Team J: Nina Sonneborn & Rebecca Conover
MATH 218 Statistical Learning Final Project
Kaggle: Airbnb

Our ultimate approach:
The Airbnb Kaggle competition gave new user and sessions data to predict which country the user's booking would be in. We used CART for prediction and cross validated the rpart complexity parameter 'cp' which essentially controls how much complexity is allowed in the model. An advantage to CART is that it does variable selection. For many categorical predictor variables, there were several levels but most of the data fell within only the few most popular levels. Thus, we re-leveled categorical variables to have less levels and used these new "pruned" variables in CART.
No matter how much we allowed a more complex tree, CART only predicted outcomes NDF (no destination found) and US. This make sense, as the training set is 58% NDF and 29% US. Ultimately, our cross validated score was very different from our submission score on Kaggle. We then realized that the train and test sets were NOT random samples of the total population. Rather they were from different times, and to account for this we would have to incorporate into our model. This is not something we got around to, but we will speculate possible approaches:

- We could study time trends in booking destinations and further incorporate the time variable into our prediction model.
- We could further investigate 'sessions.csv' to study time trends.
- We could only use training data from 2014

Extras:
Originally, we were working on Expedia's kaggle competition, hoping to leverage the EDA we did during DataFest. The Expedia competition used data about users and hotels to predict which hotel cluster a user would book. This competition proved very difficult, as the outcome variable was categorical with 100 levels. Additionally, many predictor variables were masked, and therefore hard to think about intuitively or do EDA with. For example, 'hotel_country', a categorical variable was coded as integer values, one integer per country. We performed Principal Components Analysis on the destinations data set, which included 150 continuous variables that were "latent descriptions" of hotels. We also spent a lot of time trying to do shrinkage with multinomial glmnet for variable selection, KNN for prediction, and CART for prediction, but encountered many run time issues. Thus, we switched to Airbnb, thinking that the nature of the competition was similar (multi-level categorical outcome variable, hospitality related data), so we could use similar strategies.

In Airbnb, the training set consisted of data from 2011-June 2014, while the test set had observations from 2014, thus the training set was not representative of the test set as the behavior of users changes over time. This was especially visible in the difference in the proportion of NDF or "no destination found" outcomes between the train and test sets, 58% and 68% respectively.

We also attempted a multinomial LASSO that we initially hoped to use for variable selection. This feature appears to be relatively new in the glmnet package and we found that broom does not work on this model, which made things very difficult. Nina wrote a function that took the model's betas for each variable

across a range of lambdas for each outcome variable and put this data into tidy format. The resulting visualizations were much more interpretable but because we had a plot for each outcome as opposed to a single visualization that summarized these results it wasn't interpretable enough, so we gave up on that. We also were able to cross validate the multinomial LASSO to find an optimal lambda that we hoped to make predictions on. Again, because this package is pretty new, we ran into too many speed bumps and the prediction function did not work. After some research we found that we might be able to solve this using a sparse matrix, but there was a high probability that we would run into more trouble so after this long, arduous journey with glmnet and LASSO, we decided to use CART (which also does variable selection).

In the end, while our Kaggle score alone doesn't show it, we learned a lot! And found relatable memes. https://www.facebook.com/Rmemes0/?hc_ref=SEARCH