Copenhagen
Business School
HANDELSHØJSKOLEN

Predicting Airbnb nightly prices

A regression problem using machine learning

Dissertation paper

Student: Alin-Cristian Preda *Student number*: 125118

Supervisor: Weifang Wu

Number of characters incl. spaces: 171.344
Number of pages: 76

Business Administration and Information Systems – Data Science

15.05.2020
Copenhagen, Denmark

TABLE OF CONTENTS

I. ABSTRACT

Airbnb is an on-line platform, enabling homeowners to rent out their unused space to travellers in need of accommodation. The "hosts" are free to establish their own arbitrary prices. Thusly, it becomes essential for these small-time entrepreneurs to gather clues as to how much they should charge. By making use of analytics and machine learning, is it possible to harvest the power of data, for the purpose of discovering actionable insights which enable data-driven decisions? Also, is the use of open data sources, such as the Inside Airbnb project, sufficient for this task? And is it better to employ cross-market data or simply focus on one city? The main objective of this research project was to come up with a model that can reliably predict nightly Airbnb prices of rooms and homes. A secondary goal was experimenting with new features – review comments text sentiment and listing photography quality scores - and new approaches to training data – using sets of multiple cities rather than just one market's data. The models that ended up being employed for this task were Random Forests and XG-Boost, which are quite capable of tackling supervised learning regression problems. Pre-trained neural networks and natural language processing's sentiment analysis branch were employed towards engineering new features which could add predictive power. The study of geo-spatial data through visualization was used to uncover insights into similarities and differences between markets. Existing literature written on the subject has aided in showcasing good practices, confirming universal findings, and providing inspiration for new approaches and perspectives. The scope was narrowed down to ten major European cities.

XG-Boost has proven itself the superior regression method, scoring highest across multiple approaches. Its best result offers an R2 score of 0.64, when making use of all ten cities' data and, also the engineered features. As is consistent across research, features such as a listing's capacity, its proximity to the centre and whether a place is fully rented out are some of the most important indicators of price levels. I demonstrated the potential of feature-engineering photographs and review texts. Open data sources do not account for all the variability of the prices and new features are to be sought out. I believe there are clues to be discovered in studies which focus on the social intricacies of Airbnb. It probably is worth to have a closer look at how hosts present themselves and how their image and interactions influence their potential to attract and secure "guests" for more competitive prices. Overall, I would argue that we are not quite there yet in terms of automating decision-making in this particular industry but neither have we reached the end of possibilities.

Key words: Airbnb; regression; price prediction; machine learning; housing; rental

II. INTRODUCTION

The digital revolution has managed to establish a prolific breeding ground for on-line business models. The rise of the sharing economy thusly became an inevitable shift towards financially empowering the individual. By making use of his or her own physical (but also non-physical) assets, many people today have essentially become micro-economic agents. Airbnb has become very popular recently and is probably the go-to on-line place to search for short-term accommodation. Whether you are a tourist, someone visiting friends and relatives or you're searching for a place to stay during your business trip, you'll be very likely to find something that fits both your needs and your budget. Airbnb quickly became a sensation and a (threatening) direct competitor to the traditional hotels, motels and hostels. People appreciate the service for being convenient and affordable. Tough not all Airbnb listings come cheap. There is tremendous variety in terms of location, design, accommodating capacity, pricing and amenities. For Airbnb home-owners or "hosts", as they will be referred to henceforth, this means that almost anything will do. After all, the company started off as three college students renting out an air-mattress in their apartment, in order to spare some money for paying their own rent.

Since the variety I have mentioned is so great, the question of how to approach pricing strategies inevitably arises. My view is that we can look at what already exists on the markets and use the data made available by the wonderful people of the Inside Airbnb project in order to construct statistical models capable of predicting the nightly fees to a certain extent of accuracy. Thusly, the research will focus on said data and its exploitation through machine learning and data analytics. The scope will be narrowed down to making use of data from ten major tourist destinations around the EU, from ten different countries. Some attention will also be dedicated to a mostly superficial study of text, geo-spatial and time-series data. These will enable us to paint a broader picture.

My ambitions are by no means a novel entry in the field, but I will try to bring my unique contribution in various niches related to the subject. I have made it possible to have easier access to the data and to be able to download it in large quantities through a specialized application. I have also made some progress in automating the deployment of data visualizations, fit for being used in quick exploratory data analysis. This project is also the first one, as of yet, to make use of such a varied dataset. While the majority of takes on this regression problem chose to narrow the scope down to a single city, I

opted to test different approaches: using one city, pairs of cities with similar price distributions, as well as a larger dataset encompassing a collection of very different markets. The study also introduces the novelty of making use of photograph quality as a predictive feature, albeit the access to the actual photos was very limited, but proven to be worthy of attention for future enthusiasts. The study of this subject is relevant to the data science community, as a means of showcasing ways of reinterpreting already tackled problems. It is also relevant to those who are interested in obtaining profits off of Airbnb or simply searching for ways to make their meets end, in true spirit of the sharing economy.

The following work will be structured in relevant chapters. Literature Review is a commentary on works that I have found to be augmentative to my research. That is, both directly and indirectly. There are papers focusing on regressing prices and other papers which discus different niches of Airbnb such as trust, discrimination and consumer behaviour. Following this, the Case Company Background chapter takes a look at how Airbnb was born, how it evolved in the giant it is today and how its future might unfold, amidst the global Coronavirus crisis. Airbnb's legal status and public opinion in Europe is discussed, as well as some of key features and concepts. The Data chapter discusses the structure of the data on Inside Airbnb, the scraping process, the pre-processing, feature selection and feature engineering. The next chapter focuses on Methodology. More precisely, it discusses the tools and methods used in analysis. The choice of algorithms is discussed, based on how they work and what they bring to the table. Tables and figures are presented and used to comment on listing, text and geo-spatial data. In the Results section, I present the algorithms used for machine learning, along with hyper-parameter choices and their respective performance metrics. Lastly, the final chapter is about Discussions, Conclusions and Recommendations, where I interpret my personal findings, as well as correlate this information with the results of already existing studies presented in the Literature Review.

III. LITERATURE REVIEW

If we are to judge by modern standards, at the rate that information is spreading, and at which technology is evolving, and with it, the economy and society, I would argue that Airbnb, although a relatively young company, has become a staple of our lifestyle. Contemporary society is becoming increasingly unthinkable without having a tech solution to any sort of problem or a tech alternative to doing virtually any type of business. Consequently, finding information on the subject is relatively easy. The sources of information are plentiful, yet, I must mention, unexpectedly unvaried. I mainly started out my research on scholar.google.com, and went from there. For technical information, which helped me complete the programming parts, I was backed up by my university courses, books and assignments. As alternative sources of information or inspiration, I relied on Kaggle, GitHub, Stack Exchange, Reddit, You Tube and other social platforms. I also read a lot of blogs such as Analytics Vidhya, Medium, Towards Data Science and other media outlets.

Initially, I started with looking up projects and research papers aimed at regressing nightly prices. However, the more I delved into the data, and the more I discovered in the ideas and findings of others, the more interesting subjects I discovered. Many people are concerned with different social aspects of Airbnb, such as diversity, inclusion, safety, lawfulness, community and sustainability. More abstract ideas such as trust and reputation are explored, with some thought-provoking implications. Some go for a more business-oriented approach and try to find better ways to monetize the service. Others just find the data to be a good candidate for practicing data science. Not all papers cited here will receive the same amount of attention. I will try to focus on those that inspired and which have more extraordinary content, as many have reached similar conclusions.

III.1 PRICE DETERMINANTS

The majority of sources seems to agree on a few key conclusions, which become self-evident once the researcher starts delving more deeply into the data. There are also some curios discoveries and some more curios assumptions that are being made, as well. One such paper is based on the idea of Dynamic Pricing, a strategy by which the hosts fluctuate their rates according to changes in the market. The study found out that multi-listing hosts outperform single-listing hosts by positioning the listing at a higher price than the neighbourhood average and by adopting less dynamic pricing strategies. The researchers have recommended that multi-unit hosts maintain high-price positionings

and be wary of potential negative effects of dynamic pricing strategies. Also, for single-unit hosts, they recommend relatively high price positioning, but to consider monitoring the market an ongoing process. This conclusion makes me draw a parallel, from these Airbnb hosts to more traditional businesses, where large players dominate the market and small ones either have to fight for small profit or get swallowed by the giants. Single unit hosts make up the majority of the population, and they are in greater competition with each other. They don't benefit from the same resources that multi-unit hosts do, which I believe can mitigate risks and potential losses more efficiently, by fashioning their properties into sort of an investment portfolio (Kwok, 2018).

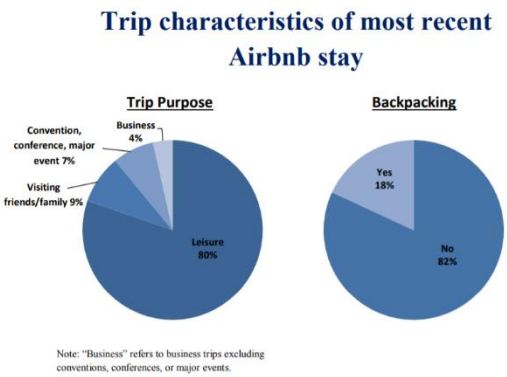
III.2 IMPORTANT FEATURES

From what I've read, almost any attempt to analyse the data quickly reaches the conclusion that, although very few acquire Superhost status, the ones that do so end up seeing more competitive prices, more bookings, more ratings, and are basically getting the most out of Airbnb. The causality needs to be explored in greater detail, though, for the correlation is clear. A study by Wang and Nicolau (which is the most cited paper I have found) found out that the Superhosts get to experience an 8.73% price increase, on average (Wang & Nicolau, 2017). According to their OLS coefficients, prices raise by 0.06% for each listing a host has counted, but the variable has a smaller effect on higher-end listings. They, unsurprisingly, like all studies, found that the further away from the centre a listing is, the cheaper it gets. The authors of the research also agree with previous findings by Gutt & Hermann and Ikkala & Lampien, which support the idea that hosts monetize their reputations. Again, they proved that renting out entire homes leads to higher pricing, but with a twist: the effect is greater for low priced listings and smaller for the expensive ones. The number of people accommodated and the provision of bathrooms, bedrooms, and real beds are all associated with higher increments in price. The increased price effect due to having wireless Internet is more impactful for low-priced listings. Offering breakfast has been found to have a significantly negative effect on prices, a finding that is noted to not be consistent with research done by the hotel industry. The authors of the study believe (as do I) that, perhaps these hosts are trying to make their inferior quality listings seem more appealing. I would add that these hosts might also be new, and inspired by hotels to include things that they think are relevant. The positive effect of free parking on the premises has been found to be more impactful for low-priced listings. Instant-booking is one feature that negatively affects the price. It is a positive feature for guests, no doubt about it. The fact that it is associated with lower prices might be part of a strategy devised by the hosts, which is based on a high occupancy rate, with lower than usual prices, which makes for a more profitable game plan. Listings in which smoking is not

prohibited do charge lower prices. Perhaps this is a conscious move on the side of the hosts, who might be smokers themselves. The authors assume their empathy with smoker guests, but I beg to differ. I suspect that it might be more of a discount for non-smokers, if smoking is ongoing on the premises or there are signs of it being a smoker's home. The results of an MDPI journal seem to conclude that "the distance to the convention center (C-Distance), the number of reviews (Reviews) and the review rating scores (Rating) are significantly connected with the Airbnb listing price." (Zhang, Chen, Han, & Yang, 2017)

"Price Determinants of Airbnb Listings: Evidence from Hong Kong" reveals the importance of certain features as price determinants. Property size, the number of bedrooms and bathrooms, the number of accommodations, and certain accommodations such as free parking were found to be associated with higher prices. Also, instantly bookable properties and those with flexible cancellation had lower than average prices, findings consistent with all other studies (Cai, Zhou, Ma, & Scott, May 2019). I am starting to remark the fact that findings such as this are so common that they basically will become a statement. The fact that studies are done with data from all over the world also bears much weight in signalling the prevalence of this trend.

III.3 SOCIAL ASPECTS OF AIRBNB

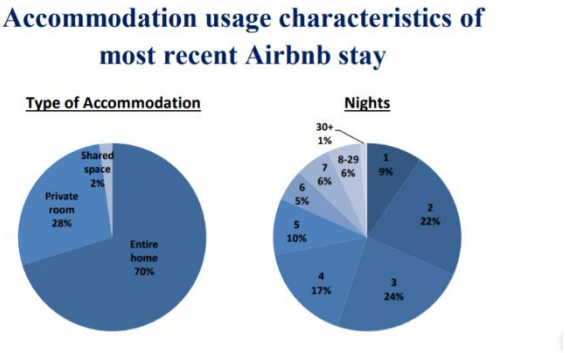


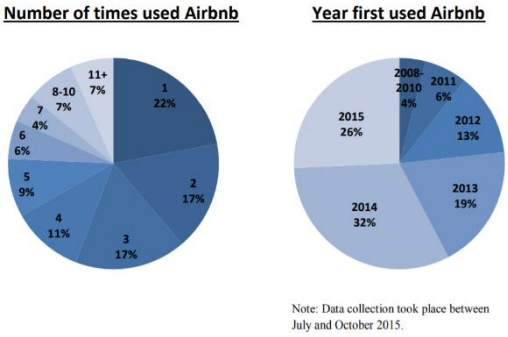
A very interesting paper uses polls to answer some questions about Airbnb. Does the business follow up on its initial flavour of shared economy, backpackers, students and air mattresses? What is your typical guest and host? Who uses this service? What do people enjoy about it? What are its strengths and competitive advantages? Basically, what is the data telling us about this phenomenon? The blue graphs are from that paper.

This particular study (Guttentag, Airbnb: Why Tourists Choose It and How They Use It, August 2016) focuses on performing some very interesting data analysis on Airbnb, which reveals some relevant information about guests. Contrary to the core ideals of the sharing economy, back-packers only represent less than 20% of guests. This could mean that guests are more alike hotel clients than their Couchsurfing counterparts. The overwhelming majority of Airbnb guests appeal to Airbnb for leisure purposes. Only a fifth go to Airbnb for other purposes such as visiting, conventions or business

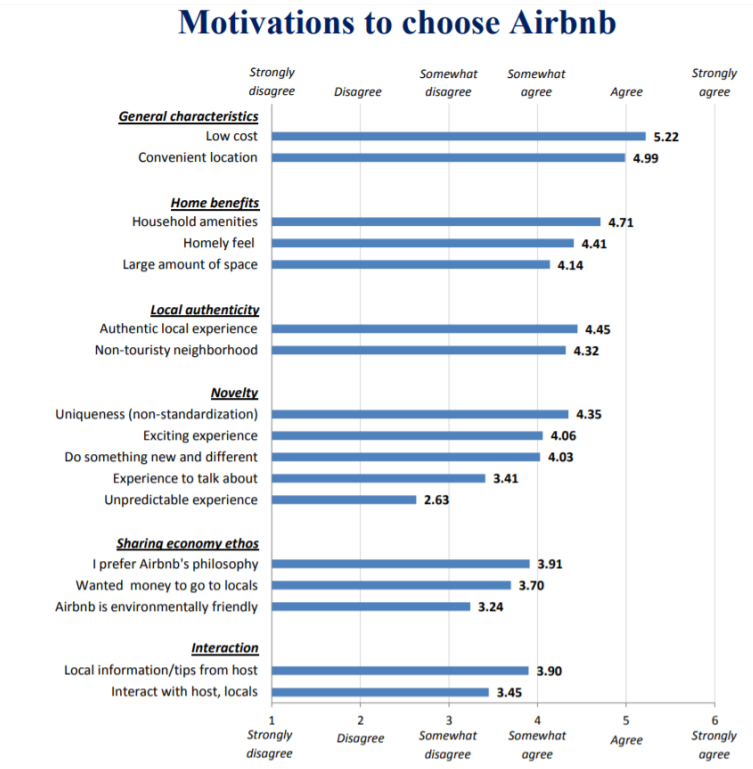
trips, which makes sense to me. It does still feel like a riskier way of doing things. I assume large companies will still appeal to hotels.

Most hosts rent out their entire place, rather than sharing a niche of their space with the guests. The study concludes that host-guest interaction is not a characteristic of this service, nor a motivator for choosing it.





Guests mainly book between 2 and 3 nights, which is consistent with my own discoveries. Most people questioned had only used Airbnb one to three times. But a significant percentage of them (7%) have used it over 11 times, which is quite high. The overwhelming majority of people had first used it after 2014. Only 4% used it before 2010. This shows us that the company is gaining increasing popularity.



As for the motivations to use Airbnb, there is a rather rich range of factors. Guests are mostly attracted by the practicality of it. The pragmatism of low cost, good locations and nice to have amenities are too tempting to say not be considered. In a secondary plan come the more experiential motivators such as: authenticity, novelty seeking. People are mostly concerned about tangible, material, practical advantages than over adventure-seeking. I assume that fact that the home is more like your own home than a hotel - because you can cook, you can perhaps smoke, you can read books and whatnot - is what people value amenities-wise.

The chart is very well thought of because it groups people's needs into general categories and presents them in descending order of their importance.

III.4 SUPERHOST STATUS

"The Impacts of Quality and Quantity Attributes of Airbnb Hosts on Listing Performance" managed to uncover that the Superhosts of their sample had, on average, twice as more reservations. The authors, Karen Xie and Zhenxing Mao, also note that a 1% increase in response rates led to five more reservations. I would like to question the exact relationship between these two variables. Surely, a host that wants more reservations, is more likely to be more active, than someone who opts to only do this occasionally. Hosts who share their space, for example, I believe to be more likely to ponder over responding. There's also the fact that response rates are only related to the promptitude of messaging exchanges, which generally just conclude the formalities of the transaction. I.e. the host accepts the guest's request for making a booking. But there is also the case of instant booking, in which reservations are basically concluded instantaneously, without the host's explicit approval. Going back to the study, the researchers note that after increasing their status on the platform, the studied hosts did receive more bookings, which is basically also guaranteed by Airbnb. Again, related to trust, hosts are encouraged by Airbnb to improve trust-signalling cues such as: experience, response rates, status, etc. Certainly, some of these cues signal desirable host qualities such as: time management, social skills and communication abilities, care, effort, efficiency, effectiveness. The study concludes that guests do not rate their perceived quality of hosts based on identity verification provided by Airbnb (Xie & Mao, July 2017). Another paper also managed to include a feature based on sentiment analysis using Text Blob. Unfortunately, there are no comments offered on the success of this endeavor. (Kalehbasti, Nikolenko, & Rezaei, July 2019).

III.5 TRUST

Trust is crucial for the success of P2P types of businesses, especially when it involves property rental and/or sharing. This is why Airbnb has implemented some trust management systems and social reputation systems. Aside from the social aspect of reputation, it is not yet clear how it translates into monetization. Economically speaking, we could be looking at the potential to earn more, by being booked more. The paper *"Price determinants on Airbnb: How reputation pays off in the sharing economy"* was conducted on an Inside Airbnb styled dataset of 86 German cities. Its authors found out that host's rating scores, host's membership age and even their photographs consistently

translated into price premiums. Interestingly, they also found out that there are certain city specific attributes that have important impact on prices (Teubner, Dann, & Hawlitschek, May 2017). They assessed these attributes' economic value by conducting a set of linear regressions, with the target being the nightly price for two persons, for two nights, and a cleaning fee, for a fairly standard Airbnb scenario. What they found out was, quite unsurprisingly, that: larger locations add to price, renting entire homes is more expensive, the distance to the city centre is an important price determinant. Also, the larger the city, the higher the prices in general, which seems only reasonable, but I haven't seen this conclusion being drawn in other papers, because they don't stretch the analysis. Deposits and strict cancellation policies were associated with higher prices. Here, one should be attentive to the correlation-causality relationship. De facto luxury properties prompt hosts to ask for deposits due to the expensive nature of the goods situated in the location. Also, there is the logic behind it that guests who desire and can afford premium, will also be able and willing to pay deposits.

A Stockholm based study highlights the importance of trustworthiness of the host. This is mainly base on the profile picture, which was found to significantly impact the pricing. The more perceived trustworthiness of the host, the higher the price can get. Interestingly, and to their surprise, the authors did not register a significant effect of the review scores, which they admitted contradicts other research. They found out that review scores on Airbnb had very low variance: 97% of listings received scores above 4.5 out of 5. They also found some evidence that hosts benefitted from being female and attractive (Ert, January 2015). The hypothesis is that, perhaps, given the failure of the online review score system, the host's profile might have stepped up to fill the role. The logic behind the assumption is that, given the absurdly high proportion of almost perfect scores, there must be some other way for guests to distinguish between hosts which can help them make decisions. The need for trust, especially in this P2P economy, manifests itself in various ways. It can lead to consumers appealing to more sensual cues, such as visual ones, when more abstract cues such as reputation systems, are inherently biased. Nevertheless, the author does not stress out these facts as being important. He just laid out the facts and the conclusions are up for debate.

III.6 SENTIMENT ANALYSIS

A group of students from the University of Washington wrote a paper in 2018, titled *"Predictive modelling on Airbnb listing prices",* Consistent with my own research, they also conducted Sentiment Analysis on listing descriptions, awarding each with a polarity score from 1-100 and, surprisingly, figured that the score was insignificant related to price. As is consistent with all research, features

such as neighbourhood, Superhost status, bedrooms, bathrooms, number of accommodates were the highest correlated with price. For this study, their neural network provided the best predictive results. The researchers also questioned whether trust varied in online marketplaces, based upon personal appearance. The number of reviews a listing had was found to be rather irrelevant (Keating, Katnic, Hahn, & Yang, 2018).

III.7 DISCRIMINATION

It is a common conception that people make up their minds about others in mere seconds. We are a rather superficial species and this is an evolutionary trait that has helped our ancestors bond with the right people and steer clear of certain individuals. A manuscript from *"Computers in Human Behavior"* has concluded that facial expressions of guests have an impact on guests browsing on Airbnb. The facial expressions have different effects on the genders. And certain expressions cannot be compensated by even unbeatable prices and/or top ratings and reputation. Apparently, *"your personal profile might jeopardize your rental opportunity*" (Fagerstrøm, Pawar, Sigurdsson, Foxall, & Yani-de-Soriano, 2017). It is established in the world of salesmen that your image is the foundation of your relationship with the clients. This relates to trust as much as to discrimination.

A 2014 Harward Business School study focused on discrimination related to Airbnb was made by Benjamin Edelman and Michael Luca. They managed to find some interesting insights, using data from New York, although the actual underlying causes of these relationships are questionable at best.

- Apparently, the host's sexual orientation and gender do not affect rental price significantly.
- Even so, real differences can unfortunately still be seen in racial segregation. Blacks do earn 12% less, even after controlling for different characteristics.
- Social networking presence is important because it validates your internet persona.
- Listing photos are important indicators of price. (Edelman & Luca, January 2014)

A very interesting idea that the authors present is potentially mind-boggling. Presuming that guests do actually discriminate based on race, is it a "flavourful" discrimination, meaning they choose non-blacks based on subjective personal taste? Or do they discriminate based on biased assumptions about the fact that the listing's host is black, such as assuming that the listing might be of inferior quality? Is there actually a difference between these two types of discrimination? The topic is sensible and beyond the scope of my research, but the findings are valuable, nonetheless.

In an article written by research journalist, Chloe Reichel, and published on journalistresrouce.org, she cites a study done in San Francisco, with Airbnb data. Some of her conclusions are valuable.

“

- *Controlling for rental characteristics such as number of bedrooms and bathrooms and cancellation policies, Hispanic and Asian hosts price their listings 15 and 11 percent lower than white hosts, respectively.*
- *Adding in a few other controls, including neighbourhood property values, area demographics and occupancy rates, this disparity was reduced slightly but still existent. After controlling for these, the data indicates Asian and Hispanic hosts charge 8 to 10 percent less than white hosts on equivalent properties.*

“ (Reichel, 2018)

This leads me to believe that listings held by hosts pertaining to a racial minority, are more generally situated further away from the centre, which explains the reduction in disparity. Still, I would have expected the discrepancy to be lower after adding the ulterior controls. Nonetheless, these with the similar study done in NY, which I referenced above.

Another study which is focused on discrimination, this time by gender, used data from five cities which span three continents, so as to be as varied as possible. The findings suggest that certain groups of hosts, like young people, Caucasians and females, are over-represented compared to the local population's composition. Also, substantial evidence was provided for the existence of homophily across all the cities, which is the preference for individuals which are similar to us in certain characteristics (Koh, Li, & Livan, March 2019). Yet again, some other research, "The impact of host race and gender on prices on Airbnb", suggests that, on average, Asian and Black male hosts earn 5% and respectively 3% less than Caucasian males for the same types of listings (Marchenko, December 2019). The author also found out that even though these minority hosts charge lower, they also face lower demand. Like me, she concludes that enough evidence has been presented towards the presence of discrimination in this industry. But although consistent, the findings fail to be conclusive.

III.8 PHOTOGRAPHY

Airbnb also offers guests the option to make use of professional photography for their listings, where possible. They claim that you can earn up to 40% more, get 24% more bookings, and charge 26% higher nightly prices with their professional photographs (Professional Photography, n.d.). The cost of the service is deducted from the host's future pay-outs, varying by home size and location. Supposedly, most hosts are able to pay it off within their next three bookings. Below, we can see some of the examples they've given. The first pictures are taken by hosts themselves and the second ones are reinterpreted by the professional photographers. In my opinion, the professional photos make the places seem larger, brighter and "warmer" than they might actually be. All in all, I also believe that appealing to Airbnb to take the photographs and be "verified" also further boosts your profile. In other words, I'm thinking that it's not just the high quality of the photos that delivers the results, but by having official photos, their algorithm probably makes your place more visible to potential guests.

A research paper from the Thirty Seventh International Conference on Information Systems, Dublin 2016 seems to suggest that, “for a room priced at $100/day, having verified photos will bring extra 9% of room booking frequency, leading an extra calendar year income of $3,285 to the host” (Zhang, Lee, Singh, & Srinivasan, 2016). The researchers deducted that part of the effect of verified photos came through their high quality. They say partly, not fully, because at the same time their research suggests that “an increase of $2,455 in calendar year income to the host, if he/she replaced his/her 15 low-quality (and unverified) photos with all high quality (and unverified) photos.” Which leads me to come back to my idea that Airbnb boosts your profile if you use their photography services, more than you would naturally would using your own professional photos. The photo illustrations in this research paper also support the idea that your photos should take wider angles and the image should be brighter and sharper to qualify as quality works. The pictures to the left of this paragraph are taken from this paper and illustrate a before and after take on the same rooms. Tripadvisor has a blog post about pictures. They claim that having at least one picture on your property profile boosts traveller engagement by 138% and your listing becomes 225% more likely to be booked (who would have thought?) (Bookings and traveler engagement driven by management actions, n.d.). Properties with at least a whooping one hundred pictures have engagement levels of over 151% and are 238% more likely to be booked than properties with no photos. While they stress out the importance of pictures, the facts they’re chosen are vague and the baseline is quite laughable: a listing with no pictures. Sadly, not much real insight from Tripadvisor.

III.9 RESULTS OF MACHINE LEARNING

A project by some Stanford University Students, done on data from Melbourne, Australia, used a plethora of machine learning models and even a few neural network variants, to try predicting nightly prices (Cai, Han, & Wu). Out of the machine learning models, Gradient Boosting has the most success. It managed to score an R2 of 0.69 on the test set and was superior to the Gradient Booster done with LASSO feature selection. Interestingly, but not surprisingly, it also managed to do a bit better than the deep learning attempts, which only reached a maximum R2 of 0.65. What is more important to note, though, is that the students incorporated text from reviews and descriptions into some original features, which led them to achieve better results than just using the vanilla features.

Similar to my research, a paper found out that the relationship between the feature vector and the price is non-linear (Kalehbasti, Nikolenko, & Rezaei, July 2019). As such, they also considered that using regression trees was appropriate. They also did feature selection, by multiple criterion. Using LASSO CV proved to be the method that got the R2 the highest. No feature selection was the worst, while manually using feature selection drastically improved model performance. Using P-values was superior to pruning features manually, but not significantly so. The best performing models, based on test-set R2 scores, were SVR, followed by Neural Net and K-means + Ridge Regression.

Another research paper manages to demonstrate, through empirical analysis, that the features using scores based on sentiment analysis of guest reviews are better indicators of price than rating scores. This is consistent with my analysis and other papers from other researchers. Another very interesting suggestion is that, even though components of the overall review score (such as cleanliness and accuracy) are better, individually, at predicting the nighty prices, they are still inferior to the predictive quality of the sentiment scores of reviews. And, they also managed to discover an unexpected effect. That is: the reviews also affected the prices that neighbouring hosts could set to their listings. They call this a "spillover effect" which helps Airbnb subliminally impose a sense of urgency for hosts to aspire to improve the quality of their services. The theoretical model suggests that when a host increases its price, its rivals also increase their price, making them strategic price complements. Out of the multidimensional components of the review score, the scores for cleanliness and accuracy have the most predictive power. This finding is sure to incentivize hosts to focus their efforts on keeping their listings clean and providing the most accurate pieces of detailed information about the services they are offering. Guests don't want to be greeted by a listing that seems unkept and that has some unspecified characteristics, that are deemed unpleasant or undesirable. The same should be understood about features which, by website description, ought to be included (breakfast or Wi-Fi, for example). In reality, sometimes they are not found or they don't meet the guests' expectations. Personally, I am tempted to bet that after things will return to business as usual, after the current Covid situation, guests will put an even greater emphasis on cleanliness and sanitation and other healthcare and perhaps, safety in general, types of amenities (Lawani, Michael, Mark, & Zheng).

An interesting Github project attempts to use neural nets for price prediction. In the end, the author concludes that this prediction problem is just one of those cases where using advanced techniques like deep learning is not a necessity. However, she notes that even her best model "only" had an R2 of 73%. The author attributes the remaining unexplained variation in the price to data that is not present, such as picture quality. And I can't deny that she is right to believe so. Lewis also presented

some very good ideas for potential directions of future work. Among them, I wish to mention: incorporating image quality as a feature; including a wider geographic area such as other cities in other countries; using NLP to make new features (Lewis, 2019).

IV. CASE COMPANY BACKGROUND

IV.1 THE LIFE OF AIRBNB

Wikipedia describes San Francisco-based unicorn Airbnb as an on-line marketplace that offers typically short-term accommodation, suited for touristic experiences. While the corporation does not actually own any real estate, it does act as an online broker. And as a broker, it receives commissions for each transaction that goes through its system. By transaction, we understand bookings made for events, experiences and lodgings (Wikipedia, 2020). The story of the company begins in August 2008 with its three founders: Brian Chesky, Nathan Blecharczyk and Joe Gebbia. The three of them thought about a way to make a few extra bucks, by turning their Loft into a Bed and Breakfast, where guests could sleep on air mattresses. Hence the name: Air, bed and breakfast (Aydin, 2020). What happened next was shocking. They made a website, airbedandbreakfast.com, and by March 2009, when they changed to their iconic new name, the men already had 10.000 users and 2.500 listings. But Airbnb only became profitable 7 long years later, in 2016, when its revenue grew by 80% in that last year. But the business is still somewhat volatile. In 2019, it reported losses of $322 million, after turning a $200 million profit in 2018. Over the years, the company has faced much scrutiny and harassment from different interest groups and authorities and also sanctions and regulations. It also had to invest heavily in safety features and to guarantee better experiences to hosts, guests, and to the communities where it is present and changing the urban housing landscape.

IV.2 EXPLAINING THE PRICING SYSTEM

Regarding additional fees, Airbnb mentions:

- **Cleaning fee**: Guests can incorporate either a cleaning fee into their nightly prices or a separate, independent cleaning fee.
- **Other fees**: Hosts can choose to add: a late check-in fee, a pet fee, or bike rental fee, etc (Airbnb , n.d.).

Airbnb already has a ‘smart-pricing’ system, which, in practice, I found to not be very reliable. Perhaps due to fact that it’s so opaque and doesn’t really explain to the host how it works and why the suggested price is so and not more or less. According to official Airbnb literature, this Smart Pricing is supposed to keep your nightly prices *“competitive as demand in your area changes”.* Its goal is to *“increase your chance of getting booked”*. They actually mention that they’ve received feedback from hosts, suggesting the prices are different from what they expected. Consequently, they gave the following vague explanation:

“

- *Lead-time: as a check-in date approaches, your price will update*
- *Market popularity: if more people are searching for homes in your area, your price will update*
- *Seasonality: as you move into, or out of high season, your price will update*
- *Listing popularity: if you get a lot of views and bookings, your price will update*
- *Listing details: if you add amenities, such as WiFi, your price will update*
- *Bookings history: as you get bookings, your future prices will be partly based on the prices you got for successful bookings. So, for instance, if you set your price higher than Smart Pricing suggests, and you get a successful booking at that price, the algorithm will update to reflect that.*
- *Review history: Your prices update as you get more positive reviews from successful stays.*
- *There are lots of factors at play—Smart Pricing even evaluates how many travelers look at your listing every day and how long they view it for! We really have built this tool to reflect factors you can’t discover just by simply comparing your listing page to others in the area.*

” (Airbnb, n.d.)

After closer inspection, it seems to me that this system is meant to rather influence the hosts to use certain pricing schemes, based on Airbnb’s local interests. With such lack of transparency, I am inclined to believe that what it’s actually doing is trying to gain more control over the market, so that the potential guests are more inclined to book the listings that bring the most profit to the company. I might be wrong with this assumption, but it is a pragmatic deduction. We must always bear in mind that the ultimate goal of any commercial enterprise is to raise profits by any means necessary. After all, Airbnb is too large of a business for it to be run by the whims of its users, without keeping them in check.

IV.3 EXPLAINING SUPERHOST STATUS

This digital badge stands for the status of "superhost", and it will automatically appear on your host profile, once and if you've reached this milestone. According to Airbnb, superhosts are *"experienced hosts who provide a shining example for other hosts, and extraordinary experiences for their guests. We check Superhosts' activity four times a year, to ensure that the program highlights the people who are most dedicated to providing outstanding hospitality."* (Airbnb, n.d.) Besides the pretty badge and the social status it brings, it also guarantees certain advantages directly from the company. You will receive more visibility on the website, which implies more earning potential. There's actually a search filter for Superhost listings, for guests who want the best of the best. And I think it's normal for Airbnb to advertise its star hosts, who are, if you think about it, some sort of indirect employees of the company.

IV.4 AIRBNB ANTI-SENTIMENT AND REGULATIONS

This section is meant to delve into the social and legal aspect of Airbnb. I shall go into more detail about the situation of some European Cities which have had past troubles with Airbnb and subsequently responded by enacting new legislation. I will also discuss a bit of economics, from the perspective of hosts in Copenhagen. Airbnb, while opening up typically less touristy areas and monetizing them more, can also backfire on the local population. According to the BBC, there is a study that found out that full-time listings can earn up to three times the median long-term rent (study done in Manhattan). There are many who feel that a trend where property owners switch from long - term tenancies to short-term ones might be very harmful. Although potentially more lucrative for the hosts, it does pose the danger of accelerating the growth of the prices for properties (and, implicitly, rents). BBC also cites a series of short-term rental restrictions applied in different cities. For example, in Amsterdam, entire home rentals have been limited to 60 nights yearly; in Berlin, the hosts need to apply for permits; in Paris, the yearly cap is 120 nights (Guttentag, What Airbnb really does to a neighbourhood, 2018). In some cities, Airbnb has become a significant part of the local housing units. One such example is Barcelona, where a 2015 study cited by the BBC had found out that around a significant 10% of all homes in Barcelona's Old Town were listed on Airbnb. Another 2014 study done in Los Angeles, California, found out that in neighbourhoods with a strong Airbnb presence, the rents increased 30% quicker than the city average. A wider US research found out that a 10% increase in Airbnb listings *"led to a 0.42% increase in rents and a 0.76% increase in house prices"*.

There's recently been a lot of fuss in Europe over this company. As documented by the BBC, ten European city councils wrote an open letter to the EU structures. In it they ask for help regarding short term renting websites, particularly Airbnb. The main issue they deal with is that the company is considered a digital information platform, not an accommodation provider, according to EU law. *"The cities—Amsterdam, Barcelona, Berlin, Bordeaux, Brussels, Krakow, Munich, Paris, Valencia and Vienna—fear such a ruling would remove a key tool they have to regulate against the worst effects of the vacation-rental industry"* (O'Sullivan, European Cities Fear They'll Lose Power To Regulate Airbnb, 2019) The cities are complaining about the fact that Airbnb holds so much valuable data that could be used for taxing purposes or against criminal offences. The problem is they're not obliged to divulge any of it.

1) **AMSTERDAM**

The University of Amsterdam published a study in 2016. They found out that over a period of 1 year, property prices increased by 0.42% *"whenever the density of Airbnb's in a square kilometre radius increased"*. Alas, Airbnb is an important player in the Netherlands' accommodation industry. It supposedly has its grip on 12% of the market share. Regarding the capital, around 5000 homes are permanently rented out. 81% of bookings are made here, out of a total of 2.6 million, which is a huge figure, even taken out of context. These properties are effectively locked out of the normal housing market. And there is this pressing housing crisis in the city, where its market supply does not meet the demand. And this naturally has the effect of driving prices up. Shockingly enough, the French Data Bureau claims that hotels in the city are 11 € cheaper on average (Stone, 2018). Airbnb is getting blamed for Amsterdam's housing crisis. The guardian confirms our findings, and states that around 22 thousand listings are offered at least for one night yearly. Sito Veracruz, Amsterdamer and urban planner, mentions that the average host earns almost 4000 € per annum if they rent their space for one full month. He is certain that Airbnb is gentrifying the place and is concerned about the price rise, which is already an issue (Zee, 2016).

Amsterdam is one of the first cities to crack down on Airbnb. The company agreed to impose a 30-day yearly limit on its rentals. It would also have to inform hosts of all rules and regulations and help Amsterdam enforce them. Also, a tourist tax on rental apartments would be collected with the help of information provided by Airbnb (Tun, 2020). In January 2020, the Amsterdam City Council had begun working on new legislation which aims to curve the dangerous growth of Airbnb on its premises. Long have the Dutchmen spoken in apocalyptic terms about the threats of this platform. According to www.dutchnews.nl, starting summer 2020, the government is set to enact new

legislation which will regulate holiday rentals. As of today, the Dutch citizens that rent their proprieties via the online platform operate in a "grey" area. Technically, it is already illegal for them to rent out properties to tourists without registering for a permit to do so. But in reality, the authorities have yet to set up clear procedures for such cases. The argument made for this move is that *"landlords are effectively removing a home from the national housing stock"* and so the country is waging war on the aggressive growth of landlord entrepreneurs. The current housing minister, Stientje van Veldhoven, stated that platforms such as Airbnb cannot be legally coerced to provide information because it is against EU guidelines, which see such rental sites as *"information platforms"* (DutchNews.nl, 2020).

2) **BERLIN**

According to Investopedia, German officials began, at one point, placing the blame on Airbnb for increasing rent prices and the house shortage crysis. A law was passed in 2014, restricting the right to Airbnb, by imposing the need to apply for permissions and setting a limit of 60 days. The lawmakers also vowed to reject 95 per cent of applications, but later, in 2018, they changed their minds. The limit was lifted on primary owner-occupied locations. A limit of 90 days was set for secondary properties (Tun, 2020).

3) **COPENHAGEN**

Now, in Denmark, we see a more relaxed approach from the government, which believes that anyone should be able to rent or sub-rent their property, given that taxes are paid, above a certain optional income threshold (generally 28k for main homes and 40k for additional homes). The new rules are a world first: hosts can share their primary home up to 70 nights yearly (which can be set by local municipalities to 100), private homes and summer houses can be shared indefinitely (Denmark Approves Forward-thinking Home Sharing Rules and Simplifies Tax, 2019). These are actually special status rules awarded to Airbnb, for agreeing to directly share information to SKAT, the tax authority. Other non-information-sharing platforms are subjected to stricter rules.

4) **LONDON**

"*Starting from early 2017, Airbnb's systems are automatically limiting entire home listings in Greater London to 90 nights per calendar year"* (I rent out my home in London; what short term rental laws apply?, n.d.). *An amendment to the city's housing law from 2015 allowed Londoners to rent for up to three months per annum, while those living outside the area of Greater London were*

granted even greater rights: 140 days per annum. Airbnb's market share in London jumped from 2.8% to 7.6% in 2017 only (Tun, 2020).

5) **MADRID**

The lawmakers in Madrid have thought of ways to decongest the city centre which is oversaturated with Airbnb listings. Their goal is to gain back territory lost to the tourism industry and to push it towards other parts of the city, to spread its benefits. The new rules state that if an apartment doesn't have its own private entrance, it can't be listed on Airbnb. This excludes any apartment in a block of flats. Alas, it applies to units which are rented out for more than 90 days per annum. Also, the requirement will not be enforced on the outskirts of Madrid. The idea is to ease the strain on the locals and infrastructure (O'Sullivan, Madrid Bans Airbnb Apartments That Don't Have Private Entrances, 2019).

6) **PARIS**

Paris officials believe that home rentals displace locals from the main city. And Paris is Airbnb's largest market, with over 60.000 listings. Investopedia records that there were crackdowns on secondary apartments in the French capital, back in 2015. These apartments were specifically only rented out for short-term stays and violated city regulations. The hosts were fined for up to a 25.000 Euros. These interventions did not prove sufficient. In 2017, officials made it so that hosts were required to officially register their listings. Mayor Anne Hidalgo even went as far as threatening to enforce major punishments, with fines going as high as 12.5 million Euros for those who activate unregistered (Tun, 2020).

7) **PRAGUE**

Under new regulations proposed by the city's mayor, Zdeněk Hřib, hosts would have to own a home and domicile in order to rent it entirely. They would also have to temporarily vacate it for the guests. Thus, tourists would generally only be occupying single rooms while living together with the hosts. For these measures, the mayor addresses the fact that this once noble city had been turned into a *"distributed hotel"* and that Airbnb would eat the city inside-out if left unregulated. Prague's institute of planning and development records tripling numbers of Airbnb listings from 2016 to 2018, with 80% of them being entire flats (Tait, 2020).

8) **VIENNA**

The financial times records that Airbnb has affected Vienna's property market, due to the drastic increase in short-term rentals. In the centre, Innere Stadt, the number of listings increased BY 42% between 2017 and 2019 (What does the rise of Airbnb mean for Vienna's property market?, n.d.).

IV.5 CORONAVIRUS

Bloomberg raises a natural question: will Airbnb be able to survive the Coronavirus outbreak or will the short-term rental platform become obsolete (Laurent, 2020)? The Financial Times reports that Airbnb's internal valuation in March 2020 had dropped below 26 billion (Airbnb lowers internal valuation by 16% to $26bn, 2020). That is 16% lower than the previous pre-corona valuation of 31 billion, when the company was going headfirst towards its IPO. Now, the company, which has been losing money even before the crysis, is rumored to consider delaying the event. The company has a fund of 260 million US dollars set aside for reimbursing its hosts who have lost big money due to cancellations. But the hosts are not very impressed with this move. They consider that it's just a publicity stunt and the money is nowhere near enough for covering their losses. Likely, many property managers who have flexible cancellation policies won't see a dime. Guests were allowed to cancel all reservations, no questions asked, no penalties, full refund.

Some think that after the quarantine is over, Airbnb's business won't just immediately bounce back to normal. Instead, it could be that there will be a preference for more traditional accommodation, at least in the short to medium term. This could be likely due to the population being scarred by the recent events, and seeking a guarantee of hygiene standards. Airbnb hosts might be indirectly forced to re-profile their listings and possibly rent them out on the long term. Citylab reports that this is a trend that is currently developing in cities such as London, Madrid and Amsterdam, where ex-short-stays are now looking for long-term tenants in the mid-term, instead of just going bankrupt (Can Airbnb Survive Coronavirus?, 2020).

v. THE DATA

Firstly, I wish clarify what type of data is available and how I might make use of it. There are 96 markets/cities which have data that's been scraped for multiple years, up to December 2019. Generally, the data has been scraped on a monthly basis, although I did discover some gaps here and there. Still, it is pretty much complete and I can choose the data of the highest quality for my analysis. The file types can be grouped into two main categories: archived and unarchived. The archived files are large and present the whole picture. The unarchived are, shall I say, snippets of the large files, which are good for making dashboards, charts and quick on-the-go data analysis. By that, I mean that they contain the same number of data points, but are heavily reduced in features, to make the files lightweight. We have three main types of data in csv format: listings (data about a listing and its owner; text, prices and tariffs, number of reviews, location, rating, host details, etc.), calendar (each listing has its own calendar which can be updated at any time with information about available dates and the nightly prices; the lists are scraped monthly) and reviews (text data concerning comments that the listings have received).

V.1 AQUIRING THE DATA - WEB SCRAPING

Sometimes the data is handed to you, like in college. Sometimes you will have to look it up on Google, check out a few sources, maybe stop at the first one, then find the download button and click on it. And other times, you'll just need to get more creative. The way I see things, there's data to be had everywhere I look. It just needs to be extracted. The aging rings that we can visually inspect on the kernel of chopped down trees record the age of a tree. Similarly, what you sometimes need is encoded in the front-end script of a webpage. In my case, tough, it's a peculiar combination of the latter two. Inside Airbnb already has the data sorted, structured, and neatly packaged and they have download links for it. But the problem is that there are just so many cities and so many files, especially of the calendar type. And for most cities, you need to manually click to show more, then scroll and click. And for each click a Windows Files Manager pop-up appears, and I need to provide it with a name that makes it easier for me to know which is which because they are all named the same. And calendar.csv.gz(48) doesn't really ring a bell to me. And then I click on save, and then I go back to insideairbnb.com and click some more, type some more names. And I would probably become

irritated after too much of this, and would limit myself to only a few cities. But Selenium, Requests and Beautiful Soup thankfully allowed me to do all this while working on different parts of the project.

Then it went even further. I became very invested with making this script perfectly automatized and saving as much time as possible. I also thought that maybe it would take me longer to make the scraper, but I considered taking the risk. The reason was that I was inspired by many on GitHub who share their work with the community, so we can advance together. That is why I think that some of the code and annotations I've made in this work could considerably help others. The data on insideairbnb.com is cited in many journals, papers, etc. There is interest for it, indeed. I have not yet seen anyone work with more than a few cities. Typically, it is focused on one. I do not know the reason why there has been such a lack of interest in processing bigger data, but for anyone out there who would like to do grand things with this website, this scraper could be a good starting point. I haven't perfected it in time to the degree I had originally planned, but I hope it's easy enough to read and understand that other could expand it or modify it or just build something new and improved. Although I have not contacted the owners of insideairbnb.com, I believe they would approve of such scrapers, should the data be used responsibly, for the purpose of science. At the moment, it can download calendar type data, from a user specified list of cities, unzip it, pre-process it, make it smaller (by only keeping a daily average of prices instead of the price for each and every listing). This was done because these files are 400MB on average, and there are 12 of them for each year, for a total of 5 years on average. Data storage and data processing time would have been otherwise affected. The story that this data can tell is of a time series function. I have managed to plot some of it in Tableau, for starters and it could make for a very interesting insight into what actually happens all around the globe. I have also explored the possibility of predicting the prices using time series analytics in R, but the results are not worth to be noted here. The data is particularly messy and we don't have enough past observations to make any reliable forecasts, in my opinion, though I am open to suggestions regarding such possibilities.

V.2 WEB-INTERACTIVE PYTHON PACKAGES

For the purpose of acquiring data through web-scraping, I have made use of multiple Python libraries built for interacting with webpages. I will describe them in more detail below.

REQUESTS "The requests library is the de facto standard for making HTTP requests in Python. It abstracts the complexities of making requests behind a beautiful, simple API so that you can focus on interacting with services and consuming data in your application" (Ronquillo, n.d.). The name says it all. Requests is the bread and butter of web crawling. It allows us to very easily make requests to a web server and retrieve data. The connection and pooling are being made "under the hood". There is not much to say about this, other than the fact that it's a standard and a must-import in all scraping projects (Requests: HTTP for Humans™, n.d.).

BEAUTIFUL SOUP *"Beautiful Soup is a Python library for pulling data out of HTML and XML files. It works with your favourite parser to provide idiomatic ways of navigating, searching, and modifying the parse tree. It commonly saves programmers hours or days of work"* (Beautiful Soup Documentation, n.d.). Beautiful Soup is great. It came out way back in 2004 and since then it has managed to help programmers, developers, web-scrapers, and, most recently, data scientists get the most out of (especially) poorly designed websites (BeautifulSoup, n.d.). It is easy to use, the documentation is rich and it boasts some pretty powerful features, with minimal effort to set it up. The only alternative I see to this package is using Regex on an HTML "soup". But that would require the programmer much time to develop. It is able to: Convert input into Unicode and output into UTF-8; Parse via both lxml and html5lib; Customize your search almost seamlessly.

SELENIUM For me, the star of the bunch is Selenium. It helped me in the past with mass-scrapings. I've based a script on Selenium that involves IMDB user reviews of movies, shows, celebrities, etc. The script only takes as input the link to the desired "entity" and scrapes all the comments. Being able automatically click on buttons which load more content on the webpage (and implicitly more data) in the browser is such a lifesaver! The package provides easy access to an API that writes functional/acceptance tests using Selenium WebDriver. The packages have access to the APIs of Web Drivers such as Tor, Firefox, and Chrome.

TWILIO I used Twilio's free API service in order to set up an alert system that sends an SMS to my phone when the scraper has finished downloading successfully, also informing me of how much time has passed. You can use other API's to send Emails or even instant messaging such as Facebook's Messenger. I found this useful because I could go out and do some other things and it would just ease my mind to know that something is being produced while I am away.

V.3 FILE SCRAPER MODUS OPERANDI

I will try to explain the way the scraper works without going into unnecessary detail. You start with the download page, *http://insideairbnb.com/get-the-data.html*. From there we can get metadata such as how many cities there are and how many of them have a "show archived data" button to click on. The data is conveniently structured as a list of tables, or table of tables, if you will. This reveals important information about the files, which we can then use to create a very orderly and robustly structured folder and file system on our computer, based on city names, file names, file types, data compiled, etc. For this, Beautiful Soup comes in handy, because it enables us to parse the underlying tree structures that visually cascade on the webpage's frontend. Next, we need to click on the *"show archived data"* buttons, using Selenium and Geckodriver

V.4 FEATURE SELECTION

Feature selection is an important step that sets us up for building powerful predictive models. This was done in multiple steps, increasing the amount of information used in decision making. Firstly, I eliminated a large proportion of the features, based on empirical findings and "common sense".

FEATURE / RECORD SELECTION, BASED ON MISSING VALUES At this stage, I only performed the analysis on Amsterdam data, to make it easier to generalize my findings and methodology across the rest of the data. The goal was to come up with a framework for automating the repetitive data munging scenario, so that I can quickly apply it to as many cities as possible. It is, of course, best to study each and every dataset on its own and make some more calculated decisions, such a task is very, very time consuming and I need to fetch as much data as possible. Hopefully, that will even out. NaN's are always such a headache and it is really important how we tackle them. I thought about looking into which features have over, say, 30% missing values, to further decide if I should completely drop them, replace the NaNs or do some feature engineering to get past the gaps.

REPLACEABLE MISSING VALUES Text features can be engineered into text length by words. Where we have NaN's, it means we have zero words. Among these features, we have: *'neighbourhood_overview', 'notes', 'transit', 'access', 'interaction', 'house_rules'*.

FEATURES WHICH CAN CLEARLY BE DROPPED Some of these have no place in a predictive model, such as links to different web-pages. Others are ambiguous and hard to establish their true sense (neighborhood_group_cleansed). While others are just irrelevant, due to their distribution which is very imbalanced (the overwhelming majority of listings do not operate under a license). Among these, we have: *'thumbnail_url', 'medium_url', 'xl_picture_url'*

'neighbourhood_group_cleansed', 'license'

FEATURES BASED ON WHICH WE CAN DROP ROWS These features' missing values are basically zero, but, arguably, these listings are not interesting due to the fact that they haven't had clients (yet). Among these, we have: *'host_response_time', 'host_response_rate', 'host_acceptance_rate'.*

FEATURES WHICH I'M NOT SURE ABOUT The missing values can easily be replaced by zeros, but I question whether they are relevant, in the context of legal restrictions being different in the cities. They are: *'weekly_price', 'monthly_price', 'security_deposit'.*

V.5 PRE-PROCESSING

As is consistent with definitions across multiple sources, such as Wikipedia, data wrangling (or data munging) is the task of altering "raw" data such that it is appropriate for ulterior processing via machines. The intent is to make it *"more appropriate and valuable for a variety of downstream purposes such as analytics"* and machine learning/deep learning (Wikipedia, n.d.). According to Trifacta, and my own view, data munging should accomplish a series of successive (from experience, I might also add, sometimes interchangeable) goals: to discover, to structure, to clean, to enrich and to validate the findings (From Data Munging to Data Wrangling, n.d.).

Another direction I'm going in with the scraper is with the listings files. Since these are also quite bulky (around 90MB on average), and since I cannot simplify them in the same manner as the calendar data, I chose to only work with the latest scraped data for each city. This should be just fine, as most listings there have been present for quite a while. There are also fresh ones and inactive ones. But the alternative would be that I take multiple files such as this and many data points would be just duplicates. Even worse, older listings files will contain outdated information about prices, reviews, ratings and so on. The goal is to do this for as many cities as I can reasonably manage and this will help gather a copious and varied amount of data, which can be used later for machine learning, to predict nightly prices. Munging these files was a more complicated task, but I think I've managed to do so quite well. Again, I want to automate demanding tasks such as this, so I will create a pipeline which downloads these types of data and processes it to a better format before I can begin working with it. I will generalize very much. For example, I make certain assumptions about which features

are important and which deserve to be dropped. Outliers are be detected and deleted according to classical statistical methods, and not through a case-by-case standard.

I will mention a few of the data munging tasks I've done on the listing files. The prices are written with $ symbols, so they are string not numbers. $ is replaced with empty spaces using regex and then the result is turned into numeric. The date needs to be changed to datetime format so we can work with it efficiently. I drop features seen as irrelevant such as "available", "adjusted_price", "minimum_nights", "maximum_nights", "listing_id". Making sure to turn the right features into categorical variables using the *astype('categorical')* method can free up a lot of physical memory. Due to the fact that three of our sampled cities do not use Euro as their currency, and the prices are set in their national currencies, it was necessary to build a function that converts the British Pound, Czech Crowns and Danish Crowns to Euro, which is used in all others.

An interesting and promising area I could go into is Natural Language Processing. And thankfully, this database provides me with a plethora of comments which can be used for better understanding what makes a listing and/or host desirable and pleasant. As it is obvious that I am missing quite a lot of information, by not studying the pictures posted on a listing, text might give some insight which I cannot find using regular data. Another idea I had was to find complementary sources of data such as Numbeo.com which has economic/consumer data available for any city on this Globe (such as consumer prices, rent prices, salaries). This type of data would be useful in two ways. 1: It can help me give some suggestions to potential hosts who want to make a profit, by taking into account certain expenses. 2: It can help me give holiday suggestions to Airbnb guests, based on my findings from Inside Airbnb and corelated with the expenses registered on Numbeo. This data would have to be scraped. I have also looked up TripAdvisor but they specifically state that they don't grant API access for academic purposes. There are some older scrapers on Github but the website has changed in the meantime. The website also looks fairly complex, so I am not convinced if it merits the effort.

V.6 FEATURE ENGINEERING

TYPES OF LOCATIONS The way I organized these categories was by logging int an Airbnb host account and actually checking out what options are available when making a new listing. After making a list of all the available options, I grouped them into more general categories, in order to reduce the number of unnecessary features.

1) APARTMENT *Apartment, Condominium, Casa particular (Cuba), Loft, Serviced apartment*

2) HOUSE *House, Bungalow, Cabin, Casa particular (Cuba), Chalet, Cottage, Cycladic house (Greece), Dammuso (Italy), Dome house, Earth house, Farm stay, Houseboat, Hut, Lighthouse, Pension (South Korea), Shepherd's hut (U.K.,France), Tiny house, Townhouse, Trullo (Italy)*

3) SECONDARY UNIT *Guesthouse, Guest suite, Farm stay*

4) UNIQUE SPACE *Barn, Boat, Bus, Camper/RV, Campsite, Castle, Cave, Dome house, Earth house, Farm stay, Houseboat, Hut, Igloo, Island, Lighthouse, Pension (South Korea), Plane, Shepherd's hut (U.K.,France), Tent, Tiny house, Tipi, Train, Treehouse, Windmill, Yurt*

5) BED AND BREAKFAST *Bed and breakfast, Casa particular (Cuba), Farm stay, Minsu (Taiwan), Nature lodge, Ryokan (Japan)*

6) BOUTIQUE AND HOTEL *Boutique hotel, Aparthotel, Heritage hotel (India), Hostel, Hotel, Nature lodge, Resort, Serviced apartment, Kezhan (China)*

AMENITIES The procedure I use for producing these more general categories of amenities is similar to the one I used for the location types. I logged into a host account and verified the official documentation. Here, some of the categories correspond to their original Airbnb counterpart, but other I made up, as I saw fit. These are “Family” and “Logistics”, whose amenities were spread out over multiple categories to which they didn’t truly feel like belonging into. The main idea was to take a different route to what most other researchers have done, which is: one-hot encoding the amenities (all of them, or some of them, selected mainly subjectively). Instead, I would group them into a few more general categories and count how many of those amenities each listing has for every category. My approach was not entirely motivated by a need to do things differently, but I was conscious about the fact that many amenities are irrelevant and one-hot encoding all of them would unnecessarily increase the complexity and runtime and clog analysis with garbage data. Also, I don’t think anyone can just guess which amenities are good to keep or not, based on subjective human experiences. As there were hundreds of amenities, with very few of them seemingly important enough to dictate price, I decided to group them by categories of amenities, as dictated by Airbnb’s policy for hosts, while adding an ‘other’ category for outliers in the data. Then, a listing would just have a few features like Basic Amenities, Additional Amenities, etc. with the value representing the number of amenities present from each category. For example: a listing could have 12/18 basic amenities and 3/10 family friendly amenities.

1) COMMON Common amenities are essentials or things that you can generally find in almost any modern homes. They are: *Essentials, Kitchen, Air conditioning, Heating, Hair dryer, Hangers, Iron, Washer, Dryer, Hot water, TV, Cable TV, Indoor fireplace ,Private entrance, Private living room, Lock on bedroom door, Shampoo, Shower gel, Bed linens, Extra pillows and blankets, Wi-fi, Ethernet connection, Pocket Wi-fi, Laptop-friendly workspace*

2) ADDITIONAL These amenities are all very nice to have. They're not essential, but almost. They are part of what makes up some of the most common first-world utilities that people have around their house. These are: *Microwave, Coffee maker, Refrigerator, Dishwasher, Dishes and silverware, Cooking basics, Oven, Stove, Bread maker, Baking sheet, Barbeque utensils, Trash can, Free parking on premises, Free street parking, Paid parking off premises, Paid parking on premises, EV charger, Gym, Pool, Hot tub, Single level home, BBQ grill, Patio or balcony, Garden or backyard, Breakfast, Beach essentials*

3) FAMILY Family oriented amenities are all about babies, children and their safety and comfort. They are: *Baby bath, Baby monitor, Babysitter recommendations, Bathtub, Changing table, Children's books and toys, Children's dinnerware, Crib, Fireplace guards, Game console, High chair, Outlet covers, Pack'n Play/travel crib, Room-darkening shades, Stair gates, Table corner guards, Window guards*

4) LOGISTICS I made up this logistics category to include amenities that are related to guests coming and going and some of the perks generous hosts have to offer. They are: *Luggage dropoff allowed, Cleaning before checkout, Long term stays allowed*

5) HOME SAFETY *Fire extinguisher,Carbon monoxide alarm,Smoke alarm,First aid kit*

6) LOCATION *Beachfront,Lake access,Ski-in/Ski-out,Waterfront*

DISTANCE TO CENTRE This new feature was calculated using latitude and longitude coordinates. Based on a listing's coordinates I could calculate its proximity to the city centre, in km. Then I took this value and divided it, for each listing, to its respective city's urban area in km2. This indicator should supposedly show how central is a location, relative to the size of the city, and thus eliminate some of the relativity and subjectivity of scale. In order to calculate the distance to the centre, I had to use a special Python module called Geopy. According to its documentation, *"Geopy makes it easy for Python developers to locate the coordinates of addresses, cities, countries, and landmarks across*

the globe using third-party geocoders and other data sources." I found this library to be rather limited in scope but quite robust in its available capabilities. I have successfully made use of it for feature engineering. Seeing as the cities I've chosen for my sample all have various shapes and sizes, I realized that distance to centre is very subjective. 5km to the centre in small Copenhagen is not the same as 5km in London, which is the largest one. According to InsideAirbnb.com, a listing's coordinates area usually not exact, and instead, point to a nearby point on the map. That's why some listings will have coordinates that are off or will share the same location with others. Therefore, it's not entirely precise, but good enough to position it relative to the centre. To quote them: *"means the location for a listing on the map, or in the data will be from 0-450 feet (150 metres) of the actual address" and "Listings in the same building are anonymized by Airbnb individually, and therefore may appear "scattered" in the area surrounding the actual address".*

QUALITY OF PHOTOGRAPHY According to experts in real Estate and probably also common sense, the pictures that are associated with a listing are a very important factor. It's the first thing that potential customers witness. It helps build an impression about the place in mere moments. I assumed that the aesthetic and technical quality of the presentation photos really do matter a lot and can enable a host to charge more, for that premium feeling that is being emanated through the browser. Although this idea showed great promise, it also came with several problems. Firstly, how am I to get my hands-on photographs of the listings? Secondly, even if I do manage this task, how am I supposed to encode the value of a photograph such that it becomes a new feature? Luckily, both of these questions have been answered. In order to download the photos, I have built another scraper. I have the URL to the listing as a variable in the original data. From there I can download the first photograph of a listing and keep sending these requests for every link in the list. I consider that only the first photograph is necessary because the rest of the photos should be of similar quality. Metadata such as the number of photos would have also been interesting, but probably not as impactful.

I came up with this idea of using something related to the pictures before encountering mentions of this in specialized literature. After considering this possibility, I found out that I was not alone in putting my faith in this possible feature. Then, the question remained: can it be done and how? I stumbled upon an article on Medium, whose title claims that they managed to use deep learning to automatically rank millions of hotel images (Lennan, https://medium.com/idealo-tech-blog/using-deep-learning-to-automatically-rank-millions-of-hotel-images-c7e2d2e5cae2, 2018). It is about how

a couple of data scientists from idealo.de (a price comparison website based in Germany), managed to implement an application which automatically ranks images for them, based on aesthetics and technical quality. They based their work on Google's NIMA research paper, which uses convolutional neural networks. The application basically incorporates two classifiers: one for aesthetics and one for quality. The scientists from idealo.de uploaded their project on GitHub (Lennan, Image Quality Assesment, n.d.), and anyone can easily use the pre-trained models or train their own, via Docker, on their local machine or in cloud. I tried to test it on some Airbnb images, and it worked very well. The predictions seemed to be very accurate for my images, pre-trained on different data as they were.

Unfortunately, my crawler was not sophisticated enough and I managed to get blocked many times. Every attempt ends up in failure after a few hundred downloads. I've tried using free proxy lists with rotating user agents, but apparently, Airbnb has constantly updating blacklists and straight away denies all requests made like this. Even if I did manage to not get blocked, though, downloading all of these photos would take a tremendous amount of time, and it's not very practical. What I did end up doing was downloading 2000 photos for listings in Amsterdam and using those to build a new feature. Then I tested this sliced dataset with and without the photos to compare the results and found out that, indeed, this would be a top feature, with a lot of predicting power associated to it.

TEXT SENTIMENT POLARITY I have at my disposal a vast amount of text data which can be used for analysis but also for creating new features to be used in prediction. I figured that the feedback a host gets from guests must be a very important factor for setting up upcoming prices. When people search for a place to stay in, many must surely at least look at the average review scores, if not read through the comment section. I certainly do, and most people I know are the same. I ended up using an off-the shelf solution for doing sentiment classification, called Vader, which has been included in the NLTK module. Vader has its quirks and it's most certainly not the smartest tool in the shed, but, generally, it does a good job of assigning a sentiment score. Based on the polarity of the general score, a listing can go from -1 (absolute negative sentiment) to 0 (absolute neutral sentiment), to 1 (absolute positive sentiment). It turns out this is also an important feature to have, with great prediction potential. I thought that maybe the way a host decides to write a summary about a listing, might have some impact on customers. Maybe some hosts use more positive language, while others are neutral. It turns out that the new feature is of minor importance and might as well be dropped

TEXT LENGTH Given the fact that text data about a listing such as name, summary, description, etc. are available, I decided to make use of it and calculate a length of text (in words) feature for every one of these. Some seem to be important in some way in determining a listing's price.

VI. METHODOLOGY

VI.1 DATA ANALYSIS

VI.1.A DATA VISUALIZATION PACKAGES

Thanks mostly due to my mandatory assignments which I had to complete for university, I noticed a simple yet important fact: the fact that, no matter the data at hand or the goals I had in mind for it, after a messy data wrangling, it was imperatively necessary to perform a thorough Exploratory Data Analysis. The thing that annoyed me the most, especially given a relatively inhumane number of variables to study, was that I had to copy and paste many chunks of the same code just to produce visualizations. This is why I was motivated to produce a better, more hassle-free and streamlined method of inspecting the data visually. Inspired by Albert Einstein's words that "everything should be made as simple as possible, but not simpler", I took it upon myself to build a collection of functions, based on some popular Python data visualization modules. These functions would enable me to spring into existence consistent graphs typically used for the purpose of understanding data. The principle is rather simple: focus on some of the most basic plots (bar chart, histogram, box plot, scatter plot, doughnut plot, table) and write functions that produce them in the user's desired style (classic bread and butter Matplotlib, aesthetically enhanced Seaborn or interactive, JavaScript-powered, Plotly). The functions would also automatically produce titles for the charts, based on the features it detected and on the type of plot used.

Although writing the code for this library of functions was time consuming and it could have seen a much more pragmatic implementation (using object-oriented-programming instead of functional programming and calling upon a class' methods), in the end, it managed to reach its goal fairly successfully. Without quantizing it, in the long term, I am sure it managed to save me many hours of monotonous copy-paste, look the documentation up, figure it out type of activities. The great thing is that I've actually used these functions for various datasets and it's just so much easier to do EDA

with such an off-the-shelf solution at my disposal. Unfortunately, by automatic most of the tasks involving building plots, I've also critically incapacitated my ability to build the 'perfect' plots for the occasion. Sometimes I would have liked to include certain types of features in my work. But given the very limited amount of input and what that input can change in a plot, more complex charts required more traditional solutions or reliance on some other user-friendly software, such as Tableau.

STYLES

Although software such as Alteryx, Tableau and Power BI exist to empower analysts, these programs do cost money, in the form of subscription fees. To a more programmer-oriented user, their pretty GUI interfaces do not suffice, though, as they take away the freedom of experimenting with full-on code customization. Fortunately, there is a range of open-source resources we can use to make the dream happen. Regarding the range of styles that my data visualizations invoke, I will write a few words about each module used and how they contrast/interact/complement each other.

MATPLOTLIB According to its official website*, "Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python"* (Matplotlib: Visualization with Python, n.d.). This is officially the most popular data visualization package available for Python and it is the one taught in schools. It forms a building block for other more refined packages.

SEABORN One such package is Seaborn, which is briefly described on its website as, *"a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics".* (seaborn: statistical data visualization, n.d.) Seaborn builds on the robustness of the former and spices it up with improved fonts, colours and other strictly aesthetic choices.

PLOTLY A more interesting and exotic choice of styling is Plotly. It is advertised as being able to be used in production to build "enterprise-ready analytic apps that sit on top of Python and R models" in *"What would typically require a team of back-end developers, front-end developers, and IT".* The library is very powerful and varied and enables customizable deployment of pretty much any type of conceivable plot. It boasts the capacity to make "interactive, publication-quality graphs". (Dash, n.d.). The plots can be hovered with the mouse in order to display underlying information. In my opinion, this library deserves further inspection, as it looks great, the documentation is extensive and exhaustive, the support is there, it's portable and it can be a lot of fun. The only downside I've noticed is that it can get pretty slow, when dealing with large amounts of data or producing multiple

visualizations simultaneously. This quote from Clare Blessen, *"Clear, effective data visualization is key to optimizing your ability to convey findings. With various packages in use such as Matplotlib, Seaborn, and Plotly, knowing the capabilities of each and the syntax behind them can become bewildering"* (Blessen, 2019), perfectly describes my predicament. Further down her blog posting on Towards Data Science, she goes into more detail about the three different styles.

VI.1.B GEO-SPATIAL DATA ANALYSIS

As there is extensive access to geodata, I believe that much valuable insight can be gained by studying it. Regarding the status of geographical location, generally, only guests with reservations which have been confirmed can see the location's true address. Otherwise, by default, an approximate location is given, in the vicinity of the actual address (How will my listings location be shown on the map?, n.d.). The plots I will reference can be found in the Appendix section, and are referenced as A#.

1) AMSTERDAM

A2 shows the distribution of listings, colour coded to Amsterdam's own Neighbourhoods. Not really much to see here, really. The centre looks packed and we can find many listings even outside the touristic areas, although they are very, very sparse, in comparison. Green listings (A1) are way above the average price. Unsurprisingly, they are mostly gathered around the city centre. This is an unsurprising finding that becomes a veritable trend with each city studied. This visualization (A3) makes use of bubbles that represent the neighbourhoods. The size of the bubbles is dictated by number of listings in that area. The numbers represent the average price of the neighbourhood. Unsurprisingly, the centre boasts both the highest density of listings and the highest prices. The outskirts are surprisingly not that cheap, tough. Most of them are still significantly more expensive than in any other European capital.

2) ATHENS

Moving on to Athens, one can't simply be left unmoved by the looks of the infrastructure here. The "web" of roads looks impressive, leading me to think that this metropolitan area is very densely populated. What makes Athens distinct from this huge urban area are simply imaginary boundaries. Looking it up on Wikipedia, I found out that, indeed, Athens is Europe's third most densely populated city (Wikipedia, n.d.). Another city found in my project sits right at the very top: Paris. This was also obvious by studying the maps, as I will show a bit later. In any case, we see the same thing here (A4):

centre is packed, and the outskirts are spread thin. The gaps we see in the graph are due to Athens' parks and historical sites. Again, price-wise, the most expensive ones gather around the centre, with listings to the NW being very cheap, around 3 times lower (A6). Density-wise, none of the neighbourhoods even comes close to the centre.

3) BERLIN

Berlin is the world capital of clubbers. The German capital is a large city but its housing here is relatively cheap. The colour coding here (A8) is based on neighbourhood groups, rather than singular neighbourhoods, because they were way too many and it got very confusing and rather irrelevant. What is most interesting about the distribution here is that we get many very distanced listings that are really, really far from the centre, which is extremely densely occupied. Again, the gaps are due to lush areas. The most expensive area is the fancier Charlottenburg (A9). While the most popular destinations are Mitte (the actual centre) and of course, Kreuzberg is on top. It's where all the fun happens. I personally been on Airbnb trips to this city several times, and I can kind of understand the sparseness around the outskirts. The city feels massive and it takes a very long time to travel from one place to another, even inside the centre.

4) COPENHAGEN

Airbnb also seems extremely popular in Copenhagen, where most neighbourhoods show a high density, especially, of course, Indre By, Vesterbro and Nørrebro, which boast the most attractions (A10). Most of the "green" prices are in Indre By, while the rest of the central area is predominantly yellow (A11). The suburbs are mostly hot red, which very cheap options.

5) LONDON

Regarding London, the distribution of apartments follows the same pattern (A13). Prices are higher in the centre and to the southwest (A14). Westminster is the priciest and most dense (A16). If we zoom in on it, the listings are practically stacking one on top of each other (A13). Interestingly, prices are 2-3 times higher on average in the centre than in the outskirts. Worth noting in London and this applies to all cities: the most reviewed listings are overwhelmingly in the city centre (A15).

6) MADRID

In Madrid we can observe that Centro is incredibly dense (A17), all the other neighbourhoods paling in comparison. There is actually no comparison. We have here over 10k listings, and the second largest one, the Salamanca neighbourhood with its almost 1.5k does not even come close (A20). Although, interestingly, Salamanca is on average twice as expensive, probably being a fancier

neighbourhood. Indeed, according to Google and Promora.com (MAP OF THE MOST EXCLUSIVE NEIGHBOURHOODS OF MADRID, n.d.), Salamanca boasts high end restaurants, designer boutiques, luxury gourmet food markets and the most coveted buildings in the capital. But there is an even more expensive neighbourhood group: San Blas – Canillejas, even though it's fairly far. I couldn't find an explanation for this, tough.

7) PARIS

On to the city of love and romance: Paris. From my findings, the French capital's Airbnb listings are, on average, the furthest away from the city centre. If we inspect the geo-data and chart it, we can understand why: the whole surface of this city is packed with Airbnb apartments (A21). Even the outskirts are full, and their density is high enough to be compared with the city centre of other capitals. It is incredible what is happening here but it's really less surprising when we think about the fantasy of Paris that we've all been raised into. If we look at the sizes of the bubbles (A23), we don't find the usual contrasts in scale. And not even in average prices. As can be seen, not all marginal areas are cheap. Some are on par with or even more costly than downtown. Checking it up on Google Maps (A33), I see that many of the major attractions are spread everywhere. The Eifel tower and Arc du Triomphe in the west explain the high value of Passy and Elysee. Paris becomes the exception to some of the rules, due to its historical landmarks. We can also see that the city is littered with pedestrian areas and points of interest, marked with light yellow on Maps. Unsurprisingly, if we look at the prices, the centre and especially the vicinity of the Tower and Arc are dark green (A22).

8) PRAGUE

As shown on Google Maps, the city's most attractive areas are all in the centre, to the left and right of Charles' Bridge. Although, on the map, the city seems large, the outskirts rather resemble villages or suburban areas, with apparently not much to do. This is why we notice this extremely high level of sparseness (A24). The de facto centre, Praha 1, is way denser than the other bubbles, and has the second highest prices, after Praha 8, which I was surprised to see was so costly, given the fact that it is far (A26). But, of course, we see most green dots concentrated near Charles' Bridge and for good reason (A25). The staple area is superb but very touristy (A34). I would not be surprised if there were more tourists there than actual Czech Republic citizen homeowners.

9) ROME

Next up we have the ancient capital of Rome. Rome is interesting due to the fact that it also opens up to the Tirenian sea, with Lido di Ostia, a popular seaside resort, bustling with tourists bathing in the

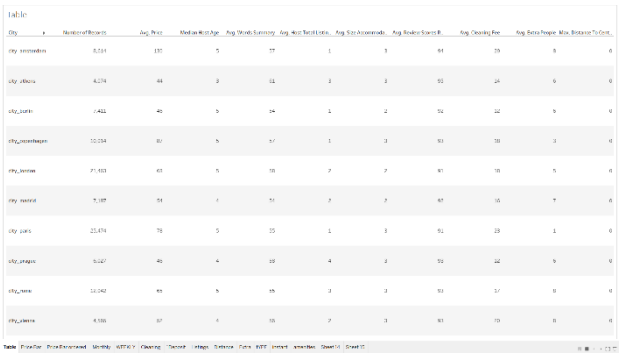
waters, eating fine Italian cuisine in traditional restaurants and tanning on the golden sands. Between this and Rome there is a large lush area. This is why, besides the general concentration in the centre, we see a unique second cluster in Lido di Ostia, although not officially part of the city premises (A27). Even so, the resort is twice as cheap as the centre, on average and has less than 700 listings, compared to the centre's impressive 16.5k which towers over all other neighbourhoods (A29). San Giovani, the second most populous one, only comes at 2.5k.

10) **VIENNA**

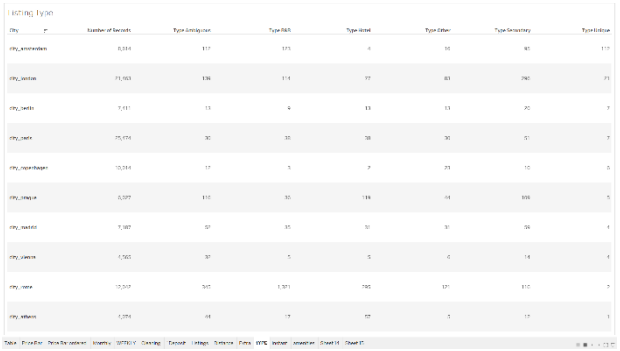
Next stop is Vienna, Austria. The listings focus on the historical centre of the city. The Innere Stadt district, the historical city centre, bustling with luxurious cafes and shops (A35) and some of the city's best attractions, has the highest average price (A32). It actually holds the highest average price of any European capital, out of the ones I am studying, at 426 € per night. It is indeed very "green" (A31) but this figure makes me think there might be something awfully wrong with the data, or we simply have a trend of luxurious accommodation going on here. According to A32, Hietzing is the second most expensive area, I think due in part to its close proximity to the famous Schonbrunn palace and the Zoo. Penzig, to the north, is also very close to the palace, but spreads out further, with more listings, probably bringing the average down. Other than that, there are Airbnb places scattered all over, similar to Berlin or Prague, but they're much cheaper. This surely is due to the fact that most attractions gather in two hubs: Innere Stadt and the Palace.

VI.1.C LISTINGS DATA ANALYSIS

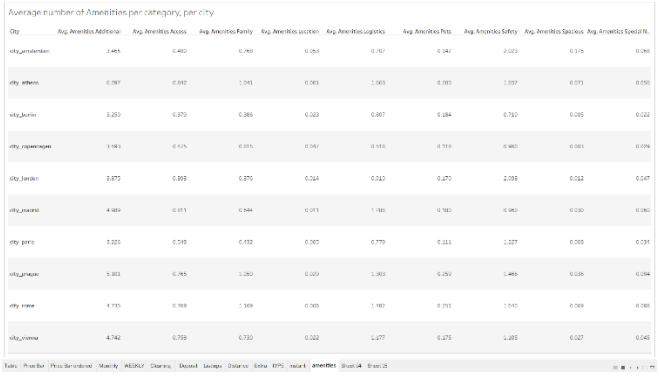
The listing files offer the highest amount of information on the general characteristics of the Airbnb reality in each and every city. By analysing the data visually, we can get a very good idea about what brings these cities together and what sets them apart. We can also develop more hypotheses, or verify existing ones. We can correlate what we find here from information elsewhere.



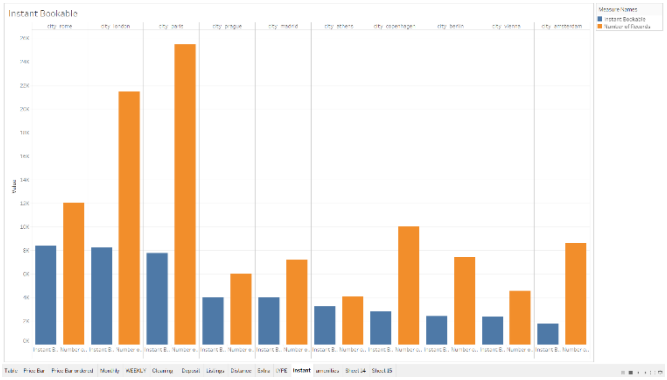
This table to the left showcases some interesting information. The ‘youngest’ hosts are in Athens, which seems to show that this European capital joined the Airbnb hype a little bit later. All cities show the same average summary description length, meaning they also probably state the same things, English being the standard language. Prague, Rome and Athens rank highest in terms of hosts with multiple listings, revealing that Airbnb in these cities might be more of a business. London and Paris have the lowest review scores. It’s probably due to the value for money principle. For example, London does accommodate a lot of its guests in what could be considered sub-standard locations. It can also be the effect of the city itself. The “Paris Syndrome” is notorious for bringing people to disappointment and disbelief after actually arriving in the city of their dreams. Overall, the guests have been most pleased with the Athens listings. The fact that the hosts are “fresher” and more oriented towards maintaining businesses might play a role in this. Plus, Greece is known to be a tourist destination that is popular with most anyone.



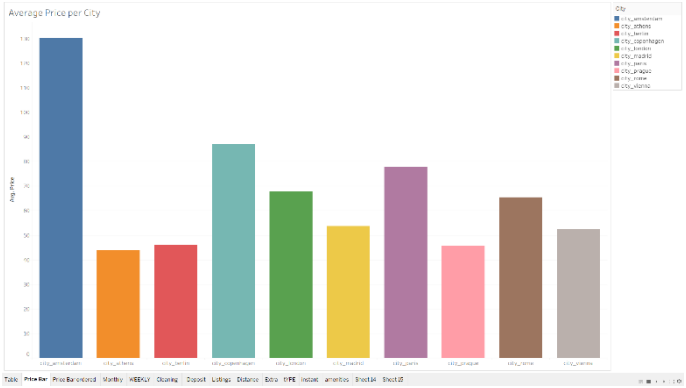
According to the table on the right, Rome is home to the most ambiguous type listings (listings which have not belong to any standard category; may be due to the fact that it’s the only city with access to warm beaches). Rome also dominates the Bed and Breakfast type of listings, by far making me think that Italian hosts like to offer authentic experiences). Again, the same is true for hotels and ‘other’, which are also dominant on the Roman market. Secondary types of accommodation are most common in London, which leads me to believe that many Londoners find advantageous to monetize any space available, in the famously crammed city. Amsterdam boasts the most ‘unique’ accommodations (such as boats, which would explain the fact that “boat” appears in the word cloud for the city’s comments section).



In terms of amenities, Athens has the highest average number of additional offerings for its guests, and also family-friendly ones. Prague also boasts the most family-oriented amenities. Amsterdam and Copenhagen have the most interesting location features. Probably due to the beaches, lakes, boats, canals and so on. Prague and Rome are the most open towards pets. Amsterdam is first in terms of safety features and is also the most spacious. This might be due to the city planning and local architecture. Prague is the best destination for guests with special needs, as the hosts there seem to be the most equipped in this sense.

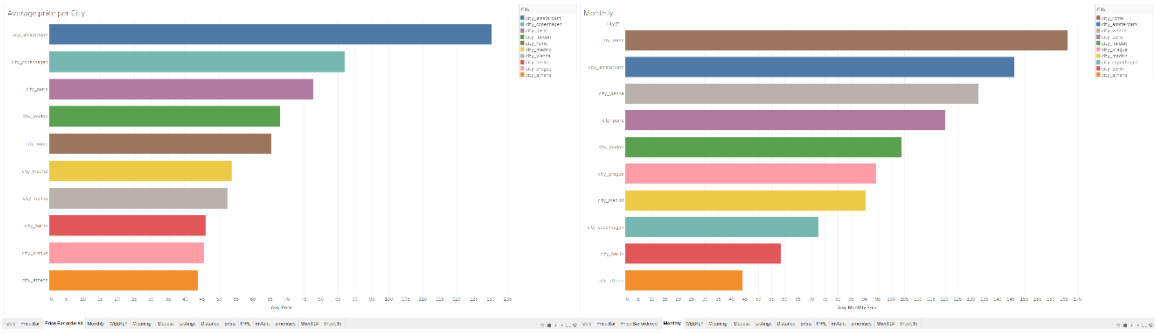


Here we can see that cities such as Athens, Madrid, Prague and Rome have the highest percentage of instantly bookable options. It makes sense, if you think about the fact that I earlier stated that Athens, Prague and Rome have some of the largest percentages of multi-listing hosts. These types of hosts run a business and it's advisable and desirable for them to be able to instantly accept requests, so that they're always booked to their maximum potential. In cities such as Amsterdam, Copenhagen, Paris and London, there are very few guests who make use of this option. I believe that this is mostly due to the fact that these cities have imposed certain restrictions. Such as: limiting the number of nights a certain type of listing can be booked yearly; having the host be required to register the location, if a set of conditions are met.

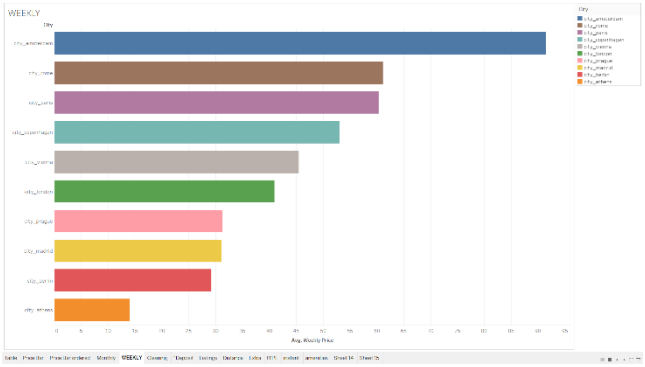


The North is indeed an expensive place to live in. It seems that Amsterdam and Copenhagen are highest in terms of pricing. Amsterdam towers at almost twice the size of the Danish capital bringing it to its own league: Tier 1. Copenhagen is closely followed by Paris, London and Rome as the Tier 2 cluster is complete. Going down to Tier three, we have Madrid and Vienna which are almost the same, then

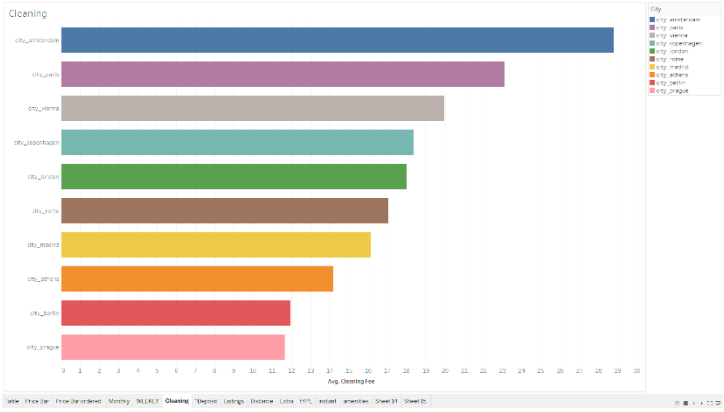
Berlin, Prague and Athens as the cheapest destinations. Here are the average nightly prices of cities, ordered in descendant fashion.



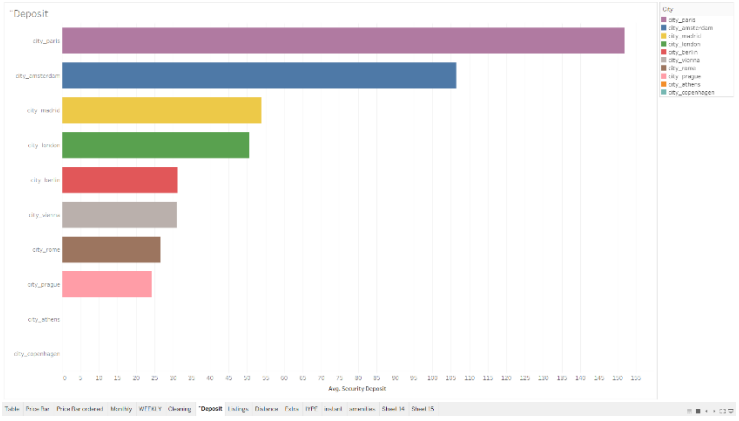
Rome, Amsterdam and Vienna have the highest monthly fees, while Athens and Berlin the lowest. But this chart does not tell the whole story. The averages here are calculated using all of the data points, many of which are zero. So, it doesn't actually show us who charges the highest monthly fees. For that, we would need to calculate the averages, only taking into account those listings that do have these prices set up.



Amsterdam has the highest weekly prices. Rome and Paris follow suit with identical prices. Athens and Berlin are, yet again, the cheapest. But the same story applies here: many don't have the option. For example, Copenhagen climbs the poll enough to reach the top this time. I'm sure this is due to the fact that not many people would rent a location for a month or more, but enough of them do have weekly offers set up, with their already high nightly prices.

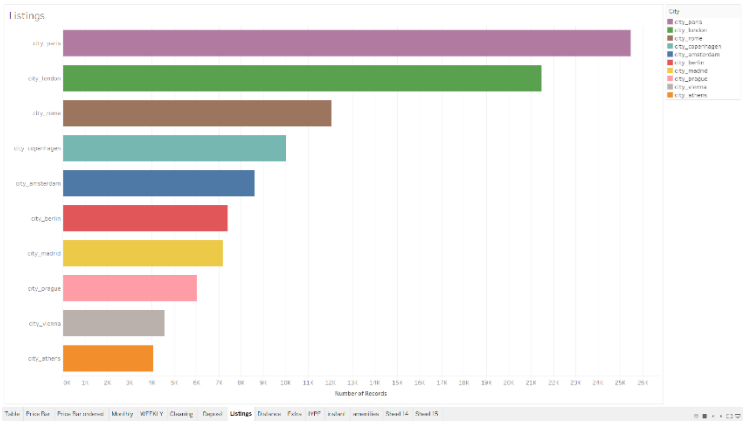


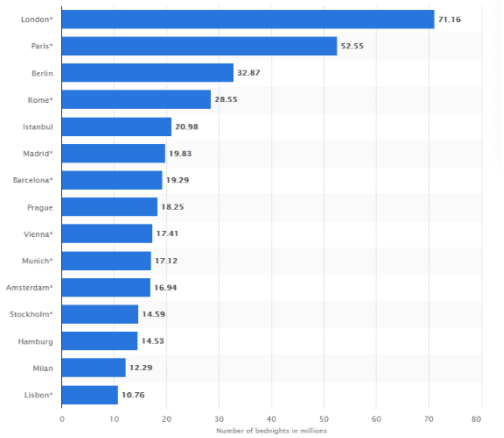
Amsterdam, Paris and Vienna rank highest in cleaning fees, while Prague, Berlin and Athens lowest. This chart is a bit more inclined towards reality. Most Airbnb listings do feature cleaning fees, which is why this top here resembles the nightly price chart more closely.

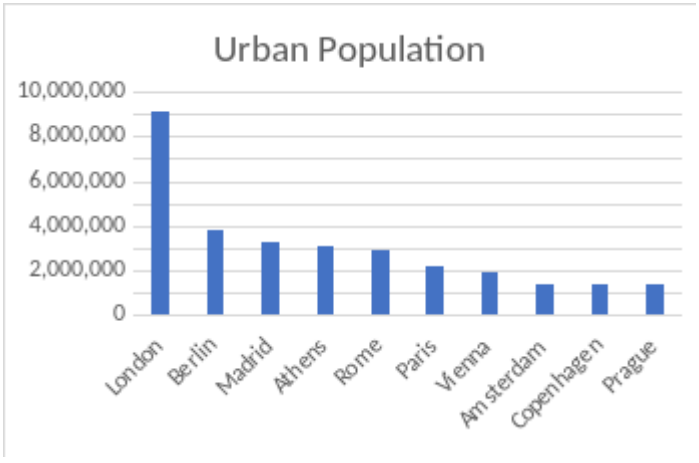


Athens and Copenhagen effectively have no deposit requirements. None of the hosts here thought of charging one. The largest deposits are payed in Paris and Amsterdam, which are some of the most expensive cities anyway. The fact that Paris here is almost off the chart, leads me to conclude that uniquely large proportion of its hosts do have this pretention. And it does hold spot #1 in terms of the number of listings it provides. When you book, not only do you pay for your stay, but also a guarantee of good behaviour.

This bar chart showcases the total number of listings per city, which gives insight into the structure of data. Some cities are overrepresented while others represent a niche minority. Is this affecting our analysis and our models' powers of prediction?

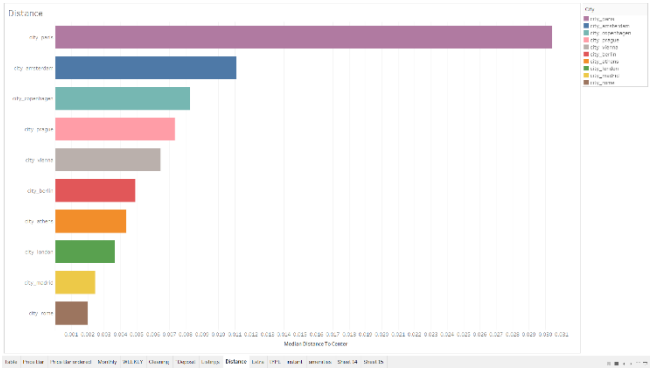




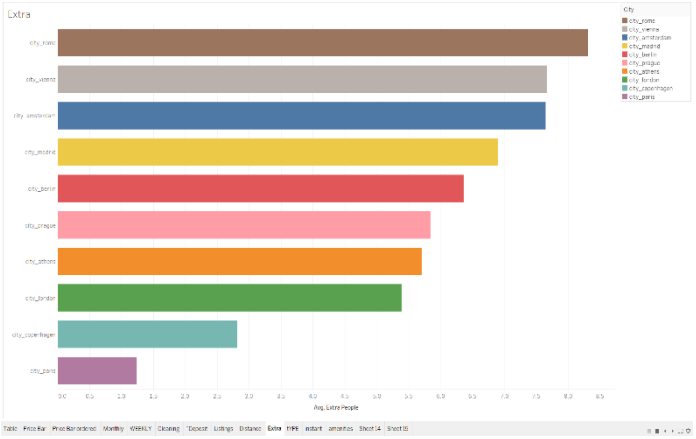


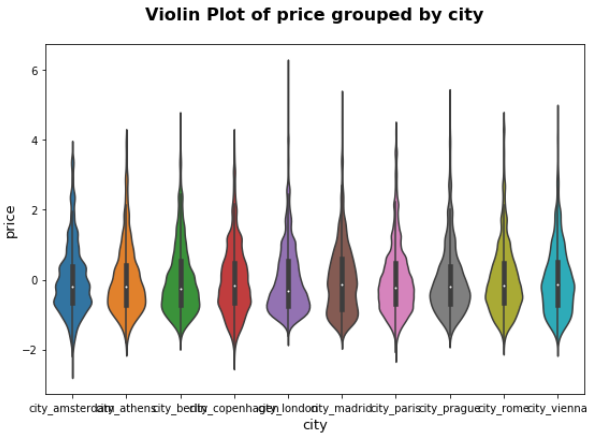
Let's see how the number of listings stands next to these cities' populations and yearly tourist turnover. According to Statista, from which the two charts above were copied, this is the top of the most visited cities in Europe (Statista, n.d.). Almost all of the cities I've selected, apart from Copenhagen and Athens, make it to the top. London and Paris have over 50 million annual visitors. Paris has a higher ratio of listings/tourists tough.

Interestingly, Paris' listings are furthest away from the centre, by far. Rome, Madrid and London's (surprise) are the most central. This peculiar anomaly can be explained if we look closely at a map of Paris, marked with the location of all of its listings. There's practically no room to throw a toothpick in that city without hitting an Airbnb. The apartments in the centre are not much denser in their distribution than their outskirt counterparts. Sort of the same thing can be said about Copenhagen, which, although not a large city by most standards, has its listings quite far from the Central Station, because they are pretty much evenly spread out.

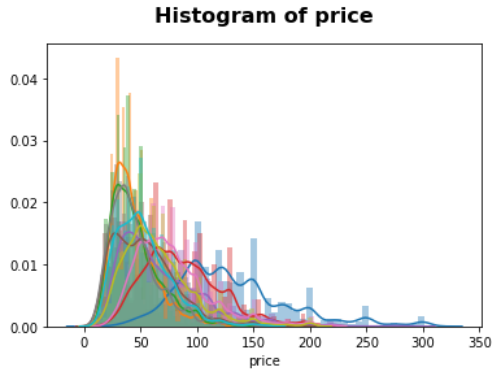


On average, the Romans charge the most for bringing extra people to the accommodation. In Paris, we could say that it's almost obsolete. Copenhagen also doesn't care much about charging more for extra people. I believe these to be economically cultural differences, dictated by how the locals regard products and services.





These violin charts showcase the distribution of normalized prices in all the cities. London (purple) and Prague (grey) have the longest tails, meaning they have the most outliers or some of the priciest listings, which do no fit the norm. Specifically, for these two, judging by the violins' broad base, the majority of listings are actually very cheap. Amsterdam's (blue) shape is very irregular, showing that we have well established price categories.

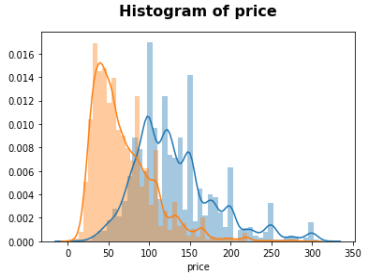


This chart shows the histograms of all the cities on top of each other. As can be seen, most distributions are right skewed, with the bulk of the values gathered towards 0. Most prices seem to be under 100 euros/night. This is just another way of viewing the violin plots or vice-versa.

After studying their histograms and box plots separately, I observed that some cities share certain similarities. Blue is the first city while yellow is the second city evoked.

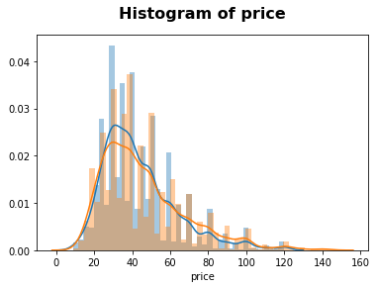
1) Amsterdam vs London

This is not the case for Amsterdam and London, which are both unique in their own way and should be studied separately. They only share a price range from 0 to around 350 euros.



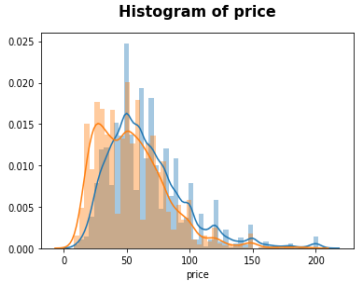
2) Athens vs Berlin

Athens and berlin are some of the cheapest cities. Their distributions are almost identical.



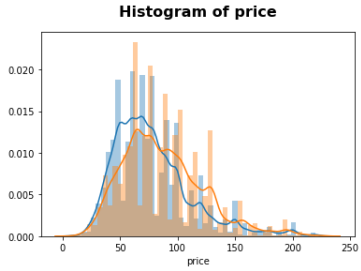
3) Rome vs Madrid

Rome and Madrid are also very similar.



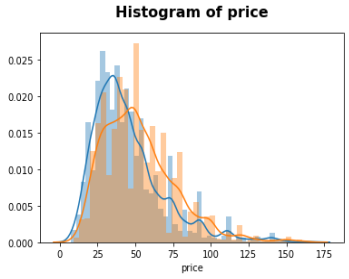
4) Paris vs Copenhagen

Copenhagen seems a bit more expensive than Paris, on average but other than that, similar.



5) Prague vs Vienna

The other two most cheap cities are Prague and Vienna, with Vienna tending to be a bit more expensive.



After this realization, I thought that it would be interesting to see if grouping cities two by two and training models like that would offer better results. Grouping cities like this could also let me include Neighbourhood as a feature and one-hot encode it, without adding too much complexity.

VI.1.D TIME SERIES DATA VISUALIZATION

I chose not to include the year 2020 in the scope of the analysis due to mainly to two major reasons.

1) MEANING OF THE CALENDAR DATA

Without a proper explanation officially offered by the website, and practically absent from the works of other researchers, I am forced to make a few of my own assumptions. A host's calendar function on Aibnb's website is not a static spreadsheet that's completed once in a while and it's set in stone. The price of a listing can realistically vary as often as the host can manually update it. The calendar files represent just the data scraped in that scraping session, whose date is recorded on the site. Approximately every month (in generally a different day of the month), the owners of the website upload the updated data frame, which varies in size according to the number of newly appeared listings and also the ones that get deleted. The calendar records the next calendar year from the date it was scraped, which means that if the ones that I use are scraped in December 2019/January 2020, those 2020 "future" prices have not yet been materialized.

2) THE CORONAVIRUS CRYSIS

COVID-19 has notoriously disrupted the global tourism industry. Both the hosts and the guests suffered heavy blows to their 2020 projections. The European Union was particularly affected by the viral crysis. Airbnb itself was practically shut down. Guests had to cancel their trips, money and experiences were lost in the process and the recorded data for 2020 effectively became even less relevant than it was before. Prices have suffered unexpectedly, surely. It will be very interesting to redo this analysis again next year to record what happened.

I would rather exclude Amsterdam from the bulk of the time series charts because its average pricing is so relatively high that it becomes very hard to interpret the rest of the cities. This is also due to the fact I am trying to cram a lot of lines, with many colour codings, which become harder to distinguish. I know from my Data Visualizations class at CBS that it is advisable to use a maximum of 5-6 categorial variables in plots, but I will try to do my best.

AVG. PRICE / YEAR (A36-A42) Unfortunately, there is a large discrepancy in the availability of the data. Some cities have only recently been added to the insideairbnb.com's database. The same can be said about the quality of the data, which can pe peculiar at its best and downright shockingly doubtful at its worst, as we'll soon see. Prague only begins in 2018 and Rome in 2017 (A36). London and Berlin follow a similar yearly trend, that is slightly lowering its prices, after initially absurd

values, if we look as back as 2015. Paris and Copenhagen make another pair, with a large increase in 2018, followed by a correction in 2019, but still on the rising trend. Madrid's prices soared from 2016 to 2017 by about 80%! Then, a slight correction, then a rise, and then soaring again from 114 € to over 230 €, which, already begins to look ridiculous. I wonder if this is right or not. Prague experiences a very steep linear growth, with projections for 2020 (A38) making it amongst the most expensive cities in Europe, above the likes of London, Rome, Paris and Copenhagen. Madrid and Athens also experience drastic growth for 2020. This makes it clear for me again that the choice to not seriously study the 2020 data was wise. Prices in Vienna are very stable and they practically follow the €75 line in the background. If we look at the full picture, with Amsterdam and all of the available years included, it looks even more menacing. Amsterdam could even be included in the other charts, if not for the 2018 and 2019 values, which are, quite literally, off the charts high. Not to mention that Paris and London also recorded extremely high values for 2015, which were halved the following year. I suspect that whoever did the scraping for these cities in those times, did so by having the prices in some other currency. Although their official currency would have still been the Euro, the user of Airbnb can choose (or is usually prompted) to use their local currency, which could be anything, I guess.

AVG PRICE / QUARTER (A43) Madrid, London and Paris are more expensive in Q2 and cheaper in Q3. The latter two are in the same price range but if you're looking to visit either of them during Q3, go for Paris. It's much cheaper. But whereas these two get even cheaper in Q4, Madrid gets more expensive. Rome is also very similar to Madrid, albeit a tone more stable overall. The rest of the cities seem to start out cheaper and progressively increase prices until they reach New Year's.

AVG PRICE / MONTH (A44-A46) Now I can see much more. Quarters don't tell us that much in this case. We're not running a production line here. London has the most dynamic pricing, with very interesting movements. We have three peaks: April, June and August. May and September are lows. Instead, May is so much more expensive to spend it in Paris! Rome and Madrid follow very similar movements, with Madrid having a distinctive spike in April which gradually comes down to its low in August. Rome's peak is in May and then it's basically the same movement. Interesting to see the prices getting higher for December, which I suspected, due to Christmas and New Year's making them popular tourist destinations. Such is the case with most of the cities, where the holidays are an occasion for city breaks. After April, Prague seems to be in full season all year, followed by a dramatic increase in prices for the end of the year. January and February are practically dead months for this city. The rest of the cities don't see much fluctuation. Not much seasonality to be detected here. Prices

are fairly constant, with subtle ups and downs, but they don't get ridiculously cheaper or pricier all of a sudden. Amsterdam is Also steadily dying until April and then has a violent upwards burst until June then slowly comes down until November, which is still much higher than spring, then sees another serious increase for the holidays. If you really want to see these cities but are on a budget, consider booking your holidays somewhere in January-March, when most cities have prices that are so low compared to season, that they might warrant some consideration.

AVG PRICE / WEEK NUMBER (A47) From studying the evolution of price based on week numbers, we can see that Madrid actually becomes the third most expensive city during week 16, which is Apr 13 – Apr 19. Some like week 22 and week 27 also produce notable spikes. I can see around 6 of these shark-fins across the year. I am guessing they might be related to some sports events or Spanish traditions. Now we can see that during weeks 33 and 34, London and Paris are practically equally priced. But whereas London is peaking, Paris is just recovering from its absolute low. The way the prices go up in Prague, right before Christmas, is shocking. Rome, Berlin and Athens' last weeks of the year are actually cheaper. These are after Christmas.

AVG PRICE / WEEK (A48-A49) Now things start to get interesting, when we go deeper, into weeks. I will exclude from my conclusions the years of 2015 and 2016, because we see some absurd movements in there. But what strikes me every time I look at this chart, is the unexpected similarity in peaks in many cities, precisely on Christmas Eve 2018. All the hosts in these cities decided to just majorly increase the price at the same time, which would be rather boring, if not for the fact that it happened in this year specifically. Was the economy doing great back then? After this particular point in time, we can see Paris, in red, and Prague, in green, having practically completely opposing trends. Prague soars and London crashes its prices. Would Brexit explain London's downfall? If we also look at the rest of the cities, it becomes rather evident that they are all either: stagnating, increasing at a slow rate or booming. If I look at Christmas 2017, the same thing was happening, but a less grandiose scale, which makes me think that there is also an underlying feature influencing prices. It might be Airbnb's own financial situation or its popularity. Or, it might be the state of the economies or the EU's or the global economy as a whole dictating these trends.

AVG PRICE / DATE (A50-A52) If we plot the daily change in prices with no filtering and just include any year and any city, we obtain a piece of abstract art. There are some movements here that look perplexing to me. Madrid looks very strange from 2016 to 2017. Actually, Madrid looks particularly strange in in its entirety. The prices fluctuate so much. Much more than any other city. A

possible exception is berlin, which after that Christmas back in 2018, shows some impossible patterns. At this stage, I am beginning to seriously question the data. I actually went through the process or re-downloading and re-processing the data and taking baby steps to monitor that it wasn't a fault on my side, but, unfortunately, this is what I've got to work with. And I don't currently have a plausible explanation for what's is going on here.

AVG PRICE / WEEKDAY (A53) As expected, weekends are more expensive; nothing else to see.

AVG PRICE / DATE (WEEK) (A54) It is just like avg. price / date, but the visualization eliminates unnecessary details, so we can better see the movement of price in time. I think the avg. price/date(week) chart is easier to follow. We can clearly see when the ups and downs were, over time, overall, across these European cities. Those 2018-2019 holidays saw the most expensive prices, and this last holiday was almost as expensive, the same as the one in-between. The key difference is that the projections for 2020 were all very high, in all cities. If these values were actually representative of the host's true intentions, then 2020 would have been a great year for Airbnb, which was set to go public. Sadly, the current crysis puts all of these conclusions under a shroud of uncertainty.

OBS: It would be very interesting to do this analysis for all of the cities and establish global, continental or even regional trends.

VI.2 NATURAL LANGUAGE PROCESSING

Although the majority of the data available is structured in some way, shape or form, and can be readily interpreted by a computer (or, with small adjustments before processing), some of the data is not. We have unstructured data, in the form of text. Lots of it, and the zero-hypothesis that we can evidently make at a first glance is that some of it holds some value. The question is, what information can we extract from it. What can it tell us? Enter Sentiment Analysis. Sentiment Analysis is sometimes also known as Opinion Mining. It is one the sub-fields of Natural Language Processing that "tries to identify and extract opinions within a given text. The aim of sentiment analysis is to gauge the attitude, sentiments and emotions of a person, given text. And it does this without taking into account semantics.

This is a task that is trickier than first expected. Even as humans, we can sometimes find it surprisingly hard to gauge the emotions of other fellow humans, when they are communicated via text. There is

no tone to be heard, no inflexion to be noticed, no volume, no posture. Sarcasm is a huge issue. Many times, we don't even realize when someone is sarcastic (saying something but meaning otherwise). Many texts are also pluri-polarized, meaning that we can find an array of emotions being displayed. For example: the movie review "The Lord of the Rings: The Return of the King Extended Version was a fantastic experience, although it could have been shorter. The cast was on-point but I was disappointed with the lack of diversity." The sentence expresses positivity (fantastic experience, on-point cast) but also negativity, a desire for more (although, but, disappointed). Even I can't precisely say if this imaginary cinephile was more pleased or turned off by the movie.

That's why machines are not expected to perform close to 100%. But at the same time, one cannot expect humans to manually label such data (taking into account duration and even subjectivity). Enter Vader: a surprisingly robust and easy to use off-the-shelf solution for performing Sentiment Analysis on texts, particularly social media, which it has been pre-trained on. VADER stands for *Valence Aware Dictionary and sEntiment Reasoner***.** Its logic is based on lexicons and rules. It was specifically designed to work with social media. The lexicon features, i.e. the words, have been labelled in accordance with their respective semantic orientation, which is either negative or positive, with neutral in-between.

According to its own dedicated research paper, "VADER performed as well as (and in most cases, better than) eleven other highly regarded sentiment analysis tools" (Hutto & Gilbert). Analytics-Vidhya has a good blog-post that I would recommend on it. It speaks of its advantages over traditional methods, such as:

"

- *Works exceedingly well on social media type text, yet readily generalizes to multiple domains*
- *It* ***doesn't require any training data*** *but is constructed from a generalizable, valence-based, human-curated gold standard sentiment lexicon*
- *It is fast enough to be used online with streaming data, and*
- *It does not severely suffer from a speed-performance tradeoff.*

" (vaderSentiment, n.d.)

Its Github portal mentions that it is fully open-sourced under MIT, and has now been successfully incorporated into the core of the NLTK package, which makes it even more convenient, as it does not

require further setting up. The package delivers results via polarity scores. The most useful one, the compound score, "is computed by summing the valence scores of each word in the lexicon, adjusted according to the rules, and then normalized to be between -1 (most extreme negative) and +1 (most extreme positive). This is the most useful metric if you want a single unidimensional measure of sentiment for a given sentence. Calling it a 'normalized, weighted composite score' is accurate."

Before moving on with my feature engineering, I wanted to play-test with the package and my actual data, to see if it has any quirks or generally delivers good results. It is worth noting that I have found some unsurprising pit-falls, but, in the end, I was easily convinced by the ease of use and "fair-enough" results. Obviously, not all the comments are going to be in English, unfortunately, which makes things a bit more complicated. Fortunately, there are workarounds, without the need to drop comments that come in other languages, which amount to over 26% of the total. We can use Nakatani Shuyo's language-detection library: langdetect. Langdetect is available on *https://github.com/shuyo/language*-detection and can be installed easily with pip. After detecting other languages, we can make use of Google's API (via googlegtrans' Translator) to translate them according to the state-of-the-art standard. After translation, we can begin to remove stop-words, punctuations and to basically normalize the corpus and make it suitable for NLTK. Unfortunately, we have lost a lot of data points again, due to missing values. For Amsterdam, for example, 1253 out of the original 9962 were removed, leaving the city with a new total of 8709. The pre-processing phase specific to the review files was again generalized and incorporated into a single script that does it the same for any city. This part of the project was heavily inspired by the work of Berliner and Kaggler Britta Bettendorf (Bettendorf, 2019). She submitted a good NLP kernel on Kaggle, from which I learned about Vader and decided to use part of her work to meet my project's needs. As her work was done on Berlin, and mine on Amsterdam, I will use both our visualizations to explain some interesting differences and similarities between the cities.

VI.2.A SENTIMENT POLARITY SCORES

A reason why it really pays off to translate the comments made in other languages is because Vader only works well with English input. Let's take a sample review from the Amsterdam dataset and its English translation and apply Vader to both variants and see what happens:

Comment in French:

*Peniche hyper confort, chaleureuse, bien équipée (et bien chauffée). Filip est attentif aux détails, et de petites attentions témoignent de son sens de l'accueil. Le quartier est tout proche du centre et de la gare. Avec les vélos mis à disposition c'est idéal. A conseiller! {'neg': 0.041, 'neu': 0.959, 'pos': 0.0, 'compound': **-0.2003**}*

Same comment in English:

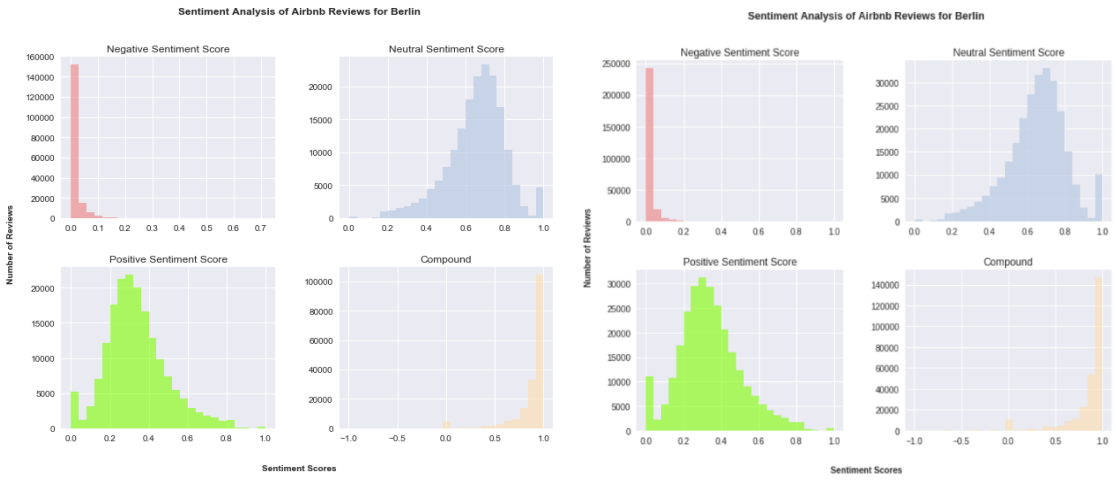
*Super comfortable, warm, well equipped (and well heated) barge. Filip is attentive to details, and small attentions testify to his sense of welcome. The area is very close to the center and the train station. With the bikes available it is ideal. To advice! {'neg': 0.0, 'neu': 0.649, 'pos': 0.351, 'compound': **0.9583**}*

Obviously, the guest was very much content with the booking and this review is clearly positive. Powerfully so. But in French the same comment is seen as pretty negative.

VI.2.B DATA VISUALIZATION OF POLARITY SCORES

The chart illustrates the histograms of each polarity score. I wish to mention that I just assume that these distributions are very similar across all cities. Can you spot the differences?

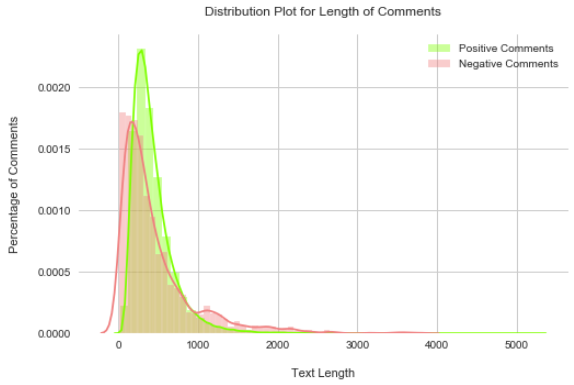
AMSTERDAM BERLIN

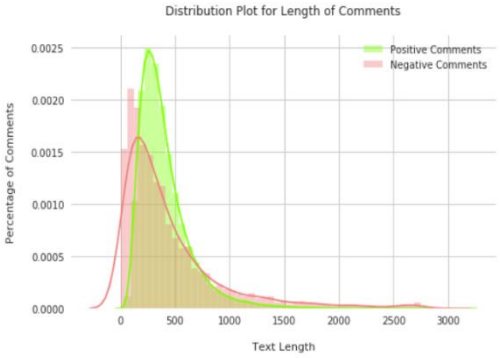


The compound score histogram heavily resembles the one of the review scores. Most guests seem to highly appreciate their stays. Negativity is generally very low, neutrality is relatively high, and

positive sentiments have a somewhat more neutral distribution. The curios thing is that the visualizations produced for the two cities are almost identical.

AMSTERDAM BERLIN

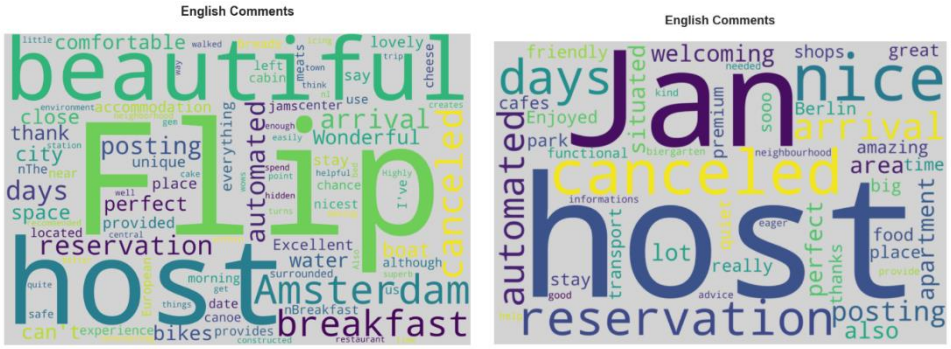




As Britta puts it, “The mode for the text length of positive comments can be found more to the right than for the negative comments, which means most of the positive comments are longer than most of the negative comments. But the tail for negative comments is thicker.” The same interpretation also applies for the comments in the Amsterdam dataset. It is due to the striking similarities between these two cities, that I have assumed that I won’t find anything of major interest by producing the same plots for all the other cities, as they would probably be very similar. Even tough that in itself is a remarkable fact, it does not constitute purpose of the project to do so.

Apparently, people generally choose to comment on the host and the city. They talk about the space, the centre, surroundings. Otherwise, the word-clouds are rather different and we can see some very peculiar and unexpected words appearing such as: flip, jams, meats, cheese, hidden, date, canoe for Amsterdam and “sooo”, ”jan” and provide for Berlin. Honestly, these visualizations don’t really tell much of a story, are hard to read and are clearly erroneous. I know for a fact that the word “cancelled” will appear in many comments because it’s automatically posted by Airbnb when a host cancels the reservation.

AMSTERDAM BERLIN



VI.3 PROFITABILITY IN COPENHAGEN

Airbnb claims that the average Copenhagener earns up to 14.400 DKK per annum, which is around 2.000 euros, half of what is earned in Amsterdam, and reflected indeed in the average prices. Airbnb also states that 94% of Copenhagen hosts share their home (Airbnb, n.d.). It also claims that Airbnb has an insignificant impact of housing prices, because *"A typical unit of housing in Copenhagen would need to be shared for up to around 246 nights per year on Airbnb to be more financially attractive to its owner than a long-term rental"*, which I think is a very bold statement that doesn't hold up and can be debunked very easily. What I am about to demonstrate is not 100% factual, and it's based more on a personal knowledge of the housing market in Copenhagen and on my experience as an Airbnb host. Also, I am basing my calculus on the information presented by SKAT. You get a deduction of 28.000 DKK on your renting income (for your main home) and the rest is being taxed at a 60% rate (SKAT, n.d.). Let's do a quick math problem. Let's say you own an average 2-room apartment in central Copenhagen and you do not pay rent. I have done the math for three different scenarios where you rent for 246 nights:

A) Nightly price = 500 DKK. Your annual brute income is **123.000** DKK. Your taxable income is 123.000 – 28.000 = 95.000. You pay 60% of that in taxes, which is 57.000 and your net earnings become **66.000** DKK = 28.000 + 38.000.

B) Nightly price = 400 DKK. Your annual brute income is **98.400** DKK. Your taxable income is 98.400 – 28.000 = 70.400. You pay 60% of that in taxes, which is 42.246 and your net earnings become **56.160** DKK = 28.000 + 28.160.

C) Nightly price = 300 DKK. Your annual brute income is **73.800** DKK. Your taxable income is 73.800 – 28.000 = 45.800. You pay 60% of that in taxes, which is 27.480 and your net earnings become **46.320** DKK = 28.000 + 18.320.

The 500 DKK price is what I would consider average for a shared space in a home. Then I went as low as 300, as a worst-case scenario. Now let's see what happens if you permanently rent that one room for 5000 DKK per month, which is pretty standard in this city.

D) Monthly price = 5000 DKK. Your annual brute income is 60.000 DKK (renting all 12 months). Your taxable income is 60.000 – 28.000 = 32.000. You pay 60% of that in taxes, which is 19.200 and your net earnings become **40.800** DKK = 28.000 + 12.800.

This is way lower than even the most pessimistic scenario with Airbnb. To get better results by comparison, we would have to lower the nightly price below 300 DKK which is only representative of very few listings. But let's not get ahead of ourselves. As a welfare state, Denmark boasts very high taxes and they don't just target income. There is also Value Added Tax which applies to commercial activities. According to SKAT, *"Currently, in Denmark, you are required to register for VAT if your turnover exceeds DKK 50,000 within a 12-month period. As such, where you supply short term accommodation and you breach this threshold you should register for VAT"* (Airbnb, 2019). Now what happens to the above scenarios if we also apply VAT, as is custom for short-term rentals? All three scenarios exceed the 50k threshold. As far as I understand, this means that you must pay VAT for each and every night, which is 20%.

A) VAT total = 24.600 DKK => Net earnings become **41.400** DKK.

B) VAT total = 19.680 DKK => Net earnings become **36.480** DKK.

C) VAT total = 14.760 DKK => Net earnings become **31.560** DKK.

Now, indeed, all of these net earning become lower than the earnings from permanent renting. But 365 typical year days minus the 246 days invoked by Airbnb leaves us with 119 days where there are no rentals. That is almost two full months of rental that we are missing. Let's say you permanently rent the room for only 10 months. How does that change things?

D) Monthly price = 5000 DKK. Your annual brute income is 50.000 DKK (renting only 10 months). Your taxable income is 50.000 – 28.000 = 23.000. You pay 60% of that in taxes, which is 13.800 and your net earnings become **37.200** DKK = 28.000 + 9200.

Not bad: 37.200 is roughly on par with scenario B, where your Airbnb nightly price is 400 DKK. But it is below what a 500 DKK price offers. So, as an apartment owner in the Danish capital, you would do well to know your accounting and taxes right and perhaps Airbnb can offer you a slightly increased profit, if you manage to attract sufficient bookings.

But things are not always so black and white. Capital and profit are desirable, of course, but what of the host's personal needs? Also, how realistic it is to share your apartment almost every day with random strangers, provided you can book it that much in the first place. On paper, renting permanently is not an obvious winner, but if we take into account the fact that you have to put in the extra work to manage the Airbnb listing, does it become worth it? I am starting to think not, at least from the perspective of someone who shares his or her home. I believe the perspective changes dramatically if we're taking about using a secondary house (which, by the way, benefits from a much higher tax deduction). But, all of this is just fiction, of course, from a legal point of view. As I've said earlier, it's only allowed to rent your primary home for 70 days, up to 100. Airbnb of this magnitude is not actually possible in this country. Perhaps there are other places, with less (or no) regulations, and lower tax rates, where you can really push the limits. It is possible tough to permanently Airbnb a secondary home. If that is the case, and taking into account the 40.000 deduction, the net earnings are around 7000 DKK (1000 €) higher, on average. Scenario D is again on par with scenario B. All things considered, it might be much safer to go for the full-rental stratagem, taking into account less tax complexity, less uncertainty, legal protection, less management to be done, etc.

A) **48.600** DKK instead of **41.400** DKK.

B) **43.680** DKK instead of **36.480** DKK.

C) **38.760** DKK instead of **31.560** DKK.

D) **44.000** DKK instead of **37.200** DKK (renting 10 months only).

Also, I believe it would be interesting to take the case of sub-renting and see if the host can make a profit or at least cover a decent amount of his/her own housing expenses. I will invoke the same 2-room apartment, which has a monthly rent of 10.000 DKK. You live in one of those of rooms, and the other is up for grabs. Should you do some Airbnb or should you look for a permanent roomie? This time, I will spare the reader the maths and just jump straight to the results. If you do it legally and rent for the maximum amount of 70 nights, you don't pay VAT, if you charge 500 per night, on average. Your net income for this figure is 30.800 DKK, which is approximately 25% of your total

annual rental costs. This is out a brute of 35.000. If you would be able to rent it for 120 nights per annum, you would have to register for VAT and end up making just over 28.000. Even without VAT, you could have only made 40.800, which is a very small increase in income, given that you rent 120 days instead of 70, almost half. In conclusion, by having half of your place being occupied for 19% of the year, you could stand to earn 25% of your rent, which I think is a really big deal for someone who doesn't necessarily need the spare room. The alternative would be to take someone in, who is constantly there, deducting 50% of your rent expenses, but by being there 100% of the time. I believe that if you are not in desperate need of that extra 25%, Airbnb is a great solution. It also offers you the opportunity to feel like a small entrepreneur, meet new people from across the world and gives you something to do, while being almost like a passive income generator.

VI.4 REGRESSION

This section will focus on the methods, models, and data used in regressing the nightly price. Furthermore, in the results chapter, I will present the results I've obtained by training these models and compare their performance, based on metrics, training time, hyperparameter tuning difficulty and interpretability. All my attempts at regressing the data has been done via two models: Random Forest regressor and the XGBoost regressor. This is because I am more familiar with these models and I enjoy many of their characteristics which I will describe below in more detail.

As per Wikipedia, regression is the process of estimating the relationships between the outcome variable and the features. A loss function is used in order to pick the model that best fits the data, according to a mathematical criterion. Example: ordinary least squares (OLS). *"In order to interpret the output of a regression as a meaningful statistical quantity that measures real-world relationships, researchers often rely on a number of classical assumptions. These often include:*

- *The sample is representative of the population at large.*
- *The independent variables are measured with no error.*
- *Deviations from the model have an expected value of zero, conditional on covariates:*
- *The variance of the residuals is constant across observations (homoscedasticity).*
- *The residuals are uncorrelated with one another. Mathematically, the variance–covariance matrix of the errors is diagonal.*

" (Wikipedia, n.d.)

VI.4.A REGRESSION PERFORMANCE METRICS

Mean Squared Error It is the average of the squared difference between the target values and the predicted values. By squaring the differences, it penalizes even the smallest errors. In practice, this leads to an exaggeration of how bad the model actually is. Even so, it presents the quality of being differentiable. Hence, it can be better used for optimization. Higher values mean worse models. It is always positive due to squaring. A perfect model would have a zero valued MSE. The metric is useful *"if we have unexpected values that we should care about. Vey high or low value that we should pay attention. [...] If we make a single very bad prediction, the squaring will make the error even worse and it may skew the metric towards overestimating the model's badness. That is a particularly problematic behaviour if we have noisy data."*

Root Mean Squared Error It is the square root of the averaged squared. The errors are squared before averaging which poses a higher penalty on larger errors. This is why RMSE is preferred when large errors are especially undesirable. This heavily depends on the specific data and goals of the project.

Mean Absolute Error It is the absolute difference between the target values and the predicted. It is supposedly more robust to outliers. It does not penalize the errors as much as MSE or RMSE. Interestingly, all of the individual differences hold equal weight. Thusly, in cases where outliers are a focus, it is not a very good metric to use.

MAPE The mean absolute percentage error measures the prediction accuracy of a forecast. It is calculated as the mean of the absolute difference between the actual and the predicted value, divided by the absolute actual value. In practice, this measure is used as a loss function or a metric in model evaluation because it has a very intuitive interpretation, when the predicted quantity is known to always be above zero (Myttenaere, Golden, Grand, & Rossi, 2016).

R^2 It is the *Coefficient of Determination*. This metric is used to compare models between them. It is very useful if we want to compare different training iterations of the same class of model, using different combinations of hyperparameters. Technically speaking, what R2's value actually tells us is how much of the variance of the underlying target variable is being explained by the predictive model. The higher the value, the better. The range of values if minus infinity to 1. When we get negative R2 it means that it means that the model is worse than predicting the mean of the target.

Adjusted R2 This metric adds a penalty to the complexity of the model, based on the number of features it uses. It is preferable to R2 because it tells us how a model is doing after adding or removing features. It is always lower than R2

VI.4.B MACHINE LEARNING ALGORITHMS

RANDOM FORESTS

Random Forests is an ensemble learning method built on decision trees. They are used for either classification or regression. They are implemented by constructing a *"multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set"* (Wikipedia, n.d.). They use bagging (otherwise known as bootstrap aggregating). Repeated bagging "selects a random sample with replacement of the training set and fits trees to these samples". This technique increases the performance of the model by decreasing variance, while keeping bias low. This takes advantage of the undisputed fact that, although single tree predictions are very noise-sensitive, the average performance of many trees is far superior (considering low correlation between the independent trees). The optimal number of trees can be found through different methods. We can use cross-validation, out-of-bag errors or even brute force, through the (Randomized) Grid Search technique. At each candidate split of each node in the learning part, the model employs a random subset of features. This is known as "feature bagging".

After years of practicing the use of random forests and decision trees, the scientific community has come to a general consensus regarding certain characteristics of these methods. In particular, "trees that are grown very deep tend to learn highly irregular patterns" i.e. overfitting the training sets. Random forests (as well as XGBoost) have the very convenient feature that they can be used to rank the importance of variables used in the prediction. Also, worth noting is that for categorical variables with different number of levels, random forests are *"biased in favour of those attributes with more levels"*. Regarding the way it handles correlation between features, then of there are groups of correlated features that are similarly important for the result, then the model will rank smaller groups above larger ones. The best thing that it does is that the trees are trained not only on different sets (through bagging), but they also make use of different features, which makes the models less likely to overfit (Yiu, 2019). *"The other main concept in the random forest is that only a subset of all the features are considered for splitting each node in each decision tree"* (Koehrsen, 2018).

XGBOOST

Being passionate about Data Science and trying to keep up-to-date with the latest trends and developments, I often find myself browsing on Kaggle, the world's number one go-to website dedicated to Machine Learning. Kaggle hosts many competitions that propose a dataset and an objective and urges individuals or teams to compete in order to deliver the most accurate models. It is no secret that the XGBoost model/package has become very popular on the platform and it has managed to overpower more traditional supervised-learning methods, in both classification and regression problems. According to its website, *"XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way [...] it is developed with both deep consideration in terms of systems optimization and principles in machine learning. The goal of this library is to push the extreme of the computation limits of machines to provide a scalable, portable and accurate library"* (Xgboost, n.d.).

To train the model, an objective function is defined, to measure how good the model is at fitting the training or validation sets. All objective functions are characterized by Training Loss and the Regularization Term. The loss, calculated via a loss function (there are several of these, depending on the type of prediction we're doing, and based on the target variable), shows how good the model is at predicting the training data. The regularization term, keeps the model complexity in check. An overly complex system is not desirable, as it loses its ability to generalize and thusly it leads to overfitting, which is a dreadful issue. *"The general principle is we want both a simple and predictive model. The trade-off between the two is also referred as bias-variance trade-off in machine learning"* (Tutorials, n.d.). *"XGBoost supports missing value by default. In tree algorithms, branch directions for missing values are learned during training."* The model also doesn't require much data pre-processing, as it does not expect normal distributions, which is very convenient, and also makes the regression performance metrics more readable. XGBoost might offer slightly different results between independent runs, *"due to non-determinism in floating point summation order and multi-threading. Though the general accuracy will usually remain the same."*

The model is better than single trees, because it actually uses an ensemble of trees (weak learners), which sums the predictions of the forest as a whole. The trees complement each other, by being built sequentially and taking note of the major pitfalls of previous iterations. *"The idea of boosting is to*

train weak learners sequentially, each trying to correct its predecessor. […] So random forests and boosted trees are really the same models; the difference arises from how we train them" (D'Souza, 2018). *"However, instead of changing the weights for every incorrect classified observation at every iteration like AdaBoost, Gradient Boosting method tries to fit the new predictor to the residual errors made by the previous predictor"* (Zhang Z., 2019). The downside is that traditional gradient boosting models are very slow because of the sequential training. And scalability is important, especially in production. Thus, XGBoost is a great alternative, being focused on both computational speed and model performance.

Other great attributes of XGBoost are:

- ***Parallelization*** *of tree construction using all of your CPU cores during training.*
- ***Distributed Computing*** *for training very large models using a cluster of machines.*
- ***Out-of-Core Computing*** *for very large datasets that don't fit into memory.*
- ***Cache Optimization*** *of data structures and algorithm to make the best use of hardware.*

VII. RESULTS

VII.1 REGRESSION

In practice, I ended up experimenting with a series of models and different data compositions, to test different hypothesis, and try to arrive to the best possible result. Thusly, I will structure this section in multiple "phases". I will briefly describe the chosen model, the choice of hyperparameters, the data used and the results. The data was split into training and test sets, with training representing 80% of the data. 20% was left out for testing the forecasts' power.

Phase 1: Testing

This was the testing phase, after creating the pipeline for cleaning the listings datasets. I decided to pick Amsterdam as the city used for creating the base regression models. The data is "raw". It contains the original data, excluding ulterior features (sentiment, photos). It can be considered that all other model iterations are to be judged by the performance of those used here. The model hyperparameters are almost default here, so I will not discuss much about it.

RandomForestRegressor(max_depth=50, random_state=1)		
METRIC	TRAINING RESULT	TESTING RESULT
MSE	258.88	1097.92
RMSE	16.09	33.13
MAE	10.91	25.27
MAPE	8.73%	21.76%
R2	90%	46%
Adj. R2	89%	32%

XGBRegressor(colsample_bytree=0.6, gamma=0, learning_rate=0.1, max_depth=7, n_estimators=50, random_state=1, l1 = 5)		
METRIC	TRAINING RESULT	TESTING RESULT
MSE	192.11	940.81
RMSE	13.86	30.67
MAE	9.58	22.59
MAPE	7.92%	18.98%
R2	92%	53%
Adj. R2	92%	41%

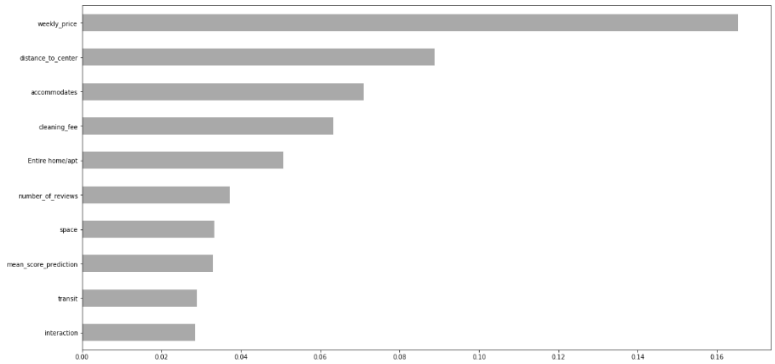
P**hase 2: Amsterdam with Photography Scores**

In Phase 2, I have kept the same models, with the same hyperparameters, in order to make a direct comparison to Phase 1. The difference is that I have added another predictive feature.

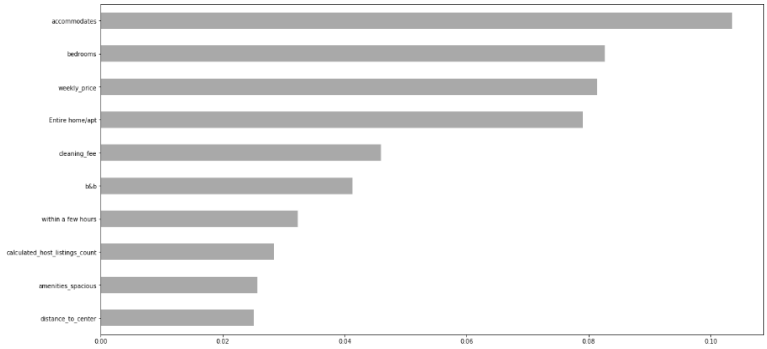
RandomForestRegressor(max_depth=50, random_state=1)		
METRIC	TRAINING RESULT	TESTING RESULT
MSE	329.34	1187.81
RMSE	18.15	34.46
MAE	12.34	26.17
MAPE	9.93%	22.27%
R2	88%	45%
Adj. R2	87%	40%

XGBRegressor(colsample_bytree=0.6, gamma=0, learning_rate=0.1, max_depth=7, n_estimators=50, random_state=1, l1 = 5)		
METRIC	TRAINING RESULT	TESTING RESULT
MSE	112.94	1110.21
RMSE	10.63	33.32
MAE	7.34	24.71
MAPE	5.97%	20.78%
R2	96%	49%
Adj. R2	95%	44%

Random forests feature importance



XGBoost feature importance



Surprisingly, XBG did not end up considering the photos as being that important. RF, on the other hand, does that. Somewhat. This leads me to think that either it's the fact that we're working with so few data points or…the pictures don't actually dictate the price much, but rather the rate of occupancy. Hnestly, by looking at what I've downloaded at least, the sample seemed to contain mostly professional photographs.

Phase 3: All the cities with one-hot encoded city feature

In this phase, I decided to test out another hypothesis: perhaps one-hot encoding the city feature would give me a boost after all. I added it to the data frame and ran some models again. Indeed, this iteration has so far proven the most successful, with the testing results coming much better this time. Also, it is important to note that further tinkering of the hyperparameters has reduced overfitting, as per the training results. The hyperparameters were not chosen randomly. I decided to start off by utilizing a technique known as Randomized Grid search, which is incorporated into Scikit Learn. *"The parameters of the estimator used to apply these methods are optimized by cross-validated search over parameter settings. In contrast to GridSearchCV, not all parameter values are tried out, but rather a fixed number of parameter settings is sampled from the specified distributions"* (scikit-learn, n.d.) After obtaining a set of hyperparameters from the randomized search, one can go into more detail with the actual Exhaustive Grid Search technique. Grid search works by taking in a series of lists of values for a list of desired hyperparameters. The search will then actively search for the combination of parameters that obtains the lowest cross validation score, which assures us that the model is good at generalizing on independent datasets, which in turn helps reduce overfitting. The metrics show that the model is able to generalize more. Getting lower results in this iteration is a good thing, as long as the gap between the training and testing phase is scaled down.

For random forests, max depth controls the growth level of the decision tree, essentially, how deep a tree goes before terminating. Min_samples_leaf controls how many samples are to be present in a leaf node, after splitting, at least. N_estimators dictates the number of single decision trees to be created. It should be noted that, after a certain threshold, performance does not go up. What does go up is training time and computational resources. Bootstrap is set to True, so that the individual trees always use different data, in order to reduce overfitting. The alternative is to make use of the full training set on each tree, which increases the bias and lessens the model's ability to generalize. Min_impurity_split is a hyperparameter that dictates a threshold for early stopping tree growth. A node is split if the impurity coefficient is above this value, otherwise becoming just a leaf.

RandomForestRegressor(bootstrap=True,max_depth=120,max_features=0.5,min_samples_leaf=8,min_impurity_split=10,n_estimators=200)		
METRIC	TRAINING RESULT	TESTING RESULT
MSE	350.12	607.85
RMSE	18.71	24.65

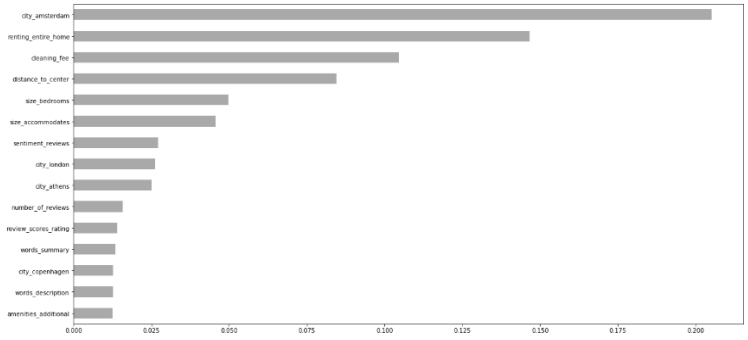
MAE	12.68	17.05
MAPE	20.46%	27.61%
R2	78%	62%
Adj. R2	78%	62%

For XGB, we share some similar hyperparameters (max_depth, controlling the depth of a tree; n_estimator dictates the number of decision trees used) and others which are unique. Colsample_bytree specifies the ratio of column subsamples to be constructed for each tree. Colsample_bynode does the same, but it happens on the level of the nodes, for each decision tree. Gamma or min_split_loss is a measure of the minimum reduction in loss function required for making another partition on a node. Ranging from 0 to infinity, the larger it is, the more conservative the model will be. In our model, Gamma is negligible, but 0.01 was found to be a bit better than 0. The learning_rate, also known as eta, shrinks the feature weights after each boosting increment, in order to make the model more conservative. Random state preserves the model's results, which vary on each re-training, if this is not specified. Subsample tells the model what percentage of the training data to randomly take for each boosting iteration, thus reducing overfitting. Next, we have two hyperparameters which dictate regularization: reg_lambda and alpha. Lambda is the L2 regularization term which is used with the feature weights. It adds squared magnitude of coefficient as the penalty term to the loss function. Our models don't make use of L2. Alpha is the L1 term, which adds an absolute value of magnitude of the coefficient as the penalty term. The result is that alpha can shrink feature's weights to absolute 0, hereby eliminating them from the model, which also helps with dimensionality reduction. Our model's alpha is 1, making it more conservative. Using an L1 term was preferable, due to the presence of many unimportant features.

XGBRegressor(colsample_bytree=0.8, colsample_bynode=1, gamma=0.01, learning_rate=0.2, max_depth=10, n_estimators=100, random_state=1, subsample=0.5, reg_lambda = 0, alpha = 1)		
METRIC	TRAINING RESULT	TESTING RESULT
MSE	225.18	585.1
RMSE	15.01	24.19
MAE	10.49	16.65
MAPE	17.54%	26.5%

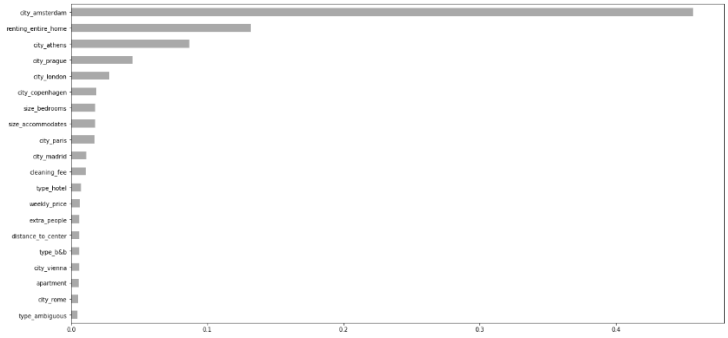
R2	86%	64%
Adj. R2	86%	63%

Random forests feature importance



RF first splits the data by Amsterdam, then entire homes, cleaning fees, distance to centre, size, sentiment and so on. I am pleased to see that the NLP sentiment is fairly well-valued in this model, as are the number of reviews, the review score, additional amenities. I am a tad surprised about the appearance of description and summary lengths, to be honest. When analysing the data at city level, there was not much difference in the average length of text between cities. Perhaps there are certain types of listings which usually write more detailed text?

XGBoost feature importance



XGB splits the data mostly by city. Again, more specifically: Amsterdam. I think it makes too much of a difference for my taste. I believe it makes the model very biased. The first feature that the trees split is this city, due to the very specific statistical proprieties that its price distributions are. This makes sense, as we know Amsterdam to be the outlier in our selection of cities. I believe this is reason enough for me to try another iteration of this phase, using all the data except for this one city. Other important features are renting the entire home, the size and the distance to the centre, which is

consistent with other findings in literature. Also, interestingly, bread and breakfast types of listings and apartments seem to make up more important categories of listing types.

Phase 4: Modelling pairs of cities

After discovering that the nightly price distributions of some pairs of cities were very similar both in shape, range, skewness and statistics such as the mean, I had an idea: what if I overdid it by making a model based on ALL cities. Maybe we just need to find ones that are similar enough so that the model can generalize better. A model for whom the city feature is not that important, perhaps? The hyperparameters remain unchanged, relative to *Phase 4*, just so we can compare the performance better. The data here arguably adds more complexity to the models because I am one-hot encoding the neighbourhood as a categorical feature. I avoided doing this in the full database models because I considered that so many neighbourhoods would add way too many features. Especially since I one-hot encoded the city. On the other hand, I know neighbourhoods are very important in dictating price, as I've seen from the maps. Either way, I left the city feature out of this because of the high correlation it has with its respective neighbourhoods.

COPENHAGEN – PARIS pair

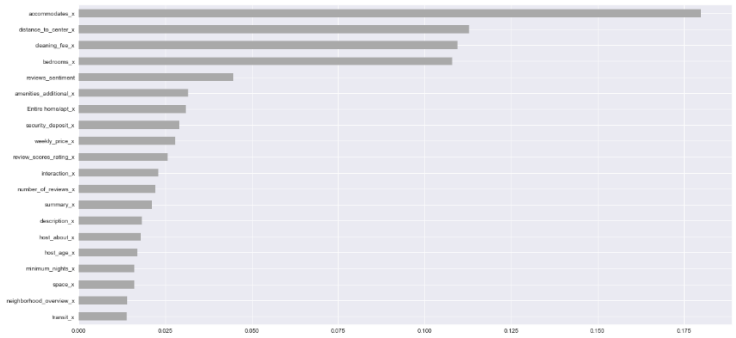
RandomForestRegressor(bootstrap=True,max_depth=120,max_features=0.5,min_samples_leaf=8,min_impurity_split=10,n_estimators=200)		
METRIC	TRAINING RESULT	TESTING RESULT
MSE	342.58	617.85
RMSE	18.51	24.86
MAE	12.88	17.52
MAPE	18.6%	23.13%
R2	69%	46%
Adj. R2	69%	44%

XGBRegressor(colsample_bytree=0.6, gamma=0, learning_rate=0.1, max_depth=6, n_estimators=100, random_state=1, verbosity=1)		
METRIC	TRAINING RESULT	TESTING RESULT
MSE	387.36	555.96
RMSE	19.68	23.58

MAE	14.24	16.76
MAPE	20.74%	23.9%
R2	65%	51%
Adj. R2	65%	50%

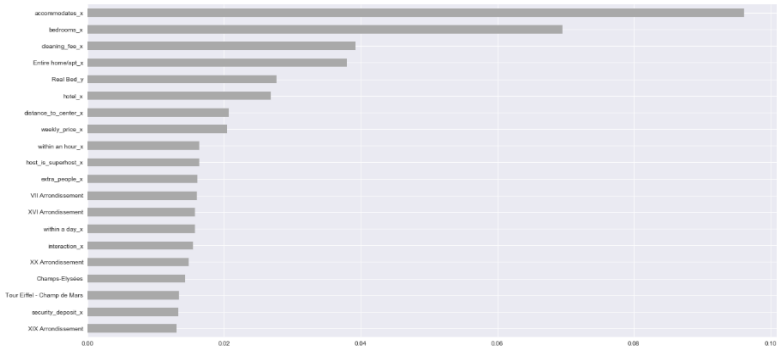
Something very interesting and frustratingly mind-boggling has happened. We can see that if we judge the performances by the MSE, RMSE and MAE metrics, there is almost no difference between these models which only take the data of two cities, and the other models, which hold all of the data. If we instead look towards the R2 and Adj. R2, the performance here, in Phase 4, is much, much lower. The smaller amount of data cannot explain the variation in the price's underlying distribution. Theoretically, for this pair of cities at least, you could just do it like this and obtain similar results to the bulkier model, but your model is not significantly better at predicting the nightly prices. Also, R2 just tells us they are bad. 50% and under is bad. Even more interestingly, adding new features with the neighbourhoods did not t keep the R2 very high, while the Adjusted R2 clearly penalizes a lot for the addition of so many neighbourhoods.

Random forest feature importance



RF takes number of accommodates as the best feature, followed by distance, cleaning fee, bedrooms, sentiment of reviews. Additional amenities score higher than if the home is rented in its entirety or not. Security deposit scores high here, perhaps because there are no such deposits in Copenhagen, but Paris has the highest deposit fees in the dataset. The length of the text in the listing is surprisingly important to RF. I was expecting to see some of the neighbourhoods here, but I was wrong. RF could not care less about them in this particular setup. It actually manages to generalize very well and what we can see in the most important features is what we would normally theorize to be important for setting the price.

XGBoost feature importance



XBG takes size as the most important features, followed by the cleaning fee, whether it's the full apartment or not. Distance is lagging behind, while sentiment does not appear at all. But here, contrary to RF, neighbourhoods make their appearance amongst the top features, all of them from Paris. We know from the visual analytics that these neighbourhoods are among the priciest in Europe: Champes Elysees, Tour Eifel. Clearly, there's some special characteristics that they hold that make them unique. It only makes sense that they are used for splitting the trees. It's also probably among the reasons why XGB does so much better than the RF. It had a good use for the new input, instead of it just ending up clogging up the model. Granted, I've also allocated a 10% higher selection of features for the XGB to use when splitting.

The results are very similar to those produced by taking in all the data, for all cities. In fact, the MAE is a bit lower here, which means that we are not producing results which merit more dedication to this strategy. R2 is much lower, which shows that, indeed, more data (even though it comes from different sources) will produce models that encapsulate more of the target variable's variation.

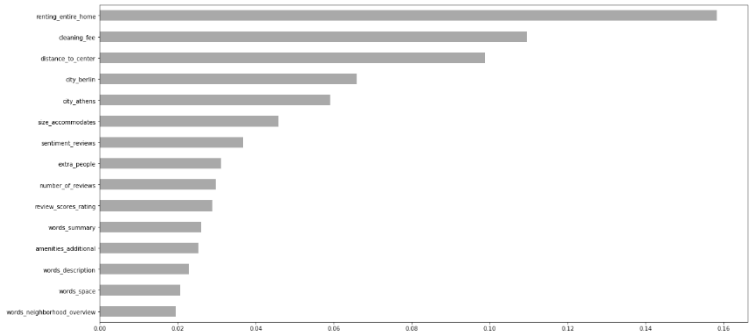
ATHENS – BERLIN pair

Pairing other cities also yields good results, in light of metrics such as MSE, RMSE and MAE. The problem is that, having less data, we obtain lower R2 scores, as a consequence. But it can still be observed that the predictions are closer to reality. Here, the scores are lower than the previous because these two cities have some of the lowest prices. Paris has the most data, which explains the much higher R2. Here, we have much less material to work with.

RandomForestRegressor(bootstrap=True,max_depth=120,max_features=0.5,min_samples_leaf=8,min_impurity_split=10,n_estimators=200)		
METRIC	TRAINING RESULT	TESTING RESULT

MSE	157.68	267.2
RMSE	12.56	16.35
MAE	9.06	12.11
MAPE	22.68%	30.78%
R2	65%	40%
Adj. R2	65%	38%

Random Forest feature importance



XGBRegressor(colsample_bytree=0.6, gamma=0, learning_rate=0.1, max_depth=6, n_estimators=100, random_state=1, verbosity=1)		
METRIC	TRAINING RESULT	TESTING RESULT
MSE	143.7	251.54
RMSE	11.99	15.86
MAE	8.85	11.69
MAPE	22.25%	29.42%
R2	68%	43%
Adj. R2	68%	42%

Phase 5: Modelling all cities minus Amsterdam

I decided to test this hypothesis: eliminating Amsterdam from the dataset will wield better results. And indeed, it did return significantly more accurate predictions, once I eliminated this outlier, all the while keeping the same model hyperparameters, for easy comparison. The model was able to generalize better and didn't take into account cities as being so important for splitting the data. What

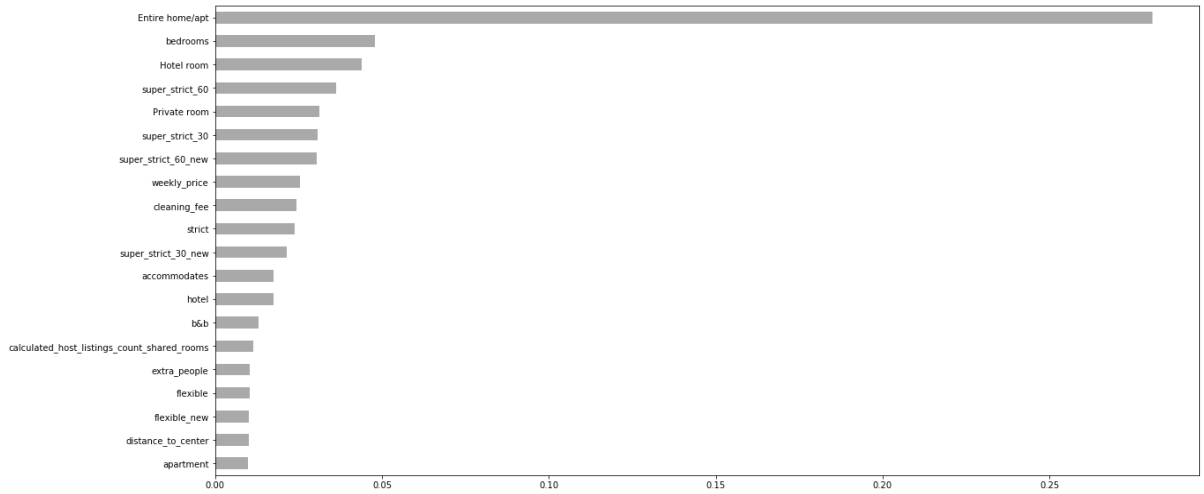
I cannot account for is the slight 1% increase in R2 in the training results. On the other hand, for the test set, R2 gets significantly lower (5%), which could be explained by the fact that the model has less data to work with.

WITH AMSTERDAM

XGBRegressor(colsample_bytree=0.8, colsample_bynode=1, gamma=0.01, learning_rate=0.2, max_depth=10, n_estimators=100, random_state=1, subsample=0.5, reg_lambda = 0, alpha = 1)		
METRIC	TRAINING RESULT	TESTING RESULT
MSE	225.18	585.1
RMSE	15.01	24.19
MAE	10.49	16.65
MAPE	17.54%	26.5%
R2	86%	64%
Adj. R2	86%	63%

WITHOUT AMSTERDAM

XGBRegressor(colsample_bytree=0.8, colsample_bynode=1, gamma=0.01, learning_rate=0.2, max_depth=10, n_estimators=100, random_state=1, subsample=1, reg_lambda = 0, alpha = 1)		
METRIC	TRAINING RESULT	TESTING RESULT
MSE	151.35	486.97
RMSE	12.3	22.07
MAE	8.74	15.47
MAPE	15.53%	26.23%
R2	87%	59%
Adj. R2	87%	58%



VII.2 CLASSIFICATION

I was curious about the unorthodox method of trying to use classification instead of regression in order to predict Airbnb nightly prices. I wanted to see what accuracy I could get and, even if it was accurate enough, how well would this suit the goal of the project. I knew about the concept of dichotomization: essentially transforming a continuous variable such that it becomes a binary categorical variable. For example, taking a range of values from 0 to 1, of they are below 0.5, they become 0, otherwise, they become 1. In my case, dichotomization would have implied transforming the prices to be either ‘low’ or ‘high’, which would not suffice for the purposes of my work. Therefore, I decided to try out transforming the price into a multi-categorical variable.

I studied all the price distributions (in Euro, standard scaled) of all the cities, for a total of ten cities. Obviously, now all distributions follow a somewhat bell curved shape, although all still retain a slight right skew, despite my efforts to eliminate some of the outliers. In any case, the mean is now 0, the standard deviation is 1 and, generally, most cities’ prices are in a range of around -2 STD to 3 or 4 STD. I designed the new price category variable to have a maximum of 8 values (1 to 7). If the price is between the minimum value and minus two STD, it belongs to category 1. If it is between minus two STD and minus 3 STD, it belongs to category 2 and so on. To keep in line with the regression models, I used RF classifiers and XGB Classifiers for this task.

Results: Training set Accuracy: 59% and Test set Accuracy: 53% using RandomForestClassifier(n_estimators = 100, criterion = 'gini', random_state = 42,max_depth = 12, max_features = 'auto', n_jobs = -1)

The results are very, very disappointing. Accuracy just above 50%, overfitting models and even if the accuracy would have been higher, the prices would have still been way off. For example: Paris and Copenhagen's STD are over 34 euros. The regression models deliver a MAE of around 16-17 euros, which is half a STD for these cities. Say the regression predicts one price as being 80 Euros. It means the price is likely to be between 64 and 96 Euro, a range of 32 Euros, lower than the Standard Deviation. In conclusion, classification for this problem is totally off and I would not recommend pursuing it further. I also thought of creating double the categories, by making the intervals be between half STDs, but, surely, this would make it even harder for the classifier to offer decent results.

VIII. DISCUSSION AND CONCLUSIONS

The results of data analysis reveal that there are both striking similarities and also key differences between different Airbnb markets. Some of these differences can be partially explained by correlating the findings with information from alternate sources. The root of other dissimilarities cannot be precisely explained using quantitative methods, but fairly educated guesses can be made related to the local culture and legal framework in place. Among the similarities that cities have in common we can mention: sharing roughly the same average summary description length; having right-skewed price distributions; having a low-season in the early winter months of the year; having a high-season at the end of the year, corresponding to winter holidays, especially New Year's Eve and Christmas; having higher prices for weekends; most listings rent out the entire home; very few listings are for shared rooms; there are very few hosts with Superhost status; the larger the city, the greater the average price; certain amenities such as free parking are better predictors (either positive or negative) of price; the overwhelming majority of comments have a positive polarity score; the same can be said about review scores; positive comments are longer on average; in their reviews, guests tend to mention things about the city, the host, the listing's surrounding attractions and transportation.

Regarding the predictive quality of our features, there are some features which consistently make it to the top of feature importance, regardless of the data used. The findings are consistent with all previous research and seem to establish a "golden" rule. It seems that there is a fail proof recipe of guaranteeing higher pricing for hosts. The larger the listing is, the higher we can charge on it. We can gauge the space by looking at a series of features: the number of accommodates, the number of bedrooms/beds. The closer a listing is to the city centre, the more expensive it gets. In general, central neighbourhoods are more expensive than those outside the kernel of the city, although this is not

always true (Some cities such as Paris have areas on the outskirts that are as or more expensive than the centre. This is probably due to historical attractions.). Renting out an entire property entitles you to charging higher nightly prices. We can agree that Superhost status is correlated with higher prices, more bookings and higher ratings although the causality is disputable. Also, hosts which own multiple listings seem to perform better than single-listing hosts. Review rating scores and the number of reviews are also relevant to the pricing schemes, although not as much as might have been guessed.

In regards to the employed machine learning models, XG-Boost consistently appears as one of the top performing predictors. Some have obtained better results with neural networks but it is not always the case. Alas, the differences in performance metrics are negligible, but the complexity involved in tuning the deep learning models, the extra hardware resources needed and the training times seem to conclude that they might not be worth the effort. R2 scores across all models and projects never seem to break the 70% barrier, which leads researchers to conclude that there are some crucial features missing from the models. Even the addition of new features which make use of photography and sentiment analysis, although useful, does not significantly affect the performance of these models. This led me to believe that there is some very important knowledge to be discovered in studies that focus on the social aspects of Airbnb. I am tempted to suspect that there should be much more attention placed on the hosts and how their avatars present themselves on the platform. Apparently, guests are, knowingly or not, biased towards certain types of hosts. Leveraging online social network presence is important. Facial expressions can make or break a booking. Evidence of discrimination seems to be consistently discovered. I also suspect that information specific to each city is out there somewhere and it could help better understand why the prices reach certain values. I couldn't make classification work for this use case but perhaps I am missing something, since I did not invest much in this approach.

Regarding the social impact of Airbnb, research seems to suggest that localities where it witnesses surges in popularity also witness rapid rates of increase in rental prices. Universally, the concentration of listings in the city centres is very large. We can also witness different types of distributions, in cities such as Paris, which are packed full of Airbnb. The culture, geography, topography and history of cities might be worth exploring if we are to understand more about the phenomenon. People seem to be attracted to characteristics which bring the most logistical advantages. Keeping it simple and balancing the quality/price report

I also wish to write a few remarks on the more technical aspects that involve picking up such as project, which might help readers with their own iterations. Generally, I've found the data to be rather messy and in need of extensive pre-processing. Even though I've managed to streamline and automate the process, I still consider that special attention should be awarded to each and every city. Ideally, EDA should precede any attempts to exclude data points. There are many peculiar outliers to be found but I believe that most of them are actually true to the listing's true profile. I did manually check up on such listings based on their URL but decided that they skewed the distributions too much. Alas, there are models that deal well with missing values and outliers. Perhaps keeping such data points could improve model performance, depending on the methodology chosen. As for the scraper, one should cautiously approach it when trying to download data en-mass, because there are some special quirks that I've mentioned, regarding some corrupted files on the website.

Regarding the profitability of practicing Airbnb hosting, I cannot make some definite conclusions. The issue is too complex to jump to statements. It really heavily depends on local regulations and taxation legislation. The only city I made a study case out of was Copenhagen. Even so, the analysis is more qualitative than quantitative and the assumptions I made are only partially backed up by facts. Even so, taking into consideration the very high taxation rates in this city, I found hosting to be appealing in certain cases. More precisely, I consider the safest alternative is to go for traditional long-term renting, because it is a stable source of income that doesn't complicate itself with social and legal issues. But lower risk also implies lower possible profitability. Outside of regulations, it is obviously much more rewarding to do short-term rentals, but the limits imposed by local governments on the number of nights per year severely limit the outcomes. An aspiring entrepreneur wishing to turn a profit out of Airbnb should search for markets which are less exigent and allow more autonomy and flexibility. Otherwise, I believe that Airbnb is ideal for those who can dispose of a room from time to time, in order to recover a considerable chunk of the rent they are paying. Needless to say, the profit margin is greatest for those who actually own property, because the costs of setting up such a "business" are much lower and it also allows for infinite possibilities of customization regarding the living space.

There are also several other directions to go in using the data from the Inside Airbnb project. The information there can be compounded with external resources to strengthen the predictive power or it can also be used to explore different research questions. For example, data from Numbeo can be used to create a model that investigates the profitability of renting out, based on consumer good prices and rent levels. Data from Trip Advisor can tell us more about what a listing's location has to offer,

in relation to whatever touristic attractions are available nearby. Research on taxation and local legislation can be used to construct a more thorough picture of the economic implications that arise. More sophisticated datasets can be built by harnessing the power of NLP and Convolutional Neural Networks that work with listing images, given the fact that these features have proven to be worthy of special interest. I believe the most important thing that is left to be done is to find new and creative ways to add explanatory features to the dataset, which could help raise R2 scores. Perhaps looking less towards the characteristics of the listing itself and more into the characteristics of the host could provide some breakthroughs. Of course, all of this would require the researcher to go out of his way to find new sources of data, which could become very time consuming. Alas, the performance that current models offer with the data at hand can be considered relatively decent. The condition is that we should not take the output of these algorithms for granted, but rather with a grain of salt. A price prediction can be a good starting point from. A host should then make use of autonomy to experiment with values which gravitate around that value and find out which one works out the best for his or her needs. After all, price is not all there is to it. We also have to think about the rate of occupancy that we want, which suits our needs. For some, lower pricing schemes, coupled with high occupancy might bring more money. For others, higher pricing with occasional renting out might be the solution.

What I would recommend to future researchers is that they try different combinations of features, and different models, in order to investigate the possibility of developing more powerful models. And even tough previous works have shown that Neural Networks do not seems to produce wonders on this type of data, I still think it could be worth it to experiment with trying out different architectures, which might wield more positive outcomes. To aspiring hosts, I have a series of advice that would help them become more competitive. Trying to reach Superhost status should not be a central goal, but it has been proven to be correlated with better performance. This, along with the apparent importance of social status on web platforms, should be enough of a motivator for property owners to do their best to be more responsive, more attentive to their guests' needs and more preoccupied with feedback. Owners should be very self-conscious of their apartment's physical characteristics, especially size and location, and adjust their pricing accordingly. Investing in providing amenities is a good thing, but one should not go over the top with special features, as they are not that important. Hosts should also keep in touch with the reality of Airbnb and inform themselves on the limits being imposed by local governments and city councils, as they can set definite thresholds for how much you can earn before tax, how long you can rent out, and even if you are legally allowed to start such an endeavour or not. Marketing is very important, as with any other type of economic adventure. But

marketing doesn't just stop at how you write about your products and services. Humans are very visual and social creatures and it seems that photographs play an important yet difficult to grasp role in shaping the success of an Airbnb listing. We know that the quality of the listing's photography is crucial, but at the same time, the way the host presents himself or herself has major implications. Research suggests that people are visually biased towards other people and the way you look might instil an array of emotions that can make or break a deal. Being a sharing economy platform, trust is important. And we've proven that review scores and comments are not relevant, due to them being overwhelmingly positive. That's why potential guests go outside of this space to seek reassurance the old-fashioned way: the way a person makes us feel through our gut instinct.

IX. References

1. (n.d.). From Medium: https://medium.com/@george.drakos62/how-to-select-the-right-evaluation-metric-for-machine-learning-models-part-1-regrression-metrics-3606e25beae0
2. Airbnb. (n.d.). *How should I choose my listing price*. From Airbnb: https://www.airbnb.com/help/article/52/how-should-i-choose-my-listings-price
3. Airbnb. (2019, January). *DENMARK – TAX CONSIDERATIONS ON SHORT-TERM LETTINGS*. From Airbnb: https://assets.airbnb.com/help/airbnb-pwc-taxguide-denmark-en.pdf
4. *Airbnb lowers internal valuation by 16% to $26bn* . (2020). From Fianancial Times: https://www.ft.com/content/02a8ca9b-1ba9-4e0a-a3d5-084dd93469bb
5. Airbnb. (n.d.). *Smart Pricing*. From Airbnb: https://blog.atairbnb.com/smart-pricing/
6. Airbnb. (n.d.). *Superhost*. From Airbnb: https://www.airbnb.com/superhost
7. Airbnb. (n.d.). *The positive impacts of home sharing in Copenhagen*. From Airbnb: https://news.airbnb.com/the-positive-impacts-of-home-sharing-in-copenhagen/
8. Airbnb. (n.d.). *What is a superhost*. From Airbnb: https://www.airbnb.com/help/article/828/what-is-a-superhost
9. Aydin, R. (2020, September 20). *How 3 guys turned renting air mattresses in their apartment into a $31 billion company, Airbnb.* Retrieved April 26, 2020 from Business Insider: https://www.businessinsider.com/how-airbnb-was-founded-a-visual-history-2016-2?r=US&IR=T
10. *Beautiful Soup Documentation*. (n.d.). From https://www.crummy.com/: https://www.crummy.com/software/BeautifulSoup/bs4/doc/
11. *BeautifulSoup*. (n.d.). From https://www.crummy.com/: https://www.crummy.com/software/BeautifulSoup/
12. Bettendorf, B. (2019, June). *NLP on Airbnb data*. From Kaggle: https://www.kaggle.com/brittabettendorf/nlp-on-airbnb-data
13. Blessen, C. (2019, November 20). *Matplotlib vs. Seaborn vs. Plotly*. From Towardsdatascience: https://towardsdatascience.com/matplotlib-vs-seaborn-vs-plotly-f2b79f5bddb
14. *Bookings and traveler engagement driven by management actions*. (n.d.). From Tripadvisor: https://www.tripadvisor.com/TripAdvisorInsights/w613
15. *Box Plot*. (n.d.). From Wikipedia: https://en.wikipedia.org/wiki/Box_plot
16. Cai, T., Han, K., & Wu, H. (n.d.). *Melbourne Airbnb Price Prediction.* Stanford : Stanford University.
17. Cai, Y., Zhou, Y., Ma, J., & Scott, N. (May 2019). Price Determinants of Airbnb Listings: Evidence from Hong Kong. *Tourism Analysis*.
18. *Can Airbnb Survive Coronavirus?* (2020, April 3). From Citylab: https://www.citylab.com/life/2020/04/coronavirus-safe-travel-airbnb-rental-business-host-bailout/608917/
19. cjhutto. (n.d.). *Vader Sentiment*. From Github: https://github.com/cjhutto/vaderSentiment
20. *Dash*. (n.d.). From Plotly: https://plotly.com/dash/
21. *Denmark Approves Forward-thinking Home Sharing Rules and Simplifies Tax*. (2019, April 4). From Airbnb: https://www.airbnb.com/help/article/1382/responsible-hosting-in-denmark
22. D'Souza, J. (2018, March 21). *A quick guide to boosting in ML* . From Medium: https://medium.com/greyatom/a-quick-guide-to-boosting-in-ml-acf7c1585cb5
23. DutchNews.nl. (2020, January 31). *Amsterdam Airbnb hosts are in a grey area after court*

ruling. From DutchNews.nl: https://www.dutchnews.nl/news/2020/01/amsterdam-airbnb-hosts-are-in-a-grey-area-after-court-ruling/
24. Edelman, B., & Luca, M. (January 2014). Digital Discrimination: The Case of Airbnb.com . *Harward Business School*.
25. Ert, E. (January 2015). Trust and Reputation in the Sharing Economy: The Role of Personal Photos on Airbnb. *Tourism Management*, 66.
26. *Exploratory Data Analysis*. (n.d.). From Wikipedia: https://en.wikipedia.org/wiki/Exploratory_data_analysis
27. Fagerstrøm, A., Pawar, S., Sigurdsson, V., Foxall, G. R., & Yani-de-Soriano, M. (2017). That Personal Profile Image Might Jeopardize Your Rental Opportunity! On the Relative Impact of the Seller's Facial Expressions upon Buying Behavior on Airbnb. *Computers in Human Behavior*.
28. *From Data Munging to Data Wrangling*. (n.d.). From Trifacta: https://www.trifacta.com/data-munging/
29. Guttentag, D. (2018, August 30). *What Airbnb really does to a neighbourhood*. From BBC News: https://www.bbc.com/news/business-45083954
30. Guttentag, D. (August 2016). *Airbnb: Why Tourists Choose It and How They Use It.* (Toronto, Ontario, Canada: Ryerson University.
31. *Histogram*. (n.d.). From Wikipedia: https://en.wikipedia.org/wiki/Histogram
32. *Homepage*. (n.d.). From Tableau: https://www.tableau.com/products/individuals?utm_medium=homepage
33. *How will my listings location be shown on the map?* (n.d.). From Airbnb: https://www.airbnb.com/help/article/2141/how-will-my-listings-location-be-shown-on-the-map
34. Hutto, C., & Gilbert, E. (n.d.). *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text.* From http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf
35. *I rent out my home in London; what short term rental laws apply?* (n.d.). From Airbnb: https://www.airbnb.com/help/article/1340/i-rent-out-my-home-in-london-what-shortterm-rental-laws-apply
36. Inside Airbnb Project. (n.d.). *About Inside Airbnb*. From Inside Airbnb: http://insideairbnb.com/about.html
37. Kalehbasti, P. R., Nikolenko, L., & Rezaei, H. (July 2019). *Airbnb Price Prediction Using Machine Learning and Sentiment Analysis.* Stanford: Stanford University.
38. Keating, J., Katnic, E., Hahn, C., & Yang, R. (2018). *PREDICTIVE MODELING ON AIRBNB LISTING PRICES.* Washington: University of Washington.
39. Koehrsen, W. (2018, August 30). *An implementation and explanation of the randon forest in Python*. From Towards Data Science: https://towardsdatascience.com/an-implementation-and-explanation-of-the-random-forest-in-python-77bf308a9b76
40. Koh, V., Li, W., & Livan, G. (March 2019). Offline biases in online platforms: a study of diversity and homophily in Airbnb. *Springer*.
41. Kwok, L. &. (2018). Pricing strategies on Airbnb: Are multi-unit hosts revenue pros? *International Journal of Hospitality Management*.
42. Laurent, L. (2020, April 2). *Will Airbnb Become Obsolete After the Coronavirus?* From Bloomberg: https://www.bloomberg.com/opinion/articles/2020-04-02/will-airbnb-become-obsolete-after-the-coronavirus
43. Lawani, A., Michael, R., Mark, T., & Zheng, Y. (n.d.). *Impact of reviews on price: Evidence from sentiment analysis of Airbnb reviews in Boston.* Boston.

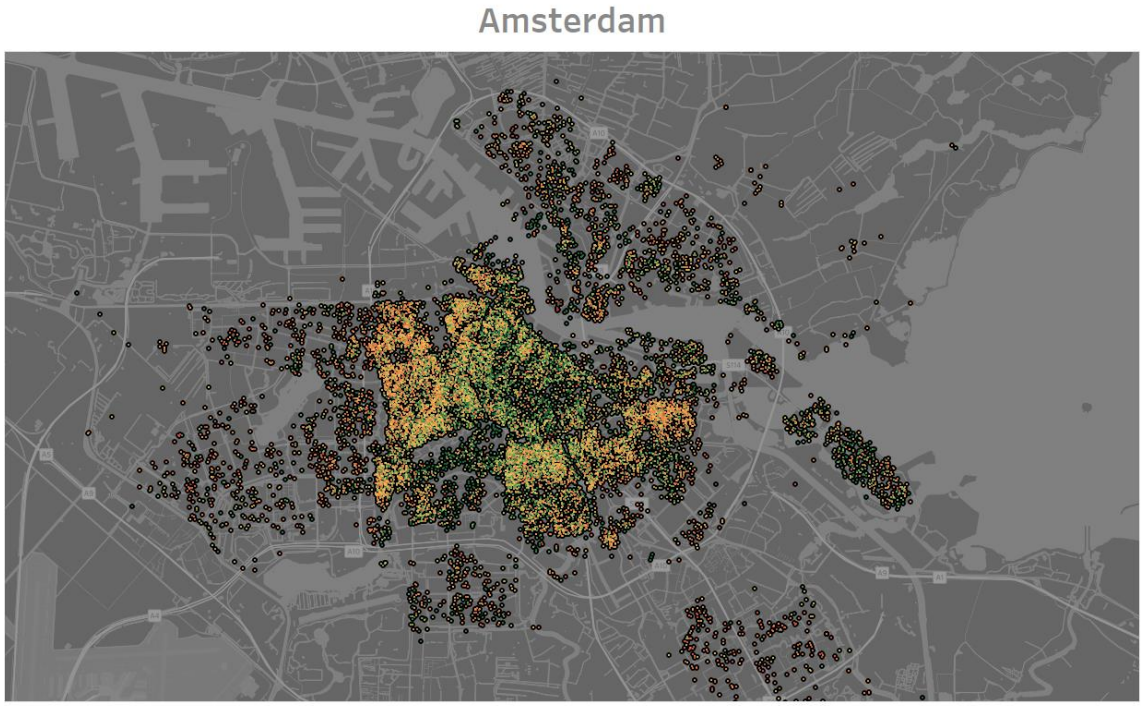
44. Lennan, C. (2018, July 19). *https://medium.com/idealo-tech-blog/using-deep-learning-to-automatically-rank-millions-of-hotel-images-c7e2d2e5cae2*. From Medium: https://medium.com/idealo-tech-blog/using-deep-learning-to-automatically-rank-millions-of-hotel-images-c7e2d2e5cae2
45. Lennan, C. (n.d.). *Image Quality Assesment*. From Github: https://github.com/idealo/image-quality-assessment/tree/master/models/MobileNet
46. Lewis, L. (2019, May 17). *Airbnb-neural-network-price-prediction*. From Github: https://github.com/L-Lewis/Airbnb-neural-network-price-prediction/blob/master/Airbnb-price-prediction.ipynb
47. *MAP OF THE MOST EXCLUSIVE NEIGHBOURHOODS OF MADRID*. (n.d.). From Promora: https://www.promora.com/en/zones/center-madrid
48. Marchenko, A. (December 2019). The impact of host race and gender on prices on Airbnb. *Journal of Housing Economics*.
49. *Matplotlib*. (n.d.). From Wikipedia: https://en.wikipedia.org/wiki/Matplotlib
50. *Matplotlib: Visualization with Python*. (n.d.). From Matplotlib: https://matplotlib.org/
51. Myttenaere, A. d., Golden, B., Grand, B. L., & Rossi, F. (2016). Mean Absolute Percentage. *Neurocomputing, Elsevier*, 1-3.
52. O'Sullivan, F. (2019, July 2). *European Cities Fear They'll Lose Power To Regulate Airbnb*. From Citylab: https://www.citylab.com/life/2019/07/vacation-rentals-europe-cities-airbnb-regulations-travel-eu/593146/
53. O'Sullivan, F. (2019, April 3). *Madrid Bans Airbnb Apartments That Don't Have Private Entrances*. From Citylab: https://www.citylab.com/life/2019/04/madrid-airbnb-apartment-vacation-rental-law-almendra-central/586291/
54. *Professional Photography*. (n.d.). From Airbnb: https://www.airbnb.com/professional_photography
55. Reichel, C. (2018, April 19). *Airbnb prices lower among minority hosts in San Francisco*. From Journalists Resource: https://journalistsresource.org/studies/society/race-society/airbnb-discrimination-hosts/
56. *Requests: HTTP for Humans™*. (n.d.). From https://requests.readthedocs.io: https://requests.readthedocs.io/en/master/
57. Ronquillo, A. (n.d.). *Python's Requests Library (Guide)*. From REALPYTHON.COM: https://realpython.com/python-requests/
58. *Scatter Plot*. (n.d.). From Wikipedia: https://en.wikipedia.org/wiki/Scatter_plot
59. scikit-learn. (n.d.). *sklearn.model_selection.RandomizedSearchCV*. From Scikit-Learn: sklearn.model_selection.RandomizedSearchCV
60. *seaborn: statistical data visualization*. (n.d.). From Seaborn.pydata: https://seaborn.pydata.org/
61. SKAT. (n.d.). *Renting out a room or property you live in*. From SKAT.dk: https://skat.dk/skat.aspx?oid=2285757
62. Statista. (n.d.). *Leading European city tourism destinations by number of bednights*. From Statista: https://www.statista.com/statistics/314340/leading-european-city-tourism-destinations-by-number-of-bednights/
63. Stone, T. (2018, April 6). *Airbnb is getting blamed for Amsterdam's housing crisis. So the city council is going to war against Airbnb*. From Citymetric: https://www.citymetric.com/business/airbnb-getting-blamed-amsterdam-s-housing-crisis-so-city-council-going-war-against-airbnb
64. Stone, T. (2018, April 6). *Airbnb is getting blamed for Amsterdam's housing crisis. So the city council is going to war against Airbnb*. From Citymetric:

https://www.citymetric.com/business/airbnb-getting-blamed-amsterdam-s-housing-crisis-so-city-council-going-war-against-airbnb
65. Tait, R. (2020, February 1). *'I'm a stranger in my own city': Prague takes on Airbnb to dam flood of tourists*. From The Guardian : https://www.theguardian.com/environment/2020/feb/01/overwhelmed-prague-tries-to-limit-airbnb-to-curb-tourism
66. Teubner, T., Dann, D., & Hawlitschek, F. (May 2017). PRICE DETERMINANTS ON AIRBNB: HOW REPUTATION PAYS OFF IN THE SHARING ECONOMY. *Journal of Self-Governance and Management Economics*.
67. Tun, Z. T. (2020, March 8). *Top Cities where Airbnb is Legal or Illegal*. From Investopedia: https://www.investopedia.com/articles/investing/083115/top-cities-where-airbnb-legal-or-illegal.asp
68. *Tutorials*. (n.d.). From Xgboost: https://xgboost.readthedocs.io/en/latest/tutorials/model.html
69. *vaderSentiment*. (n.d.). From Github: https://github.com/cjhutto/vaderSentiment
70. Wang, D., & Nicolau, J. L. (2017). Price Determinants of Sharing Economy Based Accommodation Rental: A Study of Listings from 33 Cities on Airbnb.com. *International Journal of Hospitality Management*.
71. *What does the rise of Airbnb mean for Vienna's property market?* (n.d.). From Financial Times: https://www.ft.com/content/3ec5d7da-1dd1-11ea-81f0-0c253907d3e0
72. Wikipedia. (2020, April 27). *Airbnb*. From Wikipedia: https://en.wikipedia.org/wiki/Airbnb
73. Wikipedia. (2020, April 27). *Airbnb*. Retrieved April 26, 2020 from Wikipedia: https://en.wikipedia.org/wiki/Airbnb
74. Wikipedia. (n.d.). *Data wrangling*. From Wikipedia: https://en.wikipedia.org/wiki/Data_wrangling
75. Wikipedia. (n.d.). *List of European Union cities by population density*. From Wikipedia: https://en.wikipedia.org/wiki/List_of_European_Union_cities_proper_by_population_density
76. Wikipedia. (n.d.). *Ranfom Forest*. From Wikipedia: https://en.wikipedia.org/wiki/Random_forest
77. Wikipedia. (n.d.). *Regression Analysis*. From Wikipedia: https://en.wikipedia.org/wiki/Regression_analysis
78. *Xgboost*. (n.d.). From Xgboost: https://xgboost.readthedocs.io/en/latest/
79. Xie, K., & Mao, Z. (July 2017). The Impacts of Quality and Quantity Attributes of Airbnb Hosts on Listing Performance. *International Journal of Contemporary Hospital Mangement* .
80. Yiu, T. (2019, June 12). *Understanding Random Forest*. From Towards Data Science: https://towardsdatascience.com/understanding-random-forest-58381e0602d2
81. Zee, R. v. (2016, October 6). *The 'Airbnb effect': is it real, and what is it doing to a city like Amsterdam?* From The Guardian: https://www.theguardian.com/cities/2016/oct/06/the-airbnb-effect-amsterdam-fairbnb-property-prices-communities
82. Zee, R. v. (2016, October 6). *The 'Airbnb effect': is it real, and what is it doing to a city like Amsterdam?* From The Guardian: https://www.theguardian.com/cities/2016/oct/06/the-airbnb-effect-amsterdam-fairbnb-property-prices-communities
83. Zhang, S., Lee, D., Singh, P. V., & Srinivasan, K. (2016). *How Much Is An Image Worth? An Empirical Analysis of Property's Image Aesthetic Quality on Demand at AirBNB.* Dublin: Thirty Seventh International Conference on Information Systems.
84. Zhang, Z. (2019, June 26). *Boosting Algorithms Explained*. From Towards Data Science:

https://towardsdatascience.com/boosting-algorithms-explained-d38f56ef3f30
85. Zhang, Z., Chen, R. J., Han, L. D., & Yang, L. (2017). Key Factors Affecting the Price of Airbnb Listings: *MDPI*, 11.

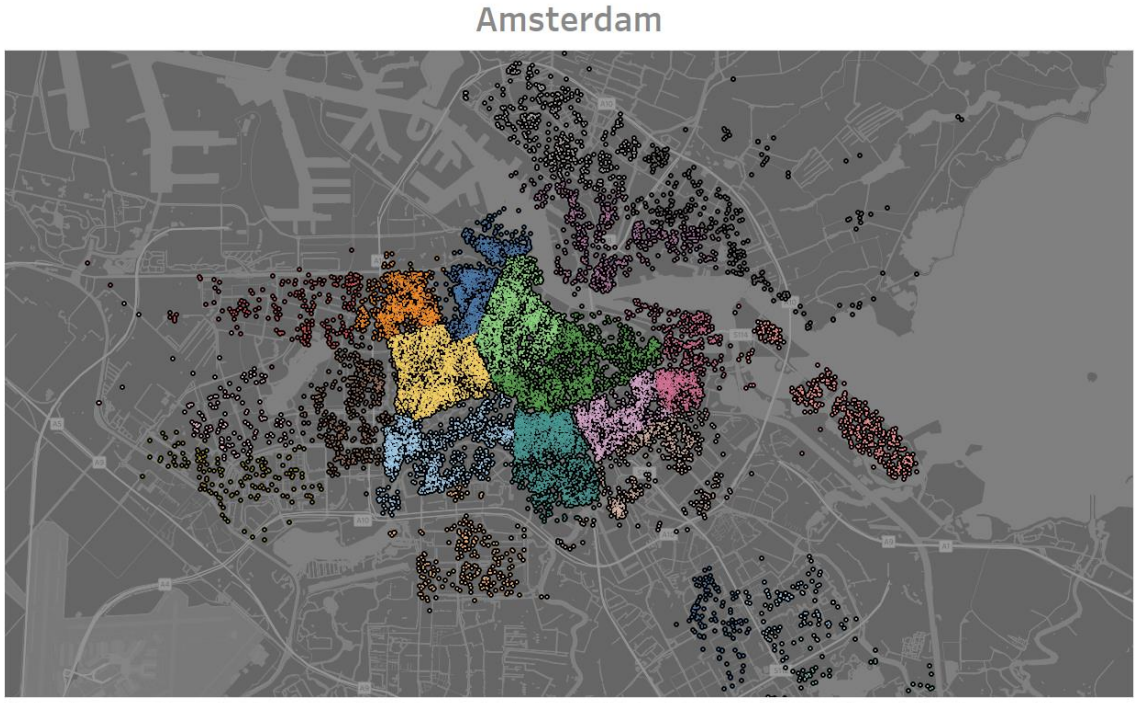
X. APPENDIX

A1



Map based on average of Longitude and average of Latitude. Color shows average of Price. Details are shown for Id. The data is filtered on City, which keeps amsterdam.

A2

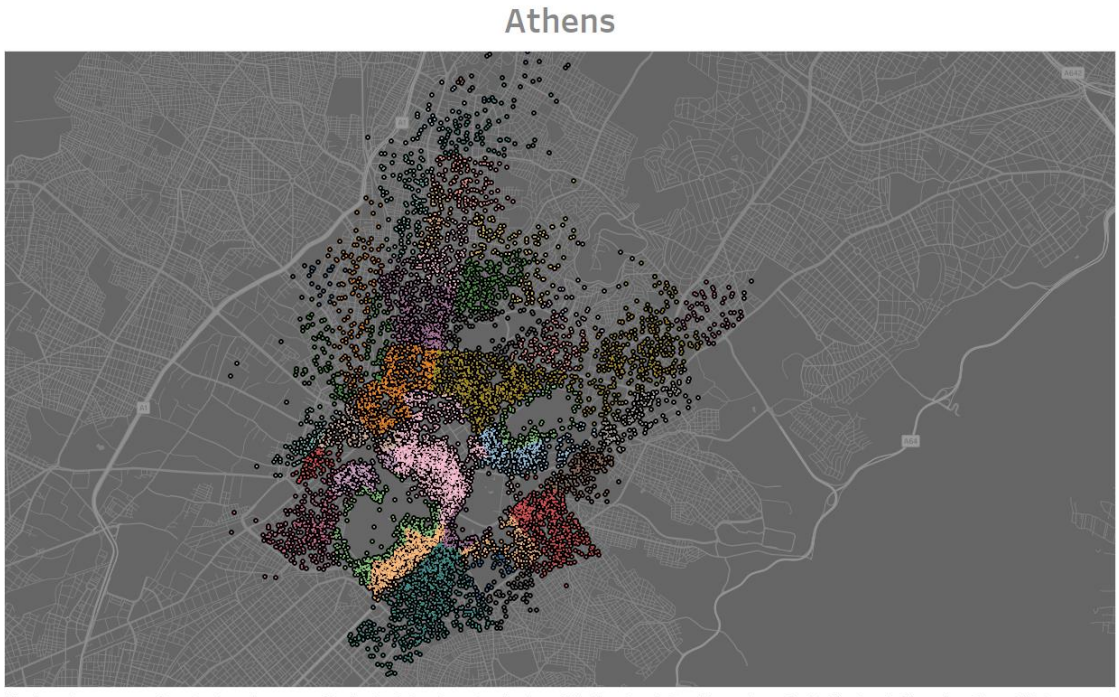


Map based on average of Longitude and average of Latitude. Color shows details about Neighbourhood. Details are shown for Id. The data is filtered on City, which keeps amsterdam.

A3

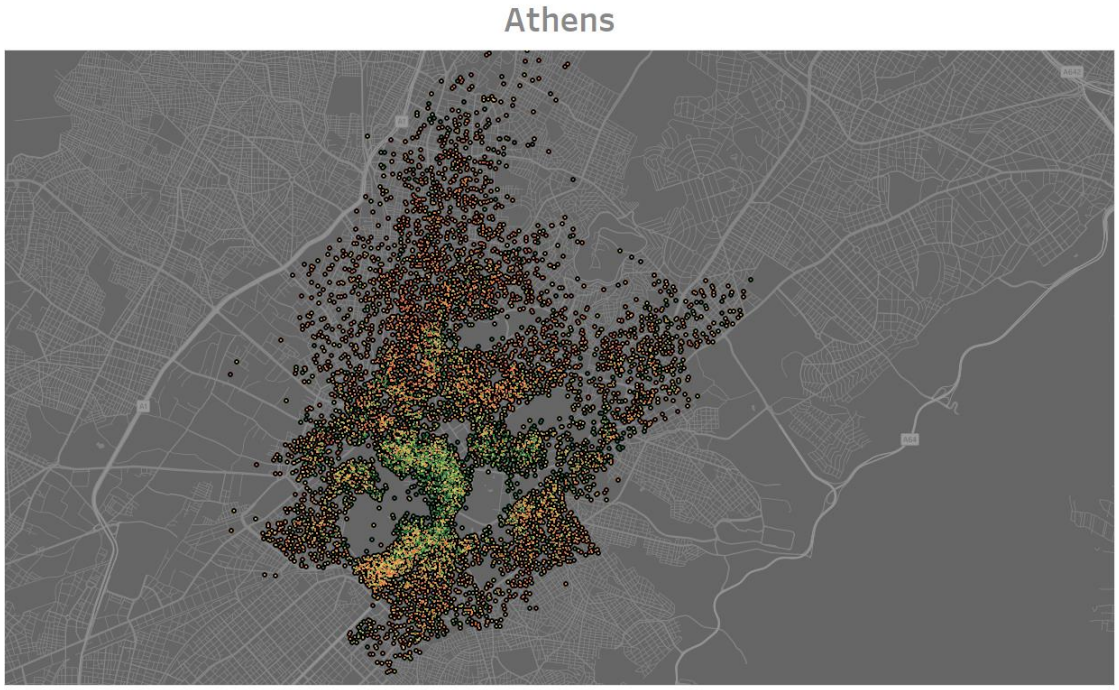
Map based on average of Longitude and average of Latitude. Color shows details about Neighbourhood. Size shows sum of Number of Records. The marks are labeled by average of Price. The data is filtered on City, which keeps amsterdam.

A4



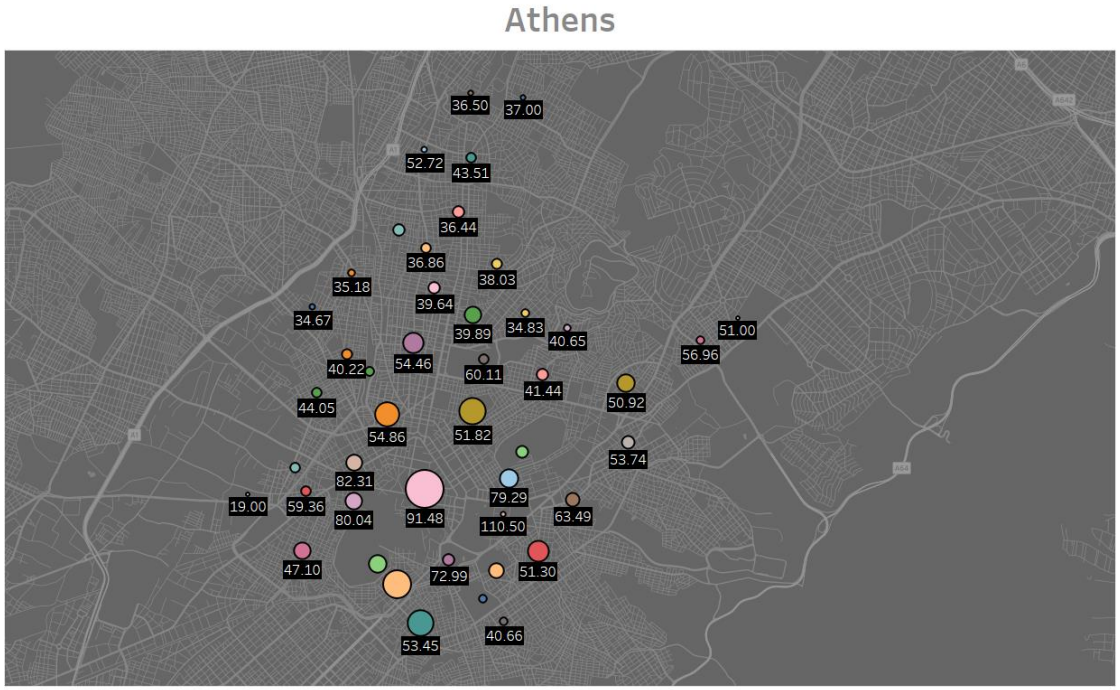
Map based on average of Longitude and average of Latitude. Color shows details about Neighbourhood. Details are shown for Id. The data is filtered on City, which keeps athens.

A5



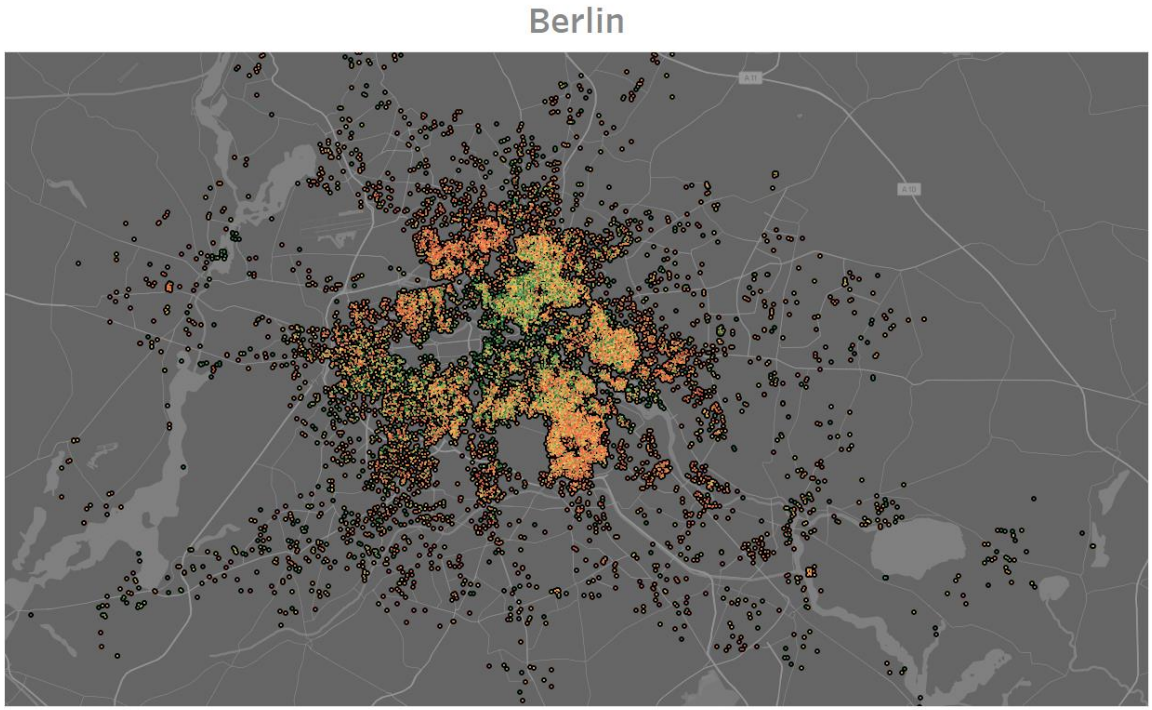
Map based on average of Longitude and average of Latitude. Color shows sum of Price. Details are shown for Id. The data is filtered on City, which keeps athens.

A6



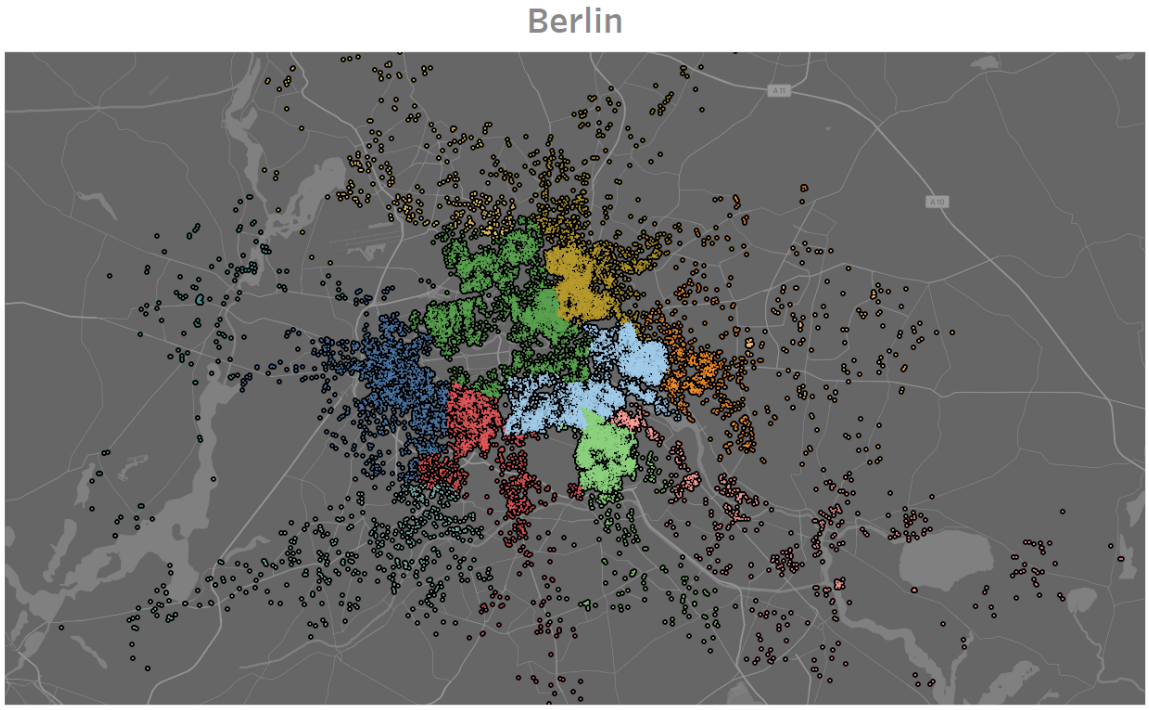
Map based on average of Longitude and average of Latitude. Color shows details about Neighbourhood. Size shows sum of Number of Records. The marks are labeled by average of Price. The data is filtered on City, which keeps athens.

A7



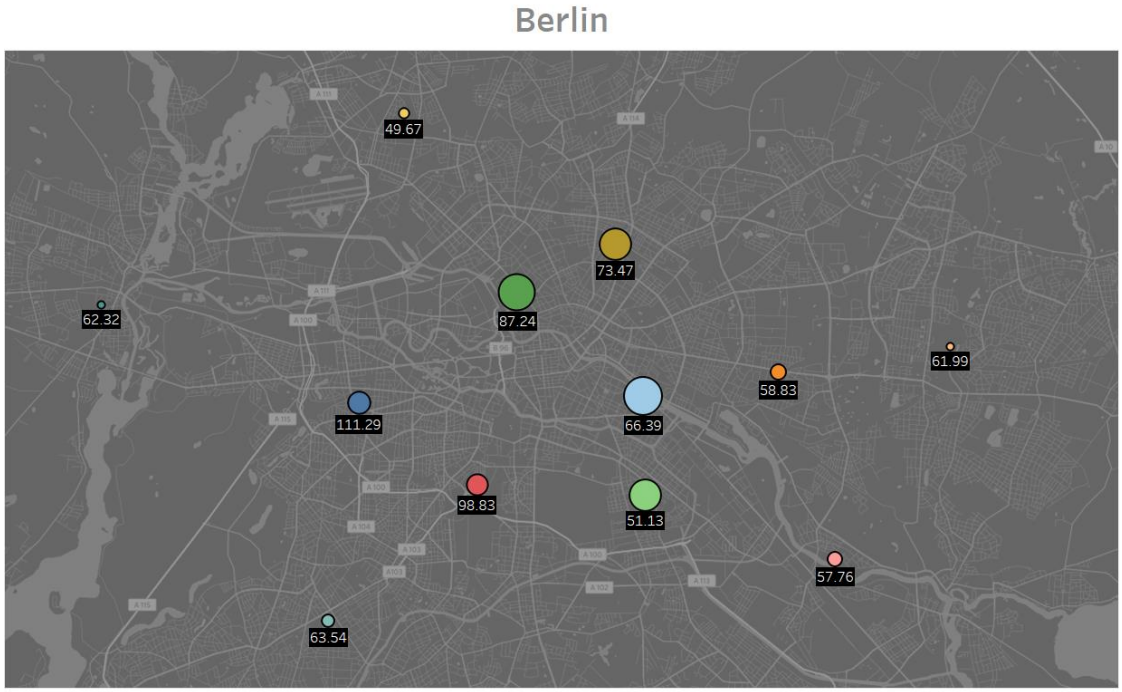
Map based on average of Longitude and average of Latitude. Color shows sum of Price. Details are shown for Id. The data is filtered on City, which keeps berlin.

A8



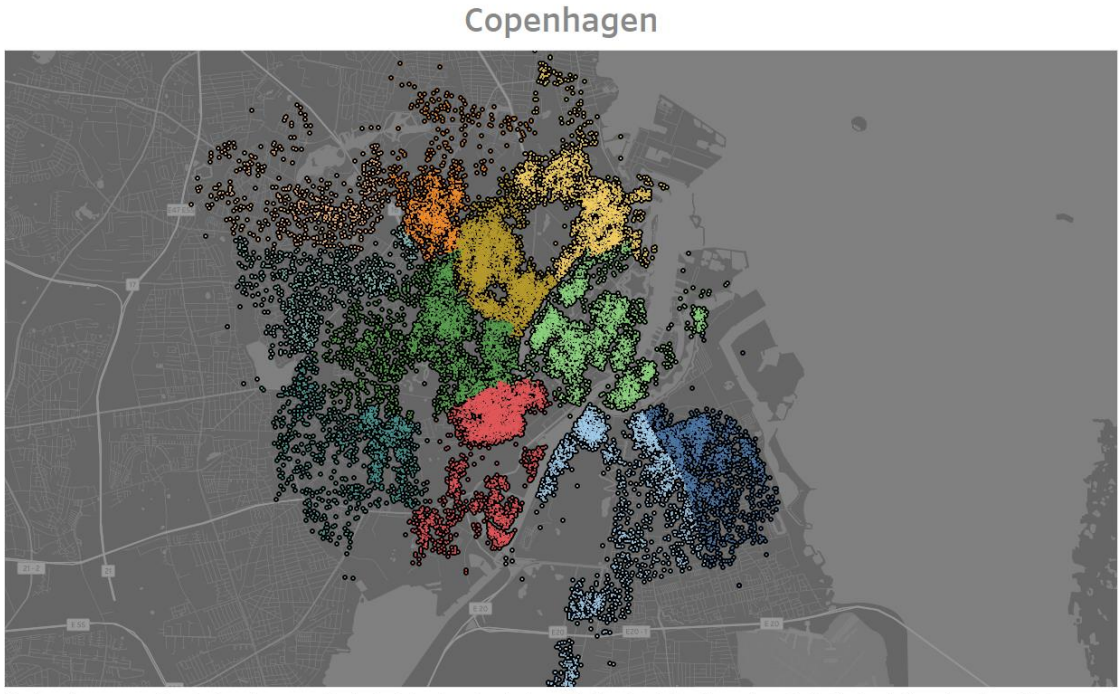
Map based on average of Longitude and average of Latitude. Color shows details about Neighbourhood Group. Details are shown for Id. The data is filtered on City, which keeps berlin.

A9



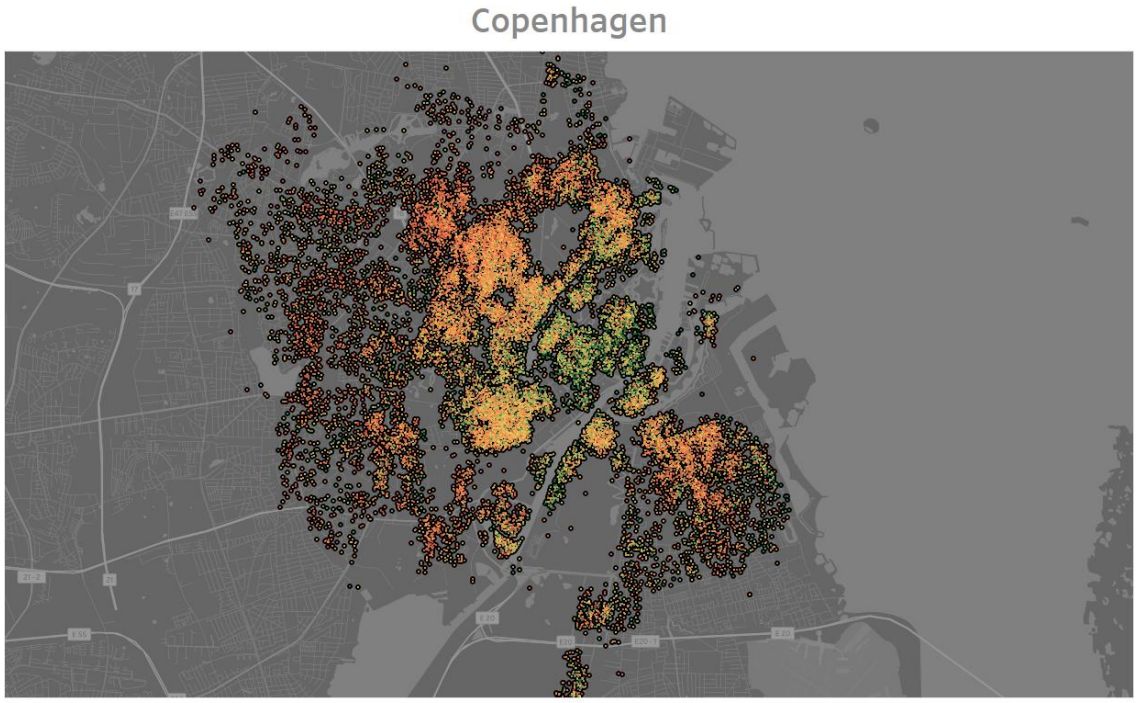
Map based on average of Longitude and average of Latitude. Color shows details about Neighbourhood Group. Size shows sum of Number of Records. The marks are labeled by average of Price. The data is filtered on City, which keeps berlin.

A10



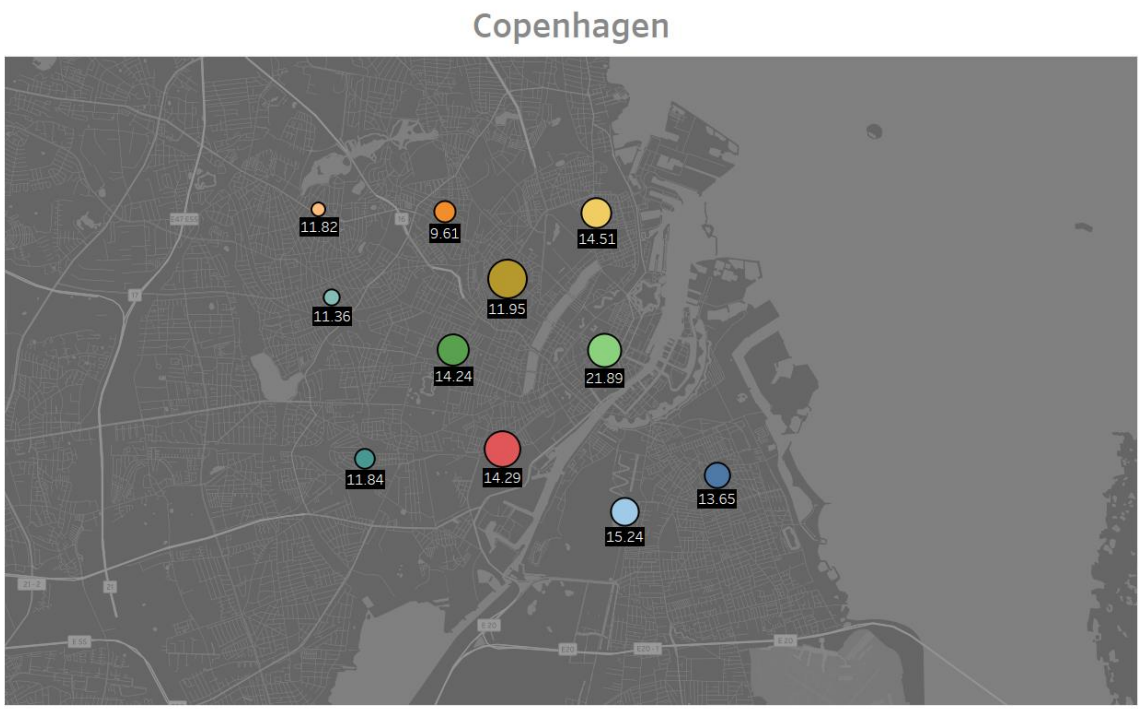
Map based on average of Longitude and average of Latitude. Color shows details about Neighbourhood. Details are shown for Id. The data is filtered on City, which keeps copenhagen.

A11



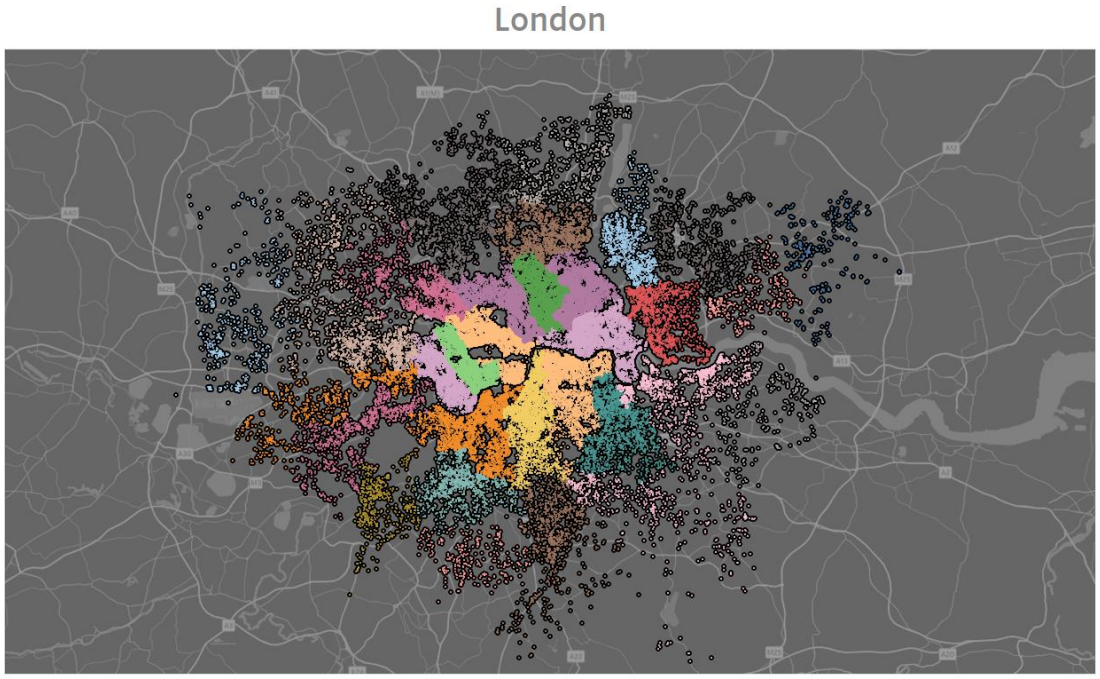
Map based on average of Longitude and average of Latitude. Color shows sum of Price. Details are shown for Id. The data is filtered on City, which keeps copenhagen.

A12



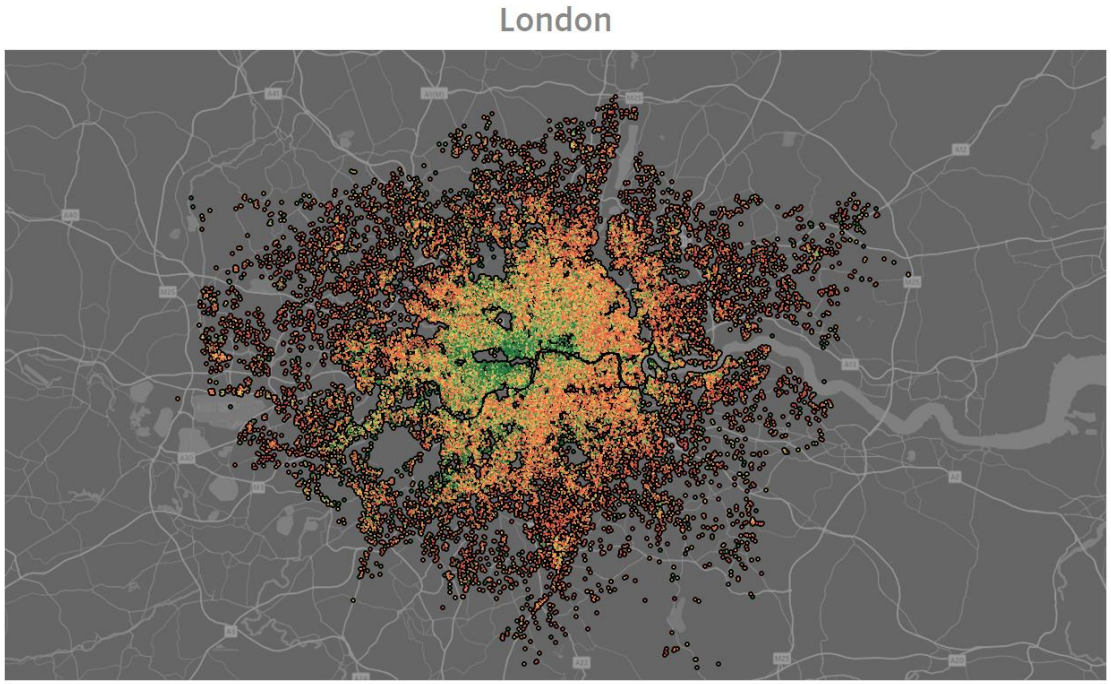
Map based on average of Longitude and average of Latitude. Color shows details about Neighbourhood. Size shows sum of Number of Records. The marks are labeled by average of Price. The data is filtered on City, which keeps copenhagen.

A13



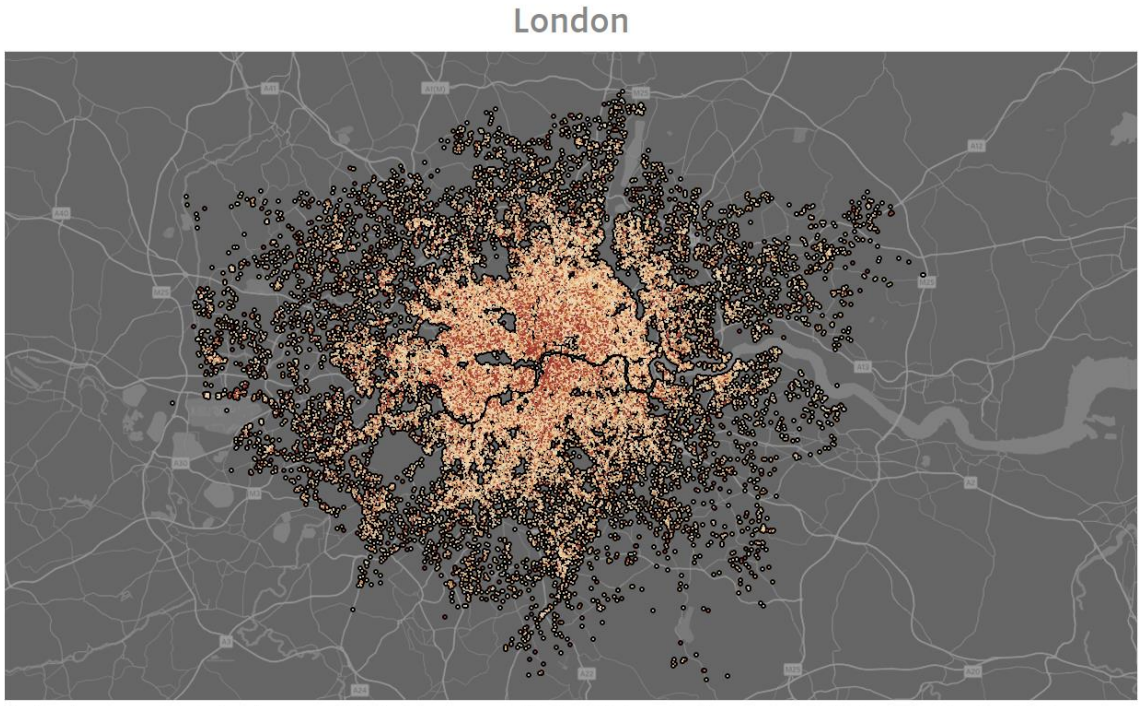
Map based on average of Longitude and average of Latitude. Color shows details about Neighbourhood. Details are shown for Id. The data is filtered on City, which keeps london.

A14



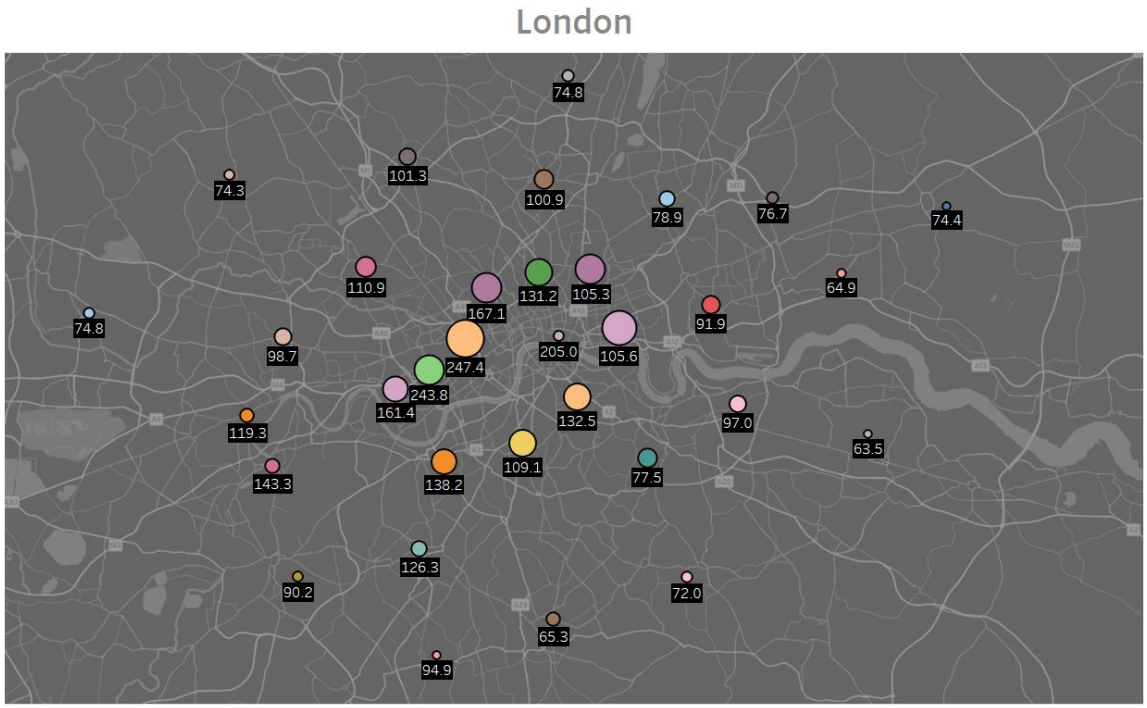
Map based on average of Longitude and average of Latitude. Color shows sum of Price. Details are shown for Id. The data is filtered on City, which keeps london.

A15



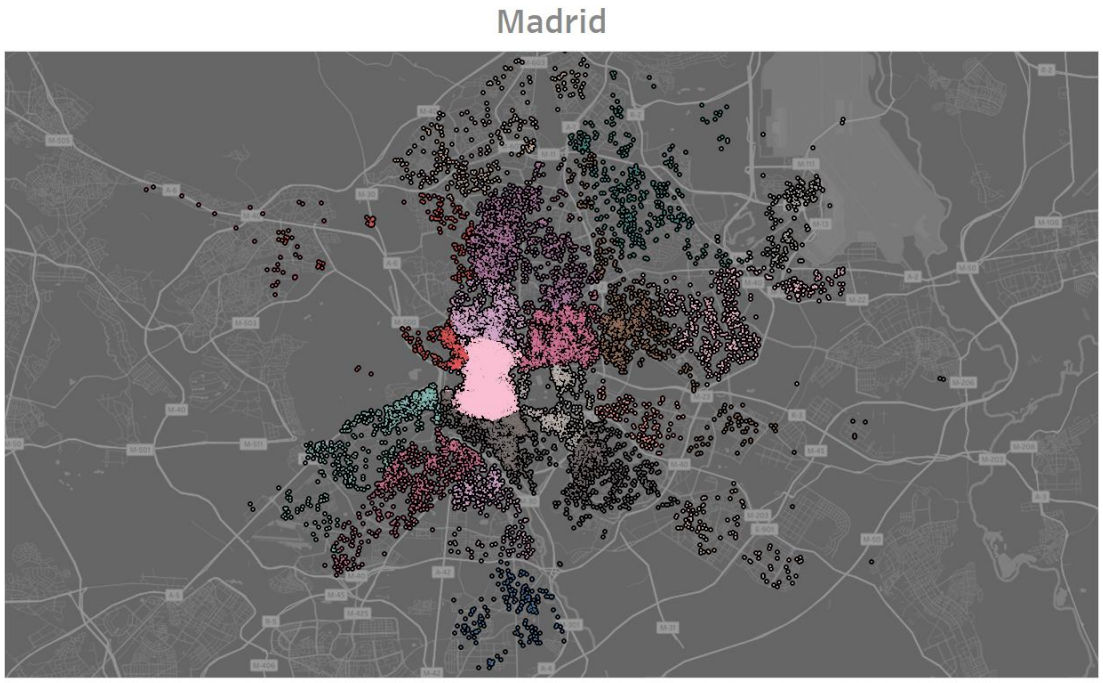
Map based on average of Longitude and average of Latitude. Color shows sum of Number Of Reviews. Details are shown for Id. The data is filtered on City, which keeps london.

A16



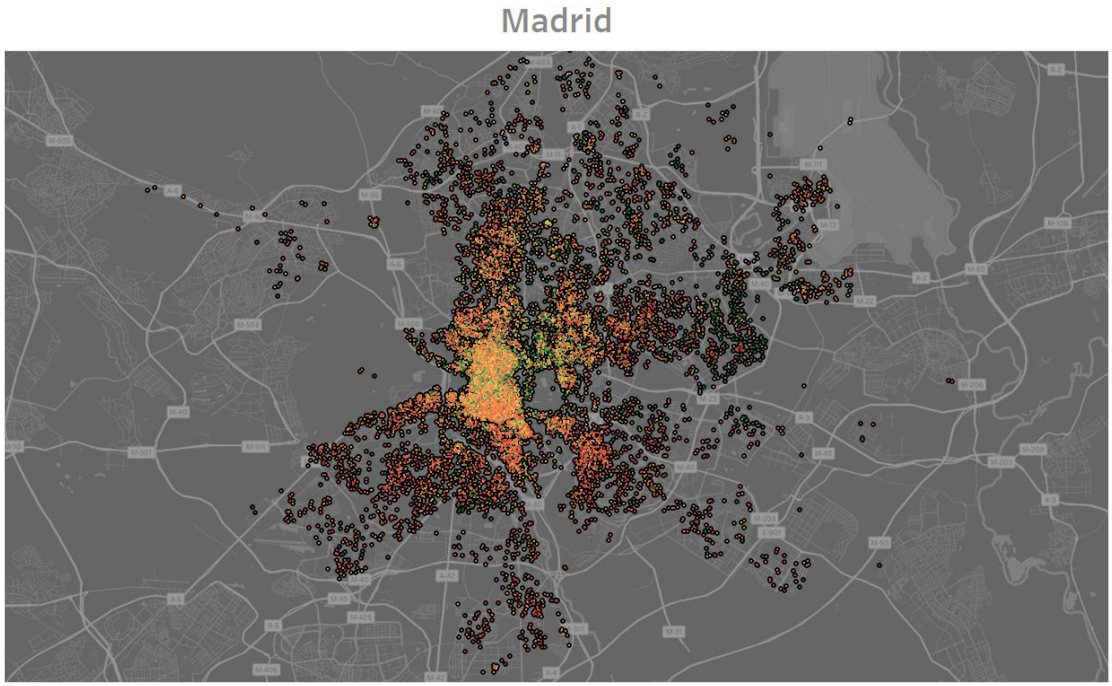
Map based on average of Longitude and average of Latitude. Color shows details about Neighbourhood. Size shows sum of Number of Records. The marks are labeled by average of Price. The data is filtered on City, which keeps london.

A17



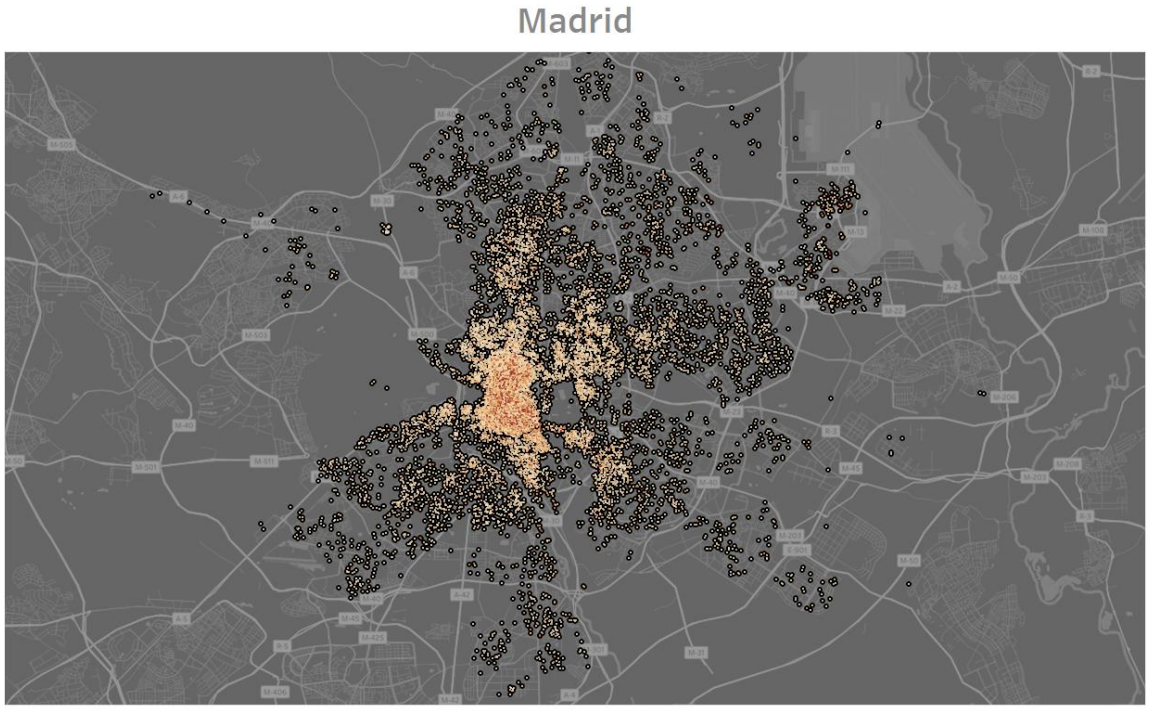
Map based on average of Longitude and average of Latitude. Color shows details about Neighbourhood Group. Details are shown for Id. The data is filtered on City, which keeps madrid.

A18



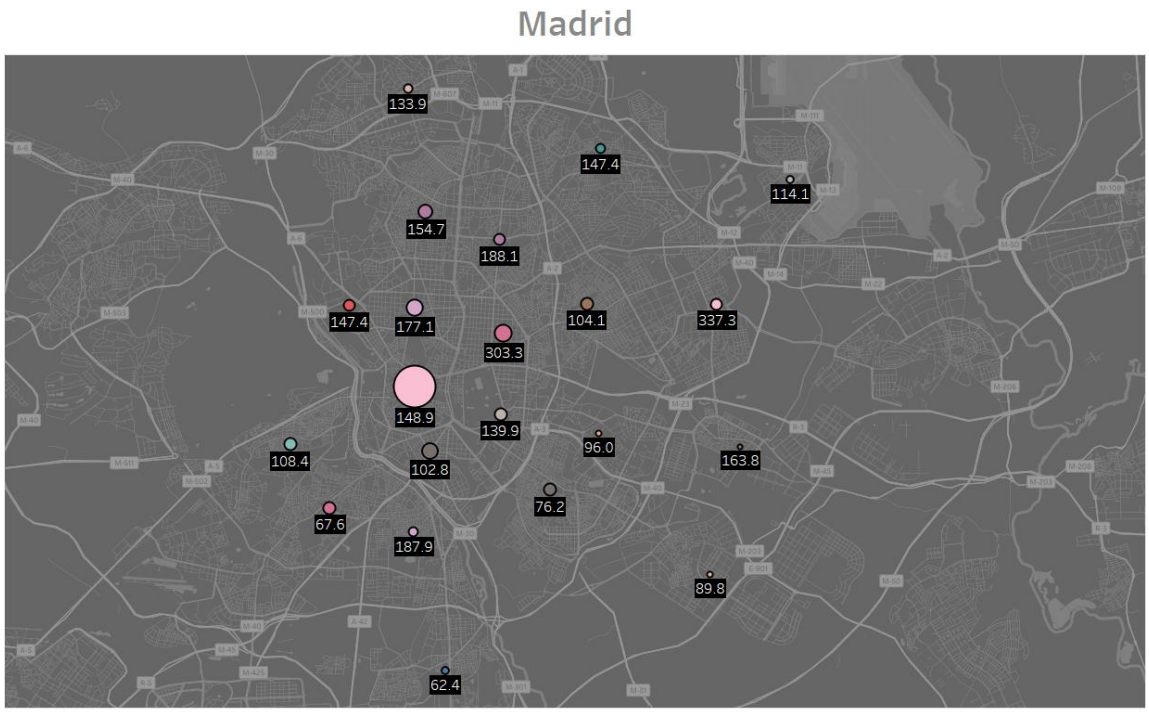
Map based on average of Longitude and average of Latitude. Color shows sum of Price. Details are shown for Id. The data is filtered on City, which keeps madrid.

A19



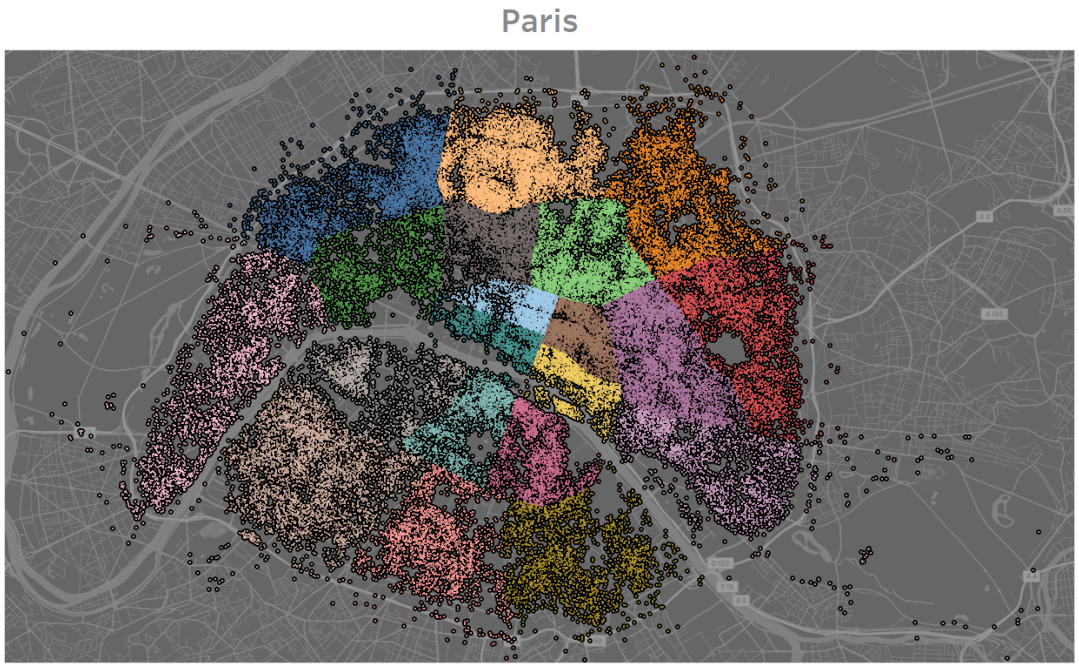
Map based on average of Longitude and average of Latitude. Color shows sum of Number Of Reviews. Details are shown for Id. The data is filtered on City, which keeps madrid.

A20



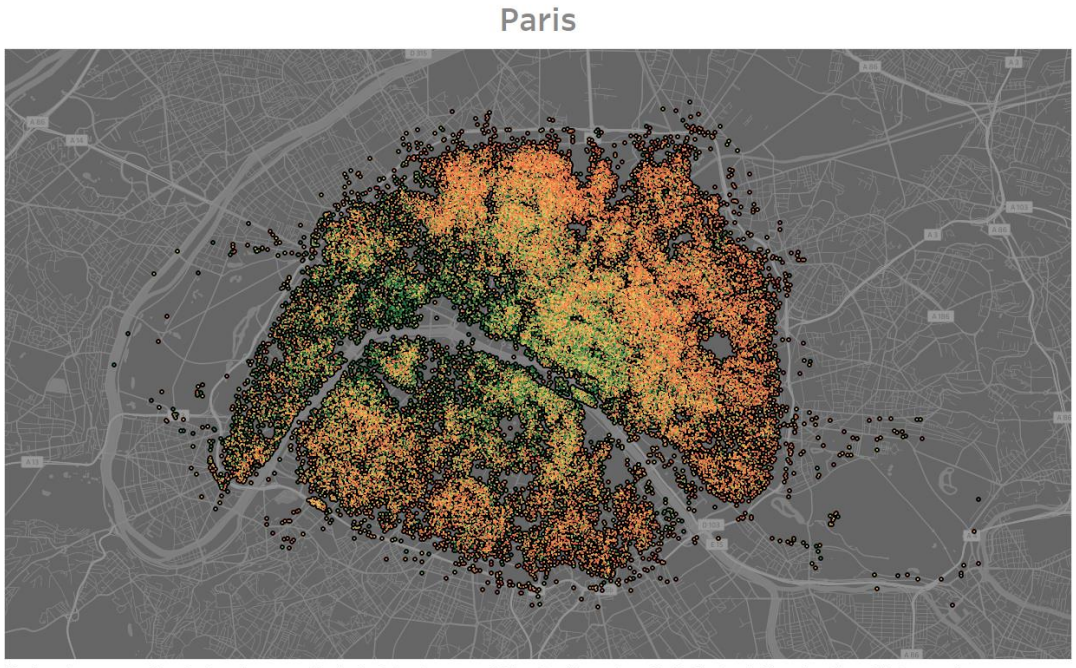
Map based on average of Longitude and average of Latitude. Color shows details about Neighbourhood Group. Size shows sum of Number of Records. The marks are labeled by average of Price. The data is filtered on City, which keeps madrid.

A21



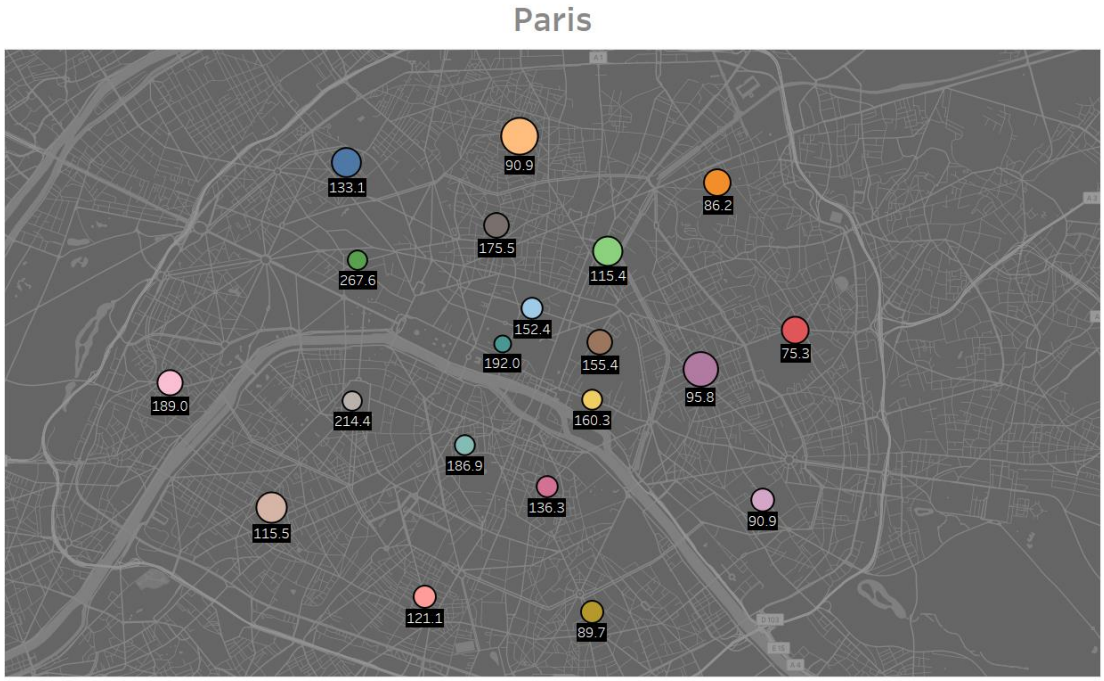
Map based on average of Longitude and average of Latitude. Color shows details about Neighbourhood. Details are shown for Id. The data is filtered on City, which keeps paris.

A22



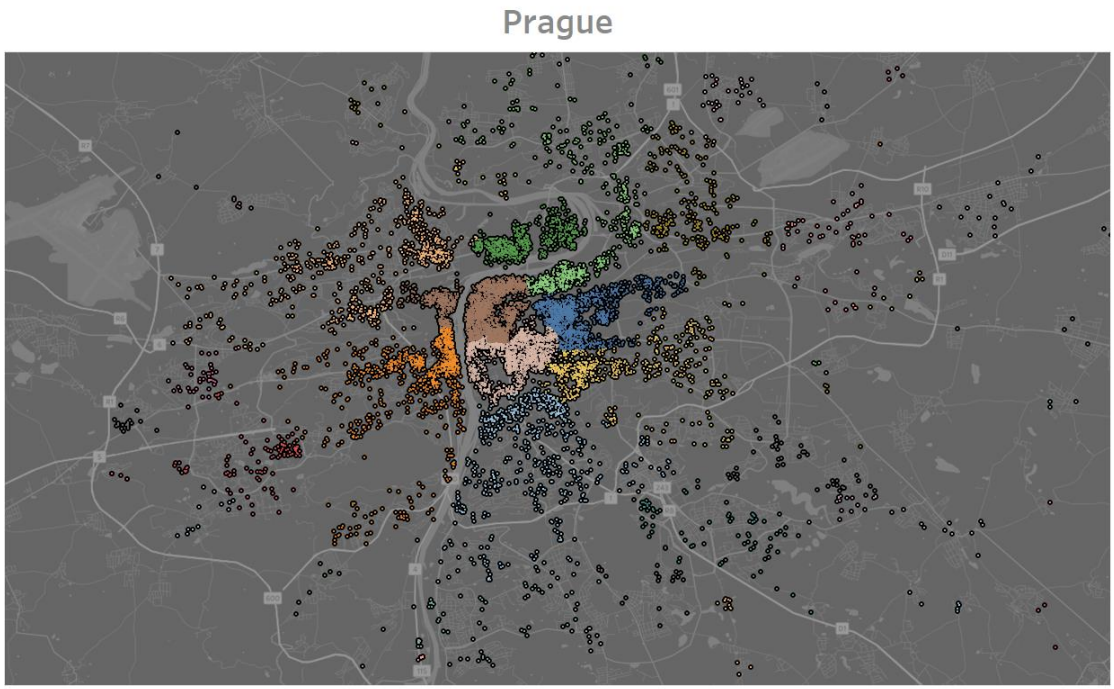
Map based on average of Longitude and average of Latitude. Color shows sum of Price. Details are shown for Id. The data is filtered on City, which keeps paris.

A23



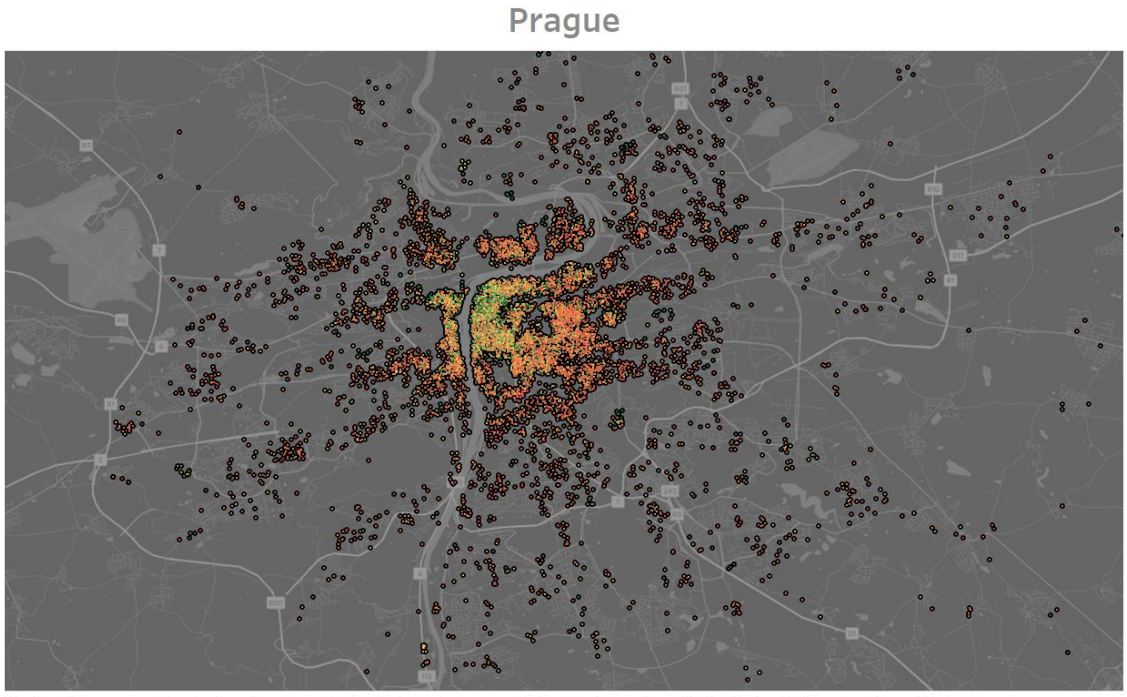
Map based on average of Longitude and average of Latitude. Color shows details about Neighbourhood. Size shows sum of Number of Records. The marks are labeled by average of Price. The data is filtered on City, which keeps paris.

A24



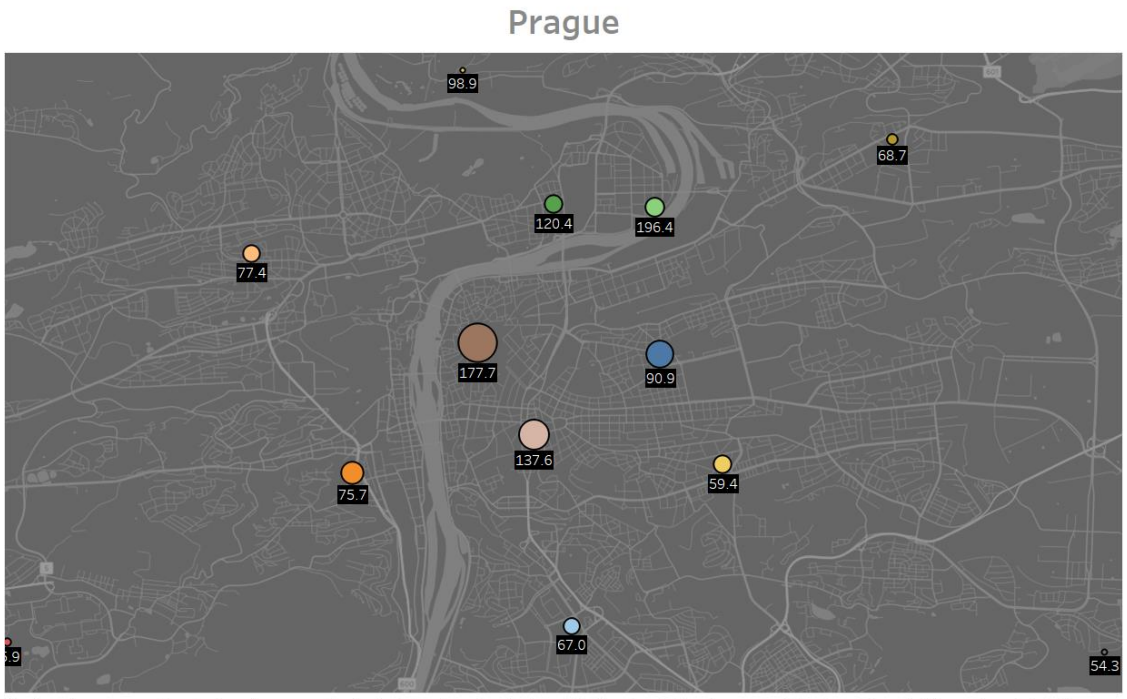
Map based on average of Longitude and average of Latitude. Color shows details about Neighbourhood. Details are shown for Id. The data is filtered on City, which keeps prague.

A25



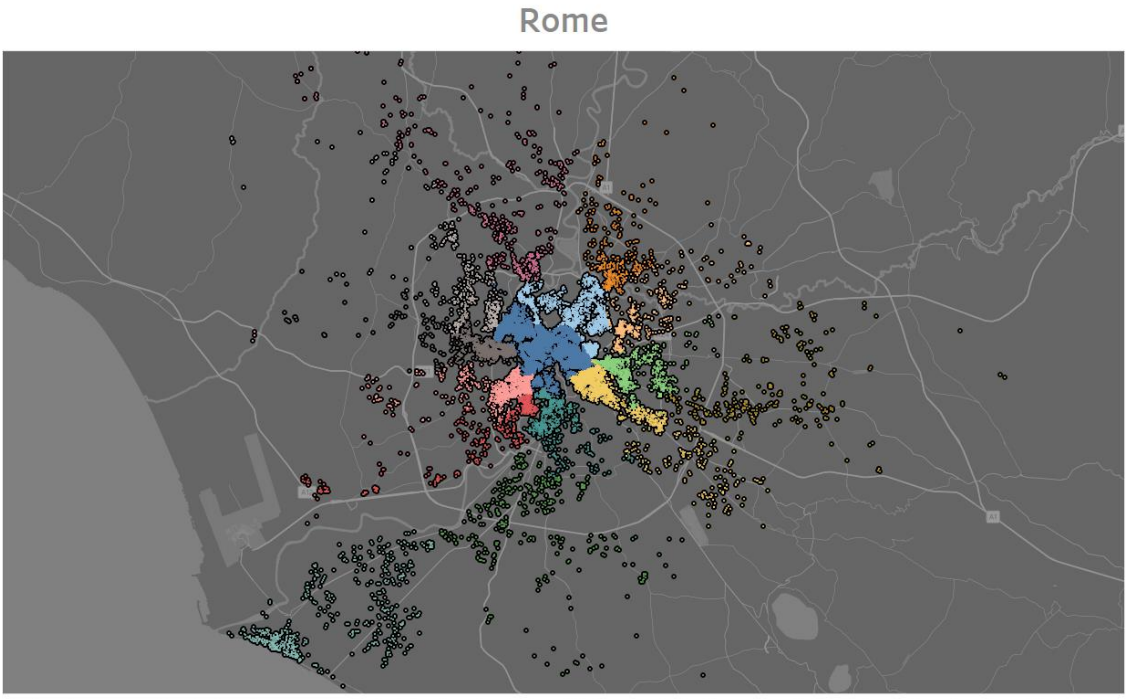
Map based on average of Longitude and average of Latitude. Color shows sum of Price. Details are shown for Id. The data is filtered on City, which keeps prague.

A26



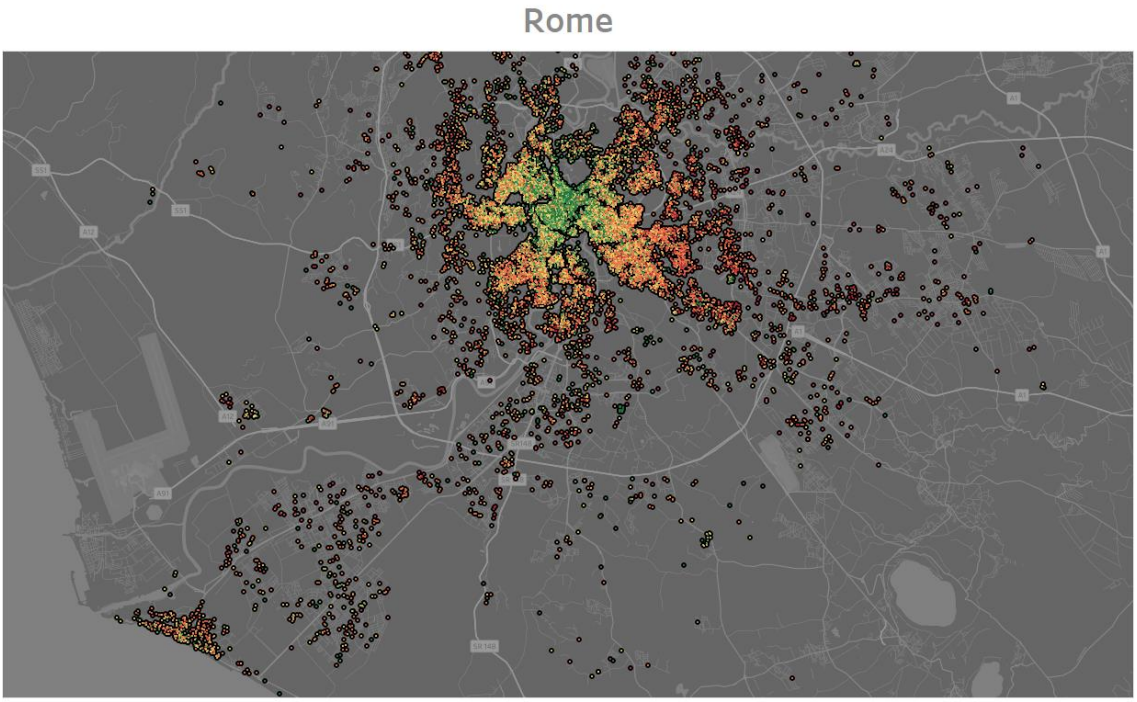
Map based on average of Longitude and average of Latitude. Color shows details about Neighbourhood. Size shows sum of Number of Records. The marks are labeled by average of Price. The data is filtered on City, which keeps prague.

A27



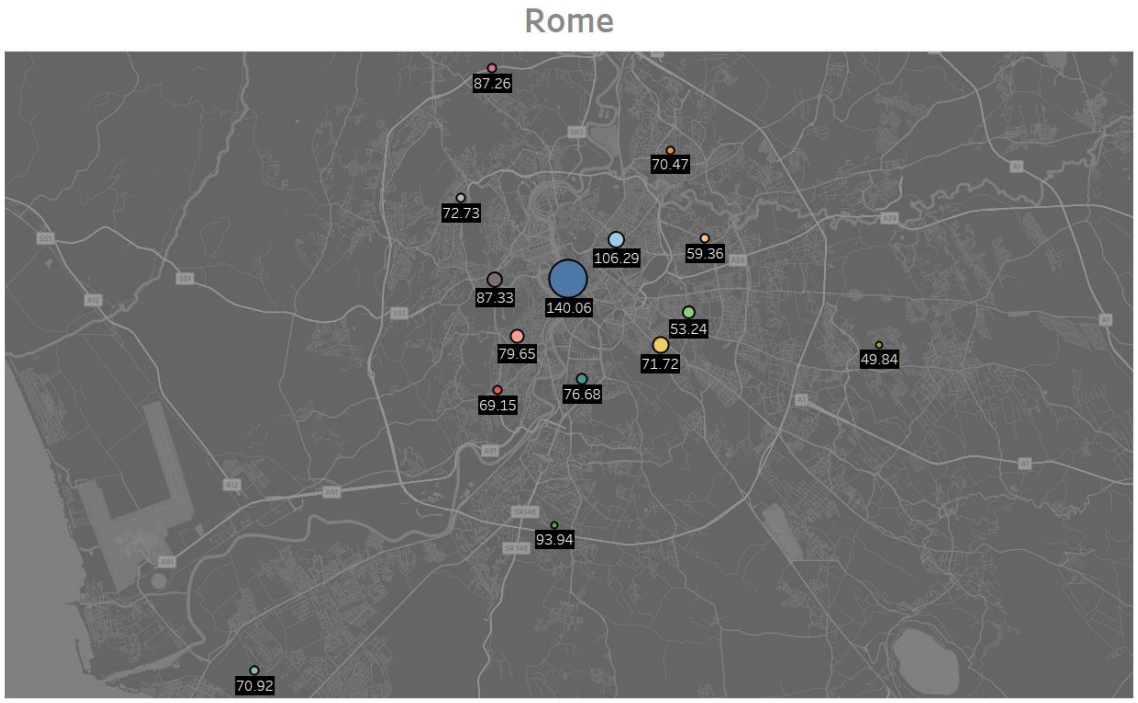
Map based on average of Longitude and average of Latitude. Color shows details about Neighbourhood. Details are shown for Id. The data is filtered on City, which keeps rome.

A28



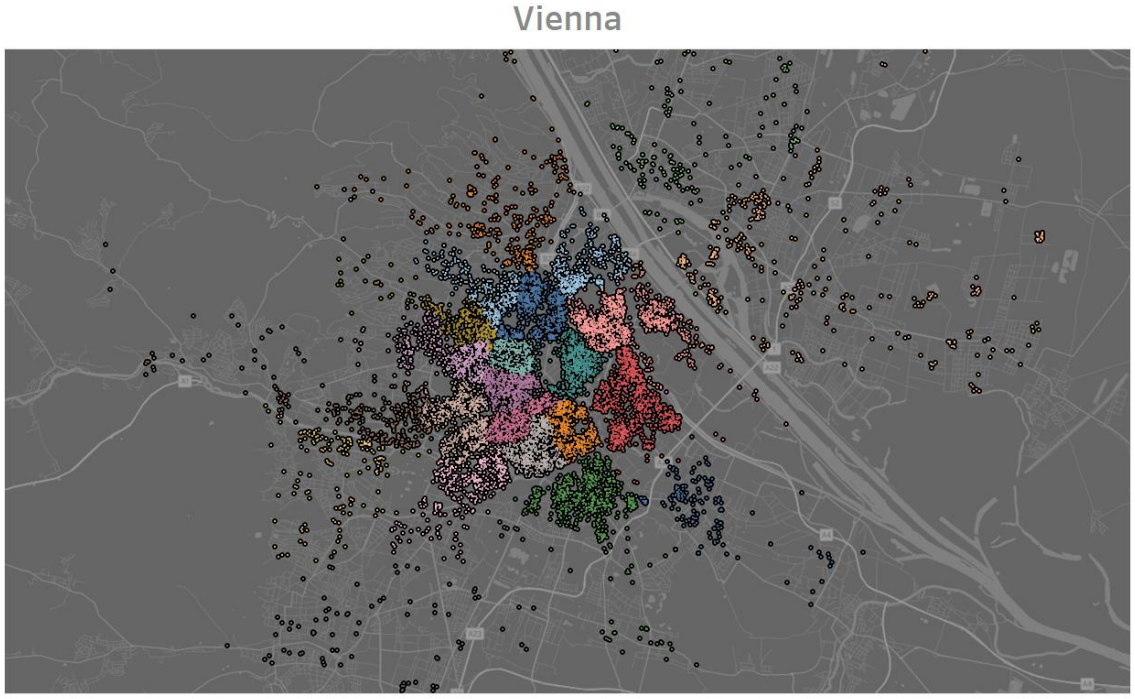
Map based on average of Longitude and average of Latitude. Color shows average of Price. Details are shown for Id. The data is filtered on City, which keeps rome.

A29



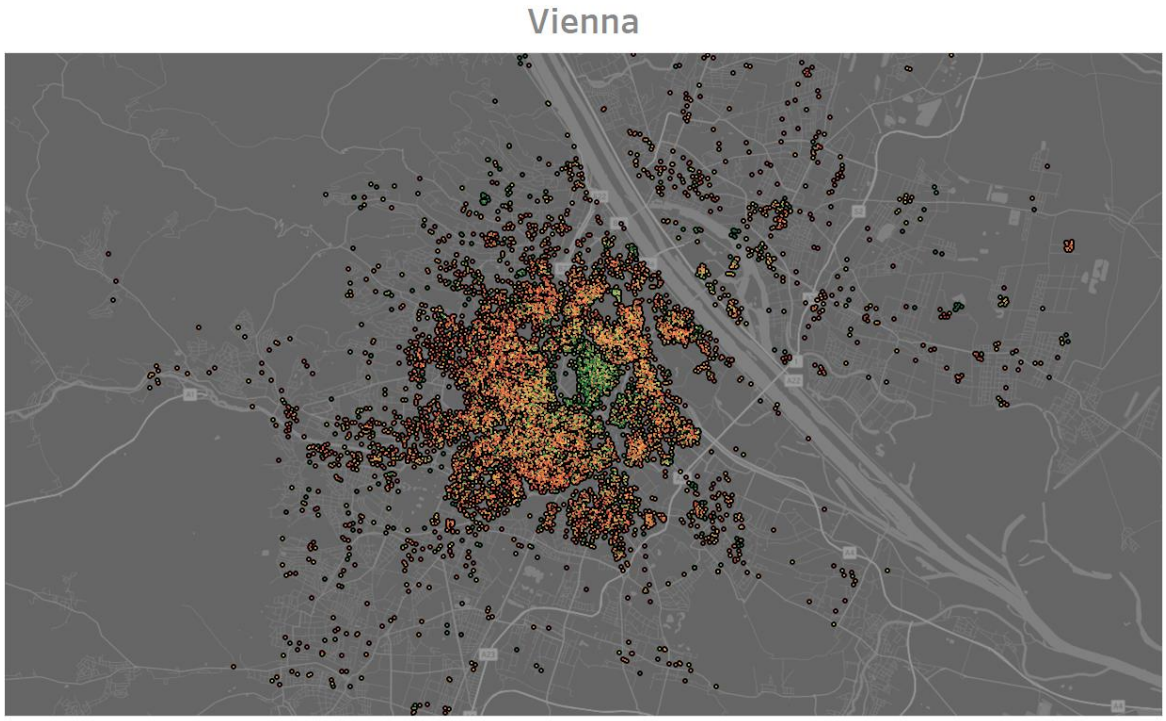
Map based on average of Longitude and average of Latitude. Color shows details about Neighbourhood. Size shows sum of Number of Records. The marks are labeled by average of Price. The data is filtered on City, which keeps rome.

A30



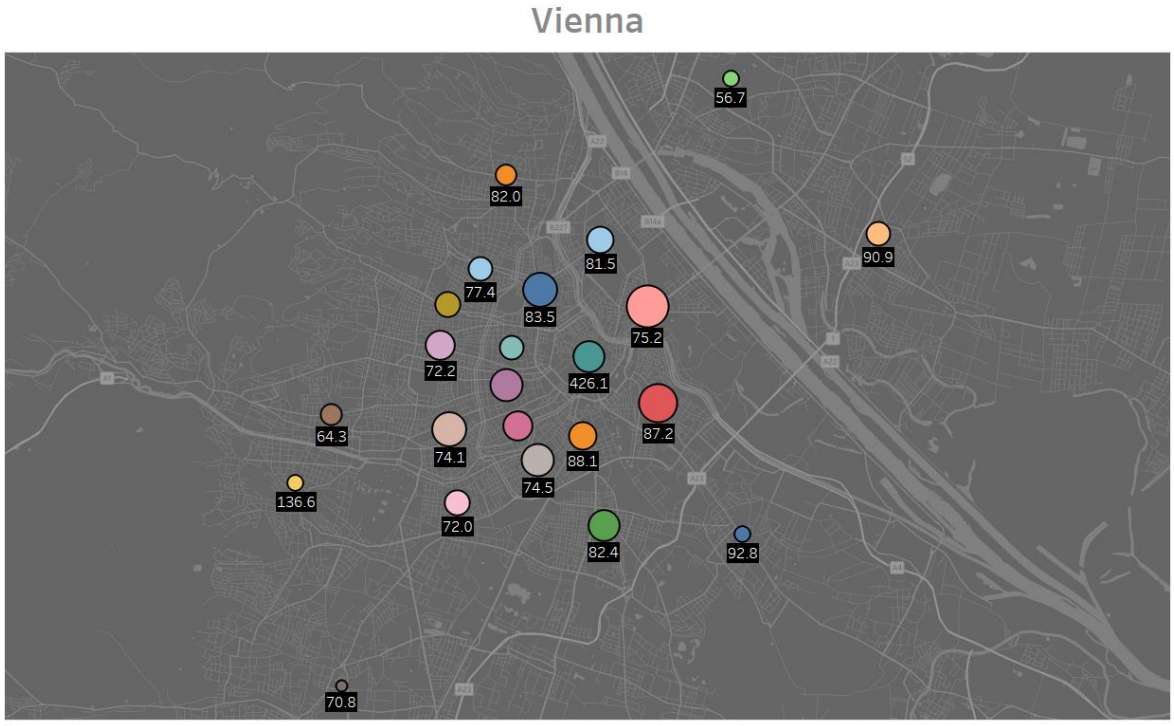
Map based on average of Longitude and average of Latitude. Color shows details about Neighbourhood. Details are shown for Id. The data is filtered on City, which keeps vienna.

A31



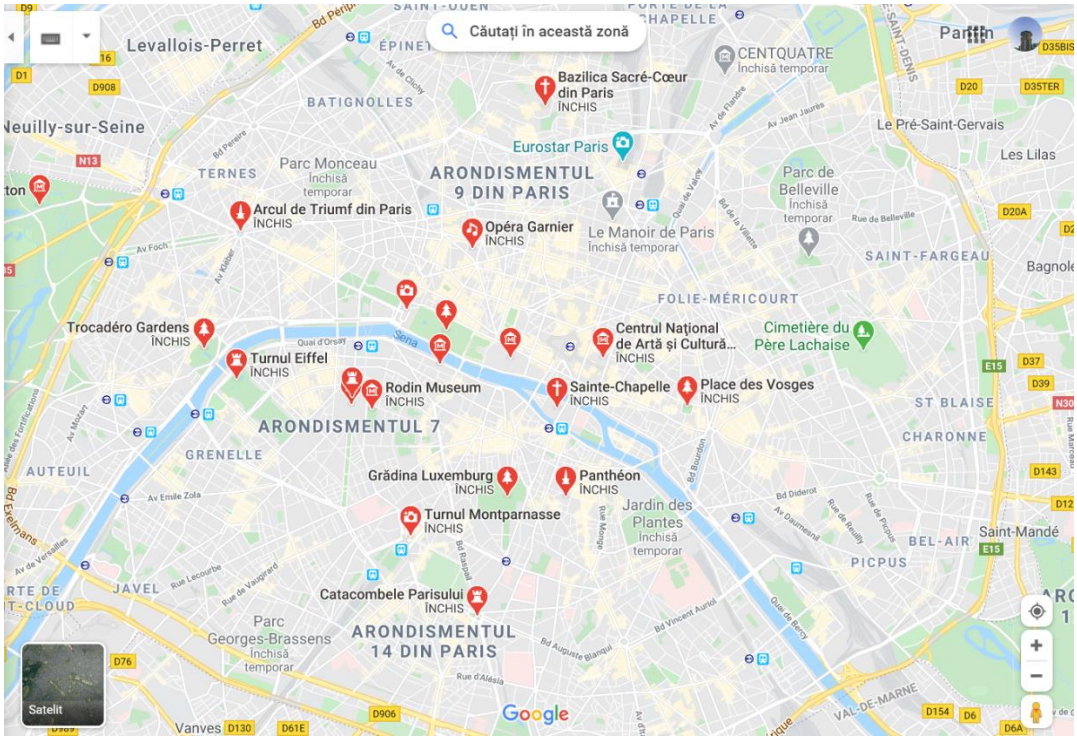
Map based on average of Longitude and average of Latitude. Color shows sum of Price. Details are shown for Id. The data is filtered on City, which keeps vienna.

A32

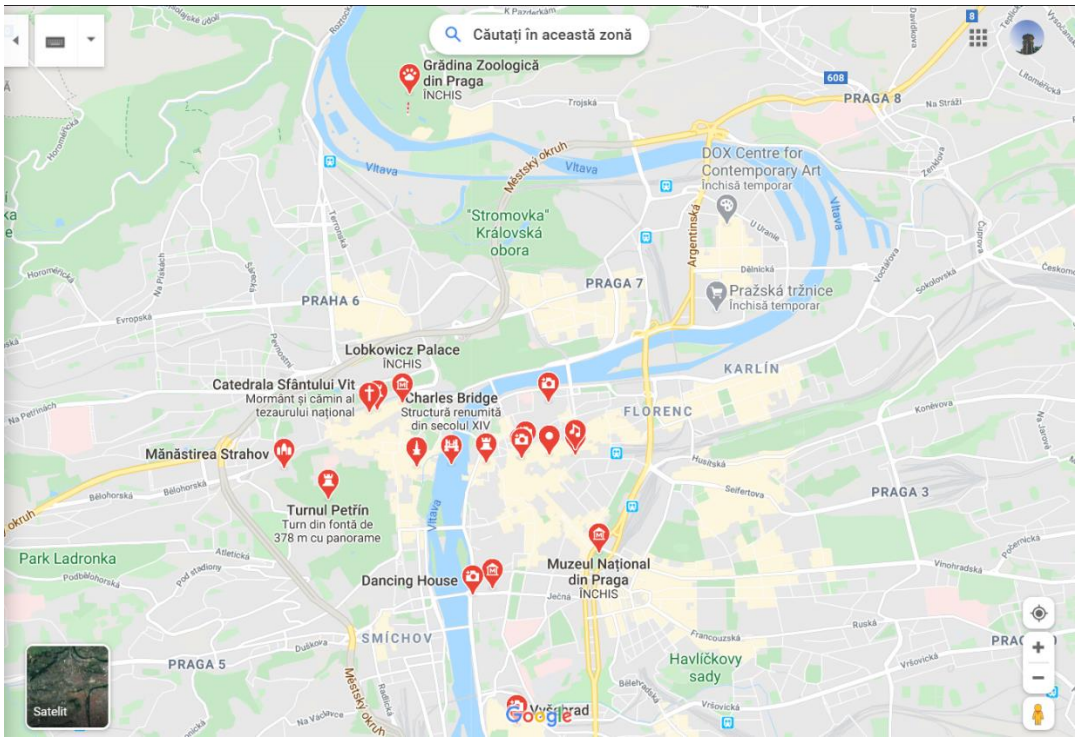


Map based on average of Longitude and average of Latitude. Color shows details about Neighbourhood. Size shows sum of Number of Records. The marks are labeled by average of Price. The data is filtered on City, which keeps vienna.

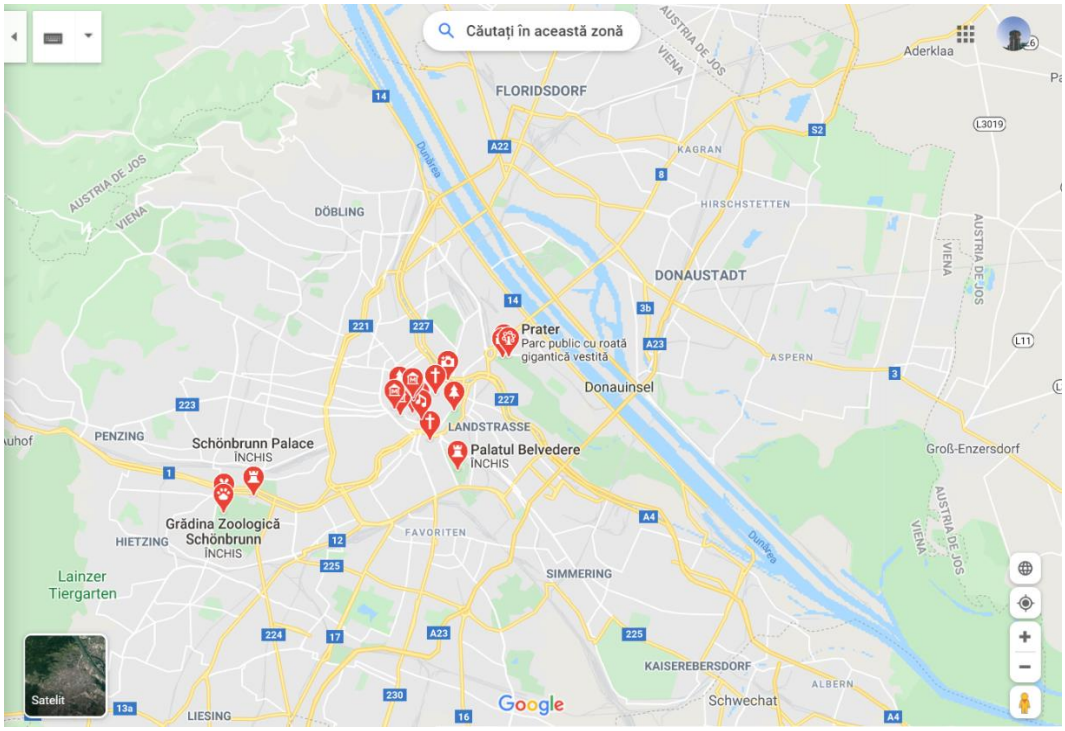
A33


Căutați în această zonă
Levallois-Perret
Bazilica Sacré-Cœur din Paris
ÎNCHIS
CENTQUATRE
Închisă temporar
Eurostar Paris
Parc Monceau
Închisă temporar
ARONDISMENTUL 9 DIN PARIS
Arcul de Triumf din Paris
ÎNCHIS
Opéra Garnier
ÎNCHIS
Le Manoir de Paris
Închisă temporar
Parc de Belleville
Închisă temporar
Le Pré-Saint-Gervais
Les Lilas
SAINT-FARGEAU
FOLIE-MÉRICOURT
Trocadéro Gardens
ÎNCHIS
Turnul Eiffel
ÎNCHIS
Rodin Museum
ÎNCHIS
Centrul Național de Artă și Cultură...
ÎNCHIS
Cimetière du Père Lachaise
Sainte-Chapelle
ÎNCHIS
Place des Vosges
ÎNCHIS
ARONDISMENTUL 7
ST BLAISE
CHARONNE
GRENELLE
AUTEUIL
Grădina Luxemburg
ÎNCHIS
Panthéon
ÎNCHIS
Turnul Montparnasse
ÎNCHIS
Jardin des Plantes
Închisă temporar
BEL-AIR
Saint-Mandé
PICPUS
JAVEL
Catacombele Parisului
ÎNCHIS
Parc Georges-Brassens
Închisă temporar
ARONDISMENTUL 14 DIN PARIS
Satelit
Vanves
Google

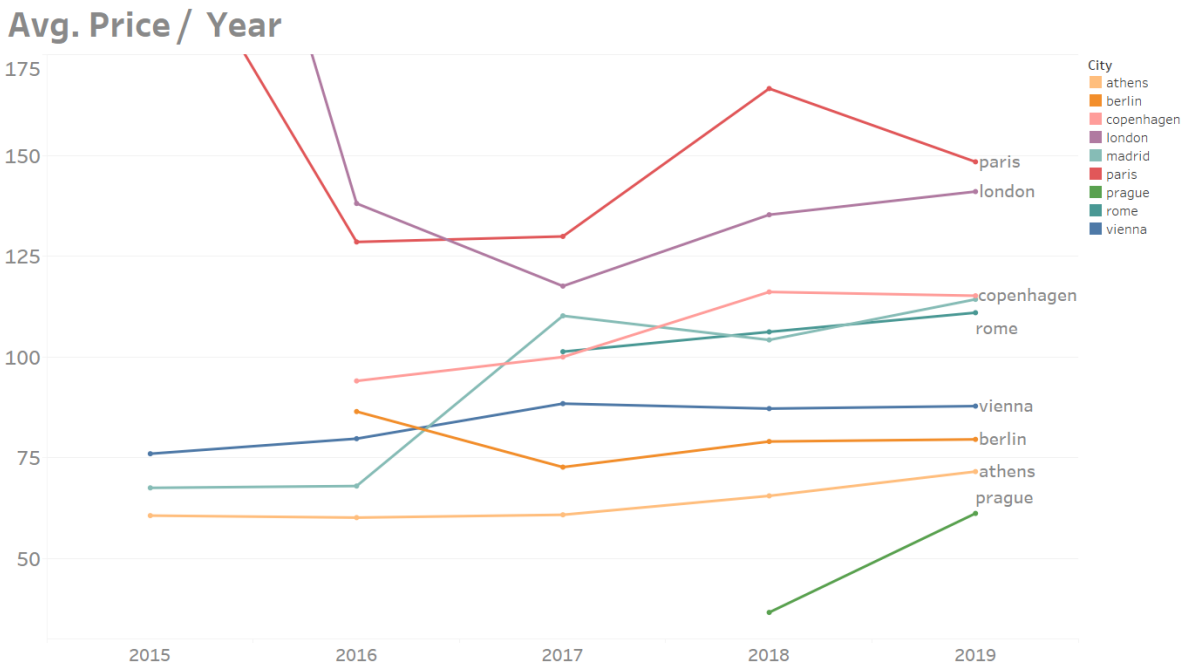
A34


Căutați în această zonă
Grădina Zoologică din Praga
ÎNCHIS
PRAGA 8
DOX Centre for Contemporary Art
Închisă temporar
Vltava
Městský okruh
"Stromovka" Královská obora
PRAGA 7
Pražská tržnice
Închisă temporar
PRAHA 6
Lobkowicz Palace
ÎNCHIS
KARLÍN
Catedrala Sfântului Vit
Mormânt și cămin al tezaurului național
Charles Bridge
Structură renumită din secolul XIV
FLORENC
Mănăstirea Strahov
Turnul Petřín
Turn din fontă de 378 m cu panorame
PRAGA 3
Park Ladronka
Muzeul Național din Praga
ÎNCHIS
Dancing House
SMÍCHOV
Havlíčkovy sady
PRAGA 5
Satelit
Google

A35

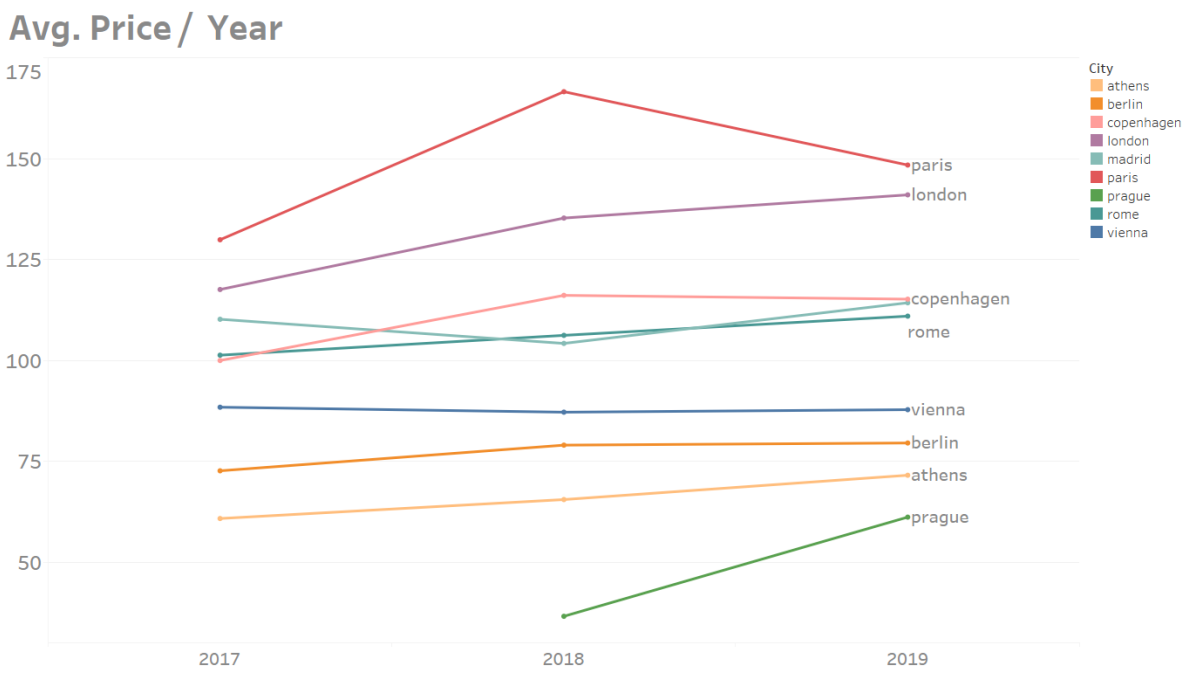


A36



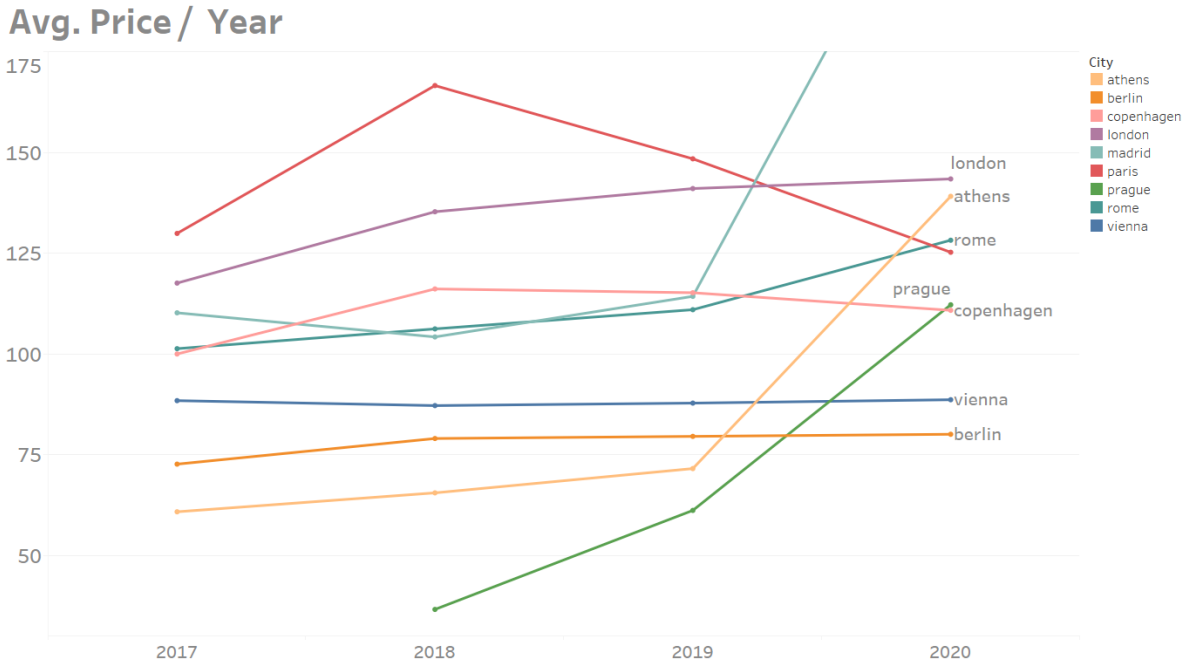
The trend of average of Average Price for Date Year. Color shows details about City. The marks are labeled by City. The view is filtered on City and Date Year. The City filter excludes amsterdam. The Date Year filter keeps 2015, 2016, 2017, 2018 and 2019.

A37



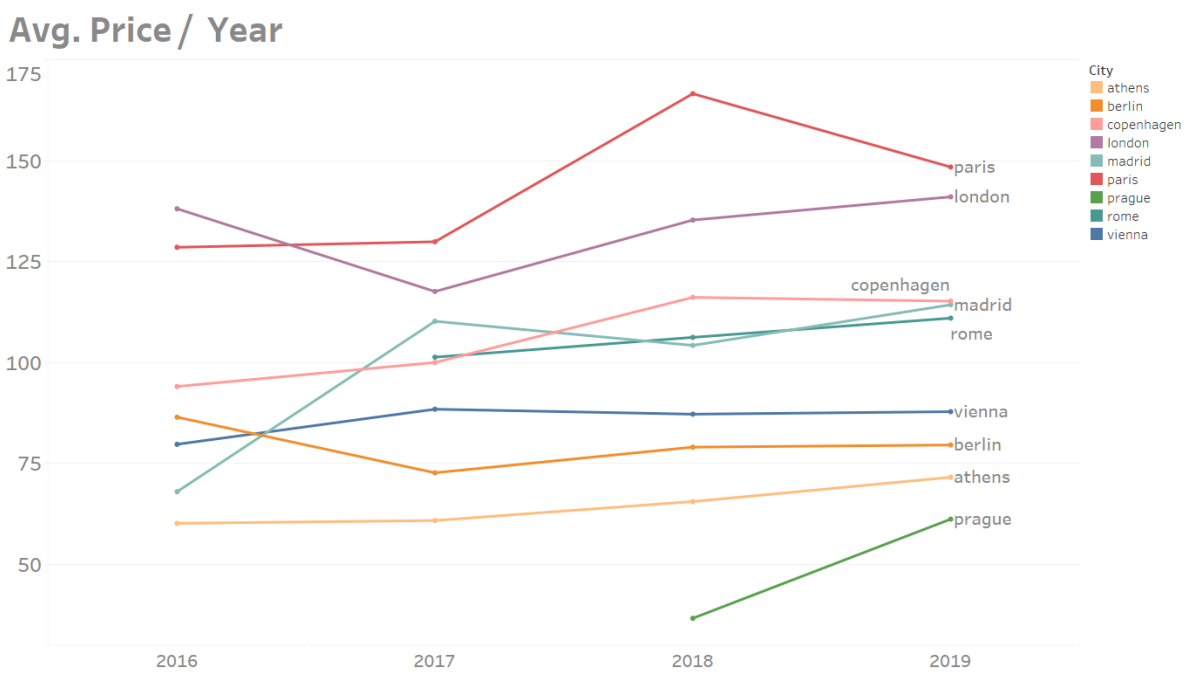
The trend of average of Average Price for Date Year. Color shows details about City. The marks are labeled by City. The view is filtered on City and Date Year. The City filter excludes amsterdam. The Date Year filter keeps 2017, 2018 and 2019.

A38



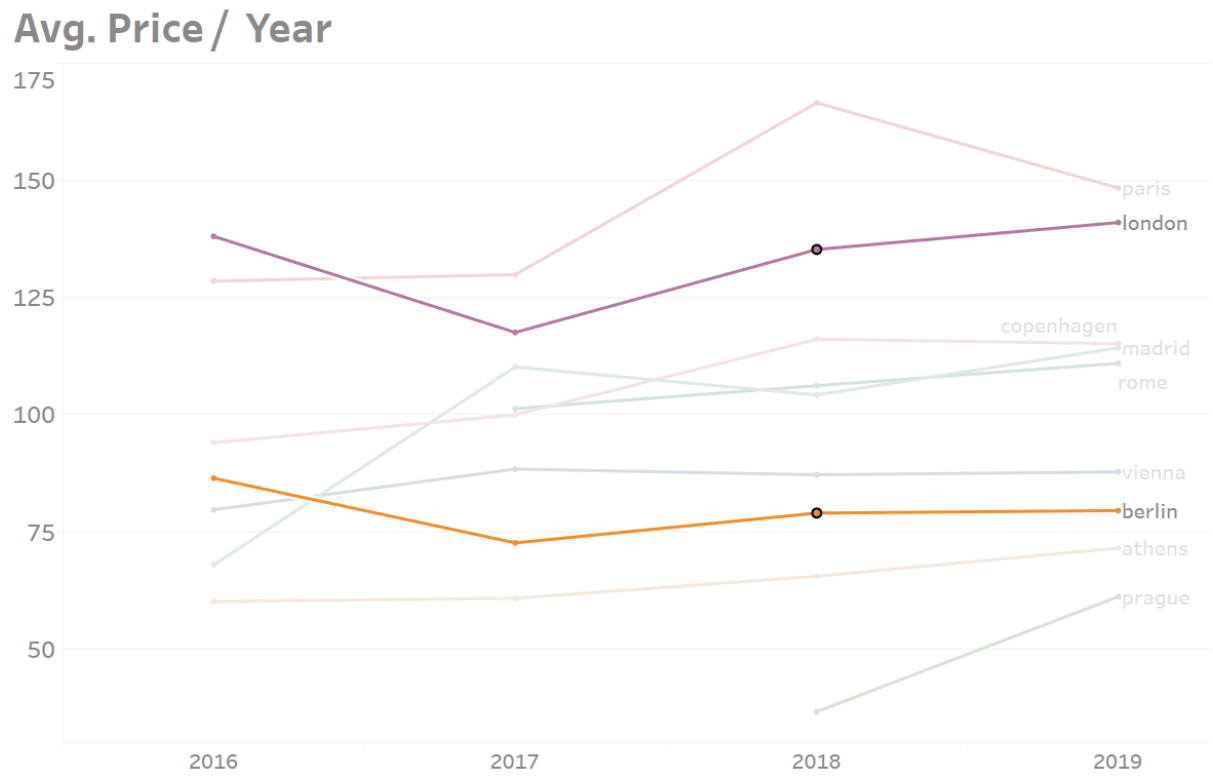
The trend of average of Average Price for Date Year. Color shows details about City. The marks are labeled by City. The view is filtered on City and Date Year. The City filter excludes amsterdam. The Date Year filter keeps 2017, 2018, 2019 and 2020.

A39

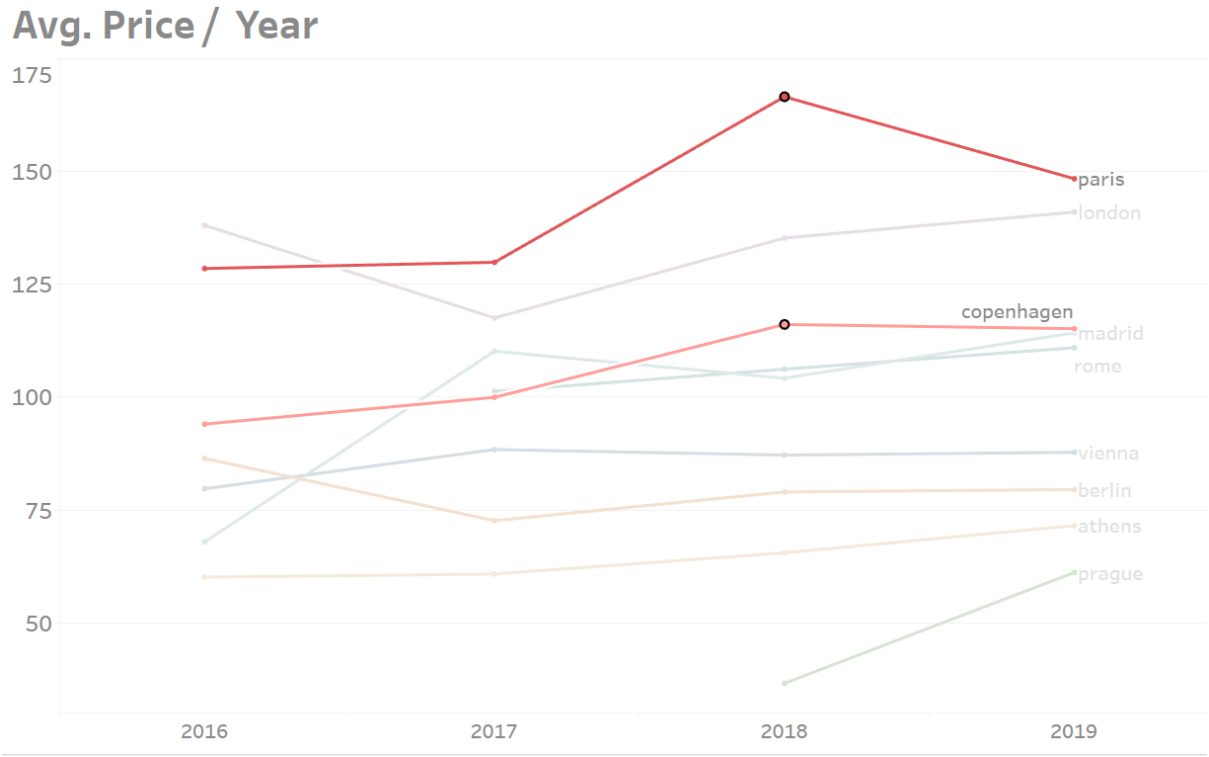


The trend of average of Average Price for Date Year. Color shows details about City. The marks are labeled by City. The view is filtered on City and Date Year. The City filter excludes amsterdam. The Date Year filter keeps 2016, 2017, 2018 and 2019.

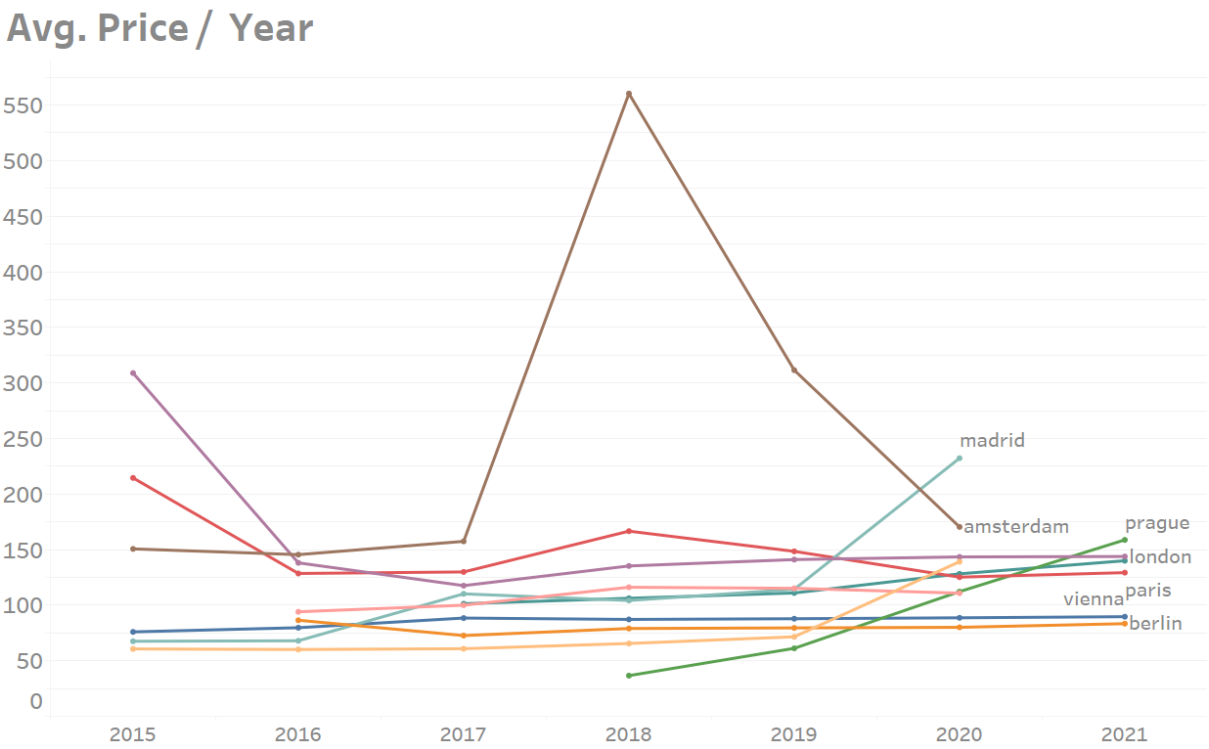
A40



A41

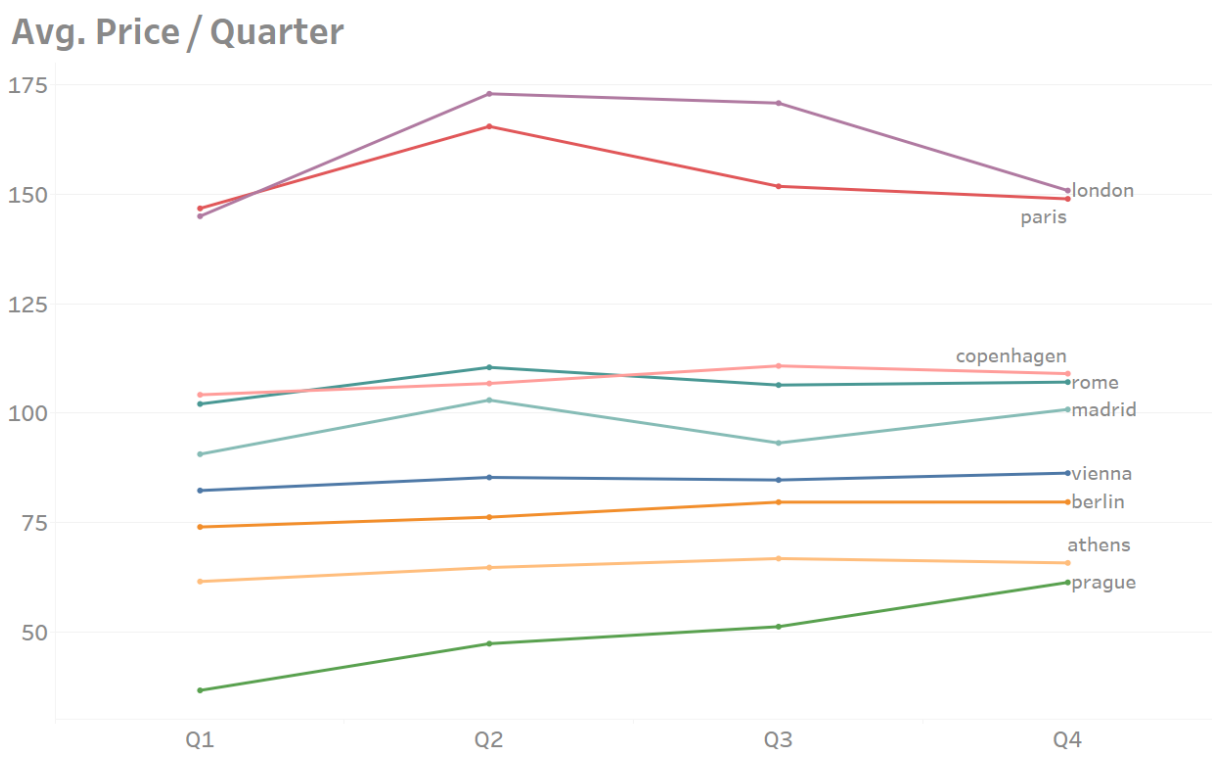


A42



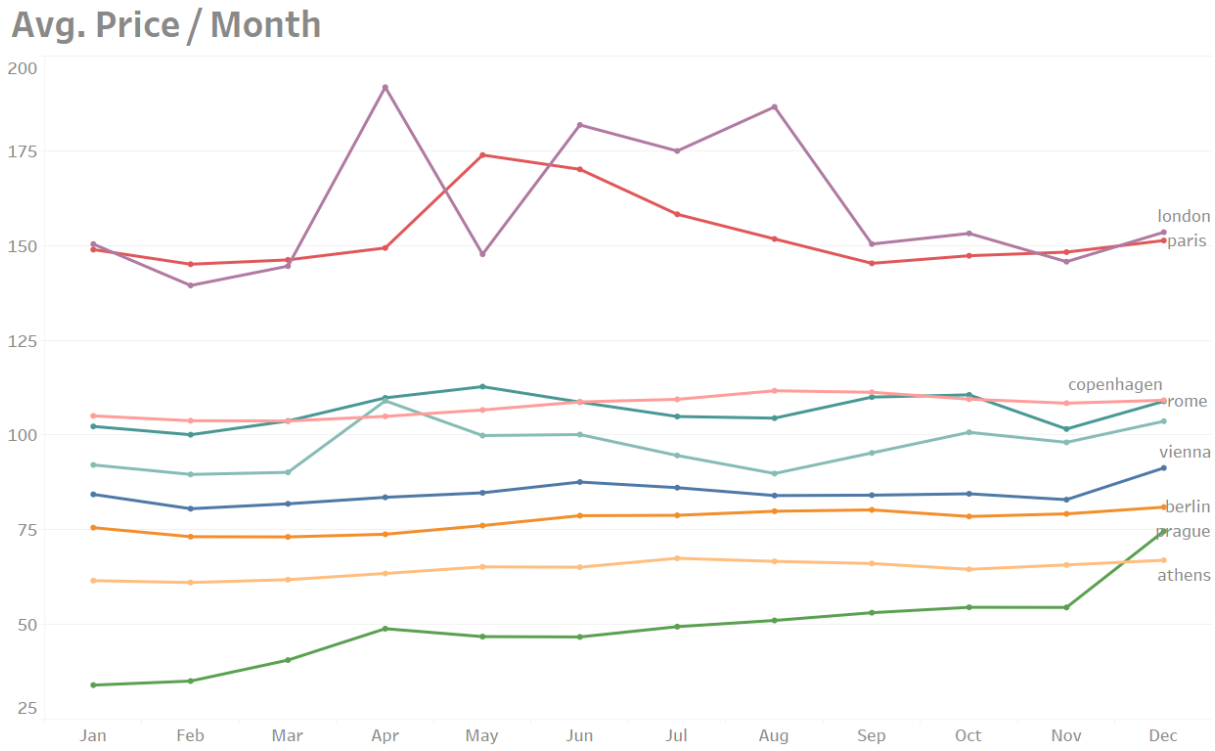
The trend of average of Average Price for Date Year. Color shows details about City. The marks are labeled by City. The view is filtered on City and Date Year. The City filter keeps 10 of 10 members. The Date Year filter keeps 7 of 7 members.

A43



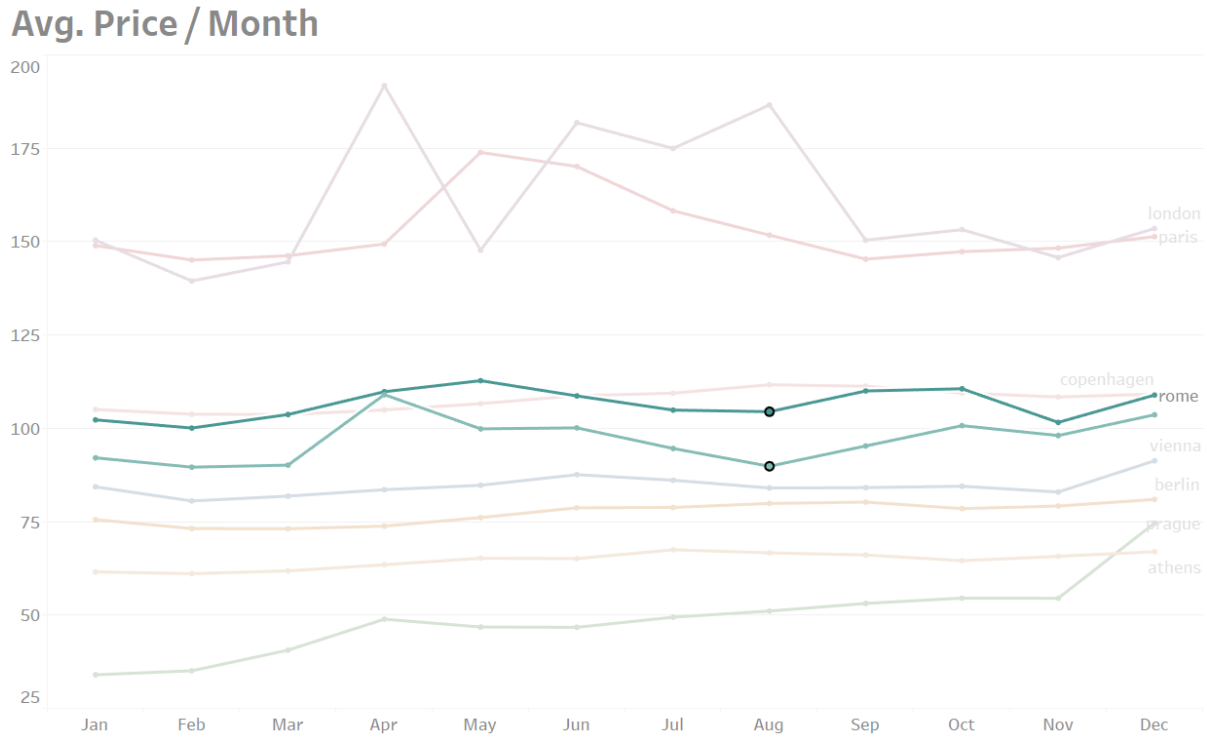
The trend of average of Average Price for Date Quarter. Color shows details about City. The marks are labeled by City. The data is filtered on Date Year, which keeps 2015, 2016, 2017, 2018 and 2019. The view is filtered on City, which excludes amsterdam.

A44

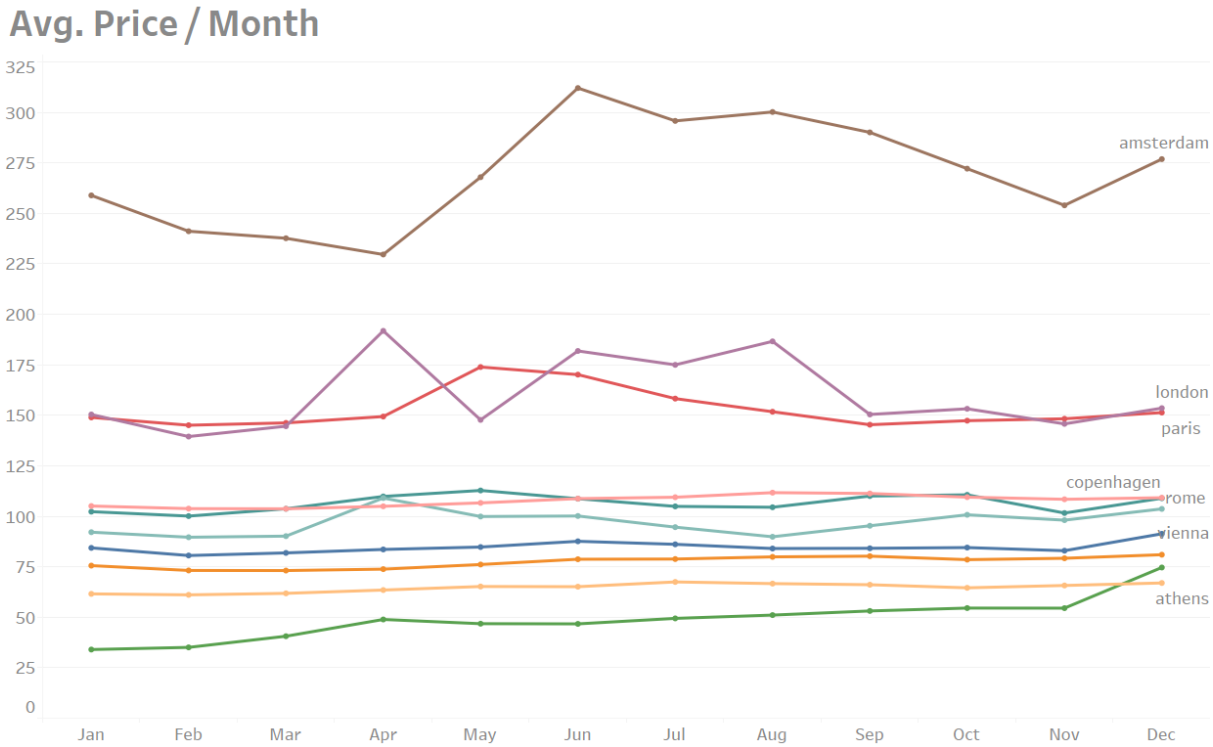


The trend of average of Average Price for Date Month. Color shows details about City. The marks are labeled by City. The data is filtered on Date Year, which keeps 2015, 2016, 2017, 2018 and 2019. The view is filtered on City, which excludes amsterdam.

A45

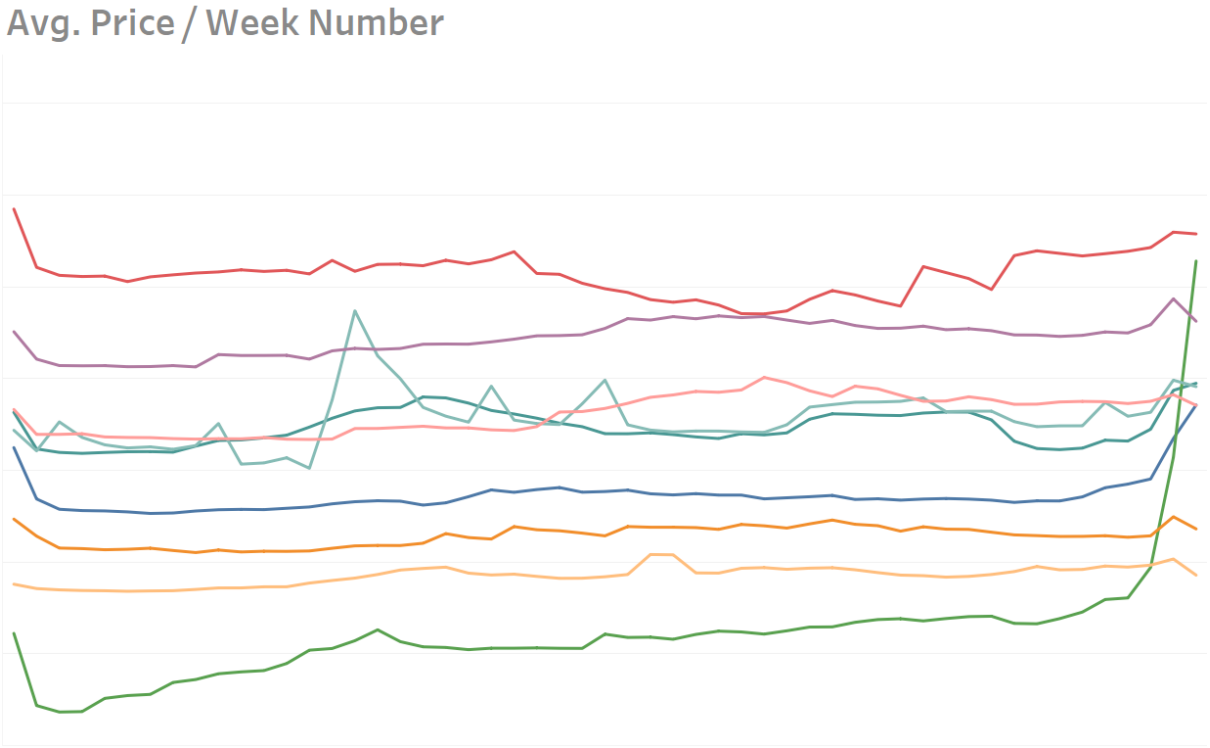


A46



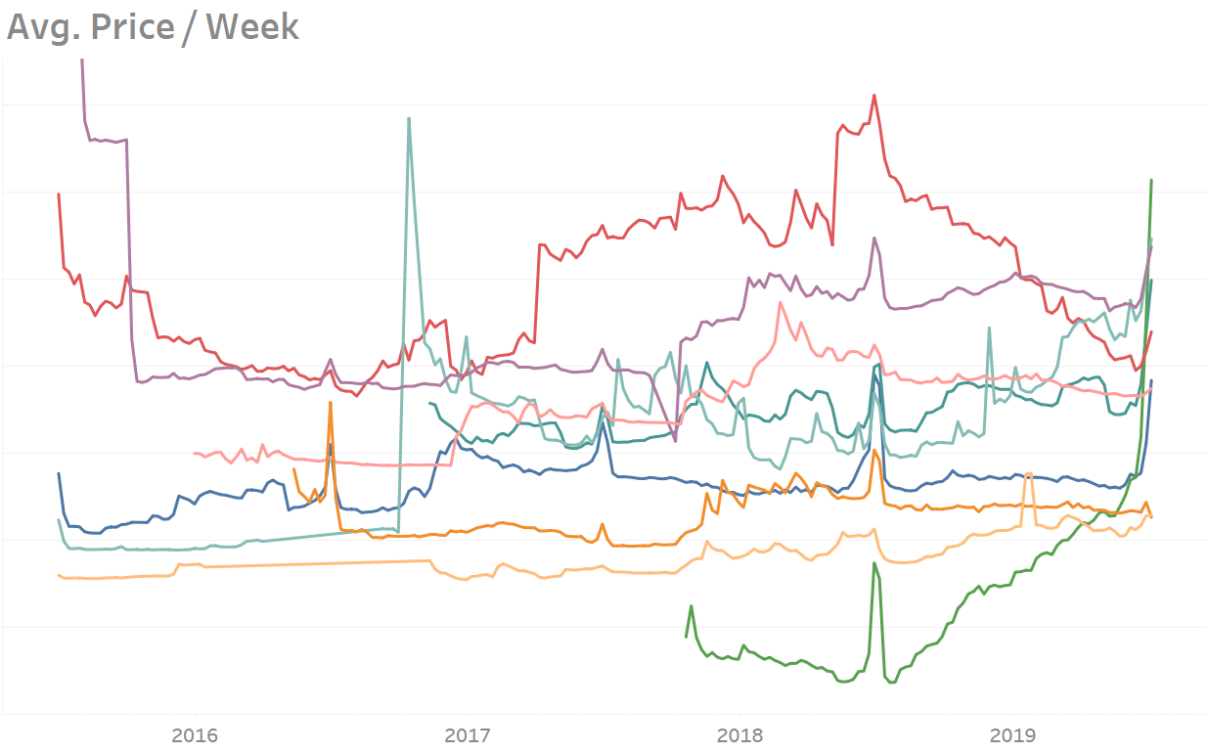
The trend of average of Average Price for Date Month. Color shows details about City. The marks are labeled by City. The data is filtered on Date Year, which keeps 2015, 2016, 2017, 2018 and 2019. The view is filtered on City, which keeps 10 of 10 members.

A47



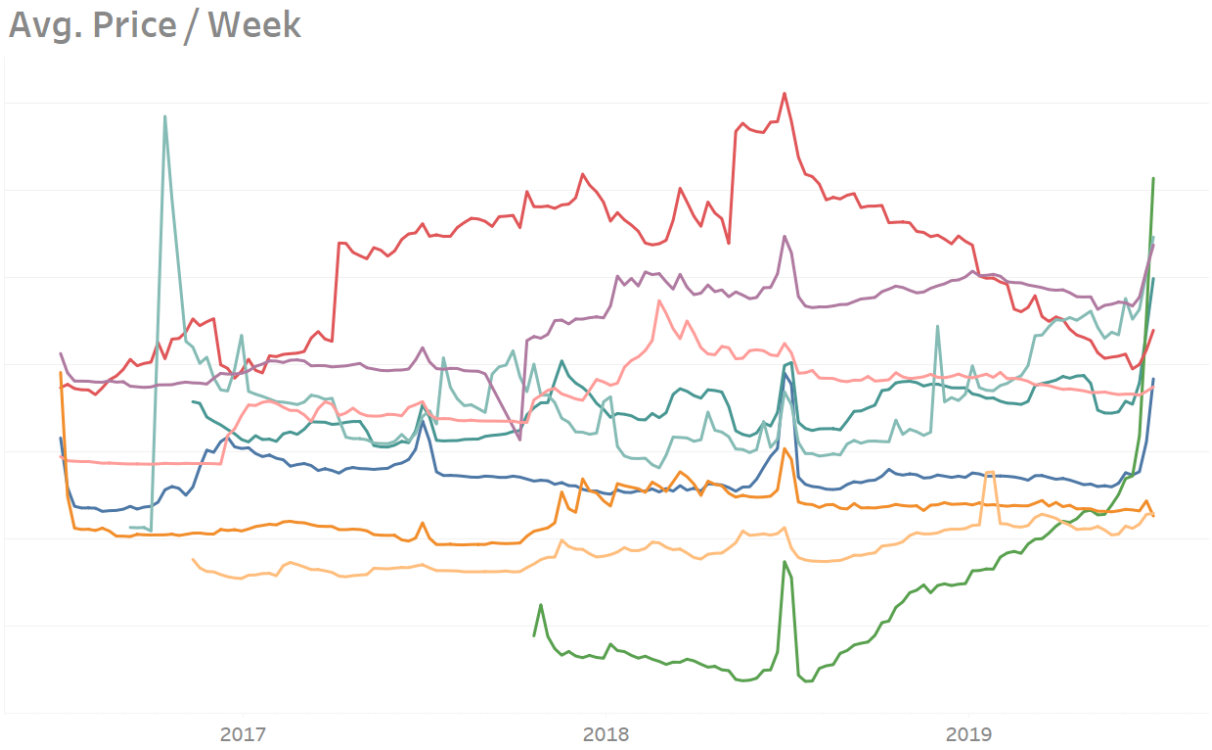
The trend of average of Average Price for Date Week. Color shows details about City. The data is filtered on Date Year, which keeps 2017, 2018 and 2019. The view is filtered on City, which excludes amsterdam.

A48



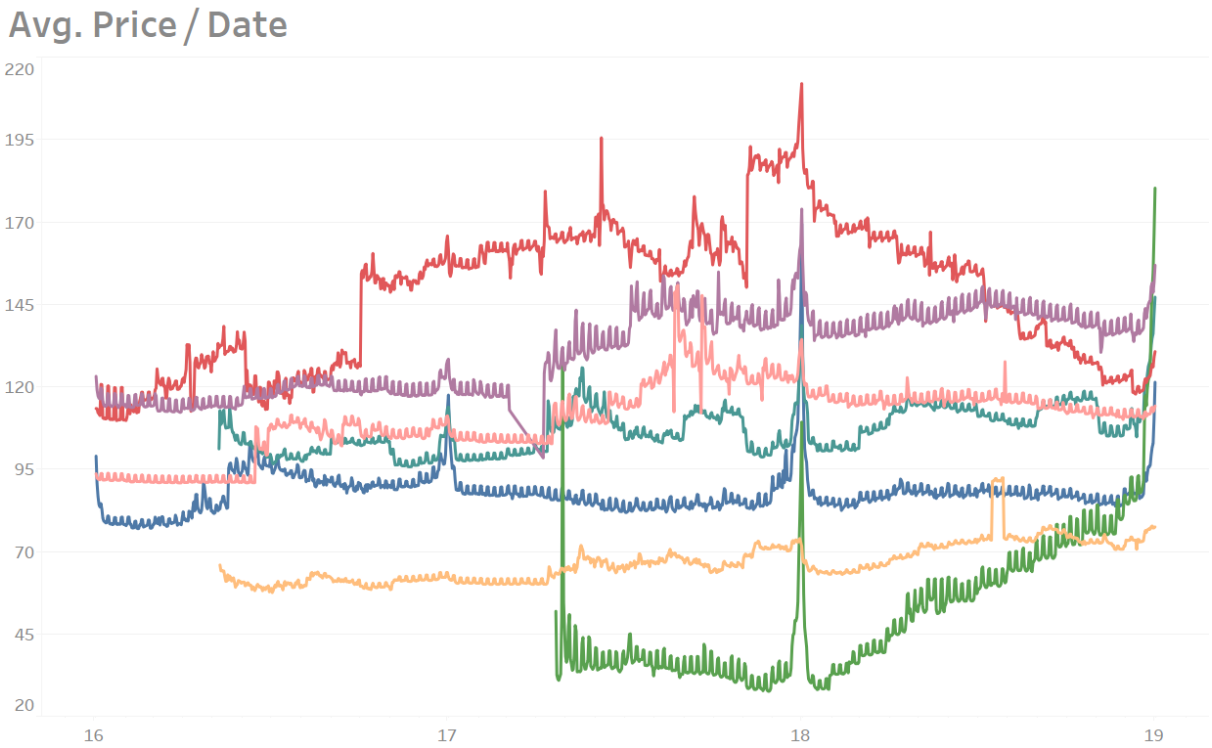
The trend of average of Average Price for Date Week. Color shows details about City. The data is filtered on Date Year, which keeps 2016, 2017, 2018 and 2019. The view is filtered on City, which excludes amsterdam.

A49



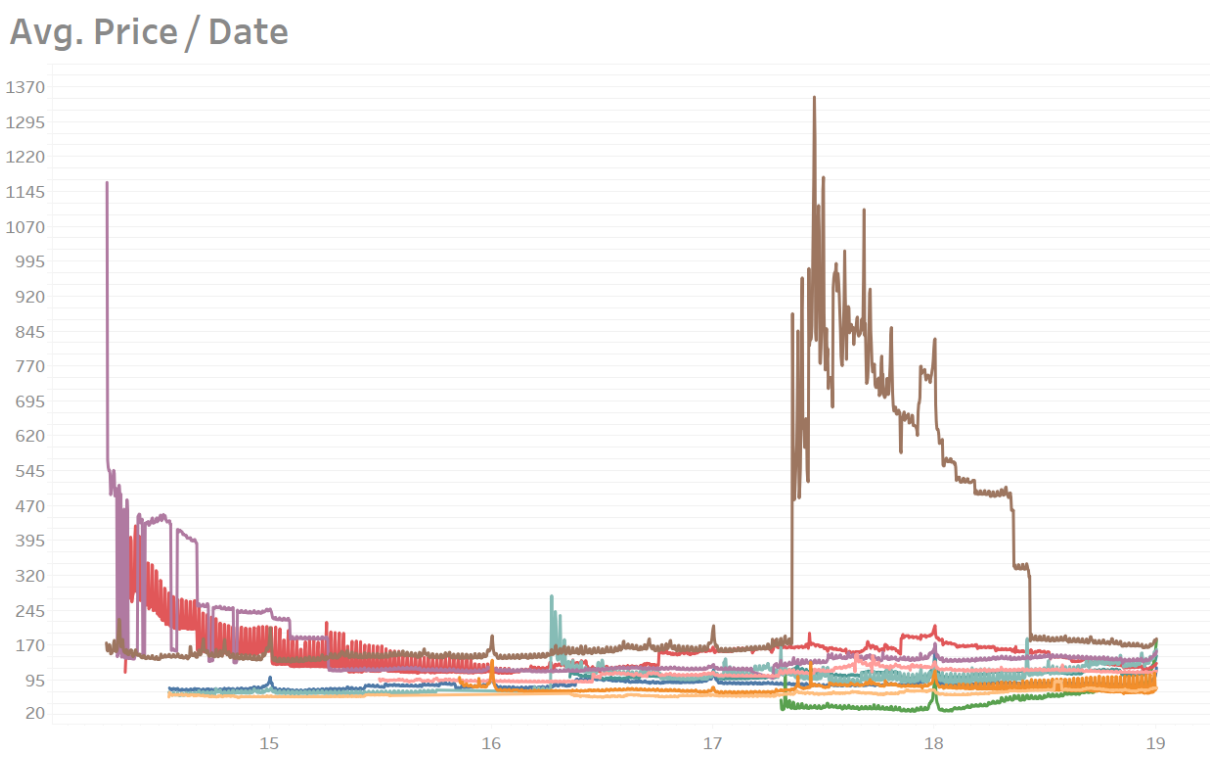
The trend of average of Average Price for Date Week. Color shows details about City. The data is filtered on Date Year, which keeps 2017, 2018 and 2019. The view is filtered on City, which excludes amsterdam.

A50



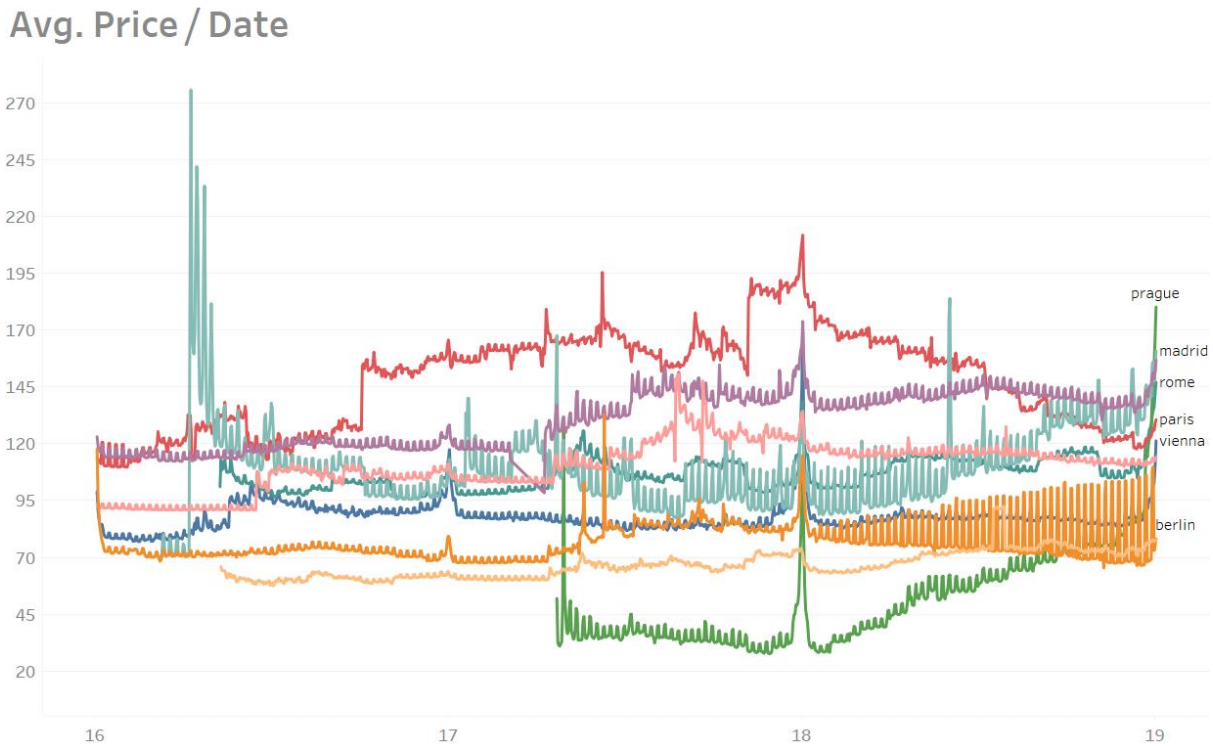
The trend of average of Average Price for Date Day. Color shows details about City. The data is filtered on Date Year, which keeps 2017, 2018 and 2019. The view is filtered on City, which excludes amsterdam, berlin and madrid.

A51



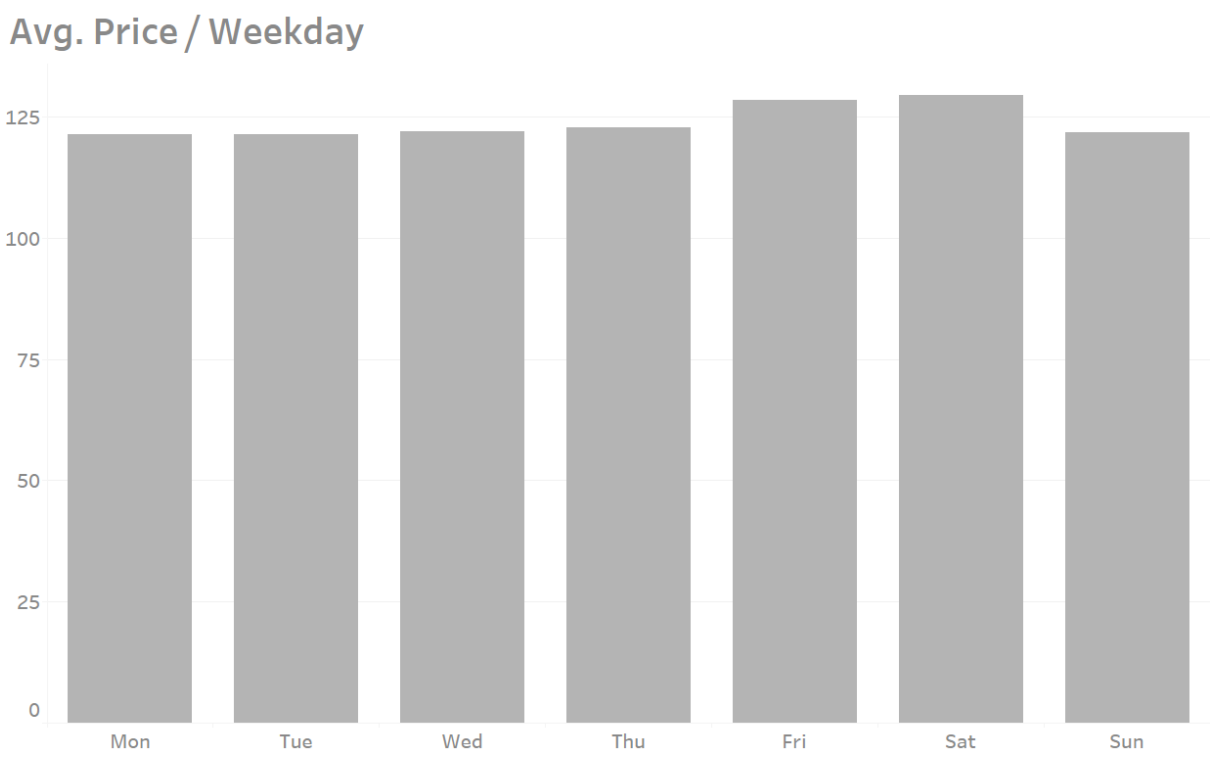
The trend of average of Average Price for Date Day. Color shows details about City. The data is filtered on Date Year, which keeps 2015, 2016, 2017, 2018 and 2019. The view is filtered on City, which keeps 10 of 10 members.

A52



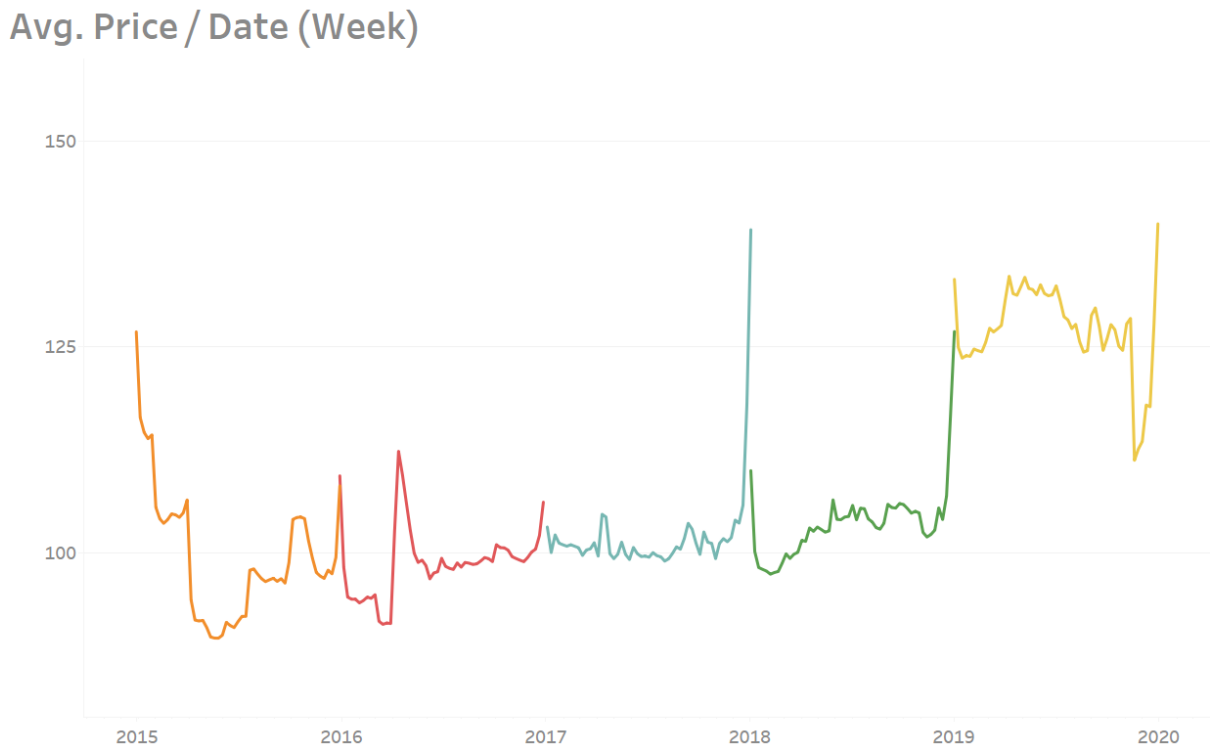
The trend of average of Average Price for Date Day. Color shows details about City. The marks are labeled by City. The data is filtered on Date Year, which keeps 2017, 2018 and 2019. The view is filtered on City, which excludes amsterdam.

A53



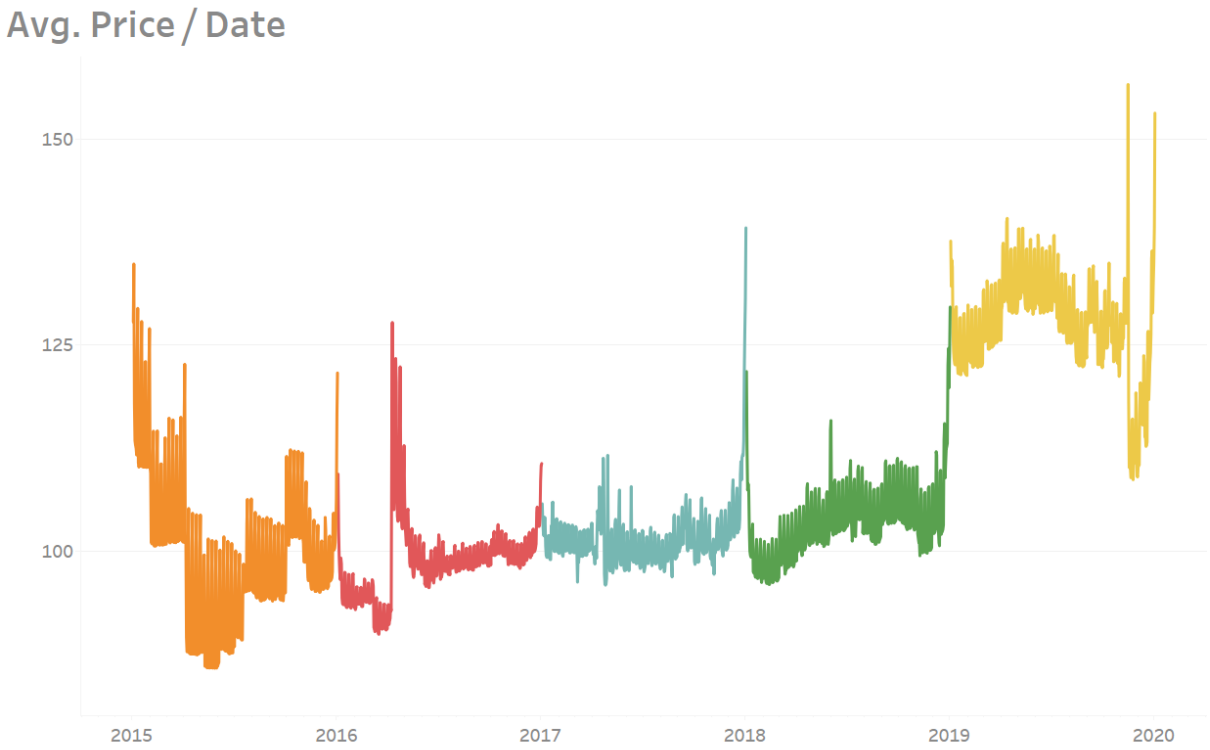
Average of Average Price for each Date Weekday. The data is filtered on City and Date Year. The City filter keeps 10 of 10 members. The Date Year filter keeps 2016, 2017, 2018 and 2019.

A54



The trend of average of Average Price for Date Week. Color shows details about Date Year. The data is filtered on City and Exclusions (DAY(Date),YEAR(Date)). The City filter excludes amsterdam. The Exclusions (DAY(Date),YEAR(Date)) filter keeps 2,035 members. The view is filtered on Date Year, which keeps 2016, 2017, 2018, 2019 and 2020.

A55



The trend of average of Average Price for Date Day. Color shows details about Date Year. The data is filtered on City, which excludes amsterdam. The view is filtered on Date Year and Exclusions (DAY(Date),YEAR(Date)). The Date Year filter keeps 2016, 2017, 2018, 2019 and 2020. The Exclusions (DAY(Date),YEAR(Date)) filter keeps 2,035 members.