analysis_iv_ii

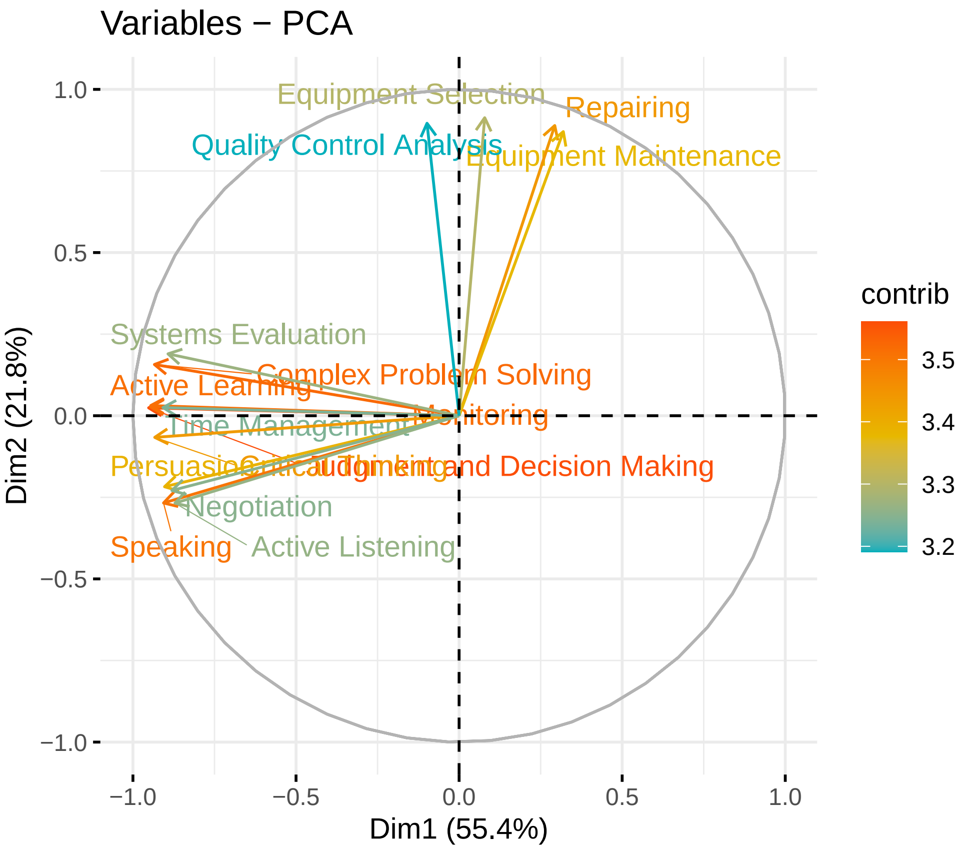
Hunter York

10/31/2020

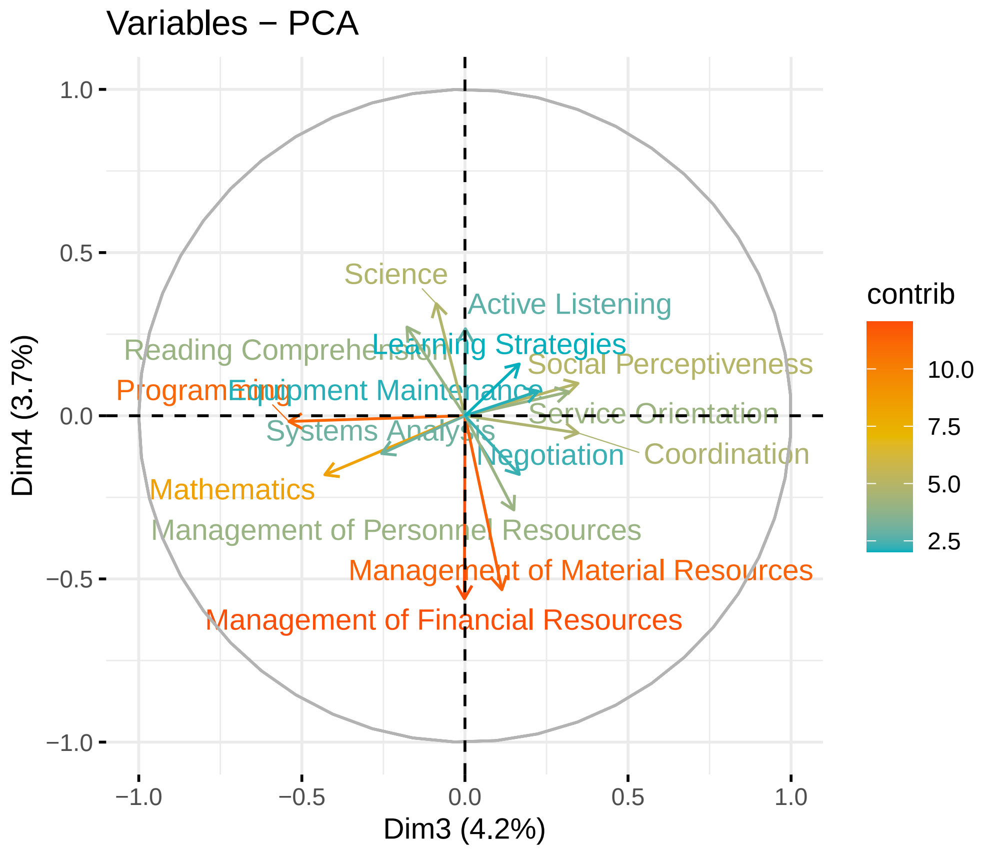
Examine how the import of skills has changed over time

Do a quick PCA to see how skills vary with respect to each other

```
# cast wide and do factor analysis
skills <- skills[OCCSOC %in% unique(acs$OCCSOC),
                 .(Data.Value = mean(Data.Value)), by = .(Element.Name, Scale.ID, OCCSOC)]


skills_wide <- dcast(skills[Scale.ID == "LV"], OCCSOC   ~Element.Name, value.var = "Data.Value")
res.pca <- prcomp(skills_wide[,2:34], scale = TRUE, center = T)

fviz_pca_var(res.pca,
             col.var = "contrib", # Color by contributions to the PC
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE ,
             select.var = list(contrib = 15)    # Avoid text overlapping
)
```

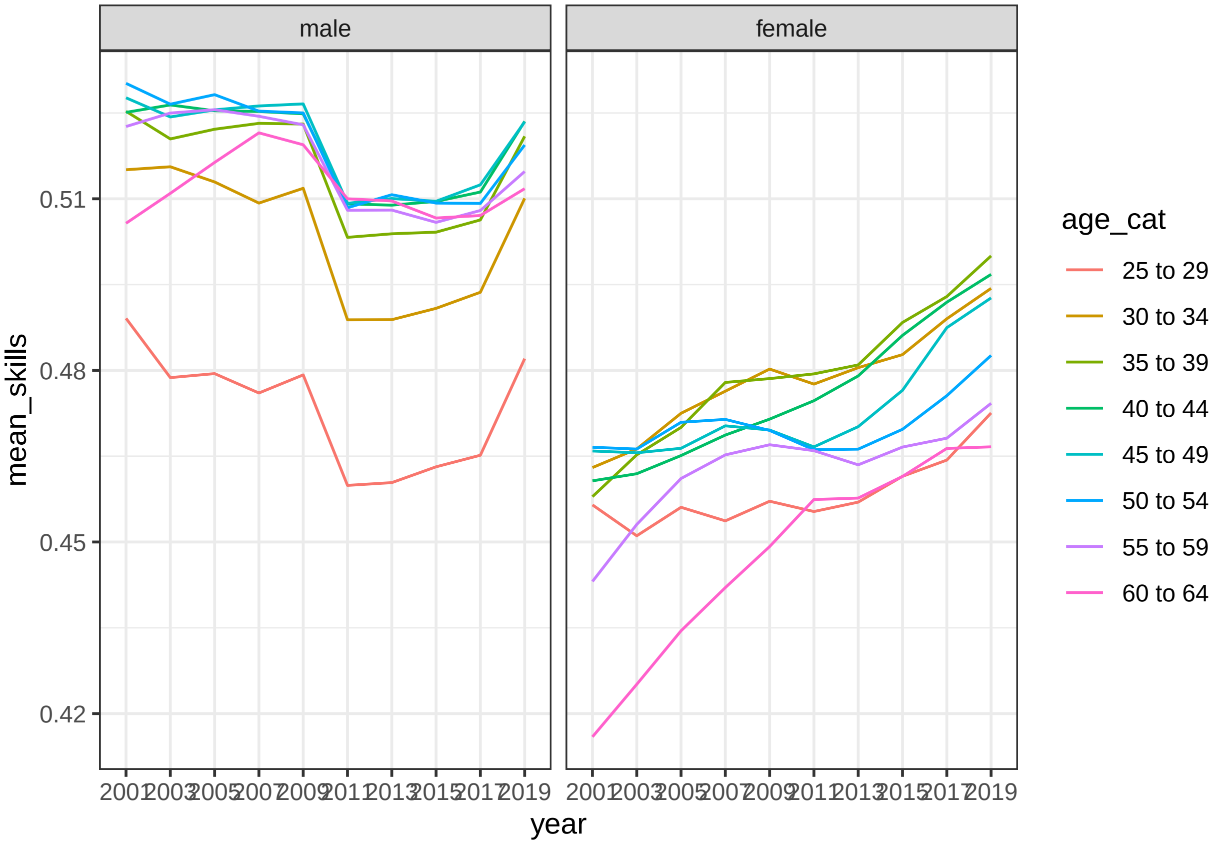


```
fviz_pca_var(res.pca,
             axes = c(3,4),
             col.var = "contrib", # Color by contributions to the PC
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE,# Avoid text overlapping,
             select.var = list(contrib = 15)
)
```

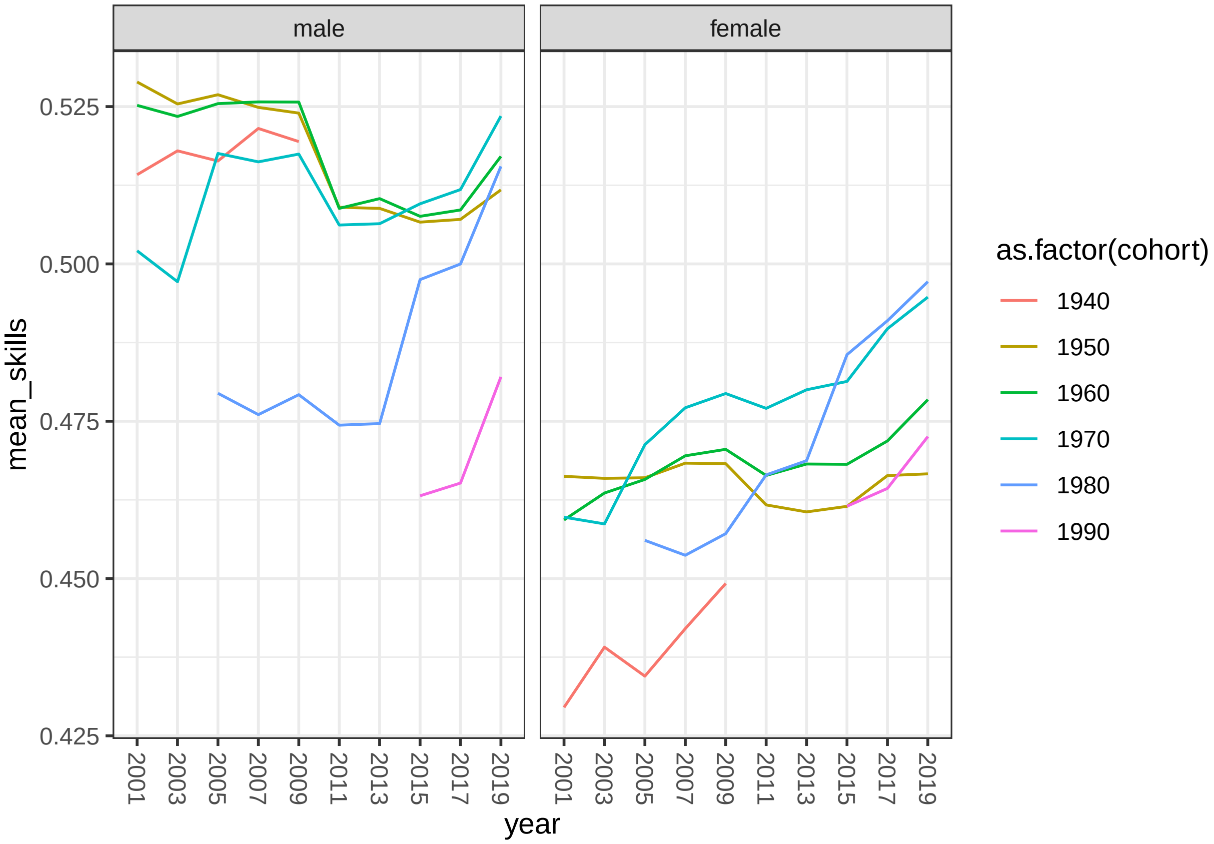


See how mean skills change across survey year using ACS data

```
# make a skills dataset that has a few interesting variables
skills_sum <- skills[Scale.ID == "LV",
                     .(average_value_skills = mean(Data.Value, na.rm = T)), by = .(OCCSOC)]


acs <- merge(acs, skills_sum, by = c("OCCSOC"), all.x = T)

skills_overview <- acs[,.(mean_skills = weighted.mean(average_value_skills, w = perwt, na.rm = T)), by =

skills_overview[, cohort := floor((as.numeric(year) - as.numeric(substr(age_cat,1,2)))/10)*10]

ggplot(skills_overview) +
  geom_line(aes(x = year, y = mean_skills, color = age_cat, group = age_cat)) +
  facet_wrap(~sex)
```



```
skills_overview[,.(mean_skills = mean(mean_skills)), by = .(cohort, year, sex)] %>%
  ggplot(.) +
  geom_line(aes(x = year, y = mean_skills, color = as.factor(cohort), group = as.factor(cohort))) +
  facet_wrap(~sex) +
  theme(axis.text.x = element_text(angle = -90, vjust = 0.5, hjust=1))
```

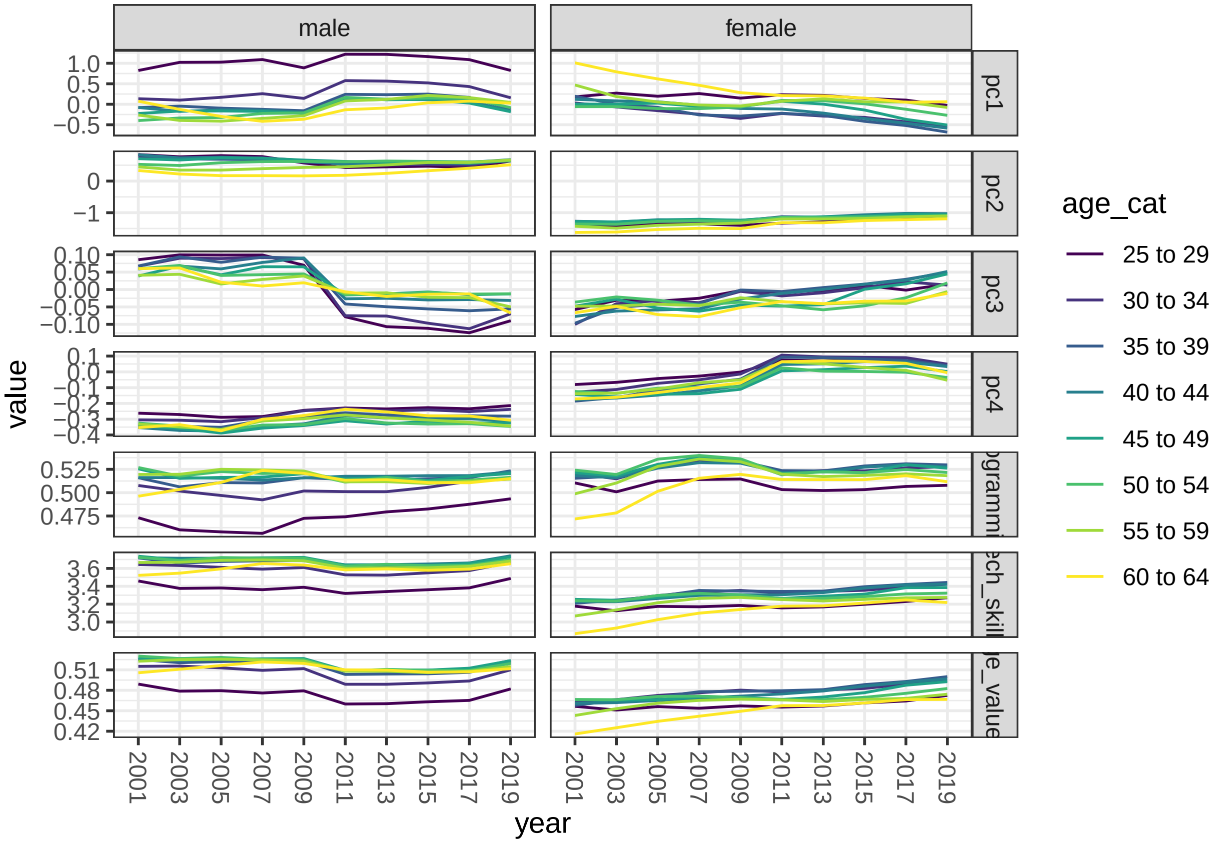


Recalculate skills by year to capture more interesting variables, like LV1 and LV2, Programming, etc

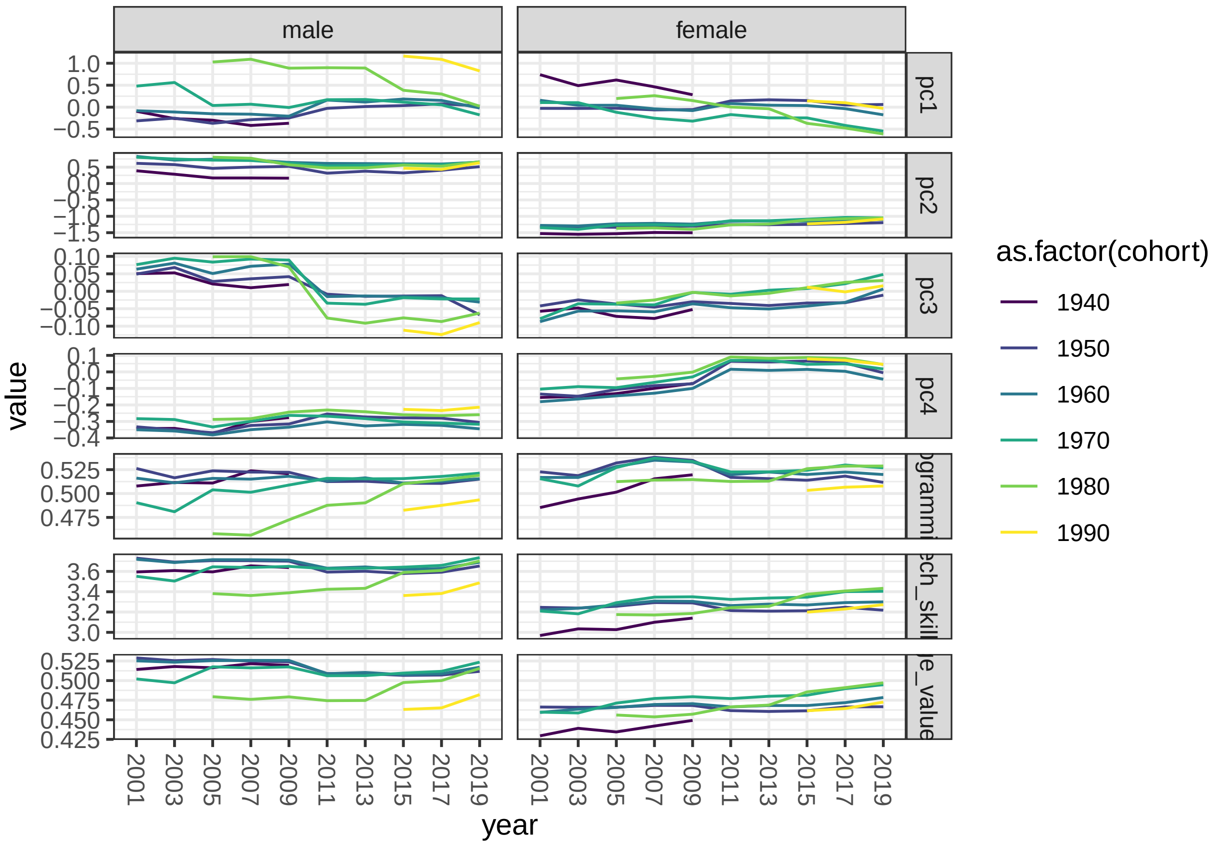
```
# skills wide
skills_sum <- skills_wide
skills_sum[, pc1 := predict(res.pca, newdata = .SD)[,1], .SDcols = names(skills_sum)]
skills_sum[, pc2 := predict(res.pca, newdata = .SD)[,2], .SDcols = names(skills_sum)]
skills_sum[, pc3 := predict(res.pca, newdata = .SD)[,3], .SDcols = names(skills_sum)]
skills_sum[, pc4 := predict(res.pca, newdata = .SD)[,4], .SDcols = names(skills_sum)]

skills_sum[, programming := Programming]
skills_sum[, tech_skills := Programming + `Complex Problem Solving` +
             `Mathematics` + Programming + Science + `Systems Analysis` +
             Troubleshooting]
skills_sum[, average_value_skills := rowMeans(.SD), .SDcols = rownames(res.pca$rotation)]
```

See how mean skills change across survey year

```
# average across years
# REMOVE THIS IF YOU WANT YEAR SPECIFIC SKILLS RATINGS
# skills_sum <- skills_sum[year == 2018,.(pc1 = mean(pc1),
#                               pc2 = mean(pc2),
#                               programming = mean(programming)), by = "OCCSOC"]

acs <- merge(acs, skills_sum[,.(OCCSOC, pc1, pc2,pc3, pc4, programming, tech_skills)], by = c( "OCCSOC"
#

skills_overview2 <- acs[,.(pc1 = weighted.mean(pc1,w = perwt, na.rm = T),
                           pc2 = weighted.mean(pc2,w = perwt,  na.rm = T),
                           pc3 = weighted.mean(pc3,w = perwt,  na.rm = T),
                           pc4 = weighted.mean(pc4,w = perwt,  na.rm = T),
                           programming = weighted.mean(programming,w = perwt,  na.rm = T),
                           tech_skills = weighted.mean(tech_skills, w = perwt, na.rm = T),
                           average_value_skills = weighted.mean(average_value_skills, w = perwt, na.rm =

skills_overview2_melt <- melt(skills_overview2, id.vars = c("year", "age_cat", "sex"))

skills_overview2_melt[, cohort := floor((as.numeric(year) - as.numeric(substr(age_cat,1,2)))/10)*10]


ggplot(skills_overview2_melt) +
  geom_line(aes(x = year, y = value, color = age_cat, group = age_cat)) +
  facet_grid(variable~sex, scales = "free") +
  scale_color_viridis_d()+
  theme(axis.text.x = element_text(angle = -90, vjust = 0.5, hjust=1))
```

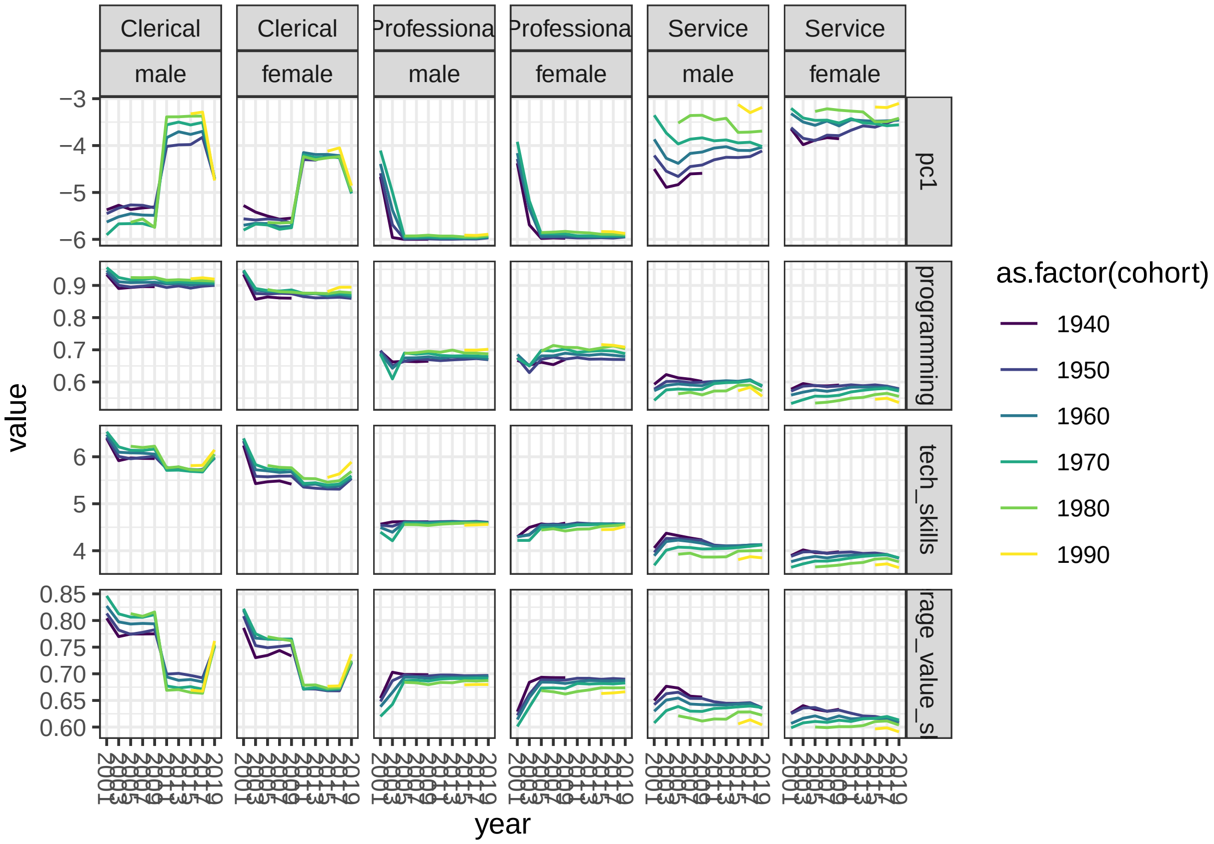


```
skills_overview2_melt[,.(value = mean(value)), by = .(cohort, sex, variable, year)] %>%
  ggplot(.) +
  geom_line(aes(x = year, y = value, color = as.factor(cohort), group = as.factor(cohort))) +
  facet_grid(variable~sex, scales = "free") +
  scale_color_viridis_d()+
  theme(axis.text.x = element_text(angle = -90, vjust = 0.5, hjust=1))
```

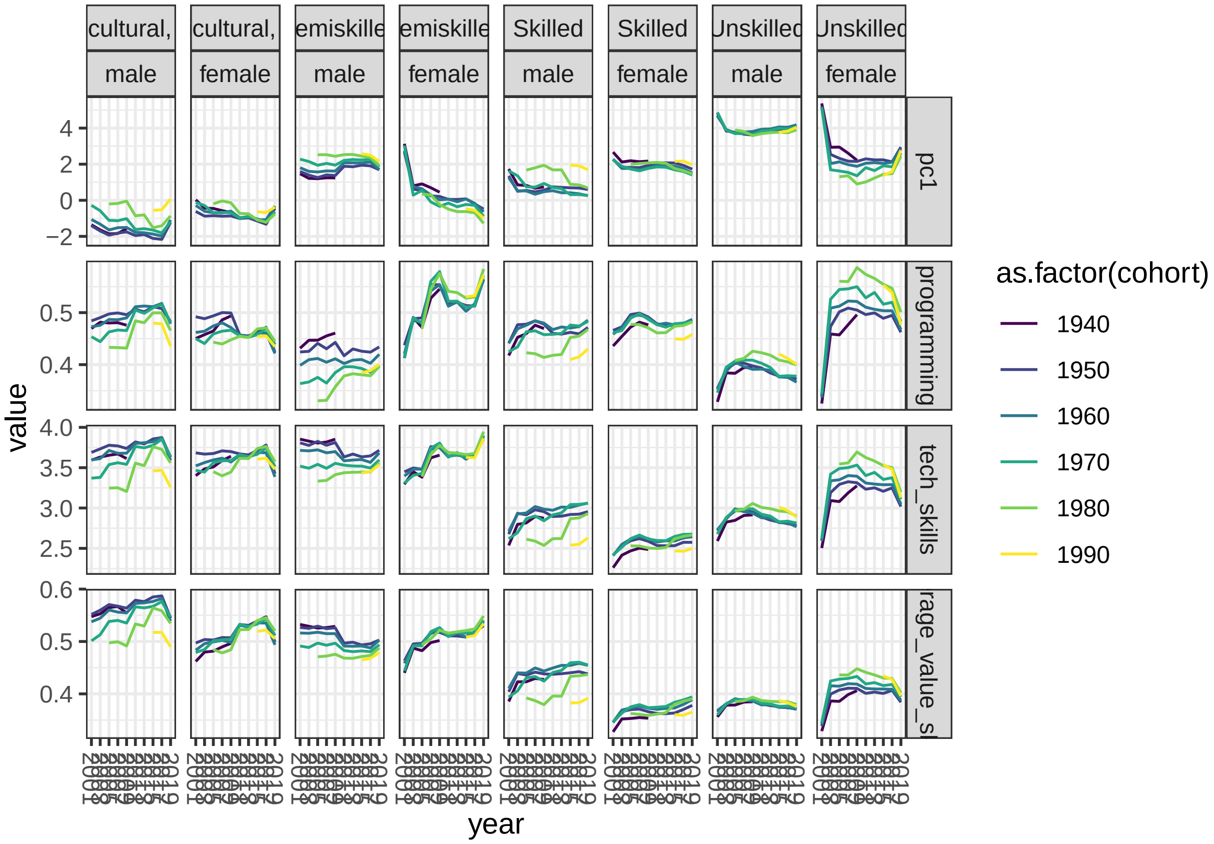


See how changes in skill level are spread over occupational grouping

```
skills_overview3 <- acs[,.(pc1 = weighted.mean(pc1,w = perwt, na.rm = T),
                           pc2 = weighted.mean(pc2,w = perwt,  na.rm = T),
                           pc3 = weighted.mean(pc3,w = perwt,  na.rm = T),
                           pc4 = weighted.mean(pc4,w = perwt,  na.rm = T),
                           programming = weighted.mean(programming,w = perwt,  na.rm = T),
                           tech_skills = weighted.mean(tech_skills, w = perwt, na.rm = T), average_valu

skills_overview3_melt <- melt(skills_overview3, id.vars = c("year", "age_cat", "sex" ,"occ_categ"))

skills_overview3_melt[, cohort := floor((as.numeric(year) - as.numeric(substr(age_cat,1,2)))/10)*10]

skills_overview3_melt <- skills_overview3_melt[,.(value = mean(value)), by = .(cohort, variable, year,

ggplot(skills_overview3_melt[variable %like% "pc1|progr|tech|averag" &
                               occ_categ %in% unique(skills_overview3_melt$occ_categ)[1:3]]) +
  geom_line(aes(x = year, y = value, color = as.factor(cohort), group = as.factor(cohort))) +
  facet_grid( variable ~occ_categ+sex, scales = "free") +
  scale_color_viridis_d()+
  theme(axis.text.x = element_text(angle = -90, vjust = 0.5, hjust=1))
```

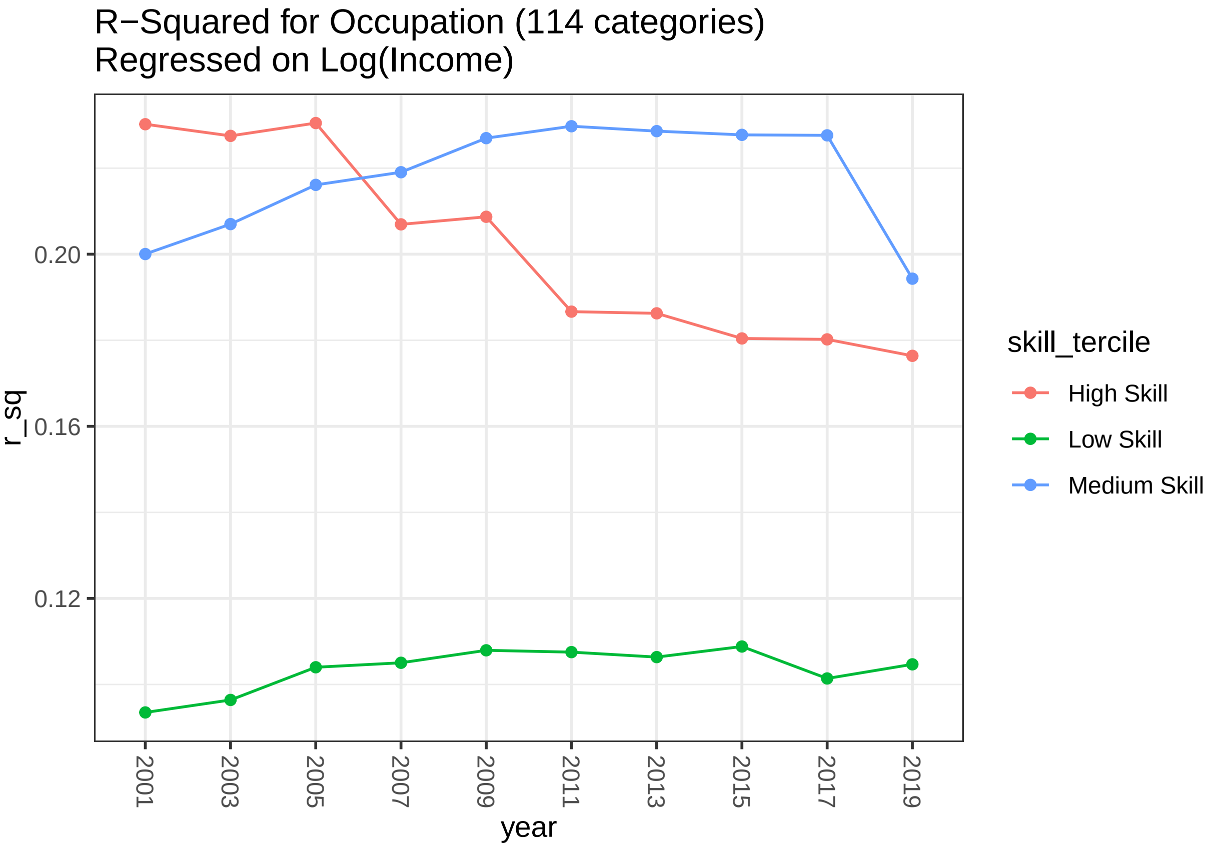


```
ggplot(skills_overview3_melt[variable %like% "pc1|progr|tech|averag" &
                               occ_categ %in% unique(skills_overview3_melt$occ_categ)[4:7]]) +
  geom_line(aes(x = year, y = value, color = as.factor(cohort), group = as.factor(cohort))) +
  facet_grid( variable ~occ_categ+sex, scales = "free") +
  scale_color_viridis_d() +
  theme(axis.text.x = element_text(angle = -90, vjust = 0.5, hjust=1))
```

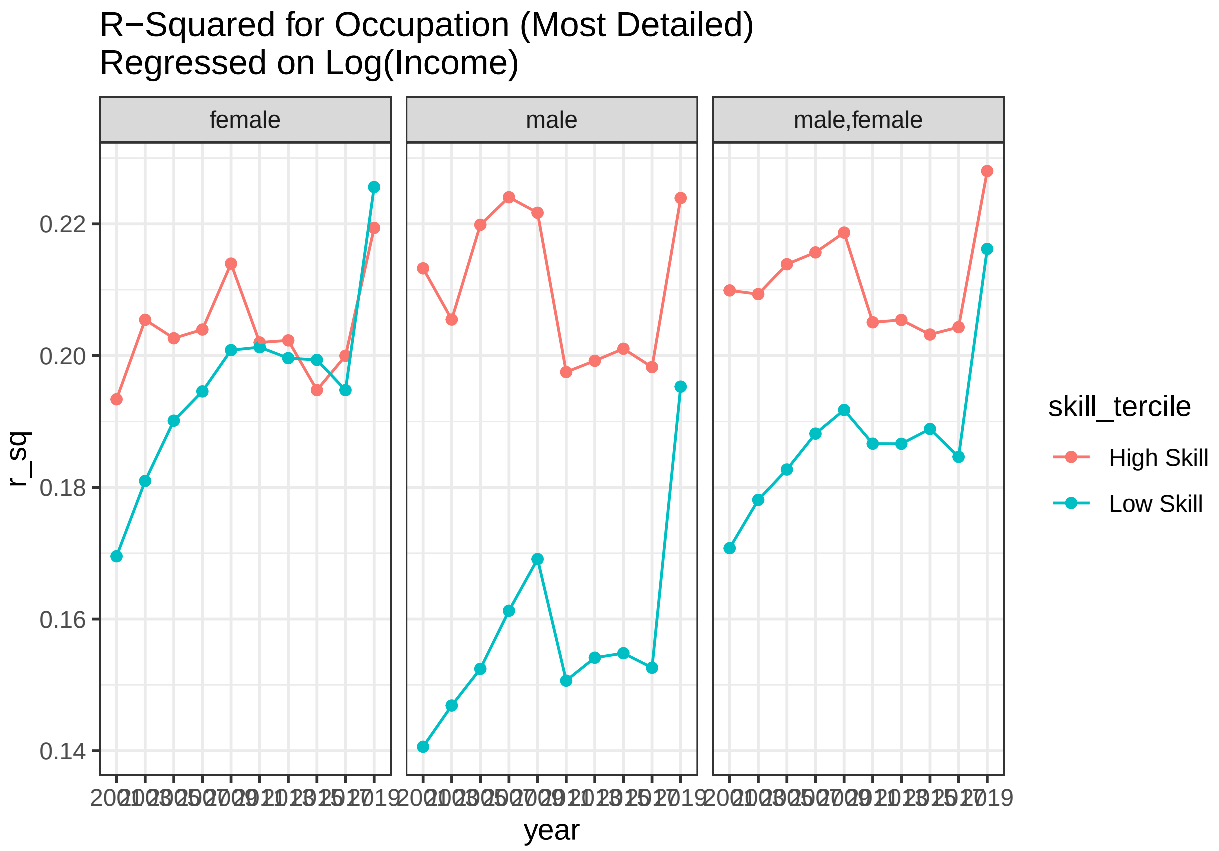


Decompose by skill tercile

```r
acs[average_value_skills >= .66, skill_tercile := "High Skill"]
acs[average_value_skills < .66 &average_value_skills >= .33, skill_tercile := "Medium Skill"]
acs[average_value_skills < .33, skill_tercile := "Low Skill"]

acs[average_value_skills > .5, skill_half := "High Skill"]
acs[average_value_skills <= .5, skill_half := "Low Skill"]

#
acs[, occsoc_sub := substr(OCCSOC,1,4)]
# see if I can recreate sakamoto's graphs
r_sq_dt <- data.table()
i <- 0
for(c.year in unique(acs$year)){
  for(c.skill in unique(acs[!is.na(skill_tercile)]$skill_tercile)){
    i <- i + 1
    #print(i)
    out <- lm(log_incwage ~ occsoc_sub, data = acs[year == c.year & skill_tercile == c.skill])
    out_dt <- data.table(year = c.year, skill_tercile = c.skill,
                         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 = skill_tercile)) +
  geom_line(aes(x = year, y = r_sq, color = skill_tercile, group = skill_tercile)) +
  labs(title = "R-Squared for Occupation (114 categories)\nRegressed on Log(Income)")+
    theme(axis.text.x = element_text(angle = -90, vjust = 0.5, hjust=1))
```



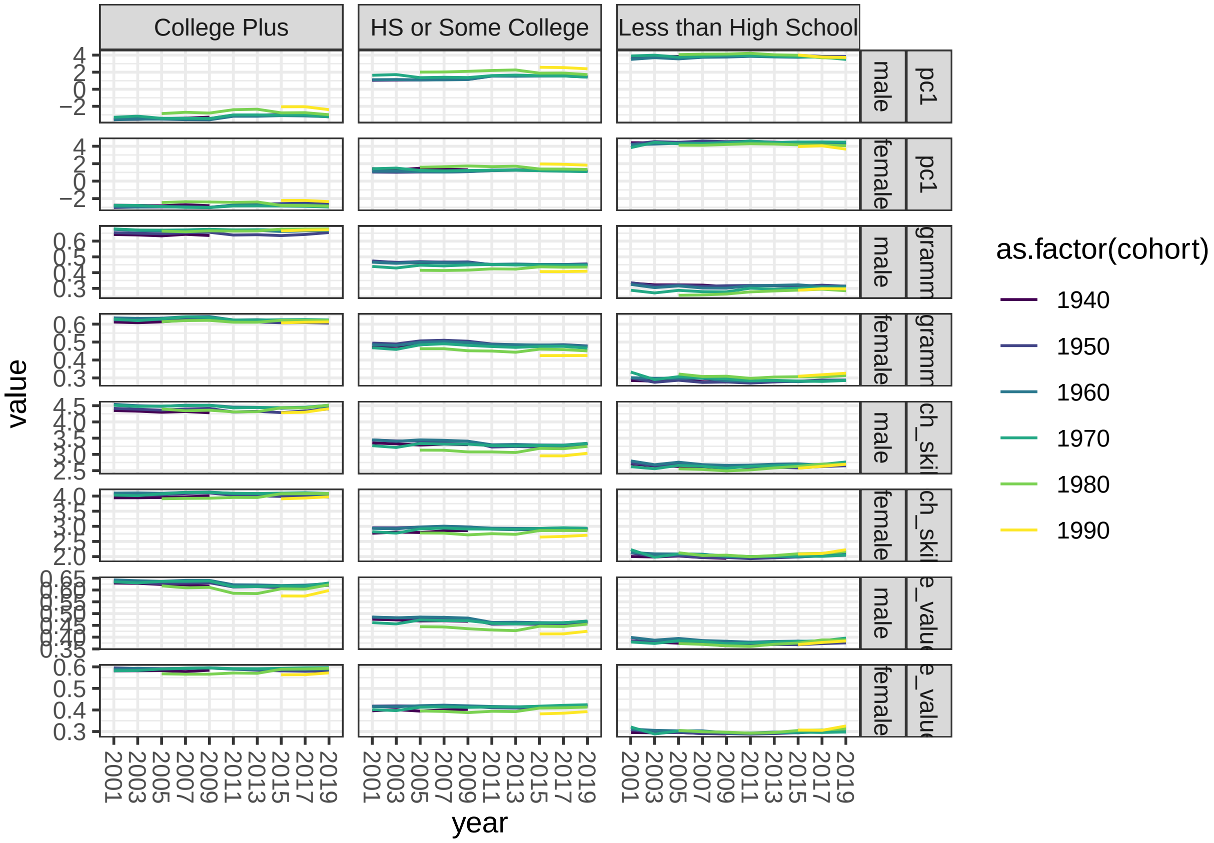
```
r_sq_dt <- data.table()
i <- 0
for(c.year in unique(acs$year)){
  for(c.skill in unique(acs[!is.na(skill_half)]$skill_half)){
    for(c.sex in list("male", "female", c("male", "female"))){
      i <- i + 1
      #print(i)
      out <- lm(log_incwage ~ OCCSOC, data = acs[year == c.year & skill_half == c.skill &
                                                    sex %in% c.sex])
      out_dt <- data.table(year = c.year, skill_tercile = c.skill,
                           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 = skill_tercile)) +
  geom_line(aes(x = year, y = r_sq, color = skill_tercile, group = skill_tercile)) +
  labs(title = "R-Squared for Occupation (Most Detailed)\nRegressed on Log(Income)") +
  facet_wrap(~sex)
```



Does inequality in skills track inequality in earnings?

Is there a return to education re: skills

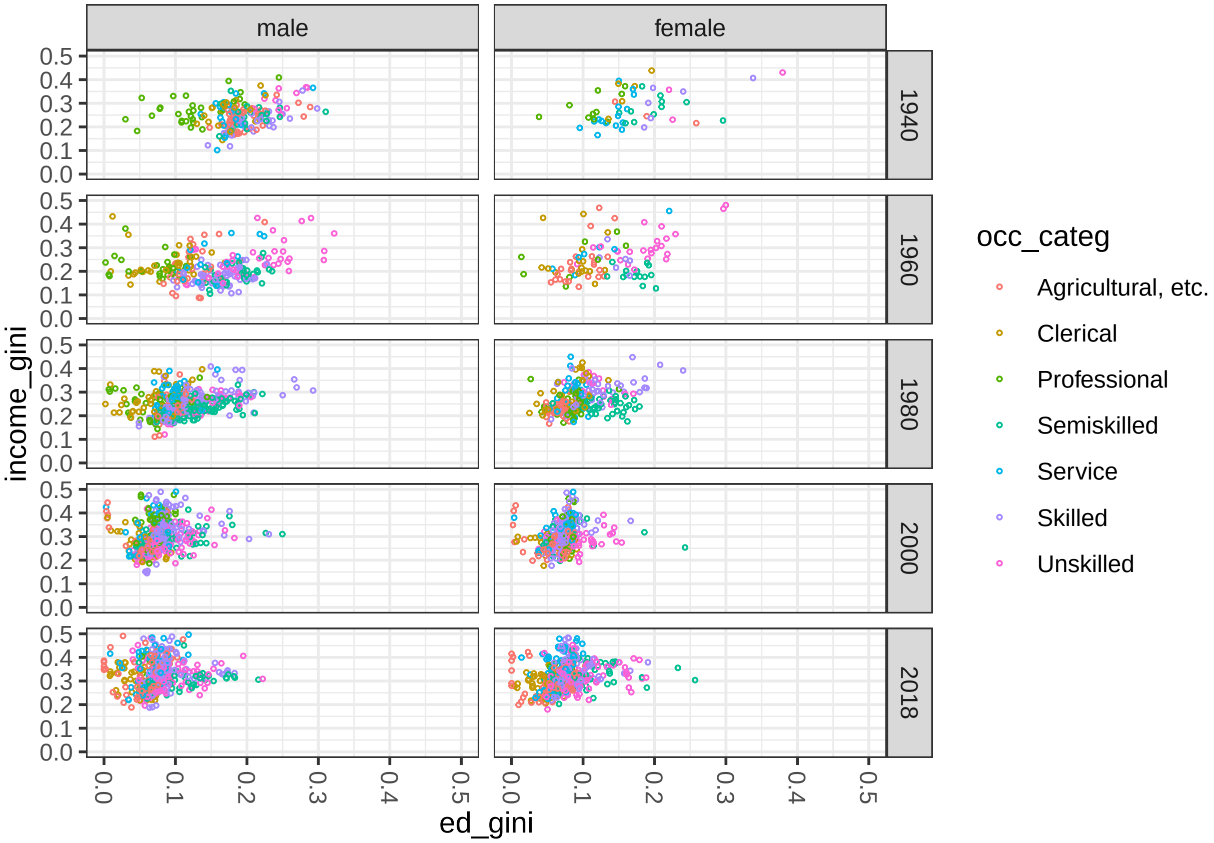
```
acs[, ed_num := as.numeric(as.character(factor(educ, labels = c(0,2.5, 6.5, 9,10,11,12, 13,14,16,18))))]
acs[ed_num <= 11, ed_categ := "Less than High School"]
acs[ed_num > 11 &ed_num < 16, ed_categ := "HS or Some College"]
acs[ed_num >= 16, ed_categ := "College Plus"]

skills_overview4 <- acs[,.(pc1 = weighted.mean(pc1,w = perwt, na.rm = T),
                           pc2 = weighted.mean(pc2,w = perwt,  na.rm = T),
                           pc3 = weighted.mean(pc3,w = perwt,  na.rm = T),
                           pc4 = weighted.mean(pc4,w = perwt,  na.rm = T),
                           programming = weighted.mean(programming,w = perwt,  na.rm = T),
                           tech_skills = weighted.mean(tech_skills, w = perwt, na.rm = T), average_valu

skills_overview4_melt <- melt(skills_overview4, id.vars = c("year", "age_cat", "sex", "ed_categ"))

skills_overview4_melt[, cohort := floor((as.numeric(year) - as.numeric(substr(age_cat,1,2)))/10)*10]

skills_overview4_melt <- skills_overview4_melt[,.(value = mean(value)), by = .(cohort, ed_categ,
                                                                               variable, year, sex)]

ggplot(skills_overview4_melt[variable %like% "pc1|progr|tech|averag"]) +
  geom_line(aes(x = year, y = value, color = as.factor(cohort), group = as.factor(cohort))) +
  facet_grid( variable + sex~ed_categ, scales = "free") +
  scale_color_viridis_d()+
  theme(axis.text.x = element_text(angle = -90, vjust = 0.5, hjust=1))
```



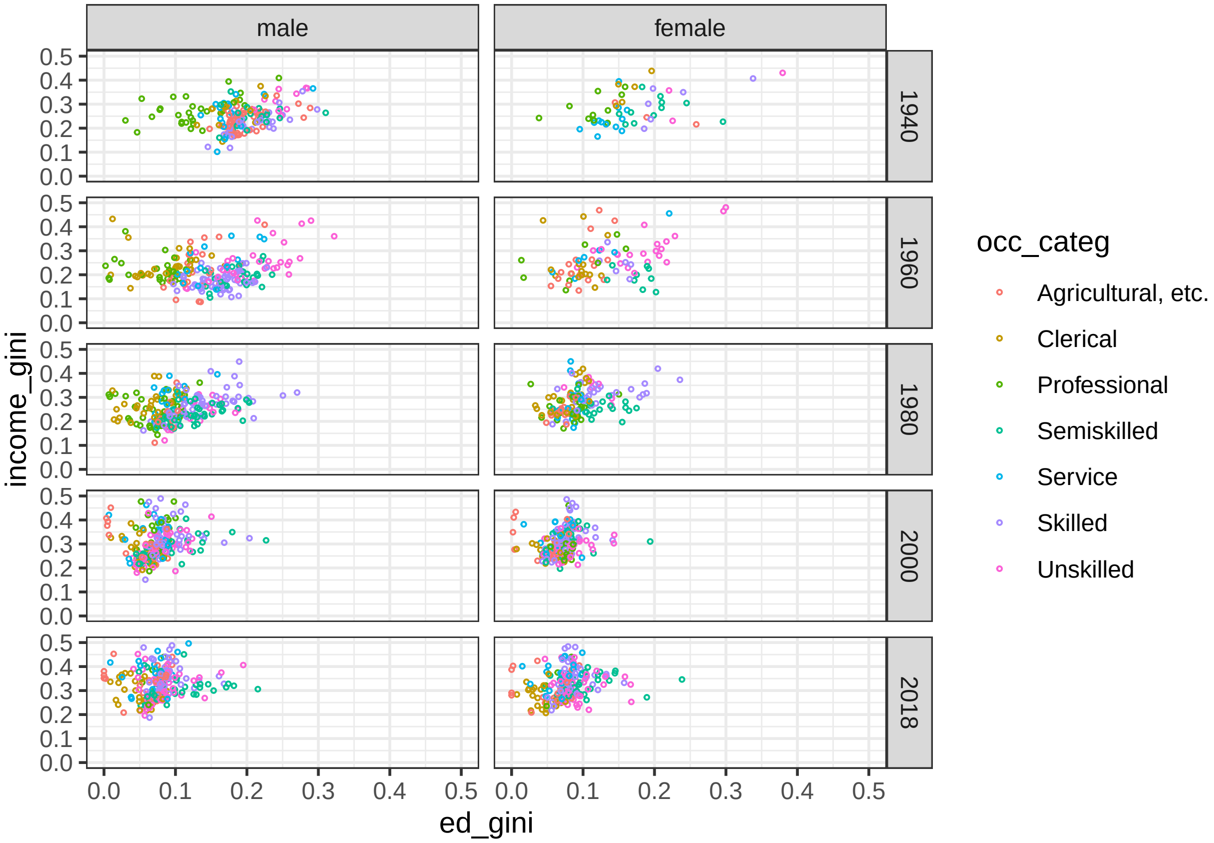
Examine how class of worker, industry, and ed affect inequality (switch to censuses)

```
## function (...)
## .Internal(options(...))
## <bytecode: 0x7fd38f1ccc78>
## <environment: namespace:base>
```

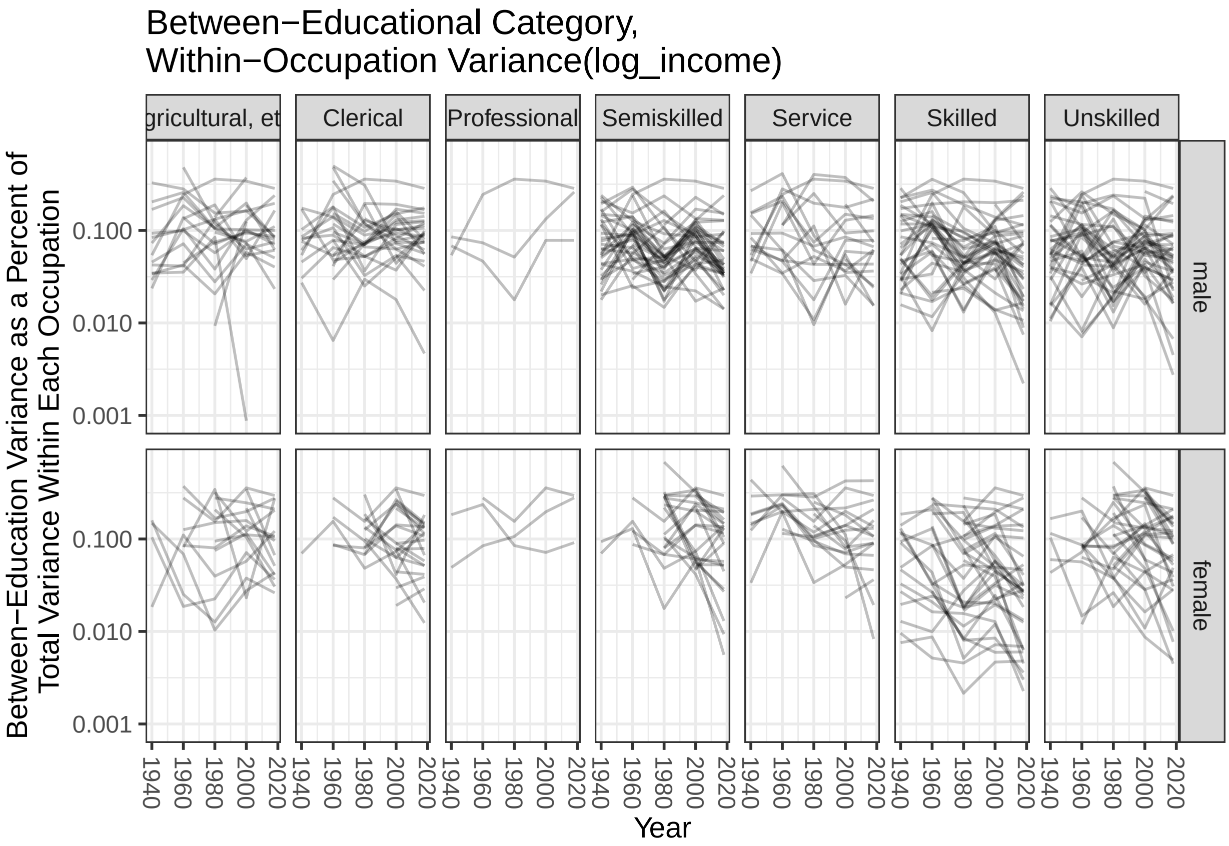
```
## track change in ed diversity across time compared to income diversity
census_1940[, ed_num := as.numeric(as.character(factor(educ, labels = c(0,2.5, 6.5, 9,10,11,12, 13,14,15
inc_ed_ineq <- census_1940[,.(ed_gini = DescTools::Gini(ed_num),
                             income_gini = DescTools::Gini(incwage),
                             N = .N), by = .(occ, year, occ_categ, sex)]

ggplot(inc_ed_ineq[N > 50]) +
  geom_point(aes(x = ed_gini, y = income_gini, color = occ_categ), size = .5, shape = 1) +
  facet_grid(year~sex) +
  xlim(0, .5) +
  ylim(0, .5) +  theme(axis.text.x = element_text(angle = -90, vjust = 0.5, hjust=1))
```



```
## track change in ed diversity across time compared to income diversity
census_1940[, ed_num := as.numeric(as.character(factor(educ, labels = c(0,2.5, 6.5, 9,10,11,12, 13,14,15
inc_ed_ineq <- census_1940[,.(ed_gini = DescTools::Gini(ed_num),
                              income_gini = DescTools::Gini(incwage),
                              N = .N), by = .(origocc1950, year, occ_categ, sex)]

ggplot(inc_ed_ineq[N > 50]) +
  geom_point(aes(x = ed_gini, y = income_gini, color = occ_categ), size = .5, shape = 1) +
  facet_grid(year~sex) +
  xlim(0, .5) +
  ylim(0, .5)
```

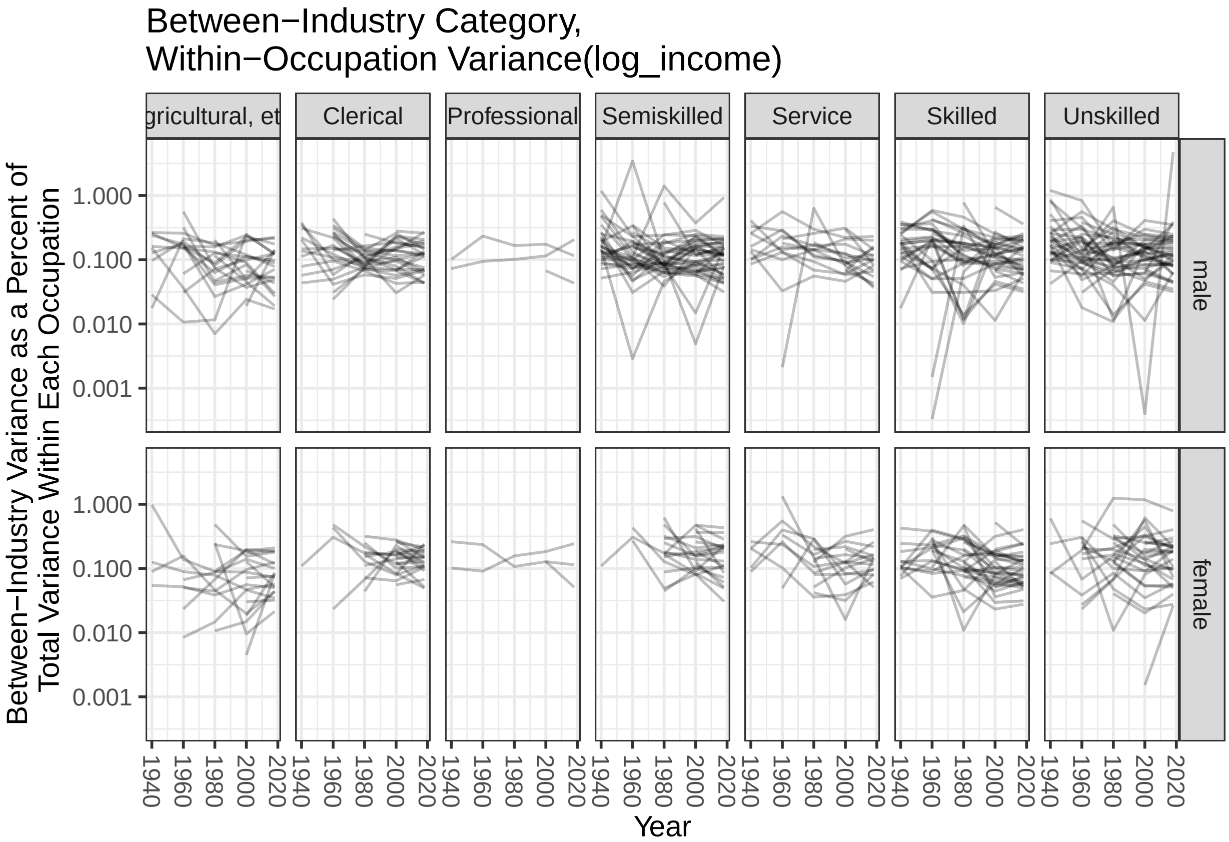


```
census_1940[,N_occ_ed := .N, by = .(origocc1950, educ, year, sex)]
inc_ed_ineq_2 <- census_1940[!is.na(log_incwage) &N_occ_ed > 1,.(within_ed_occ_var = var(log_incwage),
                                                                 N_ed_occ = .N,
                                                                 ed_occ_avg = mean(log_incwage)), by =
  merge(., census_1940[!is.na(log_incwage)&N_occ_ed > 1,.(witihin_occ_var = var(log_incwage),
                                                          N_occ = .N,
                                                          occ_avg = mean(log_incwage)), by = .(origocc1

inc_ed_ineq_2_sum  <- inc_ed_ineq_2[,.(bw_var = weighted.var(ed_occ_avg, N_ed_occ),
                                       wi_var = mean(witihin_occ_var),
                                       occ_avg = mean(occ_avg),
                                       N_occ = mean(N_occ)), by = .(origocc1950, year, sex)]

inc_ed_ineq_2_sum[, bw_perc := bw_var /(wi_var)]

census_1940[year == 2018,.(origocc1950, occ_categ)] %>% unique() %>%
  merge(., inc_ed_ineq_2_sum, by = "origocc1950") -> inc_ed_ineq_2_sum

ggplot(inc_ed_ineq_2_sum[N_occ > 50 & !origocc1950 %like% "nec|n.e.c.|NEC|missing|unknown|NA"]) +
  geom_line(aes(x = as.numeric(year), y = bw_perc, group = (origocc1950)), size = .5, alpha = .25)+
  facet_grid(sex~occ_categ) +
  scale_color_viridis_d() +
  guides(color = F) +
  ylim(0, 1) +
  scale_y_continuous(trans = "log10")+
  labs(title = "Between-Educational Category,\nWithin-Occupation Variance(log_income)",
       x = "Year",
       y = "Between-Education Variance as a Percent of \nTotal Variance Within Each Occupation") +
    theme(axis.text.x = element_text(angle = -90, vjust = 0.5, hjust=1))
```



Repeat for Industry

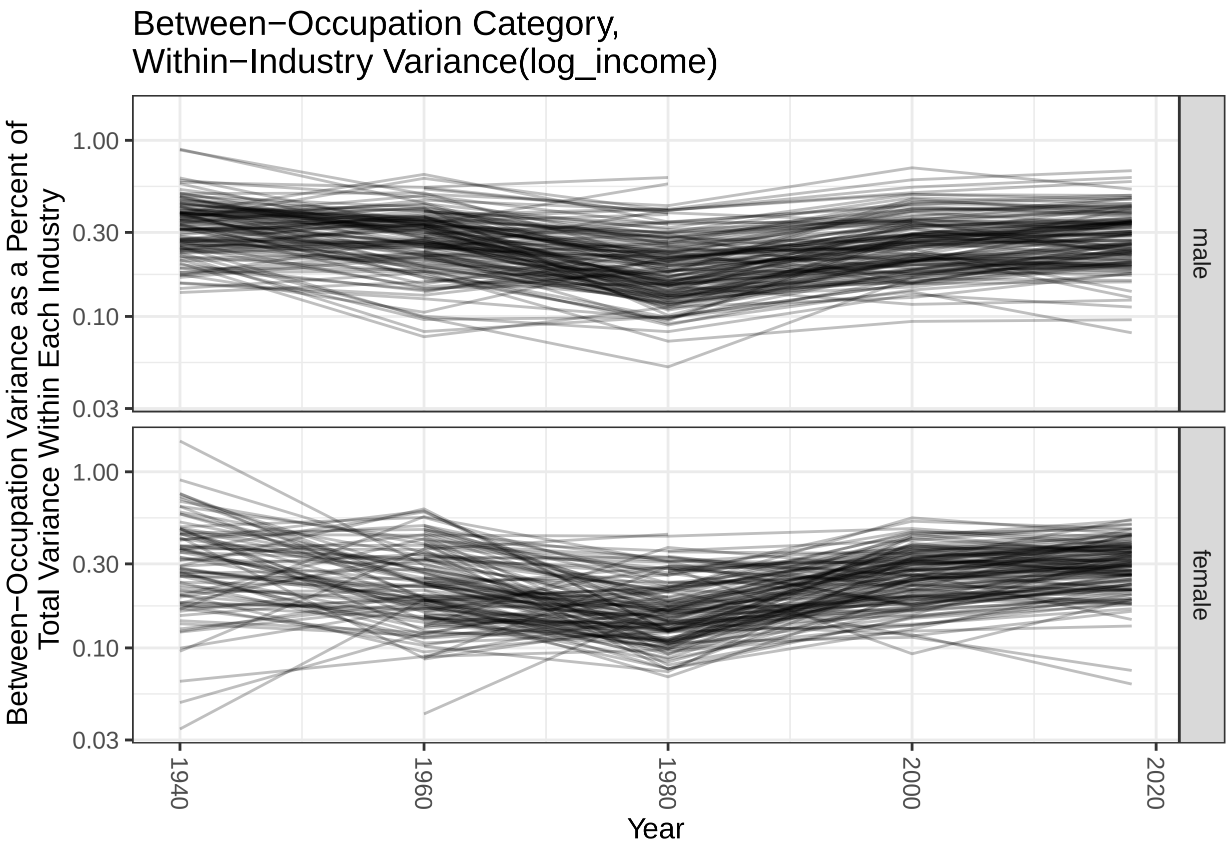
```
census_1940[,N_occ_ind:= .N, by = .(origocc1950, origind1950, year, sex)]
inc_ind_ineq_2 <- census_1940[!is.na(log_incwage) &N_occ_ind > 1,.(within_ind_occ_var = var(log_incwage
                                                                   N_ind_occ = .N,
                                                                   ind_occ_avg = mean(log_incwage)), by
  merge(., census_1940[!is.na(log_incwage)&N_occ_ind > 1,.(witihin_occ_var = var(log_incwage),
                                                           N_occ = .N,
                                                           occ_avg = mean(log_incwage)), by = .(origocc

inc_ind_ineq_2_sum  <- inc_ind_ineq_2[,.(bw_var = weighted.var(ind_occ_avg, N_ind_occ),
                                         wi_var = mean(witihin_occ_var),
                                         occ_avg = mean(occ_avg),
                                         N_occ = mean(N_occ)), by = .(origocc1950, year, sex)]

inc_ind_ineq_2_sum[, bw_perc := bw_var /(wi_var)]

census_1940[year == 2018,.(origocc1950, occ_categ)] %>% unique() %>%
  merge(., inc_ind_ineq_2_sum, by = "origocc1950") -> inc_ind_ineq_2_sum

ggplot(inc_ind_ineq_2_sum[N_occ > 50 & !origocc1950 %like% "nec|n.e.c.|NEC|missing|unknown|NA"]) +
  geom_line(aes(x = as.numeric(year), y = bw_perc, group = (origocc1950)), size = .5, alpha = .25)+
  facet_grid(sex~occ_categ) +
  scale_color_viridis_d() +
  guides(color = F) +
  ylim(0, 1) +
  scale_y_continuous(trans = "log10")+
  labs(title = "Between-Industry Category,\nWithin-Occupation Variance(log_income)",
       x = "Year",
       y = "Between-Industry Variance as a Percent of \nTotal Variance Within Each Occupation")+
    theme(axis.text.x = element_text(angle = -90, vjust = 0.5, hjust=1))
```



Repeat for Industry

```
census_1940[,N_occ_ind:= .N, by = .(origocc1950, origind1950, year, sex)]
inc_ind_ineq_2 <- census_1940[!is.na(log_incwage) &N_occ_ind > 1,.(within_ind_occ_var = var(log_incwage
                                                                   N_ind_occ = .N,
                                                                   ind_occ_avg = mean(log_incwage)), by
  merge(., census_1940[!is.na(log_incwage)&N_occ_ind > 1,.(witihin_occ_var = var(log_incwage),
                                                           N_occ = .N,
                                                           occ_avg = mean(log_incwage)), by = .(origind

inc_ind_ineq_2_sum  <- inc_ind_ineq_2[,.(bw_var = weighted.var(ind_occ_avg, N_ind_occ),
                                         wi_var = mean(witihin_occ_var),
                                         occ_avg = mean(occ_avg),
                                         N_occ = mean(N_occ)), by = .(origind1950, year, sex)]

inc_ind_ineq_2_sum[, bw_perc := bw_var /(wi_var)]
```

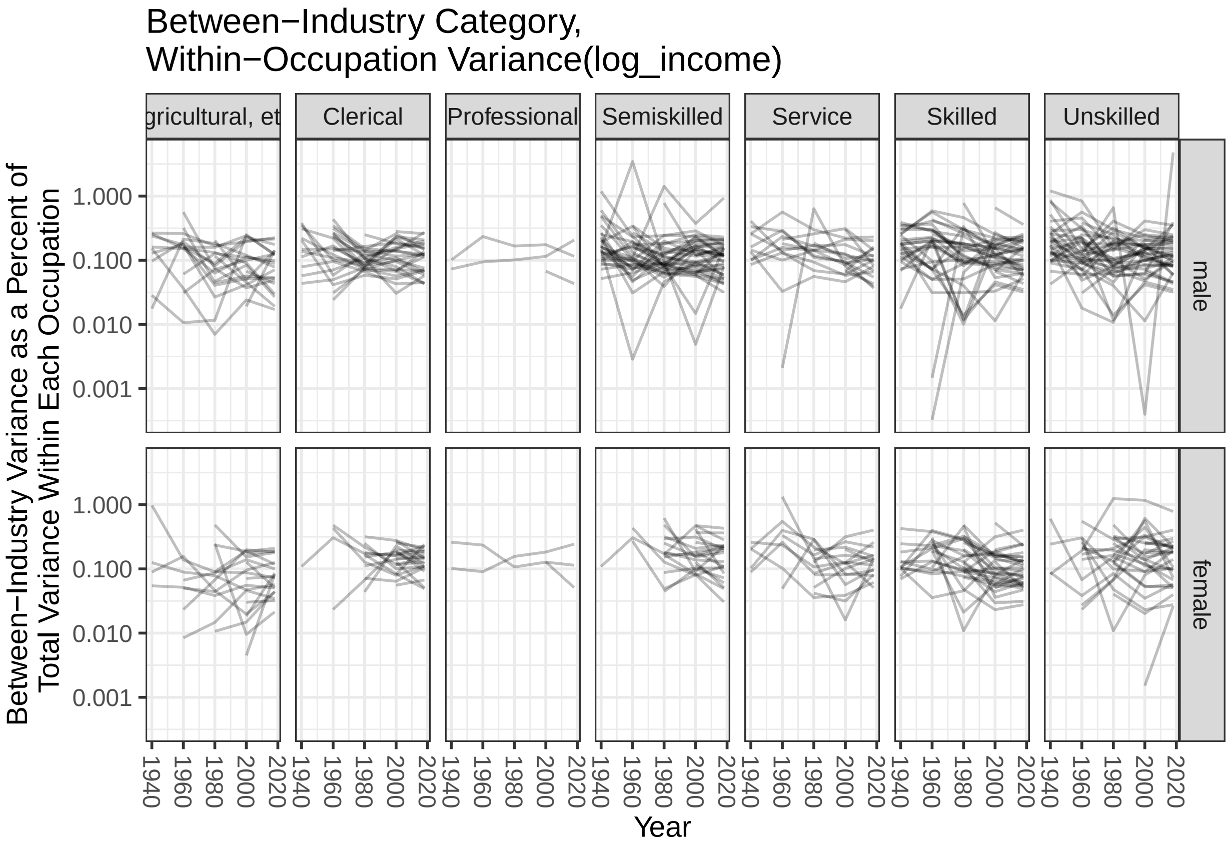
```
ggplot(inc_ind_ineq_2_sum[N_occ > 50 & !origind1950 %like% "nec|n.e.c.|NEC|missing|unknown|NA"]) +
  geom_line(aes(x = as.numeric(year), y = bw_perc, group = (origind1950)), size = .5, alpha = .25)+
  facet_grid(sex~.) +
  scale_color_viridis_d() +
  guides(color = F) +
  ylim(0, 1) +
  scale_y_continuous(trans = "log10")+
  labs(title = "Between-Occupation Category,\nWithin-Industry Variance(log_income)",
       x = "Year",
       y = "Between-Occupation Variance as a Percent of \nTotal Variance Within Each Industry")+
    theme(axis.text.x = element_text(angle = -90, vjust = 0.5, hjust=1))
```



Repeat for Industry

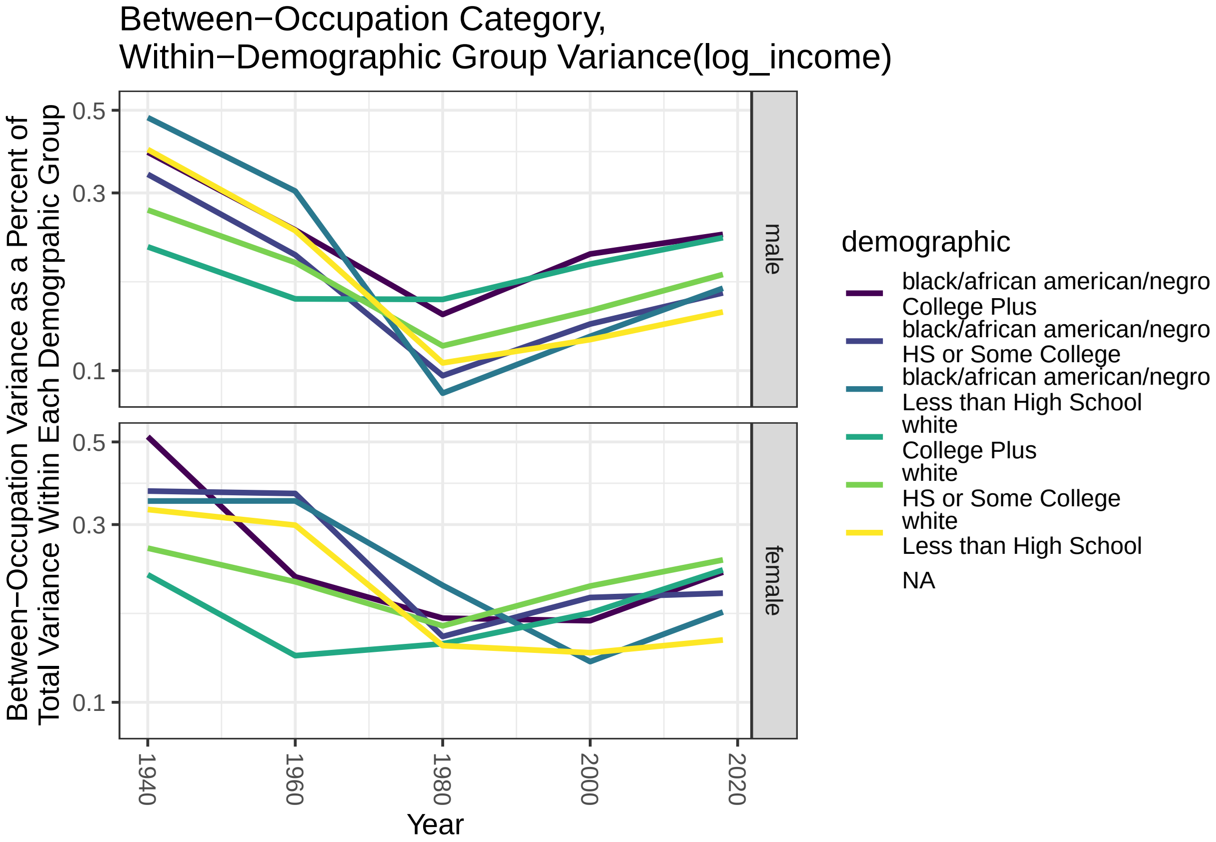
```
census_1940[,N_occ_ind:= .N, by = .(origocc1950, origind1950, year, sex)]
inc_ind_ineq_2 <- census_1940[!is.na(log_incwage) &N_occ_ind > 1,.(within_ind_occ_var = var(log_incwage
                                                                   N_ind_occ = .N,
                                                                   ind_occ_avg = mean(log_incwage)), by
  merge(., census_1940[!is.na(log_incwage)&N_occ_ind > 1,.(witihin_occ_var = var(log_incwage),
                                                           N_occ = .N,
                                                           occ_avg = mean(log_incwage)), by = .(origocc

inc_ind_ineq_2_sum  <- inc_ind_ineq_2[,.(bw_var = weighted.var(ind_occ_avg, N_ind_occ),
                                         wi_var = mean(witihin_occ_var),
                                         occ_avg = mean(occ_avg),
                                         N_occ = mean(N_occ)), by = .(origocc1950, year, sex)]
```

```
inc_ind_ineq_2_sum[, bw_perc := bw_var /(wi_var)]

census_1940[year == 2018,.(origocc1950, occ_categ)] %>% unique() %>%
  merge(., inc_ind_ineq_2_sum, by = "origocc1950") -> inc_ind_ineq_2_sum

ggplot(inc_ind_ineq_2_sum[N_occ > 50 & !origocc1950 %like% "nec|n.e.c.|NEC|missing|unknown|NA"]) +
  geom_line(aes(x = as.numeric(year), y = bw_perc, group = (origocc1950)), size = .5, alpha = .25)+
  facet_grid(sex~occ_categ) +
  scale_color_viridis_d() +
  guides(color = F) +
  ylim(0, 1) +
  scale_y_continuous(trans = "log10")+
  labs(title = "Between-Industry Category,\nWithin-Occupation Variance(log_income)",
       x = "Year",
       y = "Between-Industry Variance as a Percent of \nTotal Variance Within Each Occupation")+
    theme(axis.text.x = element_text(angle = -90, vjust = 0.5, hjust=1))
```



Repeat within demographic group

```
census_1940[ed_num <= 11, ed_categ := "Less than High School"]
census_1940[ed_num > 11 &ed_num < 16, ed_categ := "HS or Some College"]
census_1940[ed_num >= 16, ed_categ := "College Plus"]

census_1940[race %like% "black|white", demographic := paste(race, ed_categ, sep = "\n")]
census_1940[,N_occ_ind:= .N, by = .(origocc1950, demographic, year, sex)]
inc_ind_ineq_2 <- census_1940[!is.na(log_incwage) &N_occ_ind > 1,.(within_ind_occ_var = var(log_incwage
                                                                   N_ind_occ = .N,
                                                                   ind_occ_avg = mean(log_incwage)), by
  merge(., census_1940[!is.na(log_incwage)&N_occ_ind > 1,.(witihin_occ_var = var(log_incwage),
                                                           N_occ = .N,
                                                           occ_avg = mean(log_incwage)), by = .(demogra

inc_ind_ineq_2_sum  <- inc_ind_ineq_2[,.(bw_var = weighted.var(ind_occ_avg, N_ind_occ),
                                         wi_var = mean(witihin_occ_var),
                                         occ_avg = mean(occ_avg),
                                         N_occ = mean(N_occ)), by = .(demographic, year, sex)]

inc_ind_ineq_2_sum[, bw_perc := bw_var /(wi_var)]


ggplot(inc_ind_ineq_2_sum[N_occ > 50]) +
  geom_line(aes(x = as.numeric(year), y = bw_perc, group = (demographic),
                color = demographic), size = 1)+
  facet_grid(sex~.) +
  scale_color_viridis_d() +
  ylim(0, 1) +
  scale_y_continuous(trans = "log10")+
  labs(title = "Between-Occupation Category,\nWithin-Demographic Group Variance(log_income)",
       x = "Year",
       y = "Between-Occupation Variance as a Percent of \nTotal Variance Within Each Demogrpahic Group"
    theme(axis.text.x = element_text(angle = -90, vjust = 0.5, hjust=1))
```


Between−Occupation Category,
Within−Demographic Group Variance(log_income)
Between−Occupation Variance as a Percent of
Total Variance Within Each Demogrpahic Group
male
female
Year
demographic
black/african american/negro
College Plus
black/african american/negro
HS or Some College
black/african american/negro
Less than High School
white
College Plus
white
HS or Some College
white
Less than High School
NA