2018 Guam and CNMI Small Boat Fisheries Cost-Earnings Survey

Title: Data cleaning documentation

Author: Crystal Dombrow

Description: This document records the data cleaning steps for individual survey responses in the 2018 Guam and CNMI small boat fisheries cost-earning survey data. Each step was manually entered in Excel, with the resulting files located in the R project folder path *GuamCNMI_SBF > Data*. Additional data cleaning steps are described in "GuamCNMI_SBF_DataCleaning.R" in the *Data* folder.

Marianas2018_SurveyData_checkedhardcopies.xlsx:

- Chose to do data cleaning in a separate file, in case I need to reference the original responses from surveys
- Corrects survey raw data against hard copies, corrections in yellow text
- Removed VarName column
- Renamed variables according to question number. Previous variable names are included next to current names in Marianas2018_Codebook_updatedvariables.pdf
- Recoded stated ranges to midpoint value
- Recoded undefined range to stated number
- Created second survey response line for respondent #1190 who reported two sets of responses for vessel characteristics questions. Made remainder of responses NA for second boat. These are now 1190A and 1190B.

Marianas2018_SurveyData_checkedskippattern.xlsx:

- Incorporated data cleaning on trip costs and annual expenditures from Michel, documented in "Documentation for 2017 GuamCNMI Calculation.docx"
- Kept mixed bottomfish trips coded as bottomfish instead of mixed gear (Q34A, Q37A)
- Recoded all accidental "0"s in categorical variables to codebook's equivalent of 0/0% (usually a "1")
- Replaced "-1" and "-2" with "NA"; new version of RStudio returning errors with replacing -1/-2/99/77 with NA in the script across various dplyr functions
- Unless otherwise designated, Q_ refers to all variables within that question (ie Q5 = Q5A, Q5B)
- Changed "RespID" variable name to "Survey"
- Determined and removed outliers:
    - Manually recoded outliers to NA (Q25, Q26, Q30:Q33, Q35B, Q35C, Q36A, Q36C, Q36E:Q36I, Q38B, Q38C, Q39A, Q39C, Q39E:Q39I, Q40A:Q40K).
    - Author calculated outliers in R program (marked with green cells in Marianas2018_SurveyData_checkedskippattern_OUTLIERSNOTES.xlsx) then checked them against the remainder of survey responses to decide which are true outliers (marked with red font inside green cell in Marianas2018_SurveyData_checkedskippattern_OUTLIERSNOTES.xlsx)
    - Manually removed true outliers.
- Checked skip pattern for each survey:
    - Boat trips: 1, 2, 5, 6, 24, 34, 37, 46
    - Boat-based gear type: 2, 8, 21, 34, 37, 56, 57
    - Give away their catch: 16B-E, 57
    - Own boat: 13, 24, 25-32, 33, 40
    - Paid for fishing costs: 35/36, 38/39 (often didn't add up, but close)
    - Save catch as food for family: 16B-C, 23, 57 (All who skipped Q16 but gave species for Q57 filled in Q23)
    - Shore-based gear type: 3, 4, 37
    - Sold fish: 11 primary motivation, 15, 16A, 17-22, 56
    - Vendor/Independent fisher: 12, 13, 22, 24, 28 (didn't use this to check the data, but could be used to fill in blanks)

    - Catch volume: Q8, Q2, Q18, Q21, Q56-57. Checked separately for pelagics, bottomfish, and reef fish
- What should we consider invalid responses? (*including these notes for future surveys*)
    - See if there's a pattern, which portions? How much? What is filled out – any economic data? Everything except economic data? Goal of the survey is economic information.
    - Did the response provide enough information for the survey effort and the report for it to count?
    - If we're dropping something, it should be obvious.
    - We're not academic, so partial responses are okay. Err on the side of keeping them, we should be able to defend it because it contributed to the project.
- Added variable:
    - sell.fish: dummy variable that manually determines who sold fish by checking skip pattern as outlined above. (1 = yes, 0 = no)
- Responses the authors considered dropping but kept:
    - 1218: skipped trip costs
    - 1235: no mention of boat based gears but marked less than 12 trips, skipped trip costs, Q55-57, didn't sell fish, doesn't own boat – *might be along for the ride, keep it. there's some added value to what they did give us.*
    - 2005: doesn't own boat, skipped Q41-49
    - T-9: doesn't own boat, skipped trip costs, Q56-57

Q1A:
- Recoded given range to midpoint value
- Recoded "few" to 6 (rounded midpoint of range for Q1B "1")

Q1B:
- Included range (Q1B) from stated value (Q1A)
- Recoded given range to match stated value (Q1A)
- Recoded to NA when Q1A "0" response and Q2 NA responses

Q2:
- Recoded values so that survey responses = 100%
- Recoded responses to 1 (0%) for surveys reporting 0 trips in Q1B
- Recoded full sets of 1's to "NA"
- Q2H: Recoded all rod and reel responses to "Rod & reel"
- Q2I: Recoded full sets of -1's with 0 trips to -1 gears

Q3A:
- Recoded "few" to NA; respondent answered Q3B

Q3B:
- Included range (Q3B) from stated value (Q3A)
- Recoded given range to match stated value (Q3A)
- Recoded to NA when Q3A "0" response and Q4 NA responses

Q4:
- Recoded values so that survey responses = 100%
- Recoded responses to 1 (0%) for surveys reporting 0 trips in Q1B
- Recoded full sets of 1's to "NA"
- Q4K: Recoded full sets of -1's with 0 trips to -1 gears
- Recoded survey #1218 to match the shore-based gear listed in Q37

Q5:
- Recoded values so that survey responses = 100%

- Rounded up for single day trips so that survey responses = 100%; we would expect only few multi-day trips.
    - For Q5 responses of all 1's or NA's, recoded Q5A as "6"

Q7:
- Recoded "South" and "depends on marine weather" as NA

Q8:
- Recoded given range to match stated value

Q10H:
- Recoded NA's to "0" if respondent completed remainder of Q10

Q12:
- Changed misplaced NA's to 0 ("no") for partially complete answers across Q14

Q13:
- Recoded values so that survey responses = 100%
- Recoded responses of all 1's to NA
- Adjusted some responses to match remainder of survey

Q14:
- Recoded #1011, 1111, 3033 to all NA's; no response to this question
- Recoded NA's to "no" for partial responses
- Recoded NA to "yes" for Q14E responses that provided information for Q14F

Q16:
- Recoded values so that survey responses = 100%
- Recoded NA's to 1's (0% of trips) for partial responses
- Recoded responses of all 1's to NA

Q17:
- Recoded values so that survey responses = 100%
- If Q16A = 1, recoded Q17 to NA
- If Q16A = >1, recoded Q17 NA's to 1 (0%)
- Recoded responses of all 1's to NA
- Recoded "Not for Sale" to NA

Q20:
- Recoded "none", "(note along with question 20) Captain keeps most of the sales. 50% +", "Don't sell", and "0" to NA
- Recoded undefined range to stated number
- Recoded Q21A to match response in Q20B

Q21:
- If respondents sold fish, recoded responses of all 0%'s to NA's
- Recoded values so that survey responses = 100%

Q22:
- Recoded "2 – independent fisher" to 2

Q24:
- Recoded to match remainder of survey (ie, Q13 captains own their boat, Q13 crew does not own, checked against Q28 as well)

Q25:

- Recoded "11' 4" " to "11"

Q26:
- Recoded "70 twin" to "140"
- Recoded to NA: "kayak w/ motor" (#1236), "twin cumins"

Q27:
- Recoded to NA: "??", "Don't know viking marine", "Don't know/old", "IDK", "Tiderunner-fiberglass", "uses 2 boats. Don't know"

Q28:
- Recoded "no" to NA

Q30:
- Recoded to NA: "family boat"
- Recoded "gift" to "0"

Q31:
- Recoded "phone GPS" to NA

Q32:
- Recoded "10,000 (came with boat)" to "10,000"

Q33:
- Recoded "Don't own" to NA
- Recoded "Several thousand" to average value of boat owners who provided responses – 24,748

Q35:
- Recoded Q35A to match best response for Q35B-E (note: some respondents gave information for Q35B, C, and D)
- Recoded Q35E "n/a" to "NA"
- Recoded Q35E "none" to "0"

Q36:
- For partial responses, recoded skipped items to "0"
- Recoded full set of 0's to "NA"

Q37:
- Recoded "None" to NA

Q38:
- Recoded Q38A to match best response for Q38B-E (note: some respondents gave information for Q38B, C, and D)
- Recoded to match Q39 for "0" responses
- Recoded "none" to 0
- Recoded to NA: "together we went trolling", "N/A", "NA"

Q39:
- For partial responses, recoded skipped items to "0"
- Recoded full set of 0's to "NA"
- Recoded Q39H response of "owner pays" to "0"
- Q39F: Recoded "10 bags" to 40

Q40:
- For partial responses, recoded skipped items to "0"
- Recoded full set of 0's to "NA"

- Recoded Q40A response of "renter's insurance" to "0"

Q44:
- Made village names uniform in capitalization and village name

Q45:
- Recoded "All my life" to midpoint of age selected in Q43

Q46:
- Recoded "0", "undetermined", "Many moons", "most" to NA

Q48:
- Recoded Q48A to match Q48B

Q49:
- Recoded open-ended responses to categorical options, where listed

Q51:
- Recoded to NA: "24/7", "0" for respondent who marked full-time employment in Q50A
- Recoded "144" to "40"; each respondent selected that they work full time in Q50A

Q56:
- Recoded to NA: "I don't sell, I share and/or donate", none, Katsuo (kept all), YFT (no sell), Wahoo (no sell), Mahi (no sell), Wahoo (kept all), Yellowfin Tuna (no sell), not selling as of now
- Recoded "trolling" and "Spear fish" to "Pelagic" based on remainder of survey responses
- Recoded "Yellow" to "Yellowfin tuna" based on remainder of survey response
- Removed parentheses and special characters, recoded to most specific species given

Q57:
- Recoded "none", "0", "no preference" to NA

comment:
- Recoded "none" to NA

email:
- Recoded "none" to NA

Island:
- #1210 and #2005 didn't include village names and had no island name attached to them. Recoded 1210 to Guam because survey numbers adjacent to it on both sides are all Guam. Recoded 2005 to Saipan because survey numbers adjacent were from Saipan.