Sim Data Analysis

```r
library(ggplot2)
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```r
library(pscl)
```

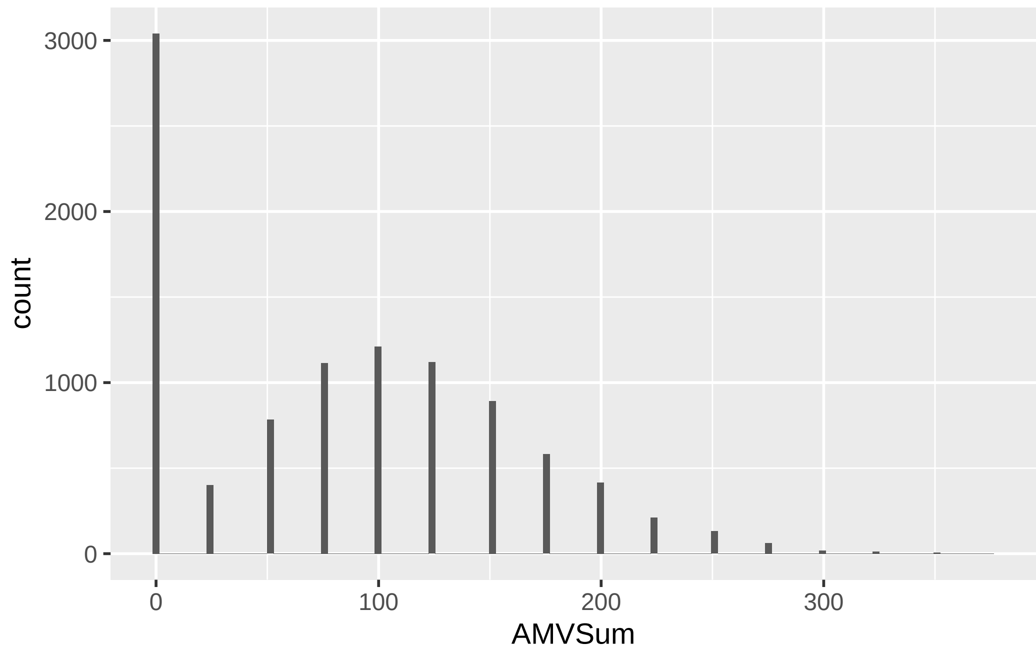
```
Classes and Methods for R developed in the
Political Science Computational Laboratory
Department of Political Science
Stanford University
Simon Jackman
hurdle and zeroinfl functions by Achim Zeileis
```

```r
library(MASS)
```

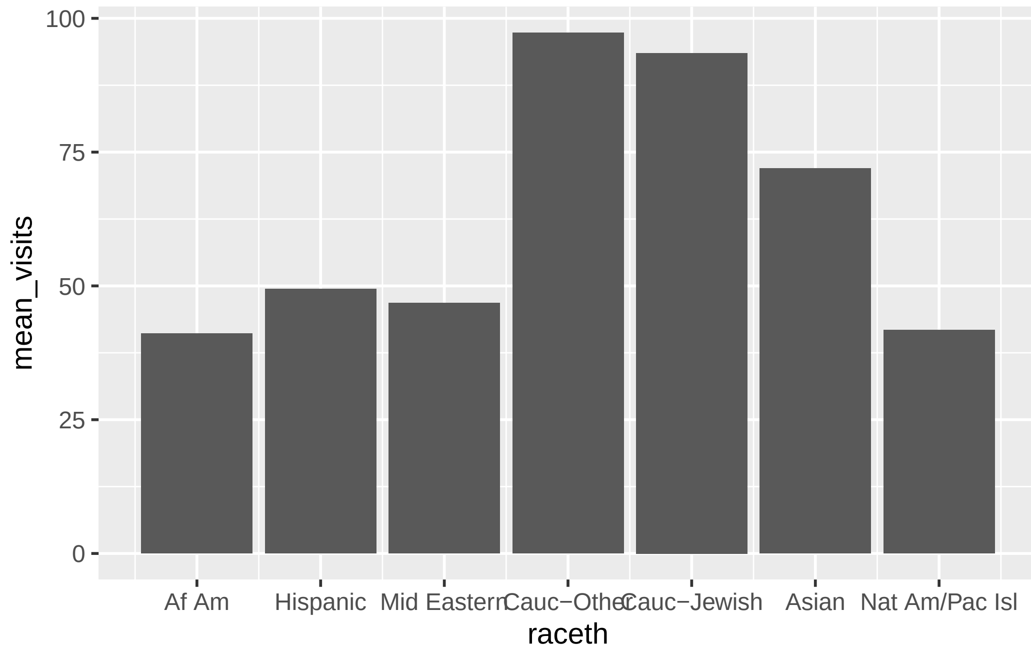
```
Attaching package: 'MASS'
```

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

    select
```

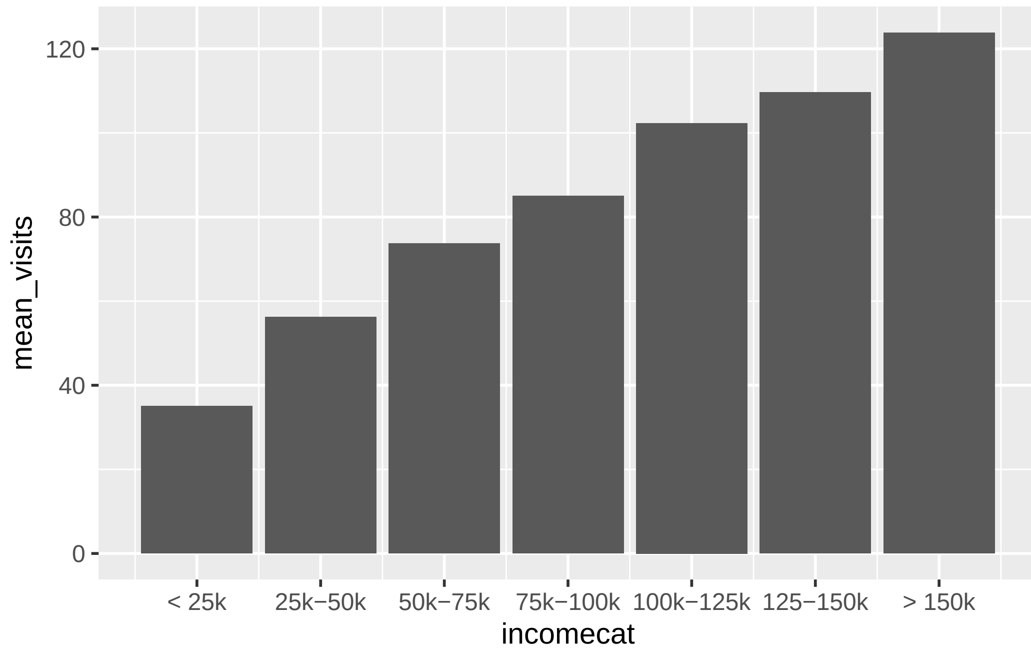
```r
library(performance)

data <- read.csv("sim_data.csv")

# Mental health visits
data |> ggplot(aes(x = AMVSum)) +
  geom_histogram(bins=round(max(data$AMVSum)/3))
```



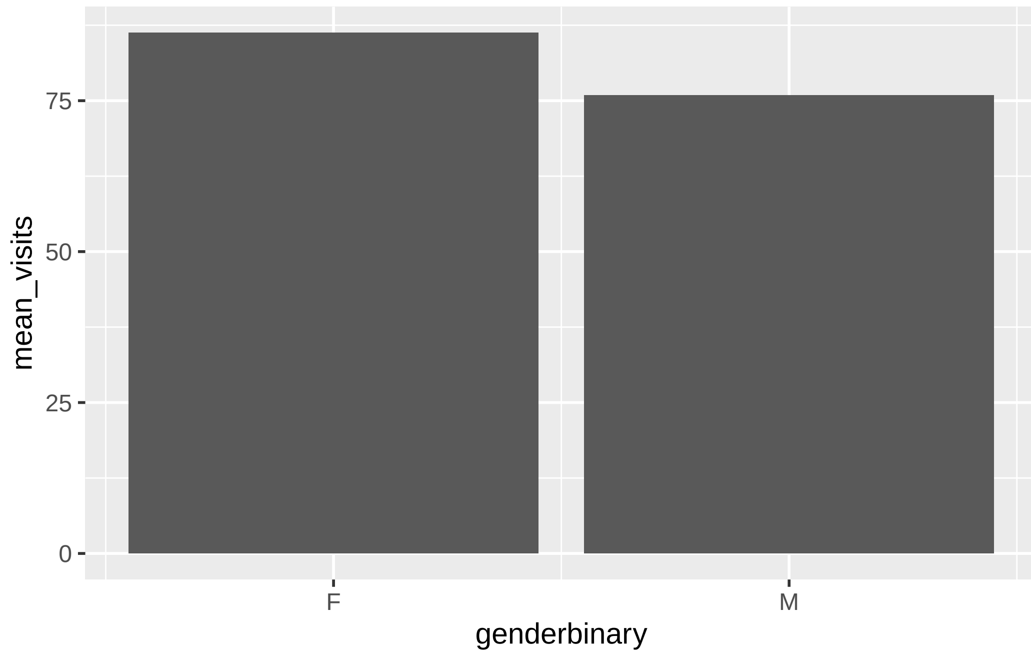

```r
# # By race
data |> group_by(raceth) |>
  summarize(mean_visits = mean(AMVSum)) |>
  ggplot(aes(x = raceth, y = mean_visits)) +
  geom_bar(stat = "identity") +
  scale_x_continuous(breaks = c(1,2,3,4,5,6,7), labels = c("Af Am", "Hispanic", "Mid Easte
```



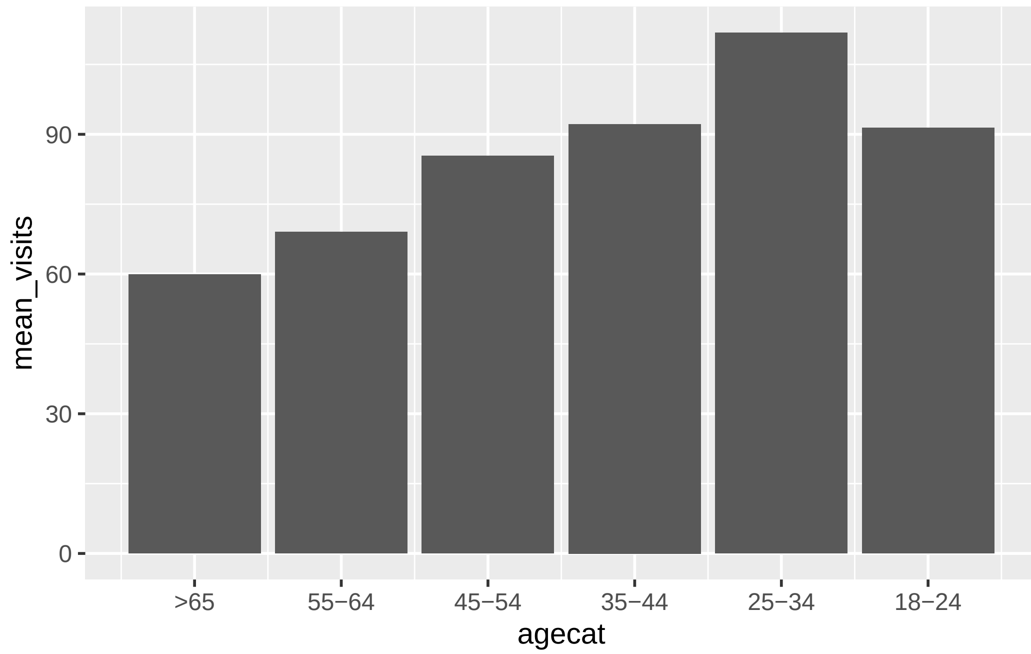

```
# By income
data |> group_by(incomecat) |>
  summarize(mean_visits = mean(AMVSum)) |>
  ggplot(aes(x = incomecat, y = mean_visits)) +
  geom_bar(stat = "identity") +
  scale_x_continuous(breaks = c(1,2,3,4,5,6,7), labels = c("< 25k", "25k-50k", "50k-75k",
```




```
# By gender
data |> group_by(genderbinary) |>
  summarize(mean_visits = mean(AMVSum)) |>
  ggplot(aes(x = genderbinary, y = mean_visits)) +
  geom_bar(stat = "identity") +
  scale_x_continuous(breaks = c(0,1), labels = c("F", "M"))
```




```
# By age
data |> group_by(agecat) |>
  summarize(mean_visits = mean(AMVSum)) |>
  ggplot(aes(x = agecat, y = mean_visits)) +
  geom_bar(stat = "identity") +
  scale_x_continuous(breaks = c(1,2,3,4,5,6), labels = c(">65", "55-64", "45-54", "35-44",
```




```r
# Model example

model1 <- glm(AMVSum ~ factor(genderbinary), family= poisson(link = "log"), data=data)
summary(model1)
```

```

Call:
glm(formula = AMVSum ~ factor(genderbinary), family = poisson(link = "log"),
    data = data)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           4.457162   0.001380 3230.99   <2e-16 ***
factor(genderbinary)1 -0.128222   0.002297  -55.82   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 792377  on 9999  degrees of freedom
Residual deviance: 789226  on 9998  degrees of freedom
AIC: 834285
```

```
Number of Fisher Scoring iterations: 5
```

```r
performance::check_overdispersion(model1)
```

```
# Overdispersion test

       dispersion ratio =     63.701
  Pearson's Chi-Squared = 636885.850
                p-value =    < 0.001

Overdispersion detected.
```

```r
performance::check_zeroinflation(model1)
```

```
# Check for zero-inflation

   Observed zeros: 3041
  Predicted zeros: 0
            Ratio: 0.00

Model is underfitting zeros (probable zero-inflation).
```

```r
exp(model1$coefficients)
```

```
          (Intercept) factor(genderbinary)1
           86.2424093             0.8796581
```

We have good reason to believe there will be a lot of zeros in the data, so we will use a zero-inflated model.

We also have a good reason to believe there is overdispersion, so we will use a negative binomial model.

```r
model1_nb <- zeroinfl(AMVSum ~ factor(genderbinary) | factor(genderbinary), data=data, dis

summary(model1_nb)
```

```
Call:
zeroinfl(formula = AMVSum ~ factor(genderbinary) | factor(genderbinary),
    data = data, dist = "negbin")

Pearson residuals:
     Min       1Q   Median       3Q      Max
-1.14829 -1.05663 -0.01203  0.68437  3.84471

Count model coefficients (negbin with log link):
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)            4.791142   0.007707  621.68  < 2e-16 ***
factor(genderbinary)1 -0.053405   0.012625   -4.23 2.34e-05 ***
Log(theta)             1.382980   0.016997   81.36  < 2e-16 ***

Zero-inflation model coefficients (binomial with logit link):
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           -0.92504    0.02841  -32.56  < 2e-16 ***
factor(genderbinary)1  0.24186    0.04422    5.47 4.51e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 3.9868
Number of iterations in BFGS optimization: 1
Log-likelihood: -4.389e+04 on 5 Df
```