Laurine Tuchagues

Data Analytics Portfolio

Tools

About me

My name is Laurine, I grew up in Marseille and live now in Berlin since 7 years, I love to travel (mostly by bike), speak fluently French, Italian, English and I currently study for a B2 in German.

Experienced in online marketing reputation management and market trend analysis, I transitioned to data analytics, blending creative problem-solving with quantitative analysis to drive strategic decisions and business outcomes. Eager to apply these skills to contribute actionable insights and foster growth within dynamic team environments.

Skills & expertise

Grouping
Cleaning and summarizing data
Descriptive analysis

Excel
SQL
Python
Tableau

Relational databases
Data wrangling
Data merging
Deriving variables

Big data
Data ethics
GitHub

Portfolio samples

1/ GameCo: a videogame company

2/ Preparing for influenza season

3/ Rockbuster Stealth: online video rental

4/ Instacart: online grocery shop

5/ Anti-money laundering projects at Pig E. Bank, San Francisco

GameCo, a Videogame company

It’s October 2016 and GameCo’s executive board is planning the marketing budget for 2017. They’re assuming that sales for the various geographic regions have stayed the same over time, and they’ve asked you to look into the data to see if this is still true. If it’s not, the marketing budget will need to be redistributed among the regions to maximize return on investment.

SKILLS/ TOOLS /PROCEDURES

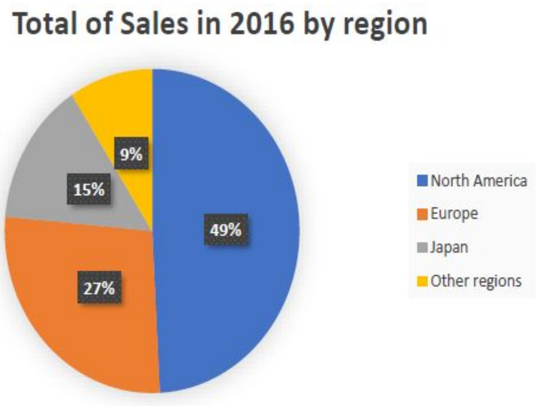
Advanced Excel
Grouping data
Summarizing data
Descriptive analysis
Visualizing results in Excel

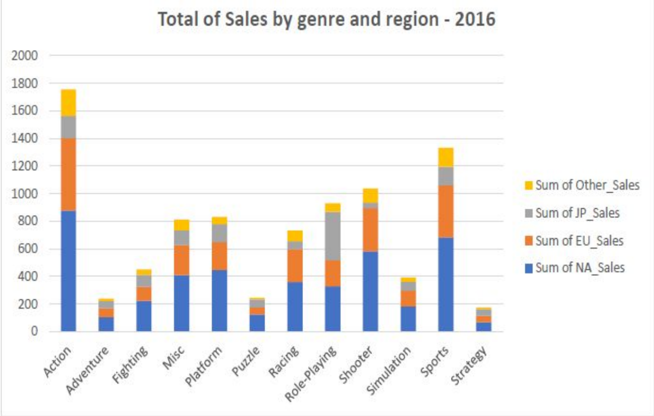
Marketing Consulting Proposal

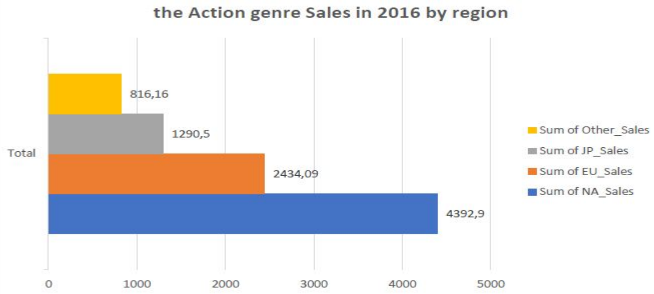
Problem Statement:
GameCo's executive board was concerned that the current distribution of sales across regions might not reflect real market demand, which could lead to a misallocation of the 2017 marketing budget. The goal was to analyze regional sales trends to ensure that marketing efforts were better aligned with sales performance in North America, Europe, Japan, and other regions.

Methods and Tools:
I used Excel for data grouping, summarizing, and visualization as it allowed for clear and immediate insights. The first step was filtering the dataset to focus on 2016 sales only. A pie chart was then used to represent market share across regions, providing an intuitive visual breakdown of each region's contribution to global sales. The descriptive analysis focused on identifying the regions with the highest sales to recommend where to concentrate marketing efforts.

Challenges:
One challenge was dealing with inconsistent data, as not all sales were categorized uniformly across regions. To address this, I standardized the data where possible and focused on the core regions for a more accurate comparison.

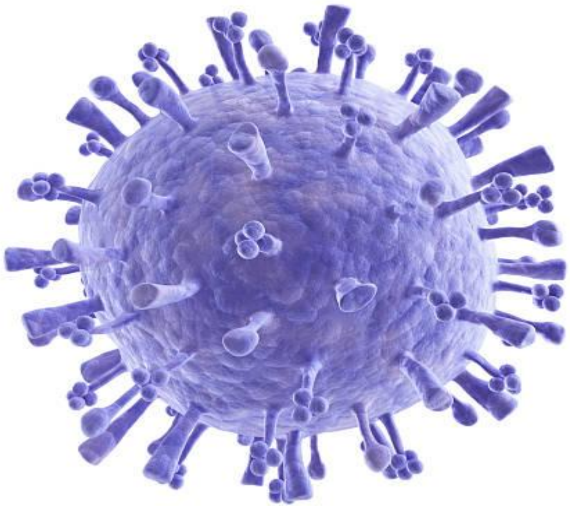






Retrospective:
Based on the analysis, I recommended that GameCo focus its marketing efforts on the Action and Sports genres in North America and Europe, while prioritizing the Role-Playing genre in Japan. In hindsight, I would have included user engagement metrics, such as hours played, to provide a more complete picture of player preferences across regions.

Preparing for Influenza Season



To help a medical staffing agency that provides temporary workers to clinics and hospitals on an as-needed basis. 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.

SKILLS/ TOOLS /PROCEDURES

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

DESCRIPTIVE ANALYSIS

Problem Statement:

The medical staffing agency needed to optimize staffing during influenza season, a period of high demand for temporary workers in clinics and hospitals. The goal was to forecast staffing needs based on historical influenza mortality data and demographics, ensuring adequate staffing levels during the peak season.

Methods and Tools:
I used Excel for data cleaning and transformation, and Tableau for visualizing mortality data. After cleaning the dataset and removing irrelevant entries, I focused on key demographic groups. Hypothesis testing was used to assess whether individuals aged 85+ were at a higher risk of influenza-related mortality. Tableau's visualization capabilities made it easy to present these results to stakeholders.

Challenges:
The biggest challenge was ensuring data accuracy, as some entries were incomplete or inconsistent. After identifying these issues, I had to decide whether to exclude or adjust certain data points to maintain the accuracy of the analysis.

DATA SPREAD

	Census 85+	Death 85+
Dataset name	Finale combine data	Finale combine data
Sample or populatior	Population	Sample
Normal distribution?	Skew right	Skew right
Variance	14932215425	303916
Standard deviation	122197	551
Mean	107846	463
Outlier percentage	5,9	3,9

t-Test: Two-Sample Assuming Unequal Variances

Variable 1: 85+ years old death rate
Variable 2: 75-84 years old death rate

	0,005	0,001
Mean	0,004	7E-04
Variance	3E-06	3E-07
Observations	458	458
Hypothesized Mean Difference	0	
df	533	
t Stat	36,46	
P(T<=t) one-tail	3E-147	
t Critical one-tail	1,648	
P(T<=t) two-tail	7E-147	
t Critical two-tail	1,964	

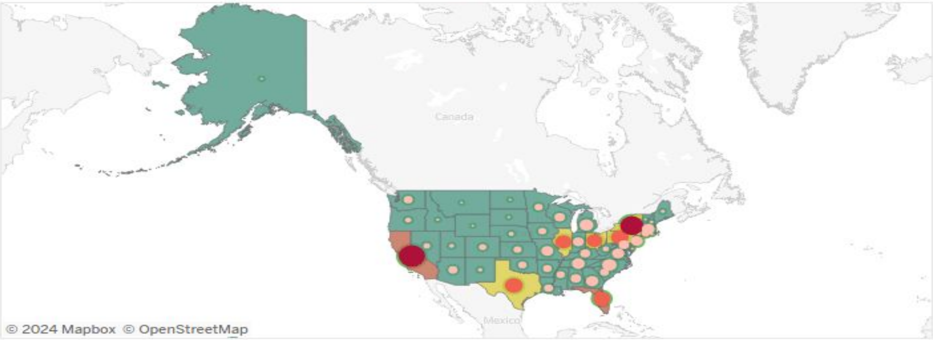
Retrospective:
The analysis confirmed that individuals aged 85+ were at a higher risk of mortality, with a p-value of 3E-147, which led to the recommendation to increase staffing in clinics serving older populations. Going forward, I would incorporate real-time outbreak data to create more dynamic staffing models. Additionally, I recognized the need to improve my predictive modeling skills for future forecasting projects.

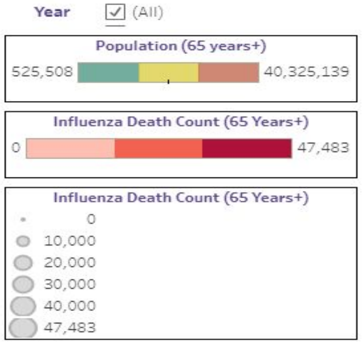
The P-value is 3E-147
The significance level is higher than the p-value : We can consider that the 85+ population is at-risk.

Presentation to stakeholders

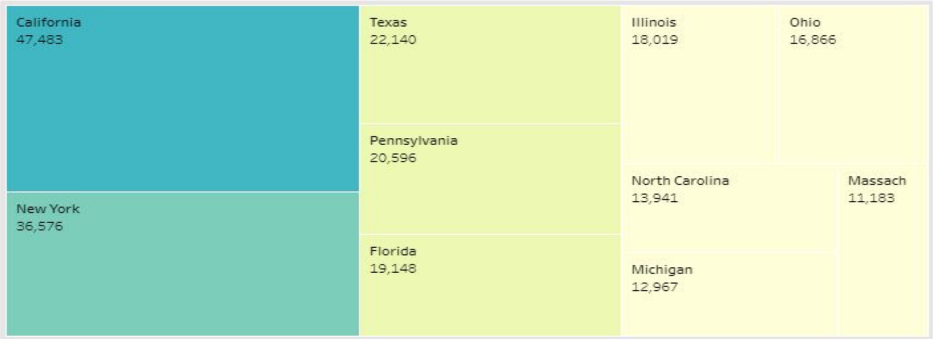
Data Visualization on Tableau

Influenza Death in the U.S (2007 - 2019)
for Age Group of 65+ Years





Influenza Mortality in the U.S. in 2009-2017
(population over 65 years)



We note that the mortality rate of the at-risk population (+65) is immediately correlated with the overall mortality rate in the states most affected by influenza.
The ten states concerned are as follows: California, New York, Texas, Pennsylvania, Florida, Illinois, North Carolina, Michigan, Ohio and Massachusetts.

Links of the full presentation:

-**Tableau Public**
-**YouTube**

Rockbuster Stealth

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 i s planning to use its existing movie licenses to launch an online video rental service in order to stay competitive.

SKILLS/ TOOLS /PROCEDURES

Create a data dictionary
Relational databases
SQL
Database querying
Filtering
Cleaning and summarizing
Joining tables
Subqueries
Common table expressions

Create a Data Dictionary

Based on an Entity Relationship Diagram (ERD)

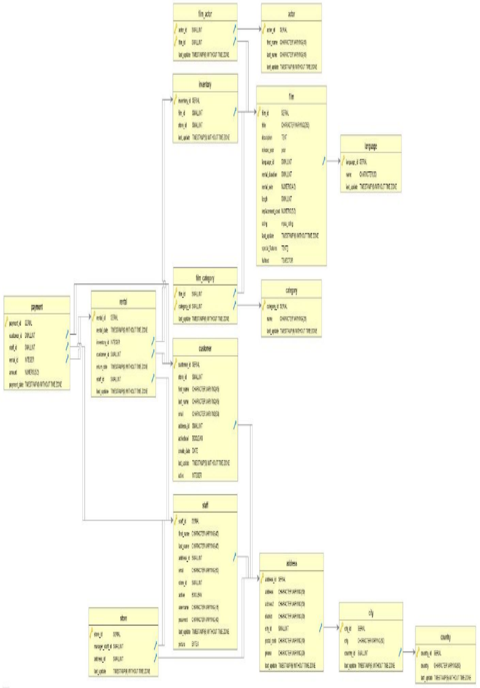
Problem Statement:

Rockbuster Stealth needed to transition from physical video rentals to an online streaming service to compete with companies like Netflix and Amazon Prime. The objective was to analyze existing customer data to inform the launch strategy, focusing on key revenue-generating countries and customer preferences.

Methods and Tools:
I used SQL to query the relational database, which contained Rockbuster's historical rental data. After loading the data into PostgreSQL, I created a data dictionary based on the entity relationship diagram (ERD). This helped ensure consistency when joining tables and creating subqueries. The analysis used common table expressions (CTEs) and subqueries to highlight regional trends and demographic preferences.

Challenges:
One challenge was identifying trends in a dataset that was based on physical rentals, as customer behaviors might differ in an online context. Additionally, analyzing non-English-speaking regions meant considering cultural differences in media consumption.

Retrospective:
The project revealed that a one-size-fits-all strategy would not work. Rockbusters content would need to be adapted for non-English-speaking regions, and I recommended expanding the content library to attract a broader audience. In the future, I would include qualitative data such as customer reviews to better understand customer sentiment and improve recommendations.



Exemple of Fact table

Rental

Column	Data Type	Description
Rental_id	Serial	Unique identifier for each rental transaction
Rental_date	Timestamp(6)without timezone	Data of rental occured
Inventory_id	Integer	Unique identifier for inventory
Customer_id	Smallint	Unique identifier for customer
Return_date	Timestamp(6)without timezone	Date of item being returned
Staff_id	Smallint	Unique identifier for staff
Last_update	Timestamp(6)without timezone	Date and time of record last updated

Exemple of Dimension table

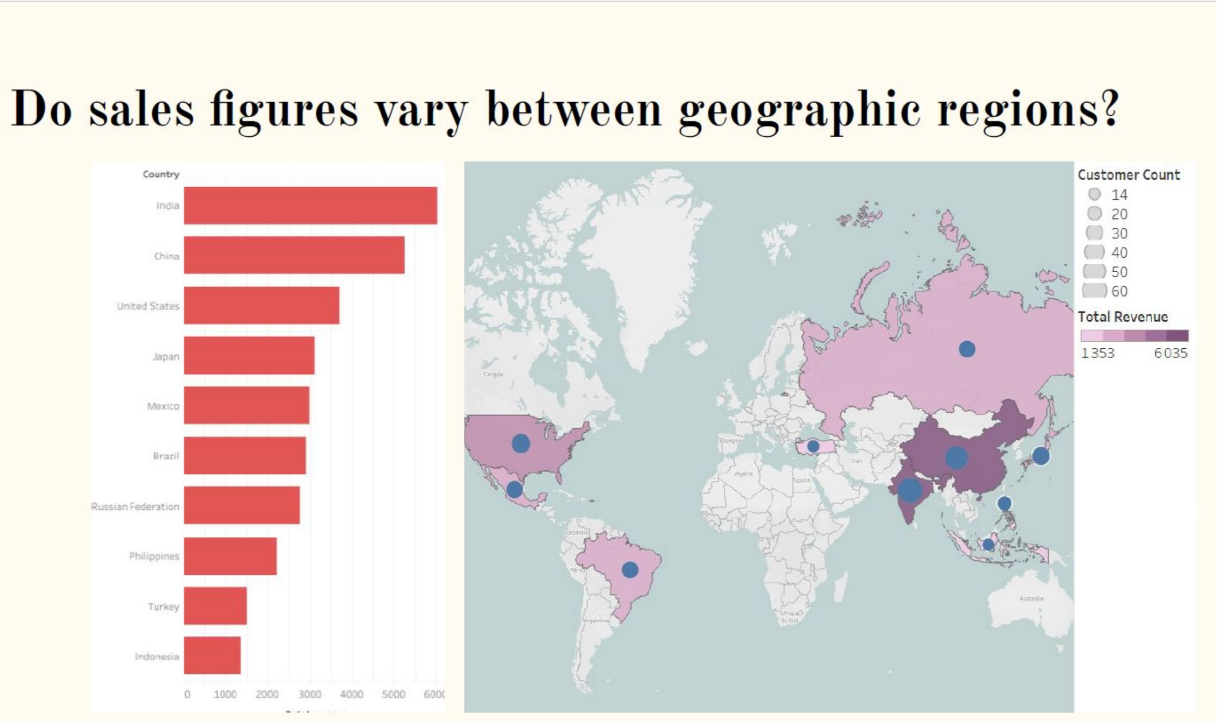
City

Column	Data type	Description
City_id	Serial	Unique identifier of a city
City	Character Varying(50)	Name of city
Country_id	Smallint	Unique identifier of a country
Last_update	Timestamp(6)without timezone	Time of last update

Presentation to Skateholders

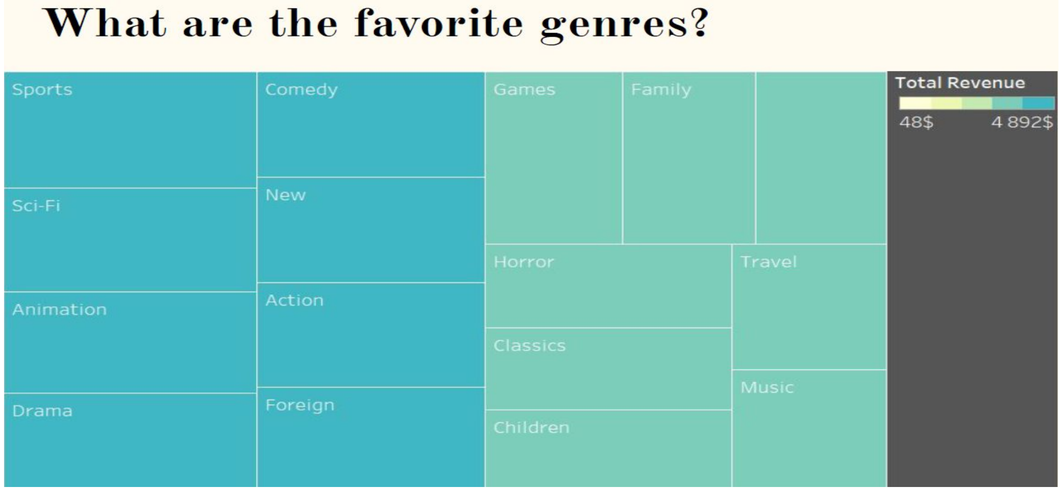
The customer base predominantly resides in non-English-speaking regions, necessitating the adaptation of our content and strategies to align with diverse cultural contexts.

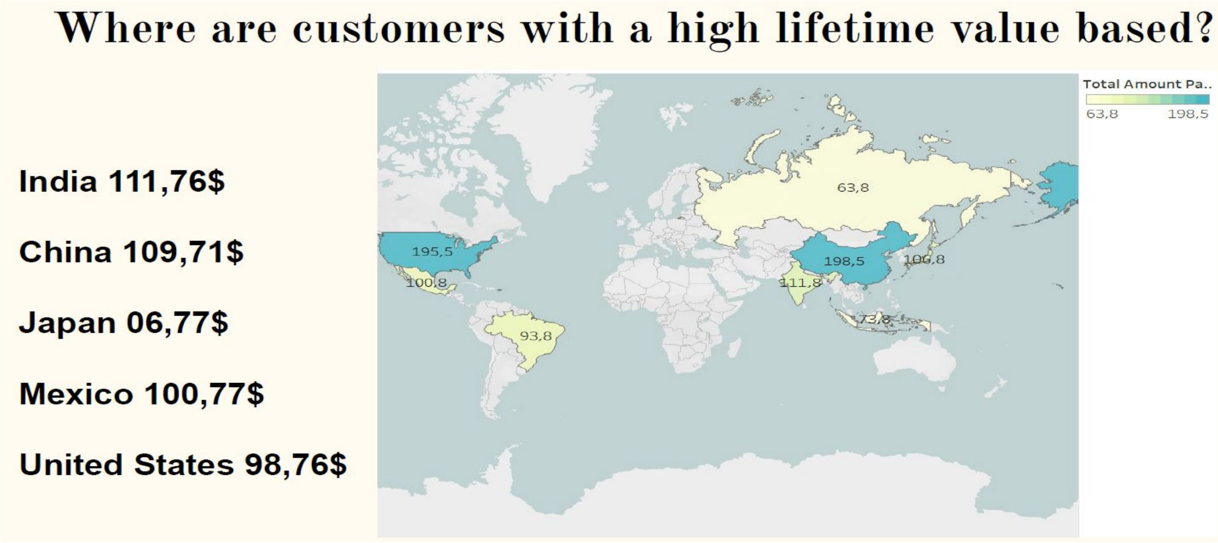
Certain key countries serve as primary revenue generators for our business, warranting concentrated marketing efforts tailored to their respective demographics and preferences.



Recommendations

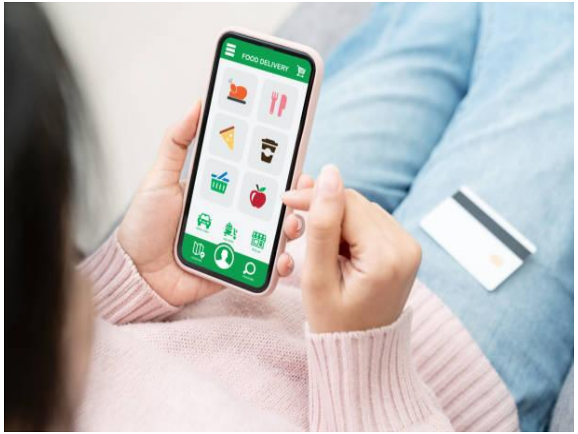
The exclusive presence of English-language films from 2006 in our library may limit our appeal to a broader audience. Exploring avenues for increased flexibility could enhance overall customer satisfaction and engagement.





Instacart
Grocery Basket

Instacart is an online grocery store that operates through an app. It already has very good sales, but they want to uncover more information about their sales patterns. My task was to perform an initial data and exploratory analysis of some of their data in order to derive insights and suggest strategies for better segmentation based on the provided criteria.



SKILLS/ TOOLS /PROCEDURES

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

Data Wrangling, derivation and aggregation

instacart

Wrangling steps

Columns dropped	Columns renamed	Columns' type changed	Comment/Reason
	order_dow		
eval_set			
	Gender		decapitilised for consistency reasons
	State		decapitilised for consistency reasons
	_merge		the original named created error while merging with orders/products data set
	Age		
First_name			not useful for analysis
surname			
		order_id (from int to str)	minimize memory consumption
		User_id (from int to str)	minimize memory consumption
		product_id (int64 to int32)	minimize memory consumption
		department_id (int64 to int8	minimize memory consumption
		orders_days_of_week (int64	minimize memory consumption
#NAME?		add_to_cart (int64 to int16	minimize memory consumption
		reordered (int64 to int16)	minimize memory consumption
		max_order (int64 to int32)	minimize memory consumption
		age (int64 to int8)	Merging process
		income (int64 to int32)	minimize memory consumption
		n_dependents (int64 to int3	minimize memory consumption

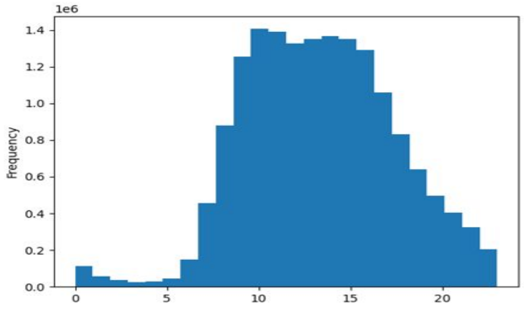
I needed to perform data wrangling, aggregation, and derivation for Instacart to clean and preprocess the raw data, ensuring it was accurate and usable for analysis. I used Python code such as **pandas** for data manipulation, **groupby()** for aggregation, and custom functions for deriving new metrics, which allowed me to uncover valuable insights from the data. This process enabled us to understand customer purchasing patterns and optimize inventory management.

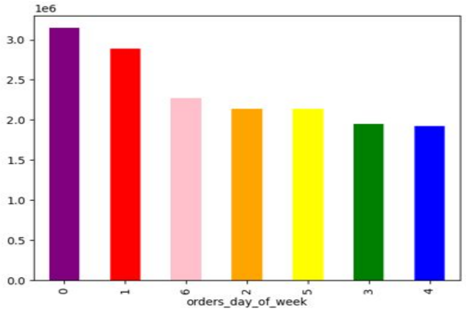
Column derivations and aggregations

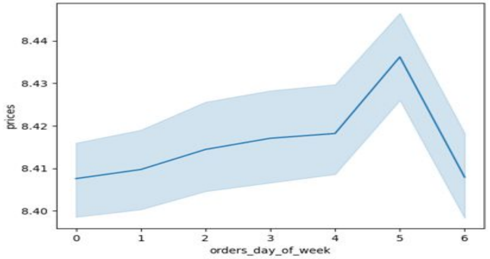
Dataset	New column	Column/s it was derived from	Conditions
orders_checked	first_order	days_since_prior_order	**["days_since_prior_order"].isnull()** True (206,209) False (3,214,874)
orders_products_me	price_range_loc	prices	**"prices" > 15** Mid-range product (21,860,860): **5 < "prices" <=15** Low-range product (10,126,321): **"prices" <= 5**
orders_products_me	busiest_days	orders_day_of_week	Busiest days (11,862,627): **if value in [0, 1]** Slowest days (7,623,037): **elif value in [3, 4]** Regularly busy (12,914,068): **else (value in [2, 5, 6])**
orders_products_me	busiest_period_of_day	order_hour_of_day	Most orders (21,114,733): **if value in [10, 11, 14, 15, 13, 12, 16, 9]** Average orders (9,996,048): **elif value in [17, 8, 18, 19, 20, 7, 21, 22]** Fewest orders (1,288,951): **else (value in [23, 6, 0, 1, 5, 2, 4, 3])**
orders_products_me	loyalty_flag	max_order (which was derived from user_id & order_number)	Loyal customer (10,282,763): **"max_order" > 40** Regular customer (15,874,128): **10 < "max_order" <= 40** New customer (6,242,841): **"max_order" <= 10**
orders_products_me	spending_flag	user_mean_product_price (which was derived from user_id & prices)	Low spender (32,280,013): **"user_mean_product_price" < 10** High spender (119,719): **"user_mean_product_price" >= 10**
orders_products_me	frequency_flag	median_days_between_orders (which was derived from user_id & days_since_prior_order)	Frequent customer (21,556,644): **"median_days_between_orders" <= 10** Regular customer

Data Visualization & Insights

Busiest days of the week and hours of the day

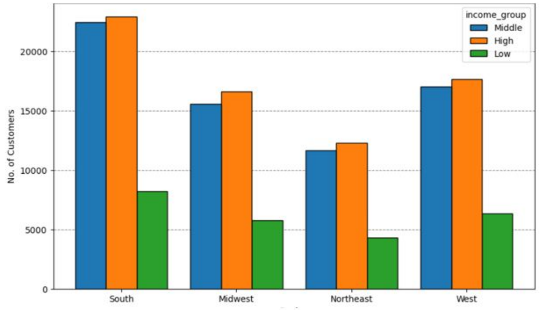


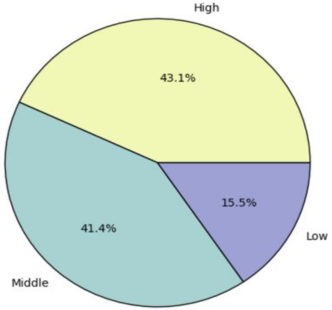




The first chart is based of the numbers of orders per hour (on 24 hours) and suggest that the busiest hours are between 9AM and 5PM. On the second column chart, which counts the numbers of orders weekly, we can observe that Saturday and Sunday are the busiest days (presented here respectively as 0 and 1)

Spending habits by income, age, and regions





In order to have a better view the profiles of the customers and their habits, we divided them in 4 regions: South, Midwest, Northeast and West.

We can notice on this column chart that most of the Instacart customers are between Middle and High income group. The Southern region is the most concerned since it count up to 23 000 of high income customers for only 12000 of the same category in the Northeast.

Problem Statement:
Instacart, a popular online grocery store, wanted to uncover more information about their sales patterns to optimize inventory management and improve customer segmentation. The goal was to perform an initial exploratory analysis of the data to derive actionable insights and suggest better customer segmentation strategies.

Methods and Tools:
I used Python for data wrangling, aggregation, and derivation. This included cleaning and preprocessing raw data to ensure accuracy. Using libraries such as `pandas`, I grouped and merged data, created new variables, and conducted exploratory analysis to identify sales trends. I also created visualizations to highlight key findings, such as the busiest ordering hours (9 AM to 5 PM) and the busiest days (Saturday and Sunday), using population flow analysis to better understand customer habits.

Challenges:
One of the challenges was handling the large volume of data and ensuring that all records were consistently formatted for analysis. Additionally, interpreting the sales patterns across different regions and income groups required careful segmentation of the data to derive meaningful insights.

Retrospective:
The analysis showed that most orders came from middle- to high-income customers, particularly in the Southern region. Going forward, I would consider integrating additional datasets, such as competitor sales data, to further improve customer segmentation and inventory management strategies. I also realized that I could benefit from deepening my knowledge of advanced clustering algorithms to refine customer profiles further.

Pig E. Bank

To increase customer retention, the sales team wants to identify the leading indicators that a customer will leave the bank. I have created a table of client attributes that I believe could indicate whether customers will leave—for example, age, estimated salary, etc. I am going to use this information to identify the top risk factors that contribute to client loss and model them in a decision tree.

SKILLS/ TOOLS /PROCEDURES

Big data
Data ethics
Data mining
Predictive analysis
Time series analysis and forecasting
Using GitHub

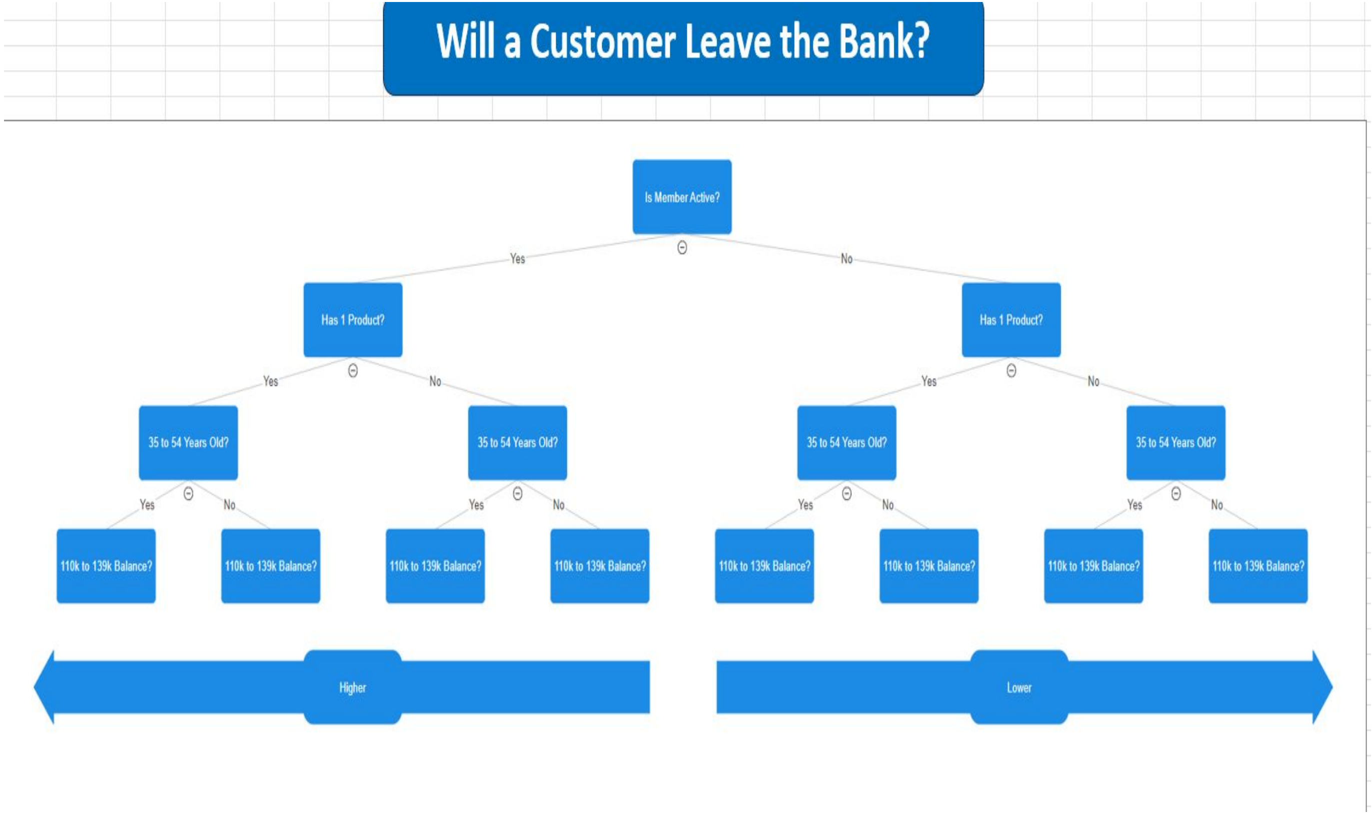
Creating a Decision Tree

Problem Statement:
Pig E. Bank sought to identify leading indicators that a customer might leave the bank in order to increase retention. Specifically, the sales team wanted to know which client attributes—such as age or balance—were predictive of customer churn. My task was to model these indicators using a decision tree and provide insights to improve customer retention.

Methods and Tools:
For this project, I used SQL for querying and data manipulation, combined with Python for building a decision tree model. The analysis involved cleaning the customer dataset, deriving relevant variables (e.g., credit score, balance, and age), and using these factors to create a predictive model. The decision tree was visualized to highlight the most important predictors of customer churn.

Challenges:
One challenge was the dataset's complexity—while some variables, such as age, showed strong correlations with customer churn, others, like credit score, were less predictive. Balancing the need for model accuracy while keeping it interpretable for the business team was also a consideration.

Retrospective:
The decision tree revealed that customers aged 35-54 with high account balances were the most likely to leave the bank. I recommended targeted retention efforts for this demographic. In retrospect, I would further explore other machine learning techniques, such as random forests or logistic regression, to see if they provide better predictions. Additionally, I would look into incorporating more behavioral data to enhance the model's accuracy.



Seasonal Weather effects on Health

This project explores the relationship between weather factors—including precipitation, humidity, temperature changes, and pressure—and the number and severity of reported health condition flare-ups.

SKILLS/ TOOLS /PROCEDURES

Tableau
Data ethics
Big Data
Time series analysis
Using GitHub

Problem Statement:
This project aimed to explore the relationship between weather patterns (including precipitation, humidity, temperature changes, and pressure) and the number and severity of reported health condition flare-ups, particularly focusing on conditions such as asthma and arthritis. The goal was to uncover potential correlations that could help predict when and where healthcare services might see increased demand due to weather-related health issues.

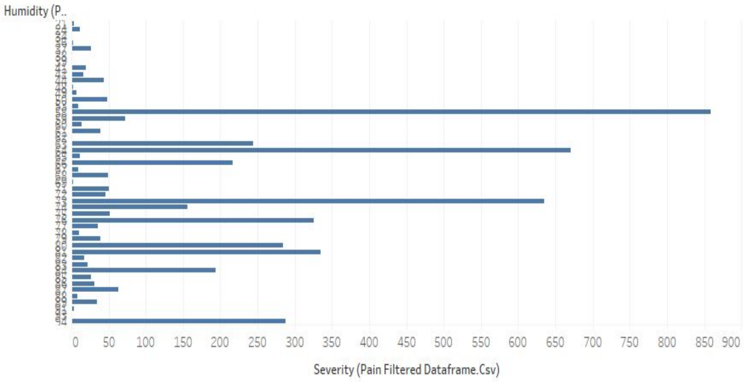
Methods and Tools:
I used Tableau for data visualization and Python for time series analysis. The dataset contained weather variables along with health data related to reported flare-ups. I cleaned the data to ensure it was suitable for analysis, and then I used time series analysis to explore whether there were statistically significant trends connecting weather conditions with health outcomes. The initial focus was on asthma and arthritis due to their known sensitivity to weather changes.

Challenges:
A key challenge was the accuracy of the data, particularly in terms of how consistently flare-up severity was reported. Additionally, the data sample for certain demographics, such as males, was limited, making it difficult to generalize findings to the broader population. There was also a lack of statistical significance in some of the initial correlations, requiring further refinement of the data analysis methods.

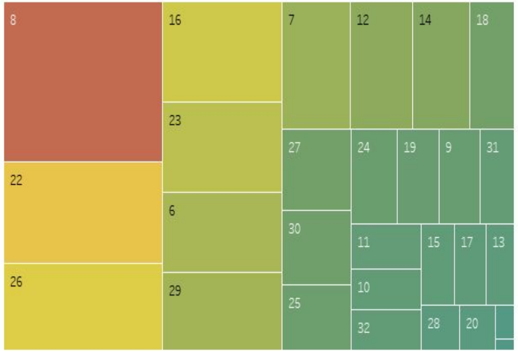
Observing Data Insights

Pain Conditions & Weather

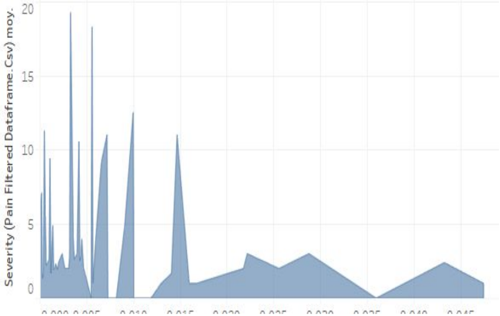
Average Severity of Pain Related Reports and Humidity



Pain Severity & Report Counts by Temperature Change



Pain Severity & Precipitation

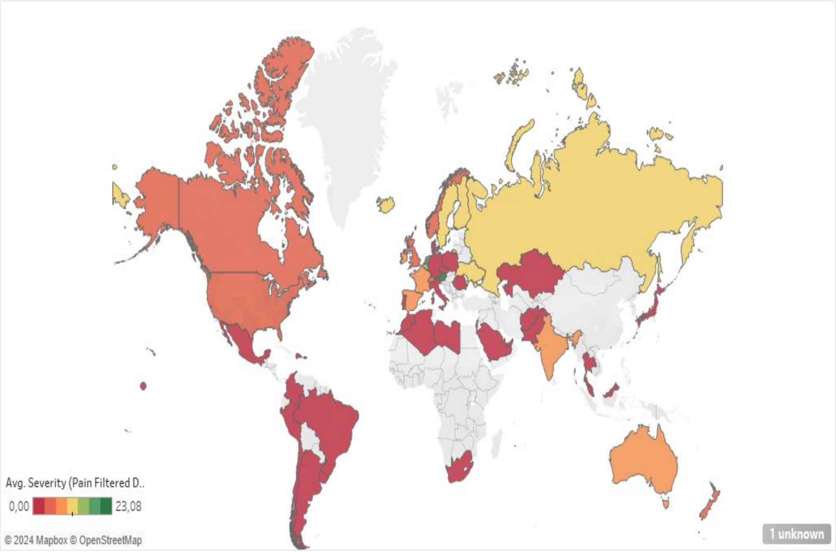


Create another approach

Retrospective:
While some potential trends emerged—such as a possible seasonal pattern in arthritis flare-ups—there was insufficient statistical significance to draw definitive conclusions. In the future, I would focus on improving data collection, particularly for male subjects, and refine the pain severity measurement. A more focused analysis, such as limiting the dataset to female participants, could also help in drawing stronger conclusions. Additionally, I would consider developing predictive algorithms to better understand how weather conditions could forecast health flare-ups, which would enhance the healthcare system's preparedness for seasonal health challenges.

Pain Conditions Report

Average Severity of Pain Condition Related Reports by Country



Average Pain Severity Reported by Month

Severity & Count of Top Pain Related Conditions

Condition (Pain Filtered Dataframe.Csv)	# Reporting	Avg. Severity (Pain Filtered D..
Fibromyalgia	166,0	6,1
Rheumatoid arthritis	48,0	8,2
Lupus	42,0	6,3
Arthritis	41,0	8,5
Headaches	35,0	4,3
Depression	33,0	9,8
joint pain	31,0	6,8
Postural Orthostatic Tachycardia Syn..	30,0	6,0
Migraine	30,0	7,3
Chronic Pain	28,0	13,1

General Retrospective for All Projects

Throughout my portfolio projects, I've developed a strong ability to turn raw data into actionable insights that drive business decisions. My experience spans multiple industries, allowing me to approach problems with both technical expertise and a clear business focus.

1. Data Preparation
I became proficient in cleaning and structuring data using Python, SQL, and Excel. This ensures the accuracy and reliability of all analyses, which is foundational for making sound business decisions.

2. Analytical Rigor
I apply rigorous analytical methods, from statistical testing to predictive modeling, to uncover deep insights. My work predicting staffing needs during flu season or identifying customer churn factors demonstrates my ability to use data to inform strategic decisions.

3. Data Storytelling
Using Tableau and Python, I turn complex datasets into clear, actionable insights. My ability to communicate findings effectively ensures that stakeholders understand and act on the analysis, speeding up decision-making.

4. Business-Focused Problem-Solving
In every project, I align my analysis with business objectives. Whether reallocating marketing budgets for GameCo or improving customer retention at Pig E. Bank, I ensure my work has a direct impact on business outcomes.

5. Overcoming Challenges and Continuous Learning
I embrace challenges and constantly learn new techniques to improve my work. From handling incomplete data to refining statistical models, I've grown stronger in advanced analytics and predictive methods, always staying up to date with new tools and technologies.

Career highlights

ONLINE MARKETING REPUTATION MANAGER / BERLIN

JULY 2018 / JANUARY 2024

[BOHEI MARKETING GMBH]

CUSTOMER EXPERT / BERLIN

JANUARY 2017 / JULY 2018

[MOVINGA/PLAYSTATION BY SYKES]

Contact

Laurine Tuchagues

laurine.tuchagues@gmail.com

(+49)15751020353