DATA5000
Artificial Intelligence Programming in Business Analytics

Causal ML & Meta-Learners

Workshop #5

DATA5000 Roadmap

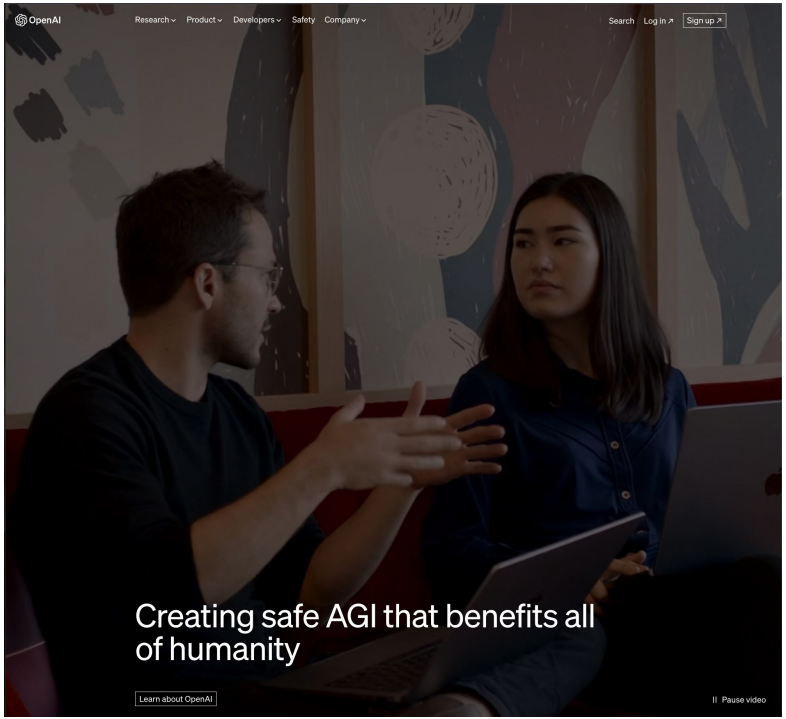
Week 1 Artificial Intelligence with Python	Week 2 AI Predictive Models	Week 3 Deep Learning	Week 4 Causal AI
Week 5 **Causal Forests & Meta-Learners**	Week 6 Advanced ML: CNN for video and image processing	Week 7 Generative AI	Week 8 Prompt Engineering
Week 9 A2: Generative AI Startup (35%)	Week 10 Introduction to Large Language Models	Week 11 Application of Large Language Models 1	Week 12 Application of Large Language Models 1
Week 13 A3: Project Report (40%)			

Our **journey** so far

Last time, in our workshop, we applied Causal Machine Learning to predict Boston House prices.

We also found that the presence of ethnicities in Boston neighborhoods is not a causal factor in the outcome for house prices.

Today, we will examine a larger dataset for house prices – the Ames Housing Dataset.



The **Structure** of Business Analytics

Descriptive Analytics

How did we get here?
Storytelling with writing & visuals

Predictive Analytics

What are the future outcomes?
Forecast Models
Predictive Models (Machine Learning)

Prescriptive / Causation Analytics

What are the causes for how we arrived at the present moment?
Causal AI inference

We master all three!

With creativity, critical and lateral thinking, and [illegible]

https://online.hbs.edu/blog/post/importance-of-business-analytics

Recall the Types of **Treatment** Effects

Average Treatment Effect (ATE)

Measures the average impact of a treatment on the outcome across the entire population.
Think: mass influence.

Local Treatment Effect (LATE)

When treatment is not random.

It focuses on the treatment effect for individuals who are affected by certain conditions or instruments (e.g., an educational program's effect on students who attend a specific school). ***Think: design to influence a subset of a population in a specific scenario.***

Heterogeneous Treatment Effect (HTE)

Accounts for variations in treatment effects across different subgroups within the population. It helps identify whether the treatment is more, or less, effective for certain groups. ***Think: did it work for some and not others? Would a technique that worked for one or more groups work for these other groups?***

Conditional Treatment Effect (CATE)

Measures the average treatment effect for a specific subgroup or condition within a population.

CATE focuses on a particular subgroup defined by specific characteristics or conditions. It provides insights into how the treatment impacts that specific group, for e.g. women. ***Think: personalisation.***

Synthetic Control Model

An important model in evaluating treatment effects is the Synthetic Control Model (SCM).

Its principal advantage is when there is no real control data available, for e.g., when investigating the effects of a new tax already implemented in a state.

SCMs use a synthetic control made up of a weighted average of the control variables from other units which have not received the treatment.

Abadie, Diamond & Hainmueller (2010) used SCM to find the effect of Proposition 99, a tobacco control program implemented in 1988 in California.

Source: Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association, 105*(490). pp. 493 – 505.

Synthetic Control Model

We will reproduce the SCM as per Abadie, Diamond & Hainmueller (2010) in Colab.

You will need to download the following data files:

prop99.csv from
https://github.com/jehangiramjad/tslib/blob/master/tests/testdata/prop99.csv

Smoking.rda from
https://github.com/johnson-shuffle/mixtape/blob/master/data/smoking.rda

Source: https://github.com/tom-beer/Synthetic-Control-Examples

Importing the Packages for SCM

Install and/or import the following packages:

```
!pip install pyreadr

import pandas as pd
import numpy as np
from matplotlib import pyplot as plt
import pyreadr
from google.colab import files
import copy
from scipy.optimize import fmin_slsqp
from sklearn.metrics import mean_squared_error
```

Importing the Data into Colab

The data is now imported into Google Colab. Once uploaded to Google colab, the files are loaded as a *Pandas* dataframe.

1. Upload prop99.csv to Colab using the ***files*** package and the ***upload*** function.
2. Load the data as a *Pandas* dataframe called **df_outcome_raw** using the ***read_csv*** function from ***Pandas***. (Hint: you can copy the path of prop99.csv by right-clicking on the uploaded file)
3. Upload smoking.rda to Colab.
4. Load the data as a *Pandas* dataframe called **rda_predictors** using the ***read_r*** function from ***pyreadr***.

Importing the Data into Colab

The data is now imported into Google Colab. Once uploaded to Google colab, the files are loaded as a *Pandas* dataframe.

The below command is used to upload the files into Colab:

```
uploaded = files.upload()
```

We then import the files as *Pandas* dataframes using the below commands:

```
df_outcome_raw = pd.read_csv('/content/prop99.csv')
```

```
rda_predictors =  pyreadr.read_r('/content/smoking.rda')
```

Synthetic Control Method in Colab

Figure 1: Trends in per-capita cigarette sales: California vs. the rest of the United States

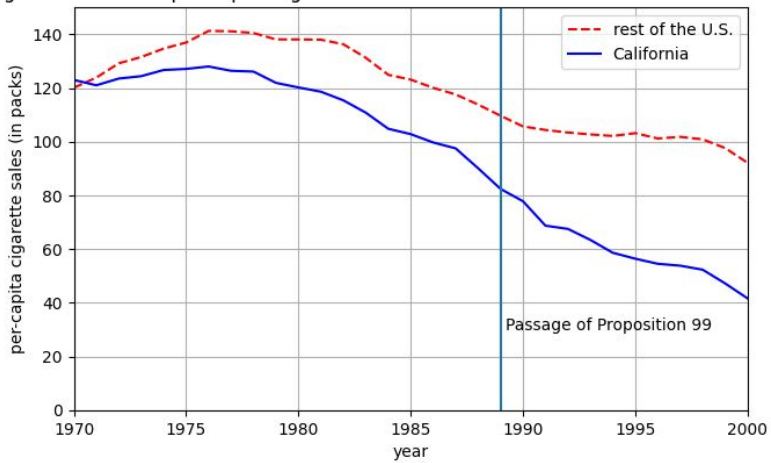
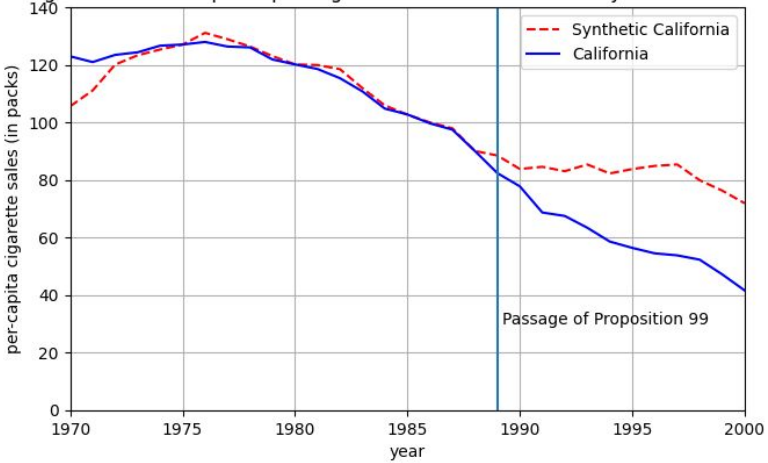
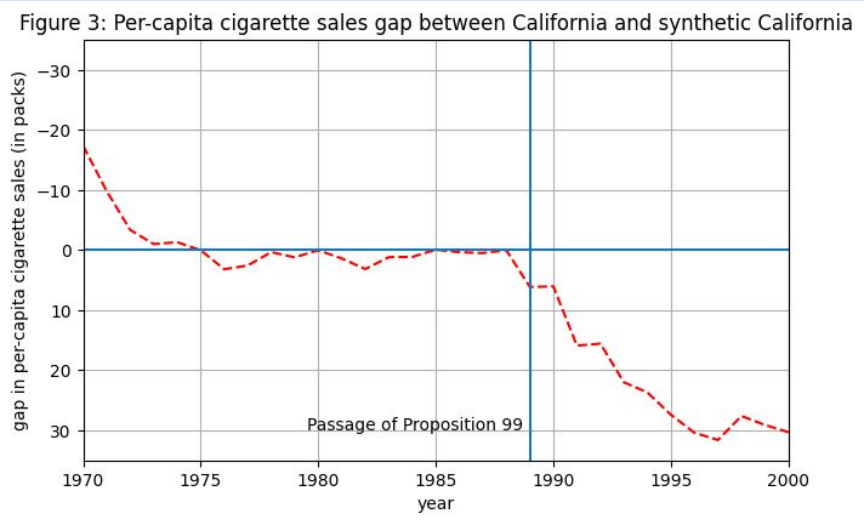


Figure 2: Trends in per-capita cigarette sales: California vs. synthetic California



SCM to Evaluate Treatment Effect

According to Figure 3, by how much did the implementation of Proposition 99 decrease cigarette sales in California by the year 2000?



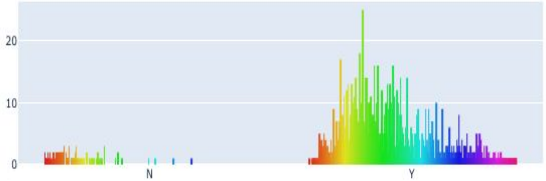
Link to notebook: https://colab.research.google.com/github/data-5000/data5000/blob/main/week_6/week_6_notebook.ipynb

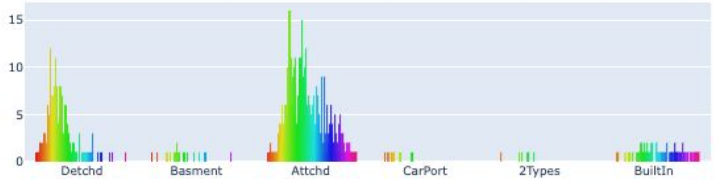
Predict Ames Iowa House Prices with **Causal ML**

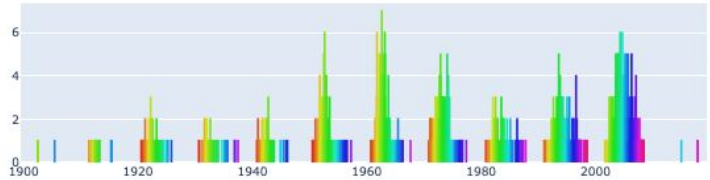
Id	MSSubClass	MSZoning	LotFrontage	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	...	SaleType	SaleCondition	SalePrice	AgeAtSale	YearsSinceRemodel	HasDeck	HasPorch	HasFireplace	HasFence	Intercept
1.0	60.0	RL	65.0	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	...	WD	Normal	208500.0	5.0	5.0	0	1	0	0	1
2.0	20.0	RL	80.0	Pave	NA	Reg	Lvl	AllPub	FR2	Gtl	...	WD	Normal	181500.0	31.0	31.0	1	0	1	0	1
3.0	60.0	RL	68.0	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	...	WD	Normal	223500.0	7.0	6.0	0	1	1	0	1
4.0	70.0	RL	60.0	Pave	NA	IR1	Lvl	AllPub	Corner	Gtl	...	WD	Abnormal	140000.0	91.0	36.0	0	1	1	0	1
5.0	60.0	RL	84.0	Pave	NA	IR1	Lvl	AllPub	FR2	Gtl	...	WD	Normal	250000.0	8.0	8.0	1	1	1	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1456.0	60.0	RL	62.0	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	...	WD	Normal	175000.0	8.0	7.0	0	1	1	0	1
1457.0	20.0	RL	85.0	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	...	WD	Normal	210000.0	32.0	22.0	1	0	1	1	1
1458.0	70.0	RL	66.0	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	...	WD	Normal	266500.0	69.0	4.0	0	1	1	1	1
1459.0	20.0	RL	68.0	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	...	WD	Normal	142125.0	60.0	14.0	1	1	0	0	1
1460.0	20.0	RL	75.0	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	...	WD	Normal	147500.0	43.0	43.0	1	1	0	0	1

1451 rows × 65 columns

We want to determine the factors that influence the house prices given 68 features.







Python **Notebook**

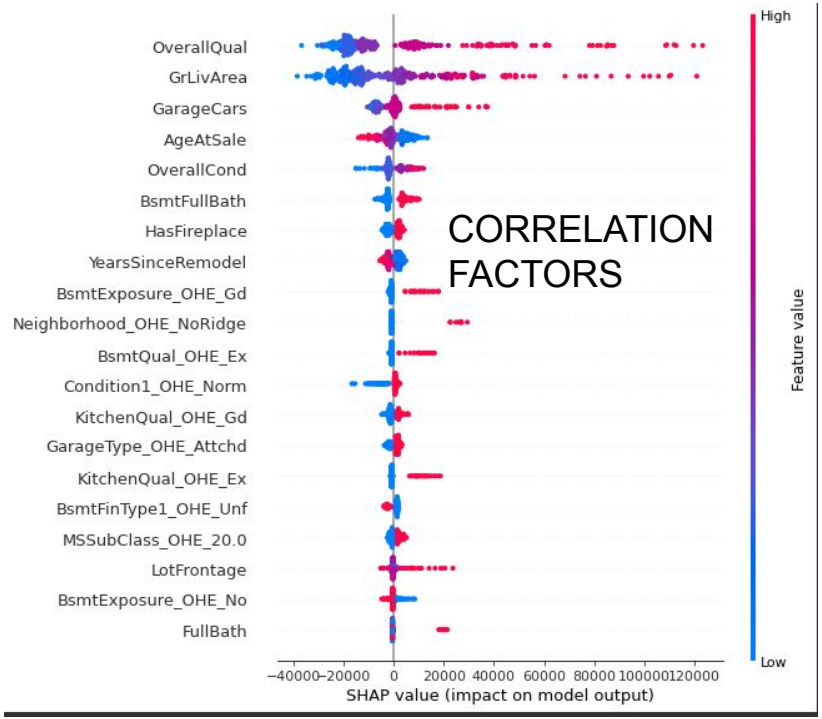
Our Ames Iowa Housing Prediction Notebook:

https://colab.research.google.com/drive/19cJCoxp3qMQLbN3tUNRnIqMxY69ZB21U?usp=sharing

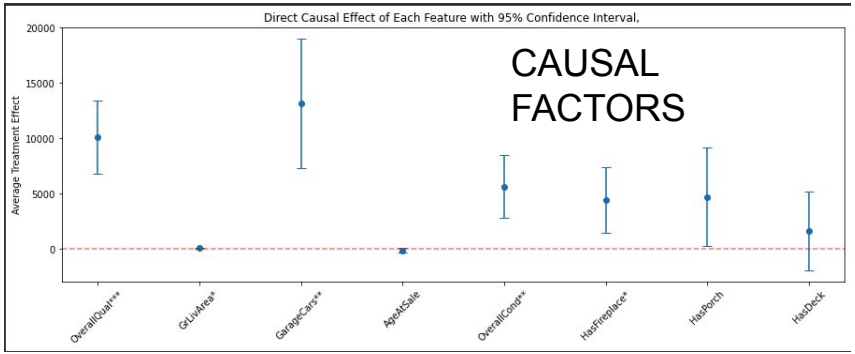
Complete the notebook and discuss the following:

1. **What were the top features found using SHAP Explainable ML?**
2. **What was predicted to be casual factors?**
3. **What happened if we controlled for fireplace?**
4. **What happened if every house had a fireplace that costs, on average, $2500?**
5. **Predict the effects on a new population of house.**

Factors that **Influence** House Prices

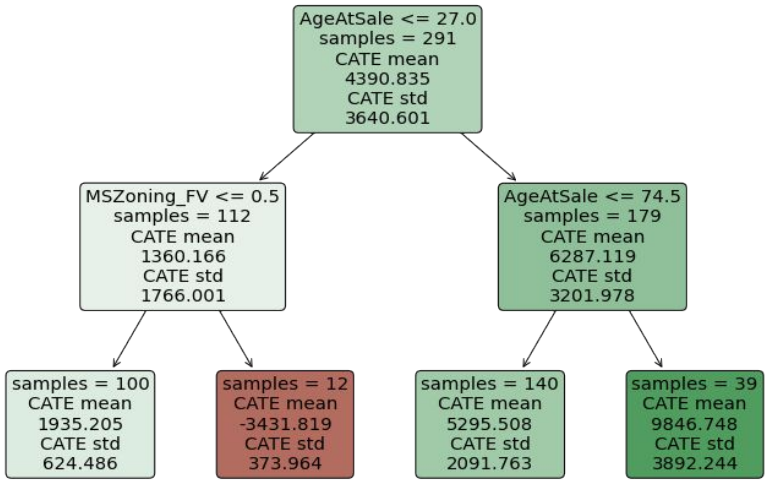


feature	feature_value	point	stderr	zstat	p_value	ci_lower	ci_upper
OverallQual	num	10103.815563	1675.148678	6.031593	1.623509e-09	6820.584485	13387.046640
GarageCars	num	13144.175268	2997.638920	4.384843	1.160696e-05	7268.910946	19019.439590
OverallCond	num	5645.799133	1457.913311	3.872520	1.077156e-04	2788.341551	8503.256716
GrLivArea	num	53.680237	16.896738	3.176959	1.488283e-03	20.563239	86.797235
HasFireplace	1v0	4391.217178	1510.683165	2.906776	3.651749e-03	1430.332582	7352.101775
HasPorch	1v0	4702.903044	2279.139024	2.063456	3.906933e-02	235.872642	9169.933446
AgeAtSale	num	-122.702114	110.966856	-1.105755	2.688327e-01	-340.193155	94.788926
HasDeck	1v0	1610.641783	1819.477098	0.885222	3.760367e-01	-1955.467800	5176.751367



Segmentation – What happened when we **controlled** for **Fireplace**?

Heterogeneity Tree for the Effects of Fireplace



Insights

From the global level, we know that the **ATE (Average Treatment Effect)** of having a fireplace is 4.4k, which means on average that having a fireplace will raise the housing price by $4.4k.

In the diagram on the left, we can see although overall fireplaces already have a positive effect on housing price, the effect is even more dramatic on houses older than 75 years old.

Policy Analysis – What is the **best policy** if we considered cost?

Insights

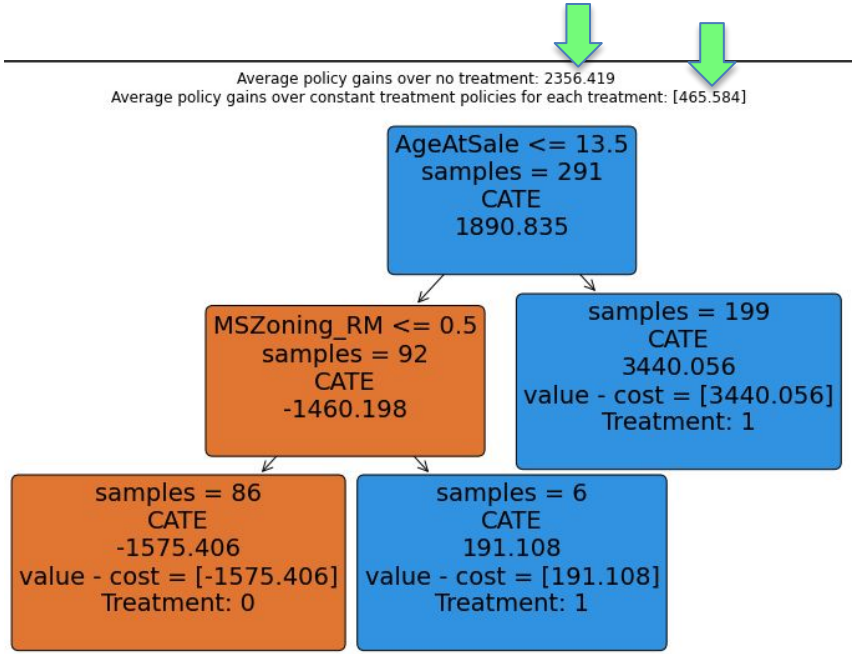
To take a step further, we'd like to know the sub-population where the treatment effect will still be positive after taking cost into consideration.

Assuming the average cost of adding a fireplace is $2,500, let us see what kind of houses have a housing price that will increase more than their cost.

You could see if we follow the recommended policy above, on average, the housing price will increase by $2,356 compared with no fireplace added.

Similarly, it will increase by $465 compared with adding a fireplace for every house.

To be more detailed, we could also output the individualized policy. In the following table, we will only print the top five houses ordered by policy gains.



Control the cost of a fireplace – observe effects

Assuming the average cost of adding a fireplace is $2,500, let us see what kind of houses have a housing price that will increase more than their cost.

Id	Treatment	Effect of treatment	Effect of treatment lower bound	Effect of treatment upper bound	MSSubClass	MSZoning	LotFrontage	Street	Alley	LotShape	...	MoSold	SaleType	SaleCondition	AgeAtSale	YearsSinceRemodel	HasDeck	HasPorch	Current treatment	HasFence	Intercept
1350.0	1	12224.503178	3758.862937	20690.143420	70.0	RM	50.0	Pave	Pave	Reg	...	12.0	WD	Normal	136.0	21.0	0	1	0	0	1
1063.0	1	9574.109709	3261.378628	15886.840790	190.0	RM	85.0	Pave	Grvl	Reg	...	9.0	WD	Normal	107.0	57.0	0	1	0	0	1
243.0	1	9482.716831	3235.816161	15729.617501	50.0	RM	63.0	Pave	NA	Reg	...	4.0	WD	Normal	106.0	56.0	0	1	0	0	1
199.0	1	8660.180927	2968.655603	14351.706250	75.0	RM	92.0	Pave	NA	Reg	...	7.0	WD	Abnorml	97.0	59.0	0	1	0	1	1
521.0	1	8618.608621	1186.794764	16050.422479	190.0	RL	60.0	Pave	Grvl	Reg	...	8.0	WD	Normal	108.0	8.0	1	1	0	0	1

5 rows × 68 columns

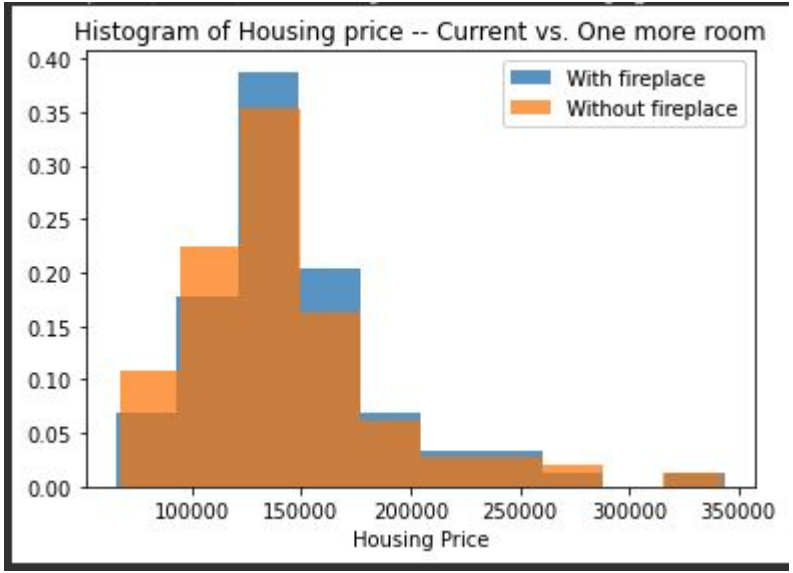
In the Treatment column, 1 corresponds to having a fireplace, and 0 corresponds to no fireplace.

What if? Counterfactual Analysis

What would happen to the house price if every home had a fireplace?

The causal analysis tool could also answer **what if** types of questions.
For a given treatment, we'd also like to know the **counterfactuals** if we intervene it in a different way.

- From the summary table we could see overall if we add a fireplace to houses without fireplaces in the test set, the average housing price for those houses will increase by about $5k.
- And the histogram shows a comparison between the current housing price distribution and the counterfactuals distribution if we added a fireplace to the fireplace-less houses in the test set.



Next, we will learn how the overall housing price changes if every house in Ames had a fireplace.

Cohort & **Local** Treatments Analysis

What is the causal effect on a new population?

From the two tables, you can see the global effect on the test set (the actual prediction) is like that of the training set.

And calculating the local effect gives you the heterogeneous treatment effect for each observation.

causes

```
1 # predict global effect
2 causal_analysis.cohort_causal_effect(x_test)
```

feature	feature_value	point	stderr	zstat	p_value	ci_lower	ci_upper
OverallQual	num	10207.265524	1694.108274	6.025155	1.689473e-09	6886.874322	13527.656727
GrLivArea	num	53.572155	16.896406	3.170624	1.521117e-03	20.455808	86.688502
GarageCars	num	13215.235723	3030.831642	4.360267	1.299037e-05	7274.914862	19155.556584
AgeAtSale	num	-139.794231	110.416091	-1.266068	2.054889e-01	-356.205793	76.617332
OverallCond	num	5723.914340	1462.849911	3.912851	9.121266e-05	2856.781200	8591.047479
HasFireplace	1v0	4390.835030	1523.103093	2.882822	3.941301e-03	1405.607823	7376.062237
HasPorch	1v0	4713.175859	2303.096997	2.046451	4.071199e-02	199.188692	9227.163026
HasDeck	1v0	1653.390857	1836.856976	0.900120	3.680566e-01	-1946.782662	5253.564375

```
1 # predict local effect for each sub-group
2 causal_analysis.local_causal_effect(x_test)
```

sample	feature	feature_value	point	stderr	zstat	p_value	ci_lower	ci_upper
0	OverallQual	num	10466.451123	1701.632579	6.150829	7.707896e-10	7131.312552	13801.589694
	GrLivArea	num	54.152219	8.332353	6.499031	8.083900e-11	37.821108	70.483330
	GarageCars	num	10319.768990	2381.310682	4.333651	1.466567e-05	5652.485818	14987.052162
	AgeAtSale	num	-276.854177	115.013736	-2.407140	1.607800e-02	-502.276958	-51.431396
	OverallCond	num	8238.320947	1069.053989	7.706179	1.296412e-14	6143.013632	10333.628263
...	...	...	...	...	...	...	...	...
290	AgeAtSale	num	-276.854177	115.013736	-2.407140	1.607800e-02	-502.276958	-51.431396
	OverallCond	num	5849.940818	3055.390550	1.914629	5.553977e-02	-138.514619	11838.396255
	HasFireplace	1v0	1339.570649	2653.292705	0.504871	6.136494e-01	-3860.787493	6539.928791
	HasPorch	1v0	6906.589930	5156.130035	1.339491	1.804109e-01	-3199.239237	17012.419098
	HasDeck	1v0	1899.860305	3984.727566	0.476785	6.335149e-01	-5910.062213	9709.782822

2328 rows × 6 columns

Transformer Foundation of Generative AI

NEXT WORKSHOP – ADVANCED DEEP LEARNING