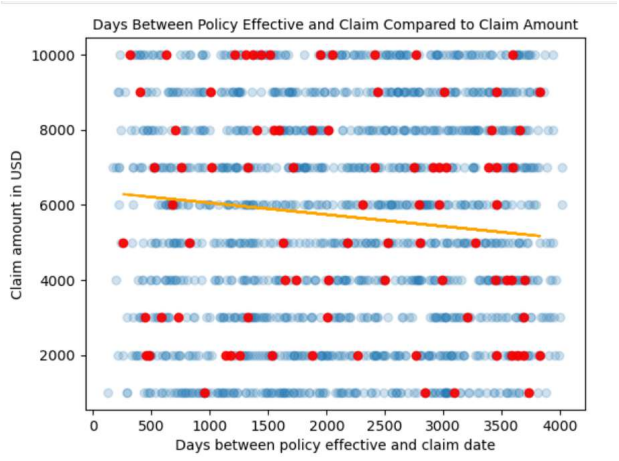
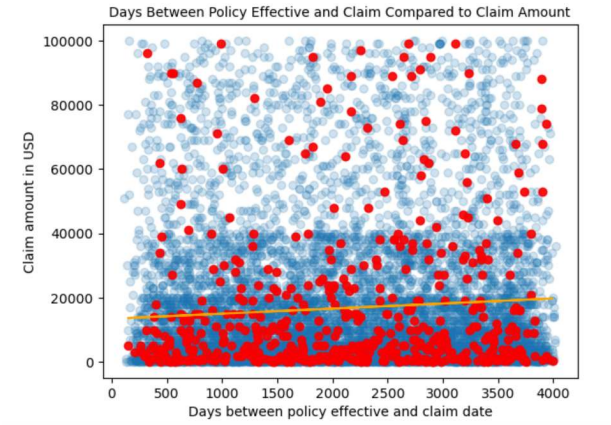
Detecting Vehicle Insurance Fraud
By: Abdul, Tyger, and Sara

Group Write-Up

For our data we used a set of datasets consisting of information on insurance agents, Insurance policies/claims, and insurance vendors that we found on Kaggle. These datasets were rated ten out of ten on kaggle so we believe they should be accurate. These data sets contained information such as the Names and IDs for the agents and vendors. For the Polices the dataset included information on the type of insurance policy, when the policy became effective, the date of the loss for the claim, the amount of money requested for the claim, the amount of damage for the accident, if the claim was approved or not, the agent and vendor associated with the claim, the location that the accident took place in as well as some information about the customer. We didn't really need to do that much data cleaning with the only things of note being to specify that we are researching motor insurance so we had to filter out all the non motor insurance claims.

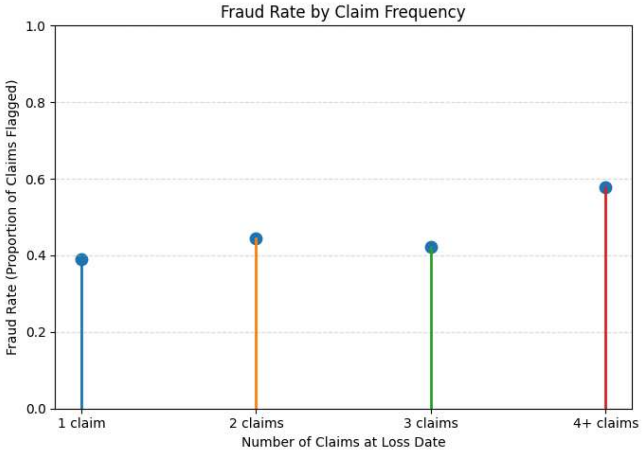
Tygers Question: One question we thought to ask was if the time between policy affective date and loss date affect the claim amount or the claim status. Our suspected answer for this question would be yes with higher claim amounts in less time resulting in more denied claims. So to test this we started by taking the start and end dates and finding the amount of days between them to do this we took the year of the data and multiplied it by the number of days in a year and the month of the day and multiplied that by thirty and added all of that together with the date for the date to get an estimate of the total number of days between the claim effective and loss date. From there we compared it to the claim amount. We then did this process a second time but only for the claims that had been denied. This produced the following graph on the right. From this graph we can see that there is a slight correlation between number of days and claim amount with the claim status. With more denied claims for an earlier number of days with a larger claim amount. This would then seem to prove our hypothesis at least slightly. The trend seems pretty minimal and if we perform this same test for all types of insurance not just motor insurance we see that the trend is the opposite with denied requests increasing as the number of days increases. With this in mind I would say there really isn't a correlation here and insurance companies are looking at the days to claim amount data to try and detect fraud.





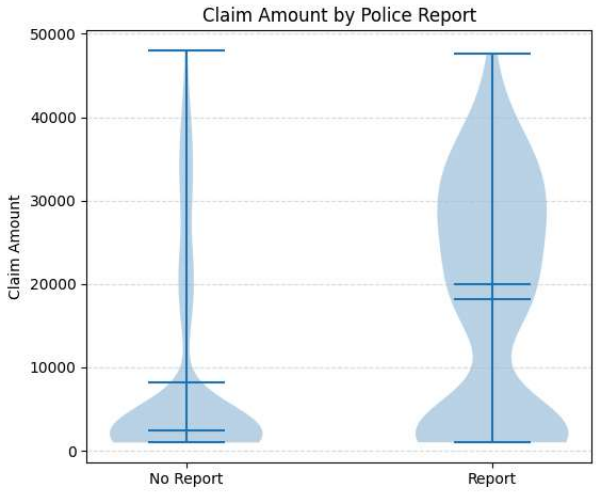
Abdul's Questions:

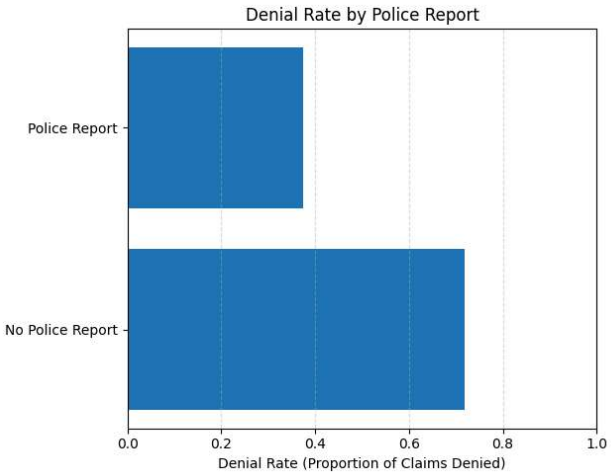
My two questions focused on different potential indicators of suspicious insurance claims. The first question I wanted to explore was whether there is a relationship between claim frequency (how many claims a customer has made) and the likelihood of fraud. My initial thought was that customers who file a lot of claims might naturally draw more attention from insurance companies, and I assumed that higher claim frequency would be associated with a higher fraud rate. To test this, I used the *CLAIMS_AT_LOSS_DATE* variable, which tells us how many claims the customer had made up to the point of the current claim. I grouped these into four categories: customers with 1 claim, 2 claims, 3 claims, and those with 4 or more claims. After doing this, I calculated what percentage of each group had been flagged for fraud investigation.



The results turned out to support the idea somewhat. Customers with only one claim had a fraud rate of around **39%**, and the rate increased slightly for customers with two or three claims. The biggest jump came with customers who had four or more claims, where the fraud rate increased to almost **60%**. I turned this into a lollipop graph, and from the visual you can see that there's a noticeable upward trend. While the increase between 1–3 claims isn't huge, the 4+ category clearly stands out, which suggests that higher claim frequency does correlate with higher fraud likelihood.

My second question looked at whether the presence of a police report affects the outcome of a claim or the amount of money that gets paid out. I assumed that having a police report would make a claim seem more legitimate, which should result in higher approval rates and maybe even higher claim amounts since insurance companies have verified documentation supporting what happened. To check this, I compared claims with police reports to those without them using two variables: *CLAIM_STATUS* (to see whether the claim was denied or approved) and *CLAIM_AMOUNT* (to see how much the payout was). I first calculated the denial rates for both groups and found a major difference. Claims without a police report were denied about **72%** of the time, while claims with a police report were denied only around **37%** of the time. This already shows a pretty big impact.

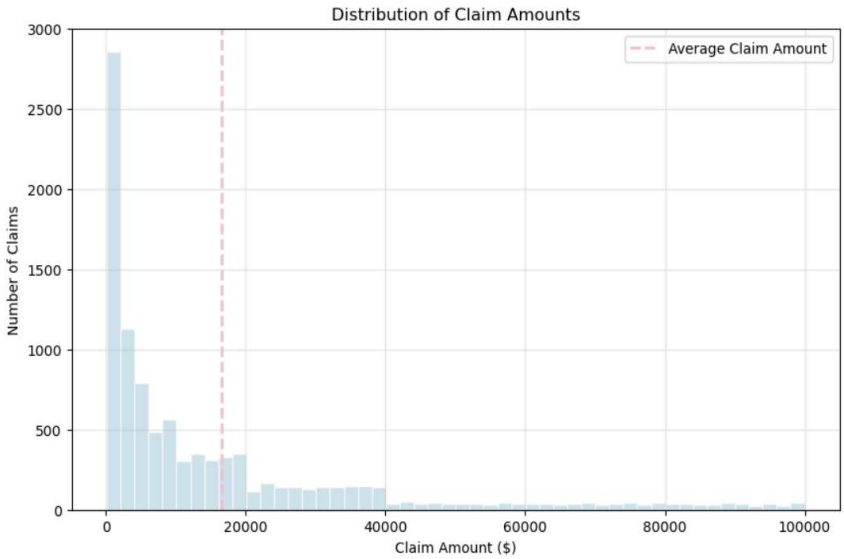




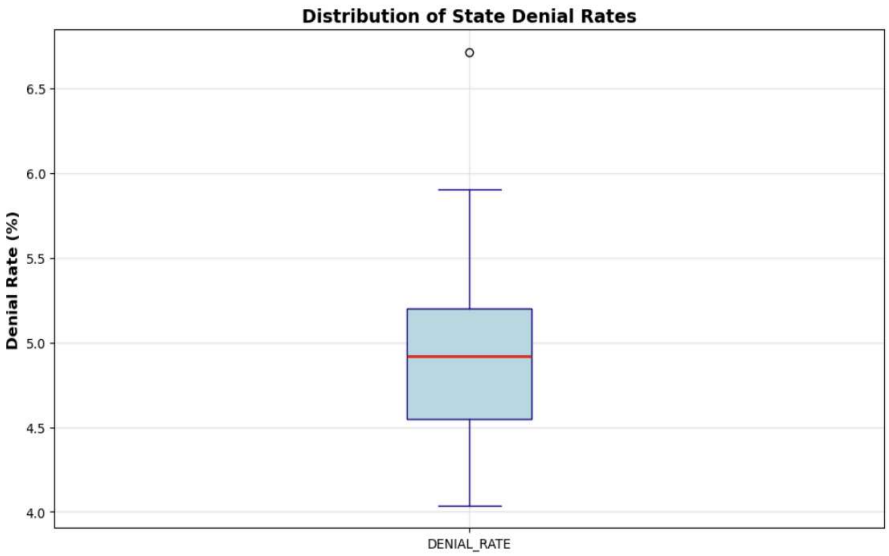
Next, I looked at claim amounts. Interestingly, claims with police reports also had much higher average claim amounts, about **$18,000** compared to around **$8,000** for claims without a police report. I visualized this using a horizontal bar chart and a violin plot. The denial rate plot clearly shows the difference between the two groups, and the violin plot shows how the distribution of claim amounts shifts upward when a police report is involved.

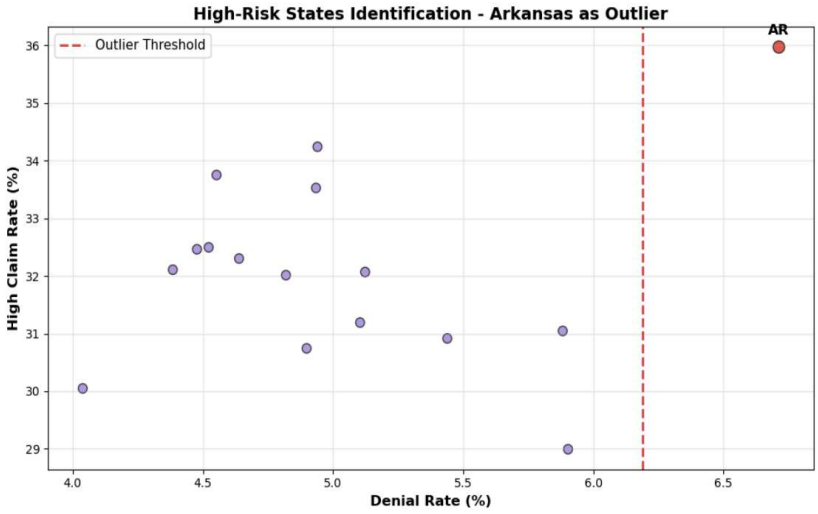
Sara's Question:

Next we wanted to ask what states baby unusually high concentrations of suspicious claim patterns. We reviewed From the 10,000 insurance claims from, we find that we have 16 states and identified two suspicious patterns: Claim amounts (status is denied ('D')), High-value claims (which are those above the average amount of $16,563 (above the average $16,563)). For each state, we calculated two key metrics: **Denial rate** (the percentage of claims denied per state) **High claim rate** (the percentage of claims above the average amount). We used the Interquartile Range (IQR) method to detect outliers and conclude that states with denial rates above Q3 + 1.5 × IQR were considered unusually high. Only one state is part of the outlier threshold of 6.188%



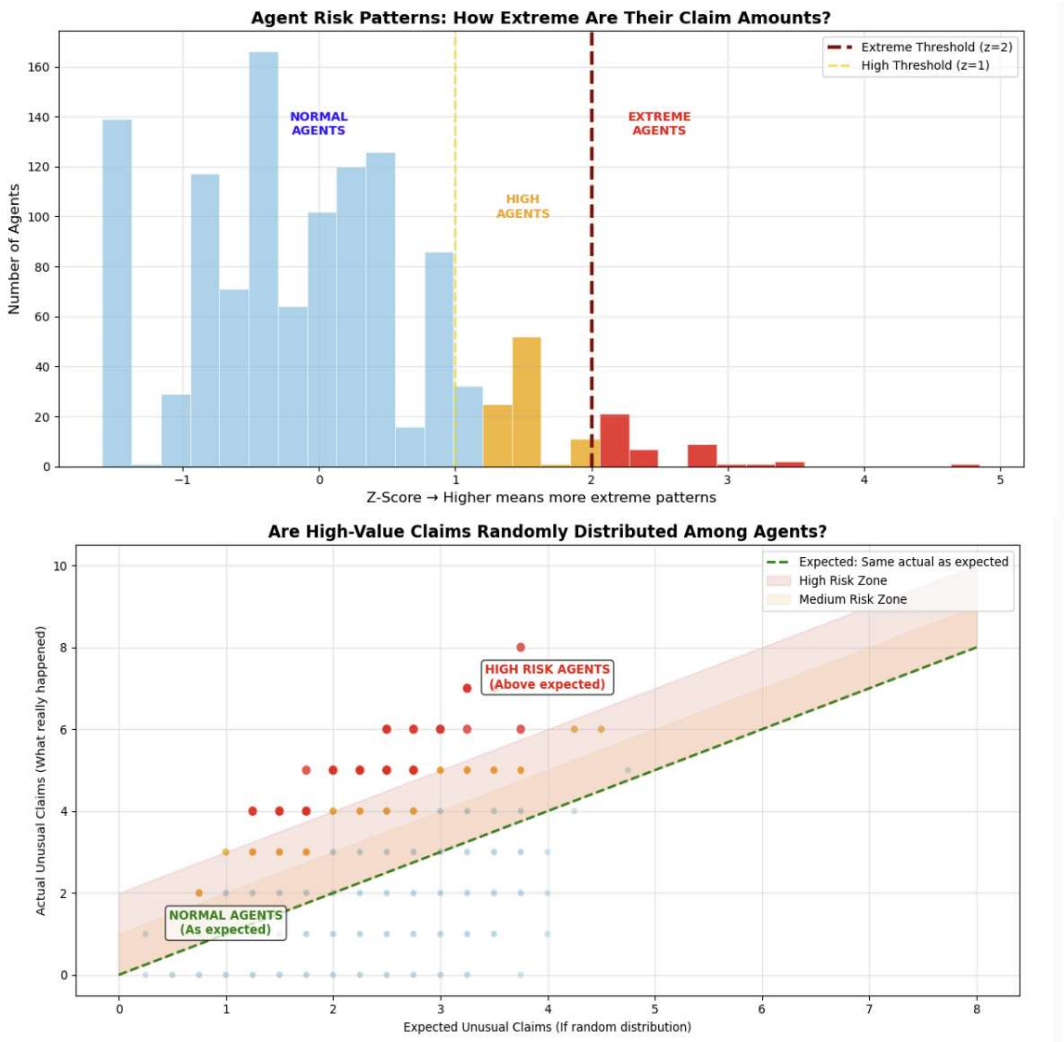
denial rate: **Arkansas (AR)** (This state presents a 6.713% denial rate and 35.972% of its claims are high-value), All other states fell within the normal range of 4.5–5.5% denial rates. The box plot showed Arkansas as the only state outside the expected range, and the scatter plot confirmed that Arkansas had both higher denial rates and more high-value claims compared to other states. While Arkansas stands out, this doesn't prove fraud is occurring. Possible explanations could include regional risk factors, stricter evaluation standards, or local demographic differences. Still, the anomaly justifies a closer review of Arkansas' claim patterns. But while Arkansas was the only outlier, our findings would be stronger with additional context. Such as local economic conditions, comparative risk analysis with neighboring states, and temporal analysis.





Another question we looked at was do the same agents regularly handle policies with unusual claim amounts? We have 1,200 insurance agents who handled the 10,000 claims. We defined an "unusual claim" as one in the top 25% of claim amounts, above $21,000 (the 75th percentile). For every agent, we calculated: Total claims handled, Number and percentage of unusual claims, A z-score to measure statistical significance, How many more unusual claims do they have compared to what would be expected by chance? We used two complementary methods: **Z-score analysis** (to see how extreme each agent's unusual claim percentage was relative to all agents),

Expected vs. Actual comparison (to verify whether agents had significantly more unusual claims than what a random distribution would predict)



The graphs showed that claim assignments are not random:

Statistical significance: 42 agents had z-scores above 2, meaning their patterns are unlikely to occur by chance (probability < 5%).
Excess claims: 59 agents had at least two more unusual claims than expected.
Some extreme cases:

- Agent 01161 handled only one claim—and it was unusual (100%).
- Agents 00010 and 01046 each had 4 out of 5 claims flagged as unusual (80%).
- Agent 00525 had eight unusual claims out of 15 total (53.3%, far above the expected 25%).

A histogram of the data showed that while most agents clustered near the expected 25% mark, a significant group appeared in the 50–100% range. Evidence suggests that certain agents receive a disproportionate share of high-value claims. Although 59 agents showed irregularities, we should focus on those with both meaningful excess claims (more than 2) and enough total cases (more than 5) to trust the pattern.

Conclusion:
Through our combined analyses, we found several meaningful patterns across the insurance claims. There is some but still very little correlation between the time between policy effective and loss date and the claims being accepted,

For claim frequency, we saw a clearer pattern: customers with a higher number of total claims, especially those with four or more; had a significantly higher fraud likelihood. This trend supports the idea that claim frequency can serve as a moderate risk factor for detecting suspicious behavior. Similarly, the presence of a police report had a major impact on claim outcomes in our data.. Claims without a police report were denied far more often, while those with police documentation not only had higher approval rates but also higher average claim amounts, suggesting insurers view these claims as more legitimate.

The only state we found with an unusually high concentration of strange claim patterns was Arkansas, and that the same agents do tend to handle unusual claims.

If we were to continue this project there are a few different ways we could continue to refine our data and expand our research such as looking at cost of living and insurance prices in Arkansas to try and find possible reasons why it had an unusually high risk. We could look at the time of the year that most policies become effective and losses occur and try and relate that to claim denial frequency. We could possibly expand our data set and by doing so not just check if the same agents handle unusual claims but expand that to the same companies or office locations. Or possibly expand into looking at different types.