Project 1 Data Mining, Knowledge Discovery
MAT 259
Jingxuan Cao

Introduction:

I am interested in discovering the trending jobs in the last few years. My plan is to search online to get the list of popular jobs in the past few years and then I will collect the checkout data of these jobs to see the result. Additionally, considering the financial crisis happened in 2008, I will also try to find out whether the financial crisis influenced these jobs. When I search online, one website give me a long list of jobs and I summarize them to: law related(340-349), finance and economy related(330-339), computer related(000-009), engineering(620-629), medical scientist & dentist(610-619). I will collect data by Dewey Decimal Classification in the database.

Query:

```
SELECT
        YEAR(cout) AS years,
        sum(if(deweyClass >= 330 and deweyClass < 340 , 1, NULL)) AS 'Econ',
        sum(if(deweyClass >= 340 and deweyClass < 350 , 1, NULL)) AS 'Law'
from spl_2016.outraw
where  (deweyClass >= 330 and deweyClass < 350) and YEAR(cout) BETWEEN 2006 and
2019
group by year(cout)
order by year(cout)
```

```
SELECT
        YEAR(cout) AS years,
        sum(if(deweyClass >= 610 and deweyClass < 620 , 1, NULL)) AS 'Medical',
        sum(if(deweyClass >= 620 and deweyClass < 630 , 1, NULL)) AS 'Engineering'
from spl_2016.outraw
where  (deweyClass >= 610 and deweyClass < 630) and YEAR(cout) BETWEEN 2006 and
2019
group by year(cout)
order by year(cout)
```

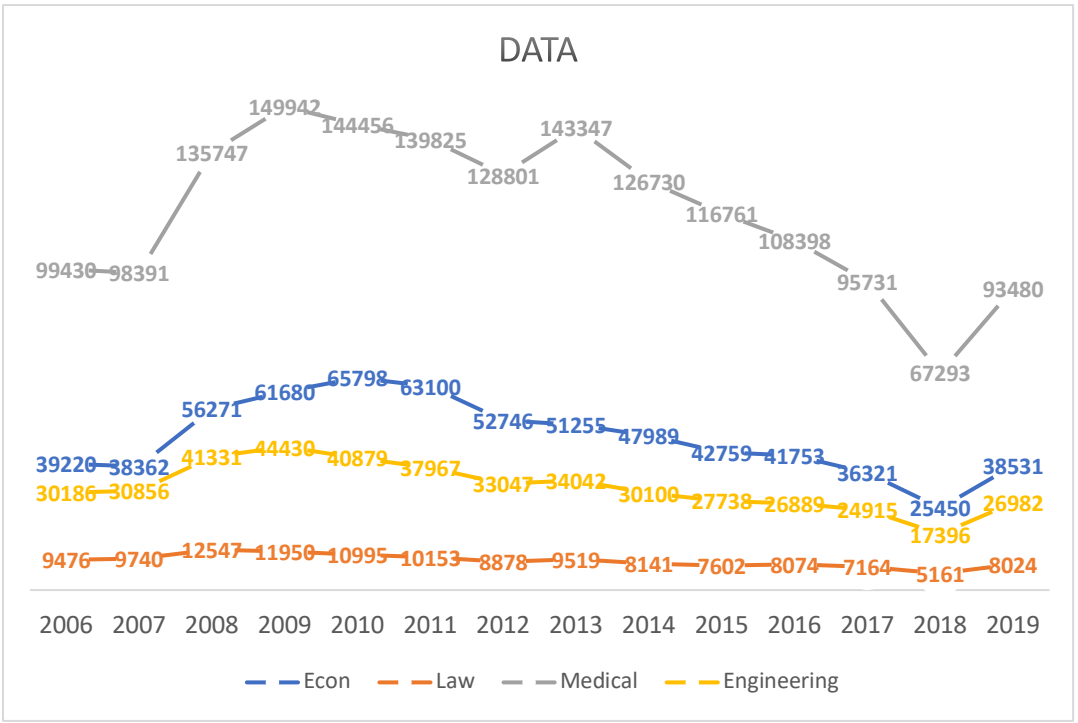
Difficulties:

When I am collecting data, I find out that I am not able to collect data from computer category that it always shows server disconnection. I cannot search for a big range of deweyClass, so I have to separate the searching for Econ, Law, Medical, Engineering.

Data:

years	Econ	Law	Medical	Engineering
2006	39220	9476	99430	30186
2007	38362	9740	98391	30856
2008	56271	12547	135747	41331
2009	61680	11950	149942	44430
2010	65798	10995	144456	40879
2011	63100	10153	139825	37967
2012	52746	8878	128801	33047
2013	51255	9519	143347	34042
2014	47989	8141	126730	30100
2015	42759	7602	116761	27738
2016	41753	8074	108398	26889
2017	36321	7164	95731	24915
2018	25450	5161	67293	17396
2019	38531	8024	93480	26982

Visualizing the data to chart:



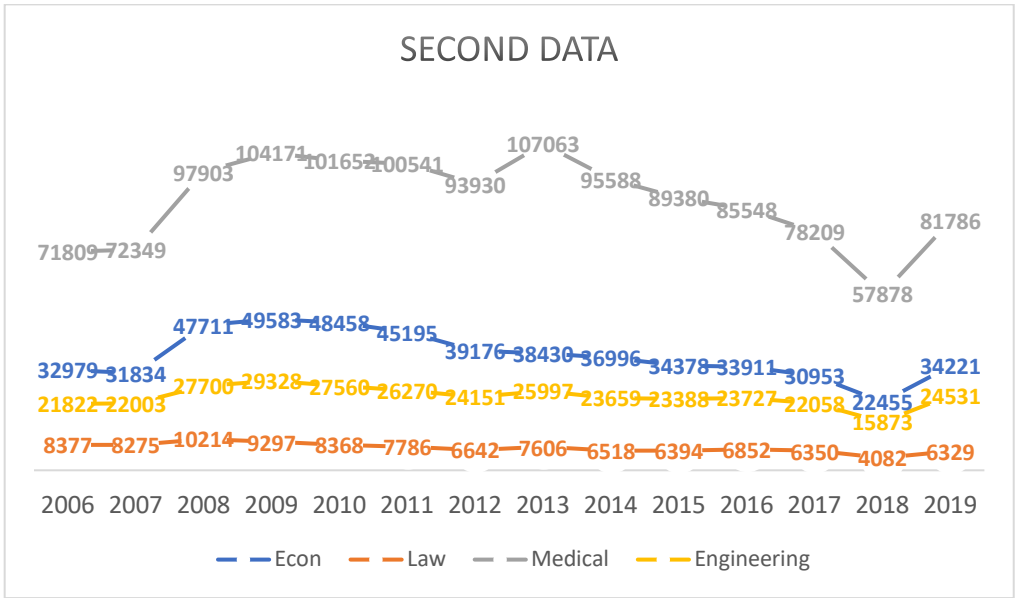
The second attempt:

I am wondering if there are some elements influence the data of each area, because I suspect that there are a lot of other media products in medical for instance DVD. Therefore, I decide to limit the counting of each area to book.

Second data:

years	Econ	Law	Medical	Engineering
2006	32979	8377	71809	21822
2007	31834	8275	72349	22003
2008	47711	10214	97903	27700
2009	49583	9297	104171	29328
2010	48458	8368	101652	27560
2011	45195	7786	100541	26270
2012	39176	6642	93930	24151
2013	38430	7606	107063	25997
2014	36996	6518	95588	23659
2015	34378	6394	89380	23388
2016	33911	6852	85548	23727
2017	30953	6350	78209	22058
2018	22455	4082	57878	15873
2019	34221	6329	81786	24531

Chart:



The different media products in Medical:

Since I know there are many media products connected with medical, I am curious about how many DVD, CD are related with it.

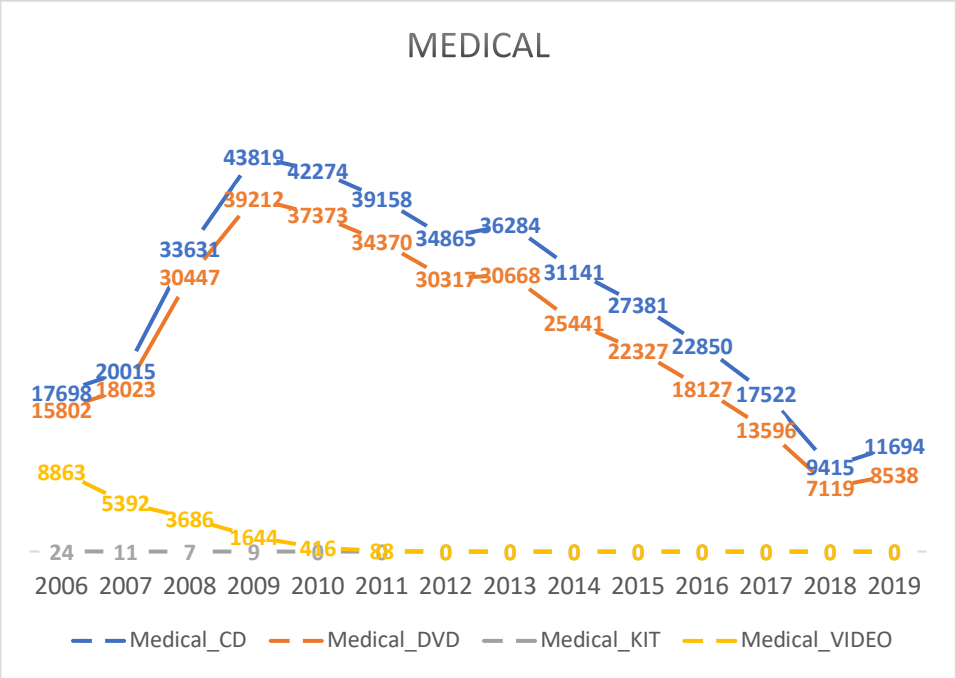
QUERY:

```
SELECT
YEAR(cout) AS years,
        sum(if(deweyClass >= 610 and deweyClass < 620 and itemType like '%cd%', 1, NULL))
AS 'Medical_CD',
        sum(if(deweyClass >= 610 and deweyClass < 620 and itemType like '%dvd%', 1,
NULL)) AS 'Medical_DVD',
        sum(if(deweyClass >= 610 and deweyClass < 620 and itemType like '%kit%', 1, NULL))
AS 'Medical_KIT',
        sum(if(deweyClass >= 610 and deweyClass < 620 and itemType like '%vhs%', 1,
NULL)) AS 'Medical_VIDEO'
from spl_2016.outraw
where  (deweyClass >= 610 and deweyClass < 630) and YEAR(cout) BETWEEN 2006 and
2019
group by year(cout)
order by year(cout)
```

Data:

years	Medical_CD	Medical_DVD	Medical_KIT	Medical_VIDEO
2006	17698	15802	24	8863
2007	20015	18023	11	5392
2008	33631	30447	7	3686
2009	43819	39212	9	1644
2010	42274	37373	NULL	416
2011	39158	34370	NULL	88
2012	34865	30317	NULL	NULL
2013	36284	30668	NULL	NULL
2014	31141	25441	NULL	NULL
2015	27381	22327	NULL	NULL
2016	22850	18127	NULL	NULL
2017	17522	13596	NULL	NULL
2018	9415	7119	NULL	NULL
2019	11694	8538	NULL	NULL

Chart:



Conclusion:

Apparently, the data of each area are decreased by eliminating other media products except books, however, the medical area is decreased the most and also the data I collected at the end prove my suspicion that medical area contains plenty of media products like DVD and CD.

The medical related books have the largest checkout number comparing with the other three categories. But considering that many people care about their health, a large number of these books may not aim for jobs. On the other side, looking at the data of engineering, economy, and law, we can tell that economy books are more popular than engineering and law. Especially from 2007 to 2008, there is a slope in the graph due to the 2007 to 2008 financial crisis that more people start to pay attention to economy and financial.

Overall, the data in library can indicate various information about local people's life. Unfortunately, because of internet, the number of people who come to library is decreasing sharply.