**Open Source Audit Data Analysis for the last 8 years**

The open source audit service that my company provided to our industry leading customers has left report data of massive volume for the last 8 years. I came up with an idea that I could elicit various insights from this big data and use them for sales and marketing purpose. It was used for sales & marketing materials. On top of that, I was able to give a speech on the trend of open source in the webinar that my company hosted.

**Speculation**

How many open sources domestic companies have used?

What kind of open sources and licenses are used in domestic companies the most?

**Conclusion**

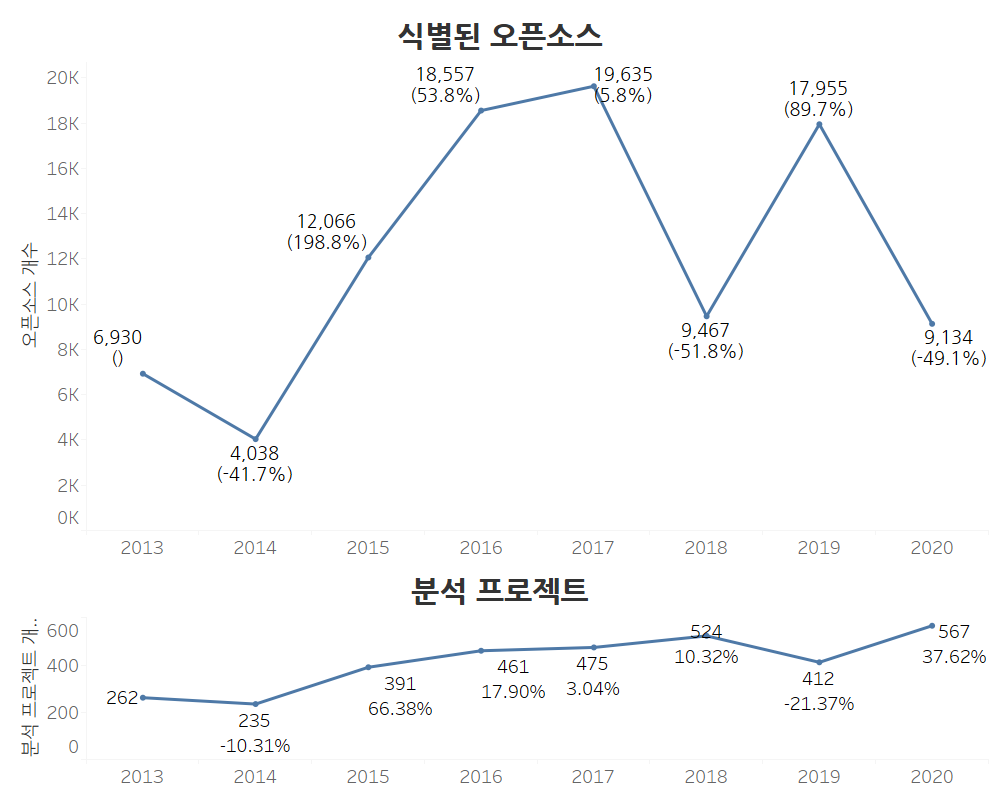
* The rate of open source use are getting bigger.
* Jquery, bootstrap, and Apache project are used the most.
* Permissive licenses are mostly used in companies, especially Apache License, MIT, and BSD.
* The trend of using copyleft keeps low, but it is important to know that we could be cautious due to the risk of open source license violation.
* Main licenses that cause license violation are GPL 2.0, and LGPL 2.1 amongst copyleft licenses.
* Amongst copyleft licenses that cause license risk, there is decrease in the use of GPL 2.0 and LGPL 2.1, but the use of more cautious licenses like GPL 3.0, LGPL 3.0, Affero GPL, and CCBY stay still.

**Workflow**

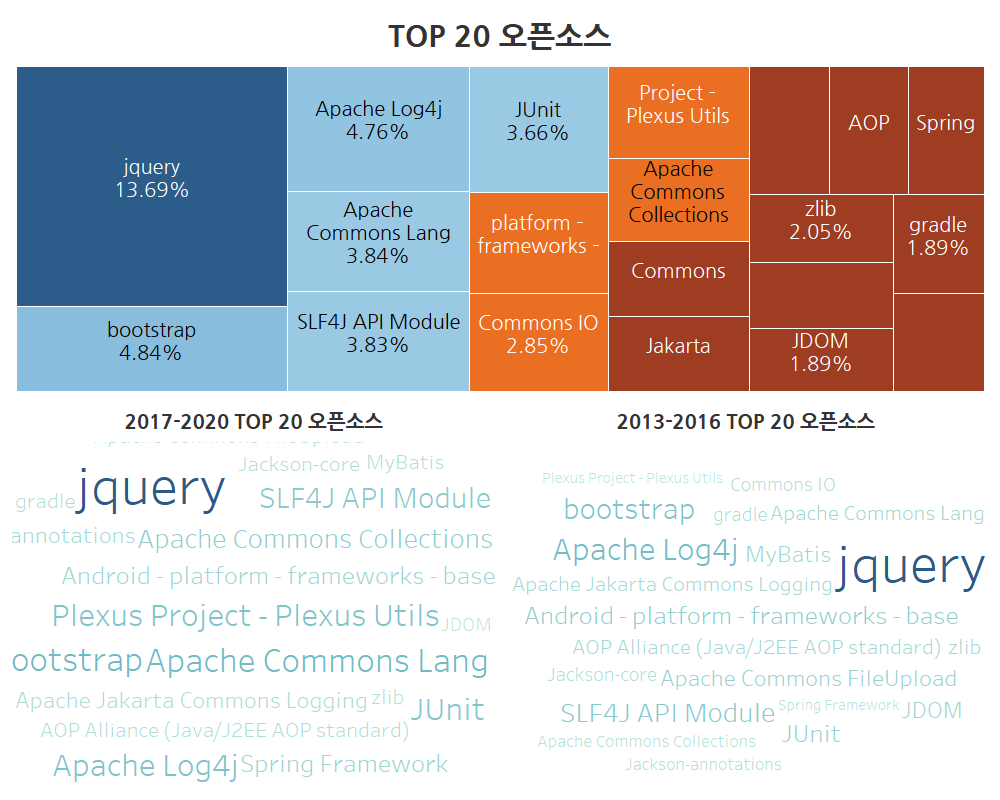
* Designing -> Cleaning -> Collecting -> Pre-processing -> Analysis -> Visualisation
* Designing: Designed data schema on what to collect
* Cleaning: Supplemented more than 3,300 audit reports with missing values
* Collecting: Created 36 automatically collecting data modules for various report templates by python
* Pre-processing : Filled missing values and unified different names for same data
* Analysis & Visualisation: Designed scenarios for sales and marketing use and visualized big data through Taebleau

**Applied Skills**

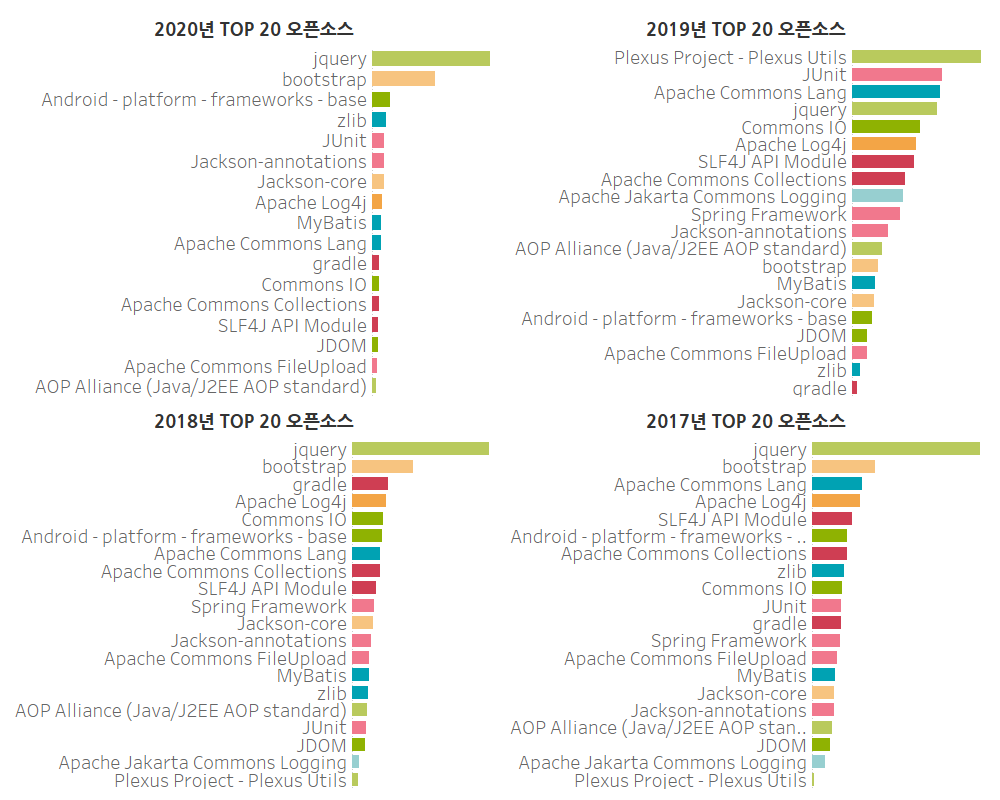
* Collecting: Python(pandas, numpy, tqdm, os, shutil
* Pre-processing: Python(pandas, numpy, matplotlib)
* Analysis and Visualisation: Taebleau



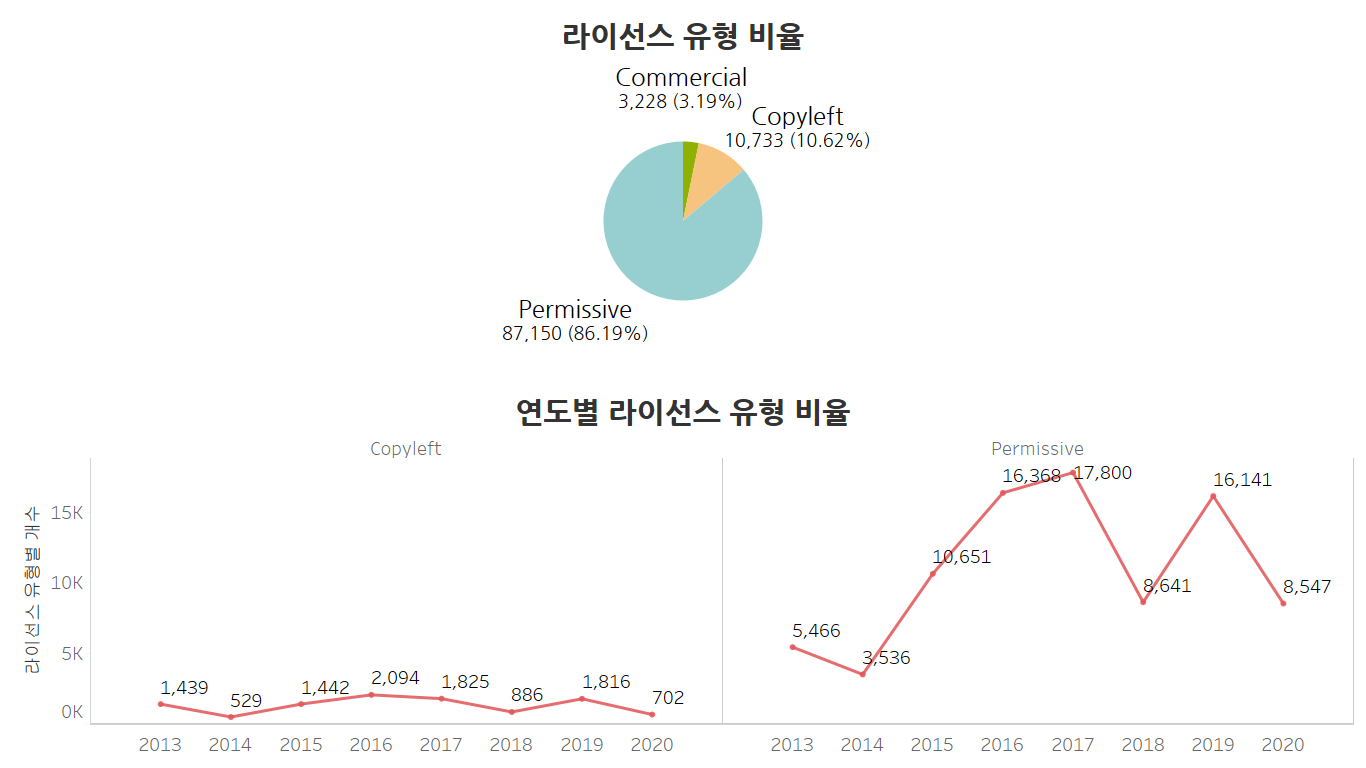
**The status of analysed projects and identified open source**



**Mostly used open source in companies**

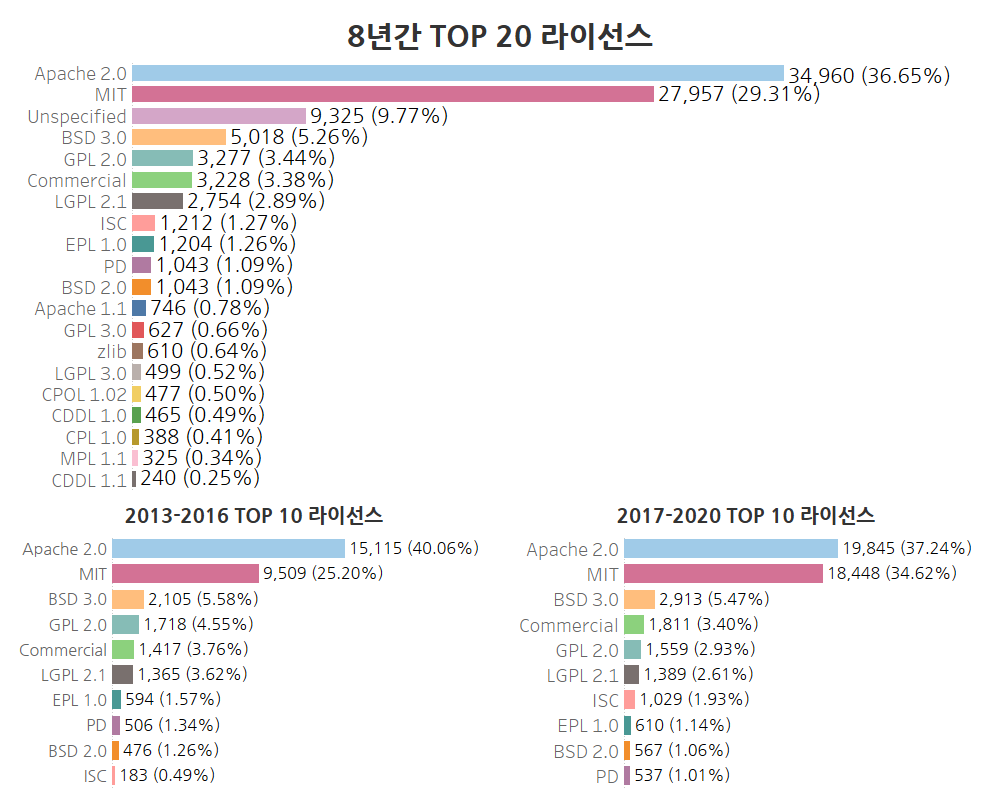


**Mostly used open source in companies by year**



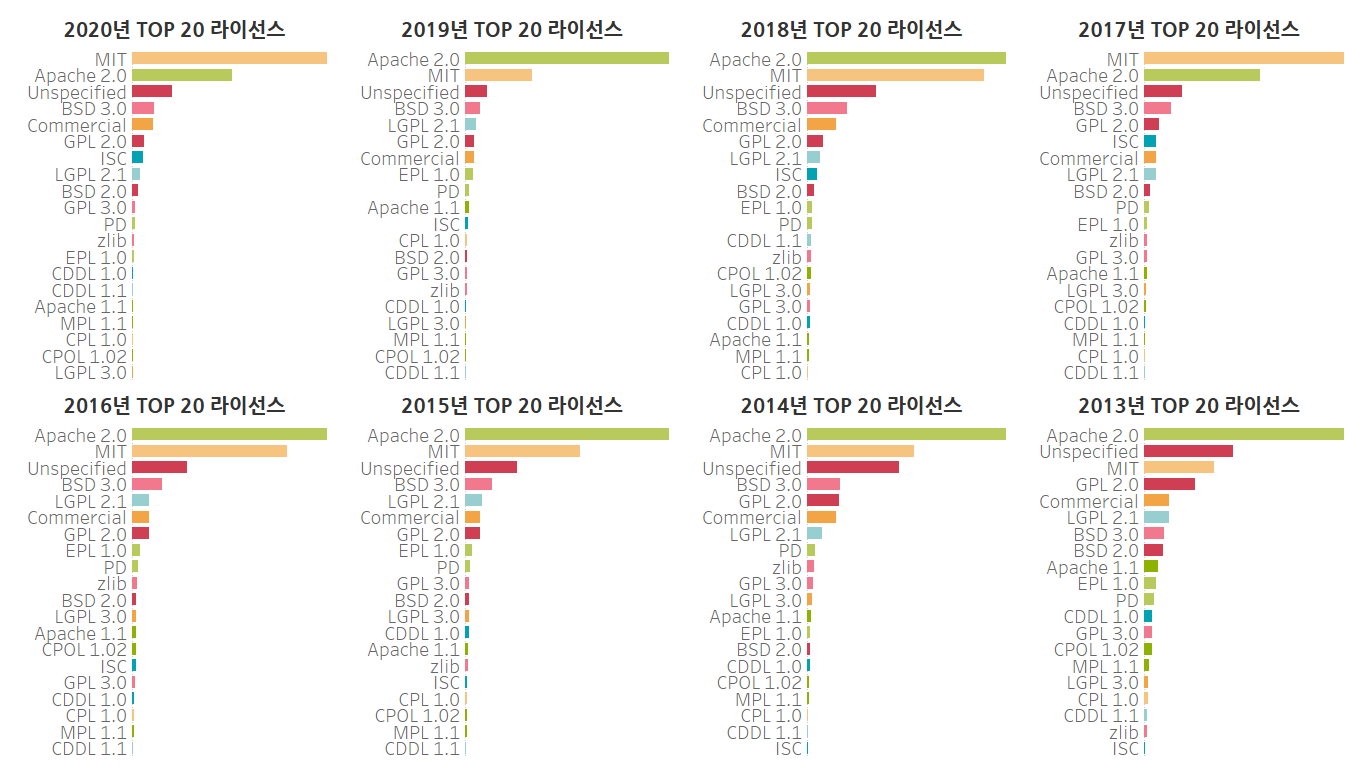
**Ratio of license type**

**Ratio of license type by year**

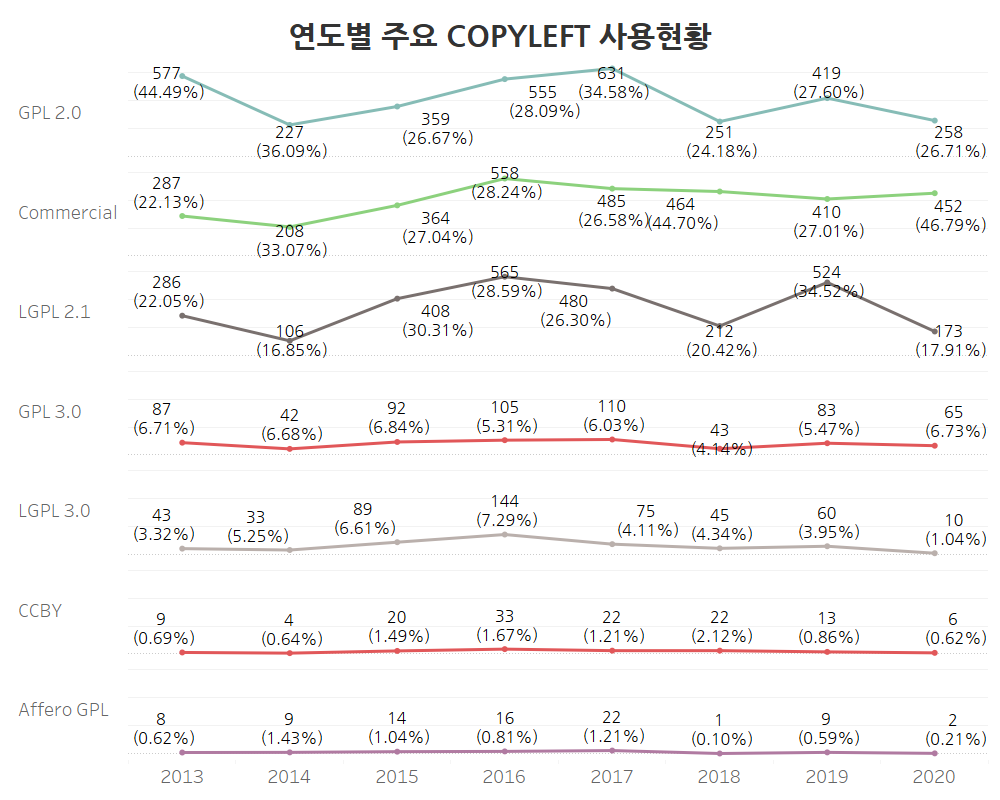


**Top 20 open source licenses for 8 years**

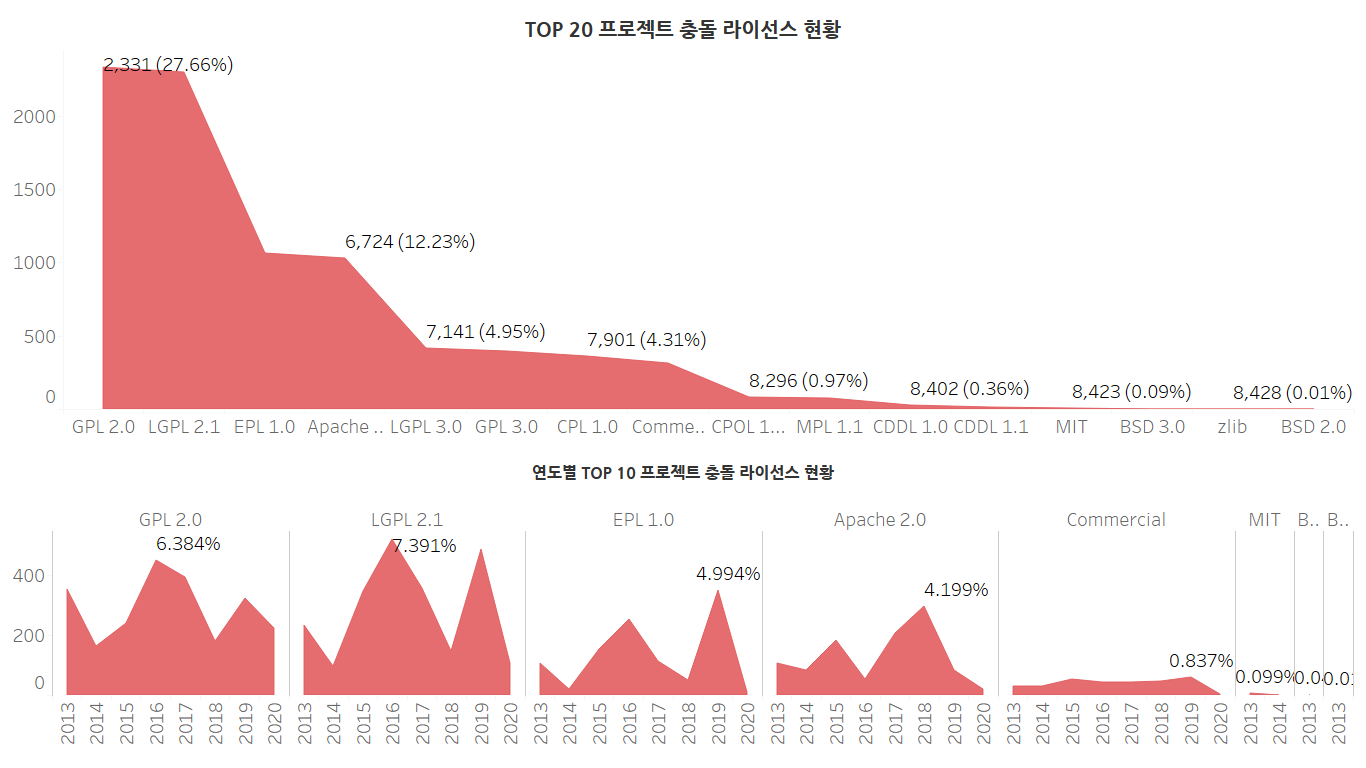
**Comparison of top 10 open source licenses between 2013 – 2016 & 2017-2020**



**Top 20 open source licenses by year**



**Main copyleft licenses by year**



**Top 20 copyleft licenses causing license risk.**

**Top 10 copyleft licenses causing license risk by year**