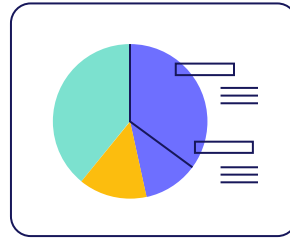
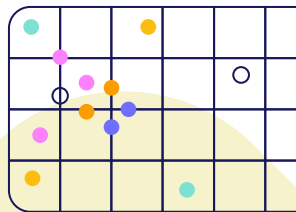


Janelle Sousley

Data Analysis Portfolio

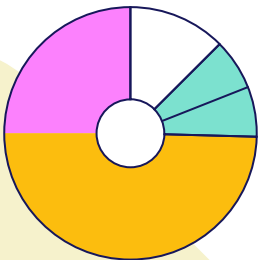


ABOUT ME

Hello! I am Janelle Sousley and I am an experienced data analyst with a proven track record in the dental industry, leveraging scientific methodologies and data-driven insights to revolutionize oral healthcare for thousands of patients.

Combining data-driven strategies with exceptional communication skills and a passion for uncovering actionable insights, I am dedicated to driving business performance and fostering success in any analytical role.

Thank you for taking the time to view my portfolio!



CONTENTS



Instacart

Python-based project
on consumer behavior,
trends, and market
analysis



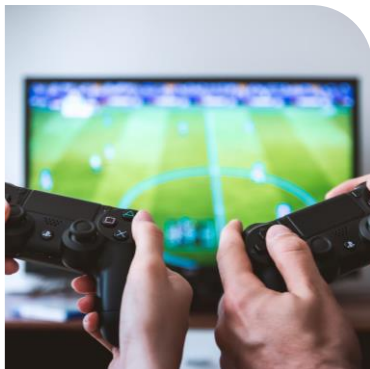
Rockbuster

International business
analysis of online video
rental services through
SQL



Influenza

National medical
staffing distribution
based on historical
trends



GameCo

Global marketing
analysis of the gaming
console industry



Pig E. Bank

Predictive analysis of
customer retention risk
for a global finance
service company



King County

Market analysis for
home sales in one of
the most populous
counties in the U.S.

INSTACART

An online grocery store that operates through an app. While already having good sales, executives want to uncover more information about their sales patterns.

OBJECTIVE

Perform an exploratory analysis of customers' behavior and sales patterns to derive insights and suggest strategies for better marketing and sales segmentations.

PROJECT DATA

- [Project Brief](#)
- [Instacart Dataset](#) provided by Kaggle
- [Customer Dataset](#) provided by CareerFoundry
- [Data Dictionary](#)

LIMITATIONS

- Data only contains records from 2017.
- Customer demographics are limited to income, age, family size, and marital status.

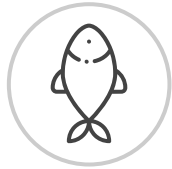
TECHNIQUES APPLIED

- Data Wrangling & Subsetting
- Data Consistency Checks
- Combining & Exporting Data
- Deriving New Variables
- Grouping Data & Aggregating Variables
- Data Visualization with Python
- Excel Reporting

TOOLS



APPROACH & METHODOLOGY



Organizing Data

Addressed mixed-type variables, missing values and duplicate values.
Wrangled data and created subsets of data.



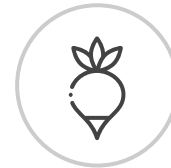
Aggregating Data

Grouped and aggregated data to supplement variable derivation.



Deriving Variables

Created variables and user-defined functions such as loyalty groups, pricing groups, etc to generate new insights for analysis.



Python Visuals & Excel

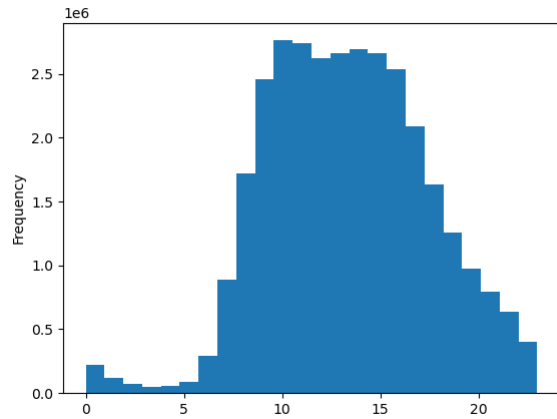
Used different libraries in Python to create visuals.
Consolidated results in Excel.



BUSINESS ANALYSIS

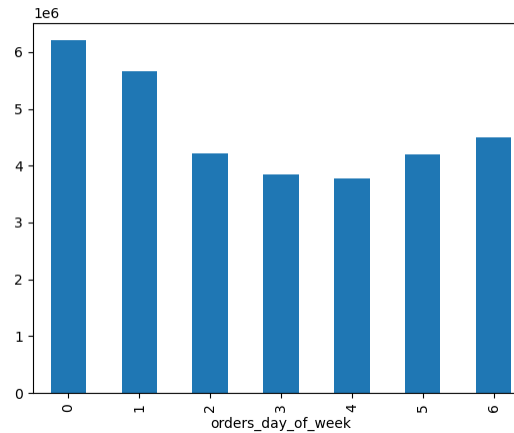


Purchase frequency highest from **9am to 4pm**.



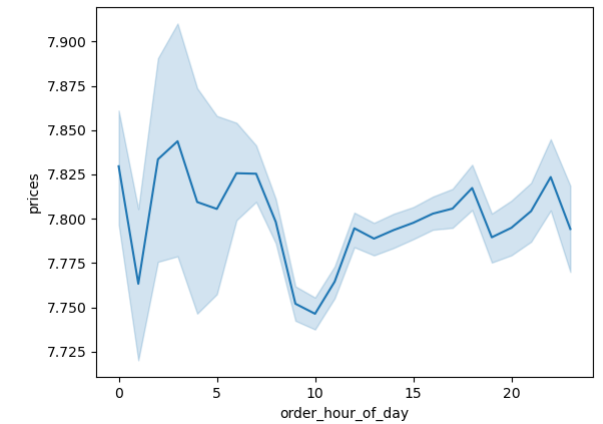
```
# creating histogram of order hour of the day
hist = ords_prods_custs['order_hour_of_day'].plot.hist(bins = 24)
```

Saturday, Sunday, and Friday are the busiest days.



```
# creating bar chart for orders day of week
ords_prods_merge['orders_day_of_week'].value_counts().plot.bar()
```

Busiest hours of the day see the **lowest prices** of goods.



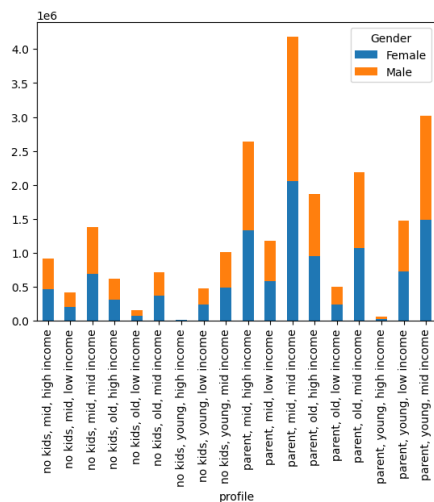
```
# Line chart for prices and order hour of day
line = sns.lineplot(data = df_2, x = 'order_hour_of_day', y = 'prices')
```



CONSUMER ANALYSIS



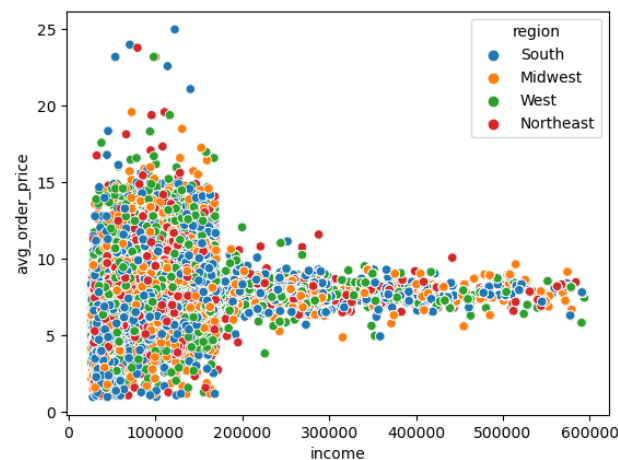
Parents with **mid income** were the highest users.
Male to female ratio was almost **50/50**.



```
# group by profile and gender
sb1 = ords_prods_custs_clean.groupby('profile')['Gender'].value_counts()

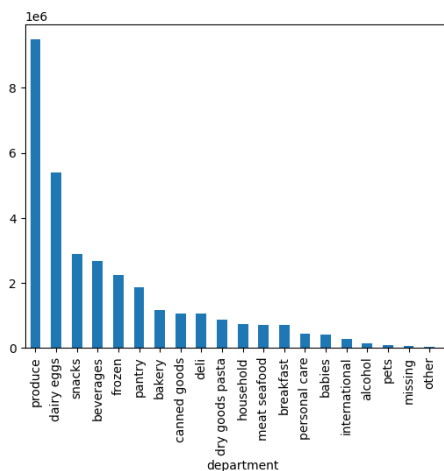
# creating stack bar of profile and gender
stackbar1 = sb1.unstack().plot.bar(stacked=True)
```

Average order **spending** was **higher** in lower to mid income, leaving untapped potential in high income customers.



```
scatter1 = sns.scatterplot(x = 'income', y = 'avg_order_price', hue = 'region',
                           data = ords_prods_custs_clean)
```

Customers **top 5 products**:
produce, dairy/eggs, snacks, beverages, frozen.



```
# creating bar chart for department counts
bar3 = ords_prods_merge['department'].value_counts().plot.bar()
```



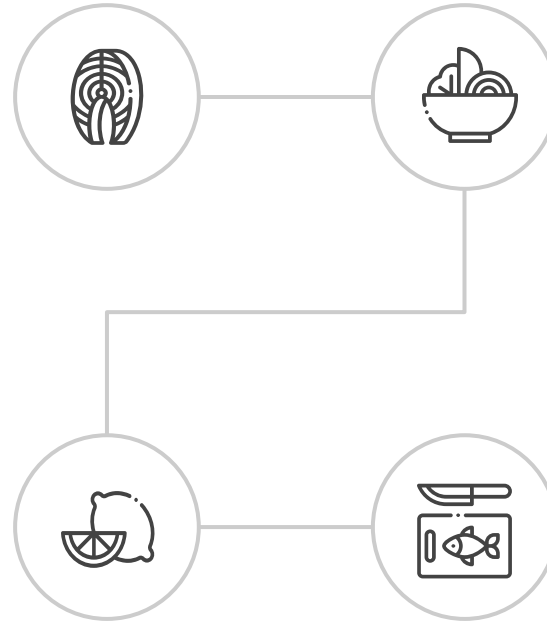
RECOMMENDATIONS

Marketing

- Saturday, Sunday, and Friday are great times to run ads as these days are when most orders are placed.
- Focusing advertisement on highest departments of produce, dairy/eggs, snacks, beverages and frozen items.
- Advertisement for lower priced deals between 10am and 5pm as this is the time most orders are placed while also being the times lower priced items are sold.

Customers

- Leverage loyalty with a loyalty rewards program.
- Further analysis of young customers with high income as they are the lowest customer base, but high potential for sales.

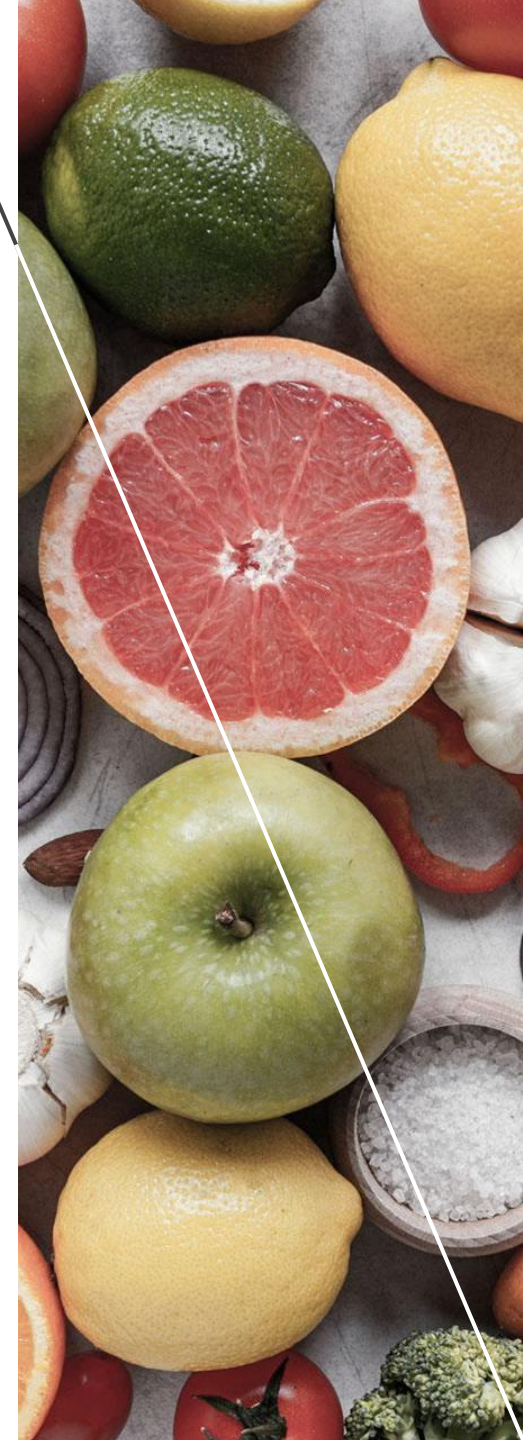


Sales


- Increased efforts and sales in high product sellers at regular intervals throughout the day to keep these items visible.
- Regular sales and advertisements in low selling departments coupled with high department sellers should help boost these items sold.

Partnerships

- Focus on new partnerships with stores that sell bulk items as it is the lowest product seller but high potential for families.
- Training for shoppers should include how to pick quality food items (especially produce, as it is the top selling product).



Rockbuster Stealth LLC

A movie rental company that used to have  stores around the world. The management team would like to use their existing movie licenses to launch on online video rental service in order to stay competitive.

OBJECTIVE

Provide insights on current business standings  to help with their 2020 launch strategy of online video services.

PROJECT DATA

- Project Brief
- Dataset provided by CareerFoundry

LIMITATIONS

- Data covers internal records of stores, customers, payments, inventory, films, and more.

TECHNIQUES APPLIED

- Relational Databases
- Data Dictionary
- Relational Databases
- SQL for Data Analysts
- Database Querying in SQL
- Filtering Data
- Summarizing & Cleaning Data in SQL
- Joining Tables of Data
- Performing Subqueries
- Common Table Expressions
- Presenting SQL Results



TOOLS



APPROACH & METHODOLOGY



Assessing Database

Created an Entity Relationship Diagram (ERD) through DbVisualizer and described the structure of the database for analysis.
Created a data dictionary for user accessibility.



Data Extraction & Summarization

Used JOINS, subqueries and common table expressions (CTE's) to provide a comprehensive understanding of business standings.

Data Cleaning

Performed common SQL commands known as CRUD (create, select/read, update, delete) to ensure clean and consistent format for filtering as well as summarizing data for output accuracy.



Data Visualization & Storytelling

Used results of SQL queries to generate visualizations and a storyboard in Tableau for the final presentation to stakeholders.



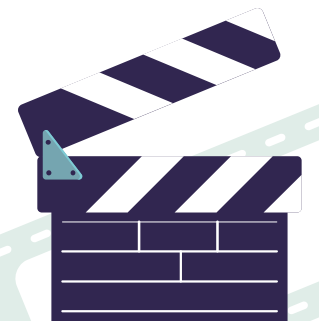
BUSINESS OVERVIEW



TOTAL FILMS	1000
TOTAL COUNTRIES	109
TOTAL ACTIVE CUSTOMERS	584
PRIMARY FILM LANGUAGE	English
TOTAL REVENUE	\$61,312
TOP 3 FILM GENRES	Sports, Sci-Fi, Animation

ROCKBUSTER FILM TRENDS

	MINIMUM	MAXIMUM	AVERAGE
RENTAL DURATION	3 days	7 days	5 days
RENTAL RATE	\$0.99	\$4.99	\$2.98
FILM LENGTH	46 minutes	185 minutes	115 minutes

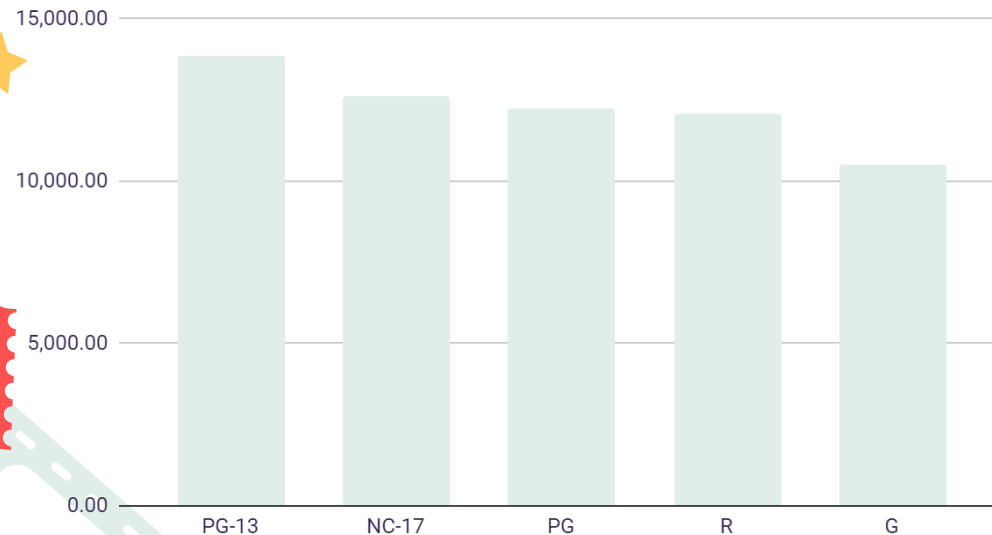


BUSINESS ANALYSIS

PG-13 had highest view rate accounting for **22%**.

```
--rating revenue
SELECT f.rating,
       SUM(p.amount)
FROM payment p
INNER JOIN rental r ON r.rental_id = p.rental_id
INNER JOIN inventory i ON i.inventory_id = r.inventory_id
INNER JOIN film f ON f.film_id = i.film_id
GROUP BY f.rating
ORDER BY SUM(p.amount) DESC
```

Revenue per Rating

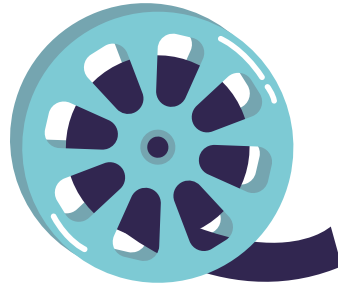


Top 3 Countries

- #1. India** 10% revenue, 10% customers (\$6,035 & 60 customers)
- #2. China** 9% revenue, 9% customers (\$5,251 revenue & 53 customers)
- #3. U.S.** 6% revenue, 6% customers (\$3,685 revenue & 36 customers)

```
-- customer count and total payment received against each country
SELECT country,
       COUNT(DISTINCT A.customer_id) AS customer_count,
       SUM(amount) AS total_payment
FROM customer A
INNER JOIN address B ON A.address_id = B.address_id
INNER JOIN city C ON B.city_id = C.city_id
INNER JOIN country D ON C.country_ID = D.country_ID
INNER JOIN payment E ON A.customer_id = E.customer_id
GROUP BY country
```

RECOMMENDATIONS



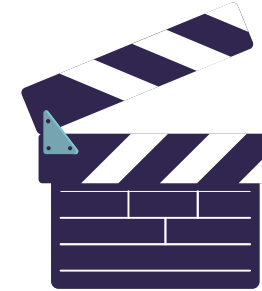
Marketing

- Upgrade marketing by building a social media presence with heavy focus on Asia and American countries.
- Media should be announcing the switch to digital platforms.



Pricing & Inventory

- Focus on select inventory with films in the PG-13 rating and sports, sci-fi and animation genres.
- Set attractive rental prices averaging around \$3.00.



Customers

- Surveys sent to existing customers to find out if streaming services would be desired and what type of content they would be interested in (genres, price points, rewards etc.).
- Leverage existing customers with tiered loyalty services including referral rewards benefits.
- Analysis on customer internet infrastructure.



Preparing for Influenza Season in the U.S.

A medical staffing agency that provides temporary workers to clinics and hospitals on an as-needed basis. They aim to allocate staff across all 50 U.S. states during influenza season.

OBJECTIVE

Analyze historical influenza trends in the U.S. to assist the medical staffing agency in the deployment of temporary healthcare personnel for the upcoming season.

PROJECT DATA

- [Project Brief](#)
- [Influenza Deaths Dataset](#) provided by [CDC](#)
- [Population Dataset](#) provided by [U.S. Census Bureau](#)
- [Influenza Visits](#) and [Lab Tests](#) Datasets provided by [CDC](#)
- [Influenza Vaccination Survey in Children](#) provided by [CDC](#)

LIMITATIONS

- 82% of influenza mortality data entries were suppressed for patient confidentiality.
- Death records identify a single underlying cause of death (influenza-related may not be counted).
- Datasets are from 2009 to 2017
- Data is limited to clinics and hospitals that chose to participate.

TECHNIQUES APPLIED

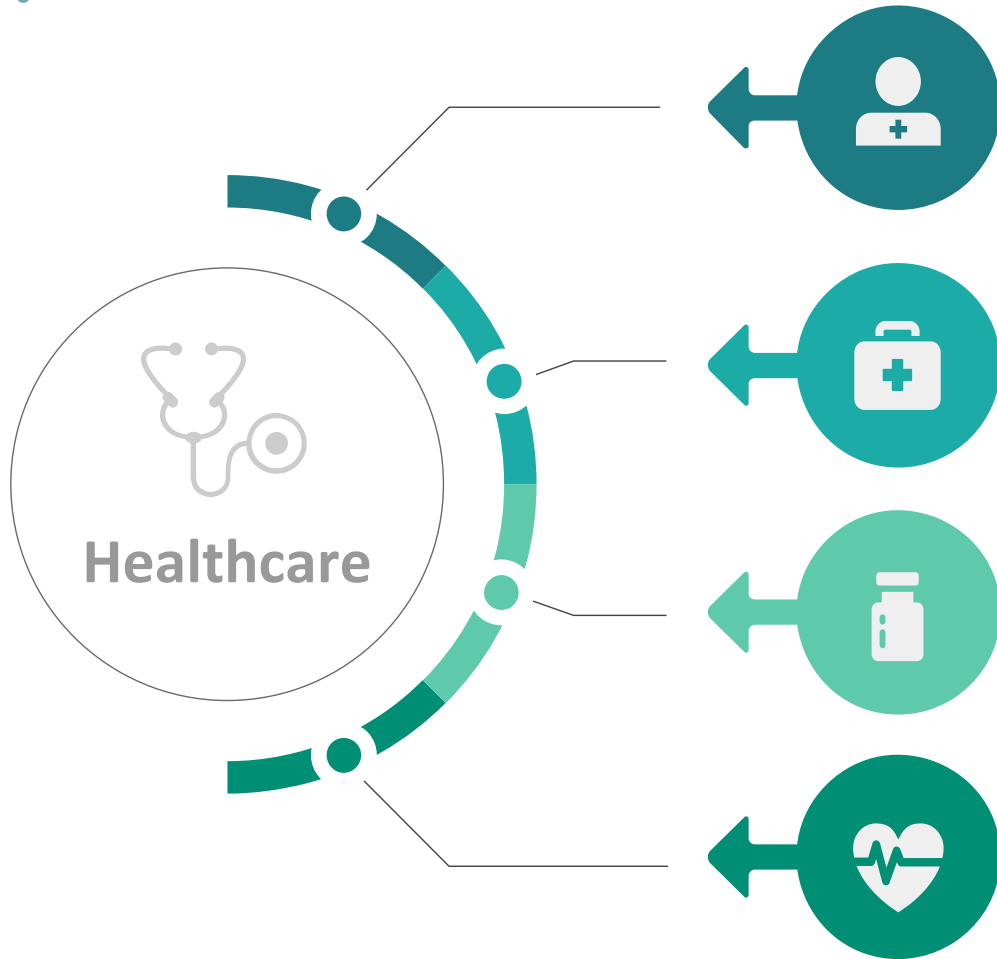
- Translating business requirements
- Data cleaning
- Data integration
- Data transformation
- Statistical hypothesis testing
- Visual analysis
- Forecasting
- Storytelling in Tableau
- Presenting results to an audience

TOOLS





APPROACH & METHODOLOGY



Designing Data Research Project

- Created business questions to help formulate a research hypothesis as a guideline for the analysis
- Prepared a project management plan to keep track of progress.

Data Preperation

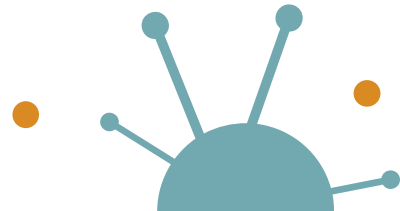
- Explored datasets for relevancy, integrity, and completeness.
- Cleaned and transformed data such as grouping, filtering, sorting, transposing etc.
- Integrated data from multiple sources and derived new variables.

Statistical Analysis & Hypthesis Testing

- Conducted visual analysis utilizing various plot types.
- Performed correlation tests between variables using linear regression.
- Tested hypothesis with two-sample t-test.

Data Visualization & Storytelling

- Designed and sequenced visuals using storytelling priciples.
- Created interactive Tableau dashboard.
- Presented findings via recoreded video presentation.



STATISTICAL ANALYSIS

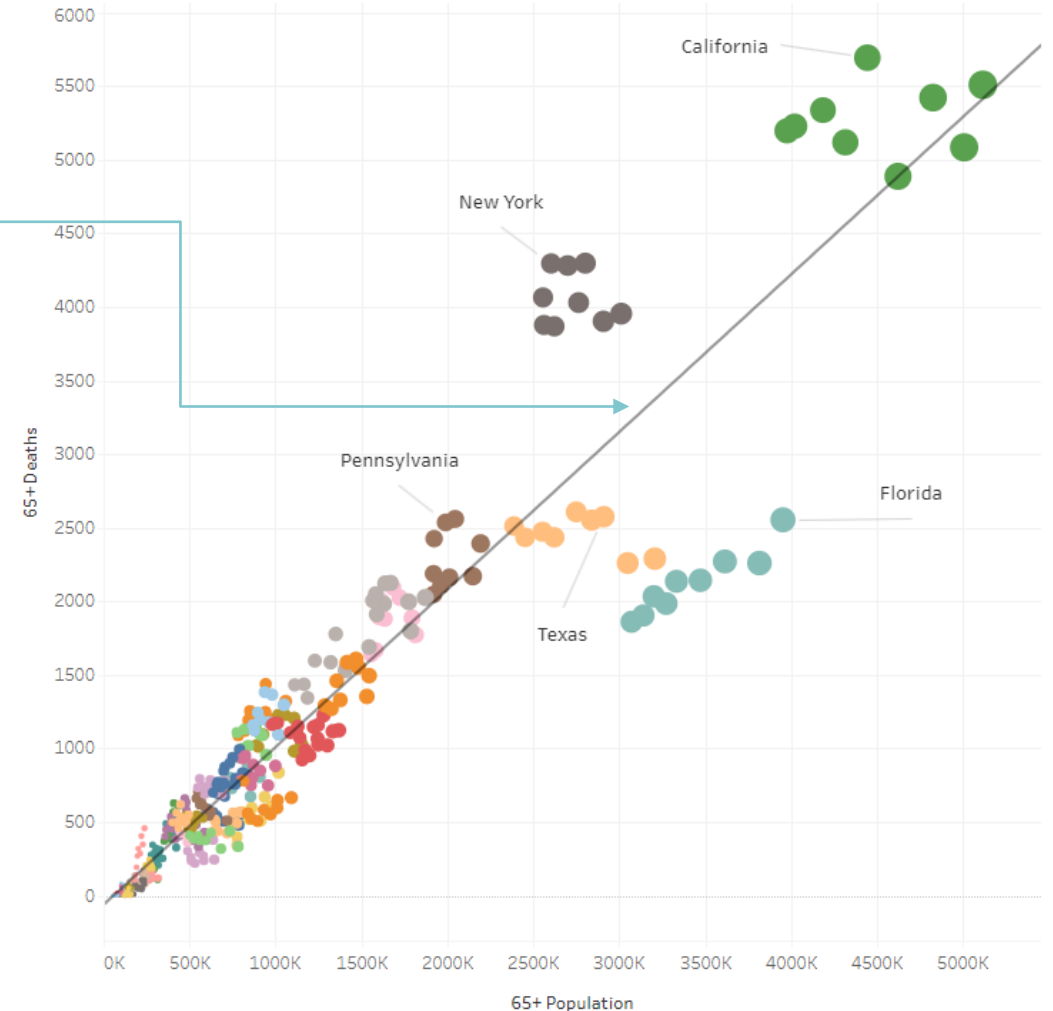
People **aged 65 and older** had the **highest mortality rate** among all age groups.

There is a **strong correlation coefficient of 0.9** between those aged 65+ and death.

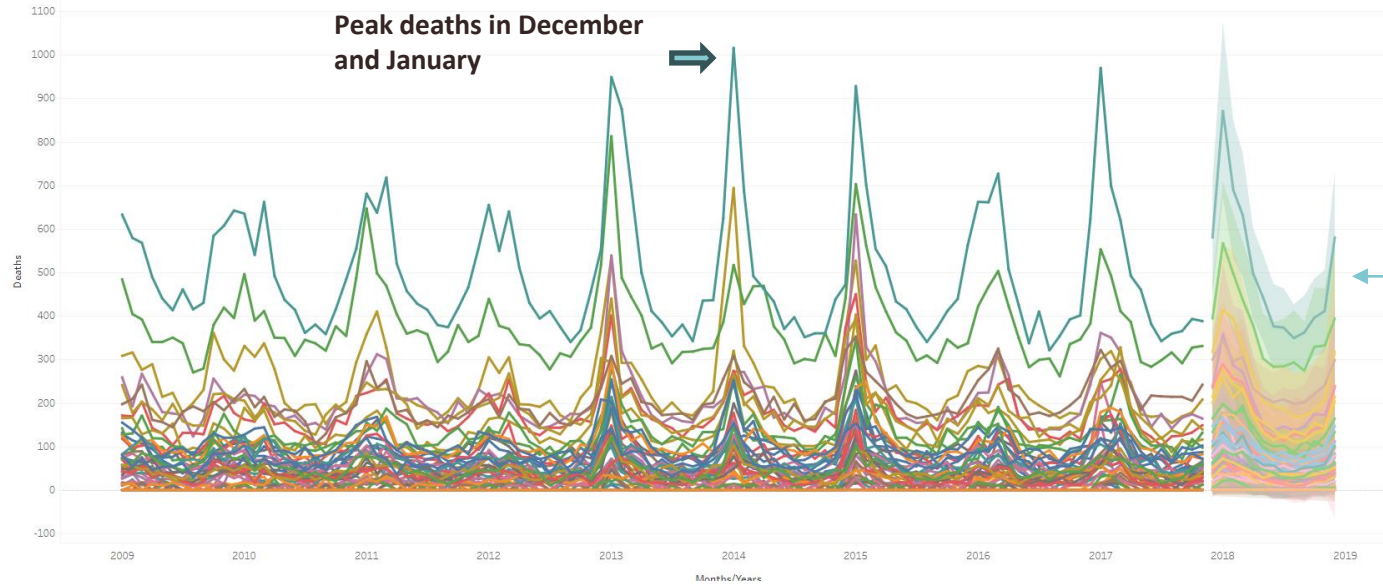
Hypothesis testing was done via inferential statistics to test confidence level of this analysis.

There is a **99% certainty** that influenza deaths for people over 65 are higher than those under 65.

t-Test: Two-Sample Assuming Unequal Variances		
	Age under 65	Age over 65
Mean	5.29602E-05	0.004584406
Variance	4.45354E-09	4.79123E-06
Observations	459	459
Hypothesized Mean Difference	0	
df	459	
t Stat	-44.33205296	
P(T<=t) one-tail	2.74E-168	
t Critical one-tail	1.648180137	
P(T<=t) two-tail	5.4743E-168	
t Critical two-tail	1.965145755	



TIMING & VULNERABLE GROUPS ANALYSIS

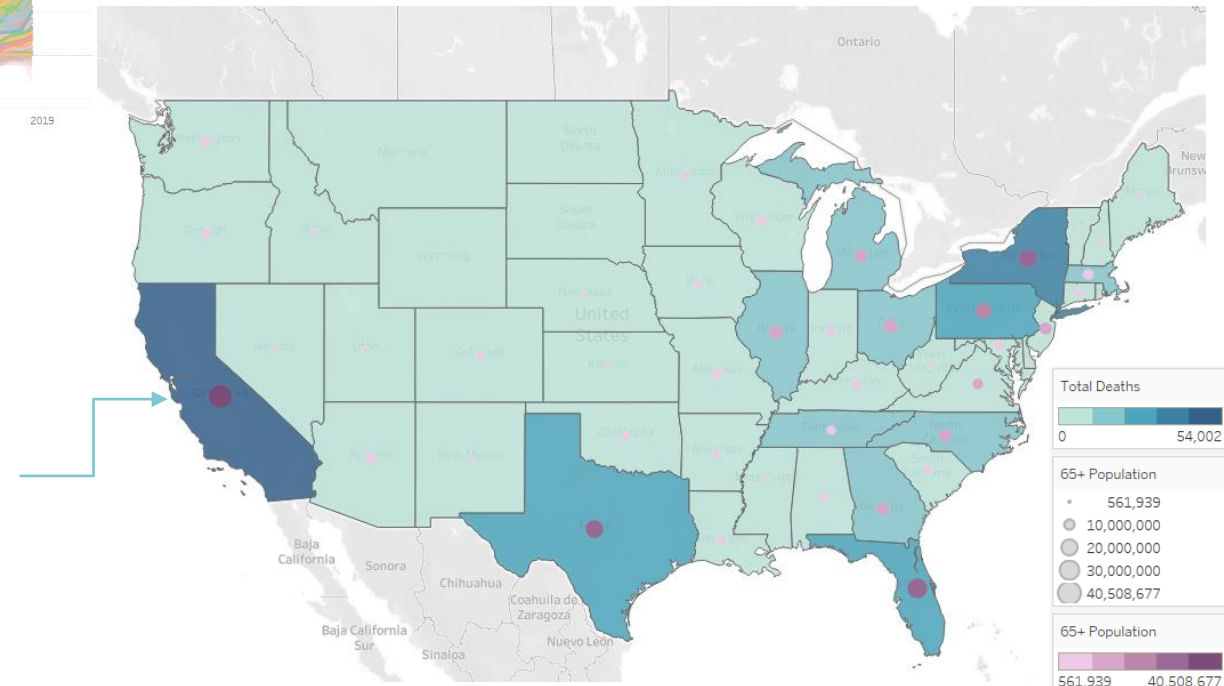


Influenza peak months are **December through February**.

Seasonality forecast shows what the death rate is predicted to be based off previous years.

California, New York, Texas and **Florida** are top states with both population over 65 and significant death rates.

California was number one in both ages over 65 and death rate.



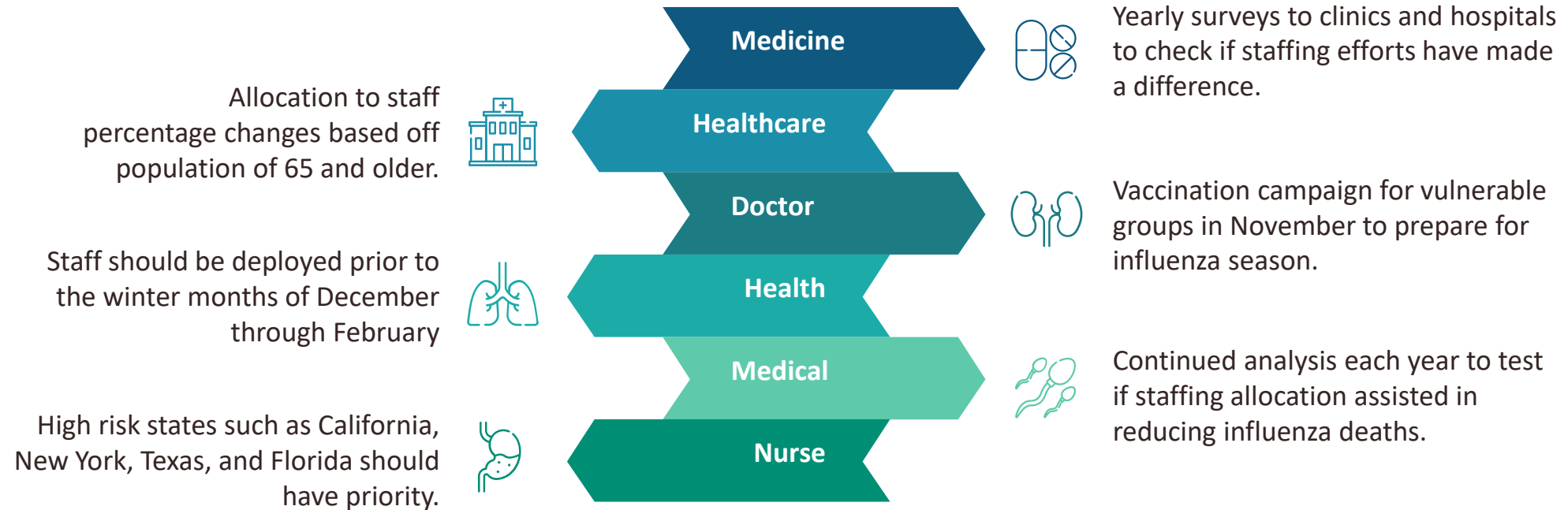
STAFF CHANGES IN PROVIDERS

Calculated total number of staff in each location and recommended allocated changes based off the states 65+ population.

Suggested Change in Providers by Population 65+ Ranked (2017)

State	# of Providers	+ or - Number of Provid..	Percentage of Change ..	% of Current Providers	Suggested % of Provid..	% of Change	Change of Providers Sc..
South Carolina	642	863	134%	0.7%	1.6%	0.9%	51
Iowa	471	466	99%	0.5%	1.0%	0.5%	50
Missouri	891	863	97%	1.0%	1.9%	0.9%	49
California	5,156	4,707	91%	5.6%	10.7%	5.1%	48
Idaho	248	204	82%	0.3%	0.5%	0.2%	47
Connecticut	628	482	77%	0.7%	1.2%	0.5%	46
Colorado	856	653	76%	0.9%	1.6%	0.7%	45
Indiana	1,122	805	72%	1.2%	2.1%	0.9%	44
Arkansas	570	385	68%	0.6%	1.0%	0.4%	43
New Jersey	1,683	956	57%	1.8%	2.9%	1.0%	42
North Carolina	1,911	1,068	56%	2.1%	3.2%	1.2%	41
Minnesota	981	511	52%	1.1%	1.6%	0.6%	40
Wisconsin	1,152	600	52%	1.3%	1.9%	0.7%	39
Ohio	2,718	742	27%	3.0%	3.8%	0.8%	38
Kentucky	1,086	253	23%	1.2%	1.5%	0.3%	37
Texas	5,104	1,080	21%	5.5%	6.7%	1.2%	36
Oklahoma	1,003	127	13%	1.1%	1.2%	0.1%	35

RECOMMENDATIONS





GameCo

A video game company interested in exploring historical sales trends in development of new games.

OBJECTIVE

Perform a descriptive analysis to gain insights into the current video game landscape for a marketing and sales' team 2017 planning.

PROJECT DATA

- [Project Brief](#)
- [Video Games Sales Dataset](#) provided by [VGChartz](#)
- [VGChartz methodology](#)

LIMITATIONS

- Dataset only tracks total number of units sold and not financial figures.
- 2016 is the latest year logged with partial records.

TECHNIQUES APPLIED

- Data Integrity, Quality, and Consistency Assessment
- Pivot Tables
- Grouping Data
- Calculated Fields
- Summarizing Data
- Descriptive Analysis
- Visualizing Results in Excel
- Presenting Results in PowerPoint

TOOLS



APPROACH & METHODOLOGY

Data Cleaning

- Removed duplicates and irrelevant values while also imputing missing figures with mean values.
- Normalized text formats and corrected typos.

Group & Summarize

- Used excel pivot tables to group and summarize data with filters and calculated fields.

Descriptive Analysis

- Calculated basic exploratory data analysis on distributions, outliers and tendencies.

Excel Visuals

- Merged dataframes and used data findings to generate visualizations using Excel charts.

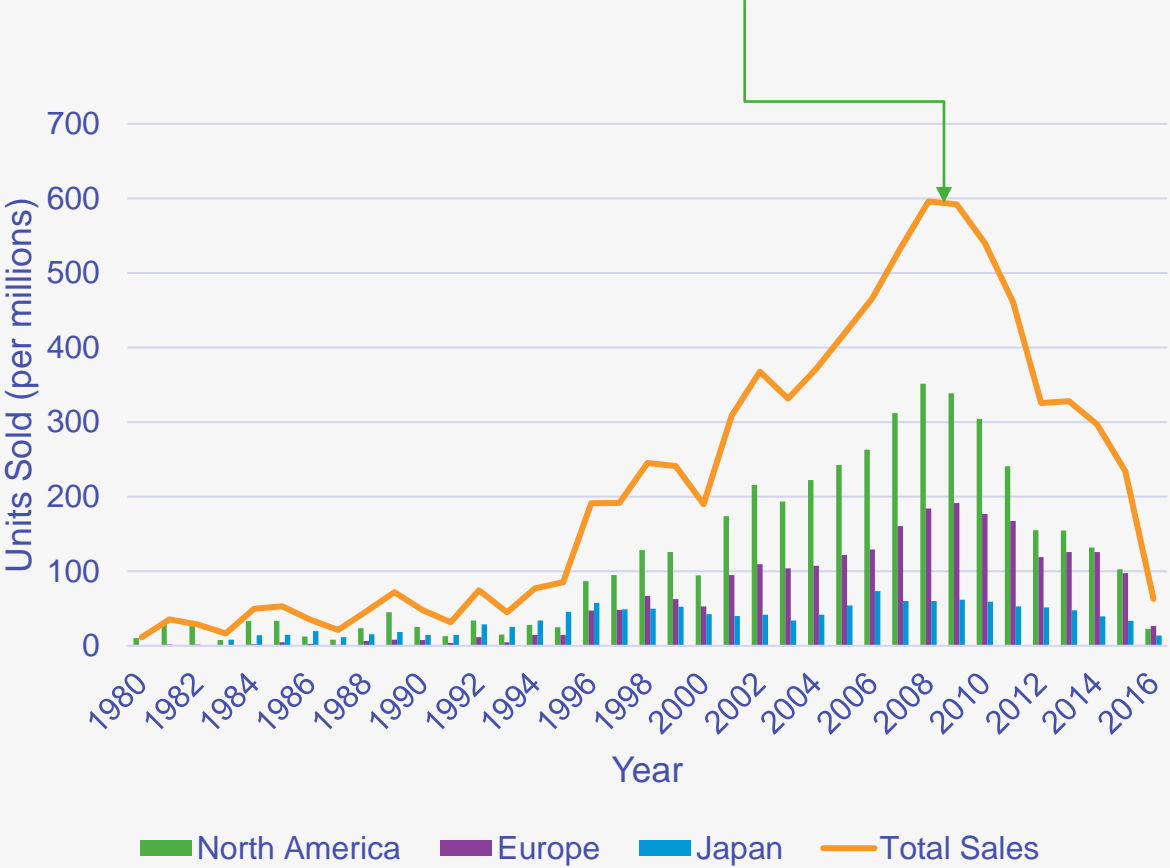




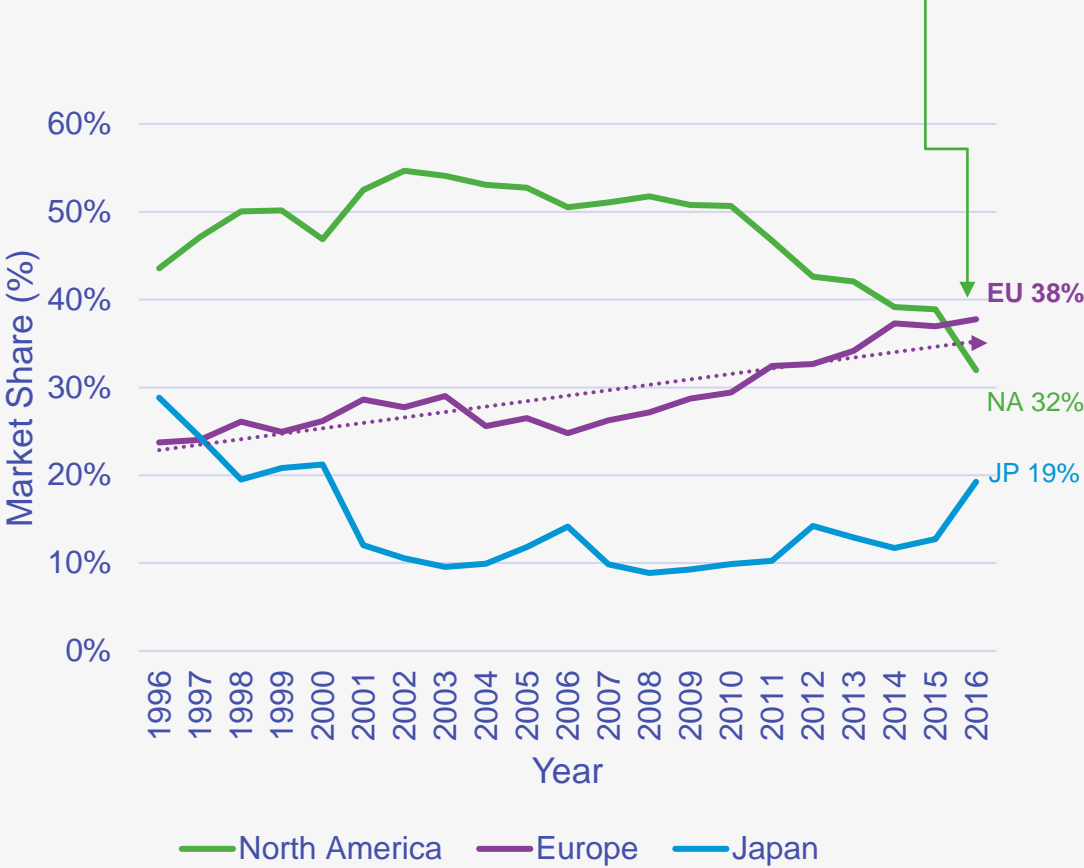
MARKET ANALYSIS



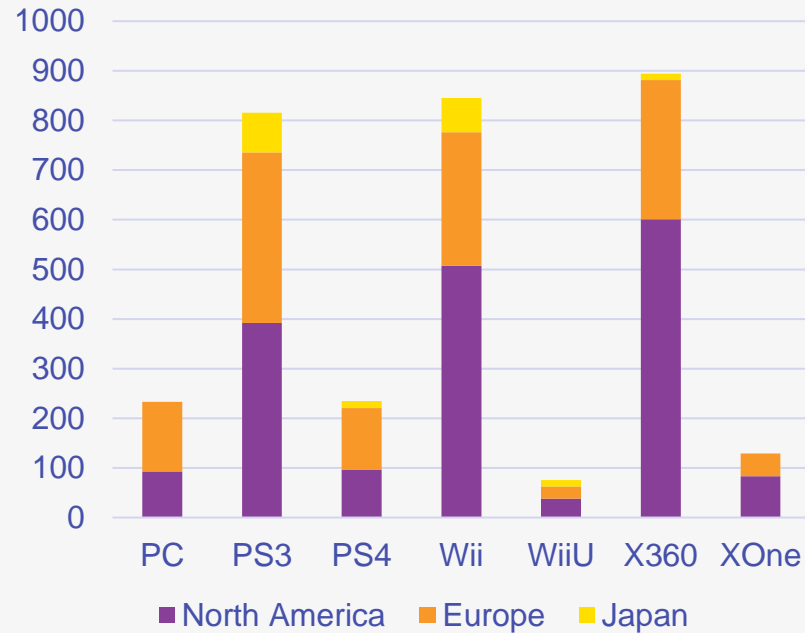
The three most popular regions had highest sales in 2008 and 2009. Since then, they have seen a steep decline.



For the first time, in 2015, Europe has overtaken North America in the highest market share.

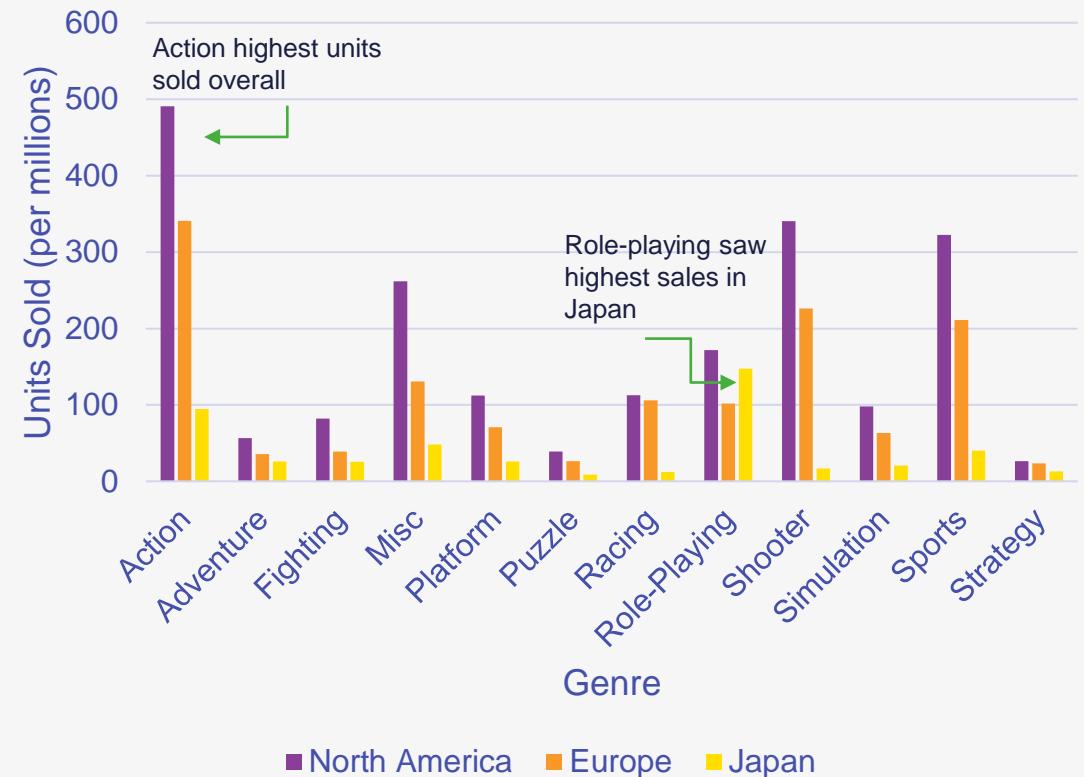


PLATFORM & GENRE ANALYSIS



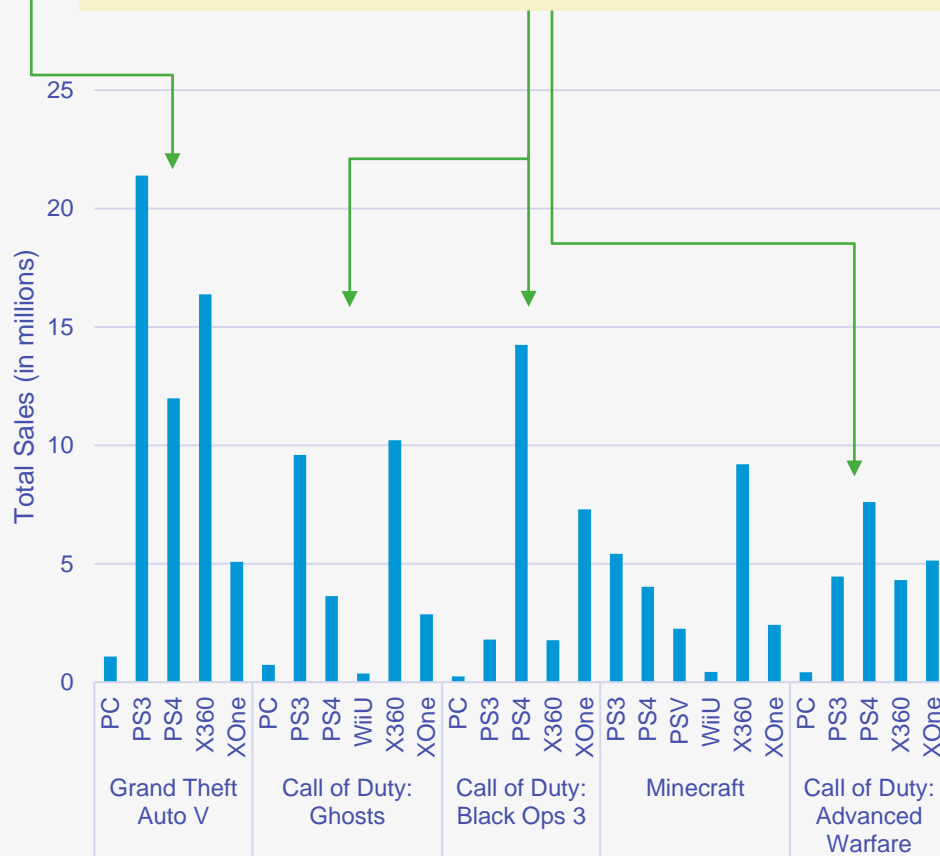
In 2016, there is a marked difference in platform preference in our top markets. Microsoft's platforms see more success in North America. Sony and Nintendo edge ahead in Europe and Japan.

Over the last ten years, Action, Shooter, and Sports genres have continued to rise in popularity across North America and Europe, while Action and Role-Playing hold steady in Japan.

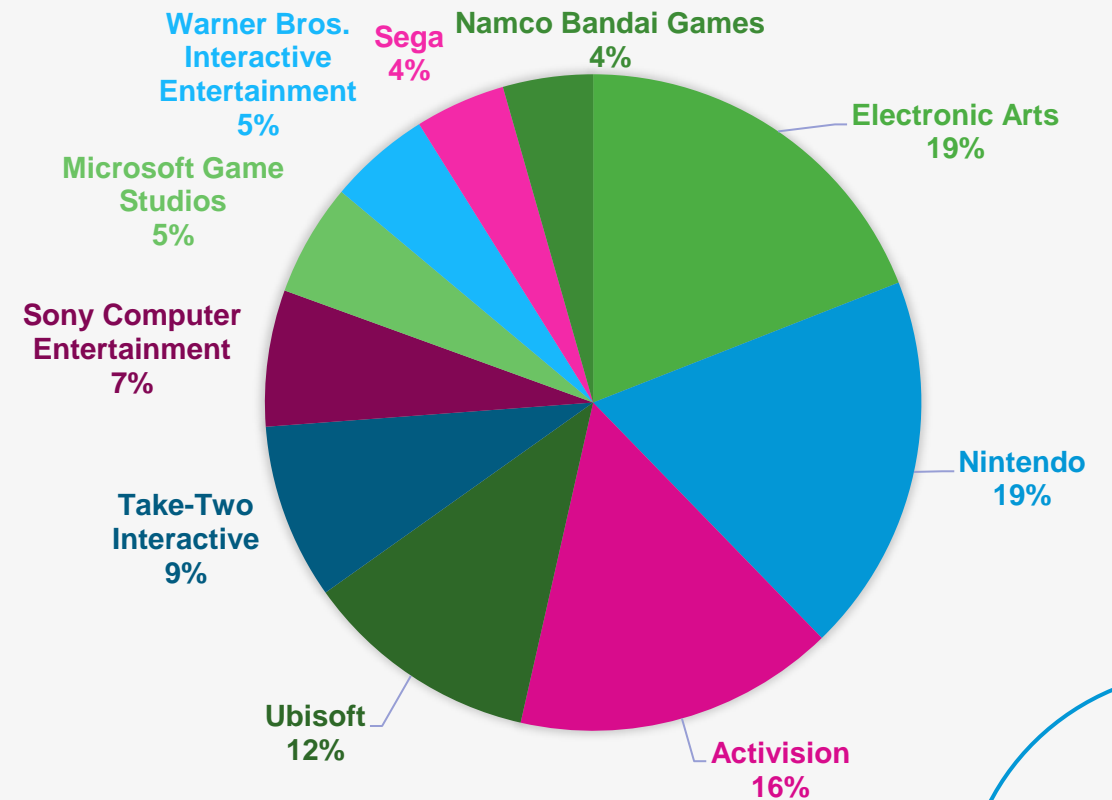


PUBLISHER & GAME ANALYSIS

Top 5 games in the last 5 years.
Grand Theft Auto V (Take-Two Interactive) takes 1st place.
Call of Duty (Activision) had 3 of the top 5 spots.



Top 10 publishers in the last 10 years.
Nintendo, Electronic Arts, Activision top 3.



RECOMMENDATIONS

Resource Allocation

- Increasing 38% of budget towards European markets as they are now the highest market share.

Target Marketing

- North America and Europe focus should be Action, Shooter and Sports genres.
- Japan target genres are Role-playing and Action.

Sales Strategies

- The steep drop in sales is most likely attributed to increased digital and mobile game markets. Sales should invest in digital distribution of future games.

Digital Analysis

- After digital media investment, further analysis should be conducted as trends may be rapidly changing.
- Adding in analysis of digital distribution services, such as Steam, to compare to physical copies of games sold.





Pig-E Bank

A global bank seeking analytical support for their customer retention team.

OBJECTIVE

Analyze historical customer data to contribute to development and optimization of predictive models in identifying client loss risk.

PROJECT DATA

- [Project Brief](#)
- [Client Database](#) provided by CareerFoundry

LIMITATIONS

- Customer demographics are limited to gender, age, and country with records of the users' account balance, estimated salary, membership status, etc.

TECHNIQUES APPLIED

- Big Data Management
- Data Ethics
- Data Mining
- Predictive Analysis
- Time Series Analysis and Forecasting

TOOLS



APPROACH & METHODOLOGY



Big Data

Defining the data characteristics (structured and unstructured) and recognizing its applications and limitations. Identifying software tools appropriate for handling big data.



Data Ethics

Identifying ethical dilemmas in managing big data in compliance with security and privacy laws



Data Mining

Performing data cleaning and descriptive statistics using pivot tables that will be utilized in generating a decision tree modeling for testing outcomes of the analysis.

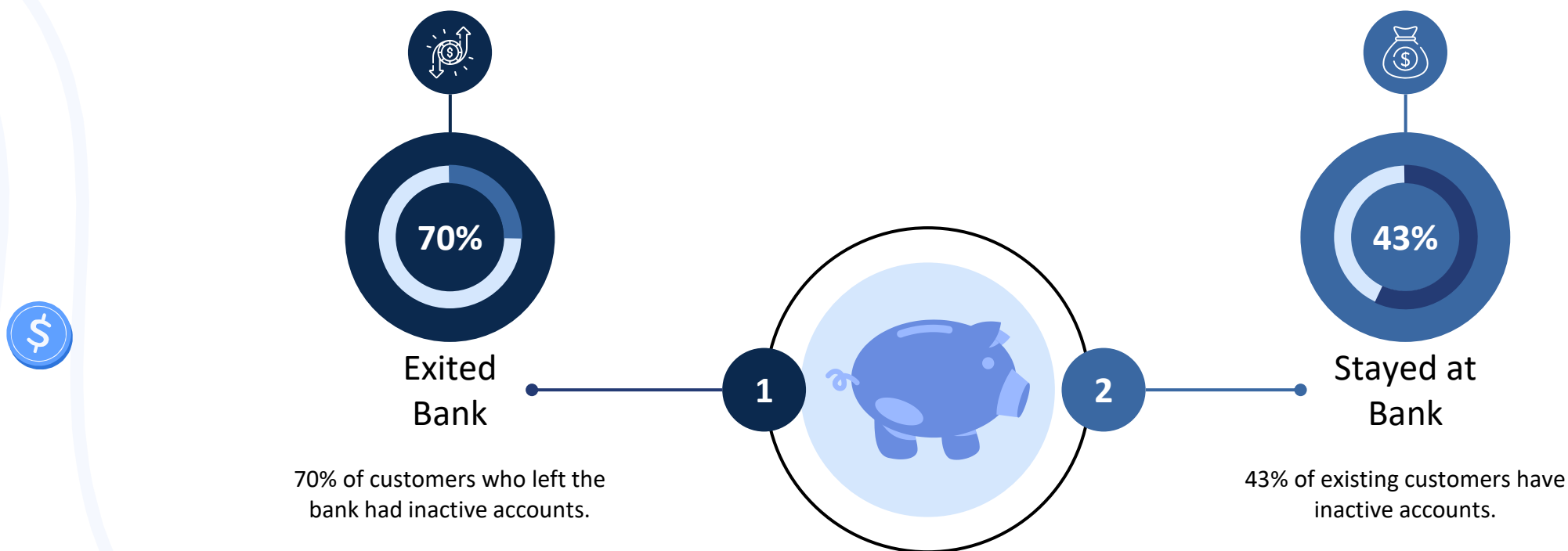


Analysis and Forecasting

Optimizing linear regression models through testing correct predictive prototypes in classifying risk factors that may have contributed to customer loss with the application of different scenarios



CUSTOMER RETENTION ANALYSIS

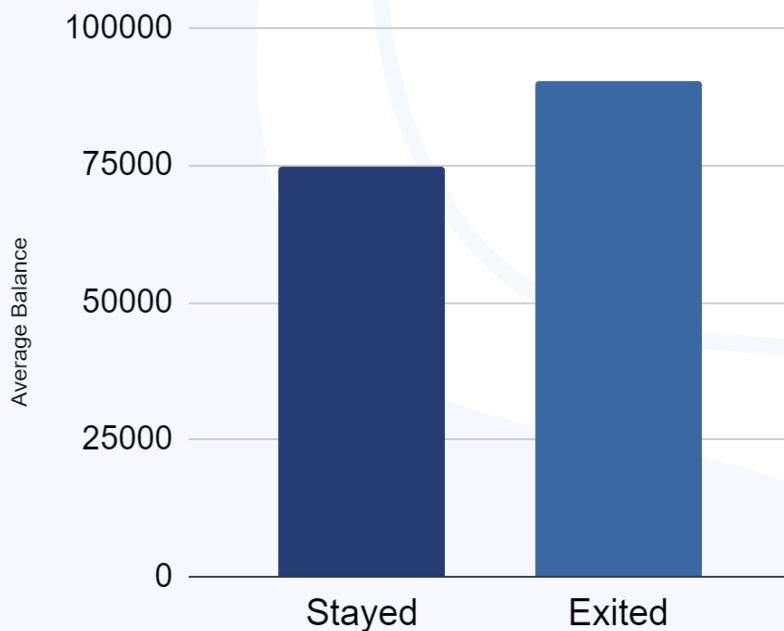




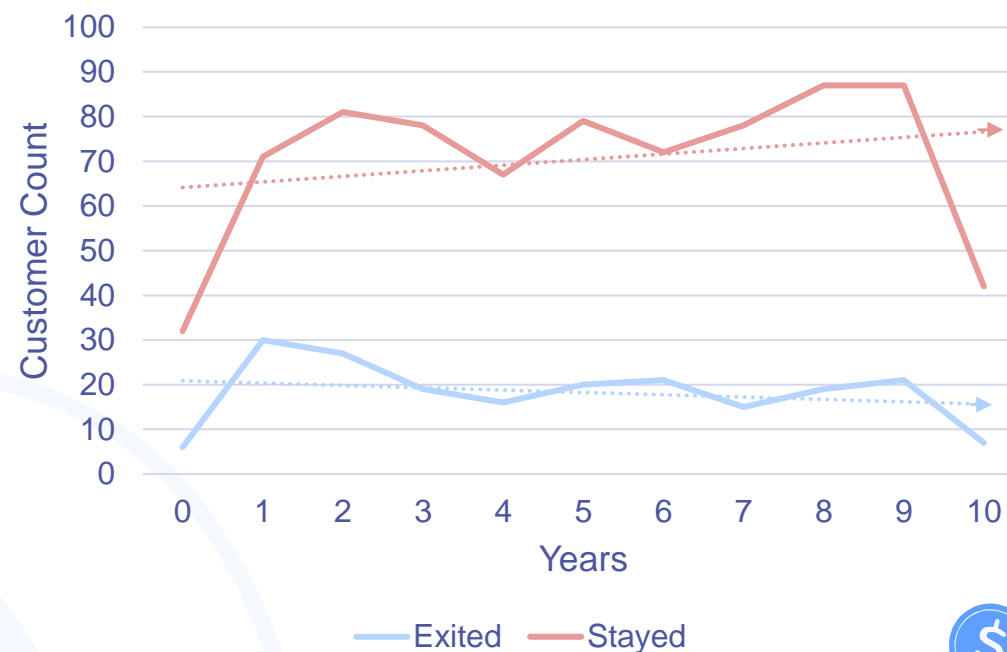
CUSTOMER RETENTION ANALYSIS



The average balance of those who **exited** the bank was **\$15.6K over** those who stayed with the bank.

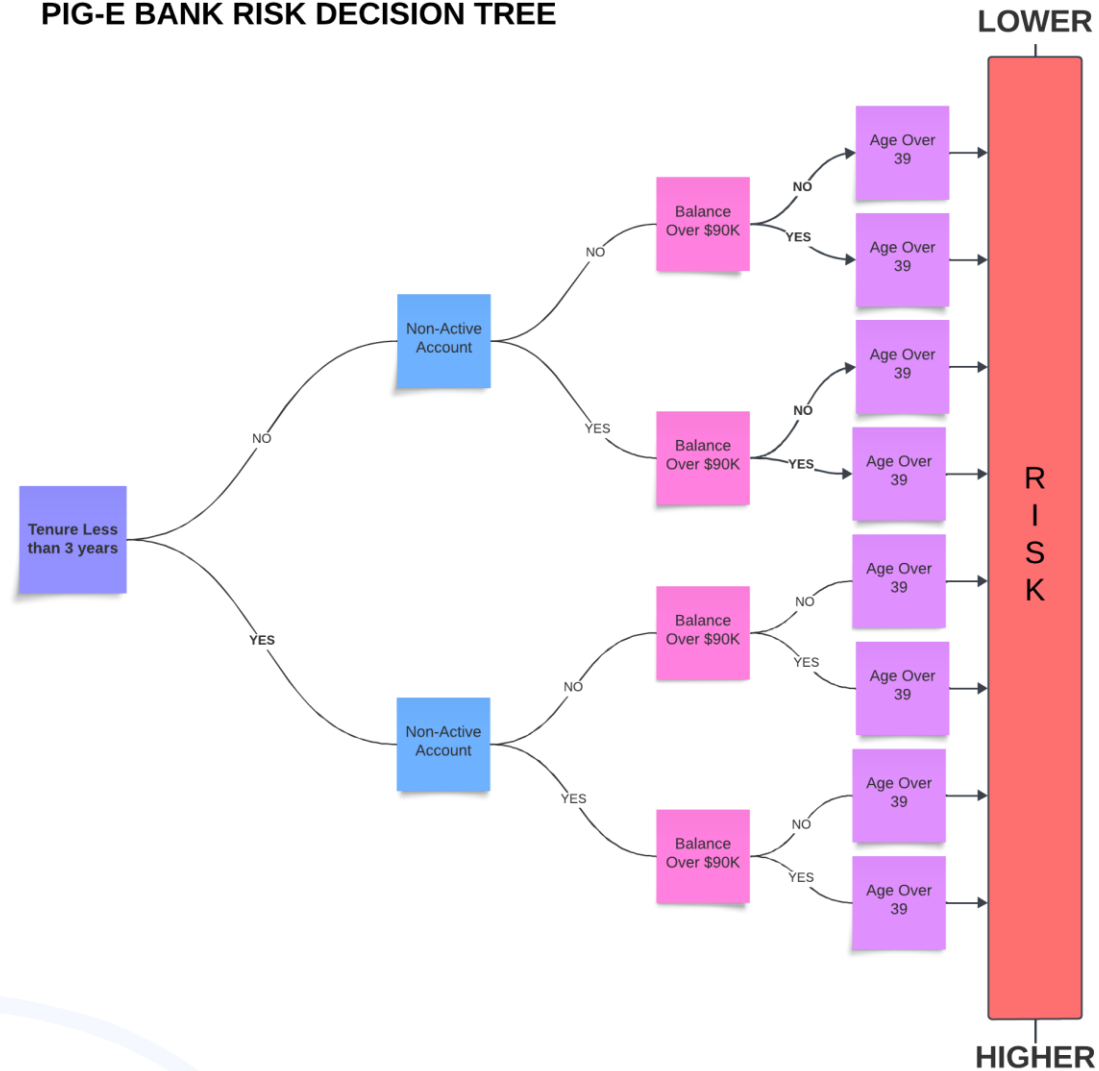


Customers who **exited** the bank saw a **downward trend after 3 years**, while customers who stayed saw an upward trend with more years.



PREDICTIVE MODEL ANALYSIS

FIG-E BANK RISK DECISION TREE

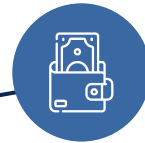


RECOMMENDATIONS



Business Development

Evaluate and adopt additional attractive banking products to target new customers and keep existing customers active.



Customer Surveys

Survey customers to gain better understanding of their banking needs, including identifying reasons for dissatisfaction.

Fraud Protection

Perhaps a lack of fraud protection is a reason customers with higher bank balance exit the bank. Higher protection may help retain high balance customers.



King County House Sales

King County is the most populous county in Washington State and the 13th most populous in the United States. There are many factors to consider when choosing a property.

OBJECTIVE

Investigate King County house sales data to determine what factors contribute to housing quality and performance so that clients can maximize their return on investment.

PROJECT DATA

- [Project Brief](#)
- [House Sales in King County, USA](#) via Kaggle
- [GIS Open Data](#) via GIS King County Government Website

LIMITATIONS

- Observations are limited to one year from May 2014 to April 2015.
- Cannot use this analysis for general house sales in Washington.

TECHNIQUES APPLIED

- Exploratory analysis through visualizations (scatterplots, correlation heatmaps, pair plots, and categorical plots).
- Geospatial analysis
- Linear Regression Analysis
- K-means Cluster Analysis
- Time-series analysis
- Present findings in a Tableau dashboard

TOOLS





APPROACH & METHODOLOGY

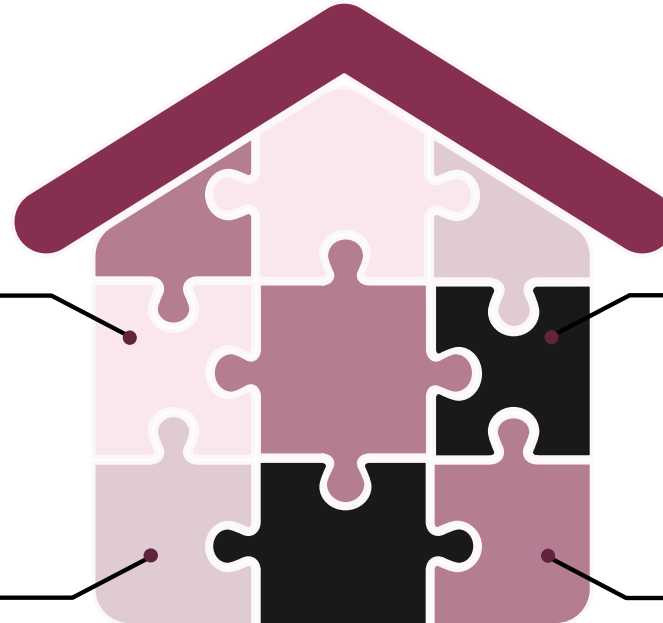


Data Sourcing and Cleaning

Sourced reliable datasets via Kaggle and government GIS sites for advanced analytics. Cleaned and prepared dataset for optimal diagnostics and defined questions to explore data content.

Exploratory Analysis

Utilize Excel and Python to find links between housing quality variables, location and price. Aggregated and grouped data for statistical analysis using techniques such as geospatial analysis, scatterplots, histograms, and more.



Advanced Modeling Techniques

Employed supervised and unsupervised machine learning models and time series analysis using Python to glean insights on the market.

Data Dashboard

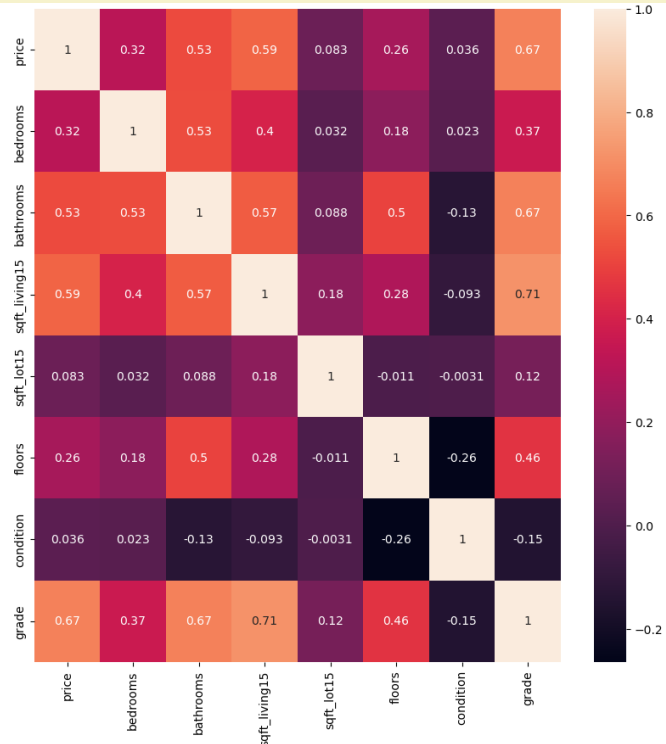
Created dahsboards to show how sales have been distributed between different cities and zipcodes. Formated a Tableau storyboard that presents significant findings of the analysis in an interactive format.



EXPLORATORY ANALYSIS

A **correlation heat map** was created to help identify initial relationships between variables such as price, square feet living space, bedrooms, grade, etc.

The analysis revealed there is a **moderate relationship** between **price** and **square feet living space** and **grade**.



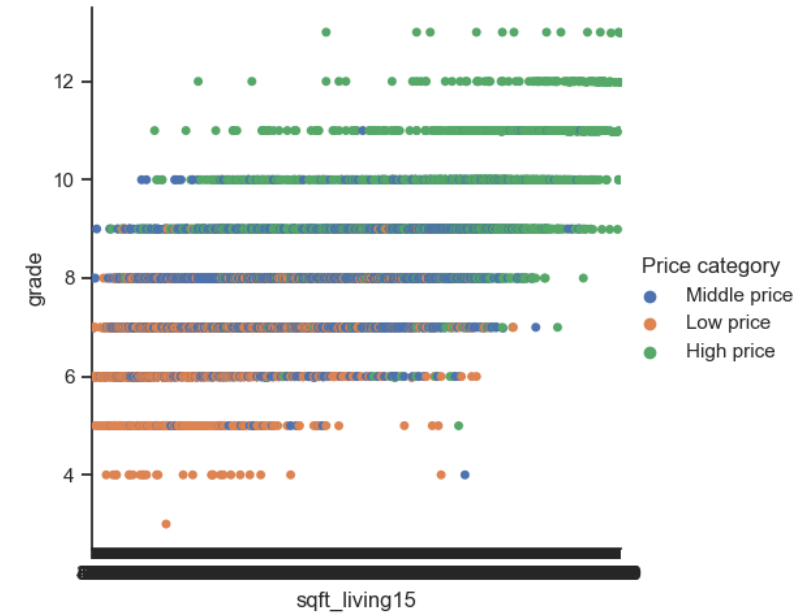
```
# Create a subplot with matplotlib
f,ax = plt.subplots(figsize=(10,10))

# Create the correlation heatmap in seaborn by applying a heatmap onto the correlation matrix and the subplots defined above.
corr = sns.heatmap(sub.corr(), annot = True, ax = ax) # The 'annot' argument allows the plot to
#place the correlation coefficients onto the heatmap.
```



Categorical plots were used to provide valuable insights with correlated variables.

We can see the **trend** between higher square feet living space, grade and different price points in the data.



```
# Create a categorical plot in seaborn using the price categories created above

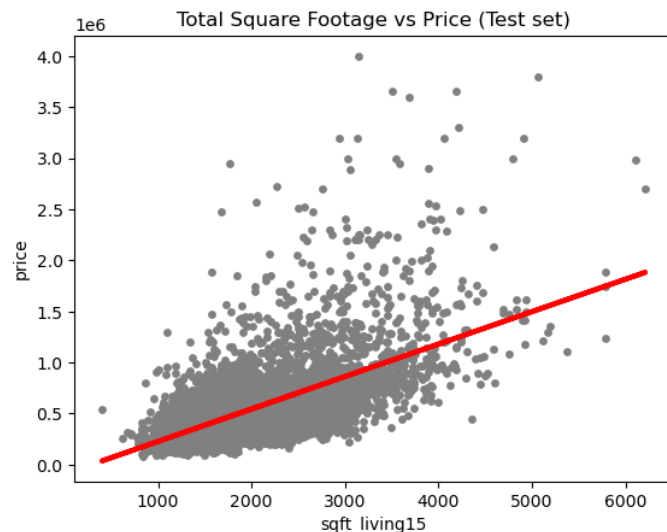
sns.set(style="ticks")
g = sns.catplot(x="sqft_living15", y="grade", hue="Price category", data=kc_house_data)
```

ADVANCED MODELING



Linear regression modeling was used to identify a positive trend.

However, the **statistical testing yielded poor results**, suggesting that this test was not enough to give an accurate prediction into the future.

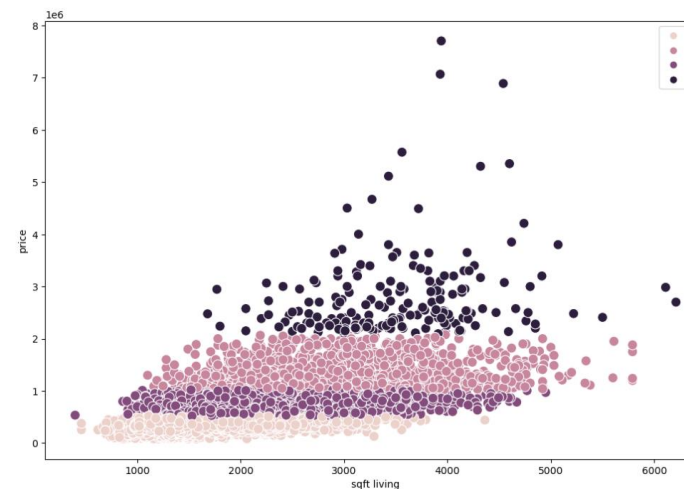


```
# Create a plot that shows the regression line from the model on the test set.

plot_test = plt
plot_test.scatter(X_test, y_test, color='gray', s = 15)
plot_test.plot(X_test, y_predicted, color='red', linewidth=3)
plot_test.title('Total Square Footage vs Price (Test set)')
plot_test.xlabel('sqft_living15')
plot_test.ylabel('price')
plot_test.show()
```

Further testing was conducted using the **k-means algorithm**, which grouped data points with similar traits into clusters.

These groups allowed for further insights into different **price points**.



```
# Plot the clusters for the "price" and "sqft_living15" variables.

plt.figure(figsize=(12,8))
ax = sns.scatterplot(x=kc_sub['sqft_living15'], y=kc_sub['price'], hue=kmeans.labels_, s=100)
# Here, you're subsetting 'X' for the x and y arguments to avoid using their labels.
# 'hue' takes the value of the attribute 'kmeans.labels_', which is the result of running the k-means algorithm.
# 's' represents the size of the points you want to see in the plot.

ax.grid(False) # This removes the grid from the background.
plt.xlabel('sqft_living') # Label x-axis.
plt.ylabel('price') # Label y-axis.
plt.show()
```

DATA DASHBOARD

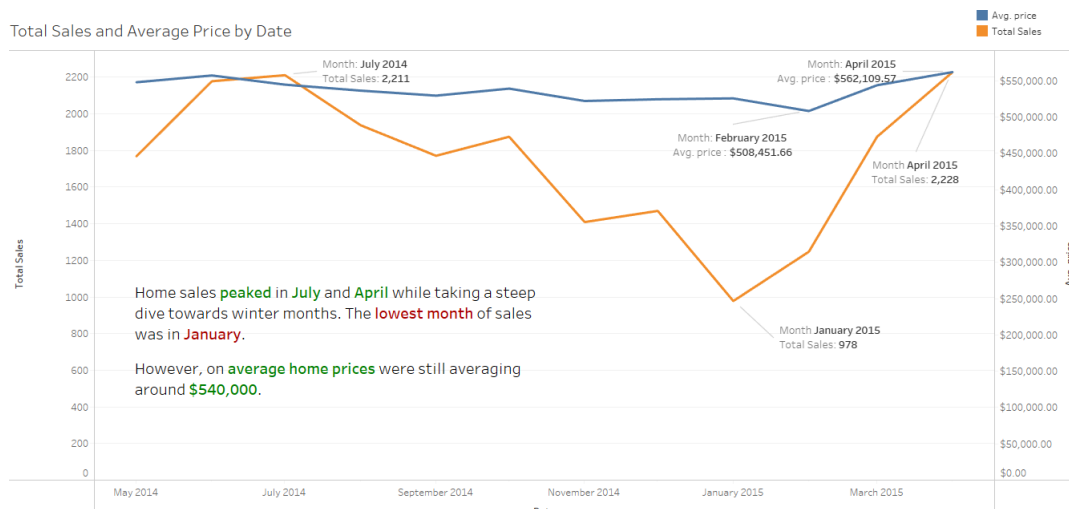
Finally, a data dashboard was created in Tableau using insights from the exploratory and advanced analysis to give insights on variables such as seasonality, city and zipcode.

House **sales peaked** from **April to July**, while average prices stayed steady.

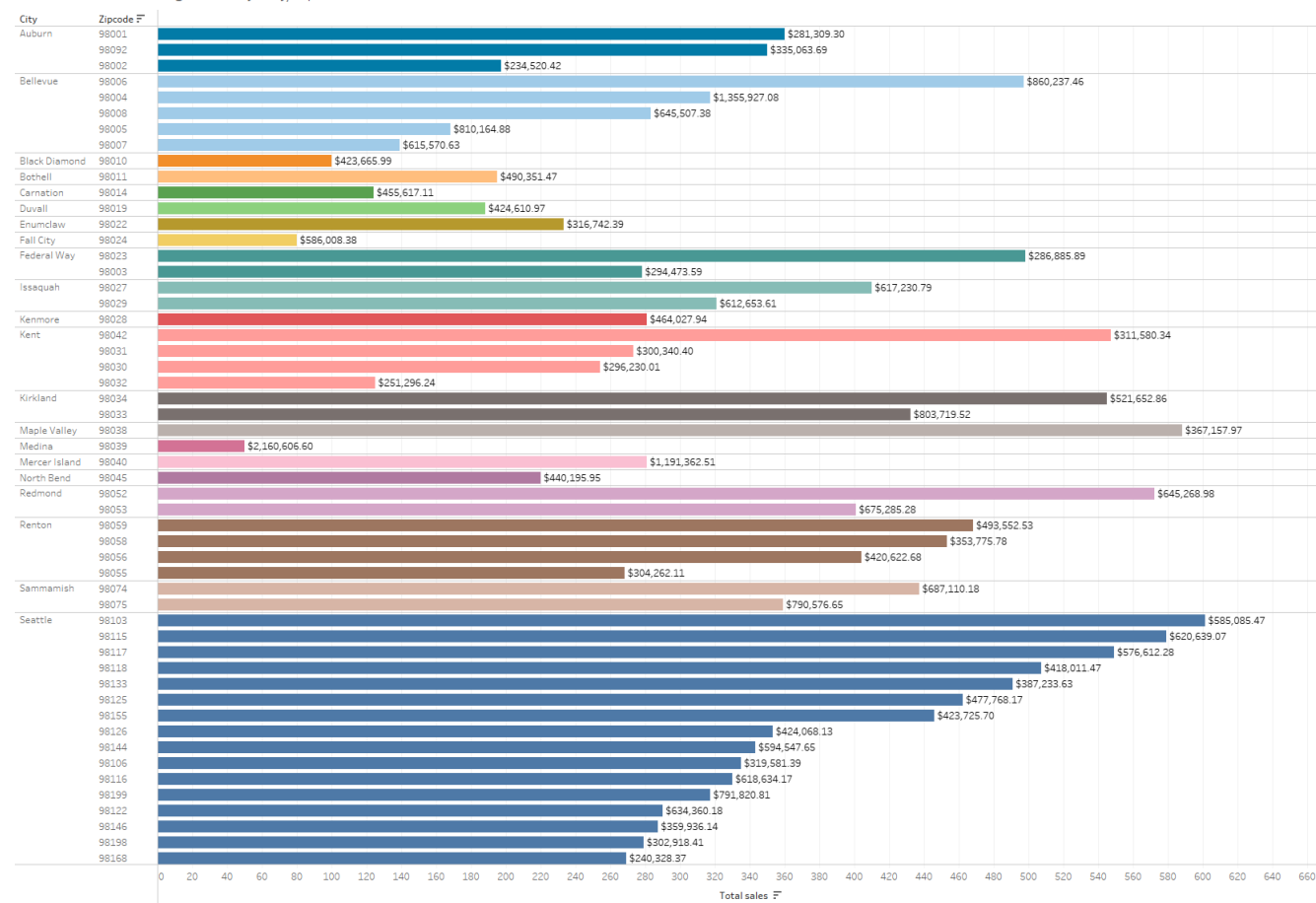
Location seems to be the greatest correlation to prices.

Seattle had **42% of sales**, however, **Medina, Mercer Island and Bellevue** had the **highest average prices per house**.

Total Sales and Average Price by Date



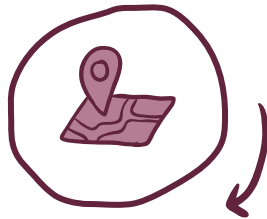
Total Sales and Average Price by City/Zipcode



RECOMMENDATIONS



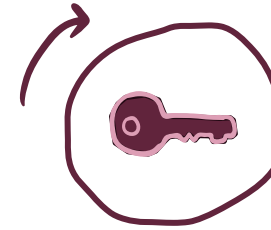
To get the best ROI, homes that are graded 8-9 with 3-4 bedrooms and average square feet of 1900 seem to correlate to higher prices.



Seattle has the highest amounts of homes sold. Medina and Mercer Island had highest average sales prices, but Bellevue, Sammamish and Redmond sold well and had more affordable houses.



Spring and Summer are higher in sales than any other months. Inventory of homes will most likely be at its highest during these times.



Further analysis must be done to compare zipcodes among cities further, including median incomes, crime rates, and schooling. This would be invaluable for finding the best investment properties.



Thanks!

JANELLE SOUSLEY



CREDITS: This presentation template was created by [Slidesgo](#), and includes icons by [Flaticon](#), and infographics & images by [Freepik](#)

