

EDA on County Policy Data

CMU MSP Team

3/20/2021

Documentation

- For the first part of the EDA, we focus on the latest cumulative deaths on **2021/02/22**
- To measure to what extent each policy was carried out in a county:
 - since we find that all schools in the same district have same public health strategies
 - for each county, we weight the count of policies by `#students` enrolled in a district/`#students` enrolled in a county
- For the second part of the EDA, we analyze the covid data on the schools' open days only (From county-level `open_date_grouped` to 2020/12/15)

Naming Conventions

- OH_K12: cleaned ohio k12 data
 - `#enrollment_1`: number of students enrolled in each district
- cases: ohio county level cases data
- county_w_dates
- county_policy: policy counts proportion to `#enrollment` at county-level
- county_data: policy data on schools' open days

Check Data

Read data

```
# libraries
library(tidyverse)
library(readxl)
library(ggplot2)
library(reshape2)
library(wesanderson)
pal <- wes_palette(name = "GrandBudapest1", type = "discrete")
# read in data
#source("step2_data_wrangle.R")
source("step3_bracket_data.R")
```

Check data quality

```
# county-wise total deaths = sum of new deaths until Feb.2021
total_deaths_by_county <- cases %>%
  group_by(COUNTY) %>%
  summarise(total_deaths = sum(NEWDEATHS, na.rm = TRUE), .groups = "drop")
```

```
# latest collection date
tail(sort(unique(cases$DATE)))
```

```
## [1] "2021-02-17" "2021-02-18" "2021-02-19" "2021-02-20" "2021-02-21"
## [6] "2021-02-22"
```

```
# latest cumulative death
latest_cumdeaths <- cases%>%
  filter(DATE == '2021-02-22')%>%
  select(COUNTY, CUMDEATHS)%>%
  unique()
```

```
# check if the latest death matches with sum of new deaths
sum(total_deaths_by_county$total_deaths != latest_cumdeaths$CUMDEATHS)
```

```
## [1] 0
```

```
cbind(total_deaths_by_county,latest_cumdeaths)%>%head(3)
```

```
##      COUNTY total_deaths  COUNTY CUMDEATHS
## 1    ADAMS           32    ADAMS         32
## 2    ALLEN          231    ALLEN         231
## 3 ASHLAND           93 ASHLAND          93
```

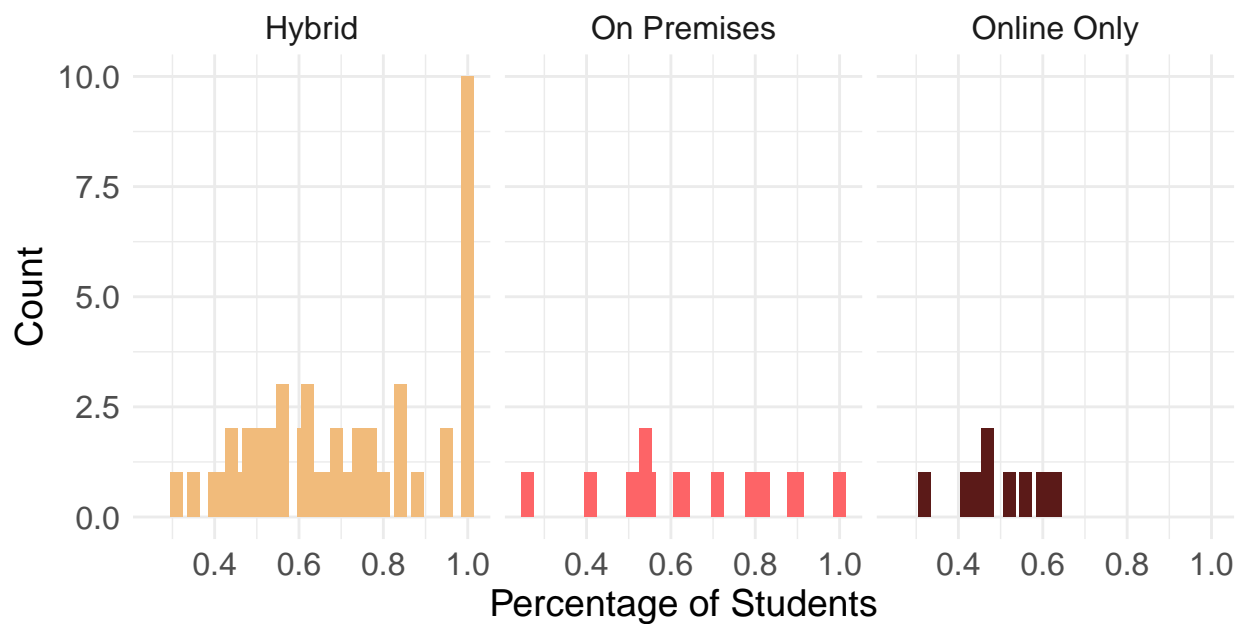
generate is online method

```
isonline_enroll <- teachingmethod_enroll%>%
  filter(teachingmethod != 'Other'&teachingmethod != 'Pending'&teachingmethod != 'Unknown')%>%
  mutate(is_online = ifelse(teachingmethod == "Online Only","Online Only","Not Online Only"))%>%
  group_by(county,is_online) %>%
  mutate(prop_online_only = sum(prop_teachingmethod))%>%
  group_by(county)%>%
  slice(which.max(prop_online_only))%>%
  select(county,is_online,prop_online_only)
```

Marginal Distribution of Number of students enrolls by Each Methods

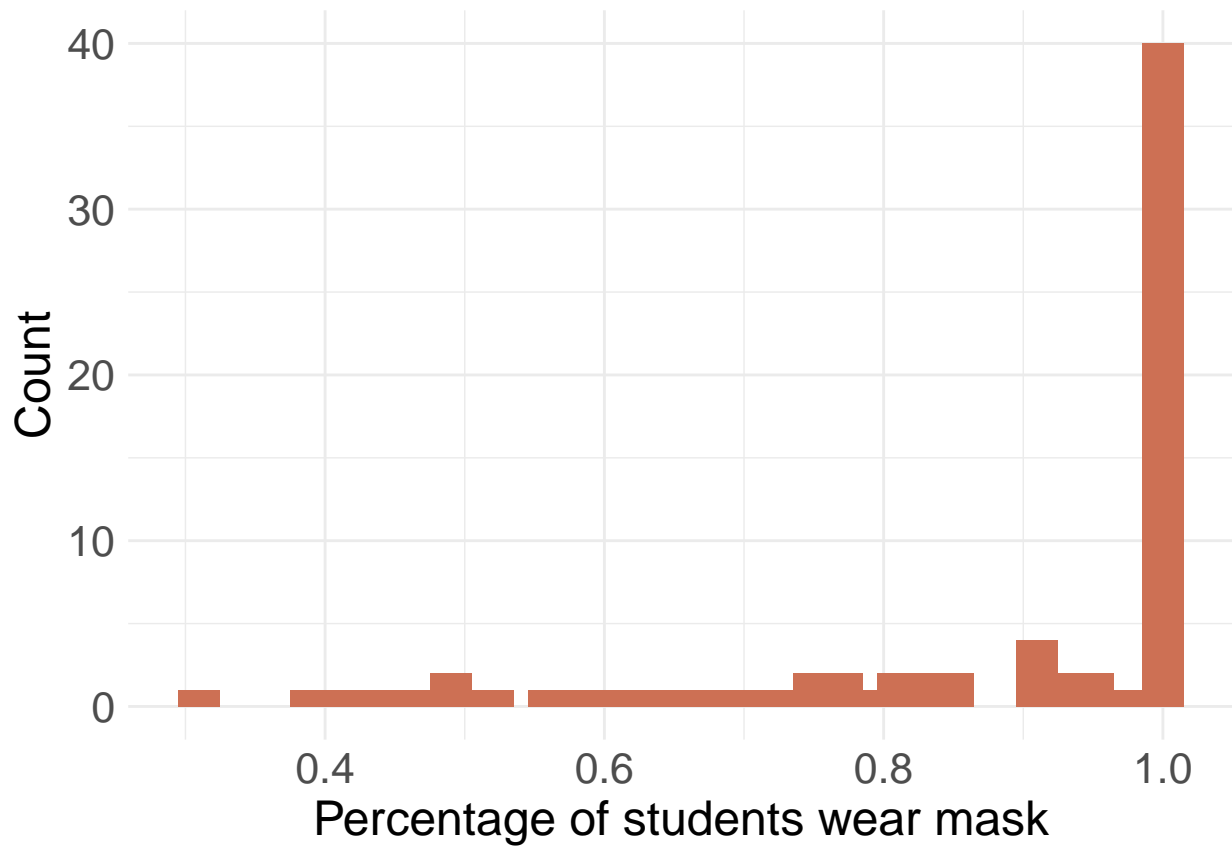
```
long_teaching%>%
  filter(teachingmethod %in% c('Online Only','On Premises','Hybrid'))%>%
  group_by(county)%>%
  slice(which.max(prop_teachingmethod))%>%
  ggplot(aes(x=prop_teachingmethod,group=teachingmethod,fill=teachingmethod))+geom_bar(stat = 'count',w
```

Conditional Distribution of Percentages of Students by Most Common Teaching Method

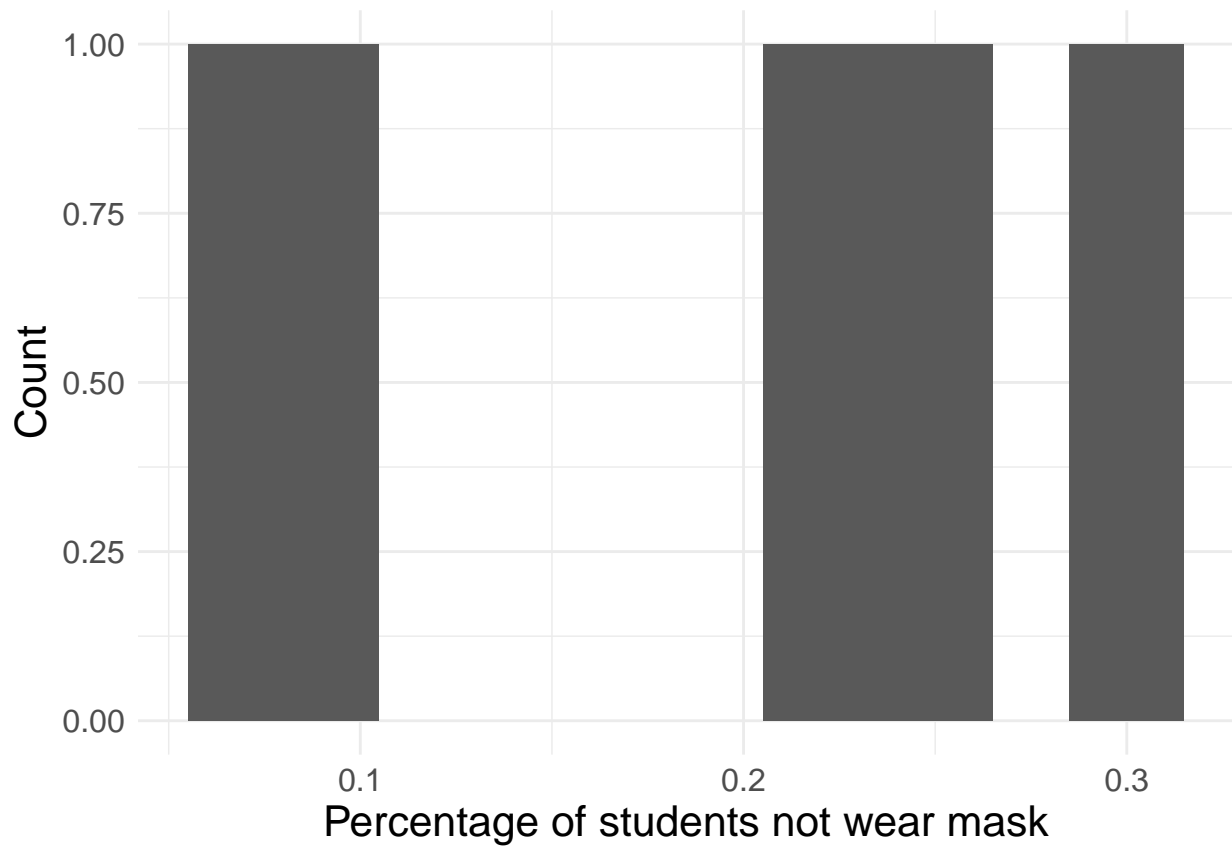


Majority Teaching Method ■ Hybrid ■ On Premises ■ Online Only

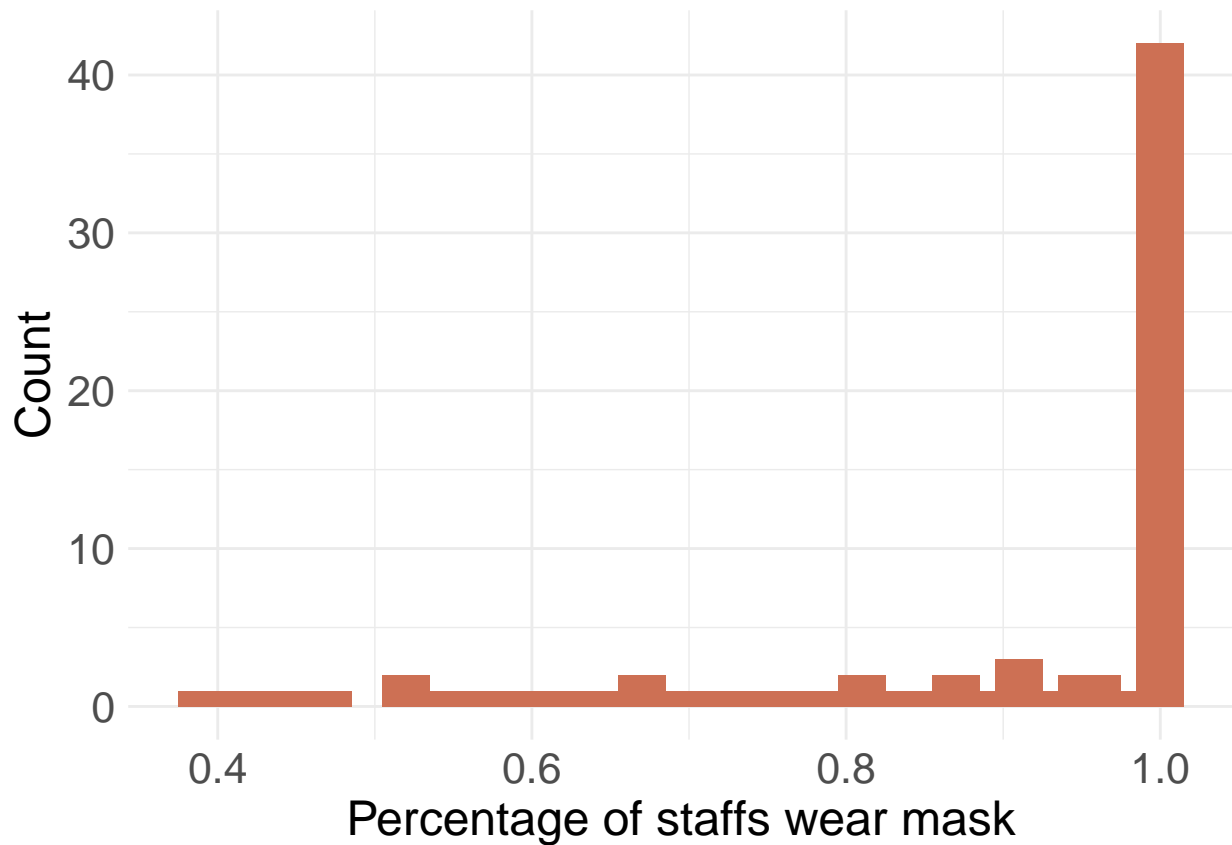
```
studentmask_enroll%>%
  filter(studentmaskpolicy=='Required for all students')%>%
  ggplot(aes(y= prop_student_mask))+geom_bar(stat = 'count',width=0.03,fill='salmon3')+labs(x = "Count")
```



```
studentmask_enroll%>%
  filter(studentmaskpolicy=='Not required')%>%
  ggplot(aes(y= prop_student_mask))+geom_bar(stat = 'count',width=0.03)+labs(x = "Count",y="Percentage of students wear mask")
```



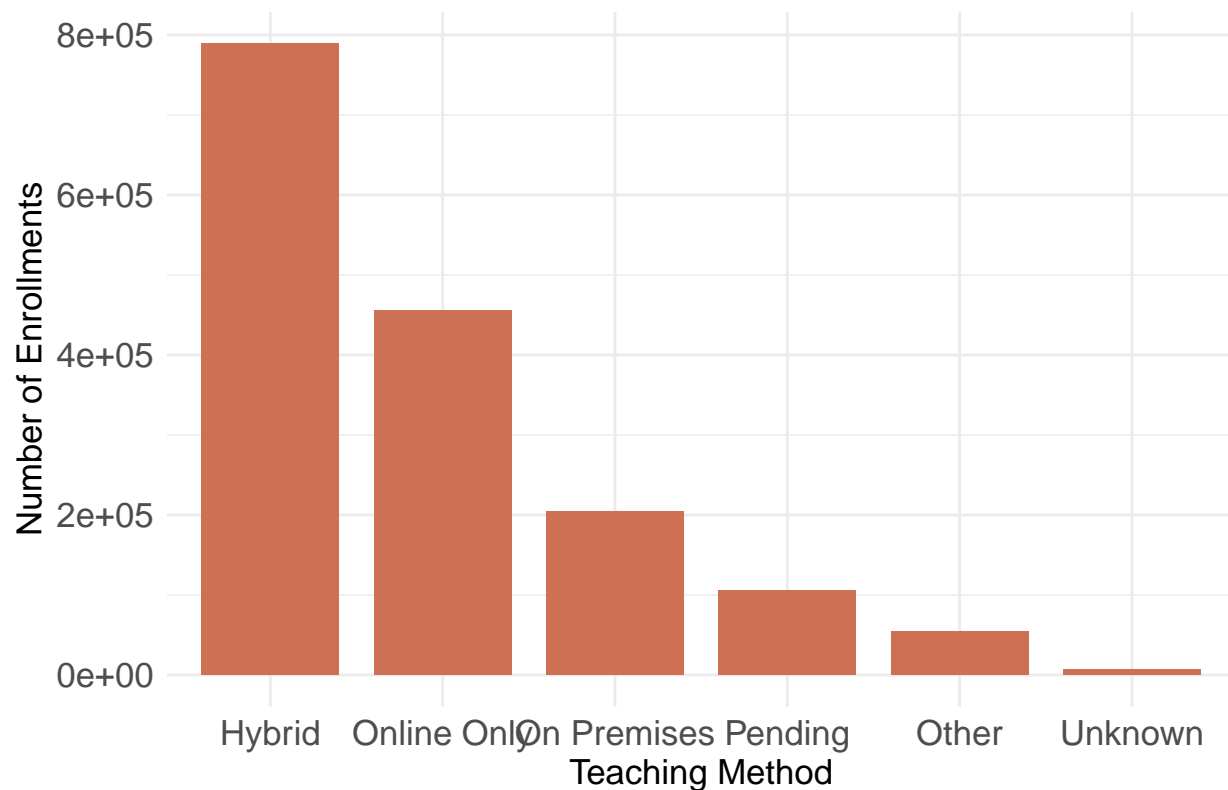
```
staffmask_enroll%>%  
  filter(staffmaskpolicy=='Required for all staff')%>%  
  ggplot(aes(y= prop_staff_mask))+geom_bar(stat = 'count',width=0.03,fill='salmon3')+labs(x = "Count",y=
```



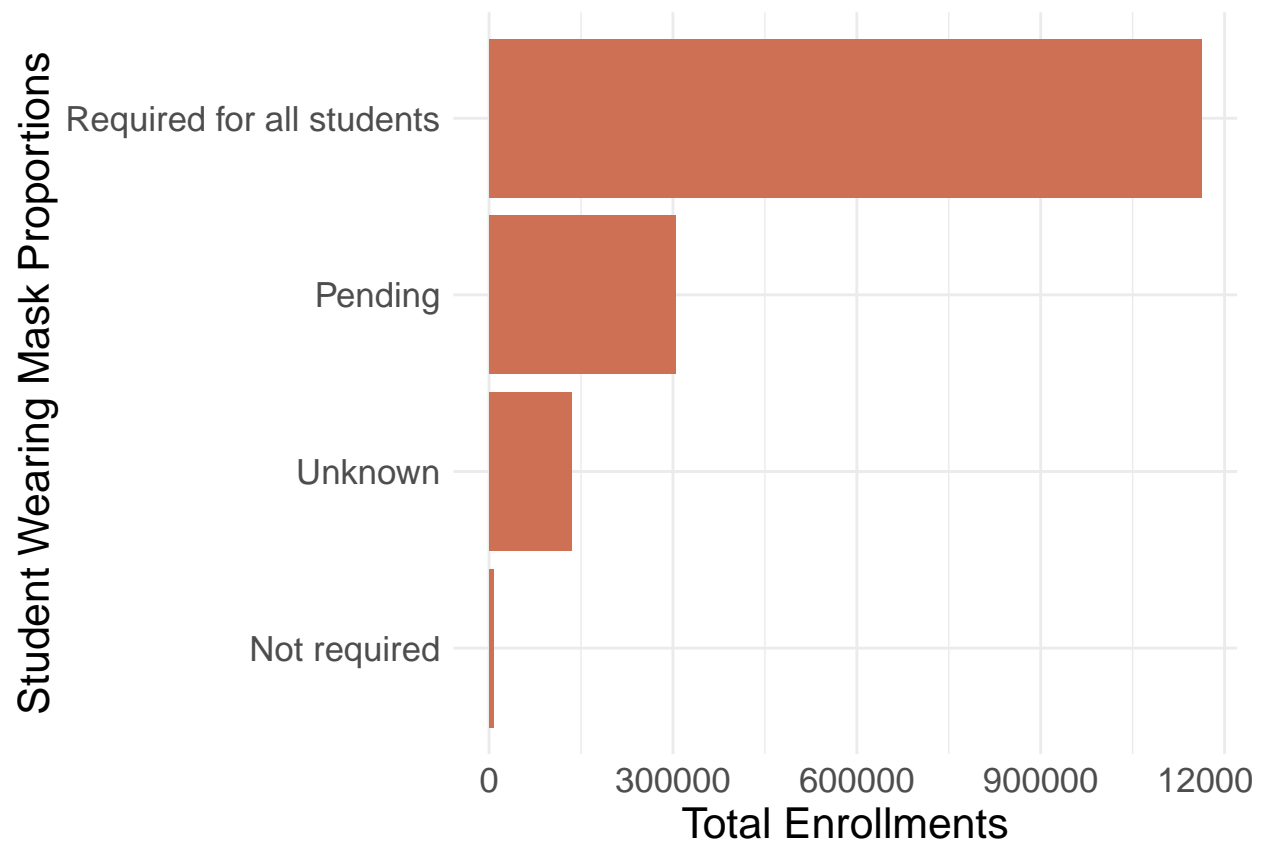
Conditional Distribution of Number of students enrolls by Each Methods

```
teachingmethod_enroll%>%
  group_by(teachingmethod)%>%
  summarise(total_teaching= sum(total_teachingmethod))%>%
  ggplot(aes(x= reorder(teachingmethod,-total_teaching),y = total_teaching))+geom_bar(stat = 'identity')
```

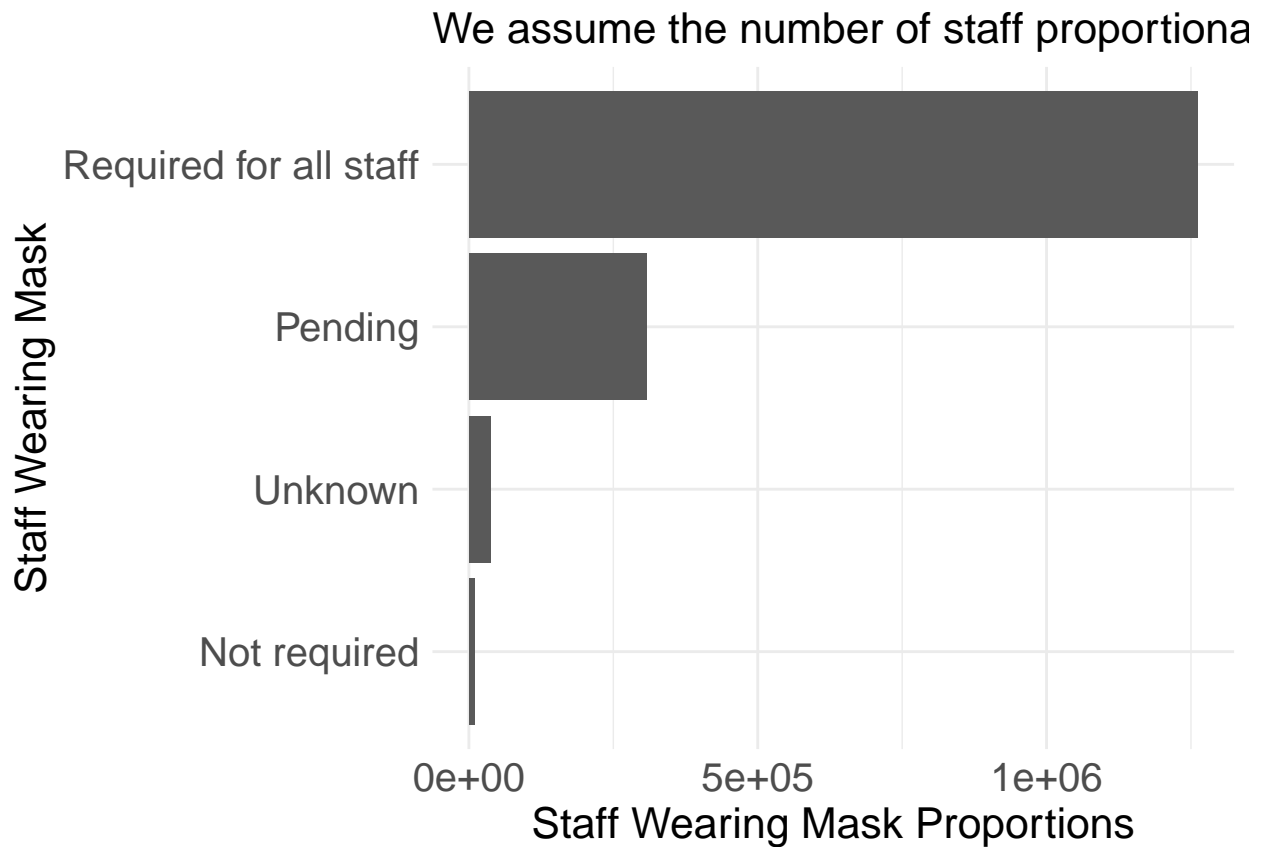
Distribution of Student Enrollments in Ohio by Teaching Method



```
studentmask_enroll1%>%
  group_by(studentmaskpolicy)%>%
  summarise(total_studentmask= sum(n_studentmask))%>%
  ggplot(aes(x= reorder(studentmaskpolicy,total_studentmask),y=total_studentmask))+geom_bar(stat = 'i
```



```
staffmask_enroll%>%
  group_by(staffmaskpolicy)%>%
  summarise(total_staffmask= sum(n_staffmask))%>%
  ggplot(aes(x= reorder(staffmaskpolicy,total_staffmask),y = total_staffmask))+geom_bar(stat = 'identity')
```

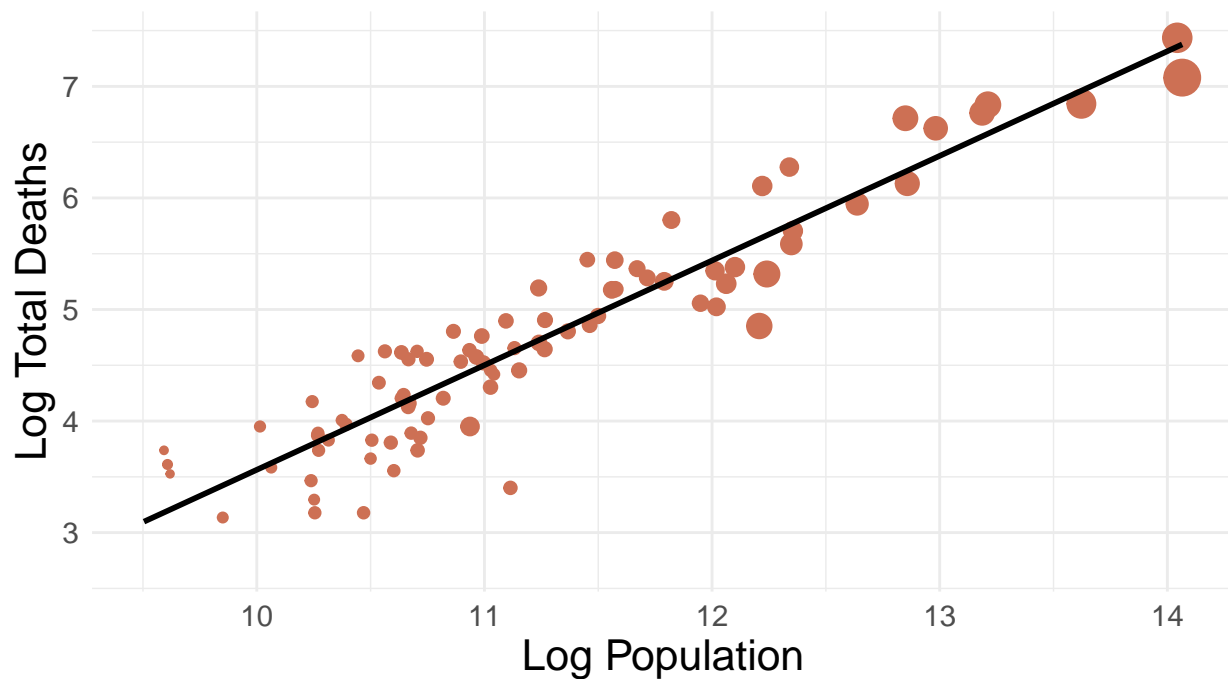



Conditional Distribution of Death Proportions

Marginal Distribution of Death v.s. Population

```
# look at influential points, size = total_county_enroll
county_policy_wide%>%
  ggplot(aes(x = log(POPULATION), y = log(CUMDEATHS), size=county_enroll)) +
  geom_point(na.rm = TRUE, colour = 'salmon3') +
  geom_smooth(method = "lm", formula = y~x, se = FALSE, na.rm = TRUE, show.legend = FALSE, colour = 'black') +
  theme_minimal() + theme(legend.position = "bottom") +
  labs(x = "Log Population", y = "Log Total Deaths",
       size = "Number of Enrollments", title = "Death against Population") + theme(text = element_text(size = 12))
```

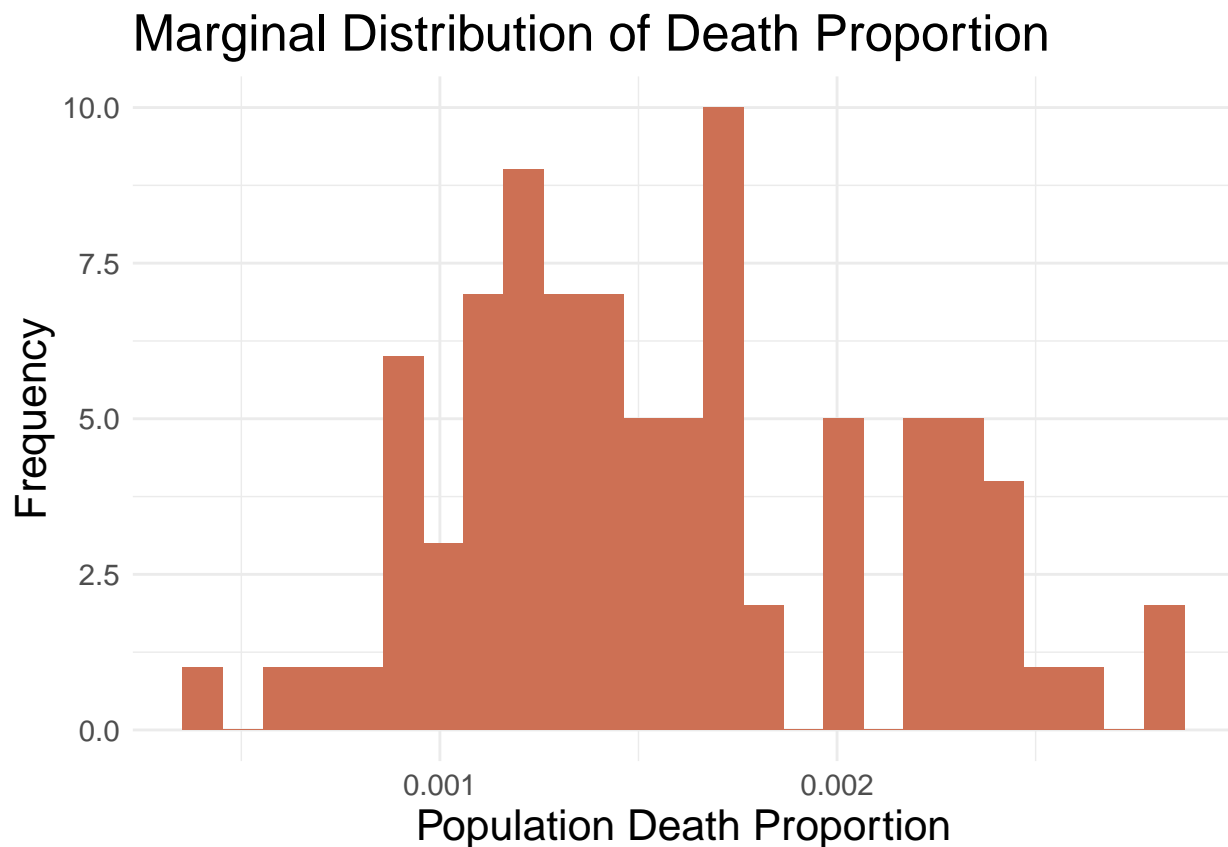
Death against Population



Number of Enrollments ● 50000 ● 100000 ● 150000

```
ggsave("DeathVSPopulation.jpg", height = 5, width = 7)
```

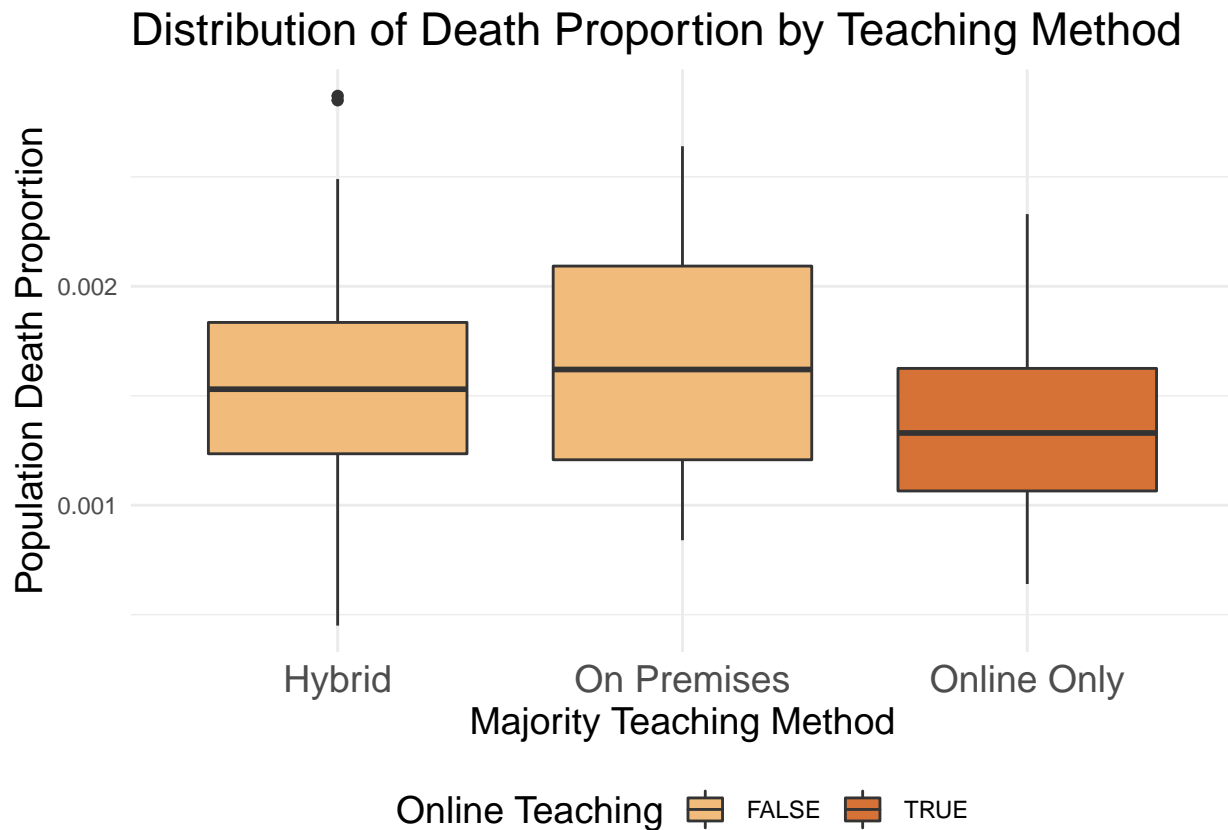
```
county_policy_wide %>%
  ggplot(aes(x = death_prop)) +
  geom_histogram(bins = 25, na.rm = TRUE, fill = 'salmon3') +
  theme_minimal() +
  labs(x = "Population Death Proportion", y = "Frequency",
       title = "Marginal Distribution of Death Proportion") + theme(text = element_text(size=14), title = c
```



```
ggsave("MargDeathProp.jpg", height = 5, width = 7)
```

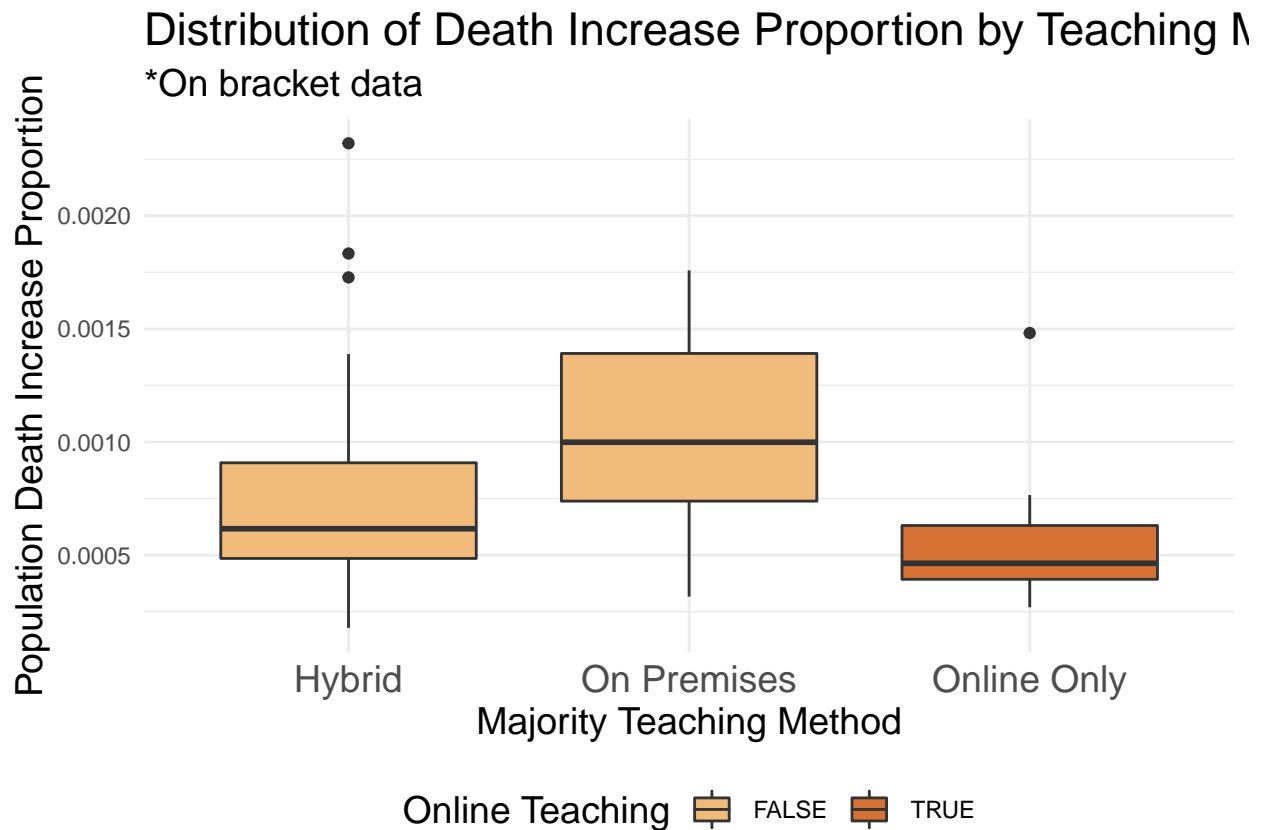
Conditional distribution of death proportions by Teaching Method

```
county_policy_wide %>%
  drop_na() %>%
  ggplot(aes(y = death_prop, x = major_teaching, fill = major_teaching == 'Online Only')) +
  geom_boxplot() + scale_fill_manual(values = pal[c(1,4)]) +
  theme_minimal() +
  theme(axis.text.x = element_text(size = 14), title = element_text(size = 14), legend.position = "bottom") +
  labs(y = "Population Death Proportion",
       title = "Distribution of Death Proportion by Teaching Method",
       x = "Majority Teaching Method", fill = 'Online Teaching')
```



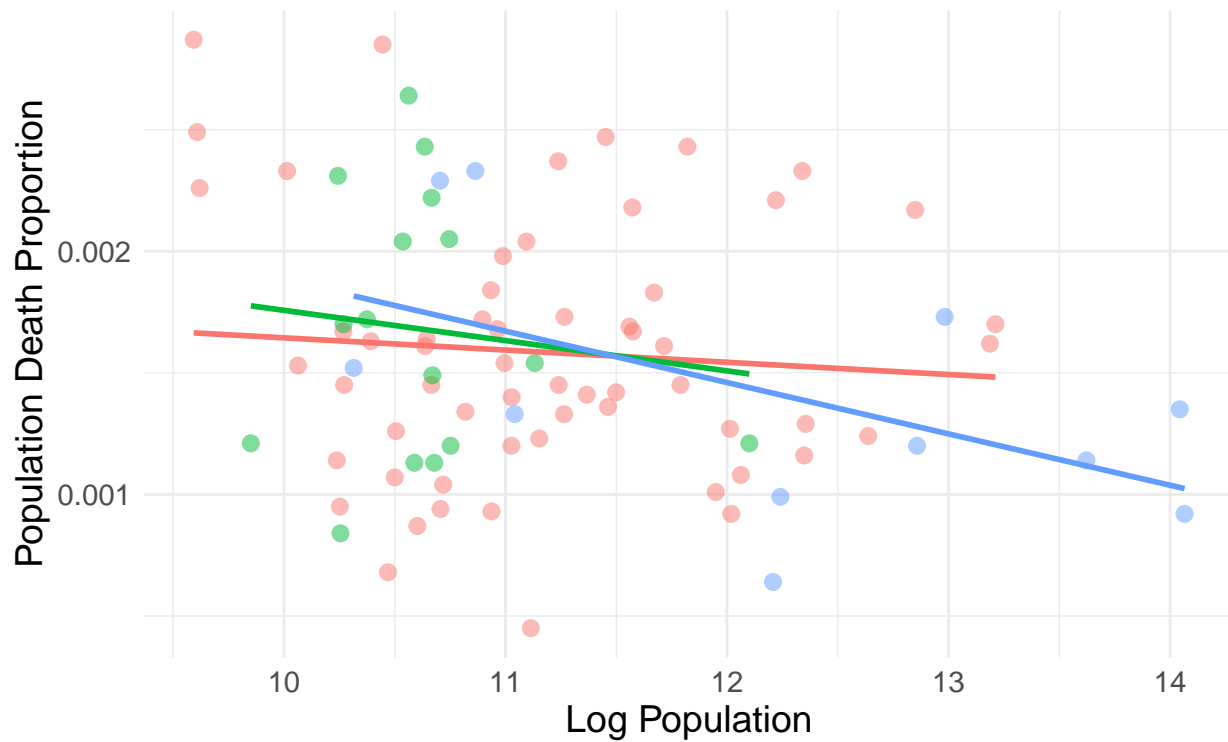
```
ggsave("DeathProp_teachingmethod.jpg", height = 5, width = 7)

bracket_county_policy_wide %>%
  drop_na()%>%
  ggplot(aes(y = deaths_increase_prop, x = major_teaching, fill = major_teaching == 'Online Only')) +
  geom_boxplot() + scale_fill_manual(values = pal[c(1,4)]) +
  theme_minimal() +
  theme(axis.text.x = element_text(size = 14), title = element_text(size = 14), legend.position = "bottom")
labs(y = "Population Death Increase Proportion",
     title = "Distribution of Death Increase Proportion by Teaching Method", subtitle = '*On bracket c',
     x = "Majority Teaching Method", fill = 'Online Teaching')
```



```
ggsave("DeathProp_bracket_teachingmethod.jpg", height = 5, width = 7)
```

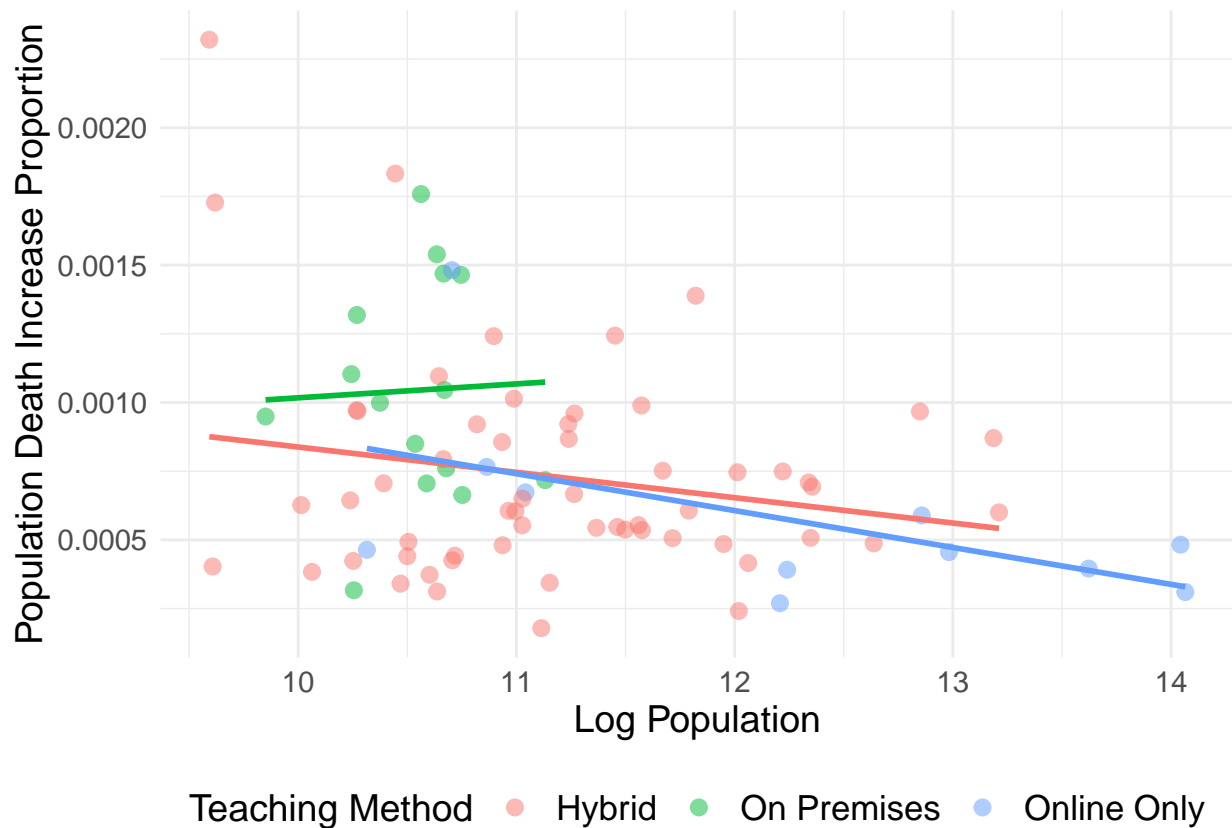
```
county_policy_wide%>%
  drop_na()%>%
  ggplot(aes(x = log(POPULATION), y = death_prop, color = major_teaching)) + geom_point(alpha = 0.5, size = 1) +
  geom_smooth(method = "lm", formula = y~x, se = FALSE, show.legend = FALSE) +
  theme_minimal() +
  theme(axis.text = element_text(size = 11), title = element_text(size = 14), legend.text = element_text(size = 11)) +
  labs(x = "Log Population", y = "Population Death Proportion",
       color = "Teaching Method")
```



Teaching Method ● Hybrid ● On Premises ● Online Only

```
ggsave("DeathProp_onlineonly.jpg", height = 5, width = 7)

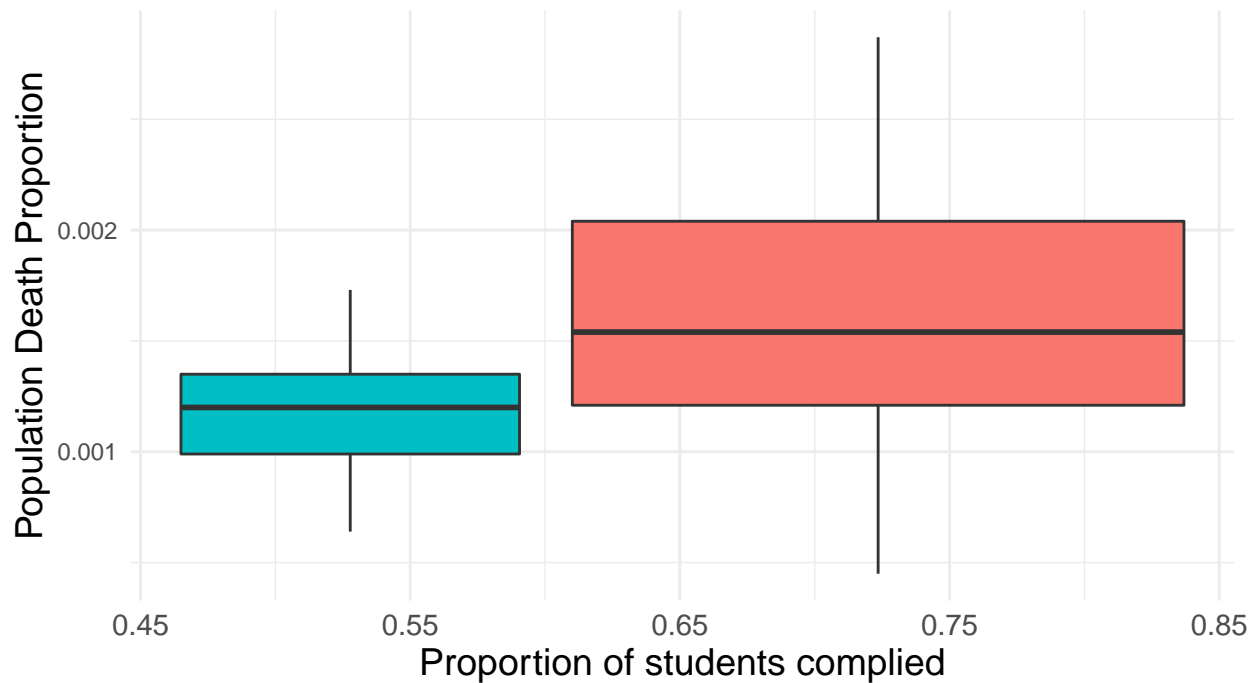
bracket_county_policy_wide%>%
  drop_na()%>%
  ggplot(aes(x = log(POPULATION), y = deaths_increase_prop, color = major_teaching)) + geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", formula = y~x, se = FALSE, show.legend = FALSE) +
  theme_minimal() +
  theme(axis.text = element_text(size = 11), title = element_text(size = 14), legend.text = element_text(size = 11)) +
  labs(x = "Log Population", y = "Population Death Increase Proportion",
       color = "Teaching Method")
```



Conditional distribution of death proportions by Online Teaching

```
county_policy_wide %>%
  right_join(isonline_enroll, by='county') %>%
  ggplot(aes(y = death_prop, x = prop_online_only, fill= is_online)) +
  geom_boxplot() +
  theme_minimal() +
  theme(axis.text.x = element_text(size = 11), title = element_text(size = 14), legend.position = "bottom")
labs(y = "Population Death Proportion",
     x = "Proportion of students complied",
     title = "Distribution of Death Proportion by Online Method",
     fill = "Majority Teaching Method")
```

Distribution of Death Proportion by Online Method



Majority Teaching Method ■ Not Online Only ■ Online Only

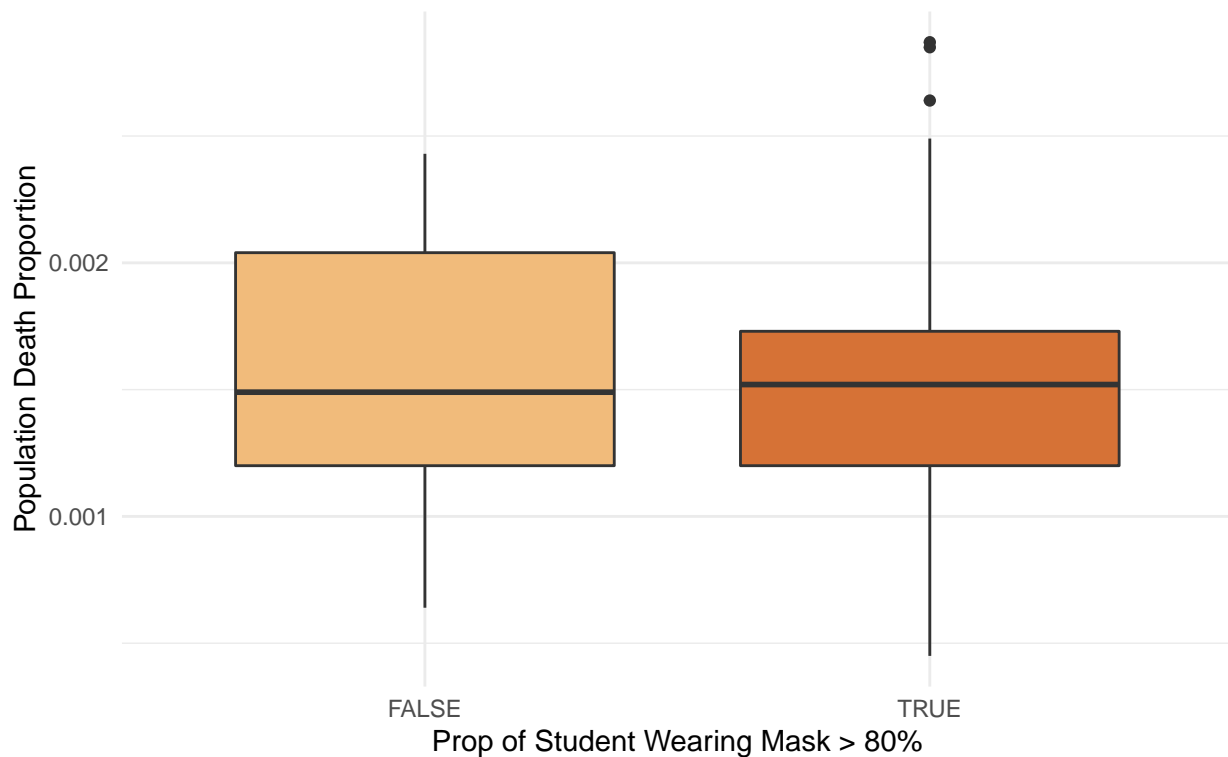
```
ggsave("DeathProp_online.jpg", height = 5, width = 7)
```

Conditional distribution of death proportions by Mask Wearing

```
long_studentmask %>%
  filter(!studentmaskpolicy%in%c('Pending','Unknown'))%>%
  group_by(county)%>%
  slice(which.max(prop_student_mask))%>%
  ggplot(aes(y = death_prop,x = (prop_student_mask>0.8),fill = (prop_student_mask>0.8))) +
  geom_boxplot() + scale_fill_manual(values = pal[c(1,4)])+
  theme_minimal() +
  labs(y = "Population Death Proportion", x='Prop of Student Wearing Mask > 80%', title = "Distribution
```


Distribution of Death Proportion by Student Wearing Mask

* Majority of students are required to wear mask for all counties

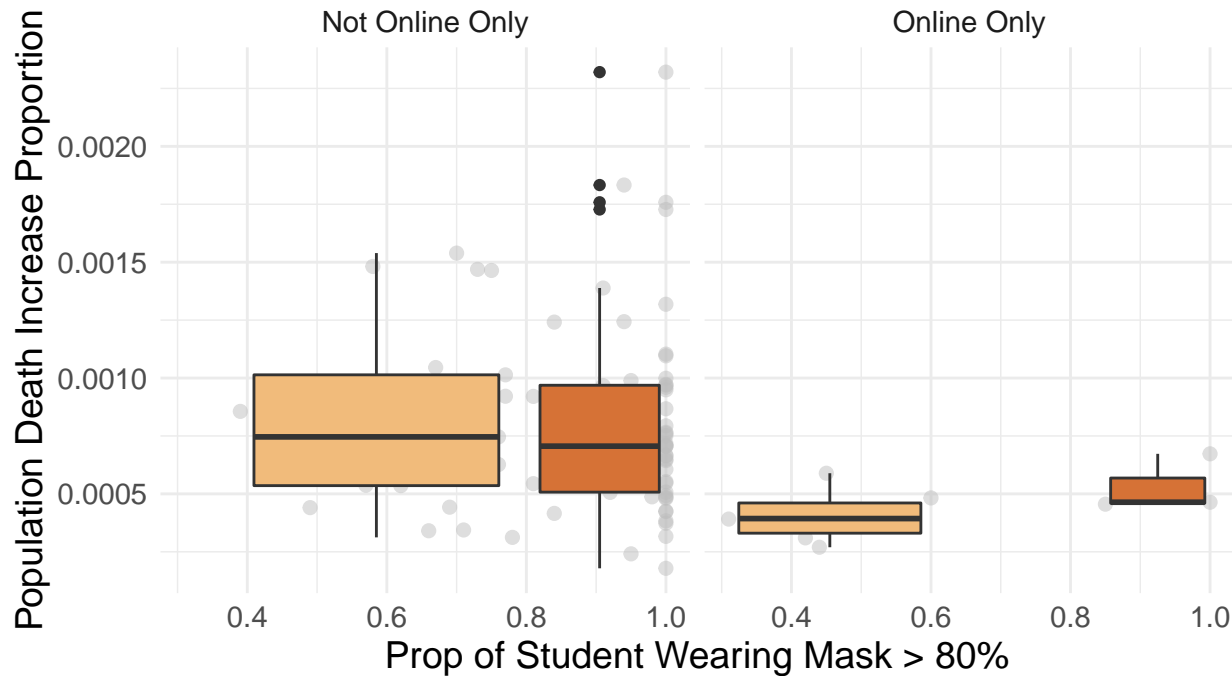


```
ggsave("DeathProp_student_mask.jpg", height = 5, width = 7)
```

```
long_bracket_isonline_mask%>%  
  drop_na()%>%  
  mutate(over_mask = prop_student_mask>0.8)%>%  
  ggplot(aes(y = deaths_increase_prop,x=prop_student_mask,group = over_mask,fill = over_mask)) +  
  geom_point(alpha=0.5,size=2,color='grey')+  
  geom_boxplot() + scale_fill_manual(values = pal[c(1,4)])+  
  theme_minimal() + facet_wrap(~is_online)+  
  labs(y = "Population Death Increase Proportion", x='Prop of Student Wearing Mask > 80%', title = "Dis
```

Distribution of Death Increase Proportion by Mask Wearing for In-person Counties

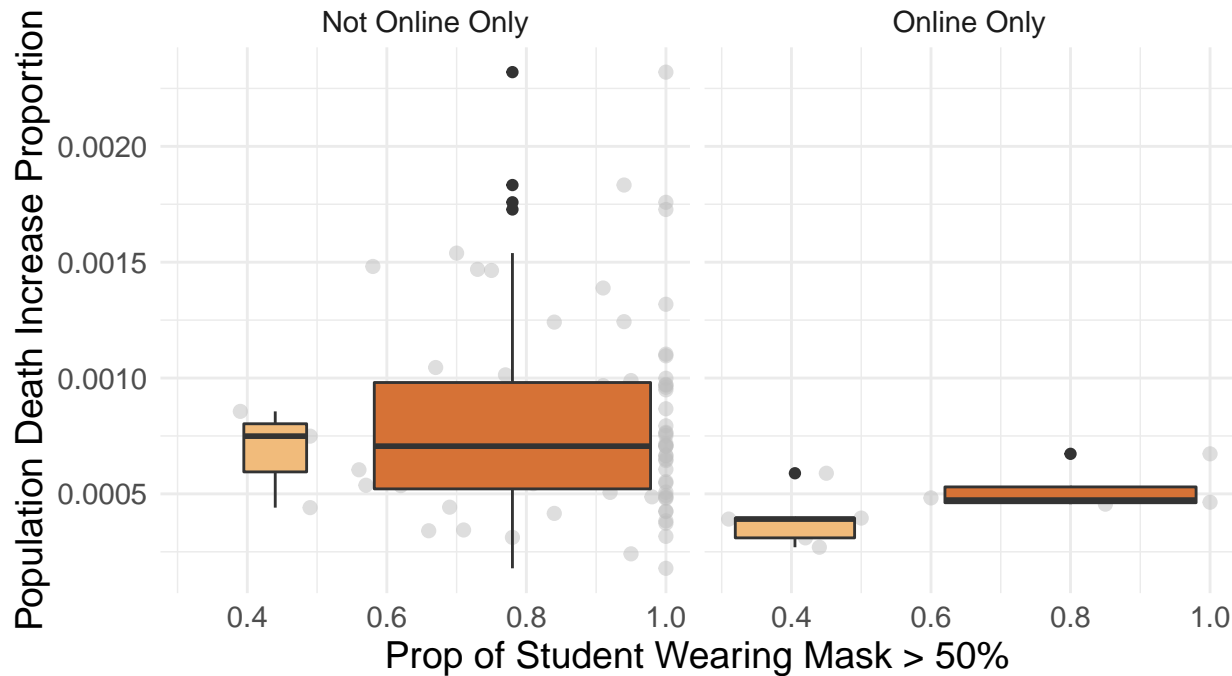
* Majority of students are required to wear mask for all counties



```
long_bracket_isonline_mask%>%
  drop_na()%>%
  mutate(over_mask = prop_student_mask>0.5)%>%
  ggplot(aes(y = deaths_increase_prop,x=prop_student_mask,group = over_mask,fill = over_mask)) +
  geom_point(alpha=0.5,size=2,color='grey')+
  geom_boxplot() + scale_fill_manual(values = pal[c(1,4)])+
  theme_minimal() + facet_wrap(~is_online)+
  labs(y = "Population Death Increase Proportion", x='Prop of Student Wearing Mask > 50%', title = "Dis
```

Distribution of Death Increase Proportion by Mask Wearing for In-person Counties

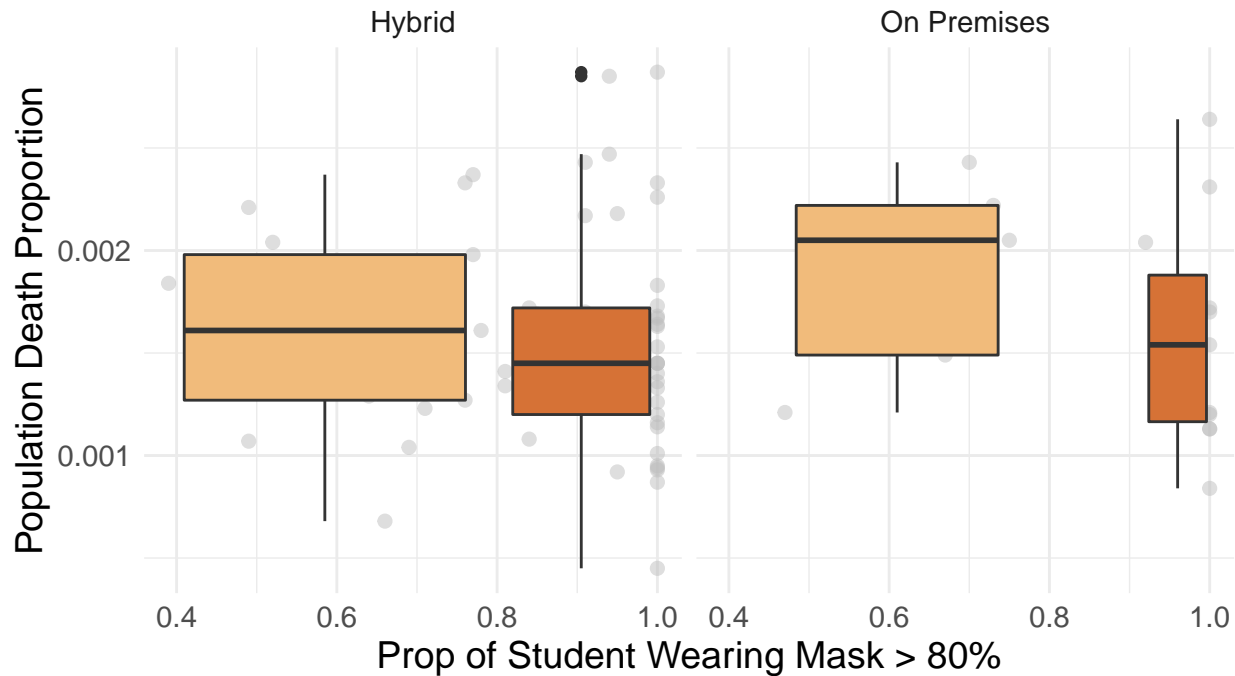
* Majority of students are required to wear mask for all counties



```
long_teaching_mask%>%
  drop_na()%>%
  filter(teachingmethod!='Online Only')%>%
  mutate(over_mask = prop_student_mask>0.8)%>%
  ggplot(aes(y = death_prop,x=prop_student_mask,group = over_mask,fill = over_mask)) +
  geom_point(alpha=0.5,size=2,color='grey')+
  geom_boxplot() + scale_fill_manual(values = pal[c(1,4)])+
  theme_minimal() + facet_wrap(~teachingmethod)+
  labs(y = "Population Death Proportion", x='Prop of Student Wearing Mask > 80%', title = "Distribution
```

Distribution of Death Proportion by Mask Wearing for In-person Counties

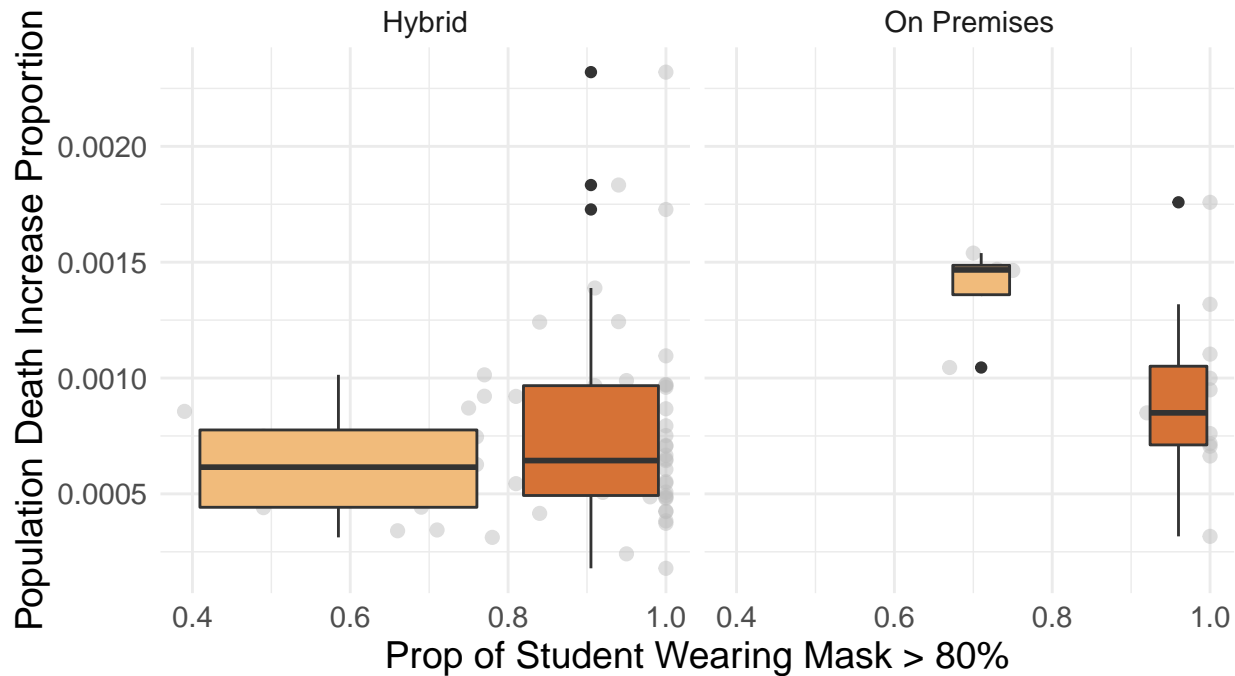
* Majority of students are required to wear mask for all counties



```
long_bracket_teaching_mask%>%
  drop_na()%>%
  filter(teachingmethod!='Online Only')%>%
  mutate(over_mask = prop_student_mask>0.8)%>%
  ggplot(aes(y = deaths_increase_prop,x=prop_student_mask,group = over_mask,fill = over_mask)) +
  geom_point(alpha=0.5,size=2,color='grey')+
  geom_boxplot() + scale_fill_manual(values = pal[c(1,4)])+
  theme_minimal() + facet_wrap(~teachingmethod)+
  labs(y = "Population Death Increase Proportion", x='Prop of Student Wearing Mask > 80%', title = "Dis
```

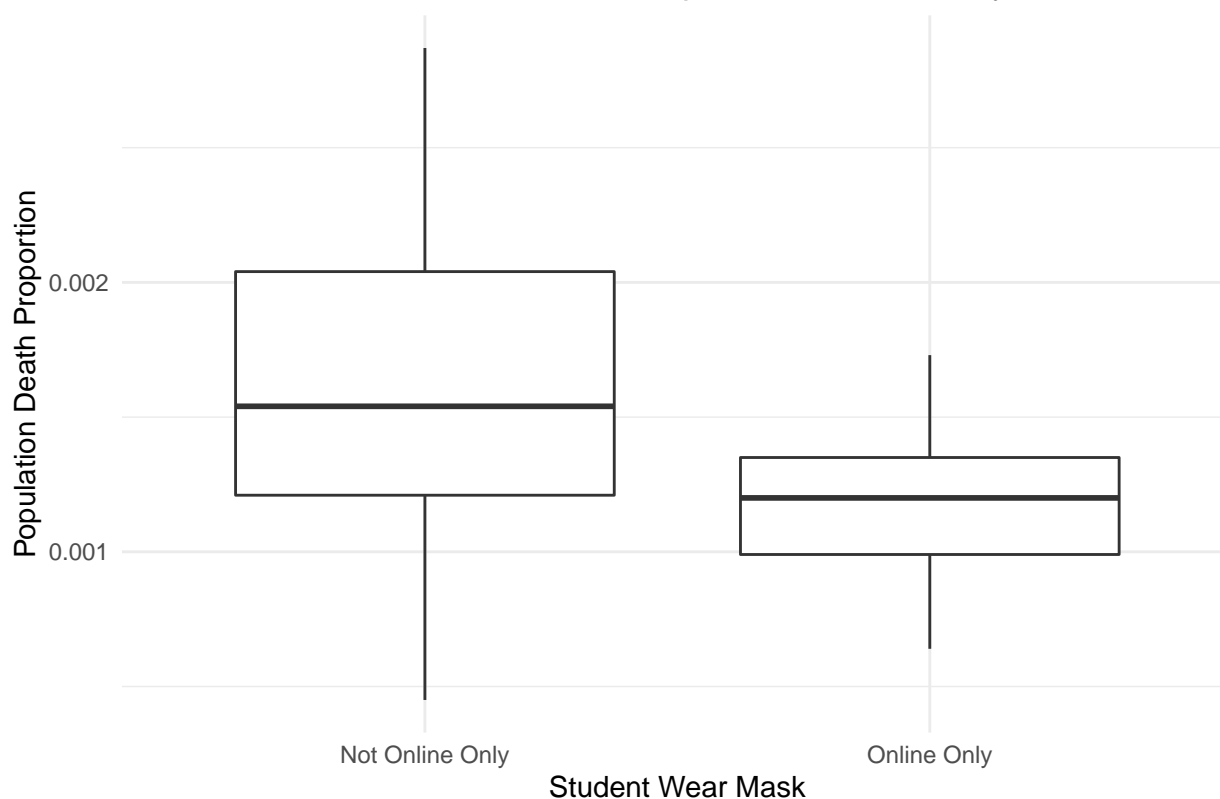
Distribution of Death Increase Proportion by Mask Wearing for In-person Counties

* Majority of students are required to wear mask for all counties



```
isonline_enroll%>%
  left_join(death_prop, by=c('county'='COUNTY'))%>%
  ggplot(aes(x = is_online, y = death_prop)) +
  geom_boxplot(na.rm = TRUE) +
  theme_minimal() +
  labs(x = "Student Wear Mask", y = "Population Death Proportion",
       title = "Conditional Distribution of Death Proportion b Online Only")
```

Conditional Distribution of Death Proportion b Online Only

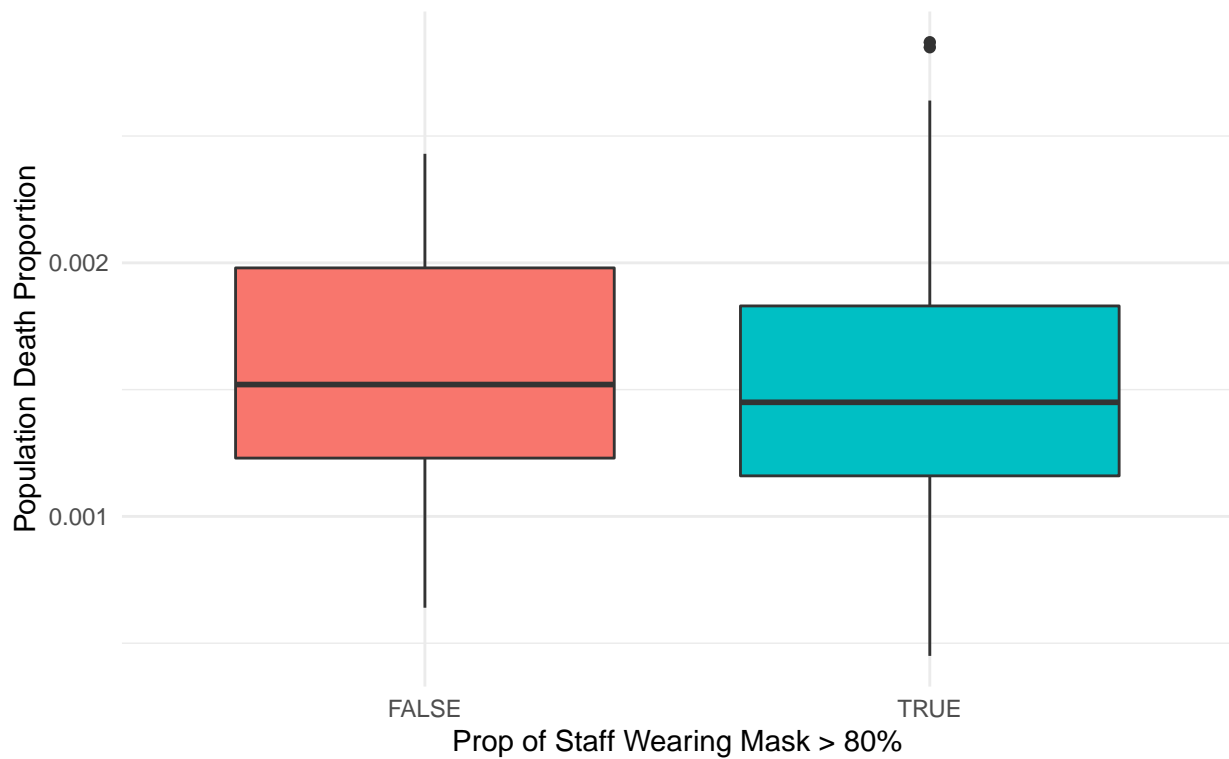


```
ggsave("DeathProp_onlineonly.jpg", height = 5, width = 7)
```

```
long_staff %>%
  filter(!staffmaskpolicy%in%c('Pending', 'Unknown'))%>%
  group_by(county)%>%
  slice(which.max(prop_staff_mask))%>%
  ggplot(aes(y = death_prop, x = (prop_staff_mask>0.8), fill = (prop_staff_mask>0.8))) +
  geom_boxplot() +
  theme_minimal() +
  labs(y = "Population Death Proportion", x='Prop of Staff Wearing Mask > 80%', title = "Distribution o
```

Distribution of Death Proportion by Staff Wearing Mask

* Majority of staffs are required to wear mask for all counties



```
ggsave("DeathProp_staff_mask.jpg", height = 5, width = 7)
```