Homework#1

Riyesh Nath

2025-09-03

GIVEN COMMAND BY PROFESSOR PLUS SOME OTHER COMMANDS TO READ THE ENTIRE DATA

```
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
```

```
load("data/d_HHP2020_24.Rdata")
d_HHP2020_24[1:10,1:6]
```

```
##    Age Gender    Education Mar_Stat income_midpoint  Race
## 1   34 female college grad  Married           62500 white
## 2   65   male some college divorced           30000 white
## 3   44 female college grad  Married          225000 other
## 4   56   male some college divorced           12500 white
## 5   57 female   adv degree    never           62500 white
## 6   44 female   adv degree  Married          125000 white
## 7   37 female   adv degree  Married           62500 Black
## 8   59   male college grad  Married           82500 white
## 9   51 female        lt hs    never           12500 Black
## 10  29 female    assoc deg  Married           40000 white
```

```
head(d_HHP2020_24,5)
```

```
##   Age Gender    Education Mar_Stat income_midpoint  Race     Hispanic
## 1  34 female college grad  Married           62500 white not Hispanic
## 2  65   male some college divorced           30000 white not Hispanic
## 3  44 female college grad  Married          225000 other not Hispanic
## 4  56   male some college divorced           12500 white not Hispanic
## 5  57 female   adv degree    never           62500 white not Hispanic
##   Number_people_HH Number_kids_HH Number_adults_HH private_health_ins
## 1                4              2                2                  0
## 2                1              0                1                  0
## 3                2              0                2                  0
## 4                2              0                2                  0
## 5                1              0                1                  0
##   public_health_ins                        work_kind
## 1                 0           employed by private co
## 2                 0                             <NA>
## 3                 0 employed by nonprofit or charity
## 4                 0                             <NA>
## 5                 0 employed by nonprofit or charity
##                            workloss income_midpoint_factor     State  Region
## 1                                no                  62500 Tennessee   South
## 2                                no                  30000   Alabama   South
## 3                                no                 225000  Michigan Midwest
## 4 yes recent household loss of work                  12500   Alabama   South
## 5                                no                  62500   Alabama   South
##      Census_division DOWN ANXIOUS WORRY INTEREST YEAR Begin_Date K4SUM
## 1 East South Central    1       4     3        1   20 2020-04-23     9
## 2 East South Central    4       3     4        4   20 2020-04-23    15
## 3 East North Central    1       1     1        1   20 2020-04-23     4
## 4 East South Central    4       4     4        4   20 2020-04-23    16
## 5 East South Central    2       2     1        2   20 2020-04-23     7
```

```
attach(d_HHP2020_24)

summary(d_HHP2020_24)
```

```
##       Age           Gender              Education           Mar_Stat
##  Min.   :17.00   male  :410536   lt hs       :  6787   Married  :556611
##  1st Qu.:39.00   female:566464   some hs     : 14934   widowed  : 54162
##  Median :52.00   trans :  1989   high school :122541   divorced :152705
##  Mean   :52.25   other :  5801   some college:210698   separated: 17850
##  3rd Qu.:65.00                   assoc deg   :103575   never    :195037
##  Max.   :88.00                   college grad:279400   NA's     :  8425
##                                  adv degree  :246855
##  income_midpoint     Race                Hispanic      Number_people_HH
##  Min.   : 12500   white:806002   not Hispanic:895979   Min.   : 1.000
##  1st Qu.: 40000   Black: 80846   Hispanic    : 88811   1st Qu.: 2.000
##  Median : 82500   Asian: 48885                         Median : 2.000
##  Mean   : 95461   other: 49057                         Mean   : 2.715
##  3rd Qu.:125000                                        3rd Qu.: 4.000
##  Max.   :225000                                        Max.   :10.000
##  NA's   :187771
##  Number_kids_HH  Number_adults_HH                    private_health_ins
##  Min.   :0.000   Min.   : 1.000   0                           : 74413
##  1st Qu.:0.000   1st Qu.: 2.000   has private health insurance:607599
##  Median :0.000   Median : 2.000   no private health insurance :149384
##  Mean   :0.623   Mean   : 2.092   NA's                        :153394
##  3rd Qu.:1.000   3rd Qu.: 2.000
##  Max.   :5.000   Max.   :10.000
##
##                   public_health_ins                             work_kind
##  0                          : 74413   employed by govt                : 96450
##  has public health insurance:302958   employed by private co          :320047
##  no public health insurance :425600   employed by nonprofit or charity: 74364
##  NA's                       :181819   self employed                   : 68547
##                                       work for family business        : 11698
##                                       NA's                            :413684
##
##                               workloss      income_midpoint_factor
##  yes recent household loss of work:171404   125000 :145006
##  no                               :794667   62500  :134183
##  NA's                             : 18719   82500  :112727
##                                             225000 : 92900
##                                             40000  : 85421
##                                             (Other):226782
##                                             NA's   :187771
##           State             Region                Census_division
##  California   : 71958   South    :317309   South Atlantic    :173111
##  Texas        : 49059   West     :310873   Pacific           :160919
##  Washington   : 37615   Northeast:151554   Mountain          :149954
##  Florida      : 33825   Midwest  :205054   West North Central:104736
##  Michigan     : 26479                      East North Central:100318
##  Massachusetts: 26236                      West South Central: 89496
##  (Other)      :739618                      (Other)           :206256
##       DOWN          ANXIOUS             WORRY          INTEREST
##  Min.   :1.00     Min.   :1.00     Min.   :1.00     Min.   :1.00
##  1st Qu.:1.00     1st Qu.:1.00     1st Qu.:1.00     1st Qu.:1.00
##  Median :1.00     Median :2.00     Median :1.00     Median :1.00
##  Mean   :1.63     Mean   :1.91     Mean   :1.72     Mean   :1.65  
##  3rd Qu.:2.00     3rd Qu.:2.00     3rd Qu.:2.00     3rd Qu.:2.00  
##  Max.   :4.00     Max.   :4.00     Max.   :4.00     Max.   :4.00  
##  NA's   :108234   NA's   :106951   NA's   :108419   NA's   :108683  
##       YEAR         Begin_Date           K4SUM      
##  Min.   :20.00   Min.   :2020-04-23   Min.   : 4.00  
##  1st Qu.:20.00   1st Qu.:2020-12-09   1st Qu.: 4.00  
##  Median :22.00   Median :2022-04-27   Median : 6.00  
##  Mean   :21.73   Mean   :2022-05-03   Mean   : 6.91  
##  3rd Qu.:23.00   3rd Qu.:2023-08-23   3rd Qu.: 8.00  
##  Max.   :24.00   Max.   :2024-07-23   Max.   :16.00  
##                                       NA's   :111831
```

```
summary(Age[Gender == "female"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   17.00   39.00   52.00   51.62   64.00   88.00
```

```
summary(Age[Gender == "male"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   17.00   40.00   54.00   53.29   67.00   88.00
```

```
summary(Age[Gender == "trans"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   17.00   26.00   31.00   36.02   41.00   88.00
```

```
summary(Age[Gender == "other"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   17.00   31.00   43.00   45.88   59.00   88.00
```

```
mean(Age[Gender == "female"])
```

```
## [1] 51.61668
```

```
sd(Age[Gender == "female"])
```

```
## [1] 15.59165
```

```
mean(Age[Gender == "male"])
```

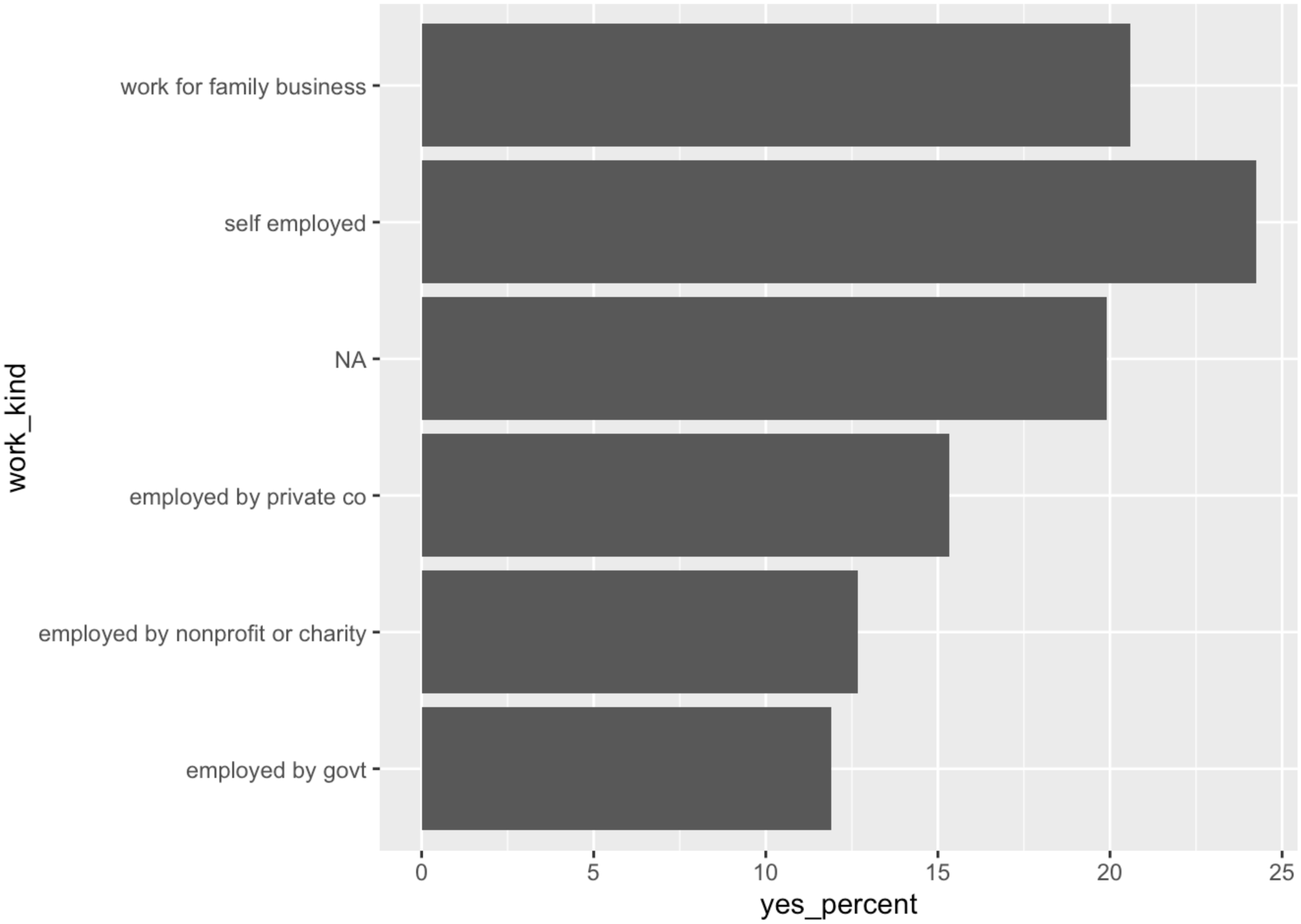
```
## [1] 53.28593
```

```
sd(Age[Gender == "male"])
```

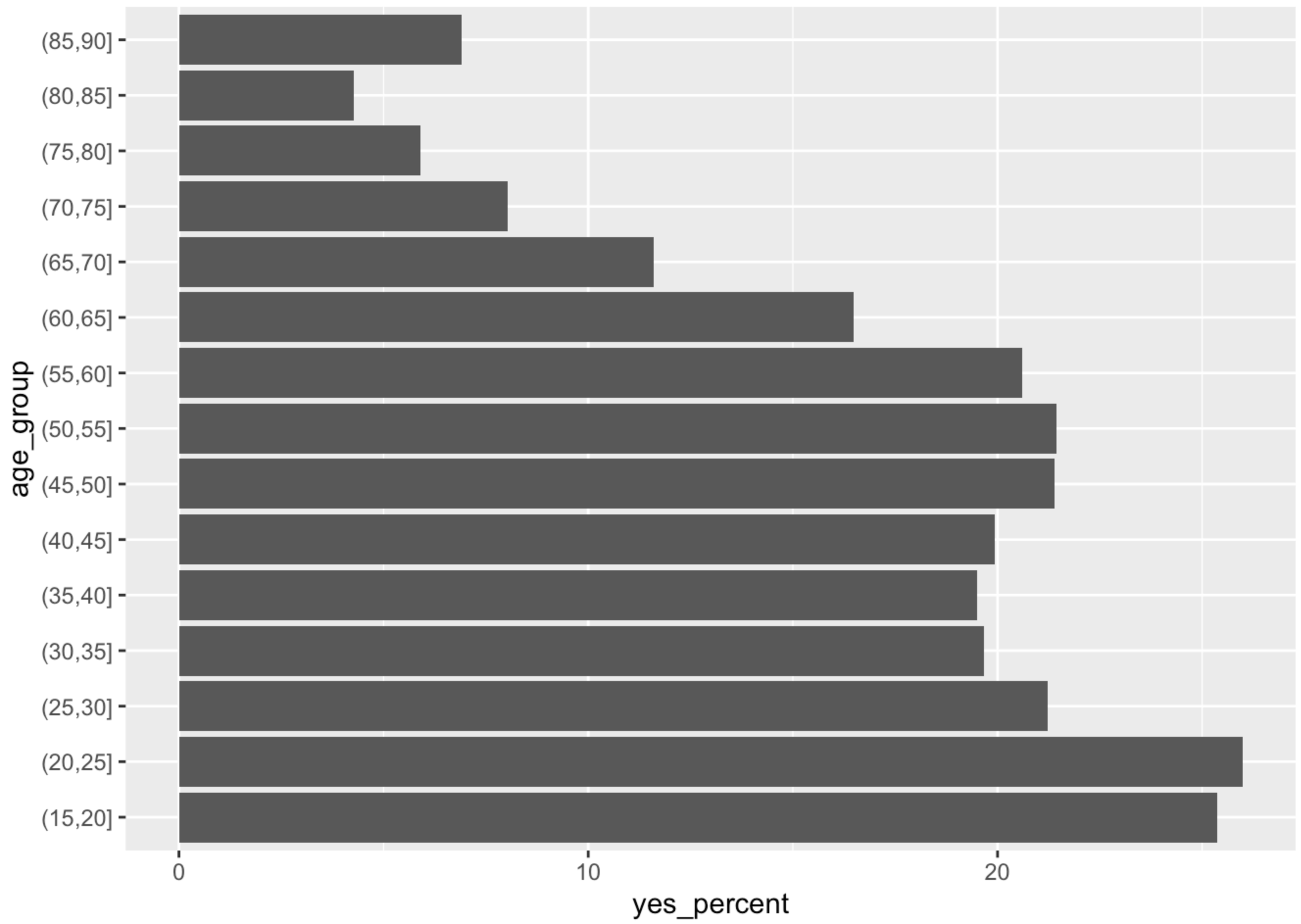
```
## [1] 16.28551
```

BASIC DATA EXPLORATION WITH A FOCUS ON JOB LOSS

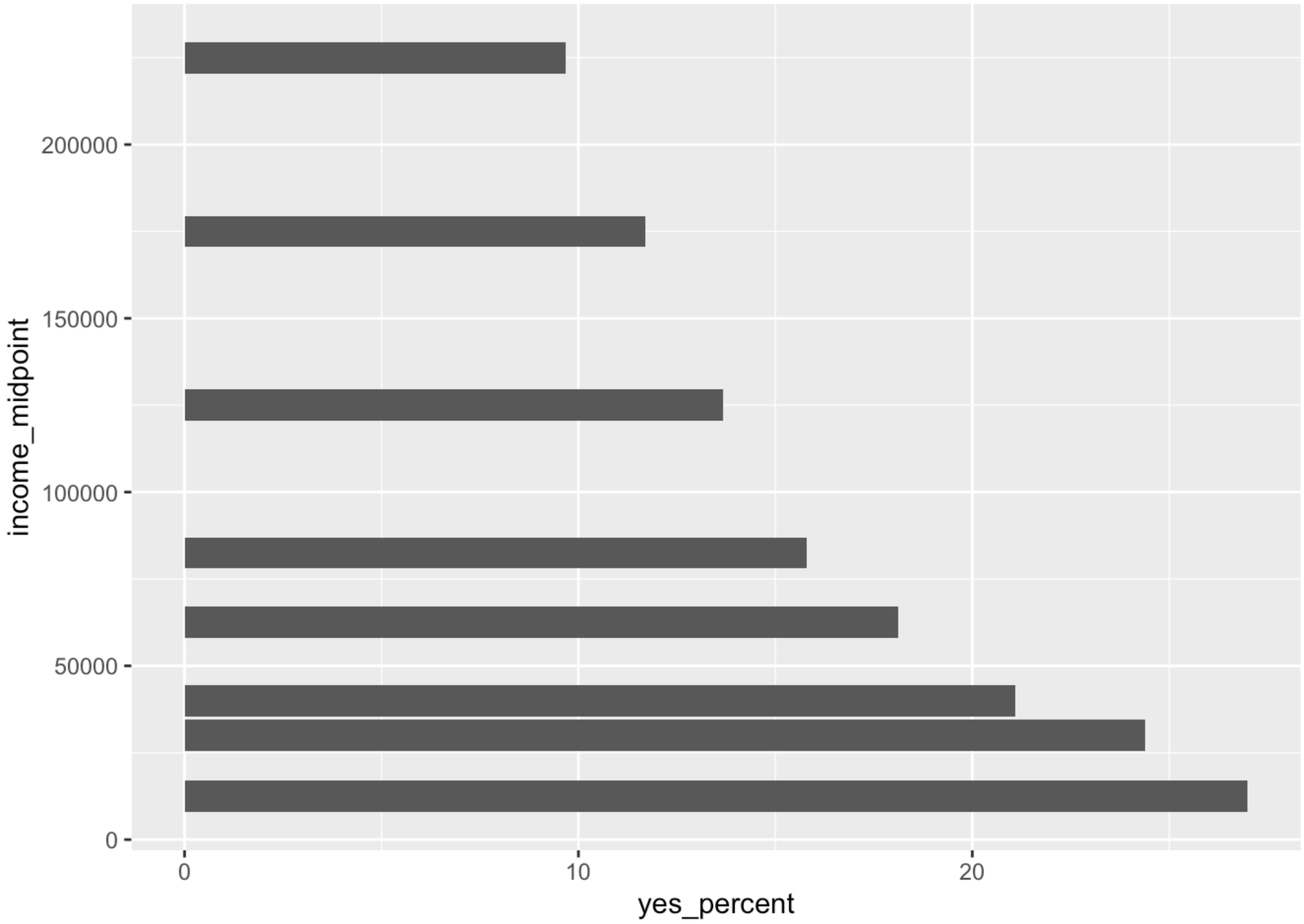
```
library(ggplot2)

# CHECK TO SEE THE PERCENT OF JOB LOSS BY SECTOR

jobloss <- d_HHP2020_24 %>%
  mutate(work_kind = if_else(is.na(work_kind), "NA", work_kind)) %>%
  group_by(work_kind) %>%
  summarise(
    total = n(),
    num_of_yes = sum(workloss == "yes recent household loss of work", na.rm = TRUE),
    yes_percent = (num_of_yes / total)*100,
    .groups = "drop"
  )

ggplot(jobloss, aes(x= work_kind, y=yes_percent)) +
  geom_col() +
  coord_flip()
```



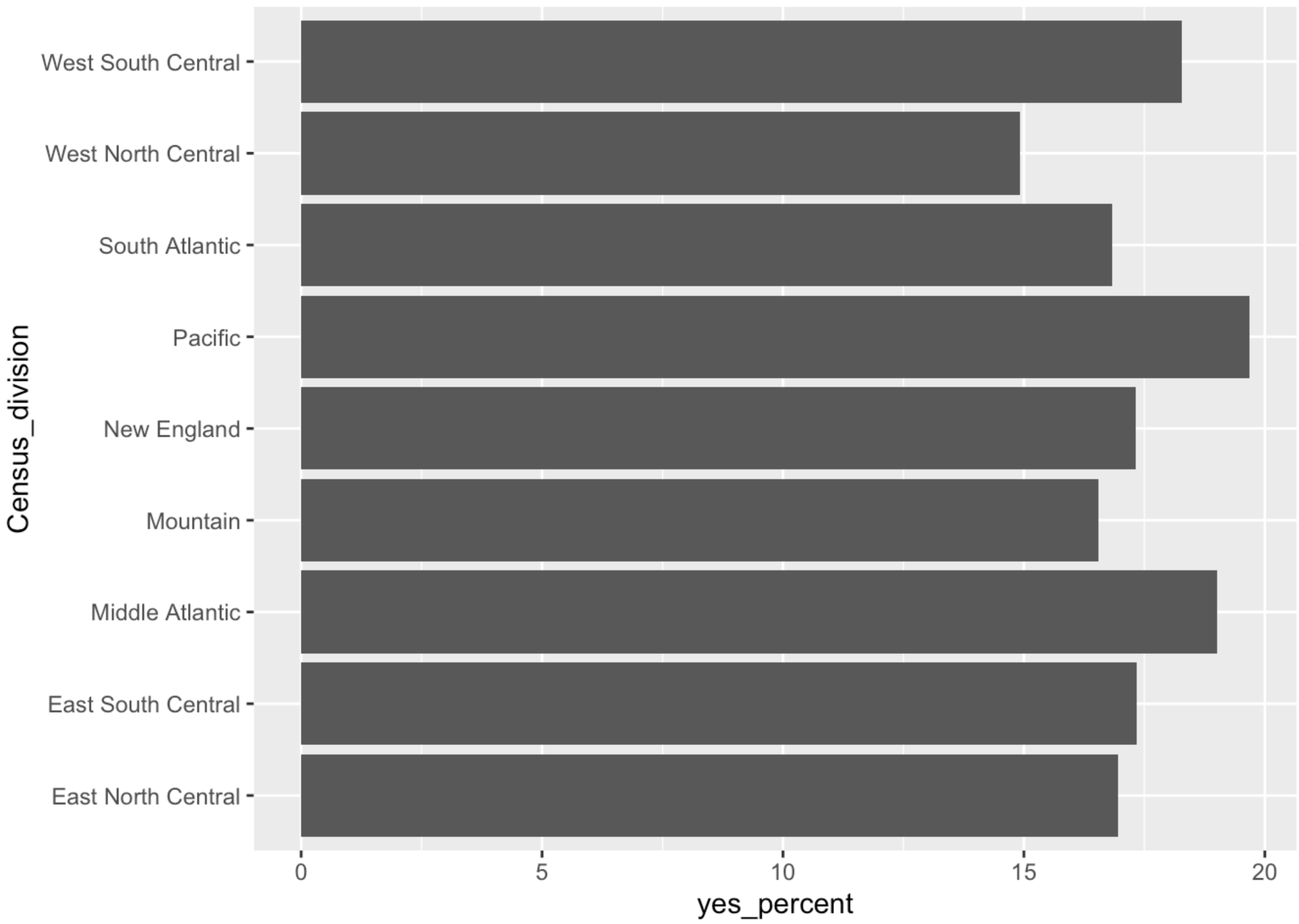
```
# CHECK BY AGE
jobloss <- d_HHP2020_24 %>%
  mutate(
    age_group = cut(Age, breaks = seq(0, 100, by = 5), right = TRUE)
    ) %>%
  group_by(age_group) %>%
  summarise(
    total = n(),
    num_of_yes = sum(workloss == "yes recent household loss of work", na.rm = TRUE),
    yes_percent = (num_of_yes / total)*100,
    .groups = "drop"
  )

ggplot(jobloss, aes(x= age_group, y=yes_percent)) +
  geom_col() +
  coord_flip()
```



```
# CHECK BY INCOME MIDPOINT
jobloss <- d_HHP2020_24 %>%
  filter(!is.na(income_midpoint)) %>%
  group_by(income_midpoint) %>%
  summarise(
    total = n(),
    num_of_yes = sum(workloss == "yes recent household loss of work", na.rm = TRUE),
    yes_percent = (num_of_yes / total)*100,
    .groups = "drop"
  )

ggplot(jobloss, aes(x= income_midpoint, y=yes_percent)) +
  geom_col() +
  coord_flip()
```



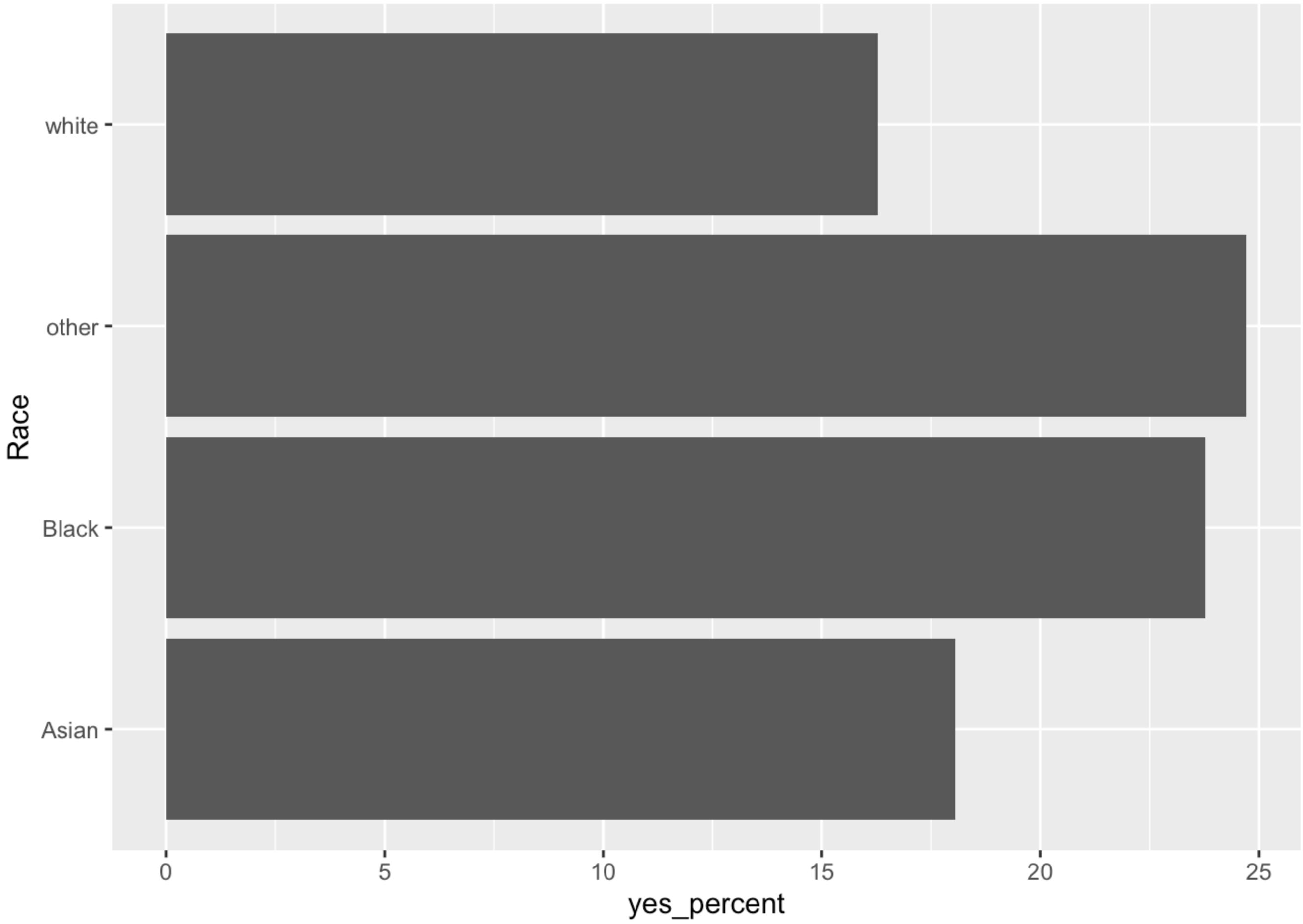
```
# CHECK BY STATE
jobloss <- d_HHP2020_24 %>%
  mutate(Census_division = if_else(is.na(Census_division), "NA", Census_division)) %>%
  group_by(Census_division) %>%
  summarise(
    total = n(),
    num_of_yes = sum(workloss == "yes recent household loss of work", na.rm = TRUE),
    yes_percent = (num_of_yes / total)*100,
    .groups = "drop"
  )

ggplot(jobloss, aes(x= Census_division, y=yes_percent)) +
  geom_col() +
  coord_flip()
```



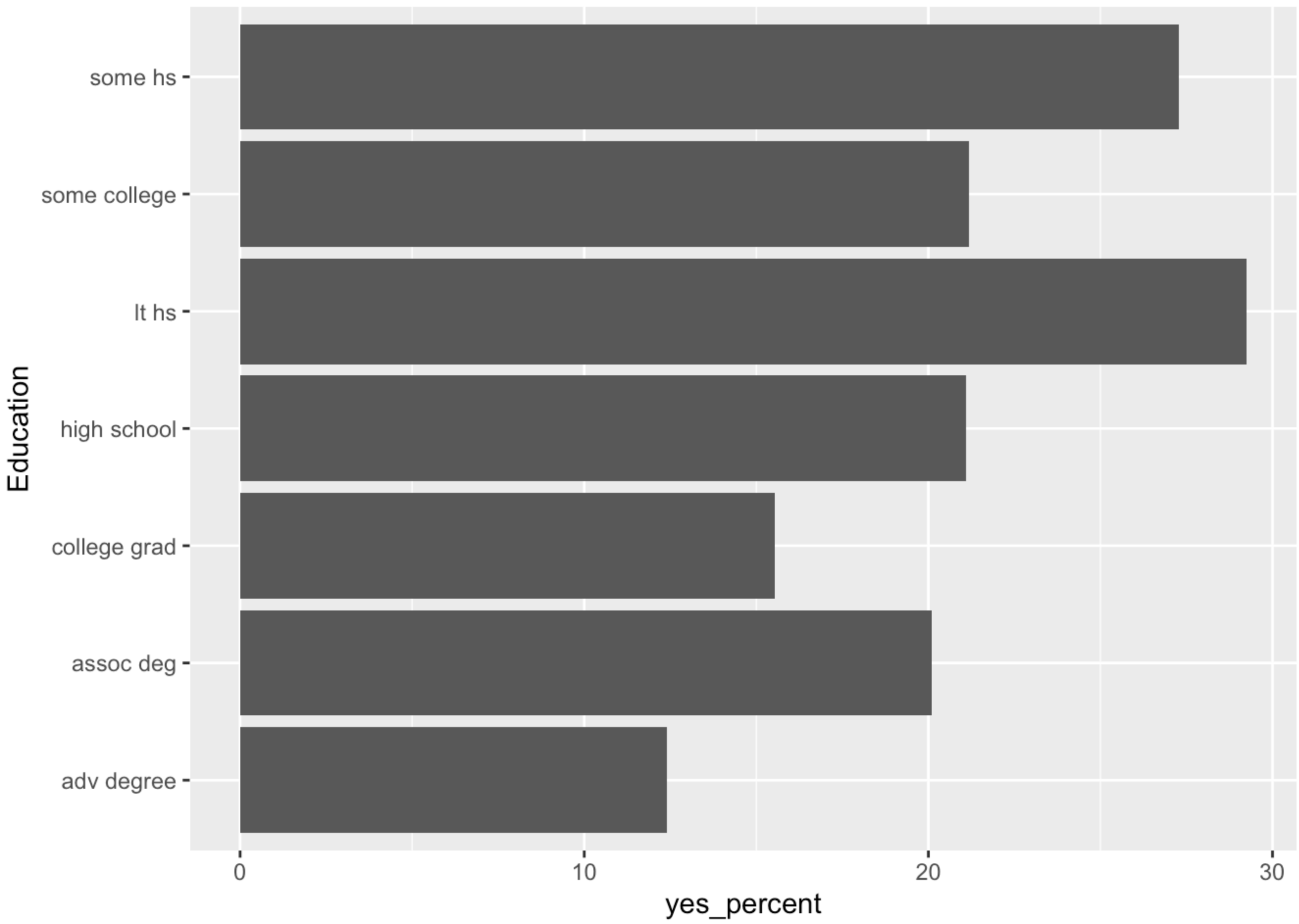
```
# CHECK BY RACE
jobloss <- d_HHP2020_24 %>%
  mutate(Race = if_else(is.na(Race), "NA", Race)) %>%
  group_by(Race) %>%
  summarise(
    total = n(),
    num_of_yes = sum(workloss == "yes recent household loss of work", na.rm = TRUE),
    yes_percent = (num_of_yes / total)*100,
    .groups = "drop"
  )
print(jobloss)
```

```
## # A tibble: 4 × 4
##   Race   total num_of_yes yes_percent
##   <chr>  <int>      <int>       <dbl>
## 1 Asian  48885       8822        18.0
## 2 Black  80846      19207        23.8
## 3 other  49057      12129        24.7
## 4 white 806002     131246        16.3
```

```
ggplot(jobloss, aes(x= Race, y=yes_percent)) +
  geom_col() +
  coord_flip()
```



```
# CHECK BY EDUCATION
jobloss <- d_HHP2020_24 %>%
  mutate(Education = if_else(is.na(Education), "NA", Education)) %>%
  group_by(Education) %>%
  summarise(
    total = n(),
    num_of_yes = sum(workloss == "yes recent household loss of work", na.rm = TRUE),
    yes_percent = (num_of_yes / total)*100,
    .groups = "drop"
  )

ggplot(jobloss, aes(x= Education, y=yes_percent)) +
  geom_col() +
  coord_flip()
```



AS WE ARE JUST IN EXPLORATION PHASE, I AM CURIOUS TO SEE HOW MUCH PERCENT OF THOSE WHO ARE BLACK, HAVE EDUCATION BELOW COLLEGE GRAD LEVEL, AND MAKE LESS THAN 60,000 LOST THEIR JOB DURING THIS PERIOD.

```
black_sixtythousand_lowerEducation <- d_HHP2020_24 %>%
  filter(Education %in% c("some hs", "some college", "lt hs", "high school"), Race == "B
lack", income_midpoint < 60000) %>%
  summarise(
    total = n(),
    num_of_yes = sum(workloss == "yes recent household loss of work", na.rm = TRUE),
    yes_percent = (num_of_yes / total)*100
  )

print(black_sixtythousand_lowerEducation)
```

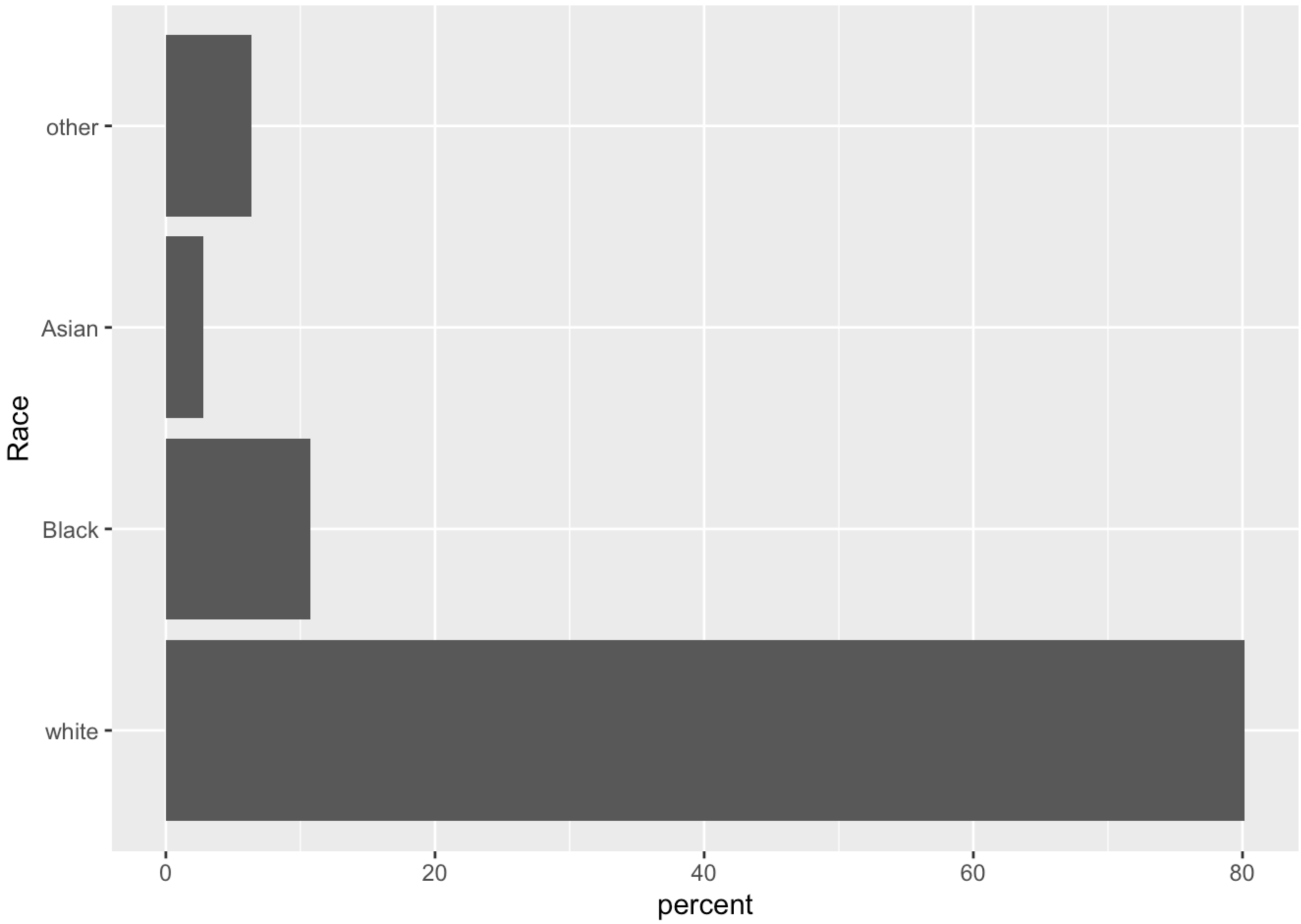
```
##   total num_of_yes yes_percent
## 1 17584       5326     30.2889
```

CURIOUS TO SEE IF MAJORITY FOR THOSE WITH LOWER EDUCATION AND WAGES ARE BLACK AMERICANS.

```
IT_HS_DATASET <- d_HHP2020_24 %>%
  filter(Education %in% c("some hs", "some college", "lt hs", "high school")) %>%
  count(Race, name = "total") %>%
  mutate(percent = 100 ∗ total / sum(total))

print(IT_HS_DATASET)
```

```
##    Race  total   percent
## 1 white 284614 80.181992
## 2 Black  38073 10.725997
## 3 Asian   9767  2.751578
## 4 other  22506  6.340433
```

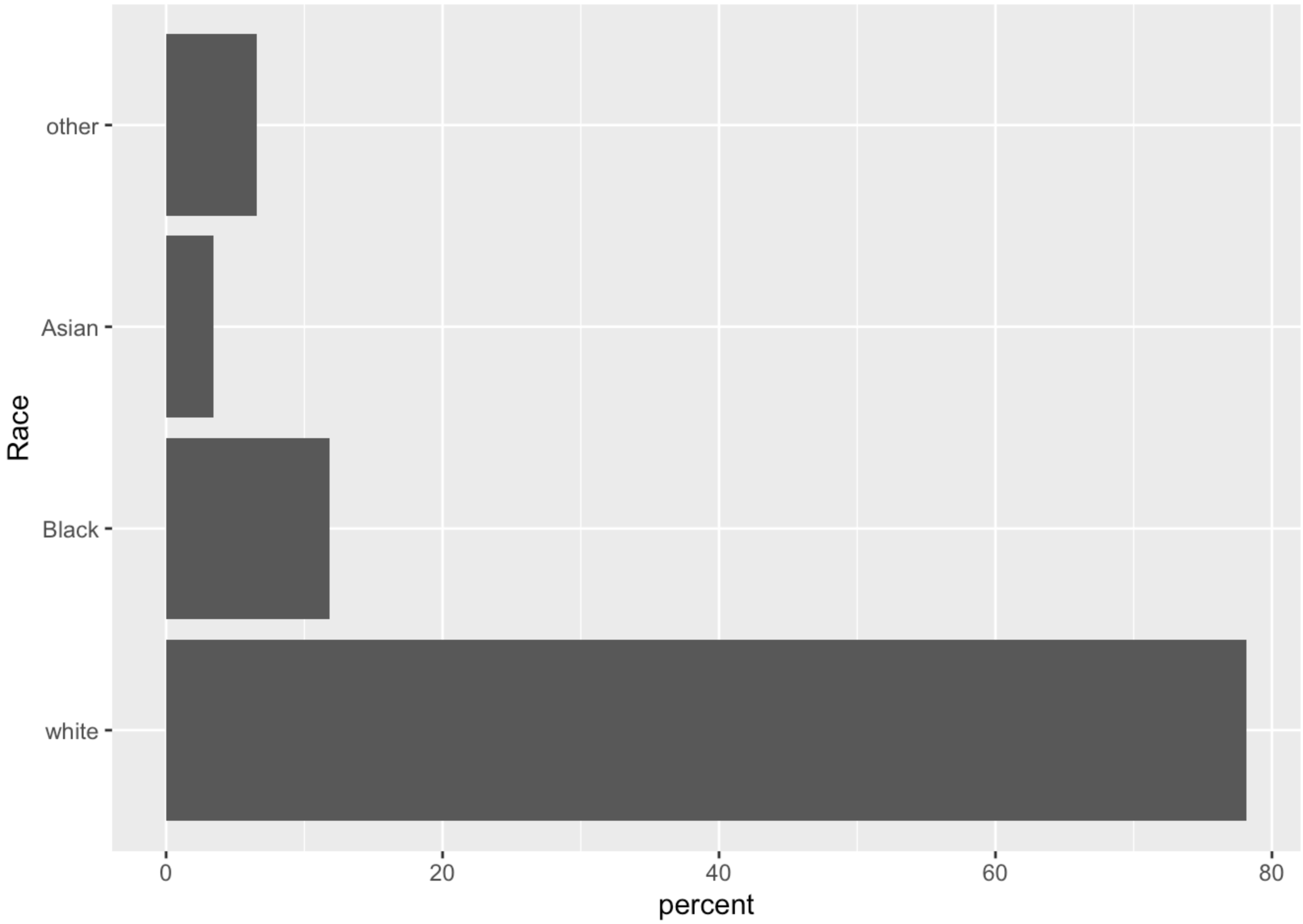
```
ggplot(IT_HS_DATASET, aes(x=Race, y=percent)) + geom_col() + coord_flip()
```



```
WAGE_DATASET <- d_HHP2020_24 %>%
  filter(income_midpoint < 60000) %>%
  count(Race, name = "total") %>%
  mutate(percent = 100 ∗ total / sum(total))

print(WAGE_DATASET)
```

```
##    Race  total   percent
## 1 white 186706 78.186402
## 2 Black  28291 11.847351
## 3 Asian   8104  3.393692
## 4 other  15695  6.572556
```

```
ggplot(WAGE_DATASET, aes(x=Race, y=percent)) + geom_col() + coord_flip()
```



LOOKING AT THIS, IT SEEMS THAT THIS MIGHT BE MORE A RATIO THAN TOTAL PERCENT DUE TO LARGE WHITE AMERICAN TOTAL. GIVEN TIME, I WOULD STUDY THIS FURTHER USING RATIO RATHER THAN JUST A BASE TOTAL.