360 Final Project

Group: Flora Shi, Belle Xu

19 April, 2021

```
stop <- read.table("stop-and-frisk.dat", header = TRUE)
```

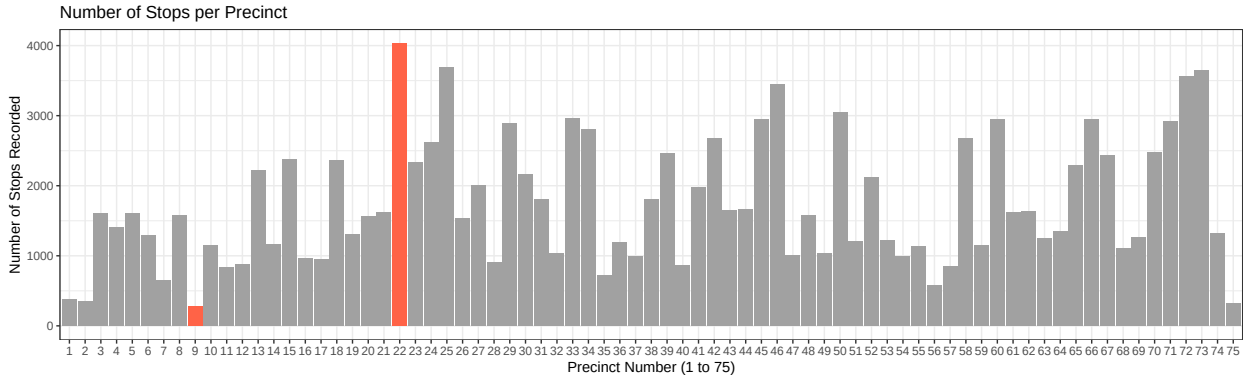
Exploratory Data Analysis

Exploratory data analysis should support project goals and help guide specification of model.

```
#make categorical variables factor
stop <- stop %>%
  mutate(precinct = factor(precinct)) %>%
  mutate(eth = factor(eth, levels = c("3", "1", "2"))) %>%
  mutate(crime = factor(crime))

#crime rate for a certain crime for each eth in a certain precinct
stop <- stop %>% mutate(crime_rate = past.arrests/pop)
#compute population proportion with in each precinct
stop <- stop %>% group_by(precinct) %>%
  mutate(total_pop = sum(pop/4))
stop <- stop %>% mutate(pop_prop = pop/total_pop)
```

```
#make a by-precinct dataframe
stop <- stop %>%
  group_by(precinct) %>%
  mutate(stops_in_precinct = sum(stops))

stops_by_precinct <- stop %>% distinct(stops_in_precinct) %>%
  pull(stops_in_precinct)
precinct <- seq(from = 1, to = 75, by = 1)
pop_by_precinct<- stop %>% group_by(precinct) %>%
  mutate(pop_in_precinct = sum(pop)) %>%
  distinct(pop_in_precinct) %>%
  pull(pop_in_precinct)
by_precinct_df <- data.frame(pop_by_precinct = pop_by_precinct,
                             stops_by_precinct = stops_by_precinct,
                             precinct = as.factor(precinct))
by_precinct_df <- by_precinct_df %>%
  mutate(toHighlight = if_else(stops_by_precinct == min(stops_by_precinct)
                               | stops_by_precinct == max(stops_by_precinct),
                               "yes", "no"))
```

```
#number of stops by precinct
ggplot(data = by_precinct_df,
       mapping = aes(x = precinct, y = stops_by_precinct, fill = toHighlight)) +
  geom_col() +
    scale_fill_manual(values= c("yes"="tomato", "no"="#a1a1a1"), guide = FALSE)+
  labs(title = "Number of Stops per Precinct", x = "Precinct Number (1 to 75)",
       y = "Number of Stops Recorded")
```



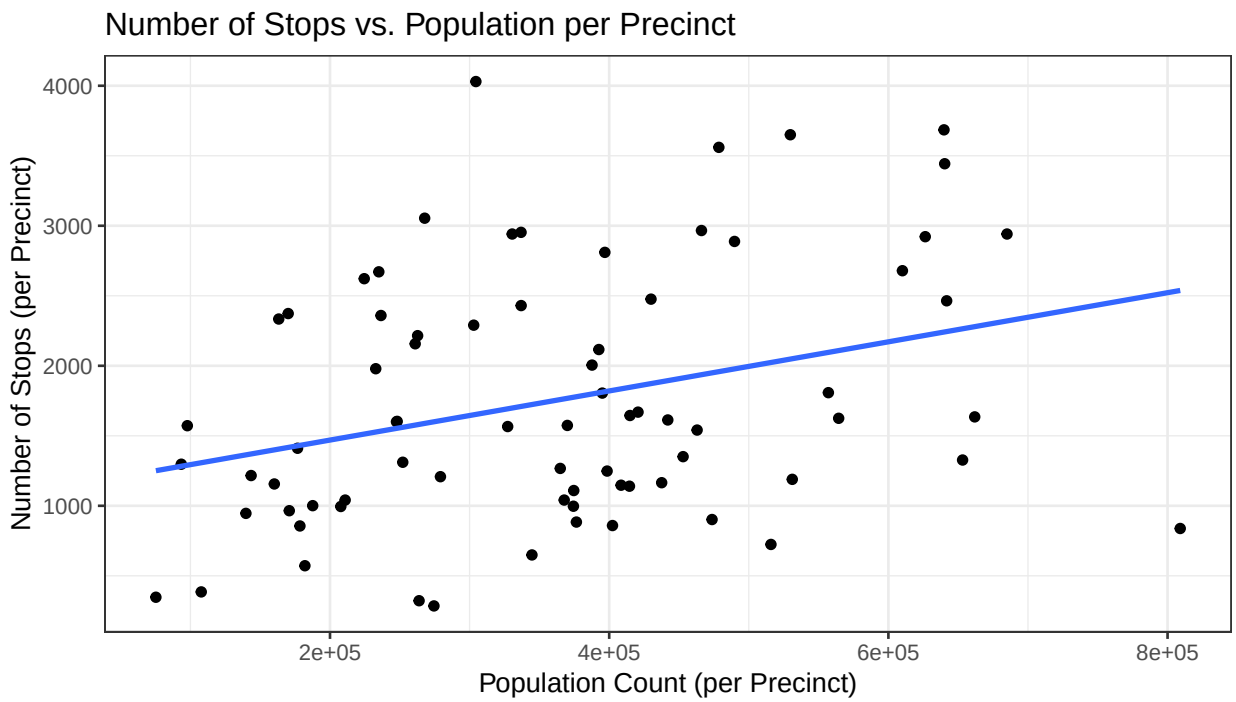

```
#Need interpretation
# get the max and min stops with function and just type it
min(by_precinct_df$stops_by_precinct)
```

```
## [1] 285
```

```
max(by_precinct_df$stops_by_precinct)
```

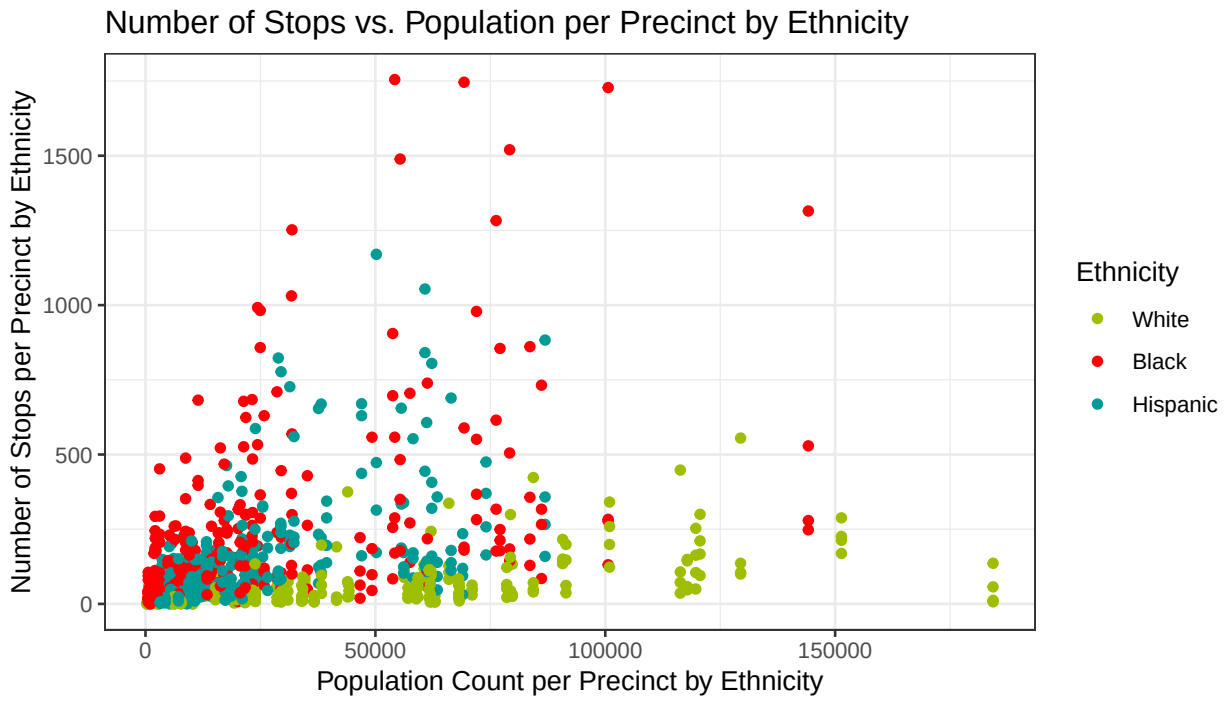
```
## [1] 4030
```

```
#number of stops vs. precinct pop
ggplot(data = by_precinct_df,
       mapping = aes(x = pop_by_precinct, y =stops_by_precinct)) +
  geom_point() +
  labs(title = "Number of Stops vs. Population per Precinct",
       x = "Population Count (per Precinct)",
       y = "Number of Stops (per Precinct)") +
  geom_smooth(method = "lm", formula = y~x, se = FALSE)
```



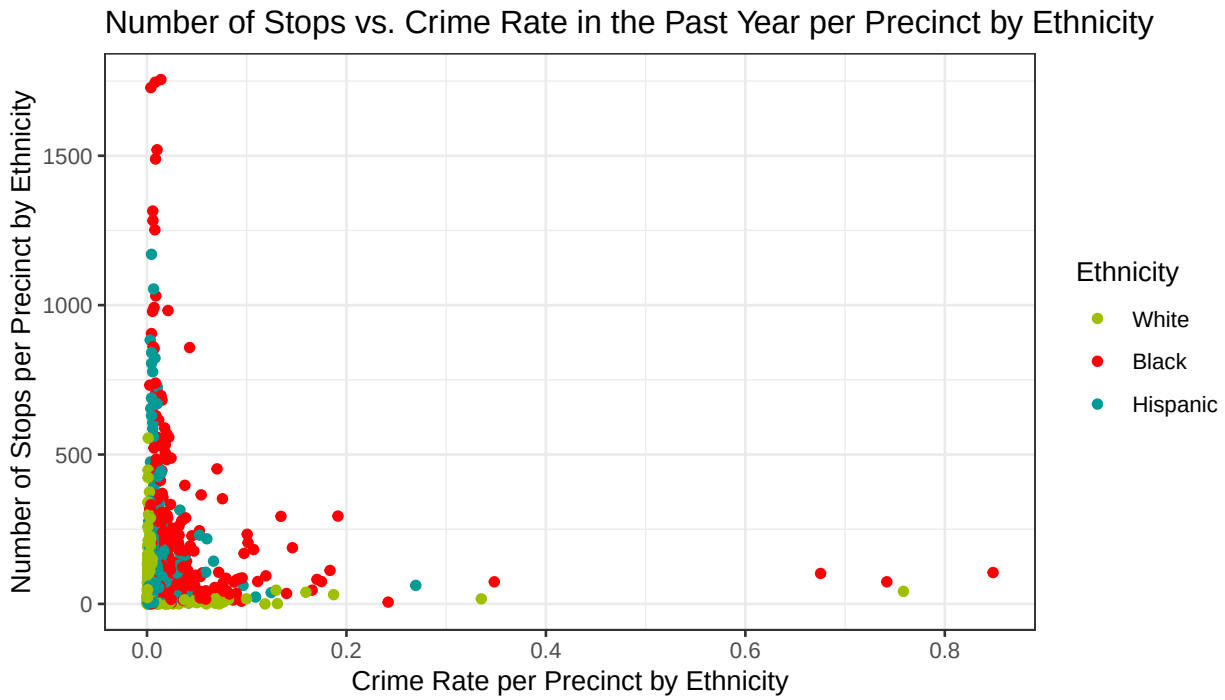

```
#no obvious trend between population in precinct vs. stops in a precinct
#maybe weakly positive
```

```
#number of stops for each ethnicity vs.
#  population for each ethnicity per precinct

ggplot(data = stop, mapping = aes(x = pop, y = stops, colour = eth)) +
  geom_point() +
  labs(title = "Number of Stops vs. Population per Precinct by Ethnicity",
       x = "Population Count per Precinct by Ethnicity",
       y = "Number of Stops per Precinct by Ethnicity",
       colour = "Ethnicity") +
  scale_color_manual(values = c("#9EBE00","#FD0006","#009B95"),
                     labels = c("White", "Black", "Hispanic"))
```




```
#There does seem to be some trend because we can see that none of the
#  points for the white population is in high number of stops
#  despite having very high population
```

```
#num of stops vs crime rate
ggplot(data = stop, mapping = aes(x = crime_rate, y = stops, colour = eth)) +
  geom_point() +
  labs(title = "Number of Stops vs. Crime Rate in the Past Year per Precinct by Ethnicity",
       x = "Crime Rate per Precinct by Ethnicity",
       y = "Number of Stops per Precinct by Ethnicity",
       colour = "Ethnicity") +
  scale_color_manual(values = c("#9EBE00","#FD0006","#009B95"),
                     labels = c("White", "Black", "Hispanic"))
```



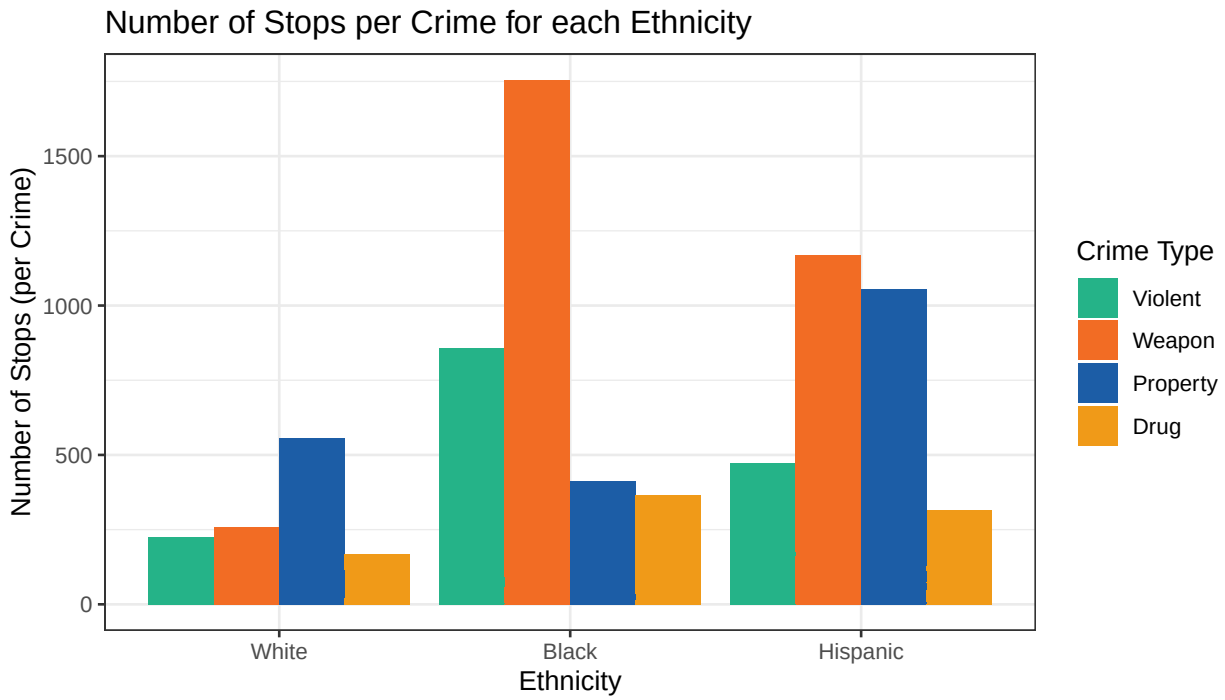

```
#crime rate is generally low with a few exceptions; however black
#  and hispanic stop count are still
#  a lot higher in low crime rate regions
```

```
#get datasets for different crime types
stop.1 <- stop %>% filter(crime == 1)
stop.2 <- stop %>% filter(crime == 2)
stop.3 <- stop %>% filter(crime == 3)
stop.4 <- stop %>% filter(crime == 4)
#investigate mean and var of each stop
Violence <- c(mean(stop.1$stops), var(stop.1$stops))
Weapon <- c(mean(stop.2$stops), var(stop.2$stops))
Property <- c(mean(stop.3$stops), var(stop.3$stops))
Drug <- c(mean(stop.4$stops), var(stop.4$stops))
df <- data.frame(Violence, Weapon, Property, Drug)
row.names(df) <- c("Mean", "Variance")
df
```

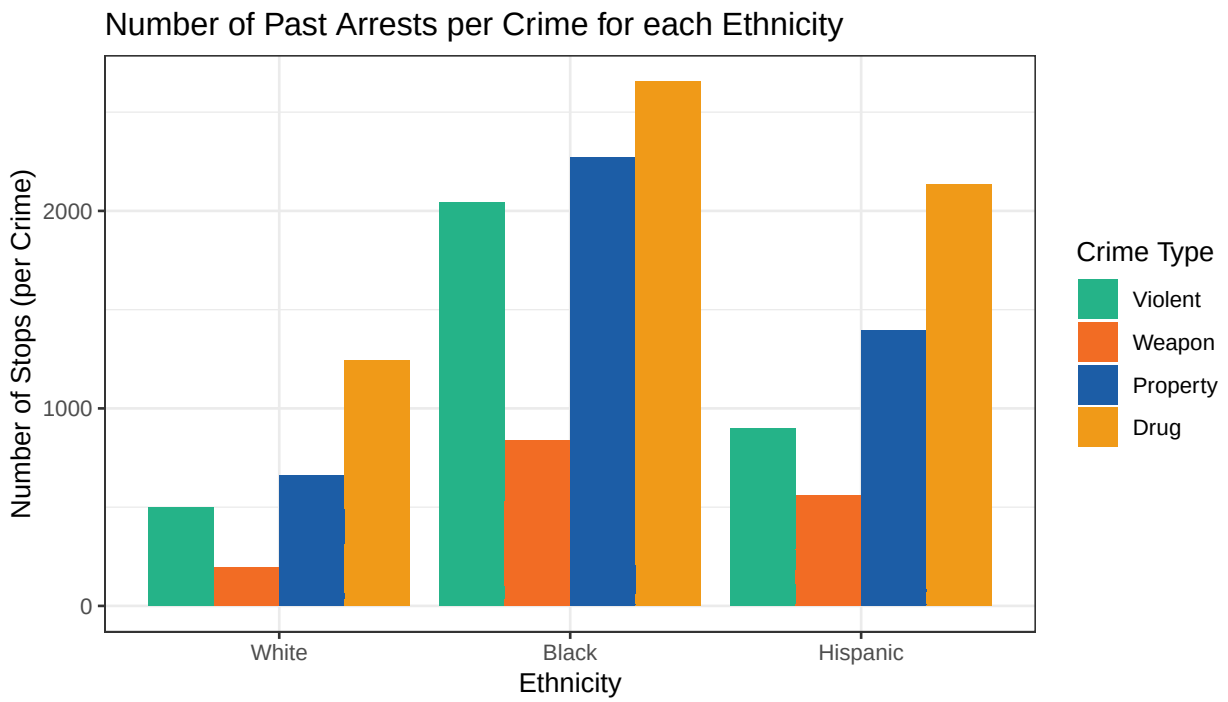
```
##            Violence      Weapon   Property       Drug
## Mean       142.5689    256.8356   117.9822   66.70222
## Variance 22489.0946 123810.0309 18919.3479 4978.23683
```

```
#number of stops per crime for each ethnicity
ggplot(data = stop, mapping = aes(x = eth, y = stops, fill = crime)) +
  geom_col(position = "dodge") +
  labs(title = "Number of Stops per Crime for each Ethnicity",
       x = "Ethnicity",
       y = "Number of Stops (per Crime)",
       fill = "Crime Type") +
  scale_x_discrete(labels=c("3" = "White", "1" = "Black",
```

```
                                "2" = "Hispanic")) +
  scale_fill_manual(values = c("#25B388","#F26C24","#1C5DA6", "#F09A18"),
                    labels = c("Violent", "Weapon",
                               "Property", "Drug"))
```




```
#number of past arrests per crime for each ethnicity
ggplot(data = stop, mapping = aes(x = eth, y = past.arrests, fill = crime)) +
  geom_col(position = "dodge") +
  labs(title = "Number of Past Arrests per Crime for each Ethnicity",
       x = "Ethnicity",
       y = "Number of Stops (per Crime)",
       fill = "Crime Type") +
  scale_x_discrete(labels=c("3" = "White", "1" = "Black",
                            "2" = "Hispanic")) +
  scale_fill_manual(values = c("#25B388","#F26C24","#1C5DA6", "#F09A18"),
                    labels = c("Violent", "Weapon",
                               "Property", "Drug"))
```




```
#from the two plots above, the stop reasons and past arrests seem to have no relation
```

Modeling

Rstan GLM

```
#glm for crime 1; done for testing only. don't repeat the code for other crimes yet.
#assuming negative binom sampling model

stan.glm.1 <- stan_glm(data = stop.1,
                  formula = stops ~ crime_rate+ pop_prop+ eth+ crime_rate:eth+ pop_prop:eth + crime_ra
                  family = neg_binomial_2,
                  seed = 360,
                  prior = cauchy(0, 2.5),
                  prior_intercept = cauchy(0, 2.5),
                  refresh = 0)
```

```
## Warning: There were 3 divergent transitions after warmup. See
## http://mc-stan.org/misc/warnings.html#divergent-transitions-after-warmup
## to find out why this is a problem and how to eliminate them.
```

```
## Warning: Examine the pairs() plot to diagnose sampling problems
```

```
#no r-hat warning = converged. no ess warning = good enough ess. apparently divergent transitions can b
```

```
# this code chunk has been omitted

# first glm but with standard normal prior
# stan.glm.2 <- update(stan.glm.1, prior = normal(0,1), prior_intercept = normal(0,1))
# first glm but with poisson family
# stan.glm.1.pois <- update(stan.glm.1, family = poisson)
# first glm but with standard normal prior & poisson family
# stan.glm.2.pois <- update(stan.glm.2, family = poisson )
```

```
#checking if omitting stuff will make the model better
stan.glm.2 <- update(stan.glm.1, formula = stops ~ crime_rate+ pop_prop+ eth+ crime_rate:eth+ pop_prop:
```

```
## Warning: There were 3 divergent transitions after warmup. See
## http://mc-stan.org/misc/warnings.html#divergent-transitions-after-warmup
## to find out why this is a problem and how to eliminate them.
```

```
## Warning: Examine the pairs() plot to diagnose sampling problems
```

```
stan.glm.3 <- update(stan.glm.2, formula = stops ~ crime_rate+ pop_prop+ eth+ crime_rate:eth)
```

```
## Warning: There were 32 divergent transitions after warmup. See
## http://mc-stan.org/misc/warnings.html#divergent-transitions-after-warmup
## to find out why this is a problem and how to eliminate them.
```

```
## Warning: Examine the pairs() plot to diagnose sampling problems
```

```
stan.glm.4 <- update(stan.glm.3, formula = stops ~ crime_rate + pop_prop + eth)
stan.glm.5 <- update(stan.glm.4, formula = stops ~ crime_rate + pop_prop)
```

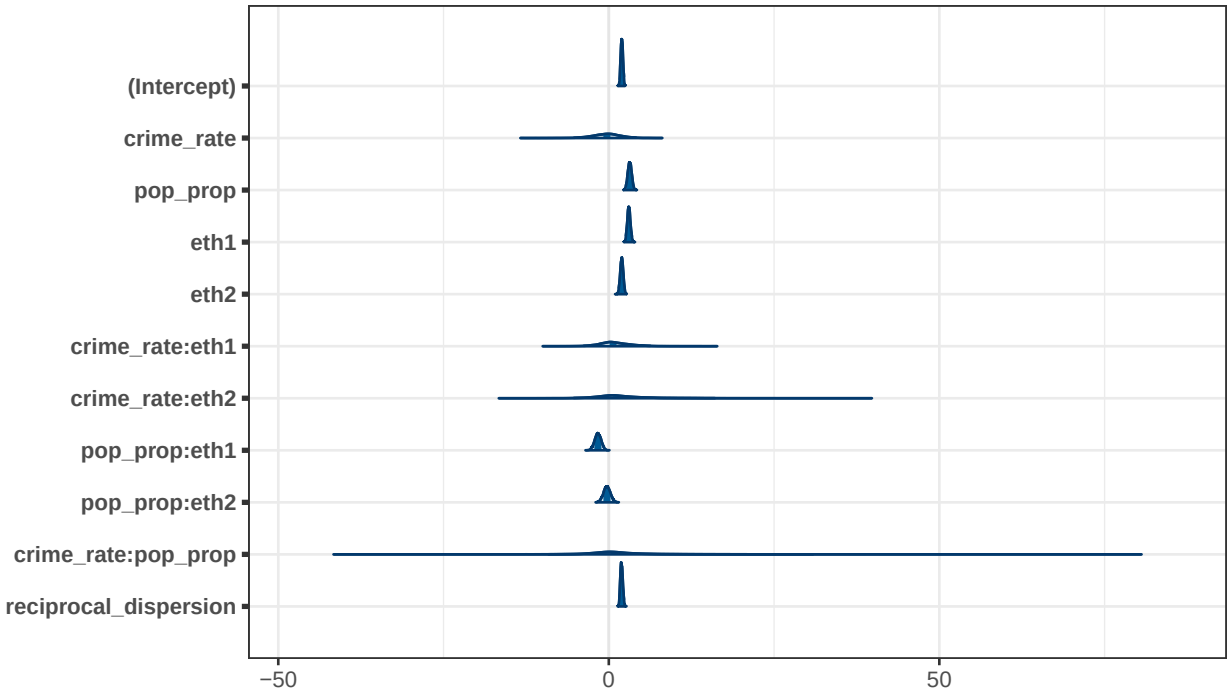
```
#checking if adding/changing stuff will make the model better
#apparently all the code below cause errors so I can't even check... Already tried with different prior

#stan.glm.6 <- update(stan.glm.1, formula = stops ~ past.arrests+ pop_prop+ eth+ past.arrests:eth+ pop_
#stan.glm.7 <- update(stan.glm.1, formula = stops ~ crime_rate+ pop+ eth+ crime_rate:eth+ pop:eth +crim
#stan.glm.8 <- update(stan.glm.1, formula = stops ~ past.arrests+ pop+ eth+ past.arrests:eth+ pop:eth +
```

```
#look at coefficients
summary(stan.glm.1)
```

```
##
## Model Info:
##  function:     stan_glm
##  family:       neg_binomial_2 [log]
##  formula:      stops ~ crime_rate + pop_prop + eth + crime_rate:eth + pop_prop:eth +
##     crime_rate:pop_prop
##  algorithm:    sampling
##  sample:       4000 (posterior sample size)
##  priors:       see help('prior_summary')
##  observations: 225
##  predictors:   10
```

```
##
## Estimates:
##                          mean   sd   10%   50%   90%
## (Intercept)             2.0    0.2  1.7   2.0   2.2
## crime_rate             -0.4    2.3 -3.1  -0.3   2.4
## pop_prop                3.2    0.3  2.8   3.2   3.5
## eth1                    3.0    0.2  2.7   3.0   3.3
## eth2                    2.0    0.2  1.7   2.0   2.2
## crime_rate:eth1         0.9    2.5 -1.9   0.7   4.0
## crime_rate:eth2         2.3    5.2 -2.5   1.2   8.8
## pop_prop:eth1          -1.7    0.5 -2.3  -1.7  -1.0
## pop_prop:eth2          -0.3    0.5 -1.0  -0.3   0.4
## crime_rate:pop_prop     3.0    9.6 -3.8   0.8  12.3
## reciprocal_dispersion   1.9    0.2  1.7   1.9   2.2
##
## Fit Diagnostics:
##            mean   sd    10%   50%   90%
## mean_PPD 150.2   14.2 132.7 149.0 168.8
##
## The mean_ppd is the sample average posterior predictive distribution of the outcome variable (for de
##
## MCMC diagnostics
##                         mcse Rhat n_eff
## (Intercept)             0.0  1.0  2409
## crime_rate              0.0  1.0  2476
## pop_prop                0.0  1.0  2528
## eth1                    0.0  1.0  2123
## eth2                    0.0  1.0  2138
## crime_rate:eth1         0.0  1.0  2562
## crime_rate:eth2         0.1  1.0  2914
## pop_prop:eth1           0.0  1.0  2178
## pop_prop:eth2           0.0  1.0  2621
## crime_rate:pop_prop     0.2  1.0  2895
## reciprocal_dispersion   0.0  1.0  3951
## mean_PPD                0.2  1.0  4252
## log-posterior           0.0  1.0  1885
##
## For each parameter, mcse is Monte Carlo standard error, n_eff is a crude measure of effective sample
```

```
#posterior distributions
mcmc_areas(as.matrix(stan.glm.1), prob = 0.95, prob_outer = 1)
```




```
round(coef(stan.glm.1), 3)
```

```
##         (Intercept)          crime_rate            pop_prop                eth1
##               1.967              -0.287               3.161               3.028
##                eth2     crime_rate:eth1     crime_rate:eth2       pop_prop:eth1
##               1.957               0.650               1.183              -1.651
##       pop_prop:eth2 crime_rate:pop_prop
##              -0.283               0.837
```

```
round(posterior_interval(stan.glm.1, prob = 0.95), 3)
```

```
##                           2.5%  97.5%
## (Intercept)              1.641  2.293
## crime_rate              -5.258  4.099
## pop_prop                 2.592  3.739
## eth1                     2.586  3.494
## eth2                     1.509  2.397
## crime_rate:eth1         -3.796  6.271
## crime_rate:eth2         -5.287 16.060
## pop_prop:eth1           -2.618 -0.725
## pop_prop:eth2           -1.296  0.743
## crime_rate:pop_prop     -9.121 30.896
## reciprocal_dispersion    1.591  2.300
```

```
#posterior predictive check
loo1 <- loo(stan.glm.1, save_psis = TRUE)
loo2 <- loo(stan.glm.2, save_psis = TRUE)
loo3 <- loo(stan.glm.3, save_psis = TRUE)
loo4 <- loo(stan.glm.4, save_psis = TRUE)
```

```
loo5 <- loo(stan.glm.5, save_psis = TRUE)
rstanarm::loo_compare(loo1, loo2, loo3, loo4, loo5) #interestingly enough loo1 is actually the best.
```

```
##            elpd_diff se_diff
## stan.glm.1   0.0       0.0
## stan.glm.2  -0.2       0.2
## stan.glm.3  -5.2       3.2
## stan.glm.4  -5.3       3.3
## stan.glm.5 -97.8      10.0
```

```
#making sure that bayesian ver of AIC (pareto k) is good
loo1
```

```
##
## Computed from 4000 by 225 log-likelihood matrix
##
##          Estimate   SE
## elpd_loo  -1229.3 19.5
## p_loo         7.4  0.8
## looic      2458.5 39.0
## ------
## Monte Carlo SE of elpd_loo is 0.1.
##
## All Pareto k estimates are good (k < 0.5).
## See help('pareto-k-diagnostic') for details.
```

```
loo2
```

```
##
## Computed from 4000 by 225 log-likelihood matrix
##
##          Estimate   SE
## elpd_loo  -1229.4 19.5
## p_loo         7.5  0.8
## looic      2458.9 39.1
## ------
## Monte Carlo SE of elpd_loo is 0.1.
##
## Pareto k diagnostic values:
##                          Count Pct.    Min. n_eff
## (-Inf, 0.5]   (good)     224   99.6%   1603
##  (0.5, 0.7]   (ok)         1    0.4%   2560
##    (0.7, 1]   (bad)        0    0.0%   <NA>
##    (1, Inf)   (very bad)   0    0.0%   <NA>
##
## All Pareto k estimates are ok (k < 0.7).
## See help('pareto-k-diagnostic') for details.
```

```
loo3
```

```
##
```

```
## Computed from 4000 by 225 log-likelihood matrix
##
##          Estimate   SE
## elpd_loo  -1234.5 19.7
## p_loo         5.9  0.7
## looic      2468.9 39.4
## ------
## Monte Carlo SE of elpd_loo is 0.0.
##
## Pareto k diagnostic values:
##                          Count Pct.    Min. n_eff
## (-Inf, 0.5]   (good)     224   99.6%   1807
##  (0.5, 0.7]   (ok)         1    0.4%   2799
##    (0.7, 1]   (bad)        0    0.0%   <NA>
##    (1, Inf)   (very bad)   0    0.0%   <NA>
##
## All Pareto k estimates are ok (k < 0.7).
## See help('pareto-k-diagnostic') for details.
```

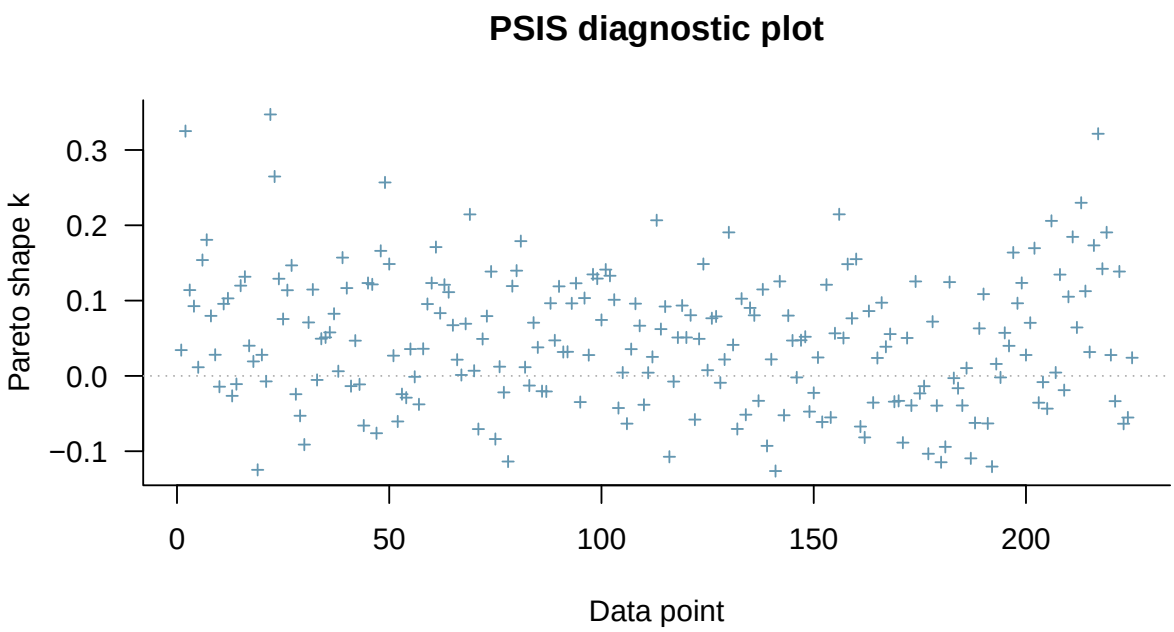
```
loo4
```

```
##
## Computed from 4000 by 225 log-likelihood matrix
##
##          Estimate   SE
## elpd_loo  -1234.5 19.7
## p_loo         5.4  0.6
## looic      2469.1 39.4
## ------
## Monte Carlo SE of elpd_loo is 0.0.
##
## All Pareto k estimates are good (k < 0.5).
## See help('pareto-k-diagnostic') for details.
```

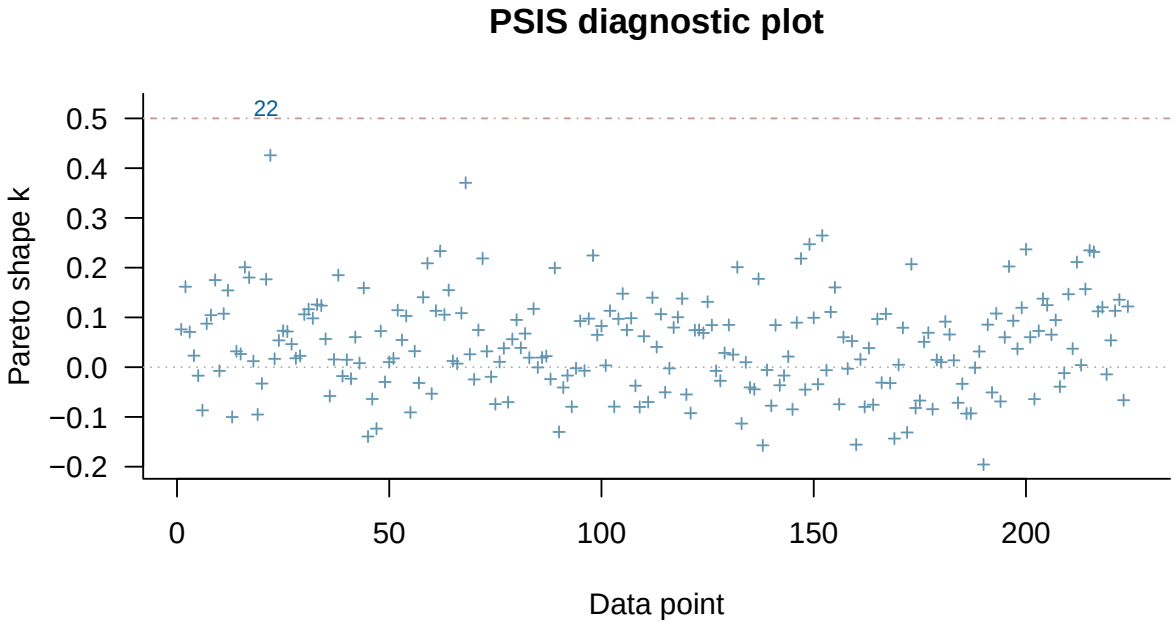
```
loo5
```

```
##
## Computed from 4000 by 225 log-likelihood matrix
##
##          Estimate   SE
## elpd_loo  -1327.0 17.0
## p_loo         3.6  0.5
## looic      2654.1 34.0
## ------
## Monte Carlo SE of elpd_loo is 0.0.
##
## All Pareto k estimates are good (k < 0.5).
## See help('pareto-k-diagnostic') for details.
```

```
#check for outliers (if there are outliers, then post pred will be sensative to 1 observation)
plot(loo1, label_points = TRUE)
```

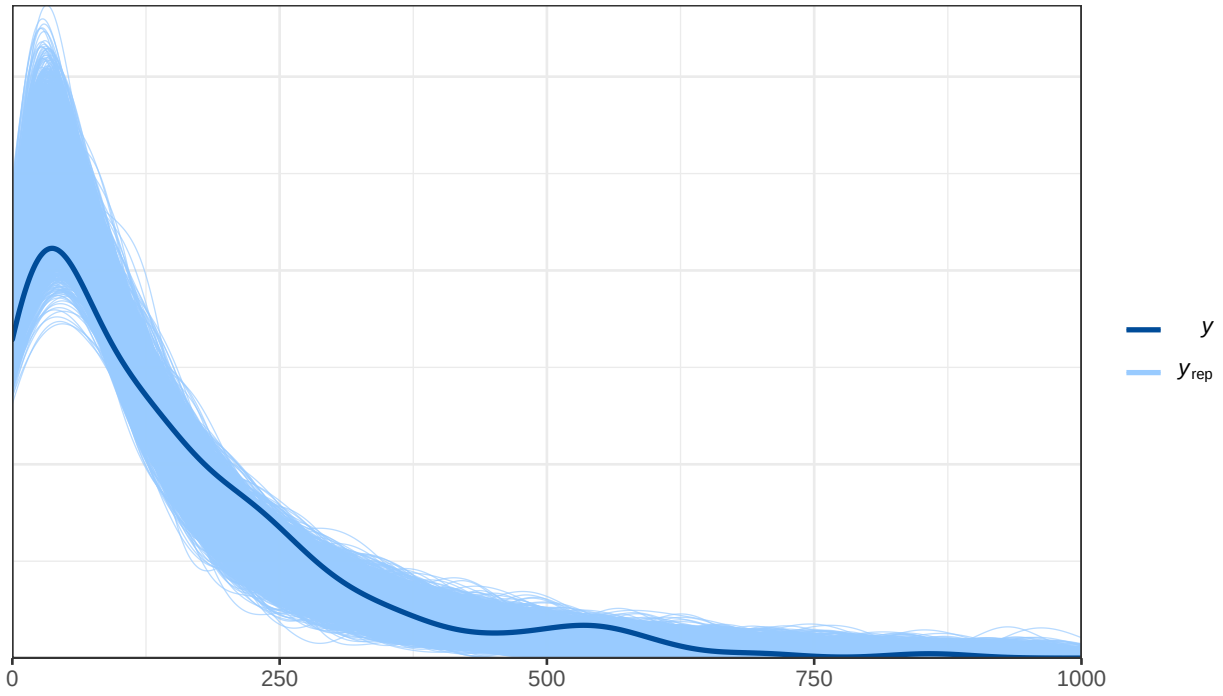



```
plot(loo2, label_points = TRUE)
```

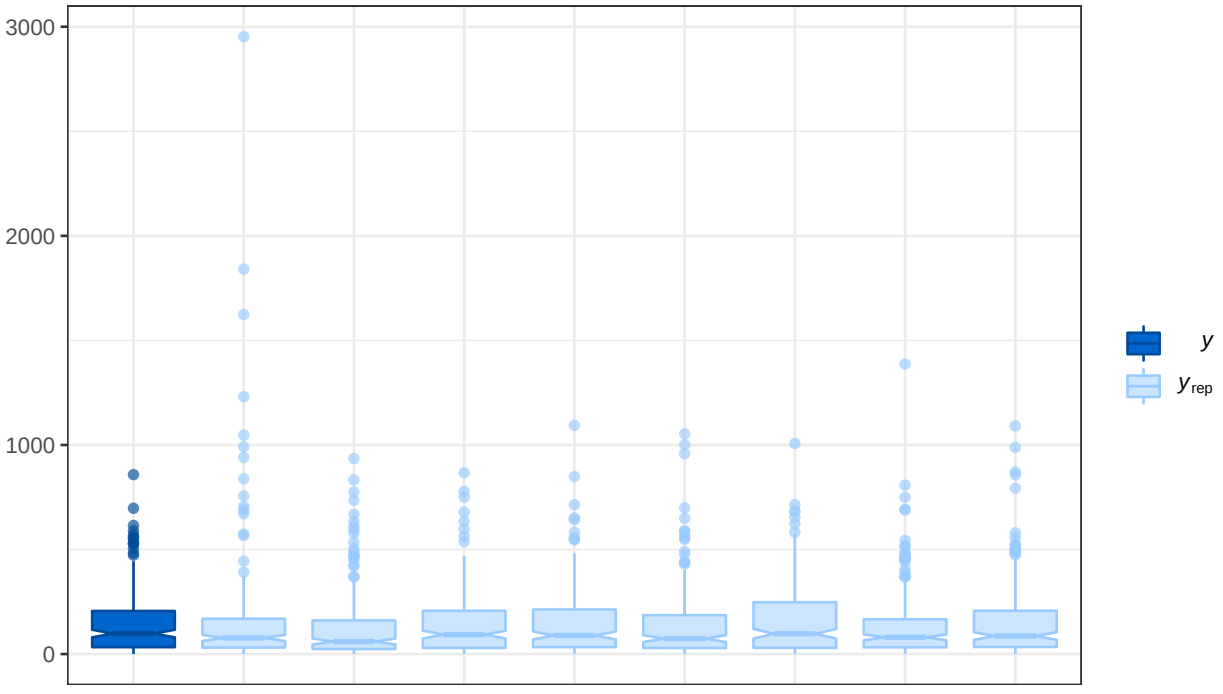



```
#individual model diagnostics - stan.glm.1
y.postpred <- posterior_predict(stan.glm.1)
color_scheme_set("brightblue")
ppc_dens_overlay(stop.1$stops, y.postpred) + xlim(0, 1000) #xlim() truncates so that we focus on the pa
```

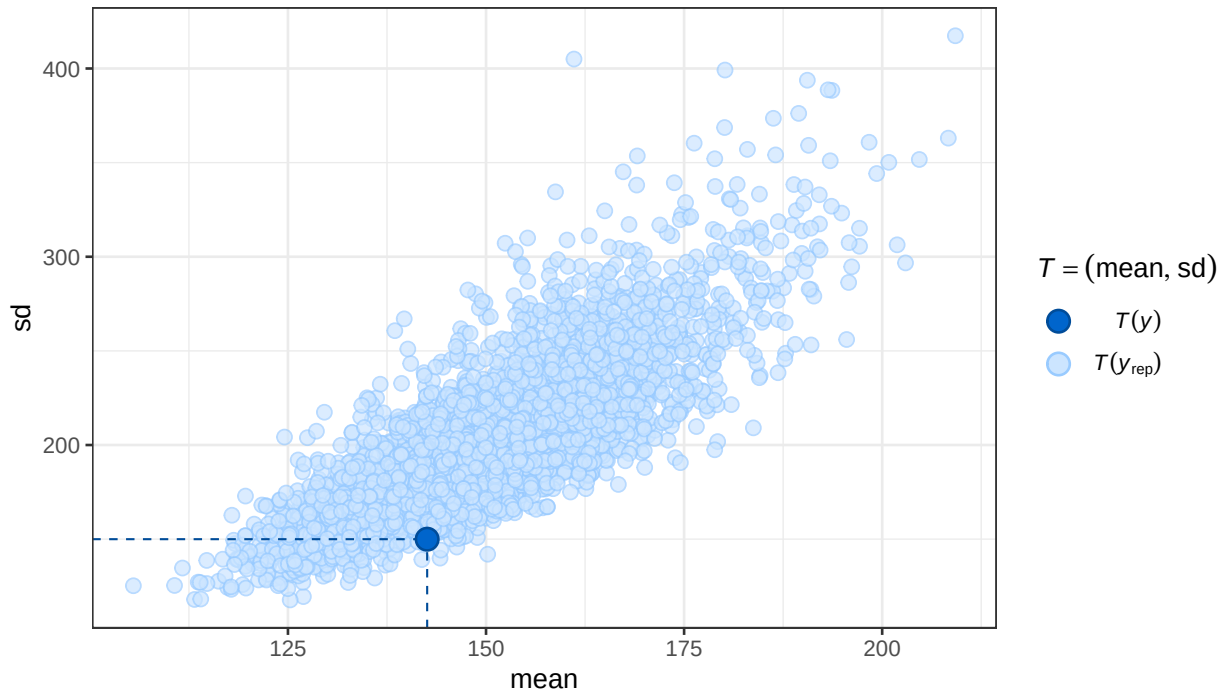
```
## Warning: Removed 8711 rows containing non-finite values (stat_density).
```



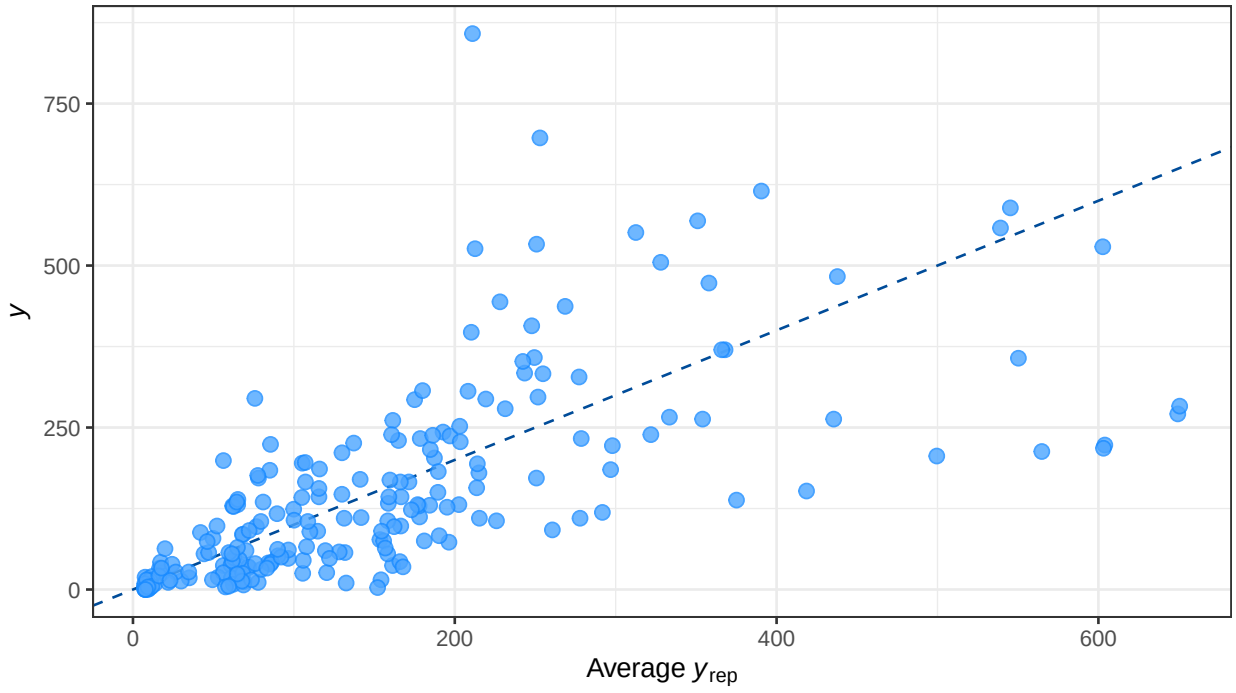

```
pp_check(stan.glm.1, plotfun = "boxplot") #check median/range
```



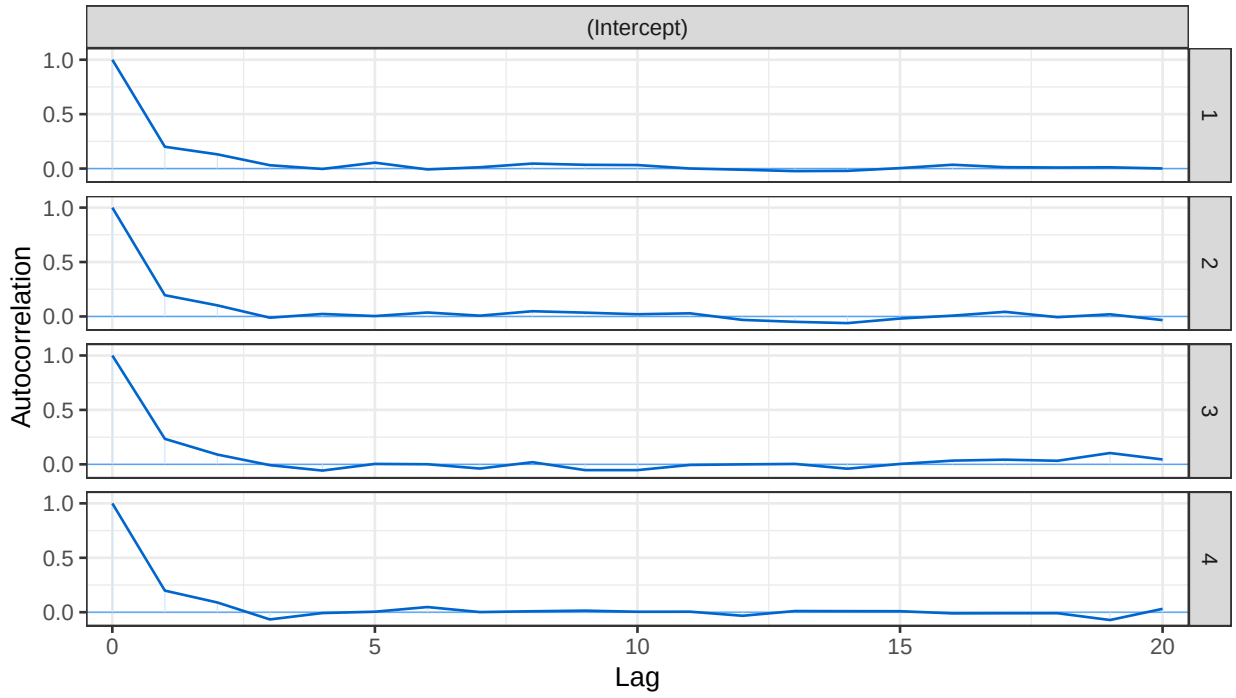

```
pp_check(stan.glm.1, plotfun = "stat_2d", stat = c("mean", "sd"))
```



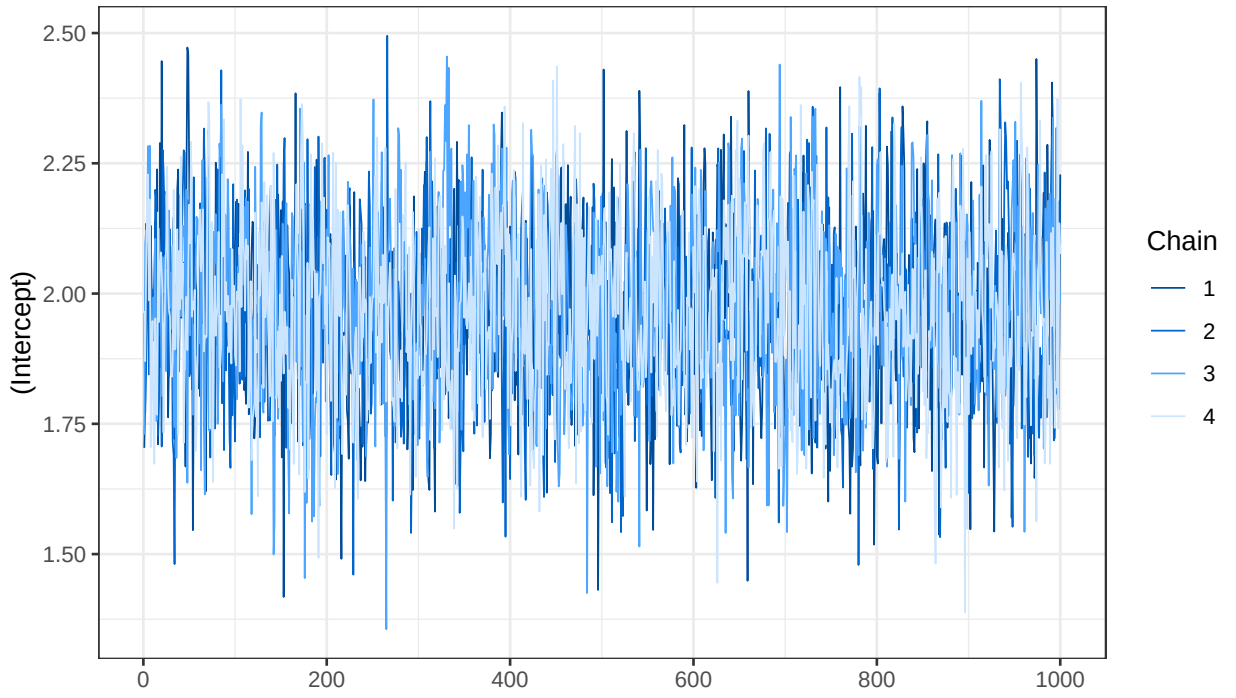

```
    # Scatterplot of two test statistics (capture the mean somewhat but
    # sd is kind of bad. at least it's not high sd)
pp_check(stan.glm.1, plotfun = "scatter_avg") # Scatterplot of y vs. average yrep (our model is good fo
```



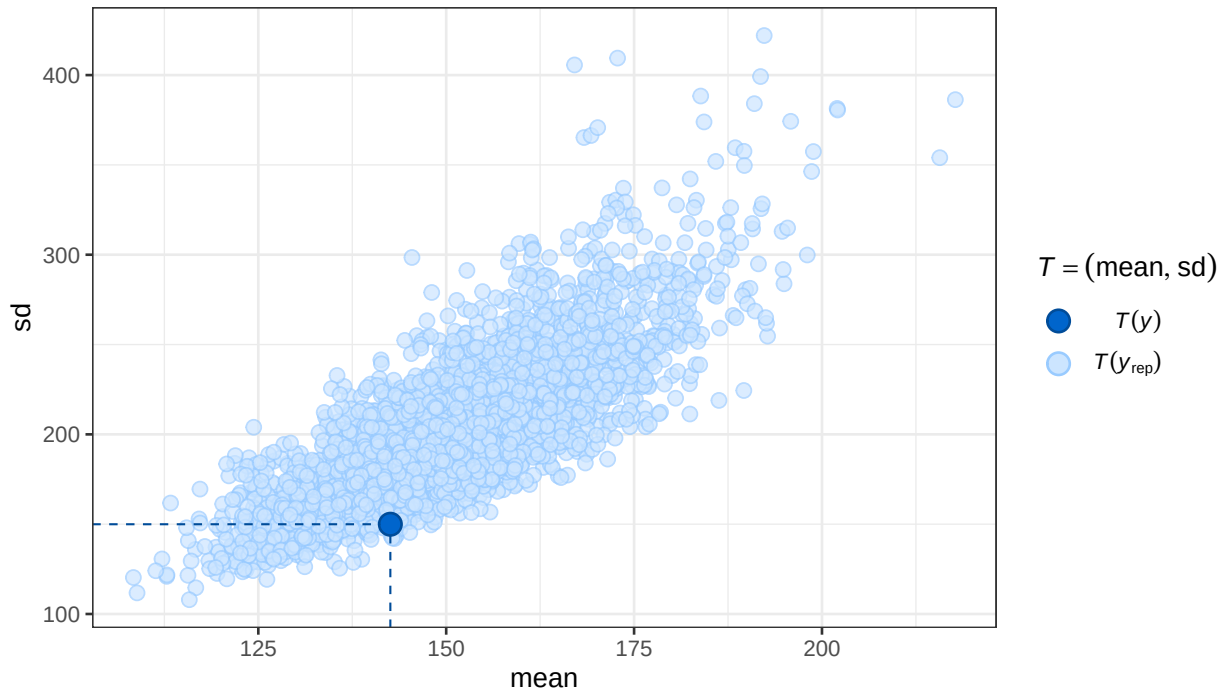

```
plot(stan.glm.1, "acf", pars = "(Intercept)")# autocorrelation by chain
```



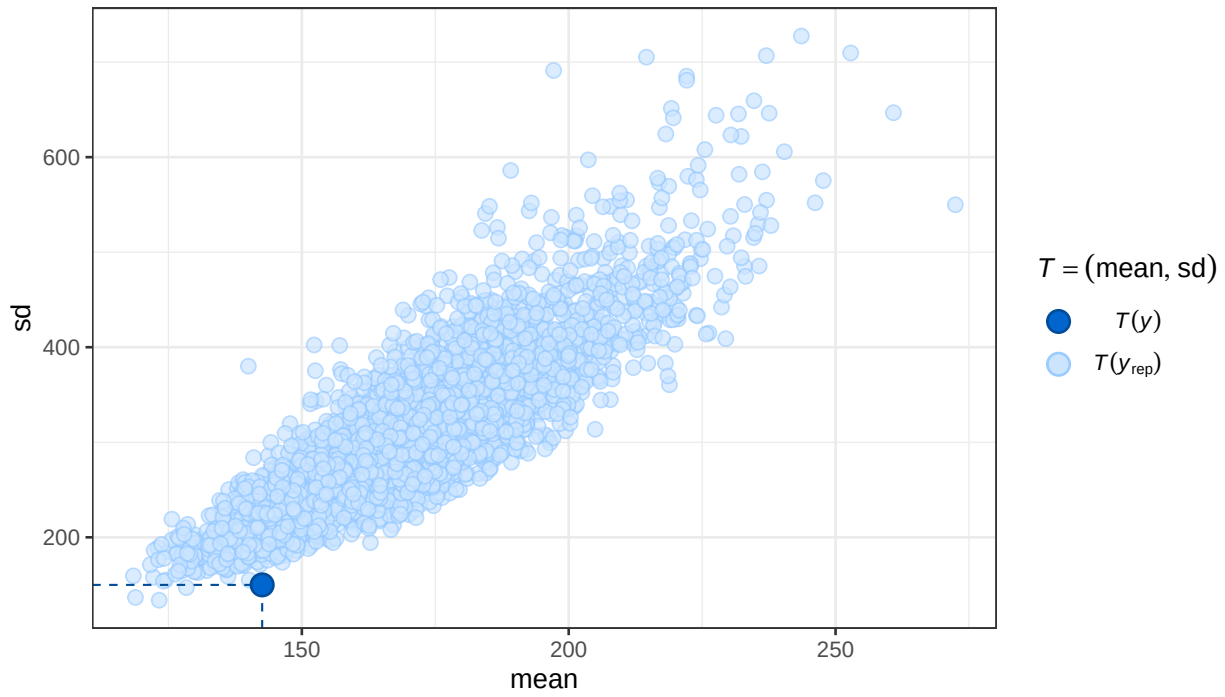

```
plot(stan.glm.1, "trace", pars = "(Intercept)") #traceplot. how to separate by chain?
```



```
pp_check(stan.glm.2, plotfun = "stat_2d", stat = c("mean", "sd"))
```




```
pp_check(stan.glm.3, plotfun = "stat_2d", stat = c("mean", "sd"))
```



```
pp_check(stan.glm.4, plotfun = "stat_2d", stat = c("mean", "sd"))
```