8 Finding the Best Distribution with Speed Over Ground (sog) and vessel_type

As the data scientist and product owner, we want to know which probability distribution and fitting AIS data to the right distribution is valuable and might give us some insight about it.

SciPy is a Python library with many mathematical and statistical tools ready to be used and applied to the data.

Columns that we are going to work with are 'sog' (Speed Over Ground) and 'vessel_type' including 'N/A', 'Search And Rescue', 'Vessel', 'Tanker', 'Cargo', 'Wing In Ground-effect', 'Passenger', 'Tug', 'Law Enforcement', and 'High Speed Craft'.

Difference Between Akaike's Information Criteria (AIC) and Bayesian Information Criteria (BIC)

There are a couple of methods to estimate parameters of a distribution based on AIS data. Maximum Likelihood Estimation (MLE), Akaike information criterion (AIC), Maximizing Bayesian Posterior Probability , etc. SciPy performs parameter estimation using MLE (documentation).

AIC and BIC are widely used in model selection criteria. AIC means Akaike's Information Criteria and BIC means Bayesian Information Criteria.

When comparing the Bayesian Information Criteria and Akaike's Information Criteria, penalty for additional parameters is more in BIC than AIC. Unlike the AIC, the BIC penalises free parameters more strongly.

Akaike's Information Criteria generally tries to find unknown model that has high dimensional reality. This means the models are not true models in AIC. On the other hand, the Bayesian Information Criteria comes across only True models. It can also be said that Bayesian Information Criteria is consistent whereas Akaike's Information Criteria is not so.

When Akaike's Information Criteria will present the danger that it would outfit. the Bayesian Information Criteria will present the danger that it would underfit. Though BIC is more tolerant when compared to AIC, it shows less tolerance at higher numbers.

Performance Metrics: Sum of Squared Errors (SSE)

Sum of Squared Errors (SSE), or SSE. The error is the difference between the observed value and the predicted value. We usually want to minimize the error. The smaller the error, the better the estimation power.

Mathematically, SST (sum of Squared total) = SSR (sum of Squared regression) + SSE (Sum of Squared Errors) documentation

Statisticians routinely use the mean squared error (MSE) as the main measure of fit. The MSE is the sum of squared errors (SSE) divided by the degrees of freedom for error.

Performance Metric does allow the use of different scoring metrics that will be discussed, but all scores are reported so that they can be sorted in ascending order (largest score is best). Some evaluation metrics (like Sum of Squared Errors) are naturally descending scores (the smallest score is best).

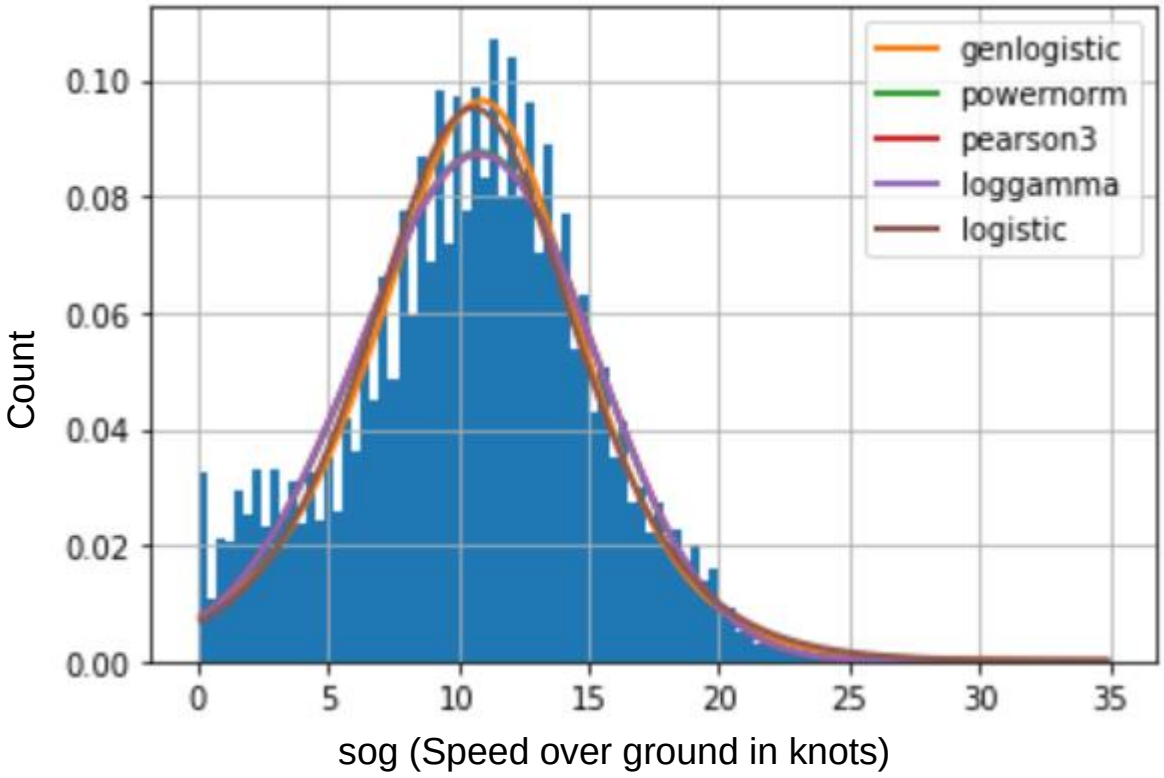
8.1 Cargo type: Top 5 Fitted Distributions

It has fitted all distributions including common distributions; gamma, lognormal, beta, burr and normal distributions. It supports all distributions supported by the Scipy library.

Calling the **summary()** method on the fitted object shows the different distributions and fit statistics such as **sumsquare_error**, **Akaike information criterion** (aic) and **Bayesian information criterion** (bic) values. By default, the summary function ranks the best five distributions based on the **sumsquare_error** values in ascending order. Additionally, it provides an illustration of different distributions fitted over a histogram.

Based on the sumsquare_error value the best distribution for the sog (Speed over ground in knots) data is the generalised logistic distribution with 197475 reports.

Plot of Top 5 Fitted Distributions (Cargo Speed: 197475 reports)



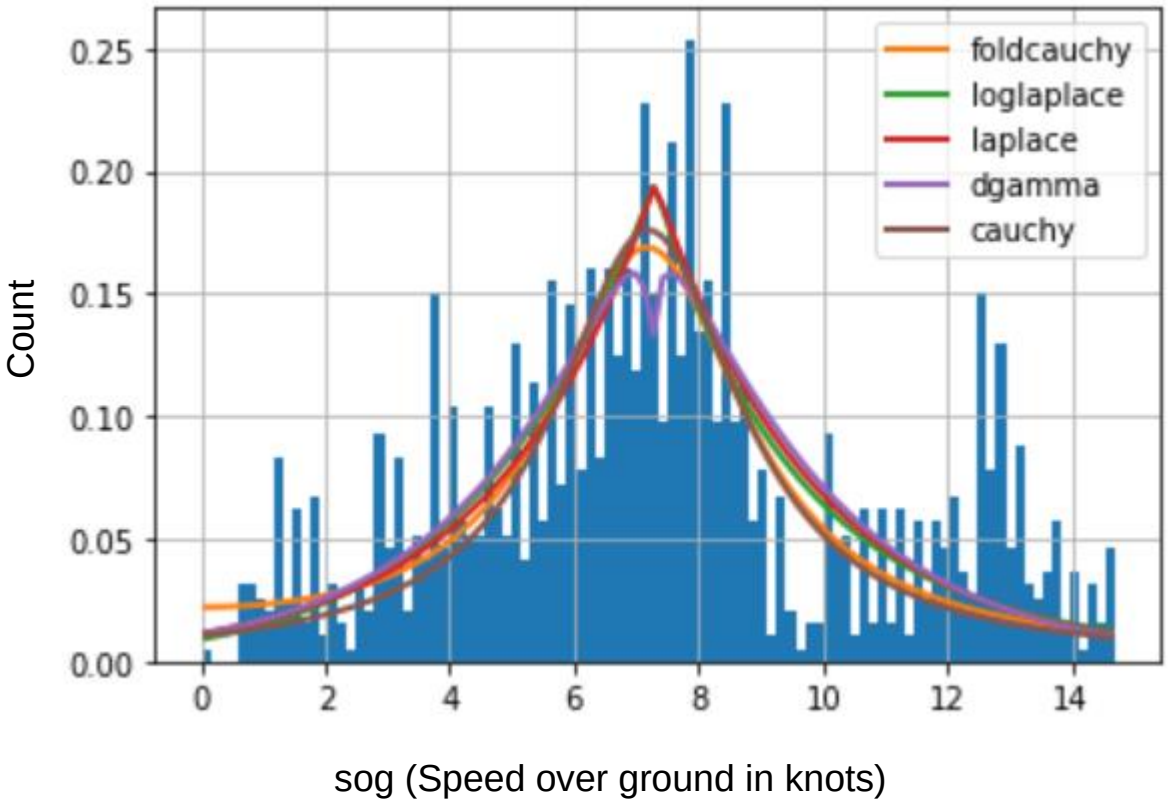
Top 5 Fitted Distributions (Cargo Speed: 197475 reports)

	sumsquare_error	aic	bic	kl_div
genlogistic	0.005878	1048.725316	-3.422171e+06	inf
powernorm	0.006401	1296.360186	-3.405346e+06	inf
pearson3	0.006494	1347.257116	-3.402498e+06	inf
loggamma	0.006511	1328.668346	-3.401980e+06	inf
logistic	0.006684	1016.553496	-3.396812e+06	inf

8.2 High Speed Craft: Top 5 Fitted Distributions:

Based on the sumsquare_error value the best distribution for the sog (Speed over ground in knots) data is the fold cauchy distribution with 1313 reports.

Plot of Top 5 Fitted Distributions (High Speed Craft Speed: 1313 reports)



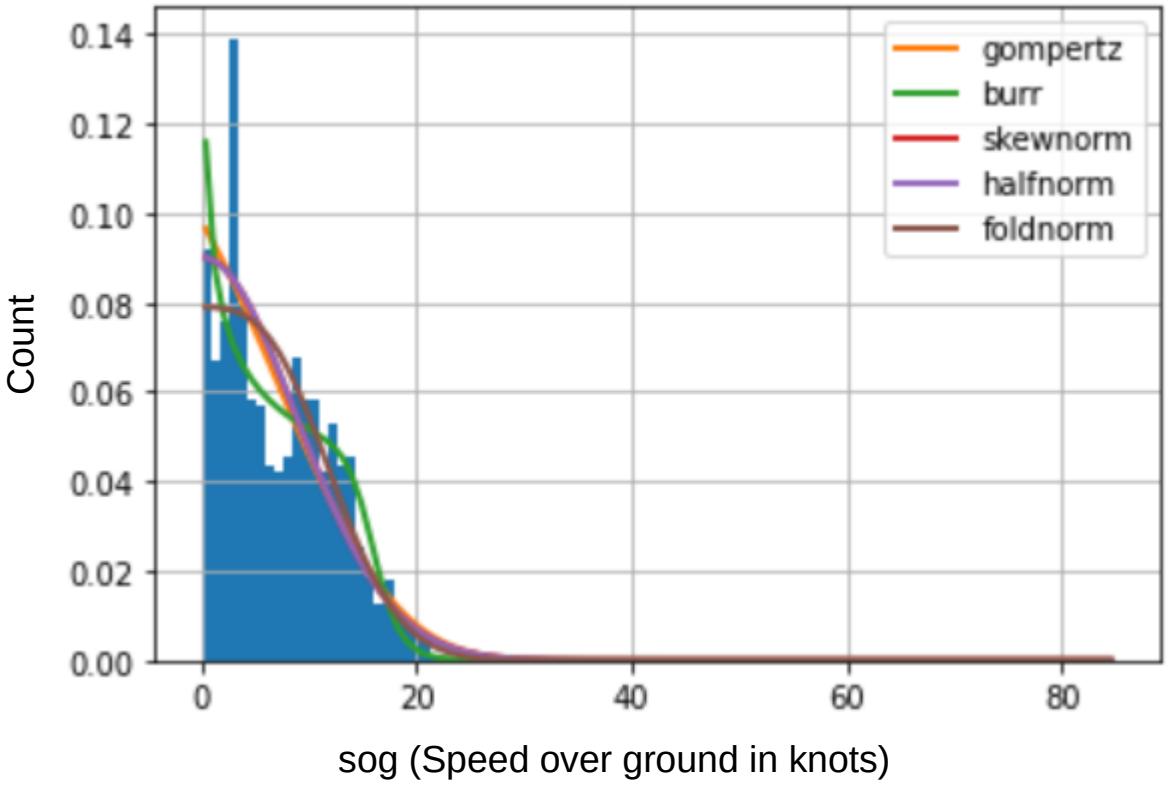
Top 5 Fitted Distributions (High Speed Craft Speed: 1313 reports)

	sumsquare_error	aic	bic	kl_div
foldcauchy	0.161229	627.471336	-11802.024396	inf
loglaplace	0.161915	621.382930	-11796.453264	inf
laplace	0.164342	617.766548	-11784.091603	inf
dgamma	0.165880	615.708546	-11764.683038	inf
cauchy	0.169487	655.033618	-11743.615290	inf

8.3 N/A type: Top 5 Fitted Distributions

Based on the sumsquare_error value the best distribution for the sog (Speed over ground in knots) data is the gompertz distribution with 17405 reports.

Plot of Top 5 Fitted Distributions (N/A (not available) Speed: 17405 reports)



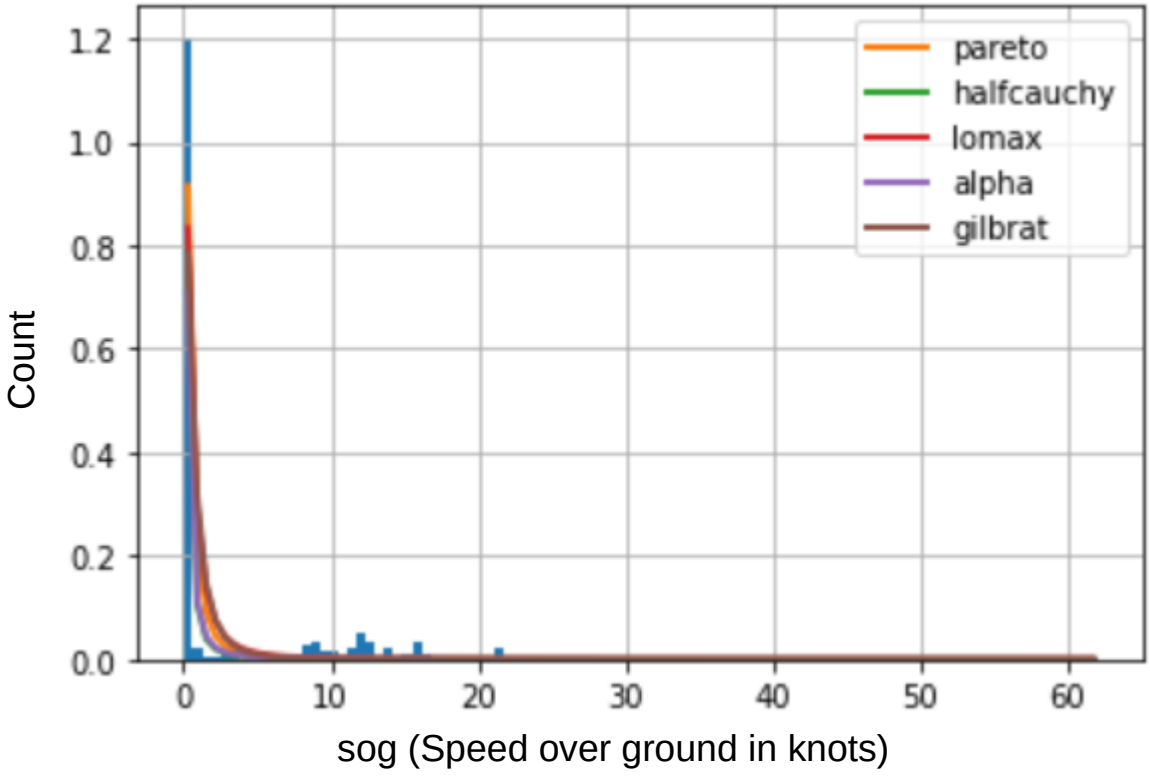
Top 5 Fitted Distributions (N/A (not available) Speed: 17405 reports)

	sumsquare_error	aic	bic	kl_div
gompertz	0.006950	8377.740017	-256408.981976	inf
burr	0.007101	2993.278582	-256024.228059	inf
skewnorm	0.007155	3566.442177	-255902.345214	inf
halfnorm	0.007155	3564.871769	-255910.510871	inf
foldnorm	0.007286	5370.514293	-255585.389152	inf

8.4 Search And Rescue type: Top 5 Fitted Distributions

Based on the sumsquare_error value the best distribution for the sog (Speed over ground in knots) data is the pareto distribution with 730 reports.

Plot of Top 5 Fitted Distributions (Search and Rescue Speed: 730 reports)



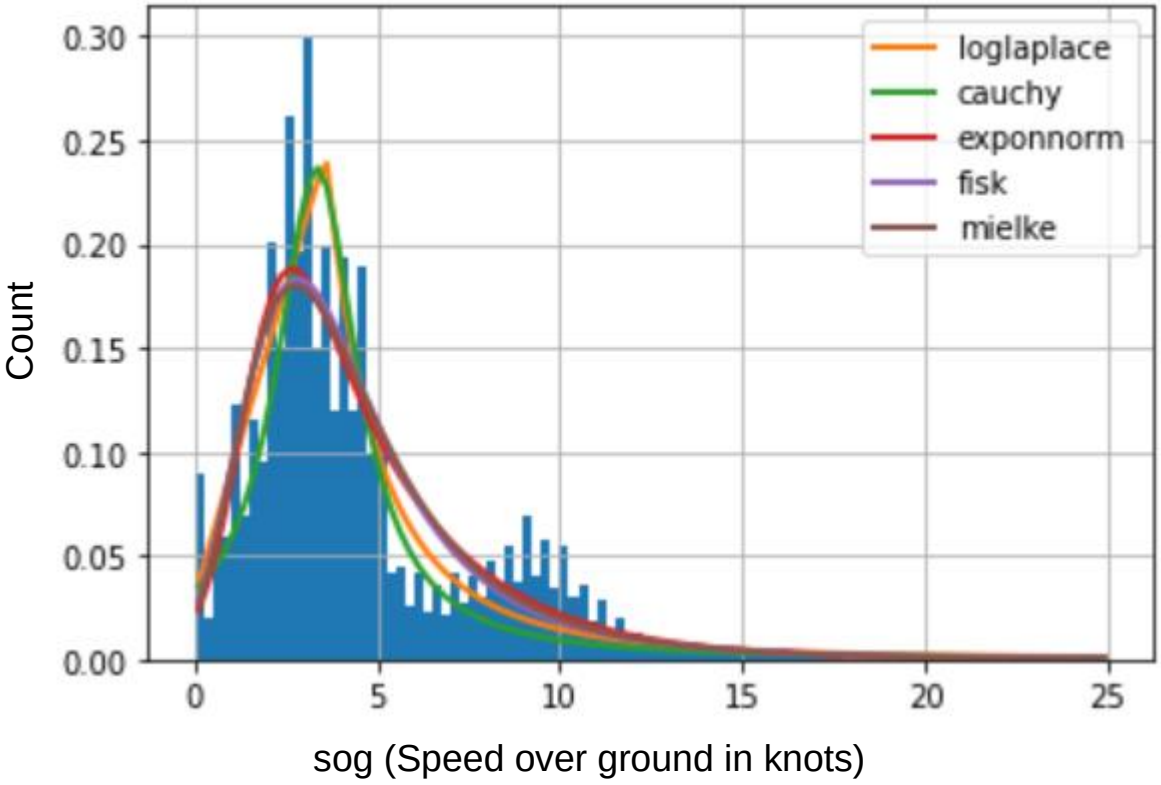
Top 5 Fitted Distributions (Search and Rescue Speed: 730 reports)

	sumsquare_error	aic	bic	kl_div
pareto	0.140894	1769.722640	-6223.758007	inf
halfcauchy	0.178963	1728.761088	-6055.758666	inf
lomax	0.241576	1693.417272	-5830.161724	inf
alpha	0.264447	1689.987882	-5764.127694	inf
gilbrat	0.297033	2170.367470	-5685.891595	inf

8.5 Vessel type: Top 5 Fitted Distributions

Based on the sumsquare_error value the best distribution for the sog (Speed over ground in knots) data is the log laplace distribution with 43235 reports.

Plot of Top 5 Fitted Distributions (Vessel Speed: 43235 reports)



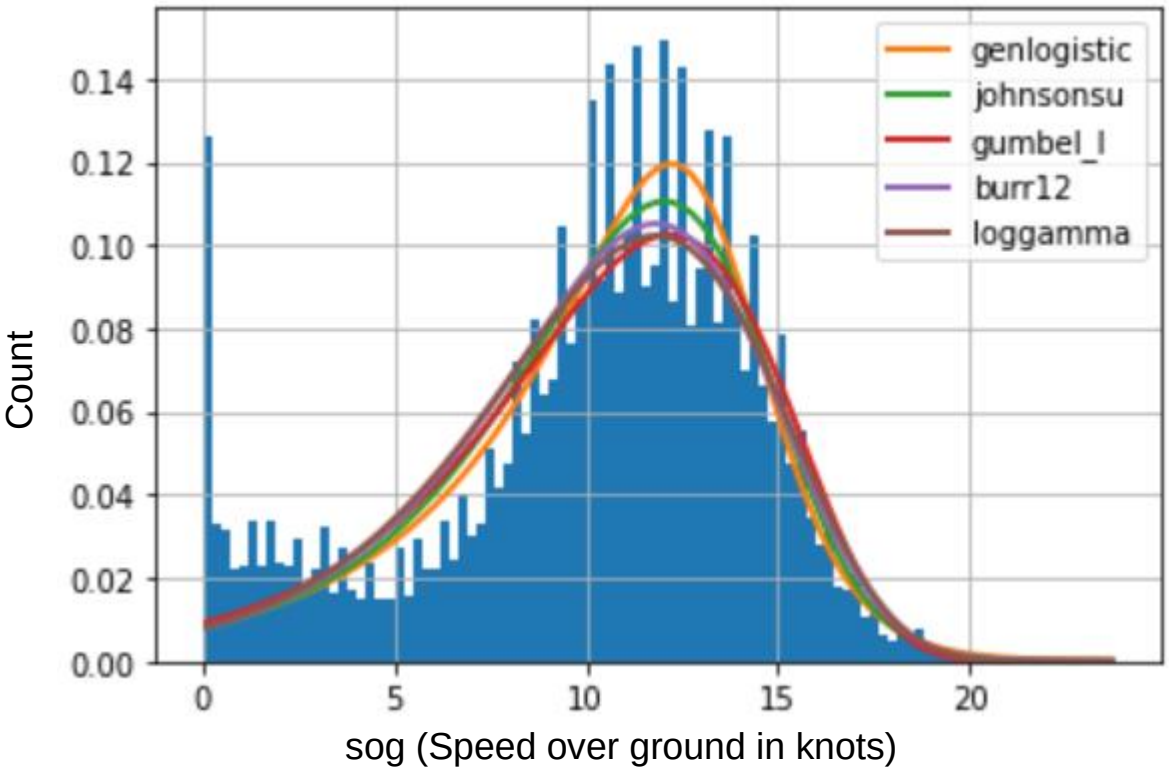
Top 5 Fitted Distributions (Vessel Speed: 43235 reports)

	sumsquare_error	aic	bic	kl_div
loglaplace	0.060473	911.950158	-582773.964349	inf
cauchy	0.063939	966.054184	-580375.163991	inf
exponnorm	0.064643	961.196859	-579891.428251	inf
fisk	0.065822	925.128579	-579109.642965	inf
mielke	0.068477	940.719641	-577389.324926	inf

8.6 Tanker type: Top 5 Fitted Distributions

Based on the sumsquare_error value the best distribution for the sog (Speed over ground in knots) data is the generalised logistic distribution with 98637 reports.

Plot of Top 5 Fitted Distributions (Tanker Speed: 98637 reports)



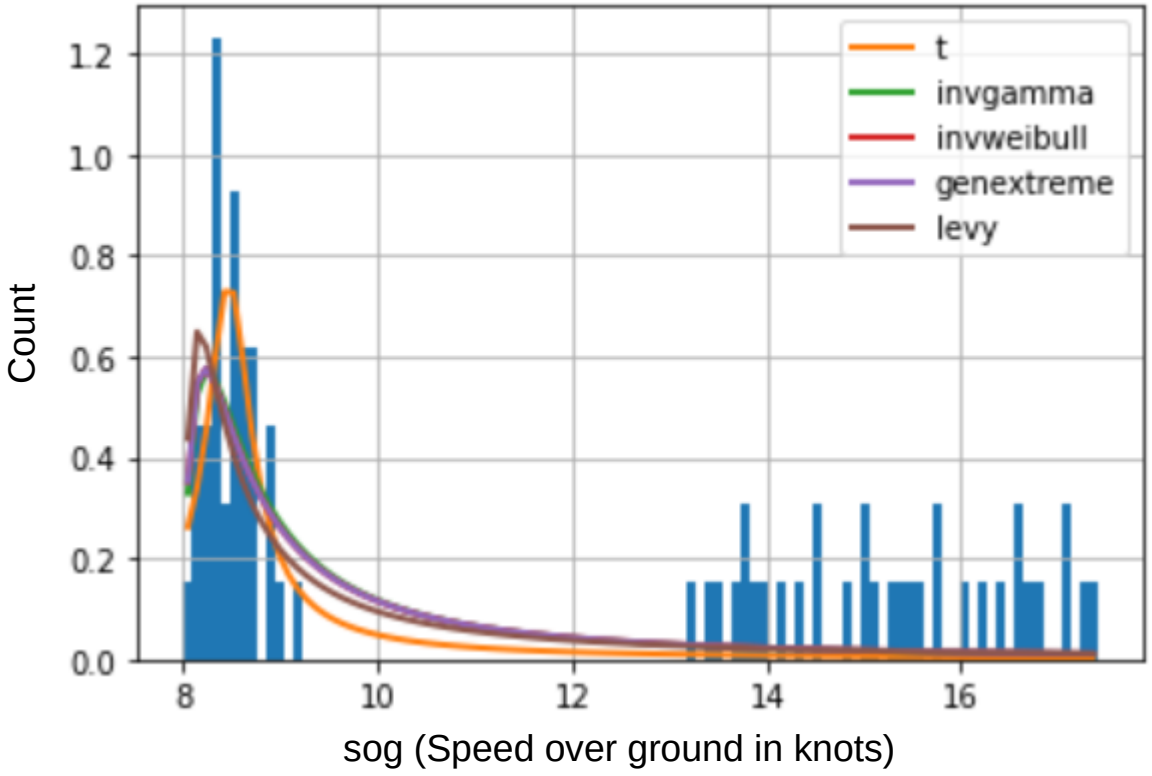
Top 5 Fitted Distributions (Tanker Speed: 98637 reports)

	sumsquare_error	aic	bic	kl_div
genlogistic	0.034156	839.469665	-1.467292e+06	0.068355
johnsonsu	0.035433	868.409323	-1.463660e+06	0.073425
gumbel_l	0.037773	1043.072531	-1.457374e+06	0.078804
burr12	0.038087	897.981730	-1.456535e+06	0.080567
loggamma	0.039458	949.069272	-1.453058e+06	0.084170

8.7 Wing In Ground-effect type: Top 5 Fitted Distributions

Based on the sumsquare_error value the best distribution for the sog (Speed over ground in knots) data is the t distribution with 69 reports.

Plot of Top 5 Fitted Distributions (Wing in Ground-effect Speed: 69 reports)



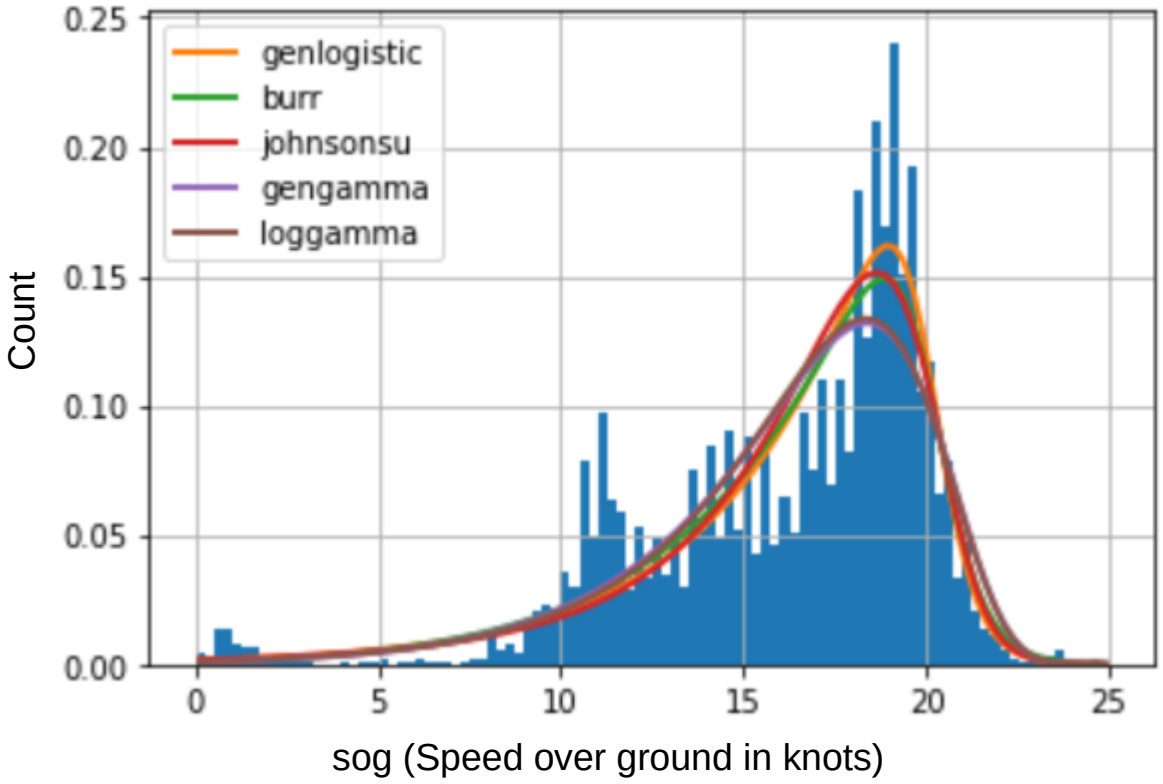
Top 5 Fitted Distributions (Wing in Ground-effect Speed: 69 reports)

	sumsquare_error	aic	bic	kl_div
t	1.940625	807.951226	-233.703317	inf
invgamma	2.338762	629.590158	-220.827132	inf
invweibull	2.357476	626.807848	-220.277210	inf
genextreme	2.357477	626.807760	-220.277195	inf
levy	2.385253	644.688406	-223.703081	inf

8.8 Passenger type: Top 5 Fitted Distributions

Based on the sumsquare_error value the best distribution for the sog (Speed over ground in knots) data is the generalised logistic distribution with 9917 reports.

Plot of Top 5 Fitted Distributions (Passenger Speed: 9917 reports)



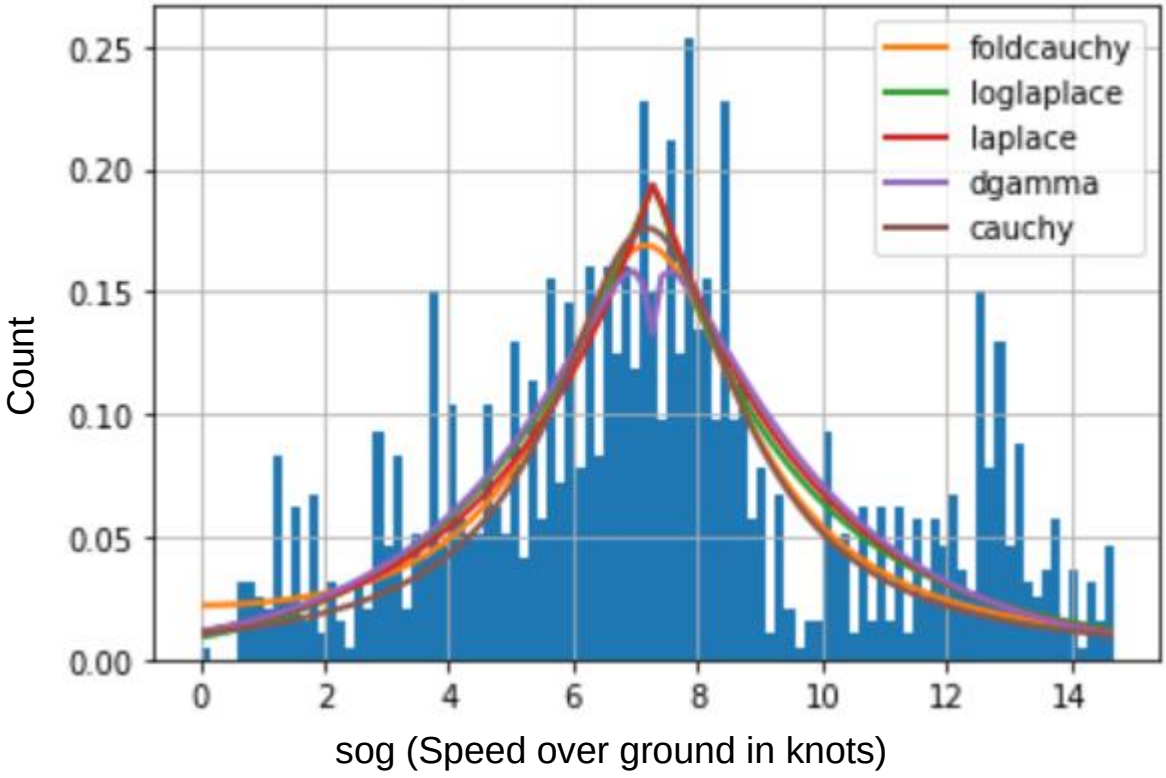
Top 5 Fitted Distributions (Passenger Speed: 9917 reports)

	sumsquare_error	aic	bic	kl_div
genlogistic	0.045619	861.505495	-121846.818390	inf
burr	0.049241	849.409724	-121079.883816	inf
johnsonsu	0.055398	864.133822	-119911.539106	inf
gengamma	0.064367	900.080787	-118423.407304	inf
loggamma	0.065229	892.701222	-118300.680579	inf

8.9 Tug type: Top 5 Fitted Distributions

Based on the sumsquare_error value the best distribution for the sog (Speed over ground in knots) data is the fold cauchy distribution with 1313 reports.

Plot of Top 5 Fitted Distributions (Tug Speed: 1313 reports)



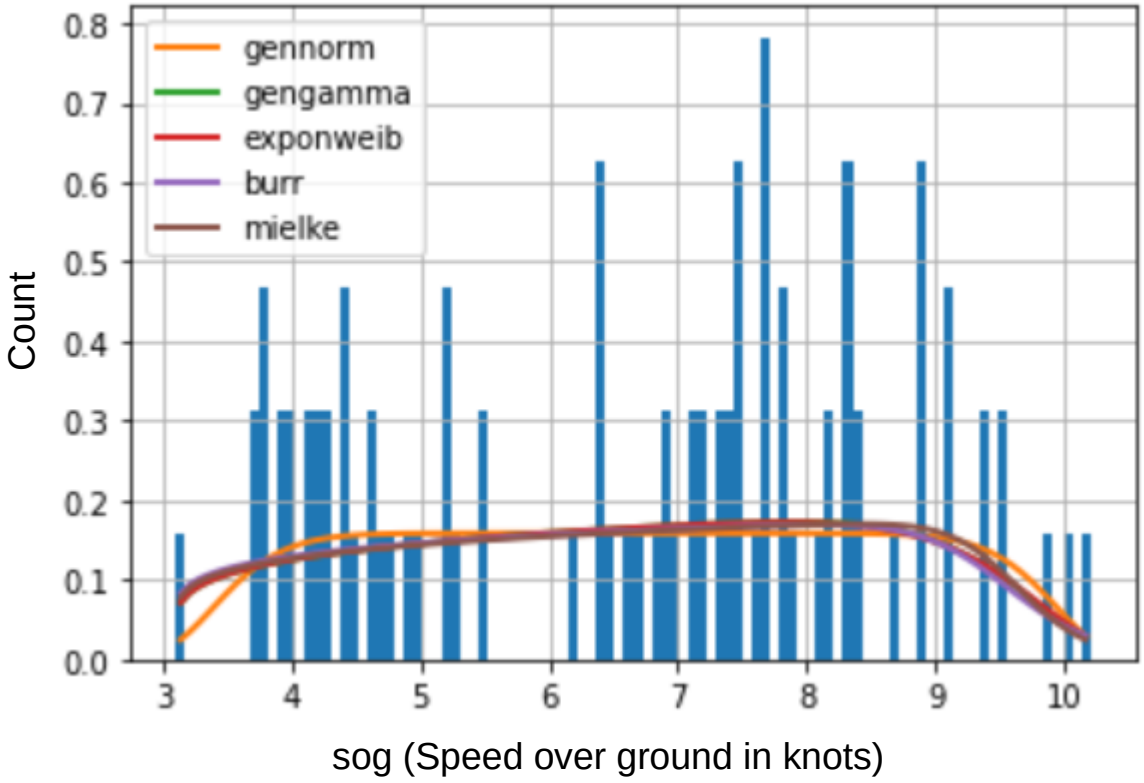
Top 5 Fitted Distributions (Tug Speed: 1313 reports)

	sumsquare_error	aic	bic	kl_div
foldcauchy	0.161229	627.471336	-11802.024396	inf
loglaplace	0.161915	621.382930	-11796.453264	inf
laplace	0.164342	617.766548	-11784.091603	inf
dgamma	0.165880	615.708546	-11764.683038	inf
cauchy	0.169487	655.033618	-11743.615290	inf

8.10 Law Enforcement type: Top 5 Fitted Distributions

Based on the sumsquare_error value the best distribution for the sog (Speed over ground in knots) data is the generalised normal distribution with 90 reports.

Plot of Top 5 Fitted Distributions (Law Enforcement Speed: 90 reports)



Top 5 Fitted Distributions (Law Enforcement Speed: 90 reports)

	sumsquare_error	aic	bic	kl_div
gennorm	3.346597	410.312495	-282.768484	inf
gengamma	3.351416	410.481051	-278.139166	inf
exponweib	3.352022	410.136474	-278.122881	inf
burr	3.352216	411.910301	-278.117698	inf
mielke	3.359504	411.186123	-277.922239	inf

9 Use Chi-squared tests on vessel temporal analysis and establish believable velocity distributions (by type)

As the data scientist and product owner, we would like to do Chi-squared (95% confidence) tests to establish the stochastic distribution (random or unique) of vessel speed by vessel type. Repeat different candidate distributions to find a good (i.e. pass) for the test data. Then verify against a different test set (e.g. a later time) to verify the hypothesis still holds.

This assumes repeating the test for each different vessel type so that we could potentially have different distributions values or distribution types for each vessel type. Learning set is the English Channel set.

Establishing the distribution is not invariant overtime would be a valid outcome for the story.

Chi-Square Test- The test is applied when you have two categorical variables from a single population. It is used to determine whether there is a significant association between the two variables ("Speed Over the Ground (SOG)" is the speed of the vessel relative to the surface of the earth and "vessel_type" (Cargo,High Speed Craft, Law Enforcement, Passenger, Search And Rescue, Tanker, Tug, Vessel, and Wing In Ground-effect).

Chi-Square test is a statistical test which is used to find out the difference between the observed and the expected data we can also use this test to find the correlation between categorical variables in our data. The purpose of this test is to determine if the difference between 2 categorical variables is due to chance, or if it is due to a relationship between them.

[30]
```
alpha = 0.05
```

+ Code + Markdown

[31]
```
# Calcualtion of Chisquare test statistics
chi_square = 0
rows = ais['sog'].unique()
columns = ais['vessel_type'].unique()
for i in columns:
    for j in rows:
        O = ais_crosstab[i][j]
        E = ais_crosstab[i]['Total'] * ais_crosstab['Total'][j] / ais_crosstab['Total']['Total']
        chi_square += (O-E)**2/E
```

[32]
```
# The p-value approach
print("Approach 1: The p-value approach to hypothesis testing in the decision rule")
p_value = 1 - stats.norm.cdf(chi_square, (len(rows)-1)*(len(columns)-1))
conclusion = "Failed to reject the null hypothesis."
if p_value <= alpha:
    conclusion = "Null Hypothesis is rejected."

print("chisquare-score is:", chi_square, " and p value is:", p_value)
print(conclusion)
```

```
Approach 1: The p-value approach to hypothesis testing in the decision rule
chisquare-score is: 218835.28278505537  and p value is: 0.0
Null Hypothesis is rejected.
```

Ref: https://onezero.blog/finding-the-best-distribution-that-fits-your-data-using-pythons-fitter-library/

How to order by Akaike Information Criterion or Bayesian Information Criterion? The KL-divergence appears to be infinite, which is worrying as KL should be low for a good fit.???