**METHODOLOGY & RESULTS**

**FLATTENING AND EXPORT TO CSV**

Used a program called SQLite Maestro to execute the query that is in the file create\_records\_flat.sql. Then exported the table to CSV (records\_flat.csv).

**EXPLORATION & MODEL**

Initially, an exploration and development of syntactically correct model was performed (contained in census\_logistic.r). Crosstables were created for exploration and based on odds-ratios and correlation, a mosaic plot of over\_50k with education\_num was created to represent one of the more important relationships in the data (education\_over\_50k\_mosaic\_plot.png). It was observed that the higher the education\_num, the proportion of over\_50k observations gets larger. This is also where it was determined that education\_num will be used over highest\_education\_lvl due to multicollinearity.

Developed a logistic regression model in R which is located in the census\_logistic\_v2.r file. This file starts with recoding missing values, creating an us\_flg variable to take the place of country\_of\_origin (to avoid quasi-complete separation), and re-assigning “never-worked” to “without-pay” for the same reason. The data is then split into training and validation using the caret.createDataPartition function. The next step is the development of the logistic regression model where capital\_gain was dropped due to quasi-complete separation, race was dropped due to insignificance. The concordance on the final model ended up being 0.89. Lastly, the validation dataset was scored and using a cutoff of 50%, the misclassification rate is 0.1665 (% predicted correct is 0.8335)