TASK

Zbiór danych (indexing.csv) prezentuje stan zaindeksowania poszczególnych domen w wyszukiwarce Google.

Ważne: dataset zawiera informacje o zaindeksowanych podstronach (*status* = *{Valid, Warning}*), jak i niezaindeksowanych: *status* = *{Excluded, Error}*.

Kolumna *status_details* przedstawiaja również **przyczynę niezaindeksowania** (jak np. “Alternate page with proper canonical tag” lub “Blocked by robots.txt”)

Celem zadania jest przedstawienie insightów ze zbioru danych. Istotne jest graficzne przedstawienie poszczególnych etapów analizy.

Przykładowymi insightami mogłyby być: lista najpopularniejszych problemów z zaindeksowaniem, opis występujących trendów, występujące korelacje.

Po skończonym zadaniu prosimy o **podsumowanie w języku angielskim.**

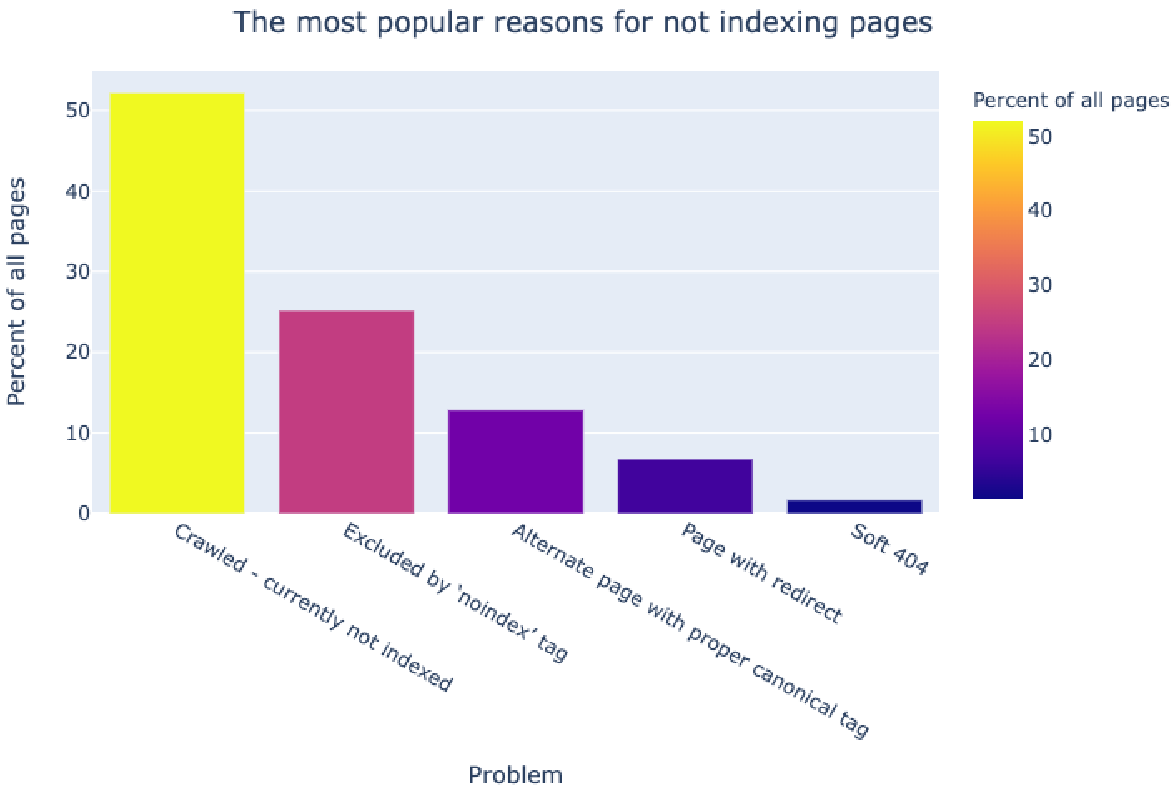
Fragment danych:

date	domain	status	status_details	pages
2021-12-06	domain1	Excluded	Alternate page with proper canonical tag	3081440
2021-12-06	domain1	Excluded	Blocked by robots.txt	43626

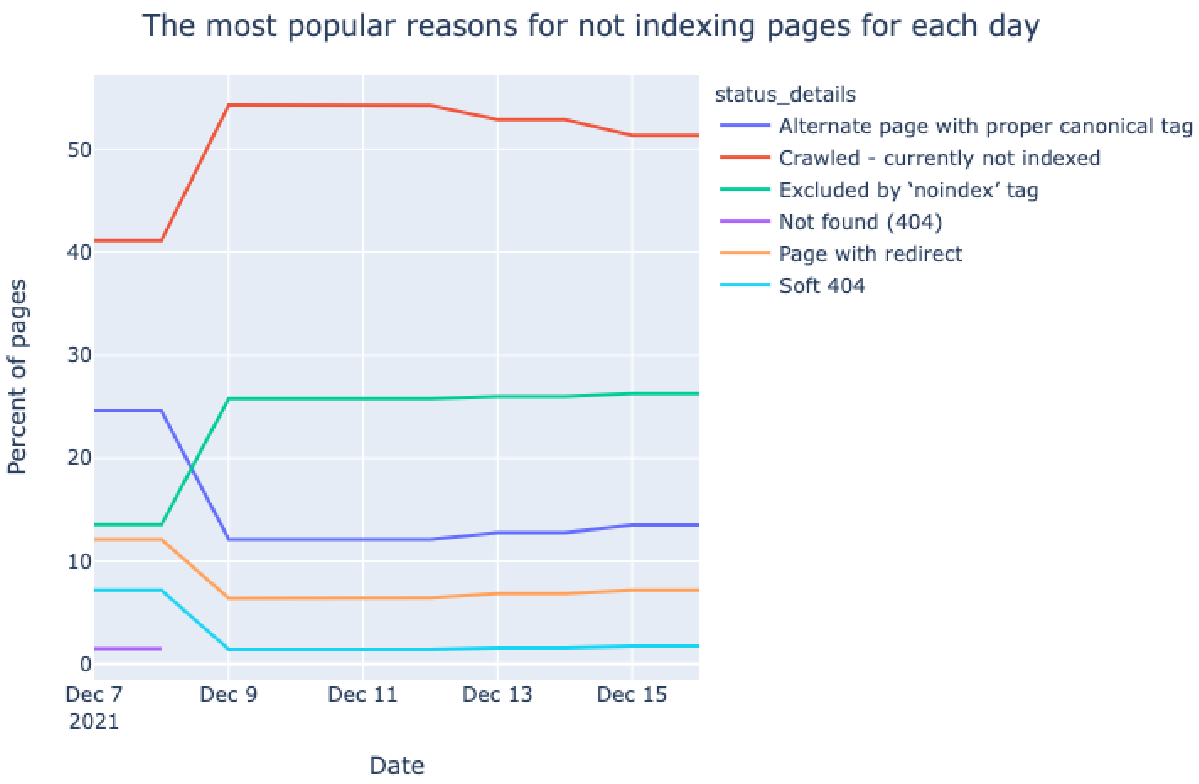
DATASET ANALYSIS

In order to analyze the data, I asked myself the following questions:

- What are the most popular reasons for not indexing the pages?
- Are there any trends over the time?
- How does the size of a domain affect the reasons for not indexing the pages?
- Are small domains more likely to be indexed?



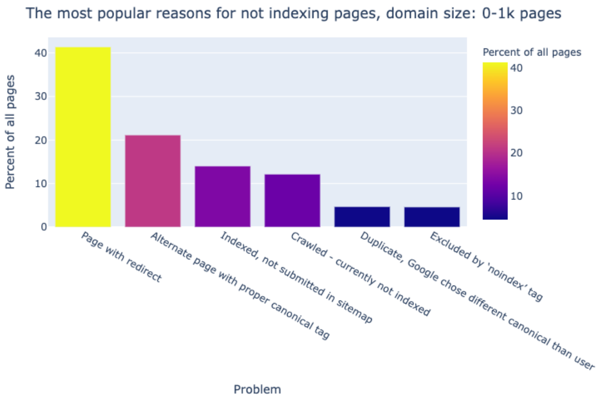
- Over half of the pages are not indexed, because they are still being **crawled**
- 1 out of 4 pages is excluded, because of **'noindex' tag**
- The TOP5 ends with status details like: **Alternate page with proper canonical tag**, **Page with redirect**, **Soft 404**

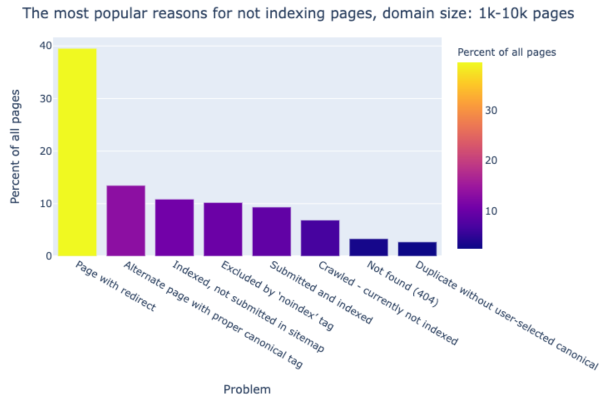


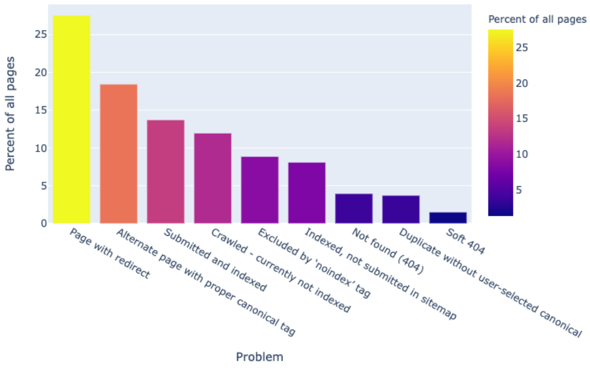
There is a reshuffle on Dec 9, due to first data coming from the biggest domains. **"Crawled - currently not indexed"** has had a downward trend since Dec 12 and this is the reason why all other *status_details* are slightly increasing.

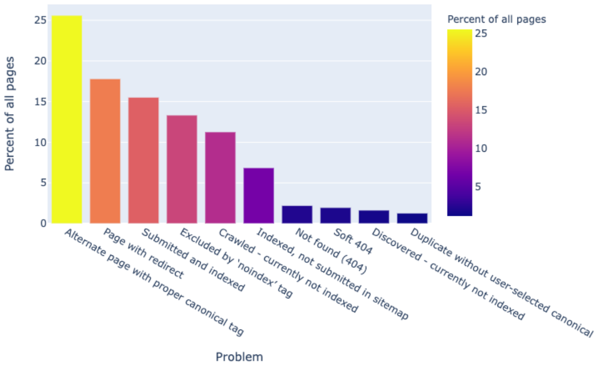
Grouping domains by size

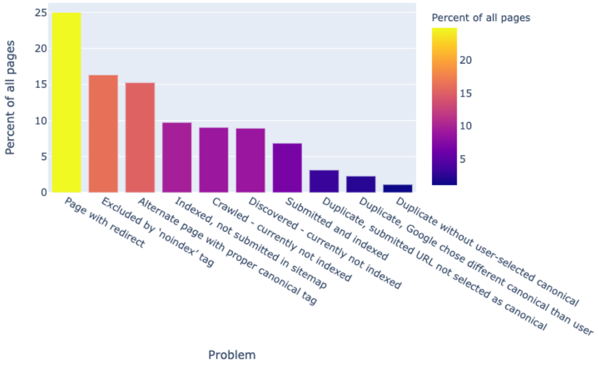
There are huge differences between sizes of domains and I thought it might be useful to categorize and split the data. The factor by which I will categorize the domains is the sum of pages per day.

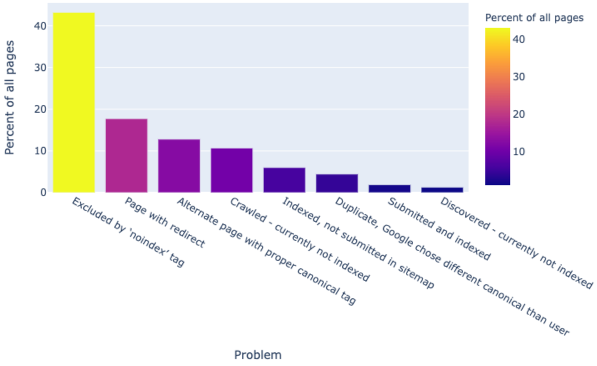


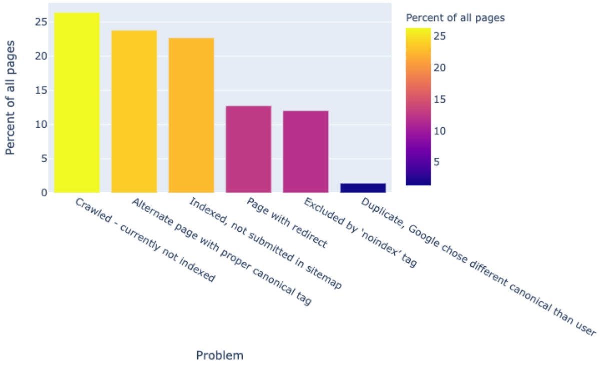


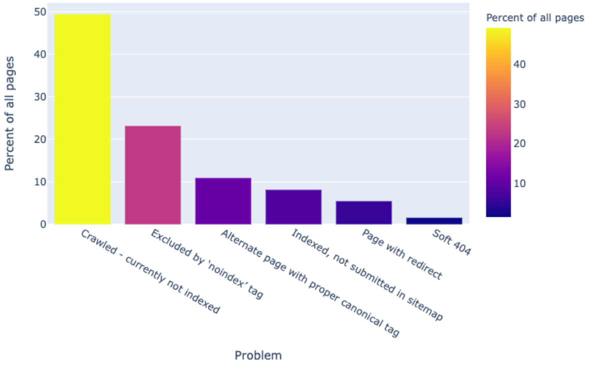

The most popular reasons for not indexing pages, domain size: 10k-100k pages


The most popular reasons for not indexing pages, domain size: 100k-1m pages


The most popular reasons for not indexing pages, domain size: 1m-10m pages


The most popular reasons for not indexing pages, domain size: 10m-100m pages


The most popular reasons for not indexing pages, domain size: 100m-1b pages


The most popular reasons for not indexing pages, domain size: 1b+ pages

- "**Page with redirect**" refers only to 6.8% of all pages, however, it is the most popular reason for not indexing pages in 4 out of 8 categories (mostly smaller domains). On the other hand, "**Page with redirect**" is only a small fraction of *status_details* in the biggest domains.
- "**Crawled - currently not indexed**" - the dominant reason for not indexing pages for the biggest domains, would have a hard time getting into the top 3 problems of the smaller domains

- "**Alternate page with proper canonical tag**" - this problem occurs in all categories with similar percentage of pages (10-25%)
- "**Excluded by 'noindex' tag**" - appears more frequently in the medium size domains, it is the second most popular *status_detail* overally, because of making up 23% of all *status_details* in the biggest domains category