Final Project Presentation

Improved prediction of salmon runs using historical salmon and weather data to enhance business planning and resource allocation

MGT 6203 Spring 2023 - Team 55

Team Information

Project context and research objective

Background & Business Justification

475,534 residents and non-resident licensed anglers fished 2.5 million days in Alaska in 2007

$1.4 billion spending on licenses and stamps, trip-related expenditures, pre-purchased packages, equipment and real estate used for fishing

Supported 15,879 jobs in Alaska and provided $545 million of income

$246 million in tax revenues generated for Alaska governments

*Based on a 2007 research paper published in 2008 on Economic Impacts and Contributions of Sportfishing in Alaska

Salmon fishing is one of the most valuable industries in Alaska and is a cornerstone of the state's economy.

Predicting fish count and providing insights into the key factors that influence salmon populations can allow businesses, local government, and the tourists to make more informed business decisions

Primary Research Objective:

Using the data our team has collected, can we create an accurate fish count prediction model?

Supporting Research Objective:

What variables have the most significant impact on fish count?

Data overview

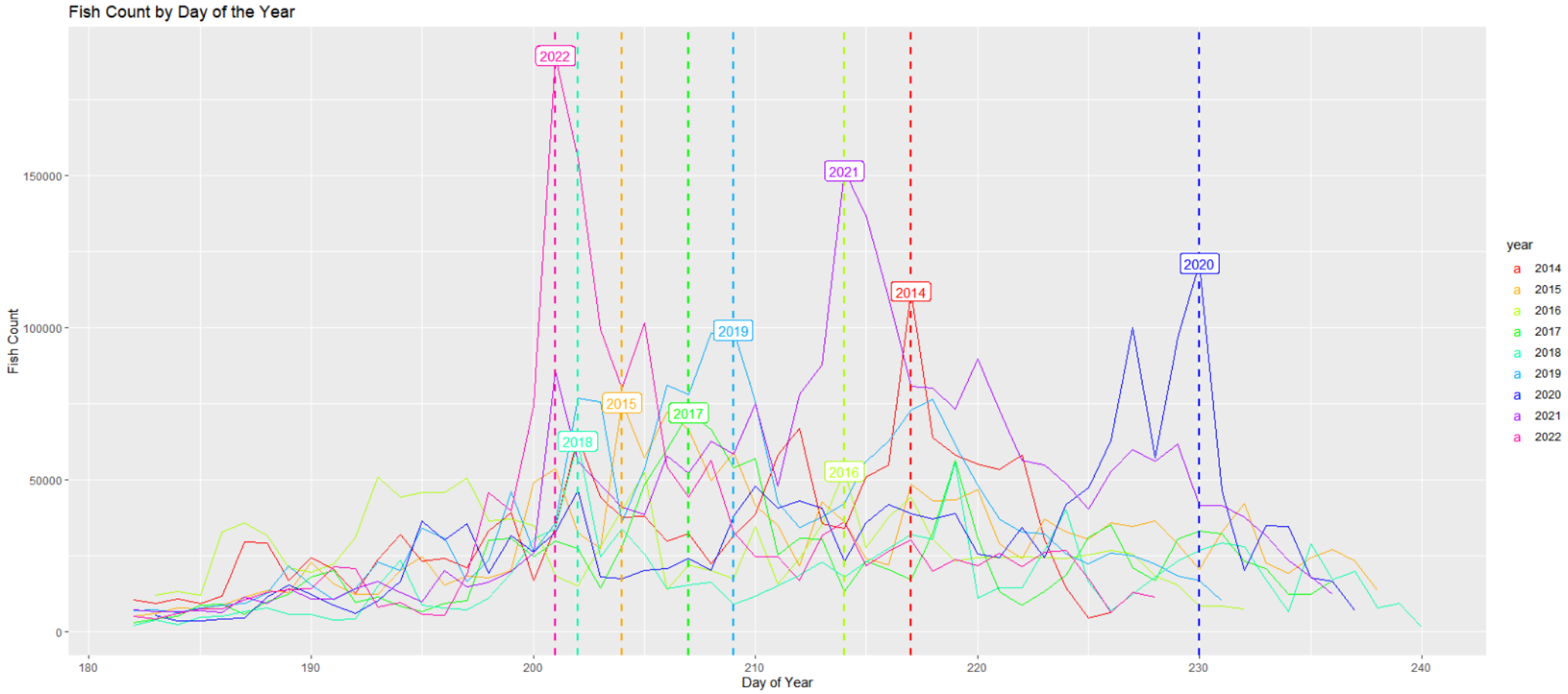
Data Sources and Cleaning

Description	Source	Unit	Data type
Daily sockeye salmon fish count at Kenai River (Late-Run Sockeye)	Alaska Department of Fish and Game	Count	Integer
Daily minimum, maximum and mean air temperatures at Kenai AP	Alaska Climate Research Center	Degree Fahrenheit	Integer
Daily Precipitation at Kenai AP	Alaska Climate Research Center	Inch	Integer
Daily river flood stage at Kenai River (Kenai Keys)	National Weather Service	Stage	Categorical
River discharge	USGS National Water Information System	ft3/s	Integer
Kenai River water temperature data at Soldotna	USGS National Water Information System	Degree Celsius	Integer
Moon Phase (split into 4 categories)	Papers with Code (Blog)	Moon Phase	Categorical
Set Net Location	Data from AF&G Biologist	Location	Categorical
Drift Net Location	Data from AF&G Biologist	Location	Categorical
Nets	Data from AF&G Biologist	Drift, Set, Both, no_nets	Categorical

Exploratory Data Analysis

Fish count peak analysis from 2014 to 2022

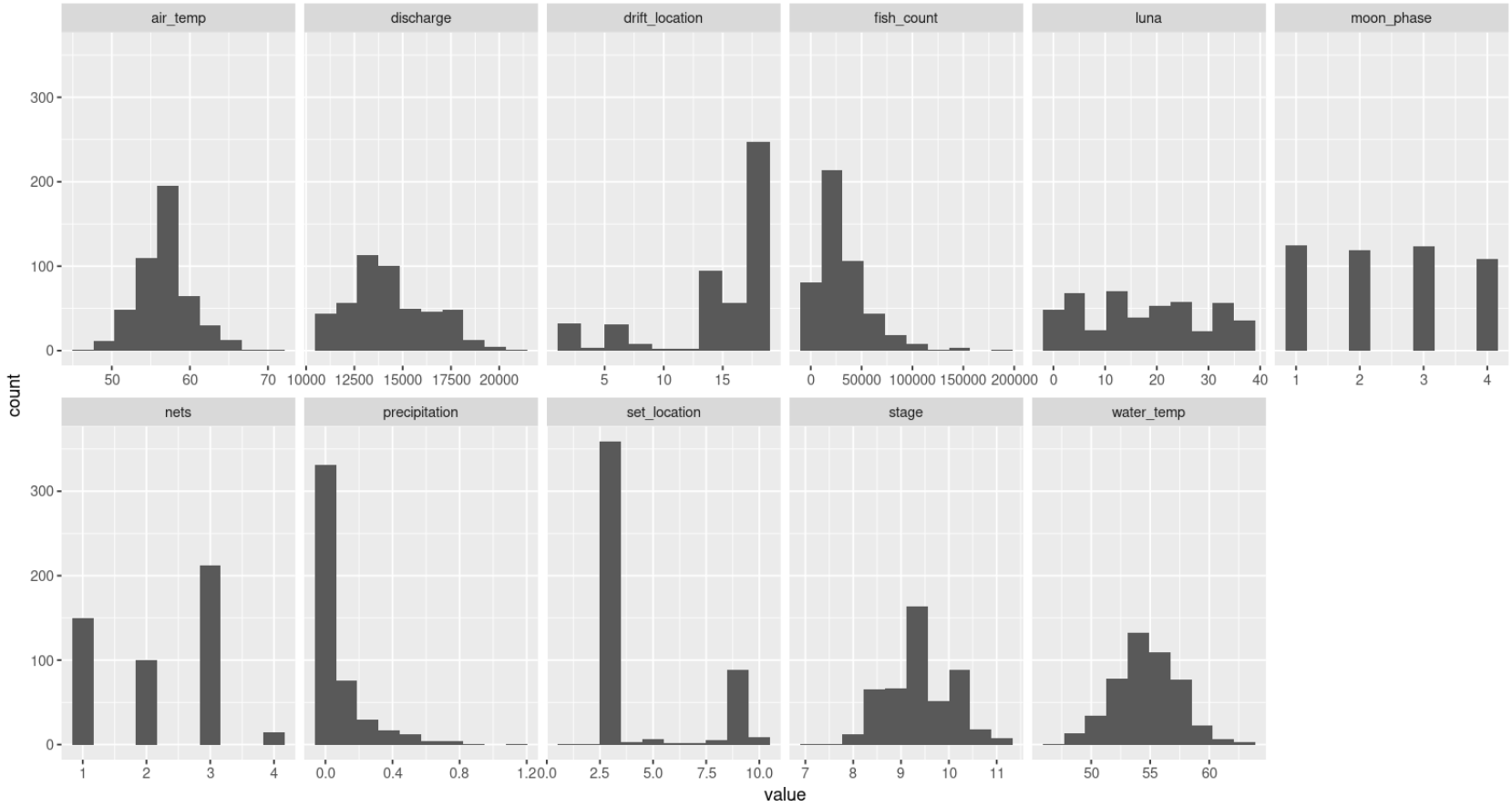
- All peaks (except for 2020) occurred between July 19 and August 8
- 7 out of the last 9 years saw peaks between July 19 and July 29



Exploratory Data Analysis

Combined distribution bar graphs of all variables

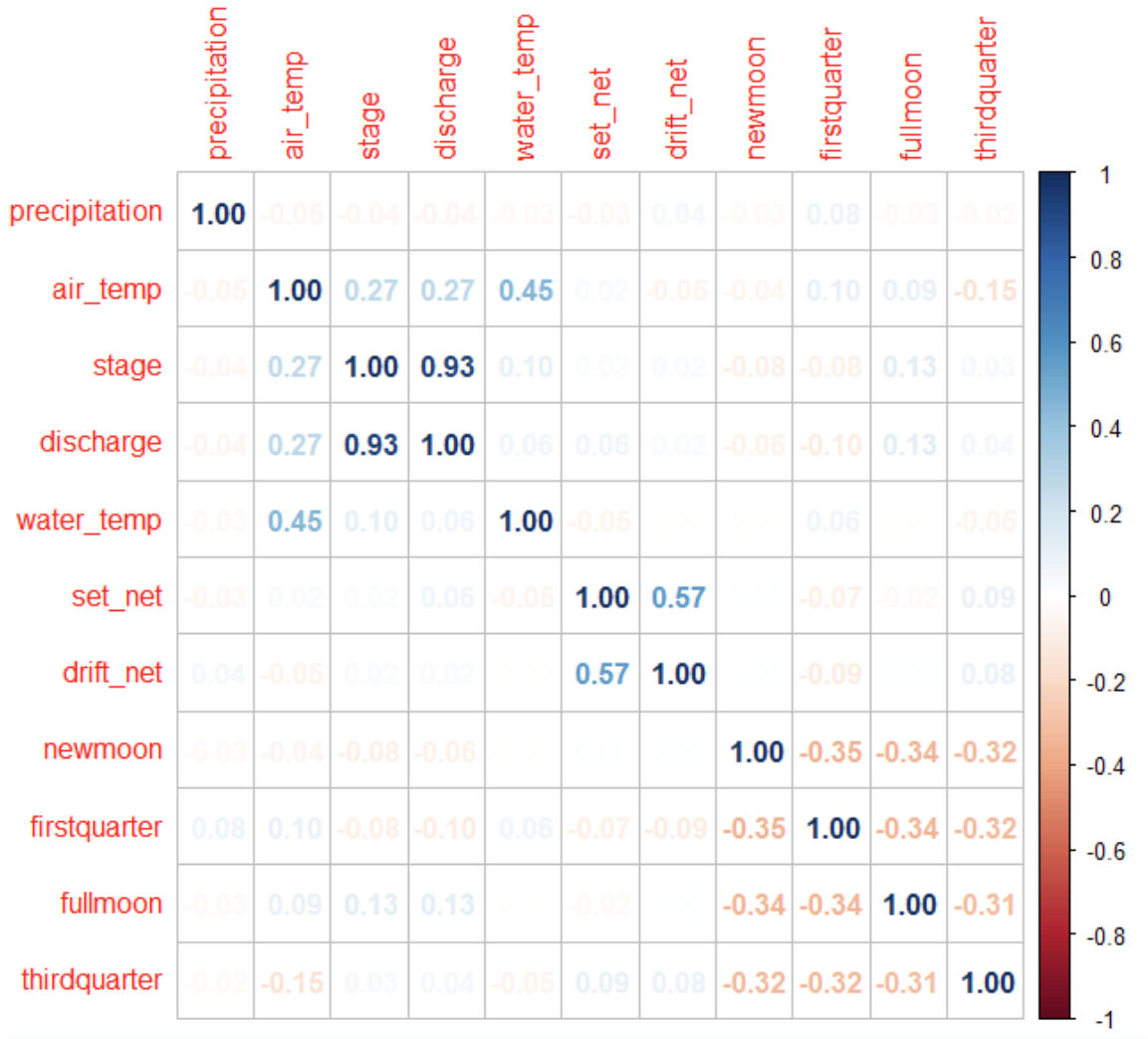
- Distributions of most attributes are not normal, except air temperature, discharge and water temperature
- Fish count and precipitation and exhibit strong negative skew
- Presence of commercial nets, moon phase and river stage exhibit a multimodal distribution



Exploratory Data Analysis

Correlation plot of numeric variables

- Very strong correlation between stage and river water discharge
- Slight to moderate correlation
  - Air temperature and water temperature
  - Set net and drift net
  - Moon phases



	precipitation	air_temp	stage	discharge	water_temp	set_net	drift_net	newmoon	firstquarter	fullmoon	thirdquarter
precipitation	1.00	-0.05	-0.04	-0.04	-0.03	-0.03	0.04	-0.03	0.08	-0.03	-0.02
air_temp	-0.05	1.00	0.27	0.27	0.45	0.02	-0.05	-0.04	0.10	0.09	-0.15
stage	-0.04	0.27	1.00	0.93	0.10	0.02	0.02	-0.08	-0.08	0.13	0.03
discharge	-0.04	0.27	0.93	1.00	0.06	0.06	0.02	-0.06	-0.10	0.13	0.04
water_temp	-0.03	0.45	0.10	0.06	1.00	-0.05	0.01	0.01	0.06	0.01	-0.05
set_net	-0.03	0.02	0.02	0.06	-0.05	1.00	0.57	0.00	-0.07	-0.02	0.09
drift_net	0.04	-0.05	0.02	0.02	0.01	0.57	1.00	0.00	-0.09	0.01	0.08
newmoon	-0.03	-0.04	-0.08	-0.06	0.01	0.00	0.00	1.00	-0.35	-0.34	-0.32
firstquarter	0.08	0.10	-0.08	-0.10	0.06	-0.07	-0.09	-0.35	1.00	-0.34	-0.32
fullmoon	-0.03	0.09	0.13	0.13	0.01	-0.02	0.01	-0.34	-0.34	1.00	-0.31
thirdquarter	-0.02	-0.15	0.03	0.04	-0.05	0.09	0.08	-0.32	-0.32	-0.31	1.00

Modelling


183.102
154.17


CONCEPTUAL SITE PLAN -
Parcel
(private property)

Linear regression

- To find the strength of relationships between our various variables and their effects on fish count
- To create a fish count predictor model

Linear Regression Output

```
Coefficients: (1 not defined because of singularities)
                                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                                       -304214.3   100993.3  -3.012 0.002740 **
log(discharge)                                     -15462.9     6631.7  -2.332 0.020158 *
log(water_temp)                                    119486.0    20248.9   5.901 7.12e-09 ***
moon_phaseFull Moon                                -6365.6     2654.2  -2.398 0.016879 *
moon_phaseNew Moon                                 -3786.5     2613.7  -1.449 0.148111
moon_phaseThird Quarter                            10623.7     2762.2   3.846 0.000137 ***
drift_locationAll                                   7705.2     4534.6   1.699 0.089970 .
drift_locationArea 3                               25520.9    20391.6   1.252 0.211389
drift_locationArea 3, KRSHA                        -1526.2    20628.6  -0.074 0.941055
drift_locationDistrict Wide except Chinitna Bay Sub -10438.8   14543.0  -0.718 0.473262
drift_locationDrift Area 1                         -19405.9    20354.7  -0.953 0.340906
drift_locationDrift Area 1 & Expanded Corridor      18288.7     9423.4   1.941 0.052908 .
drift_locationDrift Area 1&2, Ex Ken/Kas Sec        51192.0    20641.4   2.480 0.013501 *
drift_locationDrift Area 1, Ex Ken/Kas Sec          25626.7     4859.8   5.273 2.09e-07 ***
drift_locationDrift Area 1, Ex. Ken/Kas & AP sec    48042.9    14800.9   3.246 0.001258 **
drift_locationDrift Area 1, Exp. Ken/Kas, & Anchor Pt.  8215.4  8385.2   0.980 0.327741
drift_locationDrift Areas 1 & 3, Ex. Ken/Kas sec    17694.6    20644.9   0.857 0.391849
drift_locationDrift Areas 1, 3 and 4                10680.5    20401.8   0.524 0.600877
drift_locationDrift Areas 3                           923.7    20377.2   0.045 0.963863
drift_locationDrift Areas 3 & 4                      3898.2     4476.5   0.871 0.384319
drift_locationExp. Ken/Kas, & Anchor Pt.            31429.5     3191.1   9.849  < 2e-16 ***
drift_locationExpanded Kenai & Kasilof Sections     13061.5     5071.7   2.575 0.010331 *
drift_locationKasilof River Special Harvest Area    20619.3     5024.4   4.104 4.83e-05 ***
drift_locationKasilof Section                        4324.1     5679.0   0.761 0.446810
netsboth                                           -16857.0     3328.6  -5.064 5.99e-07 ***
netsdrift                                               NA         NA      NA       NA
netsset                                             14007.7     5622.3   2.491 0.013081 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20230 on 450 degrees of freedom
Multiple R-squared:  0.3486,    Adjusted R-squared:  0.3125
F-statistic: 9.635 on 25 and 450 DF,  p-value: < 2.2e-16
```

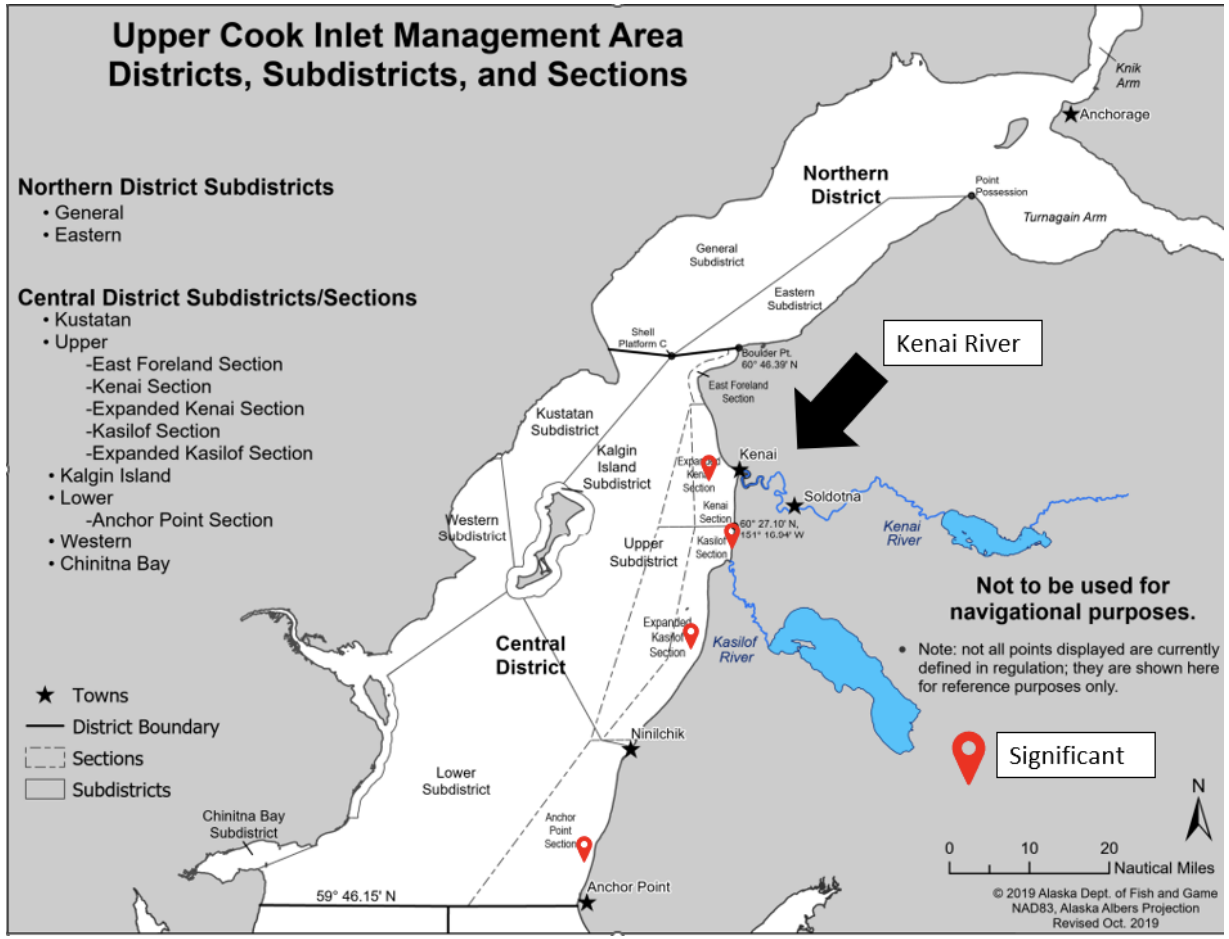
Linear regression

Water temperature

Linear regression

Statistically significant drift locations

- All
- Drift Area 1 & Expanded Corridor
- Drift Area 1&2, Ex Ken/Kas Sec
- Drift Area 1, Ex Ken/Kas Sec
- Drift Area 1, Ex Ken/Kas Sec & AP Sec
- Exp. Ken/Kas, & Anchor Pt.
- Expanded Kenai & Kasilof Sections
- Kasilof River Special Harvest Area

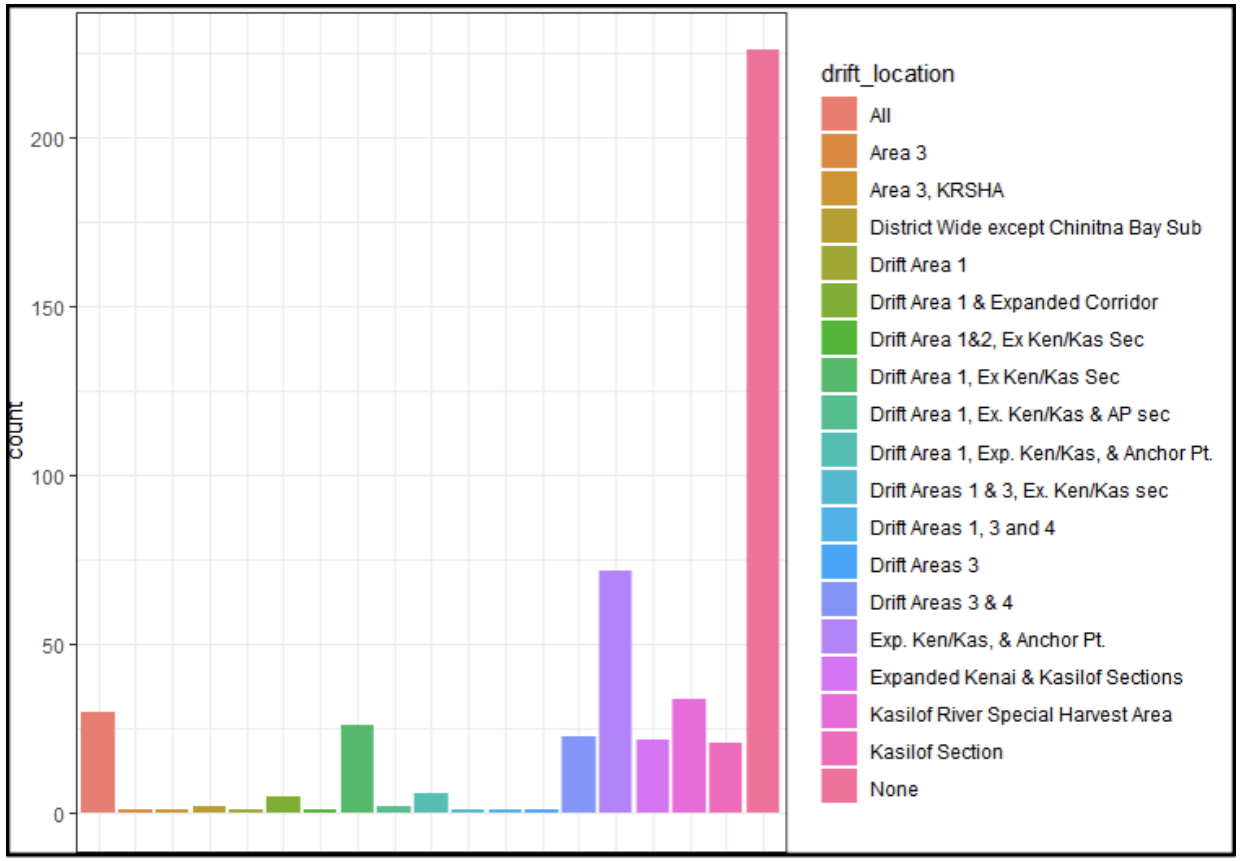


Linear regression

Explaining positive regression coefficients

- Insufficient historic data
- Significant simply because these locations have the highest fish count

Bar chart showing the count of drift fishing locations



Logistic regression

- To predict the probability of a 'high' fish count day – when the fish count was above 30,737, the mean of the dataset

Logistic regression

Interpreting the coefficients

- Presence of drift nets and water temperature were the two most significant factors
- They had a positive coefficient, implying a positive correlation with the log odds of a high fish count

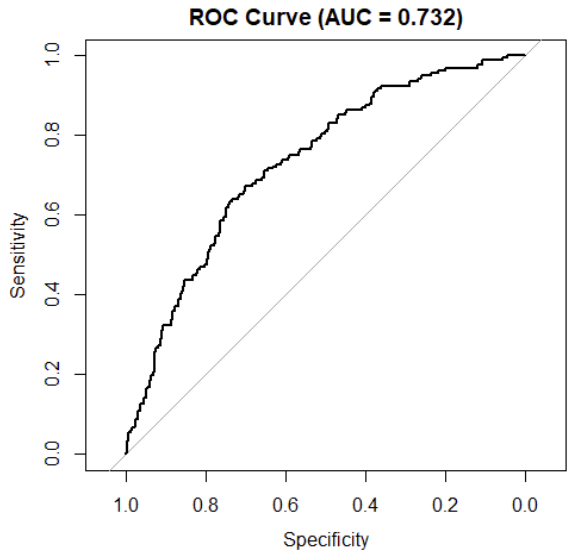
Logistic Regression Output

```
Call:
glm(formula = high_fish ~ precipitation + air_temp + water_temp +
    discharge + set_net + drift_net + moon_phase, family = "binomial",
    data = fm_logit)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0765  -0.9102  -0.6112   1.0831   2.1369

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -1.143e+01  2.457e+00  -4.653 3.27e-06 ***
precipitation           -1.365e+01  5.354e+02  -0.025   0.9797
air_temp                -6.847e-02  3.692e-02  -1.855   0.0636 .
water_temp               2.845e-01  4.787e-02   5.943 2.80e-09 ***
discharge               -7.718e-05  5.195e-05  -1.486   0.1374
set_net                 -4.552e-01  2.647e-01  -1.720   0.0855 .
drift_net                9.930e-01  2.551e-01   3.893 9.90e-05 ***
moon_phaseFull Moon     -5.519e-01  2.973e-01  -1.856   0.0634 .
moon_phaseNew Moon      -3.200e-02  2.814e-01  -0.114   0.9095
moon_phaseThird Quarter  4.390e-01  2.950e-01   1.488   0.1367
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 633.25  on 474  degrees of freedom
Residual deviance: 557.61  on 465  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 577.61

Number of Fisher Scoring iterations: 12
```

Logistic regression

ROC plot and Confusion Matrix

- AUC metric of 0.732 indicates that there is about a 73.2% chance that the model will be able to accurately distinguish between a positive and negative class
- Sensitivity of 0.044 and Specificity of 0.993 suggest that the model could predict true negatives very well but true positives poorly




```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 290 175
         1   2   8

               Accuracy : 0.6274
                 95% CI : (0.5821, 0.671)
    No Information Rate : 0.6147
    P-Value [Acc > NIR] : 0.303

                  Kappa : 0.0448

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.04372
            Specificity : 0.99315
         Pos Pred Value : 0.80000
         Neg Pred Value : 0.62366
             Prevalence : 0.38526
         Detection Rate : 0.01684
   Detection Prevalence : 0.02105
      Balanced Accuracy : 0.51843

       'Positive' Class : 1
```

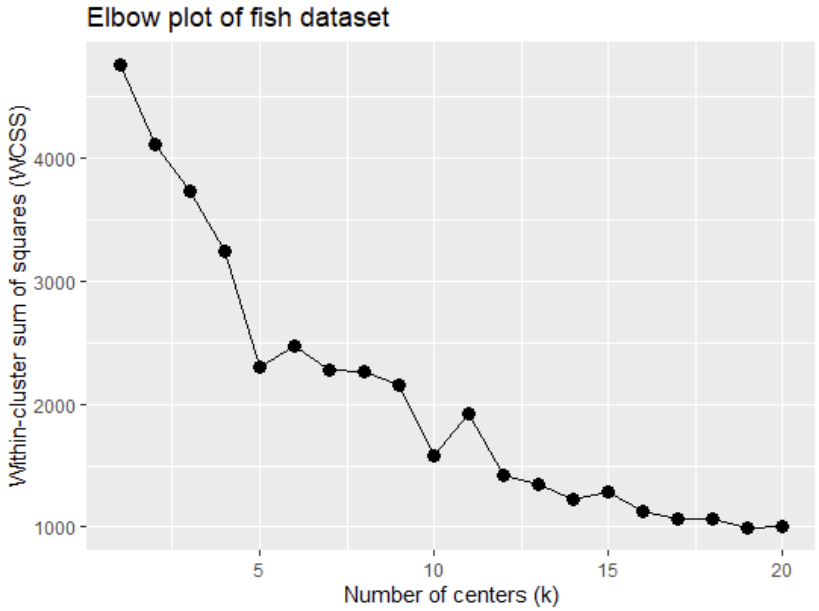
K-means clustering

- Identify clusters of data within the dataset and analyze each cluster

Interpreting the results

- Elbow plot shows us the Within Cluster Sum of Squares (WCSS) metric for each cluster which allows us to identify the marginal benefit of adding another cluster
- The optimal number of clusters is 5
- Cluster 5 had the highest mean fish count at an average of 51,084
  - air_temp and discharge had values close to the average
  - Average set_net value was 0, the lowest, and the drift_net value was 0.32

Elbow plot of K-means Clustering ran with 1-20 centers



Mean statistics of each cluster

cluster	fish_count	precipitation	air_temp	discharge	set_net	drift_net	newmoon	firstquarter	fullmoon	thirdquarter
1	28402.14	0.0000000	56.58737	14086.29	0.34677419	0.5322581	1	0.0000000	0	0.0000000
2	29141.53	0.0106383	57.15426	13818.09	0.05319149	0.2659574	0	1.0000000	0	0.0000000
3	31265.99	0.0000000	56.57792	14470.13	1.00000000	0.9870130	0	0.4025974	0	0.5974026
4	23488.30	0.0000000	57.37395	14768.91	0.32773109	0.5294118	0	0.0000000	1	0.0000000
5	51084.76	0.0000000	56.12097	14459.68	0.00000000	0.3225806	0	0.0000000	0	1.0000000

Random forest regression (RFR), Ridge regression and Lasso regression

Ridge regression and lasso regression lessen the impact of any influential factor and prevent overfitting

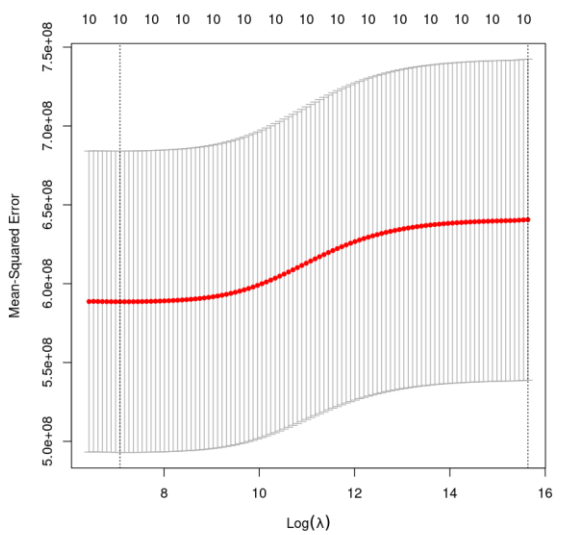
Similarly, random forest regression will give us the average R squared of hundreds of decision trees which should give us a more accurate result.

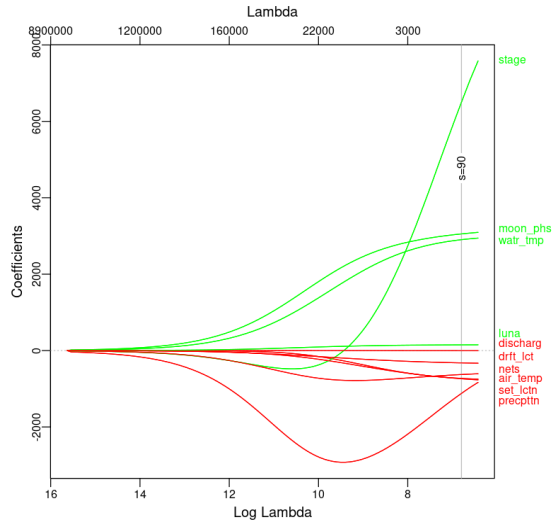
Random forest regression (RFR), Ridge regression and Lasso regression

Interpreting the graphs

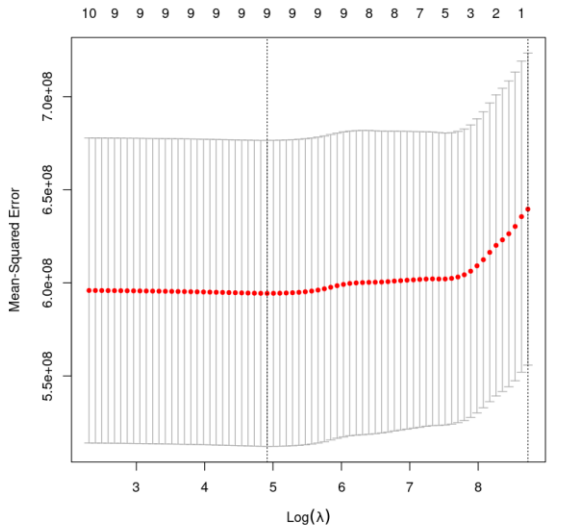
- The left graph shows how well each lambda to MSE – the lowest point in the curve indicates the optimal lambda
- The right graph is a Trace plot which visualizes how coefficient estimates change as we decrease lambda
- Green color indicates an increase in coefficient value and red indicates a decrease

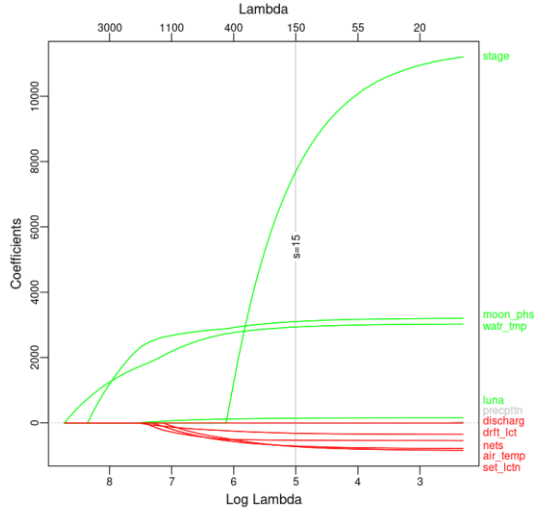
Ridge regression - MSE by lambda value and Trace plot





LASSO regression - MSE by lambda value and Trace plot





Random forest regression (RFR), Ridge regression and Lasso regression

Interpreting the coefficients

- In both ridge and lasso regressions, precipitation is insignificant
- In ridge regression, stage is insignificant
- Both models are aligned with regards to negativity vs positivity of coefficients

Optimal Advanced Regression Coefficients

	Ridge Regression	Lasso Regression
Intercept	-71123.40	-110811.77
Precipitation	~ 0	0
Air_temp	-303.44	-708.45
Stage	~ 0	7679.49
Discharge	-0.95	-3.59
Water_temp	2360.96	2932.61
Lunar	82.37	142.34
Moon_phase	2741.20	3101.13
Set_location	-215.04	-725.54
Drift_location	-162.37	-316.29
Nets	-351.49	-530.68

Discussion and key takeaways

Evaluation of models and challenges

- R-squared values across all models are too low to be used confidently
- Model output are generally consistent across models with a few exceptions

Regression Models R-Squared values

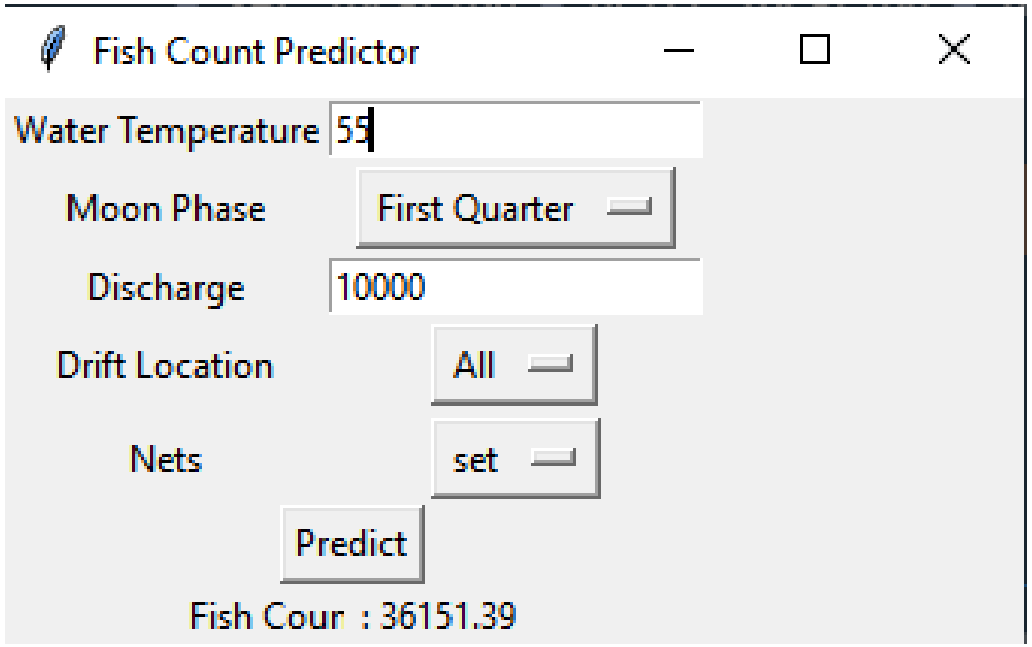
Model	R-squared
Ridge Regression	0.1238
Lasso Regression	0.1272
Random Forest Regression (RFR)	0.3548
Linear Regression	0.3486

Conclusion and future work

- None of the models are sufficiently accurate based on R-squared values - more relevant data is required
- All regression models provided insights into extent of significance of each factor on fish count
- Determined relationships between certain factors and their effect on fish count
- Developed an illustrative working tool that could be used once that model is established and created a solid foundation for future research on this topic

Illustrative fish count predictor GUI

- Users can input the water temperature, moon phase, discharge, drift location, and nets data points.
- When the user hits ”Predict” it then uses our linear regression model to give the predicted fish count



Thank you

Citations

[1] Southwick Associates, Inc., Willian J. Romberg, Allen E. Bingham, Gretchen B. Jennings and Robert A. Clark (2008, December). Economic Impacts and Contributions of Sportfishing in Alaska, 2007. Retrieved April 14, 2023, from Alaska Department of Fish and Game

[2] Horton, C. (2016, December 5). Economic impact of fishing the Kenai Peninsula up for debate. Alaska Journal. Retrieved March 12, 2023, from https://www.alaskajournal.com/community/2008-05-25/economic-impact-fishing-kenai-peninsula-debate

[3] Kramer, C. (2014, September 4). Lunar effects on salmon. Alaska Science Forum. https://www.gi.alaska.edu/alaska-science-forum/lunar-effects-salmon

[4] Alaska Department of Fish and Game (n.d.). *Fish count data search*. Fish Counts - Sport Fish - ADF&G. Retrieved March 12, 2023, from https://www.adfg.alaska.gov/sf/FishCounts/index.cfm?ADFG=main.displayResults

[5] Alaska Climate Research Center. (n.d.). Daily air temperature and precipitation at Kenai AP datasets. Retrieved March 12, 2023, from https://akclimate.org/data/data-portal/

[6] US Department of Commerce, N. O. A. A. (2020, September 23). Historical river observations database. National Weather Service. Retrieved March 12, 2023, from https://www.weather.gov/aprfc/rivobs

[7] USGS water data. USGS water data for the nation. (n.d.). Retrieved March 12, 2023, from https://waterdata.usgs.gov/nwis/

[8] Mateos, L. (n.d.). Moon phases dataset. Moon Phases Dataset | Papers With Code. Retrieved March 12, 2023, from https://paperswithcode.com/dataset/moon-phases