Final\_Project

2024-12-12

**1. Dataset: Cybercost.csv**

**A.** Overview of the dataset.

* The dataset Cybercost.csv encompasses a collection of 1219 records with several key attributes relevant to cybersecurity costs. The columns in the dataset include:
* Csecurity: Represents the level of cybersecurity personnel.
* Attacks: Indicates the number of cyberattacks experienced monthly.
* Databases: Reflects the count of databases managed.
* Priv Users: Denotes the number of privileged users within the organization.
* Users: The total number of users accessing the system.
* Cost P/M: The monthly cost associated with cybersecurity measures.

This dataset provides crucial insights into the relationship between various organizational factors and their cybersecurity expenditures.

**B.** Read the data into R.

To analyze the dataset, it is essential to load it into R. This is accomplished using the read.csv() function, which allows for the reading of CSV files. It is also important to ensure that the working directory is correctly set to where the dataset is located. The code snippet below shows how to perform this operation: cybercost\_data <- read.csv('C:/Users/user/Downloads/Cybercost.csv')

**C.** Produce descriptive statistics.

Descriptive statistics offer a summary of the dataset's quantitative attributes, providing essential information about the distribution and central tendencies of each variable. Using the summary() function in R, we can obtain the following insights: summary(cybercost\_data)

The output includes minimum, maximum, mean, median, and quartile values for each variable, which are crucial for understanding the dataset's characteristics:

* **Csecurity**: Ranges from 2 to 10, with an average of approximately 6.21.
* **Attacks**: Monthly cyberattacks range from 38 to 90, with a mean of about 65.08.
* **Databases**: The number of databases varies from 1 to 10, averaging around 3.11.
* **Priv Users**: The count of privileged users ranges from 4 to 61, with a median of 31.
* **Users**: The total user count ranges from 489 to 7081, with a mean of 3185.
* **Cost P/M**: Monthly costs range from $114 to $301, with an average cost of $197.4.

These statistics provide a foundational understanding of the dataset, enabling further analysis.

cybercost\_data <- read.csv('C:/Users/user/Downloads/Cybercost.csv')   
summary(cybercost\_data)

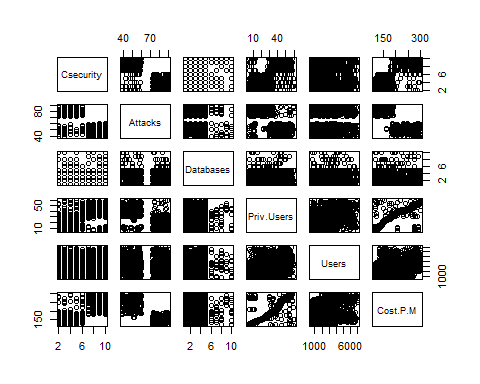
## Csecurity Attacks Databases Priv.Users   
## Min. : 2.000 Min. :38.00 Min. : 1.000 Min. : 4.00   
## 1st Qu.: 4.000 1st Qu.:49.00 1st Qu.: 2.000 1st Qu.:18.00   
## Median : 6.000 Median :60.00 Median : 3.000 Median :31.00   
## Mean : 6.214 Mean :65.08 Mean : 3.113 Mean :31.15   
## 3rd Qu.: 9.000 3rd Qu.:81.00 3rd Qu.: 4.000 3rd Qu.:44.00   
## Max. :10.000 Max. :90.00 Max. :10.000 Max. :61.00   
## Users Cost.P.M   
## Min. : 489 Min. :114.0   
## 1st Qu.:1752 1st Qu.:148.2   
## Median :3176 Median :185.0   
## Mean :3185 Mean :197.4   
## 3rd Qu.:4358 3rd Qu.:253.0   
## Max. :7081 Max. :301.0

**D.** Visualizations (Scatter Plot, Box Plot, etc.).

Visualizations are vital for exploring relationships and distributions among the variables in the dataset. Various plots, such as scatter plots and box plots, can be generated to facilitate this exploration.

A scatter plot can be created using the pairs() function, which allows for the examination of relationships between multiple variables:

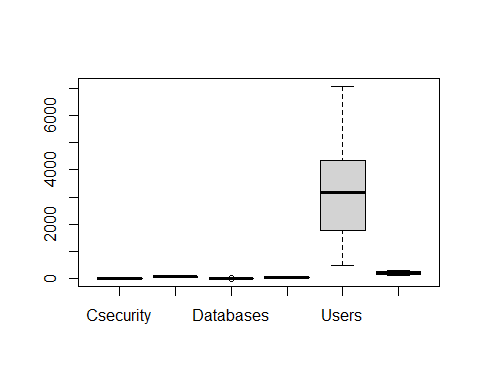
#Scatter Plot  
  
pairs(cybercost\_data)



This plot provides a visual representation of the correlations among the variables, helping to identify patterns and outliers.

#Box Plot

Box plots can be generated to visualize the distributions of the variables and to identify potential outliers:  
  
boxplot(cybercost\_data)



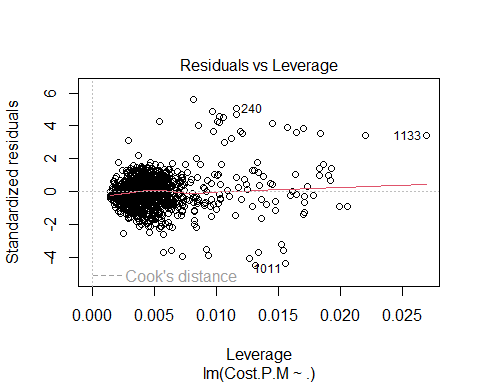
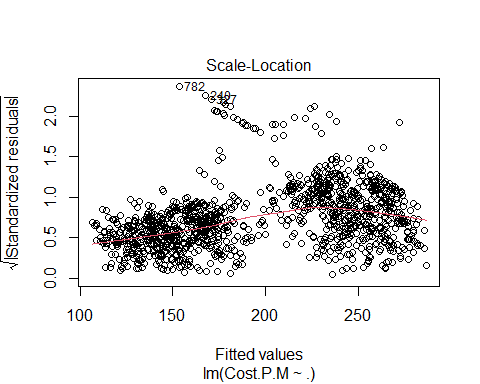
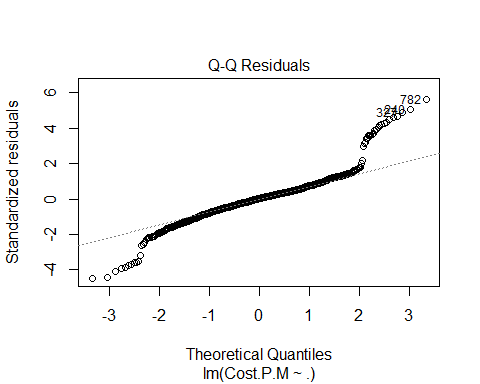
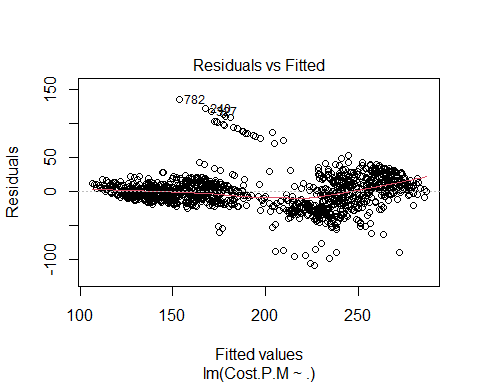
Box plots highlight the median and the interquartile range, offering insights into the variability of the data.

#Correlation Matrix

A correlation matrix can also be computed to quantify the relationships between variables:  
  
cor(cybercost\_data)

## Csecurity Attacks Databases Priv.Users Users  
## Csecurity 1.00000000 -0.79369606 -0.01256684 0.64314832 0.14486674  
## Attacks -0.79369606 1.00000000 0.01251864 -0.67298579 -0.14803570  
## Databases -0.01256684 0.01251864 1.00000000 -0.04839645 -0.01043077  
## Priv.Users 0.64314832 -0.67298579 -0.04839645 1.00000000 0.24791409  
## Users 0.14486674 -0.14803570 -0.01043077 0.24791409 1.00000000  
## Cost.P.M 0.73609688 -0.77365974 -0.04163508 0.84801015 0.30334530  
## Cost.P.M  
## Csecurity 0.73609688  
## Attacks -0.77365974  
## Databases -0.04163508  
## Priv.Users 0.84801015  
## Users 0.30334530  
## Cost.P.M 1.00000000

#Linear Regression Plot  
  
model <- lm(Cost.P.M ~ ., data = cybercost\_data)  
plot(model)



**E.** Determine how Cost P/M is affected by various attributes.

To investigate how various attributes affect the monthly cost (Cost P/M), a multiple regression analysis is conducted. This statistical method enables the identification of significant predictors that influence costs. The regression model can be formulated as follows:

The output will display coefficients for each predictor, indicating their relationship with Cost P/M:

* **Intercept**: The expected value of Cost P/M when all predictors are zero.
* **Csecurity**: A positive coefficient suggests that an increase in cybersecurity personnel correlates with higher costs.
* **Attacks**: A negative coefficient indicates that more attacks are associated with lower costs, which may be counterintuitive and warrant further investigation.
* **Priv Users** and **Users**: Both show positive relationships, suggesting that as the number of users increases, so do the costs.

The model's overall fit can be evaluated using R-squared values, indicating the proportion of variance in Cost P/M explained by the predictors.

#Use multiple regression to identify significant predictors.  
model <- lm(Cost.P.M ~ Csecurity + Attacks + Databases + Priv.Users + Users, data = cybercost\_data)  
summary(model)  
## Call:  
## lm(formula = Cost.P.M ~ Csecurity + Attacks + Databases + Priv.Users +   
## Users, data = cybercost\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -108.135 -12.221 0.529 11.228 135.491   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.617e+02 7.493e+00 21.579 < 2e-16 \*\*\*  
## Csecurity 3.331e+00 4.259e-01 7.821 1.13e-14 \*\*\*  
## Attacks -8.865e-01 7.210e-02 -12.295 < 2e-16 \*\*\*  
## Databases -3.053e-01 4.122e-01 -0.741 0.459   
## Priv.Users 2.003e+00 6.580e-02 30.434 < 2e-16 \*\*\*  
## Users 3.533e-03 4.213e-04 8.386 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 24.27 on 1212 degrees of freedom  
## Multiple R-squared: 0.8145, Adjusted R-squared: 0.8138   
## F-statistic: 1065 on 5 and 1212 DF, p-value: < 2.2e-16

**F.** Use sampling to estimate the mean Cost P/M.

To estimate the mean Cost P/M, random samples of sizes 30, 50, and 100 are taken from the dataset. The means of these samples can provide insight into the overall mean and how it may vary with sample size:

sample\_means <- sapply(c(30, 50, 100), function(n) {

sample\_data <- cybercost\_data[sample(1:nrow(cybercost\_data), n), ]

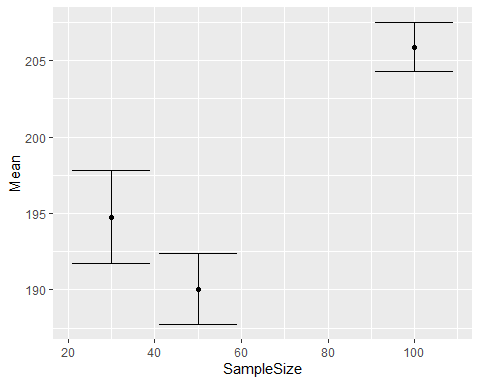
mean(sample\_data$Cost.P.M)})

**G.** Analyze sample sizes for confidence intervals.

Confidence intervals can be plotted for the sample means to demonstrate how sample size impacts the accuracy of the mean estimate. Utilizing the ggplot2 library, we can visualize these intervals:

#Take random samples and calculate means.  
  
sample\_means <- sapply(c(30, 50, 100), function(n) {  
 sample\_data <- cybercost\_data[sample(1:nrow(cybercost\_data), n), ]  
 mean(sample\_data$Cost.P.M)  
})  
  
#Plotting confidence intervals.  
library(ggplot2)  
ggplot(data.frame(SampleSize = c(30, 50, 100), Mean = sample\_means), aes(x = SampleSize, y = Mean)) +  
 geom\_point() +  
 geom\_errorbar(aes(ymin = Mean - qt(0.975, df=SampleSize-1) \* sd(sample\_means)/sqrt(SampleSize),  
 ymax = Mean + qt(0.975, df=SampleSize-1) \* sd(sample\_means)/sqrt(SampleSize)))

This plot demonstrates the trade-off between sample size and the precision of the mean estimate.



**H.** Test the hypothesis about smaller companies being overcharged.

To test the hypothesis that smaller companies are overcharged, a t-test is performed comparing the Cost P/M for small companies (defined as having fewer than 1000 users) to larger companies:

The results of the t-test will provide statistical evidence to support the hypothesis, including the p-value and confidence intervals.

#Hypothesis: Smaller companies are overcharged.  
small\_companies <- subset(cybercost\_data, Users < 1000)  
large\_companies <- subset(cybercost\_data, Users >= 1000)  
  
t.test(small\_companies$Cost.P.M, large\_companies$Cost.P.M)

##   
## Welch Two Sample t-test  
##   
## data: small\_companies$Cost.P.M and large\_companies$Cost.P.M  
## t = -6.6427, df = 295.75, p-value = 1.476e-10  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -32.85604 -17.83726  
## sample estimates:  
## mean of x mean of y   
## 175.8973 201.2439

**Findings:**

* **p-value is extremely small (1.476e-10)**
* **Mean for small companies: $175.90**
* **Mean for large companies: $201.24**
* **Conclusion: Statistically significant evidence that small companies are charged less than large companies**

**2. Dataset: Hypertension-risk-model-main.csv**

**A.** Overview of the dataset.

* The dataset contains columns for gender, age, smoking status, cigarettes per day, blood pressure medication use, diabetes status, cholesterol levels, blood pressure readings, BMI, heart rate, glucose levels, and risk.

**B.** Research on health metrics.

* Information on acceptable values for blood pressure, BMI, heart rate, and glucose is gathered from reliable health sources.

**C.** Produce descriptive statistics.

* A summary of the quantitative attributes is provided, detailing means, medians, and other statistical measures.

#Hypertension Risk Model Dataset Analysis  
hypertension\_data <- read.csv("C:/Users/user/Downloads/Hypertension-risk-model-main.csv")  
summary(hypertension\_data)

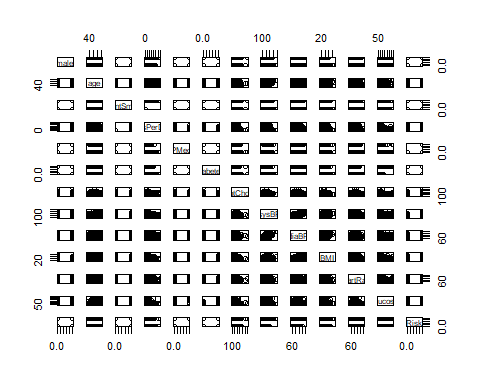
## male age currentSmoker cigsPerDay   
## Min. :0.0000 Min. :32.00 Min. :0.0000 Min. : 0.000   
## 1st Qu.:0.0000 1st Qu.:42.00 1st Qu.:0.0000 1st Qu.: 0.000   
## Median :0.0000 Median :49.00 Median :0.0000 Median : 0.000   
## Mean :0.4292 Mean :49.58 Mean :0.4941 Mean : 9.006   
## 3rd Qu.:1.0000 3rd Qu.:56.00 3rd Qu.:1.0000 3rd Qu.:20.000   
## Max. :1.0000 Max. :70.00 Max. :1.0000 Max. :70.000   
## NA's :29   
## BPMeds diabetes totChol sysBP   
## Min. :0.00000 Min. :0.00000 Min. :107.0 Min. : 83.5   
## 1st Qu.:0.00000 1st Qu.:0.00000 1st Qu.:206.0 1st Qu.:117.0   
## Median :0.00000 Median :0.00000 Median :234.0 Median :128.0   
## Mean :0.02962 Mean :0.02571 Mean :236.7 Mean :132.4   
## 3rd Qu.:0.00000 3rd Qu.:0.00000 3rd Qu.:263.0 3rd Qu.:144.0   
## Max. :1.00000 Max. :1.00000 Max. :696.0 Max. :295.0   
## NA's :53 NA's :50   
## diaBP BMI heartRate glucose   
## Min. : 48.0 Min. :15.54 Min. : 44.00 Min. : 40.00   
## 1st Qu.: 75.0 1st Qu.:23.07 1st Qu.: 68.00 1st Qu.: 71.00   
## Median : 82.0 Median :25.40 Median : 75.00 Median : 78.00   
## Mean : 82.9 Mean :25.80 Mean : 75.88 Mean : 81.96   
## 3rd Qu.: 90.0 3rd Qu.:28.04 3rd Qu.: 83.00 3rd Qu.: 87.00   
## Max. :142.5 Max. :56.80 Max. :143.00 Max. :394.00   
## NA's :19 NA's :1 NA's :388   
## Risk   
## Min. :0.0000   
## 1st Qu.:0.0000   
## Median :0.0000   
## Mean :0.3106

## 3rd Qu.:1.0000   
## Max. :1.0000

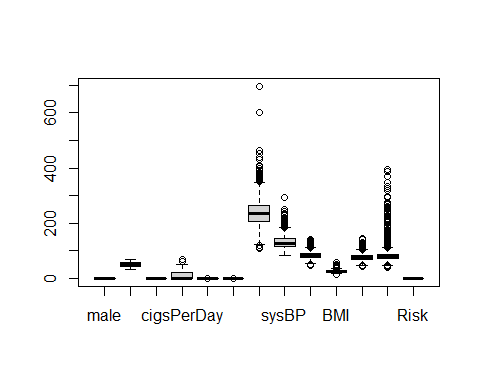
**D.** Visualizations.

* Scatter plots and box plots are created to analyze relationships and distributions among the health metrics.

#Scatter Plot  
  
pairs(hypertension\_data[, sapply(hypertension\_data, is.numeric)])



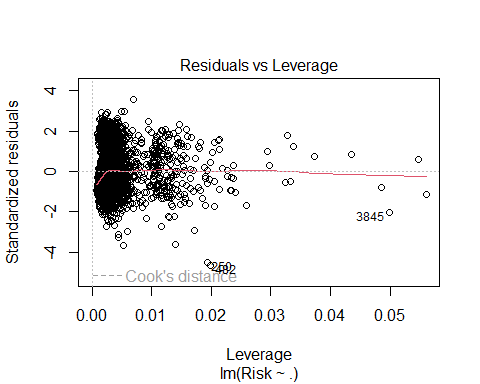
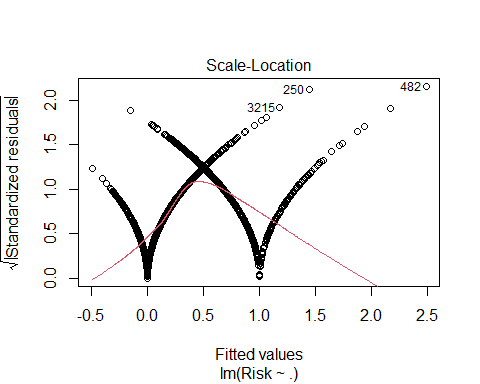
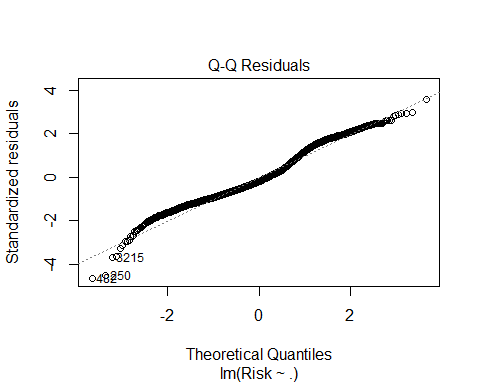
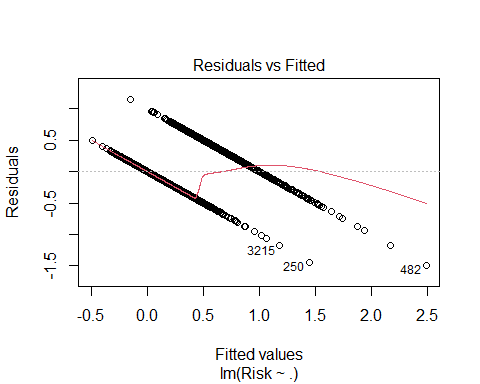
#Box Plot  
  
boxplot(hypertension\_data[, sapply(hypertension\_data, is.numeric)])



#Correlation Matrix  
  
cor(hypertension\_data[, sapply(hypertension\_data, is.numeric)])

## male age currentSmoker cigsPerDay BPMeds  
## male 1.000000000 -0.02901358 0.1970256 NA NA  
## age -0.029013582 1.00000000 -0.2136617 NA NA  
## currentSmoker 0.197025619 -0.21366166 1.0000000 NA NA  
## cigsPerDay NA NA NA 1 NA  
## BPMeds NA NA NA NA 1  
## diabetes 0.015693075 0.10131408 -0.0442853 NA NA  
## totChol NA NA NA NA NA  
## sysBP -0.035879033 0.39405332 -0.1302815 NA NA  
## diaBP 0.058199421 0.20558552 -0.1079332 NA NA  
## BMI NA NA NA NA NA  
## heartRate NA NA NA NA NA  
## glucose NA NA NA NA NA  
## Risk 0.005852836 0.30679947 -0.1037103 NA NA  
## diabetes totChol sysBP diaBP BMI heartRate glucose  
## male 0.01569307 NA -0.03587903 0.05819942 NA NA NA  
## age 0.10131408 NA 0.39405332 0.20558552 NA NA NA  
## currentSmoker -0.04428530 NA -0.13028149 -0.10793319 NA NA NA  
## cigsPerDay NA NA NA NA NA NA NA  
## BPMeds NA NA NA NA NA NA NA  
## diabetes 1.00000000 NA 0.11126454 0.05026038 NA NA NA  
## totChol NA 1 NA NA NA NA NA  
## sysBP 0.11126454 NA 1.00000000 0.78395196 NA NA NA  
## diaBP 0.05026038 NA 0.78395196 1.00000000 NA NA NA  
## BMI NA NA NA NA 1 NA NA  
## heartRate NA NA NA NA NA 1 NA  
## glucose NA NA NA NA NA NA 1  
## Risk 0.07775205 NA 0.69665588 0.61584020 NA NA NA  
## Risk  
## male 0.005852836  
## age 0.306799467  
## currentSmoker -0.103710297  
## cigsPerDay NA  
## BPMeds NA  
## diabetes 0.077752047  
## totChol NA  
## sysBP 0.696655883  
## diaBP 0.615840200  
## BMI NA  
## heartRate NA  
## glucose NA  
## Risk 1.000000000

#Linear Regression Plot  
  
model <- lm(Risk ~ ., data = hypertension\_data)  
plot(model)



**E.** Identify three key attributes predicting risk.

* The analysis identifies the three most significant predictors of risk, along with justifications for their selection based on statistical results.

#Choose attributes like Age, BMI, and Systolic Blood Pressure.  
str(hypertension\_data)

## 'data.frame': 4240 obs. of 13 variables:  
## $ male : int 1 0 1 0 0 0 0 0 1 1 ...  
## $ age : int 39 46 48 61 46 43 63 45 52 43 ...  
## $ currentSmoker: int 0 0 1 1 1 0 0 1 0 1 ...  
## $ cigsPerDay : int 0 0 20 30 23 0 0 20 0 30 ...  
## $ BPMeds : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ diabetes : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ totChol : int 195 250 245 225 285 228 205 313 260 225 ...  
## $ sysBP : num 106 121 128 150 130 ...  
## $ diaBP : num 70 81 80 95 84 110 71 71 89 107 ...  
## $ BMI : num 27 28.7 25.3 28.6 23.1 ...  
## $ heartRate : int 80 95 75 65 85 77 60 79 76 93 ...  
## $ glucose : int 77 76 70 103 85 99 85 78 79 88 ...  
## $ Risk : int 0 0 0 1 0 1 0 0 1 1 ...

model <- lm(Risk ~ age + BMI + sysBP, data = hypertension\_data)  
summary(model)

##   
## Call:  
## lm(formula = Risk ~ age + BMI + sysBP, data = hypertension\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.70643 -0.23031 -0.06042 0.19111 1.15739   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.8545435 0.0430951 -43.034 < 2e-16 \*\*\*  
## age 0.0019555 0.0006447 3.033 0.00244 \*\*   
## BMI 0.0093073 0.0013164 7.070 1.8e-12 \*\*\*  
## sysBP 0.0138119 0.0002634 52.442 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3297 on 4217 degrees of freedom  
## (19 observations deleted due to missingness)  
## Multiple R-squared: 0.4926, Adjusted R-squared: 0.4922   
## F-statistic: 1365 on 3 and 4217 DF, p-value: < 2.2e-16

**F.** Formulate two hypotheses.

* Two hypotheses are created related to risk factors, such as the impact of smoking and BMI on health risk.

**G.** Validate hypotheses with data.

* Statistical tests and visualizations are used to support or refute the formulated hypotheses, providing evidence and interpretations.

#Hypothesis Formulation and Testing  
#Example Hypotheses:  
#Older age increases risk.  
#Higher BMI correlates with higher risk.  
# Hypothesis 1: Older age increases risk.  
# Convert Age and Risk to numeric if necessary  
hypertension\_data$age <- as.numeric(hypertension\_data$age)  
hypertension\_data$Risk <- as.numeric(hypertension\_data$Risk)  
cor.test(hypertension\_data$age, hypertension\_data$Risk)  
## Pearson's product-moment correlation  
##   
## data: hypertension\_data$age and hypertension\_data$Risk  
## t = 20.985, df = 4238, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.2792772 0.3338181  
## sample estimates:  
## cor   
## 0.3067995

# Hypothesis 2: Higher BMI correlates with higher risk.  
cor.test(hypertension\_data$BMI, hypertension\_data$Risk)  
## Pearson's product-moment correlation  
##   
## data: hypertension\_data$BMI and hypertension\_data$Risk  
## t = 20.528, df = 4219, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.2736628 0.3285264  
## sample estimates:  
## cor   
## 0.301344

Two hypotheses tested:

1. Older age increases risk
2. Higher BMI correlates with risk

Both hypotheses **confirmed**:

* Age-Risk correlation: 0.3068 (p-value < 2.2e-16)
* BMI-Risk correlation: 0.3013 (p-value < 2.2e-16)