KIRSTEN CURRIE

Data Analytics Portfolio

CONTENTS

TOOLS USED

GAMECO

Finding Market Fit in the Global Video Game Industry

03

INTRO In order to launch into the global video game sales market, new game company GameCo would like to gain a sense of current global video game sales in order to establish their niche within the marketplace.

- Over 35 years of game sales data was provided; to bring focus for stakeholders, the data was filtered down to the most recent 4 years in sales.

- The top 3 regions were evaluated to provide context of genre drivers

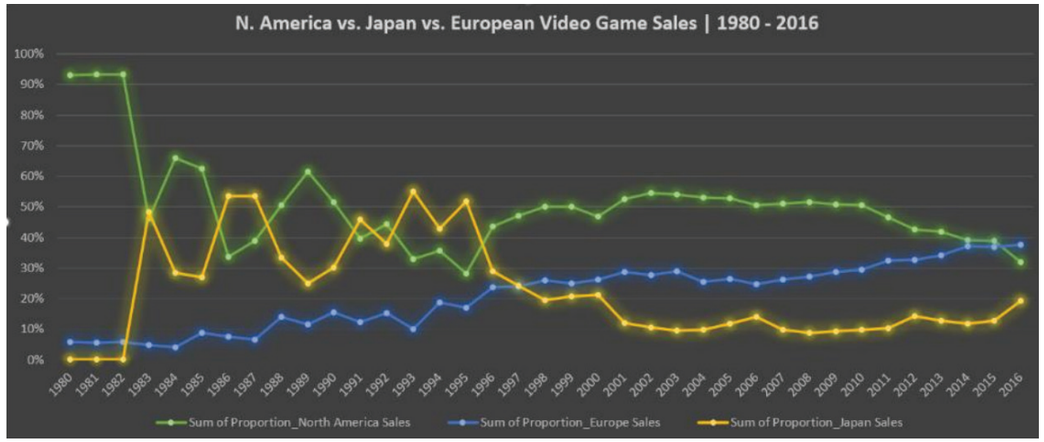


FIG. A: Overwhelming view of 35+ years in sales

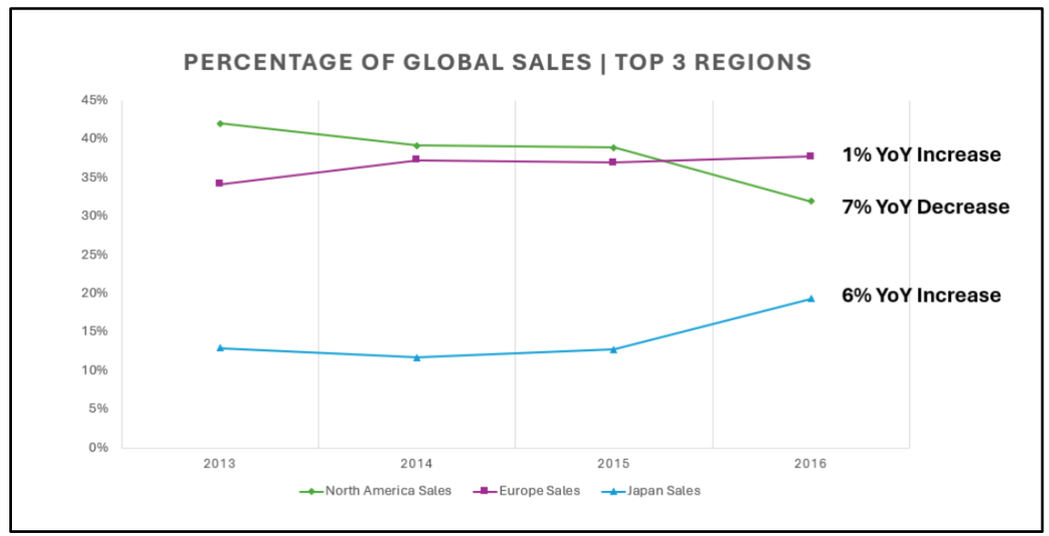


FIG. B: Data simplified to previous 4 years of top regional sales

HIGHLIGHTS A further drill-down into 2016 sales by region revealed potential genres by means of a categorical stacked bar chart. It was essential to define niche preferences by region in order to meet their unique demands.

- **Shooter** games performed strong across NA & EU with **26% of total global sales**.
- **Action** games performed well across all 3 regions and represented **28% of total global sales**.
- **Sports** were primarily strong for NA & EU and represented **21% of total global sales**.
- Though not strong in EU & NA, **Role Playing Games** represent a top pick for **Japan** game players.

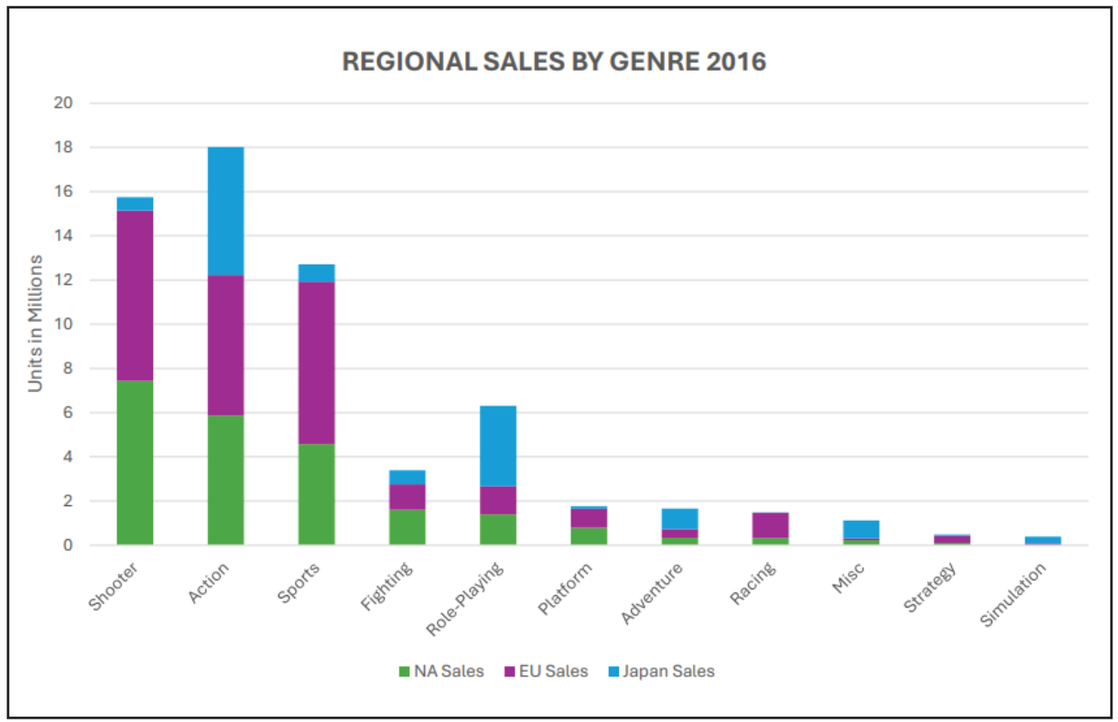


FIG. C: Stacked bar chart highlights regional preferences; a clear win for Shooter, Action, & Sports

Shooter games—a top genre for NA sales—saw a shift in top Publishers between 2015 and 2016. **Does a shakeup in major players represent market share opportunity up for grabs?**

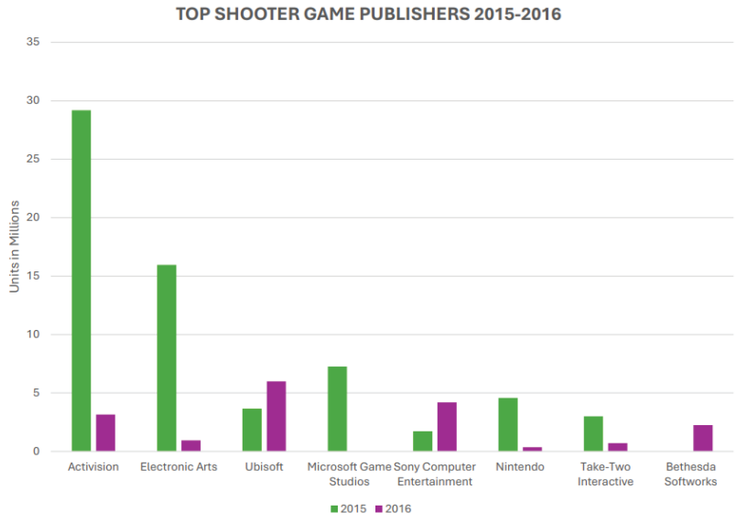


FIG. D: Comparison bar chart revealing a shift in market demand between 2015 - 2016 for Shooter games

TAKEAWAYS In order to claim their place among existing video game producers, GameCo must evaluate current offerings and determine how they will differentiate. A further evaluation of publishers within top performing genres was provided to the client to clarify possible market opportunity.

The following recommendations were provided:

- Assess a possible market entry into Shooter games within the U.S.
- Follow the shift from Action to Sports games within the E.U. & be first to market to capture this change
- Follow Japan's recent growth spike in RPG sales and offer a combination of genres within this category (e.g. mixup of RPB, Action, & Adventure games)

INFLUENZA STUDY

Helping Medical Staffing Agencies Prepare for Flu Season

07

Interactive Influenza T ableau Storyboard HERE

Live Influenza Storyboard Presentation HERE

INTRO The scope of this project involved employing CDC Influenza data in addition to the United States census to arm a theoretical medical staffing agency with knowledge on how to prepare for the upcoming flu season by knowing when and where to send their teams.

The initial goal was to **develop a hypothesis** around which age group would be the most vulnerable and therefore require the most aid.

The over 65 year old and under 5 are considered by health experts to be the most at need. However, where does the data indicate death rates hit the highest?

By exploring the correlation between death rates and age groups, it became clear that the **over 65 population had a strong, positive correlation** and therefore necessitated higher amounts of medical staff.

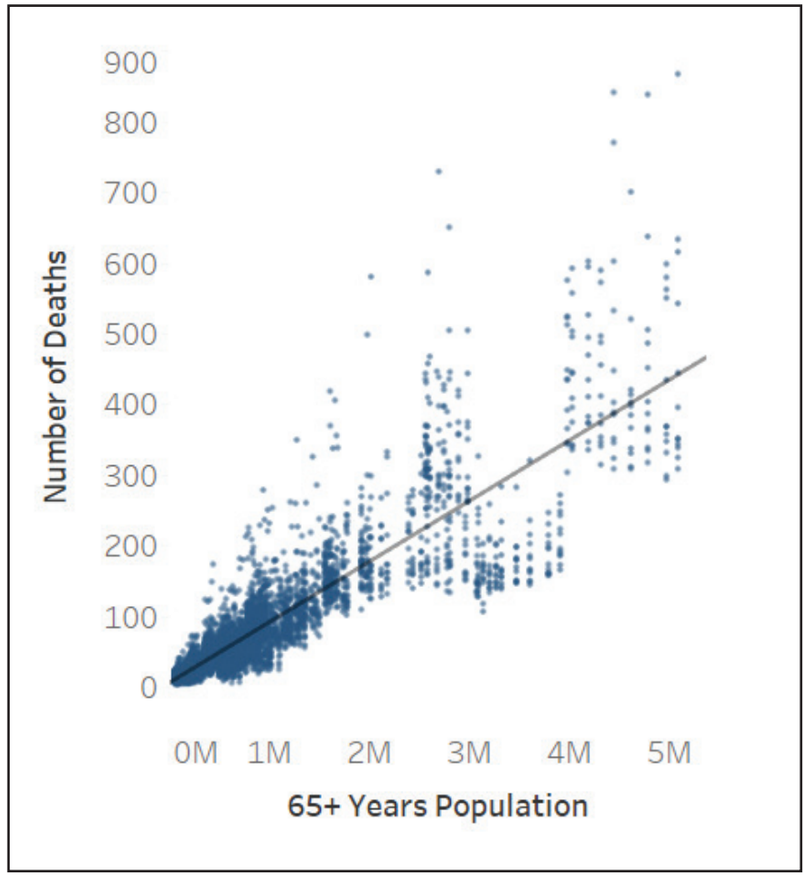


FIG. A: Compared to the <5 age group, over 65+ individuals showed higher death to population correlation.

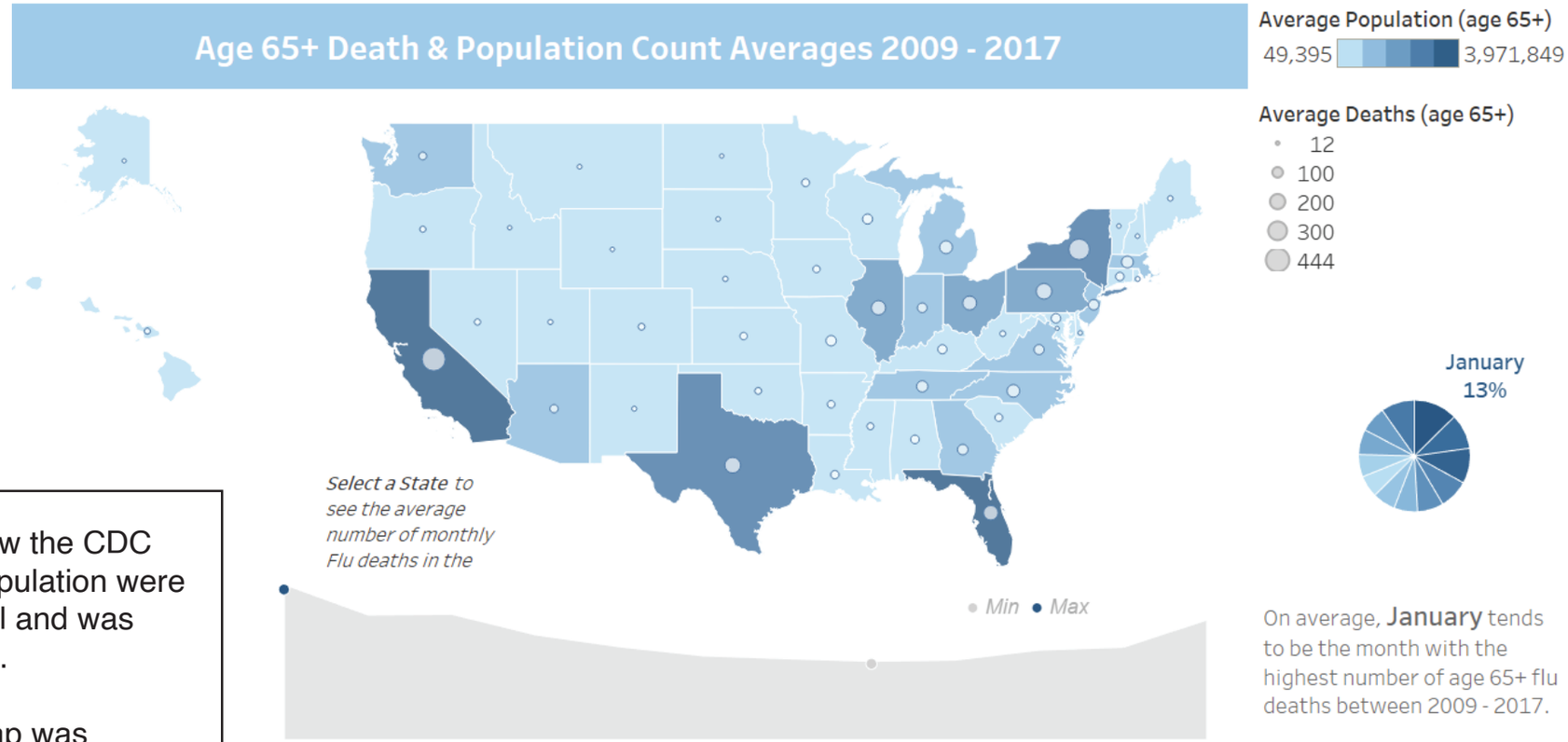


FIG. B: An interactive choropleth dashboard that explores ages 65+ Influenza deaths by state

HIGHLIGHTS In order to fully see how the CDC Influenza death rates and total U.S. population were correlated, the data was joined in Excel and was aggregated by the relevent age groups.

The following interactive choropleth map was developed to find the states with the highest death rates for those over the age 65.

By selecting each state, you will see maximum and minium number of deaths in each state between the years 2009 and 2017.

Additionally, the death counts were tied to a pie chart which illustrated how January tended to be the month with the highest amount of deaths on average (though there were some individual variances by state).

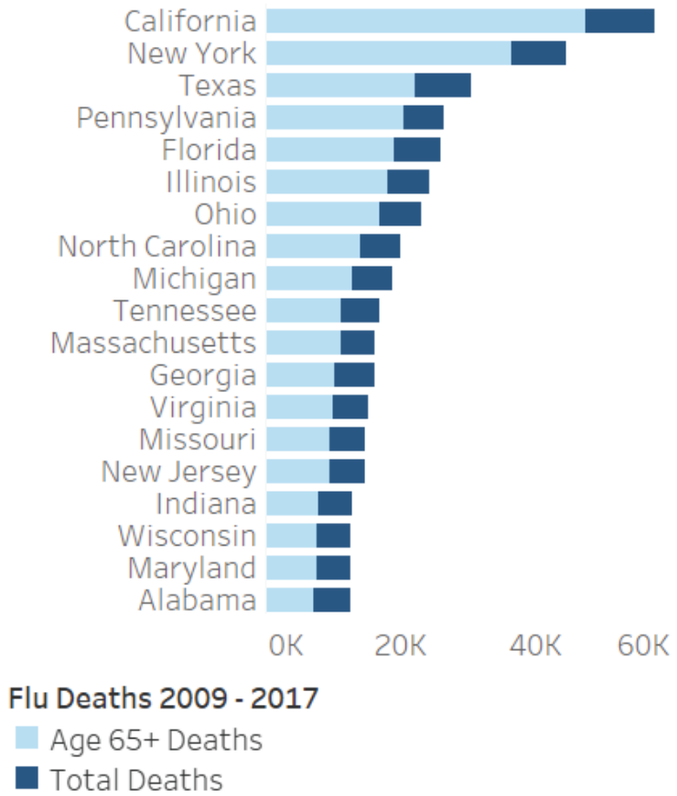


FIG. C: Bar chart that highlights states with the highest death rates (and makes it clear how the 65+ population suffers more than the other age groups)

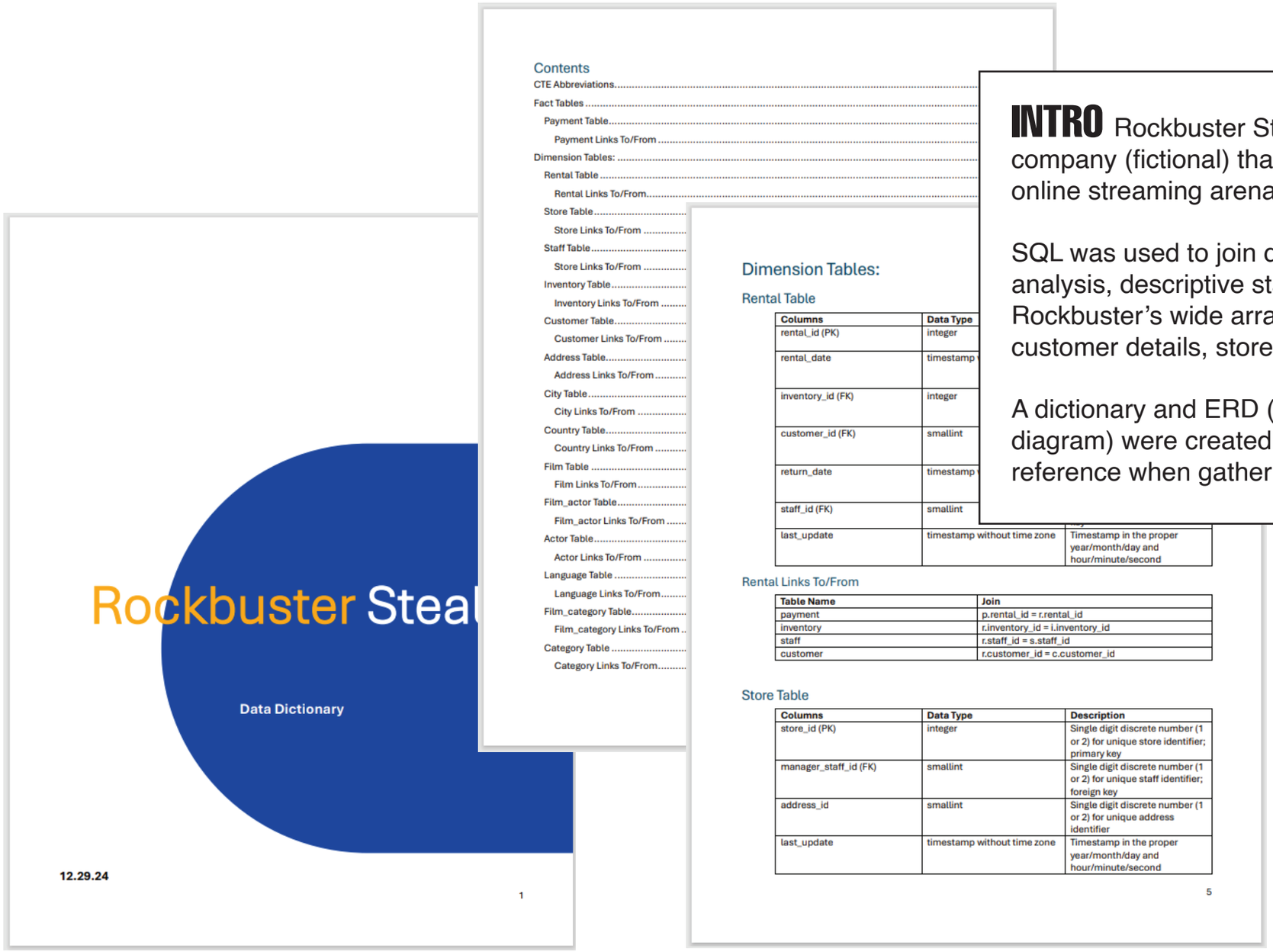
TAKEAWAYS After performing the analysis, it was clear that states like California, New York, or Texas were in the highest need of aid--in part due to their larger population sizes.

A staffing agency would need to adequately prepare its team for the peak winter months while taking care to stock and get all preparation into place during the slower months of August and September.

Oddly enough, states with mid-sized populations like Illinois or Pennsylvania had higher death counts, so as a further study, it was determined more could be done to explore the cause of these abnormalities.

Rockbuster Stealth

Turning Video Rental Data into Online Streaming Success



INTRO Rockbuster Stealth is a video rental company (fictional) that would like to enter into the online streaming arena.

SQL was used to join data, perform exploratory data analysis, descriptive statistics, and groupings with Rockbuster's wide array of tables containing customer details, store locations, and movie sales.

A dictionary and ERD (exploratory relationship diagram) were created and used as a continual reference when gathering insights from the data.

FIG. A: Pages from the Rockbuster Stealth Data Dictionary

HIGHLIGHT This entity relationship diagram was a frequently referenced tool when working on a number of joins within the Rockbuster data tables.

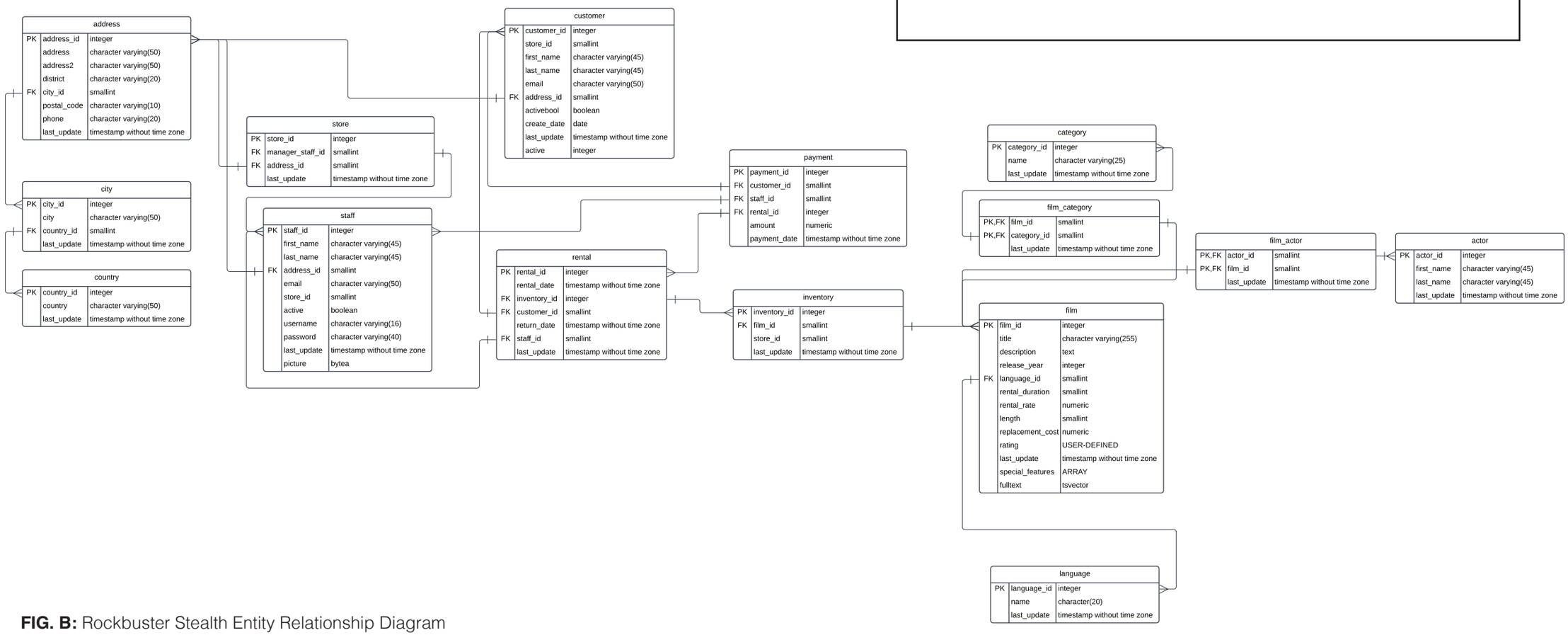


FIG. B: Rockbuster Stealth Entity Relationship Diagram

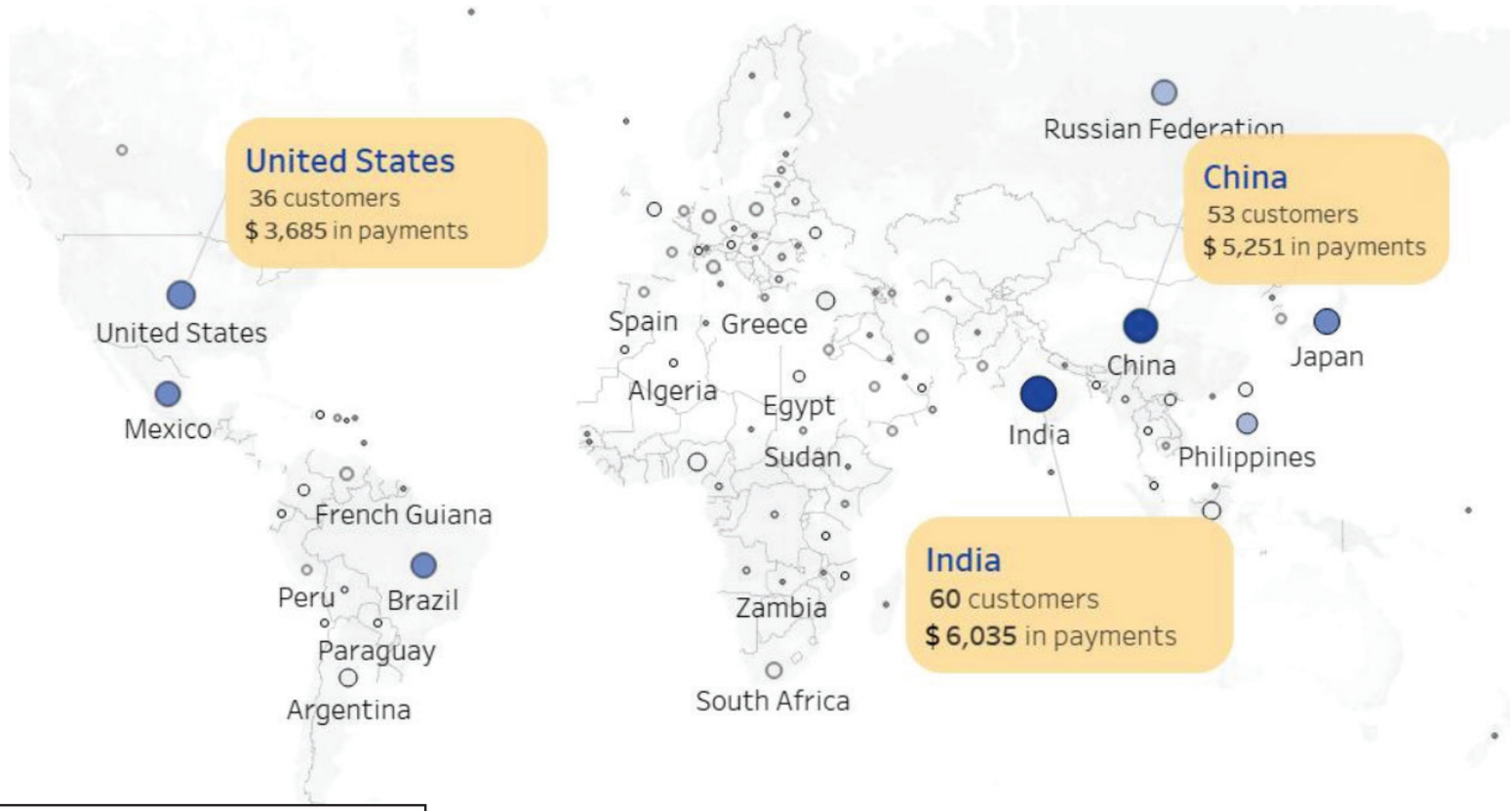


FIG. C: A proportional symbol map shows that India, the United States, and China have the highest contribution to Rockbuster's video sales.

HIGHLIGHT Data was pulled for top sales by country to illustrate where the highest demand was coming from.

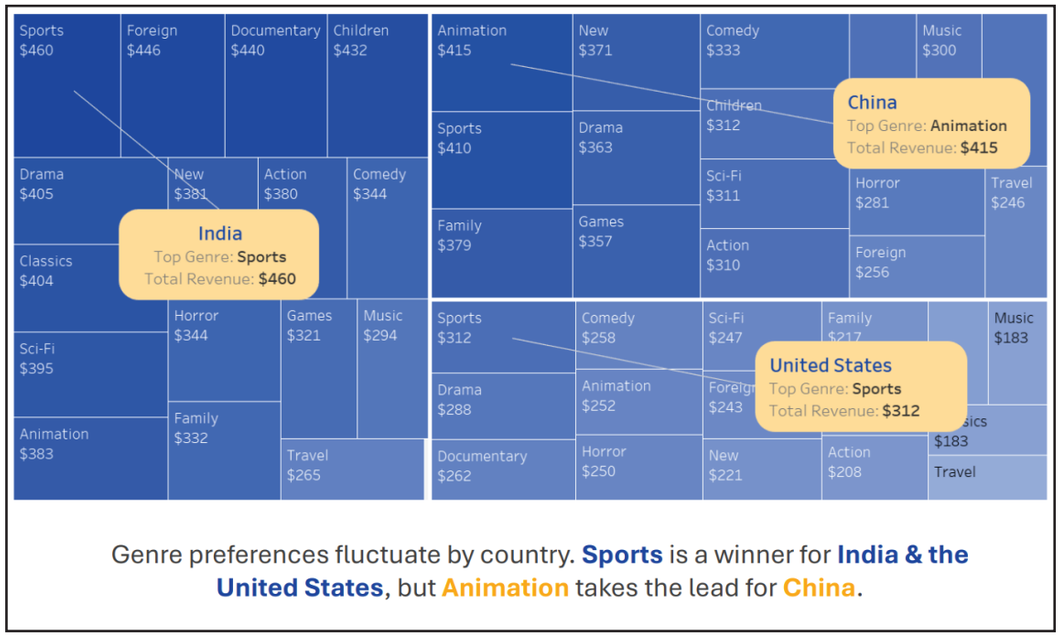


FIG. D: A data tree map illustrates film preferences by top 3 countries

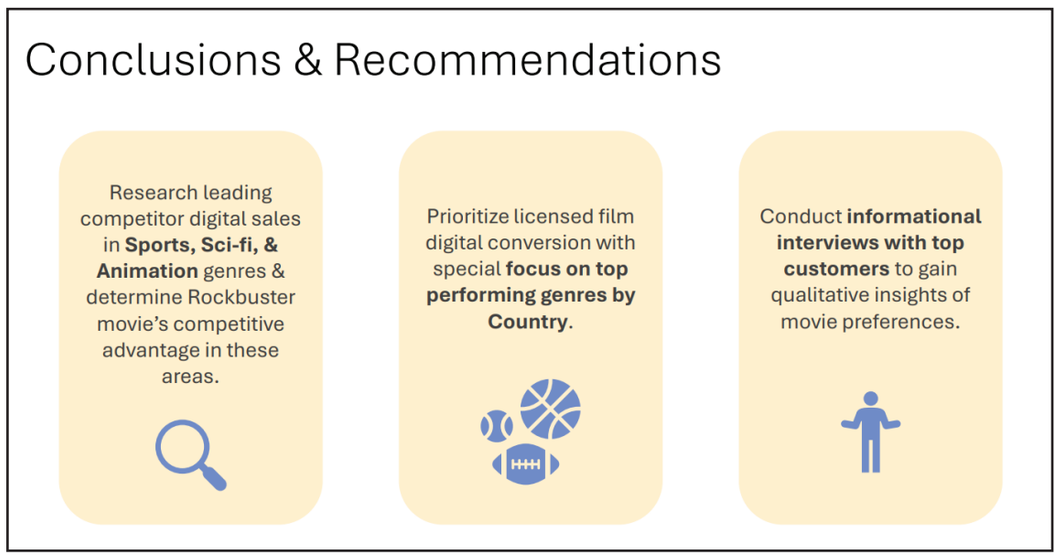


FIG. E: Concise recommendations are provided to Rockbuster to help them prepare for launching into the online streaming business.

TAKEAWAYS A variety of visuals were leveraged to help Rockbuster gain a clear understanding of where the majority of their business was deriving its sales from and how they might best accommodate those top selling regions.

Other useful information was pulled from the SQL joins such as top ten customers from the top ten cities within the top ten performing countries.

In order to help the company's chief stakeholders make strategic decisions for their company, clear and concise points were made regarding their next steps and plan of action.

Instacart Basket Analysis

Exploring Grocery Sales to Learn What's Trendy to Eat & When

17

Instacart Dataset HERE

Instacart Github Repository HERE

INTRO This was a theoretical Instacart project meant to analyze product sales via Products, Orders, and Customer datasets. Original data was accessed by Instacart and actual customer names & other variables such as prices had been imputed to protect information.

Leveraging Python using Jupyter Lab Notebooks, the Instacart data was cleaned, wrangled, and prepared for analysis.

One of the datasets contained over 32 million rows of data, which in turn required troubleshooting techniques in order to run the study smoothly. In the end, sampling 20% of the data was selected in order to conduct the analysis.

FIG. A: A frequency chart illustrates which age groups spend the most money; a clear increase is seen at the age 40 mark

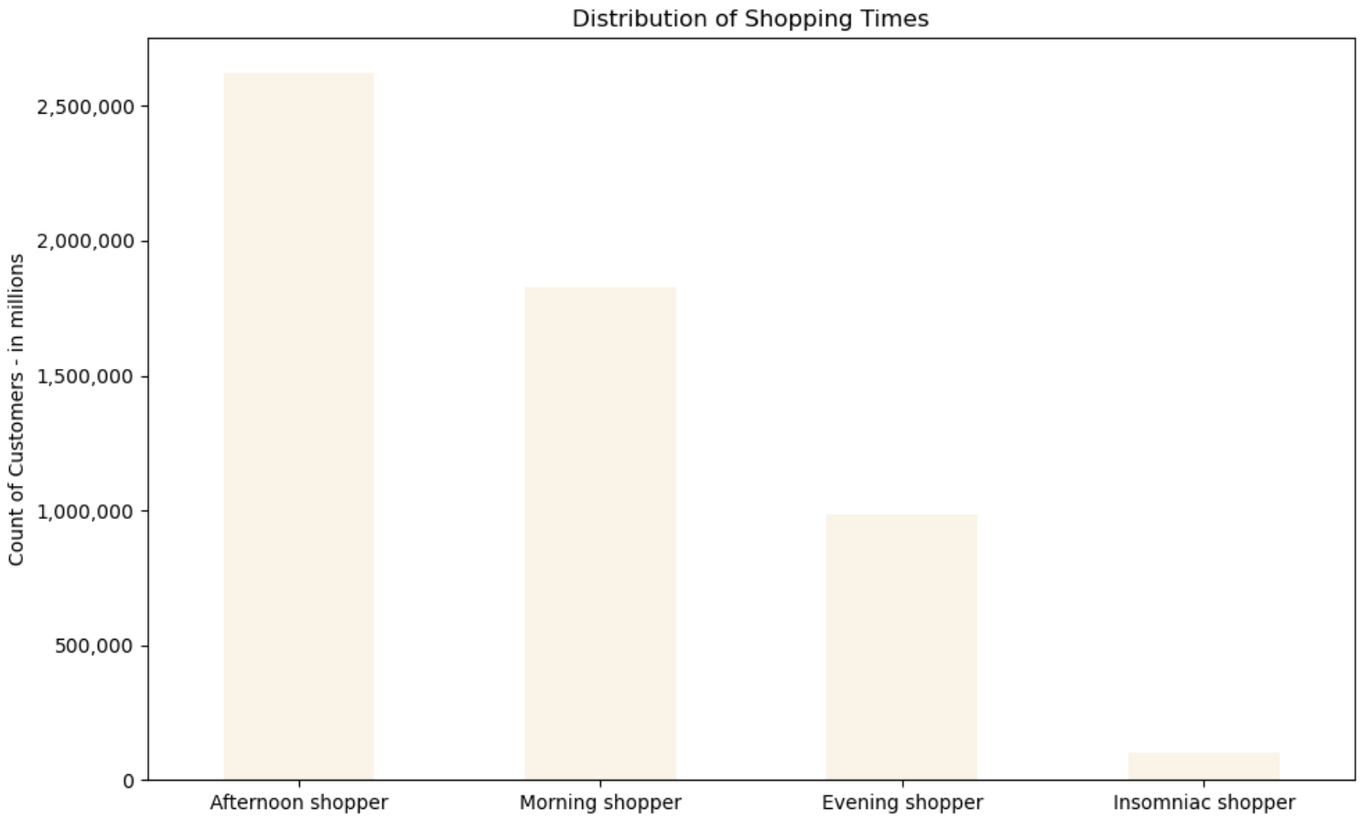


FIG. B: Here shoppers are arranged to groups by time of day they shop; though a small portion of the populous, Insomniac shoppers still must be accounted for.

HIGHLIGHT With millions of data insights at hand, you can imagine the challenge of being able to glean insights without the ability to group the variables into more meaningful categories.

With the help of data aggregation (calculating average, min, max, etc) and column flag creation (creating new data variables that consolidated age groups, customer loyalty, shoppers by time of day, etc), capturing a sense of which products customers were drawn to became much easier, and a narrative was much simpler to form.

Profile, Region, & Department Comparisons

- Age Groups vs Loyalty

[275]:
```
# Create crosstab of 'ages_flag' and 'loyalty_flag' columns

age_group_loyalty = pd.crosstab(
    ic_final['ages_flag'][ic_final['ages_flag'] != 'nan'],
    ic_final['loyalty_flag'][ic_final['loyalty_flag'] != 'nan'],
    dropna=True
)
```

Last executed at 2025-01-12 16:09:45 in 4.23s

[278]:
```
# Check output of age_group_loyalty crosstab (and add commas)

age_group_loyalty.style.format('{:,.0f}')
```

Last executed at 2025-01-12 16:09:48 in 6ms

[278]:

loyalty_flag ages_flag	Loyal customer	New customer	Regular customer
Adults 25 - 44	385,155	364,374	898,520
Middle Aged 45 - 64	383,680	364,220	890,169
Senior 65+	337,275	325,228	796,084
Under 25	139,665	134,777	332,645

FIG. C: Screenshot of a Jupyter Notebook script detaling the creation of age group by customer loyalty crosstabs

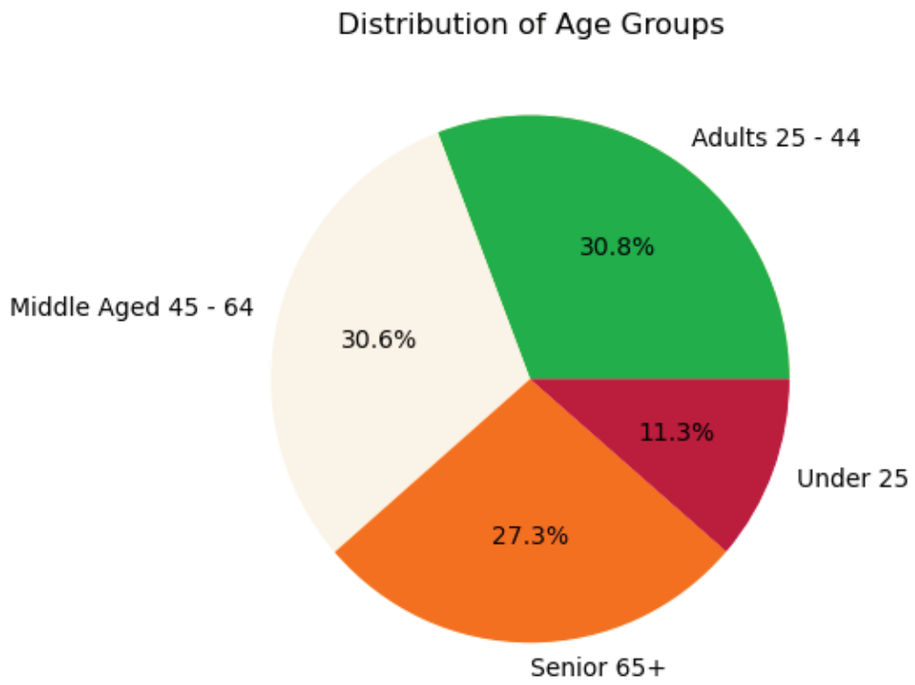
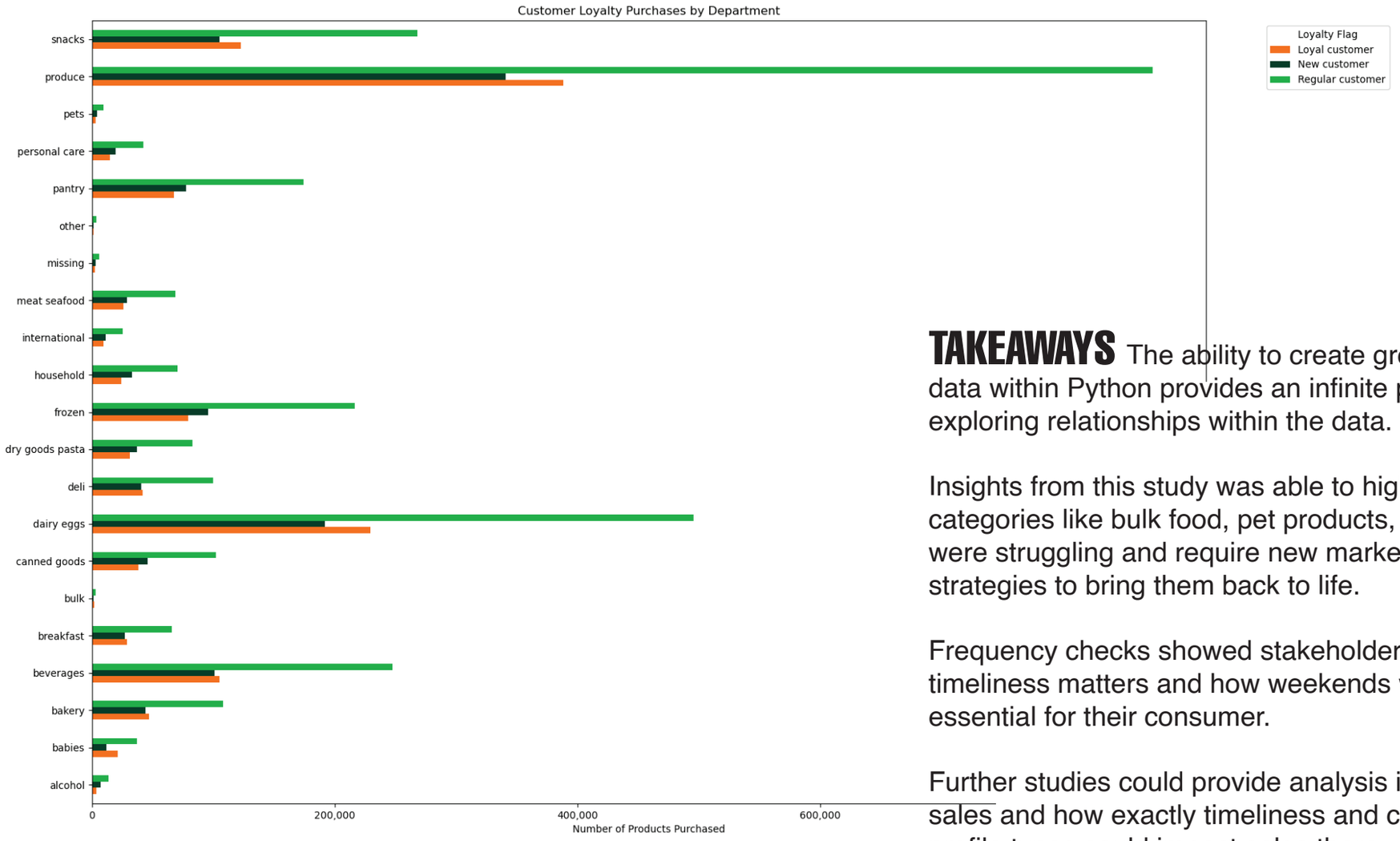


FIG. D: Customer ages were grouped into broader categories by creating new data variable columns in Python



TAKEAWAYS The ability to create groups and join data within Python provides an infinite possiblity for exploring relationships within the data.

Insights from this study was able to highlight how categories like bulk food, pet products, and alcohol were struggling and require new marketing strategies to bring them back to life.

Frequency checks showed stakeholders how timeliness matters and how weekends were essential for their consumer.

Further studies could provide analysis into regional sales and how exactly timeliness and customer profile types could impact sales there.

FIG. E: Customers are divided into loyalty segments (New, Regular, and Loyal) then categorized by their shopping preferences. Produce is a clear winner.

Cary, NC Real Estate

Discovering Property Features that Drive Real Estate Value

23

Cary Real Estate Tableau Storyboard HERE

Cary Real Estate Github Repository HERE

INTRO The following analysis was an opportunity to leverage more complex data analysis skills such as linear regression and k-means cluster analysis through machine learning.

With personal interest in Raleigh and surrounding surburb real estate, I utilized publicly available Cary, North Carolina real estate property data provided by the local government.

In depth information such as property location, zip code, geo location, price from most recent sale, acreage on deed, building square footage, land value, etc were provided.

The most useful information was land classification (residential under 10 acres, commercial gas station, recreational golf course, apartments, etc) which helped create larger property category groupings (residential, commercial, etc).

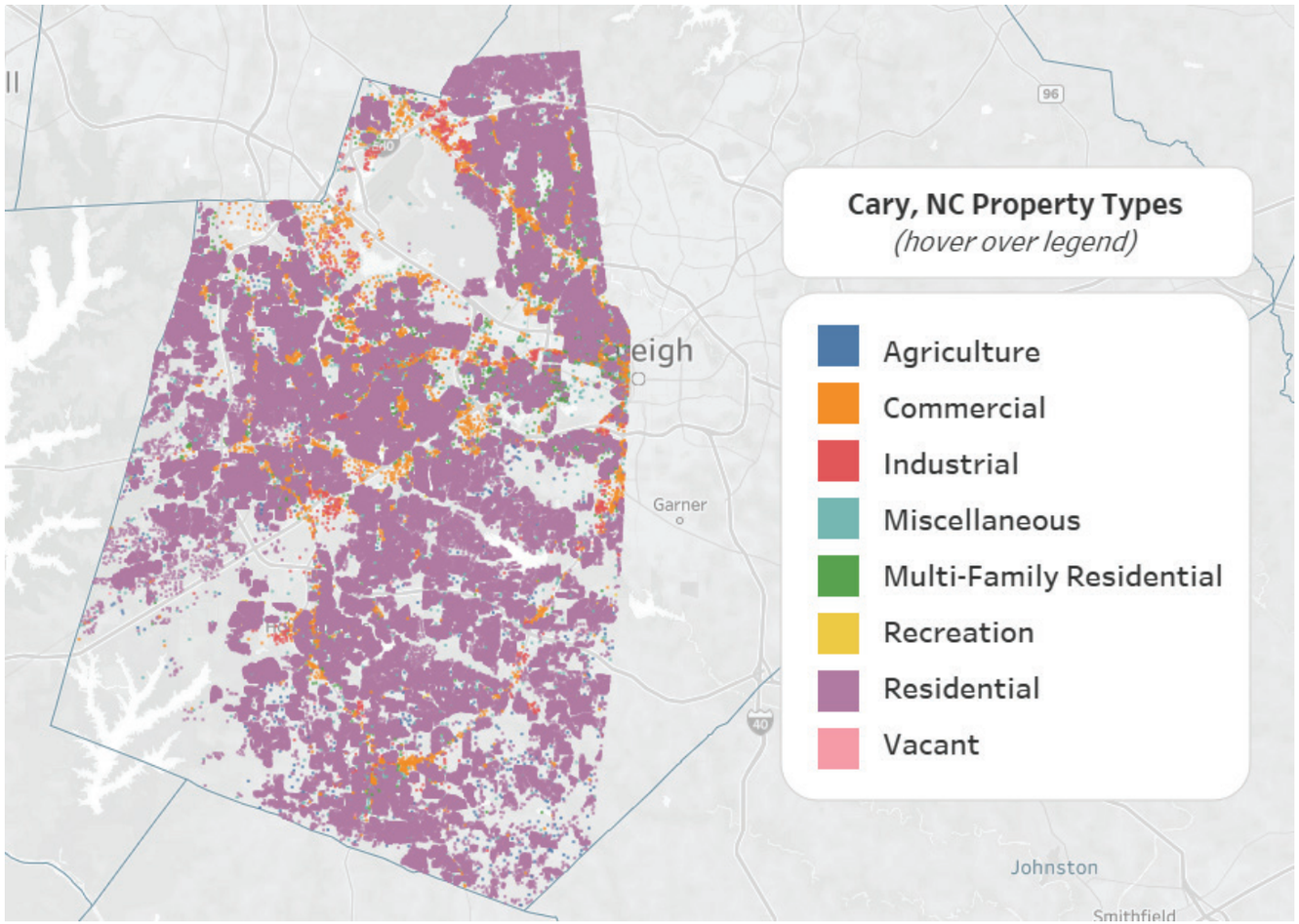


FIG. A: Interactive Tableau layer map that shows distribution of Cary, NC property types.

Find Max Value Property

```
[58]: # Find the property that is worth $738,378,563 (it's the airport..)

max_row_bldgvalue = df_2.loc[df_2['bldgvalue'] == 738378563]
max_row_bldgvalue
```

[58]:

	location	deedacres	landclass	totalstructures	totalunits	propertydesc	bldgvalue	landvalue
148502	2800 Airport Blvd	4584.26	Exempt	62.0	0.0	LOPT RALEIGH-DURHAM INTN'L AIRPORT 03-766-772	738378563	474053257

FIG. B: Exploratory data analysis led to the discovery of "outlier" values such as the Raleigh-Durham Intl. Airport (it had the highest building value).

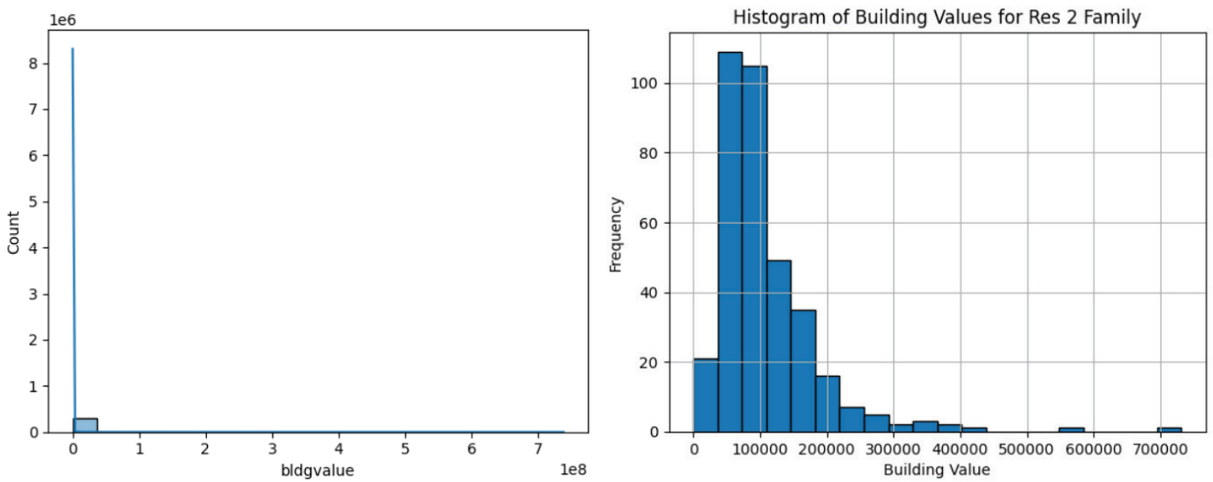


FIG. C: Left image shows histogram of building values across the entire dataset; the unusual distribution indicates outlier values. Image on the right is the same plot, however narrowing into the "2 Family Residential" property type, revealing much more informative data information.

HIGHLIGHT The initial exploratory phase showcased the importance of thoroughly understanding data outliers.

The Cary Real Estate dataset contains more than just residential properties--some of these values skewed the initial distribution analyses.

By digging in a bit further, it revealed interesting facts, such as the building with the highest value being the airport or discovering the oldest recorded historical property.

Proper review of outliers can reveal important information and indicates that grouping into property categories will help fine-tune the analysis and eliminate the need to cut out relevent information.

HIGHLIGHT From initial correlation matrix testing, the following hypothesis was developed:

The larger a property's acreage or building square footage, the higher the property value.

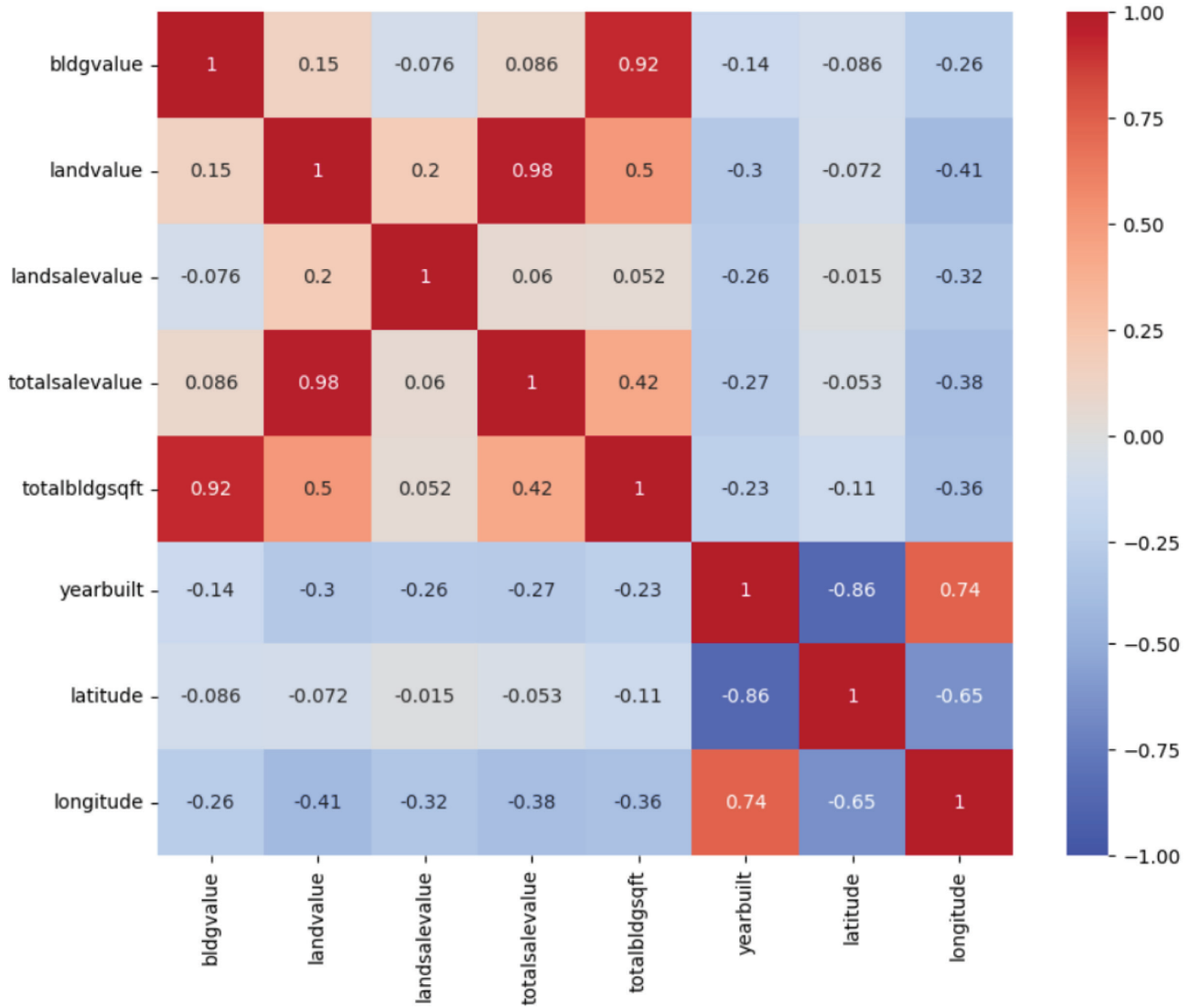


FIG. D: Acreage variable was not included in this particular correlation matrix, but land value and total building square footage seemed to have stronger potential correlations with the most recent total sale value as well as land or building value, which is why the initial hypothesis was made.

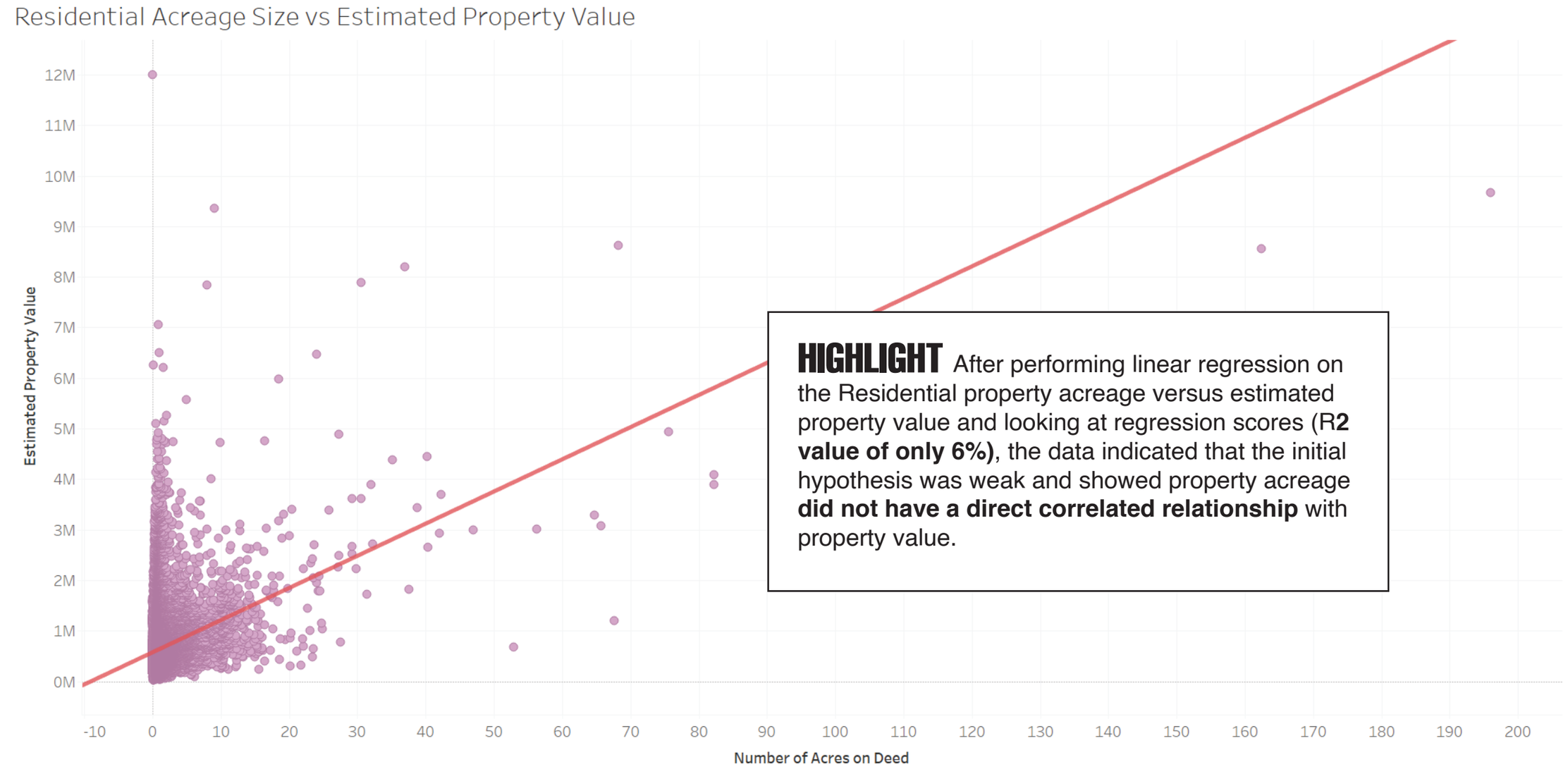


FIG. E: Number of acres (across Residential properties) versus the estimated property value (a variable derived from land and building values provided in the dataset), only had about a 6% potential correlation in the linear regression model.

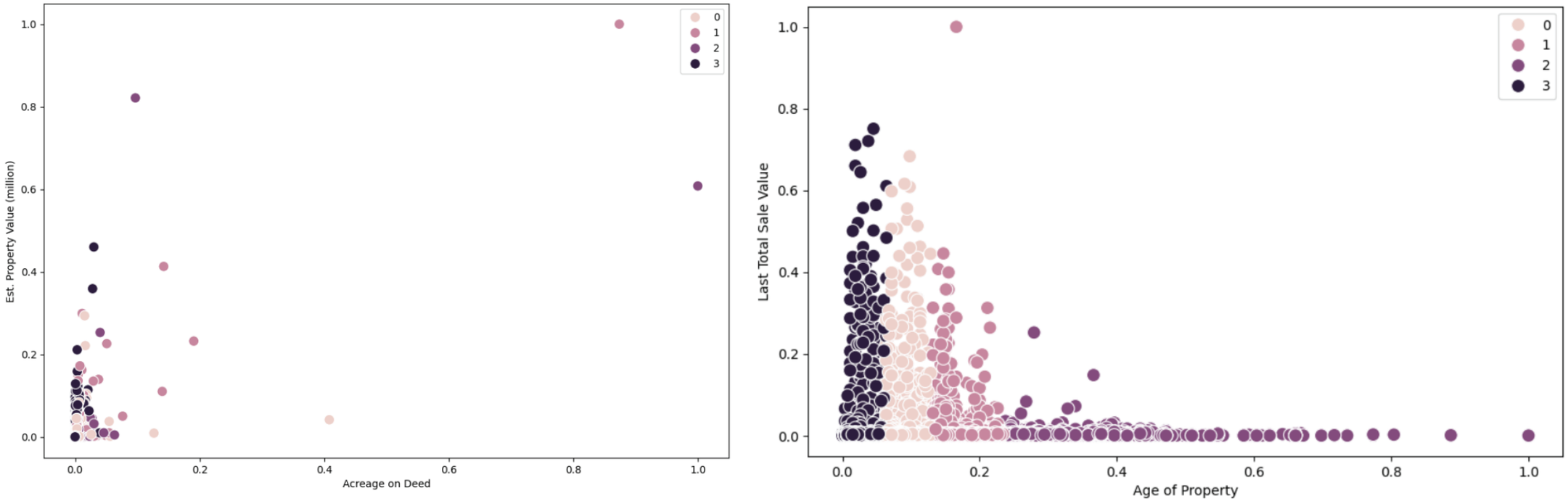


FIG. F: K-Means clustering performed on property features versus property value. Left image shows no clear clusters with the property acreage and estimated property value. Right image shows more distinct clusters by age of property compared to the most recent total property sale value.

HIGHLIGHT Since linear regression indicated no distinct direct, linear relationships between variables, a k-means cluster machine learning analysis was applied to the entire dataset (incuding all property class types) to see if there were any “hidden” relationships not initially detected. Exploration of the “property age” variable showed a possible multi-variate relationship between property age & price.

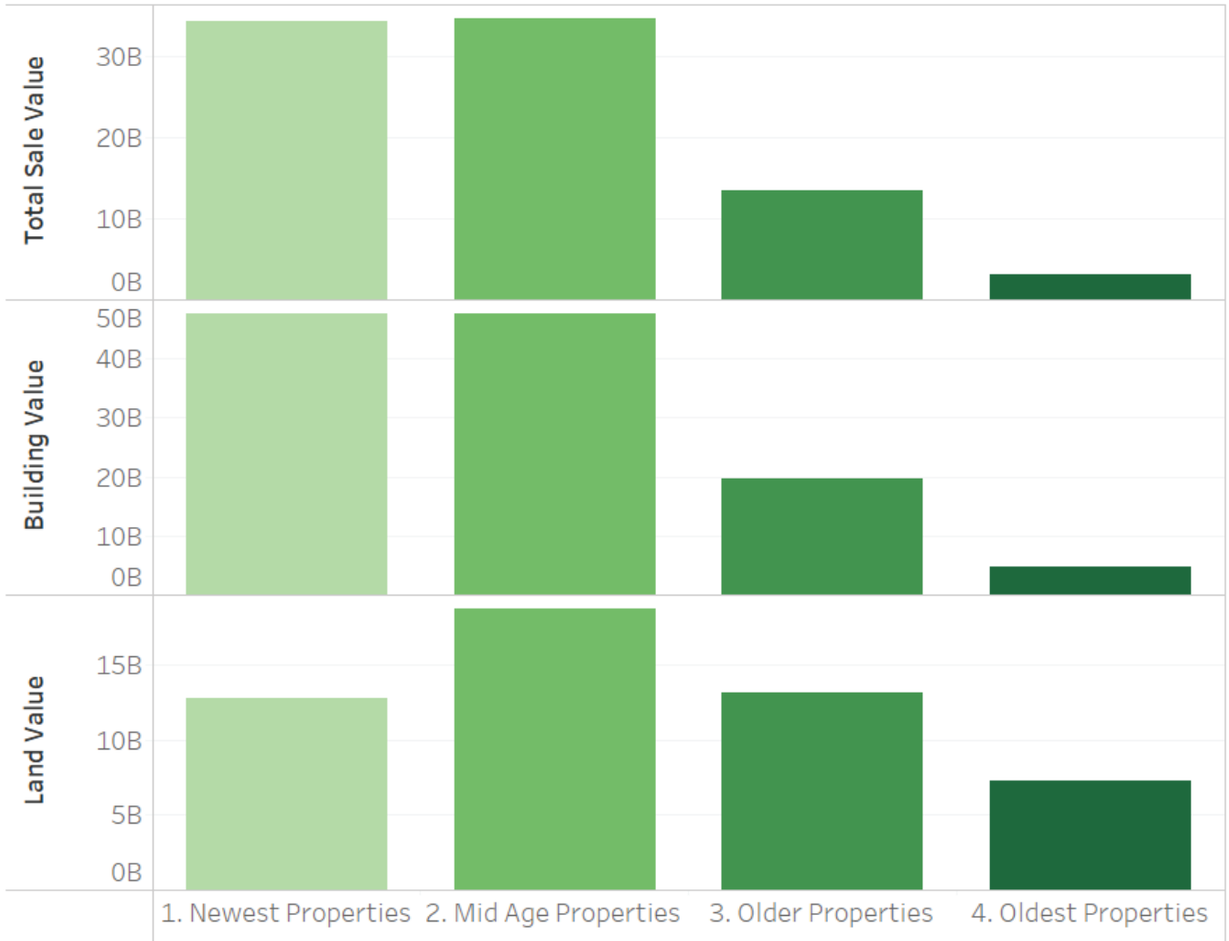


FIG. G: Bar chart comparing property age groupings (dervied from k-means cluster analysis) against total sale value, building value, and land value. Data reveals older properties have potential to be valued less than newer properties.

TAKEAWAYS Initial analysis of Cary, NC real estate indicated that **property age could have a potential, non-linear relationship to the property's value**.

Within this particular dataset (after cleaning), showed the **newest properties** had a median age of 7 years and **median most recent total sale value of $424,500**.

Oldest homes had a median value of 74 years and **median most recent total sale value of $148,000.**

Further analysis could be used to discover what features might help a home retain its value (location, recent renovation, proximity to schools/work/commerce, etc).

Property age could be further investigated as well--particularly the time period between when most recent sales took place versus property value.

Data limitation could be that older properties had a longer time period between being sold and current date of analysis compared to newer properties, resulting in a potential recency bias with the change of property values over time.

CONTACT

https://www.linkedin.com/in/kirstencurrie/

https://github.com/kirstencurrie

kirstenlynncurrie@gmail.com