Steps and Future Work:

To identify users that are likely spend money in their game after finishing the tutorial, three main phases are investigated.

First, the data is joined and load, and visualizing is done to understand data. The images are attached in the appendix at the end of this PDF.

Second, Preprocessing and feature selection are done. Duplicates and remaining Nan values are removed. Then, categorical features are changed to numerical. Next, an automatic feature selection is done using random forest algorithm. Then, data is split into train and test. Next, data is normalized using train data. Because data is very unbalanced to determine total spend, downsampling and oversampling using random sampler and SMOTE is done. Applying PCA and drop categorical columns can be done too.

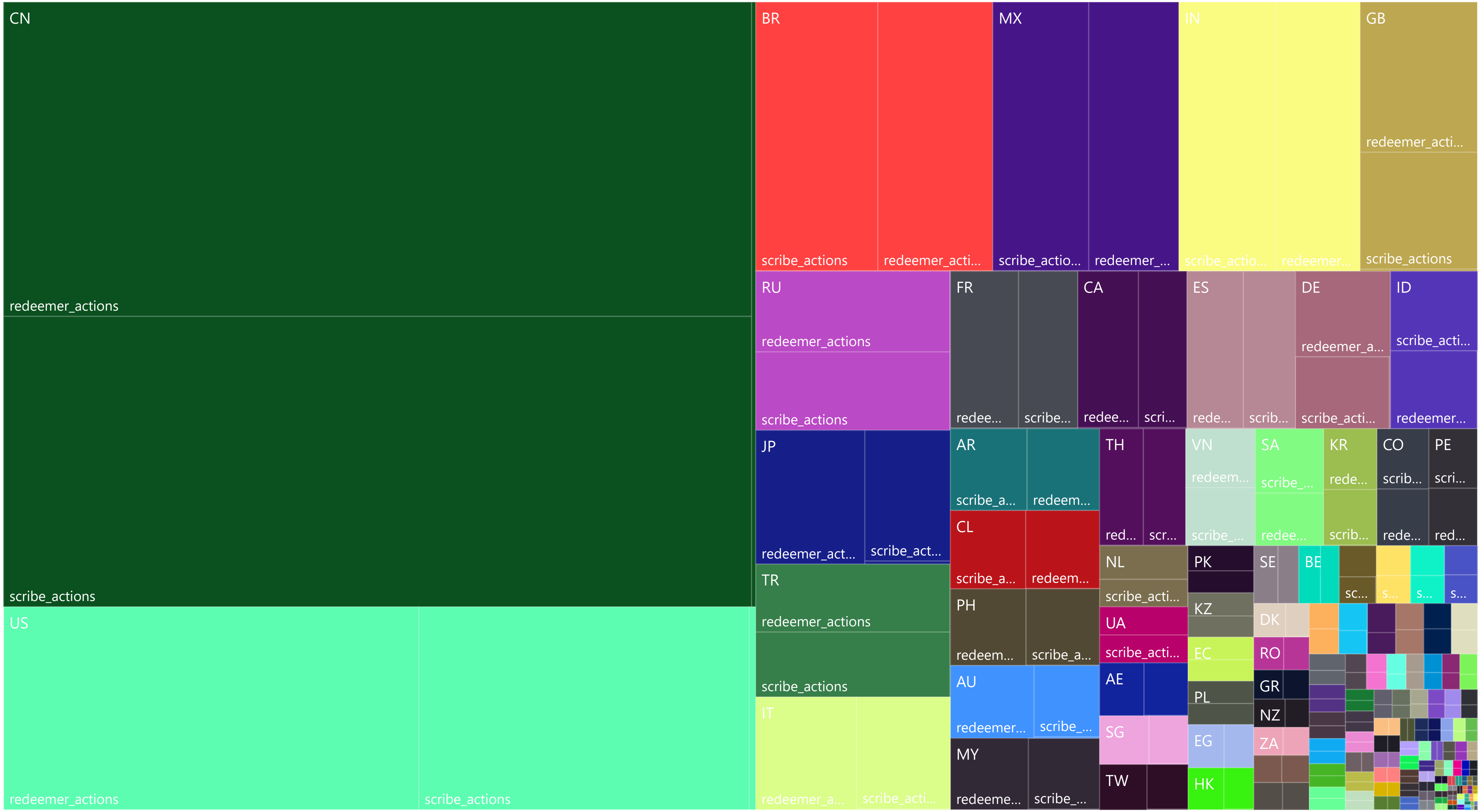
Third, Machine Learning and Deep Learning models are applied. To detect users who likely spend money two approaches are given: classification and regression. Because the game designers' question is including a suggestion for different prices then the total spend can be viewed as a continues for regression problem or it can be categorized into different ranges to deal with classification. Both approaches are done using different classification and regression models. For regression, different methods such as LSTM, XGBoost, Kernel Ridge etc. are applied and, MSE and MAE errors are reported. For classification, different methods such as LSTM, KNN etc. are applied, precision, recall and F-measure are reported.

The results show that imbalance data affect the model very much. For regression, Kernel Ridge results were better than other models. For classification, the minority class identification is a challenge. Precision, recall and F-measure for majority class was very good, however, the minority class criteria were not good generally. This is a very common problem in imbalanced data which needs more investigation with more complex models (for example hybrid models). It is interesting to mention that because the precision, recall and F-measure for the class of users who are not likely to spend money on game are very high, then we can look at the problem reverse (too many False Positive is responsible for low accuracy of negative class). In this scenario, the users who are most likely won't spend will be detected and remain users should be investigated more (a very big chunk of users will be eliminated).

For future work, anomaly detection and hyperparameter tuning using Bayesian optimization or Genetic Algorithm can be done. However, preprocessing is the most important step, and more investigation is needed on preprocessing.

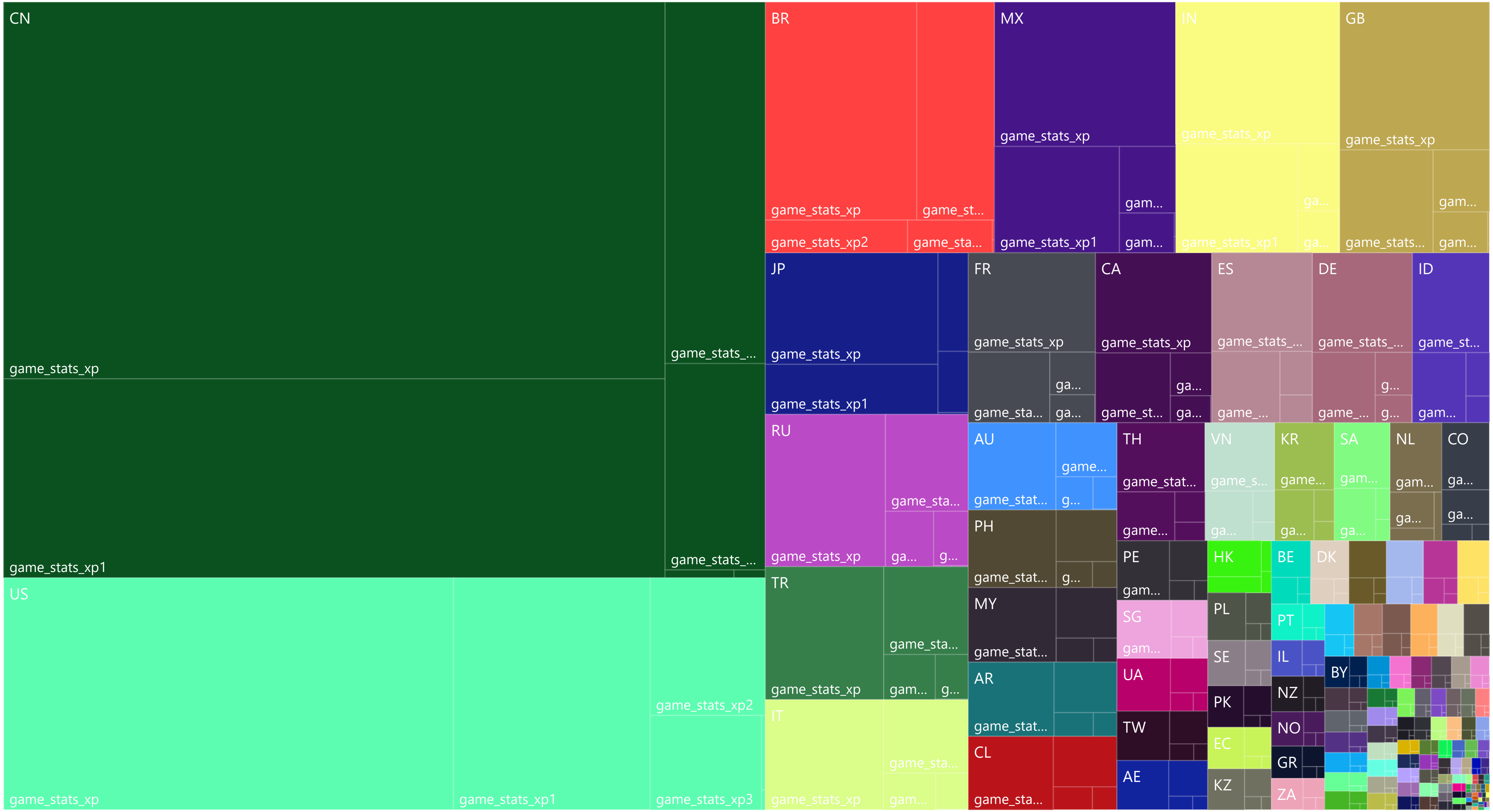
Appendix:

total_spend, redeemer_actions and scribe_actions by geo_s



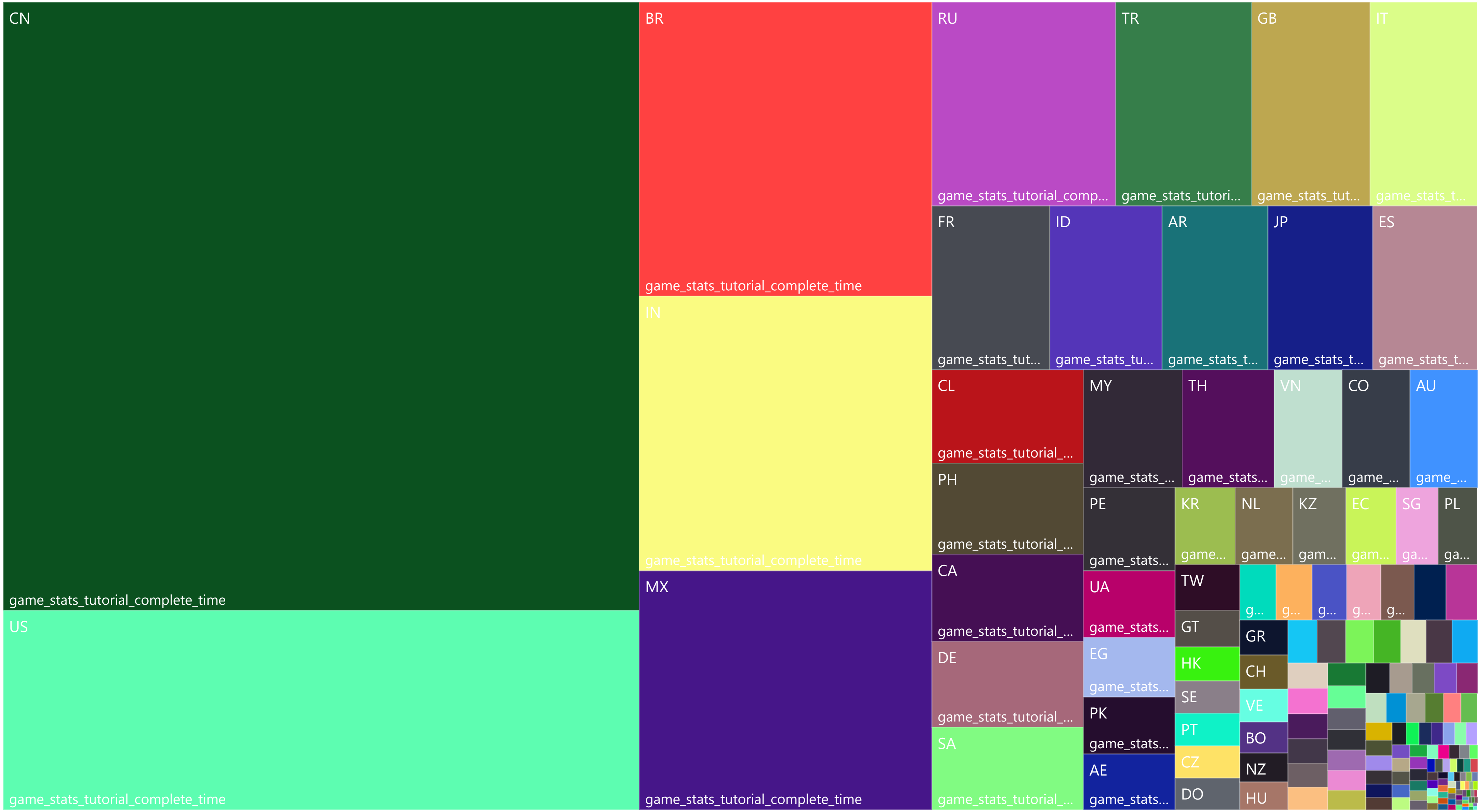
It is not surprising to see most of the people who spend money are from China and US. Brazil, Mexico, India and Great Britain are in the next positions.

game_stats_tutorial_complete, total_spend, game_stats_xp, game_stats_xp1, game_stats_xp2 and game_stats_xp3 by geo_s

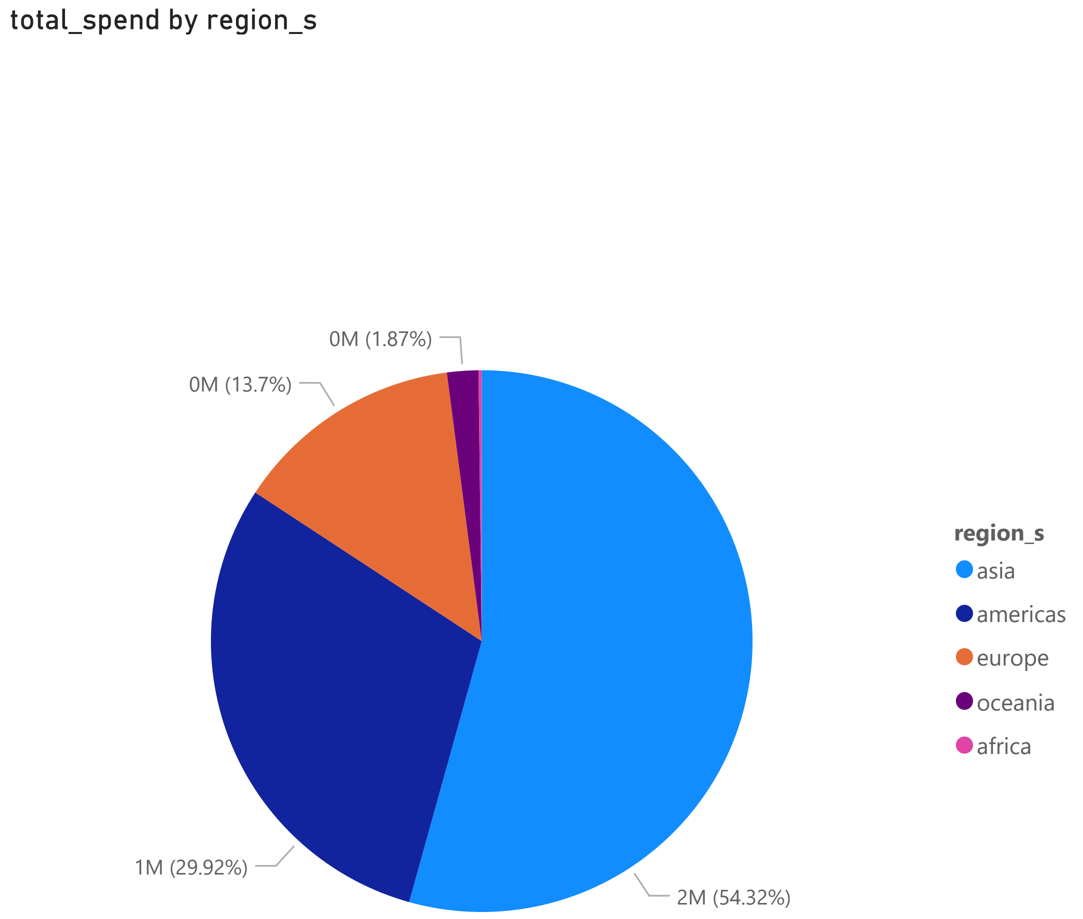


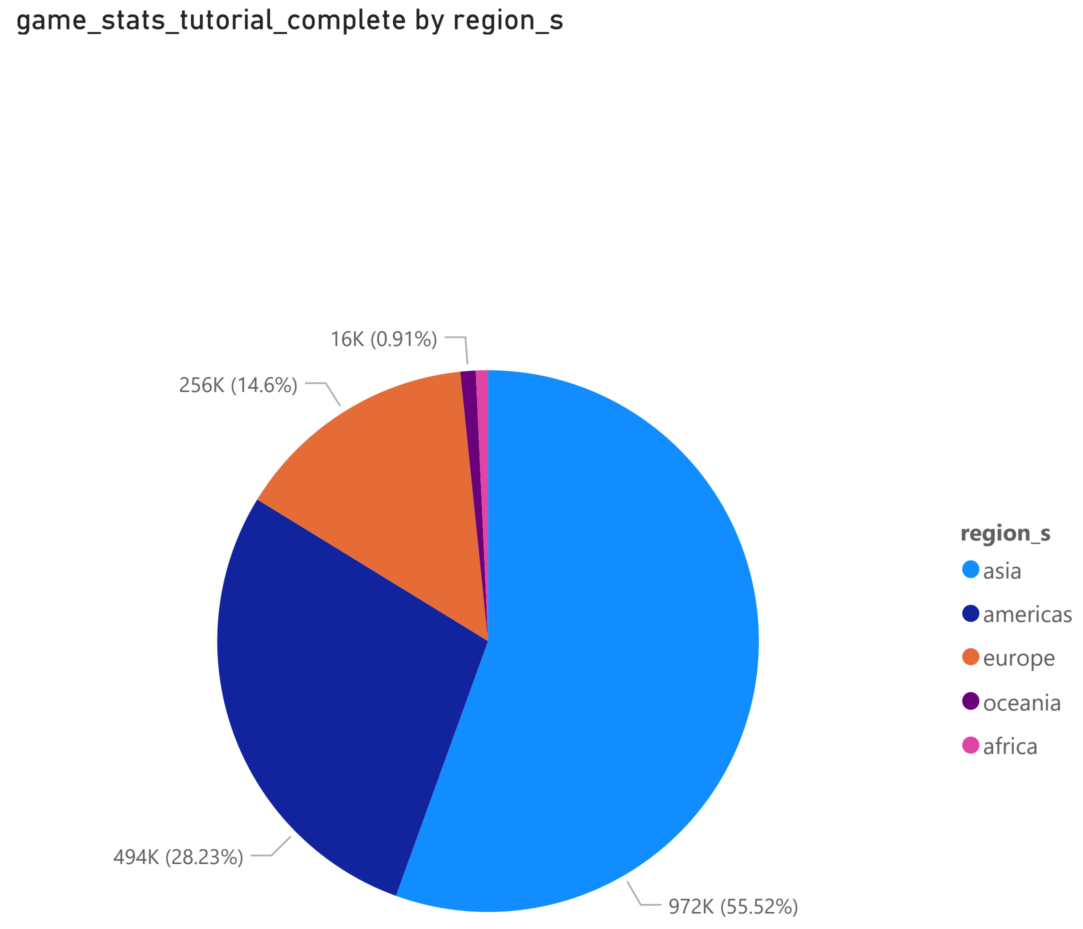
Again the same position based on player who completed tutorial.

total_spend and game_stats_tutorial_complete_time by geo_s



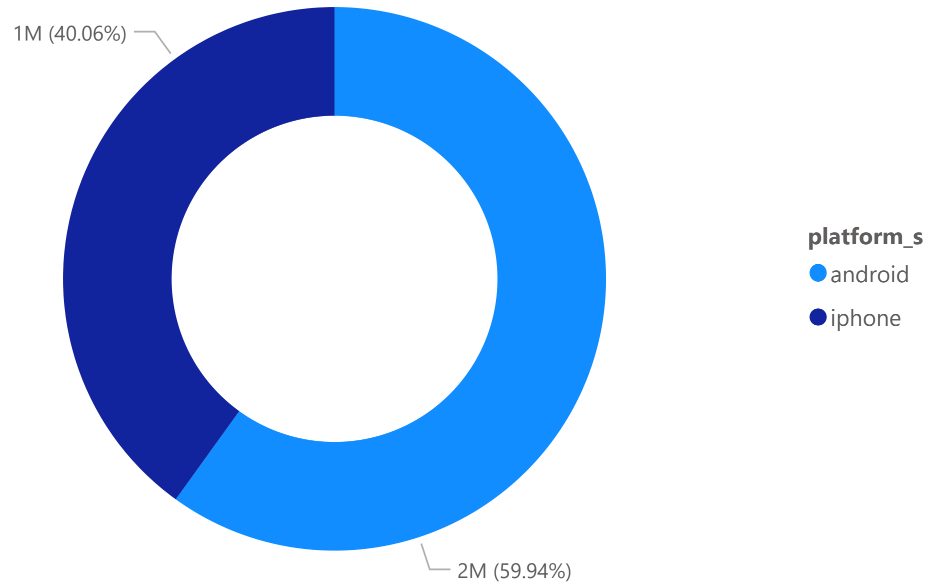
For complete time, Brazil, India and Mexico are in the positions 3rd to 5th.



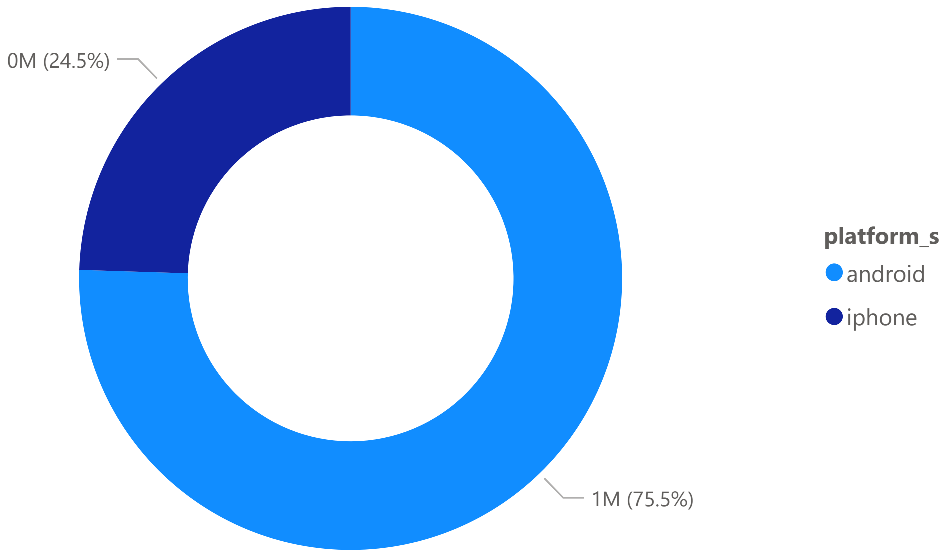


It is interesting to see total spend and users who completed tutorial are identical by region. 85% from Asia and America.

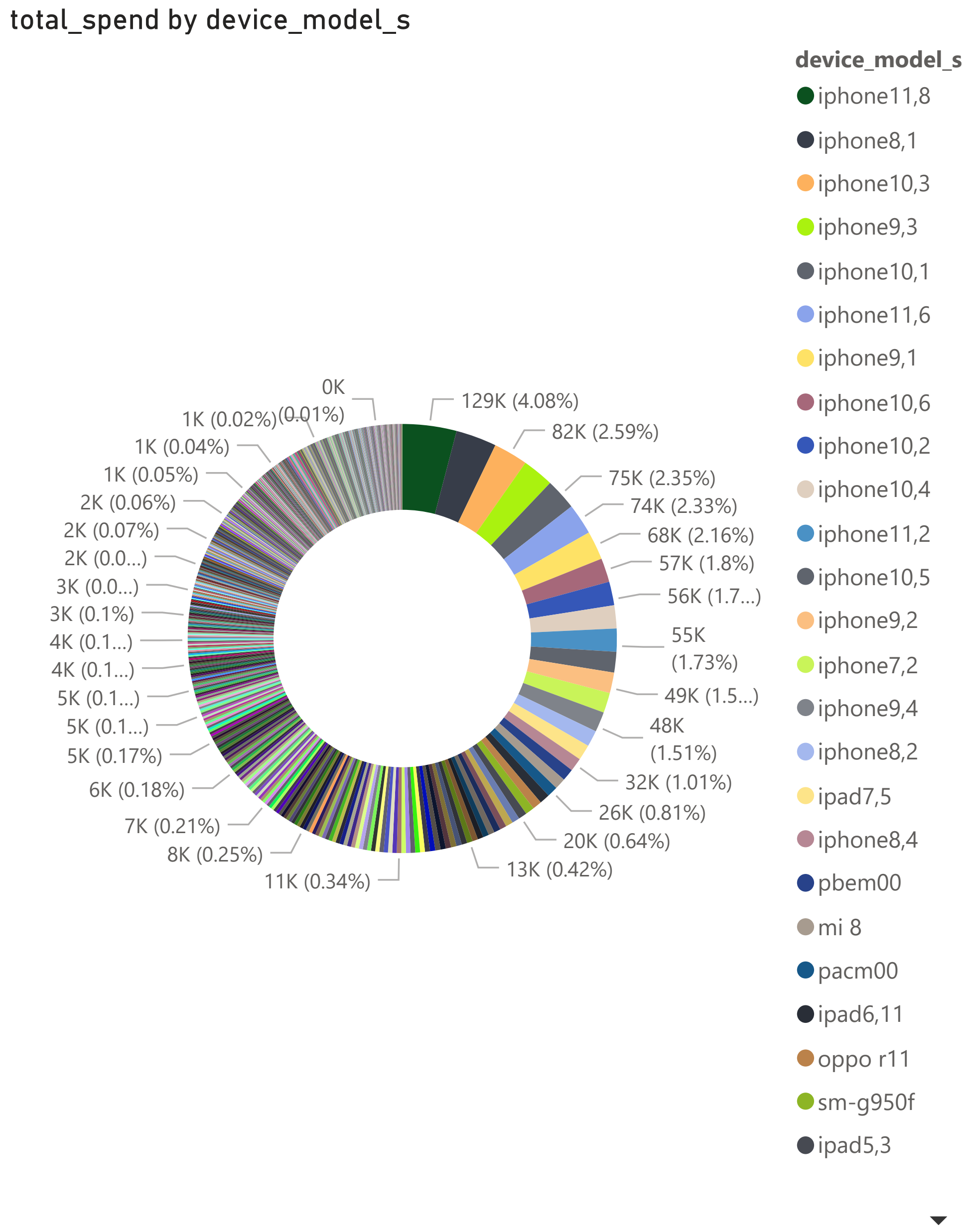
total_spend by platform_s

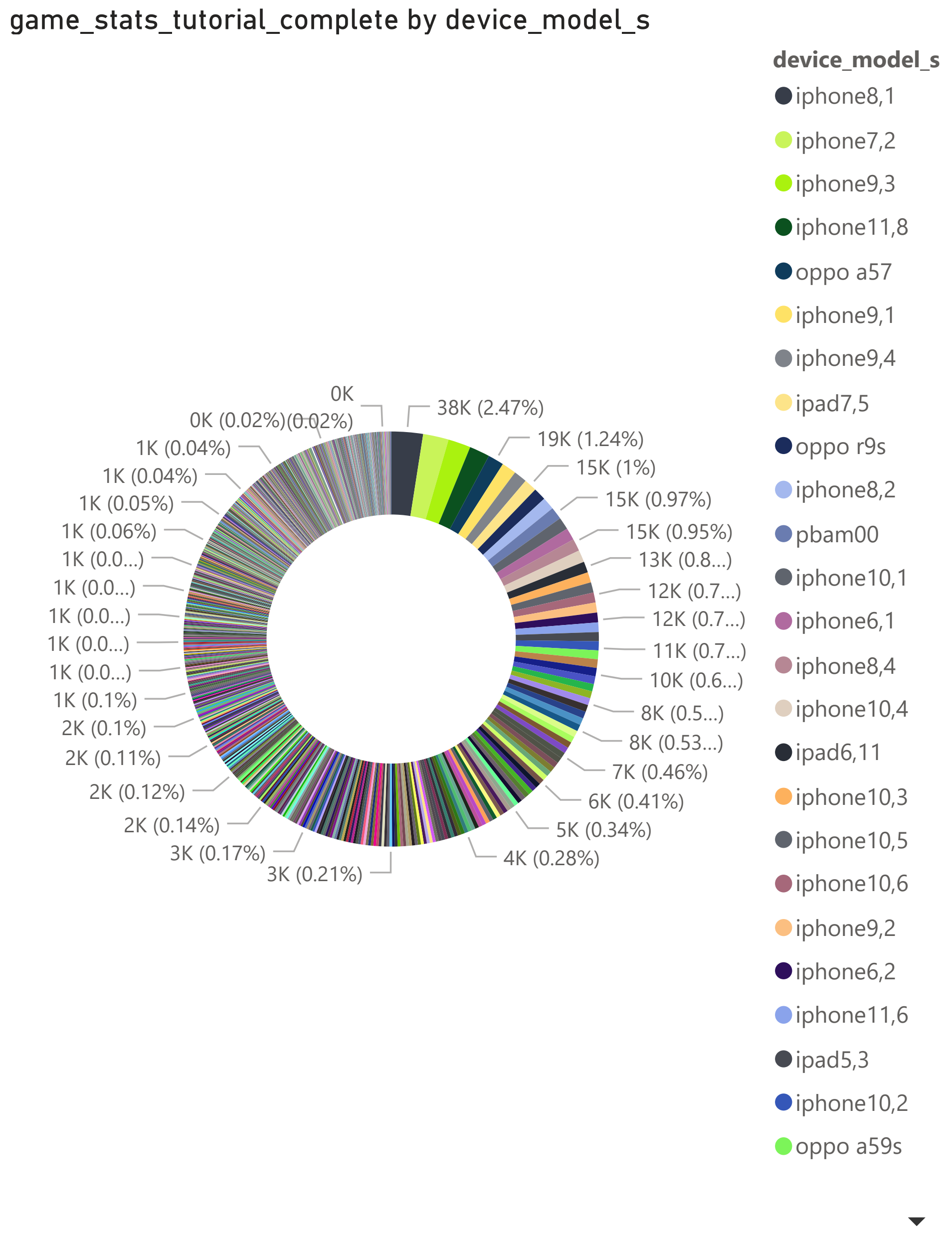


game_stats_tutorial_complete by platform_s



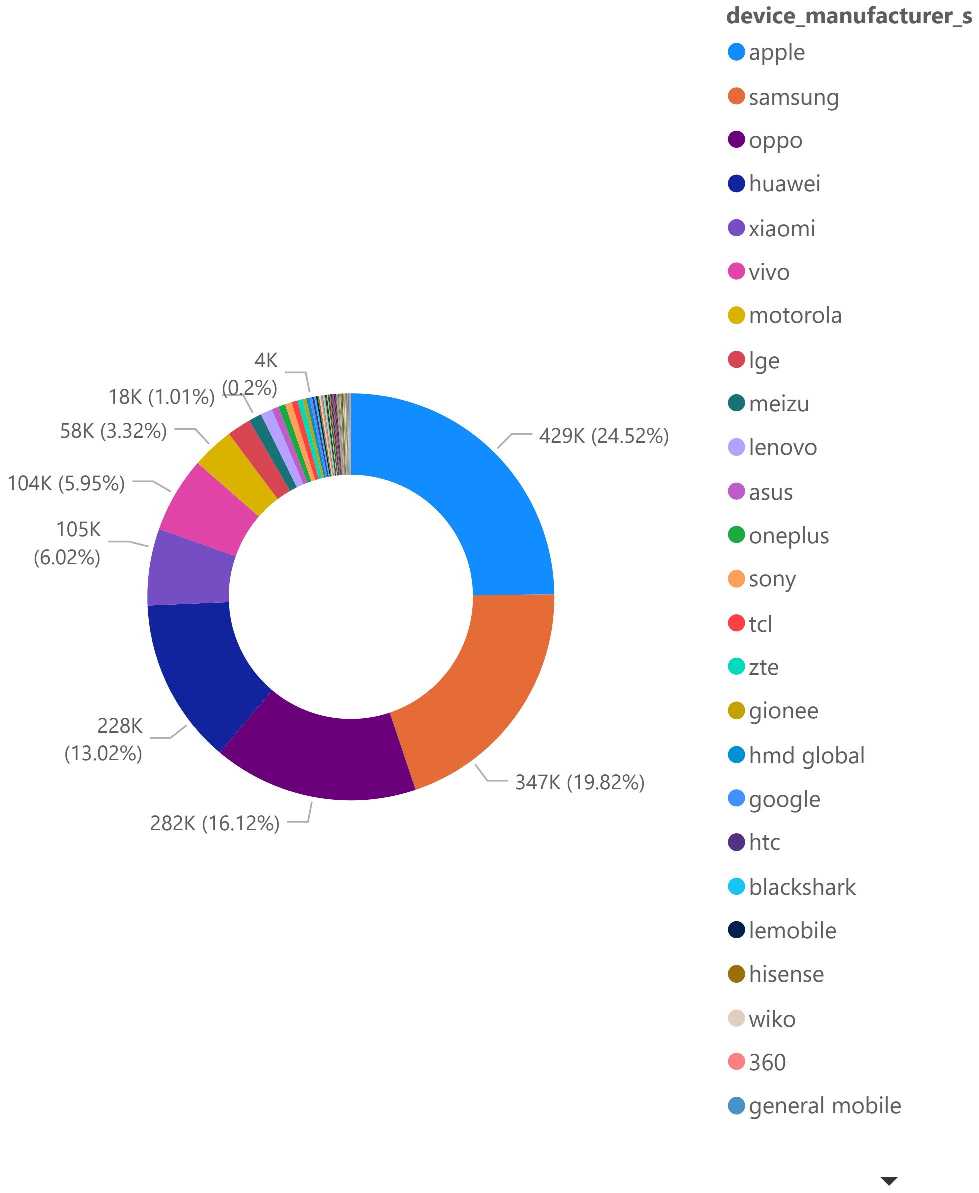
Android is the most popular among people who spend or complete tutorial.


total_spend by device_model_s

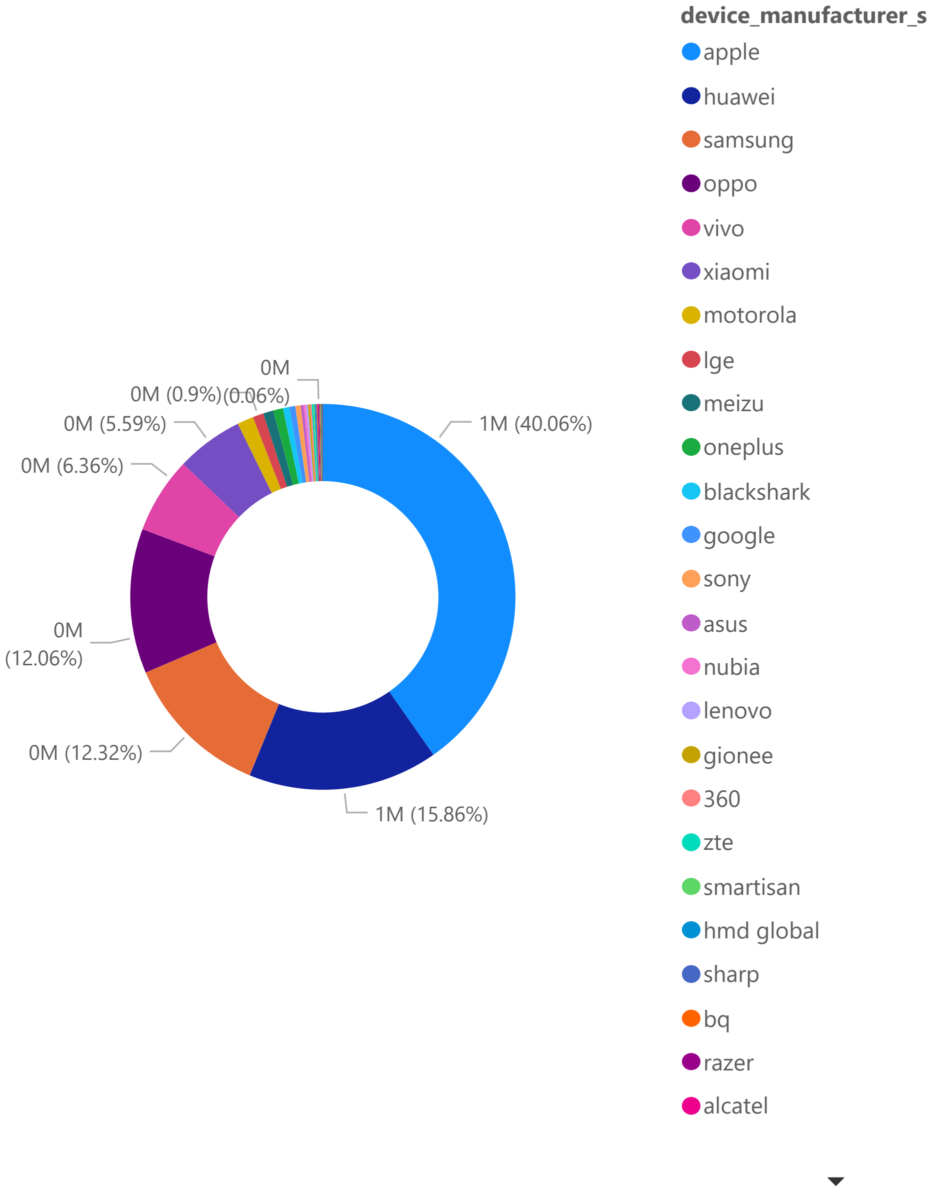

game_stats_tutorial_complete by device_model_s

Iphone users are most likely to spend and finish tutorial if we look only at device model?

game_stats_tutorial_complete by device_manufacturer_s

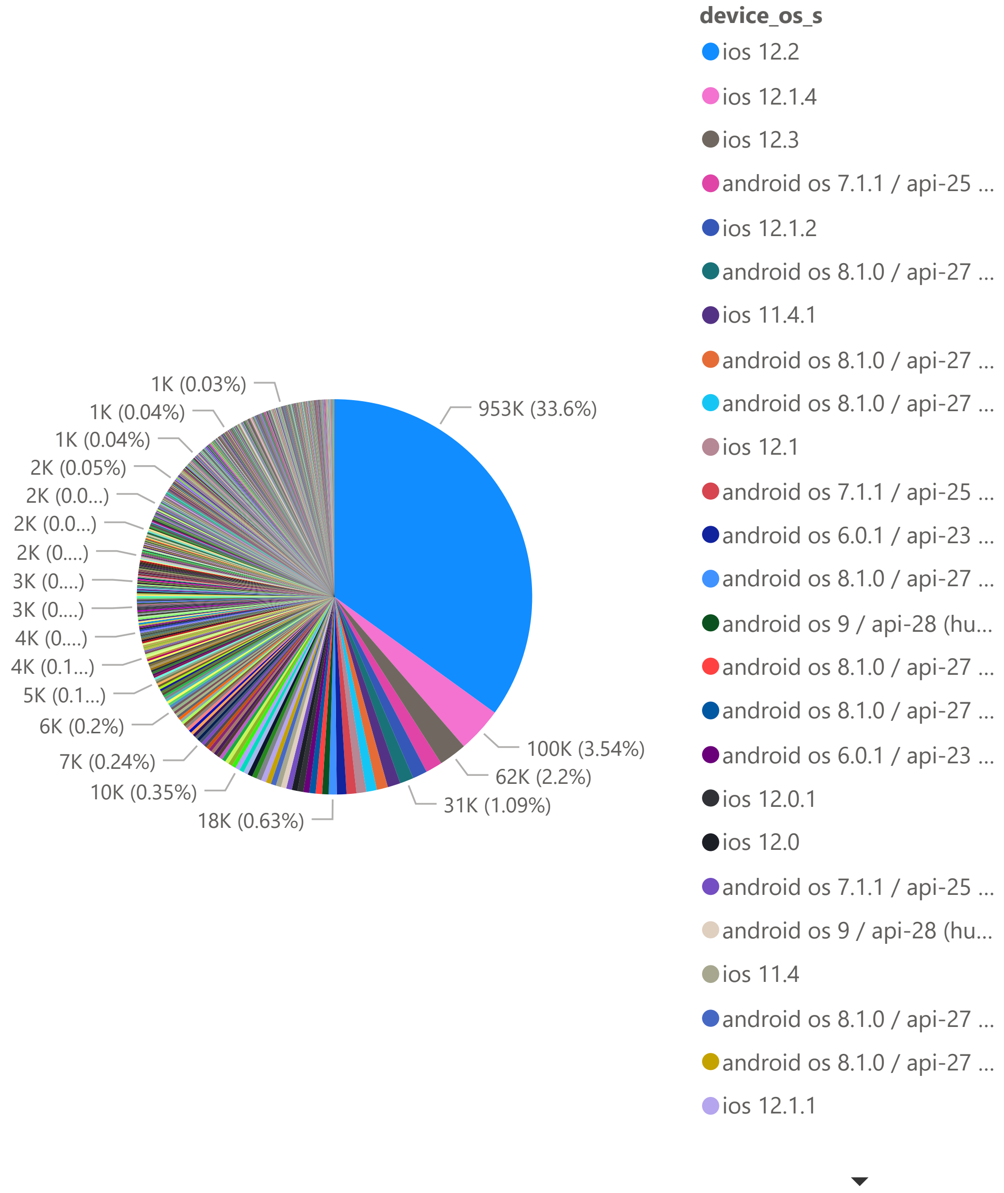


total_spend by device_manufacturer_s

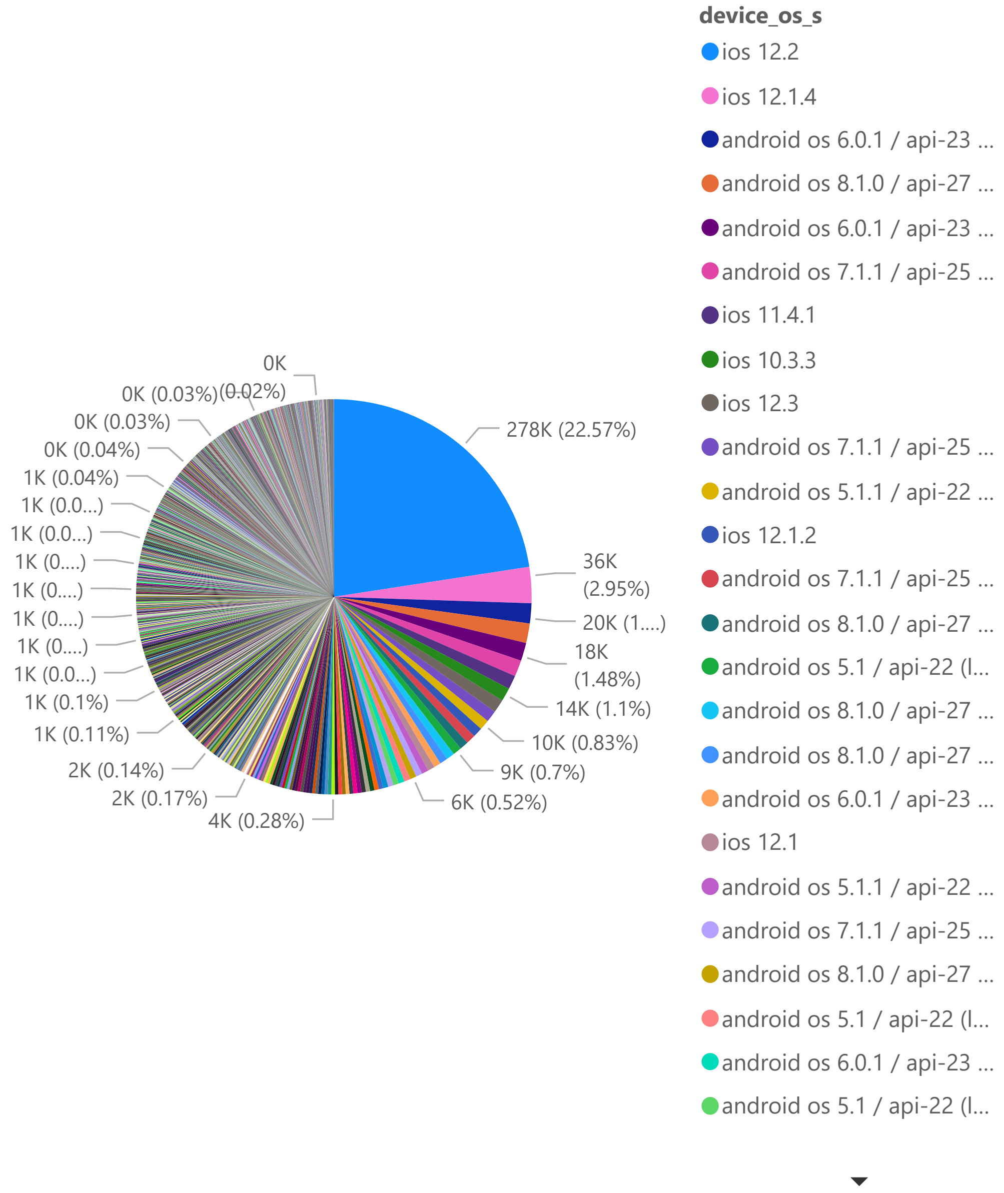


Again apple users are mostly to spend. Can we targeted them more? However, most of other phones use only android.

total_spend by device_os_s

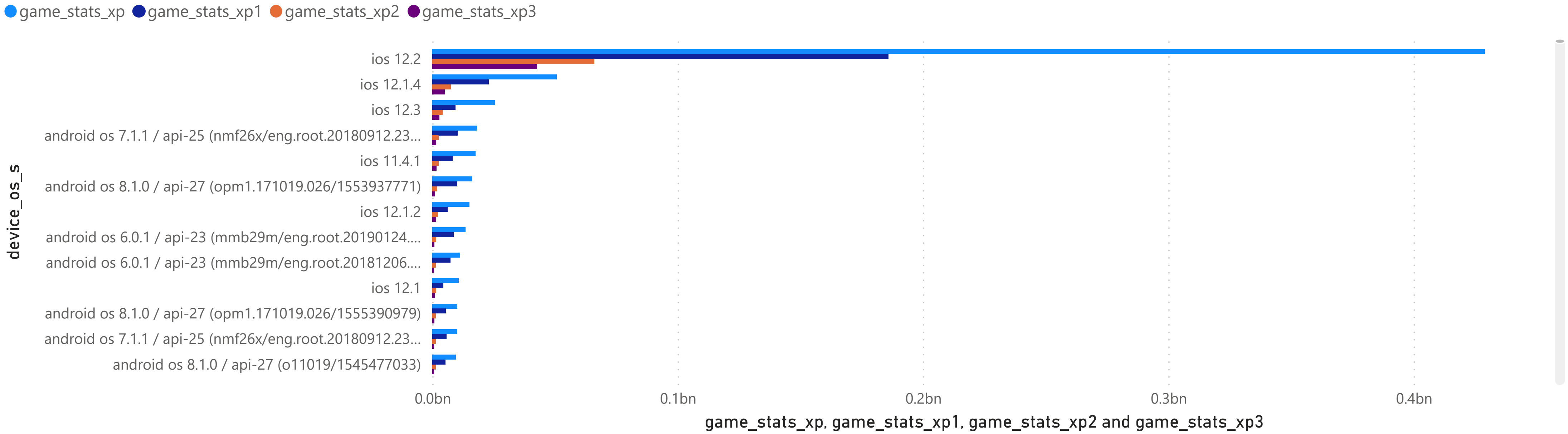


game_stats_tutorial_complete by device_os_s

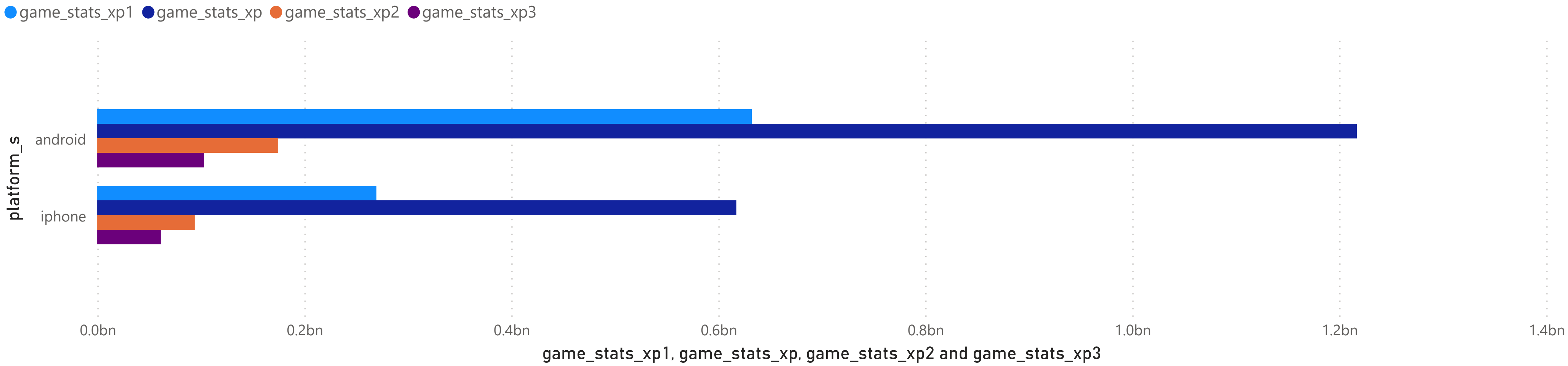


ios 12.2 the most popular for users who spend or complete the tutorial.

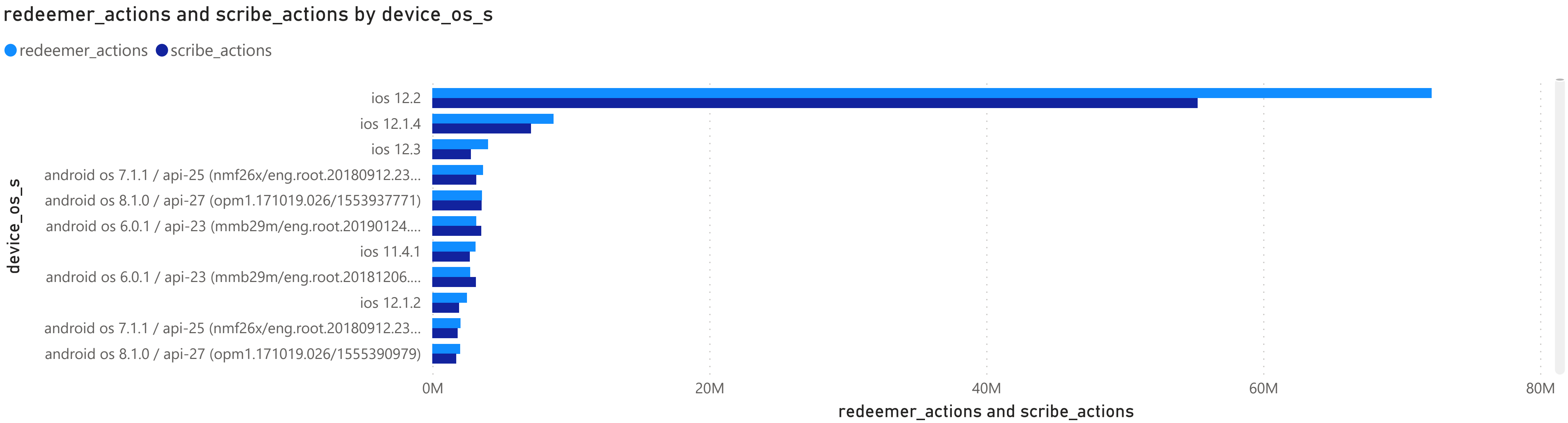
game_stats_xp, game_stats_xp1, game_stats_xp2 and game_stats_xp3 by device_os_s



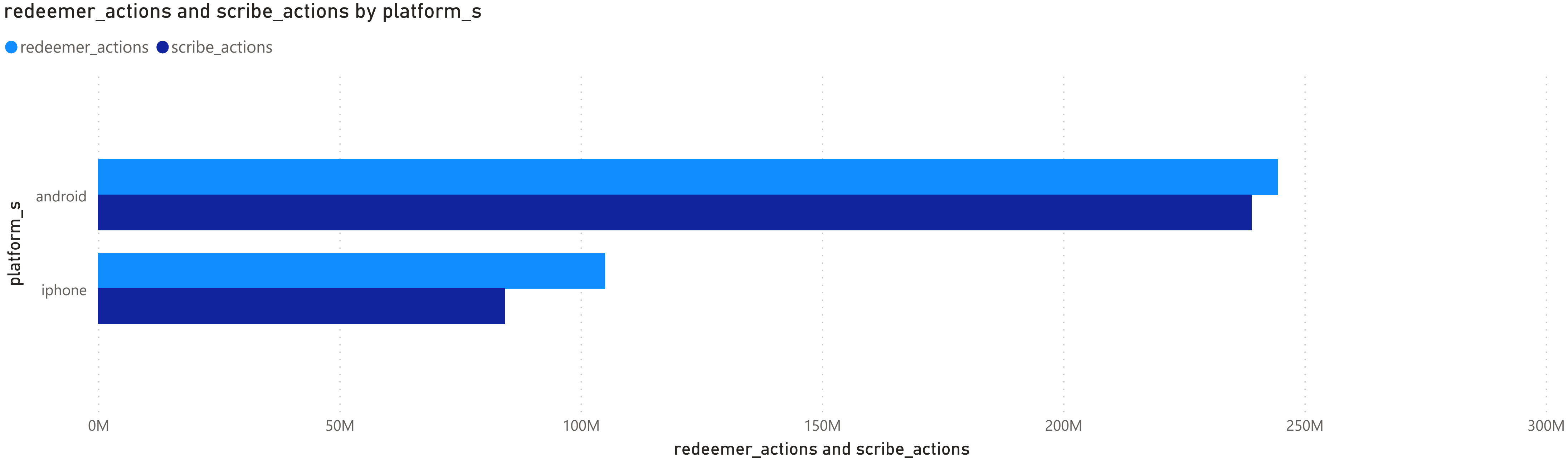
game_stats_xp1, game_stats_xp, game_stats_xp2 and game_stats_xp3 by platform_s



For game stats experiences, ios 12.2 again is by far the most used. However, compare to with platform android is far more used. It is interesting to see game_stats_xp has the most points in both graphs.



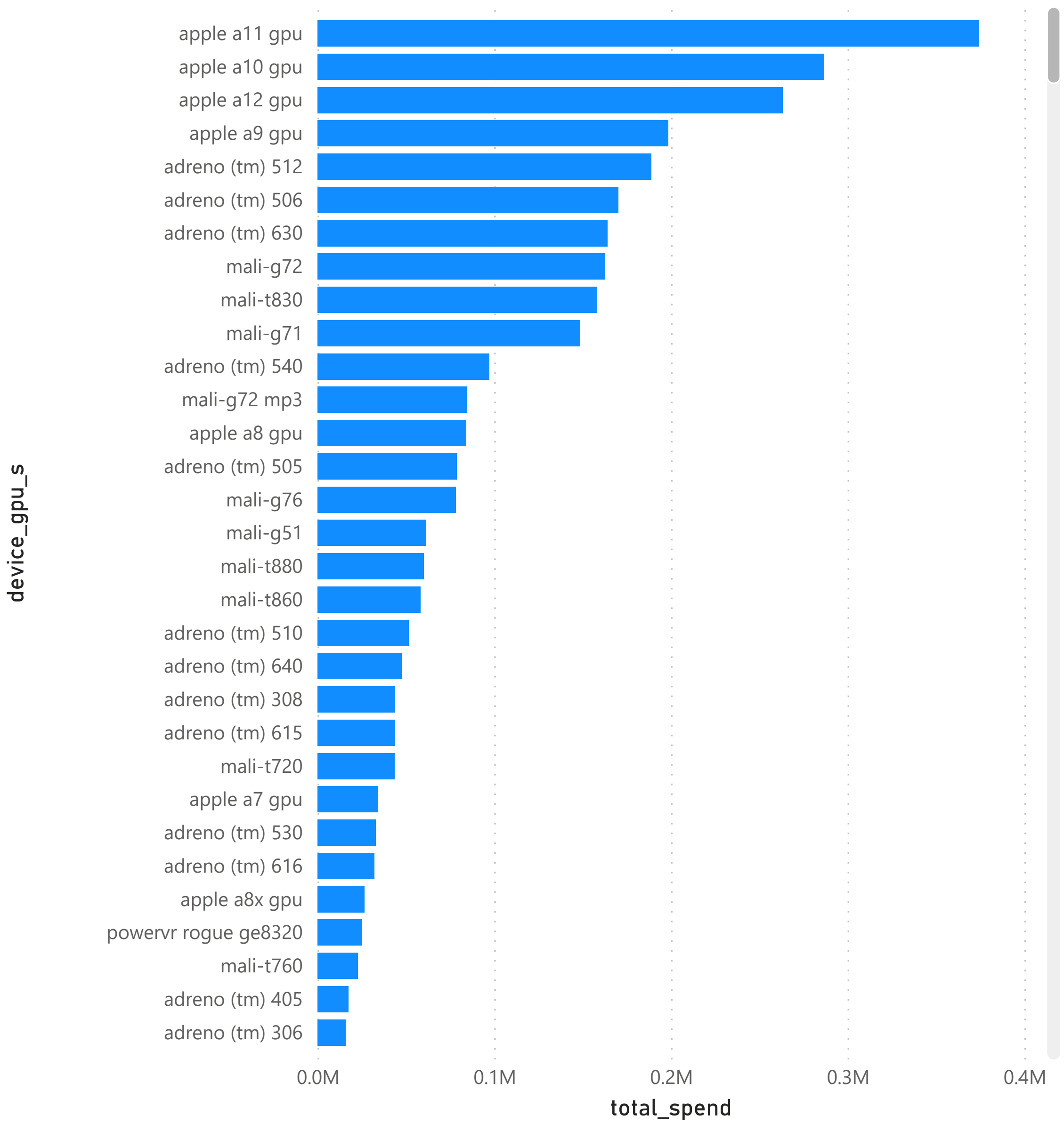
redeemer_actions and scribe_actions by device_os_s



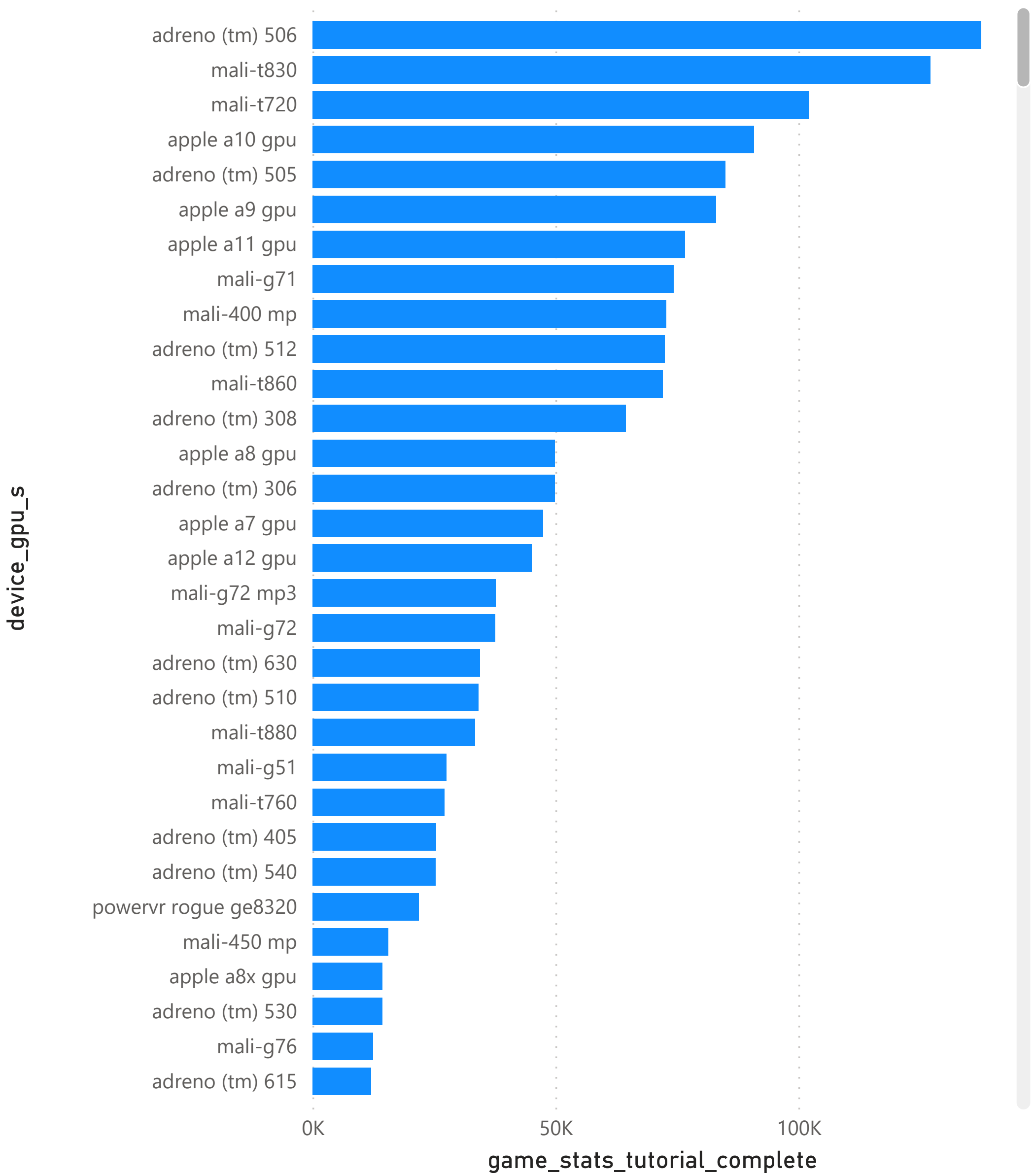
redeemer_actions and scribe_actions by platform_s

ios 12.2 far more used for redeemer actions. However, comparing with platform again android is used far more.

total_spend by device_gpu_s



game_stats_tutorial_complete by device_gpu_s



The Gpu can be investigated too. Better GPU, most likely to play game?