waste water data analysis

Section one:

problem definition and planning

- Identify problem
- List projects deliverables
- Generate success factors
- Understand each resource and other limitations
- Put together appropriate team
- Create a plan

☑ **Travis Boltz:** ~~Problem definition and team formation~~ due February 22

Waste water is a becoming a significant problem for the Great Lakes Region, especially when untreated sewage makes its way into the natural ecosystem from combined sewage overflows (CSO's). Untreated sewage unbalances the natural nutrient cycle by loading large concentrations of nutrients, like nitrogen, in the water. Algae already present in the water use these nutrients to grow exponentially to the point that they become harmful algae blooms (HAB). A HAB is dangerous for the environment and human health and usually necessitates the shut-down of water treatment plants used for drinking water. This has massive consequences for the local economy and the sustainability of natural ecosystems. One of the ways to prevent theses algae blooms is to better understand how much nitrogen is present in the natural ecosystem. The metric used to measure the total amount of nitrogen in the water is called Total Kieldahl Nitrogen (TKN). This process requires a lab to analyze a sample of water to determine the TKN present. It is not always the case that TKN was captured in present or past data. To overcome this issue, we want to predict the level of TKN present in the water using other measurements that are present in the dataset.

The team consists of the following 4 individuals:
-Travis Boltz
-Yitao Ma
-Yue Ning
-Yuqing Zhao

☑ ~~Team information~~ due February 22

Members:
Travis Boltz
Yitao Ma
Yue Ning
Yuqing Zhao

☑ **Yitao:** ~~Project proposal~~ due March 1

- Identify problem
- List projects deliverables
- Generate success factors
- Understand each resource and other limitations
- Put together appropriate team
- Create a plan

☑ **yuqing zhao:** ~~Project plan~~ due March 8

✓ Project plan

Section two:

Data preparation

- Access and combine data tables
- Summarize data
- Look for errors
- Transform data
- Segment data

☑ **Yue Ning:** ~~Data collection~~ due March 8

Raw influents data for 8 locations(2011-2020) from:
http://www.mwrd.org/irj/portal/anonymous?NavigationTarget=navurl://9f766d4f820e9482d016681c86031b76
Rainfall data: https://drive.google.com/file/d/1BHuq89bgyt7kC_Paf1_CjnWiG4uBgm2d/view

☐ Data combining due March 14

Get a table that contain all data.

☐ **Yitao:** Join data table for different attributes on date. due March 12

□ **yuqing zhao:** Concatenate date table of different years and locations. due March 14

□ Summary statistics due March 17

Get summary statistics about the whole data set.
Scattering plots will be performed for each attribute at different locations and years.

□ **Yitao:** A table of summary of each attribute. due March 15

□ **yuqing zhao:** Bar plots for each attribute. due March 16

□ **Yue Ning:** Correlation plot containing each attribute. due March 17

□ **Yue Ning:** Correlation matrix. due March 17

□ **Travis Boltz:** Data cleaning due March 20

Fix errors and remove or impute missing values.

□ **Travis Boltz:** Remove attributes depend on domain expertise. due March 18

□ **Travis Boltz:** Remove attributes showing strong correlations. due March 18

□ **Travis Boltz:** Remove attributes that have a majority of null values due March 19

□ **Travis Boltz:** Use multiple imputation for any missing values at random or missing values completely at random due March 20

Section three:

data preparation

- Summarizing data
- Exploring relationships between attributes
- Grouping the data
- Identifying non-trivial facts, patterns and trends
- Building regression models
- Building classification models

□ **Yitao:** ANOVA-test due March 23

Use ANOVA-test p-value approach to compare data for TKN (target variable) from the 8 waste water treatment plants to see if the means are equal. If the means are not equal, we will use dummy variables to differentiate between the waste water treatment plants.

□ **Yitao:** Demographic analysis due March 26

Our main target is to see the relationship between pollutants and population distribution.

Location of water plants
- Calumet WRP: 400 East 130th Street, Chicago, IL 60628.
- Stickney WRP: 6001 West Pershing Road, Cicero, IL 60804.
- O'Brien: WRP: 3500 Howard Street, Skokie, IL 60076.
- Kirie WRP: 701 Oakton Street, Des Plaines, IL 60018.
- Egan WRP: 550 South Meacham Road.
- Hanover Park WRP: 200 Sycamore Avenue Hanover Park, IL 60133.
- Lemont WRP: .

Rough steps
- Use `demography` package in R to draw a map of Chicago based on population distribution density.
- Mark locations of water plants on the map and the result will directly show us relationship.

□ **Yitao:** Graph of population density with locations labeled due March 25

□ Regression models

Use 6 different regression models to predict the total Nitrogen concentration. The general linear regression model will serve as the baseline.

□ **Yitao:** Validation methods due March 30

Apply both methods to general linear models to compare and decide which validation method to use

□ Cross-validation

Use 10-fold cross-validation technique to avoid over-fitting of the models.

□ Conventional validation

Use stratified sampling method to take 80% data as training data set and 20% as testing data set.

- **Yue Ning:** General linear regression model with testing on hold-out data. due April 1
- **Yue Ning:** General linear regression model with cross-validation. due April 1
- **Travis Boltz:** Principal components analysis (PCA) due April 2
- **yuqing zhao:** Lasso regression model due April 4
- **yuqing zhao:** Ridge regression model due April 4
- **yuqing zhao:** ElasticNet Regression model due April 4
- **yuqing zhao:** Best subset regression model due April 4

Classification model

This might not be a suitable model for describing our data, but we will see the result first.
Use 3 classification models to predict low, medium and high level of Nitrogen.

- Decision tree model
- Linear classifier-Linear Discriminant Analysis(LDA)
- Random forest model

Model evaluation due April 6

Adjusted R^2 and MSE will be used to evaluate our regression model.
Recall, precision and F1 Score will be used to evaluate the classification models(if used).

- **yuqing zhao:** Adjusted R^2 for evaluating regression models. due April 6
- **Travis Boltz:** Test MSE for evaluating regression models. due April 6
- **Yue Ning:** Recall, precision and F1 Score for evaluating classification models (if used) due April 6
- **Travis Boltz:** Use Normal Q-Q Plot to determine if residuals are normally distributed due April 5
- **Travis Boltz:** Use Scale-Location plot to check for equal variance. If there is an indication that there is unequal variance possibly perform a log transform on the target variable and primary dependent variables to correct for heteroscedasticity. due April 5
- **Yue Ning:** Use Residuals vs Leverage to check to see any points have a disproportionate influence. We will examine those points to determine if it is an outlier or due April 5
- **yuqing zhao:** Residuals vs Fitted shows if residuals have a non-linear pattern. due April 5

Section four:

Deployment

● Generate report

Project presentation due April 26

A project presentation will be delivered in class. 18 min presentation with about 20 slides.

- **yuqing zhao:** presentation part 1(data preparation) due April 26
- **Yitao:** presentation part 2(data analysis) due April 26
- **Travis Boltz:** presentation part 3(introduction and summary) due April 26

Project report due May 3

8 pages report including analysis and figures.
Code:
https://github.com/Qingzz7/18S571project

- **Yitao:** Introduction due April 30
- **Yue Ning:** Data preparation due May 1
- **Travis Boltz:** Data analysis due May 2
- □ **yuqing zhao:** Summary due May 3