**Kaggle Competition - Starter**

**Introduction**

Welcome to your first week of work at the Center for Disease Control. Time to get to work!

Due to the recent epidemic of West Nile Virus in the Windy City, we've had the Department of Public Health set up a surveillance and control system. We're hoping it will let us learn something from the mosquito population as we collect data over time. Pesticides are a necessary evil in the fight for public health and safety, not to mention expensive! We need to derive an effective plan to deploy pesticides throughout the city, and that is **exactly** where you come in!

As it's your first week on the job, let's get your development environment set up. Amongst your orientation group, we'll need you to get started by doing the following exercises. Also, see Linda in HR about getting your benefits set up. We have a GREAT health plan!

Once again, welcome to the CDC. We have high expectations for you!

**Dataset**

The dataset, along with description, can be found here: <https://www.kaggle.com/c/predict-west-nile-virus/>.

**This is also where you will be submitting your code for evaluation**. We will be using the Kaggle Leaderboard to keep track of your score. The leaderboard uses roughly 30% of the dataset to score an AUC (Area Under Curve) metric. [You can read more about the scoring metric here](https://www.kaggle.com/wiki/AreaUnderCurve).

**Requirements**

Here are your orientation assignments for your first day on the job:

* Set up a GitHub repository.
* Explore the data.
* Brainstorm a project roadmap.
* Produce a model and make predictions for where the Department of Public Health should be spraying.
* Submit your predictions to Kaggle (you can submit these as many times as you want to Kaggle) and get an AUC score.
* Create a Trello board/a Google sheet with tickets assigned to individual members of your team to keep the project organized. (Please tag whoever in your team is taking the lead on any part)

**Deliverable**

**GitHub Repo**

1. Create a GitHub repository for the group. Each member should be added as a contributor.
2. Retrieve the dataset and upload it into a directory named assets.
3. Generate a .py or .ipynb file that imports the data available data.

**EDA**

1. Describe the data. What does it represent? What types are present? What does each data points' distribution look like? Discuss these questions, and your own, with your partners. Document your conclusions.
2. What kind of cleaning is needed? Document any potential issues that will need to be resolved.

**Note:** As you know, EDA is the single most important part of data science. This is where you should be spending most of your time. Knowing your data, and understanding the status of its integrity, is what makes or breaks a project.

**The Scientific Method** *Remember High School?*

1. Start up a new document and describe the following:

* What is our problem statement?
* What can we learn from the data in order to make an educated hypothesis?
* What is our hypothesis?

**Project Planning**

1. Define your deliverable - what is the end result?
2. Break that deliverable up into its components, and then go further down the rabbit hole until you have actionable items. Document these using a project managment tool (like Trello) and/or GitHub to track things getting done.
3. Begin deciding priorities for each task. These are subject to change, but it's good to get an initial consensus. Order these priorities however you would like.

**Modeling**

1. The goal is of course to build a model and make predictions that the city of Chicago can use when it decides where to spray pesticides! Your team should have a clean Jupyter Notebook that shows your EDA process, your modeling and predictions.
2. Conduct a cost-benefit analysis. This should include annual cost projections for various levels of pesticide coverage (cost) and the effect of these various levels of pesticide coverage (benefit). *(Hint: How would we quantify the benefit of pesticide spraying? To get "maximum benefit," what does that look like and how much does that cost? What if we cover less and therefore get a lower level of benefit?)*
3. Your final submission CSV should be in your GitHub repo.

**Presentation**

* Audience: You are presenting to members of the CDC. Some members of the audience will be biostatisticians and epidemiologists who will understand your models and metrics and will want more information. Others will be decision-makers, focusing almost exclusively on your cost-benefit analysis. Your job is to convince both groups of the best course of action in the same meeting and be able to answer questions that either group may ask.
* Length: 25 - 30 minutes (a rough guideline: 2 minute intro, 15-20 minutes on model, 8-10 minutes on cost-benefit analysis, 3 minute recommendations/conclusion)

**Your project is due at 9:00 a.m. on Friday, August 11.**

**Evaluation**

The Kaggle competition should be evaluated on the following 4 points:

1. **AUC Scoring**: A clear winning group will be determined based on the AUC Scoring performed by the Kaggle Leaderboard. This is not to say that the winning group's work was the best submission. Remember, just hitting a benchmark is not enough to determine success; the following points are just as important if not more so.
2. **Clearly documented observations**: Students should have some log, whether it is a markdown file, text file, or iPython notebook, describing the observations and decisions they made along the way. This should be submitted to your instructor prior to your final presentation.
3. **Code**: All of your models, pipelines, cleaning techniques, and transformations should be properly coded and documented. Syntax is important, as data scientists are often tasked with building products that will be collaborated upon or maintained by other engineers. It is also important that no mistakes were made while pipelining data. If any data points were corrupted, the results are useless.
4. **Presentation**: Your presentation is expected to be client facing. Describe your πdata and approach as if your client is in front of you. This includes explaining the decisions made, the means by which you evaluated your decisions, and visualizations to support the story you are telling. This is a storytelling exercise, so be sure to set up, explain, and summarize your work clearly.