Disasters Data Analysis

Group4:
35780959, 36497215
35621168, 36390968

Contents

01

Background

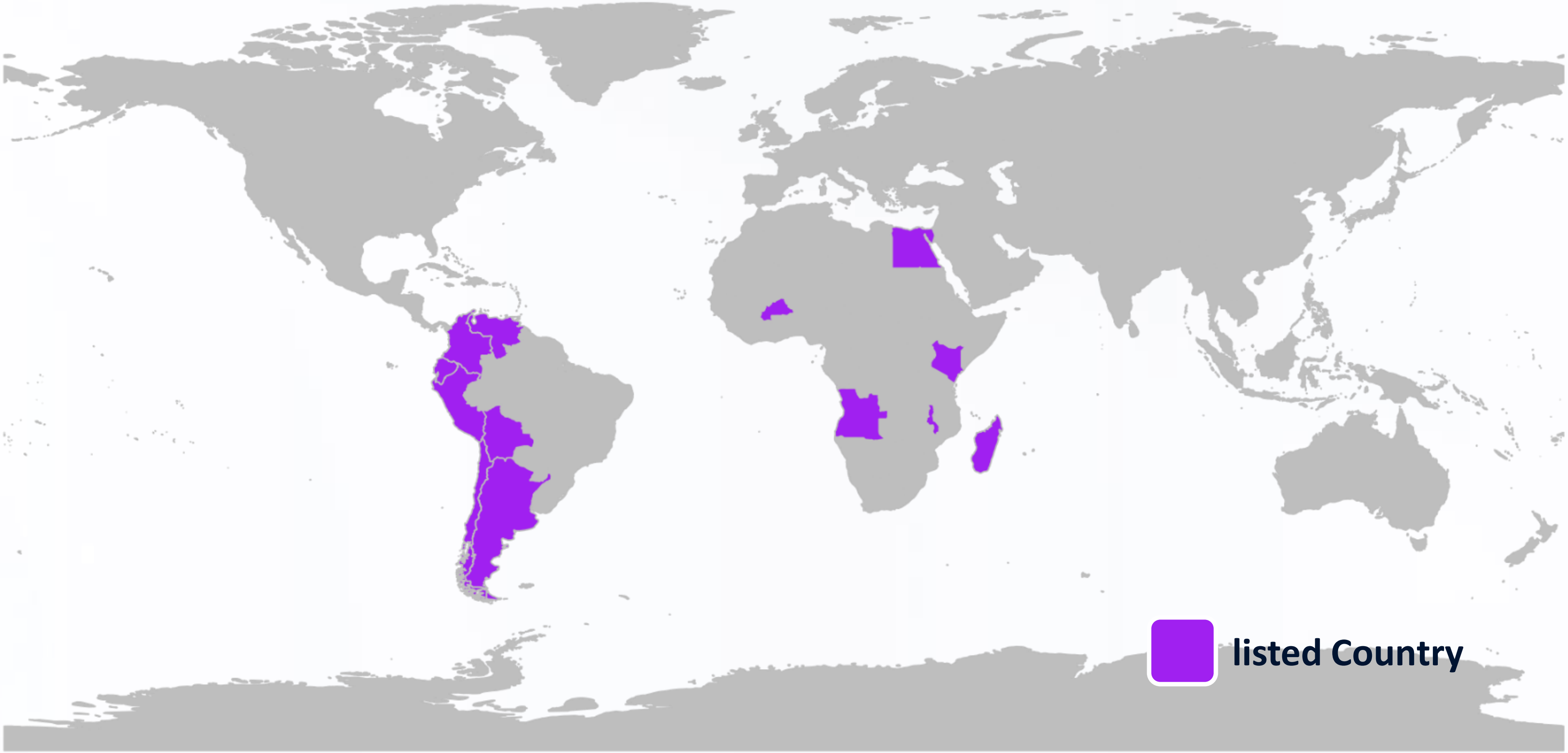
Background

Find patterns from past data

- Improve people's ability to respond to disasters
- Improve the efficiency of resource allocation

Predict future data

- Provide effective disaster prediction methods
- Reduce casualties and losses caused by disasters
- Promote mutual assistance among countries



Death, Injured, Missing, Houses Destroyed, Houses Damaged, Directly affected, Indirectly affected, Relocated, Evacuated

Disaster Data from UNDRR DesInventar Sendai is **analyzed** using **PCA**, **Classification**, and **Regression** methods.

02

Agenda

Agenda

Data cleaning

27st Nov. 35621168, 35780959

Feature Engineering & Data Visualization

29st Nov. 36497215, 36390968

PCA and Regression Modelling

1st Dec. 35780959,35621168

Classification

1st Dec. 36390968, 36497215

PPT Preparing

3st Dec. 36390968, 36497215, 35621168, 35780959

Modifing

4st Dec. 36390968, 36497215, 35621168, 35780959

03

Data Processing

Purpose: Analyze and extract useful information for the analysis target.

Data Processing Steps

01 Standardize

02 Time Filtering

03 Missing Values Imputation

04 Sort, merge columns

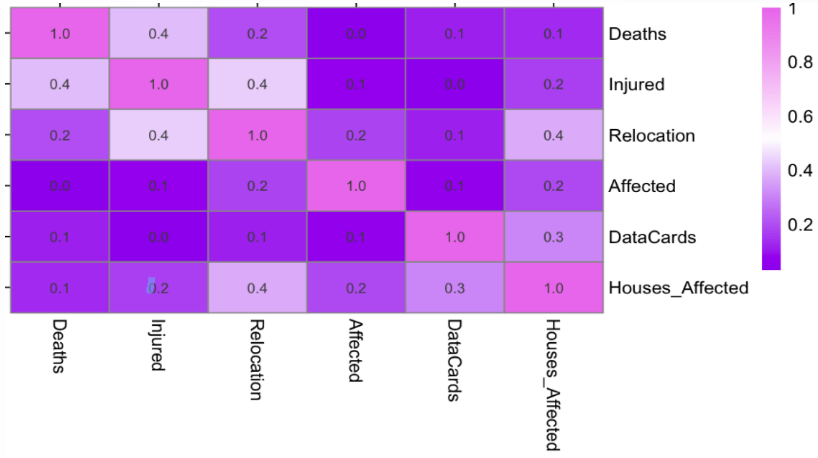
05 PCA

06 Linear& Randon Forest Regression

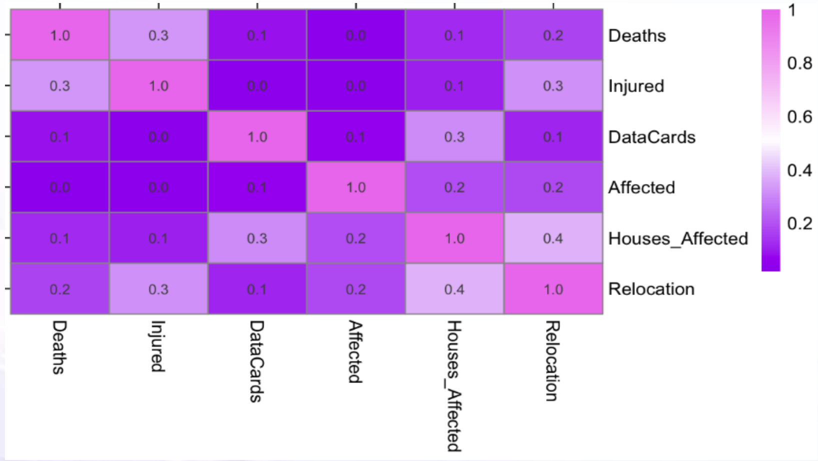
07 Classification

Data Processing

Correlation Coefficient Matrix



logarithmic transformation

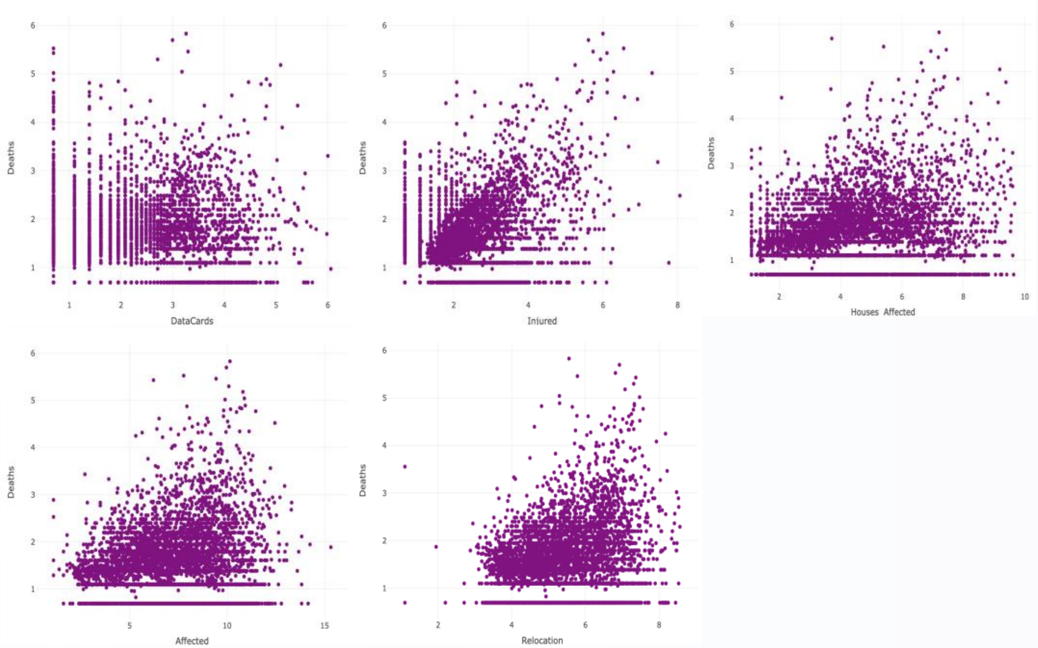


Column Descriptive Statistics

Column	Mean	Median	Variance
Country	NA	NA	NA
Year...Month	NA	NA	NA
Event	NA	NA	NA
DataCards	7.707979	2.0	6.000241e+02
Deaths	16.594033	3.0	1.248415e+05
Injured	182.794964	6.0	4.055905e+06
Missing	264.702760	3.0	8.869154e+07
Houses.Destroyed	255.687785	9.0	4.726256e+07
Houses.Damaged	981.697236	25.0	3.195978e+08
Directly.affected	4894.461261	130.0	2.205081e+09
Indirectly.Affected	21487.271648	292.0	2.351915e+10
Relocated	3284.429688	103.5	3.308388e+08
Evacuated	1581.893027	115.0	6.570118e+07

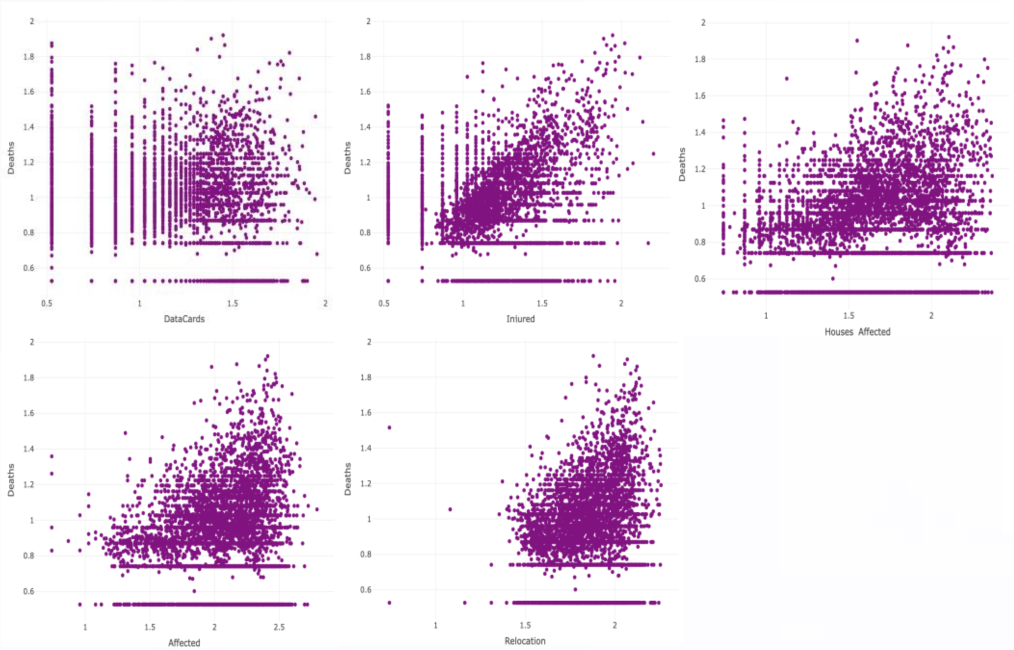
Scatter plot

Before



Logarithmic transformations reduce skewness, handle outliers, strengthen linear relationships, improve interpretability, and help model convergence.

After



We applied logarithmic transformation in feature engineering to normalize skewed data, reducing the impact of large outliers and improving subsequent classification and regression analysis.

04

Analysis

PCA

Purpose: Reduce dimensions affecting the dependent variable "DEATH" to simplify regression, while preserving as much variance (information) as possible.

PCA

Step1

Loadings Table

A matrix: 5 × 5 of type dbl

	PC1	PC2	PC3	PC4	PC5
DataCards	0.2949630	0.91159968	-0.1422123	0.2372822	-0.07386243
Injured	0.4768959	-0.34335006	-0.1971656	0.3974797	-0.67662145
Houses_Affected	0.4647159	0.06097482	0.7695872	-0.4016174	-0.16358552
Affected	0.4761309	-0.06546760	-0.5800093	-0.6370584	0.16358229
Relocation	0.4928545	-0.20758810	0.1105717	0.4675116	0.69513190

Step2

Proportion of Variance and Cumulative Proportion

```
Importance of components:
                          PC1    PC2     PC3    PC4     PC5
Standard deviation     1.8144 0.9106 0.63209 0.5211 0.45585
Proportion of Variance 0.6584 0.1658 0.07991 0.0543 0.04156
Cumulative Proportion  0.6584 0.8242 0.90414 0.9584 1.00000
```

PC processing Result:

1. Data easy to regression

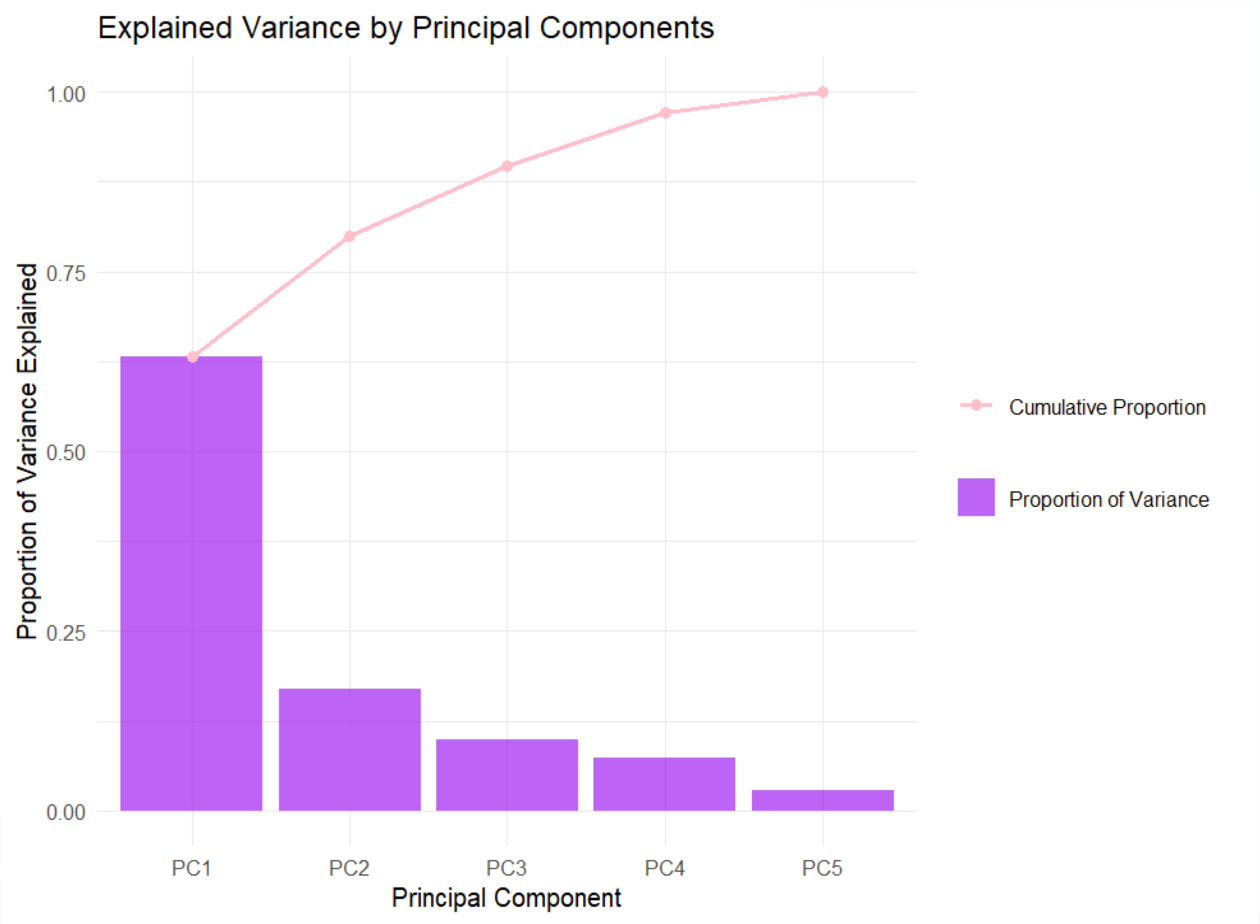
2. Reduce multicollinearity

3. Dimensionality reduction(13—>5):
PCA reduces the dimensionality of the dataset, which can simplify the model during regression and improve computational efficiency.

4. The principal components are orthogonal

PCA

Step3



Reasons for pick the first 4 PC:

1. Data integrity:

Filtering by 0.8 rate selects the first two. However, but important structural information is lost. Choosing the first four can get enough information(96%).

2. Enough independent variables

Choosing the first four provides sufficient information and enough independent variables for subsequent regression and classification models.

Regression with

Multiple linear regression

Random forest regression

Multiple Linear Regression

Step1

Deaths~PC1-PC4 Linear regression

```
Call:
lm(formula = Deaths ~ PC1 + PC2 + PC3 + PC4, data = trainData_lm)

Residuals:
    Min      1Q  Median      3Q     Max 
-87.63  -2.08  -0.95   0.90 327.37 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.49347    0.09380  79.891  < 2e-16 ***
PC1          5.40289    0.05173 104.441  < 2e-16 ***
PC2          3.01489    0.10496  28.725  < 2e-16 ***
PC3         -0.65986    0.14924  -4.422 9.88e-06 ***
PC4         -2.18125    0.17977 -12.133  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Deaths== 7.49+5.403PC1+3.015PC2-0.660PC3-2.181PC4

Conclusion:

During data cleaning, we removed most of outliers and eliminated multicollinearity. Given sufficient data and information, we suspected a more obvious nonlinear relationship, so we applied a new regression method.

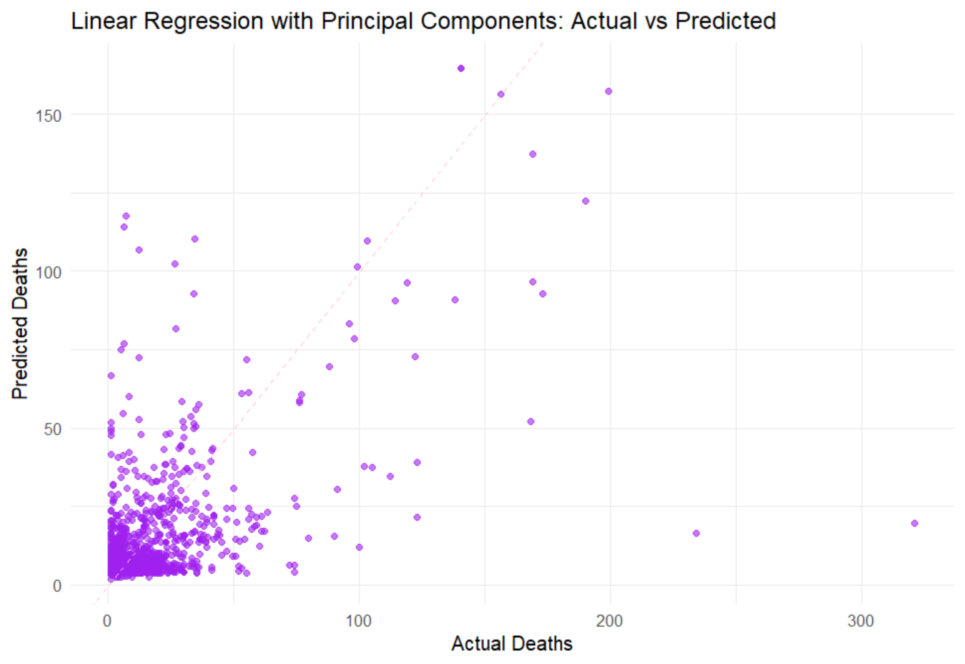
Step2

MSE & R^2 for

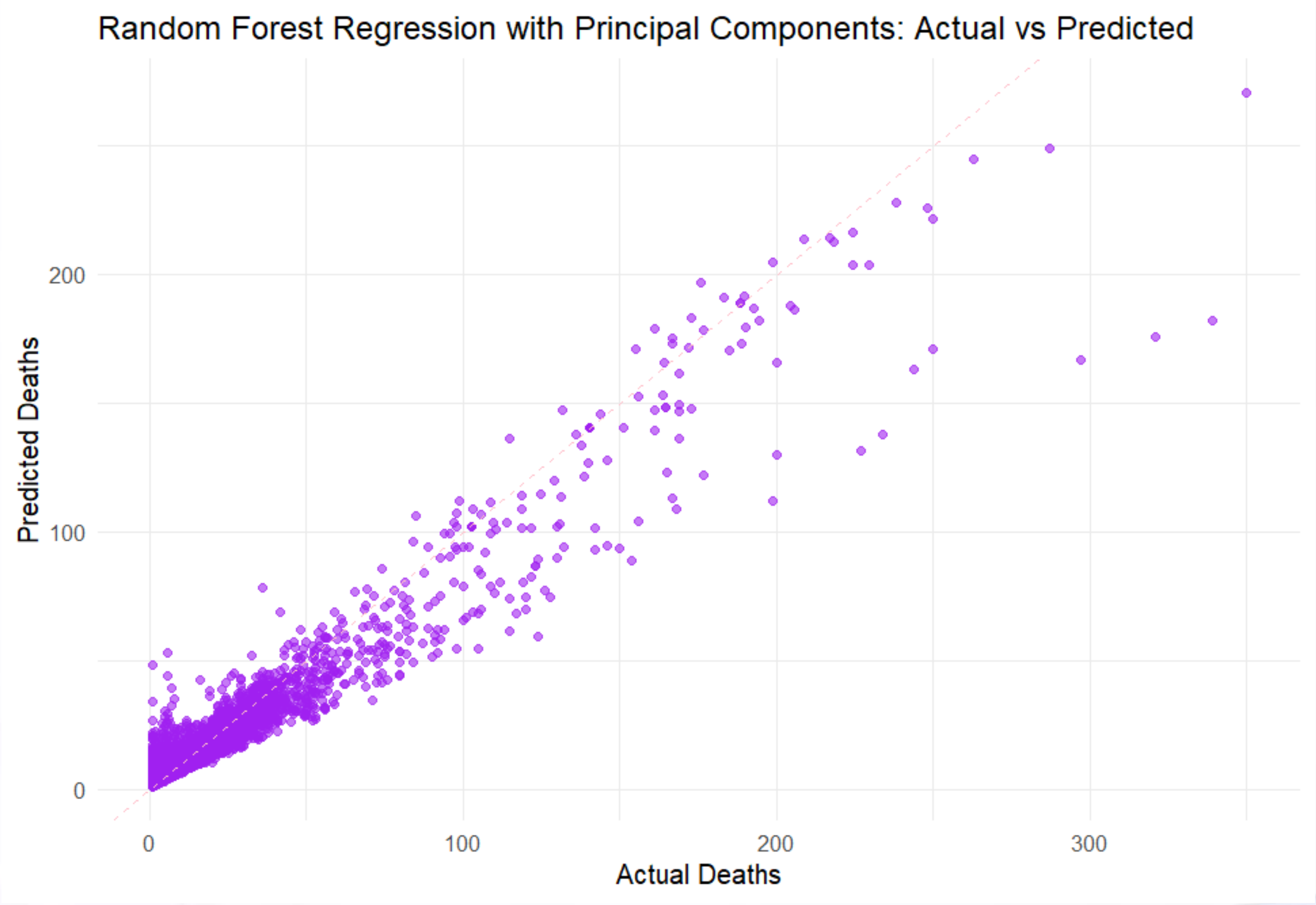
```
Linear Regression with Principal Components:
MSE:  128.2372
R-squared:  0.5043109
```

Step3

Result



Random ForestRegression




```
Random Forest with Top 4 Principal Components:
R-squared:  0.9101979
MSE:  20.78319
```

The Random Forest regression outperforms linear regression in R-squared and MSE for two main reasons:

1. **Nonlinear Relationships**:
Random Forest can capture nonlinear patterns, while linear regression makes it difficult to handle them, leading to poorer performance.

2. **Noise**:
Linear regression is more affected by noise, but Random Forest reduces noise impact through multiple samplings and independent training, improving accuracy and robustness.

Classification

We made three classifications based on continent, economy, and events, and applied three different model approaches to each classification.

```r
data$continent <- case_when(
  data$country %in% c("Angola", "Burkina Faso", "Egypt", "Kenya", "Madagascar", "Malawi") ~ "Africa",
  data$country %in% c("Argentina", "Bolivia", "Chile", "Colombia", "Ecuador", "Peru", "Venezuela") ~ "South America",
  TRUE ~ "Other")
data$economic_status <- case_when(
  data$country %in% c("Argentina", "Chile", "Colombia", "Peru") ~ "High Developeing",
  data$country %in% c("Bolivia", "Ecuador", "Egypt", "Venezuela") ~ "Medium Developing",
  data$country %in% c("Angola", "Burkina Faso", "Kenya" ,"Madagascar", "Malawi") ~ "Lower Developing",
  TRUE ~ "Other")
data$disaster_type <- case_when(
  data$event %in% c("DROUGHT", "FLOOD", "RIVERINE FLOOD", "EARTHQUAKE", "LANDSLIDE",
                    "HAILSTORM", "TORNADO", "LIGHTNING", "CYCLONE", "AVALANCHE",
                    "SNOWSTORM", "FOREST FIRE", "HEATWAVE", "MUDSLIDE", "TSUNAMI",
                    "SURGE", "STRONG WIND", "STORM", "RAINS", "ELECTRICSTORM",
                    "FLASH FLOOD", "HEAVY RAIN", "STORMY RAIN") ~ "Natural",
  data$event %in% c("FIRE", "AVIATION ACCIDENT", "PLANE CRASH", "ROAD ACCIDENT",
                    "TRAINS ACCEDNTS", "NAVIGATION ACCIDENT", "CONSTRUCTION COLLAPSE",
                    "STRUCTURAL COLLAPSE", "POLLUTION", "POLLUTION MARINE", "EXPLOSION",
                    "MATERIALS PELIGROSOS", "TERROR ATTACK", "INDUSTRIAL ACCIDENT") ~ "Human-made",
  TRUE ~ "Other"
)
```

Feature Engineering

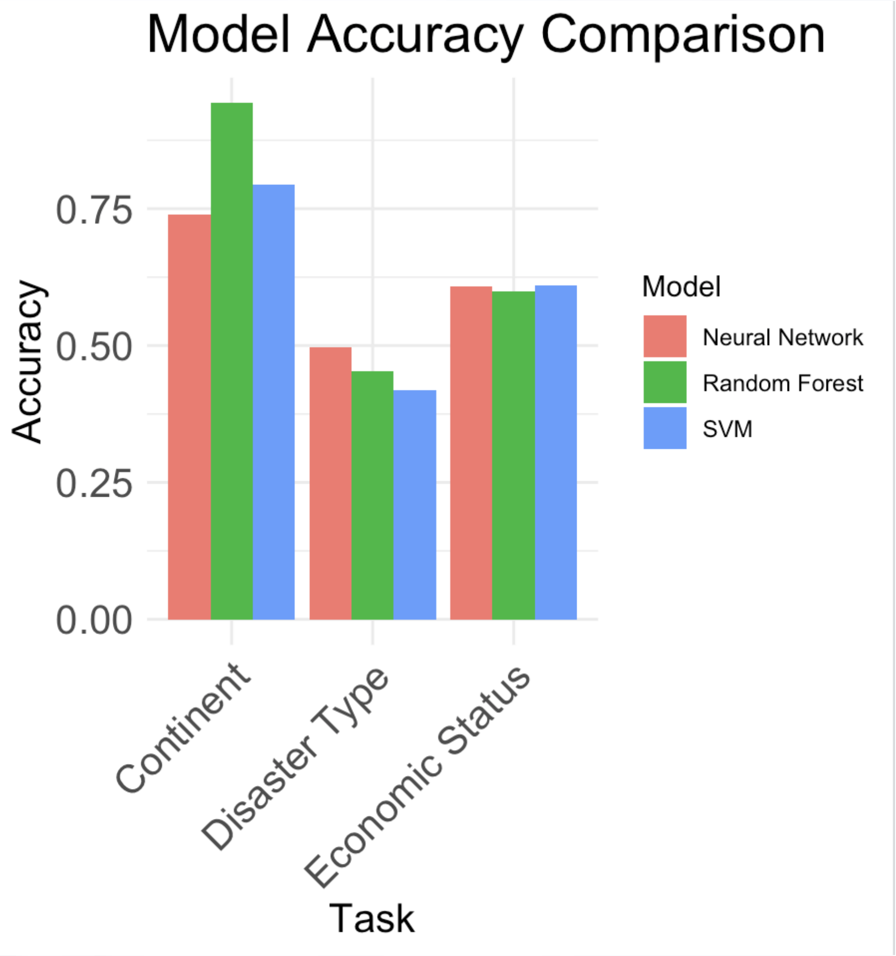
Model Construction

```r
run_models <- function(target_var, train_data, test_data, task_name) {
  # Random Forest
  rf_model <- randomForest(as.formula(paste(target_var, "~ deaths + injured + houses_affected  ")),
                           data = train_data, ntree = 500, mtry = 2)
  rf_pred <- predict(rf_model, newdata = test_data)
  rf_cm <- confusionMatrix(rf_pred, test_data[[target_var]])

  # SVM
  svm_model <- svm(as.formula(paste(target_var, "~ deaths + injured + houses_affected")),
                   data = train_data, kernel = "radial", cost = 1, gamma = 0.1)
  svm_pred <- predict(svm_model, newdata = test_data)
  svm_cm <- confusionMatrix(svm_pred, test_data[[target_var]])

  # Neural Network
  nn_model <- nnet(as.formula(paste(target_var, "~ deaths + injured + houses_affected  ")),
                   data = train_data, size = 5, decay = 0.1, maxit = 200)
  nn_pred <- predict(nn_model, newdata = test_data, type = "class")
  nn_cm <- confusionMatrix(as.factor(nn_pred), test_data[[target_var]])
```

Performance and Task Difficulty



Performance of different models

- The Random Forest model has the highest accuracy in predicting the disaster type task.

- In the prediction of disaster types, the accuracy of all three models was not very high, indicating that the task may be challenging and the relationship between data features and disaster types may be complex.

- From the chart, it can be seen that the task of predicting economic status is relatively difficult.

Model Performance Comparison

Random Forest, SVM and Neural Network

Model Performance Comparison

Task	Model	Accuracy
Disaster Type	Random Forest	0.452277657266811
Disaster Type	SVM	0.418655097613883
Disaster Type	Neural Network	0.496746203904555
Economic Status	Random Forest	0.598698481561822
Economic Status	SVM	0.609544468546638
Economic Status	Neural Network	0.608459869848156
Continent	Random Forest	0.942578548212351
Continent	SVM	0.794149512459372
Continent	Neural Network	0.739978331527627

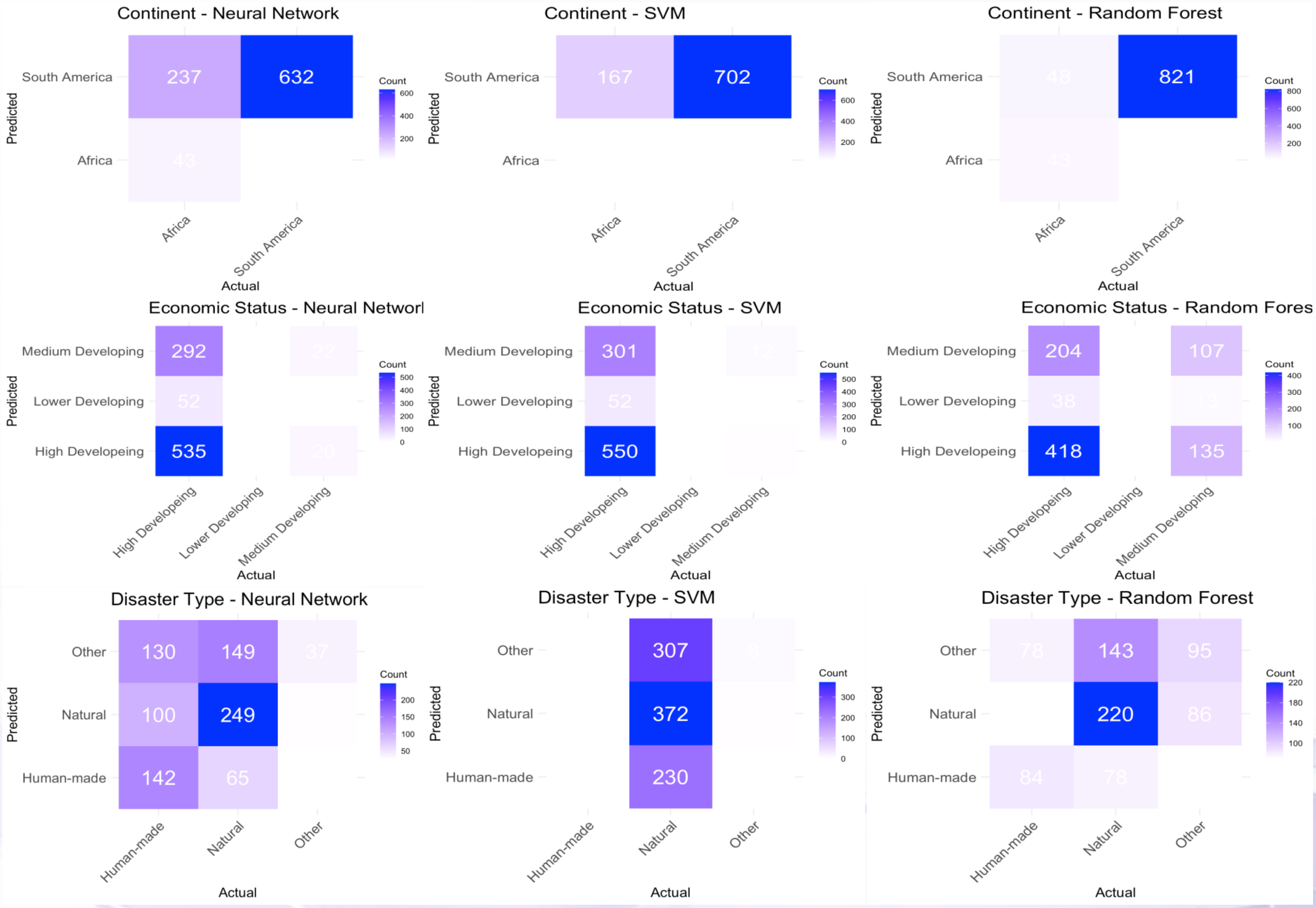
We compared the performance of three algorithms by using confusion matrices and relevant statistics.

1.For the "**continent**", the Random Forest model performed best with an accuracy of 0.9426, indicating strong classification ability.

2.For the “**economic status**”, the SVM led with 0.6095 accuracy, slightly ahead of the neural network (0.6085), suggesting that complex morphologies can be better handled.

3.For the "**disaster type**", the accuracy of all models was low, with the neural network at 0.4967, indicating difficulty in classification or possible dataset issues.

Confusion matrix



The confusion matrix, by showing the relationship between the model's predictions and the actual labels, helps us intuitively understand which categories the model performs well in and which categories it tends to make errors in.

Conclusion

1. **Disaster type and death**: Prediction models should be tailored to different disaster types, with enhanced monitoring of high-risk areas.

2. **Economic status and death**: Wealthier countries tend to have lower disaster mortality rates, while poorer ones face higher rates. Improving economic conditions can reduce disaster impact.

3. **Regional variation in deaths**: The death toll varies by region. Strengthening international cooperation can help reduce casualties.

4. **Machine learning in disaster management**: Technologies like random forest, SVM, and neural networks can improve disaster prediction and early warning, reducing casualties.

5. **Analysis of key influencing factors:** Although we did not focus on temporal characteristics in the current analysis, the impact of seasonality on disaster types is also a potentially critical factor. For example, hazards such as floods and droughts are often seasonal and prone to occur in certain specific months.

Thanks!