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:  
  
Finding 1:   
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  
  
A picture containing text, font, screenshot, line

Description automatically generated

A picture containing text, font, screenshot, line

Description automatically generated

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

A screenshot of a computer

Description automatically generated with low confidence

A close-up of numbers

Description automatically generated with low confidence

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**  
A picture containing text, font, screenshot, line

Description automatically generated  
**Cleaned Datasets**  
A picture containing text, screenshot, font, line

Description automatically generated

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**  
A picture containing text, document, font, screenshot

Description automatically generated A screenshot of a computer

Description automatically generated with low confidence

**A screenshot of a computer

Description automatically generated with low confidence A picture containing text, screenshot, document, font

Description automatically generated**

Temporal by Date **Cleaned Datasets  
A picture containing text, menu, document, screenshot

Description automatically generated A picture containing text, menu, document, screenshot

Description automatically generated**

**A screenshot of a computer

Description automatically generated with low confidence A picture containing text, document, number, paper

Description automatically generated**

Finding 5:  
Estimator of the Cumulative Distribution Function  
**Cleaned Datasets**A picture containing text, screenshot, plot, line

Description automatically generatedA picture containing text, screenshot, plot, line

Description automatically generated

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   
A picture containing text, screenshot, plot, line

Description automatically generated

Post-Policy Claims

A picture containing text, screenshot, display, plot

Description automatically generated

Cases by Counties  
**Cleaned Datasets**Pre-Policy Reimbursements

A picture containing text, plot, line, screenshot

Description automatically generated

Post-Policy Reimbursements  
A picture containing text, screenshot, plot, diagram

Description automatically generated

Finding 7:  
Distribution

**Cleaned Datasets  
Pre-Policy Claims  
1.5 x IQR Capped; No Outliers**A picture containing text, screenshot, plot, diagram

Description automatically generated

**Cleaned Datasets  
Post\_Policy Claims  
1.5 x IQR Capped; No Outliers**

A picture containing screenshot, diagram, text, plot

Description automatically generated

**Cleaned Datasets  
Pre-Policy Reimbursements  
1.5 x IQR Capped; No Outliers  
A picture containing screenshot, plot, diagram, line

Description automatically generated**

**Cleaned Datasets  
Post\_Policy Claims  
1.5 x IQR Capped; No Outliers  
A picture containing screenshot, diagram, text, plot

Description automatically generated**

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.