Data Analyst Portfolio

Eduardo Sanchez Duran

Gmail in GitHub

Welcome to my professional portfolio

- I am *Eduardo Sánchez Durán*, passionate Data Analyst with experience in marketing, international trades, and logistics.
- My work career has focused on the development and strengthening of safety and trust teams in online commercial exchange platforms.
- Play key roles in inventory and logistics management for businesses, especially in the exciting and challenging start-up environment.
- In this portfolio you will find concrete examples of my work as Data Analyst, highlighting projects that reflect my ability to face challenges and achieve exceptional results.
- I invite you to explore my portfolio and discover how my experience can add value to your organization.

PROJECTS

1. MARKETING STRATEGY FOR AN ONLINE GROCERY STORE

2. ANSWERING BUSINESS QUESTIONS FOR AN ONLINE VIDEO RENTAL COMPANY

3.PREPARING FOR FLU SEASON IN THE U.S.

4. ANTI-MONEY LAUNDERING PROJECTS AT A GLOBAL BANK

5. ANALYZING GLOBAL VIDEO GAME SALES

6. YACHT & BOATS SALES WEBSITE OPEN EXPLORATORY ANALYSIS

1.Instacart Grocery Basket Analysis

CONTEXT

Instacart is an online grocery store operating via mobile app. It is working on refining its marketing strategy by conducting targeted marketing campaigns.

OBJECTIVE

The Objective of the analysis aims to discover sales patterns and consumer behaviors for sales and marketing departments to improve their marketing strategy.

Skills & Procedures

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

Data:
Customers Data Set
Instacart Data Sets:
Data Dictionary
Citation (: "The Instacart Online Grocery Shopping Dataset 2017", Accessed from www.instacart.com/datasets/grocery-shopping-2017 via Kaggle on

Tools used:

python jupyter X GitHub

Analysis

There are distinct patterns in ordering behavior throughout the week, weekdays and weekends show different ordering behaviors. Each age group contributes uniquely to the overall distribution of orders. and the 65+ age groups exhibit notable peaks on specific days

instacart

3. Create customer profiles.

3.1. Age group profile

In [11]:
```
# Create age groups
age_bins = [18, 25, 35, 45, 55, 65, 100]
age_labels = ['18-24', '25-34', '35-44', '45-54', '55-64', '65+']
df_dept_cust_ords['age_group'] = pd.cut(df_dept_cust_ords['age'], bins=age_bins, labels=age_labels, right=False)
```

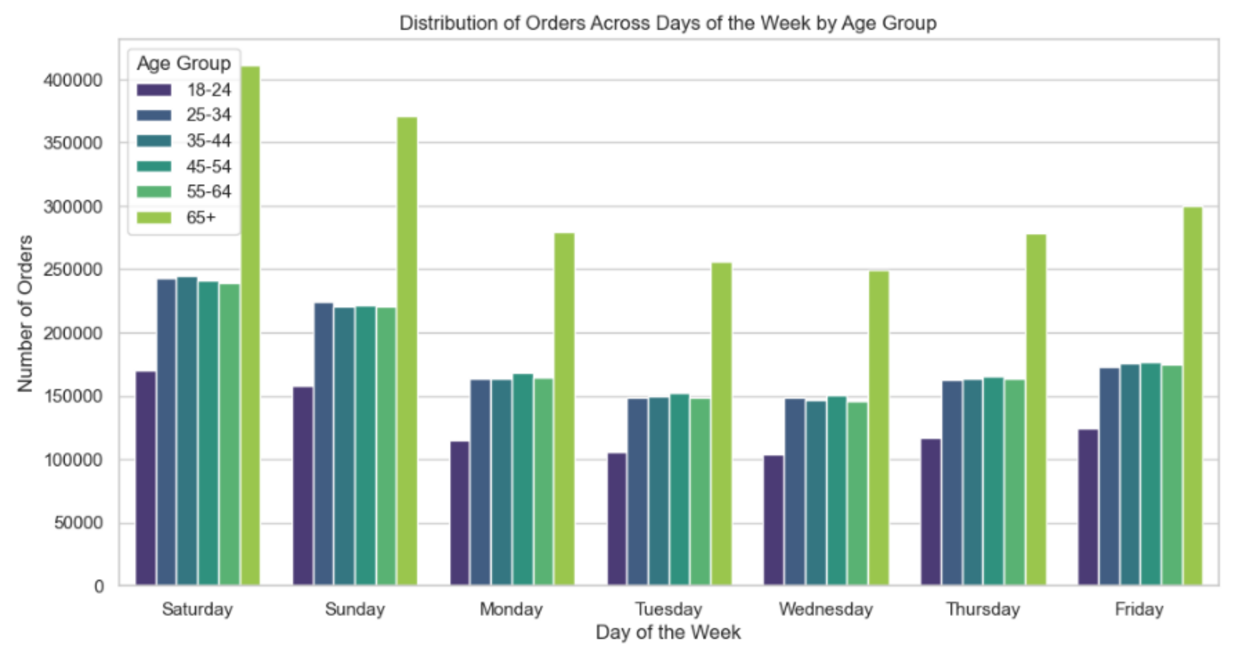
In [12]:
```
# Define a mapping from numerical values to day names
day_of_week_mapping = {
    0: 'Saturday',
    1: 'Sunday',
    2: 'Monday',
    3: 'Tuesday',
    4: 'Wednesday',
    5: 'Thursday',
    6: 'Friday'
}

# Replace numerical values with day names in the 'orders_day_of_week' column
df_dept_cust_ords['orders_day_of_week'] = df_dept_cust_ords['orders_day_of_week'].replace(day_of_week_mapping)
```

In [13]:
```
# Check unique values in the 'orders_day_of_week' column
unique_values = df_dept_cust_ords['orders_day_of_week'].unique()

# Print the unique values
print(unique_values)
```

```
['Saturday' 'Wednesday' 'Friday' 'Sunday' 'Tuesday' 'Monday' 'Thursday']
```

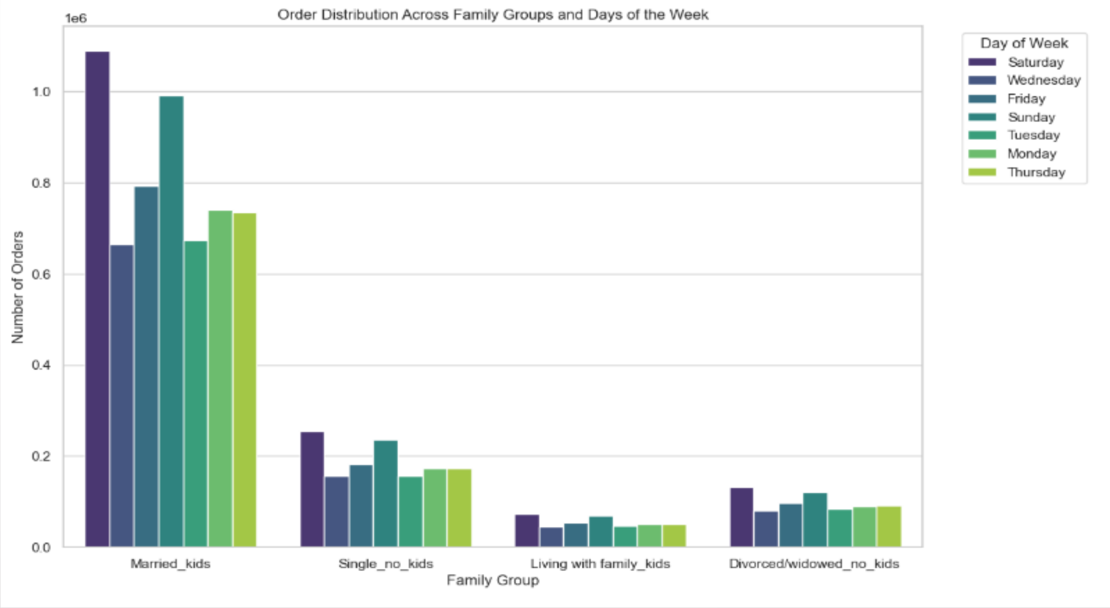
In [14]:
```
# Define the order of days of the week
days_order = ['Saturday','Sunday','Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday']
```

In [15]:
```
# Set the style for the plot
sns.set(style="whitegrid")

# Plot a grouped bar chart to show the distribution of orders for each age group across days of the week
plt.figure(figsize=(12, 6))
sns.countplot(x='orders_day_of_week', hue='age_group', data=df_dept_cust_ords, order=days_order, palette='viridis')
plt.title('Distribution of Orders Across Days of the Week by Age Group')
plt.xlabel('Day of the Week')
plt.ylabel('Number of Orders')
plt.legend(title='Age Group')
plt.show()
```



Analysis

Observing the bars corresponding to Married_kids and Single_no_kids for each day of the week, identify which days have higher order volumes for each family group. Married_kids tend to order more on Saturdays and Single_no_kids on weekdays, the marketing team can strategize promotions or targeted advertising to maximize engagement on these days.



3.3. Family Group profile

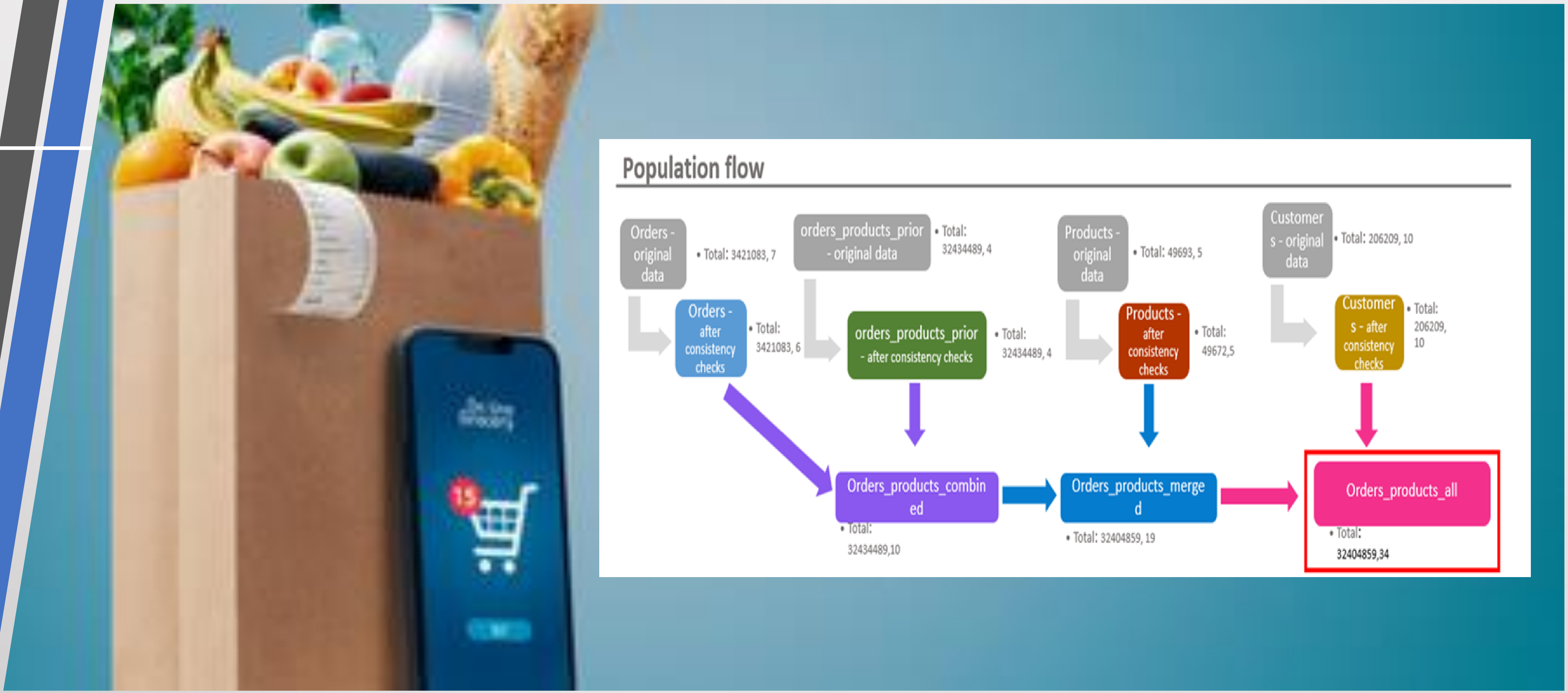
In [25]:
```
# Check the unique values in the 'family_status' column
unique_family_status = df_dept_cust_ords['family_status'].unique()

# Print the unique values
print(unique_family_status)
```

```
['married' 'single' 'living with parents and siblings' 'divorced/widowed']
```

In [26]:
```
# Define a custom function to determine family_groups
def determine_family_groups(row):
    if row['family_status'] == 'married' and row['n_dependants'] == 0:
        return 'Married_no_kids'
    elif row['family_status'] == 'married' and row['n_dependants'] >= 1:
        return 'Married_kids'
    elif row['family_status'] == 'single' and row['n_dependants'] >= 1:
        return 'Single_kids'
    elif row['family_status'] == 'single' and row['n_dependants'] == 0:
        return 'Single_no_kids'
    elif row['family_status'] == 'divorced/widowed' and row['n_dependants'] == 0:
        return 'Divorced/widowed_no_kids'
    elif row['family_status'] == 'divorced/widowed' and row['n_dependants'] >= 1:
        return 'Divorced/widowed_kids'
    elif row['family_status'] == 'living with parents and siblings' and row['n_dependants'] == 0:
        return 'Living with family_no_kids'
    elif row['family_status'] == 'living with parents and siblings' and row['n_dependants'] >= 1:
        return 'Living with family_kids'
    else:
        return 'Other'

# Apply the custom function to create the 'family_groups' column
df_dept_cust_ords['family_groups'] = df_dept_cust_ords.apply(determine_family_groups, axis=1)

# Checking for NaN values
df_dept_cust_ords['family_groups'].value_counts(dropna=False)
```

Out[26]:
```
family_groups
Married_kids                5688380
Single_no_kids              1330529
Divorced/widowed_no_kids     694276
Living with family_kids      388030
Name: count, dtype: int64
```

In [27]:
```
# Set the style for the plot
sns.set(style="whitegrid")

# Plot a grouped bar chart to show the distribution of orders_day_of_week across family groups
plt.figure(figsize=(12, 8))
sns.countplot(x='family_groups', hue='orders_day_of_week', data=df_dept_cust_ords, palette='viridis')

# Set plot labels and title
plt.xlabel('Family Group')
plt.ylabel('Number of Orders')
plt.title('Order Distribution Across Family Groups and Days of the Week')

# Show the plot
plt.legend(title='Day of Week', loc='upper right', bbox_to_anchor=(1.2, 1))
plt.show()
```


Population flow
Orders - original data
• Total: 3421083, 7
orders_products_prior - original data
• Total: 32434489, 4
Products - original data
• Total: 49693, 5
Customers - original data
• Total: 206209, 10
Orders - after consistency checks
• Total: 3421083, 6
orders_products_prior - after consistency checks
• Total: 32434489, 4
Products - after consistency checks
• Total: 49672,5
Customers - after consistency checks
• Total: 206209, 10
Orders_products_combined
• Total: 32434489,10
Orders_products_merged
• Total: 32404859, 19
Orders_products_all
• Total: 32404859,34

RECOMMENDATIONS

It see some patterns in consumer behavior, in some of the profiles created, which will allow the Marketing department to emphasize those segments, such as the hours of greatest traffic in the app, or the days of the week with the highest consumption, or some clients depending on their family status or income.

For more details you can consult the project analysis and the final report here : GitHub Repository

2.Rockbuster Stealth LLC Analysis

- Context: Rockbuster Stealth LLC is a movie rental company that used to have stores around the world. Facing stiff competition from streaming services such as Netflix and Amazon Prime, the management team is planning to use its existing movie licenses to launch an online video rental service to stay competitive.
- Objective: Answer the questions posed by the different departments, query the data using SQL. they expect data-driven answers that they can use for their 2020 company strategy

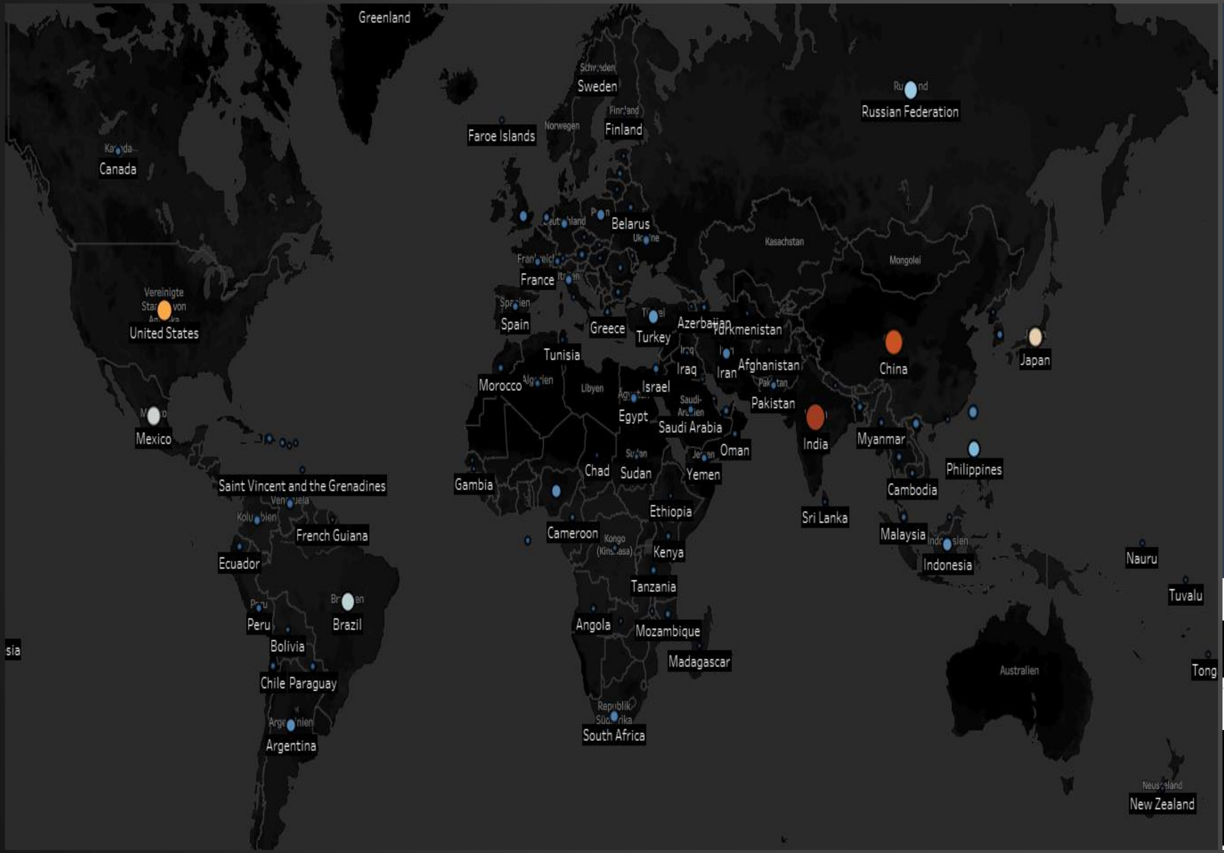
Skills & Procedures

- Relational databases
- Database querying
- Filtering
- Cleaning and summarizing
- Joining tables
- Subqueries
- Create Visualizations of SQL Results
- Create Entity Relationship Diagram (ERD)
- Create Data Dictionary

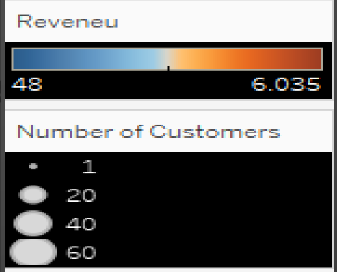
Data:
data set contains information about Rockbuster's film inventory, customers, and payments, among other things.

Tools Used:

GEOGRAPHICAL ROCKBUSTER CUSTOMER COUNT AND TOTAL PAYMENT RECEIVED IN EACH COUNTRY




```sql
-- Top 10 Cities Customer Count

SELECT D.country, C.city,
       COUNT(customer_id) AS customer_count
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
WHERE D.country IN ('India', 'China', 'United States', 'Japan', 'Mexico', 'Brazil',
                  'Russian Federation', 'Philippines', 'Turkey', 'Indonesia')
GROUP BY D.country, C.city
ORDER BY customer_count DESC
LIMIT 10;
```



Top 5 customers

Customer id	First Name	Last Name	Country	City	Total Amount Paid
225	Arlene	Harvey	India	Ambattur	$ 111.76
424	Kyle	Spurlock	China	Shanwei	$ 109.71
240	Marlene	Welch	Japan	Iwaki	$ 106.77
486	Glen	Talbert	Mexico	Acua	$ 100.77
537	Clinton	Buford	United States	Aurora	$ 98.76

```sql
-- Top 5 Customers from Top 10 Cities
WITH top_5_customers_cte AS (
    SELECT
        B.customer_id AS Customer_id,
        B.first_name AS Customer_first_Name,
        B.last_name AS Customer_Last_Name,
        E.country AS Country,
        D.city AS City,
        SUM(A.amount) AS Total_Amount_paid
    FROM payment A
        INNER JOIN customer B ON A.customer_id = B.customer_id
        INNER JOIN address C ON B.address_id = C.address_id
        INNER JOIN city D ON C.city_id = D.city_id
        INNER JOIN country E ON D.country_id = E.country_id
    WHERE city IN ('Aurora', 'Acua', 'Citrus Heights', 'Iwaki', 'Ambattur', 'Shanwei',
                   'So Leopoldo', 'Teboksary', 'Tianjin', 'Cianjur')
    GROUP BY B.customer_id, B.first_name, B.last_name, E.country, D.city
    ORDER BY Total_Amount_Paid DESC
    LIMIT 5
)

-- Main Query using the CTE
SELECT *
FROM top_5_customers_cte;
```

Recommendations

- Understand customers preferences and habits to improve the selection of movies for the online video service. Consider using customer data to offer personalized recommendations and tailor marketing efforts.

- To maximize revenue, focus on licensing and promoting movies that have historically contributed the most to revenue gain. This could involve securing exclusive rights to popular titles or ensuring a diverse catalog that appeals to a wide audience.

For more details about the development of the project, final report and final presentation, consult the following link: GitHub Repository

GitHub

3. Preparing for U.S. Influenza Season

Context: The United States has an influenza season where more people than usual suffer from the flu. Some people, particularly those in vulnerable populations, develop serious complications and end up in the hospital. Hospitals and clinics need additional staff to adequately treat these extra patients.

Objective: Determine when to send staff, and how many, to each state. As analyst, I need technical skills to analyze the data and soft skills to communicate the insights to stakeholders.

- Translating business requirements
- Data cleaning
- Data integration
- Data transformation
- Statistical hypothesis testing
- Visual analysis
- Forecasting
- Storytelling in Tableau

Tools Used :

Data Sets

1. Influenza deaths by geography
Source: CDC

2. Population data by geography, time, age, and gender
Source: US Census Bureau

3. Survey
Source: CDC (Fluview)
Download Influenza Visits Data Set
Download Lab Tests Data Set

4. Survey of flu shot rates in children
Source: CDC

1

Dependent: Flu Mortality, percent of population flu deaths.

Independent: population range age groups.

2

a. ***Null Hypotheses:*** Mortality rates do not differ for the population over 65 years of age.

b. ***Alternative Hypotheses:*** People over 65 have a higher risk of dying from the flu than those under 65.

c. ***One-tailed,*** because we aim to the most vulnerable population in general terms, it takes us in only one direction.

d. **Alpha** = 0.05.

3

T-test Stadistics

4

a. ***P-value:*** -2.6793512475552

b. The P-value its smaller than alfa = 0.05, that means the null hypothesis can be ruled out

5

It was found that people over 65 years of age have a greater risk of dying from the flu, thus rejecting the null hypothesis that assumes the opposite.

6

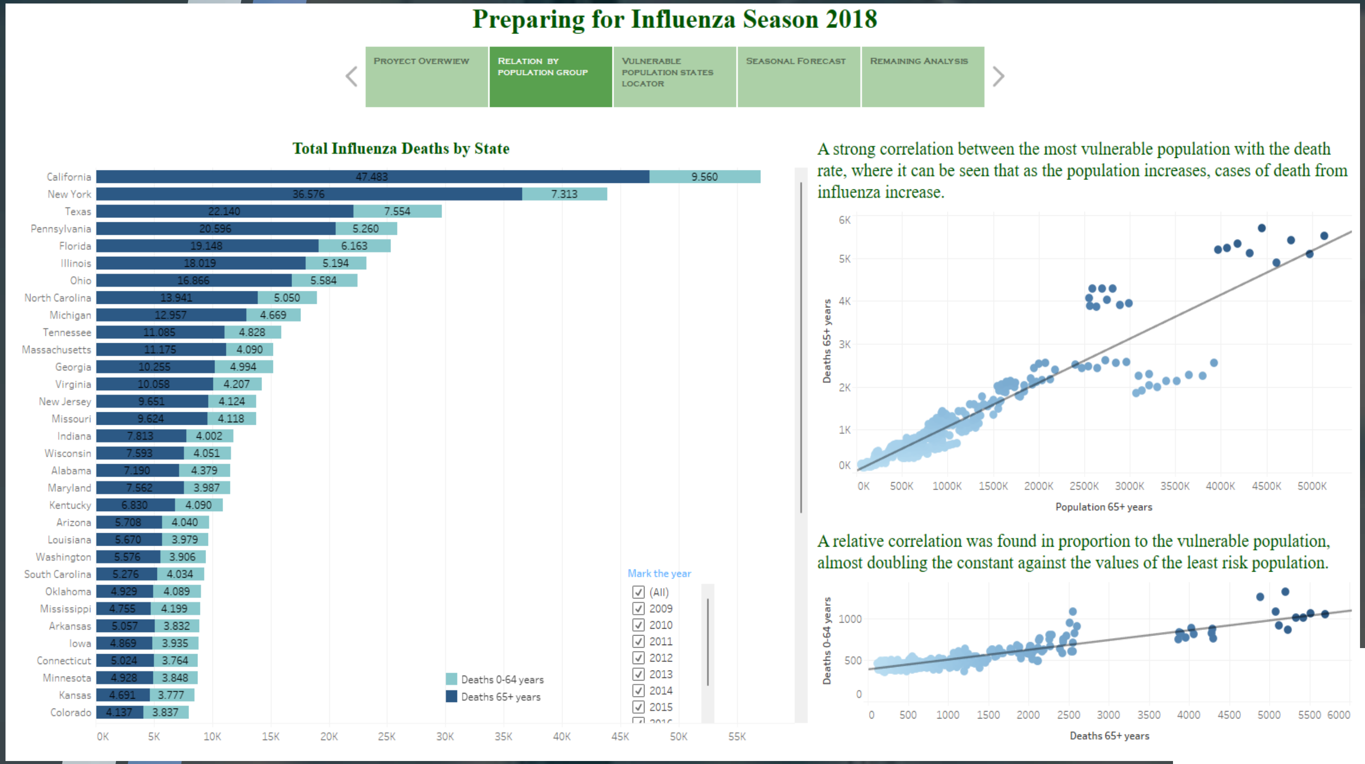
According to these results, we were able to verify the alternative hypothesis, which leads us to focus safely on the vulnerable population, and to be able to prepare a logistical program in which regions more staff will be needed for the next season.

t-Test: Two-Sample Assuming Unequal Variances

	0-64 years % Deaths	*65-85+ years % Deaths*
Mean	0.000198585	0.000246477
Variance	4.0994E-08	8.96816E-08
Observations	409	409
Hypothesized Mean Difference	0	
df	717	
t Stat	-2.679351248	
P(T<=t) one-tail	0.003772845	
t Critical one-tail	1.646981593	
P(T<=t) two-tail	0.007545691	
t Critical two-tail	1.963278089	

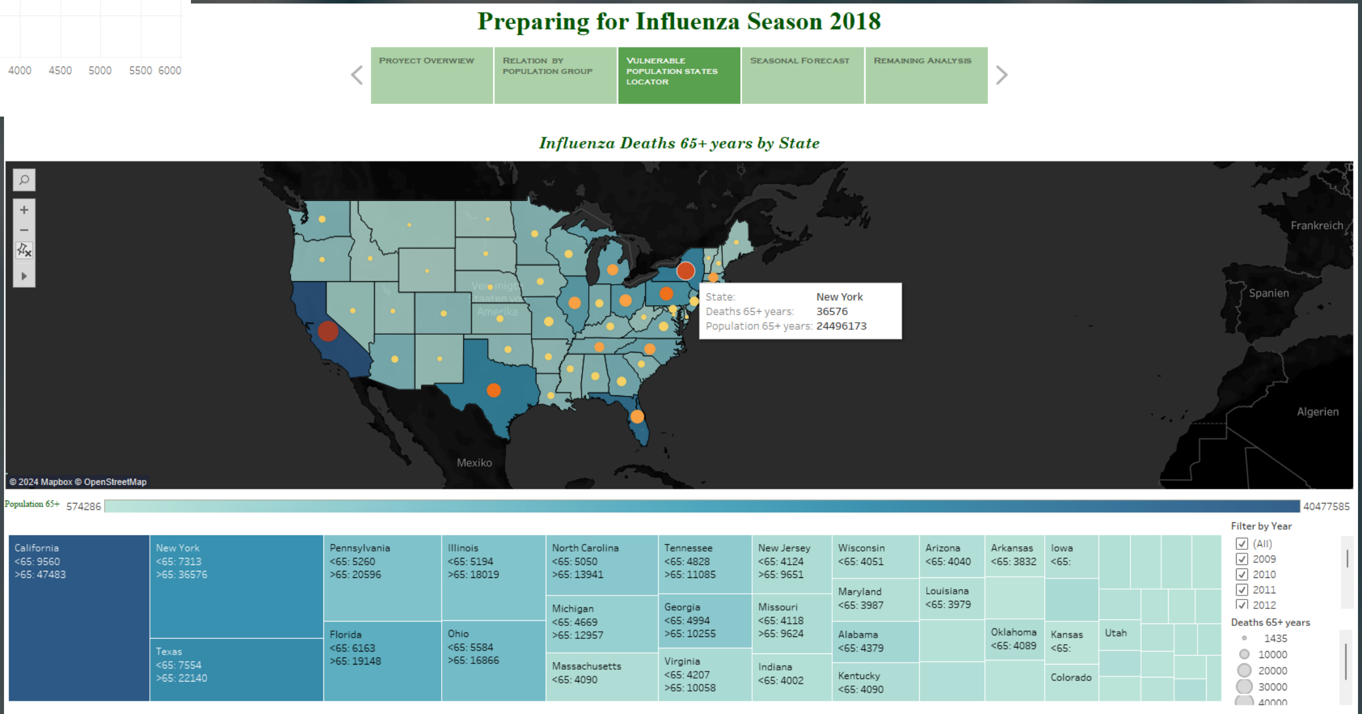
0.3 Pearson's correlation coefficient

Analysis T-test Stadistics

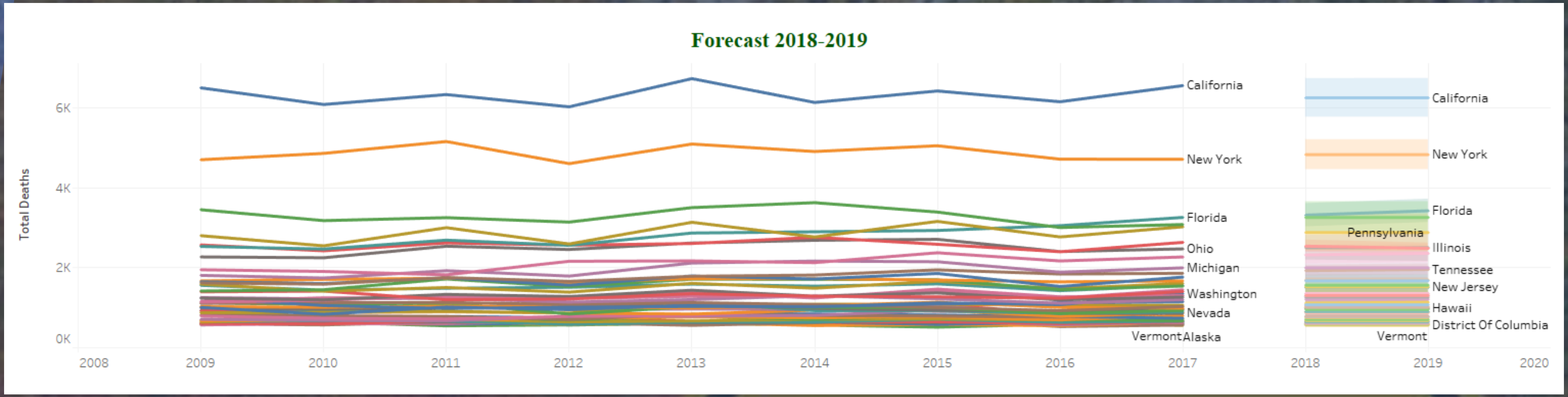


Visual Analysis

Tableau Storytelling



The most vulnerable population segments identified in higher risk areas, and finding a relationship between the variables



Conclusions

A count of historic flu deaths tells us the places where we should focus more attention, sending more medical personnel, mainly in the winter season

Prioritize states with large vulnerable populations, with high influenza mortality.

Presenting results to an audience

4.PIG e-Bank
Anti-Money Laundering Analysis

PIG e-bank is a fictitious bank created to learn the skill to develop

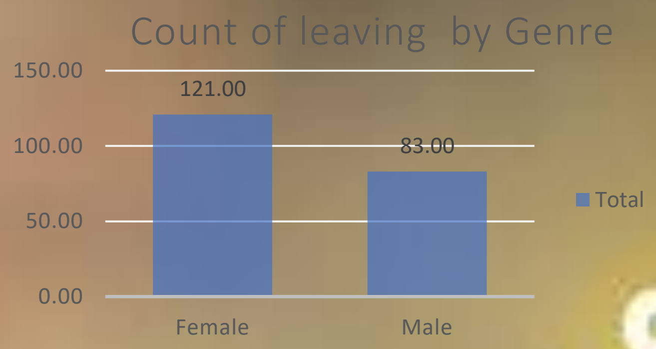
- *Objective* : To increase customer retention, the sales team wants to identify the leading indicators that a customer will leave the bank. Identify the top risk factors that contribute to client loss and model them in a decision tree.

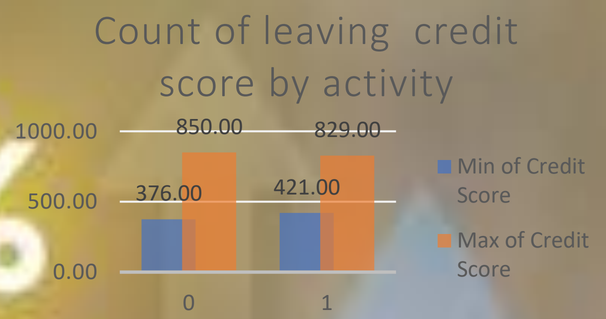
Skills

- Data ethics
- Data mining
- Predictive analysis
- Time series analysis and
- forecasting

Tools Used:

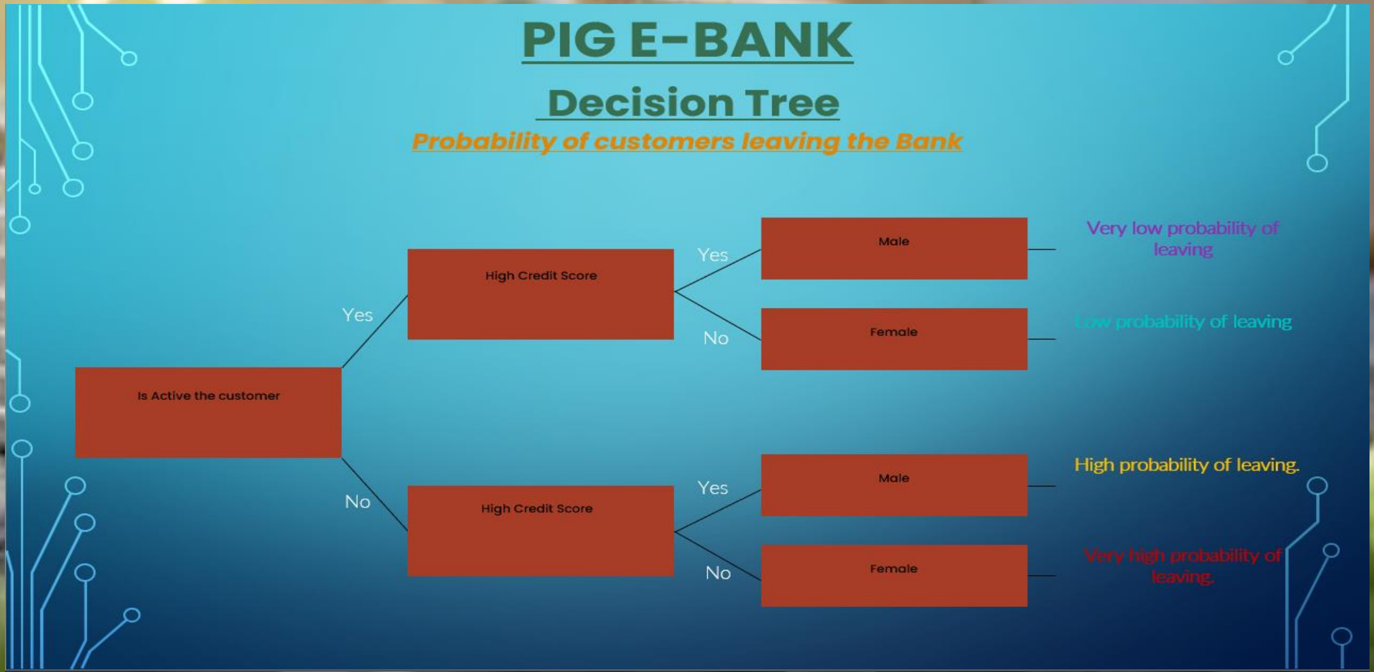
Analysis & Conclusions





The decision tree to determine the probability of customers leaving the bank. Haven't been active, Are female , low credit score.

Measures such as targeted marketing campaigns, incentives to maintain activity, or even new features in the banking app, should be taken to retain these customers



5. GameCo
Analyzing global video game sales

Context: GameCo is a global video game company planning a strategic campaign to boost sales. To inform their strategy, they aim to identify regions with the highest profit margins, the most popular platforms, best-selling video games, and historical sales trends.

Objective: Analyze sales across regions to get an idea of how games are performing in the most popular market and filter the data to look at specific subsegments.

Skills & Procedures

- Excel
- Grouping data
- Summarizing data
- Descriptive analysis
- Visualizing results in Excel
- Presenting results

Data:
vgsales_dirty.xlsx

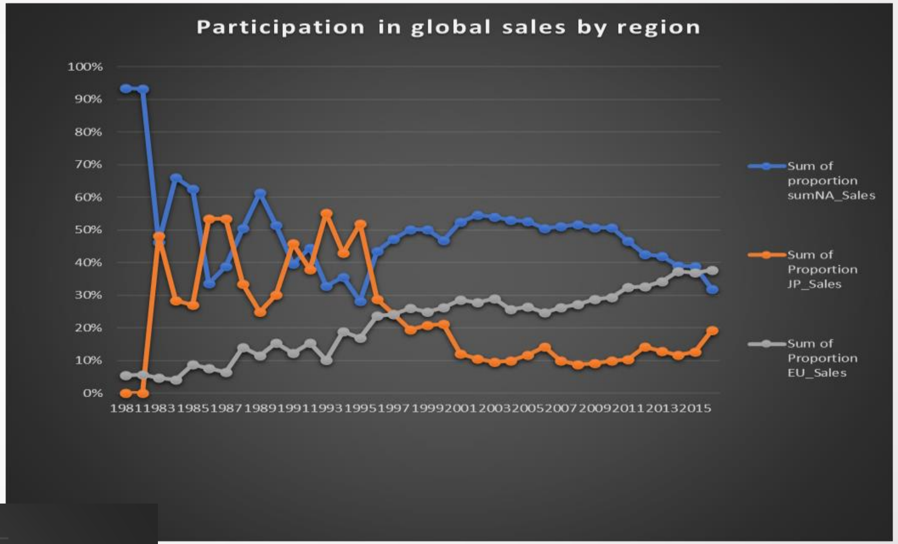
Source: VGChartz.

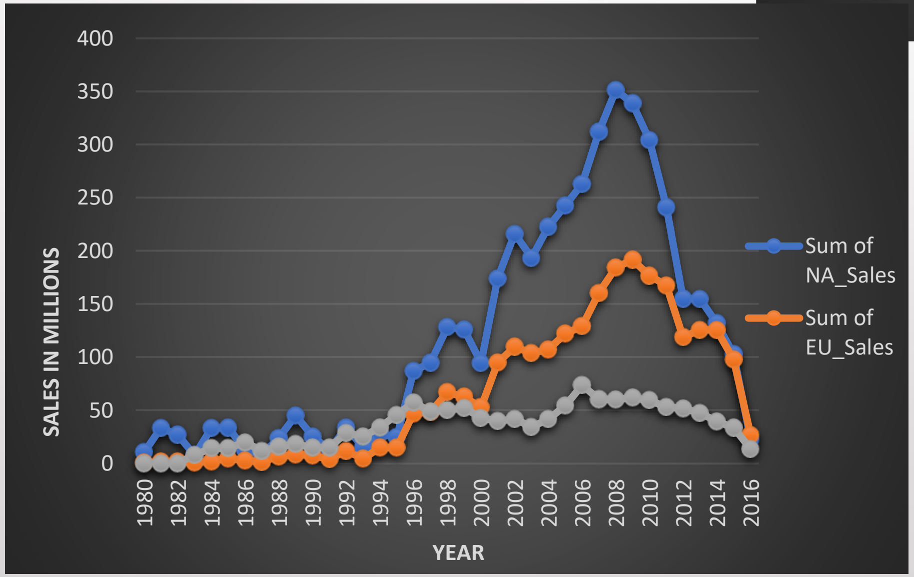
Tools Used:

Analysis & Conclusions

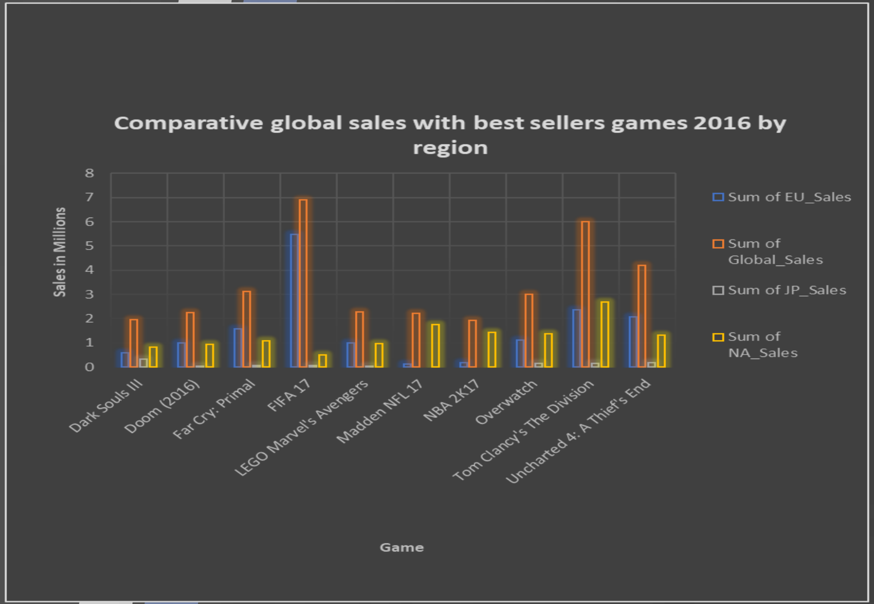
Global sales trends show divergence and negative correlation between Japan and North America, while the Euro zone exhibits a positive trend.

This challenges the expectation of uniformity, highlighting the need to focus on specific changes in each region for a nuanced analysis.





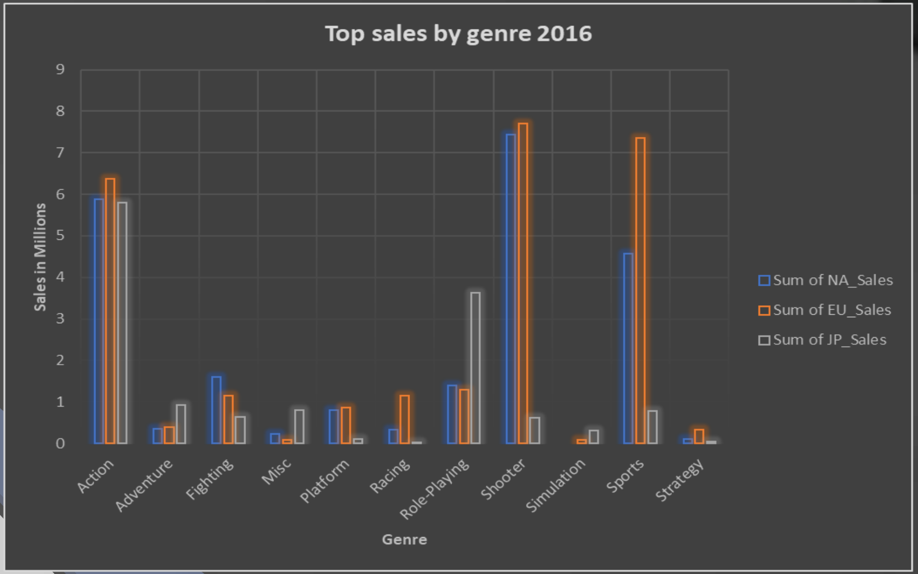
Since the 1980, North America has consistently led in video game sales. A universal boom occurred from 1996 to 2008, followed by a simultaneous downturn across all regions, persisting to the present.



Analysis & Conclusions

Final presentation

FIFA 2017 is the top-selling video game globally, with Europe leading in sales.



Reveals a consistent pattern where the top three video game genres; shooter, sports, and action, continue to dominate sales across all three regions.

6. Nearly New Nautical Analysis Yacht & Boats website.

Context

"Nearly New Nautical" it's a Yacht and Boat sales website which allows users to advertise their new and used boats for sale.

The marketing team is preparing a weekly newsletter for boat owners.

Objective

The newsletter is designed to help sellers get more views of their boats, in addition to being aware of market trends.

They would like you look at the recent data and get some ideas.

Skills & Procedures

Analysis Criteria:

- Exploratory analysis through visualizations (scatterplots, correlation heatmaps, pair plots, and categorical plots)
- Geospatial analysis using a shapefile
- Regression analysis
- Cluster analysis
- Time-series analysis
- k-means clustering
- Analysis narrative and results (presented in the Tableau Dashboard).

Data Source:

BOAT SALES ANALYSIS

- This is an external data source.
- The data is provided by Kaggle.

License:

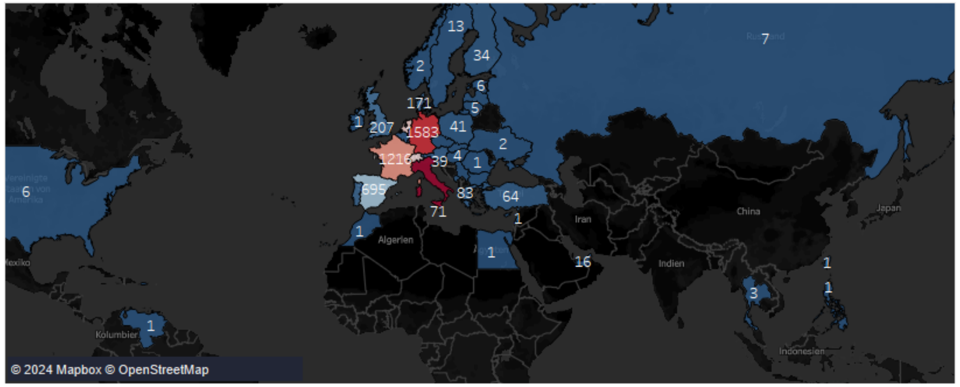
CCo: Public Domain

Tools Used:

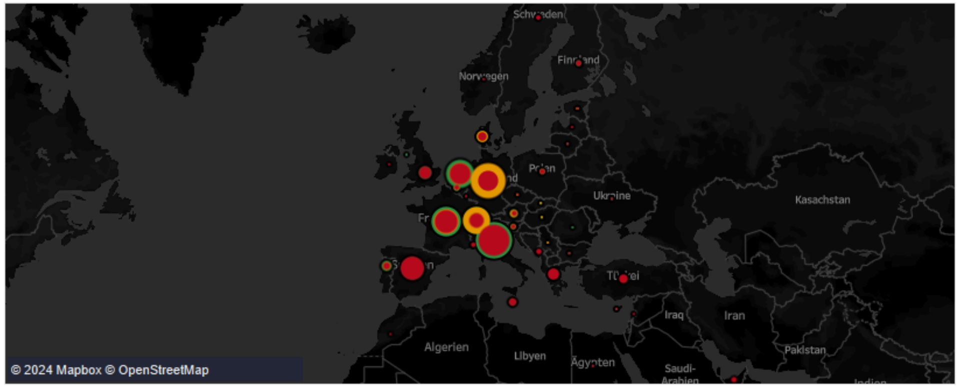
Geospatial Analysis

```
In [27]: # Setup a Folium map at a high-level zoom
         map = folium.Map(location=[0, 0], zoom_start=2)

         # Choropleth map binding Pandas DataFrames and GeoJSON geometries
         folium.Choropleth(
             geo_data=country_geo,
             data=data_to_plot,
             columns=['Country', 'Last 7 days views'],
             key_on='feature.properties.name',  # Check your GeoJSON file to find the correct key
             fill_color='YlOrBr',
             fill_opacity=0.6,
             line_opacity=0.1,
             legend_name="Last 7 days views"
         ).add_to(map)

         # Add a layer control to the map
         folium.LayerControl().add_to(map)

         # Display the map
         map
```

Last 7 days views



Countries by category of price



In the map above we can see the users by country, showing the countries with the most views on the platform, from the blue color with the least views to the red color with the most views; At the same time, on the map below we can see the countries by price category, finding the highest concentration of views and users in central Europe with a high number of visits in the countries of Germany, France and Italy, which could lead us to focus on these countries, something that could be observed is that Germany has a high number of visits, when clicking on the area you will see high volumes of advertising and low prices for the country as a whole, the opposite case for Italy, which has a lower volume of low-cost ads and more high-priced ads.

Price category

- ☑ (All)
- ☑ High price
- ☑ Low price

Price category

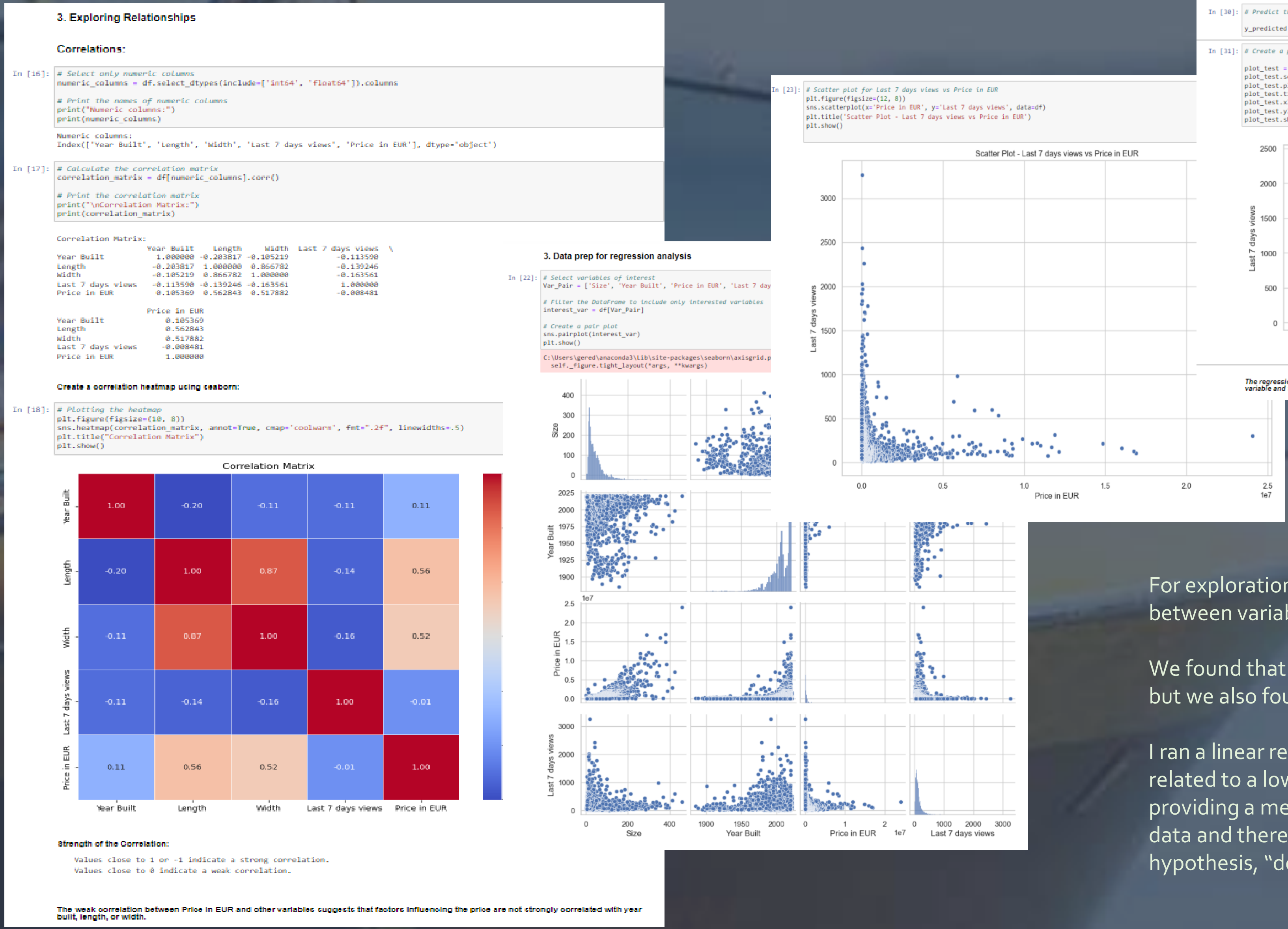
- High price
- Low price
- Middle price

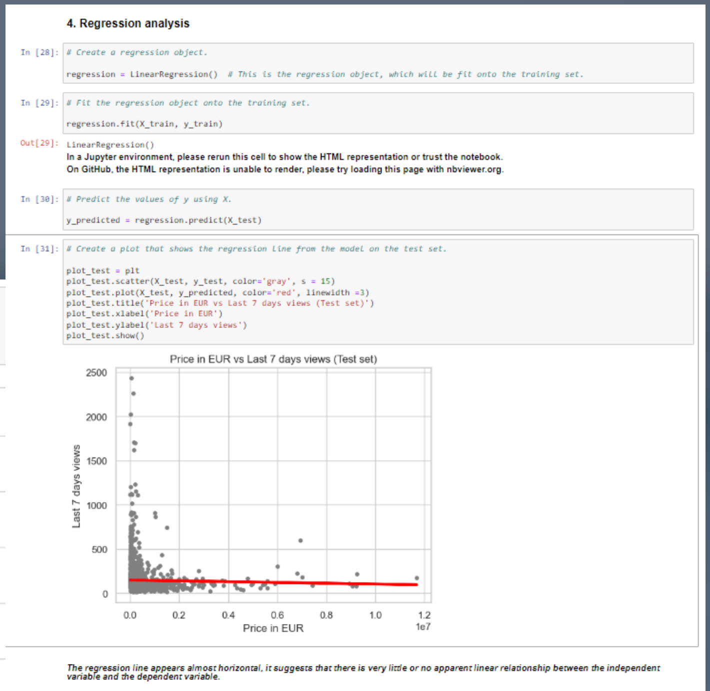
The analysis allows us to see the countries where the ads are located, showing us the views in each country.

Highlighting the countries with the greatest visualization, in the most intense red color to the faintest, and the countries in blue are those with the least visits on the website.

In turn we can see the other map where we see the countries with high price ranges low price, finding greater activity in the central European area.

Exploration Analysis and Linear Regression



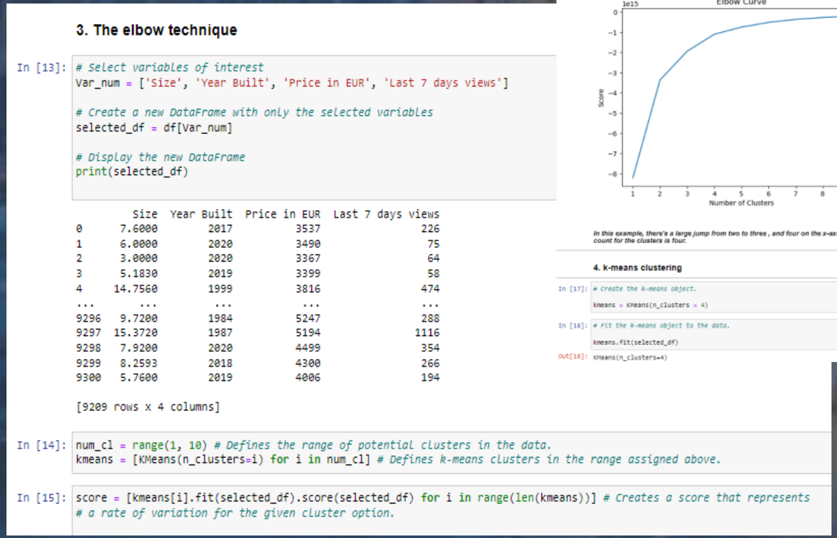


For exploration purposes, we started looking for relationships between variables last 7 days views and Price in EUR

We found that there is not dependence between the variables, but we also found some connection between each other

I ran a linear regression, which showed the increase in views related to a lower price, and it appears that the model is not providing a meaningful explanation for the variability in the data and there is not enough evidence to affirm the hypothesis, “do views increase when the price is lower?”

Clusters Analysis



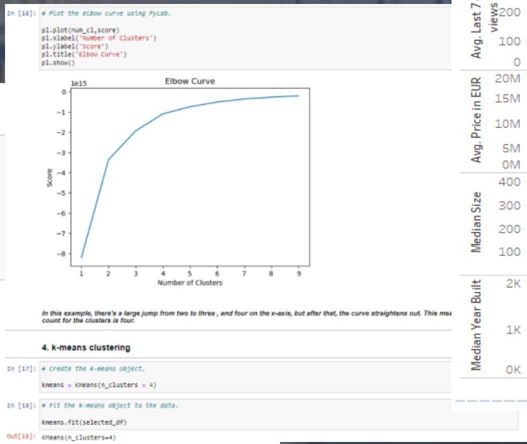
3. The elbow technique

```
In [13]: # Select variables of interest
         Var_num = ['Size', 'Year Built', 'Price in EUR', 'Last 7 days views']

         # Create a new DataFrame with only the selected variables
         selected_df = df[Var_num]

         # Display the new DataFrame
         print(selected_df)
```

	Size	Year Built	Price in EUR	Last 7 days views
0	7.6000	2017	3537	226
1	6.0000	2020	3490	75
2	3.0000	2020	3367	64
3	5.1830	2019	3399	58
4	14.7560	1999	3816	474
...	...	...	...	...
9296	9.7200	1984	5247	288
9297	15.3720	1987	5194	1116
9298	7.9200	2020	4499	354
9299	8.2593	2018	4300	266
9300	5.7600	2019	4006	194

[9209 rows x 4 columns]

```
In [14]: num_cl = range(1, 10) # Defines the range of potential clusters in the data.
         kmeans = [KMeans(n_clusters=i) for i in num_cl] # Defines k-means clusters in the range assigned above.

In [15]: score = [kmeans[i].fit(selected_df).score(selected_df) for i in range(len(kmeans))] # Creates a score that represents
         # a rate of variation for the given cluster option.
```

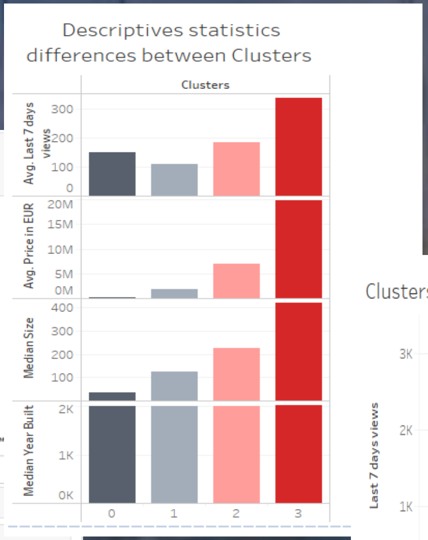


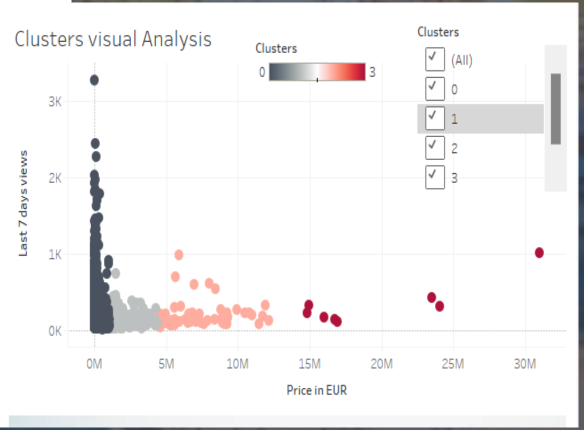
```
In [16]: # Plot the elbow curve using PyLab.

         pl.plot(num_cl,score)
         pl.xlabel('Number of Clusters')
         pl.ylabel('Score')
         pl.title('Elbow Curve')
         pl.show()
```


In this example, there's a large jump from two to three , and four on the x-axis, but after that, the curve straightens out. This mea... count for the clusters is four.

4. k-means clustering

```
In [17]: # Create the k-means object.

         kmeans = KMeans(n_clusters = 4)

In [18]: # Fit the k-means object to the data.

         kmeans.fit(selected_df)
Out[18]: KMeans(n_clusters=4)
```





Dark Gray cluster shows the lowest average price and the smallest median size, and the model year of manufacture is older.

Light Gray Cluster shows a noticeable increase in price, as its size and the model year of manufacture, it is a little more modern than the previous one, but the views decreased on average compared to the previous dark gray cluster

Pink Cluster you can notice an increase in price, and size and at the same time the year of manufacture is more recent.

Red Cluster shows high popularity, with fewer counts, showing larger size and higher price range, with high views in its few ads.

Time-series Analysis

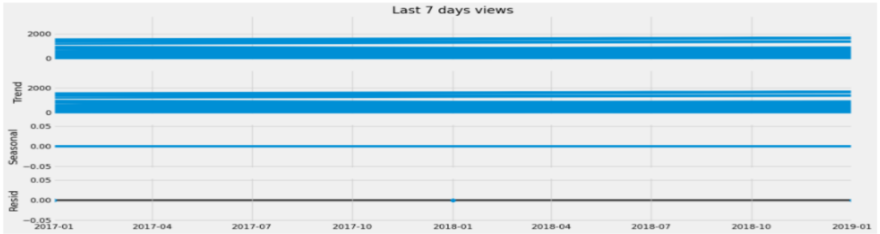
3. Time-series analysis: decomposition

In [11]:
```python
period = 1  # Specify the period (assuming yearly data)

# Infer the frequency from the existing datetime index
time_df.index.freq = pd.infer_freq(time_df.index)

# Decompose the time series using an additive model
decomposition = sm.tsa.seasonal_decompose(time_df['Last 7 days views'], model='additive', period=period)
```

In [12]:
```python
# Access the decomposed components (trend, seasonal, residual)
trend = decomposition.trend
seasonal = decomposition.seasonal
residual = decomposition.resid
```

In [13]:
```python
from pylab import rcParams

# This will define a fixed size for all special charts.
rcParams['figure.figsize'] = 18, 7
```

In [14]:
```python
# Plot the separate components

decomposition.plot()
plt.show()
```



The trend component appears as a horizontal line, it suggests that there is no clear upward or downward trend, and horizontal line for the seasonal component suggests that the seasonal variation is consistent across the entire time series, without significant fluctuations

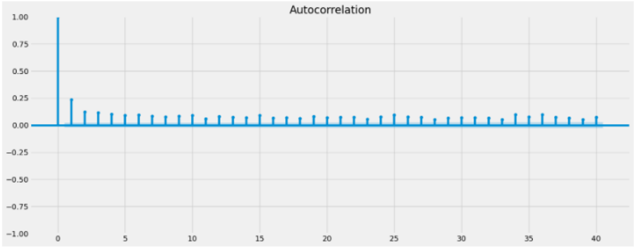
4. Testing for stationarity

In [15]:
```python
# Import the adfuller() function
from statsmodels.tsa.stattools import adfuller

# Define the function
def dickey_fuller(timeseries):
    # Perform the Dickey-Fuller test:
    print('Dickey-Fuller Stationarity test:')
    test = adfuller(timeseries, autolag='AIC')
    result = pd.Series(test[0:4], index=['Test Statistic', 'p-value', 'Number of Lags Used', 'Number of Observations Used'])
    for key, value in test[4].items():
        result['Critical Value (%s)' % key] = value
    print(result)

# Apply the test using the function on the time series
dickey_fuller(time_df['Last 7 days views'])
```

```
Dickey-Fuller Stationarity test:
Test Statistic                -8.778694e+00
p-value                        2.409420e-14
Number of Lags Used            3.500000e+01
Number of Observations Used    9.173000e+03
Critical Value (1%)           -3.431063e+00
Critical Value (5%)           -2.861855e+00
Critical Value (10%)          -2.566938e+00
dtype: float64
```

The low p-value and the test statistic being lower than the critical values suggest that I can reject the null hypothesis, indicating that the 'Last 7 days views' time series is likely stationary.

In [16]:
```python
# import the autocorrelation and partial correlation plots
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf

# Check out a plot of autocorrelations
plot_acf(time_df)
plt.show()
```



Do not provide meaningful insights, it might indicate that the data doesn't exhibit strong trends or seasonality.

The trend component appears as a horizontal line, it suggests that there is no clear upward or downward trend, and horizontal line for the seasonal component suggests that the seasonal variation is consistent across the entire time series, without significant fluctuations

Insights

Despite initial assumptions, it was found that smaller, cheaper ships tended to attract more views, while older ships with lower prices generally got fewer views.

There were some cases where increasing size, later years of manufacturing, or higher prices were correlated with increased views, but this was not consistent across all groups.

Explore additional variables that may influence views, such as marketing efforts, geographic location, or specific features highlighted in ads.

GitHub

Thank you!

Gmail

GitHub