**CS601 C Final Project**

**Name:Diksha Korat**

**UID:-U01923523**

Complete the exercises and answer the questions outlined below. Your submission must be a  
combination of what the problem statement is you are working on, # R comments, R code, and  
screen captures of your plots, and use bold print to indicate your findings. Your code must run  
successfully and you must answer all the questions to get credit. I.e. The objective is to make it  
look as closely as possible to real statistics paper / data mining analysis. Do not just give me an  
RMD file. I want either a Word, HTML, or PDF document with all the R code, graphs, etc.  
imbedded as you would see in a real statistics paper.  
Rubik: I have labeled this a combination of final and project. This course because it is fully  
online doesn’t really provide an appropriate environment for standard exams or projects. This  
activity is going to be about 15% of your grade. Just because it has simple requests like  
“numerical summary of the variables / quantitative columns” doesn’t mean you identify one  
column and give me the mean, median, standard deviation, etc. It means looks at the columns  
carefully and indicate which columns should be used for your study and why. If you just do the  
minimum you will end up with a grade like a B. If you want an A each question should be  
answered in detail. The better your discovery and justification aspects the better the grade.  
1. This exercise relates to the dataset “Cybercost.csv”  
The dataset has only a few columns.  
1. Csecurity  
2. Attacks  
3. Databases  
4. Priv Users  
5. Users  
6. Cost P/M  
A. Total number of Cybersecurity Personnel 2) Number of monthly cyberattacks 3)  
Databases 4) Number of Privilege Users 5) Total number of Users 6) Cost P/M

a. Just for your information this is not real data collected from customers. There  
are 1219 records in the dataset which is the full customer base.

B. Read the data into R. Make sure that you have the directory set to the correct location  
for the data or use file.choose().

Answer: Csecurity,Attacks,Databases,Priv Users,Users,Cost P/M

6,74,4,12,3356,132

10,43,4,45,2415,263

3,81,2,16,5899,145

9,50,4,34,3005,196

2,80,5,9,1164,131

5,76,3,10,2935,129

5,72,4,12,1678,131

6,88,2,27,4761,161

5,77,3,31,2149,184

10,42,3,40,574,225

6,90,2,31,1239,178

3,83,1,8,955,121

10,43,5,34,4194,186

8,59,2,39,4376,206

4,86,5,30,2931,179

4,80,3,21,3087,156

4,78,4,11,3637,135

4,76,1,39,3414,186

10,47,2,50,6691,282

10,55,4,36,6058,193

10,55,2,55,3581,295

5,75,1,11,6688,136

5,75,2,15,1739,142

10,57,1,47,4894,262

5,73,1,10,1611,120

7,56,1,48,3278,275

9,43,4,11,2430,207

10,45,2,43,3717,243

8,58,1,43,5496,250

8,53,3,48,845,262

3,83,1,19,1752,147

4,77,4,13,898,137

7,57,3,57,4749,299

10,58,1,54,4361,288

9,39,3,47,3155,274

4,81,2,27,5069,177

9,57,5,38,846,205

5,83,2,33,3955,186

8,40,2,42,3247,234

3,81,3,13,855,133

4,84,1,32,1647,185

2,86,5,12,4361,137

8,56,4,46,3254,264

4,87,2,34,3998,189

7,57,1,58,5073,298

5,73,2,11,687,131

2,89,4,8,943,118

5,87,4,30,1614,179

10,44,1,46,4745,258

4,79,2,23,600,161

9,48,4,47,6770,267

5,76,2,10,2209,124

10,42,2,42,2365,226

5,86,2,8,510,125

10,47,5,53,5364,287

3,82,4,6,2643,117

5,76,1,13,3188,137

2,72,1,20,3972,154

8,52,5,50,4934,283

10,50,5,45,2043,244

3,80,2,21,2254,157

10,45,2,49,5694,271

9,55,4,54,4576,287

10,57,1,32,2819,195

3,86,3,14,921,138

10,52,1,40,2321,223

7,40,2,41,775,235

9,60,5,42,3838,232

9,59,5,57,3744,298

2,39,5,30,3117,181

10,46,2,41,804,242

2,73,2,23,6539,156

10,55,5,35,3163,196

2,89,3,21,1692,156

10,56,5,35,4360,195

7,45,2,49,2814,264

8,52,2,56,1897,294

8,43,3,45,4119,248

10,43,1,49,4019,273

6,75,4,23,5720,159

3,83,5,6,1358,121

8,48,5,38,3779,209

9,60,3,44,3599,248

4,85,3,26,1319,168

2,89,5,9,2955,116

3,75,1,15,1794,142

6,88,5,29,3414,183

5,83,1,30,799,184

10,41,8,48,5699,274

4,77,5,11,1291,133

8,48,5,39,3127,207

2,81,3,28,6535,172

8,47,4,44,4099,238

5,89,4,28,4007,180

10,46,2,52,4095,288

10,49,3,38,1754,207

5,75,1,13,2752,138

2,80,4,13,4590,134

9,47,1,40,2206,224

4,85,3,25,5374,168

9,59,3,39,4294,230

9,48,4,36,6348,203

4,81,5,16,2871,142

10,55,1,49,4458,278

4,72,2,7,3556,123

4,76,5,13,2052,130

10,53,1,55,2084,292

8,54,4,46,6289,256

10,57,2,44,3866,256

5,73,4,13,5552,134

4,75,3,14,3039,140

5,75,2,28,2039,180

3,87,5,32,5299,189

3,80,3,29,523,178

3,81,2,27,6490,175

10,53,2,32,802,182

9,56,4,48,5569,274

7,59,5,39,5715,208

9,49,4,10,4177,207

3,74,10,15,3981,143

7,44,4,46,5417,264

2,52,5,25,4127,163

10,53,4,8,4002,290

5,73,4,32,877,188

7,46,3,46,5409,266

7,42,2,44,723,256

2,86,4,9,1617,120

5,76,6,17,1984,142

10,54,3,38,4266,206

7,43,2,39,3494,228

5,75,5,28,6154,178

5,75,4,29,4468,178

8,56,2,55,5167,300

10,60,4,55,3933,288

5,78,3,12,965,130

3,87,1,9,2536,129

10,56,10,35,4307,197

4,85,2,16,1862,146

10,50,4,38,2666,219

10,50,6,37,3211,197

3,75,3,17,2248,145

7,53,5,39,1363,224

9,58,3,44,2332,251

10,40,5,49,4168,275

3,74,1,57,1756,118

3,75,3,18,3302,145

9,51,1,53,3783,285

4,76,3,5,2676,117

3,90,3,28,4396,172

4,74,3,6,2105,120

5,86,1,43,4172,173

10,43,4,42,4522,225

4,87,1,31,2699,181

4,77,3,6,2203,116

4,90,3,20,2734,154

7,56,2,56,5083,289

9,48,4,44,3217,238

2,87,2,12,770,130

4,81,3,22,5502,154

3,82,2,9,2098,123

2,90,2,7,2527,117

10,40,4,45,2550,264

9,48,1,35,608,204

6,43,4,48,4659,274

4,87,5,20,4347,156

2,46,4,54,5399,287

10,52,5,41,5390,225

4,86,3,27,2821,179

10,49,3,44,1951,253

5,50,2,59,885,220

10,40,1,50,4900,278

6,87,2,12,3396,125

10,55,5,55,4043,298

10,53,1,38,2891,202

4,73,3,17,2577,148

3,50,5,42,969,245

2,76,5,16,2421,143

4,87,3,28,5204,180

4,78,5,61,3250,148

3,78,8,30,3389,184

7,53,10,43,2385,224

7,51,2,55,4833,294

4,76,4,5,2510,115

2,79,1,13,691,134

10,48,2,53,2763,279

2,82,4,18,1417,148

2,74,5,21,837,155

8,46,1,30,580,185

5,75,2,10,1414,131

5,88,4,19,668,152

7,59,1,55,4135,291

4,86,1,35,3258,188

10,45,3,47,3969,270

10,57,3,38,908,201

10,54,1,37,637,204

10,38,1,37,4530,200

8,47,3,55,5372,287

9,40,2,48,3898,273

8,57,3,53,3160,289

4,83,5,7,2414,123

10,45,4,48,4777,272

8,41,5,52,4418,284

10,48,4,50,5211,279

10,40,1,52,6514,287

10,56,3,51,4139,280

10,47,1,53,4104,285

10,43,4,33,3239,189

9,39,1,6,4447,277

9,49,4,37,3665,204

9,58,1,47,3158,262

3,88,4,31,5950,186

4,78,1,25,3107,162

3,78,1,10,3045,123

6,87,4,30,2043,182

5,85,2,25,4773,159

3,76,5,6,1465,115

4,89,6,10,1453,131

3,74,3,21,606,148

9,50,5,34,3789,196

4,83,4,6,1947,117

8,44,2,44,2803,248

5,90,4,14,1688,134

3,52,5,26,1455,164

5,76,4,16,5686,142

3,55,1,46,7049,248

5,76,1,33,6118,186

10,58,4,47,5039,275

3,82,2,16,3333,147

7,56,3,55,3604,288

10,56,5,53,3474,294

2,80,4,8,3758,123

5,74,4,11,2687,129

3,82,1,6,804,114

3,82,1,28,2058,175

5,76,1,33,3679,185

3,86,6,26,3388,170

4,78,2,25,6414,162

5,86,2,51,2740,176

5,90,5,28,4447,181

9,53,2,5,3805,290

5,84,4,23,5572,159

10,48,5,32,4790,185

8,43,2,48,1910,261

10,45,3,33,5685,244

2,81,1,10,2623,125

9,56,5,42,996,259

7,59,3,56,4137,293

C. Produce a descriptive statistical summary of quantitative attributes in the data set.

Answer:- Csecurity Attacks Databases

Min. : 2.000 Min. :38.00 Min. : 1.000

1st Qu.: 4.000 1st Qu.:49.00 1st Qu.: 2.000

Median : 6.000 Median :60.00 Median : 3.000

Mean : 6.214 Mean :65.08 Mean : 3.113

3rd Qu.: 9.000 3rd Qu.:81.00 3rd Qu.: 4.000

Max. :10.000 Max. :90.00 Max. :10.000

Priv.Users Users Cost.P.M

Min. : 4.00 Min. : 489 Min. :114.0

1st Qu.:18.00 1st Qu.:1752 1st Qu.:148.2

Median :31.00 Median :3176 Median :185.0

Mean :31.15 Mean :3185 Mean :197.4

3rd Qu.:44.00 3rd Qu.:4358 3rd Qu.:253.0

Max. :61.00 Max. :7081 Max. :301.0

D. Scatter Plot, Box Plot, correlate, and linear and multiple linear regression plot all the quantitative attributes in the data set.

Answer:-A screenshot of a scatter plot matrix

Description automatically generated

A box plot with text overlay

Description automatically generated

A diagram of a data analysis

Description automatically generated with medium confidence

A graph of a linear regression plot

Description automatically generated

E. Now comes the interesting part of the study. Your job is determining how is cost P/M  
determined. You are to use all your skills that you have learned during this course to  
determine what attribute(s) are best. Note: one variable (attribute) may not be enough  
to determine how the cost is determined.

1. Answer:- Exploratory Data Analysis: Conduct a thorough exploratory data analysis to understand the distribution, range, and relationships between the "Cost P/M" attribute and other quantitative attributes. This can involve calculating summary statistics, creating visualizations such as scatter plots and box plots, and identifying any patterns or trends.
2. Correlation Analysis: Calculate the correlation coefficients between "Cost P/M" and all other quantitative attributes in the dataset. Identify attributes with high positive or negative correlations with "Cost P/M". These attributes may have a strong relationship with the cost.
3. Multiple Linear Regression: Perform a multiple linear regression analysis with "Cost P/M" as the dependent variable and all other quantitative attributes as independent variables. This will help identify the attributes that have a significant impact on the cost when considered together.
4. Feature Selection Techniques: Utilize feature selection techniques such as stepwise regression, forward selection, or backward elimination to identify the subset of attributes that best predict the "Cost P/M". These techniques help identify the most relevant attributes while minimizing redundancy.
5. Domain Knowledge: Consider any domain-specific knowledge or expertise that may provide insights into the factors influencing the cost. For example, if you have knowledge about the cybersecurity industry, you may have an idea of which attributes are likely to impact the cost.
6. Machine Learning Algorithms: Apply machine learning algorithms such as decision trees, random forests, or gradient boosting to identify the most important attributes for predicting the "Cost P/M". These algorithms can provide insights into the relative importance of different attributes.

F. This dataset is the complete customer list. This means what you determined above is  
the real mean. Now we want to use your other skills you learned during this semester  
(i.e. sampling, inference, confidence intervals) to figure out the mean without using the  
entire customer list.

Answer:- To estimate the mean of the "Cost P/M" attribute without using the entire customer list, we can utilize sampling, inference, and confidence intervals. Here's an approach you can follow:

Random Sampling: Take multiple random samples from the customer list. The sample size will depend on the level of precision and confidence desired. For example, you can start with a sample size of 30, 50, 100, and so on.

Calculate Sample Mean: For each sample, calculate the mean of the "Cost P/M" attribute. This will give you an estimate of the mean based on each sample.

Central Limit Theorem: As the sample size increases, the distribution of sample means will approach a normal distribution, regardless of the shape of the population distribution. This is known as the Central Limit Theorem.

Confidence Interval: Calculate the confidence interval for the mean using the sample means. The confidence interval provides a range of values within which the true population mean is likely to fall. The level of confidence is typically set at 95% (or any desired level).

Interpretation: Interpret the confidence interval. It will give you an estimate of the range within which the true population mean is likely to lie. The narrower the interval, the more precise the estimate.

G. Take multiple random samples of the customer list. First try with 30, 50, 100, etc. till  
you can determine how big a sample you need so that with 95% confidence level that  
you are able to determine a sample means that are similar to the real mean. Plot the  
sample means, confidence levels, real means, etc. on the same plots as they did in your  
textbook for each random sample.

Answer: To determine the sample size needed to estimate the mean of the "Cost P/M" attribute with a 95% confidence level, we can take multiple random samples from the customer list and calculate the sample means. Here's an approach you can follow:

Set the Confidence Level: Define the desired confidence level, which in this case is 95%.

Random Sampling: Take random samples from the customer list with different sample sizes. Start with a sample size of 30, then increase it to 50, 100, and so on.

Calculate Sample Means: For each sample, calculate the mean of the "Cost P/M" attribute. This will give you an estimate of the mean for each sample.

Calculate Confidence Intervals: Calculate the confidence interval for each sample mean using the appropriate formula. The confidence interval provides a range within which the true population mean is likely to fall.

Plot the Results: Create a plot to visualize the sample means, confidence intervals, and the real mean. You can use a line plot or a bar plot to display the sample means, and error bars to represent the confidence intervals. Include the real mean as a reference point on the plot.

Determine Optimal Sample Size: Examine the plot and observe how the sample means and confidence intervals change with different sample sizes. Look for a sample size where the confidence intervals are relatively narrow and include the real mean within their range. This indicates that the sample mean is similar to the real mean with a 95% confidence level.

By following this approach and plotting the results, you can determine the sample size needed to estimate the mean of the "Cost P/M" attribute with a 95% confidence level. Remember to document your findings and include the plots in your analysis.

H. Prove or disprove the following hypothesis “Companies that are smaller (<1000) are  
being overcharged (i.e. number of employees))”. Note: hypothesis many times are  
proven incorrect. See if you can make up another hypothesis that you can prove or  
disprove. In this case since you have the entire customer list you can actually verify if  
the hypothesis calculations done with the sample are correct.

Answer:- To prove or disprove the hypothesis that "Companies that are smaller (<1000) are being overcharged (i.e. number of employees)," we can analyze the data and compare the cost per month (Cost P/M) for companies with fewer than 1000 employees to those with 1000 or more employees. Here's an approach you can follow:

Data Preparation: Divide the dataset into two groups based on the number of employees. One group should include companies with fewer than 1000 employees, and the other group should include companies with 1000 or more employees.

Descriptive Analysis: Calculate the mean cost per month (Cost P/M) for each group separately. Compare the means to see if there is a noticeable difference between the two groups.

Hypothesis Testing: Perform a hypothesis test to determine if the difference in means between the two groups is statistically significant. You can use a t-test or another appropriate statistical test for this comparison. The null hypothesis would be that there is no significant difference in the cost per month between the two groups.

Interpretation: Based on the results of the hypothesis test, you can either accept or reject the null hypothesis. If the p-value is less than the chosen significance level (e.g., 0.05), you can conclude that there is a significant difference in the cost per month between companies with fewer than 1000 employees and those with 1000 or more employees.

Additionally, you can make up another hypothesis to test using the same approach. For example, you could hypothesize that "Companies in a specific industry have a higher cost per month compared to companies in other industries." Follow a similar process of data preparation, descriptive analysis, hypothesis testing, and interpretation to prove or disprove this hypothesis.

Remember to document your findings, including the hypothesis, analysis steps, and results, in your analysis.

I. Obviously, the idea here is to show off what you have learned during the semester. So,  
the more the project looks like a real paper that a “statistician/data analyst” would  
make the better the grade.

Answer:- Absolutely! To make your project look like a real paper that a statistician or data analyst would produce, here are some key elements to consider:

Title and Abstract: Start with a clear and concise title that reflects the focus of your study. Follow it with an abstract that provides a brief overview of the project, including the objectives, methods, and key findings.

Introduction: Begin with an introduction that provides background information on the dataset and the problem you are addressing. Clearly state the research question or objective of your analysis.

Literature Review: Review relevant literature and studies related to your research question. Discuss previous findings, theories, and methodologies that are applicable to your analysis. This demonstrates your understanding of the existing knowledge in the field.

Methodology: Describe the methodology and techniques used in your analysis. Explain the data collection process, any preprocessing steps, and the statistical methods employed. Justify your choices and provide references to established methodologies.

Data Description: Provide a detailed description of the dataset, including the variables, their definitions, and any transformations or cleaning performed. Include summary statistics, such as means, medians, and standard deviations, for the quantitative attributes.

Data Analysis: Present the results of your analysis using appropriate statistical techniques. Include visualizations, such as scatter plots, box plots, histograms, and correlation matrices, to illustrate relationships and patterns in the data. Use clear and concise language to explain your findings.

Hypothesis Testing: Clearly state the hypotheses you are testing and describe the statistical tests used to evaluate them. Present the results of the hypothesis tests, including test statistics, p-values, and any relevant confidence intervals. Interpret the findings and discuss their implications.

1. Discussion: Interpret and discuss the results in the context of your research question. Address any limitations or potential sources of bias in your study. Compare your findings to previous research and offer insights or explanations for any discrepancies.
2. Conclusion: Summarize the key findings of your analysis and their implications. Discuss the practical and theoretical implications of your results. Highlight any recommendations for future research or areas for further investigation.
3. References: Include a comprehensive list of references for all sources cited in your project. Follow a consistent citation style, such as APA or MLA, and ensure that all references are properly formatted.
4. Formatting and Presentation: Use a professional and consistent formatting style throughout your document. Pay attention to grammar, spelling, and punctuation. Include section headings, subheadings, and a table of contents to improve readability.

2. This exercise relates to the dataset “Hypertension-risk-model-main.csv”

The dataset has a few columns.  
1. male 1 / female 0  
2. Age  
3. Current Smoker yes 1 / no 0  
4. Cigarettes per day  
5. Takes Blood Pressure Pills yes 1 / no 0  
6. Has diabetes yes 1 / no 0  
7. Total Cholesterol  
8. Systolic Blood Pressure Reading  
9. Diastolic Blood Pressure Reading  
10. Body Mass Index (BMI)

1. Resting Heart Rate  
   12. Glucose  
   13. Risk (What you are trying to predict based on the above information)

A. Read the data into R. Make sure that you have the directory set to the correct location  
for the data or use file.choose().

Answer:- To determine what are considered good values for blood pressure, BMI (Body Mass Index), heart rate, and glucose, you can refer to reputable sources such as medical websites, research papers, or guidelines from health organizations. Here are some commonly accepted ranges for these measurements:

Blood Pressure:

Normal blood pressure is typically defined as systolic pressure below 120 mmHg and diastolic pressure below 80 mmHg.

Prehypertension: Systolic pressure between 120-139 mmHg or diastolic pressure between 80-89 mmHg.

Hypertension (high blood pressure): Systolic pressure of 140 mmHg or higher, or diastolic pressure of 90 mmHg or higher.

BMI (Body Mass Index):

Underweight: BMI below 18.5

Normal weight: BMI between 18.5 and 24.9

Overweight: BMI between 25 and 29.9

Obesity: BMI of 30 or higher

Heart Rate:

Resting heart rate for adults is typically between 60 and 100 beats per minute (bpm).

Athletes or individuals who are physically fit may have lower resting heart rates, often below 60 bpm.

Glucose:

Fasting blood glucose levels are commonly used to assess diabetes risk.

Normal fasting blood glucose levels are typically below 100 mg/dL (5.6 mmol/L).

Prediabetes: Fasting blood glucose levels between 100-125 mg/dL (5.6-6.9 mmol/L).

Diabetes: Fasting blood glucose levels of 126 mg/dL (7.0 mmol/L) or higher.

C. Produce a descriptive statistical summary of quantitative attributes in the data set.

Answer:- male age currentSmoker

Min. :0.0000 Min. :32.00 Min. :0.0000

1st Qu.:0.0000 1st Qu.:42.00 1st Qu.:0.0000

Median :0.0000 Median :49.00 Median :0.0000

Mean :0.4292 Mean :49.58 Mean :0.4941

3rd Qu.:1.0000 3rd Qu.:56.00 3rd Qu.:1.0000

Max. :1.0000 Max. :70.00 Max. :1.0000

cigsPerDay BPMeds diabetes

Min. : 0.000 Min. :0.00000 Min. :0.00000

1st Qu.: 0.000 1st Qu.:0.00000 1st Qu.:0.00000

Median : 0.000 Median :0.00000 Median :0.00000

Mean : 9.006 Mean :0.02962 Mean :0.02571

3rd Qu.:20.000 3rd Qu.:0.00000 3rd Qu.:0.00000

Max. :70.000 Max. :1.00000 Max. :1.00000

NA's :29 NA's :53

totChol sysBP diaBP

Min. :107.0 Min. : 83.5 Min. : 48.0

1st Qu.:206.0 1st Qu.:117.0 1st Qu.: 75.0

Median :234.0 Median :128.0 Median : 82.0

Mean :236.7 Mean :132.4 Mean : 82.9

3rd Qu.:263.0 3rd Qu.:144.0 3rd Qu.: 90.0

Max. :696.0 Max. :295.0 Max. :142.5

NA's :50

BMI heartRate glucose

Min. :15.54 Min. : 44.00 Min. : 40.00

1st Qu.:23.07 1st Qu.: 68.00 1st Qu.: 71.00

Median :25.40 Median : 75.00 Median : 78.00

Mean :25.80 Mean : 75.88 Mean : 81.96

3rd Qu.:28.04 3rd Qu.: 83.00 3rd Qu.: 87.00

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. Scatter Plot, Box Plot, correlate, and linear and multiple linear regression plot all the  
quantitative attributes in the data set.

Answer:- A diagram of a scatter plot matrix

Description automatically generated

A diagram of a box plot

Description automatically generated

**A graph of different types of diabetes

Description automatically generated with medium confidence**

A rectangular graph with black lines

Description automatically generated

E. This dataset is easier than the Cybercost example above to find attributes that will  
predict risk. Clearly you will find many attributes that will predict risk. The idea here is  
if you could choose three attributes which three attributes would you choose that give  
you the best prediction of risk and in what order. Example: (male, age, smoker). Why  
did you choose those three (i.e. justify your answer tables, graphs)

Answer:- To identify the three attributes that give the best prediction of risk in the dataset, we can use various techniques such as exploratory data analysis, correlation analysis, and regression modeling. Here's an approach to determine the three attributes and their order of importance:

Exploratory Data Analysis: Begin by conducting exploratory data analysis to understand the relationship between each attribute and the target variable (risk). This can involve creating visualizations such as scatter plots, box plots, and histograms to identify any patterns or trends.

Correlation Analysis: Calculate the correlation coefficients between each attribute and the risk variable. Identify attributes with high positive or negative correlations with risk. These attributes may have a strong relationship with the target variable.

Regression Modeling: Perform regression analysis, such as linear regression or logistic regression, to assess the predictive power of each attribute individually. Evaluate the significance of each attribute's coefficient and its impact on the prediction of risk.

Feature Selection Techniques: Utilize feature selection techniques, such as stepwise regression or LASSO regression, to identify the subset of attributes that provide the best prediction of risk. These techniques help identify the most relevant attributes while minimizing redundancy.

Model Evaluation: Assess the performance of different models using metrics such as R-squared, AIC, BIC, or accuracy, depending on the type of regression model used. Compare the models with different combinations of attributes to determine the three attributes that yield the best prediction of risk.

Justification: Justify the selection of the three attributes based on their statistical significance, correlation with risk, and their individual predictive power. Provide tables, graphs, and statistical summaries to support your justification.

F. Make up two hypotheses about the above data. Example: If a person smokes, he is at  
risk. (don’t use this one)

Answer:- Certainly! Here are two hypotheses that you can test using the given dataset:

Hypothesis 1: Individuals with higher body mass index (BMI) are at a higher risk of hypertension.

Justification: BMI is commonly associated with cardiovascular health, and higher BMI values have been linked to an increased risk of hypertension in previous research.

Hypothesis 2: Individuals with higher resting heart rates have a higher risk of hypertension.

Justification: Resting heart rate is an indicator of cardiovascular fitness and can be influenced by factors such as physical activity and overall health. Previous studies have suggested that higher resting heart rates may be associated with an increased risk of hypertension.

G. Prove or disprove your two hypotheses. (i.e. tables and graphs)

Answer:- To prove or disprove the two hypotheses, we can perform statistical analysis and examine tables and graphs to evaluate the relationship between the variables. Here's an approach to test the hypotheses and present the results:

Hypothesis 1: Individuals with higher body mass index (BMI) are at a higher risk of hypertension.

Statistical Analysis: Perform a logistic regression analysis with hypertension as the dependent variable and BMI as the independent variable. Assess the significance of the BMI coefficient to determine if there is a relationship between BMI and hypertension.

Tables and Graphs: Present a table summarizing the logistic regression results, including the coefficient, p-value, and odds ratio for BMI. Additionally, create a graph, such as a bar plot or box plot, to compare the distribution of BMI between individuals with and without hypertension.

Interpretation: Based on the statistical analysis and visualizations, discuss whether the results support or refute the hypothesis. If the coefficient for BMI is significant and the odds ratio indicates an increased risk of hypertension with higher BMI values, it would support the hypothesis.

Hypothesis 2: Individuals with higher resting heart rates have a higher risk of hypertension.

Statistical Analysis: Perform a logistic regression analysis with hypertension as the dependent variable and resting heart rate as the independent variable. Evaluate the significance of the resting heart rate coefficient to determine if there is a relationship between heart rate and hypertension.

Tables and Graphs: Present a table summarizing the logistic regression results, including the coefficient, p-value, and odds ratio for resting heart rate. Additionally, create a graph, such as a box plot or scatter plot, to compare the distribution of resting heart rate between individuals with and without hypertension.

Interpretation: Analyze the statistical analysis and visualizations to determine if the results support or refute the hypothesis. If the coefficient for resting heart rate is significant and the odds ratio indicates an increased risk of hypertension with higher heart rates, it would support the hypothesis.

H. Obviously, the idea here is to show off what you have learned during the semester. So,  
the more the project looks like a real paper that a “statistician/data analyst” would  
make the better the grade.

Answer:- Absolutely! To make your project look like a real paper that a statistician or data analyst would produce, here are some key elements to consider:

Title and Abstract: Start with a clear and concise title that reflects the focus of your study. Follow it with an abstract that provides a brief overview of the project, including the objectives, methods, and key findings.

Introduction: Begin with an introduction that provides background information on the research question or problem you are addressing. Clearly state the objectives of your analysis and the hypotheses you are testing.

Literature Review: Review relevant literature and studies related to your research question. Discuss previous findings, theories, and methodologies that are applicable to your analysis. This demonstrates your understanding of the existing knowledge in the field.

Methodology: Describe the methodology and techniques used in your analysis. Explain the data collection process, any preprocessing steps, and the statistical methods employed. Justify your choices and provide references to established methodologies.

Data Description: Provide a detailed description of the dataset, including the variables, their definitions, and any transformations or cleaning performed. Include summary statistics, such as means, medians, and standard deviations, for the quantitative attributes.

Data Analysis: Present the results of your analysis using appropriate statistical techniques. Include visualizations, such as scatter plots, box plots, histograms, and correlation matrices, to illustrate relationships and patterns in the data. Use clear and concise language to explain your findings.

Hypothesis Testing: Clearly state the hypotheses you are testing and describe the statistical tests used to evaluate them. Present the results of the hypothesis tests, including test statistics, p-values, and any relevant confidence intervals. Interpret the findings and discuss their implications.

Discussion: Interpret and discuss the results in the context of your research question. Address any limitations or potential sources of bias in your study. Compare your findings to previous research and offer insights or explanations for any discrepancies.

Conclusion: Summarize the key findings of your analysis and their implications. Discuss the practical and theoretical implications of your results. Highlight any recommendations for future research or areas for further investigation.

References: Include a comprehensive list of references for all sources cited in your project. Follow a consistent citation style, such as APA or MLA, and ensure that all references are properly formatted.

Formatting and Presentation: Use a professional and consistent formatting style throughout your document. Pay attention to grammar, spelling, and punctuation. Include section headings, subheadings, and a table of contents to improve