Andy Francis

Data Analyst Portfolio

03/27/2023

About Me

Mission Statement

I aim to become Data Analyst in the United States and to assist companies in making data-driven decisions. I am looking to be part of a company that values my contributions while allowing me independence and ownership over my work.

My 10+ years of work as a **collegiate professor** combined with my **data analytics certification** will allow me to work well as a team member, consistently meet established deadlines, and to effectively communicate my work. I bring a **detail-oriented**, **people focused** and **data-driven mindset**, with **excellent management**, **research**, **collaboration**, and **communication skills**.

Contents

Chocolate Taste Analysis

Objective:

To perform an exploratory analysis of chocolate taste rating data in order to derive insights and suggest strategies for better tasting chocolate.

Data:

Open data was sourced through Kaggle here

Skills:

- Python coding
- Data wrangling and merging
- Grouping Data
- Populations flows
- Data Aggregation
- Data Visualization

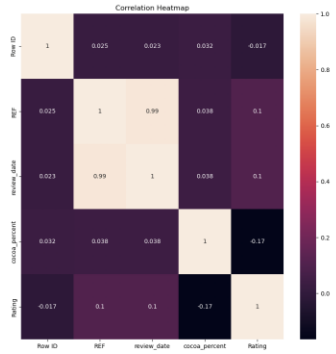
Tools:

View Python scripts and final report here

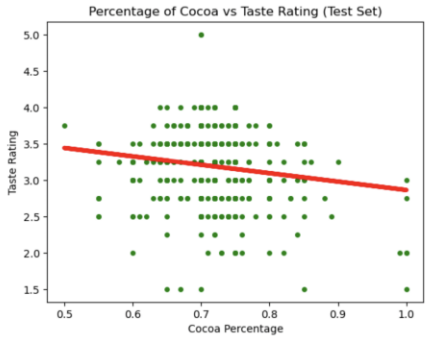
Initial Analysis

- Scanned data using python to check for irregularities and locate viable information for analysis
- Kept a detailed log of all cleaning, wrangling, and merging steps, noting the steps taken to transform the dataset
- Used Python to check for correlations and perform a regression analysis

```
# Create a subplot with matplotlib and a correlation heatmap in seaborn
f,ax = plt.subplots(figsize=(10,10))
corr = sns.heatmap(sub.corr(), annot = True, ax = ax).set(title='Correlation Heatmap')
```

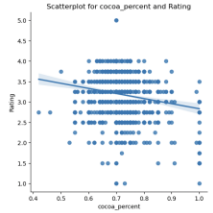



```
#Create plot that shows the regression line from the model
plot_test = plt
plot_test.scatter(X_test, Y_test, color='green', s = 15)
plot_test.plot(X_test, Y_predicted, color='red', linewidth =4)
plot_test.title('Percentage of Cocoa vs Taste Rating (Test Set)')
plot_test.xlabel('Cocoa Percentage')
plot_test.ylabel('Taste Rating')
plot_test.show()
```

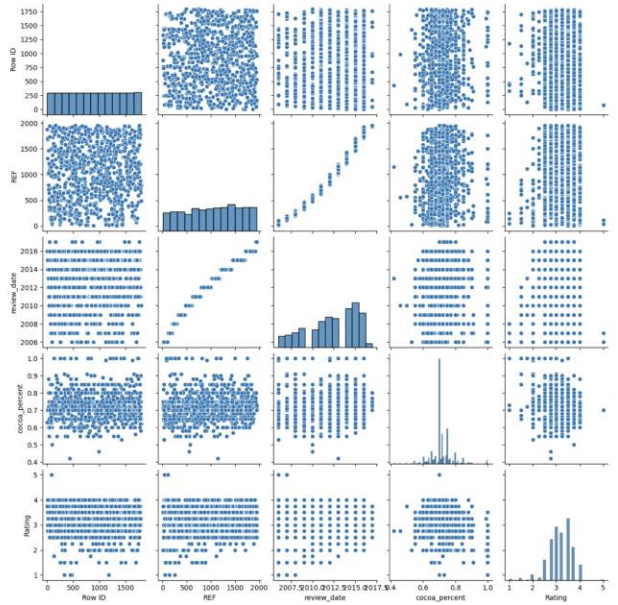



```
#Create scatterplot for 'cocoa_percent' and 'Rating' in seaborn
sns.lmplot(x = 'cocoa_percent', y = 'Rating', data = df).set(title='Scatterplot for cocoa_percent
```

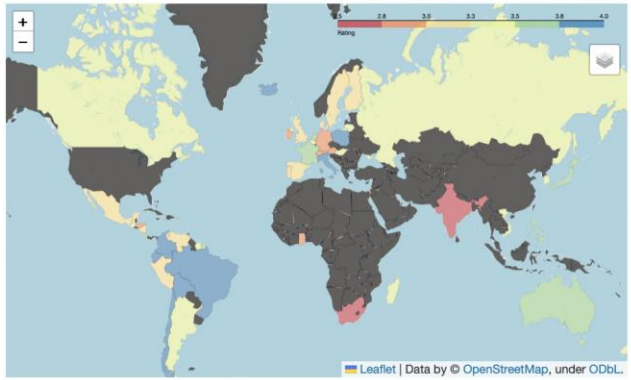
```
<seaborn.axisgrid.FacetGrid at 0x7fcdf27d6eb0>
```




```
#Create a pair plot using the subset created earlier
g = sns.pairplot(sub)
```



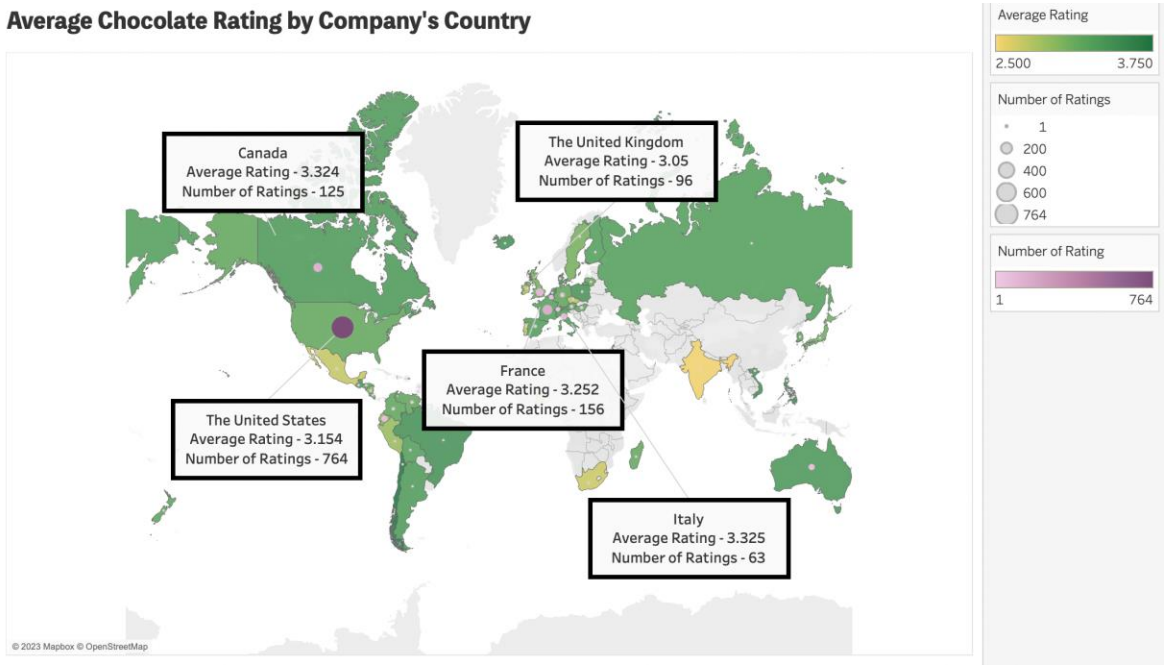

```
#Create choropleth map
map = folium.Map(location = [100, 0], zoom_start = 1.5)

folium.Choropleth(
    geo_data = world_geo,
    data = data_to_plot,
    columns = ['company_location', 'Rating'],
    key_on = 'feature.properties.name',
    fill_color = 'Spectral', fill_opacity=0.6, line_opacity=0.1,
    legend_name = "Rating").add_to(map)

folium.LayerControl().add_to(map)

map
```

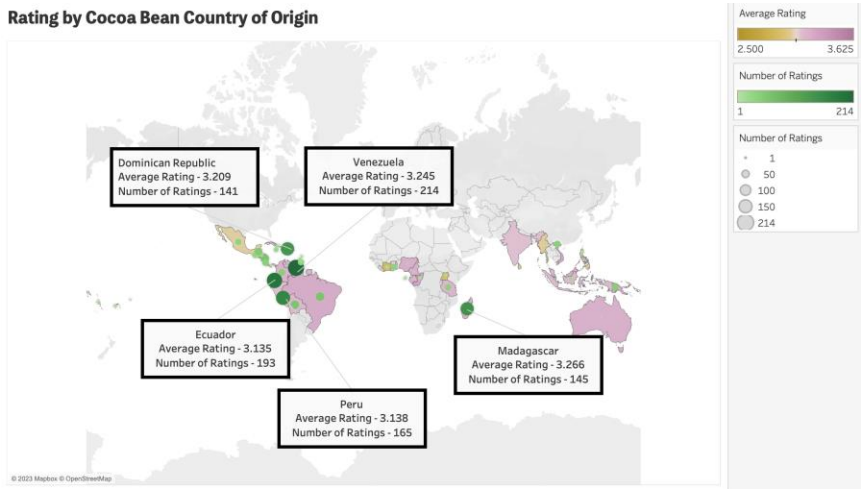


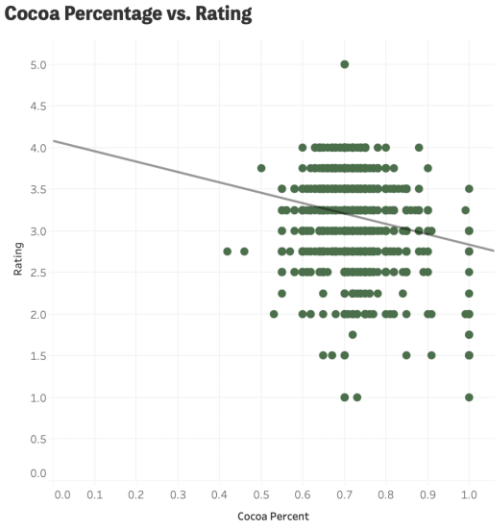
View Python scripts and final report here

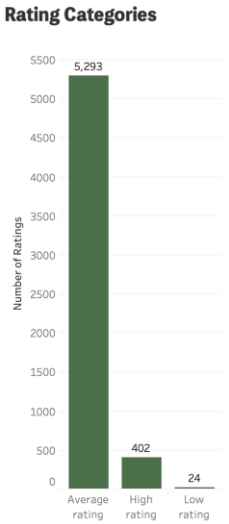
Consumer Profiling and Market Analysis

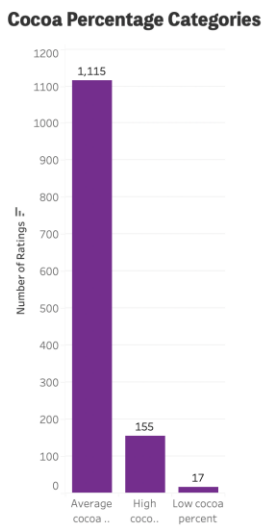
- Created a storyboard to showcase my results.
- All visualizations were created using Tableau Public.











View Python scripts and final report here

Conclusions

Observations

1) Canada and Italy are producing the best tasting chocolate on average

2) Cocoa beans from Madagascar and Venezuela are used in the best tasting chocolate on average

3) There is no correlation between cocoa percentage used in the chocolate and the taste rating

Recommendations

1) Purchase from those countries for the best average tasting chocolate

2) Purchase from companies who source beans from these countries

3) There is no need to purchase chocolate purely based on cocoa percentage if taste is the primary goal

View Python scripts and final report here

An online grocery store that operates through an app

Objective:

To perform an exploratory analysis of Instacart's sales data in order to derive insights and suggest strategies for better segmentation.

Data:

Customer Data Set
Data Dictionary

Skills:

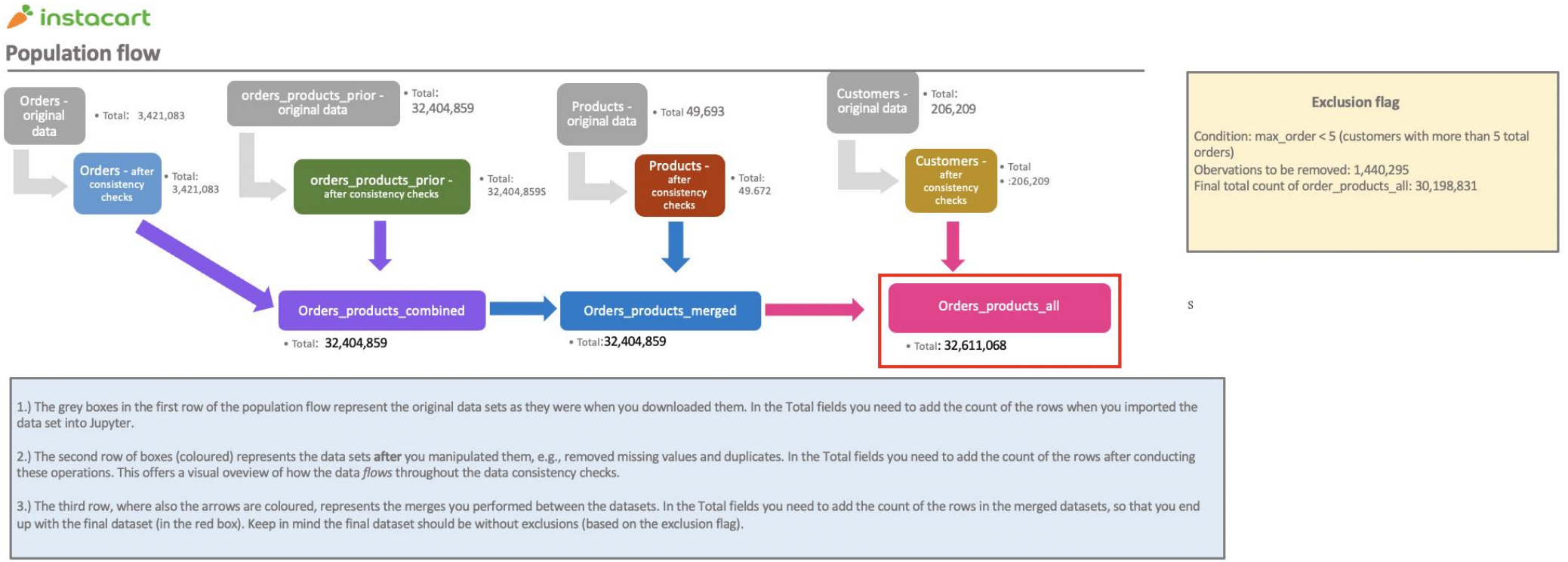
- Python coding
- Data wrangling and merging
- Grouping Data
- Populations flows
- Data Aggregation
- Data Visualization

Tools:

View Python scripts and final report here

Initial Analysis

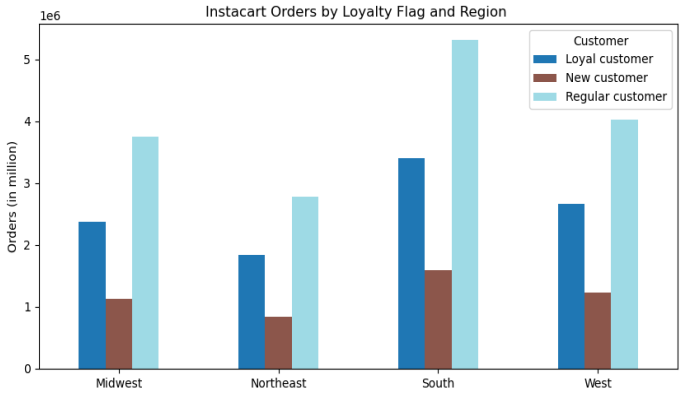
- Scanned data using python to check for irregularities and locate viable information for analysis
- Kept a detailed log of all cleaning, wrangling, and merging steps, noting the steps taken to transform the dataset
- Merged dataframes are utilized to address the stakeholder's specific questions.

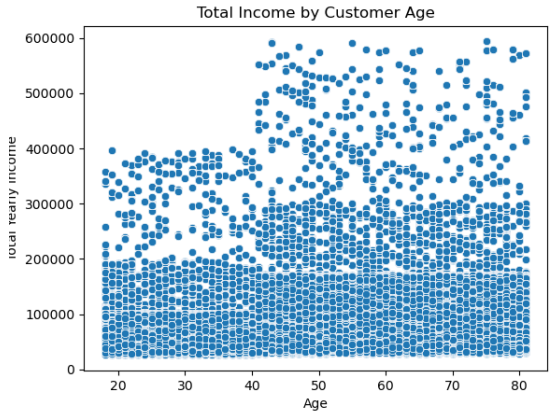


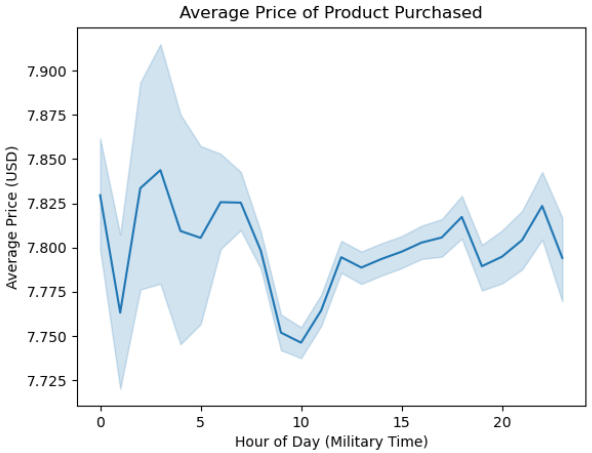
View Python scripts and final report here

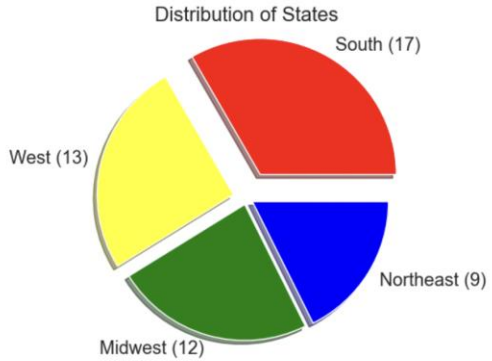
Consumer Profiling and Market Analysis

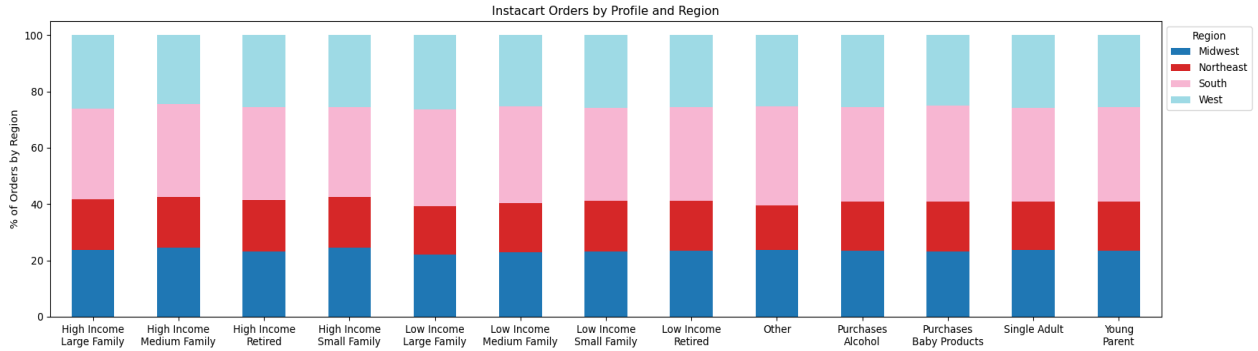
- Created crosstabs from merged dataframes aid in the understanding of the correlation between variables.
- All visualizations were created using matplotlib











View Python scripts and final report here

Conclusions

Questions

1) The sales team needs to know what the busiest days of the week and hours of the day are (i.e., the days and times with the most orders) in order to schedule ads at times when there are fewer orders.
2) They also want to know whether there are particular times of the day when people spend the most money, as this might inform the type of products they advertise at these times.
3) Instacart has a lot of products with different price tags. Marketing and sales want to use simpler price range groupings to help direct their efforts.
4) What's the distribution among users in regards to their brand loyalty (i.e., how often do they return to Instacart)?
5) Are there differences in ordering habits based on a customer's loyalty status?
6) Are there differences in ordering habits based on a customer's region?
7) What differences can you find in ordering habits of different customer profiles?

Recommendations

1) 1) The busiest days for orders are Saturday and Sunday, The slowest days for orders are Tuesday and Wednesday. The busiest hours for orders are between 09:00 and 16:00. The slowest hours orders occur between 01:00 and 05:00.
2) The average price is consistent regardless of the time of day is is purchased (between $7.75 and $7.85 USD).
3) Price ranges have been groups into High Range Products (> $15.00), Mid Range Products (<= $15.00 and > $5.00) and Low Range Products (> $5.00). Mid Range Products are the best selling across regions.
4) Regular Customers (Max order <= 40 & > 10) have the highest numbers across regions, followed by Loyal Customers (Max order > 40) and New Customers (Max order <= 10)
5) There is no difference in ordering habits between customer's loyalty status.
6) There is no difference in ordering habits between customer's region.
7) All customer profiles order produce and dairy/eggs the most. Differences in ordering habits between the different customer profiles are not apparent in my analysis.

View Python scripts and final report here

Rockbuster Stealth, LLC.

A fictional online video rental company

Objective:

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 Rockbuster Stealth management team is planning to use its existing movie licenses to launch an online video rental service in order to stay competitive.

Data:

Mock data set provided by CareerFoundry

Skills:

- Data Dictionary and Entry Relational Diagram (ERD)
- Relational Databases and SQL
- Database querying and filtering
- Common table expressions
- Subqueries
- Data cleaning and visualization

Tools:

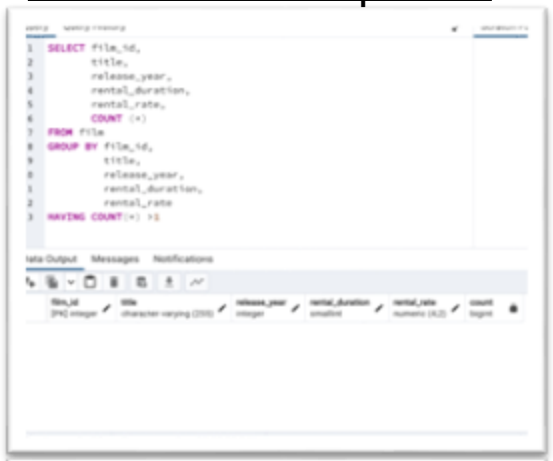
View SQL queries and final presentation here

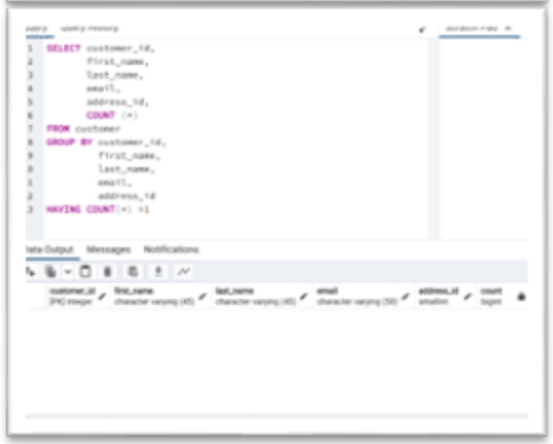
Initial Analysis

- Created an Entry Relational Diagram (ERD) and data dictionary using DbVisualizer
- Performed CRUD functions to ensure data does not contain duplicates and/or missing values.

ERD

CRUD function queries



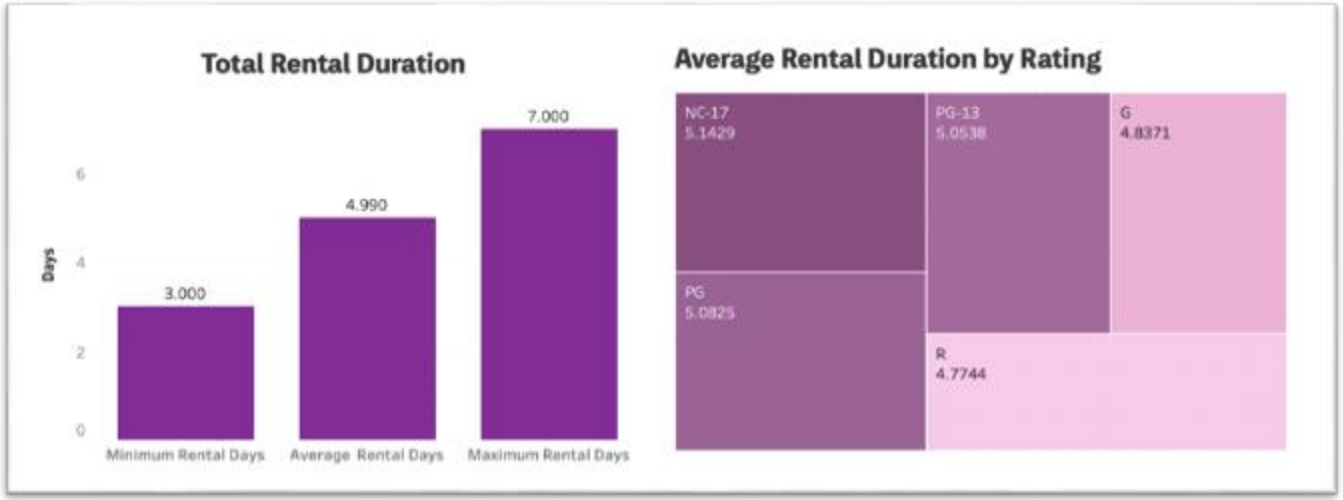


View SQL queries and final presentation here

Consumer Behavior Analysis

- Subqueries, common table expressions, and joining tables were used to extract relevant records.
- Query results were saved as .csv files to use to create visualizations in Tableau.

```
Query  Query History
1  WITH average_total_amount_paid_cte AS
2  (SELECT A.customer_id, A.first_name, A.last_name, D.city, E.country,
3  SUM(B.amount)AS total_amount_paid
4  FROM Customer A
5  INNER JOIN payment B ON A.customer_id = B.customer_id
6  INNER JOIN address C ON A.address_id = C.address_id
7  INNER JOIN city D ON C. city_id = D. city_id
8  INNER JOIN country E on D. country_id = E. country_id
9  WHERE D.city IN ('Aurora','Bhusawal','Shivaouri','Cianjur','Kuwana',
10                 'Acua','Saint Louis','So Leopoldo','Iwaki','Eskisehir')
11 GROUP BY A.customer_id, A.first_name, A.last_name, D.city, E.country
12 ORDER BY total_amount_paid DESC
13 LIMIT 5)
14 SELECT AVG(total_amount_paid) AS average_paid
15 FROM average_total_amount_paid_cte

Data Output  Messages  Notifications
average_paid
numeric
104.162
```



View SQL queries and final presentation here

Conclusions

Observations:

1) Sports, Animation, Action, and Family are Rockbuster's highest-grossing genres.

2) Films rated NC-17, PG, and PG-13 have the highest average rental duration.

3) India, China, and the United States have the highest revenue

4) More customers = higher revenue

Recommendations

1) Acquire more films in these genres in Rockbuster's inventory.

2) Acquire more films with these rating in Rockbuster's inventory. Higher average rental duration = more revenue

3) Invest more resources in these countries.

4) Create rewards programs to keep customers active.

View SQL queries and final presentation here

Preparing for Influenza Season

Predicting National Medical Staffing Needs

Objective:

The analysis will help plan for influenza season, a time when additional staff are in high demand. The final results will examine trends in influenza and how they can be used to proactively plan for staffing needs across the country.

Data:

CDC Influenza deaths by geography, time, age, and gender
United States Census population data by geography

Skills:

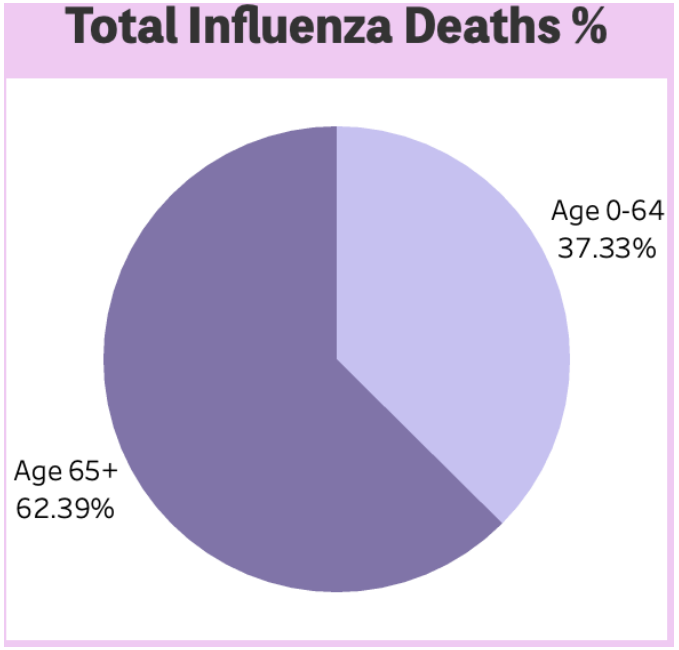
- Data cleaning, profiling and integrity
- Data transformation and Integration
- Statistical analysis and hypothesis testing using both Excel and Tableau functions
- Data visualizations and Storyboard creation using Tableau

Tools:

View final presentation video here

Initial Analysis

- **Hypothesis**: If a state has a higher population age 65+, then the mortality rate will increase.
- Performed statistical analyses in Excel for vulnerable populations (age 65+)
- Conducted T-tests on the research hypothesis to confirm the connection between vulnerable populations and total influenza deaths



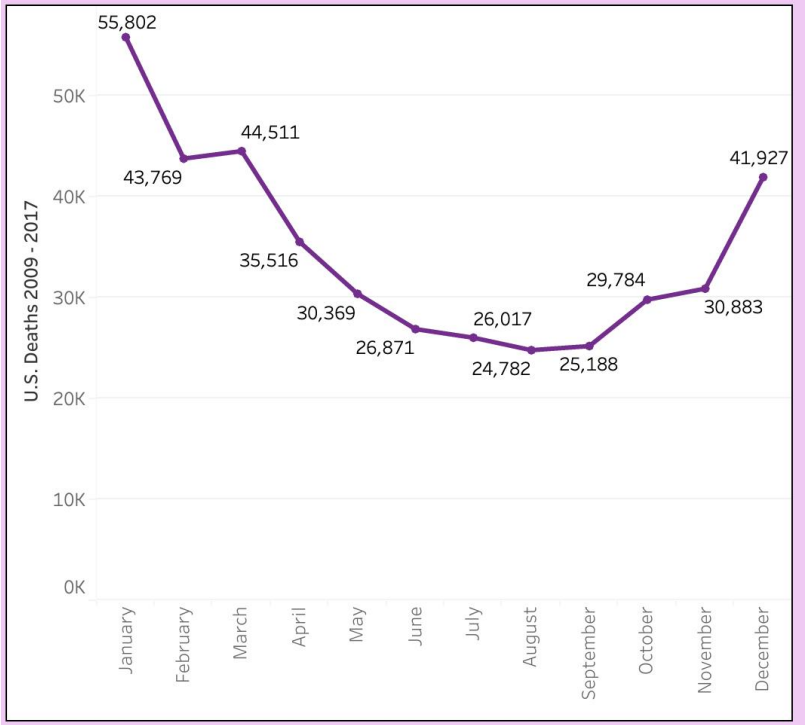
	Population Age 65+	Total Deaths Age 65+
Mean	514,404	897
Standard Deviation	561,278	791
Correlation Coefficient	0.81	

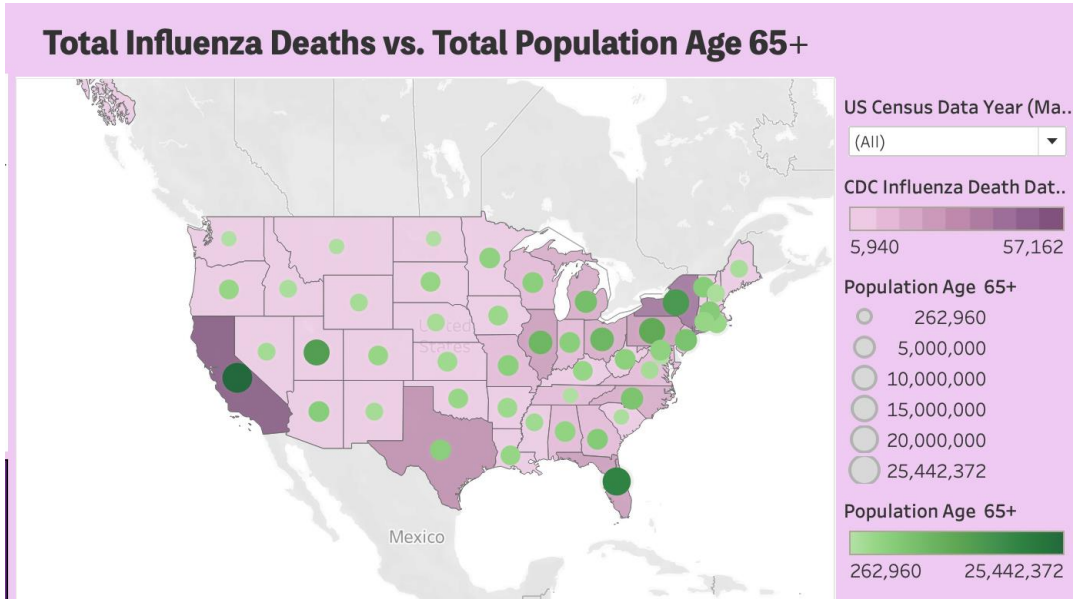
	Total Deaths (High population % of age 65+)	Total Deaths (Low population % of age 65+)
Mean	0.05%	0.04%
Variance	3.15077E-07	8.43182E-08
Observations	258	210
Hypothesized Mean Difference	0	
df	401	
t Stat	4.200763665	
P(T<=t) one-tail	1.64045E-05	
t Critical one-tail	1.648662397	
P(T<=t) two-tail	3.2809E-05	
t Critical two-tail	1.965897465	

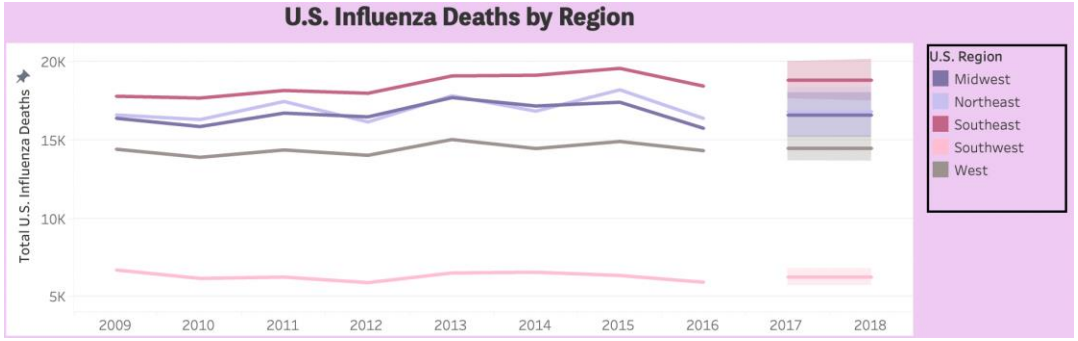
View final presentation video here

Time Series Analysis

- The line chart shows that most influenza deaths occur during the winter months
- Density mapping shows that California, New York, and Texas have the most influenza deaths
- Forecasting show that influenza deaths will follow the same trends for 2017 - 2018







View final presentation video here

Conclusions

Observations

1) 34% of total U.S. influenza deaths occur during the winter months. (December, January, February)

2) California, Texas, and New York have the highest number of influenza deaths annually.

3) 62% of all influenza deaths were elderly citizens (age 65+).

4) Not all states were able to report accurate influenza data.

Recommendations

1) Additional medical staffing will be required nationwide during these months.

2) These states will require the most additional staffing.

3) States with high elderly populations should be monitored due to the increased influenza risk for that population.

4) States should be consulted to ensure that they are able to accurately report their influenza data.

View final presentation video here

GameCo

A fictional company that distributes video games worldwide

Objective:

Perform a 10-year descriptive analysis to gain insights into the current video game sales trends for the marketing and sales teams' 2017 budget planning.

Data:

[GameCo mock data 1980 – 2016]

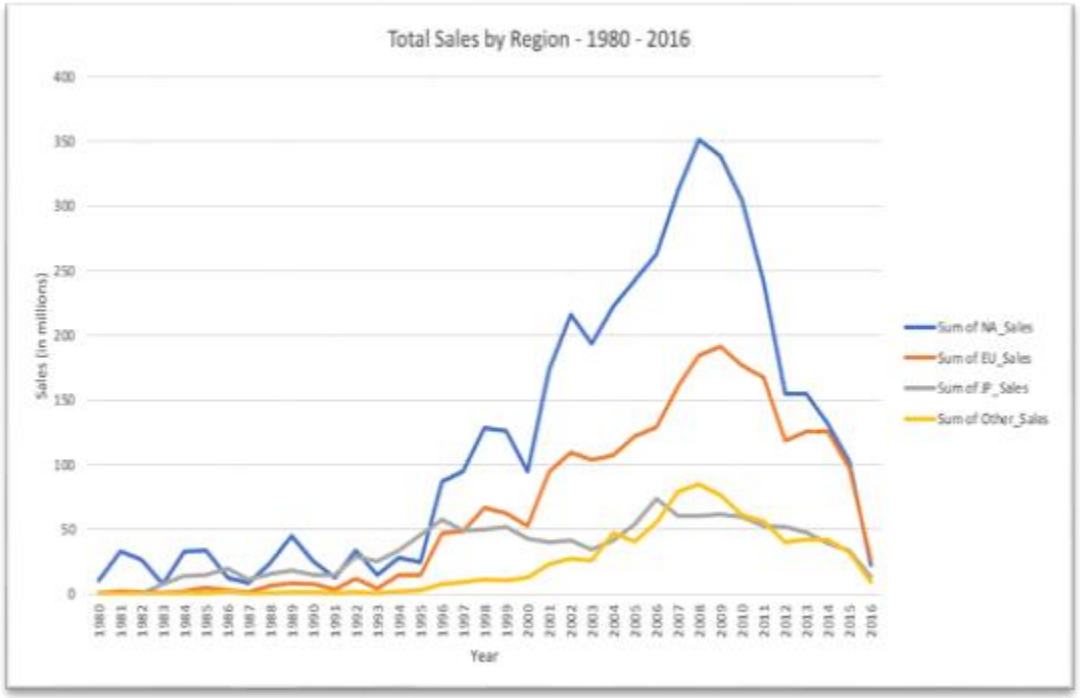
Skills:

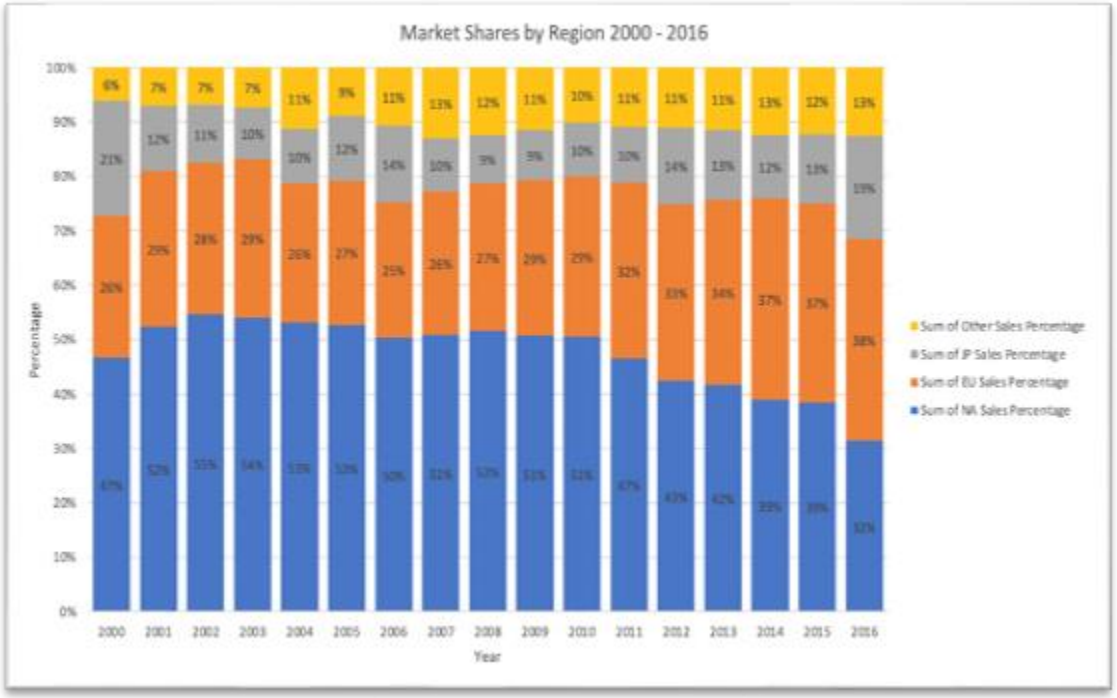
- Data cleaning, profiling and integrity
- Data transformation and Integration
- Data grouping and summarization
- Calculated fields using pivot tables
- Data visualizations in Excel

Tools:

Initial Analysis

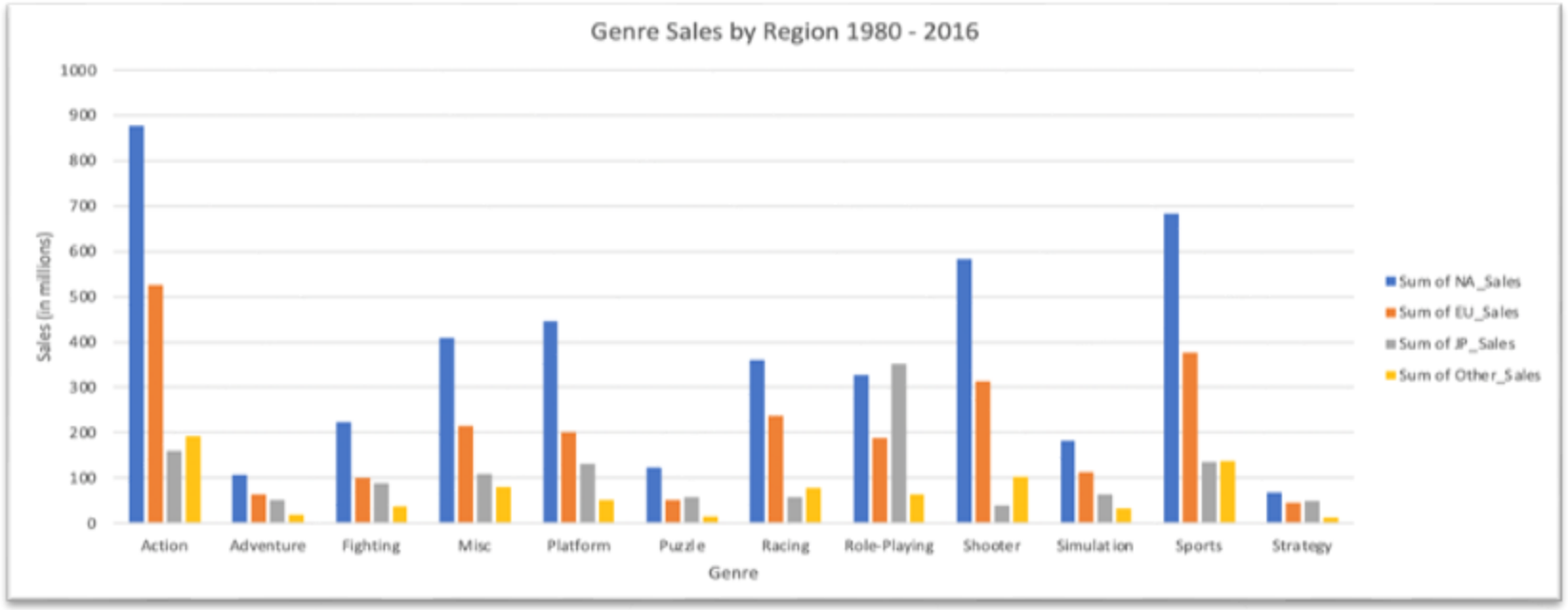
- Market shares were determined to show the regional trends of GameCo's total sales.
- Analysis shows a clear decline in sales, contradicting the company's current business understanding.

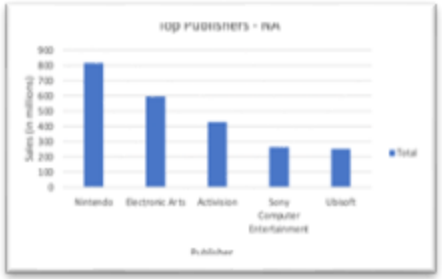


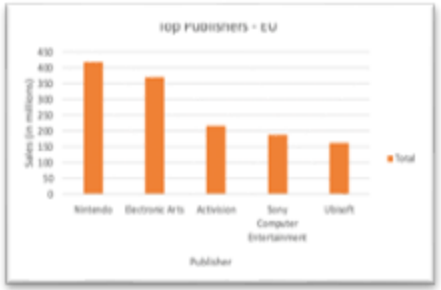


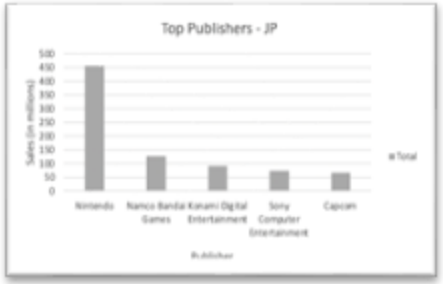
Market Analysis

- Action, Sports, and Shooter games dominate the markets of North America and Europe, as well as globally.
- Nintendo is the strongest publisher globally, dominating every region except Other.









Conclusions

Observations:

1) Regional sales have not remained consistent over time, as the market shares of total sales have changed.

2) The market share of North America has decreased by 19% since 2009 (matching the global sales trends). Sales in the EU hold 38% of the market share in 2016.

3) Japan's most popular publishers are different than the other regions. Nintendo, EA, Activision, and Sony were in the top 4 publishers of all other regions.

4) Action games have outperformed all other genres globally. The only region where action games were outperformed was Japan.

Recommendations

1) GameCo should not continue to make marketing decisions based on the assumptions that sales trends remain consistent.

2) The EU should have priority in marketing funds since it holds the greatest market share.

3) The marketing strategy in Japan should be different than that of other regions, since its top genres and publishers differ from the others.

4) GameCo should consider pushing more action titles in North America, the European Union and Other. GameCo should continue to market role playing games in Japan.

Contact Information

Andy Francis

Email: andyfrancismusic@gmail.com

Cell: (989) 289 - 7827