Given the installs of the subset of apps with a higher value of installs, we can see where consumers are eager to install new apps. This helps developers outside and inside of the workings of the Play Store to determine where they want to channel their efforts.

What is a good app to develop to reach the highest number of consumers based on the categories or genres consumers are more likely to install from the Google Play Store? The higher number of installs means more consumers will stick with an app for the long-term, so seeing where the highest number of installs is happening now is a great way to target the highest opportunities for new developers.

In trying to determine the answer to this question, we are going to need to undergo prediction. We used the rating, category, and genre as the input columns for the prediction models for our first try. It seemed after our PowerBI visualizations, however, that not only were many ratings missing for apps in our dataset, but most of the others has ratings above four. As such, there was a strong correlation, so we decided to remove rating.

First, we created a naïve bayes mining structure. This showed quite clearly that Rating was the best indicator of sales before we removed it, but after we removed it, Category was the strongest indicator. Next, we tried a decision tree structure. This gave us information on which categories were avoided.

According to the decision tree model mining model viewer, an app is more likely to be purchased if the rating is between 4.2 and 4.6, especially if the category is not “Medical.” So, it seems developers do not want to develop medical apps. This seems to make sense due to less of the population having a use for specialized medical apps.

After tried without the rating, the highest categories for probability of installing an app were, in order, family, finance, and education. In addition, not only were medical apps less likely to be installed,

but games were less likely to be installed. My suspicion is that this is due to oversaturation in the games market.

The naïve bayes model clearly shows that category has the biggest impact on app installations. Both models according to the lift charts below are, as it goes, terrible at predicting installations as shown according to the lift chart below. Using the information from the chart, however, we can see that the Naïve Bayes model is slightly better at predicting.