


# Final Projects

[Start Assignment](#)

**Due** Dec 15 by 11:59pm    **Points** 20    **Submitting** a file upload

The goal of your final project is to impress your peers and your instructors by utilizing some of the tools you have learned this semester. The assignment is to communicate an engaging public policy use case of predictive modeling by showing off both your analytical skills and your ability to convert those skills into a relevant policy use case.

You are *required* to work in pairs ([signup sheet](#) )

([https://docs.google.com/spreadsheets/d/1D2pDhWJDqNW4AihVhdT-tNDj99\\_Qg8oMNs8xPPF17Ac/edit?usp=share\\_link](https://docs.google.com/spreadsheets/d/1D2pDhWJDqNW4AihVhdT-tNDj99_Qg8oMNs8xPPF17Ac/edit?usp=share_link))

). Both team members should work on the model, but each member will present a different part of the deliverable. Each pair will receive a grade that comprises both parts of the project. If you have trouble finding a partner by the end of class on the day of the assignment, consult a TA for help.

Please focus on cross-validation (random, spatial and time, where appropriate) and on goodness of fit indicators (accuracy and generalizability) **that relate directly to the business process**. You may have to make up the business process, but please consider social, economic etc. costs/benefits.

Date	Schedule	Due
11/17	Project Introduced	
11/24	<i>Thanksgiving – No Class</i>	
12/1	Prepare to discuss your (1) chosen project, (2) use case, (3) methods, (4) data, and (5) some app wireframe ideas in lab. Have some exploratory analysis done by this point.	Deliverable 1 (choose partner & project option)
12/8	Presentations in class	Deliverable 2 (in-class presentation)
12/15	Final markdown and video due at noon.	Deliverable 3 (recorded presentation & markdown)

## Deliverables

**Deliverable 1** (Due 12/1): Have a partner chosen, pick a project option, develop it into a use case, and be prepared in lab to talk for a few minutes about these questions:

- What is the project use case?
- How could data make a difference in answering this question? Do you have a sense for the business as usual decision making?
- What datasets have you identified to help you answer this question?

- What kind of model would you build and what is the dependent variable? How will you validate this model (cross-validation & goodness of fit metrics that relate to the business process)?
- How do you think that stakeholders would want to consume this data?
- What are the use cases for your app and what should the app do?
- What are some results of your exploratory analysis and any introductory modeling?

**Deliverable 2** (Due 12/8): Team member 1 will present a 4 minute, 12 slide, in-class '**PechaKucha**' presentation that 'sells' us on the idea of this fancy new planning app that you've designed to solve an important problem. This is a presentation where the slides are set to change automatically, **every 20 seconds**. This timing is an absolute requirement.

Submit your presentation file to the "Final Class Presentation" assignment on Canvas so we can queue it up for class.

1. Spend ~50% of your presentation time on exploratory analysis, the model's methodology, and the model's results. The expectation is that you will have preliminary model or a modeling idea at this stage, which you will sharpen by the time the assignment is due.
2. The other 50% of presentation time should be selling us the project. Focus on questions like: What is the use case? Who is the user? How does the app put the model into the hands of a non-technical decision maker? Who is creating the app? Have you created something that is usable by the client?

Remember – sell it to us. What should come first - the model or the app? Don't forget to constantly remind the audience about the use case to keep your solution relevant.

**Deliverable 3** (Due 12/15 @ 11:59 pm): Your team will submit two things: (1) a recorded presentation and (2) the markdown.

Team member 1 will have to record a **pechakucha presentation** (4-min max & 20 seconds per slide), then upload it on Youtube. Submit that video link (1) in the comments of your Canvas submission and (2) add the link to your markdown.

Team member 2 will be responsible for a **markdown** write up that would allow someone to replicate your analysis. *Please show and fold your code blocks*. Make sure your markdown is polished, proofread, and all code & visuals are accessible. Submit this markdown to Canvas. At a minimum, please hit on the below components:

1. Motivate the analysis – “What is the use case; why would someone want to replicate your analysis and why would they use this approach?”
2. Describe the data you used.
3. Describe your exploratory analysis using maps and plots.
  - I expect to see high quality data visualizations
4. What is the spatial or space/time process?
5. Describe your modeling approach and show how you arrived at your final model.
6. Validate your model with cross-validation and describe how your predictions are useful (accuracy vs. generalizability).
7. Provide additional maps and data visualizations to show that your model is useful.
8. Talk about how your analysis meets the use case you set out to address.

9. What could you do to make the analysis better?

## Project Options

### Project option 1 – Predict heroin overdose events to better allocate prevention resources

The City of Mesa has a dataset of heroin overdose locations. Using these data extracted from the city's [Open Data portal](https://data.mesaaz.gov/Fire-and-Medical/Fire-and-Medical-Opioid-Overdose-Incidents/qufy-tzv6) [↗](https://data.mesaaz.gov/Fire-and-Medical/Fire-and-Medical-Opioid-Overdose-Incidents/qufy-tzv6) (<https://data.mesaaz.gov/Fire-and-Medical/Fire-and-Medical-Opioid-Overdose-Incidents/qufy-tzv6>), your job will be to estimate a geospatial risk prediction model, predicting overdoses as a function of environmental factors like crime, 311 and inspections. You should validate your model against a kernel density, as we have did in class. Also, you should try to train your model from one time period (long enough to have enough data) and test it on an out of out of sample test set time period (the following year, for instance). Note the fact that the data have some accuracy diminished to make them more anonymous – think about how this plays into your prediction and use case.

You can also undertake this project using [similar data from Cincinnati, Ohio](https://github.com/sydnng/Cincinatti_Overdose_Data) [↗](https://github.com/sydnng/Cincinatti_Overdose_Data) ([https://github.com/sydnng/Cincinatti\\_Overdose\\_Data](https://github.com/sydnng/Cincinatti_Overdose_Data)).

Think critically about how you might offer these predictions to a public health official in your app. What do they want to know? Also remember that while your predictions are about overdose, it may be safe to assume that these are also places where people are just using heroin.

### Project option 2 – Predict food inspection failures in Chicago to better allocate inspectors

The Chicago Health Department wants to come up with a better way to allocate their limited health inspectors across the many food establishments in the City. How well can you predict if a food establishment will fail a health inspection? Can you figure out an interesting way to use the model to help the Health Department prioritize their inspections? This will use a logistic regression.

Specifically, your goal is to estimate a model using [inspection data](https://data.cityofchicago.org/Health-Human-Services/Food-Inspections/4ijn-s7e5) [↗](https://data.cityofchicago.org/Health-Human-Services/Food-Inspections/4ijn-s7e5) (<https://data.cityofchicago.org/Health-Human-Services/Food-Inspections/4ijn-s7e5>) from one year to predict for the next. Does your model work better for certain kinds of establishments? Certain types of neighborhoods? Find your data on the Chicago Open Data Site.

### Project option 3 – Predict permit applications in Philadelphia:

Rising home values are threatening Philadelphia's relatively affordable cost of living. Can you model the last ten years of 'development' (defined as [the issuance of various permits](https://data.phila.gov/visualizations/li-building-permits) [↗](https://data.phila.gov/visualizations/li-building-permits) (<https://data.phila.gov/visualizations/li-building-permits>)) over the last ten years? Use your domain knowledge to determine a use case related to a spatial forecast of permits. Imagine a user who might find utility in such a prediction - what kind of policy intervention, program or business model might make use of it?



Use some of your knowledge and opinions about urban and community dynamics to guide your project. What is the best type of permit to indicate development - zoning changes? construction permits? How do data at the community level help you model this phenomenon or assess model outputs?

### Project option 4 - Forecast Metro train delays in and around NYC:

An amazing [dataset](https://www.kaggle.com/pranavbadami/nj-transit-amtrak-nec-performance?select=2018_11.csv) [↗](https://www.kaggle.com/pranavbadami/nj-transit-amtrak-nec-performance?select=2018_11.csv) ([https://www.kaggle.com/pranavbadami/nj-transit-amtrak-nec-performance?select=2018\\_11.csv](https://www.kaggle.com/pranavbadami/nj-transit-amtrak-nec-performance?select=2018_11.csv)) has popped up on Kaggle recently that list origin/destinations delays for Amtrak and NJ Transit trains. Can you predict train delays? Consider the time frame that it would be useful to have such predictions.




Predicting 5 minutes out is not going to be as useful as 2-3 hours out. Consider training on a month and predicting for the next week or two. Consider time/space (train line, county etc.) cross validation. Many app use cases here.

### Project option 5 – Forecasting wildfire risk for a region in California:





With climate change, the State of California is exhibiting increased threat of wildfire. No doubt fire risk is a function of climate and weather, but also a host of time-invariant, spatial variables such as vegetation, elevation, land cover and more. Your challenge is to integrate California's [Fire Perimeter](https://frap.fire.ca.gov/frap-projects/fire-perimeters/) , data for 2-3 or years with [other](https://frap.fire.ca.gov/mapping/gis-data/) , fire data, vegetation, land cover data, elevation data and other, to estimate fire risk. Can you use spatial cross-validation to validate this model?

There are multiple possible model approaches here. For an app, granted none of us are forestry experts, but can you design a fire management app that prioritizes where naturalist should clear brush, do burns, etc. Maybe, this is an app aimed at insurance companies or homeowners?




### Project option 6 – Forecast Airbnb Prices in Amsterdam

You can predict home prices – how about apartment rents? Using a scraped Airbnb [dataset](https://www.kaggle.com/erikbruin/airbnb-amsterdam?select=listings.csv) , [predict property prices based on property characteristics and other neighborhood level data](https://www.kaggle.com/erikbruin/airbnb-amsterdam?select=listings.csv) , [data](https://maps.amsterdam.nl/open_geodata/?LANG=en) . For your app, do not reinvent the Airbnb app, but think about a consumer facing or public-policy focused use case. Consider the fact that regulation of Airbnb is a contentious issue in tight real estate markets and the service has been outlawed in some cities.

### Project option 7: Forecasting parking demand

What drives parking demand (revenues)? If you knew, could you create a tool that would predict parking demand over time and space. Using San Francisco open [parking data](https://data.sfgov.org/Transportation/SFMTA-Parking-Meter-Detailed-Revenue-Transactions/imvp-dq3v) , [locations](https://data.sfgov.org/Transportation/SFMTA-Parking-Meter-Detailed-Revenue-Transactions/imvp-dq3v) , [forecast parking demand as a function of built environment, neighborhood and road characteristics](https://data.sfgov.org/Transportation/Parking-Meters/8vzz-qzz9). Lots of app use cases possible depending if the user is public or private sector. There are *105m rows here* – so you'll have to use the API to download a small slice of the data. You can read a literature review of concepts in parking economics and algorithmic estimations of occupancy and demand here - ([Fichman, 2016](https://www.researchgate.net/publication/309231344_An_Evaluation_of_Pittsburgh%27s_Dynamically-Priced_Curb_Parking_Pilot) , [This is a particularly challenging project, and grading will account for that degree of difficulty](https://www.researchgate.net/publication/309231344_An_Evaluation_of_Pittsburgh%27s_Dynamically-Priced_Curb_Parking_Pilot)). I have created [a relatively rough occupancy estimation script](https://github.com/mafichman/parking_estimations) , [but you will probably need to update it to make it more accurate](https://github.com/mafichman/parking_estimations).

### Project Option 8: Forecast train occupancy levels

Can you forecast train occupancy for various OD pairs? There is a really great Kaggle [dataset](https://www.kaggle.com/c/train-occupancy-prediction/data)  (<https://www.kaggle.com/c/train-occupancy-prediction/data>) (the data are [here](https://www.kaggle.com/c/train-occupancy-prediction/discussion/27828#latest-156701)  (<https://www.kaggle.com/c/train-occupancy-prediction/discussion/27828#latest-156701>)) as a training and test csv and there is a station location csv). Note that there are three occupancy outcomes, low, medium and high, so you are going to have a three-way confusion matrix. Can you create an app that would help transportation planners do a better job planning the system? [This](https://github.com/GillesVandewiele/KaggleTrainOccupancy/blob/master/line_info.csv)  ([https://github.com/GillesVandewiele/KaggleTrainOccupancy/blob/master/line\\_info.csv](https://github.com/GillesVandewiele/KaggleTrainOccupancy/blob/master/line_info.csv)) table shows which stations are on which lines.

Final Rubric		
Criteria	Ratings	Pts
Use_Case_Motivation		2 pts
Exploratory_Analysis		4 pts
Model_And_Validation		4 pts
Code_Base_Replicability		1 pts
Conclusion		1 pts
Live_Presentation		3 pts
Video_Presentation		5 pts
		Total Points: 20