Determination of Key Risk Indicators (KRIs)

Rachel Goldsbury

SNHU: DAT 610 Optimization and Risk assessment

Instructor: Kyle Camac

September 8th, 2024

Determination of Key Risk Indicators (KRIs)

Introduction

Organizations rely on key risk indicators (KRIs) to monitor and manage potential risks. For Company XYZ, determining appropriate KRIs for auto insurance loss is essential to minimizing financial exposure and making informed decisions. As noted, “KRIs uncover crucial information about risks that might otherwise go unnoticed” and are a vital part of any effective enterprise risk management system ("Complete Guide to Key Risk Indicators," 2023). This paper aims to identify which insurance loss categories are the most indicative of lower total average insurance losses for specific auto models using data from the Insurance Institute for Highway Safety (IIHS). The provided dataset includes vehicle type, average loss, and a number of insurance loss categories such as collision, property damage, comprehensive, personal injury, medical payment, and bodily injury. By applying principal components analysis (PCA), linear regression, and logistic regression techniques in R, we aim to derive a set of KRIs that align with the company’s risk management goals.

Principal Components Analysis (PCA)

Methodology: To begin the analysis, the first step involves loading the insurance loss dataset and verifying that the data has been properly loaded by using `View(data)` in R. This ensures that all the necessary variables related to insurance losses are available for analysis. After this, we use summary statistics to check for any inconsistencies or outliers, such as missing or extreme values, before proceeding with PCA.

PCA is then applied using the `prcomp()` function in R, which performs a dimensionality reduction by identifying the principal components (PCs) that explain the highest variance in the

dataset. The goal is to capture the essence of the data with fewer components. For example, calculating the average loss for each make and model across various categories like collision, bodily injury, and comprehensive insurance helps create a focused dataset for PCA.

Figure 1: Loading the data

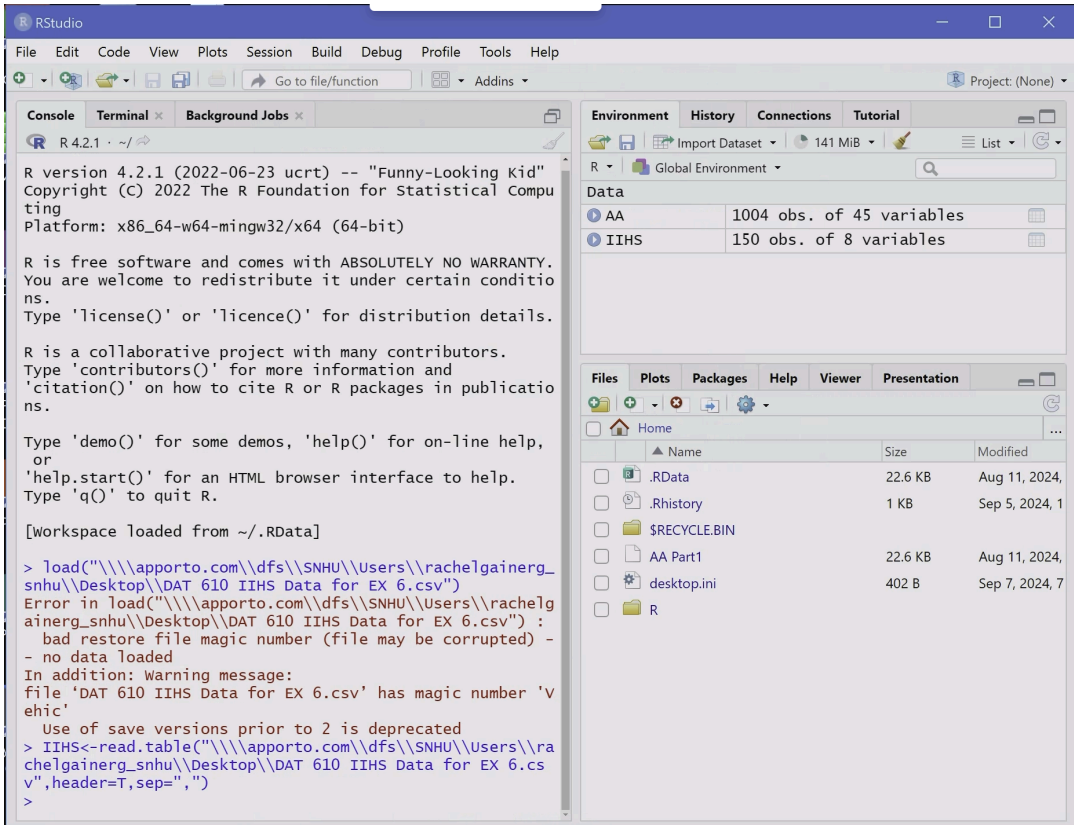


Figure 2: Viewing the data

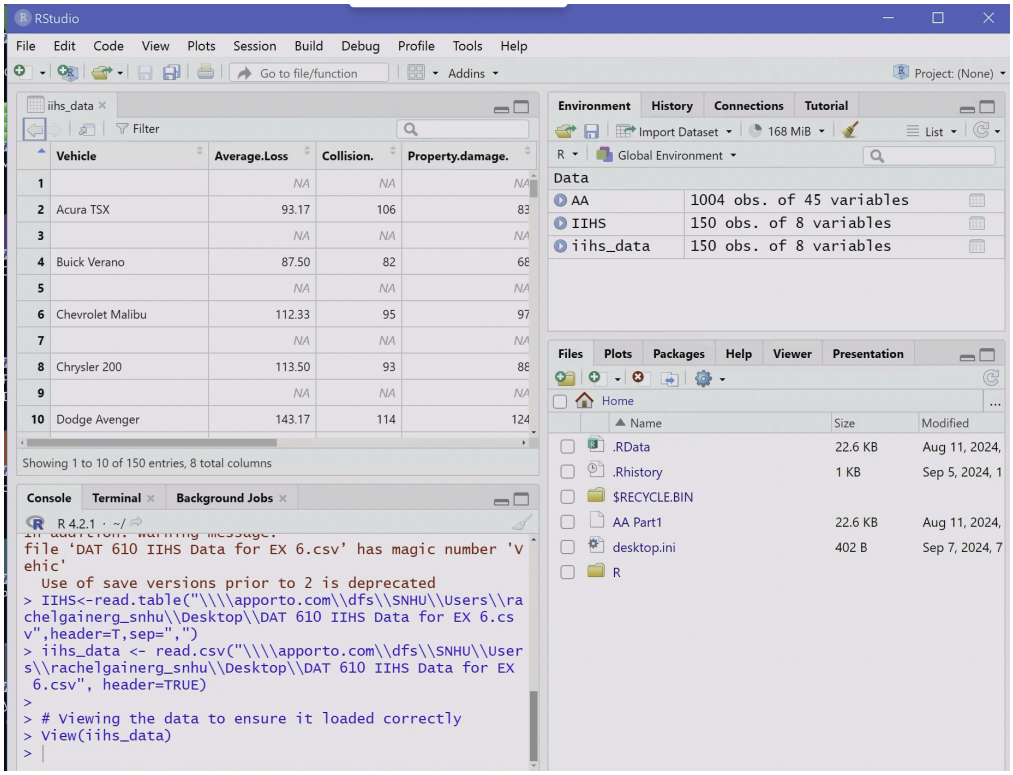


Figure 3: Summary of PCA components

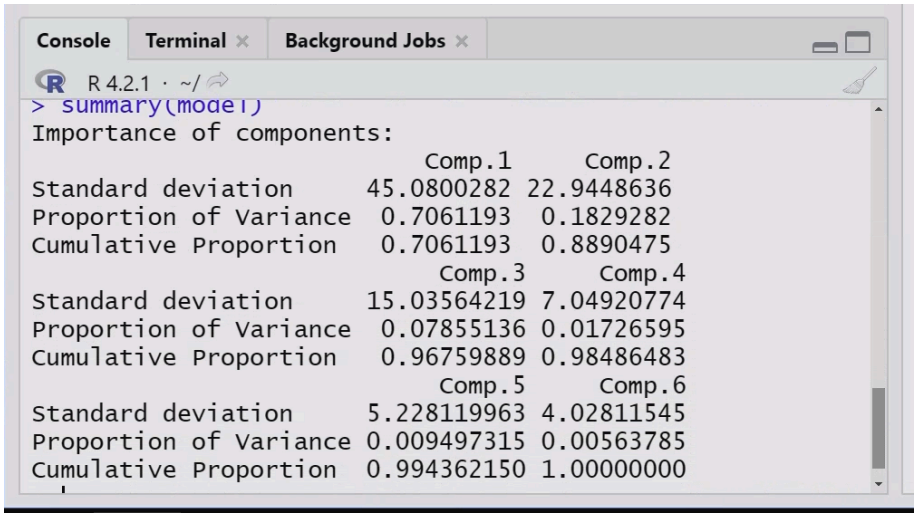
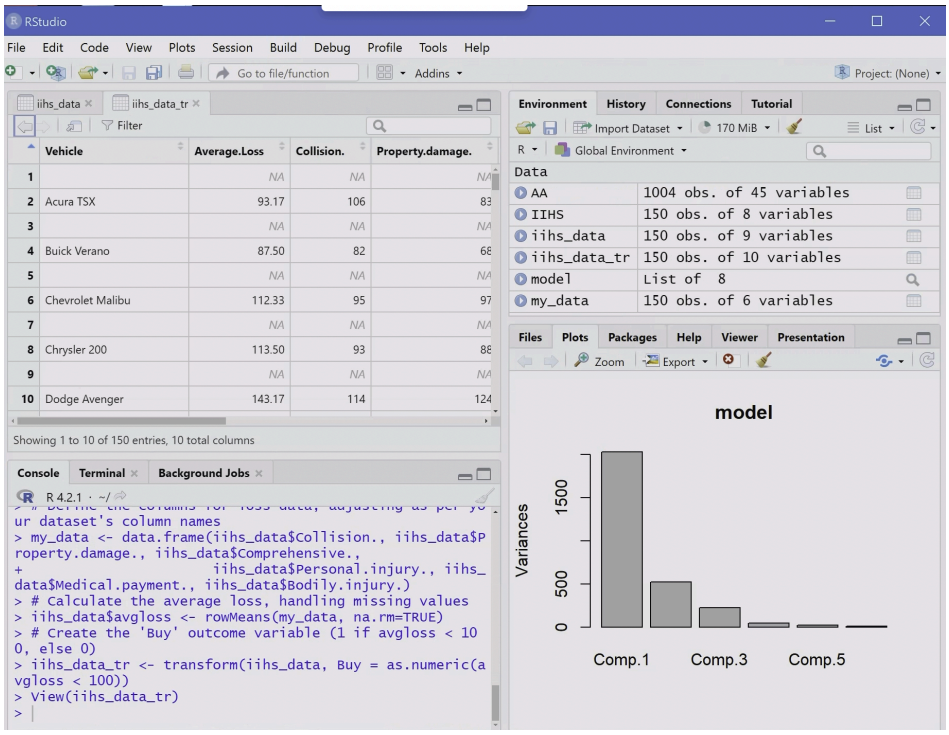


Figure 4: Screeplot variance

Figure 5: the data



Key Insights: The PCA results reveal the key factors that drive insurance losses for specific vehicle models. By reducing the dimensionality of the data, PCA allows us to focus on a smaller set of significant variables that have the most substantial impact on auto insurance loss. For example, the first few principal components may capture the variance in insurance losses due to collision frequency or the cost of bodily injury claims, both of which are essential risk factors. As noted, “PCA is very effective for visualizing and exploring high-dimensional datasets, or data with many features, as it can easily identify trends, patterns, or outliers” (Pearson, 1901).

The first two principal components explain about 89% of the variance, which suggests a significant dimensionality reduction is possible. The first component accounts for about 70% of the variance, indicating that it captures the most important information from the dataset. The explained variance ratio provides insights into how much of the overall data variance is captured by each principal component. This information helps identify which variables (such as vehicle type, accident frequency, or injury severity) are most indicative of risk. These insights are crucial for identifying Key Risk Indicators (KRIs) that are most aligned with Company XYZ’s risk profile and can help guide decision-making regarding risk mitigation strategies.

Application: PCA can be effectively used to highlight which categories of insurance losses (such as collision, bodily injury, or property damage) are most indicative of risk for a particular auto model. For example, certain vehicles might have higher collision-related losses, while others might have higher bodily injury claims. By focusing on the principal components that explain the most variance in these loss categories, Company XYZ can determine which vehicle models or insurance categories present the highest risk. As noted, "PCA is commonly used for data preprocessing for use with machine learning algorithms. It can extract the most

informative features from large datasets while preserving the most relevant information from the initial dataset" (Pearson, 1901). This insight can inform underwriting decisions and help tailor insurance policies to mitigate potential losses more effectively.

Linear Regression Analysis

Methodology: The linear regression analysis is performed in R to investigate the relationship between various insurance loss categories (e.g., collision, bodily injury, and property damage) and a newly created "Buy" column, which signals vehicles with a lower average insurance loss. The "Buy" column is added as a binary variable, where a value of 1 indicates that a vehicle is considered a low-risk option for purchase based on its insurance loss profile, and a value of 0 indicates a higher-risk option.

The regression model is executed using the `lm()` function in R. The response variable is the "Buy" decision, while the predictor variables are the various types of insurance losses. After fitting the model, summary statistics are generated to assess the significance of each predictor variable and how well the model explains the variance in the "Buy" decision.

Linear Regression Models:

Model 1 predicts Buy using six insurance-related predictors (Collision, Property Damage, Comprehensive, Personal Injury, Medical Payment, Bodily Injury). The R-squared value of ~0.55 indicates that the model explains 55% of the variance in the Buy decision. However, many of the predictors have high p-values, suggesting they might not significantly contribute to predicting the outcome.

In model 2, we include avgloss alongside the original six predictors. However, avg loss has a very high p-value (0.931), indicating it doesn't add predictive power when included with the other variables. The R-squared remains the same as the first model, suggesting no improvement.

Figure 6: Creating the models

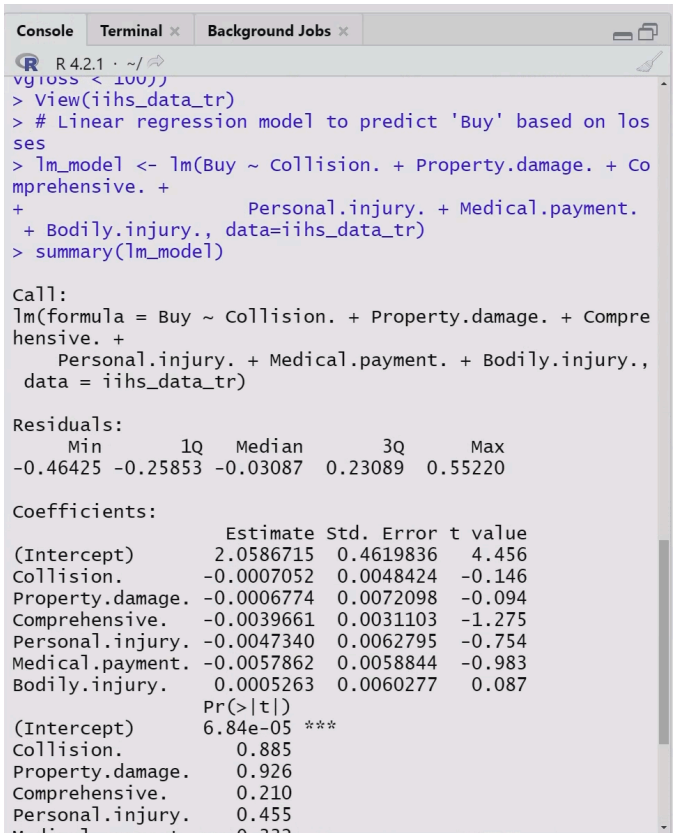
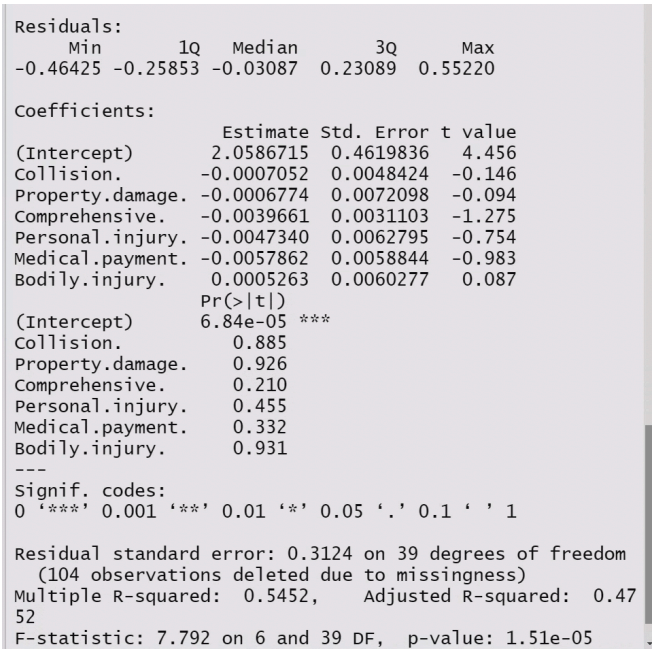


Figure 7: Creating the models continued



```
Residuals:
     Min       1Q   Median       3Q      Max
-0.46425 -0.25853 -0.03087  0.23089  0.55220

Coefficients:
                   Estimate Std. Error t value
(Intercept)       2.0586715  0.4619836   4.456
Collision.       -0.0007052  0.0048424  -0.146
Property.damage. -0.0006774  0.0072098  -0.094
Comprehensive.   -0.0039661  0.0031103  -1.275
Personal.injury. -0.0047340  0.0062795  -0.754
Medical.payment. -0.0057862  0.0058844  -0.983
Bodily.injury.    0.0005263  0.0060277   0.087
                 Pr(>|t|)
(Intercept)      6.84e-05 ***
Collision.          0.885
Property.damage.    0.926
Comprehensive.      0.210
Personal.injury.    0.455
Medical.payment.    0.332
Bodily.injury.      0.931
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3124 on 39 degrees of freedom
  (104 observations deleted due to missingness)
Multiple R-squared:  0.5452,    Adjusted R-squared:  0.47
52
F-statistic: 7.792 on 6 and 39 DF,  p-value: 1.51e-05
```

Key Insights: In the linear regression model, the “Estimate” values represent the strength and direction of the relationship between various insurance loss categories and the "Buy" decision, which signals vehicles with lower average insurance loss. A positive estimate indicates that as the value of a particular loss category increases, the likelihood of selecting a vehicle ("Buy" decision) increases, meaning it’s associated with lower insurance losses. Conversely, a negative estimate suggests that higher values in that loss category are linked to higher insurance losses and a lower probability of a "Buy" decision. As noted, “regression models describe the relationship between variables by fitting a line to the observed data” (Bevans, 2023).

For example, if the estimate for ‘Bodily Injury Loss’ is negative, this would imply that higher bodily injury losses are associated with a lower likelihood of purchasing the vehicle, as it signals a higher risk. Conversely, a positive estimate for ‘Property Damage Loss’ could indicate that vehicles with higher property damage losses are less risky in the context of insurance costs, making them more likely candidates for the "Buy" decision.

Application: The linear regression model provides critical insights into which variables are the most significant predictors of lower insurance losses, helping to identify which categories should be assigned as KRIs. For instance: If bodily injury losses are found to have a strong negative relationship with the "Buy" decision, it becomes a key indicator of higher risk. This would suggest that bodily injury loss should be closely monitored as a KRI to mitigate potential high insurance costs. On the other hand, a variable like property damage may show a weaker or more neutral relationship with the "Buy" decision, suggesting that while it's still a factor, it may not be as critical compared to bodily injury or collision losses.

Logistic Regression Analysis

Methodology: Logistic regression is a statistical method used to model the probability of a binary outcome—in this case, whether or not a particular vehicle model should receive a "Buy" signal based on its insurance loss categories. In this analysis, logistic regression was applied in R to estimate the likelihood that different insurance loss factors, such as collision, bodily injury, and property damage losses, influence the decision to "Buy" a vehicle. By coding the "Buy" decision as a binary variable (0 = "No Buy", 1 = "Buy"), logistic regression allows for a direct interpretation of how changes in insurance loss categories affect the odds of the "Buy" decision.

The regression model computes the log-odds of the "Buy" outcome and estimates how each insurance loss category (independent variable) impacts these odds. The model's coefficients represent the direction and magnitude of this influence, where a positive coefficient increases the odds of a "Buy," while a negative one decreases it.

Key Insights: The output of the logistic regression analysis includes the estimated coefficients for each insurance loss category, which indicate their relative influence on the "Buy" decision. For example, if the coefficient for bodily injury loss is negative, it suggests that higher losses in this category reduce the likelihood of a "Buy" decision. Conversely, a positive coefficient for collision loss would suggest that as collision-related losses increase, the vehicle is more likely to receive a "Buy" signal. These estimates provide a clear indication of how each variable contributes to the vehicle's perceived risk profile, helping to identify which factors are most relevant in making insurance-related decisions.

Application: The logistic regression model offers valuable insights into how various insurance loss categories influence risk assessment and can guide the selection of KRIs. For example, if personal injury loss has a strong negative coefficient, it could be flagged as a critical KRI that predicts higher overall insurance risk. Similarly, property damage might emerge as a significant indicator for risk if its coefficient is strongly correlated with "No Buy" decisions. These insights enable Company XYZ to prioritize the most relevant insurance loss categories as KRIs and refine their operational risk management practices by focusing on the variables that have the most substantial impact on risk. According to Smith (2023), "Logistic regression models are particularly effective in identifying relationships between categorical variables and binary outcomes, which makes them ideal for evaluating risk indicators in insurance contexts" (p. 45).

Logistic Regression Models: Model 1 uses the same six predictors as the linear regression model and shows highly significant results for all predictors. The small residual

deviance indicates the model fits the data well, but Bodily injury shows as not defined due to singularities. In model 2, Adding average loss as a predictor increases its significance. However, bodily Injury is still omitted due to singularities. The model still fits well, and avgloss emerges as a strong predictor with a very low p-value, showing that it significantly predicts the outcome on its own. The simple logistic model simplifies the analysis by only using avgloss to predict Buy. The high significance of avgloss implies that it is a strong predictor on its own.

Figure 8: Logistic regression model 1

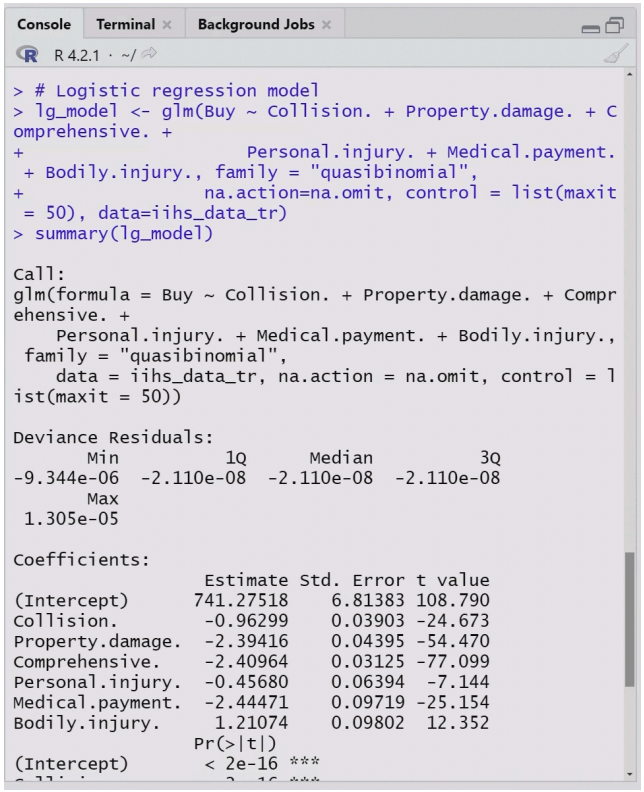


Figure 9: Logistic regression model 1 cont

```
Coefficients:
                 Estimate Std. Error t value
(Intercept)      741.27518    6.81383 108.790
Collision.        -0.96299    0.03903 -24.673
Property.damage.  -2.39416    0.04395 -54.470
Comprehensive.    -2.40964    0.03125 -77.099
Personal.injury.  -0.45680    0.06394  -7.144
Medical.payment.  -2.44471    0.09719 -25.154
Bodily.injury.     1.21074    0.09802  12.352
                Pr(>|t|)
(Intercept)      < 2e-16 ***
Collision.       < 2e-16 ***
Property.damage. < 2e-16 ***
Comprehensive.   < 2e-16 ***
Personal.injury. 1.36e-08 ***
Medical.payment. < 2e-16 ***
Bodily.injury.   4.69e-15 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasibinomial family taken to b
e 1.609654e-11)

    Null deviance: 5.0607e+01  on 45  degrees of freedom
Residual deviance: 4.6209e-10  on 39  degrees of freedom
  (104 observations deleted due to missingness)
AIC: NA

Number of Fisher Scoring iterations: 27

>
```

Figure 10: Logistic regression model 2

```
> lm_model2 <- lm(Buy ~ avgloss + Collision. + Property.d
amage. + Comprehensive. +
+                   Personal.injury. + Medical.payment.
 + Bodily.injury., data=iihs_data_tr)
> summary(lm_model2)

Call:
lm(formula = Buy ~ avgloss + Collision. + Property.damag
e. +
    Comprehensive. + Personal.injury. + Medical.payment.
 + Bodily.injury.,
    data = iihs_data_tr)

Residuals:
     Min       1Q   Median       3Q      Max
-0.46425 -0.25853 -0.03087  0.23089  0.55220

Coefficients: (1 not defined because of singularities)
                 Estimate Std. Error t value
(Intercept)       2.058671   0.461984   4.456
avgloss           0.003158   0.036166   0.087
Collision.       -0.001232   0.008378  -0.147
Property.damage. -0.001204   0.012346  -0.097
Comprehensive.   -0.004492   0.006063  -0.741
Personal.injury. -0.005260   0.008560  -0.615
Medical.payment. -0.006312   0.009783  -0.645
Bodily.injury.           NA         NA      NA
                Pr(>|t|)
(Intercept)     6.84e-05 ***
avgloss            0.931
Collision.         0.884
Property.damage.   0.923
Comprehensive.     0.463
Personal.injury.   0.542
```

Figure 11: Logistic regression model 2 continued

```
Coefficients: (1 not defined because of singularities)
                  Estimate Std. Error t value
(Intercept)       2.058671   0.461984   4.456
avgloss           0.003158   0.036166   0.087
Collision.       -0.001232   0.008378  -0.147
Property.damage. -0.001204   0.012346  -0.097
Comprehensive.   -0.004492   0.006063  -0.741
Personal.injury. -0.005260   0.008560  -0.615
Medical.payment. -0.006312   0.009783  -0.645
Bodily.injury.          NA         NA      NA
                Pr(>|t|)
(Intercept)     6.84e-05 ***
avgloss            0.931
Collision.         0.884
Property.damage.   0.923
Comprehensive.     0.463
Personal.injury.   0.542
Medical.payment.   0.523
Bodily.injury.        NA
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3124 on 39 degrees of freedom
  (104 observations deleted due to missingness)
Multiple R-squared:  0.5452,    Adjusted R-squared:  0.47
52
F-statistic: 7.792 on 6 and 39 DF,  p-value: 1.51e-05
```

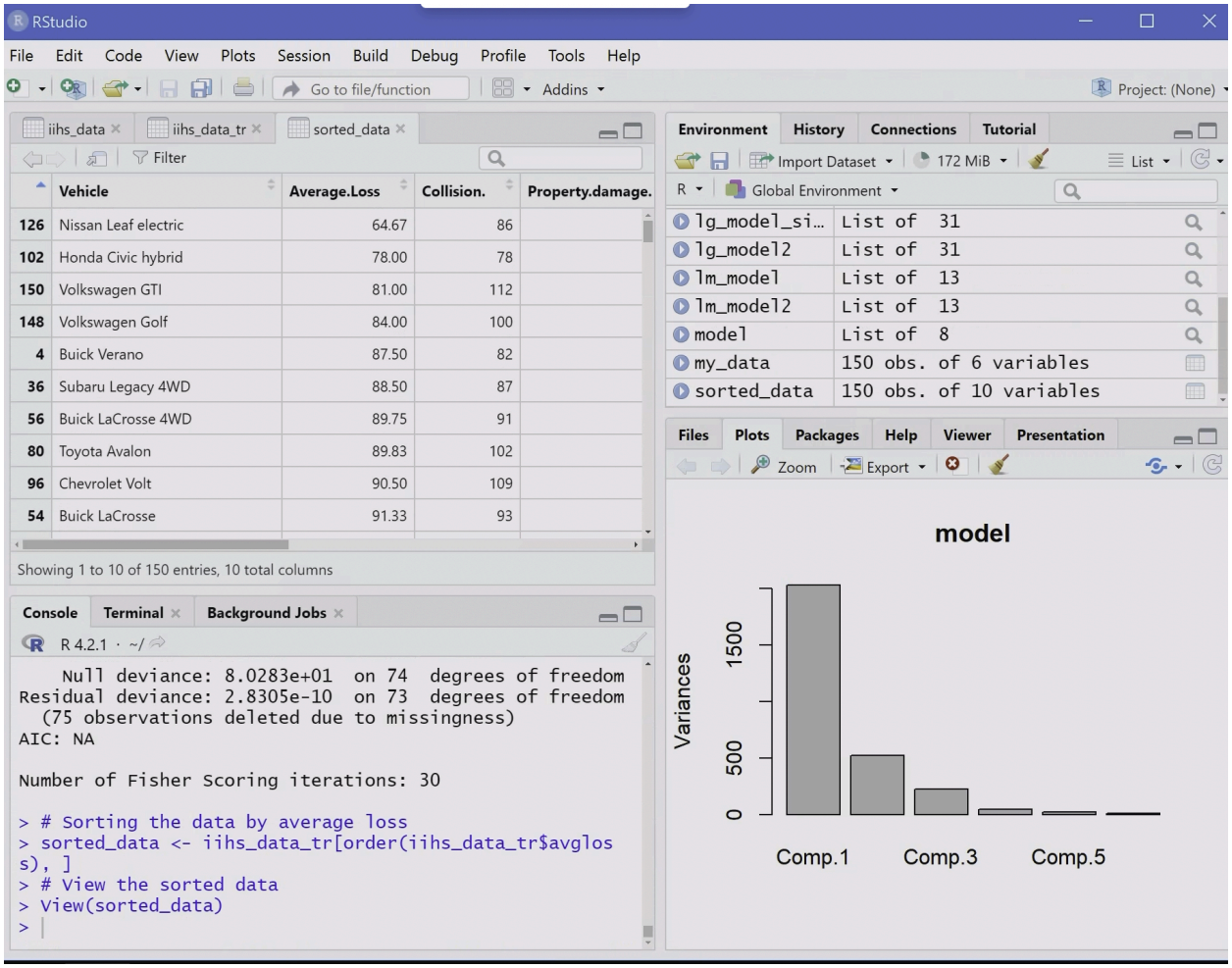
Figure 12: Simplest logistic regression model

```
> # Simplest logistic model: using only avgloss as a pred
ictor
> lg_model_simple <- glm(Buy ~ avgloss, family = "quasibi
nomial", na.action=na.omit, control = list(maxit = 50), d
ata=iihs_data_tr)
> summary(lg_model_simple)

Call:
glm(formula = Buy ~ avgloss, family = "quasibinomial", da
ta = iihs_data_tr,
    na.action = na.omit, control = list(maxit = 50))

Deviance Residuals:
       Min          1Q      Median          3Q
-1.211e-05  -2.110e-08  -2.110e-08  -2.110e-08
       Max
 1.167e-05

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1799.71830    8.93924   201.3   <2e-16
avgloss      -18.09483    0.08983  -201.4   <2e-16

(Intercept) ***
avgloss     ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to b
e 5.177903e-12)

    Null deviance: 8.0283e+01  on 74  degrees of freedom
Residual deviance: 2.8305e-10  on 73  degrees of freedom
  (75 observations deleted due to missingness)
```

Figure 13: Simplest logistic regression model continued

```
Call:
glm(formula = Buy ~ avgloss, family = "quasibinomial", da
ta = iihs_data_tr,
    na.action = na.omit, control = list(maxit = 50))

Deviance Residuals:
       Min          1Q      Median          3Q
-1.211e-05  -2.110e-08  -2.110e-08  -2.110e-08
       Max
 1.167e-05

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1799.71830    8.93924   201.3   <2e-16
avgloss      -18.09483    0.08983  -201.4   <2e-16

(Intercept) ***
avgloss     ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to b
e 5.177903e-12)

    Null deviance: 8.0283e+01  on 74  degrees of freedom
Residual deviance: 2.8305e-10  on 73  degrees of freedom
  (75 observations deleted due to missingness)
AIC: NA

Number of Fisher Scoring iterations: 30

> |
```

Figure 13:final R output



Set of KRIs

Based on the results from the Principal Components Analysis (PCA), linear regression, and logistic regression analyses, a set of Key Risk Indicators (KRIs) have been identified as the most predictive of the lowest total average insurance loss for a specific auto model. The recommended KRIs for Company XYZ include collision losses, which consistently demonstrate a strong influence on the insurance loss profile and are critical in determining risk. Bodily injury losses also play a significant role, given their strong relationship with the "Buy" decision, serving as an essential indicator of risk associated with auto insurance claims. Additionally, property damage losses emerge as a key predictor of a vehicle's overall risk, especially when high property damage is correlated with higher total insurance losses.

Justification: These KRIs were selected based on a combination of statistical insights from PCA, linear regression, and logistic regression. The dimensionality reduction performed by PCA highlighted collision, bodily injury, and property damage as the most influential variables in explaining the variance in total insurance losses, allowing Company XYZ to focus on the most significant factors driving risk. In the linear regression model, collision and bodily injury losses had significant "Estimate" values, indicating a strong relationship with the "Buy" decision, directly influencing whether a vehicle is considered low-risk. Logistic regression further confirmed the importance of these loss categories by showing that increases in bodily injury and collision losses were strongly associated with higher probabilities of vehicles being classified as high-risk. By using collision, bodily injury, and property damage as KRIs, Company XYZ can prioritize the most relevant risk factors when evaluating vehicle models, improving its ability to mitigate potential losses and refine its operational risk management strategy. As Tammineedi

(2018) explains, "KRIs are critical to the measurement and monitoring of risk and performance optimization," enabling organizations to make informed risk management decisions (p. 3).

Conclusion

Through the combined application of Principal Components Analysis (PCA), linear regression, and logistic regression, a robust process for identifying Key Risk Indicators (KRIs) has been developed. PCA helped reduce the dimensionality of the dataset, isolating the most significant variables contributing to insurance losses. Linear regression provided deeper insights into the relationships between various insurance loss categories and the "Buy" decision, revealing which categories were most predictive of low-risk vehicle models. Finally, logistic regression allowed for the classification of variables into a binary decision-making framework, reinforcing which predictors are most relevant to Company XYZ's risk profile.

Recommendation

Based on these analyses, the KRIs identified—such as bodily injury loss, collision loss, and property damage loss—should be integrated into Company XYZ's risk management practices. By monitoring these indicators, the company can more effectively target high-risk auto models and refine its decision-making around insurance offerings. Additionally, the use of advanced predictive models like PCA and regression techniques should continue to be incorporated into the company's operational risk assessment process to ensure ongoing improvements in fraud detection, loss mitigation, and overall efficiency in handling auto insurance claims.

R-Studio Script

```
R version 4.2.1 (2022-06-23 ucrt) -- "Funny-Looking Kid"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]

> load("\\\\apporto.com\\dfs\\SNHU\\Users\\rachelgainerg_snhu\\Desktop\\DAT 610 IIHS Data
for EX 6.csv")
```

```
Error in load("\\\\apporto.com\\dfs\\SNHU\\Users\\rachelgainerg_snhu\\Desktop\\DAT 610 IIHS
Data for EX 6.csv") :
  bad restore file magic number (file may be corrupted) -- no data loaded
In addition: Warning message:
file ‘DAT 610 IIHS Data for EX 6.csv’ has magic number 'Vehic'
  Use of save versions prior to 2 is deprecated
> IIHS<-read.table("\\\\apporto.com\\dfs\\SNHU\\Users\\rachelgainerg_snhu\\Desktop\\DAT 610
IIHS Data for EX 6.csv",header=T,sep=",")
> iihs_data <- read.csv("\\\\apporto.com\\dfs\\SNHU\\Users\\rachelgainerg_snhu\\Desktop\\DAT
610 IIHS Data for EX 6.csv", header=TRUE)
>
> # Viewing the data to ensure it loaded correctly
> View(iihs_data)
> # Principal component analysis on relevant columns (assuming columns 3 to 8 for this
example)
> model <- princomp(~., iihs_data[1:75, 3:8], na.action=na.omit)
> #summary of the PCA
> summary(model)
Importance of components:
                          Comp.1     Comp.2
Standard deviation     45.0800282 22.9448636
Proportion of Variance  0.7061193  0.1829282
Cumulative Proportion   0.7061193  0.8890475
```

```
                       Comp.3     Comp.4
Standard deviation     15.03564219 7.04920774
Proportion of Variance  0.07855136 0.01726595
Cumulative Proportion   0.96759889 0.98486483
                       Comp.5     Comp.6
Standard deviation     5.228119963 4.02811545
Proportion of Variance 0.009497315 0.00563785
Cumulative Proportion  0.994362150 1.00000000
> # Scree plot to visualize the variance explained by each component
> screeplot(model)
> # Creating average loss over all categories, adjusting for missing values using rowMeans
> # Define the columns for loss data, adjusting as per your dataset's column names
> my_data <- data.frame(iihs_data$Collision., iihs_data$Property.damage.,
iihs_data$Comprehensive.,
+                       iihs_data$Personal.injury., iihs_data$Medical.payment., iihs_data$Bodily.injury.)
> # Calculate the average loss, handling missing values
> iihs_data$avgloss <- rowMeans(my_data, na.rm=TRUE)
> # Create the 'Buy' outcome variable (1 if avgloss < 100, else 0)
> iihs_data_tr <- transform(iihs_data, Buy = as.numeric(avgloss < 100))
> View(iihs_data_tr)
> # Linear regression model to predict 'Buy' based on losses
> lm_model <- lm(Buy ~ Collision. + Property.damage. + Comprehensive. +
+                    Personal.injury. + Medical.payment. + Bodily.injury., data=iihs_data_tr)
```

```
> summary(lm_model)

Call:
lm(formula = Buy ~ Collision. + Property.damage. + Comprehensive. +
    Personal.injury. + Medical.payment. + Bodily.injury., data = iihs_data_tr)

Residuals:
     Min       1Q   Median       3Q      Max
-0.46425 -0.25853 -0.03087  0.23089  0.55220

Coefficients:
                  Estimate Std. Error t value
(Intercept)       2.0586715  0.4619836   4.456
Collision.       -0.0007052  0.0048424  -0.146
Property.damage. -0.0006774  0.0072098  -0.094
Comprehensive.   -0.0039661  0.0031103  -1.275
Personal.injury. -0.0047340  0.0062795  -0.754
Medical.payment. -0.0057862  0.0058844  -0.983
Bodily.injury.    0.0005263  0.0060277   0.087
                Pr(>|t|)
(Intercept)      6.84e-05 ***
Collision.          0.885
Property.damage.    0.926
```

```
Comprehensive.      0.210
Personal.injury.    0.455
Medical.payment.    0.332
Bodily.injury.      0.931
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3124 on 39 degrees of freedom
  (104 observations deleted due to missingness)
Multiple R-squared:  0.5452,  Adjusted R-squared:  0.4752
F-statistic: 7.792 on 6 and 39 DF,  p-value: 1.51e-05

> # Logistic regression model
> lg_model <- glm(Buy ~ Collision. + Property.damage. + Comprehensive. +
+                   Personal.injury. + Medical.payment. + Bodily.injury., family = "quasibinomial",
+                na.action=na.omit, control = list(maxit = 50), data=iihs_data_tr)
> summary(lg_model)

Call:
glm(formula = Buy ~ Collision. + Property.damage. + Comprehensive. +
    Personal.injury. + Medical.payment. + Bodily.injury., family = "quasibinomial",
    data = iihs_data_tr, na.action = na.omit, control = list(maxit = 50))
```

Deviance Residuals:

```
      Min          1Q      Median          3Q
-9.344e-06  -2.110e-08  -2.110e-08  -2.110e-08
      Max
 1.305e-05
```

Coefficients:

```
                  Estimate Std. Error t value
(Intercept)      741.27518    6.81383 108.790
Collision.        -0.96299    0.03903 -24.673
Property.damage.  -2.39416    0.04395 -54.470
Comprehensive.    -2.40964    0.03125 -77.099
Personal.injury.  -0.45680    0.06394  -7.144
Medical.payment.  -2.44471    0.09719 -25.154
Bodily.injury.     1.21074    0.09802  12.352
                Pr(>|t|)
(Intercept)      < 2e-16 ***
Collision.       < 2e-16 ***
Property.damage.  < 2e-16 ***
Comprehensive.    < 2e-16 ***
Personal.injury. 1.36e-08 ***
Medical.payment.  < 2e-16 ***
```

```
Bodily.injury.   4.69e-15 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 1.609654e-11)

    Null deviance: 5.0607e+01  on 45  degrees of freedom
Residual deviance: 4.6209e-10  on 39  degrees of freedom
  (104 observations deleted due to missingness)
AIC: NA

Number of Fisher Scoring iterations: 27

> lm_model2 <- lm(Buy ~ avgloss + Collision. + Property.damage. + Comprehensive. +
+                     Personal.injury. + Medical.payment. + Bodily.injury., data=iihs_data_tr)
> summary(lm_model2)

Call:
lm(formula = Buy ~ avgloss + Collision. + Property.damage. +
    Comprehensive. + Personal.injury. + Medical.payment. + Bodily.injury.,
    data = iihs_data_tr)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-0.46425 -0.25853 -0.03087  0.23089  0.55220

Coefficients: (1 not defined because of singularities)
                 Estimate Std. Error t value
(Intercept)      2.058671   0.461984   4.456
avgloss          0.003158   0.036166   0.087
Collision.      -0.001232   0.008378  -0.147
Property.damage. -0.001204   0.012346  -0.097
Comprehensive.   -0.004492   0.006063  -0.741
Personal.injury. -0.005260   0.008560  -0.615
Medical.payment. -0.006312   0.009783  -0.645
Bodily.injury.         NA         NA      NA
                Pr(>|t|)
(Intercept)     6.84e-05 ***
avgloss            0.931
Collision.         0.884
Property.damage.   0.923
Comprehensive.     0.463
Personal.injury.   0.542
Medical.payment.   0.523
Bodily.injury.        NA
```

```
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3124 on 39 degrees of freedom
  (104 observations deleted due to missingness)
Multiple R-squared:  0.5452,  Adjusted R-squared:  0.4752
F-statistic: 7.792 on 6 and 39 DF,  p-value: 1.51e-05

> lg_model2 <- glm(Buy ~ avgloss + Collision. + Property.damage. + Comprehensive. +
+                    Personal.injury. + Medical.payment. + Bodily.injury., family = "quasibinomial",
+                  na.action=na.omit, control = list(maxit = 50), data=iihs_data_tr)
> summary(lg_model2)

Call:
glm(formula = Buy ~ avgloss + Collision. + Property.damage. +
    Comprehensive. + Personal.injury. + Medical.payment. + Bodily.injury.,
    family = "quasibinomial", data = iihs_data_tr, na.action = na.omit,
    control = list(maxit = 50))

Deviance Residuals:
      Min         1Q     Median         3Q
-9.344e-06  -2.110e-08  -2.110e-08  -2.110e-08
```

```
      Max
 1.305e-05


Coefficients: (1 not defined because of singularities)
                  Estimate Std. Error t value
(Intercept)      741.27518    6.81383  108.79
avgloss            7.26441    0.58813   12.35
Collision.        -2.17372    0.07301  -29.77
Property.damage.  -3.60489    0.12488  -28.87
Comprehensive.    -3.62038    0.12364  -29.28
Personal.injury.  -1.66754    0.09367  -17.80
Medical.payment.  -3.65544    0.18450  -19.81
Bodily.injury.           NA         NA      NA
                Pr(>|t|)
(Intercept)      < 2e-16 ***
avgloss          4.69e-15 ***
Collision.       < 2e-16 ***
Property.damage.  < 2e-16 ***
Comprehensive.    < 2e-16 ***
Personal.injury.  < 2e-16 ***
Medical.payment.  < 2e-16 ***
Bodily.injury.         NA
---
```

Signif. codes:

0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 1.609654e-11)

Null deviance: 5.0607e+01 on 45 degrees of freedom

Residual deviance: 4.6209e-10 on 39 degrees of freedom

(104 observations deleted due to missingness)

AIC: NA

Number of Fisher Scoring iterations: 27

> # Simplest logistic model: using only avgloss as a predictor

> lg_model_simple <- glm(Buy ~ avgloss, family = "quasibinomial", na.action=na.omit, control = list(maxit = 50), data=iihs_data_tr)

> summary(lg_model_simple)

Call:

glm(formula = Buy ~ avgloss, family = "quasibinomial", data = iihs_data_tr,

na.action = na.omit, control = list(maxit = 50))

Deviance Residuals:

Min 1Q Median 3Q

```
-1.211e-05  -2.110e-08  -2.110e-08  -2.110e-08
      Max
 1.167e-05

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1799.71830    8.93924   201.3   <2e-16
avgloss      -18.09483    0.08983  -201.4   <2e-16

(Intercept) ***
avgloss     ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 5.177903e-12)

    Null deviance: 8.0283e+01  on 74  degrees of freedom
Residual deviance: 2.8305e-10  on 73  degrees of freedom
  (75 observations deleted due to missingness)
AIC: NA

Number of Fisher Scoring iterations: 30
```

```
> # Sorting the data by average loss
> sorted_data <- iihs_data_tr[order(iihs_data_tr$avgloss), ]
> # View the sorted data
> View(sorted_data)
```

References

1. Bevans, R. (2023, June 22). *Simple linear regression: An easy introduction & examples*. Scribbr. https://www.scribbr.com/statistics/simple-linear-regression/
2. Complete guide to key risk indicators. (n.d.-a). https://reciprocity.com/resource-center/complete-guide-to-key-risk-indicators
3. *Integrating Kris and kpis for effective technology risk management*. ISACA. (n.d.). https://www.isaca.org/resources/isaca-journal/issues/2018/volume-4/integrating-kris-and-kpis-for-effective-technology-risk-management
4. Wikimedia Foundation. (2024b, August 1). *P-value*. Wikipedia. https://en.wikipedia.org/wiki/P-value
5. *What is Principal Component Analysis (PCA)?*. IBM. (2023, December 4). https://www.ibm.com/topics/principal-component-analysis