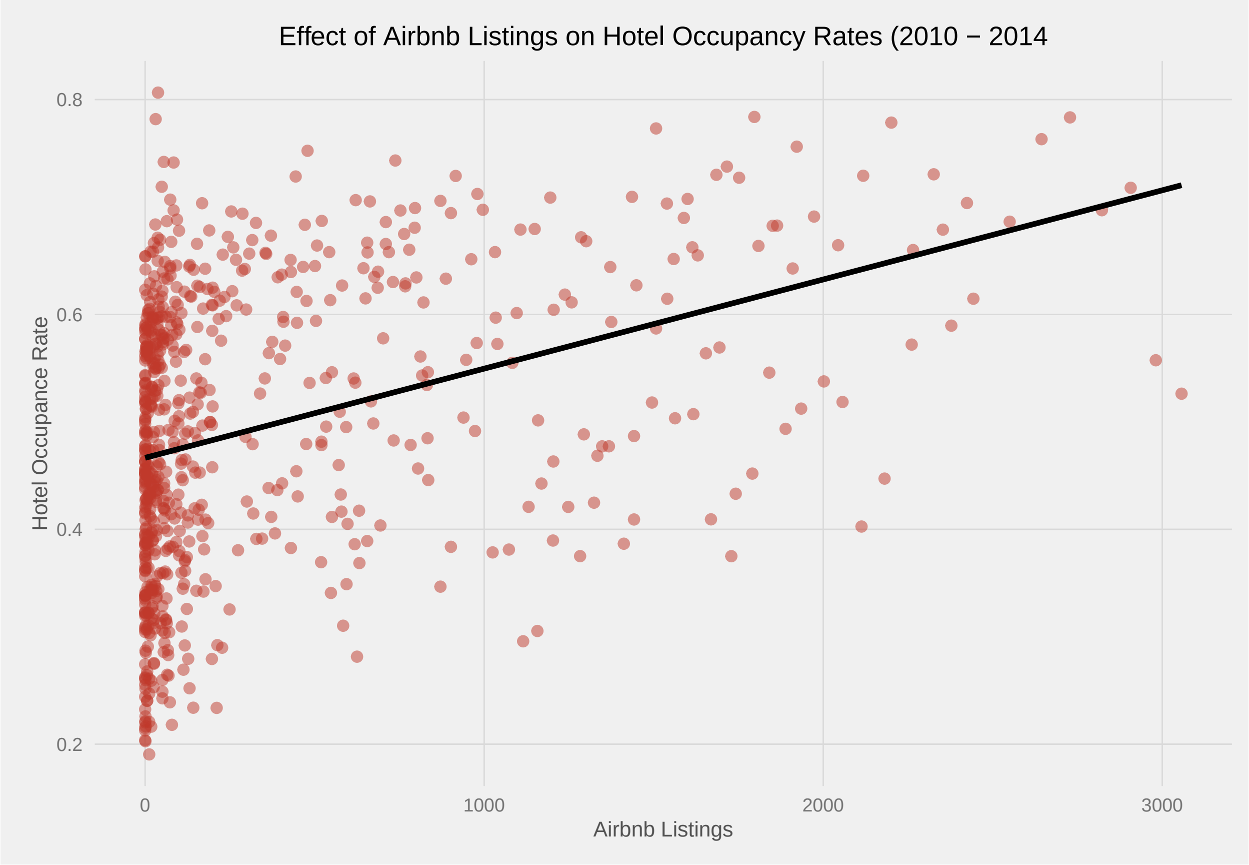
Pair Assignment No. 3 - Data Snapshot

Introduction to Collaborative Social Science Data Analysis

Daniel J Murphy & Paulo H Kalkhake

November 9, 2016



# Introduction While the effects of Airbnb on Berlin apartment prices has generated much discussion, significantly less attention has been paid to Airbnb's effect on the hotel industry. Airbnb claims that they are largely complementary to hotels, since most Airbnb listings are found outside of the main hotel district in a given city. In Berlin for example, 77% of Airbnb listings are outside of the major hotel districts.

However, a paper by Zervas, Proserpio, and Byers (2016) found that the rise of Airbnb had a negative effect on hotel revenue in the state of Texas. This shows that, while Airbnb may be complementary, they also compete with the incumbent hotel industry. In this paper we will seek to illustrate the magnitude of the "Airbnb effect" on hotels in Berlin. To that end, two hypotheses will guide our thinking.

H1: The higher the Airbnb supply in a given district in Berlin, the lower the hotel occupancy rate will be in that same district.

H2: Since Airbnb's listings are concentrated in districts with low hotel density, the effect of substitution will be more pronounced in those districts.

Data & Variables

Our data comes from three different sources, the Statistical Information System Berlin/Brandenburg (SBB) (StatIS-BBB 2016), the Federal Statistical Office and the statistical offices of the Länder (FSO)[1] (Germany 2015), and *InsideAirbnb.com* (Cox 2016). In order to conduct our analysis we neede to clean, merge and manipulate these data sets.

From the Statistical Information System Berlin/Brandenburg, we collected monthly data on the number of overnight stays and the number of guests to arrive at their accommodations in the reporting period in Berlin (StatIS-BBB 2016). The surveys are carried out at the beginning of each month and refer to the reporting period of the previous month. The results are organized regionally according to districts and municipalities, allowing us to have specific data for each of the twelve districts in Berlin. We also gathered data for yearly household income groups and the number of employed and unemployed people per district. Based on the data, we calculated a yearly average household income and unemployment rate per district.

From the Regional Database of Germany, we collected data on the supply of hotel beds in each district in a given year. Unfortunately this data is only recorded annually. Hence, our analysis must assume that the number of beds in each hotel stay constant throughout the year. For our main dependent variable, we used data on the number of beds generally available in each district per day in a given year and the monthly data on the number of overnight stays from guests to compute a variable as a proxy for hotel occupancy (as a percentage).

From *InsideAirbnb.com*, we scraped data on 15,368 listings, i.e. apartments or rooms, for Berlin from August, 2008 until October, 2015. Amongst 92 variables for each listing covering topics ranging from room price to information on the host, the data includes (1) the neighbourhood of each listing, (2) the date that an Airbnb host signed up, and (3) the date of the last review of each listing.

Our most significant methodological challenge is the absence of precise listing availability during our period of interest, as it is not directily available in the data. However, in keeping with the Zervas, Proserpio, and Byers (2016) methodology, we used data on when the host became Airbnb member as a proxy for market entry. We then construct a variable for cumulative supply based on this information, i.e. the total number of listings in a district belonging to users that have signed up on Airbnb in prior to that month. This is

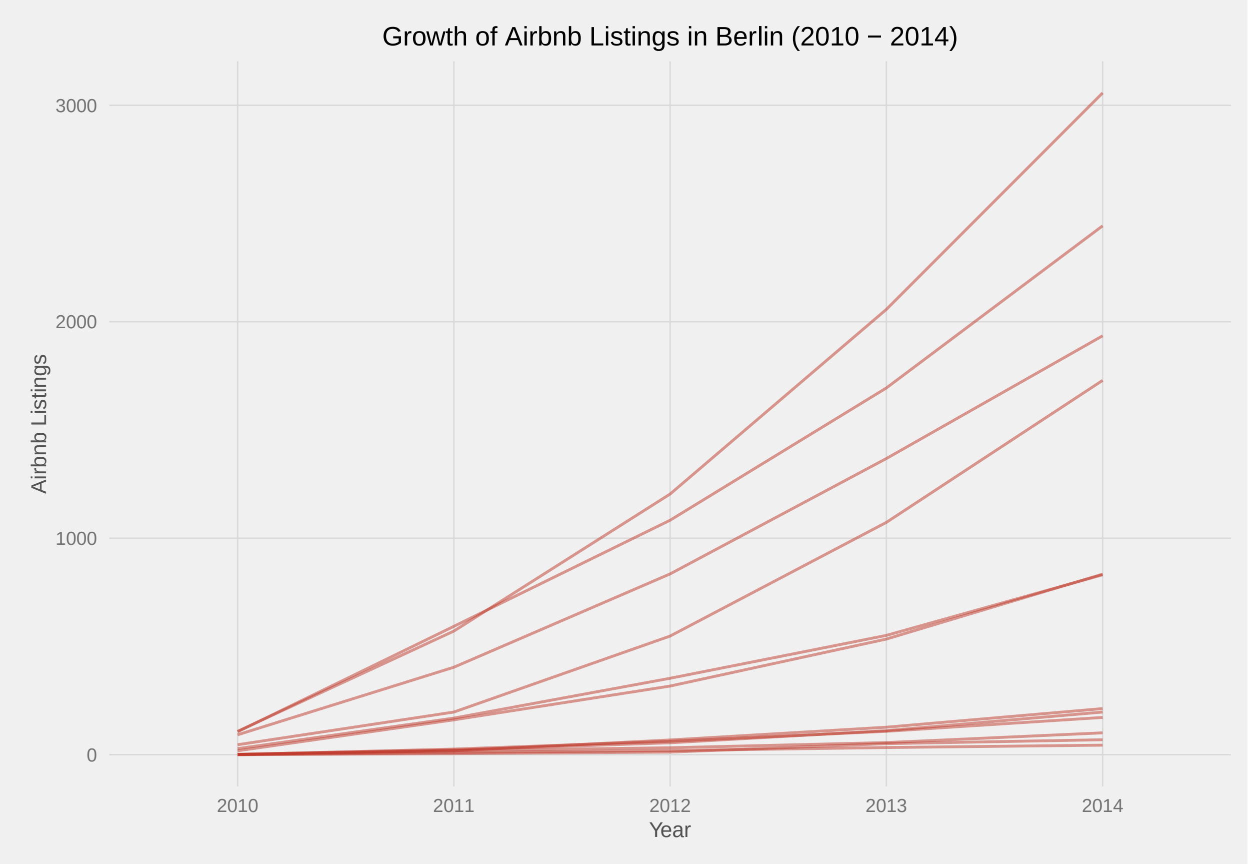
[1] Both databases use JAVA-based website, which did not allow direct web scraping. The data was manually downloaded.

not ideal, as this proxy does not take into account whether or not a listing was available in a given month. However, Airbnb itself is also unable to produce exact supply data. This is because owners do not accurately update their listings' availability.

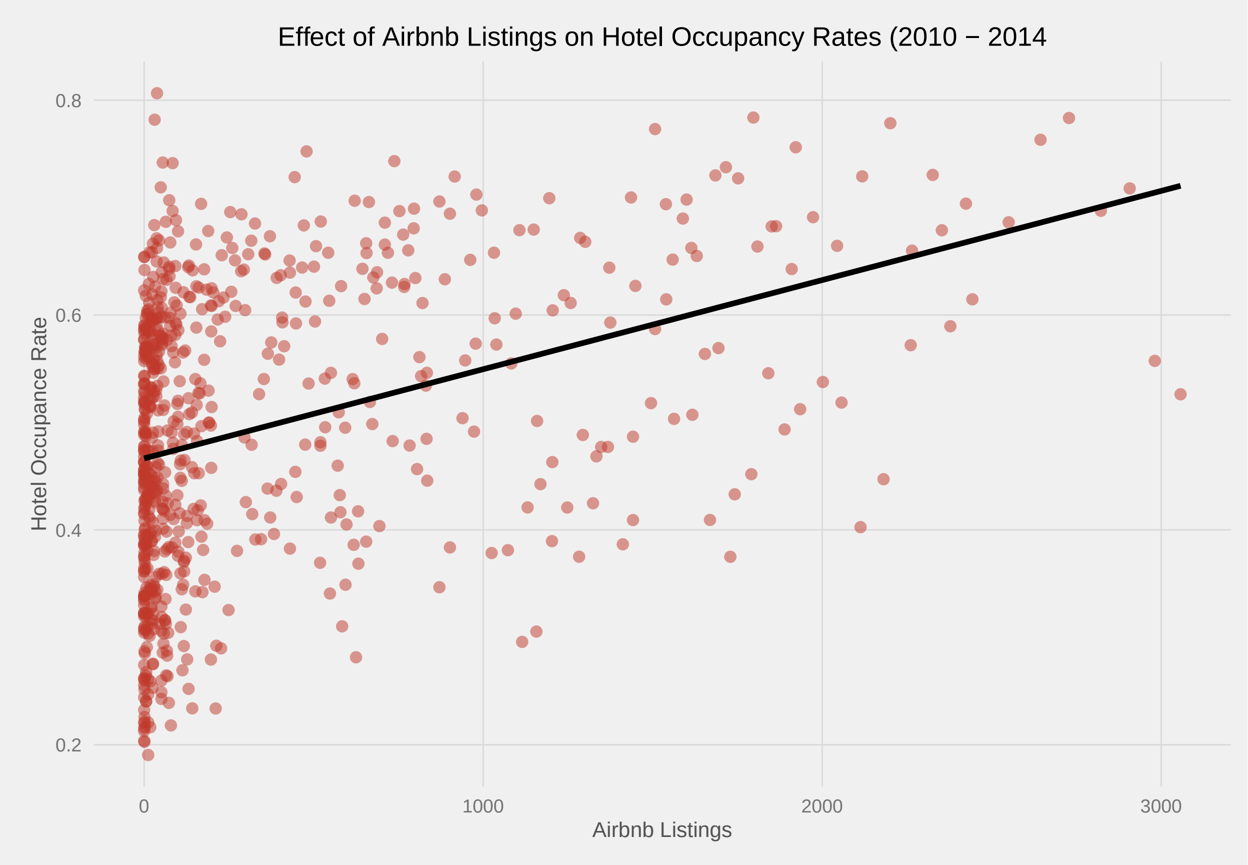
The final data set covers 720 monthly obversations across tewlve districts in Berlin and covers the time period between 2010 and 2014.

Descriptive Analysis

Airbnb‘s popularity amongst users and hosts has increased substantially since it was founded in August 2008. As Berlin's popularity has increased over that same period, many Berliners began listing apartments on Airbnb. The growth in Airbnb listings has not been equal across all neighbourhoods in Berlin, but it has been positive in every year from 2010 to 2014. That trend has accelerated each year, with more and more Berliners listing apartments on the site.



Upon plotting Airbnb supply against hotel room occupancy, we were surprised to see a clear positive correlation. However, we realized the importance of accounting for general demand in an area as it becomes more attractive to tourists. Zervas, Proserpio, and Byers (2016) accomplished this by controlling for passengers listing the local airport as their final destination. We will incorporate a similar statistic in our final analysis, as tourism in Berlin has increased dramatically in recent years. Given confirmation from Zervas et al. that Airbnb does indeed compete with hotels, we believe that this increase in Berlin's popularity is largely responsible for the positive correlation we observe here.



Inferential Analysis

Thus for further analysis of the data, we included the number of guests per month for each district.

```
##
## Call:
## lm(formula = occup_rate ~ AB_supply + guests, data = analysis_simple)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.25737 -0.09085 -0.00090  0.09797  0.34762
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.457e-01  5.557e-03  80.216  < 2e-16 ***
## AB_supply   4.568e-05  8.591e-06   5.317 1.41e-07 ***
## guests      4.586e-07  4.955e-08   9.256  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1167 on 717 degrees of freedom
## Multiple R-squared:  0.2226, Adjusted R-squared:  0.2205
## F-statistic: 102.7 on 2 and 717 DF,  p-value: < 2.2e-16
```

Even when including guests in our model we find that Airbnb supply has a highly significant positive effect on hotel occupancy, with a p-value on the .01% level. This surprised us as well, and warrants further exploration of potential explanations in our final analysis.

Bibliography

Cox, Murray. 2016. “Inside Airbnb.” http://insideairbnb.com/get-the-data.html.

Germany, Regional Database. 2015. https://www.regionalstatistik.de.

StatIS-BBB. 2016. https://www.statistik-berlin-brandenburg.de/webapi/jsf/dataCatalogueExplorer.xhtml.

Zervas, Georgios, Davide Proserpio, and John Byers. 2016. “The Rise of the Sharing Economy: Estimating the Impact of Airbnb on the Hotel Industry.” *Boston U. School of Management Research Paper*, no. 2013-16.