1. Introduction

The dataset that we are looking at revolves around insurance claims after getting into car accidents, and whether or not such claims were fraudulent. The data has a plethora of information regarding characteristics of the driver, the car, and information pertaining to where and when the accident took place. Some of the information that is being looked at is the make of the vehicle, price of the vehicle, age of the vehicle, age of the driver, sex, marital status, driving rating, past claims, and much more. After all of the demographic information, there is also an indicator as to whether or not the insurance company found the claim to be fraudulent or not. All of this information will be useful in understanding if there are any inherent characteristics regarding an accident that makes certain drivers, cars, situations, etc. more prone to insurance claim fraud. This data can specifically be helpful for insurance companies, in that this clustering application can help companies properly price policies. Specifically our data can bring to light risk factors for certain drivers, "[a]utomobile insurance policy pricing relies on rating factors to assess the exposure to loss associated with an insurance policy. These factors are used to separate the lower risk drivers and vehicles from the higher ones, and largely form the basis of what an individual is charged on an auto insurance policy" (Consumer Cost of Automobile Insurance 2). The factors that we plan to look at in our data are congruent with what the American Academy of Actuaries also look at: "Driver age and/or years of driving experience, Gender, Marital status, Driving record (tickets and accidents), Claims history, Miles driven, Vehicle make and model" (Consumer Cost of Automobile Insurance 3) and will provide useful information in determining what types of premiums insurance companies can and should place upon their policies. The clustering of this information will provide companies or policy makers with the information on which types of people/cars are more likely going to commit fraud or even get into an accident in general. In another sense, this application of our data can be helpful in regards to risk mitigation. "Insurance companies have a legal and moral obligation to shareholders and their policyholders to challenge and resist payment of fraudulent claims. To do this, they employ well-trained fraud analysts and investigators, using a variety of data resources and forensic analysis to ferret out fraud" (The Impact of Insurance Fraud). This legal and moral obligation can be upheld with our data, in that we are able to give companies accurate information and trends. As well, "risk analysis is not just based on short-term horizons…This focus on understanding long term impacts allows decision makers to better understand the typical range within which outcomes are expected to lie…" (Risk management– an actuarial approach 3). In leveraging our dataset, companies (as an example) can adopt a proactive stance in regards to risk management by utilizing trends and clusterings. Our data offers a deeper understanding of long-term impacts and enables decision makers with the ability to anticipate and mitigate risks over periods of time.

Based on the research motivation we selected, each type of clustering technique has a bit of variance in terms of how useful they would be. In terms of well-separated clusters, this would be highly useful because well-separated clusters could indicate that there are distinct characteristics with things like demographic information, car details, or other factors that tend towards fraud. The people using our data could utilize these well-separated clusters to justify and

identify patterns of behavior when it comes to fraudulent claims. For example, if one cluster is highly characteristic, insurers can allocate more resources towards determining if premiums for that group of people with similar characteristics should be increased. For singleton outlier clusters, this could show unusual claims that deviate from typical patterns. This could provide companies with something to look for in the future to see if this pattern develops further. Or on the other hand, this information could be just an outlier occurrence that is not representative of anything further, but again could be used as a way to guide further exploration to see if this pattern comes up in the future. Fuzzy clusterings would likely be extremely useful in detecting overlapping characteristics of fraudulent behavior. For instance, it is unlikely that just being male or female identifying would give an insurance company enough information as to whether or not premiums should be raised, but a combination of categories would likely give a better indication as to what types of people/cars would commit insurance fraud. Fuzzy clustering allows for a less rigid form of understanding the characteristics in our data. A dendrogram displaying nested cluster relationships might not be as helpful in understanding and producing actionable insights for fraud detection, but it could be useful in showing hierarchical patterns and relationships between different groups within the data. This could potentially be reproduced in subsets of people with similar characteristics.

Bibliography

Anitra. "The Impact of Insurance Fraud." Pharmacists Mutual Insurance Company, 15 May 2020, phmic.com/impact-insurance-fraud/.

"Consumer Cost of Automobile Insurance." American Academy of Actuaries, Apr. 2021, www.actuary.org/sites/default/files/2021-04/ConsumerCostOfAutoInsurance.IB_.4.21.pdf.

"Risk Management – An Actuarial Approach." Institute and Faculty of Actuaries, June 2017, www.actuaries.org.uk/documents/risk-management-actuarial-approach.

2. Dataset discussion

Dataset Display:

```
import pandas as pd
df=pd.read_csv("fraud_oracle.csv")
df.head()
```

There are originally 15420 rows and 33 columns in this dataset before cleaning.

This is the link for where we got the dataset, it is from Kaggle:
https://www.kaggle.com/datasets/shivamb/vehicle-claim-fraud-detection/data

Each row/observation represents an insurance claim made in the years 1994-96 based on a car accident that the person either caused or got into. Within each row there is an indicator as to whether or not fraud was detected in said claim.
The author of the data did not release how the data was collected, but this data is also referenced in an Oracle research about detecting fraud, so we believe the data to be reputable.
This dataset could definitely not encompass ALL types of observations that would be helpful in determining fraudulent claims. There are inherently so many other characteristics about people that can be determinants in whether or not they are likely to commit fraud - for example criminal history, past fraud, credit score, etc.
The answer to the prior question is really important because of course, characteristics that demonstrate a trend in what types of people commit fraud, does not necessarily mean they will commit fraud. Since we are only analyzing data from limited years, towns, and characteristics, we naturally will not have as large of an interpretation because we are simply limited as to how much data we are given. With more data, there is likely more accuracy in which types of traits are prevalent in fraud, so with this, the people using our data will likely need to take a deeper look into their clients. Our data can help them establish specific trends to look for, but can not inherently give them a “diagnoses” as to who will commit fraud.

3. Basic Dataset Cleaning and Exploration

There were no missing values in our dataset.

Before doing the outlier identification and noise identification, we have decided to clean the data by removing all of the information that isn’t from the year 1994. Since we have 3 years of data, and a lot of rows, we want to remove the other years so our data is a little bit easier to cluster. In this case our analysis will focus strictly on the year 1994. We have reduced our dataset to 6142 rows this way. We also wanted to get rid of certain columns that we will likely not use in our analysis. The columns that we decided to drop were Month, WeekOfMonth, DayOfWeek, AccidentArea, DayOfWeekClaimed, MonthClaimed, WeekOfMonthClaimed, PolicyType, PolicyNumber, PoliceReportFiled, WitnessPresent, AgentType, NumberOfSuppliments because the aim of our project is to see if there are any characteristics about the person or the vehicle that make them tend more towards fraud, and not information about when and where the claim was submitted.

```
yeardf = df[df['Year'] == 1994]


columns_to_drop = ['Month', 'WeekOfMonth', 'DayOfWeek','AccidentArea','DayOfWeekClaimed','MonthClaimed','WeekOfMonthClaimed','PolicyType','PolicyNumber','PoliceReportFiled',
'WitnessPresent','AgentType','NumberOfSuppliments']
yeardf.drop(columns=columns_to_drop, inplace=True)
yeardf
```

✓ 0.4s Python

```
cleaneddata = yeardf[yeardf['Age'] != 0]
cleaneddata
```

It appears that there aren't any evident outliers/noise. We got rid of the points where age was 0 and I think we are good to move on to descriptive statistics. It makes sense that we won't have noise/outliers.

```
from sklearn.preprocessing import StandardScaler

df1=StandardScaler().fit_transform(cleaneddata)
df2=pd.DataFrame(df1, columns=cleaneddata.columns)

df2.head()
```

More confirmation that there is no noise - we ran the DBSCAN Algorithim.

4. Basic Descriptive Analytics Before using any unsupervised learning algorithms, you should learn more about your dataset by performing some basic descriptive analytics.	
OPTION 1 **If your dataset is a structured dataset (ie. not image, audio, time-series data etc.), do the following.** * For your numerical attributes, calculate basic summary statistics about each attribute. * For any categorical attributes (including the pre-assigned class labels, if your dataset has any) count up the number of observations of each type. * Determine if there exist are any strong pairwise relationships between the variables in your dataset.	2

`yeardf["Make"].value_counts()`
The counts of each type of make seems to primarily focused Pontiac, Toyota, Honda, and Mazda and the lowest amounts are in Mercedes, Porche, BMW, and Jaguar.

yeardf["Sex"].value_counts()
The counts of Sex show us that mainly males are involved in accidents.

yeardf["MaritalStatus"].value_counts()

The counts of Marital Status tells us that most of the accidents happen to people that are married with the next highest category being those who are single and a much lower count in those who are divorced or widows.

yeardf["Fault"].value_counts()
The Fault counts tells us that fault for the accidents mainly lie with the policy holders than third party's.

yeardf["VehicleCategory"].value_counts()
The Vehicle Category tells us that accidents happen with mostly Sedan and Sport vehicles with a small amount in Utility vehicles

yeardf["VehiclePrice"].value_counts()
The Vehicle Price counts tell us that the price of most the vehicles start from 20000 to 39000 and then vehicles exceeding 69000. While less counts reside in prices less than 20000 and vehicles ranging in price between 40000 to 69000

yeardf["PastNumberOfClaims"].value_counts()
The counts from the Past Number Of Claims tells us that most accidents have either none all the way to 4 previous claims where a smaller amount have more than 4 previous claims.

yeardf["AgeOfVehicle"].value_counts()
The counts of the Age of Vehicles tells us that most of the vehicles in the dataset are 6 years and older and the lowest counts are vehicles of 2 to 3 years old.

yeardf["AgeOfPolicyHolder"].value_counts()
The counts of the Age Of PolicyHolder tells us that most of the counts belong to policyholders who age is between 31 to 50 and the lowest counts in those who are from 18 to 25 years old

yeardf["AddressChange_Claim"].value_counts()
The Address Change Claim tells us that most of the counts belong to those who had no change in address and a much smaller amount in those who changed their address within 6 months to 8 years.

yeardf["NumberOfCars"].value_counts()
The counts of the Number Of Cars tells us that most of the accidents happen to owners who have only one vehicle and owners of 8 or more vehicles make up a small percentage of the counts.

yeardf["BasePolicy"].value_counts()
The Base Policy counts tell us that between Collision, Liability, and All Perils are evenly distributed but slightly lean more towards collision

There does not seem to be any strong pairwise relationship between variables as shown in our pairplot.

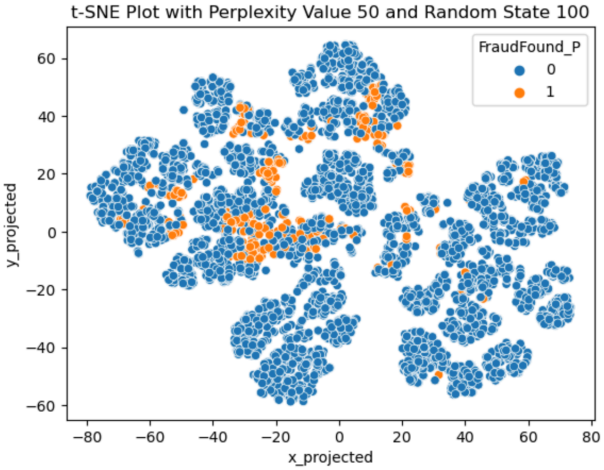
5.Scaling Decisions From your analyses conducted here, discuss whether you should scale the dataset or not. Explain why or why not. If you choose to scale, then do so in this section here.

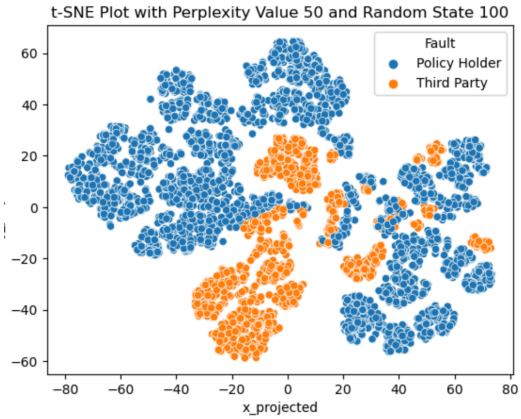
We decided not to scale our data because we only have a few numeric variables, and they are all within one digit of each other, so they would not have super strong pulls/impact on the data this way.

6. Clusterability and Clustering Structure Questions Does your analysis suggest that the dataset clusterable? (The answer to this should be yes). Explain why.

GOWER'S

After running the TSNE algorithm, it does appear that our dataset is clusterable. Right now, it appears that fraud is localized to about 2 of the clusters, and some of our clusters do not appear to have fraud at all.

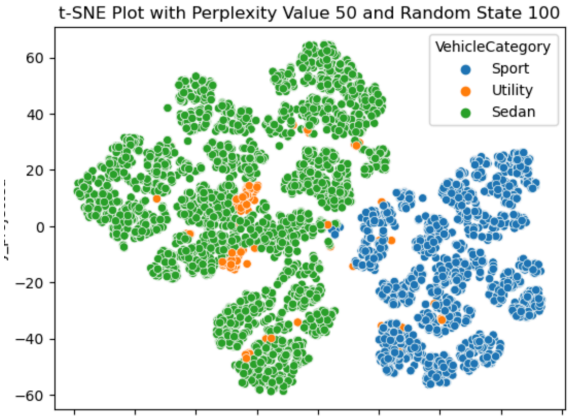


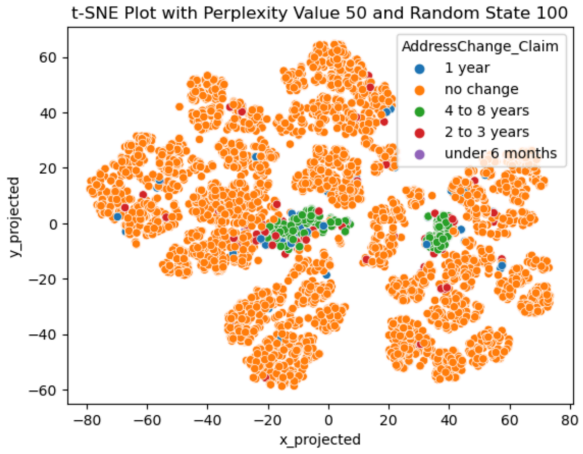


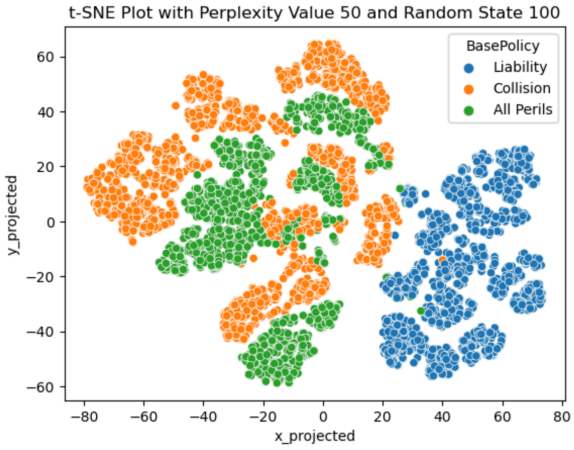
It looks like there are about 3 inherent clusters in our dataset with a lot of nested subclusters in them. It appears that the shsape of our data is somewhat spherical and well balanced in size. The clusters definitely have separation (not super well separated, but our TSNE plot can't really distinguish this to begin with).

This TSNE plot is not going to tell us if there are outliers/noise (but we concluded above that there probably won't be) - as well this plot won't tell us how far apart our clusters are separated and the distance of objects or points in the initial dataset. It also won't be able to tell us the centroid or summary statistics, as well as the sparsity of the data. We have a lot of attributes that are at play at once so we will have to further analyze the association between

some of the attributes and the clusters that are going on- but it does appear for some of the clusters that there are more clear attribute distinctions than others. For example, the cluster in the middle bottom has no fraud found, which can indicate a lot based on our goals. We can visualize if there are attributes at play by running the TSNE algorithim again but coloring by certain attributes. It looks like whoever is at fault for the accident could have some indication of whether or not there was fraud. It seems that if a third party was at fault, it tends to fall into clusters that are not consistent with fraud. It also seems that type of vehicle could provide something interesting - Sports vehicle's seem to be in the cluster with little fraud, Whereas Utility and Sedan could have associations with Fraud. Address Change can also be an attribute that is playing into the clustering structure of our data. It looks like those that have changed their address in the last 4-8 years could have instances of fraud as seen in the middle of the plot. And finally type of insurance could be influencing the clustering structure. It seems that there is no fraud when the base policy is Liability, and most fraud would come from collision or all perils. Based on all of these attributes demonstrating something about the clustering structure, and potentially how fraud can be determined in this case, we will want to keep a look out for some of these traits when we use clustering algorithms for the rest of our data.







7.

The first clustering algorithm that we chose to work with was k-prototypes. The reason for this is that since our data is mixed, we wanted to work with an algorithm that would compliment our data. We initially thought of separating our data based on only numeric information, but we didn't think this would fully encompass our research goals, so we wanted to work with an algorithm that was going to allow us to see if there was information about both the person and the car that relates to fraud that we could analyze. Our research goals were seeing if there were inherent characteristics of a person or a vehicle that would allow us to make some sort of link between them and the association of fraud. If we were to split and limit our data we would not be able to reach our goals, so we wanted to keep both categorical and numerical information in our data. Based on the TSNE plot and DBSCAN above we believe that most of the ideal properties are met to run this algorithm. It appears that our clustering structure is evident: roughly spherical, decently well separated. As well, we did not see the appearance of outliers/noise in our data which means that the k-prototypes algorithm is likely to work well for our data.

The second clustering algorithm that we chose to work with was hierarchical agglomerative clustering with gower's distance. The reason for this is that since our data is mixed, we wanted to work with an algorithm that would compliment our data. As well, when we ran the TSNE plot, we saw a lot of nested clusters within our data, and we thought that an agglomerative clustering method would really reveal to us what those nested clusters could be. This also works well with our research goals because there could be larger characteristics that coincide with fraud that break down into smaller more distinct characteristics. An example (that could occur we haven't actually noticed this) would be gender being a large defining characteristic that breaks down into make of car. We believe that most of the ideal properties are met to run this dataset because like we said before there are no outliers/noise in our dataset. As well, the nature of the separation and nested clusterings make this algorithm really interesting to use because it hopefully reveals something interesting about our data. In our algorithm running we are going to use multiple types of linkage to see if our data is best clustered using a certain type of HAC.

A fuzzy clustering of the dataset such as using the Fuzzy C-means clustering would give us the cluster membership scores for our dataset but likely is not going to be that helpful in determining

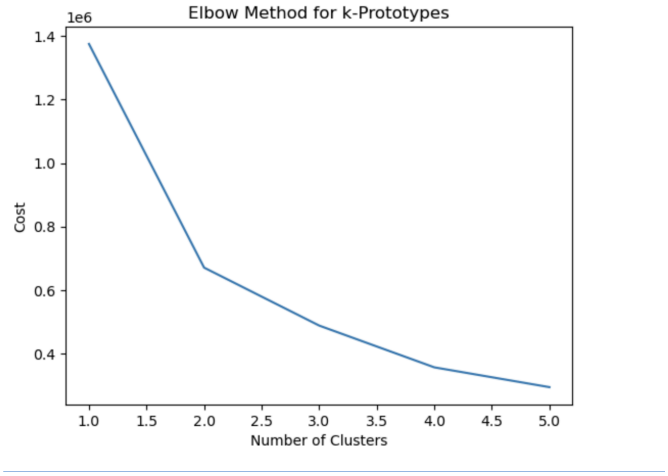
fraud. Since there are so many attributes to our dataset and we are only interested in seeing two outcomes, fraud or not fraud, fuzzy clustering would not be especially useful.

8. Clustering the Dataset and Post-Cluster Analysis for Algorithm 1	
Parameter Selection * Select the parameters that you intend to use for this clustering algorithm. Explain and show your work for why you selected these particular parameters. * If one of your parameters is the cluster number, consult two methods for choosing this cluster number. Note any differences suggested by these two methods and discuss why they may have been different.	2
Clustering the Data Cluster the dataset with this algorithm and the parameters that you chose. Make sure to use a random state for non-deterministic algorithms.	0.5
Clustering Algorithm Results Presentation Present and discuss the results from each of these algorithms to the reader of your report in an insightful way that relates back to your original motivation for performing the unsupervised learning analysis. For instance: • If your clustering is a hard assignment, you can color-code your t-SNE plot with the cluster labels. • If your clustering algorithm is a hierarchical clustering algorithm, give the dendrogram and explain the nested relationships. • If your clustering is a fuzzy clustering (or has cluster membership scores), you can plot K (# of clusters) t-sne plots, and color code each plot by the cluster membership score for the kth clusters.	1
Additional Cluster Exploration * How far apart are each of your clusters from each other? (Use a cluster-sorted similarity matrix) * How cohesive and well-separated are EACH of your clusters? (Use a silhouette plot. Hint: it's the barplot with the silhouette scores of each observations.) * Are any of your clusters more sparse than others? (Hint: KNN Distance plot) * Shortcomings: Based on the properties of your dataset and the clusters, you have any reason to believe that your cluster-sorted similarity matrix and your silhouette plot are not the best techniques to use to measure the cohesion and separation of your clusters? Explain.	3
Finding the "Inherent" Clusters * Do you have any evidence to suggest that your clustering "split" one of the "inherent" clusters that exists in this dataset? Explain. Hint: You can use a t-SNE plot, cluster sorted-similarity matrix, or dendrogram (if you used HAC) to check this. * Do you have any evidence to suggest that one of the clusters in your clustering contains two or more of the "inherent" clusters that exists in this dataset? Or in other words, do you have any evidence to suggest that there exists a meaningful separation of points in one of your clusters? Explain. Hint: You can use a t-SNE plot, cluster sorted-similarity matrix, or dendrogram (if you used HAC) to check this.	1.5

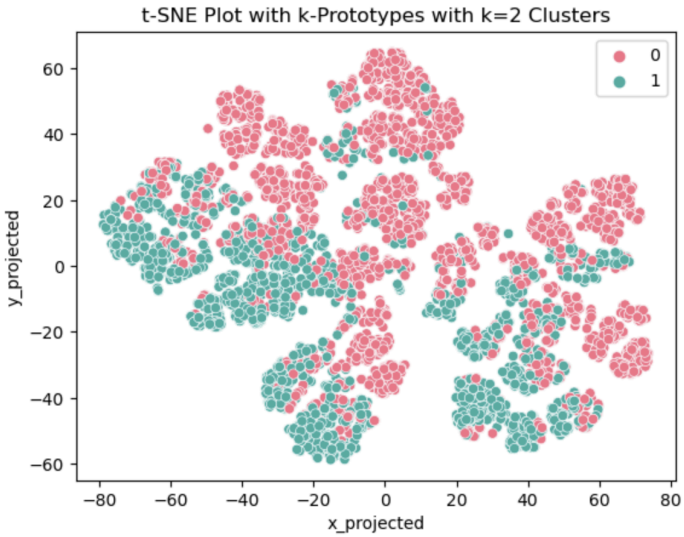
Describing Each of the Clusters [QUALITIATIVELY THE MOST IMPORTANT PART OF THE PROJECT!]
Finally, describe what type of attribute values and attribute relationships characterize each of the resulting clusters in your final clustering. You can choose at least one of these options (or pick multiple options to learn more).

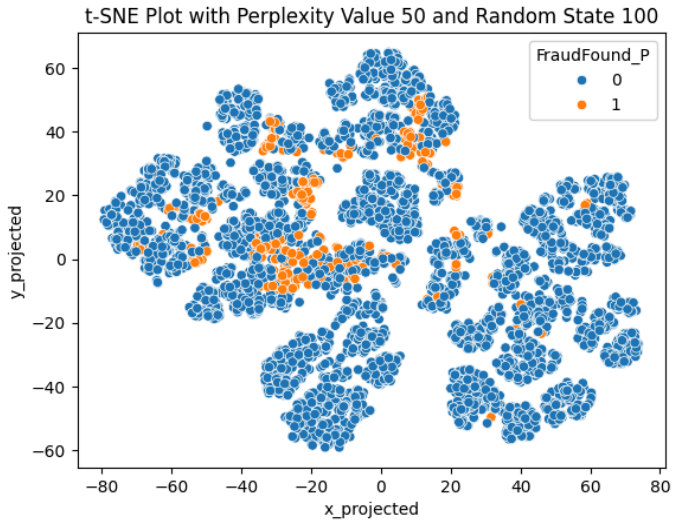
• Option 1 (don't use this if your dataset is an image dataset):
  * Create a side-by-side boxplots visualization for each numerical attribute in your dataset (where each cluster label is given a boxplot).
  * Create a side-by-side barplot visualization for each categorical attribute in your dataset (where each cluster label appears on the x-axis).
o Use these plots to thoroughly describe which type of attribute values characterize each of the resulting clusters in your final clustering.

• Option 2 (if you used a prototype-based clustering algorithm):
  * If your clustering algorithm is a prototype-based clustering algorithm, display (visualize if it's an image dataset) and compare each of the prototypes of the clusters.
  * Use these prototypes to thoroughly describe which type of attribute values characterize each of the resulting clusters in your final clustering.

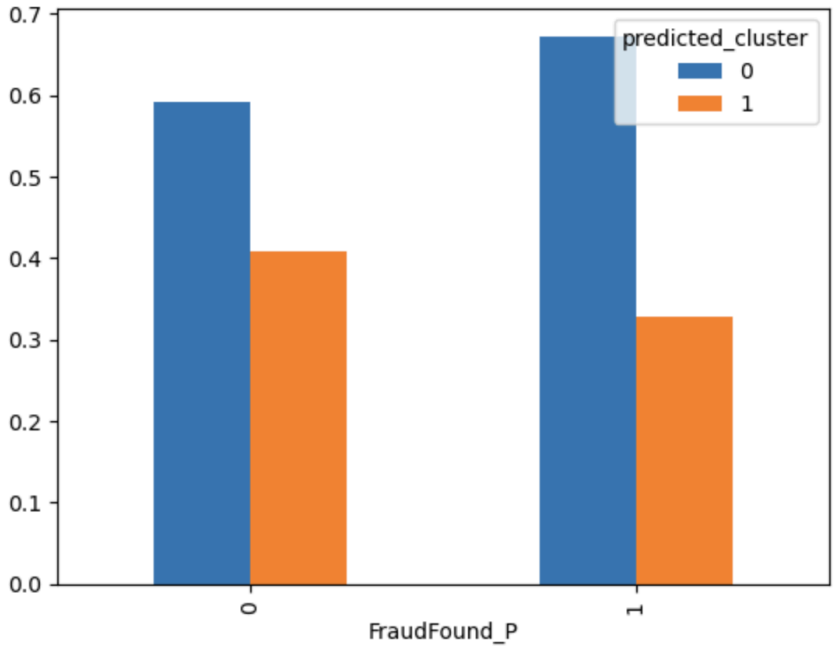


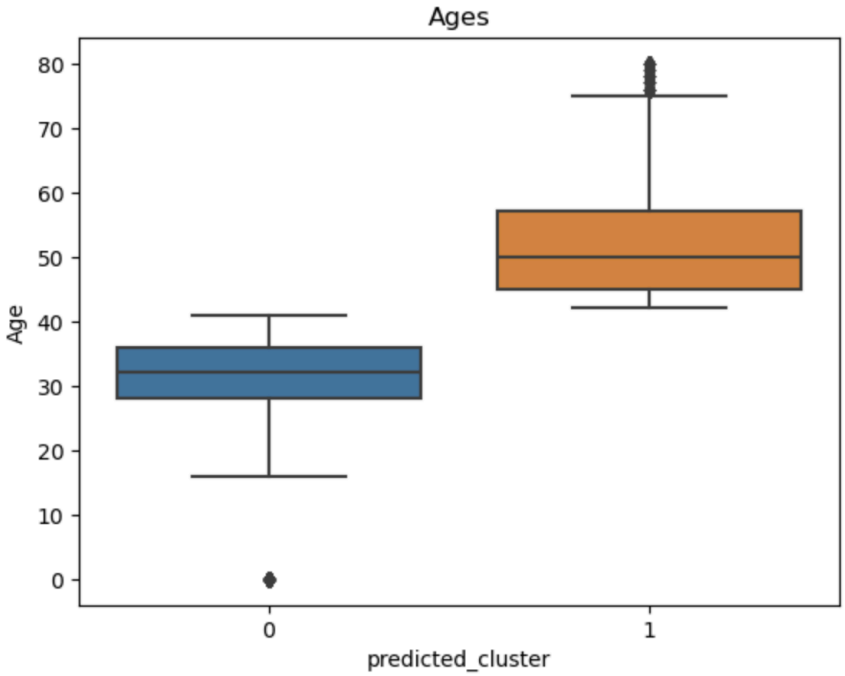
The first algorithm we decided to try out is K_prototype. For the number of clusters, I chose 2 because the t-SNE plot suggests the existence of two main clusters. I also made an elbow plot that clearly indicates an elbow point at k=2.





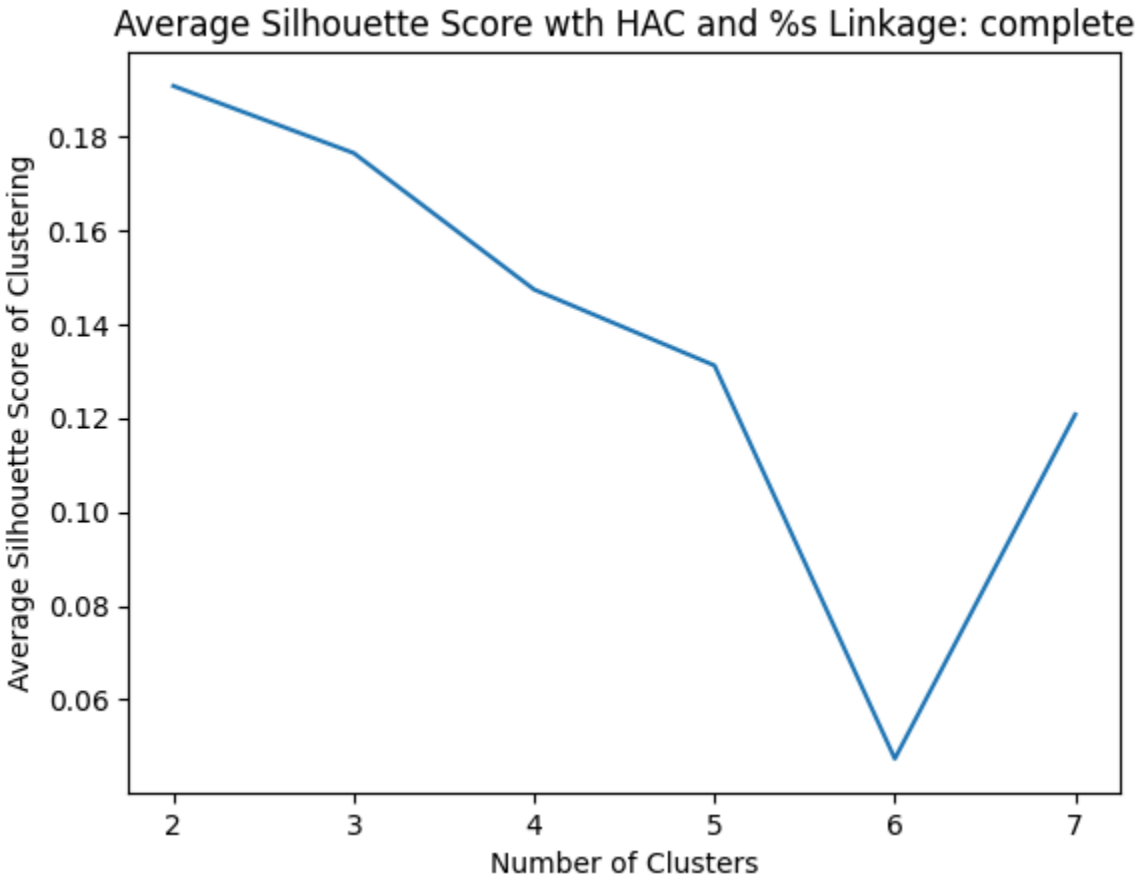
For the result, we plot T-SNE against our cluster labels. Since we are using a different metric to measure distance, it is not very helpful to make a fair comparison. Therefore the T-SNE plot may not accurately reflect whether there exists meaningful separation between K-prototype clusters.

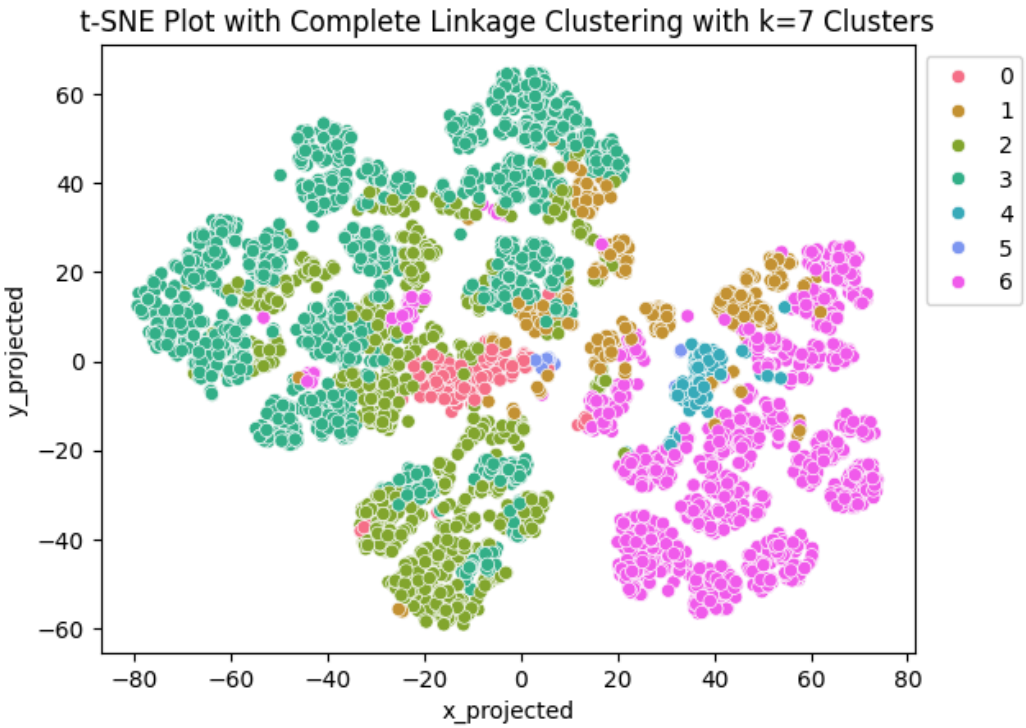

predicted_cluster
0
1
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
1
FraudFound_P

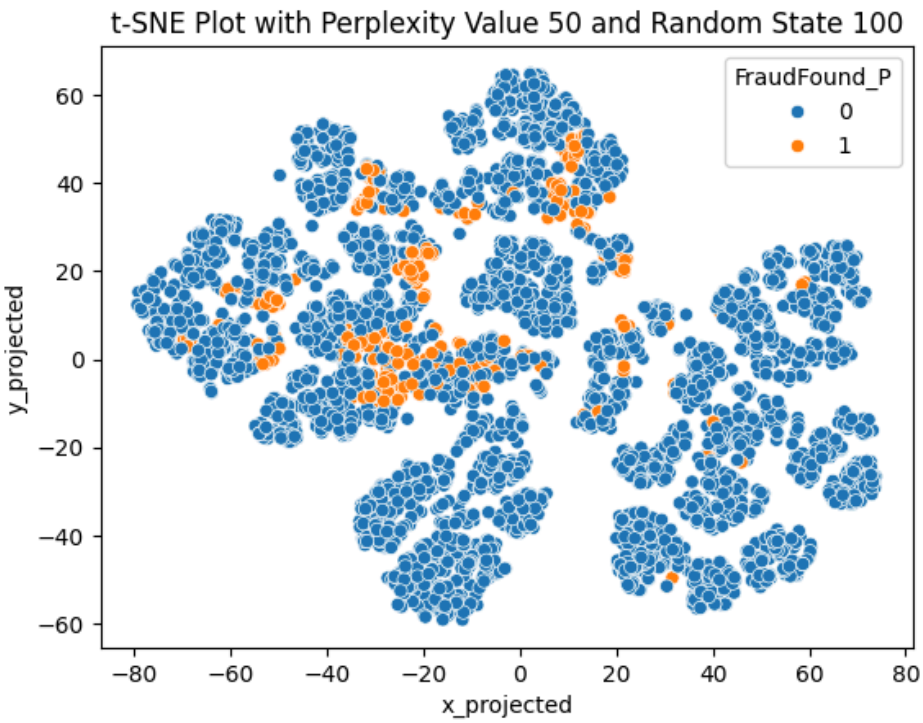

Ages
80
70
60
50
40
30
20
10
0
Age
0
1
predicted_cluster

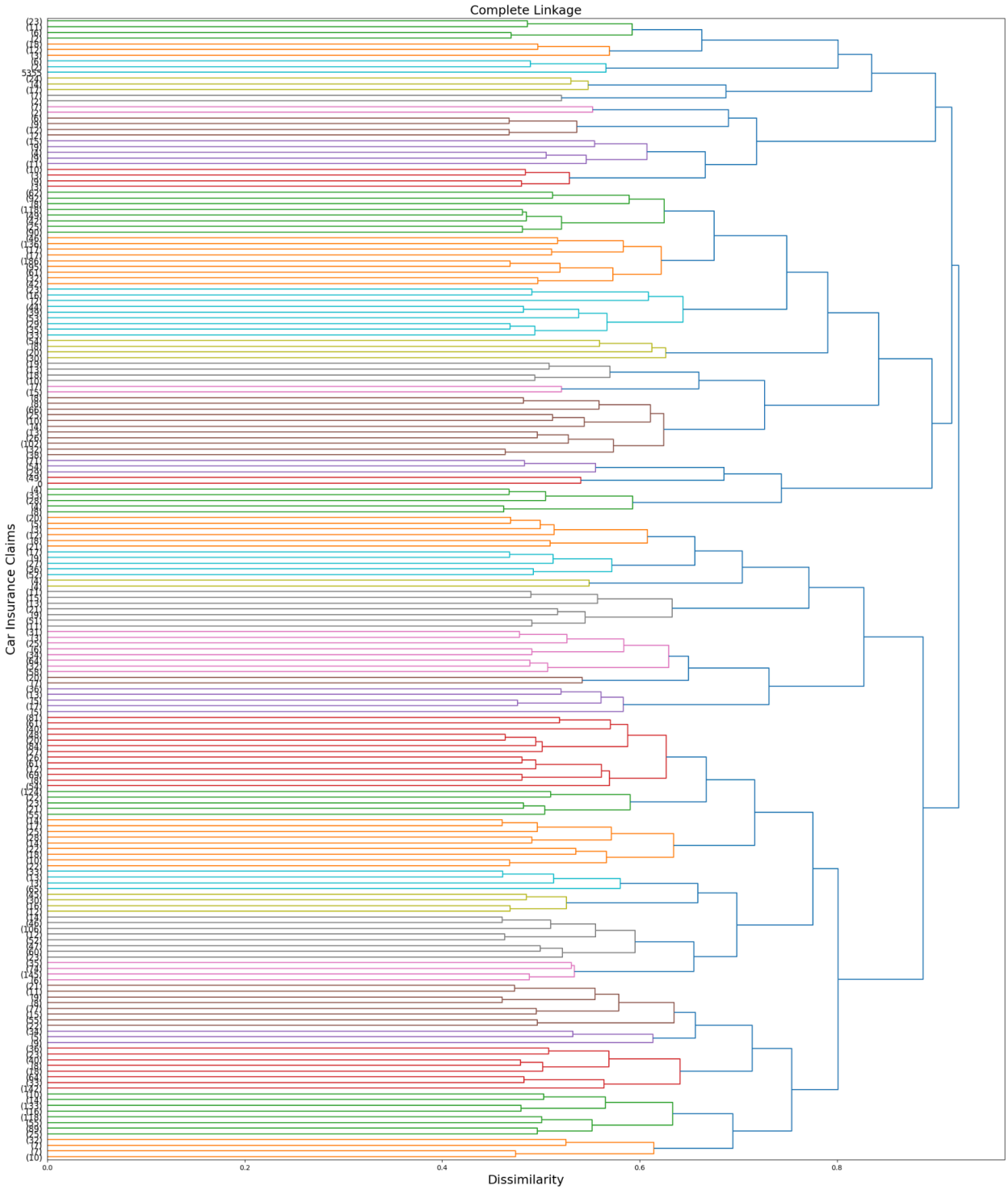
After plotting the dendrograms for HAC with single, complete, average, and ward linkage, we found that complete has the most evenly distributed observations in each cluster. However, when looking at the the average silhouette score plot,

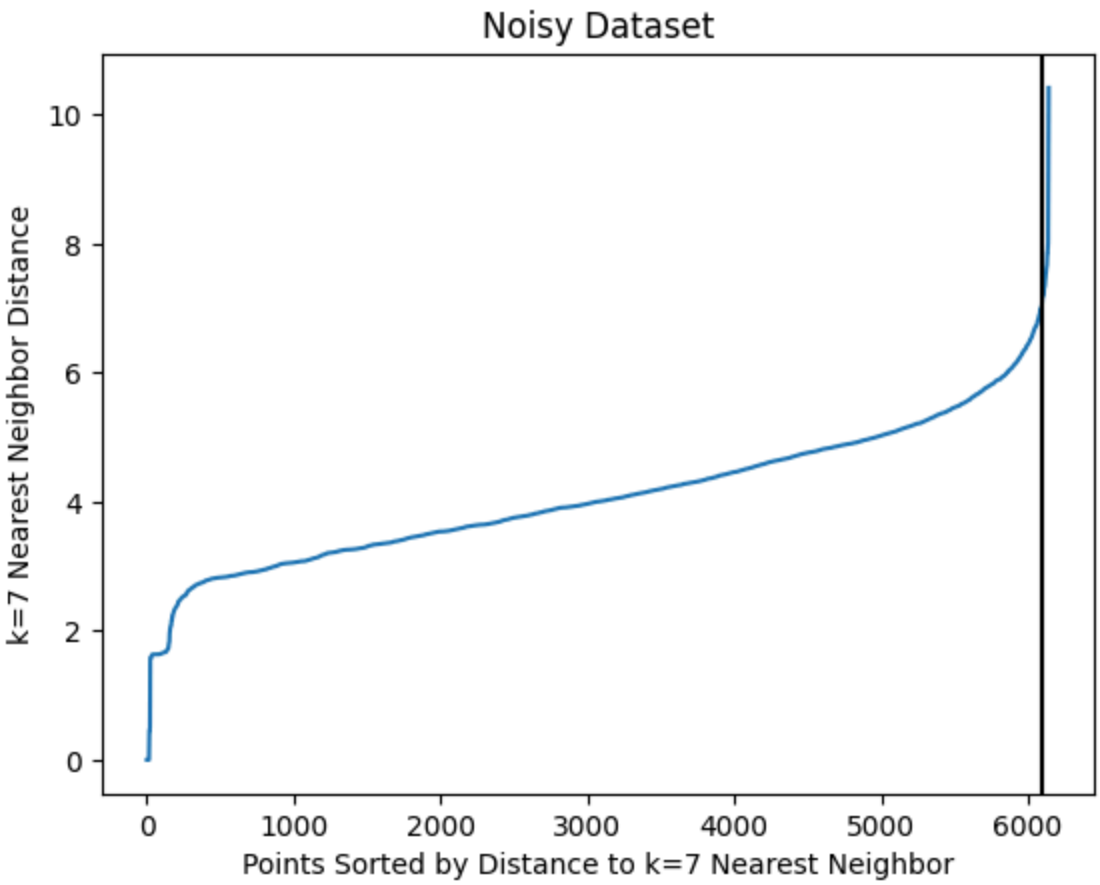
Reasoning behind choosing linkage function with number of clusters



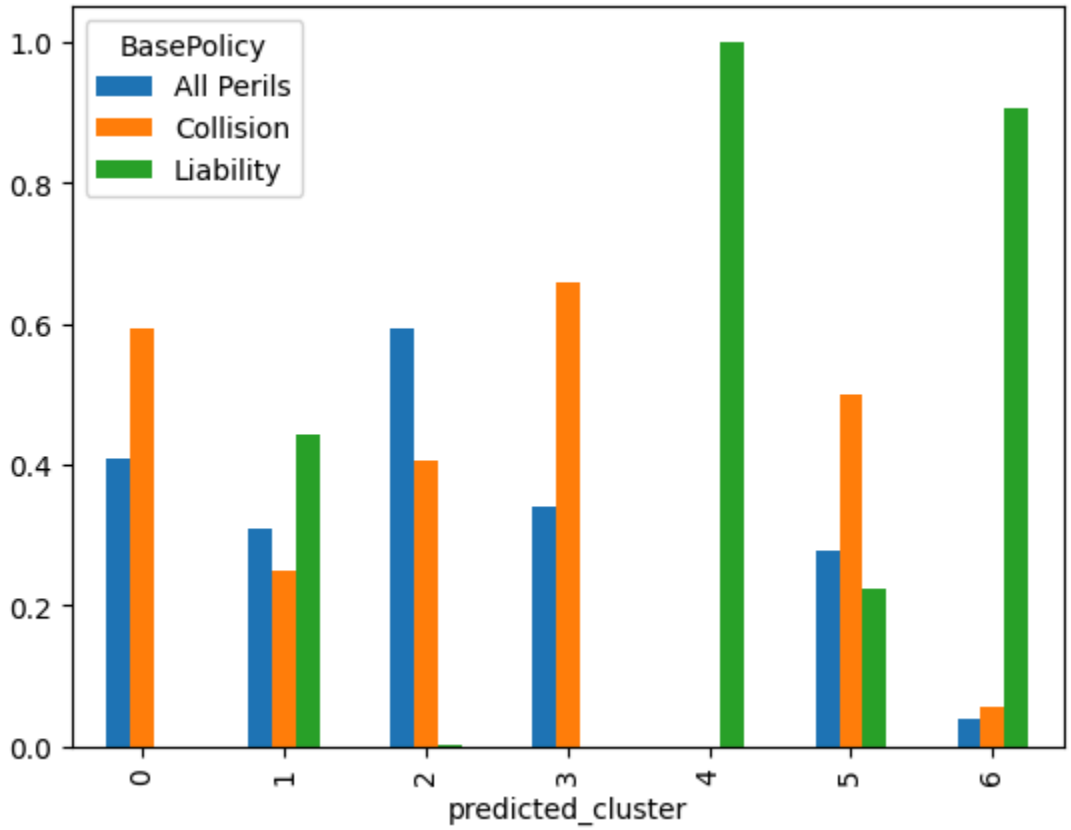

t-SNE Plot with Complete Linkage Clustering with k=7 Clusters
y_projected
x_projected
60
40
20
0
−20
−40
−60
−80
−60
−40
−20
0
20
40
60
80
0
1
2
3
4
5
6

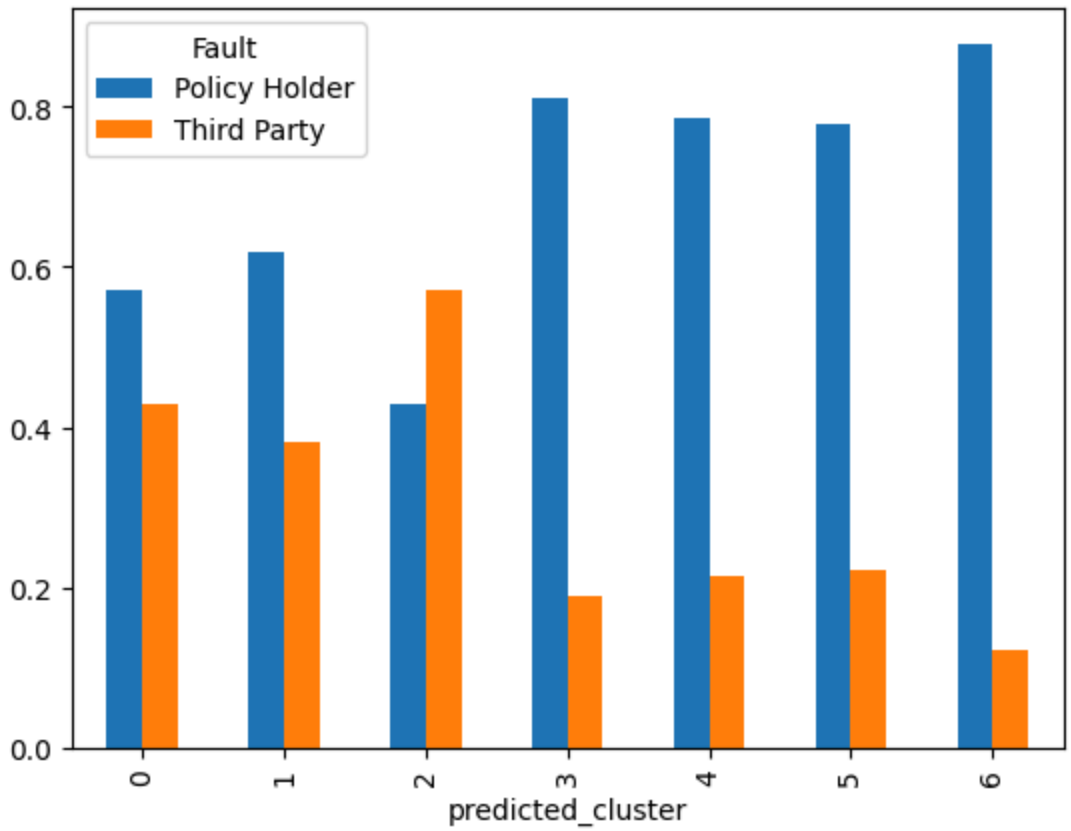

t-SNE Plot with Perplexity Value 50 and Random State 100
FraudFound_P
0
1
y_projected
x_projected
60
40
20
0
−20
−40
−60
−80
−60
−40
−20
0
20
40
60
80

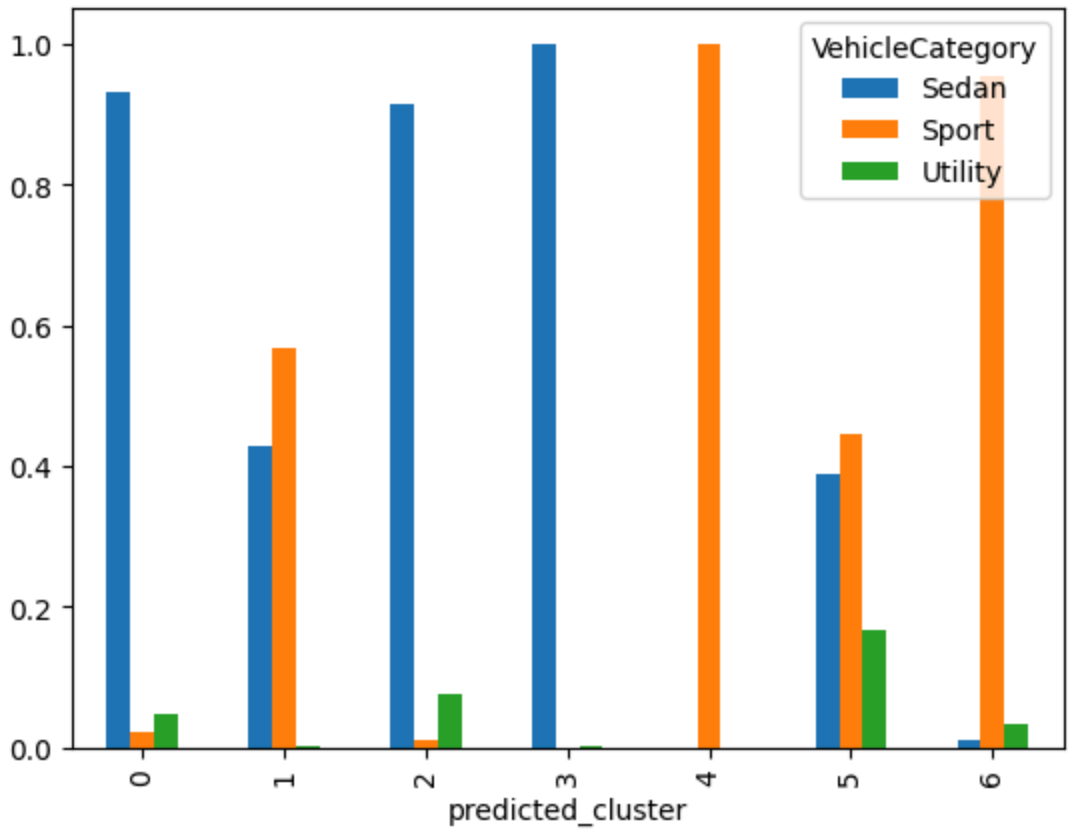

Complete Linkage
Car Insurance Claims
Dissimilarity
0.0
0.2
0.4
0.6
0.8

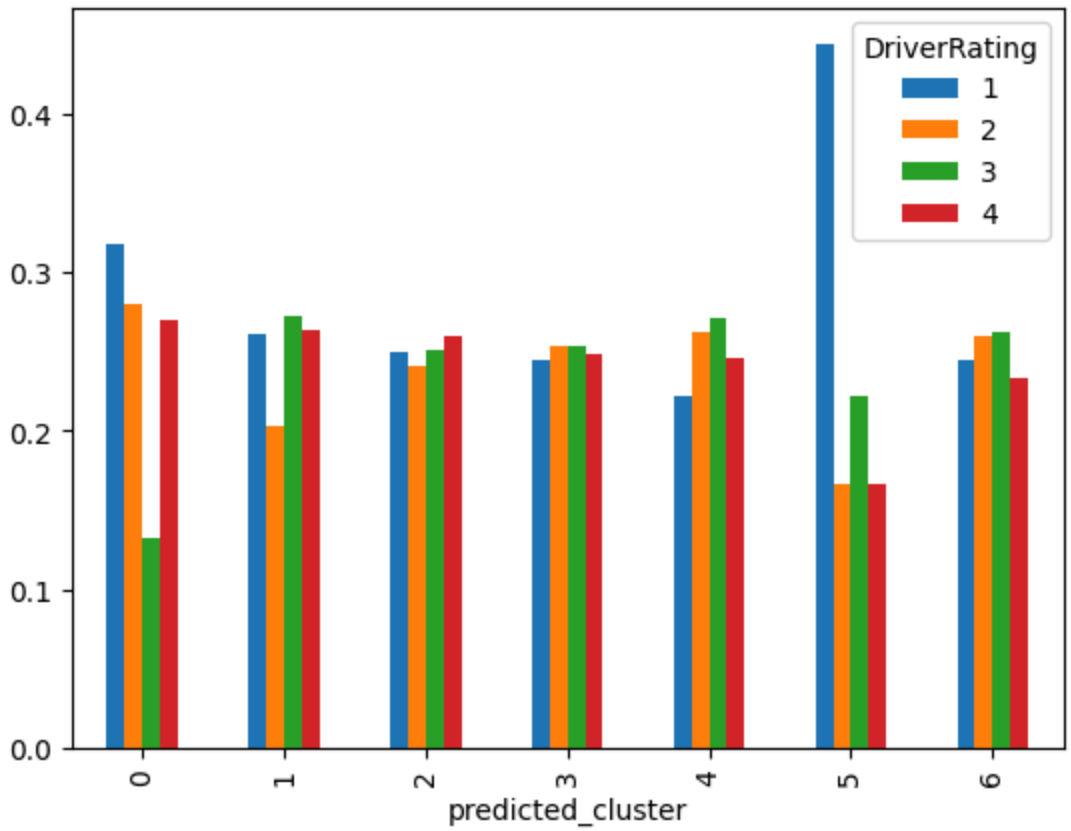

Noisy Dataset
k=7 Nearest Neighbor Distance
0
2
4
6
8
10
0
1000
2000
3000
4000
5000
6000
Points Sorted by Distance to k=7 Nearest Neighbor

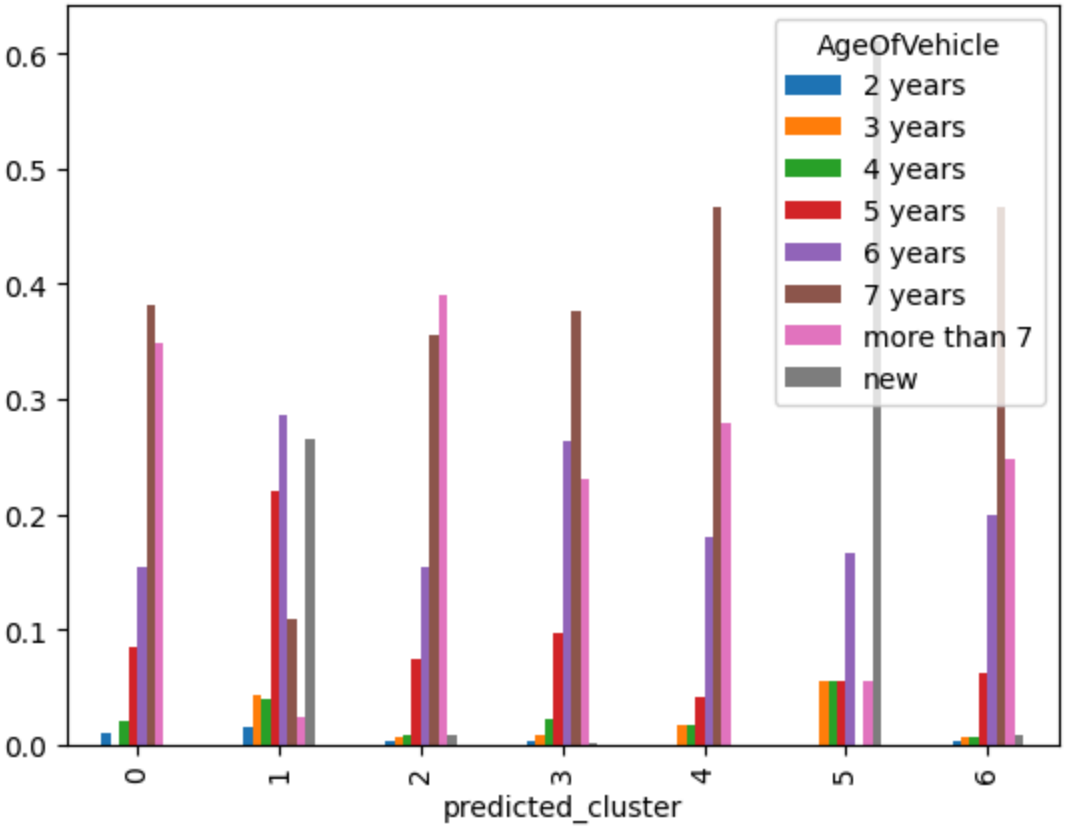
Results

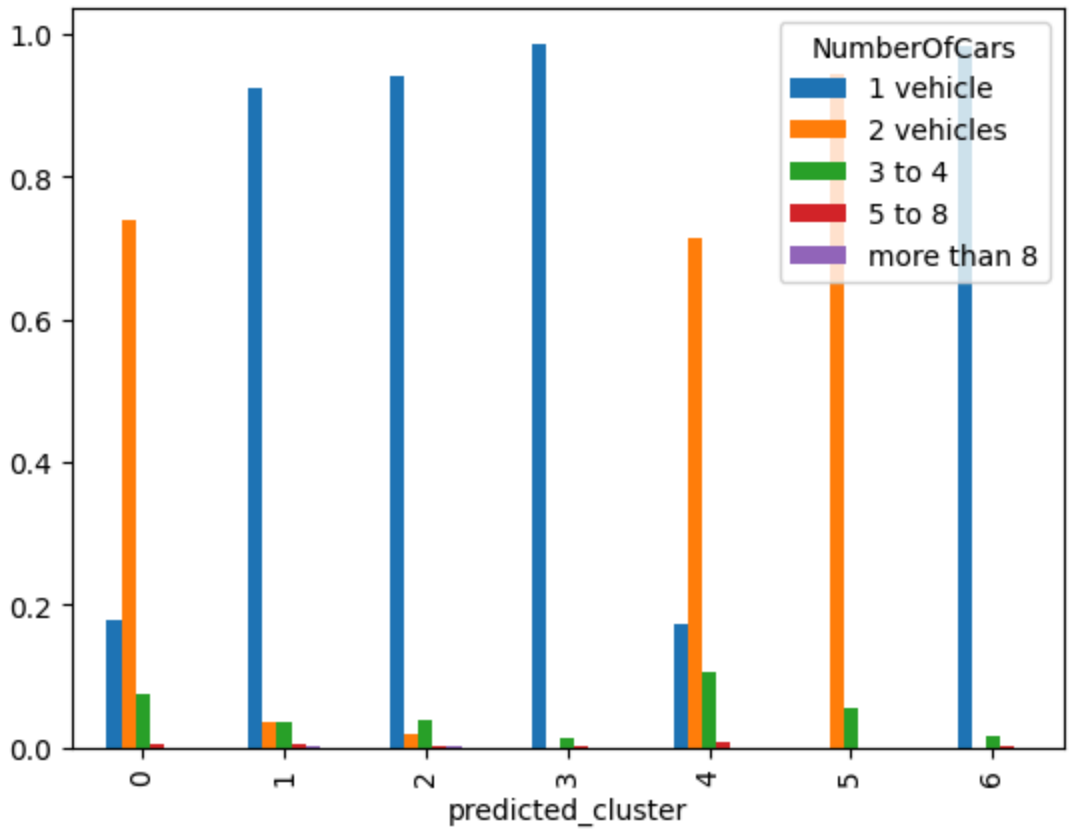

BasePolicy
All Perils
Collision
Liability
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
predicted_cluster

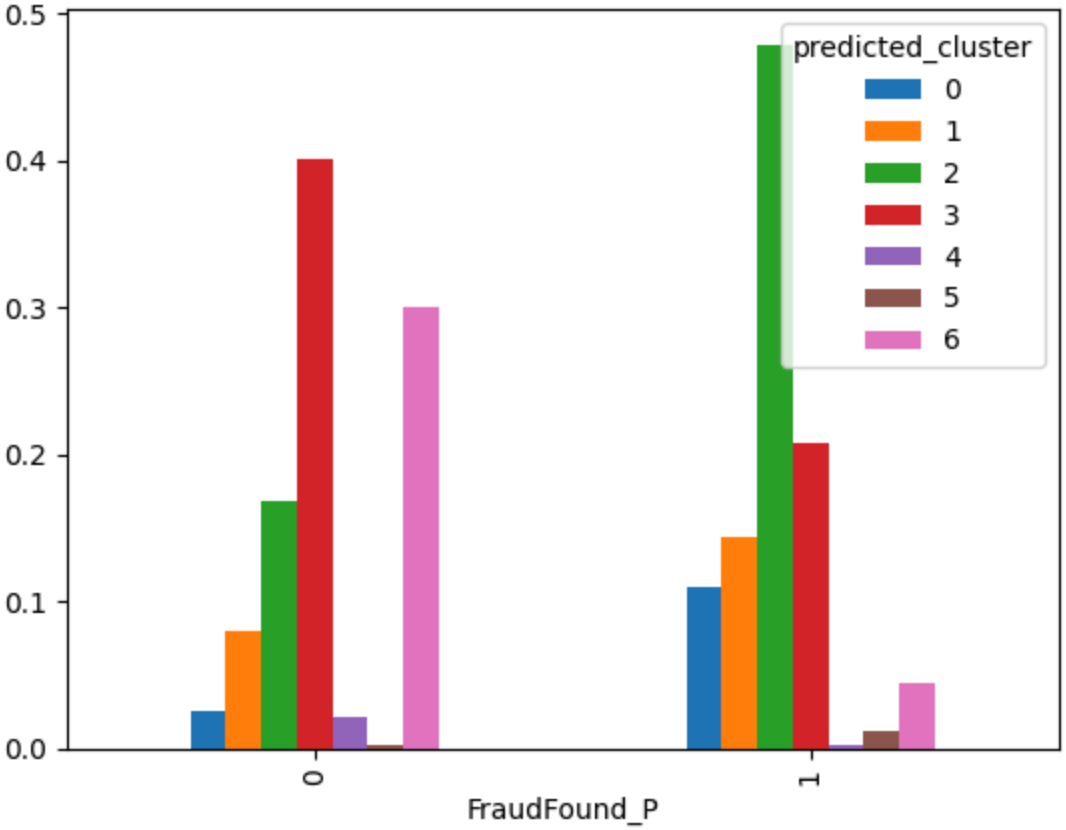

Fault
Policy Holder
Third Party
0.0
0.2
0.4
0.6
0.8
0
1
2
3
4
5
6
predicted_cluster

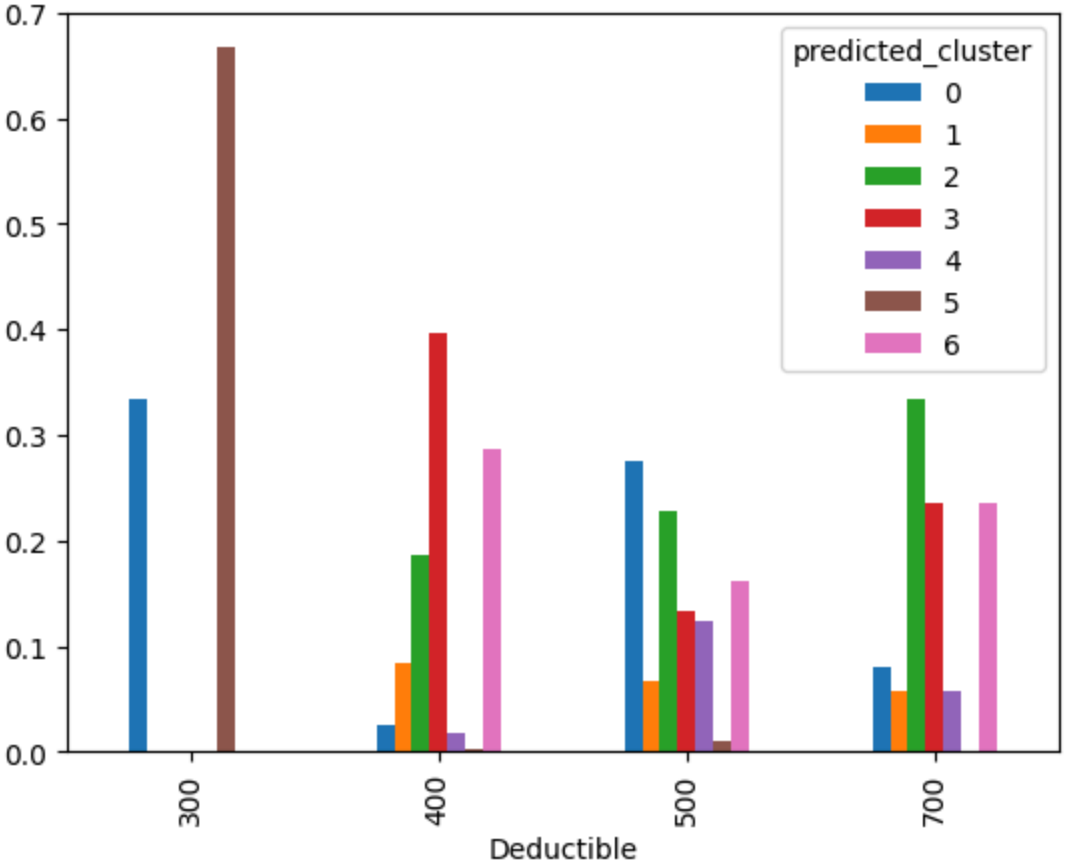

VehicleCategory
Sedan
Sport
Utility
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
predicted_cluster

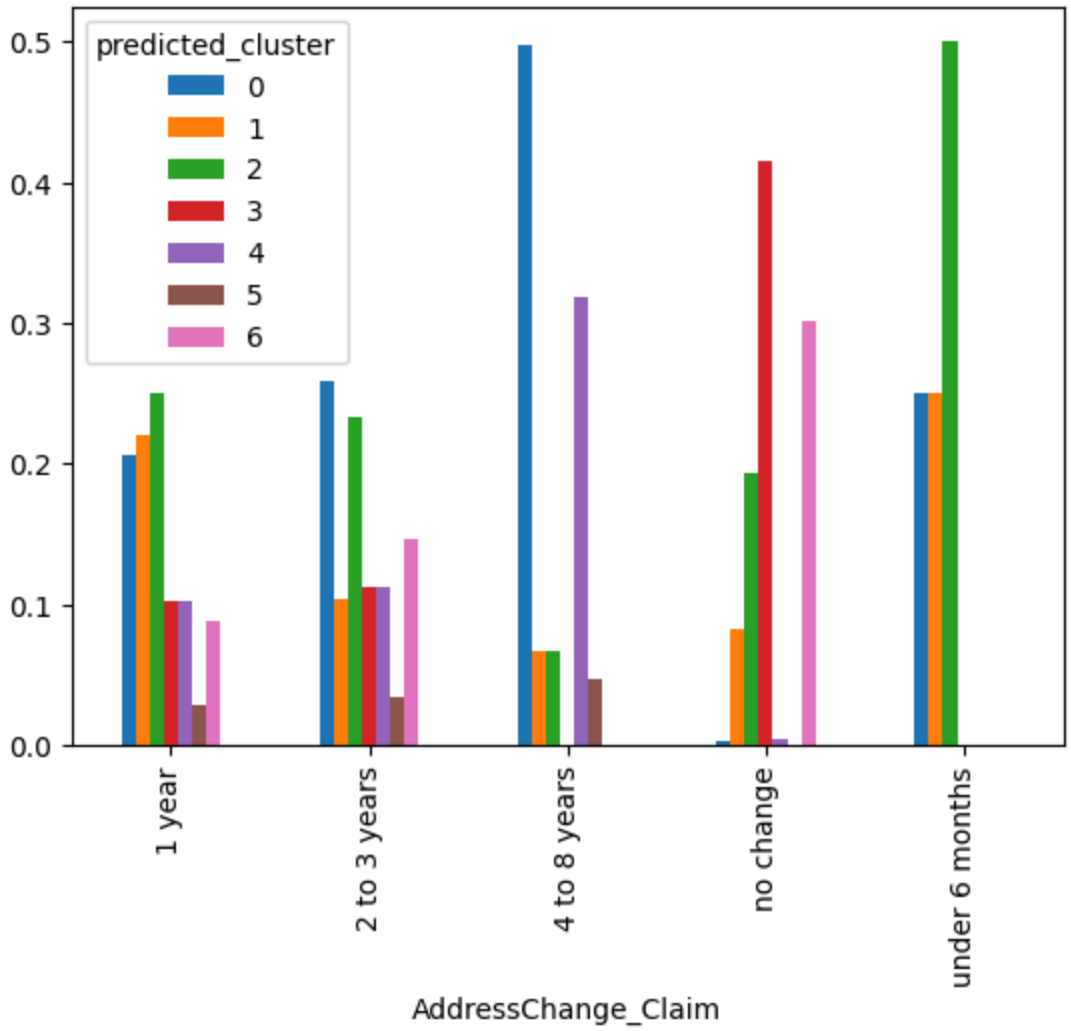

DriverRating
1
2
3
4
0.4
0.3
0.2
0.1
0.0
0
1
2
3
4
5
6
predicted_cluster


AgeOfVehicle
2 years
3 years
4 years
5 years
6 years
7 years
more than 7
new
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
1
2
3
4
5
6
predicted_cluster


NumberOfCars
1 vehicle
2 vehicles
3 to 4
5 to 8
more than 8
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
predicted_cluster


0.5
0.4
0.3
0.2
0.1
0.0
predicted_cluster
0
1
2
3
4
5
6
0
1
FraudFound_P


predicted_cluster
0
1
2
3
4
5
6
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
300
400
500
700
Deductible



10. Comparison

Clustering Comparison: WHAT was different?

* Were the results of these two clustering algorithms different? Calculate and interpret the adjusted rand score to measure how different overall your two partition-based clusterings were. If one of your clusterings was not a partition, convert it into one.

DO THIS

* In addition, more specifically describe what types of observations were different between these two clusterings and how they were different. t-SNE plots color-coded by the cluster labels can help you describe this

Based on the two TSNE plots that we found in each of our clustering algorithms, there were a lot of differences happening between our two forms of clusters. Firstly, for the K-prototype method, only two clusters were found to be the inherent form of our dataset, but for HAC 7 clusters were

found. As well, the traits that made up these clusters were entirely different. Where the k-prototype clustering appears to be separated by types of ages, for the HAC clustering, we were able to distinguish some clusters that were localized by whether or not fraud was in the cluster. As well, for the k-prototypes method, there are bound to be some errors because the TSNE plot was created with the gowers distance, but k-prototypes does not use gowers distance as the metric. Therefore HAC in general might match the inherent clustering structure a bit better than the k-prototype algorithm does.

Clustering Comparison: Ability to Identify the "Inherent" Clusters
* Do you have any reason to believe that one of these clusterings was better able to identify all the "inherent" clusters that exist in this dataset? Explain.

- While on the one hand I think k-prototype did a better job at identifying the inherent two clusters, HAC did a much better job in identifying the nested nature of the clusters that would better help with our research goals. The TSNE plot suggested about 2-3 big clusters with a lot of nested clusters. The k-prototype did identify to use 2 clusters, but it appears that these clusters were formed more so on a scaling issue with the age, rather than a true identification of the natural clusters. The HAC did a much better job in identifying the types of clusters that we were trying to look for. The were more distinct and had more defining characteristics to them. As well, HAC was able to understand nested nature of the dataset because of the agglomerative nature of the data.

Clustering Comparison: Clustering Trustworthiness/Usefulness
* Based on your analysis insights, do you have any reason to "trust" the results of one of your clustering
algorithms more when it comes to reflecting a meaningfully separated/usable set of clusters? Explain.

- I would trust the HAC clustering method more than the k-prototype method specifically because we believe the prototype method could have been impacted by a scaling issue, as well the gower's distance of the TSNE plot versus a different distance metric for k-prototype could have really impacted the usability of our clusters. As well, the k-prototype method didn't appear to have distinct information about them, and were therefore not super helpful with our research insights.

Clustering Comparion: Clustering Insights
* Does clustering algorithm #1 reveal any insights about the data that algorithm #2 does not? If so, what are they?

Clustering algorithm one reveals some issues with our method that clustering 2 did not. While I do not think this clustering method is helpful for our research goals, it is helpful for us to learn from and to fix for further analysis.

* Does clustering algorithm #2 reveal any insights about the data that algorithm #1 does not? If so, what are they?

Clustering algorithm two reveals a lot more for our research goals than the first algorithm does. Clustering algorithm two allowed us to actually make potential insights with our data. The first of which being base policies effect on potential fraud. Secondly the vehicle type. Then the person at fault in the accident. As well the deductible. All of these factors we found to be in play with whether or not fraud is found in an insurance claim. There were noticeable differences in these clusters in line with the TSNE plot that helped us understand our research goal, and how we can further understand fraud and where it could potentially come from.

11. Conclusion

Summarize:
Our report was to investigate and identify characteristics between people and vehicles that could lead us to believe there are instances of insurance fraud. To do this, we started by cleaning out insignificant columns and interpreting the remaining columns of the dataset by obtaining their summary statistics. We decided that scaling our dataset would not be a good idea due to the numerical and categorical split of our dataset and began cluster cluster analysis. We first started by doing a TSNE algorithm and found our dataset is indeed clusterable and then chose to proceed with k-protoypes and hierarchical agglomerative clustering with gower's distance to cluster our dataset because we wanted to work within a mixed dataset. Both clustering algorithms worked with our dataset and gave us useful insights in line with our research goals. They worked so well that we did not feel the need to do fuzzy clustering on top of our two previously used algorithms as we did not deem it useful enough to help further our research goal for this dataset. After running our algorithms it appears that the characteristics that are most consistent with fraud include All Perils insurance policy, third party liability, higher deductible, and sedan vehicles. Whereas characteristics that are more inherent with no fraud include liability insurance policy, lower deductibles, and sports cars. We will delve more into what this means in our recommendation.

Clustering algorithm 2:

- Base Policy: All perils (most expensive)
  liability / collision less fraud
- Fault: Third Party
- Higher deductible more fraud
- Sedan more likely to be involved in fraud
- Sport less likely to be involved in fraud

Recommendation:

These insights would be helpful in determining what types of rates and risks an insurance company can take on with certain clients. Based on our clustering findings we saw that all perils insurance policies are more likely to have instances of fraud. This in general makes sense because with all perils insurance rather than you having to prove something should be covered, an insurance company has to prove that the loss is not covered. This means there is a lot more of a gray area with what can be paid out with insurance claims. On the other hand, we saw that liability insurance is less consistent with fraud - meaning that in instances where you are paying for when you are at fault, there is less likely to be fraud on those claims. Our recommendation for insurance companies on this issue would be to look closely at the claims being made on all-perils policies to see if the claim is really a loss, or if it is just potential fraud. Consistent with this is the idea that third party fault is more likely to have fraud on it. Just like the liability insurance, the policy holder being liable is less likely to have fraud. This could be because the policy holder could exaggerate what types of claims they are making about the third party being at fault. As an example we have seen recently, some cars will purposely slam on the brakes to be rear ended to commit insurance fraud, i.e. a third party was at fault, but the claim was fraudulent. We also saw that higher deductibles are likely to have insurance fraud. This makes sense based on this study "The most common insurance fraud activity and one that contributes a significant portion of dollar losses is the practice of padding claim amounts in the event of a loss. One of the largest issues insurance companies face is that policyholders often do not perceive insurance claim padding as an unethical behavior. However, very little research has examined the factors that contribute to such perceptions. Considering how consumers often attempt to justify fraudulent behavior from a fairness perspective, the present work examines how the amount of the deductible in an insurance claim situation can influence feelings of fairness and ethicality. The results of an experimental study show that higher deductible amounts result in stronger perceptions that insurance claim padding is fair to the insurance company, weaker perceptions that the behavior is unethical, and higher proposed claim award amounts" (Miyazaki 589). Basically this study is saying that with higher deductibles, there is more room for padding, and more room for fraudulent behavior to be masked with the high deductibles. Our recommendation here would be to really look at how deductibles are being priced, and how those with high deductibles are using their rates. The final characteristic that we found noteworthy was that

sedans were more likely to be involved in fraud, and sports cars less likely to be involved in fraud. This makes sense because according to Progressive, “sedans tend to have a higher bodily injury claims frequency than SUVs, so they may be considered riskier and therefore more expensive to insure”. Our recommendation based on this would be to take the risk factor of sedans into consideration when pricing and assessing risk of the driver. We think that a lot of what we uncovered would be useful information for the insurance company to take into consideration or use when they are pricing rates or assessing risk of their clients. While these are just trends, and not casual links, it can still be useful - because actuarial science in general is assessing trends and forecasting what they think will happen.

We also would recommend that the person if they have more data uses a clustering structure that has nested properties since our HAC with gower’s distance worked a lot better than our K-Prototypes.

Miyazaki, A.D. Perceived Ethicality of Insurance Claim Fraud: Do Higher Deductibles Lead to Lower Ethical Standards?. J Bus Ethics 87, 589–598 (2009). https://doi.org/10.1007/s10551-008-9960-4

Corporation, Progressive. “Cheapest Cars to Insure.” Progressive, 3 Nov. 2023, www.progressive.com/answers/cheapest-cars-to-insure/.

Shortcomings/Caveats:

Obviously this analysis is based on trends rather than causal links. While there are indeed some characteristics that we deemed to produce more fraud than others, this will not give us a definitive answer for insurance companies. As well, there are so many other forms of characteristics about a person/car that were not taken into consideration when doing our clustering (because it was not in our data). Some shortcomings of our clustering analysis was that we were limited to only a few algorithms that could cluster mixed data. If our data had had more numerical information, we probably could have chosen different algorithms that would have clustered our data better. On the same note, since we had so many characteristics at play, there was bound to be instances of fraud/not fraud in almost every cluster. This makes our analysis a little more speculative than we would like, and we had to base our conclusions off of most of what the data was telling us, not all.

Future Work:

For future analysis, it would be interesting to collect further information about different characteristics of the people/cars. It would be interesting to know if the person has committed fraud before, if they’ve been involved in a lot of accidents recently, maybe some other

demographic information about them like how much they make. I also would be interested to run analysis on solely numerical data to see if different clustering algorithms could give us more nuanced findings. As well, I would also like to explore more data from more years, and if possible more recent data. Due to time constraints and the ability of our computer to run algorithms, we were limited to roughly 6000 data points to cluster, but on a larger scale or with more data, I would be interested to see where our analysis could take us!