Data Exploration: Contextual Influences

Yao Yu

November 11, 2021

In this Data Exploration assignment we will again be exploring the Nationscape dataset (Tausanovitch and Vavreck 2020), which was used in Reny and Newman's (2021) study of the effecs of the protests after George Floyd's killing.

Unlike previous assignments, however, you will be asked to take a bigger role in defining the research question and identifying the specific data that you would need to use. *This is practice for operationalizing questions of the type you will do for your research project.*

Throughout the assignment, we will provide a running example of how you might approach the tasks. For you own work, please do not use either this example or the Geroge Floyd protests.

Note: Because this assignment is a bit different, you are require to do all of the questions (although non-data science students can skip question 7). This is to ensure that you have enough material for you blog post.

If you have a question about any part of this assignment, please ask! Note that the actionable part of each question is **bolded**.

Developing a Research Question about Contextual Influences

Data Details:

- File Name: `vars_data.xlsx`
- Source: This file shows what variables are covered in each wave of the Nationscape Data Set (Tausanovitch and Vavreck 2020). We will be using data from the survey itself in other parts of the exercise, but which specific files and variables will be up to you! Therefore, we don't present them in depth here.

Variable Name	Variable Description
`Date`	The date of the wave of the Nationscape survey
`response_id`	This and all other variables are the names of variables in the Nationscape data; the cells are 1 if that variable was included in that week's survey and 0 otherwise

```
#L oad the data summarizing variable availability
NationscapeVars_1 <- read_xlsx('vars_data.xlsx',sheet = 1)
NationscapeVars_2 <- read_xlsx('vars_data.xlsx',sheet = 2)
```

Now let's get the data from two sheets into one data set.

```
NationscapeVars <- full_join(NationscapeVars_1,NationscapeVars_2) %>%
  replace(is.na(.),0)
```

```
## Joining, by = c("Date", "response_id", "start_date", "right_track", "economy_better", "interest", "r
```

Question 1

Contextual influences are all about the fleeting events that shift our attitudes and behavior. These can be something we personally experience, like encountering people on the street (Sands 2017) or voting at a school (Berger et al. 2008). But they can also be events we are exposed to by press coverage like Supreme Court decision (Tankard and Paluck 2017) or even emotions evoked by press coverage (as was experimentally modeled by Zeitzoff 2014). For this exercise we will think about events that people in a given state or across the country would plausibly have been exposed to via news coverage. **Think about events that happened between July 2019 and July 2020. Maybe this is something that made national news or maybe it was something that received a lot of coverage in your home state or region. Write down an example or two that you might be interested in considering. Use Google Trends to confirm that there was a spike in interest, as demonstrated by an increase in Google searches, in your event and include a screenshot or a hyperlink to your results.** Try entering a relevant search term and then using a "Custom time range" (one of the drop down options instead of the default ""Past 12 months") to make your visualization.

Example

The teaching staff were interested in thinking about the effects of the back-to-back mass shootings that took place on August 3 and 4, 2019 in El Paso, Texas and Dayton, Ohio, respectively. A Google Trends visualization confirms a dramatic spike in search interest for "shooting" in early August 2019.

The event that I'm considering is the California Kincade wildfires in October 2019.

Question 2

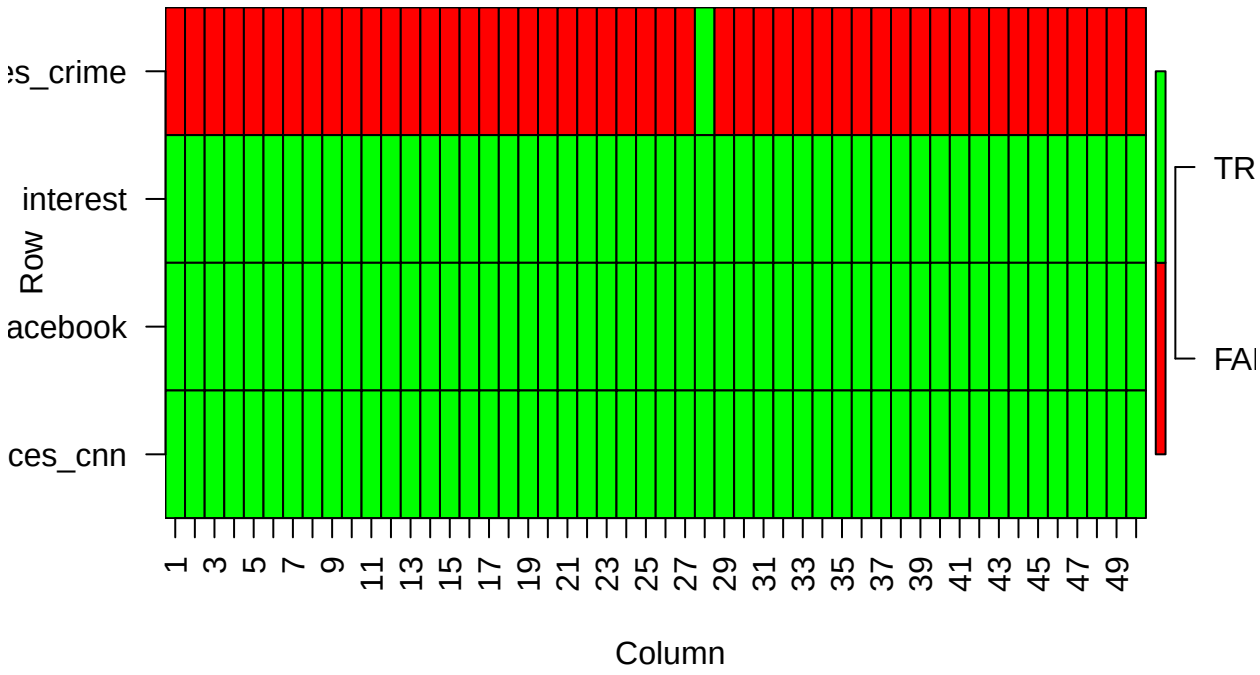
Think about some outcomes of interest to you that might have been affected by the contextual influence of the event that you chose. Look in the Nationscape data for variables that fit your outcomes or are reasonable proxies for those outcomes. The variable names in the data you have loaded are pretty informative, but use the full data folder you downloaded earlier to look in the codebooks for more complete descriptions of the variables and how they are measured. There is a codebook in each week's folder; you can look at any week's codebook to get a sense of the variables that are common across the survey waves. **Make sure that your variables are present in the data for the time period in which you want to look for contextual effects. Present these results in a plot.**

Example

We might think that news coverage of mass shootings would induce anxiety. Given some of readings last week, we wondered if respondents would become more interested in the news due to anxiety. There aren't any obvious proxies for anxiety, but there are several good variables to gauge political interest (the `interest` variable) and information seeking (the variables of the form `news_sources_xxxx`). Increases in these variables would be consistent with a noted effect of anxiety on the search for inforamtion. It also seems plausible that the shootings might elevate the priority people place on crime as an issue worthy of national attention. `extra_priorities_crime` seems appropriate here.

We can check the availability of these data using a heatmap.

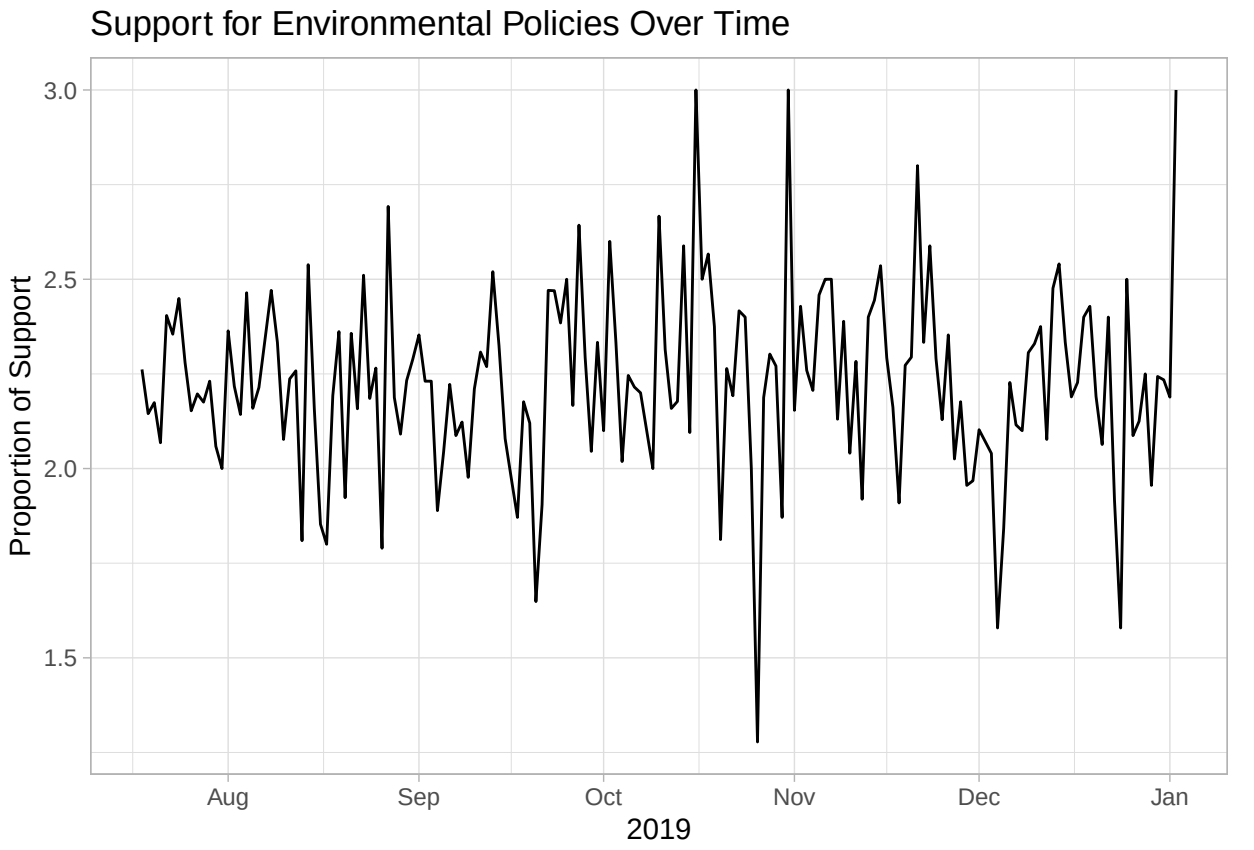
```
heat_data <- NationscapeVars %>% mutate(across(.cols = everything(), as.logical)) %>% select(c(extra_pri

plot(t(as.matrix(heat_data)), col = c('red','green'), las = 2)
```



This isn't the prettiest plot. But it quickly shows us that the political interest and news consumption variables have coverage throughout the data, whereas the crime question was only asked in one week.

```
# Collect the file names and select which ones we want
file_names_1 <- list.files("Nationscape-DataRelease_WeeklyMaterials_DTA/phase_1_v20200814/") %>%
  .[1:24]
file_names_2 <- list.files("Nationscape-DataRelease_WeeklyMaterials_DTA/phase_2_v20200814/") %>%
  .[1:24]

# Reading in all phase one weeks
# selecting the variables we want
phase_1 <- map_dfr(.x = file_names_1,
                   ~read_dta(file = str_c("Nationscape-DataRelease_WeeklyMaterials_DTA/phase_1_v2020081
                     select(start_date, environment, cap_carbon, green_new_deal,
                            age, gender,household_income, education))

phase_1_clean <- phase_1 %>%
  mutate(across(.cols = everything(), ~na_if(., 999))) %>%
  mutate(across(.cols = everything(), ~na_if(., 888)))

wildfire_eda <- phase_1_clean %>%
  mutate(date = as.Date(start_date),
         environment = ifelse(environment == 2, 0, environment),
         cap_carbon = ifelse(cap_carbon == 2, 0, cap_carbon),
         green_new_deal = ifelse(green_new_deal == 2, 0, green_new_deal),
         green_avg = environment + cap_carbon + green_new_deal) %>%
  drop_na(green_avg) %>%
  group_by(date) %>%
  summarize(prop_support_green_avg = sum(green_avg) / n(),
            .groups = "drop") %>%
```

```
  ggplot(aes(x = date, y = prop_support_green_avg)) +
  geom_line() +
  theme_light() +
  labs(
    title = "Support for Environmental Policies Over Time",
    x = "2019",
    y = "Proportion of Support"
  )

wildfire_eda
```




```
# ggplotly(wildfire_eda)
```

The variables I chose to evaluate are:

`environment`: Make a large-scale investment in technology to protect the environment (Agree or Disagree)

`cap_carbon`: Cap carbon emissions to combat climate change (Agree or Disagree)

`green_new_deal`: Enact a Green New Deal (Agree or Disagree)

I took an average of these three responses and created a new variable `green_avg`.

From the graph looking at a support for environment policies, we can see that aside from the outlier in early October, the next highest moment of support is at the end of October, exactly when the California Kincade wildfire starts.

Question 3

Based on what you have thought about and the data you have found, clearly state a specific research question and a hypothesis. Which channel (or channels) through which situational factors can affect political behavior does your hypothesis implicate? (In class, we talked about rational choice, priming, and emotional channels.) The research question should not be obvious ahead of time (although you should have a theoretical expectation or competing hypotheses); it should be be descriptive, correlational, or causal in nature; and it should be answerable with the data you have available. Make sure your research question is specific; don't confuse the research question with a broader, motivating question that might be used to get people interested in your topic.

Example

Since we don't have enough data to consider crime as a national policy priority, we will focus on information search. Our research question is "Were the early August 2019 mass shootings associated with increased interest in and consumption of political news?" This might fit under a broader motivational question of "Does news coverage of violent events lead to information seeking by causing anxiety?" but we don't have the ability to answer such a broad, causal question using only the Nationscape data.

Our hypothesis is that the August 2019 mass shootings were associated with increased political interest and news consumption, especially in the states where those shootings took place. This is an example of how situational context could influence political behavior through the emotional channel, although we cannot directly test the role of anxiety.

Question: Does news about wildfires make Americans more supportive of environmental policies?

Question 4

In academic and professional settings, peer feedback, especially early in a project, can force you to clarify your thinking and be an important source of ideas. It's also important to be able to give a quick 'elevator pitch' for your project (so named because it can be delivered in no more time than an elevator ride). We've randomly assigned you into groups to share your ideas so far and get your peers' input about sources you should read, different ways to approach your analysis, or questions about your hypotheses. **Get together in your groups, have everyone give their project's 'elevator pitch,' and gather feedback from your peers. Write at least one thing you took away from this session.** The next couple of questions will ask you to try to use the data to answer your research question and test your hypotheses, so be sure to brainstorm good ways to approach those tasks.

I learned that this attitude might also vary depending on partisanship and which news outlets people are getting their news from. For example, Republicans who are more likely to get their news from Fox might not see any shift in support towards environmental policies if they don't really cover the wildfires. Democrats, on the other hand, who watch networks like NBC and CNN that likely covered the wildfire, might be more supportive of environmental policies since the news networks covered the wildfire.

Question 5

No research project exists in a vacuum. As you get ready for your final projects, we want you to practice finding, summarizing, and citing related literature. **Identify at least two academic articles that might provide some background for your research question. List the complete source citations and include links to the articles you found.** Google Scholar (https://scholar.google.com/) or Hollis (https://hollis.harvard.edu/) are good places to look for these.

Example

O'Brien and Taku (2022) find in an experiment on US undergraduates that reading news coverage about mass shootings increases anxiety. Joslyn and Haider-Markel (2018) show using survey evidence that people

who experienced higher anxiety in the wake of the 2016 Orlando shooting changed their policy beliefs and perceptions of institutions. Our research examines the middle step in this causal chain: information search. Were the August 2019 mass shootings associated with increased interest in and consumption of political news?

(Your response can just be a list of articles, but feel free to expand on it as we did above if you so choose.)

References:

Joslyn, Mark R., and Donald P. Haider-Markel. "The Direct and Moderating Effects of Mass Shooting Anxiety on Political and Policy Attitudes." Research & Politics, (July 2018). https://doi.org/10.1177/2053168018794060.

O'Brien, Colin, and Taku, Kanako. "Alpha and beta changes in anxiety in response to mass shooting related information." Personality and Individual Differences, Volume 186, Part A, (2022). https://doi.org/10.1016/j.paid.2021.111326.

Hazlett and Mildenberger (2020) find that "wildfires increased support for costly, climate-related ballot measures by 5 to 6 percentage points for those living within 5 kilometers of a recent wildfire, decaying to near zero beyond a distance of 15 kilometers". However, these effects are only seen in Democratic areas and not Republican areas, supporting my prior hypothesis. Crow et al. (2017) evaluate how news media coverage on two wildfires in Colorado helped shift policy narrative on mitigating wildfire risk through policy change.

HAZLETT, C., & MILDENBERGER, M. (2020). Wildfire Exposure Increases Pro-Environment Voting within Democratic but Not Republican Areas. American Political Science Review, 114(4), 1359-1365. doi:10.1017/S0003055420000441

Crow, D. A., Berggren, J., Lawhon, L. A., Koebele, E. A., Kroepsch, A., & Huda, J. (2017). Local media coverage of wildfire disasters: An analysis of problems and solutions in policy narratives. Environment and Planning C: Politics and Space, 35(5), 849–871. https://doi.org/10.1177/0263774X16667302

Question 6

Read in the data from the weeks surrounding your event of interest and test your hypothesis. This can be something straightforward like a difference-in-means or you can plot a visualization of the data. Just take one of the approaches we have used in class before to get an initial sense for if the data provide evidence of the contextual effects you theorized. Note that you might have to do a fair bit of data cleaning in order to do this. Pay particular attention to how missing data are coded.

Example

First we will need to load and compile the data for the several weeks surrounding the mass shootings we are investigating.

```
# load the weekly survey files

Jul18 <- read_dta('ns20190718.dta') %>%
  remove_all_labels() # we need to remove the labels in order for these files to be joined together
Jul25 <- read_dta('ns20190725.dta')%>%
  remove_all_labels()
Aug01 <- read_dta('ns20190801.dta') %>%
  remove_all_labels()
Aug08 <- read_dta('ns20190808.dta')%>%
  remove_all_labels()
Aug15 <- read_dta('ns20190815.dta')%>%
  remove_all_labels()
```

```
# join them all together
Summer2019 <- full_join(Jul18,Jul25) %>%
  full_join(., Aug01) %>%
  full_join(., Aug08) %>%
  full_join(., Aug15)
```

```
## Joining, by = c("response_id", "start_date", "right_track", "economy_better", "interest", "registrati
## Joining, by = c("response_id", "start_date", "right_track", "economy_better", "interest", "registrati
## Joining, by = c("response_id", "start_date", "right_track", "economy_better", "interest", "registrati
## Joining, by = c("response_id", "start_date", "right_track", "economy_better", "interest", "registrati
```

```
# recode NAs
Summer2019 <- Summer2019 %>%
  mutate(across(.cols = everything(), ~na_if(., 999))) %>%
  mutate(across(.cols = everything(), ~na_if(., 888)))
```

It might be helpful to create a variable indicating whether observations took place after your event of interest. We do this using the below code.

```
# create an indicator variable for surveys administered after the mass shootings
Summer2019 <- Summer2019 %>%
  mutate(treated = if_else(start_date > as.Date('2019-08-04'), TRUE, FALSE))
```

As a first cut, we can try a difference in means. Don't forget to check the effect size.

```
difference_in_means(interest ~ treated, data = Summer2019 %>% filter(as.Date('2019-07-31') < start_date)
```

```
## Design:  Standard
##            Estimate Std. Error  t value   Pr(>|t|)   CI Lower    CI Upper
## treated -0.06314646 0.02083279 -3.03111 0.00244478 -0.1039845 -0.02230844
##               DF
## treated 7592.752
```

```
cohen.d(interest ~ treated, data = Summer2019 %>% filter(as.Date('2019-07-31') < start_date) %>% filter
```

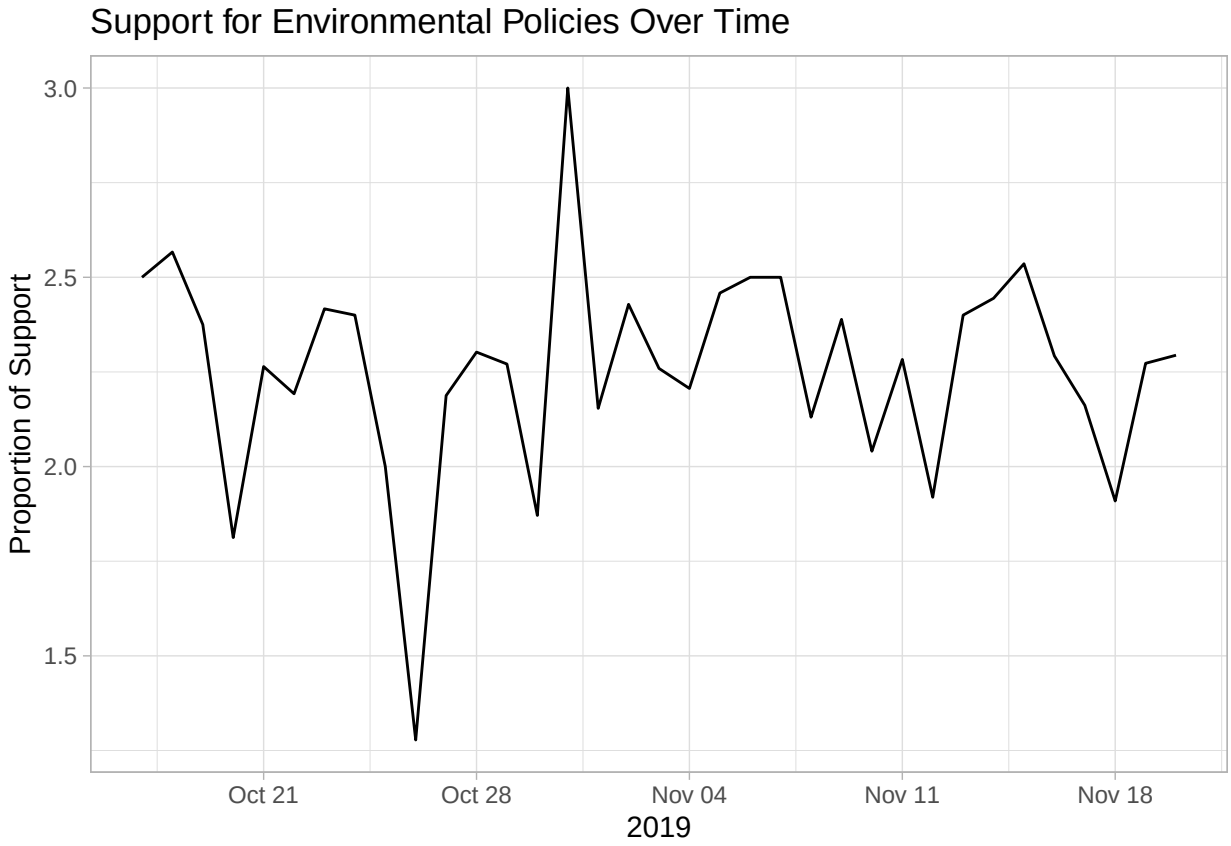
```
## Warning in cohen.d.formula(interest ~ treated, data = Summer2019 %>%
## filter(as.Date("2019-07-31") < : Cohercing rhs of formula to factor
```

```
##
## Cohen's d
##
## d estimate: 0.06951593 (negligible)
## 95 percent confidence interval:
##      lower      upper
## 0.02453589 0.11449598
```

Even though there is a statistically significant difference in means, the effect size is negligible. And becuase the Nationscape survey asks different groups of people the same questions every week, some of the change in measured opinion is probably just due to sampling error instead of significant shifts in the population. At least based on this test, there doesn't seem to be evidence supporting our hypothesis. That's ok! It could be because this is a pretty crude test of our theory, or we might just have been wrong. It certainly warrants more investigation, but disconfirming hypotheses is an important part of how science moves forward.

```
# load the weekly survey files
wildfire_weeks <- file_names_1 %>% .[14:18]

wildfire_data <- map_dfr(.x = wildfire_weeks,
                  ~read_dta(file = str_c("Nationscape-DataRelease_WeeklyMaterials_DTA/phase_1_v2020081
```

```
                    select(start_date, environment, cap_carbon, green_new_deal,
                           age, gender, household_income, education))

# recode NAs
wildfire_data <- wildfire_data %>%
  mutate(across(.cols = everything(), ~na_if(., 999))) %>%
  mutate(across(.cols = everything(), ~na_if(., 888))) %>%
  mutate(date = as.Date(start_date),
         treatment = (date > as.Date("2019-10-23"))) %>%
  select(-start_date) %>%
  mutate_at(vars(environment, cap_carbon, green_new_deal,
                 gender, household_income, education),
            ~as.numeric(.)) %>%
  mutate(environment = ifelse(environment == 2, 0, environment),
         cap_carbon = ifelse(cap_carbon == 2, 0, cap_carbon),
         green_new_deal = ifelse(green_new_deal == 2, 0, green_new_deal),
         green_avg = environment + cap_carbon + green_new_deal)

wildfire_data %>%
  drop_na(green_avg) %>%
  group_by(date) %>%
  summarize(prop_support_green_avg = sum(green_avg) / n(),
            .groups = "drop") %>%
  ggplot(aes(x = date, y = prop_support_green_avg)) +
  geom_line() +
  theme_light() +
  labs(
    title = "Support for Environmental Policies Over Time",
    x = "2019",
    y = "Proportion of Support"
  )
```




```
difference_in_means(green_avg ~ treatment, data = wildfire_data)
```

```
## Design:  Standard
##              Estimate Std. Error    t value  Pr(>|t|)   CI Lower   CI Upper
## treatment -0.05516903 0.07373277 -0.7482294 0.4547319 -0.2000916 0.08975354
##                  DF
## treatment 428.8646
```

```
cohen.d(green_avg ~ treatment, data = wildfire_data)
```

```
## Warning in cohen.d.formula(green_avg ~ treatment, data = wildfire_data):
## Cohercing rhs of formula to factor
```

```
##
## Cohen's d
##
## d estimate: 0.05124252 (negligible)
## 95 percent confidence interval:
##       lower       upper
## -0.08491321  0.18739825
```

From the difference in means and Choen's d calculations, we can see that there appears to be a negligible effect that the Kincade Fire drastically changed attitudes towards a green new deal. It might be interesting to plot the dates of every major wildfire in 2019 to see if it's just a very rapid spike up and then returns to normal or if this spike is just a coincidence.

Question 7: DATA SCIENCE QUESTION

Extend your work from the previous question to consider other factors, like the possibility of heterogenous treatment effects, confounding variables, or use a more sophisticated approach to statistical inference, like regression discontinuity in time.

```
library(mice)
```

```
##
## Attaching package: 'mice'
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```
# Imputing missing data
wildfire_data_imputed <- wildfire_data %>%
  mice(printFlag = FALSE) %>%
  complete() %>%
  as_tibble()
```

```
## Warning: Number of logged events: 20
```

```
fit_raw <- lm(green_avg ~ treatment + age + gender + household_income + education,
              data = wildfire_data, na.action = na.omit)

fit_imputed <- lm(green_avg ~ treatment + age + gender + household_income + education,
                  data = wildfire_data_imputed)

stargazer(fit_raw, fit_imputed, header = FALSE,
          dep.var.labels = c("Support for Environmental Policies"),
          covariate.labels = c("Treatment Wildfire", "Age", "Gender", "HH Income", "Education"),
          title = "Support for Environmental Policies as a Function of Treatment Wildfire and controls")
```

From this model, we can see that the treatment variable of the wildfire is actually associated with a slight decrease in support for a environmental policies on average while controlling for other demographic variables. Both models using data with and without imputation for missing vales both show very similar results. Overall, it seems to suggest that there is no significant effect of wildfires and more support for environmental policies on average. It should also be noted that there were far too many missing variables to be using imputation in this example, so the model with the imputed data should not be used for any serious reflections (purely pedagogical).

Table 2: Support for Environmental Policies as a Function of Treatment Wildfire and controls

	Dependent variable:	
	Support for Environmental Policies	
	(1)	(2)
Treatment Wildfire	−0.068	0.001
	(0.075)	(0.014)
Age	−0.010***	−0.007***
	(0.002)	(0.0003)
Gender	−0.228***	−0.225***
	(0.063)	(0.011)
HH Income	−0.012**	−0.004***
	(0.005)	(0.001)
Education	0.053***	0.026***
	(0.016)	(0.003)
Constant	2.826***	2.707***
	(0.169)	(0.029)
Observations	1,181	32,303
R^2	0.036	0.024
Adjusted R^2	0.032	0.024
Residual Std. Error	1.056 (df = 1175)	1.004 (df = 32297)
F Statistic	8.691*** (df = 5; 1175)	158.471*** (df = 5; 32297)
Note:		*p<0.1; **p<0.05; ***p<0.01