help

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.0     v forcats 0.5.1
```

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

```r
library(ggcorrplot)
library(kableExtra)      # for printing tables
```

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

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

```r
dropout_data = read_tsv("../data/clean/dropout-data.tsv")
```

```
## Rows: 1247 Columns: 27

## -- Column specification ---------------------------------------------------
## Delimiter: "\t"
## chr  (3): fips, state, county
## dbl (24): num_of_schools, male, white, black, asian, hispanic, married, sing...

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```
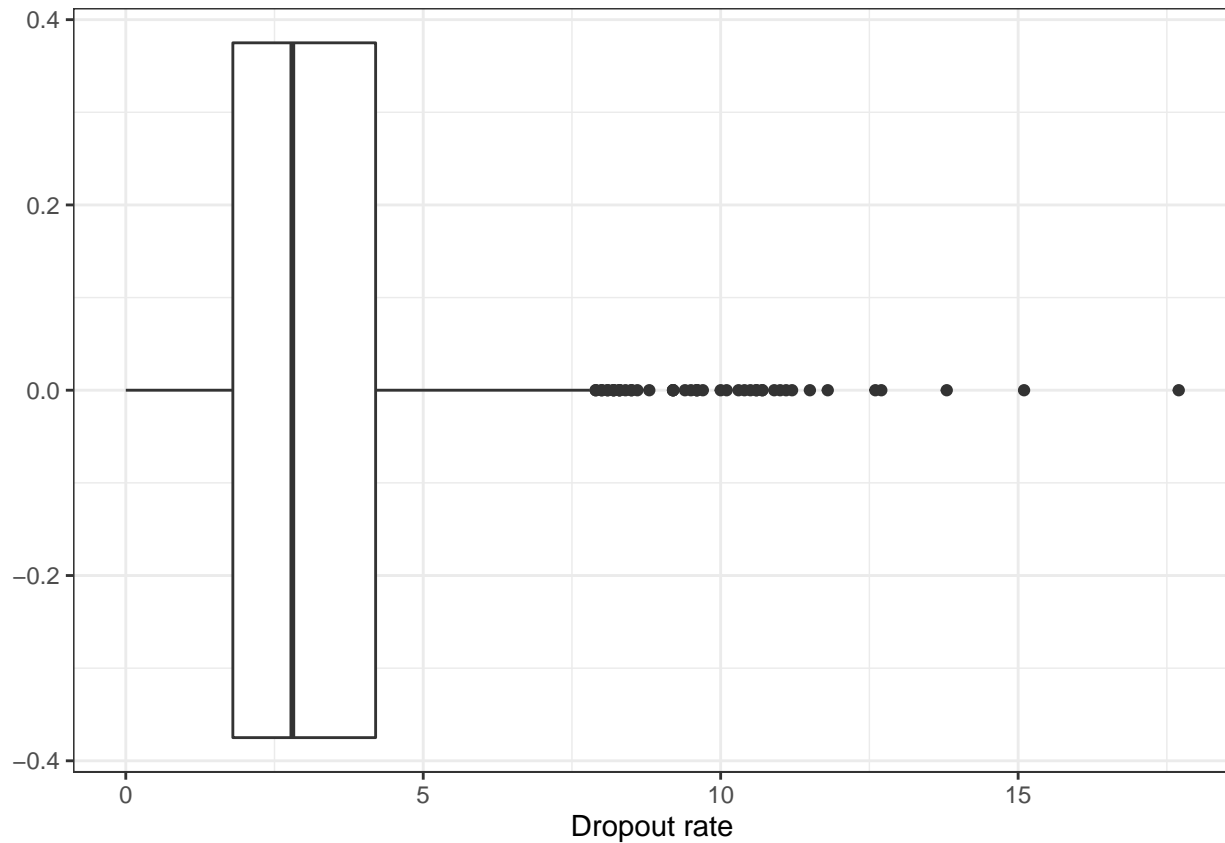
```r
# calculate median dropout rate
median_dropout_rate = dropout_data %>%
  summarise(median(dropout_rate)) %>%
  pull()
```

```
# remove extreme outlier
dropout_data = dropout_data %>% filter(dropout_rate < 40)
```

```
# create histogram of dropout rate
dropout_data %>%
  ggplot(aes(x=dropout_rate)) +
  geom_boxplot() +
  labs(x = "Dropout rate") +
  theme_bw() + theme(legend.position = "none")
```



```
dropout_data %>%
  arrange(desc(dropout_rate)) %>%
  select(county, state, dropout_rate) %>%
  head(10) %>%
  rename(County = county,
         State = state,
         `Dropout Rate` = dropout_rate) %>%
  kable(format = "latex", row.names = NA,
        booktabs = TRUE, digits = 2,
        caption = "The selected counties with the highest dropout rates") %>%
  kable_styling(position = "center")
```

Table 1: The selected counties with the highest dropout rates

County	State	Dropout Rate
Red River Parish	Louisiana	17.7
Morehouse Parish	Louisiana	15.1
Denver County	Colorado	13.8
Huerfano County	Colorado	12.7
Madison Parish	Louisiana	12.6
Mahnomen County	Minnesota	11.8
Lake County	Colorado	11.5
Iberville Parish	Louisiana	11.2
Union Parish	Louisiana	11.1
Beltrami County	Minnesota	11.0

```r
# Demographics
demographic_variables = dropout_data %>%
  select(male,
         white,
         black,
         asian,
         hispanic,
         married,
         single_mom,
         #bachelors_only,
         advanced_educ,
         exp_1_5,
         exp_6_10,
         exp_11_20,
         exp_21_30,
         exp_gt_30)

demographic_cor = cor(demographic_variables)

# Economic
economic_variables = dropout_data %>%
  select(gini_index,
         per_capita_income,
         welfare,
         #mean_salary,
         mbsal_1_5,
         mbsal_6_10,
         mbsal_11_20,
         mbsal_21_30,
         mbsal_gt_30)

economic_cor = cor(economic_variables)

# total
total_variables = dropout_data %>%
  select(gini_index,
         per_capita_income,
```
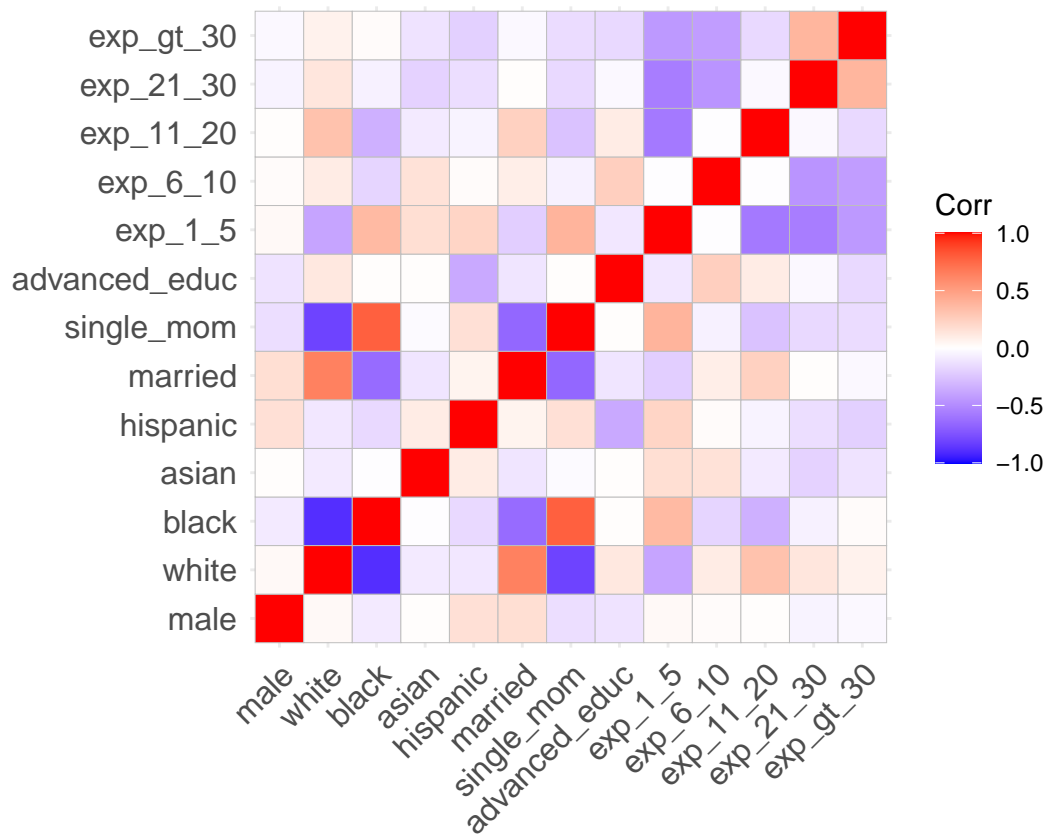
```
         welfare,
         #mean_salary,
         mbsal_1_5,
         mbsal_6_10,
         mbsal_11_20,
         mbsal_21_30,
         mbsal_gt_30,
         male,
         white,
         black,
         asian,
         hispanic,
         married,
         single_mom,
         #bachelors_only,
         advanced_educ,
         exp_1_5,
         exp_6_10,
         exp_11_20,
         exp_21_30,
         exp_gt_30)
total_cor = cor(total_variables)
```
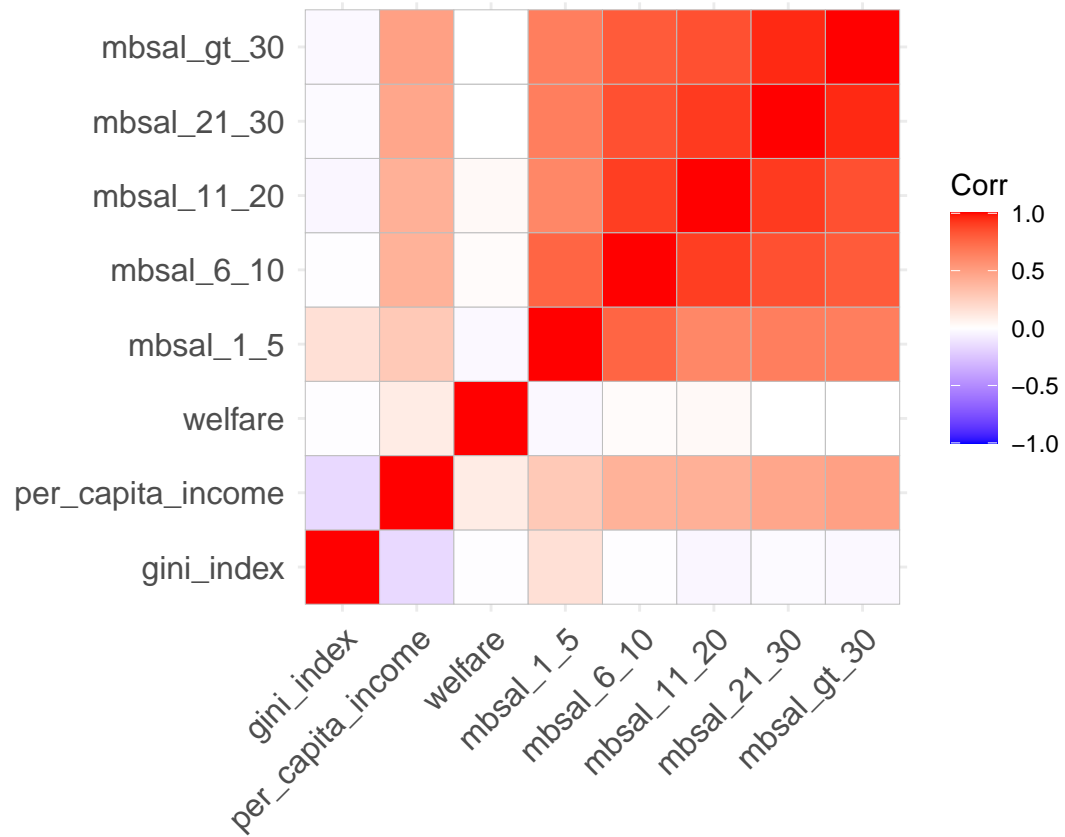
```
ggcorrplot(demographic_cor)
```

```
ggcorrplot(economic_cor)
```



```
ggcorrplot(total_cor)
```