DAVID EY

DATA ANALYTICS PORTFOLIO

PROJECTS

GameCo

Analyze global video game sales to plan new game development

Influenza Preparation

Use historical flu & census data to allocate medical staff

Rockbuster Stealth

Determine movie acquisition strategy for a new streaming service

Instacart

Initial data & exploratory analysis for insights and better segmentation

Pig E. Bank

Assess client & transaction risk, build and optimize models to assist a compliance program

Airbnb Berlin: Pricing Factors & Effects on Local Market

Determine price and rating factors, effects on local markets, impact of recent regulation

GAMECO: VIDEO GAMES SALES ANALYSIS

BACKGROUND

GOAL

Perform a descriptive analysis of a video game sales data set to foster a better understanding of how GameCo's new games might fare in the market.

TOOLS

Excel
PowerPoint

DATASET

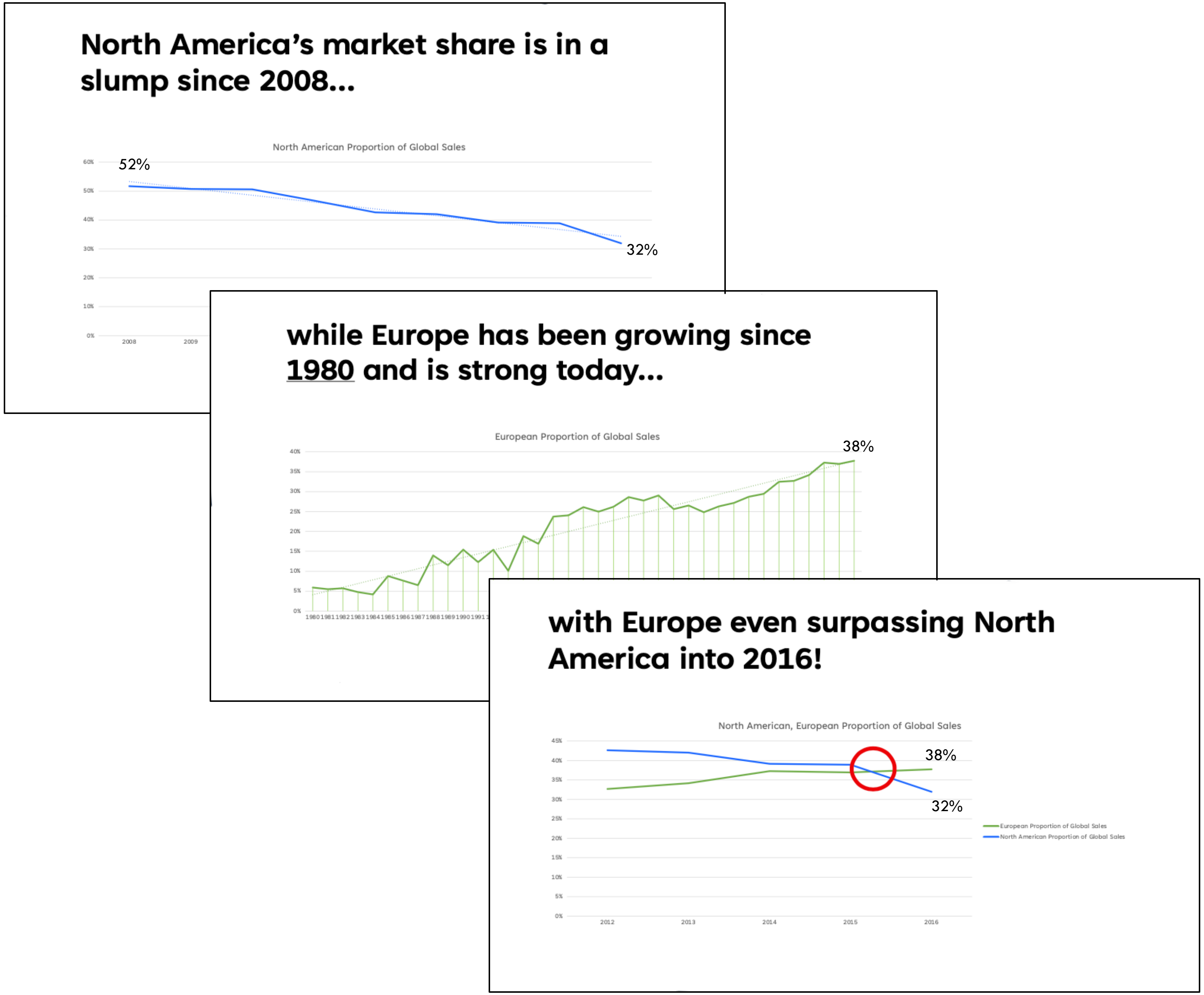
[VGChartz video game sales data (1980-2016)]

SKILLS USED

Grouping data
Summarizing data
Descriptive analysis
Visualizing results
Presenting results

GAMECO: SALES SECTION

In market share, Europe is the primary market for GameCo's investments, growing consistently since 1980 and is now the largest (at **38%**). This investment should come at the expense of North America, on a consistent downtrend since 2008 (**52%** then, **32%** now).



North America's market share is in a slump since 2008...

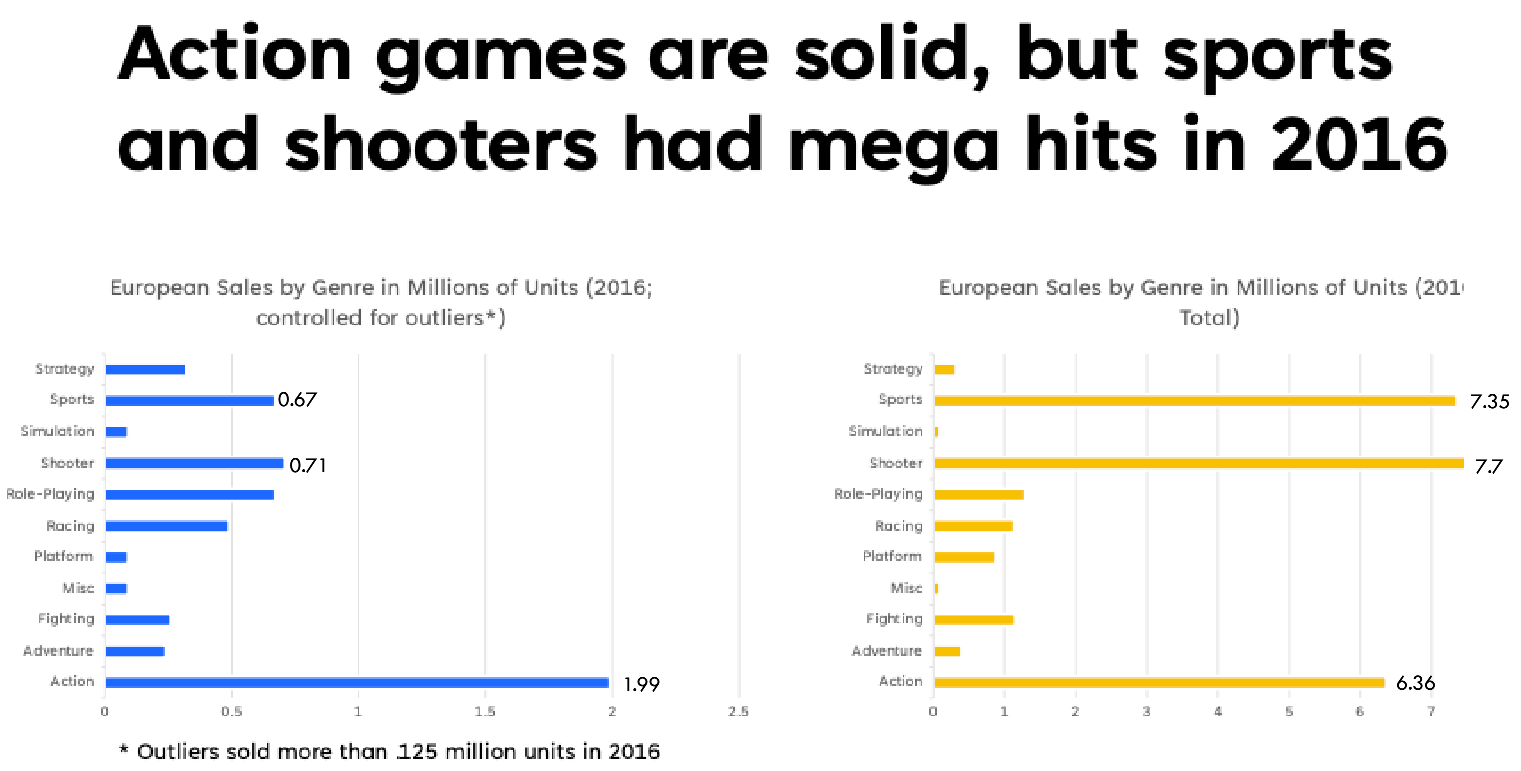
while Europe has been growing since 1980 and is strong today...

with Europe even surpassing North America into 2016!

GAMECO: MARKETING SECTION

In Europe, as the most important market, publish Action (1.99m units sold in 2016), Shooters (0.71m) and Sports (0.67m) games. These are the strongest categories even when excluding statistical outliers.

GameCo should evaluate its capability to compete with major franchises in these genres



Action games are solid, but sports and shooters had mega hits in 2016

European Sales by Genre in Millions of Units (2016; controlled for outliers*)

European Sales by Genre in Millions of Units (201
Total)

* Outliers sold more than .125 million units in 2016

PROJECT LINKS

PROJECT BRIEF

FULL PRESENTATION (PDF)

FULL PRESENTATION (PPTX)

PROJECT REFLECTIONS

INFLUENZA PREPARATION: ALLOCATE MEDICAL STAFF FOR FLU SEASON

BACKGROUND

GOAL

Analyze temporal and geographic flu trends to plan for staffing needs across the US and lower the mortality rate

TOOLS

Excel
Tableau

DATASET

[US Census Data (2009-2017)]

[Flu-Related Death Counts (2009-2017)]

SKILLS USED

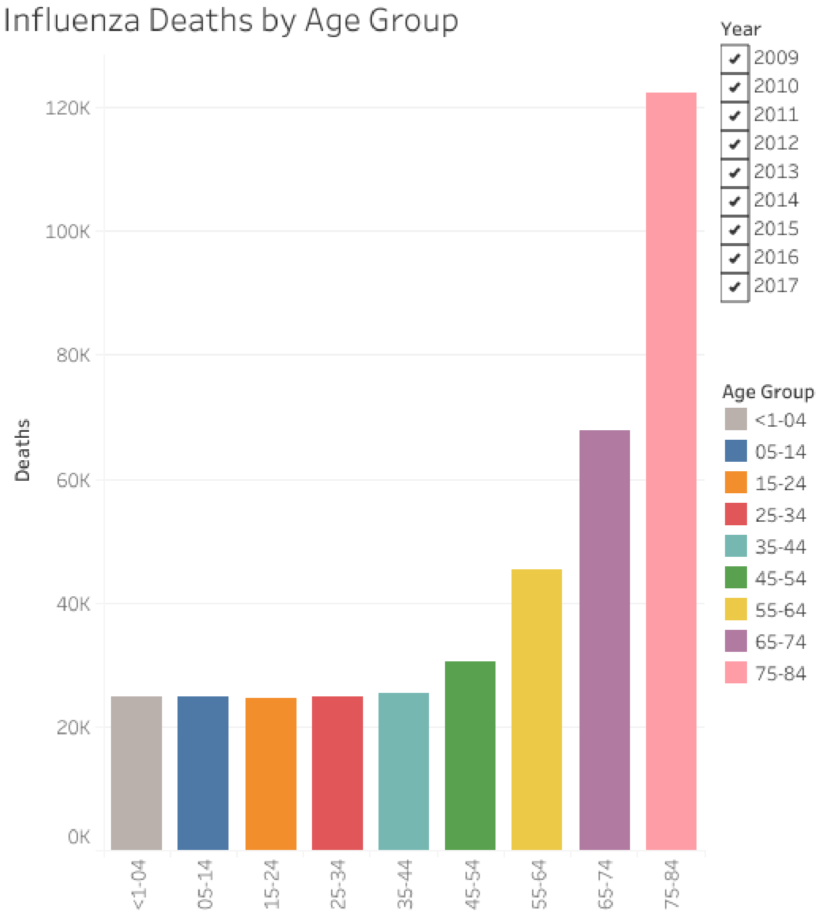
Translating business requirements
Data cleaning
Data integration
Data transformation
Statistical hypothesis testing
Visual analysis
Forecasting
Storytelling in Tableau
Presenting results to an audience

FLU SEASON: APPROACH

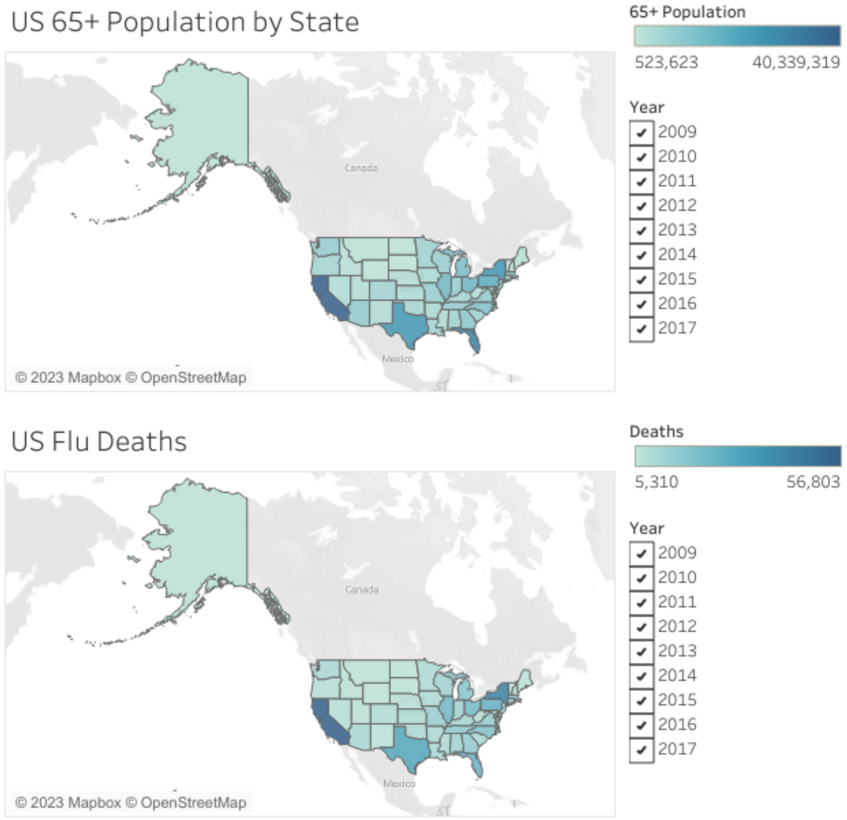
Approach:

Mortality rates from the flu were used to determine where staff would be most needed.

Exploratory analysis revealed that the higher age groups **(especially 65+)** were much more likely to die from influenza than younger groups **(.99 correlation between 65+ age and mortality)**


Influenza Deaths by Age Group

View in Tableau


US 65+ Population by State

US Flu Deaths

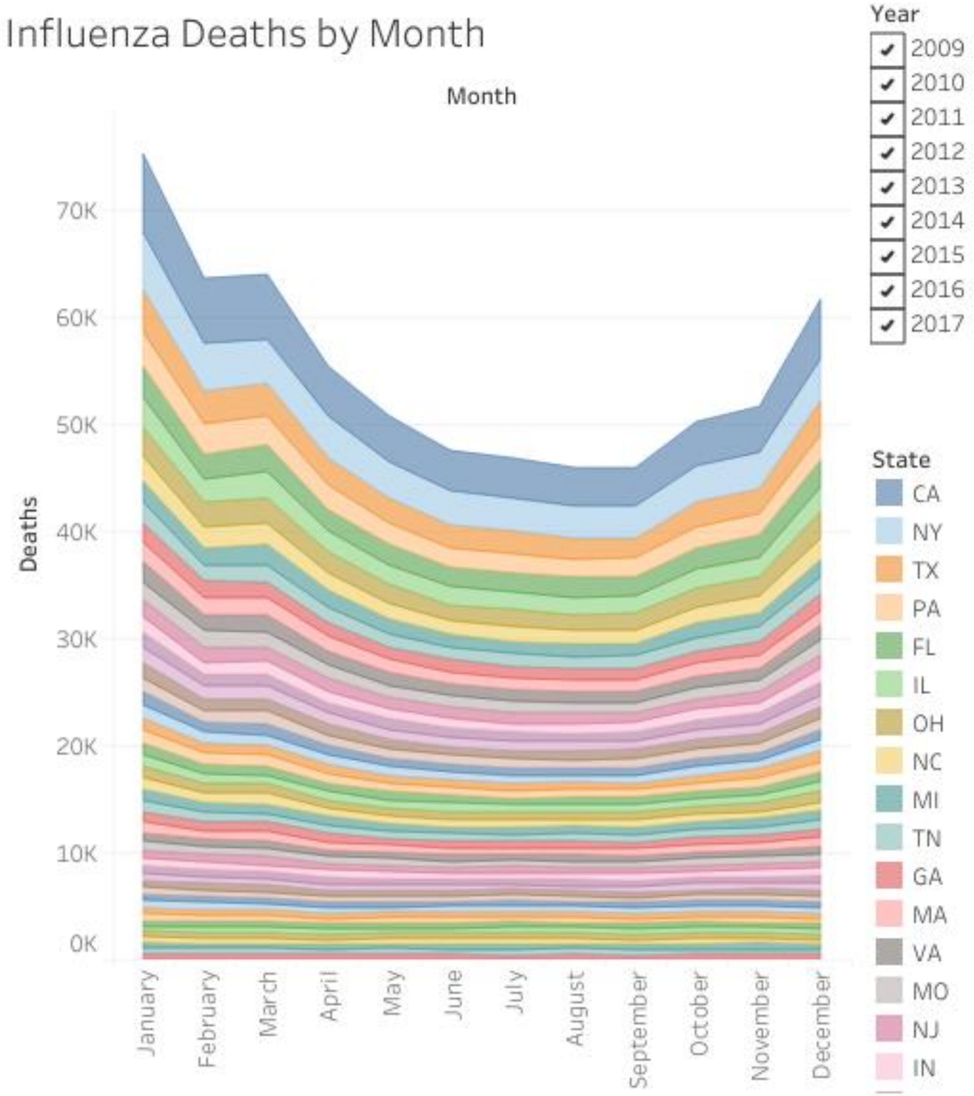
Visualized: the 65+ aged population and flu mortality heatmaps are very similar, demonstrating this correlation

View in Tableau

FLU SEASON: WHEN?

Overall, flu season occurs simultaneously throughout the US, running from September to May.

States stressed by the flu will be stressed simultaneously; staff allocations will be based on most need.

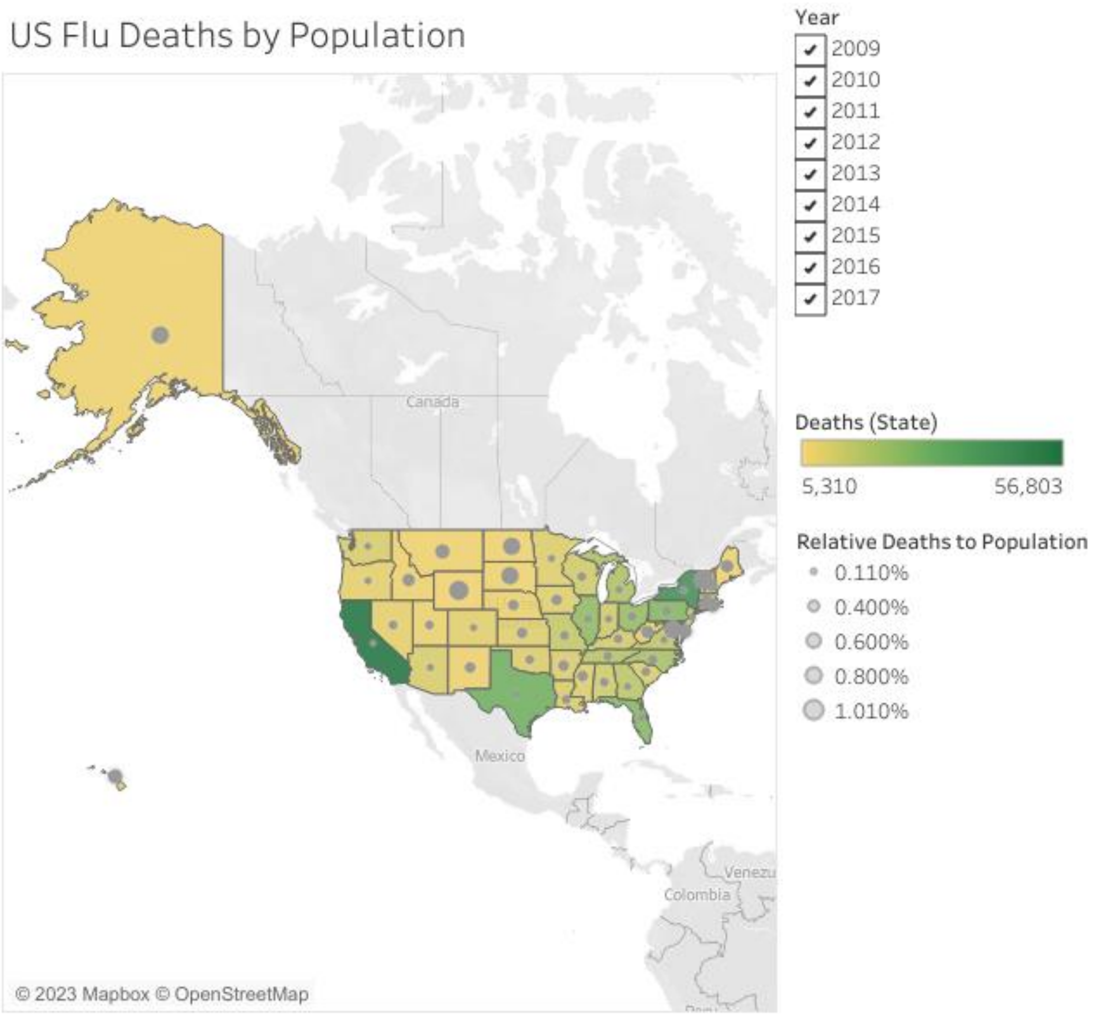


View in Tableau

FLU SEASON: WHERE TO SEND STAFF?

Staff ultimately need to be deployed where there is a greater mortality (CA, NY, TX, etc.), which namely tends to be where a larger 65+ and larger general population tends to be.

At the same time some states have a disproportionally large amount of deaths relative to their population (WY, ND, SD, etc.) and would also need a relatively higher amount of staff.



View in Tableau

PROJECT LINKS

PROJECT BRIEF

TABLEAU LINK

PRESENTATION (PPTX)

PRESENTATION (PDF)

VIDEO PRESENTATION

ROCKBUSTER STEALTH: MOVIE ACQUISITION STRATEGY

BACKGROUND

GOAL

Utilize existing data to inform the strategic approach of a video rental store franchise in introducing a new online service, with a focus on revenue, customer base, and geography.

TOOLS

PostgreSQL
Tableau
DbVisualizer

DATASET

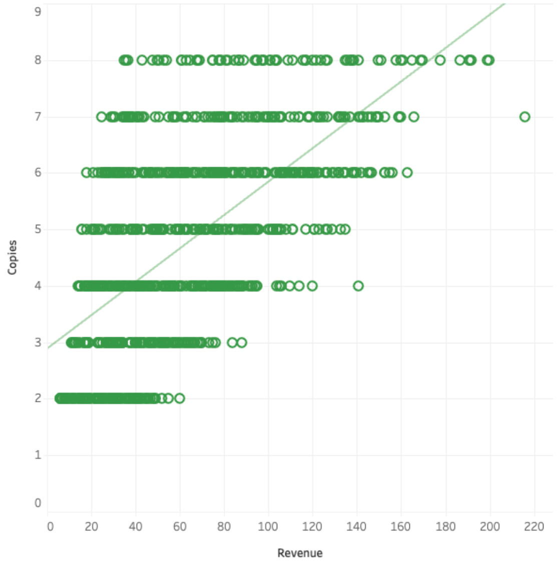
Fictional company Rockbuster's inventory, customer, and payment data

SKILLS USED

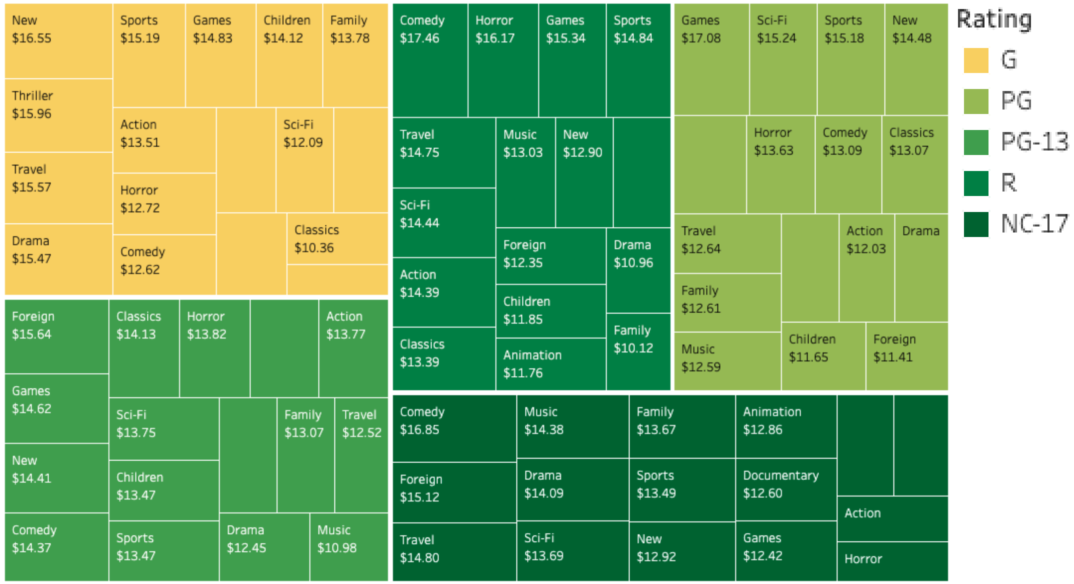
Relational databases
SQL
Database querying
Filtering
Cleaning and summarizing
Joining tables
Subqueries
Common table expressions

ROCKBUSTER: REVENUE INSIGHTS

Copies to Revenue Correlation



Revenue per Copy by Genre and Rating



Based on revenue & inventory data, there is a strong positive correlation (**0.62**) between the copies in inventory and the revenue a movie generates

Rockbuster should license movies generating above average revenue per copy (**$13.41**) for the streaming service

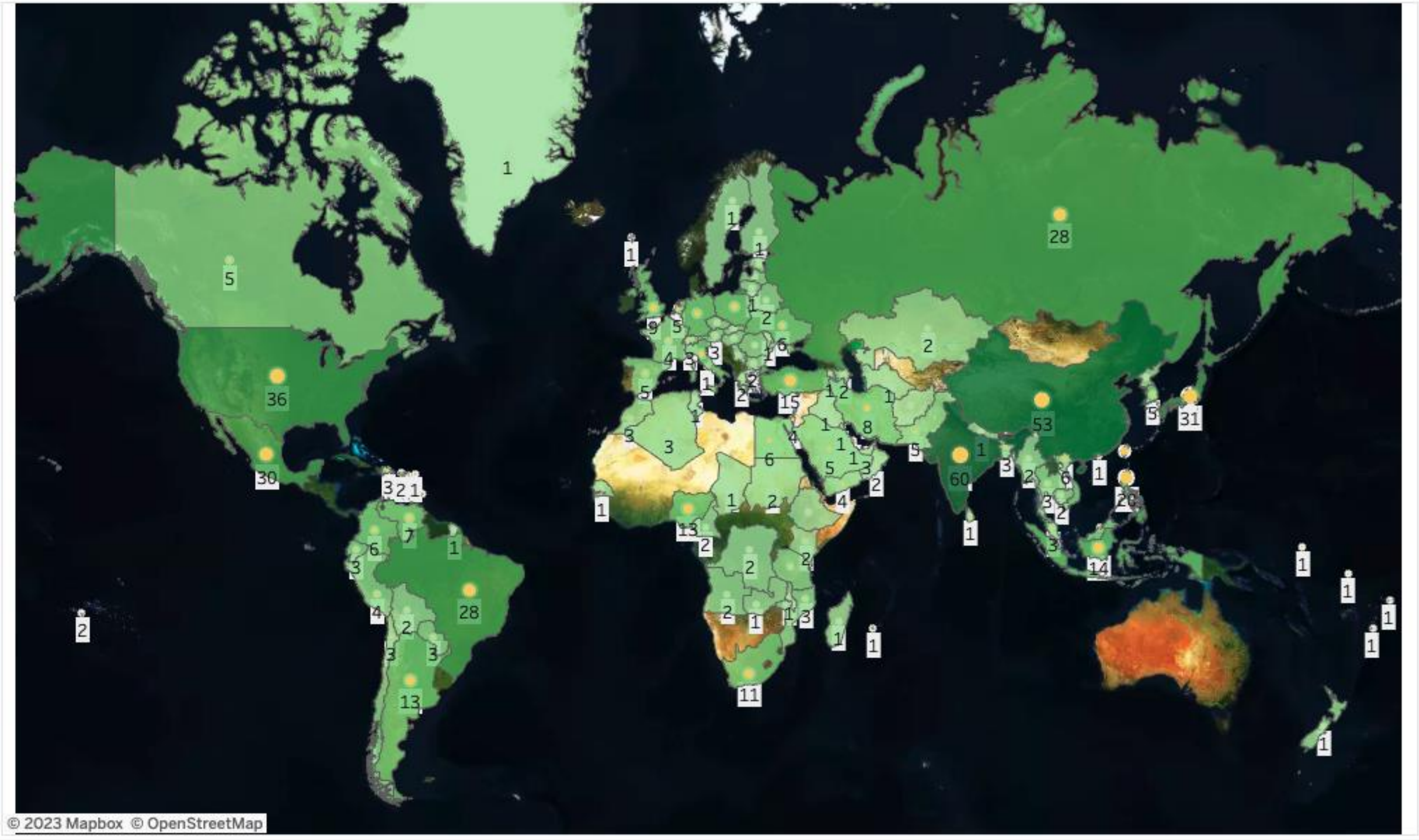
Different rating and genre combinations brought in different revenue values: New PG-13 generate best revenue per copy (**$16.55**)

ROCKBUSTER: CUSTOMER COUNTS AND REVENUE

Country	Customers	Revenue
India	**60**	**$ 2,367.29**
China	**53**	**$ 2,089.80**
United States	36	$ 1,344.31
Japan	31	$ 1,185.61
Mexico	30	$ 1,143.73
Brazil	28	$ 1,144.78
Russian Federation	28	$ 1,044.89
Philippines	20	$ 812.40
Turkey	15	$ 581.86
Indonesia	14	$ 548.92

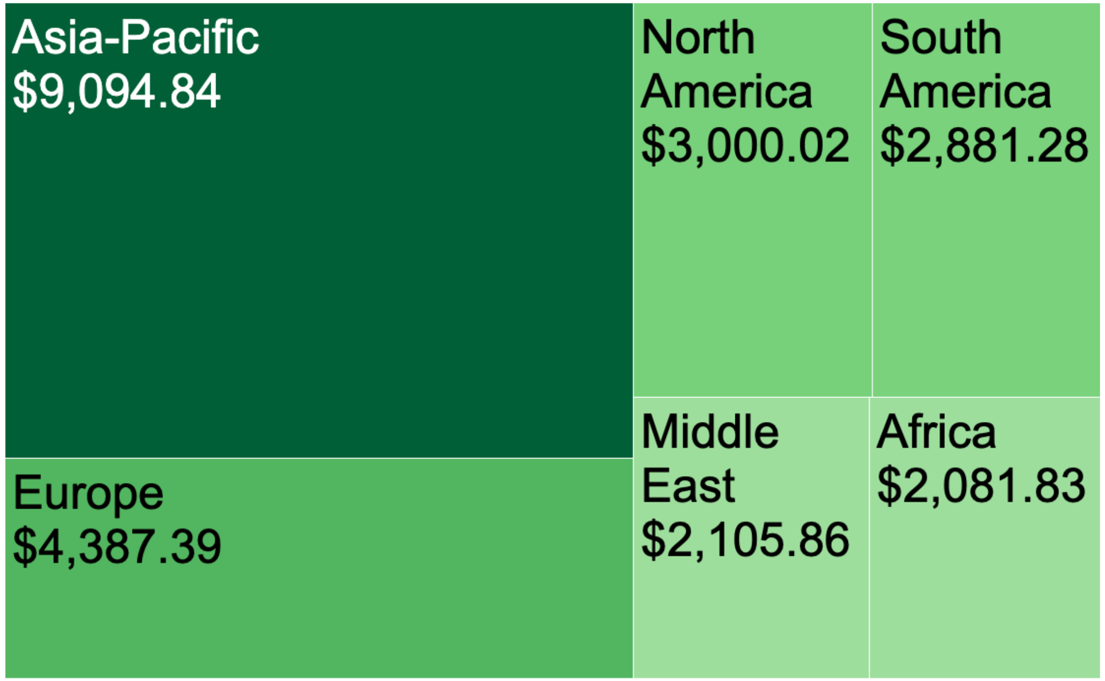
India and China are the largest countries by far in terms of customers and revenue.

Customers and Revenue by Country



ROCKBUSTER: REGIONS AND RECOMMENDATIONS

Revenue by Region



- Focus especially on Asia-Pacific as this has been Rockbuster Stealth's strongest consumer base, but…
- Keep internet accessibility in mind! Some of our customers may be better served by physical locations due to poor internet access

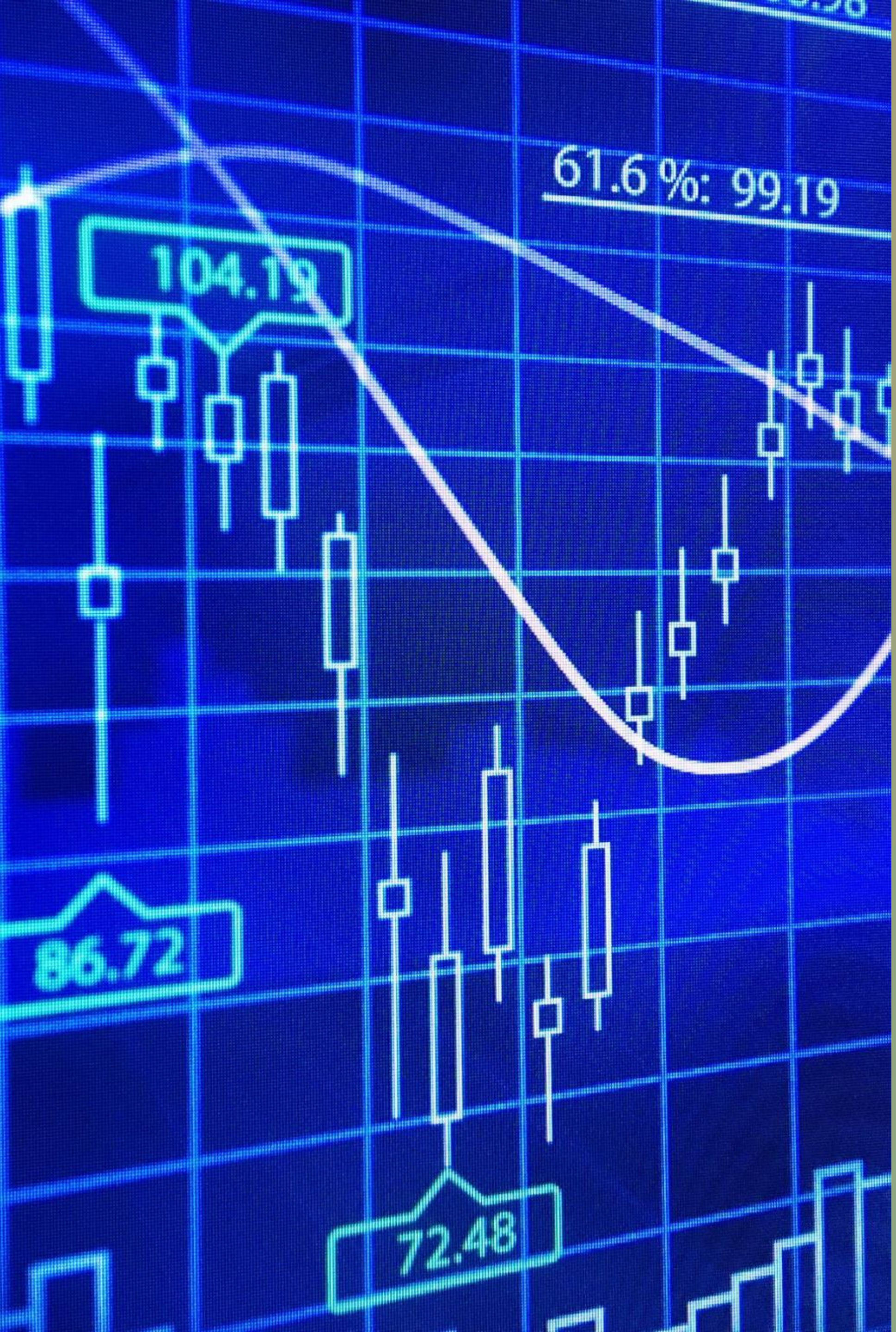
PROJECT LINKS

PROJECT BRIEF

PRESENTATION (PPTX)

PRESENTATION (PDF)

DATA DICTIONARY



INSTACART: INITIAL DATA & EXPLORATORY ANALYSIS

BACKGROUND

GOAL

Perform an initial data and exploratory analysis on sales information in order to derive insights and suggest strategies for better segmentation

TOOLS

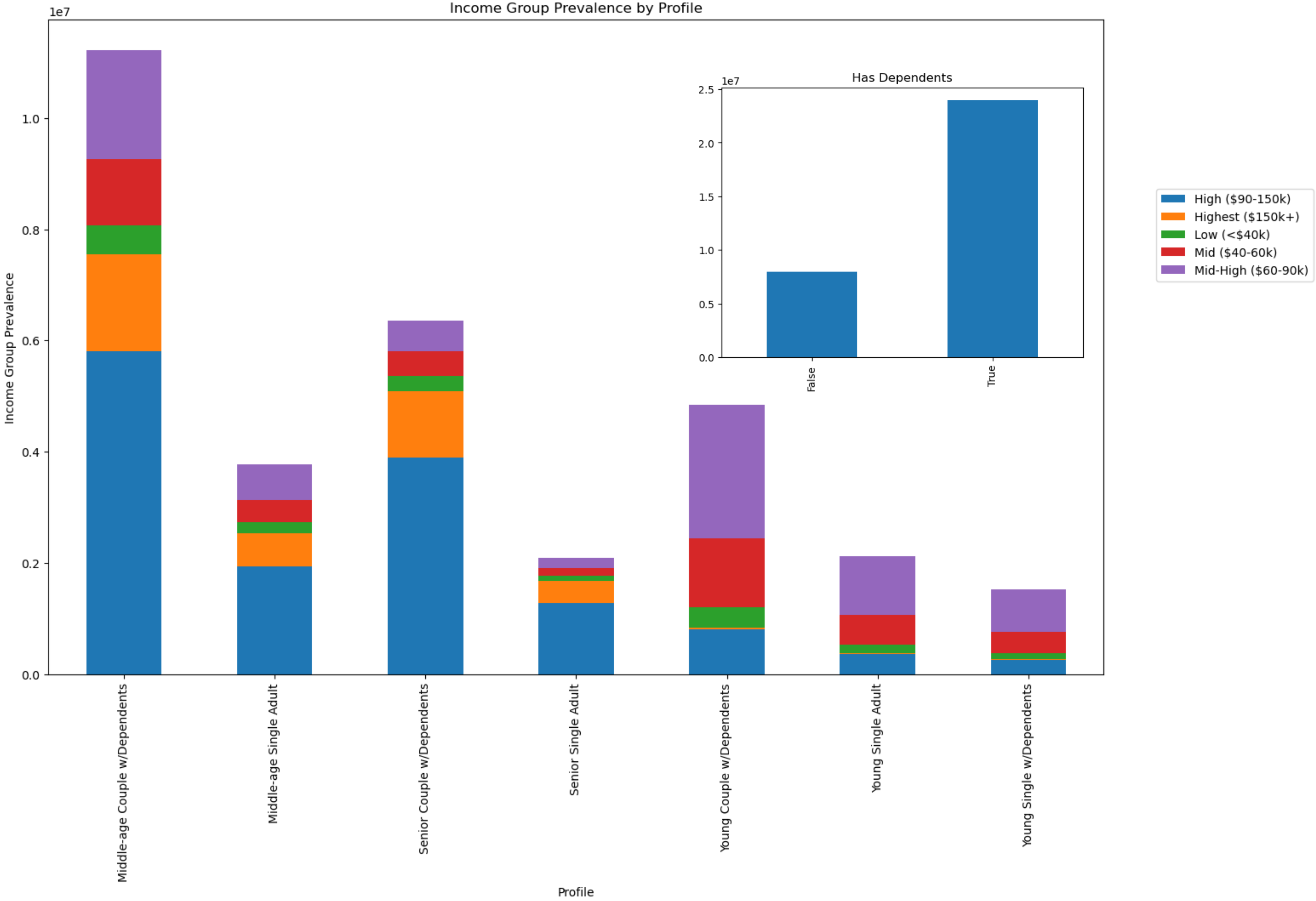
Python
Jupyter
Pandas
NumPy
Matplotlib
Seaborn

DATASET

The Instacart Online Grocery Shopping Dataset 2017

SKILLS USED

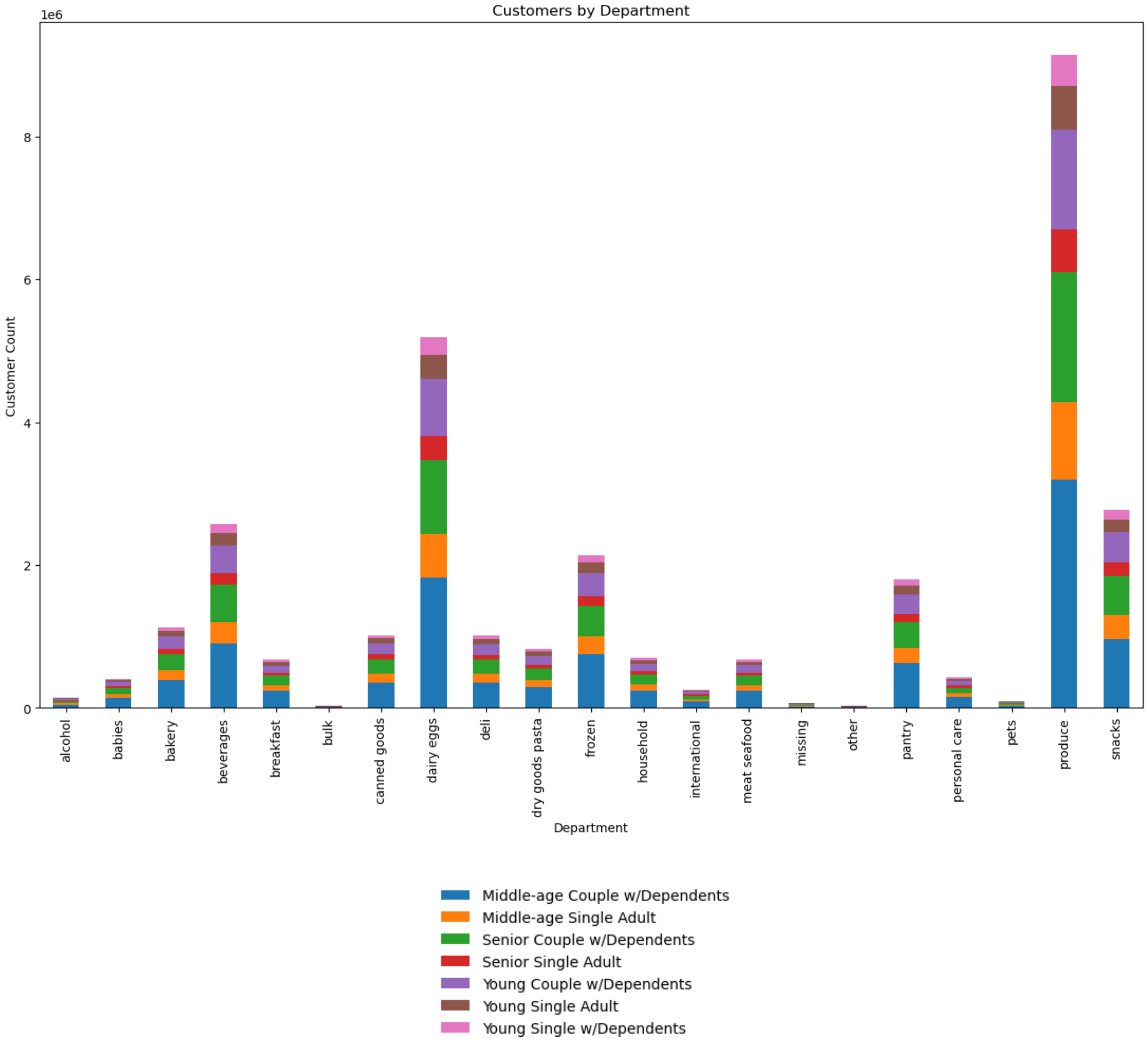
Data wrangling
Data merging
Deriving variables
Grouping data
Aggregating data
Reporting in Excel
Population flows



INSTACART: CUSTOMER ORDER INSIGHTS

Instacart's orders are primarily from those w/dependents earning >$60k, **61% of all orders.**

75% of Instacart's orders come from those w/dependents.

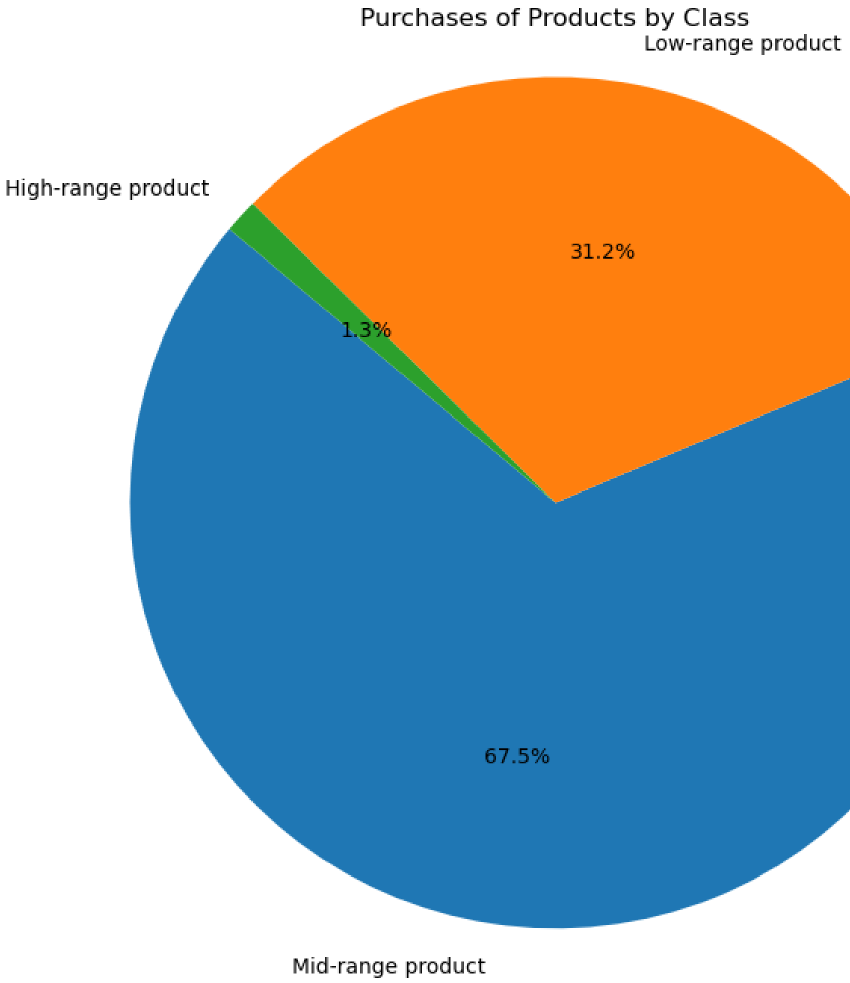


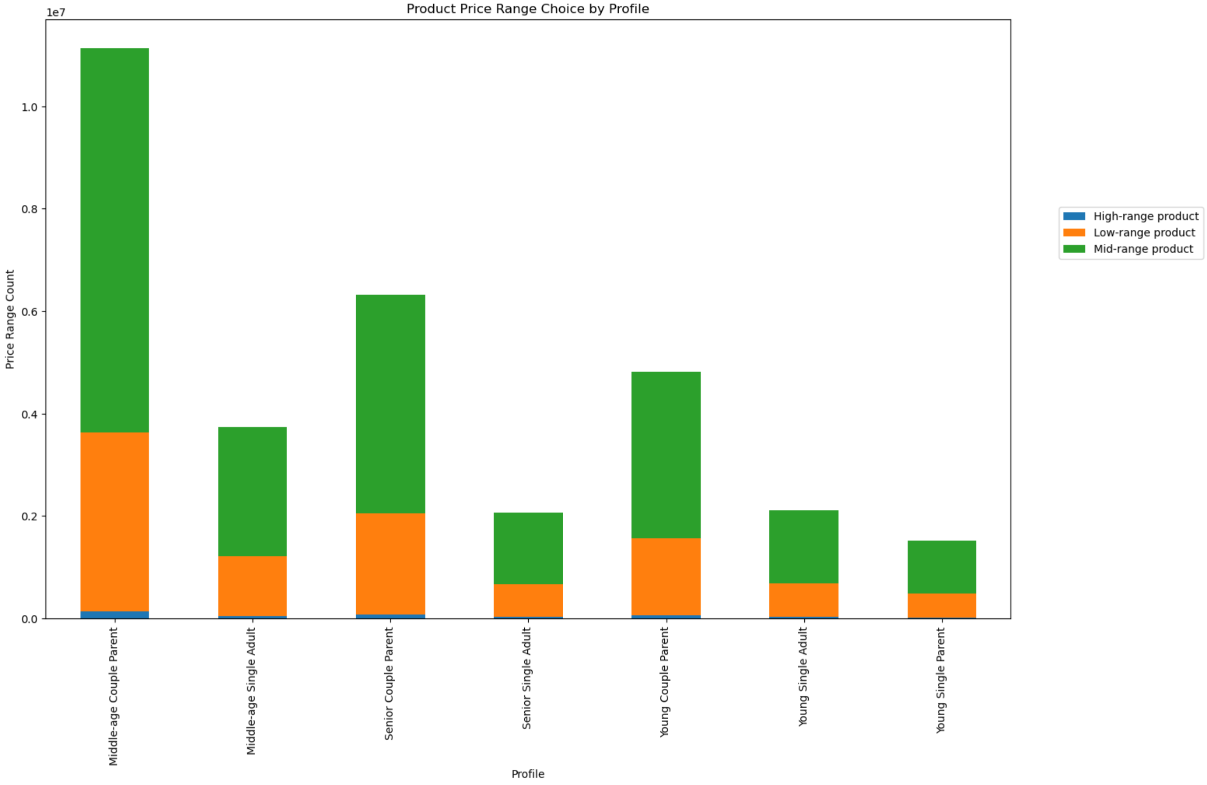
INSTACART: PURCHASING BEHAVIOR

Dairy eggs (5.9m) and produce (9.1m) are by far the two largest categories purchased in orders while the meat/seafood is pretty tiny (679k), which could suggest that these customers are more likely to be vegetarian.

However, the canned goods (662k), snacks (2.8m), and frozen (2.1m) departments are also sizable and would best be broken down into further categories in order to make the best judgment there.

INSTACART: PRODUCT CLASSES





Most Instacart orders (67.5%) are for mid-range products ($5-15), a newly derived category.

INSTACART: PROJECT LINKS

PROJECT BRIEF

GITHUB

EXCEL REPORT

PIG E. BANK: ASSESSING RISK

BACKGROUND

GOAL

Identify factors to predict the likelihood of customers leaving the bank

TOOLS

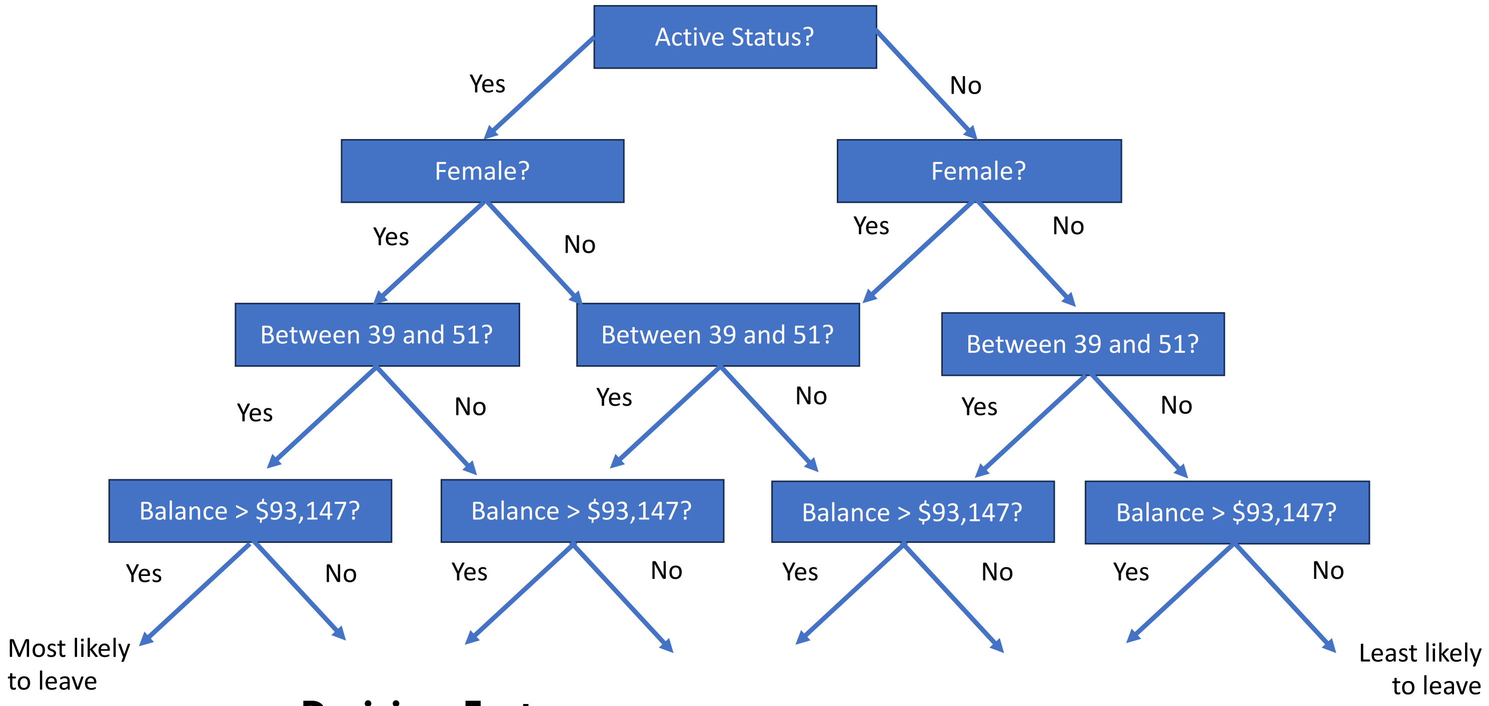
Excel
GitHub

SKILLS USED

Big data
Data ethics
Data mining
Predictive analysis
Time series analysis
Forecasting
Using GitHub

PIG E. BANK: LEAVER CHART

Pig E Bank Leaver Decision Chart



Active Status?

Yes

No

Female?

Female?

Yes

No

Yes

No

Between 39 and 51?

Between 39 and 51?

Between 39 and 51?

Yes

No

Yes

No

Yes

No

Balance > $93,147?

Balance > $93,147?

Balance > $93,147?

Balance > $93,147?

Yes

No

Yes

No

Yes

No

Yes

No

Most likely to leave

Least likely to leave

Decision Factors

1. **Lacking active status:** 70% of leavers do not have active status

2. **Being a woman:** about 60% of leavers are women

3. **Being between age 39 and 51:** about 55% of leavers are of this age

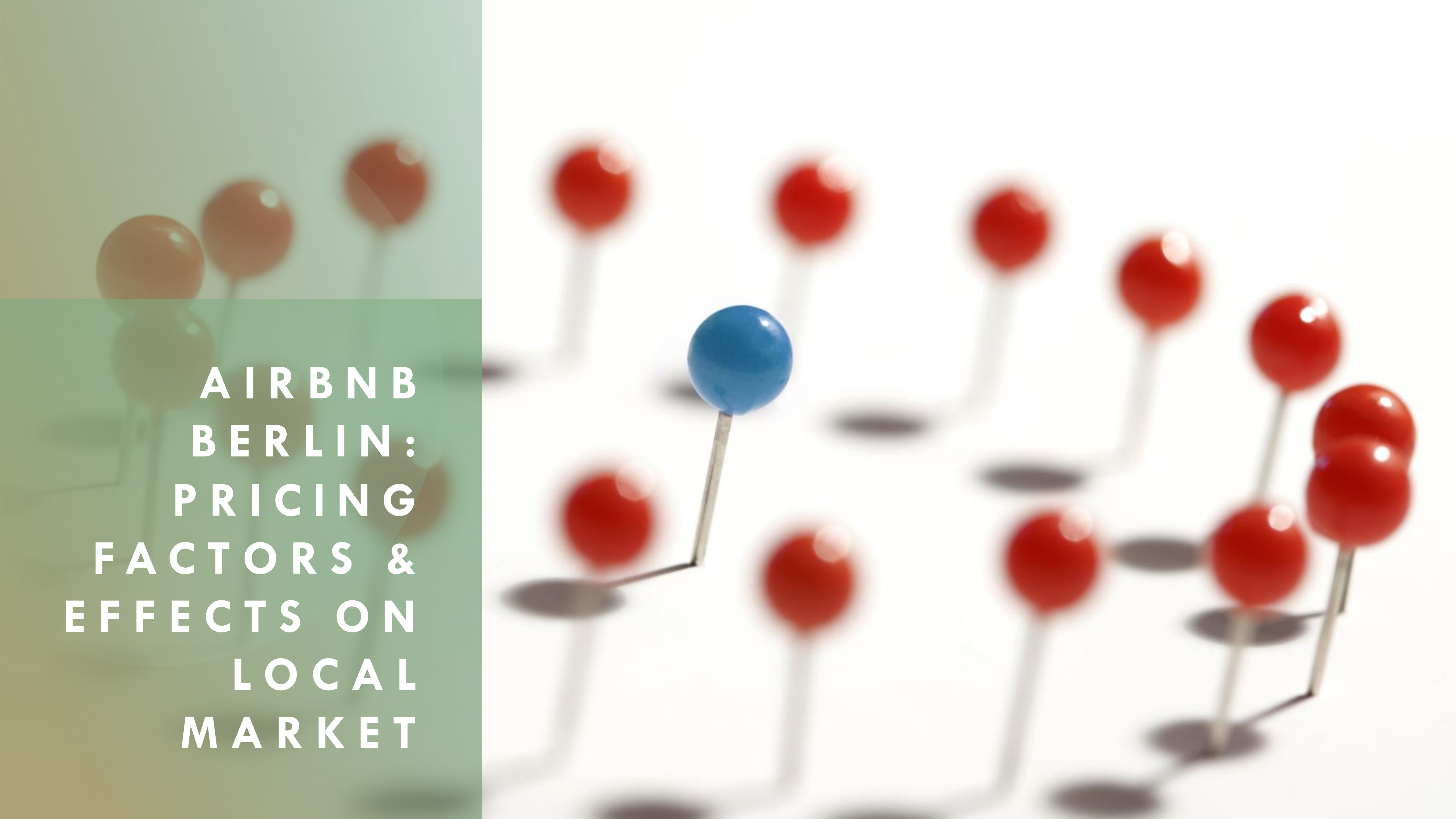
4. **Have a higher balance:** leavers' median and average balances ($93,147) are 20% higher than those who remain ($74,830).

PIG E. BANK : PROJECT LINKS

PROJECT BRIEF

EXCEL DATA

AIRBNB BERLIN: PRICING FACTORS & EFFECTS ON LOCAL MARKET



BACKGROUND

GOAL

Conduct time series, geospatial, and clustering analysis on data from Airbnb to determine price and rating factors, judge effects on local markets, determine effect of recent regulation

TOOLS

Python
Jupyter
Pandas
NumPy
Matplotlib
Seaborn
SciPy
Tableau
Excel

DATASET

Data assembled from Insideairbnb.com

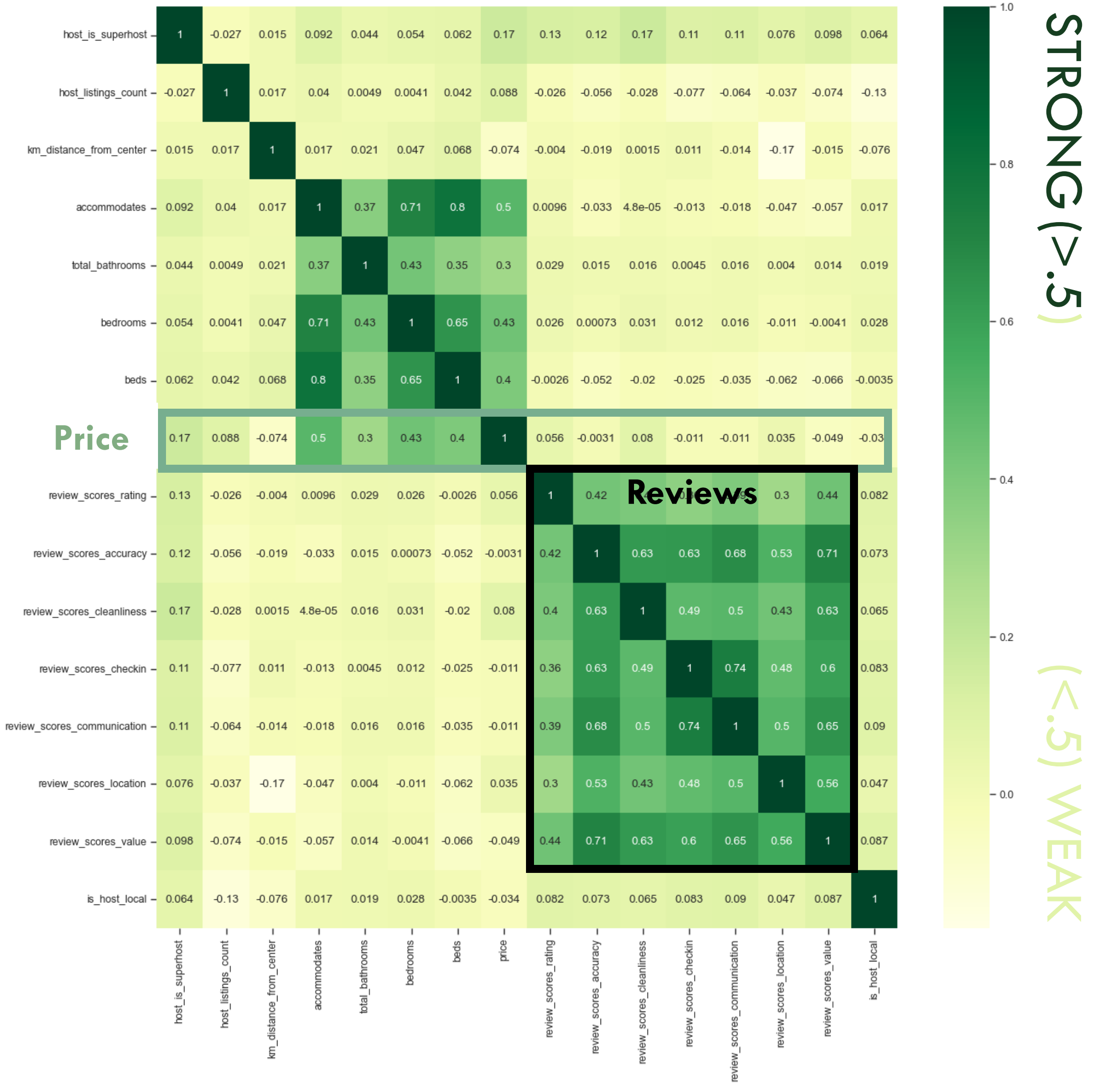
SKILLS USED

Sourcing open data
Data wrangling & cleaning
Geospatial analysis
Dickey-Fuller Test
Autocorrelations
Regression analysis
Clustering
Time series data sourcing and analysis
Supervised and unsupervised machine learning

PRICE & RATING CORRELATION

From correlation, the only quantitative factors affecting price were accommodation capacity and related comfort factors (bed counts, bedrooms, bathrooms, etc.), and at weak levels (**0.5** and lower)

Ratings only seemed to be correlated with each other



	host_is_superhost	host_listings_count	km_distance_from_center	accommodates	total_bathrooms	bedrooms	beds	price	review_scores_rating	review_scores_accuracy	review_scores_cleanliness	review_scores_checkin	review_scores_communication	review_scores_location	review_scores_value	is_host_local
host_is_superhost	1	-0.027	0.015	0.092	0.044	0.054	0.062	0.17	0.13	0.12	0.17	0.11	0.11	0.076	0.098	0.064
host_listings_count	-0.027	1	0.017	0.04	0.0049	0.0041	0.042	0.088	-0.026	-0.056	-0.028	-0.077	-0.064	-0.037	-0.074	-0.13
km_distance_from_center	0.015	0.017	1	0.017	0.021	0.047	0.068	-0.074	-0.004	-0.019	0.0015	0.011	-0.014	-0.17	-0.015	-0.076
accommodates	0.092	0.04	0.017	1	0.37	0.71	0.8	0.5	0.0096	-0.033	4.8e-05	-0.013	-0.018	-0.047	-0.057	0.017
total_bathrooms	0.044	0.0049	0.021	0.37	1	0.43	0.35	0.3	0.029	0.015	0.016	0.0045	0.016	0.004	0.014	0.019
bedrooms	0.054	0.0041	0.047	0.71	0.43	1	0.65	0.43	0.026	0.00073	0.031	0.012	0.016	-0.011	-0.0041	0.028
beds	0.062	0.042	0.068	0.8	0.35	0.65	1	0.4	-0.0026	-0.052	-0.02	-0.025	-0.035	-0.062	-0.066	-0.0035
Price	0.17	0.088	-0.074	0.5	0.3	0.43	0.4	1	0.056	-0.0031	0.08	-0.011	-0.011	0.035	-0.049	-0.03
review_scores_rating	0.13	-0.026	-0.004	0.0096	0.029	0.026	-0.0026	0.056	1	0.42	[illegible]	[illegible]	[illegible]	0.3	0.44	0.082
review_scores_accuracy	0.12	-0.056	-0.019	-0.033	0.015	0.00073	-0.052	-0.0031	0.42	1	0.63	0.63	0.68	0.53	0.71	0.073
review_scores_cleanliness	0.17	-0.028	0.0015	4.8e-05	0.016	0.031	-0.02	0.08	0.4	0.63	1	0.49	0.5	0.43	0.63	0.065
review_scores_checkin	0.11	-0.077	0.011	-0.013	0.0045	0.012	-0.025	-0.011	0.36	0.63	0.49	1	0.74	0.48	0.6	0.083
review_scores_communication	0.11	-0.064	-0.014	-0.018	0.016	0.016	-0.035	-0.011	0.39	0.68	0.5	0.74	1	0.5	0.65	0.09
review_scores_location	0.076	-0.037	-0.17	-0.047	0.004	-0.011	-0.062	0.035	0.3	0.53	0.43	0.48	0.5	1	0.56	0.047
review_scores_value	0.098	-0.074	-0.015	-0.057	0.014	-0.0041	-0.066	-0.049	0.44	0.71	0.63	0.6	0.65	0.56	1	0.087
is_host_local	0.064	-0.13	-0.076	0.017	0.019	0.028	-0.0035	-0.034	0.082	0.073	0.065	0.083	0.09	0.047	0.087	1

CLUSTERING

Through **unsupervised machine learning**, we derived four different clusters primarily based on price and listing counts, helping us determine that:

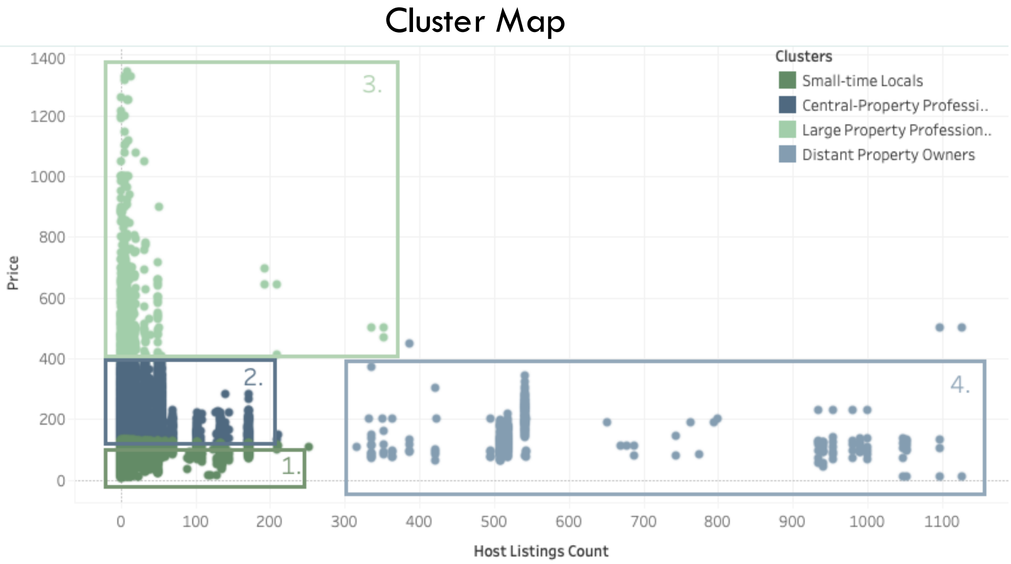
- **Small-time Locals** (79% of listings) are leaving the platform (33% between Sep 2022–Jun 2023), likely driven by a March 2023 licensing requirement
- Other, more professionalized clusters tended to stay with the platform

1. Small-time Locals

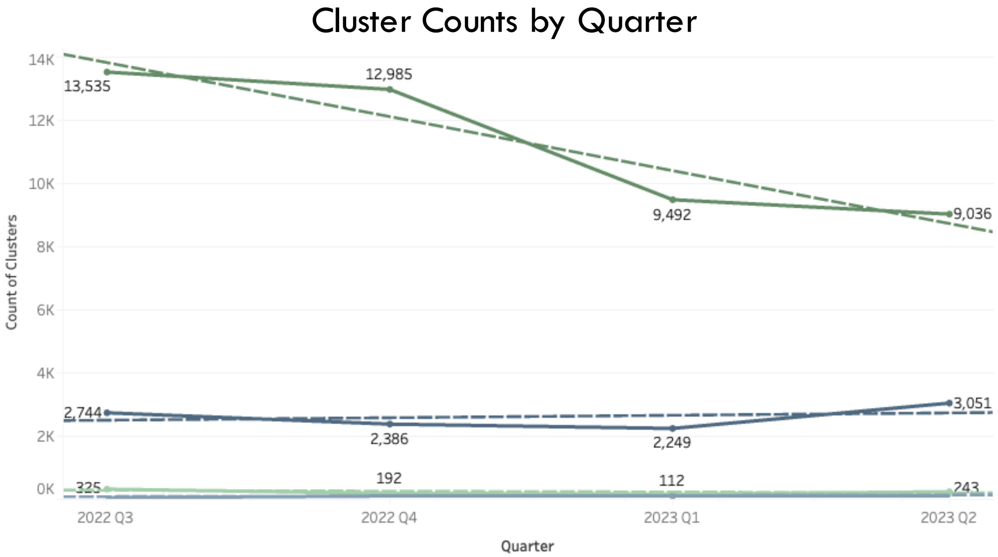
2. Central Property Professionals

3. Large Property Professionals

4. Distant Property Owners



Cluster Map

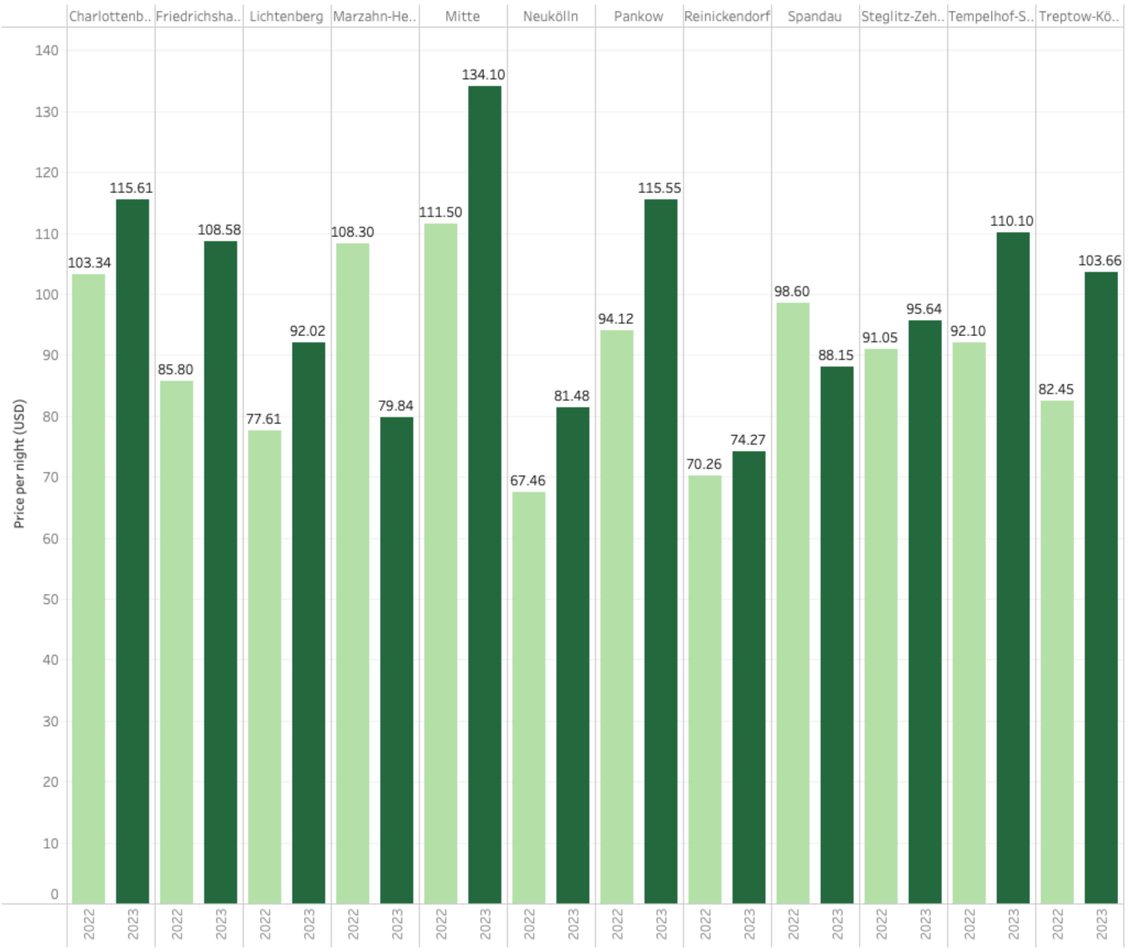


Cluster Counts by Quarter

PRICE EFFECTS

- Airbnb prices stayed flat 2022-2023, but **Small-time Locals** charged the lowest prices (**$63 per night**)
- As 33% left the platform, the average cost of Airbnb rose 11.6%, exceeding local rental price increases by 8%, German national inflation by 9.6%
- Airbnb presence is known to drive up local rental prices and this will continue, but prices will stabilize once the Small-Time Local exodus is complete

Price per Night by Neighborhood and Year



PROJECT LINKS

PROJECT BRIEF

TABLEAU LINK

GITHUB

THANK YOU