analysis_ii_time_series

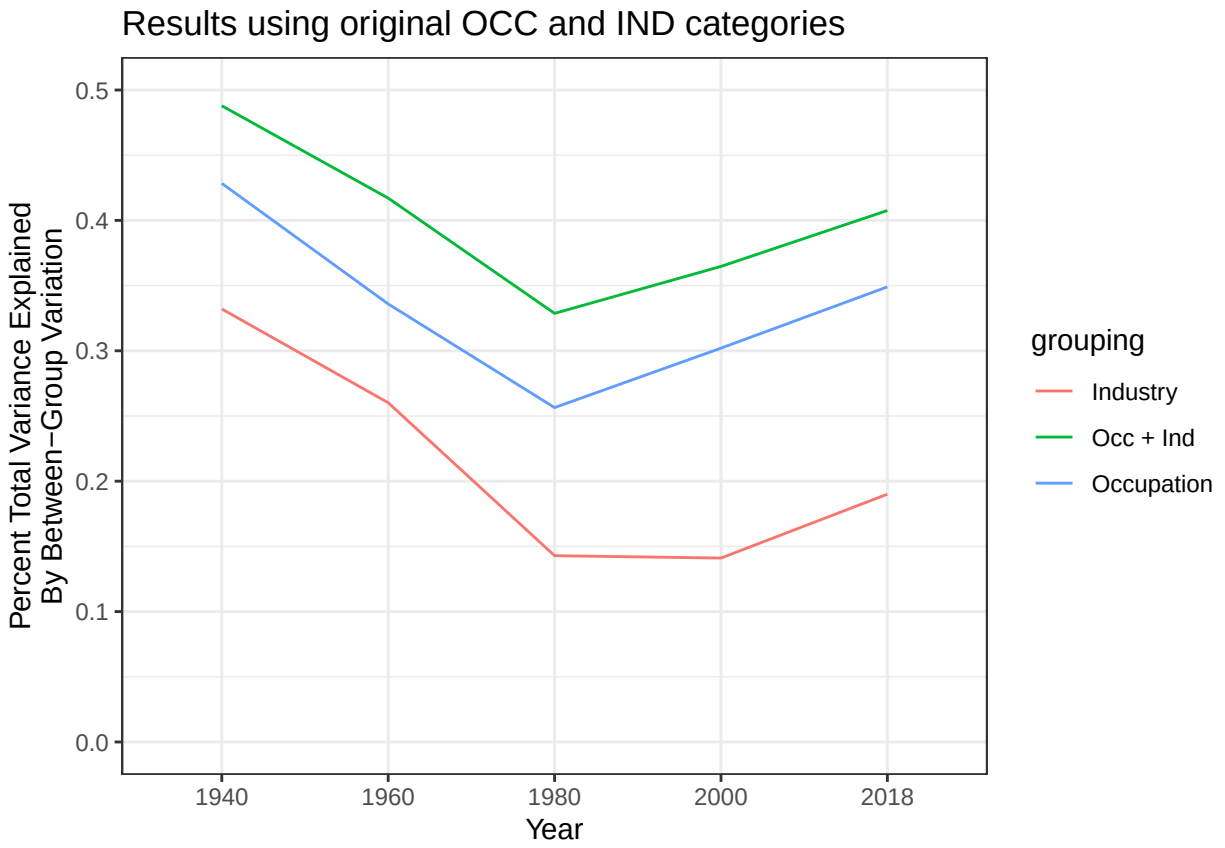
Hunter York

10/10/2020

Time Trends: Decomposition of Variation in Log Earnings across 80 Years

```
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)
}

#  create another function to loop over data and
#  calculate occ, ind, and occ + ind var

dem_var_gettr <- function(c.dat2, c.by_vars_2){
  occ_only <- get_vars(c.dat2,
                       c.by_vars = c("occ"),
                       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("ind"),
                       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("occ", "ind"),
                      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_gettr(census_1940, c.by_vars_2 = c("year"))


plot_dt <- temp

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

plot_dt[, within_perc :=
          avg_within_var/
          (avg_within_var+avg_between_var)]
plot_dt[, between_perc :=
          avg_between_var/
          (avg_within_var+avg_between_var)]
plot_dt[, bw_wi_perc_ratio := between_perc/within_perc]

# ggplot(plot_dt)+
#   geom_line(aes(x = age_start, y = f_stat, color = grouping)) +
#   facet_grid(urban~sex) +
#   geom_hline(yintercept = 1, linetype = "dashed")

# cast long
plot_dt_long <- melt(plot_dt, id.vars = c("year",
                                          "grouping"),
                     measure.vars = c("within_perc",
                                      "between_perc",
                                      "bw_wi_perc_ratio",
                                      "ms_wi",
                                      "ms_bw"))

gg1 <- ggplot(plot_dt_long[variable %like% "between_perc"]) +
  geom_line(aes(x = year, y = value,
                color = grouping, group = grouping))+
  labs(x = "Year", y = "Percent Total Variance Explained\nBy Between-Group Variation",
       title = "Results using original OCC and IND categories") +
  ylim(0, .5)
print(gg1)
```




```
#
```

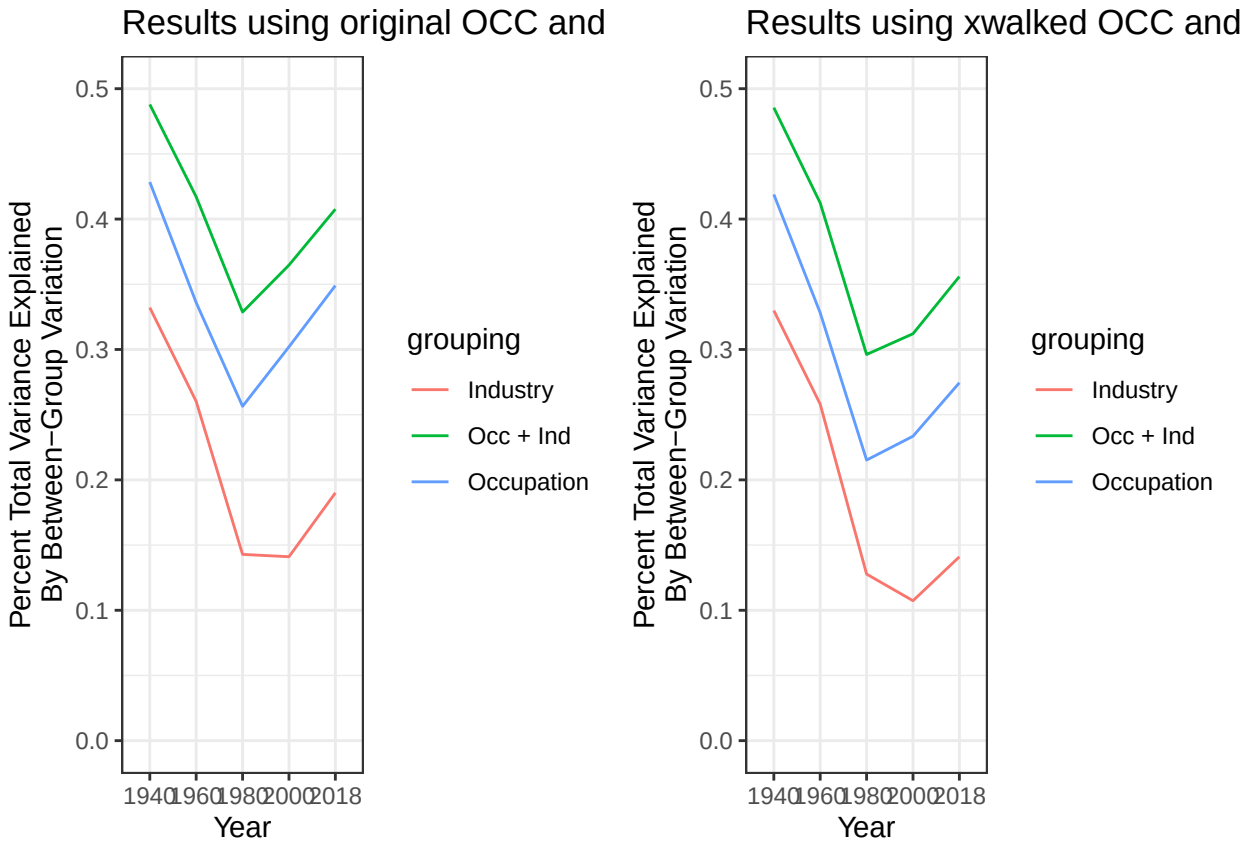
Now do it with a standardized industry variable and with standardized occupation variable

This uses census-to-census crosswalks. ACS 2018 xwalk values seem to be off. (50 should be 51 or 52 for managers, for instance). https://usa.ipums.org/usa/volii/occ2018.shtml

```
# 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"), al
```

```r
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]


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

gg2 <- ggplot(plot_dt2_long[variable %like% "between_perc"]) +
  geom_line(aes(x = year, y = value,
```

```
                color = grouping, group = grouping))+
  labs(x = "Year", y = "Percent Total Variance Explained\nBy Between-Group Variation",
       title = "Results using xwalked OCC and 1950IND"
       ) +
  ylim(0, .5)

library(gridExtra)
grid.arrange(gg1, gg2, nrow = 1)
```




```
#
```

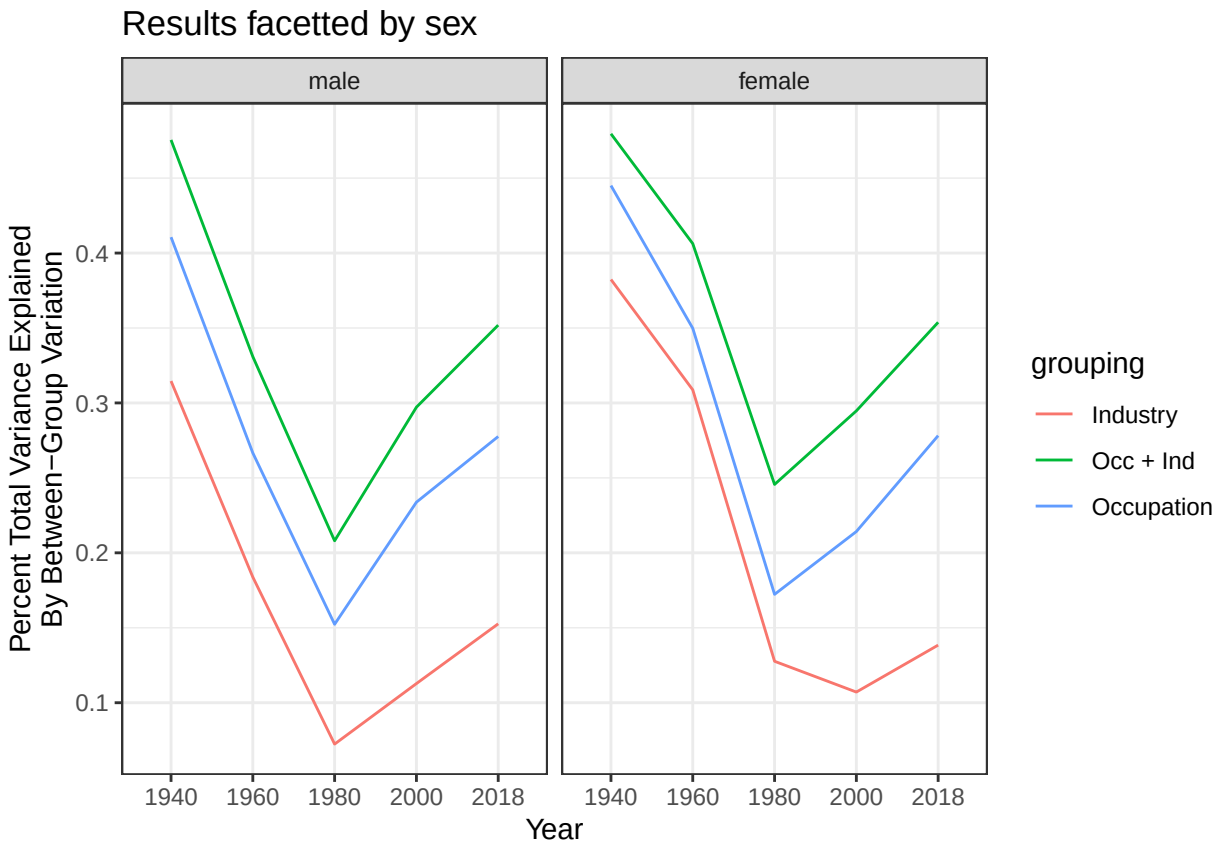
Tables of numbers of occ and ind code by decade

```
census_1940[,.(occ = length(unique(occ)),
               ind = length(unique(ind)),
               xwalk_occ_1950 = length(unique(OCC1950)),
               census_ind_1950 = length(unique(origind1950)),
               census_occ_1950 = length(unique(origocc1950))), by = .(year)]
```

```
##    year occ ind xwalk_occ_1950 census_ind_1950 census_occ_1950
## 1: 1940 228 133            215             121             215
## 2: 1960 294 151            267             145             267
## 3: 1980 504 232            220             143             220
## 4: 2000 509 264            187             134             187
## 5: 2018 529 269            156             131             174
```

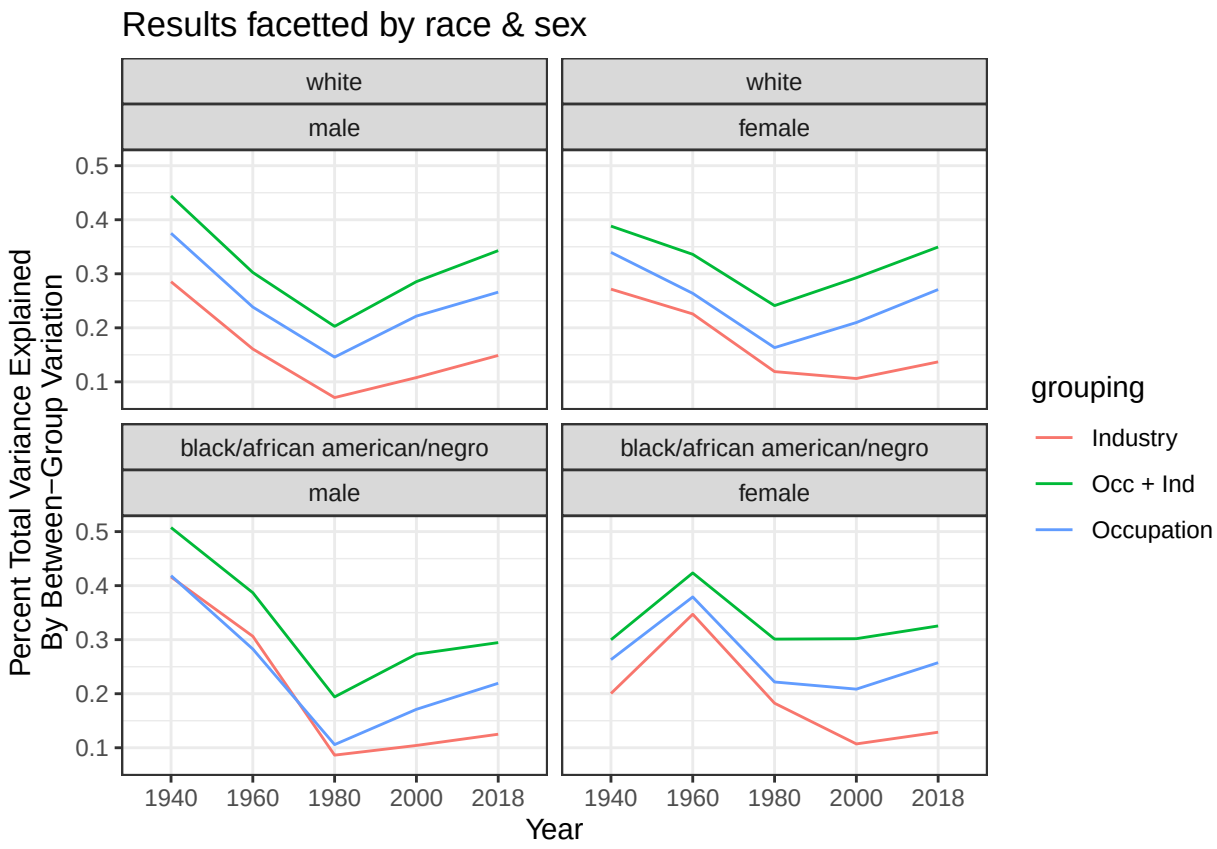
Subanalyses by sex

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


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]


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

gg3 <- ggplot(plot_dt2_long[variable %like% "between_perc" ]) +
  geom_line(aes(x = year, y = value,
                color = grouping, group = grouping))+
  facet_wrap(~sex)+
  labs(x = "Year", y = "Percent Total Variance Explained\nBy Between-Group Variation",
       title = "Results facetted by sex")

print(gg3)
```




```
temp <- dem_var_gettr2(census_1940[race %like% "white|black"], c.by_vars_2 = c("year", "sex", "race"))


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]


# cast long
plot_dt3_long <- melt(plot_dt2, id.vars = c("year",
                                            "sex",
                                          "grouping",
                                          "race"),
                      measure.vars = c("within_perc",
                                       "between_perc",
                                       "bw_wi_perc_ratio",
                                       "ms_wi",
```

```
                                        "ms_bw"))

ggplot(plot_dt3_long[variable %like% "between_perc" ]) +
  geom_line(aes(x = year, y = value,
                color = grouping, group = grouping))+
  facet_wrap(race~sex)+
  labs(x = "Year", y = "Percent Total Variance Explained\nBy Between-Group Variation",
       title = "Results facetted by race & sex")
```



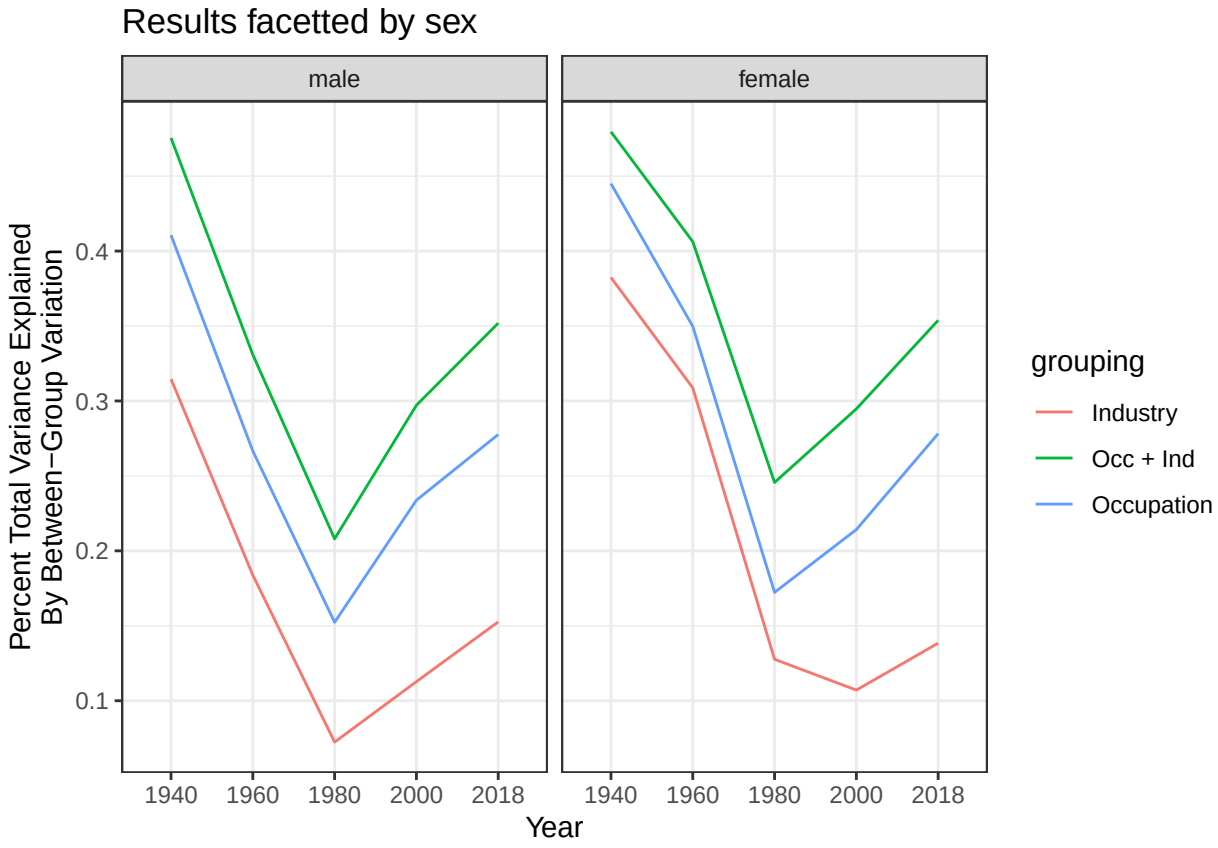
Explore the extent to which changes in job bins affects results

Subset to the least common denominator for all years

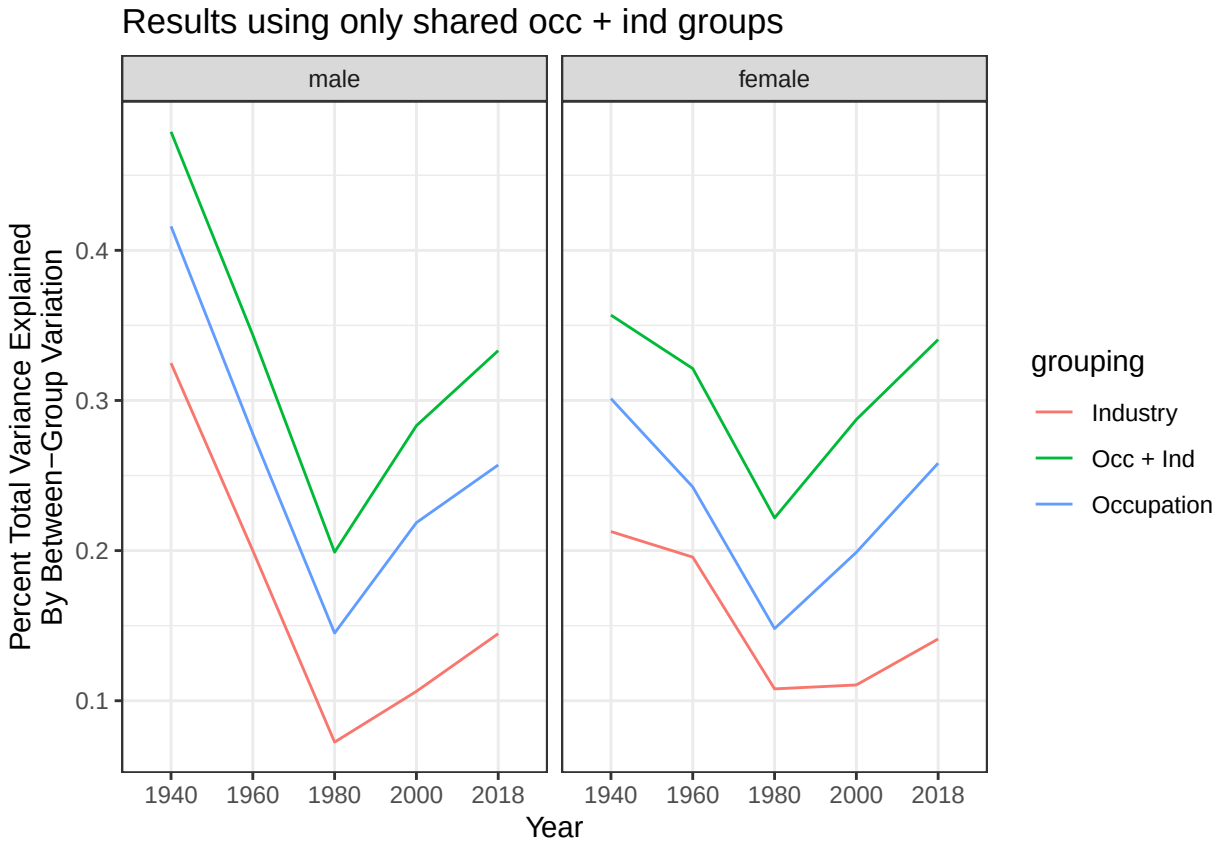
```
common_occ <- Reduce(intersect, list(unique(census_1940[year == 1940, origocc1950]),
                        unique(census_1940[year == 1960, origocc1950]),
                        unique(census_1940[year == 1980, origocc1950]),
                        unique(census_1940[year == 2000, origocc1950]),
                        unique(census_1940[year == 2018, origocc1950])))

common_ind <- Reduce(intersect, list(unique(census_1940[year == 1940, origind1950]),
                        unique(census_1940[year == 1960, origind1950]),
                        unique(census_1940[year == 1980, origind1950]),
                        unique(census_1940[year == 2000, origind1950]),
                        unique(census_1940[year == 2018, origind1950])))
```

```
temp <- dem_var_gettr2(census_1940[origocc1950 %in% common_occ &
                                     origind1950 %in% common_ind], c.by_vars_2 = c("year", "sex"))


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]


# cast long
plot_dt3_long <- melt(plot_dt2, id.vars = c("year",
                                            "sex",
                                           "grouping"
                                           ),
                      measure.vars = c("within_perc",
                                       "between_perc",
                                       "bw_wi_perc_ratio",
                                       "ms_wi",
                                       "ms_bw"))

gg4 <- ggplot(plot_dt3_long[variable %like% "between_perc" ]) +
  geom_line(aes(x = year, y = value,
                color = grouping, group = grouping))+
  facet_wrap(~sex)+
  labs(x = "Year", y = "Percent Total Variance Explained\nBy Between-Group Variation",
       title = "Results using only shared occ + ind groups")
```

Original (1950 Standard Occupation Codings, all jobs)



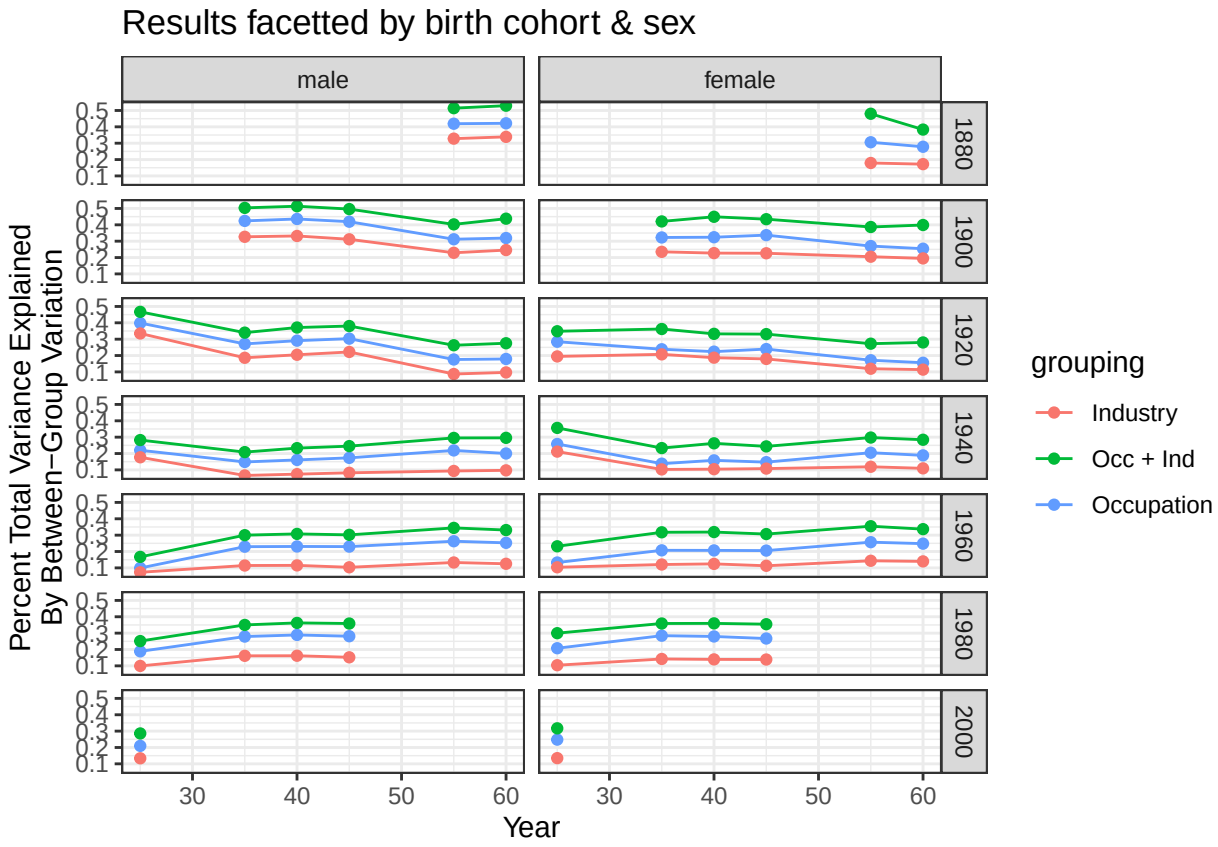
Subset (1950 Standard Occupation Codings, common jobs to all years)



Visualize Cohort Effects, using data with common jobs only

```
temp <- dem_var_gettr2(census_1940[origocc1950 %in% common_occ &
                                   origind1950 %in% common_ind], c.by_vars_2 = c("year", "sex", "age_


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]


# cast long
plot_dt3_long <- melt(plot_dt2, id.vars = c("year",
                                            "sex",
                                          "grouping",
```

```
                                        "age_cat"),
                    measure.vars = c("within_perc",
                                     "between_perc",
                                     "bw_wi_perc_ratio",
                                     "ms_wi",
                                     "ms_bw"))

plot_dt3_long[, birth_cohort :=  round(((round(as.numeric(as.character(year))/10) * 10) - as.numeric(su

ggplot(plot_dt3_long[variable %like% "between_perc" & birth_cohort %in% seq(1800,2000,20)]) +
  geom_line(aes(x = as.numeric(substr(age_cat,1,2)), y = value,
                color = grouping, group = grouping))+
    geom_point(aes(x = as.numeric(substr(age_cat,1,2)), y = value,
                color = grouping, group = grouping))+
  facet_grid(birth_cohort~sex)+
  labs(x = "Year", y = "Percent Total Variance Explained\nBy Between-Group Variation",
       title = "Results facetted by birth cohort & sex")
```

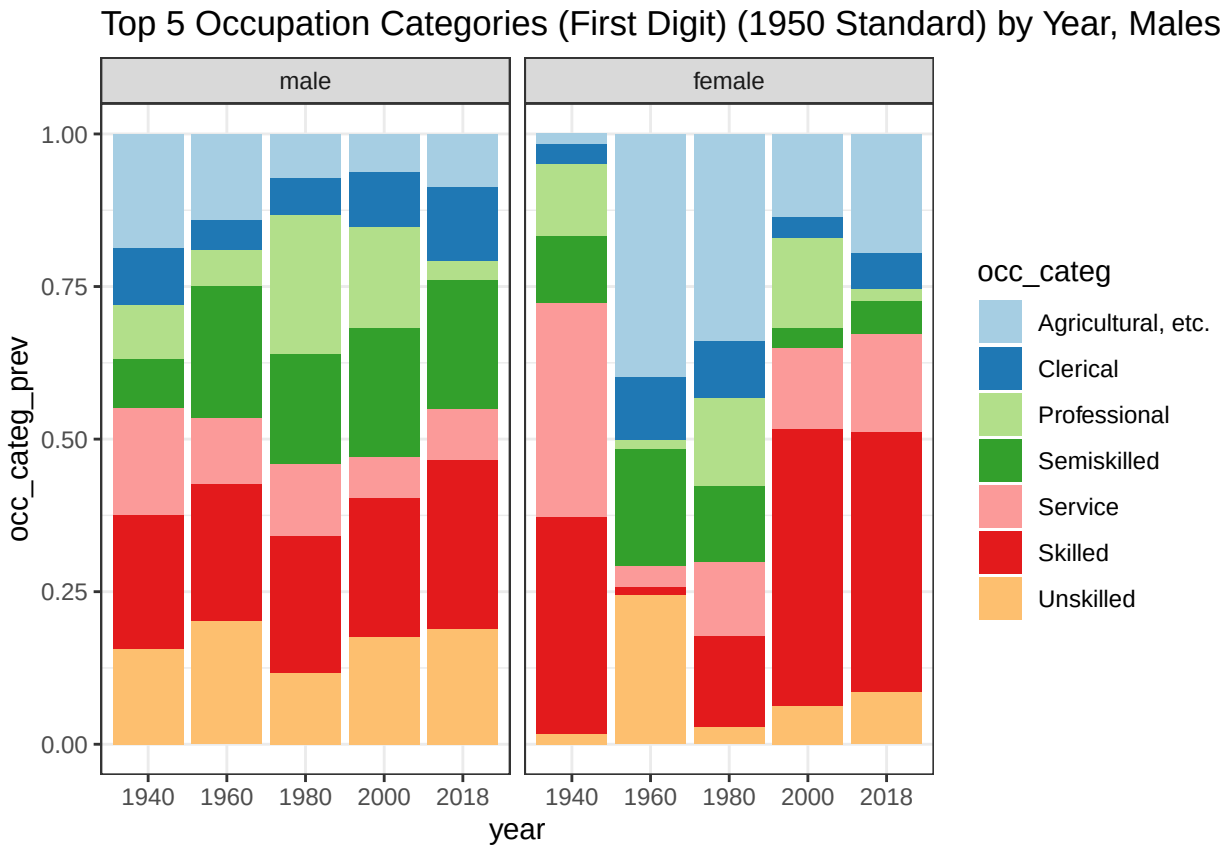


Briefly explore composition of labor market

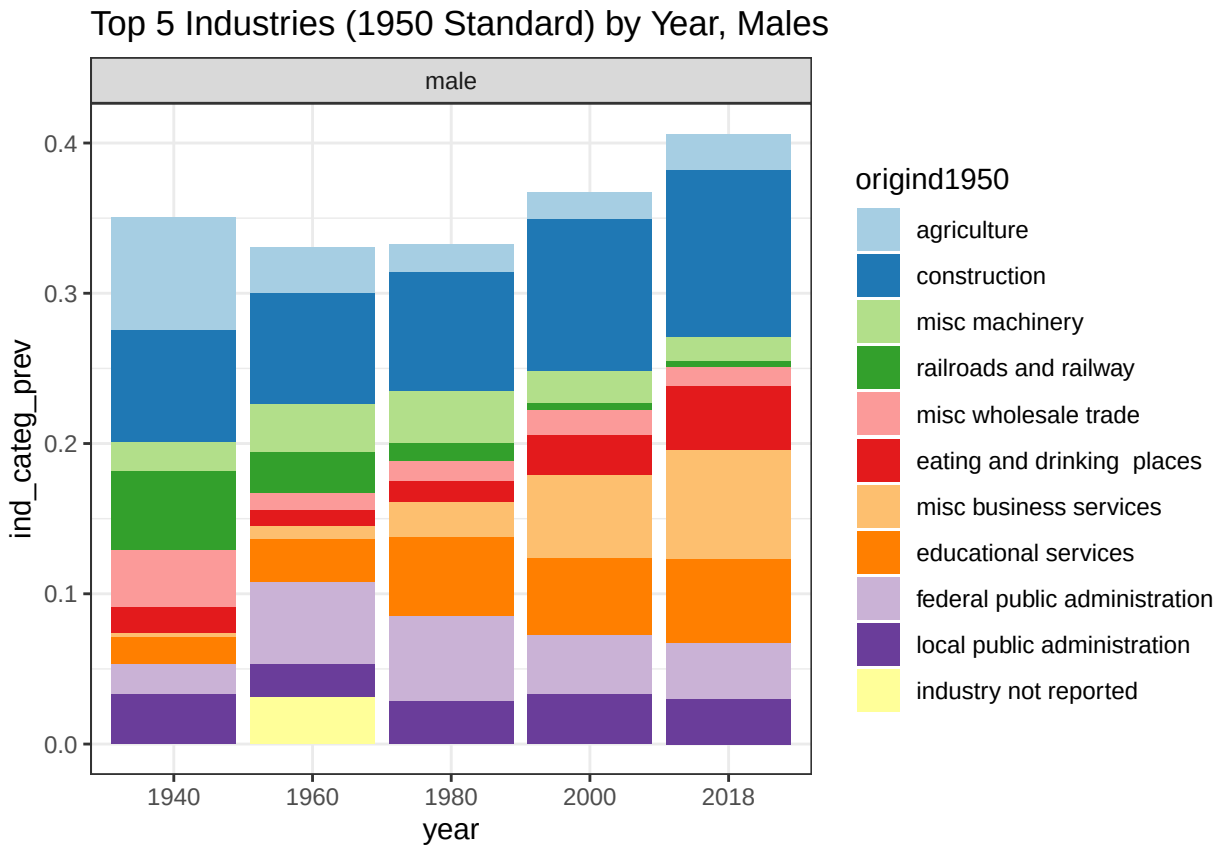
This uses the first prefix of the occ codes

```
census_1940[,sum_perwt := (sum(perwt)), by = .(year, sex)]
occ_categ <- census_1940[,.(occ_categ_prev = (sum(perwt))/mean(sum_perwt)), by = .(year, sex, occ_categ

ggplot(occ_categ) +
```

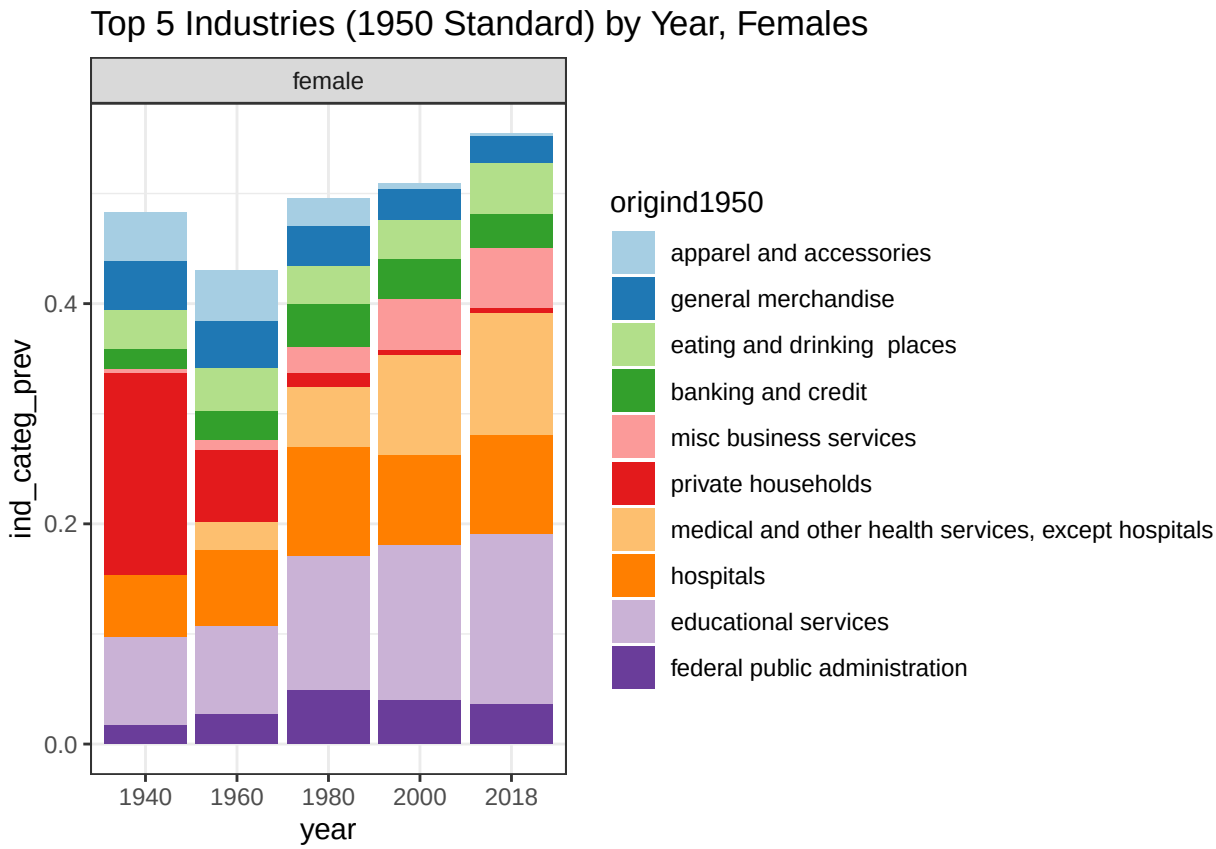
```
  geom_bar(aes(x = year,y=occ_categ_prev, fill = occ_categ),
               stat = "identity", position = "stack") +
  facet_grid(~sex) + scale_fill_brewer(palette = "Paired") +
  ggtitle("Top 5 Occupation Categories (First Digit) (1950 Standard) by Year, Males")
```



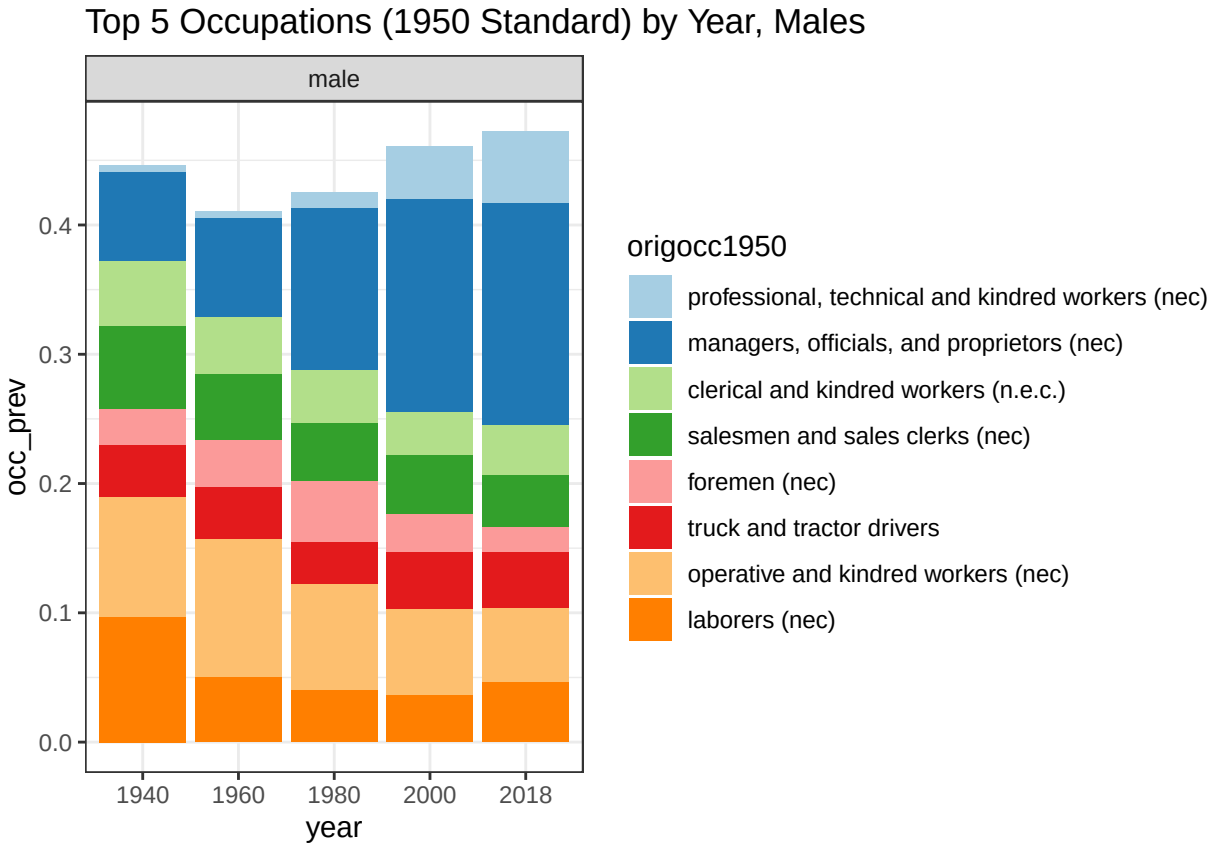

```
# calculate top 10 industries and occs by year
ind_categ <- census_1940[,.(ind_categ_prev = (sum(perwt))/mean(sum_perwt)), by = .(year, sex, origind195
ind_categ[, rank := frankv(ind_categ_prev, order = -1), by = .(year, sex)]
top_10 <- ind_categ[rank %in% 1:5 & sex == "male", unique(origind1950)]

ggplot(ind_categ[origind1950 %in%  ind_categ[rank %in% 1:5 & sex == "male", unique(origind1950)] & sex =
  geom_bar(aes(x = year,y=ind_categ_prev, fill = origind1950),
               stat = "identity", position = "stack") +
  facet_grid(~sex) + scale_fill_brewer(palette = "Paired") +
  ggtitle("Top 5 Industries (1950 Standard) by Year, Males")
```



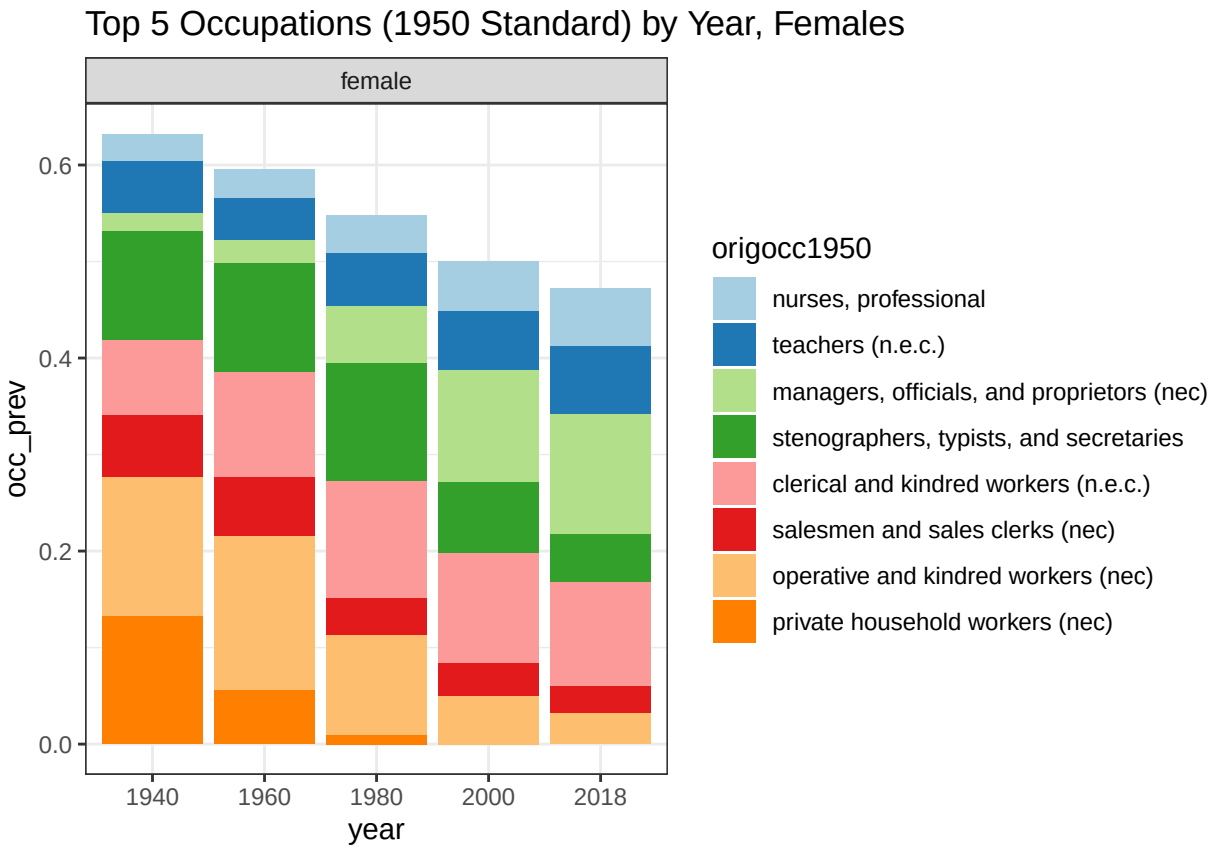

```
ggplot(ind_categ[origind1950 %in%  ind_categ[rank %in% 1:5 & sex == "female", unique(origind1950)] & sex
  geom_bar(aes(x = year,y=ind_categ_prev, fill = origind1950),
             stat = "identity", position = "stack") +
  facet_grid(~sex) + scale_fill_brewer(palette = "Paired") +
  ggtitle("Top 5 Industries (1950 Standard) by Year, Females")
```




```
# calculate top 10 occupations and occs by year
occ_categ <- census_1940[,.(occ_prev = (sum(perwt))/mean(sum_perwt)), by = .(year, sex, origocc1950)]
occ_categ[, rank := frankv(occ_prev, order = -1), by = .(year, sex)]
top_10 <- occ_categ[rank %in% 1:5 & sex == "male", unique(origocc1950)]

ggplot(occ_categ[origocc1950 %in%  occ_categ[rank %in% 1:5 & sex == "male", unique(origocc1950)] & sex =
  geom_bar(aes(x = year,y=occ_prev, fill = origocc1950),
              stat = "identity", position = "stack") +
  facet_grid(~sex) + scale_fill_brewer(palette = "Paired") +
  ggtitle("Top 5 Occupations (1950 Standard) by Year, Males")
```

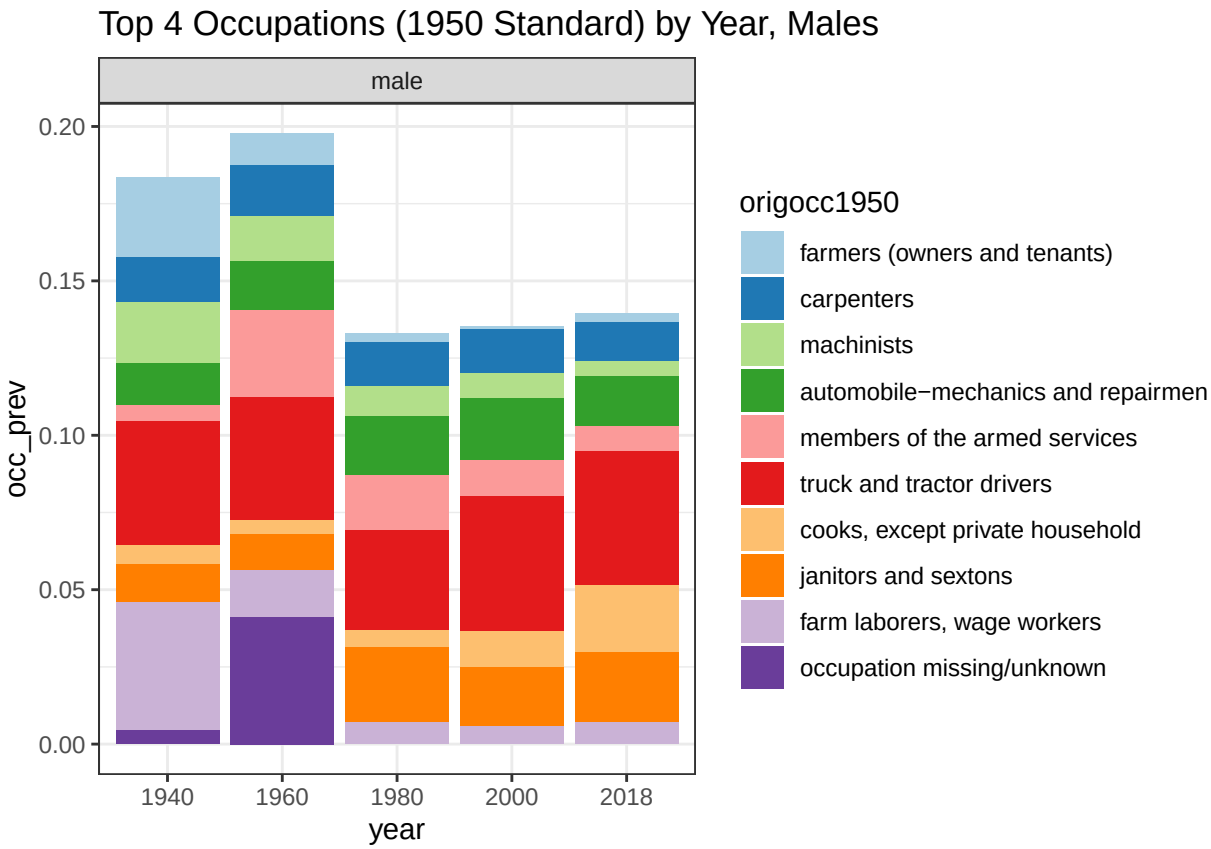



```
ggplot(occ_categ[origocc1950 %in%  occ_categ[rank %in% 1:5 & sex == "female", unique(origocc1950)] & sex
  geom_bar(aes(x = year,y=occ_prev, fill = origocc1950),
             stat = "identity", position = "stack") +
  facet_grid(~sex) + scale_fill_brewer(palette = "Paired") +
  ggtitle("Top 5 Occupations (1950 Standard) by Year, Females")
```

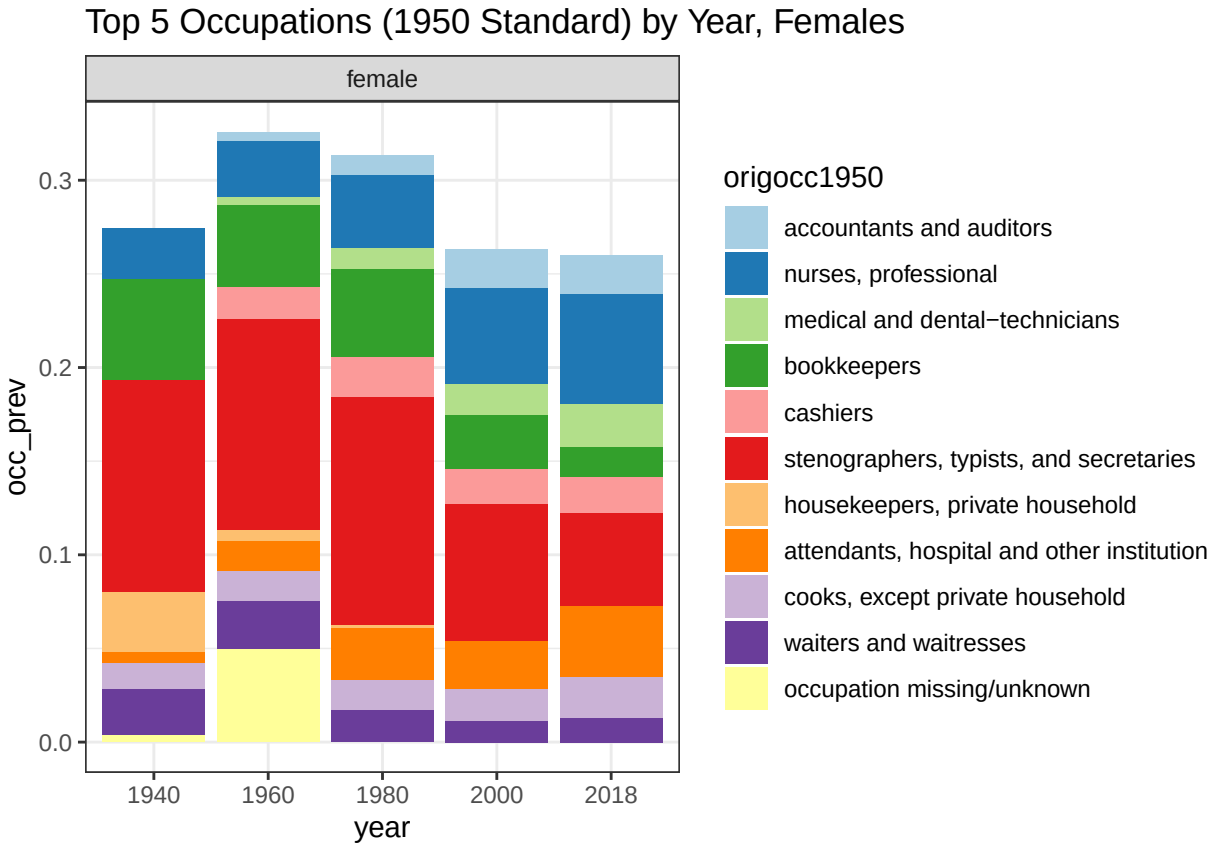


Excluding "NEC" designations

```
# calculate top 10 occupations and occs by year
occ_categ <- census_1940[!origocc1950 %like% "n.e.c.|nec|NEC|N.E.C.",.(occ_prev = (sum(perwt))/mean(sum
occ_categ[, rank := frankv(occ_prev, order = -1), by = .(year, sex)]
top_10 <- occ_categ[rank %in% 1:5 & sex == "male", unique(origocc1950)]

ggplot(occ_categ[origocc1950 %in%  occ_categ[rank %in% 1:4 & sex == "male", unique(origocc1950)] & sex =
  geom_bar(aes(x = year,y=occ_prev, fill = origocc1950),
             stat = "identity", position = "stack") +
  facet_grid(~sex) + scale_fill_brewer(palette = "Paired") +
  ggtitle("Top 4 Occupations (1950 Standard) by Year, Males")
```




```
ggplot(occ_categ[origocc1950 %in%  occ_categ[rank %in% 1:5 & sex == "female", unique(origocc1950)] & sex
  geom_bar(aes(x = year,y=occ_prev, fill = origocc1950),
               stat = "identity", position = "stack") +
  facet_grid(~sex) + scale_fill_brewer(palette = "Paired") +
  ggtitle("Top 5 Occupations (1950 Standard) by Year, Females")
```


Top 5 Occupations (1950 Standard) by Year, Females
female
occ_prev
0.0
0.1
0.2
0.3
1940
1960
1980
2000
2018
year
origocc1950
accountants and auditors
nurses, professional
medical and dental−technicians
bookkeepers
cashiers
stenographers, typists, and secretaries
housekeepers, private household
attendants, hospital and other institution
cooks, except private household
waiters and waitresses
occupation missing/unknown