Identifying the Most Influential Factors Affected to the Motor Insurance Claim Using Some Dimensionality Reduction Techniques

For the Bachelor of Science Honours Degree

in

Financial Mathematics and Industrial Statistics

By

K.L.G.P. Kasundi

SC/2020/11793

Supervisor:

Dr. (Mrs.) E.J.K.P Nandani

Department of Mathematics

University of Ruhuna

Matara.

2024

Acknowledgement

I would like to express my deepest gratitude to my supervisor Dr. (Mrs.) E.J.K.P Nandani to her unending guidance and support for completing this research, and also to the supervisory board for giving the necessary advices to me.

Moreover, I would like to thank for the instructors for sharing their valuable insight and experiences and supporting me in this study.

Furthermore, I would like to appreciate my batch mates for sharing their knowledge with me during this study.

Finally, this report stands as a testament of the collective effort and collaboration of all those who mentioned above. Without their valuable contribution this report would have not been possible.

Contents

List of Tables

List of Figures

Abstract

Insurance is crucial in today's modern world, particularly for automobiles, which are highly risky liabilities. Insurance companies introduce motor insurance to cover damage in accidental or disaster situations, claiming to recover the risk on their policy limits. It is in the interest of every motor insurance company to identify the major impacts of the motor insurance claim process. The objective of this research is to examine and analyze the factors affecting motor insurance claims using Principal Component Analysis (PCA) and Factor Analysis for Mixed Data (FAMD) as the robust analytical tools. For that, factors such as vehicle characteristics were considered. Results from the PCA and FAMD reveal the components that contribute most to the variance in claims of motor insurance. According to the obtained results for the model using only numerical factors with PCA, PC2 is significant. The model using both numerical and categorical variables which are encoded by giving weights, resulted that the vehicles assigned for personal use and having higher values typically have lower claim amounts, whereas vehicles assigned for commercial use and having lower values have higher claim amounts. Also, according to the regression model based on FAMD, certain brands of vehicles may be linked to greater claim amounts also, more expensive vehicles tend to have higher claim amounts.

***Key words**: Claim, Motor Insurance Industry, Principal Component Analysis, Vehicle Characteristics, Factor Analysis for Mixed Data*

CHAPTER ONE
INTRODUCTION

1.1. Background of the study

The motor insurance industry, a cornerstone of the financial services sector, operates in a dynamic and evolving environment. The process of determining the costs of vehicle insurance claims is intricate and impacted by several interrelated variables. In the past, insurers have computed rates using large datasets that included information on vehicle features, driver demographics, and past claims history. However, it can be difficult to precisely quantify and price risks due to the complex interactions between these factors.

The way insurers handle risk assessment has changed dramatically in recent years due to the introduction of sophisticated analytical methodologies. Principal Component Analysis (PCA) is one of these approaches that is particularly effective in revealing the underlying structure of intricate datasets. Also, factor analysis for mixed data (FAMD) is a statistical method for analyzing datasets with a mix of several variable kinds, such as continuous and categorical variables. Motor insurance claims are impacted by a wide range of factors, from personal driving habits to external socioeconomic conditions. Using PCA and FAMD allows for the identification of the key elements that determine claim costs.

Even though it's critical to comprehend the variables affecting the cost of auto insurance claims, a thorough and methodical investigation with cutting-edge methods like PCA and FAMD is still a mostly uncharted area. Previous research frequently finds it difficult to separate the complex interactions that exist within the data, which leaves a gap in our understanding of the most important aspects. This lack of knowledge is a problem not just for insurance firms looking to improve their pricing tactics but also for legislators hoping to promote equity and transparency in the motor insurance industry. Against this backdrop, our research uses PCA and FAMD to analyze the intricate dynamics of motor insurance claim pricing to close this gap. Our goal is to uncover the most important variables to offer insightful information that will help improve pricing models, which will eventually help policyholders and insurance companies alike. Examining these variables via the prism of PCA not only advances current risk assessment

techniques but also advances the industry's overarching objective of using advanced analytics to improve decision-making in the dynamic field of motor insurance.

1.2. Literature Review

There are three types of motor insurance coverage, third-party plus fire and theft, own damage cover for vehicle damage or theft and mandatory third-party cover for liabilities. A car should have both own damage and third-party insurance to guarantee full coverage. Multiple administrative layers are involved in the management of motor insurance claims, including review, investigation, adjustment, and payment or denial by the insurer. A motor insurance claim is a request for payment from the insurer in accordance with the terms of the policy. (Vaughan & Vaughan, 1995). The complexity of claims varies. Some settle quickly while others take years to resolve. According to (Kengere, Kituyi, & Ntwali, 2020), a claim is an assertion that the insurer must carry out its end of the bargain. When finding the factors affecting motor insurance claim processing time, the analysis was done by using both quantitative and qualitative methods and the quantitative data analysis includes descriptive statistics which includes frequency distribution, mean, standard deviation and regression. (Lemma, 2019)

Principal Component Analysis (PCA) is a popular statistical technique that is well-known for its ability to find patterns in a variety of fields and reduce the dimensionality of data. PCA helps to clarify variable relationships by removing important patterns from complicated datasets. (Mishra, et al., 2017) This research contributes to a larger body of work that acknowledges PCA as an effective method for deciphering latent structures and streamlining high-dimensional data. PCA is widely used by researchers and practitioners because of its capacity to improve data analysis and interpretation by removing unnecessary information.

1.3. Problem Statement

Determining appropriate claim costs is essential for the sustainability of insurers and the equity of policyholders in the complicated and changing automobile insurance market. Even with the application of advanced statistical models, it is still difficult to understand the complex interactions between a wide range of factors affecting the cost of motor insurance claims.

Previous research frequently fails to provide a thorough knowledge of the most important elements influencing the cost of motor insurance claims because it uses traditional techniques that are unable to effectively sort through the complex web of variables.

Therefore, it is imperative to close this knowledge gap by using advanced analytical methods to reveal hidden patterns and correlations in complicated datasets, such as Principal Component Analysis (PCA), factor analysis for mixed data (FAMD). Inaccuracies in pricing models have the potential to undermine the core values of the insurance sector by putting insurers under financial duress and creating possible injustices for policyholders. It is critical to address this issue if automobile insurance policies are to continue to advance toward a more accurate, egalitarian, and data-driven framework.

By giving a clearer knowledge of the factors influencing claim pricing and insights that might improve industry procedures, the use of PCA and FAMD offers a methodical approach to solving this problem.

1.4. Objectives

This study's main goal is to use Principal Component Analysis (PCA) and factor analysis for mixed data (FAMD) to methodically determine and rank the most important variables influencing the cost of auto insurance claims. Through dimensionality reduction of an extensive dataset, our goal is to extract the key elements that substantially influence claim price variability. Through this research, we hope to improve the accuracy of pricing models for motor insurance, offer insurers useful information, support industry best practices, influence policy decisions and expand the use of PCA and FAMD in the field of insurance analytics.

- Assembling a complete data set

  Compile and arrange an extensive dataset encompassing many aspects of motor insurance, such as driver demographics, vehicle attributes, driving patterns, and past claims information.

- Identify key influencing factors

  Systematically examining PCA and FAMD data to determine and rank the most important variables that affect the cost of auto insurance claims. This includes being aware of the relative importance and weight of each major claim price fluctuation component.

- Calculate the effect on price models

  Evaluate how the listed contributing variables affect the current models of motor insurance pricing. Examine how these variables might be included in pricing models to increase the precision and accuracy of claim price forecasts.

- Provide actionable recommendations for insurance companies

Converting study results into suggestions that insurance firms can use. delivering practical insights that help insurers adjust and enhance their pricing plans in light of a more comprehensive comprehension of the key determinants.

CHAPTER TWO
METHODOLOGY

2.1. Research Approach

The research approach of this study would be involving a combination of quantitative and multivariate statistical methods, pointing in on the efficient examination of mathematical information to uncover powerful factors influencing motor insurance claim prices. These quantitative and mixed strategies are appropriate for distinguishing examples, circumstances and logical results connections inside datasets. So, the principal component analysis (PCA) and factor analysis for mixed data (FAMD) act as a crucial role in identifying the relationships between variables.

By joining quantitative analysis for the numerical data with the interpretability of PCA and multivariate statistical methods for the both numerical and categorical data with the interpretability of FAMD, this exploration means to contribute significant bits of knowledge to the comprehension of engine protection guarantee costs, helping both scholarly world and industry partners. The picked approach lines up with laid out philosophies in these explorations and expands on the perceived utility of both PCA and FAMD in separating significant examples from complex datasets.

2.2. Data Collection

The data collection of this study wishes to remain anonymous in order to protect the proprietary and concealed nature of the source. The information gathered includes specifics about automobile insurance claims and related factors that were obtained straight from the insurance provider. Certain information about the data's source cannot be made public due to legal restrictions and privacy issues. A total of 135 data points were obtained for analysis, providing a substantial dataset for investigating the influential factors affecting motor insurance claim prices. The gathered data includes both categorical and numerical variables relevant to the study's goals.

The obtained data for this study as follows.

Variable name	Meaning
Vehicle price	Value of the vehicle involved in the accident
Brand	Company name of the vehicle
Vehicle age	Period from the year of manufacture of the vehicle to the date of accident
Vehicle use	Purpose of using the vehicle
License revoked	Validity invalidity of driving license after the accident happened
Claim price	Amount of claim

Table 1: Variable Description

Variable name	Variable Type	Missing Data Indicators
Vehicle price	Numerical	N/A
Brand	Categorical	N/A
Vehicle age	Numerical	N/A
Vehicle use	Categorical	N/A
License revoked	Categorical	N/A
Claim price	Numerical	N/A

Table 2: Data Dictionary Table

The categories of these categorical variables are as follows.

Variable	Categories
Brand	Honda, Toyota, Suzuki, Hyundai, Micro, Nissan, Renault, Maruti, Tata, DFSK, Yamaha, Lanka Ashok Leyland, Mercedes, Benz, Zotye, Mitsubishi Fuso, Mistubishi, Bajaj, Mazda, Isuzu, Mahidra, KIA, Daihatsu
Vehicle use	Commercial, Private
License revoked	Yes, No

Table 3: Categories of the Categorical Variables

2.3. Methods of Data Analysis

In this study, the two main techniques for data analysis have been used to thoroughly investigate the key variables influencing the cost of auto insurance claims. The first method is the application of Principal Component Analysis (PCA). At first, PCA is only used to numerical data related to the motor insurance claim prices. Then, both numerical and categorical data are

included to the analysis, with numerical values being used to encode the categorical variables. The second method involved in this study is Factor Analysis for Mixed Data (FAMD). Particularly, datasets with both numerical and categorical variables can be subjected to FAMD.

2.3.1. Principal Component Analysis

The one of the methods for data analysis in this study is Principal Component Analysis (PCA), a well-established technique widely recognized for its efficacy in reducing reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. (Mishra, et al., 2017)

Karl Pearson created Principal Component Analysis (PCA) in 1901, which served as the foundation for multivariate analysis. In 1933, Harold Hotelling independently established PCA. In the middle of the 20th century, the technique gained popularity and earned widespread recognition for dimensionality reduction. PCA became computationally feasible in the 1970s with the introduction of computers. Sirovich and Kirby used "eigenfaces" to show off the versatility of PCA in facial identification in 1987. These days, PCA is an essential method for feature extraction and dimensionality reduction in data science, machine learning, and other fields. Its progression over time is indicative of its continuing importance in data analysis.

Goals of PCA are,

- Reducing the size of the data set by retaining only this significant information are the objectives of PCA
- Make the data set's description simpler
- Examine how the variables and observations are organized.
- Summarize the data without much loss of information by limiting the number of dimensions.
- This method used in image compression

The step-by-step process of PCA involves the following key stages.

1) Standardization of variables

In this step, transform data to have the mean of 0 and standard deviation of 1.

$$Scaled\ value = \frac{Value - Mean}{Standard\ Deviation}$$

This method less affected by the presence of outliers and will help to bring values to a comparable range with other variables.

2) Calculation of covariance matrix

Calculating the covariance matrix between the variables in the dataset is necessary for Principal Component Analysis (PCA). The degree of covariance or correlation between every pair of variables is represented by the covariance matrix, which is a square matrix.

The covariance matrix for a dataset containing 'p' variables is a 'p x p' matrix where every element indicates the covariance between two related variables.

Assume that, there is a dataset with three variables x, y and z. Then, the covariance matrix has 3 rows and 3 columns. Related to this, the covariance matrix is,

$$\begin{matrix} CoV(x,x) & CoV(x,y) & CoV(x,z) \\ CoV(x,y) & CoV(y,y) & CoV(y,z) \\ CoV(x,z) & CoV(y,z) & CoV(z,z) \end{matrix}$$

for, $CoV(x,y) = \frac{1}{n-1} \sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y})$ where,

n = number of observations

x_k = k^{th} element of x variable

$\bar{x}$ = mean of variable x

$\bar{y}$ = mean of variable y

3) Determination of eigenvalues and eigenvectors

Eigen Values:

The amount of variance that each principal component accounts for is represented by an eigenvalue. They show how much variety is captured by each component. Principal components with higher eigenvalues hold onto more details about the original variables. Eigenvalues in PCA are frequently arranged in descending order to indicate how significant each principal component is in explaining the dataset's overall variability.

If the covariance matrix is A, then the characteristic equation of the covariance matrix is,

$\det(A - \lambda I)$.

Eigen Vectors:

In the original variable space, eigenvectors correspond to the direction of maximum variance. The weights or coefficients given to the initial variables in the process of creating the principal components are represented by each eigenvector, which corresponds to a particular eigenvalue. The unit length of the normalized eigenvectors indicates the degree and direction of each variable's influence on the principal components.

4) Selection of principal components

In principle Component Analysis (PCA), criteria are set usually based on the size of the eigenvalues assigned to each component, in order to pick the principal components. Selecting which principal components to keep in order to capture the dataset's most important variance depends on this phase.

The amount of variance explained by each primary component is shown by the eigenvalues. Retaining primary components with larger eigenvalues is typically the standard practice because of their greater overall variance contribution. Greater eigenvalues suggest that more information about the original variables is retained in the corresponding principal components.

Typically, the eigenvalues are ranked in descending order and the top 'k' principal components such as 95% or 99%, that cumulatively capture a significant fraction of the overall variance are chosen. This guarantees that the majority of the variability in the dataset is represented by the components that were kept.

2.3.2. Factor Analysis for Mixed Data

Factor Analysis for Mixed Data (FAMD) is a principal component method that committed to explore data with both continuous and categorical variables. The FAMD algorithm can be thought of as a hybrid of multiple correspondence analysis (MCA) and principal component analysis (PCA). In differently, it functions as MCA for qualitative factors and as PCA for quantitative data. It can be seen as a combination between PCA and MCA. (Husson, 2023)

More exactly, the continuous variables are scaled to unit variance and the categorical variables are transformed into a decomposition data table (crisp coding) and then scaled using the specific scaling of MCA. This guarantees that the effects of continuous and categorical variables are balanced in the analysis. It means that both variables are on an equal state for determining the

dimensions of variability. This method allows the study of similarities between individuals and the study of relationships between all variables, taking into account confounding variables. It also provides graphical outputs such as the representation of the individuals, the correlation circle for the continuous variables and representations of the categories of the categorical variables, and also specific graphs to visualize the relationship between both types of variables.

Factor analysis is an appropriate statistical technique used to find common patterns among observed variables in datasets to reveal underlying structures. Traditional factor analysis methods might not be appropriate when working with mixed data, which contains both continuous and categorical variables. Factor Analysis for Mixed Data (FAMD) is a useful method in these situations.

Key Concepts in FAMD

1) Types of variables

    - Continuous Variables: Variables with Numerical Values
    - Categorical Variables: Variables with Discrete categories or Groups

2) Characteristics

    - In traditional factor analysis assumes that all variables are continuous
    - Combining continuous and categorical variables requires a specialized technique

FAMD Process

Traditional factor analysis is expanded by FAMD to handle mixed data sources. Factor analysis is used for Continuous variables and Multiple Correspondence Analysis (MCA) is used for Categorical data.

- Analysis of Factors in Continuous Variables:

  With FAMD, latent factors that liable for variance in continuous variables can be found, just like in traditional factor analysis. Important output parameters are communalities, factor loadings, and eigenvalues.

- Multiple Correspondence Analysis (MCA) for categorical variables:

Categorical variables are subjected to MCA to identify relationships between categories. It transforms the categorical data into numerical values, so factor analysis can be done on them.

- Combining the results:

  Combined factor scores from the above two studies present a complete picture of the underlying structure of mixed data.

Interpretations of above Process

- Factor Loadings:

  Factor loadings for continuous variables show the strength and direction of the relationship between the variable and the latent factor.

  Describe the locations of categorical points on the MCA plot for categorical variables.

- Eigenvalues:

  The eigenvalue indicates the variance responsible for each factor. Greater impact is indicated by higher eigenvalues.

Practical Applications

- Model fit assessment:

  Evaluate model fit with mixed data-fit indices such as chi-square, CFI, and RMSEA.

- Software implementation:

  R (with the 'FactoMineR' package) or Python (with the 'Prince' library) are two examples of specific statistical software packages used to implement FAMD.

CHAPTER THREE

RESULTS

In this section, the results obtained from the principal component analysis and factor analysis will be discussed.

3.1. Principal Component Analysis

3.1.1. PCA only with numerical factors

PCA offers a way to minimize dimensionality in the setting of numerical components while maintaining important information. This method works especially well for deciphering hidden patterns in high-dimensional data, finding important numerical factors, and simplifying complicated datasets. We examine the use of PCA to numerical components in this study, emphasizing its value in obtaining significant insights and enabling a more effective examination of numerical variables. In this section, numerical factors affecting the claim are presented using PCA.

The numerical variables in this research are value, vehicle age and Claim were transformed into a PCA. According to that, the two values 1.0158 and 0.9840 represents the standard deviations of the two principal components named PC1 and PC2, respectively. The standard deviation quantifies the degree of variation present in the data along a certain component. Greater variability in the data is indicated by a greater standard deviation. So, the eigenvalues obtained as a preliminary step in this process are 1.0158 and 0.9839 and the covariance matrix is,

$$\begin{pmatrix} & \text{Value} & \text{Vehicle age} \\ \text{Value} & 4.985\text{e}+12 & -453971.807 \\ \text{Vehicle age} & -4.539\text{e}+05 & 40.890 \end{pmatrix}$$

Additionally, the following table shows the obtained principal components values.

	PC1	**PC2**
Value	0.7071068	-0.7071068
Vehicle Age	-0.7071068	-.0.7071068

Table 4: Principal Components

According to that,

PC1:

- The “Value” variable has a positive loading of 0.7071068 on PC1. It positively contributes PC1’s creation.
- The “Vehicle Age” variable has a negative loading of -0.7071068 on PC1 which means it negatively contributes PC1’s.
- This implies that “Value” has a positive effect on PC1 whereas “Vehicle Age” has a negative effect on this combination of factors.

PC2:

- The “values” variable has a negative loading of -0.7071068 on PC2.
- The “Vehicle Age” variable also has a negative loading of -0.7071068 on PC2.
- This implies both variables contribute negatively to the formation of PC2.

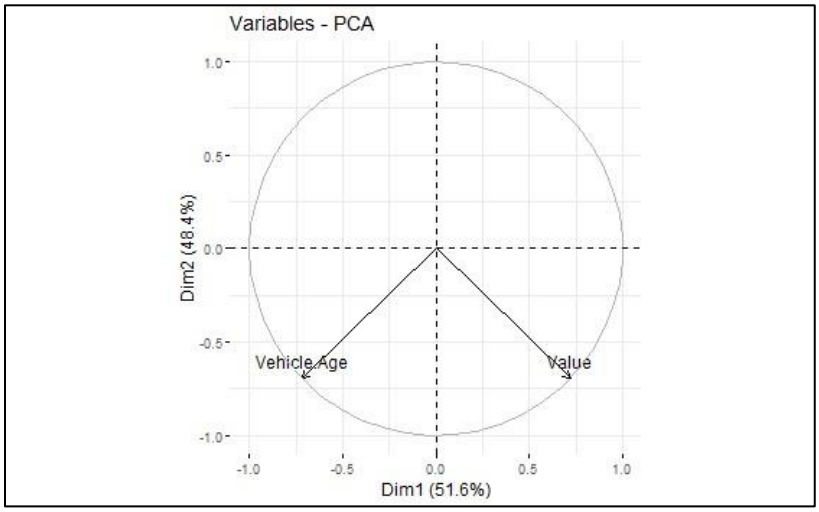


Figure 1: Biplot

The above biplot shows that a dramatic fall occurs when you add additional principal components (PCs), with the first PC1 capturing the majority of the variation in the data. This suggests that PC1 captures the essential information in the data most succinctly. Subsequent PCs contribute increasingly less, suggesting they may not be as essential for understanding the data's structure.

The importance of above these principal components is as follows.

	PC1	PC2
Proportion of variance	0.5159	0.4841
Cumulative proportion	0.5159	1.0000

Table 5: Importance of Principal Components

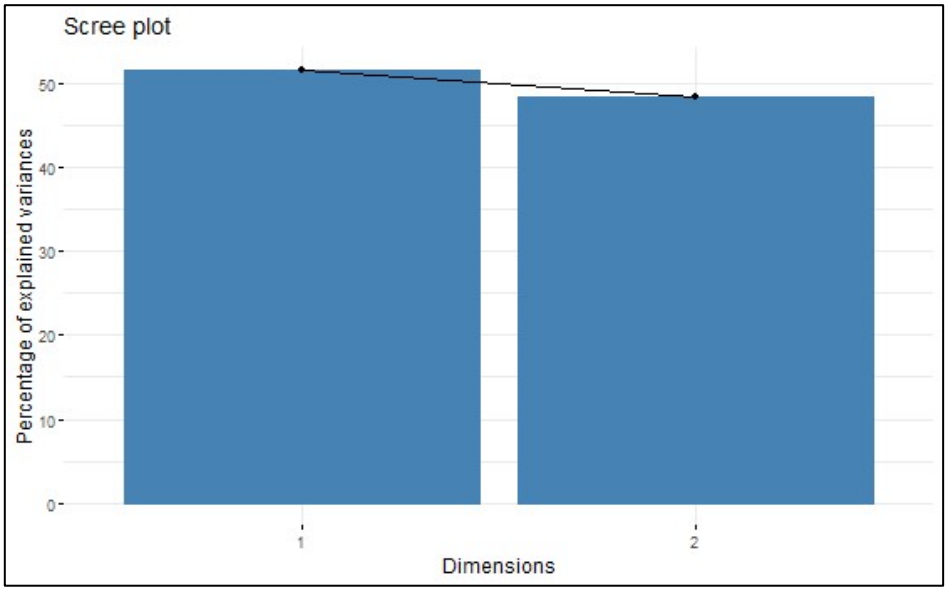


Figure 2: Scree Plot

The percentage of variation explained by each PC is shown graphically in the above figure 2. The scree plot's line indicates a sharp decrease in the proportion of variance explained following the first two PCs. This implies that most of the relevant information in the data is captured by the first two PCs and that adding more PCs wouldn't provide much additional benefit. Principal component 1 and 2 are dominant in understanding data structure since they account for all of the variation in the data. PC1 reigns supreme, holding a large share of explained variance. This is further supported by the scree plot, which shows that these two elements successfully condense the important information found in the data.

Additionally, the following table shows the obtained regression output.

	Coefficient	Std. Error	P value
Intercept	45890	5900	1.87E-12
PC1	4974.0	5830.0	0.3951
PC2	-12991.0	6018.0	0.0327
Residual Standard Error		68960 on 129 df	
Multiple R Squared		0.04974	
P Value		0.2474	

Table 6: Regression Model

Despite having a relatively low overall explanatory power, the regression model with PC1 and PC2 as predictors shows a statistically significant connection with the response variable. PC1's effect seems to be inconclusive, whereas PC2's looks to be very detrimental.

3.1.2. PCA with both numerical and categorical data

As the dataset contained both numerical and categorical data and PCA primarily relies on numerical data, the categorical variables (vehicle number, vehicle use and license revoked) were transformed into a PCA compatible format using encoding. With this method, new binary features are generated for every unique category that exists within a categorical variable.

According to that, the eigenvalues obtained as a preliminary step in this process are 1.3389, 1.1086, 0.9691, 0.9149 and 0.4494 and the covariance matrix is,

$$\begin{pmatrix} & Vehicle.Number & Value & Vehicle.Age & Vehicle.Use & License.revoked \\ Vehicle.Number & 3.563e-01 & -8.516e+04 & -3.018e+00 & -1.492e-03 & -3.427e-03 \\ Value & -8.516e+04 & 4.985e+12 & -4.539e+05 & -1.849e+05 & 4.857e+04 \\ Vehicle.Age & -3.018e+00 & -4.539e+05 & 4.089e+01 & -7.462e-02 & 1.326e-02 \\ Vehicle.Use & -1.492e-03 & -1.849e+05 & -7.462e-02 & 2.417e-01 & -1.194e-02 \\ License.revoked & -3.427e-03 & 4.857e+0 & 1.326e-02 & -1.194e-02 & 7.540e-02 \end{pmatrix}$$

And also, the following table shows the obtained principal components values.

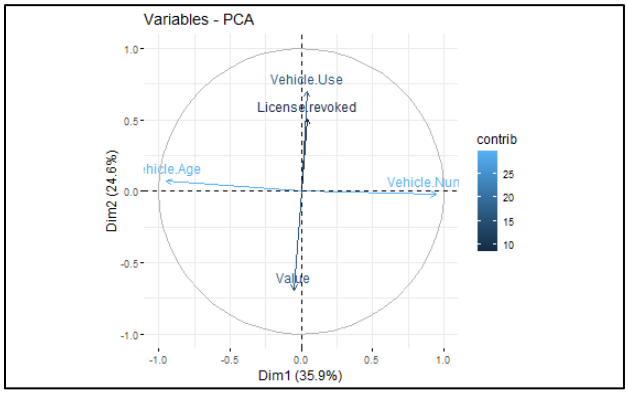
	PC1	**PC2**	**PC3**	**PC4**	**PC5**
Vehicle number	0.70697	-0.01645	0.015301	-0.06387	0.703994
Vehicle value	-0.03795	-0.62533	-0.366728	0.68141	0.093289
Vehicle age	-0.70491	0.06330	0.005550	-0.07435	0.702511
Vehicle use	0.02824	0.62990	0.278264	0.72311	0.045916
License revoked	0.03231	0.45596	- 0.887589	-0.05640	-0.007619

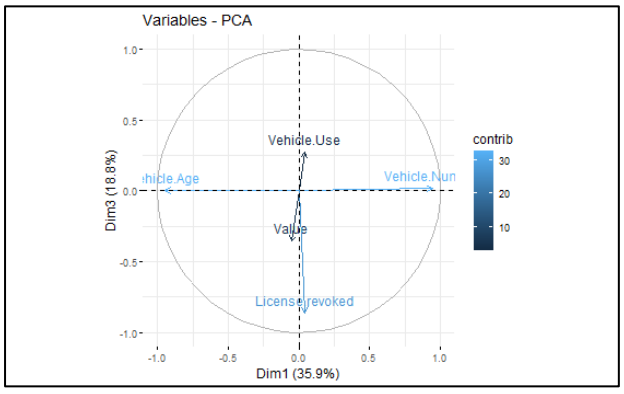
Table 7: Principal Components

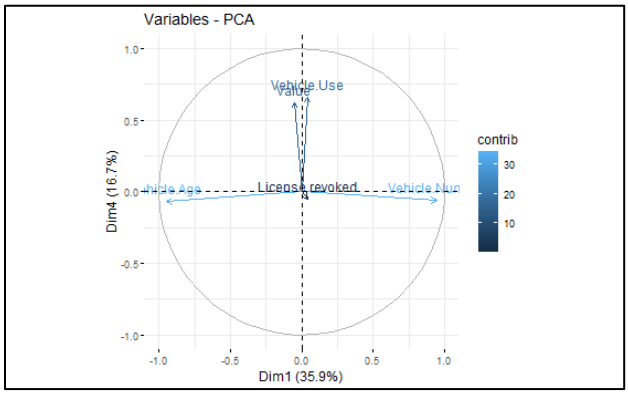
Since PC1 is primarily driven by the positive influence of the vehicle number and the negative influence of the vehicle age, it appears that vehicles with the most recent numbers tend to be newer. This is how the combined influence of the vehicle number and age is reflected in the first principal component. PC2 illustrates the difference between the value and purpose of vehicles, showing that vehicles with higher values are used for private purposes and those with lower values are used for commercial purposes. Understanding the dynamics of value and usage patterns, this suggests a possible trade-off between affordability and usage intensity.

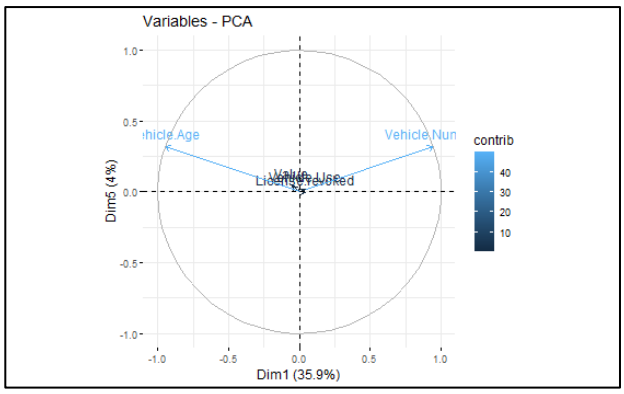
License cancellation after the accident happened has a significant impact on the PC3, indicating that it represents a particular class of vehicles with prior license violations. The feature may draw attention to possible risk factors connected to vehicles whose licenses have been revoked. The PC4 records the relationship between vehicle value and use, just like the PC2 does. On the other hand, it presents an alternative subtlety that might highlight particular vehicle usage patterns across a range of values. Similar to PC1, PC5 represents the relationship between the age and vehicle number. But in contrast to PC1, it might represent a weaker or secondary relationship. This part adds to our understanding of the relationship between vehicle age and number by offering more information about how the two factors interact.

The following biplots show these relationships between these factors and principal components better.









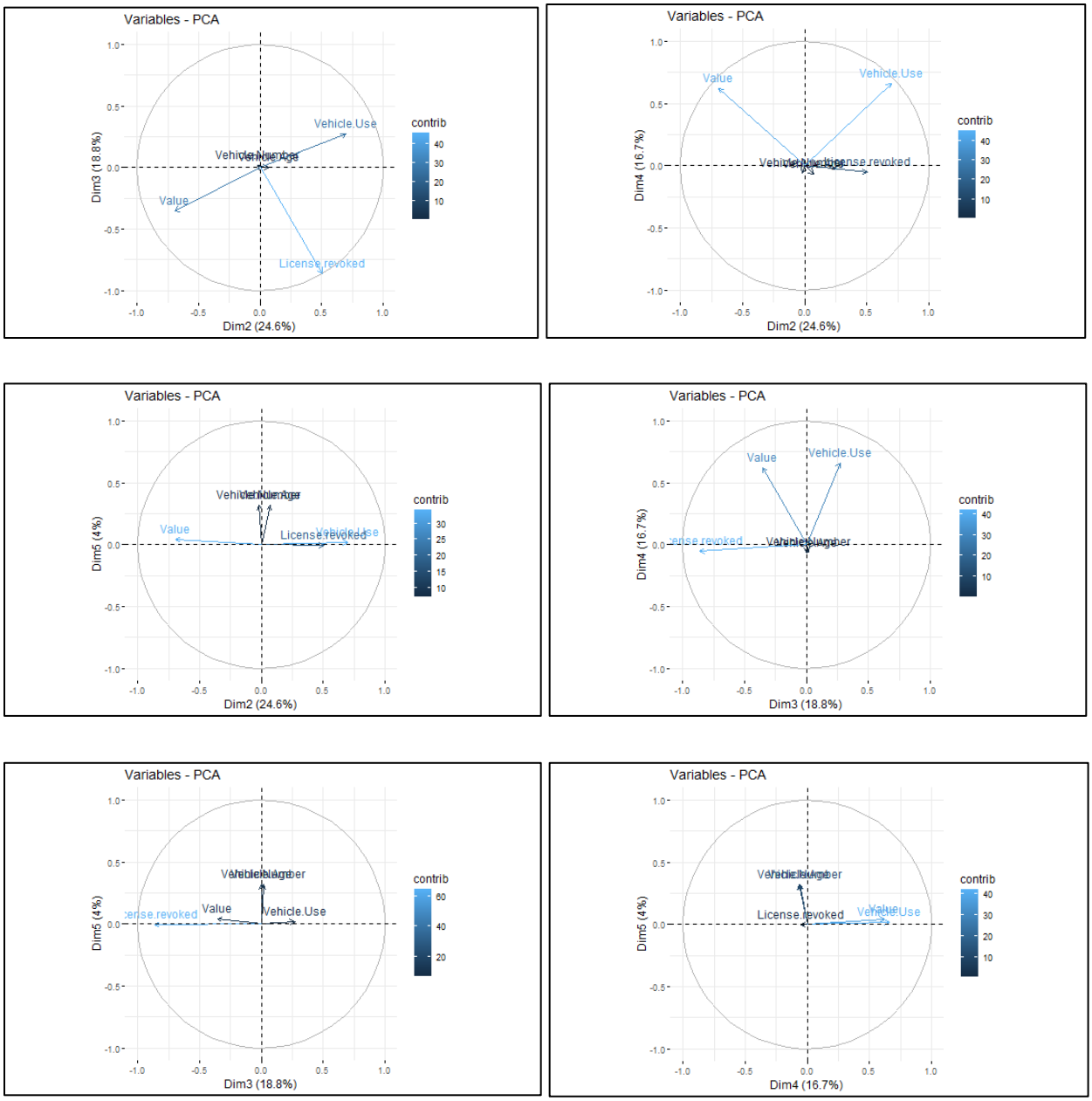


Figure 3: Biplots

The importance of above these principal components is as follows.

	PC1	PC2	PC3	PC4	PC5
Proportion of Variance	0.3585	0.2458	0.1878	0.1674	0.04039
Cumulative Proportion	0.3585	0.6043	0.7922	0.9596	1.00000

Table 8: Importance of Principal Components

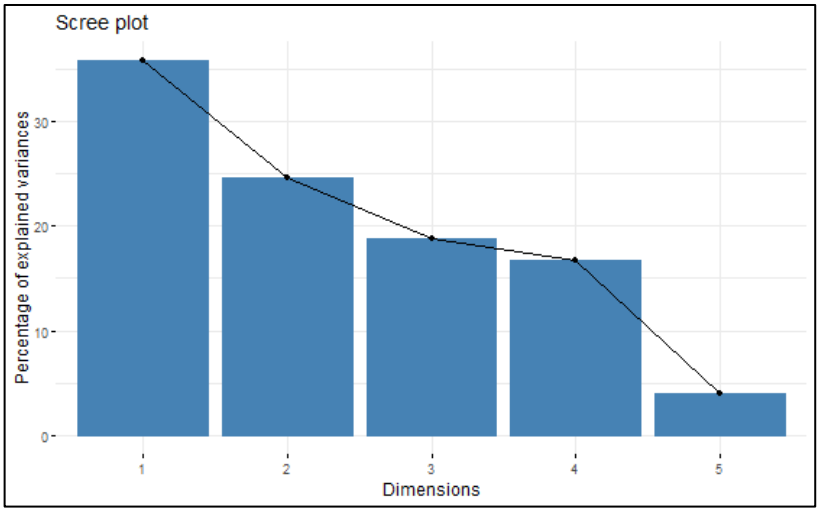


Figure 4: Scree Plot

According to the above table 7 and figure 4, the leading component, PC1, captures a dominant 35.85% of the variance, staking its claim to be the most accurate interpreter of the underlying structure of the data right away. PC2 closely follows its lead with a significant contribution of 24.58%, adding to the understanding with its distinct viewpoint. When combined, these two elements reveal more than 60% of the data, demonstrating their critical importance in the data exploration process.

Furthermore, PC3 shows up, providing an extra 18.78% of the variance and shedding more light on the complex data tangle. PC4 joins with16.74%, not as prominent as the other PCs, but still contributing to a complete picture of the data's features. Lastly, even with a limited 4.04% contribution, PC5 makes sure that every aspect of the data is represented in the overall story. So, the effect of these obtained principal components on the motor insurance claim is described by the following model.

	Coefficient	Std. Error	P value
Intercept	45889.5	5935.5	2.65E-12
Dim. 1	-4297.0	4449.7	0.3360
Dim. 2	-11690.1	5373.9	0.0314
Dim. 3	317.3	6147.3	0.9589
Dim. 4	5627.0	6511.1	0.3891
Dim. 5	7707.0	13257.7	0.5620
Residual Standard Error		68960 on 129 df	
Multiple R Squared		0.04974	
P Value		0.2474	

Table 9: Regression Model

The above tables illustrate that the multiple R-squared value of 0.04974 and it indicates the model's overall fit is weak. This means that the first five principal components account for only 4.97% of the variation in motor insurance claims. Here, p value also greater than 0.05 which implies the lack of a statistically significant relationship between the PCs and the claim amount.

And also, the non-significant p values of PC1, PC3, PC4 and PC5 suggest that there are no meaningful relationships for the claim amount with these components. However, PC2 is statistically significant with the lower p value of 0.0314 while having a stronger possible inverse relationship with the claim amount. As the PC2 shows that vehicles with higher values are used for commercial purposes and those with lower values are used for private purposes, this clearly emphasize that the vehicles with higher values and used for commercial purposes (represented by higher PC2 scores) tend to have lower claim amounts. Conversely, vehicles with lower values and used for private purposes (represented by lower PC2 scores) tend to have higher claim amounts.

However, some changes in the behavior of each variable in the principal components are caused by changing the encoded order, but it does not affect the final result. For an example, while encoding the license revoked variable, the numerical values of 'no' and 'yes' categories given as 1 and 2 respectively and when those values are changed into 1 and 0, the behavior of this in each component changed conversely as shown in follows but the final result obtained as same as before.

	PC1	**PC2**	**PC3**	**PC4**	**PC5**
Vehicle number	0.70697	-0.01645	0.015301	-0.06387	0.703994
Vehicle value	-0.03795	-0.62533	-0.366728	0.68141	0.093289
Vehicle age	-0.70491	0.06330	0.005550	-0.07435	0.702511
Vehicle use	0.02824	0.62990	0.278264	0.72311	0.045916
License revoked	**-0.03231**	**-0.45596**	**0.887589**	**0.05640**	**0.007619**

Table 10: principal Components

3.2. Factor Analysis for Mixed Data

A principal component method identified as factor analysis of mixed data (FAMD) is used to examine a set of data that includes both quantitative and qualitative variables. In this case, the categorical variables (vehicle number, vehicle use, brand and license revoked) as well as the numerical variables (vehicle age and value) were used to find the most effecting factors to the motor insurance claim price using FAMD.

Here, the obtained eigen values are 2.7396, 2.5700, 2.3303, 2.2366 and 2.0.

And also, the following table shows the obtained factor analysis components values.

	Dim. 1	**Dim. 2**	**Dim. 3**	**Dim. 4**	**Dim. 5**	**Dim. 6**
Vehicle number	0.997	9.741e-01	0.984	0.978	1.000e+00	1.000e+00
Vehicle value	0.002	6.589e-01	0.063	0.005	6.628e-31	0.000e+00
Vehicle age	0.863	6.380e-05	0.003	0.0002	1.388e-30	1.883e-32
Vehicle use	0.001	1.665e-01	0.227	0.194	4.399e-31	1.080e-30
License revoked	0.003	5.192e-02	0.294	0.237	2.394e-30	5.140e-32
Brand	0.871	7.183e-01	0.756	0.819	1.000e+00	1.000e+00

Table 11: Components

A general "vehicle profile" associated with size, category and market segment may be captured by factor 1, which has a strong correlation with vehicle number, brand, and age. The weak relationship between vehicle value and other factors can probably be accounted for. While the second component's precise meaning is still unknown in the absence of additional context, it is somewhat correlated with the brand, vehicle value and number of the vehicle. Revocation of vehicle use has little bearing on this.

Although the dimension 3 and 4 show somewhat different aspects of usage patterns or risk factors, they both show moderate associations with vehicle number, use, and revoked status. In these dimensions, age and brand have less of an impact. The vehicle number and brand for dimension 5 and both for dimension 6 are the only individual variables that are related to the highly specific dimensions 5 and 6. They could be distinct patterns that defy easy explanations from other sources.

In this FAMD process, the importance of above these principal components are as follows.

	Dim. 1	Dim. 2	Dim. 3	Dim. 4	Dim. 5	Dim. 6
% of variance	2.2273	2.0894	1.8945	1.8183	1.6260	1.6260
Cumulative % of variance	2.2273	4.3167	6.2113	8.0297	9.6558	11.2818

Table 12: Importance of Variance

With a combined explanation of more than 4% of the variance, the first two dimensions are clearly the most important. While dimension 2 offers a substantial amount of extra information, dimension 1 probably captures the essential underlying factor. Contributions from dimensions 3 and 4 are significant (about 3.7%), indicating less significant but still significant factors. Lastly, dimension 5 and 6 only contribute a small amount (2.3%), which might help to explain certain small variations in the data.

These things illustrate better in the below figure.

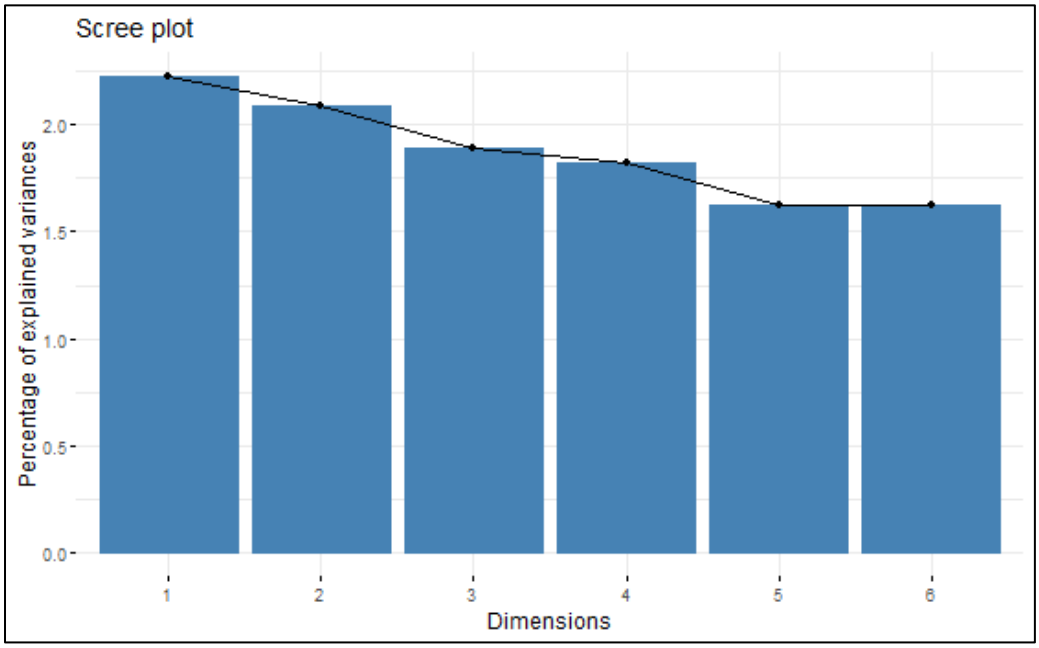


Figure 5: Scree plot

Thus, the following model describes how these obtained components impact the motor insurance claim prices.

	Coefficient	Std. Error	P value
Intercept	45889.5	58339.4	1.38E-12
Dim. 1	-3016.2	3527.9	0.3942
Dim. 2	9960.9	3642.5	0.0071
Dim. 3	3320.5	3825.2	0.3869
Dim. 4	-405.5	3904.5	0.9174
Dim. 5	-4127.5	4129.1	0.3193
Dim. 6	-621.0	4129.1	0.1331
Residual Standard Error			67850 on 129 df
Multiple R Squared			0.08739
P Value			0.06456

Table 13: Regression Model

According to the table 12, the multiple R-squared of 0.087 indicates a relatively weak overall fit, suggesting that the six dimensions extracted by FAMD explain only 8.7% of the variance in motor insurance claim prices. Also, the p value of 0.06456 is slightly higher than the common threshold of 0.05. It suggests that the overall evidence for a relationship between the dimensions and claim prices is not very strong, but it's still worth considering.

According to the p values of each dimension, dimension 1, 3, 4, 5 and 6 are not statistically significant. However, the coefficient of dimension 2 is positive and statistically significant with p value of 0.0071. This indicates a positive relationship with claim prices, meaning higher scores on dimension 2 are associated with higher claim amounts which means certain brands of vehicles may be linked to greater claim amounts because of their price, performance or typical driving habits. Also, more expensive vehicles tend to have higher claim amounts.

CHAPTER FOUR
DISCUSSION

This study focuses on the impact of principal component analysis (PCA) and factor analysis of mixed data (FAMD) on understanding the factors influencing motor insurance claim prices. Based on the methodology, the discussion is divided into three sections as Factor Analysis for Mixed Data (FAMD), PCA with both numerical and categorical data and PCA only with numerical factors.

In the principal component analysis (PCA) performed by using only the numerical factors value, vehicle age on claim price, PC1 and PC2 were found to be dominant, accounting for 60% of the variability in the data. PC1 indicates a trade-off between value and vehicle age by reflecting their combined influence. PC2 shows a possible trade-off between affordability and usage intensity by contrasting value and vehicle use. Significant results were obtained from the regression model utilizing PC1 and PC2, with PC2 negatively affecting claim amounts.

PCA identified dominant components like PC1, highlighting the positive influence of vehicle number and the negative influence of vehicle age by incorporating numerical and categorical variables. PC2 emphasized the difference between a car's worth and use. License revocation had a significant impact on PC3, indicating a particular category with potential risk factors. PC4 captured the relationship between vehicle value and use, similar to PC2 in the process. The relationship between the age and number of the vehicle was further refined by PC5 and PC6.

The regression model based on this case reveals a limited overall fit, with only 4.97% of the variation in motor insurance claims explained by the first five principal components. PC2 is the only one of these that is statistically significant (p-value = 0.0314) and shows a significant correlation with claim amounts. PC2 proposes an inverse relationship in which vehicles assigned for personal use and having higher values typically have lower claim amounts, whereas vehicles assigned for commercial use and having lower values have higher claim amounts. So, it is clear as generally, the vehicles used for commercial purposes have higher values of claim prices. However, generally, the vehicles with higher values have higher claim amounts. As this model not good for fit because it’s significant error, we can’t agree with some of these conclusions.

As well as, by combining quantitative and qualitative factors, the FAMD identified dimensions that contributed to the explanation of 8.7% of the variation in the cost of auto insurance claims. The most important dimensions were the first two, which represented overall car profiles and relationships with number, brand and value. Further information about usage trends or risk factors was given by dimensions 3 and 4.

The regression model based on FAMD, dimensions indicated a weak overall fit, with dimension 2 showing a positive and statistically significant relationship with claim amounts. That is certain brands of vehicles may be linked to greater claim amounts because of their price, performance or typical driving habits. Also, more expensive vehicles tend to have higher claim amounts.

Here, the slight difference in the p value of the regression model of FAMD which indicate the slightly weak overall fit, is happened because the claim prices are not much accurate as they are estimated claim prices.

CHAPTER FIVE
CONCLUSION

Although PCA and FAMD offer significant insights into the intricate dynamics affecting the motor insurance claim prices, the regression models that result from these analyses have a limited overall fit. The only statistically significant component found in the models is PC2, which struggles to make up for a significant portion of the variability in claim amounts, especially when it comes to PCA. The results indicate that using these models alone to predict claim prices should be done with caution as they might not totally take advantage of every relevant factor. The limitations of each approach, such as encoding sensitivity and low explanatory power, should be carefully considered, and further exploration and improvement are necessary. Additionally, the study highlights the need for incorporating additional variables to enhance the predictive capabilities of the models and achieve a more comprehensive understanding of the factors influencing motor insurance claims.

Appendix

Appendix 1: R code for PCA only with numerical factors

```
setwd("E:/Documents/PEHARA/Uni/3rd Year/Semester 1/Case Study II (MFM3151)")
Data_Nu <- read.csv("Numerical Only.csv")
Data_Nu
library(readr)
library(dplyr)
library(stats)
library(ggplot2)
library(FactoMineR)
library(factoextra)
scaled_features_Nu <- scale(Data_Nu)
scaled_features_Nu
cov_claim_nu <- Data_Nu[, 1:2]
claim_covariance_matrix_nu <- cov(cov_claim_nu)
print(claim_covariance_matrix_nu)
# Assuming 'scaled_features_Nu' is your dataset and you want to exclude the 3th column
columns_to_exclude_Nu <- 3
columns_to_include_Nu <- setdiff(seq_len(ncol(scaled_features_Nu)), columns_to_exclude_Nu)
# Perform PCA
PCA_Claim_nu <- prcomp(scaled_features_Nu[, columns_to_include_Nu], center =
TRUE, scale. = TRUE)
PCA_Claim_nu
summary(PCA_Claim_nu)
fviz_pca_var(PCA_Claim_nu,axes = c(1, 2),col.var = "contrib")
fviz_screeplot(PCA_Claim_nu)
pcs_claim_nu <- as.data.frame(PCA_Claim_nu$x)
ols.data_claim_nu <- cbind(Data_Nu$Claim, pcs_claim_nu)
ols.data_claim_nu
Dependent_claim_Nu <- ols.data_claim_nu[, 1]  # Dependent variable
Independent_claim_Nu <- ols.data_claim_nu[, 2:3]  # Independent variables (PC1 and PC2)
```

```
lmodel_claim_nu <- lm(Dependent_claim_Nu ~ ., data = Independent_claim_Nu)
lmodel_claim_nu
summary(lmodel_claim_nu)
```

Appendix 2: R code for PCA with both numerical and categorical data

```
setwd("E:/Documents/PEHARA/Uni/3rd Year/Semester 1/Case Study II (MFM3151)")
Data_all <- read.csv("claim All.csv")
Data_all
library(readr)
library(dplyr)
library(stats)
library(ggplot2)
library(FactoMineR)
library(factoextra)
scaled_features_All <- scale(Data_all)
cov_claim_all <- Data_all[, 2:6]
claim_covariance_matrix_all <- cov(cov_claim_all)
print(claim_covariance_matrix_all)
# Assuming 'scaled_features_All' is your dataset and you want to exclude the 1st column
columns_to_exclude_All <- 1
columns_to_include_All<- setdiff(seq_len(ncol(scaled_features_All)), columns_to_exclude_All)
# Perform PCA
PCA_Claim_All <- prcomp(scaled_features_All[, columns_to_include_All], center =
TRUE, scale. = TRUE)
PCA_Claim_All
summary(PCA_Claim_All)
fviz_pca_var(PCA_Claim_All,axes = c(1, 2), col.var = "contrib")
fviz_screeplot(PCA_Claim_All)
pcs_claim_All <- as.data.frame(PCA_Claim_All$x)
ols.data_claim_All <- cbind(Data_all$Claim, pcs_claim_All)
ols.data_claim_All
```

```
Dependent_claim_All <- ols.data_claim_All[, 1]  # Dependent variable
Independent_claim_All <- ols.data_claim_All[, 2:6]  # Independent variables PC1 to PC5
lmodel_claim_All <- lm(Dependent_claim_All ~ ., data = Independent_claim_All)
lmodel_claim_All
summary(lmodel_claim_All)
```

Appendix 3: R code for FAMD

```
library(FactoMineR)
library(factoextra)
library(readxl)
Claim_N_C <- read_excel("E:/Documents/PEHARA/Uni/3rd Year/Semester 1/Case Study II (MFM3151)/Claim N&C.xlsx")
head(Claim_N_C)
res.famd <- FAMD(Claim_N_C,
                 sup.var = 1,  ## Set the target variable "Churn" as a supplementary variable, so it is not included in the analysis for now
                 graph = FALSE,
                 ncp=6)
res.famd
fviz_screeplot(res.famd)
pcs_claim_famd <- as.data.frame(res.famd$ind$coord)
ols.data_claim_famd <- cbind(Claim_N_C$Claim, pcs_claim_famd)
ols.data_claim_famd
Dependent_claim_famd <- ols.data_claim_famd[, 1]  # Dependent variable
Independent_claim_famd <- ols.data_claim_famd[, 2:7]  # Independent variables
lmodel_claim_famd <- lm(Dependent_claim_famd ~ ., data = Independent_claim_famd)
lmodel_claim_famd
summary(lmodel_claim_famd)
```

References

Chelaru-Centea, N. (2019, August 19). *Intelligence Refinery.* Retrieved from https://nextjournal.com/pc-methods/calculate-pc-mixed-data

Eric. (2023, March 15). *Data Analytics Blog.* Retrieved from Applications of Principal Components Analysis in Finance: https://www.aptech.com/blog/applications-of-principal-components-analysis-in-finance/

Gessese, Y. B. (2018). *The effect of motor insurance claim management on customer satisfaction at ethiopian insurance corporation.* School of Motor Insurance Claim, Department of Business Administration. Addis Ababa, Ethiopia: St. Mary's University.

Husson, F. (2023, October 13). *rdrr.io.* Retrieved from FactoMineR: Multivariate Exploratory Data Analysis and Data Mining: https://rdrr.io/cran/FactoMineR/man/FAMD.html

Kengere, A., Kituyi, A., & Ntwali, A. (2020, September). Claims Management and Financial Performance of Insurance Companies in Rwanda: A Case of SONARWA General Insurance Company Ltd. *Journal of Financial Risk Management, Vol. 9*, No. 3.

Lemma, S. (2019). *Factors Affecting Motor Insurance Claim Processing Time: The Case of Awash Insurance Company S.C.* School of Graduate Studies. Addis Ababa, Ethiopia: St. Mary's University.

Mahmood, M. S. (2021, July 12). *Towards Data Science.* Retrieved from Medium: https://towardsdatascience.com/factor-analysis-of-mixed-data-5ad5ce98663c

Mishra, S. P., Sarkar, U., Taraphder, S., Datta, S., Swain, D., Saikhom, R., . . . Laishram, M. (2017). Multivariate Statistical Data Analysis- Principal Component Analysis (PCA). *International Journal of Livestock Research*, 60-78.

Vaughan, E. J., & Vaughan, T. (1995). *Essential of Risk Management ans Insurance* (11th ed.). USA.