Final Project

Group 1

2023-11-19

Roles

- Tyson: Data tidying, creating dataframes for modeling and analysis, collaborating on modeling and analysis, minor role in presentation
- Junyoung: Data exploration for coalescent data, presentation and script
- Ikjoo: Data exploration for data by ethnicity, organizing final rmd file, presentation and script
- Wonjun (Jason): Organizing group collaboration efforts and meetings, collaborating on modeling, presentation and script
- Areum: Data exploration for data by gender, collaborating on analysis, presentation and script

Tyson

```r
# Data Tidying
# This code is separating rows in the wages_tidy data frame
# based on whether the "demographic" column contains
# "black," "hispanic," or "white," creating a new column
# "ethnicity" in the process.
wages_ethnicity <- wages_tidy %>%
 filter(grepl(c("black|hispanic|white"), demographic)) %>%
 separate(demographic, into = c("ethnicity", "demographic"),
          sep = "_", extra = "merge") %>%
 arrange(ethnicity, demographic, year)

wages_ethnicity_inverse <- wages_tidy %>%
 filter(!grepl(c("black|hispanic|white"), demographic)) %>%
 mutate(ethnicity = NA) %>%
 relocate(ethnicity, .after = year)
wages_ethnicity_combined <- wages_ethnicity %>%
 bind_rows(wages_ethnicity_inverse) %>%
 arrange(ethnicity, demographic, year)
```

```r
# Data Tidying
# Similar process as creating the "ethnicity".
# Created new row called "Gender" and combined with "wages_ethnicity_combined"
wages_gender <- wages_ethnicity_combined %>%
 filter(grepl(c("men|women"), demographic)) %>%
 separate(demographic, into = c("gender", "demographic"),
          sep = "_", extra = "merge") %>%
 arrange(ethnicity, gender, demographic, year)
```

```r
wages_gender_inverse <- wages_ethnicity_combined %>%
 filter(!grepl(c("men|women"), demographic)) %>%
 mutate(gender = NA) %>%
 relocate(gender, .after = ethnicity)
wages_gender_combined <- wages_gender %>%
 bind_rows(wages_gender_inverse) %>%
 arrange(ethnicity, gender, demographic, year)
```

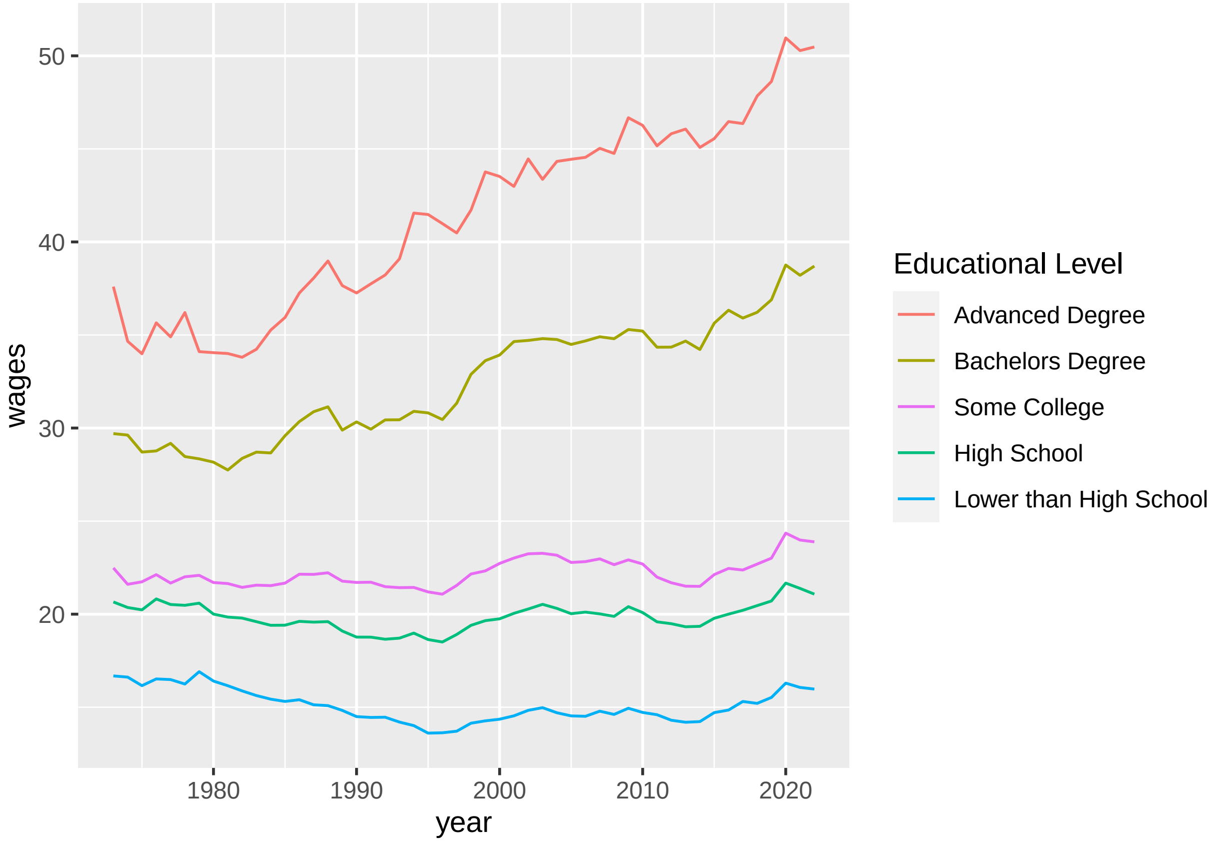
```r
# Data Tidying
# Reordering the demographic for better visualization in the graph
wages_final <- wages_sep

wages_final$demographic <- factor(wages_final$demographic,
                                levels = c("advanced_degree",
                                           "bachelors_degree",
                                           "some_college",
                                           "high_school",
                                           "less_than_hs"),
                                labels = c("Advanced Degree",
                                           "Bachelors Degree",
                                           "Some College",
                                           "High School",
                                           "Less than High School"),
                                ordered = TRUE)
```

##Junyoung

```r
# Separated certain column(year, demographic, wages).
# Preprocessing step for the further analysis.
overall_income <- read.csv("wages_sep.csv") %>%
  select(year, demographic, wages)
```

```r
# The mean overall wages by "year" and "demographic"  between 1973 ~ 2022.
aggregated_overall_income <- aggregate(wages ~ year
                                       + demographic,
                                        data = overall_income, FUN = mean)
```

```r
# Line plot based on data: "aggregated overall income".
# Utilized scale_color_discrete function for better visualization
library(ggplot2)
aggregated_overall_income %>%
  ggplot() +
  geom_line(
    mapping = aes(
      x = year,
      y = wages,
      color = demographic
    )) + scale_color_discrete(name = "Educational Level",
                              breaks =c("advanced_degree",
                                        "bachelors_degree",
                                        "some_college",
                                        "high_school",
                                        "less_than_hs"),
                              labels = c("Advanced Degree",
                                         "Bachelors Degree",
                                         "Some College",
                                         "High School",
                                         "Lower than High School"))
```



Areum

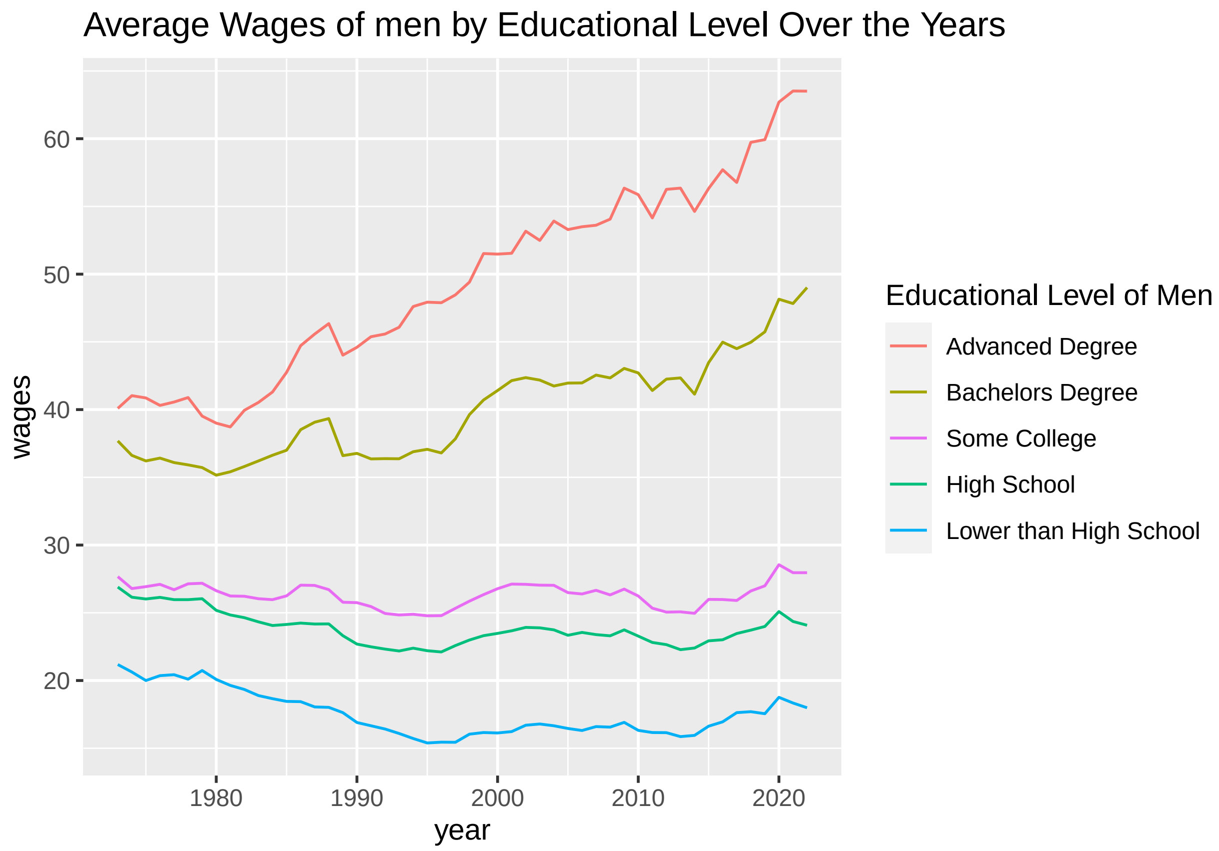
```
# Data of men
wbe_men <- wages_by_education %>%
  select(year,men_less_than_hs:men_advanced_degree)
```

```
wbe_women <- wages_by_education %>%
  select(year,women_less_than_hs:women_advanced_degree)
```

```
wbe_men2 <-wbe_men %>%
  pivot_longer(cols =2:men_advanced_degree, names_to = 'educational_level',
               values_to = 'wages' )
```

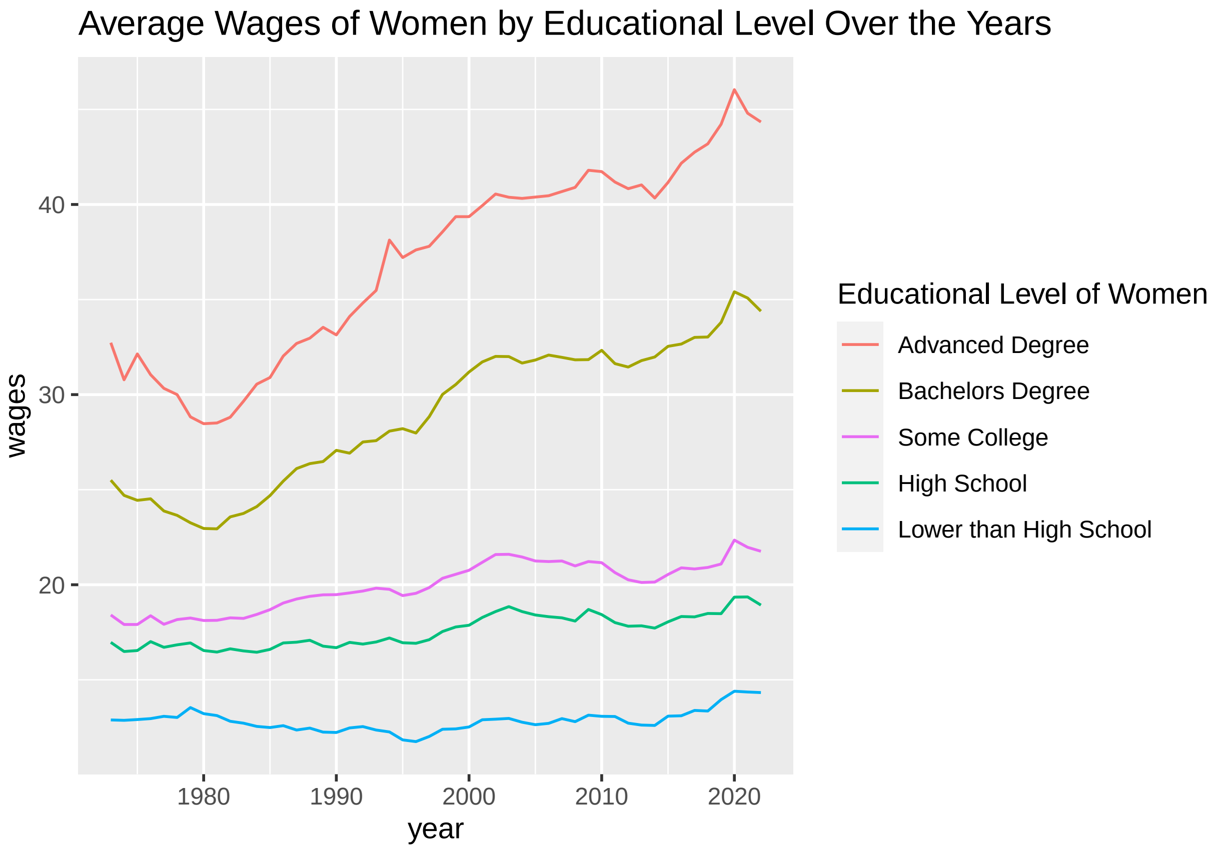
```
# Data Exploration
# Wages of men overtime based on educational level
```

```
wbe_men2 %>%
  ggplot()+
  geom_line(
    mapping = aes(x = year, y = wages,
                  color = educational_level)
  )+
  scale_color_discrete(name = "Educational Level of Men",
                       breaks =c("men_advanced_degree",
                                "men_bachelors_degree",
                                "men_some_college",
                                "men_high_school",
                                "men_less_than_hs"),
                       labels = c("Advanced Degree",
                                 "Bachelors Degree",
                                 "Some College",
                                 "High School",
                                 "Lower than High School"))+
  labs(title="Average Wages of men by Educational Level Over the Years"
  )
```



```
wbe_women2 <-wbe_women %>%
  pivot_longer(cols =2:women_advanced_degree, names_to = 'educational_level',
               values_to = 'wages' )
```

```
# Data Exploration / Line plot
# Wages of women overtime based on educational level
wbe_women2 %>%
  ggplot()+
  geom_line(
    mapping = aes(x = year, y = wages,
                  color = educational_level))+
  scale_color_discrete(name = "Educational Level of Women",
                       breaks =c("women_advanced_degree",
                                 "women_bachelors_degree",
                                 "women_some_college",
                                 "women_high_school",
                                 "women_less_than_hs"),
                       labels = c("Advanced Degree",
                                  "Bachelors Degree",
                                  "Some College",
                                  "High School",
                                  "Lower than High School"))+
  labs(title="Average Wages of Women by Educational Level Over the Years"
  )
```



Ikjoo

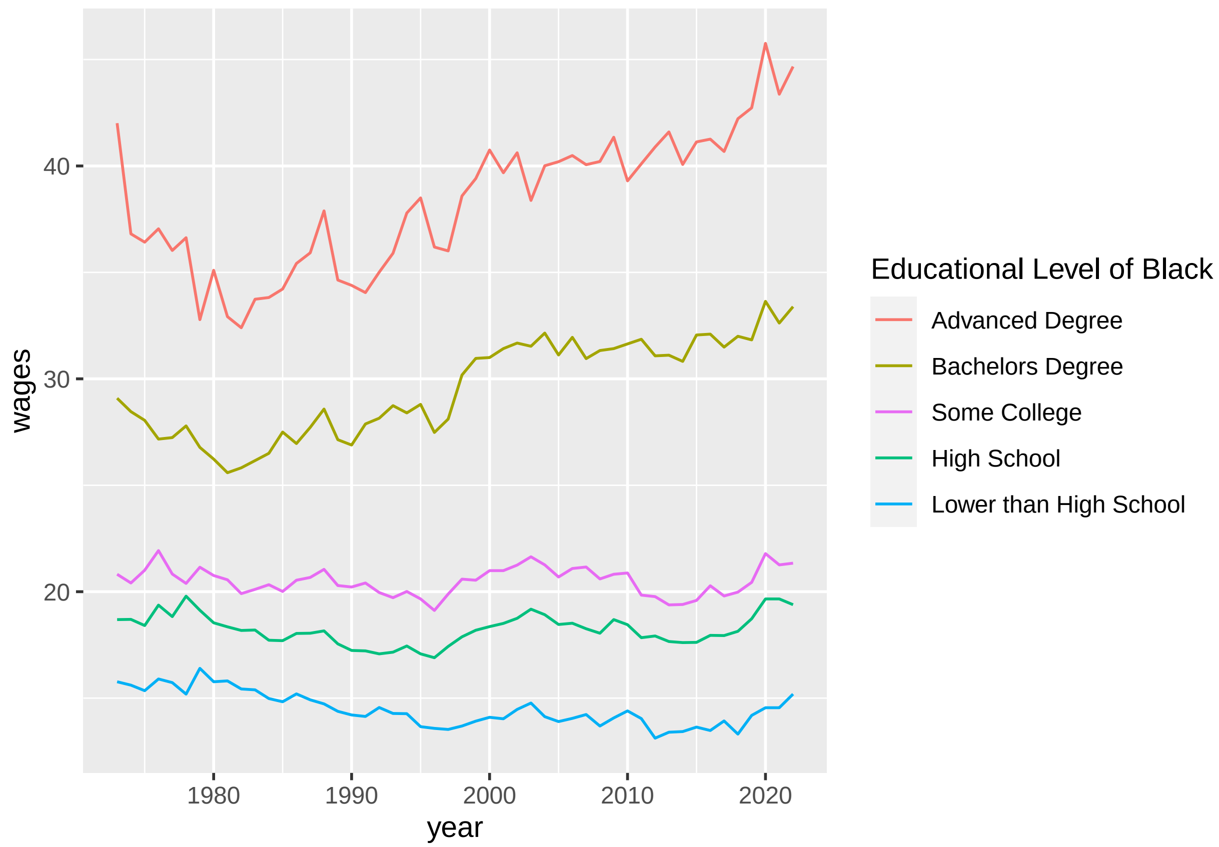
```
# Data of White people
wbe_white <- wages_by_education %>%
  select(year,white_less_than_hs:white_advanced_degree)
```

```
# Data of Black people
wbe_black <- wages_by_education %>%
  select(year,black_less_than_hs:black_advanced_degree)
```

```
# Data of Hispanic People
wbe_hispanic <- wages_by_education %>%
  select(year,hispanic_less_than_hs:hispanic_advanced_degree)
```

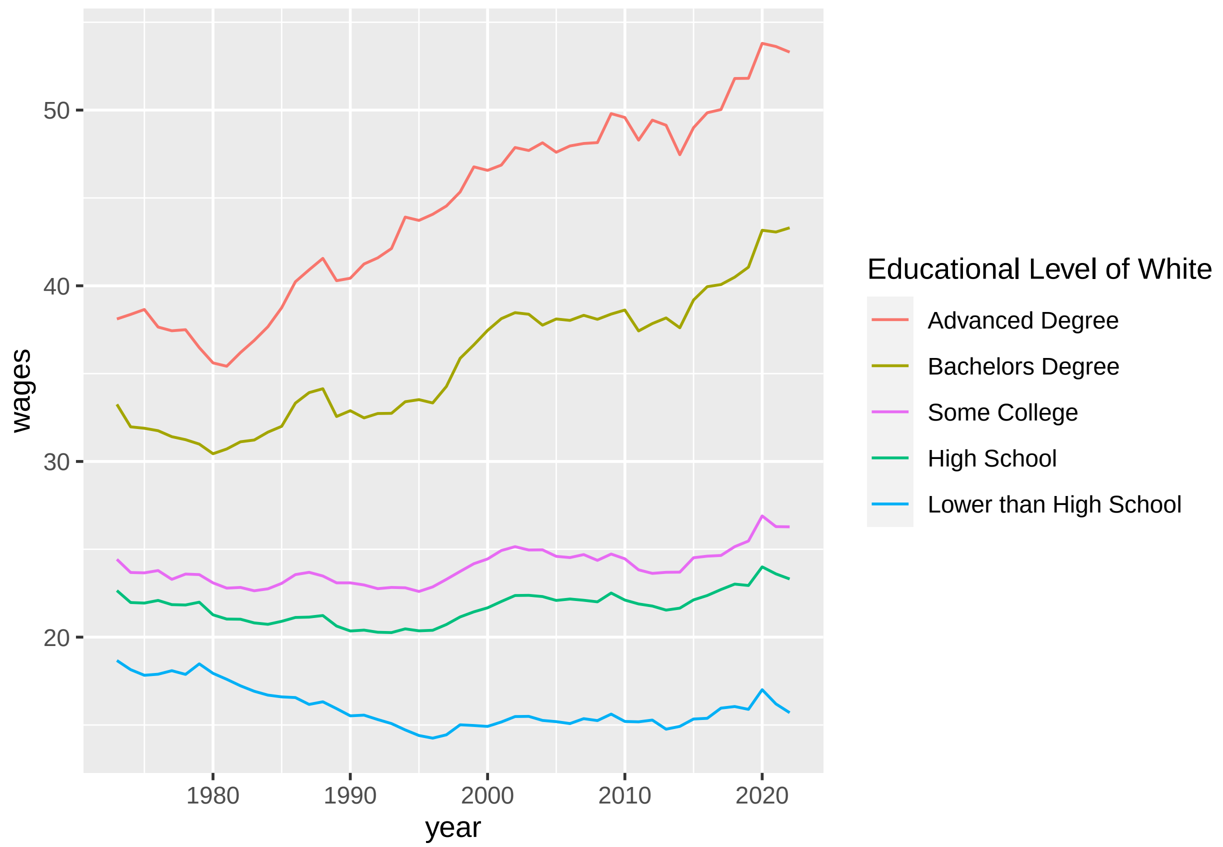
```
# Black people divided in their level of education
wbe_black_education <-wbe_black %>%
pivot_longer(cols =2:black_advanced_degree, names_to = 'educational_level',
             values_to = 'wages')
```

```
# Data Exploration / Line plot
# Wages of Black People overtime based on educational level
wbe_black_education %>%
  ggplot()+
  geom_line(mapping = aes(x = year,
                          y = wages,
                          color = educational_level)) +
  scale_color_discrete(name = "Educational Level of Black",
                       breaks =c("black_advanced_degree",
                                "black_bachelors_degree",
                                "black_some_college",
                                "black_high_school",
                                "black_less_than_hs"),
                       labels = c("Advanced Degree",
                                 "Bachelors Degree",
                                 "Some College",
                                 "High School",
                                 "Lower than High School"))
```



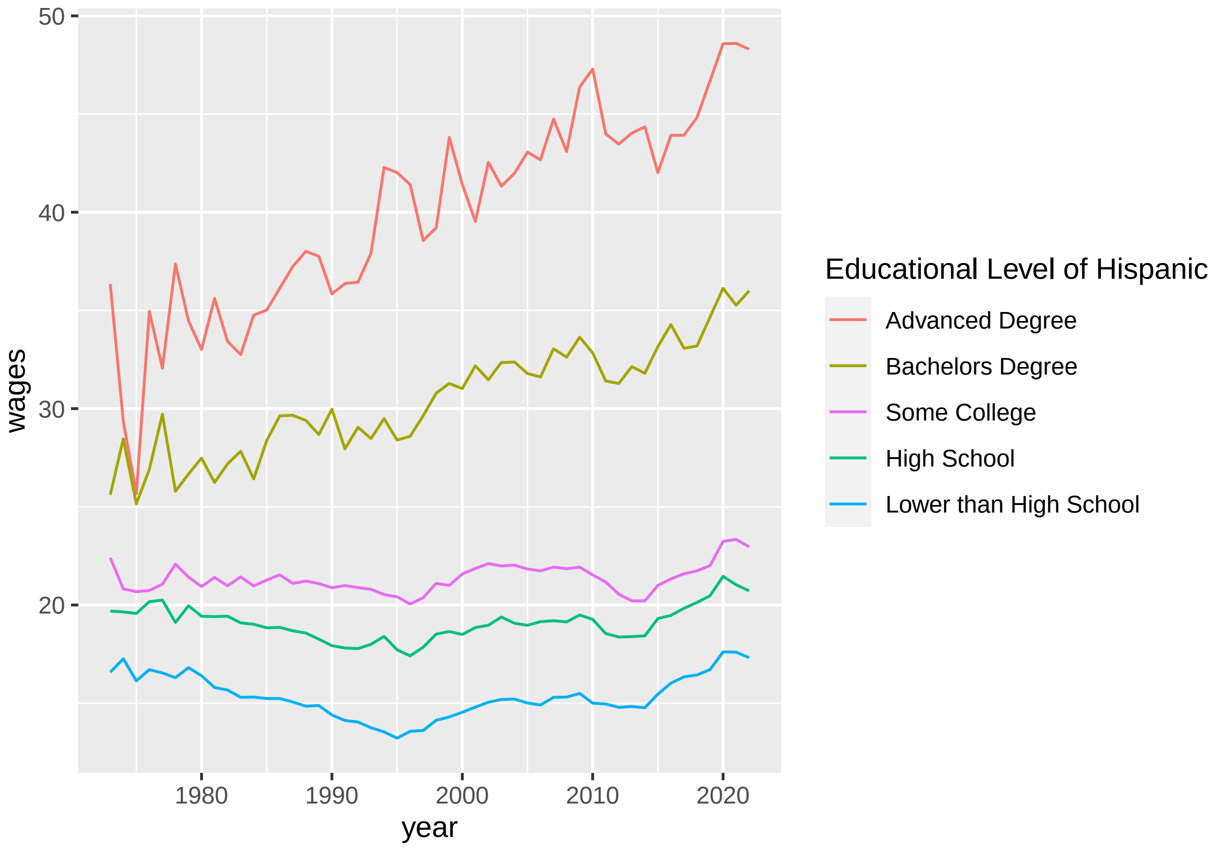
```
# White people divided in their level of education
wbe_white_education <-wbe_white %>%
pivot_longer(cols =2:white_advanced_degree, names_to = 'educational_level',
             values_to = 'wages')
```

```
wbe_white_education %>%
  ggplot()+
  geom_line(mapping = aes(x = year,
                          y = wages,
                          color = educational_level))+
              scale_color_discrete(name = "Educational Level of White",
                                   breaks =c("white_advanced_degree",
                                             "white_bachelors_degree",
                                             "white_some_college",
                                             "white_high_school",
                                             "white_less_than_hs"),
                                   labels = c("Advanced Degree",
                                              "Bachelors Degree",
                                              "Some College",
                                              "High School",
                                              "Lower than High School"))
```



```r
# Hispanic people divided in their level of education
wbe_hispanic_education <-wbe_hispanic %>%
pivot_longer(cols =2:hispanic_advanced_degree, names_to
             ='educational_level',
             values_to = 'wages')
```

```r
# Data Exploration / Line Plot
# Wages of Hispanic overtime based on educational level
wbe_hispanic_education %>%
  ggplot()+
  geom_line(mapping = aes(x = year,
                          y = wages,
                          color = educational_level))+
  scale_color_discrete(name = "Educational Level of Hispanic",
                                 breaks =c("hispanic_advanced_degree",
                                           "hispanic_bachelors_degree",
                                           "hispanic_some_college",
                                           "hispanic_high_school",
                                           "hispanic_less_than_hs"),
                                 labels = c("Advanced Degree",
                                            "Bachelors Degree",
                                            "Some College",
                                            "High School",
                                            "Lower than High School"))
```



Wonjun (Jason) Lee

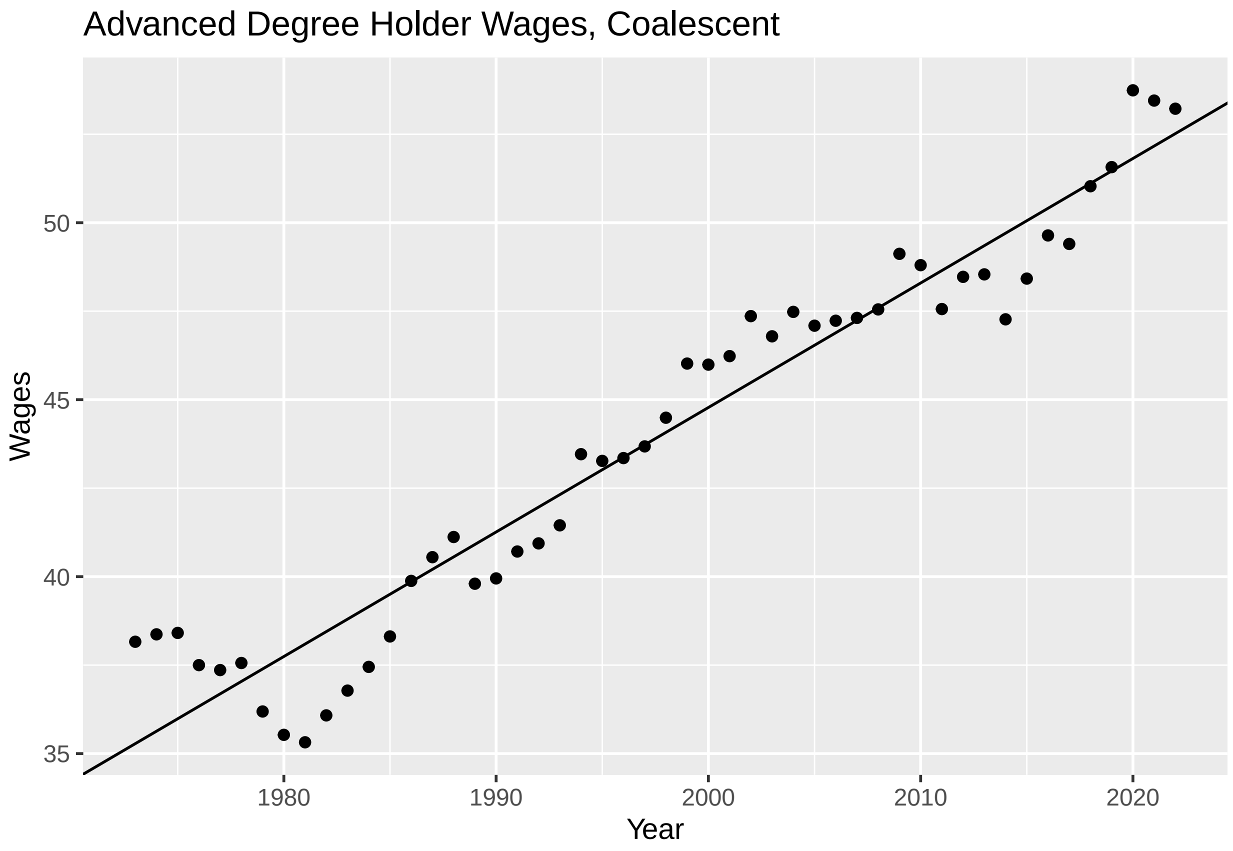
Extrapolating Wages

```
wages_ad_fil <- wages_sep %>%
  filter(
  is.na(ethnicity)
  & is.na(gender)
  & demographic == "advanced_degree"
  )

wages_ad_lm <- lm(
    wages ~ year, data = wages_ad_fil
  )
```

Created the scatter plot with regression line for each specific educational level. The gender and ethnicity has not been considered.

Demonstration

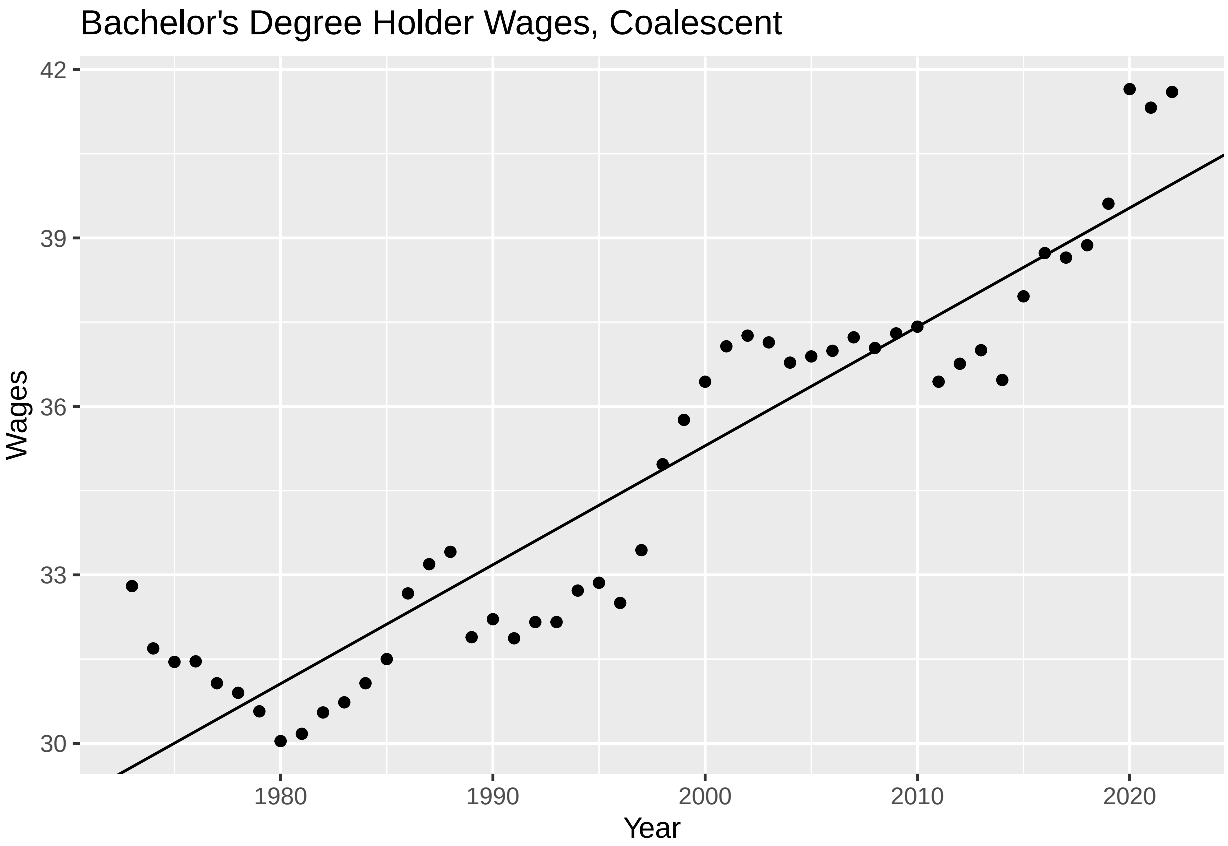
```r
wages_ad_fil %>%
  ggplot() +
  geom_point(
    mapping = aes(
      x = year,
      y = wages
    )
  ) +
  geom_abline(
    slope = wages_ad_lm$coefficients[2],
    intercept = wages_ad_lm$coefficients[1]
  ) +
  labs(
    x = "Year",
    y = "Wages",
    title = "Advanced Degree Holder Wages, Coalescent"
  )
```



```r
wages_bd_fil <- wages_sep %>%
  filter(
  is.na(ethnicity)
  & is.na(gender)
  & demographic == "bachelors_degree"
  )
```

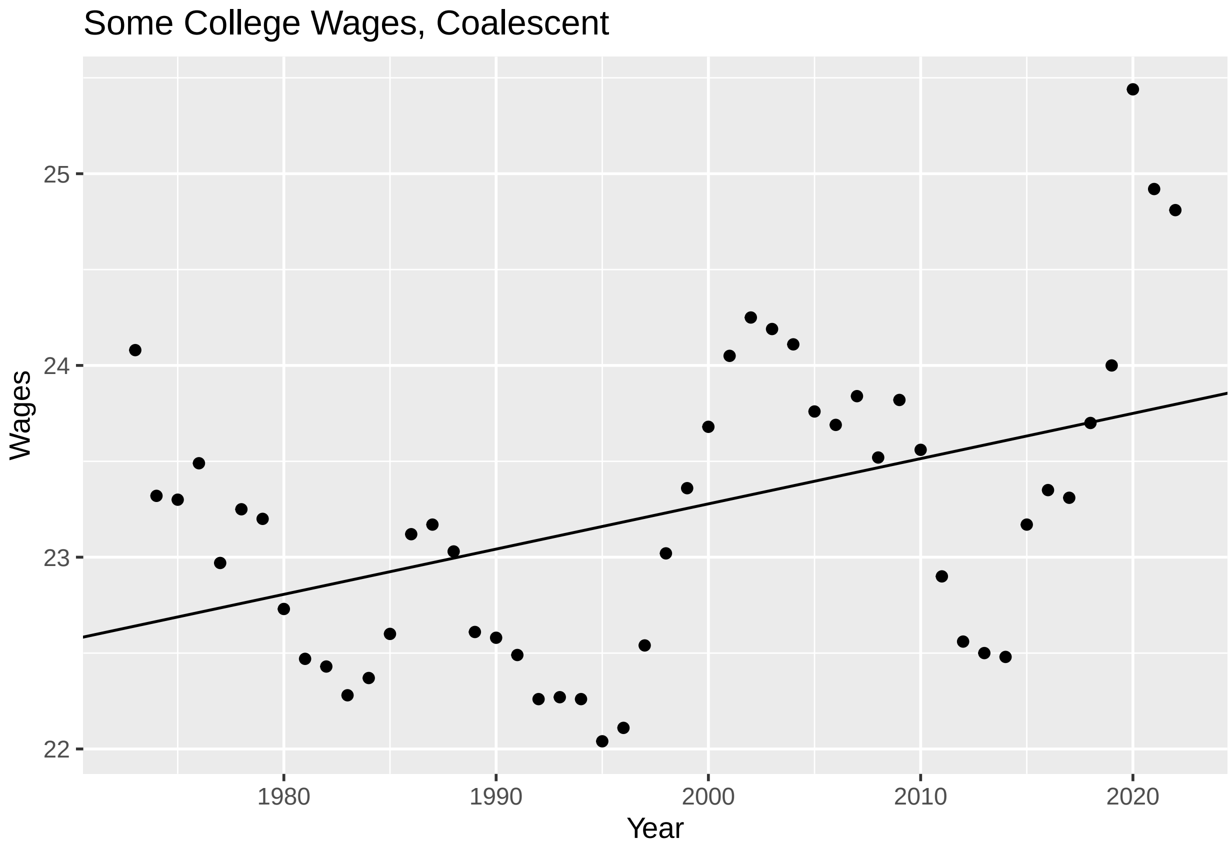
```
wages_bd_lm <- lm(
    wages ~ year, data = wages_bd_fil
  )
```

```
wages_bd_fil %>%
  ggplot() +
  geom_point(
    mapping = aes(
      x = year,
      y = wages
    )
  ) +
  geom_abline(
    slope = wages_bd_lm$coefficients[2],
    intercept = wages_bd_lm$coefficients[1]
  ) +
  labs(
    x = "Year",
    y = "Wages",
    title = "Bachelor's Degree Holder Wages, Coalescent"
  )
```



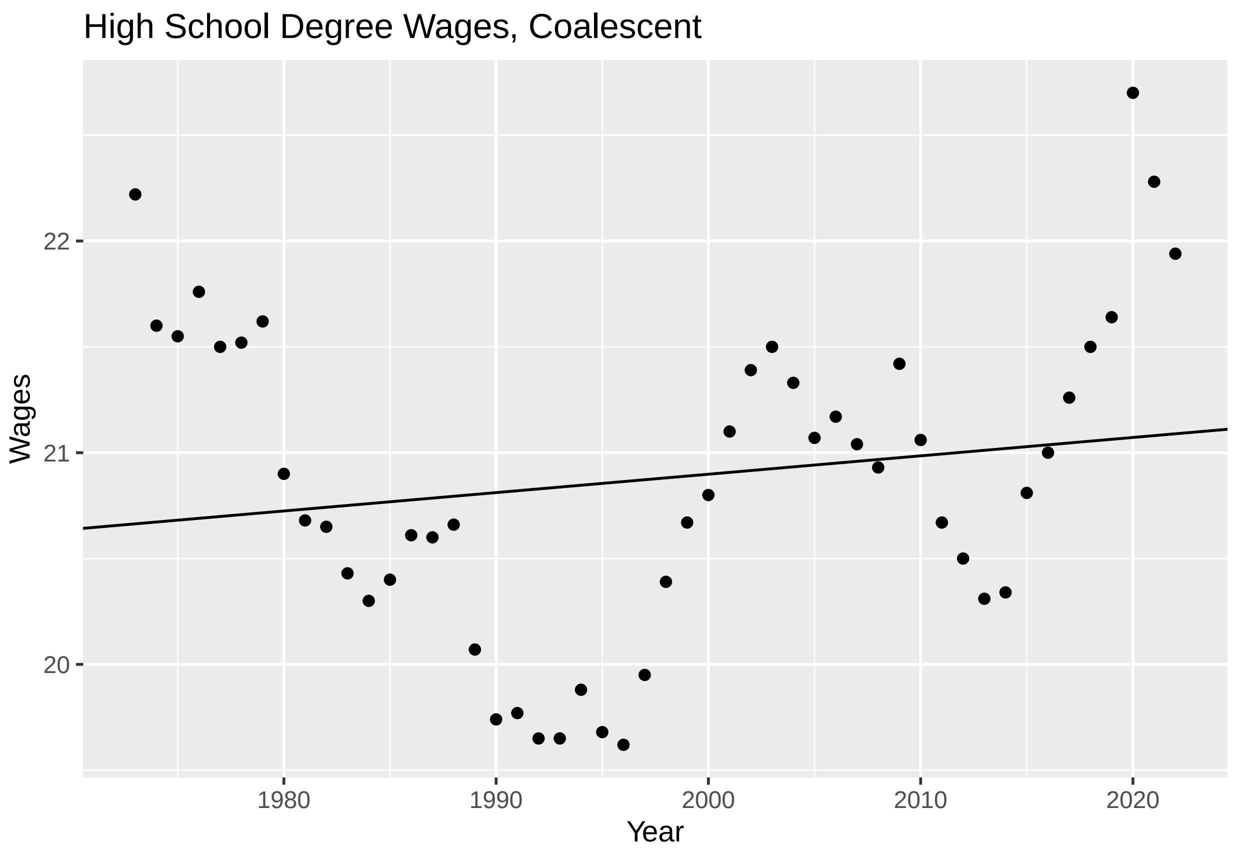
```
wages_sc_fil <- wages_sep %>%
  filter(
  is.na(ethnicity)
  & is.na(gender)
  & demographic == "some_college"
  )

wages_sc_lm <- lm(
    wages ~ year, data = wages_sc_fil
  )
```

```
wages_sc_fil %>%
  ggplot() +
  geom_point(
    mapping = aes(
      x = year,
      y = wages
    )
  ) +
  geom_abline(
    slope = wages_sc_lm$coefficients[2],
    intercept = wages_sc_lm$coefficients[1]
  ) +
  labs(
    x = "Year",
    y = "Wages",
    title = "Some College Wages, Coalescent"
  )
```



```
wages_hs_fil <- wages_sep %>%
  filter(
  is.na(ethnicity)
  & is.na(gender)
  & demographic == "high_school"
  )

wages_hs_lm <- lm(
    wages ~ year, data = wages_hs_fil
  )
```

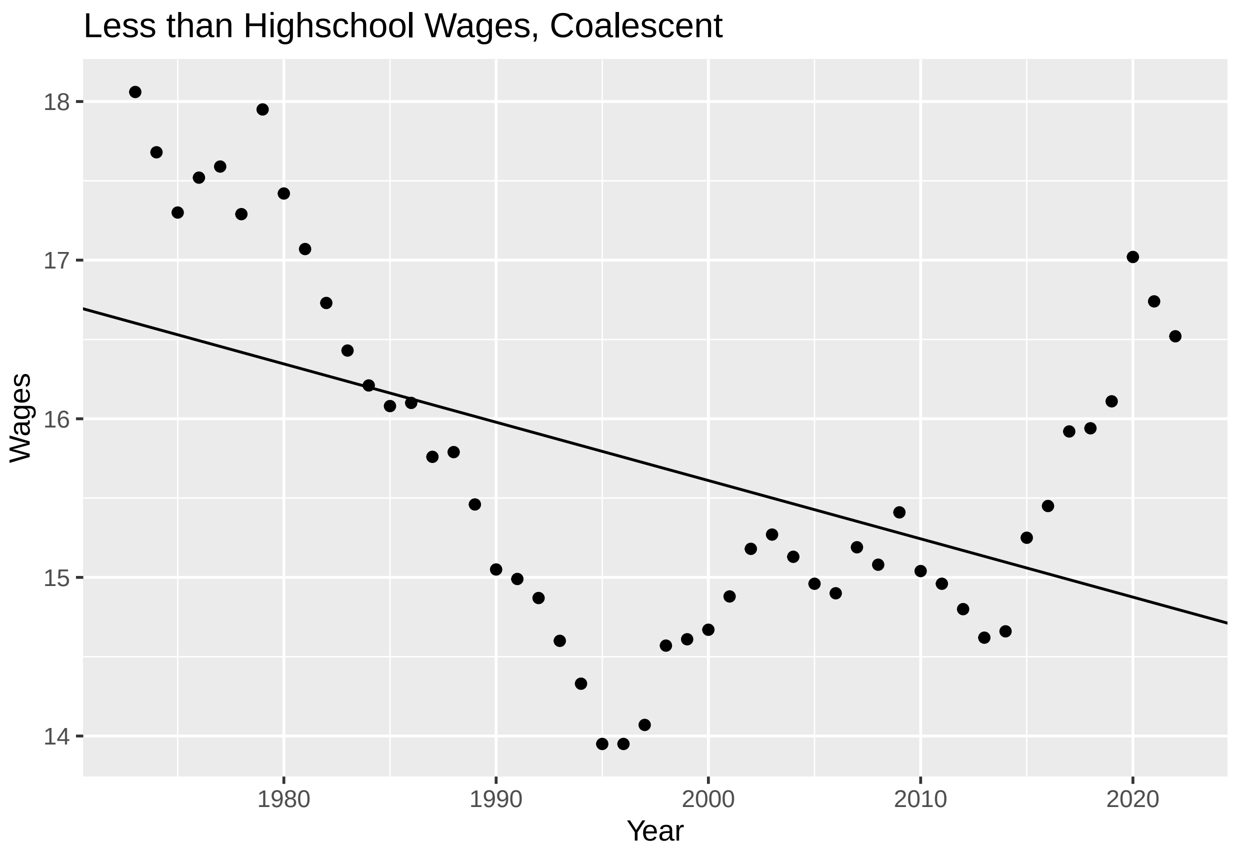
```
wages_hs_fil %>%
  ggplot() +
  geom_point(
    mapping = aes(
      x = year,
      y = wages
    )
  ) +
  geom_abline(
    slope = wages_hs_lm$coefficients[2],
    intercept = wages_hs_lm$coefficients[1]
  ) +
  labs(
    x = "Year",
    y = "Wages",
    title = "High School Degree Wages, Coalescent"
  )
```



```
wages_lhs_fil <- wages_sep %>%
  filter(
  is.na(ethnicity)
  & is.na(gender)
  & demographic == "less_than_hs"
  )

wages_lhs_lm <- lm(
    wages ~ year, data = wages_lhs_fil
  )
```

```
wages_lhs_fil %>%
  ggplot() +
  geom_point(
    mapping = aes(
      x = year,
      y = wages
    )
  ) +
  geom_abline(
    slope = wages_lhs_lm$coefficients[2],
    intercept = wages_lhs_lm$coefficients[1]
  ) +
```

```r
  labs(
    x = "Year",
    y = "Wages",
    title = "Less than Highschool Wages, Coalescent"
  )
```



```r
# The prediction of future wages based on educational level
wages_extra <- data.frame(
year=c(2023:2040)
)
wages_extra$advanced_degree <- predict(wages_ad_lm, wages_extra)
wages_extra$bachelors_degree <- predict(wages_bd_lm, wages_extra)
wages_extra$some_college <- predict(wages_sc_lm, wages_extra)
wages_extra$high_school <- predict(wages_hs_lm, wages_extra)
wages_extra$less_than_hs  <- predict(wages_lhs_lm, wages_extra)
```

Collaboration between Tyson and Areum

```r
sal_growth <- extra_adj_sal %>%
  mutate(growth_rate = 0)
for(m in 2:50) {
  for(n in 0:4) {
    sal_growth$growth_rate[m+50*n] = (sal_growth$adj_salary[m+50*n] / sal_growth$adj_salary[m+50*n-1])
  }
}
```

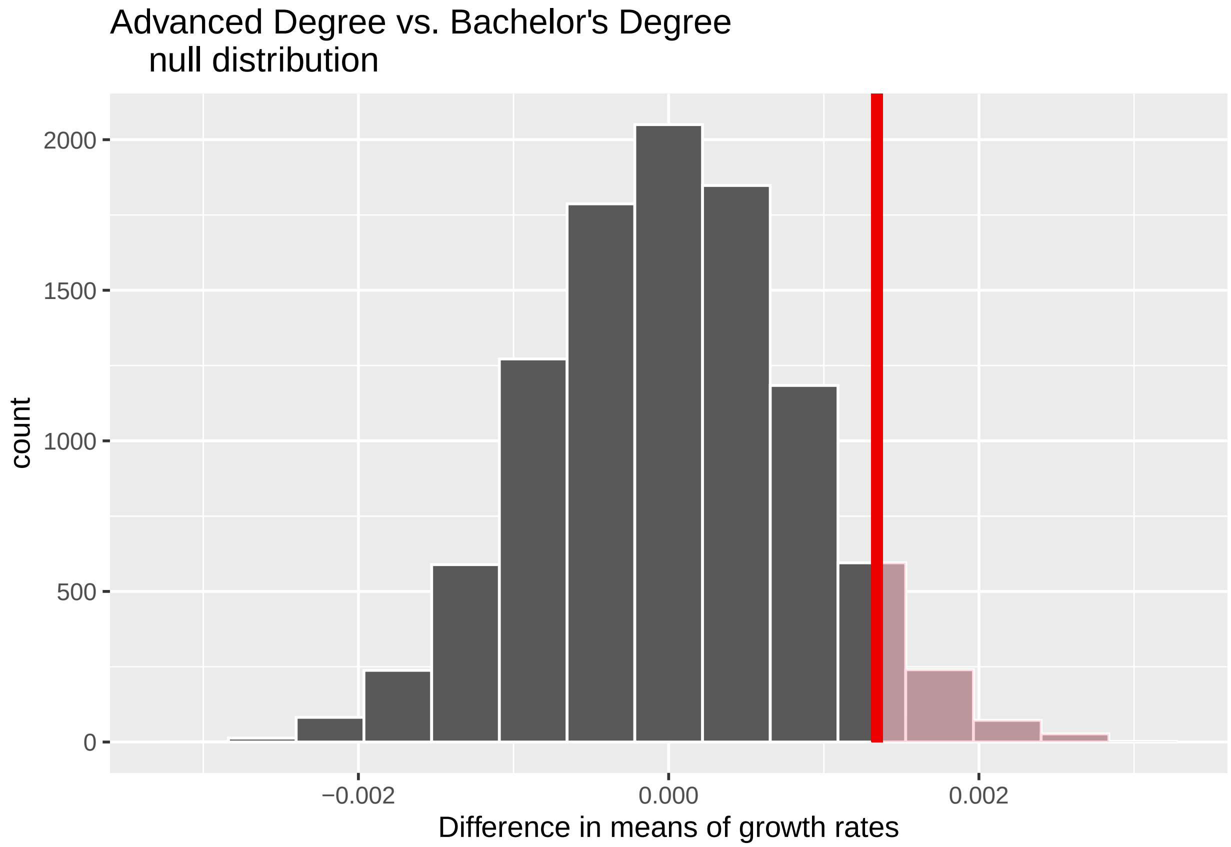
```
sal_growth <- sal_growth %>%
select(c("degree", "growth_rate")) %>%
filter(growth_rate != 0)
```

```
set.seed(111)
# Advanced Degree vs. Bacehlor's Degree
ad_vs_bd_df <- sal_growth %>%
  filter(degree %in% c("advanced_degree", "bachelors_degree"))

ad_vs_bd_null <- ad_vs_bd_df %>%
  specify(growth_rate ~ degree) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 10000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("advanced_degree", "bachelors_degree"))

ad_vs_bd_obs_stat <- ad_vs_bd_df %>%
  specify(growth_rate ~ degree) %>%
  calculate(stat = "diff in means", order = c("advanced_degree", "bachelors_degree"))
```

```
ad_vs_bd_null %>%
  get_p_value(obs_stat = ad_vs_bd_obs_stat, direction = "right")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1  0.0561
```

```
set.seed(111)
ad_vs_bd_null %>%
  visualize() +
  shade_p_value(obs_stat = ad_vs_bd_obs_stat, direction = "right")+
  labs(
    title = "Advanced Degree vs. Bachelor's Degree
    null distribution",
    x= "Difference in means of growth rates"
  )
```

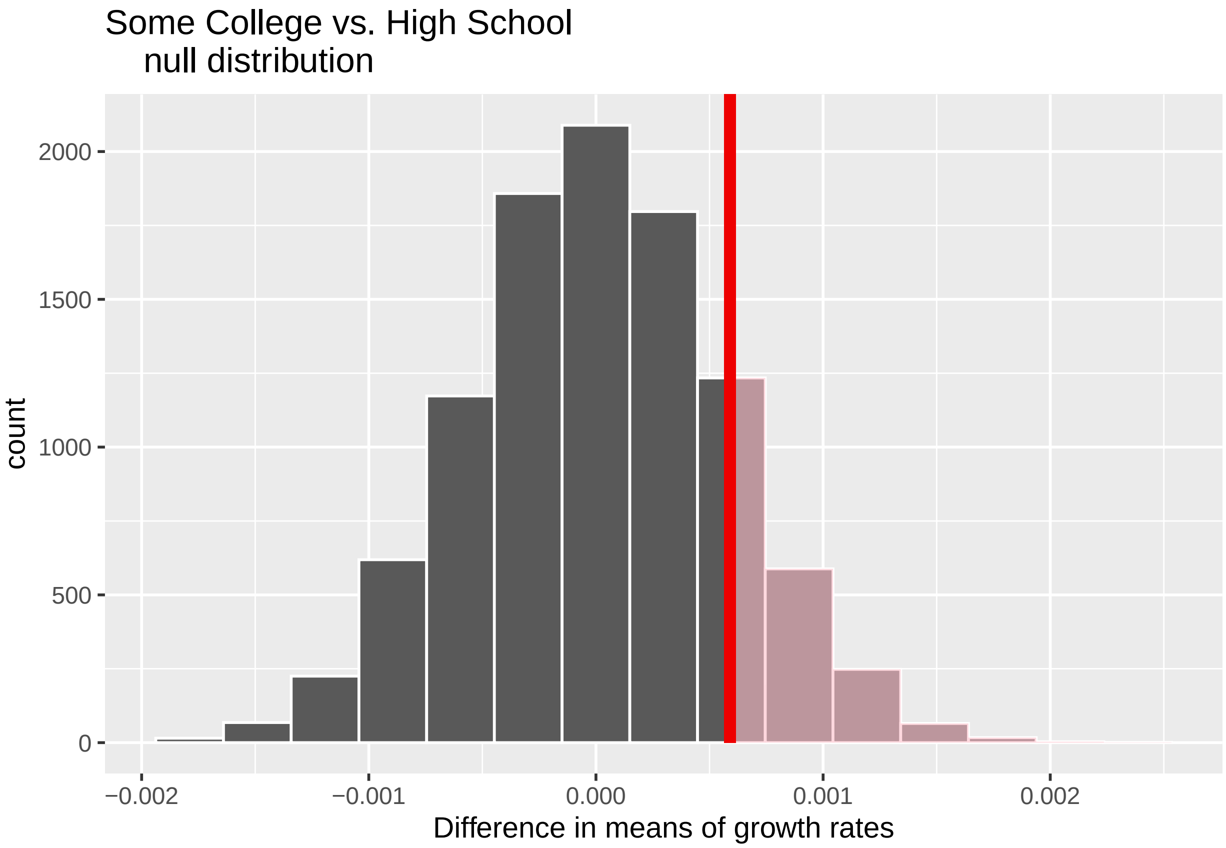


```r
# Some College vs. High School
set.seed(123)
sc_vs_hs_df <- sal_growth %>%
  filter(degree %in% c("some_college", "high_school"))

sc_vs_hs_null <- sc_vs_hs_df %>%
  specify(growth_rate ~ degree) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 10000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("some_college", "high_school"))

sc_vs_hs_obs_stat <- sc_vs_hs_df %>%
  specify(growth_rate ~ degree) %>%
  calculate(stat = "diff in means", order = c("some_college", "high_school"))
```

```r
sc_vs_hs_null %>%
  get_p_value(obs_stat = sc_vs_hs_obs_stat, direction = "right")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.149
```

```r
set.seed(123)
sc_vs_hs_null %>%
  visualize() +
  shade_p_value(obs_stat = sc_vs_hs_obs_stat, direction = "right")+
  labs(
    title = "Some College vs. High School
    null distribution",
    x= "Difference in means of growth rates"
  )
```



```r
# Advanced Degree vs. High School
set.seed(124)
ad_vs_hs_df <- sal_growth %>%
  filter(degree %in% c("advanced_degree", "high_school"))

ad_vs_hs_null <- ad_vs_hs_df %>%
  specify(growth_rate ~ degree) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 10000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("advanced_degree", "high_school"))

ad_vs_hs_obs_stat <- ad_vs_hs_df %>%
  specify(growth_rate ~ degree) %>%
  calculate(stat = "diff in means", order = c("advanced_degree", "high_school"))
```

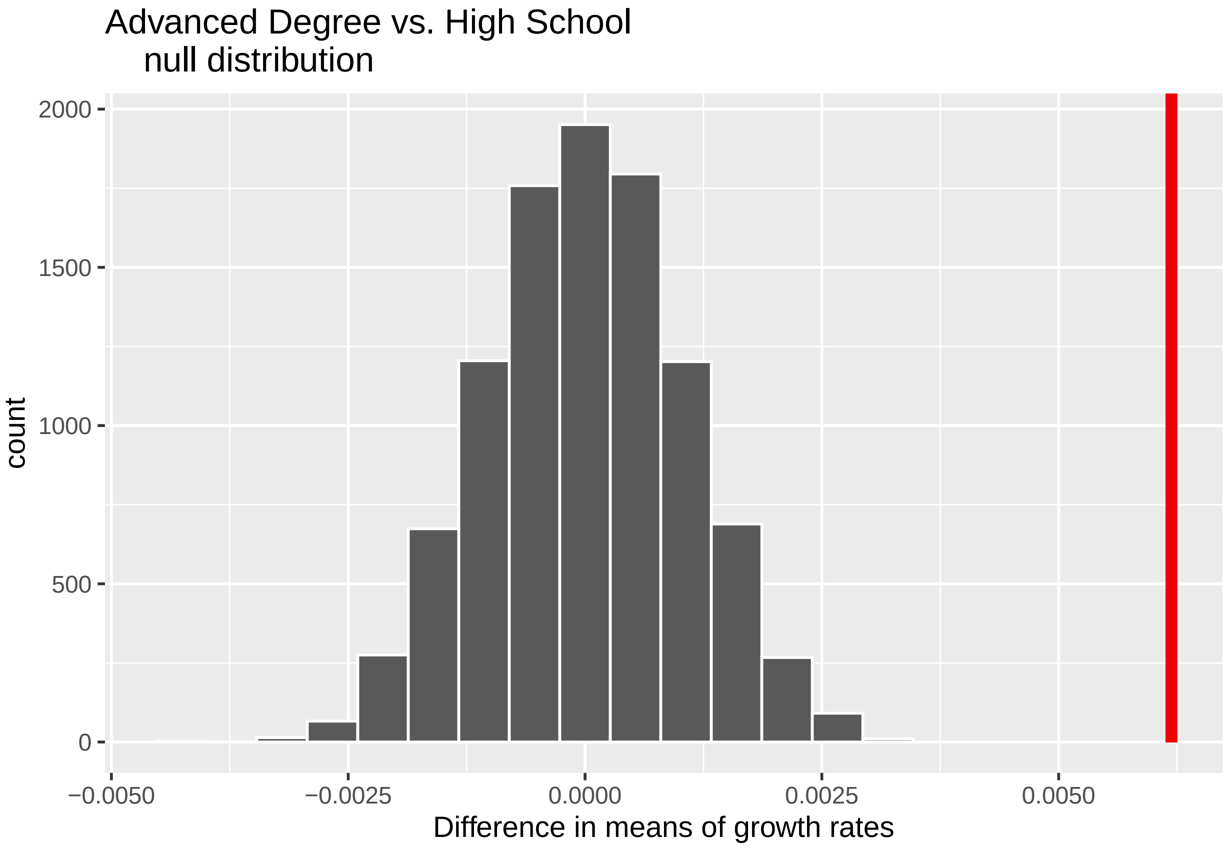
```
ad_vs_hs_null %>%
  get_p_value(obs_stat = ad_vs_hs_obs_stat, direction = "right")
```

```
## Warning: Please be cautious in reporting a p-value of 0. This result is an
## approximation based on the number of `reps` chosen in the `generate()` step.
## See `?get_p_value()` for more information.
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1       0
```

```
set.seed(124)
ad_vs_hs_null %>%
  visualize() +
  shade_p_value(obs_stat = ad_vs_hs_obs_stat, direction = "right")+
  labs(
    title = "Advanced Degree vs. High School
    null distribution",
    x= "Difference in means of growth rates"
  )
```

```
## Warning in min(diff(unique_loc)): no non-missing arguments to min; returning
## Inf
```



```r
# Advanced Degree vs. Some College
set.seed(125)
ad_vs_sc_df <- sal_growth %>%
  filter(degree %in% c("advanced_degree", "some_college"))

ad_vs_sc_null <- ad_vs_sc_df %>%
  specify(growth_rate ~ degree) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 10000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("advanced_degree", "some_college"))

ad_vs_sc_obs_stat <- ad_vs_sc_df %>%
  specify(growth_rate ~ degree) %>%
  calculate(stat = "diff in means", order = c("advanced_degree", "some_college"))
```

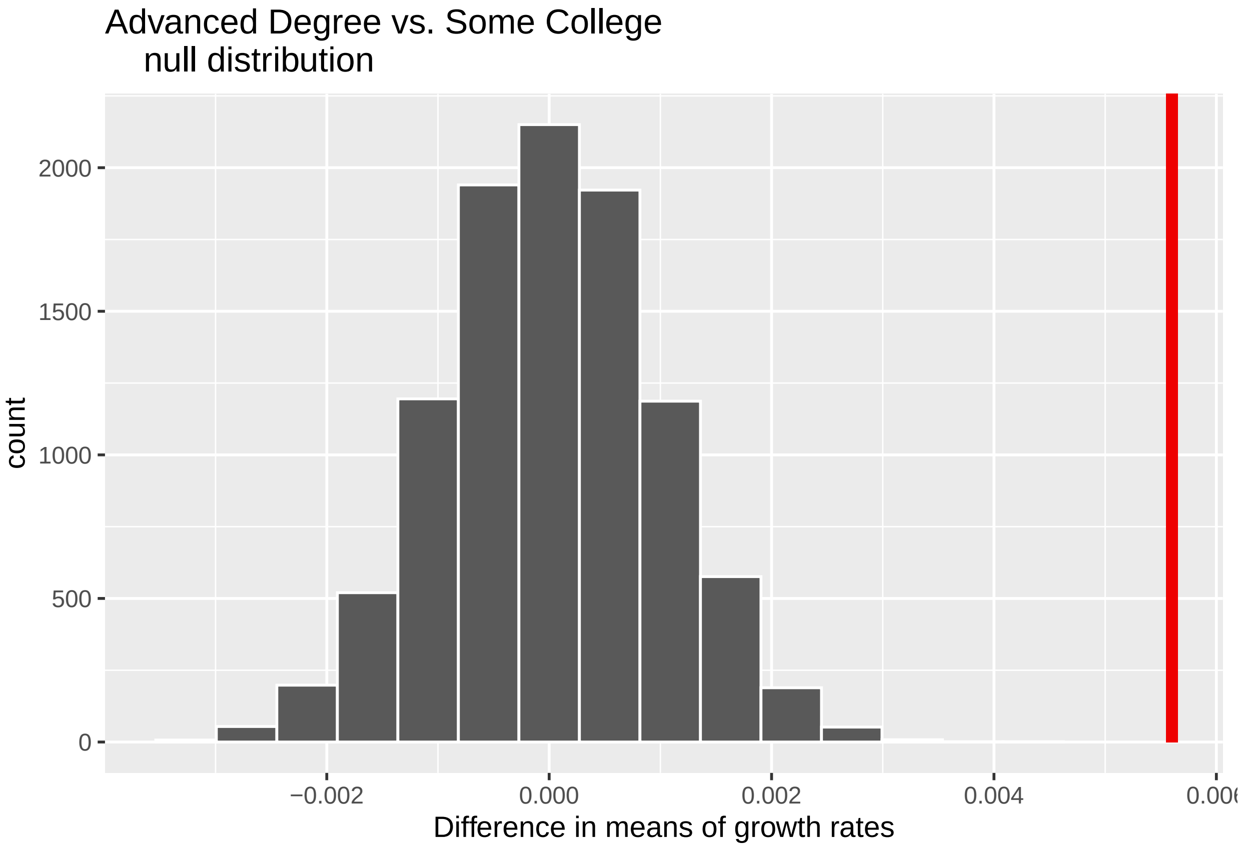
```r
ad_vs_sc_null %>%
  get_p_value(obs_stat = ad_vs_sc_obs_stat, direction = "right")
```

```
## Warning: Please be cautious in reporting a p-value of 0. This result is an
## approximation based on the number of `reps` chosen in the `generate()` step.
## See `?get_p_value()` for more information.
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1       0
```

```r
set.seed(125)
ad_vs_sc_null %>%
  visualize() +
  shade_p_value(obs_stat = ad_vs_sc_obs_stat, direction = "right")+
  labs(
    title = "Advanced Degree vs. Some College
    null distribution",
    x= "Difference in means of growth rates"
  )
```

```
## Warning in min(diff(unique_loc)): no non-missing arguments to min; returning
## Inf
```



```
# Bachelor's Degree vs. Some College
set.seed(126)
bd_vs_sc_df <- sal_growth %>%
  filter(degree %in% c("bachelors_degree", "some_college"))

bd_vs_sc_null <- bd_vs_sc_df %>%
  specify(growth_rate ~ degree) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 10000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("bachelors_degree", "some_college"))

bd_vs_sc_obs_stat <- bd_vs_sc_df %>%
  specify(growth_rate ~ degree) %>%
  calculate(stat = "diff in means", order = c("bachelors_degree", "some_college"))
```

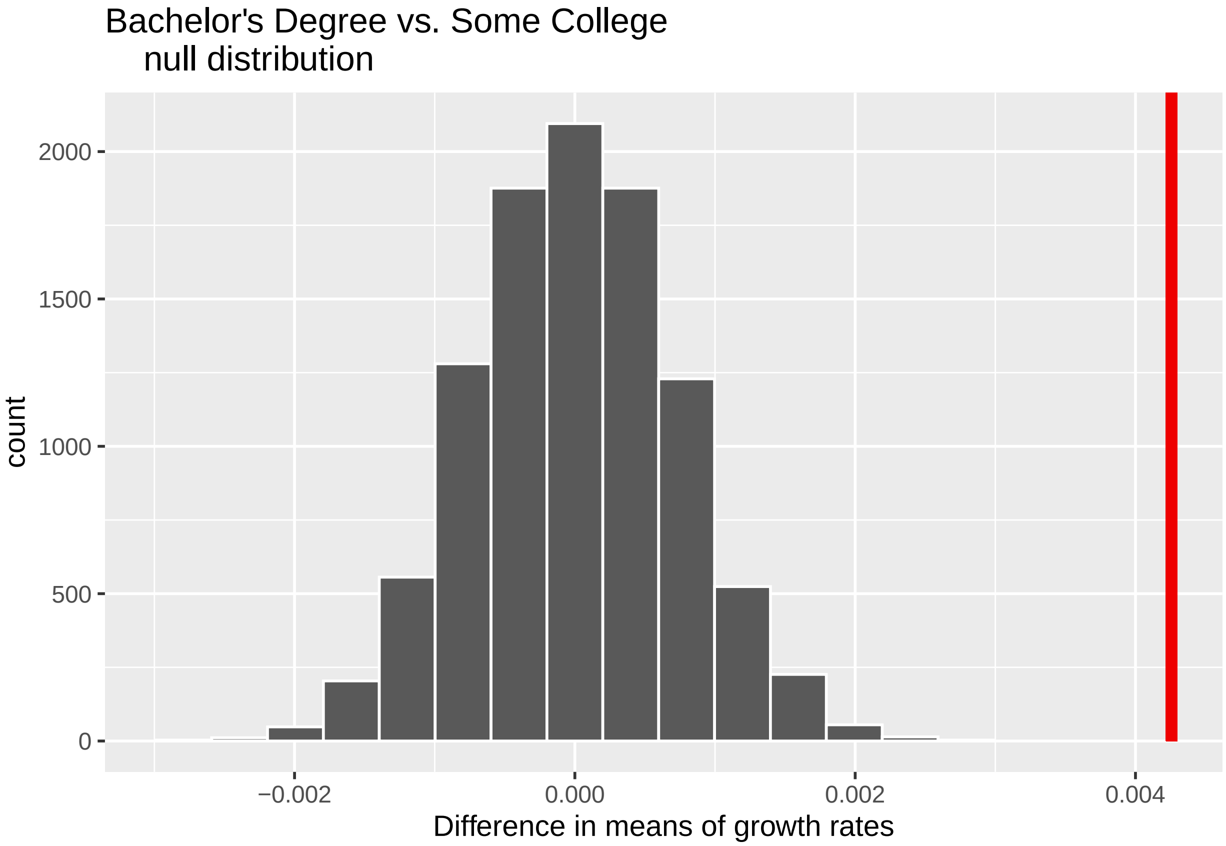
```
bd_vs_sc_null %>%
  get_p_value(obs_stat = bd_vs_sc_obs_stat, direction = "right")
```

```
## Warning: Please be cautious in reporting a p-value of 0. This result is an
## approximation based on the number of ‘reps‘ chosen in the ‘generate()‘ step.
## See ‘?get_p_value()‘ for more information.
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1        0
```

```
set.seed(126)
bd_vs_sc_null %>%
  visualize() +
  shade_p_value(obs_stat = bd_vs_sc_obs_stat, direction = "right")+
  labs(
    title = "Bachelor's Degree vs. Some College
    null distribution",
    x= "Difference in means of growth rates"
  )
```

```
## Warning in min(diff(unique_loc)): no non-missing arguments to min; returning
## Inf
```



```
set.seed(127)
# Bachelor's Degree vs. High School
bd_vs_hs_df <- sal_growth %>%
  filter(degree %in% c("bachelors_degree", "high_school"))

bd_vs_hs_null <- bd_vs_hs_df %>%
  specify(growth_rate ~ degree) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 10000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("bachelors_degree", "high_school"))
```

```
bd_vs_hs_obs_stat <- bd_vs_hs_df %>%
  specify(growth_rate ~ degree) %>%
  calculate(stat = "diff in means", order = c("bachelors_degree", "high_school"))
```

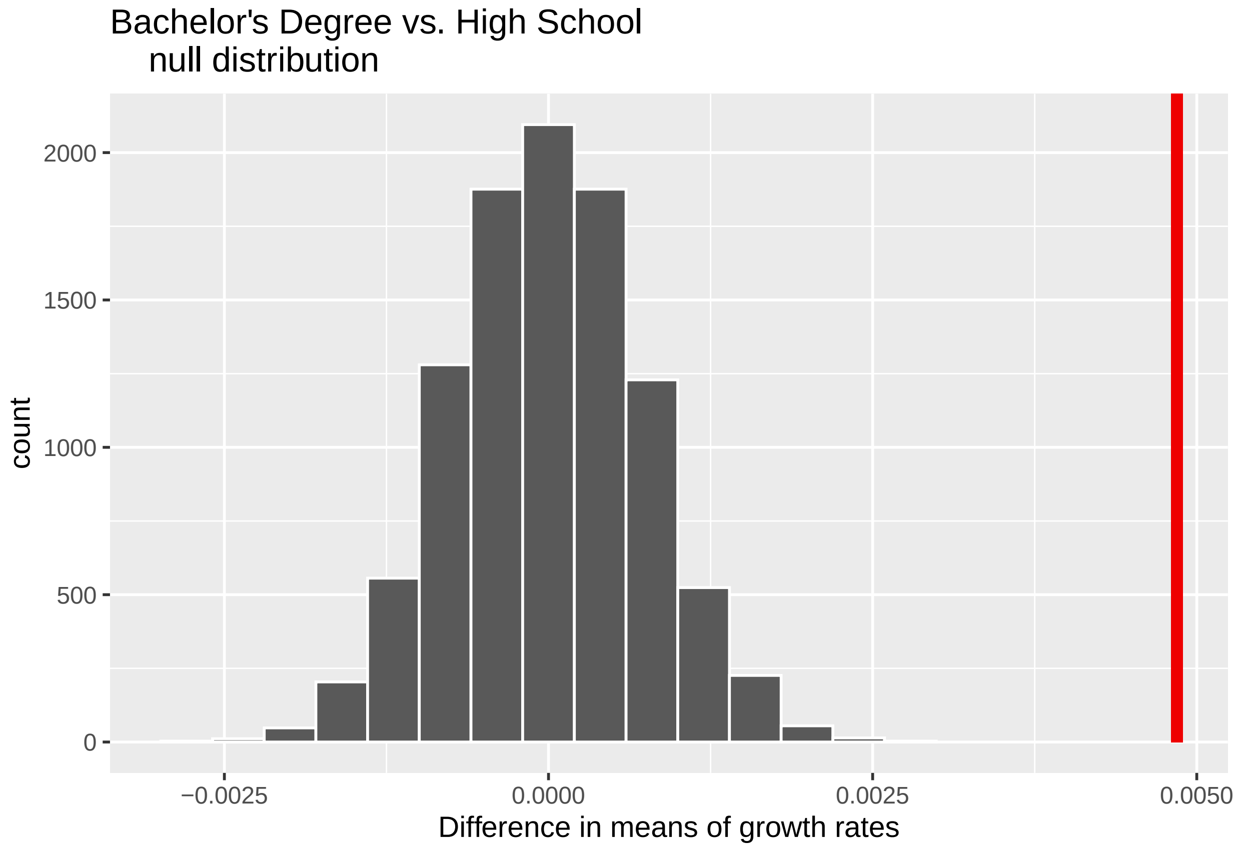
```
bd_vs_hs_null %>%
  get_p_value(obs_stat = bd_vs_hs_obs_stat, direction = "right")
```

```
## Warning: Please be cautious in reporting a p-value of 0. This result is an
## approximation based on the number of ‘reps‘ chosen in the ‘generate()‘ step.
## See ‘?get_p_value()‘ for more information.
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1       0
```

```
set.seed(127)
bd_vs_sc_null %>%
  visualize() +
  shade_p_value(obs_stat = bd_vs_hs_obs_stat, direction = "right")+
  labs(
    title = "Bachelor's Degree vs. High School
    null distribution",
    x= "Difference in means of growth rates"
  )
```

```
## Warning in min(diff(unique_loc)): no non-missing arguments to min; returning
## Inf
```


Bachelor's Degree vs. High School
null distribution
count
2000
1500
1000
500
0
−0.0025
0.0000
0.0025
0.0050
Difference in means of growth rates