Post 1 - Data Exploration with Star Plots

Nicholas Lai

October 26, 2017

Introduction

Star plots are a fun method to represent multivariate data in a way that's easy to view and categorize. They appeal to a basic, human appreciation of patterns and similarities in a way that a table or an index statistic just doesn't (even if those methods are often better for serious analysis).

Star plots are especially interesting when you have a list of objects described by some number of variables. The shape of an object's star plot reveals interesting properties at a glance, and similarly shaped objects can reveal relationships just as readily.

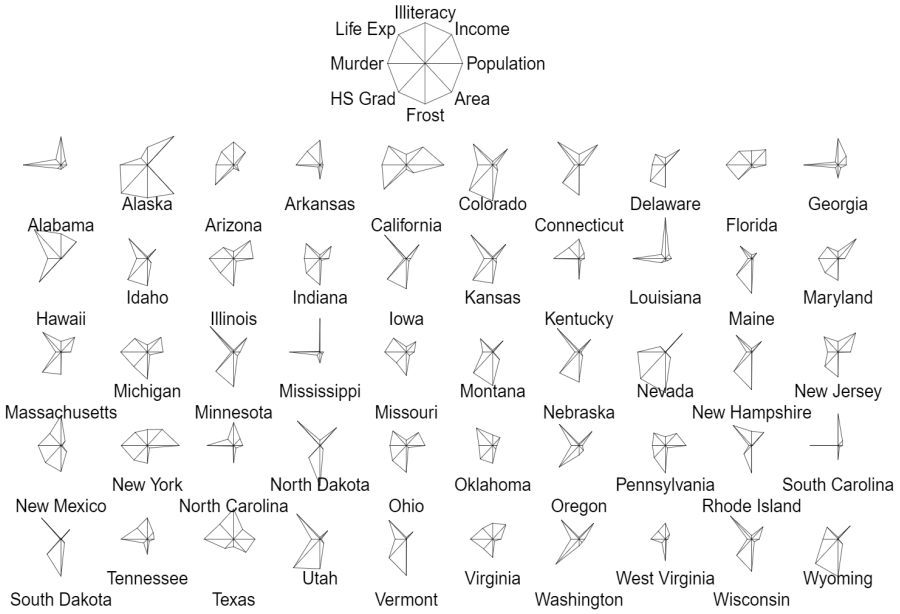
My hope is that after you finish this post, you will come to appreciate how visualizing data in star plots helps identify attributes of and similarities between objects and understand the limitations and pitfalls of using them.

Shape and Patterns in Star Plots

Shape

In star plots, variables describing a class of objects are arranged radially, and the height at each variable axis represents the value of that variable for the given object.

As mentioned in the introduction, star plots really shine when they represent objects of a certain type that are described by a number of variables common to the objects. For example, in class, we saw star plots of the 50 states described by a number of statistics about their population and climate:



The 50 States as Stars

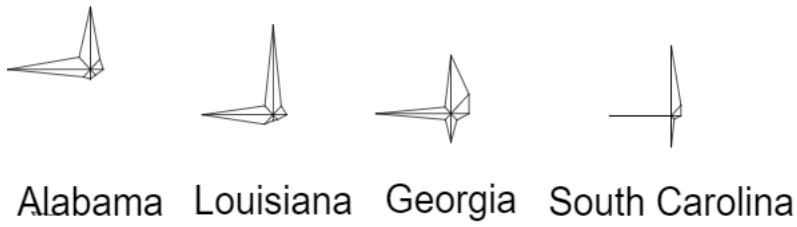
Observe the shape of the state of Mississippi:

Star Plot of Mississippi

What you notice right away is that it has a very striking mirrored L-shape. By looking at the legend above, observe that the axes that Mississippi is very high in are illiteracy and murder, both of which make the state a worse place to live in. From the shape of the graph alone, we gained valuable insight about the general properties of the object at a glance.

Patterns

Now, notice that many other states have a similar mirrored L shape:



All of these are southern states that border one another! And we got a relationship between these states just by inspection of the star plots of states. This ability to display patterns with shape is one of the huge advantages of star plots, and this makes them a great tool for data exploration.

Star Plots Example - Pokemon

When I was a kid, I was really into Pokemon, a game series in which you collect the eponymous creatures and battle other Pokemon with them. One question that a fan of the Pokemon games might ask is: which are the best Pokemon?

All Pokemon can be described by six characteristics: HP, Attack, Defense, Sp.Atk, Sp.Def, and Speed. (If you are at all interested in what these variables mean, you can read the data dictionary provided in the references link for this dataset. All you really need to know is that higher numbers in any of these variables means the Pokemon is stronger.)

There's this fun dataset from Kaggle that has these attributes (along with others):

```
#Reading in the data file into a data frame
poke_stats <- read.csv('./data/Pokemon.csv')

#Show first few rows of data
head(poke_stats)
```

```
##   X.                  Name Type.1 Type.2 Total HP Attack Defense Sp..Atk
## 1  1             Bulbasaur  Grass Poison   318 45     49      49      65
## 2  2               Ivysaur  Grass Poison   405 60     62      63      80
## 3  3              Venusaur  Grass Poison   525 80     82      83     100
## 4  3 VenusaurMega Venusaur  Grass Poison   625 80    100     123     122
## 5  4            Charmander   Fire          309 39     52      43      60
## 6  5            Charmeleon   Fire          405 58     64      58      80
##   Sp..Def Speed Generation Legendary
## 1      65    45          1     False
## 2      80    60          1     False
## 3     100    80          1     False
## 4     120    80          1     False
## 5      50    65          1     False
## 6      65    80          1     False
```

Let's trim down the size of the dataset, as this is too busy to compare with starplots visually.

```
#Filtering pokemon to generation 1 only
poke_stats %>% filter(Generation == 1) -> poke_stats
#Filter down to the last 10 pokemon
poke_stats <- poke_stats[156:166,]
```

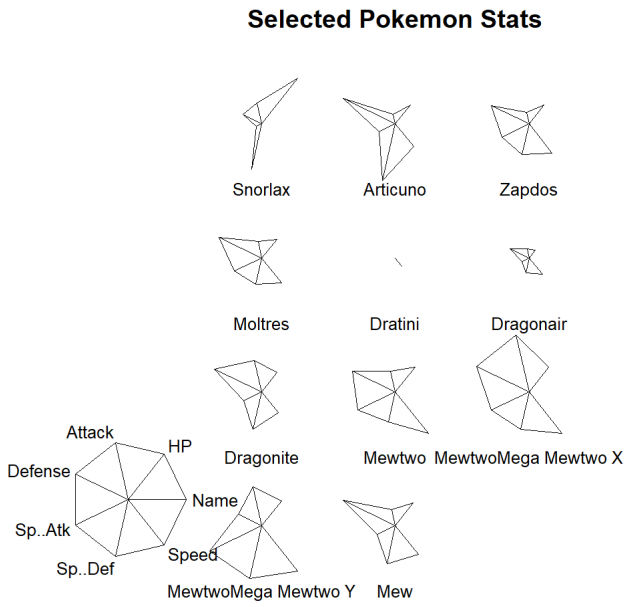
We are only interested in the stats listed above, so let's drop the other variables:

```
#Selecting numeric columns
poke_stats <- select(poke_stats, Name, HP, Attack,
                     Defense, Sp..Atk, Sp..Def, Speed)
#Show beginning of remaining data frame
head(poke_stats)
```

```
##          Name  HP Attack Defense Sp..Atk Sp..Def Speed
## 156   Snorlax 160    110      65      65     110    30
## 157  Articuno  90     85     100      95     125    85
## 158    Zapdos  90     90      85     125      90   100
## 159   Moltres  90    100      90     125      85    90
## 160   Dratini  41     64      45      50      50    50
## 161 Dragonair  61     84      65      70      70    70
```

Now, we are ready to plot the data:

```
#Make the name column a vector
poke_stats$Name <- as.vector(poke_stats$Name)
#Plot the Pokemon data as stars
stars(poke_stats, labels = poke_stats$Name,
      key.loc = c(0, 2.75),
      main = "Selected Pokemon Stats")
```



Using the previously discussed method of looking at shapes and patterns, we can quickly observe some things about the Pokemon in question.

Since we know that having higher stats makes Pokemon stronger (naively), the larger the starplot is, the better the Pokemon is! By quick observation, Dratini seems to be a weak Pokemon, while Mega Mewtwo X seems to be the strongest.

Also, notice how Dragonair and Dragonite have roughly the same shape of starplot. This is because they have an evolutionary relationship with each other (Dragonair evolves into the better Dragonite under some conditions)! So just by the shape of the star plots, we were able to see a connection between two objects and uncover a pattern.

Broader Applications

Even though the last example was about Pokemon, which you may or may not have heard of, you can apply this process to any invdividuals if you have the data about them!

For example, you could, instead, plot NBA teams with variables describing their performance, as we did in HW#3. This technique is generalizable for any such grouping!

Limitations of Star Plots

Lots of Objects

When there are too many objects to compare, looking at star plots loses much of its appeal, as it will quickly become time consuming to check each individual plot and compare them when the number of plots is large.

Directionality of Data

When using star plots, it is a best practice to make sure your variables are “good” or “bad” in the same direction. For example, in the 50 states stats from class, murder rate and illiteracy are bad when high, while income and life-expectancy are good. So, it becomes harder to interpret the star plot of the data at a glance.

Lack of Mathematical Rigor

One object's star plot having a similar shape as another object's star plot is not a mathematically rigorous argument for similarity. Star plots are helpful to eyeing a trend, not proving that one is there.

Misleading Size

Higher values of variables that happen to be next to each other on the star plot create a larger area than if those higher values were between lower values instead. This can lead to interpretation errors.

Conclusion

Star plots have plenty of shortcomings, as detailed above. But they are a visually intuitive way to represent patterns in data, and are a wonderful tool for data exploration. Even though they are not by themselves an argument for a trend between objects or abnormality of an object, they are a clue to finding those things. And they're pretty too!

References

https://github.com/ucb-stat133/stat133-fall-2017/blob/master/slides/15-principal-components1.pdf - The 50 States Example

https://www.kaggle.com/abcsds/pokemon - The Pokemon Dataset

http://www.math.yorku.ca/SCS/sugi/sugi16-paper.html#H1_5:Star - Brief discussion on directionality of data in star plots

http://www.itl.nist.gov/div898/handbook/eda/section3/starplot.htm - Brief discussion of size restriction on star plot method of visualization

http://info.slis.indiana.edu/~katy/S637-S11/cleveland84.pdf - Very important paper in the field of Data Visualization. Details how information transfer error increases with certain representations of data.

https://www.rdocumentation.org/packages/graphics/versions/3.4.0/topics/stars - Documentation of the R star plot function

http://i-ocean.blogspot.com/2008/08/clock-this.html - Example of applications of star plots