#Environment Preparation

```r
# Install necessary packages
# (if not already installed, uncomment the install.packages lines)
# Base R package, usually pre-installed
# install.packages("stats")
# For creating detailed data visualizations and plots
# install.packages("ggplot2")
# For efficient data manipulation with functions
# install.packages("dplyr")
# For working with dates and times, including parsing and formatting
# install.packages("lubridate")

# Load necessary libraries for data analysis and visualization
library(stats)
library(ggplot2)
library(dplyr)
library(lubridate)


CES2020 <- read.csv("data/CES20_Common_OUTPUT_vv_small.csv")

CES2020 <- CES2020 %>%
  mutate(CC20_410_n = as.numeric(as.factor(CES2020$CC20_410)))

table(CES2020$voted_R)
```

```
## < table of extent 0 >
```

```r
CES2020 <- CES2020 %>%
  mutate(voted_R = case_when(
    CC20_410_n == 2 ~ 1,
    CC20_410_n == 5 ~ 0,
    CC20_410_n %in% c(1, 3, 4, 7) ~ NA_real_,
    TRUE ~ NA_real_
  ))

table(CES2020$voted_R)
```

```
##
##     0     1
## 26188 17702
```

```r
CES2020 <- CES2020 %>%
  mutate(male = if_else(gender == "Male", 1, 0))

CES2020 <- CES2020 %>%
  mutate(CC20_302_ind = case_when(
    CC20_302 == "Gotten much better" | CC20_302 == "Gotten somewhat better" ~ 1,
    CC20_302 == "Stayed about the same" ~ 0,
    CC20_302 == "Gotten much worse" | CC20_302 == "Gotten somewhat worse" ~ -1,
    TRUE ~ NA_real_ # Handling any other or missing values
  ))
```

```
CES2020 <- CES2020 %>%
  mutate(CC20_303_ind = case_when(
    CC20_303 == "Increased a lot" | CC20_303 == "Increased somewhat" ~ 1,
    CC20_303 == "Stayed about the same" ~ 0,
    CC20_303 == "Decreased a lot" | CC20_303 == "Decreased somewhat" ~ -1,
    TRUE ~ NA_real_ # Handling any other or missing values
  ))

CES2020 <- CES2020 %>%
  mutate(abortion_position = case_when(
    # Liberal position
    CC20_332a == "Support" & CC20_332f == "Oppose" ~ 1,
    # In-between position
    CC20_332a == "Oppose" & CC20_332f == "Oppose" ~ 0,
    # Conservative position
    CC20_332a == "Oppose" & CC20_332f == "Support" ~ -1,
    # Exclude invalid/unconstrained responses
    CC20_332a == "Support" & CC20_332f == "Support" ~ NA_real_
  ))


CES2020$age <- 2020 - CES2020$birthyr

table(CES2020$age)
```

```
##
##   18   19   20   21   22   23   24   25   26   27   28   29   30   31   32   33
##  584  618 1144  916  878  879  940 1124 1056 1168 1248 1309 1037  960  996  948
##   34   35   36   37   38   39   40   41   42   43   44   45   46   47   48   49
## 1027 1162 1014 1014 1039 1014 1165 1019 1006  962  933  687  721  707  699  840
##   50   51   52   53   54   55   56   57   58   59   60   61   62   63   64   65
## 1003  962  932  934  989 1199 1235 1292 1328 1331 1201 1200 1077 1373 1225 1044
##   66   67   68   69   70   71   72   73   74   75   76   77   78   79   80   81
##  911  814  990 1049  879  700  792  834  716  617  522  529  500  388  317  272
##   82   83   84   85   86   87   88   89   90   91   92   93   94   95
##  209  190  154  120  109   76   55   42   23   23   12   10    6    2
```

```
CES2020$age_group <- cut(CES2020$age,
  breaks = c(18, 30, 45, 65, Inf),
  labels = c("18-29", "30-44", "45-64", "65+"),
  right = FALSE
)

initial_model <- lm(voted_R ~ male + age_group + educ + abortion_position + CC20_302_ind + CC20_303_ind

summary(initial_model)
```

```
##
## Call:
## lm(formula = voted_R ~ male + age_group + educ + abortion_position +
##     CC20_302_ind + CC20_303_ind + factor(pid3) + factor(race) +
##     factor(ideo5), data = CES2020)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.22671 -0.15664  0.00602  0.10759  1.20059
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      0.402720   0.084397   4.772 1.83e-06 ***
## male                             0.006369   0.002697   2.361 0.018212 *
## age_group30-44                   0.021603   0.005088   4.246 2.18e-05 ***
## age_group45-64                   0.045394   0.004825   9.408  < 2e-16 ***
## age_group65+                     0.039257   0.005088   7.715 1.24e-14 ***
## educ4-year                      -0.021848   0.004655  -4.693 2.70e-06 ***
## educHigh school graduate         0.014472   0.004803   3.013 0.002590 **
## educNo HS                       -0.006649   0.011409  -0.583 0.560032
## educPost-grad                   -0.038321   0.005060  -7.573 3.73e-14 ***
## educSome college                -0.005306   0.004795  -1.107 0.268469
## abortion_position               -0.106542   0.002478 -43.003  < 2e-16 ***
## CC20_302_ind                     0.116685   0.001995  58.498  < 2e-16 ***
## CC20_303_ind                     0.024663   0.002105  11.716  < 2e-16 ***
## factor(pid3)Independent          0.202775   0.003706  54.717  < 2e-16 ***
## factor(pid3)Not sure             0.207943   0.012759  16.297  < 2e-16 ***
## factor(pid3)Other                0.251075   0.007527  33.356  < 2e-16 ***
## factor(pid3)Republican           0.417848   0.004804  86.974  < 2e-16 ***
## factor(race)Black               -0.085978   0.009326  -9.220  < 2e-16 ***
## factor(race)Hispanic             0.003362   0.009525   0.353 0.724142
## factor(race)Middle Eastern      -0.029285   0.036941  -0.793 0.427931
## factor(race)Native American      0.060315   0.017279   3.491 0.000482 ***
## factor(race)Other                0.073138   0.012847   5.693 1.26e-08 ***
## factor(race)Two or more races    0.052270   0.012360   4.229 2.35e-05 ***
## factor(race)White                0.035376   0.008319   4.252 2.12e-05 ***
## factor(ideo5)Conservative        0.091896   0.083824   1.096 0.272955
## factor(ideo5)Liberal            -0.254838   0.083902  -3.037 0.002388 **
## factor(ideo5)Moderate           -0.151819   0.083844  -1.811 0.070191 .
## factor(ideo5)Not sure           -0.111887   0.084253  -1.328 0.184190
## factor(ideo5)Very conservative   0.068707   0.083857   0.819 0.412598
## factor(ideo5)Very liberal       -0.259092   0.083929  -3.087 0.002023 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.2648 on 41173 degrees of freedom
##   (19797 observations deleted due to missingness)
## Multiple R-squared:  0.709,  Adjusted R-squared:  0.7088
## F-statistic:  3460 on 29 and 41173 DF,  p-value: < 2.2e-16
```

```r
# WE LATER USED A TABLE WE MADE
```

```r
fte_poll_data <- read.csv("data/president_polls_538.csv")

fte_poll_data <- fte_poll_data %>%
  filter(candidate_name %in% c("Kamala Harris", "Donald Trump"))

consolidated_data <- fte_poll_data %>%
  group_by(question_id) %>%
  summarize(
    Harris_pct = mean(ifelse(candidate_name == "Kamala Harris", pct, NA), na.rm = TRUE),
    Trump_pct = mean(ifelse(candidate_name == "Donald Trump", pct, NA), na.rm = TRUE)
  )

consolidated_data <- fte_poll_data %>%
  # Filter to include only question_ids with both candidates present
  filter(candidate_name %in% c("Kamala Harris", "Donald Trump")) %>%
  group_by(question_id) %>%
  # Keep only question_ids with both Harris and Trump rows
  filter(n_distinct(candidate_name) == 2) %>%
  # Summarize to calculate percentages and keep other columns
  summarize(
    Harris_pct = mean(pct[candidate_name == "Kamala Harris"], na.rm = TRUE),
    Trump_pct = mean(pct[candidate_name == "Donald Trump"], na.rm = TRUE),
    across(-c(candidate_name, pct), first)
  ) %>%
  ungroup()

consolidated_data$r_spread <- consolidated_data$Trump_pct - consolidated_data$Harris_pct

# fte_poll_quality <- consolidated_data$numeric_grade

# lm_numeric_grade <- lm(r_spread ~ fte_poll_quality , data=consolidated_data)
# summary(lm_numeric_grade)

# Install and load lubridate
if (!require(lubridate)) install.packages("lubridate")
library(lubridate)

# Clean, convert, and filter
consolidated_data <- consolidated_data %>%
  mutate(start_date = mdy(trimws(start_date))) %>% # Clean and convert to Date
  filter(start_date >= as.Date("2024-07-01")) # Filter for dates after June 30, 2024


consolidated_data <- consolidated_data %>%
  mutate(week_number = as.numeric(strftime(start_date, format = "%U")) + 1)

table(consolidated_data$week_number)
```

```
##
##  27  28  29  30  31  32  33  34  35  36  37  38  39  40  41
##   8  43  41 175  74 125  99 140  96 125 139 191  83  50   1
```

```
consolidated_data <- consolidated_data %>% filter(week_number >= 35)
consolidated_data
```

```
## # A tibble: 685 x 53
##    question_id Harris_pct Trump_pct poll_id pollster_id pollster     sponsor_ids
##          <int>      <dbl>     <dbl>   <int>       <int> <chr>        <chr>
##  1      207058         46        50   87917        1890 SoCal Strat~ 2152,2170
##  2      207085         49        47   87919        1886 Quantus Pol~ 2184
##  3      207186       47        45      87920         568 YouGov       352
##  4      207317       47.5      51.2    87935        1102 Emerson      960
##  5      207318       50.1      49      87944        1102 Emerson      960
##  6      207319       51.2      47.6    87936        1102 Emerson      960
##  7      207320       49.4      49.4    87938        1102 Emerson      960
##  8      207321       49.4      48.8    87937        1102 Emerson      960
##  9      207322       49        49.9    87943        1102 Emerson      960
## 10      207323       49.1      49.8    87939        1102 Emerson      960
## # i 675 more rows
## # i 46 more variables: sponsors <chr>, display_name <chr>,
## #   pollster_rating_id <int>, pollster_rating_name <chr>, numeric_grade <dbl>,
## #   pollscore <dbl>, methodology <chr>, transparency_score <dbl>, state <chr>,
## #   start_date <date>, end_date <chr>, sponsor_candidate_id <int>,
## #   sponsor_candidate <chr>, sponsor_candidate_party <chr>,
## #   endorsed_candidate_id <lgl>, endorsed_candidate_name <lgl>, ...
```

```r
consolidated_data_states <- consolidated_data %>%
  group_by(week_number, state) %>%
  summarize(
    r_spread = mean(r_spread, na.rm = TRUE),
    pct_Harris = mean(Harris_pct, na.rm = TRUE),
    pct_Trump = mean(Trump_pct, na.rm = TRUE)
  )

consolidated_data_states <- consolidated_data_states %>%
  group_by(week_number) %>%
  mutate(national_spread = if_else(state == "", r_spread, NA_real_))


consolidated_data_states <- consolidated_data_states %>%
  mutate(national_spread = if_else(state != "", first(national_spread[!is.na(national_spread)]), nationa

consolidated_data_states <- consolidated_data_states %>%
  mutate(state_relative_spread = if_else(state != "", r_spread - national_spread, NA_real_))



final_fte_poll_data <- consolidated_data_states %>%
  group_by(state) %>%
  summarize(
    r_spread = mean(r_spread, na.rm = TRUE),
    national_spread = mean(national_spread, na.rm = TRUE),
    state_relative_spread = mean(state_relative_spread, na.rm = TRUE)
  )


pred_fte_final_data <- final_fte_poll_data %>%
  mutate(fundementals_pred = 46.3)

pred_fte_final_data <- pred_fte_final_data %>%
  mutate(state_Trymo_share = fundementals_pred + state_relative_spread)

pred_fte_final_data
```

```
## # A tibble: 39 x 6
##    state        r_spread national_spread state_relative_spread fundementals_pred
##    <chr>           <dbl>           <dbl>                 <dbl>             <dbl>
##  1 ""              -3.00           -3.00                NaN                 46.3
##  2 "Alaska"        7.97            -3.39                 11.4               46.3
##  3 "Arizona"       0.868           -3.17                  4.04              46.3
##  4 "Arkansas"     15               -2.11                 17.1               46.3
##  5 "California"  -25.8             -2.70                -23.1               46.3
##  6 "Colorado"    -12.6             -2.99                 -9.64              46.3
##  7 "Connecticu~  -16               -3.47                -12.5               46.3
##  8 "Delaware"    -18.4             -3.62                -14.7               46.3
##  9 "Florida"       5.33            -3.17                  8.50              46.3
## 10 "Georgia"       0.866           -3.17                  4.03              46.3
## # i 29 more rows
## # i 1 more variable: state_Trymo_share <dbl>
```

```
# We then averged this out with CURRENT 538 polls
# We estimated state_relative_spread for states NOT in this data
```