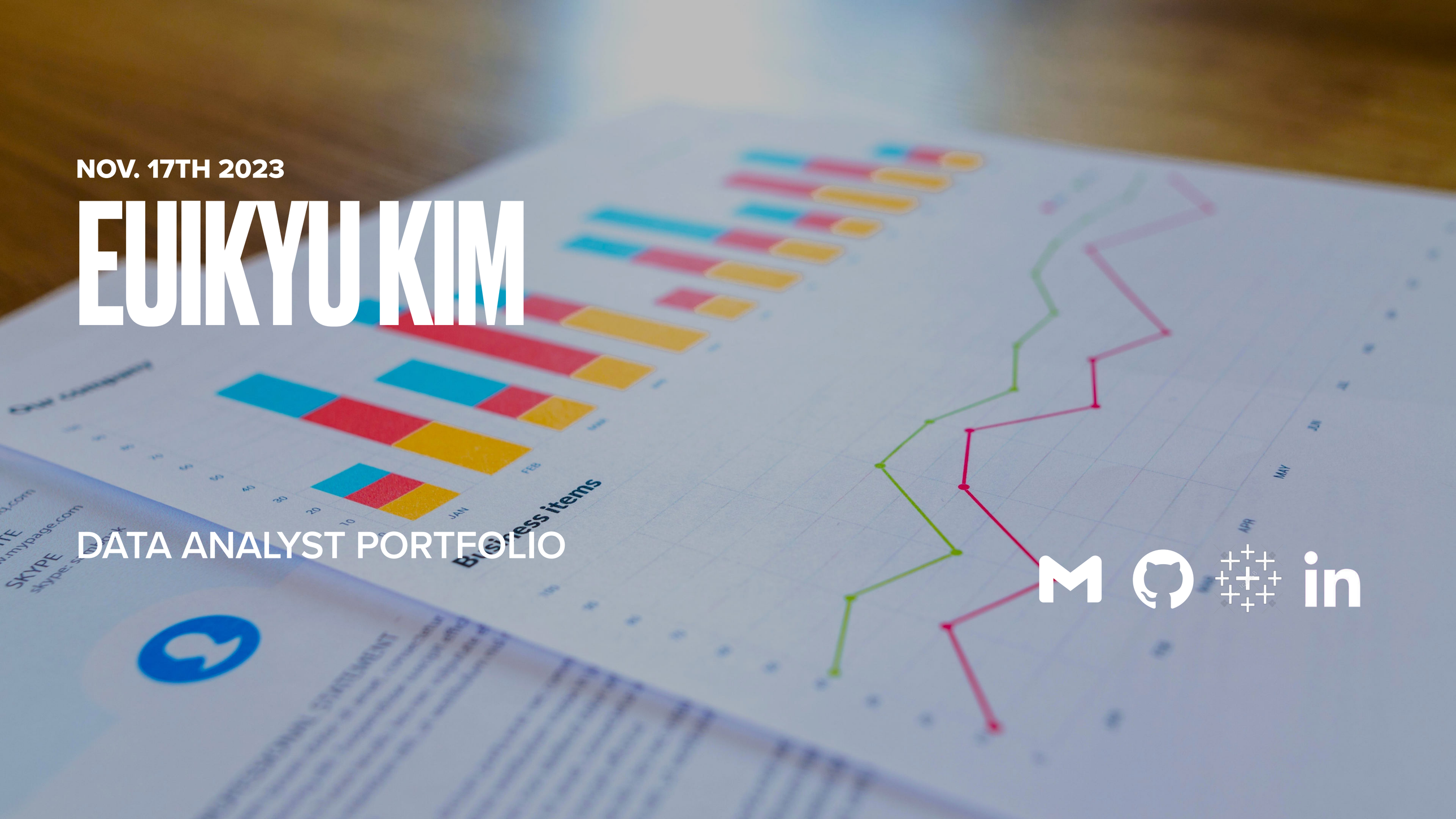
NOV. 17TH 2023

EUIKYU KIM

DATA ANALYST PORTFOLIO



Welcome to my Portfolio

Hi I'm Euikyu Kim

"My goal is to have a sustainable impact on the organization's success with the use of my extensive research, compelling visualizations, and project coordination/management abilities. I aim to make a difference by leveraging my accumulated analytical skill sets in discovering innovative solutions, through a data-driven approach fueled by creativity, and critical thinking.

My professional experience in business operations in multiple departments - big and small organization settings has deepened my understanding of the importance of well-organized, competent workflow procedures influenced by data-rich decisions. Therefore, my desire and determination to bring positive changes are what pushed me to evolve my career map."

Kim Euikyu

NOTE: THIS PORTFOLIO COMPRISES EXCERPTS FROM RESEARCH STUDIES. THESE ARE NOT FULL-DETAILED REPORTS. HYPERLINKS ARE PROVIDED TO ACCESS FULL VERSIONS.

PROJECT EXCERPTS

INSTACART
Python-based project on Consumer Behavior Trends and Marketing Analysis

ROCKBUSTER STEALTH
International Business Analysis of Online Video Rental Services through SQL

PREPARING FOR INFLUENZA SEASON IN THE U.S.
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

HOUSE SALES
Regression, cluster, and time-series analysis to find factors determining house prices in King County, WA

OBJECTIVE

Produce an exploratory analysis of consumer behavior and sales patterns in sustaining information that will benefit sales and marketing departments in developing tactical promotional efforts.

PROJECT & DATA

- Project Brief
- Customer Data Set, Customer Orders, Products, and Department I Modified Open Source from Instacart
- Data Dictionary I Link provided by CareerFoundry

LIMITATIONS

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

TECHNIQUES APPLIED

- Data Cleaning: Wrangling and Subsetting
- Data Consistency Checks
- Combining and Exporting Data
- Deriving New Variables
- Grouping Data and Aggregating Variables
- Python Visualization and Excel Report

TOOLS

Excel Anaconda Panda Jupyter Pandas matplotlib NumPy seaborn

EXCERPT

INSTACART

An online grocery delivery service that allows users to order from local stores. Executives are curious to learned about different consumer groups and understand their purchasing patterns for marketing campaigns

NOTE: Instacart is a real company that's made its data publicly available online. However, the content of this project has been fabricated for the purposes of this Achievement.

1. ORGANIZING DATA

Tracking data irregularities like mixed-type variables, missing and/or duplicate values for ratification. Methodically noting steps trailing the transformation of the data frame will allow other users to identify procedural actions taken.

2. DERIVING VARIABLES

Using conditional logic in the form of if-statements, user-defined functions, the loc() function, and for-loops in deriving new columns.

3. AGGREGATING DATA

Creating flags and placing new columns for a summary of descriptive analysis with groupby() function.

4. PYTHON VISUALIZATION & EXCEL REPORT

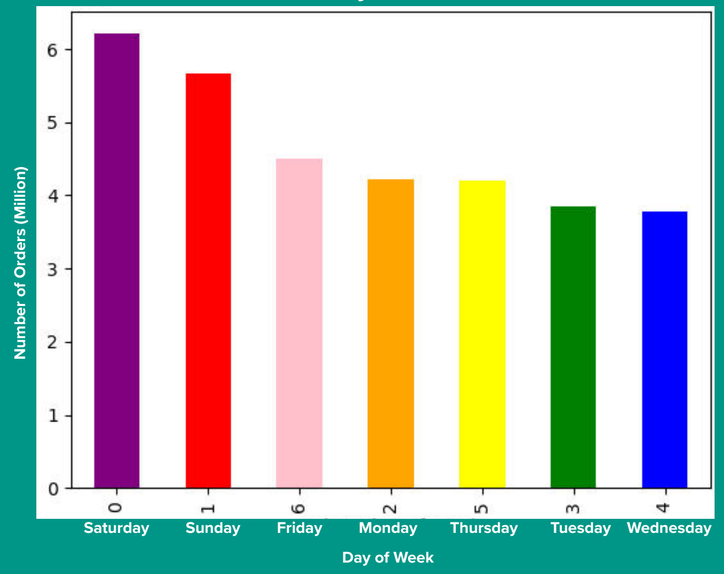
A merged data frame is utilized to generate a compelling visualization addressing the stakeholders' questions.

BUSINESS ANALYSIS

FULL REPORT

```
bar = ords_prods_merge['orders_day_of_week'].
        value_counts().sort_index().plot.bar(color =
        ['purple','red','pink','orange','yellow','green','blue'])
```

Busiest Day of the Week



“The Busiest Days of the Week are Saturday, Sunday, and Friday”

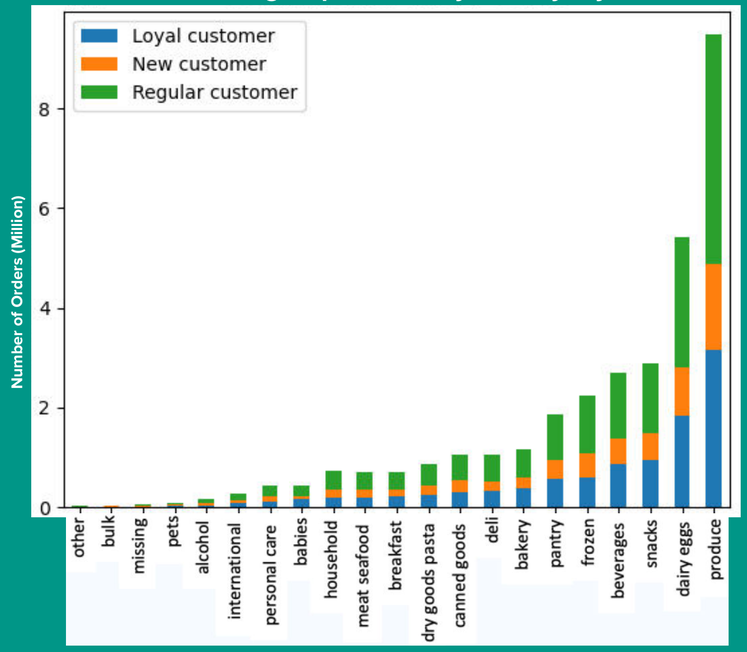
```
line_hr_by_loyal = pivot_data.plot()
plt.title("Price by Hour of day & Loyalty status")
plt.xlabel("Hour of day")
plt.ylabel("Number of Orders (million)")
plt.legend(loc ='upper left', bbox_to_anchor=(0.5, 1))
```

Hourly Average Prices of the Day by the Loyalty Group

"While loyal and regular customers have a big difference in average price depends on time, new customers have a relatively small difference."

```
stkd_dept_by_loyal = dept_by_loyal.sort_values(by = ['Loyal customer'],
                     ascending = True).plot.bar(stacked = True)
plt.title("Department by Loyalty status")
plt.xlabel("Hour of day")
plt.ylabel("Number of Orders (million)")
plt.legend(loc ='upper left', bbox_to_anchor=(0, 1))
```

The Most Selling Department by the Loyalty Status



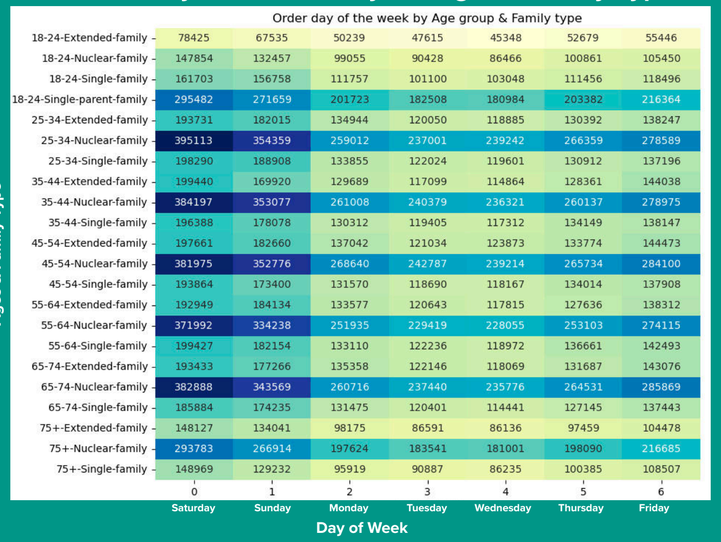
“The Most Selling Department by the Loyal Status is Produce”

CONSUMER SEGMENT ANALYSIS

FULL REPORT

```
plt.figure(figsize = (12, 8))
hmap_day_by_age_fam = sns.heatmap(pivot_tab, cmap =
                        'YlGnBu', annot = True, fmt = 'd', cbar = True)
plt.title("Order day of the week by Age group & Family type")
plt.xlabel("Day of week")
plt.ylabel("Age group - Family type")
plt.show()
```

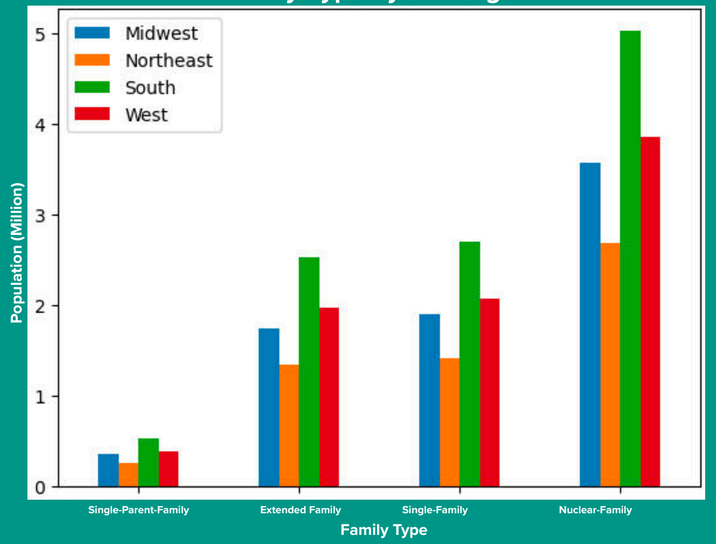
Busiest Day of the Week by the Ages & Family Type



"The Busiest Days of the Week are Saturday and Sunday"

```
bar_fam_type_by_region = fam_type_by_region.sort_values(by
= ['West'], ascending = True).plot.bar()
plt.title("Family type by Region")
plt.xlabel("Family type")
plt.ylabel("Population (million)")
plt.legend(loc ='upper left', bbox_to_anchor=(0, 1))
```

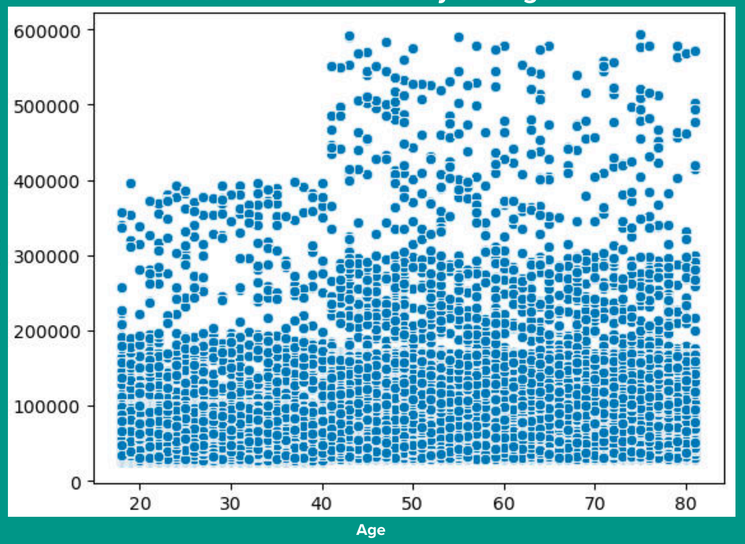
Family Type by the Region



"Nuclear-Family customers have the highest portion in all regions of the U.S. followed by the Single, Extended, and Single-Parent-Family"

```
scttr_age_income = sns.scatterplot (x = 'age', y = 'income', data = df_merge)
```

Income Distribution by the Ages



20 to 40 years old : 0 to 200,000 ———— very many
200,000-4,000,000 — not many
over 4,000,000 ——— none
40 to 80 years old : 0-200,000 ————— very many
200,000-3,000,000 — many
over 3,000,000 ——— not many

CUSTOMER TREND

RECOMMENDATION

ADVERTISEMENT

- Run ads after 6pm, when people finish work and orders begin to decrease.
- Run ads during peak order times to challenge and set new sales records.
- Enhance advertising by introducing special events like discounts or Buy One Get One (BOGO) offers during low-order periods.

REGION

Explore the Northeast region market which has a smaller share but has high population density and convenient public transportation with innovative and trendy offline stores such as Amazon Go.

PRODUCT

- Focus marketing efforts on organic products between 9am and 4pm when orders are at their highest.
- Develop marketing initiatives with a focus on vegetarians.

CUSTOMER LOYALTY

Maintaining a customer base where the proportion follows the sequence of Regular > Loyal > New is indicative of a healthy and effective customer management strategy.

AVERAGE ORDER PRICE

Implement additional strategies between 8 to 11am when both regular and loyal customers tend to place orders with the lowest average order prices.

OBJECTIVE

Provide insights on the current business standing to support the company's 2020 marketing campaign for a new product launch.

PROJECT & DATA

- Project Brief
- Dataset | Source by CareerFoundry
- Data Dictionary

LIMITATIONS

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

TECHNIQUES APPLIED

- Relational Databases
- Entity Relationship Diagram (ERD)
- Creating a Data Dictionary
- Database Querying
- Filtering Data
- Data Cleaning and Summarizing
- Jointing Tables
- Subqueries
- Common Table Expressions

TOOLS

Excel PowerPoint Tableau PostgreSQL

1. ASSESSING DATABASE

Initial review and crafting of an Entity Relationship Diagram (ERD) through DbVisualizer describes the structure of the database for analysis and capturing it through a data dictionary for user accessibility.

2. DATA CLEANING

Performing CRUD functions ensures a clean and consistent format for filtering as well as summarizing data for output accuracy.

3. DATA EXTRACTION & SUMMARIZATION

Extracting records using Joining Tables, Subqueries, and Common Table Expressions (CTEs) commands allow an extensive comprehension of the current business standings.

4. DATA VISUALIZATION & STORYTELLING

SQL results are converted to CSV files to generate visualization with ease for the final presentation and tableau storyboard.

FULL REPORT

BUSINESS OVERVIEW

Analysis says ...

- Average revenue per movie: $64
- Average rental duration: 5 days
- Revenue per movie title range: $6-$216
- Three rental rates: $0.99/day-$4.99/day
- Total revenue (2007): $61,312

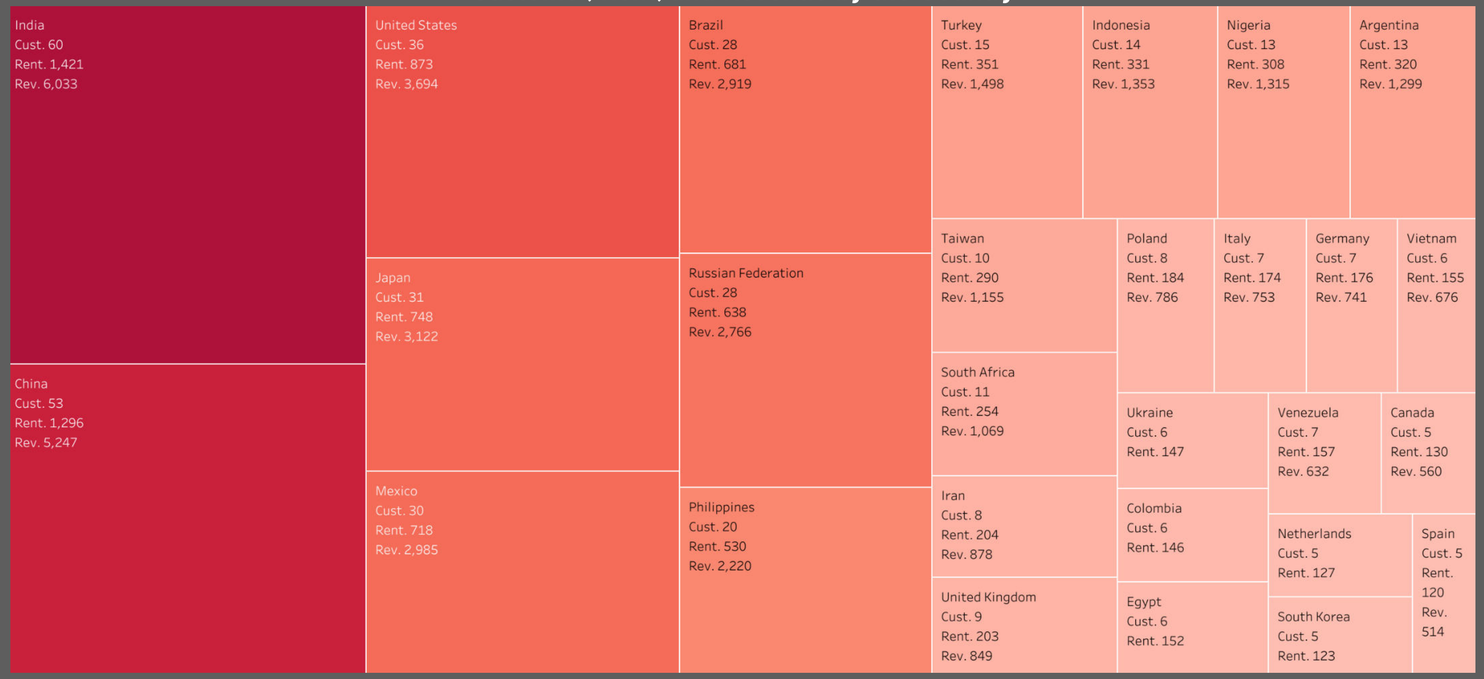
There are ...

- 5 film ratings
- 16 genres
- 1,000 movie licenses
- in 4,581 inventories
- 16,044 movie rentals
- 584 active customers
- in 108 countries

BUSINESS ANALYSIS

FULL REPORT

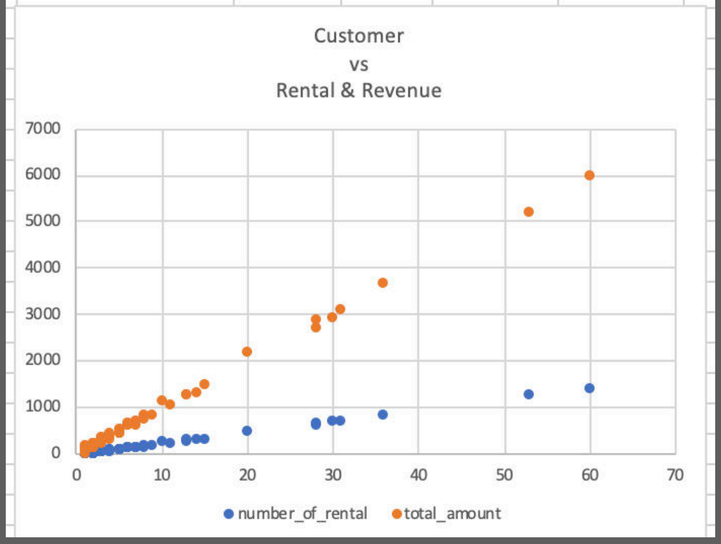
Customer, Rent, and Revenue by the Country



"Asian countries, including India, China, Japan, the Philippines, and Indonesia, account for 30% (178) of customers, 27% (4,326) of rentals, and 29% ($17,975) of revenue in the global sales."

Correlation between Customer vs Rental & Revenue

CORRELATION		VALUE
Customer vs Rental	=CORREL(B:B,C:C)	0.9989993
Customer vs Revenue		0.9989474



" = CORREL ('number_of_customer', 'total_amount')"
-> Customer vs Rental : 0.9989993 "
= CORREL ('number_of_customer', 'number_of_rental')"
-> Customer vs Revenue : 0.9989474

ONLINE VIDEO RENTAL SERVICES

RECOMMENDATION

REGION

- Priority Market: Asia (India, China, Japan, Philippines, Indonesia)
- Second-Priority Market: Americas (United States, Mexico, Brazil)
- Third Opportunity Market: Europe (Turkey, Germany, etc.)
- Markets to Pioneer: Middle East, Africa, Oceania

FEES & DURATIONS

- On a $2.98 for 5 days basis
- Vary by the popularity of the movie

CUSTOMER PLAN

Reward point program

INFRASTRUCTURE

Consider the network infrastructure in developing countries and major target markets such as India, China, the Philippines, and Indonesia. e.g. Low resolution, downloads, offline services, etc

NEXT STEP

- Analysis of competitor companies.
- Local research on target markets.

OBJECTIVE

Analyze influenza trends in the U.S. in order to support a medical staffing agency in strategically planning the deployment of temporary healthcare personnel for the upcoming influenza season.

PROJECT & DATA

- Project Brief
- Influenza Deaths I Source by CDC
- U.S. Population I Source by U.S. Census Bureau
- Influenza Visit I Source by CDC (Fluview)

LIMITATIONS

- Denza mortality data entries were suppressed for patient confidentiality. Imputed estimates are used for calculations.
- Death records identify a single underlying cause of death (influenza-initiated may not be counted).
- Data sets are from 2009 to 2017.

TECHNIQUES APPLIED

- Designing a Data Research Project
- Data Profiling and Integrity
- Data Cleaning
- Data Transformation and Integration
- Excel: Pivot Tables, VLOOKUP, PowerPivot
- Statistical Analysis and Hypothesis Testing
- Data Visualization and Storytelling (Tableau)

TOOLS

Excel Word Tableau

EXCERPT

PREPARING FOR INFLUEZA SEASON IN THE U.S.

A medical staffing agency aims to allocate temporary workforces in the most impacted areas to mitigate the influenza outbreak for the coming season. With limited resources to hire new employees, they are determined to support the healthcare system across all 50 states.

1. DESIGNING DATA RESEARCH PROJECT

Interpreting business requirements to data questions leads to a research hypothesis that serves as a guideline for the analysis. A project management plan is prepared to keep track of progress.

2. DATA PREPARATION

Exploring the datasets for information relevancy, integrity, completeness, etc. will help produce valuable insights. Then, transforming and integrating multiple data discloses the influenza case developments that will warrant the planning phase.

3. STATISTICAL ANALYSIS & HYPOTHESIS TESTING

Performing statistical methods that detect critical age populace targeted by influenza will model the next steps in preparation for the staff distribution. The relations of multiple variables are then confirmed through t-testing.

4. DATA VISUALIZATION & STORYTELLING

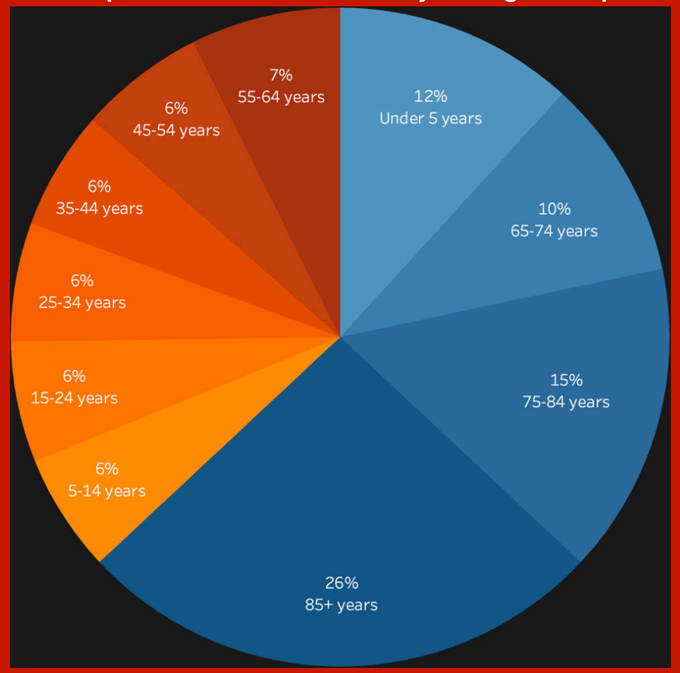
Data results are utilized for a compelling tableau presentation disclosing influenza trends and vulnerable age populations.

STATISTICAL ANALYSIS & HYPOTHESIS TESTING

FULL REPORT

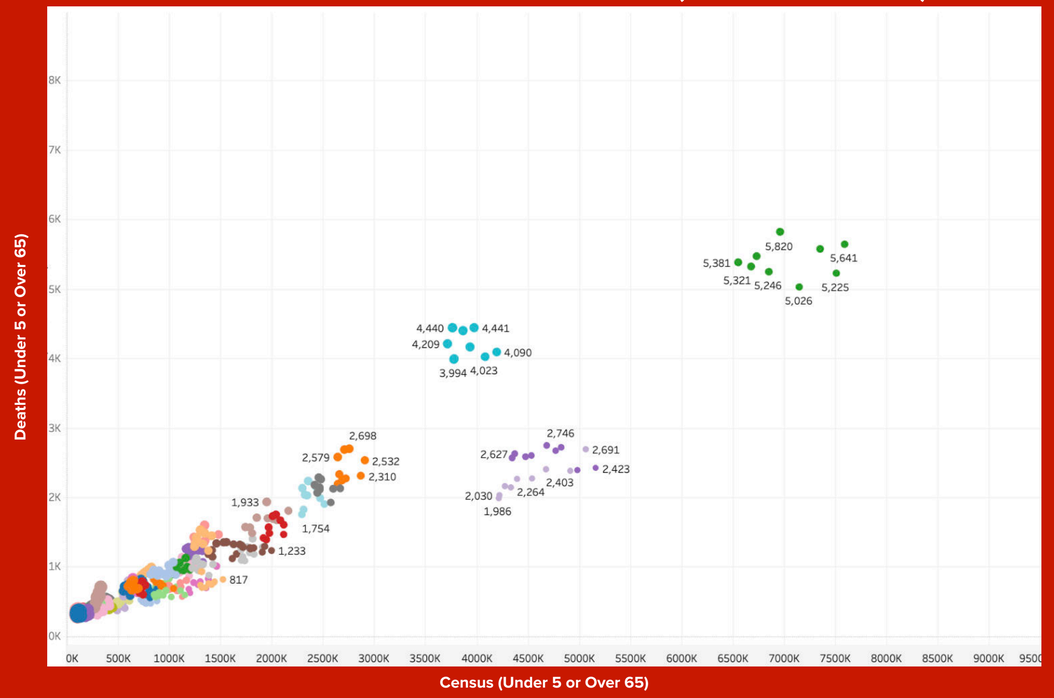
HYPOTHESIS:
If the proportion of age under 5 or age over 65 increases among influenza patients, then the influenza mortality rate will also increase.

Proportion of the Flu Death by the Age Group



"63% of all influenza deaths occur in those under the age of 5 or over 65 years."

Correlation Between Flu Deaths & Census (Under 5 or Over 65)

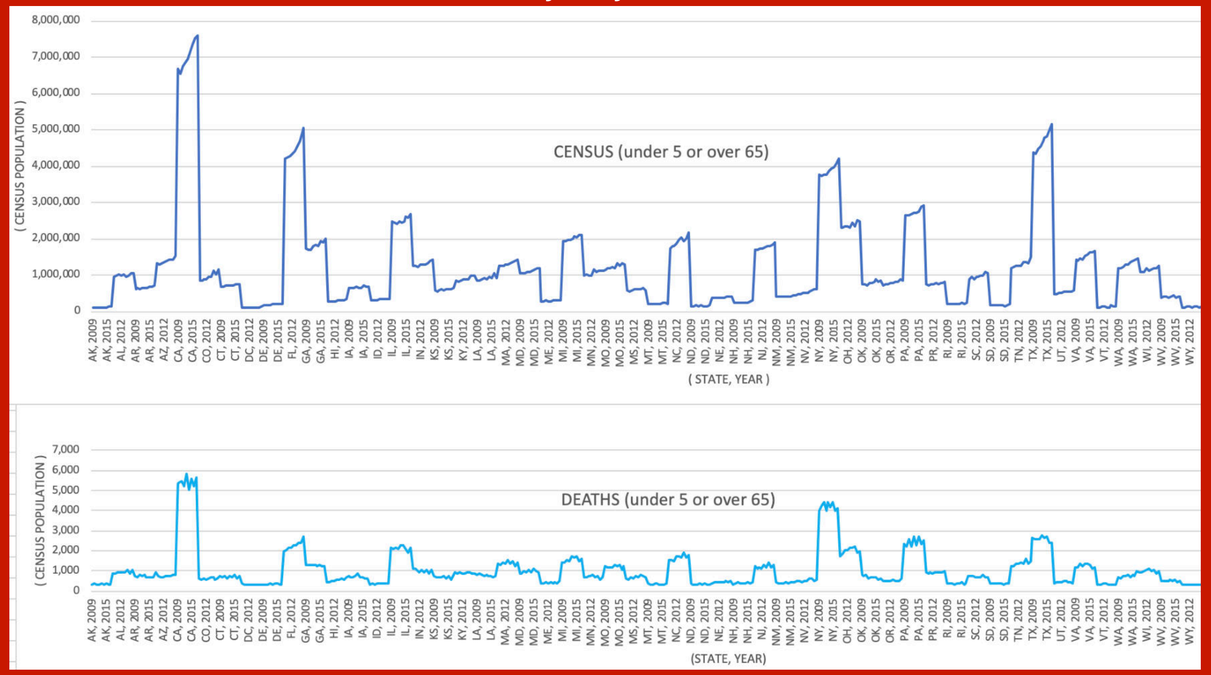


"They have a 0.94 Correlation Coefficient which means the very strong correlation"

TIME SERIES ANALYSIS

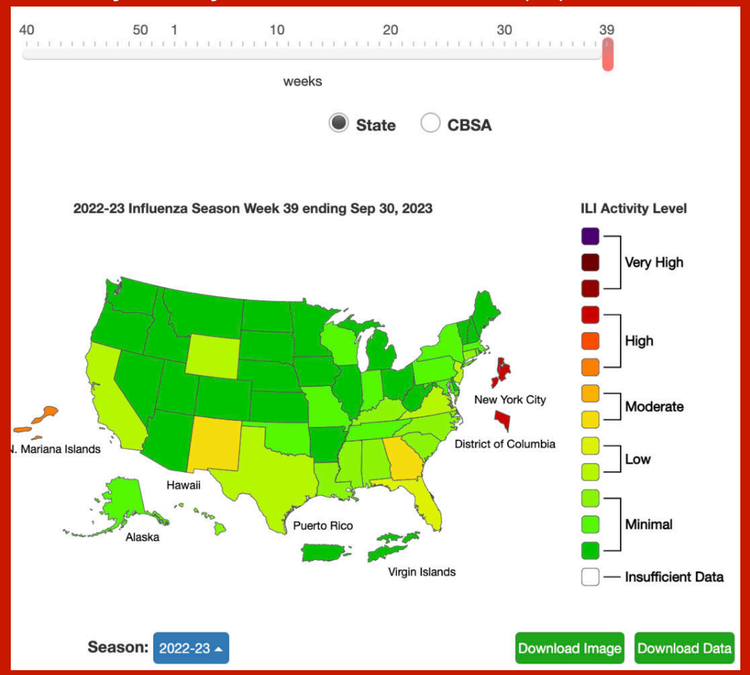
FULL REPORT

Time Series Analysis by the States and Year



“While the upper chart shows the movement of the CENSUS population, the lower chart shows the movement of the influenza deaths. Both charts are based on the group of ages under 5 or over 65 by state and year.”

Monthly Activity of Influenza-Like Illness (ILI) in the U.S.



Flue Season: Begin on October
Peak on December
Reduce in January (NY, DC reduce in June)
Frequency: East > West
South > North
West & North-Central begin late and reduce early

NATIONAL MEDICAL STAFFING

RECOMMENDATION

REGION

Densely populated metropolitans and their regions.
(e.g. CA, NY, TX, PA, FL, etc)

SUBJECT

Vulnerable age group.
(e.g. Under 5 years, over 65 years)

VACCINATION

Identify the specific start and end date of the flu season for each State based on the information from the website, '*Monthly Activity of Influenza-Like Illness (ILI) in the U.S.*' by the Centers for Disease Control & Prevention, CDC.

NEXT STEP

- Preferential vaccination to the medical staff, population in metropolitan regions, or vulnerable age groups.
- Induce habituation of quarantine behavior through health-related campaigns. (e.g. Social distancing, Masks, hand washing, or Sanitizing)
- Determine how to collect data for monitoring and effectiveness of the solution.

OBJECTIVE

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

PROJECT & DATA

- Project Brief
- Data sourced (1980-2016) from VGChartz.
- VGChartz's data collection methodology.

LIMITATIONS

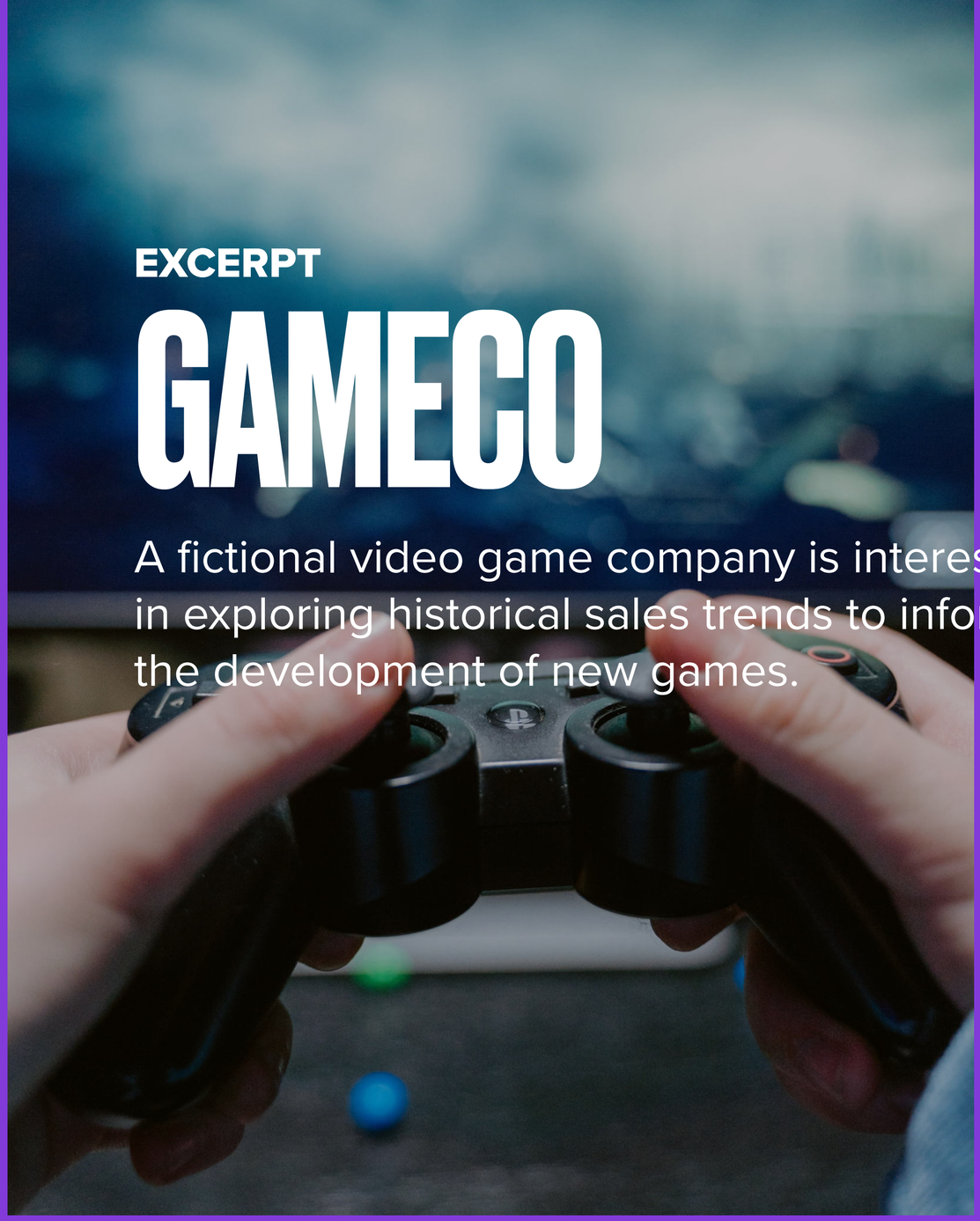
- Tracks the total number of units sold (not financial figures) to the retail stores.
- 2016 is the latest year logged with partial records.

TECHNIQUES APPLIED

- Data Integrity, Quality, and Consistency Assessment
- Data Cleaning
- Pivot Tables
  - Data Grouping & Summarizing
  - Calculated Fields
- Descriptive Analysis
- Excel Visualization Results
- PowerPoint Presentation

TOOLS

Excel PowerPoint



1. DATA CLEANING

Preparing data by removal of duplicate and irrelevant values, imputing missing figures with mean values. Normalizing text formats for accessibility.
(i.e. corrections on typos and names with special characters, and so on).

2. GROUP & SUMMARIZING DATA

Using excel pivot tables to a group and summarize data with categorical filters enhances customize views.

3. DESCRIPTIVE ANALYSIS

An application of basic exploratory data analysis measures central tendency, distribution, and outliers.

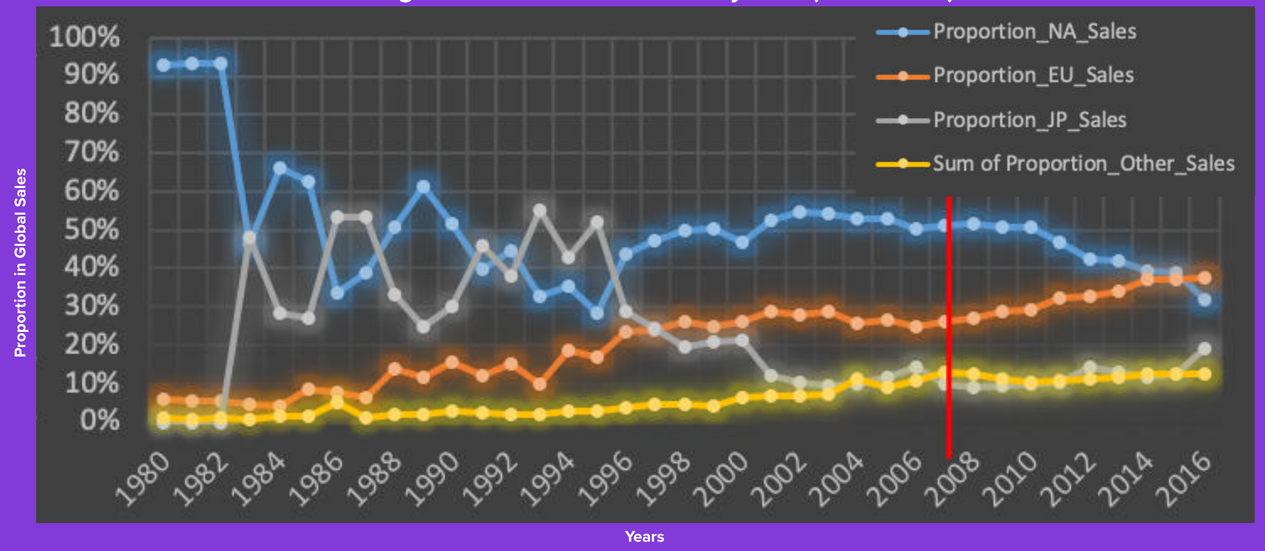
4. EXCEL VISUALIZATION

Data findings are converted to charts illustrating geographical trends by segment. Suitable graphs are used for the final presentation report.

GLOBAL SALES TRENDS

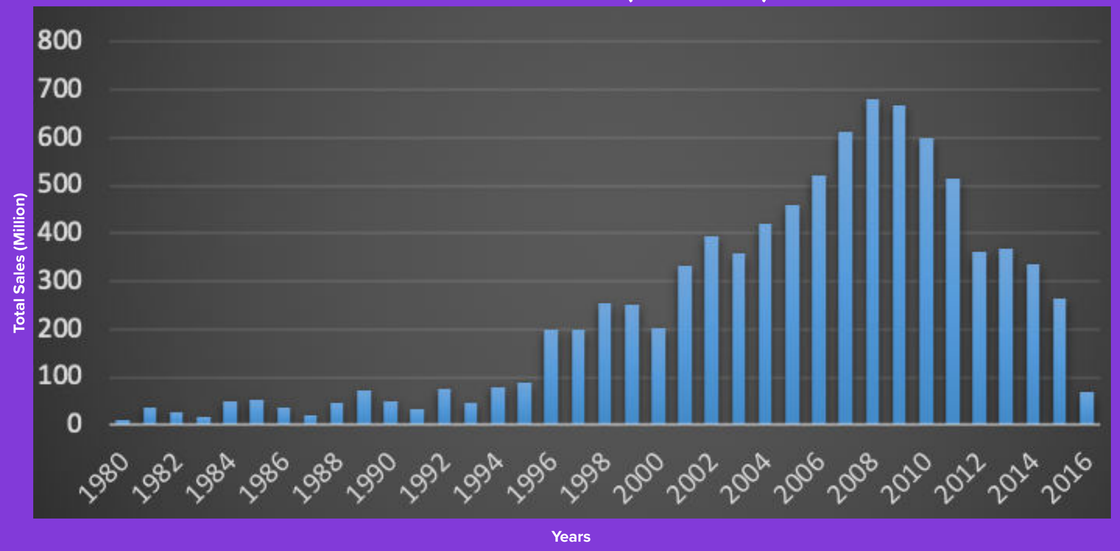
FULL REPORT

Changes of Global Sales Share by Year (1980-2016)



Sales in North 'North America' and 'Japan' flows very oppositely.
Sales in the 'EU' and 'Other' have steadily risen, albeit slightly. (EU > Other)
Sales in 'EU' overtook 'North America' for the first time since 1980, the first year of this dataset.
North America's global share ranges from 28% (1995) to 93% (1980-81).

Global Sales in Total (1980-2016)

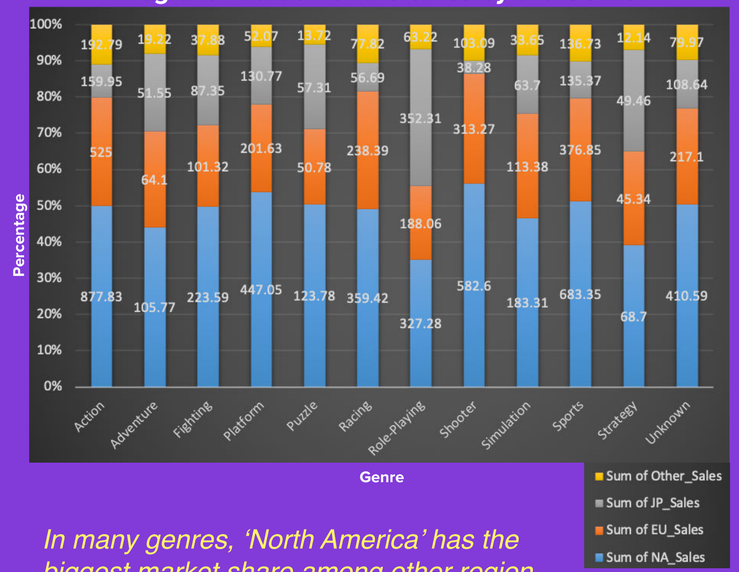


Global sales have fallen sharply since 2008 till now, October 2016.

MARKET ANALYSIS

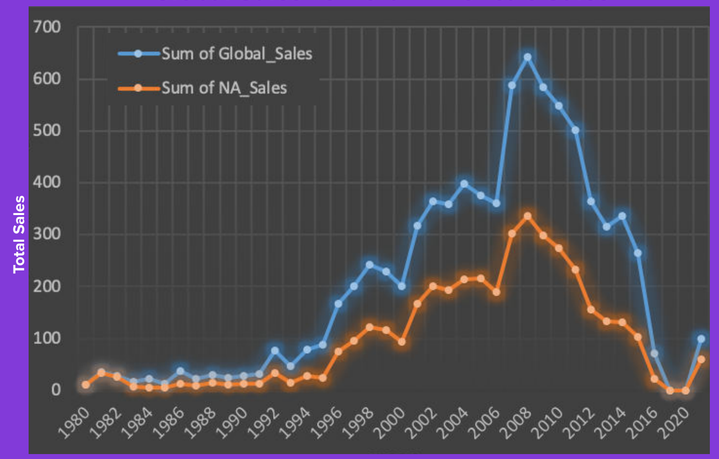
FULL REPORT

Regional Portion of the Sales by the Genre



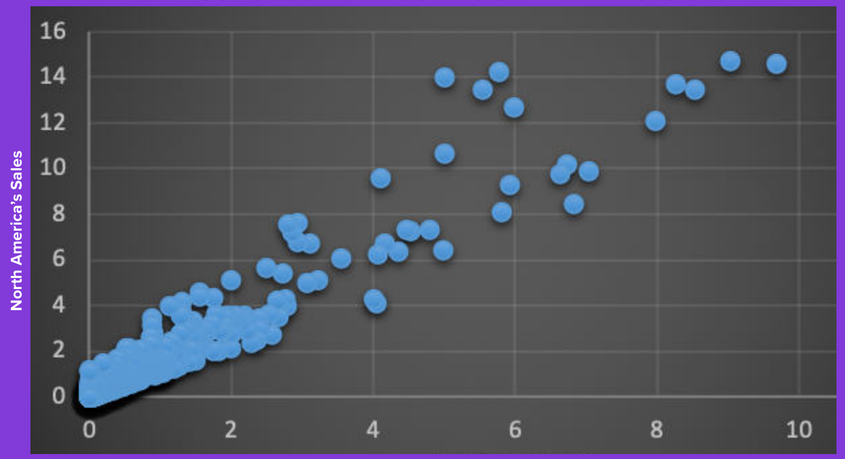
In many genres, 'North America' has the biggest market share among other region. such as Platforms, Shooters, and Sports.

Global Sales vs North America's Sales



*'North America' sales accounts for **a large portion** of the **Global** sales.*

Correlation between North America & Global



Sales in 'NA' and the Global has a strong positive correlation.

GLOBAL MARKET

RECOMMENDATION

REGION

Focus on North America and Europe with active genre.

SALES DEVELOPMENT

- Expand the market to developing countries with large populations such as China, India, and Southeast Asia.
- Consider the field of mobile or online games for populations with low purchasing power.

CHALLENGER

- Develop characters like the Super Mario brothers or Pokemon.
- Collaboration with existing characters such as Lego or Disney.

NEXT STEP

- Collect data on developing countries with large populations such as China, India, and Southeast Asia.
- Analyze major competitors.
- Research the field of mobile or online games.

OBJECTIVE

Contribute to the development and optimization of models to enhance the efficiency of the bank's compliance program in identifying client loss risk factors and flagging suspicious transaction behavior.

PROJECT & DATA

- Project Brief
- Client Dataset I provided by CareerFoundry.

LIMITATIONS

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

TECHNIQUES APPLIED

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

TOOLS

Excel

1. BIG DATA

Preparing data by removal of duplicate and irrelevant values, imputing missing figures with mean values. Normalizing text formats for accessibility.
(i.e. corrections on typos and names with special characters, and so on).

2. DATA ETHICS

Using excel pivot tables to a group and summarize data with categorical filters enhances customize views.

3. DATA MINING

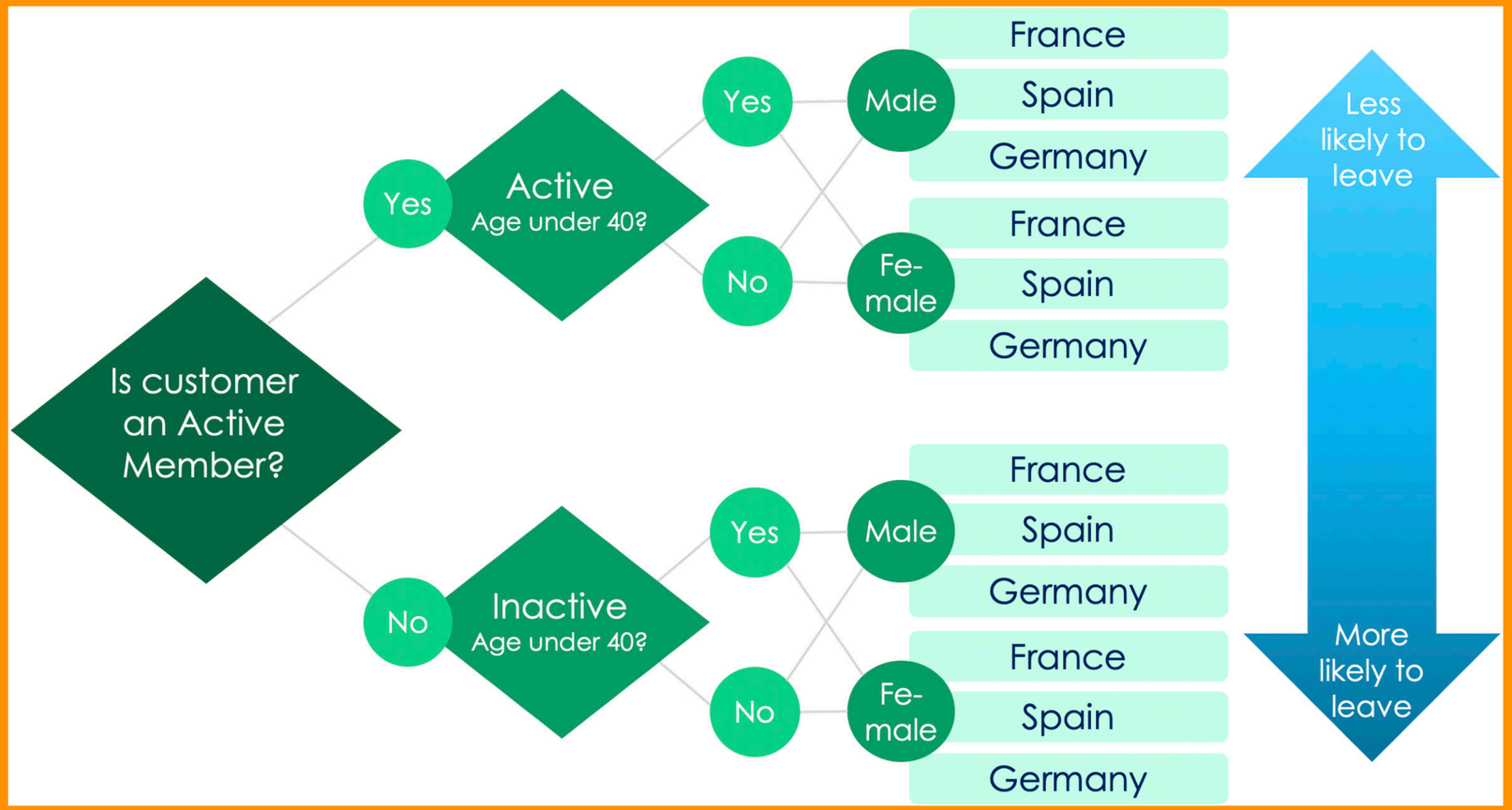
An application of basic exploratory data analysis measures central tendency, distribution, and outliers.

4. PREDICTIVE, TIME SERIES ANALYSIS AND FORECASTING

DOptimizing 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.

PREDICTIVE MODEL ANALYSIS

DATA SET



- *Analysis of client data showed behavioral and demographic variables associated with clients exiting the bank.*

- *Non-members tend to leave the bank more than members.*
- *Females tend to leave the bank more than males.*
- *Germany has the highest rate of exit while France has the lowest rate of exit the bank.*
- *Older customers tend to leave the bank more than younger customers.*
- *Low-credit customers tend to leave the bank more than the high-credit customers.*
- *Customers with high balances tend to leave the bank more than customers with low balances.*
- *High-income customers tend to leave the bank more than low-income customers.*

GLOBAL MARKET

RECOMMENDATION

TARGET CUSTOMERS

- Inactive Member
- Female
- Germany
- Older Age
- Lower Credit Score
- Higher Balance
- Higher Salary

FRAUDULENCE CONTROL

As the number of clients increases, invest time and effort in controlling fraudulence.

NEXT STEP

- Implement banking strategies aimed at driving product sales within the target markets of France and Spain, targeting potential clients and business communities.
- Perform a marketing research analysis on the German market to delve deeper into the factors contributing to the elevated customer attrition risk.

OBJECTIVE

By employing regression, cluster, and time-series analysis, this project aims to find factors determining house prices in King County, WA contributing to understanding of the house market.

PROJECT & DATA

- Project Brief
- Dataset 1 (House sale prices for King County, 2014-2015) from kaggle.com
- Dataset 2 (Real estate market indicators in the U.S., 1996-2023) from data.nasdaq.com

LIMITATIONS

- Conclusions may be limited if the dataset lacks size, scope, or representativeness.
- External influences may not be fully considered, impacting the validity of recommendations.

TECHNIQUES APPLIED

- Data Cleaning & Sourcing
- Exploratory Analysis
  - Linear Relationship
  - Linear Regression
- Cluster & Spatial Analysis
- Time Series Analysis
- Storytelling in Tableau

TOOLS

Excel PowerPoint

1. DATA CLEANING & WRANGLING

Preparing data by removal of duplicate and irrelevant values, imputing missing figures with mean values. Normalizing text formats for accessibility. (i.e. corrections on typos and names with special characters, and so on).

2. CREATING SUBSETS & CLUSTERS

Creating a smaller data set from a whole data set based on a particular filter. Grouping data points in a meaningful way in order to identify similar subgroups (clusters) within the data.

3. EXPLORATORY VISUAL ANALYSIS

To identify relationships between variables.
i.e.) Correlations, Scatterplots, Pair plots, and Categorical plots.

4. SUPERVISED / UNSUPERVISED MACHINE LEARNING

Feed enough data into a machine that it will predict that outcome, linear regression.
Let the algorithms search for patterns in the data to categorize the data points into groups.

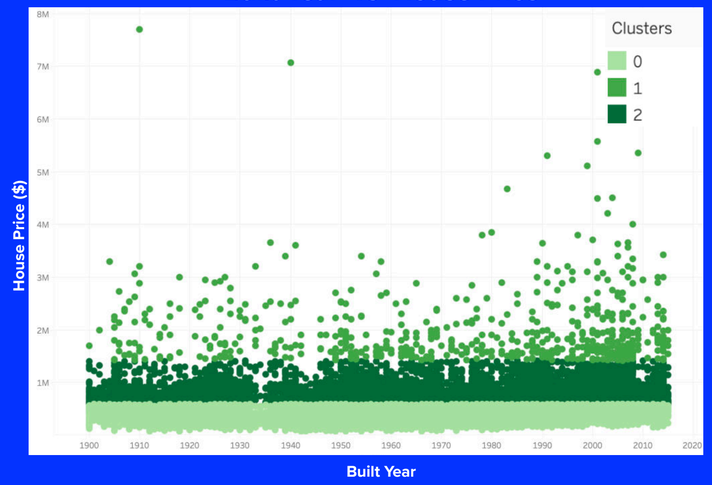
EXPLORATORY ANALYSIS

FULL REPORT

‘Living Space’ vs ‘House Price’

- An increase of 0.00175628 units in 'sqft_living' to 'price.'
- 50% of the variability in 'sqft_living' influences the 'price.'
- The 'sqft_living' is a statistically predictor of 'price.'

‘Built Year’ vs ‘House Price’

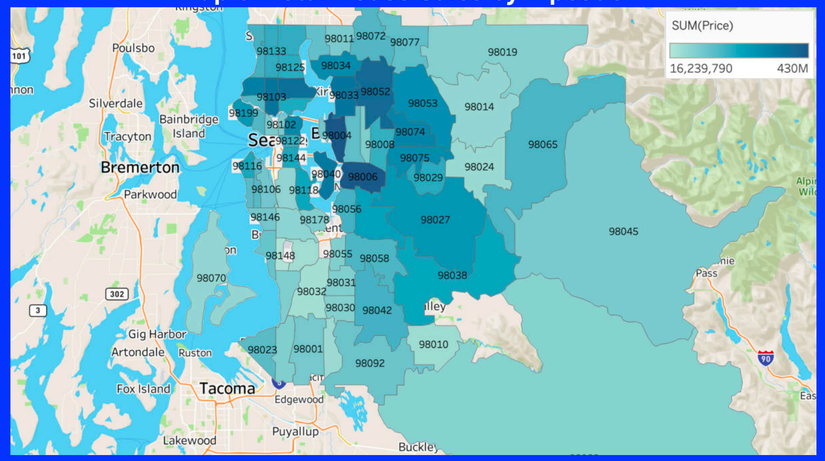


CLUSTER 0 (Light Green): Houses that built between 1900s to 2020 are upto about $580K.

CLUSTER 1 (Green): Houses that built between 1900s to 2020 are between about $1,425K and over.

CLUSTER 2 (Dark Green): Houses that built between 1900s to 2020 are between about $600K to $1,300K.

Map of Total House Sales by Zipcode

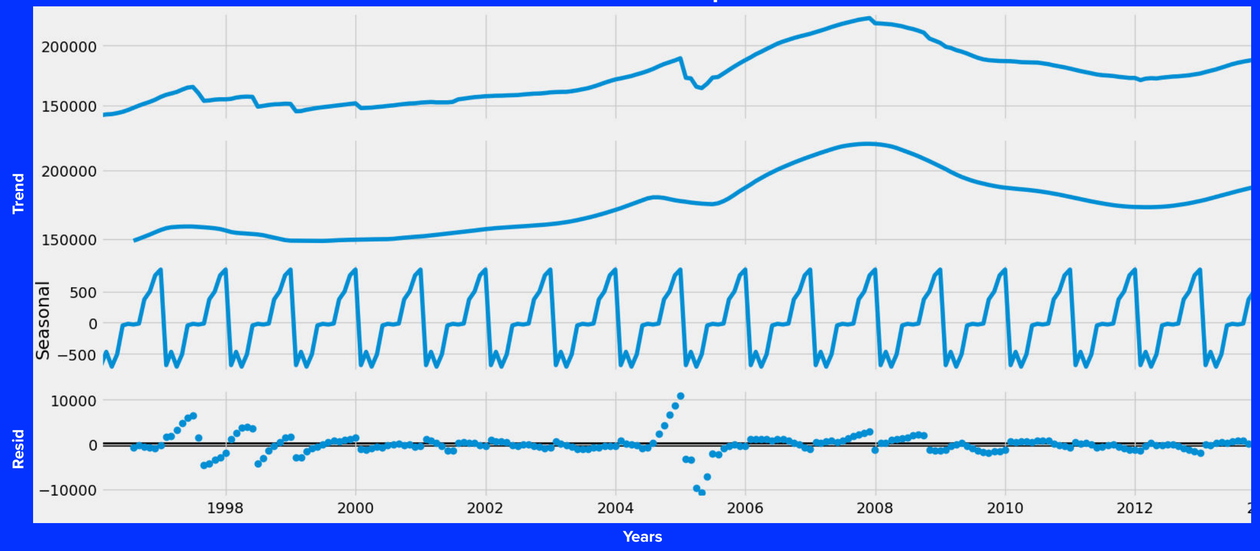


Upon examining the map, it becomes evident that the total house prices in proximity to the water are relatively higher than those in other areas of King County.

TIME SERIES ANALYSIS

FULL REPORT

Time Series Decomposition

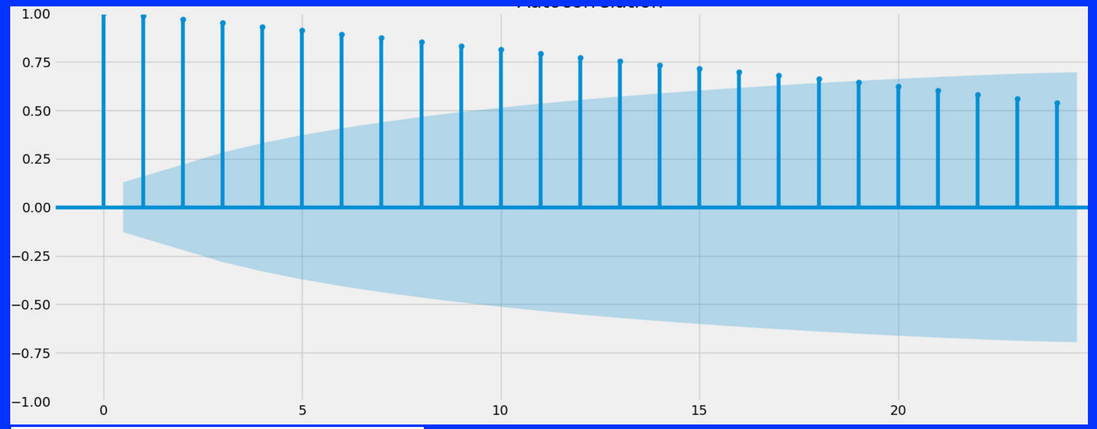


The trend analysis reveals a steady rise in house prices since 1996, with a notable peak in 2008 linked to the subprime mortgage crisis. Post-crisis, the trend rebounded from 2012, indicating a recovery.

Seasonality patterns show price increases from spring to fall, contrasting with winter decreases due to reduced demand.

Residual variations highlight a significant peak in 2004, followed by a correction in 2005 and a more stable market post-2006.

Autocorrelation




```
Dickey-Fuller Stationarity test:
Test Statistic                  -1.206771
p-value                          0.670676
Number of Lags Used              2.000000
Number of Observations Used    230.000000
Critical Value (1%)             -3.459106
Critical Value (5%)             -2.874190
Critical Value (10%)            -2.573512
dtype: float64
```

The test results suggest that the data is likely non-stationary because the p-value is relatively high, and the test statistic is not more negative than the critical values at common significance levels (1%, 5%, 10%). This implies that the data may exhibit trends, seasonality, or other non-stationary characteristics.

HOUSE SALES

RECOMMENDATION

SALES DEVELOPMENT

- Focus on high-performing areas, specifically targeting waterfront properties with a marketing strategy
- Enhance property grades through renovations, upgrades, or new construction.
- Educate potential buyers on factors influencing property value, such as waterfront location, living space size, grade, and age.
- Collaborate with local experts, including real estate professionals, appraisers, and contractors, to gain insights into property values, market trends, and improvement strategies.

LIMITATIONS

- Lack of data size (2014-2015).
- External influences like economic shifts, policy changes, or unforeseen events may not be fully considered.

NEXT STEP

- Increase dataset size and diversity.
- Account for economic indicators, policy changes, and external events.
- Tailor analysis to specific regions, conducting separate analyses.