```
# Load the data set.

load("polls2020.RData")

# Categorize the functionailty and data types.

head(polls2020)
```

A data.frame: 6 × 9

	state	candidate_name	startdate	enddate	polls
	<chr>	<chr>	<chr>	<chr>	<cl
1	Wyoming	Joseph R. Biden Jr.	10/20/2020	11/1/2020	SurveyMonk
2	Wyoming	Joseph R. Biden Jr.	10/18/2020	10/31/2020	SurveyMonk
3	Wyoming	Joseph R. Biden	10/17/2020	10/30/2020	SurveyMonk

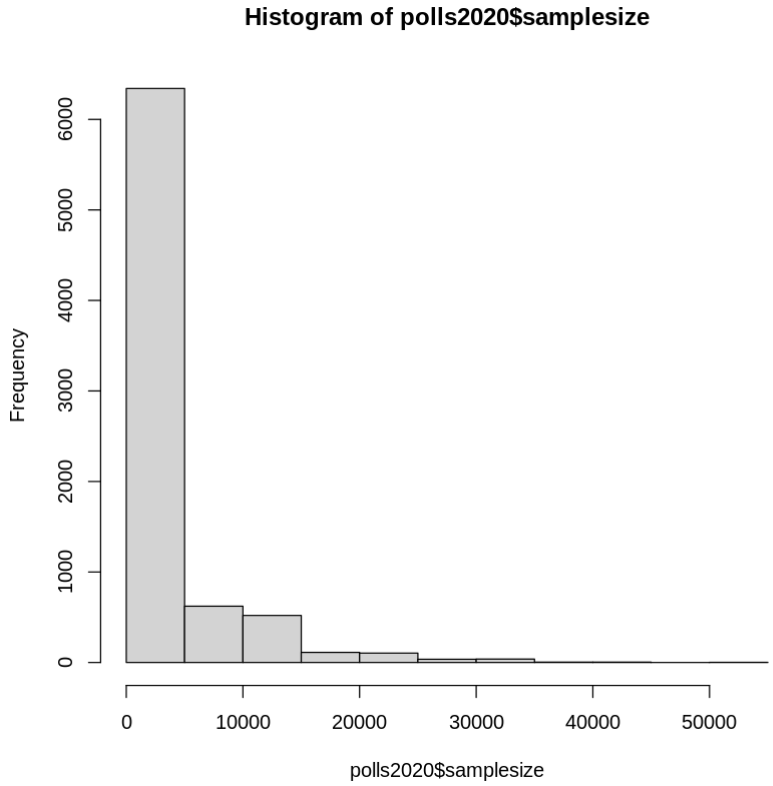
```
# Dimensions
```

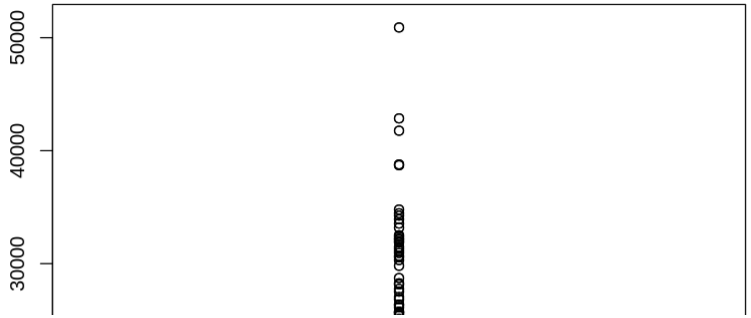
```
# Dimensions

dim(polls2020)


#Summary of the data set

summary(polls2020)
```

```
7786 · 9
   state          candidate_name      startdate          enddate
 Length:7786        Length:7786        Length:7786        Length:7786
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

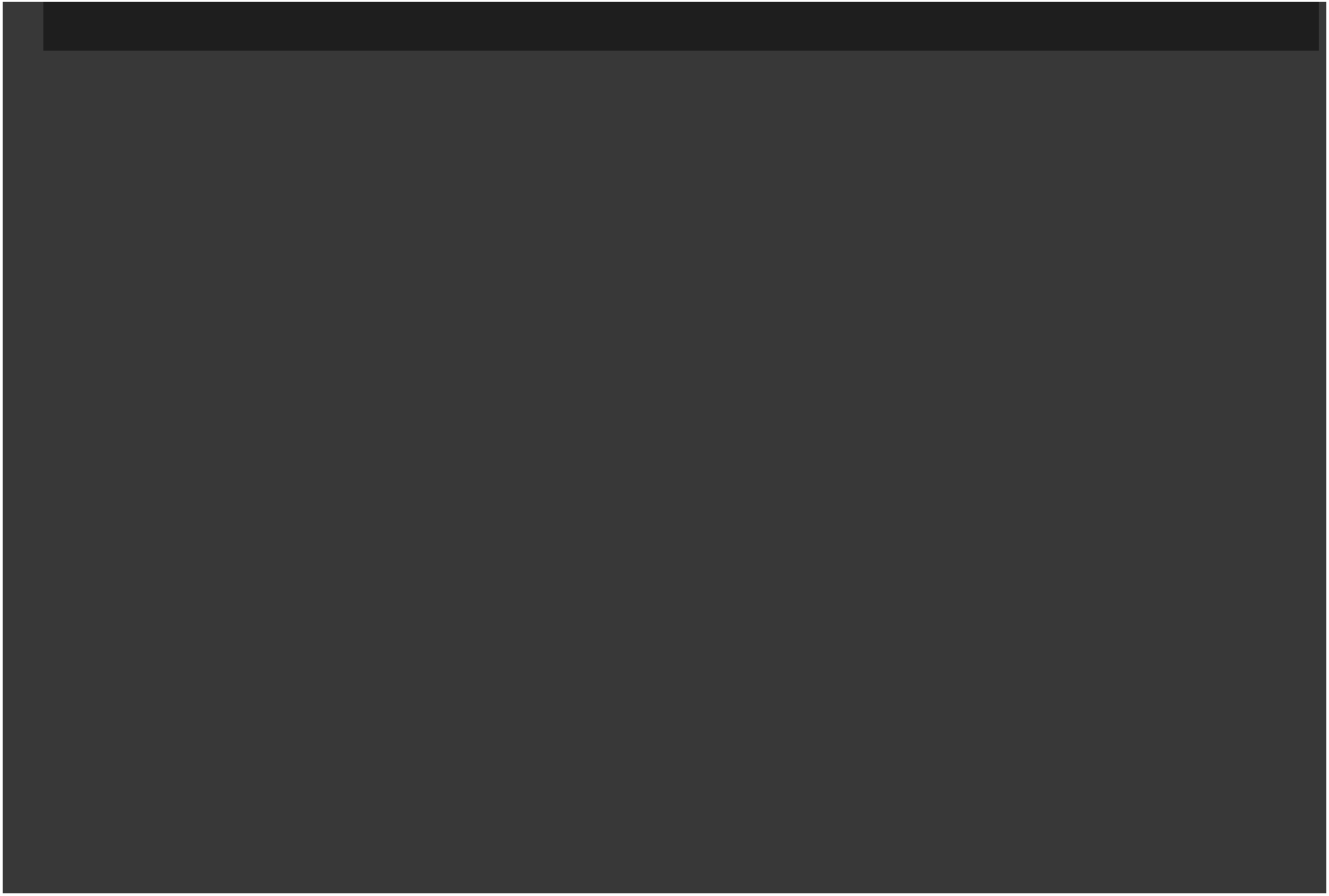
```
#Visualize

install.packages("plyr")

library(plyr)

hist(polls2020$samplesize)

boxplot(polls2020$samplesize)
```

Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)






```
# Correct Errors

polls2020$state < toupper(polls2020$state)

polls2020$candidate_name <- toupper(polls2020$candidate_name)

polls2020$pollster <- toupper(polls2020$pollster)

# Visualize
head(polls2020)
```



```
TRUE · TRUE · TRUE · TRUE · TRUE · TRUE · TRUE · TRUE · TRUE ·
TRUE · TRUE · TRUE · TRUE · TRUE · TRUE · TRUE · TRUE · TRUE ·
TRUE · TRUE · TRUE · TRUE · TRUE · TRUE · TRUE · TRUE · TRUE ·
TRUE · TRUE · TRUE · TRUE · TRUE · TRUE · TRUE · TRUE · TRUE ·
TRUE · TRUE · TRUE · TRUE · TRUE · TRUE · TRUE · TRUE · TRUE ·
```

```
# correct Errors

library(stringr)

polls2020$pollster <- str_trim(polls2020$pollster)

polls2020$candidate_name <- str_trim(polls2020$candidate_name)

head(polls2020)
```

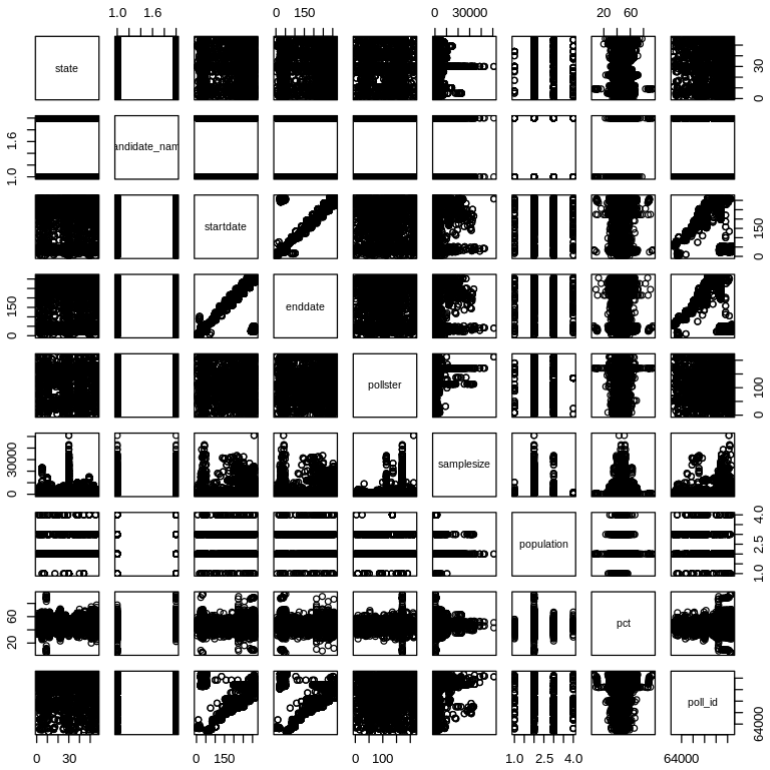
A data.frame: 6 × 9

state	candidate_name	startdate	enddate	p
<chr>	<chr>	<chr>	<chr>	

```
#Replace Outliers
replace <- boxplot.stats(polls2020$samplesize)
polls2020$samplesize[polls2020$samplesize %in% replace] <- median(polls2020$samplesize)


# Remove Missing Values and replace them with 0
sum(is.na(polls2020))

any(is.na(polls2020))

polls2020[is.na(polls2020)] <- 0

sum(is.na(polls2020))

any(is.na(polls2020))
```

2

```
# Visualzie

plot(polls2020)
```



20s completed at 6:40 AM