Project Summary – Fraud Data Analysis & Segmentation

Dataset: Canadian Anti-Fraud Centre (CAFC), ~329k reports (2021–2025). **Goal:** Explore large-scale fraud and cybercrime reports, clean and prepare the dataset, and apply statistical/ML methods to identify meaningful victim and fraud profiles.

Note: This document is the main analysis notebook (Colab). An additional Exploratory Data Analysis (EDA) Q&A notebook is attached as a supplement, showing descriptive fraud breakdowns by theme, province, age, gender, and solicitation method.

1. Data Preparation

- Downloaded raw CAFC dataset (70MB+).
- Cleaned mixed French/English duplicates, standardized currency, parsed dates.
- Extracted and imputed victim age midpoints (30% missing), using stratified random imputation by fraud category.
- Constructed numeric + categorical features: dollar losses, number of victims, solicitation methods, fraud themes, gender, complaint type, and province.

2. Exploratory Data Analysis

- **Univariate:** Distribution of losses, age, gender, and complaint type.
- **Bivariate:** Fraud theme × gender, solicitation method × loss, age × report type.
- **Geography:** Province-level fraud intensity per 100k population.
- **Temporal:** Monthly reports with 12-month rolling averages, quarterly loss trends.
- **Highlights:**
    - Investments and Romance scams caused the largest total losses (>C$1.4B combined).
    - Per-capita risk highest in Quebec, Manitoba, and Yukon.
    - Romance fraud peaked in ages 60–69, with women showing higher aggregate losses.

3. Segmentation & Clustering

- Preprocessed data via one-hot encoding, scaling, and dimensionality reduction (PCA → UMAP).
- Applied **K-Means** and **HDBSCAN** for unsupervised segmentation.
  - K-Means suggested 3 stable macro-clusters (silhouette ≈0.40).
  - HDBSCAN found 42 finer sub-clusters, ~28% noise filtered out.
- Clusters interpreted by top fraud types, solicitation channels, and dollar loss distributions.
- Example: Cluster enriched in *Extortion/Phishing* with "Direct Call/Text" methods vs. Cluster dominated by *Investment scams* with higher losses.

4. Skills Demonstrated

- **Data wrangling:** Pandas, NumPy, regex, datetime, imputation.
- **EDA & visualization:** Plotly, seaborn, ECDFs, heatmaps, rolling trends.
- **Unsupervised ML:** PCA, UMAP, K-Means, HDBSCAN, cluster validity metrics.
- **Analytical communication:** Markdown summaries, plots, and interpretable tables.

5. Outcomes

- Produced actionable fraud profiles linking demographics, solicitation methods, and loss intensity.
- Demonstrated ability to take a raw, messy national dataset and transform it into insights using advanced statistical and machine learning techniques.
- Ready to adapt the workflow for operational fraud detection, policy insights, or business risk analytics.

Project Summary – Exploratory Fraud Q&A (EDA Supplement)

Dataset: Canadian Anti-Fraud Centre (CAFC), ~329k reports (2021–2025).
Goal: Provide a stakeholder-friendly exploratory analysis of fraud and cybercrime reports, answering key descriptive questions with clear statistics and tables.

1. Key Questions Answered

- **What are the most common fraud types?**
  - Identity Fraud (23% of reports), Extortion (9.6%), Phishing (8.7%).
- **Which scams cause the largest losses?**
  - Investments (C$1.18B), Romance (C$256M), Spear Phishing (C$246M).
  - Highest average loss per case: Investments, Spear Phishing, Timeshare.
- **Who is most affected?**
  - Romance scams peak in ages **60–69**, especially for women (C$130M vs men C$69M).
  - Complaint type split shows meaningful differences across fraud themes (χ^2 test $p<0.00001$).
- **How do fraudsters reach victims?**
  - Major channels by loss: Internet-Social (C$717M), Internet (C$596M), Email (C$305M), Direct Call (C$252M).
  - Door-to-door is rare but has the **highest average loss per case** (~C$51k).
- **Where are the geographic hot spots?**
  - Ontario leads in both reports (96k) and total losses (C$871M).
  - British Columbia and Alberta also high in losses; Quebec ranks 2nd in report volume (67k).
  - Adjusted for population, Quebec and Manitoba show high per-capita exposure.

2. Skills Demonstrated

- **Data summarization:** Cross-tabbing by theme, method, age, gender, and province.
- **Statistical testing:** Chi-square for complaint type distributions.
- **Stakeholder communication:** Direct Q&A style with concise tables and percentages.
- **Geographic insights:** Provincial ranking by reports and dollar losses.

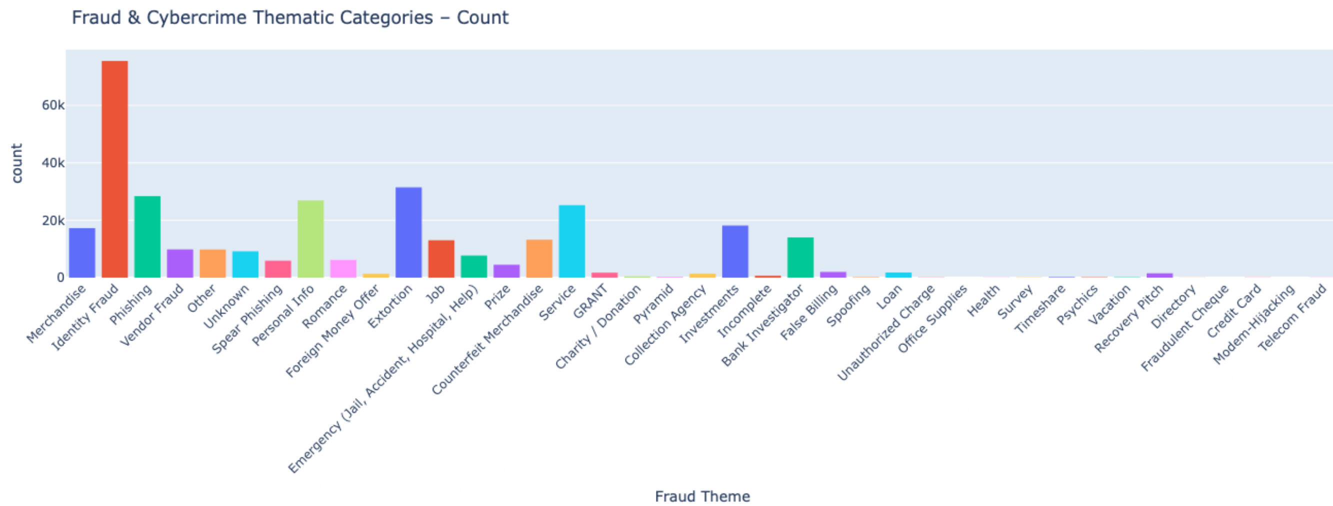
3. Outcomes

- Clear profiles of **high-impact fraud types** (Investments, Romance, Spear Phishing).
- Identified **vulnerable demographics** (seniors, women in romance scams).
- Exposed **channel risk differences** (door-to-door vs online).

Exploratory Questions to Answer

What are the most common types of fraud in Canada?

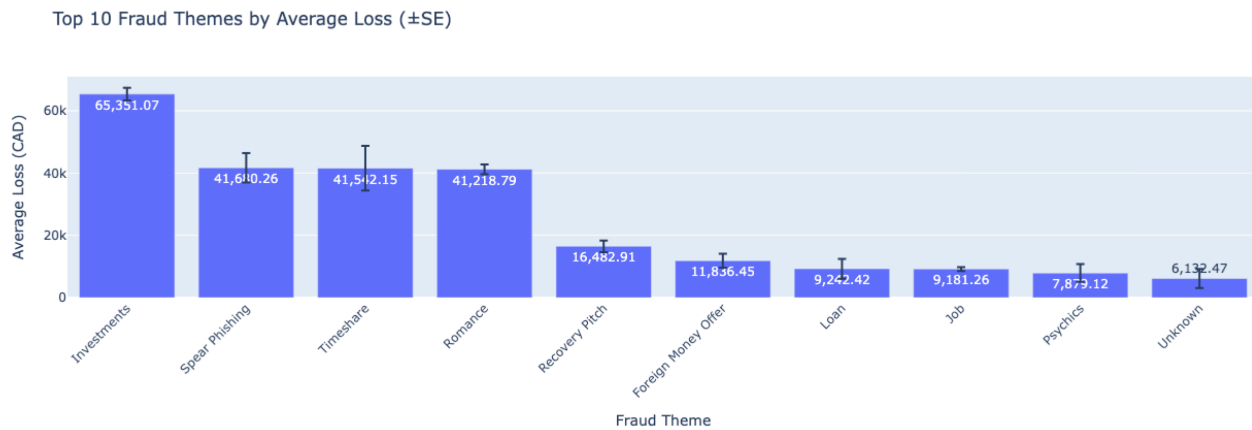
Identity fraud (23%), extortion (9.57%), phishing (8.65%), personal info (8.1%)

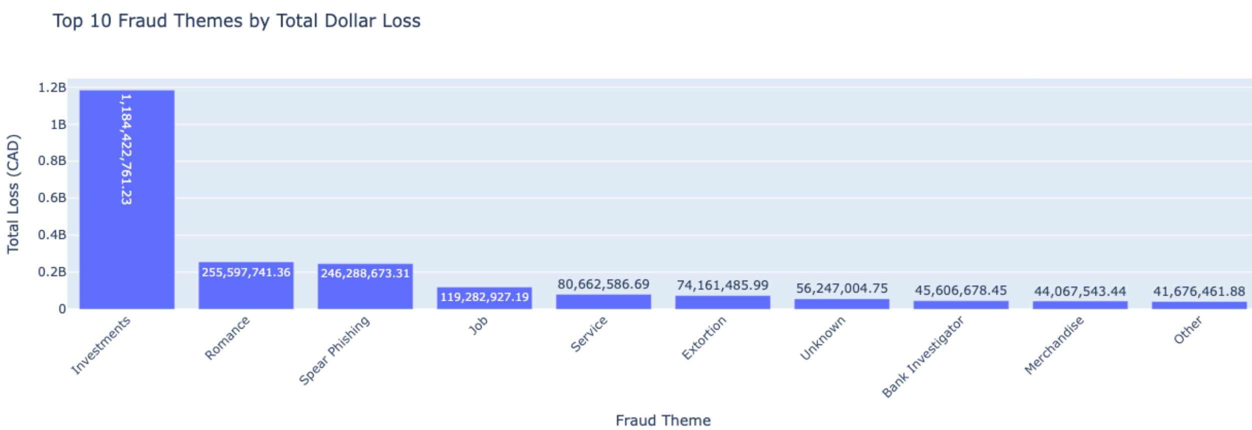


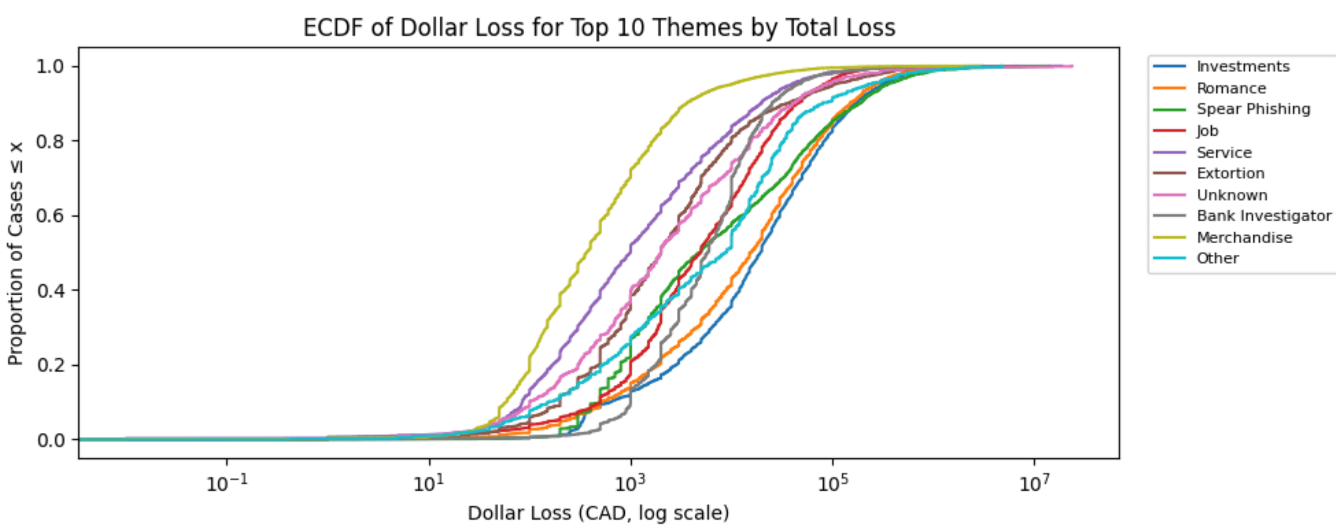
Which fraud types lead to the highest financial losses?

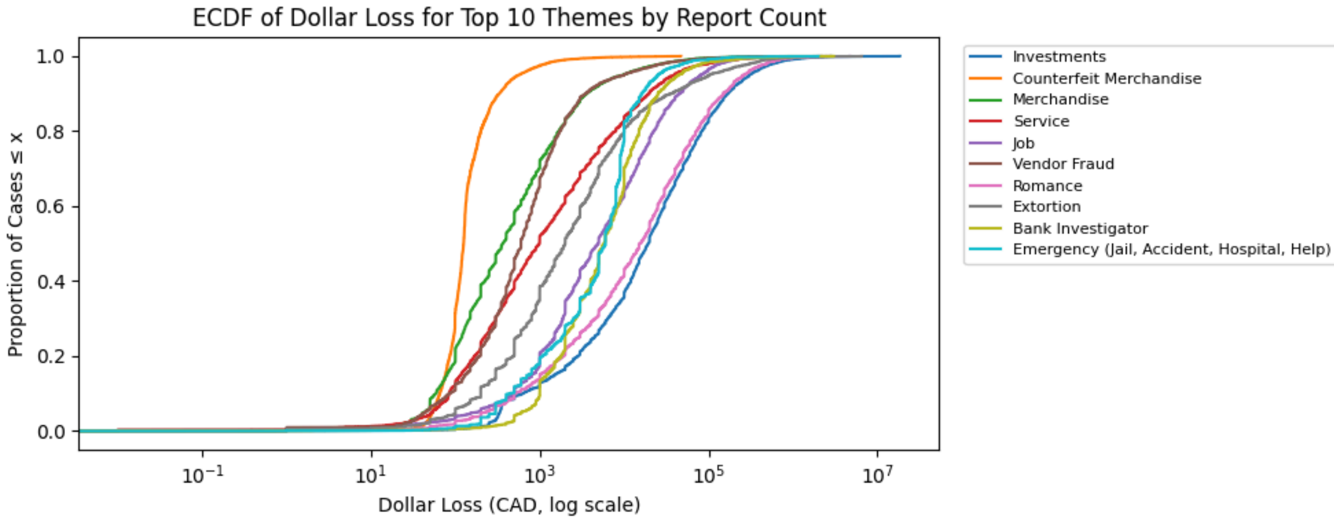
Top 10 Fraud Themes by TOTAL and Average Loss

Fraud Theme	Total Loss (CAD)	Fraud Theme	Average Loss (CAD)
Investments	1,184,422,761.23	Investments	65,351.07
Romance	255,597,741.36	Spear Phishing	41,680.26
Spear Phishing	246,288,673.31	Timeshare	41,542.15
Job	119,282,927.19	Romance	41,218.79
Service	80,662,586.69	Recovery Pitch	16,482.91
Extortion	74,161,485.99	Foreign Money Offer	11,836.45
Unknown	56,247,004.75	Loan	9,242.42
Bank Investigator	45,606,678.45	Job	9,181.26
Merchandise	44,067,543.44	Psychics	7,879.12
Other	41,676,461.88	Unknown	6,132.47

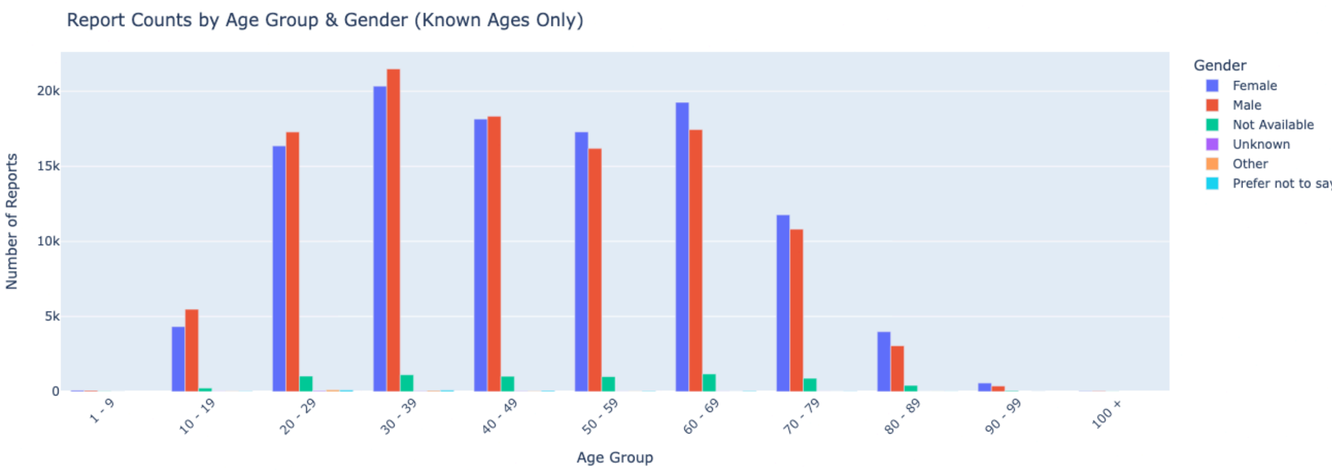

Top 10 Fraud Themes by Average Loss (±SE)
Average Loss (CAD)
65,351.07
41,218.79
9,181.26
Investments
Spear Phishing
Timeshare
Romance
Recovery Pitch
Foreign Money Offer
Loan
Job
Psychics
Unknown
Fraud Theme

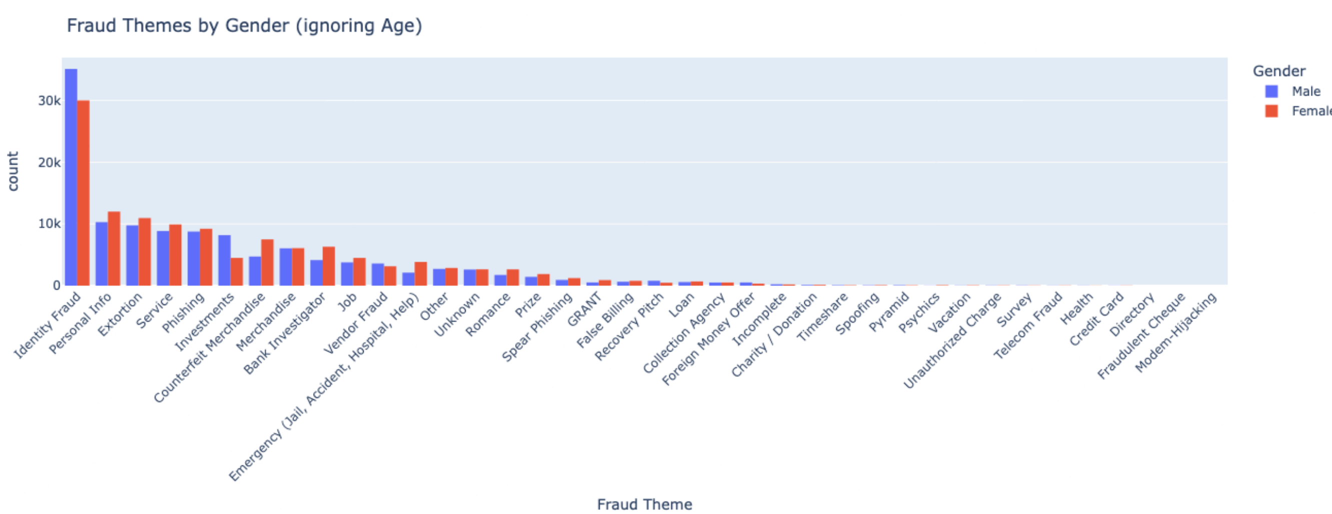

Top 10 Fraud Themes by Total Dollar Loss
Total Loss (CAD)
1,184,422,761.23
255,597,741.36
246,288,673.31
119,282,927.19
80,662,586.69
74,161,485.99
56,247,004.75
45,606,678.45
44,067,543.44
41,676,461.88
Investments
Romance
Spear Phishing
Job
Service
Extortion
Unknown
Bank Investigator
Merchandise
Other
Fraud Theme


ECDF of Dollar Loss for Top 10 Themes by Total Loss
Proportion of Cases ≤ x
Dollar Loss (CAD, log scale)
Investments
Romance
Spear Phishing
Job
Service
Extortion
Unknown
Bank Investigator
Merchandise
Other



How does fraud vary across age groups and genders?

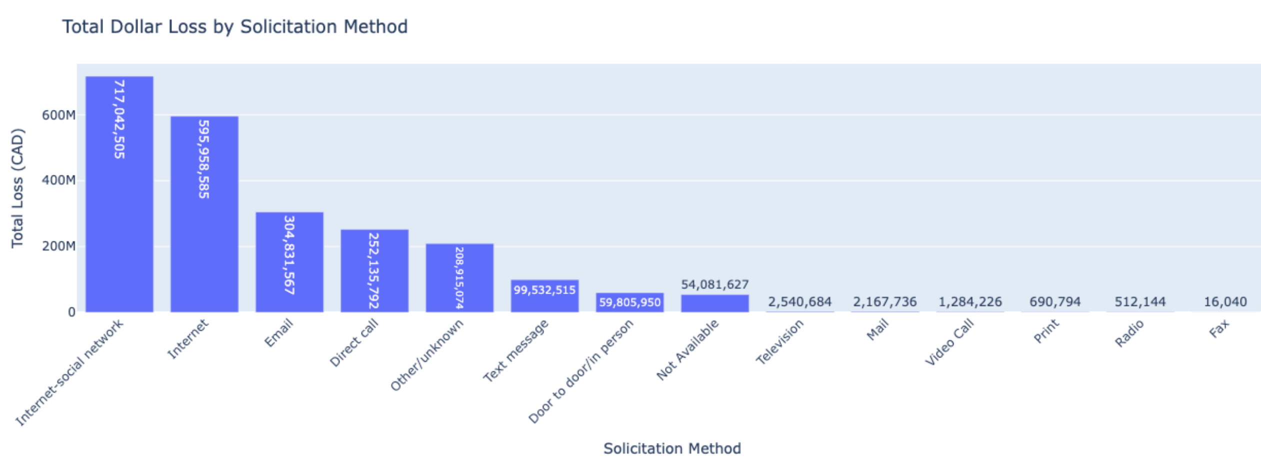


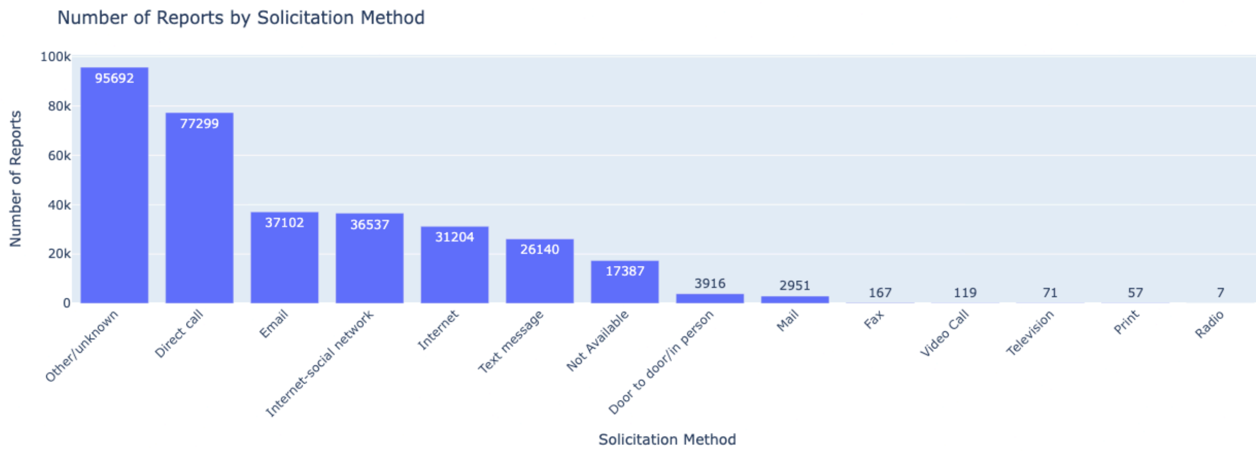


What are the most common fraudster solicitation methods?

Solicitation Method – Reports, Total Loss, Avg Loss (positive only)

Solicitation Method	reports	total_loss	avg_loss_pos
Direct call	77299	252,135,791.54	24,313.96
Door to door/in person	3916	59,805,950.29	51,335.58
Email	37102	304,831,567.39	47,078.23
Fax	167	16,039.98	5,346.66
Internet	31204	595,958,584.70	31,172.64
Internet-social network	36537	717,042,504.87	29,289.76
Mail	2951	2,167,736.14	15,483.83
Not Available	17387	54,081,627.27	4,706.84
Other/unknown	95692	208,915,074.13	63,269.25
Print	57	690,794.48	36,357.60
Radio	7	512,143.78	102,428.76
Television	71	2,540,684.38	55,232.27
Text message	26140	99,532,515.20	26,345.29
Video Call	119	1,284,226.30	22,932.61





Are there geographic hot spots for certain types of fraud?

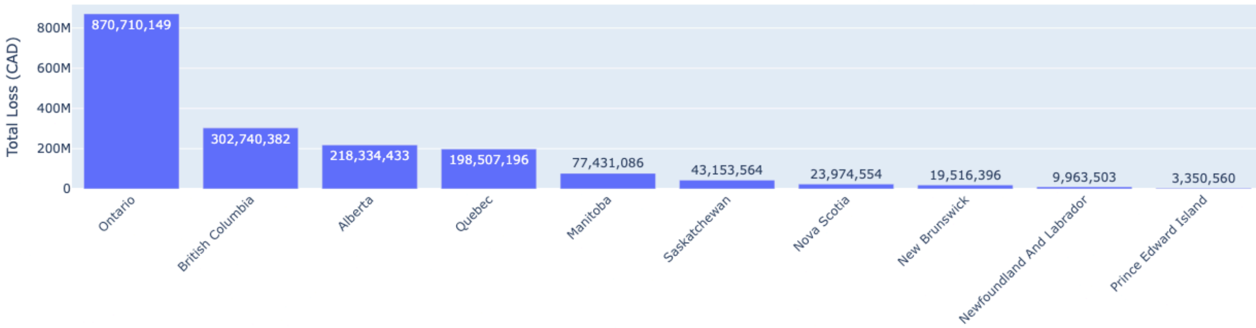
Top 10 Provinces by Number of Reports and Total Dollar Loss

Province/State	reports	Province/State	total_loss
Ontario	96150	Ontario	870,710,149.23
Quebec	67041	British Columbia	302,740,382.25
British Columbia	31604	Alberta	218,334,432.59
Alberta	26189	Quebec	198,507,196.26
Manitoba	9014	Manitoba	77,431,085.95
Saskatchewan	5889	Saskatchewan	43,153,564.20
Nova Scotia	5334	Nova Scotia	23,974,554.28
New Brunswick	4408	New Brunswick	19,516,396.28
Newfoundland And Labrador	1912	Newfoundland And Labrador	9,963,503.26
Prince Edward Island	857	Prince Edward Island	3,350,559.50

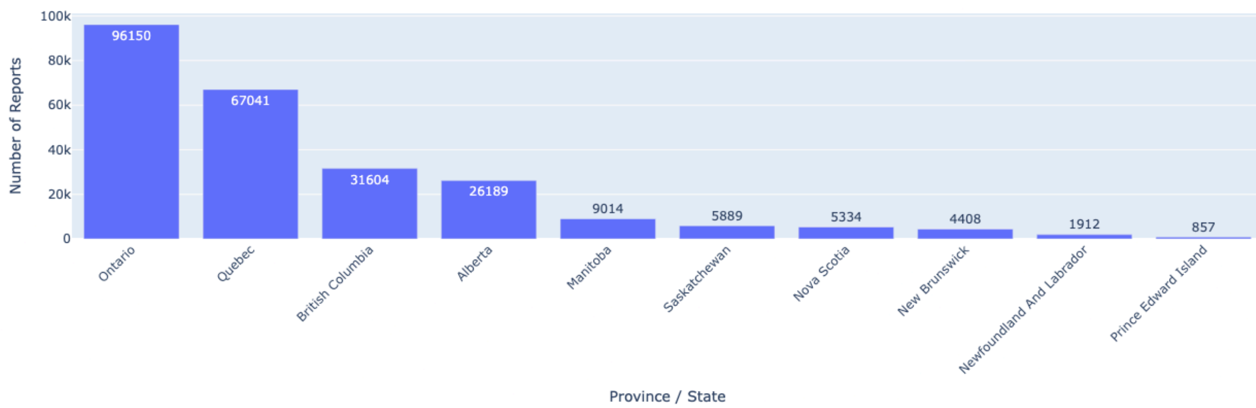
What are the top fraud categories in each province?

Total Dollar Loss (CAD) in Top 10 Provinces

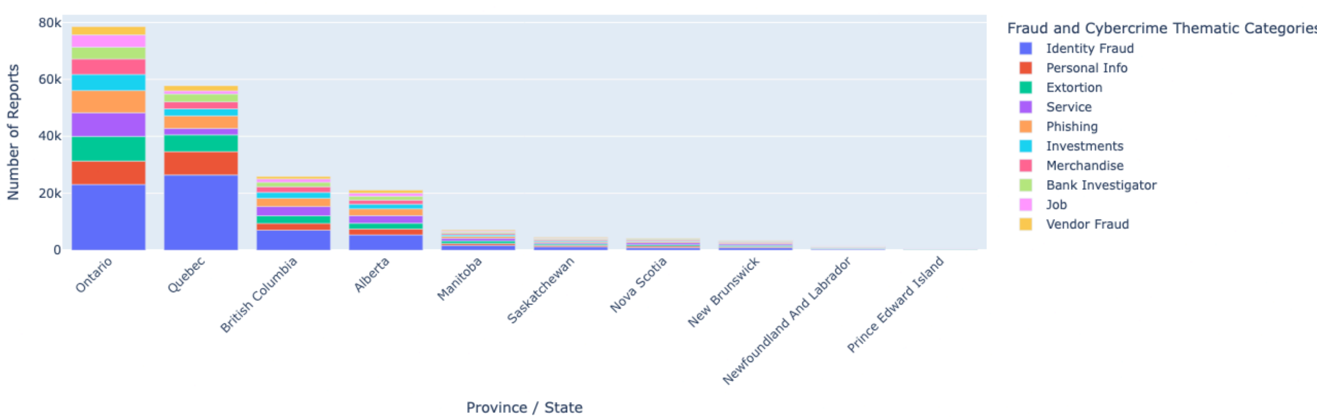
Total Dollar Loss in Top 10 Provinces

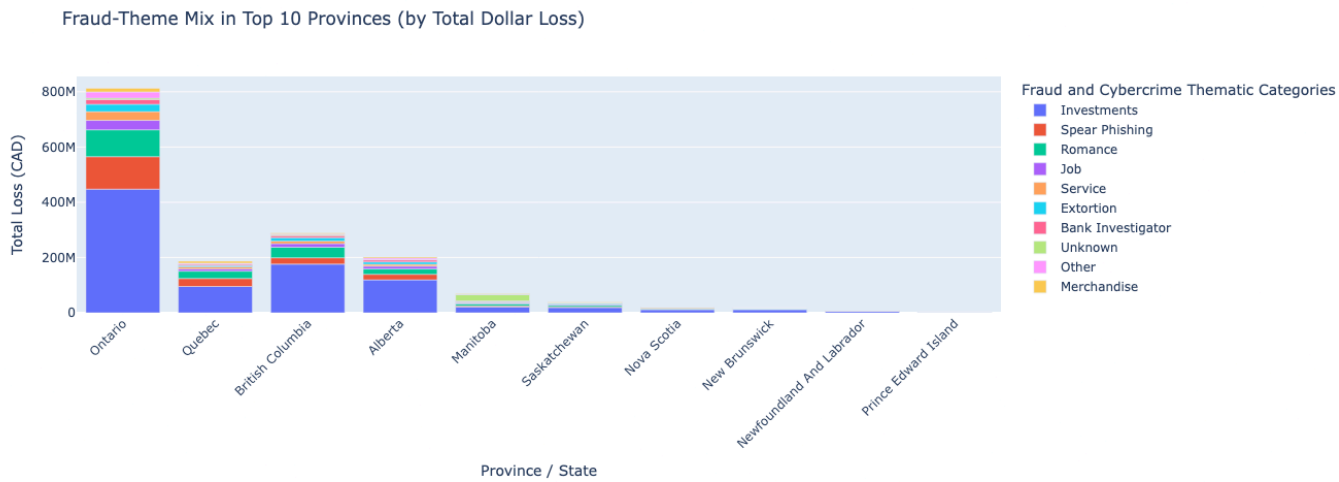


Top 10 Provinces by Report Count



Fraud-Theme Mix in Top 10 Provinces (Top 10 Themes Only)





Side Questions:

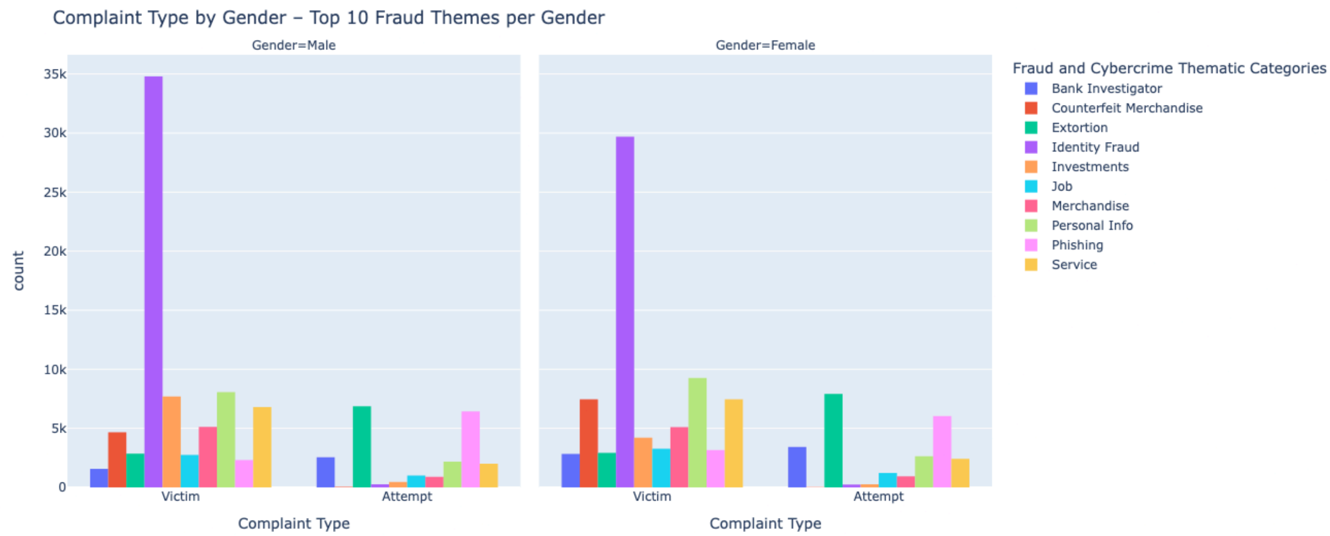
1- romance x age x gender: when romance fraud is the highest aka at what age is each gender most susceptible to romance fraud?
- total loss $ by gender
- susceptibility
2- complaint type (victim or attempt) by sex, age, category

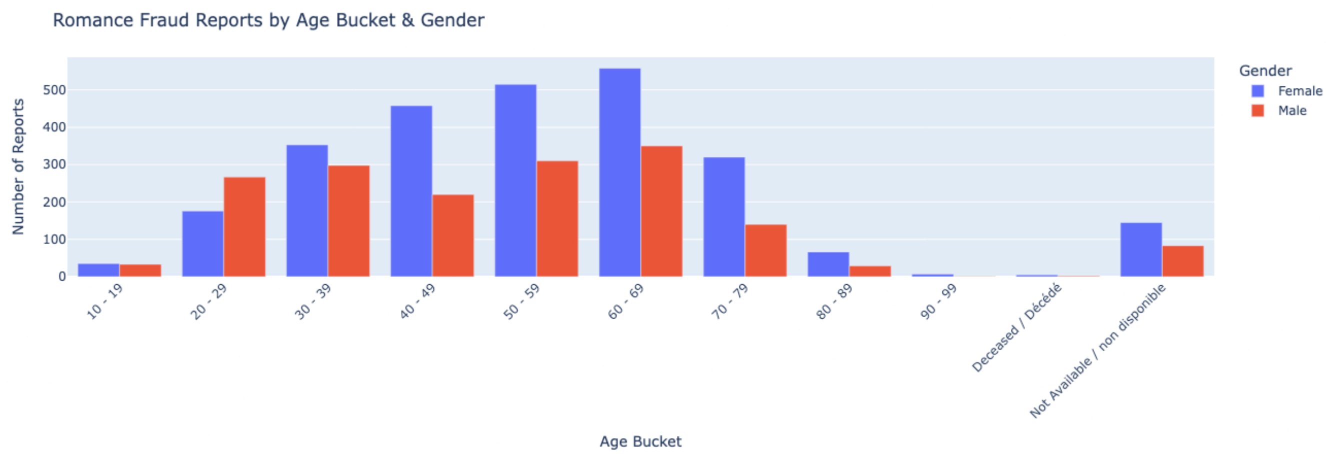
Romance Fraud – Peak Age Bucket & Total Loss by Gender

Gender	peak_age_bucket	reports	total_loss
Female	60 - 69	558	129,870,493.35
Male	60 - 69	350	68,990,292.37

Complaint Type × Gender (Filtered to Victim / Attempt & Male / Female)

Complaint Type	Male	Female
Victim	85143	85938
Attempt	31869	35518





Are specific complaint types (Victim, Attempt) more frequent in some fraud types?

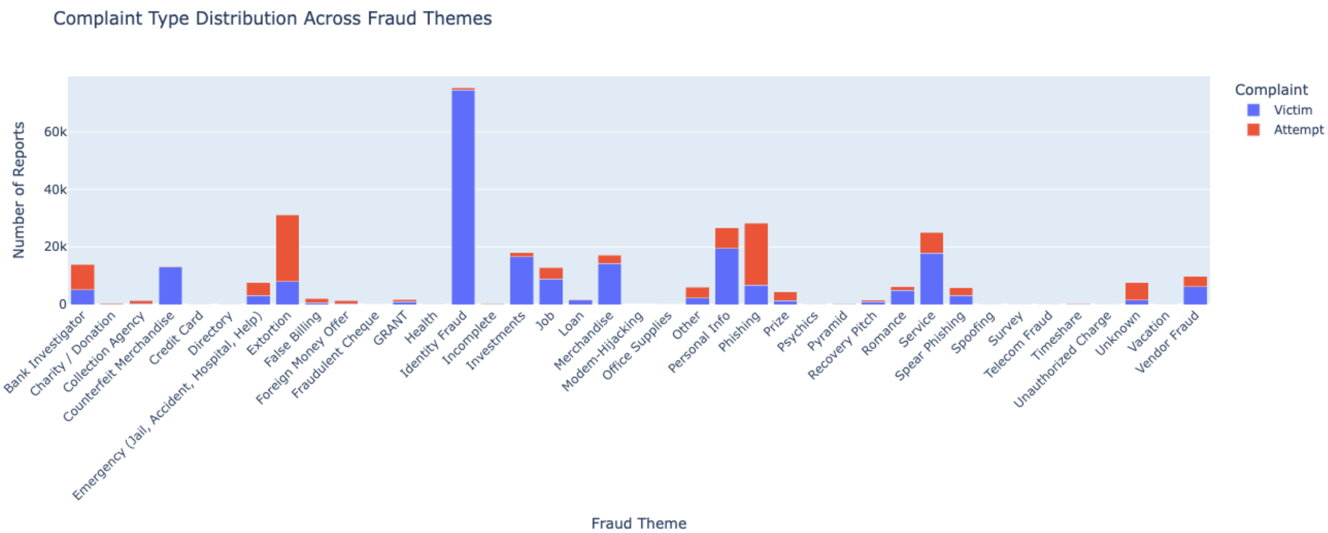
Chi-square test across themes: $\chi^2=122058.7$, dof=38, $p< 0.00001$

Victim vs Attempt Counts by Fraud Theme

Fraud and Cybercrime Thematic Categories	Attempt	Victim	Victim %
Bank Investigator	8835	5120	0.366894
Charity / Donation	259	160	0.381862
Collection Agency	1180	209	0.150468
Counterfeit Merchandise	157	13066	0.988127
Credit Card	4	54	0.931034
Directory	38	11	0.22449

Fraud and Cybercrime Thematic Categories	Attempt	Victim	Victim %
Emergency (Jail, Accident, Hospital, Help)	4419	3227	0.422051
Extortion	23181	7966	0.255755
False Billing	1313	683	0.342184
Foreign Money Offer	1062	274	0.20509
Fraudulent Cheque	3	1	0.25
GRANT	762	976	0.561565
Health	52	24	0.315789
Identity Fraud	732	74633	0.990287
Incomplete	226	93	0.291536
Investments	1283	16672	0.928544
Job	3985	8907	0.690894
Loan	345	1448	0.807585
Merchandise	2891	14211	0.830955
Modem-Hijacking	0	1	1
Office Supplies	7	1	0.125
Other	3660	2307	0.386626
Personal Info	7154	19504	0.731638
Phishing	21628	6732	0.237377
Prize	3107	1371	0.306163
Psychics	43	104	0.707483
Pyramid	49	230	0.824373
Recovery Pitch	624	902	0.591088
Romance	1295	4869	0.789909
Service	7116	17942	0.716019
Spear Phishing	2672	3138	0.540103
Spoofing	38	39	0.506494
Survey	93	19	0.169643
Telecom Fraud	43	44	0.505747
Timeshare	51	148	0.743719
Unauthorized Charge	8	91	0.919192
Unknown	6029	1599	0.209622

Fraud and Cybercrime Thematic Categories	Attempt	Victim	Victim %
Vacation	42	93	0.688889
Vendor Fraud	3652	6115	0.626088



Remaining questions

- [ ] What percent of reports have missing key fields (age, gender, loss)?

Predictive Modeling

- [ ] Can we predict whether a fraud report will lead to financial loss?
- [ ] Which features (e.g., method, age, fraud type) are most predictive?
- [ ] Are there fraud types with disproportionately high loss-to-case ratios?
- [ ] Is loss amount associated with certain demographic combinations?

Clustering & Pattern Mining

- [ ] Are there recognizable fraud/victim personas (e.g., “retired investor scammed via phone”)?
- [ ] Can we group reports based on shared patterns (fraud type + method + victim traits)?
- [ ] Do certain fraud types cluster by province or language?
- [ ] Do some solicitation methods target specific age groups?

Policy & Impact Insights

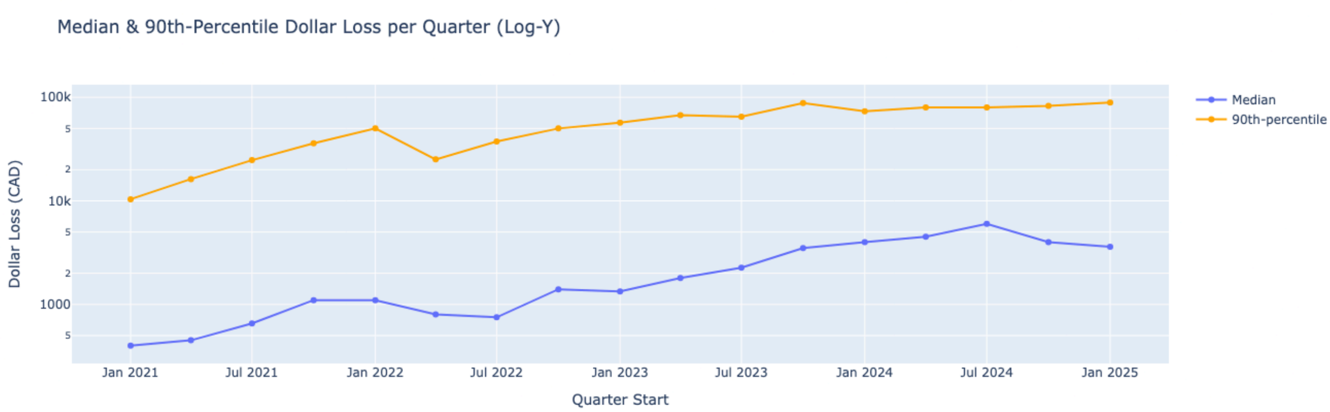
- [ ] What actionable findings could help target public awareness campaigns?
- [ ] Are there underserved or high-risk groups that could be protected?
- [ ] Which fraud types should receive investigative priority based on loss impact?

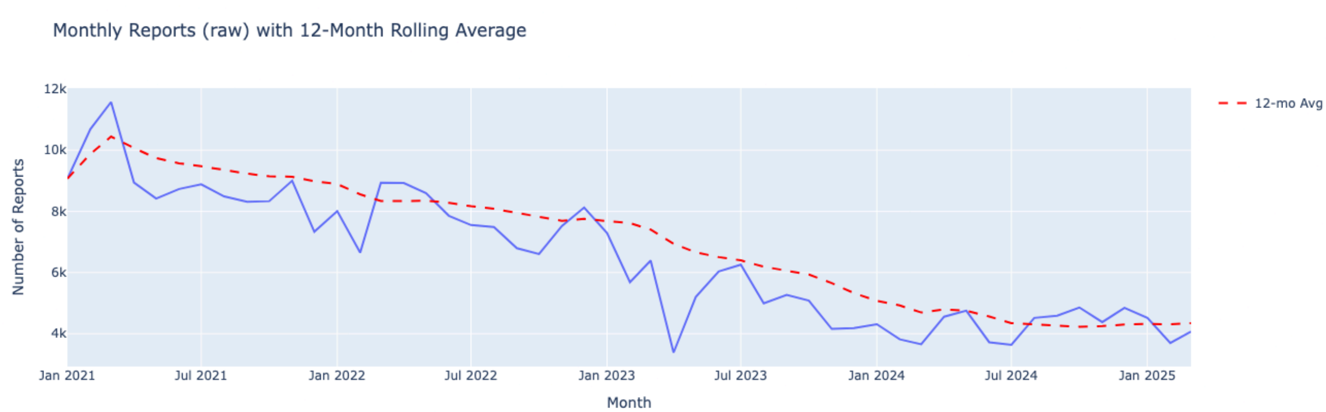
- [ ] Can we identify regions or fraud styles where enforcement/prevention may be weak?

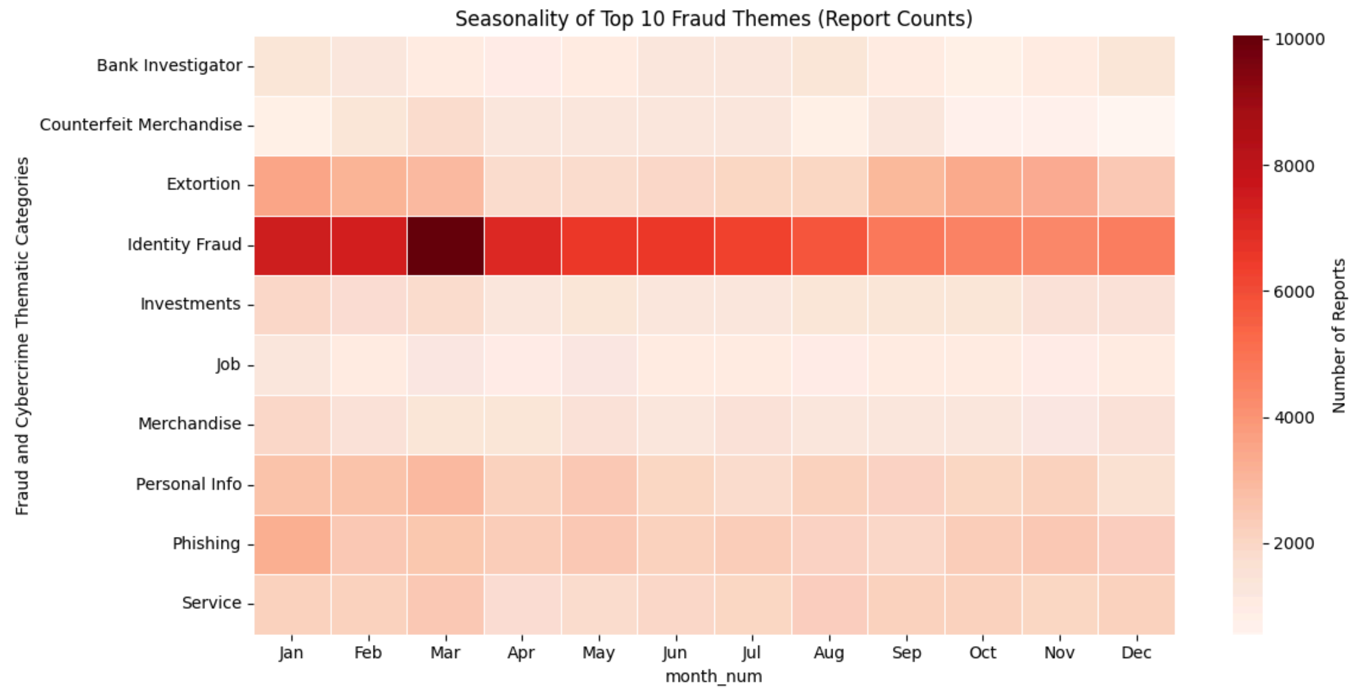
Analysis Ideas (post first EDA)

1. Temporal Dynamics

Idea	What to show	Why it matters
MonthlyTrend	Line chart of total reports per month since 2021; overlay 12-month rolling average	Detect seasonality or growth spikes
ThemeSeasonality	Faceted heat-map (month × fraud theme) of report counts	Pinpoint holiday scams, tax-season phishing, etc.
LossTrend	Line chart of **median** (or 90th-pct) Dollar Loss per quarter	Reveals whether fraudsters are netting bigger scores over time

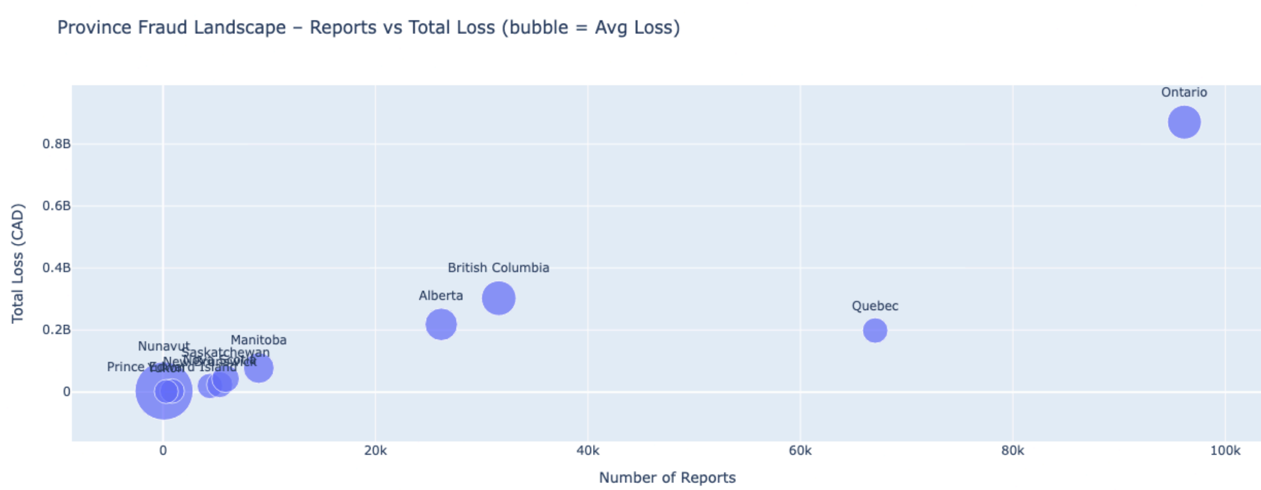




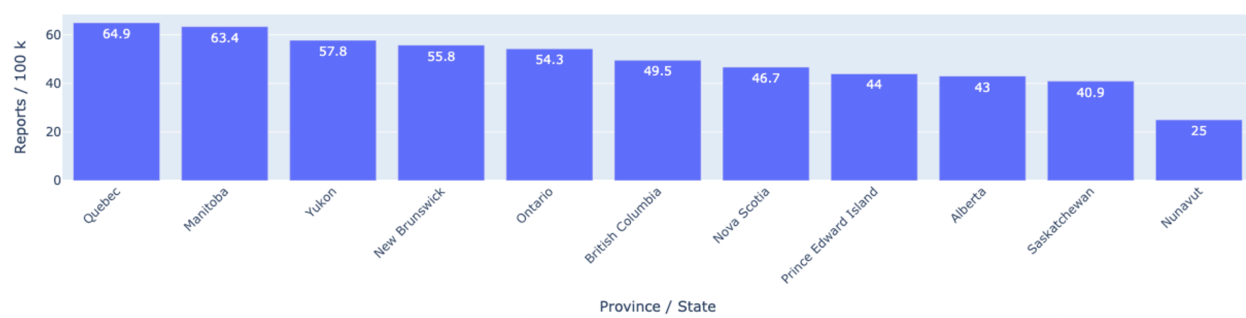


2. Geographic Deep-Dives

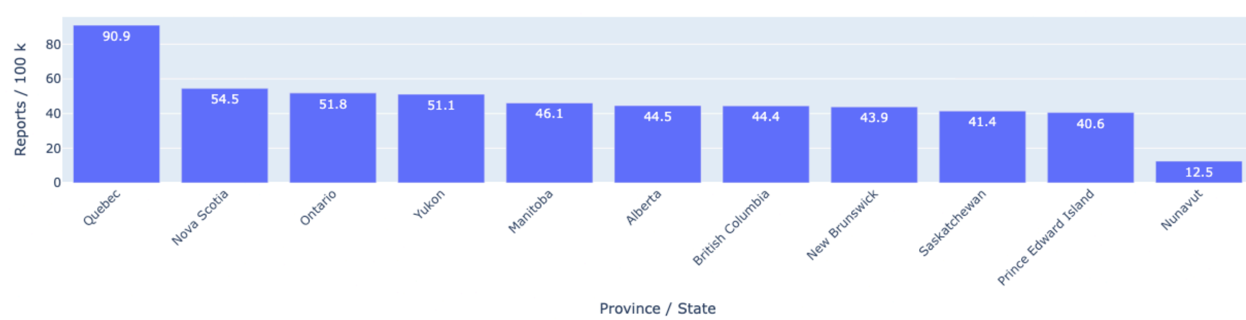
Idea	What to show	Why it matters
Per-Capita Hot Spots	Per province reports vs total loss	Adjusts for population size; flags true outliers
Theme-Specific Maps	Mini maps of top 3 themes' per-capita intensity	Targeted regional interventions
Loss vs Count Bubble	Scatter: province (x=reports, y=total loss, bubble=size=avg loss)	Separates "many small" vs "few huge" provinces



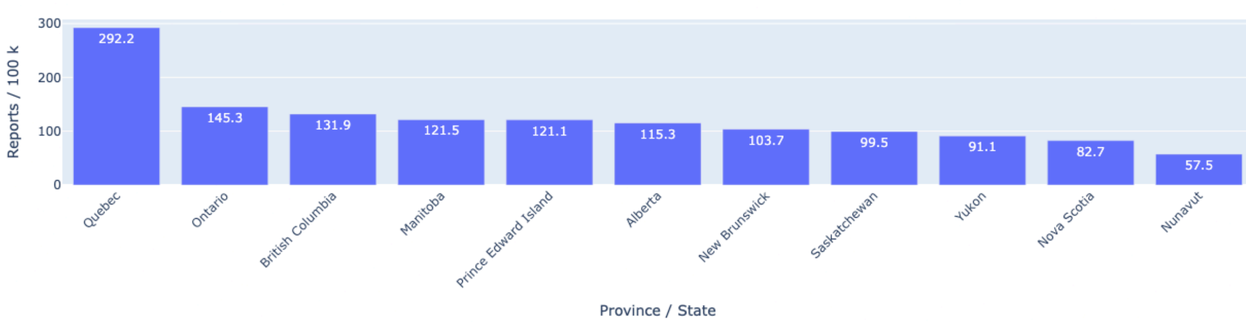
Extortion – Reports per 100 k Population (2025)



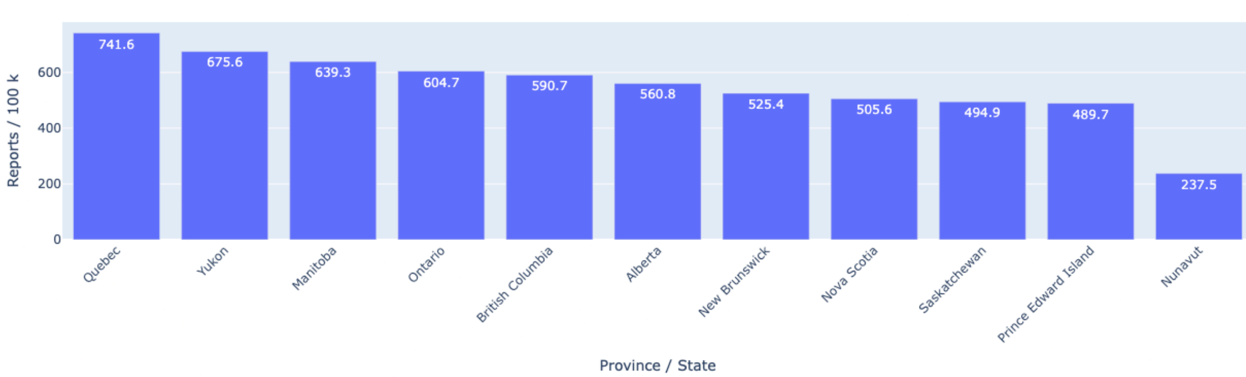
Personal Info – Reports per 100 k Population (2025)



Identity Fraud – Reports per 100 k Population (2025)



Fraud Reports per 100 k Population (2025)



3. Demographic Interactions

Idea	What to show	Why it matters
Age × Solicitation	Stacked bar or mosaic showing which methods hit each age bucket	Tailor awareness campaigns (e.g., phone scams vs seniors)
Gender × Loss Boxen	Boxen plot (log-Y) of Dollar Loss split by gender	Identifies risk profile differences
Non-Person Victims	Table: Business/Deceased counts & losses by fraud theme	Gauge corporate fraud vs individual fraud impact

4. Behavioural & Structural Patterns

Idea	What to show	Why it matters
Method → Theme Sankey	Sankey diagram from Solicitation Method to Fraud Theme	Visualise common “channels” leading into scam types
Number of Victims vs Loss	2-D density or scatter (log-log)	Exposes whether multi-victim events correlate with higher losses
Outlier Drill-down	Table of top 1 % loss cases with IDs, theme, province	Case studies for law-enforcement attention

5. Missingness & Data Quality

Idea	What to show	Why it matters
Missingness Heat-map	Seaborn matrix sorted by theme	See if certain fraud themes systematically skip age/loss
MCAR vs MAR Tests	For Language, Province, Loss missingness	Decide if imputation is safe or biasing

6. Unsupervised Segmentation

Idea	What to show	Why it matters
UMAP + HDBSCAN	2-D projection coloured by discovered cluster	Unearth latent victim/fraud personas
Feature Importance (SHAP) for Clusters	Bar of top drivers per cluster	Tells you *why* clusters differ

7. Predictive Modeling Prep

Idea	What to show	Why it matters
Binary Classifier: Will a report incur $ loss?	ROC curve, precision-recall, SHAP	Proactive alerting; understand drivers
Loss Severity Regression	Quantile regression (median, 90th-pct)	Predict potential exposure even if loss not yet known
Pipeline w/ KNN Imputer & One-Hot	Code + cross-val metrics	Ensures no leakage & clean feature handling

8. Risk-Weighted KPIs

Idea	What to show	Why it matters
Loss per Report Ratio	Bar: (total_loss / reports) by theme	Highlights “high payoff” scams
Theme Gini Coefficient	Inequality of losses within each theme	Shows whether a few huge cases dominate totals

9. Interactive Dashboard Plan

Idea	Component	Data source
Global Filters	Year, Province, Theme	Parquet
KPI Cards	Total Loss, Reports, Median Loss	Live aggregates
Linked Plots	Map, Time-series, ECDF	Cached rollups
Download Button	CSV of filtered subset	UI convenience

10. Documentation & Reporting

Idea	Deliverable	Purpose
Methodology Note	Obsidian page summarising cleaning & imputation decisions	Transparency for reviewers
One-pager Infographic	Canva or Figma visual of key stats	Exec storytelling

Idea	Deliverable	Purpose
Repro Steps	Makefile or `run_all.sh` to rebuild parquet, run EDA	Easy hand-off to teammates