1. Consider the Parana dataset from the geoR library (data(parana), help(parana)) and Identify the region of interest, design, response variable and covariates, if any. Plot what the map would look like.

Exploratory Data Analysis:

Geographically, the dataset's 143 data points range from east (150.122 to 768.509) to north (70.360 to 461.968).With a median of 269.92, the data values range from a minimum of 162.77 to a maximum of 413.70.Additional information labeled "loci.paper" is included in the dataset.

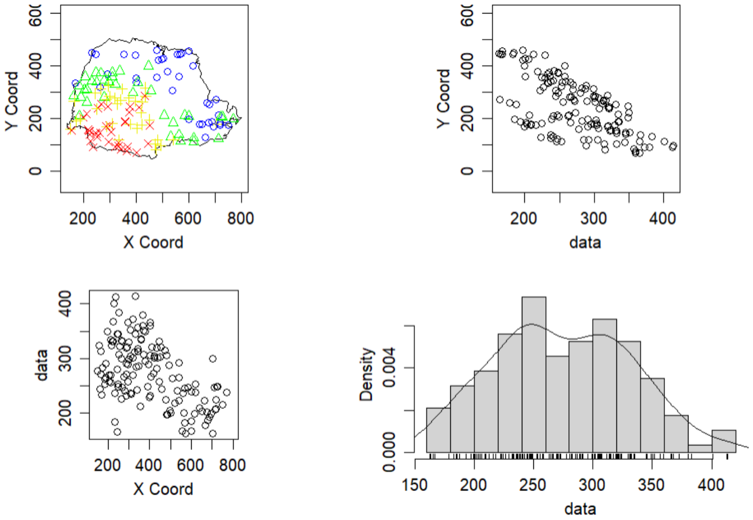


Figure 1: shows the distribution of the data

The below graph shows the distribution of log-silica percentages throughout the Paraná basin, with colored dots denoting values that fall or rise above the median. Red points are displayed above the median, and blue points are displayed below it. These categories are indicated by a legend. This facilitates the rapid identification of basin regions with comparatively higher or lower silica content.

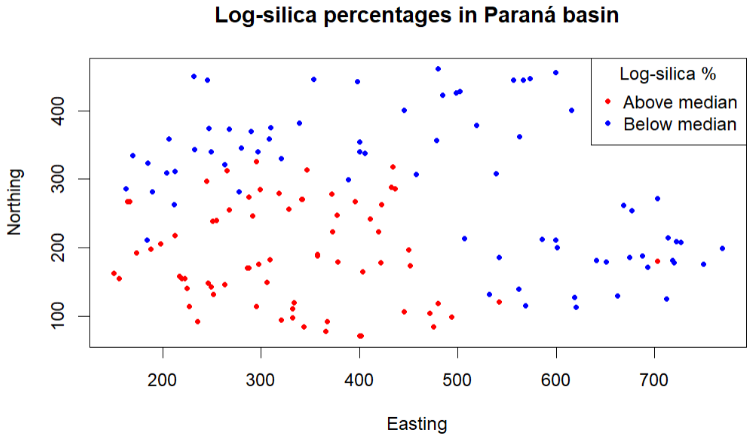


Figure 2: Log-silica percentages in Parana basin

Here below is the Parana data points with in the border of the Paranam the data points in red colors.

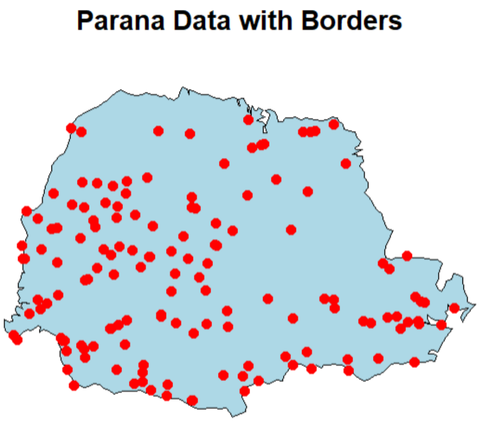
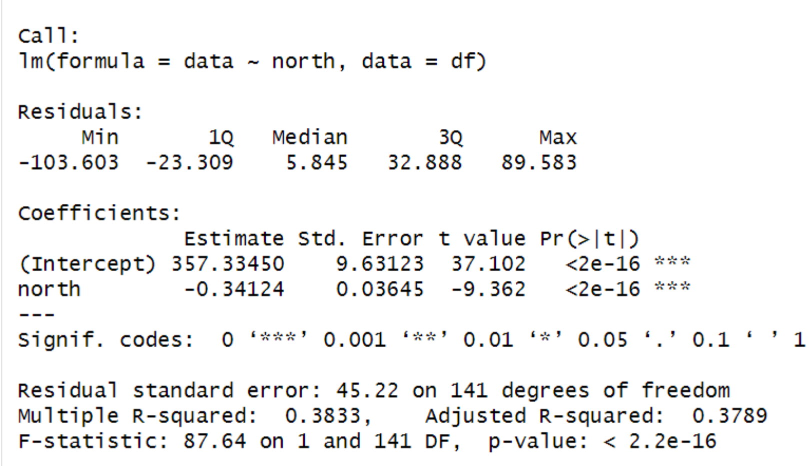


Figure 3: Parana Data with borders

2. Consider elevation data(elevation) as a simple linear regression problem, with elevation as the response and north-south coordinate as the explanatory variable. Fit the model and examine the residuals of the model. Do you consider that a more sophisticated model is necessary for the analysis of spatial variation? Do you consider that a more sophisticated model is necessary for the analysis of spatial variation, justify your answer.



```
Call:
lm(formula = data ~ north, data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-103.603  -23.309    5.845   32.888   89.583 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 357.33450    9.63123  37.102   <2e-16 ***
north        -0.34124    0.03645  -9.362   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 45.22 on 141 degrees of freedom
Multiple R-squared:  0.3833,	Adjusted R-squared:  0.3789 
F-statistic: 87.64 on 1 and 141 DF,  p-value: < 2.2e-16
```

Figure 4: Simple Linear Regression

The north-south coordinate (north) was used as the explanatory variable and elevation (data) as the response variable in a straightforward linear regression model. With a standard error of roughly 9.63, the analysis places the intercept at roughly 357.33, which is the point on the y-axis where the regression line crosses the north-south coordinate when it is zero. With a standard error of 0.036 and a slope coefficient of roughly -0.341 for the north, there appears to be a little negative correlation between height and the north-south coordinate. This means that elevation declines as one moves northward (or southbound, depending on the coordinate system orientation). The R-squared result shows that approximately 38.33% of the variance in elevation is explained by this model.These results are statistically significant because the p-values for the slope and the intercept are both less than 0.001. The quartile spread of the residuals, which range from -103.603 to 89.583, indicates some variability around the fitted line, which may necessitate additional research into any non-linearity of variance problems in the data.

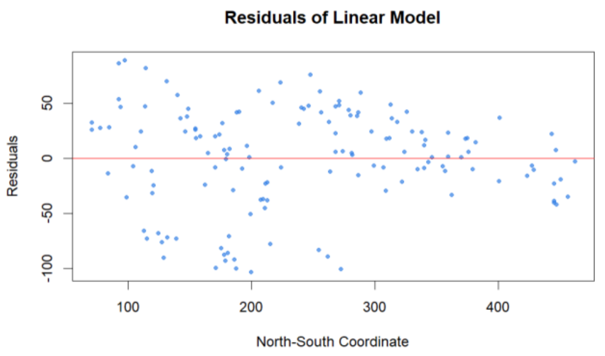


Figure 5: Residual Linear Model

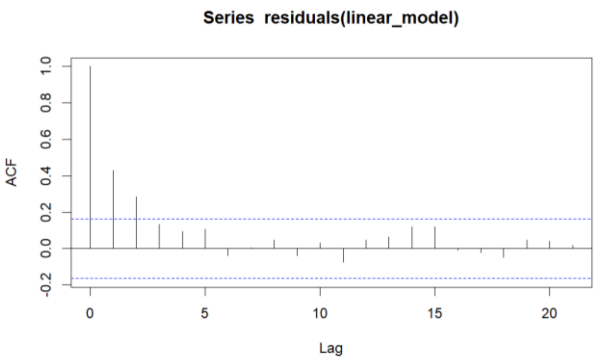
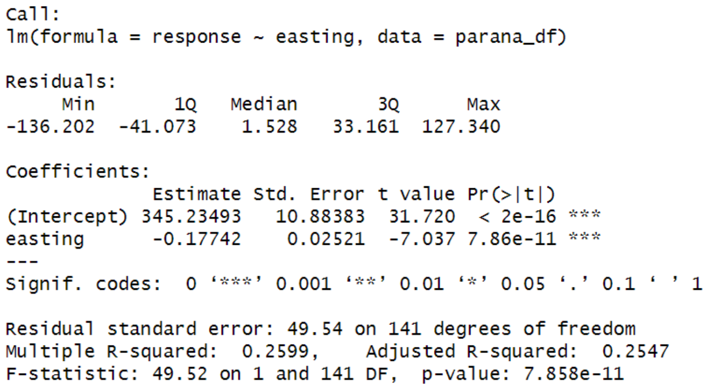


Figure 6: Series residuals(linear_model)

Yes, in order to analyze spatial variation, a more complex model is required because: The large range of residuals indicates that the underlying patterns and complexities may not be fully captured by the simple linear model. With an R-squared value of 0.3833, the current model is unable to explain a sizable amount of the variation in the data.Simple linear models do not take autocorrelation into account when analyzing spatial data, which might result in results that are skewed or ineffective.

3. Consider the following models for a data set, whose response Yi = 1, 2, ..., n associated with a sequence of xi positions along a spatial axis of one x-dimension.

Model A:



```
Call:
lm(formula = response ~ easting, data = parana_df)

Residuals:
     Min       1Q   Median       3Q      Max 
-136.202  -41.073    1.528   33.161  127.340 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 345.23493   10.88383  31.720  < 2e-16 ***
easting      -0.17742    0.02521  -7.037 7.86e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 49.54 on 141 degrees of freedom
Multiple R-squared:  0.2599,	Adjusted R-squared:  0.2547 
F-statistic: 49.52 on 1 and 141 DF,  p-value: 7.858e-11
```

Figure 7: Model A regression model

A straightforward linear regression model in which the response variable is forecast using the dataset's "easting" location. The response variable shrinks as the easting value rises, according to the negative coefficient for "easting" (-0.17742). The model explains just around 26% of the response's variability, which suggests that other factors may also be impacting the response. Nevertheless, it is statistically significant (p < 2e-16 for the intercept and p < 7.86e-11 for easting).

Here below is the model A best fitted line in the data.

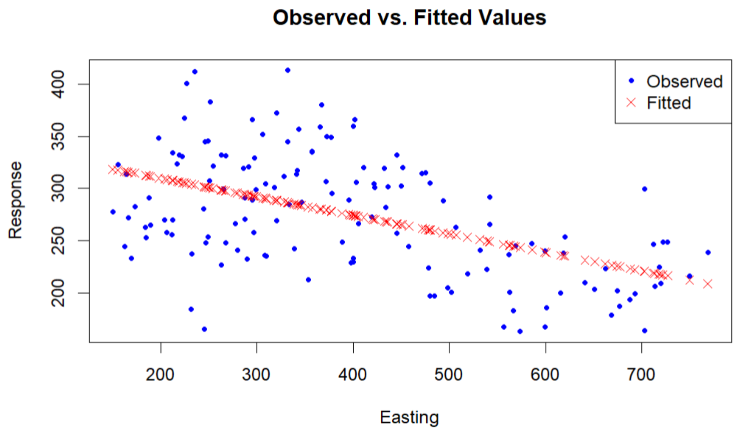
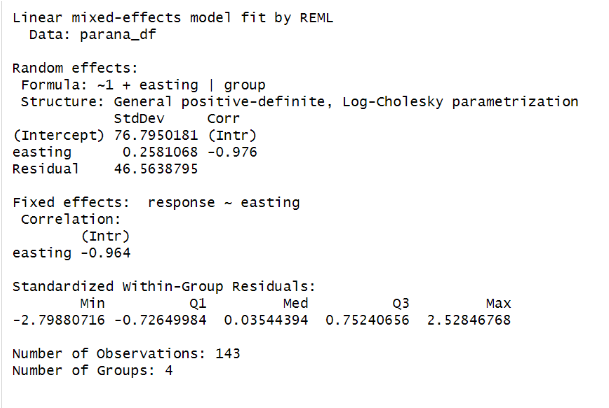


Figure 8: Linear Regression Model A observed vs Fitted values

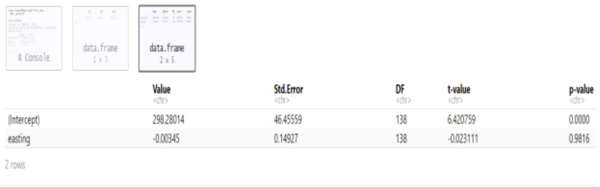
Model B:

The 'easting' variable into groups of quartiles and develops a linear mixed-effects model to take into consideration the variation in slopes and intercepts within these groups. With this configuration, the model is able to represent both the random fluctuations over various spatial regions and the direct effects of "easting."



```
Linear mixed-effects model fit by REML
  Data: parana_df

Random effects:
 Formula: ~1 + easting | group
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev     Corr  
(Intercept) 76.7950181 (Intr)
easting      0.2581068 -0.976
Residual    46.5638795       

Fixed effects:  response ~ easting
 Correlation: 
        (Intr)
easting -0.964

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max 
-2.79880716 -0.72649984  0.03544394  0.75240656  2.52846768 

Number of Observations: 143
Number of Groups: 4 
```

Figure 9: Linear Regression Model B



	Value	Std.Error	DF	t-value	p-value
	<chr>	<chr>	<chr>	<chr>	<chr>
(Intercept)	298.28014	46.45559	138	6.420759	0.0000
easting	-0.00345	0.14927	138	-0.023111	0.9816

2 rows

Figure 10: Intercept and easting

The intercept and slope of "easting" in the linear mixed-effects model output exhibit substantial random effects variability, with a large negative correlation (-0.976) between them. The response variance that the model is unable to explain is indicated by the residual standard deviation, which stands at 46.56. The model, which uses data from 143 observations split into 4 groups, indicates intricate, variable correlations between "easting" and the reaction across various spatial regions, emphasizing how crucial it is to take spatial heterogeneity into account when doing analyses.

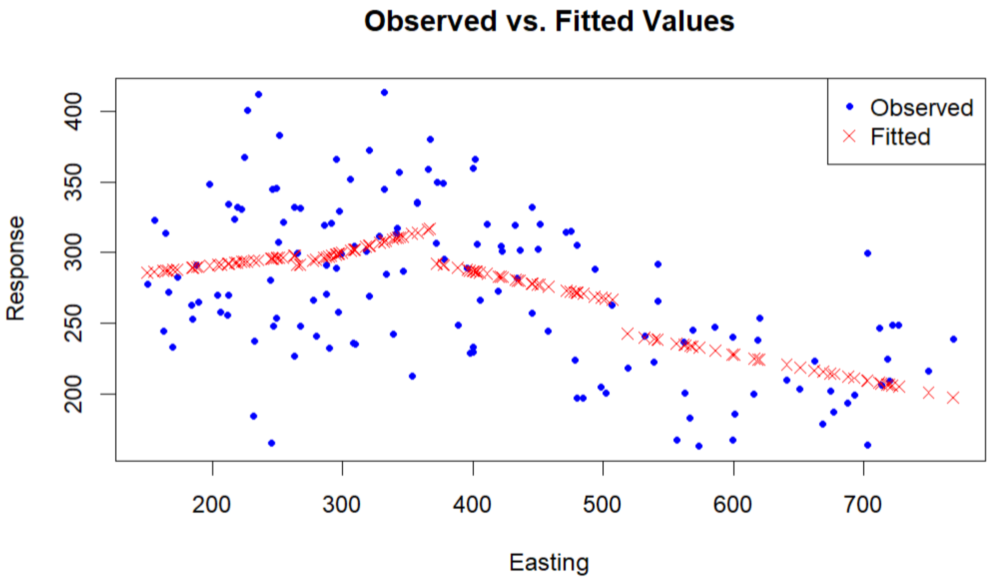
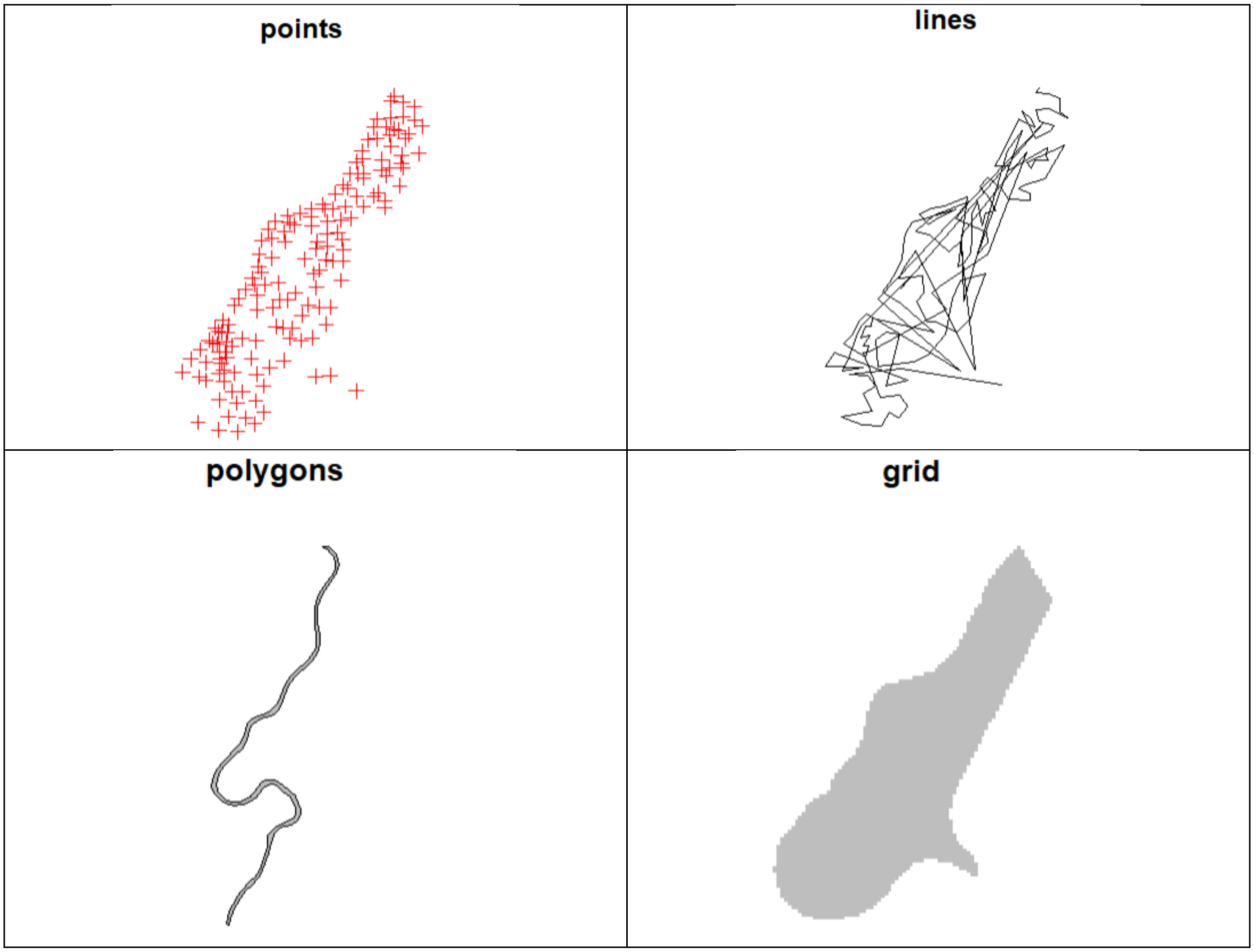


Figure 11: Model b Linear Regression

4.Look for and reproduce figure 3.2 from the book “Applied Spatial Data Analyses with R”.
Model A:



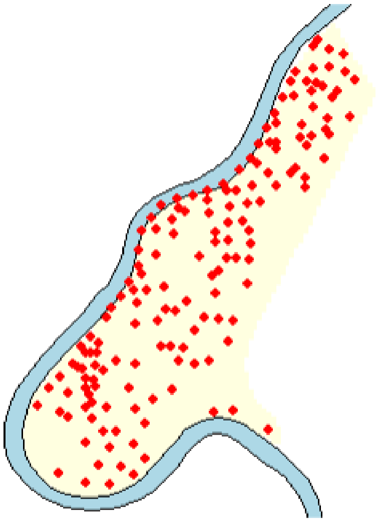


Figure 12: Plot

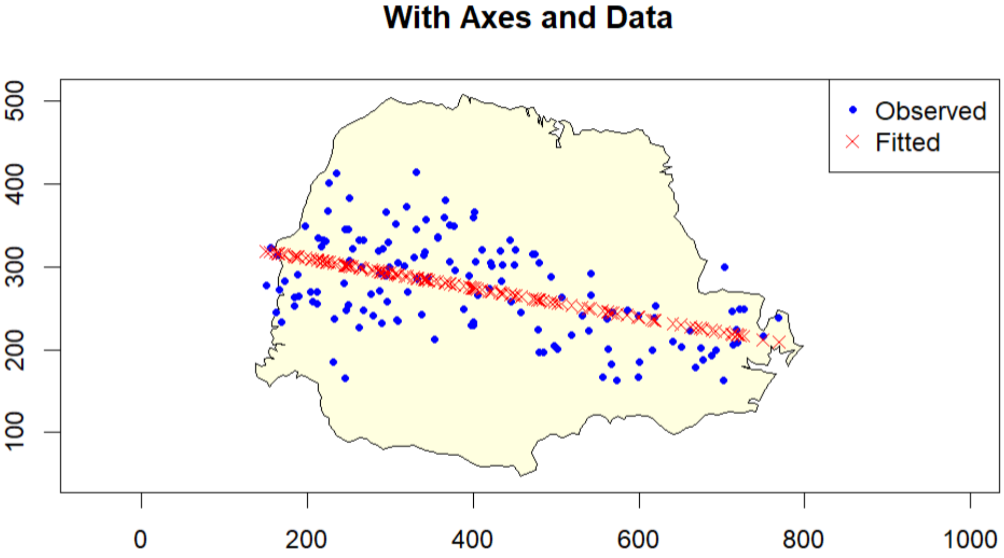


Figure 13: Model A plot in Parana Region

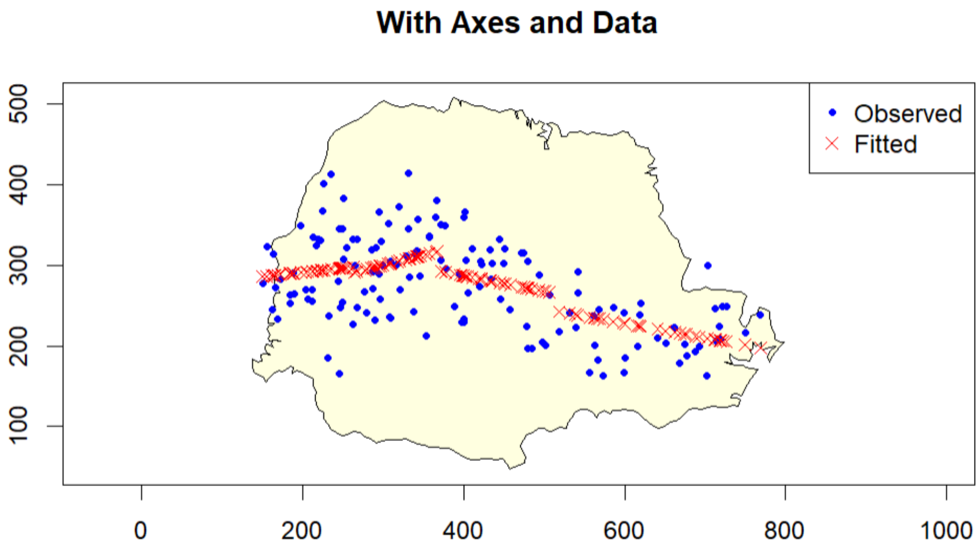
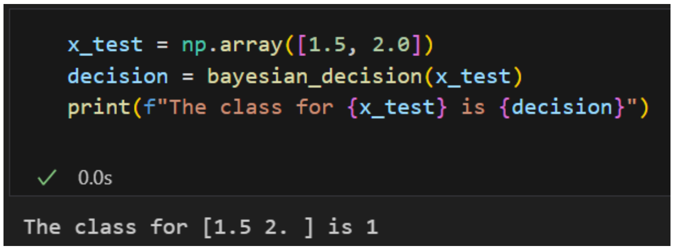


Figure 14: Model B plot in Parana Region

5. Consider the following.

a) Calculate the optimal Bayesian classifier for a symmetric cost function.

The class for [1.5 2.0] is 1," verifies that the point [1.5, 2.0] [1.5,2.0] is correctly identified by the Bayesian classifier as more likely belonging to the class $Y = 1$ Y=1, based on its closeness to the mean [1, 2] [1,2] in the multivariate normal distribution space. Under symmetric cost functions and identical prior probability for each class, this classification is ideal.



```
x_test = np.array([1.5, 2.0])
decision = bayesian_decision(x_test)
print(f"The class for {x_test} is {decision}")
```

✓ 0.0s

```
The class for [1.5 2. ] is 1
```

Figure 15: Bayesian classifier test

b) Generate samples (X, Y) of size n = 50, 100, 500.

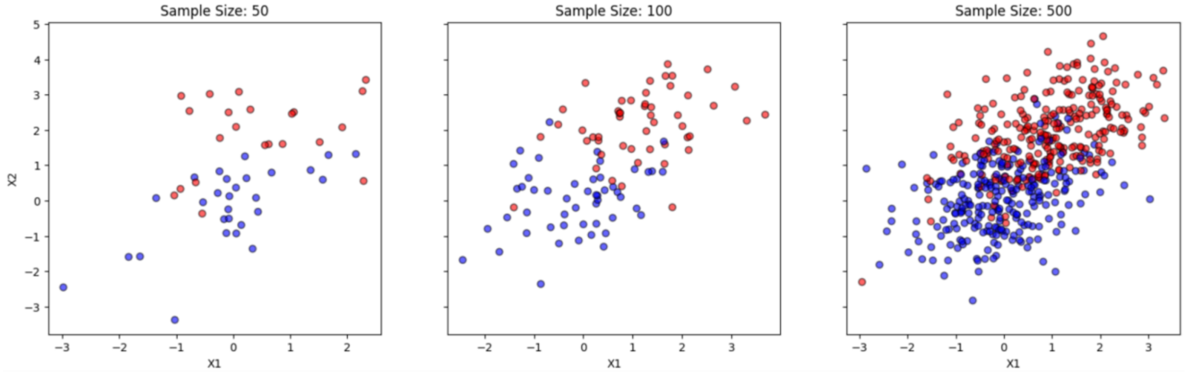
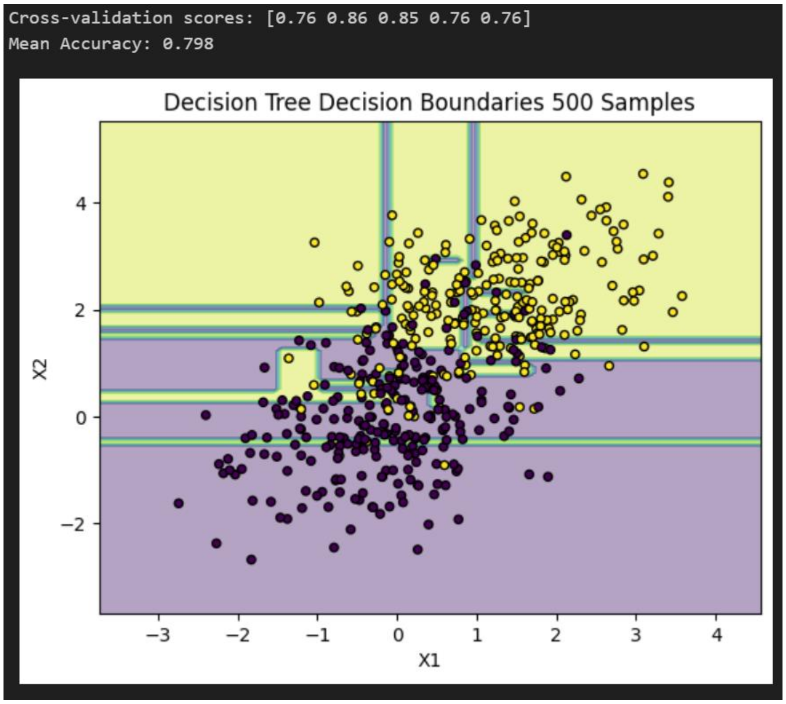


Figure 16: Generated the Random sample 50, 100, 500

c) Find and discuss suitable decision trees for these data and compare the best models obtained with cross-validation with the optimal Bayesian classifier, both visually and quantitatively.

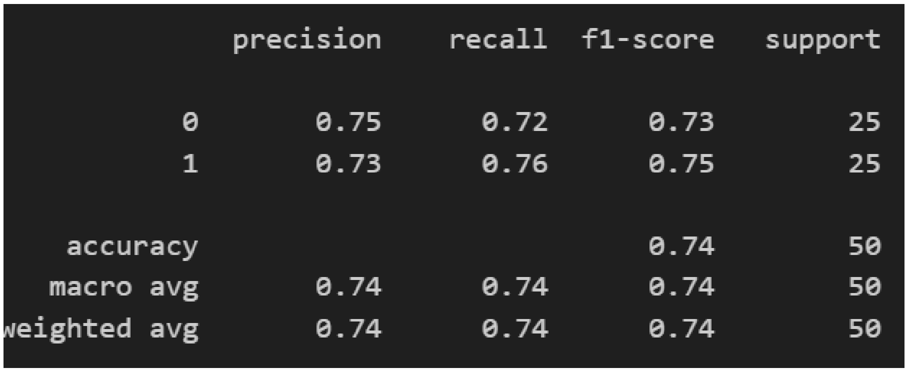
For 500 samples:



```
Cross-validation scores: [0.76 0.86 0.85 0.76 0.76]
Mean Accuracy: 0.798
```

Figure 17: Decision tree plot with accuracy for 500 samples

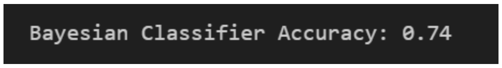
Across class 0 and class 1, the decision tree classifier performed similarly, achieving an accuracy of 74% with virtually comparable precision, recall, and F1-scores for both classes on a dataset of 50 samples.



```
              precision    recall  f1-score   support

           0       0.75      0.72      0.73        25
           1       0.73      0.76      0.75        25

    accuracy                           0.74        50
   macro avg       0.74      0.74      0.74        50
weighted avg       0.74      0.74      0.74        50
```

Figure 18: confusion matrix for 500 samples

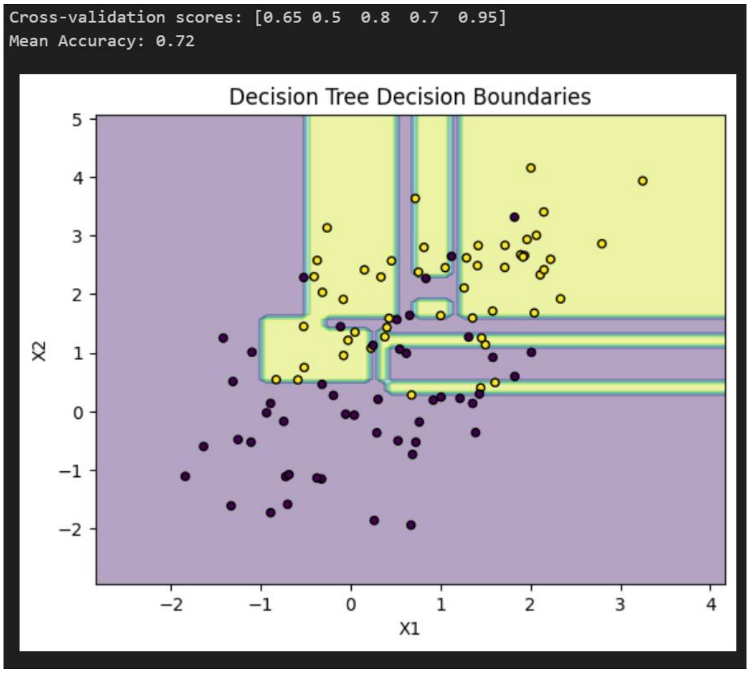
Using cross-validation, we created, trained, and assessed a decision tree model. Next, we visualized the model's decision boundaries and compared its performance—both visually and quantitatively—with that of the best Bayesian classifier.



```
Bayesian Classifier Accuracy: 0.74
```

Figure 18: Accuracy

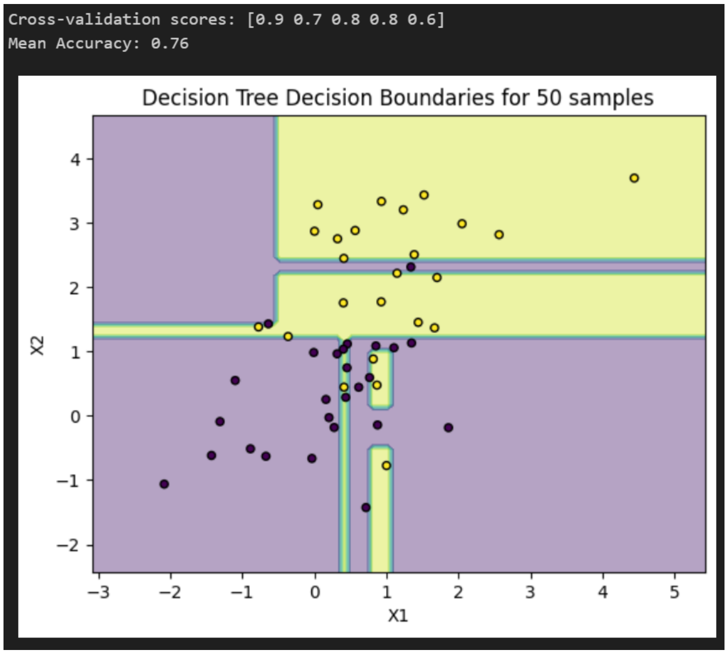
For 100 samples



```
Cross-validation scores: [0.65 0.5  0.8  0.7  0.95]
Mean Accuracy: 0.72
```

Figure 20: Decision Tree plot on 100 samples

For 50 Samples



```
Cross-validation scores: [0.9 0.7 0.8 0.8 0.6]
Mean Accuracy: 0.76
```

Figure 21: Decision Tree plot on 50 Samples