initial-eda

Lillian Clark

11/8/2020

```
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------- tidyverse 1.3.0 --
```

```
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
## v purrr   0.3.4
```

```
## -- Conflicts ------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(broom)
library(knitr)
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

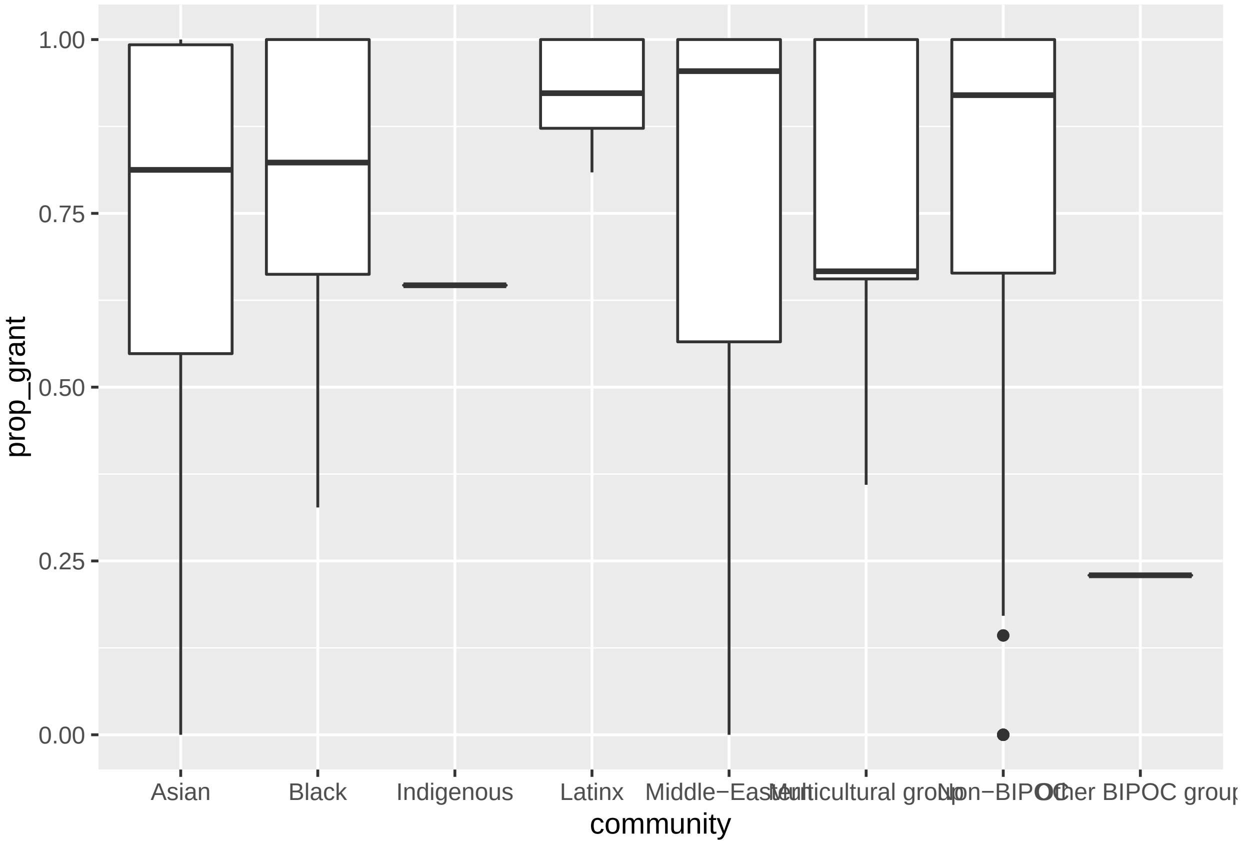
```
prog <- read.csv("data-labeled/programming.csv")
prog <- prog[-1]
budget <- read.csv("data-labeled/budget.csv")
budget <- budget[-1]
```

```
sofc <- read.csv("data-labeled/sofc.csv")
sofc <- sofc[-1]
budget_unfilt <- read.csv("data-labeled/filtered-budget-from-source.csv")
budget_unfilt <- budget_unfilt[-1]
```

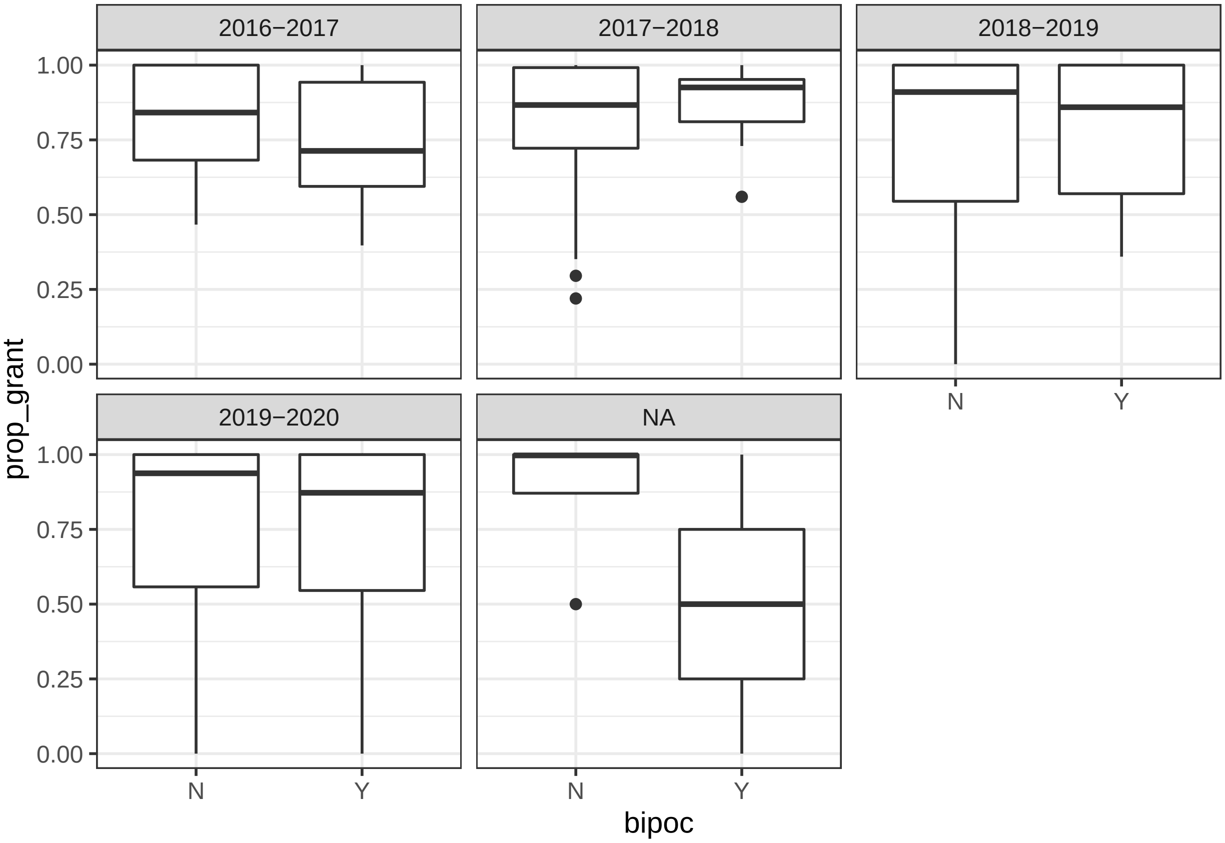
```
make_plots <- function(df) {
  plot1 <- ggplot(df, aes(x = community, y = prop_grant)) +
    geom_boxplot()
    theme_bw()

  plot2 <- ggplot(df, aes(x = bipoc, y = prop_grant)) +
    geom_boxplot() +
    facet_wrap(. ~ schoolyr) +
    theme_bw()

  plot3 <- ggplot(df, aes(x = community, y = prop_grant)) +
    geom_boxplot() +
    coord_flip() +
    facet_wrap(. ~ schoolyr) +
    theme_bw()

  plot4 <- ggplot(df, aes(x = community, y = prop_grant)) +
    geom_point(alpha = 0.3)+
    theme_bw()

  plot5 <- ggplot(df, aes(x = prop_grant)) +
    geom_histogram(aes(fill = factor(community, levels=c("Asian", "Black",
                                                         "Indigenous", "Latinx",
                                                         "Middle-Eastern",
                                                         "Multicultural group",
                                                         "Other BIPOC group",
                                                         "Non-BIPOC"))),
                   position = "stack", color = "white") +
    scale_fill_discrete(name = "community") +
    theme_bw()

  plot6 <- ggplot(df, aes(x = prop_grant)) +
    geom_histogram() +
    facet_wrap(. ~ community) +
    theme_bw()

  return(list(plot1, plot2, plot3, plot4, plot5, plot6))
}
```

```
make_plots(prog)
```

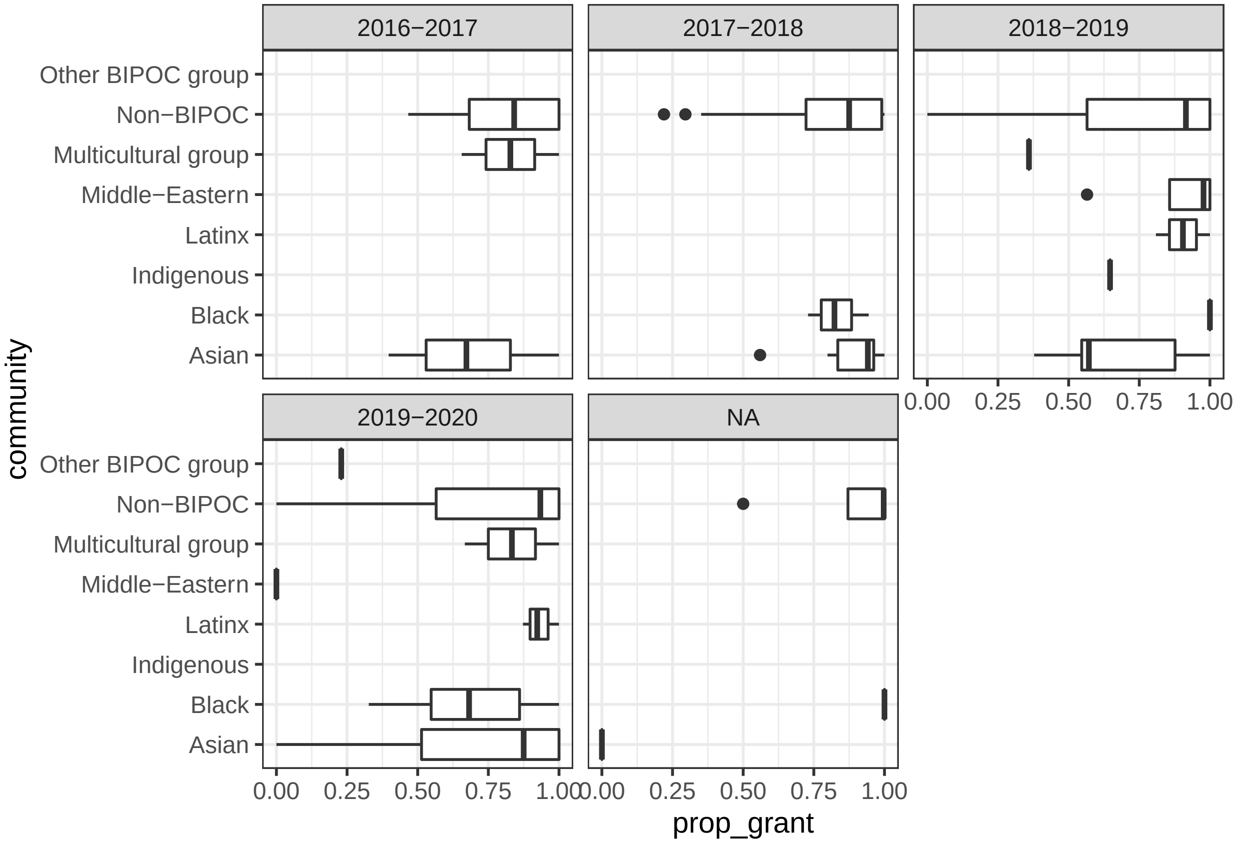
```
## [[1]]
```


prop_grant
1.00
0.75
0.50
0.25
0.00
Asian
Black
Indigenous
Latinx
Middle−Eastern
Multicultural group
Non−BIPOC
Other BIPOC group
community

```
##
## [[2]]
```

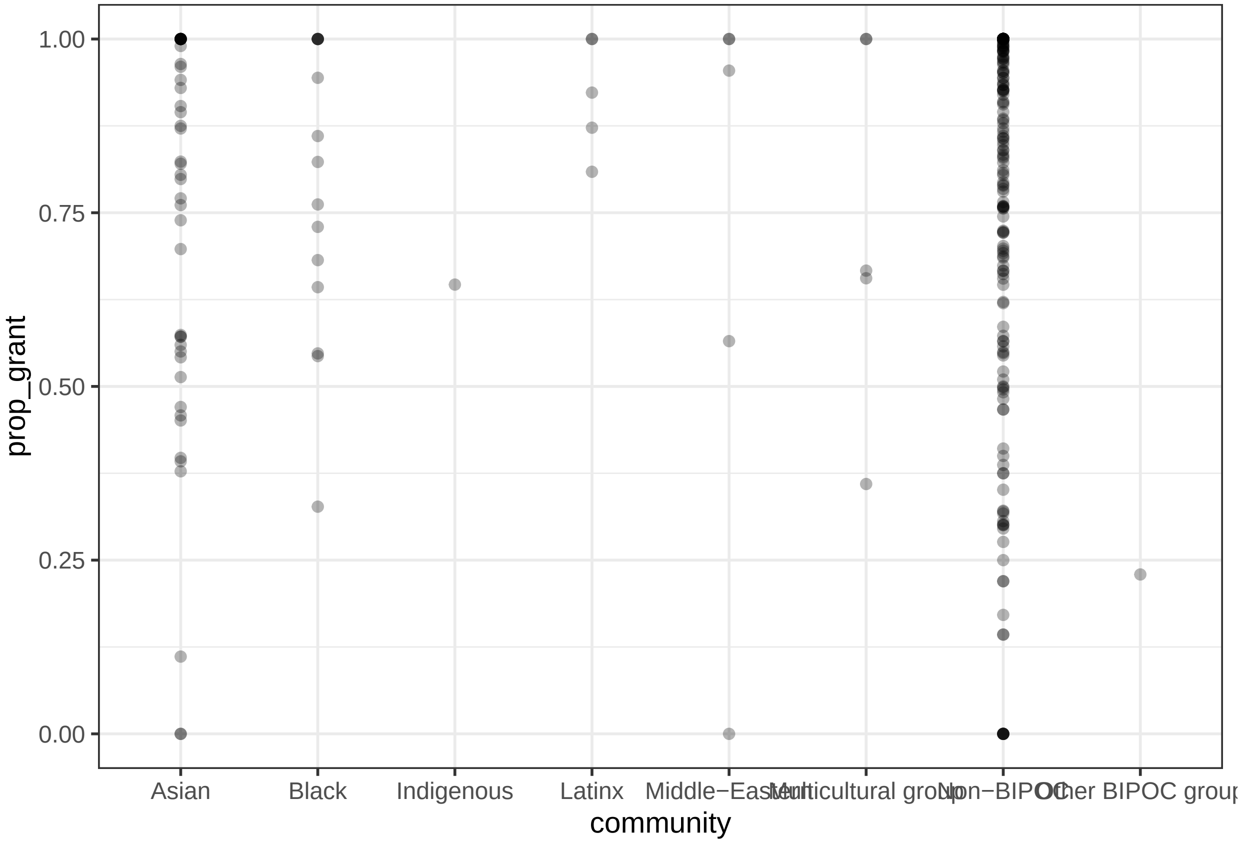

2016−2017
2017−2018
2018−2019
2019−2020
NA
prop_grant
1.00
0.75
0.50
0.25
0.00
N
Y
bipoc

```
##
## [[3]]
```



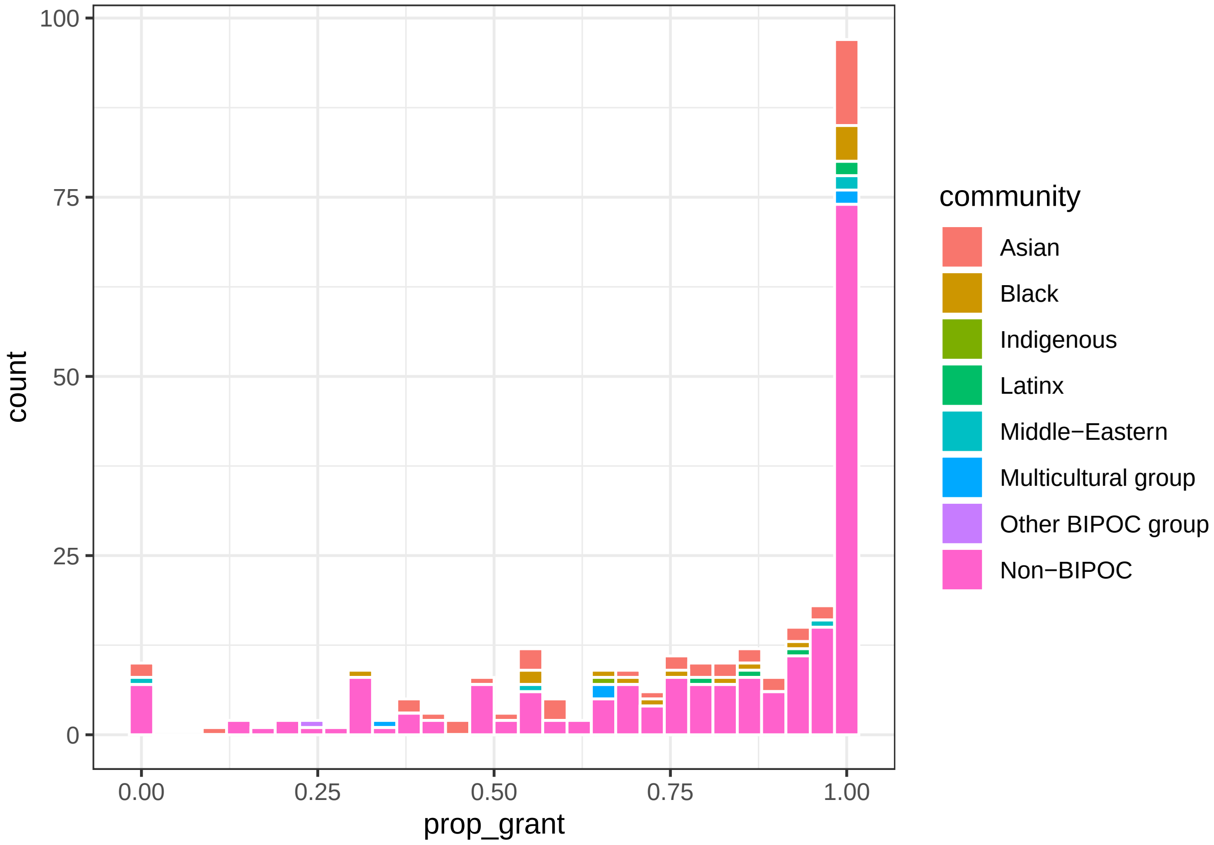
2016−2017
2017−2018
2018−2019
2019−2020
NA
Other BIPOC group
Non−BIPOC
Multicultural group
Middle−Eastern
Latinx
Indigenous
Black
Asian
community
0.00
0.25
0.50
0.75
1.00
prop_grant

```
##
## [[4]]
```



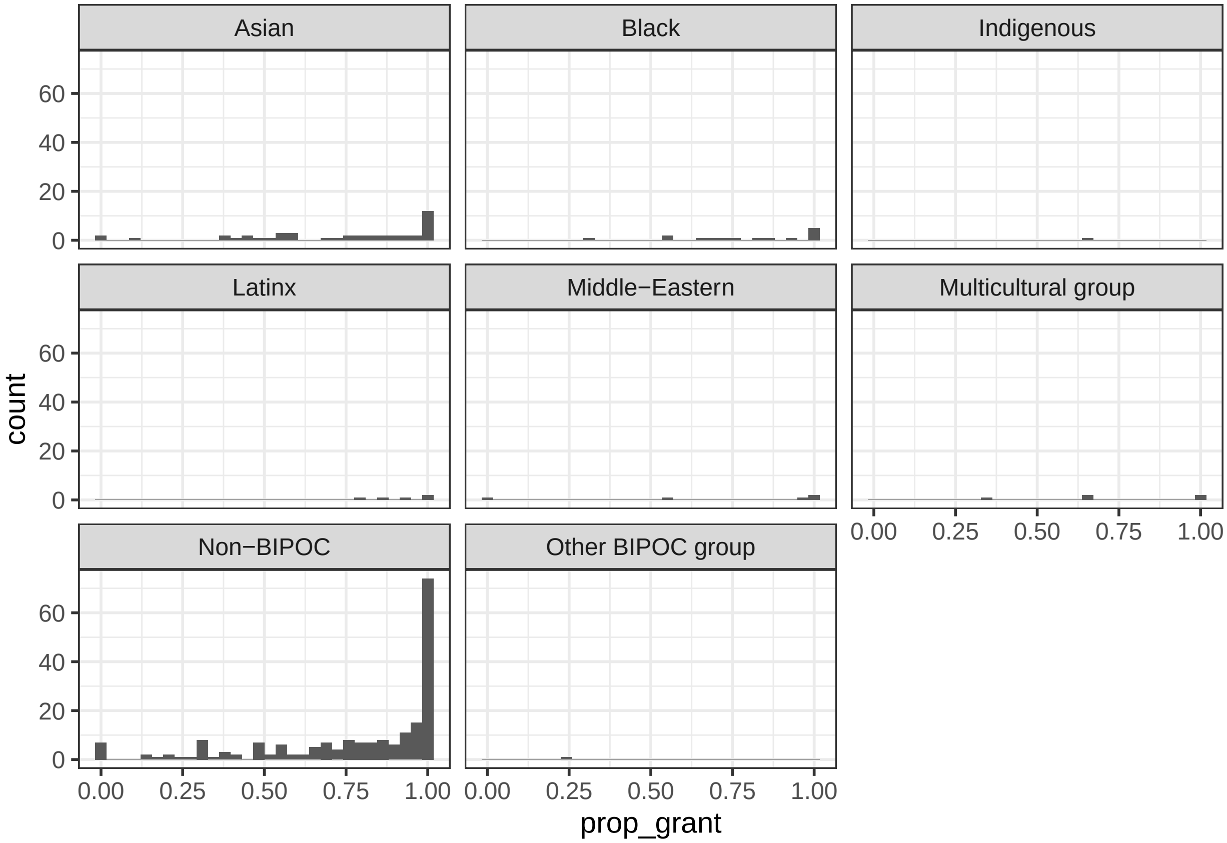
```
##
## [[5]]
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



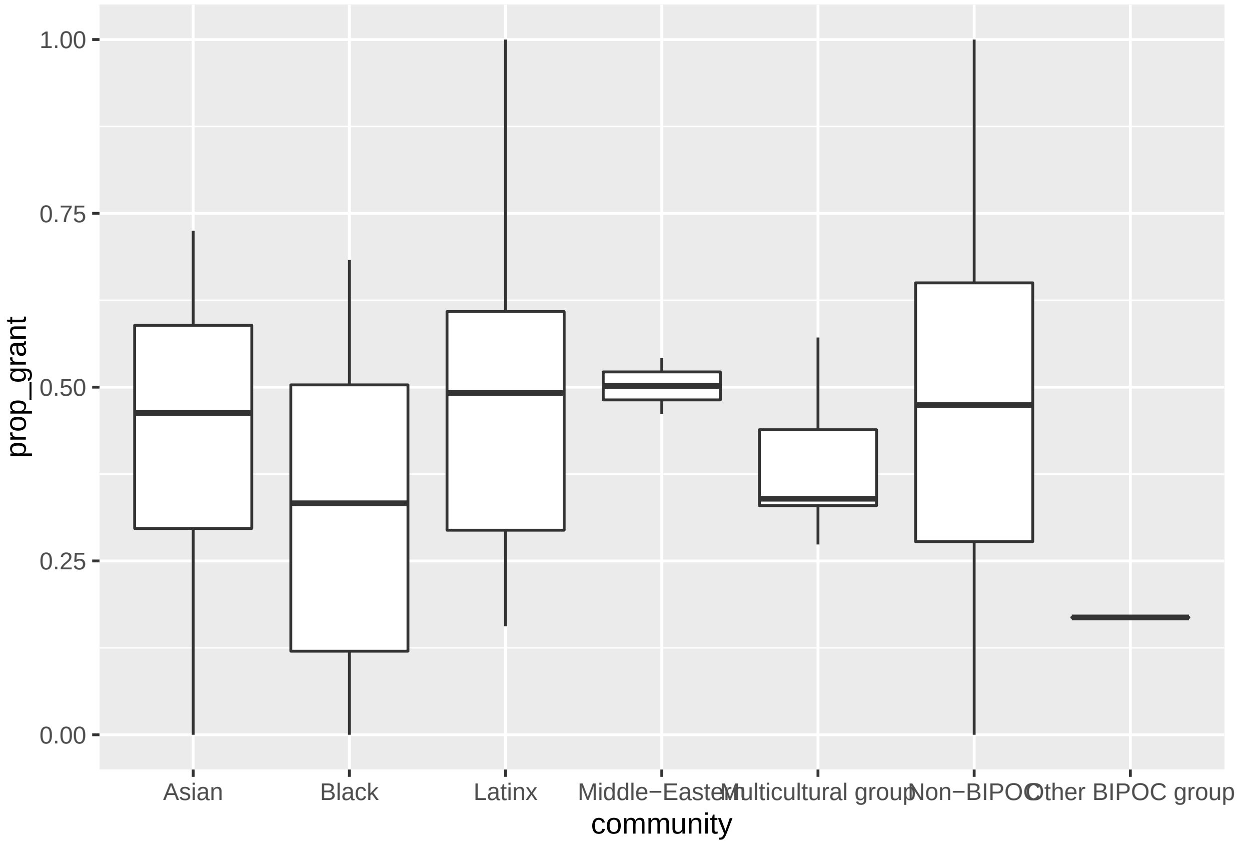
```
##
## [[6]]
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```

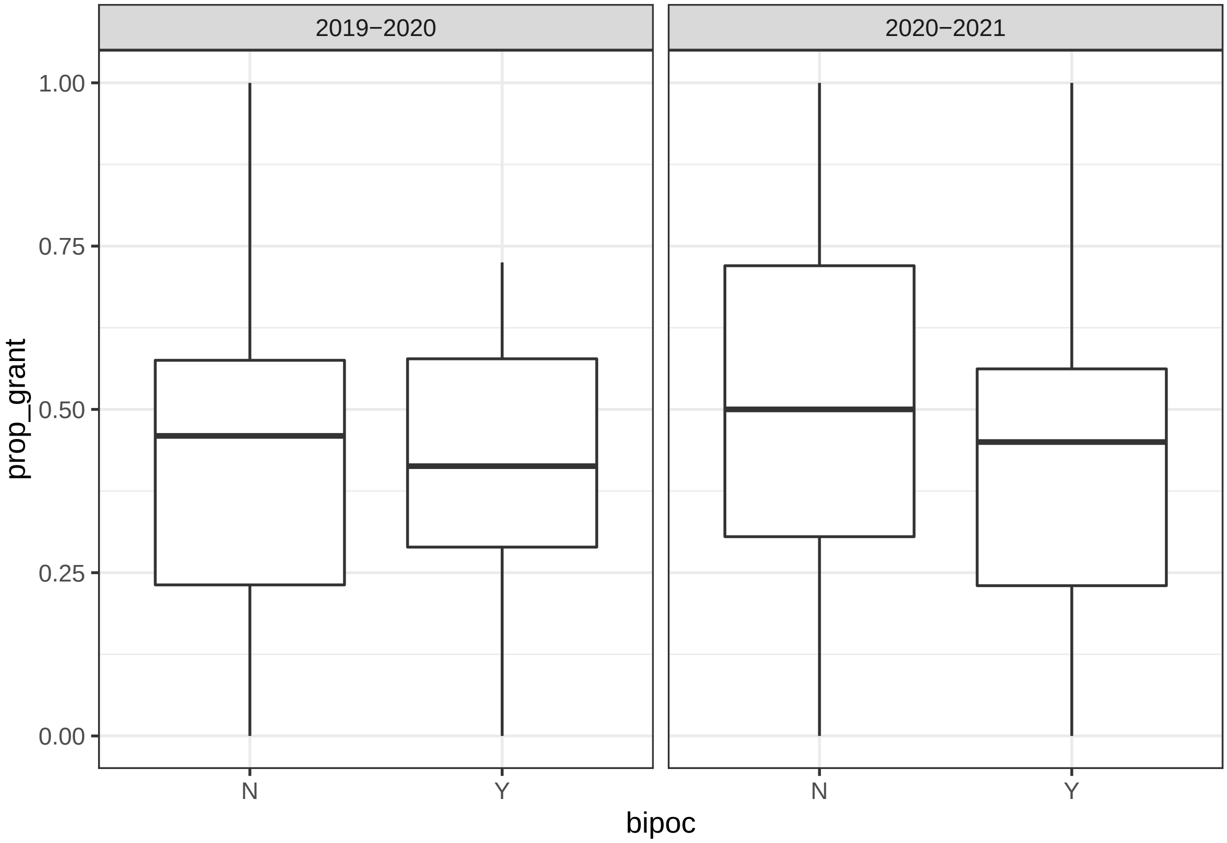


```
make_plots(budget)
```

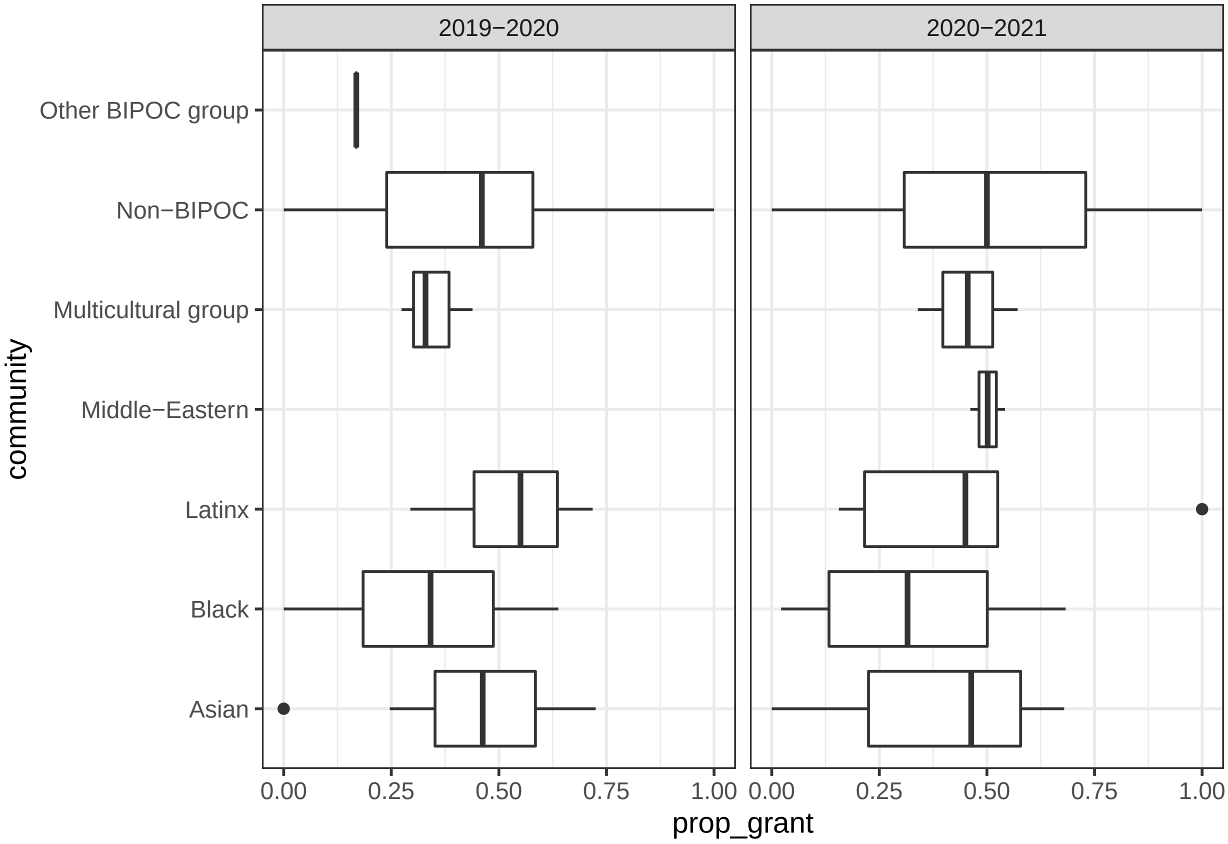
```
## [[1]]
```


1.00
0.75
0.50
0.25
0.00
prop_grant
Asian
Black
Latinx
Middle−Eastern
Multicultural group
Non−BIPOC
Other BIPOC group
community

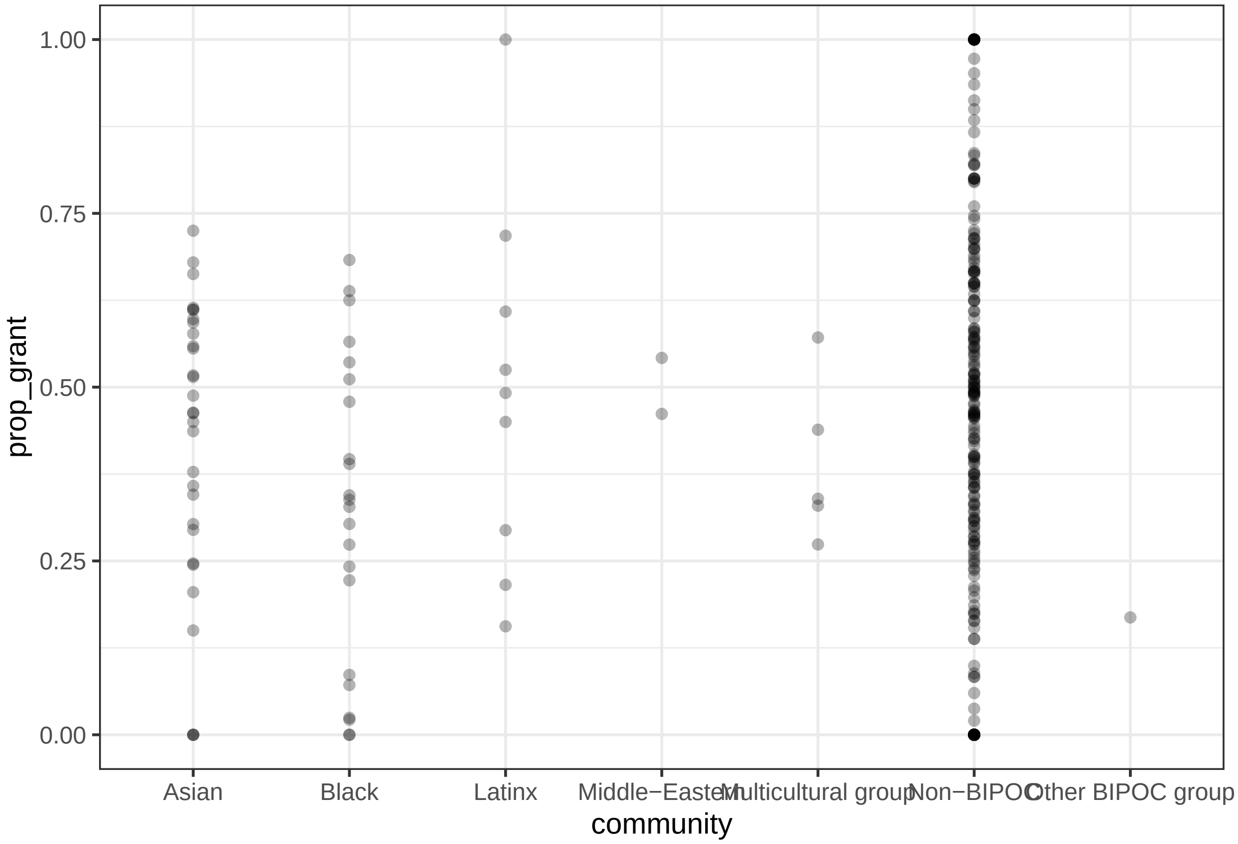
```
##
## [[2]]
```



```
##
## [[3]]
```

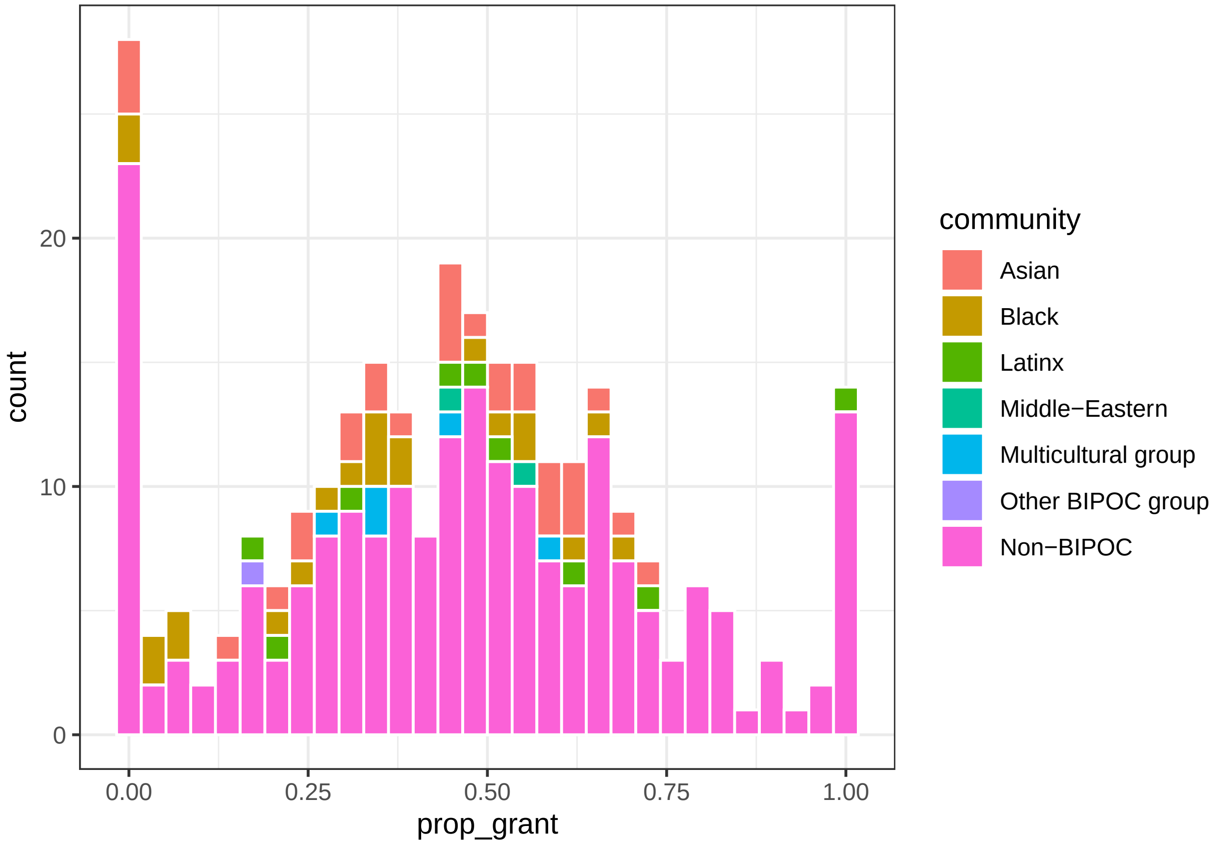

2019−2020
2020−2021
community
Other BIPOC group
Non−BIPOC
Multicultural group
Middle−Eastern
Latinx
Black
Asian
0.00
0.25
0.50
0.75
1.00
0.00
0.25
0.50
0.75
1.00
prop_grant

```
##
## [[4]]
```



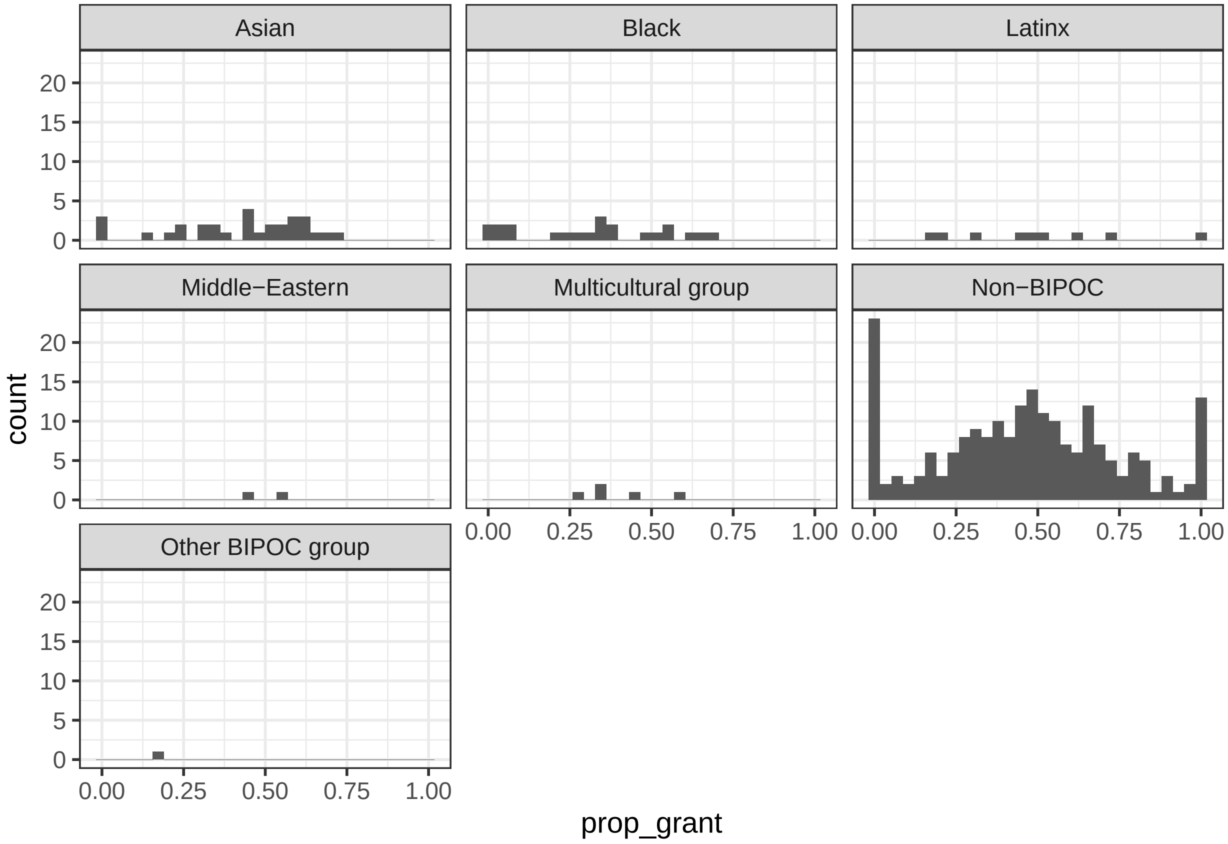
```
##
## [[5]]
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



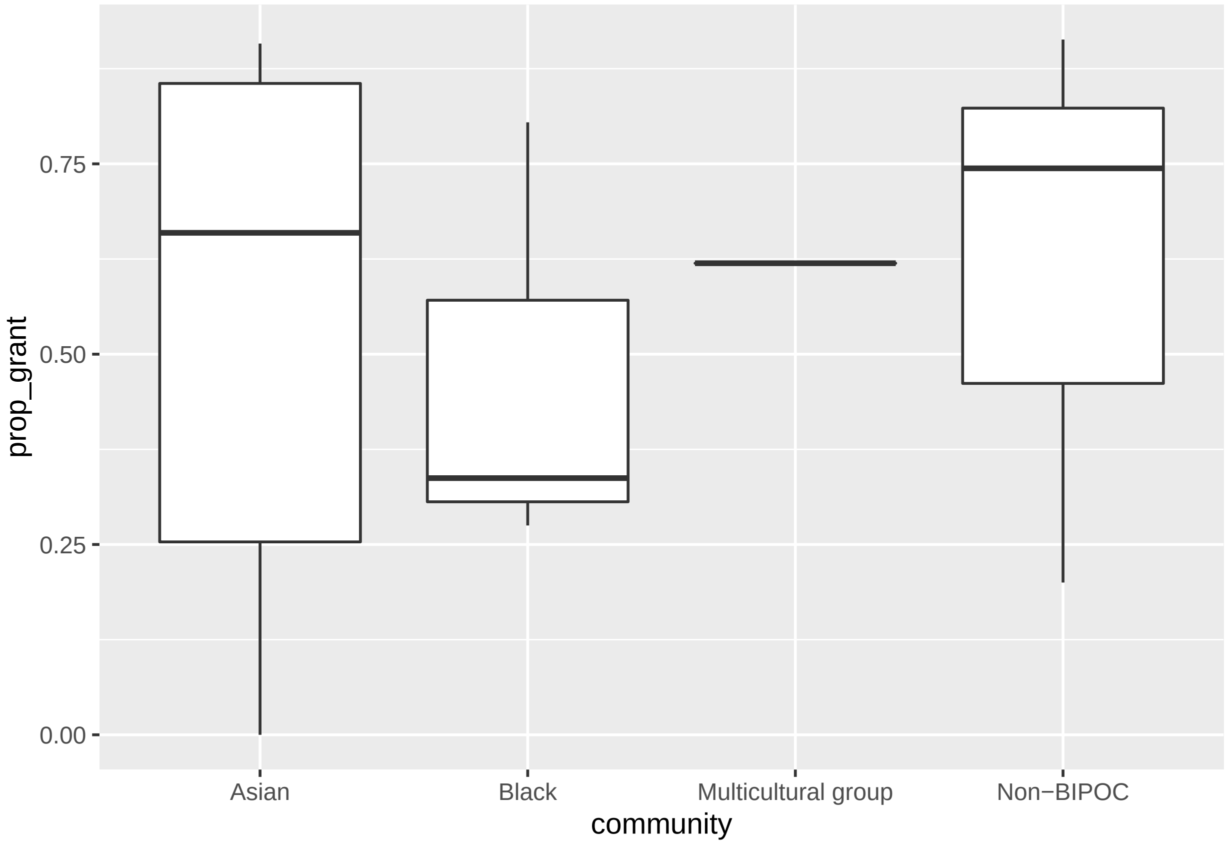
```
##
## [[6]]
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```

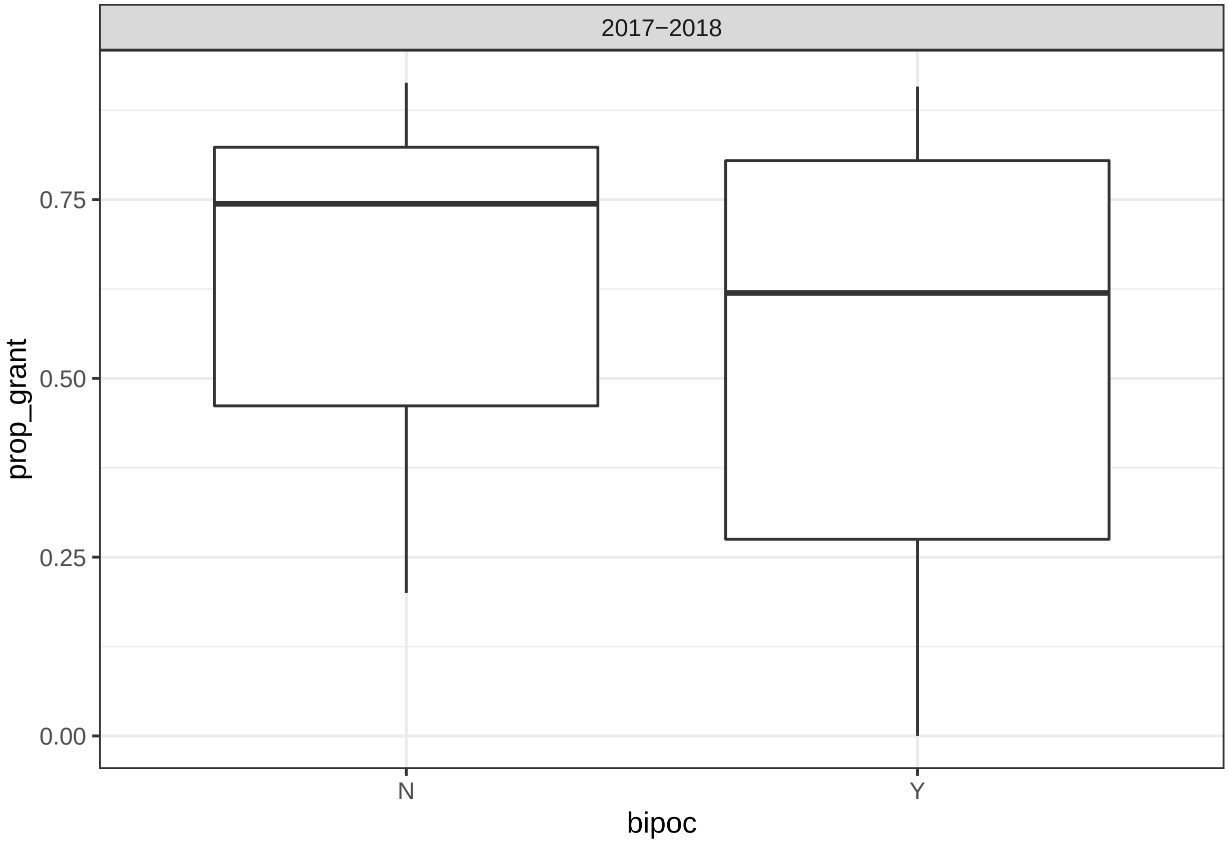


```
make_plots(sofc)
```

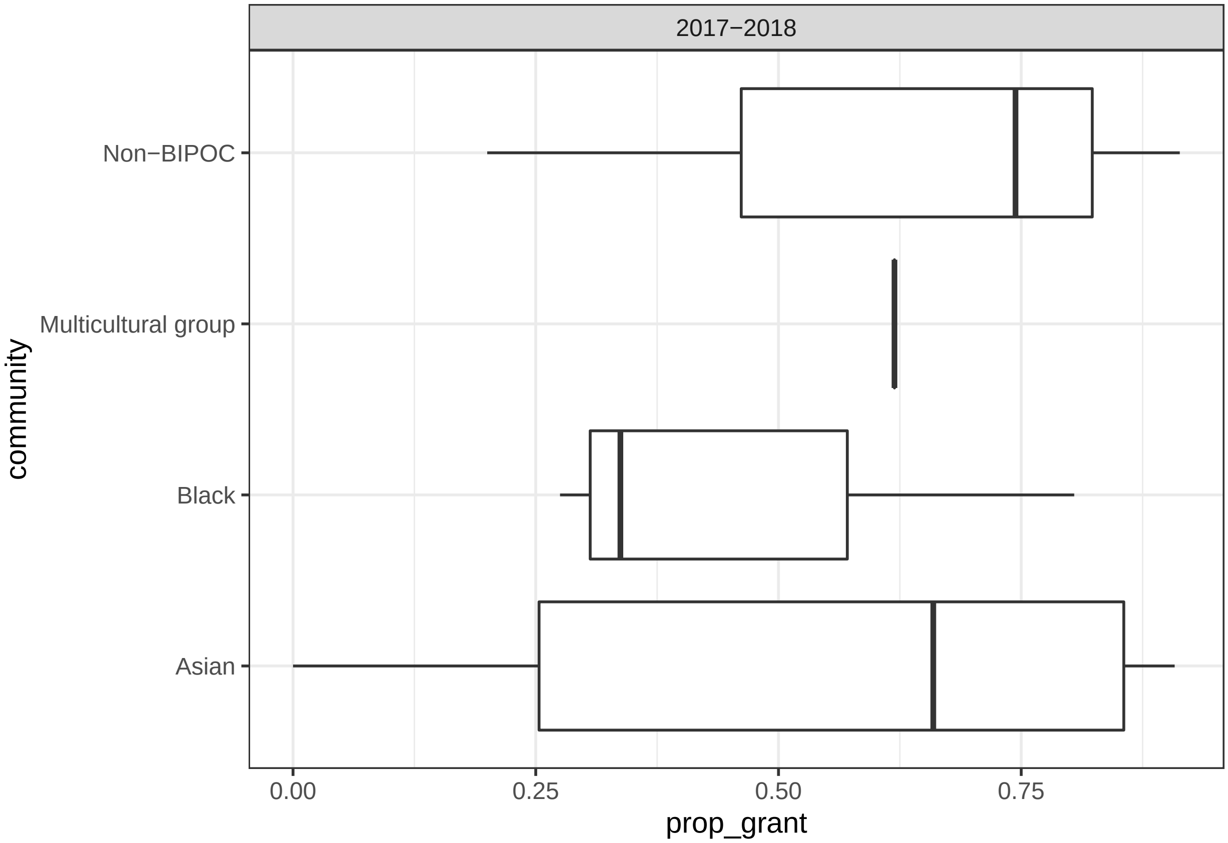
```
## [[1]]
```


prop_grant
0.00
0.25
0.50
0.75
Asian
Black
Multicultural group
Non−BIPOC
community

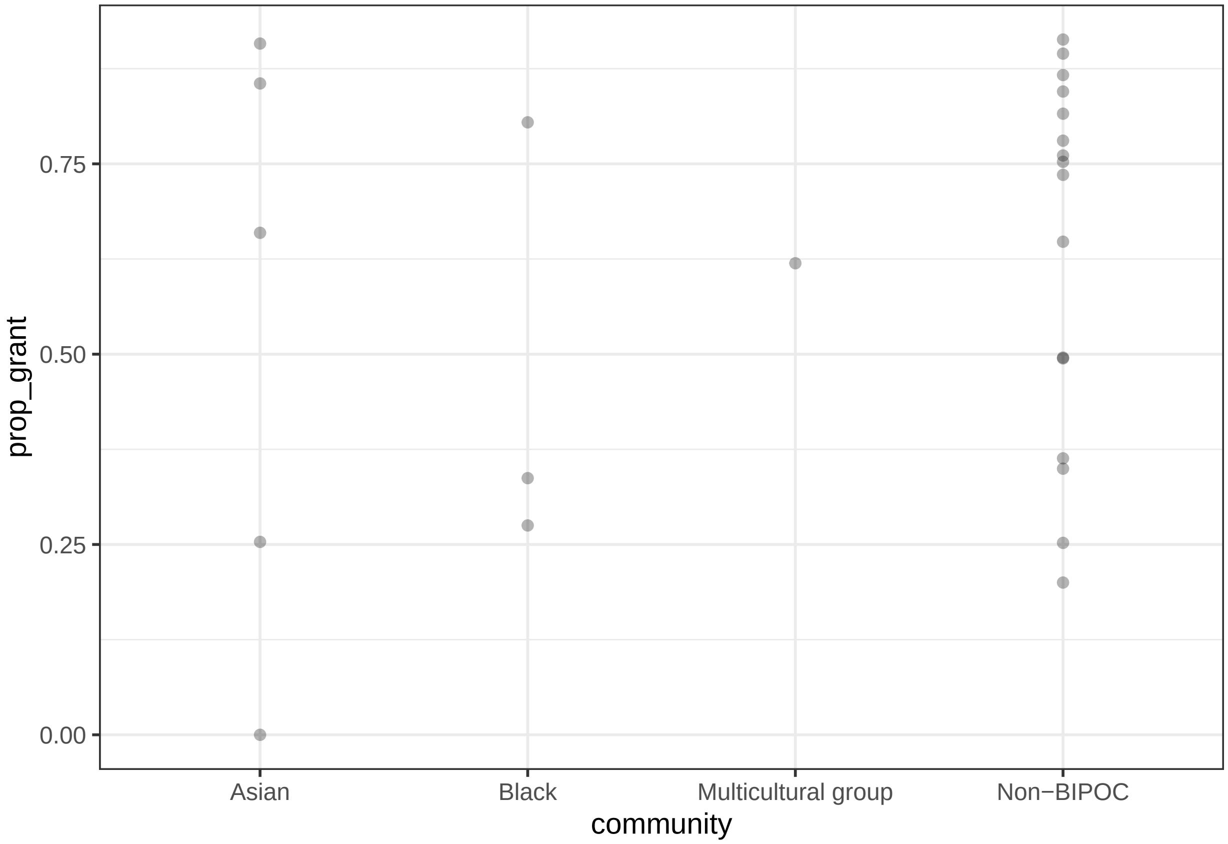
```
##
## [[2]]
```



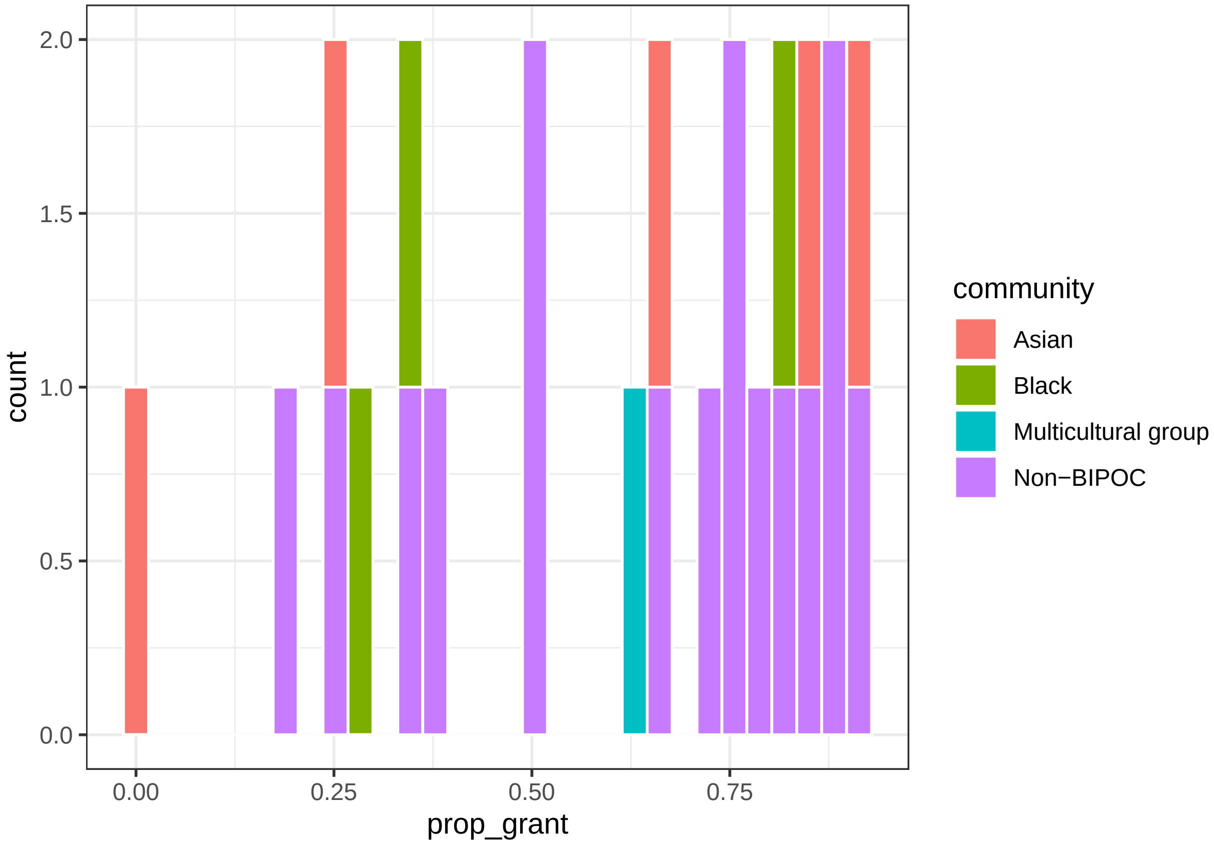
```
##
## [[3]]
```


2017−2018
Non−BIPOC
Multicultural group
Black
Asian
community
0.00
0.25
0.50
0.75
prop_grant

```
##
## [[4]]
```

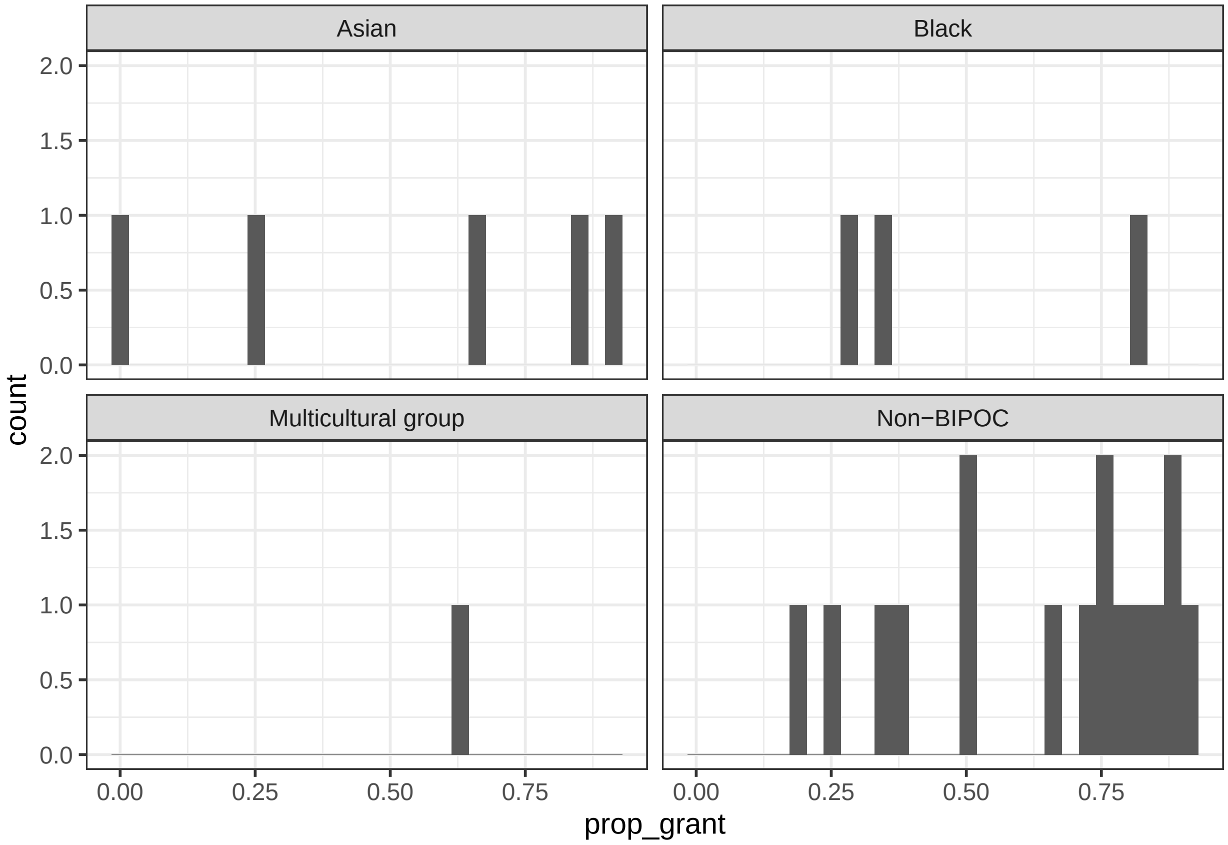


```
##
## [[5]]

## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



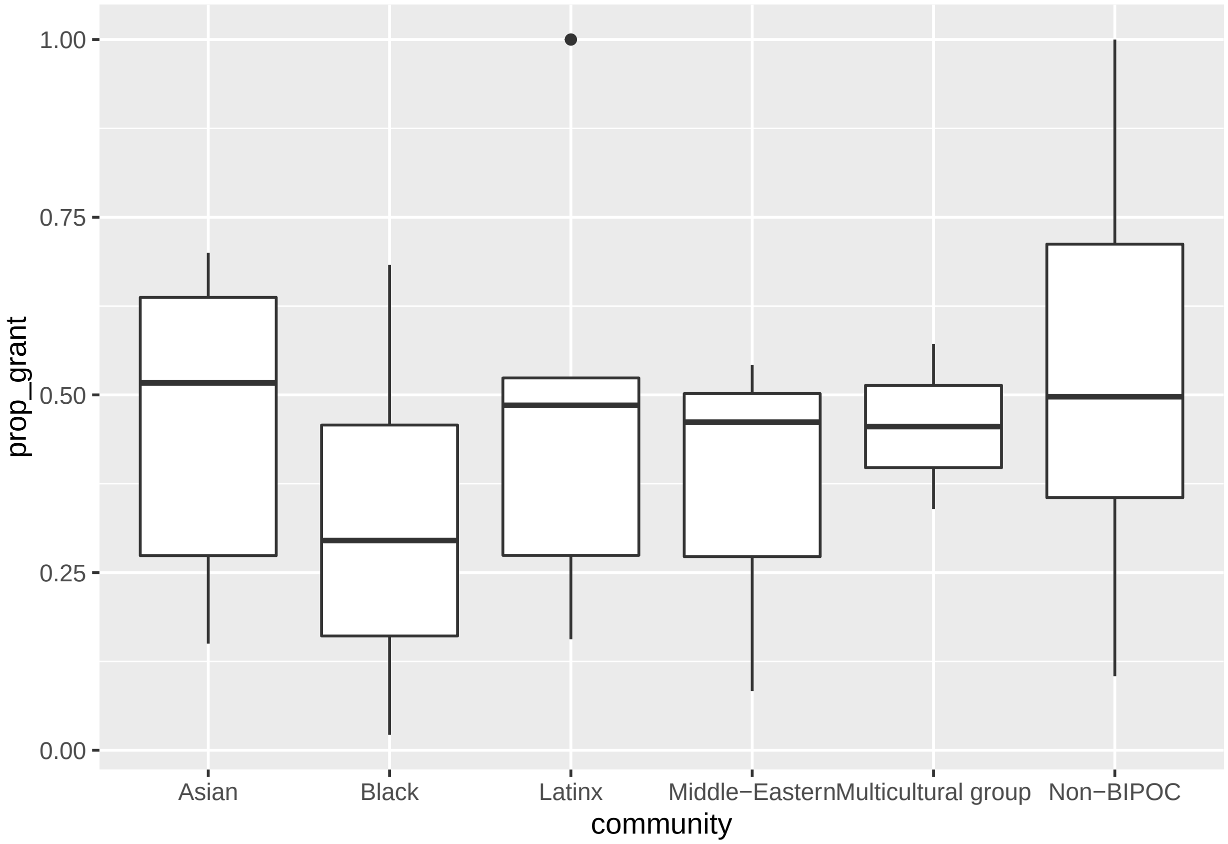
```
##
## [[6]]
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```

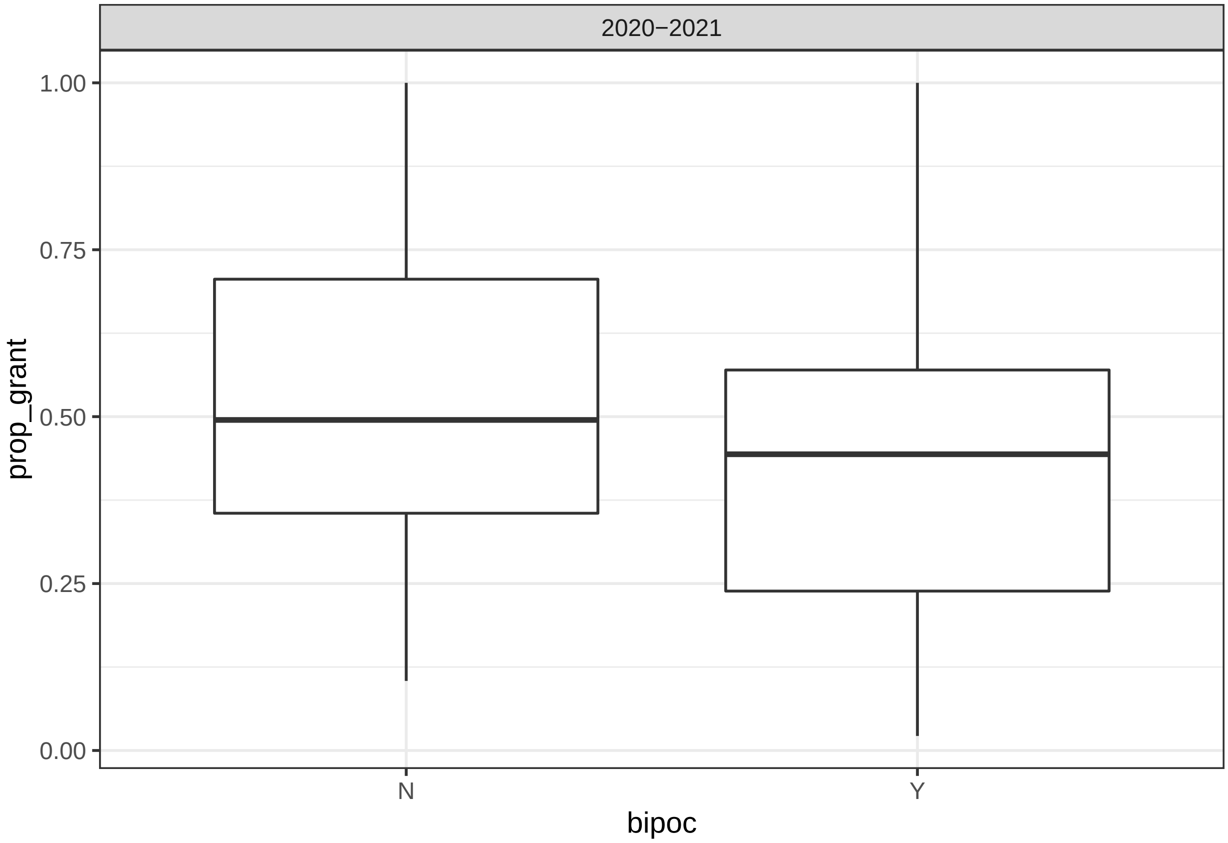


```r
make_plots(budget_unfilt)
```

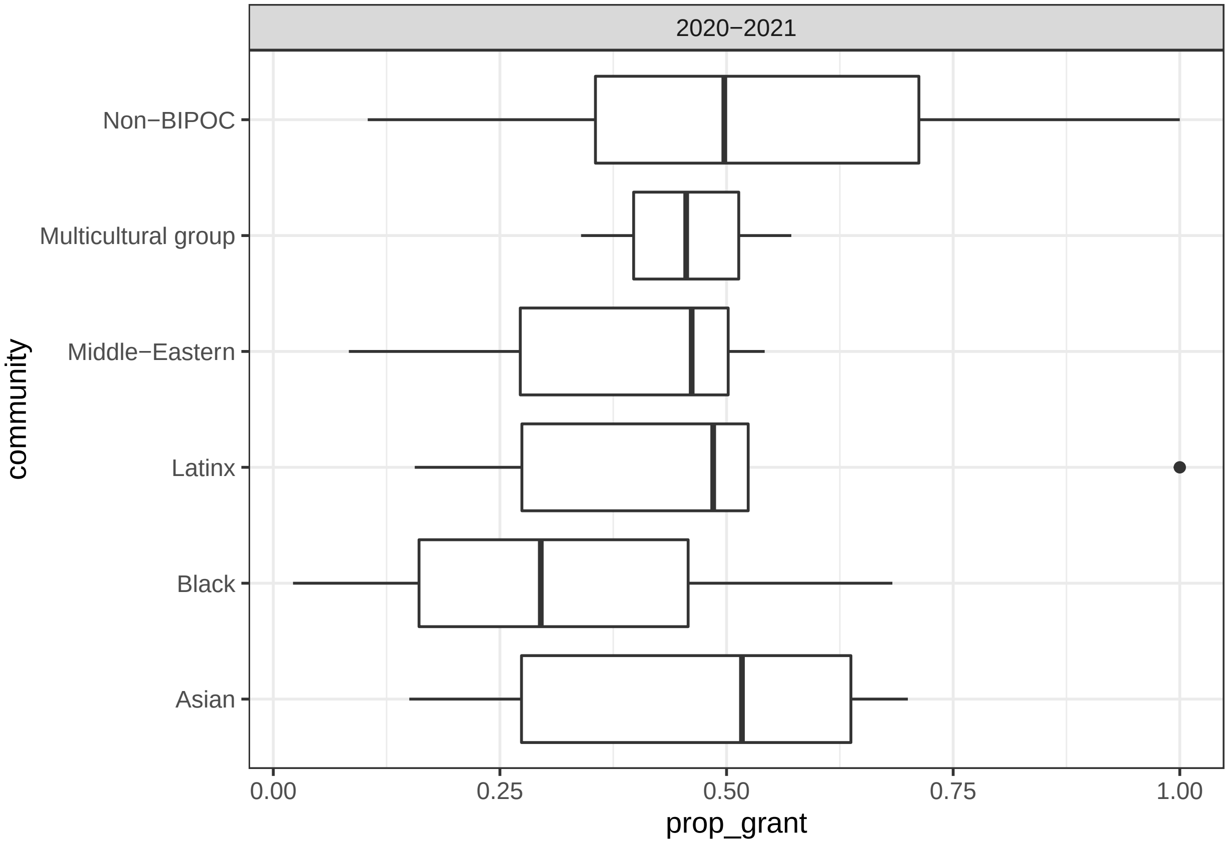
```
## [[1]]
```


1.00
0.75
0.50
0.25
0.00
prop_grant
Asian
Black
Latinx
Middle−Eastern
Multicultural group
Non−BIPOC
community

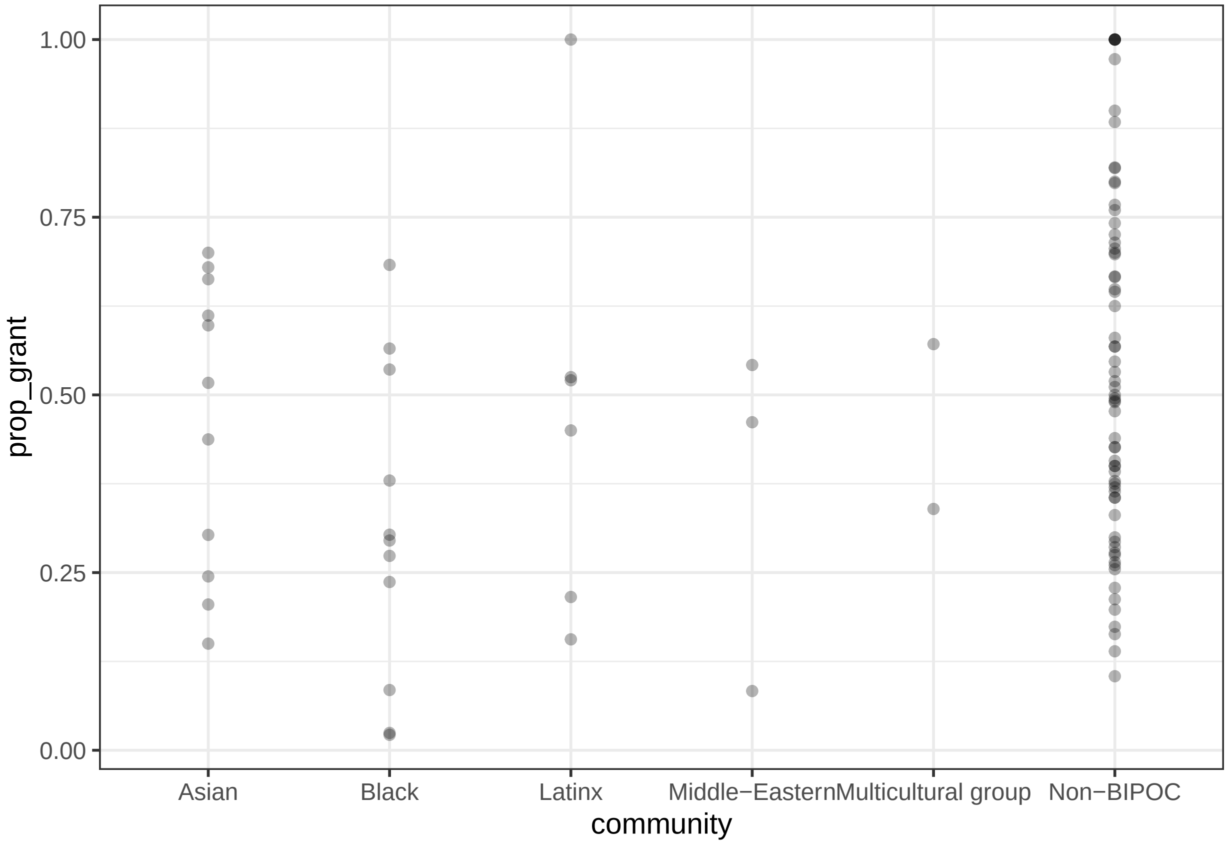
```
##
## [[2]]
```



```
##
## [[3]]
```

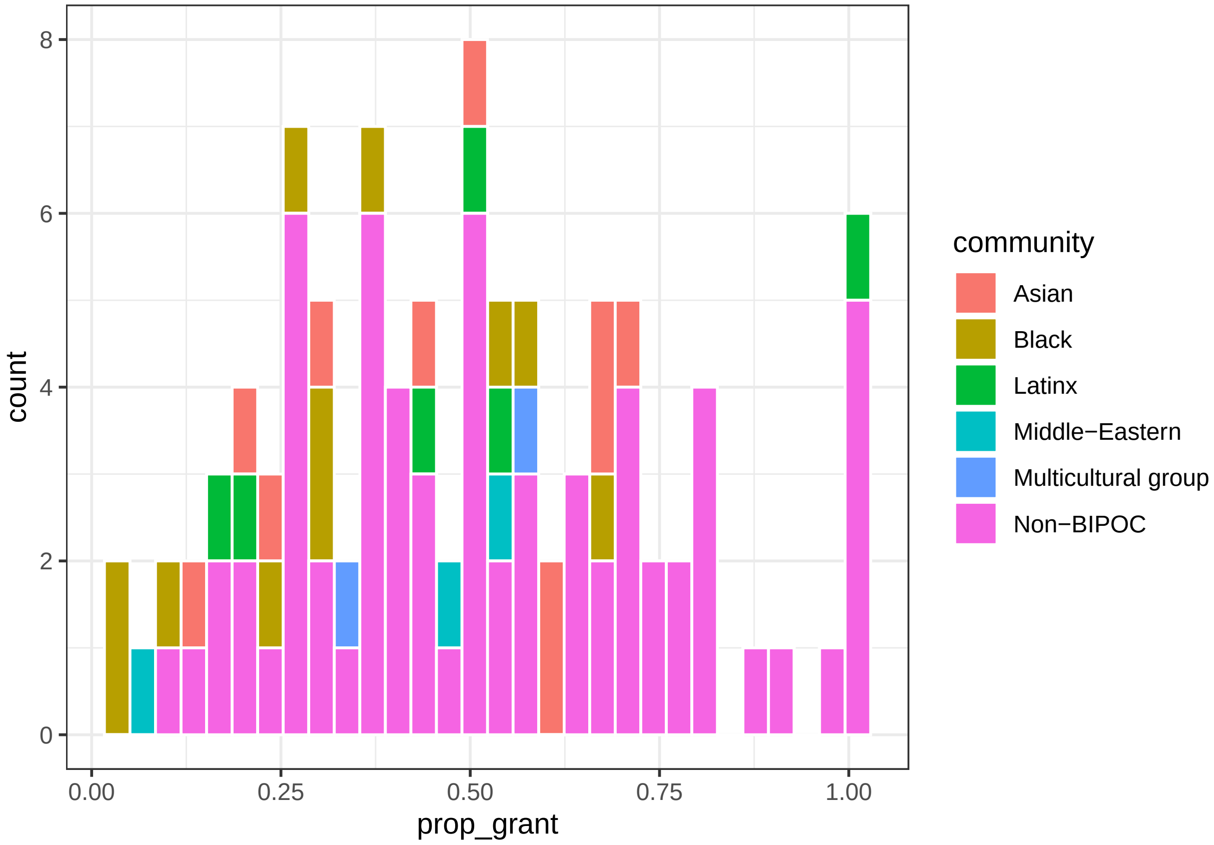

2020−2021
Non−BIPOC
Multicultural group
Middle−Eastern
Latinx
Black
Asian
community
0.00
0.25
0.50
0.75
1.00
prop_grant

```
##
## [[4]]
```



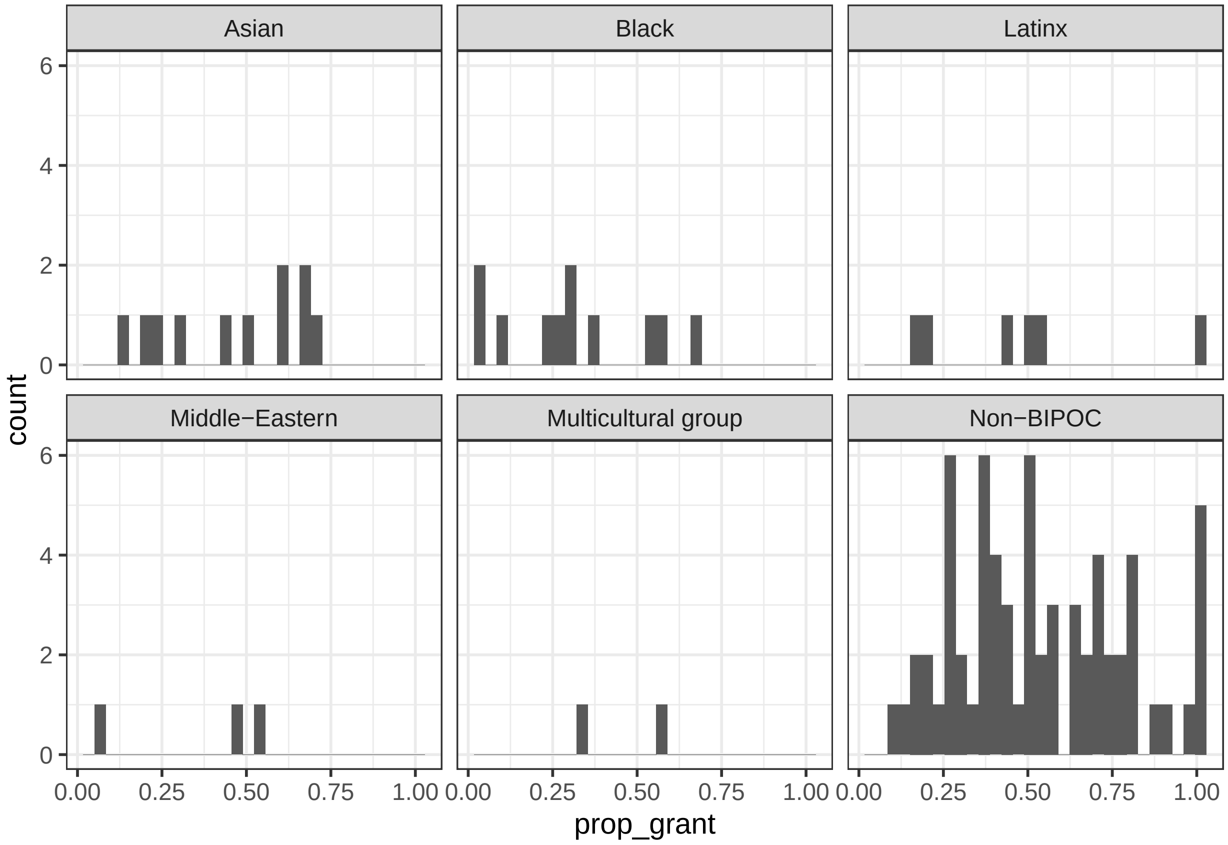
```
##
## [[5]]
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



```
##
## [[6]]
```

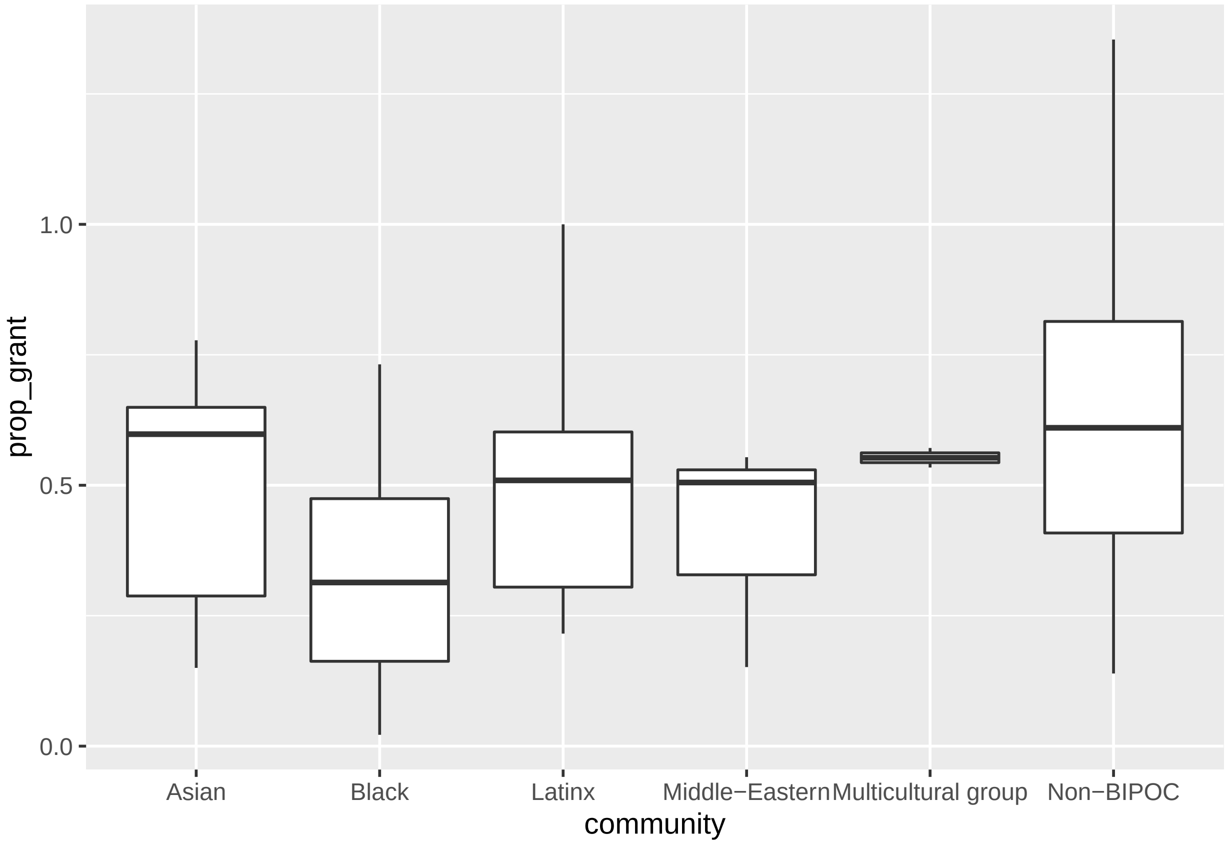
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



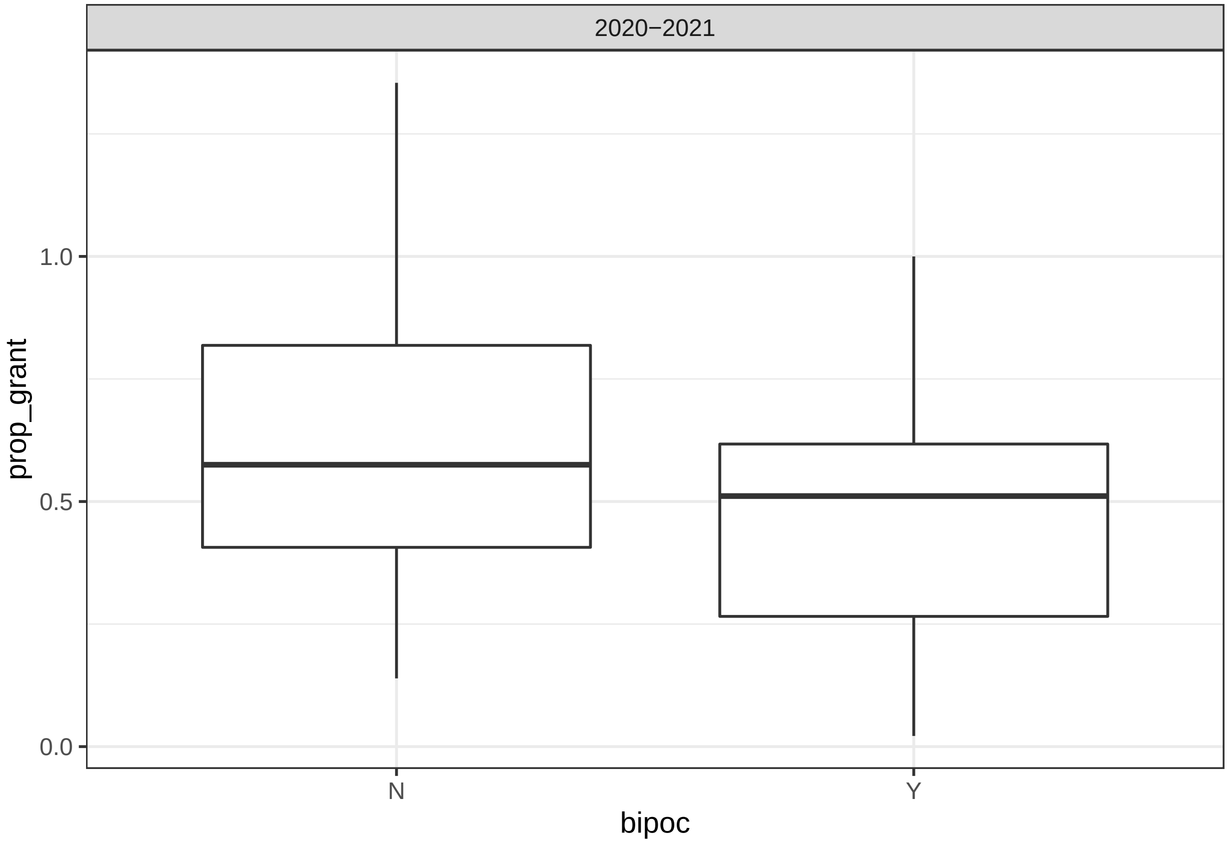
```r
budget_filt <- budget_unfilt %>%
  select(-req, -prop_grant) %>%
  rename(req = req_filt, prop_grant = prop_grant_filt)
```

```r
make_plots(budget_filt)
```

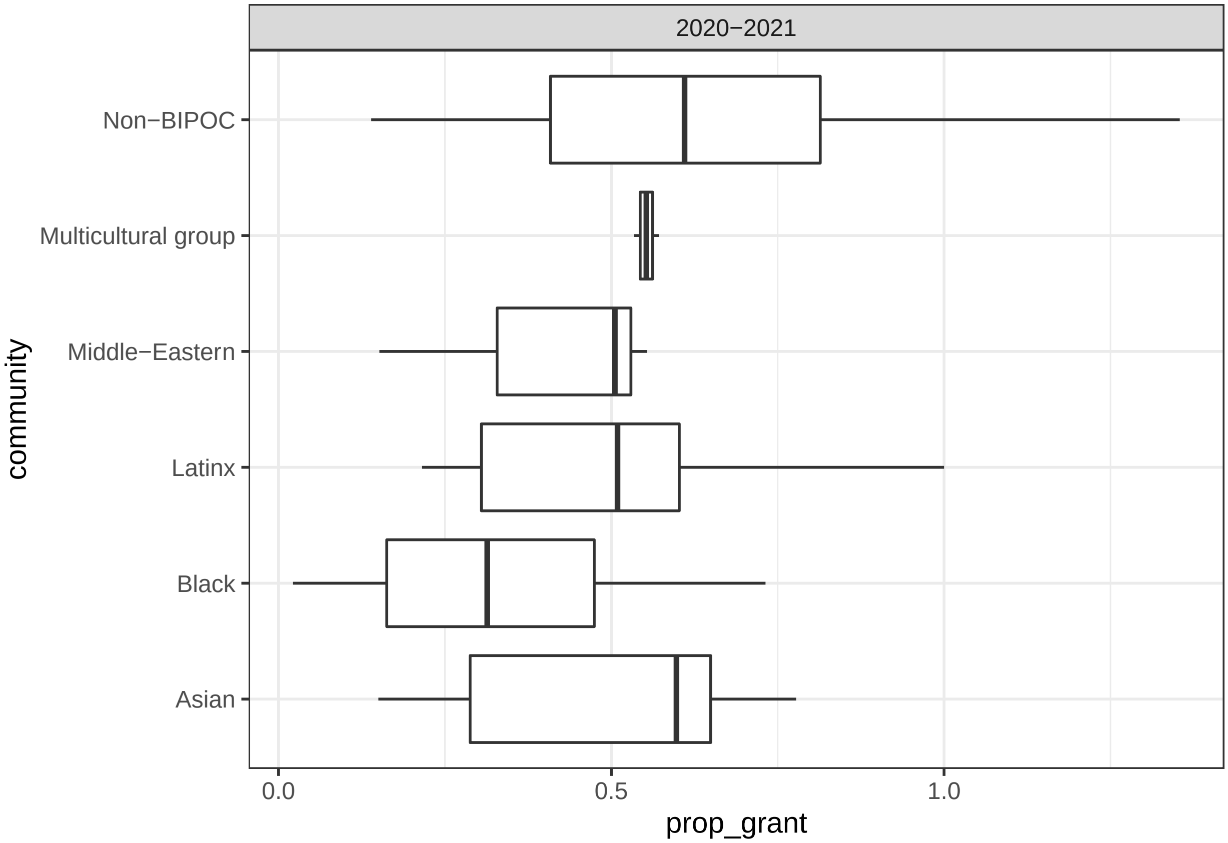
```
## [[1]]
```


prop_grant
1.0
0.5
0.0
Asian
Black
Latinx
Middle−Eastern
Multicultural group
Non−BIPOC
community

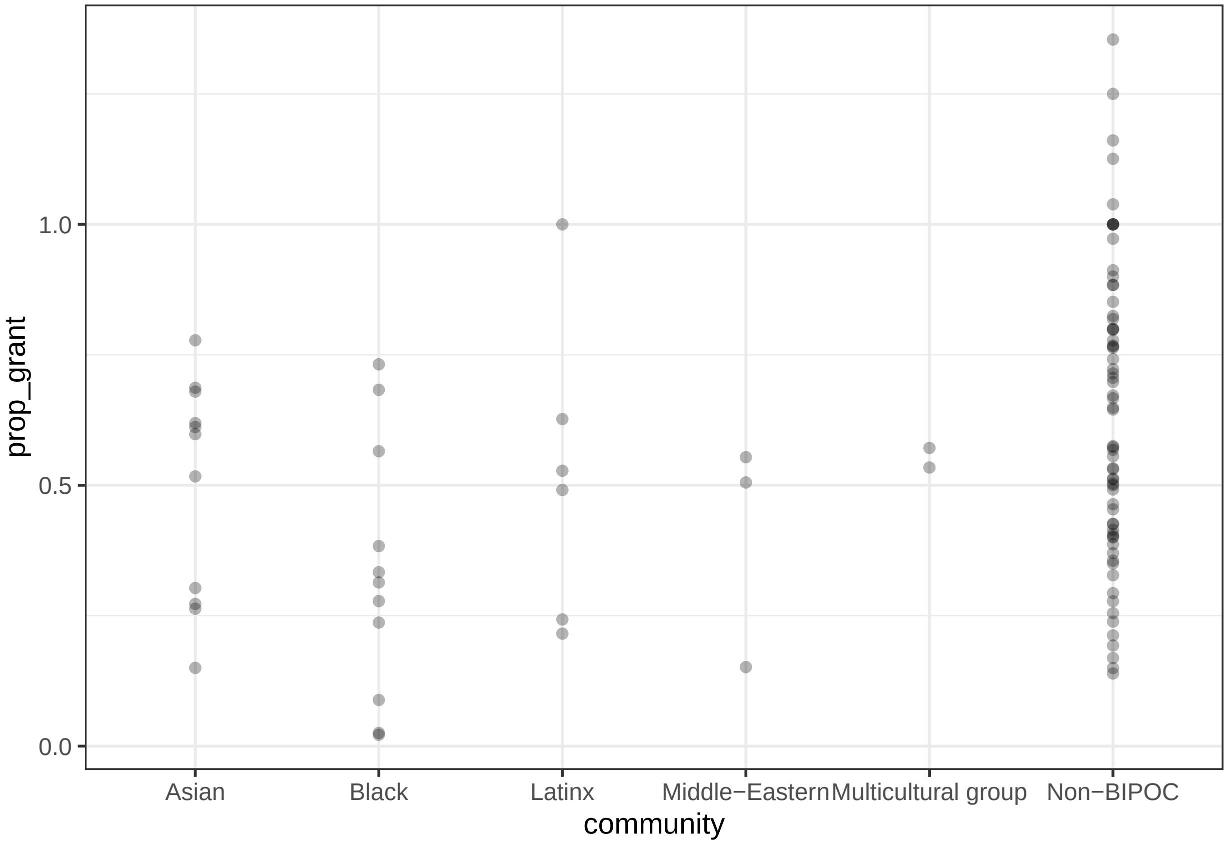
```
##
## [[2]]
```



```
##
## [[3]]
```

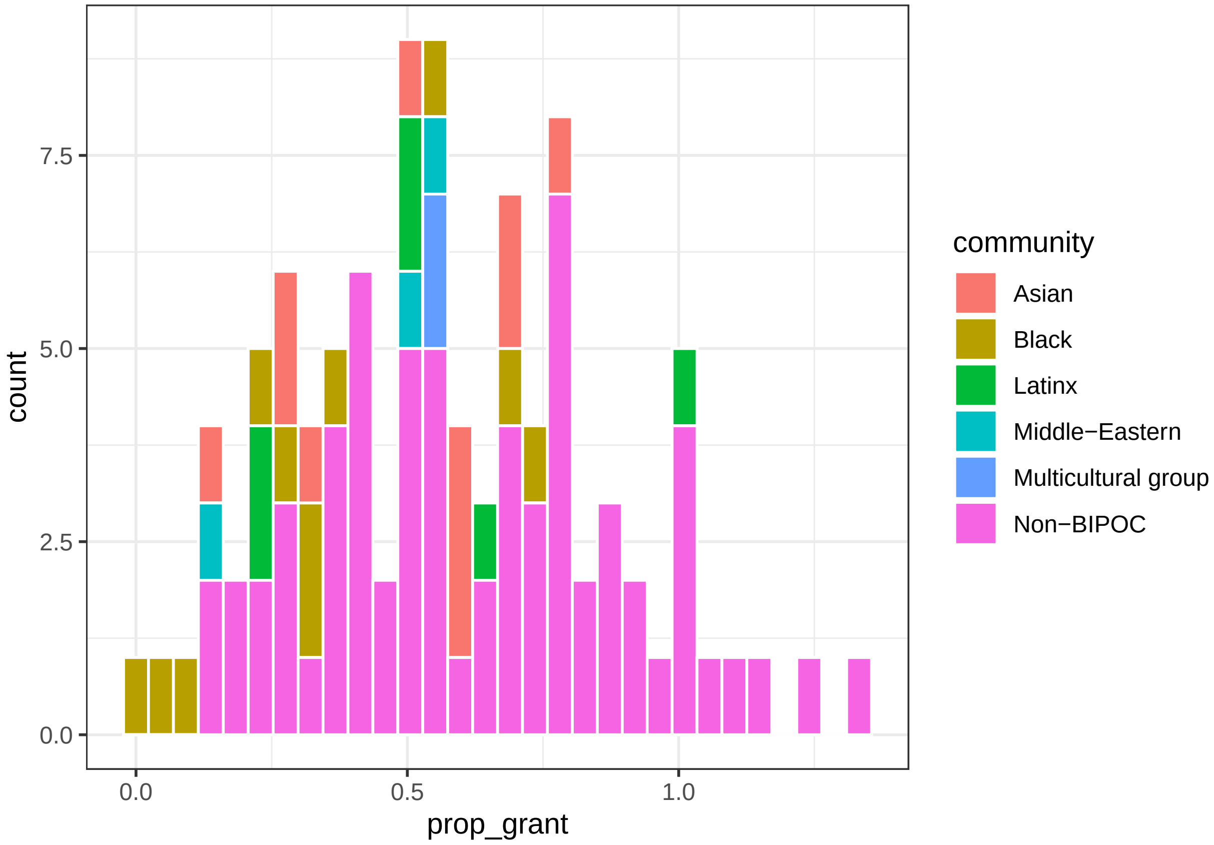

2020−2021
Non−BIPOC
Multicultural group
Middle−Eastern
Latinx
Black
Asian
community
0.0
0.5
1.0
prop_grant

```
##
## [[4]]
```



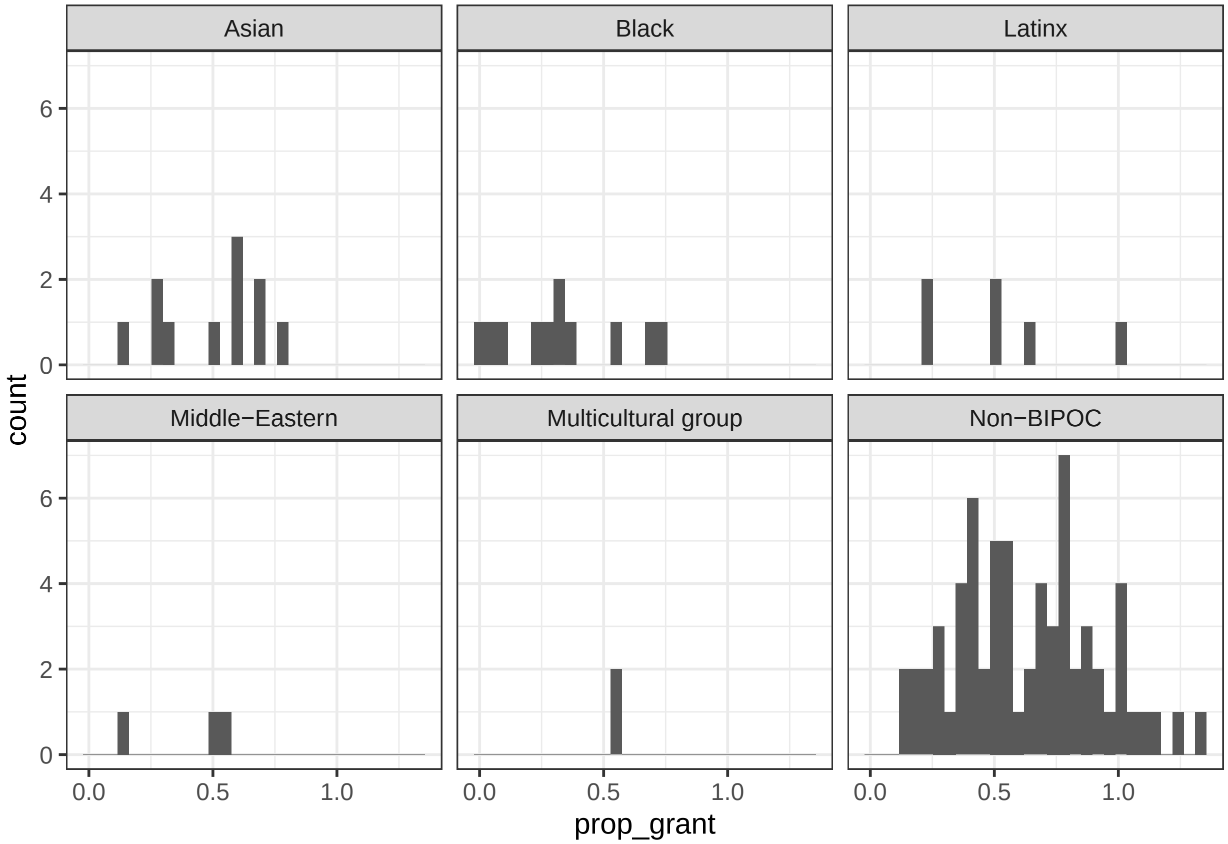
```
##
## [[5]]
```

```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```

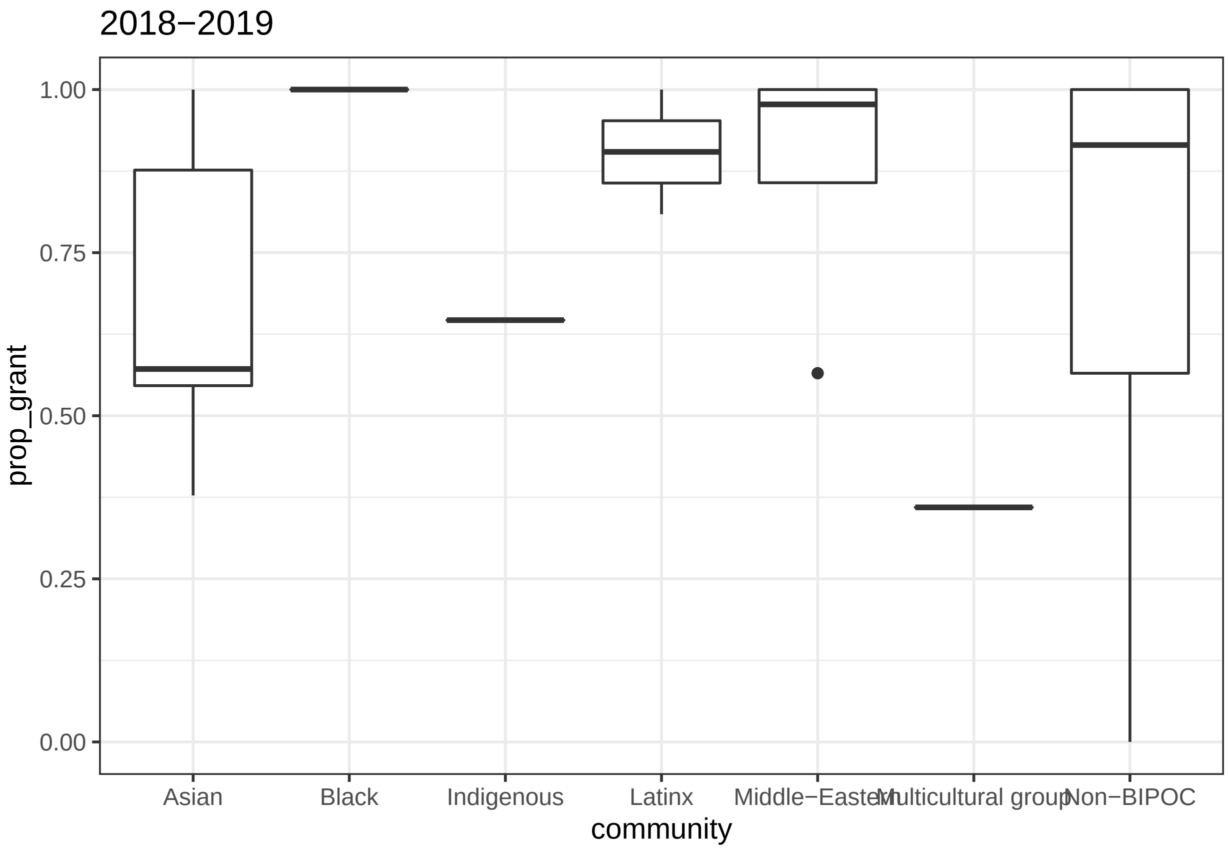


```
##
## [[6]]
```

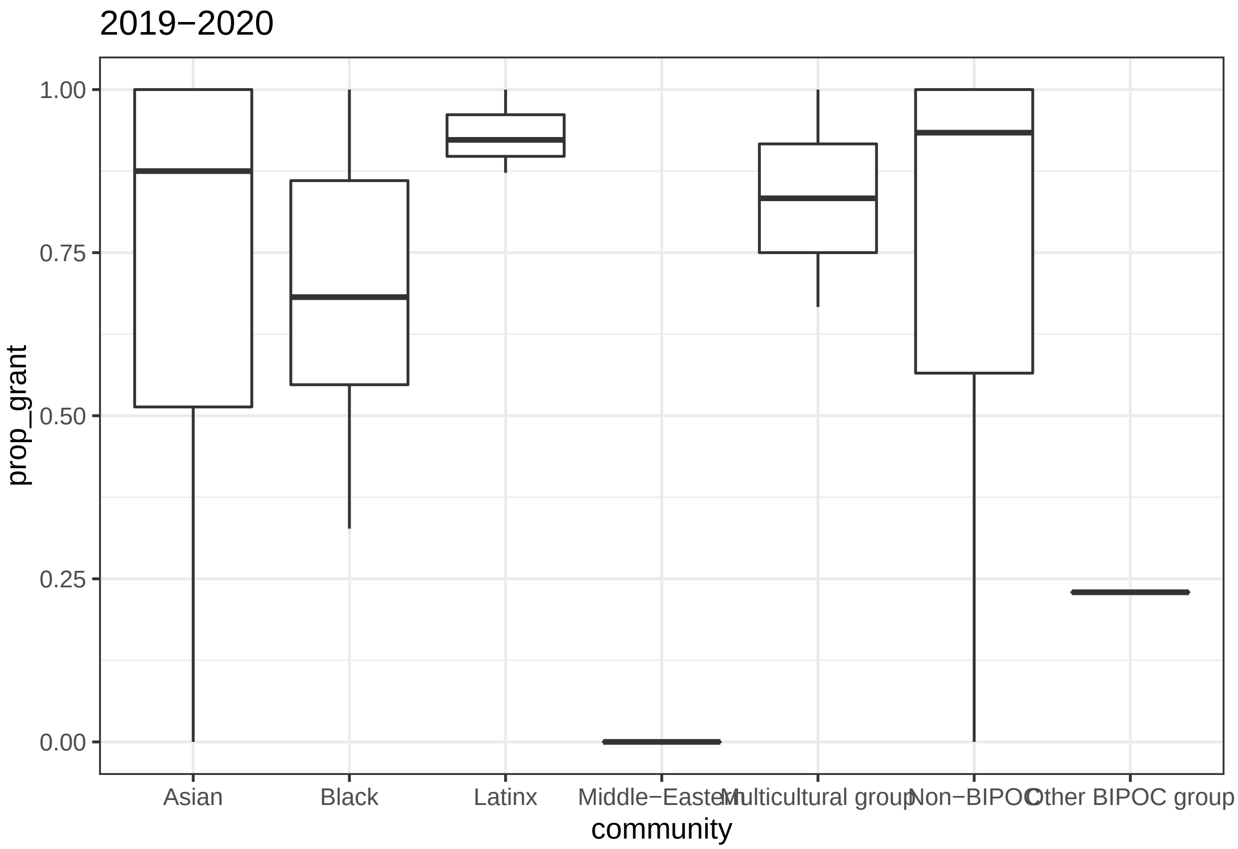
```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



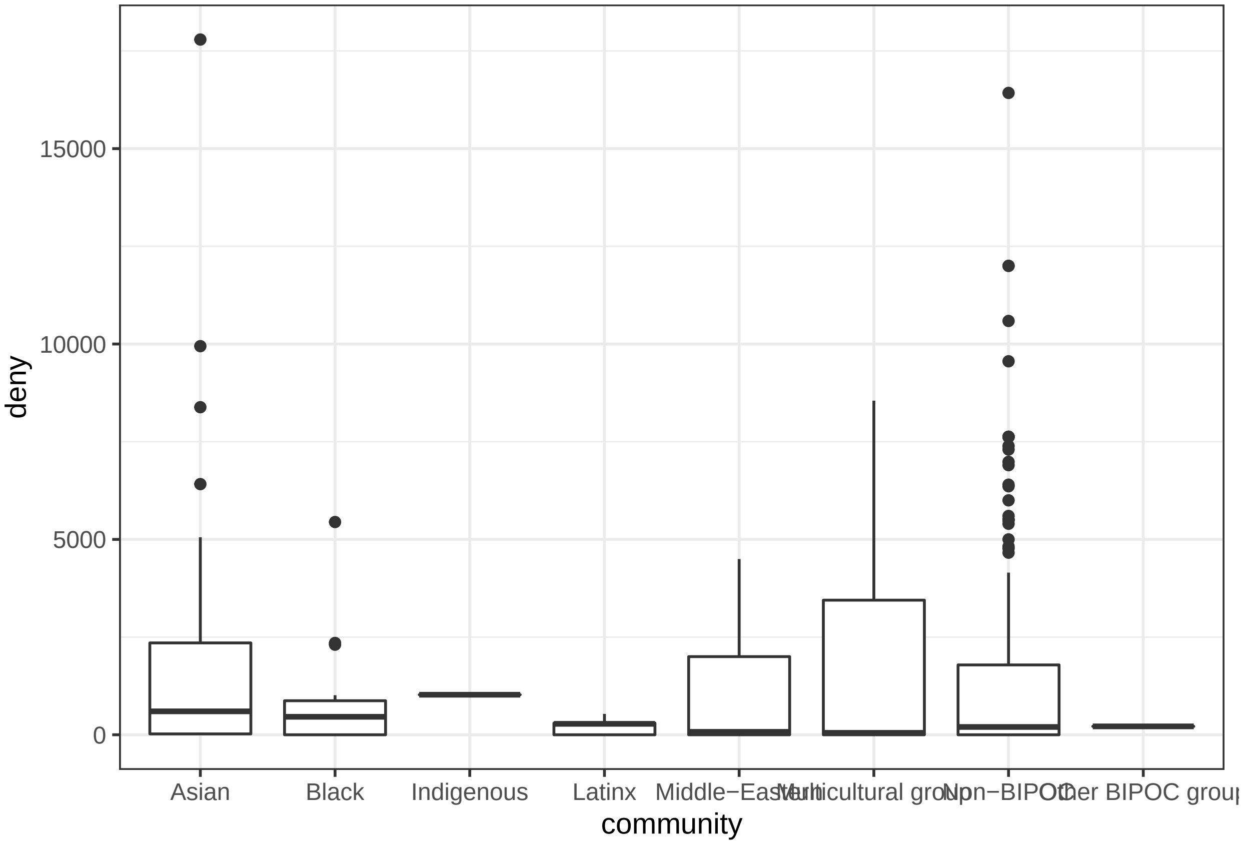
```
prog %>%
  filter(schoolyr == "2018-2019") %>%
  ggplot(aes(x = community, y = prop_grant)) +
  labs(title = "2018-2019") +
  geom_boxplot() +
  theme_bw()
```



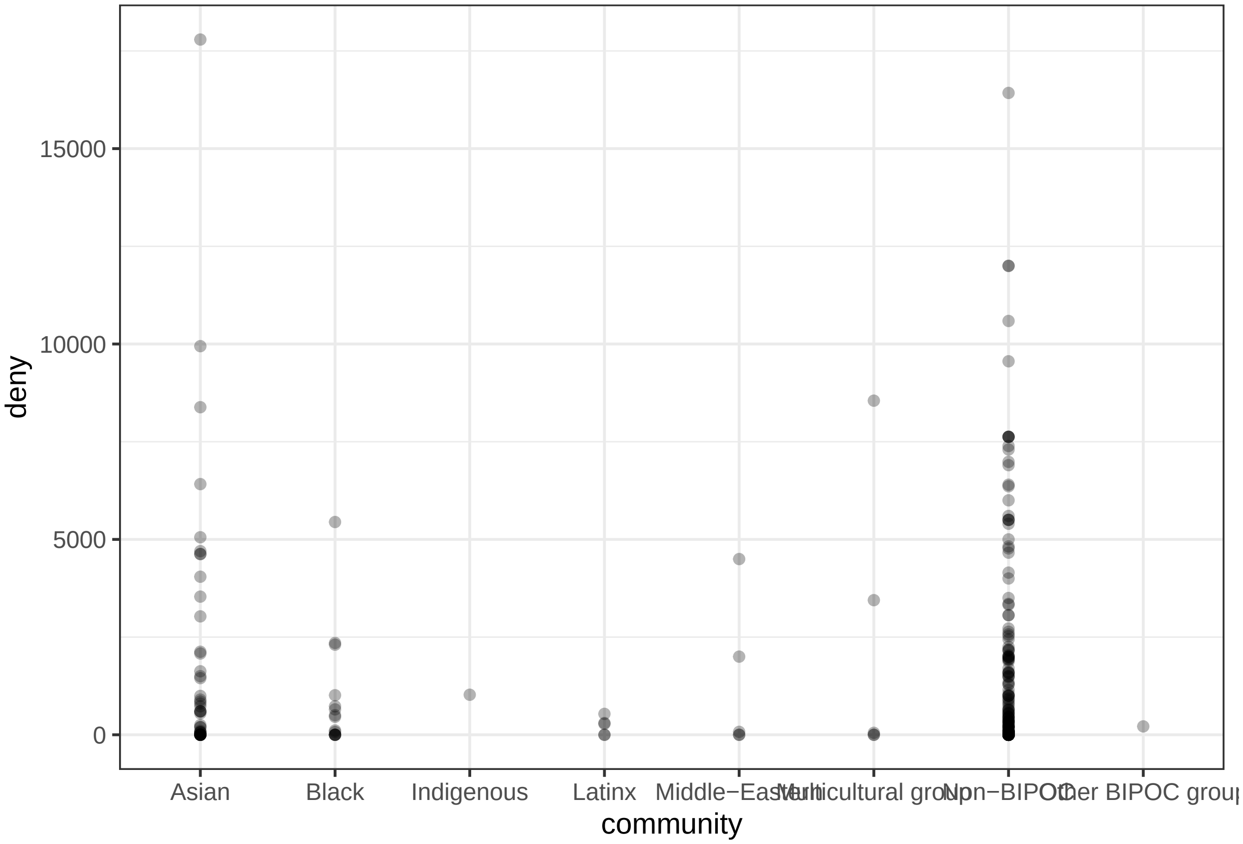
```
prog %>%
  filter(schoolyr == "2019-2020") %>%
  ggplot(aes(x = community, y = prop_grant)) +
  labs(title = "2019-2020") +
  geom_boxplot() +
  theme_bw()
```



```
ggplot(prog, aes(x = community, y = deny)) +
  geom_boxplot() +
  theme_bw()
```

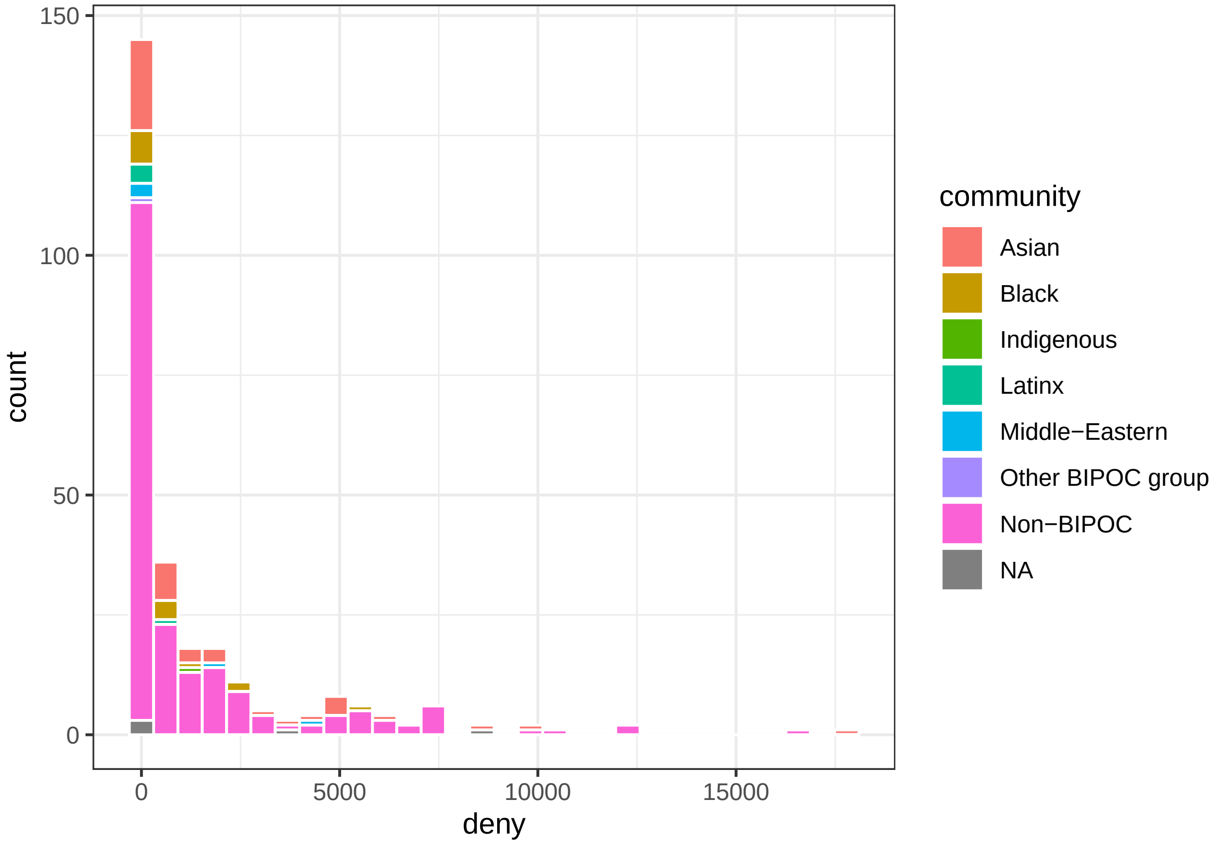


```r
ggplot(prog, aes(x = community, y = deny)) +
  geom_point(alpha = 0.3) +
  theme_bw()
```



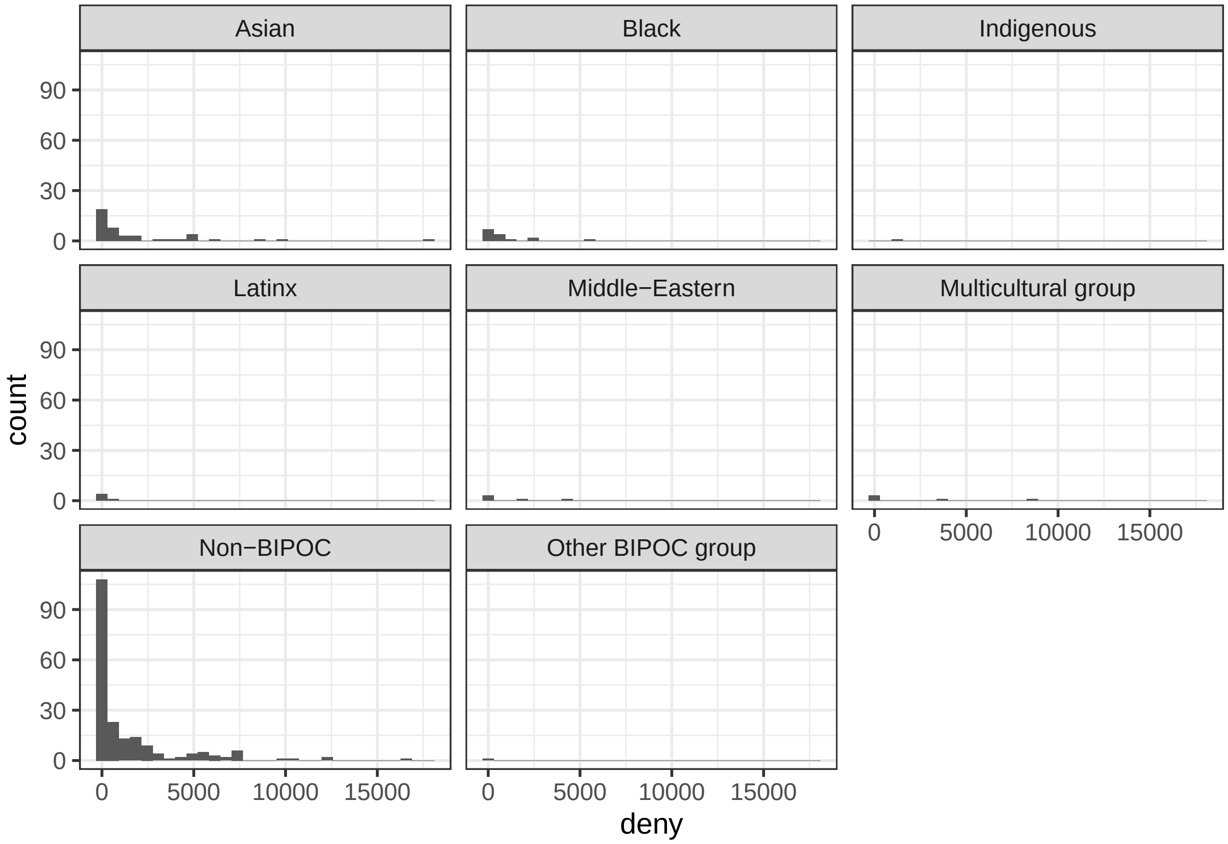
```
ggplot(prog, aes(x = deny)) +
  geom_histogram(aes(fill = factor(community, levels=c("Asian", "Black", "Indigenous", "Latinx", "Middl
               position = "stack", color = "white") +
  scale_fill_discrete(name = "community") +
  theme_bw()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
ggplot(prog, aes(x = deny)) +
  geom_histogram() +
  facet_wrap(. ~ community) +
  theme_bw()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
aggregate(prog$prop_grant, list(prog$org), mean) %>%
  arrange(desc(x)) %>%
  head(10)
```

```
##                                     Group.1 x
## 1                          Acapella Council 1
## 2                     Amnesty International 1
## 3                  Asian American Alliance  1
## 4  Asian Intervarsity Christian Fellowship 1
## 5                                Brownstone 1
## 6                       CrossFit Blue Devil 1
## 7                             Devilish Keys 1
## 8                       Duke Amandla Chorus 1
## 9                              Duke Archery 1
## 10                      Duke Chinese Dance  1
```

```
aggregate(prog$prop_grant, list(prog$community), mean) %>%
  arrange(desc(x))
```

```
##               Group.1         x
## 1              Latinx 0.9208275
## 2               Black 0.7907999
## 3           Non-BIPOC 0.7826702
## 4 Multicultural group 0.7363779
## 5               Asian 0.7292228
## 6      Middle-Eastern 0.7039526
## 7          Indigenous 0.6465517
## 8   Other BIPOC group 0.2293907
```

```
aggregate(prog$grant, list(prog$org), sum) %>%
  arrange(desc(x)) %>%
  head(10)
```

```
##                            Group.1        x
## 1               Blue Devils United 53714.07
## 2       Asian Students Association 35721.00
## 3              Blue Devils United  33139.00
## 4             Duke Catholic Center 30565.61
## 5             Duke Chinese Theater 28713.25
## 6           Duke Conservation Tech 25905.50
## 7                         TEDxDuke 25895.00
## 8     National Panhellenic Council 23179.35
## 9                        Duke Diya 21137.75
## 10 Singapore Students Association 20680.00
```

```
aggregate(prog$grant, list(prog$community), sum) %>%
  arrange(desc(x))
```

```
##               Group.1         x
## 1           Non-BIPOC 659775.13
## 2               Asian 147190.64
## 3               Black  47569.71
## 4 Multicultural group  21480.28
## 5      Middle-Eastern  14175.00
## 6              Latinx  12215.00
## 7          Indigenous   1875.00
## 8   Other BIPOC group     64.00
```

```
aggregate(prog$deny, list(prog$community), sum) %>%
  arrange(desc(x))
```

```
##               Group.1         x
## 1           Non-BIPOC 289559.14
## 2               Asian  88525.03
## 3               Black  13633.00
## 4 Multicultural group  12045.00
## 5      Middle-Eastern   6572.00
## 6              Latinx   1114.99
## 7          Indigenous   1025.00
## 8   Other BIPOC group    215.00
```

```
# ANOVA for programming funds
model_bipoc <- lm(prop_grant ~ bipoc, data = prog)
kbl(model_bipoc %>% tidy(conf.int=TRUE),digits=3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.777	0.020	38.242	0.000	0.737	0.816
bipocY	-0.013	0.035	-0.363	0.717	-0.083	0.057

```r
kbl(tidy(aov(model_bipoc)),digits=3)
```

term	df	sumsq	meansq	statistic	p.value
bipoc	1	0.010	0.010	0.132	0.717
Residuals	273	20.823	0.076	NA	NA

```r
model_comm <- lm(prop_grant ~ community,data=prog)
kbl(model_comm %>% tidy(conf.int=TRUE),digits=3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.729	0.042	17.554	0.000	0.647	0.811
communityBlack	0.062	0.082	0.747	0.455	-0.101	0.224
communityIndigenous	-0.083	0.279	-0.297	0.767	-0.631	0.466
communityLatinx	0.192	0.130	1.473	0.142	-0.064	0.448
communityMiddle-Eastern	-0.025	0.130	-0.194	0.846	-0.281	0.231
communityMulticultural group	0.007	0.130	0.055	0.956	-0.249	0.263
communityNon-BIPOC	0.053	0.046	1.164	0.245	-0.037	0.144
communityOther BIPOC group	-0.500	0.279	-1.794	0.074	-1.049	0.049

```r
kbl(tidy(aov(model_comm)),digits=3)
```

term	df	sumsq	meansq	statistic	p.value
community	7	0.559	0.080	1.051	0.396
Residuals	267	20.274	0.076	NA	NA

```r
# ANOVA for budget funds 2019-2021
model_bipoc <- lm(prop_grant ~ bipoc, data = budget)
kbl(model_bipoc %>% tidy(conf.int=TRUE),digits=3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.462	0.019	24.934	0.000	0.425	0.498
bipocY	-0.057	0.037	-1.546	0.123	-0.129	0.015

```r
kbl(tidy(aov(model_bipoc)),digits=3)
```

term	df	sumsq	meansq	statistic	p.value
bipoc	1	0.170	0.170	2.39	0.123
Residuals	276	19.602	0.071	NA	NA

```r
model_comm <- lm(prop_grant ~ community,data=budget)
kbl(model_comm %>% tidy(conf.int=TRUE),digits=3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.421	0.049	8.667	0.000	0.326	0.517
communityBlack	-0.100	0.075	-1.333	0.184	-0.247	0.047
communityLatinx	0.074	0.101	0.731	0.465	-0.125	0.273
communityMiddle-Eastern	0.080	0.195	0.413	0.680	-0.303	0.463
communityMulticultural group	-0.031	0.129	-0.240	0.811	-0.284	0.222
communityNon-BIPOC	0.043	0.052	0.827	0.409	-0.059	0.145
communityOther BIPOC group	-0.253	0.271	-0.933	0.351	-0.786	0.280

```r
kbl(tidy(aov(model_comm)),digits=3)
```

term	df	sumsq	meansq	statistic	p.value
community	6	0.549	0.091	1.289	0.262
Residuals	271	19.223	0.071	NA	NA

```r
# ANOVA for budget funds 2020-2021
budget2021 <- budget %>%
  filter(schoolyr == "2020-2021")

model_bipoc <- lm(prop_grant ~ bipoc, data = budget2021)
kbl(model_bipoc %>% tidy(conf.int=TRUE),digits=3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.518	0.028	18.188	0.000	0.462	0.575
bipocY	-0.108	0.054	-1.996	0.048	-0.215	-0.001

```r
kbl(tidy(aov(model_bipoc)),digits=3)
```

term	df	sumsq	meansq	statistic	p.value
bipoc	1	0.294	0.294	3.983	0.048
Residuals	124	9.160	0.074	NA	NA

```r
model_comm <- lm(prop_grant ~ community,data=budget2021)
kbl(model_comm %>% tidy(conf.int=TRUE),digits=3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.400	0.071	5.672	0.000	0.260	0.540
communityBlack	-0.078	0.111	-0.702	0.484	-0.299	0.143
communityLatinx	0.069	0.141	0.492	0.624	-0.210	0.349
communityMiddle-Eastern	0.102	0.206	0.495	0.621	-0.305	0.509
communityMulticultural group	0.056	0.206	0.270	0.788	-0.352	0.463
communityNon-BIPOC	0.122	0.076	1.607	0.111	-0.028	0.273

```r
kbl(tidy(aov(model_comm)),digits=3)
```

term	df	sumsq	meansq	statistic	p.value
community	5	0.504	0.101	1.353	0.247
Residuals	120	8.950	0.075	NA	NA

```r
# ANOVA for budget funds 2019-2020
budget2021 <- budget %>%
  filter(schoolyr == "2019-2020")

model_bipoc <- lm(prop_grant ~ bipoc, data = budget2021)
kbl(model_bipoc %>% tidy(conf.int=TRUE),digits=3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.418	0.024	17.487	0.000	0.370	0.465
bipocY	-0.017	0.049	-0.355	0.723	-0.114	0.080

```r
kbl(tidy(aov(model_bipoc)),digits=3)
```

term	df	sumsq	meansq	statistic	p.value
bipoc	1	0.008	0.008	0.126	0.723
Residuals	150	9.924	0.066	NA	NA

```r
model_comm <- lm(prop_grant ~ community,data=budget2021)
kbl(model_comm %>% tidy(conf.int=TRUE),digits=3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.443	0.067	6.659	0.000	0.311	0.574
communityBlack	-0.121	0.100	-1.214	0.227	-0.318	0.076
communityLatinx	0.085	0.145	0.588	0.558	-0.201	0.372
communityMulticultural group	-0.096	0.163	-0.587	0.558	-0.418	0.226
communityNon-BIPOC	-0.024	0.071	-0.339	0.735	-0.164	0.116
communityOther BIPOC group	-0.274	0.266	-1.031	0.304	-0.800	0.252

```r
kbl(tidy(aov(model_comm)),digits=3)
```

term	df	sumsq	meansq	statistic	p.value
community	5	0.243	0.049	0.733	0.6
Residuals	146	9.690	0.066	NA	NA

```r
# ANOVA for SOFC programming totals (right now this is only 2017-2018)
model_bipoc <- lm(prop_grant ~ bipoc, data = sofc)
kbl(model_bipoc %>% tidy(conf.int=TRUE),digits=3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.635	0.067	9.419	0.00	0.496	0.775
bipocY	-0.112	0.112	-0.995	0.33	-0.344	0.121

```r
kbl(tidy(aov(model_bipoc)),digits=3)
```

term	df	sumsq	meansq	statistic	p.value
bipoc	1	0.072	0.072	0.99	0.33
Residuals	23	1.675	0.073	NA	NA

```r
model_comm <- lm(prop_grant ~ community,data=sofc)
kbl(model_comm %>% tidy(conf.int=TRUE),digits=3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.535	0.126	4.260	0.000	0.274	0.797
communityBlack	-0.063	0.205	-0.307	0.762	-0.490	0.364
communityMulticultural group	0.084	0.308	0.273	0.787	-0.556	0.724
communityNon-BIPOC	0.100	0.144	0.696	0.494	-0.199	0.400

```
kbl(tidy(aov(model_comm)),digits=3)
```

term	df	sumsq	meansq	statistic	p.value
community	3	0.090	0.030	0.38	0.769
Residuals	21	1.657	0.079	NA	NA

```
# ANOVA for budget funds from source unfiltered
model_bipoc <- lm(prop_grant ~ bipoc, data = budget_unfilt)
kbl(model_bipoc %>% tidy(conf.int=TRUE),digits=3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.529	0.030	17.515	0.000	0.469	0.589
bipocY	-0.112	0.052	-2.175	0.032	-0.214	-0.010

```
kbl(tidy(aov(model_bipoc)),digits=3)
```

term	df	sumsq	meansq	statistic	p.value
bipoc	1	0.280	0.280	4.729	0.032
Residuals	97	5.753	0.059	NA	NA

```
model_comm <- lm(prop_grant ~ community,data = budget_unfilt)
kbl(model_comm %>% tidy(conf.int=TRUE),digits=3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.464	0.073	6.339	0.000	0.319	0.610
communityBlack	-0.155	0.104	-1.497	0.138	-0.361	0.051
communityLatinx	0.013	0.123	0.109	0.913	-0.231	0.258
communityMiddle-Eastern	-0.102	0.158	-0.645	0.520	-0.416	0.212
communityMulticultural group	-0.009	0.187	-0.048	0.962	-0.380	0.362
communityNon-BIPOC	0.069	0.079	0.868	0.387	-0.088	0.226

```
kbl(tidy(aov(model_comm)),digits=3)
```

term	df	sumsq	meansq	statistic	p.value
community	5	0.541	0.108	1.833	0.114
Residuals	93	5.492	0.059	NA	NA

```
# ANOVA for budget funds from source filtered
model_bipoc <- lm(prop_grant ~ bipoc, data = budget_filt)
kbl(model_bipoc %>% tidy(conf.int=TRUE),digits=3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.628	0.034	18.620	0.000	0.561	0.695
bipocY	-0.176	0.058	-3.063	0.003	-0.291	-0.062

```
kbl(tidy(aov(model_bipoc)),digits=3)
```

term	df	sumsq	meansq	statistic	p.value
bipoc	1	0.694	0.694	9.385	0.003
Residuals	97	7.173	0.074	NA	NA

```
model_comm <- lm(prop_grant ~ community,data=budget_filt)
kbl(model_comm %>% tidy(conf.int=TRUE),digits=3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.498	0.082	6.087	0.000	0.336	0.661
communityBlack	-0.165	0.116	-1.428	0.157	-0.395	0.065
communityLatinx	0.019	0.138	0.140	0.889	-0.254	0.293
communityMiddle-Eastern	-0.095	0.177	-0.535	0.594	-0.446	0.256
communityMulticultural group	0.055	0.209	0.262	0.794	-0.360	0.469
communityNon-BIPOC	0.133	0.088	1.501	0.137	-0.043	0.308

```
kbl(tidy(aov(model_comm)),digits=3)
```

term	df	sumsq	meansq	statistic	p.value
community	5	1.019	0.204	2.766	0.022
Residuals	93	6.849	0.074	NA	NA