SSPs Human Capitals - National

James Millington

Libraries

```
library(tidyr)
library(dplyr)
```

```

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
library(readxl)
library(ggplot2)
library(ggrepel)
library(ggiraph)
```

Load Data

```
load_ssp <- function(filepath,varlab){
  #for each SSP data sheet
  for(n in 1:5){
    ssp <- paste0("ssp", n)
```

```r
    sheet <- read_excel(filepath, sheet = ssp, na='<Null>')

    #pivot longer, add ssp id column
    lsheet <- sheet %>%
      pivot_longer(
        cols = !OID:unit,
        names_to='year',
        values_to='value'
      ) %>%
      mutate(ssp=n)

    #calculate ranks of values
    lranks <- lsheet %>%
      group_by(year) %>%
      mutate(unit='rank',
             #value=min_rank(desc(value)))   #high rank number is low value
             value=min_rank(value))   #high rank number is high value

    #combine into single dataframe
    ldat <- rbind(lsheet,lranks)
    if(n > 1){
      alldat <- rbind(alldat,ldat)
    } else { alldat <- ldat}
  }
  alldat <- mutate(alldat,variable=varlab)
  return(alldat)
}
```

```r
edu <- load_ssp("data/edu.xlsx", "edu")
ma <- load_ssp("data/ma.xlsx", "ma")
gdp <- load_ssp("data/gdp.xlsx", "gdp")
health <- load_ssp("data/health.xlsx", "health")
gini <- load_ssp("data/gini.xlsx", "gini")
wap <- load_ssp("data/wap.xlsx", "wap")
tec <- load_ssp("data/tec.xlsx", "tec")

countries <-edu %>%
  select(OID, GID_0) %>%
  distinct()
```

Calculations

```
#root mean square error (using median)
rmse_med <- function(x) {
  med = median(x, na.rm=TRUE)
  return(sum(sqrt((x-med)^2)))
}

#range function
rangesr <- function(x) {
  max = max(x, na.rm=TRUE)
  min = min(x, na.rm=TRUE)
  return(diff(c(min,max)))
}

out_all = countries
for(nm in list(edu,ma, gdp, health, gini, wap, tec)){
  rmsemed_nm = paste0(nm$variable[1],"_rmsemed")
  medmed_nm = paste0(nm$variable[1],"_medmed")
  summed_nm = paste0(nm$variable[1],"_summed")
  sumrng_nm = paste0(nm$variable[1],"_sumrng")
  summax_nm = paste0(nm$variable[1],"_summax")
  summin_nm = paste0(nm$variable[1],"_summed")

  out_all <- nm %>%
  #edu %>%
    filter(unit=='rank') %>%
    group_by(GID_0, ssp) %>%
    #calc median rank across years, for each ssp, for each country
    summarise(medrank = median(value, na.rm=TRUE),
              maxrank = max(value, na.rm=TRUE),
              minrank = min(value, na.rm=TRUE),
              rangerank = rangesr(value)
              ) %>%
    group_by(GID_0) %>%
    summarise(!!rmsemed_nm := rmse_med(medrank), #calc rmse (median) for ssp medians
              !!medmed_nm := median(medrank),    #calc median of ssp medians
              !!summed_nm := sum(medrank),       #calc sum of ssp medians
              !!sumrng_nm := sum(rangerank),     #calc sum of ssp ranges
              !!summax_nm := sum(maxrank),       #calc sum of ssp maxs
              !!summin_nm := sum(minrank),       #calc sum of ssp mins
              ) %>%
```

```r
    left_join(out_all, ., by='GID_0')
}
```

```
Warning: There were 440 warnings in `summarise()`.
The first warning was:
i In argument: `maxrank = max(value, na.rm = TRUE)`.
i In group 16: `GID_0 = "AND"`, `ssp = 1`.
Caused by warning in `max()`:
! no non-missing arguments to max; returning -Inf
i Run `dplyr::last_dplyr_warnings()` to see the 439 remaining warnings.
```

```
`summarise()` has grouped output by 'GID_0'. You can override using the
`.groups` argument.
```

```
Warning: There were 520 warnings in `summarise()`.
The first warning was:
i In argument: `maxrank = max(value, na.rm = TRUE)`.
i In group 16: `GID_0 = "AND"`, `ssp = 1`.
Caused by warning in `max()`:
! no non-missing arguments to max; returning -Inf
i Run `dplyr::last_dplyr_warnings()` to see the 519 remaining warnings.
```

```
`summarise()` has grouped output by 'GID_0'. You can override using the
`.groups` argument.
`summarise()` has grouped output by 'GID_0'. You can override using the
`.groups` argument.
```

```
Warning: There were 440 warnings in `summarise()`.
The first warning was:
i In argument: `maxrank = max(value, na.rm = TRUE)`.
i In group 16: `GID_0 = "AND"`, `ssp = 1`.
Caused by warning in `max()`:
! no non-missing arguments to max; returning -Inf
i Run `dplyr::last_dplyr_warnings()` to see the 439 remaining warnings.
```

```
`summarise()` has grouped output by 'GID_0'. You can override using the
`.groups` argument.
`summarise()` has grouped output by 'GID_0'. You can override using the
`.groups` argument.
```

```
`summarise()` has grouped output by 'GID_0'. You can override using the
`.groups` argument.
`summarise()` has grouped output by 'GID_0'. You can override using the
`.groups` argument.
```

```
out_2100 = countries
for(nm in list(edu,ma, gdp, health, gini, wap, tec)){
  rmsemed_nm = paste0(nm$variable[1],"_rmsemed")  #calc rmse (median) for ssp medians
  med_nm = paste0(nm$variable[1],"_med")
  rng_nm = paste0(nm$variable[1],"_rng")
  max_nm = paste0(nm$variable[1],"_max")
  min_nm = paste0(nm$variable[1],"_min")
  sum_nm = paste0(nm$variable[1],"_sum")

  out_2100 <- nm %>%
  #edu %>%
    filter(unit=='rank', year==2100)  %>%
    group_by(GID_0) %>%
    summarise(!!rmsemed_nm := rmse_med(value),       ##calc rmse (median) for ssp ranks
              !!med_nm := median(value, na.rm=TRUE), #calc median of ssp 2100 ranks
              !!sum_nm := sum(value),                #calc sum of ssp 2100 ranks
              !!rng_nm := rangesr(value),            #calc range of ssp 2100 ranks
              !!max_nm := max(value, na.rm=TRUE),    #calc sum of ssp 2100 rank maxs
              !!min_nm := min(value, na.rm=TRUE),    #calc sum of ssp 2100 rank mins
              ) %>%

    left_join(out_2100, ., by='GID_0')
}
```

```
Warning: There were 88 warnings in `summarise()`.
The first warning was:
i In argument: `edu_rng = rangesr(value)`.
i In group 4: `GID_0 = "AND"`.
Caused by warning in `max()`:
! no non-missing arguments to max; returning -Inf
i Run `dplyr::last_dplyr_warnings()` to see the 87 remaining warnings.

Warning: There were 104 warnings in `summarise()`.
The first warning was:
i In argument: `ma_rng = rangesr(value)`.
i In group 4: `GID_0 = "AND"`.
```

```
Caused by warning in `max()`:
! no non-missing arguments to max; returning -Inf
i Run `dplyr::last_dplyr_warnings()` to see the 103 remaining warnings.
```

```
Warning: There were 88 warnings in `summarise()`.
The first warning was:
i In argument: `health_rng = rangesr(value)`.
i In group 4: `GID_0 = "AND"`.
Caused by warning in `max()`:
! no non-missing arguments to max; returning -Inf
i Run `dplyr::last_dplyr_warnings()` to see the 87 remaining warnings.
```

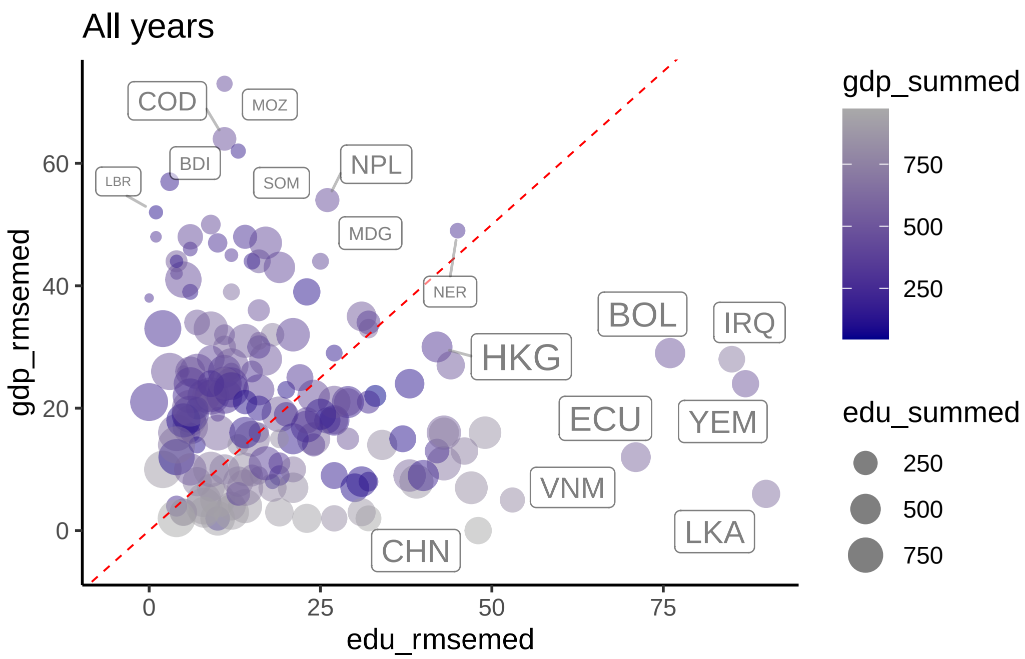
Plots

All years

Education

```
p <- out_all %>%
  drop_na() %>%
  ggplot(aes(x=edu_rmsemed, y=gdp_rmsemed, size=edu_summed, colour=gdp_summed)) +
  geom_point_interactive(alpha=0.5) +
  #geom_point_interactive(aes(tooltip = GID_0, data_id=GID_0),alpha=0.5) +
  ggtitle("All years") +
  scale_colour_gradient(low="darkblue",high="darkgrey")+
  xlim(-5, NA) +
  ylim(-5, NA) +
  geom_abline(intercept = 0, slope = 1,
              linewidth = 0.35,colour='red', linetype='dashed') +
  geom_label_repel(aes(label = GID_0),
                  alpha=0.5,
                  max.overlaps=15,
                  box.padding   = 0.35,
                  point.padding = 0.5,
                  segment.color = 'grey50',
                  show.legend = FALSE,
                  color='black') +
  theme_classic()

p
```

```
Warning: ggrepel: 154 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



```
#girafe(ggobj = p)
```

More grey is higher overall GDP ranking (e.g. USA is in bottom left)

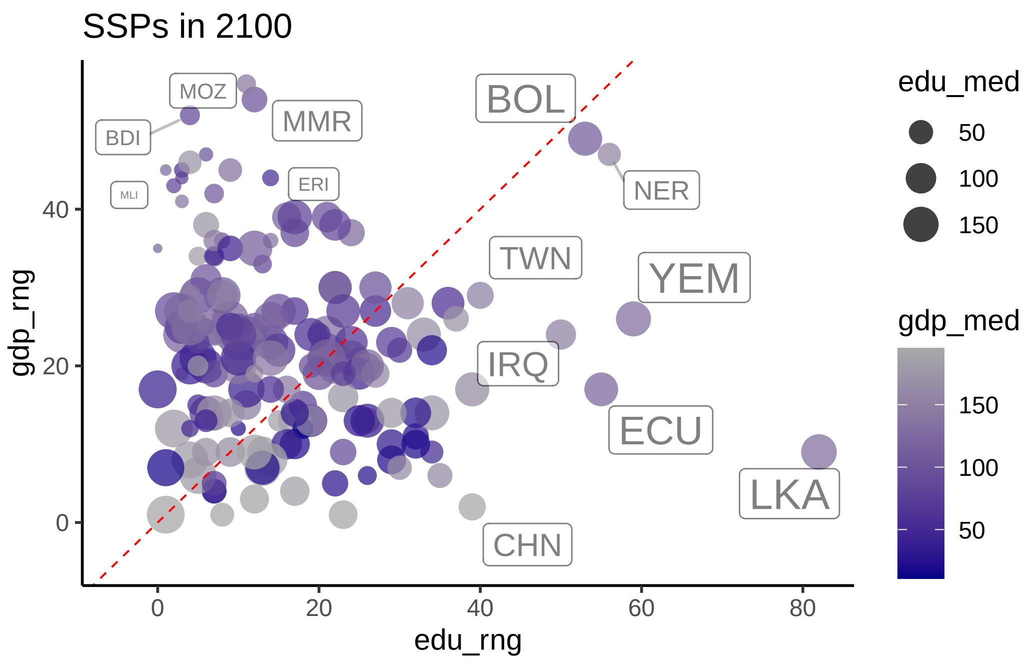
2100

```
#setup template https://stackoverflow.com/a/16727357

p2100 <-
  list(
    geom_point(alpha=0.75),
    ggtitle("SSPs in 2100"),
    scale_colour_gradient(low="darkblue",high="darkgrey"),
    xlim(-5, NA),
    ylim(-5, NA),
    geom_abline(intercept = 0, slope = 1,
                linewidth = 0.35,colour='red', linetype='dashed'),
    geom_label_repel(aes(label = GID_0),
                    alpha=0.5,
                    max.overlaps=15,
                    box.padding   = 0.35,
                    point.padding = 0.5,
                    segment.color = 'grey50',
                    show.legend = FALSE,
                    color='black'),
    theme_classic()
  )
```

Education

```
out_2100 %>%
  drop_na() %>%
  ggplot(aes(x=edu_rng, y=gdp_rng, size=edu_med, colour=gdp_med)) +
  p2100
```

```
Warning: ggrepel: 157 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



More grey is higher overall GDP ranking (e.g. USA is in bottom left)

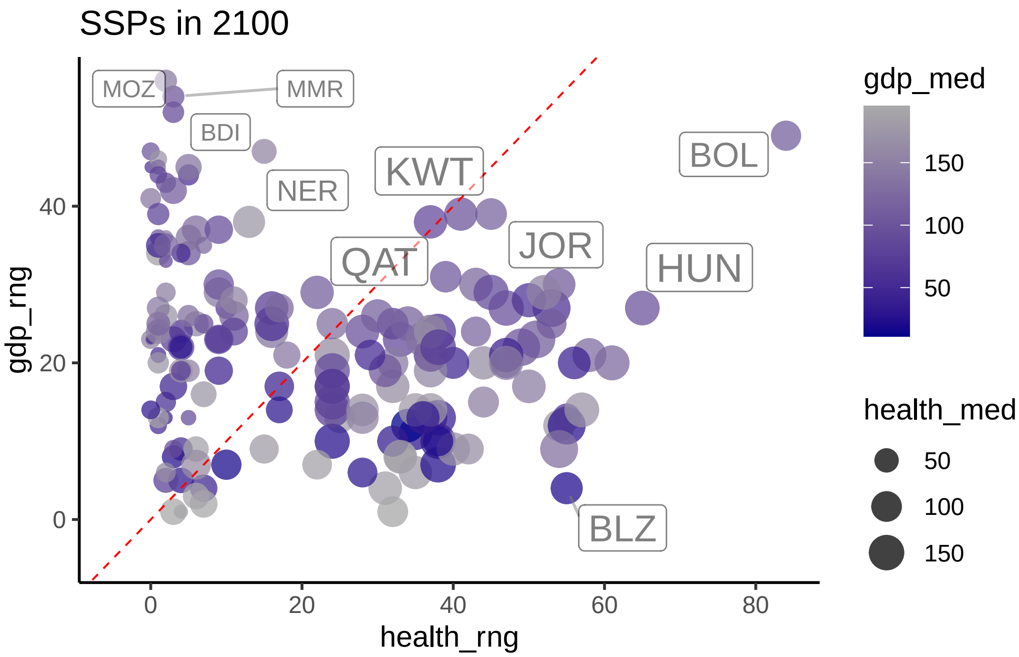
Interpretation

- China has low GDP range (always ranked well in 2100), but relatively variable Educational rank
- USA always ranked well on both indicators
- Sri Lanka (LKA) has a relatively consistent GDP ranking (quite high), but highly variable education ranking
- Mozambique, Burundi, Myanmar have variable GDP ranking (intermediate), but relatively consistent (poor) education ranking

Health

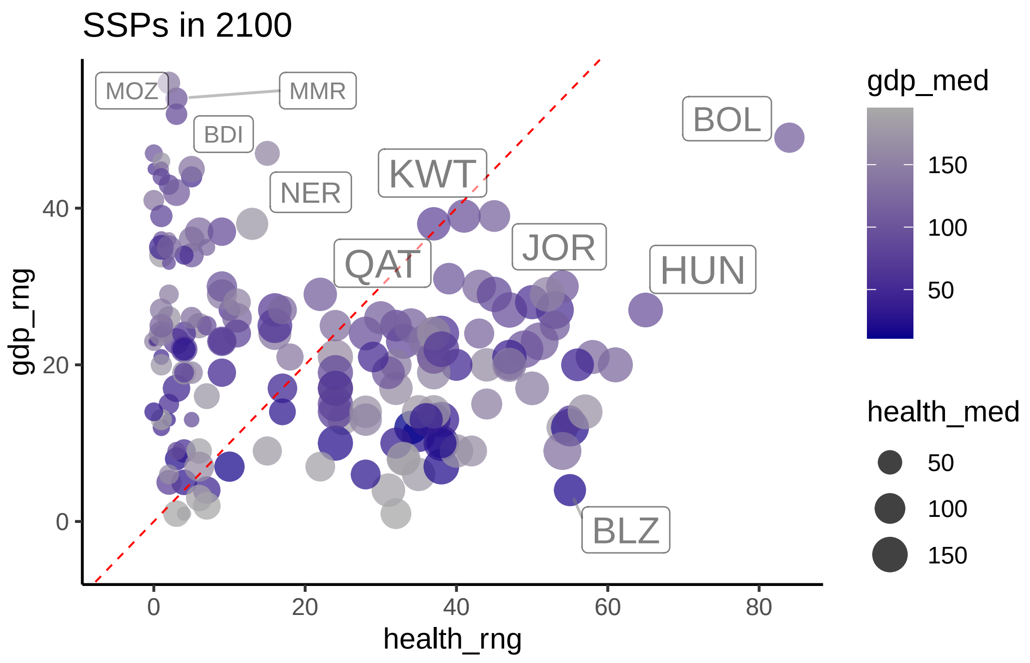
```
out_2100 %>%
  drop_na() %>%
  ggplot(aes(x=health_rng, y=gdp_rng, size=health_med, colour=gdp_med)) +
  p2100
```

```
Warning: ggrepel: 160 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



```
out_2100 %>%
  drop_na() %>%
  ggplot(aes(x=health_rng, y=gdp_rng, size=health_med, colour=gdp_med)) +
  p2100
```

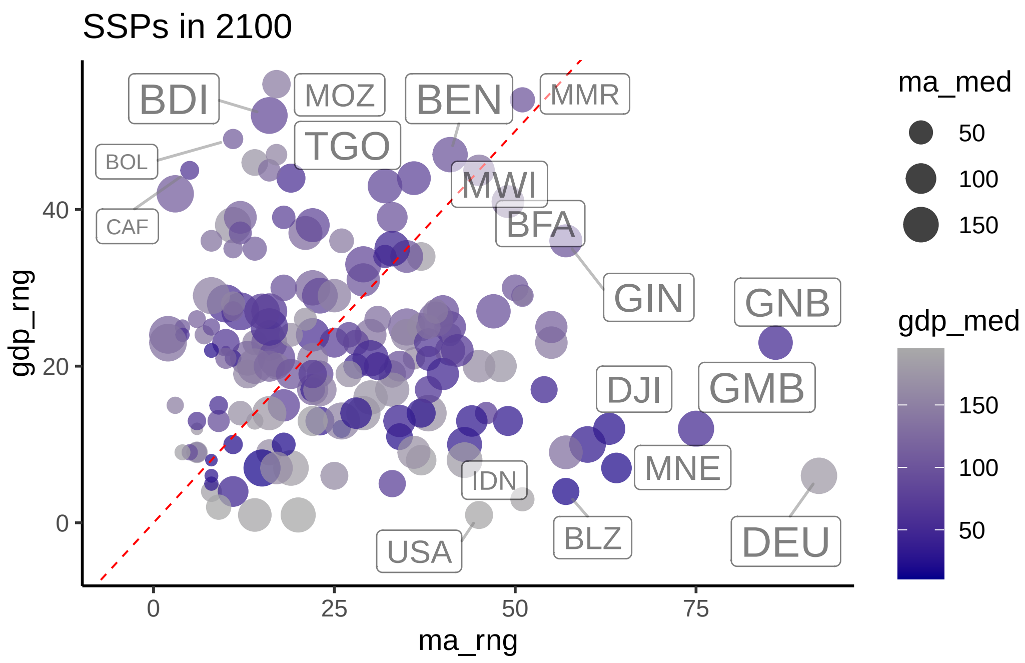
```
Warning: ggrepel: 160 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



Market Access

```
out_2100 %>%
  drop_na() %>%
  ggplot(aes(x=ma_rng, y=gdp_rng, size=ma_med, colour=gdp_med)) +
  p2100
```

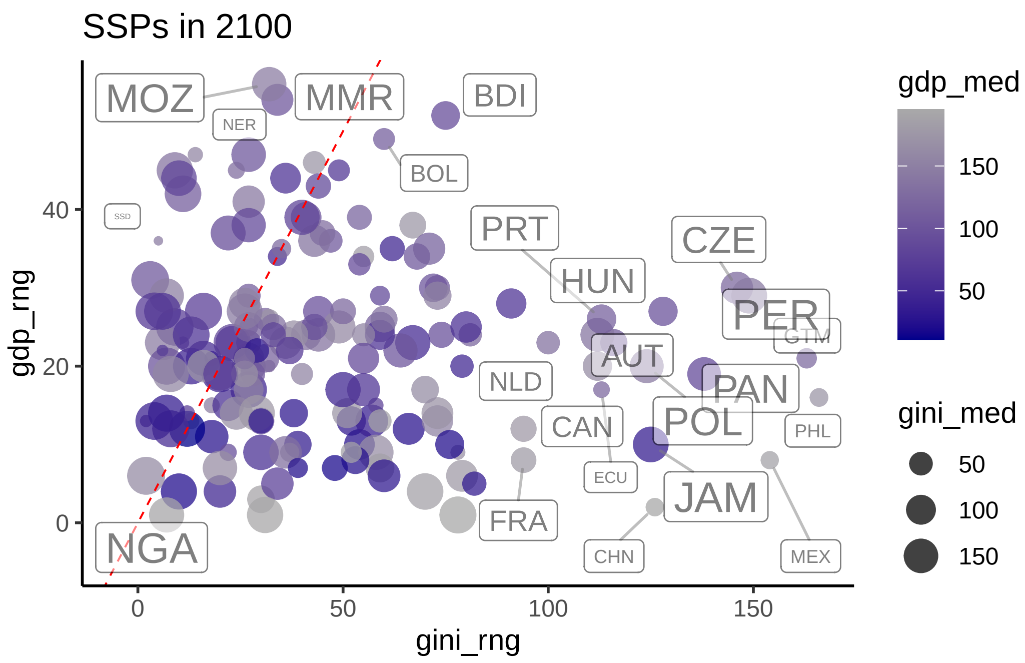
```
Warning: ggrepel: 152 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



Gini

```
out_2100 %>%
  drop_na() %>%
  ggplot(aes(x=gini_rng, y=gdp_rng, size=gini_med, colour=gdp_med)) +
  p2100
```

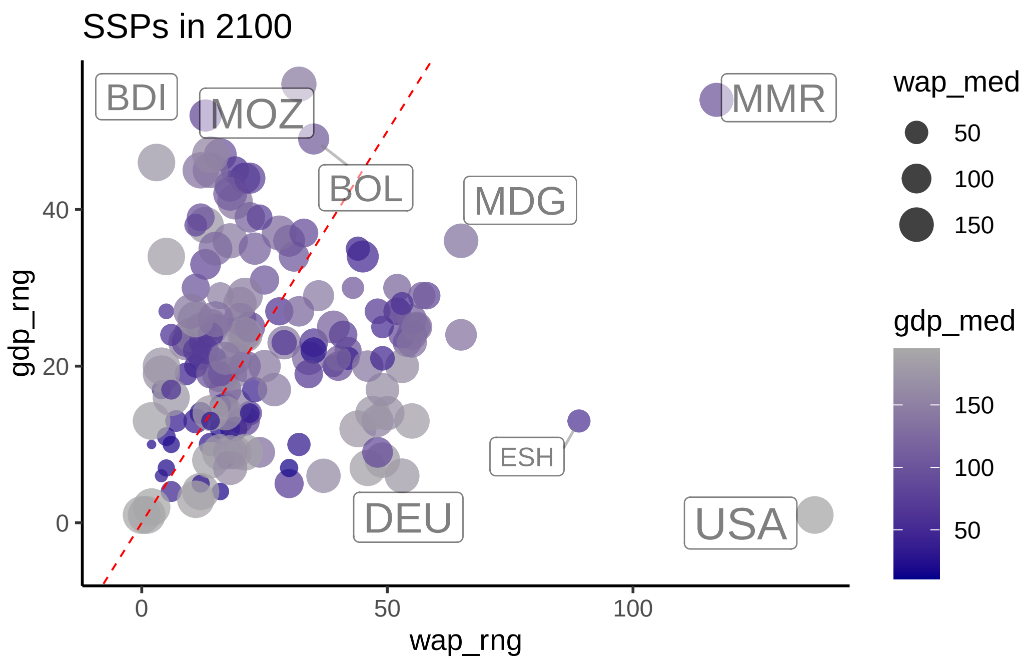
```
Warning: ggrepel: 147 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



Working Age Population

```
out_2100 %>%
  drop_na() %>%
  ggplot(aes(x=wap_rng, y=gdp_rng, size=wap_med, colour=gdp_med)) +
  p2100
```

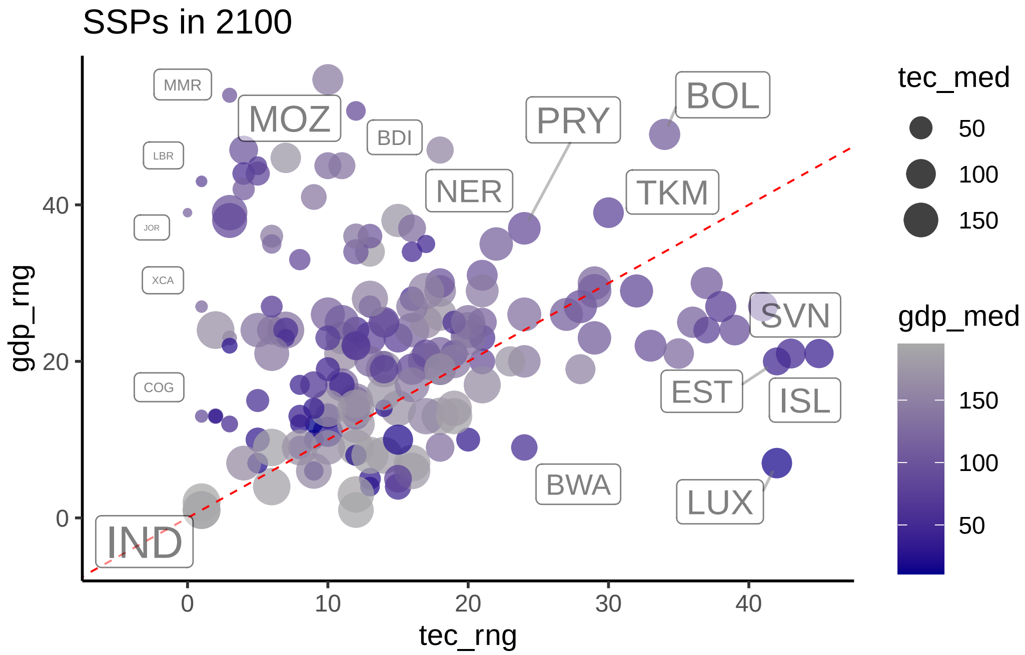
```
Warning: ggrepel: 162 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



Energy

```
out_2100 %>%
  drop_na() %>%
  ggplot(aes(x=tec_rng, y=gdp_rng, size=tec_med, colour=gdp_med)) +
  p2100
```

```
Warning: ggrepel: 153 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



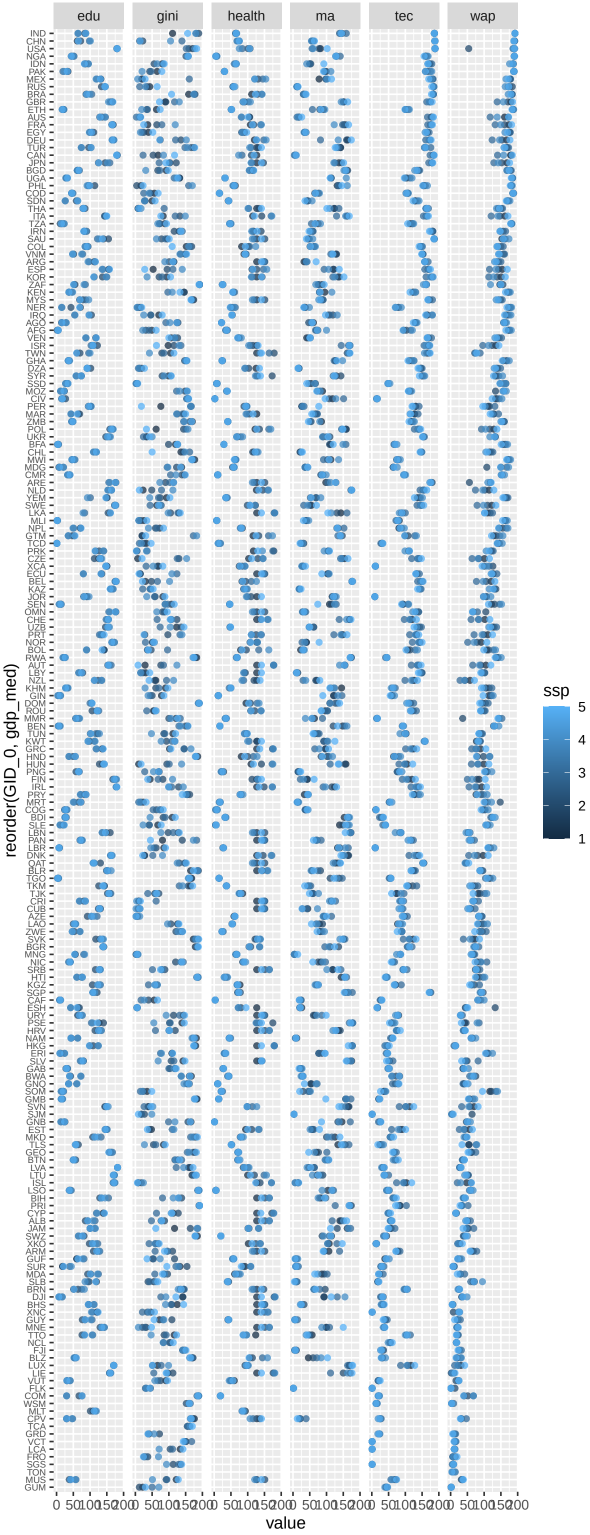
Experimental Another way to plot these (if GDP is always the comparator) might be to show countries ranked on y-axis by gdp (med rank) then points on xaxis for the other variable value (for each ssp), then facet on x for variables (but this then does not show countries with variable vs non-variable gdp)

However, the below shows that with 200+ countries it's hard to get this looking good. It does show a string relationship between GDP and WAP (and to some degree energy).

```
#create data for 2100 to trial gdp ranked plot

ranks_2100 = edu
counter = 1
for(nm in list(edu,ma, health, gini, wap, tec)){

  if(counter == 1){
    ranks_2100 <- filter(nm, unit=='rank', year==2100)
  } else {
    ranks_2100 <- bind_rows(ranks_2100, filter(nm, unit=='rank', year==2100))
  }
  counter = counter + 1
}
```

```
#create gdp rank data
gdp_2100 <-
  out_2100 %>%
  select(OID, GID_0, gdp_med)

#join ranks for our metrics to gdp
gdp_ranks_2100 <-
  left_join(gdp_2100, ranks_2100, by='GID_0',suffix=c("",".y")) %>%
  select(-ends_with(".y"))
```

```
#plot for EDU
gdp_ranks_2100 %>%
  drop_na() %>%
  arrange(gdp_med) %>%  #order countries by median SSP GDP
  #filter(variable=='edu') %>%
  ggplot(aes(x=value, y=reorder(GID_0,gdp_med), colour=ssp)) +
  geom_point(alpha=0.75) +
  theme(axis.text.y = element_text(size=rel(0.65))) +
  facet_grid(.~variable)
```


edu
gini
health
ma
tec
wap
IND
CHN
USA
NGA
IDN
PAK
MEX
RUS
BRA
GBR
ETH
AUS
FRA
EGY
DEU
TUR
CAN
JPN
BGD
UGA
PHL
COD
SDN
THA
ITA
TZA
IRN
SAU
COL
VNM
ARG
ESP
KOR
ZAF
KEN
MYS
NER
IRQ
AGO
AFG
VEN
ISR
TWN
GHA
DZA
SYR
SSD
MOZ
CIV
PER
MAR
ZMB
POL
UKR
BFA
CHL
MWI
MDG
CMR
ARE
NLD
YEM
SWE
LKA
MLI
NPL
GTM
TCD
PRK
CZE
XCA
ECU
BEL
KAZ
JOR
SEN
OMN
CHE
UZB
PRT
NOR
BOL
RWA
AUT
LBY
NZL
KHM
GIN
DOM
ROU
MMR
BEN
TUN
KWT
GRC
HND
HUN
PNG
FIN
IRL
PRY
MRT
COG
BDI
SLE
LBN
PAN
LBR
DNK
QAT
BLR
TGO
TKM
TJK
CRI
CUB
AZE
LAO
ZWE
SVK
BGR
MNG
NIC
SRB
HTI
KGZ
SGP
CAF
ESH
URY
PSE
HRV
NAM
HKG
ERI
SLV
GAB
BWA
GNQ
SOM
GMB
SVN
SJM
GNB
EST
MKD
TLS
GEO
BTN
LVA
LTU
ISL
LSO
BIH
PRI
CYP
ALB
JAM
SWZ
XKO
ARM
GUF
SUR
MDA
SLB
BRN
DJI
BHS
XNC
GUY
MNE
TTO
NCL
FJI
BLZ
LUX
LIE
VUT
FLK
COM
WSM
MLT
CPV
TCA
GRD
VCT
LCA
FRO
SGS
TON
MUS
GUM
reorder(GID_0, gdp_med)
0 50100150200
0 50100150200
0 50100150200
0 50100150200
0 50100150200
0 50100150200
value
ssp
5
4
3
2
1