STAT 650 Project

Colin Busby

2022-10-07

Research Question/Data Analysis Goals

https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

The Heart Disease data set came from Kaggle, while the actual data comes from the 2020 annual CDC survey data of 400k adults related to their health status. The dimensions of the data set are 319,795 observations of 18 variables. The primary variable of interest is `HeartDisease`, which indicates by `Yes/No` whether the subject had heart disease. The remaining variables are various factors such as Race, whether or not they smoke or drink alcohol, etc.

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

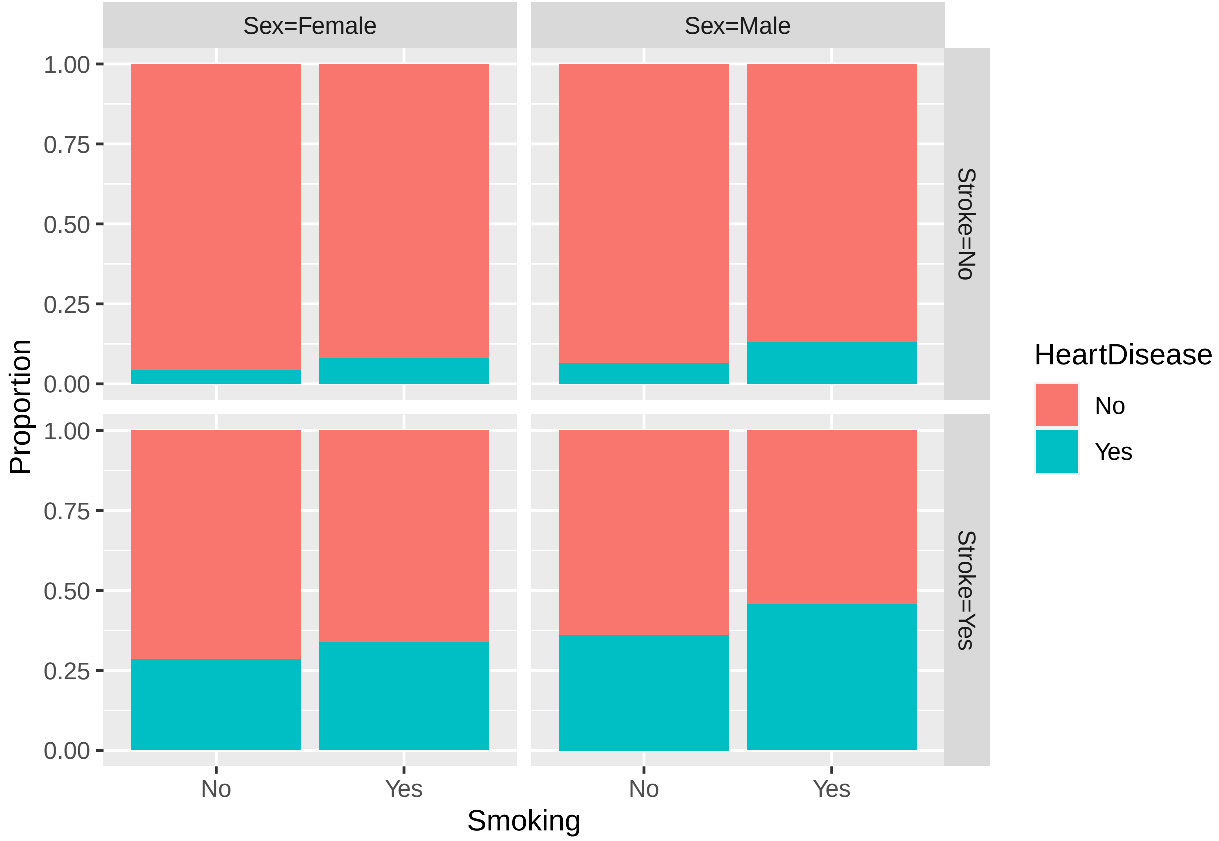
```r
heart.clean <- read_csv("heart_2020_cleaned.csv")
```

```
## Rows: 319795 Columns: 18
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (14): HeartDisease, Smoking, AlcoholDrinking, Stroke, DiffWalking, Sex, ...
## dbl  (4): BMI, PhysicalHealth, MentalHealth, SleepTime
##
## 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.
```

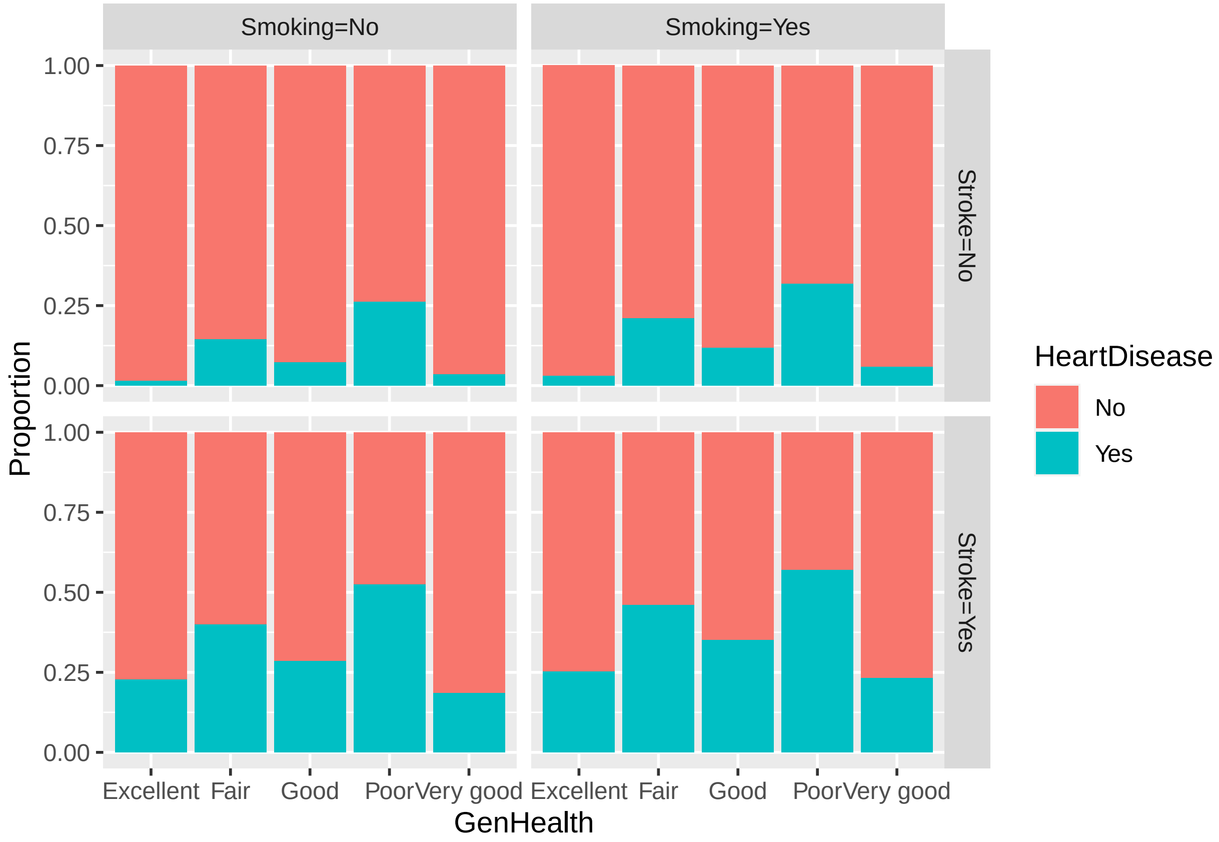
```r
# Note: slice_sample is leftover from playing around with sampling, 319795 is entirety of data set.
set.seed(867)
heart.sample <- slice_sample(heart.clean, n = 319795)
glimpse(heart.sample)
```

```
## Rows: 319,795
## Columns: 18
## $ HeartDisease     <chr> "No", "Yes", "No", "No", "No", "No", "No", "No", "No"~
## $ BMI              <dbl> 22.59, 24.21, 31.58, 29.75, 23.91, 19.65, 21.93, 21.7~
## $ Smoking          <chr> "No", "Yes", "No", "No", "Yes", "Yes", "No", "No", "N~
## $ AlcoholDrinking  <chr> "No", "No", "No", "No", "No", "No", "Yes", "No", "No"~
## $ Stroke           <chr> "No", "No", "No", "No", "No", "No", "No", "No", "No",~
## $ PhysicalHealth   <dbl> 0, 0, 10, 0, 15, 30, 0, 0, 30, 4, 0, 0, 0, 0, 0, 0, 0~
## $ MentalHealth     <dbl> 0, 0, 3, 0, 15, 0, 30, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, ~
## $ DiffWalking      <chr> "No", "No", "Yes", "No", "No", "Yes", "No", "No", "Ye~
## $ Sex              <chr> "Female", "Male", "Female", "Male", "Female", "Female~
## $ AgeCategory      <chr> "70-74", "75-79", "80 or older", "60-64", "35-39", "6~
## $ Race             <chr> "White", "White", "White", "White", "White", "Hispani~
## $ Diabetic         <chr> "No", "No", "Yes", "Yes", "No", "Yes", "No", "No", "N~
## $ PhysicalActivity <chr> "Yes", "Yes", "No", "Yes", "Yes", "No", "Yes", "Yes",~
## $ GenHealth        <chr> "Excellent", "Excellent", "Fair", "Good", "Good", "Fa~
## $ SleepTime        <dbl> 8, 7, 6, 7, 6, 6, 8, 8, 8, 5, 6, 8, 8, 7, 6, 8, 6, 6,~
## $ Asthma           <chr> "No", "No", "No", "No", "No", "Yes", "No", "No", "No"~
## $ KidneyDisease    <chr> "No", "No", "No", "No", "No", "Yes", "No", "No", "No"~
## $ SkinCancer       <chr> "No", "No", "Yes", "No", "No", "No", "No", "No", "No"~
```

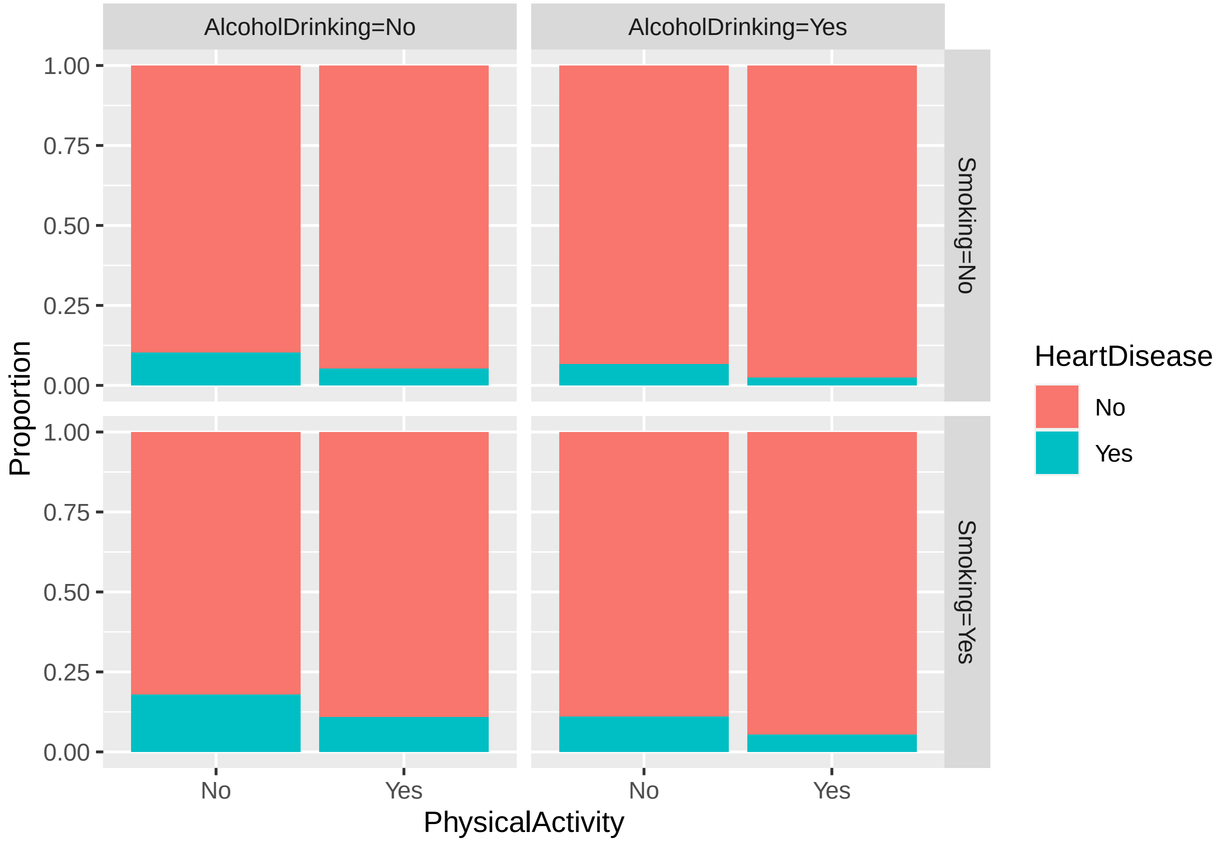
```
ggplot(heart.sample) +
  geom_bar(aes(x=Smoking, fill = HeartDisease), position = "fill")  +
  facet_grid(Stroke ~ Sex,labeller = purrr::partial(label_both,sep = "=")) +
  ylab("Proportion")
```



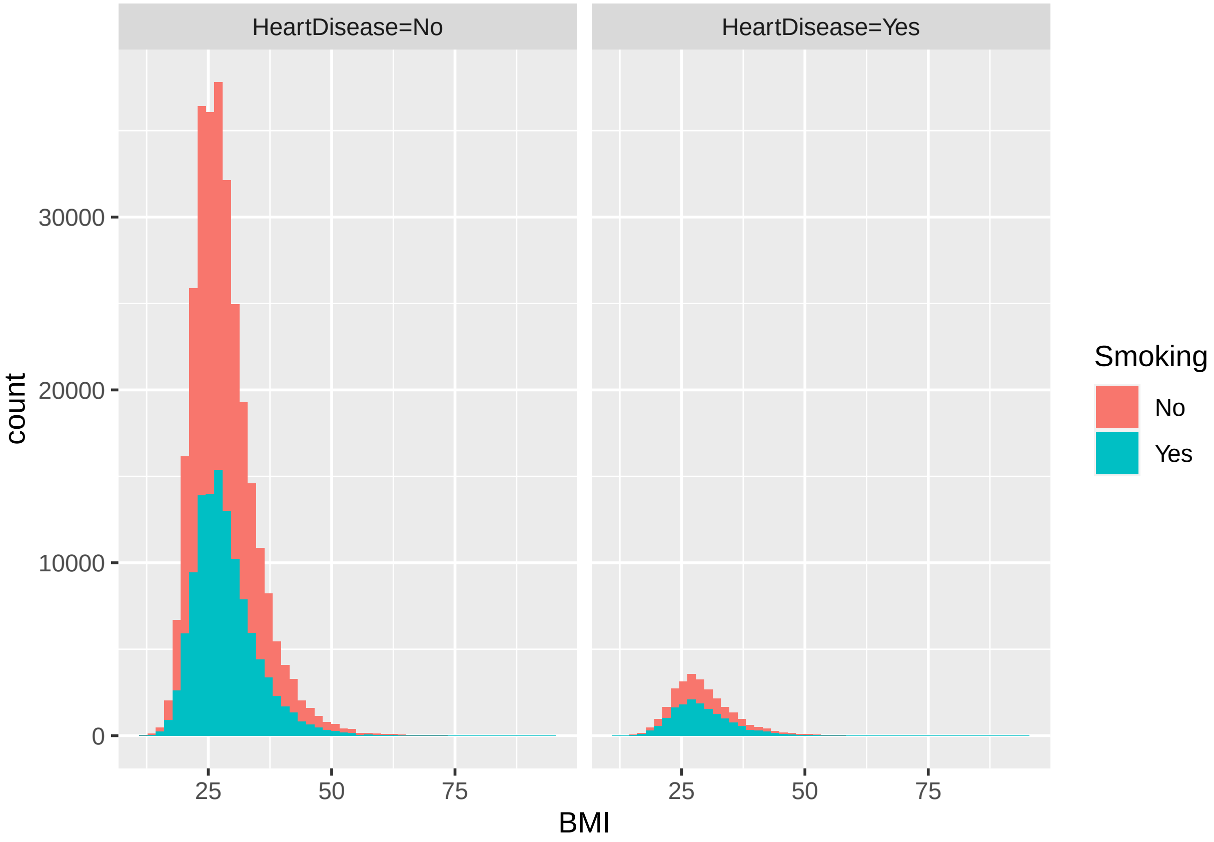
```
ggplot(heart.sample) +
  geom_bar(aes(GenHealth, fill = HeartDisease), position = "fill") +
  facet_grid(Stroke ~ Smoking,labeller = purrr::partial(label_both,sep = "=")) +
  ylab("Proportion")
```



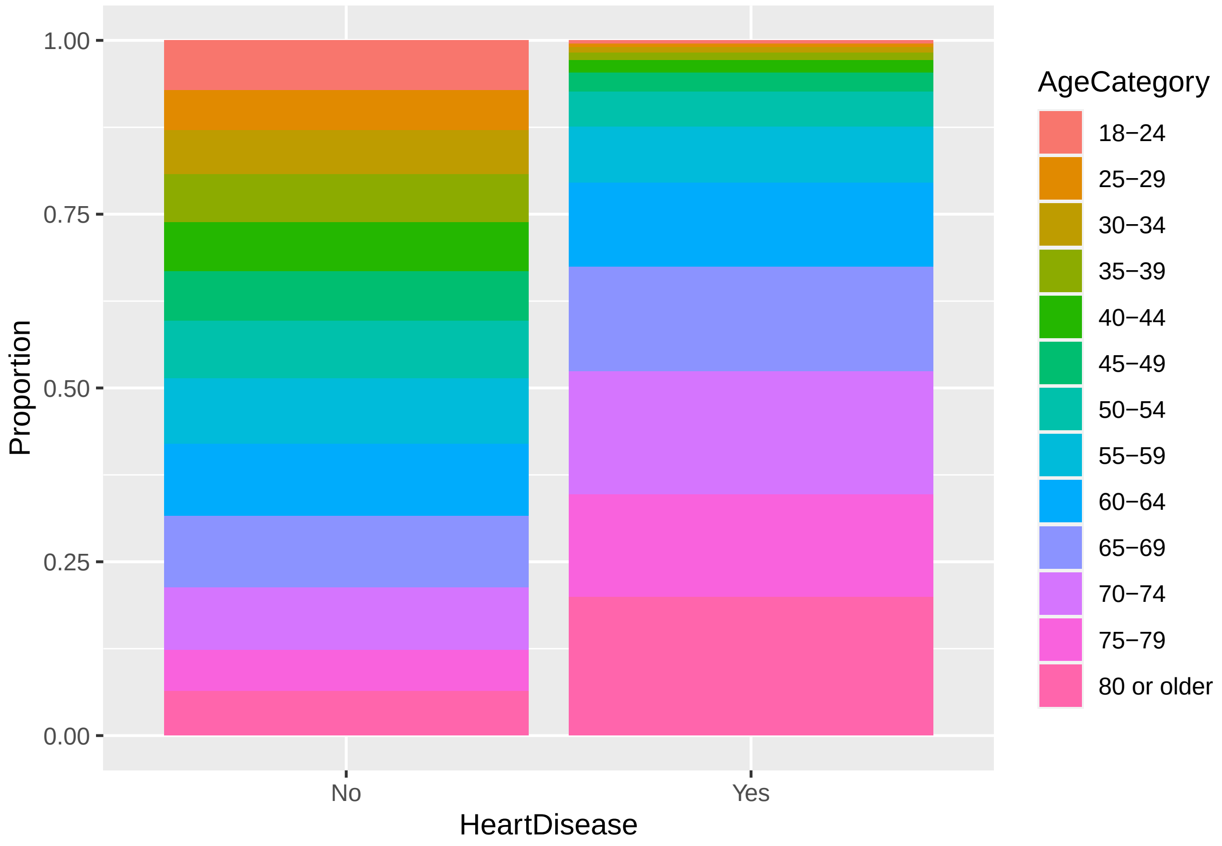
```
ggplot(heart.sample) +
  geom_bar(aes(x = PhysicalActivity, fill = HeartDisease),position = "fill") +
  facet_grid(Smoking ~ AlcoholDrinking, labeller = purrr::partial(label_both,sep = "=")) +
  ylab("Proportion")
```



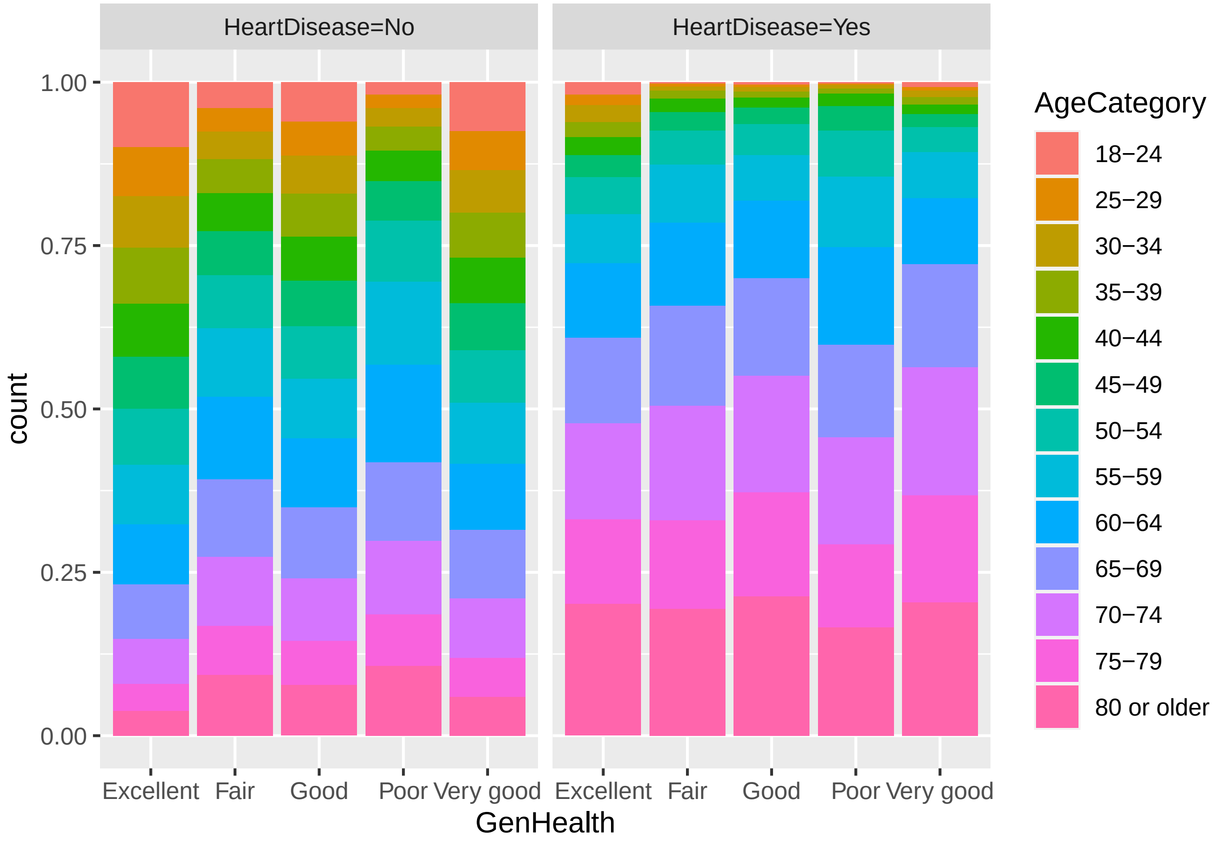
```
ggplot(heart.sample) +
  geom_histogram(aes(BMI, fill=Smoking), bins = 50) +
  facet_wrap(vars(HeartDisease),labeller = purrr::partial(label_both,sep = "="))
```



```
ggplot(heart.sample) +
  geom_bar(aes(HeartDisease, fill = AgeCategory), position = "fill",) +
  ylab("Proportion")
```



```
ggplot(heart.sample) +
  geom_bar(aes(GenHealth, fill = AgeCategory), position = "fill") +
  facet_wrap(vars(HeartDisease),labeller = purrr::partial(label_both,sep = "="))
```


HeartDisease=No
HeartDisease=Yes
count
1.00
0.75
0.50
0.25
0.00
Excellent
Fair
Good
Poor
Very good
GenHealth
AgeCategory
18−24
25−29
30−34
35−39
40−44
45−49
50−54
55−59
60−64
65−69
70−74
75−79
80 or older