Predicting Future Sales

Eric Huynh

April 2 Springboard Cohort

Problem & Objective

- Poorly stocked retail stores ruin the customer experience and satisfaction.
- Poor customer experience reduces customer loyalty which in turn reduces repeat business.
  - Also drives consumers to online shopping where it is much more difficult to keep customers from churning to larger corporations that can sell merchandise at much lower prices.
- Objective:
  - Create a model that can reliably predict the future sales of a product at a shop for up to a month.

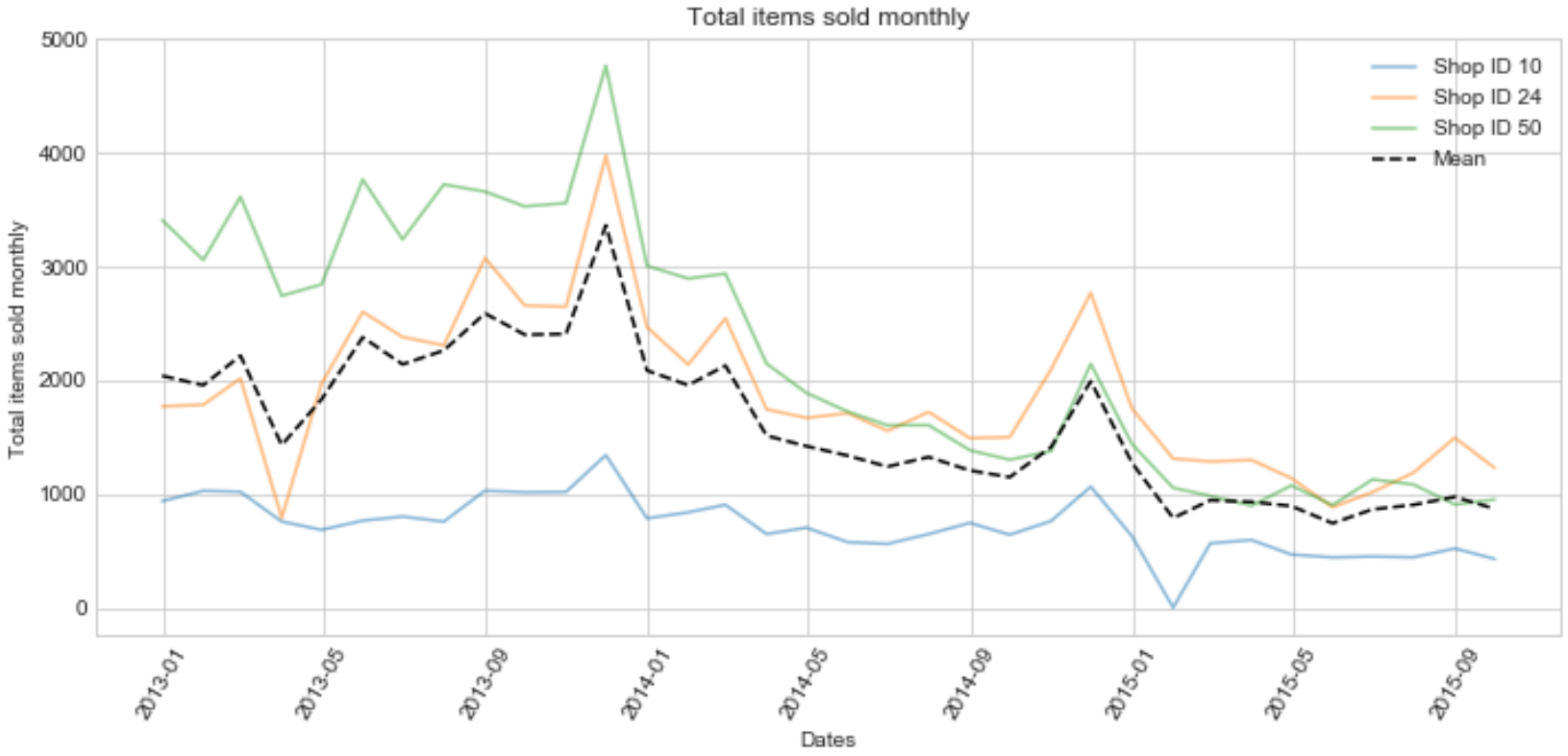
Potential Clients

- Individual Retail Stores:
  - Helps predict which products are still desired and capable of creating profitable sales
  - Keeps track of the efflux of each product to ensure properly stocked inventory
- Corporate Retail Stores:
  - Same applications as individual retail stores, but at higher magnitudes
  - Properly stocked retail stores are necessary to building corporation reliability
- Production Companies:
  - Create a priority list to focus manufacturing for generating higher sale values
  - If expanded with geographical information, companies can focus shipments to particular areas with better sale rates

Data from Kaggle Competition

- Training Data:
  - Consists of daily entries from Jan. 1st 2013 to Oct. 31st 2015.
- Supplemental Datasets:
  - “items.csv” – contains information on each of the 22,170 unique items
  - “item_categories.csv” – contains information for 84 unique item categories
  - “shops.csv” – contains information for 60 unique shops
- Test Data:
  - Consists of two categorical columns: shop_id and item_id

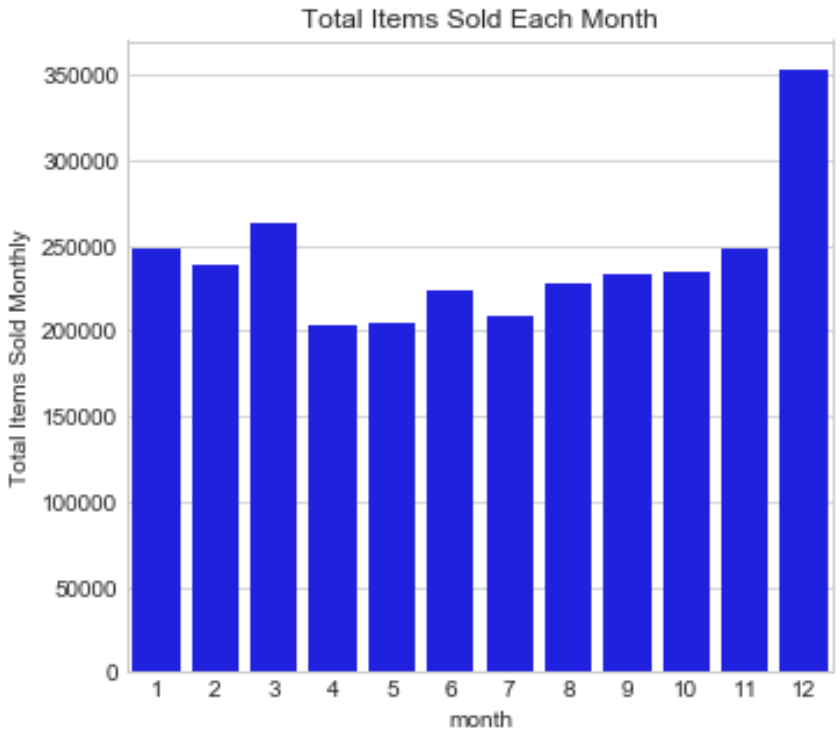
Time-Series

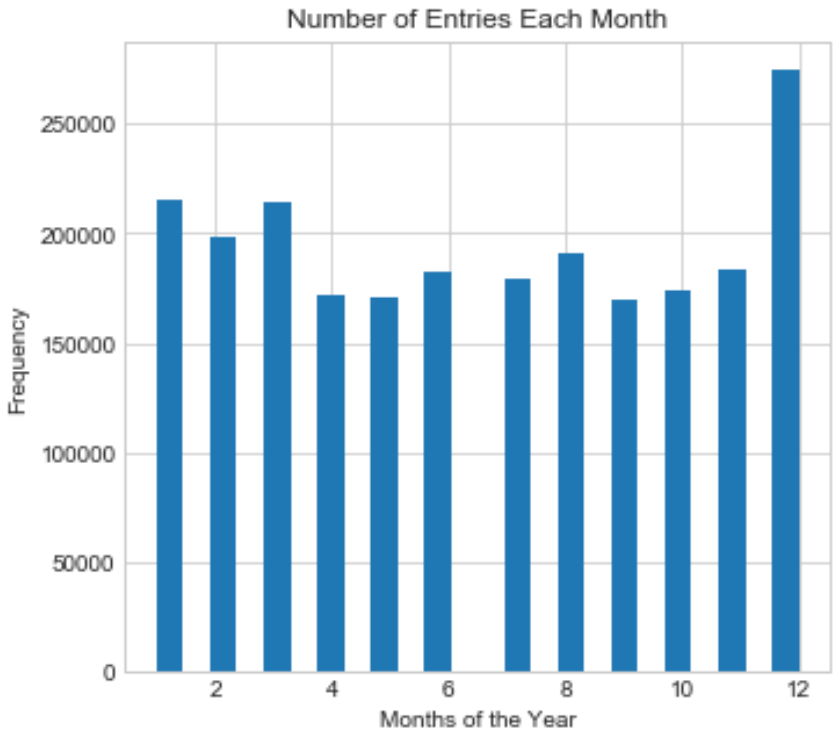


Key Points:

- Apparent **decline** of sales over time.
- **Reoccuring peak** in December of 2013 and 2014.
- May be due to seasonal promotion around that time (potentially Christmas and New Years)

Seasonal Trend



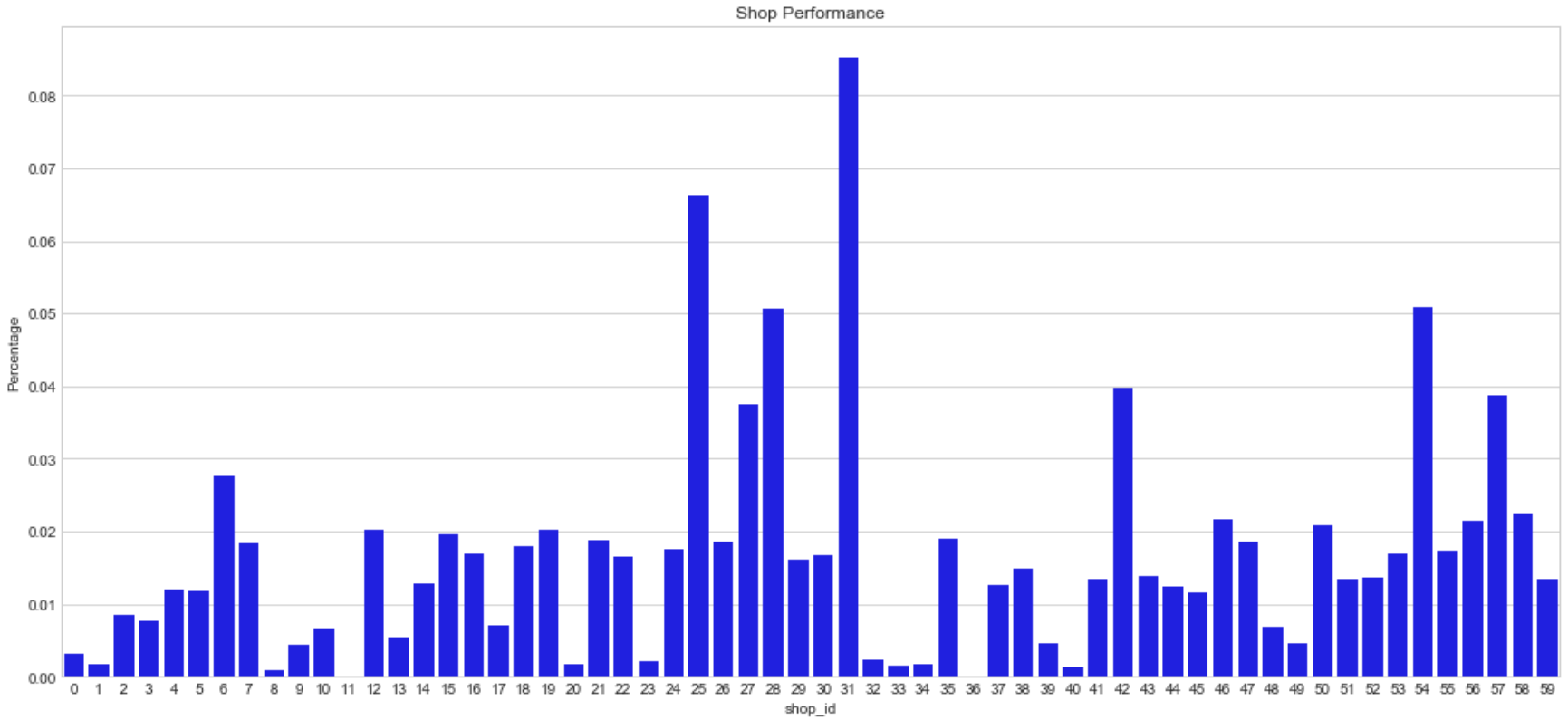


NOTE: Data from 2013 to 2014 only

Key Points:

- **Increased** total sales and total entries during December.
- Increase is reflective of data seen in time-series

Shop Performance

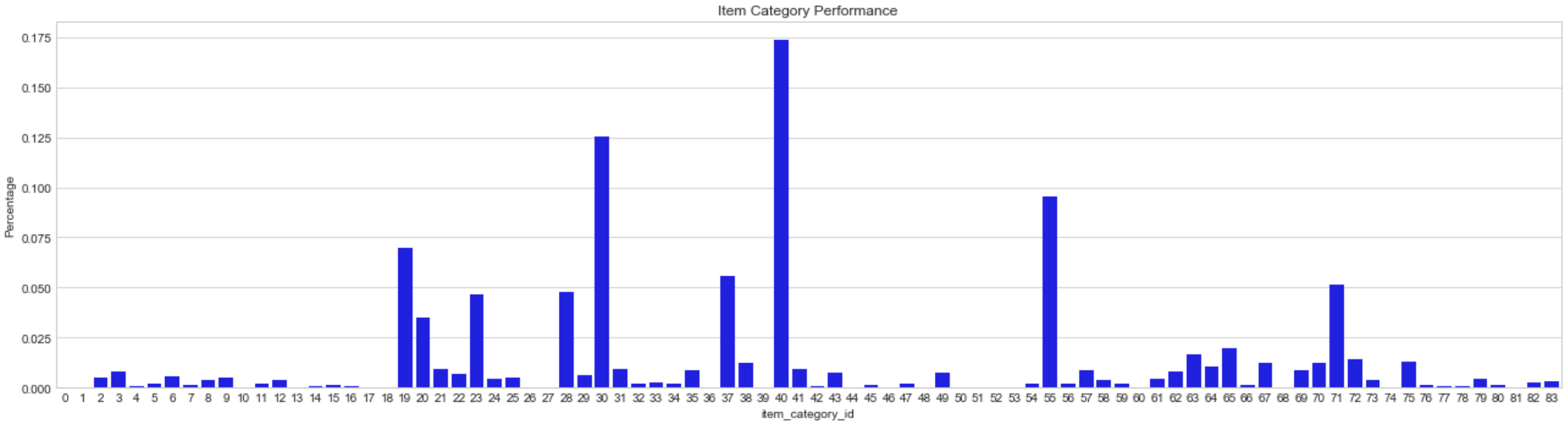


Key Points:

- Many shops performed at the **same level**
- Small number of shops performed significantly better than the rest.

- The shops showing similar levels of performance will be grouped

Item Category Performance

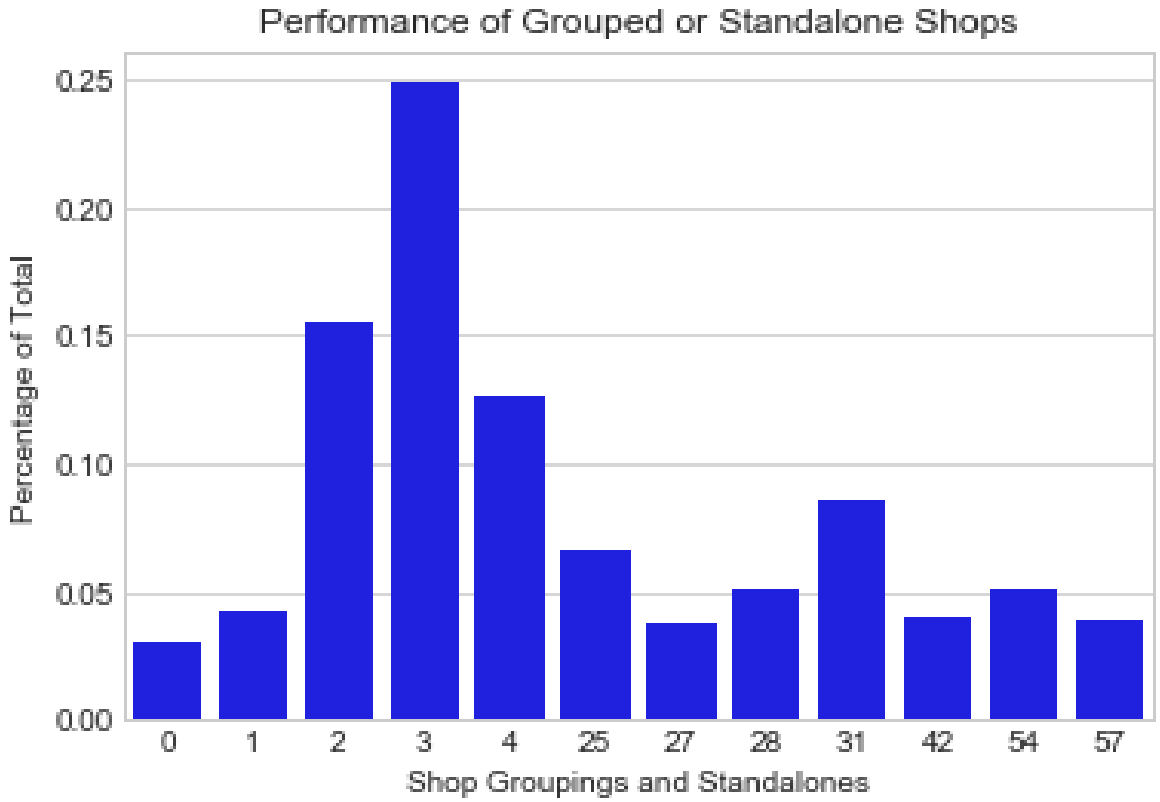


Key Points: (similar to shop performance)

- Many item categories performed at the **same level**
- Small number of item categories performed significantly better than the rest.

- Similarly performing item categories will be grouped

Shop Groupings



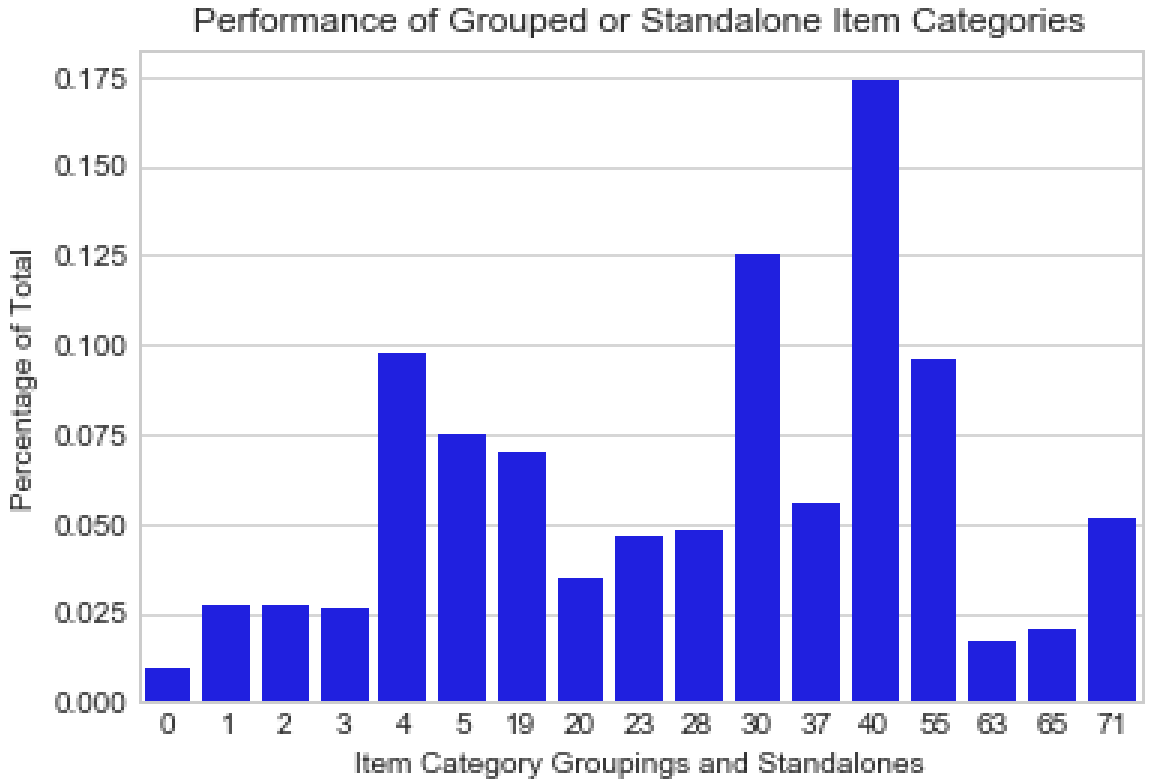
- **60** unique shop IDs -> **5 grouped** shop IDs + **8 standalone** shop IDs

Group 1: 14 shops performing at 0% - 0.5%
Group 2: 6 shops performing from 0.5% - 1%
Group 3: 12 shops performing from 1% - 1.5%
Group 4: 14 shops performing from 1.5% - 2%
Group 5: 6 shops performing from 2% - 2.5%

Statistical Validation: (t-test between the closet performing groups; p-value < 0.05)

- Weakest t-score: -6.7
- Weakest p-value: $1.1 * 10^{-6}$

Item Category Groupings



- **84** unique category IDs -> **6 grouped** category IDs + **11 standalone** category IDs

Group 1: 29 categories performing at 0% - 0.15%
Group 2: 14 categories performing from 0.15% - 0.3%
Group 3: 7 categories performing from 0.3% - 0.45%
Group 4: 5 categories performing from 0.45% - 0.6%
Group 5: 12 categories performing from 0.6% - 1%
Group 6: 6 categories performing from 1% - 1.5%

Statistical Validation: (t-test between the closet performing groups; p-value < 0.05)

- Weakest t-score: -5.97
- Weakest p-value: $2.57 * 10^{-5}$