

Open in app ↗

Sign up

Sign In



You have 2 free member-only stories left this month. [Sign up](#) for Medium and get an extra one.

★ Member-only story

Uncovering the Limitations of Traditional DiD Method

Dealing with Multiple Time Periods and Staggered Treatment Timing



Nazlı Alagöz · [Follow](#)

Published in Towards Data Science

11 min read · Feb 21



Listen



Share



Cover image, generated by the author using [NightCafé](#)

Difference-in-Differences (DiD) is a popular statistical method for estimating the causal impact of interventions in observational studies by comparing the outcome

difference between two groups before and after treatment. Most DiD guides focus solely on the canonical DiD setup where there are only two periods and two groups (treated and nontreated).

However, in many real-world applications of DiD, there are multiple time periods and variations in the treatment timing. **Recent studies on DiD show that DiD may give significantly misleading estimates of the treatment effects in these situations.** In certain scenarios, the treatment effect estimates may show an opposite sign compared to the actual treatment effect.

In this article, I'll discuss an important issue that can occur in the canonical DiD setup when staggered treatment timing and multiple time periods are present. I'll also present solutions to address this issue. It's important to note that while I will focus solely on this one issue in DiD, for a more comprehensive overview of other potential challenges, you can refer to [my previous article](#). Additionally, I'll provide further resources at the end of this article for those who wish to delve deeper into DiD issues.

Lockdowns and Music Consumption Example

To provide an example, let's consider a hypothetical scenario. Imagine that we run a music streaming service that operates in various countries. We want to investigate the effects of Covid-19 lockdowns on music consumption in these countries. By examining the impact of reduced mobility, we can gain insights into whether listening to music is associated with certain activities, such as commuting, as opposed to working from home.

Since we cannot manipulate the implementation of lockdowns, it's not feasible to conduct an A/B test to examine their effects. Therefore, we must rely on observational data. In this case, we utilize the varying times that lockdowns were imposed in the countries included in our dataset. In this example, the treatment is the implementation of a lockdown. For this toy example, I simulated a dataset, the details of which can be found [in my previous article](#) and [this Gist](#). All the analysis code is also available in [this Gist](#).

```
rm(list = ls())
library(data.table) # Fast data frames
library(fastDummies) # Create dummy variables
library(fixest) # Fixed-effects regression
library(kableExtra) # Make nice tables
library(bacondecomp) # Goodman-Bacon Decomposition
library(did) # Difference-in-differences package by Callaway & Sant'Anna
source('sim_data.R') # Import data simulation functions and utilities

data <- sim_data() # Simulate the dataset

# EDA and Analysis -----
select_cols <- c('unit', 'period', 'cohort_period', 'treat', 'hrs_listened')

kable(head(data[, ..select_cols]), 'simple')
```

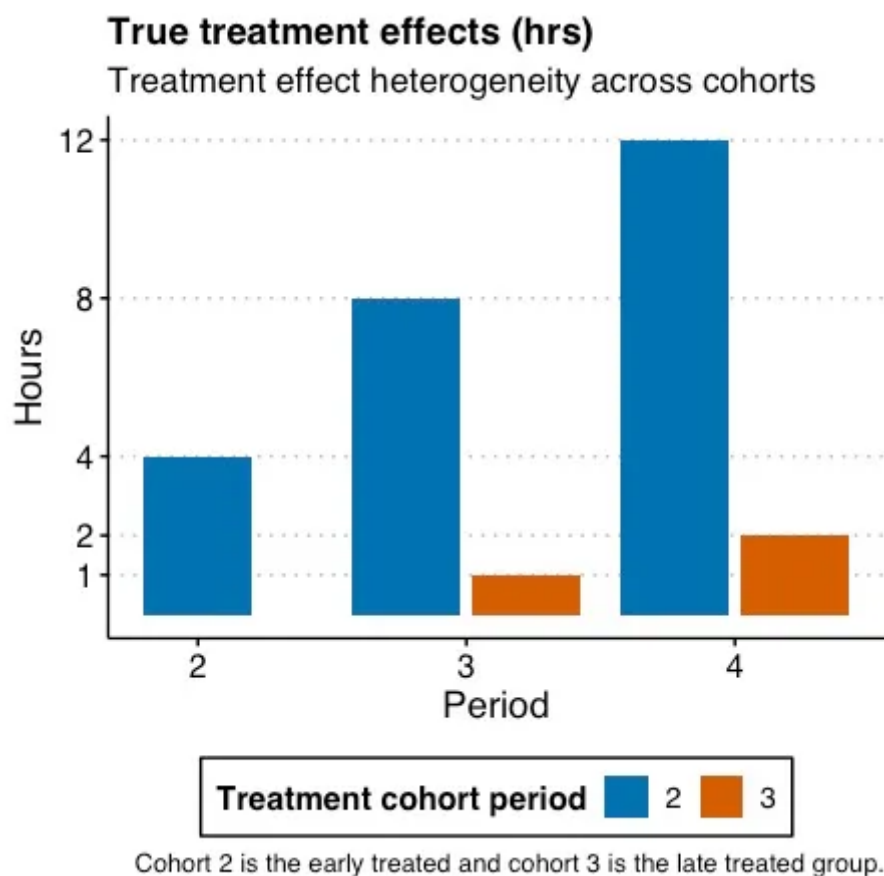
unit	period	cohort_period	treat	hrs_listened
1	0	2	0	75.11986
1	1	2	0	79.75711
1	2	2	1	93.40679
1	3	2	1	74.52117
1	4	2	1	106.00396
2	0	2	0	100.32866

A snapshot of the data for selected columns, image by the author.

We have data on 1000 units or customers for each period that they are observed. `cohort_period` indicates in which period a unit is treated and therefore which treatment cohort they belong to. A unit is treated (`treat = 1`) when `cohort_period` \geq `period`. `hrs_listened` is the outcome of interest indicating the total music consumption (hours). There are two cohorts in the dataset: the early-treated cohort and the late-treated cohort. The early cohort is treated in period 2 and the late cohort is treated in period 3. In total, there are five periods, starting from period 0 and ending with period 4.

As we have an observational dataset, where treatment is not randomly assigned, a simple difference-in-means approach cannot be employed to estimate the treatment effect. Instead, we aim to distinguish the treatment effects from customer- and season-related factors, and to achieve this, we utilize a DiD framework.

Just before we get into DiD, although this is not possible in real-life applications, in this simulated dataset I know the true treatment effects for each treatment cohort and period. Below, I visualize these as they will be necessary to evaluate the estimated treatment effects in the next steps.



True treatment effects per cohort and period, image by the author.

As seen in this graph, the true treatment effects for both cohorts are positive for all the periods. The treatment effect increases over time for both cohorts. However, overall the treatment effect is greater and increases quite a lot over the treated periods for cohort 2 compared to cohort 3. Thus, there is heterogeneity in treatment effects across treatment cohorts and periods.

Canonical DiD

Suppose that we are not aware that having multiple periods and staggered treatment can lead to misleading estimates when using the canonical DiD method. Thus, naively, we decide to use the canonical DiD setup to account for seasonality in the

listening patterns and customer-specific effects in our observational dataset. We use a DiD setup like this [1]:

$$Y_{it} = \alpha_i + \gamma_t + \beta^{dd} D_{it} + \epsilon_{it}$$

Canonical DiD, image by the author.

Y_{it} is the outcome of interest. α_i is the unit-fixed effects that controls for time-constant unit characteristics. γ_t is the time-fixed effects that controls for time trends or seasonality. D_{it} is the treatment dummy at time t for unit i . ϵ_{it} is the random error. The coefficient of interest β^{dd} indicates the treatment effect.

We estimate the treatment effect using canonical DiD setup in R:

```
formula <- as.formula('hrs_listened ~ treat')
canonical_did <- feols(formula,
                        data = data, panel.id = "unit",
                        fixef = c("unit", "period"), cluster = "unit")
summary(canonical_did)
```

```
OLS estimation, Dep. Var.: hrs_listened
Observations: 5,000
Fixed-effects: unit: 1,000, period: 5
Standard-errors: Clustered (unit)
      Estimate Std. Error   t value Pr(>|t|)
treat -0.472155   0.678929  -0.695441  0.48694
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 9.1397      Adj. R2: 0.086608
                Within R2: 1.066e-4
```

Canonical DiD estimates, image by the author.

The estimated treatment effect is -0.47 (though not statistically significant)! But how can this be when we have positive (and generally big) treatment effects for each treated cohort and period? **The reason for this is that DiD estimator is a weighted average of all possible two-group/two-period DiD estimators in the data [2].** In other words, the canonical DiD estimate can be decomposed into a weighted average two-group x two-period treatment estimates. This is called **Goodman-Bacon**

decomposition. Let's get the weights and estimates used to calculate the canonical DiD estimates using 'bacondecomp' package [2]:

```
# Goodman-Bacon Decomposition
bacon_decomp <- bacon(formula, data, id_var="unit", time_var='period', quietly
```

treated	untreated	estimate	weight	type
3	2	-4.531235	0.5	Later vs Earlier Treated
2	3	3.586925	0.5	Earlier vs Later Treated

Goodman-Bacon Decomposition, image by the author.

In the image above, I show the Goodman-Bacon decomposition for our canonical DiD estimate. Indeed, if we sum these estimates up with their respective weights we get exactly the treatment effect estimated by the canonical DiD: $0.5 \times -4.53 + 0.5 \times 3.59 = -0.47$. Since we have 2 groups and 2 periods where the treatment indicator changes we have 2 comparisons.

Let's examine this table in detail to see what the issue is. We start with the second comparison where the treatment estimate is 3.59. Here, the control group ('untreated') is the late-treated group, cohort 3. The 'treated' group is the early-treated group, cohort 2. I am putting the 'treated' and 'untreated' in quotes because as you might remember all groups in our dataset are eventually treated. 'treated' and 'untreated', here, rather refer to groups used as **treatment** and **control** groups by the canonical DiD estimator.

Let's move on to the first comparison highlighted in red where the estimate is -4.53. **Here, the early-treated group (cohort 2) is used as the control group for the late-treated group by the canonical estimator!** This comparison does not make much sense. Still, everything would work fine if the treatment effects were constant across cohorts and periods. In this application, as well as in many others, this is not the case. Since the treatment effects are higher and dynamically increasing for the early-treated group, in comparison it seems as if the treatment effect for the late-treated group is negative! **The comparisons where the early-treated group is used as the control group for the late-treated groups are called forbidden comparisons [2].**

How to address this problem?

The main solution to this problem is to not restrict the treatment effect to a single estimate and to carefully choose the control group. In the next steps, we are going to see how to do this. First, I am going to demonstrate how to address this problem without a particular DiD package. Later, I will use an R package that's specifically designed for DiD with multiple time periods.

Addressing the issue without relying on a specific DiD package

First, let's address this problem without relying on a specific DiD package. I know that there are two things that I need to do to get a good estimate of the treatment effect: (1) **not restrict the treatment effect estimation to a single coefficient**, (2) **make sure that I have a good control group** [3][4].

In essence, it is necessary to account for variations in treatment effects across different cohorts and periods. Moreover, it's crucial to ensure that there are untreated observations available for each period being assessed for treatment effects. This is because using treated observations as controls can lead to significantly misleading results, as previously noted.

As you remember, I only have treated groups in this dataset: the cohort treated in period 2 (early-treated) and the cohort treated in period 3 (late-treated). Clearly, I cannot estimate any treatment effects for the late-treated cohort as there are no not-yet-treated observations to use as controls when they are treated.

There is more hope for the early-treated group as when they were treated the late-treated group was not yet treated. This means that we can certainly estimate a treatment effect in period 2 for the early treated cohort. However, we cannot estimate any treatment effects for further periods as there are no untreated observations from period 3 onward. This is why we will drop periods that have no untreated units. Let's code this up.

```
# Drop periods that have no untreated units  
data <- data[period < 3]
```

Now, it is time to estimate the only treatment effect that we can estimate. We will only estimate a treatment effect for the early treated group in period 2.

```

# Create dummy variables
data <- data %>%
  dummy_cols(select_columns = c("cohort_period", "period"))

interact_covs <- 'cohort_period_2:period_2'

# Regression
formula <- as.formula(paste0('hrs_listened ~ ', interact_covs))
model <- feols(formula,
  data = data, panel.id = "unit",
  fixef = c("unit", "period"), cluster = "unit")
summary(model)

```

```

OLS estimation, Dep. Var.: hrs_listened
Observations: 3,000
Fixed-effects: unit: 1,000, period: 3
Standard-errors: Clustered (unit)

```

	Estimate	Std. Error	t value	Pr(> t)
cohort_period_2:period_2	3.58692	0.754562	4.75365	2.2908e-06 ***

```

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 8.1097      Adj. R2: 0.015167
                Within R2: 0.010741

```

Regression results after accounting for staggered treatment, image by the author.

As seen from the results above, the estimated treatment effect is much more sensible this time: 3.6. This means that the lockdown resulted in a 3.6 hours increase in music consumption for this cohort in this period. The point estimate is not exactly equal to the true treatment effect (approx. 4 hrs) because of the noise in the data.

Addressing the issue using 'did' package

As an alternative to doing everything manually let's use the [did_package](#) by [Callaway and Sant'Anna](#) [3]. What we need to do is to use `att_gt` function to estimate the treatment effects at cohort- and period-level using the right control group. Below, the code is given. One thing here that is important is that you need to specify the `control_group` as 'notyettreated' because this function by default tries to find an untreated group to use as the control group.


```
# did package
out <- att_gt(ymname = "hrs_listened",
             gname = "cohort_period",
             idname = "unit",
             tname = "period",
             xformula = ~1,
             data = data,
             est_method = "reg",
             control_group = 'notyettreated'
)
out
```

Group-Time Average Treatment Effects:

Group	Time	ATT(g,t)	Std. Error	[95% Simult. Conf. Band]
2	1	0.3796	0.9287	-1.5529 2.3120
2	2	3.3971	0.8918	1.5414 5.2529 *

Signif. codes: `*' confidence band does not cover 0

P-value for pre-test of parallel trends assumption: 0.67471

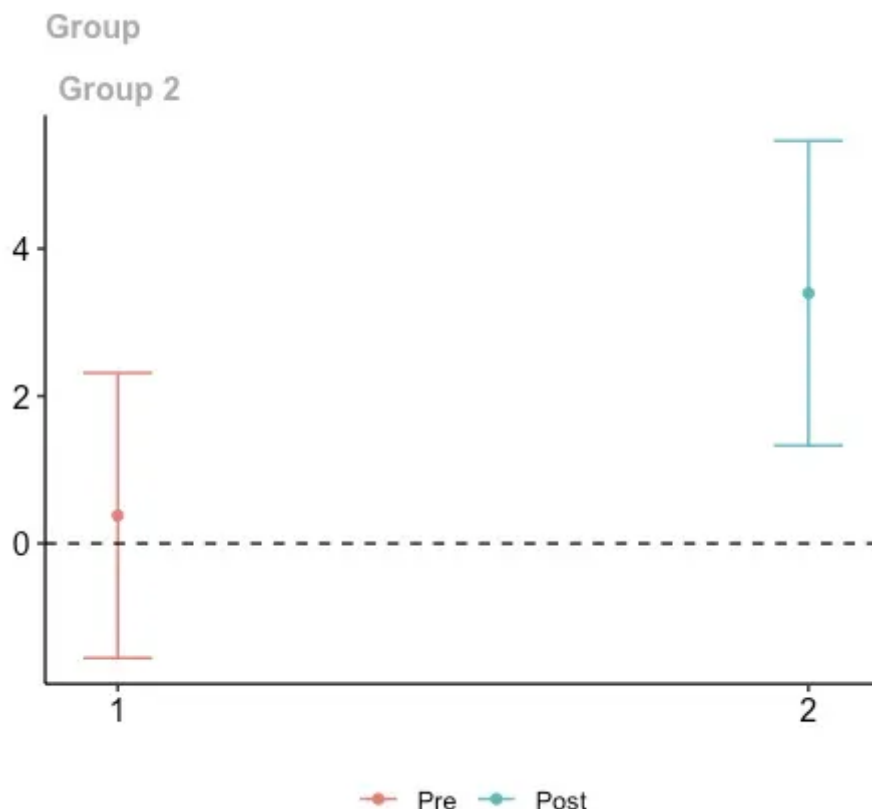
Control Group: Not Yet Treated, Anticipation Periods: 0

Estimation Method: Outcome Regression

did package results, image by author.

`att_gt` function estimates cohort-period-specific treatment effects, i.e., $ATT(g, t)$. We are interested in the treatment effect for cohort 2 in period 2 and this is given as 3.4 hrs. This is almost the same as the treatment effects we estimated without relying on this package. There can be some differences in the estimates due to the exact process of estimation. The 'treatment effects' for the period before the treatment is also reported in the first line and this is not statistically significant as expected because in this case, I know that there are no systemic changes to the outcome variable prior to the intervention. We can also graph these results with a single line of code:

```
ggdid(out) # graph the results
```



Visualizing did package results, image by author.

This graph is also useful to check the overtime trends pre- and post-treatment. This type of visualization comes especially handy when estimating treatment effects for many cohorts and periods, though in this case I only have one cohort and two periods for which I can make estimations.

Using this package has additional advantages compared to my manual method relevant here. As long as you specify the required variables for the `att_gt` function you don't need to do much else. You don't even need to drop the periods without untreated observations as this package already takes this into account and estimates effects only for periods where there is a valid control group. Another nice advantage is that the package as default reports uniform confidence intervals that accounts for multiple hypothesis testing (and this results in wider confidence bands due to a higher critical value used). The differences in the exact estimates between the two methods are due to the exact estimation method is not exactly the same.

Coming back to our example, we see that the lockdown has a positive effect on music consumption (though we were able to estimate this only for one cohort in one period). This indicates that actually music listening is complementary to staying at home.

Conclusions

Here are the key takeaways from this article:

- Canonical DiD can lead to misleading estimates in applications where there are multiple periods and variations in treatment timing.
- To prevent this issue, one can use estimators that account for multiple periods and variations in treatment timing.
- These estimators are suitable for staggered treatment contexts because they allow for flexible treatment effects and only estimate treatment effects for periods in which there is a valid control group.
- This is not the only issue that might come up when doing DiD analysis. For other issues, [see my previous post on event studies](#).

References

- [1] Angrist, J. D., & Pischke, J. S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.
- [2] Goodman-Bacon, Andrew. (2021) “[Difference-in-differences with variation in treatment timing](#).” *Journal of Econometrics* 225.2: 254–277.
- [3] Callaway, B., & Sant’Anna, P. H. (2021). [Difference-in-differences with multiple time periods](#). *Journal of Econometrics*, 225(2), 200–230.
- [4] Wooldridge, J. M. (2021). [Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators](#). Available at SSRN 3906345.

Other Useful Resources on DiD

Videos:

[Pedro H.C. Sant’Anna — “Difference-in-Differences with Multiple Time Periods”](#)

Andrew Goodman-Bacon “Difference-in-Differences with Variation in Treatment Timing”

A nice paper that summarizes recent DiD literature:

Roth, J., Sant’Anna, P. H., Bilinski, A., & Poe, J. (2022). What’s trending in difference-in-differences? A synthesis of the recent econometrics literature. *arXiv preprint arXiv:2201.01194*.

A compilation of many DiD-related resources.

Thank you for reading!

If you liked the post and would like to see more of my articles consider following me.

***Disclaimer:** I write to learn so it might be that you spot an error in the article or code. If you do so, please let me know.*

[Data Science](#)[Causal Inference](#)[Causal Data Science](#)[Statistics](#)[Hands On Tutorials](#)[Follow](#)

Written by Nazlı Alagöz

133 Followers · Writer for Towards Data Science

PhD candidate in Quantitative Marketing. I write on causal inference, data science and work related topics.
www.linkedin.com/in/nazli-m-alagoz

More from Nazlı Alagöz and Towards Data Science



Nazlı Alagöz in Towards Data Science

Crossing the Bridge: A Comparison of Data Science in Academia and Industry

A Ph.D. student's exploration of the surprising parallels between academic and industrial data science

🌟 · 8 min read · May 29

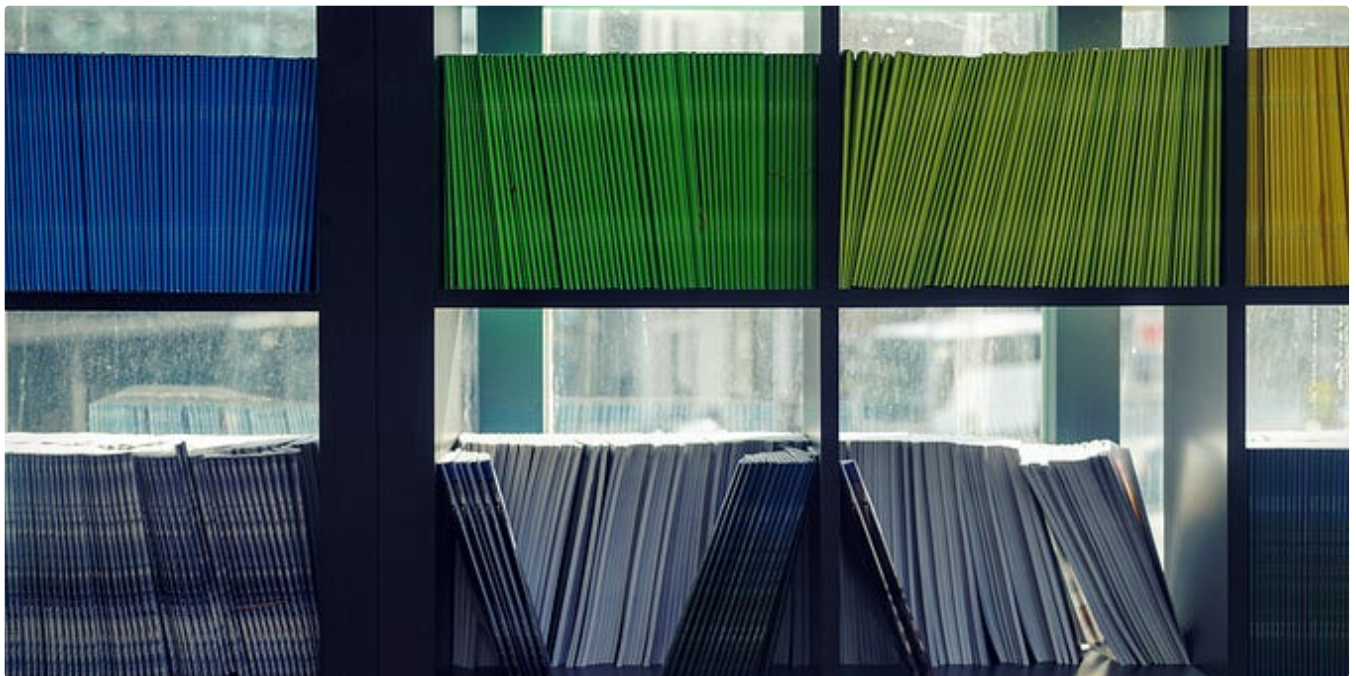


225



2





Jacob Marks, Ph.D. in Towards Data Science

How I Turned My Company's Docs into a Searchable Database with OpenAI

And how you can do the same with your docs

15 min read · Apr 25



3.6K



46



Leonie Monigatti in Towards Data Science

Getting Started with LangChain: A Beginner's Guide to Building LLM-Powered Applications

A LangChain tutorial to build anything with large language models in Python

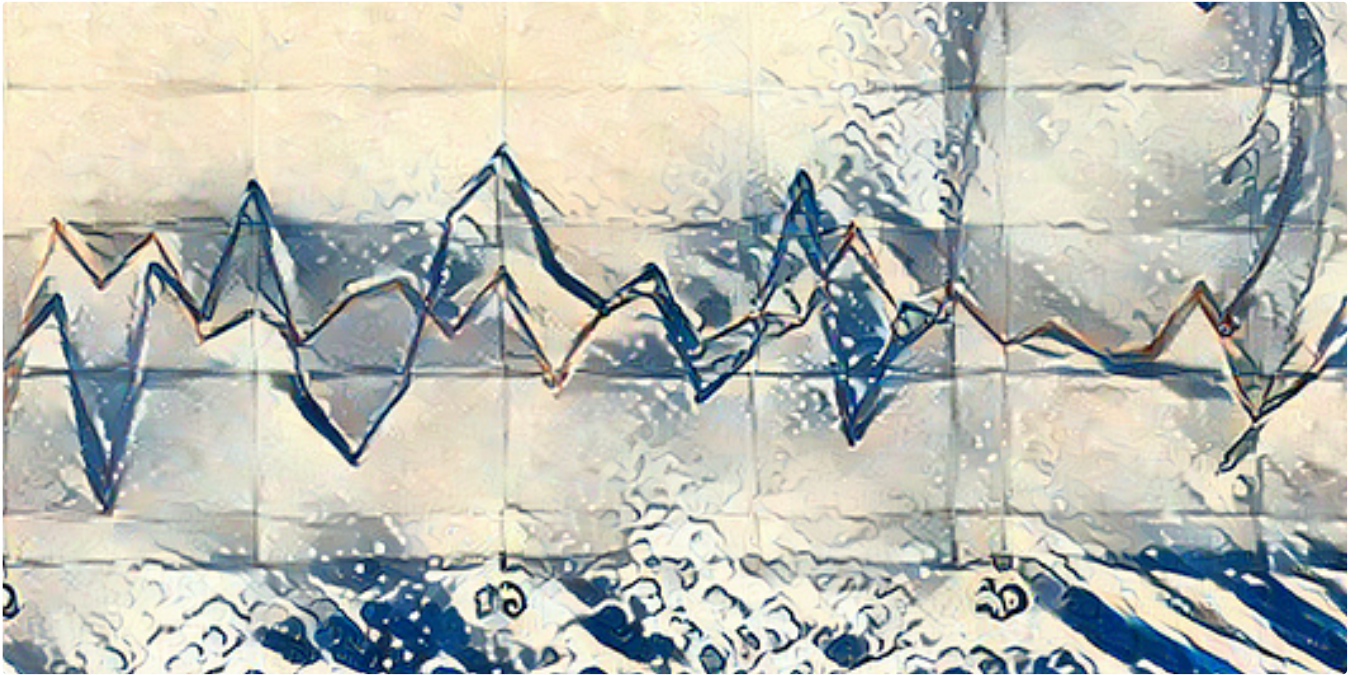
🌟 · 12 min read · Apr 25



2.6K



20



Nazlı Alagöz in Towards Data Science

SynthDiD 101: A Beginner's Guide to Synthetic Difference-in-Differences

On the method's advantages and disadvantages, demonstrated with the synthdid package in R

🌟 · 8 min read · Apr 26



171



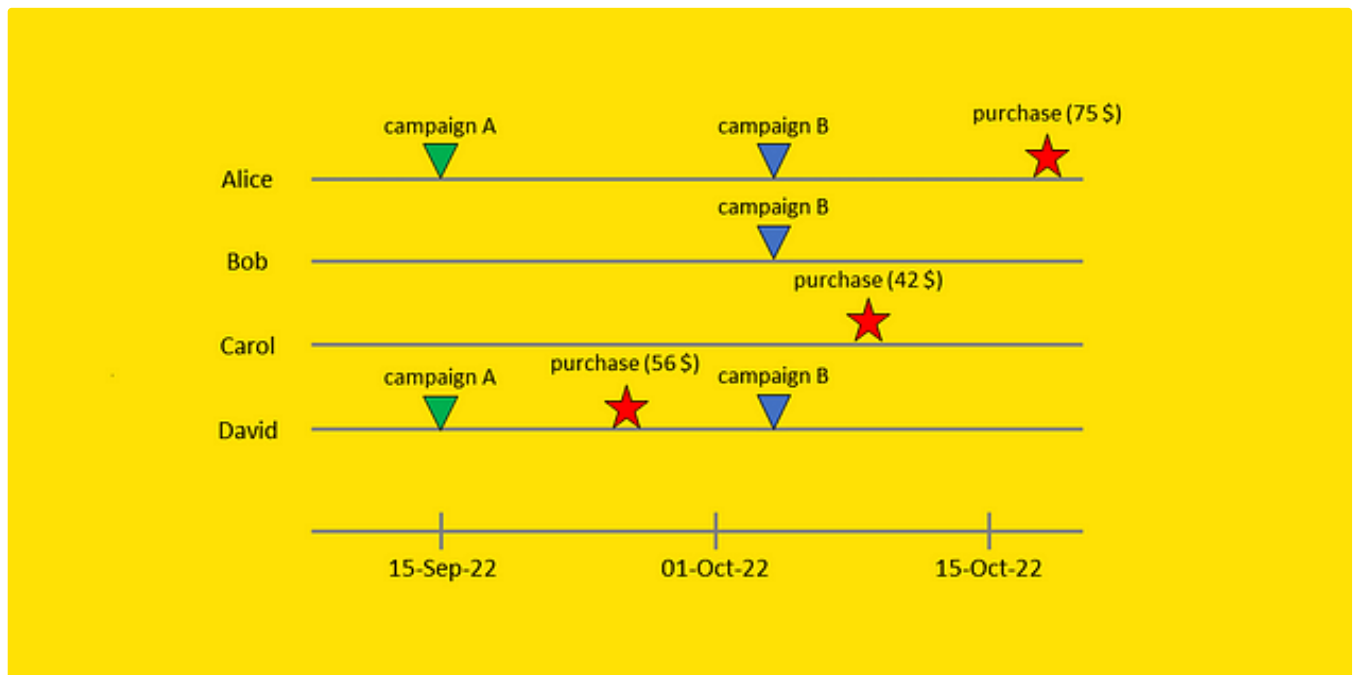
1



See all from Nazlı Alagöz

See all from Towards Data Science

Recommended from Medium



 Samuele Mazzanti in Towards Data Science

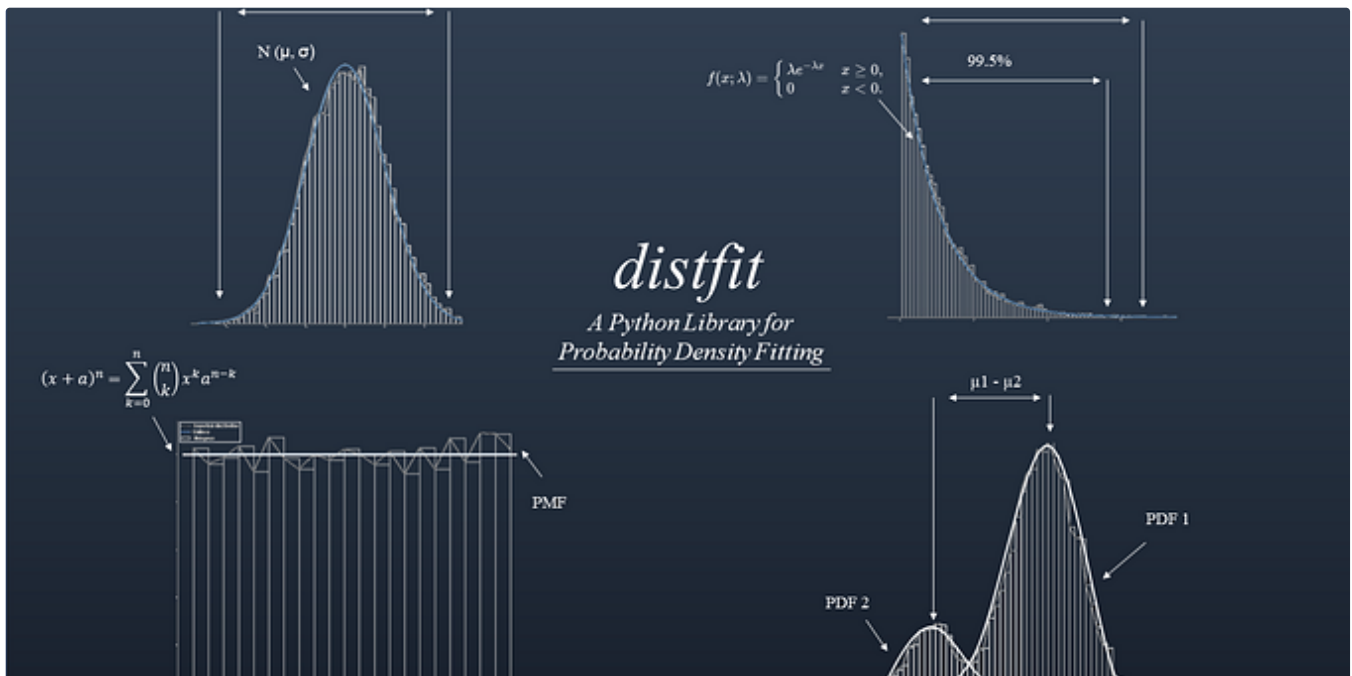
Using Causal ML Instead of A/B Testing

In complex environments, Causal ML is a powerful tool because it is more flexible than A/B Testing, and it doesn't require strong...

🌟 · 9 min read · Nov 29, 2022

 923  15






 Erdogan Taskesen in Towards Data Science

How to Find the Best Theoretical Distribution for Your Data

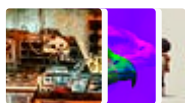
Knowing the underlying data distribution is an essential step for data modeling and has many applications, such as anomaly detection...

★ · 19 min read · Feb 4

 955  10



Lists



What is ChatGPT?

9 stories · 84 saves



Stories to Help You Grow as a Software Developer

19 stories · 99 saves



Stories to Help You Level-Up at Work

19 stories · 76 saves



Staff Picks

341 stories · 96 saves



Nazlı Alagöz in Towards Data Science

Event Studies for Causal Inference: The Dos and Don'ts

A guide to avoiding the common pitfalls of event studies

★ · 17 min read · Dec 18, 2022



207



Jiahui Wang in Towards Data Science

Introduction of Four Types of Item Similarity Measures

Covers how to choose the similarity measure when item embeddings are available

🌟 · 5 min read · Feb 18



52



Luís Roque in Towards Data Science

Primer on Bayesian Deep Learning

Probabilistic Deep Learning

🌟 · 8 min read · Feb 1



100



2





Nazlı Alagöz in Towards Data Science

Crossing the Bridge: A Comparison of Data Science in Academia and Industry

A Ph.D. student's exploration of the surprising parallels between academic and industrial data science

🌟 · 8 min read · May 29



225



2



See more recommendations