will use the listings density as the only Airbnb metric to use in a predictive model. It produced the most telling, and significant set of discoveries. With such strong regressor results, we also expect sound predictive results.

Model Fit for Airbnb Review Density

Tables 8 and 9 show the least squares multiple linear regression results for all eight cities.

	Austin		Los Angeles		Manhattan		New Orleans	
Indip. var	<i>p</i> -val	β	<i>p</i> -val	β	<i>p</i> -val	β	<i>p</i> -val	β
Hotel Count		-0.03 ↓		-0.00		-0.00		-0.05 ↓
POI Count		-0.00		0.04 ↓	***	0.28 █		0.13 █
Bus Stop Count	.	0.10 █	***	-0.08 ↓		-0.03 ↓		0.01 ↓
Talent Index		-0.08 ↓	***	0.46 █		-0.10 ↓	**	0.31 █
Income Diversity Index		-0.05 ↓	**	0.07 ↓	***	0.25 █		-0.24 █
Bohemian Index	***	0.19 █	***	0.28 █	***	0.37 █	***	0.27 █
Race Diversity Index	**	-0.14 ↓		0.01 ↓	*	-0.15 ↓	.	0.12 ↓
Unemployment Ratio	.	0.08 ↓		0.03 ↓		0.00		0.01 ↓
Poverty by Income		-0.03 ↓	*	0.10 ↓		0.07 ↓	.	0.19 █
Young Persons Prop		0.06 ↓		0.04 ↓	***	0.25 █		0.06 ↓
Owner Occupied		0.08 ↓	***	-0.22 ↓		0.03 ↓		-0.02 ↓
Median Income		-0.11 ↓	*	-0.09 ↓	***	-0.36 █	***	-0.60 █
Median Value		0.07 ↓	.	0.05 ↓		0.07 ↓	.	0.18 █
Median Age		0.03 ↓		-0.01 ↓	*	-0.13 ↓	*	0.12 ↓
Distance to Center	***	-0.71 █	***	-0.36 █	***	-0.22 ↓	***	-0.29 █
<i>Adjusted R-squared</i>		0.75		0.62		0.60		0.49
<i>Moran's Test</i>		0.03		0.05		0.04		0.01

Table 5.5: Regression Analysis - Review Density I

	Oakland		San Diego		San Francisco		Seattle	
Indip. var	<i>p</i> -val	β	<i>p</i> -val	β	<i>p</i> -val	β	<i>p</i> -val	β
Hotel Count	*	-0.16 ↓		-0.03 ↓		-0.05 ↓	.	-0.12 ↓
POI Count		-0.01 ↓		0.02 ↓	**	0.19 █	**	0.33 █
Bus Stop Count	.	0.14 █	*	0.13 █	**	0.20 █		-0.04 ↓
Talent Index	*	0.48 █	***	0.36 █		0.11 ↓	*	-0.32 █
Income Diversity Index		0.10 ↓	*	0.13 █		0.02 ↓		0.02 ↓
Bohemian Index	*	0.18 █	.	0.08 ↓	***	0.30 █	*	0.14 █
Race Diversity Index		-0.05 ↓	**	-0.14 ↓	**	-0.19 ↓		-0.05 ↓
Unemployment Ratio		-0.01 ↓	**	0.14 █	*	-0.15 ↓		-0.10 ↓
Poverty by Income		-0.12 ↓	*	0.13 █		0.05 ↓		-0.07 ↓
Young Persons Prop	.	0.19 █	***	0.33 █		0.09 ↓	*	0.35 █
Owner Occupied		-0.10 ↓		-0.02 ↓	**	0.28 █		-0.05 ↓
Median Income	*	-0.40 █	*	-0.22 ↓	*	-0.24 ↓		0.16 ↓
Median Value		0.11 ↓	.	0.12 ↓		0.06 ↓		0.10 ↓
Median Age		0.28 █	***	-0.19 ↓	.	-0.14 ↓		0.08 ↓
Distance to Center		-0.16 ↓	***	-0.39 █	**	-0.28 ↓	***	-0.48 █
<i>Adjusted R-squared</i>		0.66		0.65		0.59		0.60
<i>Moran's Test</i>		0.01		0.03		0.05		0.01

Table 5.6: Regression Analysis - Review Density II

5.2 Prediction Results

In this section, we present and attempt to reason the results for the prediction analysis for Airbnb adoption. This section is split into two parts, firstly, we run our prediction over all four ordinal categories for listing density: *VLP*, *LP*, *HP*, *VHP*, capturing the entire dataset. Next, we predict only the extreme ends of Airbnb penetration (*VLP* and *VHP*), as this offers the most cogent suggestion that Airbnb’s urban-spatiality may be predicted. For both classifiers, we compare the results to a benchmark.

Tables 5.7 and 5.8 show the accuracy, precision and recall results for 1) Support Vector Machine with radial basis kernel ((**SVM**), 2) Random Forests (**RF**), 3) Multinomial Logistic Regression (**MLR**) and 4) Naive Bayes (**NB**). "**F**.<performance measure>" denotes the results run on all variables. "**B**.<performance measure>" denotes the results run only on the benchmark. For both classifiers, the benchmark is the algorithm run only on the variable *distance to center*.

The tables were compiled as follows. First, for each city, the accuracy, precision and recall were calculated for each category, for each algorithm using the one-vs-all method as described in section 4.4.2. Then, the results were averaged, yielding a single valued result for each algorithm relative to the model (i.e. full model or benchmark).

5.2.1 Quartile Classification

The results, as given in Table 5.7, are depictive of a model that serves as a sound predictor of Airbnb’s spatial penetration, at least relative to the benchmark.

One of the research questions we ask is whether the results are indicative of the model’s ability to generalise across cities. Given the wide ranging socio-demographic spectrum of cities that make up the dataset, the consistency of results over each eight folds is suggestive that the model will predict unseen cities at a similar rate to our findings. By taking a slight leap of faith, we may infer that the model will stratify quartile-classes correctly 50% of the time in cities where Airbnb has an established presence.

In comparison to a random benchmark, our model outperforms it by around 100% for all four algorithms, whereby a random benchmark would perform at 25%. Our benchmark yields an accuracy ranging between 0.25 and 0.39, which in all cases performs well under our full model. The precision and recall estimates (metrics denoting how useful and complete the findings are) have similar values to accuracy. The full model’s performance surpasses the benchmark by analogous margins for both precision and recall. For further information showing the consistency of performance of the algorithms for this dataset, please find the attached, more descriptive inter-city results in the appendices section.

VLP and VHP were, unsurprisingly, the categories that were the most accurate, precise and had the best recall. Presumably, this is because the correlative disparity is greatest, and thus easiest to predict, at the tails of the prediction distribution. Since the labels are ordinal by nature, VLP and VHP ‘sit’ next to only one other class, whereas LP and HP are sandwiched between two classes, and are thus more likely to be confused.

	F.Accuracy	B.Accuracy	F.Precision	B.Precision	F.Recall	B.Recall
SVM	0.45	0.25	0.47	0.30	0.45	0.25
RF	0.49	0.39	0.50	0.43	0.48	0.39
MLR	0.50	0.35	0.50	0.36	0.49	0.35
NB	0.48	0.34	0.50	0.36	0.48	0.34

Table 5.7: Quartile Classification Results Summary

5.2.2 Extreme Classification

Despite the four-category models classifying well above the benchmarks in all instances, the results only predict around 50% of the time. For regulators and policy makers, the findings do not provide strong enough evidence that one may predict the Airbnb adoption in enough situations to create policy based on the findings. They are more interested in creating policies that help capture and predict the ‘extremes’ of Airbnb adoption. Thus we turn to analysing the predictive results of the floor and ceiling of listings density. Here, we are only interested in predicting areas of very high or very low penetration.

The findings show far improved results upon the previous classification attempt. The average accuracy of prediction is around 90%, meaning that on new, unseen city data, the classifier will correctly forecast results nine times out of ten. As there are only two categories, a random benchmark here would work at 50%. Our computed benchmark, though generally quite high, is less consistent and in all cases noticeably smaller than our findings.

Of all cities, Oakland was predicted correctly the most. In fact, three of the four algorithms predicted Oakland’s accuracy at over 98%. The city that was most difficult to predict was Manhattan, where the algorithm with the highest accuracy performed at 84%. This is perhaps to be expected, given that Manhattan is the most densely populated urban area in America¹. We conjecture that, due to high population concentrations, the Manhattan tract captures a multitude of diverse demographic characteristics that cancel out expected patterns. Alternatively, it may simply be the case that the variable with greatest significance in other models, distance to center, has a relatively low influence in Manhattan. Indeed, Manhattan’s benchmark is indicative of this.

We look back at our final research question, which asked how possible it would be to predict Airbnb penetration in unseen cities. Over all cities chosen in this study, the result is unanimous; Airbnb’s spatial penetration may be predicted to impressive accuracy in large, urbanised, American areas where the company’s presence is mature.

	F.Accuracy	B.Accuracy	F.Precision	B.Precision	F.Recall	B.Recall
SVM	0.88	0.54	0.89	0.60	0.88	0.73
RF	0.90	0.81	0.91	0.82	0.90	0.86
MLR	0.90	0.73	0.91	0.79	0.90	0.83
NB	0.86	0.68	0.87	0.77	0.86	0.80

Table 5.8: Outer Quartile Classification Results Summary

¹www.en.wikipedia.org/wiki/List_of_United_States_cities_by_population_density

Chapter 6

Conclusion

6.1 Summary

This study provides quantifiable evidence that Airbnb’s spatial penetration may be determined by a discrete set of social and geographic area characteristics. Through regression and classification techniques, we have compiled findings and established approaches that will enable future policy makers to design more relevant regulations and track Airbnb’s continued growth. This is the first time Airbnb’s urban adoption has been measured on a national scale, and proved that such patterns exist between listings.

Considering only the listing density of Airbnb in American cities, we have determined that the most statistically significant variables across the U.S. exhibited a mixture of geographic and demographic categories. Airbnb’s adoption is most strongly correlated with areas that have higher proportions of young, talented and creative individuals (a ‘creative class’) residing in more touristic and gentrified areas. Interestingly, the results of this study are somewhat synonymous to outcomes of the geo-spatial study in London [18, 19]. The overlap between Florida and Clifton’s social research in the UK and U.S., respectively, provides an explanation [16, 17]. We observe that ‘creative class’ of both nations is immensely similar, and hypothesize that it is the same groups of people who are more prone to adopting new technologies, such as Airbnb.

Unfortunately, the price may not so easily be predicted. However, despite low confidence scores in some cities, there are clear patterns found through regression. The variables talent index and distance from city center maintain obvious, cross city patterns that offer a clear insight between listing price and social dynamics and geographic characteristics, indicative of higher prices existing in more gentrified areas. The case for this argument is compounded by the constant positive correlation between bohemian index and price. Future work may use these findings to develop predictive pricing models.

To yield more contextual and useful interpretation of the results, we must first rigorously evaluate our process and findings.

6.2 Critical Evaluation

Despite capturing relationships that hold across all cities used in the study, the approach was not without its limitations. First and foremost, multiple linear regression cannot determine casual relationships among the independent variables in the model. Thus, while we can justifiably argue that the variables used in the classification step *predict* Airbnb’s listing density, we may not make the case that they *cause* it. One may use Principal Components Analysis (PCA) to attempt to isolate possibly correlated variables into a subset of linearly uncorrelated ‘principal components’. This technique would remove collinearity between variables and allow us to argue causation, but only in combination with a temporal analysis, taken place over some years. Given the time-scope of this project, such a process would have been impractical.

Secondly, our choice of cities was limited by the data source used to compile Airbnb information¹. Though our city choice reflects a wide-ranging socio-demographic subset that captures much of American culture, it is certainly not ideal. Four of the eight cities reside in California, and one must question their relative cultural variation. Of the sources available at the time of collection, the chosen cities were the right choice, however were the study to be repeated, we would recommend adding Washington D.C., Chicago and Nashville to the dataset.

¹insideairbnb.com

As a minor note, the inorganic shapes of many U.S. cities make it more difficult to conduct spatial-autocorrelation tests. As it is only a small proportion of tracts in each city that skew the data, we expect this had little impact on the final Moran's test value.

Finally, we must consider the data used in the study. Despite using the most recent data for the U.S. census, it is currently almost six years old. Thus, much of the work is based on the assumption that our dataset does not vary much from the reality of current demographics in the selected cities. Having considered this at the beginning of the project, we concluded that our assumption was justified. Of course, there will be differences in data, but socio-demographic shifts are often a slow, drawn out process and not a sudden disruption.

6.3 Implications

This project has implications that are both theoretical and practical. The theoretical overtone is rooted in urban studies; our findings suggest groups of people in close proximity that display a common set of characteristics have similar tendencies that impact the adoption of Airbnb. Our findings provide evidence that the 'creative class' is still very much alive. As the results give evidence of nationwide patterns in urban areas, the practical implications are that legislation may be designed for a national scale by regulators and policy makers. This marks the first study that confirms the existence of cross city patterns and gives strong foundation upon which to build new regulations for the sharing economy.

6.4 Future Work

The obvious next step would be to conduct the same analysis with temporal constraints to unravel the cause-effect relationship between Airbnb and urban characteristics. A study that of this vain would offer serious insight into Airbnb's expansion and how the relationship between Airbnb adoption and area characteristics have developed over time. As an extension of this, we ask what the impact of major events would have on Airbnb's demand. For example, how would Airbnb adoption change with the annual SXSW festival in Austin, Texas? Other future work may focus on host characteristics and measure how they compare with the our findings. To better understand the

cause-effect relationship in this context, one must track the growth of Airbnb in cities in which adoption is not yet matured. Since the U.S. market for Airbnb is in its maturity stages, we ask how adoption would manifest itself in cities outside of the U.S. where Airbnb is in its introductory stage in the product lifecycle.

This study uses Airbnb as a case study in point for the sharing economy. It would be intriguing as to whether it is the same group of people who are early adopters for other platforms in peer-to-peer market places such as Uber or TaskRabbit.

Bibliography

- [1] Boyd Cohen and Jan Kietzmann. Ride on! mobility business models for the sharing economy, September 2014.
- [2] New York State Attorney General Eric T. Schneiderman. Airbnb in the city, 2014.
- [3] UK Office for National Statistics. Travel trends 2015, 20 May, 2015.
- [4] PwC. Growth is in the air: But it’s coming down to earth. *UK hotels forecast 2016*, 2016.
- [5] Jeremy Rifkin. The zero marginal cost society: The internet of things, the collaborative, 2014.
- [6] J. Yannis Bakos. Reducing buyer search costs: Implications for the electronic marketplace. *Management Science*, Volume 43(Issue 12), December 1997.
- [7] Aliza Fleischer Eyal Ert and Nathan Magen. Trust and reputation in the sharing economy: The role of personal photos on airbnb, January 2016.
- [8] John W. Byers Boston University Georgios Zervas, Davide Proserpio. A first look at the online reputation of airbnb, where every stay is above average, 2015.
- [9] Benjamin Edelman and Michael Luca. Digital discrimination:the case of airbnb.com, January 10, 2014.
- [10] The Economist. The rise of the sharing economy: Peer to peer rental, Mar 9th, 2013.
- [11] Johanna Interian. Up in the air: Harmonising the sharing economy through airbnb regulations, 2016.
- [12] Brad Turtle. Can we stop pretending the sharing economy is all about sharing?, 2014.

- [13] Arvind Malhotra and Marshall Van Alstyne. The dark side of the sharing economy and how to lighten it. *Communications of the ACM*, Volume 57(Issue 11):Pages 24–27, November 2014.
- [14] Davide Proserpio Georgios Zervas and John W. Byers. The rise of the sharing economy: Estimating the impact of airbnb on the hotel industry, 2013 (updated 2016).
- [15] Richard Florida. Bohemia and economic geography, 1993.
- [16] Richard Florida. The rise of the creative class, 2002.
- [17] Nick Clifton. The creative class in the uk: An initial analysis. *Geografiska Annaler: Series B, Human Geography*, Volume 90(Issue 1):Pages 63–82, March 2008.
- [18] Eleanor Hughes. Understanding the geography of airbnb. Master’s thesis, UCL Department of Computer Science, 2015.
- [19] Davide Prosperio Daniele Quercia Giovanni Quattrone, Licia Capra and Mirco Musolesi. Who benefits from airbnb: Regulating the sharing economy, 2016.
- [20] C. E. Gehlke and Katherine Biehl. Certain effects of grouping upon the size of the correlation coefficient in census tract material. *Journal of the American Statistical Association*, Volume 29:Pages 169–170, March 1934.
- [21] Mark Melnik Alvaro Lima. Boston: Measuring diversity in a changing city, 2013.
- [22] Philip Meyer and Shawn McIntosh. The usa today index of ethnic diversity. *International Journal of Public Opinion Research*, 4(1):51–58, 1992.
- [23] L. Jost. Entropy and diversity. *Ecology & Organismal Biology*, 113(2):363–375, May 2006.
- [24] Richard Florida. Economic geography of talent, 2002.
- [25] Michael Zweig. What’s class got to do with it, american society in the twenty-first century, 2004.
- [26] Joe M. Bohlen and George M. Beal. The diffusion process. *Agriculture Extension Service, Iowa State College*, Special Report(Number 18):Page 56, May 1957.

- [27] Kang. Michael Batty. Camille, Roth. Soong Moon and Barthelemy Marc. Structure of urban movements: Polycentric activity and entangled hierarchical flows, 2010.
- [28] Theodore Micceri. The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105(1):156–166, 1989.

Appendix A

Data Collection

Note that all code is attached in a Google Drive folder. However, we list it here for completeness in the document. Please follow the following link for all datasets, scripts and relevant material.

<https://drive.google.com/folderview?id=0BwcQZklv7An-bGthSWdCWmRXSGs&usp=sharing>

Bus Stops and PoI's

To calculate the number of bus stops and points of interest for all tracts, given the raw latitude and longitude coordinate pairs of all data points, we undertake the following two step process. First for each data point, we convert it to its respective tract identification number. Next, we count the number of tracts that belong to each unique tract identification number. The process for collecting bus stops changed from city to city, as in some instances OpenStreetMap did not have enough data, in these cases we had to extract latitude longitude pairs from the addresses of each bus stop, before converting them to tract identification numbers, this is the same procedure that took place in collating hotel data. The following two code snippets highlight the former approach. Code for the latter part of the code may be found in the attached Google Drive folder. The *city* name must match the file that exists in the same folder which contains the raw coordinate pairs of the given data points.


```

1 import requests
2 import urllib2
3 import untangle
4 import csv
5
6 city = 'LosAngeles'
7 stopsReader = city + '_stops.csv'
8 csvWriter = city + 'busstopsTract.csv'
9
10 # Extract latitude and longitude for the raw CSV file
11 def returnLatLonList():
12     lat_lon_list = []
13     with open(stopsReader, 'rU') as csvfile:
14         csvreader = csv.DictReader(csvfile)
15         for row in csvreader:
16             # position_pair = row['Location'] # Latitude-Longitude pair of coordinates
17             latitude = row['stop_lat']
18             longitude = row['stop_lon']
19             dict = {'Latitude':latitude, 'Longitude':longitude}
20             lat_lon_list.append(dict)
21     return lat_lon_list
22
23
24 # Function to return the block associated with the given Latitude-Longitude
    coordinates
25 def returnBlock(latitude, longitude):
26     url = 'http://data.fcc.gov/api/block/2010/find?latitude=' + str(latitude) + '&
        longitude=' + str(longitude)
27     try:
28         file = urllib2.urlopen(url)
29     except urllib2.HTTPError, e:
30         logging.error('HTTPError = ' + str(e.code))
31     except urllib2.URLError, e:
32         logging.error('URLError = ' + str(e.reason))
33     except httpplib.HTTPException, e:
34         logging.error('HTTPException')
35     except Exception:
36         import traceback
37     data = file.read()

```

```

38     file.close()
39     doc = untangle.parse(data)
40     block = doc.Response.Block['FIPS']
41     return block
42
43
44 # Write out the new data to a CSV
45 def writeCsv():
46     busstop_list = returnLatLonList()
47     with open(csvWriter, 'w') as csvfile:
48         fieldnames = ['Tract']
49         writer = csv.DictWriter(csvfile, fieldnames=fieldnames)
50         for cs in busstop_list:
51             lat = cs['Latitude']
52             lon = cs['Longitude']
53             tract = returnBlock(lat, lon)
54             writer.writerow({'Tract': tract})
55
56
57 # Main function
58 if __name__ == '__main__':
59     writeCsv()

```

Listing A.1: Code to convert raw **Bus Stop** and **Point of Interest** data to tract ID

The next code snippet counts the number of data points that exist in each unique GEOID.

```

1 import csv
2
3 city = 'LosAngeles'
4 cultureFile = city + 'busstopsTract.csv'
5 hotelcountwriter = city + 'BusstopCount.csv'
6 geoid_list_csv = '../' + city + 'GEOIDs.csv'
7
8 # List the Tract ID's (GEOID's)
9 def readGeoIDCsv():
10     geoid_list = []
11     with open(geoid_list_csv, 'rU') as csvfile:
12         csvreader = csv.DictReader(csvfile)
13         for row in csvreader:

```

```

14     geoid = row['GEOID']
15     geoid_list.append(geoid)
16     return geoid_list
17
18 # Read CSV and count unique ID's
19 def readCsv():
20     geoid_list = readGeoIDCsv()
21     geoid_list_count = []
22     with open(cultureFile, 'rU') as csvfile:
23         csvreader = csv.DictReader(csvfile)
24         for row in csvreader:
25             geoid = row['GEOID']
26             geoid = geoid[:-4]
27             print geoid
28             if geoid in geoid_list:
29                 count = 1
30                 if len(geoid_list_count)>0:
31                     match = False
32                     for cs in geoid_list_count:
33                         if (cs['geoid'] == geoid):
34                             cs['count'] = cs['count'] + 1
35                             match = True
36                     if match == False:
37                         dict = {"geoid":geoid, "count":1}
38                         geoid_list_count.append(dict)
39             else:
40                 dict = {"geoid":geoid, "count":1}
41                 geoid_list_count.append(dict)
42     return geoid_list_count
43
44 # Write out the new data to a csv
45 def writeCsv():
46     geoid_list_count = readCultureCsv()
47     with open(hotelcountwriter, 'w') as csvfile:
48         fieldnames = ['GEOID', 'Count']
49         writer = csv.DictWriter(csvfile, fieldnames=fieldnames)
50         for cs in geoid_list_count:
51             geoid = cs['geoid']
52             count = cs['count']

```

```

53     writer.writerow({ 'GEOID': geoid , 'Count': count })
54
55
56 # Main function
57 if __name__ == '__main__':
58     writeCsv()

```

Listing A.2: Counting the number of **Bus Stop** and **Point of Interest** tract ID's/GEOID's data to tract ID

Distance to Center

The script below calculates the distance walked between two points and outputs it to a CSV file, which is then appended to the dataset for the city in question. We take the average of the distance there and the distance back.

```

1  # Los Angeles Center: 34.053405, -118.242768
2
3  import csv
4  import urllib2 , json
5  import untangle
6
7  city = 'LosAngeles'
8  filecsv = city + 'Tracts.csv'
9  center_writer = city + 'DistanceToCenter.csv'
10
11 # Read Coordinates of each tract in city
12 def readCsv():
13     coordinates_list = []
14     with open(filecsv , 'rU') as csvfile:
15         csvreader = csv.DictReader(csvfile)
16         for row in csvreader:
17             geoid = row['GEOID']
18             latitude = row['LAT']
19             longitude = row['LON']
20             dict = { 'GEOID': geoid , 'Lat': latitude , 'Lng': longitude }
21             coordinates_list.append(dict)
22     return coordinates_list

```

```

23
24 # Compute distance between center and tract (manually pass in coords)
25 def distanceBetweenCoordinates(coordinatesFrom, coordinatesTo):
26     url = 'http://router.project-osrm.org/table?loc=' + str(coordinatesFrom[0]) + ',' +
        + str(coordinatesFrom[1]) + '&loc=' + str(coordinatesTo[0]) + ',' + str(
            coordinatesTo[1])
27     response = urllib2.urlopen(url)
28     data = json.loads(response.read())
29     # Take average distance as they differ from starting location
30     distanceTo1 = data['distance_table'][0][1]
31     distanceTo2 = data['distance_table'][1][0]
32     averageDistance = (distanceTo1+distanceTo2)/2
33     return averageDistance
34
35 # Write distance to csv file
36 def writeCsv():
37     coordinates_list = readCsv()
38     with open(center_writer, 'w') as csvfile:
39         fieldnames = ['GEOID', 'Latitude', 'Longitude', 'DistanceToCenter']
40         writer = csv.DictWriter(csvfile, fieldnames=fieldnames)
41         writer.writeheader()
42         for cs in coordinates_list:
43             distance = distanceBetweenCoordinates([cs['Lat'], cs['Lng']], [34.053405,
                -118.242768])
44             writer.writerow({'GEOID': cs['GEOID'], 'Latitude': cs['Lat'], 'Longitude': cs[
                'Lng'], 'DistanceToCenter': distance})
45             print 'GEOID: ' + cs['GEOID']
46             print 'Lat: ' + str(cs['Lat'])
47             print 'Lng: ' + str(cs['Lng'])
48             print 'Distance: ' + str(distance)
49
50 if __name__ == '__main__':
51     writeCsv()

```

Listing A.3: Code to calculate the distance between two pairs of coordinates

Hotels

The following code uses the Google Maps API to convert a given address to coordinates and

then to the associated U.S. tract ID. The hotel addresses are scraped using the Chrome extension *kiminolabs*.

```
1 import requests
2 import urllib2
3 import untangle
4 import csv
5 import json
6 from googlemaps import Client
7
8 city = 'LosAngeles'
9 hotelfile = city + 'HotelsRaw.csv'
10 api_key = "AIzaSyD0qUuCPPEJ2Q5Qn3h0OfPMzsO9zb0Pcdo"
11 hotelWriter = city + 'Hotels.csv'
12
13 # Read in addresses of the hotels
14 def readCsv():
15     hotel_address_list = []
16     with open(hotelfile, 'rU') as csvfile:
17         csvreader = csv.DictReader(csvfile)
18         for row in csvreader:
19             address = row['property1']
20             dict = {'Address': address}
21             hotel_address_list.append(dict)
22     return hotel_address_list
23
24 # Convert Addresses to coordinates
25 def returnLatLon():
26     hotel_address_list = readCsv()
27     lat_lon_list = []
28     hotelCount = 0
29     for address in hotel_address_list:
30         address['Address'] = address['Address'] + ', ' + 'Los Angeles'
31         print address['Address']
32
33     api_response = requests.get('https://maps.googleapis.com/maps/api/geocode/json?
34     address={0}&key={1}'.format(address, api_key))
35     api_response_dict = api_response.json()
36     try:
```

```

36     if api_response_dict['status'] == 'OK':
37         hotelCount+=1
38         latitude = api_response_dict['results'][0]['geometry']['location']['lat']
39         longitude = api_response_dict['results'][0]['geometry']['location']['lng']
40         dict = {'Latitude':latitude, 'Longitude':longitude}
41         lat_lon_list.append(dict)
42     except Exception as e:
43         print str(e)
44         pass
45     return lat_lon_list
46
47 # Return the block associated with the given Latitude-Longitude coordinates
48 def returnBlock(latitude, longitude):
49     url = 'http://data.fcc.gov/api/block/2010/find?latitude=' + str(latitude) + '&
50         longitude=' + str(longitude)
51     try:
52         file = urllib2.urlopen(url)
53     except urllib2.HTTPError, e:
54         logging.error('HTTPError = ' + str(e.code))
55     except urllib2.URLError, e:
56         logging.error('URLError = ' + str(e.reason))
57     except httplib.HTTPException, e:
58         logging.error('HTTPException')
59     except Exception:
60         import traceback
61     data = file.read()
62     file.close()
63     doc = untangle.parse(data)
64     block = doc.Response.Block['FIPS']
65     return block
66
67 # Write out to CSV
68 def writeCsv():
69     lat_lon_list = returnLatLon()
70     with open(hotelWriter, 'w') as csvfile:
71         fieldnames = ['Block', 'Latitude', 'Longitude']
72         writer = csv.DictWriter(csvfile, fieldnames=fieldnames)
73         for cs in lat_lon_list:
74             latitude = cs['Latitude']

```

```

74     longitude = cs['Longitude']
75     block = returnBlock(latitude, longitude)
76     # geoid = "14000US"+block[:11] # required for other cities
77     writer.writerow({'Block':block, 'Latitude':latitude, 'Longitude':longitude})
78     print str(block)
79
80
81 # Main function
82 if __name__ == '__main__':
83     writeCsv()

```

Listing A.4: Converting and amassing **Hotel** coordinates

Appendix B

Chloroplasts and Other Maps

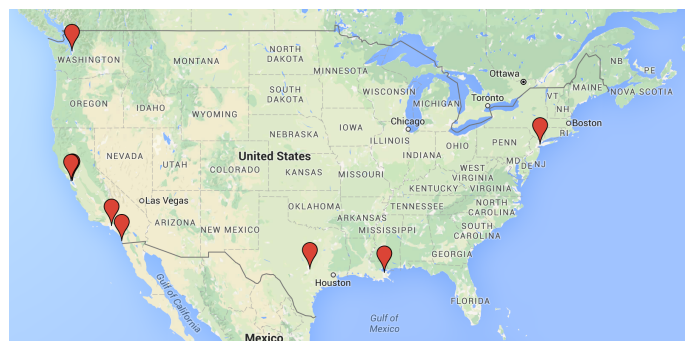


Figure B.1: Map of Chosen Cities

Appendix C

Regression Analysis

Code

The *R* code below is a simplified version of the method used to read in and normalise, run a linear regression and stepwise linear regression, and the Moran's test on each model. We simplify the code for readability. The full code file can be found in the Google Drive folder as linked to above.

```
1 # Read in Transformed Data
2 austin.in = read.csv("/Users/andrewgretores/.../san_diego_combined_data.csv")
3
4 # Compute Z-scores and convert to data-frame
5 aus.zscores <- scale(austin.in, center=TRUE, scale=TRUE)
6 aus.zscores <- as.data.frame(aus.zscores)
7
8 # Run the Linear Regression on each full model
9 fit.lm.one <- lm(Aus.Airbnb.Price ~ Aus.Hotels ... Aus.Distance.To.Center, data= aus
   .zscores) # test against all variables
10 ...
11
12 # Stepwise Lienar Regression for each model
13 step1 <- stepAIC(fit.lm.one, direction="both") \\
14 ...
15
16 # Moran's Test
17
```

```

18 # Residuals for reach model
19 residf1 = fit.lm.one$residuals
20 ...
21
22 # Compute Inverse Distance Matrix
23 reslat.dists <- as.matrix(dist(cbind(latlon $LON, latlon $LAT))) # Distance matrix
24 reslat.dists.inv <- 1/reslat.dists # Take inverse of distance matrix
25 diag(reslat.dists.inv) <- 0 # Replace diagonal with 0
26
27 # Full Regression Moran I (for each full and stepwise model)
28 Moran.I(residf1, reslat.dists.inv)
29 ...

```

Listing C.1: Regression and Stepwise and Moran's I Implementation in R

Appendix D

Classification Analysis

Code

The code that yielded the accuracy, precision and recall results for classification is shown below.

```
1 library(ipred) # Recursive Partitioning and Regression Trees
2 library(party) # Conditional Inference Trees
3 library(e1071) # SVM
4 library(ROCR) # ROC Curvescla
5 library(rpart)
6 library(randomForest)
7 library(adabag)
8 library(gbm)
9
10 cities = c('Aus', 'LA', 'NO', 'Oak', 'SD', 'Sea', 'Man', 'SF')
11 modelCount = '8' # May change number of cities used
12
13 accuracy.Results.Matrix <- data.frame(
14     Aus=double(),
15     LA=double(),
16     NO=double(),
17     Oak=double(),
18     SD=double(),
19     Sea=double(),
```

```

20         Man=double() ,
21         SF=double() ,
22         stringsAsFactors=FALSE)
23
24 precision.Results.Matrix <- data.frame(Model=character() ,
25         VHP=double() ,
26         VLP=double() ,
27         stringsAsFactors=FALSE)
28
29 recall.Results.Matrix <- data.frame(Model=character() ,
30         VHP=double() ,
31         VLP=double() ,
32         stringsAsFactors=FALSE)
33
34 opt.Gamma.Cost.Matrix <- data.frame(
35         opt=character() ,
36         Aus=double() ,
37         LA=double() ,
38         NO=double() ,
39         Oak=double() ,
40         SD=double() ,
41         Sea=double() ,
42         Man=double() ,
43         SF=double() ,
44         stringsAsFactors=FALSE)
45
46 # Optimal cost and gamma parameters (previously computed)
47 Gamma.cost.Matrix.Filled <- matrix(c(0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 80.0,
48         200.0, 120.0, 80.0, 40.0, 80.0, 40.0, 40.0), nrow=2, ncol=8, byrow = TRUE)
49 opt.Gamma.Cost.Matrix[1,1] <- "Gamma"
50 opt.Gamma.Cost.Matrix[2,1] <- "Cost"
51
52 # Loop over all cities
53 for(i in 1:length(cities))
54 {
55     city = cities[i]
56
57     # Import testing and training data
58     testData.in = read.csv(paste("/Users/andrewgretores/Documents/University College

```

```

    London/Dissertation/StatisticalAnalysis/Predictive_Analysis/4ClassExtremes/
    TestingSets/TestAB2_",city,".csv",sep=""))
58 trainData.in = read.csv(paste("/Users/andrewgreatorex/Documents/University College
    London/Dissertation/StatisticalAnalysis/Predictive_Analysis/4ClassExtremes/
    TrainingSets/", modelCount,"Excl",city,"Training.csv",sep=""))
59
60 testData <- data.frame(testData.in)
61 trainData <- data.frame(trainData.in)
62
63 testData.match <- testData
64 trainData.match <- trainData
65 expl.var.location <- length(testData.match[1,])
66
67 # -----#
68
69 # SVM - radial basis (default kernel)
70
71 # Uncomment the following code to find optimal gamma and cost parameters
72
73 # obj.s.r <- tune.svm(Airbnb_Count_km2~., data = trainData.match, gamma=seq(.1,2,
    by=.2), cost=seq(40,200, by=40))
74 # gamma.s.r <- obj.s.r $best.parameters[,1] # optimised gamma
75 # cost.s.r <- obj.s.r $best.parameters[,2] # optimised cost
76 # gamma.s.r <- Gamma.cost.Matrix.Filled[1,i]
77 # cost.s.r <- Gamma.cost.Matrix.Filled[2,i]
78
79 opt.Gamma.Cost.Matrix[1,i+1] <- gamma.s.r
80 opt.Gamma.Cost.Matrix[2,i+1] <- cost.s.r
81
82 svm.model <- svm(Airbnb_Count_km2 ~ ., data=trainData.match, cost=cost.s.r, gamma=
    gamma.s.r)
83 svm.pred <- predict(svm.model, testData.match[, -expl.var.location])
84
85 # -----#
86
87 # Random Forests
88
89 obj.rf <- tuneRF(trainData.match[, -expl.var.location],trainData.match[, expl.var.
    location], stepFactor=0.5)

```

```

90 rf.model <- randomForest(Airbnb_Count_km2 ~ ., data=trainData.match, ntree=500)
91 rf.pred <- predict(rf.model, newdata=testData.match[, -expl.var.location], type="
    class")
92
93 # -----#
94
95 # Multinomial Logistic Regression
96
97 mlr.model <- multinom(Airbnb_Count_km2 ~ ., data=trainData.match)
98 mlr.pred <- predict(mlr.model, newdata=testData.match[, -expl.var.location], type="
    class")
99
100 # -----#
101
102 # Naive Bayes
103
104 nb.model <- naiveBayes(Airbnb_Count_km2 ~ ., data = trainData.match)
105 nb.pred <- predict(nb.model, newdata=testData.match[, -expl.var.location], type="
    class")
106
107 # -----#
108
109 s = table(pred = svm.pred, true = testData.match[, expl.var.location])
110 rf = table(pred = rf.pred, true = testData.match[, expl.var.location])
111 mlr = table(pred = mlr.pred, true = testData.match[, expl.var.location])
112 nb = table(pred = nb.pred, true = testData.match[, expl.var.location])
113
114 s.perf = classAgreement(s)
115 rf.perf = classAgreement(rf)
116 mlr.perf = classAgreement(mlr)
117 nb.perf = classAgreement(nb)
118
119 # -----#
120
121 accuracy.s <- sum(diag(s)) / sum(s)
122 accuracy.rf <- sum(diag(rf)) / sum(rf)
123 accuracy.mlr <- sum(diag(mlr)) / sum(mlr)
124 accuracy.nb <- sum(diag(nb)) / sum(nb)
125

```

```

126 precision.s <- diag(s) / rowSums(s)
127 precision.rf <- diag(rf) / rowSums(rf)
128 precision.mlr <- diag(mlr) / rowSums(mlr)
129 precision.nb <- diag(nb) / rowSums(nb)
130
131 recall.s <- (diag(s) / colSums(s))
132 recall.rf <- (diag(rf) / colSums(rf))
133 recall.mlr <- (diag(mlr) / colSums(mlr))
134 recall.nb <- (diag(nb) / colSums(nb))
135
136 svm.city = paste(city, "_svm.rad", sep="")
137 rf.city = paste(city, "_rf", sep="")
138 mlr.city = paste(city, "_mlr", sep="")
139 nb.city = paste(city, "_nb", sep="")
140
141 precision.Results.Matrix[i*4-3,1] <- svm.city
142 precision.Results.Matrix[i*4-2,1] <- rf.city
143 precision.Results.Matrix[i*4-1,1] <- mlr.city
144 precision.Results.Matrix[i*4,1] <- nb.city
145 for(cat in 1:length(precision.s))
146 {
147   precision.Results.Matrix[i*4-3,cat+1] <- precision.s[cat]
148   precision.Results.Matrix[i*4-2,cat+1] <- precision.rf[cat]
149   precision.Results.Matrix[i*4-1,cat+1] <- precision.mlr[cat]
150   precision.Results.Matrix[i*4,cat+1] <- precision.nb[cat]
151 }
152
153 recall.Results.Matrix[i*4-3,1] <- svm.city
154 recall.Results.Matrix[i*4-2,1] <- rf.city
155 recall.Results.Matrix[i*4-1,1] <- mlr.city
156 recall.Results.Matrix[i*4,1] <- nb.city
157 for(cat in 1:length(recall.s))
158 {
159   recall.Results.Matrix[i*4-3,cat+1] <- recall.s[cat]
160   recall.Results.Matrix[i*4-2,cat+1] <- recall.rf[cat]
161   recall.Results.Matrix[i*4-1,cat+1] <- recall.mlr[cat]
162   recall.Results.Matrix[i*4,cat+1] <- recall.nb[cat]
163 }
164

```



```

165 accuracy.Results.Matrix[1,i] <- accuracy.s
166 accuracy.Results.Matrix[2,i] <- accuracy.rf
167 accuracy.Results.Matrix[3,i] <- accuracy.mlr
168 accuracy.Results.Matrix[4,i] <- accuracy.nb
169
170 }
171
172 precision.results.out = paste("/Users/andrewgreatorex/Documents/University College
    London/Dissertation/StatisticalAnalysis/Predictive_Analysis/Results_4Class_
    Extremes/Predictive/", modelCount, "_precisionResults_fourmodel2.csv", sep="")
173
174 recall.results.out = paste("/Users/andrewgreatorex/Documents/University College
    London/Dissertation/StatisticalAnalysis/Predictive_Analysis/Results_4Class_
    Extremes/Predictive/", modelCount, "_recallResults_fourmodel2.csv", sep="")
175
176 accuracy.results.out = paste("/Users/andrewgreatorex/Documents/University College
    London/Dissertation/StatisticalAnalysis/Predictive_Analysis/Results_4Class_
    Extremes/Predictive/", modelCount, "_accuracyResults_fourmodel2.csv", sep="")
177
178 write.table(precision.Results.Matrix, file=precision.results.out, sep=",", eol="\n",
    row.names=FALSE, col.names=TRUE)
179
180 write.table(recall.Results.Matrix, file=recall.results.out, sep=",", eol="\n", row.
    names=FALSE, col.names=TRUE)
181
182 write.table(accuracy.Results.Matrix, file=accuracy.results.out, sep=",", eol="\n",
    row.names=FALSE, col.names=TRUE)

```

Listing D.1: Classification Implementation in R

Column Combinations

A final analysis conducted that we decided not to include in the main body of this document was to identify which features had the greatest positive impact on the classification of results. By iterating over every unique permutation of features, the statistical significance of each combination was computed. Perhaps unsurprisingly, the results of this analysis indicated that the combination of Young Persons Proportion and Bohemian Index were the combination that yielded the greatest accuracy result. Note that because of the large number of computations, we only decided to use one algo-

rithm; SVM. The results of the analysis and code may be found in the attached Google Drive folder.

Results - Full Classification

Tables D.1 through to D.6 show the accuracy, precision and recall results for the full model and full benchmark model. This set of tables give greater insight into the performance of the model on individual cities. The most interesting result is that in Austin, Texas, the benchmark is extremely close to the full model, implying that the distance to center is an extremely important factor in this case.

	Aus	LA	NO	Oak	SD	Sea	Man	SF
SVM	0.29	0.41	0.48	0.54	0.51	0.47	0.46	0.45
RF	0.46	0.51	0.55	0.58	0.45	0.51	0.41	0.47
MLR	0.51	0.47	0.52	0.62	0.52	0.45	0.43	0.47
NB	0.44	0.51	0.48	0.53	0.46	0.52	0.38	0.49

Table D.1: Four Class Classifier - Accuracy Results By City

	Aus	LA	NO	Oak	SD	Sea	Man	SF
SVM	0.22	0.27	0.26	0.28	0.30	0.27	0.22	0.23
RF	0.45	0.30	0.36	0.43	0.36	0.45	0.34	0.46
MLR	0.48	0.30	0.31	0.29	0.46	0.35	0.33	0.27
NB	0.44	0.39	0.28	0.26	0.42	0.28	0.38	0.26

Table D.2: Four Class Classifier - Benchmark Accuracy Results By City

	Aus	LA	NO	Oak	SD	Sea	Man	SF
SVM	0.32	0.48	0.47	0.54	0.53	0.49	0.44	0.48
RF	0.47	0.55	0.55	0.59	0.48	0.53	0.39	0.49
MLR	0.55	0.49	0.55	0.63	0.54	0.43	0.40	0.45
NB	0.44	0.53	0.48	0.55	0.49	0.51	0.34	0.47

Table D.3: Four Class Classifier - Precision Results By City

	Aus	LA	NO	Oak	SD	Sea	Man	SF
SVM	0.35	0.28	0.33	0.46	0.28	0.33	0.21	0.22
RF	0.48	0.35	0.36	0.48	0.51	0.46	0.36	0.48
MLR	0.44	0.29	0.33	0.47	0.47	0.36	0.30	0.20
NB	0.38	0.41	0.28	0.14	0.43	0.47	0.42	0.33

Table D.4: Four Class Classifier - Benchmark Precision Results By City

	Aus	LA	NO	Oak	SD	Sea	Man	SF
SVM	0.29	0.41	0.48	0.55	0.52	0.47	0.46	0.45
RF	0.46	0.51	0.54	0.59	0.45	0.49	0.41	0.46
MLR	0.48	0.47	0.52	0.62	0.52	0.45	0.43	0.47
NB	0.45	0.51	0.48	0.54	0.46	0.52	0.38	0.50

Table D.5: Four Class Classifier - Recall Results By City

	Aus	LA	NO	Oak	SD	Sea	Man	SF
SVM	0.25	0.27	0.26	0.27	0.30	0.27	0.22	0.22
RF	0.45	0.30	0.35	0.44	0.37	0.45	0.34	0.47
MLR	0.48	0.30	0.30	0.29	0.46	0.35	0.33	0.28
NB	0.44	0.39	0.28	0.25	0.42	0.28	0.38	0.26

Table D.6: Four Class Classifier - Benchmark Recall Results By City

Results - Extreme Classification

In the same format as above, tables D7-D12 show the accuracy, precision and recall results for the extreme model (two class classifier) and the extreme benchmark model. Notably, Austin's accuracy is much improved (excluding the SVM) indicating that it is far easier to predict Austin's lower and upper penetrations of Airbnb.

	Aus	LA	NO	Oak	SD	Sea	Man	SF
SVM	0.77	0.79	0.93	1.00	0.88	0.85	0.84	0.94
RF	0.91	0.83	0.92	0.98	0.88	0.88	0.85	0.93
MLR	0.94	0.82	0.93	0.98	0.86	0.91	0.85	0.90
NB	0.90	0.87	0.89	0.93	0.85	0.82	0.75	0.91

Table D.7: Extreme Class Classifier - Accuracy Results By City

	Aus	LA	NO	Oak	SD	Sea	Man	SF
SVM	0.65	0.50	0.59	0.48	0.53	0.55	0.50	0.56
RF	0.95	0.62	0.83	0.93	0.69	0.90	0.68	0.89
MLR	0.99	0.63	0.72	0.67	0.73	0.76	0.66	0.69
NB	0.98	0.64	0.68	0.57	0.78	0.60	0.61	0.56

Table D.8: Extreme Class Classifier - Benchmark Accuracy Results By City

	Aus	LA	NO	Oak	SD	Sea	Man	SF
SVM	0.78	0.82	0.94	1.00	0.89	0.88	0.87	0.94
RF	0.91	0.85	0.92	0.98	0.89	0.90	0.86	0.93
MLR	0.94	0.83	0.93	0.98	0.86	0.92	0.85	0.92
NB	0.90	0.88	0.90	0.93	0.85	0.84	0.75	0.93

Table D.9: Extreme Class Classifier - Precision Results By City

	Aus	LA	NO	Oak	SD	Sea	Man	SF
SVM	0.74	0.46	0.69	0.52	0.56	0.62	0.51	0.67
RF	0.96	0.64	0.85	0.94	0.72	0.90	0.69	0.91
MLR	0.99	0.65	0.82	0.81	0.73	0.84	0.68	0.81
NB	0.98	0.64	0.81	0.78	0.78	0.78	0.64	0.77

Table D.10: Extreme Class Classifier - Benchmark Precision Results By City

	Aus	LA	NO	Oak	SD	Sea	Man	SF
SVM	0.77	0.79	0.93	1.00	0.88	0.85	0.84	0.94
RF	0.91	0.83	0.92	0.98	0.88	0.88	0.84	0.92
MLR	0.94	0.82	0.93	0.98	0.86	0.91	0.84	0.90
NB	0.90	0.87	0.89	0.93	0.85	0.82	0.75	0.91

Table D.11: Extreme Class Classifier - Recall Results By City

	Aus	LA	NO	Oak	SD	Sea	Man	SF
SVM	0.65	0.49	0.59	0.51	0.53	0.55	0.51	0.57
RF	0.95	0.62	0.83	0.92	0.69	0.90	0.68	0.89
MLR	0.99	0.64	0.72	0.64	0.73	0.76	0.66	0.68
NB	0.98	0.64	0.67	0.54	0.78	0.60	0.61	0.55

Table D.12: Extreme Class Classifier - Benchmark Recall Results By City