MAT 265
Assignment 6
Ilia Nikiforov

Outlier detection – incorrect check-in times

Abstract

In this report I focus on entries with incorrectly classified check in times (earlier than check out times). I explore overall yearly trends in those anomalies, use cross tabs to classify them by both check in and check out, identify most extreme cases with largest discrepancies, and investigate cases with **both** check in time and check out time classified incorrectly.

Report

For the first query, I start with the summary of cases when check-in times occur earlier than check-out times aggregated by year.

QUERY 1

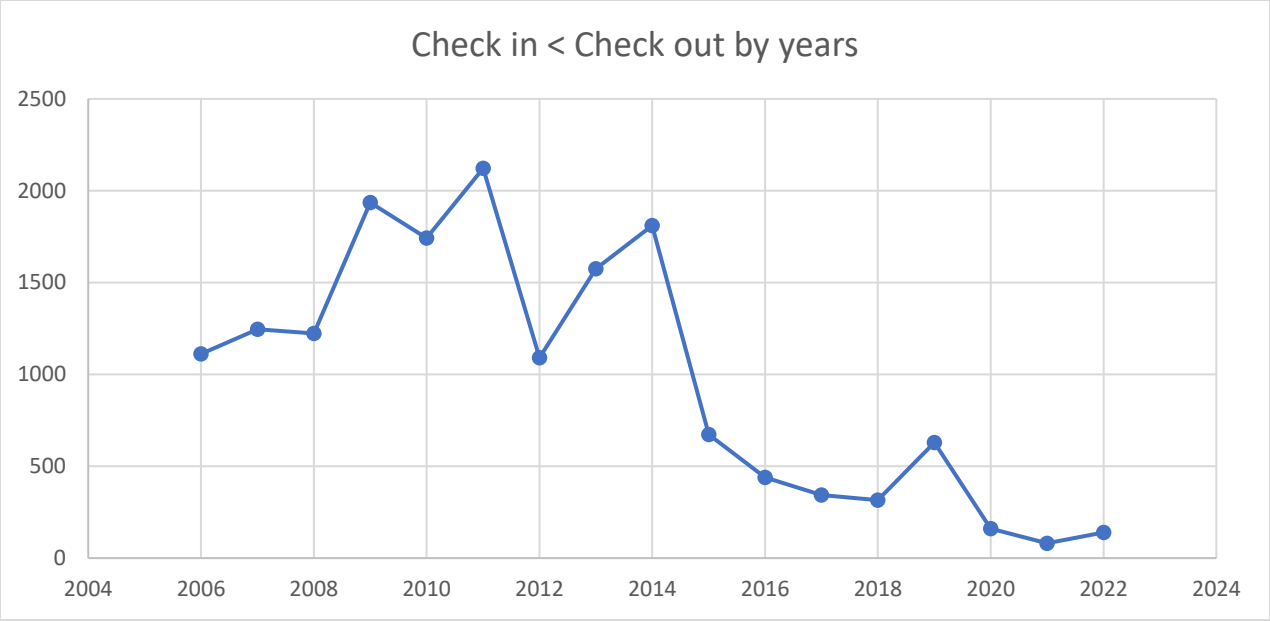
```
select

year(cout) as years,

count(*)

from spl_2016.inraw

where cin<cout

group by 1
```

RESULT

See cin_less_cout_years.csv.

The following chart demonstrates:

1. The number of those cases decreased with time. Perhaps SPL improved their processes and were able to catch most of the technical errors that lead to check-in times being earlier than check-out times.
2. There are no entries with “placeholder” date of 1970-01-01.



For the next query, I will aggregate both check-ins and check-outs by day of the week. The purpose here is to catch whether there is a difference between those times of more than 1 day.

QUERY 2

```
select

DAYNAME(cin) as day_cin,

DAYNAME(cout) as day_cout,

count(*)

from spl_2016.inraw

where cin<cout

group by 1,2
```

RESULT:

See cin_less_cout_days.csv.

The following cross tab (made in Excel, it's faster) shows that all anomalies occurred on the same date – there are no cases when check-in time was more than 24 hours earlier than check-out times.

		Check out							
	Sum of count(*)	**Column Labels**							
	Row Labels	**Sunday**	**Monday**	**Tuesday**	**Wednesday**	**Thursday**	**Friday**	**Saturday**	**Grand Total**
Check in	Sunday	1424							1424
	Monday		2740						2740

Tuesday			2927					2927
Wednesday				3005				3005
Thursday					2675			2675
Friday						1930		1930
Saturday							1941	1941
Grand Total	**1424**	**2740**	**2927**	**3005**	**2675**	**1930**	**1941**	**16642**

For the next query, I will repeat the same procedure, but this time aggregate data by hours.

QUERY 3

```
select

hour(cin) as cin_hours,

hour(cout) as cout_hours,

count(*)

from spl_2016.inraw

where cin<cout

group by 1,2
```

RESULT:

See cin_less_cout_hours.csv.

The crosstab below shows the following:

1. In most cases, check-in times were logged for the same hour as check-out times (the difference is therefore less than 60 minutes).
2. In the top right part of the crosstab: there is a small number of cases when check-in time is more than 1 hour earlier compared to check-out times. For most of those cases, check-in time was logged in the range of 9am-12pm. The most dramatic difference is 10 hours: check out hour is 19, while check in hour is 9.
3. The fields highlighted in yellow correspond to the times outside of working hours. Let's assume that check-out hours are logged into the system automatically when someone gives a book to a librarian and librarian scans the code. That means that in general, check-out times must be correct in the database. The yellow fields indicate that a check-out happened before 10am and after 8pm, meaning that the library was closed, and it was not possible to check out books. Therefore, yellow cells highlight a clear technical issue on **both** ends: incorrect check in, and incorrect check out. Let's also ignore the check outs that happened at 20pm – maybe someone was late and checked the book out at 20.01 – that seems like a plausible real life scenario.

		Check out hour																	
		0	**4**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**23**
Check in hour	0	3																	
	4		1																
	7			2															
	8				87														
	9					334	14	1	3	3	1	3	3	2	2	1			
	10						1150	18	2	1	2	5	2	7	5	2			
	11							1498	22	1	1	2	3	2		2			
	12								1746	34	12	17	29	19	4	15			
	13									1803	33								
	14										1883	26	1	1	2				
	15											2110	37	1					
	16												1971	31		2			
	17													1813	14				
	18														906	20			
	19															917	6		
	20																2		
	21																	1	
	23																		1

outside working hours

For next queries, I will closer inspect these entries that indicate both check in and check out error. First, I will see if there are any distinct items with the curse of wrong time:

QUERY 5:

```
select

distinct(itemNumber),

title,

itemType,

count(*)

from spl_2016.inraw

where cin<cout and hour(cout) not in (10,11,12,13,14,15,16,17,18,19,20)

group by 1,2,3

order by 4 desc

limit 10
```

RESULT:

See top_items_with_cin_cout_errors.csv.

It seems that there is no particular item that most often has issues with check in and check out times simultaneously. Interestingly, there are a few “test items” – clearly used in some technical diagnostics process on purpose.

itemNumber	title	itemType	count(*)
2370795	Snow wolf	acbk	3
609299	Encouraging the heart a leaders guide to rewarding and recognizing others	acbk	2
948660	Moses on management 50 leadership lessons from the greatest manager of all time	acbk	2
2219177	Huai yun sheng bing zen me ban 80 ge yun qi ji bing fang zhi de zui jia cuo shi	acbk	2
2782287	Test Title TagSys Folio 220	acbk	2
796882	Carry me back	acbk	2
2262548	test item	dcillb	2
2471150	Pizza kittens	jcbk	1
4185287	Come dance with me	accd	1
5331022	Hunting badger	acbk	1

Next, I will see how many of these are actually attributed to test items, by classifying titles into test vs normal items:

QUERY 6:

```
select

case when

title like 'test%' or 'Test%' or '%test%' then 'test_item'

else 'normal_item' end as test_vs_normal,

count(*)

from spl_2016.inraw

where cin<cout and hour(cout) not in (10,11,12,13,14,15,16,17,18,19,20)

group by 1
```

RESULT:

See test_vs_normal_items.csv.

Clearly, test items are very few – most of the items that have both check in and check out issues simultaneously are actually normal library items.

test_vs_normal	count(*)
normal_item	424
test_item	5

Conclusion

In this report I've focused on items that have an anomaly check in time earlier than check out time. I found that this discrepancy is never higher than 24 hours, and in most cases is less than 60 minutes. Most of the cases with discrepancy of more than 1 hour are "checked in" before 12pm. There is a small number of cases that have both incorrect check in time and incorrect check out time (occurred outside of normal working hours). Most of those "worst" cases are regular library books, with a few "test items" which can probably be ignored.