Modeling Invasive Carp Passage Risk in the Mississippi River

Caleb Bush

cbush6@bellarmine.edu

2/9/2026

Introduction

The chosen dataset is from 2017-2018 Telemetry data for Asian carp and native fish species at Lock and Dam 19 in the Upper Mississippi River Basin. This was found on U.S Geological Survey (USGS) website. I chose this dataset because I am very interested in the outdoors and wildlife. I also know that invasive species can be very dangerous to a habitat's structure. So, looking into something like this would be very interesting for me. The dataset uses tagged fish's movement for data collection providing numeric and categorical data. This report will describe the variable and their attributes. It will also explore the relationship between those variables alongside their distributions. The report is going to identify missing data and explain why these missing values occur. The use of tables and graphs will show patterns in the data as well as correlation. Lastly the report will discuss some challenges and how they will be approached.

Dataset Description

Attribute Name	Definition	Data Type	Range / Valid Values	% Missing
Date	Date data were collected for lock and fish observations.	Interval	Dates from 3/15/2017 to 9/30/2018	0%
Week	Week of the year corresponding to the observation date.	Ordinal	1–52	0%
Season	Season of observation (SPR, SUM, FAL, WINT).	Nominal	SPR, SUM, FAL, WINT	0%
Temp	Water temperature in degrees Celsius.	Ratio	−2.18 to 29.0 °C	0%
Stage.Ft	River stage height in feet.	Ratio	2.34–20.71 ft	0.03%
Stage.m	River stage height in meters.	Ratio	0.713–6.312 m	0.03%
D.lock.n	Number of downstream lockages per day.	Ratio	0–14	0%
U.lock.n	Number of upstream lockages per day.	Ratio	0–12	0%
Tot.lock.n	Total lockages per day.	Ratio	0–23	0%
Rec.D.n	Downstream recreational lockages per day.	Ratio	0–6	0%
Rec.U.n	Upstream recreational lockages per day.	Ratio	0–5	0%
Rec.Tot.n	Total recreational lockages per day.	Ratio	0–10	0%
Barge.D.n	Downstream commercial barge lockages.	Ratio	0–10	0%
Barge.U.n	Upstream commercial barge lockages.	Ratio	0–11	0%
Barge.Tot.n	Total commercial barge lockages.	Ratio	0–16	0%
TRANSMITTERID	Unique transmitter ID for tagged fish.	Nominal	Alphanumeric IDs	3.91%

DeployDate	Date fish transmitter was deployed.	Interval	Tagging dates 2015–2016	3.91%
Species	Fish species tagged.	Nominal	LKSG, BHCP, GSCP, BUSK, SVCP, etc.	3.91%
Length	Fish length (mm).	Ratio	422–1410 mm	12.39%
Weight	Fish weight (grams).	Ratio	290–23620 g	4.50%
AGENCY	Agency responsible for tagging.	Nominal	MDC, FWS, WIU, etc.	3.91%
Deploy.loc	Location fish was tagged.	Nominal	Pools 16–24, MMR, tributaries	3.91%
Up.passage	Indicator of upstream passage during study.	Nominal	0 = no, 1 = yes	3.91%
Down.passage	Indicator of downstream passage.	Nominal	0 = no, 1 = yes	3.91%
UpPass.2017	Upstream passage occurred in 2017.	Nominal	0 or 1	0%
UpPass.2018	Upstream passage occurred in 2018.	Nominal	0 or 1	0%
Up2017.date	Date of upstream passage in 2017.	Interval	Apr–Aug 2017	94.77%
Up2018.date	Date of upstream passage in 2018.	Interval	Mar–Sep 2018	94.25%
START_DATETIME	Start time of fish residency event.	Interval	Date-time values	3.91%
END_DATETIME	End time of residency event.	Interval	Date-time values	3.91%
RESIDENCEEVENT	Unique ID for each residency event.	Nominal	1–212	3.91%
DURATION.sec	Duration of residency event in seconds.	Ratio	60–902100 sec	3.91%
DURATION.min	Duration in minutes.	Ratio	1–15035 min	3.91%
log.DUR.min	Log-transformed duration (minutes).	Interval	0–4.177	3.91%
NUMRECS	Number of detections during event.	Ratio	2–54722	3.91%

Table 1. Description of every variable showing their type, range of values, and percent of data missing.

Dataset Summary Statistics

Variable	Mean	Median	SD	Min	Max
Week	27.42	29	16.01	1	52
Temp (°C)	12.54	10.13	11.09	-2.18	29
Stage.Ft	6.95	5.11	3.98	2.34	20.71
Stage.m	2.12	1.56	1.21	0.71	6.31
D.lock.n	3.42	3	2.92	0	14
U.lock.n	2.98	3	2.60	0	12
Tot.lock.n	6.40	7	4.95	0	23
Rec.D.n	0.51	0	0.97	0	6
Rec.U.n	0.34	0	0.83	0	5
Rec.Tot.n	0.85	0	1.65	0	10
Barge.D.n	2.91	3	2.52	0	10
Barge.U.n	2.64	3	2.29	0	11
Barge.Tot.n	5.55	6	4.21	0	16
Length (mm)	834.04	760	190.92	422	1410
Weight (g)	6794.60	5190	3729.62	290	23620
Up.passage	0.08	0	0.26	0	1
Down.passage	0.11	0	0.32	0	1
UpPass.2017	0.05	0	0.22	0	1
UpPass.2018	0.06	0	0.23	0	1

RESIDENCEEVENT	32.95	16	40.20	1	212
DURATION.sec	9169.82	2820	30989.10	60	902100
DURATION.min	152.83	47	516.49	1	15035
log.DUR.min	1.63	1.67	0.73	0	4.18
NUMRECS	169.80	9	1706.63	2	54722

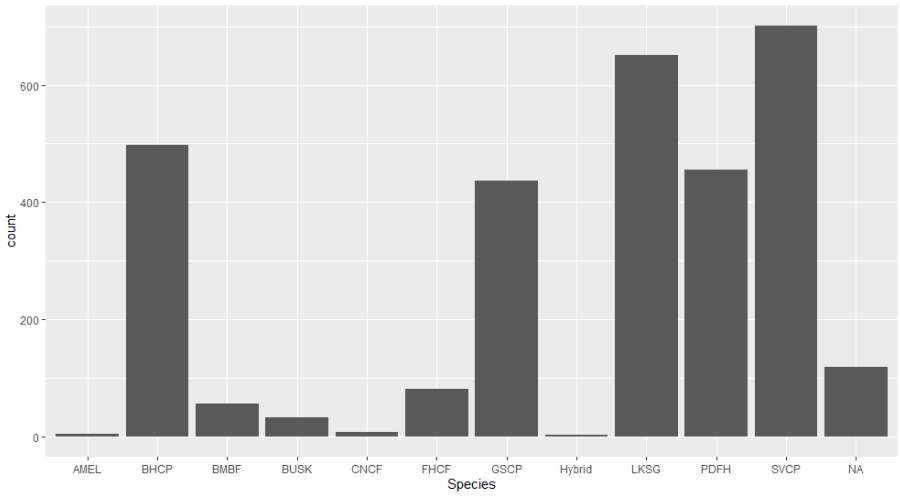
Table 2. Summary Statistics for numeric and continuous variables. Measures include mean, median, standard deviation, minimum, and maximum values.

Species Code	Species Name	Count (n)
SVCP	Silver Carp	701
LKSG	Lake Sturgeon	651
BHCP	Bighead Carp	497
PDFH	Paddlefish	455
GSCP	Grass Carp	437
FHCF	Flathead Catfish	81
BMBF	Bigmouth Buffalo	56
BUSK	Blue Sucker	33
CNCF	Channel Catfish	7
AMEL	American Eel	4
Hybrid	Hybrid Carp	2
Missing	Unknown species	119

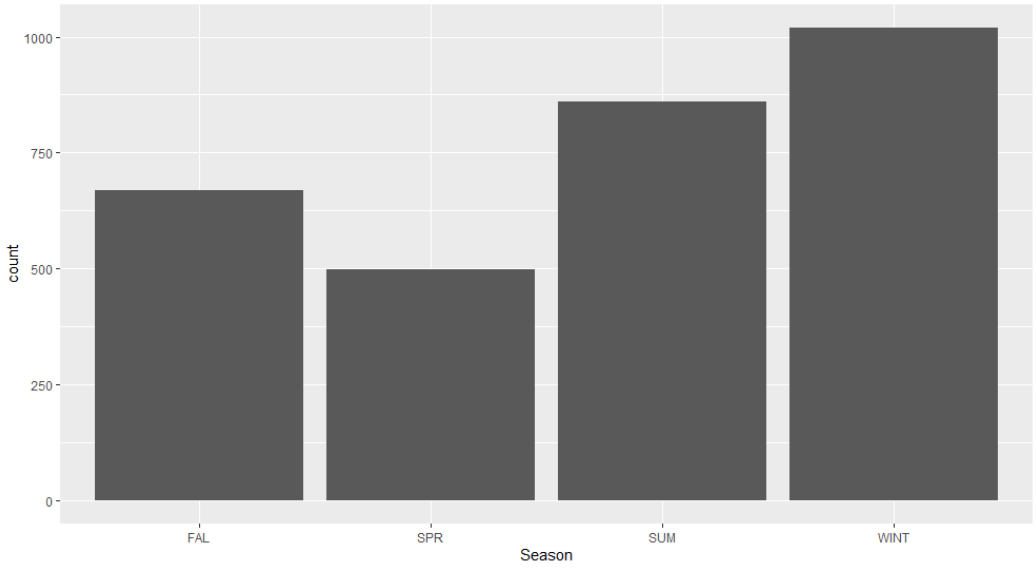
Table 3. The invasive carp are the focus of the study, so they have the most coverage. The 119 missing values are something that will need work.

The data anomalies in this set would be that the variables Up2017.date and Up2018.date have approximately 95% missing data. Other notable variables are length with 12% missing and several other fish identification variables being around 4% missing. The main outlier is in duration with the maximum being 15,035 minutes. This in turn is skewing the duration variables as well.

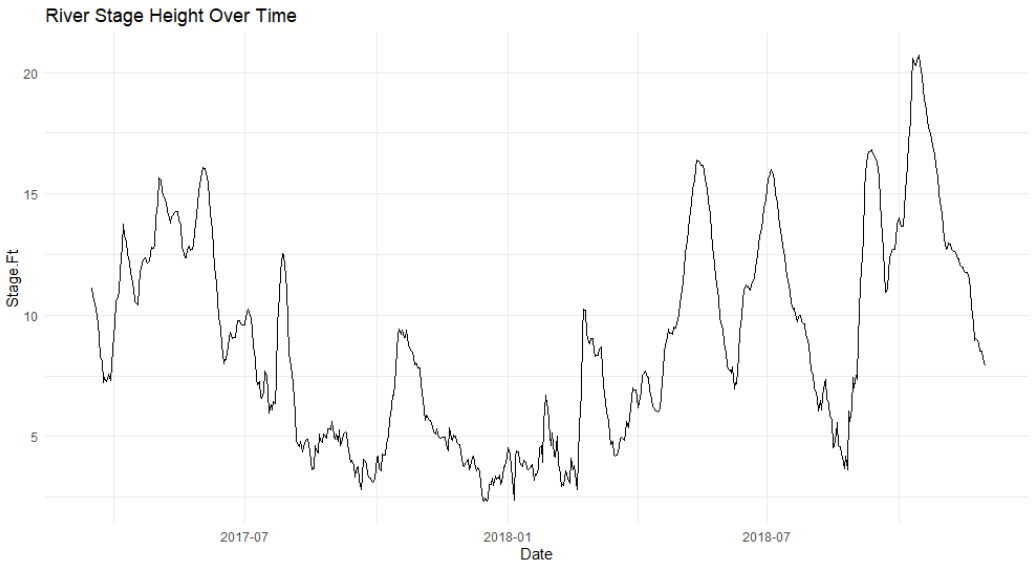
Dataset Graphical Exploration



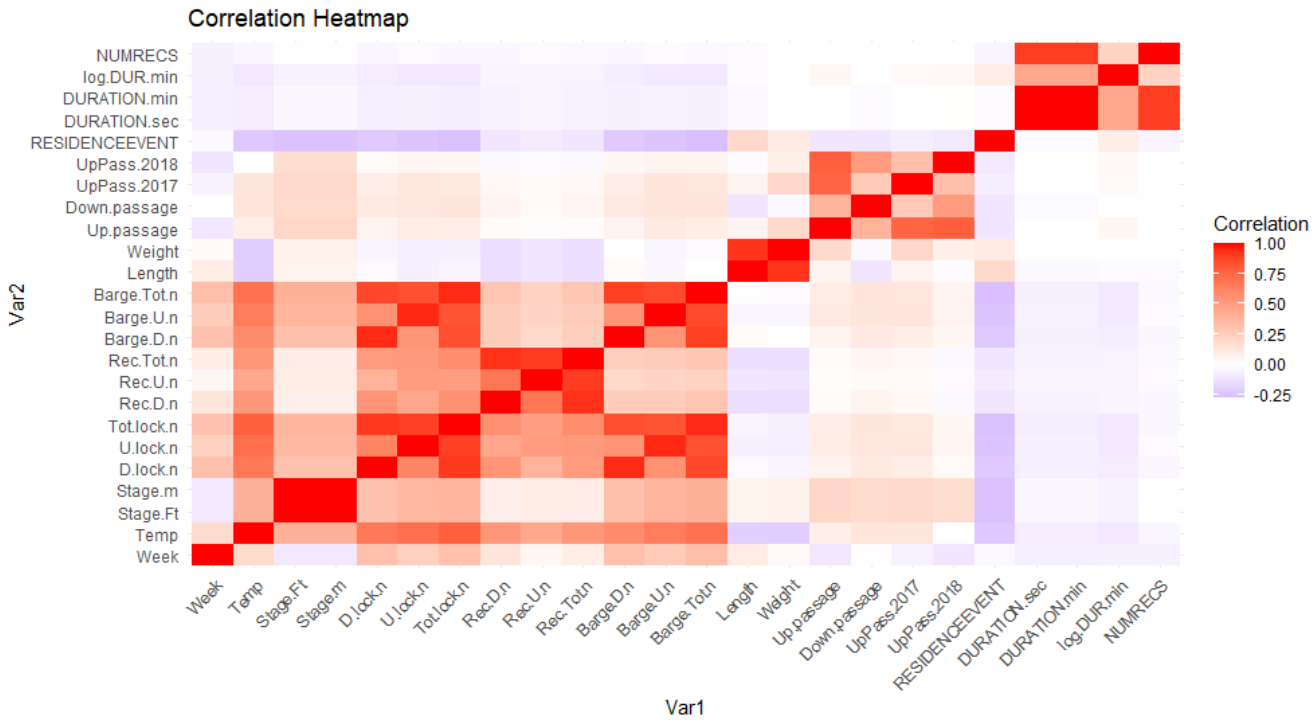
This bar graph shows the distribution of fish species from the data. With silver carp, lake sturgeon, bighead carp, grass carp, and paddlefish being the most frequently recorded.



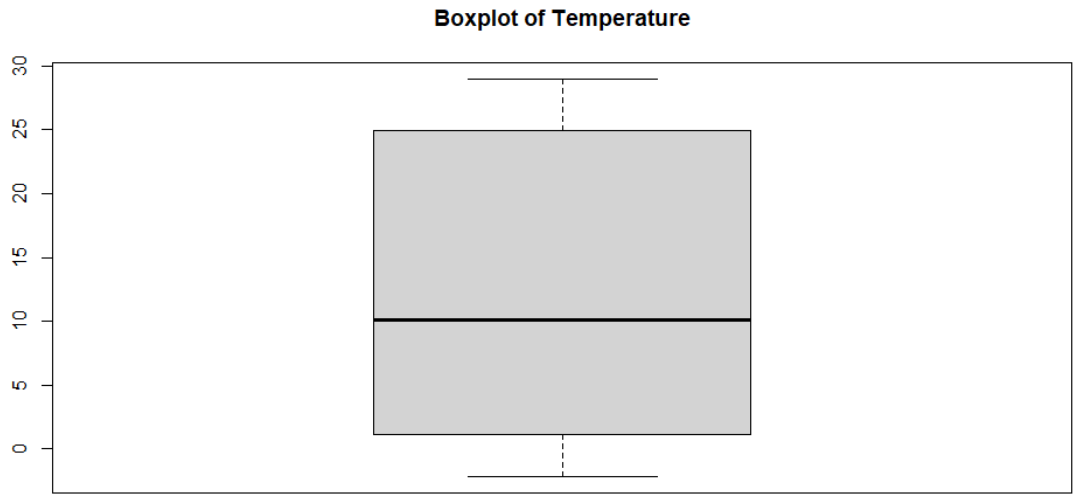
This bar graph shows the distribution of the seasons. You can see winter was the season with the highest activity.



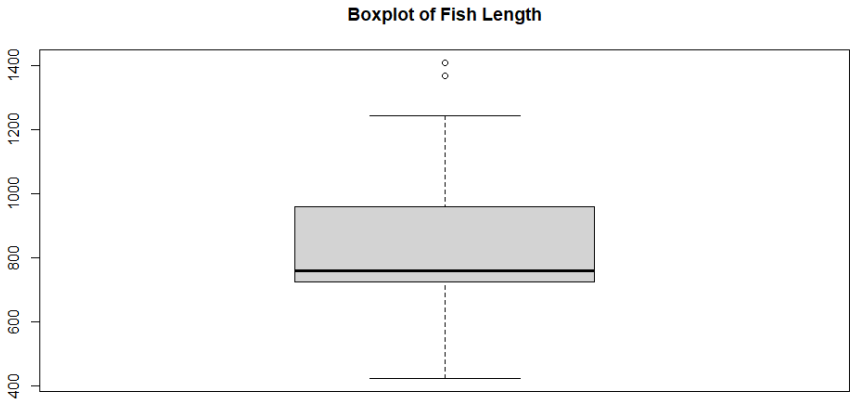
This graph shows this distribution of the height of the river stage over this time which can be useful for comparison to fish movement.



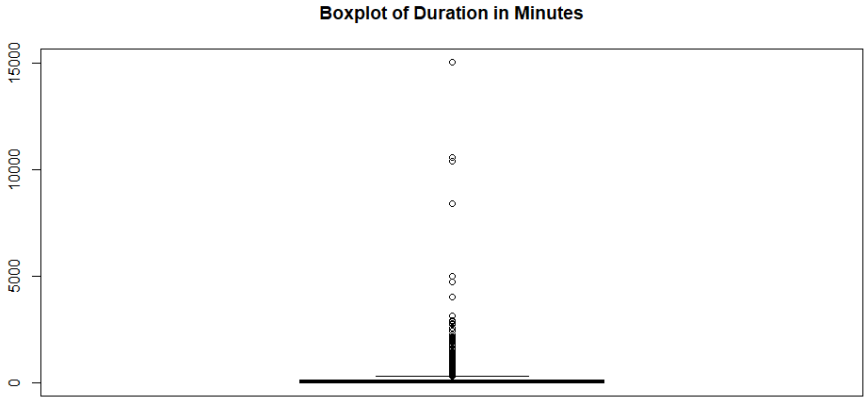
This correlation heatmap shows the relationship between the variables. The passage indicators have mass missing data so their relationship to other variables at this does not mean a whole lot in this figure.



This boxplot shows the distribution of water temperature and how it's not very skewed.



This boxplot of fish length shows how there are 2 major outliers skewing this data.



In this boxplot of duration, you can see how many outliers there are that will need to be taken care of as they are heavily skewing the data.

Summary of Findings

The main findings from doing this analysis are that water stage height and water temperature seem to be key variables for prediction. Also, that missing data is a problem along with very large outliers. More specifically 2017 and 2018 upstream passage. The potential problem causing this is that there was only a recording of when passage occurred so converting the missing values to a 0 for no passage might work. Another notable variable is length, with 12% of its data missing. Imputation would work best for this since 12% is far too large to drop. Using the median for the imputation would be ideal to help with large and small values skewing the data. As seen from one

of the boxplots above duration has many outliers. Removing these will make this variable more suitable for use than it is now since it is so skewed. There are also many variables with around 4% of their data missing. Since many variables have this percentage of data missing checking to see if they are all missing data from the same row could shed light on whether to drop the data from those columns or do some sort of imputation. There is also some imbalance in species passage as some fish are rarer than others so species count may not be true to other parts of the river. Some columns are redundant in their measuring. For example, there are two columns measuring the stage height but in different units. Removing one of these would be good since they are doing the same thing. Overall, the dataset has good data that just needs some preprocessing before any modeling is done.