MGT 6203 Group Project Proposal Template

**Please edit the following template to record your responses and provide details on your project plan.**

**TEAM INFORMATION (1 point)**

**Team #:** 55

**Team Members:**

*[Insert background information: Name, professional background, education background, previous*  *analytics related projects you have worked on]*

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Profession background** | **Education background** | **Previous analytics related projects you have worked on** |
| Hayden  GT id: hblackburn3 | 2 years experience as financial data analyst and 2 years experience data quality engineer for large manufacturing company | BS in Finance and Marketing | Multiple data cleaning projects utilizing python to automate work or clean raw data, linear regression on premier league data |
| Richard Li  GT id: rli465 | 5 years of experience working in semiconductor manufacturing, currently working for a data science interviewing startup | BS in Industrial and Systems Engineering | Have worked on a forecasting model for semiconductor wafers going through the production line. |
| Binh Vu  GT id: bvu38 | 2 years of experience working in aerospace/manufacturing industry | background in Mechanical Engineering/CS | Courses projects: predicting housing prices and card game simulation. |
| Lijia Cheng  GT id: lcheng94 | Experienced public accounting manager specializing in financial statements and internal controls audit of companies | BBA (Accountancy) from NUS  Chartered Accountant (Singapore)  Financial Risk Manager (GARP) | Using analytics to tackle crowd management problems |
| Alexandra  GT id: aprokhorova3 | Risk analyst in a tech consulting company | B.A. in Linguistics  B.S. in Electrical Engineering  Data Science Bootcamp | **Predicting Business Success:** Built a [classification model](https://github.com/sasha-talks-tech/yelp-classification) predicting restaurant star ratings based on foot traffic and attributes.  **Yelp Review Data:** Performed NLP [analysis](https://github.com/sasha-talks-tech/NLP_Yelp) on Yelp restaurant reviews. Created a scalable, production-ready automated text pre-processing pipeline using multiple years of user-generated data in order to enable topic modeling and entity detection workflows.  **Oracle Hackathon at Data Science Go Conference:** Trained a regression model to predict air quality using data on chemical compound concentration, accompanied with visualizations. Our team earned an award for the best insights.  **UC Berkeley Hackathon at TextXD Conference:** Trained an [NLP model](https://github.com/sasha-talks-tech/police-misconduct) to analyze data on crimes committed by police officers in California. Analysis was accompanied with data visualizations and narrative.  **Women Through the Lens of the New York Times:** Completed NLP-centric data journalism [research](https://towardsdatascience.com/strong-women-through-the-lens-of-the-new-york-times-f7f7468a2645) on gender equality and representation in The New York Times, by using topic modeling, frequent term visualization and sentiment analysis. |

**OBJECTIVE/PROBLEM (5 points)**

**Project Title:** Investigating the impact of certain environmental conditions on salmon fish runs in Kenai River (Alaska)

**Background Information on chosen project topic:**

One of the most iconic industries in Alaska is salmon fishing, which has been a cornerstone of the state's economy and culture for generations. Every year, millions of salmon swim up the state's rivers and streams to spawn, drawing in fishermen from around the world participate in the fishing opportunities that Alaska has to offer.

One of our team member’s family is from Alaska and they fish there regularly. The fish and game department uses sonar technology to track the incoming salmon and post daily counts after each day. It appears random how one day 1,000 salmon will run through the sonar buoys and on the next 90,000 will run. We want to find out if we can through regression, \_\_\_\_\_\_\_\_, \_\_\_\_\_\_\_\_\_ understand and find relationships between weather data and the fish run counts. We will use Alaska Fish & Game’s data along with a dataset for weather (temperature, weather conditions: overcast/sunny/stormy, water temperature, (tide data, moon phase?), and precipitation) to see if we can predict the best conditions for salmon to make them run up the river and the amount that will run.

Climate change is expected to have a significant impact on the timing and abundance of salmon runs, as changes in water temperature, river flows, and precipitation patterns can all affect the spawning and migration of salmon. Weather patterns such as storms and droughts can also have a significant impact on salmon populations, both directly and indirectly through changes in water levels and temperatures.

**Problem Statement (clear and concise statement explaining purpose of your analysis and investigation):**

Investigate the impact of different conditions (temperature, weather conditions: overcast/sunny/stormy, water temperature, (tide data, moon phase?) on the number of salmon that will run at XX in Alaska

are most significant for salmon to make a run of (over x amount)?

**State your Primary Research Question (RQ):**

What is the impact of certain conditions (temperature, weather conditions: overcast/sunny/stormy, water temperature) on fish count?

**Add some possible Supporting Research Questions (2-4 RQs that support problem statement):**

1. Does tide or moon phase have any effect on the fish count?
2. Can we predict the number of salmon that will run on a given day?
3. Does rain/location/species have any effect on the fish count?

**Business Justification:** **(Why is this problem interesting to solve from a business viewpoint? Try to quantify the financial, marketing or operational aspects and implications of this problem, as if you were running a company, non-profit organization, city or government that is encountering this problem.)**

Understanding the factors that influence salmon populations is critical to the sustainable management of Alaska’s fisheries and ensures both commercial fishing and tourism industries can continue to thrive. There are several groups of people that would benefit from accurate fish count predictions, including:

* Alaska Fish and Game (city/government) - make available data that is the most relevant to locals and visitors; using predictive fish count data, local city/government could designate certain time of the year as the fishing season vs not to prevent overfishing and maintain sustainable fish populations
* Local fisheries – allow fisheries to plan for fishing on days with the best predicted fishing conditions, assuming fisheries do not go out fishing every day with limited manpower and resources
* Local businesses selling fishing-related products or tourism-related businesses (accommodations, restaurants) - allow local businesses to better predict high and low seasons to better manage supply to meet anticipated demand
* Visitors - allow visitors to make advance plans for when they will visit Alaska based on probabilities for fish run on a given day; This would help tourists not get burnt out fishing all day and not catching anything thus, making their experience more enjoyable and create a chance they will return to Alaska in the future.

**DATASET/PLAN FOR DATA (4 points)**

**Data Sources (links, attachments, etc.):**

* Fish count data: <https://www.adfg.alaska.gov/sf/FishCounts/index.cfm?ADFG=main.displayResults&COUNTLOCATIONID=40&SpeciesID=420>
* Weather data (Air temperature, precipitation): <https://akclimate.org/data/data-portal/>
* Water temperature data
* Tide / Moon phase data: [**https://www.usharbors.com/harbor/alaska/kenai-city-pier-ak/tides/#:~:text=Tides%20in%20Kenai%20City%20Pier%2C%20AK%20for%20Today%20%26%20Tomorrow&text=Next%20high%20tide%20is%202,tomorrow%20is%208%3A31%20AM**](https://www.usharbors.com/harbor/alaska/kenai-city-pier-ak/tides/%23:~:text=Tides%2520in%2520Kenai%2520City%2520Pier,%2520AK%2520for%2520Today%2520&%2520Tomorrow&text=Next%2520high%2520tide%2520is%25202,tomorrow%2520is%25208:31%2520AM)

**Data Description (describe each of your data sources, include screenshots of a few rows of data):**

Data dictionary?

Example:



**Key Variables: (which ones will be considered independent and dependent? Are you going to create new variables?** **What variables do you hypothesize beforehand to be most important?)**

Dependent: Number of fish detected

Independent:

* Precipitation amount
* Weather conditions: sunny, cloudy, stormy
* Water temperature
* Air temperature: daily min/max
* Water level
* Water flow
* Tide data? Not sure yet what the factors from this would be – tides work with moon phases so perhaps the factor is which moon phase? Would be cool if that ended up having a significant relationship

**APPROACH/METHODOLOGY (8 points)**

**Planned Approach (In paragraph(s), describe the approach you will take and what are the models you will try to use? Mention any data transformations that would need to happen. How do you plan to compare your models? How do you plan to train and optimize your model hyper-parameters?))**

Data prep:

* Join fish count data with weather data
* Clean data & handle missing values
  + Transform categorial data into dummy variables
* Exploratory data analysis
  + Check for correlation/multi-collinearity
  + Distribution
  + Visualize data – scatter plots/histograms, etc

Potential models:

* Regression Model:
  + Linear Regression, random forest regression, Ridge regression (high correlation), Lasso ---> Fish Population
* Optimization:
  + Cross-validation
* Clustering
  + Find out if there are data points that are grouped together and further analyze what’s similar or different about these clusters
* CUSUM
  + Change detection to find out how many runs in a season

Look at localize temperature vs localize fish populations

**Anticipated Conclusions/Hypothesis (what results do you expect, how will you approach lead you to determining the final conclusion of your analysis) Note: At the end of the project, you do not have to be correct or have acceptable accuracy, the purpose is to walk us through an analysis that gives the reader insight into the conclusion regarding your objective/problem statement**

**What business decisions will be impacted by the results of your analysis? What could be some benefits?**

* Fishing industry is key driver of tourism in Alaska, can use this to predict number of tourists coming to alaska

**PROJECT TIMELINE/PLANNING (2 points)**

**Project Timeline/Mention key dates you hope to achieve certain milestones by:**

**Appendix (any preliminary figures or charts that you would like to include):**