WebDB

MHetman

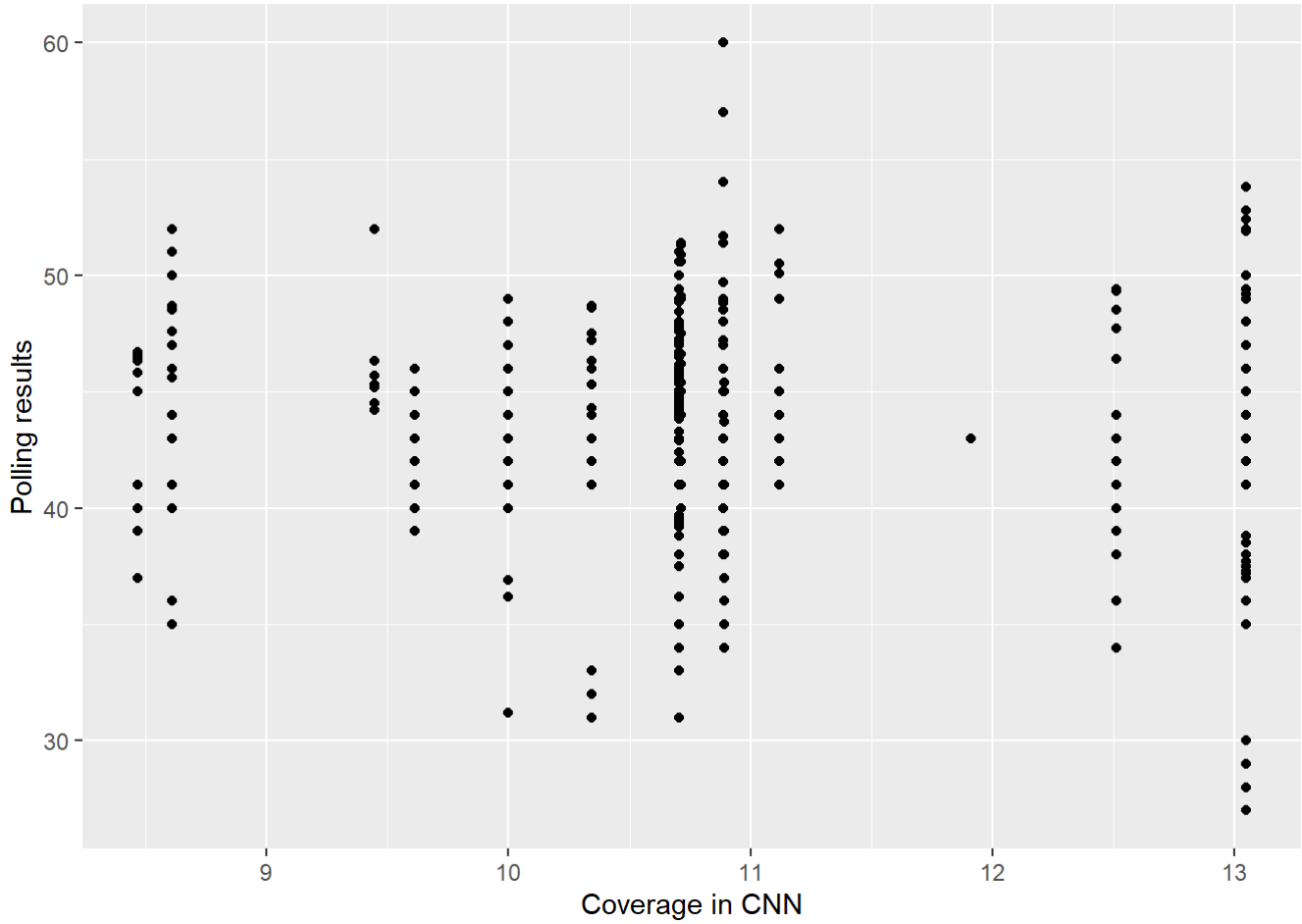
5 12 2019

Data preprocessing:

```
polls_file <- file.path("dataset", 'president_polls.csv')
candidates <- list.files("dataset/candidates", full.names = TRUE)

polls_dt <- read.csv(polls_file) %>% as.data.table()

read_append <- function(file) {
  dt <- fread(file)
  dt[, filename := basename(file)]
  return(dt)
}

candidates_mentions_dt <- lapply(candidates, read_append) %>% rbindlist
names(candidates_mentions_dt)[names(candidates_mentions_dt) == "filename"] <- "Candidate"
candidates_mentions_dt$Candidate <- gsub(".csv","",candidates_mentions_dt$Candidate)

firstup <- function(x) {
  substr(x, 1, 1) <- toupper(substr(x, 1, 1))
  x
}

candidates_mentions_dt$Candidate <- lapply(candidates_mentions_dt$Candidate, firstup)
```

Correlation between Trump and CNN:

```
cnn_trump_mentions <- candidates_mentions_dt[Series == 'CNN' & Candidate == 'Trump', c(1,3)]
cnn_trump_mentions <- separate(cnn_trump_mentions, 1, into = c('month', 'year'), sep = '/')
trump_polls_result <- polls_dt[answer == 'Trump', c("pct", "start_date")]
trump_polls_result <- separate(trump_polls_result, 'start_date', into = c('day', 'month', 'ye
ar'), sep = '/')
trump_polls_result$year = paste0("20", trump_polls_result$year)

trump_polls_cnn_infulence <- sqldf("
  SELECT *
  FROM trump_polls_result JOIN cnn_trump_mentions
  ON (trump_polls_result.month = cnn_trump_mentions.month + 1 AND trump_polls_result.year = c
nn_trump_mentions.year)
  OR (trump_polls_result.month = 1 AND cnn_trump_mentions.month = 12 AND trump_polls_result.y
ear = cnn_trump_mentions.year + 1)
")
ggplot(trump_polls_cnn_infulence, aes(Value, pct)) + geom_point() + scale_x_continuous(name=
"Coverage in CNN") + scale_y_continuous(name="Polling results")
```




```
cor(x = trump_polls_cnn_infulence$Value, y = trump_polls_cnn_infulence$pct, method = "pearso
n")
```

```
## [1] -0.1057554
```