Analyzing Regional Water Temperature and Salmon Population Spawning Dynamics Across Pacific Northwest Sites.

By: Nicole Navarijo, Julie Ramsey, Jack Thomas and Prashant C

Introduction

This project blends environmental science with data engineering, leveraging ETL workflows and database design to provide actionable insights on salmon populations in the Pacific Northwest. Let’s dive in!"

The Problem

The Pacific Northwest is home to various salmon species, and their populations are deeply affected by environmental factors like water temperature. Scientists and environmentalists need data-driven insights to monitor these populations and understand how climate change impacts them.

It's no secret that North Pacific Salmon have been experiencing serious population decline in recent years. Climate change being singled out as the main suspected cause along with pollution, heavy logging, and the building of hydroelectric dams. With salmon being a species that lives in the ocean but travels upstream to spawn, this is a huge problem.

Many sources cite warming water temperatures as one of the main threats that impact yearly spawning trends and it is hypothesized that if the upstream rivers are too warm, many salmon perish on their journey and never make it to their spawning grounds. One such theory points out that retreating glaciers might be leading cause for the lack of cold water in these upstream rivers.

So, we in our group asked: How can we harness data to uncover trends and correlations between water temperature and salmon populations across the Northwestern United States?"

The Dataset and ETL Workflow

To answer this question, we worked with several datasets containing information on water temperature, fish populations, and regional metadata. The raw data procured from our two sources was pretty straightforward with only a few duplicate rows and null values.

The more recent data seemed to be more incomplete however, like it was still being finalized.

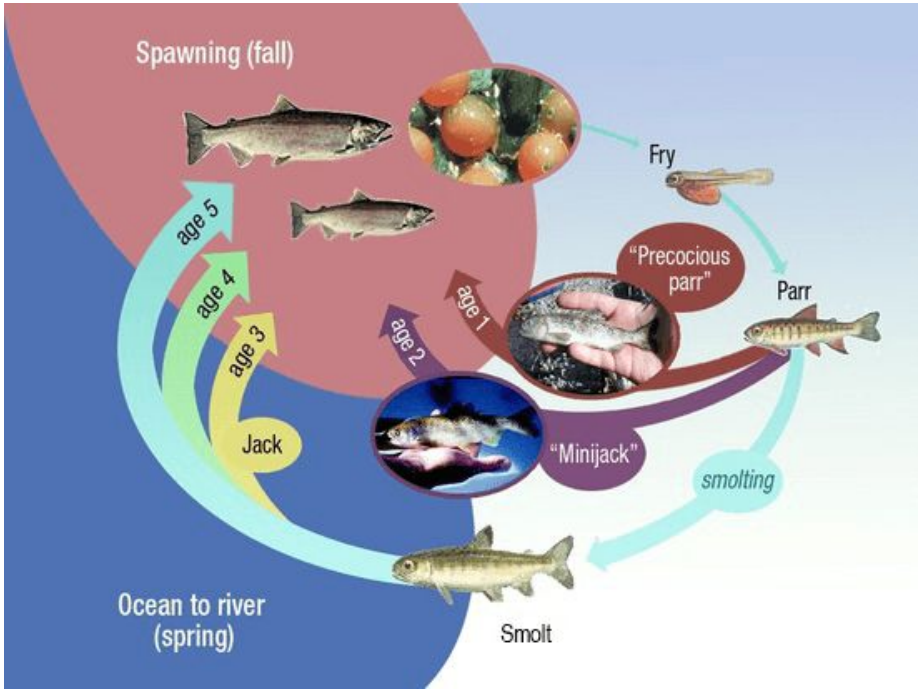
Because of this we needed to constrain our analysis only to years 2013-2019

Data Collection

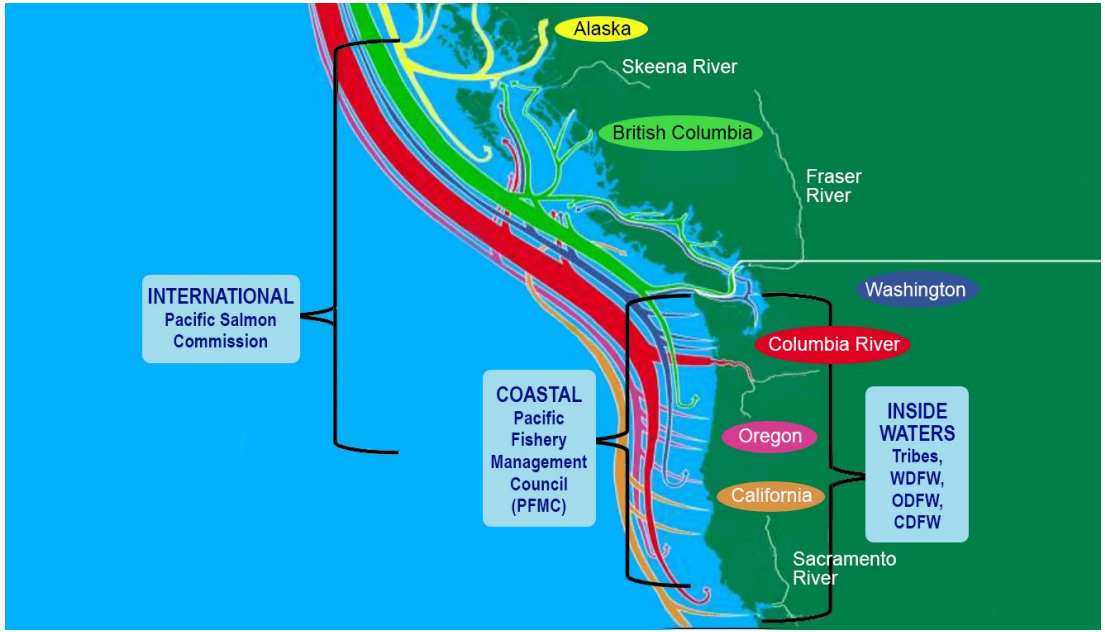
1. *Population Data*: This dataset was sourced from the NOAA and contains historical records of yearly salmon population counts for all species/subspecies, as well parent river, and run seasons. It quantifies salmon population in the form of “Spawners,” meaning the number of mature salmon returning inland to spawn. Spawn counts are made during each species/subspecies run-season which may occur during different times of the year.

2. *Temperature Data*: This data was collected through a series of queries on a USGS public access database. The data collected from each monitoring site comprised of the daily average water temperature, 365 days per year between 2013-2019. The monitoring site locations from which we collected data were selected based on location within each watershed region and their record “completeness” during our desired timeframe. Locations are equally spread throughout each region on the main waterways, allowing us insight into lower, middle, and upstream temperature data.

Development Lifecycle



Migration Routes



Regional Visuals

This Dash application presents an interactive map highlighting the three regions where our data was collected: the Oregon Coast, Willamette/Lower Columbia, and Interior Columbia.

For each region, the map displays the recorded salmon species and their population numbers for the selected year.

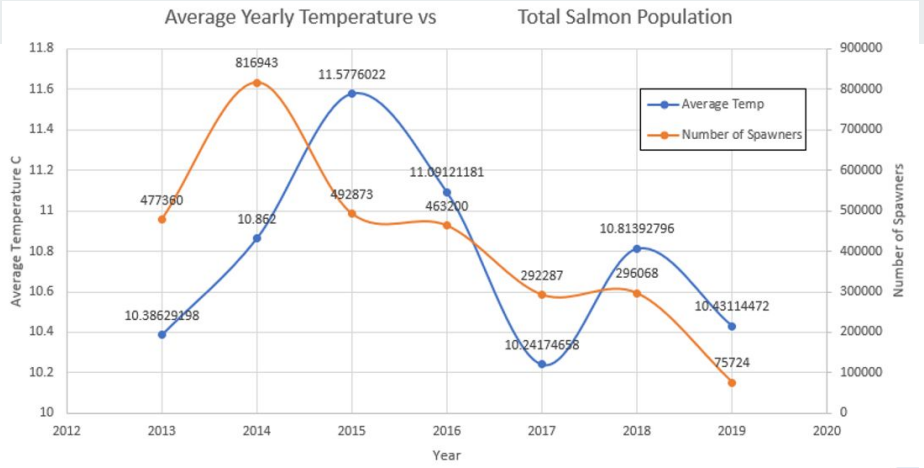
You can use the dropdown menu to choose a specific year and explore the data and hovering over each point will display its population information.

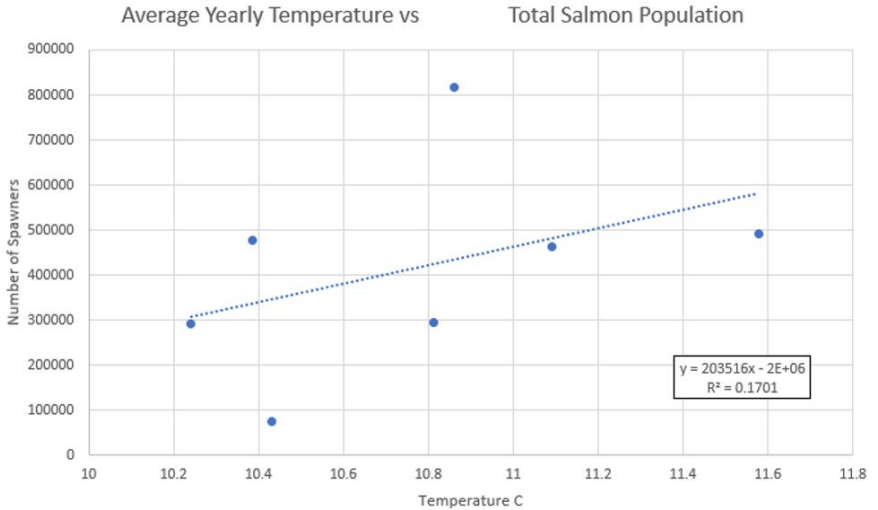
View Interactive Map

Data Exploration

- Temperature data for each year was averaged and plotted against the yearly number of spawners recorded.
- Not the trend we expected to see based on previous research, which was to see temperature and population inversely related.
- Instead we see a slight positive correlation which indicates warmer water is actually better for survival (Unlike the literature suggests).
- Statistical analysis done using a significance correlation coefficient test utilizing a Pearson coefficient.
    - Degrees of freedom = 5
    - Alpha = 0.05
    - Two-tailed

Region	All Regions
Correlation coefficient	0.412487485
T-value calculated	1.012499723
Critical T-value	2.57
Significant?	not significant
Null Hypothesis	accept





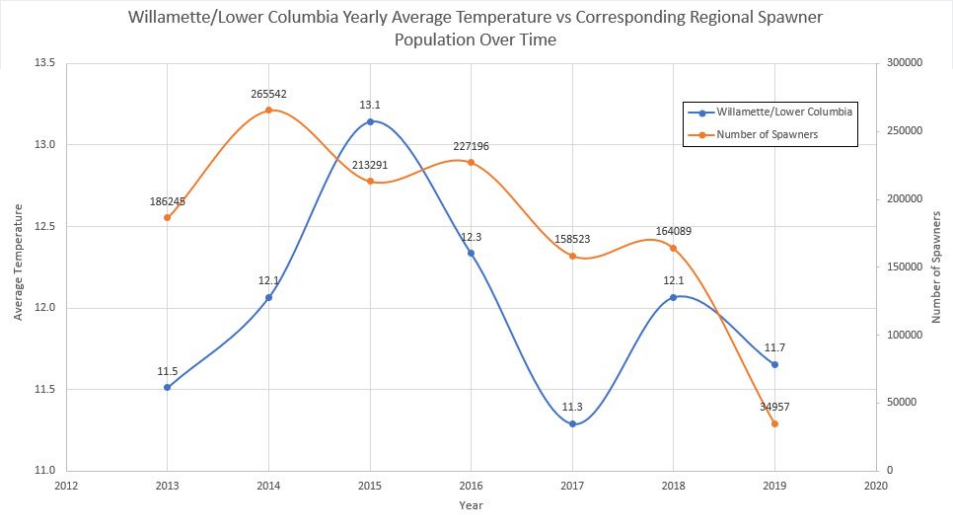
Data Exploration by Region

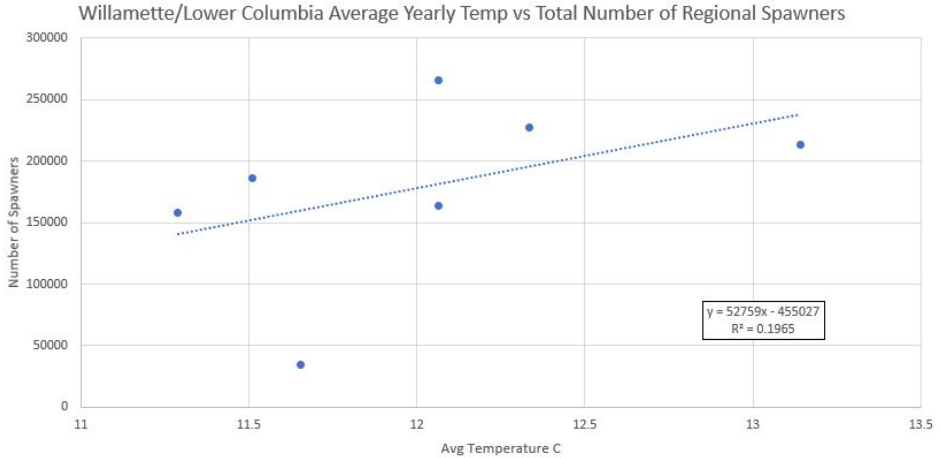
- Due to the previous statistical result, we decided to evaluate each region individually to gain more resolution.
- We were met with the same results for all 3 regions. (A weak positive correlation)

Statistical parameters are the same as the previous test

- Degrees of freedom = 5
- Alpha = 0.05
- Two-tailed

Region	Willamette/Lower Columbia
Correlation coefficient	0.443262075
T-value calculated	1.105726133
Critical T-value	2.57
Significant?	not significant
Null Hypothesis	accept



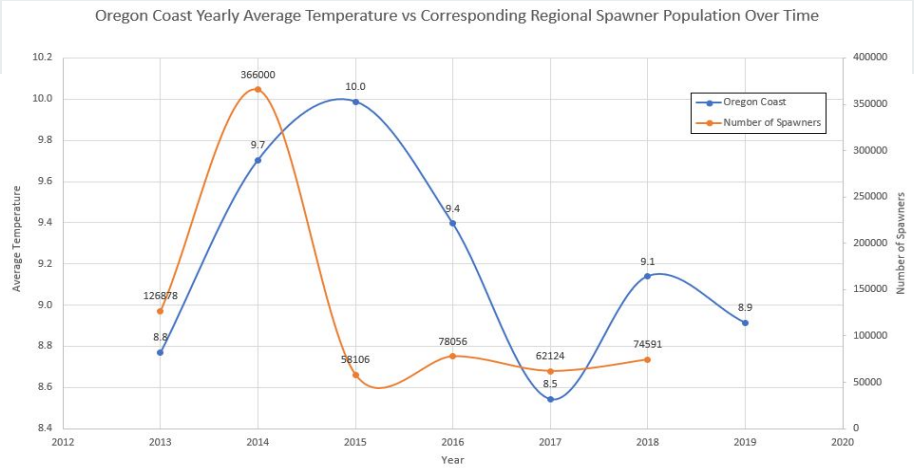


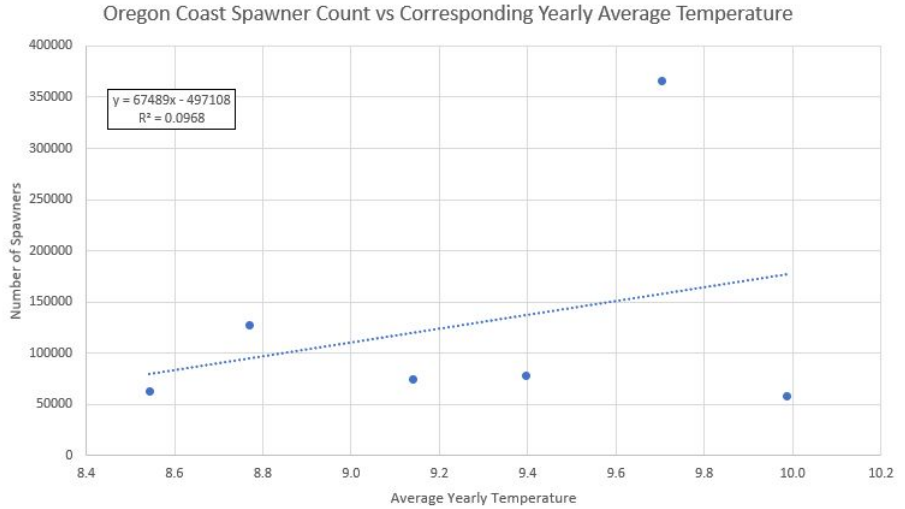
Data Exploration by Region

Statistical parameters are the same as the previous test

- Degrees of freedom = 5
- Alpha = 0.05
- Two-tailed

Region	Oregon Coast
Correlation coefficient	0.311163774
T-value calculated	0.732128937
Critical T-value	2.57
Significant?	not significant
Null Hypothesis	accept



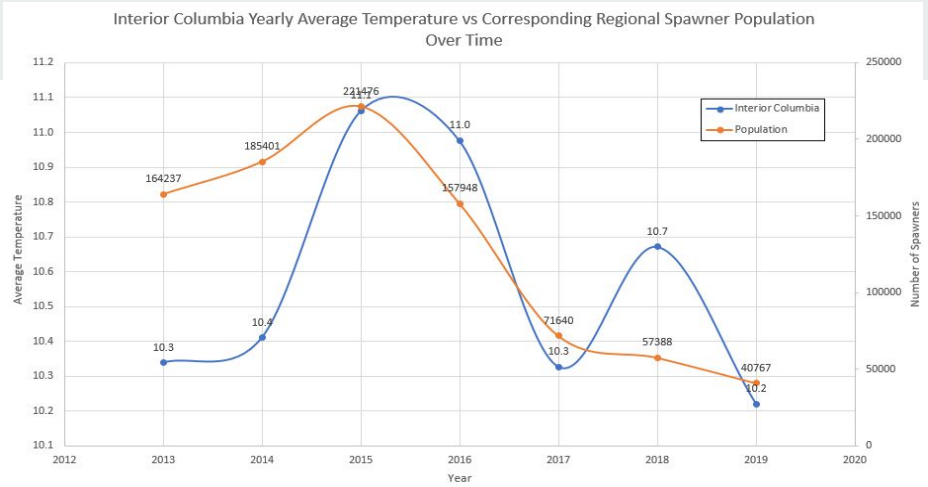


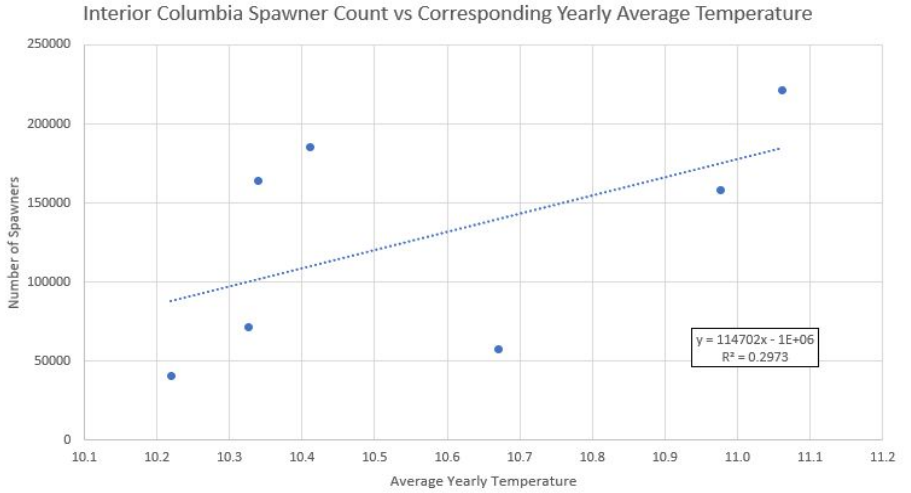
Data Exploration by Region

Statistical parameters are the same as the previous test

- Degrees of freedom = 5
- Alpha = 0.05
- Two-tailed

Region	Interior Columbia
Correlation coefficient	0.545250138
T-value calculated	1.454437631
Critical T-value	2.57
Significant?	not significant
Null Hypothesis	accept





Some Takeaways

- Data seems to indicate a positive correlation rather than a negative one. NOT SIGNIFICANT
- Accept the null hypothesis: “There is no correlation between the overall average yearly water temperature and changes in the salmon population.”
- Different approach might be needed
- Temperature constantly fluctuates, we need to find a way to accommodate for that.

New Strategy:

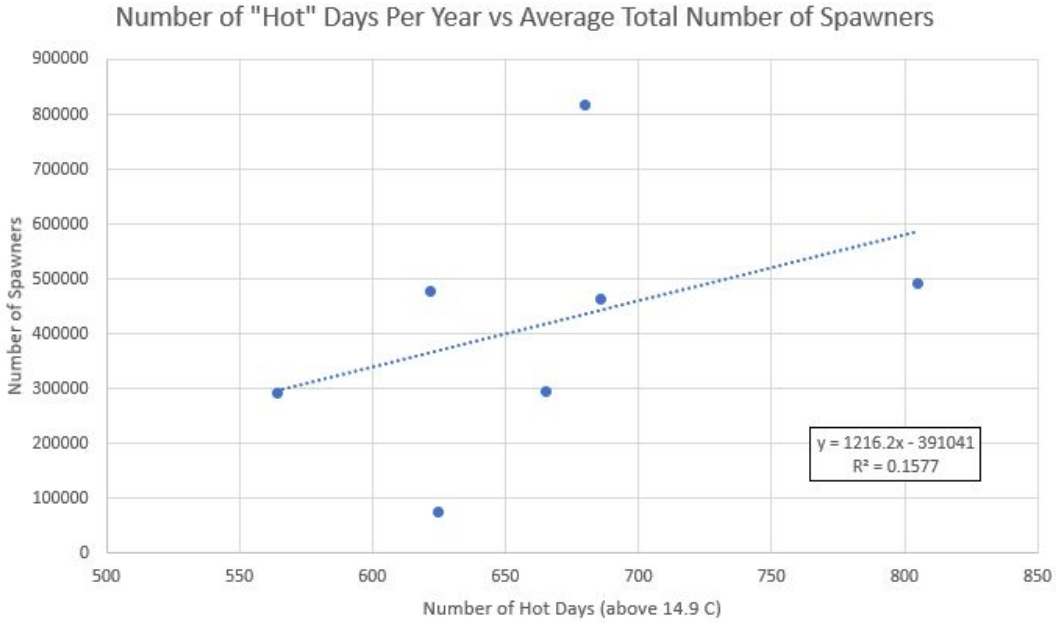
- Salmon thrive when temperature is below 15 degrees Celsius.
- Potential for long/short seasons skewing the data but not affecting the average.
- Plot the number of days that were hotter than 15 C in a given year.

Data Exploration continued :(

- Unfortunately we saw the same results for the overall average and the average of all 3 regions.
- Weak positive correlation

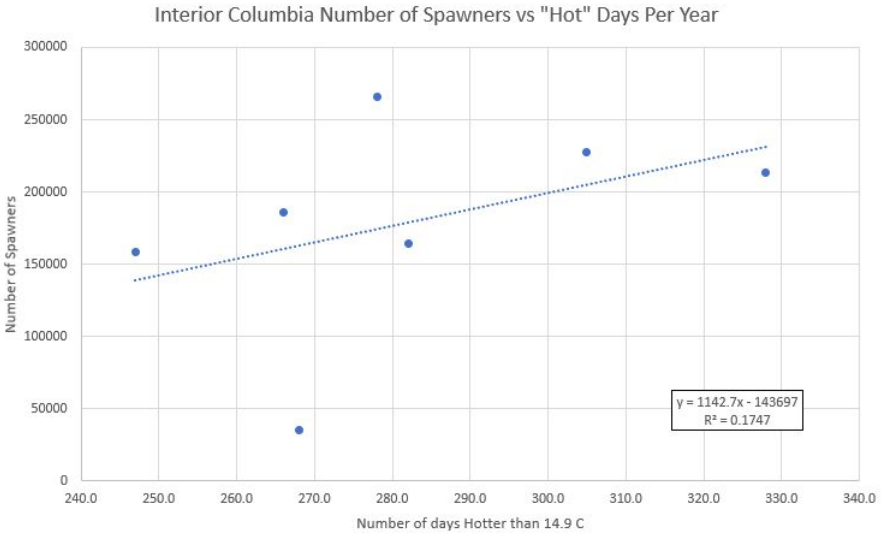
Statistical parameters are the same as the previous test

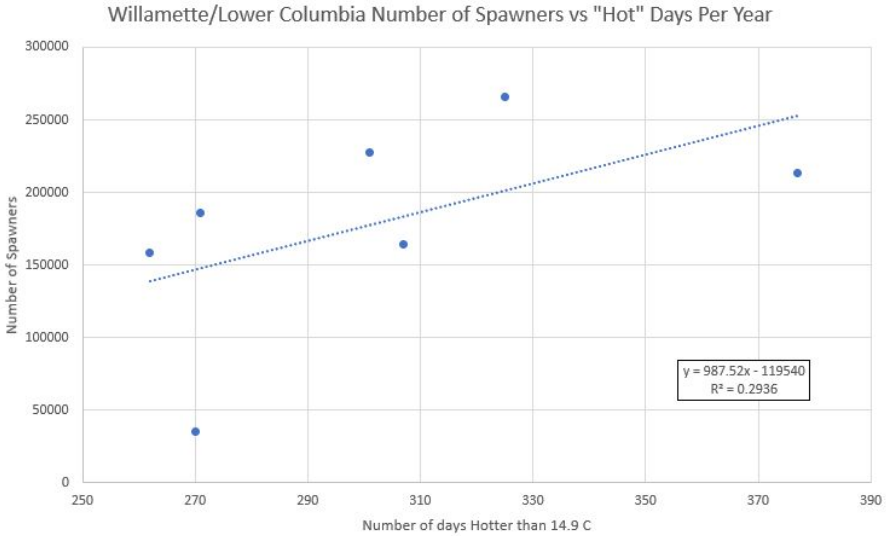
- Degrees of freedom = 5
- Alpha = 0.05
- two-tailed



Region	All Regions(# of hot days)
Correlation coefficient	0.397053281
T-value calculated	0.96735936
Critical T-value	2.57
Significant?	not significant
Null Hypothesis	accept

Data Exploration and Visualization :(





Region	Interior Columbia (# of hot days)
Correlation coefficient	0.417940154
T-value calculated	1.028694291
Critical T-value	2.57
Significant?	not significant
Null Hypothesis	accept

Region	Willamette/Lower Columbia(# of hot days)
Correlation coefficient	0.541885958
T-value calculated	1.441716837
Critical T-value	2.57
Significant?	not significant
Null Hypothesis	accept

Potential Hypothesis:

- Small sample size: n=7
  - Limited by only one population record per year over 7 years

- Seasonal data resolution might be more important than previously thought and it might not be so straightforward
  - ex. if “hot” spring weather leads to early snowmelt, then cold water sources are scarce in the fall which is when most major runs are.

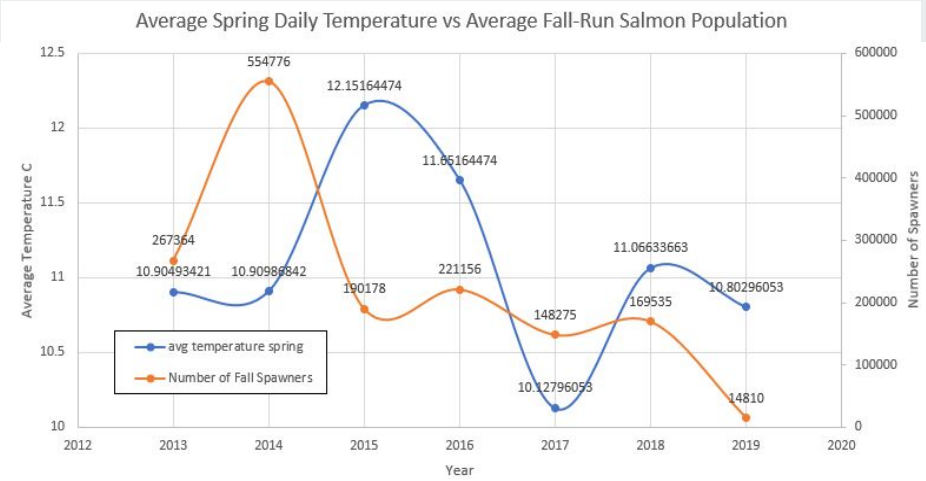
- Delayed effect:
  - Immediately after spawning, salmon die. The effect of suboptimal spawn might not be observed until years later.
  - Offspring spend first 3-4 years of life near spawning grounds.

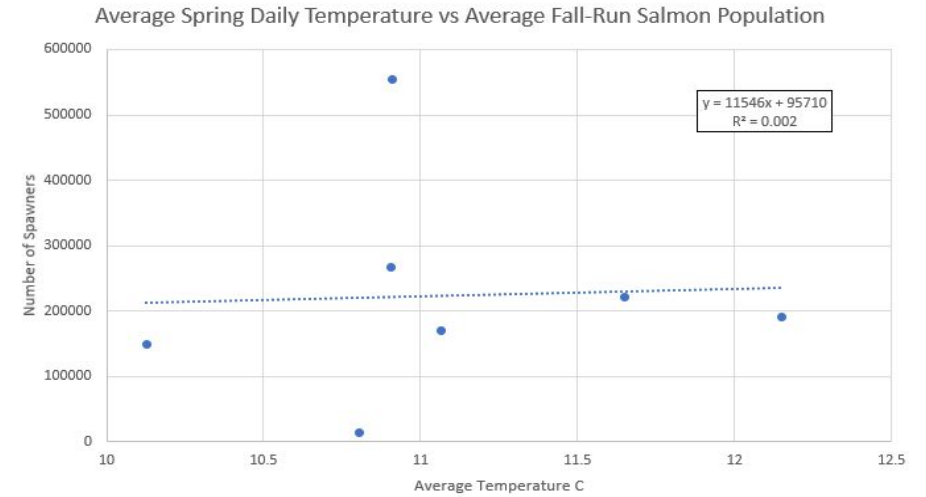
Data Exploration (final effort)

- Average spring temperature plotted against only “Fall-run” species

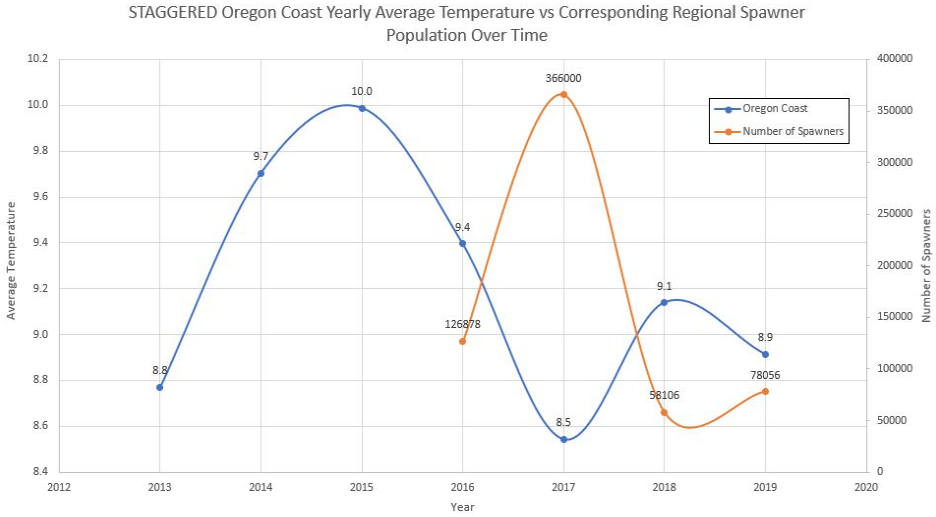
- Initially looked promising but nothing significant (womp womp)

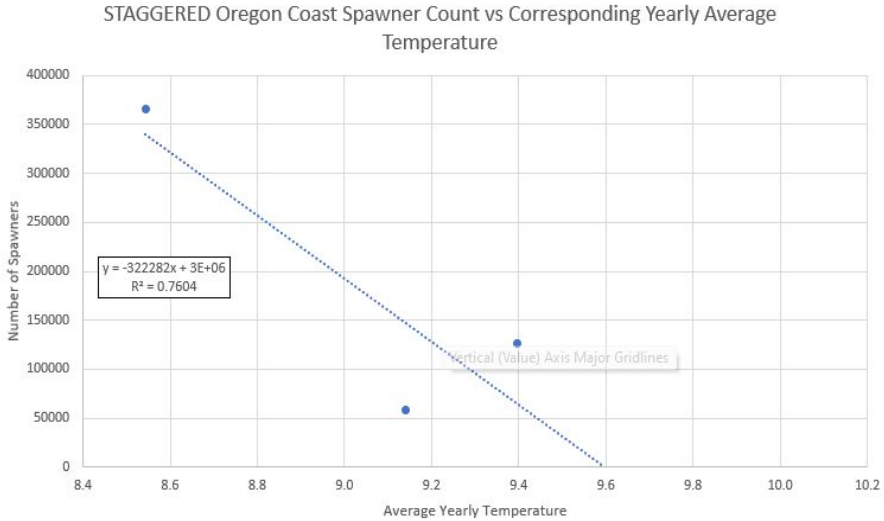
Region	Spring Temp/Fall Run
Correlation coefficient	0.045113188
T-value calculated	0.100978963
Critical T-value	2.57
Significant?	not significant
Null Hypothesis	accept





Data Exploration (delayed effect hypothesis)





- Staggered the data set by 4 years
- Is this something or is this nothing?
- Probably nothing since I just made it up
- No stats data because n=3

Statistics Summary

Average Temperature Data				
Region	Interior Columbia	Oregon Coast	Willamette/Lower Columbia	All Regions
Correlation coefficient	0.545250138	0.311163774	0.443262075	0.412487485
T-value calculated	1.454437631	0.732128937	1.105726133	1.012499723
Critical T-value	2.57	2.57	2.57	2.57
Significant?	not significant	not significant	not significant	not significant
Null Hypothesis	accept	accept	accept	accept

Count of "Hot" days Data				
Region	Interior Columbia (# of hot days)	Oregon Coast(# of hot days)	Willamette/Lower Columbia(# of hot days)	All Regions(# of hot days)
Correlation coefficient	0.417940154	-0.029779831	0.541885958	0.397053281
T-value calculated	1.028694291	-0.066619273	1.441716837	0.96735936
Critical T-value	2.57	2.57	2.57	2.57
Significant?	not significant	not significant	not significant	not significant
Null Hypothesis	accept	accept	accept	accept

Region	Spring Temp/Fall Run
Correlation coefficient	0.045113188
T-value calculated	0.100978963
Critical T-value	2.57
Significant?	not significant
Null Hypothesis	accept

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Final Takeaways

Sample size limitation: Our correlation coefficient test gives us insight about the strength and direction of the linear relationship between x and y. However, the reliability of the linear model also depends on how many observed data points are in the sample.

Overall, we are going to accept our null hypothesis stating "There is no correlation between water temperature and changes in the salmon population."

There are likely lots of factors at play that carry equal if not more influence over the diminishing salmon population.

- Air temperature
- Logging (stirs up silt which kills eggs)
- Water allocation (dams, farming, mining, new construction developments)
- Pollution
- Climate change (drought or fires)
- Disease (parasites)

Future Endeavors

Other metrics: air temperature, dissolved oxygen, flow rate, and gauge height

Our next endeavor would likely try to incorporate some of these additional metrics and how they play a role. Our main dataset still has so many relationships to discover if we wanted to zoom in on specific species, in a specific spot where an unusual event may have occurred.

For example what if:

Hot outside air -> melts snow early -> makes water cold -> low “fall-run” stream height/flow -> low population

Current trend: cold water -> low population (not significant)

Ethics and Considerations

🚨 **Accuracy and Data Integrity** 🚨

- **Consideration**: Ensuring the accuracy of the data is critical because inaccurate information could lead to incorrect conclusions and, consequently, harmful decisions.
- **Ethical Action**: Double-check data sources, apply rigorous data cleaning, and document all transformations to maintain data integrity. Provide transparency in your methodology so that others can replicate or validate your results.

🚨 **Environmental Impact** 🚨

- **Consideration:** The analysis focuses on environmental data, which is highly sensitive and can have far-reaching implications for conservation and sustainability efforts.
- **Ethical Action**: Highlight the importance of using the data to support positive environmental actions, such as protecting vulnerable salmon populations or mitigating climate change effects.

🚨 **Impact on Indigenous and Local Communities** 🚨

- **Consideration:** Many salmon populations are in regions with Indigenous communities who may rely on these fish for cultural, economic, or subsistence purposes. Environmental decisions influenced by your data could have direct effects on these communities..
- **Ethical Action**: Consider the broader social implications of your findings.

Recommendations:

2. Clear Communication.
3. Compliance and Sensitivity

Conclusion

- In conclusion, this project showcases the power of data engineering to tackle real-world problems. By combining ETL workflows, a structured database, and insightful visualizations, we've uncovered valuable insights into the impact of water temperature on salmon populations. Thank you for your time, and I'm happy to answer any questions

Resources

CSV Source:
https://or.water.usgs.gov/cgi-bin/grapher/graph_all_setup.pl?basin_id=umpqua&site_id=14317450#step2
https://www.webapps.nwfsc.noaa.gov/apex/f?p=261:40::::RP,40:G_CURRENT_ARCHIVE:53

Images: https://www.fisheries.noaa.gov/

Python Documentation and Tutorials

Python Libraries and Packages: matplotlib, plotly, dash, pandas, json, shapely, geopandas

Development Tools: Visual Studio Code, QGIS mapping software

Version Control: Git and Github

Collaboration and Communication: Slack and Zoom

Instructional Staff: Travis Hopkins (Instructor), Kian Layson (TA)