Answers 6.1

1. Data Source

Data requirements

The data set must:

- Be open source
- Come from an authentic/authoritative source
- Include non-anonymized column names
- Be no more than 3 years old (up to a maximum of 10 years if you've found a perfect data set for your needs and no newer data is available)
- Contain at least 2 continuous variables (excluding index or ID variables, dates, years, etc.)
- Contain at least 2 categorical variables (excluding index or ID variables, dates, years, etc.)
- Contain at least 1,500 rows
- Include a geographical component with at least 2 different values (e.g., countries, continents, U.S. states, cities, latitude and longitude values—anything you can visualize on a map!)

Summary of data source

Open data set from 'kaggle': Snow Crab Geospatial Data (1975-2018) – Link to 'kaggle'

- The data was collected from NOAA (National Oceanic and Atmospheric Administration, U.S. Department of Commerce). It contains catch per unit effort data of commercial snow crab landings in the Alaskan Eastern Bering Sea. The catch per unit effort is an indirect measure of the abundance of a target species.

Can be considered internal data from an official United States government website, therefore owned by the government of the U.S., this data can be assigned as trustworthy. Although, technically, it's external data provided by a third-party organization.

Source of variables

id: Internal, administrative
latitude: Internal, administrative
longitude: Internal, administrative
year: Internal, administrative
name: Internal, administrative
sex: Internal, administrative
bottom depth: Internal, administrative
surface_temperature: Internal, administrative
bottom_temperature: Internal, administrative
haul: Internal, administrative
cpue: Internal, administrative

Content

id: Primary key id
latitude: The latitude (in decimal degrees) at the start of the haul.
longitude: The longitude (in decimal degrees) at the start of the haul.
year: Year the specimen was collected.
name: The common name of the marine organism associated with the scientific name.
sex: Gender of crab.
bottom depth: Meters (in m). Weighted average depth (in m) and is calculated by adding gear depth to net height.
surface temperature: Surface temperature (in tenths of a degree Celsius).
bottom temperature: Average temperature (in tenths of a degree Celsius) measured at the maximum depth of the trawl.
haul: This number uniquely identifies a haul within a cruise. It is a sequential number, in chronological order of occurrence.
cpue: Catch number per area the net swept in number/square nautical mile.

2. Data Profile

Data profile of raw/original data

Variables & data types

Variables	*Time-variant/-invariant*	*Structured/Unstructured*	*Qualitative/Quantitative*	*Qualitative: Nominal/Ordinal Quantitative: Discrete/Continuous*
id	Time-invariant	Structured	Qualitative	Nominal
latitude	Time-invariant	Structured	Quantitative	Continuous
longitude	Time-invariant	Structured	Quantitative	Continuous
year	Time-variant	Structured	Qualitative	Ordinal
name	Time-invariant	Structured	Qualitative	Nominal
sex	Time-invariant	Structured	Qualitative	Nominal
bottom depth	Time-invariant	Structured	Quantitative	Discrete
surface_temperature	Time-variant	Structured	Quantitative	Continuous
bottom_temperature	Time-variant	Structured	Quantitative	Continuous
haul	Time-invariant	Structured	Qualitative	Nominal
cpue	Time-variant	Structured	Quantitative	Discrete

Cleaning of the data

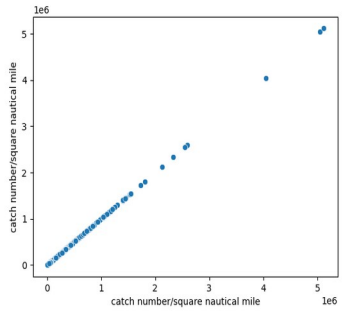
Using Jupyter there were no missing values found, no duplicates and no mixed-type columns.

Renaming of column 'cpue' into 'catch number/square nautical mile'.

Descriptive analysis

Variable	Min	Max	Mean	Std
catch number/square nautical mile	52	5 117 962	32 875	115 427

Min and max values of the numerical variables don’t look quite off or clearly suspicious. However, the scatterplot of variable 'catch number/square nautical mile' proved different, while 3 potential outliers were found.



The majority of values appear below 3 mio, with quite a gab before the next one. Hence, the 3 potential outliers were removed.

3. Defining Questions

Clarifying questions

1. Why do catch number values vary so much?

2. Did catch numbers change between 1975-2018?

Adjoining questions

1. Did catch numbers change geographically by region?

2. Did catch numbers increase or decrease over the years?

Funneling questions

1. Is there a certain trend in catch numbers between 1975-2018?

2. Is there a seasonal pattern/seasonality which may impact catch numbers?

3. Are there any special events which may impact the catch number?

Elevating questions

1. Were there any changes in fishing policies throughout the years, which may impact catch numbers?

2. Were there any natural disasters, which may have impacted the catch numbers?