Exercise 1: Categorize the following examples of spatial data as to their data type: geostatistical data, lattice data, point pattern data, or non-of-the-above. Explain your answers. Also, if whats being described is point pattern data, explain whether it's marked or unmarked point pattern data.

a. Distribution of oaks and pines in a forest stand.

Solution:
A Distribution of oaks and pines in a forest stand as a collection of longitude and latitude pairs along with information on species, oaks or pine, would be an example of marked point pattern data. Here the location is the data, along with the extra dimension of species classification.

b. Number of squirrel nests in the trees in (a).

Solution:
In this case I assume the trees in (a) where denoted as point pattern data, in which case adding another dimension like number of squirrel nest, and thus the data would still be marked point pattern data.

c. Percentage of Republican voters in each state in the continental U.S;

Solution:
Here we are recording a percentage at each state, the data are some sort of aggregate for an entire region so this would be an example of lattice data.

d. Concentration of mineral in soil;

Solution:
This data seems like it would be used to predict where we might find deposits of a certain mineral in a given area. I would image the goal would be to produce a smooth map of mineral concentration, for something like prediction in unsampled locations. In which case I would say this is an example of geostatistical data. I could also see this data being used for some sort of cluster analysis of different minerals, to try an discover why certain minerals form in certain environments, in which case location would be part of the data like in point pattern data.

e. Amount of snowfall during January at 50 locations in Alaska;

Solution:
It seems like the goal of the data would be to estimate a smooth map of the snowfall throughout Alaska, for prediction purposes. The locations may effect the quality of that estimate but are not particularly relevant (like they would be in point pattern data). This suggests to me that this is an example of geostatistical data.

f. Elevation in the foothills of the Allegheny mountains;

Solution:
It seems like elevation is often assumed to be a continuous quantity and in most cases this data is used to produce a smooth(debatable) map of the elevation. I would suppose this is an example of geostatistcal data.

g. Locations of animals across time, using values reported by an electronic tag.

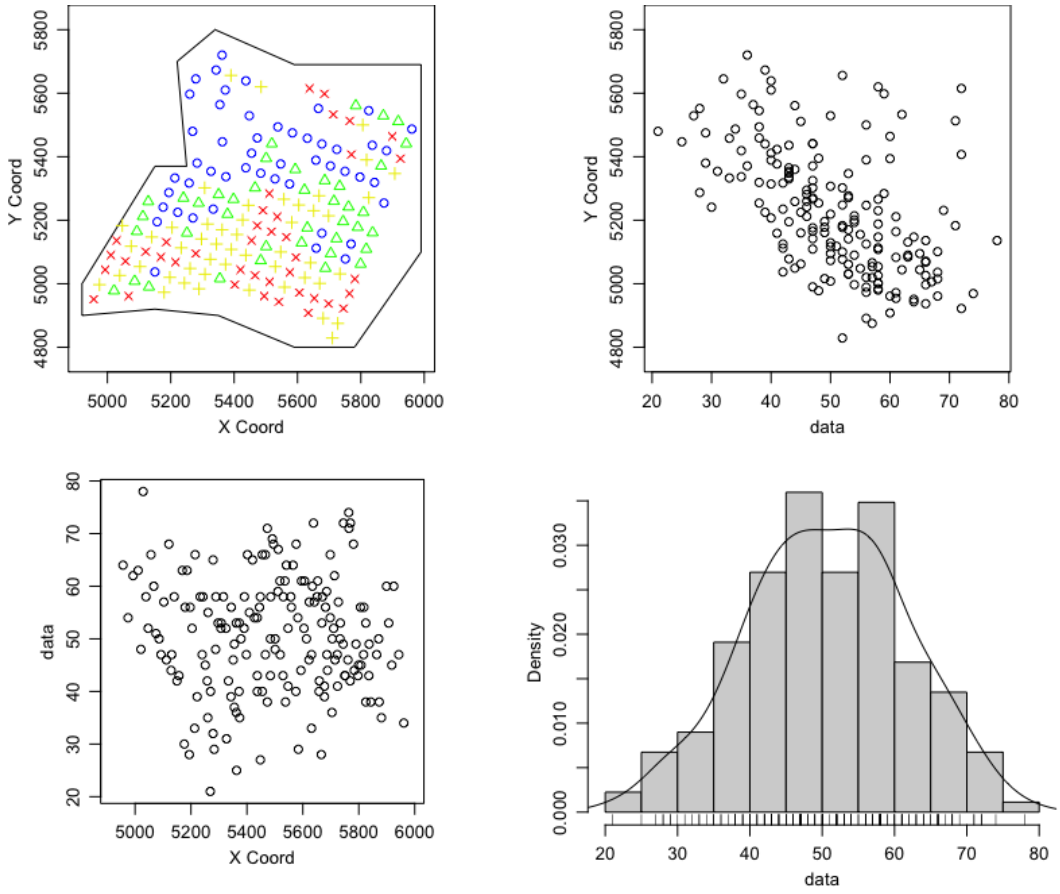
Solution:
It seems like the data would look like lat, long pairs with a dimension added for time. However the data is very dependent, in some cases we are making multiple observations on the same animal over time, and the location data is not entirely independent. If I had to put this in a category I would say marked point pattern data, but I think this would likely be treated as non-of-the-above.

Exercise 2: Plot using R.

a. Plot the data set ca20, using the plot function as illustrated on page 17 of the lecture notes.

Solution:
The following code produces the plot from page 17 of the lecture
Code:

```
> install.packages('geoR')
> library(geoR)
> plot(ca20)
```



b. What is this data set? You can find documentation on it using the R command help(parana).

Solution:
The data contains 178 observations of calcium content extracted from soil samples with 20cm depth. The area of study was divided into 3 separate sub regions. Generally the data are calcium content, location (in long, lat pair) and elevation. The area of each sub region was also recorded. Documentation from the help(ca20) function describes the data structures which store the data.

c. Describe the general features of this data set, e.g. N-S or E-W trend, if any; where are the values higher? Lower? WHat is the range of values in this data set? Describe the shape of the distribution.

Solution:
Looking at the first sub plot(top left) which shows the distribution of calcium samples across the sample region and recalling the hardcoded color values for the ordinal scale used to describe the data we see that there is a clear upward trend in

the N-S direction. This is corroborated by the partial dependence subplot(top right) which shows a clear trend as the latitude decreases. The second partial dependence plot(bottom left) shows no clear trend with the longitude of the sample and it's calcium content. The final subplot(bottom right) shows that the values for calcium concentration from our sample look approximately normal, with good symmetry. We should note that our sample's distribution could easily be bimodal if the sample region were two E-W transects separated by a substantial distance in the N-S direction. Beyond that unequal sampling on these transects would cause left or right skews.

d. If the data needs to be log-transformed, include the 4-panel plot you get after transforming, and discuss the resulting plots; does it look as though this is an appropriate transformation.

Solution:
The data does not appear to be skewed. I don't think log-transforming this data is necessary.

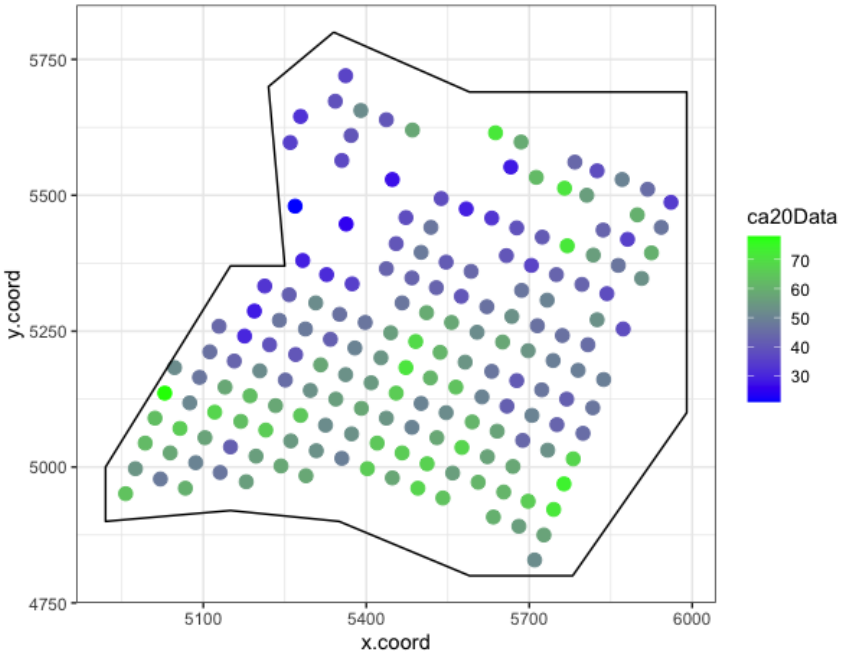
e. Create ggplot versions of the four panel plots for the data on the original scale. Try changing some colors, font sizes, dot sizes.

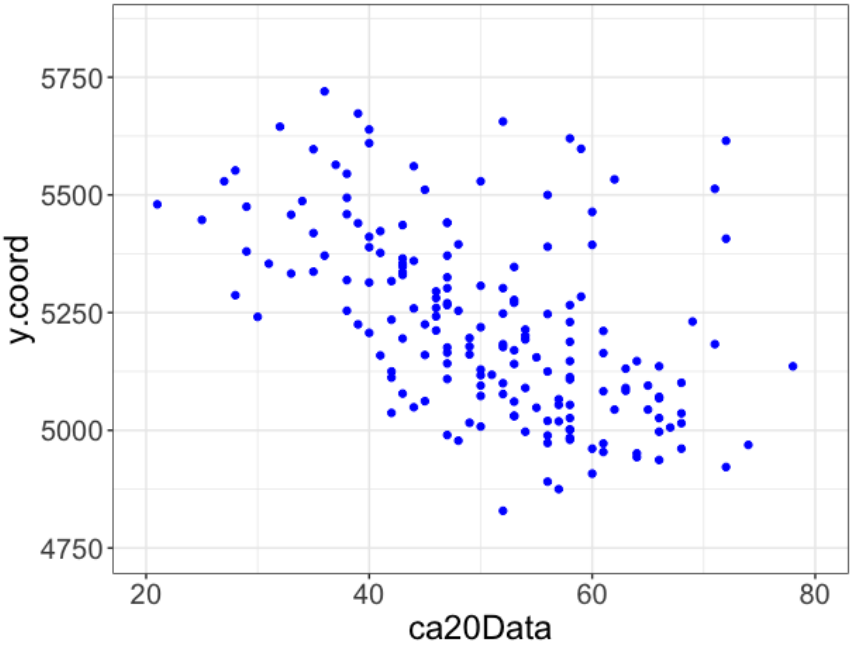
Solution:

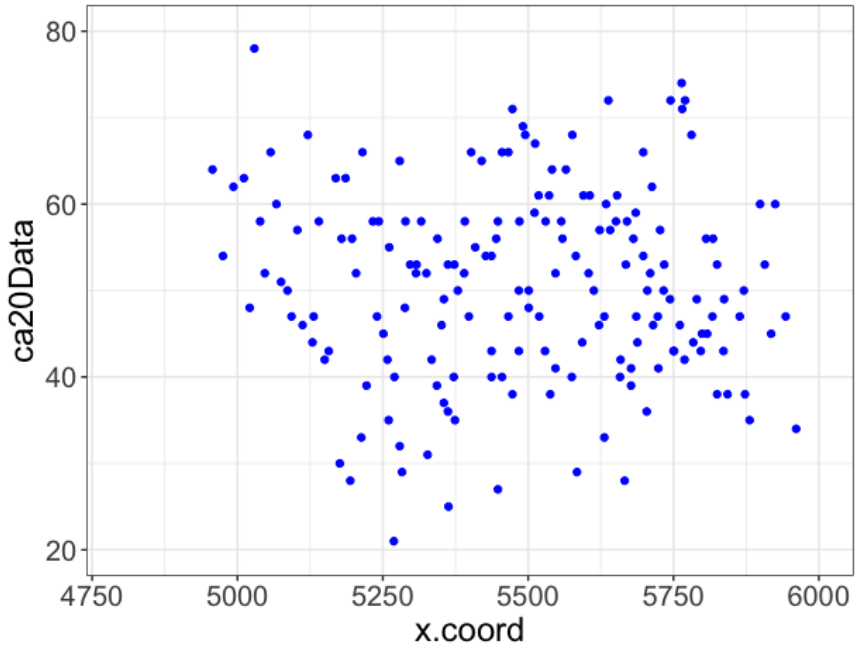
Code:

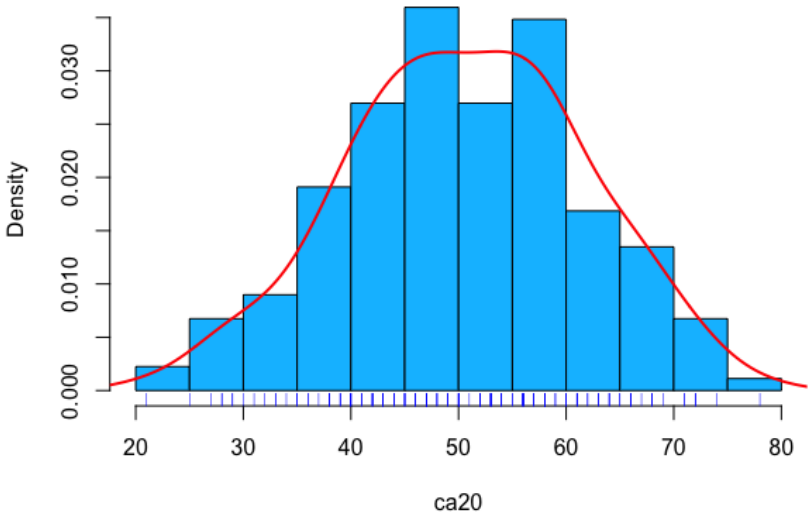
```
## Extract names of structures in ca20 object
names(ca20)

## Create New dataframe
mydata <- data.frame(
  x.coord = ca20$coords[,"east"],
  y.coord = ca20$coords[,"north"],
  ca20Data = ca20$data)
## Create Borders Dataframe
myborders <- data.frame(
  xx = ca20$border[,"east"],
  yy = ca20$border[,"north"])

## First subplot
ggplot(mydata) +
    geom_point(aes(x.coord,y.coord, color = ca20Data), size = 3) +
    coord_fixed(ratio = 1) +
    scale_color_gradient(low = "blue", high = "green") +
    geom_path(data = myborders, aes(xx,yy)) +
```

```
    theme_bw() # b/w background

## Pull min-max for axis limits
min(mydata$ca20Data)
max(mydata$ca20Data)
min(mydata$x.coord)
max(mydata$x.coord)
min(mydata$y.coord)
max(mydata$y.coord)

## Second subplot
ggplot(mydata, aes(ca20Data, y.coord)) +
    geom_point(size = 1.5, color = "blue") +
    theme_bw() +
    theme (axis.title = element_text(size = 18)) +
    theme(axis.text=element_text(size=15)) +
    xlim(20,80) +
    ylim(4750,5850)


## Third subplot
ggplot(mydata, aes(x.coord, ca20Data)) +
  geom_point(size = 1.5, color = "blue") +
  theme_bw() +
  theme (axis.title = element_text(size = 18)) +
  theme(axis.text=element_text(size=15)) +
  ylim(20,80) +
  xlim(4800,6000)

## Fourth subplot
hist(mydata$ca20Data, prob=TRUE, main="",
     xlab="ca20", col="deepskyblue")
rug(mydata$ca20Data, side = 1, col = "blue")
lines(density(mydata$ca20Data), col = 'red', lwd = 2)
```


5750
5500
5250
5000
4750
y.coord
5100
5400
5700
6000
x.coord
ca20Data
70
60
50
40
30


5750
5500
5250
5000
4750
y.coord
20
40
60
80
ca20Data

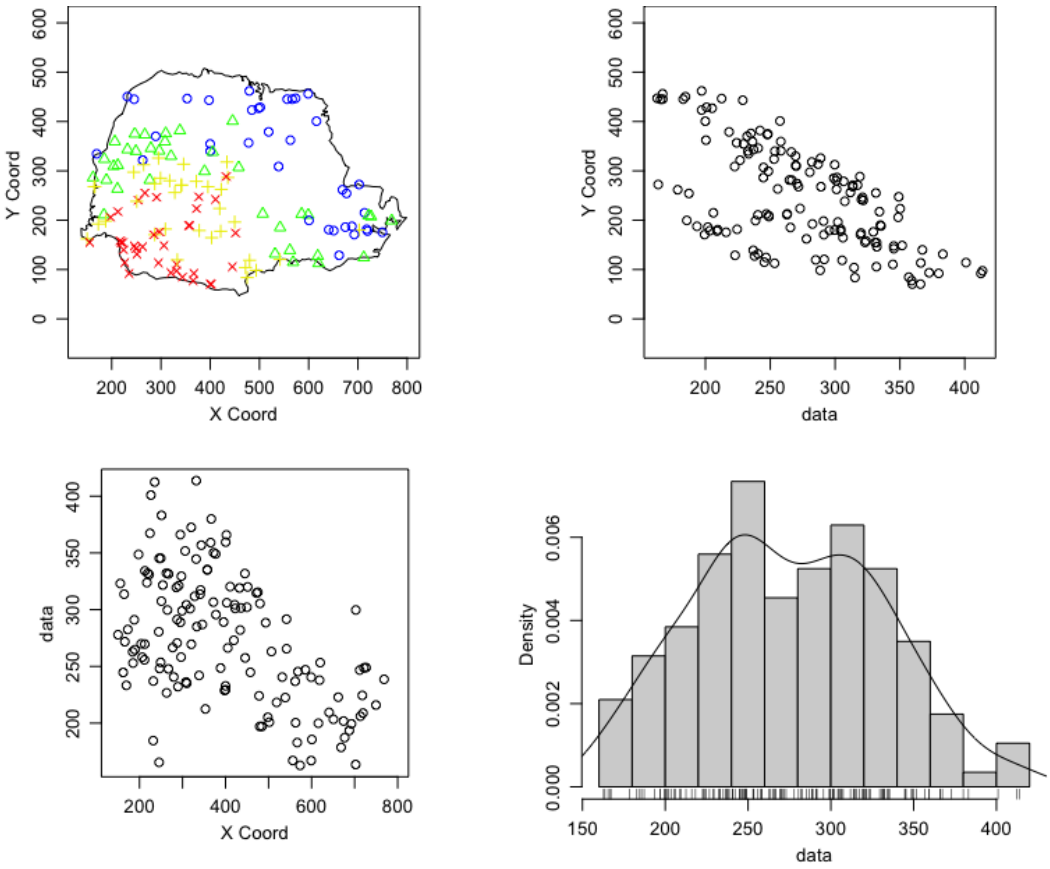

ca20Data
80
60
40
20
4750
5000
5250
5500
5750
6000
x.coord


Density
0.030
0.020
0.010
0.000
20
30
40
50
60
70
80
ca20

Exercise 3: Repeats part (a)-(d) for the parana data set.

a. Plot the data set parana, using the plot function as illustrated on page 17 of the lecture notes.

Solution:



b. What is this data set? You can find documentation on it using the R command help(parana).

Solution:
This data set is average rainfall records at 143 different recording stations during the May-June spring season. This information comes from the help(parana) command.

c. Describe the general features of this data set, e.g. N-S or E-W trend, if any; where are the values higher? Lower? What is the range of values in this data set?

Solution:
The first subplot shows us the distribution of weather stations, recalling the default color scale we know that the lowest values for average rainfall are in the north east region. We can see that as we move from the north east to south west the average rainfall increases. This general trend is supported by the partial dependce plots in the second and third subplots (upper right and lower left respectively). The final subplot (lower right) shows the distribution of our data sample. We can see that it appears norma. and symmetric with a small dip near the mean, which might suggests we are looking at a bimodal sample.

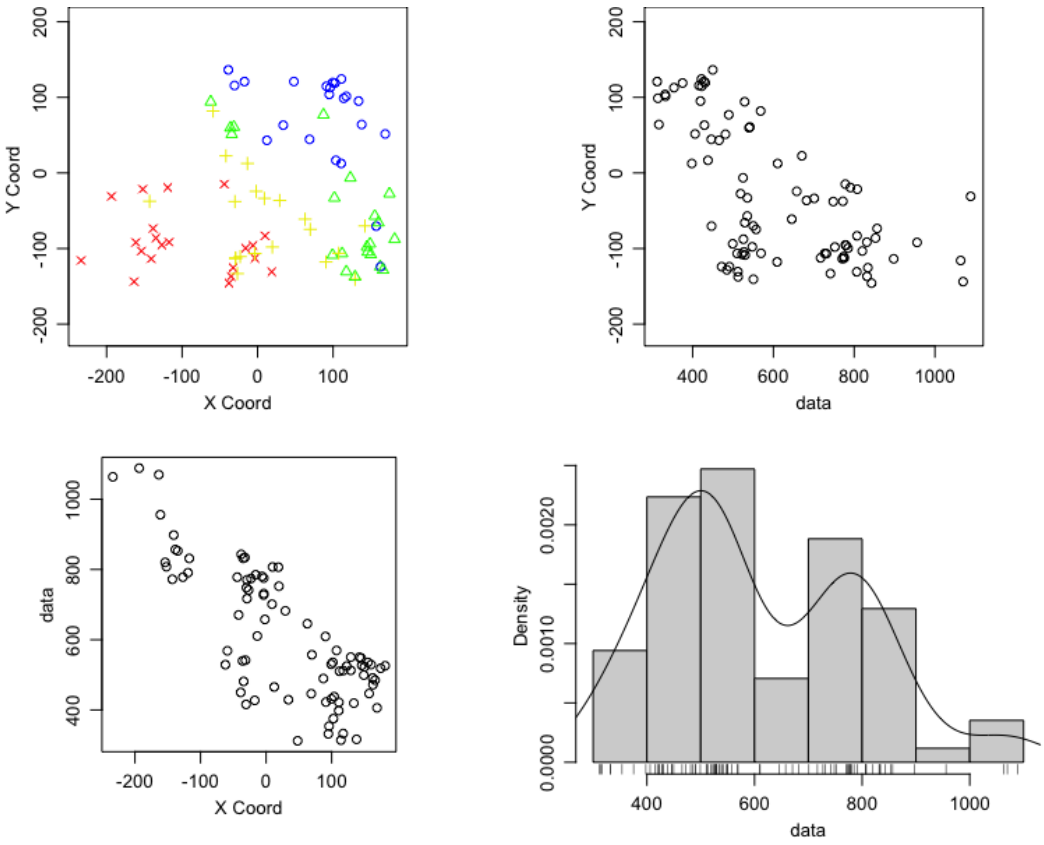
d. If the data needs to be log-transformed, include the 4-panel plot you get after transforming, and discuss the resulting plots; does it look as though this is an appropriate transformation.

Solution:
The data does not appear to be skewed. I don't think log-transforming this data is necessary.

Exercise 4: Repeats part (a)-(d) for the wolfcamp data set.

a. Plot the data set wolfcamp, using the plot function as illustrated on page 17 of the lecture notes.

Solution:



b. What is this data set? You can find documentation on it using the R command help(wolfcamp).

Solution:
This data set contains piezometric head measurements for the Wolfcamp aquifer in Texas. This is a measure of pressure in an aquifer that is computed from the height of water in a tube. The data comes in long, lat pairs and head measurements.

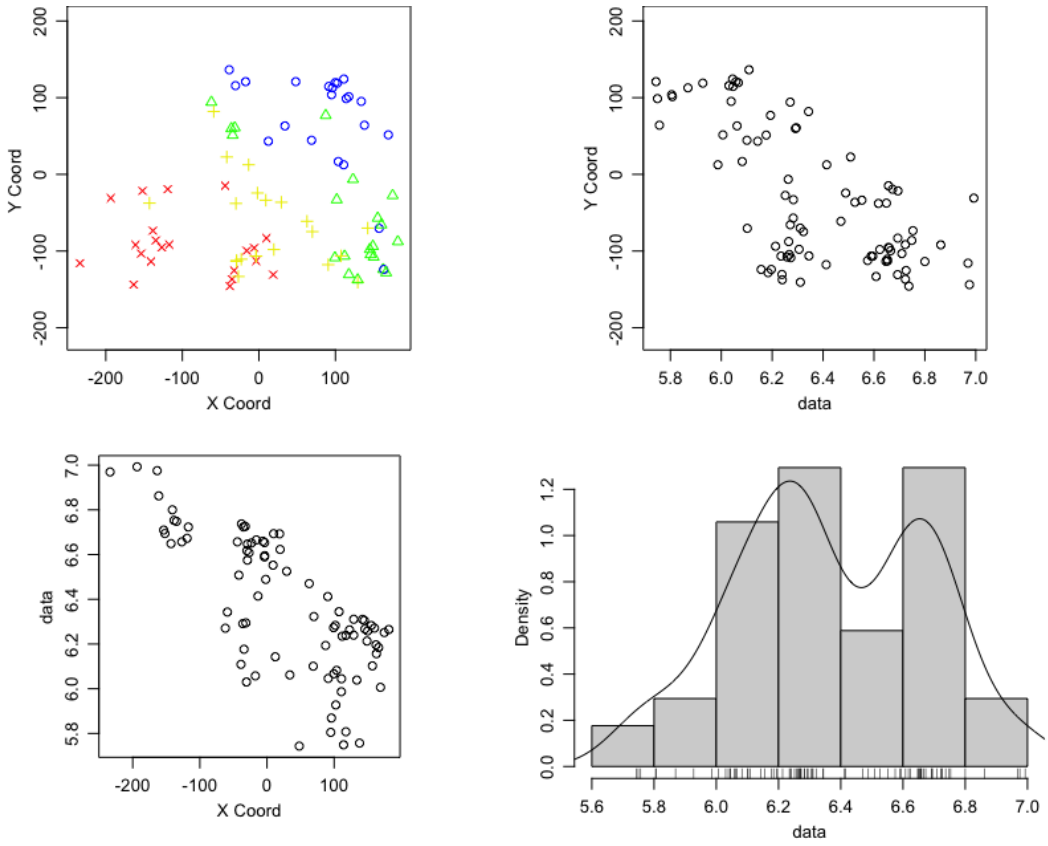
c. Describe the general features of this data set, e.g. N-S or E-W trend, if any; where are the values higher? Lower? What is the range of values in this data set?

Solution:
The first subplot(top left) shows us the distribution of measurement stations for the Wolfcamp aquifer. It shows a clear trend of increasing pressure as we move in the south west directions. The next two subplots show this trend in partial dependence plots. The final subplot(bottom right) shows a slight right skew in the distribution of pressure measurements, along with possible bimodal-ness.

d. If the data needs to be log-transformed, include the 4-panel plot you get after transforming, and discuss the resulting plots; does it look as though this is an appropriate

transformation.

Solution:
The data does appear to be right skewed and looks like it could benefit from a log transform. Doing so we get the following plot,



The distribution of the data looks to be more centered and symmetric but is still exhibiting some bimodal-ness.

Exercise 5: We sometimes use the function $f(d) = exp(-d/\lambda)$ to express the correlation between observations made at locations separated by the distance d, where $d \geq 0$. Here, λ is a positive constant, which we treat as a known constant.

a. Show that $f(d) \geq 0$ and that $f(d) \leq 1$.

Solution:
First note that the exponential function is continuous and differentiable across it's entire domain, that being the all real numbers. We also know that function $f(x) = e^x$ in particular has a range from $(0, \infty)$. Let $x = -d/\lambda$ for some positive constant λ and note that $f(d)$ has the same range as $f(x)$, therefore it follows that there exist values of d such that $f(d) > 0$ and $f(d) \leq 1$.

b. Let $\lambda = 1$. Find d so that $f(d) = .05$.

Solution:
Substituting $\lambda = 1$ to our function $f(d)$ we get, $f(d) = exp(-d)$. Solving for d, when $f(d) = .05$,

$$\begin{aligned} .05 &= e^{-d} \\ ln(.05) &= -d \\ -ln(.05) &= d \end{aligned}$$

We get that the minimum distance between observations in order for them to be almost uncorrelated is $-ln(.05) \approx 3$.

c. Let $\lambda = 5$. Find d so that $f(d) = .05$

Solution:
Substituting $\lambda = 5$ to our function $f(d)$ we get, $f(d) = exp(-d/5)$. Solving for d, when $f(d) = .05$,

$$\begin{aligned} .05 &= e^{-d/5} \\ 5ln(.05) &= -d \\ -5ln(.05) &= d \end{aligned}$$

We get that the minimum distance between observations in order for them to be almost uncorrelated is $-5ln(.05) \approx 15$.

d. True/ False explain briefly. The larger λ is, the quicker correlations die out with increased distance between observations.

Solution:
False. Increasing λ actually has the opposite effect on the slope of the function. Consider differentiating the function $f(d) = e^{d/\lambda}$. We would get, $f'(d) = \frac{-1}{\lambda} e^{d/\lambda}$. The $\frac{-1}{\lambda}$ term shows us that the slope will decrease faster as λ decreases.

e. If d is measured in km, what are the units for λ?.

Solution:
Note that we are using $f(d)$ as a measure of correlation as a function of distance. Correlation is a unitless measure. Therefore we would want λ to be in the same units as d to cancel them out.

Exercise 6: We sometimes use the function $f(d) = exp(-d\lambda)$ to express correlation between observations made at locations separated by the distance d, $d \geq 0$. Here, λ is a positive constant, which we treat as a known constant.

a. Let $\lambda = 5$. Find d so that $f(d) = .05$. Compare this with your answer from 5c.

Solution:
Substituting $\lambda = 5$ and solving for d we get,

$$\begin{aligned} .05 &= e^{-d5}, \\ ln(.05) &= -d5, \\ -\frac{ln(.05)}{5} &= d. \end{aligned}$$

Therefore we get a value for $d = -\frac{ln(.05)}{5} \approx .5$. This value is substantially less than the one given in 4c, which is as expected given our explanation in 4c and our understanding of the derivative of the exponential function.

b. If d is measured in km, what are the units for λ?

Solution:
Similarly to 4e, we want to make our measure of correlation unitless. In order to do that for the function $f(d) = exp(-d\lambda)$ λ must have units in $1/km$.

Exercise 7: Logarithms using different bases.

a. Find $log_{10}(x)$ where $x = 2, 3, 4, 5$. Using R; - signif(log10(2:5), 4)

Solution:
Using the stated R code we get the following values,

$$log_{10}(2) = .3010$$

$$log_{10}(3) = .4771$$

$$log_{10}(4) = .6021$$

$$log_{10}(5) = .6990$$

b. Find $ln(x)$ where $x = 2, 3, 4, 5$. Using R; - signif(log(2:5), 4).

Solution:
Using the stated R code we get the following values,

$$ln(2) = 0.6931$$

$$ln(3) = 1.0990$$

$$ln(4) = 1.3860$$

$$ln(5) = 1.6090$$

c. Find the ratios, $log_{10}/ln(x)$, where $x = 2, 3, 4, 5$. Using R; - signif(log10(2:5)/log(2:5), 4)

Solution:
Using the stated R code we get the $log_{10}/ln(x) = .4343$ for all values of x.

d. Find $log_{10}(e)$

Solution:
Doing so in R we get, $log_{10}(e) = 0.4342945$.

e. Find the ratios, $ln(x)/log_{10}$, where $x = 2, 3, 4, 5$. Using R; - signif(log(2:5)/log10(2:5), 4)

Solution:
Using the stated R code we get the $ln(x)/log_{10} = 2.303$ for all values of x.

f. Find ln(10)

Solution:
Doing so in R we get, $ln(10) = 2.302585$