Statistical Tests

Load and filter the data

```r
library(tidyverse)
library(scales)
library(DT)
library(forcats)
library(lubridate)
library(gridExtra)
library(broom)

source('../helpers.R')

# already filtered to only interesting divisions
stops <- read_rds('../../processed_data/prepared_stops.rds')
# filter to only BLACKS, ASIANS, HISPANICS, and WHITES
# only during INTERTWILIGHT TIME
# only VEHICLE STOPS

filt_stops <- stops %>%
  filter(
    !weekdays(STOP_DT) %in% c('Saturday', 'Sunday')
  )

dim(filt_stops) # ~ 43k stops
```

```
## [1] 43652    22
```

Group the data

```r
race_sum <- filt_stops %>%
  group_by(DIV1_DESC, is_daylight) %>%
  summarize(
    total_stops = n(),
    white_stops = sum(DESCENT_DESC == 'WHITE'),
    black_stops = sum(DESCENT_DESC == 'BLACK'),
    hispanic_stops = sum(DESCENT_DESC == 'HISPANIC'),
    # calculate distribution % of race across day / night stops
    white_perc = white_stops / total_stops,
    black_perc = black_stops / total_stops,
    hispanic_perc = hispanic_stops / total_stops
  )

race_diffs <- race_sum %>%
  select(div = DIV1_DESC, contains('perc')) %>%
  mutate(
    white_diff = lag(white_perc) - white_perc,
    black_diff = lag(black_perc) - black_perc,
    hisp_diff = lag(hispanic_perc) - hispanic_perc
```

```
) %>%
filter(!is.na(white_diff)) %>%
ungroup() %>%
mutate(
    div_by_white = fct_reorder(div, white_diff, .desc = T),
    div_by_black = fct_reorder(div, black_diff, .desc = T),
    div_by_hisp = fct_reorder(div, hisp_diff, .desc = T)
)


white_p <- race_diffs %>%
  gen_race_diff_plot(race_diffs$div_by_white, race_diffs$white_diff, 'Whites')
black_p <- race_diffs %>%
  gen_race_diff_plot(race_diffs$div_by_black, race_diffs$black_diff, 'Blacks')
hispanic_p <- race_diffs %>%
  gen_race_diff_plot(race_diffs$div_by_hisp, race_diffs$hisp_diff, 'Hispanics')
```

Note: Divisions that stop more often during visible hours are at the top.

```
prop_diff_plot <- grid.arrange(
  white_p, black_p, hispanic_p, nrow = 1,
  top = "Differences in Proportion of Night Stops - Day Stops",
  bottom = 'Divisions at the top stop more during visible hours than non-visible'
)
```


Differences in Proportion of Night Stops - Day Stops

Divisions at the top stop more during visible hours than non-visible
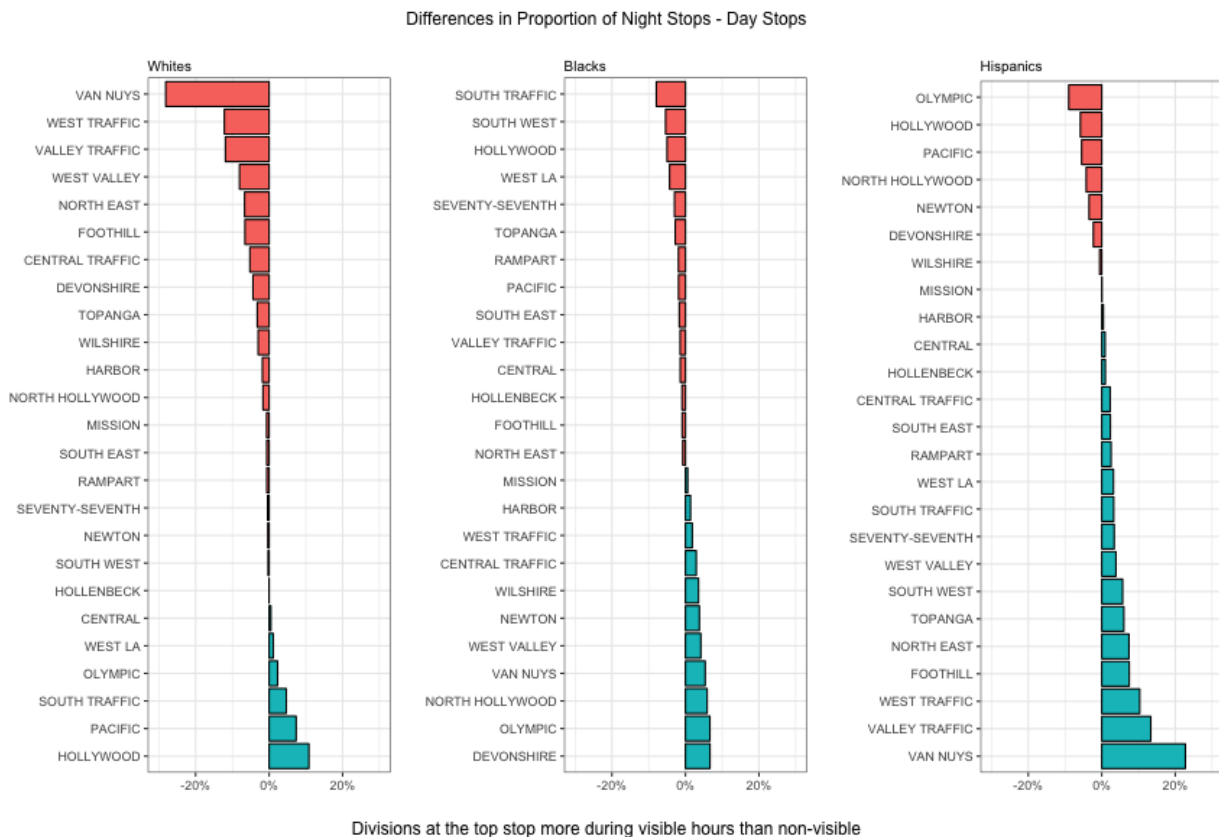
Perform Proportion Test

H_a: Day Proportion > Night Proportion

```r
white_prop_tests <- prop_test_all_divs(race_sum, 'white_stops') %>%
  mutate(is_significant = p.value < 0.05)
black_prop_tests <- prop_test_all_divs(race_sum, 'black_stops') %>%
  mutate(is_significant = p.value < 0.05)
hisp_prop_tests <- prop_test_all_divs(race_sum, 'hispanic_stops') %>%
  mutate(is_significant = p.value < 0.05)

white_plot <- gen_race_prop_test_plot(white_prop_tests, 'white_stops', 'White')
black_plot <- gen_race_prop_test_plot(black_prop_tests, 'black_stops', 'Blacks')
hisp_plot <- gen_race_prop_test_plot(hisp_prop_tests, 'hispanic_stops', 'Hispanics')
```
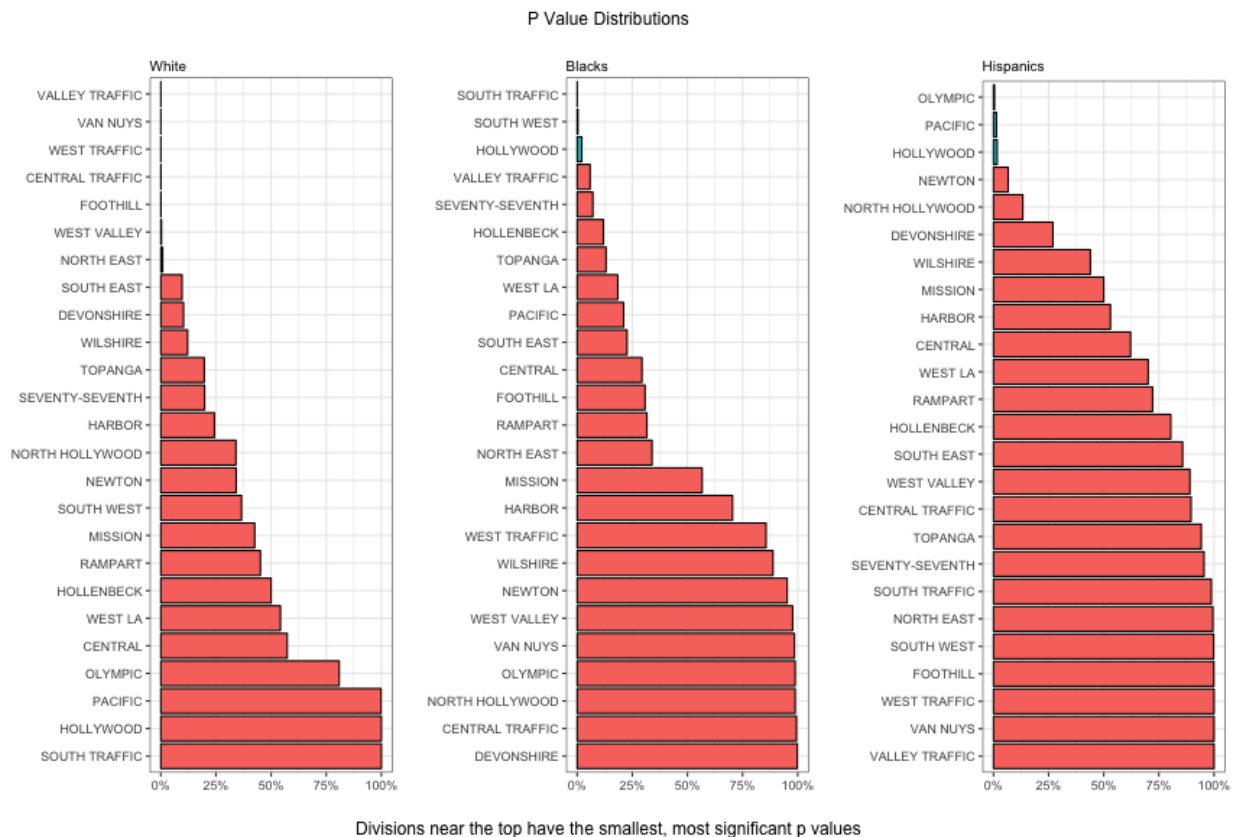
Again, Divisions listed near the top show significant p values; That is, the blue divisions would be divisions that we would want to analze more closely for appearances of racial discrimination.

```r
grid.arrange(
  white_plot, black_plot, hisp_plot,
  nrow = 1, top = "P Value Distributions",
  bottom = 'Divisions near the top have the smallest, most significant p values'
)
```



```r
white_df <- white_prop_tests %>%
  select(division, race, p.value) %>%
  mutate(is_bonferroni_significant = p.value < 0.05 / length(unique(division)))
black_df <- black_prop_tests  %>%
  select(division, race, p.value) %>%
```

```r
  mutate(is_bonferroni_significant = p.value < 0.05 / length(unique(division)))
hisp_df <- hisp_prop_tests %>%
  select(division, race, p.value) %>%
  mutate(is_bonferroni_significant = p.value < 0.05 / length(unique(division)))


# bonferroni correction
bonferroni_divisions <- bind_rows(white_df, black_df, hisp_df) %>%
  filter(is_bonferroni_significant) %>%
  select(-is_bonferroni_significant) %>%
  arrange(p.value) %>%
  mutate(division = fct_reorder(division, p.value),
         race = fct_recode(race,
           blacks = 'black_stops',
           whites = 'white_stops',
           hispanics = 'hispanic_stops'))

bonferroni_divisions %>%
  ggplot(aes(division, p.value, fill = race)) +
  geom_bar(stat = 'identity', color = 'black') +
  theme_bw() +
  theme(legend.position = 'bottom', legend.title = element_blank()) +
  coord_flip() +
  scale_y_continuous(trans = 'log10') +
  labs(
    x = '', y = 'log10(p value)',
      title = 'Whites in Traffic Divisions Would Appear the Most Targeted',
    subtitle = 'smaller, more significant p values <--> larger, less significant p values')
```

Whites in Traffic Divisions Would Appear the Most Targeted

smaller, more significant p values <--> larger, less significant p values