EDA with scatterplots

Peter Hoff

STAT 423

Applied Regression and Analysis of Variance

University of Washington

Things to learn

Data analysis objectives

- description
- prediction
- inference

Tools for plotting

- scatterplots
- jittering
- boxplots
- dichotomization
- color and plotting characters
- transformations
- scaling
- lines

Seattle weather data

```
weather[1:5,1:8]
```

```
##         Date Max_Temperature_F Mean_Temperature_F Min_TemperatureF
## 1 10/13/2014                71                 62               54
## 2 10/14/2014                63                 59               55
## 3 10/15/2014                62                 58               54
## 4 10/16/2014                71                 61               52
## 5 10/17/2014                64                 60               57
##   Max_Dew_Point_F MeanDew_Point_F Min_Dewpoint_F Max_Humidity
## 1              55              51             46           87
## 2              52              51             50           88
## 3              53              50             46           87
## 4              49              46             42           83
## 5              55              51             41           87
```

```
dim(weather)
```

```
## [1] 365  21
```

```
colnames(weather)
```

```
##  [1] "Date"                      "Max_Temperature_F"
##  [3] "Mean_Temperature_F"        "Min_TemperatureF"
##  [5] "Max_Dew_Point_F"           "MeanDew_Point_F"
##  [7] "Min_Dewpoint_F"            "Max_Humidity"
##  [9] "Mean_Humidity"             "Min_Humidity"
## [11] "Max_Sea_Level_Pressure_In" "Mean_Sea_Level_Pressure_In"
## [13] "Min_Sea_Level_Pressure_In" "Max_Visibility_Miles"
## [15] "Mean_Visibility_Miles"     "Min_Visibility_Miles"
## [17] "Max_Wind_Speed_MPH"        "Mean_Wind_Speed_MPH"
## [19] "Max_Gust_Speed_MPH"        "Precipitation_In"
## [21] "Events"
```

Daily min and max temperatures

Minimum temperature typically occurs near dawn.

Maximum temperature typically occurs in the afternoon.

Daily min and max temperatures

Minimum temperature typically occurs near dawn.

Maximum temperature typically occurs in the afternoon.

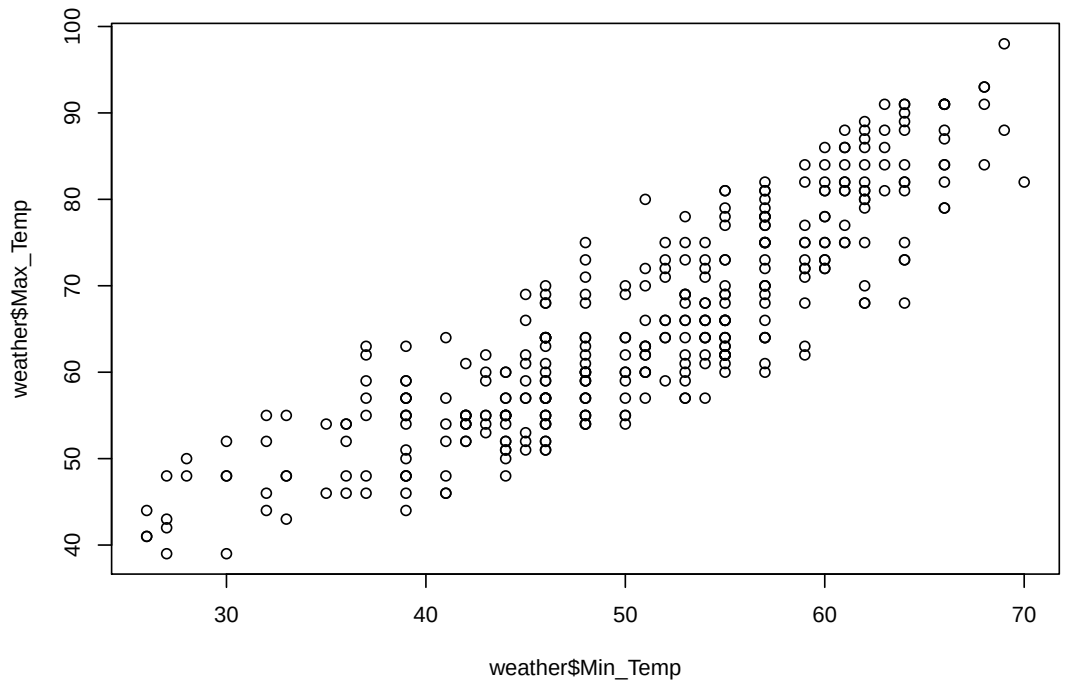
Question: For a given min temp, what is the distribution of max temps?

Daily min and max temperatures

Minimum temperature typically occurs near dawn.

Maximum temperature typically occurs in the afternoon.

Question: For a given min temp, what is the distribution of max temps?

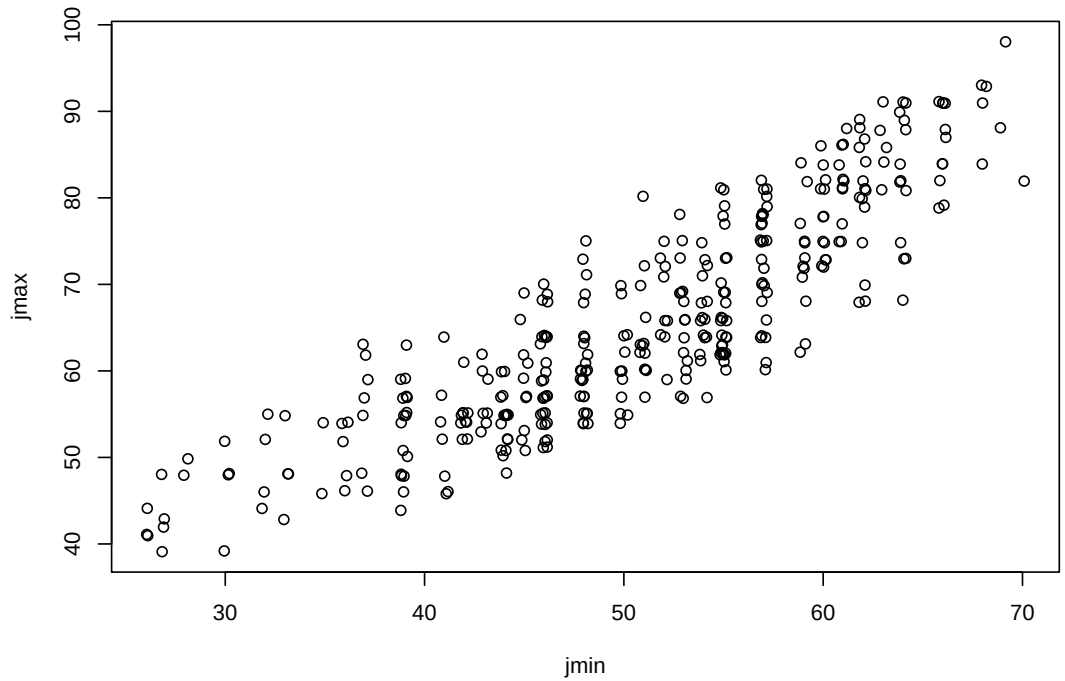


Daily min and max temperatures

Minimum temperature typically occurs near dawn.

Maximum temperature typically occurs in the afternoon.

Question: For a given min temp, what is the distribution of max temps?

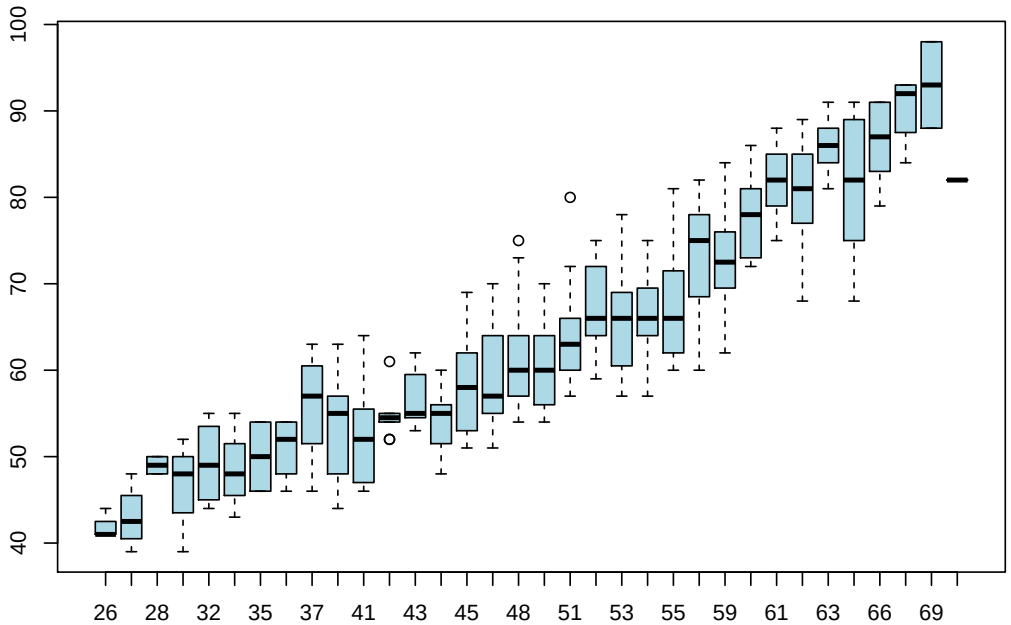


Daily min and max temperatures

Minimum temperature typically occurs near dawn.

Maximum temperature typically occurs in the afternoon.

Question: For a given min temp, what is the distribution of max temps?

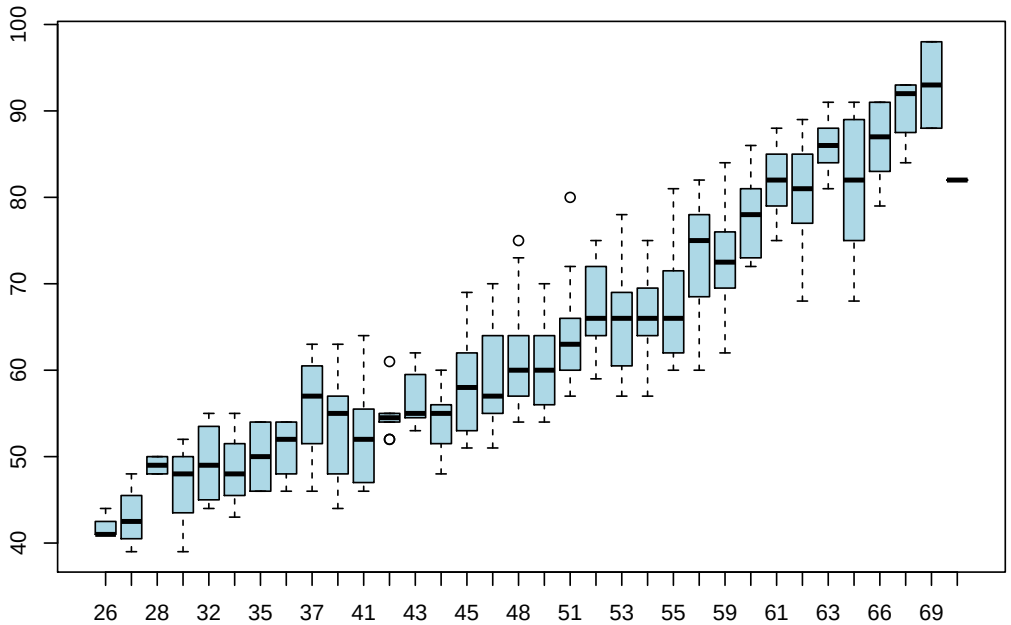


Daily min and max temperatures

Minimum temperature typically occurs near dawn.

Maximum temperature typically occurs in the afternoon.

Question: For a given min temp, what is the distribution of max temps?



scatterplot, jittering, boxplots

Data analysis questions

Descriptive:

- What are the sample means and sd's of min and max temps?
- What is the sample correlation between min and max temps?

Data analysis questions

Descriptive:

- What are the sample means and sd's of min and max temps?
- What is the sample correlation between min and max temps?

Predictive:

- Given a min temp, what is the expected max temp?
- Given a min temp, what is the range of likely max temps?
- Given a min temp, what is the probability that max temp $> t$?

Data analysis questions

Descriptive:

- What are the sample means and sd's of min and max temps?
- What is the sample correlation between min and max temps?

Predictive:

- Given a min temp, what is the expected max temp?
- Given a min temp, what is the range of likely max temps?
- Given a min temp, what is the probability that max temp $> t$?

Inferential:

- Is the expected max temp linear in min temp (in this range)?
- If so, what is the linear relationship?
- Is the variance of max temp constant across levels of min temp?

Data analysis questions

Descriptive:

- What are the sample means and sd's of min and max temps?
- What is the sample correlation between min and max temps?

Predictive:

- Given a min temp, what is the expected max temp?
- Given a min temp, what is the range of likely max temps?
- Given a min temp, what is the probability that max temp $> t$?

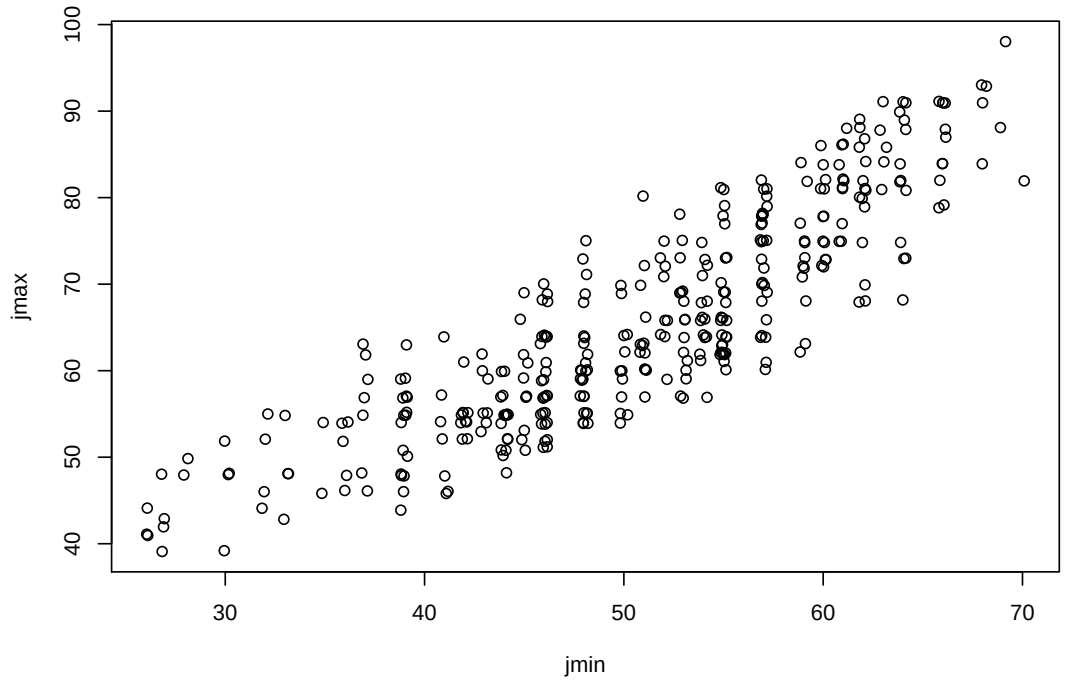
Inferential:

- Is the expected max temp linear in min temp (in this range)?
- If so, what is the linear relationship?
- Is the variance of max temp constant across levels of min temp?

description, prediction, inference

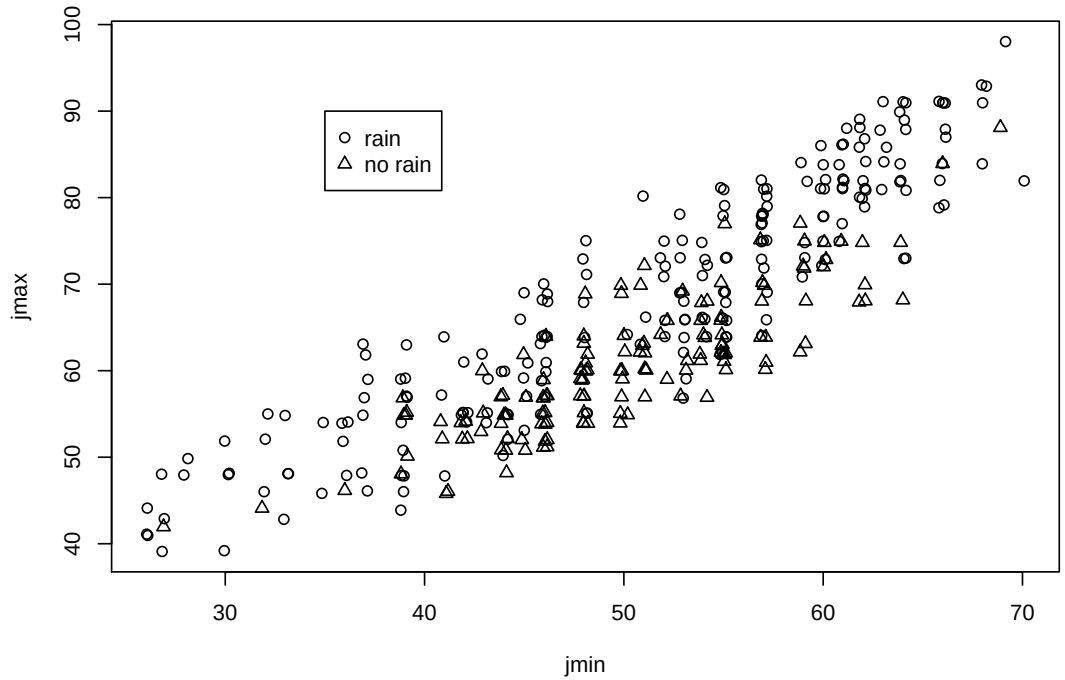
Rain and temperature

What is the effect of rain?



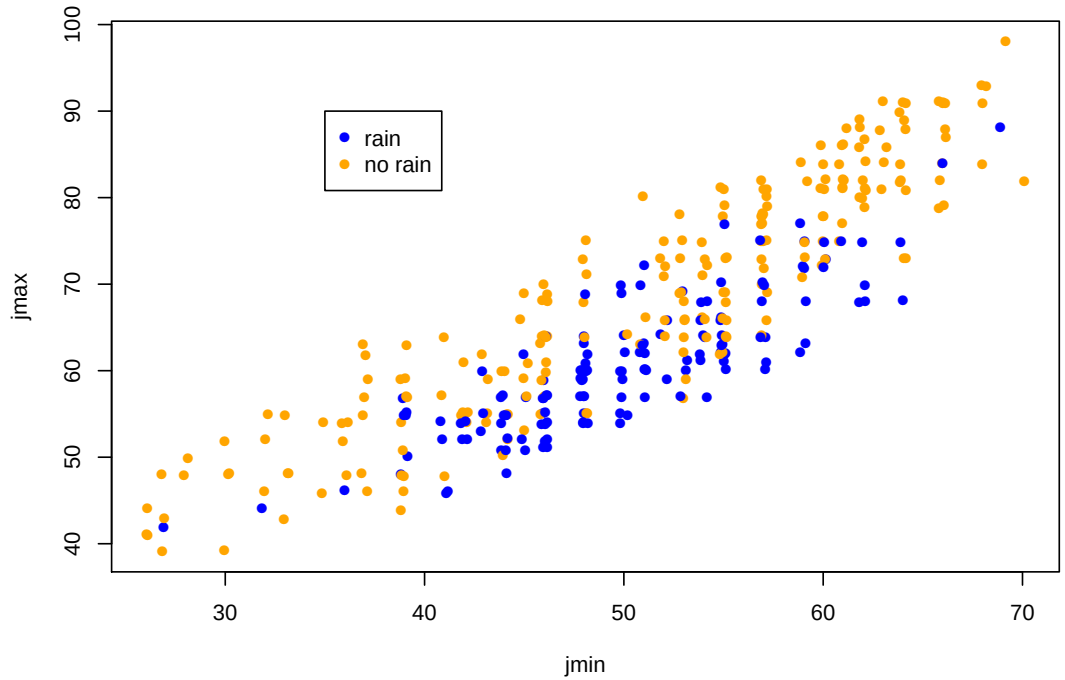
Rain and temperature

What is the effect of rain?



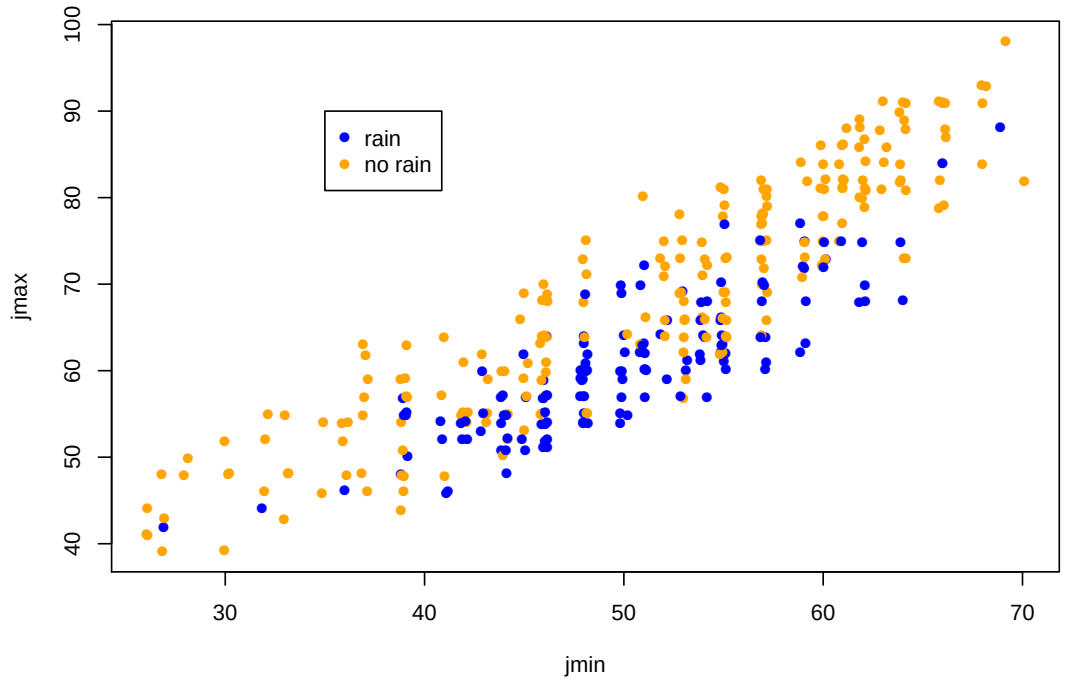
Rain and temperature

What is the effect of rain?



Rain and temperature

What is the effect of rain?



dichotomization, plotting characters, plotting colors

Data analysis questions

Descriptive:

- How does rain change sample means, sds and correlation?

Data analysis questions

Descriptive:

- How does rain change sample means, sds and correlation?

Predictive:

- What are the likely values of max temp, given min temp and rain?

Data analysis questions

Descriptive:

- How does rain change sample means, sds and correlation?

Predictive:

- What are the likely values of max temp, given min temp and rain?

Inferential:

- Is there an effect of rain on the relationship between min and max temp?

Bike share data

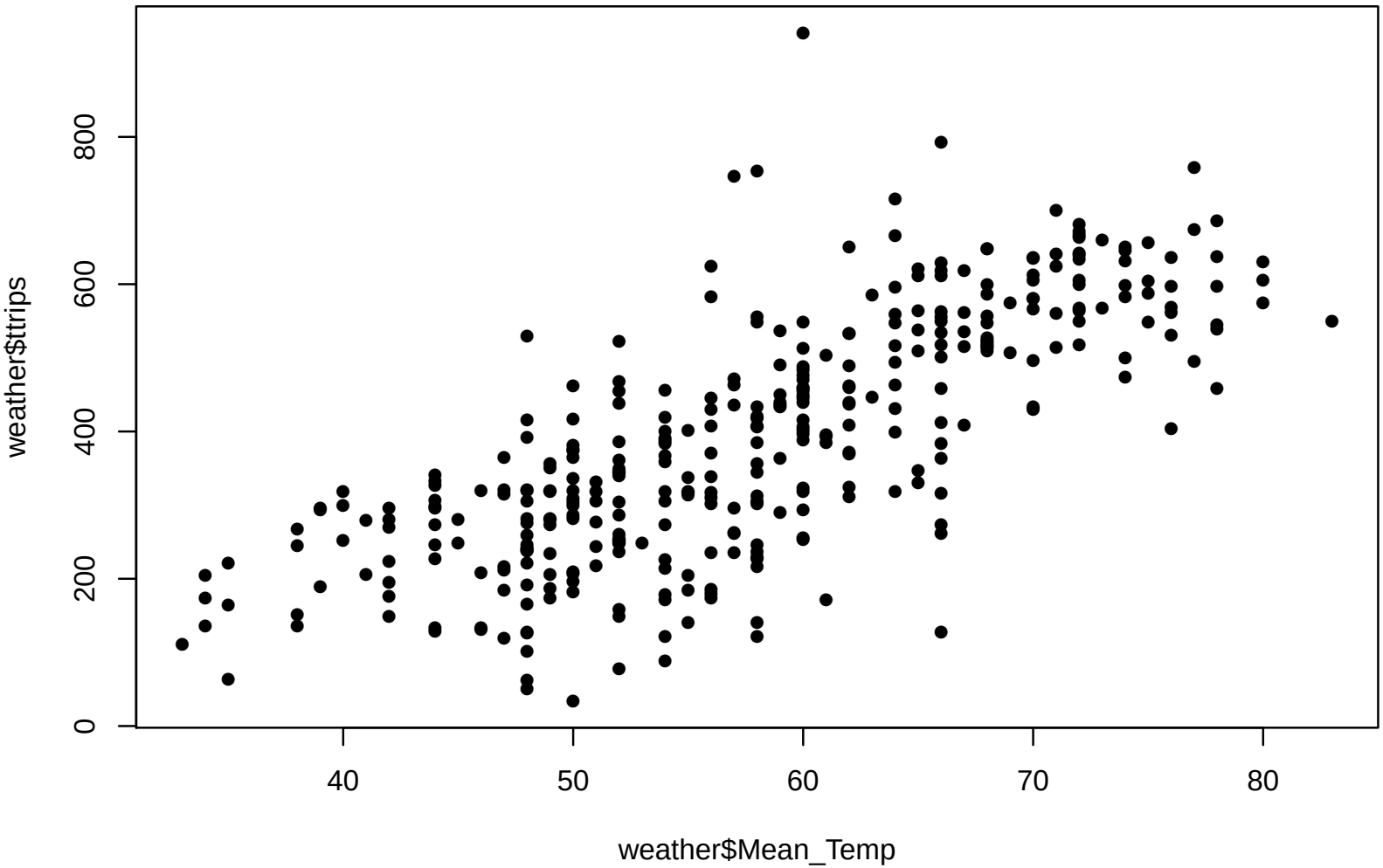
Pronto bike share data:

https://www.prontocycleshare.com/datachallenge

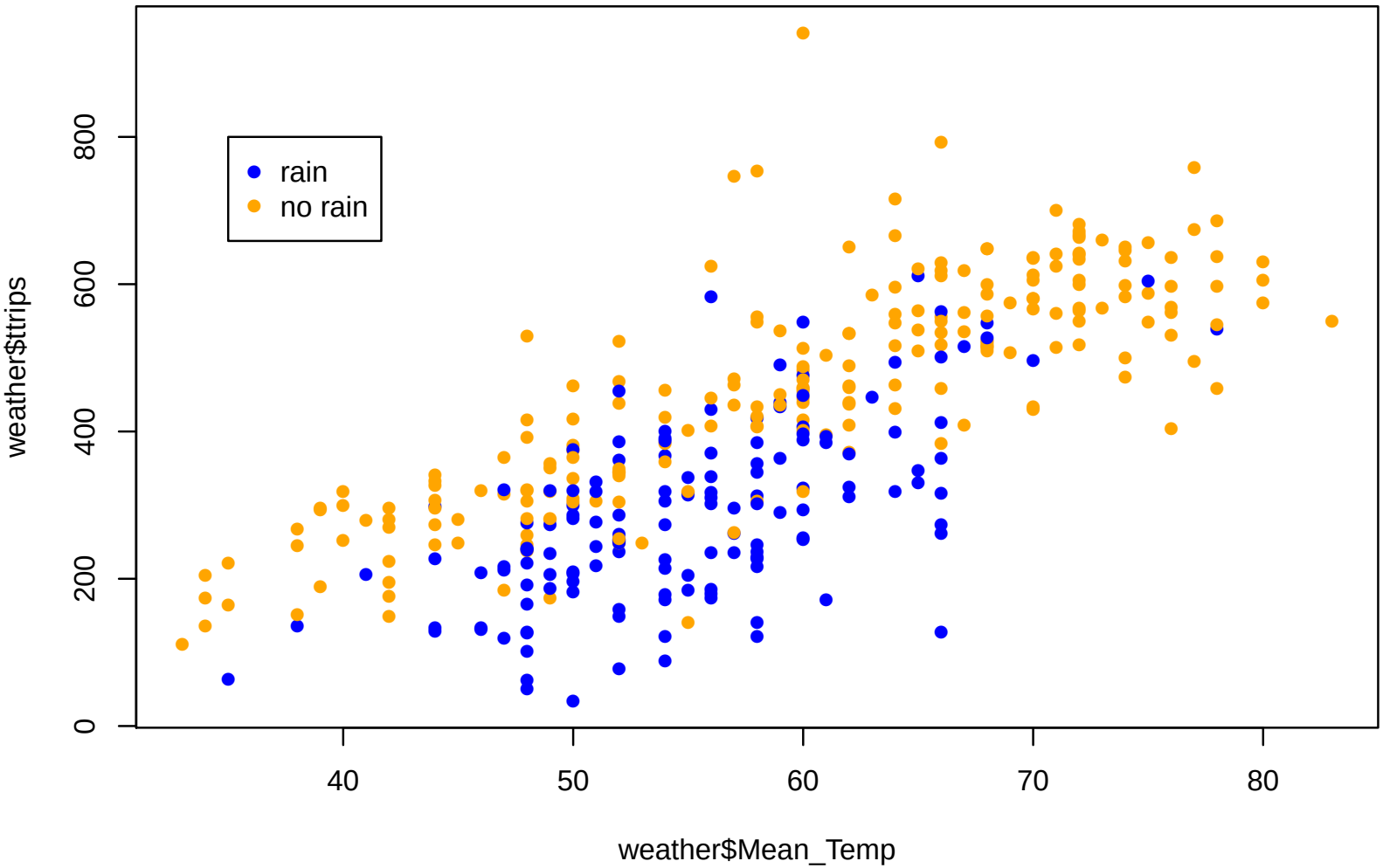
```
weather[1:10,c("Date","Mean_Temperature_F","Precipitation_In","ttrips")]
```

```
##          Date Mean_Temperature_F Precipitation_In ttrips
## 1  10/13/2014                 62             0.00    409
## 2  10/14/2014                 59             0.11    491
## 3  10/15/2014                 58             0.45    313
## 4  10/16/2014                 61             0.00    395
## 5  10/17/2014                 60             0.14    294
## 6  10/18/2014                 64             0.31    399
## 7  10/19/2014                 64             0.00    666
## 8  10/20/2014                 60             0.44    389
## 9  10/21/2014                 58             0.10    357
## 10 10/22/2014                 58             1.43    141
```

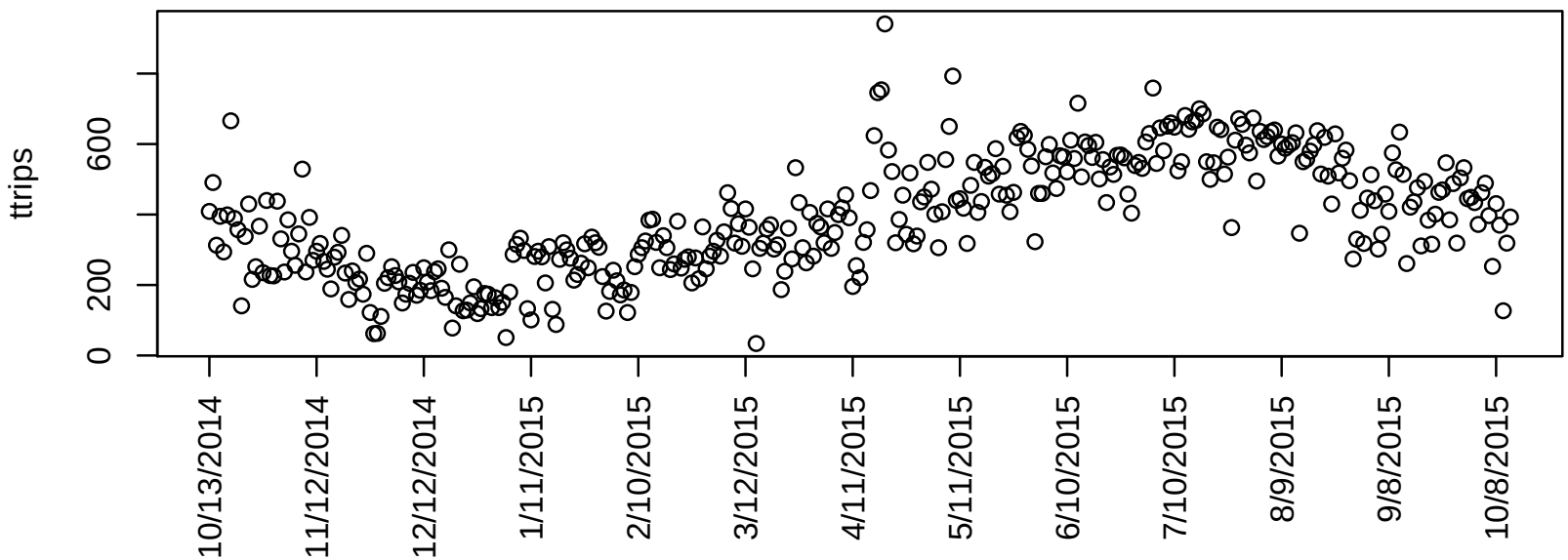
Bike share data



Bike share data



Bike share data



Data analysis questions

Descriptive:

- What are the sample means, sds and correlations between trips and temp?
- How does the rainy day data look compared to the sunny day data?

Data analysis questions

Descriptive:

- What are the sample means, sds and correlations between trips and temp?
- How does the rainy day data look compared to the sunny day data?

Predictive:

- Given mean temp, what is the range of likely number of trips?
- Given mean temp and rain, what is the range of likely number of trips?

Data analysis questions

Descriptive:

- What are the sample means, sds and correlations between trips and temp?
- How does the rainy day data look compared to the sunny day data?

Predictive:

- Given mean temp, what is the range of likely number of trips?
- Given mean temp and rain, what is the range of likely number of trips?

Inferential:

- Is the expected number of trips linear in the mean temp?
- Is there an effect of rain on the expected number of trips?

Data analysis questions

Descriptive:

- What are the sample means, sds and correlations between trips and temp?
- How does the rainy day data look compared to the sunny day data?

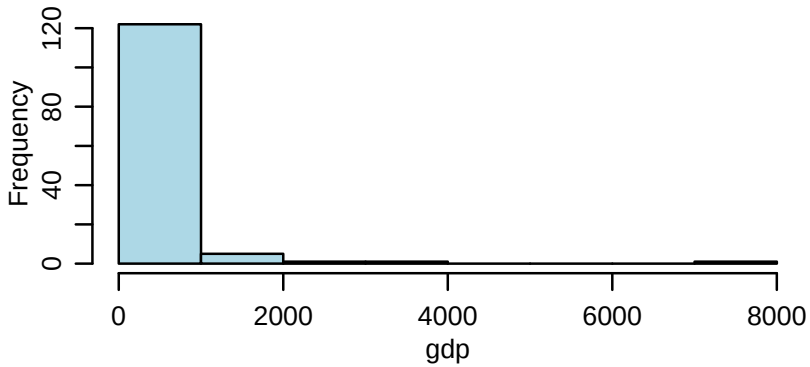
Predictive:

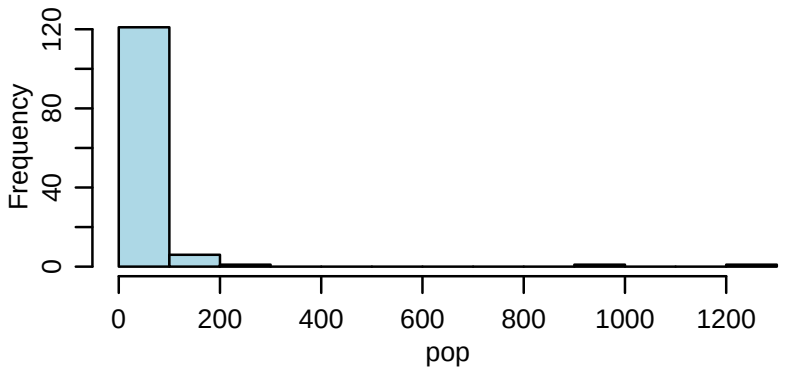
- Given mean temp, what is the range of likely number of trips?
- Given mean temp and rain, what is the range of likely number of trips?

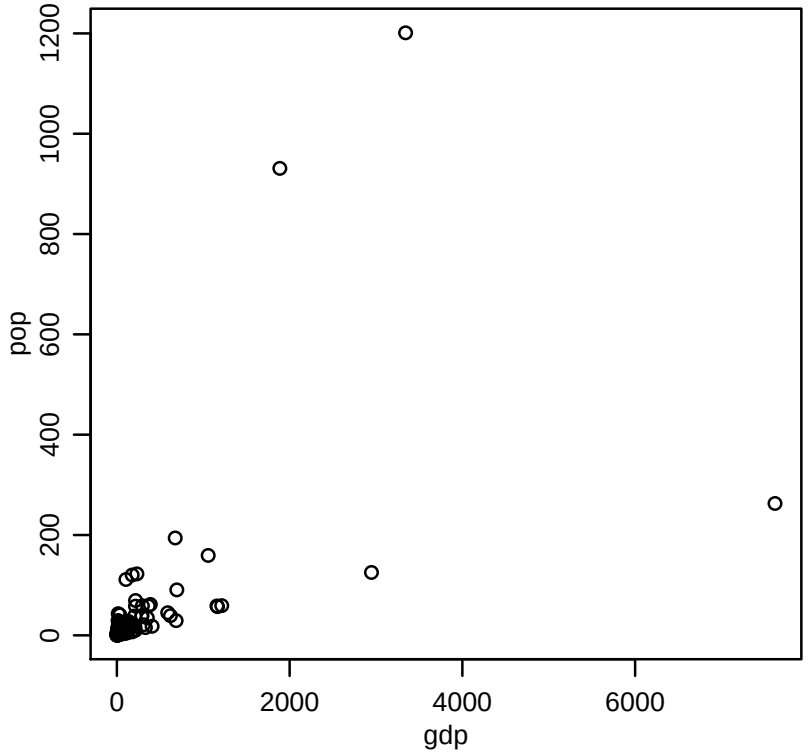
Inferential:

- Is the expected number of trips linear in the mean temp?
- Is there an effect of rain on the expected number of trips?

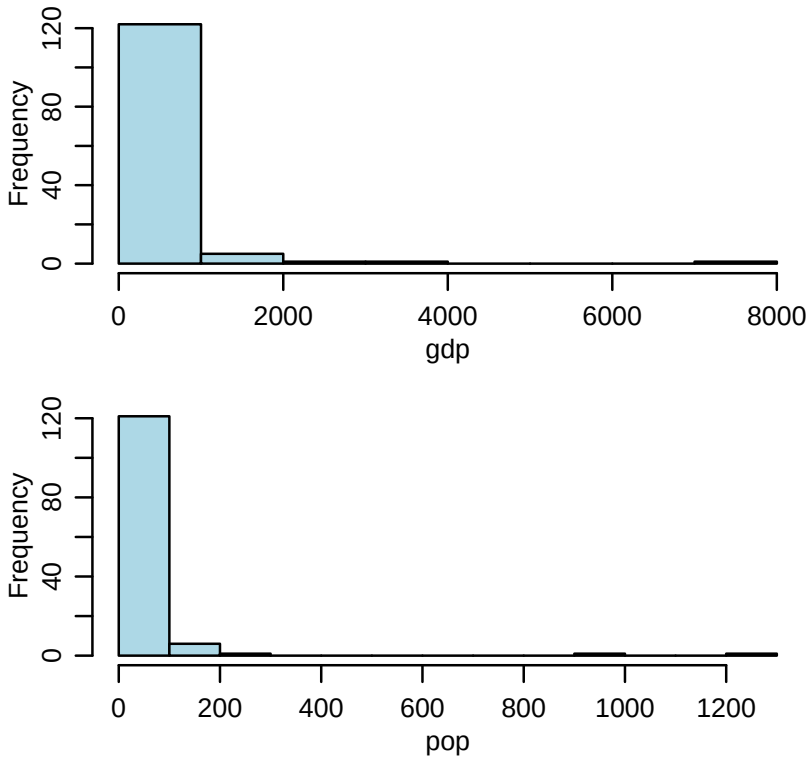
GDP, population and polity

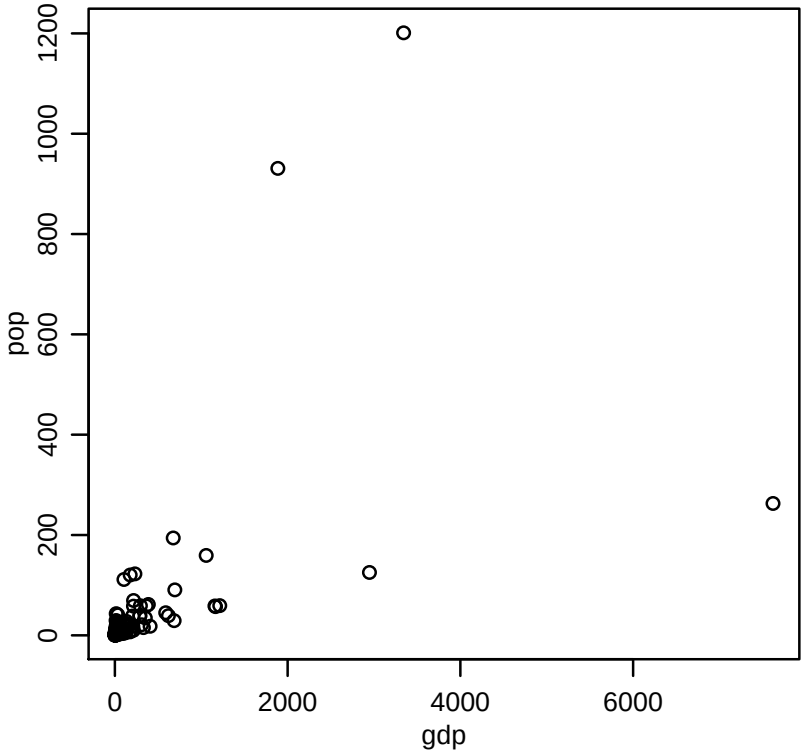






GDP, population and polity





Why is this not a good plot?

GDP and population

Q: How can we visualze all of the data?

GDP and population

Q: How can we visualze all of the data?

A: Transform the data - shrink large values more than small values.

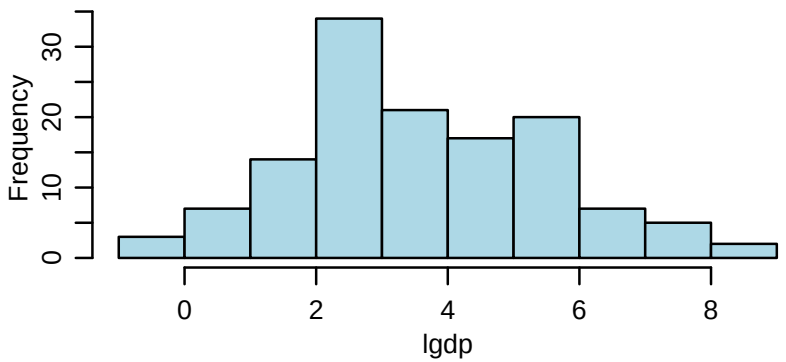
```
lpop<-log(pop)
lgdp<-log(gdp)
```

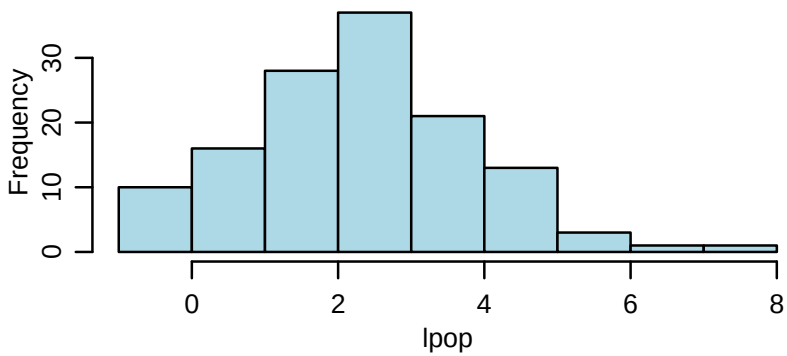
GDP and population

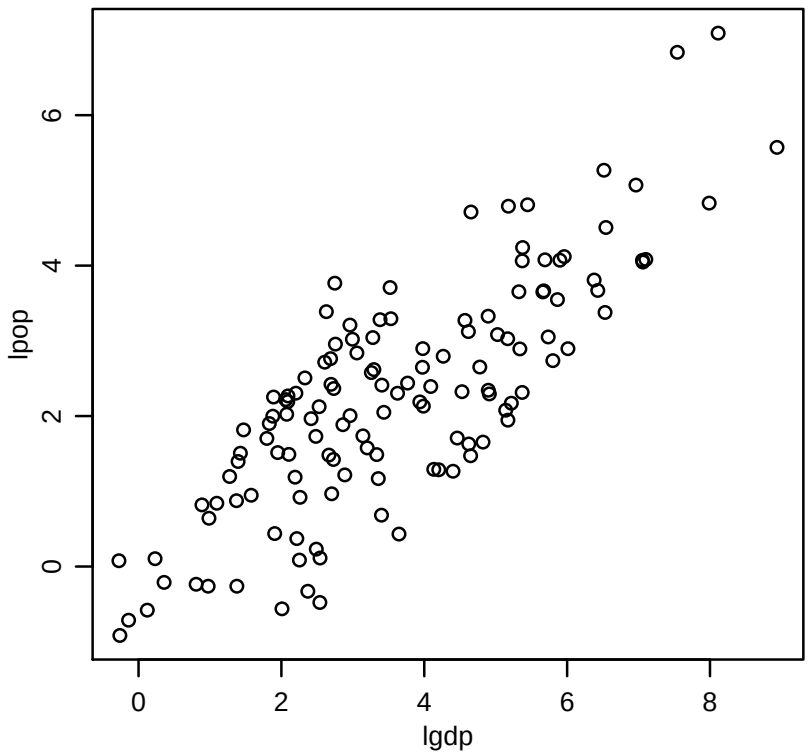
Q: How can we visualze all of the data?

A: Transform the data - shrink large values more than small values.

```
lpop<-log(pop)
lgdp<-log(gdp)
```



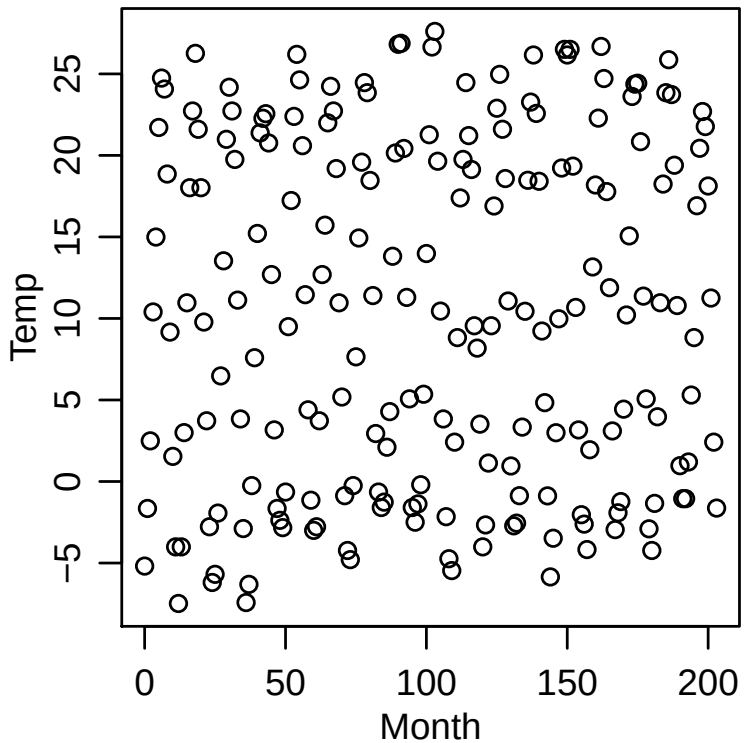




transformations

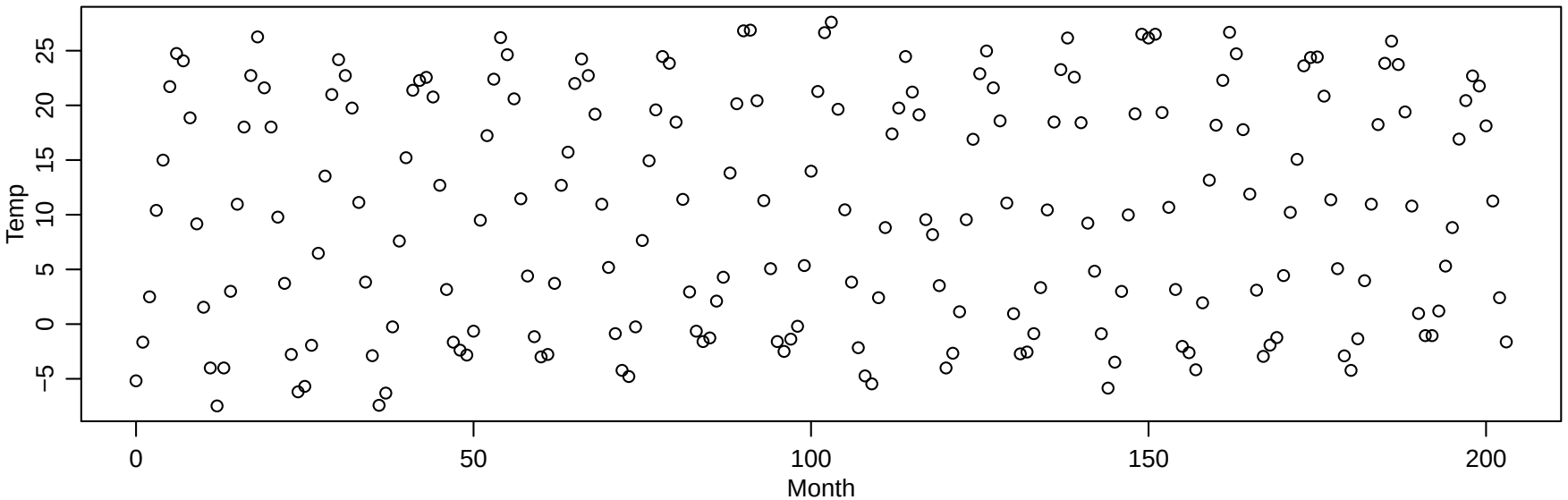
Soil temperature

Monthly soil temperature in Mitchell, Nebraska (1976-1992)



Soil temperature

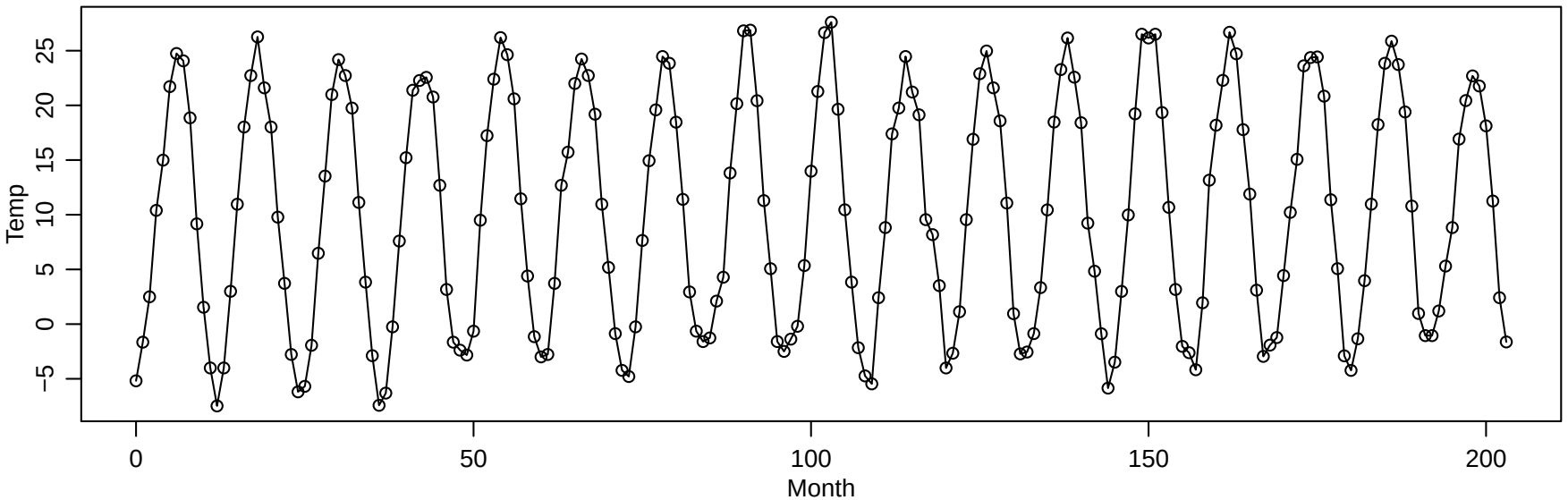
Monthly soil temperature in Mitchell, Nebraska (1976-1992)



scaling

Soil temperature

Monthly soil temperature in Mitchell, Nebraska (1976-1992)



scaling , lines

What we learned

Data analysis objectives

- description
- prediction
- inference

Tools for plotting

- scatterplots
- jittering
- boxplots
- dichotomization
- color and plotting characters
- transformations
- scaling
- lines