1. Non English terms - Thai language



The Thai characters simply mean LCD/OLED, TV, inches etc. and won’t provide useful information to identify the KATABAN name. Hence translation is unnecessary.

Solution: Modify the web scraping code to better handle the situation when have Thai characters to scrape more useful information

2. Suffix like AKXXT, GXXP eg. QA75QN85CAKXXT, Samsung OLED QA77S95CAGXXP

Solution: in python code, add a cleaning step to remove suffix before matching

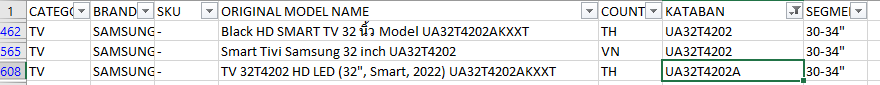
3. Character encoding issue

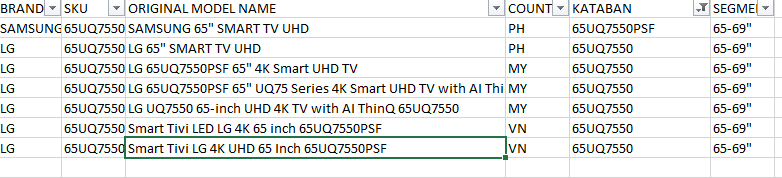


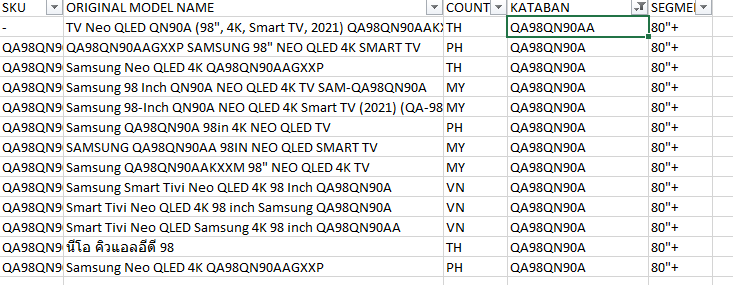


There is a space at the end of the KATABAN name

Solution: manually clean up

4. KATABAN inconsistency





Solution: manually clean the KATABAN name