
RStudio Basics

1

Load the data set and get the names of all of the variables included. Identify two interesting variables.

```
#load data
load("data/countries2.rda")

#get names of the variables
names(countries2)
```

```
## [1] "hdi_rank"      "wbcountry"      "ccode"
## [4] "hdi"           "lifexp"          "mnschool"
## [7] "gini1019"      "femexp"          "malexp"
## [10] "fem_mnschool"  "male_mnschool"  "gender_inequality"
## [13] "matmort"       "teen_fert"      "fem_leg"
## [16] "fem_seced"     "male_seced"     "fem_labor"
## [19] "male_labor"    "chg_pop"        "urban"
## [22] "fert1520"      "inf_no_dtp"     "inf_no_measel"
## [25] "infant_mort"   "kid_mort"       "TB100k"
## [28] "health_exp"    "sec_ed"         "educ_exp"
## [31] "jail100k"      "homicide100k"   "fem_suicie"
## [34] "male_suicide"  "food_def"       "net_mig"
## [37] "internet"      "mobile_phone"   "docs10k"
## [40] "hosp10k"       "rural_electric" "fem_loc_gvt"
## [43] "fem_finance"   "redlist"        "skilled1019"
## [46] "mil_exp"       "pop19_M"        "gdp_billions"
## [49] "gdp_pc"
```

This data set contains aggregate variables and data about the economy and regarding population indexes. Some of the variables are numerical, while others are categorical.

One potential dependent variable to study could be “kid_mort”, or infant’s mortality.

```
summary(countries2$kid_mort)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.     NA's
##  1.695   7.056  16.619  28.168  43.628 121.530         2
```

We could study infant’s mortality as related to the independent variable “health_exp”, or expenditures for health-related costs.

```
summary(countries2$health_exp)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.     NA's
##  1.181   4.499   6.286   6.620   8.193  17.143         9
```

2

List the number of all the variables.

```
#View(countries2)
dim(countries2)
```

```
## [1] 195 49
```

The numbers reflect the rows and columns. Therefore the data set 'countries2' contains 195 countries and 49 variables.

3

Create a new object, calculated as the percent in the state who approve of the Biden's performance MINUS the percent in the state who disapprove of Biden's performance. Sort the data set by the new variable and list the six highest and lowest states.

```
library(readxl)
Approve21 <- read_excel("data/Approve21.xlsx")
#View(Approve21)

#calculate proportion of approval
Approve21$net_approve <- (Approve21$Approve) - (Approve21$Disapprove)

#Sort the data set by ``Approve21$net_approve`` and list the six highest and lowest states.
Approve21 <- Approve21[order(Approve21$net_approve),]

#six highest
head(Approve21)
```

```
## # A tibble: 6 x 6
##   state      stateab Approve Disapprove Neither net_approve
##   <chr>      <chr>    <dbl>    <dbl>    <dbl>    <dbl>
## 1 West Virginia WV         23         72         5        -49
## 2 Wyoming     WY         26         68         5        -42
## 3 Oklahoma    OK         29         65         6        -36
## 4 Idaho       ID         30         65         5        -35
## 5 North Dakota ND         31         64         5        -33
## 6 Alabama     AL         31         63         6        -32
```

The six highest states are known to be very conservative, and they are primarily southern states. In this region, people are known to prefer Republican candidates over time.

```
#six lowest
tail(Approve21)
```

```
## # A tibble: 6 x 6
##   state      stateab Approve Disapprove Neither net_approve
##   <chr>      <chr>    <dbl>    <dbl>    <dbl>    <dbl>
## 1 Rhode Island RI         57         37         7         20
```

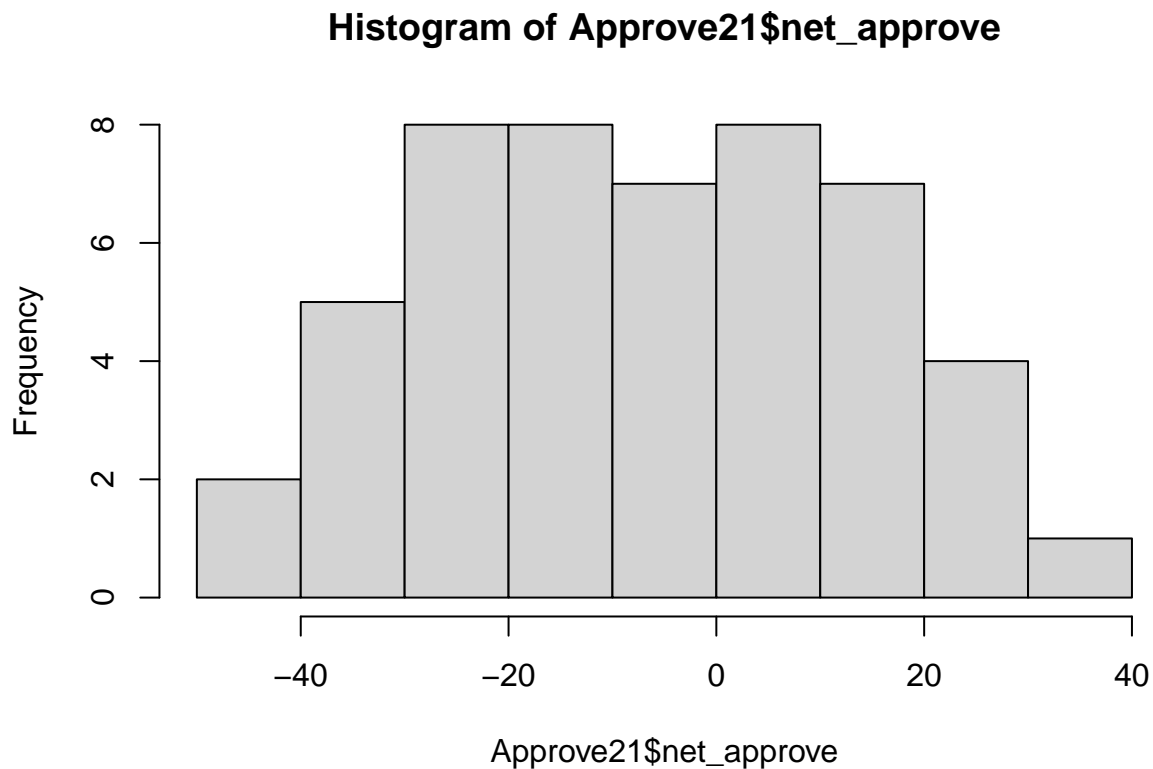
## 2	California	CA	57	35	8	22
## 3	Vermont	VT	58	36	7	22
## 4	Maryland	MD	59	33	8	26
## 5	Hawaii	HI	61	33	6	28
## 6	Massachusetts	MA	63	30	7	33

The six lowest states are primarily coast states and northern states, which are known to overwhelmingly vote liberal and will have potentially favored President Biden during the elections.

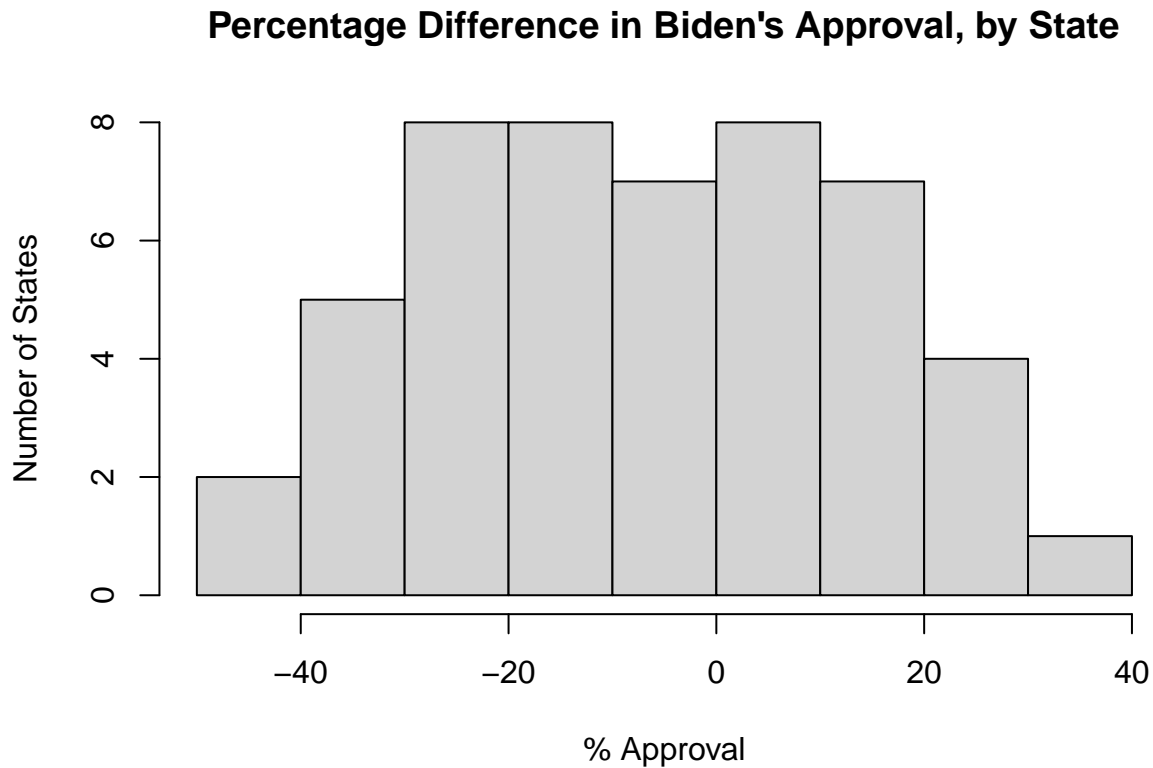
4

Produce a histogram of a variable and describe what you see.

```
#create histogram
hist(Approve21$net_approve)
```



```
#create labels
hist(Approve21$net_approve,
      main="Percentage Difference in Biden's Approval, by State", #Graph title
      ylab="Number of States", #vertical axis label
      xlab="% Approval") #Horizontal axis label
```



The histogram highlights two distinct sets of states. These either have a 10% to 20% of people that support President Biden, or they have a 10% to 20% of people who disapprove of him. This means that there are some states, those with a negative percentage difference, who will lean toward having a majority of Republican voters, while those with a positive percentage difference will have a majority of Democrat voters. This histogram denotes the political polarization that American politics is currently living in. People either moderately approve or disapprove of a candidate, but no political candidate captures an overwhelming amount of support across the board.