Exploratory Analysis of Credit Card Fraud Data

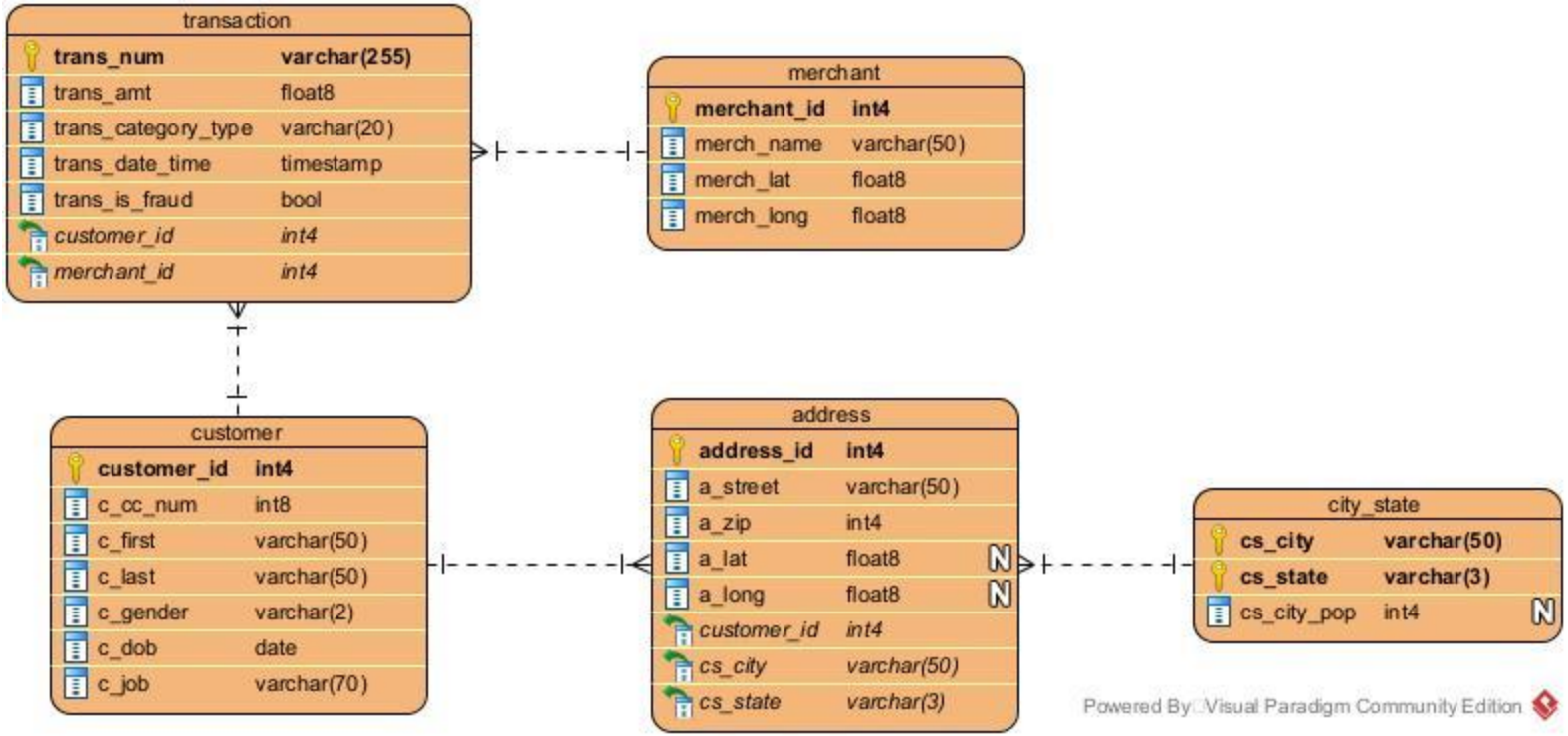
Presented by:
OBIEME, Esther
SANTIAGO, Ma. Angelica

Table of Contents

Data Set

	Column	Data_Type	Max_Character_Length	Min_Character_Length	Null_Value_Count	Null_%	Distinct_Value_Count	Min_Value	Max_Value	Average	Sample_Values
0	trans_date_trans_time	object	16	16	0	0.00%	226,976	NaN	NaN	None	[17/10/2020 21:31, 30/10/2020 15:25]
1	cc_num	float64	None	None	0	0.00%	904	6.04E+10	4.99E+18	4.17839E+17	[6538440000000000.0, 5559860000000000.0]
2	merchant	object	43	13	0	0.00%	693	NaN	NaN	None	[fraud_Macejkovic-Lesch, fraud_Hudson-Grady, fraud_Bradtke PLC]
3	category	object	14	4	0	0.00%	14	NaN	NaN	None	[grocery_net, misc_net, grocery_pos, travel, personal_care]
4	amt	float64	None	None	0	0.00%	37,256	1.00	22,768.11	69.39	[80.19, 75.24, 63.03, 8.09, 3.18]
5	first	object	11	3	0	0.00%	341	NaN	NaN	None	[Charles, Jodi, Destiny, Michael, Michael]
6	last	object	11	2	0	0.00%	471	NaN	NaN	None	[Santos, Harris, Vance, Bell, Curry]
7	gender	object	1	1	0	0.00%	2	NaN	NaN	None	[F, M, F, F, F]
8	street	object	35	12	0	0.00%	924	NaN	NaN	None	[137 Adam Dale, 08236 Kim Hill, 358 Pruitt Square]
9	city	object	25	3	0	0.00%	849	NaN	NaN	None	[Brooklyn, Waupaca, Mendon, Cascade Locks, Clay Center]
10	state	object	2	2	0	0.00%	50	NaN	NaN	None	[TX, MN, AR, LA, OK]
11	zip	int64	None	None	0	0.00%	912	1.26E+03	9.99E+04	48842.63	[78040, 47987, 31605, 56668, 96103]
12	lat	float64	None	None	0	0.00%	910	2.00E+01	6.57E+01	38.54	[45.0632, 32.153, 41.0646, 41.8467, 31.7706]
13	long	float64	None	None	0	0.00%	910	-1.66E+02	-6.80E+01	-90.23	[-97.2092, -98.9656, -75.2811, -102.7413, -91.4539]
14	city_pop	int64	None	None	0	0.00%	835	2.30E+01	2.91E+06	88221.89	[19090, 2097, 3402, 965, 22305]
15	job	object	59	3	0	0.00%	478	NaN	NaN	None	[Science writer, Call centre manager, Social researcher]
16	dob	object	10	10	0	0.00%	910	NaN	NaN	None	[20/07/1984, 05/07/1984, 24/11/1994, 13/01/1960, 27/09/1970]
17	trans_num	object	32	32	0	0.00%	555,719	NaN	NaN	None	[5130dd961ff7ce0edac70d92f235863a]
18	unix_time	int64	None	None	0	0.00%	544,760	1.37E+09	1.39E+09	1380678865	[1388296019, 1381355566, 1376766833, 1387020171, 1373093323]
19	merch_lat	float64	None	None	0	0.00%	546,490	1.90E+01	6.67E+01	38.54	[40.393163, 35.460916, 41.419746, 29.392323, 41.096626]
20	merch_long	float64	None	None	0	0.00%	551,770	-1.67E+02	-6.70E+01	-90.23	[-80.125188, -93.706401, -90.657289, -72.652867, -84.579786]
21	is_fraud	int64	None	None	0	0.00%	2	0.00E+00	1.00E+00	0	[0, 0, 0, 0, 0]

Entity Relationship Diagram



Data Exploration

Query 1

Multi-Table Join – Exploring Transaction Details

	customer_id	c_job	c_gender	age	cs_city	cs_state	a_zip	trans_date_time	merch_name	trans_category_type	trans_amt	trans_is_fraud
0	74078269	Mechanical engineer	M	52	Columbia	SC	29209	6/21/2020 12:14	Kirlin and Sons	personal care	2.86	FALSE
1	18702918	Sales professional, IT	F	30	Altonah	UT	84002	6/21/2020 12:14	Sporer-Keebler	personal care	29.84	FALSE
2	95127892	Librarian, public	F	50	Bellmore	NY	11710	6/21/2020 12:14	Swaniawski, Nitzsche and Welch	health fitness	41.28	FALSE
3	59734903	Set designer	M	33	Titusville	FL	32780	6/21/2020 12:15	Haley Group	misc pos	60.05	FALSE
4	81017923	Furniture designer	M	65	Falmouth	MI	49632	6/21/2020 12:15	Johnston-Casper	travel	3.19	FALSE
...	...	...	...	...	...	...	...	...	...	...	...	...
555714	91660892	Town planner	M	54	Luray	MO	63453	12/31/2020 23:59	Reilly and Sons	health fitness	43.77	FALSE
555715	51585285	Futures trader	M	21	Lake Jackson	TX	77566	12/31/2020 23:59	Hoppe-Parisian	kids pets	111.84	FALSE
555716	99872181	Musician	F	39	Burbank	WA	99323	12/31/2020 23:59	Rau-Robel	kids pets	86.88	FALSE
555717	28750882	Cartographer	M	55	Mesa	ID	83643	12/31/2020 23:59	Breitenberg LLC	travel	7.99	FALSE
555718	58471486	Media buyer	M	27	Edmond	OK	73034	12/31/2020 23:59	Dare-Marvin	entertainment	38.13	FALSE

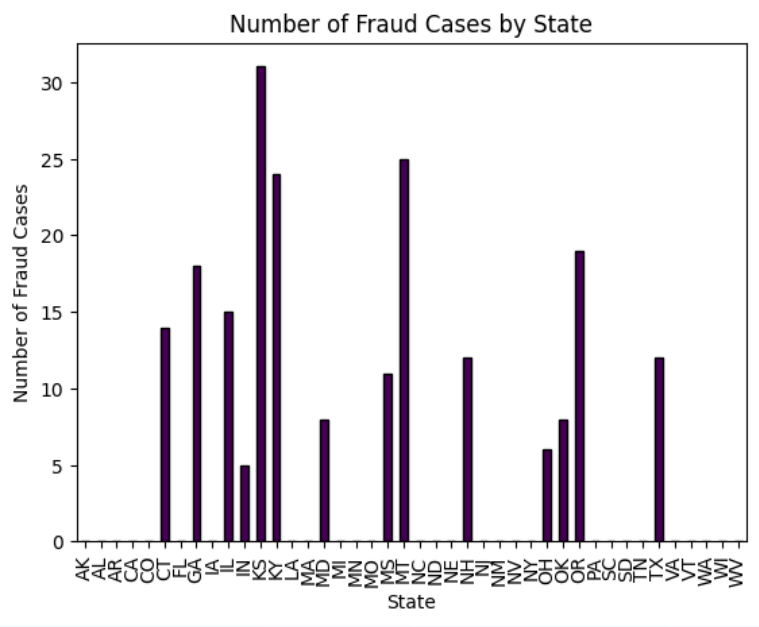
	customer_id	c_job	c_gender	age	cs_city	cs_state	a_zip	trans_date_time	merch_name	trans_category_type	trans_amt	trans_is_fraud
0	163011	Engineer, control and instrumentation	F	40	Bridgeport	NJ	8014	12/27/2020 15:58	Kozey-McDermott	travel	22,768.11	FALSE
1	49735544	Higher education careers adviser	F	26	Jay	FL	32565	12/22/2020 21:30	Corwin-Romaguera	travel	21,437.71	FALSE
2	64061725	Hospital doctor	M	65	Pembroke	NC	28372	11/27/2020 14:54	Kovacek Ltd	travel	19,364.91	FALSE
3	89989635	Health physicist	F	49	Newhall	CA	91321	9/21/2020 12:02	Johnston-Casper	travel	16,837.08	FALSE
4	83765334	Surgeon	M	21	Conway	NH	3818	12/16/2020 21:16	Boyer-Haley	travel	16,339.26	FALSE
...	...	...	...	...	...	...	...	...	...	...	...	...
555714	78534997	Structural engineer	F	35	Philadelphia	PA	19149	7/31/2020 12:43	Dooley Inc	shopping pos	1	FALSE
555715	94198392	Podiatrist	F	20	Beaver Falls	PA	15010	7/31/2020 19:32	Bernier, Volkman and Hoeger	misc net	1	FALSE
555716	66347357	Film/video editor	F	30	West Sayville	NY	11796	8/10/2020 14:34	Metz-Boehm	shopping pos	1	FALSE
555717	93323632	Colour technologist	F	37	Cowlesville	NY	14037	8/3/2020 20:46	Goldner-Lemke	entertainment	1	FALSE
555718	277372	Planning and development surveyor	F	38	Lake Oswego	OR	97034	8/3/2020 19:59	Goyette Inc	shopping net	1	FALSE

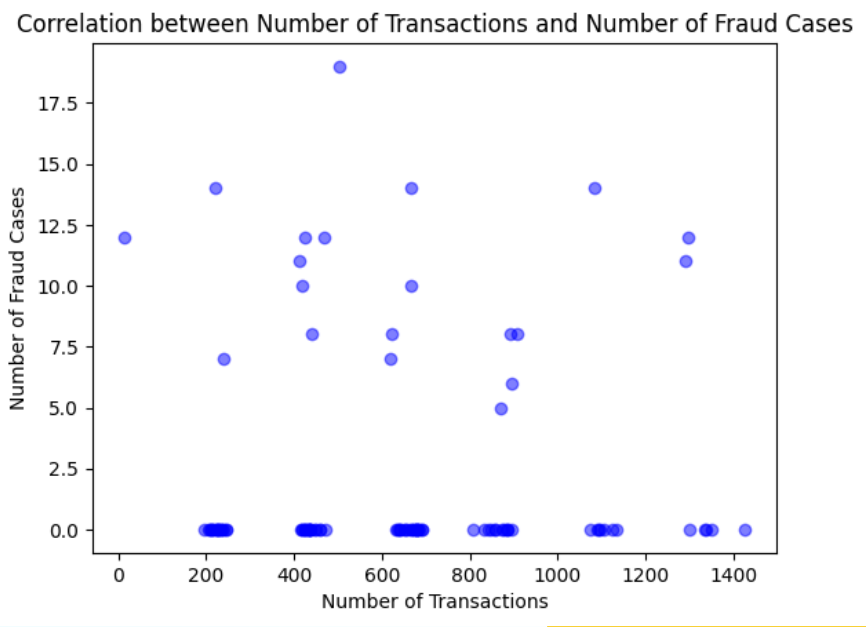
```
SELECT
    c.customer_id,
    c.c_job,
    c.c_gender,
    EXTRACT(YEAR FROM AGE('2020-12-31', c_dob)) AS
    age,
    a.cs_city,
    a.cs_state,
    a.a_zip,
    t.trans_date_time,
    m.merch_name,
    t.trans_category_type,
    t.trans_amt,
    t.trans_is_fraud
FROM customer c
    JOIN address a ON c.customer_id = a.customer_id
    JOIN transaction t ON c.customer_id =
    t.customer_id
    JOIN merchant m ON t.merchant_id = m.merchant_id
ORDER BY
    t.trans_date_time;
```

Query 2

Creating a View – Identifying Frequent Customers

	customer_id	c_gender	age	c_job	a_zip	cs_city	cs_state	num_of_transaction	total_amount	no_of_fraud
0	2279629	M	24	Pensions consultant	40077	Westport	KY	1428	$98,991.93	0
1	916035	F	48	Designer, exhibition/display	92585	Sun City	CA	1352	$117,944.10	0
2	6779390	F	45	Firefighter	46254	Indianapolis	IN	1339	$114,631.63	0
3	9794581	F	32	Tax inspector	65072	Rocky Mount	MO	1334	$74,756.38	0
4	8548529	F	35	Regulatory affairs officer	21872	Whaleyville	MD	1301	$66,376.27	0
...	...	...	...	...	...	...	...	...	...	...
95	10359342	F	68	Magazine journalist	62266	New Memphis	IL	211	$10,807.69	0
96	4294505	F	84	Biochemist, clinical	25832	Daniels	WV	209	$12,838.53	0
97	7150172	M	33	Comptroller	92101	San Diego	CA	204	$15,149.16	0
98	10180169	M	81	Estate manager/land agent	50527	Curlew	IA	195	$11,102.84	0
99	3603551	M	18	Chemical engineer	66958	Morrowville	KS	12	$7,993.74	12





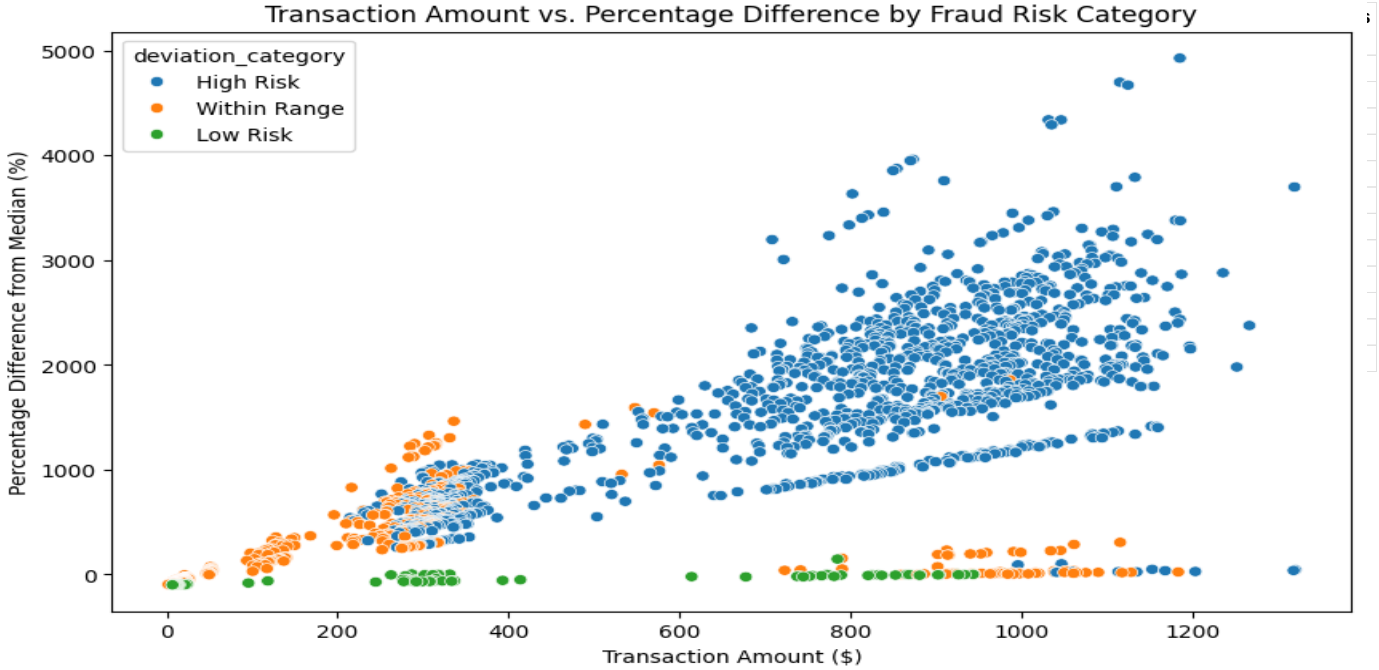

```
CREATE VIEW top_customers
AS
SELECT
    c.customer_id,
    c.c_gender,
                                    E('2020-12-31',


                                    AS

                                    total_amount,
                                    s_is_fraud =
                                    L END ) AS


                                    N c.customer_id

                                    customer_id =


                                    ob,
                                    a.cs_city,
```

Query 3

CTE & Materialized View – Analyzing Transaction Patterns

```
WITH median_transaction AS (
SELECT
    c.customer_id,
    PERCENTILE_CONT(0.5) WITHIN GROUP (ORDER BY
    t.trans_amt) AS median_trans_amt
FROM
    customer c
    JOIN transaction t ON t.customer_id =
    c.customer_id
GROUP BY
    c.customer_id
) SELECT
    c.customer_id,
    c.c_gender,
    c.c_job,
    m.merch_name,
    t.trans_category_type,
    t.trans_date_time,
    t.trans_amt,
    med.median_trans_amt,
    ((t.trans_amt - med.median_trans_amt)/
    med.median_trans_amt) * 100 AS
    percentage_diff, t.trans_is_fraud
FROM
    customer c
    JOIN transaction t ON c.customer_id =
    t.customer_id
    JOIN merchant m ON t.merchant_id =
    m.merchant_id
    JOIN median_transaction med ON
    c.customer_id = med.customer_id
ORDER BY
    percentage_diff DESC;
```

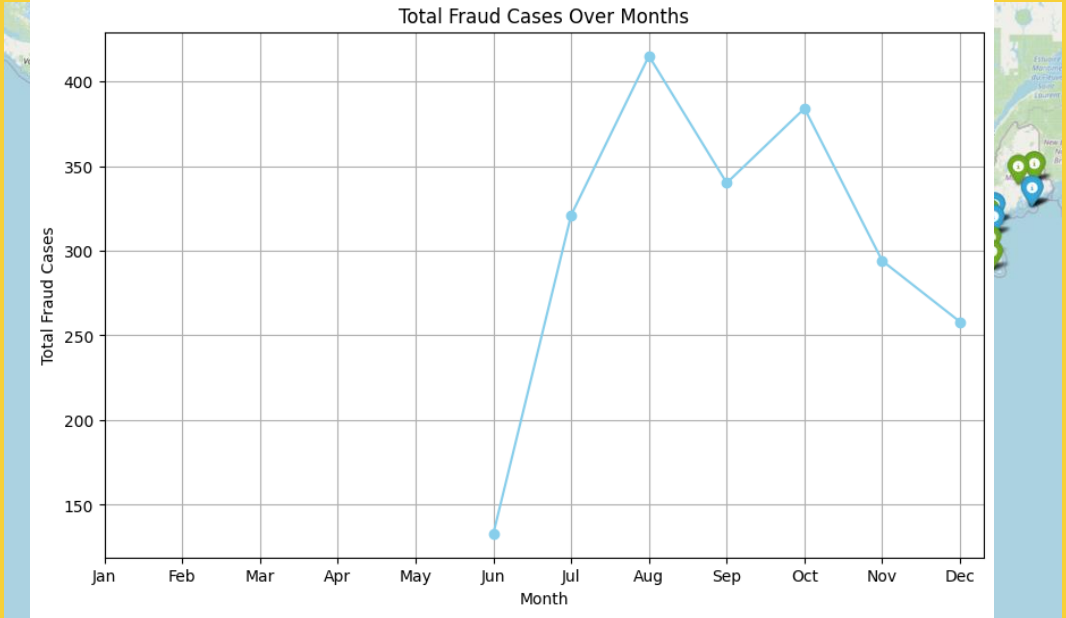
	customer_id	c_gender	c_job	merch_name	trans_category_type	trans_date_time	trans_amt	median_trans_amt	percentage_diff	trans_is_fraud
0	83765334	M	Surgeon	Boyer-Haley	travel	12/16/2020 21:16	$16,339.26	$22.96	High Risk	FALSE
1	163011	F	Engineer, control and instrumentation	Kozey-McDermott	travel	12/27/2020 15:58	$22,768.11	$35.46	High Risk	FALSE
2	49735544	F	Higher education careers adviser	Corwin-Romaguera	travel	12/22/2020 21:30	$21,437.71	$50.43	High Risk	FALSE
3	50479522	F	Counsellor	Hagenes, Hermann and Stroman	travel	7/3/2020 18:13	$13,149.15	$32.38	High Risk	FALSE
4	9488913	M	Biomedical engineer	Schroeder, Wolff and Hermiston	travel	10/19/2020 18:11	$12,969.90	$32.89	High Risk	FALSE
...	...	...	...	...	...	...	...	...	...	...
555714	19654975	M	Armed forces training and education officer	Luettgen PLC	gas transport	7/30/2020 0:52	$7.84	$774.14	Within Range	TRUE
555715	87292974	M	Commissioning editor	Huels-Nolan	gas transport	12/2/2020 0:07	$9.16	$955.50	Within Range	TRUE
555716	69296734	M	Historic buildings inspector/conservation officer	Raynor, Feest and Miller	gas transport	9/1/2020 9:19	$7.99	$837.83	Within Range	TRUE
555717	19654975	M	Armed forces training and education officer	Prohaska-Murray	gas transport	7/31/2020 3:43	$7.00	$774.14	Within Range	TRUE
555718	34461572	M	Database administrator	Mraz-Herzog	gas transport	12/22/2020 3:56	$6.60	$867.05	Within Range	TRUE



Query 4

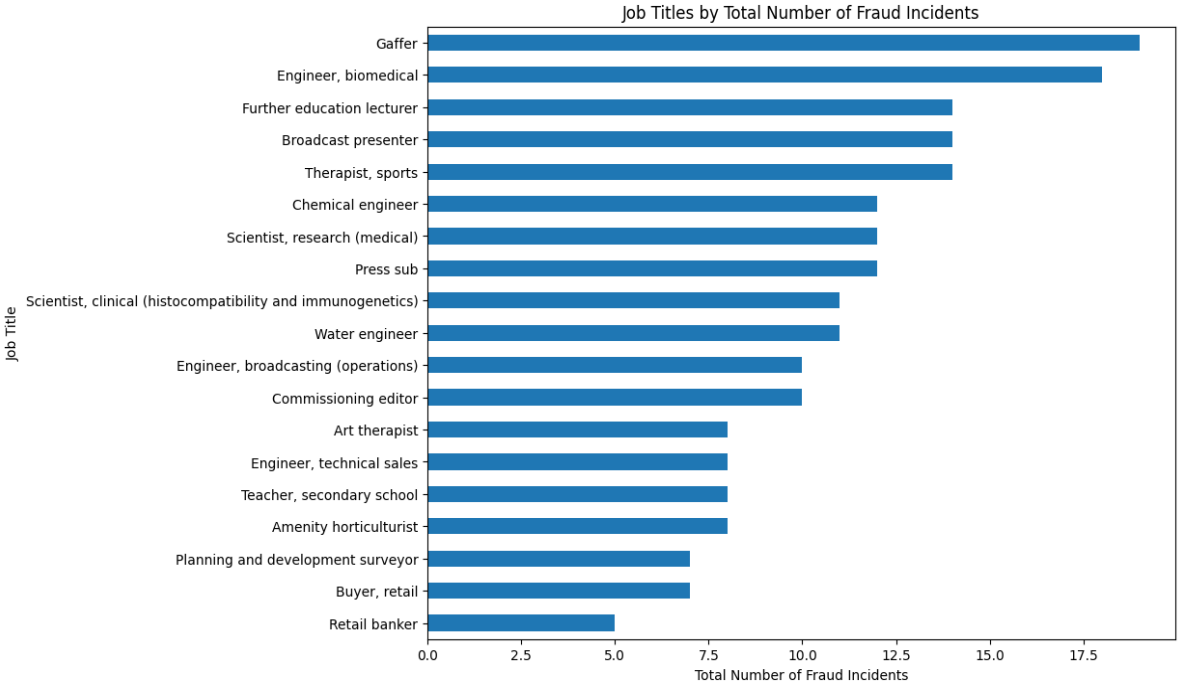
Aggregation – Profiling High-Risk Transactions

trans_category_type	transaction_month	total_fraud_cases	total_male_involved	total_female_involved	med_fraud_amt	merch_name	merch_lat	merch_long
grocery pos	8	6	1	5	$330.49	Cole PLC	40.325204	-79.237092
shopping net	7	6	3	3	$963.38	Jast Ltd	37.963992	-82.926133
shopping net	10	5	1	4	$977.08	Mosciski, Ziemann and Farrell	38.709403	-104.012066
grocery pos	10	5	2	3	$317.37	Padberg-Welch	44.732656	-109.444699
grocery pos	10	5	3	2	$291.24	Kiehn Inc	32.988015	-80.955716
...	...	...	...	...	...	...	...	...
kids pets	10	1	0	1	$20.61	Nolan-Williamson	34.849935	-86.922332
kids pets	10	1	0	1	$21.23	Roberts, Daniel and Macejkovic	31.212812	-85.081825
kids pets	10	1	0	1	$20.73	Streich, Rolfson and Wilderman	39.29604	-76.368333
kids pets	10	1	0	1	$21.09	Weimann-Lockman	41.881426	-83.316455
kids pets	11	1	1	0	$21.55	Berge-Hills	38.653037	-84.419288




```sql
SELECT
    tr.trans_category_type,
    EXTRACT(MONTH FROM tr.trans_date_time) AS
    transaction_month,
    COUNT(CASE WHEN tr.trans_is_fraud = True THEN
    1 END) AS total_fraud_cases,
    COUNT(CASE WHEN tr.c_gender = 'M' THEN 1 END)
    AS total_male_involved,
    COUNT(CASE WHEN tr.c_gender = 'F' THEN 1 END)
    AS total_female_involved,
    PERCENTILE_CONT(0.5) WITHIN GROUP (ORDER BY
    tr.trans_amt) AS med_fraud_amt,
    tr.merch_name,
    m.merch_lat,
    m.merch_long
FROM
    trans_cust_demo tr
    JOIN transaction t ON tr.trans_num =
    t.trans_num
    JOIN merchant m ON t.merchant_id =
    m.merchant_id
WHERE
tr.trans_is_fraud = True
GROUP BY
    tr.trans_category_type, transaction_month,
    tr.merch_name, m.merch_lat, m.merch_long
HAVING
    COUNT(CASE WHEN tr.trans_is_fraud = True THEN
    1 END) >= 1
ORDER BY
    total_fraud_cases DESC;
```

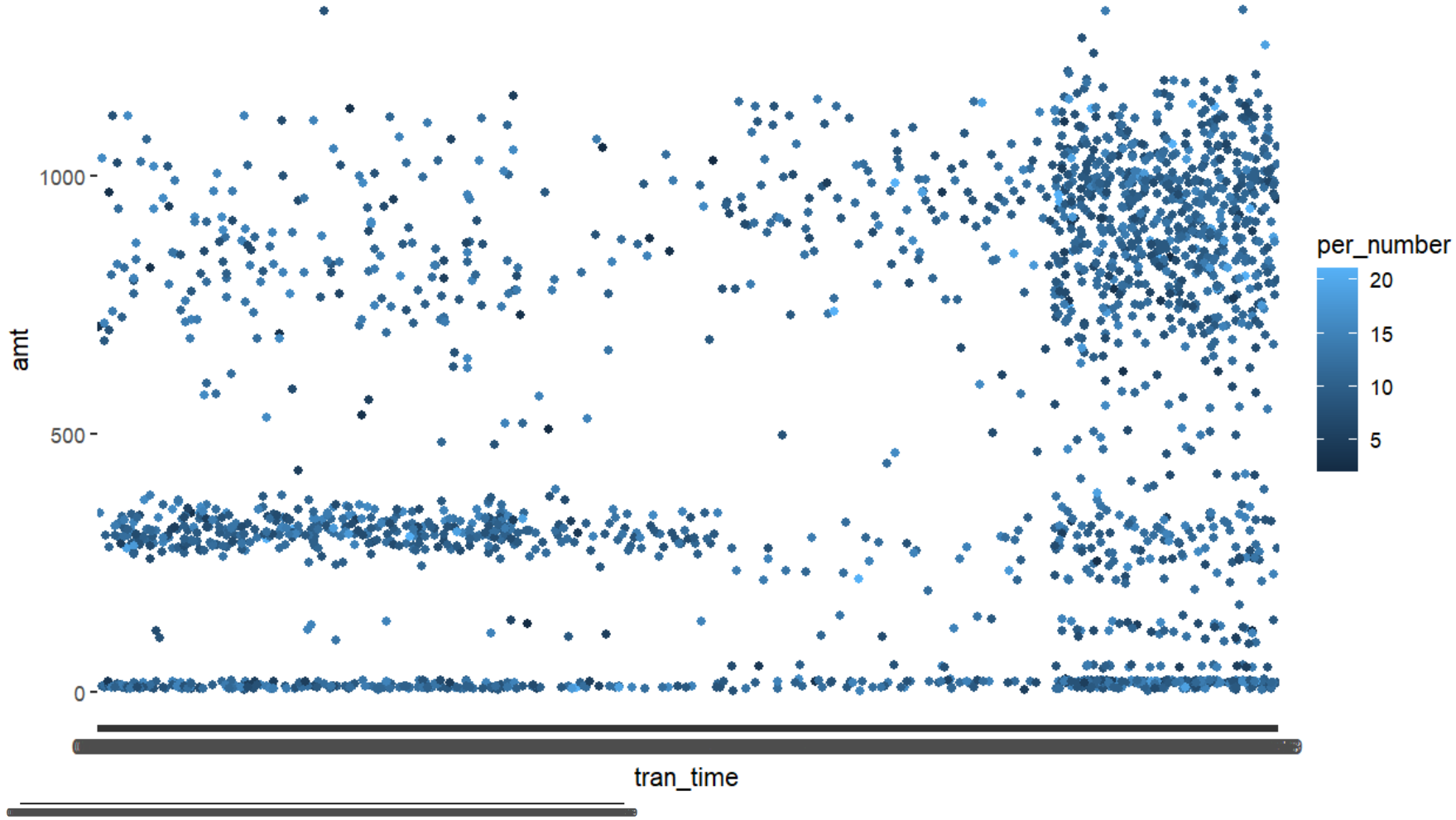
customer_id	age	c_job	c_gender	cs_city	cs_state	total_trans_amt	no_of_fraud
5936811	23	Gaffer	F	Burrton	KS	$12,002.06	19
7303638	56	Further education lecturer	M	Turner	MT	$6,863.58	14
9533425	21	Therapist, sports	F	Smiths Grove	KY	$5,839.30	14
4279521	43	Broadcast presenter	F	Preston	CT	$5,782.20	14
3603551	18	Chemical engineer	M	Morrowville	KS	$7,993.74	12
10580441	52	Engineer, biomedical	F	Stayton	OR	$8,195.87	12
1529474	44	Press sub	F	Thrall	TX	$5,685.68	12
5077091	65	Scientist, research (medical)	M	Belmont	NH	$5,489.25	12
146774	54	Scientist, clinical (histocompatibility	F	Ridgeland	MS	$3,950.54	11

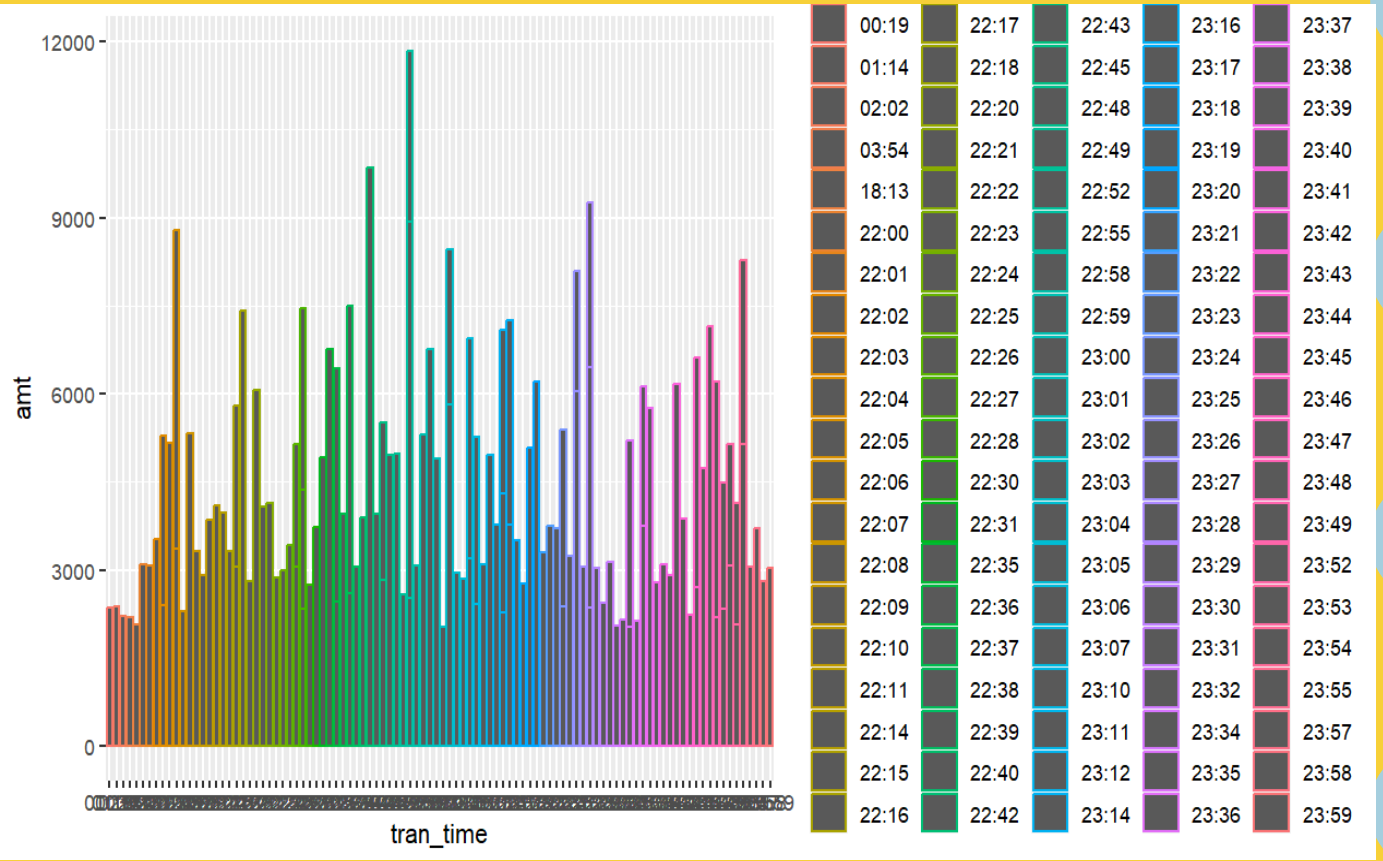


Query 5

Sub-query – Fraud Victims' demographics

```
WITH cust_demo AS (
  SELECT
      c.customer_id,
      EXTRACT(YEAR FROM AGE('2020-12-31', c_dob)) AS
      age,
      c.c_job,
      c.c_gender,
      a.cs_city,
      a.cs_state
  FROM
      customer c
      JOIN address a ON c.customer_id = a.customer_id
  GROUP BY
      c.customer_id, a.cs_city, a.cs_state,
      c.c_gender, c.c_job, age
  ) SELECT
      cd.*,
      SUM(t.trans_amt) AS total_trans_amt,
      COUNT(*) AS no_of_fraud
  FROM
      cust_demo cd
      JOIN transaction t ON cd.customer_id =
      t.customer_id
      JOIN merchant m ON t.merchant_id = m.merchant_id
  WHERE
      t.trans_is_fraud = True
      AND cd.customer_id IN (SELECT customer_id FROM
      top_customers)
  GROUP BY
      cd.customer_id, cd.age, cd.c_job, cd.cs_state,
      cd.cs_city, cd.c_gender
  ORDER BY
      no_of_fraud DESC;
```


amt
1000
500
0
tran_time
per_number
20
15
10
5


12000
9000
6000
3000
0
amt
tran_time
00:19
01:14
02:02
03:54
18:13
22:00
22:01
22:02
22:03
22:04
22:05
22:06
22:07
22:08
22:09
22:10
22:11
22:14
22:15
22:16
22:17
22:18
22:20
22:21
22:22
22:23
22:24
22:25
22:26
22:27
22:28
22:30
22:31
22:35
22:36
22:37
22:38
22:39
22:40
22:42
22:43
22:45
22:48
22:49
22:52
22:55
22:58
22:59
23:00
23:01
23:02
23:03
23:04
23:05
23:06
23:07
23:10
23:11
23:12
23:14
23:16
23:17
23:18
23:19
23:20
23:21
23:22
23:23
23:24
23:25
23:26
23:27
23:28
23:29
23:30
23:31
23:32
23:34
23:35
23:36
23:37
23:38
23:39
23:40
23:41
23:42
23:43
23:44
23:45
23:46
23:47
23:48
23:49
23:52
23:53
23:54
23:55
23:57
23:58
23:59

Performance Tuning

```
SELECT
     tr.trans_category_type,
     EXTRACT(MONTH FROM tr.trans_date_time) AS transaction_month,
     COUNT(CASE WHEN tr.trans_is_fraud = True THEN 1 END) AS total_fraud_cases,
     COUNT(CASE WHEN tr.c_gender = 'M' THEN 1 END) AS total_male_involved,
     COUNT(CASE WHEN tr.c_gender = 'F' THEN 1 END) AS total_female_involved,
     PERCENTILE_CONT(0.5) WITHIN GROUP (ORDER BY tr.trans_amt) AS med_fraud_amt,
     tr.merch_name,
     m.merch_lat,
     m.merch_long
FROM
     trans_cust_demo tr
     JOIN transaction t ON tr.trans_num = t.trans_num
     JOIN merchant m ON t.merchant_id = m.merchant_id
WHERE
     tr.trans_is_fraud = True
GROUP BY
     tr.trans_category_type, transaction_month, tr.merch_name, m.merch_lat, m.merch_long
HAVING COUNT(CASE WHEN tr.trans_is_fraud = True THEN 1 END) >= 1
ORDER BY
     total_fraud_cases DESC;
```

Performance Tuning

WITHOUT INDEX

	Node
1	Sort (cost=21901.04..21902.85 rows=722 width=109) (rows=1445 loops=1)
2	Aggregate (cost=21048.37..21866.76 rows=722 width=109) (rows=1445 loops=1)
	Filter: (count(CASE WHEN tr.trans_is_fraud THEN 1 ELSE NULL::integer END) >= 1)
	Rows Removed by Filter: 0
3	Incremental Sort (cost=21048.37..21772.86 rows=2167 width=88) (rows=2145 loops=1)
4	Nested Loop Inner Join (cost=21048.07..21675.69 rows=2167 width=88) (rows=2145 loops=1)
5	Gather Merge (cost=21047.66..21300.05 rows=2167 width=52) (rows=2145 loops=1)
6	Sort (cost=20047.64..20049.9 rows=903 width=52) (rows=715 loops=3)
7	Nested Loop Inner Join (cost=0.42..20003.31 rows=903 width=52) (rows=715 loops=3)
8	Seq Scan on trans_cust_demo as tr (cost=0..13451.5 rows=903 width=81) (rows=715 loops=3)
	Filter: trans_is_fraud
	Rows Removed by Filter: 184525
9	Index Scan using transaction_pkey on transaction as t (cost=0.42..7.26 rows=1 width=37) (rows=1 loops=2145)
	Index Cond: ((trans_num)::text = (tr.trans_num)::text)
10	Memoize (cost=0.41..0.47 rows=1 width=20) (rows=1 loops=2145)
	Buckets: **Batches**: **Memory Usage**: 66 kB
11	Index Scan using merchant_pkey on merchant as m (cost=0.4..0.46 rows=1 width=20) (rows=1 loops=557)
	Index Cond: (merchant_id = t.merchant_id)

WITH INDEX idx_trans_fraud

	Node
1	Sort (cost=13378.27..13380.07 rows=722 width=109) (rows=1445 loops=1)
2	Aggregate (cost=12525.6..13343.99 rows=722 width=109) (rows=1445 loops=1)
	Filter: (count(CASE WHEN tr.trans_is_fraud THEN 1 ELSE NULL::integer END) >= 1)
	Rows Removed by Filter: 0
3	Incremental Sort (cost=12525.6..13250.09 rows=2167 width=88) (rows=2145 loops=1)
4	Nested Loop Inner Join (cost=12525.3..13152.92 rows=2167 width=88) (rows=2145 loops=1)
5	Gather Merge (cost=12524.89..12777.27 rows=2167 width=52) (rows=2145 loops=1)
6	Sort (cost=11524.87..11527.12 rows=903 width=52) (rows=715 loops=3)
7	Hash Inner Join (cost=414.7..11480.53 rows=903 width=52) (rows=715 loops=3)
	Hash Cond: ((t.trans_num)::text = (tr.trans_num)::text)
8	Seq Scan on transaction as t (cost=0..10188.5 rows=231550 width=37) (rows=185240 loops=3)
9	Hash (cost=387.61..387.61 rows=2167 width=81) (rows=2145 loops=1)
	Buckets: 4096 **Batches**: 1 **Memory Usage**: 290 kB
10	Index Scan using idx_trans_fraud on trans_cust_demo as tr (cost=0.42..387.61 rows=2167 width=81) (rows=2145 loops=1)
	Index Cond: (trans_is_fraud = true)
11	Memoize (cost=0.41..0.47 rows=1 width=20) (rows=1 loops=2145)
	Buckets: **Batches**: **Memory Usage**: 66 kB
12	Index Scan using merchant_pkey on merchant as m (cost=0.4..0.46 rows=1 width=20) (rows=1 loops=557)
	Index Cond: (merchant_id = t.merchant_id)

Conclusion

- The data exploration undertaken has shed light on apparent patterns in fraudulent activity, although it is crucial to note that causation does not imply correlation. This foundational analysis is a step towards more advanced analytics needed to identify the factors that genuinely influence fraud.
- The inclusion of additional data sets could provide further clarity on fraud patterns, such as crime rates in areas where credit card fraud has occurred or specific events that may impact fraudulent activities during certain periods.
- Data manipulation for visualization preparation has been facilitated by the use of Common Table Expressions (CTEs), materialized views, and CASE statements.
- Python has proven to be an effective tool in streamlining the process of extracting and transforming data into meaningful results, demonstrating its utility in simplifying complex analytical tasks.

Thank you!