Starting with the google search: **Vancouver Development Permit Data**

**Found** [opendatasoft](https://data.opendatasoft.com/explore/dataset/issued-building-permits%40vancouver/export/)

**Has 35444 records of Permits**

**Satisfied the conditions:**

* The data is refreshed daily
* It contains the Permit number
* It contain the Permit's description
* It contain the Permit's applicant name
* It can scraped or accessed via an api

Needed to find a website that displayed a permit’s status:

Found [PermitSearchByNumber](https://plposweb.vancouver.ca/Public/Default.aspx?PossePresentation=PermitSearchByNumber)

**Satisfied the conditions:**

* It contain the Permit's status

Well these 2 had to be put together. So, I wrote a script that can interact with the website that gives the status using Selenium. It takes an average of 2.41 seconds for 50 permits. I have added the python script and the xlsx with Permit numbers and their status below.



The status also shows if a permit’s job has been completed or if the permit has expired. So while the first scraping is going to be time consuming (35444 records. ~35 hours), the scraping from the next iteration is narrowed down since most of the records are either Expired or Completed. (58% of the most recent permits have either been completed or have expired. This percentage increases with older permits). This cuts down the processing time immensely.

Points to improve the code (Things I couldn’t handle under the time constraints):

1. Source the opendatasoft data from API instead of Selenium Automation(The formatting has newline and delimiter issues)
2. Include Ongoing and archived applications. <https://www.shapeyourcity.ca/development>
3. Scrape the most recent status. Some have multiple records on single permit number. Analyse all the records of that single permit number and pick the most recent.