Unconscious Bias: An Exploration of Airbnb Open Data

Yinyu Ji, Emily Lu, Sam Woo, Valerie Zhang

December 6, 2019

I. Introduction

Unconscious racial bias is a well-documented phenomenon in many different contexts. Studies suggest that "unconscious bias may be related to clinical decision making and may predict poor patient-physician interaction" (Haider, Adil H., et al. 2011). Police are also more likely to check the vehicles of African-American motorists than white motorists for drugs (Knowles, Persico, and Todd 2001). Furthermore, "stereotypes about Black Americans may influence perceptions of intent during financial negotiations" (Kubota, et al. 2013); indeed, "participants are willing to discriminate against Black proposers even at a cost to their own financial gain."

Given the evidence that racial biases affect financial decisions, we are interested in the potential role of unconscious bias in a particular context: rental lodging. We believe that this is a fascinating context to explore due to the various choices that customers seeking lodging must make when choosing among locations and properties. In this study, we take a unique approach in studying unconscious bias by analyzing observational data collected by the popular online lodging marketplace, Airbnb. The platform has about 150 million users as of 2019[1] and has challenged the traditional hotel industry. Along with their first name or nickname, hosts on Airbnb publicly list their properties, varying from a single bedroom to entire mansions, with various details about the neighborhood and property. Recent research suggests that Airbnb's "beneficiaries are disproportionately white and high-wealth households" (Bivens 2019, p. 3), since white and high-wealth households are more likely to own nonprimary residential property.

In this study, we analyze a 2019 Airbnb dataset containing data about listings in New York City in combination with the dataset *Demographic aspects of first names* (Tzioumis 2018). We use various models including linear regression, backwards stepwise regression, tree-based models, and a mixed effects model to explore the relationship between features of the property listing and reviews per month, our response variable. In particular, we are interested in determining whether there is an association between the apparent ethnicity of a host's first name and their popularity of their listings on Airbnb, using reviews per month as a proxy for "popularity" of a property. Our results suggest that the host's apparent ethnicity may affect the choice of users to book and review the host's property, and that unconscious biases may affect the choices that users on the platform make when making a reservation.

II. Data

We combined data from two sources into one dataset. The first source is the 2019 NYC Airbnb Open Data found on Kaggle[2] and second source is *Demographic aspects of first names*. We

[1] For more statistics about Airbnb: https://muchneeded.com/airbnb-statistics/.

[2] Airbnb data available here: https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data?fbclid=IwAR34rWUhZD4nkHLnXgG9Os2G_saSCsI7hG1-BN8hwC49hhO1aWQxdNLcnjs. Though

chose the subset of Airbnb data due to its availability on Kaggle and because New York City is one of the most racially diverse cities in the world, allowing us to gather host first names of various ethnicities. The original dataset contains 47,905 unique listings and various information about the listing. *Demographic aspects of first names* is a list of 4,250 first names and information on their respective count and proportions across six mutually exclusive racial and Hispanic origin groups (consistent with the Census Bureau), collected from mortgage applications.

We focus on host first names to determine ethnicity as the literature suggests that stereotypical first names is associated with implicit bias, at least in the context of the classroom (Conaway and Bethune 2015). Importantly, the dataset contains only first names of the hosts, which is what customers see on the platform. Furthermore, we focus on the variable *reviews_per_month* due to research that suggests that online customer reviews significantly impact hotel business performance (Ye, Law, and Gu 2009). Thus we believe that reviews per month is a suitable proxy for a host's success and popularity.

Variable Name	Type	Description
log_rpm	double	Response variable; log of number of reviews per month.
neighbourhood_group	factor	Location (Bronx, Brooklyn, Manhattan, Staten Island, Queens)
room_type	factor	Listing space type (Entire home/apt, Private room, Shared room)
price	integer	Price in dollars
minimum_nights	integer	Amount of nights minimum
calculated_host_listings_count	integer	Amount of listings per host
availability_365	integer	Number of days when the listing is available for booking
ethnicity	factor	Predictor variable of interest, see *Data Wrangling*

Data Wrangling Methods:

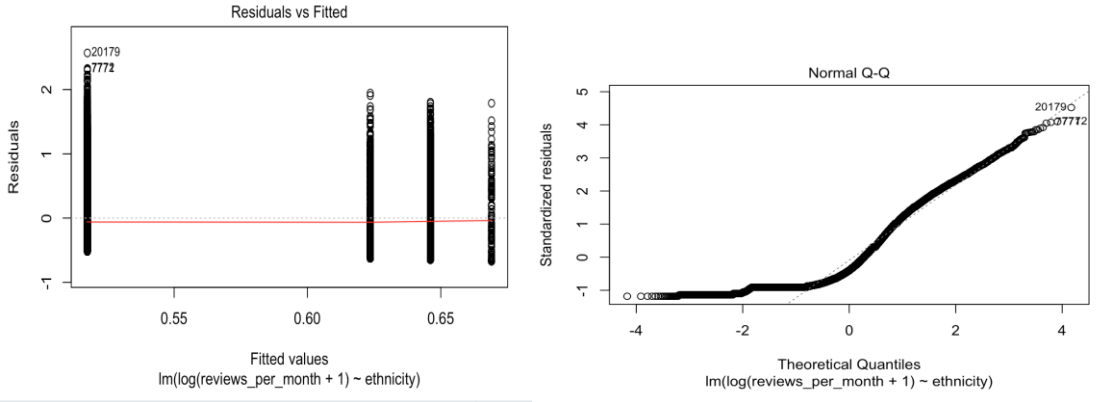
Of the variables contained in the Airbnb dataset, we dropped the following when working with the data: *number_of_reviews, description, id, date, obs*. These variables are intuitively poor predictors of the response, *reviews_per_month*; they are unique to each listing and are unlikely to explain any overall patterns in *reviews_per_month*. Based on our EDA, we decided to log transform *reviews_per_month* because it is extremely right-skewed (see R appendix); we imputed any log(0) with 0 to preserve the listing data points with 0 reviews per month.

we retrieved the data from Kaggle, the data is public by Airbnb and the original source can be found here: http://insideairbnb.com/. This particular dataset is collected from New York City in 2019 and was posted 3 months ago.

We also dropped any users whose names did not appear in *Demographic aspects of first names*: this includes any host names with more than one word were dropped (as some listings have multiple hosts), as well as individuals with more obscure names. These data points are hard to interpret, and it is logical to believe that these hosts are less likely to be typecast, justifying their removal. To create the *ethnicity* predictor variable, we assigned each first name to the race with the highest percentage of representation in the *Demographic* dataset. The problem in labelling each host with a race is that first names are highly variable (common first names appear in all races). This method guarantees that every host first name is associated with one race, allowing for better interpretability. Intuitively, it is likely that an individual will only associate a name with one race.

III. Models

Because of space constraints, It suffices to check the linear regression assumptions for our baseline model, which attempts to predict log(reviews_per_month) with ethnicity.



Because we only have a categorical variable, the linearity assumption is automatically met. The constant variance assumption is violated. Based on the QQ plot, the normality assumption is definitely not met. Thus, because the assumptions are not met, we should be wary about the results obtained through linear models.

Simple Linear Regression and Backwards Stepwise Regression

The simple linear regression model included all of our predictors , which are the following variables: *neighbourhood_group, room_type, price, minimum_nights, calculated_host_listings_count,* and *availability_365*. At the 0.05 significance level, we found that factor indicators for neighbourhood group (*neighbourhood_groupBrooklyn, neighbourhood_groupManhattan, neighbourhood_groupQueens*), minimum nights, calculated host listings count, and availability were significant.

Simple Linear Regression 1 (*ethnicity* NOT included)	**Coef. Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	-3.383e-01	6.002e-02	-5.636	1.76e-08 ***

neighbourhood_groupBrooklyn	-4.227e-01	5.963e-02	-7.089	1.39e-12 ***
neighbourhood_groupManhattan	-3.770e-01	5.978e-02	-6.307	2.90e-10 ***
neighbourhood_groupQueens	-1.343e-01	6.326e-02	-2.124	0.0337 *
neighbourhood_groupStaten Island	-1.434e-01	1.094e-01	-1.311	0.1898
room_typePrivate room	8.887e-03	1.725e-02	0.515	0.6065
room_typeShared room	2.613e-02	5.613e-02	0.466	0.6415
price	-2.135e-05	3.296e-05	-0.648	0.5172
minimum_nights	-5.380e-03	3.937e-04	-13.666	< 2e-16 ***
calculated_host_listings_count	-1.236e-02	8.984e-04	-13.754	< 2e-16 ***
availability_365	2.791e-03	6.828e-05	40.872	< 2e-16 ***

This linear regression model was then used as the starting model for backwards stepwise regression. We cannot directly interpret the p-value results of a backwards stepwise regression as they are artificially inflated by the variable selection process. To make inferences using backwards stepwise regression, we start using a linear model using all of the predictors excluding the predictor variable of interest (here, *ethnicity*).

After performing backwards variable selection, the following predictors remained and are therefore significant: *neighbourhood_group, minimum_nights, calculated_host_listings_count,* and *availability_365*.

Backwards Stepwise Regression 1 (*ethnicity* NOT included)	**Coef. Estimate**	**Std. Error**
(Intercept)	-0.3338703	0.0589911
neighbourhood_groupBrooklyn	-0.4246815	0.0595996
neighbourhood_groupManhattan	-0.3813673	0.0596000
neighbourhood_groupQueens	-0.1349074	0.0632512
neighbourhood_groupStaten Island	-0.1445586	0.1093798
minimum_nights	0.0003931	0.0003931
calculated_host_listings_count	-0.0123542	0.0008975
availability_365	0.0000679	0.0000679

Then, we add back *ethnicity* to generate a “full” linear regression model (see R Appendix for model summary). This model is then used to generate another backwards stepwise model. As we can see, our factor indicators for *ethnicity* remain in the model after variable selection, suggesting that *ethnicity* is a statistically significant predictor of reviews per month.

Backwards Stepwise Regression 2 (*ethnicity* NOT included)	**Coef. Estimate**	**Std. Error**
(Intercept)	-3.713e-01	5.927e-02
neighbourhood_groupBrooklyn	-4.036e-01	5.970e-02
neighbourhood_groupManhattan	-3.619e-01	5.969e-02
neighbourhood_groupQueens	-1.281e-01	6.327e-02
neighbourhood_groupStaten Island	-1.332e-01	1.093e-01
minimum_nights	-5.350e-03	3.929e-04
calculated_host_listings_count	-1.218e-02	8.972e-04
availability_365	2.775e-03	6.788e-05
ethnicityHispanic	1.691e-01	3.012e-02
ethnicityBlack	1.663e-01	1.235e-01
ethnicityAPI	2.050e-01	5.213e-02

Regularization

With many variables in the linear models (multiple types of neighborhood groups and room types), we decided to regularize to reduce the complexity of the resulting model. This was done through utilizing the Lasso and Ridge methods where we could see which variables were the most “significant” in predicting our response, *log_rpm*. We started with the“full” model with all the variables included and then implemented lasso/ridge with lambda chosen by cross validation. We used the *glmnet* library with the function *cv.glmnet*(). The resulting equations were:

LASSO

log_rpm = -0.6949548 + (0.143193219)**neighbourhood_groupQueens* +
(-0.003312989)**minimum_nights* +
(-0.005319552)**calculated_host_listings_count* +
(0.002304800)**availability_365* + (0.012289586)**ethnicityHispanic*

Ridge

log_rpm = -6.27e-01 + (-8.34e-02)**neighbourhood_groupBrooklyn* +
(-5.70e-02)**neighbourhood_groupManhattan* +
(1.76e-01)**neighbourhood_groupQueens* +
(2.67e-01)**neighbourhood_groupStaten Island* +
(2.94e-02)**room_typePrivate room* + (8.36e-02)**room_typeShared room* +
(1.64e-05)**price* + (-3.81e-03)**minimum_nights* +
(-6.41e-03)**calculated_host_listings_count* +
(1.92e-03)**availability_365* + (1.39e-01)**ethnicityHispanic* +
(2.15e-01)**ethnicityBlack* + (1.53e-01)**ethnicityAPI*

As expected, there were fewer variables remaining in the Lasso model than the Ridge and original model. The “important” variables included both neighborhood groupQueen and ethnicityHispanic which were variables that related to our discussion of racial bias. However, with this model formula, we cannot determine the relative importance within the remaining variables as the variables were not standardized. The results of the Lasso suggests that Hispanic hosts have different experiences with the average number of reviews that they get compared to the White hosts (they get slightly more reviews). The Ridge model contained all the variables still but with much smaller coefficients than the simple model as expected. In this model, isHispanic also had a positive association with *log_rpm* supporting the claim above regarding Hispanic hosts having more reviews than White hosts.

Tree-based methods (Decision Tree, Random Forest)
The decision tree and random forest both included essentially all of the predictors, eliminating the *neighborhood* variable from the rf model as the randomForest function cannot take categorical variables with more than 53 categories. In the rf model, we applied the following adjustments: (1) our decision to set `mtry` to 8 was achieved through cross-validation, (2) our adjustment to `maxnodes` was achieved through minimizing the difference between the RMSEs for train and test sets to minimize model overfitting.

```
tree = rpart(log_rpm~.,data=train,control=list(maxdepth = 20))
rf = randomForest(log_rpm~.-neighbourhood,data=train, mtry = 8,maxnodes=50)
```

Decision trees have no probabilistic assumptions on both predictors and the response variables (unlike linear regression models). The modelling is based on simply reducing error with discrete steps in each branch of the tree. The lack of assumptions in decision tree models carries on to the Random Forest model which is simply an ensemble of decision trees.

Mixed Effects
We first found potential confounders with ethnicity through variable selection: I used backwards variable selection with ethnicity as the lower bound. The formula was as follows:

```
log_rpm ~ neighbourhood + minimum_nights + calculated_host_listings_count +
            availability_365 + ethnicity
```

For the mixed model approach, we clustered based on neighborhood because it is reasonable to assume that predictors for hosts in the same neighborhood are correlated. (Here we did not cluster by burrough/neighbourhood group because that would have resulted in too few clusters.) We include ethnicity as both a fixed effect and random effect because in order to infer whether the overall average effect is significant and to control for its effect. From Lecture 23 Slide 31, a variable should be included as a random effect to control for its effect. We chose a smaller subset by only using variables remaining in the backwards variable selection as random

effects: this makes sense because the variable selection yields variables that are related to the response variable, which makes them potential confounders with inactivity. (Recall that confounding variables with inactivity are correlated with both the response, log_rpm, and inactivity.) The formula is therefore as follows:

Model	Estimate	Std Error	T Value
Intercept	0.05864	0.02363	2.482
Hispanic	0.29436	0.11493	2.561
Black	0.17633	0.48365	0.365
API	0.32521	0.16192	2.008

Because the magnitude for the t values for Hispanic and API is greater than 2, it follows the Hispanic and API log_rpm is significantly different from that of whites at the 0.05 level. (Although we do not know the degrees of freedom, consulting a t-test critical values table suggests that 2 is a cutoff value at the 0.05 level). The magnitude for Black is 0.365, suggesting that the log_rpm for Black is not significantly different from White.

We will not consider the Anova F-test with a baseline model, as it results in convergence problems. As discussed with Kevin in office hours previously, using Anova to determine whether or not a predictor is significant runs into convergence problems. The t-test analysis is rigorous enough.

IV. Model Selection and Analysis

Selection Criteria

Model	Train RMSE	Test RMSE
Simple Linear Regression w/ Ethnicity	1.325468	1.324521
Backwards Stepwise Regression w/ Ethnicity at start	1.325481	1.324546
Ridge	1.331079	1.331102
LASSO	1.330574	1.330796
Decision Tree	1.198253	1.200165
Random Forest	1.155058	1.163688
Mixed Effects	1.385671	1.391124

In our examination of the respective RMSE calculated from the train and test sets, we hope to minimize RMSE's values overall, as well as their differences between the train and test set. The random forest minimizes RMSE's overall, and the difference between train and test is minimal in order to reduce model overfitting.

Additionally, the linear regression, regularization, and mixed effects models all rely on the assumptions for linear regressions. (The mixed effects model first uses variable selection, which uses a linear model.) However, as discussed earlier, the normality assumption and constant variance assumptions are violated. Therefore, it makes sense to prioritize other methods learned in class, such as decision trees and random forests, when generating a predictive model.

V. Conclusion

From evaluating the RMSE's, it follows that the random forest model offers the optimal predictive model. Again, the model is of the following form:

```
rf = randomForest(log_rpm ~.-neighbourhood,data=train, mtry = 8,maxnodes=50)
```

It is difficult for us to graphically capture the effects of *ethnicity* within a tree-based model, for which standard plots rely upon quantitative thresholds. There is an argument to be made about adjusting our main predictor variable, *ethnicity*, to be a quantitative variable based on how we create the variable through other data wrangling methods. However, in order to maintain consistency across our models and to evaluate our models consistently, we chose not to do this.

From the linear regression, regularization, and mixed models approaches, we see that *ethnicity* appears to be significant when predicting reviews per month (*log_rpm*). From the backward stepwise approach, we saw that *ethnicity* remains in the model after variable selection. From the lasso approach, we concluded that Hispanic hosts have slightly more reviews on a log level than White hosts. Finally, for the mixed effects model, we concluded that Hispanic and API *log_rpm* is significantly different from that of whites at the 0.05 level. Thus, the multitude of these interpretations makes our conclusion that ethnicity as a whole (or at least having certain ethnicities) is significant to predicting reviews_per_month more robust. Putting these pieces together, we now interpret the findings in terms of the larger context of our data. Ethnicity is significant to predicting reviews_per_month, and because reviews_per_month is a reasonable proxy for popularity, we can conclude that an Airbnb host's popularity is influenced by their ethnicity.

However, although *ethnicity* is a significant predictor, it may not be the most significant predictor. The many models we fitted can help us compare the significance between predictors: for the Lasso approach, the magnitude of the "Queens" coefficient is ten times that of "Hispanic," which in turn is much larger than the other potential confounders, `minimum_nights`, `calculated_host_listings_count`, and `availability_365`. The largest magnitudes within the Ridge model are associated with *neighborhood* and *ethnicity*, also suggesting the potential significance of *neighborhood* in predicting *log_rpm*. The mixed effects model allows us to control for *neighborhood*: after controlling for *neighborhood*, we see that being API and Hispanic

is significantly different from being white. Therefore, we believe that our study adds to the literature that *ethnicity* is in fact associated with a customer's economic decisions; we showed that *ethnicity* implied by first name is associated with the log reviews per month of Airbnb listings.

Moving forward, we hope to address the limitations of this approach. Since the data we used in this study was observational, we cannot make causal inferences regarding the impact of host name ethnicity on hosting popularity or success. In an ideal study, we would be able to control for potential confounders that may impact a customer's decisions while browsing Airbnb to isolate the effect of ethnicity. These confounders may include even the time of day a customer is viewing listings and the customer's own socio-economic or racial background. We would also try to collect a response variable that better reflects hosting popularity, such as review scores or sentiment. Additionally, the different racial categories were not equally represented in our data, which may lead to uncertainty in our results. Further studies into this subject may investigate the implicit biases of Airbnb customers based on the reviewer's own self-identified ethnicity. We may also want to explore the effects of host full or last names on reviewer perception.

VI. References

Conaway, Wendy, and Sonja Bethune. "Implicit Bias and First Name Stereotypes: What Are the Implications for Online Instruction?." *Online Learning* 19, no. 3 (2015): 162-178.

Haider, Adil H., et al. "Association of unconscious race and social class bias with vignette-based clinical assessments by medical students." *Jama* 306.9 (2011): 942-951.

Knowles, John, Nicola Persico, and Petra Todd. "Racial bias in motor vehicle searches: Theory and evidence." *Journal of Political Economy* 109.1 (2001): 203-229.

Kubota, Jennifer T., et al. "The price of racial bias: Intergroup negotiations in the ultimatum game." *Psychological science* 24.12 (2013): 2498-2504.

Tzioumis, Konstantinos (2018) Demographic aspects of first names, Scientific Data, 5:180025 [dx.doi.org/10.1038/sdata.2018.25].

Ye, Qiang, Rob Law, and Bin Gu. "The impact of online user reviews on hotel room sales." International Journal of Hospitality Management 28, no. 1 (2009): 180-182.

Bivens, Josh. “The economic costs and benefits of Airbnb”. *Economic Policy Institute*. (2019): 3-4.