NYU Center for Data Science

DS-UA 112
Introduction to Data Science

Week 6: Lecture 1
Charts – Visualizations of Datasets

NYU Center for Data Science

How can we generate understandable images from tables?

DS-UA 112
Introduction to Data Science

Week 6: Lecture 1
Charts – Visualizations of Datasets

Adapted from Nolan, Cleveland, Dalessandro

Announcements

- Please check Week 6 agenda on NYU Classes
  - Homework 3
    - Question 3
  - Lab 5
- Remember to post to Piazza

See @240 on Piazza for more information about Question 3a

Announcements

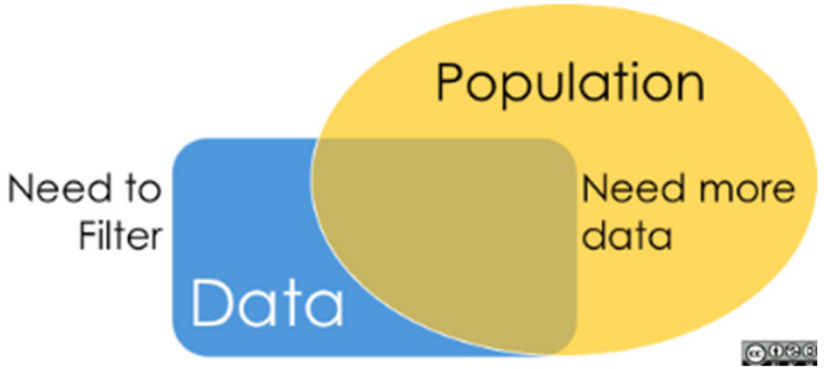
- Midterm
  - Wednesday March 11
    4:55-6:10pm
  - Reference Sheet
- Practice
  - Exam
  - Problems
  - Examples

Bring your reference sheet along with our reference sheet

Review

Recording any changes to the data helps track its properties

- Properties of Data in Tables
  - Qualitative and Quantitative
  - Scope
  - Granularity
  - Temporality
  - Faithfulness



Review

Choose the type of each variable

	Quantitative Continuous	Quantitative Discrete	Qualitative Nominal	Qualitative Ordinal
Number of Siblings				
Presidential Approval Rating				
CO2 Level				
Weight				
Borough of NYC				
Year in College				
Zipcode				
Crime (Felony, Violation, Misdemeanor)				

Review

Imputation means replacing missing data with substitute data

- Missing Values
  - Replace with average
  - Replace with values from another dataset
  - Replace with random values
  - Replace with last present value
  - Drop the records

16	Jeremy	Male	9/21/2010	5:56 AM
17	Shawn	Male	12/7/1986	7:45 PM
18	Diana	Female	10/23/1981	10:27 AM
19	Donna	Female	7/22/2010	3:48 AM
20	Lois	NaN	4/22/1995	7:18 PM
21	Matthew	Male	9/5/1995	2:12 AM
22	Joshua	NaN	3/8/2012	1:58 AM
23	NaN	Male	6/14/2012	4:19 PM
24	John	Male	7/1/1992	10:08 PM

Review

Remember to use the exclamation point in cells with commands

- File Size
  - ls: list contents of folder
  - cat: display contents of file
  - head: show first ten lines
  - tail: show last ten lines
  - du: measure size of file
  - wc: count number of lines
  - file: display file type

```
!du -sh data
```

```
28K     data
```

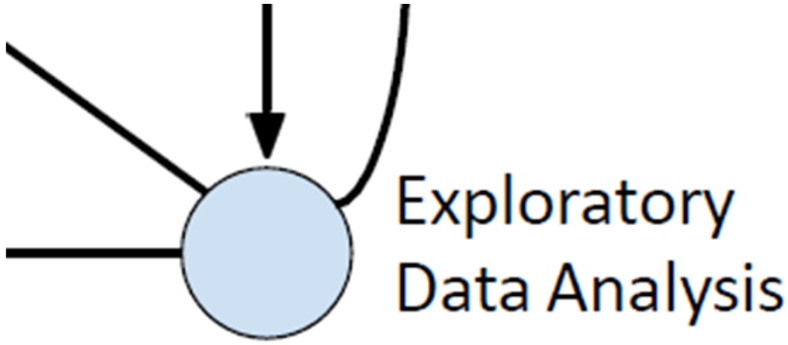
```
!du -sh data/*
```

```
12K     data/more_data
4.0K    data/movies_100_rows.csv
4.0K    data/movies.csv
```

Review

DEMO

- Exploratory Data Analysis
  - Look at data and descriptions of the columns.
  - Examine columns. Examine groups of related rows.
  - Visualize and summarize data. Apply transformations.
  - Validate assumptions about data. Identify and address inaccuracies.
  - Record all changes



Review

DEMO

- Explora
  - Loo
    of
  - E
    g
  -
  - Validate ass
    data. Identify and ad
    inaccuracies.
  - Record all changes

```
# RECIPROCITY
#
# Use of these data implies an agreement to reciprocate.
# Laboratories making similar measurements agree to make their
# own data available to the general public and to the scientific
# community in an equally complete and easily accessible form.
# Modelers are encouraged to make available to the community,
# upon request, their own tools used in the interpretation
# of the ESRL data, namely well documented model code, transport
# fields, and additional information necessary for other
# scientists to repeat the work and to run modified versions.
# Model availability includes collaborative support for new
# users of the models.
```

ysis

Agenda

- Types of Charts
  - Single Variable
  - Multiple Variables
- Approaches to Visualizations
  - Scale
  - Colors
  - Conditioning
  - Transformations

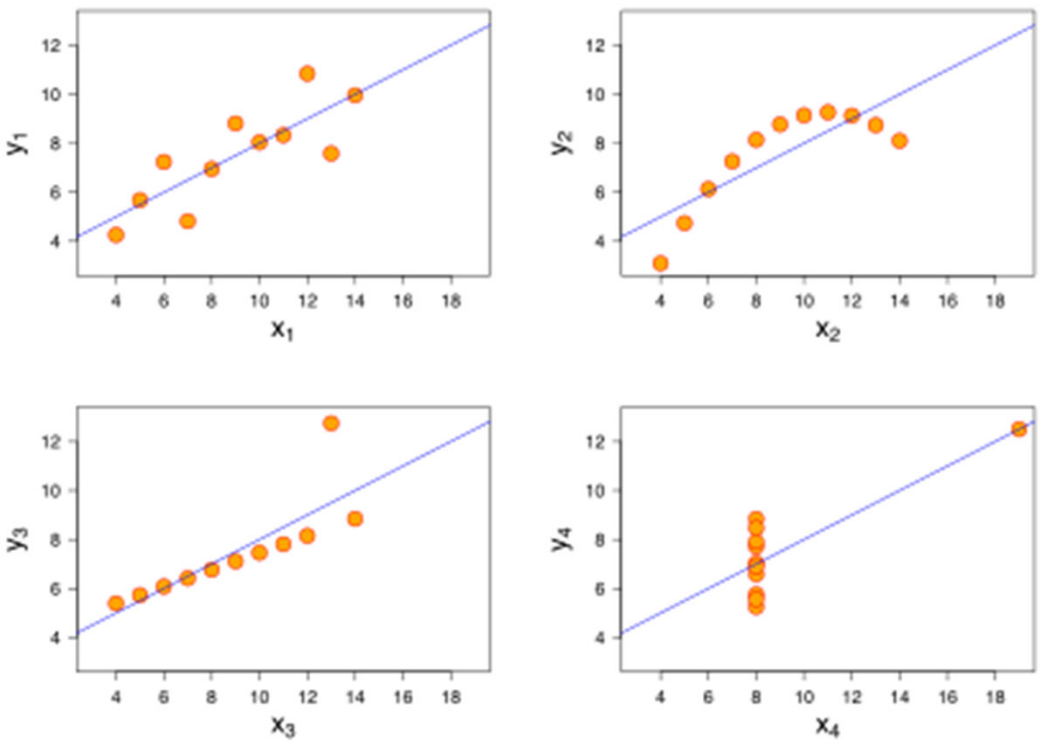
References

- Nolan, Lau, Gonzalez (Chapter 6)

Why Visualization?

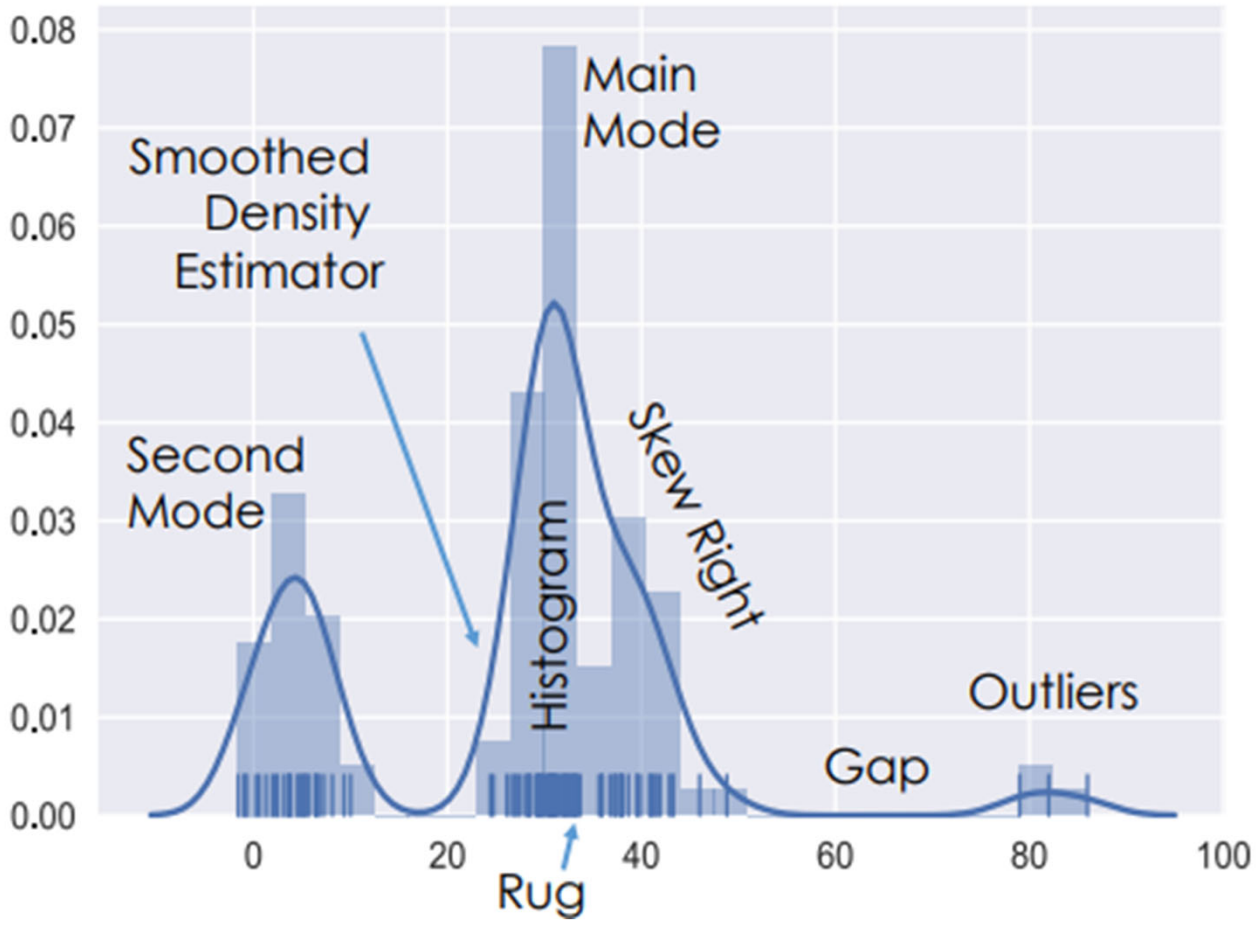
- We have different approaches to summarizing tables
- Numbers like average and standard derivation sometime omit details
- Visualizations can help us to quantify information

Property	Value
Mean of x	9
Sample variance of x	11
Mean of y	7.50
Sample variance of y	4.125



Why Visualization?

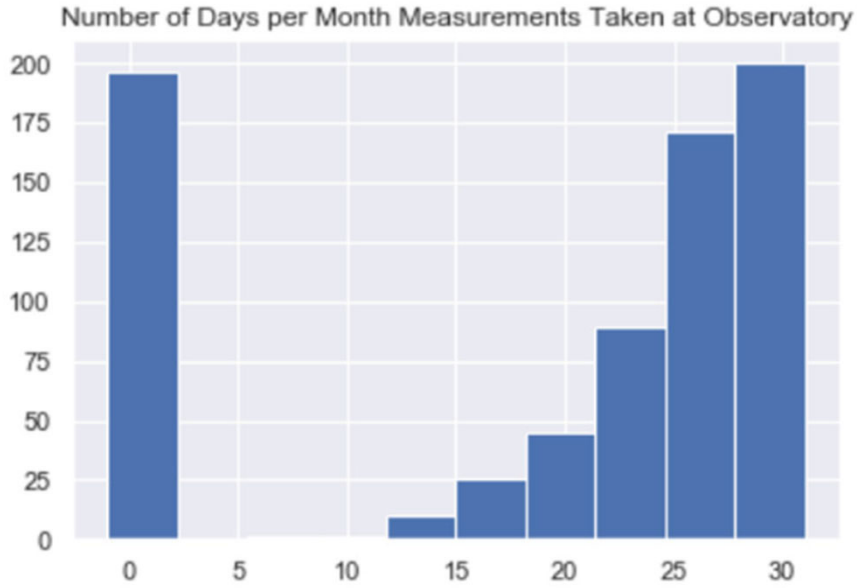
- We describe properties of numbers in charts much like we describe properties of numbers in tables



Exercise

What are the functions for these charts in seaborn?

- Which of the following plots can be used to depict a single qualitative variable?
  1. ~~histograms~~
  2. bar chat
  3. box plots
  4. scatter plots
  5. line chart

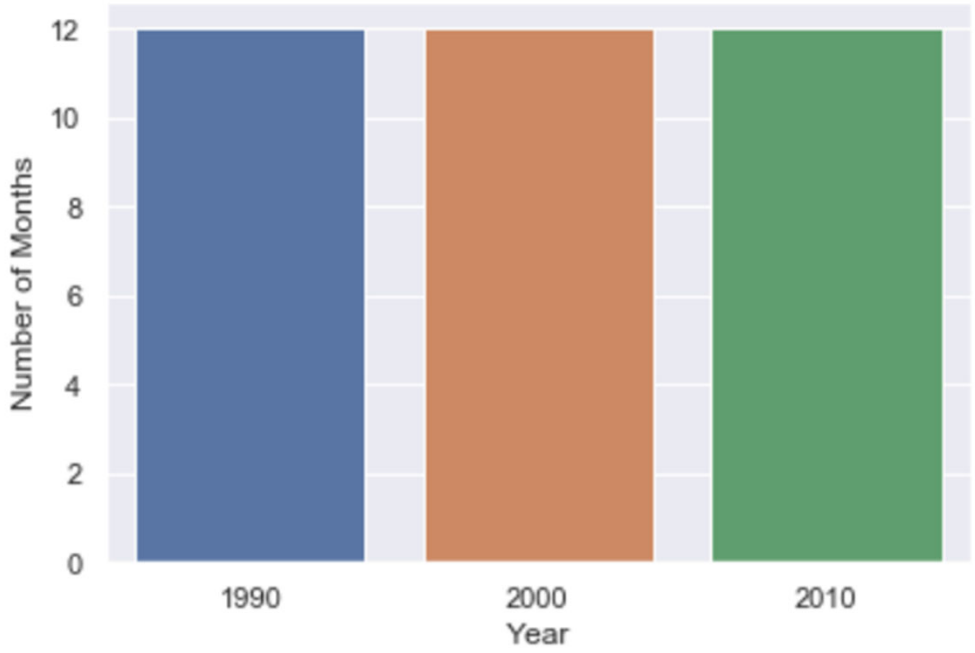


Exercise

What are the functions for these charts in seaborn?

- Which of the following plots can be used to depict a single qualitative variable?

1. histograms
2. bar chat
3. box plots
4. scatter plots
5. line chart

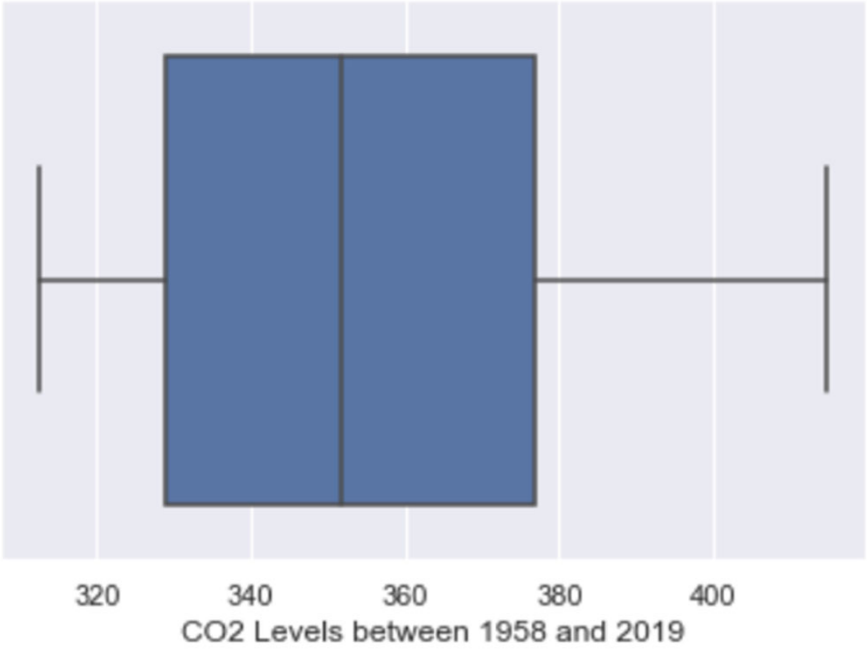


Exercise

What are the functions for these charts in seaborn?

- Which of the following plots can be used to depict a single qualitative variable?

1. histograms
2. bar chat
3. ~~box plots~~
4. scatter plots
5. line chart

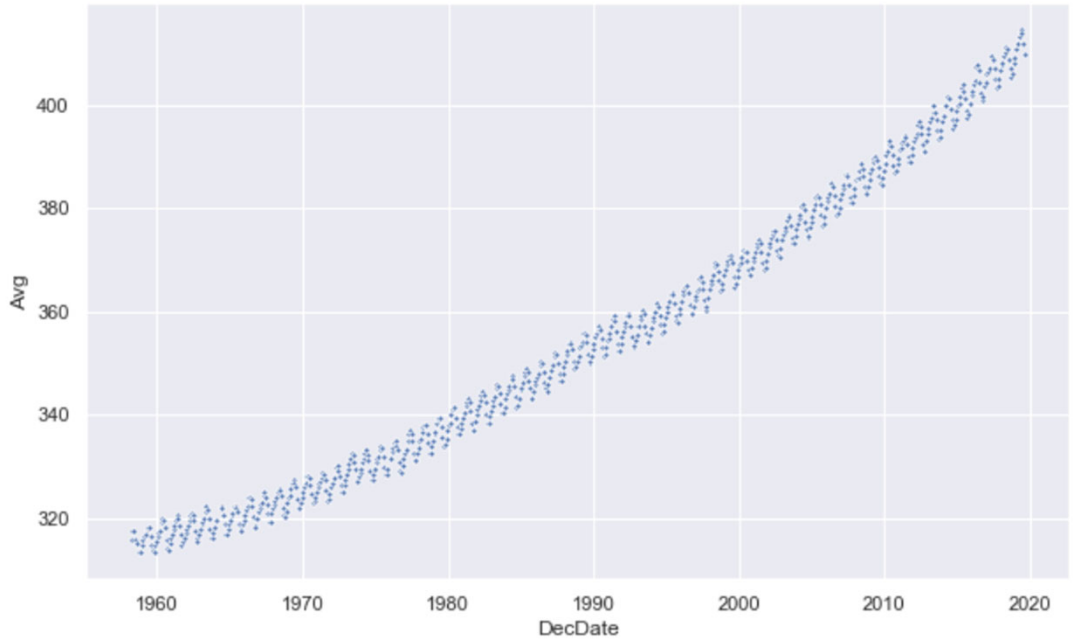


Exercise

What are the functions for these charts in seaborn?

- Which of the following plots can be used to depict a single qualitative variable?

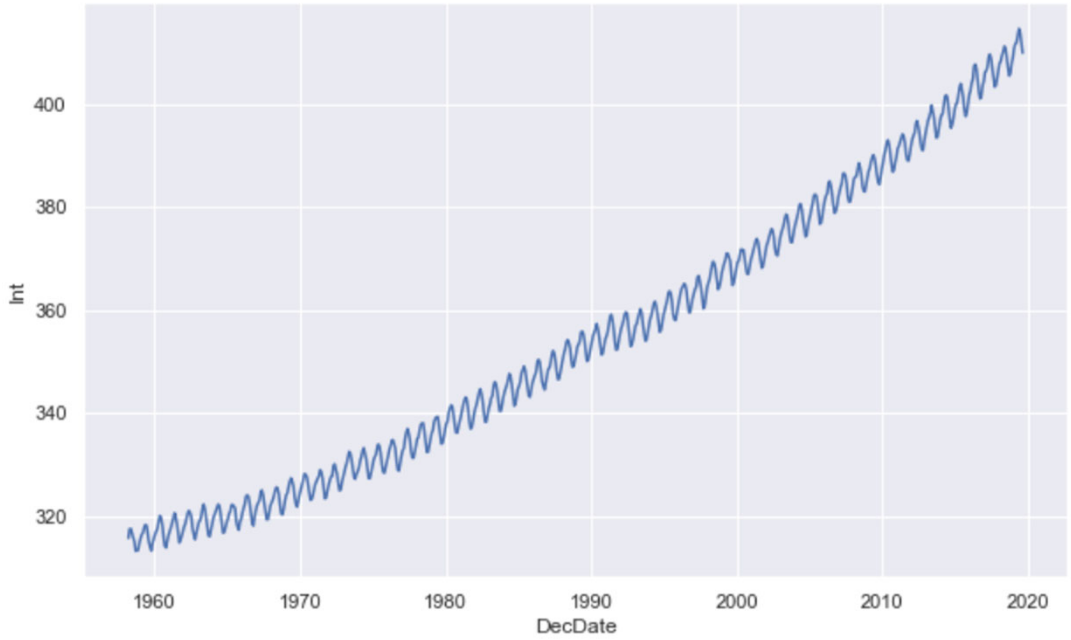
1. histograms
2. bar chat
3. box plots
4. ~~scatter plots~~
5. line chart



Exercise

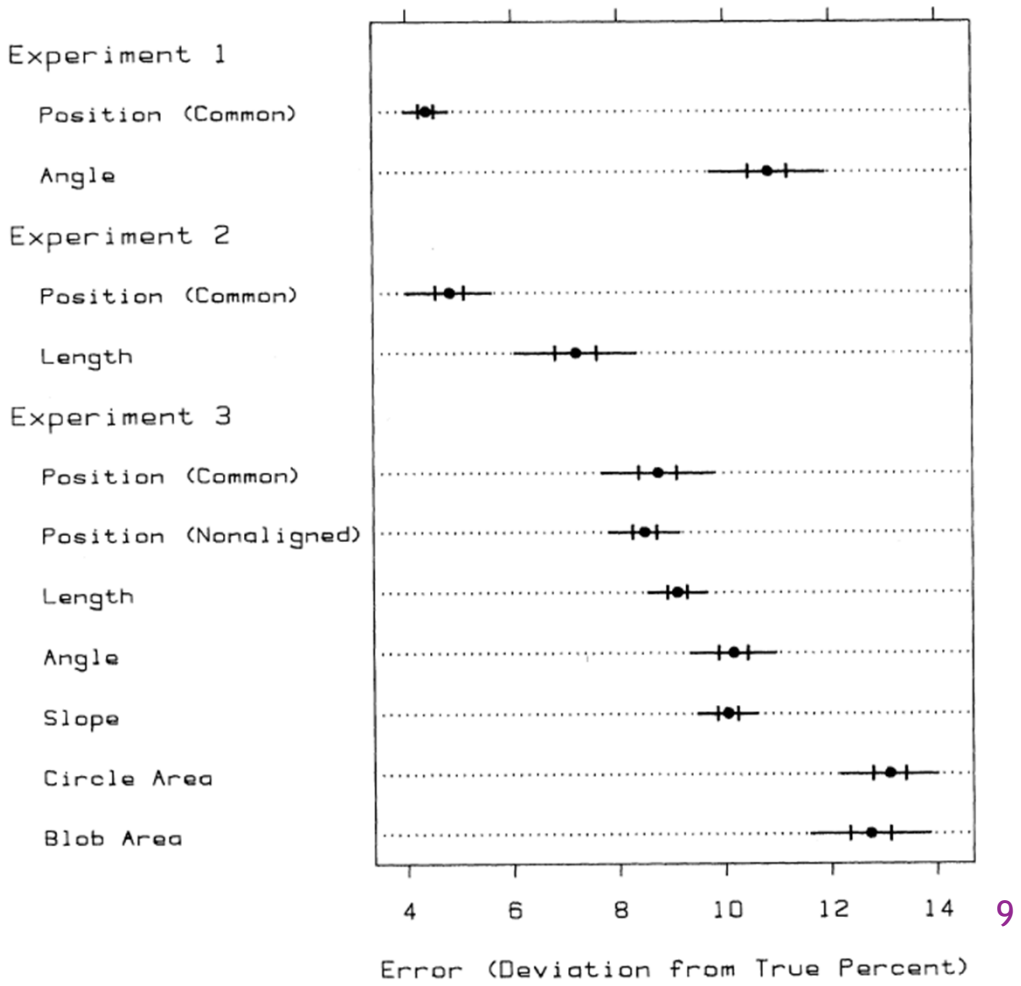
What are the functions for these charts in seaborn?

- Which of the following plots can be used to depict a single qualitative variable?
  1. histograms
  2. bar chat
  3. box plots
  4. scatter plots
  5. ~~line chart~~



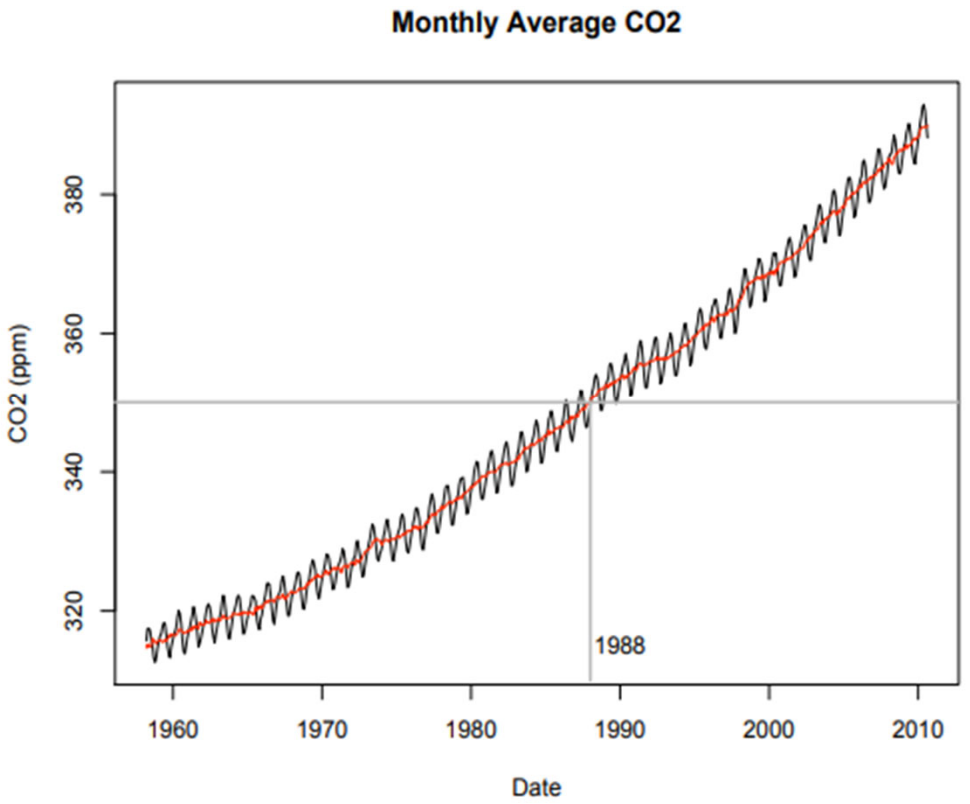
Perception

- Some aspects are more understandable than others depending on accuracy of perception
  1. Position
  2. Length
  3. Angle
  4. Area
  5. Shading



Perception

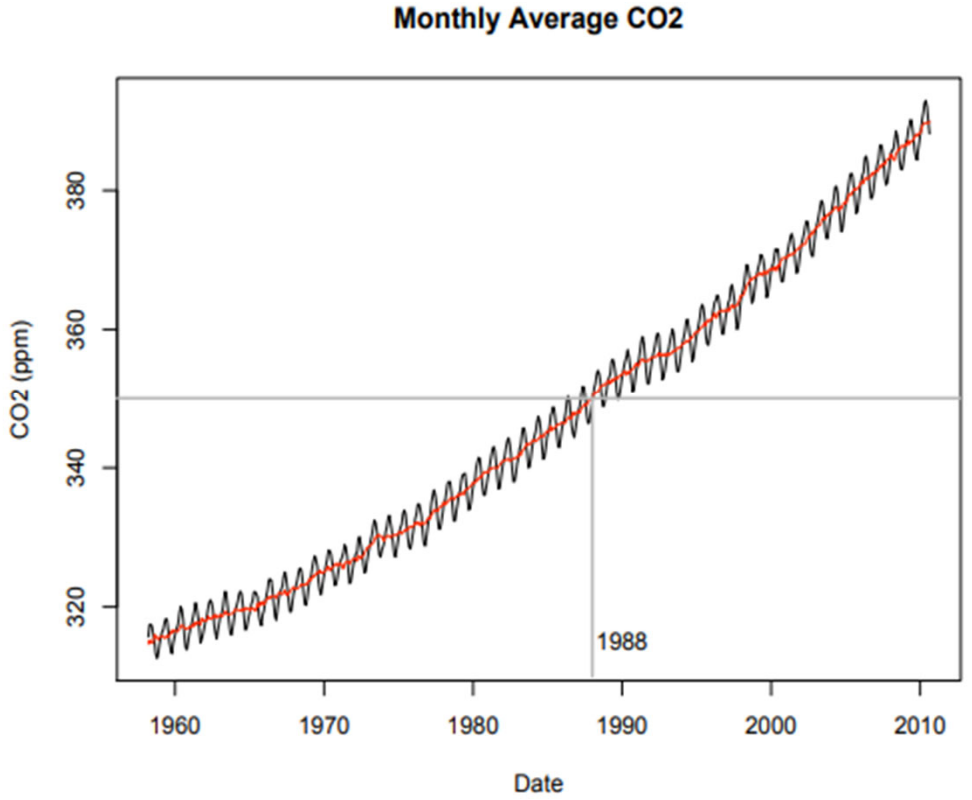
- Questions
  - What are the differences in values?
    - What are the differences in rate of change?
  - What is the scope of the data?
    - How would the numbers differ at other observatories?



Single Chart

What are the differences in values?

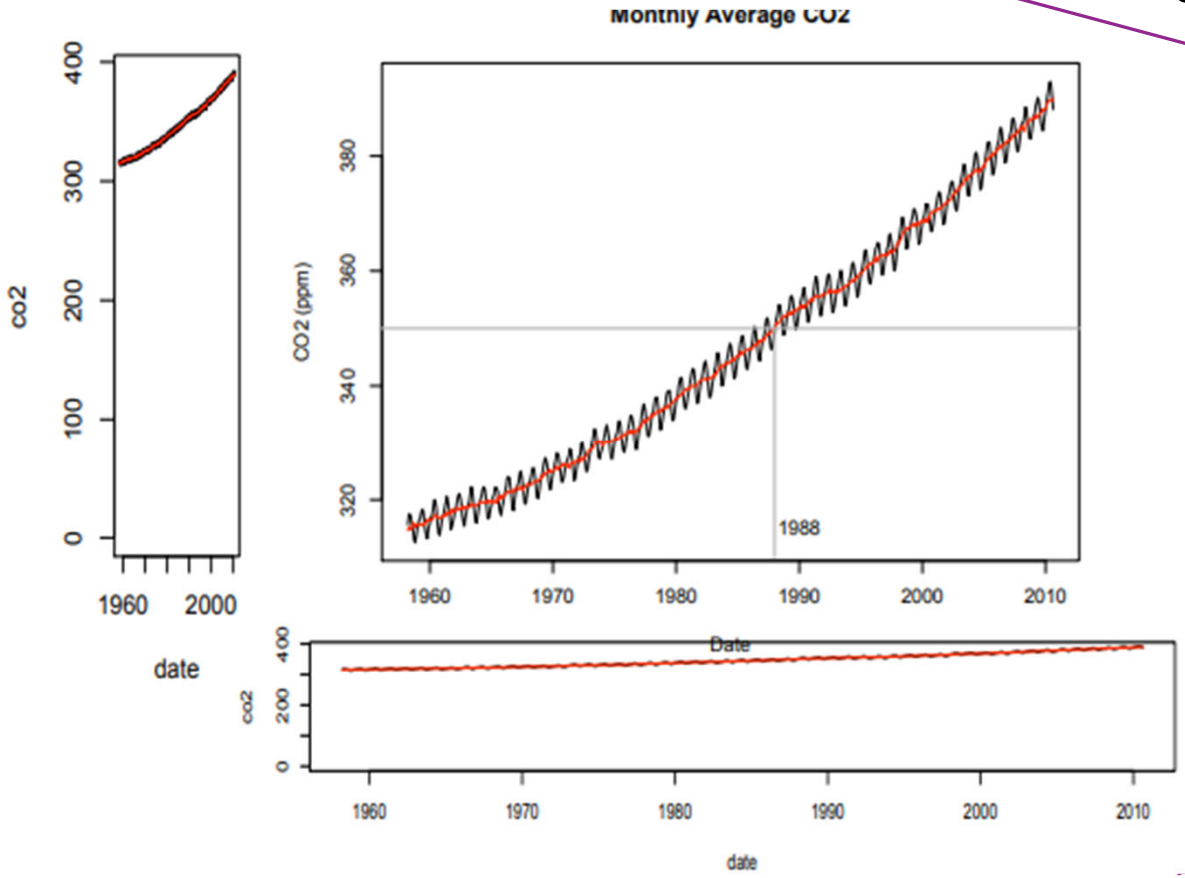
- Ratio of height to width approximately 1
  - Trend line inclining at about 45 degree
- Values of CO2 ranges from about 300ppm to 400ppm
  - Value 0 excluded from chart



Scale

What are the differences in values?

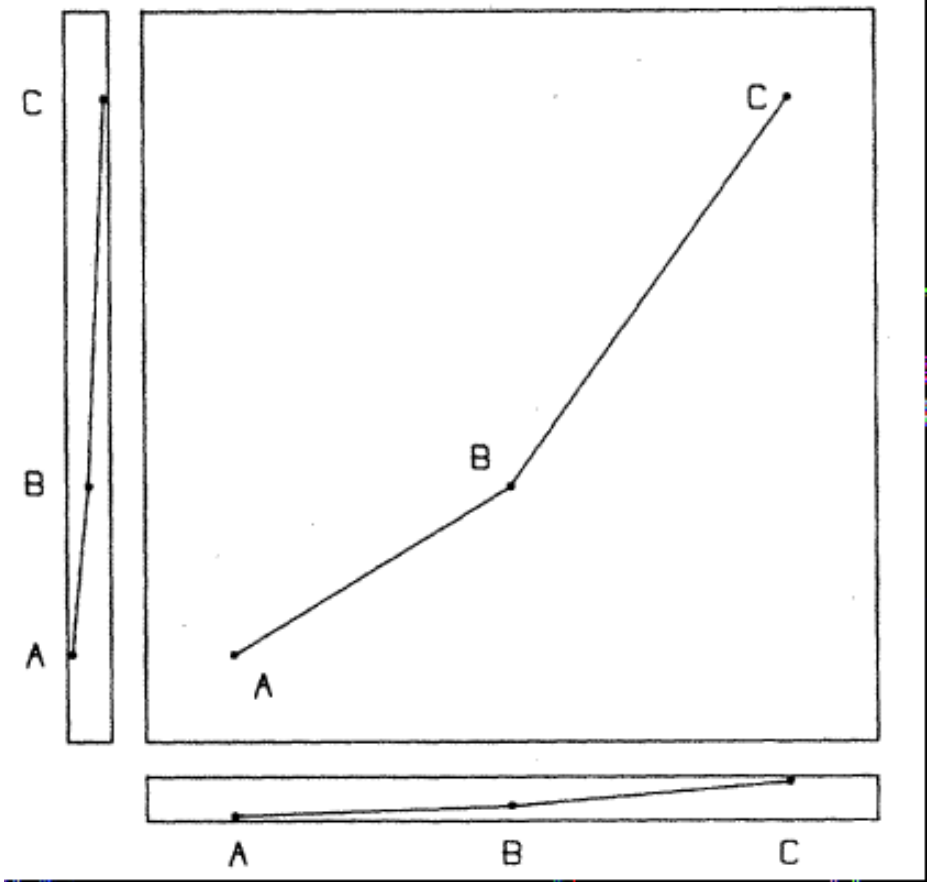
- Would including 0 affect the scale of the chart?
  - Maintaining the incline of 45 degrees about mean adding empty space
  - Filling empty space would mean setting the incline to 0 degrees



Scale

What are the differences in values?

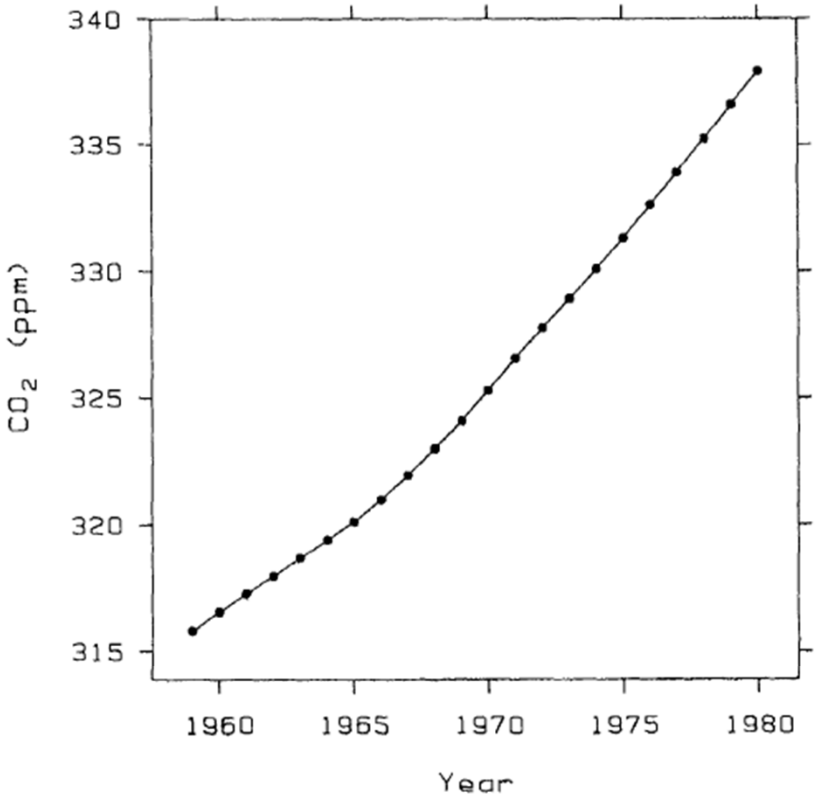
- Would including 0 affect the scale of the chart?
  - Maintaining the incline of 45 degrees about mean adding empty space
  - Filling empty space would mean setting the incline to 0 degrees



Scale

What are the differences in values?

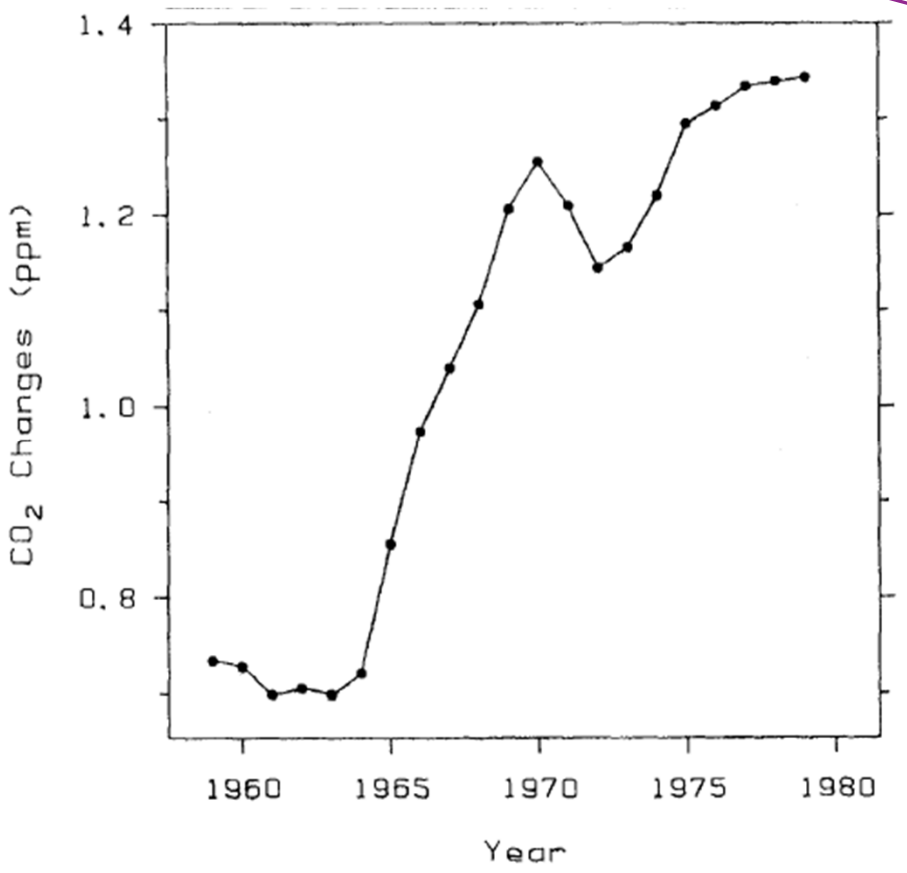
- Could we guess the rate of change each year from the chart?
  - Rate appears constant from 1957 to 1965 and 1967 to 1985
  - Latter rate seems slightly larger than former rate



Scale

What are the differences in values?

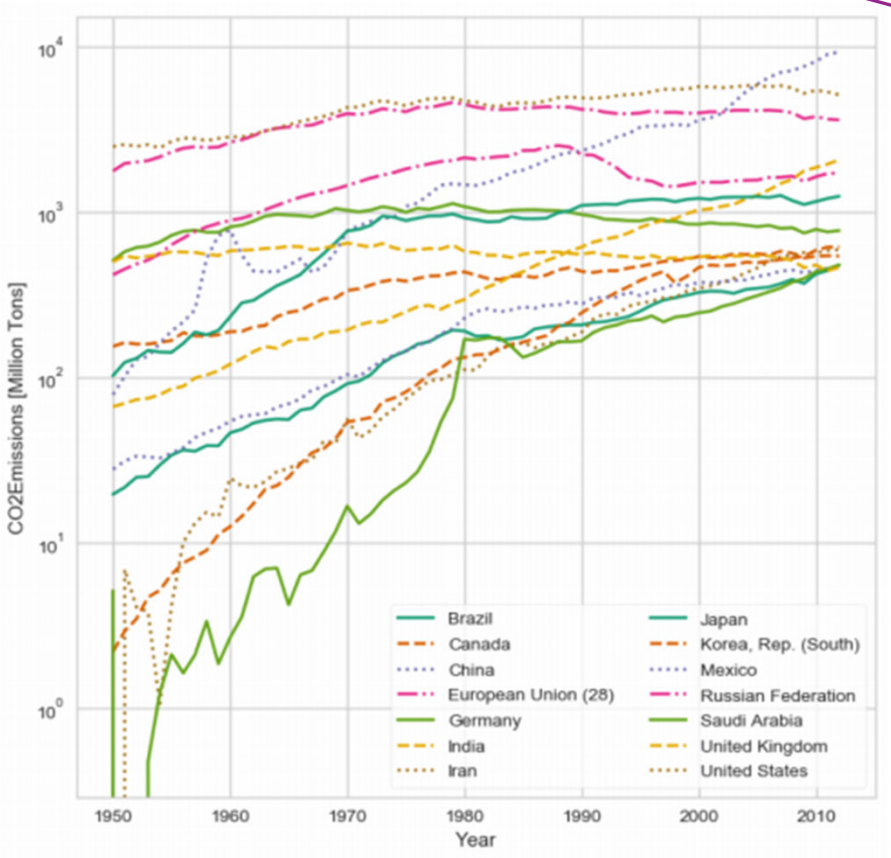
- Could we guess the rate of change each year from the chart?
  - Calculating the rate of change shows the increase over time
  - Rate of change has doubled between 1965 and 1985
- Position more understandable than slope



Multiple Charts

What is the scope of the data?

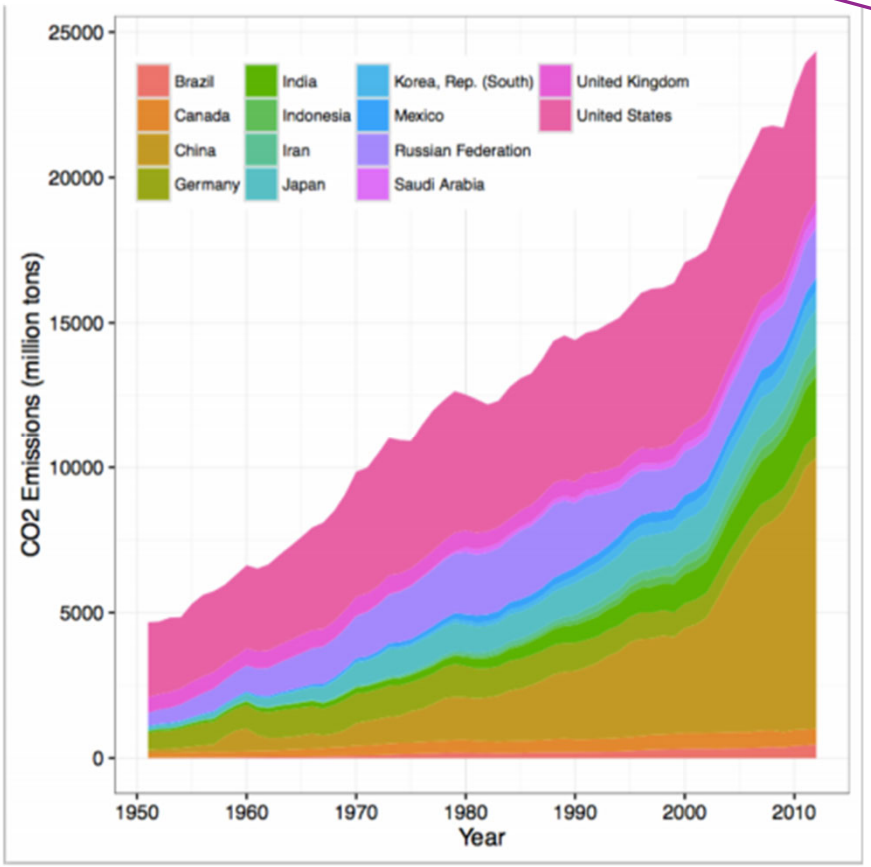
- Observations from Mauna Loa Observatory cannot measure emissions from different countries
- Stratified measurements across countries show differences in CO2 output



Multiple Charts

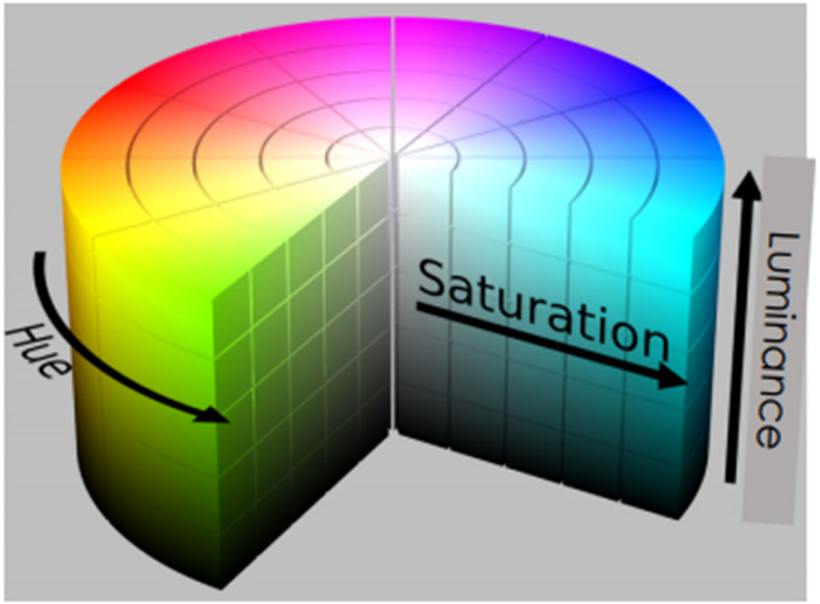
What is the scope of the data?

- Shading obscures changes between UK, Saudi Arabia and US
- Superposition of charts make India and Indonesia indistinguishable
- Cannot condition on China to track difference with Russia
- Values for Canada need to be transformed to match scale



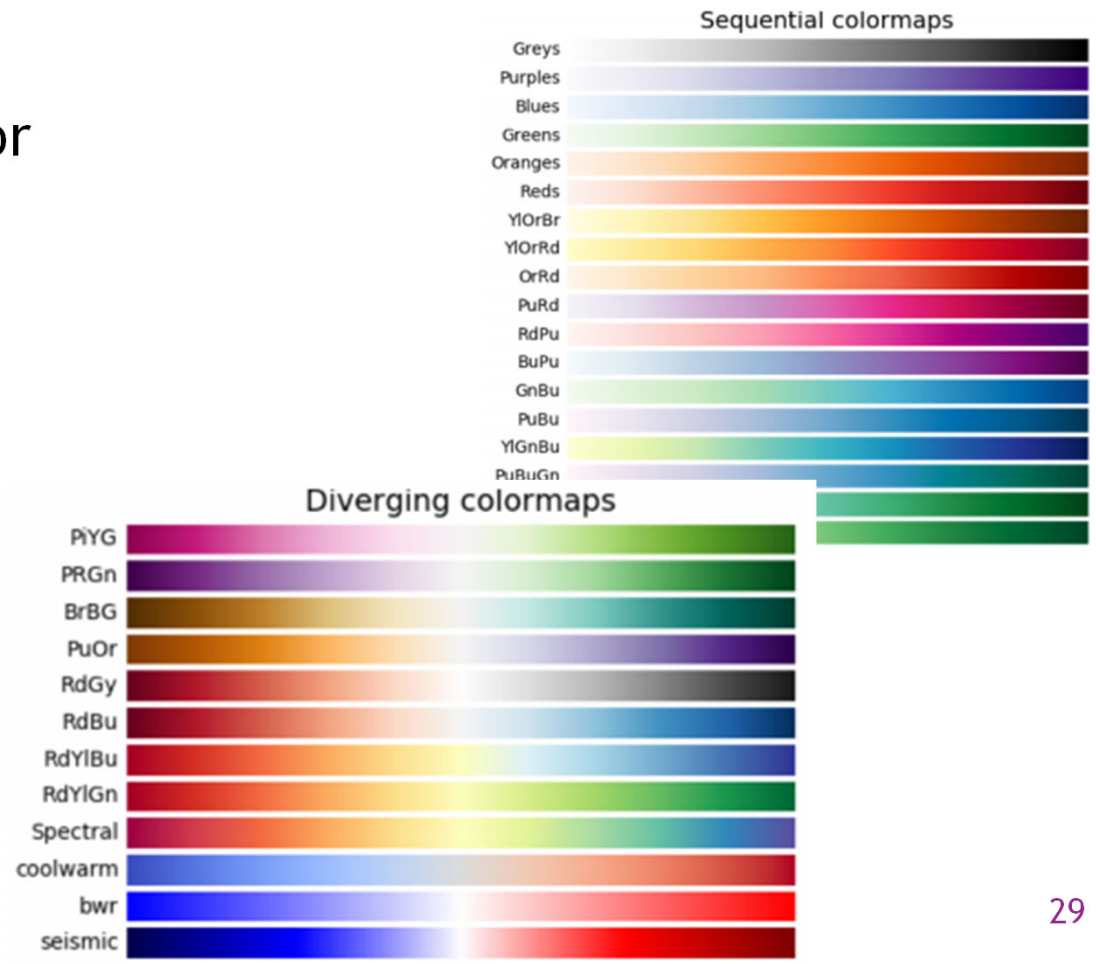
Color

- Different hues can be indistinguishable
- Color saturation can be distracting
- Lighter colors tend to make areas look larger than darker colors



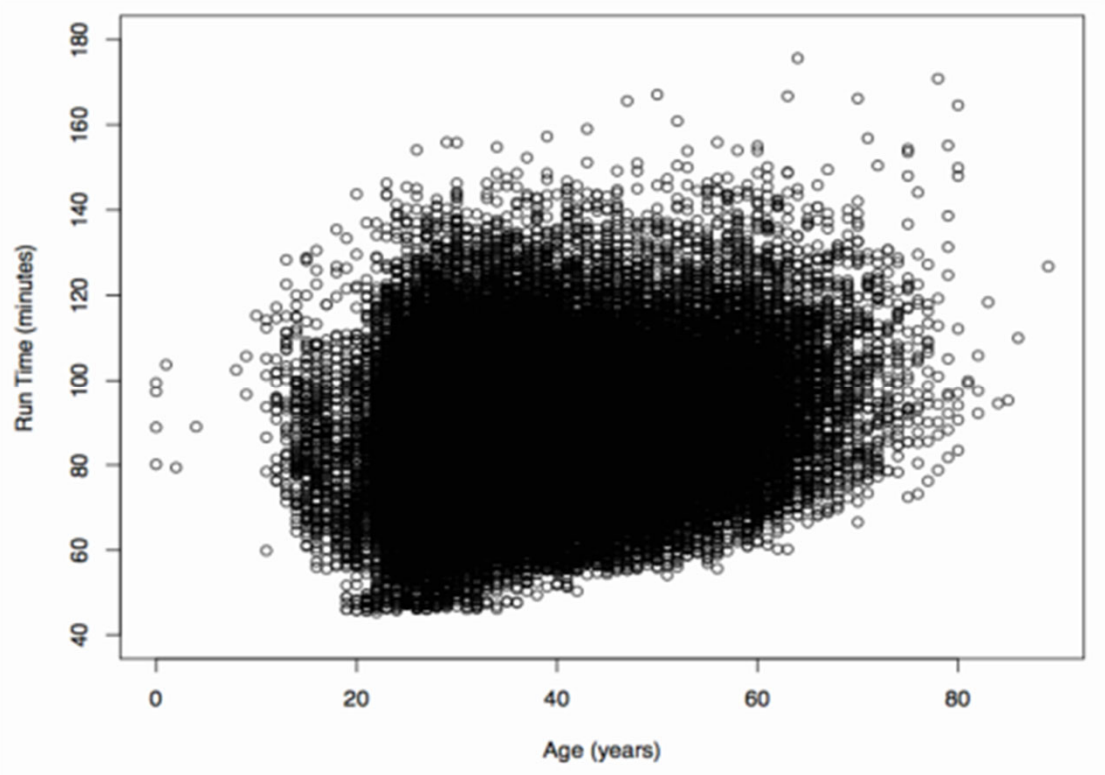
Color

- Qualitative
  - Choose distinguishable color scheme
- Quantitative
  - Choose color scheme that reflects the size of the numbers
  - Sequential for different emphasis on high and low
  - Diverging for different emphasis on medium and high / low



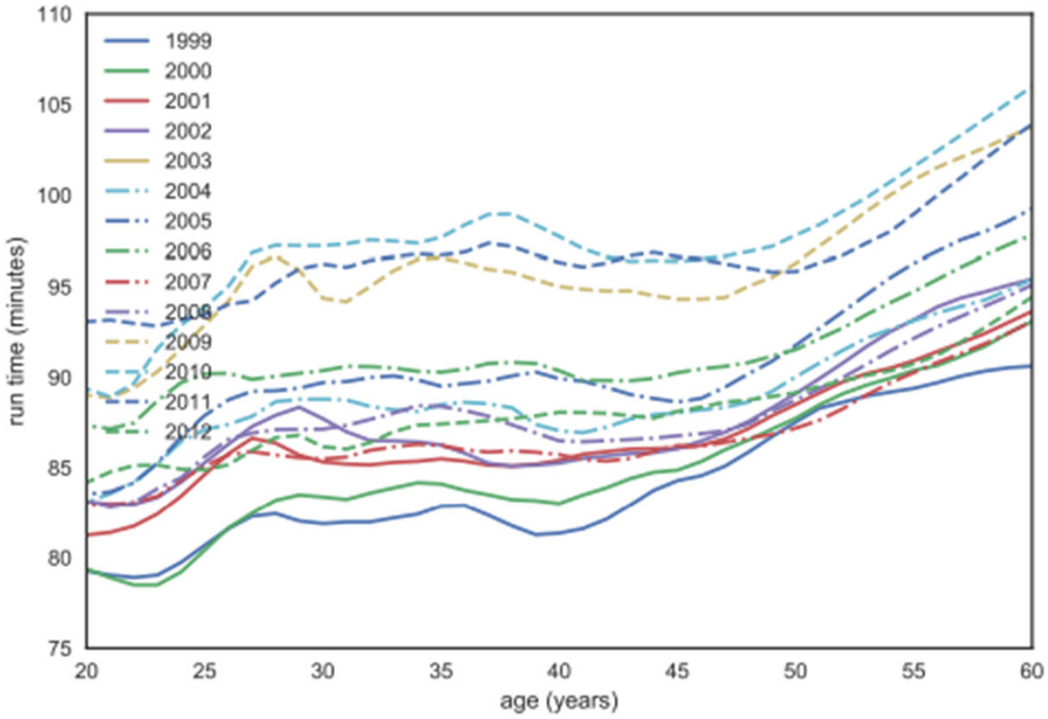
Superposition

- If one chart explains a trend in another chart, then you could try to them on top of each other (superimpose)
- Otherwise try to plot the charts side by side (juxtapose)



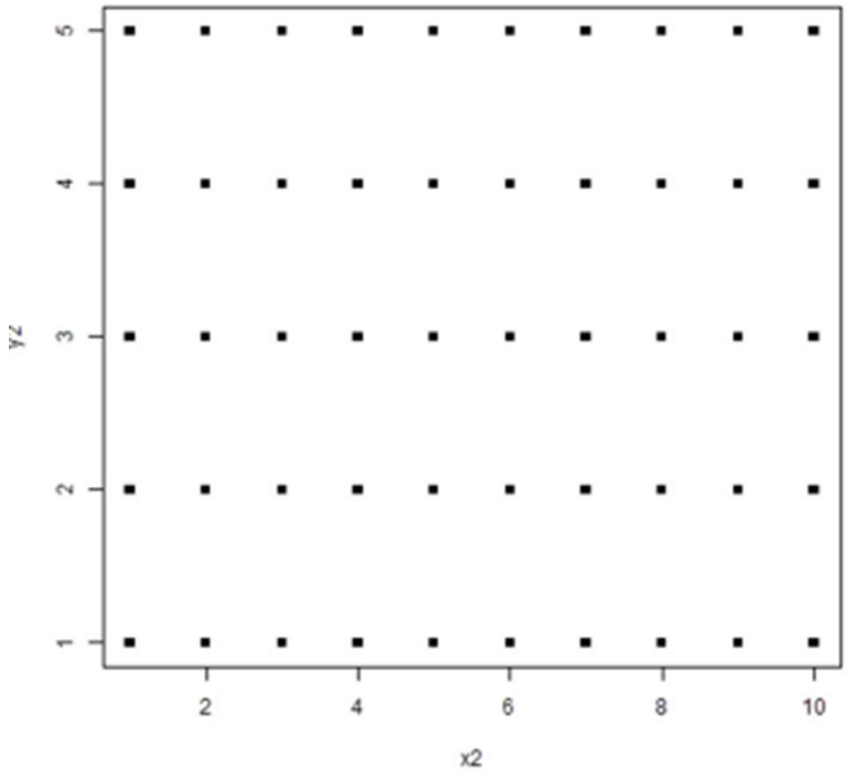
Superposition

- If one chart explains a trend in another chart, then you could try to them on top of each other (superimpose)
- Otherwise try to plot the charts side by side (juxtapose)



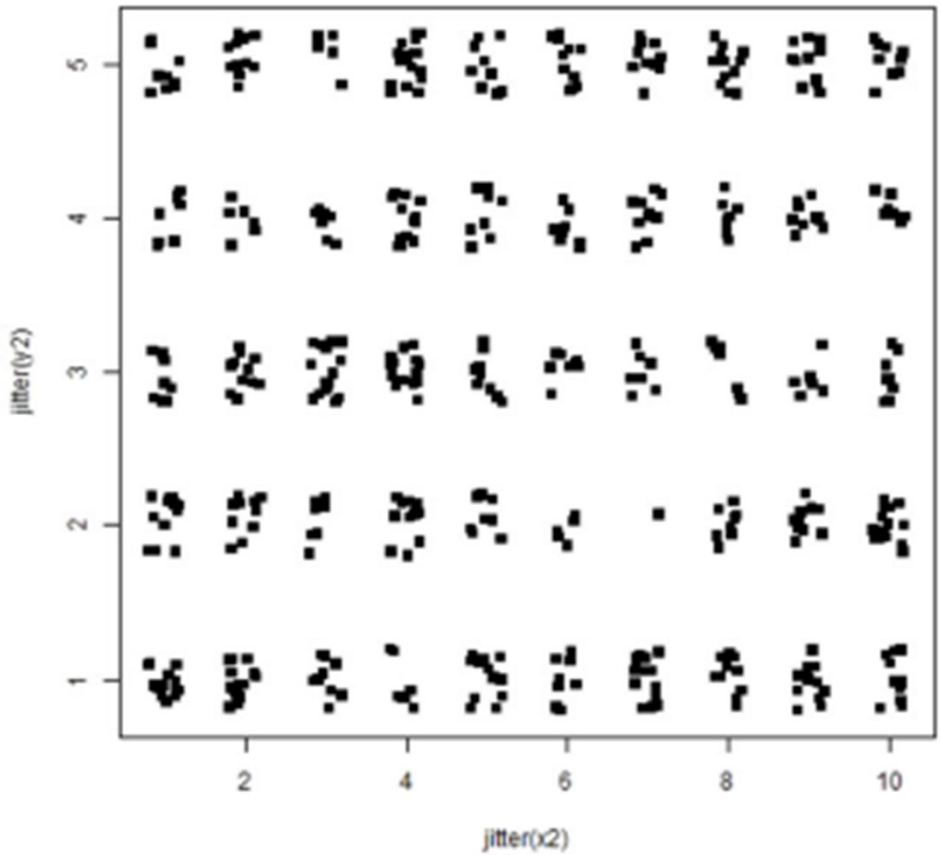
Over-plotting

- With too many charts in the same figure the information gets obscured through over-plotting
- If you can split the charts across different figures then try to jitter the data



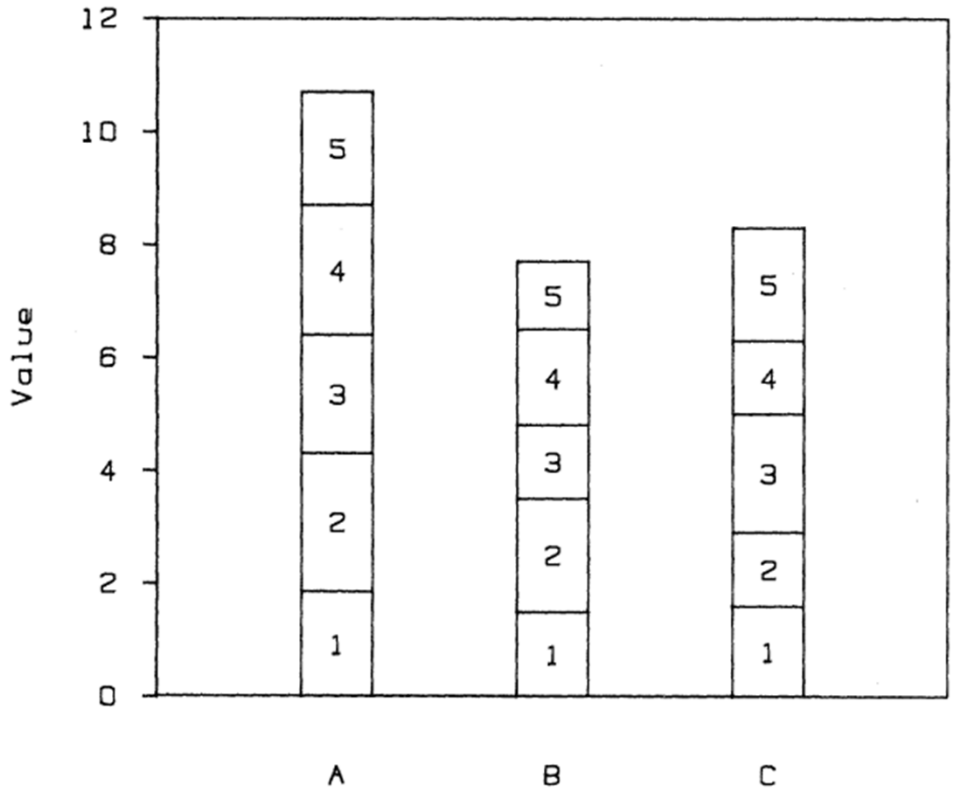
Over-plotting

- With too many charts in the same figure the information gets obscured through over-plotting
- If you can split the charts across different figures then try to jitter the data



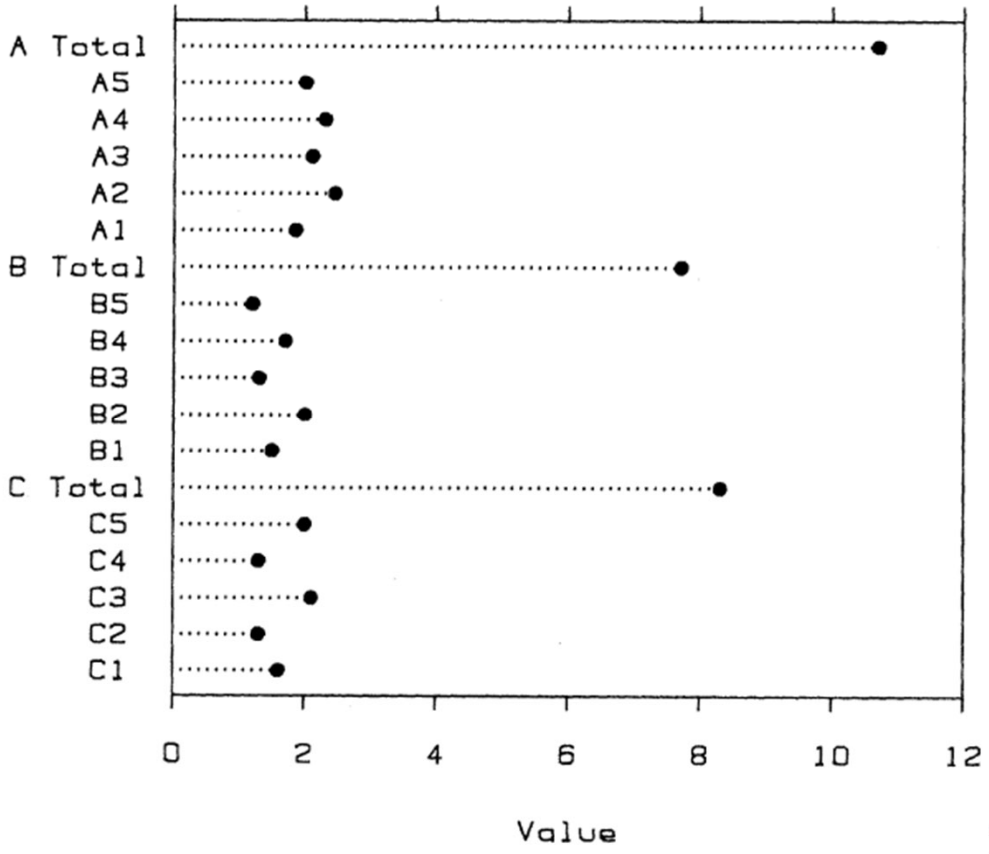
Conditioning

- Differences between categories can get obscured without a consistent reference
- Stacking regions of the chart on top of each other leads to a jiggling baseline that prevents comparison



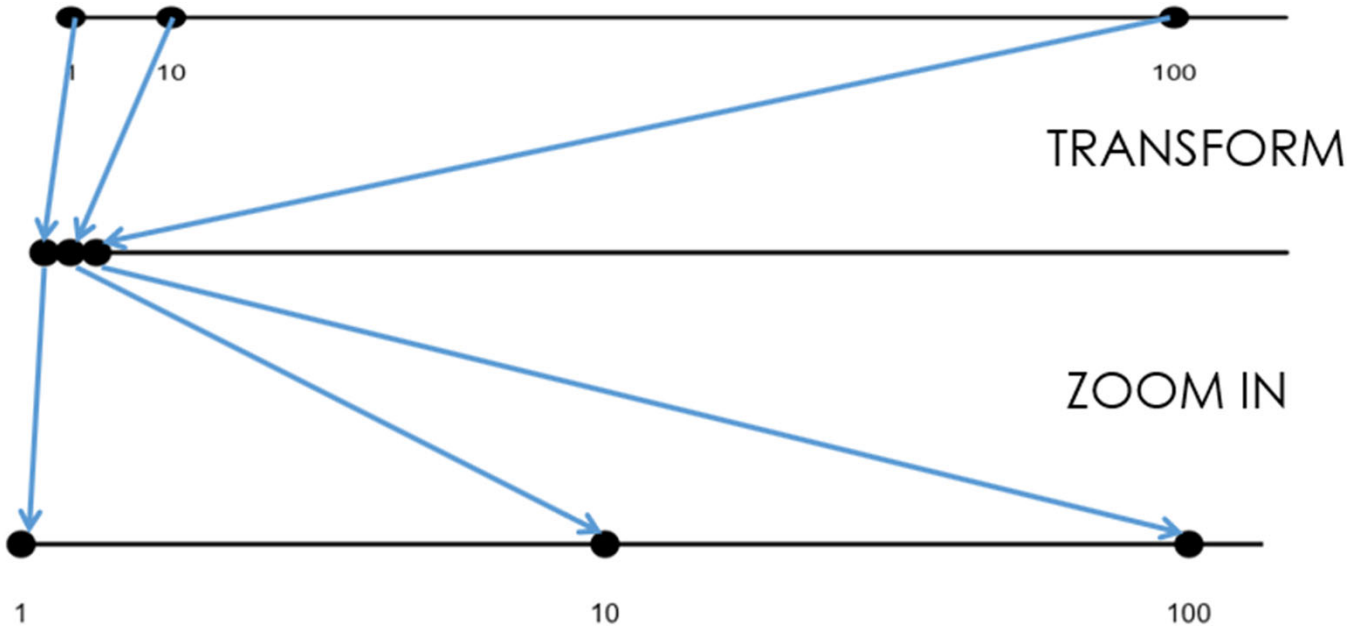
Conditioning

- Differences between categories can get obscured without a consistent reference
- Stacking regions of the chart on top of each other leads to a jiggling baseline that prevents comparison



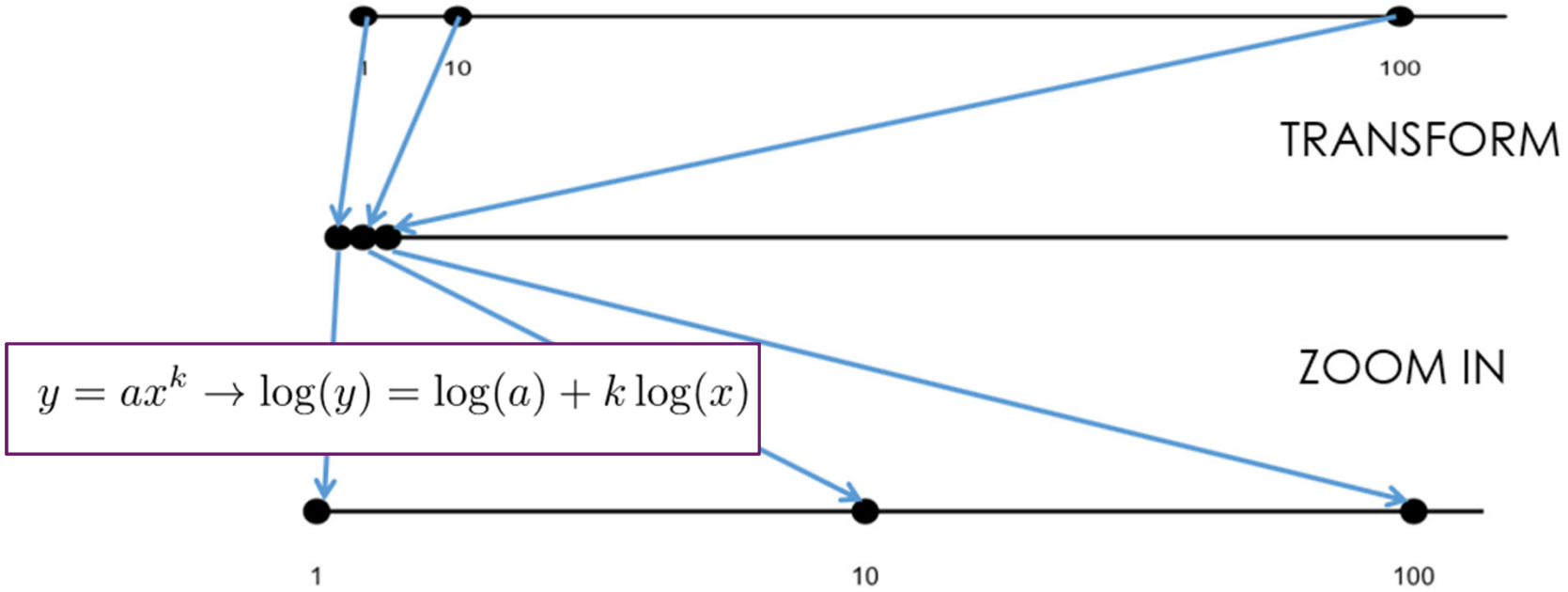
Transformation

- Many transformations to the data can improve visibility of the numbers
- Logarithms bring large numbers together but keep small numbers the same



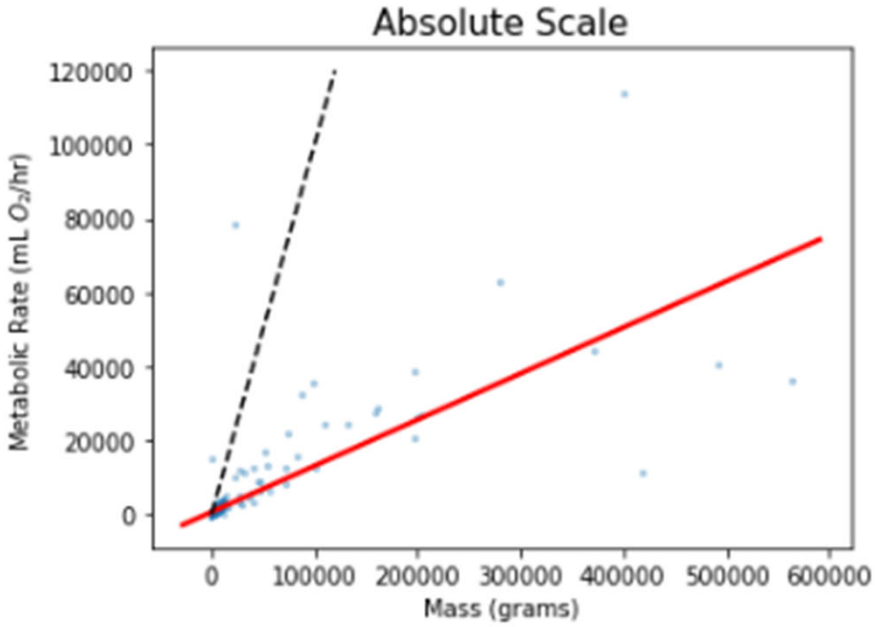
Transformation

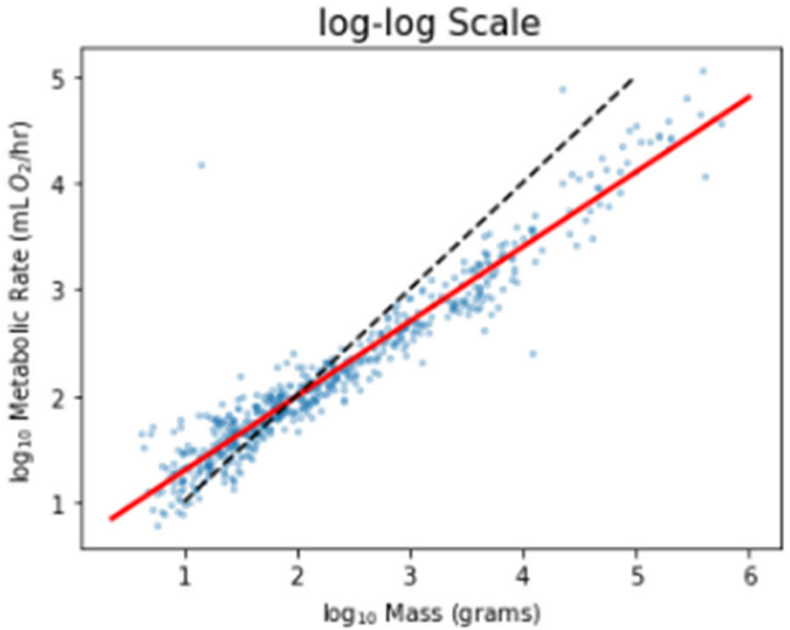
- Many transformations to the data can improve visibility of the numbers
- Logarithms bring large numbers together but keep small numbers the same



Transformation

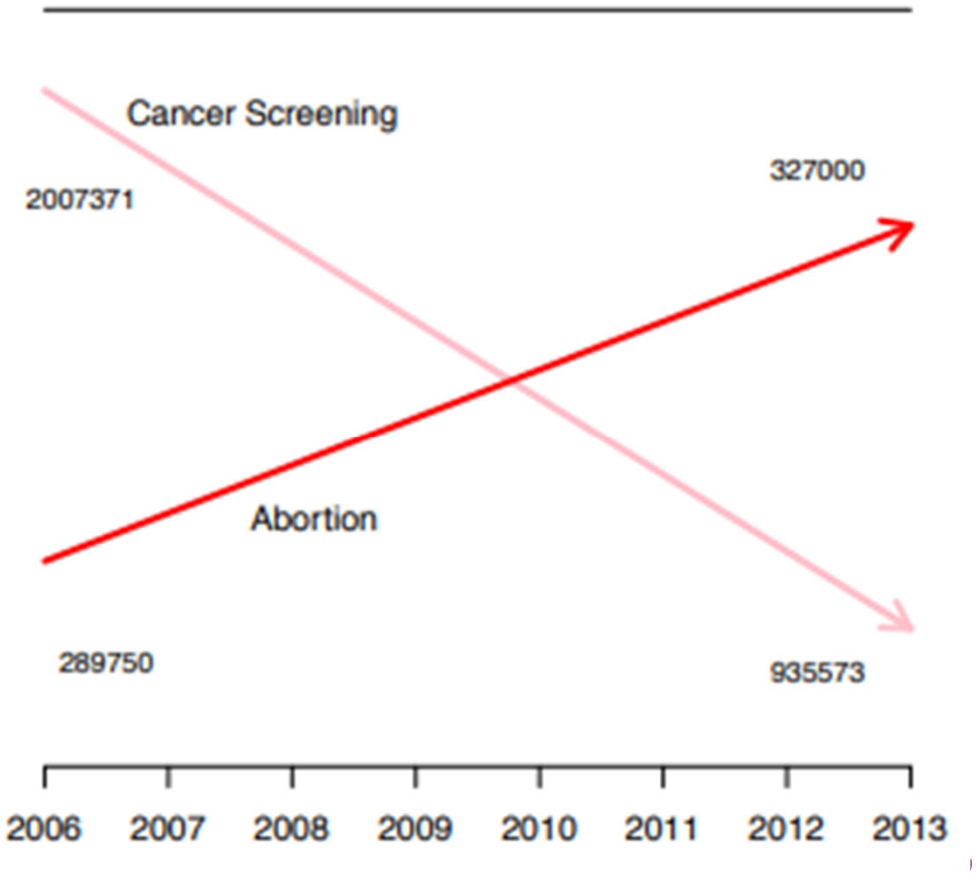
- Many transformations to the data can improve visibility of the numbers
- Logarithms bring large numbers together but keep small numbers the same





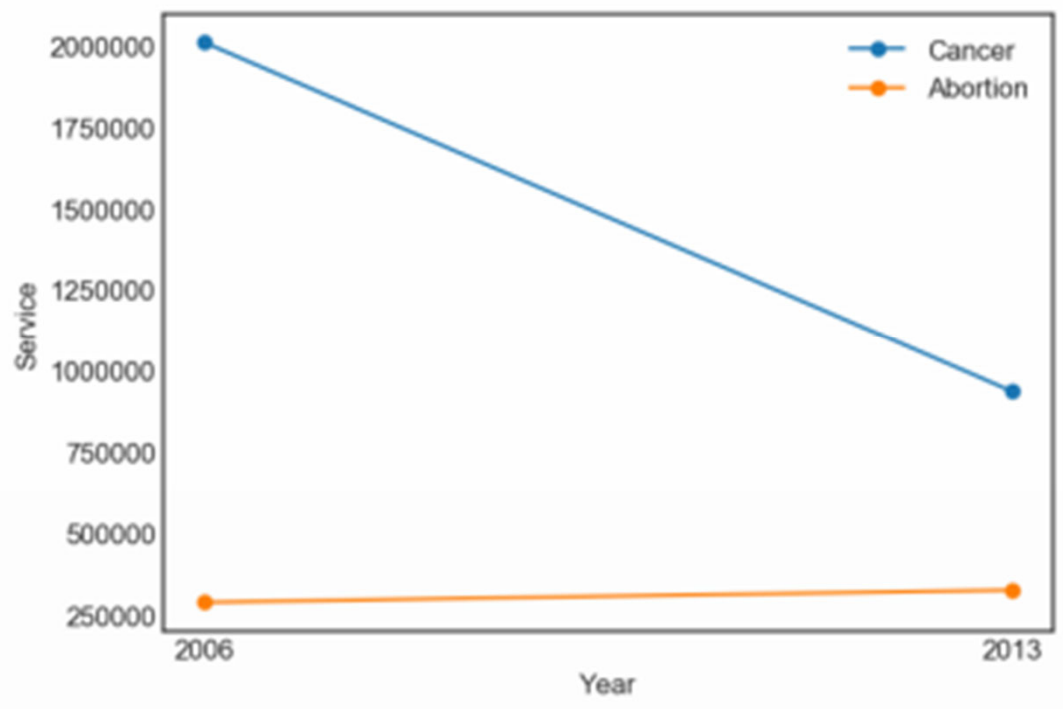
Exercise

- When plotting data, which of the following approaches to axis scaling should be avoided?
  1. Using different scales for variables on the same axis
  2. Changing the scale in middle of axis
  3. Standardizing the scales for the same axis
  4. Maintaining a consistent scale through the axis



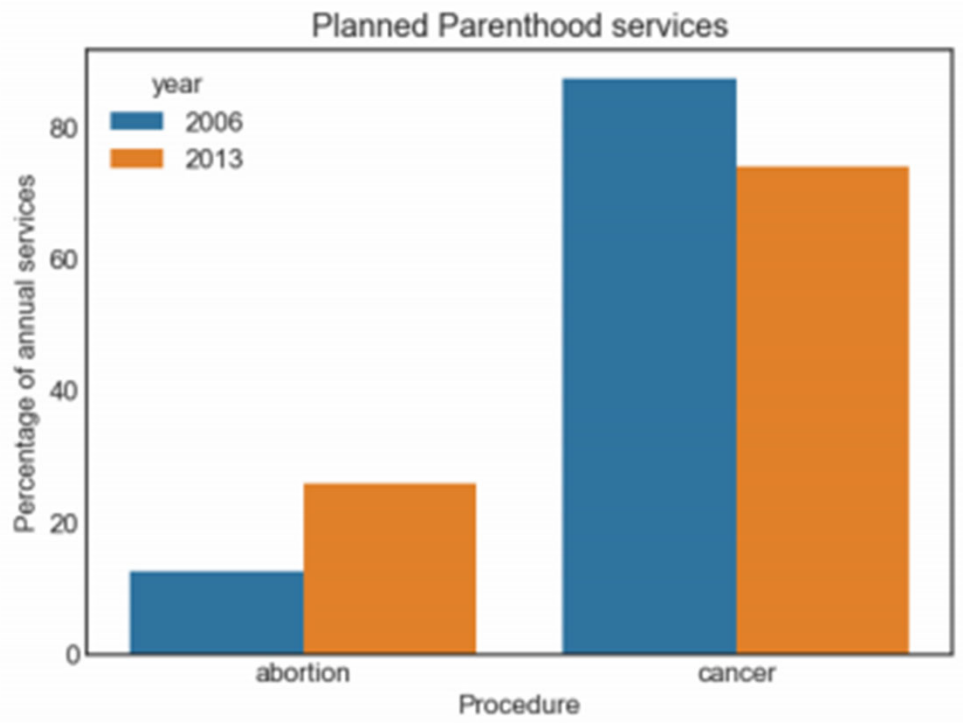
Exercise

- When plotting data, which of the following approaches to axis scaling should be avoided?
  1. Using different scales for variables on the same axis
  2. Changing the scale in middle of axis
  3. Standardizing the scales for the same axis
  4. Maintaining a consistent scale through the axis



Exercise

- When plotting data, which of the following approaches to axis scaling should be avoided?
  1. Using different scales for variables on the same axis
  2. Changing the scale in middle of axis
  3. Standardizing the scales for the same axis
  4. Maintaining a consistent scale through the axis



Summary

- Types of Charts
  - Single Variable
  - Multiple Variables
- Approaches to Visualizations
  - Scale
  - Conditioning
  - Colors
  - Transformations

Goals

- Choosing Appropriate Charts based on Data Type
- Helpful Practices for Understandable Images
- Transforming Datasets

Questions

- Questions on Piazza?
  - Please provide your feedback along with questions
- Question for You!

How can we match patterns of characters in strings?

Regular Expression Tutorial
https://docs.python.org/2/howto/regex.html

Questions

- Que
  - [
- (

How can we mate
patterns of characters in
strings?

https://doc x.html