analysis_v

Hunter York

11/18/2020

Compositional Changes

Plot top 35 Occupations change in % share over time

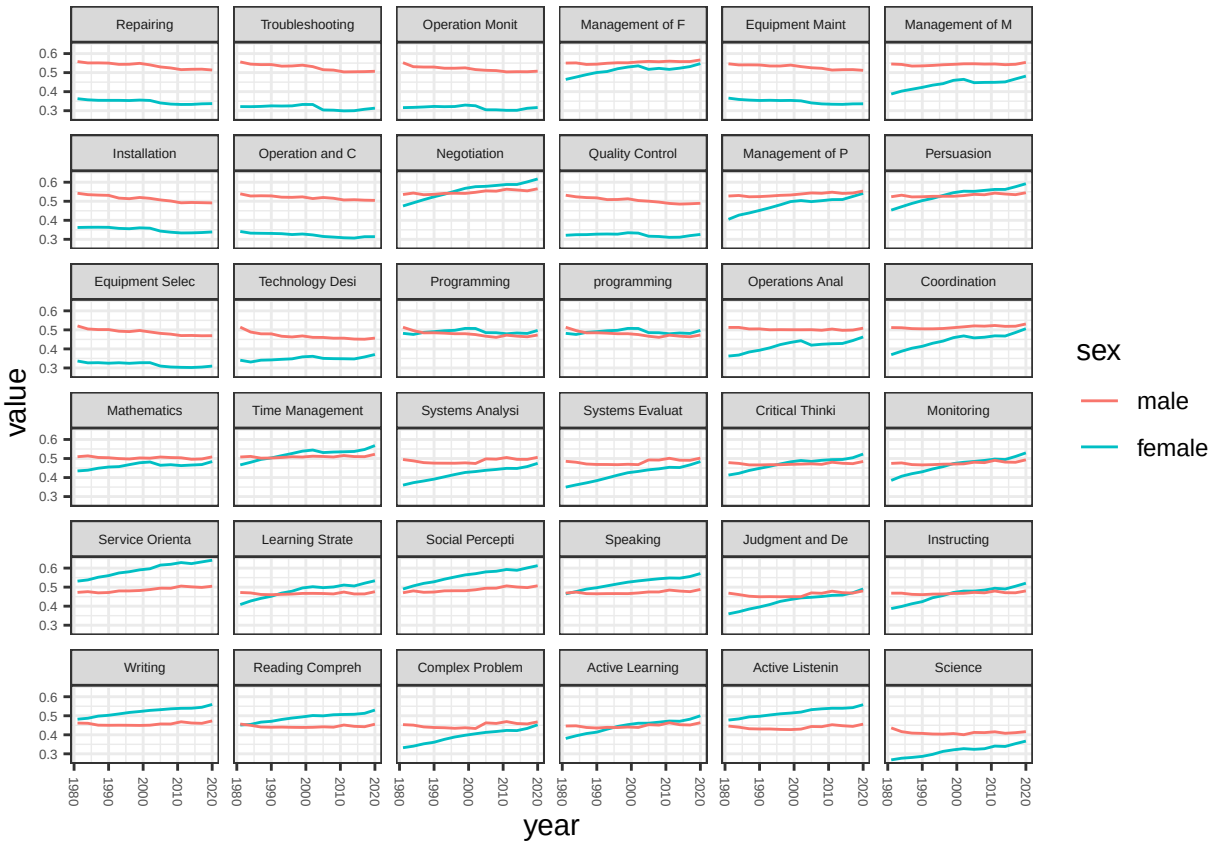
```
occ_share <- acs[!is.na(`CPS Occupational Title`),.(N = .N), by = .(`CPS Occupational Title`, OCC2010, y
occ_share[,percent_share := N/sum(N), by = .(year, sex)]
occ_share[, year_rank := frankv(percent_share, order = -1L), by = .(year, sex)]
occ_share <- occ_share[`CPS Occupational Title` %in% occ_share[year %in% c(1981, 2020) & year_rank <= 20
occ_share[, `CPS Occupational Title` := paste0(str_sub(`CPS Occupational Title`,1,10), "...\n",
                                               str_sub(`CPS Occupational Title`,-10,-1))]
occ_share[, `CPS Occupational Title` := factor(`CPS Occupational Title`, levels = occ_share[year == 2020
ggplot(occ_share) +
  geom_line(aes(x = year, y = percent_share*100, group = paste(`CPS Occupational Title`, sex), color = s
  facet_wrap(~`CPS Occupational Title`)+
    theme(axis.text.x = element_text(angle = -90, vjust = 0.5, hjust=1))+
    scale_y_continuous(trans = "log10")+
  theme(strip.text = element_text(size = 5), axis.text = element_text(size = 5))
```

Repeat for younger ages (25-34)

```
occ_share <- acs[!is.na(`CPS Occupational Title`) & as.numeric(as.character(age)) <= 34,.(N = .N), by =
occ_share[,percent_share := N/sum(N), by = .(year, sex)]
occ_share[, year_rank := frankv(percent_share, order = -1L), by = .(year, sex)]
occ_share <- occ_share[`CPS Occupational Title` %in% occ_share[year %in% c(1981, 2020) & year_rank <= 2
occ_share[, `CPS Occupational Title` := paste0(str_sub(`CPS Occupational Title`,1,10), "...\n",
                                               str_sub(`CPS Occupational Title`,-10,-1))]
occ_share[, `CPS Occupational Title` := factor(`CPS Occupational Title`, levels = occ_share[year == 202
ggplot(occ_share) +
  geom_line(aes(x = year, y = percent_share*100, group = paste(`CPS Occupational Title`, sex), color = s
  facet_wrap(~`CPS Occupational Title`)+
    theme(axis.text.x = element_text(angle = -90, vjust = 0.5, hjust=1)) +
  scale_y_continuous(trans = "log10") +
  theme(strip.text = element_text(size = 5), axis.text = element_text(size = 5))
```

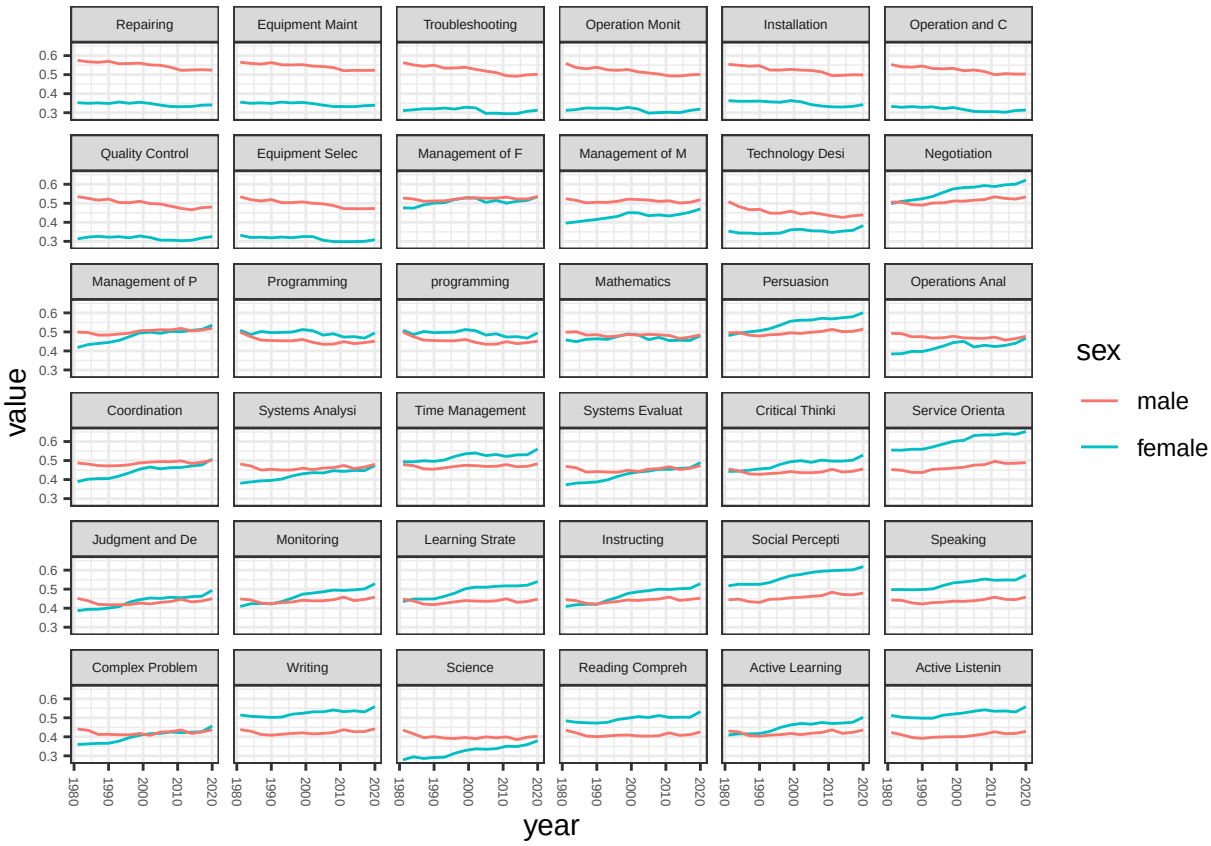
Plot average value for each skill over time, all ages

```
skill_share <- acs[, lapply(.SD, weighted.mean, w = asecwt, na.rm = T), .SDcols = vars, by = .(sex, yea
skill_share <- melt(skill_share, id.vars = c("year", "sex"))
skill_share[, variable := substr(variable, 1,15)]
skill_share[, variable := factor(variable, levels = skill_share[year == 1981 & sex == "male"] %>% .[ord
ggplot(skill_share[!variable %like% "pc|average|tech"]) +
  geom_line(aes(x = year, y = value, group = paste0(sex, variable), color = sex))+
facet_wrap(~variable) +
  theme(axis.text.x = element_text(angle = -90, vjust = 0.5, hjust=1)) +
  theme(strip.text = element_text(size = 5), axis.text = element_text(size = 5))
```



Repeat for younger ages (25-34)

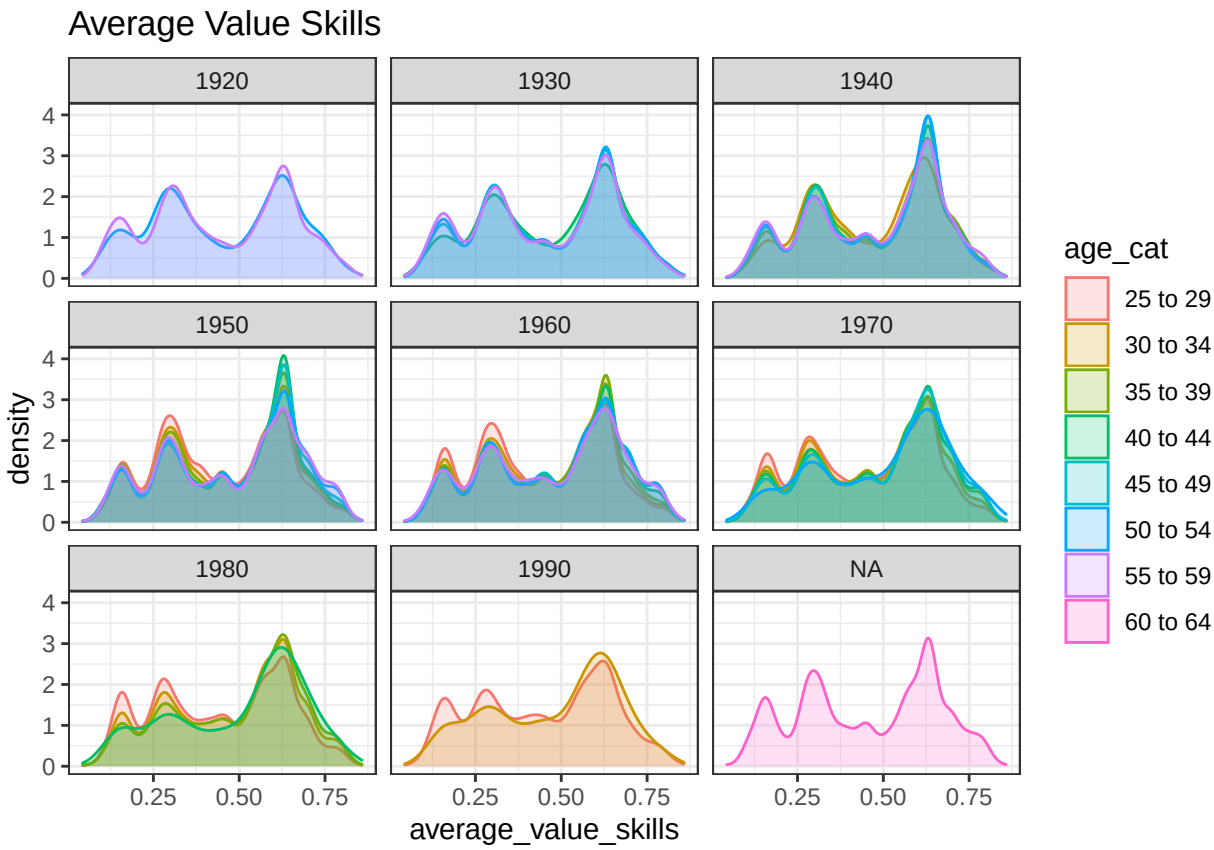
```
skill_share <- acs[as.numeric(as.character(age)) <= 34, lapply(.SD, weighted.mean, w = asecwt, na.rm = T
skill_share <- melt(skill_share, id.vars = c("year", "sex"))
skill_share[, variable := substr(variable, 1,15)]
skill_share[, variable := factor(variable, levels = skill_share[year == 1981 & sex == "male"] %>% .[ord
ggplot(skill_share[!variable %like% "pc|average|tech"]) +
  geom_line(aes(x = year, y = value, group = paste0(sex, variable), color = sex))+
facet_wrap(~variable) +
  theme(axis.text.x = element_text(angle = -90, vjust = 0.5, hjust=1)) +
  theme(strip.text = element_text(size = 5), axis.text = element_text(size = 5))
```



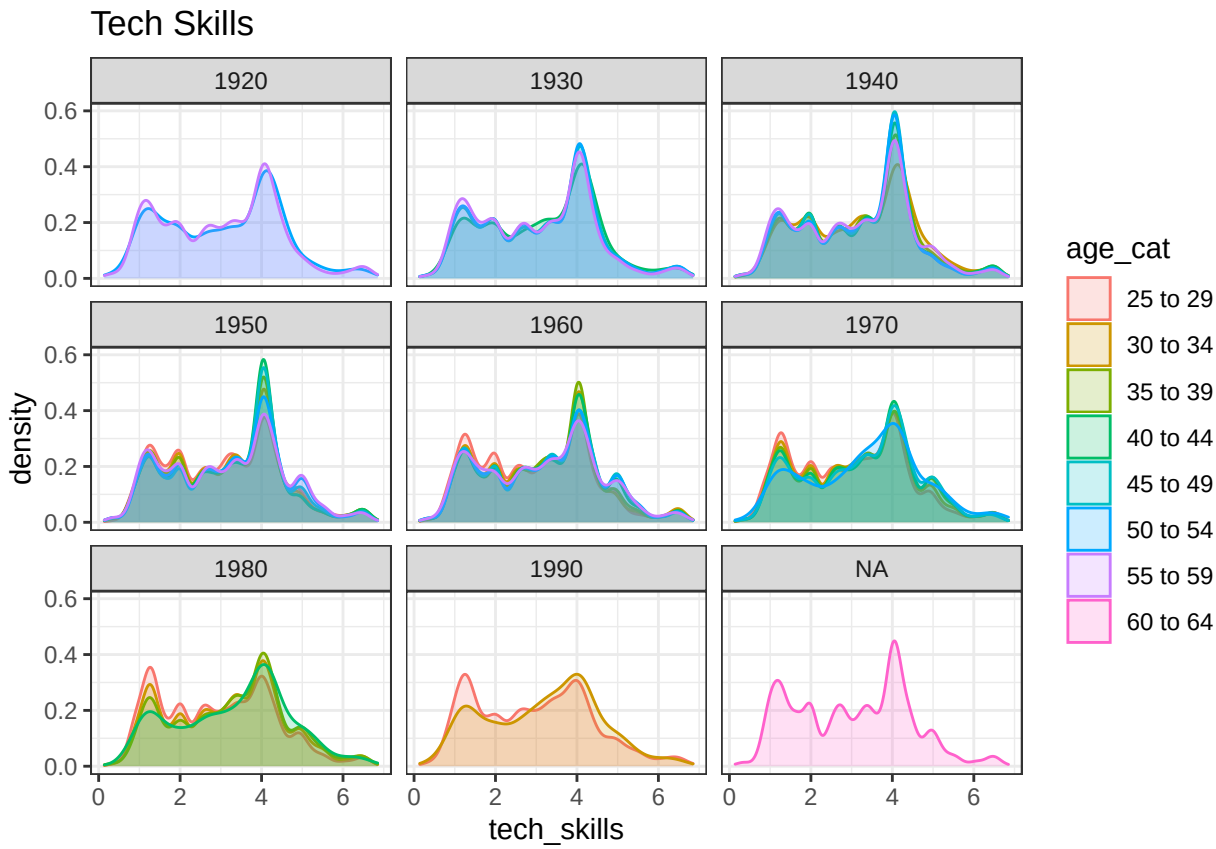
See if movement between occupations clusters by skill, at the cohort level

First, visualize movement between occupaitons

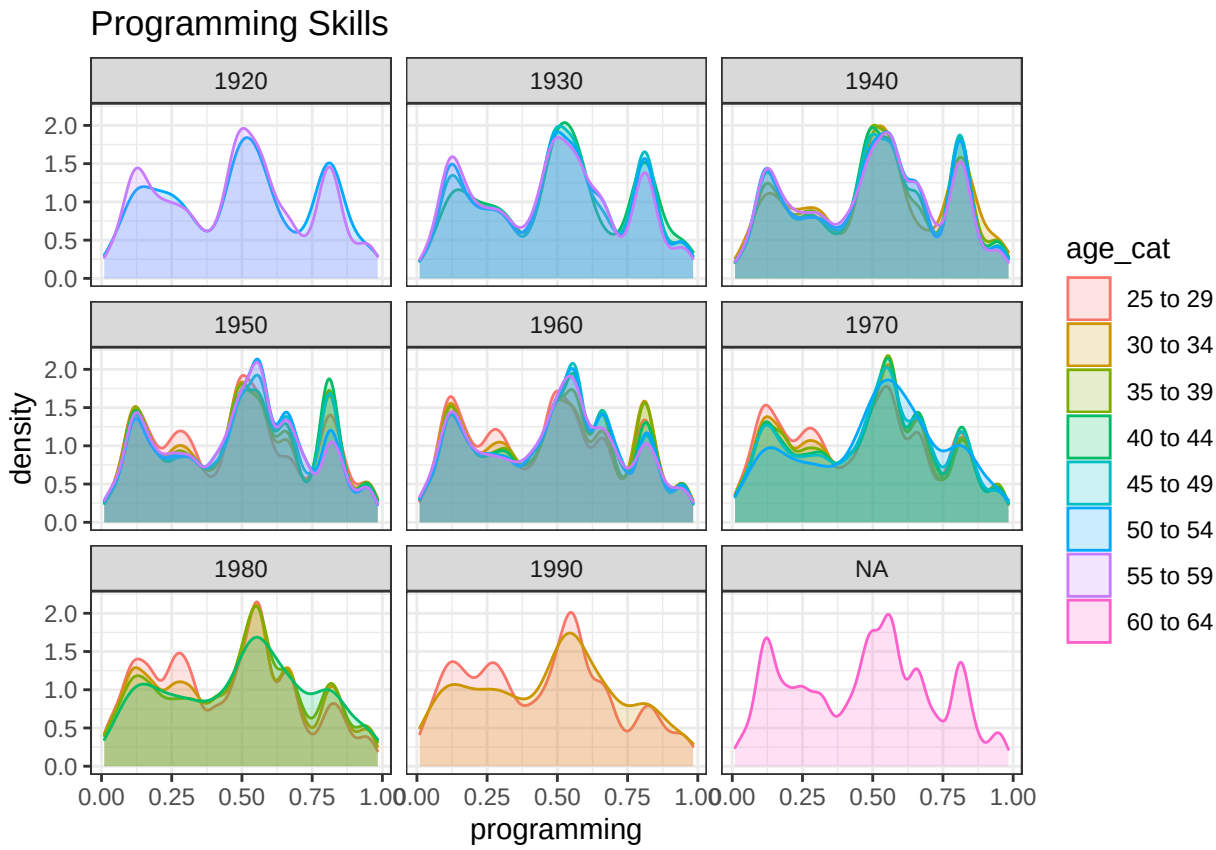
```
acs[as.numeric(as.character(age)) <= 59, cohort := floor((year - as.numeric(as.character(age)))/10)*10]

ggplot(acs) +
  geom_density(aes(x = average_value_skills, color = age_cat, fill = age_cat), alpha = .2) +
  facet_wrap(~cohort) +
  ggtitle("Average Value Skills")
```



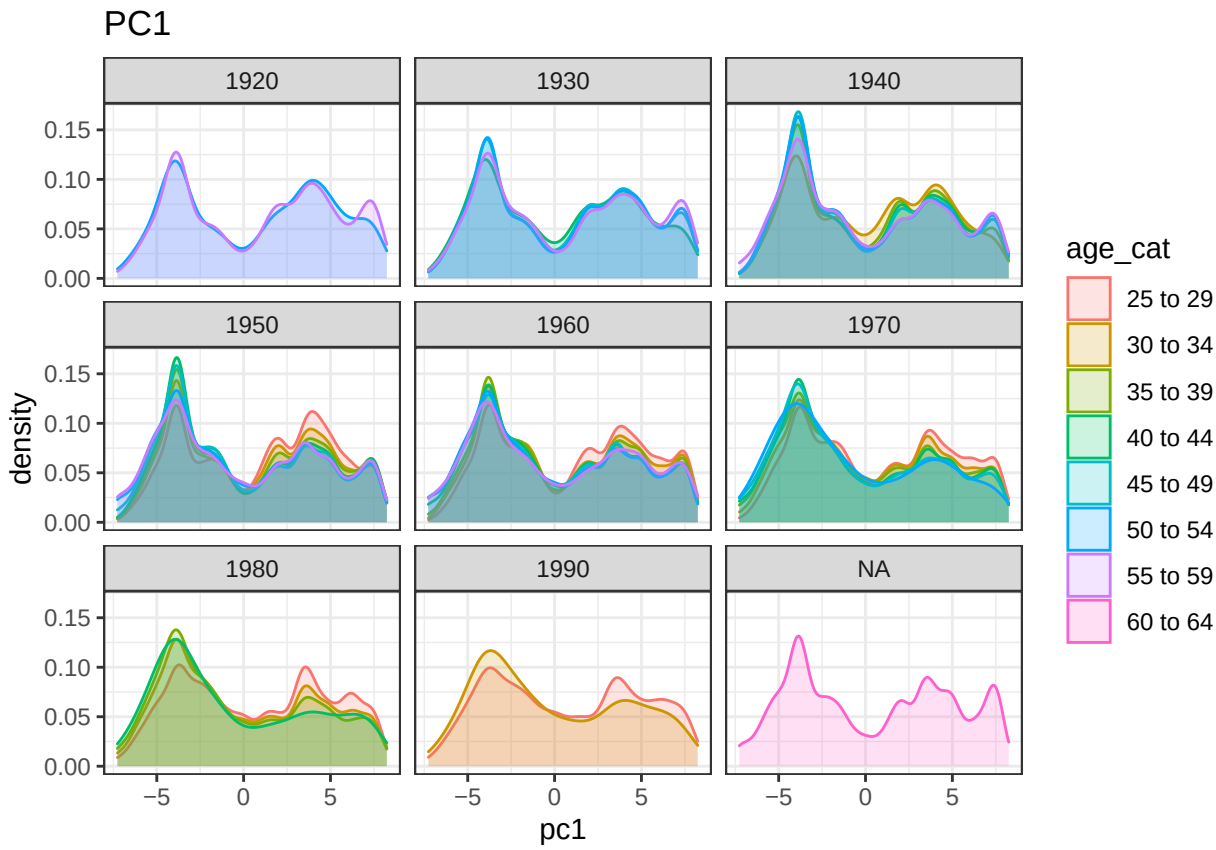

```
ggplot(acs) +
  geom_density(aes(x = tech_skills, color = age_cat, fill = age_cat), alpha = .2) +
  facet_wrap(~cohort) +
  ggtitle("Tech Skills")
```



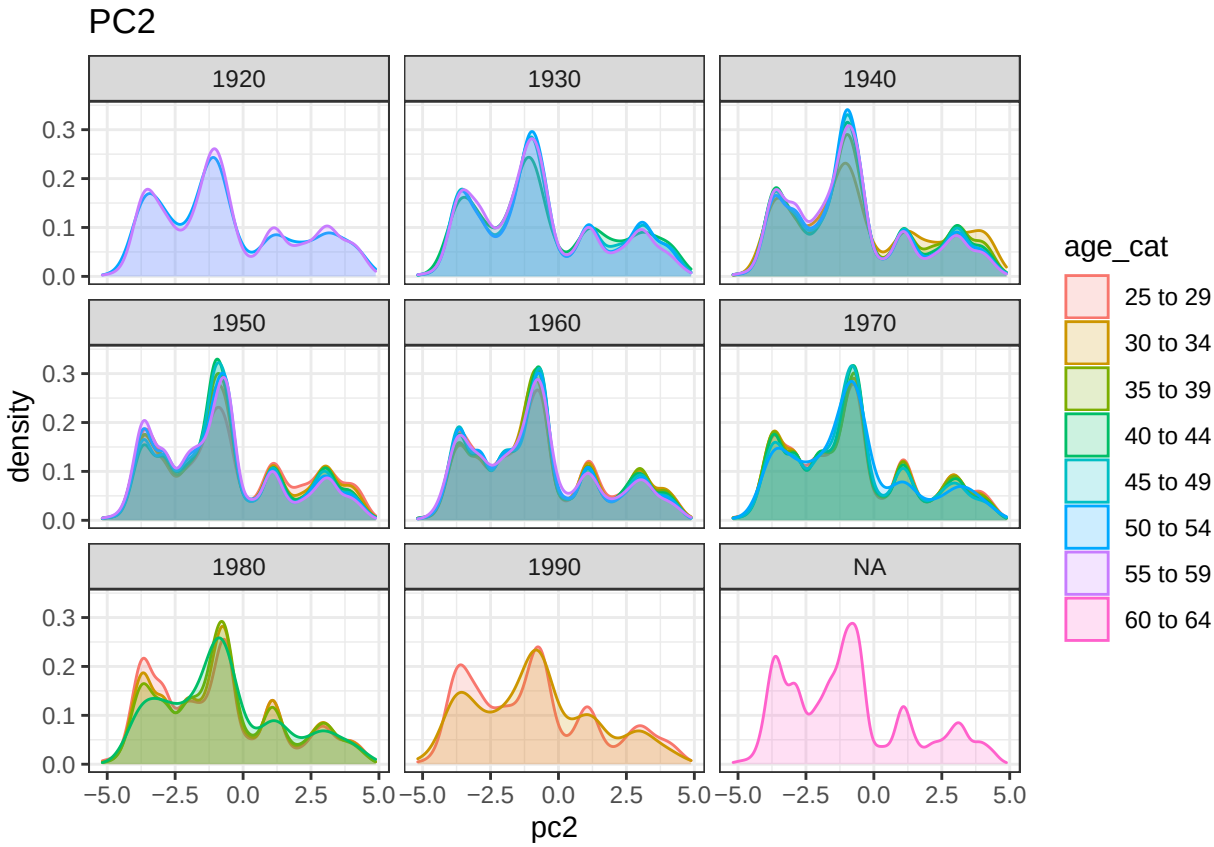

```
ggplot(acs) +
  geom_density(aes(x = programming, color = age_cat, fill = age_cat), alpha = .2) +
  facet_wrap(~cohort) +
  ggtitle("Programming Skills")
```



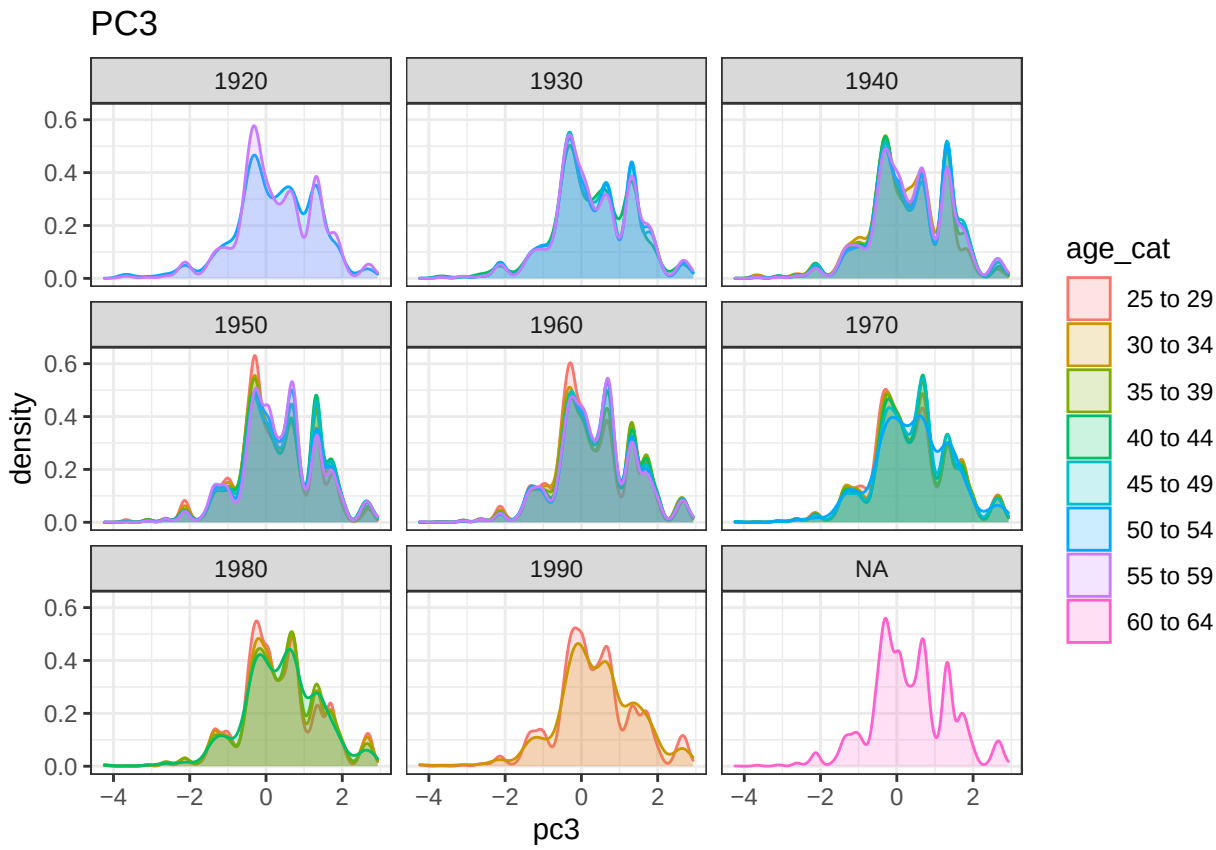

```
ggplot(acs) +
  geom_density(aes(x = pc1, color = age_cat, fill = age_cat), alpha = .2) +
  facet_wrap(~cohort) +
  ggtitle("PC1")
```



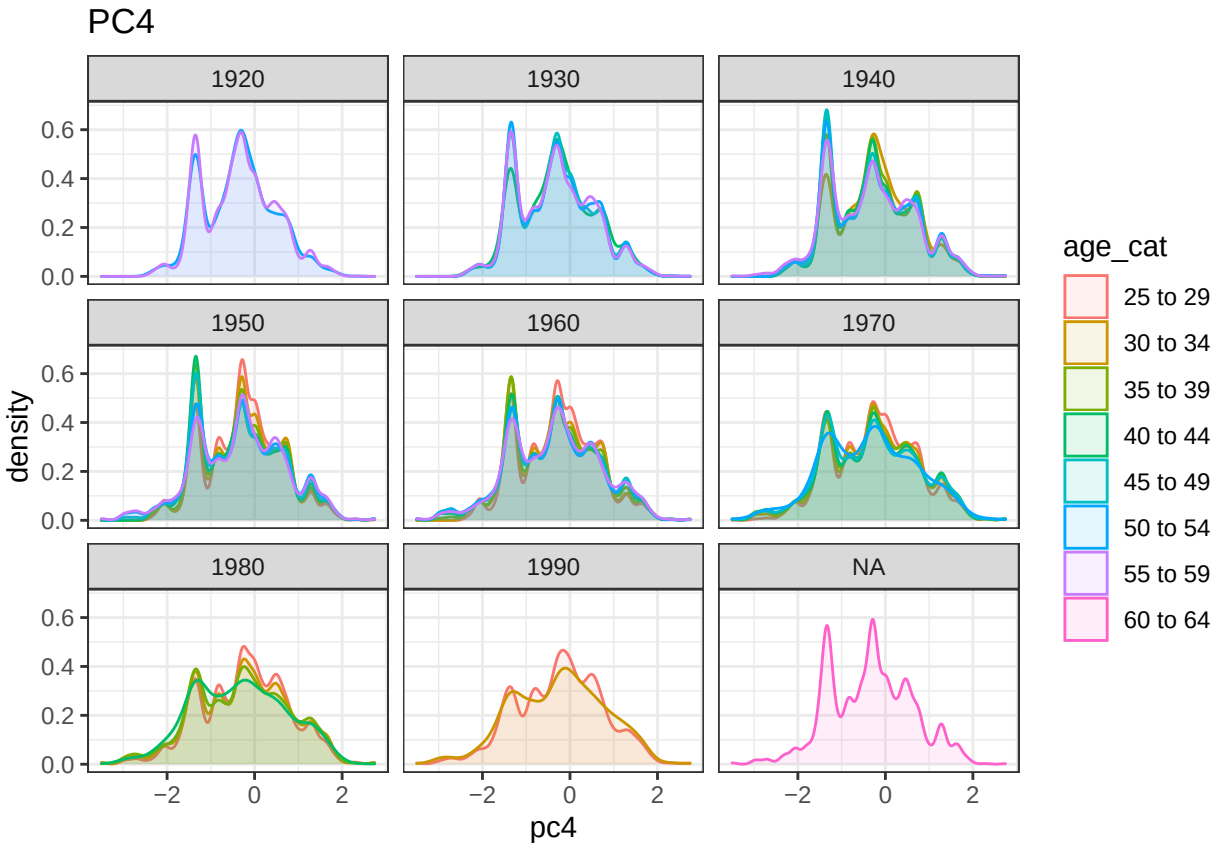

```
ggplot(acs) +
  geom_density(aes(x = pc2, color = age_cat, fill = age_cat), alpha = .2) +
  facet_wrap(~cohort) +
  ggtitle("PC2")
```




```
ggplot(acs) +
  geom_density(aes(x = pc3, color = age_cat, fill = age_cat), alpha = .2) +
  facet_wrap(~cohort) +
  ggtitle("PC3")
```




```
ggplot(acs) +
  geom_density(aes(x = pc4, color = age_cat, fill = age_cat), alpha = .1) +
  facet_wrap(~cohort) +
  ggtitle("PC4")
```

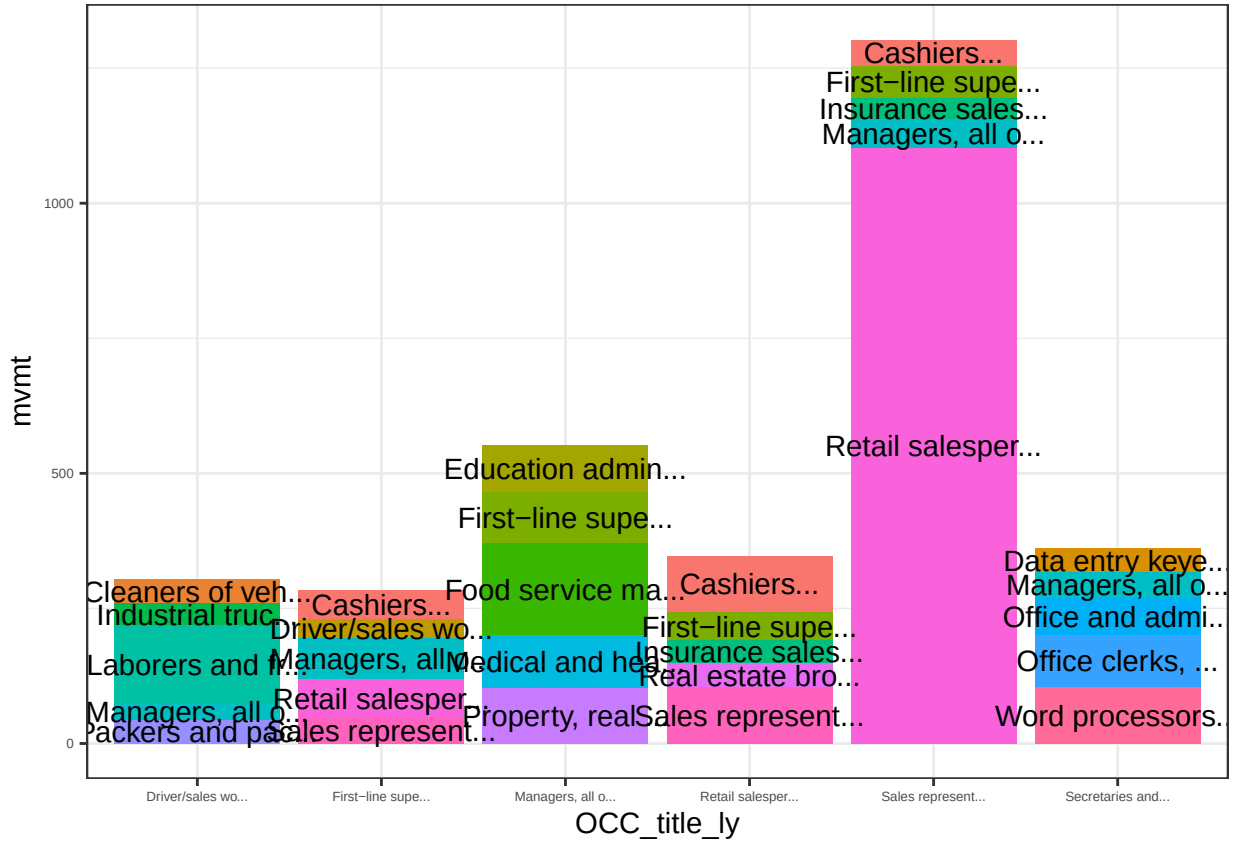


See if movement between occupations clusters by skill, at the individual level

First, visualize movement between occupaitons

```
# merge on both new and old jobs
setnames(acs, vars, paste0(vars, "_current"))
setnames(acs, "CPS Occupational Title","CPS Occupational Title_current")
# merge it all
acs[, OCC10LY := as.character(OCC10LY)]
acs <- merge(acs, skills_final,
             all.x = T,
             by.x = "OCC10LY",
             by.y = "CPS Code")

setnames(acs, vars, paste0(vars, "_ly"))
setnames(acs, "CPS Occupational Title","CPS Occupational Title_ly")

# subset to places where people have moved jobs
acs_moved <- acs[OCC2010 != OCC10LY]

# calculate flows
acs_flows <- acs[!is.na(`CPS Occupational Title_current`) &
                   !is.na(`CPS Occupational Title_ly`),.(mvmt = .N), by = .(OCC2010, OCC10LY, `CPS Occu
```

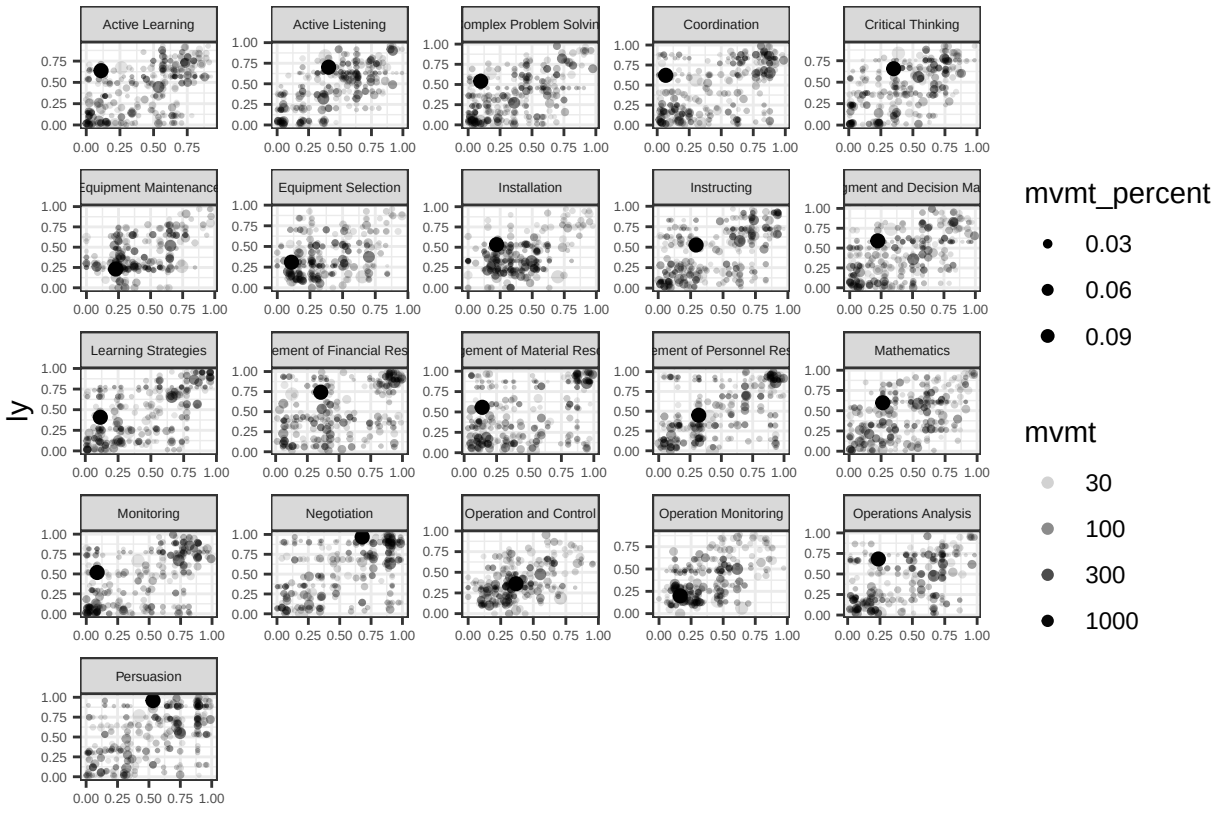
```
# graph top movment
temp <- acs_moved[, .N, by = OCC10LY] %>% .[order(N, decreasing = T)] %>% .[1:7, OCC10LY]
acs_flows[,rank := frankv(mvmt, order = -1L, ties.method = "first"), by = OCC10LY]
acs_flows[, OCC_title_ly := paste0(str_sub(`CPS Occupational Title_ly`, 1, 15),
                                   "...")]
acs_flows[, OCC_title_current:= paste0(str_sub(`CPS Occupational Title_current`, 1, 15),
                                   "...")]

acs_flows[, OCC_title_current := factor(OCC_title_current)]

ggplot(acs_flows[OCC10LY %in% temp & rank %in% 2:6]) +
  geom_bar(aes(x = OCC_title_ly, y = mvmt, fill = OCC_title_current), position = position_stack(),
           stat = "identity") +
  geom_text(aes(x = OCC_title_ly, y = mvmt,label = OCC_title_current, group = OCC_title_current),
            position = position_stack(vjust = 0.5)) +
  guides(fill = F) +
  theme(strip.text = element_text(size = 5), axis.text = element_text(size = 5))
```



See how these cluster by skill

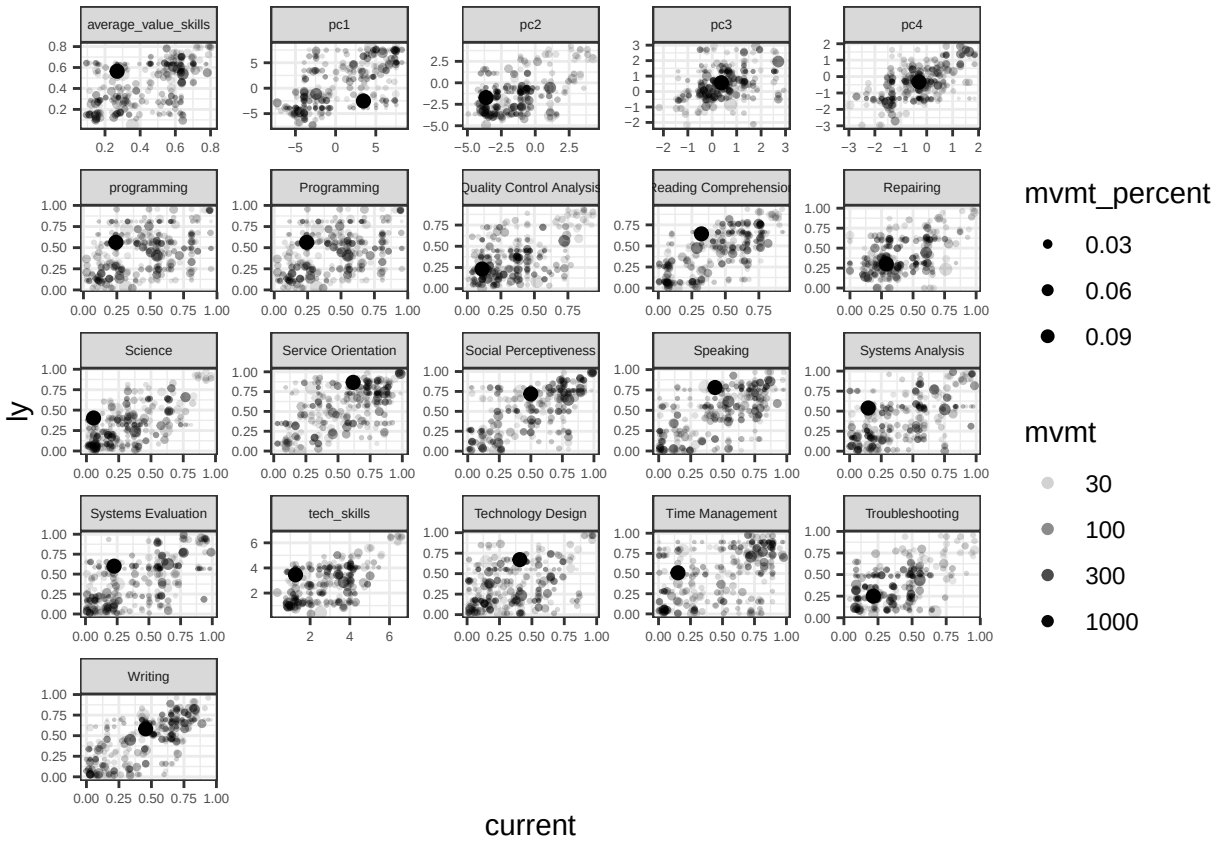
```
skills_subs <- acs[!is.na(`CPS Occupational Title_current`) &
                     !is.na(`CPS Occupational Title_ly`),.SD, .SDcols = names(acs)[names(acs) %like% "_

acs_flows_merged <- merge(acs_flows, skills_subs)
```

```
acs_flows_merged_melt <- melt(acs_flows_merged, id.vars = c("CPS Occupational Title_current",
                                                            "CPS Occupational Title_ly",
                                                            "OCC2010", "OCC10LY",
                                                            "mvmt", "rank"))

acs_flows_merged_melt[, value := as.numeric(value)]
acs_flows_merged_melt <- acs_flows_merged_melt[!is.na(value)]

# cast wide
acs_flows_merged_melt[, current := ifelse(variable %like% "current", "current", "ly")]
acs_flows_merged_melt[, variable := gsub("_current|_ly", "", variable)]
acs_flows_merged_wide <- dcast(acs_flows_merged_melt, ... ~ current, value.var = "value")

# rate standardize
acs_flows_merged_wide[, mvmt_percent := mvmt/sum(mvmt), by = .(OCC10LY, variable)]
# plot
ggplot(acs_flows_merged_wide[OCC2010 != OCC10LY &
                               mvmt > 20 & variable %in% unique(acs_flows_merged_wide$variable)[1:21]]) +
  geom_point(aes(x = current, y = ly, size = mvmt_percent, alpha = mvmt)) +
  facet_wrap(~variable, scales = "free") +
  scale_alpha_continuous(trans = "log10") +
  scale_size_continuous(range = c(0, 2)) +
  theme(strip.text = element_text(size = 5), axis.text = element_text(size = 5))
```




```
ggplot(acs_flows_merged_wide[OCC2010 != OCC10LY &
                               mvmt > 20 & variable %in% unique(acs_flows_merged_wide$variable)[22:42]]) +
  geom_point(aes(x = current, y = ly, size = mvmt_percent, alpha = mvmt)) +
```

```
  facet_wrap(~variable, scales = "free") +
  scale_alpha_continuous(trans = "log10")  +
  scale_size_continuous(range = c(0, 2)) +
  theme(strip.text = element_text(size = 5), axis.text = element_text(size = 5))
```



See if we can recreate the sakomoto graphs from last week

```
acs2 <- data.table(readstata13::read.dta13("../inputs/cps_00004.dta"))
acs2 <- acs2[year %in% seq(1981, 2020, 6)]
#acs2 <- acs2[year %in% c(1981,2020)]

#acs2 <- acs2[
  # acs2$empstat == "at work" &
  #                            (acs2$schlcoll %like% "does not attend|niu"| is.na(acs2$schlcoll)) &
  #                            as.numeric(as.character(acs2$age)) %in% 25:64 &
                             #acs2$wkswork1 >= 30 & ### !!!!! ahhhh
                             #acs2$incwage != 99999999,]
                            # acs2$incwage > 0,]
acs2[, log_incwage := log(incwage + 1)]
acs2[, ed_num := as.numeric(as.character(factor(educ,levels = levels(acs2$educ), labels = c(0,0,0,2.5,1
                                                                                            5,6,7.5, 7,8,9,10,11,11, 11, 11,12,
                                                                                            13,14,14,15,15,15,16,16,17,17,18,18,18,

acs2[ed_num <= 14, ed_categ := "Less than HS"]
acs2[ed_num > 14, ed_categ := "College Plus"]
```

```
# r_sq_dt <- data.table()
# i <- 0
# for(c.year in unique(acs2$year)){
#   for(c.educ_yr in unique(acs2[!is.na(ed_categ)]$ed_categ)){
#     for(c.sex in list("male", "female", c("male", "female"))){
#       i <- i + 1
#       print(i)
#       out <- lm(log_incwage ~ as.factor(occ2010), data = acs2[year == c.year & ed_categ == c.educ_yr &
#                                                               sex %in% c.sex])
#       out_dt <- data.table(year = c.year, ed_categ = c.educ_yr,
#                            sex = paste0(c.sex, collapse = ","),
#                            r_sq = summary(out)$r.squared)
#       r_sq_dt <- rbind(r_sq_dt, out_dt, fill = T)
#     }
#   }
# }
#
# ggplot(r_sq_dt) +
#   geom_point(aes(x = year, y = r_sq, color = ed_categ)) +
#   geom_line(aes(x = year, y = r_sq, color = ed_categ, group = ed_categ)) +
#   labs(title = "R-Squared for Occupation (Most Detailed)\nRegressed on Log(Income), all Cases") +
#   facet_wrap(~sex)
```

Drop missing

```
acs2 <- acs2[
  # acs2$empstat == "at work" &
  #                            (acs2$schlcoll %like% "does not attend|niu"| is.na(acs2$schlcoll)) &
  #                            as.numeric(as.character(acs2$age)) %in% 25:64 &
                              #acs2$wkswork1 >= 30 & ### !!!!! ahhhh
                              acs2$incwage != 99999999,]
                             # acs2$incwage > 0,]

# r_sq_dt <- data.table()
# i <- 0
# for(c.year in unique(acs2$year)){
#   for(c.educ_yr in unique(acs2[!is.na(ed_categ)]$ed_categ)){
#     for(c.sex in list("male", "female", c("male", "female"))){
#       i <- i + 1
#       print(i)
#       out <- lm(log_incwage ~ as.factor(occ2010), data = acs2[year == c.year & ed_categ == c.educ_yr &
#                                                               sex %in% c.sex])
#       out_dt <- data.table(year = c.year, ed_categ = c.educ_yr,
#                            sex = paste0(c.sex, collapse = ","),
#                            r_sq = summary(out)$r.squared)
#       r_sq_dt <- rbind(r_sq_dt, out_dt, fill = T)
#     }
#   }
# }
#
# ggplot(r_sq_dt) +
#   geom_point(aes(x = year, y = r_sq, color = ed_categ)) +
```

```
#   geom_line(aes(x = year, y = r_sq, color = ed_categ, group = ed_categ)) +
#   labs(title = "R-Squared for Occupation (Most Detailed)\nRegressed on Log(Income), all Cases") +
#   facet_wrap(~sex)
```

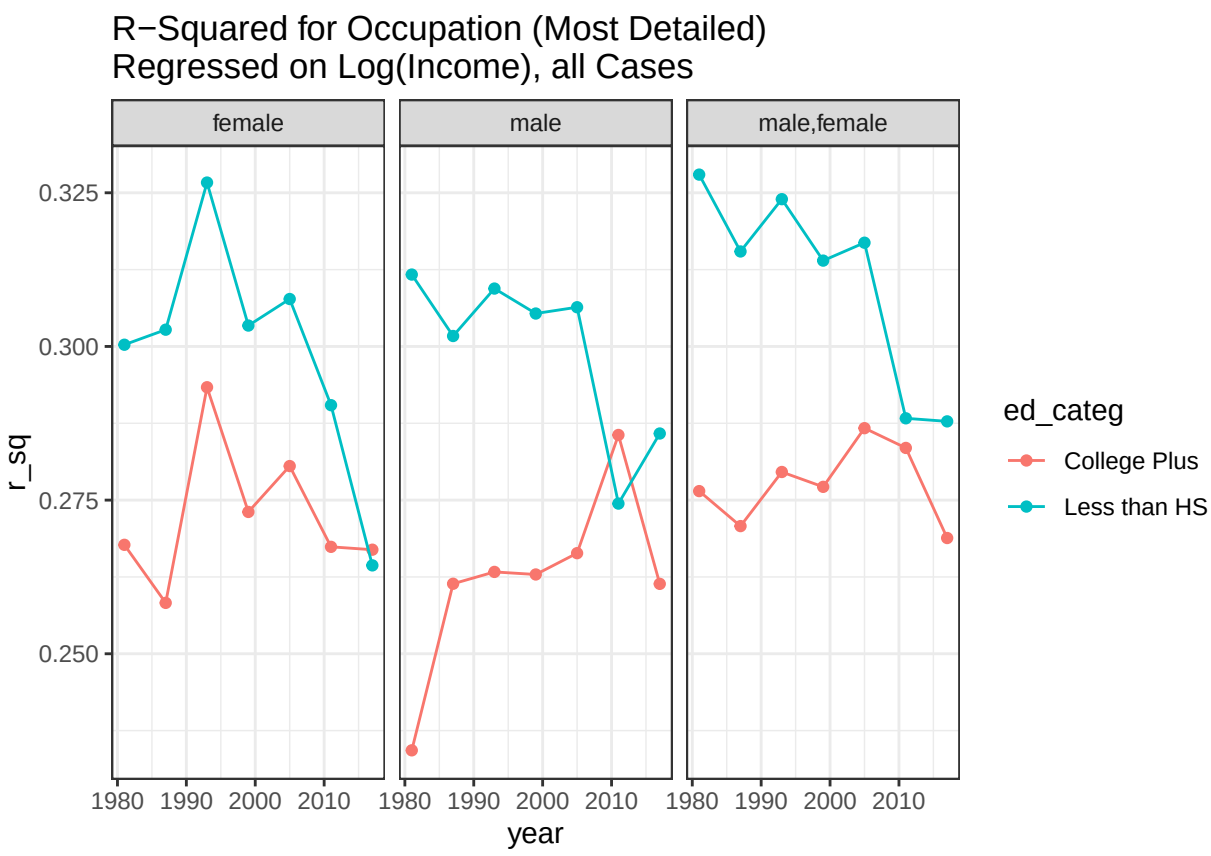
Drop missing + 0 earners

```
acs2 <- acs2[
  # acs2$empstat == "at work" &
  #                             (acs2$schlcoll %like% "does not attend|niu"| is.na(acs2$schlcoll)) &
  #                             as.numeric(as.character(acs2$age)) %in% 25:64 &
                              #acs2$wkswork1 >= 30 & ### !!!!! ahhhh
                              acs2$incwage != 99999999 &
                              acs2$incwage > 0,]

r_sq_dt <- data.table()
i <- 0
for(c.year in unique(acs2$year)){
  for(c.educ_yr in unique(acs2[!is.na(ed_categ)]$ed_categ)){
    for(c.sex in list("male", "female", c("male", "female"))){
      i <- i + 1
      print(i)
      out <- lm(log_incwage ~ as.factor(occ2010), data = acs2[year == c.year & ed_categ == c.educ_yr &
                                                              sex %in% c.sex])
      out_dt <- data.table(year = c.year, ed_categ = c.educ_yr,
                           sex = paste0(c.sex, collapse = ","),
                           r_sq = summary(out)$r.squared)
      r_sq_dt <- rbind(r_sq_dt, out_dt, fill = T)
    }
  }
}
```

```
## [1] 1
## [1] 2
## [1] 3
## [1] 4
## [1] 5
## [1] 6
## [1] 7
## [1] 8
## [1] 9
## [1] 10
## [1] 11
## [1] 12
## [1] 13
## [1] 14
## [1] 15
## [1] 16
## [1] 17
## [1] 18
## [1] 19
## [1] 20
## [1] 21
## [1] 22
```

```
## [1] 23
## [1] 24
## [1] 25
## [1] 26
## [1] 27
## [1] 28
## [1] 29
## [1] 30
## [1] 31
## [1] 32
## [1] 33
## [1] 34
## [1] 35
## [1] 36
## [1] 37
## [1] 38
## [1] 39
## [1] 40
## [1] 41
## [1] 42
```

```
ggplot(r_sq_dt) +
  geom_point(aes(x = year, y = r_sq, color = ed_categ)) +
  geom_line(aes(x = year, y = r_sq, color = ed_categ, group = ed_categ)) +
  labs(title = "R-Squared for Occupation (Most Detailed)\nRegressed on Log(Income), all Cases") +
  facet_wrap(~sex)
```



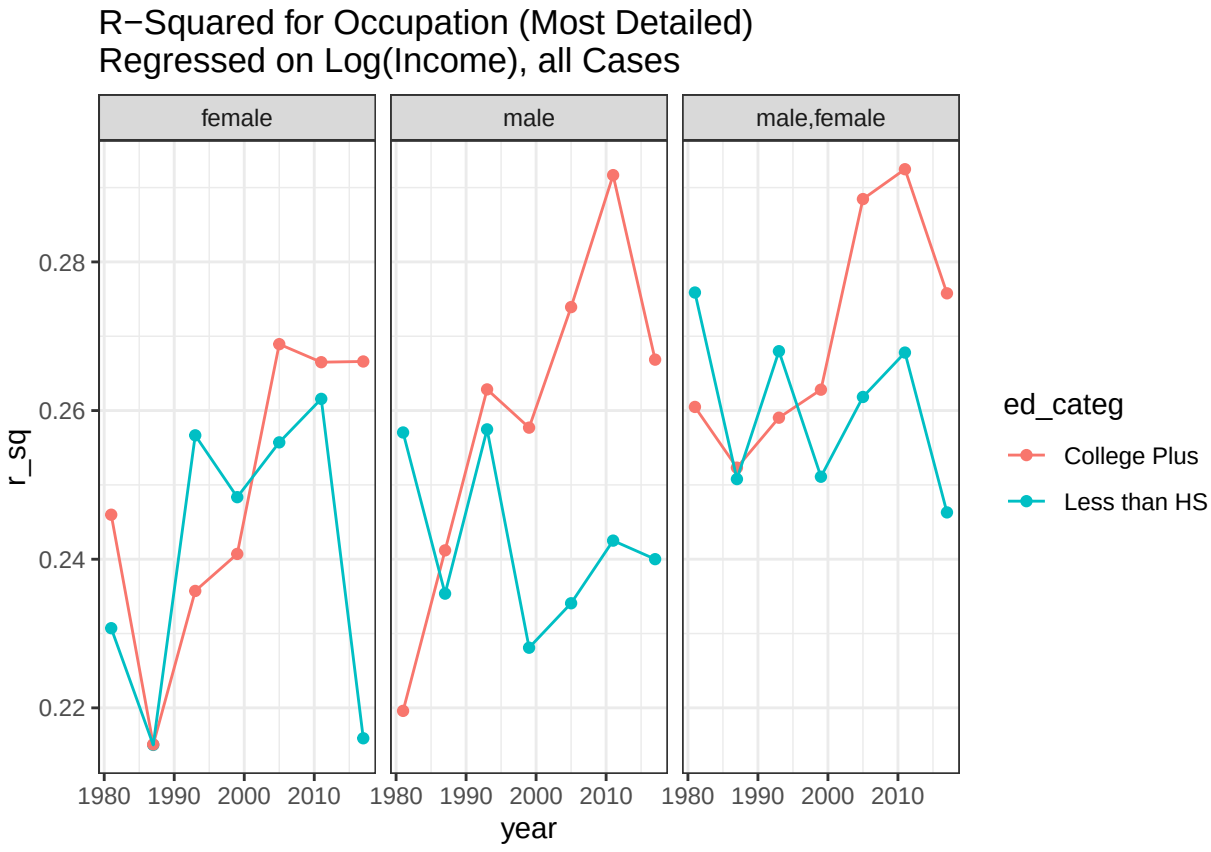
Drop missing + 0 earners and part timers

```
acs2 <- acs2[
  # acs2$empstat == "at work" &
  #                              (acs2$schlcoll %like% "does not attend|niu"| is.na(acs2$schlcoll)) &
  #                              as.numeric(as.character(acs2$age)) %in% 25:64 &
                               acs2$wkswork1 >= 30 & ### !!!!! ahhhh
                               acs2$incwage != 99999999 &
                               acs2$incwage > 0,]

r_sq_dt <- data.table()
i <- 0
for(c.year in unique(acs2$year)){
  for(c.educ_yr in unique(acs2[!is.na(ed_categ)]$ed_categ)){
    for(c.sex in list("male", "female", c("male", "female"))){
      i <- i + 1
      print(i)
      out <- lm(log_incwage ~ as.factor(occ2010), data = acs2[year == c.year & ed_categ == c.educ_yr &
                                                              sex %in% c.sex])
      out_dt <- data.table(year = c.year, ed_categ = c.educ_yr,
                           sex = paste0(c.sex, collapse = ","),
                           r_sq = summary(out)$r.squared)
      r_sq_dt <- rbind(r_sq_dt, out_dt, fill = T)
    }
  }
}
```

```
## [1] 1
## [1] 2
## [1] 3
## [1] 4
## [1] 5
## [1] 6
## [1] 7
## [1] 8
## [1] 9
## [1] 10
## [1] 11
## [1] 12
## [1] 13
## [1] 14
## [1] 15
## [1] 16
## [1] 17
## [1] 18
## [1] 19
## [1] 20
## [1] 21
## [1] 22
## [1] 23
## [1] 24
## [1] 25
## [1] 26
## [1] 27
```

```
## [1] 28
## [1] 29
## [1] 30
## [1] 31
## [1] 32
## [1] 33
## [1] 34
## [1] 35
## [1] 36
## [1] 37
## [1] 38
## [1] 39
## [1] 40
## [1] 41
## [1] 42
```

```
ggplot(r_sq_dt) +
  geom_point(aes(x = year, y = r_sq, color = ed_categ)) +
  geom_line(aes(x = year, y = r_sq, color = ed_categ, group = ed_categ)) +
  labs(title = "R-Squared for Occupation (Most Detailed)\nRegressed on Log(Income), all Cases") +
  facet_wrap(~sex)
```



Compositional Biases induced by dropping part time workers and 0 earners + missing

```
acs2 <- acs2[
   acs2$empstat %like% "at work|armed|has job" &
                              (acs2$schlcoll %like% "does not attend|niu"| is.na(acs2$schlcoll)) &
```

```
                              as.numeric(as.character(acs2$age)) %in% 25:64 &
                             acs2$wkswork1 >= 30 & ### !!!!! ahhhh
                             acs2$incwage != 99999999 &
                             acs2$incwage > 0,]


r_sq_dt <- data.table()
i <- 0
for(c.year in unique(acs2$year)){
  for(c.educ_yr in unique(acs2[!is.na(ed_categ)]$ed_categ)){
    for(c.sex in list("male", "female", c("male", "female"))){
      i <- i + 1
      print(i)
      out <- lm(log_incwage ~ as.factor(occ2010), data = acs2[year == c.year & ed_categ == c.educ_yr &
                                                              sex %in% c.sex])
      out_dt <- data.table(year = c.year, ed_categ = c.educ_yr,
                           sex = paste0(c.sex, collapse = ","),
                           r_sq = summary(out)$r.squared)
      r_sq_dt <- rbind(r_sq_dt, out_dt, fill = T)
    }
  }
}
```

```
## [1] 1
## [1] 2
## [1] 3
## [1] 4
## [1] 5
## [1] 6
## [1] 7
## [1] 8
## [1] 9
## [1] 10
## [1] 11
## [1] 12
## [1] 13
## [1] 14
## [1] 15
## [1] 16
## [1] 17
## [1] 18
## [1] 19
## [1] 20
## [1] 21
## [1] 22
## [1] 23
## [1] 24
## [1] 25
## [1] 26
## [1] 27
## [1] 28
## [1] 29
## [1] 30
## [1] 31
```

```
## [1] 32
## [1] 33
## [1] 34
## [1] 35
## [1] 36
## [1] 37
## [1] 38
## [1] 39
## [1] 40
## [1] 41
## [1] 42
```

```
ggplot(r_sq_dt) +
  geom_point(aes(x = year, y = r_sq, color = ed_categ)) +
  geom_line(aes(x = year, y = r_sq, color = ed_categ, group = ed_categ)) +
  labs(title = "R-Squared for Occupation (Most Detailed)\nRegressed on Log(Income), Partial Cases") +
  facet_wrap(~sex)
```