Visualisation: Modelling the World

Richie Morrisroe

July 16, 2019

Structure

- This talk is an approach to visualisation
- Not many absolutes
- assumptions of vision
- Assumptions of Statistical Graphics
- Understanding data with Visualisation
- Communicating to others with Visualisation

What is Visualisation?

- a tool for understanding the world
- a way to communicate a particular perspective on data
- an adjunct to thought

Why Visualisation?

- The eye is really really good at finding patterns in pictures
- in fact, it's so good that it can find patterns that aren't even there

Figure: What do you see?

m

The importance of perspective

- You can see one of two things in the previous image
- Which of them can depend on what you expect to see
- It can also depend on what your environment contains

Muller-Lyer

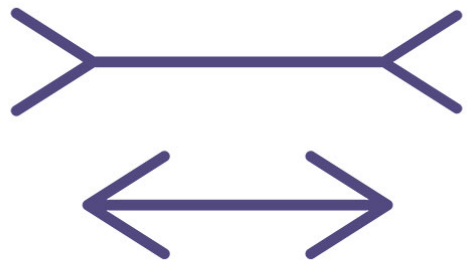


Figure: Which line is longer?

This illusion doesn't affect everyone similarly

- Europeans and Americans are more susceptible
- Africans are less susceptible
- Possibility that it is due to presence of right angles in urban environments
- appears to be a small difference between urban and rural dwellers

Who cares?

- Shows that how we interpret stimuli is not tabula rasa
- When you gaze into the image, the image also gazes into you...
- We bring our own perception and previous associations into any image [1]

[1] anything really, but we're talking about images here.

When to use Visualisation?

Always

Running Example

- Property Price Register
  - Kinda a crappy dataset
  - No cleaning or checking done by the authority
  - lots of craziness (1 apartment for 18.6mn)

Property Price Register

- We used Google's geocoding service to get more details on each observation
- I updated Shane Lynn's script and ran it on the data up till October 2018
- I also typically break out properties sold for greater than 1e6, as they are often multiple-unit sales (and there's little to no automated way of figuring this out) [2]
- Lots of manual fixing required
- the irish text definitely doesn't help

[2] please someone in the audience suggest a better idea

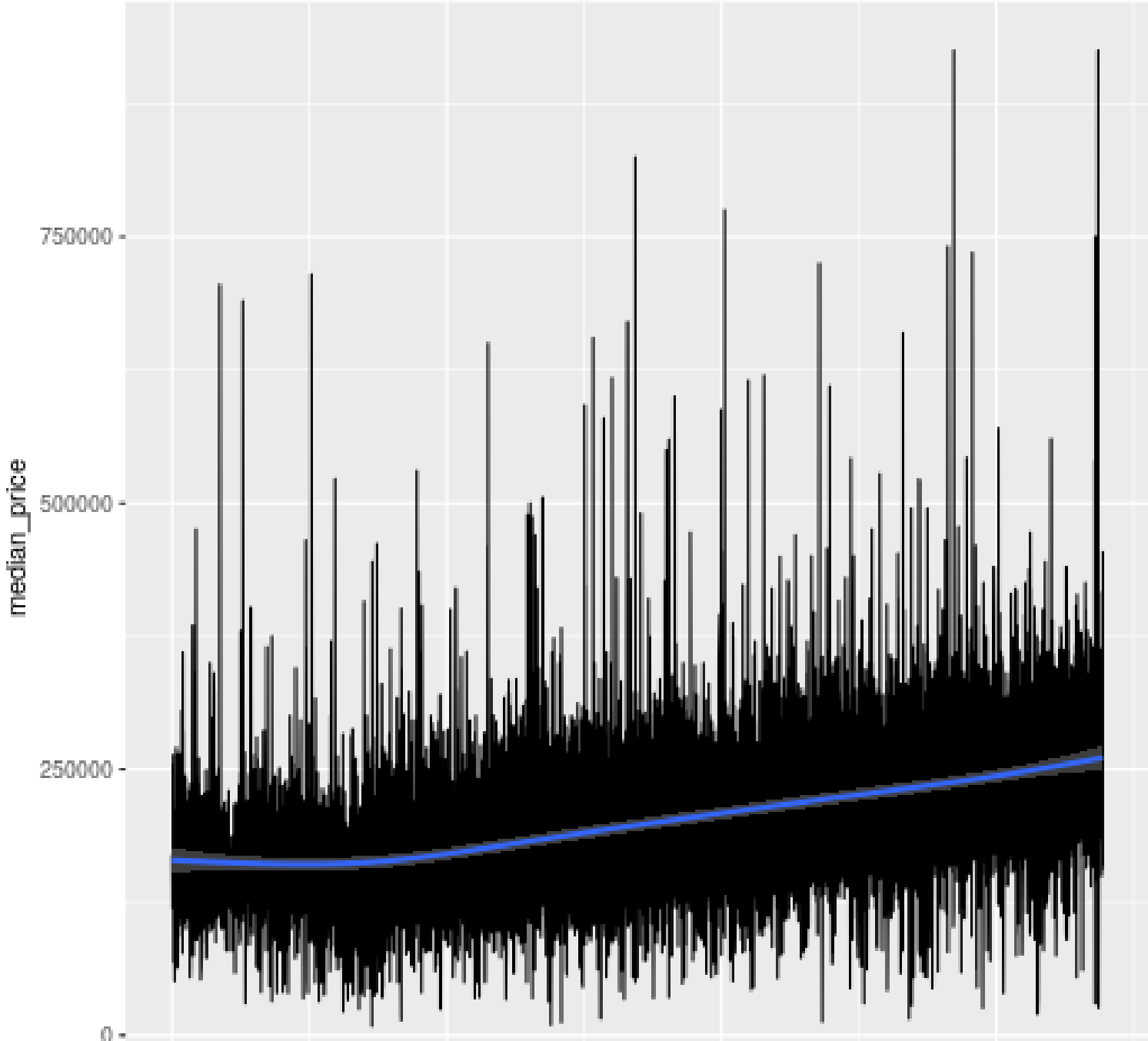
Assumptions of Statistical Graphics

- there are many
- in this section, I'd like to subvert them, in order to make you think

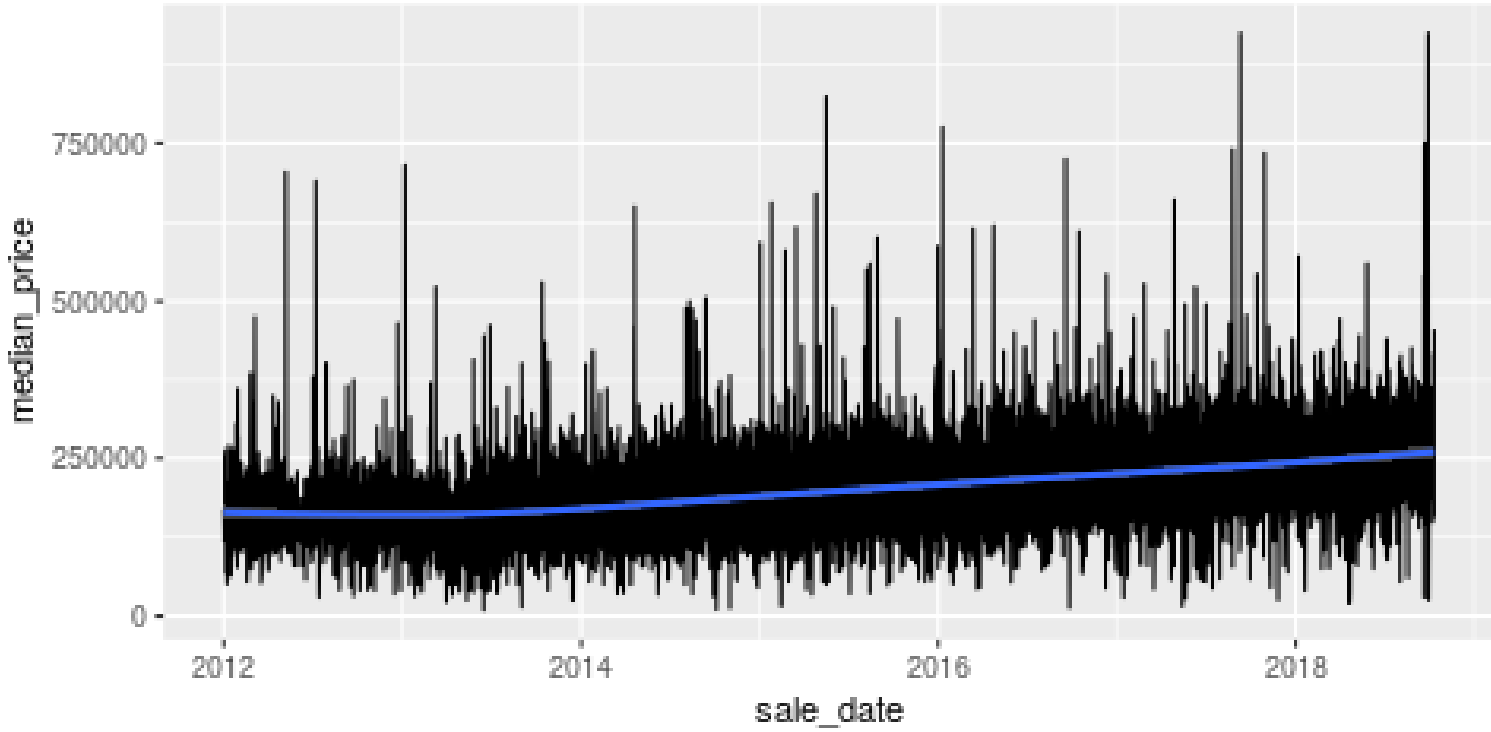
Line Graphs

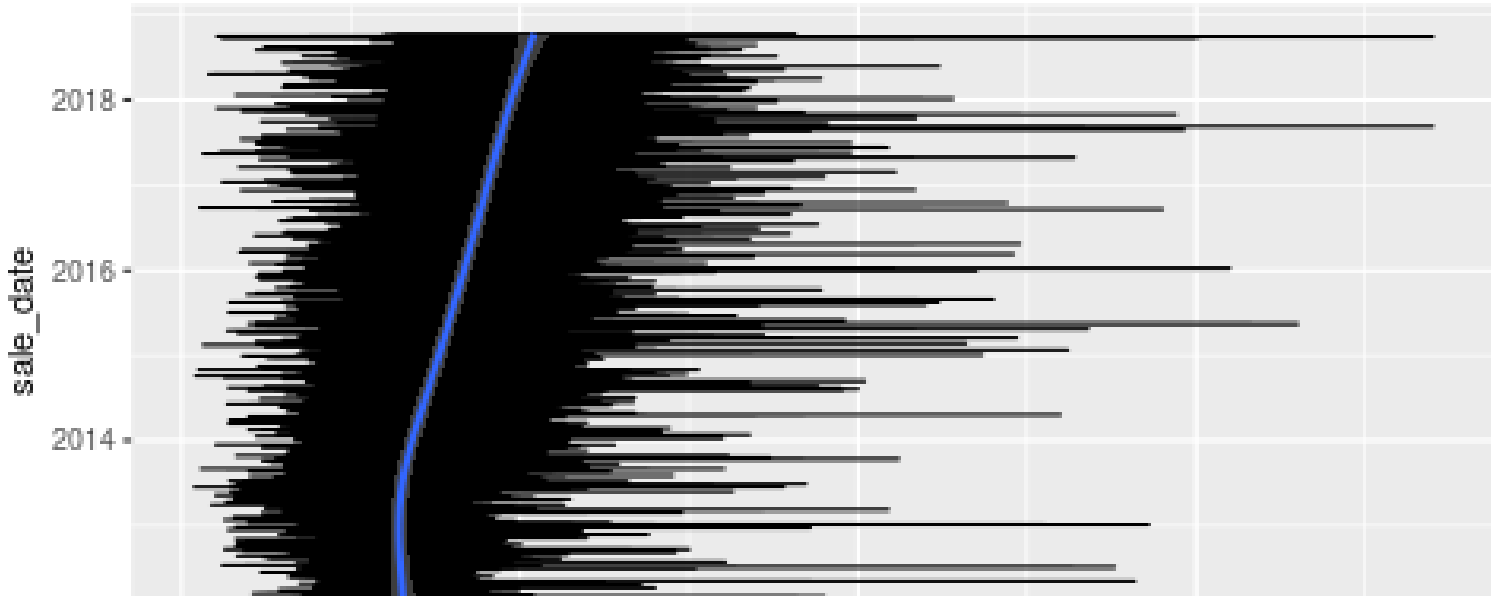
- Normally represent time
- scatterplots don't (always) have the same assumptions
- what is the deepest assumption?

Median Property Price by Day, Ireland 2011-18

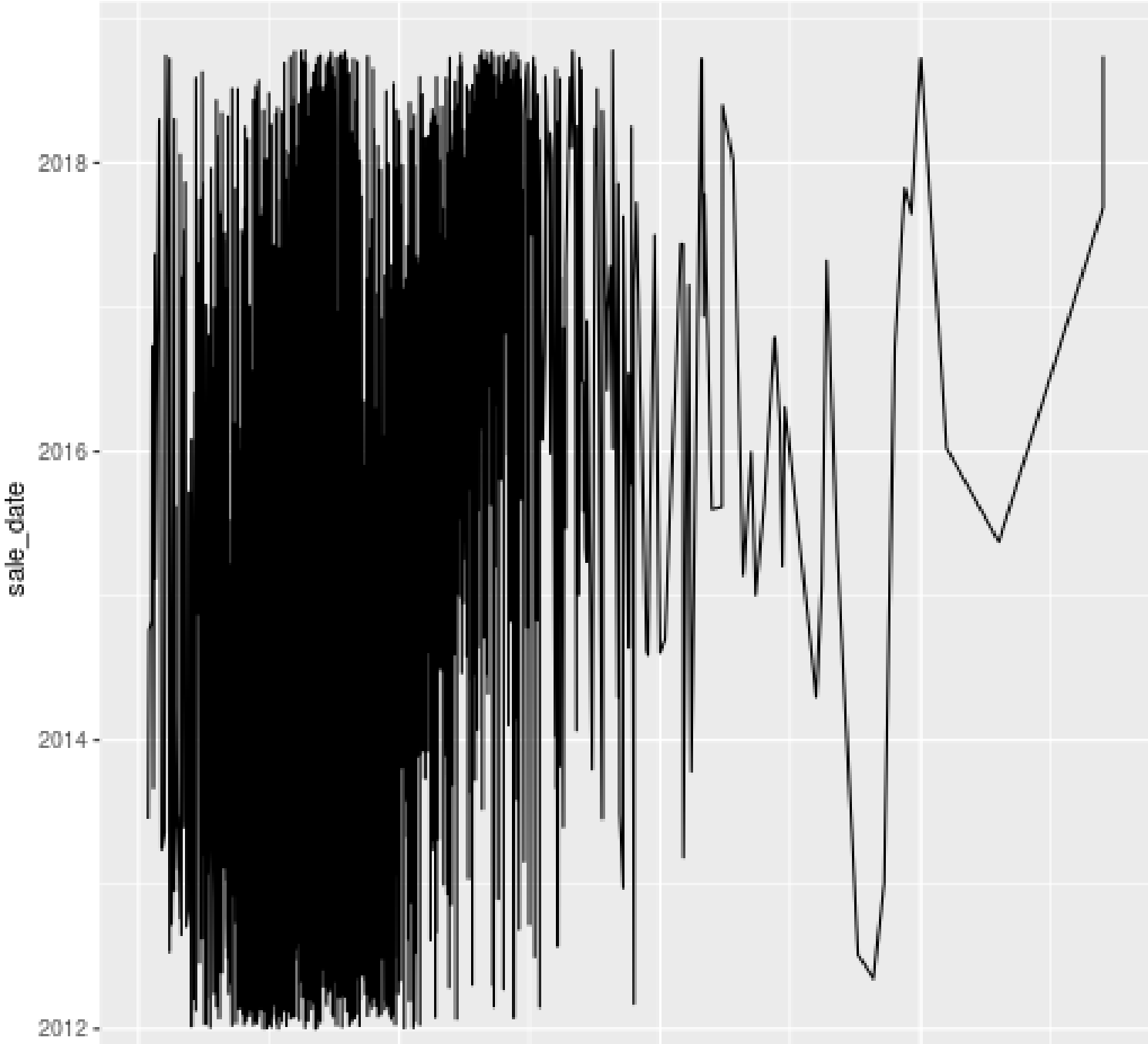


Flipped Line Chart

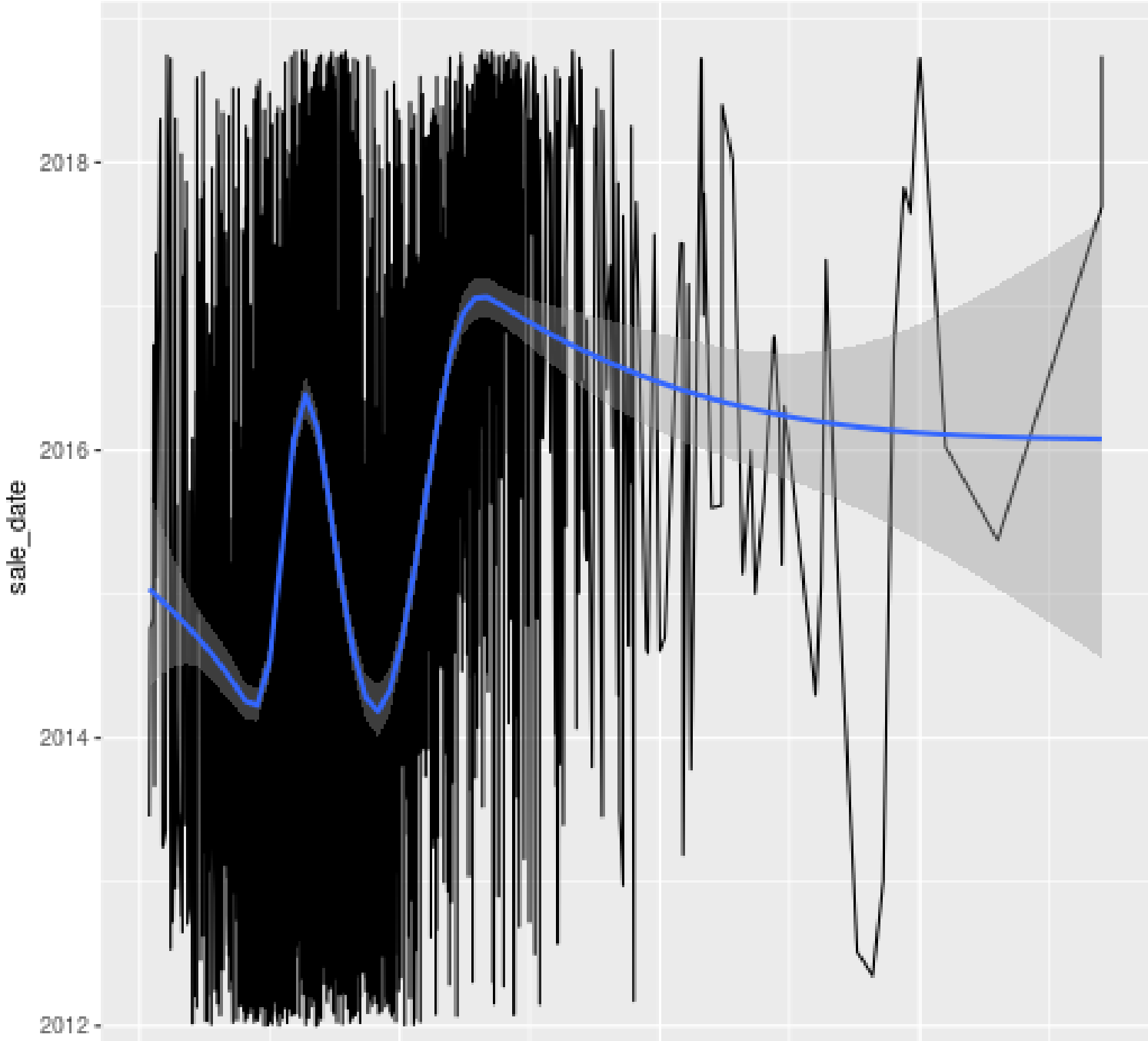




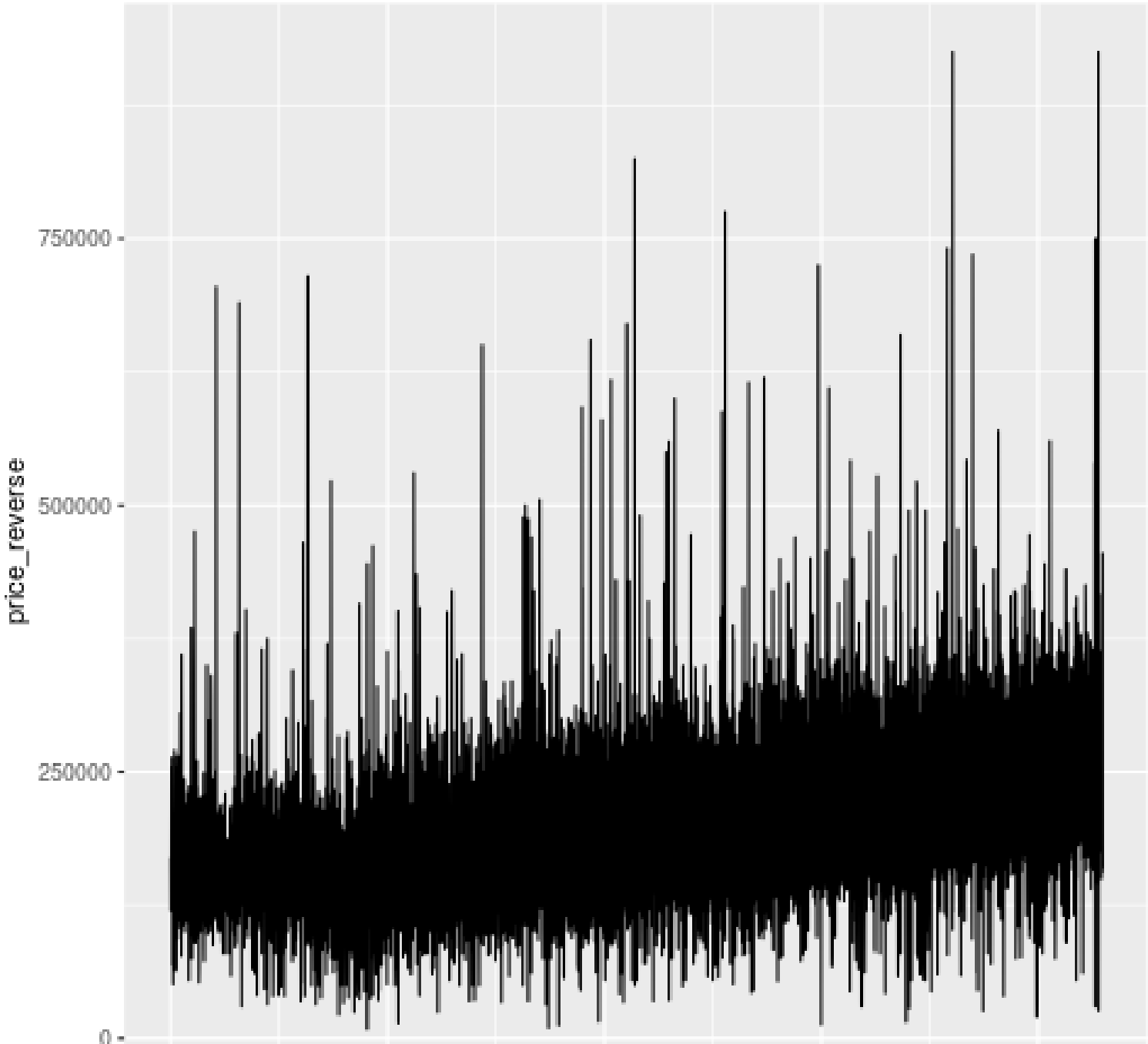
F-ing Line Chart



Abusing Standard Assumptions



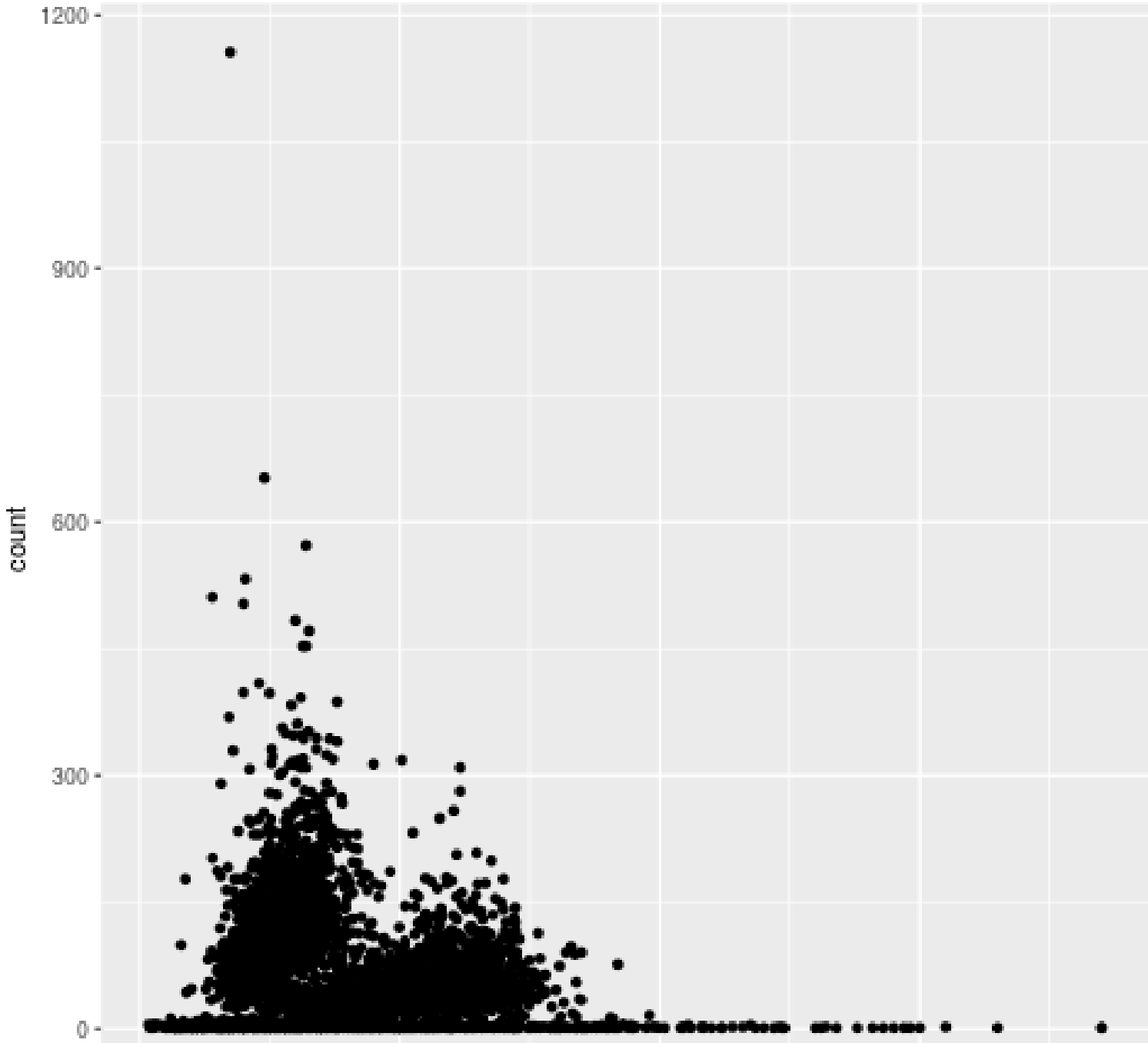
Backwards Line Chart



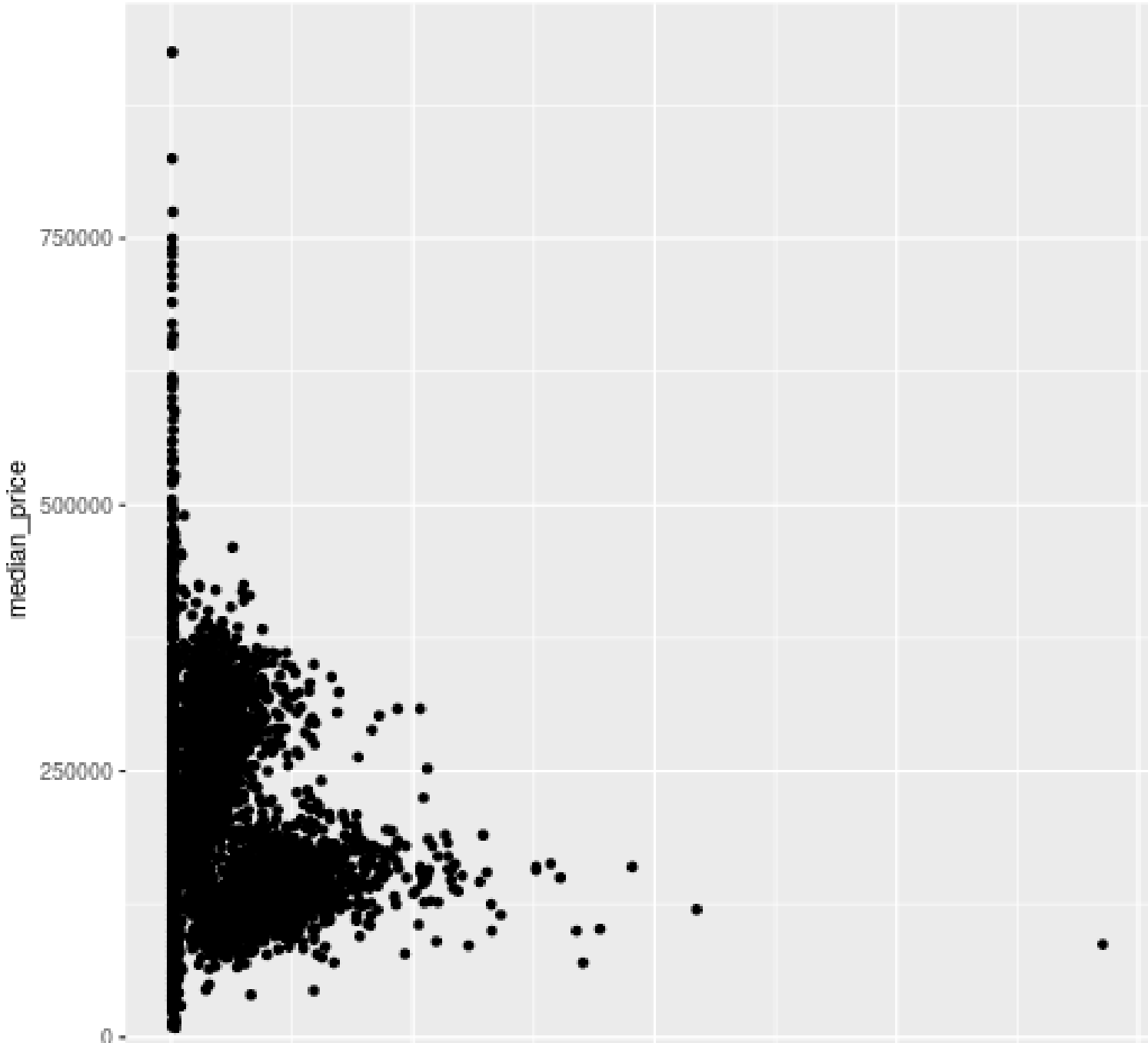
Scatter plot

- Also encodes a set of base assumptions
- points nearer to each other in space are more related
- more orientation issues

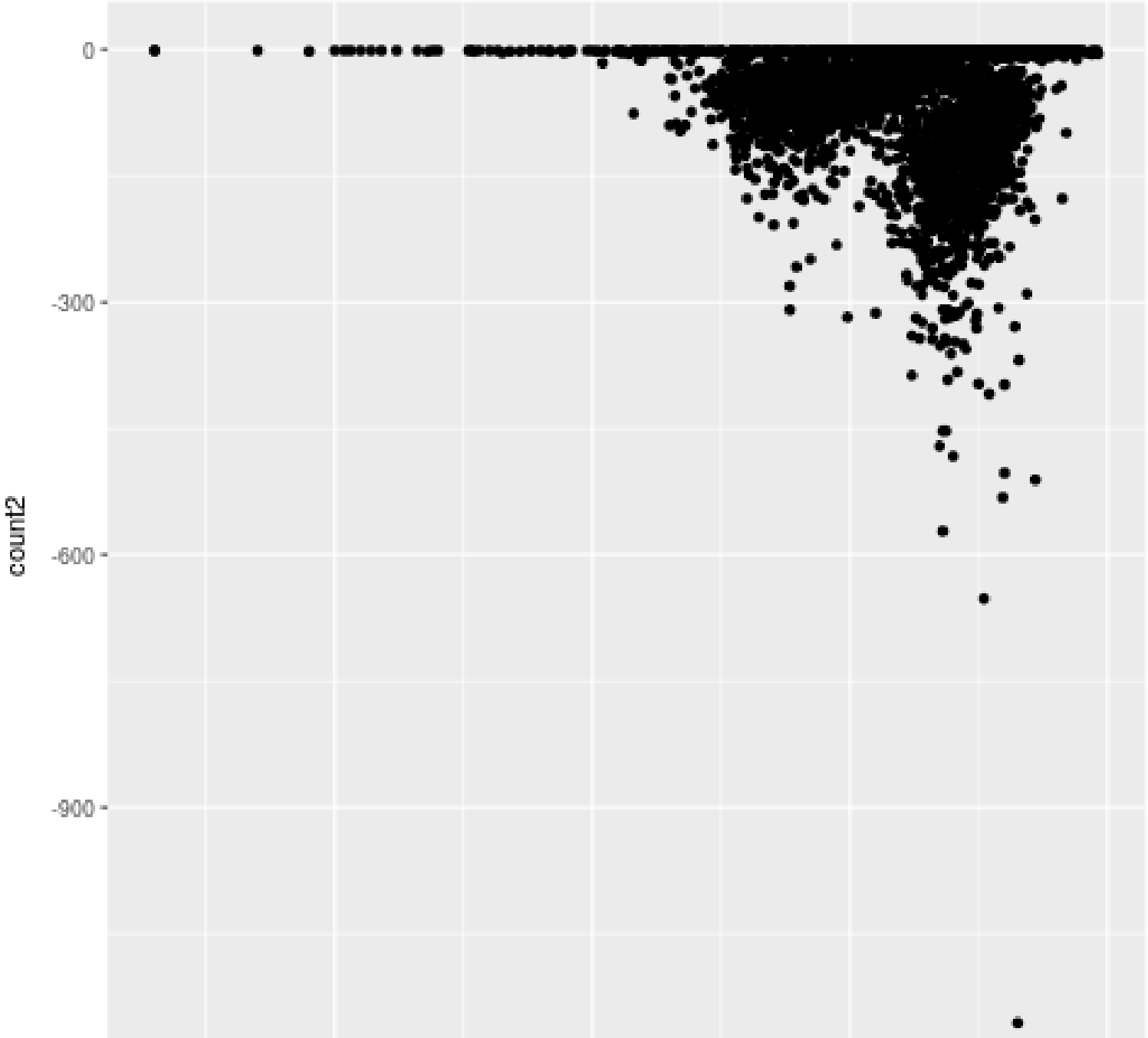
Standard Scatter



Flipped Scatter



Other side



What does this tell us?

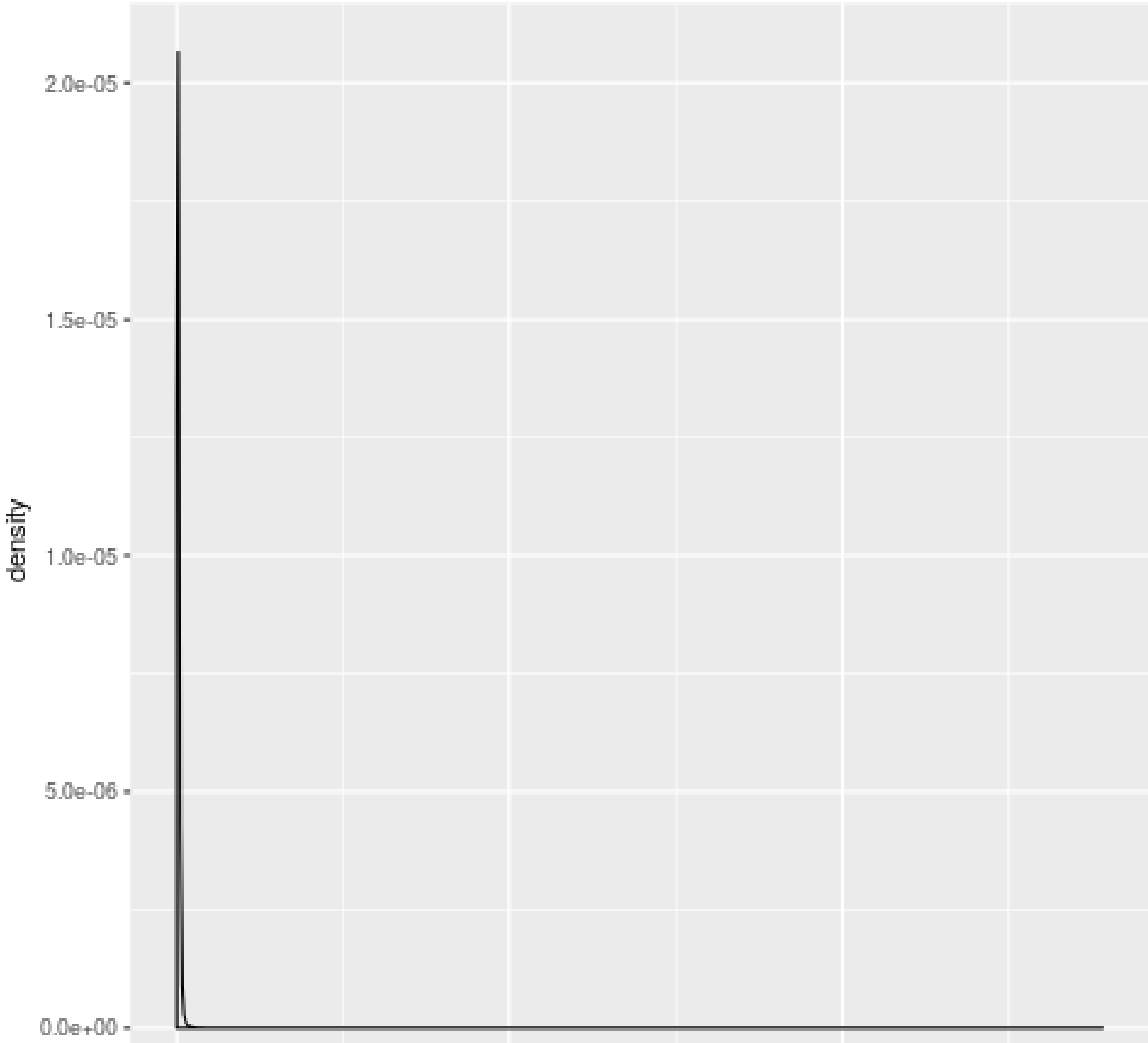
- We have a base level of assumptions that we bring to graphics (especially statistical graphics)
- Most of these appear to have been formed by Descartes
- When these assumptions are subverted, expect problems

Simple Statistical Graphics

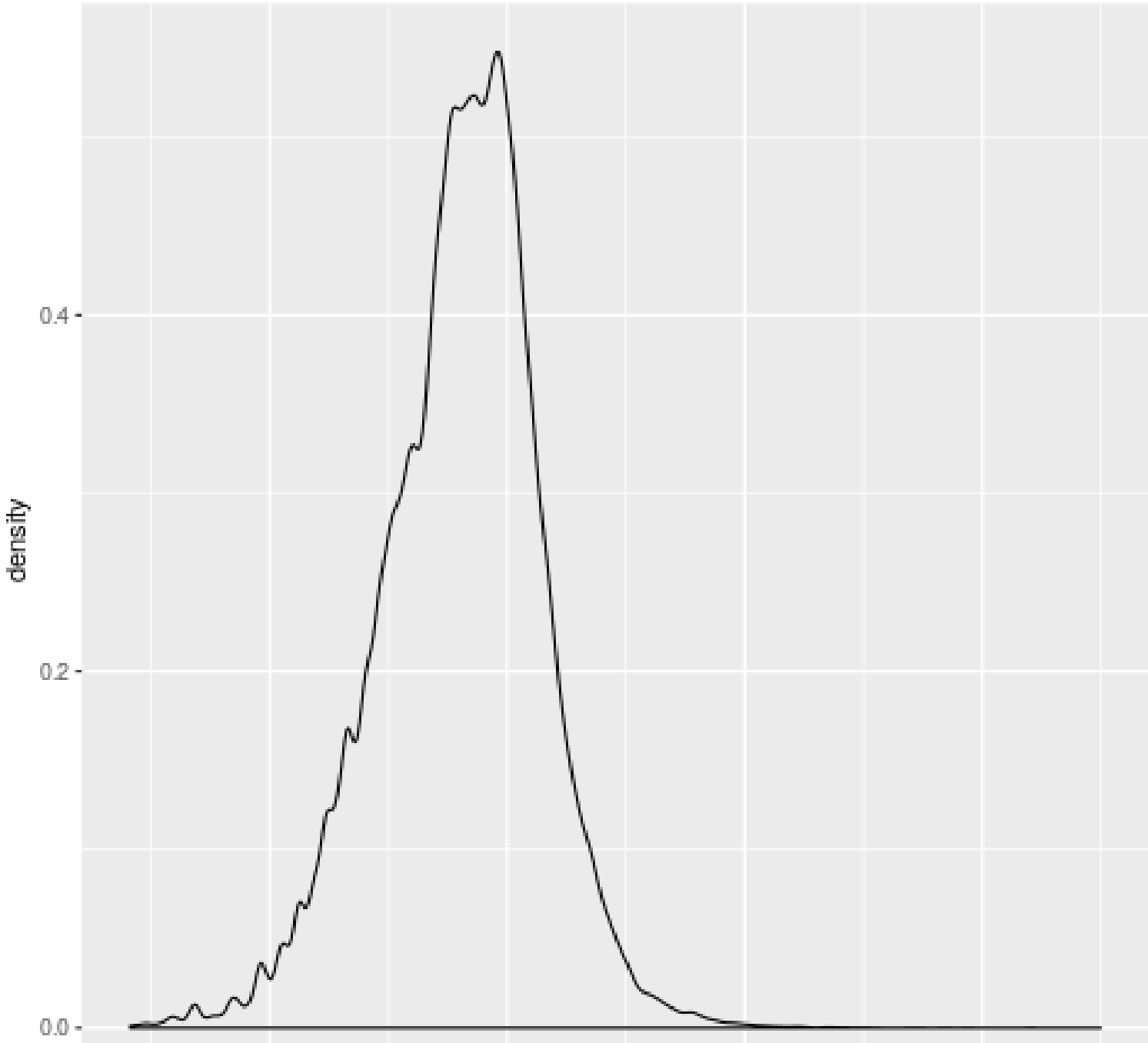
- Graphs excel at showing relations between things
- Consider the difference between quantiles of a variable, and a density plot
- For example, the price of houses:

0%	5079
10%	55000
20%	85000
30%	115000
40%	145000
50%	175000
60%	214000
70%	255505
80%	315000
90%	430000
100%	139165000

Density Plot



Better Density Plot



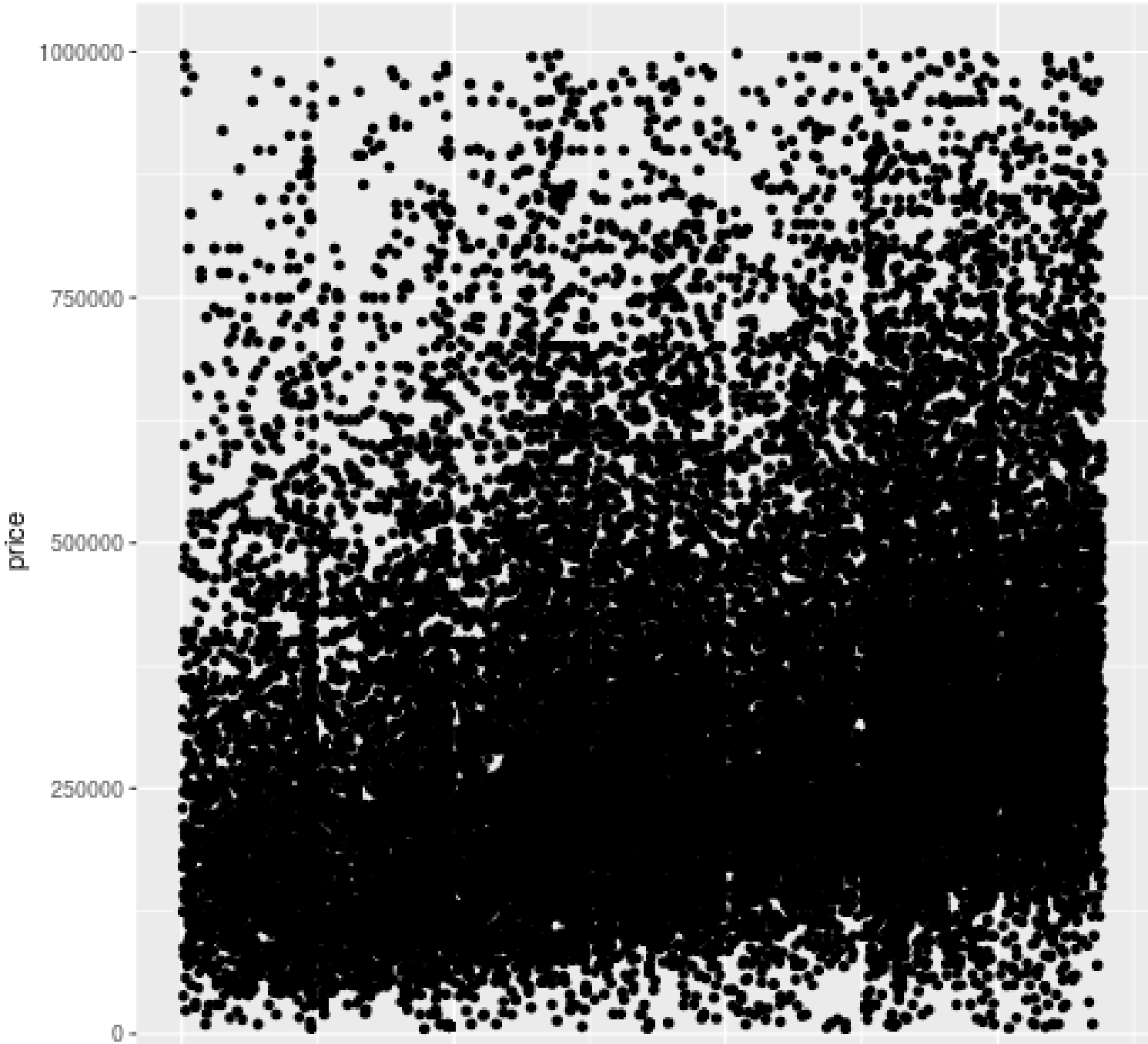
Transformations

- Useful to get a better sense of the data
- Have a bunch of assumptions (what's the log of -1)
- Can be used to deceive very, very easily
- Really really useful in everyday practice

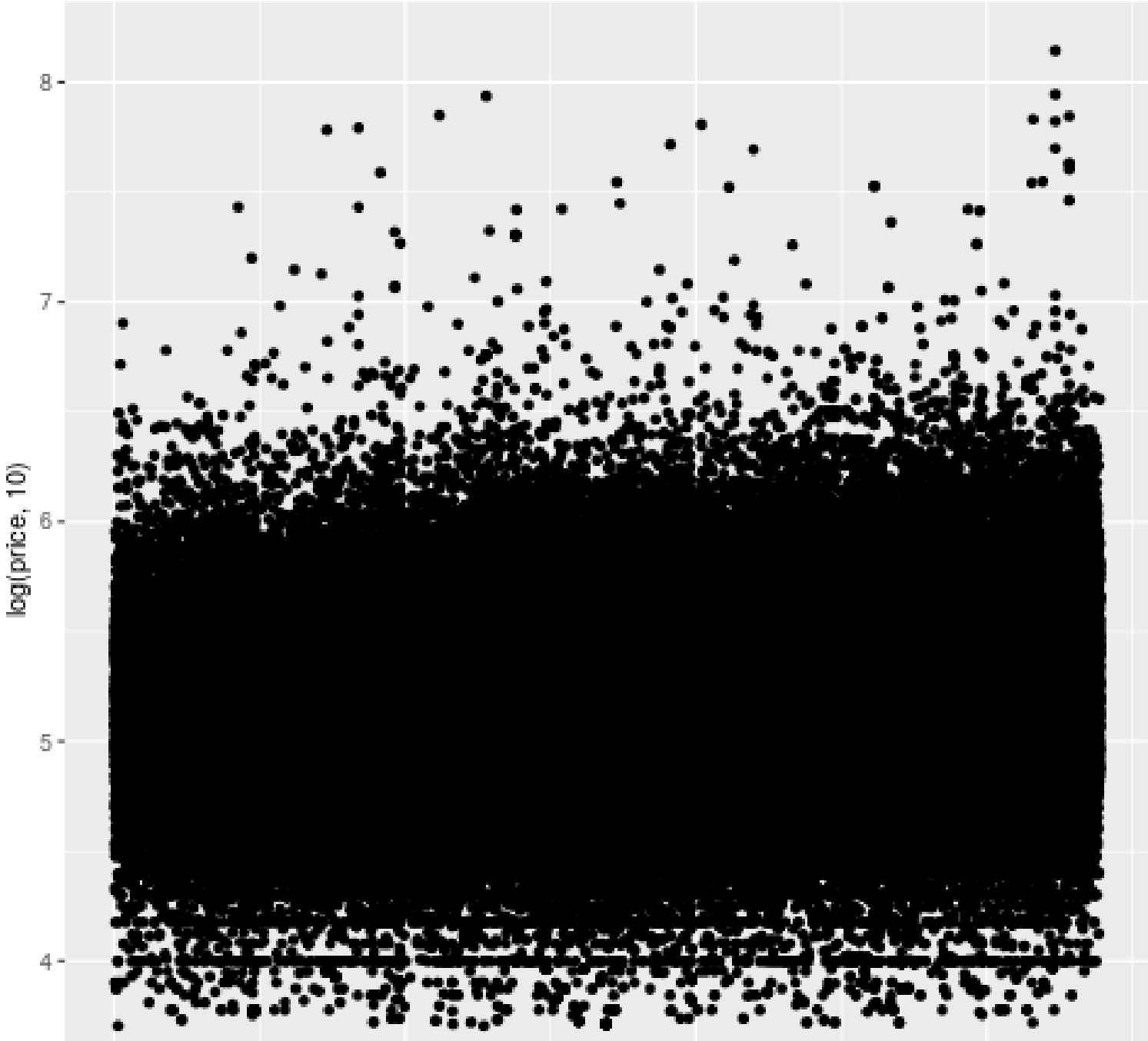
Getting the sense of things

- Picking the right visualisation for the data is important

Sampling and Plotting

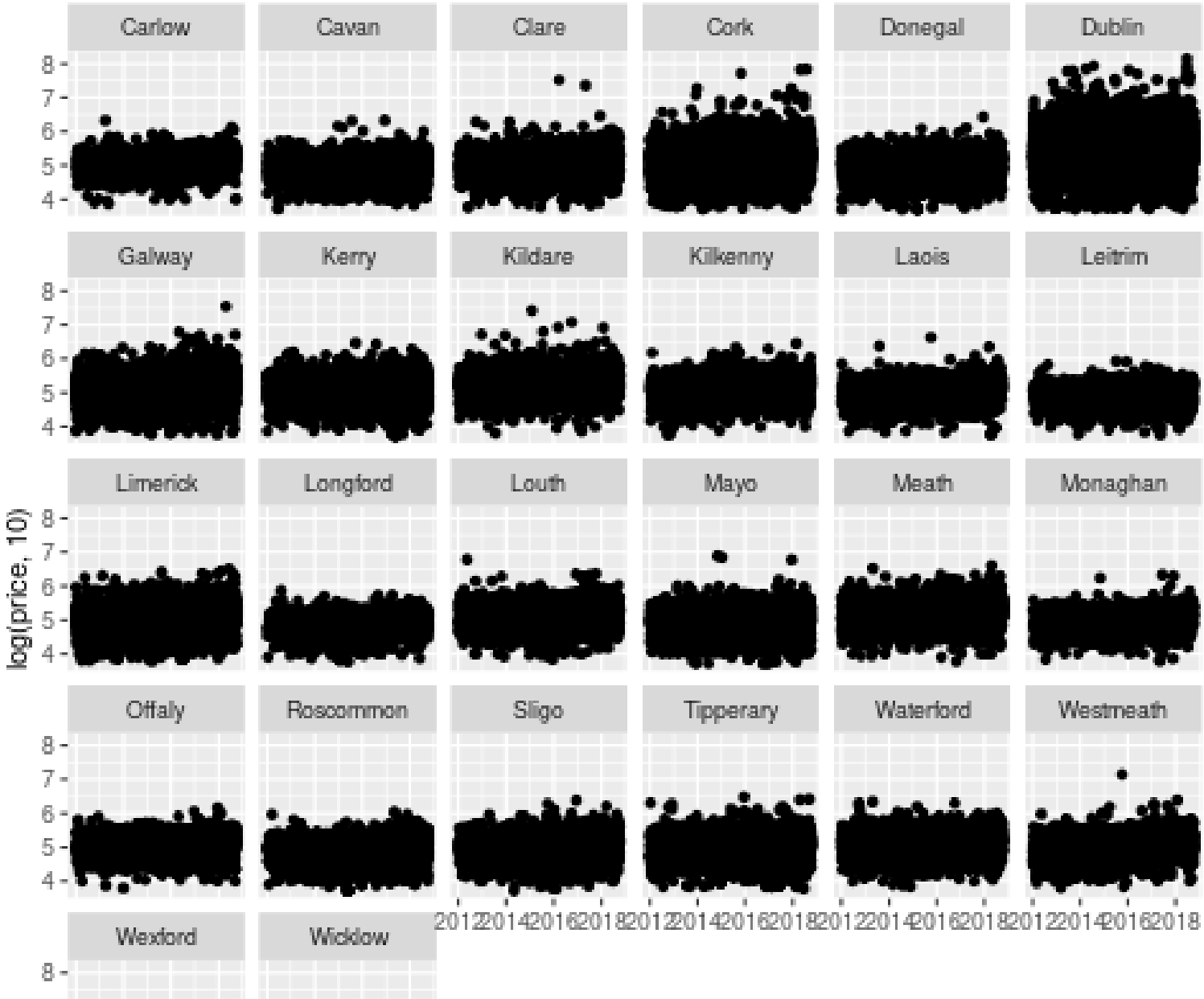


Transformations Help



No data is an island

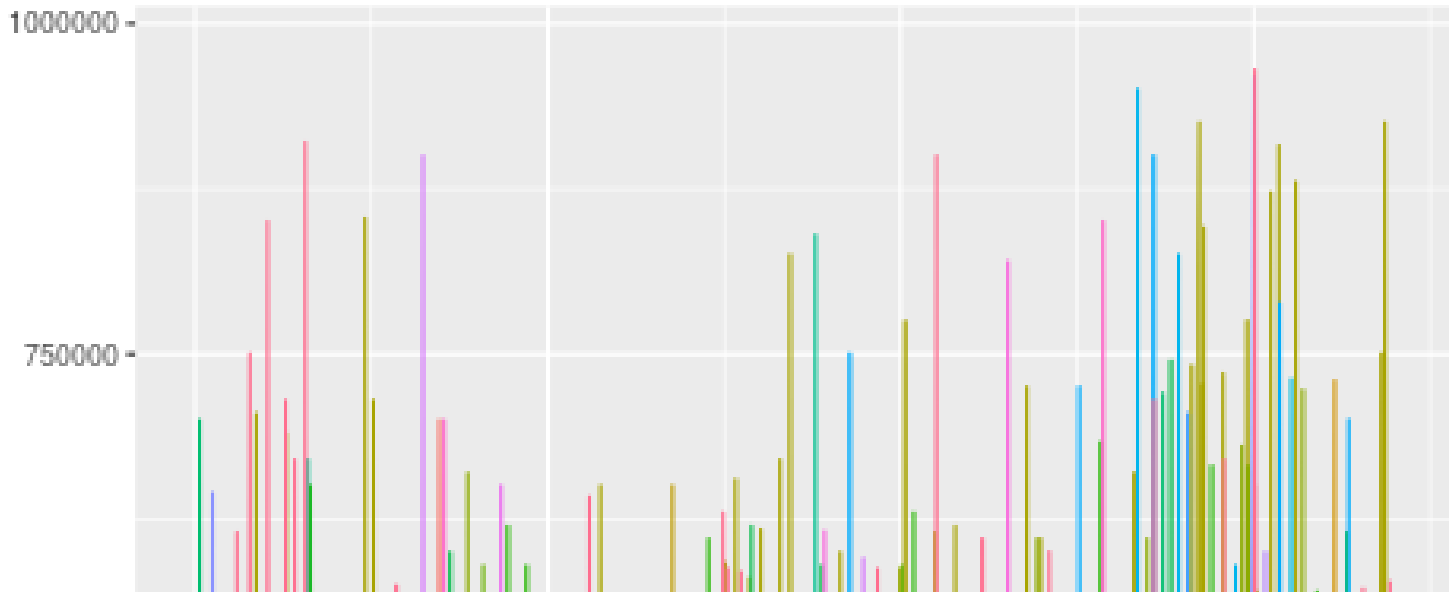
- The first obvious thing is to split by county, right?



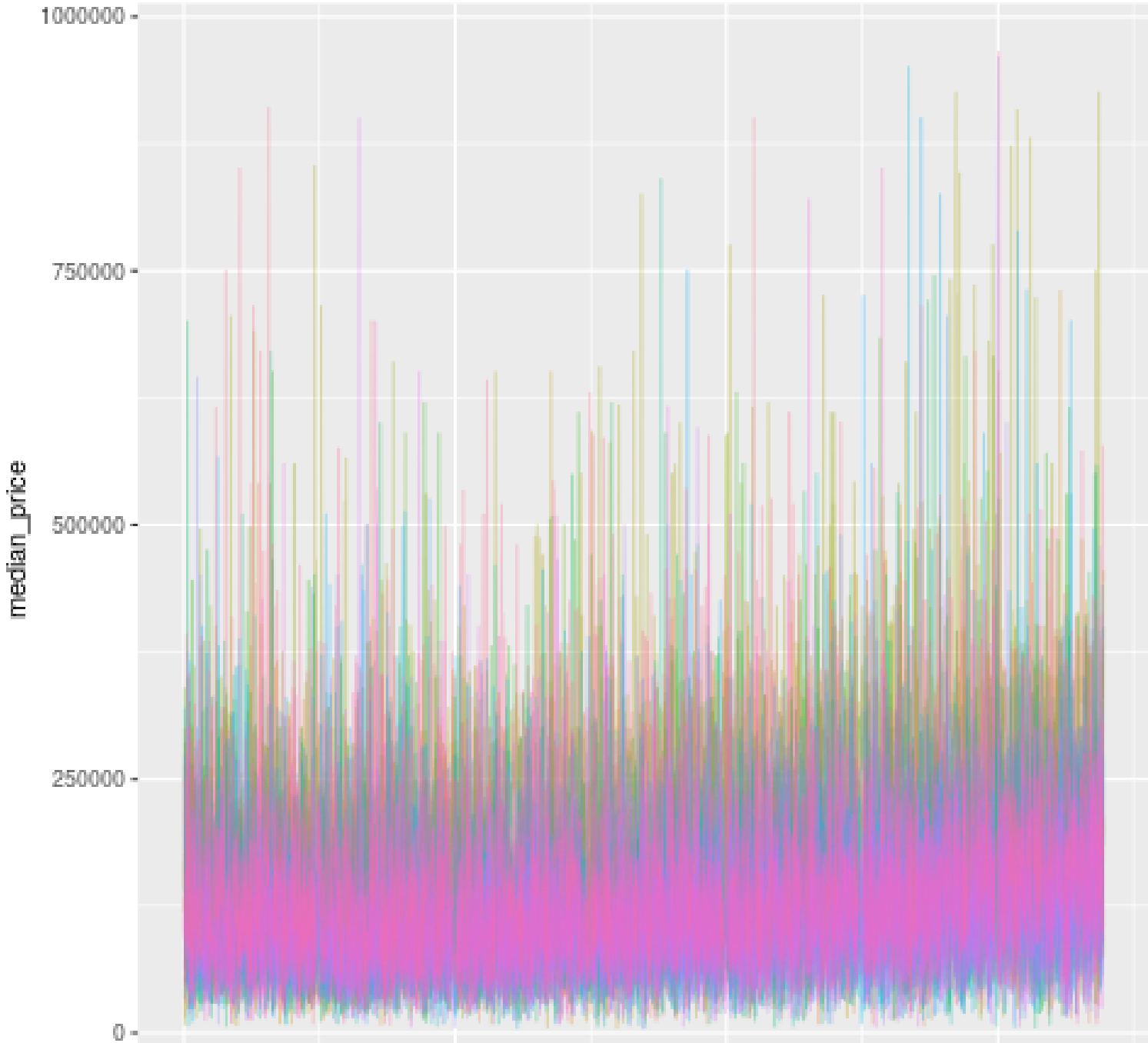
Summarisation

- The obvious answer is summarisation

```
county_daily <- ppr_gc2 %>% group_by(sale_date, ppr_county,
  summarise(count=n(), min_price=min(price),
            median_price=median(price),
            max_price=max(price)) %>%
  mutate(min_to_median=min_price/median_price,
         max_to_median=max_price/median_price,
         max_to_min=max_price/min_price)
ggplot(county_daily, aes(x=sale_date, y=median_price, colou
```



Reducing Alpha kinda works. . .



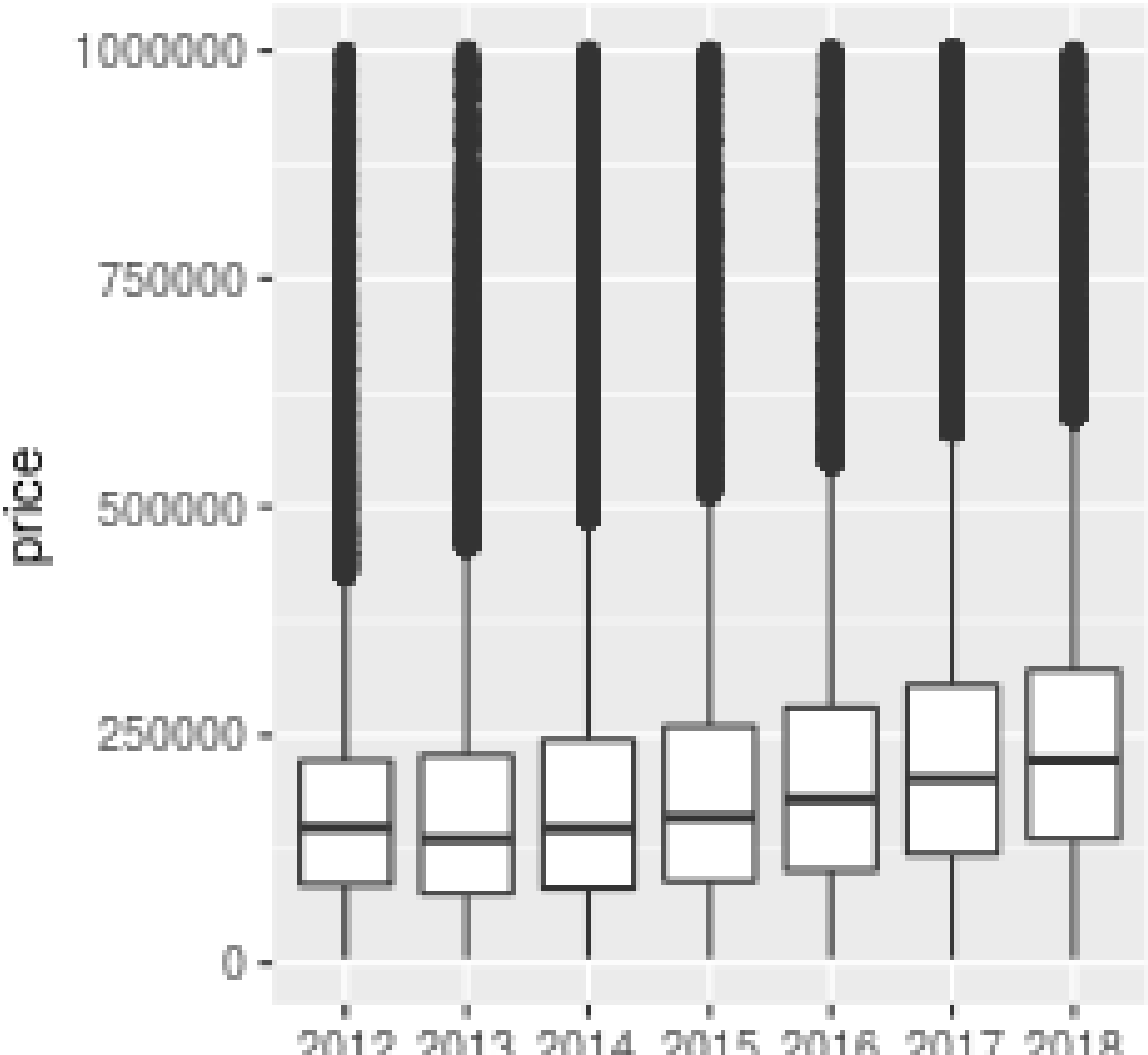
A redundant faceting variable

- We just group by a higher level variable

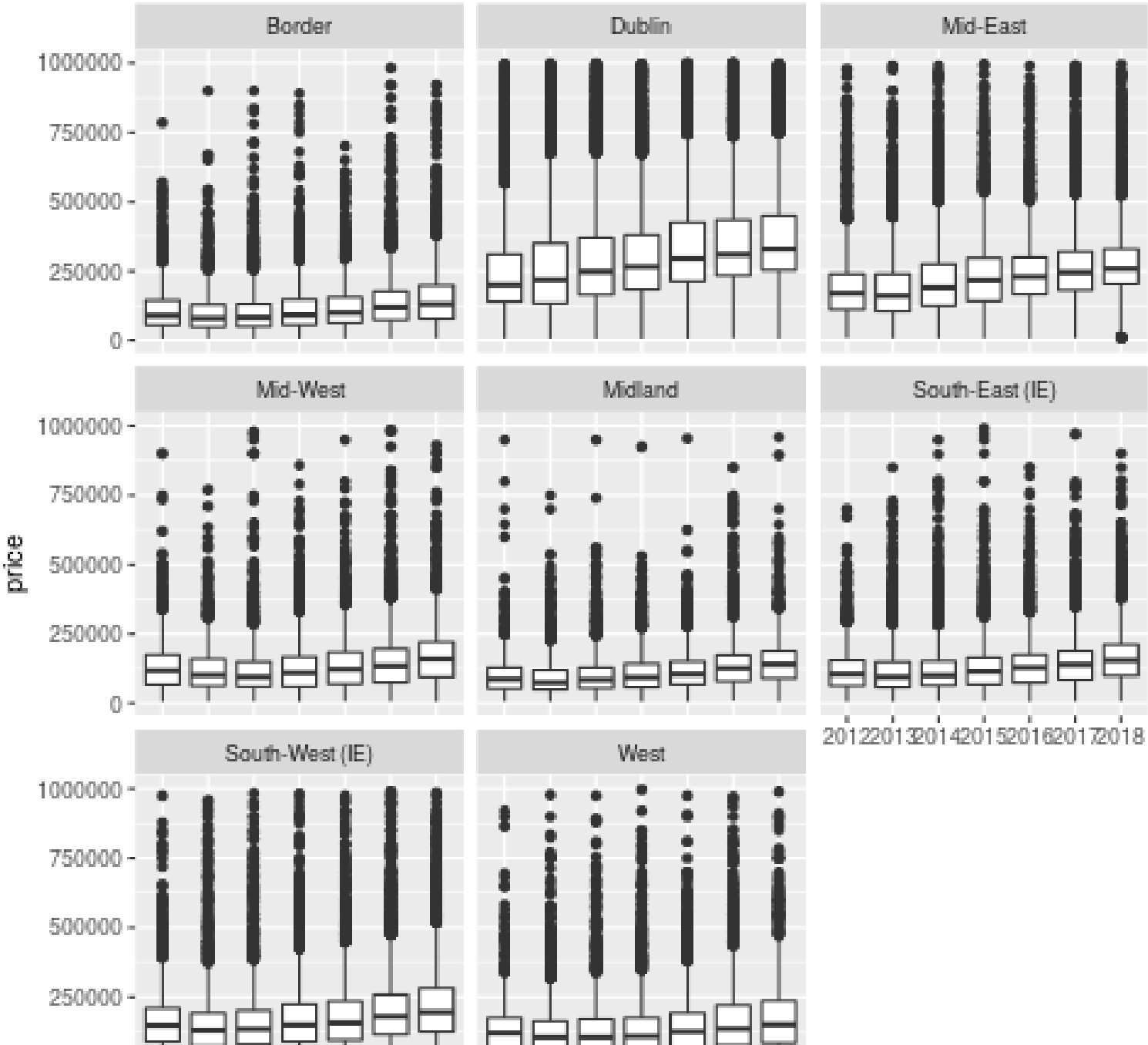
WTF?

- This is one of the major advantages of visualisation:
  - it helps to (dis)confirm your assumptions
  - given that we have too many lines in the various groupings,we know that somethng has gone horribly wrong
  - in this case, it's a mismatch between two different types of data

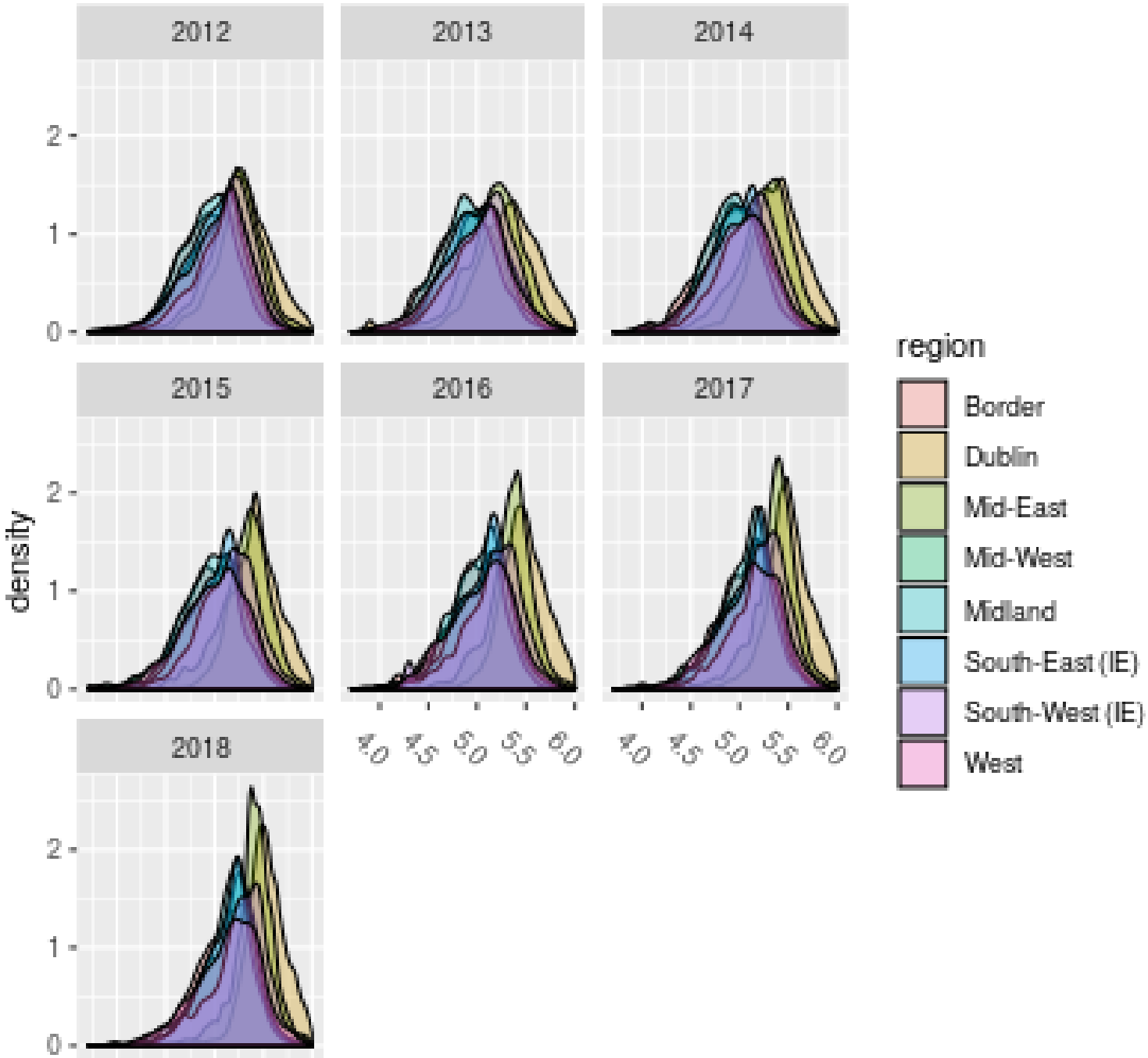
Distributions (i.e. boxplots)



Faceting, redux



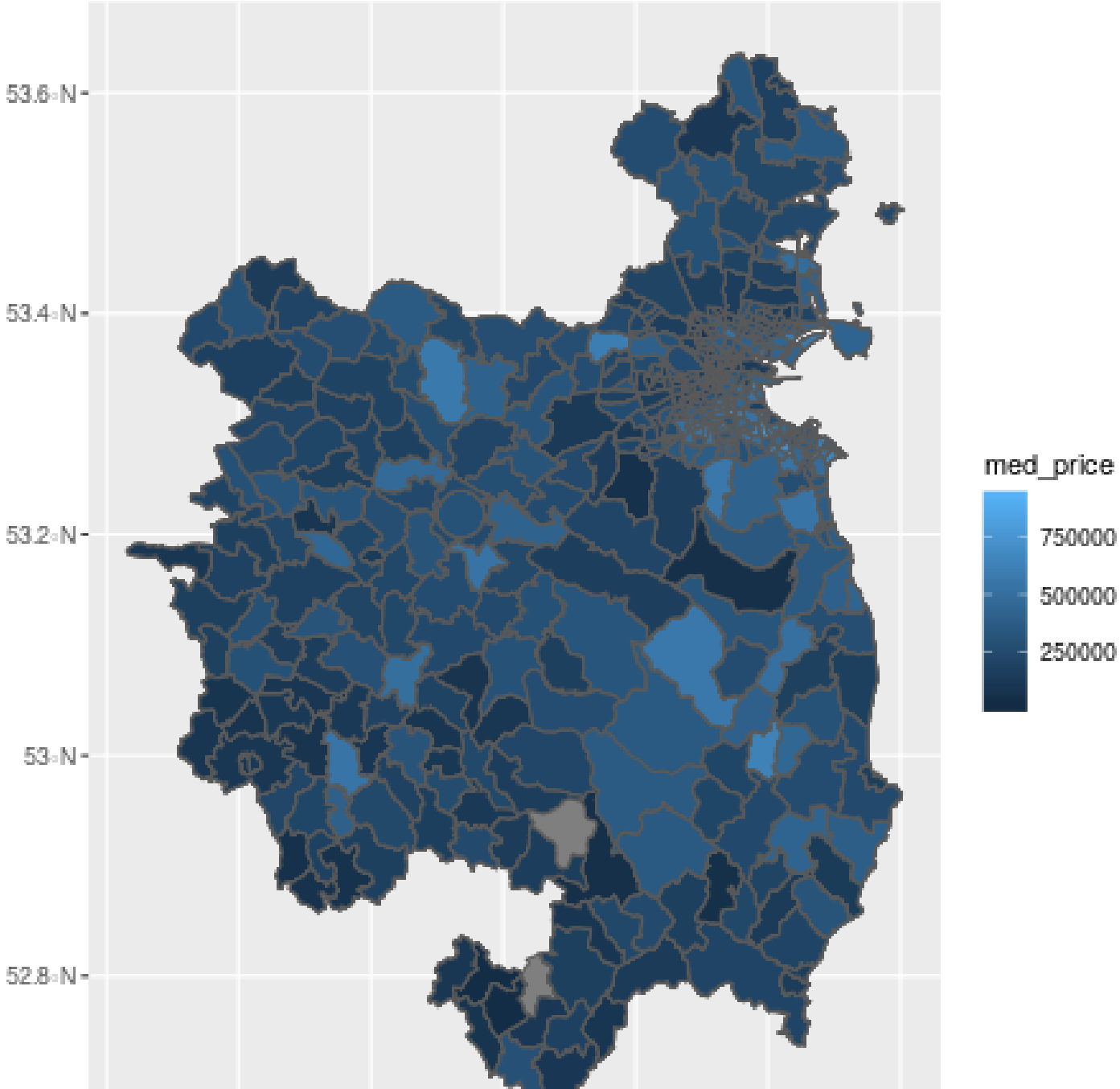
Distributions over Time, Redux



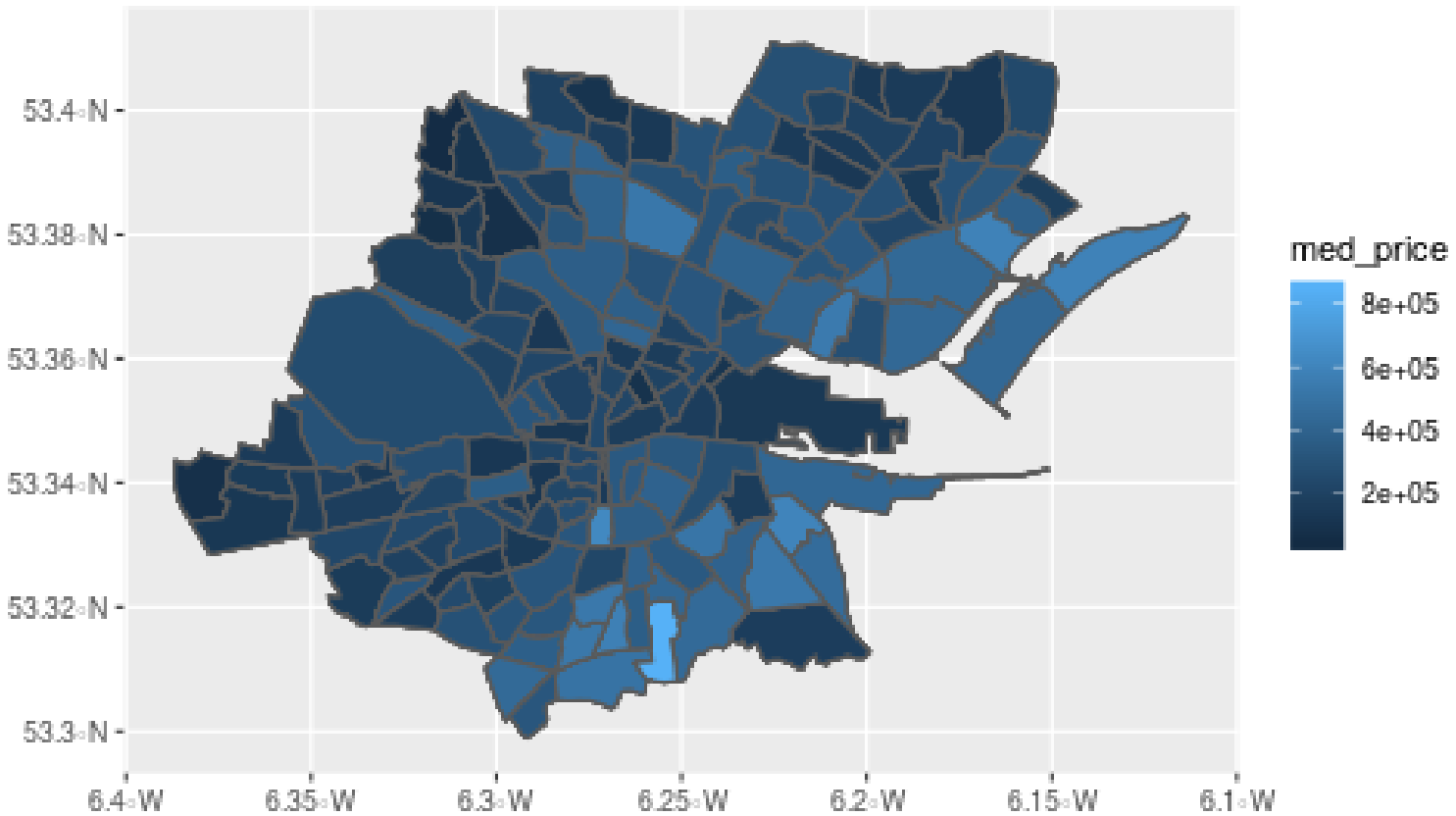
Spatial vs Temporal

- line plots vs maps
- time versus space
- both provide insight into
- pick one, difficult to do both

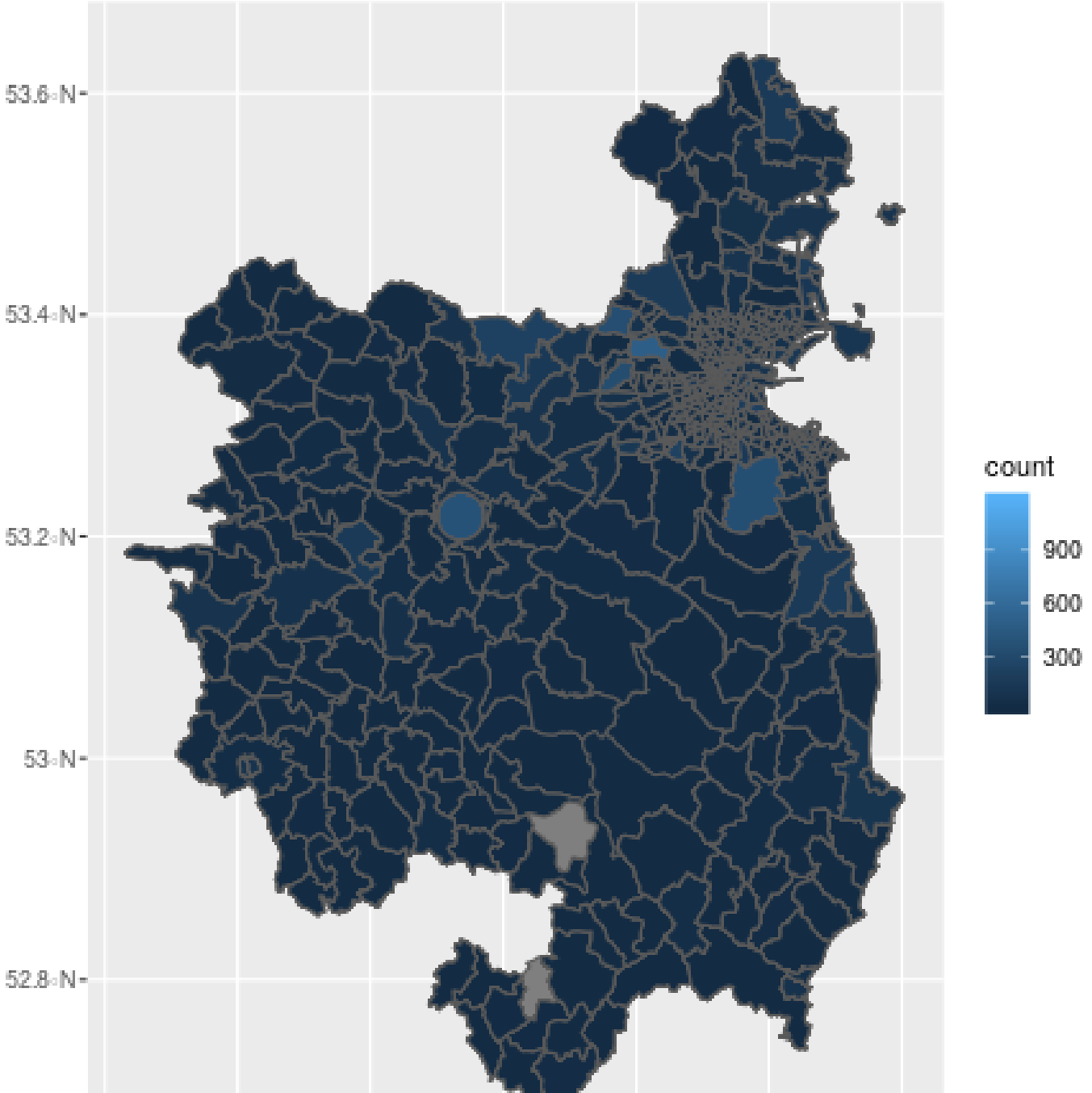
Line plots ignore space, maps ignore time



Dirty Oul Town



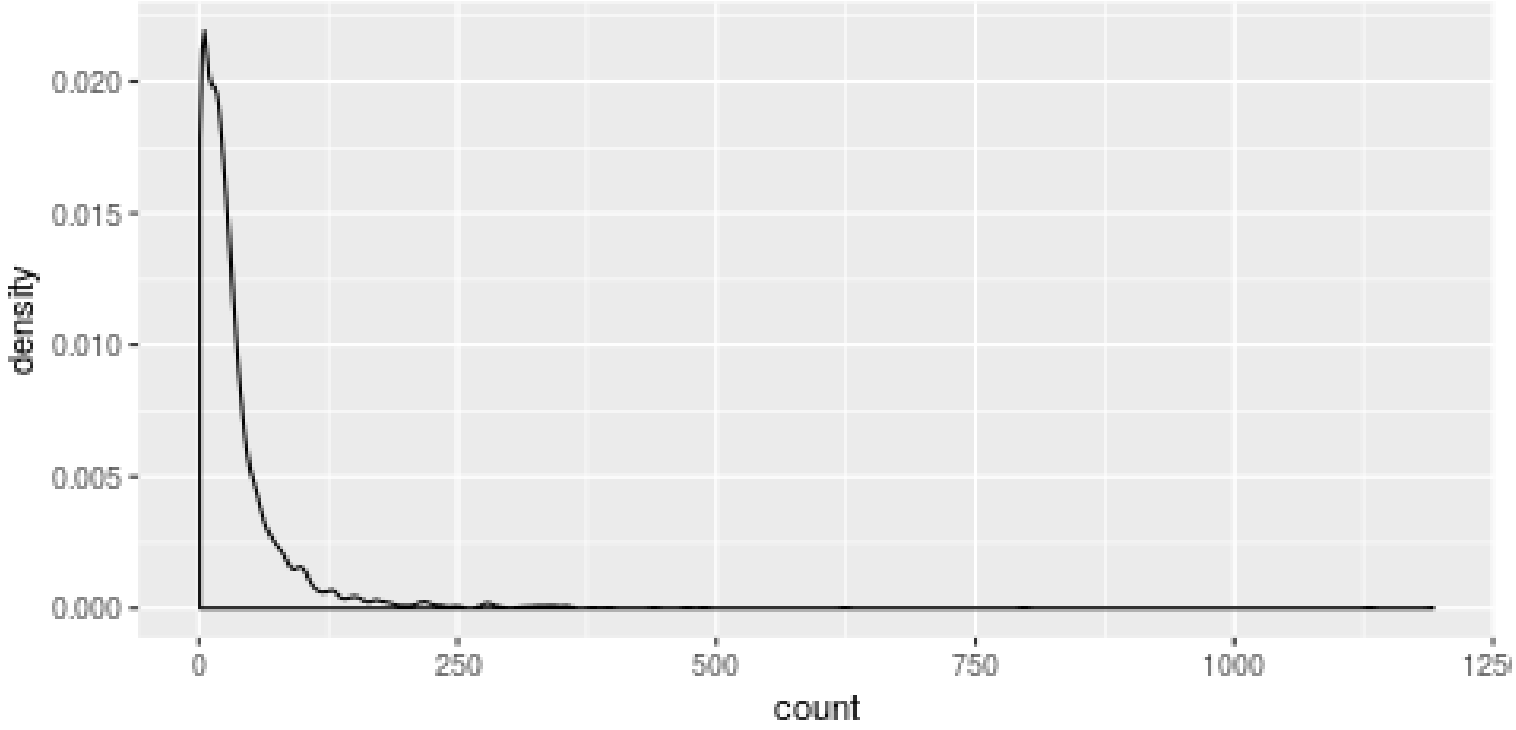
Counts tell a different story

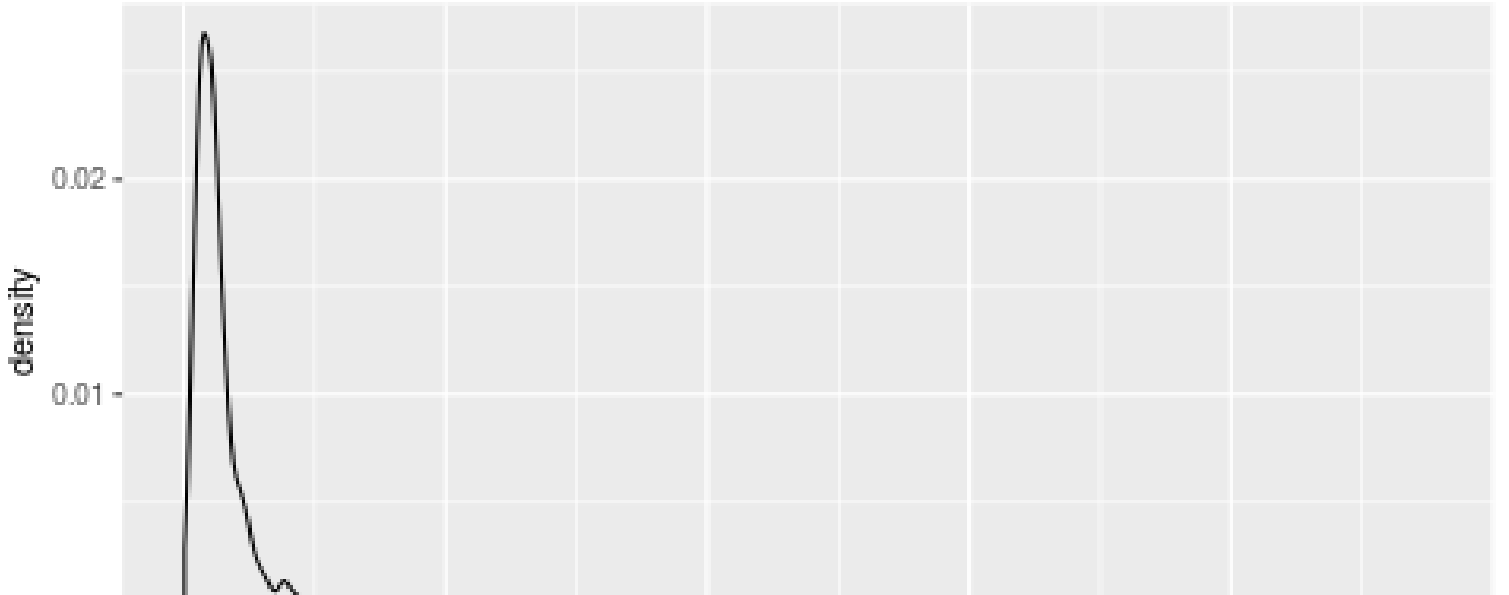


Dublin City (again)

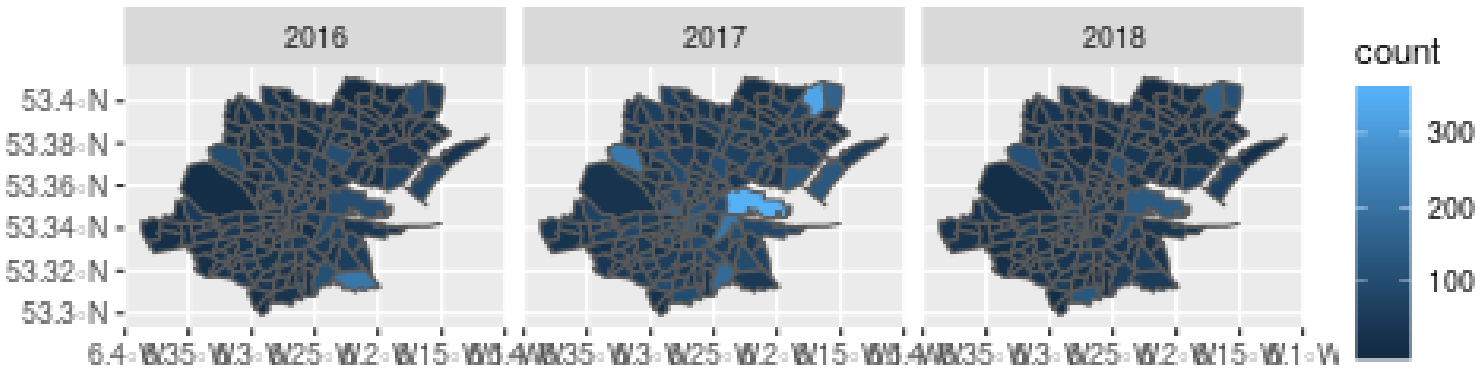
```
filter(elec_m_sf, COUNTYNAME=="Dublin City") %>% ggplot(ele
```

Density Plots to help maps

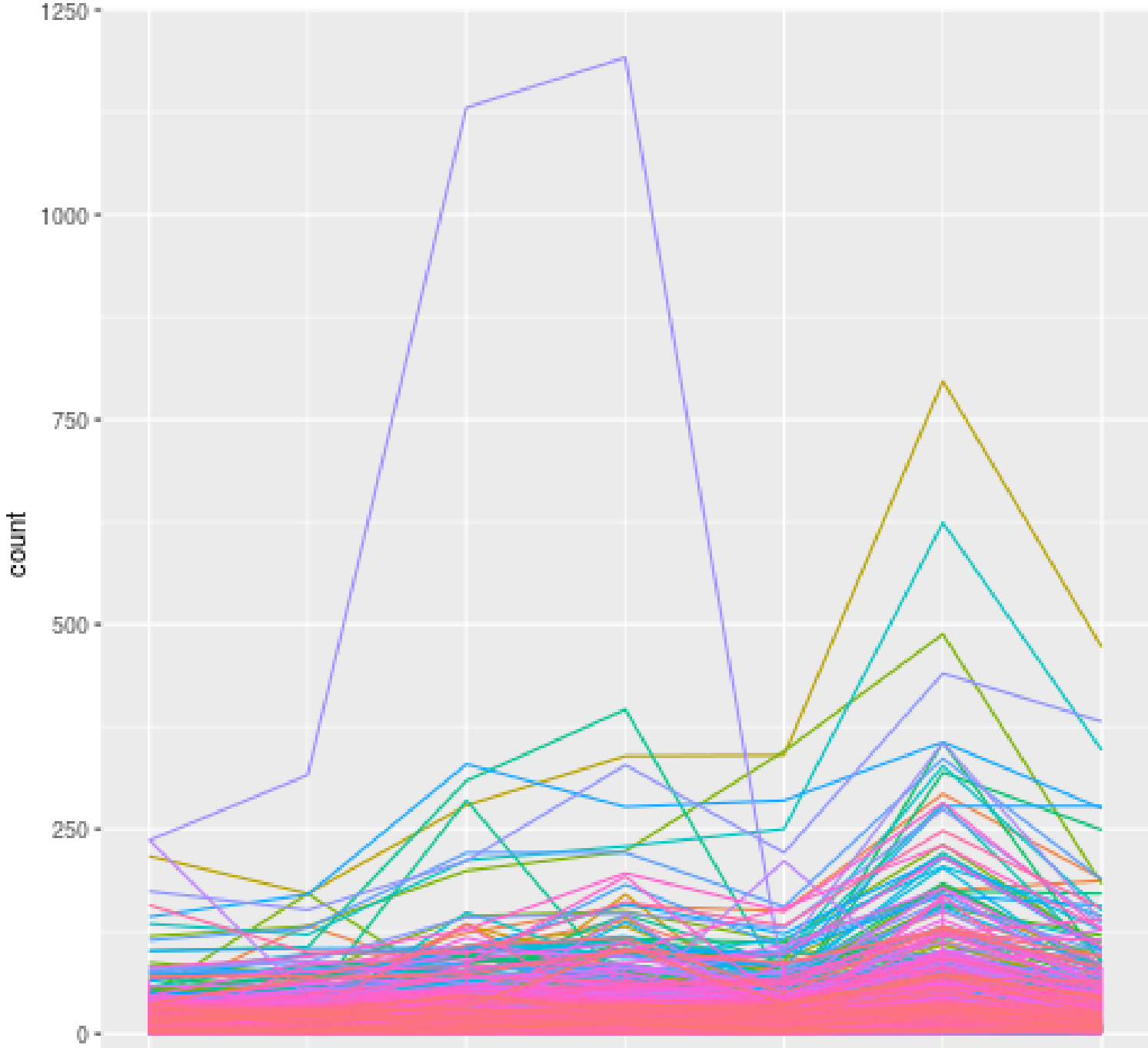




Maps over Time



Lines for Time



Interactivity and Dashboards

- Can show both time and space
- for reporting, these are essential
- Much more effort from a software-engineering perspective [3]

[3] for me, at least

Performative vs Presentation

- Two types of graphs:
  - for yourself
  - for other people (and different audiences need different things)

Performative Graphics

- These are used to help you understand a problem
- typically created in an iterative fashion
- often move from data transformation to visualisation and back again (like this talk)

Presentation Graphs

- To some extent, your job with presentation visualisations is to tell a story
- hopefully, it will be nuanced, but that isn't a requirement [4]
- Often good to show smooths as opposed to raw data
- raw data is often ugly
- need for care here, as this should only be done where there is a clear effect

[4] and in fact, it may be better to remove all nuance from the presentation and provide a longer document with all the failed approaches and hacking needed to actually reproduce your results

Advice

- As few as possible
- One clear message
- Repeat yourself
- Remove nuance

As few as possible

- There should be no extraneous graphs
- Each graph should have a clear purpose
- Smooths are really effective

One Clear Message

- You should only be telling one story at a time
- People are easily confused
- Especially in an oral presentation
- Backup docs should contain nuance

Repeat Yourself

- This is the key to helping people retain information
- This is easier once you know the story
- Say what you want to say, say it, then say what you said

Remove Nuance

- This varies by audience
- Salespeople may just want the results
- colleagues may want to see the code
- most people just want a high level explanation
- Nuance should be present, just not in a presentation

Conclusions

- Everyone bring assumptions to visualisations
- Make sure that you take advantage of this
- Visualisation is primarily a tool for communicating with yourself
- Iterative process, even bad graphs can teach you something
- Secondarily, it's a tool for communicating with others
- When using visualisations with others, keep it simple