Jeffrey Frankenfeld

Data Analytics Portfolio

Projects & Tools

- GameCo
- Influenza
- Rockbuster
- Instacart
- Pig E. Bank
- Crop Yield Analysis

GameCo Project Overview

Objective

As a data analyst for GameCo, I was tasked with performing a descriptive analysis of a video game dataset to provide insights that will inform the development of new games and help understand how they may perform in the market. This scenario takes place in 2016.

Key Questions

1. Are certain types of games more popular than others?
2. What other publishers will likely be the main competitors in certain markets?
3. Have any games decreased or increased in popularity over time?
4. How have their sales figures varied between geographic regions over time?

Data

- Data Source: VGChartz (link)
- File: excel .CSV
- From 1980 to 2016

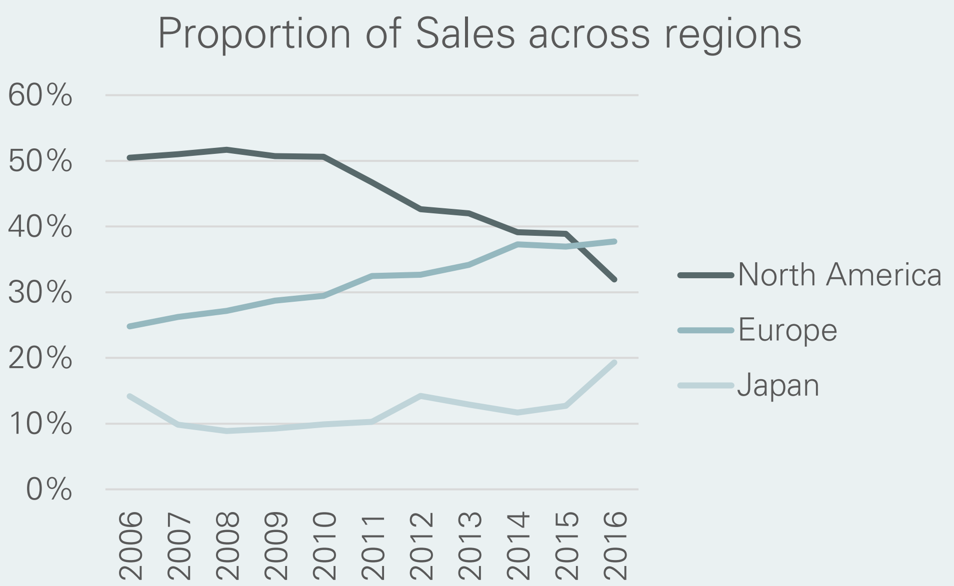
Skills

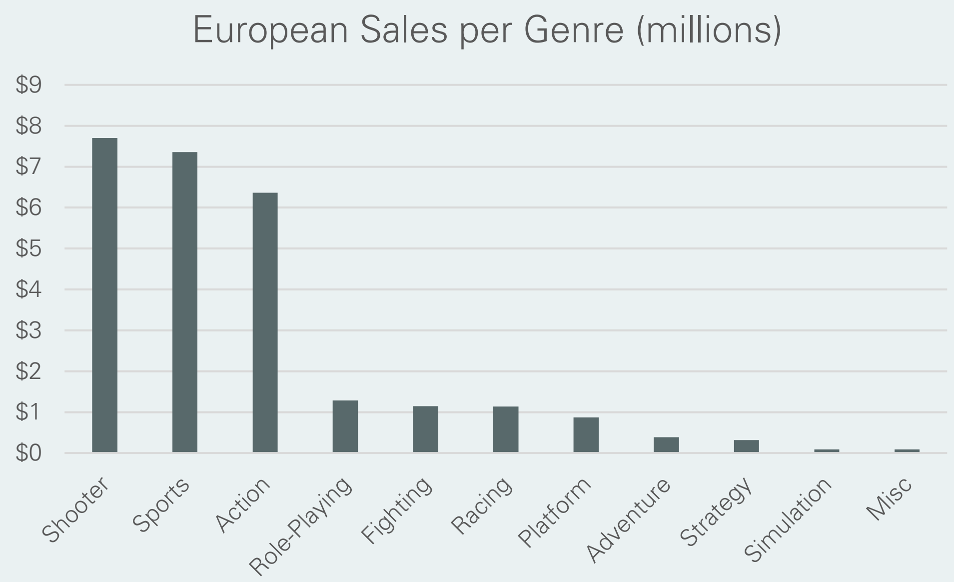
- Data cleaning
- Pivot tables
- Descriptive analysis
- Insight development
- Excel visualizations

Challenges

- Converting business requirements into actionable insights
- Creating a story with the data available

Analysis





This analysis highlights the geographic sales across regions over the past decade and showing the top performing genres from the Europe market in 2016.

The graphs above showcase the ability to pull insights from data and dive into a specific regions market based on the initial exploratory data analysis.

Summary

Insights

Less games available

- In the past decade, the amount of games released per year have decreased by 66%

European Gamers have taken over

- For the first time since collecting this data, the European video game sales have surpassed North American and Japan sales owning 38% of the share.

Shooter and Sport games dominance

- ‘Shooter’ and ‘Sport’ games provide the largest return on investment as they contribute on average 200K in sales per game.

Recommendations

Focus on game development to increase selection and availability

Pour money in the European gamer market to capitalize on their growing market share

Promote ‘Sport’ and ‘Shooter’ games, more specifically Call of Duty, Battlefield Series, FIFA Series, and WII Sports Series.

Influenza Season Project Overview

Objective

In this project, I analyzed trends in influenza to help a medical staffing agency proactively plan for staffing needs during flu season across all 50 U.S. states.

The goal was to determine when and how many temporary staff should be allocated to hospitals and clinics to meet the increased demand for healthcare workers.

Requirements

- Provide data to support a staffing plan, informing the timing and distribution of medical personnel across the U.S.
- Investigate seasonality of influenza across states.
- Prioritize states with large vulnerable populations.
- Identify data limitations that hinder analysis.

Data

- Influenza deaths by geography
    - Source: CDC
- Population data by geography, time, age, and gender
    - Source: US Census Bureau

Skills

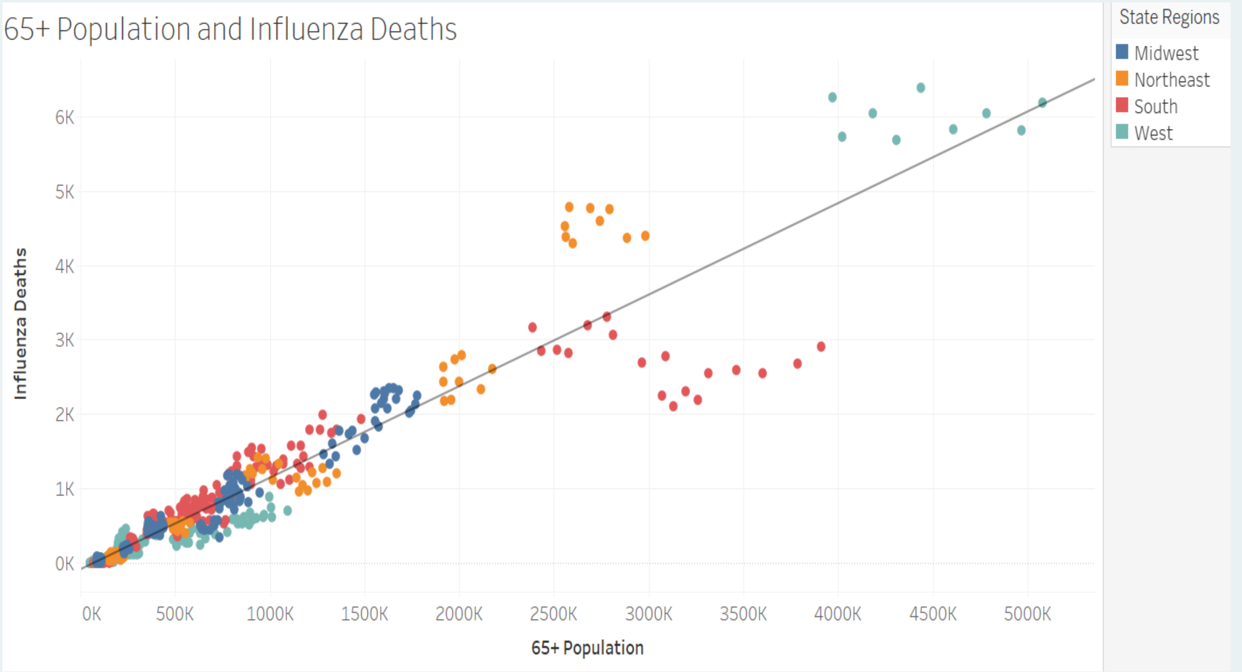
- Data cleaning, integration, and transformation
- Statistical hypothesis testing
- Visual creation using Tableau
- Storytelling

Challenges

- Distinguishing between relevant and irrelevant data.
- Identifying variables with a potential relationship and testing for correlation.

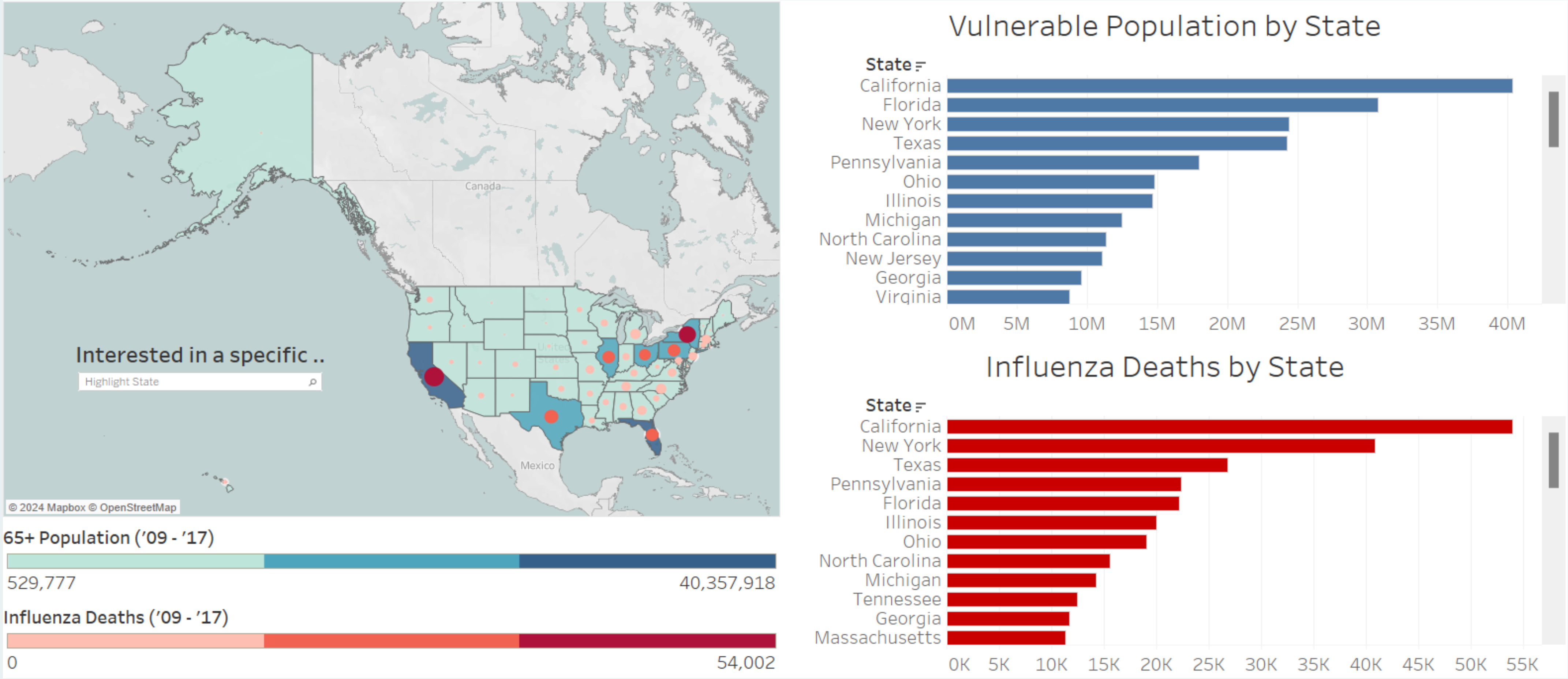
Statistical Testing

- Null Hypothesis: The number of influenza deaths is the same for individuals who are 65+ and individuals who are less than 65 years old.
    - H0:μ 65+ pop=μ 0-64 pop
- Alternative Hypothesis: The number of influenza deaths is higher for individuals who are 65+ compared to individuals less than 65 years old.
    - Ha:μ 65+ pop > μ 0-64 pop
- Standard-level significance: Alpha = 0.05
- **P-Value: 6.96013E-45**



By performing p-value testing, I can reject the null hypothesis and show there is enough evidence that more influenza deaths occur in the 65+ population.

State Analysis



The spatial analysis features regional information on the vulnerable population (65+) and influenza deaths by state. This showcases my ability to create complex visualizations with customizable filters.

Summary

Insights

Vulnerable Population

- Through statistical testing, an individual is 65 and older has greater chance of mortality from influenza.

High Priority States

- The same seven states had the largest vulnerable populations and highest influenza death counts.

Peak Influenza Season

- Between December and February is peak influenza season.

Recommendations

States with a high vulnerable population should increase their medical staffing during the influenza season.

Consider these states the 'High Priority' states and increase staffing to 2.5 physicians per 1,000 people per the World Health Organization.

Deploy medical staff during the Fall season to acclimate to facility and have tenure to beginning of March.

Rockbuster Project Overview

Objective

I worked as a data analyst for Rockbuster Stealth, a movie rental company transitioning to an online video rental service to compete with streaming platforms.

My role involved loading the company's data into a relational database, using SQL to perform analyses, and providing insights to support the launch strategy, while presenting finding for management.

Key Questions

1. Which movies contributed the most/least to revenue gain?
2. What was the average rental duration for all videos?
3. Which countries are Rockbuster customers based in?
4. Do sales figures vary between geographic regions?

Data

- Data set containing film inventory, customers, payments and other information
  - Size: ~3MB

Skills

- Relational databases
- SQL
- Data dictionary creation
- Database querying
- Data filtering
- Data cleaning & summarizing
- Joining, subqueries, CTEs

Challenges

- Creating complex queries

Descriptive Statistics

Total Films: 1,000

584 Active Customers in 108 Countries

Film Language: English

Average Rental Duration: 5 days

Total Revenue: $61,312

Average Movie Rating: PG-13

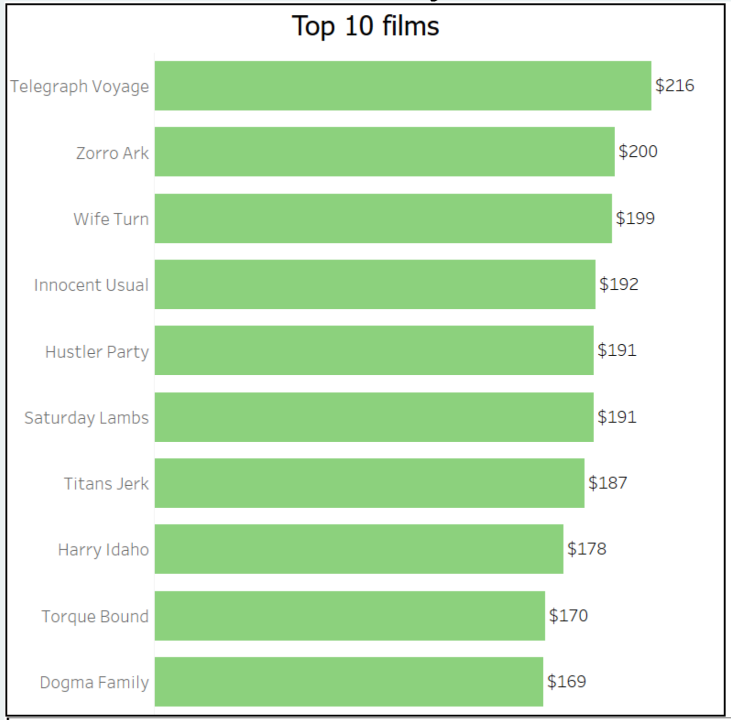
Average Replacement Cost: $19.98

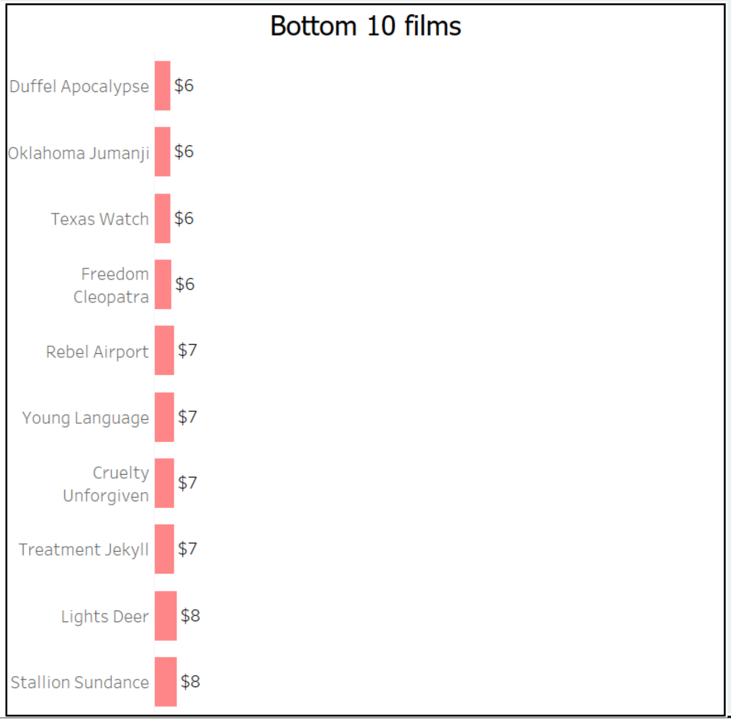
Average Rental Rate: $2.98

```sql
SELECT
    COUNT(film_id) AS total_films,
    MIN(release_year) AS min_release_year,
    MAX(release_year) AS max_release_year,
    ROUND(AVG(release_year)) AS avg_release_year,
    MIN(rental_duration) AS min_rental_duration,
    MAX(rental_duration) AS max_rental_duration,
    ROUND(AVG(rental_duration)) AS avg_rental_duration,
    MIN(rental_rate) AS min_rental_rate,
    MAX(rental_rate) AS max_rental_rate,
    ROUND(AVG(rental_rate),2) AS avg_rental_rate,
    MIN(length) AS min_length,
    MAX(length) AS max_length,
    ROUND(AVG(length)) AS avg_length,
    MIN(replacement_cost) AS min_replacement_cost,
    MAX(replacement_cost) AS max_replacement_cost,
    ROUND(AVG(replacement_cost),2) AS avg_replacement_cost,
    MODE() WITHIN GROUP (ORDER BY B.name) AS mode_language,
    MODE() WITHIN GROUP (ORDER BY A.rating) AS mode_rating,
    MODE() WITHIN GROUP (ORDER BY A.special_features) AS mode_special_features
FROM film A
JOIN language B ON A.language_id = B.language_id;
```

The query above showcases the SQL skills developed through this project utilizing descriptive aggregates and joining capability.

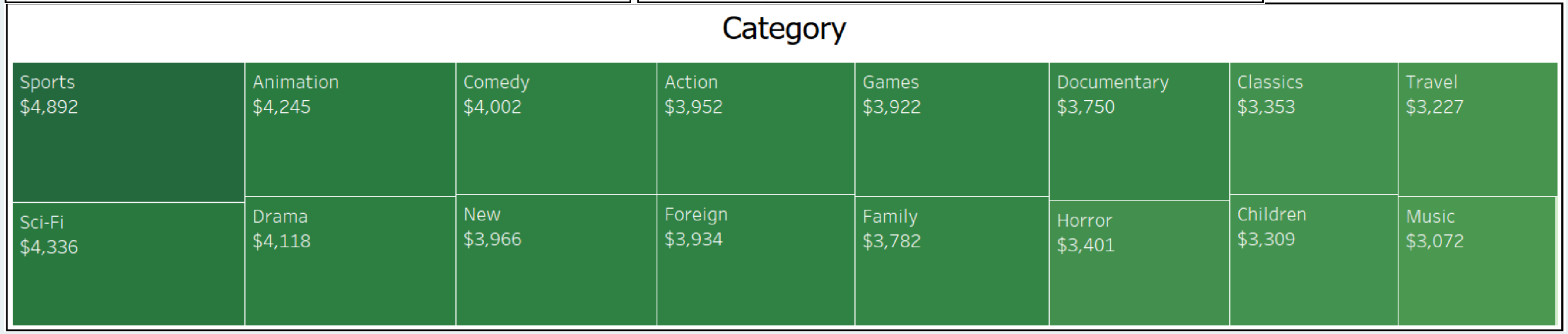
Film Analysis





The visuals show the top and bottom performing films and how each category performs with the use of a treemap.

This highlights my ability to answer the key questions in visuals that tell a story, not just numbers.



Summary

Recommendations

Increase Content Diversity

- Expand movie collection to have multiple languages and release dates. Rockbuster is currently isolating themselves by only have movies in English.

Membership Loyalty & Referral Program

- Acknowledge the high performing customers with early release and discounted rates. Add a referral program / code to incentivize word-of-mouth promotion.

Focus on top performing countries

- Concentrate marketing efforts in the top performing countries; such as India, China, and the United States.

Promote Popular Categories

- Concentrate marketing efforts on Sports, Sci-Fi, Animation, and Drama categories as these currently bring in the highest revenue.

Instacart Project Overview

Objective

As a data analyst for Instacart, I conducted an exploratory analysis of sales data with the use of Python to uncover patterns and provide insights for better customer segmentation.

My analysis informed a targeted marketing strategy aimed at reaching the right customer profiles with relevant products to improve sales performance.

Key Questions

1. What are the busiest days of the week and hours of the day?
2. Are there specific times of day when spending is highest, to inform product advertising?
3. How can simple price range groupings be used to optimize marketing and sales efforts?
4. Are certain types of products more popular than others?
5. How do different customer types and their ordering behaviors vary?

Dataframes

- Customers: used for purchasing patterns
- Orders: used for spending times and habits
- Products: used for product popularity and sales analysis
- Departments: used as a data dictionary for other dataframes

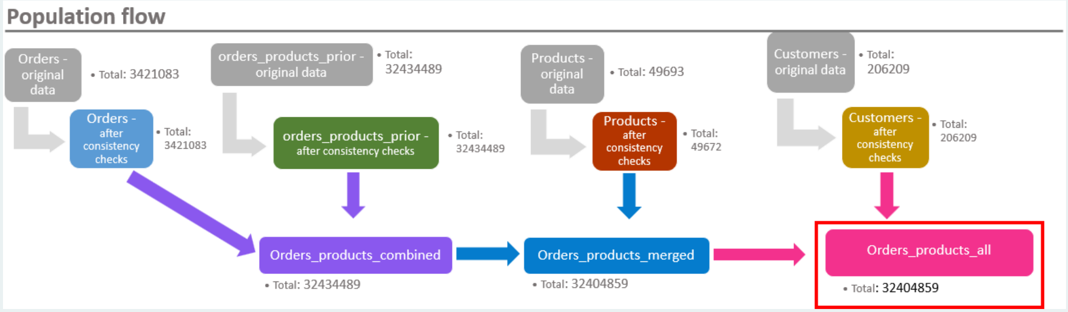
Skills

- Python
- Data wrangling & cleaning
- Data merging
- Deriving variables
- Grouping data
- Aggregating data
- Population Flows

Challenges

- Handling dataframes with more than 30 million records while being efficient with computer storage.
- Learning visualization creation using seaborn and scipy.

Python Coding




```python
# creating a loop statement
result = []

for value in ords_prods_all["department_id"]:
  if value == 18:
    result.append("baby_purchase")
  elif value == 8:
    result.append("pet_purchase")
  else:
    result.append("neither")
```

```python
# creating "baby_or_pet" column based on results above
ords_prods_all['baby_or_pet'] = result
```

```python
# creating age_group" column
ords_prods_all.loc[ords_prods_all['age'] <= 29, 'age_group'] = 'young_adult'
```

```python
ords_prods_all.loc[(ords_prods_all['age'] > 29) & (ords_prods_all['age'] <= 49), 'age_group'] = 'adult'
```

```python
ords_prods_all.loc[(ords_prods_all['age'] > 49) & (ords_prods_all['age'] <= 64), 'age_group'] = 'older_adult'
```

```python
ords_prods_all.loc[ords_prods_all['age'] > 64, 'age_group'] = 'senior'
```

If-statements

```python
# finding missing values
df_prods.isnull().sum()
```

```
product_id        0
product_name     16
aisle_id          0
department_id     0
prices            0
dtype: int64
```

```python
# creating subset of missing values
df_nan = df_prods[df_prods['product_name'].isnull() == True]
```

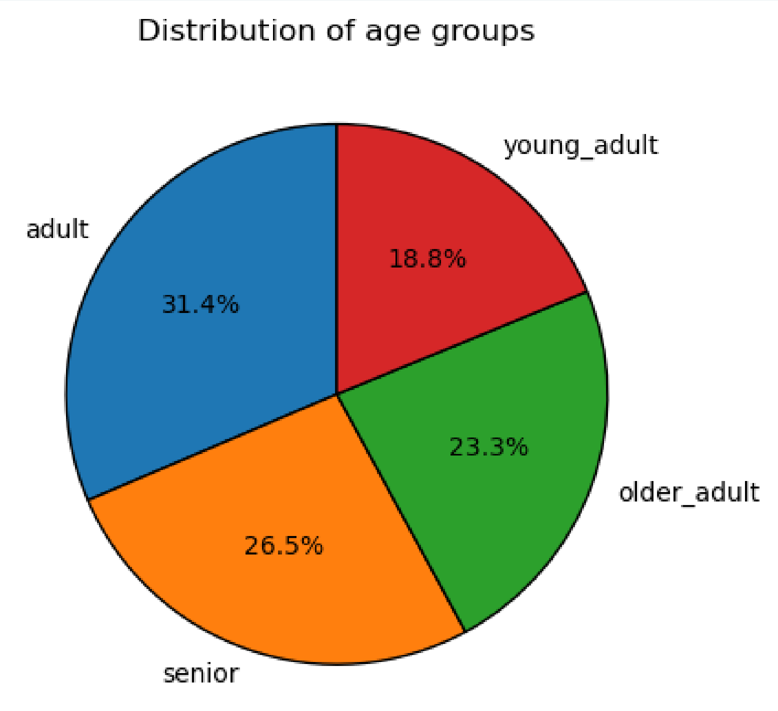
Data Cleaning

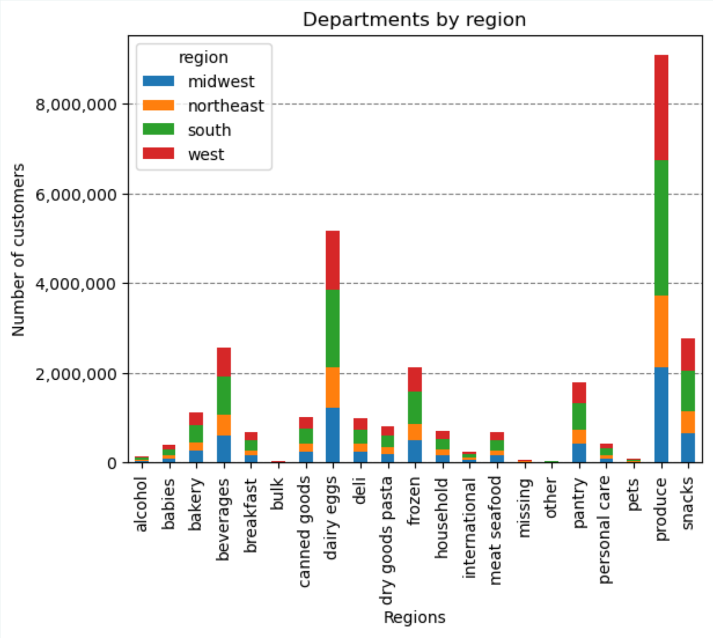
```python
for col in df_ords.columns.tolist():
  weird = (df_ords[[col]].map(type) != df_ords[[col]].iloc[0].apply(type)).any(axis = 1)
  if len (df_ords[weird]) > 0:
    print (col)
else : print ("no mixed-type data")
```

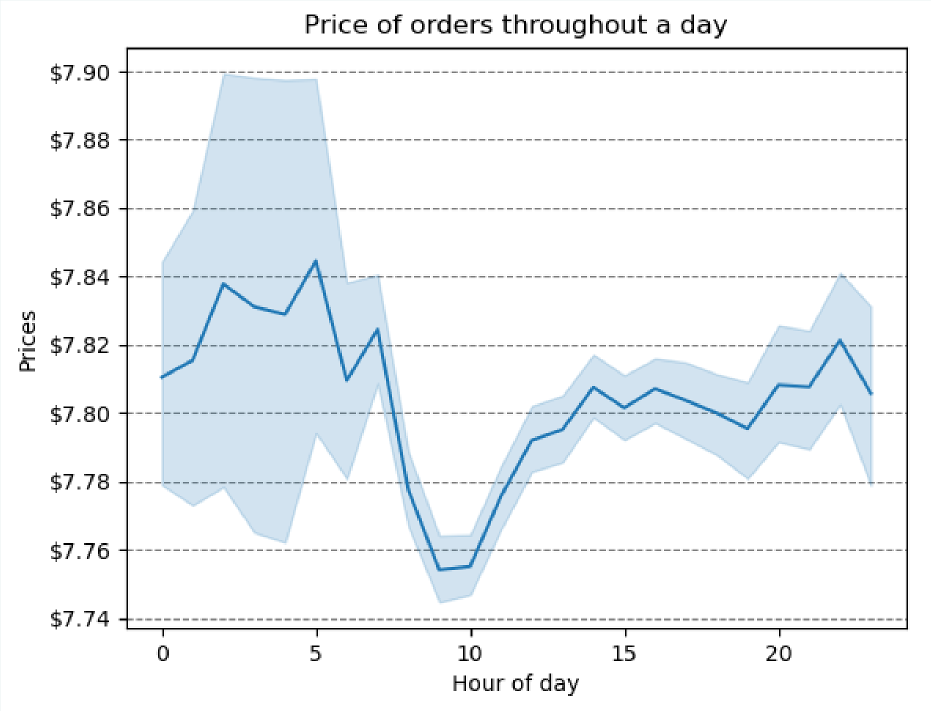
```python
line = sns.lineplot(data = line_chart_df, x = 'order_hour_of_day', y = 'prices')
plt.title('Price of orders throughout a day')
plt.ylabel('Prices')
plt.xlabel('Hour of day')
plt.gca().set_axisbelow (True)
plt.grid(axis = 'y', color = 'grey', linestyle = 'dashed')
plt.gca().yaxis.set_major_formatter(format('${x:,.2f}'))
```

Visualization creation

Analysis







Python and Jupyter notebooks were used to analyze sales patterns and develop customer profiles. This provided valuable insights to understand customer behavior, product performance, and improve segmentation strategies.

Summary

Insights

Busiest days and hours

- The busiest days of the week are Saturday, followed by Sunday.
- The busiest time of day is between 10am – 3pm

Product Price Ranges

- Mid-range products: majority of products in Instacart catalog (pantry, personal care, beverages).
- Low-range products: 2nd highest products in Instacart catalog (snacks, produce).
- High-range products: lowest number of products in Instacart (meat & seafood).

Customer Profiles

- Young adults count for the least amount of customers.
- Low-income counts for least amount of customers.
- 'Married' user is the base customer.

Recommendations

Implement marketing campaigns during slower weekday hours to boost orders and revenue, with messaging focused on quicker delivery times.

Mid-range products: Offer limited-time discounts on mid-range products, especially on items that customers buy regularly

Low-range products:

- implement 'add-on' suggestions when customers are about to check out.

- Create a 'buy more, save more' campaign.

High-range products: Create 'loyalty program incentives' for purchasing high-range products that are redeemable on future purchases.

Young adults: promote trendy products such as organics or plant-based foods

Low income: Create 'value pack campaigns to offer significant savings or develop flexible payment options to make larger purchases more manageable.

Married user:

- Support the 'base' by developing family meal kits and bundles

- Create 'baby essentials' campaign with subscription services for diapers, wipes and formula.

Pig E. Bank Project Overview

Objective

As a data analyst for a global bank, I provided analytical support to prevent money laundering by assessing client information and transaction risk.

My work involved addressing data ethics and applying techniques like predictive analysis, data mining, and time-series forecasting.

I then developed a Github profile to showcase my past projects to potential recruiters and employers.

Key Question

1. Identify the top risk factors that contribute to client loss and model them in a decision tree.

Data

- Fictional client data set from CareerFoundry
  - Contains 991 records of client bank data

Skills

- Big data
- Data ethics
- Data mining
- Predictive analysis
- Time-series Analysis
- Github

Challenges

- Identifying data ethics and how to mitigate / prevent it.

Customer Attrition Rate

Contributing Factors

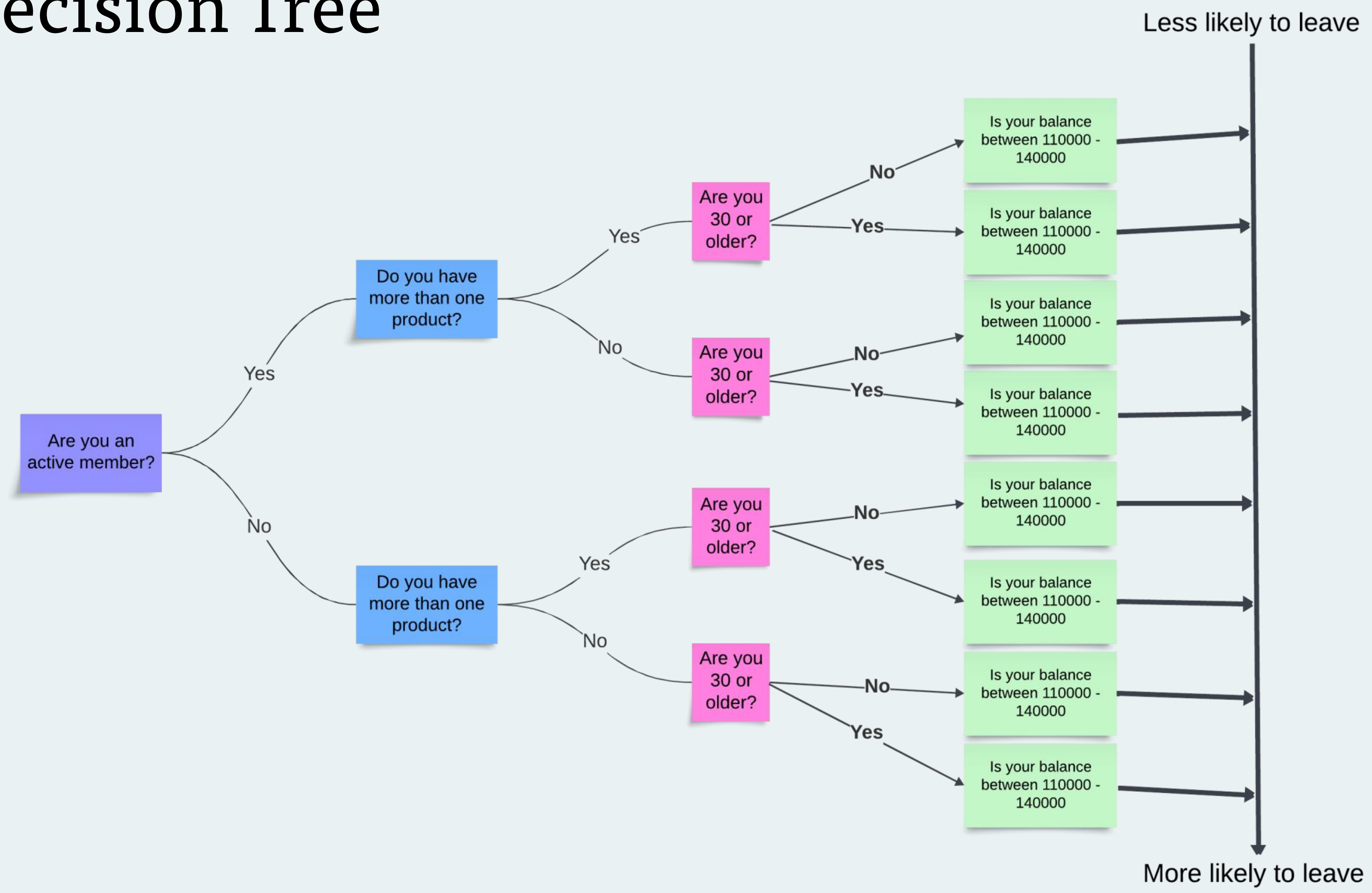
1) Is the customer an active member?

2) Does the customer have more than one bank product?

3) Is the customer 30 and older?

4) Is the customer’s balance within $110,000 - $140,000?

Decision Tree



Summary

Insights

1) Activity Status

- 70% of prior members were not active at the bank before leaving.
- This is the primary indicator if an individual will leave the bank.

2) Bank Products

- The more products (accounts, credit cards, loans) a client has, they are more likely to stay.
- Almost 70% of prior members had a single product before leaving the bank.

3) Age

- Current customers between 30 – 39 has the largest share of clients.
- The bank should promote bank products to this age group if they want to increase their retention rate.

4) Bank Balance

- Bank balances between $110,000 - $140,000 consist of 35% of former clients.
- When a customers balance meets this range, offer products such as investment advice or CDs so they will stay at the bank and invest that amount instead of taking it elsewhere.

Crop Yield Project Overview

Objective

This project analyzes agricultural data to predict crop yield based on factors like rainfall, fertilizer use, and irrigation. It includes exploratory data analysis, hypothesis testing, and predictive modeling, with results visualized through Tableau dashboards and stories.

My analysis informed a targeted marketing strategy aimed at reaching the right customer profiles with relevant products to improve sales performance.

Key Questions

1. Which variables have the strongest correlation with the crop yield?
2. Does the use of fertilizers improve yield for crops
3. Are yields significantly higher for crops that are irrigated vs non-irrigated crops?
4. Is there a certain region that produces more crop than others?
5. Does weather condition have an effect on crop yield?

Dataframes

- **Crop Dataset:** sole data set used for the entirety of the project

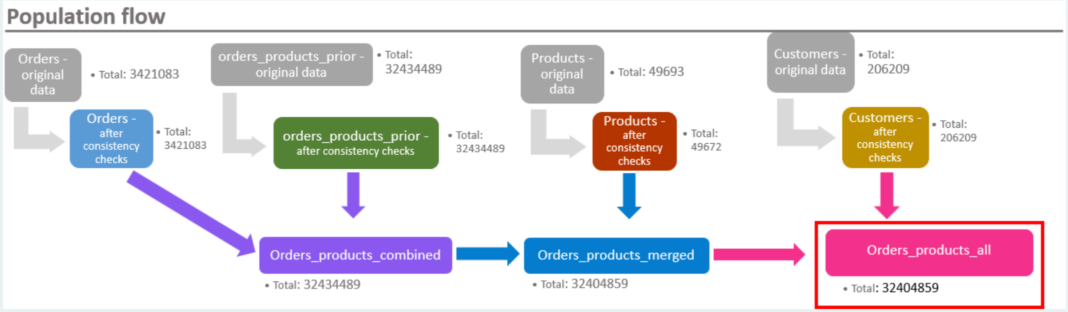
Skills

- Python
- Data wrangling & cleaning
- Deriving variables
- Grouping data
- Aggregating data
- Linear Regression
- Cluster Analysis

Challenges

- Deriving valuable insights from an open-sourced data set where there was no guarantee of said insights
- Learning linear regression and cluster analysis using sklearn library

Python Coding




```python
# creating a loop statement
result = []

for value in ords_prods_all["department_id"]:
  if value == 18:
    result.append("baby_purchase")
  elif value == 8:
    result.append("pet_purchase")
  else:
    result.append("neither")
```

```python
# creating "baby_or_pet" column based on results above
ords_prods_all['baby_or_pet'] = result
```

```python
# creating age_group" column
ords_prods_all.loc[ords_prods_all['age'] <= 29, 'age_group'] = 'young_adult'
```

```python
ords_prods_all.loc[(ords_prods_all['age'] > 29) & (ords_prods_all['age'] <= 49), 'age_group'] = 'adult'
```

```python
ords_prods_all.loc[(ords_prods_all['age'] > 49) & (ords_prods_all['age'] <= 64), 'age_group'] = 'older_adult'
```

```python
ords_prods_all.loc[ords_prods_all['age'] > 64, 'age_group'] = 'senior'
```

If-statements

```python
# finding missing values
df_prods.isnull().sum()
```

```
product_id        0
product_name     16
aisle_id          0
department_id     0
prices            0
dtype: int64
```

```python
# creating subset of missing values
df_nan = df_prods[df_prods['product_name'].isnull() == True]
```

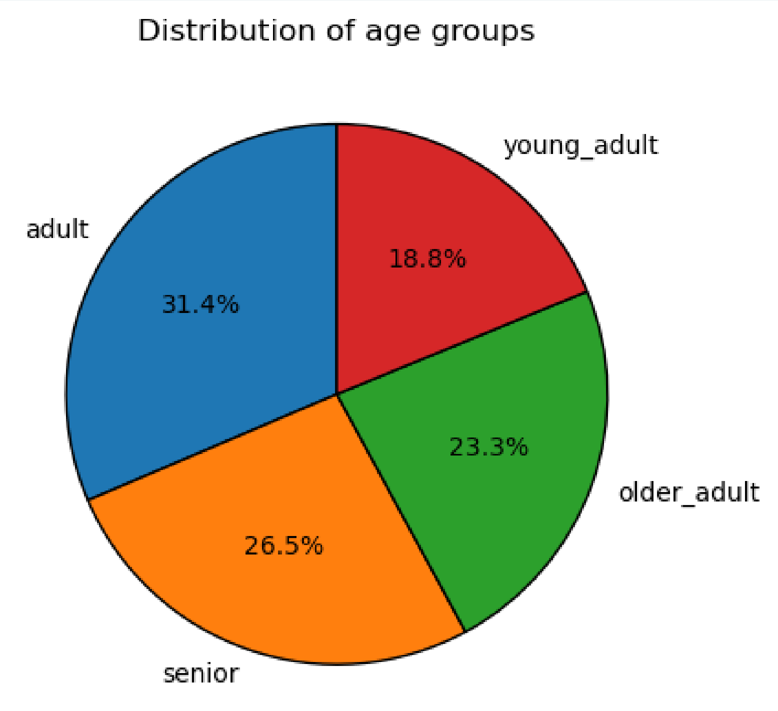
Data Cleaning

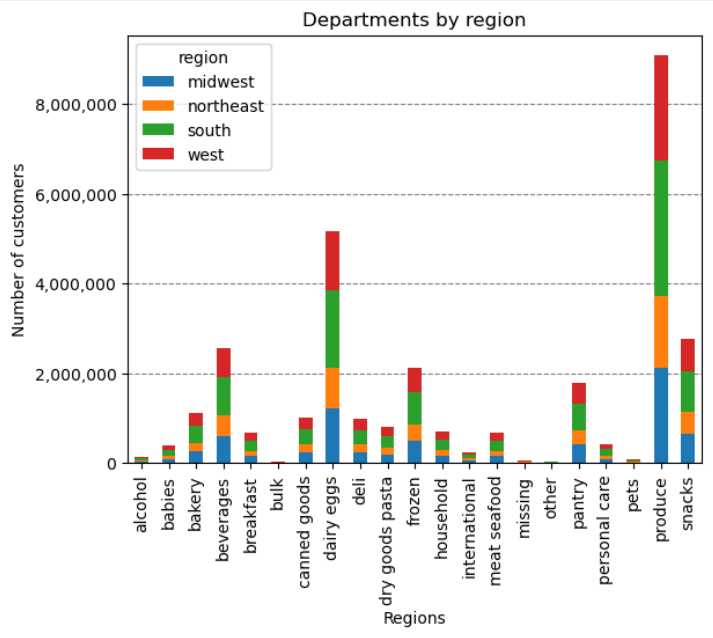
```python
for col in df_ords.columns.tolist():
  weird = (df_ords[[col]].map(type) != df_ords[[col]].iloc[0].apply(type)).any(axis = 1)
  if len (df_ords[weird]) > 0:
    print (col)
else : print ("no mixed-type data")
```

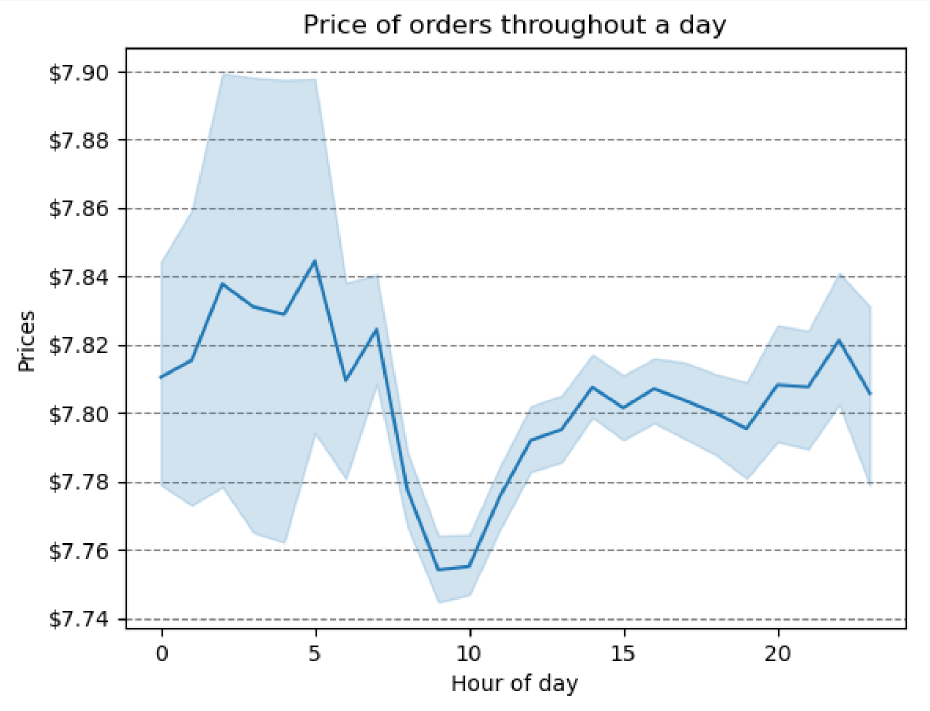
```python
line = sns.lineplot(data = line_chart_df, x = 'order_hour_of_day', y = 'prices')
plt.title('Price of orders throughout a day')
plt.ylabel('Prices')
plt.xlabel('Hour of day')
plt.gca().set_axisbelow (True)
plt.grid(axis = 'y', color = 'grey', linestyle = 'dashed')
plt.gca().yaxis.set_major_formatter(format('${x:,.2f}'))
```

Visualization creation

Analysis







Python and Jupyter notebooks were used to analyze sales patterns and develop customer profiles. This provided valuable insights to understand customer behavior, product performance, and improve segmentation strategies.

Summary

Insights

Busiest days and hours

- The busiest days of the week are Saturday, followed by Sunday.
- The busiest time of day is between 10am – 3pm

Product Price Ranges

- Mid-range products: majority of products in Instacart catalog (pantry, personal care, beverages).
- Low-range products: 2nd highest products in Instacart catalog (snacks, produce).
- High-range products: lowest number of products in Instacart (meat & seafood).

Customer Profiles

- Young adults count for the least amount of customers.
- Low-income counts for least amount of customers.
- ‘Married’ user is the base customer.

Recommendations

Implement marketing campaigns during slower weekday hours to boost orders and revenue, with messaging focused on quicker delivery times.

Mid-range products: Offer limited-time discounts on mid-range products, especially on items that customers buy regularly

Low-range products:

- implement ‘add-on’ suggestions when customers are about to check out.

- Create a ‘buy more, save more’ campaign.

High-range products: Create ‘loyalty program incentives’ for purchasing high-range products that are redeemable on future purchases.

Young adults: promote trendy products such as organics or plant-based foods

Low income: Create ‘value pack campaigns to offer significant savings or develop flexible payment options to make larger purchases more manageable.

Married user:

- Support the ‘base’ by developing family meal kits and bundles

- Create ‘baby essentials’ campaign with subscription services for diapers, wipes and formula.

Thank you

Jeffrey Frankenfeld

303-725-2012

jeff-frankenfeld@outlook.com

Click to connect!