ANALYZING SURVEY DATA IN R

Visualizing a categorical variable

Kelly McConville
Assistant Professor of Statistics

NHANES: visualizing race

```
library(dplyr)
tab_unw <- NHANESraw %>%
   group_by(Race1) %>%
   summarize(Freq = n()) %>%
   mutate(Prop = Freq/sum(Freq)) %>%
   arrange(desc(Prop))
```

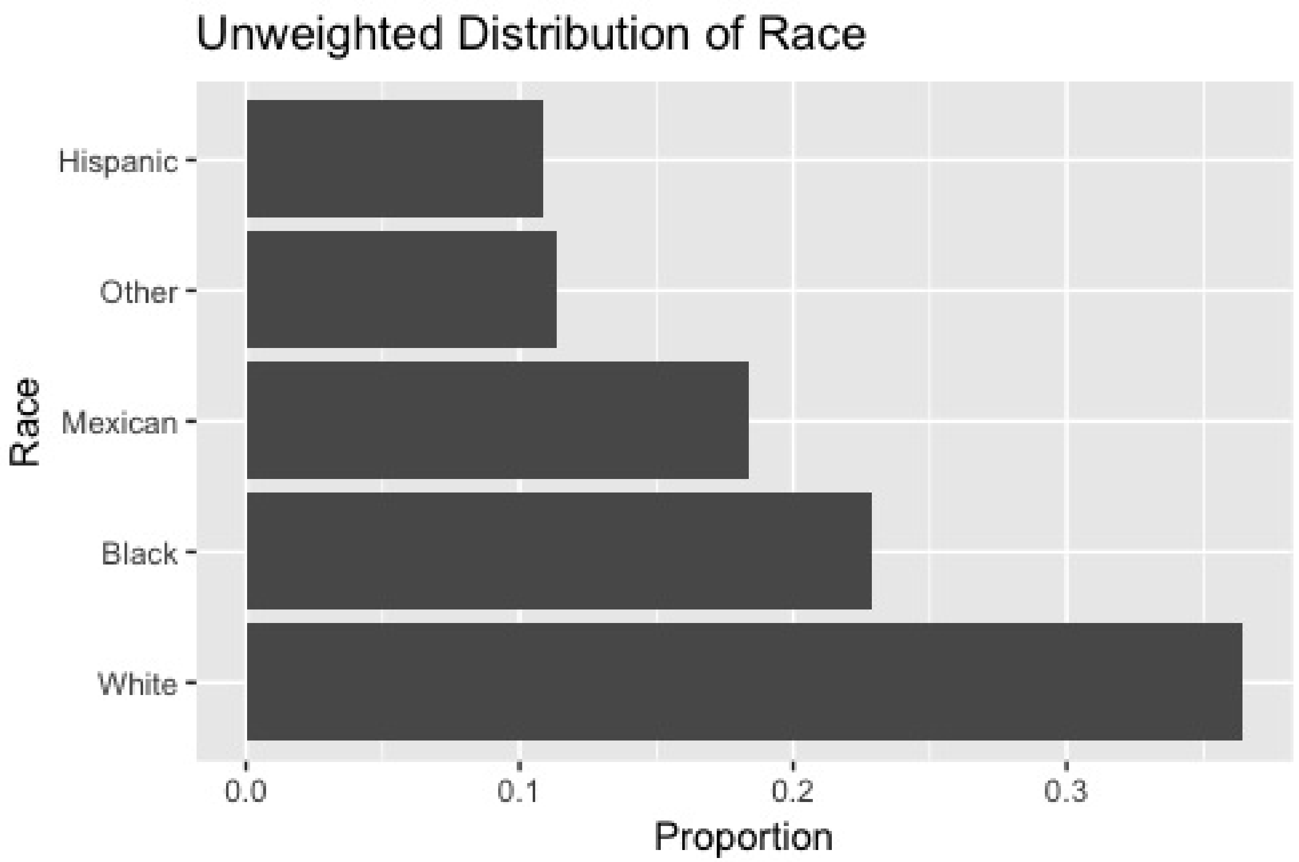
```
tab_unw
```

```
# A tibble: 5 x 3
     Race1  Freq      Prop
    <fctr> <int>     <dbl>
 1    White  7393 0.3643128
 2    Black  4640 0.2286503
 3  Mexican  3739 0.1842507
 4    Other  2312 0.1139309
 5 Hispanic  2209 0.1088553
```

NHANES documentation

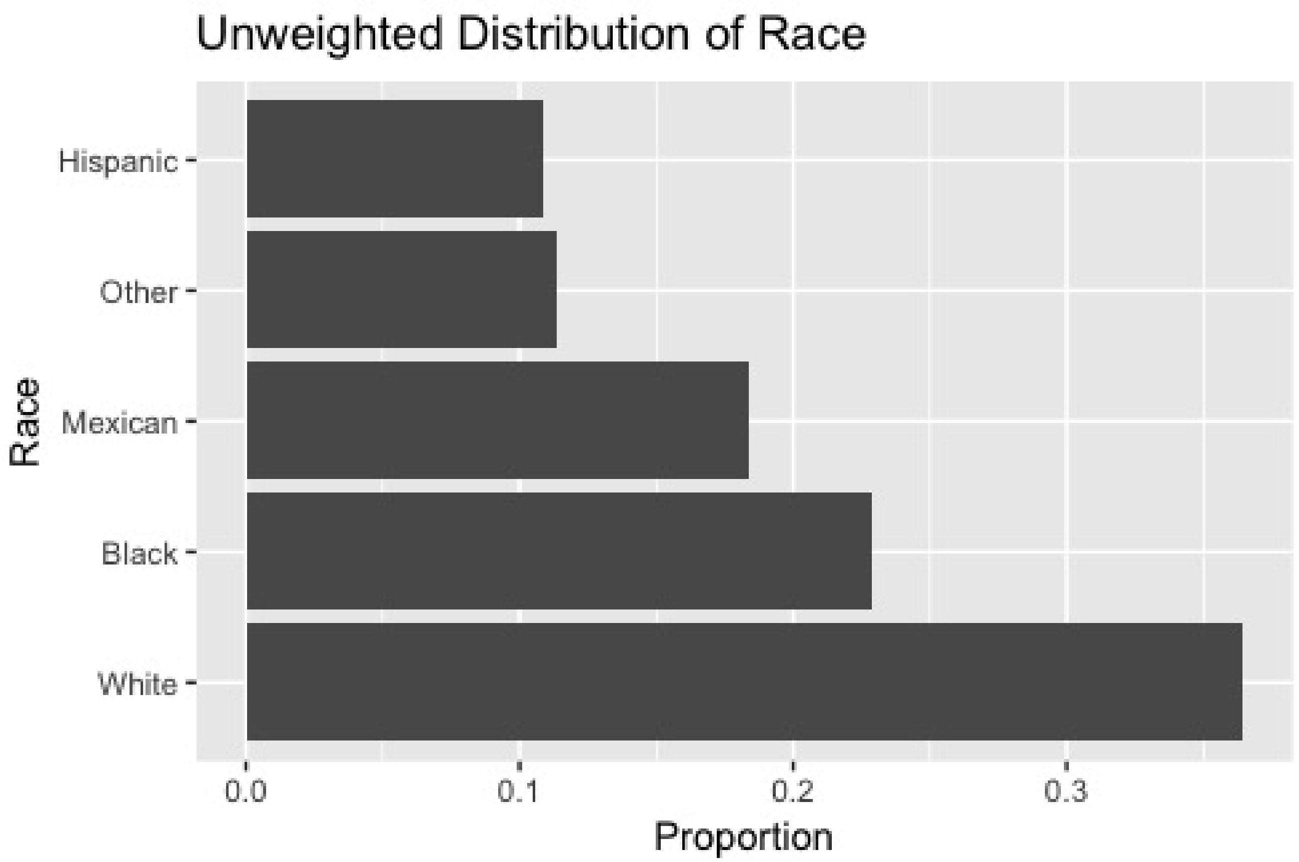
NHANES: visualizing race

```
library(ggplot2)
ggplot(data = tab_unw, mapping = aes(x = Race1, y = Prop)) +
  geom_col() +
  coord_flip() +
  scale_x_discrete(limits = tab_unw$Race1) # Labels layer omitted
```



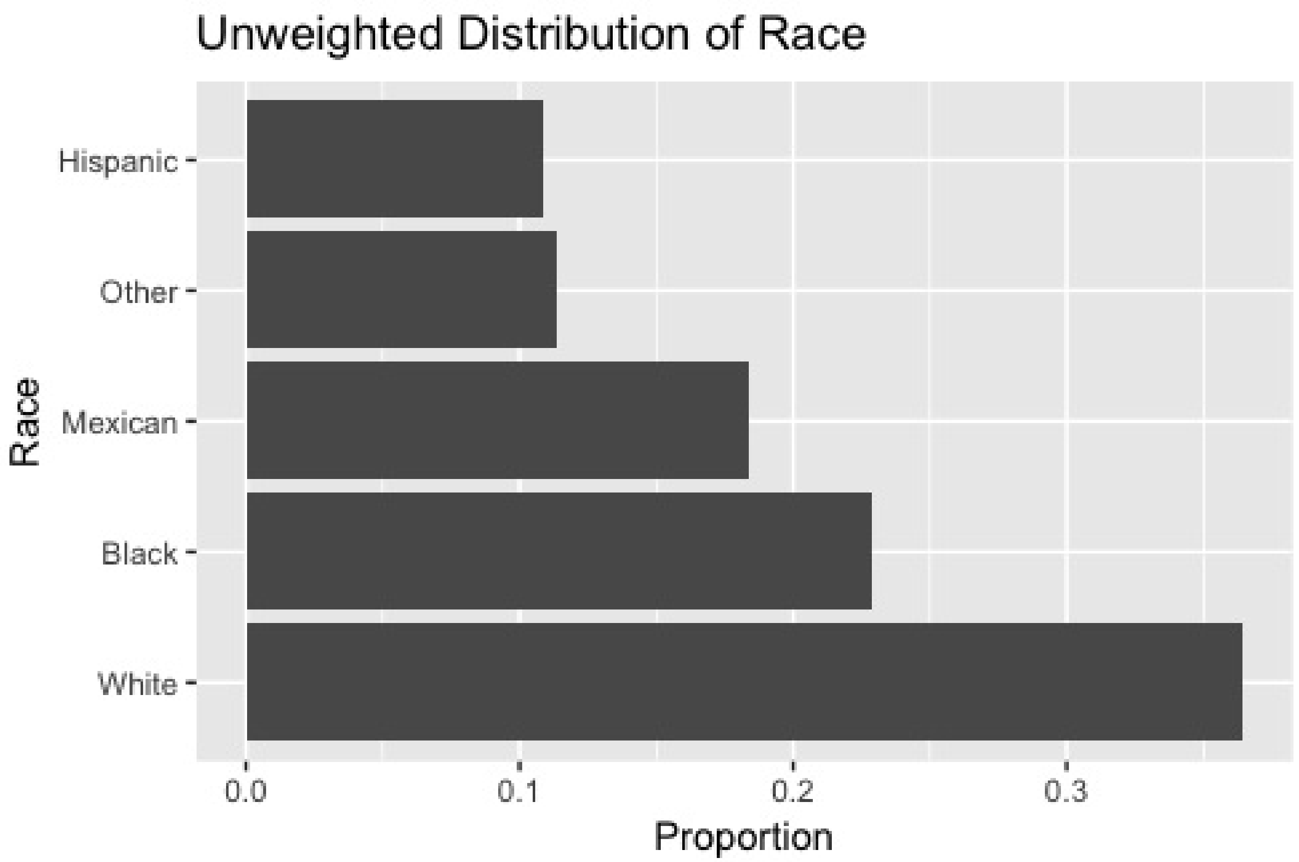
NHANES: visualizing race

```
library(ggplot2)
ggplot(data = tab_unw, mapping = aes(x = Race1, y = Prop)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  scale_x_discrete(limits = tab_unw$Race1) # Labels layer omitted
```



NHANES: visualizing race

```r
library(ggplot2)
ggplot(data = tab_unw, mapping = aes(x = Race1, y = Prop)) +
  geom_col() +
  coord_flip() +
  scale_x_discrete(limits = tab_unw$Race1) # Labels layer omitted
```



NHANES: visualizing race

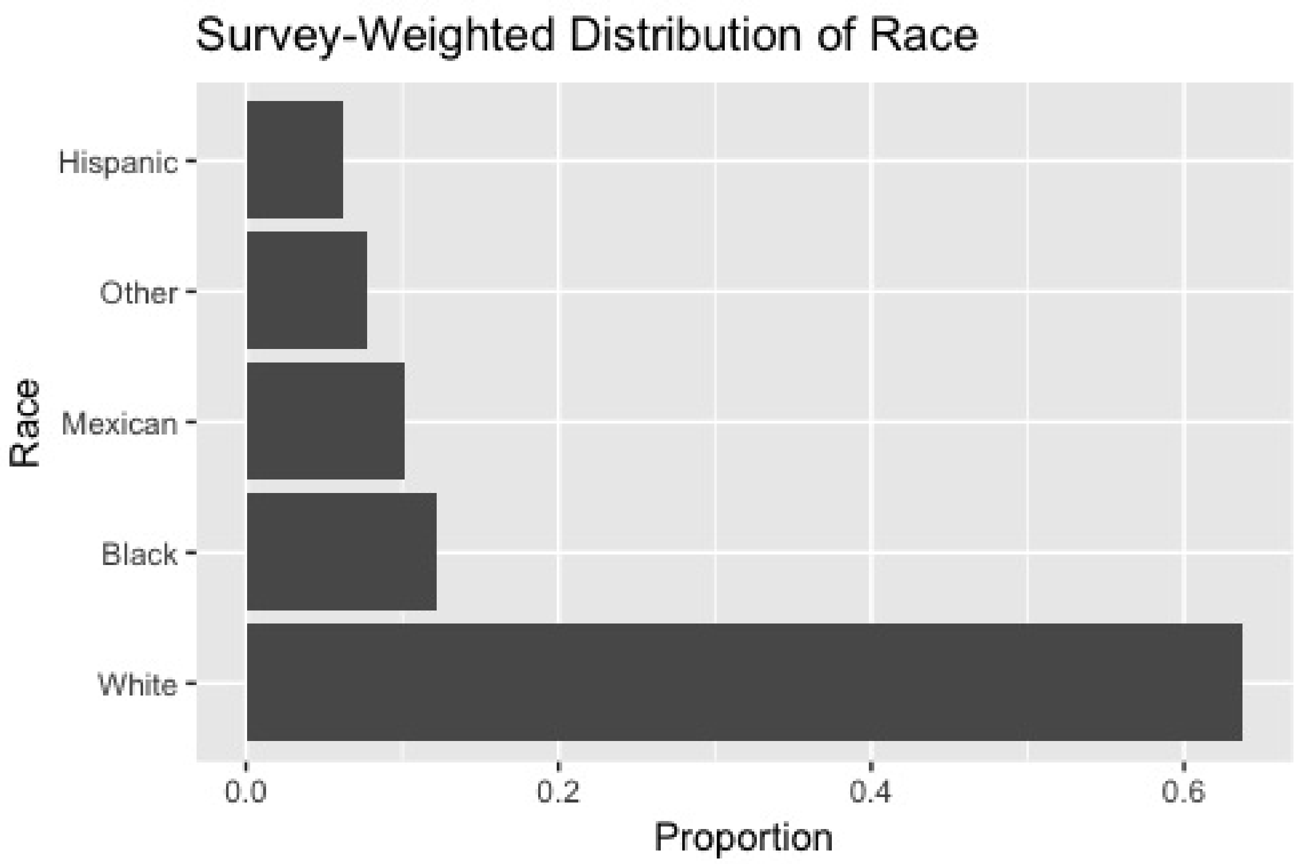
```
tab_w <- svytable(~Race1, design = NHANES_design) %>%
   as.data.frame() %>%
   mutate(Prop = Freq/sum(Freq)) %>%
   arrange(desc(Prop))
```

```
tab_w
```

```
     Race1      Freq       Prop
1    White 193966274 0.63748664
2    Black  37241616 0.12239773
3  Mexican  30719158 0.10096112
4    Other  23389002 0.07686994
5 Hispanic  18951150 0.06228456
```

NHANES: visualizing race

```
ggplot(data = tab_w, mapping = aes(x = Race1, y = Prop)) +
  geom_col() +
  coord_flip() +
  scale_x_discrete(limits = tab_w$Race1) # Labels layer omitted
```



ANALYZING SURVEY DATA IN R

Let's practice!

ANALYZING SURVEY DATA IN R

Exploring two categorical variables

Kelly McConville
Assistant Professor of Statistics

NHANES: race and diabetes

```
svytable(~Diabetes,
         design = NHANES_design)
```

```
Diabetes
       No       Yes
275814034  24335536
```

```
tab_w <- svytable(~Race1 + Diabetes,
                  design = NHANES_design)
tab_w
```

```
          Diabetes
Race1             No       Yes
  Black     32697528   4003497
  Hispanic  17258245   1370393
  Mexican   27886500   2081657
  White    177088354  14708094
  Other     20883407   2171895
```

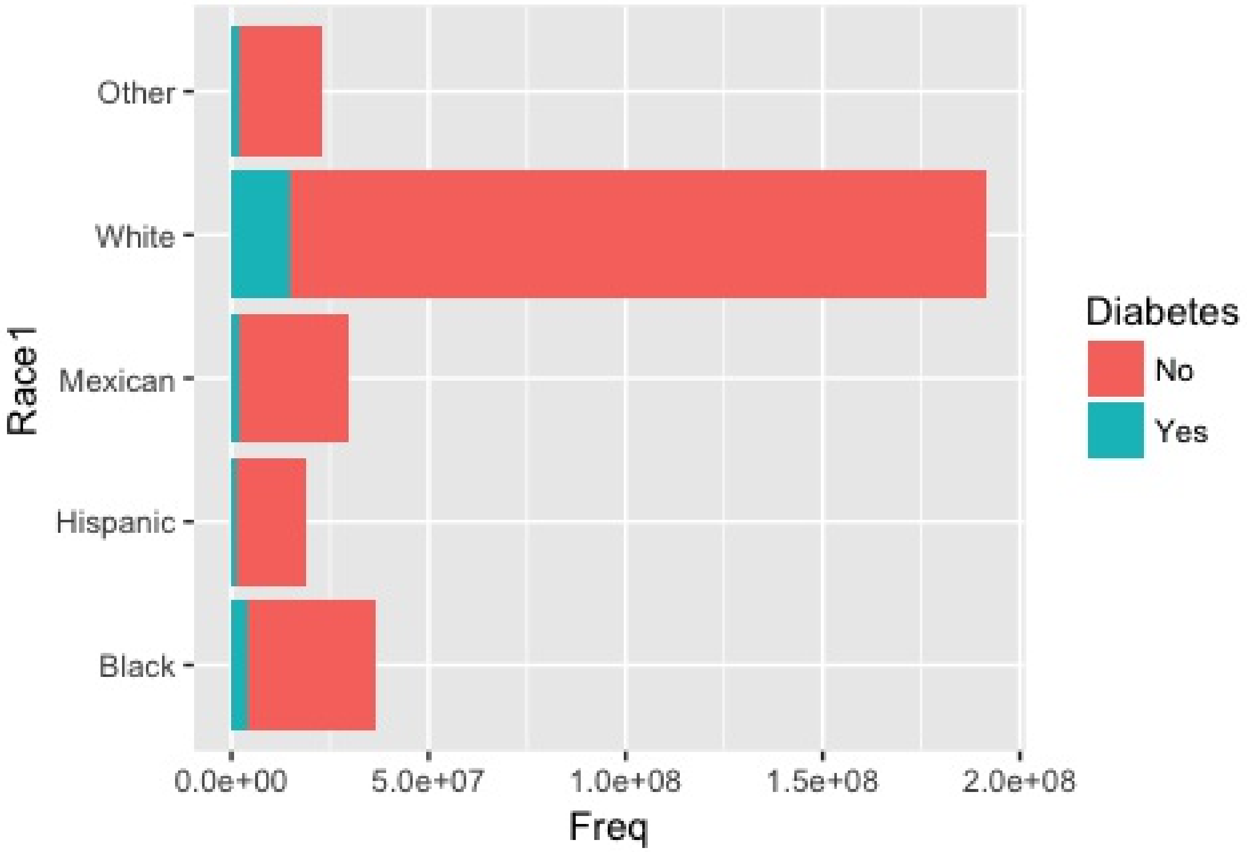
NHANES: race and diabetes

```
tab_w <- as.data.frame(tab_w)
tab_w
```

```
      Race1 Diabetes      Freq
1     Black       No  32697528
2  Hispanic       No  17258245
3   Mexican       No  27886500
4     White       No 177088354
5     Other       No  20883407
6     Black      Yes   4003497
7  Hispanic      Yes   1370393
8   Mexican      Yes   2081657
9     White      Yes  14708094
10    Other      Yes   2171895
```

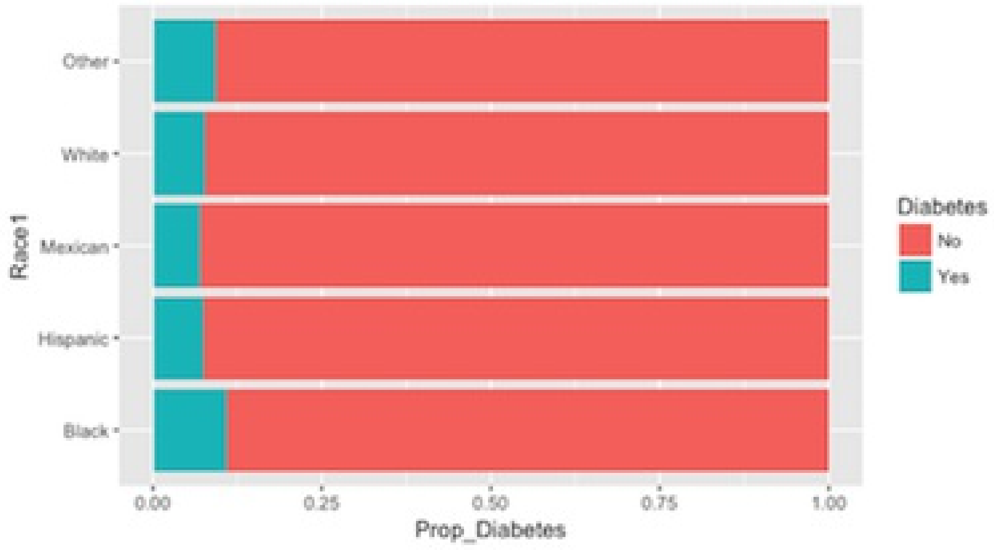
```
ggplot(data = tab_w,
       mapping = aes(x = Race1,
                     fill = Diabetes,
                     y = Freq)) +
  geom_col() + coord_flip()
```

NHANES: race and diabetes



NHANES: race and diabetes

```
ggplot(data = tab_w,
       mapping = aes(x = Race1,
                     y = Freq,
                     fill = Diabetes)) +
  geom_col(position = "fill") +
  coord_flip()
```



ANALYZING SURVEY DATA IN R

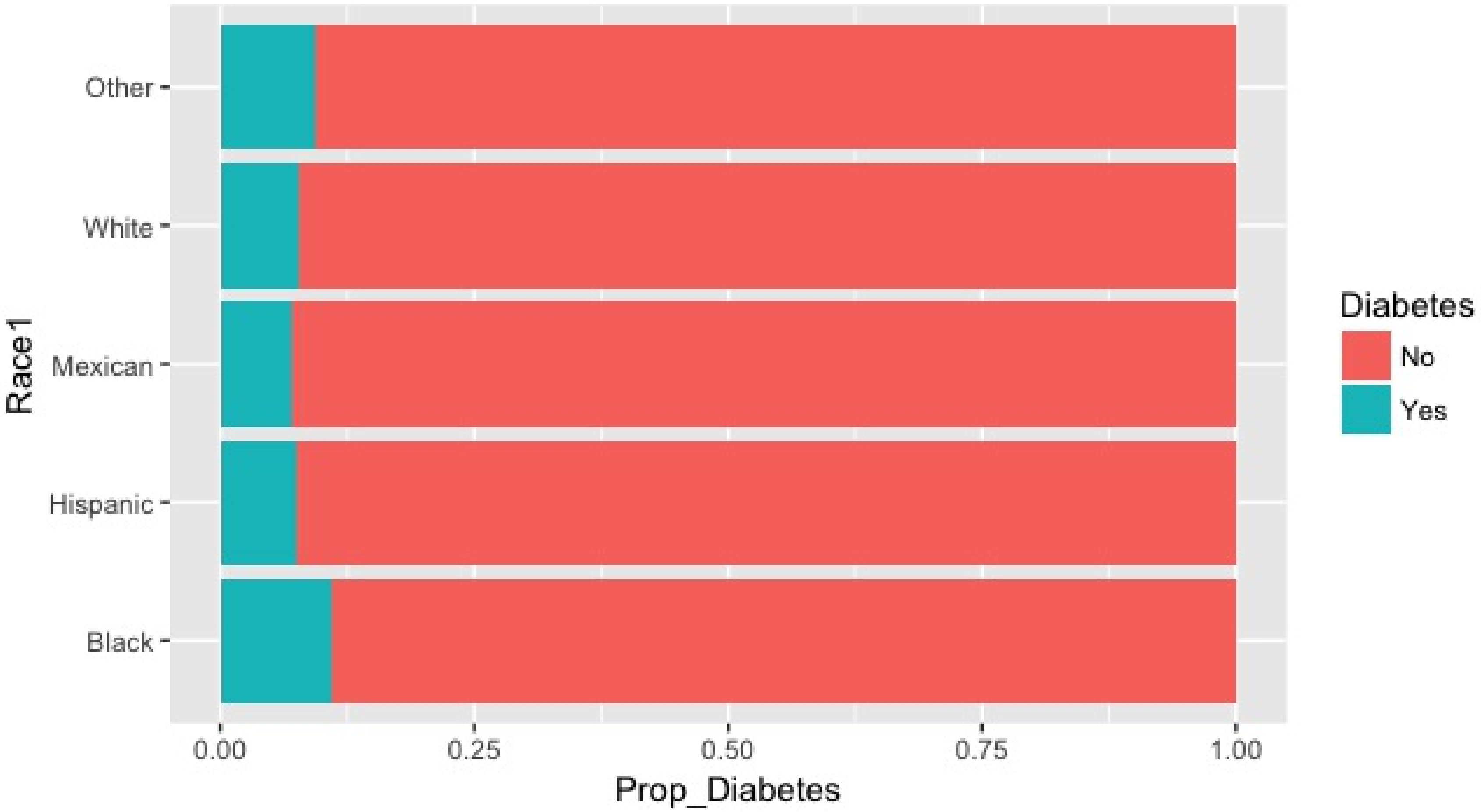
Let's practice!

ANALYZING SURVEY DATA IN R

Inference for categorical variables

Kelly McConville
Assistant Professor of Statistics

NHANES: Race and Diabetes



Inference: Chi-square Test

Null Hypothesis: Prevalence of diabetes is not associated with race.

Alternative Hypothesis: Prevalence of diabetes is associated with race.

```
svychisq(~Race1 + Diabetes, design = NHANES_design, statistic = "Chisq")
```

```


    Pearsons X^2: Rao & Scott adjustment

data:  svychisq(~Race1 + Diabetes, design = NHANES_design,
statistic = "Chisq")
X-squared = 37.708, df = 4, p-value = 0.0001177
```

ANALYZING SURVEY DATA IN R

Let's practice!