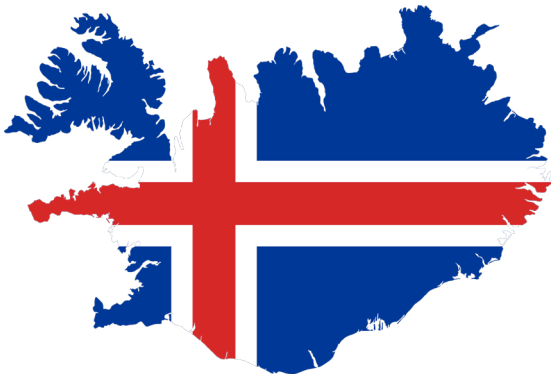
Iceland Fisheries Monitoring

Final Report

February 2023

Team
Joseph Hancuch
Sean Johnson
Daniel Noel
Dylon Polcik
Mark Schumacher

Table of Contents



Executive Summary

Overview

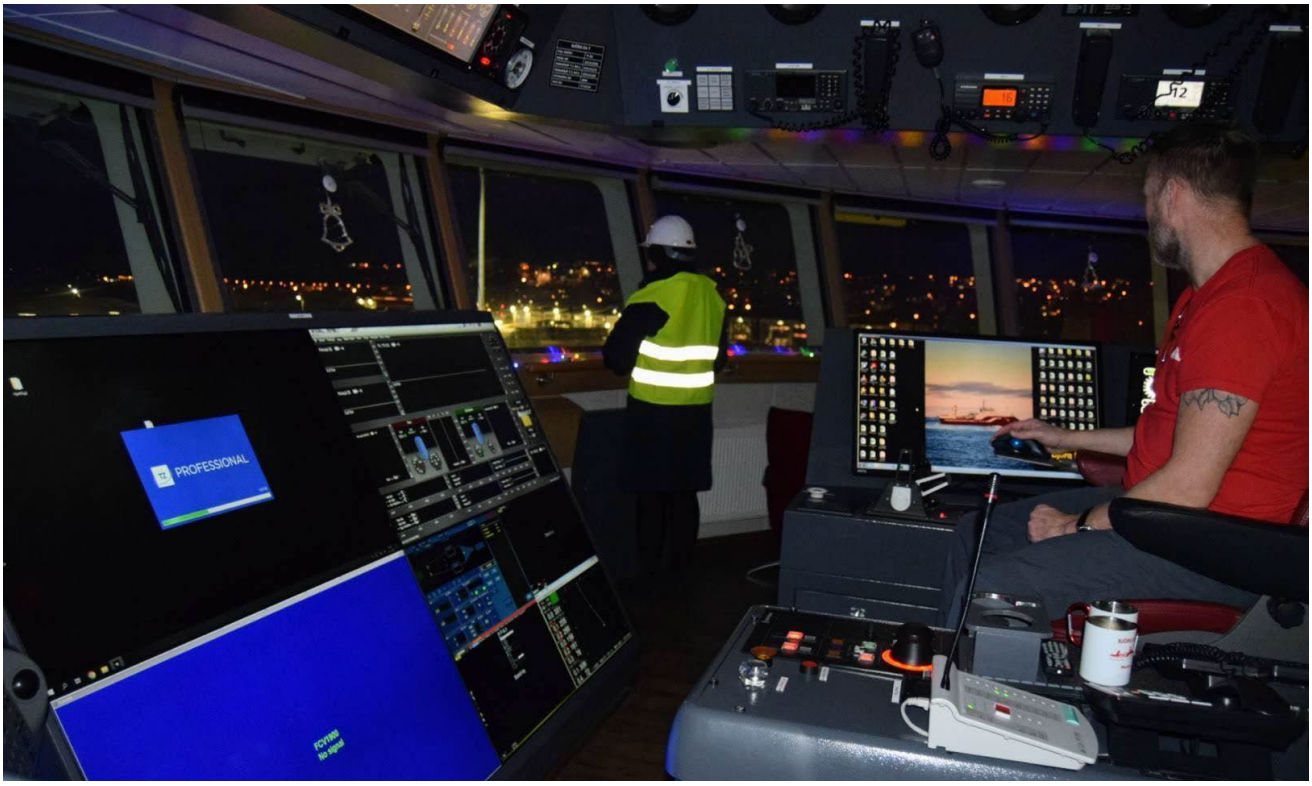
The Icelandic seafood industry exists among fiercely competitive foreign markets where sustainability certification requirements are key to having access to the best markets. Industry stakeholders including wholesale buyers, investors, and consumers will pay increased attention not only to the management and conduct of actors using the resources, but to the sustainable use of the resources themselves. There is a lot of work to be done for Icelandic fisheries to be a global leader in ensuring and demonstrating responsible management of the resource as all production components are electronically traceable, thus creating a competitive advantage over other nations.

There are many ways to ensure proof of responsible behavior of actors using the resources. Two main approaches are available: a labor-intensive traditional route that entails government inspectors being onboard in a high proportion of fishing trips (a path taken in many parts of the world), or, a path involving the automation of surveillance. The traditional approach is costly and requires doubling or tripling the number of inspectors. Even with an increase in inspectors, coverage of the inspections would remain low. Instead, the industry should adopt an approach that involves the automation of monitoring of the handling and processing of marine catches with access to near real-time data. By focusing more on automation and, ideally, by using data generated by the industry, more comprehensive and cost-efficient surveillance could be built that ensures better monitoring of the value chain. By doing that, the industry could take large steps towards the UN targets, (12,13,14) by demonstrating responsible behavior in the management of the resource. This will help producers to meet increased demands on European markets, and some kind of certification such as a Digital Product Passport (https://gs1.eu/activities/digital-product-passport/).

It is likely that the responsibility to prove sustainable management of fisheries will be shifted from public bodies (Fiskistofa) to fisheries companies. This is in line with developments in food production. Following the Escherichia coli breakout in the US and the Bovine Spongiform Encephalopathy (BSE) in Western Europe in the early 1990's and the resulting erosion of public trust in food safety systems, there was worldwide adoption of the Hazard Analysis Critical Control Point (HACCP). HACCP implemented a quality certification system based on reversing the burden of proof from government agencies to producers.

Business Case

The path forward to ensuring the sustainability of Iceland fisheries needs to be driven by technology. Data collected in the Icelandic fisheries sector combined with data science techniques could be used to strengthen monitoring of seafood and thus strengthen the competitive position of the Icelandic fisheries sector. As a first step toward shifting the burden of proof to the industry, there is an opportunity for Fiskistofa to demonstrate the use of publicly available landings data combined with data science techniques to assess risk and control surveillance. Since 2008, the number of staff in the Fiskistofa Surveillance Department has been reduced from 43 to its current staff of 28, including only 18 field inspectors. By implementing a technology driven approach, Fiskistofa will increase both its coverage and effectiveness.



Goals

The aim of this project is to develop tools for evaluating the risk of discard by irresponsible parties, and to deploy on-site inspectors in a more focused manner. The focus of this capstone project will be on analyzing landings data and examining the relationships between various data elements, including the market prices of non-targeted species and the presence of inspectors.

The Approach

Data Sources

We have received two large databases in csv format from Fiskistofa that are the source of the data for our analysis. The first database contains records of 286,701 landings of fishing vessels in Iceland from January 1, 2017, through December 2022 for a total of 1,249,661 rows total.

The second database is a record of all financial transactions and pricing involving the Icelandic Fishing Industry. The pricing database contains 1,566,399 transactions involving trades of fish between financial principles involving Icelandic fish.

A “landing” in the fishing industry is the part of the fish catch that is brought to shore. A problem in the fishing industry is that landings may represent the only record of a catch, as part of a catch can be discarded at sea and will go unrecorded.

The landings database contains several supplementary files that will we use to develop data models which include:

- Records of all vessel landings that received surveillance for Fiskistofa inspectors
- Mappings of all recognized terminology used to describe Fish species
- Mappings of all recognized terminology used to describe ShipTypes and Gear used in the Fishing industry
- Mappings of all Fish species to taxonomies
- Details related to how the catch totals apply to specific fishing quotas

The pricing database contains specific transaction information related to purchase of fish from Icelandic fishing companies:

- The dates of and amounts in Icelandic Krona of each transaction between fish companies and fish processors.
- The contracts related to landings which relate to specific fish species and landingsIds

Data Cleaning and Validation

To make the data digestible by data science tools, we have spent significant time validating the two large databases to confirm that values will become the basis for a valid data model. We have taken many steps to “clean” the data and produce viable reports:

- We learned the number of records exceeds Excel’s capacity, so we only use Excel on smaller record sets.
- We have figured out the coding signatures required for the extended character sets of the Icelandic alphabet.
- We ran into problems with comma as a delimiter when it is used as a decimal and within company names. A better delimiter would be a character that does not naturally occur in the dataset (such as a caret (^) or sharp (#)).
- In some fields we experienced very high negative values or other values that do not correspond with logical values given our understanding of the data field. We have thrown out a very small number of records that do not comply with any logical understanding of the data. Long term, we may make some data governance recommendations based on these extreme values.

- Because of the significance of species for the overall problem of eliminating discard, we have created mappings for a large number of alternate species names that appear in the landings table, but do not have a corresponding name in the Fiskistofa look up tables.
- Similarly, the surveillance landing table did not map cleanly into the landings table leading to a challenge with inspections with multiple landings and quite a few missing landings.
- We continue to confront issues related to the diversity of data and what data we can safely exclude from our analysis. For example, the data contains conventions for testing vessels, ship types and gear. A large amount of “sport fishing” falls under the domain of the Fiskistofa, but is not relevant for our analysis. We recently noticed issues related to Iceland’s whaling fleet which we managed to exclude by removing them from the analysis.

It is important to note that we decided to focus on the landings data and avoid the complexity of working with Iceland's quota system. While the quota system is a crucial factor in the fishing industry, it would have required more time than the ten weeks allowed for this project to fully understand and incorporate it into our analysis. By focusing on the landings data, we were able to complete our analysis within the time constraints, while still capturing the relevant information for our study.

As part of this process, we have completed a document titled Data Governance Notes which records our experiences working with the data. While overall the data is in very good shape, these notes document a few of the stumbling blocks and this Data Governance document might be helpful for anyone performing similar work in the future.

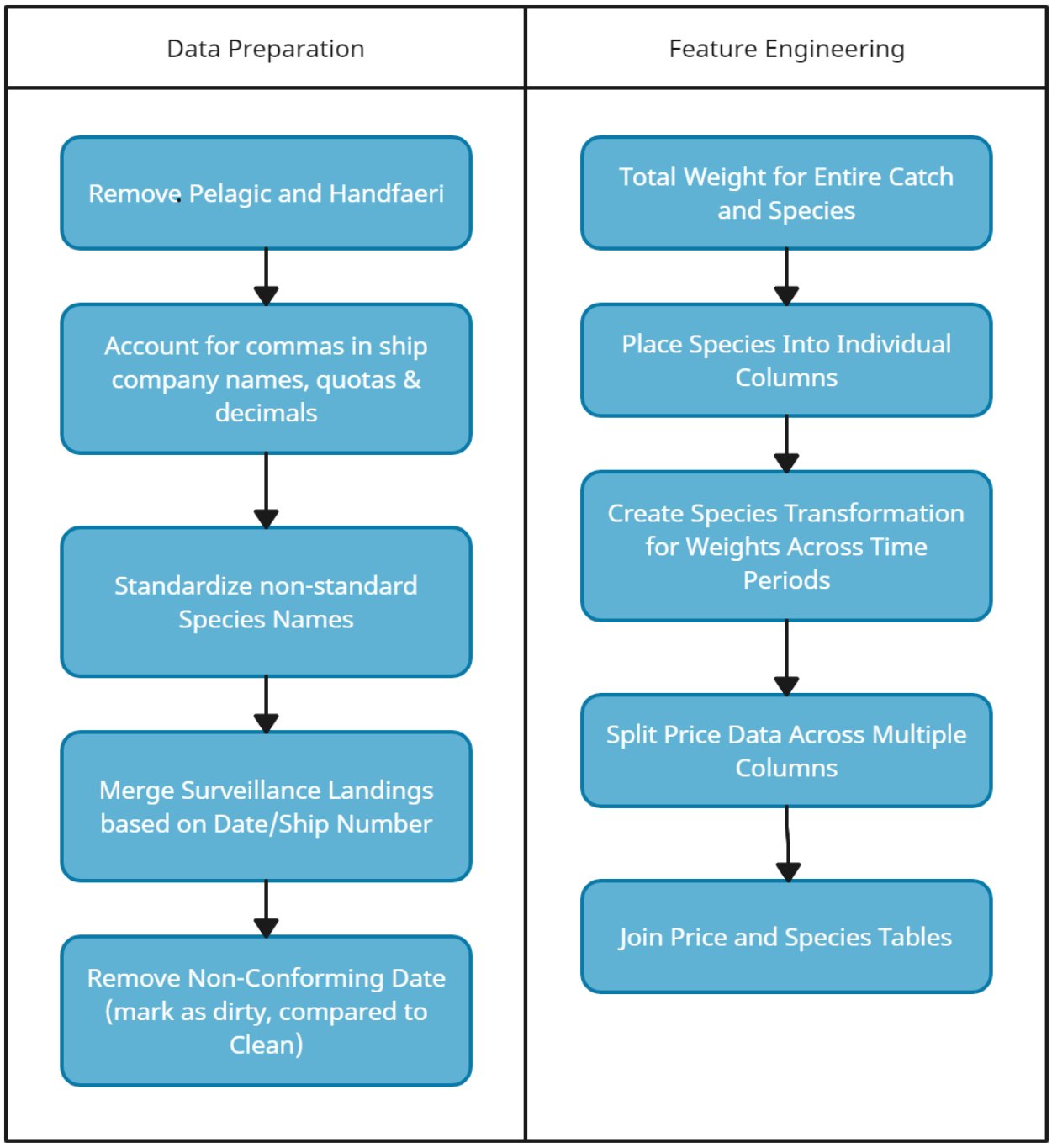
Data Exclusions

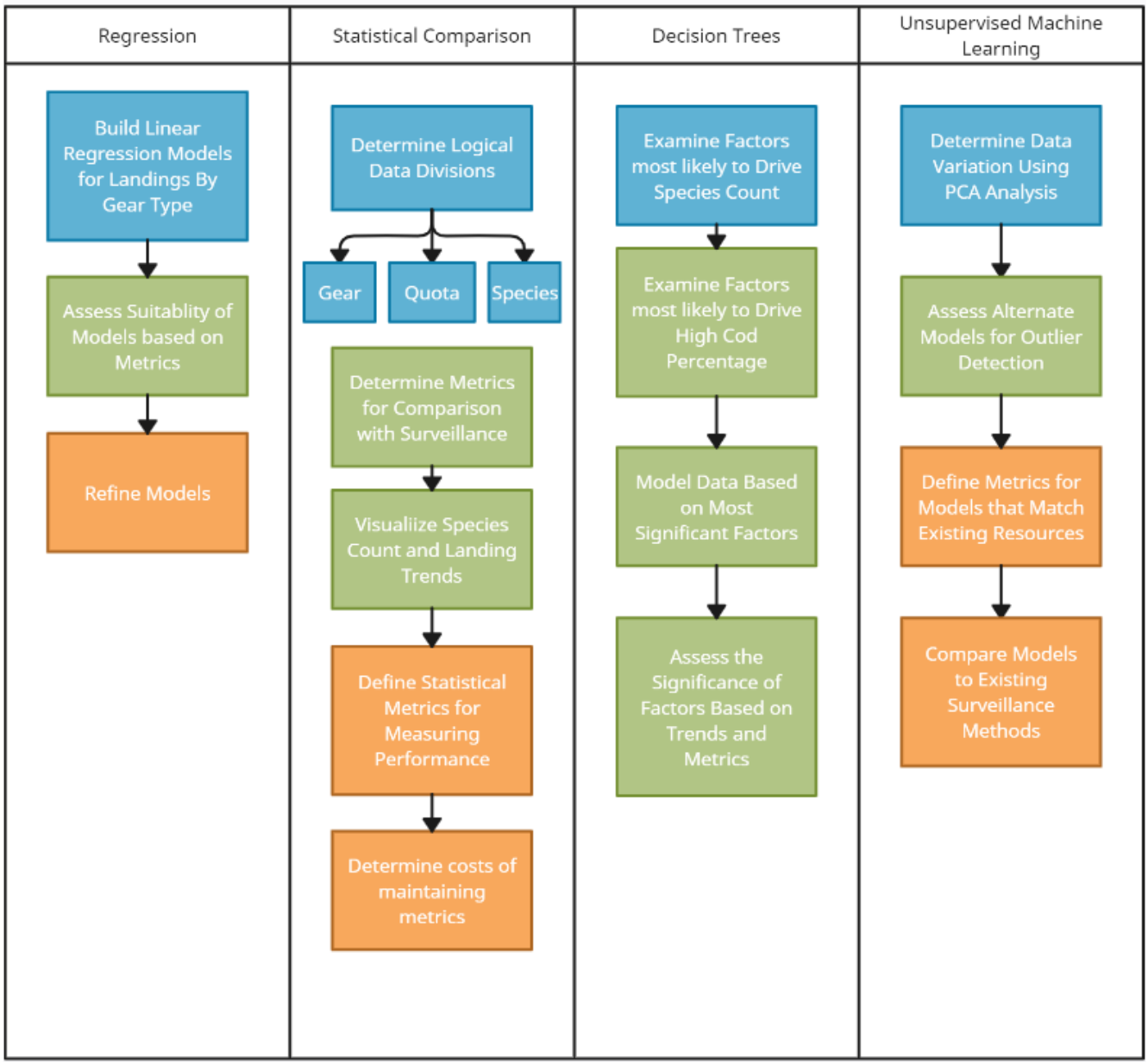
The main focus of this project is on commercial cod landings. Landings data related to recreation and sports fishing as well as Pelagic fishing has been excluded from the dataset that will be used for this study.

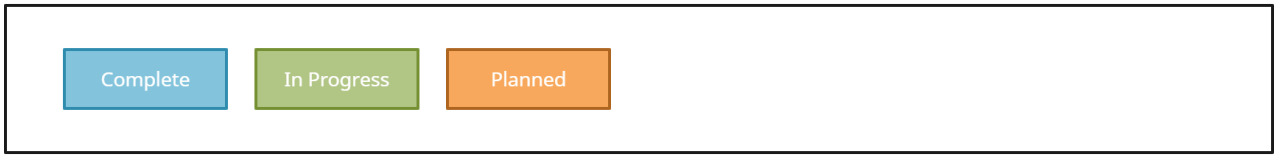
Pelagic fishing in Iceland involves use of different gear targets and different species than the rest of the industry. Pelagic fishing practices were identified in the data based on Nót and Flotvarpa gear types and were removed from the dataset.

Landings related to Handfæri gear type were also excluded.

Model Process Flow




Regression
Statistical Comparison
Decision Trees
Unsupervised Machine Learning
Build Linear Regression Models for Landings By Gear Type
Assess Suitablity of Models based on Metrics
Refine Models
Determine Logical Data Divisions
Gear
Quota
Species
Determine Metrics for Comparison with Surveillance
Visualiize Species Count and Landing Trends
Define Statistical Metrics for Measuring Performance
Determine costs of maintaining metrics
Examine Factors most likely to Drive Species Count
Examine Factors most likely to Drive High Cod Percentage
Model Data Based on Most Significant Factors
Assess the Significance of Factors Based on Trends and Metrics
Determine Data Variation Using PCA Analysis
Assess Alternate Models for Outlier Detection
Define Metrics for Models that Match Existing Resources
Compare Models to Existing Surveillance Methods


Complete
In Progress
Planned

Analysis of Data

Exploratory Data Analysis

An exploratory data analysis (EDA) was conducted on both the landings and pricing data. The purpose of an EDA is to identify errors in the data and assess the level of data cleaning necessary, start to understand patterns, detect outliers, and develop some initial insights that help to inform which variables to include in modeling.

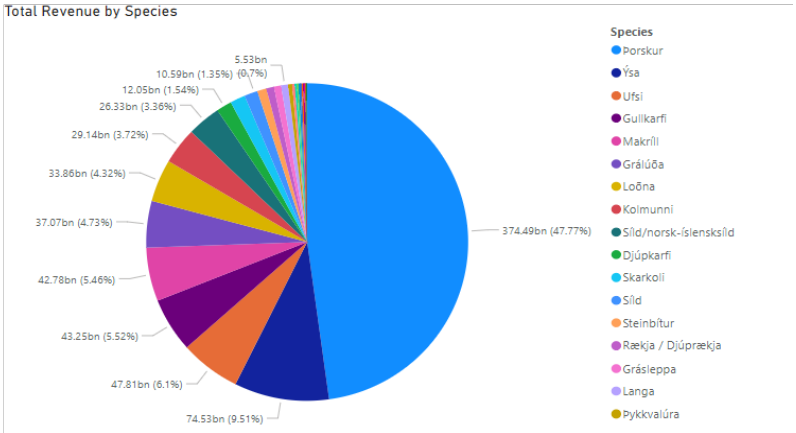
General Observations

- **The Iceland seafood market is dominated by a few species, particularly, the þorskur (Atlantic cod)**, a fast growing, omnivorous ground fish. Þorskur make up the largest percentage of landings weight as well as revenue in the landings and pricing databases. It is important to understand the relationship of the percentage of cod in a catch to the percentages for other species.
- **Similar gear types should result in catches of similar characteristics across different vessel types.** The most common gear types used by the Icelandic fishing industry are Botnvarpa and Lina.
- **The metric to start with in understanding discard behavior appears to be the cod percentage of total landing.** The assumption would be that as the cod percentage increases (and percentages of other species decreases), the existence of risk of discard increases.
- **The presence of inspectors drives significantly different cod percentages in landings.**
- **Dot plots have proven to be an effective way of illustrating species count trends in landings.** The trend from 2017 to 2022 is toward more species represented in landings.

Cod Percentage of Landing as a Benchmark

Although the overall eco-system of Iceland's fisheries includes over 60 species, the market for Icelandic seafood is dominated by a relatively small number of species. From 2017 to 2022, Porskur (Cod) generated 47.8 % of the total revenue. The top 7 species comprise 83.4 % of total revenue (see Figure 1 below).

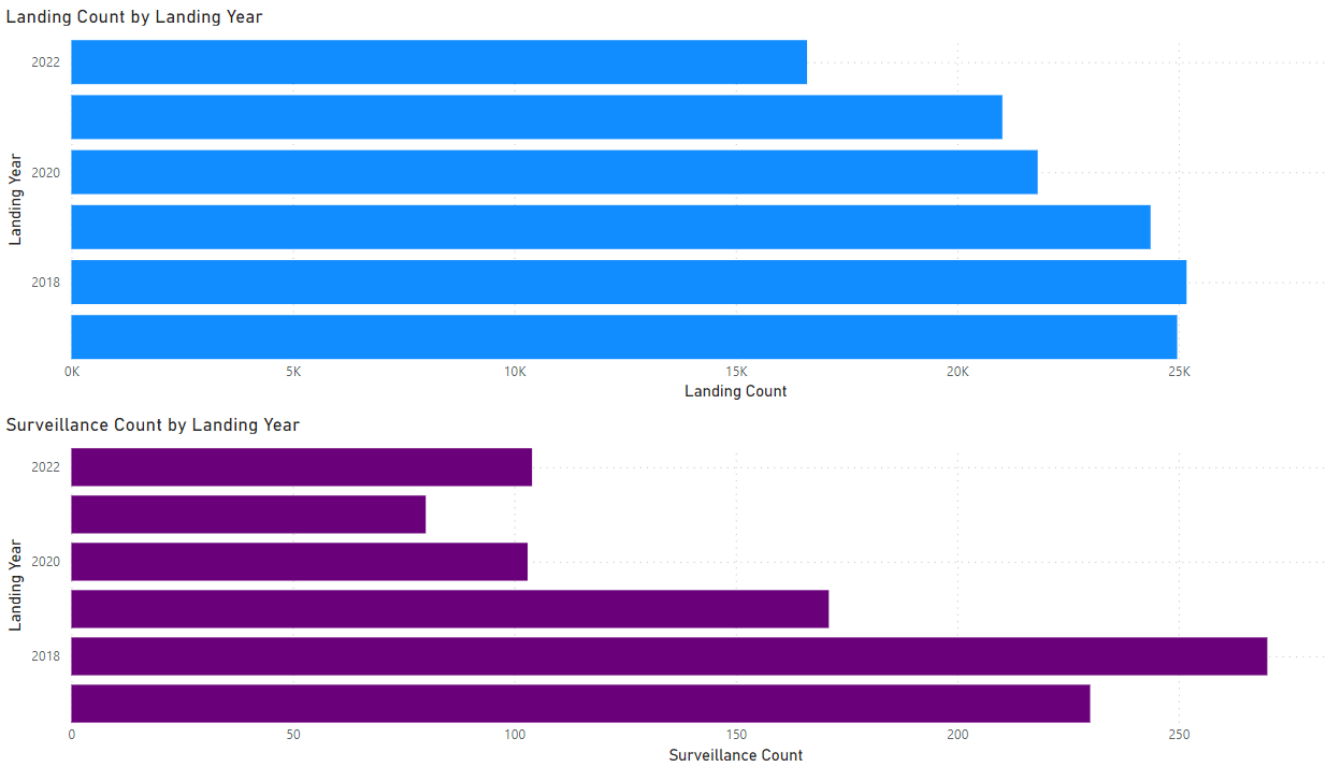
Figure 1 - Total Revenue by Species – 2017-2022



Impact of Inspections

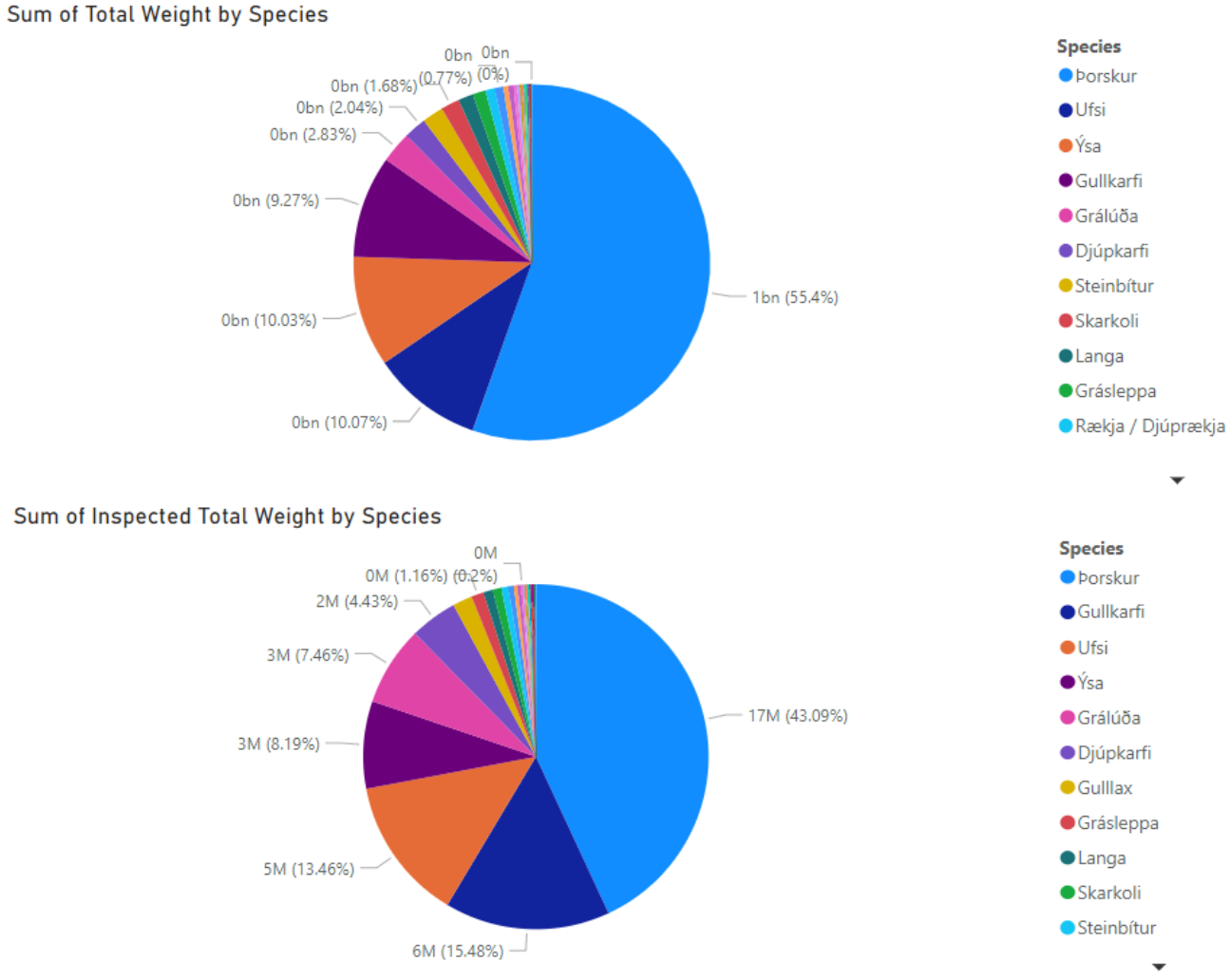
The landings database from Fiskistofa (the fishing Directorate of Iceland) contains thousands of landings from 2017 until late 2022. After removing many landings such as recreation and sports fishing from our dataset, we have 133,936 landings per year with an average of 22,322 per year. Of these landings, the total number of inspected landings is 1,333. As the chart below indicates, the number of inspected landings has dropped sharply from a peak of 341 in 2018 to a pandemic low of 147 in 2020. In this data sample the ratio of inspected landings to overall landings is 1 out of every 104.8 landings. Further complicating our analysis was the table structure used to manage inspections which did not contain landingIds for each inspection, but instead relied on LandingDate as a way of mapping these values. Consequently, we were not able to manage a large number of surveillance landings which was particularly relevant for certain gear types.

Figure 2 - Inspections Compared to Landings – 2017-2022



The following Pie charts show the division of Landings Weight by Species for all the landings in Iceland compared to a pie chart of the comparable vessels that were inspected during the same period (Figure 3).

Figure 3 - Species Variety in Landings, Non-Inspected vs. Inspected 2017-2022

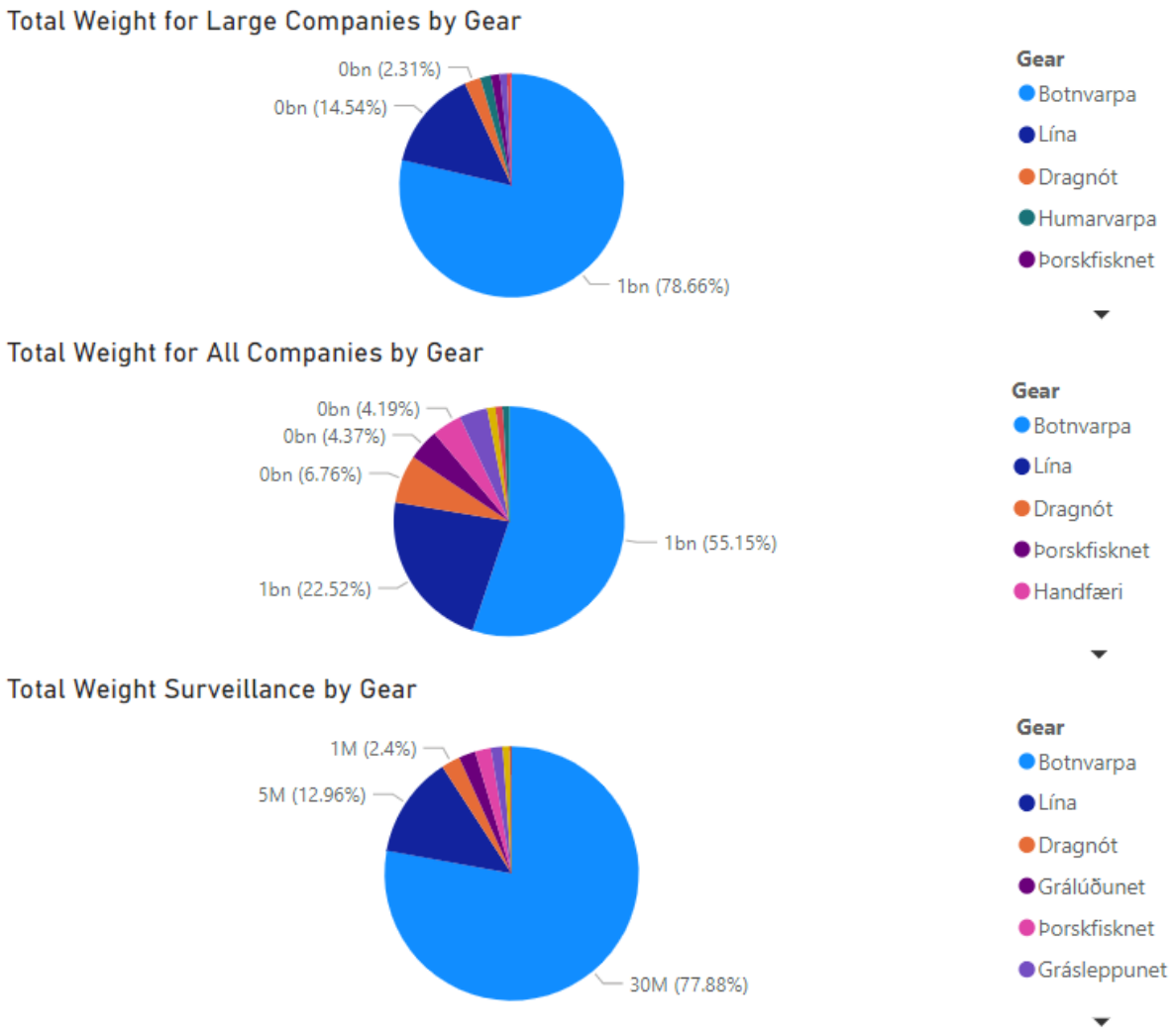


These pie charts illustrate a fundamental problem which is the purpose of this examination: **the presence of inspectors leads to different catch totals for different fish.** In the case of our data, the overall proportion of Cod represented in the total landing weight shrank from 55.4% without inspectors to 43.1% when inspectors are present.

The explanation for the difference in observed species count between inspected species counts and actual counts is an age-old problem known as discard. Fisherman, regularly discard (or throw back) those species that are not desirable. Many papers have been written on the subject, but no method has managed to reliably determine the actual numbers – though conventional estimates assume discard is anywhere from 5% to 15% globally.

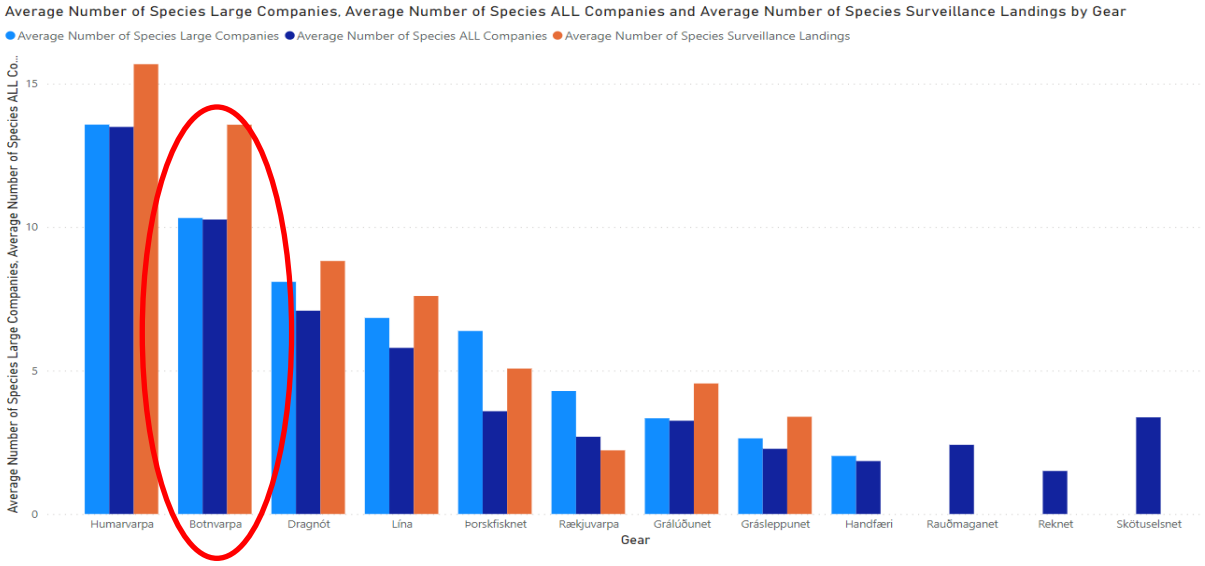
The Icelandic Fish Directorate is attempting to correct this behavior through many measures including surveillance inspection, drones, public information campaigns, and accurate data modeling. The most important aspect of the type of species that should be present in any landing is the gear type that is used by the vessel. The most experienced members of Fiskistofa believe that similar gear type should result in catches of similar characteristics across different vessel types. The following diagrams show the most common gear types used by the Icelandic fishing industry are Botnvarpa and Lina (Figure 4):

Figure 4 - Total Landings Weight by Gear Type 2017-2022



For the Botnvarpa gear type, inspected landings contain an average of 13.57 different species compared to an average of 10.32 species for all landings (see circled section of Figure 5).

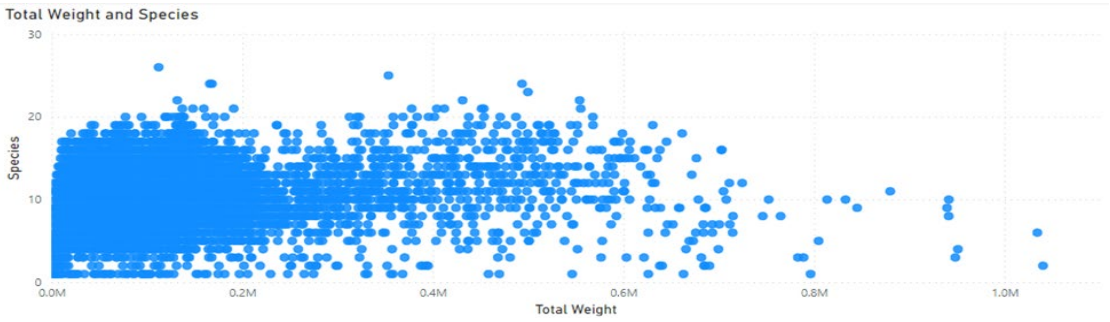
Figure 5 - Average Number of Species by Gear Type 2017-2022



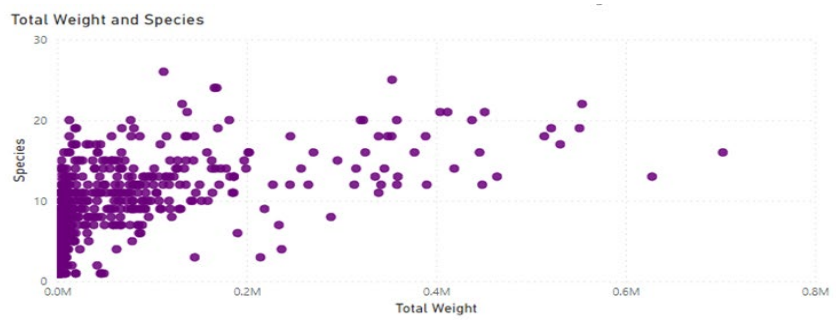
Breaking out this data further, we do see patterns emerging when we graph a comparison of species counts to total weight of inspected landings compared to all landings. **In the presence of an inspector, the number of species tends to increase with the weight of the catch.** The graphs in Figure 6 illustrate this phenomena.

Figure 6 - Total Landings Weight vs. Number of Species

All Landings



Inspected Landings

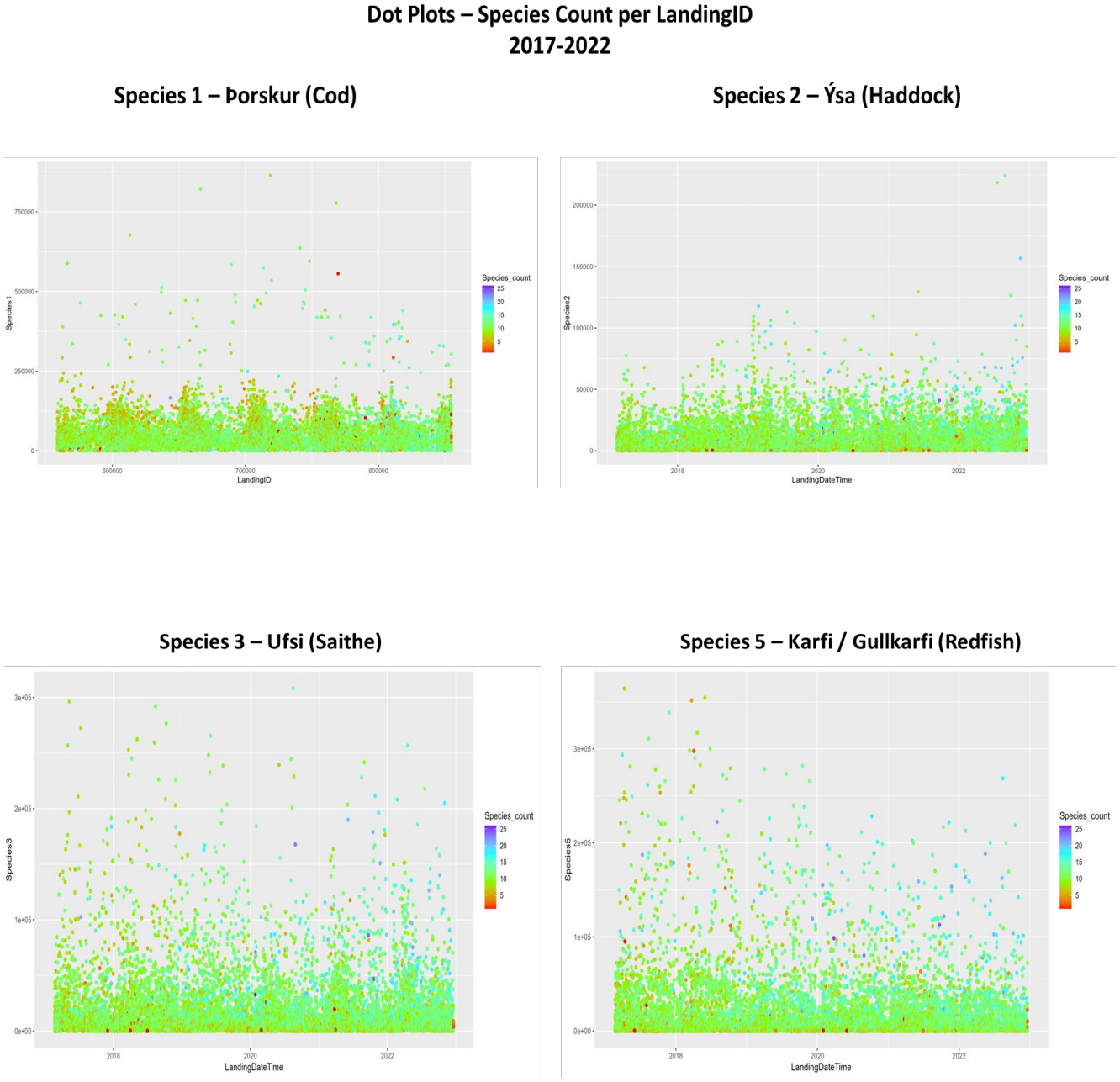


Dot Plots - Illustrating the Mix of Species in Landings

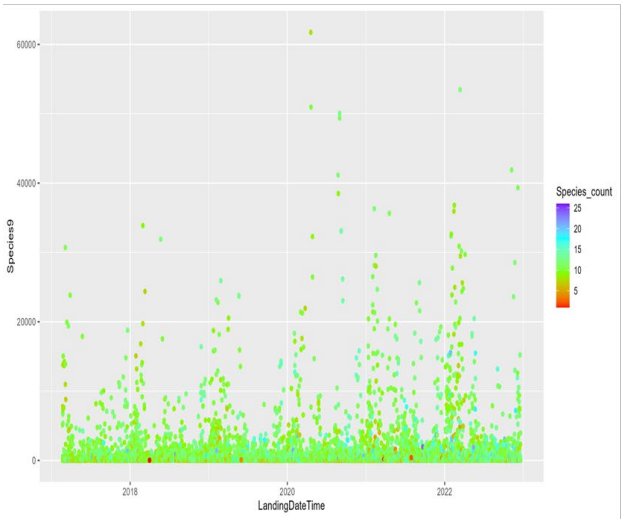
Dot Plots proved to be an effective means of highlighting clusters and outliers for species counts in landings data. In each of the dot plot graphs below, each dot represents all landings from 2017 to 2022 that contained the species, plotted by LandingID and weight of the total landing across all species in the catch. The color of the dot represents the number of species in the landing. The rainbow color pallet assigns a blue/purple coloring to more diverse landings and assigns orange/red coloring to less diverse landings.

A common trend across graphs is that the dots trend toward the blue/purple color and away from orange/red, a representation of the fact that more species are represented in landings as we move into 2022. This is likely an indication that discard is decreasing.

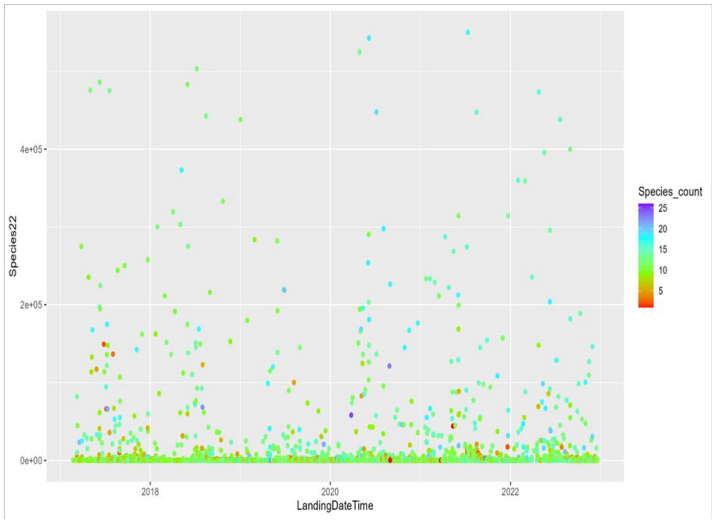
Figure 7 - Dot Plots



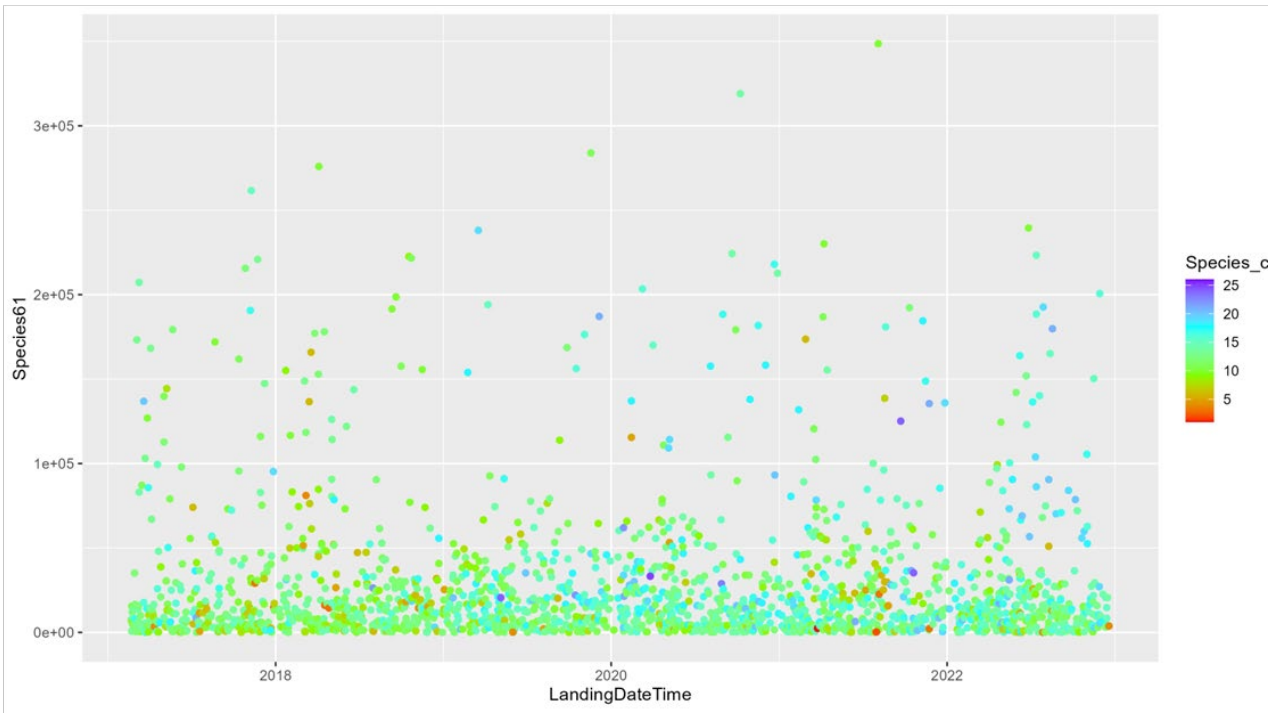
Species 9 – Steinbítur (Atlantic wolffish)



Species 22 – Grálúða / Svarta Spraka (Greenland Halibut)



Species 61 – Djúpkarfi (Deepwater Redfish)

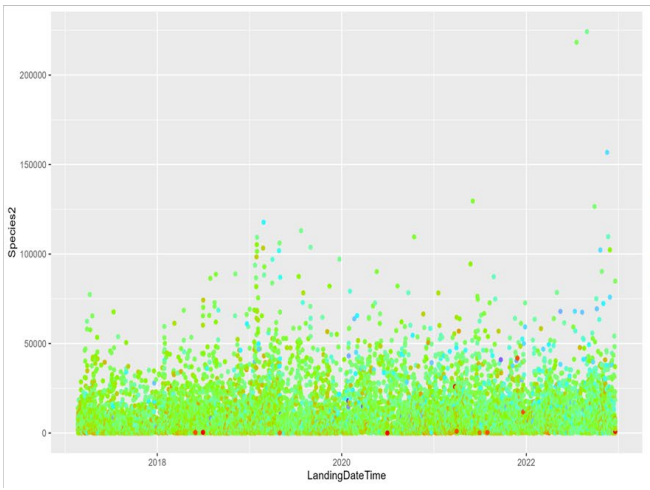


In the second set of dot plots below, non-surveilled and surveilled dot plots are show side-by-side for species 19 and 61. You will notice that not only are there zero red dots (representing low species counts) but most of the dot are aqua and blue (representing high species counts). **When comparing surveilled landings to non-surveilled landings, there is much more species diversity in inspected landings and likely less discard.**

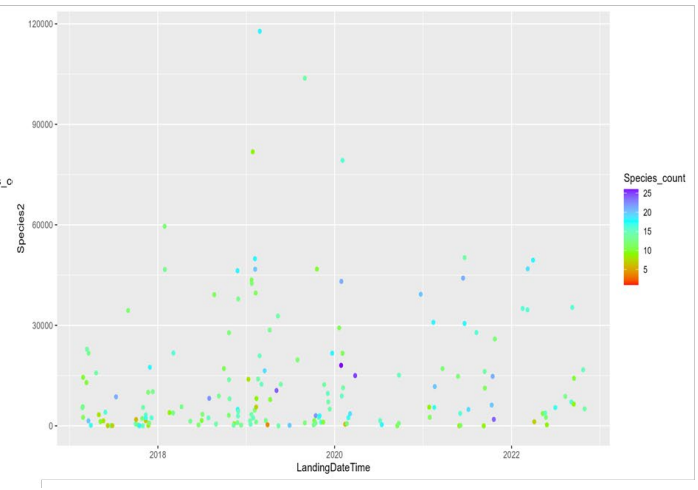
Dot Plots - Species Count per LandingID
Non-Surveilled vs Surveilled
2017-2022

Species 2 – Ýsa (Haddock)

Non-Surveilled

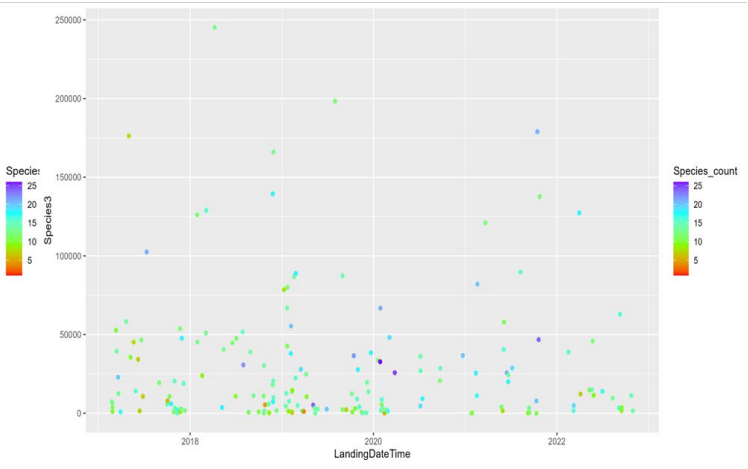


Surveilled



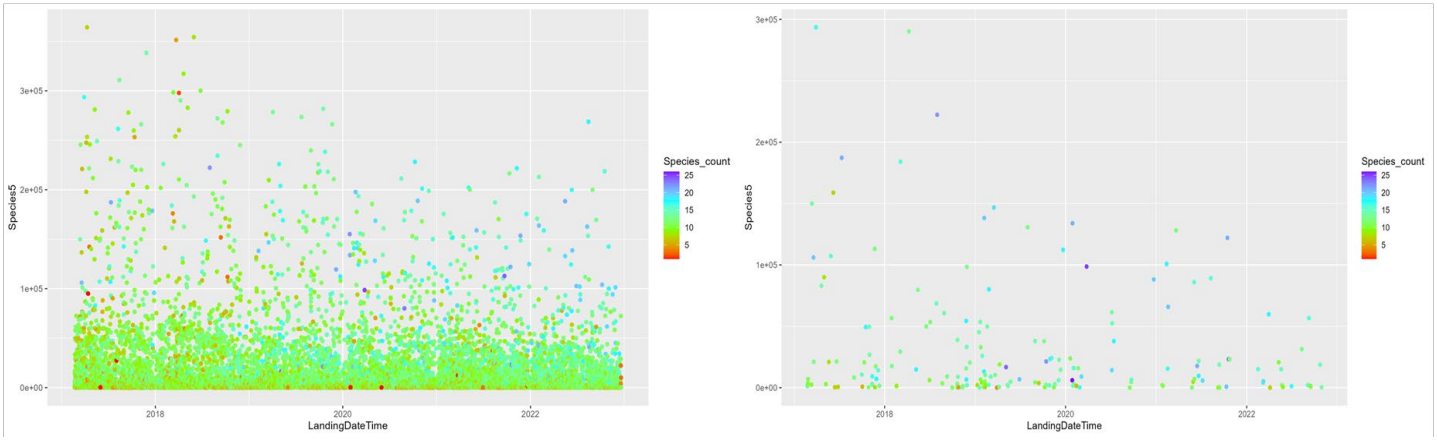
Species 3 – Ufsi (Saithe)

Non-Surveilled



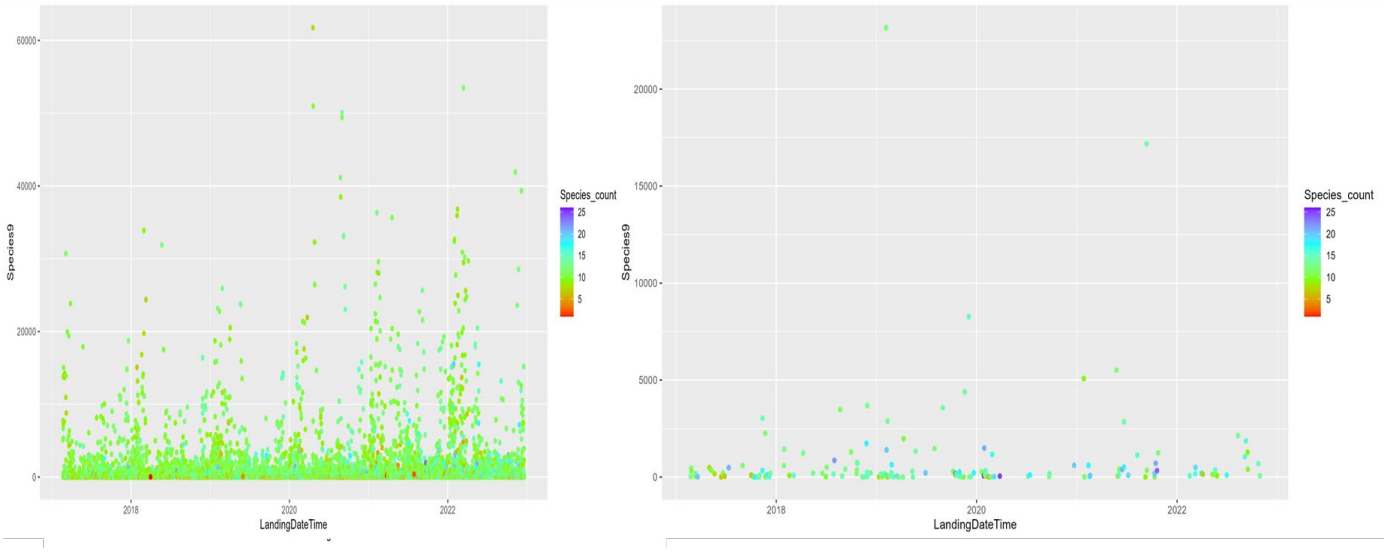
Species 5 – Karfi / Gullkarfi (Redfish)

Non-Surveilled | **Surveilled**



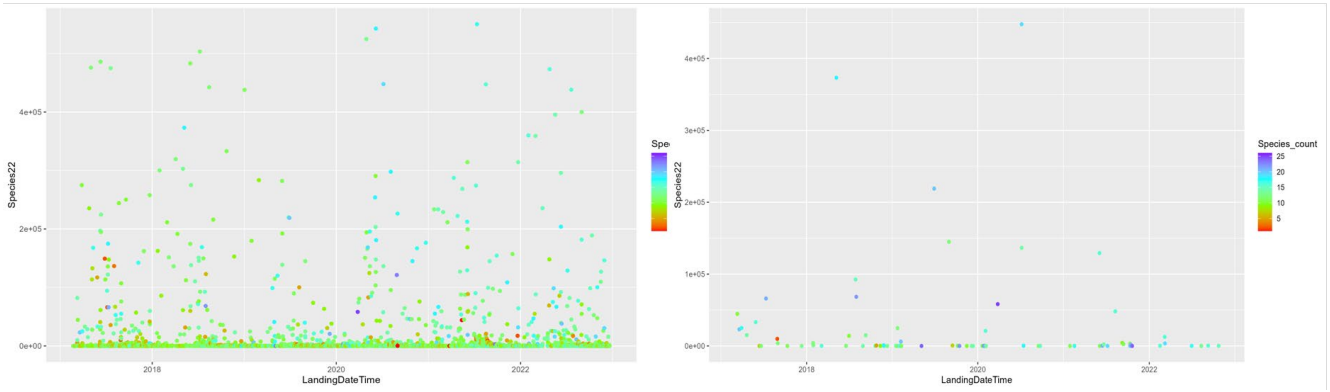
Species 9 – Steinbítur (Atlantic wolffish)

Non-Surveilled | **Surveilled**



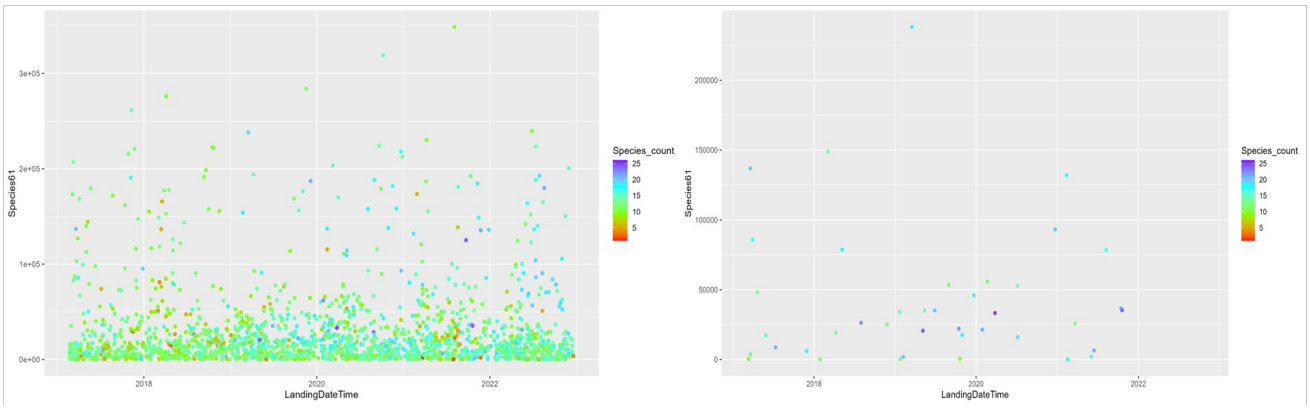
Species 22 – Grálúða / Svarta Spraka (Greenland Halibut)

Non-Surveilled | **Surveilled**



Species 61 – Djúpkarfi (Deepwater Redfish)

Non-Surveilled **Surveilled**



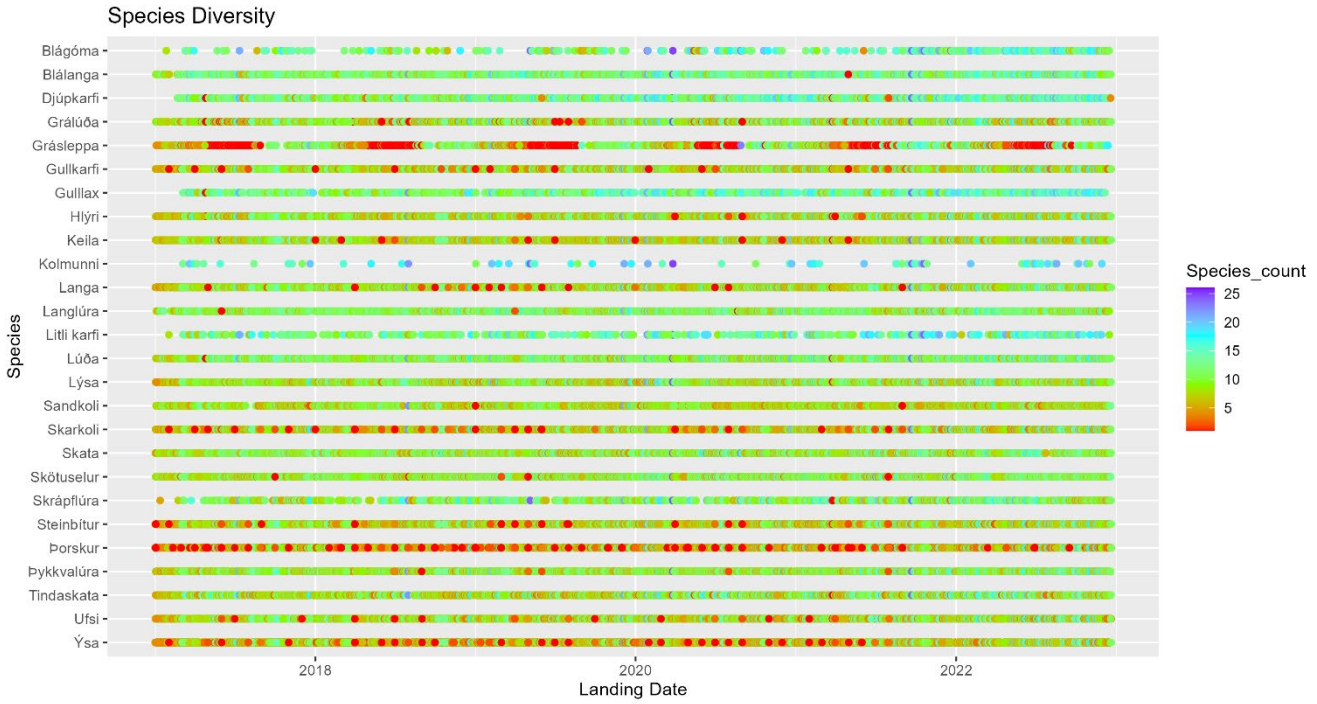
Slice Charts – Illustrating Species Diversity

Below are three variations of species slice charts that were developed to show both the overall stack of diversity in landings as well as the health of the fishery ecosystem from 2017 through 2022. Similar to the dot plots above, the color of the dot indicates the overall species count of the landings. The top two slice charts contain data from all landings, while the bottom chart only contains data from landings where an inspector was present.

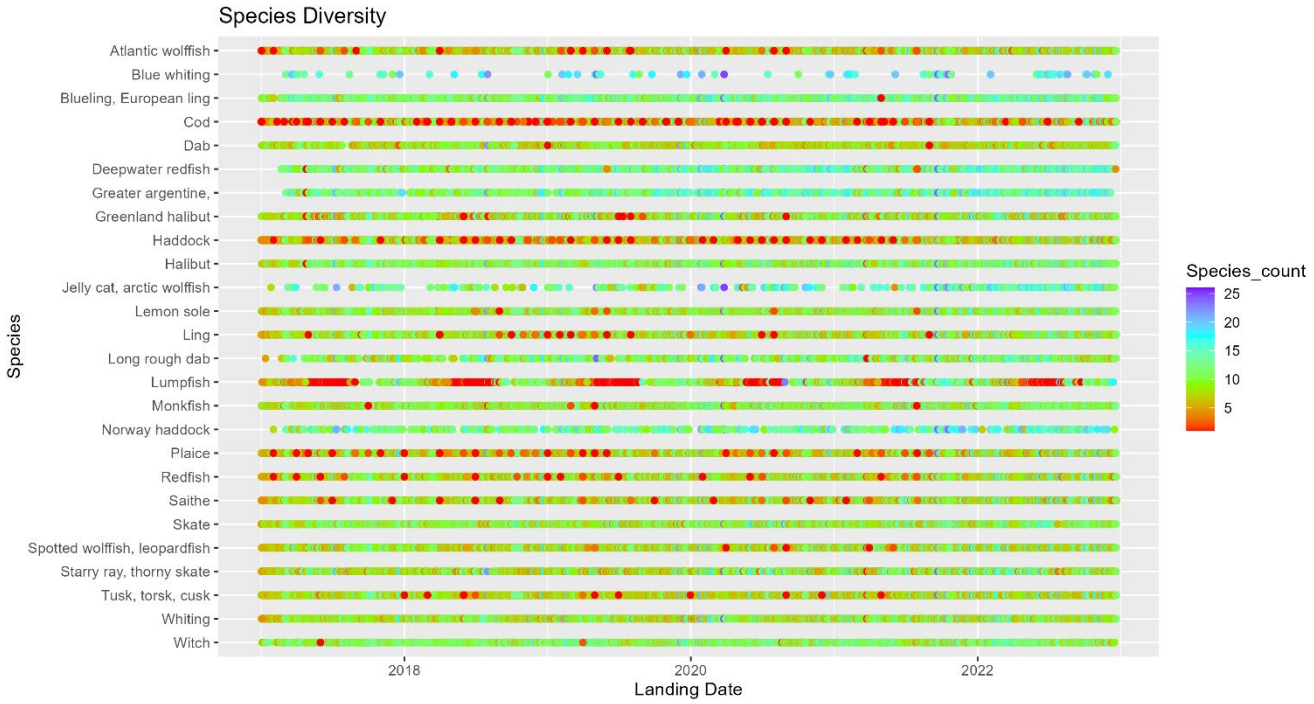
The charts provide a high level, one page view of the diversity of landings. As was evident in the above dot plots, there is less "red" on the right side of the graphs, an indication that species diversity is trending favorably, an indication of a decrease in discarding.

Slice Charts – 2017 to 2022
All Landings Versus Inspected Landings

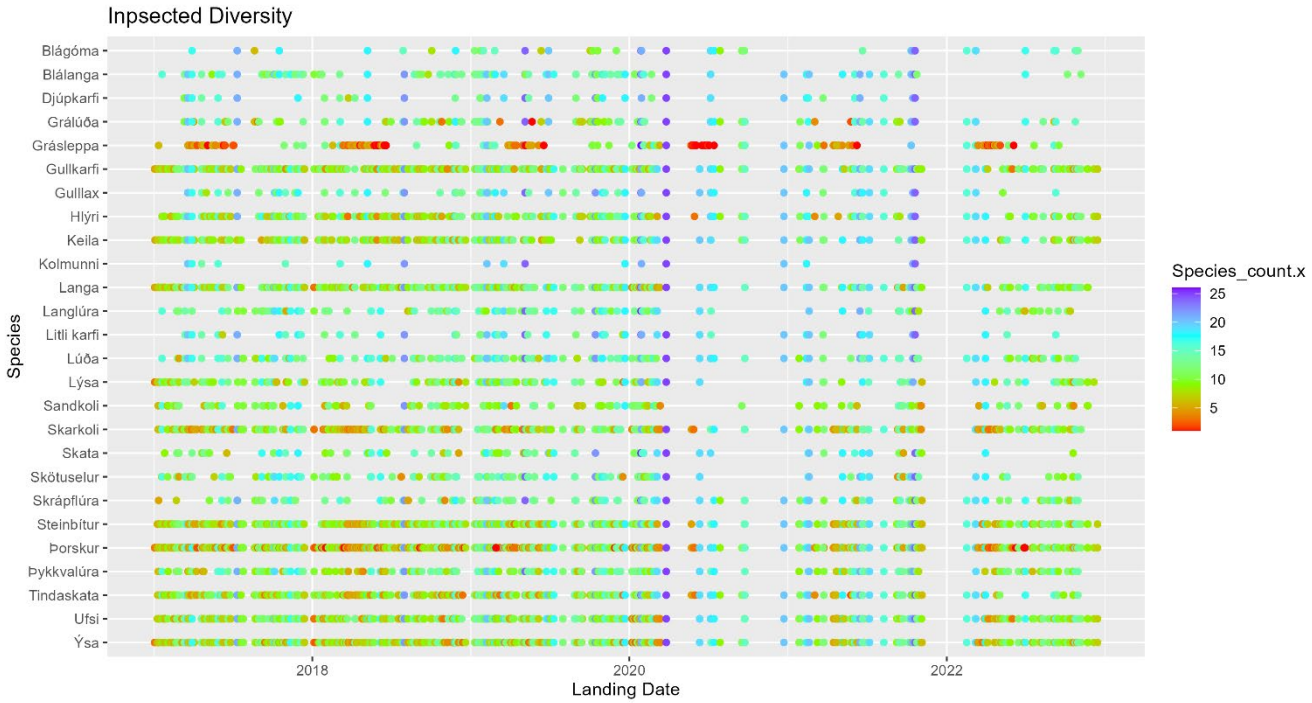
All Landings – Icelandic Species Names



All Landings – English Species Names



Inspected Landings – Icelandic Species Names



Feature Engineering

After the data was cleaned, we performed feature engineering. This involved creating new columns from the existing data to capture additional information that could be useful for our analysis. For example, we created a column for the total weight of all species caught during each landing, as well as columns for dates and quarters for the total landing. For the pricing dataset this included calculating the amount of kg a company had caught during the year to that point along with calculating numerous pricing statistics in an attempt to determine what previous pricing information was correlated with the price paid for a specific species.

At this point in the data extraction process, we developed a series of reports which we presented to the Fishing Directorate in Iceland. They noticed that we were combining two different types of fishing that were skewing our results. Pelagic fishing (i.e. ground fish) are caught in large numbers using different gear in summer months. The techniques used for pelagic fishing differ greatly from the fishing industry in general. With the guidance from Iceland we decided to divide the data by gear type as an initial step in our processing methodology.

Final Data Review

The primary table for species analysis included information on 133,936 landings with 256 features. Of these landings, 958 occurred in the presence of inspectors. However, this number is lower than the total number of inspections due to some inspections being unable to match landing IDs, as well as inspections for Pelagic vessels and vessels involved in fishing types beyond the scope of this

investigation. The pricing data was then merged with the landings/inspections dataset. This data included the amount of kg of each species a company had caught up to that point in the year to get an idea of where the company stood in terms of their quota limit. The pricing data also attempted to capture price incentives that existed for certain species. To do so, features were included that calculated the unweighted mean of the price per kg sold for a species for the previous 21, 14, and 7 days. While weighted mean, median, maximum, and minimum price per kg for a species were also created and tested, they didn't demonstrate as high of a correlation with the actual price sold compared to the unweighted mean.

In addition to the pricing data, features were included that tried to capture the current market price for a species. The time period for these means started the day before the fishing company sold their catch, as the price a company gets for their catch isn't available when they are out at sea. This allowed for a better understanding of the market price for a species. Overall, the pricing data and market price features provided valuable information for understanding the dynamics of the fishing industry.

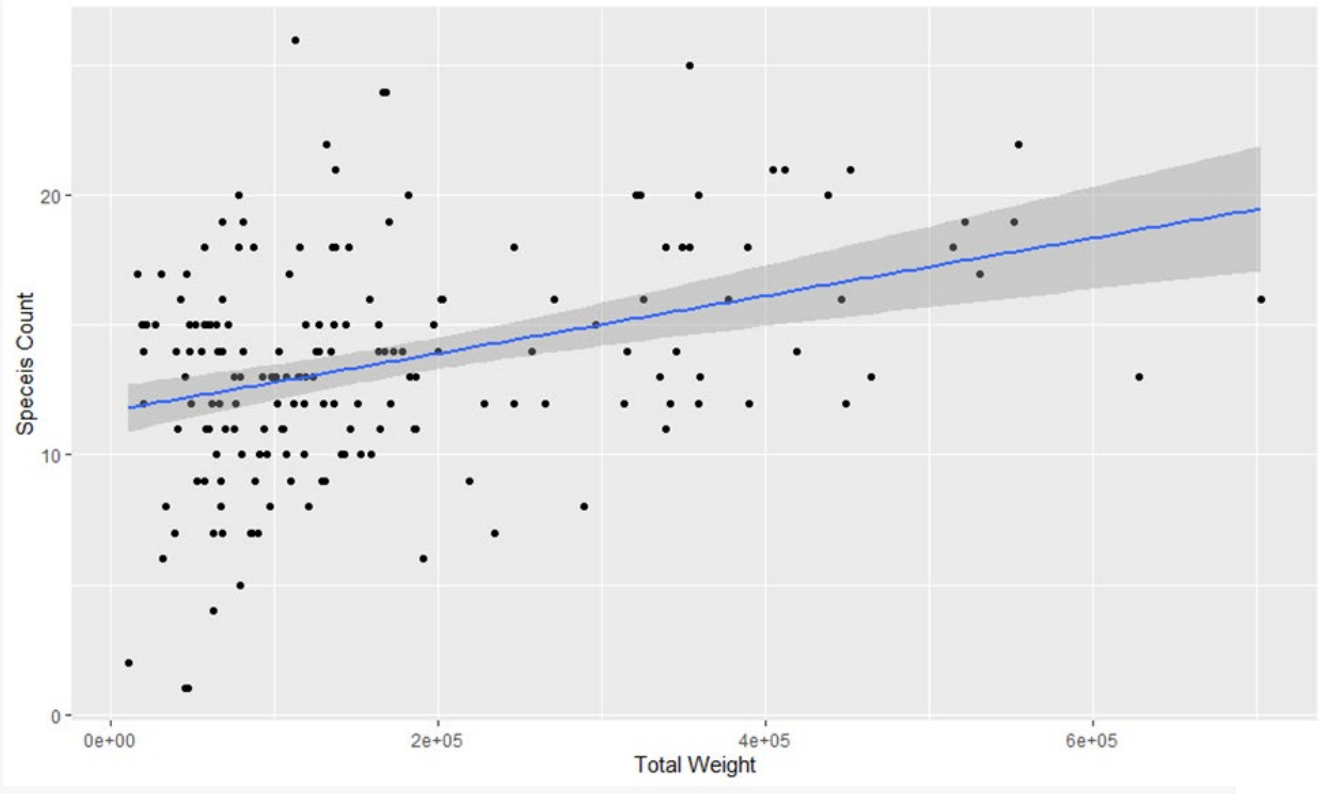
Regression Modeling

Linear models, such as basic linear regression, rely on the assumptions of linearity and homoscedasticity of the data. However, in this problem, we have a limited number of data points and a significant amount of heterogeneity within the data, making it a challenge to model the relationships in the given data.

To develop a predictive model, we must start with data that is known to be accurate that can thus be used to predict and model the existing relationships. For the most common Botnvarpa gear type, we have records of 177 surveillance inspections from 2017 until the present. **The data does not show much correlation between variables suggesting it will be very difficult to develop a predictive model**.

The graph below (Figure 7) shows the data points derived from modeling the total weight of a landing compared to the number of species on each landing. In this case, we did find a statistically significant relationship between total weight and the number of species. However, the high mean squared error and the low r-squared value suggests the surveillance data is not at the point where we can begin developing effective linear models based on our current understanding of the data.

Figure 7 – Graph of Botnvarpa Landings to Species Count



In this case the simple linear model produced the following ANOVA table with a relatively low R squared value of 0.126

```
Coefficients:

Estimate Std. Error t value Pr(>|t|)
(Intercept)        1.170e+01  4.816e-01  24.291  < 2e-16 ***
total_weight_all 1.112e-05  2.207e-06   5.037 1.17e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 4.059 on 175 degrees of freedom
Multiple R-squared:  0.1266,  Adjusted R-squared:  0.1216
F-statistic: 25.37 on 1 and 175 DF,  p-value: 1.172e-06
```

Attempting to extend this model to other types of linear regression and machine learning is unlikely to be effective due to the small sample size which increases the chances of overfitting and poor generalization, as there may not be enough information to accurately model the relationship between the variables. Additionally, the heterogeneity of the data may lead to non-linear relationships and unequal variances in the error terms.

Interestingly, the regression model with the best r squared value was composed of the following variables several of which contained statistical significance:

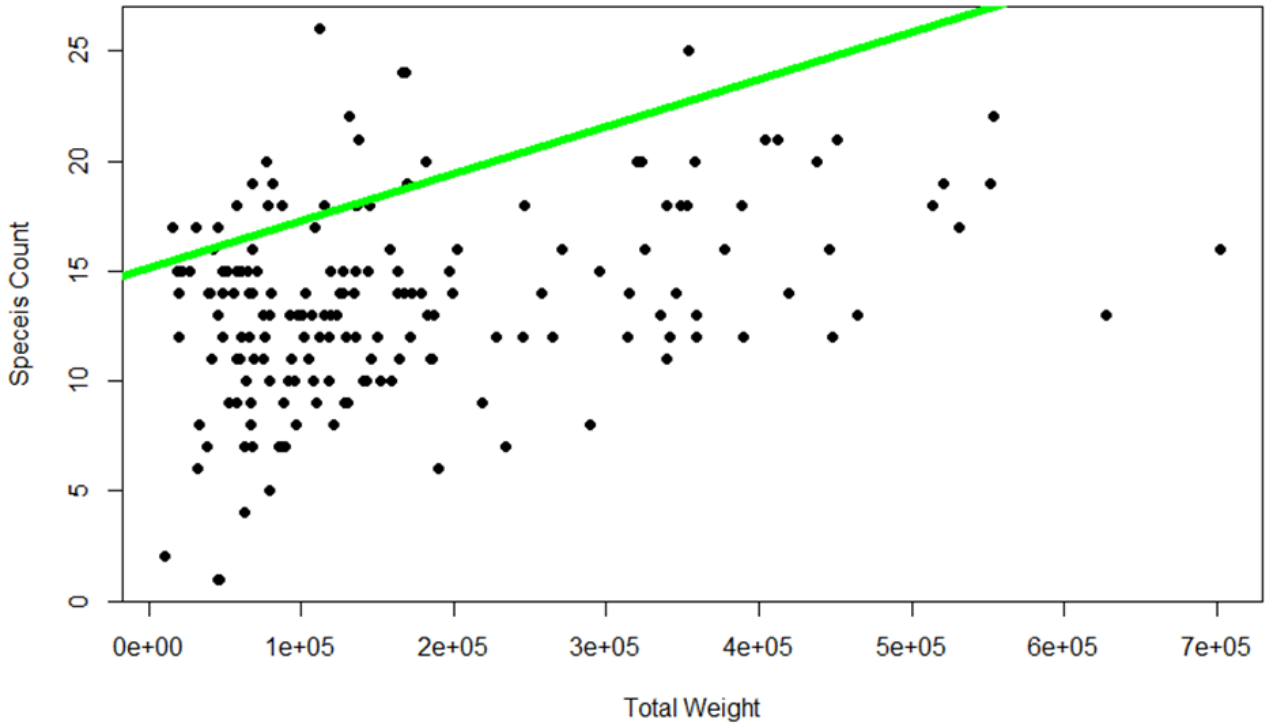
```
Coefficients:

Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.514e+01  1.058e+00  14.305  < 2e-16 ***
total_weight_all  2.142e-05  5.135e-06   4.171 0.000105 ***
Species1          -2.006e-05  1.259e-05  -1.593 0.116647
Species2           5.962e-06  2.790e-05   0.214 0.831532
Species3          -3.803e-05  1.207e-05  -3.151 0.002590 **
Species23         -2.034e-04  5.955e-05  -3.417 0.001176 **
Species21          3.274e-03  2.495e-03   1.312 0.194692
LandingsMonth    -1.667e-01  9.753e-02  -1.710 0.092796 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 2.518 on 57 degrees of freedom
  (112 observations deleted due to missingness)
Multiple R-squared:  0.5046,  Adjusted R-squared:  0.4438
F-statistic: 8.296 on 7 and 57 DF,  p-value: 6.011e-07
```

This model also contains a very high error rate and puts the overall predictive line much higher than the prior model (Figure 8):

Figure 8 - Graph of Botnvarpa Landings to Species Count



Attempting to model the data based solely on the total weight in the catch or other variables with statistical correlation will likely result in a high rate of error. This is because the complexity of the relationship between the variables is not captured by a simple linear regression model, leading to unreliable and potentially misleading results.

At the present time, linear models are not appropriate for this problem given the limited sample size and the heterogeneity of the data. Alternative modeling techniques, such as non-linear regression or tree-based methods, may be more suitable for accurately capturing the relationships between the variables.

While it does not appear possible to model the data at the present time, such a model of surveillance data may emerge at some point in the future based on improved knowledge or tracking of fishing gear or conditions. The limited data set makes it difficult to extract meaningful relationships.

The surveillance sample does suggest some relationships that are potentially useful for future monitoring. On the low end of the species count, few landings exist with fewer than 10 species – especially at higher weight levels. Fiskistofa could potentially develop a metric related to this gear type for landings with species counts that fall outside of the surveillance observations.

Regression and Classification Tree Models

General Observations

- Tree models are a type of supervised machine learning that can predict numerical or categorical values.
- Decision trees, random forests, and gradient boosting are the three types of tree modeling techniques used in the analysis of landings data.
- Decision trees are simple but are the least accurate, while random forests and gradient boosting are more accurate but complex.
- The tree models were applied to the Icelandic fishing industry to predict species count, identify commonalities among inspected trips, and classify whether a certain species of fish was caught on a given trip.
- The model for predicting species count was not very predictive due to the most impactful variables being those realized after the trip had already occurred.
- The classification tree model for identifying commonalities among inspected trips had a near-perfect accuracy, but was unable to identify common threads among inspected trips due to the minimal amount of inspection data available.
- A binomial classification tree model was able to predict whether a certain species of fish was caught on a given trip, with the corresponding accuracy metrics and most predictive variables for each species' model shown in a summary table.

Overview of Tree Models

Tree models are a type of supervised machine learning that utilize a series of if-then statements to make a prediction of either numerical or categorical values. A model that predicts numerical values is referred to as a regression tree model, while the models that predict categorical values are classification tree models.

The landings data will be explored through the use of three different tree modeling techniques:

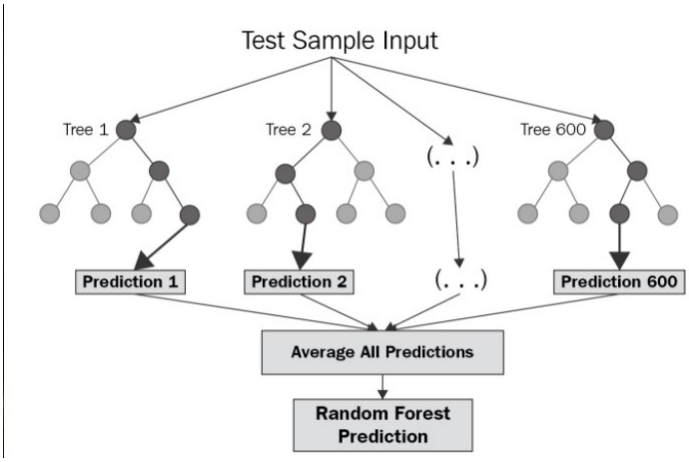
1. Decision trees
2. Random forests
3. Gradient boosting

Decision trees are the simplest of the tree models, while random forests and gradient boosting are more complex forms of decision trees which would add scale to decision trees by using more advanced iterative techniques. The design process includes dividing a given dataset into "training" and "test" sets which are used as their name suggests. The algorithm builds a decision tree based on the data available in the training set and predicts either the class or value of the target variable. The resulting decision tree model can then be evaluated based on the test data which presents the issue of overfitting the model. An example visual for a decision tree predicting whether Gullkarfi were caught on a given trip is shown in Figure 10.

Random forests and gradient boosting are more accurate than decision trees but are difficult to visualize. Random forests are an ensemble learning method that assembles a variety of decision trees and takes the average of the outputs. An important aspect of tree models, including random

forests, is their ability to determine the most impactful variables for predicting the target variable. Below is a crude model displaying the random forest technique at work (Figure 9).

Figure 9 – Random Forest Technique



Gradient boosting models are considered one of the most common machine learning algorithms due to their speed and accuracy, even when working with large datasets. In the simplest terms, the technique iterates upon itself while adjusting weights to minimize the overall error of the model. Like the random forests above, there are techniques to return the most commonly used predictor variables in the gradient boosting model.

Applying Tree Models to the Icelandic Fishing Industry

The first tree model that was tested attempted to predict species count for a given landing. For the landings data excluding Pelagic and Handfæri records, the biggest predictors of species count are whether certain marginal species are caught on a given trip. This fact itself is not surprising, although some insight can be gained by examining which species the model is utilizing to predict species count. Since species such as cod and haddock are brought in on most landings, these species are not predictive to species count. Instead, it is the species that are brought in on landings with higher species counts that can predict the total species count for the trip such as Gullkarfi and Ufsi. Unfortunately, the variables predictive of species count were all realized after the trip had already occurred, so there is minimal predictive value to this model. While the ideal goal is to have the ability to predict the species count of a given trip prior to the trip having taken place, based only on factors such as time of year, fishing area, harbor, etc., eliminating the element of species weights from the data did not lead to a predictive model with an acceptable accuracy metric.

There are a limited number of inspected trips relative to the landings dataset as a whole, but identifying the commonalities among inspected trips is an important consideration. Assessing common characteristics of inspected trips might provide insight as to what degree these inspections are discouraging discard. A classification decision tree model was assembled, and the resulting model was over 99% accurate while only using six predictor variables, but a further analysis into the intricacies of the model explains why this is not necessarily a positive result. All but two of the non-inspected trips were classified correctly, but only twelve inspections were correctly classified by the model. The model can claim a near-perfect accuracy due to the insignificant amount of inspected trips. Unfortunately, classification tree models are unable to identify common threads among inspected trips.

The final relationship to be explored through tree models is a binomial classification tree model denoting whether a certain species of fish was caught on a given trip. This model was able to provide the most insights, as it displayed an acceptable accuracy metric while signaling important predictive factors. The decision tree for predicting whether Gullkarfi was caught is shown in Figure 10 which indicates that species count is the greatest predictor. Blue boxes signify landings for which the Gullkarfi was caught. The tree demonstrates that if four or less species are caught on a given trip, it is likely that Gullkarfi was not one of them, while if five or more species were caught, Gullkarfi is probably among them.

This model was accurate for approximately 88% of records in the data upon which the model was trained and tested on. In the training dataset, about half of the landings had Gullkarfi in its catch while the other half did not, ensuring that the same issue for the inspection classification model is not persisting here. Ultimately, the insight gained from this decision tree can help verify whether or not egregious displays of discard are not occurring. For example, if seven or more species are caught without using a Dragnót, 96.5% of these trips will contain Gullkarfi in the catch, so it is cause for concern if this is not the case.

This exercise was repeated for the six most common fish species that are seen in the landings data, with Þorskur excluded since it appears in the majority of landings and there is unlikely to be actionable insights derived from such a model. The accuracy metrics with respect to the test data and the most predictive variables for each of the six most common fish are shown in the summary table in Figure 11.

Figure 10 – Decision Tree for Predicting Whether Gullkarfi Was Caught

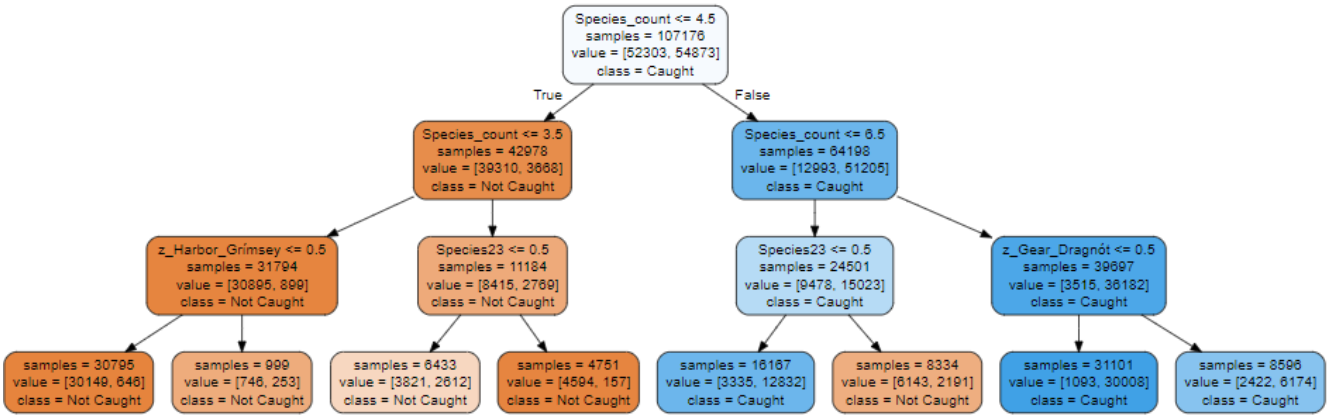


Figure 11 - Summary Table for Predicting Whether Common Fish Are Caught

Species	DT Accuracy	RF Accuracy	GB Accuracy	Most Predictive Variables
Ýsa	90.54%	95.91%	94.38%	Species Count, Þorskur Weight, Lína Gear
Ufsi	82.59%	91.31%	89.64%	Species Count, Gullkarfi Weight, Þorskfisknet Gear
Gullkarfi	87.99%	93.74%	92.50%	Species Count, Skarkoli Weight, Hlýri Weight
Gulllax	99.56%	99.58%	99.61%	Djúpkarfi Weight, Species Count, Gullkarfi Weight
Grálúða	94.66%	96.86%	96.95%	Hlýri Weight, Þorskur Weight, Species Count
Djúpkarfi	99.04%	99.41%	99.37%	Gulllax Weight, Keila Weight, Gullkarfi Weight

Unsupervised Modeling

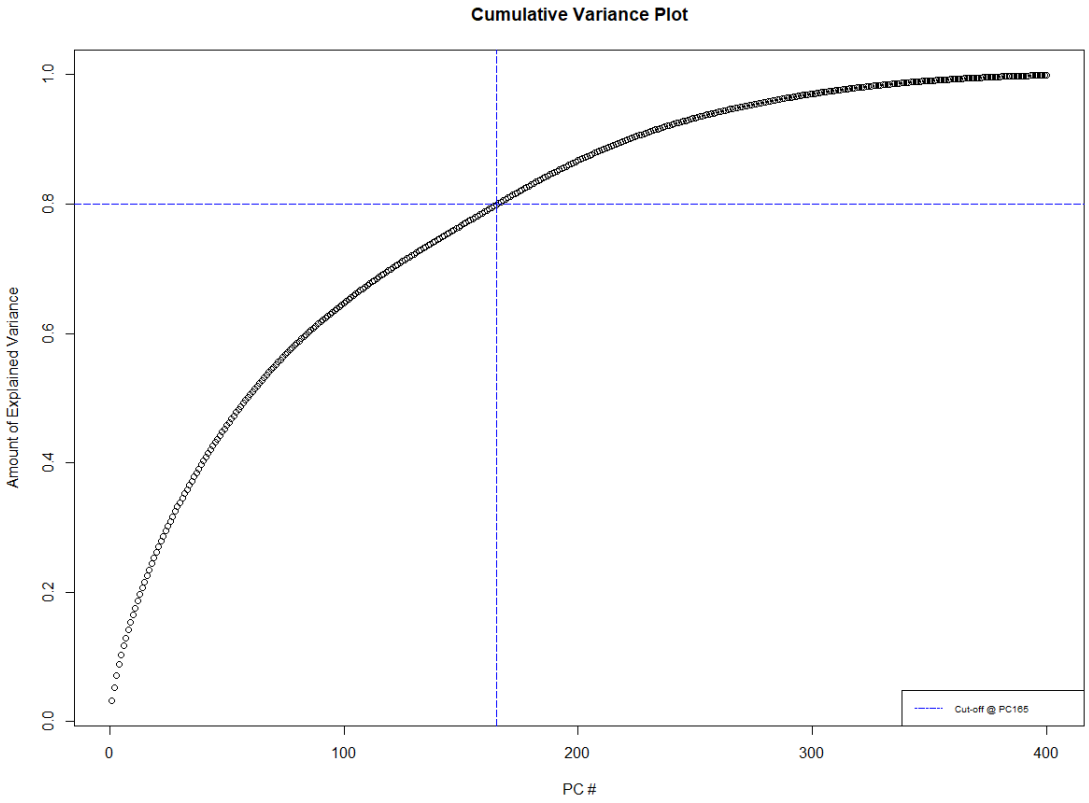
Summary

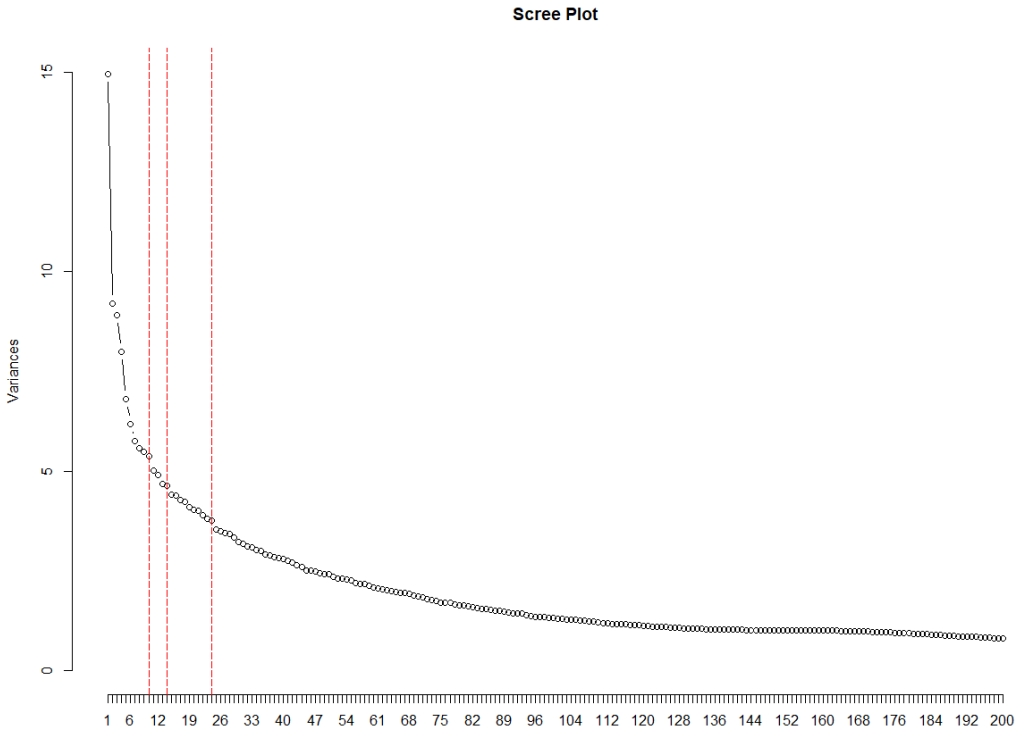
- Principal component analysis (PCA) is a dimension reduction technique. This allows us to visualize information from numerous variables in a single vector. We can see that there is some segmentation based upon the number of species caught indicating there are distinct groupings of fishing trips. Trips with a low species count are commonly found in the left-hand corner of the first PCA chart indicating there is a relationship within the dataset that these trips share. We can see even clearer groupings when examining the size of the firms and size appears to be correlated negatively with the number of species caught on a trip. This may indicate that the larger firms engage in specialization, where each trip is pursuing a single species and any bycatch is species living in the same zone as the target species.
    - Two downsides:
        - PCA only calculates linear combinations and potentially the dataset on non-linearities leading to low variance explained. Diffusion maps will be implemented to conduct nonlinear dimension reduction.
        - PCA does not allow for interpretability of the vector relationships.

The problem of detecting discard is that we don't have an actual measurement of the issue. Unsupervised models are utilized to make sense of the underlying dynamics within the data and to categorize the trips. Principal components analysis is utilized to summarize the information contained in our several hundred columns into a much smaller number of columns. This type of analysis can help illuminate trends and important relationships within the data and inform modeling decisions.

Principal Components Analysis (PCA) is a technique to reduce dimensionality**.** The data we utilized originally had each type of fish caught on a trip as the lowest level of observation. To make the trip itself the lowest level of observation, the dataset had to be reshaped, increasing the number of columns to several hundred.[1] The first eigenvector (or principal component) has the largest variance of all of the possible linear combinations of the columns in the data set. Each following principal component has the maximum amount of variance subject to being orthogonal to the previous principal components. All the principal components are uncorrelated.[2]

[1] https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202
[2] Elements of Statistical Learning



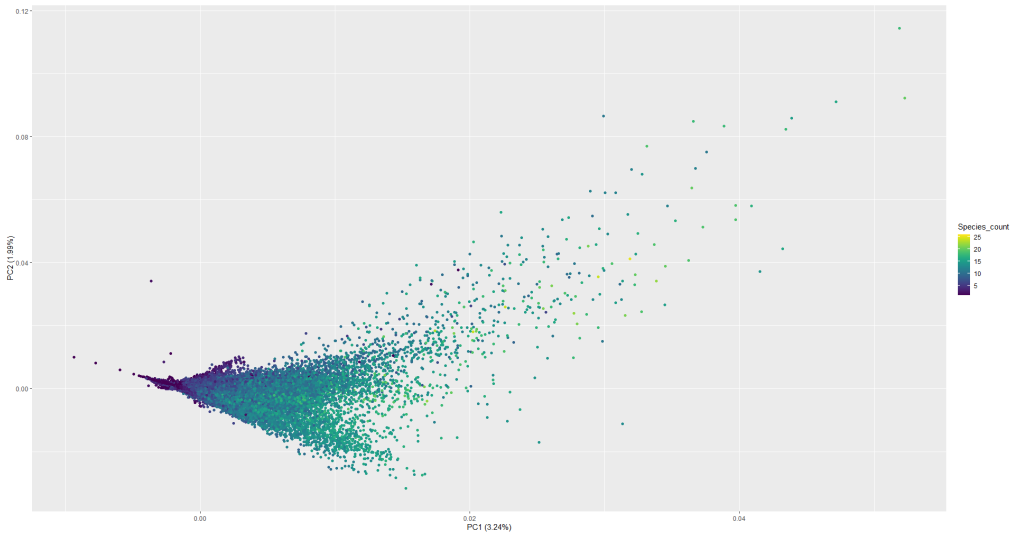


The two charts above, Cumulative Variance Plot and Scree Plot are two ways to visualize the variance captured by each eigenvector. Most notability, for cumulative variance, it would take 165 eigenvectors to explain 80% of the variance in the dataset. Usually, PCA can explain 80% of the variance in far less than 50 eigenvectors.

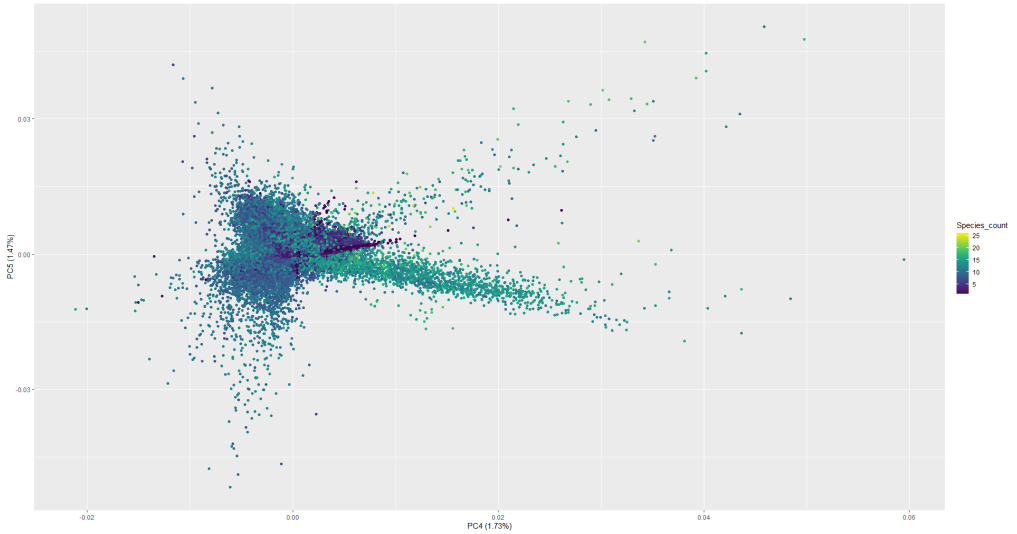
To determine the number of eigenvectors to down select, we use both the visual Scree Plot test and the Guttman-Kaiser Criterion. For the Scree Test, we select all the eigenvectors up to where a sharp change in the slope occurs. This is a more subjective test and three different potential points are identified in the above Scree Plot marked with the dotted red lines. The Guttman Kaiser Criterion in short selects an eigenvalue whose covariance is greater than 1.[3] In this case, the Guttman-Kaiser Criterion indicates we should select 156 eigenvectors.

A potential issue with PCA is that it is examining linear combinations. This means if you have large non-linearities within your dataset, PCA is likely to not provide a good explanation of the variance. This could be the reason why it takes over a 100 eigenvectors to get to a cumulative variance explained of 80% and the Guttman-Kaiser Criterion indicating 156 eigenvectors are needed.

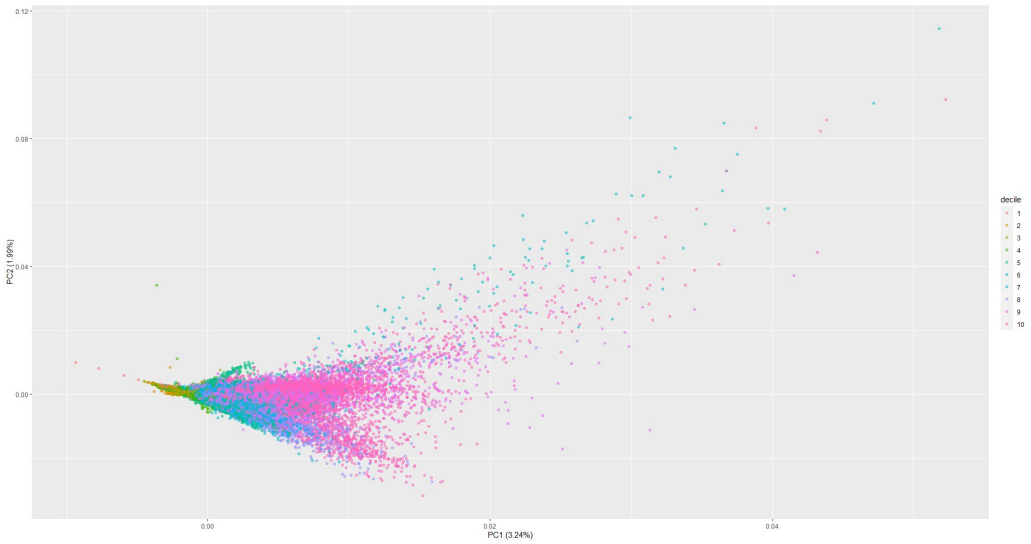
However, using PCA biplots can help explain some of the linear variation we are seeing within the dataset. Examining the Scree Plot, a vast majority of the eigenvectors explaining 5% or more of the variance reside from PC1 to PC10. The two biplots below show the plotting of Principal Component 1 (PC1) against Principal Component 2 (PC2) and Principal Component 4 (PC4) against Principal Component 5 (PC5). In these plots, we can start to see some segmentation of the population of trips by the number of species caught. However, all four PC's don't explain much variation but it's promising. **We can see that there is some segmentation based upon the number of species caught, indicating there are distinct groupings of fishing trips. Trips with a low species count are commonly found in the right-hand corner of the first PCA chart, indicating there is a relationship of variation within the dataset that these trips share.**



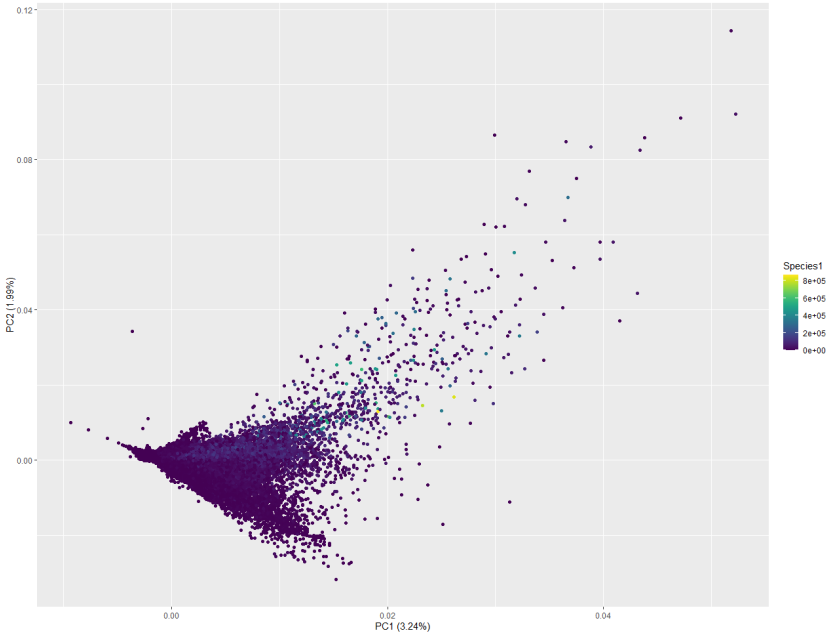
[3] https://biologydirect.biomedcentral.com/articles/10.1186/1745-6150-2-2

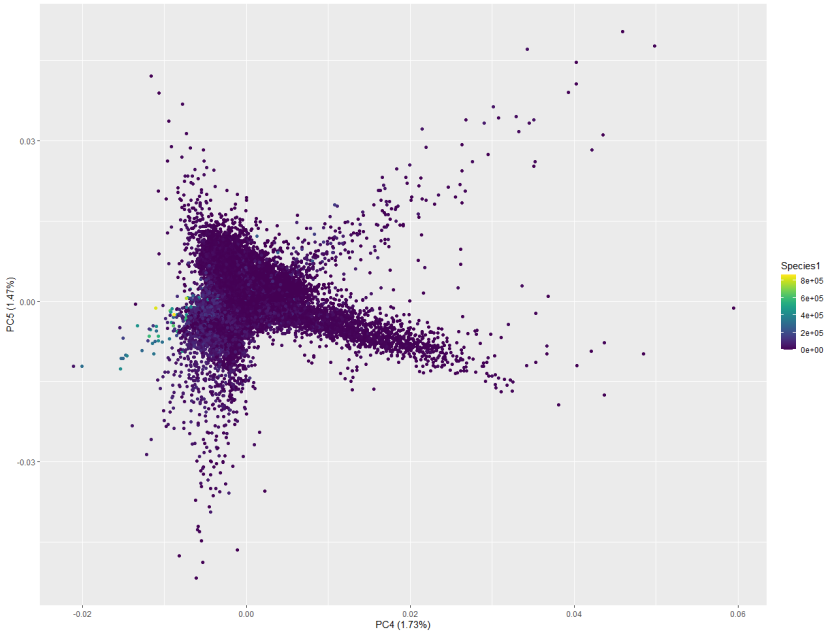


Second, we examine the PCA results by size of the fishing company. Decile 10 represents the largest companies while Decile 1 is the smallest set of companies. There appears to be a clear pattern of dispersal. The largest firms are clustered in the left-hand side of the graph, while smaller firms appear to the right of the graph. We can also see that there appears a relationship between the size of the company and the number of species caught on a trip. It appears **the larger the company, the less likely the trip is to have a high species count.** Potentially, the larger firms engage in specialization where each trip might target a single species. We need to do more investigation based on company size.



Additionally, we examined if we could separate trips where cod is caught compared to other trips. Cod is considered the main target due to its high demand and subsequent high prices per kg. Our EDA shows that most of the trips where Cod are caught include fewer than 2 other species.





While using PCA has unearthed several linear relationships, it is clear there are more non-linearities in the underlying data. This is evident through both the first two principal components only accounting for roughly 5.5% of the variance within the dataset and requiring 165 principal components to capture 80% of the variation in the dataset. It is preferred to have a higher value for the first two principal components and a much lower number of principal components to reach 80% variation. There are multiple techniques that handle non-linear dimensional reduction. The group tried diffusion maps, isomap, and kernel-based PCA to name a few.[4] The issue the group ran into is the prohibitive amount of RAM required to compute these nonlinear dimension reduction techniques. These techniques routinely required more than 32GB of RAM. Utilizing nonlinear dimension reduction techniques is an area that with more time and greater computing power could yield important insights.

Zero Boat Sampling Methodology

Inspections create trip observations where we can say with certainty no discard occurred. We can then use these observations to create a “zero” trip or a profile of what an average trip should look like. As more and more trips are inspected, these observations begin to form a sample statistic (mean, distribution, standard deviation, etc.) which is representative of the population.[5] As the number of inspected observations increase relative to the population of trips, the confidence in which a typical trip looks like increases as well. The impact is trips can be flagged that don’t conform to the expected distribution of values for a trip. An important caveat is that these trips must be randomly selected. Otherwise, the resulting “zero” ship is representative of the current inspection selection criteria rather than the population of trips.

When deciding how many trips need to be inspected in order to generate a sample representative of the population, the confidence level and the acceptable margin of error need to be determined. A

[4] https://cran.r-project.org/web/packages/dimRed/vignettes/dimensionality-reduction.pdf
[5] https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_probability/BS704_Probability12.html

confidence level indicates the probability the sample statistic is representative of the population statistic given the margin of error and is expressed in percentages. For example, the 95% confidence level for a sample statistic represents the probability that the sample statistic is representative of the population statistic given margin of error. Within the audit literature, the generally acceptable confidence level value is 95%.[6,7] Determining the acceptable margin of error is a trade off between by resource constraints and the desired accuracy of the sample statistic. The margin of error is directly influenced by the number of trips inspected. The higher the number of trips, the lower the margin of error and the more precise the sample statistics are.

Below is a table that shows the trade off between the margin of error and the number of inspected trips given a 95% confidence level and 21,026 trips in 2021. Comparing the population statistic to the sample statistic, in this case the mean number of species caught on a trip in 2021, it is evident how as the desired margin of error decreases, inspections increase, and the resulting sample mean is closer to the population mean.

The 13% margin of error row is a good example of sampling and the uncertainty that accompanies it. The sample mean is only .2 off of the population mean. However, if we repeated the sampling for 13% margin of error, it could look closer to 12% or 14% margin of error where the sample mean is 1.2 and 1 respectively off from the population mean.

Confidence Level	Margin of Error	Number of Inspections	Sample: Mean Number of Species Caught on a Trip	Population: Mean Number of Species Caught on a Trip
95%	18%	30	8.4	5.8
95%	17%	34	7.1	5.8
95%	16%	38	6.2	5.8
95%	15%	43	6.9	5.8
95%	14%	49	6.8	5.8
95%	**13%**	**57**	**6.0**	**5.8**
95%	12%	67	7.2	5.8
95%	11%	80	6.4	5.8

[6]https://www.occ.gov/publications-and-resources/publications/comptrollers-handbook/files/sampling-methodologies/pub-ch-sampling-methodologies.pdf
[7]https://us.aicpa.org/content/dam/aicpa/publications/accountingauditing/keytopics/downloadabledocuments/sampling_guide_technical_notes.pdf

95%	10%	96	6.0	5.8
95%	9%	118	6.7	5.8
95%	8%	150	6.4	5.8
95%	7%	195	6.2	5.8
95%	6%	264	6.4	5.8
95%	5%	378	6.0	5.8
95%	4%	584	6.0	5.8
95%	3%	1016	6.2	5.8
95%	2%	2156	6.0	5.8
95%	1%	6593	5.9	5.8

The method above assumes that the number of species caught is distributed homogeneously across the population. In reality, it is likely values that the fishing directorate is interested in are heterogeneously distributed due to different characteristics of trips such as the size of the firm, type of gear being used, time of year, fishing location, etc. In order to obtain a more accurate sample statistic, the stratified sampling technique can be used. Stratified sampling divides the population into mutually exclusive groups, also known as strata.[8] The proportion of observations in each strata compared to the population number of observations determine how many samples need to be drawn from each strata.

For example, using the 2021 trips, trips can be grouped based upon size of the firm (terciles based on 2020 kilograms caught) and gear used. Each of these groups (or strata) are mutually exclusive. A trip can only belong to one of these strata. Once the trips have been allocated to a stratum, the proportion of total trips the strata consist of can be calculated. The number of inspections required to be drawn from each stratum is then proportionally assigned. If we choose a 10% margin of error, this requires 96 inspections. Thus, each of the 96 inspections is allocated to a stratum depending on the stratum's proportion of the population counts.

Gear	Tercile	Count	Percent of Total	Number of Inspections Required

[8] https://cals.arizona.edu/classes/rnr321/Ch4.pdf

Botnvarpa	Smallest Third	46	0%	0
Botnvarpa	Middle Third	80	0%	0
Botnvarpa	Largest Third	2417	12%	12
Dragnót	Smallest Third	225	1%	1
Dragnót	Middle Third	2159	11%	10
Dragnót	Largest Third	616	3%	3
Grálúðunet	Smallest Third	3	0%	0
Grálúðunet	Middle Third	0	0%	0
Grálúðunet	Largest Third	50	0%	0
Grásleppunet	Smallest Third	2428	12%	12
Grásleppunet	Middle Third	58	0%	0
Grásleppunet	Largest Third	0	0%	0
Humarvarpa	Smallest Third	0	0%	0
Humarvarpa	Middle Third	21	0%	0
Humarvarpa	Largest Third	104	1%	0
Lína	Smallest Third	1739	9%	8
Lína	Middle Third	3691	18%	18
Lína	Largest Third	3007	15%	14
Rækjuvarpa	Smallest Third	16	0%	0
Rækjuvarpa	Middle Third	110	1%	1

Rækjuvarpa	Largest Third	155	1%	1
Rauðmaganet	Smallest Third	28	0%	0
Rauðmaganet	Middle Third	0	0%	0
Rauðmaganet	Largest Third	0	0%	0
Skötuselsnet	Smallest Third	105	1%	1
Skötuselsnet	Middle Third	0	0%	0
Skötuselsnet	Largest Third	0	0%	0
Þorskfisknet	Smallest Third	1454	7%	7
Þorskfisknet	Middle Third	523	3%	3
Þorskfisknet	Largest Third	996	5%	5

The “zero” boat proposal would change the role of inspections from an enforcement mechanism to a sampling mechanism to inform the “zero” boat enforcement strategy. In order to use inspections to develop the “zero” boat enforcement strategy and to still use inspections to target trips deemed risky for discard, the fishing directorate would have to run two different inspection programs. It would not be possible to combine the programs since for the “zero” boat methodology inspections need to be randomly determined.

Having two inspection programs where one randomly selects trips and other selects based upon a risk profile is similar to how the U.S. Internal Revenue Service (IRS) functions. The IRS uses a model to determine which taxpayers to audit based upon likelihood the taxpayer is evading taxes. The IRS cannot use these audited taxpayers as a sample representative of the population since they are not randomly selected. In order to calculate statistics representative of the entire taxpayer population, the IRS randomly selects taxpayers to audit under the National Research Program.[9]

[9] https://www.irs.gov/irm/part4/irm_04-022-001

Conclusions

Key Insights

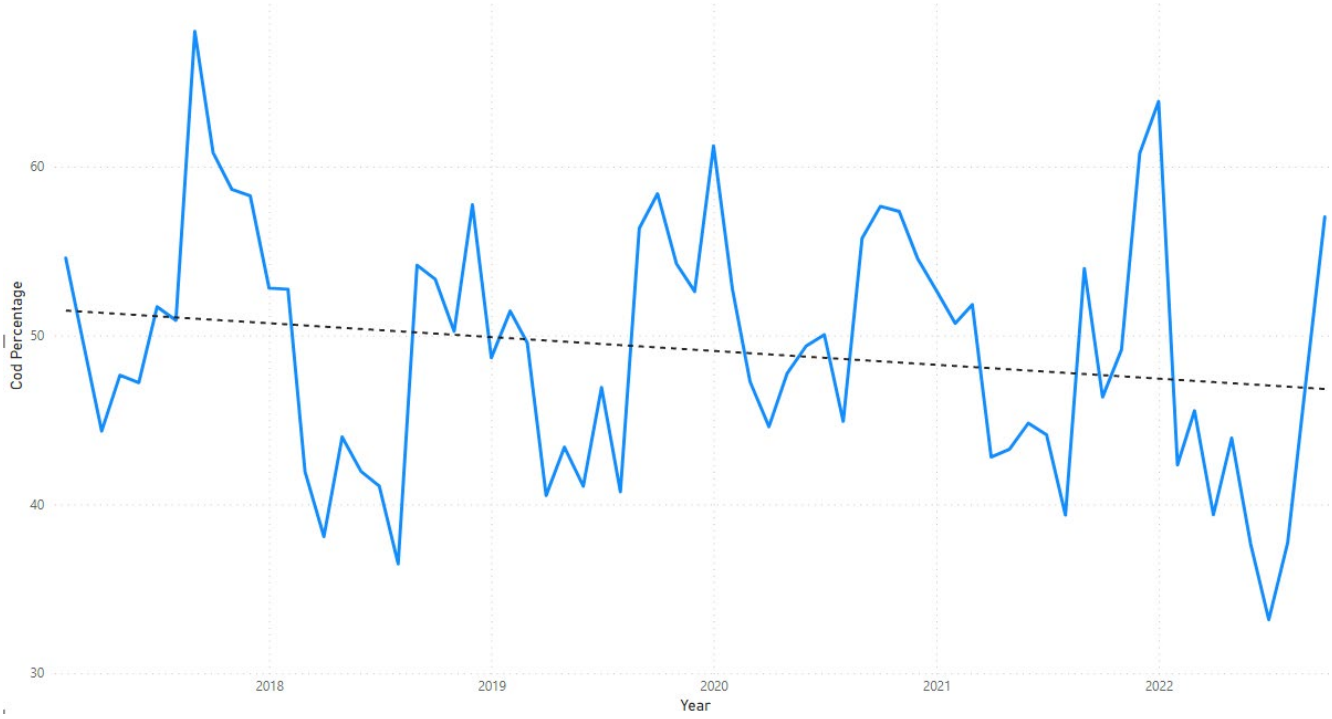
In our study, we aimed to address the problem of discard in the fishing industry by measuring its prevalence and trends using data science techniques. However, discard is a challenging problem to measure since the actual "catch" is not observed. Despite this, we were able to gain insights into the trends in the representation of species in landings.

The questions we focused on included the following:

- What are the indicators of discard and can we make intelligent guesses and estimates about the prevalence of this practice?
- Based on the indicators, what is the trend in discard? Are current methods aimed at reducing discard having an impact? Do we see trends in the data related to educational campaigns, inspectors, and drones?
- Finally, when (if ever) will we be able to conclude that discard no longer exists?

Cod percentage of total landing and species counts in landings are the best indicators of discard. These metrics are good benchmarks for discard because cod fishing is central to the Icelandic fishing industry at an aggregate level and provides significant financial incentives. Below is a chart which shows the monthly trend in cod percentage of total weight of landings from 2017 through 2022. Though there has been significant fluctuations by month, the overall trend in cod percentage is a decrease, indicative of a decrease in discard.

Cod Percentage Monthly Trend - 2017-2022

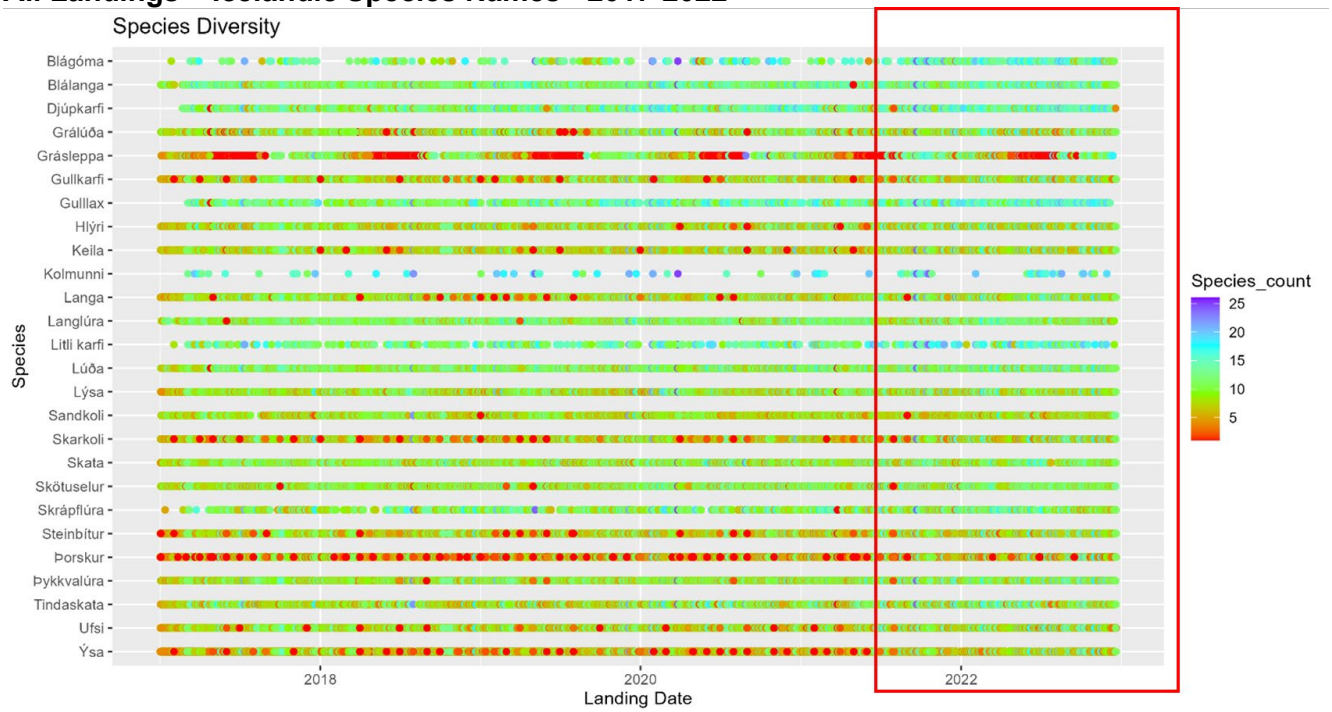


We found that dotplots and slice charts are effective tools to monitor the overall health of the ecosystem and fishing behaviors over time. Particularly in the dot plots and slice charts, we noticed decreases in cod percentages starting in 2021 that coincided with the implementation of drones.

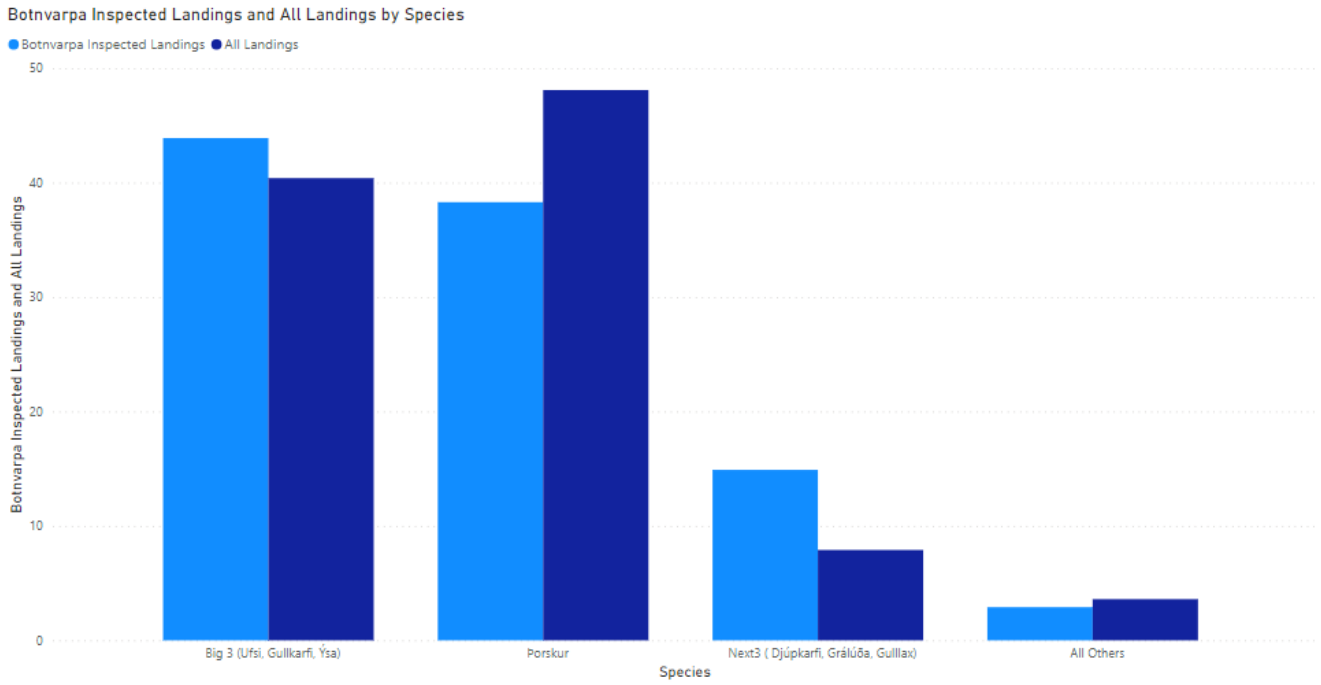
The graph which is most indicative is the slice chart shown below. As previously mentioned in this report, each dot is colored to represent the species count in the total catch. The area in the box,

which incidentally coincides with the introduction of drones in 2021, is significantly more green/blue and less red for most species, representing more species diversity in catches.

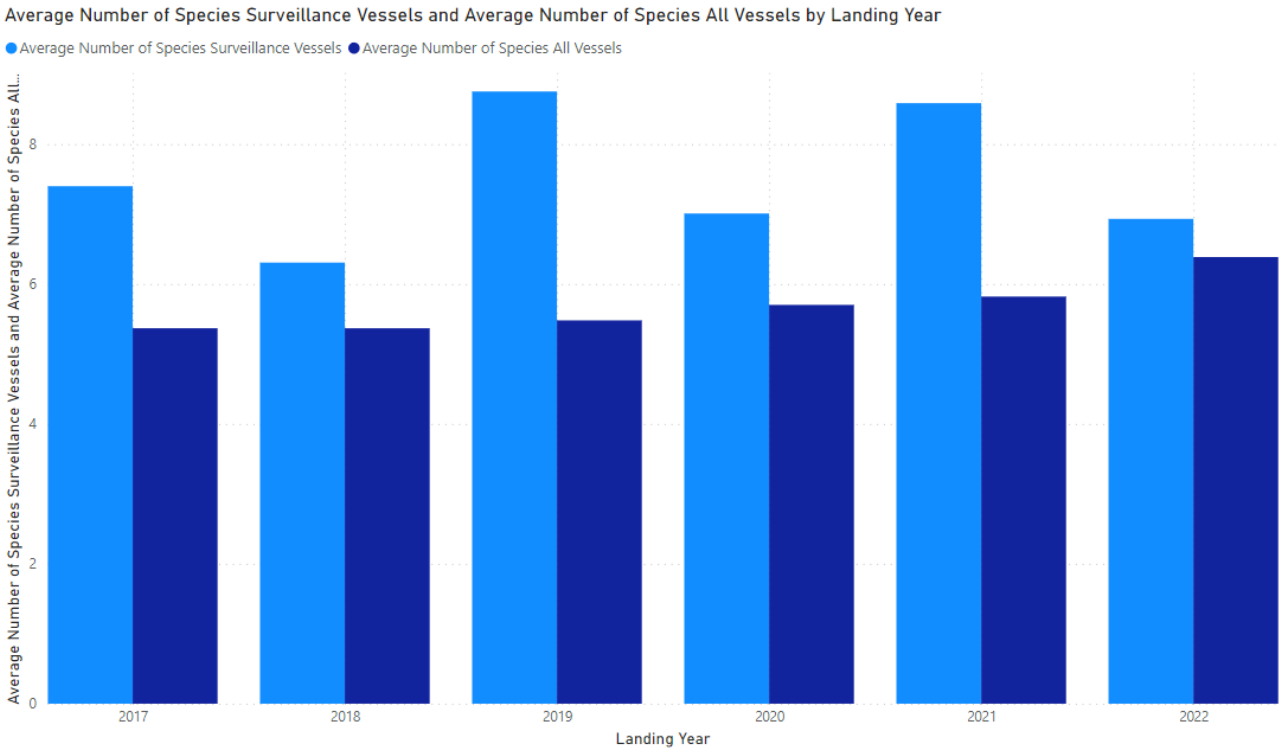
All Landings – Icelandic Species Names - 2017-2022



The presence of inspectors leads to markedly different reported catch totals in landings. As shown below, cod percentages of landed weight are lower (and percentages for other species are higher) in the presence of inspectors for landings that involved botnvarpa gear.



It should also be stated that there is an overall positive trend in the average number of species represented in inspected landings from 2017 to 2022 (see dark blue bars in the graph below).



We were not able to draw any statistically significant conclusions from the pricing data. We were not able to correlate fluctuations in price with activity. We suspect this may be the result of contracts between the fishing companies and the buyers, but we did not have visibility to the contractual commitments.

We were not able to build a "zero boat" baseline that would have predictive value for catches. Our attempts to model individual landings with tree classification, regression and unsupervised machine learning produced very high error rates. Our discussions with Fiskistofa employees suggested it might be possible to build a zero boat model, but such a model would require modeling elements of the Icelandic fishing industry that were beyond the scope of this study - such as the quota system or specific aspects of processing on fishing vessels. In order to build a zero boat baseline, we need to develop a better understanding of the “random” nature of current surveillance inspections and the extent to which the surveillance dataset is reflective of the overall fishing industry.

Recommendations

Data Governance/Management

Based on our findings, we recommend including LandingsID in pricing data. LandingsID is a key field that allows for the joining of data from multiple sources such as landings data and pricing data.

We recommend using something other than comma delimiter. As mentioned earlier in the data cleansing and validation section of this report, we ran into problems with comma as a delimiter when it is used as a decimal and within company names. A better delimiter would be a character that does not naturally occur in the dataset (such as a caret (^) or sharp (#)).

It would be helpful to aggregate pelagic and groundfish in the data to make segmentation for analysis easier.

We would suggest developing an export methodology (queries, text files, schema/data dictionary) that will allow the Fiskistofa team to send data to data scientist researchers. We hope that our work can serve as a foundation for future data analytics.

Operational Recommendations

We suggest pursuing technology-driven alternative means of inspection, including the use of drones, on board cameras, and data analytics to understand discard. Despite having fewer inspectors in 2008, inspection methods appear to have had a positive impact on the cod percentages, indicating a reduction in discard. In one of our discussions, it was suggested that the presence of birds could be an indication of discard, so the incorporation of satellite footage on birds might be something to pursue.

Fiskistofa inspection methods have the discard problem trending in a favorable direction, and they should continue down this path. Recent surveillance techniques including the use of drones in 2021 had an observable positive impact on diversity of species and cod percentages in landings.

Fiskistofa should potentially consider the statistical implications of their inspection strategy, including confidence interval and margin of error goals across gear types.

Finally, we recommend following and participating in discussions concerning a zero boat predictive model for discard. Iceland should engage with other nations who are progressing on this front. Techniques such as regression models, classification trees, or machine learning tools have the potential to be predictive. It is also important to consider how the zero boat model is received by the fishing industry and whether it is possible to develop any type of modeling around landing data.

Dashboard Visualization and Mobile App

Data Visualization

To make our team's findings

actionable by Fiskistofa, we plan to deliver an interactive Tableau dashboard that will enable Fiskistofa to easily identify opportunities to reallocate their resources to combat manipulation of the quota system.

Due to the declining number of Fiskistofa inspectors, it is increasingly important to develop methods that allow the agency to allocate scarce inspector resources efficiently and effectively. The primary visualization used in our dashboard is an extension of the Slice charts referenced in the Exploratory Data Analysis section but with additional layers of filtering and the ability to drill down into other key metrics.

The dashboard allows the end user to filter based on the following criteria:

- Inspected - whether the ship was inspected or not
- Gear - primary focus has been on Botnvarpa and Lina gear types but user has control over including/excluding any gear
- Top Species - this flag will only display the top 25 species of fish according to percentage of weight caught
- Heiti (Species) - allows the user to explicitly include/exclude any fish
- Ship Company
- Harbor
- Landing Date

In addition to the slice chart, the main dashboard contains trending landing weight, colored according to species diversity in the catch to assist in identifying catches with lower species diversity that may indicate discard/quota manipulation. Filtering the trending landing weight visual is also possible by selecting (clicking) any of the species from the slice chart.

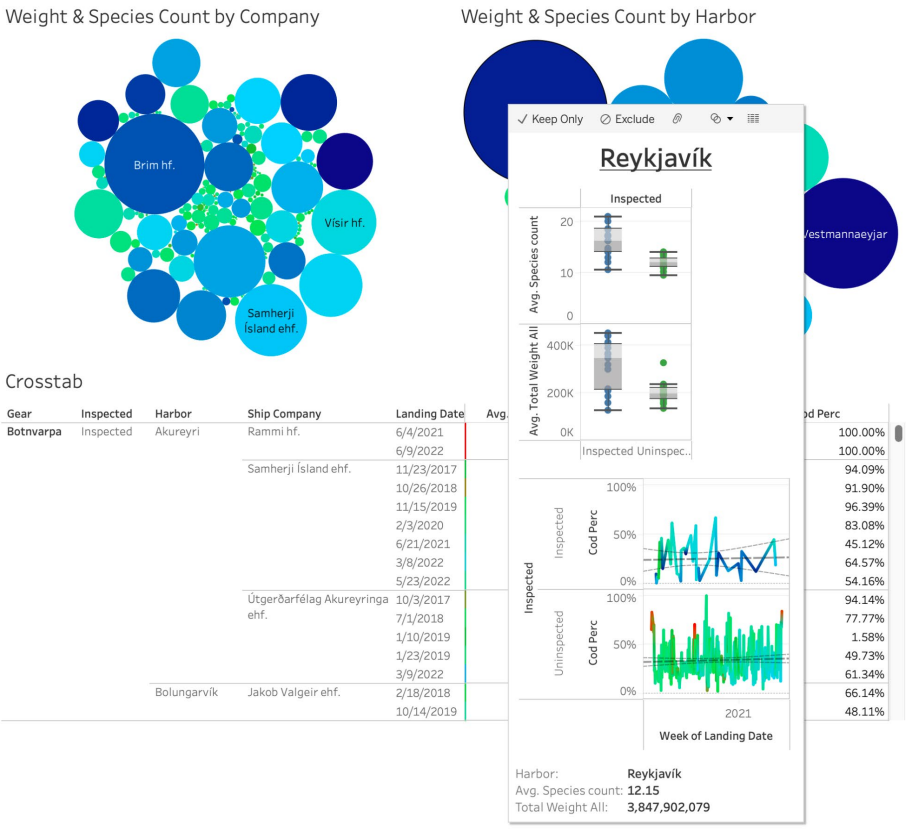
The secondary dashboard contains bubble charts for harbor and ship company, with bubble size displaying total weight and the shading of the bubbles displaying species diversity. This allows the end user to capture a high level understanding of the volume at each Harbor or from each company accounting for the filtering conducted based on the criteria above. Below the bubble charts contain crosstabs where the end user can drill down from Gear > Inspected/Uninspected > Harbor > Ship Company > Landing Date. Key metrics can be rolled up or drilled down into based on this hierarchy to understand how trends look at various levels of the hierarchy. These metrics are catch volume (kg), species diversity, and cod percentage (% of total catch weight coming from cod).

The bubble charts also allow the user to select individual bubbles to display box plots of species diversity and catch volume broken out by inspected/uninspected for the selected harbor/company. Additionally trending cod percentage will be displayed to enable visibility into whether inspected ships

are identifying higher risk ships over time and whether surveillance and inspections result in lower cod percentage due to enforcement.

The dashboard, as well as the mobile application (below) are housed in Tableau. As such there are three methods of delivery:

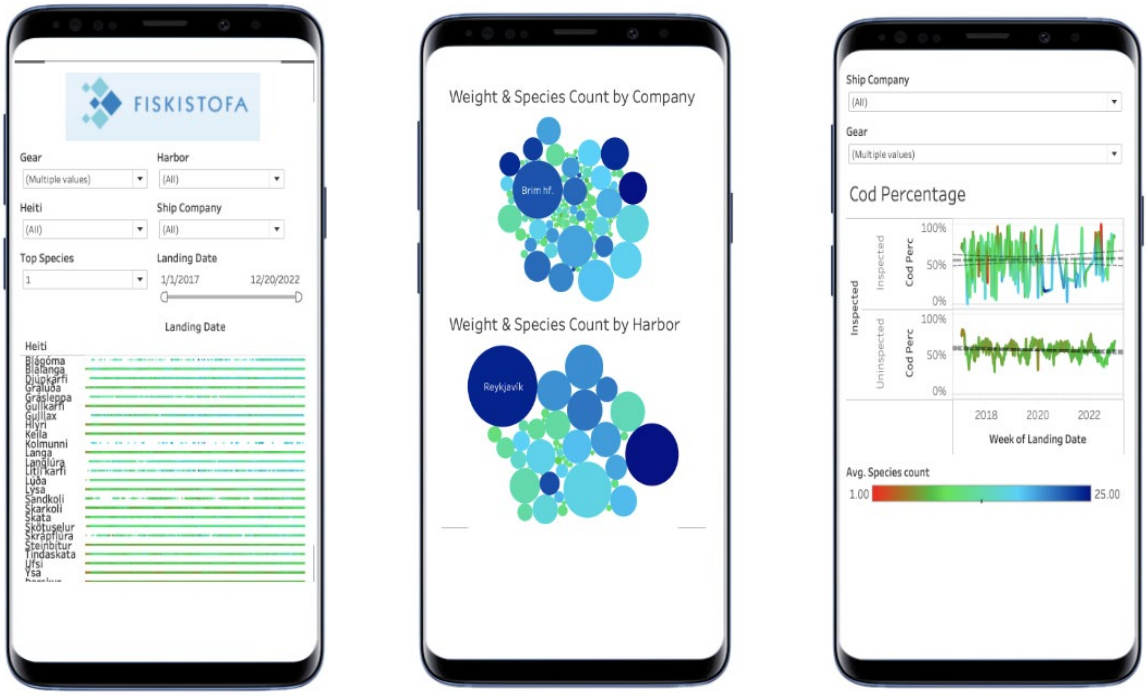
- Published to Fiskistofa's Tableau server
- Published to Tableau Public server
- Delivered as Tableau Packaged File



As we do not have publishing access to Fiskistofa's server and privacy concerns prevent us from publishing to a public server, we will deliver the dashboard as a packaged file. This will allow for maximum flexibility once Fiskistofa extends the project for further development. The packaged file format will also allow us to deliver the transformed data file that powers the dashboard to enable Fiskistofa the opportunity to continue leveraging the metrics identified throughout our project. This format will also allow for viewability using Tableau Reader, which is free to download and use to consume packaged Tableau dashboards.

Mobile App Development

The mobile version of the dashboard will largely be an extension of the most important aspects of our primary dashboard - the slice charts, bubble charts, and inspection breakdowns. While some of the interactive aspects of the primary dashboard will not be able to be extended to the mobile version, we envision this being used either on the go or in the field, to allow quick snapshots of activity.



Project Plan

Activity	Weeks										
	Pre-Course Start	1	2	3	4	5	6	7	8	9	10
Project Definition											
Team Selection											
Define Team Roles											
Project Goals Assignment											
Exploratory Data Analysis											
Model Build and Refinement											
Initial Findings/Exec Summary											
Model Selection/Recommendations											
Dashboad/App Design											
Final Report											
Final Presentation											
		Activity Duration				Deliverable Due					

Project Definition - With assistance from Fiskistofa team, identify the data set, business scenario, analytical direction.

Team Selection - Recruit the appropriate team for the project with similar interests and complementary skills.

Define Team Roles - Identify team strengths and interests to best support the project and delineate the scope of work for each role.

Project Goals - Establish the business case, proposed analysis, project timeline, and outcomes.

Exploratory Data Analysis - Conduct a preliminary analysis of the data to gain insights.

Model Build and Refinement - Develop and improve various models to predict the risk of discard in landings of catches

Initial Findings and Executive Summary - Deliver a report stating the problem, the approach taken, data analysis and preliminary conclusions.

Model Selection and Recommendations - The models will be evaluated via a variety of performance metrics as well as interpretability to derive recommendations to the client as well as begin dashboard and app design.

Dashboard and App Design - Deliver a functional dashboard and a preliminary beta application.

Final Report - Compose a comprehensive report explaining our work including our analysis, conclusions, and business recommendations.

Final Presentation - Deliver the final report in an online presentation, illustrating our problem, our analysis, takeaways and client recommendations. The app and dashboard will be showcased via a demonstration.

Team

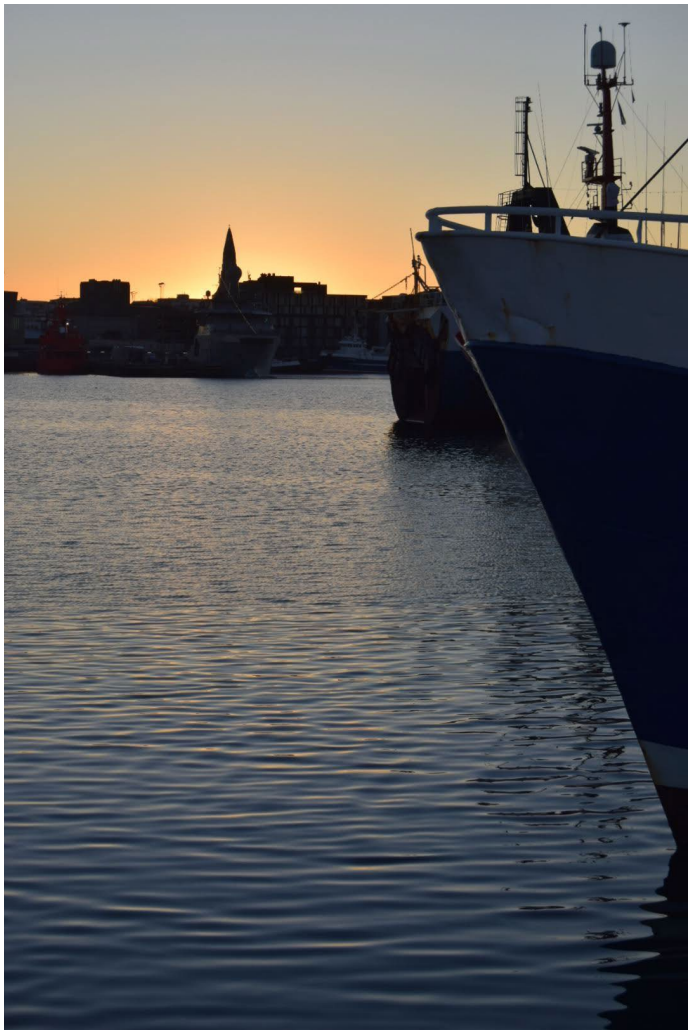
Joseph Hancuch is an economist with the U.S. Treasury Department focusing on international taxation and the spatial impact of the U.S. tax system.

Sean Johnson is a data analyst for Northern Trust primarily assisting with client implementation. Sean has limited professional experience, having graduated a year and a half ago with an undergraduate degree in Mathematics. Sean brings an analytical mind and will contribute to cleaning and exploring data, as well as developing predictive models.

Daniel Noel has worked in the field of software development for thirty years - most recently as a Principal Software Engineer for Sphera Corporation which developed tools for energy compliance. Dan has worked primarily in the Microsoft .NET stack with focus on the middle tier and back-end databases.

Dylon Polcik has been a marketing science professional for ten years focusing on experimentation and causal inference - most recently at Snap Inc.

Mark Schumacher is a finance and operations strategist, with over 20 years of experience at Accenture and EY. His goal is to help organizations utilize the power of data to build efficient, resilient and responsible business models.



Appendix

Codebase

The code base that was used in this project consisted of both python and R files. The main purpose of the code was to perform data transformation on the .csv files supplied by Iceland.

These .csv files were transformed using a series of 50 to 100 code files, each of which applied a different level of transformation to the data. This process was necessary to clean and standardize the data, making it easier to analyze and visualize.

In addition to the data transformation, the code base also attempted to create charts and data visualizations using R and PowerBI. This was done to help make the insights from the data more understandable and accessible to a wider audience.

The entire code base was deployed to Github, which provided a secure and centralized location for storing and sharing the code. This was an important aspect of the project as it allowed multiple team members to collaborate and work on the code at the same time, ensuring that everyone was working with the most up-to-date version.

Overall, the combination of Python and R programming languages, along with the use of popular libraries and tools such as Pandas, NumPy, ggplot2, dplyr, and PowerBI, allowed us to effectively transform and analyze the .csv data supplied by Iceland. The resulting code base provided a flexible and scalable solution for processing and exploring large datasets, enabling us to gain valuable insights into the data.

Software and Analytics Tools

- Python 3.5
- RStudio
- Jupyter Lab
- Microsoft Power BI
- Tableau Desktop
- Microsoft Excel
- GitHub

Data Dictionary

Landings Data

Variable Name	Variable Type	Brief Description
ShipNr	integer	Number between 1 and 9999 denoting the ship number; a negative

		number is a test ship
LandingDate	character	Date of the ship's landing
UnderSize	integer	Relevant to the quota; this tag pertains to fish not large enough to qualify under the quota system
Species	character	The Icelandic name for the species of fish caught on the landing
AmountKg	integer	The weight of fish caught in kilograms
IcePercentage	integer	Some fish are brought to shore packed with ice and this figure corresponds to the percentage of ice in such cases
Harbor	character	The harbor in Iceland where the ship made landfall
QuotaType	character	The quota system is divided into different subsystems and this field designates the type of quota to which this catch is applied
GuttedOrNot	character	Many ships gut the fish on board to promote quicker processing on shore
FishingArea	character	The area where the fish was caught (typically Icelandic fishing waters)
Gear	character	Corresponds to the type of gear used to catch the fish
StorageMethod	character	The method of storage; the majority are Iced (Ísað)
ShipType	character	The type of ship used in the landing
WeightingType	character	The method of weighing; most are Harbor Scale (Hafnarvog) or License Holder (Vigtunarleyfishafi)
FullProcess	integer	Unclear; not used in analysis
ShipCompany	character	The entity responsible for fishing
ShipOwner	character	The owner of the ship, often the same as the ship company
Receiver	character	The receiver of fish for processing
Length	numeric/float	Unclear; not used in analysis

Pricing Data

Variable Name	Variable Type	Brief Description
LandingDate	character	Data of the ship's landing
ShipNR	integer	Number between 1 and 9999 denoting the ship number; a

		negative number is a test ship
ShipName	character	Name of the ship
Company	character	The entity responsible for fishing
SaleType	character	The type of sale; most are Markets (Markaðir)
FishingGear	character	Corresponds to the type of gear used to catch the fish
Species	character	The Icelandic name for the species of fish caught on the landing
GuttedUngutted	character	Many ships gut the fish on board to promote quicker processing on shore
CoolingMethod	character	The method of cooling in storage; the majority are Iced (Ísað)
QuotaUse Afdrif	character	The quota system is divided into different subsystems and this field designates the type of quota to which this catch is applied
ProcessType Radstofun	character	Unclear; not used in analysis
Price kr/kg	integer	Price of fish in Icelandic Króna per kilogram
Amount kg	integer	Amount of fish sold in kilograms
TotalPrice	integer	The total price of fish dependent on the amount and price columns
Buyer	character	The buyer of the fish
STADA_SOLU	integer	Unclear; not used in analysis

Description of Variables

Numerical Variables in Landings Data

	count	mean	std	min	25%	50%	75%	max
ShipNr	1249660.00	3127.51	1880.68	78.00	2069.00	2685.00	2905.00	9999.00
LandingID	1249660.00	705984.95	87053.09	554967.00	630052.00	707673.00	778241.25	854826.00
UnderSize	1249660.00	0.08	0.27	0.00	0.00	0.00	0.00	1.00
AmountKg	1249660.00	6278.37	72016.70	1.00	30.00	188.00	874.00	3448130.00
IcePercentage	978764.00	1.67	55.36	-9650.00	2.00	3.00	12.00	99.94
FullProcess	1249660.00	0.02	0.12	0.00	0.00	0.00	0.00	1.00
Length	1249660.00	21.57	16.09	0.00	10.40	14.50	28.90	113.40

Categorical Variables in Landings Data

	count	unique	top	freq
LandingDate	1249660	2168	08.05.2017	1897
Species	1249660	126	Þorskur	403886
Harbor	1249660	69	Ólafsvík	93867
QuotaType	1249660	8	Til vinnslu	1057158
GuttedOrNot	1249660	37	óslægt	810263
FishingArea	1249660	8	Ísland	1240903
Gear	1249660	23	Handfæri	268519
StorageMethod	1249660	5	Ísað	1223681
ShipType	1249660	14	Krókaaflamarksbátur	465337
WeighingType	1249660	4	Hafnarvog	634819
ShipOwner	1249660	1507	Skinney-Þinganes hf.	30509
ShipCompany	1249660	1526	Nesfiskur ehf.	42647
Receiver	1249660	644	Fiskmarkaður Íslands hf.	274202

Numerical Variables in Pricing Data

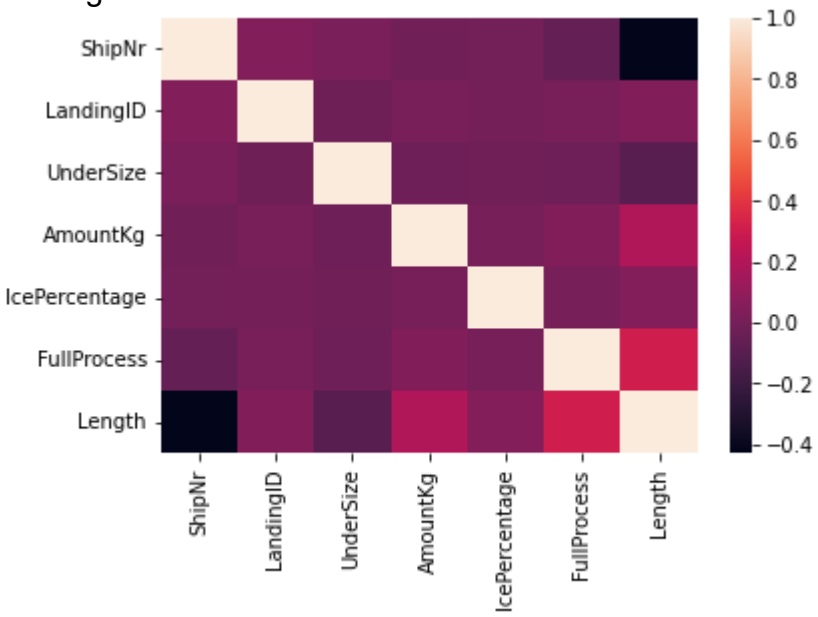
	count	mean	std	min	25%	50%	75%	max
ShipNR	1566206.00	3263.04	1999.99	78.00	2090.00	2704.00	2929.00	9999.00
Price kr/kg	1565845.00	241.32	167.45	0.00	133.00	220.00	317.00	29697.00
Amount kg	1566206.00	1530.55	25868.93	-1669130.00	32.00	125.00	503.00	3262101.00
TotalPrice	1565845.00	500665.64	3801607.11	-48404770.00	6536.00	41552.00	179692.00	523121197.50
STADA_SOLU	1566206.00	3.00	0.00	3.00	3.00	3.00	3.00	3.00

Categorical Variables in Pricing Data

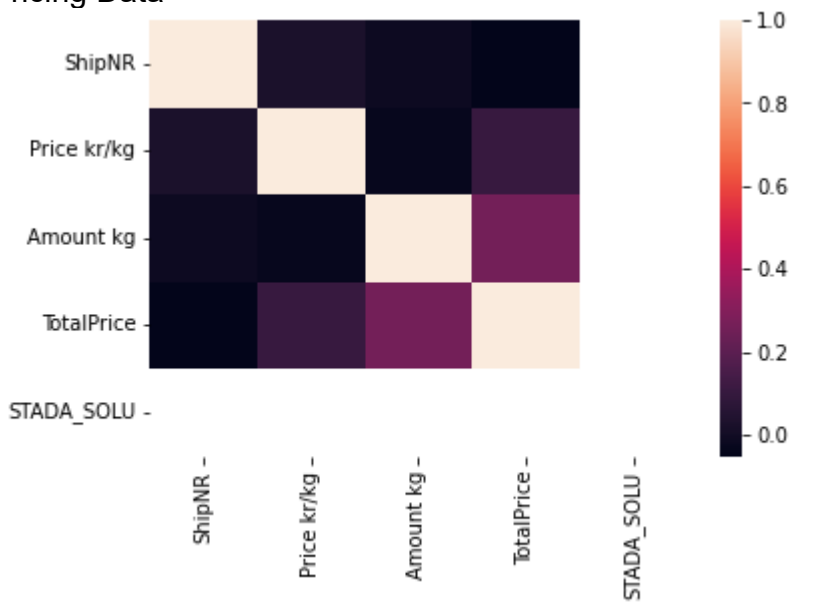
	count	unique	top	freq
LandingDate	1249660	2168	08.05.2017	1897
Species	1249660	126	Þorskur	403886
Harbor	1249660	69	Ólafsvík	93867
QuotaType	1249660	8	Til vinnslu	1057158
GuttedOrNot	1249660	37	óslægt	810263
FishingArea	1249660	8	Ísland	1240903
Gear	1249660	23	Handfæri	268519
StorageMethod	1249660	5	Ísað	1223681
ShipType	1249660	14	Krókaaflamarksbátur	465337
WeighingType	1249660	4	Hafnarvog	634819
ShipOwner	1249660	1507	Skinney-Þinganes hf.	30509
ShipCompany	1249660	1526	Nesfiskur ehf.	42647
Receiver	1249660	644	Fiskmarkaður Íslands hf.	274202

Correlation Plots

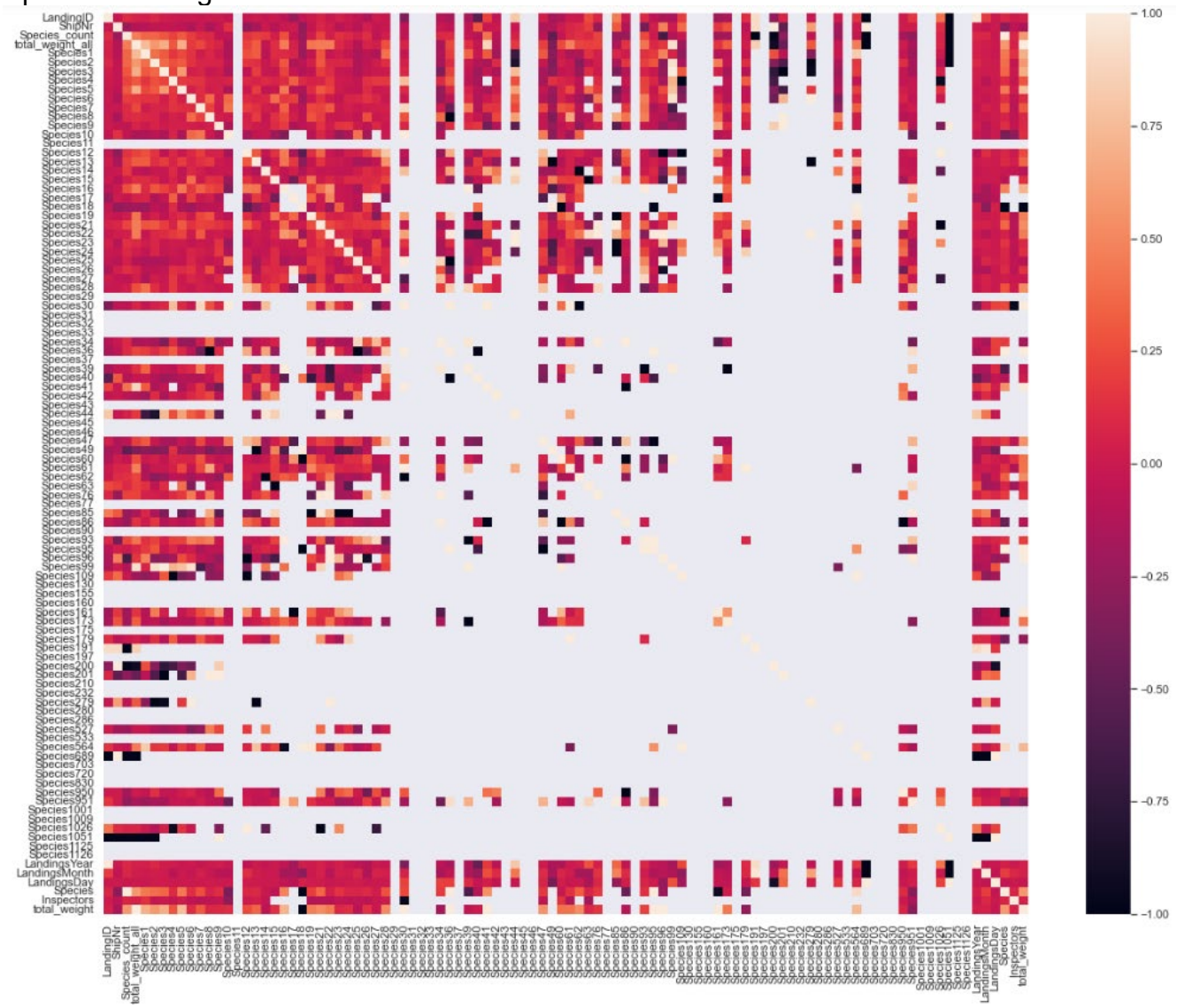
Landings Data



Pricing Data



Species Landings Data



Classification Model Metrics

TP = True Positive

FP = False Positive

TN = True Negative

FN = False Negative

Accuracy: Total number of correct predictions divided by total number of observations

Precision: The fraction of relevant instances among all retrieved instances

Precision = TP / (TP + FP)

Recall: The fraction of retrieved instances among all relevant instances; also referred to as "sensitivity"

Recall = TP / (TP + FN)

AUC is the area under the ROC curve. The AUC represents a model's ability to differentiate between positive and negative classes. An area of 1.0 represents a model that made all predictions perfectly. An area of 0.5 represents a model as good as random.

Samherji Newsletter Article

Mark Schumacher and Dan Noel traveled to Iceland December 11-21 to work with the Fiskistofa team. The purpose of the trip was to define goals and objectives of the project, gather relevant data, and learn about the challenges faced by Fiskistofa as well as the industry in responsibly managing Iceland's fisheries, a vital natural resource important to the Iceland economy as well as global food sources. Mark and Dan along with the Fiskistofa team observed the landing of a catch of 40 tons of cod at the Akureyri harbor. The next morning, the group toured the Samherji fish processing plant to observe the processing of the 40 tons of cod. Below is an article that appeared in the Samherji company newsletter.

SAMHERJI
ICE FRESH SEAFOOD

CONTACT | ENGLISH

THE COMPANY FISH FARMING LAND PROCESSING SALES SHIPPING COMPANY

An American university project in data science extends to Akureyri

Ögmundur Knútsson, director of the fisheries department, Mark Schumacher, master's student, Dóróthea Jónsdóttir, project director of the Department of Fisheries, and Daniel Noel, master's student.

Two Master's students in *Data Sciences*, Mark Allen Schumacher and Daniel Robert Noel at Northwestern University in Chicago, USA, are in Akureyri in connection with a project organized by the Norwegian Fisheries Agency. Cooperation with Samherja and Útgerðarfélag Akureyringa on the project was requested, as the level of technology in the companies' ships and processing houses is high.

.

Ögmundur Knútsson, head of the fisheries department, welcomes the collaboration. He says that the information will be used, among other things, to promote and develop a variety of controls in the industry, which will increase consumer confidence in Icelandic seafood.

Kristján Vilhelmsson, manager of Samherja's shipping division, says that with increased technology, new possibilities are created.

It is important that everyone works together

"For fisheries companies, a large amount of information will be generated that can be used both to develop controls as well as to strengthen the science behind the advice of the Norwegian Maritime Authority. The Ministry of Food has launched a project related to the mapping of information collected in the

Mark Schumacher, master's student, Ólafur Pálmi Guðnason, development and data manager of the Fisheries Agency, Daniel Noel, master's student and Ögmundur Knútsson, fisheries agency manager

process from fishing to consumers. Part of this project is then to look at how information from the industry can be used to improve surveillance and science. It is important that all parties work together to strengthen and enhance the image of Icelandic seafood in competitive markets. Part of that is to ensure that supervision and the science behind advice is of the highest quality," says Ögmundur Knútsson.

A huge amount of information gathered every day of the year

Kristján Vilhelmsson, managing director of Samherja's shipping division, says that it has not been up to companies in the industry to work with the government and the academic community.

"We have a huge amount of information that is collected automatically throughout the year, from the beginning of fishing until the fish is taken out of the country. Of course, this data can be used in a variety of ways. With the public sector and the scientific community making

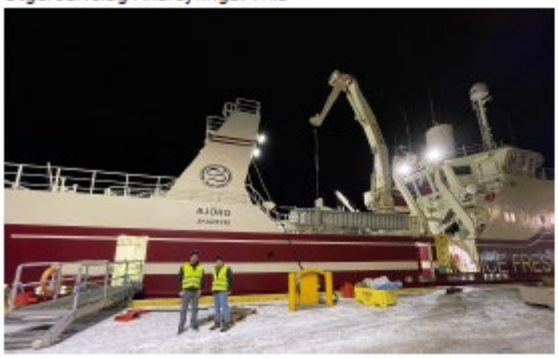
Magnús Finnsson represented Samherji

use of our data, it is likely that the system will be simplified and made more efficient. With increased technology, new possibilities are created, which, for example, provide an opportunity to reduce unnecessary costs and further ensure the reliability of fishing and processing."

Followed up with landing from Björgu EA and processing of products in ÚA

"Master's students from Northwestern University, who have extensive knowledge in this field of data science, contacted Fiskistofa to explore the possibility of doing their final project with us, which is a pleasure and a certain recognition. Four students are involved in this project and two of them are currently in Akureyri to do preliminary research. They observed the landing of catches from the trawler Björgu EA and then all aspects of the fish processing activities of Útgerðarfélag Akureyringa. This

Útgerðarfélag Akureyringa. This



is an exciting project in many ways, and I welcome this collaboration," says Ögmundur Knútsson, director of the fisheries office.

Ögmundur Knútsson, director of the fisheries department, Dóróthea Jónsdóttir, project manager of the Department of Fisheries, Mark Schumacher, master's student, and Daniel Noel, master's student

Article weblink:
https://www.samherji.is/is/frettir/bandariskt-haskolaverkefni-i-gagnavisindum-teygir-sig-til-akureyrar

Note: Photo Credits
All photos in this document with the exception of the processing plant photos in the Samherji article were taken by Mark Schumacher, Dan Noel, and Dóróthea Jónsdóttir of Fiskistofa.