Multivariate Analysis of Vehicle Insurance Claims

Fifi Zhang/23011135

School of Mathematical and Computational Sciences, Massey University

161.762: Multivariate Analysis for Big Data

Matthew Pawley

June 7th, 2023

Contents

Executive Summary

The aim of this paper is to gain insights into vehicle insurance claims by utilizing a range of multivariate analysis methods. Principle Component Analysis (PCA) and Factor Analysis (FA) are conducted to analyze the associations among customer, vehicle, policy, incident, and claim-related variables. It reveals that age and months as a customer are highly correlated and the total claim amount, vehicle claim, property claim, and injury claim are closely related to each other. Next, Clustering and Multiple Correspondence Analysis (MCA) are utilized to investigate customer segments. Then, Canonical Correlation Analysis (CCA) is performed to analyze the correlation between incident-related variables and the claim amount-related variables, which indicates that the incident-related variables such as the incident hour of the day and the number of vehicles involved are highly correlated with claim amount-related variables such as vehicle claim amount, property claim amount, and injury claim amount. Finally, Canonical Discriminant Analysis (CDA), Stepwise Discriminant Analysis (SDA), and Quadradic Discriminant Analysis (QDA) are deployed to analyze the incident severity groups differences and the classification accuracy. It shows that the most important variables in separating the incident severity groups are vehicle claim, the number of vehicles involved, and incident hour of the day, and the classification is not accurate since the misclassification rate is as large as 0.60.

1 Introduction

The report focuses on multivariate analysis in the field of vehicle insurance claims. Risk, customer, and cost are the three pillars of an insurance company, so it is crucial to better understand the customer's characteristics and vehicle insurance claims patterns (Abdelhadi et al., 2005). A total set of eight multivariate methods have been used in the study, including unconstrained analysis such as Principal Component Analysis (PCA), Factor Analysis (FA), Clustering, and constrained analysis like Multiple Correspondence Analysis (MCA), Canonical Correlation Analysis (CCA), Canonical Discriminant Analysis (CDA), Stepwise Discriminant Analysis (SDA), and Quadradic Discriminant Analysis (QDA). PCA and FA are used to analyze the relationship between customer, policy, incident-related variables, and claim amounts variables such as vehicle claim amount, property claim amount, and injury claim amount. Clustering and MCA are utilized to investigate the customer segments according to customer-related variables such as age, sex, education level, hobby, and auto-make. CCA is conducted to get insights into how the customer, policy, vehicle, and incident-related variables are correlated with the sets of vehicle claims amounts variables and how many correlations exist. CDA is used to see how the incident severity groups are different and SDA is used to obtain the key drivers to separate the groups. Finally, I perform QDA to see how accurately the observations be classified into incident severity groups.

The methods section describes the dataset and methodologies used in this report. The results sections illustrate the multivariate analysis outcome of the research questions. The discussion and conclusion summarize the key findings and future improvements.

2 Methods

2.1 Dataset Description

The vehicle insurance claim dataset is acquired from the Kaggle website, which shows claims data from 1000 auto insurance policies of a US insurer in 2015. It contains 35 variables and can be categorized into five groups which are customer-related variables (i.e. age, months as a customer, insured sex, insured education level, insured occupation, insured hobbies), vehicle-related variables (i.e. auto make, auto model, auto year), policy-related variables (i.e. policy number, policy annual premium, capital gains, capital loss), incident-related variables (i.e. Incident hour of the day, number of vehicles involved, bodily injuries, incident severity, incident date, incident location), and claim-related variables (i.e. total claim amount, injury claim, vehicle claim, property claim, fraud reported). The detailed information of variables is shown in Table 1.

Table 1: variable description

category	variable name	data type	description	format/value
customer	months_as_customer	Num	total monts of being customer of the insurer	328
	age	Num	the insured age	48
	insured_sex	Char	the insured sex	MALE
	insured_education_level	Char	the insured education level	MD
	insured_occupation	Char	the insured occupation	craft-repair
	insured_hobbies	Char	the insured hobby	sleeping

	insured_relationship	Char	the insured social relationship	husband
vehicle	auto_make	Char	auto make brand	Saab
	auto_model	Char	auto make brand model	92x
	auto_year	Num	auto production year	2004
policy	policy_annual_premium	Num	annual policy premium	1406.91
	capital-gains	Num	the profits earned from the policy	53300
	capital-loss	Num	the loss in value less than its original purchase price	0
	policy_number	Num	the policy number, primary key	521585
	policy_bind_date	Num	the date buying the policy	2014/10/17
	policy_state	Char	the state where the policy belongs to	OH
incident	incident_hour_of_the_day	Num	the hour the incident happens	5
	number_of_vehicles_involved	Num	the number of cars involved in the incident	1
	bodily_injuries	Num	the number of injuries in the incident	1
	witnesses	Num	the number of witnesses in the incident	2
	incident_date	Num	the data of incident	2015/1/25
	incident_type	Char	the type of incident	single vehicle collision, multiple vehicle collision, vehicle theft, parked car
	collision_type	Char	the type of collision	front collision, side collision, rear collision,
	incident_severity	Char	the severity of incident	major damage, minor damage, total loss, trivial damage
	authorities_contacted	Char	the authorities contacted in the incident	Police
	incident_state	Char	the state where the incident happened	SC

	incident_city	Char	the city where the incident happened	Columbus
	incident_location	Char	the incident location	9935 4th Drive
	property_damage	Char	whether there is a property damage	YES
	police_report_available	Char	whether there is a police report	YES
claim	total_claim_amount	Num	the total claim amount	71610
	injury_claim	Num	injury claim amount	6510
	property_claim	Num	property claim amount	13020
	vehicle_claim	Num	vehicle claim amount	52080
	fraud_reported	Char	whether it is a fraud claim	Y

2.2 Methodologies

1. Principle Component Analysis (PCA)

PCA is a dimensionality reduction technique used to transform data into a new coordinate system with principal components (Besse & Ramsay, 1986). It works best to analyze the structure of high-dimensional data, reduce noise, and visualize patterns. PCA aims to explain the maximum amount of variance in a dataset by identifying linear combinations of variables (principal components) that capture the most variation in the data. It assumes that the observed variables are directly related to a small number of underlying factors. It is worth mentioning that the principal components are orthogonal, and it is suitable to use PCA when the number of variables is less than the number of observations. Besides, avoid using them when there are lots of zeros and data are highly skewed.

2. Factor Analysis (FA)

FA is also a dimensionality reduction technique to model the underlying structure of a set of variables. It aims to identify the common factors that are responsible for the correlation between variables. FA assumes that each variable is influenced by multiple underlying factors. The factors can be orthogonal or oblique. Like PCA, it is suitable for the dataset with more variables than the observations.

3. Clustering

Clustering refers to unsupervised learning techniques used to group similar data points based on their features. It is a distance-based method, so it is sensitive to the scaling. The common methods include K-means and hierarchical clustering.

4. Multiple Correspondence Analysis (MCA)

CA is an exploratory data analysis technique for visualizing the associations between categorical variables. It contains simple correspondence analysis and multiple correspondence analysis.

5. Canonical Correlation Analysis (CCA)

CCA is used to measure the linear relationship between two sets of variables. It needs the response variables and predictor variables. The research questions are like “How a set of variables x are correlated with the other set of variables y?” and “How many correlations of x and y exist?”

6. Canonical Discriminant Analysis (CDA)

CDA is a dimensionality reduction technique for classification that seeks to maximize the separation between classes. The axis needs to meet two criteria simultaneously that are maximizing the distance between the group centroids and minimizing the variation within each class. It's used in pattern recognition, image analysis, and feature selection.

7. Stepwise Discriminant Analysis (SDA)

SDA is also a dimensionality reduction technique used to reduce variables and find the key drivers to discriminate groups.

8. Quadratic Discriminant Analysis (QDA)

QDA is a supervised classification technique that models the decision boundary between classes and evaluates expected error rates using quadratic functions. It assumes groups have different covariance.

3 Results

3.1 Data preprocessing

I conduct data cleaning before performing multivariate analysis which is handling the missing value by imputing them with the mean value in numerical variables and the most frequently occurred value in categorical variables.

3.2 Multivariate Analysis

3.2.1 Principle Component Analysis (PCA)

PCA is utilized to explore the relationship among 17 numerical variables, including 2 customer-related variables such as age and months as a customer, 1 vehicle-related variable like auto-year, 5 policy-related variables like policy annual premium, 5 incident-related variables like the incident hour of the day, the number of vehicles involved, and bodily injuries, and 4 claim-related variables like total claim amount.

1. Correlation matrix

Table 2 shows the interrelationships among these variables. It indicates that there is a highly positive correlation between “age” and “months_as_customer” with correlation coefficients as high as 0.92, as shown in red in Table 2. What’s more, It indicates that there is highly positive correlation between “total_claim_amount” and “vehicle_claim”, “property_claim”, and “injury_claim” with the coefficients of 0.81, 0.81, and 0.98 respectively, which is mainly because the total claim amount is the aggregate value of the other three. If these highly correlated variables are included in the dataset in an ordinary regression analysis, it would lead to multicollinearity and overfitting problem, so I drop the “months_as_customer” and “total_claim_amount” variables in the latter analysis.

Table 2: correlation matrix of the customer, vehicle, policy, incident, and claim-related variables

variables	**age**	**auto_year**	**bodily_injuries**	**capital-gains**	**capital-loss**	**incident_hour_of_the_day**	**injury_claim**	**months_as_customer**	**number_of_vehicles_involved**	**policy_annual_premium**	**property_claim**	**total_claim_amount**	**vehicle_claim**	**witnesses**
age	1.00	0.00	-0.02	-0.01	0.01	0.09	0.08	**0.92**	0.02	0.01	0.06	0.07	0.06	0.05

auto_year	0.00	1.00	-0.02	0.03	-0.06	0.02	-0.01	0.00	0.03	-0.05	-0.01	-0.04	-0.04	0.05
bodily_injuries	-0.02	-0.02	1.00	0.06	-0.02	-0.03	0.05	-0.01	0.01	0.03	0.04	0.05	0.04	-0.01
capital-gains	-0.01	0.03	0.06	1.00	-0.05	-0.02	0.03	0.01	0.06	-0.01	0.00	0.02	0.02	-0.02
capital-loss	0.01	-0.06	-0.02	-0.05	1.00	-0.03	-0.05	0.02	-0.01	0.02	-0.02	-0.04	-0.03	-0.04
incident_hour_of_the_day	0.09	0.02	-0.03	-0.02	-0.03	1.00	0.17	0.07	0.12	0.00	0.18	0.22	0.22	0.01
injury_claim	0.08	-0.01	0.05	0.03	-0.05	0.17	1.00	0.07	0.22	-0.02	0.56	**0.81**	0.72	-0.02
months_as_customer	**0.92**	0.00	-0.01	0.01	0.02	0.07	0.07	1.00	0.01	0.01	0.03	0.06	0.06	0.06
number_of_vehicles_involved	0.02	0.03	0.01	0.06	-0.01	0.12	0.22	0.01	1.00	-0.05	0.22	0.27	0.27	-0.01
policy_annual_premium	0.01	-0.05	0.03	-0.01	0.02	0.00	-0.02	0.01	-0.05	1.00	-0.01	0.01	0.02	0.00
property_claim	0.06	-0.01	0.04	0.00	-0.02	0.18	0.56	0.03	0.22	-0.01	1.00	**0.81**	0.73	0.05
total_claim_amount	0.07	-0.04	0.05	0.02	-0.04	0.22	**0.81**	0.06	0.27	0.01	**0.81**	1.00	**0.98**	-0.01
vehicle_claim	0.06	-0.04	0.04	0.02	-0.03	0.22	0.72	0.06	0.27	0.02	0.73	**0.98**	1.00	-0.02
witnesses	0.05	0.05	-0.01	-0.02	-0.04	0.01	-0.02	0.06	-0.01	0.00	0.05	-0.01	-0.02	1.00

2. Choose the number of eigenvectors

According to the scree plot as shown in Figure 1 and eigenvalues of the correlation matrix as shown in Table 3, I choose five principal components based on the criterion of eigenvalues greater than 1.0.

Figure 1 shows the principal component's eigenvalues and the proportion of variance they account for, which is the visualization of Table 3. The elbow occurs at the third principal component, which means that there are two large eigenvalues accounting for approximately 50% of the variation and three eigenvalues approaching 1.0.

Figure 1: the scree and variance plots of PCA

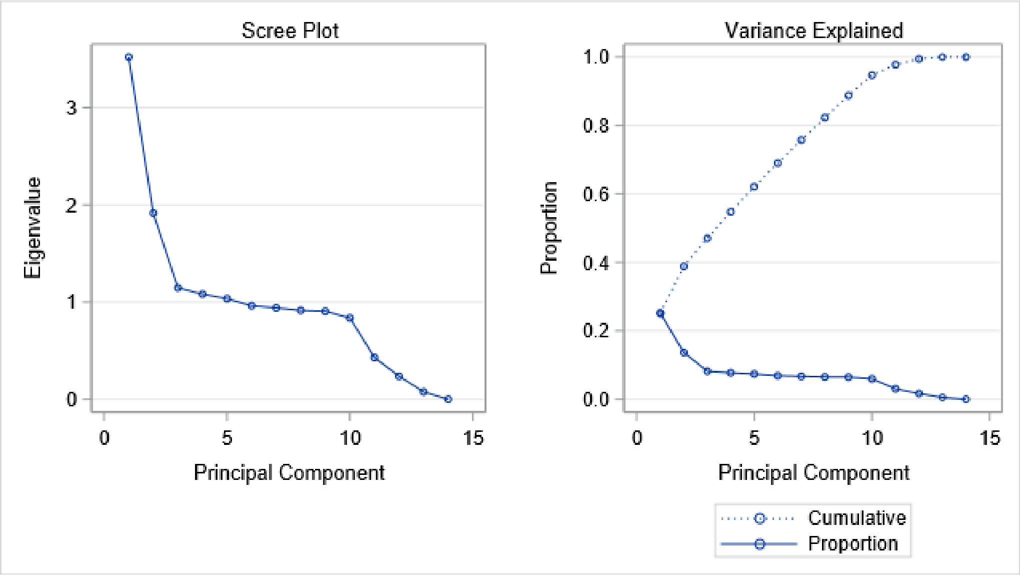


Table 3 shows that eight components are needed to account for 82% of the variation, but only the first five eigenvalues are greater than 1.0. In this case my criterion of choosing the number of eigenvectors is eigenvalues greater than 1.0, thus I keep five components.

Table 3: eigenvalues of the correlation matrix

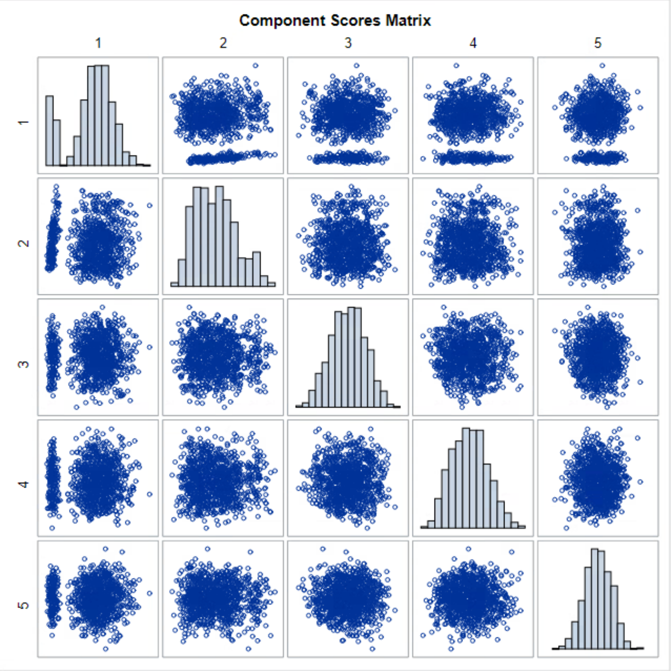
dimension	Eigenvalue	Proportion	Cumulative
1	**3.52**	0.25	0.25
2	**1.92**	0.14	0.39
3	**1.15**	0.08	0.47
4	**1.08**	0.08	0.55
5	**1.03**	0.07	0.62
6	0.96	0.07	0.69
7	0.94	0.07	0.76
8	0.91	0.07	0.82
9	0.91	0.06	0.89
10	0.84	0.06	0.95
11	0.43	0.03	0.98
12	0.24	0.02	0.99
13	0.08	0.01	1.00

14	0.00	0.00	1.00

3. Component scores matrix

Figure 2 shows the distribution of each principal component score and a scatter plot of each pair of component scores. These are the observations in the data set and their scores on the first five components. Score plots are useful for finding patterns in the observations such as outliers or groupings of observations. Comparing component 1 to components 2 to 5, it shows a horizontal line pattern and scattered points. It may suggest a strong association among the observations regarding the underlying structure represented by component 1.

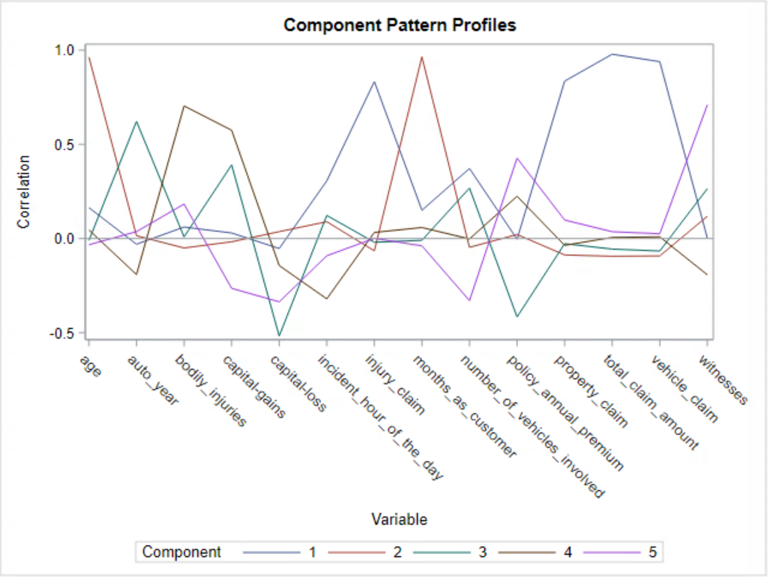
Figure 2: the component scores matrix



4. Component pattern profiles

Figure 3 shows the correlation between each variable and each extracted component. For instance, the blue line representing Component 1 is most highly correlated with the injury claim amount, property claim amount, vehicle claim amount, and total claim amount, while Component 2 which is the orange line is highly correlated with age and months as a customer and Component 4 as shown the brown line is highly correlated with bodily injuries and capital gains.

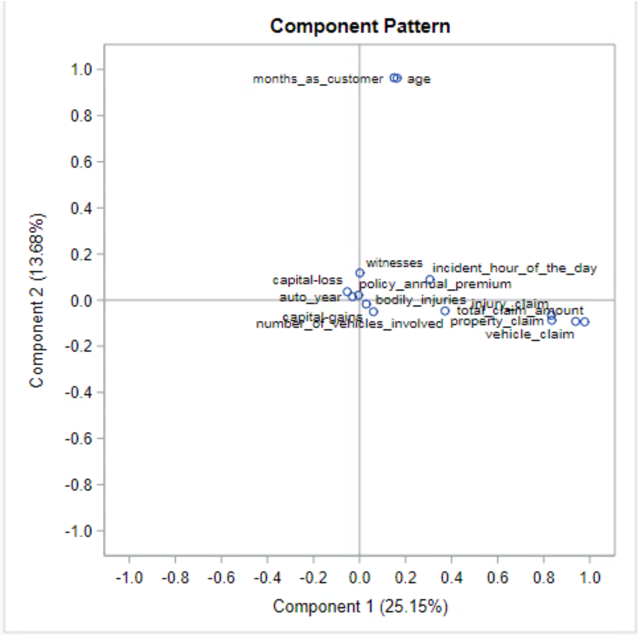
Figure 3: the component pattern profiles



5. Component plot

Figure 4 also shows similar variable-component associations. For example, on the right side of the plot, the claim amount-related variables have high associations with component 1 but have no relation with component 2. On the contrary, on the top of the plot, age and months as a customer are highly correlated with component 2 while having no relation with component 1.

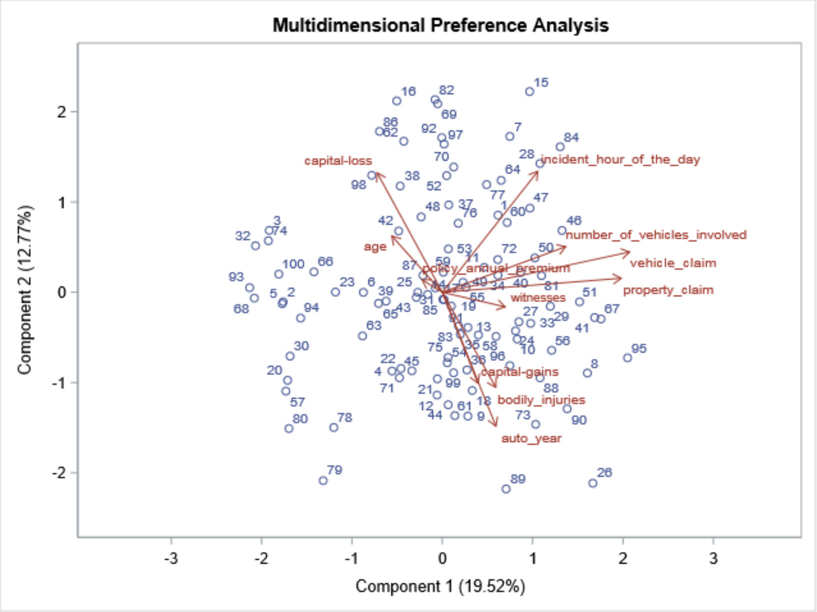
Figure 4: component plot of Component 1 and Component 2



6. Biplot

I randomly choose 100 observations to draw a biplot of these variables, and it shows similar results. For instance, age and capital loss are highly correlated due to the same direction and the same with claim amount-related variables. Furthermore, bodily injuries, capital gains, and auto years are closely related to each other.

Figure 5: biplot of Component 1 and Component 2



s

3.2.2 Factor Analysis (FA)

FA is also conducted to analyze the relationship among customer, policy, vehicle, incident, and claim amount-related variables. Both eigenvalues and scree plots suggest one factor. Table 4 shows the first eigenvalue account for more than 100% of the shared variance and in Figure 6 the elbow also occurs at the second eigenvector.

Table 4: eigenvalues of the reduced correlation matrix

dimension	Eigenvalue	Proportion	Cumulative
1	**2.10**	1.05	1.05
2	0.17	0.09	1.14
3	0.14	0.07	1.21
4	0.08	0.04	1.25
5	0.04	0.02	1.27
6	-0.01	0.00	1.27
7	-0.03	-0.01	1.25
8	-0.06	-0.03	1.22
9	-0.07	-0.04	1.19
10	-0.09	-0.04	1.15
11	-0.12	-0.06	1.08
12	-0.17	-0.08	1.00

Figure 6: the scree and variance plot

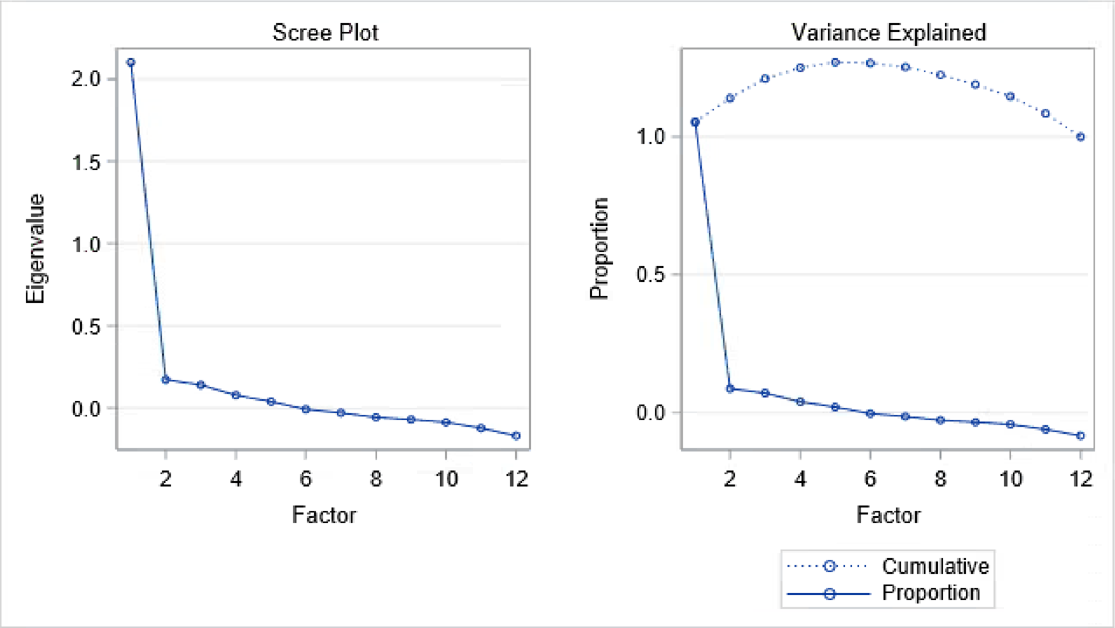


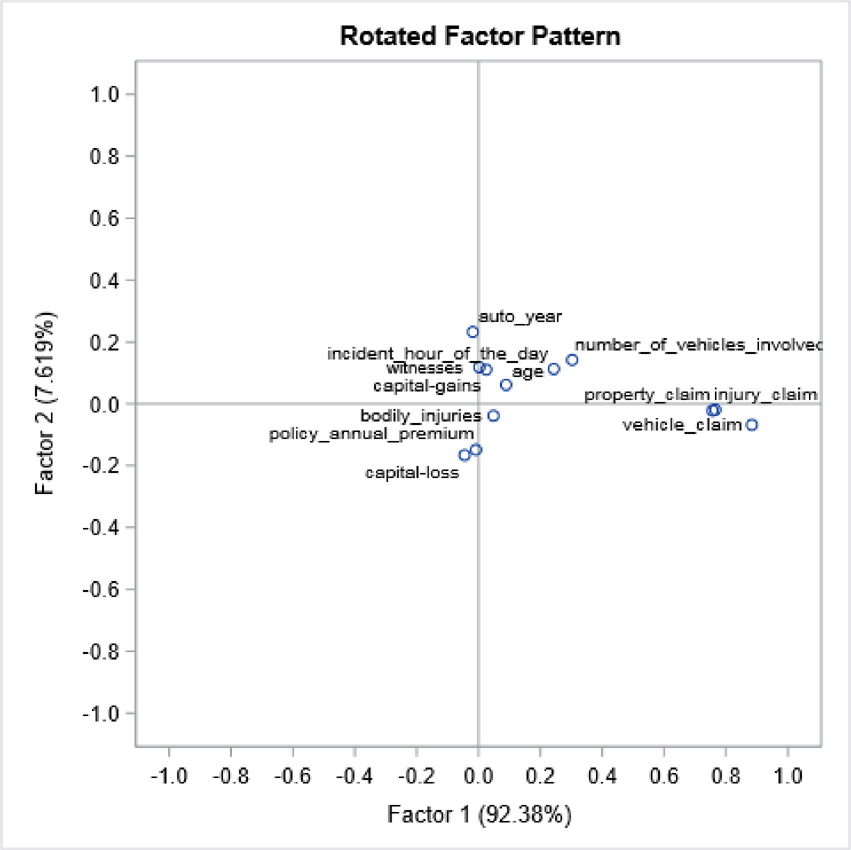
Table 5 shows the standardized regression coefficients predicting variables from the factor, which indicates that there is a high correlation between claim amount-related variables and factor 1, with coefficients of 0.76, 0.77, and 0.78 respectively, as shown in the red. Figure 7 plots the rotated factor pattern, which also illustrates the same relationship, as shown on the right side.

Table 5: factor pattern

variables	Factor1
age	0.09
auto_year	-0.02
bodily_injuries	0.05
capital-gains	0.02
capital-loss	-0.04
incident_hour_of_the_day	0.24
injury_claim	**0.76**
number_of_vehicles_involved	0.30
policy_annual_premium	-0.01

property_claim	**0.77**
vehicle_claim	**0.88**
witnesses	0.00

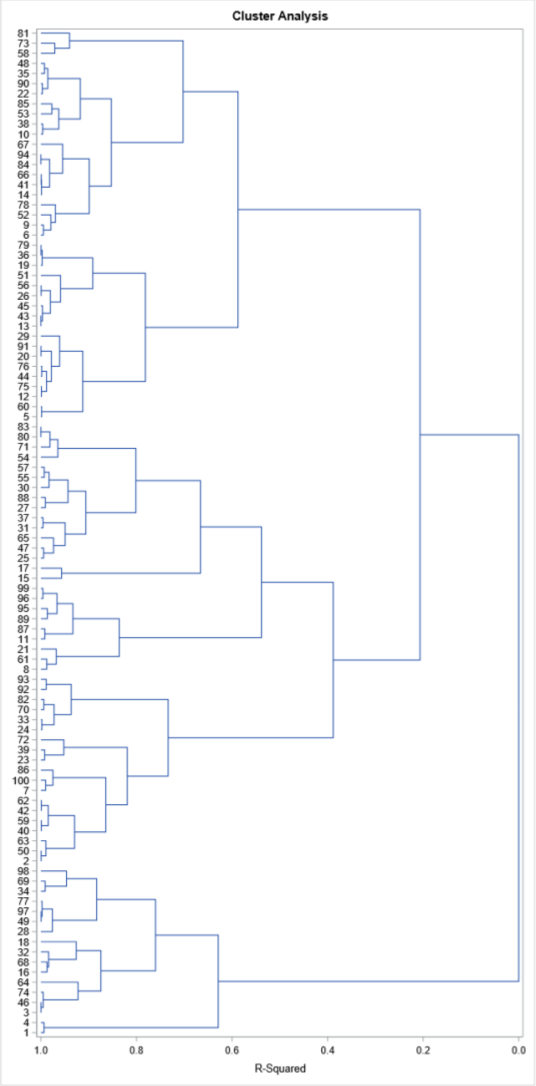
Figure 7: rotated factor pattern



3.2.3 Clustering

Clustering is used to analyze customer segment patterns based on numerical variables such as age, auto year, and policy annual premium. As shown in Figure 8, as the number of branches grows to the left from the root, the R square approaches 1, and the first five clusters account for about 95% of the variation. In other words, five clusters explain nearly all the variation, but it is not clear to see which segments can be divided into.

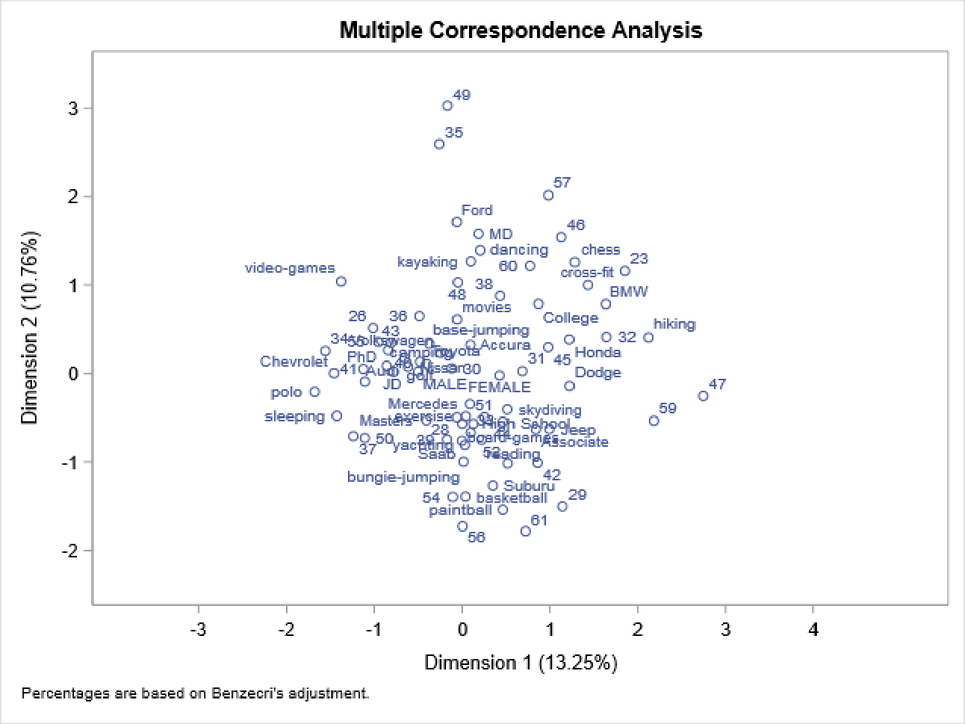
Figure 8: the dendrogram of customer-related variables



3.2.4 Multiple Correspondence Analysis (MCA)

MCA is utilized to see the associations between customer-related variables such as age, sex, education level, hobby, and auto-make. We can observe specific patterns in the plot. For example, individuals with Ph.D. or JD degrees show a strong interest in camping, and their preferred vehicle choices often include Volkswagen, Chevrolet, Nissan, and Audi. Furthermore, individuals with a high school education level tend to enjoy activities such as skydiving, yachting, exercising, and driving Mercedes cars. Additionally, MDs (Medical Doctors) have a passion for dancing and driving a Ford. Individuals around the age of 55 love basketball and paintball, and they frequently opt for Subaru as their own car.

Figure 9: multiple correspondence analysis

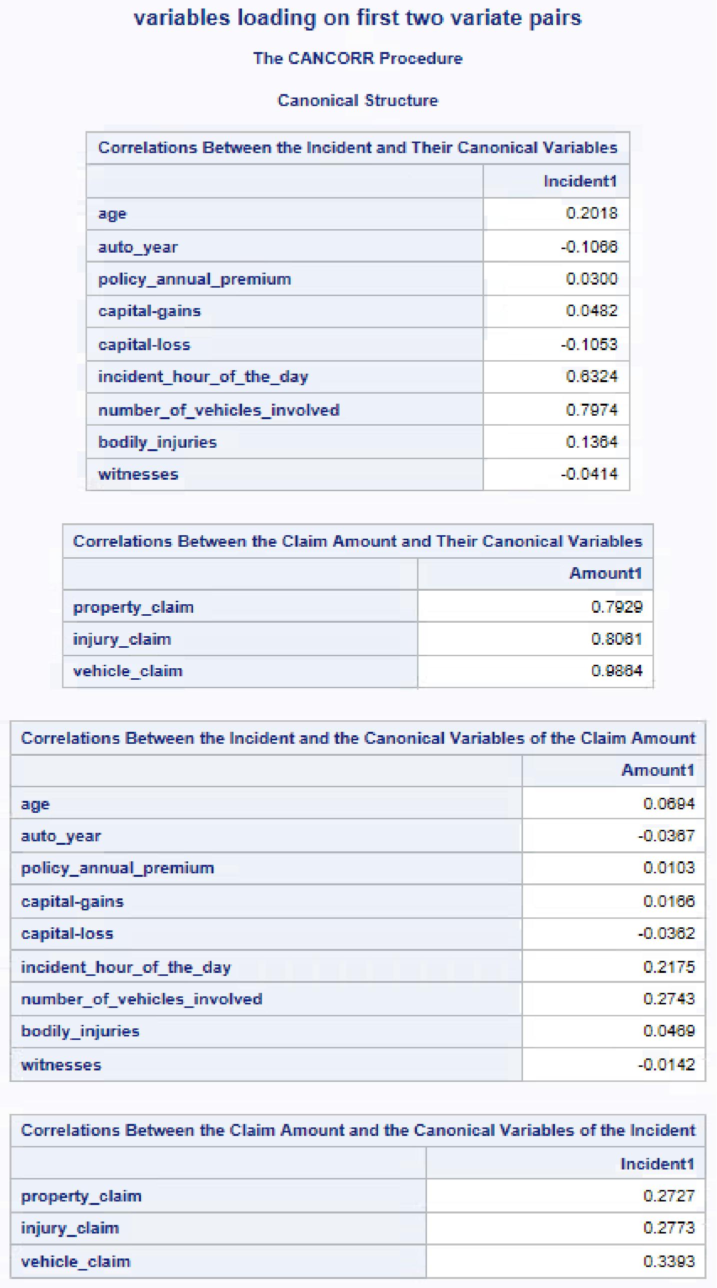


3.2.5 Canonical Correlation Analysis (CCA)

CCA is utilized to analyze how the sets of incident-related variables are correlated with the claim amount-related variables. The results in following tables and figures indicate that there are certain associations between incident features and claim amounts.

Table 6 contains four sub-tables. The first table shows the correlation between the original incident variables and the Incident variate. It indicates that the Incident canonical variate seems to be highly correlated with the incident hour of the day and the number of vehicles involved, with coefficients of 0.63 and 0.79 respectively. Thus, the Incident appears to be the extent of incident damage. The second table shows the correlation between injury, property, vehicle claim amount variables, and Amount variate, which appears high correlation. The third table shows a similar pattern emerging from the first table. For example, the incident hour of the day and the number of vehicles involved are relatively highly correlated with Amount variate, with the coefficients of 0.21 and 0.27 respectively. The last table also shows the same patterns as the second table. Overall, the incident-related items (the incident hour of the day and the number of vehicles involved) that are strongly correlated with the Incident variate are also highly correlated with the corresponding Amount variate. Therefore, there are certain associations between incident features and claim amounts.

Table 6: canonical structure of variables loading on first two variate pairs



variables loading on first two variate pairs

The CANCORR Procedure

Canonical Structure

Correlations Between the Incident and Their Canonical Variables	
	Incident1
age	0.2018
auto_year	-0.1066
policy_annual_premium	0.0300
capital-gains	0.0482
capital-loss	-0.1053
incident_hour_of_the_day	0.6324
number_of_vehicles_involved	0.7974
bodily_injuries	0.1364
witnesses	-0.0414

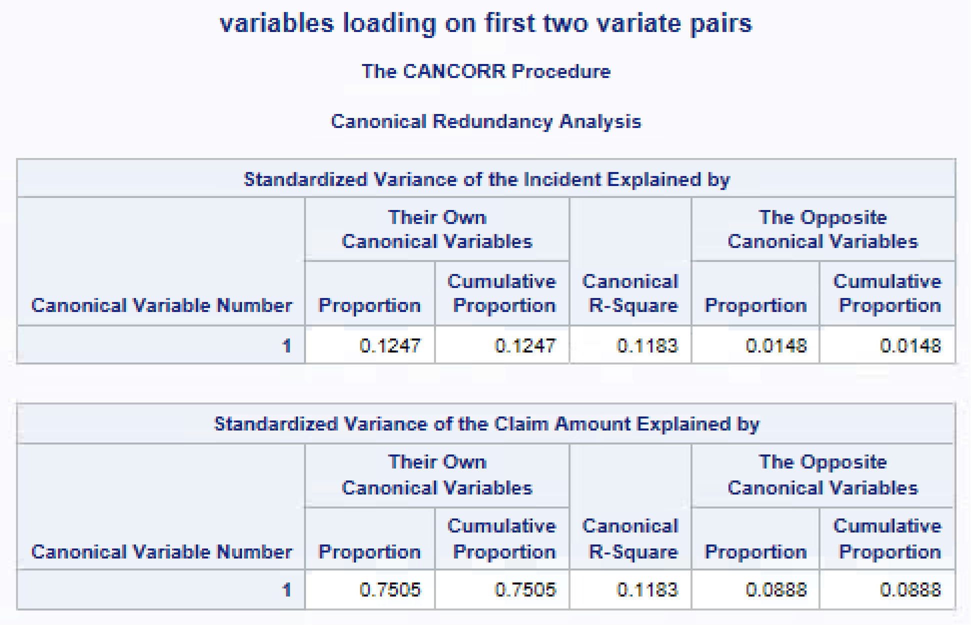
Correlations Between the Claim Amount and Their Canonical Variables	
	Amount1
property_claim	0.7929
injury_claim	0.8061
vehicle_claim	0.9864

Correlations Between the Incident and the Canonical Variables of the Claim Amount	
	Amount1
age	0.0694
auto_year	-0.0367
policy_annual_premium	0.0103
capital-gains	0.0166
capital-loss	-0.0362
incident_hour_of_the_day	0.2175
number_of_vehicles_involved	0.2743
bodily_injuries	0.0469
witnesses	-0.0142

Correlations Between the Claim Amount and the Canonical Variables of the Incident	
	Incident1
property_claim	0.2727
injury_claim	0.2773
vehicle_claim	0.3393

Table 7 shows that about 12% of the variance in the incident variables is explained by their own canonical variate while there is only 1.5% is explained by the opposite variate. Besides, 75% of the variance in the claim amount variables is explained by their own variate while only 9% is explained by the opposite variate.

Table 7: standardized variance of canonical variates



variables loading on first two variate pairs

The CANCORR Procedure

Canonical Redundancy Analysis

Standardized Variance of the Incident Explained by

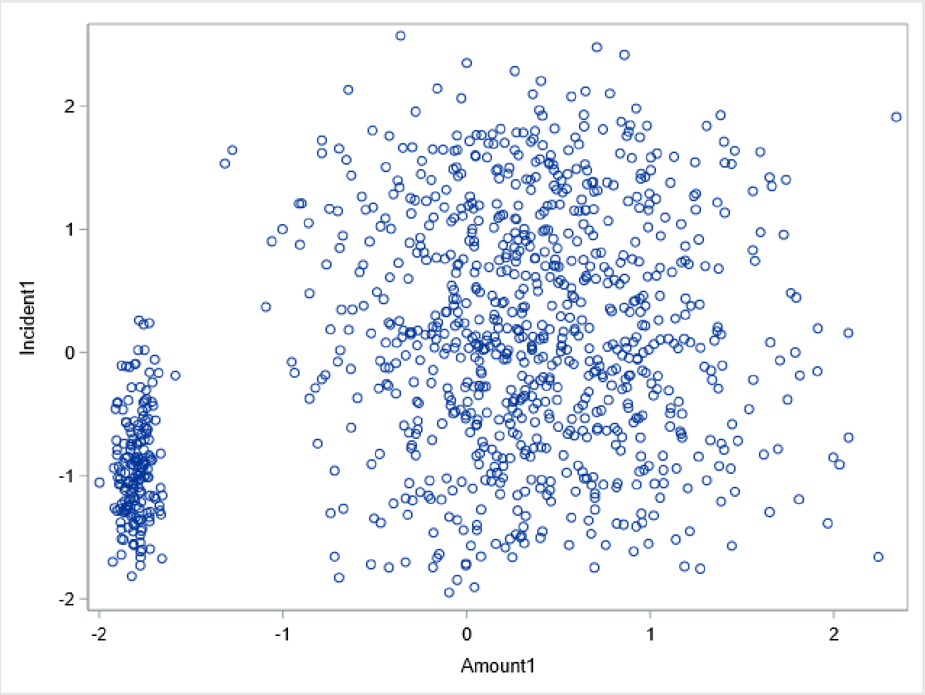
Canonical Variable Number	Their Own Canonical Variables		Canonical R-Square	The Opposite Canonical Variables	
	Proportion	Cumulative Proportion		Proportion	Cumulative Proportion
1	0.1247	0.1247	0.1183	0.0148	0.0148

Standardized Variance of the Claim Amount Explained by

Canonical Variable Number	Their Own Canonical Variables		Canonical R-Square	The Opposite Canonical Variables	
	Proportion	Cumulative Proportion		Proportion	Cumulative Proportion
1	0.7505	0.7505	0.1183	0.0888	0.0888

In Figure 10, the vertical line pattern and the scattered data cloud pattern refer to different aspects of the relationship between these two canonical variate pairs. The former indicates a strong relationship between the canonical variate pairs, in other words, a high degree of linear association between incident-related variables and claim amount-related variables. However, the scattered data cloud suggests a weaker relationship.

Figure 10: scatter plot of canonical variate pairs



3.2.6 Canonical Discriminant Analysis (CDA)

CDA is deployed to see how the incident severity groups differ based on numerical variables. Table 8 shows the proportion of classes in incident severity with the major damage, minor damage, total loss, and trivial damage accounting for 0.28, 0.35,0.28, and 0.09 respectively.

Table 8: class level information

incident severity	Frequency	Proportion
Major Damage	276	**0.28**
Minor Damage	354	**0.35**
Total Loss	280	**0.28**
Trivial Damage	90	**0.09**

In Table 9, upon exploring the variables that are closely related to the first discriminant function (injury claim, property claim, and vehicle claim), it suggests that these variables are all claim amount variables. Thus, it appears that the claim amount variables are responsible for an important portion of the discrimination among the three groups. What’s more, examining the second discriminant function, it implies that the variables pertaining to capital loss and witnesses are responsible for an important portion of the discrimination among the groups.

Table 9: pooled within the canonical structure

Variable	Can1	Can2	Can3
age	0.05	0.09	**0.51**
auto_year	0.02	0.08	0.13
bodily_injuries	0.00	0.26	0.10
capital-gains	0.04	-0.35	0.27
capital-loss	-0.05	**0.49**	0.21
incident_hour_of_the_day	0.24	-0.28	**-0.48**
injury_claim	**0.63**	-0.12	0.35
number_of_vehicles_involved	0.36	0.10	-0.08
policy_annual_premium	0.03	-0.19	0.12
property_claim	**0.66**	0.34	0.09
vehicle_claim	**0.95**	0.02	0.06
witnesses	0.00	0.46	-0.35

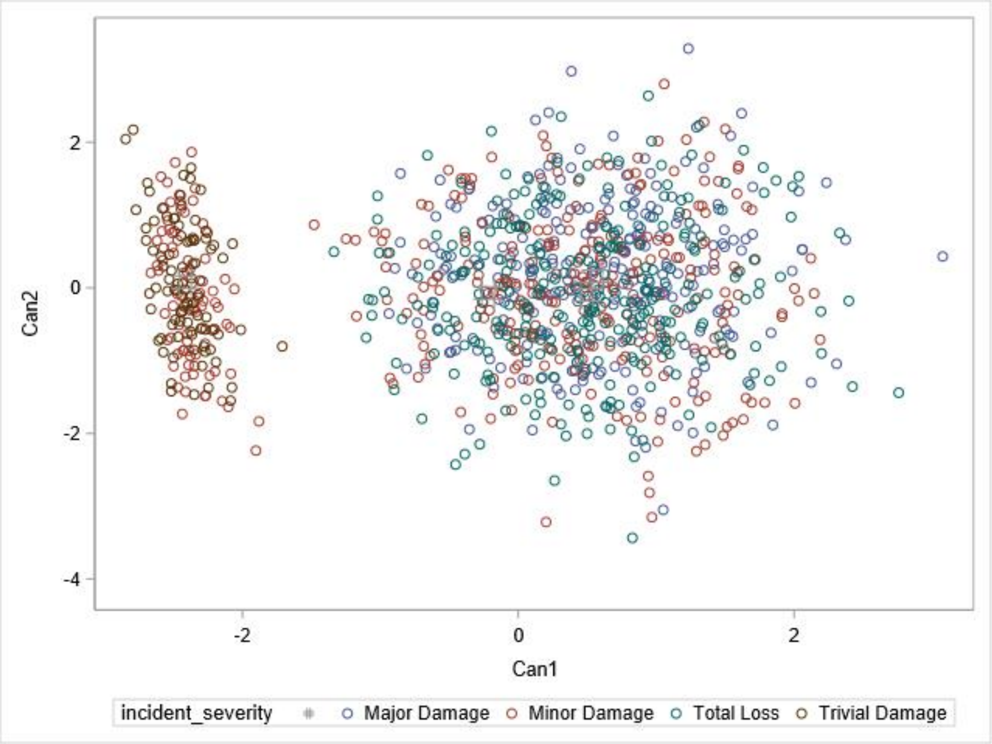
Table 10 shows the means of the subgroups on each of the canonical discriminant functions. It indicates that the first canonical discriminant function seems to be mostly separating the Major and the Trivial damage, and the second canonical discriminant function appears to be separating the Major damage from the Total loss group.

Table 10: class means on canonical variables

incident_severity	Can1	Can2	Can3
Major Damage	**0.57**	**0.16**	0.01
Minor Damage	-0.22	-0.07	0.10
Total Loss	0.50	**-0.10**	-0.11
Trivial Damage	**-2.42**	0.08	-0.10

Figure 11 plots the group centroids in the space defined by the canonical discriminant functions to get an idea of how well the group centroids are separated. The same results can be found in Table 10.

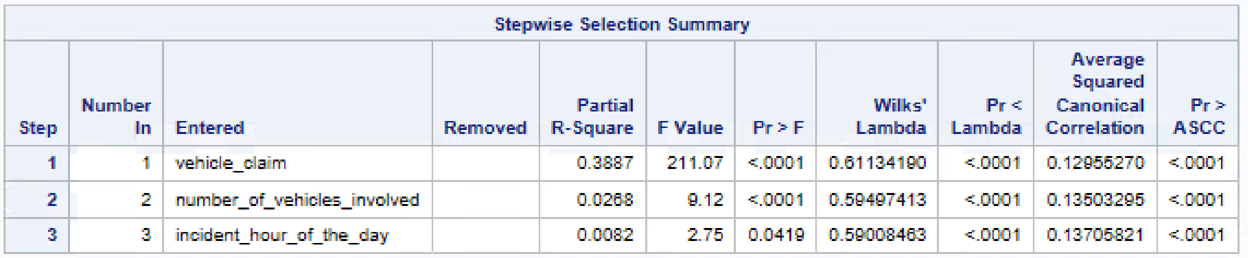
Figure 11: scatter plot of canonical discriminant functions



3.2.7 Stepwise Discriminant Analysis (SDA)

SDA is used to find the key drivers to separate the incident severity groups. The result shows that vehicle claim, number of vehicles involved, and incident hour of the day are the most important variables impacting the separating of incident severity groups.

Table 11: stepwise selection summary



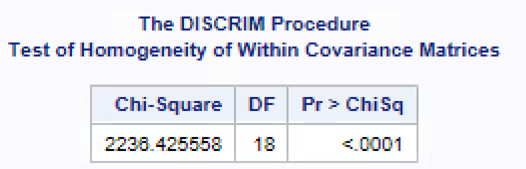
Stepwise Selection Summary

Step	Number In	Entered	Removed	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda	Average Squared Canonical Correlation	Pr > ASCC
1	1	vehicle_claim		0.3887	211.07	<.0001	0.61134190	<.0001	0.12955270	<.0001
2	2	number_of_vehicles_involved		0.0268	9.12	<.0001	0.59497413	<.0001	0.13503295	<.0001
3	3	incident_hour_of_the_day		0.0082	2.75	0.0419	0.59008463	<.0001	0.13705821	<.0001

3.2.8 Quadratic Discriminant Analysis (QDA)

QDA is utilized to analyze how accurately can observations be classified into incident severity groups. Since the Chi-square value is significant at 0.05 level, as shown in Table 12, the null hypothesis that the covariance matrices are homogeneous can be rejected. Therefore, QDA is more suitable than Fisher Linear Discriminant Analysis in this case.

Table 12: the test of homogeneity of the within-covariance matrices

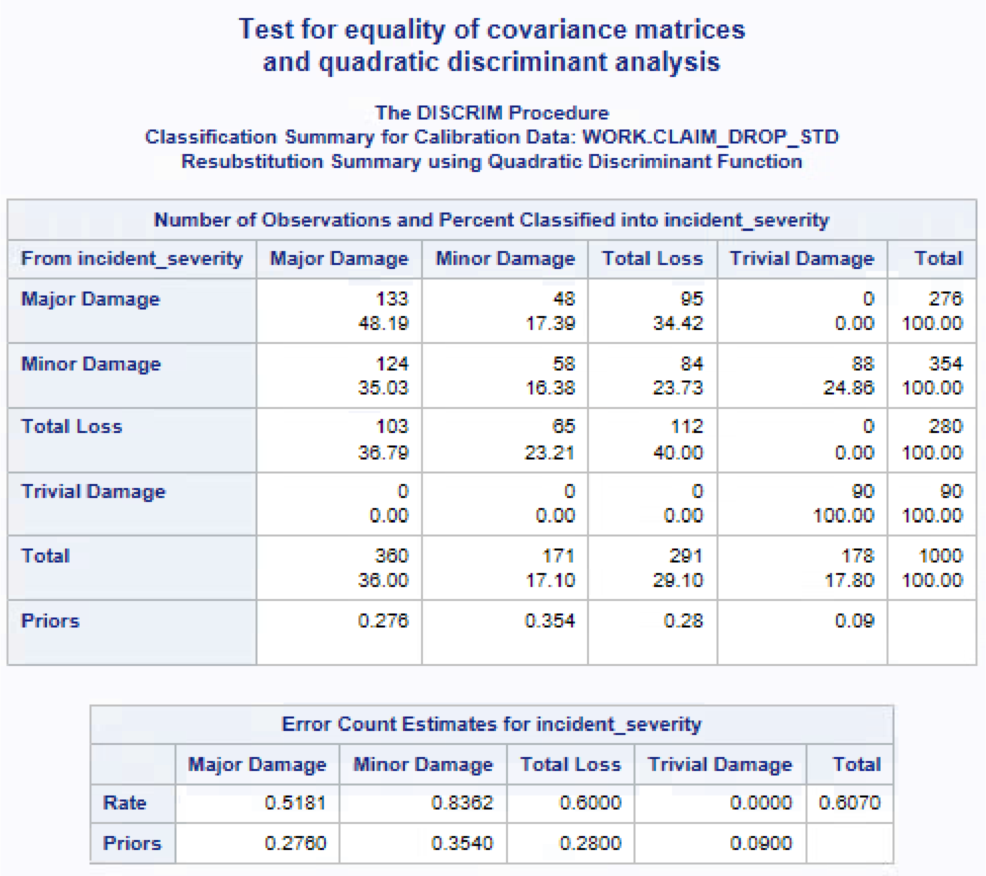


The DISCRIM Procedure
Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
2236.425558	18	<.0001

Table 13 reveals that the total misclassification rate is 0.6, which is relatively high. Specifically, out of a total of 276 cases of major damage, 133 cases are correctly classified, resulting in an error rate of 0.51. What's more, of minor damage, only 58 out of 354 cases are correctly classified, indicating a high error rate of 0.83. Additionally, the error rate for total loss is 0.60, while trivial damage is correctly classified due to its low proportion.

Table 13: the misclassification count and estimated error count



Test for equality of covariance matrices
and quadratic discriminant analysis

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.CLAIM_DROP_STD
Resubstitution Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into incident_severity					
From incident_severity	Major Damage	Minor Damage	Total Loss	Trivial Damage	Total
Major Damage	133 48.19	48 17.39	95 34.42	0 0.00	276 100.00
Minor Damage	124 35.03	58 16.38	84 23.73	88 24.86	354 100.00
Total Loss	103 36.79	65 23.21	112 40.00	0 0.00	280 100.00
Trivial Damage	0 0.00	0 0.00	0 0.00	90 100.00	90 100.00
Total	360 36.00	171 17.10	291 29.10	178 17.80	1000 100.00
Priors	0.276	0.354	0.28	0.09	

Error Count Estimates for incident_severity					
	Major Damage	Minor Damage	Total Loss	Trivial Damage	Total
Rate	0.5181	0.8362	0.6000	0.0000	0.6070
Priors	0.2760	0.3540	0.2800	0.0900	

4 Discussion

Eight different multivariate analysis methods are employed to address the research questions across the four categories. Firstly, PCA and FA are utilized to analyze the associations among customer, vehicle, policy, incident, and claim-related variables. The five different outputs of PCA, including the correlation matrix, the component scores matrix, the component pattern profiles, the component plot of Component 1 and Component 2, and the biplot, reveal consistent findings, indicating that variables such as age and months as a customer are closely related to each other. Furthermore, the total claim amount, vehicle claim, property claim, and injury claim are highly correlated with each other, which is verified by the rotated factor pattern in FA.

Secondly, clustering and CA are used to explore the customer segments. Clustering results, as demonstrated by the dendrogram, reveal the possibility of forming five clusters based on variables such as age, auto year, and policy annual premium. However, it does not explicitly indicate the specific segments that can be distinguished. To gain further insights, an additional analysis using MCA is conducted, incorporating variables such as age, sex, education level, and hobby. This analysis reveals interesting insights, including the fact that individuals holding Ph.D. or JD degrees exhibit strong interests in camping, often preferring vehicles such as Volkswagen, Chevrolet, Nissan, and Audi while individuals with a high school education level tend to engage in activities such as skydiving, yachting, exercising, and driving Mercedes cars. These findings can be leveraged to inform the strategic planning of customer segments for the insurance company.

Thirdly, CCA is deployed to analyze how the incident-related variables are correlated with the claim amount-related variables. Three outputs, including canonical structure, the standardized variance of canonical variates, and scatter plot of canonical variate pairs, reveal similar outcomes that the incident-related variables like the incident hour of the day and the number of vehicles involved are highly correlated with claim amount-related variables such as vehicle claim amount, property claim amount and injury claim amount.

Finally, CDA, SDA, and QDA are performed to analyze how the incident severity groups differ and how accurately can observations be classified into incident severity groups. CDA reveals that the claim amount variables are the key to separating the major damage and trivial damage group. Besides, capital loss and witnesses are vital for the discrimination between major damage and total loss group. SDA reveals that the most important variables in separating the incident severity groups are vehicle claim, the number of vehicles involved, and incident hour of the day. Lastly, QDA reveals that the classification is not accurate if new observations come in since the misclassification rate is as large as 0.6.

5 Conclusion

The study aims to get a better understanding of the vehicle insurance claim-related variables based on a comprehensive set of eight different multivariate analysis methods. Through techniques such as PCA, FA, Clustering, MCA, CCA, CDA, SDA, and QDA, I investigate the associations among variables, examine customer segments, identify the key drivers to separate incident severity groups, and assessed incident severity groups classification accuracy. The findings highlight the importance of understanding the complex interplay between various variables in the insurance domain. Further research can build upon these findings to enhance customer segmentation and claims risk assessment in the field of vehicle insurance claims.

References

Abdelhadi, S., Elbahnasy, K., & Abdelsalam, M. (2005). A PROPOSED MODEL TO PREDICT AUTO INSURANCE CLAIMS USING MACHINE LEARNING TECHNIQUES. . . *Vol.*, *22*.

Besse, P., & Ramsay, J. O. (1986). Principal components analysis of sampled functions. *Psychometrika*, *51*(2), 285–311. https://doi.org/10.1007/BF02293986