Mia Jerphagnon

DATA SCIENCE ALLIANCE

Summer 2024 Retrospective

Internship Goals

Develop technical skills

Python, Tableau, machine learning models

Understand real-world applications

Business solutions and data-driven decisions

Learn Responsible Data Science

Best practices for Fairness, Transparency, Privacy, and Veracity

Build professional relationships

Networking with and getting advice from people in the industry

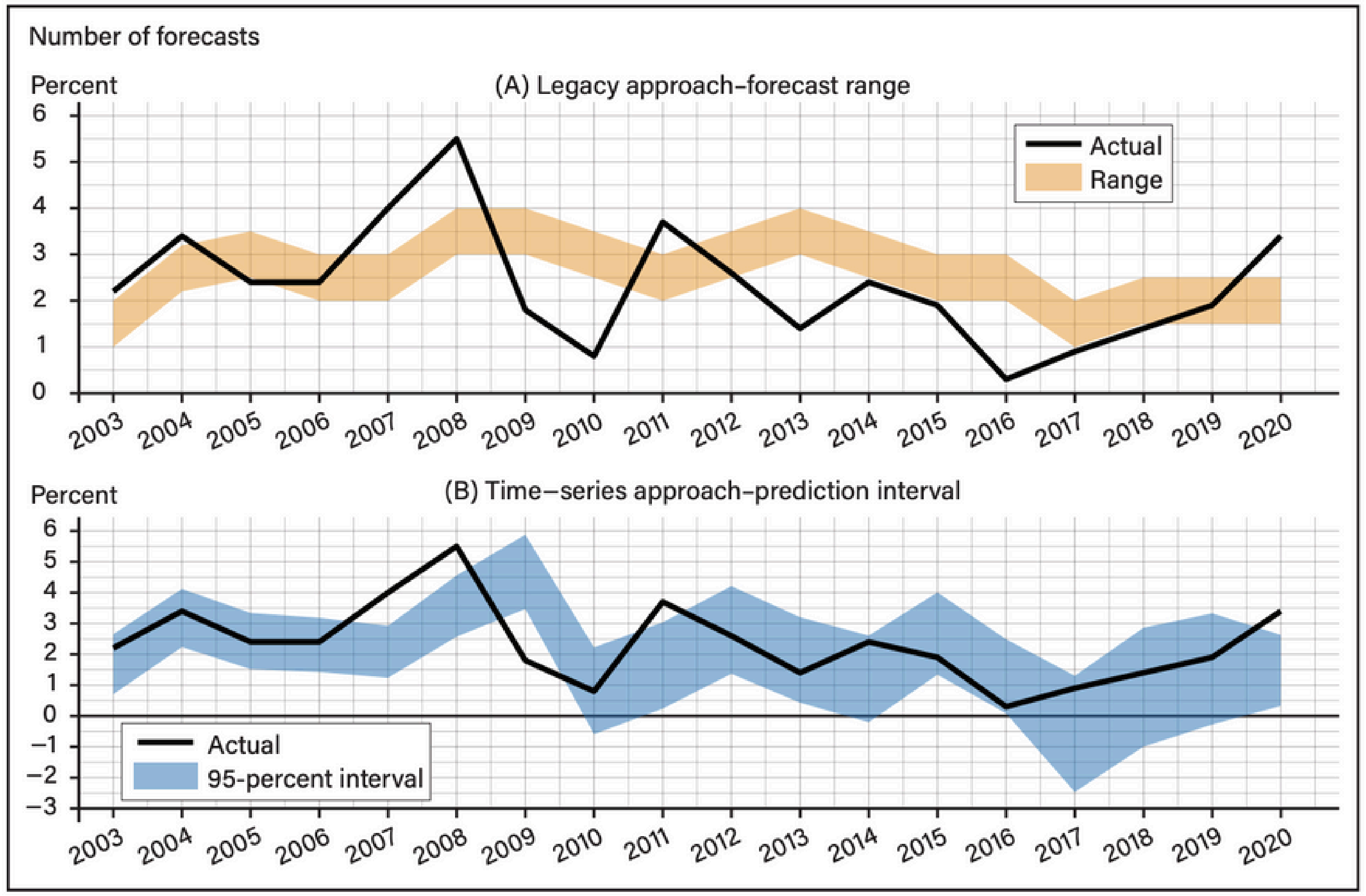
Food Bank Project & Key Takeaways

Literature reviews, forecasting food insecurity and demand, and the Tableau executive summary dashboard

Lit Reviews

Time-Series Methods for Forecasting and Modeling Uncertainty in the Food Price Outlook

Figure 2
Actual annual percent change in all food prices, forecast range, and the 95-percent prediction interval, 2003–20



Predicting hotspots of unsheltered homelessness using geospatial administrative data and volunteered geographic information

Predictors	
Provider-based administrative data[1]	Sum of homeless services street outreach case encounters in each quarter/year, lagged by one year; 2018–2019.
Volunteered geographic information[1]	Sum of 311 calls from residents about homeless encampments in each quarter/year, lagged by one year 2018–2019.
Expert reports[1]	Expert generated estimated counts of the number of unsheltered people per tract from the 2019–2020 LAHSA PIT Count planning sessions.
Neighborhood sociodemographic characteristics[2]	% Non-Hispanic Black; Poverty rate; % Housing vacancy
Built environment features[2]	Access to services (i.e., number of shelters within half a mile of the tract centroid); % Commercial land; Distance (miles) to the Central Business District (log transformed)
Persistent trends in unsheltered homelessness[1]	PIT Count of total unsheltered persons in the previous year; 2018–2019

1) Time-varying variable; 2) Time-invariant variables, corresponding to the year 2019.

Lit Reviews

Identification of factors related to food insecurity and the implications for social determinants of health screenings

Multivariable Association Between Select Characteristics and Food Security Status

Variable	Estimate (SE)	OR (95% CI)
Sociodemographic Characteristics		
Education level of high school or less	0.315 (−0.137, 0.766)	1.37 (0.872, 2.151)
Receives SNAP benefits	**0.591 (0.129, 1.052)**	**1.805 (1.138, 2.864)**
Experienced significant changes in life in past year	**0.63 (0.196, 1.064)**	**1.877 (1.216, 2.899)**
Health Information		
Chronic disease	**0.462 (0.041, 0.883)**	**1.587 (1.042, 2.417)**
Food Shopping and Dietary Behaviors		
Store 1 shopping frequency in months	**0.099 (0.032, 0.166)**	**1.104 (1.033, 1.18)**
Shops with own car	**−0.589 (−1.038, −0.14)**	**0.555 (0.354, 0.869)**
2010 Healthy Eating Index score	−0.02 (− 0.042, 0.002)	0.98 (0.959, 1.002)
Beliefs about Food Shopping and Diet		
I have enough time to shop for fresh and healthy foods	−0.265 (− 0.6, 0.07)	0.767 (0.549, 1.072)
It is convenient for me to purchase fresh and healthy food	**−0.304 (− 0.605, − 0.003)**	**0.738 (0.546, 0.997)**
Eating a fresh and healthy diet is affordable	− 0.17 (− 0.425, 0.084)	0.844 (0.654, 1.088)
It is easy to eat a fresh and healthy diet	**− 0.317 (− 0.58, − 0.056)**	**0.728 (0.56, 0.946)**

California Health Interview Survey 2024 Questionnaire Topics

FoodAPS National Household Food Acquisition and Purchase Survey

ADLTFSRAW

Variable: ADLTFSRAW	Definition: Adult food security score—30-day measure				Type: Numeric
	Raw score based on values of FOODSECUREQ1-FOODSECUREQ10				
	N	Min	Max	Mean	#Missing (.)
	4,826	0	10	1.721923	0

ADLTFSCAT

Variable: ADLTFSCAT	Definition: Adult food security status—30-day measure			Type: Numeric
	Classification based on value of ADLTFSRAW: 0 = high food security, 1-2 = marginal food security; 3-5 = low food security; 6-10 = very low food security			
	Value	Count	Percent	Value description
	1	2,522	52.26	High food security
	2	960	19.89	Marginal food security
	3	785	16.27	Low food security
	4	559	11.58	Very low food security

FOODSUFFICIENT

Variable: FOODSUFFICIENT	Definition: Respondent description of food sufficiency within last 30 days			Type: Numeric
	Value	Count	Percent	Value description
	1	2,452	50.81	Enough of the kinds of food we want to eat
	2	1,892	39.20	Enough, but not always the kinds of food we want to eat
	3	365	7.56	Sometimes not enough to eat
	4	117	2.42	Often not enough to eat

Key Takeaways from Lit Reviews

Clear structure & focus

Project goal, project overview, key findings, data used, RDS

Challenges & biases in data

Gaps in data, data collection bias

Use of statistical models

Logistic regression, time series econometric methods

Application of findings to project

Total food demand for entire zip code:
FD = (FB - SB) x # of households in zip

Food Insecurity: `fi_estimation_state_annual.ipynb`

1: Data Prep & Merging

Merged dfs, calculated shares, converted Year to factor var., loaded additional vars., extracted year from date

```
R-squared:                  0.6329
R-squared (Between):       -4.6351
R-squared (Within):         0.8693
R-squared (Overall):       -3.0075
```

2: Run FA regression

Unemployment
Poverty rate
Median income
Hispanic households
Black households
Homeowner pop.
Disability pop.

Entity & time effects

3: Multi-collinearity

Calculated VIF

#1: Homeowners share and median income had high VIFs and p-values

#2: CPI, rent price, electricity price, and gas price had high VIFS

4: New regressions

Adding CPI decreased R2. Model had high p-values, too.

Adding CalFresh variable increased R2 (within) to 0.88, but p-val > 0.01.

Adding cost per meal further decreased R2.

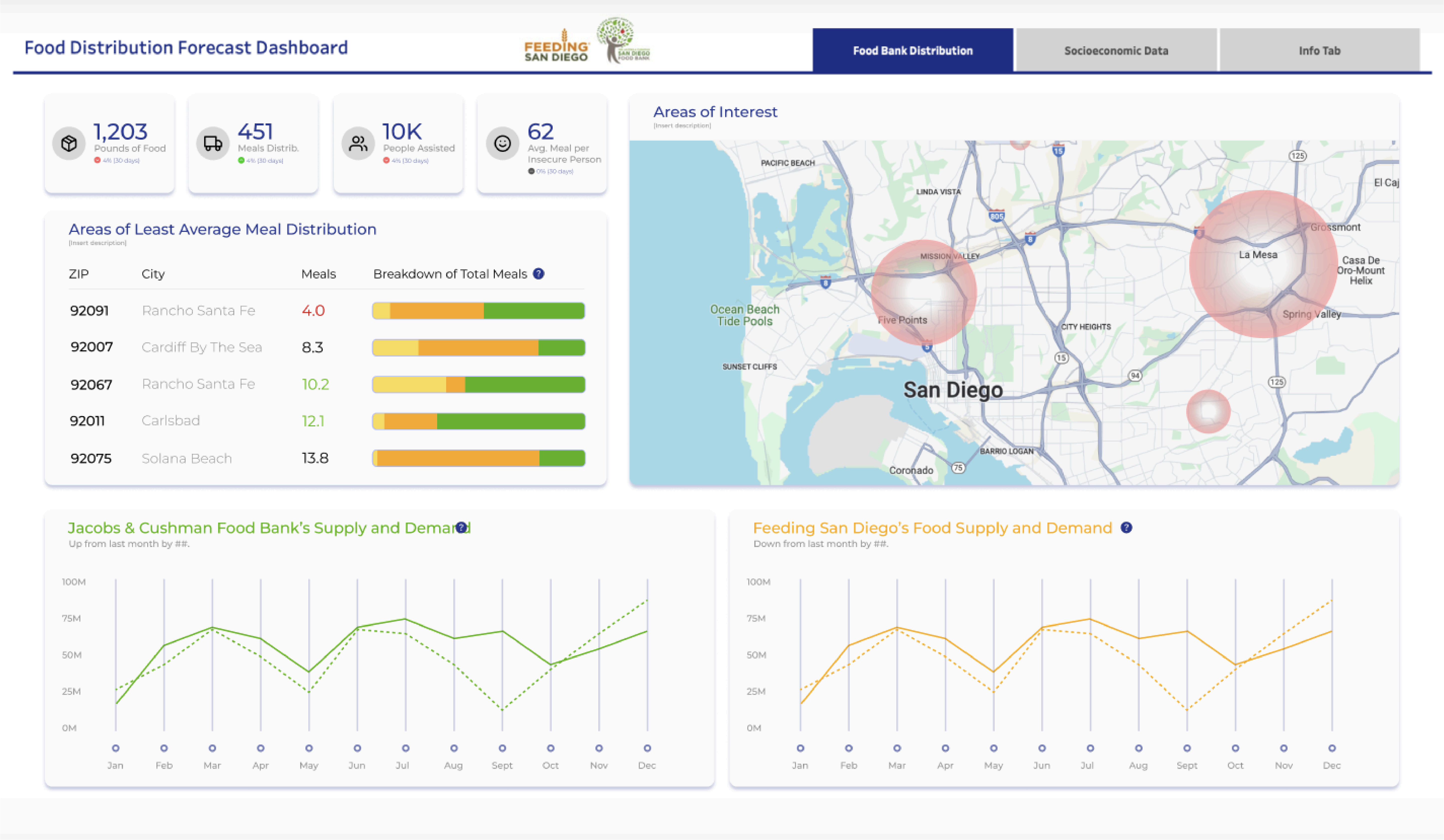
5: Analysis & takeaways

Model with the FA variables and CalFresh had the best results.

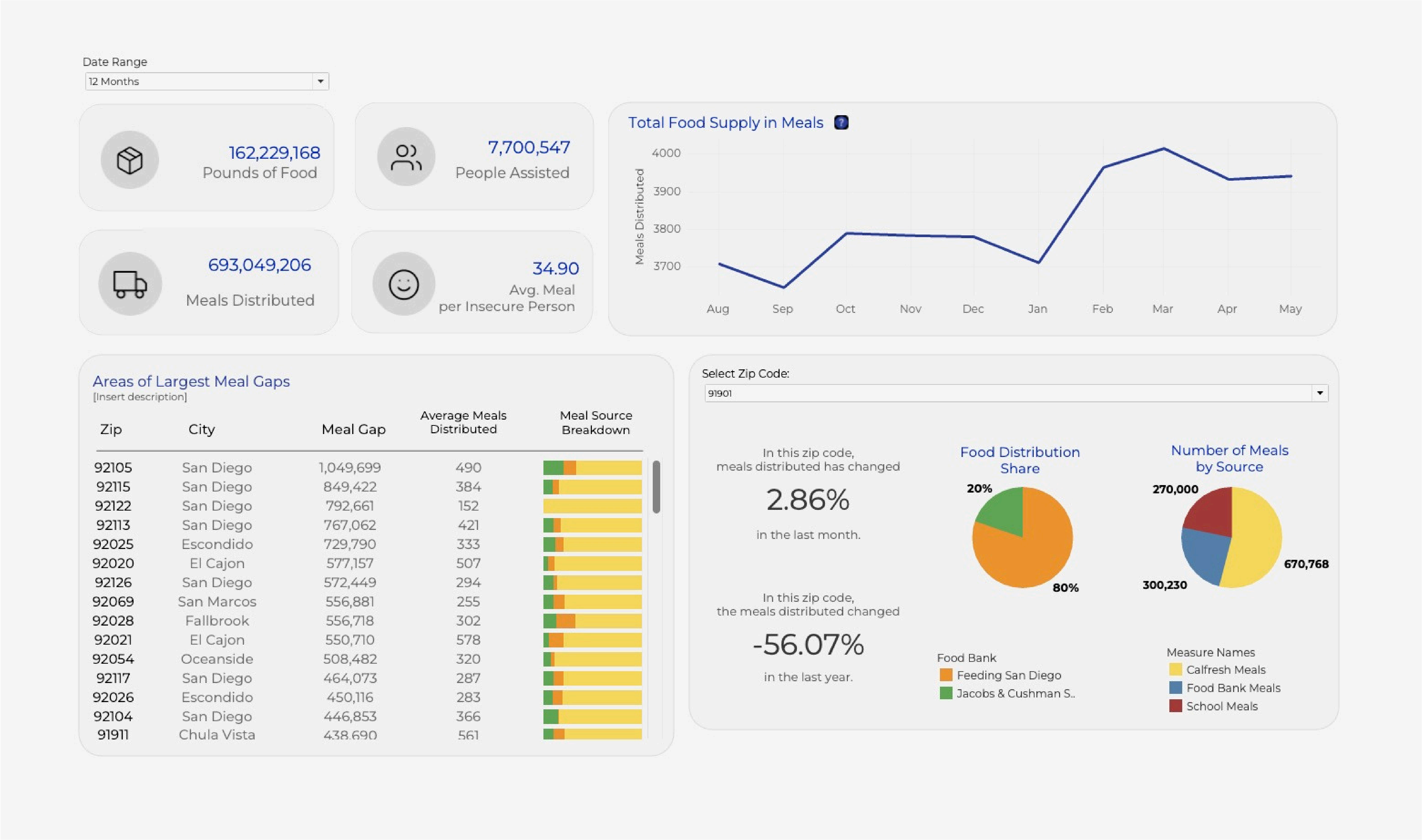
Should we remove older years which decreased R2?

Is R2 the best/only accuracy measure?

Brainstorm & Design Food Bank Dashboard



Learn & Draft on Tableau



Final Draft for Stakeholders

Food Distribution Forecast Dashboard

FEEDING SAN DIEGO | SAN DIEGO FOOD BANK

At a Glance | Food Distribution | Socioeconomic Data | Info Tab

Food Assistance Overview in San Diego County

1 Month

Feeding San Diego and Jacobs & Cushman San Diego Food Bank served with:

7,202,752 Pounds of Food

200,076 People Assisted

Together with CalFresh assistance and School Meals, both food banks served with:

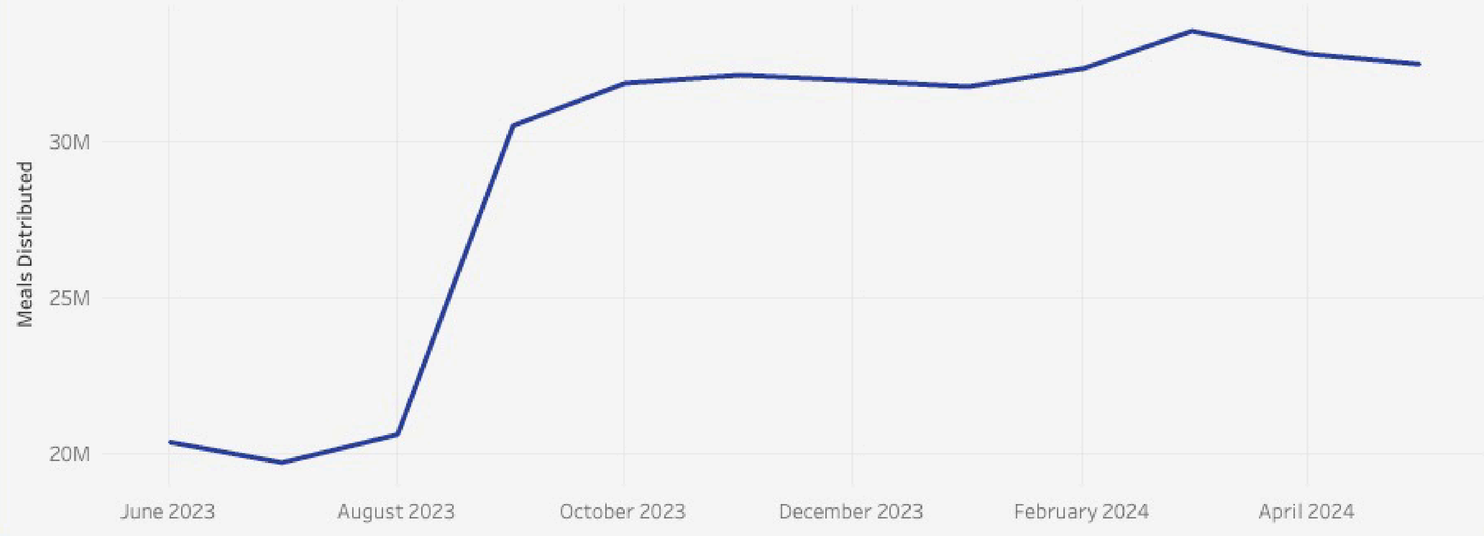
32,492,779 Meals Distributed

56.14 Avg. Meals per Person in Poverty

Areas of Largest Meal Gaps

Zip Code	City	Meal Gap	Avg. Meals per Person in Poverty	Distribution Source
92122	San Diego	437,994	20.1	
92115	San Diego	211,195	48.4	
92025	Escondido	210,656	45.5	
92069	San Marcos	199,868	40.1	
92104	San Diego	199,266	34.6	
92126	San Diego	187,208	42.4	
92054	Oceanside	186,180	40.5	
92028	Fallbrook	180,179	43.3	
92011	Carlsbad	176,710	11.1	
92117	San Diego	167,689	40.5	
92105	San Diego	166,617	53.3	

Total Food Supply in Meals Across San Diego County



Select Zip Code for detailed food distribution information:

92115

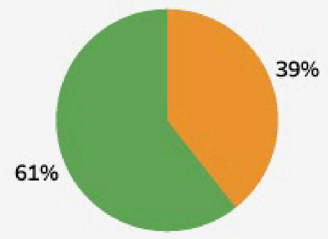
Change in distribution since **last month:**

0.22%

Change in distribution since this month **last year:**

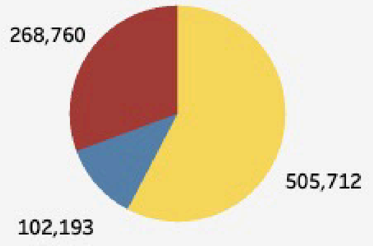
4.64%

Food Bank Distribution Share



Food Bank
- Feeding San Diego
- Jacobs & Cushman SDFB

Number of Meals by Source



Meal Source
- Calfresh
- Food Banks
- School Meals

Food Demand: `food_demand_8_20.ipynb`

1: Finding data

Decennial county-level poverty data, annual zip-code level poverty data, annual county-level wage data, county-level labor data

2: Loading & processing

Cleaning untidy data was a huge task.
Melting
Concatenation
Extracting Year from csv file
Missingness
Different granularity
Aggregation

3: Debug & debug

Mixed data types created problems, e.g. Year was stored as both a str & int within the same column.
Required new data to to connect zip code to county.
After melting, counties inc. SD were missing.

4: Analysis & takeaways

Finding data is a ginormous task all on its own.

Cleaning untidy data, using it with other datasets, and debugging are challenging and time-consuming.

This is why tidy data practices are important!

Be patient preparing data for running models. Doing it correctly will help in the long run and lead to more credible, unbiased results.

Challenges & Areas of Improvement

Data Quality Issues
Handling incomplete, inconsistent, or noisy data is challenging, not to mention finding data that suits our project. I hope to further improve robust data cleaning and preprocessing techniques.

Mia Jerphagnon

Presenting findings at weekly meetings
I want to work on being more well-spoken and concise when sharing the work I completed each week.

Mia Jerphagnon

Successes

Networking
Getting to know everyone in the DSA team and professionals at the discourses was an amazing experience! I was able to build confidence and feel comfortable talking to unfamiliar faces.

Mia Jerphagnon

Responsible Data Science
I loved learning about the RDS Framework and discussing how we can best implement and explain best practices.

Mia Jerphagnon

Consider stakeholders & consumers's needs when delivering data as a product

Mia Jerphagnon

Learning Tableau and developing other skills!
I was able to get a comprehensive understanding of Tableau and its various features and customizations.

Mia Jerphagnon

Collaborating within a professional setting
It was great learning the protocols of working for an organization and how to collaborate outside of a classroom setting.

Mia Jerphagnon