Capstone Requirements:

* Choose a Data Set for Capstone
* Write a half page proposal of what data set they will be looking at and what their focus will be
* Identify the business problem/idea they are analyzing
* Do exploratory data analysis
* Utilize Tableau and/or Excel to create visualizations and search for data insights
* Create a Model to support business idea
* Create a 5-7 minute presentation presenting a data insight

Notes

* Presentation and Tableau Visualizations should each be portfolio pieces
* Goal of presentations is to be comfortable talking about project to non technical audience

**Info for Final Capstone:**

During the week of December 9th, the class will spend time exploring the potential datasets for their final capstones (see list below). Each student must decide which dataset they want to work with for final capstone **before** Tuesday December 18th. Final capstone presentation are scheduled for Tuesday Jan 14th, with TTP graduation scheduled for Wednesday January 15th. We will also have a preliminary final capstone presentation, date TBD.

Please reach out with links to other specific datasets and topics you might have ideas about for capstone ASAP to discuss. Slack me with any questions.

**Target length for Final Presentations: 5-7 minutes**.

*The datasets*

* [http://insideairbnb.com/get-the-data.html](https://slack-redir.net/link?url=http%3A%2F%2Finsideairbnb.com%2Fget-the-data.html) (must use cities outside of NYC)
* [https://www.kaggle.com/olistbr/brazilian-ecommerce](https://slack-redir.net/link?url=https%3A%2F%2Fwww.kaggle.com%2Folistbr%2Fbrazilian-ecommerce) (this is Capstone 3 dataset, see below)
* [https://github.com/albahnsen/ML\_RiskManagement/blob/master/datasets/cost\_sensitive\_classification\_churn.csv.zip](https://slack-redir.net/link?url=https%3A%2F%2Fgithub.com%2Falbahnsen%2FML_RiskManagement%2Fblob%2Fmaster%2Fdatasets%2Fcost_sensitive_classification_churn.csv.zip) (see also: [https://github.com/albahnsen/ML\_RiskManagement/blob/master/exercises/10\_CS\_Churn.ipynb](https://slack-redir.net/link?url=https%3A%2F%2Fgithub.com%2Falbahnsen%2FML_RiskManagement%2Fblob%2Fmaster%2Fexercises%2F10_CS_Churn.ipynb))
* [https://www.kaggle.com/c/ieee-fraud-detection/data](https://slack-redir.net/link?url=https%3A%2F%2Fwww.kaggle.com%2Fc%2Fieee-fraud-detection%2Fdata) (See #8 for better data)
* [https://www.kaggle.com/mlg-ulb/creditcardfraud](https://slack-redir.net/link?url=https%3A%2F%2Fwww.kaggle.com%2Fmlg-ulb%2Fcreditcardfraud) (See #8 for better data)
* [https://www.instacart.com/datasets/grocery-shopping-2017](https://slack-redir.net/link?url=https%3A%2F%2Fwww.instacart.com%2Fdatasets%2Fgrocery-shopping-2017) (===>see also: [https://www.kaggle.com/c/instacart-market-basket-analysis](https://slack-redir.net/link?url=https%3A%2F%2Fwww.kaggle.com%2Fc%2Finstacart-market-basket-analysis) )
* [https://www.kaggle.com/c/elo-merchant-category-recommendation/data](https://slack-redir.net/link?url=https%3A%2F%2Fwww.kaggle.com%2Fc%2Felo-merchant-category-recommendation%2Fdata)
* [https://www.kaggle.com/c/home-credit-default-risk/data](https://slack-redir.net/link?url=https%3A%2F%2Fwww.kaggle.com%2Fc%2Fhome-credit-default-risk%2Fdata)
* [https://www.kaggle.com/c/zillow-prize-1/overview](https://slack-redir.net/link?url=https%3A%2F%2Fwww.kaggle.com%2Fc%2Fzillow-prize-1%2Foverview) (data on characteristics of homes for sale)
* [https://www.kaggle.com/c/coupon-purchase-prediction/overview](https://slack-redir.net/link?url=https%3A%2F%2Fwww.kaggle.com%2Fc%2Fcoupon-purchase-prediction%2Foverview)
* [https://www.kaggle.com/c/kkbox-churn-prediction-challenge/overview](https://slack-redir.net/link?url=https%3A%2F%2Fwww.kaggle.com%2Fc%2Fkkbox-churn-prediction-challenge%2Foverview)

Note: datasets #4 and #5 are likely too anonymized for you to do your best to data work. If you were leaning towards those I'd like to suggest dataset #8 instead, where you need to predict whether or not someone will default on their home loan.

Note: You do have the ability to continue with the module 3 capstone dataset for your final capstone. HOWEVER, you'd need to teach yourself to do a customer lifetime value model using survival analysis and the Python *lifelines* package as part of your capstone. It's a STRETCH goal, but google-able, and doable in three weeks. (see, for example, [https://towardsdatascience.com/survival-analysis-intuition-implementation-in-python-504fde4fcf8e](https://slack-redir.net/link?url=https%3A%2F%2Ftowardsdatascience.com%2Fsurvival-analysis-intuition-implementation-in-python-504fde4fcf8e)[)](https://slack-redir.net/link?url=https%3A%2F%2Ftowardsdatascience.com%2Fsurvival-analysis-intuition-implementation-in-python-504fde4fcf8e)))

For the final capstone, pretend you are working for the company whose data you are studying. Your final presentation is a data analysis presented to **management**. Give us insights in your presentation. Tell us a story about our company. (Follow this article closely about what a data analyst does [https://hbr.org/2018/12/what-great-data-analysts-do-and-why-every-organization-needs-them](https://slack-redir.net/link?url=https%3A%2F%2Fhbr.org%2F2018%2F12%2Fwhat-great-data-analysts-do-and-why-every-organization-needs-them) )

You can use python or excel or sql as you desire to do extensive data exploration deep

dives, to find the story you want to tell.

You **must** use Tableau for visualizations.

Models: For your final capstone you must find some question that you want to try to model. Some of these datasets lead themselves to models, e.g. churn and fraud detection are naturally logistic regression models. Others, like airbnb and instacart, you will have to use your data exploration and business knowledge to choose out what you want to study. The kaggle-defined instacart task is to predict which previously purchased products will be in a user’s next order, but you should NOT answer that question because linear models are not powerful enough to answer it. For instacart, please choose another question to make a model on. (edited)