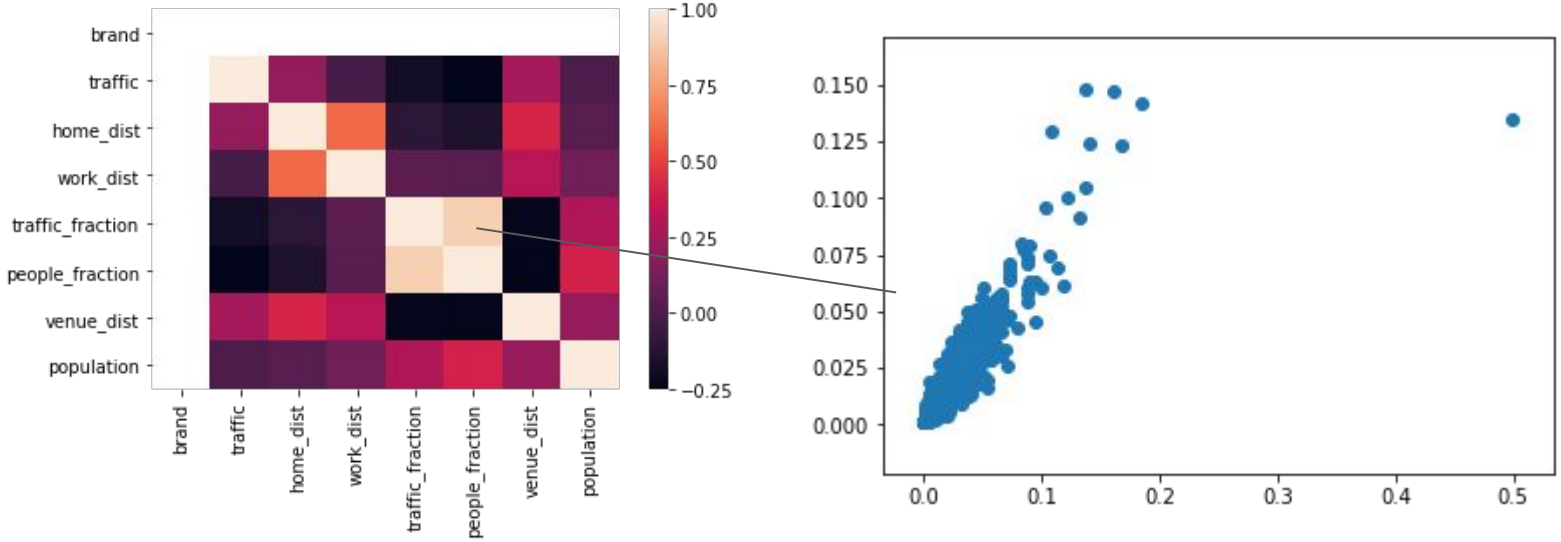
What we’re interested in

- Correlation between grocery shopping trends and interest in e-commerce, i.e. “e-commerce penetration”
- Correlation between e-commerce penetration and popularity of other forms of technology, i.e “tech penetration”
- What could it mean in regions where tech penetration is high, but e-commerce penetration is low?
- Ultimate goal: find some practical & novel insight for e-commerce companies to identify key areas of opportunity

Methodology - what got us thinking

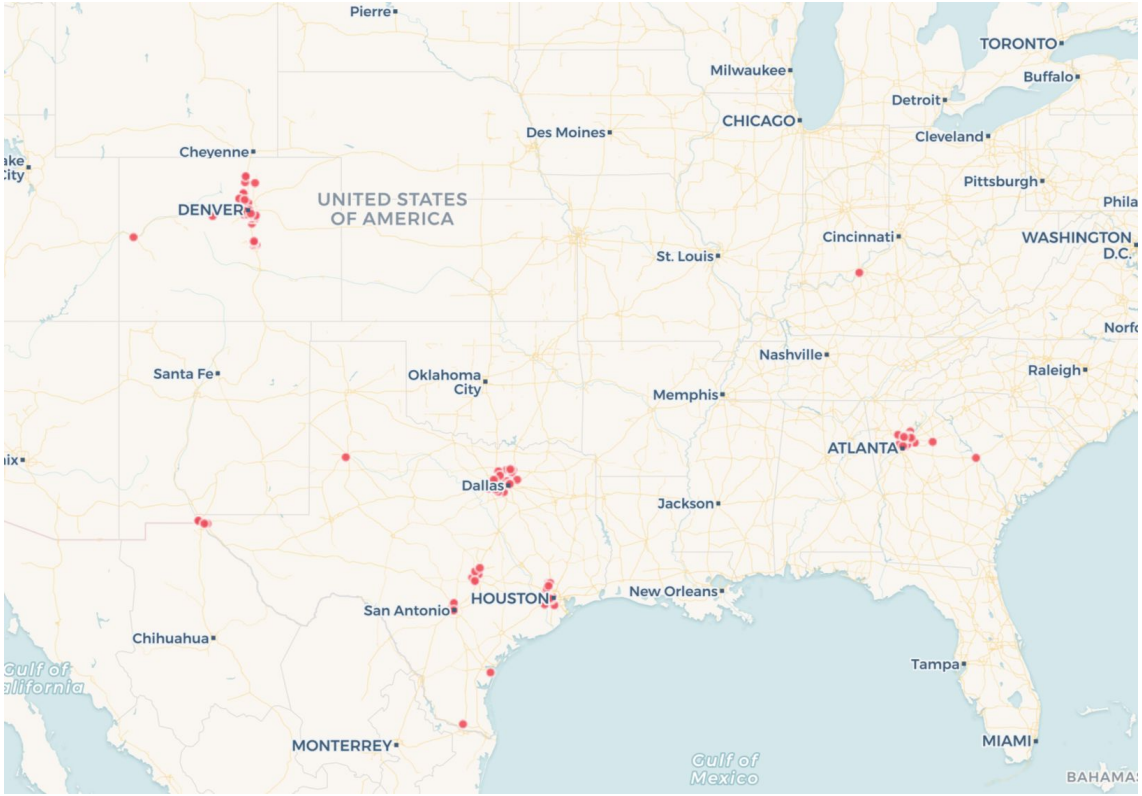
-There's a high correlation between traffic_fraction and people_fraction (to be expected)--but what could it mean that it's *this* high? It appears that there are certain regions with very few grocery shoppers; what could they be up to?



What got us thinking (2)

-Data is highly clustered; so it would make sense to explore trends by cities

-Moreover: little to no clustering of brands within cities



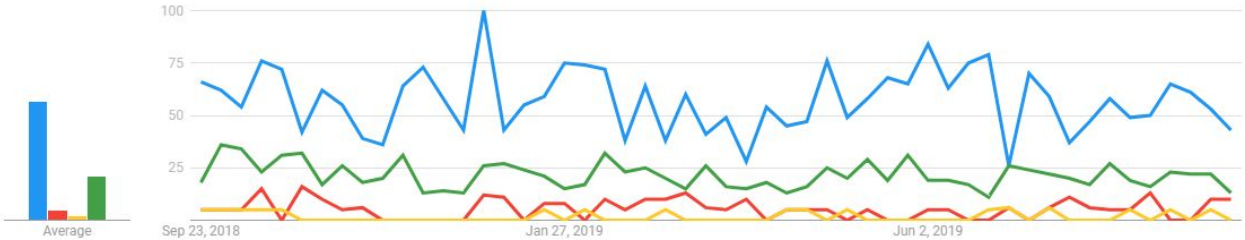
Methodology - alternatives to groceries (e-commerce)

-We decided to look for correlations between frequency of grocery shopping and interest in e-commerce by Census tract region

-Data on e-commerce is limited (Amazon and Instacart are particularly sneaky); so we got creative

-Collected Google trends search data for 8 most popular cities to gauge interest in e-commerce brands

Google Trends interest over time for Instacart, Amazon fresh, Peapod, and Shipt in Austin, TX (over the past year)

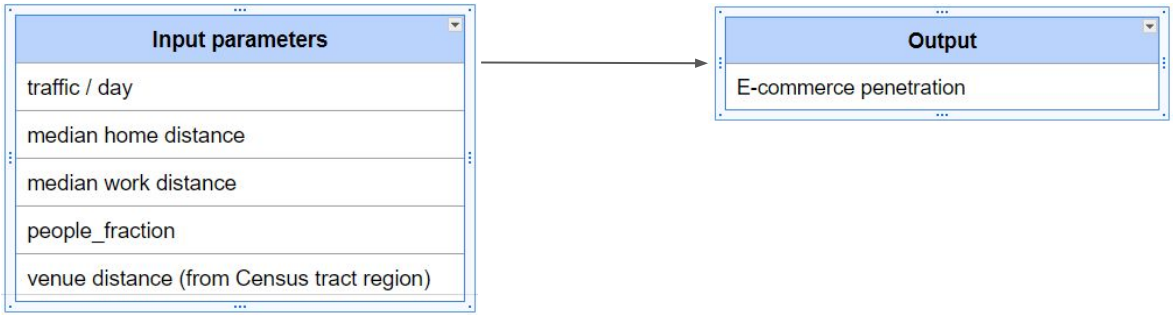


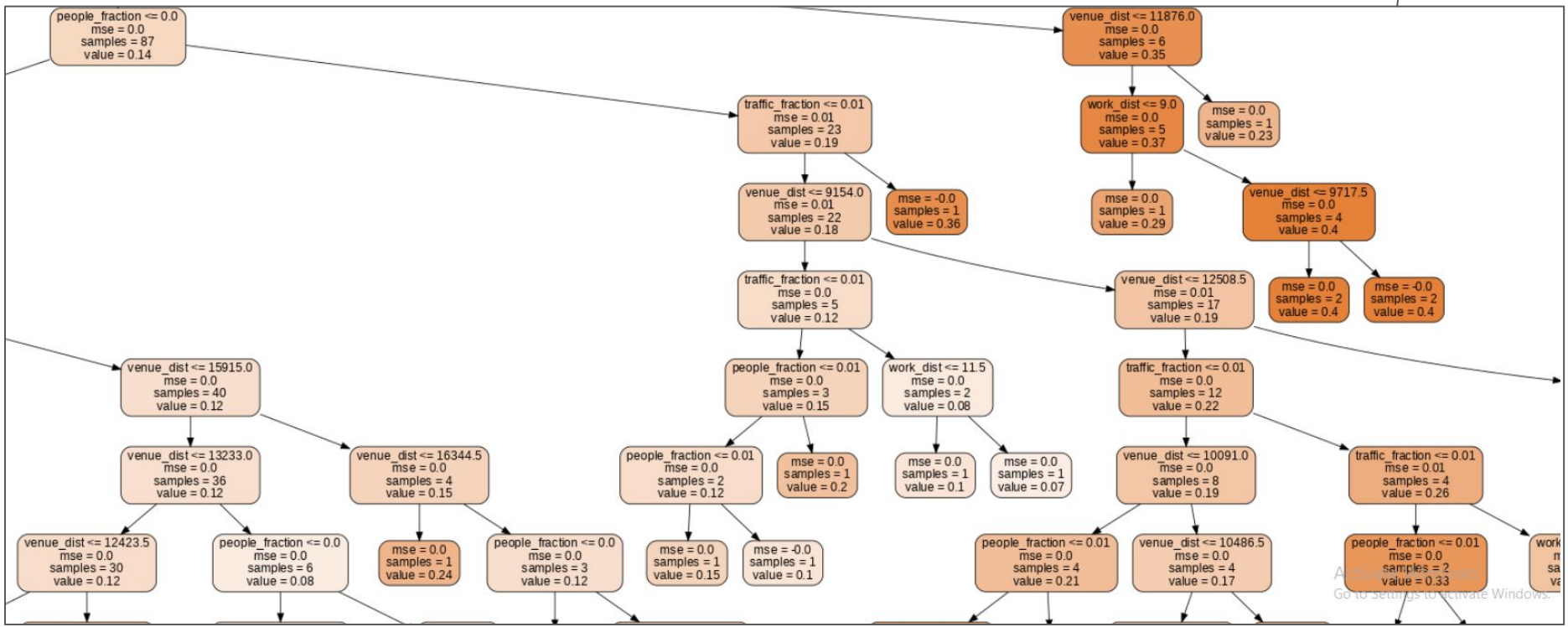
Methodology - finding “e-commerce penetration”

-We tried several models (decision tree regressor, random forest regressor, multi-layer perceptron) to determine the relationship between the factors in the grocery dataset and the relative interest in e-commerce in each region

-Each data point is a unique (venue_id, fips_code) pair; ~few hundred to 1000 / city

-Goal: given data on grocery shopping behavior & trends of a particular region, estimate the relative interest in e-commerce within that region




people_fraction <= 0.0
mse = 0.0
samples = 87
value = 0.14
venue_dist <= 11876.0
mse = 0.0
samples = 6
value = 0.35
traffic_fraction <= 0.01
mse = 0.01
samples = 23
value = 0.19
work_dist <= 9.0
mse = 0.0
samples = 5
value = 0.37
mse = 0.0
samples = 1
value = 0.23
venue_dist <= 9154.0
mse = 0.01
samples = 22
value = 0.18
mse = -0.0
samples = 1
value = 0.36
mse = 0.0
samples = 1
value = 0.29
venue_dist <= 9717.5
mse = 0.0
samples = 4
value = 0.4
traffic_fraction <= 0.01
mse = 0.0
samples = 5
value = 0.12
venue_dist <= 12508.5
mse = 0.01
samples = 17
value = 0.19
mse = 0.0
samples = 2
value = 0.4
mse = -0.0
samples = 2
value = 0.4
venue_dist <= 15915.0
mse = 0.0
samples = 40
value = 0.12
people_fraction <= 0.01
mse = 0.0
samples = 3
value = 0.15
work_dist <= 11.5
mse = 0.0
samples = 2
value = 0.08
traffic_fraction <= 0.01
mse = 0.0
samples = 12
value = 0.22
venue_dist <= 13233.0
mse = 0.0
samples = 36
value = 0.12
venue_dist <= 16344.5
mse = 0.0
samples = 4
value = 0.15
people_fraction <= 0.01
mse = 0.0
samples = 2
value = 0.12
mse = 0.0
samples = 1
value = 0.2
mse = 0.0
samples = 1
value = 0.1
mse = 0.0
samples = 1
value = 0.07
venue_dist <= 10091.0
mse = 0.0
samples = 8
value = 0.19
traffic_fraction <= 0.01
mse = 0.01
samples = 4
value = 0.26
venue_dist <= 12423.5
mse = 0.0
samples = 30
value = 0.12
people_fraction <= 0.0
mse = 0.0
samples = 6
value = 0.08
mse = 0.0
samples = 1
value = 0.24
people_fraction <= 0.0
mse = 0.0
samples = 3
value = 0.12
mse = 0.0
samples = 1
value = 0.15
mse = -0.0
samples = 1
value = 0.1
people_fraction <= 0.01
mse = 0.0
samples = 4
value = 0.21
venue_dist <= 10486.5
mse = 0.0
samples = 4
value = 0.17
people_fraction <= 0.01
mse = 0.0
samples = 2
value = 0.33

Methodology - the other side of the coin

How can we make use of this data?

-It’s worthwhile to find areas with e-commerce penetration that differs from our expectations; but first, we need to have some expectations!

-“Tech penetration”: the relative popularity/adoption of technology trends (that are similar to e-commerce) by region

    -e.g. social media & internet usage, computer/cell phone ownership

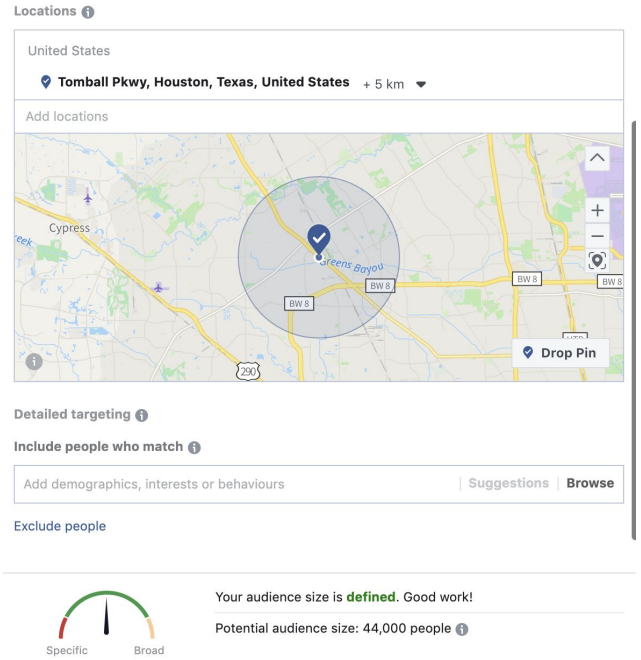
-We expect areas with the infrastructure and culture of tech in place, to be the ones adopting the e-commerce trend

Methodology - determining tech penetration

-Facebook data on usage by location: set up an advertising page & created ads to get estimates of facebook user numbers within store vicinities

-Manually collected this data for 57 selected stores to obtain an idea of tech penetration in surrounding areas, where we also had e-commerce penetration data (Google trends)

-Census statistics on internet access & computer ownership



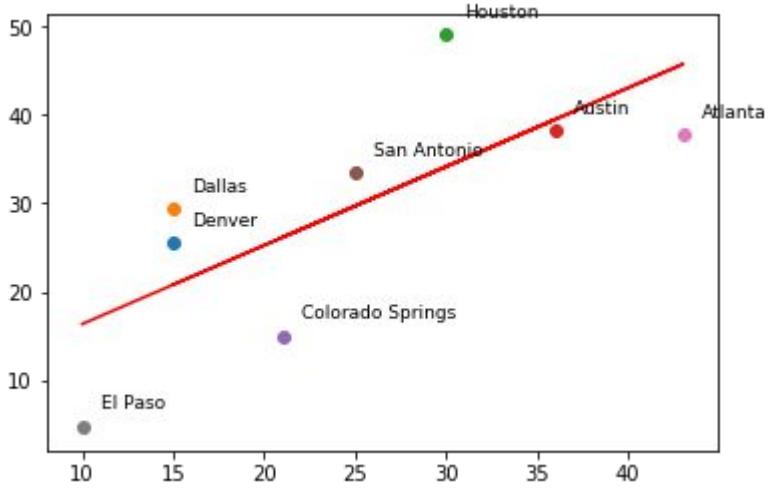
Methodology - finally, the grand finale

-We regress e-commerce penetration with tech penetration to spot any outliers

-Areas with e-commerce *lower* than expected by tech penetration could be targets for rapid e-commerce growth potential! (as the infrastructure & culture is already in place)

-First-order sanity checks--clear linear correlation between x & y; bigger cities have higher tech penetration

-Cities of interest: Dallas, Houston, Denver (so get on it, e-commerce companies!)



Data sources

Google trends: https://trends.google.com/trends/?geo=US

Facebook public data: https://www.facebook.com/business/tools/ads-manager

Census stats on computer ownership/internet access:
https://www.census.gov/data/tables/2012/demo/computer-internet/computer-use-2012.html