Discovering Most Prevalent Factors in Dengue Disease Outbreak.

Dallin Johnson

Department of Biostatistics

University of Kansas

December 22, 2022

Summary / Abstract

In this statistical analysis, a dataset exploring a dengue outbreak was explored. The purpose of this study was to create a multiple logistic regression equation that could best predict if an individual was with or without dengue. In order to predict this dependent variable, multiple independent variables were tested, analyzed, and implemented into a series of equations.

Data on the dengue outbreak in Mexico was collected from a single study containing 196 participants. Various aspects of each of these participants lives were collected as well, including: identification number, age, socio-economic status, which sector they lived in, and if they owned a savings account. These dependent variables were then analyzed further to see how well they could predict the disease outcome.

From the p-values that were calculated from this dataset, it was concluded that the p-values of age in years and the sector in which the people lived were statistically significant. In other words, age in years and the sector were reliable independent variables in calculating the dependent variable of whether an individual had the disease. Another conclusion that was drawn from the data was that the diseased individuals had an older median aged that those without the disease and that the disease was more prominent in sector two than one.

Introduction

In 1988, a study on a dengue epidemic/outbreak that was occurring on the pacific coast of Mexico took place[1]. That study included 196 people[1] that were selected from two distinct sectors of the same city. Multiple categorical variables were collected on each of the participants of the study in order to gather more insight of which factors could be of great influence whether a citizen of this city had dengue or not.

The six variables that were collected for said study were: id, age in years, socio-economic status, sector, savings account, and disease[1]. ID was a number that was correlated to each of the individuals of the study and was not considered as an independent variable that could be used in the proposed logistic regression equation. Age of the participants was recorded. Socio-economic status placed each individual in one of the three groups: upper class, middle class, and lower class. Sector recorded whether the individual was from sector one or sector two from the city. Data on whether the citizen had a savings account was recorded. Finally, whether or not the individual had dengue is the dependent variable that this statistical analysis was attempting to calculate.

As stated previously, the main focus of this statistical analysis was to see if it was possible to predict if an individual had dengue or not based on the independent variables that were collected on each individual. The most important factor that was assessed in this study was the reliability of each individual independent variable as an indicator to whether or not it could be used to estimated/ predict the outcome of the disease variable.

As disease by group was assessed, this study was able to see which groups were effected the most by the disease. An assessment of the number of individuals in each group based on variable was then viewed in order to understand which groups had a fair amount of representation and which groups did not. After this information was quantified and visualized using different models, a final equation for the main question for this study was then created. The final equation for the study was calculated with all of the independent variables in use. After concluding that some of the variables would not be the most accurate to use, a then more accurate, reduced model was produced in order to ensure that the outcome would lead to the best prediction for whether or not a citizen in this city would have dengue.

Methods

Age Models and Figures

In order to see the differences and interpret the data accurately, simple models that displayed the differences between diseased individuals and individuals without disease were created. The first step was to analyze the age predictor's relationship with the disease.

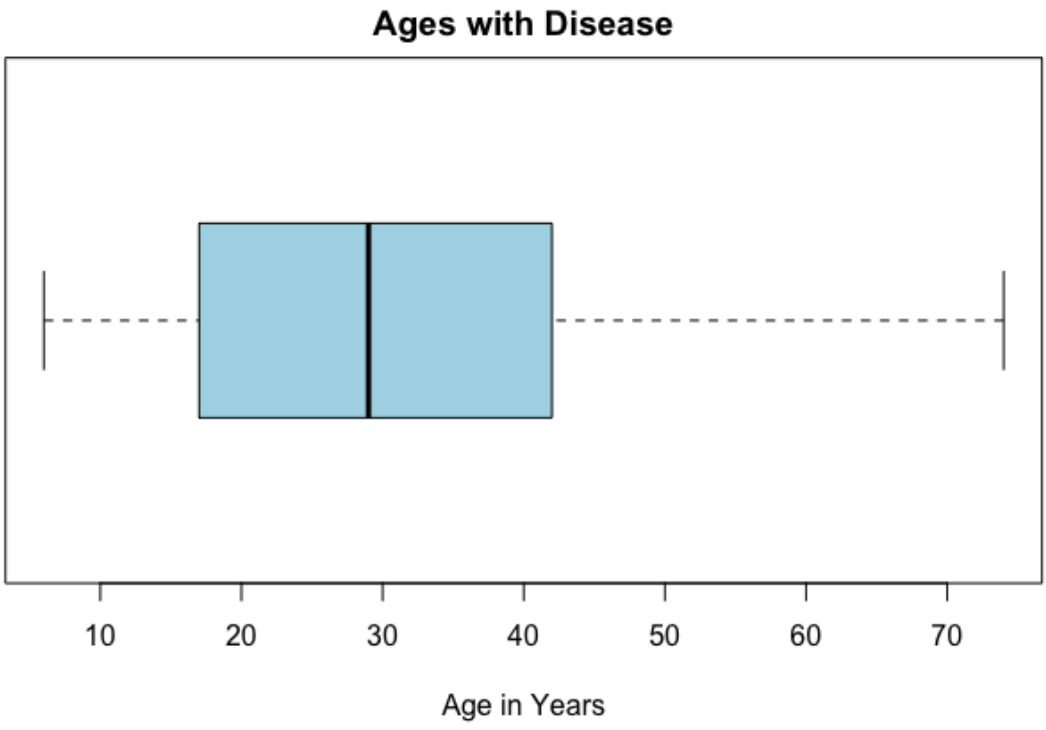

Figure 1: Ages of Diseased Citizens

As shown by the Figure 1 and calculated for this analysis, the median age of the participants that had dengue was 29, with the first quantile at 17 years of age and the third quantile at 42 years of age.

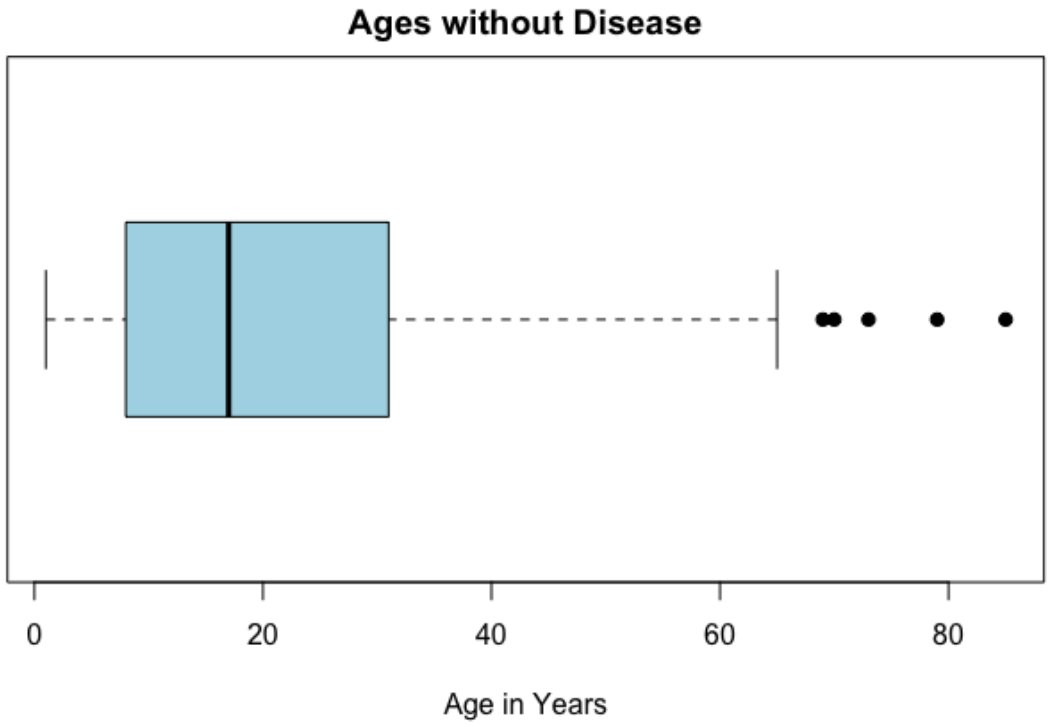

Figure 2: Ages of Citizens without Dengue

Figure 2 illustrates a different story than Figure 1. The first and third quartile, along with the median age was lower for citizens without the disease compared to those who did. The median age of individuals without the disease was 17, while the ages of the first and third quartile were 8 and 31 respectively.

An analysis was then carried out to see the proportion of diseased individuals and people without disease as it related to each of the predictor variables. For example, with the sector variable, two models were created for the figure. The first model was a model that visualized the number of diseased individuals in each sector, while the second model was one that depicted the number of people without disease in each sector.

Number of Surveyed Individuals Models and Figures

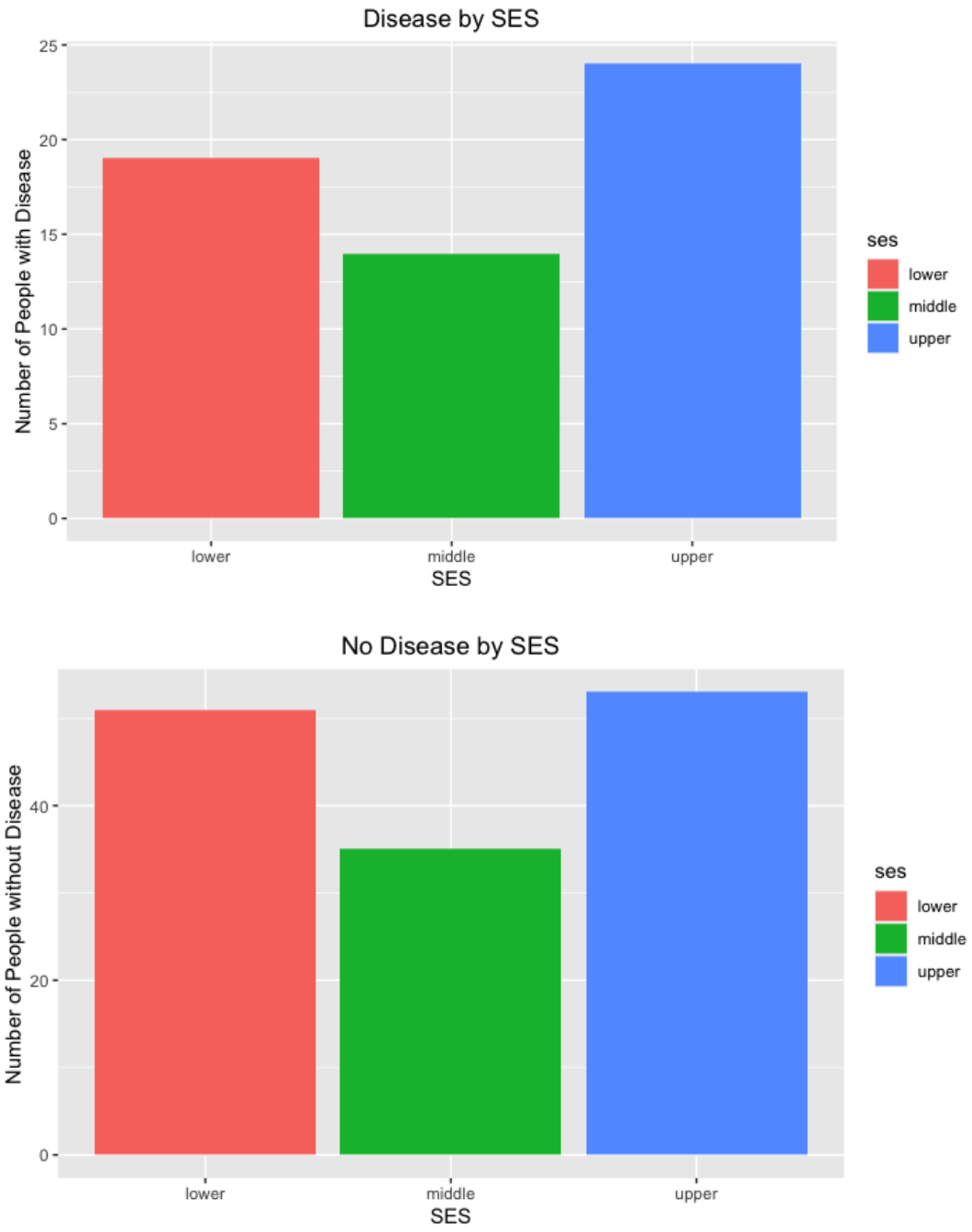

Figure 3: Citizens by Socio-Economic Status

Nothing very abnormal was noticed about the number of citizens with and without disease in each socio-economic class.

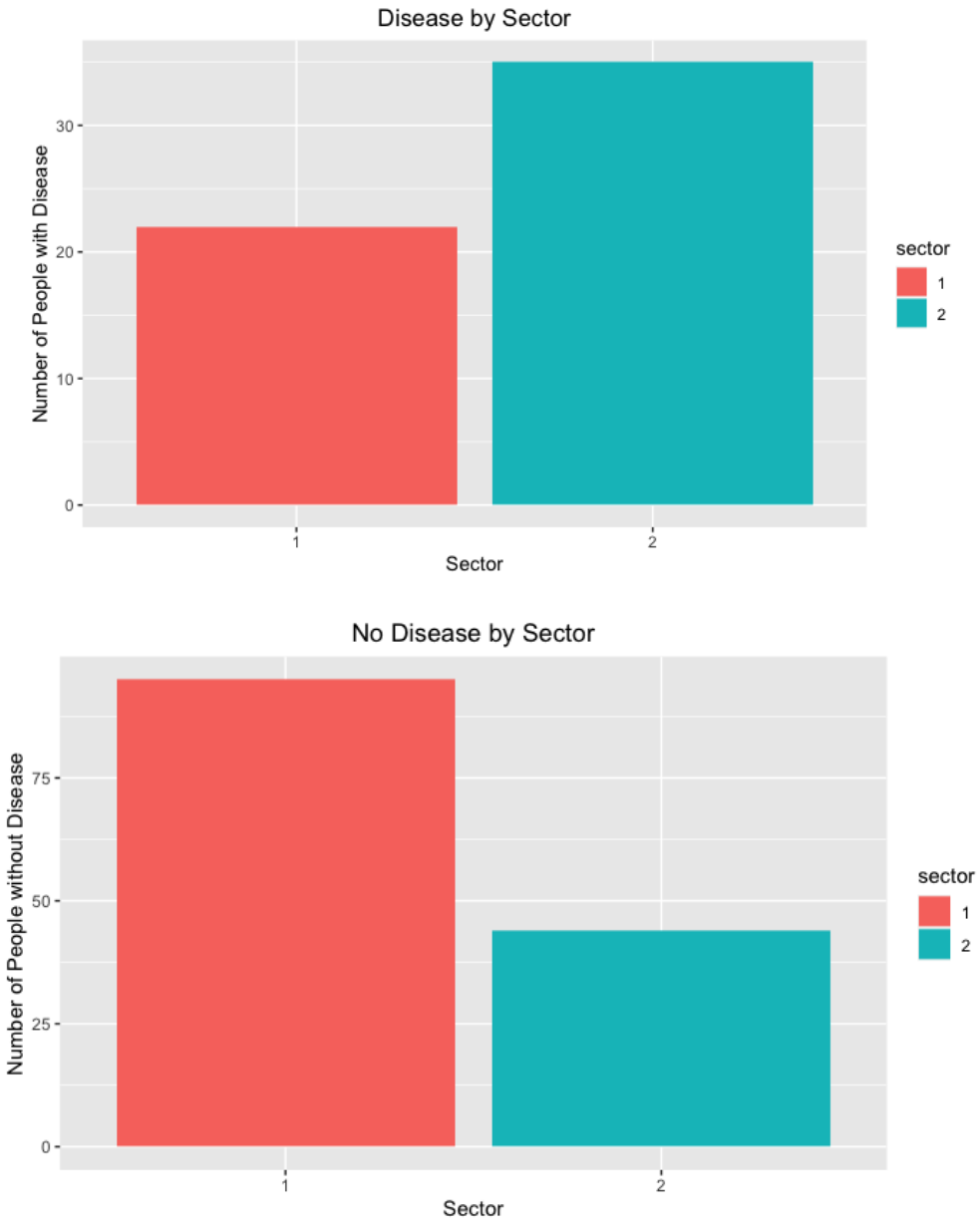

Figure 4: Disease by Sector

The number of citizens that were surveyed from both sectors were fairly close, one thing that was interesting about this bar plot was the difference in how much more the dengue outbreak had affected sector 2 over sector 1.

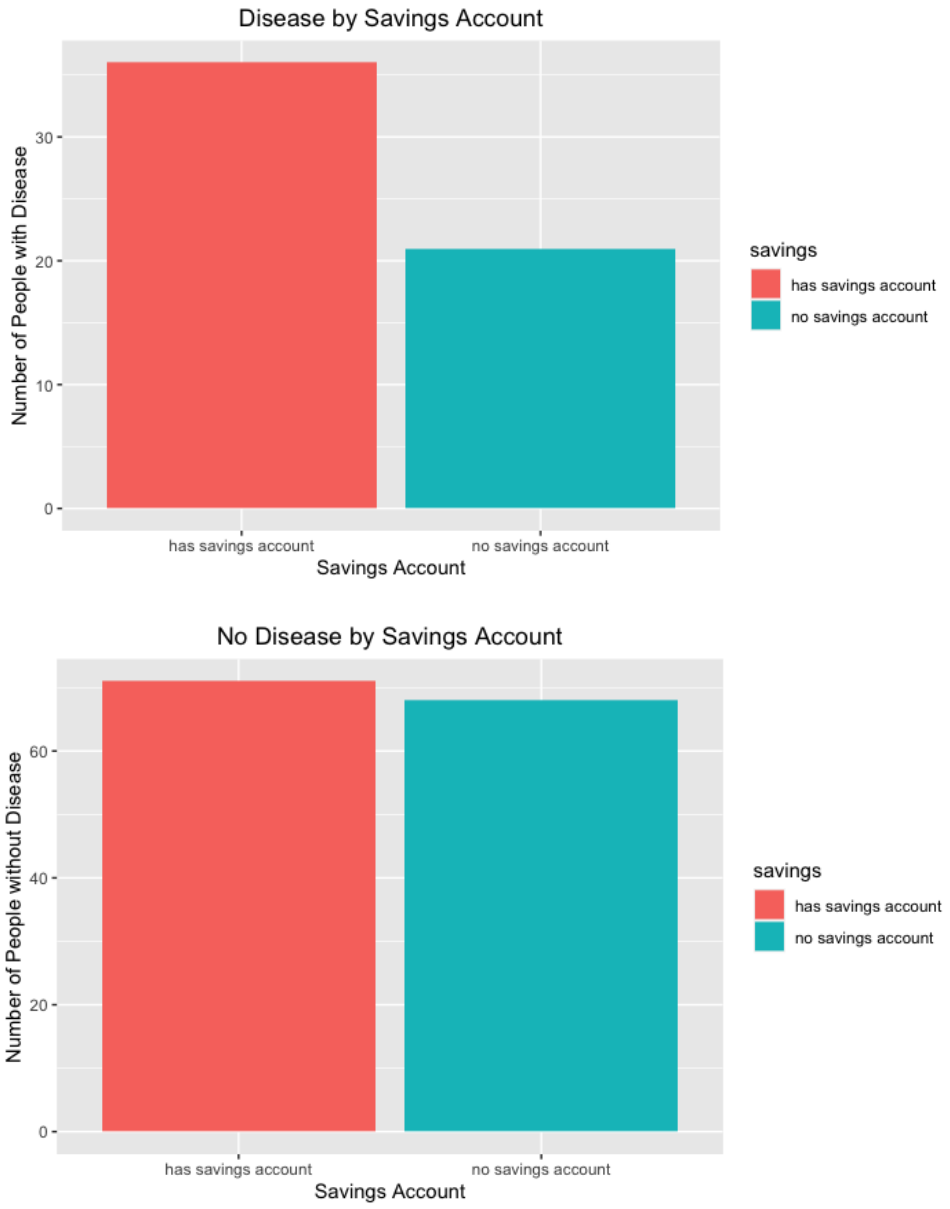

Figure 5: Disease by Savings Account

Nothing out of the ordinary was observed for the two models that comprise Figure 5. Only aspect that was noted was the similar number of people without disease who had a savings account compared to those who did not have a savings account.

Next part of the analysis dealt with the number of people in each group of the variables. For example, comparing the number of individuals that were surveyed from sector 1 versus those that were surveyed from sector 2. This part of the analysis was completed in order to identify any group that may have been over or underrepresented.

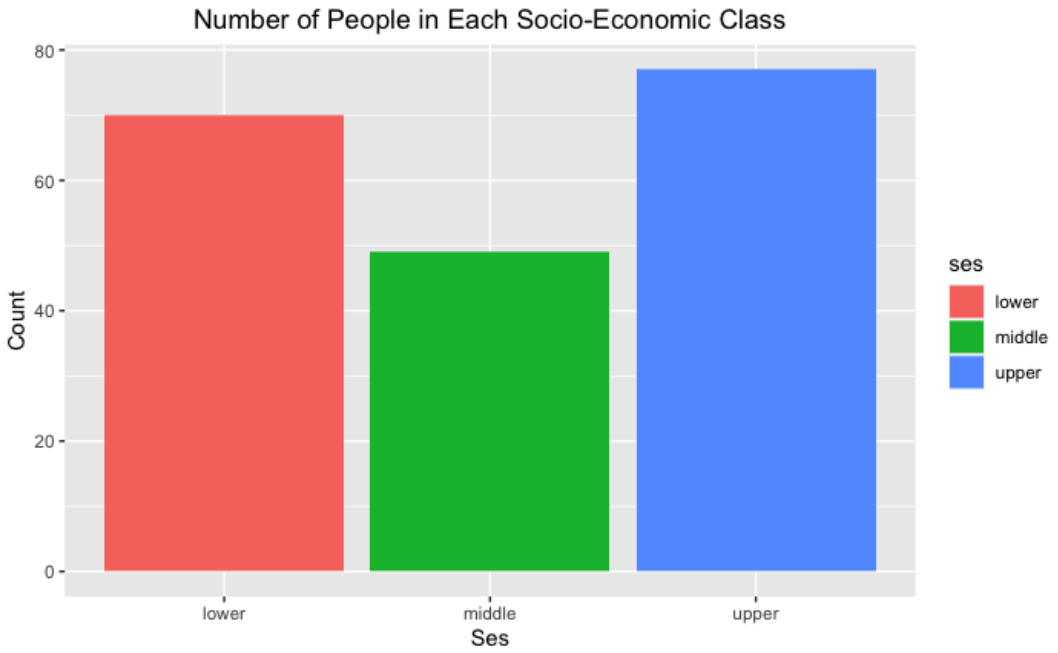

Figure 6: People Surveyed by Socio-Economic Status

Nothing too unusual was observed with Figure 6 data. However, the middle class status was definitely underrepresented within this study.

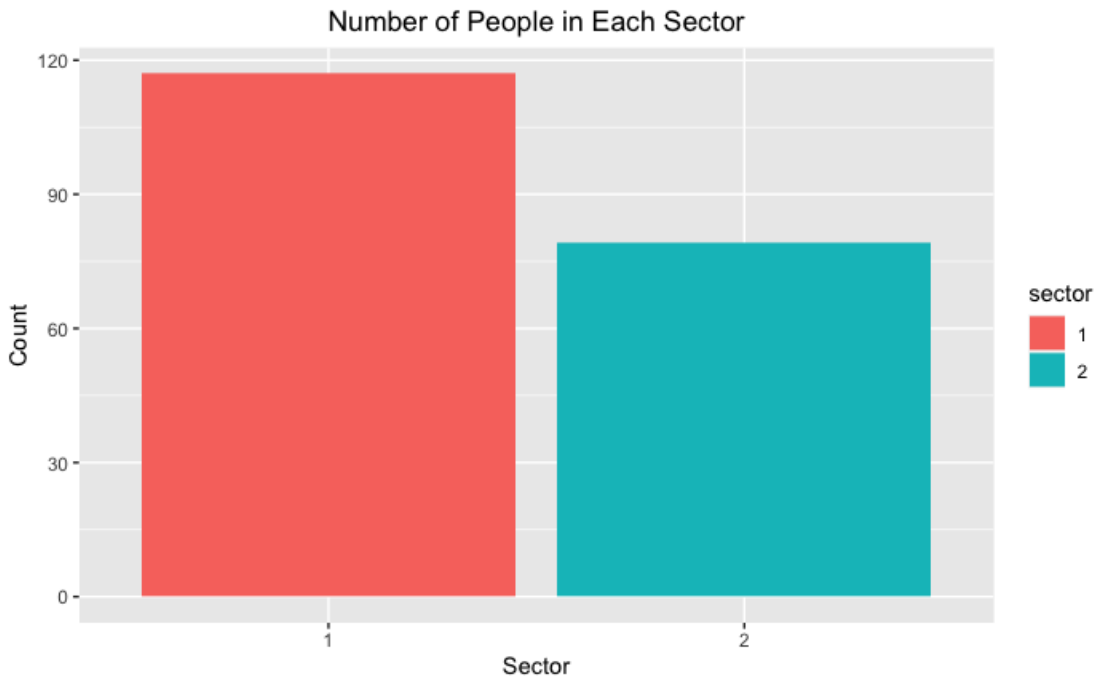

Figure 7: People Surveyed by Sector

There were definitely more citizens that were being surveyed for the study that were from sector 1 than from sector 2.

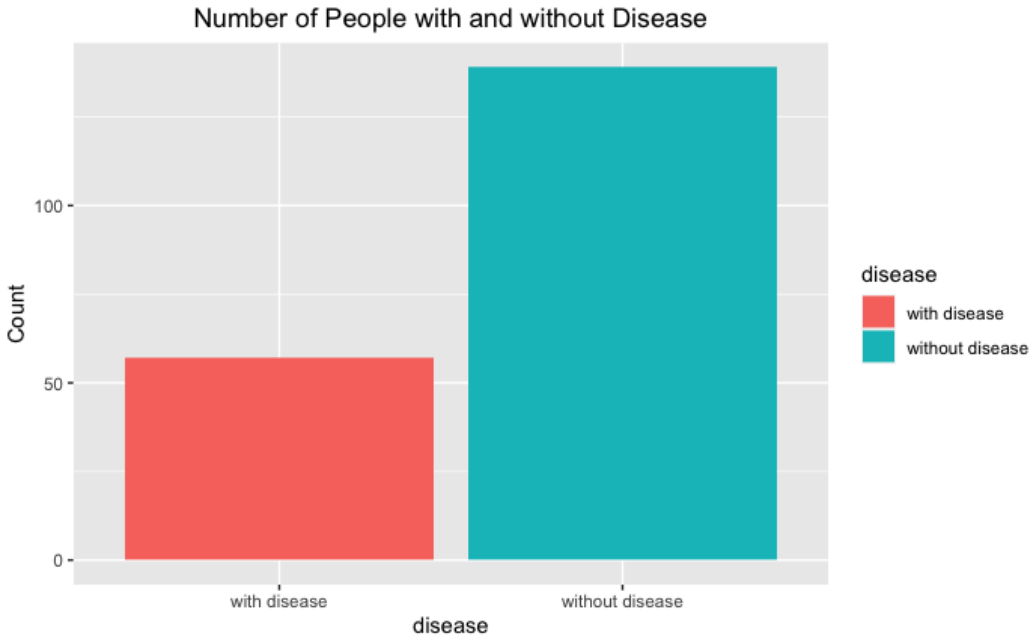

Figure 8: People Surveyed by Disease

Figure 8 demonstratest that there were far more people that were without dengue that those that did have the disease.

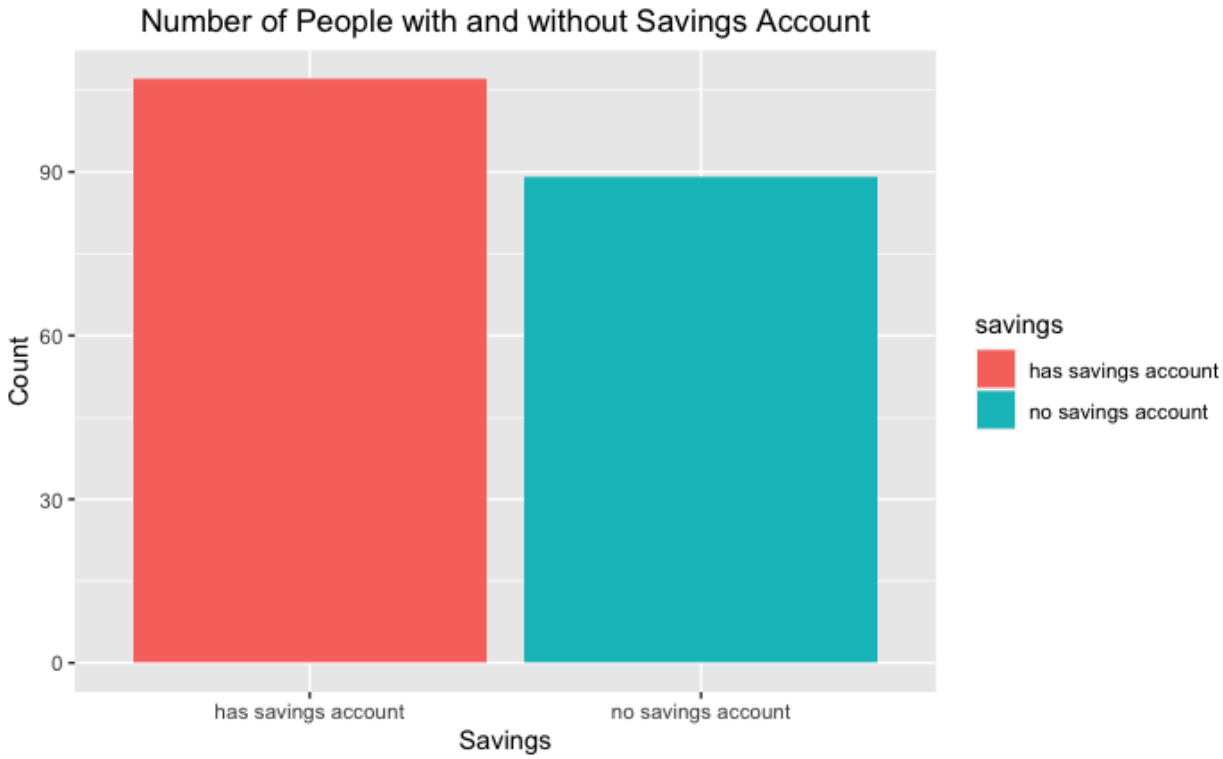

Figure 9: People Surveyed by Savings Account

Again, there was nothing unusual with the number of surveyed citizens from the initial study as shown in Figure 9.

Before attacking the main question that was proposed for this statistical analysis, these models helped shine light and tell the audience more about the data. After creating these visuals, it was then time to dive into creating a multiple logistic regression equation that could help predict disease outcome.

Multiple Logistic Regression Equation

First, a multiple logistic regression equation was created using all of the predictor variables to estimate their correlation and the power that they had when predicting if a citizen would have dengue. The estimated equation was as follows:

$$Y_i = \beta_0 - \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} - \beta_4 X_{i4} - \beta_5 X_{i5}$$

Equation with exact values:

$$Y_i = 2.076617 - 0.02728 X_{i1} + 0.202055 X_{i2} + 0.237633 X_{i3} - 1.249464 X_{i4} - 0.040692 X_{i5}$$

The equation shown above show a multiple logistic regression equation that can help predict if a citizen will have dengue or not. When all of the independent(X) values are put into this equation, the answer will come out as Y_i, which is the log odds that someone will have dengue.

Table 1: Covariates in Full Model

Covariate	Symbol in Equation	P-Value
Age in Years	X_{i1}	0.002813
Socio-Economic Status: Middle	X_{i2}	0.659925
Socio-Economic Status: Upper	X_{i3}	0.583789
Sector 2	X_{i4}	0.000466

No Savings Account	X_{i5}	0.918266
Not a Covariate but the Intercept	β_0	7.84e-05

All p-values for the covariates and the y-intercept are given in the table above. Most of the p-values were greater than 0.05, therefore unreliable. For the covariates that don't apply to this model, for example, if a person had a savings account, then the "no savings account" covariate would be set to zero. This covariate is set to zero due to it being inapplicable to this equation. In order to create a model that would be more accurate, a new logistic equation was created with only the covariates that had a p-value that was less than 0.05, thus being statistically significant. This new equation only included age in years and the sector as covariates in predicting disease, this equation is shown here:

$$Y_i = \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2}$$

Equation for reduced model with exact values:

$$Y_i = 2.15966 - 0.02681 X_{i1} + 1.18169 X_{i2}$$

Table 2: Covariates in Reduced Model

Covariate	**Symbol in Equation**	**P-Value**
Age in Years	X_{i1}	0.002813
Sector 2	X_{i2}	0.659925
Intercept (Not a Covariate)	β_0	0.583789

In the reduced model, it was observed that the p-values for covariates and the intercept were all lower once the unreliable covariates were eliminated. Thus, this model is reasonable more accurate that the model with all of the covariates. Plus, the AIC for this reduced model was calculated to be 217.64, which was lower(meaning it is more accurate) than the AIC that was calculated for the full model(223.21).

Conclusion

What can be concluded from the data was that the most reliable covariates for predicting disease were age in years and which sector the individuals live in. These conclusions were gathered from multiple aspects of the analysis. The p-values for these two covariates were the only p-values that were less than 0.05 for the full and reduced models. In addition to the p-values being lower for these covariates, when a reduced model was produced with only these two covariates as the predictor variables, the AIC value was lower, meaning that the reduced model was the more accurate of the two.

The p-values for the other covariates were very high for this analysis. Suggesting that they were not only inaccurate, but very inaccurate. There may be multiple factors that are contribute to this high p-value which was further discussed in the discussion section.

Discussion

The whole process of the original study was a great concept. Using multiple predictor variables to understand if someone would have dengue is a great idea. However, if a further study were to be conducted there are two main issues that would need to be resolved. These issues pertain to the sample size and one of the predictor variables.

A sample size of 196 is way to small to accurately represent the dengue outbreak that affected multiple cities of the pacific coast of Mexico. If a further study were to be conducted, a larger sample size is necessary.

Another aspect of the the study that could be improved would be the selection of the predictor variables. Particularly, the savings account variable. All of the other variables were correlated to disease outbreak. Age groups, economic class, and area lived in can definitely affect the outbreak of a disease. However, whether someone has a savings account or not does not make much sense in how it could tell whether someone would be more likely to have dengue or not.

References

[1] Dantes HG, Koopman JS, Addy CL, Zarate ML, Marin MA, Longini Júnior IM, Guttierez ES, Rodriguez VA, Garcia LG, Mirelles ER. Dengue epidemics on the Pacific Coast of Mexico. Int J Epidemiol. 1988 Mar;17(1):178-86. doi: 10.1093/ije/17.1.178. PMID: 3384535.

Appendix

#Packages needed for data evaluation and cleaning the data.

```
disease_data <- data.frame(read_csv("~/Downloads/disease - Sheet1.csv"))

## Rows: 196 Columns: 6

## ── Column specification
──────────────────────────────────────────────────────────────────────────────

## Delimiter: ","

## dbl (6): id, ageyrs, ses, sector, disease, savings

##

## ℹ Use `spec()` to retrieve the full column specification for this data.

## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.

disease_data[disease_data$ses == 1,]$ses <- "upper"

disease_data[disease_data$ses == 2,]$ses <- "middle"

disease_data[disease_data$ses == 3,]$ses <- "lower"

disease_data[disease_data$disease == 1,]$disease <- "with disease"

disease_data[disease_data$disease == 0,]$disease <- "without disease"

disease_data[disease_data$savings == 1,]$savings <- "has savings account"

disease_data[disease_data$savings == 0,]$savings <- "no savings account"



disease_data$ses <- as.factor(disease_data$ses)

disease_data$sector <- as.factor(disease_data$sector)

disease_data$disease <- as.factor(disease_data$disease)
```

```
disease_data$savings <- as.factor(disease_data$savings)

disease_data$ageyrs <- as.integer(disease_data$ageyrs)

#View(disease_data)



#Check data types and if there are NA

str(disease_data)

## 'data.frame':  196 obs. of  6 variables:

##  $ id            : num  1 2 3 4 5 6 7 8 9 10 ...

##  $ ageyrs : int  33 35 6 60 18 26 6 31 26 37 ...

##  $ ses           : Factor w/ 3 levels "lower","middle",..: 3 3 3 3 1 1 1 2 2 2 ...

##  $ sector : Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 1 1 ...

##  $ disease: Factor w/ 2 levels "with disease",..: 2 2 2 2 1 2 2 1 1 2 ...

##  $ savings: Factor w/ 2 levels "has savings account",..: 1 1 2 1 2 2 2 1 2 2 ...

sum(is.na(disease_data))

## [1] 0
```

#An Overview of Diseases.(Reread and check all of this code.)

```
#Boxplot of Disease and Ages

#With Disease

disease_age_yes <- disease_data %>%

 filter(disease == "with disease")

boxplot(disease_age_yes$ageyrs, main = "Ages with Disease", pch = 19, col = "lightblue", horizontal =
TRUE, xlab = "Age in Years")
```

```
summary(disease_age_yes$ageyrs)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.

##    6.0    17.0    29.0    32.6    42.0    74.0

#Without Disease

disease_age_no <- disease_data %>%

 filter(disease == "without disease")

boxplot(disease_age_no$ageyrs, main = "Ages without Disease", pch = 19, col = "lightblue", horizontal = TRUE, xlab = "Age in Years")

summary(disease_age_no$ageyrs)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.

##    1.00    8.00   17.00   22.14   31.00   85.00

#Boxplot of Diseases in Each Category


#SES

#W/ disease

ggplot(disease_age_yes, aes(x = ses, fill = ses)) +

 geom_bar() +

 ggtitle("Disease by SES") +

 theme(plot.title = element_text(hjust = 0.5)) +

 xlab("SES") +

 ylab("Number of People with Disease")

#W/O disease
```

```
ggplot(disease_age_no, aes(x = ses, fill = ses)) +
 geom_bar() +
 ggtitle("No Disease by SES") +
 theme(plot.title = element_text(hjust = 0.5)) +
 xlab("SES") +
 ylab("Number of People without Disease")
#Sector
#W Disease
ggplot(disease_age_yes, aes(x = sector, fill=sector)) +
 geom_bar() +
 ggtitle("Disease by Sector") +
 theme(plot.title = element_text(hjust = 0.5)) +
 xlab("Sector") +
 ylab("Number of People with Disease")
#W/O Disease
ggplot(disease_age_no, aes(x = sector, fill=sector)) +
 geom_bar() +
 ggtitle("No Disease by Sector") +
 theme(plot.title = element_text(hjust = 0.5)) +
 xlab("Sector") +
 ylab("Number of People without Disease")
```

```
#Savings

#W Disease

ggplot(disease_age_yes, aes(x = savings, fill=savings)) +

 geom_bar() +

 ggtitle("Disease by Savings Account") +

 theme(plot.title = element_text(hjust = 0.5)) +

 xlab("Savings Account") +

 ylab("Number of People with Disease")

#W/O Disease

ggplot(disease_age_no, aes(x = savings, fill=savings)) +

 geom_bar() +

 ggtitle("No Disease by Savings Account") +

 theme(plot.title = element_text(hjust = 0.5)) +

 xlab("Savings Account") +

 ylab("Number of People without Disease")
```

#Number of People in Different Categories.

```
#ses

ggplot(disease_data, aes(x = ses, fill = ses)) +

 geom_bar() +

 ggtitle("Number of People in Each Socio-Economic Class") +

 theme(plot.title = element_text(hjust = 0.5)) +
```

```
 xlab("Ses") +
 ylab("Count")
#sector
ggplot(disease_data, aes(x = sector, fill = sector)) +
 geom_bar() +
 ggtitle("Number of People in Each Sector") +
 theme(plot.title = element_text(hjust = 0.5)) +
 xlab("Sector") +
 ylab("Count")
#disease
ggplot(disease_data, aes(x = disease, fill = disease)) +
 geom_bar() +
 ggtitle("Number of People with and without Disease") +
 theme(plot.title = element_text(hjust = 0.5)) +
 xlab("disease") +
 ylab("Count")
#seaving
ggplot(disease_data, aes(x = savings, fill = savings)) +
 geom_bar() +
 ggtitle("Number of People with and without Savings Account") +
 theme(plot.title = element_text(hjust = 0.5)) +
```

```
 xlab("Savings") +
 ylab("Count")
```

#Creating the Model for the Data. Actual Multiple Logistic Regression Equation.

```
#Logistic model with all variables
fit <- glm(disease ~ ageyrs + ses + sector + savings, data = disease_data, family = "binomial")
summary(fit)
##
## Call:
## glm(formula = disease ~ ageyrs + ses + sector + savings, family = "binomial",
##     data = disease_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0918  -1.0134   0.5630   0.8309   1.6614
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          2.076617   0.525834   3.949 7.84e-05 ***
## ageyrs              -0.027280   0.009132  -2.987 0.002813 **
## sesmiddle            0.202055   0.459199   0.440 0.659925
## sesupper             0.237633   0.433750   0.548 0.583789
```

```
## sector2                    -1.249464   0.357009  -3.500 0.000466 ***
## savingsno savings account -0.040692   0.396540  -0.103 0.918266
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##         Null deviance: 236.33  on 195  degrees of freedom
## Residual deviance: 211.21  on 190  degrees of freedom
## AIC: 223.21
##
## Number of Fisher Scoring iterations: 4
plot(fit)
#Estimated Multiple Logistic Regression Equation
#Yi = 2.076617-0.027280Xi1+0.202055Xi2+0.237633Xi3-1.249464Xi4--0.040692Xi5 + Ei(error)
#The answer you will get for Yi is the log(odds) answer.
#Xi1 = age in years it what you would plug in here
#Xi2 = 1 if ses middle, 0 if not middle
#Xi3 = 1 is ses upper, 0 if not (0 for both this and Xi2 when ses is lower)
#Xi4 = 1 if sector 2, 0 if sector 1
#Xi5 = 1 if no savings account, 0 if they do have a savings account
```

```
#P-Values suggest that age and sector are reliable variables for my analysis.



#New logistic model with only age and sector as variables

fit_reduced <- glm(disease ~ ageyrs + sector, data = disease_data, family = "binomial")

summary(fit_reduced)

##

## Call:

## glm(formula = disease ~ ageyrs + sector, family = "binomial",

##     data = disease_data)

##

## Deviance Residuals:

##     Min      1Q  Median      3Q     Max

## -2.0275  -1.0093   0.5606   0.8199   1.6839

##

## Coefficients:

##             Estimate Std. Error z value Pr(>|z|)

## (Intercept)  2.15966   0.34388   6.280 3.38e-10 ***

## ageyrs      -0.02681    0.00865  -3.100 0.001936 **

## sector2     -1.18169    0.33696  -3.507 0.000453 ***

## ---

## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##

## (Dispersion parameter for binomial family taken to be 1)

##

##         Null deviance: 236.33  on 195  degrees of freedom

## Residual deviance: 211.64  on 193  degrees of freedom

## AIC: 217.64

##

## Number of Fisher Scoring iterations: 3

plot(fit_reduced)

#Equation for the logistic model with only age and sector as variables

#Yi = 2.15966 - 0.02681Xi1 - 1.18169Xi2 + Ei(error)

#P-values are lower for Beta0 and the two variables in this reduced model compared to the full.
Suggesting that this model id more accurate.

#The AIC is lower for the reduced model, also showing that it is the more accurate model.
```