Yan Zhang’s
Data Analytics Portfolio

List of Projects

- Māori Language Immersion in New Zealand Schools
- RockBuster Stealth LLC Marketing Transition Analysis
- Instacart Grocery Market Analysis
- Preparing for Flu Season
- Structured Topic Modelling of China’s TVET Assessment Policy
- GameCo Global Video Game Sales
- Peer Misconduct across School and Workplace Settings

Project 1: Māori Language Immersion in New Zealand Schools

Sourcing open data

Exploring relationships

Geographical visualisations

Supervised maching learning

Unsupervised machine learning

Time series analysis

Tools:

Introduction

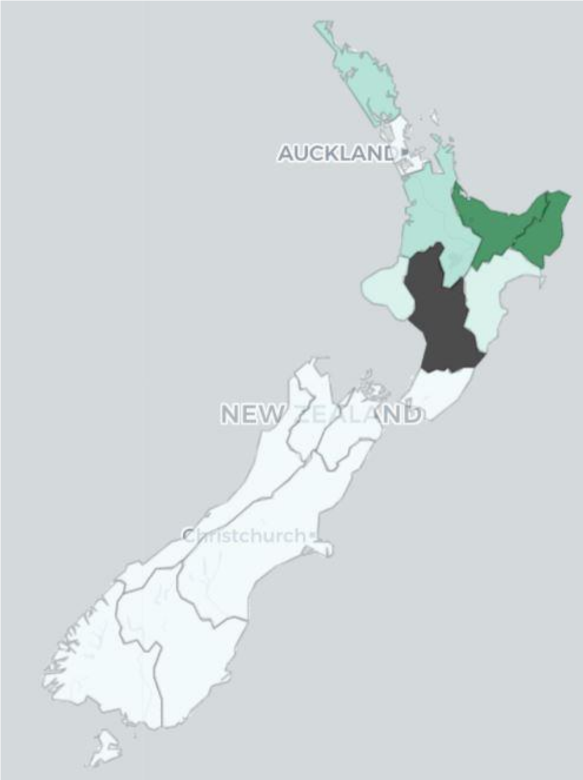
- This project examines how Māori language immersion education, aimed at both linguistic and cultural revitalisation, is implemented across New Zealand schools to inform education policy.
  - Examines where Māori immersion schools are located across the country.
  - Tests whether participation in immersion education is associated with ethnicity or socio-economic context.
  - Identifies clusters or patterns of schools that share similar profiles of ethnicity and decile.
  - Tracks how Māori immersion education has grown or shifted across years.

Data Source

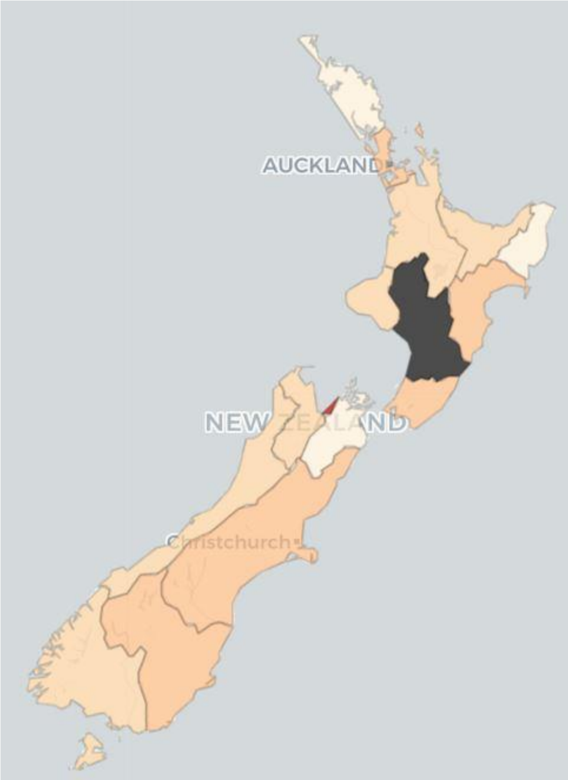
  - New Zealand government official website, School Rolls by School (2010–2024)

- Data Limitations
  - Because the dataset is aggregated at the school level, it cannot support individual-level or causal analysis.

Immersion across NZ Regions



High Immersion Area



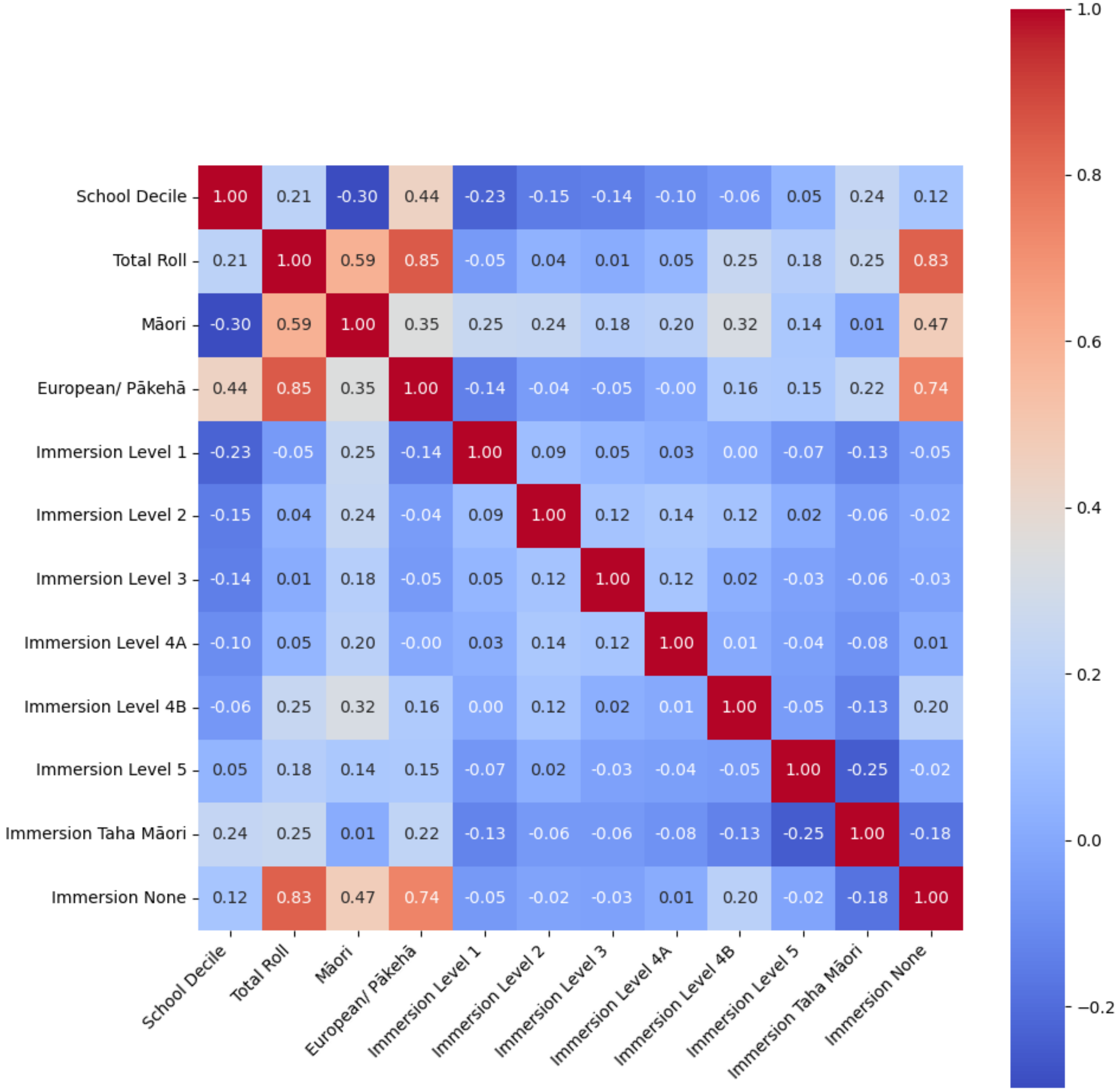
Low Immersion Area

Click to view the interactive maps

- Regions such as **Gisborne** and **Bay of Plenty** show higher coverage of Level 1 Māori language immersion. In contrast, non-immersion is more prevalent **across the country**.

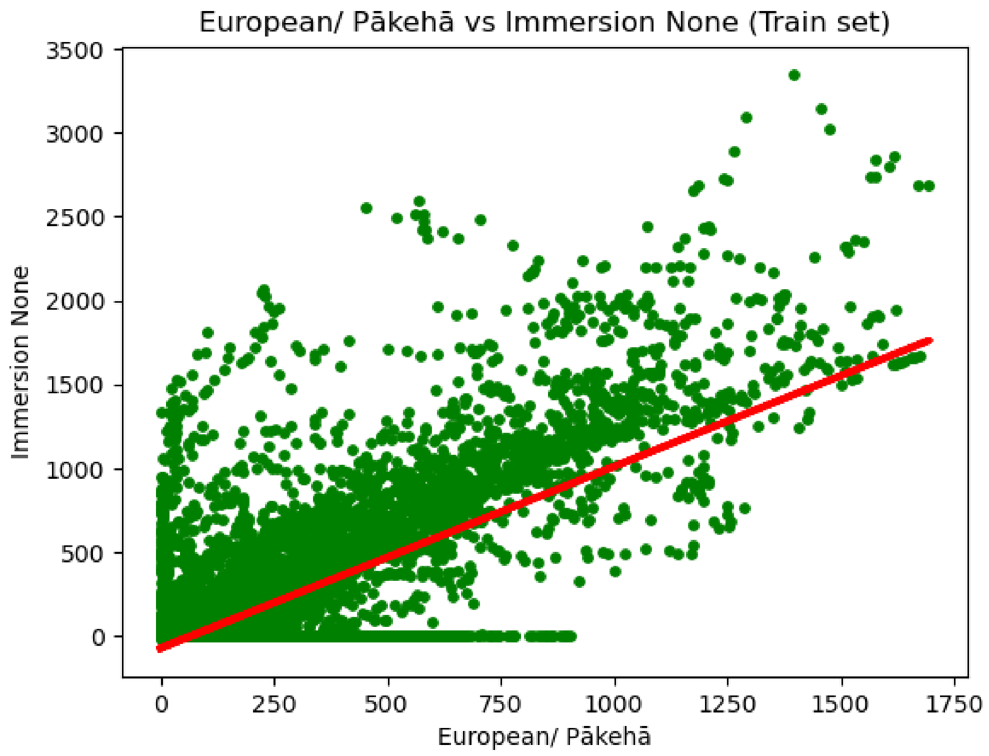
Correlations

- The variables of **school decile, ethnicity, and Māori language immersion** show different range of correlations.
  - **Immersion None** aligns with **European/Pākehā** (*r* = 0.74).
  - This suggests that schools with a higher proportion of **European/Pākehā students** tend to have **fewer students enrolled in Māori language immersion programmes**.



	School Decile	Total Roll	Māori	European/ Pākehā	Immersion Level 1	Immersion Level 2	Immersion Level 3	Immersion Level 4A	Immersion Level 4B	Immersion Level 5	Immersion Taha Māori	Immersion None
School Decile	1.00	0.21	-0.30	0.44	-0.23	-0.15	-0.14	-0.10	-0.06	0.05	0.24	0.12
Total Roll	0.21	1.00	0.59	0.85	-0.05	0.04	0.01	0.05	0.25	0.18	0.25	0.83
Māori	-0.30	0.59	1.00	0.35	0.25	0.24	0.18	0.20	0.32	0.14	0.01	0.47
European/ Pākehā	0.44	0.85	0.35	1.00	-0.14	-0.04	-0.05	-0.00	0.16	0.15	0.22	0.74
Immersion Level 1	-0.23	-0.05	0.25	-0.14	1.00	0.09	0.05	0.03	0.00	-0.07	-0.13	-0.05
Immersion Level 2	-0.15	0.04	0.24	-0.04	0.09	1.00	0.12	0.14	0.12	0.02	-0.06	-0.02
Immersion Level 3	-0.14	0.01	0.18	-0.05	0.05	0.12	1.00	0.12	0.02	-0.03	-0.06	-0.03
Immersion Level 4A	-0.10	0.05	0.20	-0.00	0.03	0.14	0.12	1.00	0.01	-0.04	-0.08	0.01
Immersion Level 4B	-0.06	0.25	0.32	0.16	0.00	0.12	0.02	0.01	1.00	-0.05	-0.13	0.20
Immersion Level 5	0.05	0.18	0.14	0.15	-0.07	0.02	-0.03	-0.04	-0.05	1.00	-0.25	-0.02
Immersion Taha Māori	0.24	0.25	0.01	0.22	-0.13	-0.06	-0.06	-0.08	-0.13	-0.25	1.00	-0.18
Immersion None	0.12	0.83	0.47	0.74	-0.05	-0.02	-0.03	0.01	0.20	-0.02	-0.18	1.00

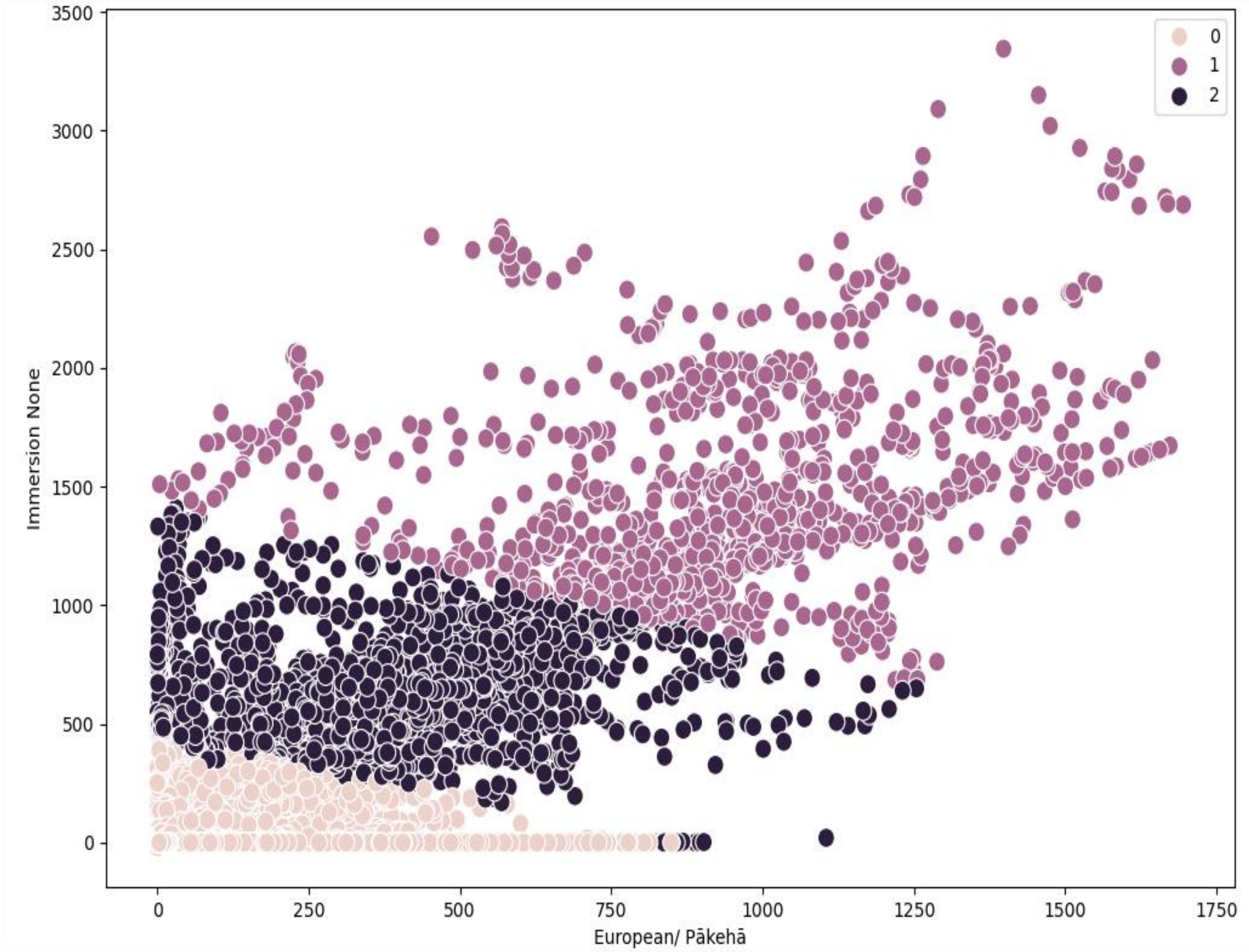
Supervised Machine Learning – Regression



- **European/Pākehā (%)** is a strong predictor of **Immersion None** (%), explaining **55% of the variance** (R^2 = 0.546, moderate fit).
- This suggests that schools with higher European/Pākehā populations tend to have lower Māori immersion participation, revealing a **demographic disparity**.
- The strong correlation between the two variables also suggests **potential multicollinearity**, which should be considered in further modelling.

Unsupervised Machine Learning – Clustering

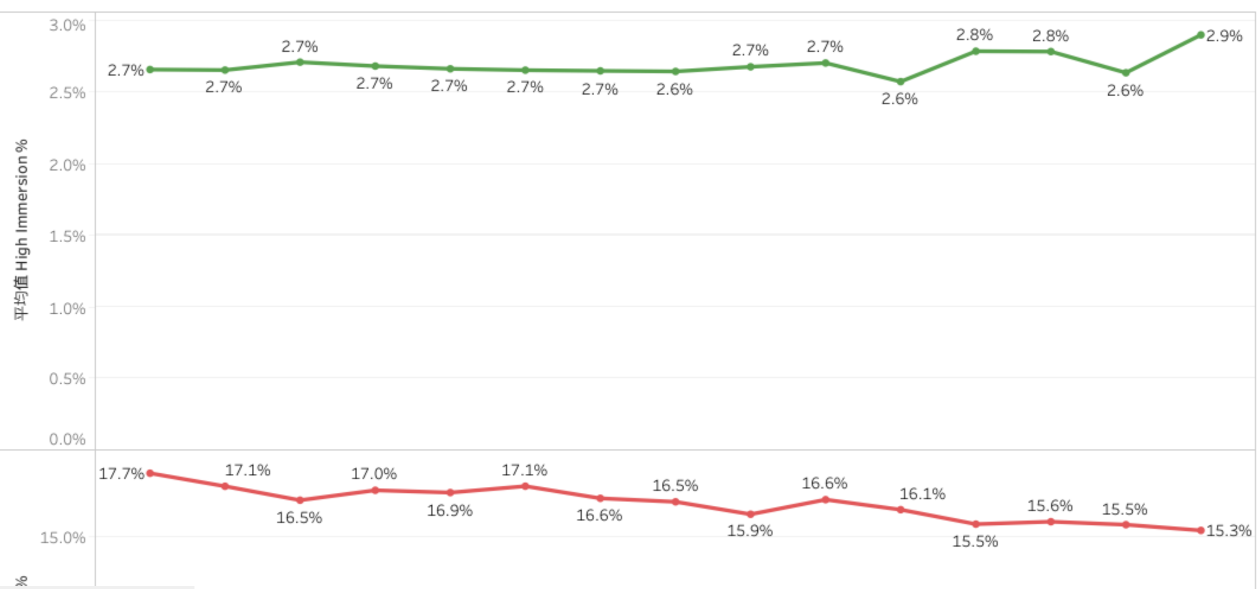
- Purple Cluster
  - High-Decile, European/Pākehā-Dominant
- Dark Purple Cluster
  - Mid-Decile, Mixed Composition
- Pink Cluster
  - Low-Decile, Higher Māori Immersion



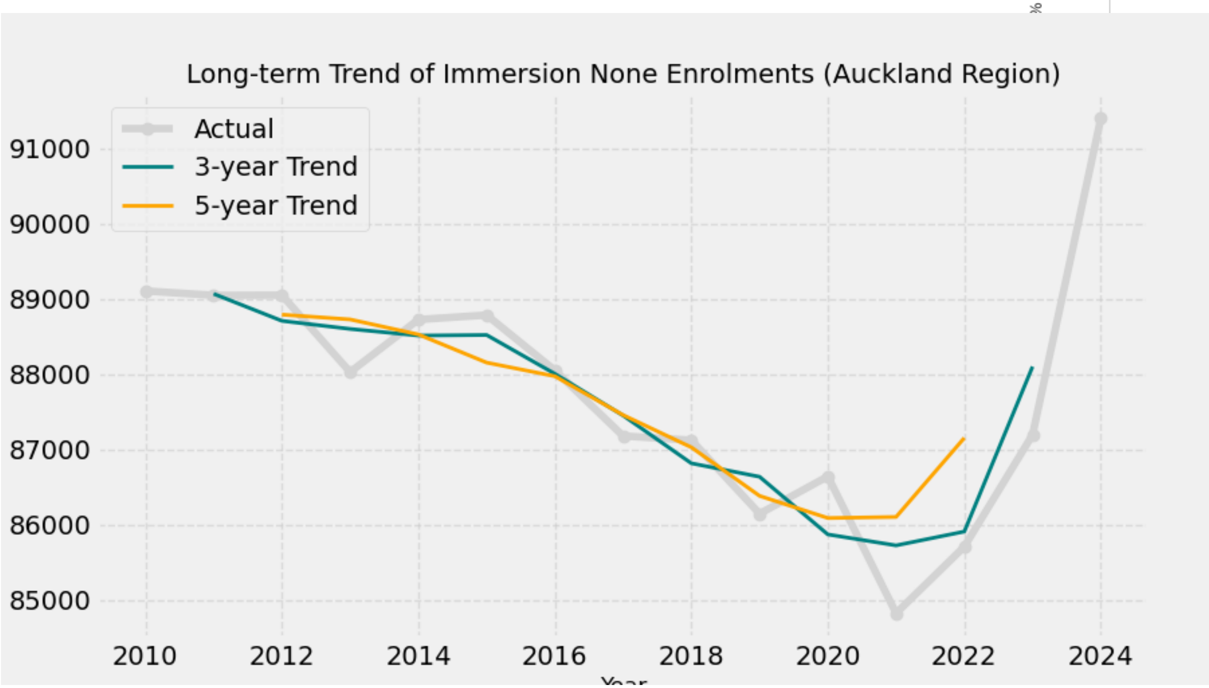
Time Series

- **Auckland:** Gradual improvement but still low immersion participation.

Avg Immersion % in Auckland



Long-term Trend of Immersion None Enrolments (Auckland Region)



Conclusion

- **Key Findings:**
  - High-Decile European/Pākehā schools show low participation.
  - Māori immersion is strongest in low-decile, Māori-majority schools.
  - Auckland shows gradual growth; Bay of Plenty maintains high immersion rates.
- **Recommendations:**
  - Expand Māori-medium options in high-decile areas.
  - Strengthen iwi–school partnerships to support te reo Māori.
  - Provide teacher training in Māori language and pedagogy.
  - Monitor participation by region and decile to ensure equity and progress.

🔗 GitHub Repository

🔗 Tableau Dashboard

Project 2: Rockbuster Stealth Marketing Transition Analysis

Data storage and structure

Summarising and cleaning data

Filtering data

Joining tables

Subqueries and CTE

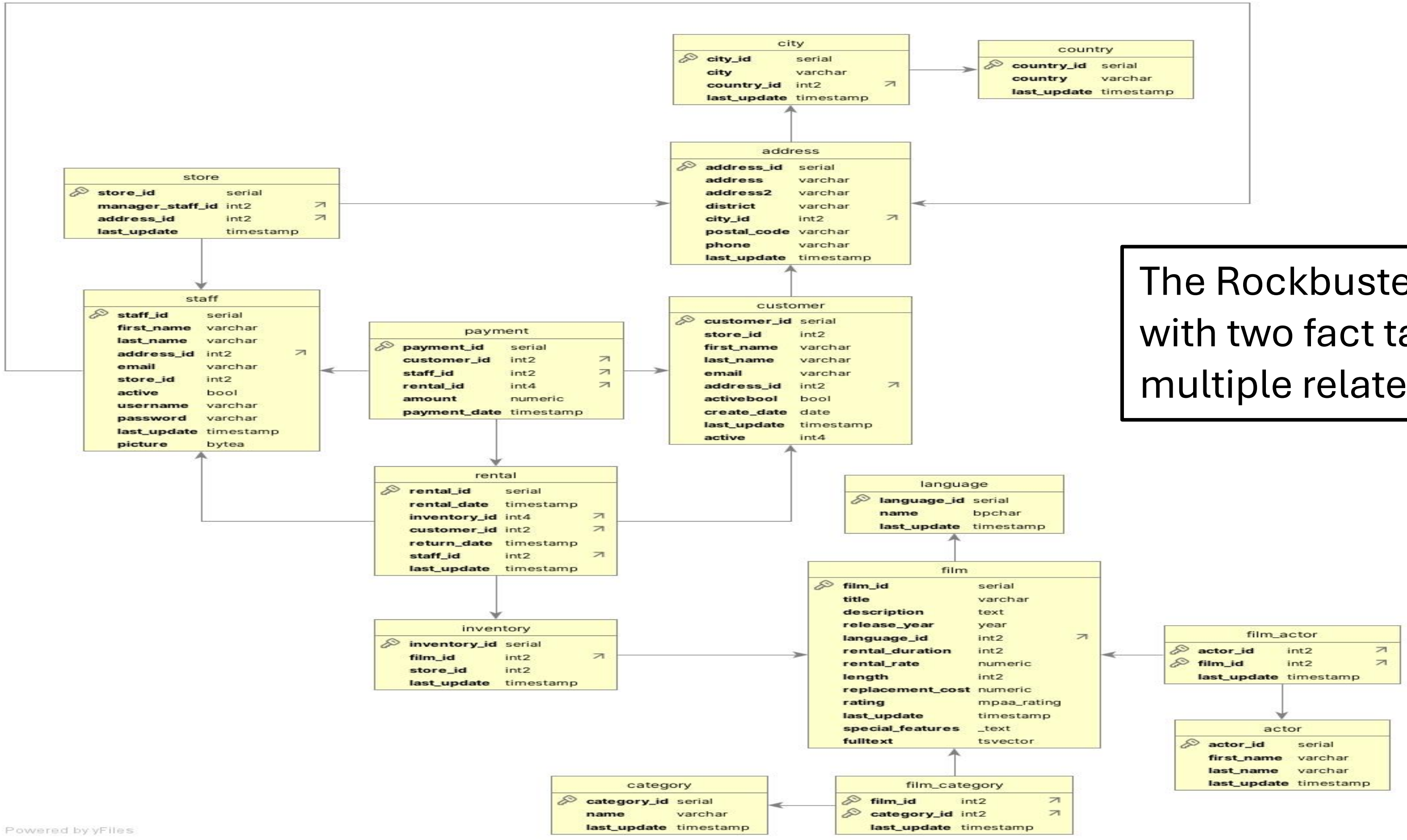
Presenting results

Tools:

Introduction

- **Context**
  - Rockbuster Stealth LLC is a movie rental company that used to have stores around the world. Facing stiff competition from streaming services, it is planning to launch an online video rental service in order to stay competitive.
- **Questions:**
  - Which countries are Rockbuster customers based in?
  - Which movies contributed the most/least to revenue gain?
  - What was the average rental duration for all videos?
  - Where are customers with a high lifetime value based?
  - Do sales figures vary between geographic regions?
- **Data**
  - Rockbuster internal records (films, rentals, customers, payments)

Entity Relationship Diagram



city	
city_id	serial
city	varchar
country_id	int2
last_update	timestamp

country	
country_id	serial
country	varchar
last_update	timestamp

store	
store_id	serial
manager_staff_id	int2
address_id	int2
last_update	timestamp

address	
address_id	serial
address	varchar
address2	varchar
district	varchar
city_id	int2
postal_code	varchar
phone	varchar
last_update	timestamp

staff	
staff_id	serial
first_name	varchar
last_name	varchar
address_id	int2
email	varchar
store_id	int2
active	bool
username	varchar
password	varchar
last_update	timestamp
picture	bytea

payment	
payment_id	serial
customer_id	int2
staff_id	int2
rental_id	int4
amount	numeric
payment_date	timestamp

customer	
customer_id	serial
store_id	int2
first_name	varchar
last_name	varchar
email	varchar
address_id	int2
activebool	bool
create_date	date
last_update	timestamp
active	int4

rental	
rental_id	serial
rental_date	timestamp
inventory_id	int4
customer_id	int2
return_date	timestamp
staff_id	int2
last_update	timestamp

language	
language_id	serial
name	bpchar
last_update	timestamp

film	
film_id	serial
title	varchar
description	text
release_year	year
language_id	int2
rental_duration	int2
rental_rate	numeric
length	int2
replacement_cost	numeric
rating	mpaa_rating
last_update	timestamp
special_features	_text
fulltext	tsvector

inventory	
inventory_id	serial
film_id	int2
store_id	int2
last_update	timestamp

film_actor	
actor_id	int2
film_id	int2
last_update	timestamp

actor	
actor_id	serial
first_name	varchar
last_name	varchar
last_update	timestamp

category	
category_id	serial
name	varchar
last_update	timestamp

film_category	
film_id	int2
category_id	int2
last_update	timestamp

The Rockbuster database have a snowflake schema with two fact tables (payment and rental) and multiple related dimension tables

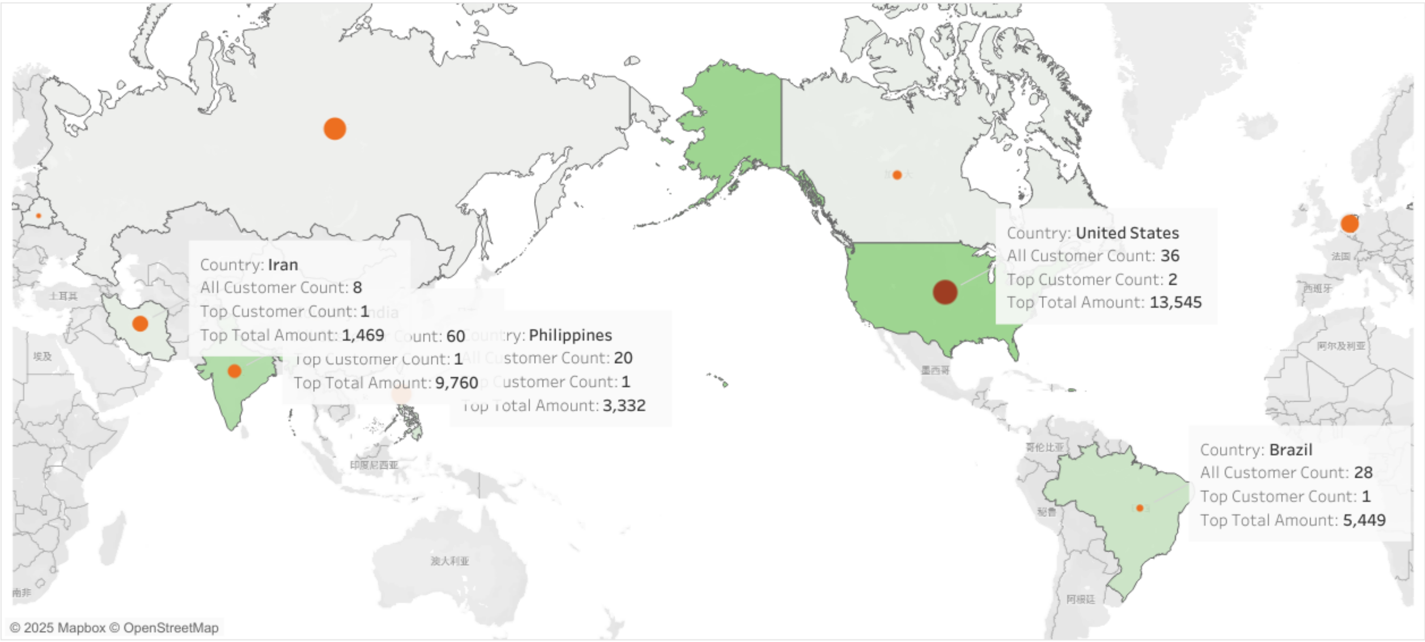
Descriptive Statistics of Data

The dataset includes **599 customers** from **600 cities** across **109 countries**. It contains **1,000 films** spanning **20 genres**, generating a total revenue of **$61,312.04**.

	Average	**Maximum**	**Minimum**
Rental Duration (Day)	5	7	3
Rental Rate ($)	3	5	1
Length(Min)	115	185	46
Replacement Cost ($)	10	30	20

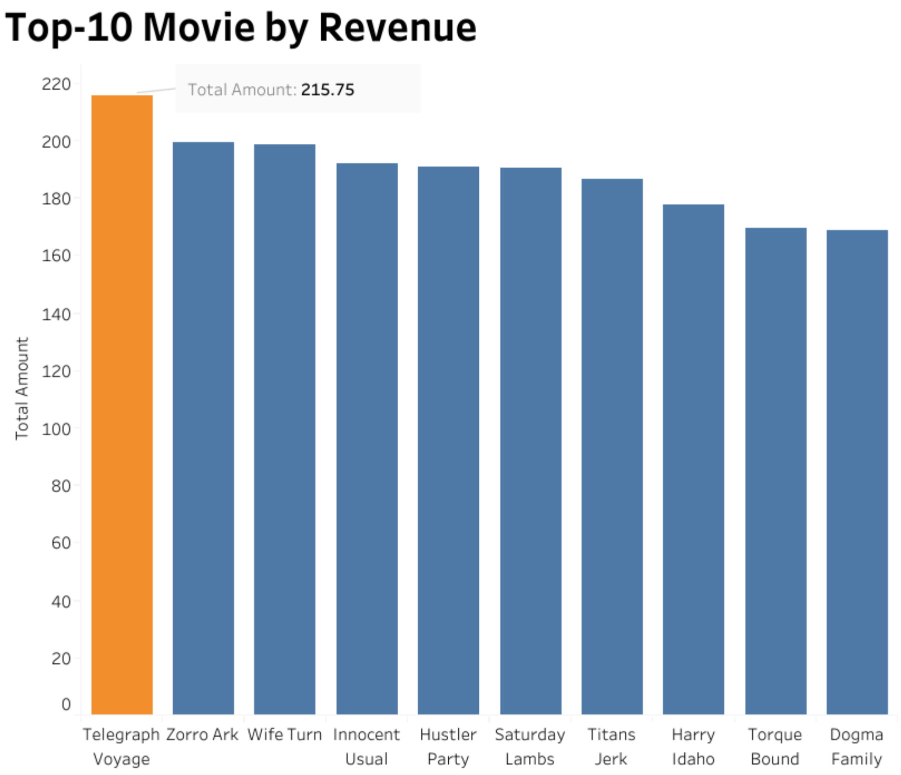
High Value Countries

Top value countries



- United States leads with the highest revenue (13,545) and the largest customer base (36 customers).
- India ranks second (9,760) with 28 customers, reflecting strong spending levels.
- Netherlands, Belarus, and Runion each appear in the top rankings due to a single high-spending customer.
- The rankings emphasize individual top spenders rather than total customer volume.

Top Movies

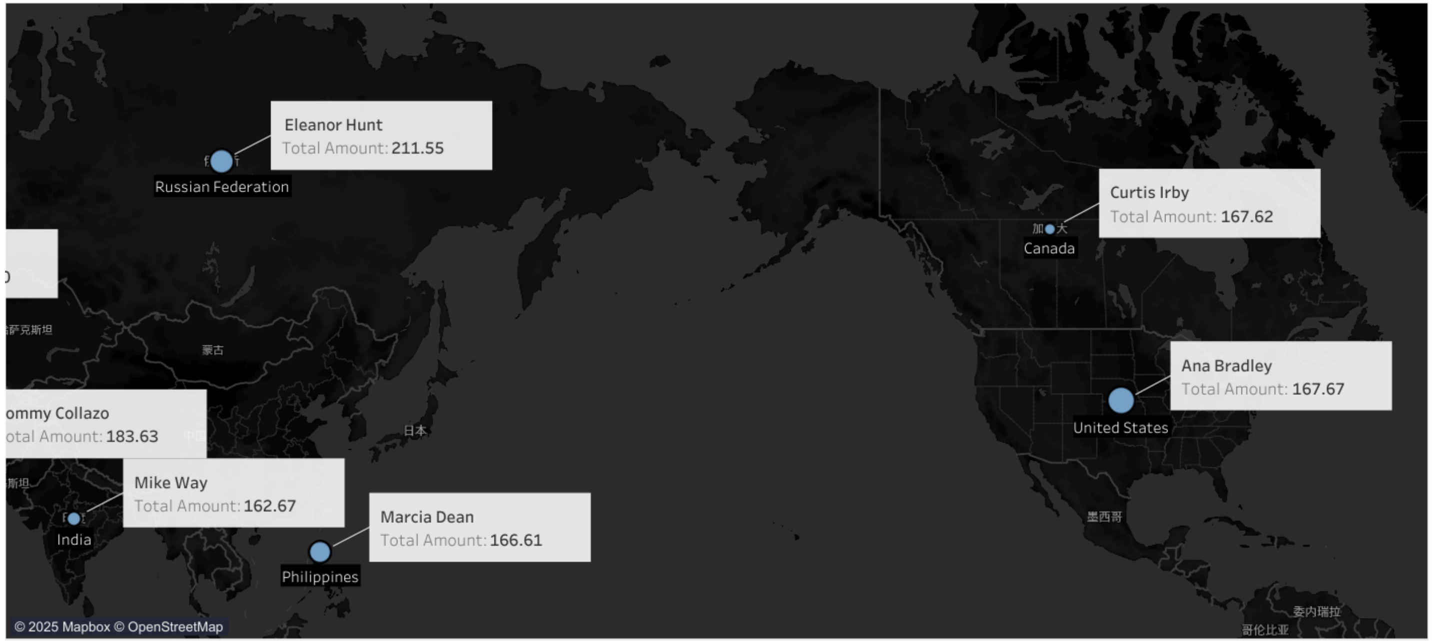


Top-10 Movie by Revenue

- Telegraph Voyage (Music genre) generated the highest revenue at 215.75.The list includes a diverse range of genres, showing broad customer preferences.
- Comedy and Drama appear multiple times, indicating strong and consistent appeal.
- Revenue for the top 10 movies is closely grouped (between 168 and 215), reflecting competitive performance across titles.
- No single genre dominates, suggesting diversified revenue sources across movie categories.

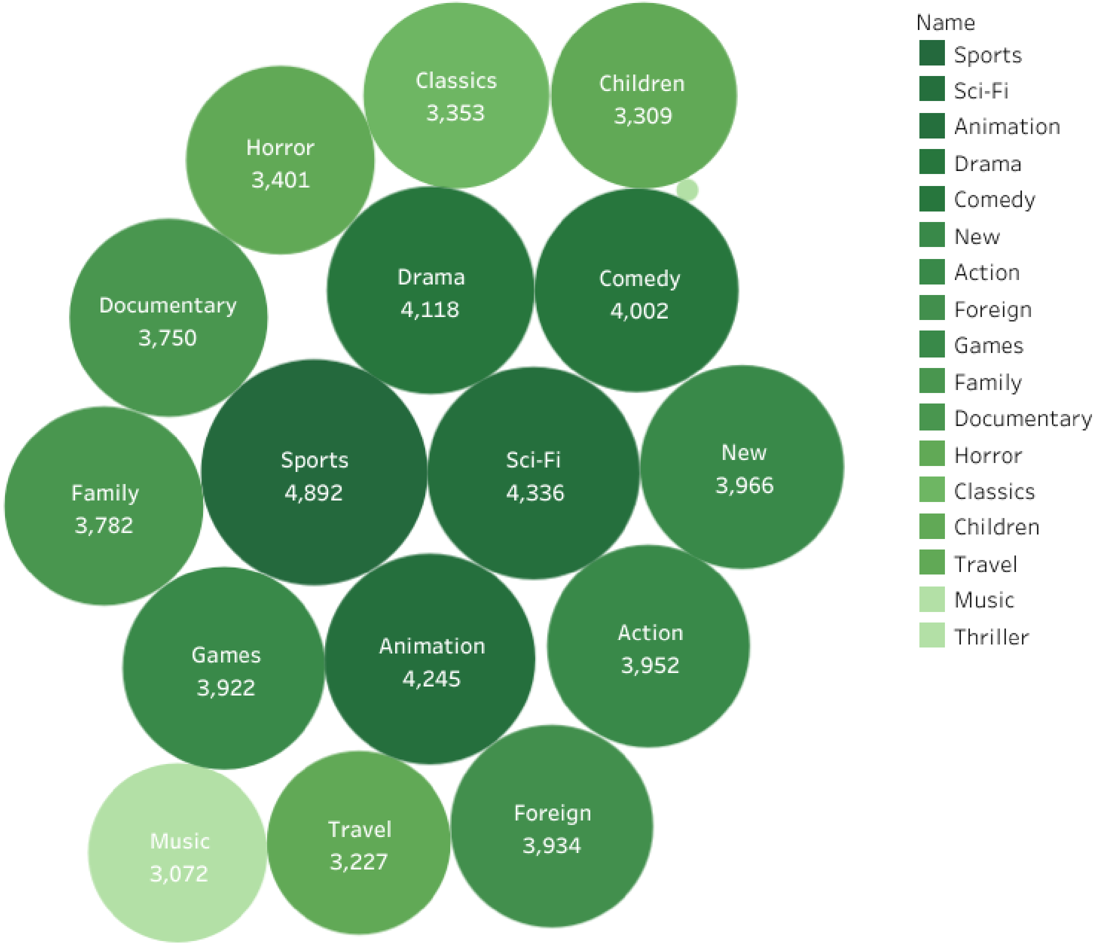
High Value Customers

Top customers



- Eleanor Hunt (Russian) is the highest spender with 211.55.
- Karl Seal (USA) follows closely with 208.58.
- Top customers come from multiple countries, reflecting a global customer base.
- Revenue among the top 10 ranges from 162 to 211, showing a strong group of high-value customers.

Revenue by Genre



Globally, the Sports, Sci-Fi, and Animation genres were the most profitable, whereas Children, Music, and Thriller generated the lowest revenues.

Recommendations

1. Focus on High-Value Markets

Prioritize marketing and customer engagement strategies in the United States and Brazil to capitalize on high spending and strong customer bases.

2. Target Niche High-Spenders

Develop tailored retention programs for individual high-value customers in countries with low customer volume but high spending power. Personalized offers, loyalty benefits, or exclusive content can increase long-term value.

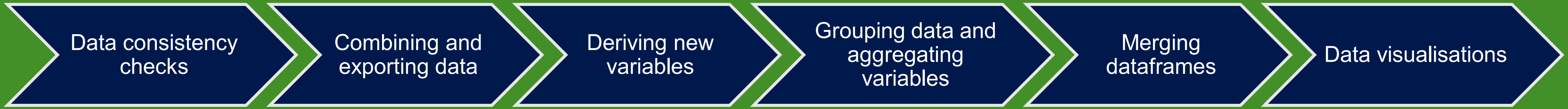
3. Invest in Proven Genres

Allocate content investment toward Comedy and Drama due to consistent performance. Use Music and Documentary genres strategically to attract niche, high-spending audiences.

🔗 Tableau Storyboard

🔗 GitHub Repository

Project 3: Instacart Grocery Market Analysis



Data consistency checks

Combining and exporting data

Deriving new variables

Grouping data and aggregating variables

Merging dataframes

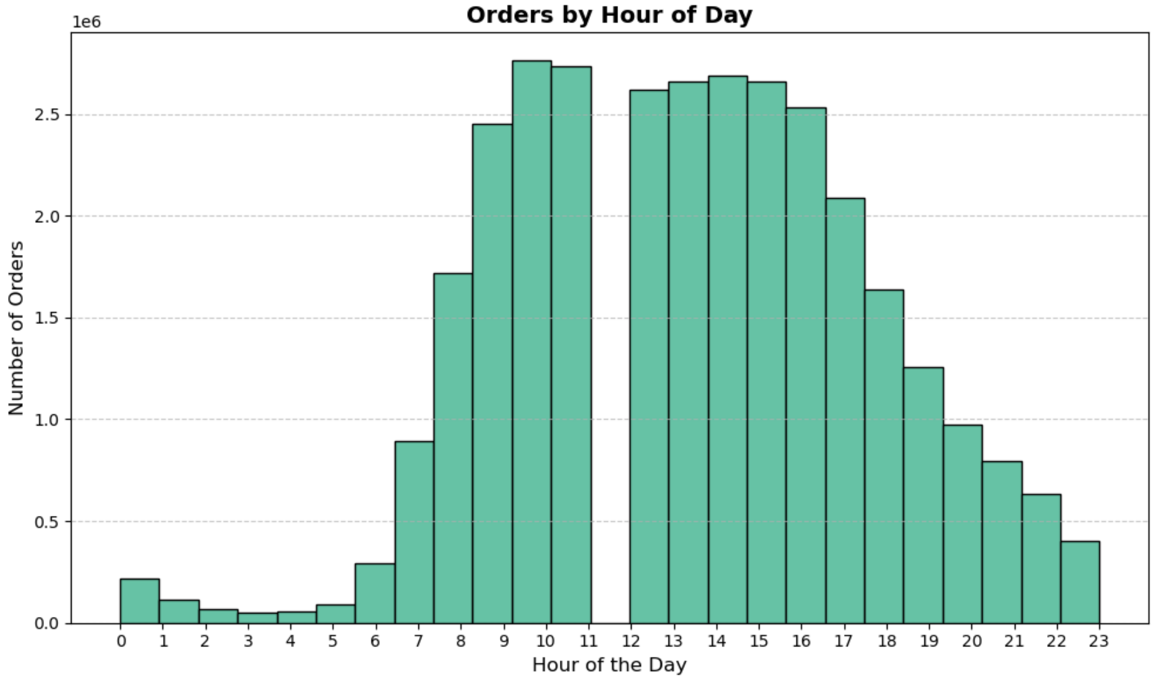
Data visualisations

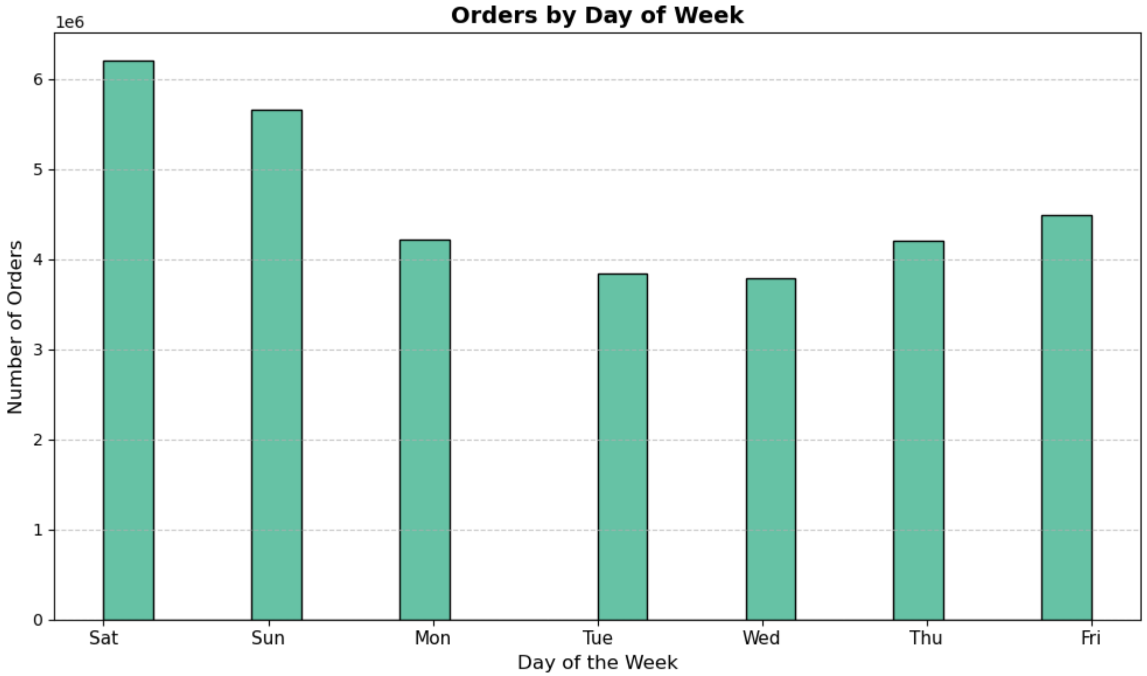
Tools:

Introduction

- **Context**
  - Instacart, an online grocery store that operates through an app. Instacart already has very good sales, but they want to uncover more information about their sales patterns.
- **Objectives**
  - Identify busiest days and hours, overall spending patterns, and popular product categories.
  - Uncover key patterns in Instacart customers’ purchasing behaviour to support data-driven marketing and sales decisions.
  - Segment customers by loyalty, region, age, income, and family status to understand diverse purchasing habits.
- **Data**
  - Accessed from Kaggle.

Order Numbers by Hour and Day

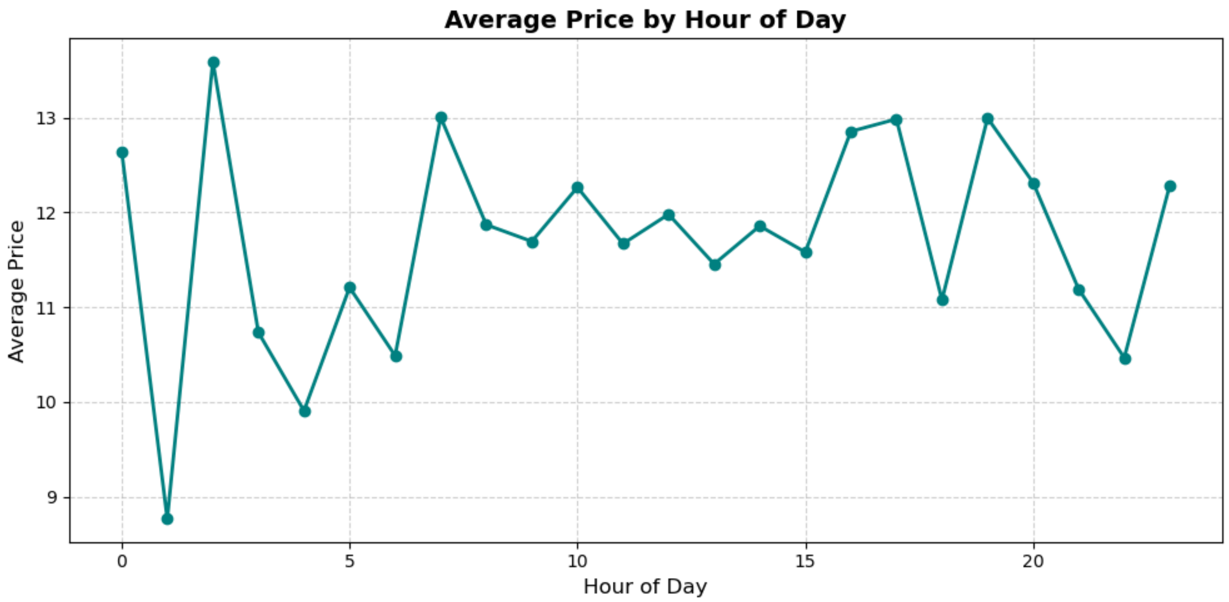


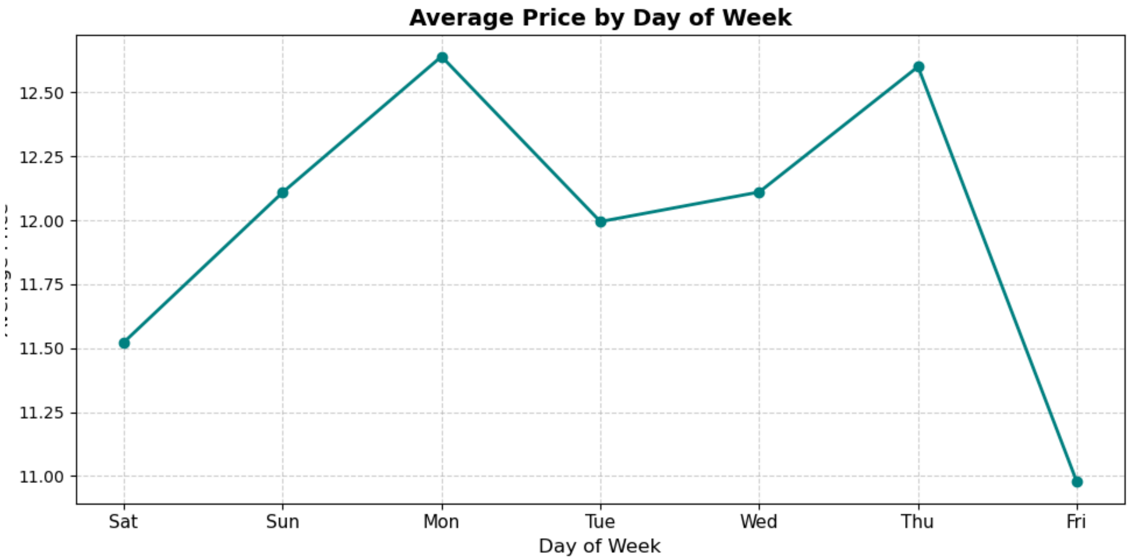


- Most orders were placed between 9:00 a.m. and 5:00 p.m., whereas order activity was minimal between midnight and 6:00 a.m.

- Orders peaked on Saturday and Sunday, while Tuesday and Wednesday were the quietest days.

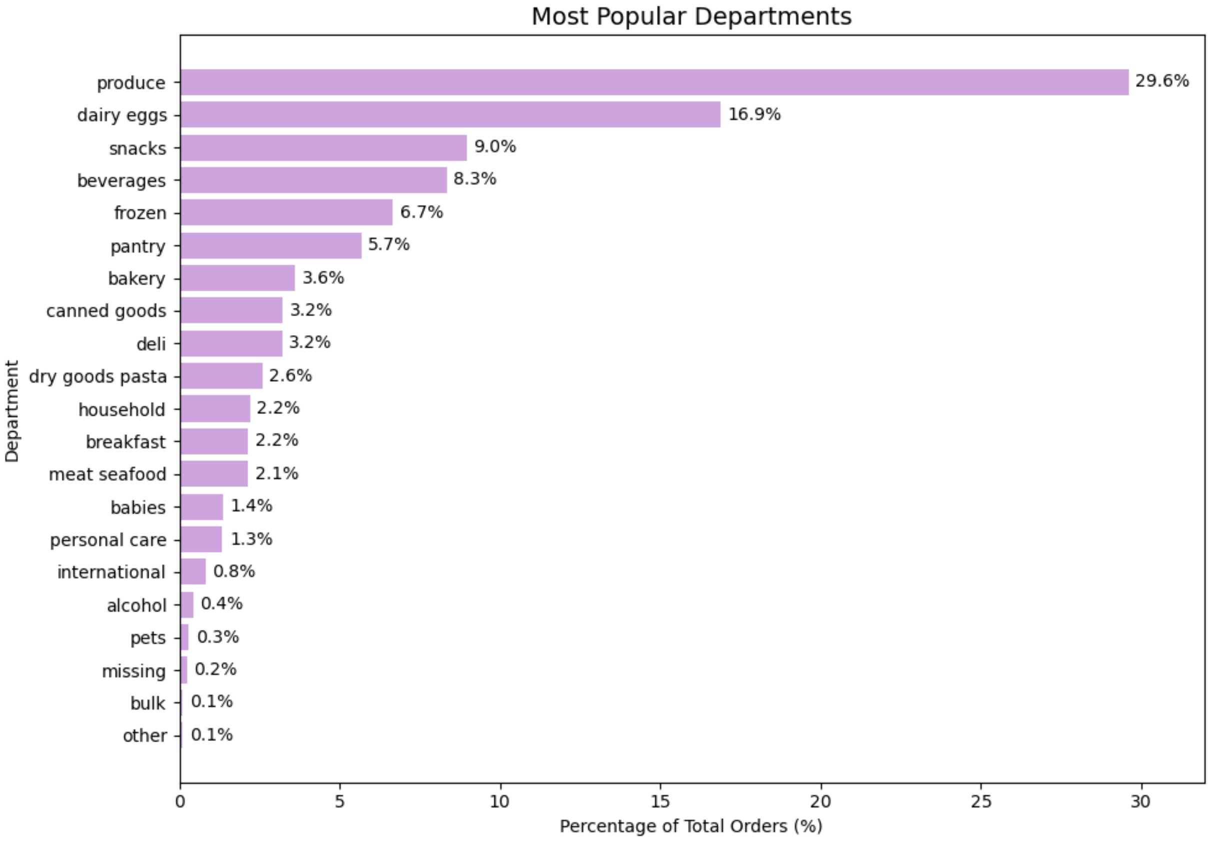
Spending by Hour and Day


Average Price by Hour of Day


Average Price by Day of Week

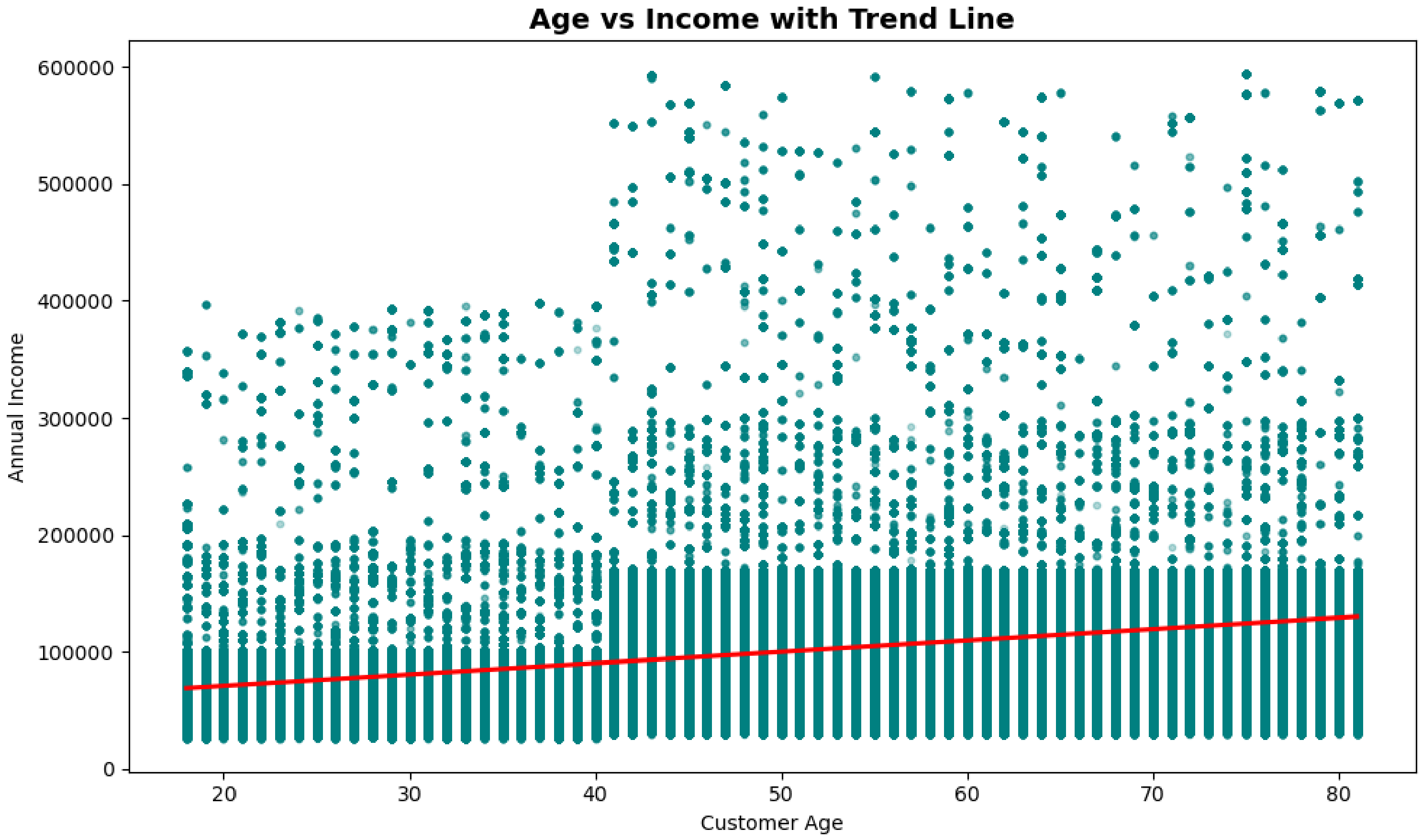
- Average spending peaked at approximately 3:00 a.m. and reached their lowest point around 1:00 a.m..
- Average spending peaked on Monday and reached their lowest point on Friday.

Order Numbers by Department



Produce, dairy eggs, snacks, beverages, frozen, and pantry were the most popular departments.

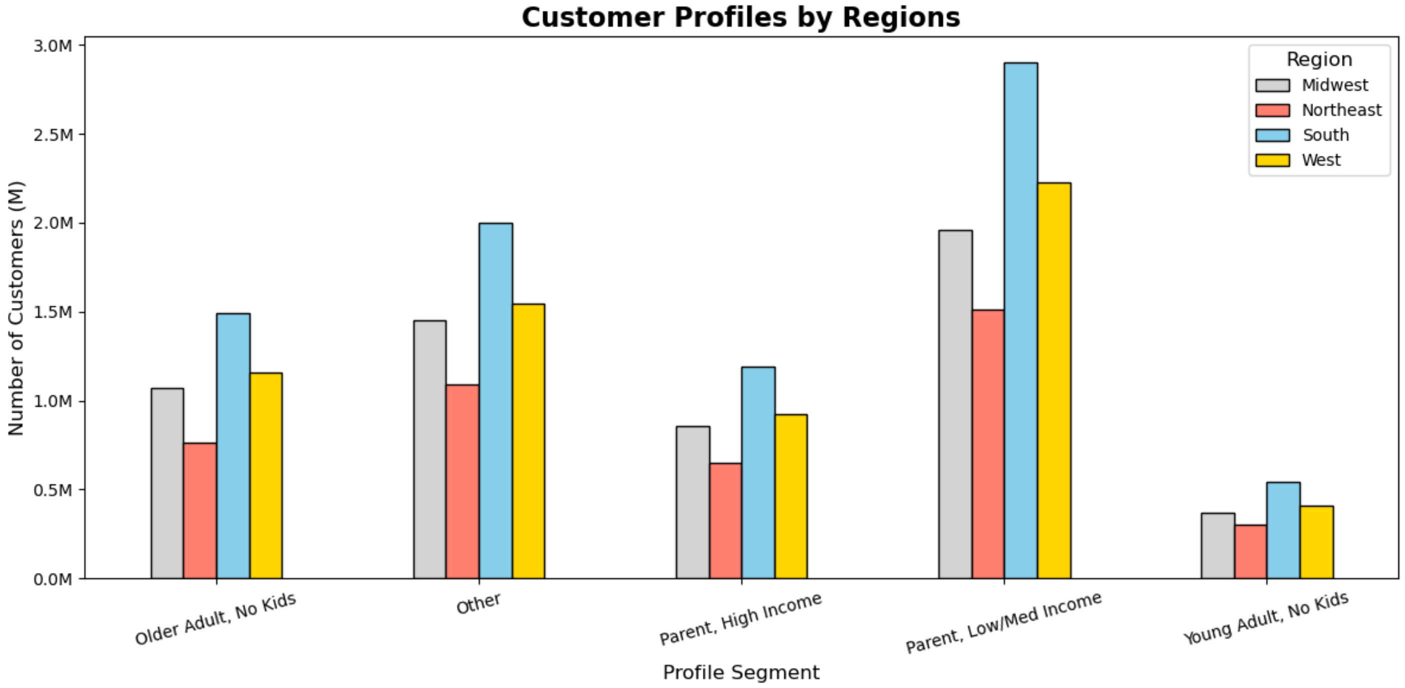
Income vs Age



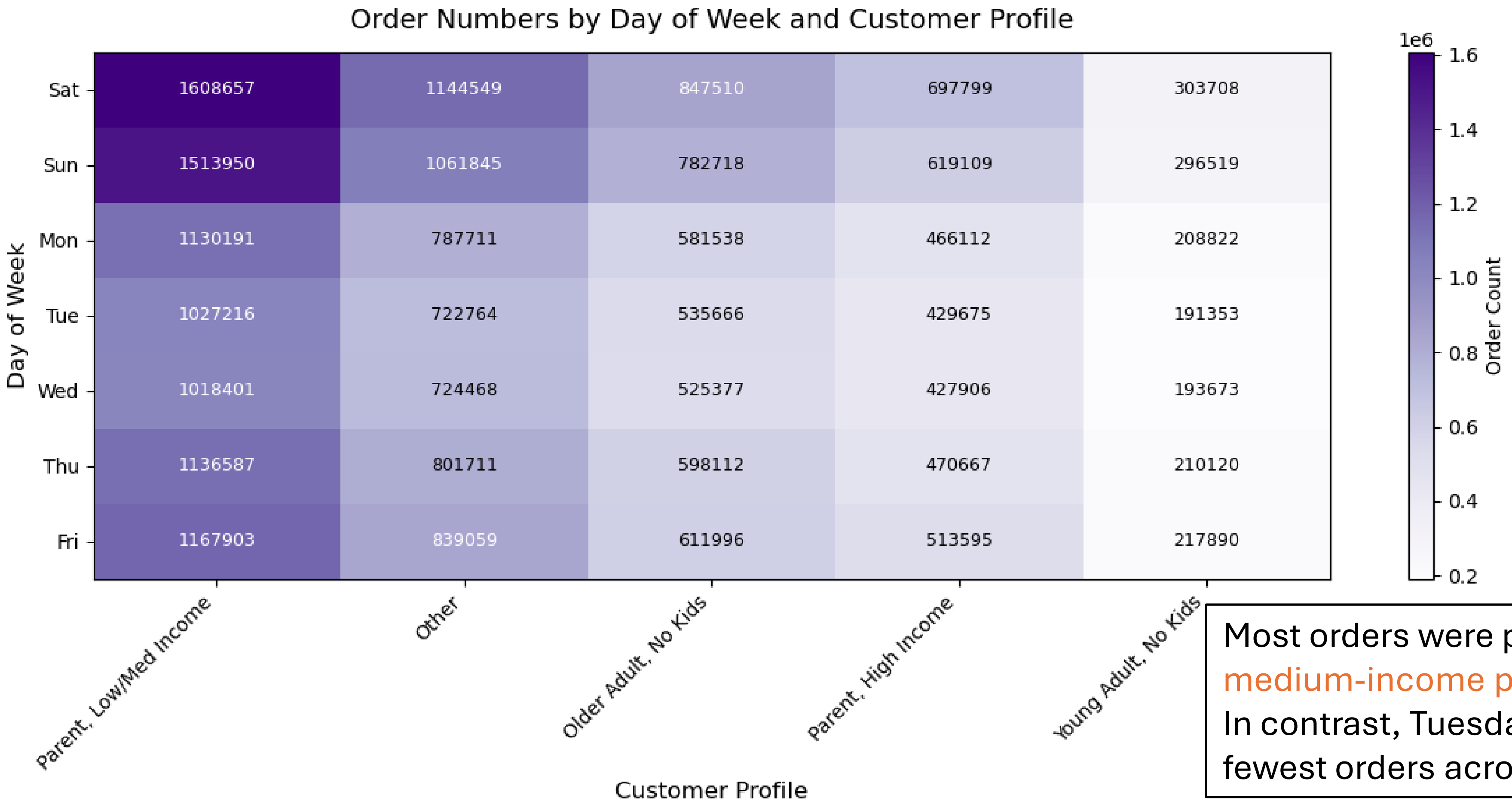
Income grew gradually with age.

Customer Profiles by Region

- Most customers were located in the South, followed by the West, Midwest, and Northeast.
- Low- and medium-income parents were the most frequent shoppers, followed by parents with one child.



Order Numbers by Day and Customer Profile



Order Numbers by Day of Week and Customer Profile

Day of Week	Parent, Low/Med Income	Other	Older Adult, No Kids	Parent, High Income	Young Adult, No Kids
Sat	1608657	1144549	847510	697799	303708
Sun	1513950	1061845	782718	619109	296519
Mon	1130191	787711	581538	466112	208822
Tue	1027216	722764	535666	429675	191353
Wed	1018401	724468	525377	427906	193673
Thu	1136587	801711	598112	470667	210120
Fri	1167903	839059	611996	513595	217890

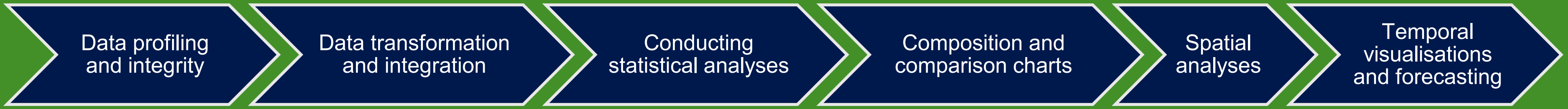
GitHub Repository

Most orders were placed at weekends by low- and medium-income parents.
In contrast, Tuesdays and Wednesdays saw the fewest orders across all customer profiles.

Conclusions

- **Key Findings**
  - Orders peak on weekends and drop midweek (Tues & Wed).
  - Spending fluctuates between 3 PM–10 PM, with Fridays showing the lowest activity.
  - Produce, dairy, snacks, and pantry staples are the top-selling categories.
  - Most customers are regular or new, mainly family-oriented, low- to medium-income parents.
- **Recommendations**
  - Schedule ads and promotions on slower days (Tues/Wed) and evenings (3–10 PM).
  - Maintain strong inventory for high-demand staples.
  - Introduce loyalty programs to retain and grow the customer base.
  - Use family-focused, value-driven marketing and expand outreach in underrepresented regions.

Project 4: Preparing for the Flu Season



Data profiling and integrity

Data transformation and integration

Conducting statistical analyses

Composition and comparison charts

Spatial analyses

Temporal visualisations and forecasting

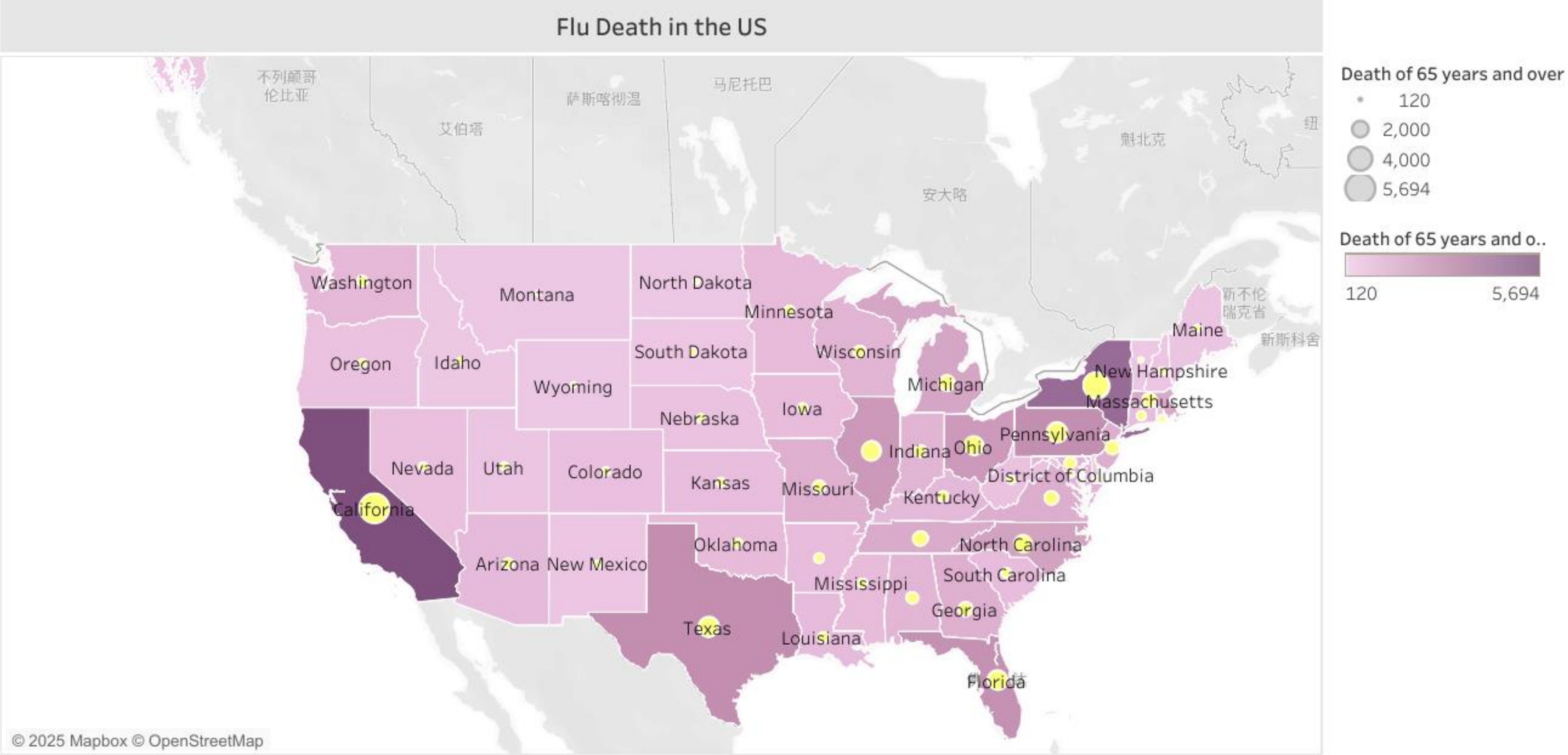
Tools:

Introduction

- **Objectives**
  - Identify the vulnerable age groups that are prone to flu deaths;
  - Find out the states where vulnerable age groups are located;
  - Determine the seasonal patterns of flu and see if that varies by state.

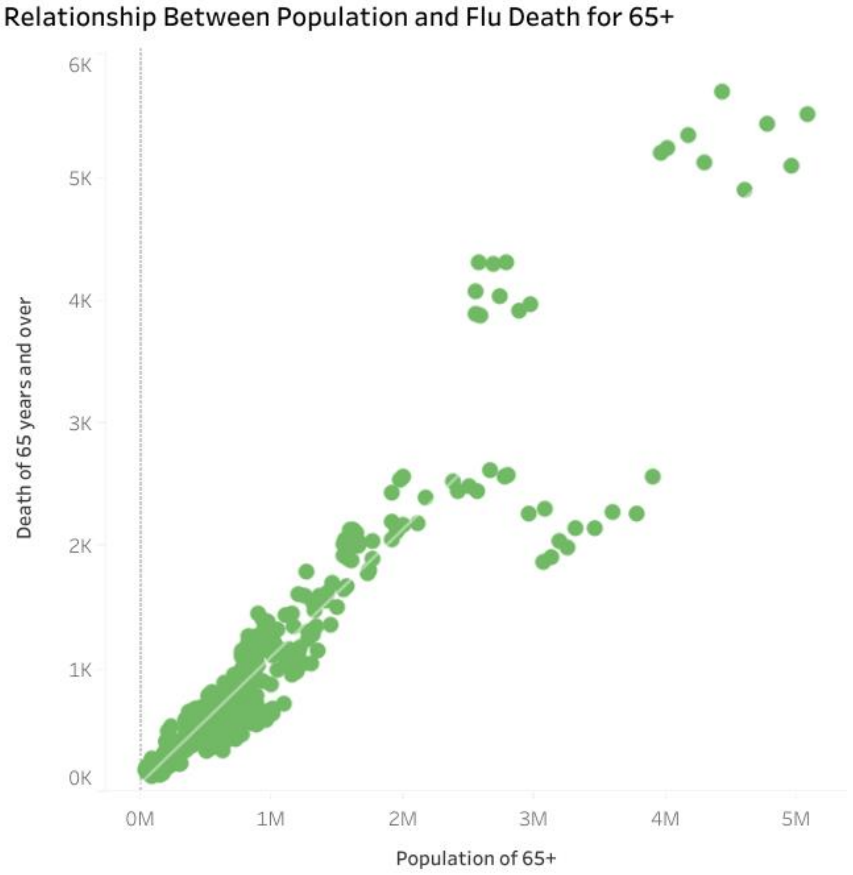
- **Data Overview**
  - US Census Data (2009 – 2017) from US Census Bureau
  - Influenza Death Data (2009 – 2017) from Centres for Disease Control and Prevention (CDC)

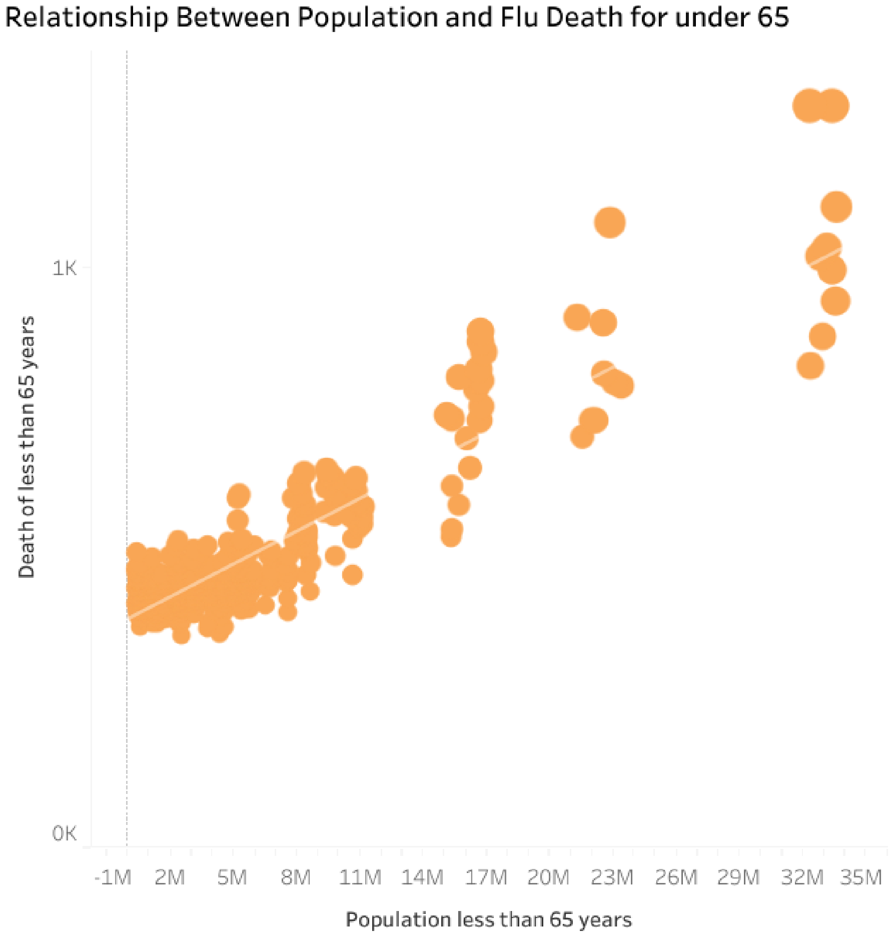
Distribution of Flu Death in US


Flu Death in the US

© 2025 Mapbox © OpenStreetMap

The Relationship between Population and Flu Deaths

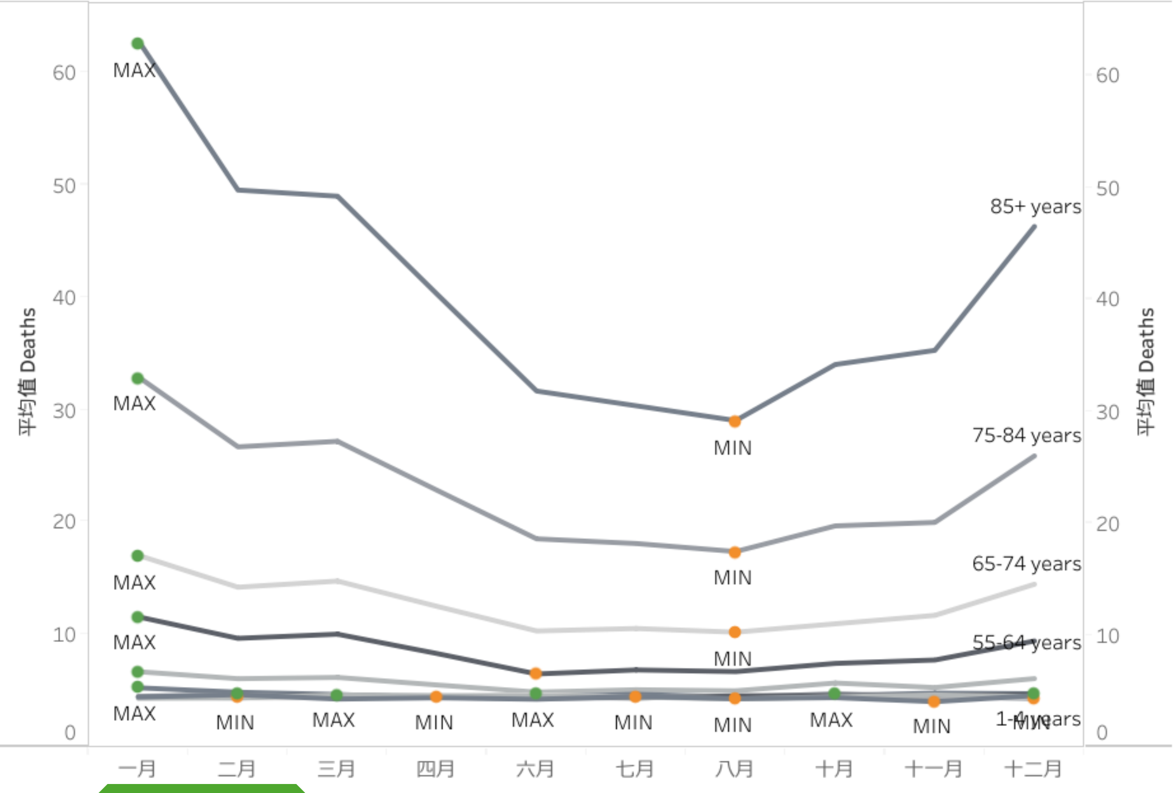




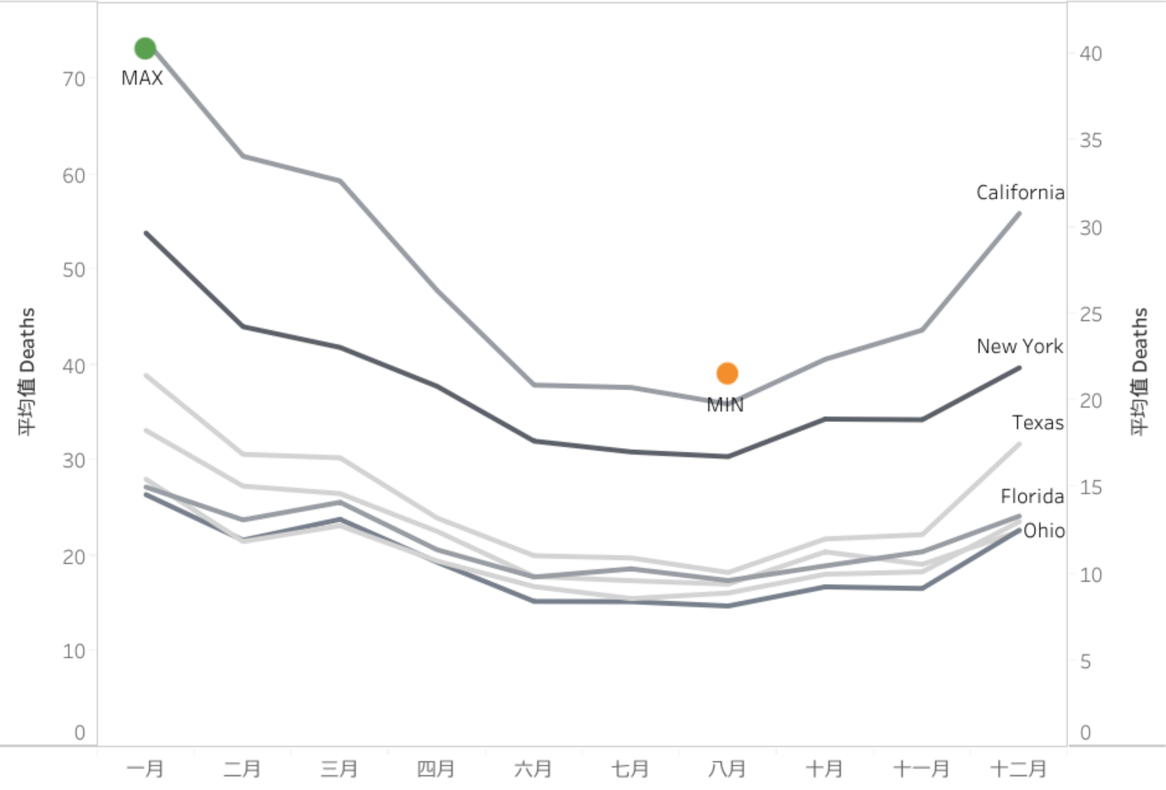
There is a strong relationship between population size and flu deaths. The highest-risk areas are those with a large proportion of residents aged 65 and older.

When is High Season?

Average Flu Death by Month and Age Group



Average Flu Death by Month in High-Risk States



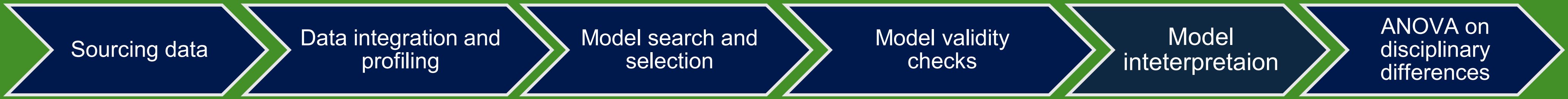
December, January, February, and March were the peak seasons across states.

Conclusions

- **Findings**
  - 65+ are much more vulnerable to flu deaths than younger populations.
  - California, Florida , New York , Texas , Pennsylvania , Illinois, and Ohio had the largest 65+ population on average.
  - Winter and early spring are the peak seasons for flu deaths.
- **Recommendations**
  - Targeted vaccination reminders particularly to individuals aged 65+
  - Extra staffing and resource allocation in high-risk states
  - Seasonal staffing planning beginning from October

Tableau Storyboard

Project 5: Structured Topic Modelling of China’s TVET Assessment Policy

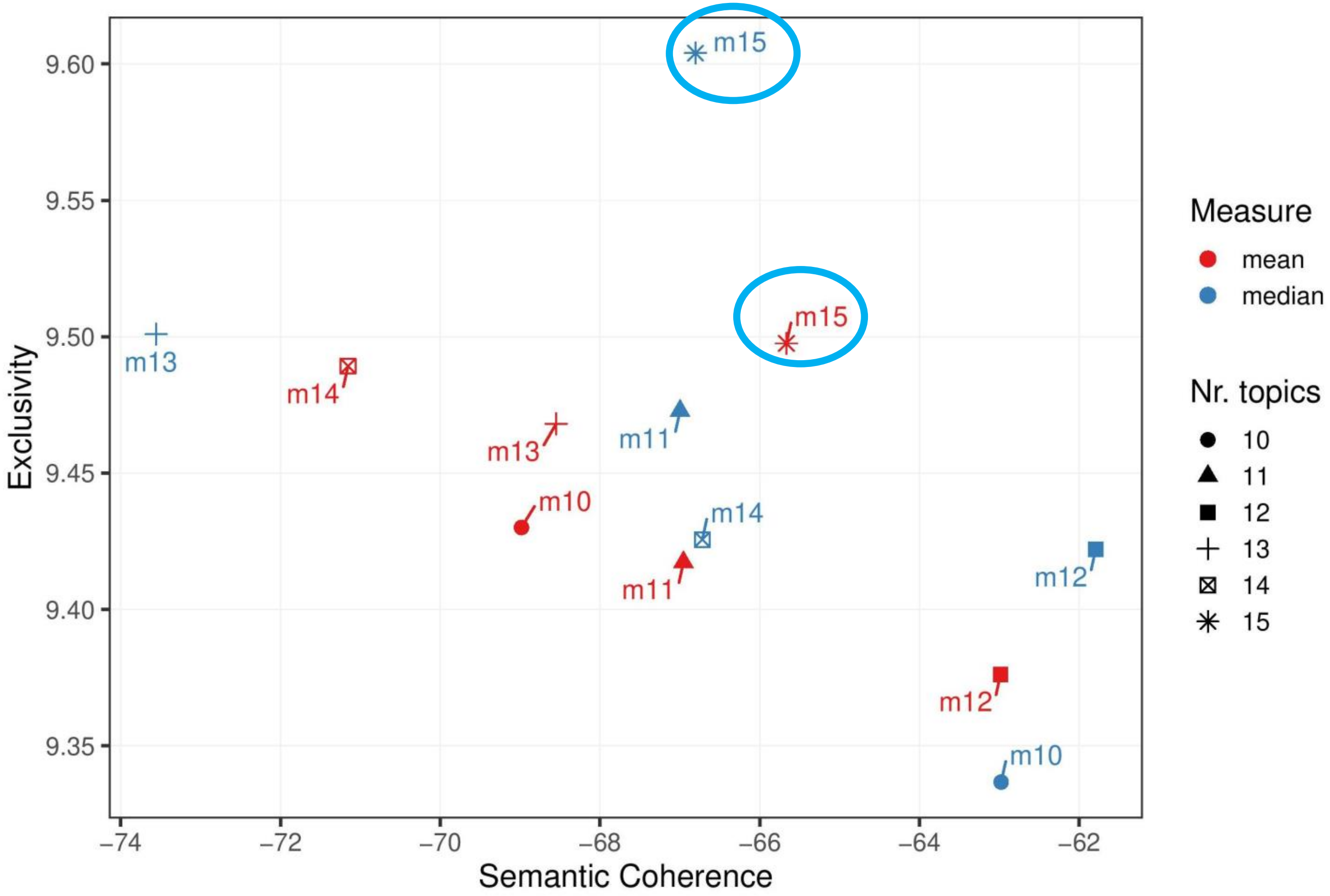

Sourcing data
Data integration and profiling
Model search and selection
Model validity checks
Model inteterpretaion
ANOVA on disciplinary differences

Tools:

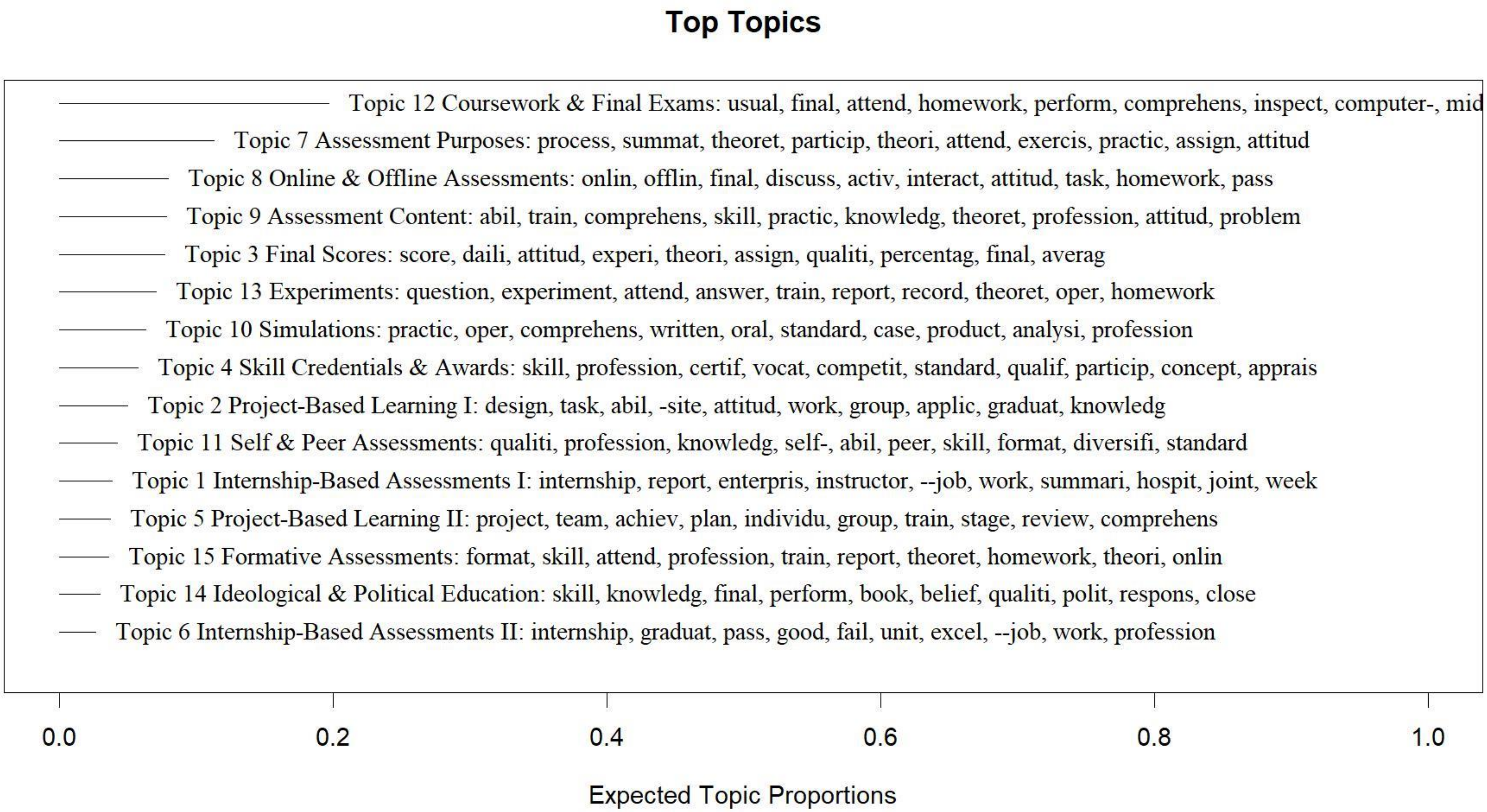
Introduction

- **Objectives**
  - This project demonstrates how Structural Topic Modelling (STM) can serve as a scalable alternative by modeling topics that vary with document-level metadata (e.g., discipline). We show end-to-end preprocessing, model search, quality evaluation, interpretation, and validation on a corpus of 166 policy documents.
- **Data**
  - Talent Development Plans from China’s tertiary vocational institutions

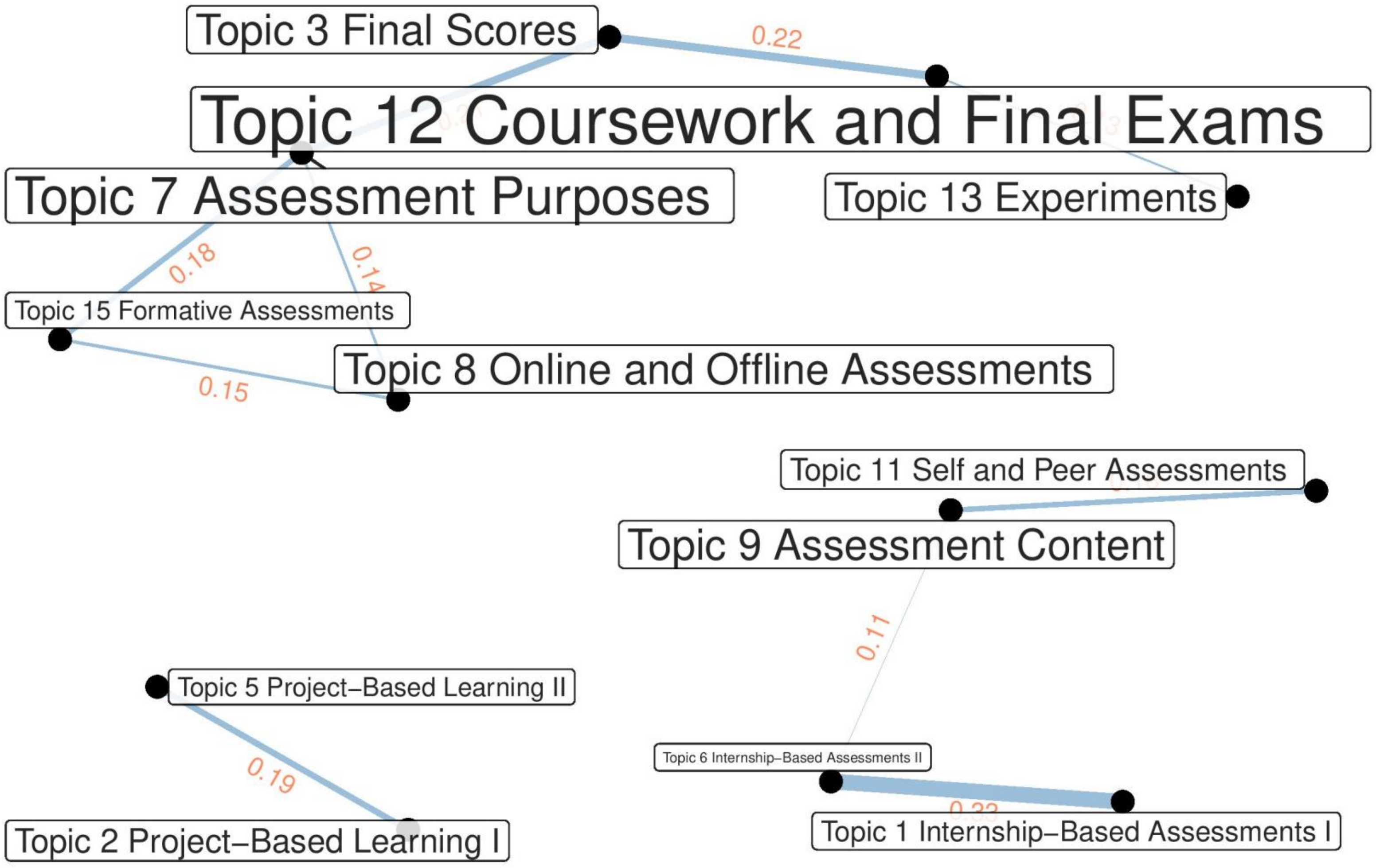
Model Search



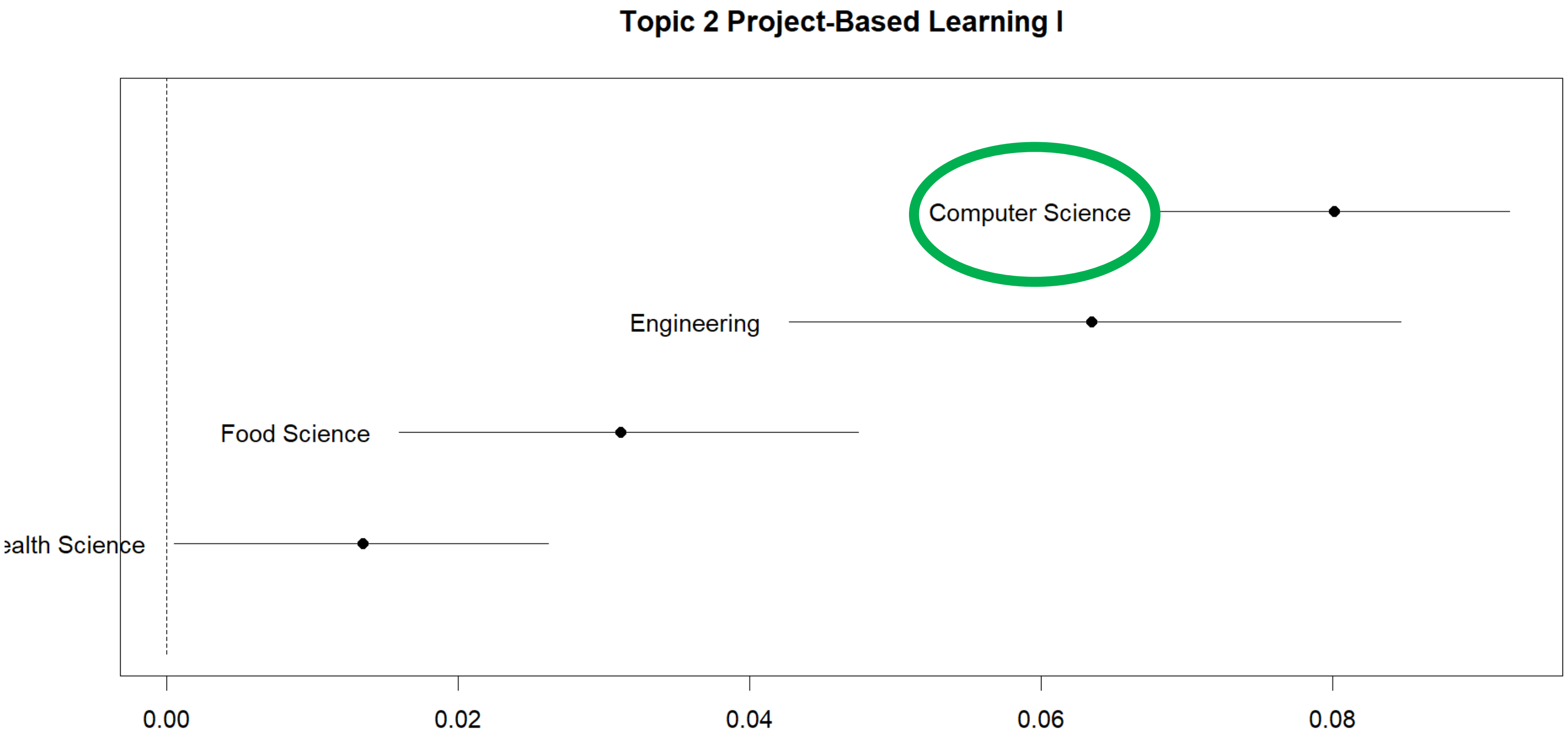
Topic Labels, Words and Proportions



Topic Correlation


Topic 3 Final Scores
0.22
Topic 12 Coursework and Final Exams
Topic 7 Assessment Purposes
Topic 13 Experiments
0.18
0.14
Topic 15 Formative Assessments
Topic 8 Online and Offline Assessments
0.15
Topic 11 Self and Peer Assessments
Topic 9 Assessment Content
0.11
Topic 5 Project–Based Learning II
0.19
Topic 6 Internship–Based Assessments II
0.33
Topic 2 Project–Based Learning I
Topic 1 Internship–Based Assessments I

Topic 2 Document-Topic Proportion by Discipline

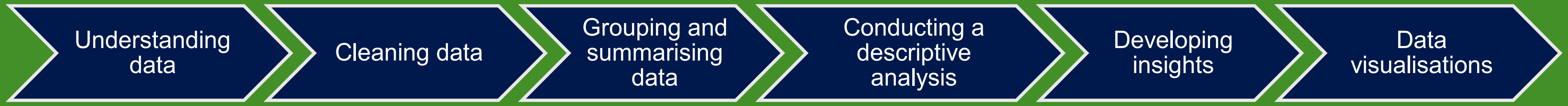


Conclusions

- **Key Findings**
  - Common assessment types: Coursework, Exams, Projects, and Self/Peer Assessments.
  - Discipline differences observed (e.g., Project Design in Computer Science).
  - Formative assessments often overlap with summative grading, reducing feedback value.
- **Recommendations**
  - Diversify assessments: Blend exams, coursework, and applied tasks.
  - Tailor to disciplines: Match formats with subject-specific practices.
  - Clarify assessment purpose: Separate formative from summative to enhance learning impact.

GitHub Repository

Project 6: GameCo Global Video Game Sales



Understanding data

Cleaning data

Grouping and summarising data

Conducting a descriptive analysis

Developing insights

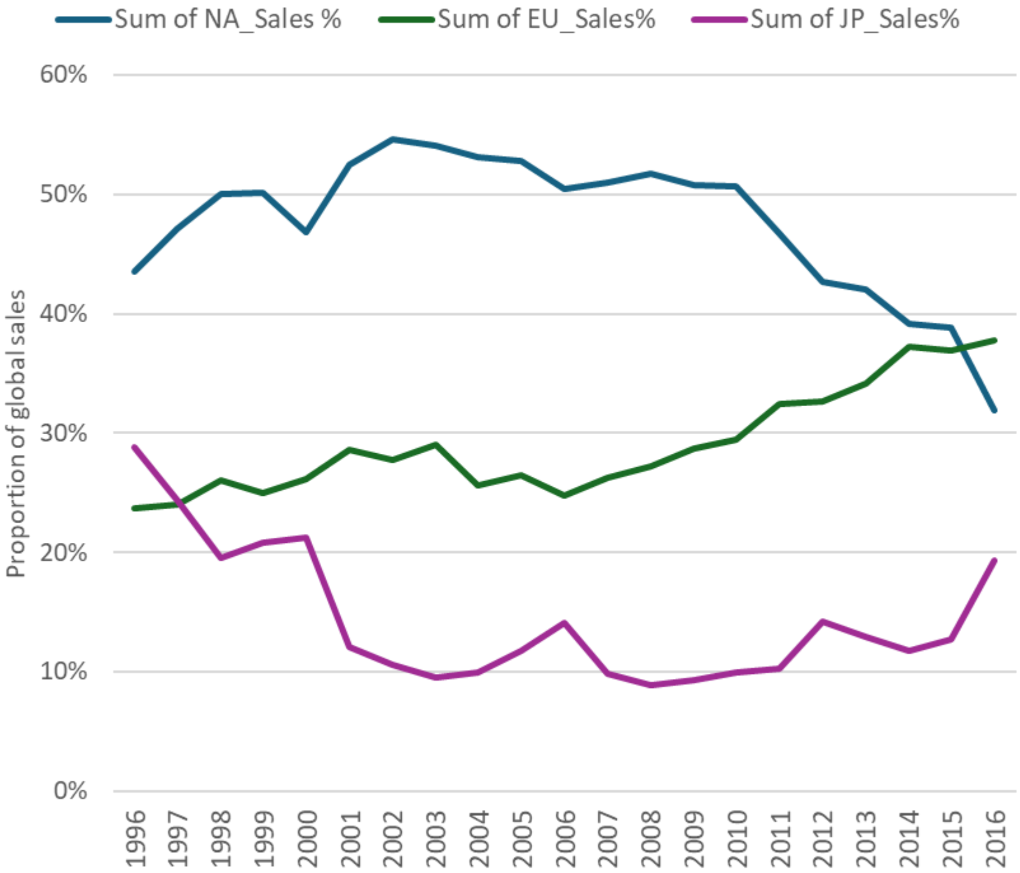
Data visualisations

Tools:

Introduction

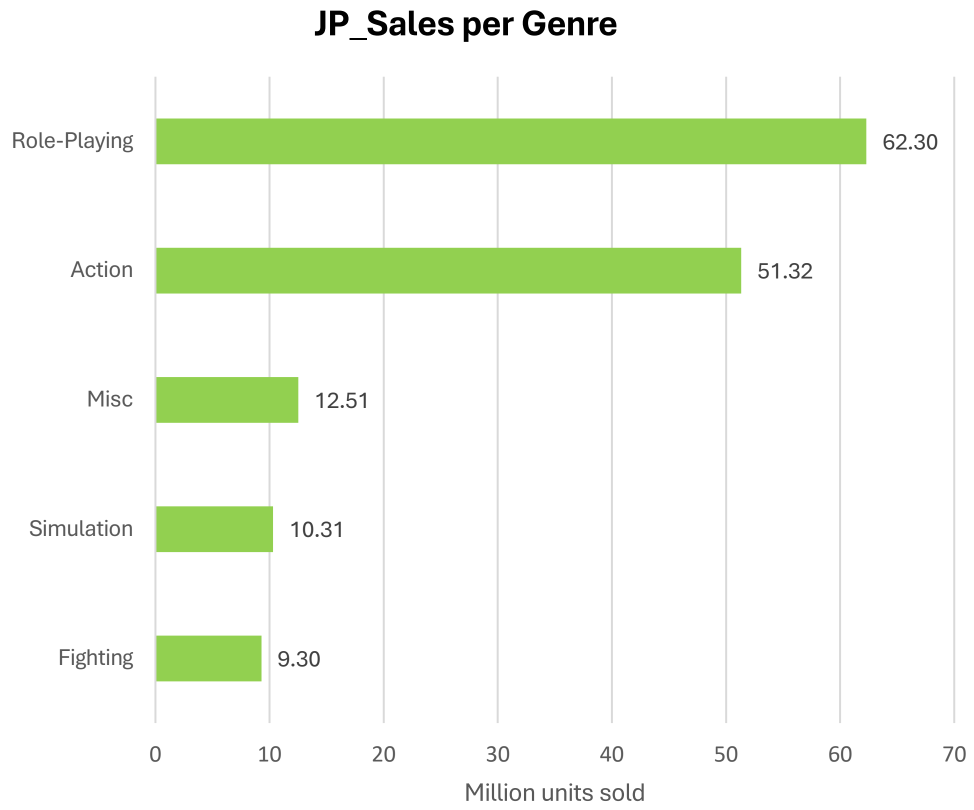
- **Context:**
  - A new video game company, GameCo, which wants to use data to inform the development of new games
- **Objectives:**
  - To find any difference in sales figures between geographic regions over time
  - To identify the publishers that are likely be the main competitors in certain markets
  - To understand the types of games that are more popular than others
  - To find games that decreased or increased in popularity over time
- **Data Set:**
  - The data set was drawn from the website VGChartz.

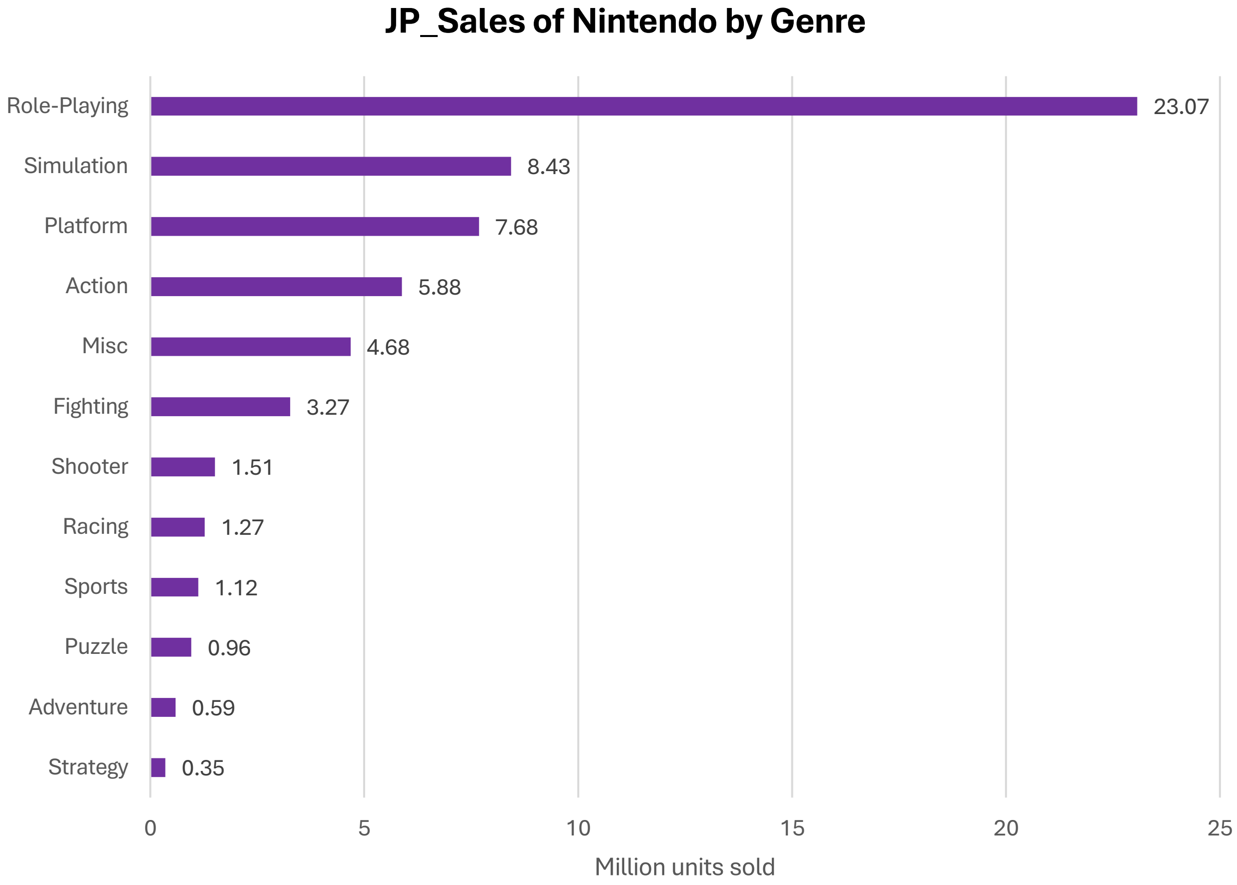
Sales between 1996 and 2016



- North America > Europe > Japan

Genre: Nintendo in JP



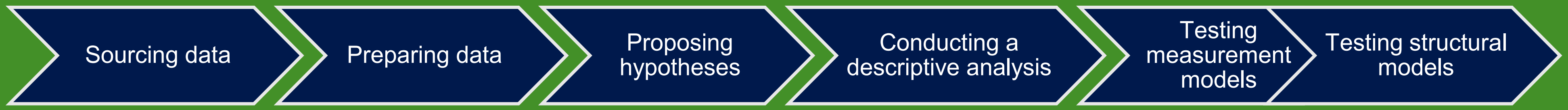


Conclusions

- **Key Findings**
  - The JP market was promising than NA and EU, where the global sales percentage showed limited growth.
  - There is still much market space for role-playing and action in addition to Nintendo’s share.

- **Recommendations**
  - Focus on the JP market, which shows strong growth potential and is less saturated by dominant competitors.
  - Specialise in role-playing and action games, genres with room for growth.

Project 7:
Peer Misconduct across School and Workplace Settings



Sourcing data

Preparing data

Proposing hypotheses

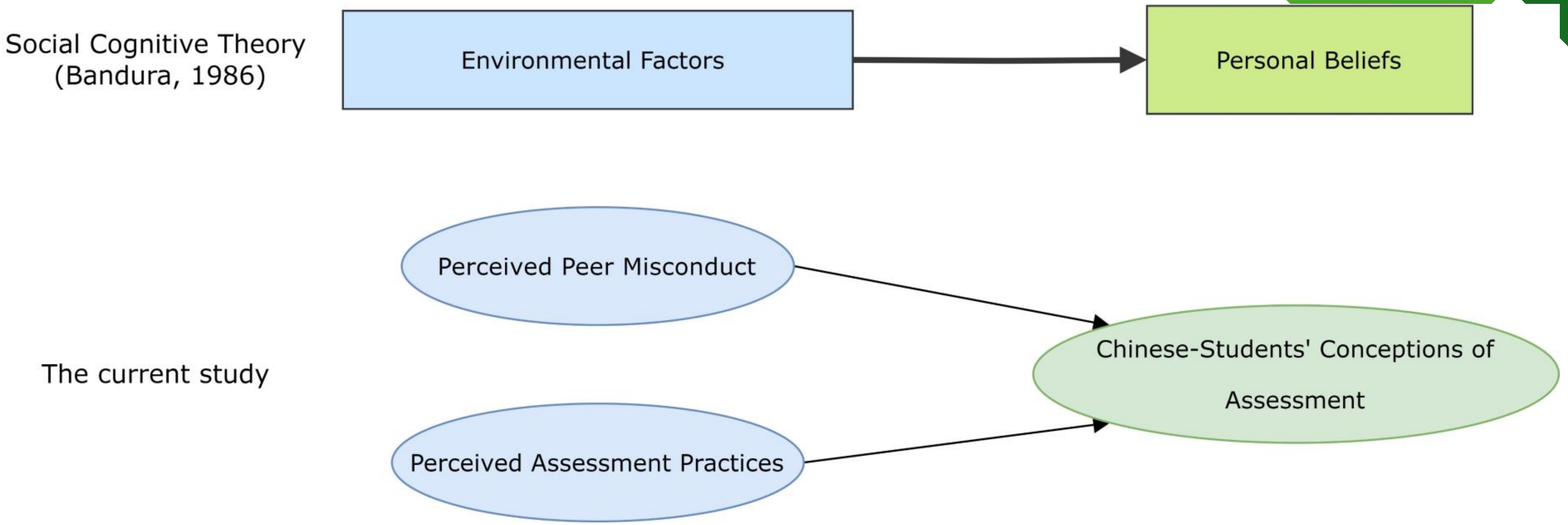
Conducting a descriptive analysis

Testing measurement models

Testing structural models

Tools:

Hypothesis



Social Cognitive Theory (Bandura, 1986)

Environmental Factors

Personal Beliefs

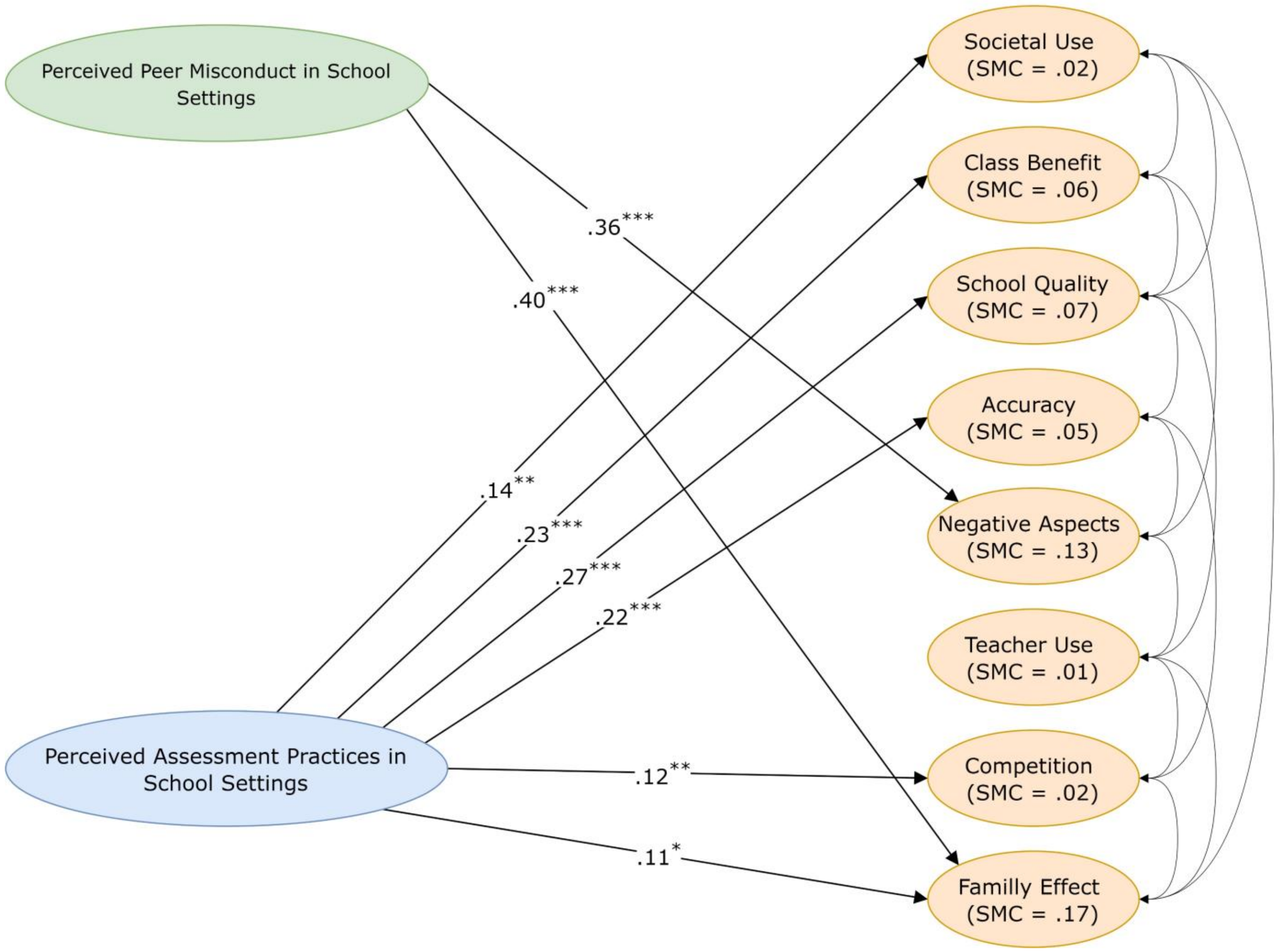
The current study

Perceived Peer Misconduct

Perceived Assessment Practices

Chinese-Students' Conceptions of Assessment

Structural Models – In School Settings



Perceived Peer Misconduct in School Settings

Perceived Assessment Practices in School Settings

Societal Use (SMC = .02)

Class Benefit (SMC = .06)

School Quality (SMC = .07)

Accuracy (SMC = .05)

Negative Aspects (SMC = .13)

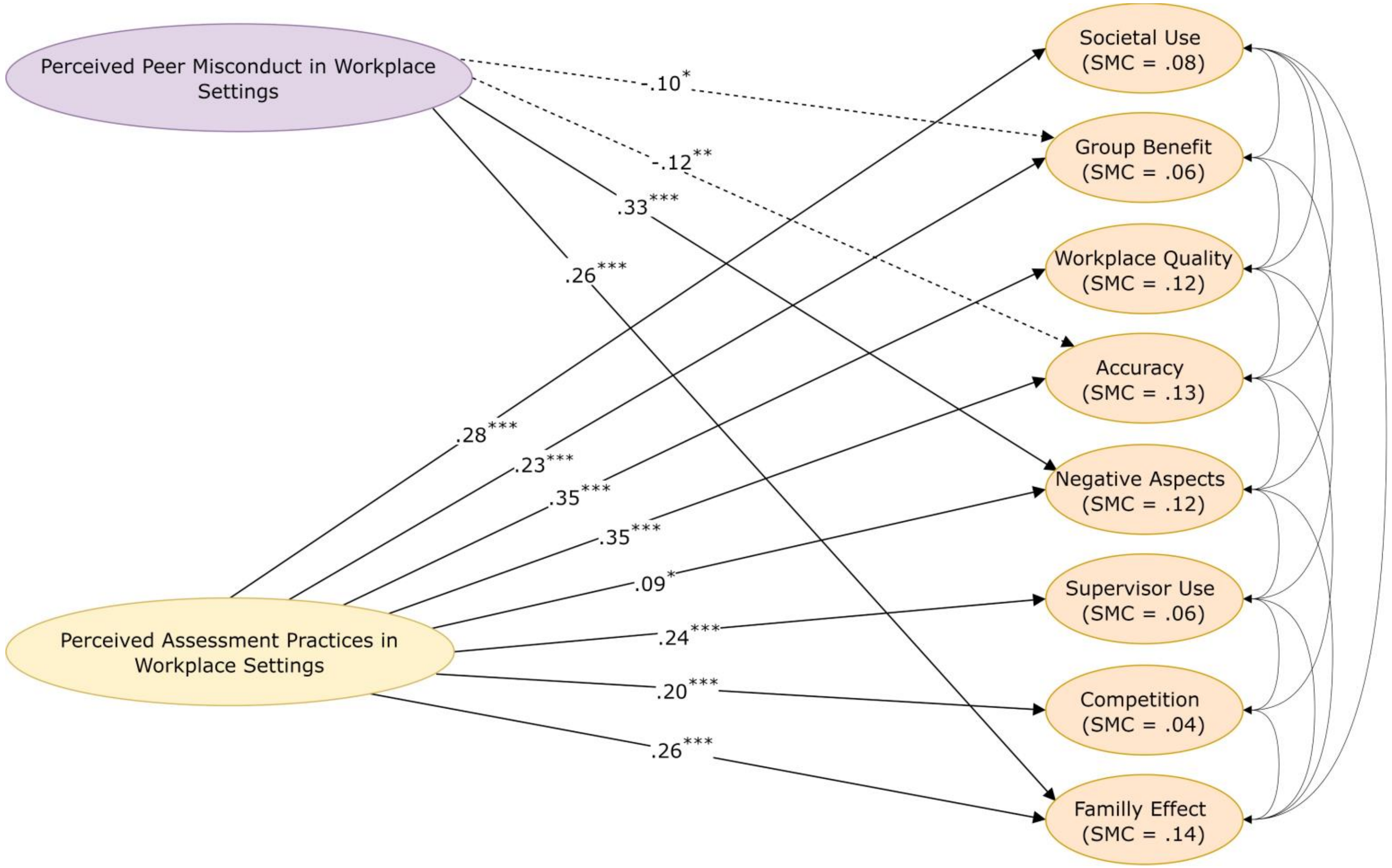
Teacher Use (SMC = .01)

Competition (SMC = .02)

Familly Effect (SMC = .17)

Model	*n*	MLRχ^2(*df*)	MLRχ^2/*df* (*p*)	SRMR	RMSEA$_r$(90%CI)	CFI$_r$	Gamma hat	AIC
1	563	1361.51(851)	**1.60(*p* = .206)**	**.042**	**.035(.031, .038)**	**.964**	**.95**	43348.75

Structural Models – In Workplace Settings



Perceived Peer Misconduct in Workplace Settings

Perceived Assessment Practices in Workplace Settings

Societal Use (SMC = .08)

Group Benefit (SMC = .06)

Workplace Quality (SMC = .12)

Accuracy (SMC = .13)

Negative Aspects (SMC = .12)

Supervisor Use (SMC = .06)

Competition (SMC = .04)

Familly Effect (SMC = .14)

Model	*n*	MLRχ^2(*df*)	MLRχ^2/*df* (*p*)	SRMR	RMSEA$_r$(90%CI)	CFI$_r$	Gamma hat	AIC
1	567	1664.55 (986)	**1.69 (*p* = .194)**	**.040**	**.037 (.034, .041)**	**.961**	**.95**	48335.89

Conclusions

- **Key Findings**
  - Students experienced different assessment practices across school and workplace settings, but their underlying beliefs about assessment remained consistent.
  - Perceptions of assessment practices were closely linked to students' beliefs about the purpose and value of assessment in both settings.
  - Observing peer misconduct reinforced negative attitudes toward assessment (e.g., devaluation of assessment grades).
  - Observing peer misconduct also reinforced the belief that assessment exists to fulfil family expectations.
- **Recommendations**
  - Establish a structured mentorship system involving school instructors and workplace supervisors to support student interns' transition from education to professional society.
  - Implement an orientation across both school and workplace ethics that includes discussions, case studies, and simulated exercises to raise ethical awareness.
  - Develop clear regulations and policies to address serious violations, including possible disqualification from graduation for severe misconduct.

Thank you!

Yan Zhang

✉ sancheungedu@outlook.com

☎ 022-492-8823

GitHub

Tableau