analysis_III_time_series_abs

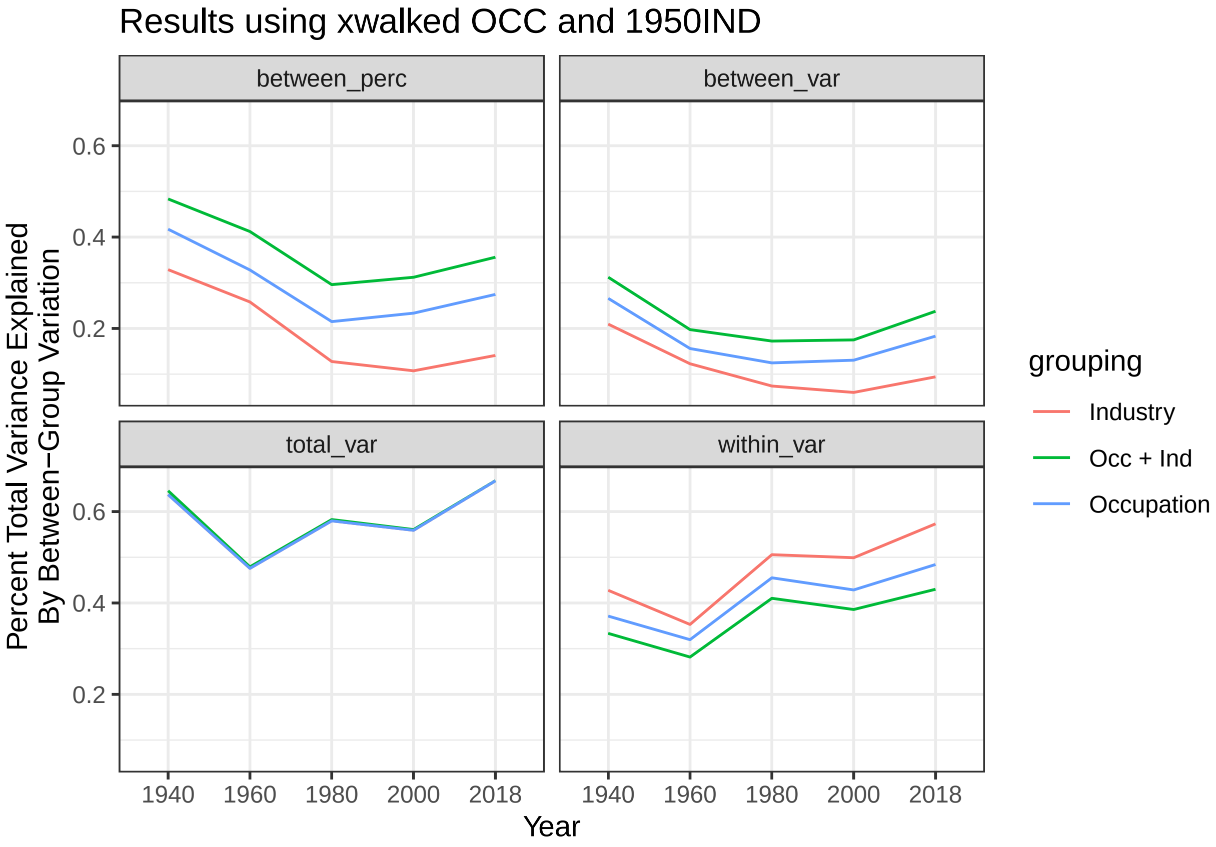
Hunter York

10/26/2020

Crosswalk

```
# load in xwalk
xwalk <- data.table(read_excel("../ref/Census_integrated_occ_crosswalks.xlsx"))
xwalk_long <- melt(xwalk, id.vars = c("OCC1950", "Occupation category description"))
setnames(xwalk_long, c("OCC1950", "OCC1950_desc", "year", "orig_occ"))
xwalk_long[as.character(year) =="ACS 2000-02", year := "2000ACS"]
xwalk_long[as.character(year)  == "ACS 2003-", year := "2018"]

# copy 1950 vals to 1940 for nowxwalk_long
xwalk_long[, year := as.character(year)]
xwalk_long[year == 1950] %>%
  .[, year := 1940] %>%
  rbind(., xwalk_long) -> temp
xwalk_long[, orig_occ := as.numeric(orig_occ)]
census_1940[, occ := as.numeric(occ)]
# merge on census
census_1940 <- merge(census_1940, xwalk_long, by.y = c("year", "orig_occ"), by.x = c("year", "occ"), all

get_vars <- function(c.data, c.by_vars,c.by_vars_2, c.var_interest){
  out_dt_1 <- c.data[,.(w_i_ss = weighted.var(get(c.var_interest), perwt) * .N,
                        N = .N,
                        k = max(.GRP)),
                     by = c(c.by_vars, c.by_vars_2)]
  out_2 <- c.data[,.(tot_ss=weighted.var(get(c.var_interest), perwt) * .N), by = c.by_vars_2]
  out_dt_1 <- merge(out_dt_1, out_2, by = c.by_vars_2)
  out_dt_1 <- out_dt_1[!is.na(tot_ss)& !is.na(w_i_ss) & !is.nan(tot_ss)& !is.nan(w_i_ss) &
                          !is.infinite(tot_ss)& !is.infinite(w_i_ss),
                       .(avg_within_var = sum(w_i_ss),
                         avg_total_var = mean(tot_ss),
                         avg_between_var = mean(tot_ss) -sum(w_i_ss),
                         N = sum(N),
                         k = length(unique(N[!is.na(w_i_ss)]))),
                       by = c.by_vars_2]
  return(out_dt_1)
}

dem_var_gettr2 <- function(c.dat2, c.by_vars_2){
  occ_only <- get_vars(c.dat2,
                       c.by_vars = c("origocc1950"),
                       c.by_vars_2 = c.by_vars_2,
                       c.var_interest = "log_incwage")
  occ_only[, grouping := "Occupation"]
  ind_only <- get_vars(c.dat2,
                       c.by_vars = c("origind1950"),
                       c.by_vars_2 = c.by_vars_2,

                       c.var_interest = "log_incwage")
    ind_only[, grouping := "Industry"]

  occ_ind <- get_vars(c.dat2,
                      c.by_vars = c("origocc1950", "origind1950"),
                      c.by_vars_2 = c.by_vars_2,
                      c.var_interest = "log_incwage")
    occ_ind[, grouping := "Occ + Ind"]


  out_dt <- rbindlist(list(occ_only, ind_only, occ_ind ))
  return(out_dt)
}


temp <- dem_var_gettr2(census_1940, c.by_vars_2 = c("year"))


plot_dt2 <- temp

plot_dt2[, f_stat := (avg_between_var/(k-1))/((avg_within_var)/(N-k))]
plot_dt2[, ms_bw := (avg_between_var/(k-1))]
plot_dt2[, ms_wi := (avg_within_var/(N-k))]

plot_dt2[, within_perc :=
          avg_within_var/
          (avg_within_var+avg_between_var)]
plot_dt2[, between_perc :=
          avg_between_var/
          (avg_within_var+avg_between_var)]
plot_dt2[, bw_wi_perc_ratio := between_perc/within_perc]
plot_dt2[, total_var := avg_total_var/N]
plot_dt2[, within_var := avg_within_var/N]
plot_dt2[, between_var := avg_between_var/N]


# cast long
plot_dt2_long <- melt(plot_dt2, id.vars = c("year",
                                          "grouping"),
                     measure.vars = c("within_perc",
                                      "between_perc",
                                      "between_var",
                                      "bw_wi_perc_ratio",
                                      "total_var",
                                      "within_var",
                                      "ms_wi",
                                      "ms_bw"))
```

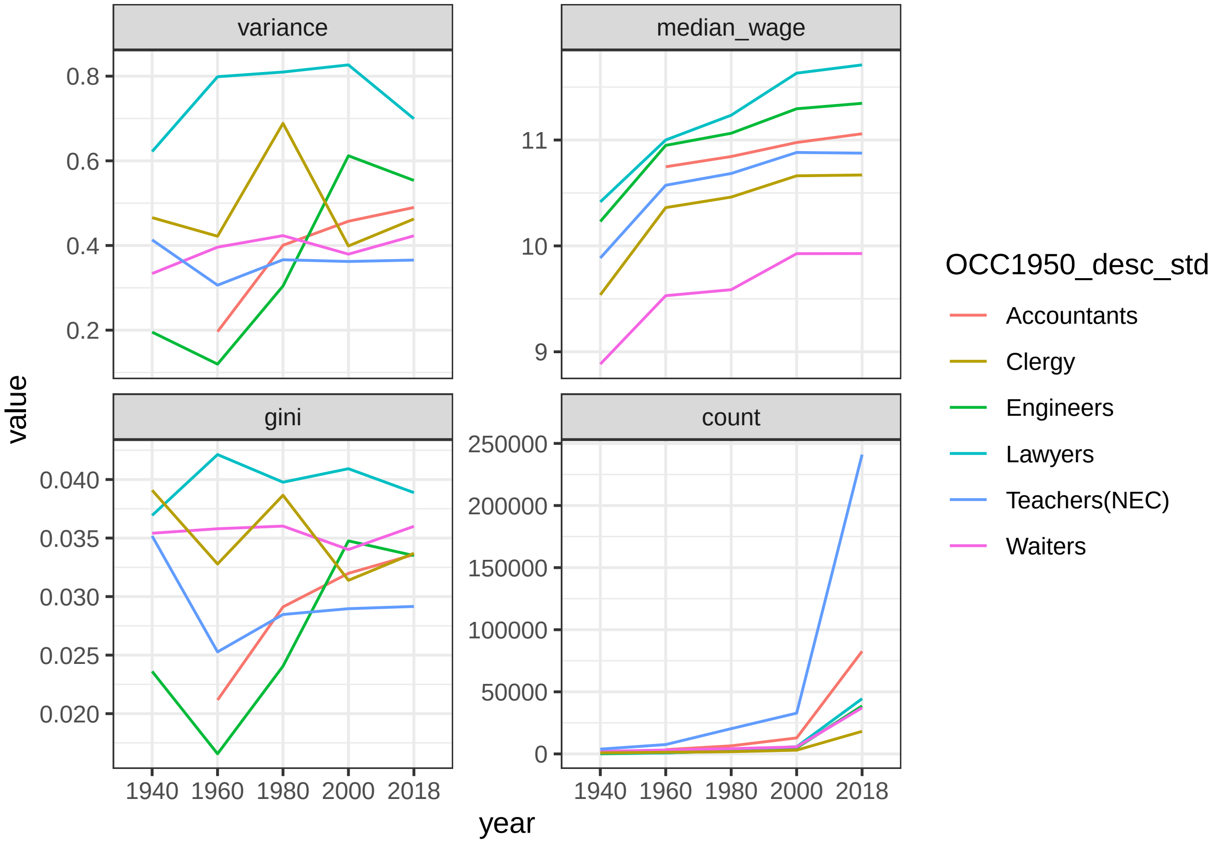
```r
gg2 <- ggplot(plot_dt2_long[variable %like% "between_perc|total_var|within_var|between_var"]) +
  geom_line(aes(x = year, y = value,
                color = grouping, group = grouping))+
  facet_wrap(~variable)+
  labs(x = "Year", y = "Percent Total Variance Explained\nBy Between-Group Variation",
       title = "Results using xwalked OCC and 1950IND"
       )
print(gg2)
```



Task 2 - Use some specific occupations to examine speciic trajectories

I'm using six relatively stable single occupational designations from the 1950 codings: teachers, industrial engineers, waiters/waitresses, accountants/auditors, lawyers and judges, clergymen.

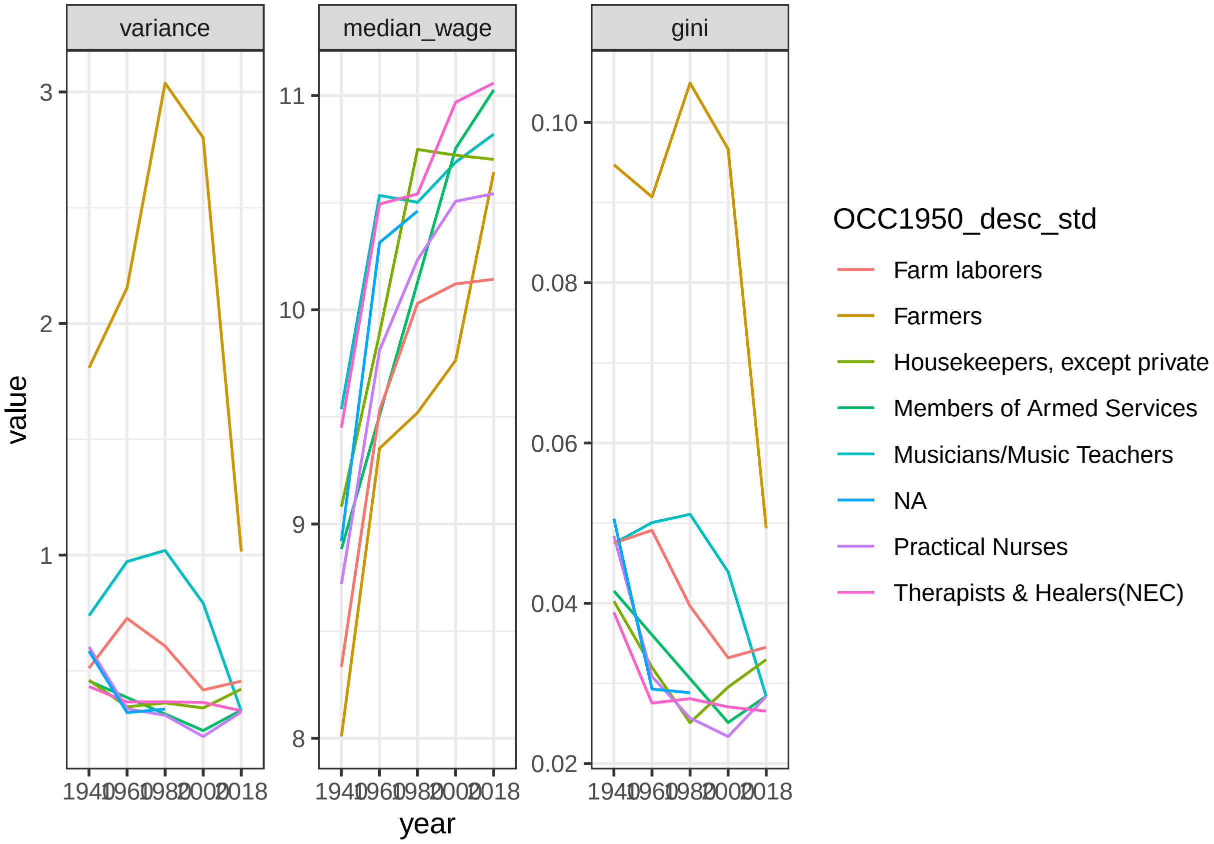
```r
census_subset <- census_1940[OCC1950 %in% c(0, 784, 55, 45, 9, 93) & !is.na(origocc1950)]

# calculate variance by job, year
vars_by_job_year <- census_subset[,.(variance = var(log_incwage),
                                     median_wage = median(log_incwage),
                                     gini = DescTools::Gini(log_incwage),
                                     count = .N), by = .(OCC1950, year)]

# merge on descs
temp <- data.table(OCC1950 =as.character(c(0, 784, 55, 45, 9, 93)), OCC1950_desc_std = c("Accountants",
vars_by_job_year <- merge(vars_by_job_year, temp)

# plot
vars_by_job_year %>% melt(., id.vars = c("OCC1950_desc_std", "OCC1950", "year")) %>%
ggplot(.) +
  geom_line(aes(x = year, y = value, group = OCC1950, color = OCC1950_desc_std)) +
  facet_wrap(~variable, scales = "free_y")
```



Task 3 - do the above analysis for all occupations and see if there are any major outliers in terms of increased or decrease equality

Increased equality

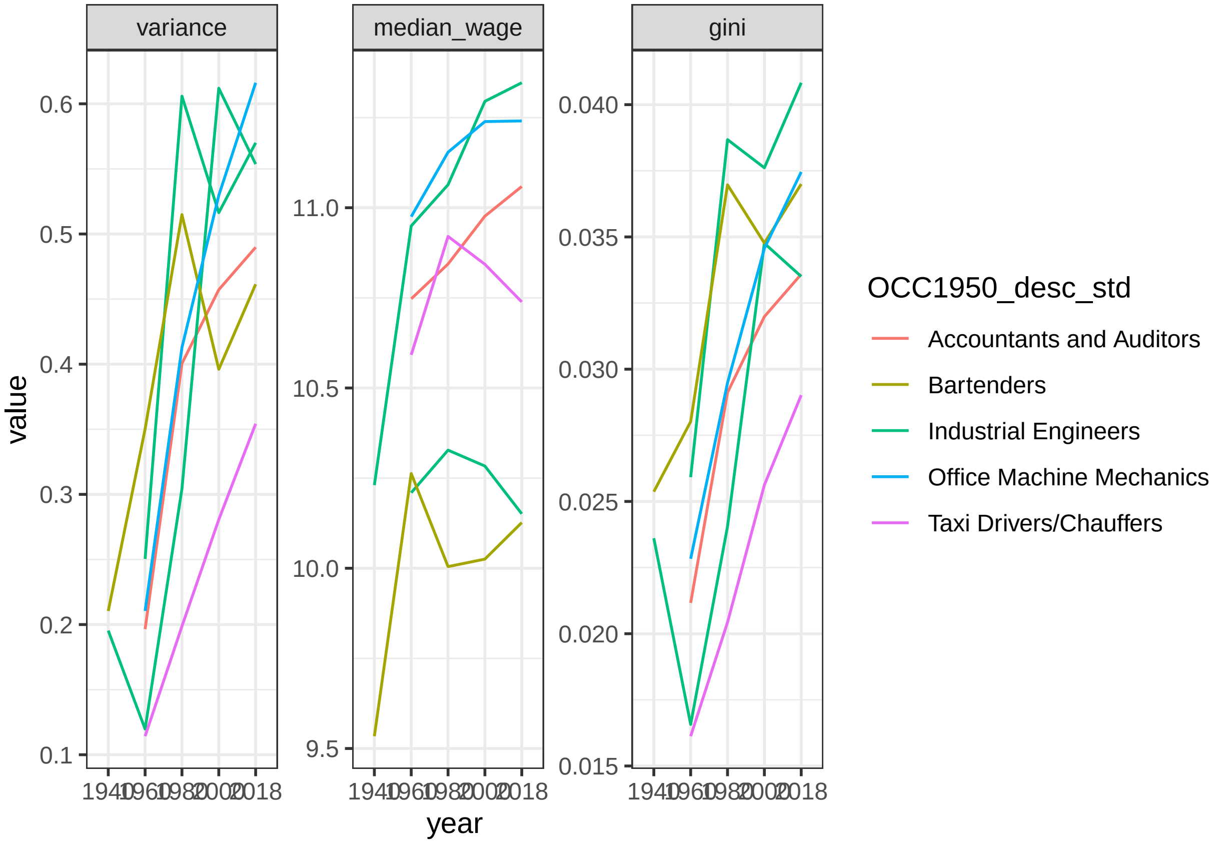
```
# calculate variance by job, year
vars_by_job_year <- census_1940[,.(variance = var(log_incwage),
                                   median_wage = median(log_incwage),
                                   gini = DescTools::Gini(log_incwage),
                                   count = .N), by = .(OCC1950, year)]

vars_by_job_year[,gini_diff := max(gini) - min(gini), by = .(OCC1950)]
vars_by_job_year[, median_count := median(count[year <= 2000]), by = .(OCC1950)]

# merge on descs
temp <- data.table(OCC1950 =as.character(c(57, 97, 100, 595, 764, 781, 820, 999)), OCC1950_desc_std = c
vars_by_job_year <- merge(vars_by_job_year, temp)

# plot
vars_by_job_year %>% melt(., id.vars = c("OCC1950_desc_std", "OCC1950", "year")) %>%
  .[variable %like%"var|wage|^gini$"] %>%
ggplot(.) +
  geom_line(aes(x = year, y = value, group = OCC1950, color = OCC1950_desc_std)) +
  facet_wrap(~variable, scales = "free_y")
```



Decreased equality

```
# calculate variance by job, year
vars_by_job_year <- census_1940[,.(variance = var(log_incwage),
                                   median_wage = median(log_incwage),
                                   gini = DescTools::Gini(log_incwage),
                                   count = .N), by = .(OCC1950, year)]

vars_by_job_year[,gini_diff := gini[year == min(year)] - gini[year == max(year)], by = .(OCC1950)]
vars_by_job_year[, median_count := median(count[year <= 2000]), by = .(OCC1950)]
vars_by_job_year <- vars_by_job_year[gini_diff < -0 & median_count > 500]

# merge on descs
temp <- data.table(OCC1950 =as.character(c(45, 750, 0, 81, 551, 682)), OCC1950_desc_std = c("Industrial

vars_by_job_year <- merge(vars_by_job_year, temp)
```

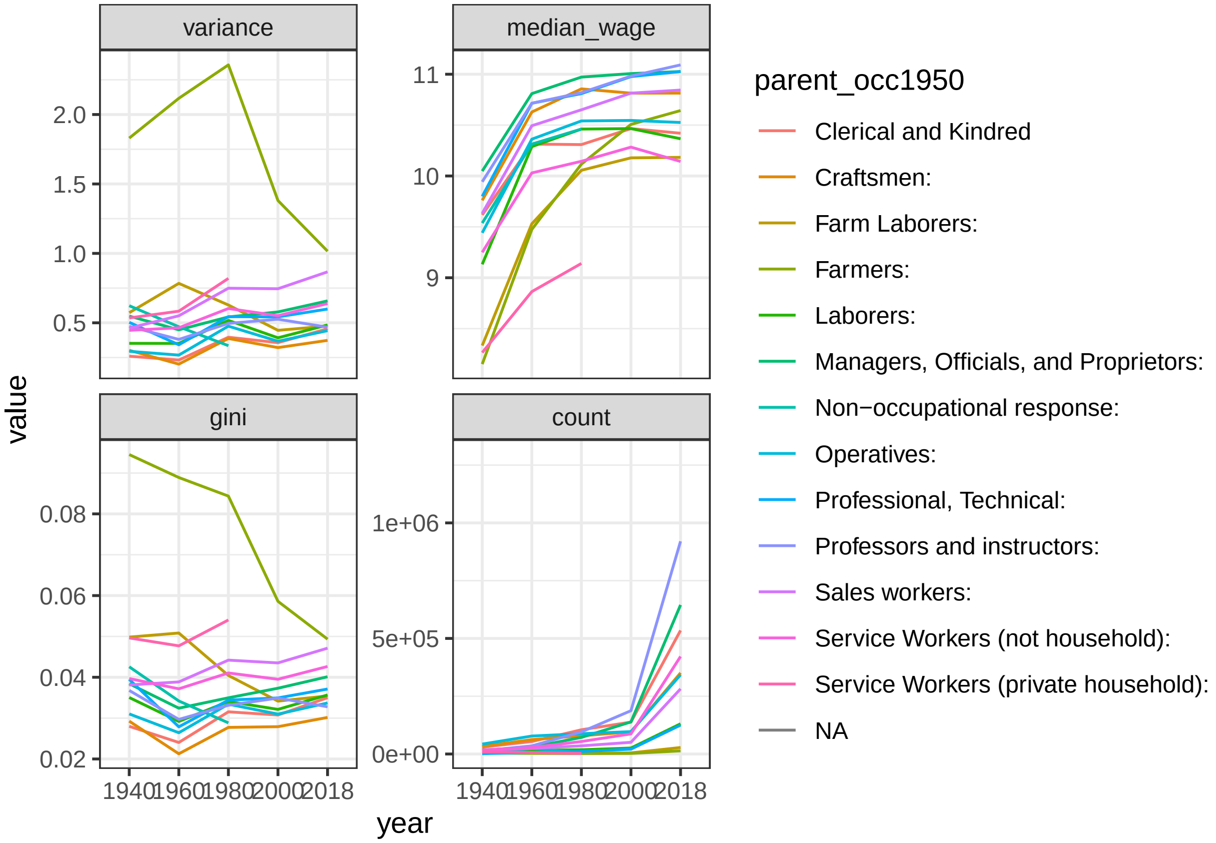
```r
# plot
vars_by_job_year %>% melt(., id.vars = c("OCC1950_desc_std", "OCC1950", "year")) %>%
  .[variable %like%"var|wage|^gini$"] %>%
ggplot(.) +
  geom_line(aes(x = year, y = value, group = OCC1950, color = OCC1950_desc_std)) +
  facet_wrap(~variable, scales = "free_y")
```



Aggregate up to less granular job titles using census hierarchy and repeat

```r
# create crosswalk for this
base_xwalk <- read_excel("../ref/Census_integrated_occ_crosswalks.xlsx") %>% data.table()
base_xwalk <- base_xwalk[,.(OCC1950, `Occupation category description`)]
base_xwalk[OCC1950 == "#",parent_occ1950 := `Occupation category description`]
for(i in 2:nrow(base_xwalk)){
  if (is.na(base_xwalk[i, parent_occ1950])){
    base_xwalk[i, parent_occ1950 := base_xwalk[i-1, parent_occ1950]]
  }
}
census_1940 <- merge(census_1940, unique(base_xwalk[OCC1950 != "#",.(OCC1950, parent_occ1950)]), by = "


# calculate variance by job, year
vars_by_job_year <- census_1940[,.(variance = var(log_incwage),
                                   median_wage = median(log_incwage),
                                   gini = DescTools::Gini(log_incwage),
                                   count = .N), by = .(parent_occ1950, year)]
```

```r
# plot
vars_by_job_year %>% melt(., id.vars = c("parent_occ1950", "year")) %>%
ggplot(.) +
  geom_line(aes(x = year, y = value, group = parent_occ1950, color = parent_occ1950)) +
  facet_wrap(~variable, scales = "free_y")
```

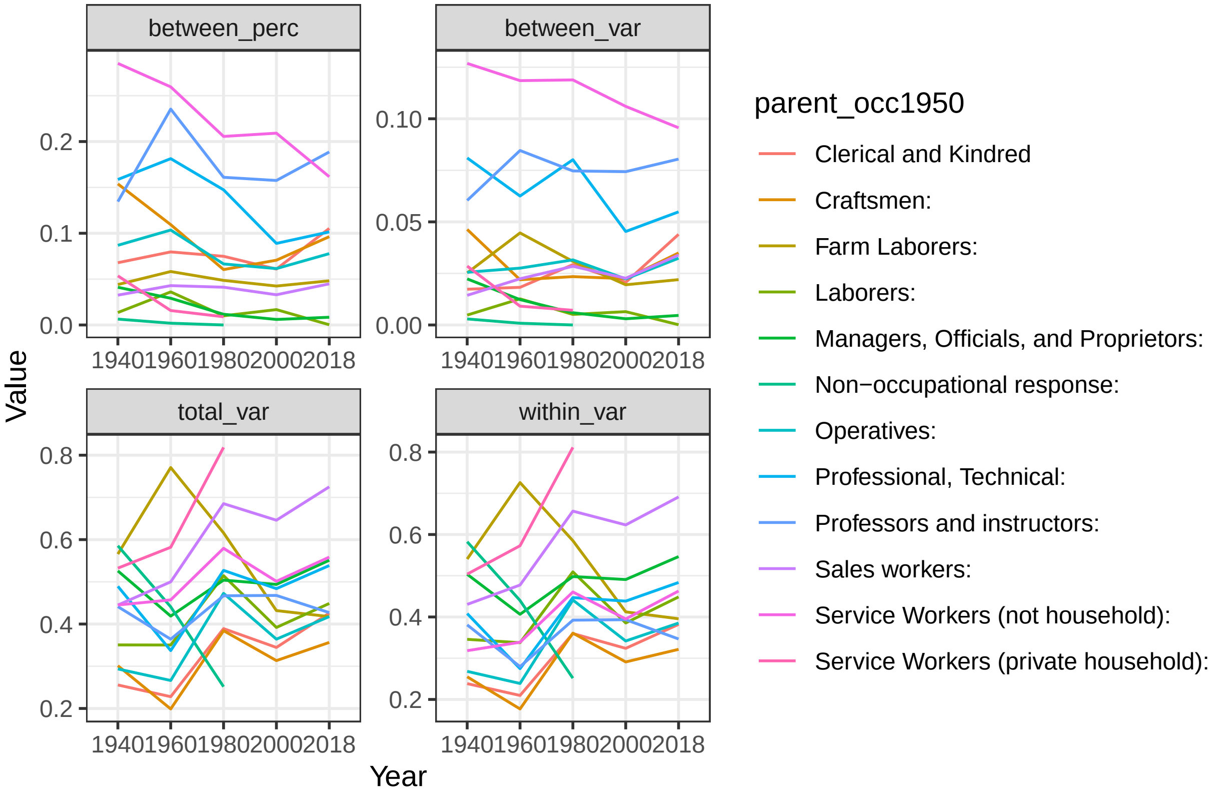


Characterize within and between occupation heterogeneity within each of these groupings

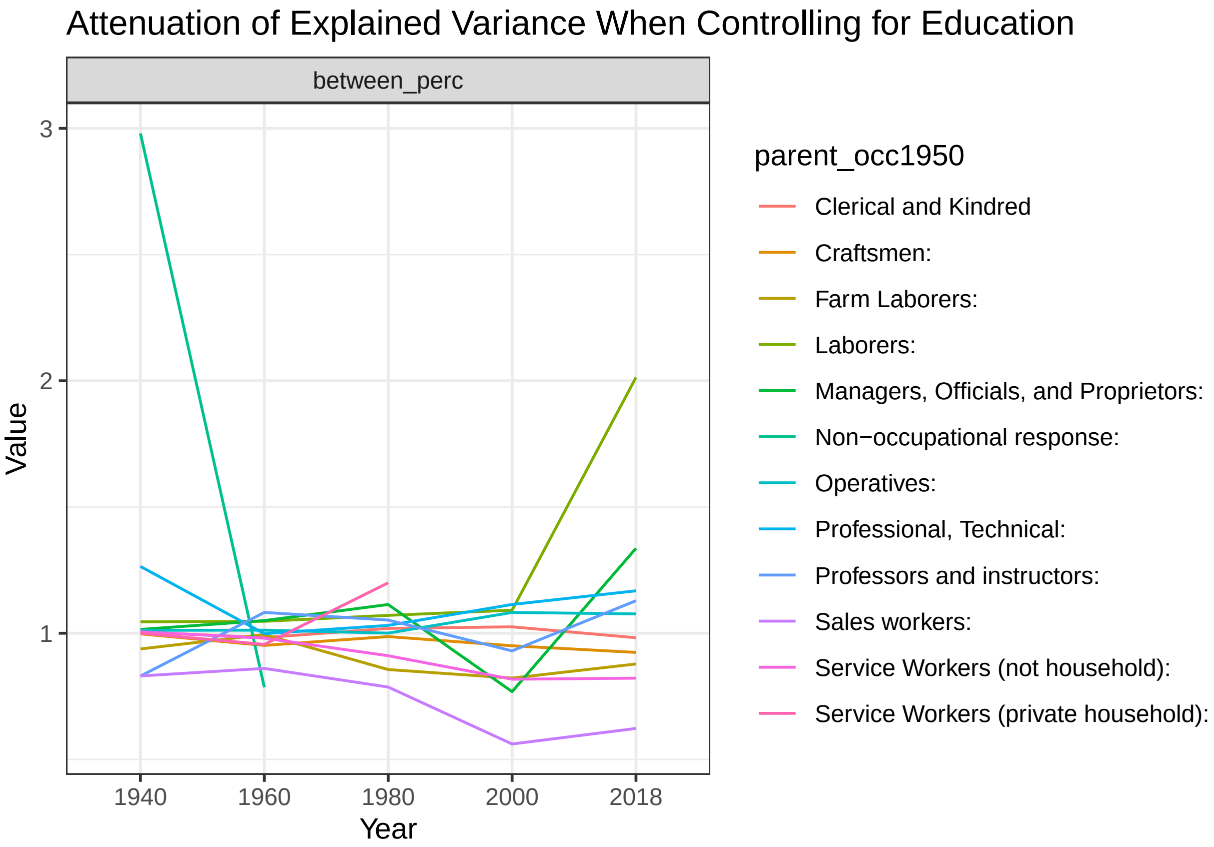
```r
temp <- dem_var_gettr2(census_1940, c.by_vars_2 = c("year", "parent_occ1950"))


plot_dt2 <- temp

plot_dt2[, f_stat := (avg_between_var/(k-1))/((avg_within_var)/(N-k))]
plot_dt2[, ms_bw := (avg_between_var/(k-1))]
plot_dt2[, ms_wi := (avg_within_var/(N-k))]

plot_dt2[, within_perc :=
          avg_within_var/
          (avg_within_var+avg_between_var)]
plot_dt2[, between_perc :=
          avg_between_var/
          (avg_within_var+avg_between_var)]
plot_dt2[, bw_wi_perc_ratio := between_perc/within_perc]
plot_dt2[, total_var := avg_total_var/N]
plot_dt2[, within_var := avg_within_var/N]
plot_dt2[, between_var := avg_between_var/N]


# cast long
plot_dt2_long <- melt(plot_dt2, id.vars = c("year",
                                          "grouping",
                                          "parent_occ1950"),
                     measure.vars = c("within_perc",
                                      "between_perc",
                                      "between_var",
                                      "bw_wi_perc_ratio",
                                      "total_var",
                                      "within_var",
                                      "ms_wi",
                                      "ms_bw"))

gg2 <- ggplot(plot_dt2_long[variable %like% "between_perc|total_var|within_var|between_var" & grouping
  geom_line(aes(x = year, y = value,
                color = parent_occ1950, group = parent_occ1950))+
  facet_wrap(~variable, scales = "free")+
  labs(x = "Year", y = "Value",
       title = "Heterogeneity Explained Within Census Category by Occupation"
       )
```

Repeat by education

```
census_1940[as.numeric(educ) %in% 0:7, ed := "Less than High School"]
census_1940[as.numeric(educ)  %in% 8:10, ed := "Some College"]
census_1940[as.numeric(educ)  %in% 11, ed := "4-Year Degree"]
census_1940[as.numeric(educ)  %in% 12, ed := "Post-Bachelors"]
census_1940[, ed := factor(ed, levels = c("Less than High School","Some College","4-Year Degree", "Post
)]


temp <- dem_var_gettr2(census_1940, c.by_vars_2 = c("year", "parent_occ1950", "ed"))


plot_dt2 <- temp

plot_dt2[, f_stat := (avg_between_var/(k-1))/((avg_within_var)/(N-k))]
plot_dt2[, ms_bw := (avg_between_var/(k-1))]
plot_dt2[, ms_wi := (avg_within_var/(N-k))]

plot_dt2[, within_perc :=
          avg_within_var/
          (avg_within_var+avg_between_var)]
plot_dt2[, between_perc :=
          avg_between_var/
          (avg_within_var+avg_between_var)]
plot_dt2[, bw_wi_perc_ratio := between_perc/within_perc]
plot_dt2[, total_var := avg_total_var/N]
plot_dt2[, within_var := avg_within_var/N]
plot_dt2[, between_var := avg_between_var/N]


# cast long
plot_dt3_long <- melt(plot_dt2, id.vars = c("year",
                                            "grouping",
                                            "parent_occ1950",
                                            "ed",
                                            "N"),
                     measure.vars = c("within_perc",
                                      "between_perc",
                                      "between_var",
                                      "bw_wi_perc_ratio",
                                      "total_var",
                                      "within_var",
                                      "ms_wi",
                                      "ms_bw"))

plot_dt3_long <- plot_dt3_long[,.(value = weighted.mean(value, N)), by = .(year, grouping, parent_occ195

#merge on dataset that didn't control for ed
setnames(plot_dt2_long, "value", "value_uncontrolled")
plot_dt3_long <- merge(plot_dt3_long, plot_dt2_long)
plot_dt3_long[, value_ratio := value/value_uncontrolled]


gg3 <- ggplot(plot_dt3_long[variable %like% "between_perc|total_var|within_var|between_var" & grouping
  geom_line(aes(x = year, y = value,
                color = parent_occ1950, group = parent_occ1950))+
  facet_wrap(~variable, scales = "free")+
  labs(x = "Year", y = "Value",
       title = ""
       )
print(gg3)
```

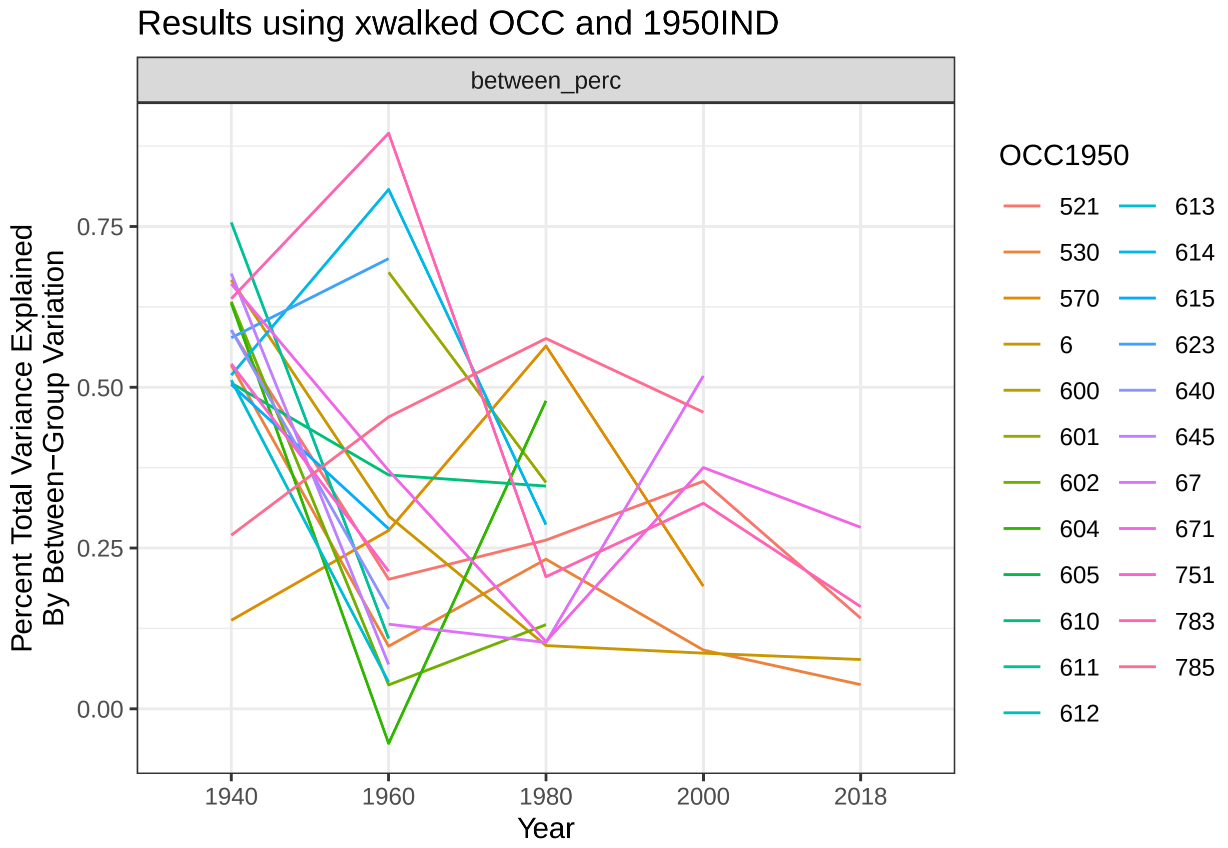


```
gg4 <- ggplot(plot_dt3_long[variable %like% "between_perc" & grouping == "Occupation" & parent_occ1950
    geom_line(aes(x = year, y = value_ratio,
                  color = parent_occ1950, group = parent_occ1950))+
  facet_wrap(~variable, scales = "free")+
  labs(x = "Year", y = "Value",
       title = "Attenuation of Explained Variance When Controlling for Education"
       )
print(gg4)
```



See which occupations adding industry helps explain variation

```
temp <- dem_var_gettr2(census_1940, c.by_vars_2 = c("year", "OCC1950"))


plot_dt2 <- temp

plot_dt2[, f_stat := (avg_between_var/(k-1))/((avg_within_var)/(N-k))]
plot_dt2[, ms_bw := (avg_between_var/(k-1))]
plot_dt2[, ms_wi := (avg_within_var/(N-k))]

plot_dt2[, within_perc :=
          avg_within_var/
          (avg_within_var+avg_between_var)]
plot_dt2[, between_perc :=
          avg_between_var/
          (avg_within_var+avg_between_var)]
plot_dt2[, bw_wi_perc_ratio := between_perc/within_perc]
plot_dt2[, total_var := avg_total_var/N]
plot_dt2[, within_var := avg_within_var/N]
plot_dt2[, between_var := avg_between_var/N]


# cast long
plot_dt2_long <- melt(plot_dt2, id.vars = c("year",
                                            "grouping",
                                          "OCC1950"),
                     measure.vars = c("within_perc",
                                      "between_perc",
                                      "between_var",
                                      "bw_wi_perc_ratio",
                                      "total_var",
                                      "within_var",
                                      "ms_wi",
                                      "ms_bw"))
plot_dt2_long[variable %like% "between_perc" & grouping == "Industry" & value > .5, unique(OCC1950)] ->
gg2 <- ggplot(plot_dt2_long[variable %like% "between_perc" & grouping == "Industry" & OCC1950 %in% temp]
  geom_line(aes(x = year, y = value,
                color = OCC1950, group = OCC1950))+
  facet_wrap(~variable)+
  labs(x = "Year", y = "Percent Total Variance Explained\nBy Between-Group Variation",
       title = "Results using xwalked OCC and 1950IND"
       )

print(gg2)
```



```
census_1940[OCC1950 %in% temp, .(OCC1950, OCC1950_desc)] %>% unique()
```

```
##     OCC1950                                                   OCC1950_desc
##  1:     521                               Engravers, except photoengravers
##  2:     530                                                       Glaziers
##  3:     570                         Pattern and model makers, except paper
##  4:     570                                                           <NA>
##  5:       6                                                        Authors
##  6:     600                                      Apprentice auto mechanics
##  7:     601                              Apprentice bricklayers and masons
##  8:     602                                          Apprentice carpenters
##  9:     604                           Apprentice machinists and toolmakers
## 10:     604                                                           <NA>
## 11:     605                              Apprentice mechanics, except auto
## 12:     610                           Apprentice plumbers and pipe fitters
## 13:     611                          Apprentices, building trades (n.e.c.)
## 14:     612                       Apprentices, metalworking trades (n.e.c.)
## 15:     612                                                           <NA>
## 16:     613                                   Apprentices, printing trades
## 17:     614                            Apprentices, other specified trades
## 18:     615                               Apprentices, trade not specified
## 19:     623                            Boatmen, canalmen, and lock keepers
## 20:     640 Fruit, nut, and vegetable graders, and packers, except factory
## 21:     645                                                      Milliners
## 22:      67                                                 Mathematicians
## 23:     671                                   Photographic process workers
## 24:     751                                                     Bootblacks
## 25:     783                               Ushers, recreation and amusement
## 26:     785                         Watchmen (crossing) and bridge tenders
## 27:     785                                                           <NA>
##     OCC1950                                                   OCC1950_desc
```