Memo: Analysis of ETB Theft Claims and Reimbursement Transactions
Audience: CDSS RADD Deputy Director
Author: Steven Archuleta
Date: 18 May 2023

Introduction:

Not unlike the diverse array of residents who constitute California's diffuse demographics, the standard deviation of an aggregated set of values can be measured by assessing the variations that exist within the individuality of the numbers themselves. These variations reflect the degree of inconsistency within the dataset, much like the daily stretching or squeezing of the economic realities faced by low-income beneficiaries of the CDSS Electronic Benefits System. Beneath this well-established distribution of general welfare lies a 21st-century cyber network of criminals who not only pilfer funds from the coffers of the Department of Social Services but also undermine the goodwill of California taxpayers. In this analysis, I am pleased to present the findings of my research on the volume and value of theft claims, as well as reimbursements, before and after the policy change implemented on January 27, 2023. The aim of this analysis is to provide insights into the monetary aspects of theft claims and reimbursements and shed light on the impact of the policy change.

Assumptions and Limitations of the Data:

When working with data and conducting data analysis, it is crucial to consider the assumptions and limitations inherent in the datasets. The raw claims file, with 345,587 observations, and the raw reimbursements file, with 120,020 observations, provide valuable insights, but it is important to acknowledge their limitations. These datasets represent a subset of the overall population and may not capture every possible scenario. Additionally, the data's accuracy, completeness, and reliability can impact the analysis outcomes, and it highlights the importance of the data's volume, value, variety, velocity, and veracity. Data cleaning and preprocessing steps were undertaken to address these factors, but it is essential to interpret the results with an understanding of the assumptions and limitations associated with this particular dataset.

Methodology:

The analysis involved a comprehensive approach that utilized various data analytical tools and methodologies to ensure accurate and meaningful insights. The following tools were employed: Google Colab, Jupyter Notebook, MySQL, and Excel. Each tool served a specific purpose in the analysis workflow.

Google Colab, with its collaborative capabilities and powerful libraries, provided an ideal environment for data exploration, cleaning, preprocessing, and visualization. It facilitated the examination of data distributions through histograms and boxplots, allowing for the identification of outliers using the 1.5*IQR rule.

MySQL, a widely adopted relational database management system, was utilized for executing queries and extracting relevant data for analysis. Its efficient querying capabilities enabled the retrieval of specific data based on defined criteria, enhancing the precision of the analysis.

Excel, a versatile spreadsheet tool, played a vital role in exploring the raw data types and handling .csv files. It offered flexibility in saving and reading data files, contributing to the seamless integration of data processing steps.

The methodology for this analysis comprised two essential steps: Data Cleaning and Preprocessing, and Exploratory Data Analysis. During the Data Cleaning and Preprocessing phase, the raw dataset underwent a comprehensive series of procedures to ensure its suitability for analysis and future machine learning modeling. This phase involved assessing the shape of the dataset and optimizing data types to improve memory usage. Thorough attention was given to addressing missing values, eliminating duplicates, and conducting meticulous exploration of unique feature values. Outliers were handled through imputation or capping techniques, and the dataset was split based on the policy date change of January 27, 2023. Statistical summaries were computed on pre and post policy dataframes, and the datasets were standardized for compatibility with pickle files, facilitating their utilization in various supervised regression models, supervised logistic models, or unsupervised clustering algorithms. These steps played a pivotal role in enhancing data quality, accuracy, and compatibility for subsequent classification and/or predictive modelling. It is essential to emphasize that the effectiveness of these steps hinges on the quality and completeness of the original raw data. By diligently addressing challenges related to data cleaning and preprocessing, the analysis aims to generate reliable and robust insights that will form the basis for informed decision-making processes.

Results:

<u>Finding 1:</u>

TOTAL CLAIMS (raw datasets):

The total value of claims reported before the policy change (pre-policy) was $91,512,819.47

The total value of claims reported after the policy change (post-policy) was $14,669,931.82.

TOTAL CLAIMS (`cleaned datasets`):

```
Total value of claims before the policy change: $ 81,374,247.18
Total value of claims after the policy change:  $ 13,555,037.66
```

TOTAL REIMBURSEMENTS (raw datasets):

The total value of reimbursements reported before the policy change (pre-policy) was $48,896,250.00

The total value of reimbursements reported after the policy change (pre-policy) was $25,368,774.00.

TOTAL REIMBURSEMENTS (`cleaned datasets`):

```
Total value of reimbursements before the policy change: $ 48523597.35
Total value of reimbursements after the policy change: $ 25196071.60
```

The analysis of the cleaned datasets reveals that the total value of claims before the policy change (pre-policy) was $81,374,247.18, which represents a decrease of approximately 11.1% compared to the raw data. Similarly, the total value of claims after the policy change (post-policy) was $13,555,037.66, indicating a reduction of about 7.2%.

In terms of reimbursements, the cleaned datasets show that the total value of reimbursements before the policy change was $48,523,597.35, reflecting a decrease of around 0.8% compared to the raw data. The total value of reimbursements after the policy change was $25,196,071.60, representing a reduction of approximately 0.8%.

Finding 2:

MEDIAN CLAIMS (raw datasets):

The median value of claims reported before the policy change (pre-policy) was $145.73

The median value of claims reported after the policy change (post-policy) was $198.36

AVERAGE CLAIMS (`cleaned datasets`):

```
Average value of claims before the policy change: $285.01
Average value of claims after the policy change:  $321.45
```

	Dataset	count	mean	std	min	25%	50%	75%	max
adj_amt	Pre-Policy Raw Claims	301179.0	303.85	909.28	0.00	44.79	145.73	400.00	99048.0
adj_amt	Post-Policy Raw Claims	44408.0	330.34	546.10	0.01	71.96	198.36	460.61	62093.0

	Dataset	count	mean	std	min	25%	50%	75%	max
adj_amt	Pre-Policy Cleaned Claims	285512.0	285.01	311.81	0.00	42.45	159.64	420.0	986.32
adj_amt	Post-Policy Cleaned Claims	42168.0	321.45	333.48	0.01	71.56	200.00	495.0	1130.16

The statistical findings provide Insights into the effects of the policy change on the adjusted amount of claims (adj_amt).

Looking at the Pre-Policy Raw Claims dataset, before the policy change, the average adjusted amount (adj_amt) was $303.85, with a high variability indicated by a standard deviation of $909.28. After the policy change, in the Post-Policy Raw Claims dataset, the average adjusted amount decreased to $330.34, and the variability decreased as well, as shown by the lower standard deviation of $546.10.

Similarly, in the Pre-Policy Cleaned Claims dataset, before the policy change, the average adjusted amount was $285.01, with a standard deviation of $311.81. After the policy change, in the Post-Policy Cleaned Claims dataset, the average adjusted amount increased slightly to $321.45, and the variability decreased with a lower standard deviation of $333.48.

From these findings, we can infer that the policy change has had an impact on the adjusted amounts of claims. In both the raw and cleaned datasets, there is a decrease in the variability of adjusted amounts after the policy change. This suggests that the policy change may have led to more consistent adjusted amounts in the claims data, as indicated by the decrease in the standard deviation.

However, these findings are based on the analysis of the adj_amt feature alone and may not capture the full impact of the policy change on other aspects of the claims data. Further analysis and consideration of additional factors would be necessary to fully understand the effects of the policy change on claims variability.

MEDIAN REIMBURSEMENTS (raw datasets):

The total value of reimbursements reported before the policy change (pre-policy) was $635.68.

The total value of reimbursements reported after the policy change (pre-policy) was $588.66.

AVERAGE REIMBURSEMENTS (`cleaned datasets`):

```
Total value of reimbursements before the policy change: $ 48523597.35
Total value of reimbursements after the policy change: $ 25196071.60
```

	Dataset	count	mean	std	min	25%	50%	75%	max
benefit_amount	Pre-Policy Raw Reimbursements	76920.0	635.68	409.71	0.01	281.00	658.0	900.0	5000.0
benefit_amount	Post-Policy Raw Reimbursements	43100.0	588.63	430.26	0.01	250.11	516.0	880.0	9000.0

	Dataset	count	mean	std	min	25%	50%	75%	max
benefit_amount	Pre-Policy Cleaned Reimbursements	75810.0	640.068558	399.641462	0.01	295.605	660.0	905.0	1819.0925
benefit_amount	Post-Policy Cleaned Reimbursements	42886.0	587.512746	418.909989	0.01	251.490	516.0	880.0	1822.7650

The statistical findings provide insights into the effects of the policy change on the benefit amount of reimbursements (benefit_amount).

In the Pre-Policy Raw Reimbursements dataset, before the policy change, the average benefit amount was $635.68, with a high variability indicated by a standard deviation of $409.71. After the policy change, in the Post-Policy Raw Reimbursements dataset, the average benefit amount decreased slightly to $588.63, and the variability decreased as well, as shown by the lower standard deviation of $430.26.

Similarly, in the Pre-Policy Cleaned Reimbursements dataset, before the policy change, the average benefit amount was $640.07, with a standard deviation of $399.64. After the policy change, in the Post-Policy Cleaned Reimbursements dataset, the average benefit amount decreased slightly to $587.51, and the variability decreased with a lower standard deviation of $418.91.

From these findings, we can infer that the policy change has had an impact on the benefit amounts of reimbursements. In both the raw and cleaned datasets, there is a decrease in the variability of benefit amounts after the policy change. This suggests that the policy change may have led to more consistent benefit amounts in the reimbursements data, as indicated by the decrease in the standard deviation.

However, these findings are based on the analysis of the benefit_amount feature alone and may not capture the full impact of the policy change on other aspects of the reimbursements data. Further analysis and consideration of additional factors would be necessary to fully understand the effects of the policy change on reimbursements variability.

Finding 3:
Time-based Comparison:
Raw Datasets

CATEGORY	COUNT	TOTAL AMOUNT
Claims - Before Policy Change:	301179	$91,512,819.47
Claims - After Policy Change:	44408	$14,669,931.82
Reimbursements - Before Policy Change:	76919	$48,896,250.00
Reimbursements - After Policy Change:	43096	$25,368,774.00

Cleaned Datasets

	Category	Count	Total Amount
0	Claims - Before Policy Change	285512	$81,374,247.18
1	Claims - After Policy Change	42168	$13,555,037.66
2	Reimbursements - Before Policy Change	75810	$48,523,597.35
3	Reimbursements - After Policy Change	42886	$25,196,071.60

The raw dataset provides information on the volume of claim activity and reimbursement activity before and after the policy change. Before the policy change, there were 301,179 claims filed, resulting in a total amount of $91,512,819.47. After the policy change, the number of claims reduced to 44,408, with a total amount of $14,669,931.82. Similarly, for reimbursements, there were 76,919 reimbursements before the policy change, totaling $48,896,250.00. After the policy change, the number of reimbursements decreased to 43,096, with a total amount of $25,368,774.00.

The cleaned dataset provides a more refined view, taking into account the specific data points related to the claims and reimbursements. Before the policy change, there were 285,512 claims, resulting in a total amount of $81,374,247.18. After the policy change, the number of claims decreased to 42,168, with a total amount of $13,555,037.66. Similarly, for reimbursements, there were 75,810 reimbursements before the policy change, totaling $48,523,597.35. After the policy change, the number of reimbursements reduced to 42,886, with a total amount of $25,196,071.60.

Given that the post-policy period only includes three months of data compared to a longer period for the pre-policy period, it is important to interpret the findings with caution. A year-over-year analysis may provide a more comprehensive understanding of the impacts of the policy change on the volume of claims and reimbursements. Additionally, due to the complexity of the relationships between claims and reimbursements, further exploration using machine learning algorithms could offer valuable insights.

Finding 4:

Temporal by Date

Raw Datasets

DATE	COUNT	AMOUNT	CATEGORY	POLICY PERIOD
2	39800	21148473.74	Claims	Before Policy Change
1	32924	17351818.79	Claims	Before Policy Change
3	32749	15826408.76	Claims	Before Policy Change
10	13304	2715230.31	Claims	Before Policy Change
9	13118	2561357.63	Claims	Before Policy Change
8	13109	2672738.92	Claims	Before Policy Change
6	13054	2833728.45	Claims	Before Policy Change
4	13045	2642380.24	Claims	Before Policy Change
5	13010	2640267.46	Claims	Before Policy Change
7	12982	2604928.22	Claims	Before Policy Change
11	8873	1613984.65	Claims	Before Policy Change
12	8126	1648912.53	Claims	Before Policy Change
13	7394	1327469.6	Claims	Before Policy Change
15	6332	1049697.75	Claims	Before Policy Change
16	6184	1068442.22	Claims	Before Policy Change
14	6107	1170808.54	Claims	Before Policy Change
18	5997	1007335.91	Claims	Before Policy Change
17	5572	825830.21	Claims	Before Policy Change
19	5246	832878.22	Claims	Before Policy Change
23	4953	810636.62	Claims	Before Policy Change
22	4869	911274.41	Claims	Before Policy Change
21	4827	714185.76	Claims	Before Policy Change
20	4452	728838.75	Claims	Before Policy Change
24	3570	547634.65	Claims	Before Policy Change
30	3511	789004.38	Claims	Before Policy Change
29	3466	737656.98	Claims	Before Policy Change
25	3331	494648.79	Claims	Before Policy Change
26	3279	644835.28	Claims	Before Policy Change
28	3180	598719.34	Claims	Before Policy Change
27	2780	506728.23	Claims	Before Policy Change
31	2035	485964.13	Claims	Before Policy Change

	DATE	COUNT	AMOUNT	CATEGORY	PAY PERIOD
2		6534	3493356.11	Claims	After Policy Change
1		5729	3004079.97	Claims	After Policy Change
3		5052	2402392.93	Claims	After Policy Change
6		2563	554809.27	Claims	After Policy Change
7		2367	481899.59	Claims	After Policy Change
9		2218	502853.61	Claims	After Policy Change
5		2194	427281.24	Claims	After Policy Change
10		2027	455730.65	Claims	After Policy Change
4		2022	409693.42	Claims	After Policy Change
8		2011	347870.58	Claims	After Policy Change
11		1246	215828.58	Claims	After Policy Change
13		1098	217640.78	Claims	After Policy Change
12		962	203305.93	Claims	After Policy Change
30		884	319395.3	Claims	After Policy Change
27		875	176032.37	Claims	After Policy Change
31		821	254155.99	Claims	After Policy Change
29		810	236357.04	Claims	After Policy Change
14		638	102157.8	Claims	After Policy Change
28		633	138570.18	Claims	After Policy Change
15		563	108322.29	Claims	After Policy Change
16		560	99076.69	Claims	After Policy Change
19		465	93626.81	Claims	After Policy Change
17		442	77546.99	Claims	After Policy Change
18		427	90458.13	Claims	After Policy Change
20		251	45849.69	Claims	After Policy Change
21		229	43595.49	Claims	After Policy Change
26		219	64188.17	Claims	After Policy Change
23		163	25176.27	Claims	After Policy Change
25		162	35474.18	Claims	After Policy Change
22		134	25038.85	Claims	After Policy Change
24		109	18166.92	Claims	After Policy Change

DATE	COUNT	AMOUNT	CATEGORY	POLICY PERIOD
19	3493	2289171	Reimbursements	Before Policy Change
15	3355	2114446	Reimbursements	Before Policy Change
9	3320	2235217	Reimbursements	Before Policy Change
21	3309	2022599	Reimbursements	Before Policy Change
13	3229	2140776	Reimbursements	Before Policy Change
18	3195	2127115	Reimbursements	Before Policy Change
23	3185	1875610	Reimbursements	Before Policy Change
16	3159	2038269	Reimbursements	Before Policy Change
14	3111	2047362	Reimbursements	Before Policy Change
20	3089	2022850	Reimbursements	Before Policy Change
12	3065	2074173	Reimbursements	Before Policy Change
22	3034	1832232	Reimbursements	Before Policy Change
17	2917	1835718	Reimbursements	Before Policy Change
8	2665	1800775	Reimbursements	Before Policy Change
10	2591	1707297	Reimbursements	Before Policy Change
7	2584	1689509	Reimbursements	Before Policy Change
6	2407	1628989	Reimbursements	Before Policy Change
28	2217	1348402	Reimbursements	Before Policy Change
26	2160	1268895	Reimbursements	Before Policy Change
29	2098	1198758	Reimbursements	Before Policy Change
11	2059	1457119	Reimbursements	Before Policy Change
5	2054	1393384	Reimbursements	Before Policy Change
25	1945	1172345	Reimbursements	Before Policy Change
30	1939	1060881	Reimbursements	Before Policy Change
27	1870	1133292	Reimbursements	Before Policy Change
24	1696	1025700	Reimbursements	Before Policy Change
2	1654	994972	Reimbursements	Before Policy Change
4	1638	1005081	Reimbursements	Before Policy Change
3	1522	951899	Reimbursements	Before Policy Change
1	1511	949362	Reimbursements	Before Policy Change
31	848	454052	Reimbursements	Before Policy Change

DATE	COUNT	AMOUNT	CATEGORY	POLICY PERIOD
10	2363	1438088	Reimbursements	After Policy Change
16	2202	1300968	Reimbursements	After Policy Change
9	2200	1405894	Reimbursements	After Policy Change
8	2120	1312350	Reimbursements	After Policy Change
22	2113	1192907	Reimbursements	After Policy Change
15	2091	1266833	Reimbursements	After Policy Change
14	2056	1172922	Reimbursements	After Policy Change
17	2051	1167114	Reimbursements	After Policy Change
7	2020	1336970	Reimbursements	After Policy Change
27	2009	1115921	Reimbursements	After Policy Change
3	1926	1213773	Reimbursements	After Policy Change
23	1856	1100591	Reimbursements	After Policy Change
21	1793	1070358	Reimbursements	After Policy Change
13	1774	1073950	Reimbursements	After Policy Change
28	1706	959379	Reimbursements	After Policy Change
6	1665	1089084	Reimbursements	After Policy Change
24	1625	941495	Reimbursements	After Policy Change
2	1612	938494	Reimbursements	After Policy Change
30	1585	864303	Reimbursements	After Policy Change
1	1330	701121	Reimbursements	After Policy Change
29	1034	507898	Reimbursements	After Policy Change
31	995	531824	Reimbursements	After Policy Change
20	942	543377	Reimbursements	After Policy Change
18	641	362065	Reimbursements	After Policy Change
4	411	254095	Reimbursements	After Policy Change
25	274	166738	Reimbursements	After Policy Change
11	229	115620	Reimbursements	After Policy Change
19	198	98123	Reimbursements	After Policy Change
12	152	70986	Reimbursements	After Policy Change
26	83	36490	Reimbursements	After Policy Change
5	40	19043	Reimbursements	After Policy Change

Temporal by Date

Cleaned Datasets

	DATE	COUNT	AMOUNT	CATEGORY	POLICY PERIOD
1	2	37660	1.916976e+07	Claims	Before Policy Change
0	1	31366	1.605140e+07	Claims	Before Policy Change
2	3	30779	1.434405e+07	Claims	Before Policy Change
9	10	12723	2.353139e+06	Claims	Before Policy Change
7	8	12523	2.330626e+06	Claims	Before Policy Change
8	9	12491	2.349499e+06	Claims	Before Policy Change
4	5	12428	2.327090e+06	Claims	Before Policy Change
5	6	12257	2.301271e+06	Claims	Before Policy Change
3	4	12244	2.272919e+06	Claims	Before Policy Change
6	7	12163	2.272820e+06	Claims	Before Policy Change
10	11	8469	1.406288e+06	Claims	Before Policy Change
11	12	7652	1.345255e+06	Claims	Before Policy Change
12	13	6940	1.116303e+06	Claims	Before Policy Change
14	15	5980	9.493908e+05	Claims	Before Policy Change
15	16	5907	9.688010e+05	Claims	Before Policy Change
13	14	5827	9.244922e+05	Claims	Before Policy Change
17	18	5747	8.477631e+05	Claims	Before Policy Change
16	17	5313	7.744622e+05	Claims	Before Policy Change
18	19	5046	6.959147e+05	Claims	Before Policy Change
22	23	4684	6.777853e+05	Claims	Before Policy Change
20	21	4627	6.497966e+05	Claims	Before Policy Change
21	22	4591	6.498902e+05	Claims	Before Policy Change
19	20	4301	6.115485e+05	Claims	Before Policy Change
23	24	3468	4.672895e+05	Claims	Before Policy Change
28	29	3325	6.479707e+05	Claims	Before Policy Change

	DATE	COUNT	AMOUNT	CATEGORY	POLICY PERIOD
1	2	37660	1.916976e+07	Claims	Before Policy Change
0	1	31366	1.605140e+07	Claims	Before Policy Change
2	3	30779	1.434405e+07	Claims	Before Policy Change
9	10	12723	2.353139e+06	Claims	Before Policy Change
7	8	12523	2.330626e+06	Claims	Before Policy Change
8	9	12491	2.349499e+06	Claims	Before Policy Change
4	5	12428	2.327090e+06	Claims	Before Policy Change
5	6	12257	2.301271e+06	Claims	Before Policy Change
3	4	12244	2.272919e+06	Claims	Before Policy Change
6	7	12163	2.272820e+06	Claims	Before Policy Change
10	11	8469	1.406288e+06	Claims	Before Policy Change
11	12	7652	1.345255e+06	Claims	Before Policy Change
12	13	6940	1.116303e+06	Claims	Before Policy Change
14	15	5980	9.493908e+05	Claims	Before Policy Change
15	16	5907	9.688010e+05	Claims	Before Policy Change
13	14	5827	9.244922e+05	Claims	Before Policy Change
17	18	5747	8.477631e+05	Claims	Before Policy Change
16	17	5313	7.744622e+05	Claims	Before Policy Change
18	19	5046	6.959147e+05	Claims	Before Policy Change
22	23	4684	6.777853e+05	Claims	Before Policy Change
20	21	4627	6.497966e+05	Claims	Before Policy Change
21	22	4591	6.498902e+05	Claims	Before Policy Change
19	20	4301	6.115485e+05	Claims	Before Policy Change
23	24	3468	4.672895e+05	Claims	Before Policy Change
28	29	3325	6.479707e+05	Claims	Before Policy Change

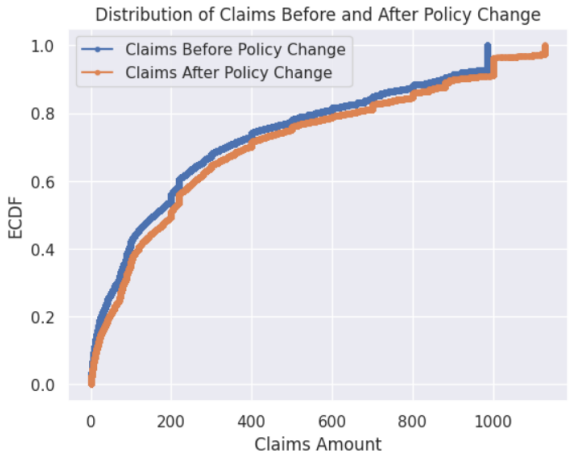
	DATE	COUNT	AMOUNT	CATEGORY	POLICY PERIOD
18	19	3458	2.269802e+06	Reimbursements	Before Policy Change
14	15	3304	2.098581e+06	Reimbursements	Before Policy Change
8	9	3273	2.223630e+06	Reimbursements	Before Policy Change
20	21	3253	2.005508e+06	Reimbursements	Before Policy Change
12	13	3192	2.123984e+06	Reimbursements	Before Policy Change
17	18	3153	2.110646e+06	Reimbursements	Before Policy Change
22	23	3128	1.862174e+06	Reimbursements	Before Policy Change
15	16	3127	2.025851e+06	Reimbursements	Before Policy Change
13	14	3077	2.033875e+06	Reimbursements	Before Policy Change
19	20	3049	2.005694e+06	Reimbursements	Before Policy Change
11	12	3034	2.061664e+06	Reimbursements	Before Policy Change
21	22	2965	1.812564e+06	Reimbursements	Before Policy Change
16	17	2869	1.827657e+06	Reimbursements	Before Policy Change
7	8	2636	1.788093e+06	Reimbursements	Before Policy Change
9	10	2539	1.691181e+06	Reimbursements	Before Policy Change
6	7	2526	1.675330e+06	Reimbursements	Before Policy Change
5	6	2371	1.620712e+06	Reimbursements	Before Policy Change
27	28	2184	1.338022e+06	Reimbursements	Before Policy Change
25	26	2133	1.260803e+06	Reimbursements	Before Policy Change
28	29	2047	1.181470e+06	Reimbursements	Before Policy Change
10	11	2042	1.444922e+06	Reimbursements	Before Policy Change
4	5	2038	1.381261e+06	Reimbursements	Before Policy Change
24	25	1917	1.164459e+06	Reimbursements	Before Policy Change
29	30	1907	1.053704e+06	Reimbursements	Before Policy Change
26	27	1846	1.121858e+06	Reimbursements	Before Policy Change

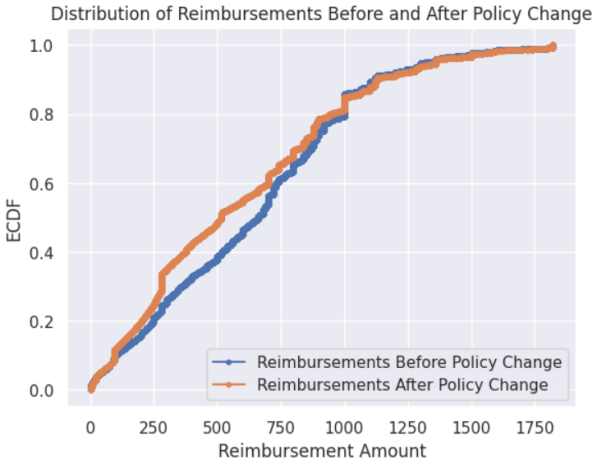
	DATE	COUNT	AMOUNT	CATEGORY	POLICY PERIOD
9	10	2355	1429502.520	Reimbursements	After Policy Change
8	9	2193	1396549.165	Reimbursements	After Policy Change
15	16	2189	1286833.380	Reimbursements	After Policy Change
7	8	2110	1303450.590	Reimbursements	After Policy Change
21	22	2098	1188798.290	Reimbursements	After Policy Change
14	15	2088	1259921.170	Reimbursements	After Policy Change
16	17	2048	1164483.290	Reimbursements	After Policy Change
13	14	2042	1163631.700	Reimbursements	After Policy Change
6	7	2012	1326421.085	Reimbursements	After Policy Change
26	27	1998	1112412.020	Reimbursements	After Policy Change
2	3	1914	1200042.055	Reimbursements	After Policy Change
22	23	1853	1089482.670	Reimbursements	After Policy Change
20	21	1785	1060020.875	Reimbursements	After Policy Change
12	13	1770	1068397.845	Reimbursements	After Policy Change
27	28	1688	950638.375	Reimbursements	After Policy Change
5	6	1652	1081923.375	Reimbursements	After Policy Change
23	24	1624	937762.000	Reimbursements	After Policy Change
1	2	1607	931325.515	Reimbursements	After Policy Change
29	30	1575	859137.890	Reimbursements	After Policy Change
0	1	1324	692507.995	Reimbursements	After Policy Change
28	29	1015	503459.360	Reimbursements	After Policy Change
30	31	982	528500.010	Reimbursements	After Policy Change
19	20	937	539772.880	Reimbursements	After Policy Change
17	18	640	361015.520	Reimbursements	After Policy Change
3	4	411	253919.825	Reimbursements	After Policy Change

Finding 5:

Estimator of the Cumulative Distribution Function

Cleaned Datasets





To compare the distribution of Claims before and after the policy change, ECDF plots can be utilized. An ECDF (Empirical Cumulative Distribution Function) represents the cumulative probability distribution of a dataset. For the Claims feature, two ECDFs can be generated - one for Claims before the policy change and another for Claims after the policy change. The x-axis of the ECDF represents the Claims amount, while the y-axis represents the cumulative probability.

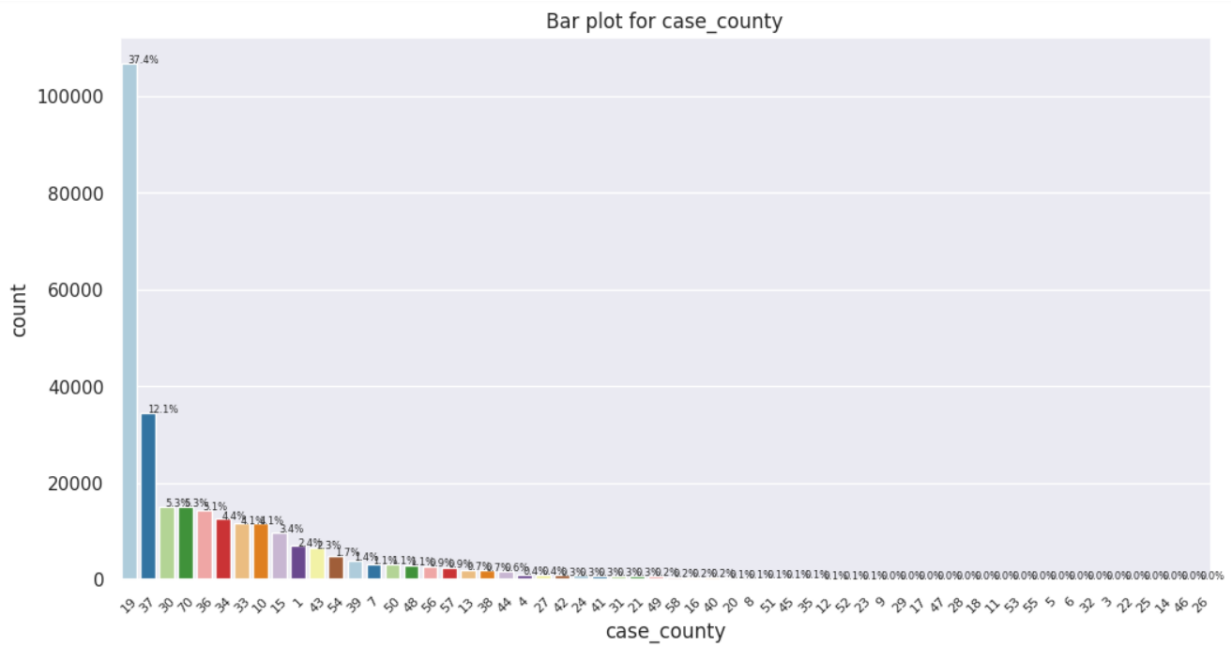
Similarly, the distribution of Reimbursements before and after the policy change can also be compared using ECDF plots. By computing the ECDF for the Reimbursements feature, the cumulative probability distribution for Reimbursements before the policy change and Reimbursements after the policy change can be visualized. This allows the observation of any differences in the distribution patterns.

When interpreting these ECDF plots, it is important to understand the axes. The x-axis represents the variable being analyzed, such as Claims amount or Reimbursements amount. The y-axis represents the cumulative probability, indicating the proportion of values that are less than or equal to a certain value on the x-axis.

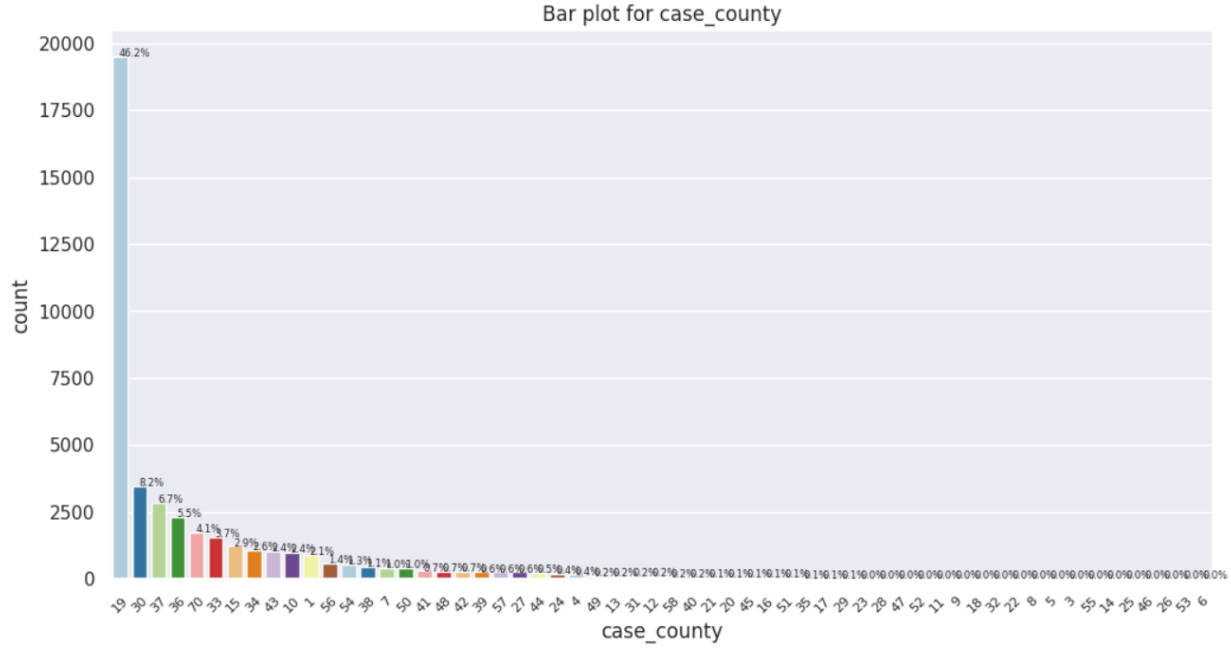
Additionally, continuous numeric variables like adj_amt and benefit_amount can be considered as alternatives to the Claims and Reimbursements features. By comparing the ECDFs of these variables between the two subsets (before and after the policy change), valuable insights can be gained regarding the differences in these metrics across the two time periods.

Finding 6:
Cases by Counties
Cleaned Datasets

Pre-Policy Claims



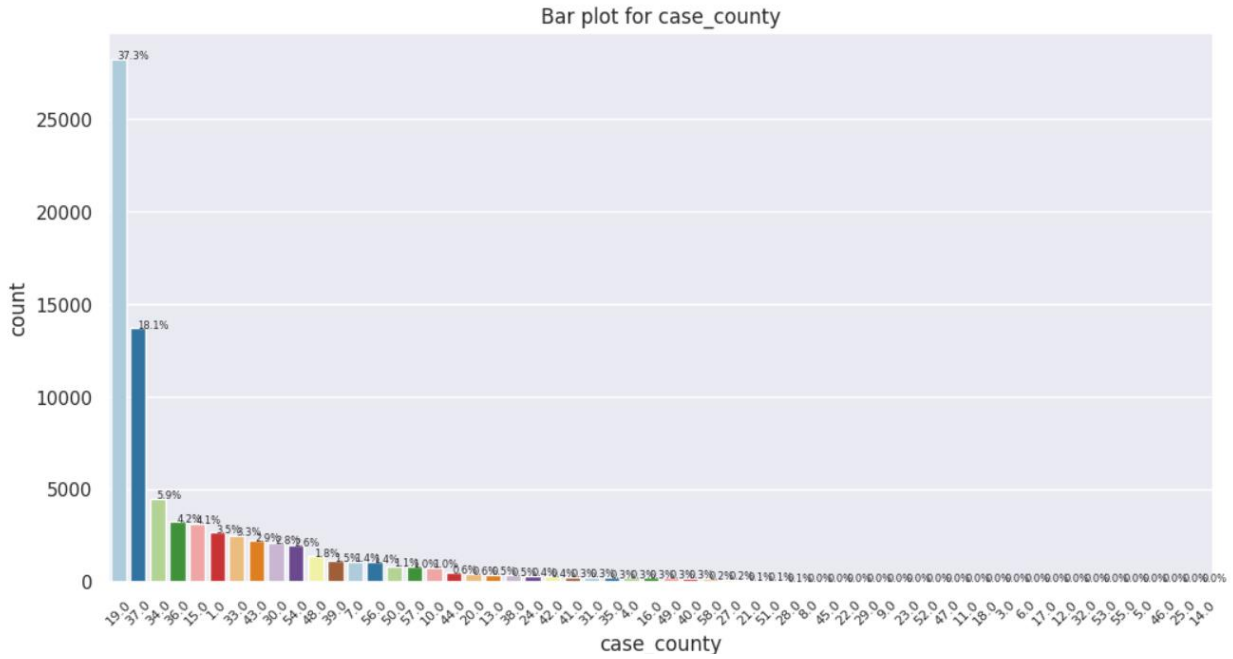
Post-Policy Claims



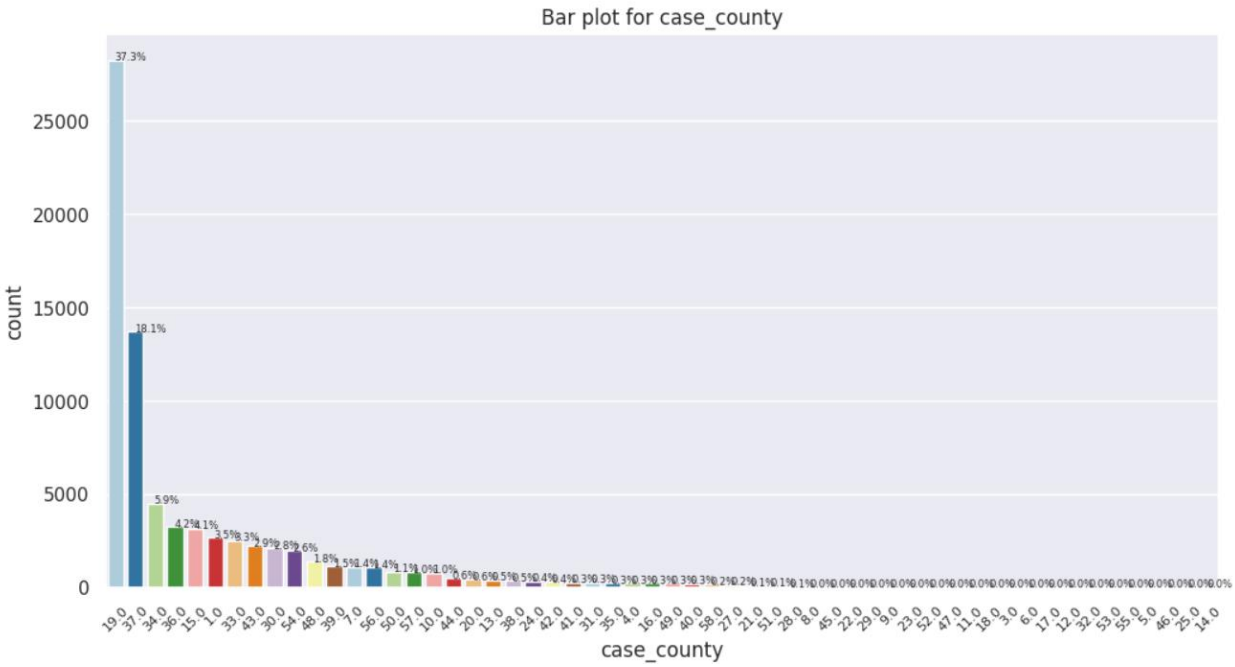
Cases by Counties

Cleaned Datasets

Pre-Policy Reimbursements

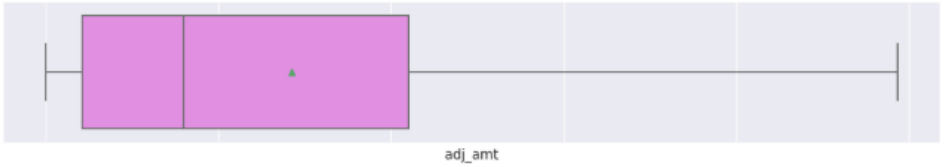


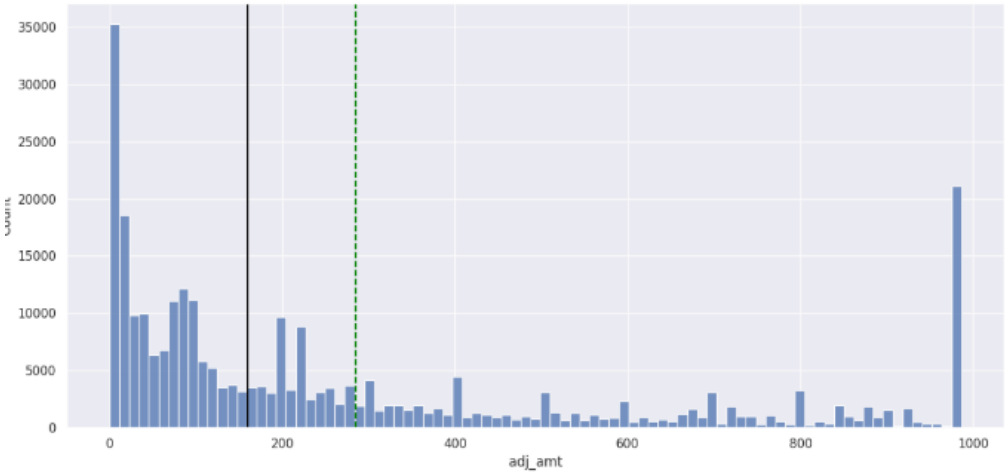
Post-Policy Reimbursements



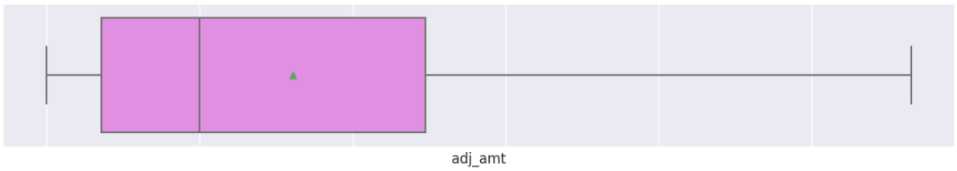
Finding 7:
Distribution

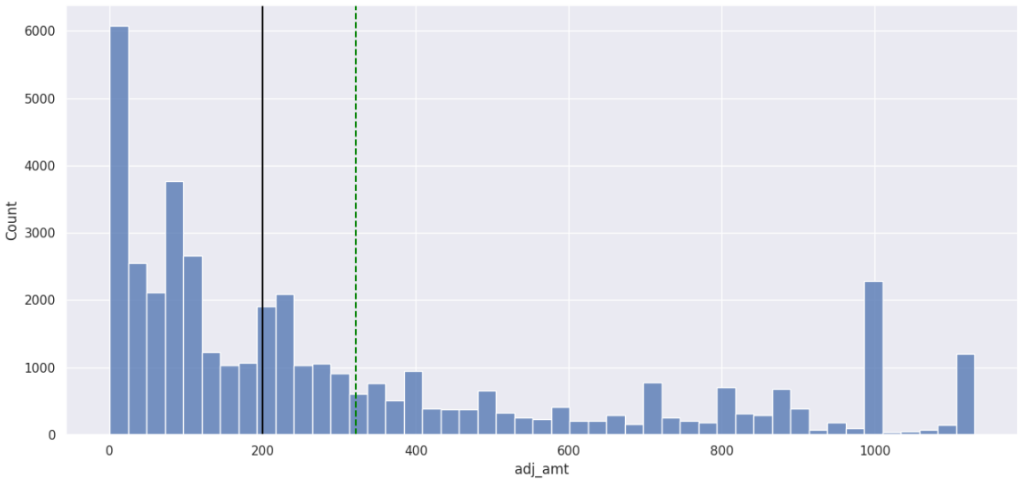
Cleaned Datasets
Pre-Policy Claims
1.5 x IQR Capped; No Outliers



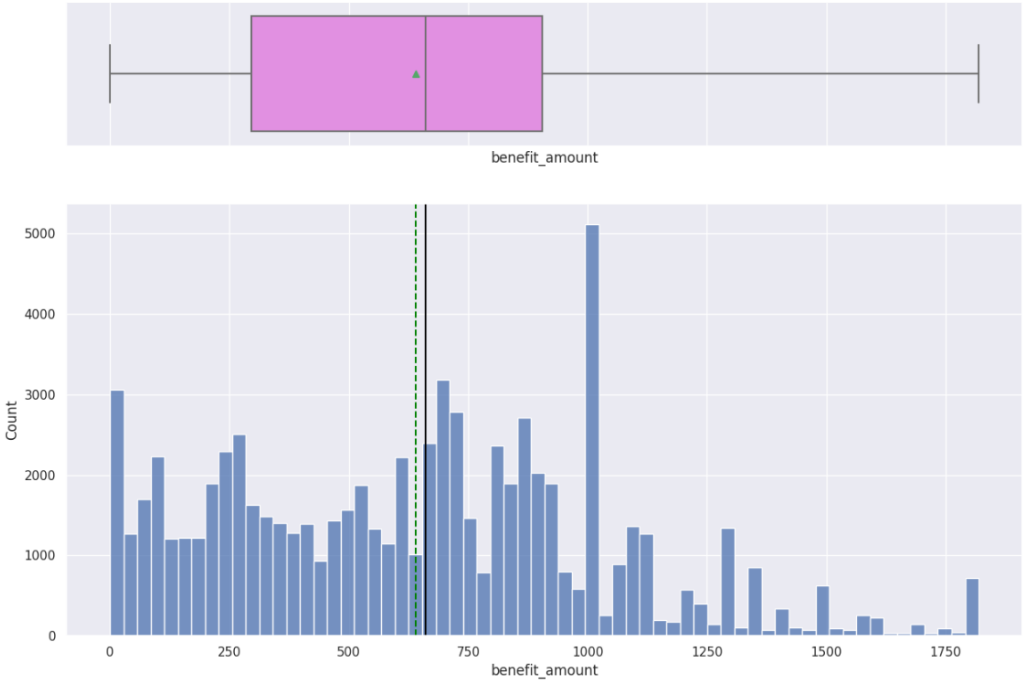


Cleaned Datasets
Post_Policy Claims
1.5 x IQR Capped; No Outliers

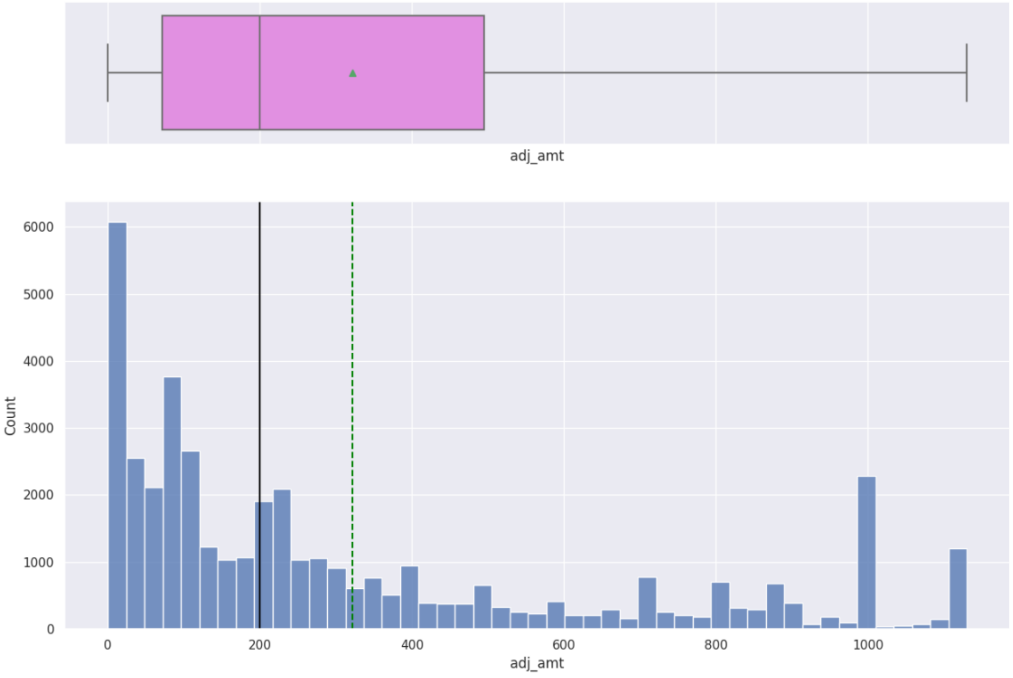




Cleaned Datasets
Pre-Policy Reimbursements
1.5 x IQR Capped; No Outliers



Cleaned Datasets
Post_Policy Claims
1.5 x IQR Capped; No Outliers



Conclusions and Recommendations:

Based on the analysis conducted, several key findings emerge regarding the effects of the policy change on claims and reimbursements. The cleaned datasets reveal a decrease in the total value of claims and reimbursements after the policy change, indicating a potential impact of the policy on the overall volume of activity. Furthermore, the analysis of median and average claims and reimbursements suggests variations in the adjusted amounts and benefit amounts, indicating potential changes in the distribution of these values.

The findings are complemented by the histogram and boxplots, which provide a visual representation of the distribution of claim and reimbursement values before and after the policy change. The histogram illustrates the frequency of values within specific ranges, while the boxplots depict the median, quartiles, and outliers. By examining these visualizations alongside the summary statistics, a comprehensive understanding of the impact of the policy change on the distribution of claims and reimbursements can be obtained.

It is important to note that the observed changes in claims and reimbursements may be influenced by factors beyond the policy change. Other factors such as changes in reporting behavior, social ethnic and socioeconomic trends, or external events might have contributed to the observed patterns. Therefore, it is essential to consider these factors when interpreting the results and drawing conclusions.

Regarding data reliability and representativeness, it is fair to acknowledge that the findings are based on the assumption that the provided dataset is reliable and representative of the entire population. However, potential data quality issues or biases within the dataset could affect the accuracy and generalizability of the results. Therefore, caution should be exercised when drawing conclusions based solely on the analyzed dataset.

In light of these findings, several recommendations can be made for further analysis and decision-making:

Conduct a year-over-year analysis: Since the post-policy period consists of only three months of data, it is recommended to perform a year-over-year analysis to compare the volume and value of claims and reimbursements. This will provide a more comprehensive understanding of the long-term effects of the policy change.

Evaluate data quality and completeness: Given the potential impact of data quality on the analysis results, it is recommended to assess the reliability and completeness of the dataset. Conducting data quality checks, identifying potential biases, and addressing any data issues will enhance the accuracy and robustness of the analysis.

Apply advanced analytical techniques: To gain more insights from the dataset, advanced analytical techniques such as machine learning algorithms can be employed. These techniques can uncover complex relationships and patterns in the data, providing a more in-depth understanding of the impact of the policy change on claims and reimbursements.

These insights, herein, can inform decision-making processes and guide future policies or interventions in the context of theft claims and reimbursements.