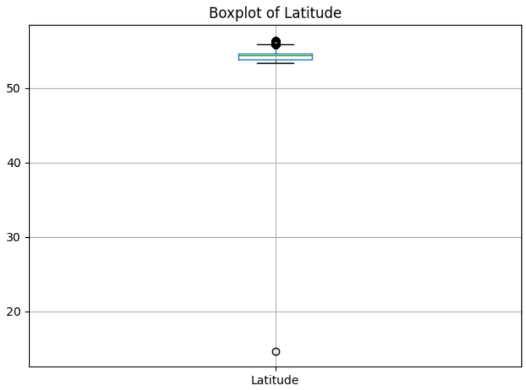
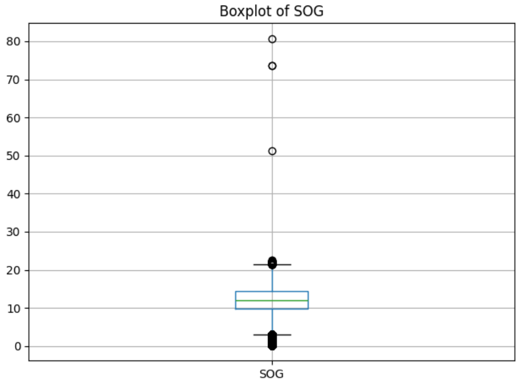
Data analysis

This report focuses only on outliers, missing values and other things needing to be cleaned – distribution is analyzed in profile.html.

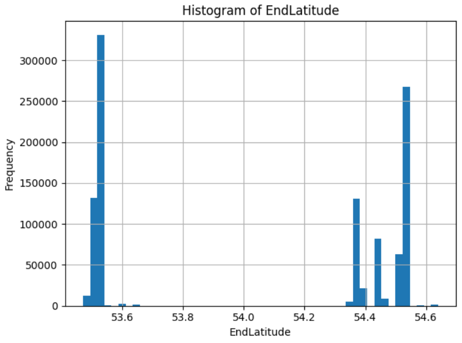
1. 116 213 duplicates
2. Breadth and length: 10924 zeros
3. Latitude: one outlier

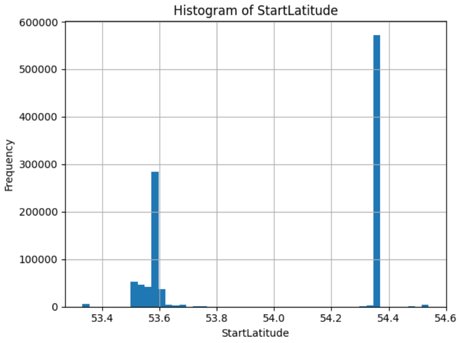


4. SOG: 3 examples with superspeed ships

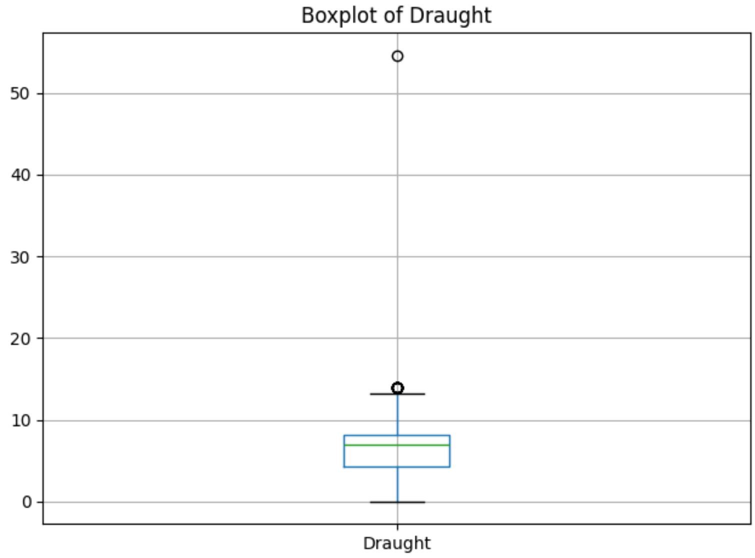


5. Why do starting and ending latitudes and longitudes differ?
   Because they start in different docks (at least lat’s suppose so)

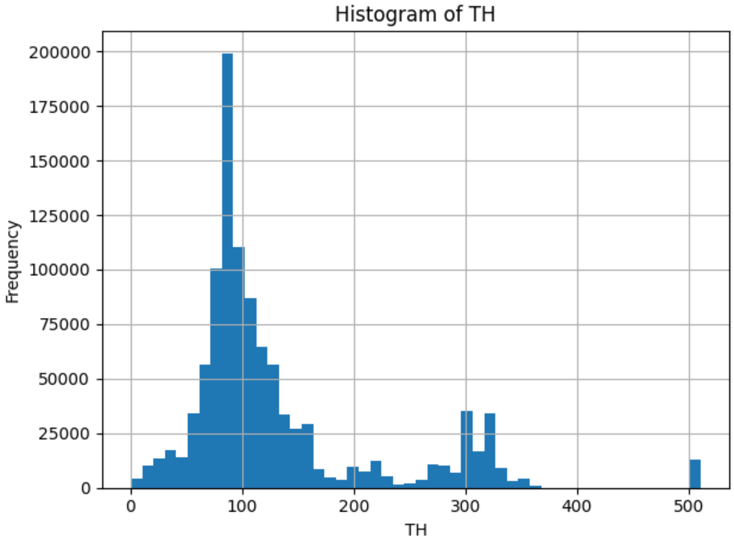




6. Draught: sometimes a ship becomes Titanic (or submarine). 16 270 missing values – to be deleted. 0 and 0,1 to be removed



7. TH: should only have values 0-360, but has 500. Over 360 needs to be removed



How to interprete box plots: clear outliers are errors and need to be deleted. Not so clear outliers are probably possible and will be checked later, but error-treshold will be over them.

Coordinates: more than 5 away from median are errors. (not anomalies, but errors)
Values constant for a trip changing during the trip
Length and Breadth: 0 are errors
Draught: 0 and 0,1 are errors
SOG: over 40 are errors
COG and TH: valid values are between 0 and 360

ANOMALIES:

Long stop in the middle of the sea (more than certain distance from harbor (start and end destinations))

Frequent or drastical course change (Over X degrees on this Latitude/Longitude in comparision to other examples/trips)

Frequent or drastical speed change (Over X knots, more than certain distance from harbor)

Frequent or drastical draught change (Over certain distance)

Latitude and longitude that are more than certain distance from usual coords on the same trip route

Starting and ending point slightly differ from usual values of startPort and endPort