Labour demand obtained through online job adverts data collection

Gueorguie Vassilev
Head of Skills and Human Development
Office for National Statistics

22 May 2025

Office for National Statistics

Official Sensitive (if required)

Contents

- What do users want to know
- Context and existing outputs
- Users, know the data
- Methods
- Benefits
- Conclusions and future work

What users want to know

Questions we're being asked - demand

- What is the top job in demand in my area?
- What are the growing jobs in demand right now?
- What skills are employers looking for? What are the UK's emerging skills?
- What sectors are hiring for occupation X, and where?
- How many AI jobs are there out there? What is the demand for AI skills?
- How many green jobs are in demand, and where?

Questions we’re being asked – shortages and inequalities?

- What jobs in demand are not being filled? Which are the shortage occupations?
  - How can we train people for them (i.e. what skills are needed for them?)
- How do job markets differ locally? How do people’s opportunities in joining the labour market differ locally?

Existing outputs

Potted history of ONS involvement with online job adverts

- 2020 – started tracking weekly changes to index of demand
- 2020 – iterative improvements to data
- 2021-22 – experimentation – one-off requests
- 2022 and 2023 – beginning to publish granular geographic dissemination
- 2024 – occupational demand first experimental output

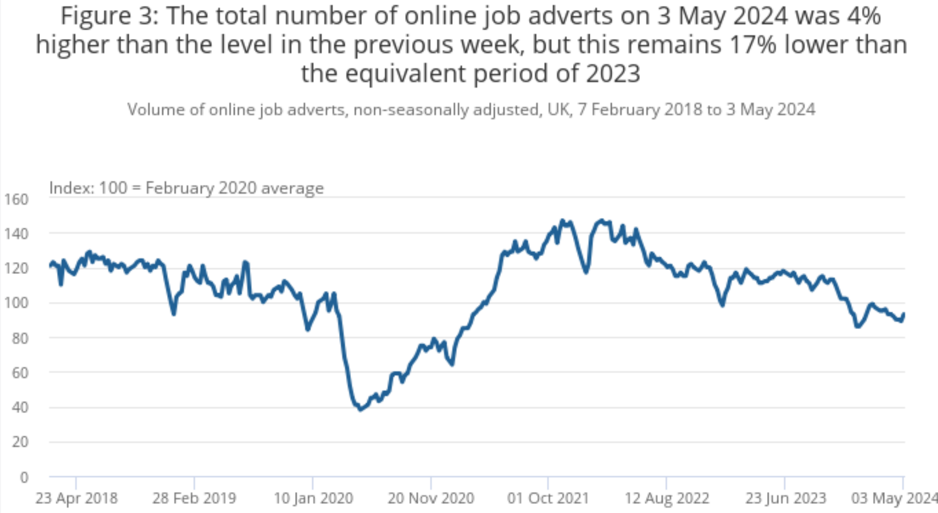


Figure 3: The total number of online job adverts on 3 May 2024 was 4% higher than the level in the previous week, but this remains 17% lower than the equivalent period of 2023

Volume of online job adverts, non-seasonally adjusted, UK, 7 February 2018 to 3 May 2024

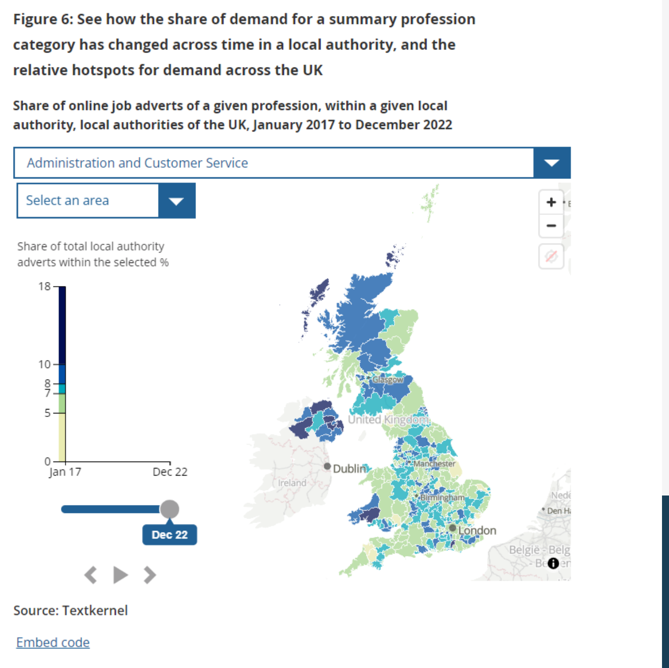


Figure 6: See how the share of demand for a summary profession category has changed across time in a local authority, and the relative hotspots for demand across the UK

Share of online job adverts of a given profession, within a given local authority, local authorities of the UK, January 2017 to December 2022

Administration and Customer Service

Select an area

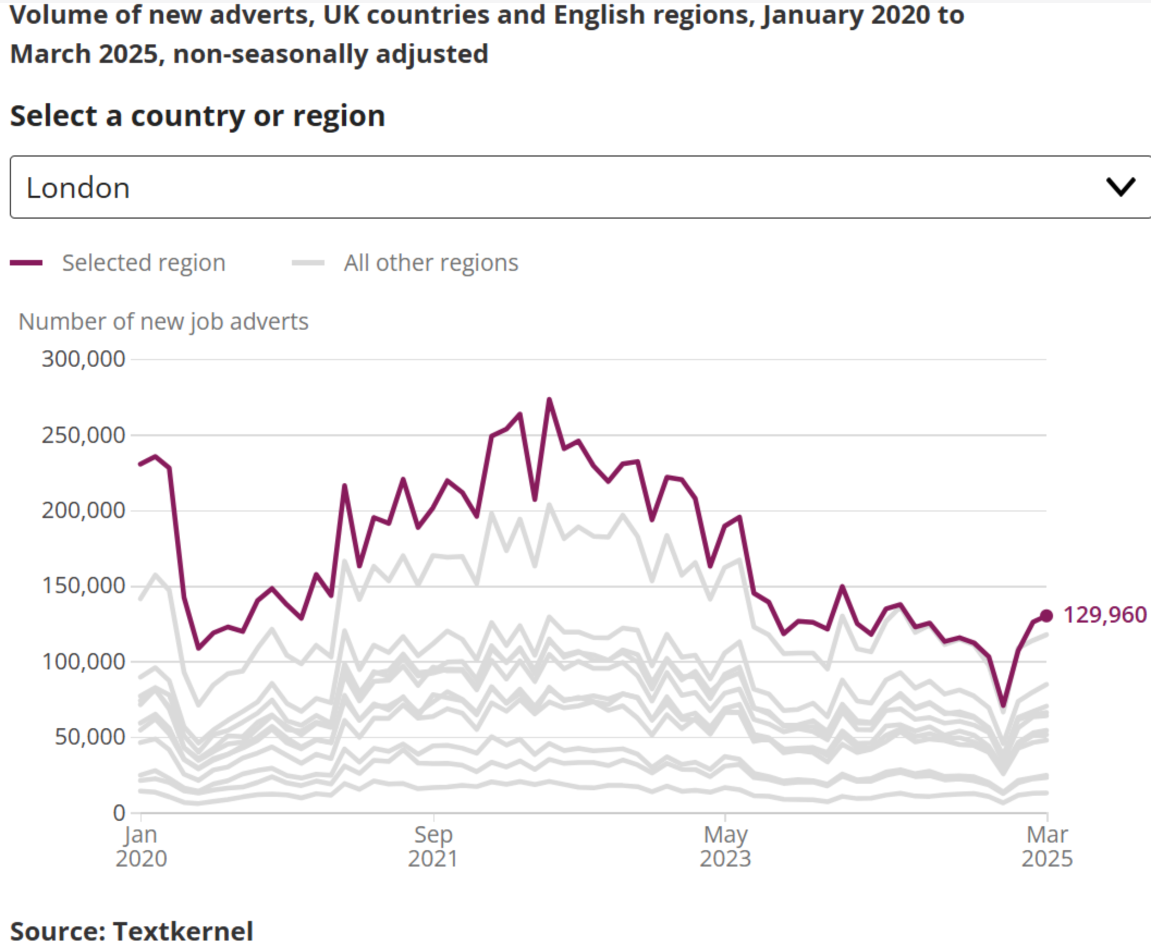
Source: Textkernel

Embed code

Current outputs

- Regular volumes up to by occupation at 4-digit SOC 2020 and local authority

  - two monthly publications



Volume of new adverts, UK countries and English regions, January 2020 to March 2025, non-seasonally adjusted

Select a country or region

London

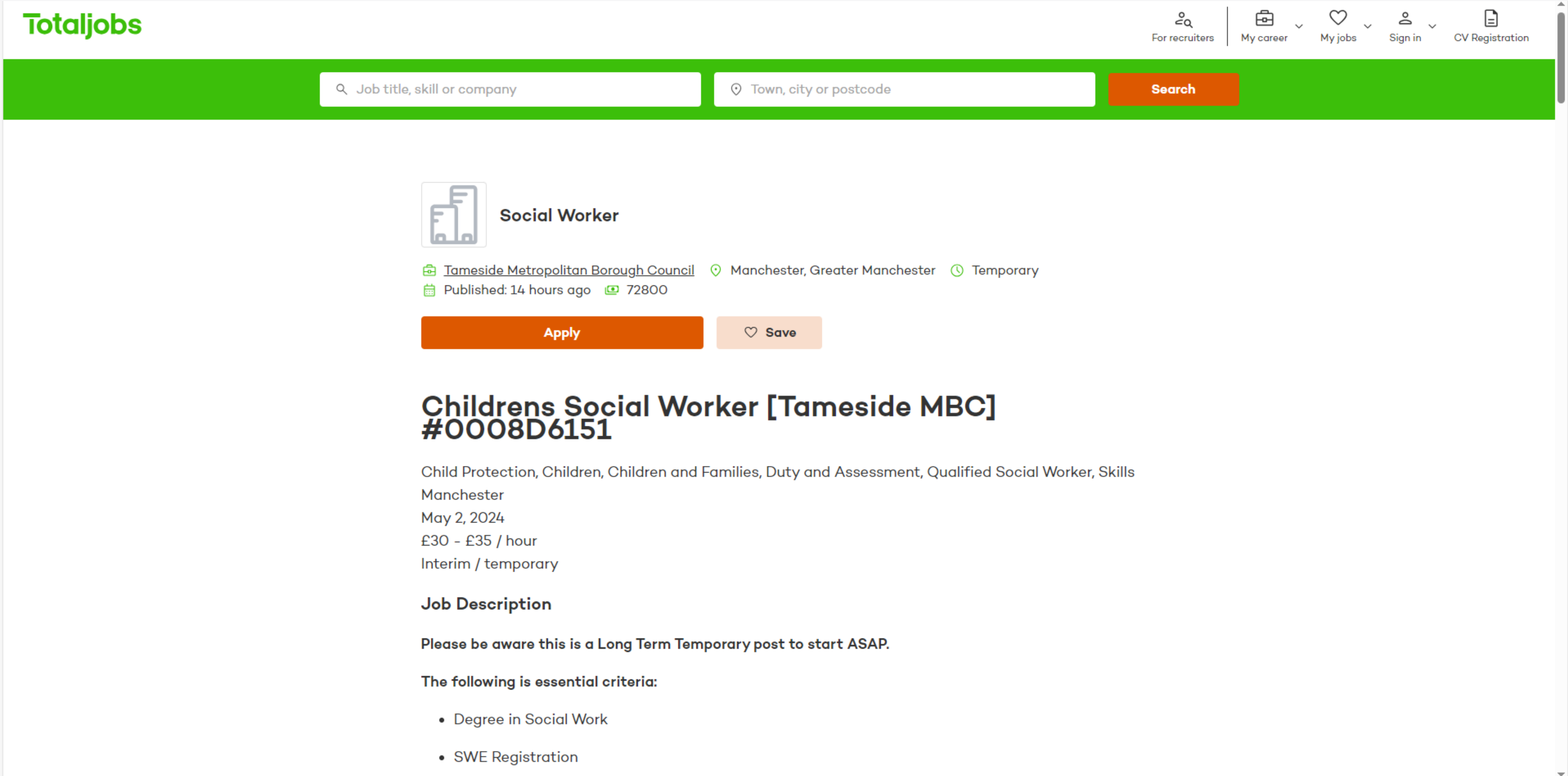
Source: Textkernel

- Continuing to derive ad-hocs: zero-hours contracts, salaries

Users, know the data

Take into account the whole data collection process

Real life example – extract on 7 May 2024



Social Worker

Tameside Metropolitan Borough Council · Manchester, Greater Manchester · Temporary
Published: 14 hours ago · 72800

Childrens Social Worker [Tameside MBC] #0008D6151

Child Protection, Children, Children and Families, Duty and Assessment, Qualified Social Worker, Skills
Manchester
May 2, 2024
£30 - £35 / hour
Interim / temporary

Job Description

Please be aware this is a Long Term Temporary post to start ASAP.

The following is essential criteria:

- Degree in Social Work
- SWE Registration

Collection considerations

Factor	Why care?
Web-scraping	Potentially misleading estimates
Data validation	Unusable data Missing data
Variable categorisation	Lack of transparency of methods
Data updates	Misinterpreted estimates - durations
Dataset combination	Missing data
Multi-source – duplicates	Potentially misleading estimates Accounting for market whims
Multi-source – classification	Inconsistent results Microdata version control

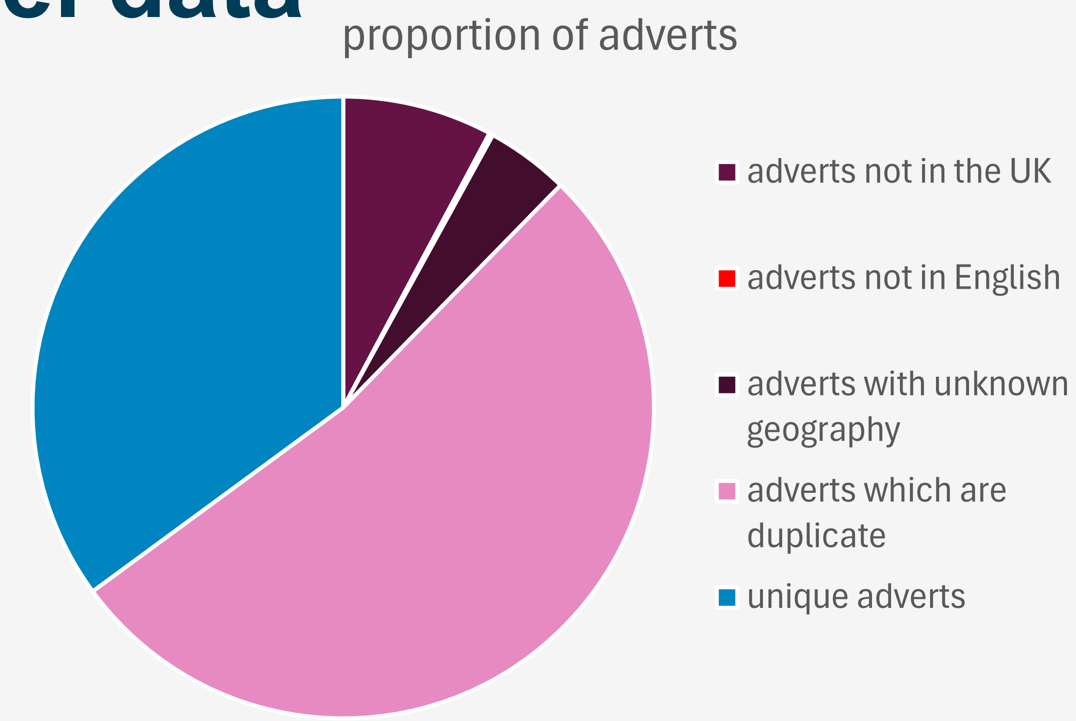
Using the data

Overview of data processing

- Data validation of raw data received
- Pre-process to get to right population
- Extract variables from text
- Aggregate
- Sense-check

Pre-processing Textkernel data

- Removing irrelevant adverts
- Deduplicating



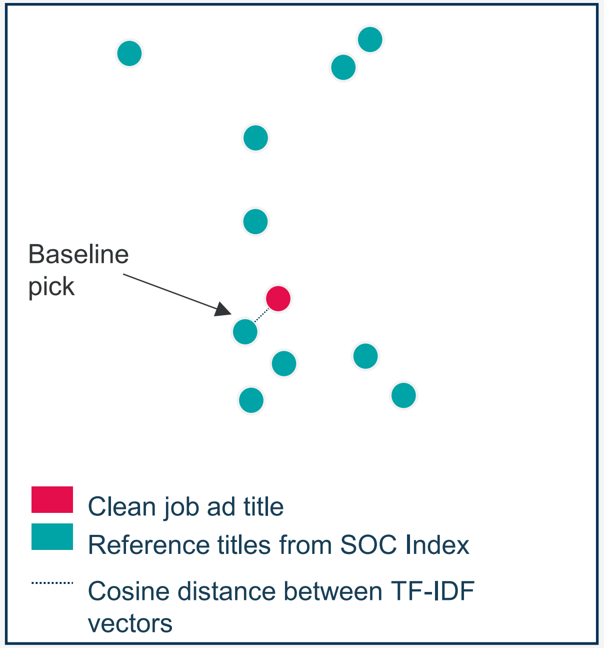
- Time-specific web-scraping issues -> imputing expiry dates
- Optimising efficiency for short runs

SOC Allocation bespoke method

- Cleaning job titles
- Convert each title to character n-gram
- Cosine similarity to each SOC index entry

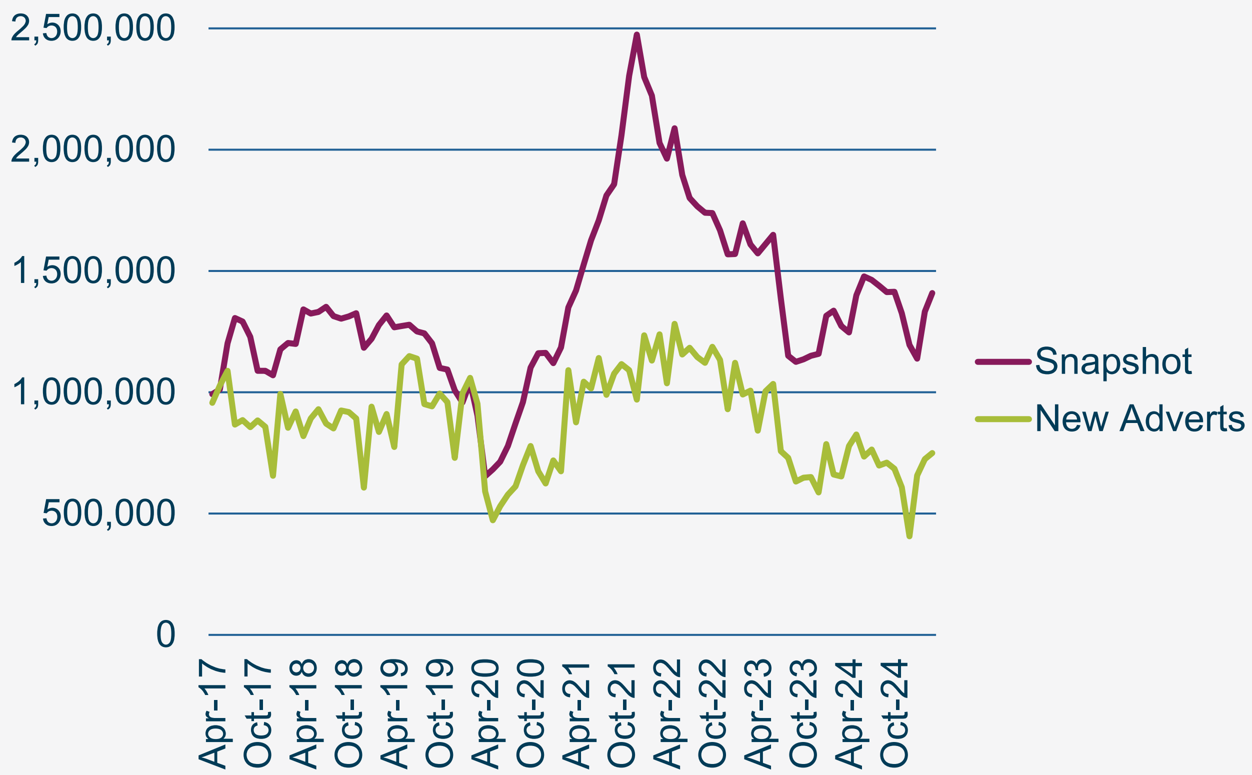
(~30,000 titles)

- Similarity score
- Evaluated accuracy against manual labelled (manual disagreements)


Baseline pick
🟥 Clean job ad title
🟩 Reference titles from SOC Index
Cosine distance between TF-IDF vectors

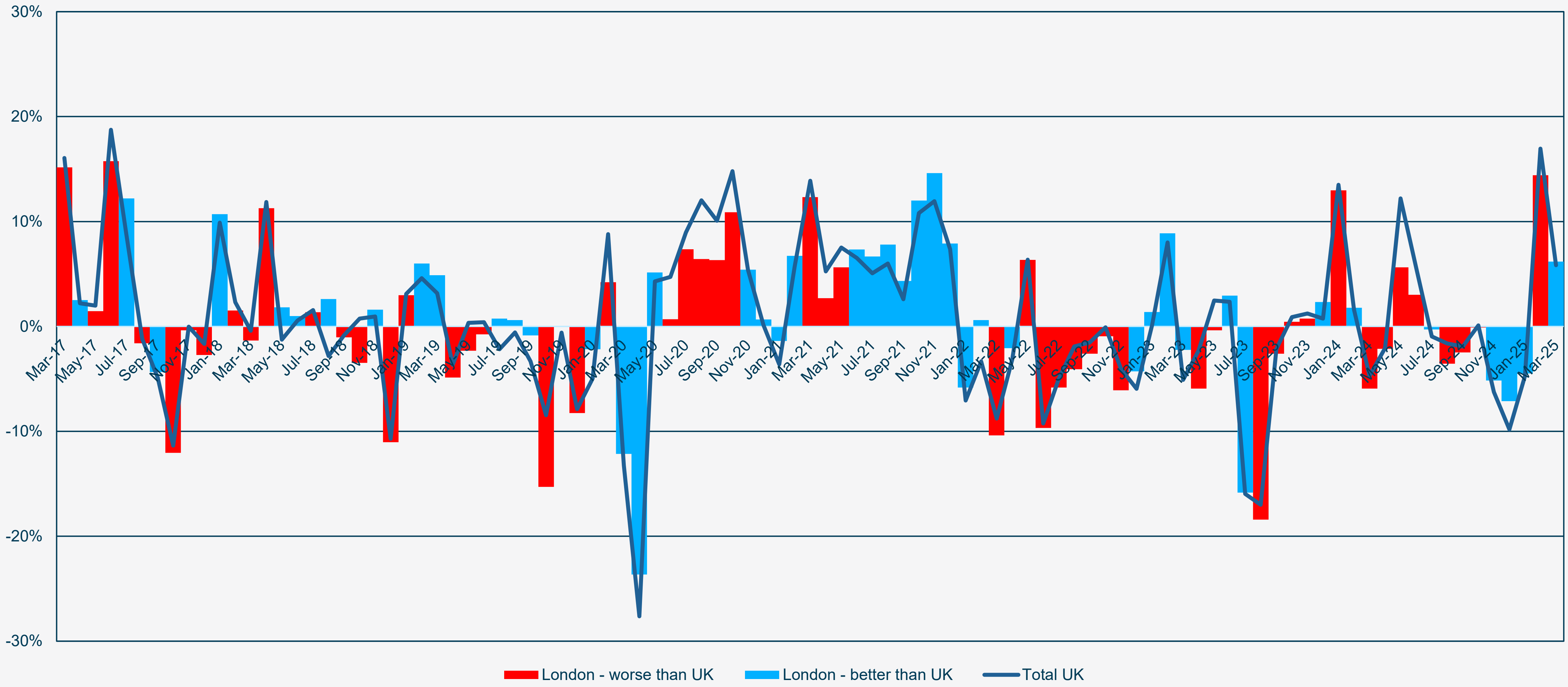
Aggregating the data

- How to summarise
- Frequency considerations
- Disaggregations
- Revisions



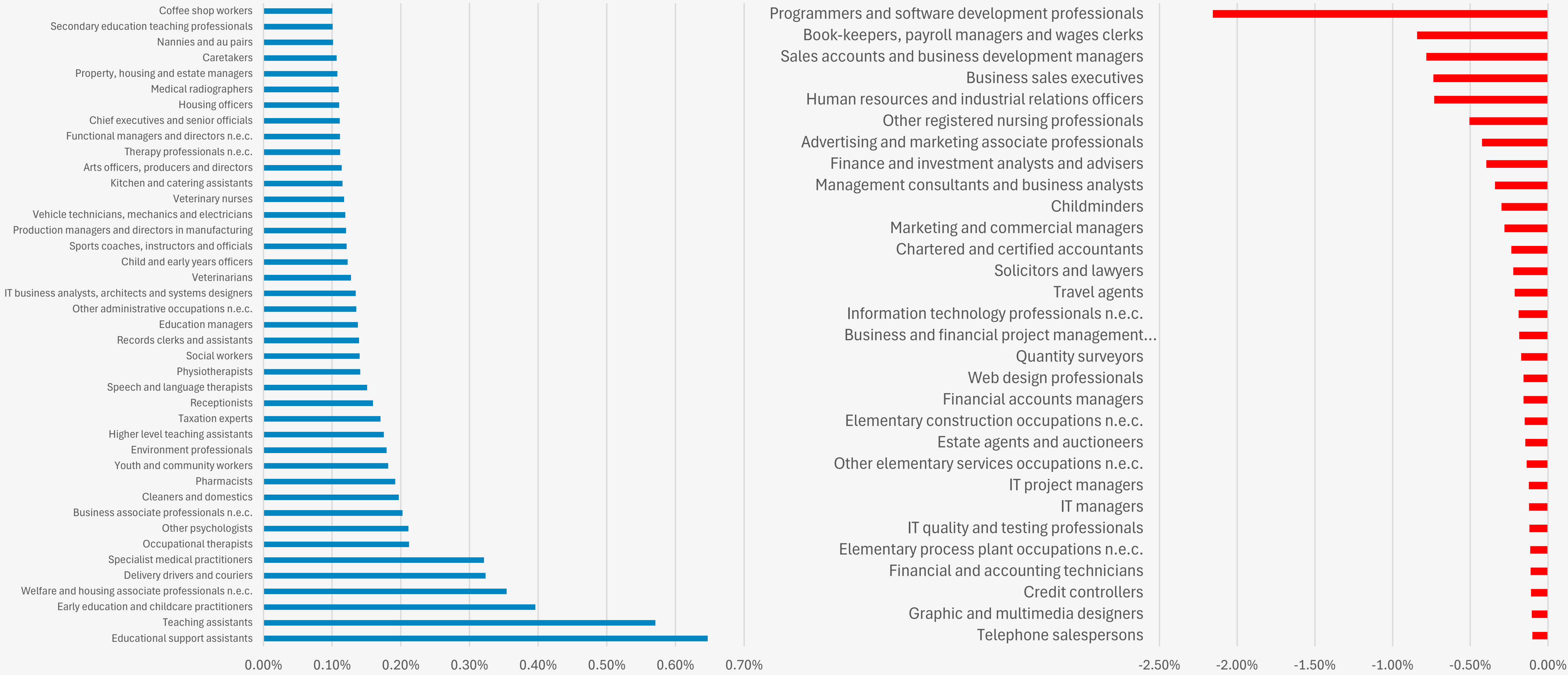
Benefits - analysing the data

London trends



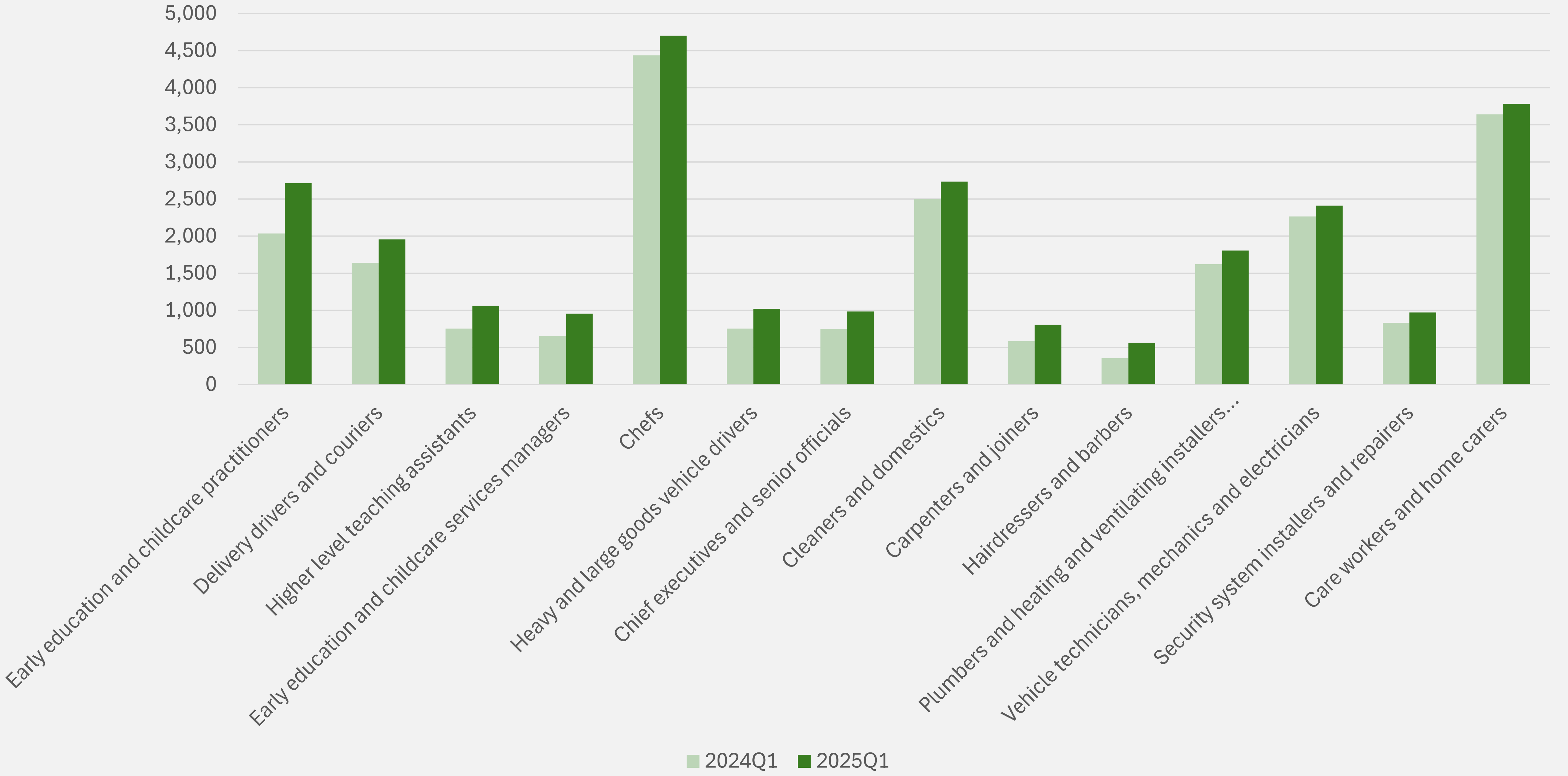
Office for National Statistics

Occupational trends in London since 2017 to 2025

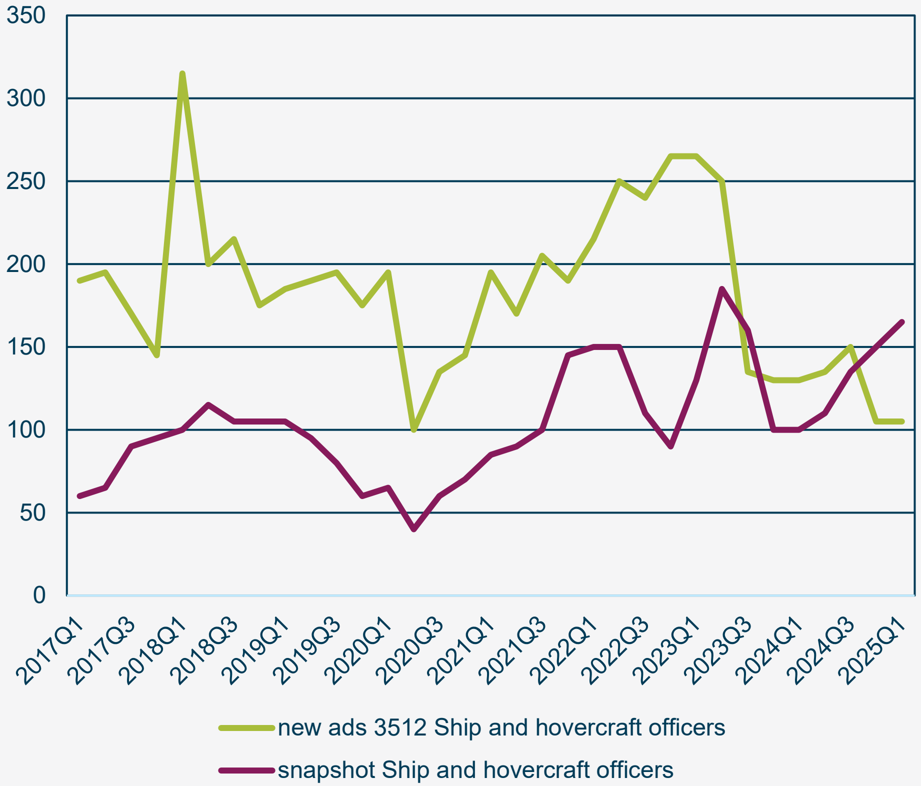


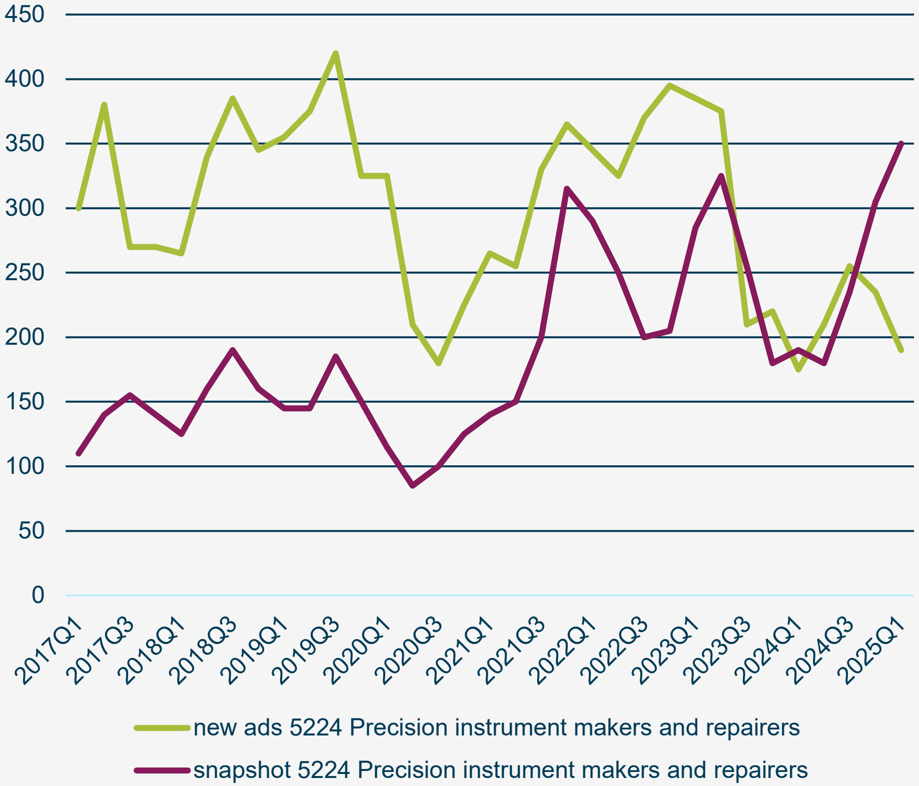
Office for National Statistics

Recent occupational trends in London



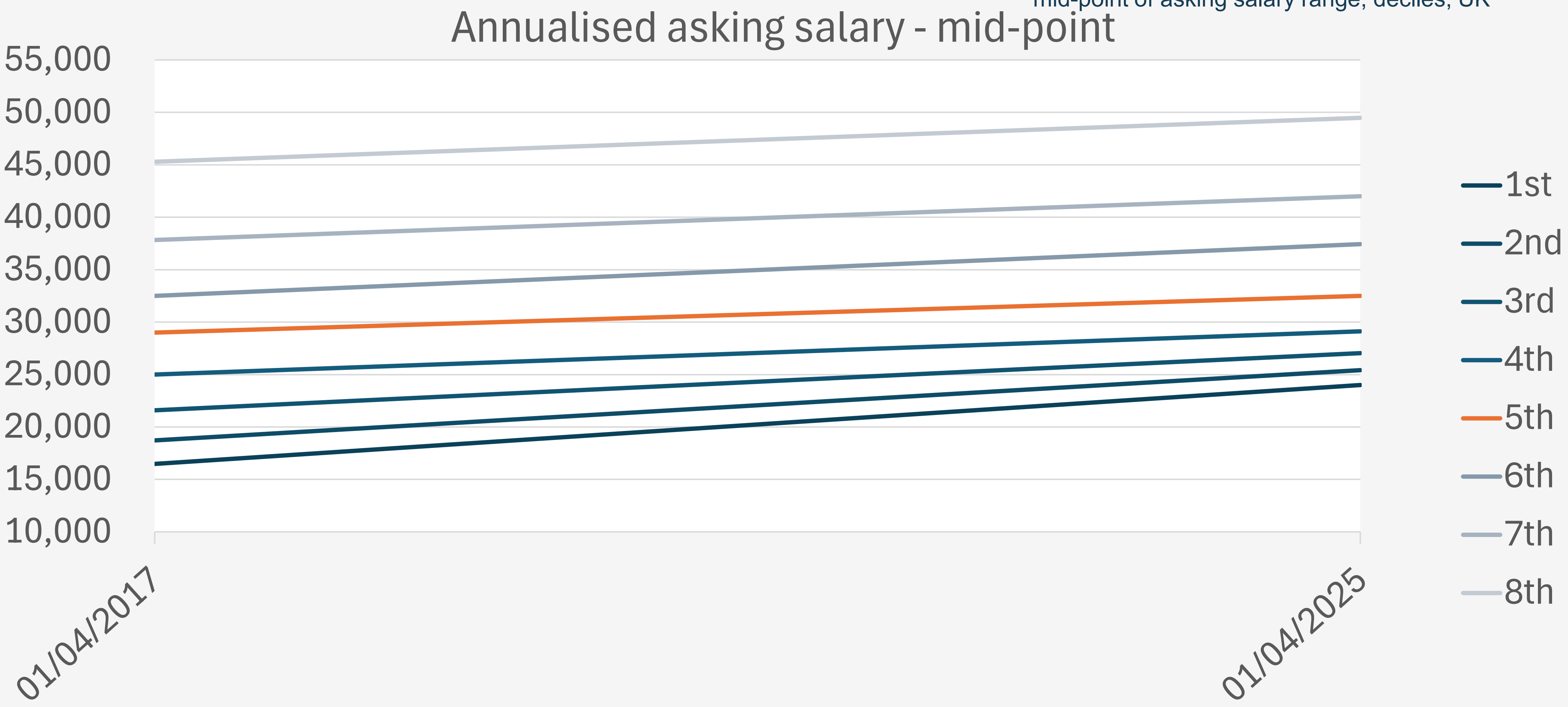
Potential occupations being hard to fill in London





Tracking UK's asking salaries and their distribution

Annual asking salary for all ads live in the month, mid-point of asking salary range, deciles, UK



Official Sensitive – Not Yet Published

Conclusions and future work

Takeaways for commercial data

- Know the data
- Traditional NLP methods still valuable for classification, especially for processing at scale

ONS next steps with job advert data:

- Continue user feedback and engagement
- Planned SOC improvements, further metrics
- Outputs by skills, data sharing, representability

Any questions?

Thank you

Get in touch: economic.wellbeing@ons.gov.uk

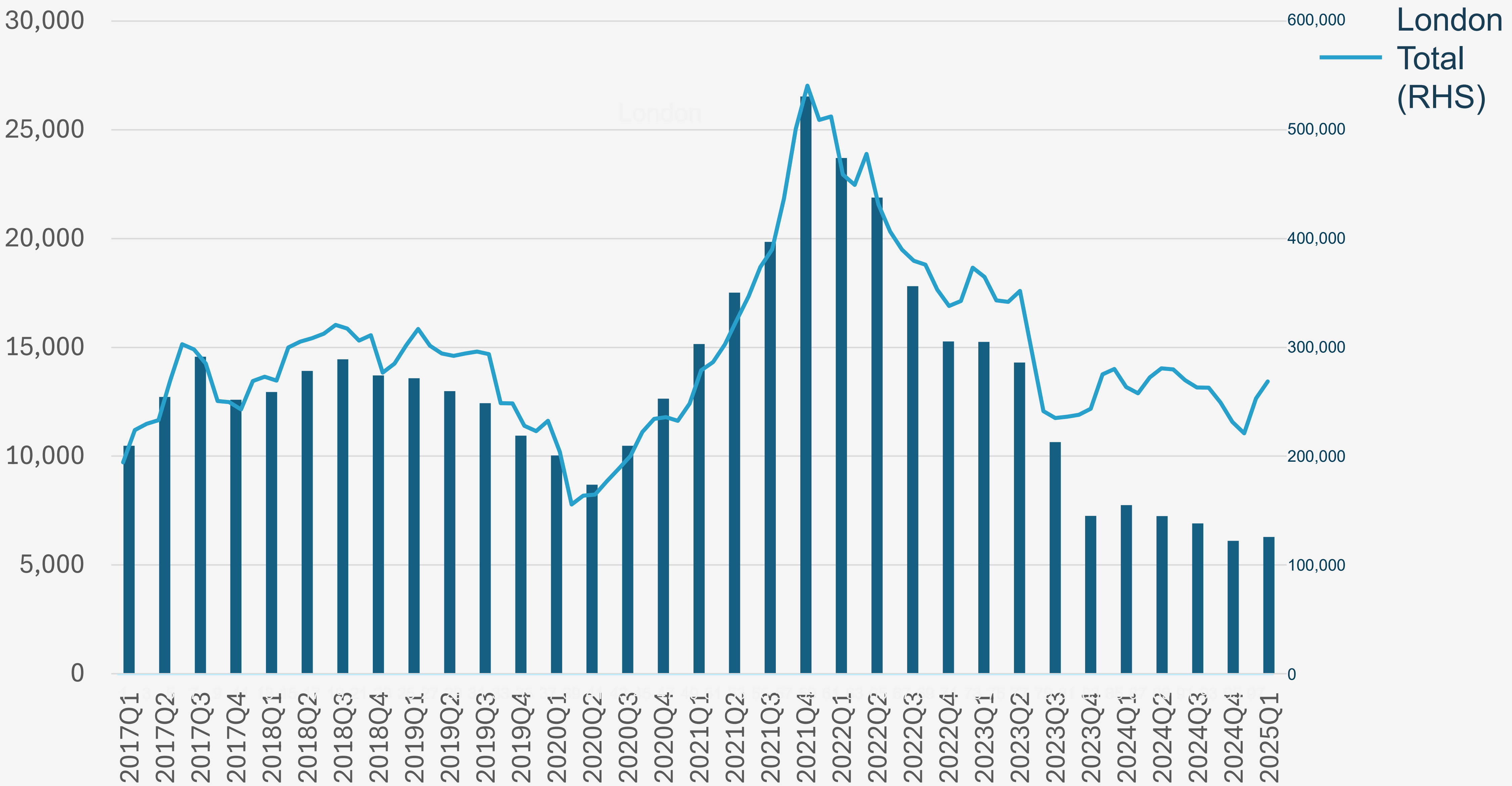
Annex – further SOC method thinking

SOC Allocation – further planned improvements

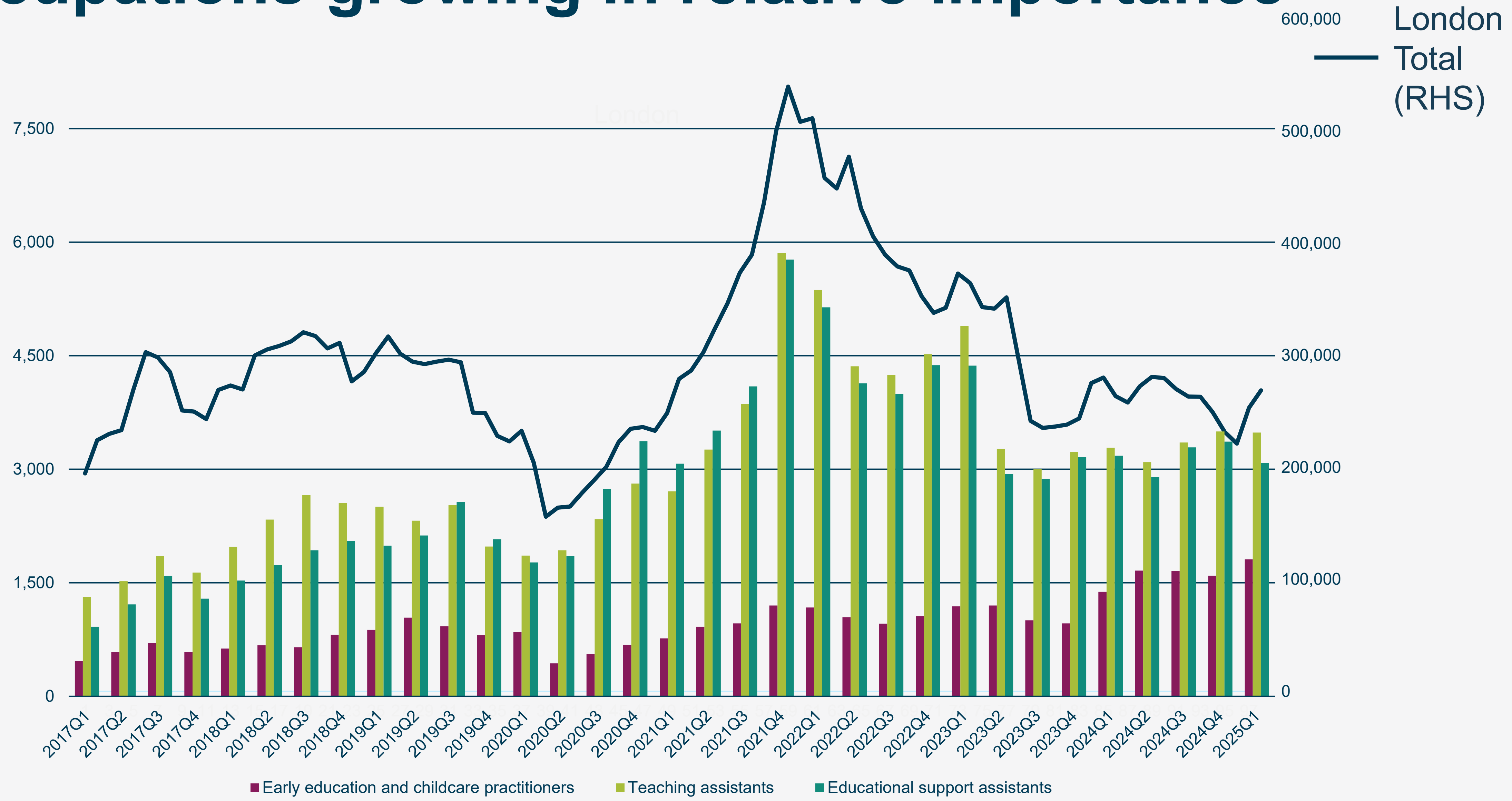
- Use high confidence baseline as training data
- Weak supervised learning on job descriptions – individual 4-digit SOC node models, combining n-grams into one overall vocabulary
- One vs rest logistic regression model
- Ensemble method for overall combination of many SOCs
- Potential for Large Language Model embeddings to make previous research ‘old school’ – BUT needs to be production-usable and cheap

Annex – more data to hand if needed

Programmers and software development professionals



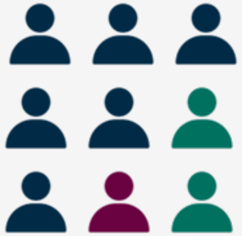
Top occupations growing in relative importance



Annex – more explanations to methods if needed

Accuracy assessment

- Manually labelling – triple coded



28% Unanimous SOC (easy ads, models should get these right)

70% Majority SOC (more robust given coder disagreement)

100% Plausible SOC (at least one match to a human coder)

- Evaluated at 1-digit to 4-digit

Performance comparison

Metric	Precision against Plausible SOC		
Method	Current method	Supervised Learning (preliminary)	Combined (preliminary)
Coverage	**100%	**100%	70%
1-digit SOC (major group)	80%	*80%	*90%
2-digit SOC (sub-major group)	75%	*70%	*90%
3-digit SOC (minor group)	70%	*65%	*85%
4-digit SOC (unit group)	65%	*65%	*80%

A note on industry

- Some online job advert aggregator sources have “SIC”
- Methods tend to be based on identifying the company, then matching to an industry
  - Job adverts don’t always have useful company info (or just recruiter company)
  - Matching company to industry is incomplete (it should be the reporting unit)
- Hence – currently industry data from job adverts not recommended

Annex – further elaborations on user requests

Office for National Statistics

Topics and metrics of interest

Topic: / Metric:	New Adverts	Snapshots	Salaries	Durations	Duplicates
Demand	✕	✕			
Shortages		✕	✕	✕	✕
Inequalities		✕	✕		