Introduction to data science

Patrick Shafto

Department of Math and Computer Science

Plan for today

- Updates for next week!
    - Class canceled Monday
    - Guest lecture on Thurs by a data scientist (former actual scientist) at a local startup (that is likely hiring!)
- HW: Bechel test
- More viz!
- HW for Sunday

HW10

- Propose a question for your project
- Explain why it is an interesting question
- What data will you use to answer it?
- How will you operationalize the question?
- How will you confirm your results?
- What are assumptions / limitations?
- What is the statement that you will be able to make after the analysis?
- Who is in your group?

- Homework

- Read the Bethel test and think about it!

Example of complete analysis

- http://nbviewer.jupyter.org/github/brianckeegan/Bechdel/blob/master/Bechdel_test.ipynb

Think about the rubric!
Also think about cleaning / exploratory data analysis

	Capstone 4
Interpretation *Ability to explain information presented in mathematical forms (e.g., equations, graphs, diagrams, tables, words)*	Provides accurate explanations of information presented in mathematical forms. Makes appropriate inferences based on that information. *For example, accurately explains the trend data shown in a graph and makes reasonable predictions regarding what the data suggest about future events.*
Representation *Ability to convert relevant information into various mathematical forms (e.g., equations, graphs, diagrams, tables, words)*	Skillfully converts relevant information into an insightful mathematical portrayal in a way that contributes to a further or deeper understanding.
Calculation	Calculations attempted are essentially all successful and sufficiently comprehensive to solve the problem. Calculations are also presented elegantly (clearly, concisely, etc.)
Application / Analysis *Ability to make judgments and draw appropriate conclusions based on the quantitative analysis of data, while recognizing the limits of this analysis*	Uses the quantitative analysis of data as the basis for deep and thoughtful judgments, drawing insightful, carefully qualified conclusions from this work.
Assumptions *Ability to make and evaluate important assumptions in estimation, modeling, and data analysis*	Explicitly describes assumptions and provides compelling rationale for why each assumption is appropriate. Shows awareness that confidence in final conclusions is limited by the accuracy of the assumptions.
Communication *Expressing quantitative evidence in support of the argument or purpose of the work (in terms of what evidence is used and how it is formatted, presented, and contextualized)*	Uses quantitative information in connection with the argument or purpose of the work, presents it in an effective format, and explicates it with consistently high quality.

- The Bechdel test (/ˈbɛkdəl/ BEK-dəl) asks whether a work of fiction features at least two women who talk to each other about something other than a man. The requirement that the two women must be named is sometimes added.

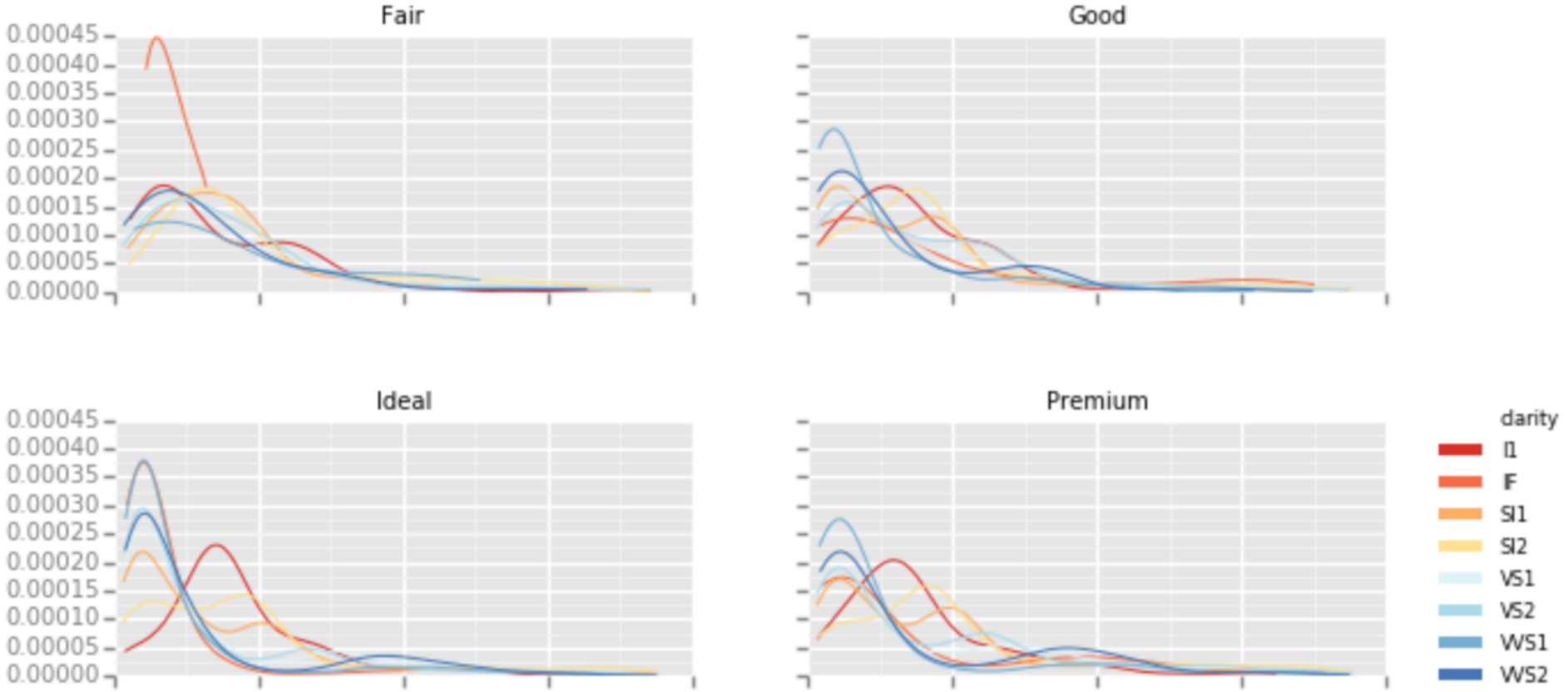
from wiki

ggplot

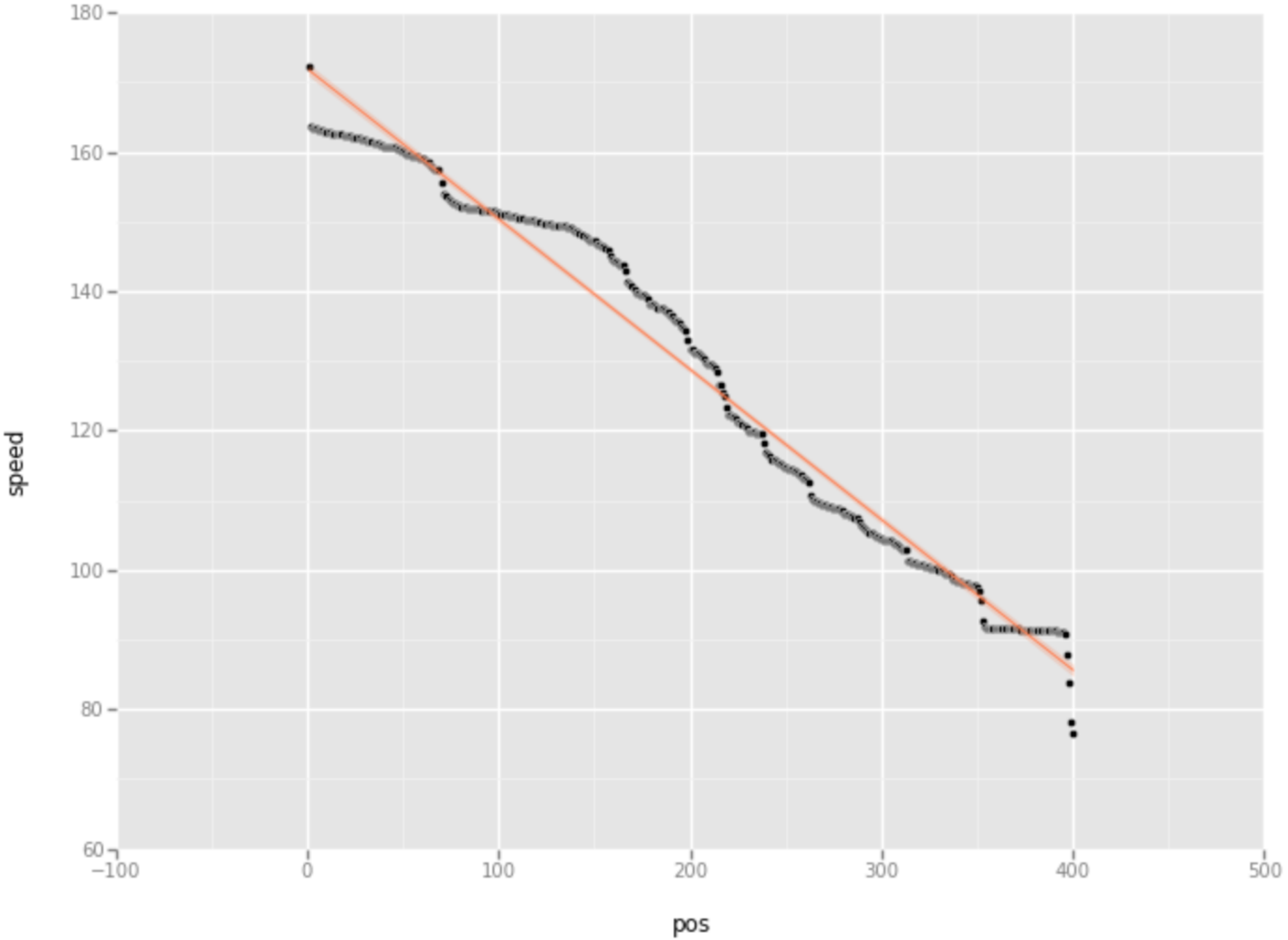
- yhat ggpy
- for more examples, look up ggplot2 for R

In [1]:
```
%matplotlib inline
from ggplot import *
import pandas as pd
import numpy as np
```

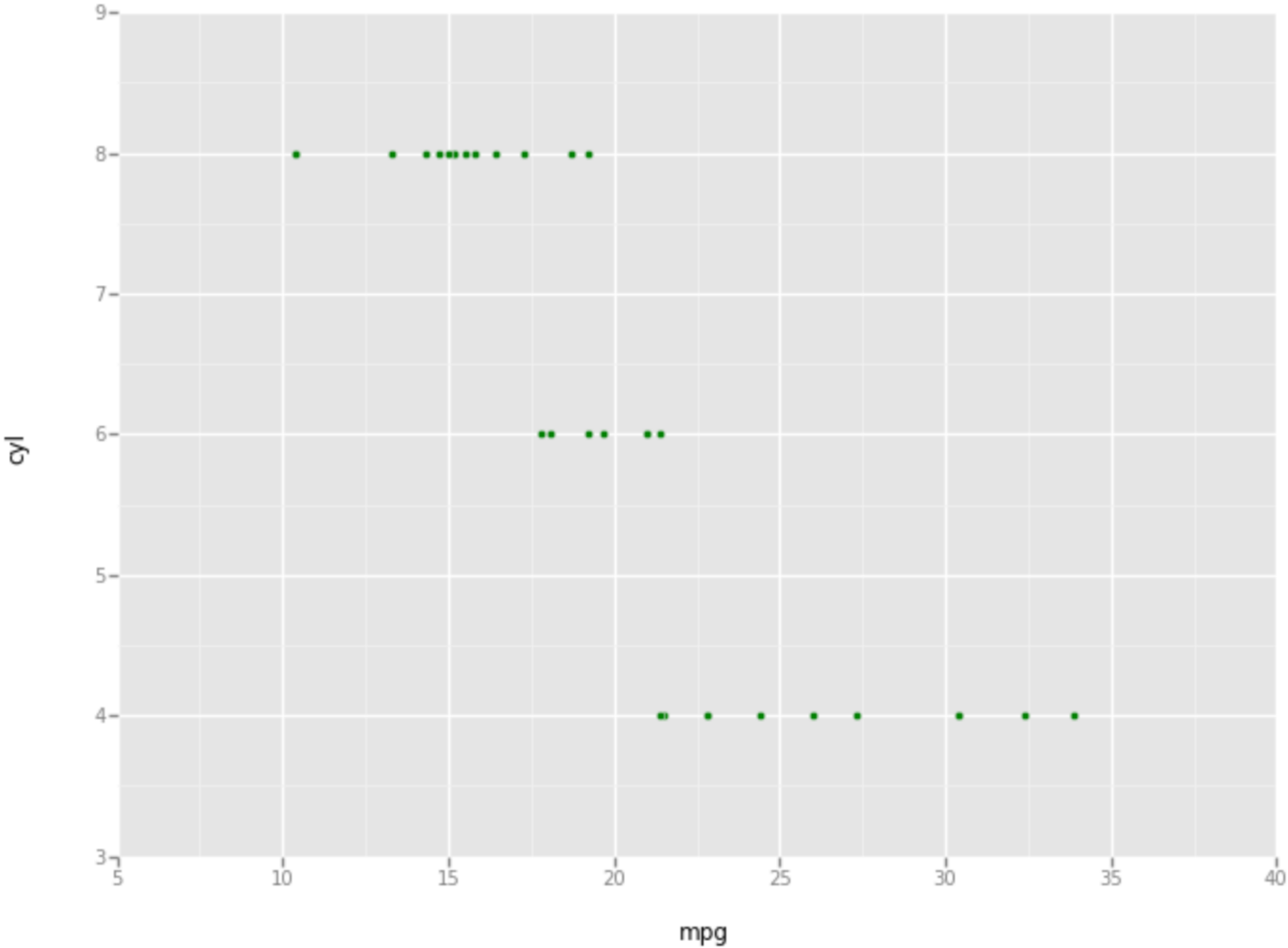
In [8]:
```
ggplot(diamonds, aes(x='price', color='clarity')) + \
    geom_density() + \
    scale_color_brewer(type='div', palette=7) + \
    facet_wrap('cut')
```



In [2]:
```
ggplot(pigeons, aes(x='pos', y='speed')) + \
    geom_point() + \
    stat_smooth(se=True, color='coral')
```



In [5]:
```
ggplot(mtcars, aes(x='mpg', y='cyl')) + geom_point(color='green')
```



Bokeh

- http://bokeh.pydata.org/en/latest/docs/user_guide/quickstart.html
  - Getting Started
  - Up to datetime axes
- More examples (check out the server apps!):
  - http://bokeh.pydata.org/en/latest/docs/gallery.html#gallery-server-examples

In class work

- Take the bokeh texas example and modify to be another state (try NJ!)
  - http://bokeh.pydata.org/en/latest/docs/gallery/texas.html
  - Comment each line of code!

HW10

- Propose a question for your project
- Explain why it is an interesting question
- What data will you use to answer it?
- How will you operationalize the question?
- How will you confirm your results?
- What are assumptions / limitations?
- What is the statement that you will be able to make after the analysis?
- Who is in your group?