Racial Discrimination in the Online Consumer Marketplace

A Study on Airbnb

Hanying Mo

May 6, 2016
Stat 157
Professor Aldous

I. Introduction

Today's online marketplaces are increasingly shifting towards systems that reduce the anonymity of buyers and sellers in an effort to build trust between the two parties, most commonly done through online profiles. While traditional marketplaces such as Amazon and eBay continue to use customer accounts and only ask for a valid form of payment, social media websites such as Facebook emphasize the "real name" policy. This consequently allows other websites to use Facebook as a form of verification for a person's identity, which includes many personal details. While this helps facilitate trust in the marketplace, the shift towards online profiles creates an unintended problem: racial discrimination. By asking users for their real names, profile photos, and descriptions, Airbnb effectively builds a marketplace that allows for hosts and guests to choose listings or guests, respectively, based on a variety of characteristics.

While Airbnb is only one platform that facilitates accommodations bookings that depends on user-generated listings, with other platforms including Couchsurfing, Vbro, and Homeaway, it is an increasingly powerful one. The still-private company was founded in 2008 and as of November 2015, it offers 2,000,000 listings worldwide. This is more than three times as many as Marriott's 535,000 rooms worldwide. With a current valuation of approximately $25 billion, Airbnb is a dominant player and now, because of its size, is more representative of the hospitality industry and other cultural or socio-economic forces that influence the overall industry. A study on Airbnb differs from a study on Marriott because for traditional accommodations companies, a profile is not necessary, simply a form of payment. Therefore, because a human host has to approve each booking request, it could also better represent people's perceptions of an acceptable booking. Especially as the entire online consumer industry is shifting towards reduced anonymity, the impact of racial discrimination on platforms could have a lasting effect on both Airbnb and other platforms.

Existing research establishes that there is racial discrimination on Airbnb. My project will focus on establishing whether there is a correlation between the host's race the listing price, and I will attempt to model how much these factors impact how much a hosts can charge. I will do so by looking at data of Asian American and White hosts in the Oakland/Berkeley area.

II. Description of Airbnb

Airbnb is an online platform that allows hosts to rent out houses, apartments, and shared/private rooms within an apartment. Airbnb facilitates communication and payment between a host and a guest. While the service is free for a guest, Airbnb charges hosts a 3% listing fee upon a successful reservation. Its advantage over existing platforms such as Homeaway is that it only charges for a booking, not just to list. Based on a system of online profiles and reviews, Airbnb has created a large community of people, both hosts and guests, and reports serving over 40 million guests in more than 190 countries. It is also one of the top ten highest-valued private companies in the United States currently.

III. Basis of Study and Source of Data

The basis of my project is from two previously-conducted studies. Benjamin Edelman and Michael Luca, both from Harvard Business School, published the paper "Digital Discrimination: The Case of Airbnb.com" in 2014, which found that non-black hosts charge approximately 12% more than black hosts for the equivalent rental, even when controlling for a variety of variables relating to the perceived popularity of a property, including price, location, and size. The difference in rates was statistically significant, with p-value < 0.01. However, while it could be argued that hosts themselves set the price and therefore a reduced price is not necessarily indicative of racial discrimination, Airbnb is still an online marketplace based on supply and demand, and African-American hosts could feel like they need to charge a lower price in order to entice people to book their listing.

Nevertheless, another study conducted in 2016 by Benjamin Edelman, Michael Luca, and Dan Svirsky, professors from Harvard Business School and the Harvard Economics Department found that requests from guests with distinctively African-American names are roughly 16% less likely to be accepted than identical guests with distinctively White names. The difference persists whether the host is African-American or White, male or female. The difference also persists whether the host shares the property with the guest or not, and whether the property is cheap or expensive.

The combination of the two studies show that there is racial discrimination, both from the hosts and the guests' sides. However, published data that includes information on race is very difficult to come by, because Airbnb itself does not ask for a host's race. The aforementioned studies were able to collect data by using Amazon Turk Workers to identify then validate a host's race from profile photos, but unfortunately did not publish their data. One website (http://insideairbnb.com/get-the-data.html) provides data on various listings in many cities but does not provide data on race. However, I was able to find a dataset of Berkeley/Oakland Airbnb and Homeaway listings from September 2015 from another university study. This data set includes "Number of Bathrooms," "Number of Bedrooms," "Number of Occupants," "Cost per Week," "Price per Room," "Race," which are Asian-American and White. I am using this data rather than scrubbing InsideAirbnb.com's data for race, so race is accurately represented since this dataset had three researchers who were able to cross-validate each others' categorization of race. For this project, I will stay consistent with the terminology of the existing reports.

I decided to use this data set because as Berkeley is a demographically-diverse area, and I was curious as to see whether the findings in the aforementioned Harvard studies could be replicated between two other racial groups, White and Asian-American. I also wanted to focus on how much each racial group could charge. Furthermore, I also compiled all Oakland area Airbnb listings from InsideAirbnb, which I also plotted as a form of sanity check and to compare the smaller dataset to its entire population. As Berkeley students, the Airbnb listing data is directly relevant to us because it represents the housing market, and findings could be indicative of additional discrimination in the area.

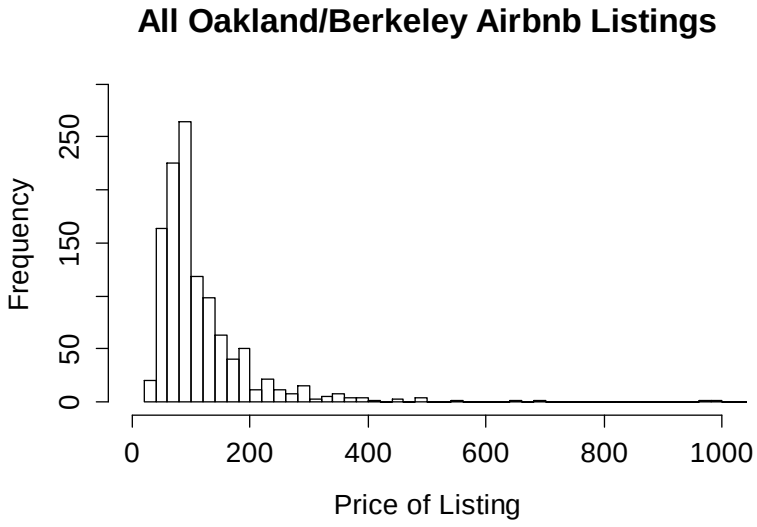
I went through the original data set and selected only Airbnb listings. I further expanded on this data set by looking up the data for a host's response time, response rate, and number of reviews to see if that would correlate to the host's listing price as well, because the results from the Harvard study indicated that discrimination held even when controlled for these factors. I then checked that all the listings were still existing. The original dataset was compiled by a simple random sample from the Airbnb and Homeaway website, so the random sampling is still relevant since I am looking only at Airbnb and added more attributes instead of listings.

My final data set has the columns "Number of Bathrooms," "Number of Bedrooms," "Number of Occupants," "Cost per Week," "Price per Room," "Race," "Response Rate," "Response Time," and "Number of Reviews of Property." I also looked at Price per Night because some homeowners offer a weekly discount, and I wanted to take this into account.

My project will focus on establishing whether there is a correlation between the host's race and the listing price, and I will attempt to model how much these factors impact how much a hosts can charge.

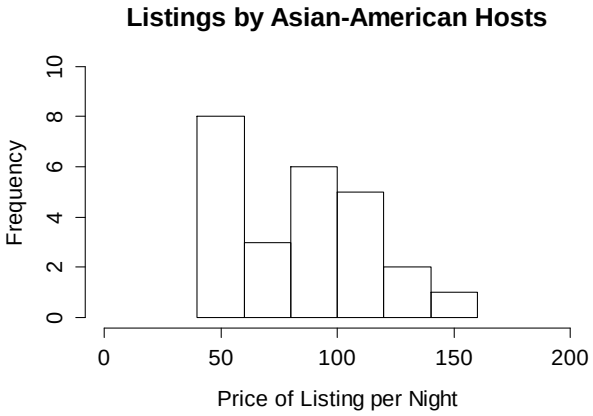
IV. Exploration of Data

I mainly used R to conduct my analysis. I began by plotting the distribution of all Airbnb listings in the Oakland/Berkeley area, to get a sense of the collective Airbnb offerings before delving into our analysis on racial discrimination in the area.

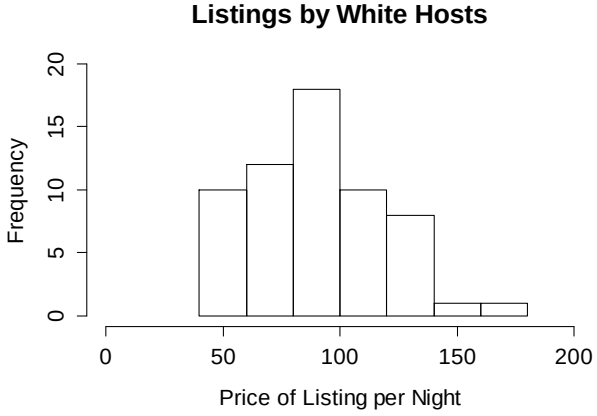


The distribution is heavily skewed right, with most of the values between $0 and $300. This is per-night data.

The data I collected has 85 Airbnb listings from the Berkeley area, with 25 from Asian-American hosts and 60 from White hosts. Again, I plot the two distributions of listing price per night in the Airbnb, and calculate summary statistics.




```
> summary(asian$perRoom)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   40.0    60.0    83.0    85.2   107.0   145.0
```




```
> summary(white$perRoom)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  41.00   70.25   87.50   91.70  111.20  180.00
```

From looking at the distributions, we see that Asian Americans, on average, have a lower mean and median compared to White hosts. Interestingly enough, the distribution of listings for Asian-American hosts has its highest peak centered around 50, with a dip soon after and another peak centered around 90. This is very interesting because you would expect it to more closely follow the distribution of the listings by White hosts or the population, but instead it shows that Asian American hosts tend to offer both lower prices and prices around the mean. The distribution of listings by White hosts is slightly skewed right, which is consistent with the population data.

The median of Asian-American hosts is 83.0, the median of White hosts is 87.5, and the means are 85.2 and 91.7, respectively. However, is the difference in this price statistically significant? I run a few t-tests as well.

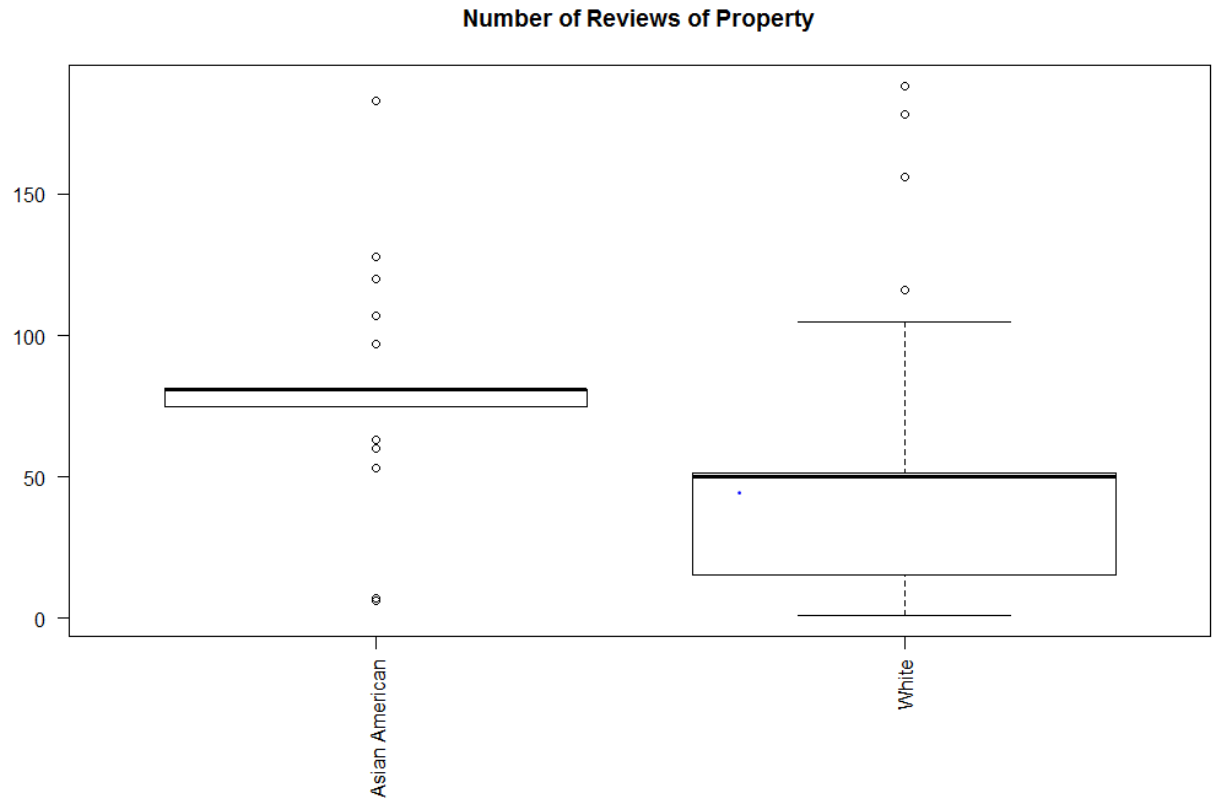
```
> t.test(asian$price, white$price)

        Welch Two Sample t-test

data:  asian$price and white$price
t = -2.628, df = 82.554, p-value = 0.01024
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -509.8378  -70.5489
sample estimates:
mean of x mean of y
 700.6400  990.8333
```

Looking at the t-test, we see a p-value of 0.01024, which is statistically significant at a significance-level of 0.05. This seems to show that there is a significant difference between the prices that Asian American hosts can charge compared to whites. Especially given a data set of 85 listings, this is a sizeable difference.

Consequently, after figuring there is a difference between prices offered by Asian-American and White hosts, I wanted to see what other factors could possibly be impacting the difference in price. Therefore, I run some additional summary statistics.

```
> summary(asian[ , c(1:4, 7:9) ])
   bathrooms       bedrooms        people          price            rRate           rTime          reviews      
 Min.   :1.0   Min.   :0.00   Min.   :1.00   Min.   : 330.0   Min.   :0.8000   Min.   : 1.00   Min.   :  6.00  
 1st Qu.:1.0   1st Qu.:1.00   1st Qu.:2.00   1st Qu.: 553.0   1st Qu.:0.9800   1st Qu.: 1.00   1st Qu.: 75.00  
 Median :1.0   Median :1.00   Median :2.00   Median : 616.0   Median :0.9800   Median : 4.00   Median : 81.00  
 Mean   :1.2   Mean   :1.28   Mean   :2.76   Mean   : 700.6   Mean   :0.9824   Mean   : 3.76   Mean   : 81.24  
 3rd Qu.:1.0   3rd Qu.:2.00   3rd Qu.:3.00   3rd Qu.: 770.0   3rd Qu.:1.0000   3rd Qu.: 4.00   3rd Qu.: 81.00  
 Max.   :3.0   Max.   :3.00   Max.   :8.00   Max.   :1700.0   Max.   :1.0000   Max.   :24.00   Max.   :183.00  

> summary(white[ , c(1:4, 7:9) ])
   bathrooms       bedrooms        people          price            rRate           rTime          reviews      
 Min.   :1.000   Min.   :0.0   Min.   : 1.0   Min.   : 315.0   Min.   :0.8000   Min.   : 1.000   Min.   :  1.00  
 1st Qu.:1.000   1st Qu.:1.0   1st Qu.: 2.0   1st Qu.: 552.2   1st Qu.:0.9500   1st Qu.: 3.000   1st Qu.: 17.25  
 Median :1.000   Median :1.0   Median : 2.5   Median : 760.0   Median :0.9500   Median : 3.000   Median : 50.00  
 Mean   :1.383   Mean   :1.6   Mean   : 3.4   Mean   : 990.8   Mean   :0.9533   Mean   : 6.917   Mean   : 49.77  
 3rd Qu.:2.000   3rd Qu.:2.0   3rd Qu.: 4.0   3rd Qu.:1152.5   3rd Qu.:1.0000   3rd Qu.: 7.000   3rd Qu.: 50.75  
 Max.   :4.000   Max.   :5.0   Max.   :10.0   Max.   :3500.0   Max.   :1.0000   Max.   :24.000   Max.   :188.00  
```

The number of bathrooms, number of bedrooms, people (occupancy), response rate, and response time don't seem to vary that much, so I didn't feel the need to visualize the data. However, they are all important in the pricing model, so I will take that into account later.



While it is statistically significant that Asian-American hosts charge a lower rate on Airbnb compared to White hosts, this could be due to a variety of factors. As I mentioned before, because

Airbnb is an online marketplace, it's governed by the rules of supply and demand. However, as the Harvard researchers showed, there is significant racial discrimination both on the hosts' and guests' sides. Consequently, there may be underlying social or cultural dynamics that would influence Asian American hosts to either post a lower listing price or see that guests would only book once they offer a lower listing price.

Looking at the number of reviews, which typically correlates to the number of times a host has actually rented out a listing and received a review, we see that Asian American hosts have a higher number of reviews compared to White hosts. Therefore, we see that Asian American hosts could be offering a lower rate in order to rent out their property more, which correlates to why they have more reviews on their property. While factors such as number of bathrooms, number of bedrooms, and people (occupancy) show the desirability of a property based on its size, the number of reviews most closely follows how many guests actually booked and stayed at the listing.

Therefore, I also want to look at the potential correlation between prices and the number of reviews.

```
> cor.test(air$price, air$reviews)

        Pearson's product-moment correlation

data:  air$price and air$reviews
t = -3.3461, df = 83, p-value = 0.001233
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.5197030 -0.1420813
sample estimates:
       cor
-0.3447656
```

Based on a p-value of 0.001233, there seems to be a statistically significant correlation between the price of the listing and the number of reviews. The sample correlation value is very interesting – it seems like there is a slight negative correlation between prices and number of reviews with r = -0.345. This is consistent with the visualization of reviews, where the higher the number of reviews, the lower the price is offered.

Now that I know that there is a statistically significant difference between the prices that Asian American and White hosts charge, it is interesting to see if there is an optimum pricing model that would 1) help Asian American hosts charge more per night or 2) help White hosts rent out properties at a higher rate. For the scope of this research project, I will focus on part 1 by building a pricing model that takes into account a host's race, and figure out how much revenue Asian American hosts are missing out on by not charging higher prices

V. Linear Regression

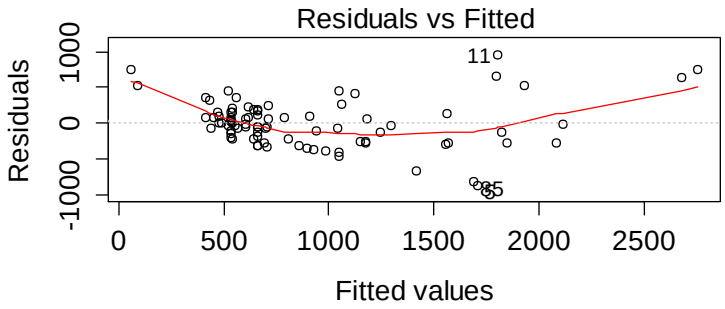
I ran a linear model on the data, accounting for every variable except the number of bathrooms, which didn't vary enough and did not significantly impact the number of people who could stay in the property. The following is the output:

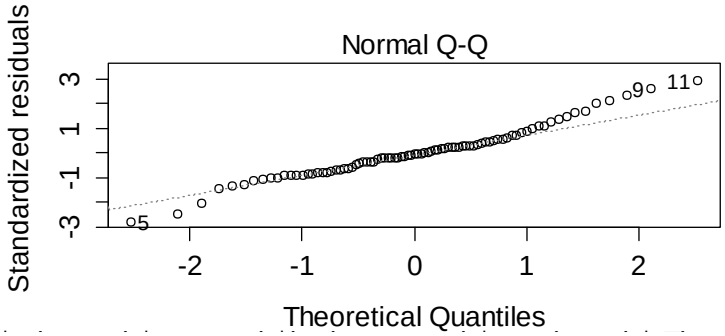
```
Call:
lm(formula = air$price ~ air$race + air$bedrooms + air$people +
    air$rTime + air$reviews + air$rRate, data = air)

Residuals:
    Min      1Q  Median      3Q     Max
-869.87 -218.16  -13.83  145.65  960.95

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1098.118    702.441   1.563 0.122033
air$race      -34.290     87.125  -0.394 0.694968
air$bedrooms  274.244     69.475   3.947 0.000172 ***
air$people    120.002     33.664   3.565 0.000626 ***
air$rTime       3.640      5.739   0.634 0.527751
air$reviews    -1.669      1.019  -1.637 0.105654
air$rRate    -940.075    730.466  -1.287 0.201917
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 337.6 on 78 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.7313,    Adjusted R-squared:  0.7106
F-statistic: 35.38 on 6 and 78 DF,  p-value: < 2.2e-16
```

Effectively, our model becomes:

Price = - 34.29 * (race) + 274.244 * (bedrooms) + 120.002 * (people) + 3.640 * (response time) - 1.669 * (reviews) – 940.075 * (response rate)

The model has an okay fit, with R^2 = 0.7106. Not surprisingly, the range of residuals is quite high. We see the fit with the following plots:





VI. Calculating Price Differential

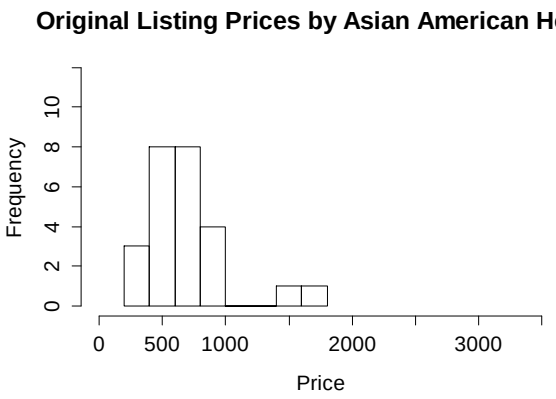
However, I am more curious about the difference in how much Asian American hosts can charge compared to White hosts. From the above model, it looks like bedrooms and number of people are the biggest indicators of price, yet we still see a statistically significant difference in the price charged by hosts. I want to fit a linear model to only the data from listings from White hosts, to see how they price their listings, then apply it to the listing details of Asian American hosts.

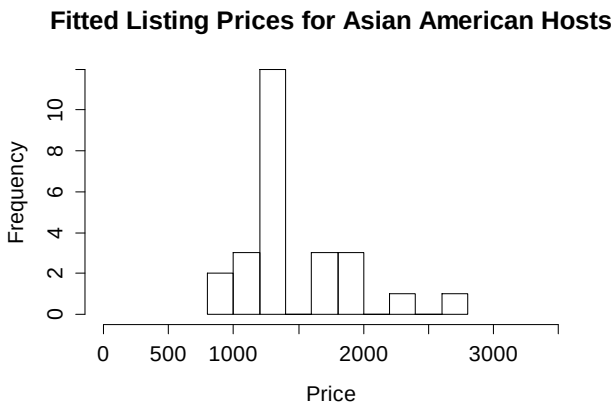
Fitting a model just to the White hosts:

```
Call:
lm(formula = price ~ bedrooms + people + rTime + reviews + rRate,
    data = white)

Residuals:
    Min      1Q  Median      3Q     Max
-890.48 -237.61    8.65  162.14  883.14

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  679.117    782.854   0.867 0.389513
bedrooms     273.055     78.814   3.465 0.001048 **
people       148.359     38.619   3.842 0.000324 ***
rTime          2.772      6.130   0.452 0.652952
reviews       -2.122      1.126  -1.885 0.064765 .
rRate       -569.751    802.998  -0.710 0.481048
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 336.6 on 54 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.7908,    Adjusted R-squared:  0.7714
F-statistic: 40.83 on 5 and 54 DF,  p-value: < 2.2e-16
```

This has a slightly better fit, with R^2 = 0.7714. We also see that reviews have a stronger influence on the price. We then use this model to calculate how much an Asian American host would be able to charge, and plotted the results below.

```
s = s.model$coefficients

fitmodel = s[1] + s[2] * asian$bedrooms + s[3] * asian$people + s[4] * asian$rTime
+ s[5] * asian$reviews + s[6] * asian$rRate
```





It's very evident that the fitted listing prices have a very different distribution compared to the original listing prices. The entire distribution is shifted towards the right, around a higher mean/median. The maximum value is larger as well. So it seems that Asian American hosts can charge a lot more, based on the attributes of their property.

I took this value and also calculated numerically how much more they could charge, by subtracting how much they were charging from the fitted values.

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
173.6   557.2   684.7   747.9   875.0  1848.0
```

So it seems on average, an Asian-American host can charge $747.9 more per week than they do currently (this was based on per-week data). The average listing price from our data is $700.64, so this difference would mean they could double their listing price. For our 25 Asian American hosts, this totals to $18,697.07 per week, not an insignificant amount.

VII. Limitations and Final Comments

Given my limitation as a single researcher on this project, the data that I collected - while it does contain enough data points – could be better. I would like to collect more data to see if this could be replicated across other ethnicities as well. I was mainly limited by not being able to categorize a host's race on my own; the Harvard researchers used Amazon Turk Workers to identify and validate results, and the initial source of my data had three researchers to validate. I did my best to mitigate this by scrubbing that raw data set to be only the data I needed and further expanded on it by doing my own research. With more data, I would like to see if my calculations on how much more Asian-Americans could earn would be more accurate.

My project looked at the probability a host could charge a higher price based on a number of attributes, including race, but it would be interesting to look at the flip-side as well: what is the chance you could predict a host's race based on the price and other property attributes?

While there is evidence of racial discrimination on Airbnb, it's difficult to see the driver of this discrimination. It could be due to cultural differences, other pricing mechanisms, personal preferences, or other factors, but that would require more accurate data on race, which Airbnb does not provide. With enough time, you could also compare other races and other cities as well, to see if the results are consistent. To its credit, Airbnb publicly states that it has a policy against discrimination. However, as we advance in our sharing economy, it will be interesting how Airbnb and other online marketplaces can design their platforms so both hosts and guests can receive the optimal experience, in regards to both pricing and availability.

VIII. Works Cited

Edelman, Benjamin G. and Luca, Michael. Digital Discrimination: The Case of Airbnb.com (January 10, 2014). Harvard Business School NOM Unit Working Paper No. 14-054. Available at SSRN: http://ssrn.com/abstract=2377353

Edelman, Benjamin G; Luca, Michael; Svirsky, Dan. Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment. January 6, 2016. Harvard Business School, Harvard Department of Economics. http://www.benedelman.org/publications/airbnb-guest-discrimination-2016-01-06.pdf

Data Sources:
http://dx.doi.org/10.7910/DVN/UBDYOO
http://insideairbnb.com/get-the-data.html