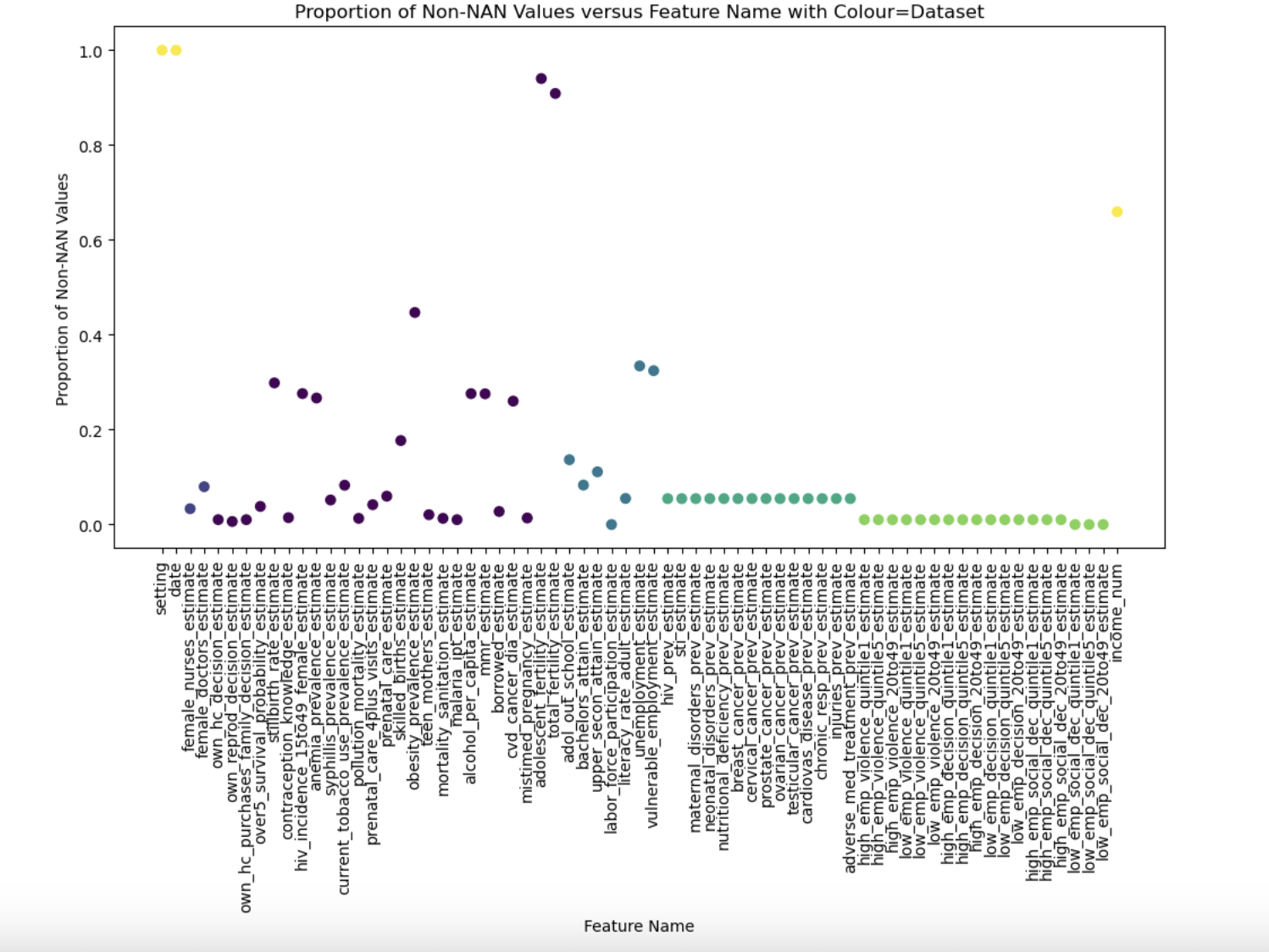
**Agenda:**

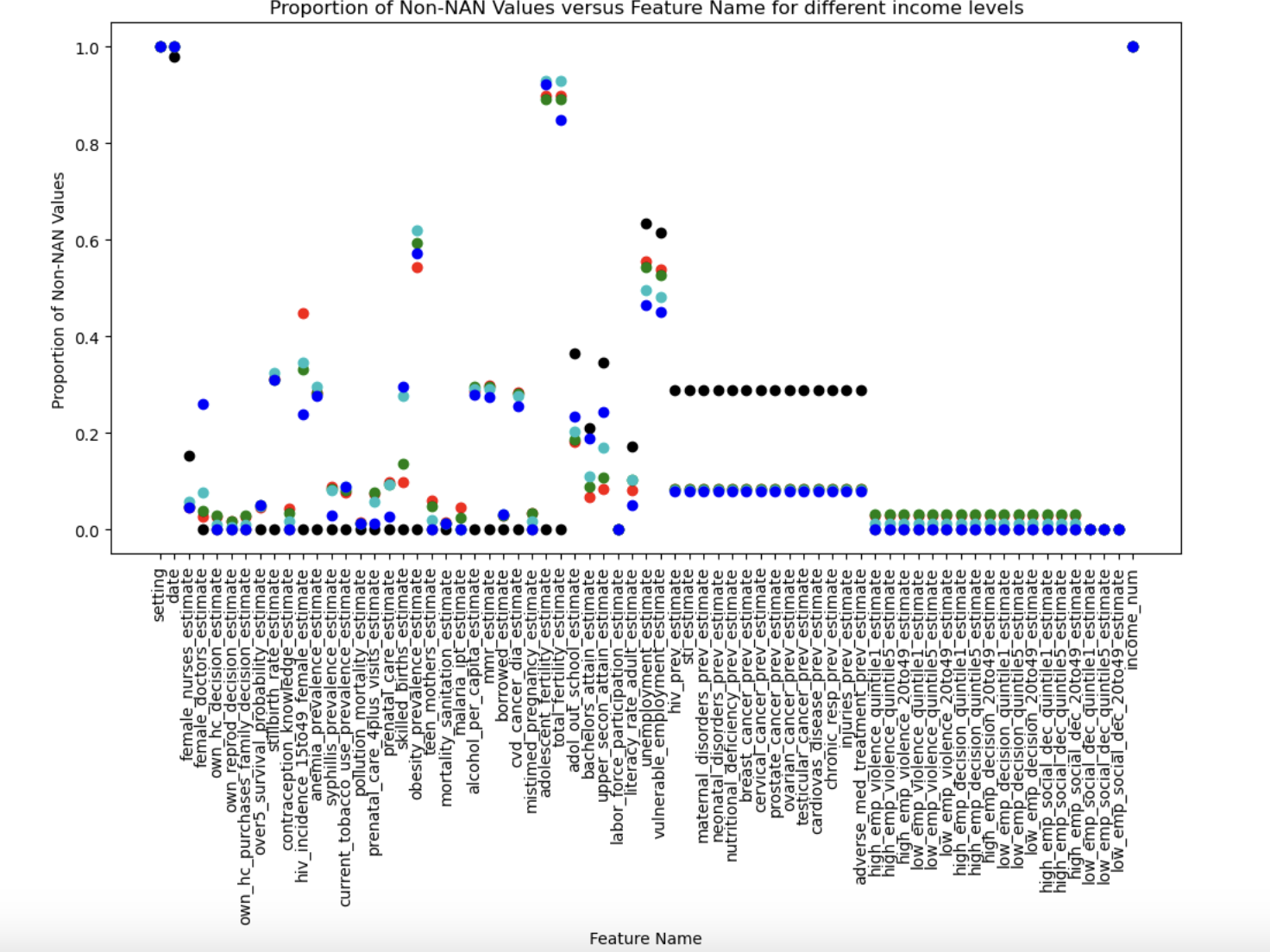
* Visualisation of missing data by feature, year, and country income level
* Linear and polynomial interpolation/extrapolation methods
* Kolmogorov-smirnov test
  + Normal
  + Log-normal
  + Logistic
  + Uniform

**Visualising the Distribution of Missing Data**

The following plot shows the proportion of non-NAN values between 2024 and 1960 for each feature (with all countries considered). The colours of the datapoints represent the different datasets that the information comes from (purple = women’s health dataset from the World Bank, etc.). Only ~21% of features were present in the dataset at a proportion of more than 10%

****

The following plot again shows the proportion of non-NAN values per feature, but this time grouped by country income level. Blue = highest income, light blue = upper middle income, green = lower middle income, red = low income, and black = unknown (size = 4). Income level did not appear to be the sole determinant of missing values.



The following plot shows how the proportion of non-NAN values has increased over time, with large increases occurring after 2000. You can see which years within which the majority of countries report data, as well as the drop off in data reporting during COVID.

A graph of a graph with blue dots

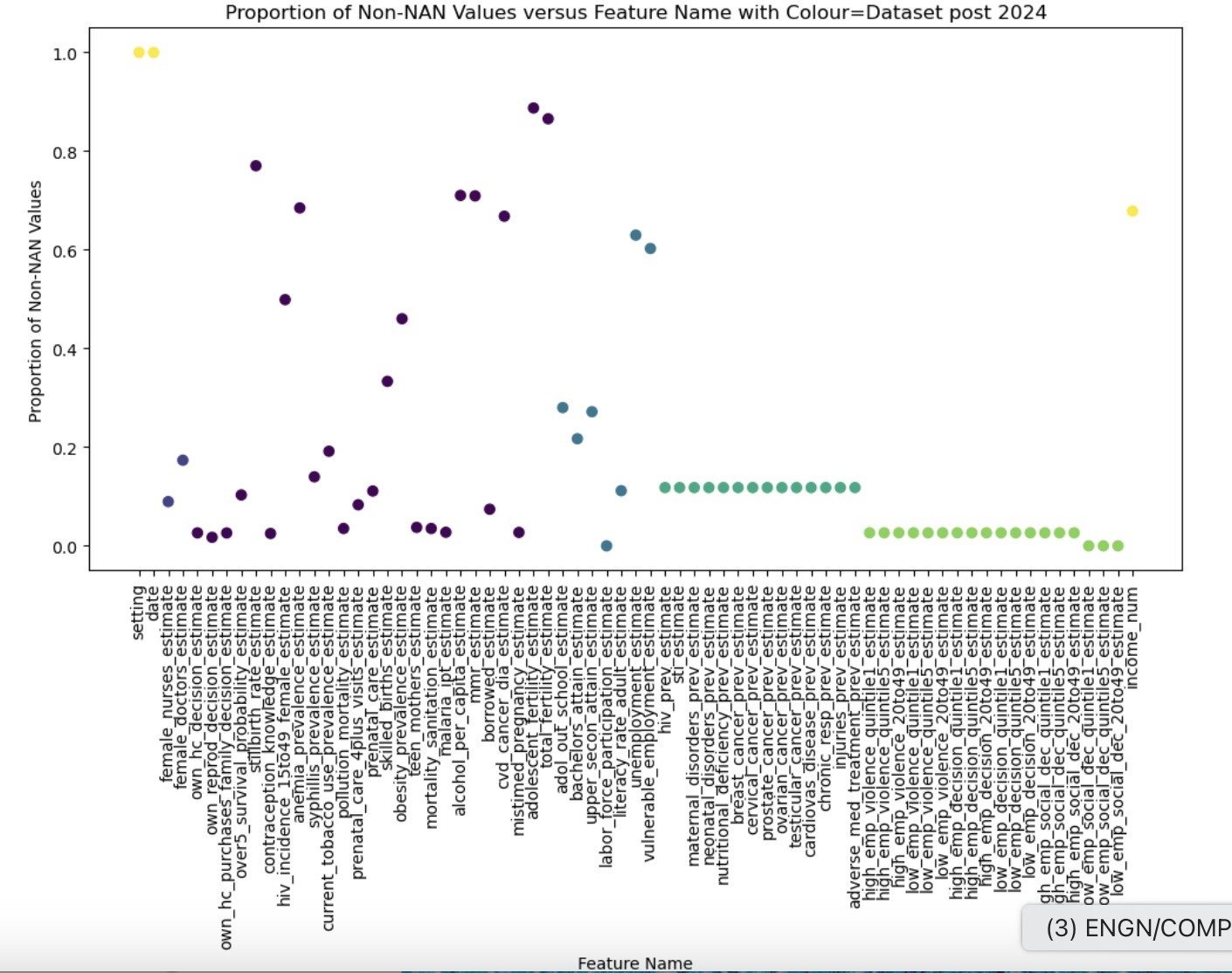
AI-generated content may be incorrect.

The following plot shows the distribution of missing data per year for countries of different income levels. It shows that the increase in data reporting is not country specific.

A graph of a number of dots

AI-generated content may be incorrect.

The following plot shows the proportion of non-NAN values for each feature across all countries post 2000. There are some higher proportions, but still many features with lots of missing data. These are features that are likely to have only started being collected much more recently.



**Iterative approach**

To curate a higher quality dataset, I iteratively removed rows or columns when they contained a higher proportion of NAN values than my threshold. The following results are for all countries with data from 2017 (with no interpolation/ extrapolation).

Columns and then rows:

* Dropped columns with >75% NAN values and then rows with >25% NAN values
  + Only ~9% NAN values
  + It had 131 rows (countries) and 19 columns (features)
* Dropped columns with > 50% NAN values and then rows with > 50% NAN values
  + Only ~14% NAN values
  + It had 237 rows (countries) and 15 columns (features)
* Dropped columns with >25% NAN values and then rows with > 75% NAN values
  + Only ~6.6% NAN values
  + It had 265 rows (countries) and 9 columns (features)

Rows and then columns:

* Dropped rows with > 25% NAN values and then columns with > 75% NAN values
  + 0.0%
  + It had 1 row (countries) and 13 columns (features)
* Dropped rows with > 50% NAN values and then columns with > 50% NAN values
  + Only ~18.6% NAN values
  + It had 6 rows (countries) and 43 columns (features)
* Dropped rows with >75% NAN values and then columns with >25% NAN values
  + Only ~3.2% NAN values
  + It had 98 rows (countries) and 17 columns (features)

Therefore, it appears that:

* *Getting rid of rows* first reduces the number of rows in the final dataset more than getting rid of the columns first reduces the number of columns.
* Dropping columns with >75% NAN values and then rows with >25% NAN values included the highest number of features with a reasonable number of countries.

**For a specific country, looking at the distribution of specific features with overlapping plots of imputed data**

I chose to use Afghanistan as my example. (I have similar interpolations for countries from all four income levels). The following graphs show purely imputation with data that demonstrates different trends.

A graph of a line

AI-generated content may be incorrect.

A group of graphs with numbers and numbers

AI-generated content may be incorrect.

A graph of data in different positions

AI-generated content may be incorrect.

Linear Extrapolation:

Using the following formula, I performed linear imputation on Afghanistan’s chronic respiratory prevalence data. As you can see, the data does not appear to fit a purely linear trend. The smaller gradient than expected is likely due to the smaller gradient between the two leftmost raw datapoints.

A graph with numbers and a number of data

AI-generated content may be incorrect.A graph of an orange dotted line

AI-generated content may be incorrect.

Polynomial:

However, using a polynomial fit may substantially overestimate the data values.

A graph of a number of years

AI-generated content may be incorrect.

**Kolmogorov-smirnov test**

I used the Kolmogorov-smirnov testto determine whether any column in the linearly or polynomially interpolated Afghanistan data fit the normal, log-normal, logistic, or uniform distribution. Unfortunately, the test indicated that the data did not fit these distributions, which makes sense when looking at their trends.

**To-Dos:**

* Discuss how to best manage missing data.
* Discuss other probability distributions that may better fit the data.