Team Bread: Josh Campbell, Lucy Yuan, Fengling Hu, Audrey Cheng, Laboni Hoque

# Datafest Write-up

In approaching the Expedia data, our group was interested in modeling how Expedia could improve their advertising techniques. We focused our analysis on the United States because there were many sources of external data about the United States available to us. First, we decided to explore click density. We created this variable by dividing the number of clicks per county by the population to find the density of clicks. Click density is a measure of the amount of popularity and traffic Expedia experiences on its site, so this variable would be of interest to Expedia, as people can only book travel plans if they actually end up on Expedia's site. We explored this variable with supplemental data from the 2010 US Census and the CoolClimate study conducted at UC Berkeley (http://coolclimate.berkeley.edu/data).

We decided to try to model the click density in a county by focusing on different characteristics of the county, eventually selecting average household income, average household value, CO2 produced through transport, and international migration as predictor variables using stepwise regression. We used a multiple linear regression model to attempt to model the click and obtained a moderate R^2 value of approximately 0.281. Thus, there is a significant percentage of the variation in click density that cannot be explained by the external variables about the county. Expedia can therefore influence the unexplained variation found in click density through advertising efforts, either through brand-based promotion or click-based advertising.

We show a graphical depiction of click densities per county represented on the United States map. Darker areas indicate higher levels of Expedia usage, and lighter areas indicate places with less Expedia usage, scaled by population densities. Expedia can use these data to explore areas where advertising should be more prevalent. For example, even when we account for low population density in the Rocky Mountain region of the United States, there is a still a very low amount of click density. Expedia could choose to focus more expenditure in that area. However, on the east and west coasts, we see large click densities. Expedia likely has strong brand recognition in these areas and may be able to reduce advertising efforts in these regions.

Then, we decided to analyze if the channel through which people accessed Expedia and the time of day people clicked were independent, using a chi-square analysis. We found that there was sufficient evidence to reject the null hypothesis, so the channel and time of day people clicked were different. We see that, at different times of day, people are accessing different channels in different proportions, so Expedia can focus its advertising differently throughout the day. For example, it can spend a greater percentage of its revenue on channel 293 in the very early morning (4am-8am), as many more users come from that channel during this time period. In addition, we analyzed if the channel and region of the United States people were in were independent using a chi-square analysis. Again, we rejected the null hypothesis and concluded that differences between channels likely do exist by region. For example, looking at the graph, we see channel 231 is more often accessed in the Northeast and West than in other regions, so advertising can be focused on that channel. (If 231 is direct access to Expedia via "Expedia.com", this further supports that brand recognition is strong in these regions). However, since there were so many observations in this dataset, these differences between channels for regions and times may or may not be practically significant. Through our analysis, we have explored several ways Expedia can improve its advertising by focusing in on different channels, regions, and demographic characteristics.