Welcome to DATA 202!

K Arnold

Welcome to DATA 202!

As you enter...

- Make a **name card** (name on front *and back*)
- Sit next to someone you **don't know well**
- Introduce yourself. *Share either*:
  - Something you're passionate about outside of this class: a cause, a subject, a hobby, etc.
  - Something you're excited about for this semester

What might this mean for us working with data?

> Seek good and not evil, that you may live,
> and so the Lord, the God of hosts, will be with you, just as you have said. Hate evil and love good, and establish justice in the gate;
> it may be that the Lord, the God of hosts, will be gracious to the remnant of Joseph.
> ...
> I hate, I despise your festivals, and I take no delight in your solemn assemblies. Even though you offer me your burnt offerings and grain offerings, I will not accept them, and the offerings of well-being of your fatted animals I will not look upon. ...
> But let justice roll down like water and righteousness like an ever-flowing stream.

Amos 5:14-15, 21-24 NRSV

Opening Prayer

From the apostle Paul's letter to the Philippians:

> This is my prayer:
> that your love may abound more and more
> in knowledge and depth of insight,
> so that you may be able to discern what is best and may be pure
> and blameless for the day of Christ, filled with the fruit of
> righteousness that comes through Jesus Christ—to the glory and
> praise of God.

What is this course?

- **visualization**: communicating data to humans
- **modeling** and validation: using data to perform tasks
- but first, **data wrangling**: How to transform data into the structure you need for visualization and modeling

...using **computing** in the R language (`#rstats`) and occasionally other tools

Where are we going?

- Introduction and Toolset (week 1)
- Data Visualization Design and Implementation (weeks 2 and 3)
- Data Wrangling (single-table and multiple-table) (weeks 4 and 5)
- Data Tidying (week 6)
- Midterm project: redesign and recreate a visualization (weeks 7 and 8)
- Making Estimates using Data (week 9)
- Making Predictions using Data (Supervised Learning) (weeks 10 and 11)
- Supervised Learning using Trees (weeks 12 and 13)
- Effective and Interpretable Models (week 14)
- Final Project (week 15)

Discussions on *ethics* and *perspectives* are woven in throughout the course.

Optional material

- Clustering (Unsupervised Learning)
- Databases and APIs
- Text Data
- Geospatial Data
- Audio and Image Data

Feedback from Last Year

"What aspects of this course most helped your learning?"

> **Lab**. **Homework**. "forced our brains to work."

"What additional or different things could you have done to enhance your learning?"

> "gone to more office hours", "gone over the exercises and homework more"

Structure

Overall, *flipped*.

- Read and watch *at home* (at 2x speed if you want)
- Solve problems *together in lab*

Weekly Rhythm

In Class

- **Monday** class: discuss a previous assignment
- **Wednesday** class: work on lab exercises
- **Friday** class: weekly quiz, finish lab, start on homework

At home

- **Monday**: Preparation and Lecture quizzes due
- **Tuesday**: Previous week's labs due
- **Wednesday**: Post or reply in a discussion forum
- **Thursday**: Previous week's homework due

At-home activities are due at end-of-day, scheduled so you don't have to work on weekends

Ways to participate

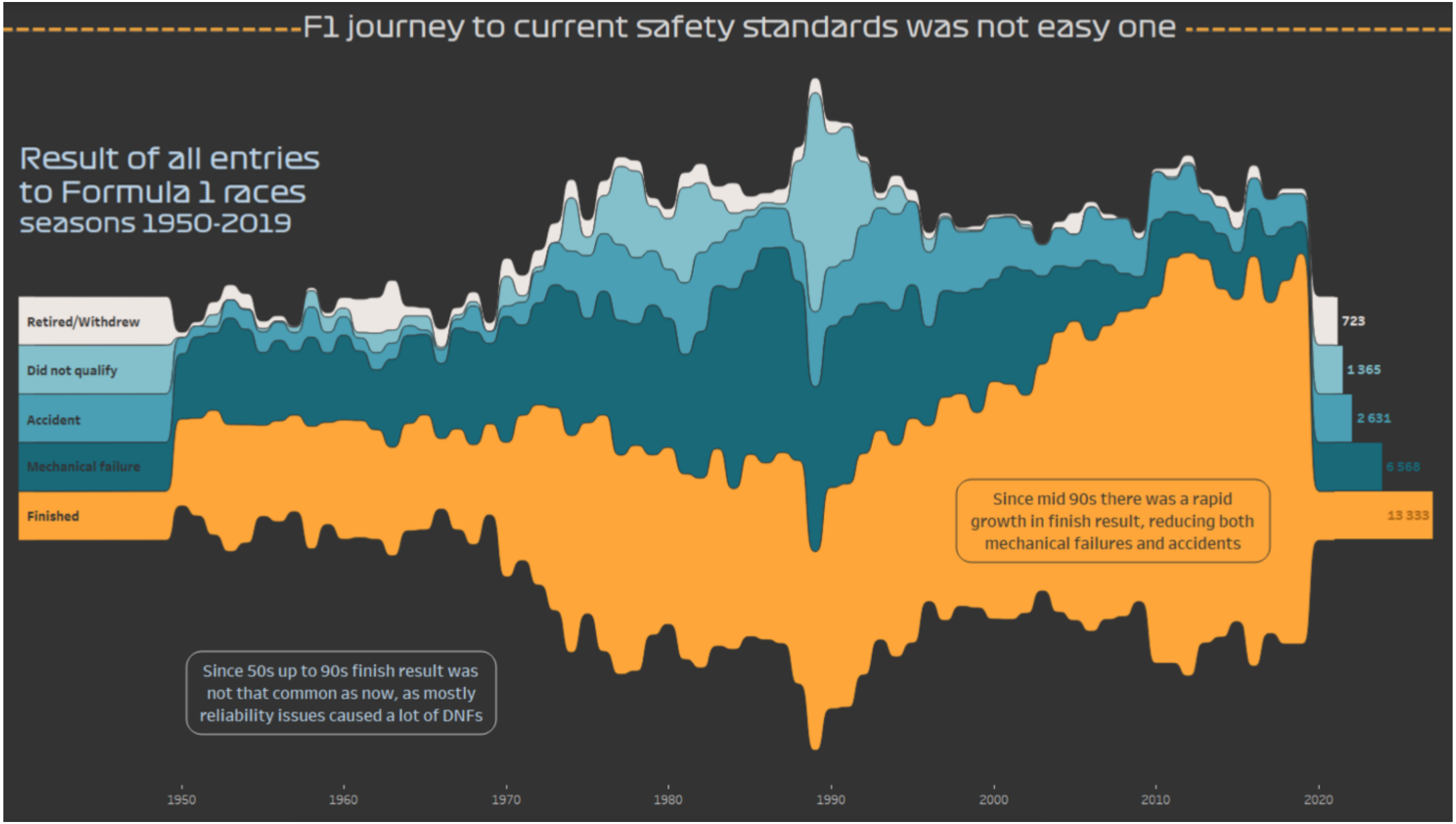
- Hot-seat coding
- Hot-seat explaining
- Present an explanation
- Present a homework or lab
- Give feedback on project presentations

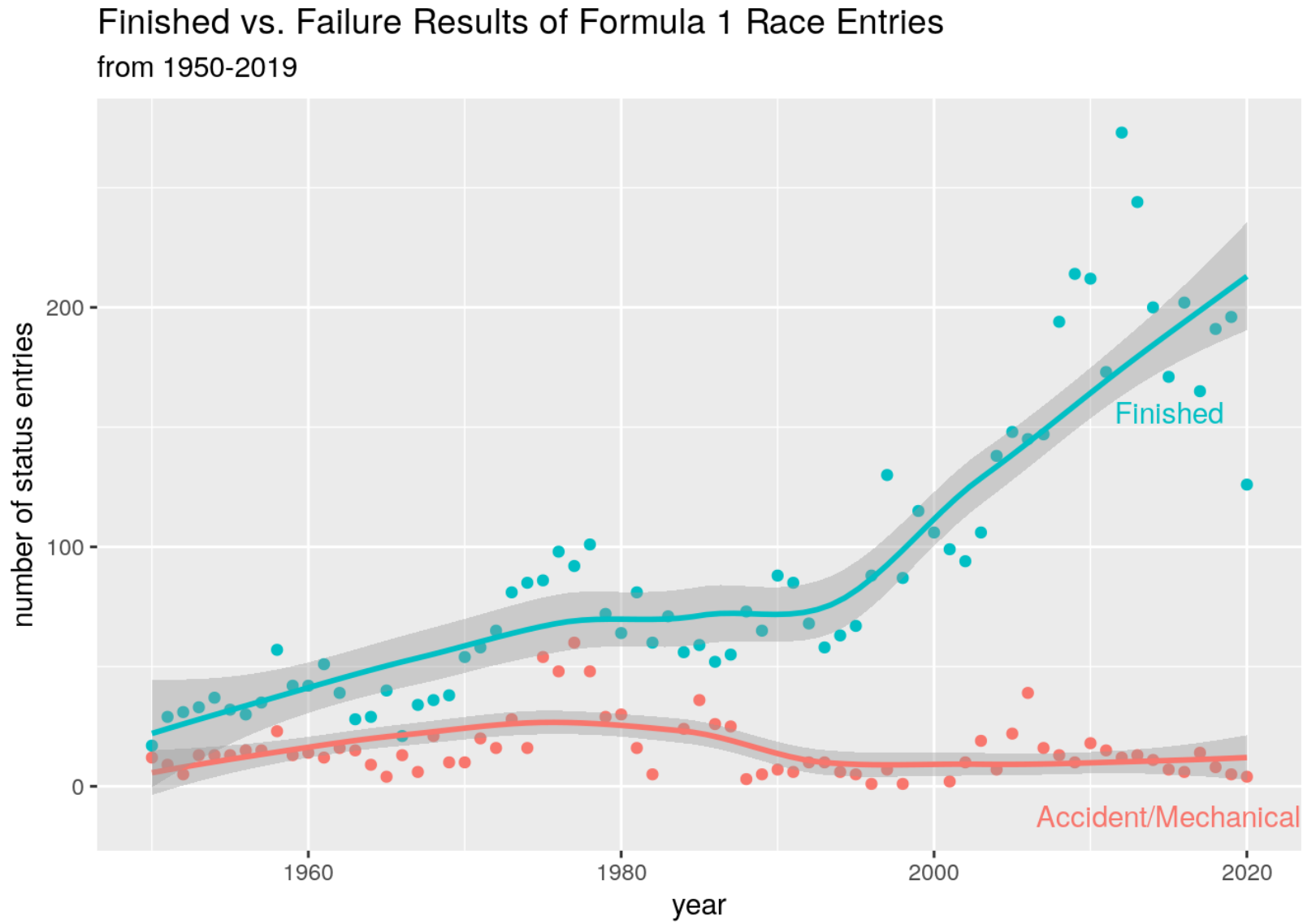
Intentional imperfection: presenters *should make mistakes* so we can all learn.

Technology We'll Use

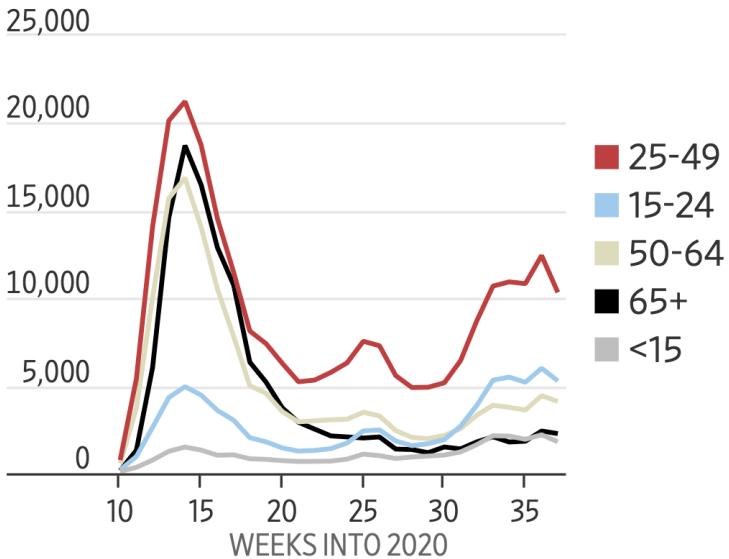
- **RStudio**: A powerful environment for working with data
  - Most students will use Calvin's installation `https://rstudio.calvin.edu/`
  - You can also install it on your own computer like I do
- Ed Discussion (**EdStem**)
  - Organized Q&A, tips, announcements
  - Anonymous questions (fear not!)
  - Ask questions in class without interrupting

Example Projects


F1 journey to current safety standards was not easy one
Result of all entries to Formula 1 races seasons 1950-2019
Retired/Withdrew
Did not qualify
Accident
Mechanical failure
Finished
723
1 365
2 631
6 568
13 333
Since mid 90s there was a rapid growth in finish result, reducing both mechanical failures and accidents
Since 50s up to 90s finish result was not that common as now, as mostly reliability issues caused a lot of DNFs
1950
1960
1970
1980
1990
2000
2010
2020

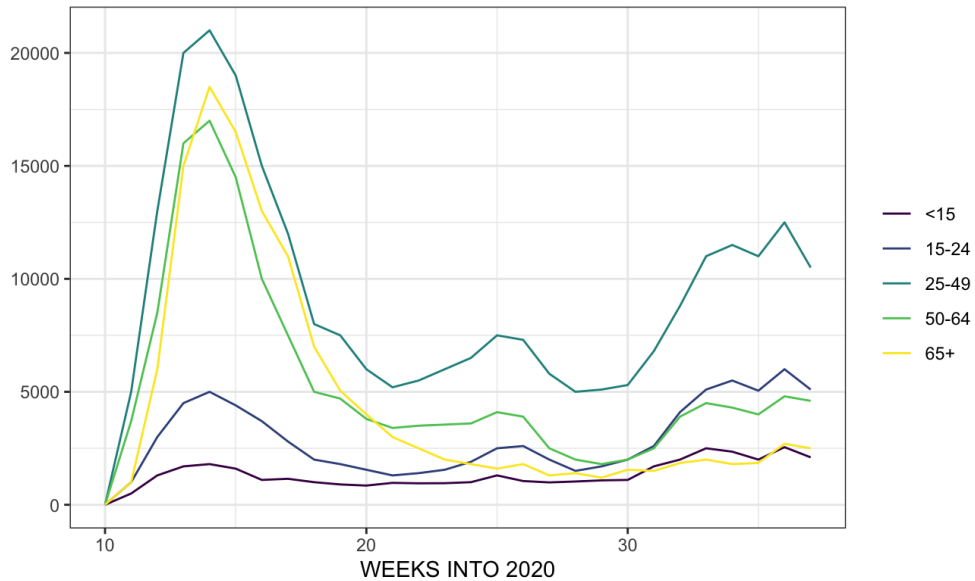

Finished vs. Failure Results of Formula 1 Race Entries
from 1950-2019
number of status entries
200
100
0
1960
1980
2000
2020
year
Finished
Accident/Mechanical

Confirmed Covid-19 cases in 17 European countries by age group



Note: Countries include Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, Germany, Iceland, Ireland, Latvia, Luxembourg, Malta, Netherlands, Norway, Poland, Portugal and Sweden
Source: European Surveillance System, part of the European Center for Disease Prevention and Control

Confirmed Covid-19 cases in 17 European countries by age group



Note: Countries include Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, Germany, Iceland, Ireland, Latvia, Luxembourg, Malta, Netherlands, Norway, Poland, Portugal and Sweden
Source: European Surveillance System, part of the European Center for Disease Prevention and Control

Example final projects

- Predict how much a used car will sell for
- Forecast how much electricity will be used
- Predict how much a plane flight will cost

Our Goals

- *Skill*: how to do these things
- *Knowledge*: understanding the underlying concepts
- *Dispositions*: wisdom in practicing these skills

Data Dispositions

Humility

Challenge: data feels powerful, people listen to what you use it to say.

So we will practice:

- Citing all sources (for both data and process)
- Acknowledging limitations
- Noticing and reporting our analysis decisions and possible alternatives
- Validation of results

Integrity

It’s tempting to say something that isn’t entirely true, or to manipulate the collection/analysis/reporting process to yield the answer you want.

So we will practice:

- Evaluating claims that others use data to make
- Clearly articulating our analysis decisions and rationale
- Reproducibility
- Using exploratory analytics to validate data against assumptions

Hospitality

We can choose to use our tools to elucidate and clarify, rather than obscure.

So we will practice:

- Clear visual communication
- Clarity of code and process
- Writing explanations that are accessible and appropriate to audience.

Compassion and Justice

Data Science can both cause harm and reveal it.

So we will:

- Study examples of how data might cause harm
- Study examples of how harm might be mitigated or revealed

A worked example

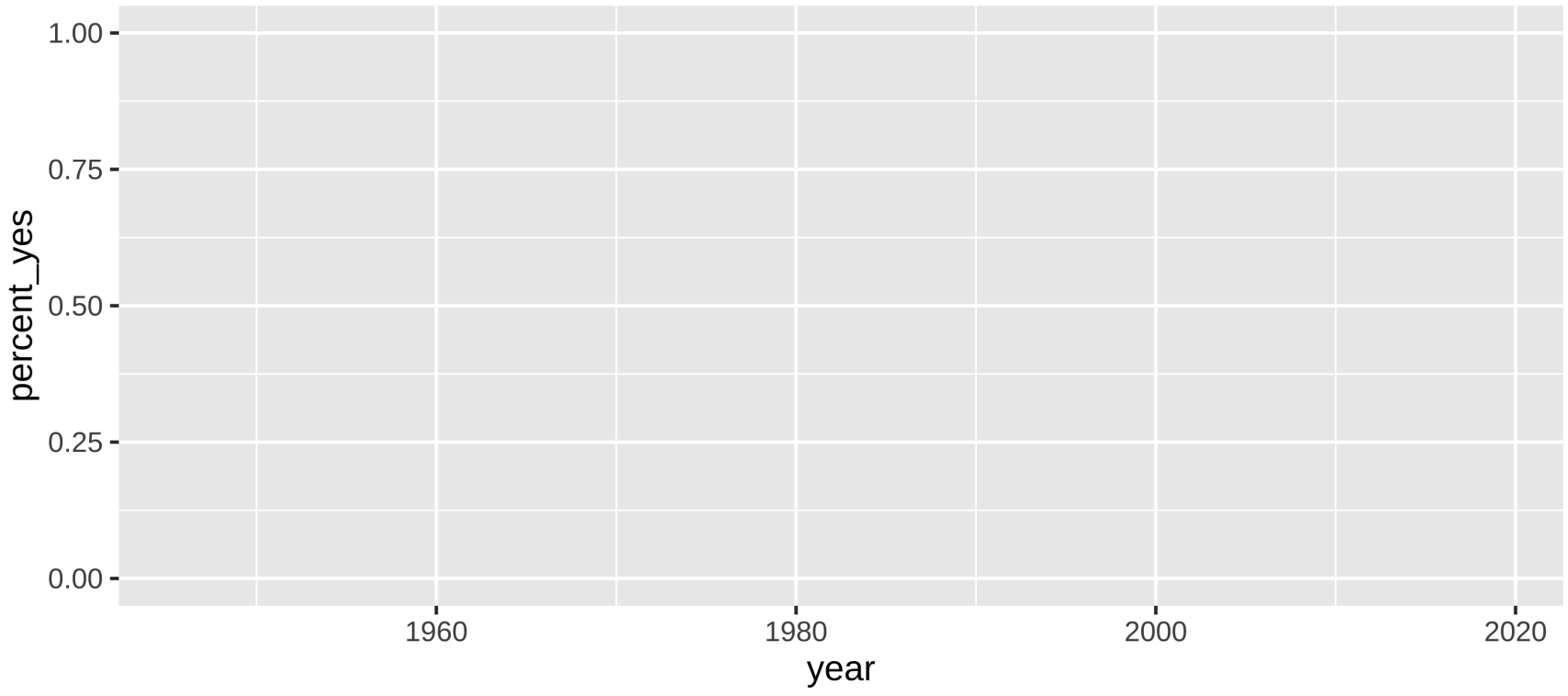
```
un_uk_us_tr
```

```
## # A tibble: 1,212 × 4
## # Groups:   country, year [219]
##   country  year issue                percent_yes
##   <chr>   <dbl> <fct>                      <dbl>
## 1 Turkey   1946 Colonialism               0.8
## 2 Turkey   1946 Economic development      0.6
## 3 Turkey   1946 Human rights              0
## 4 Turkey   1947 Colonialism               0.222
## 5 Turkey   1947 Economic development      0.5
## 6 Turkey   1947 Palestinian conflict      0.143
## # … with 1,206 more rows
```

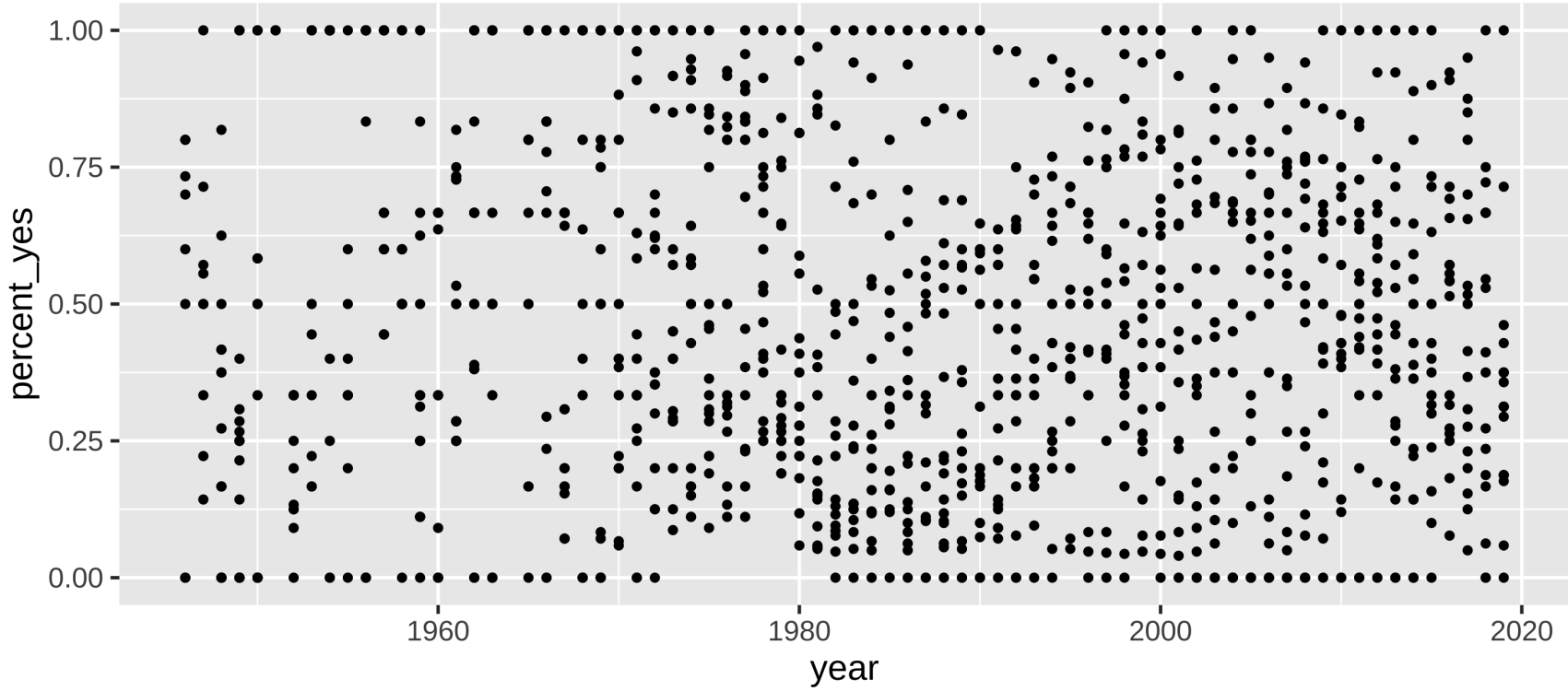
Erik Voeten "Data and Analyses of Voting in the UN General Assembly" Routledge Handbook of International Organization, edited by Bob Reinalda (published May 27, 2013)

```
ggplot(un_uk_us_tr)
```

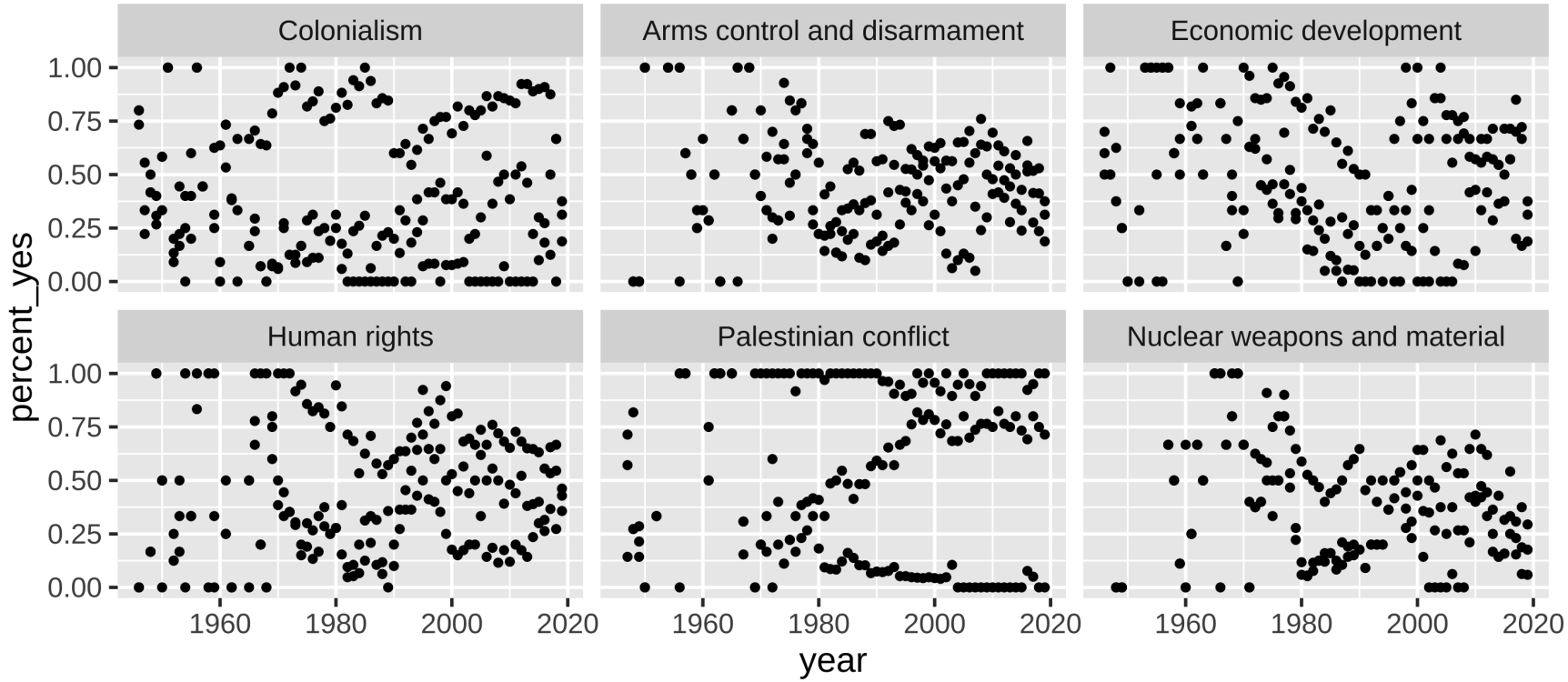
```
ggplot(un_uk_us_tr,
       aes(x = year, y = percent_yes))
```



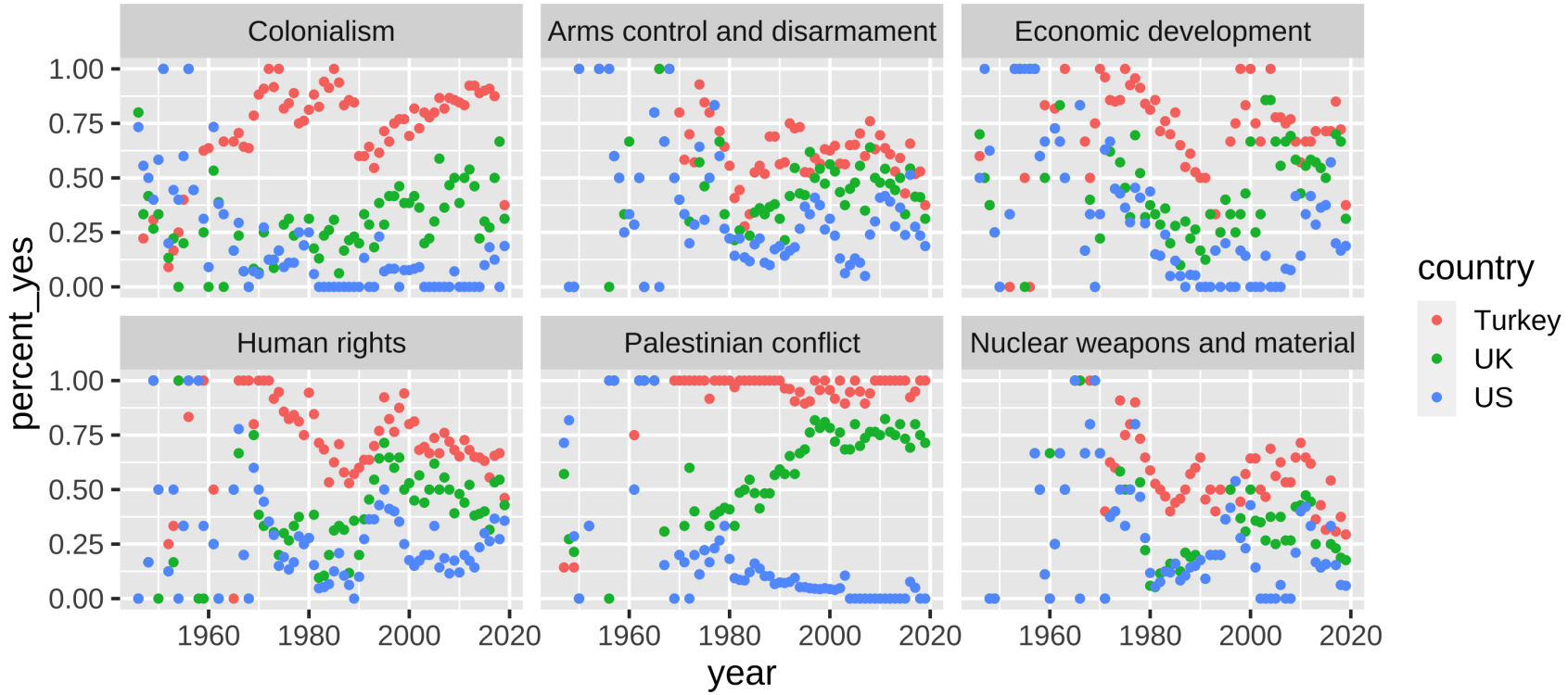
```
ggplot(un_uk_us_tr,
       aes(x = year, y = percent_yes)) +
  geom_point()
```



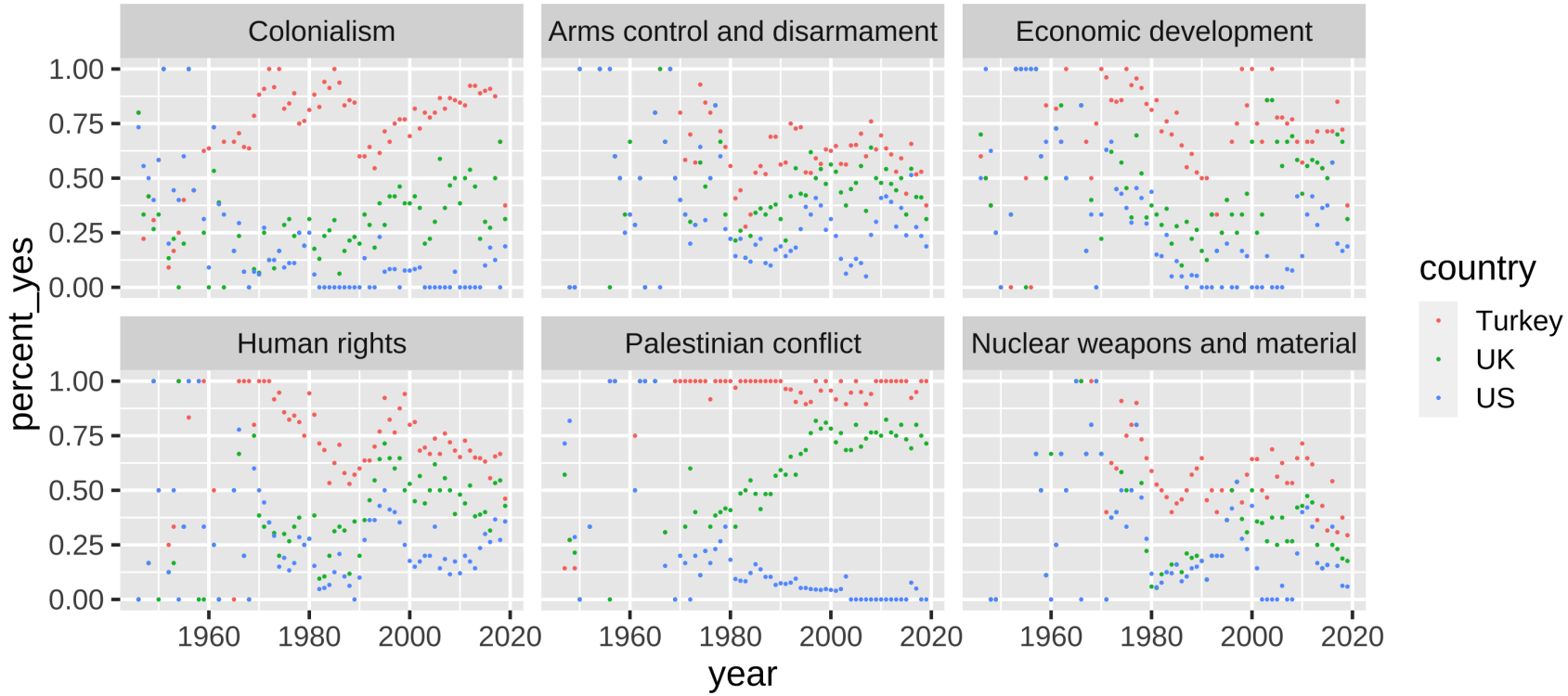
```
ggplot(un_uk_us_tr,
       aes(x = year, y = percent_yes)) +
  geom_point() +
  facet_wrap(~issue)
```



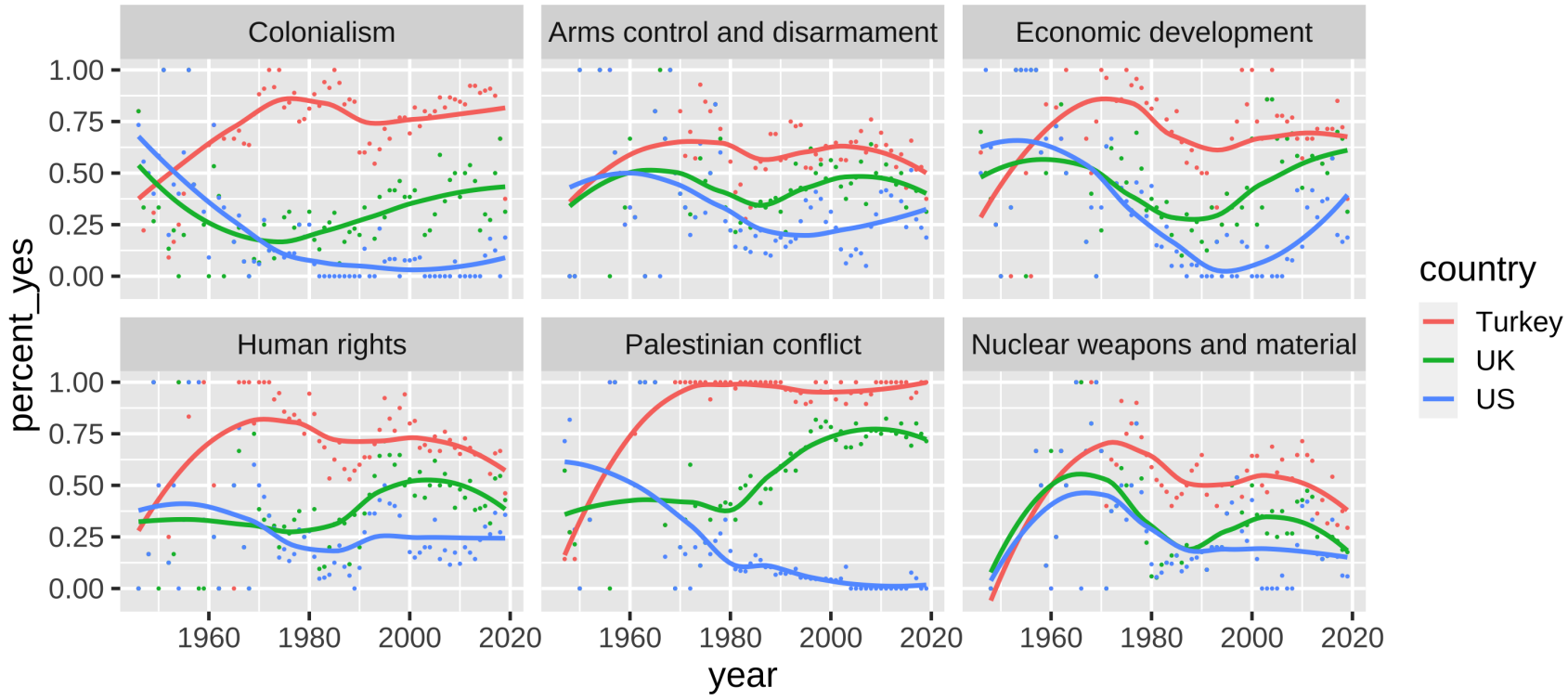
```
ggplot(un_uk_us_tr,
       aes(x = year, y = percent_yes, color = country)) +
  geom_point() +
  facet_wrap(~issue)
```



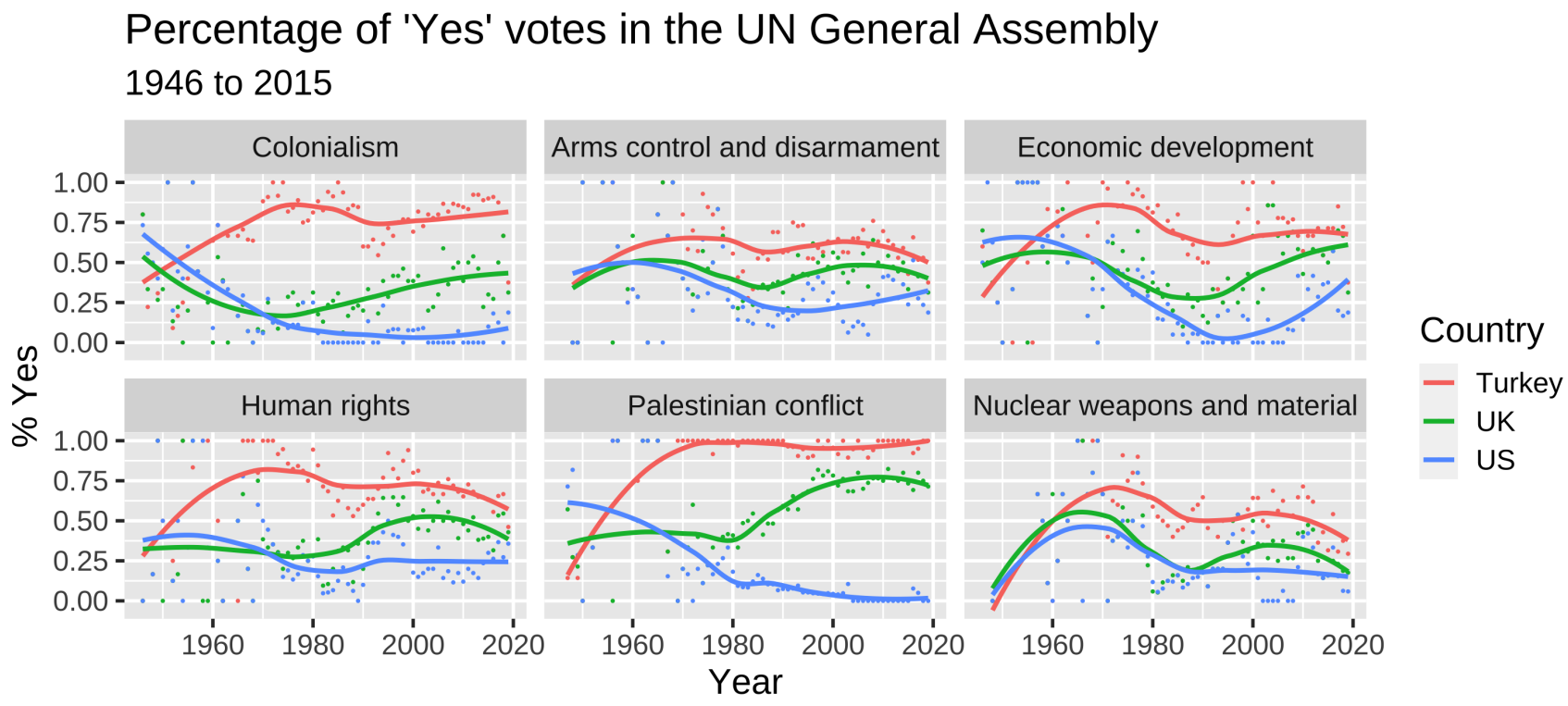
```
ggplot(un_uk_us_tr,
       aes(x = year, y = percent_yes, color = country)) +
  geom_point(size = .25) +
  facet_wrap(~issue)
```



```
ggplot(un_uk_us_tr,
       aes(x = year, y = percent_yes, color = country)) +
  geom_point(size = .25) +
  geom_smooth(method = "loess", se = FALSE) +
  facet_wrap(~issue)
```



```
ggplot(un_uk_us_tr,
       aes(x = year, y = percent_yes, color = country)) +
  geom_point(size = .25) +
  geom_smooth(method = "loess", se = FALSE) +
  facet_wrap(~issue) +
  labs(
    title = "Percentage of 'Yes' votes in the UN General Assembly",
    subtitle = "1946 to 2015",
    y = "% Yes", x = "Year", color = "Country"
  )
```



```r
ggplot(un_uk_us_tr,
       aes(x = year, y = percent_yes, color = country)) +
  geom_point(size = .25) +
  geom_smooth(method = "loess", se = FALSE) +
  facet_wrap(~issue) +
  labs(
    title = "Percentage of 'Yes' votes in the UN General Assembly",
    subtitle = "1946 to 2015",
    y = "% Yes", x = "Year", color = "Country"
  ) +
  scale_y_continuous(labels = label_percent())
```