# dataPreprocessing

February 10, 2022

## 1 Data preprocessing

### 1.0.1 Import libraries

[1]:
```
import copy
import numpy as np
import pandas as pd
from datetime import datetime
```

### 1.0.2 Read data

[2]:
```
orig_cout_df = pd.read_csv("Magic_Checkouts_Inraw_2012-2021.csv")
orig_cout_df.head()
```

[2]:

| | id | itemNumber | bibNumber | cout | cin | collcode | itemtype | barcode | title | callNumber | deweyClass | subj |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 55353049 | 4281565 | 2747292 | 2012-01-09 11:01:00 | 2012-01-09 11:06:00 | nanf | acbk | 10073481748 | Magical mathematics the mathematical ideas tha… | 793.85 D5403M 2012 | 793.85 | NaN |
| 1 | 55385010 | 3352322 | 2547761 | 2012-01-09 19:30:00 | 2012-01-10 17:08:00 | ncnf | jcbk | 10063272545 | Amazing magic tricks apprentice level | J793.8 B266A02 2009 | 793.80 | NaN |
| 2 | 55385036 | 2800764 | 2415516 | 2012-01-09 19:30:00 | 2012-01-10 17:08:00 | nynf | jcbk | 10057005547 | ultimate book of card magic tricks | YA 793.8 L857U 2006 | | |
| 3 | 55385068 | 3811326 | 2653782 | 2012-01-09 19:30:00 | 2012-01-10 17:08:00 | ncnf | jcbk | 10070852222 | Magic up your sleeve amazing illusions tricks … | J793.8 B3885M 2010 | | |
| 4 | 55433477 | 4147770 | 2724101 | 2012-01-09 15:49:00 | 2012-01-13 17:07:00 | ncrdr | jcbk | 10074013979 | Magic tricks | ER MCKAY | | |

```
2      793.80   NaN
3      793.80   NaN
4         NaN   NaN
```

### 1.0.3 Remove unnecessary columns

[3]:
```
title_df = orig_cout_df.drop(columns=['itemNumber', 'bibNumber', 'collcode',␣
↪'itemtype', 'barcode', 'callNumber', 'deweyClass', 'subj'])
title_df.head()
```

[3]:
```
         id                 cout                  cin  \
0  55353049  2012-01-09 11:01:00  2012-01-09 11:06:00
1  55385010  2012-01-09 19:30:00  2012-01-10 17:08:00
2  55385036  2012-01-09 19:30:00  2012-01-10 17:08:00
3  55385068  2012-01-09 19:30:00  2012-01-10 17:08:00
4  55433477  2012-01-09 15:49:00  2012-01-13 17:07:00

                                              title
0  Magical mathematics the mathematical ideas tha…
1              Amazing magic tricks apprentice level
2                 ultimate book of card magic tricks
3  Magic up your sleeve amazing illusions tricks …
4                                       Magic tricks
```

### 1.0.4 Add title id column

[4]:
```
titles = []
for rcdi in range(len(title_df)):
    if title_df.loc[rcdi, 'title'] not in titles:
        titles.append(title_df.loc[rcdi, 'title'])
for rcdi in range(len(title_df)):
    title_df.loc[rcdi, 'title_id'] = int(titles.index(title_df.loc[rcdi,␣
↪'title']))
title_df['title_id'].astype(int)
print("Number of titles: ", len(titles))
```

```
Number of titles:  64
```

[5]:
```
title_df.head()
```

[5]:
```
         id                 cout                  cin  \
0  55353049  2012-01-09 11:01:00  2012-01-09 11:06:00
1  55385010  2012-01-09 19:30:00  2012-01-10 17:08:00
2  55385036  2012-01-09 19:30:00  2012-01-10 17:08:00
3  55385068  2012-01-09 19:30:00  2012-01-10 17:08:00
4  55433477  2012-01-09 15:49:00  2012-01-13 17:07:00

                                              title  title_id
```

```
0  Magical mathematics the mathematical ideas tha…       0.0
1              Amazing magic tricks apprentice level       1.0
2                 ultimate book of card magic tricks       2.0
3  Magic up your sleeve amazing illusions tricks …       3.0
4                                      Magic tricks       4.0
```

### 1.0.5 Add year, month, day, and duration column

in hours

[6]:
```
time_df = copy.deepcopy(title_df)
for rcdi in range(len(time_df)):
    cout = datetime.strptime(time_df.loc[rcdi, 'cout'], "%Y-%m-%d %H:%M:%S")
    cin = datetime.strptime(time_df.loc[rcdi, 'cin'], "%Y-%m-%d %H:%M:%S")
    delta_t = cin - cout
    time_df.loc[rcdi, 'year'] = cout.year
    time_df.loc[rcdi, 'month'] = cout.month
    time_df.loc[rcdi, 'day'] = cout.day
    time_df.loc[rcdi, 'duration'] = delta_t
    time_df.loc[rcdi, 'rent_hours'] = round(delta_t.days*24 + delta_t.seconds/
 ↪3600, 1)
del rcdi
time_df.head()
```

[6]:
```
         id                 cout                  cin  \
0  55353049  2012-01-09 11:01:00  2012-01-09 11:06:00
1  55385010  2012-01-09 19:30:00  2012-01-10 17:08:00
2  55385036  2012-01-09 19:30:00  2012-01-10 17:08:00
3  55385068  2012-01-09 19:30:00  2012-01-10 17:08:00
4  55433477  2012-01-09 15:49:00  2012-01-13 17:07:00

                                               title  title_id    year  month  \
0  Magical mathematics the mathematical ideas tha…       0.0  2012.0    1.0
1              Amazing magic tricks apprentice level       1.0  2012.0    1.0
2                 ultimate book of card magic tricks       2.0  2012.0    1.0
3  Magic up your sleeve amazing illusions tricks …       3.0  2012.0    1.0
4                                      Magic tricks       4.0  2012.0    1.0

   day        duration  rent_hours
0  9.0 0 days 00:05:00         0.1
1  9.0 0 days 21:38:00        21.6
2  9.0 0 days 21:38:00        21.6
3  9.0 0 days 21:38:00        21.6
4  9.0 4 days 01:18:00        97.3
```

### 1.0.6 Output dataset

[7]:
```
magic_df = time_df.drop(columns=['cout', 'cin'])
magic_df['year'].astype(int)
magic_df['month'].astype(int)
magic_df['day'].astype(int)
magic_df.head()
```

[7]:

| | id | title | title_id | year | month | day | duration | rent_hours |
|---|---|---|---|---|---|---|---|---|
| 0 | 55353049 | Magical mathematics the mathematical ideas tha… | 0.0 | 2012.0 | 1.0 | 9.0 | 0 days 00:05:00 | 0.1 |
| 1 | 55385010 | Amazing magic tricks apprentice level | 1.0 | 2012.0 | 1.0 | 9.0 | 0 days 21:38:00 | 21.6 |
| 2 | 55385036 | ultimate book of card magic tricks | 2.0 | 2012.0 | 1.0 | 9.0 | 0 days 21:38:00 | 21.6 |
| 3 | 55385068 | Magic up your sleeve amazing illusions tricks … | 3.0 | 2012.0 | 1.0 | 9.0 | 0 days 21:38:00 | 21.6 |
| 4 | 55433477 | Magic tricks | 4.0 | 2012.0 | 1.0 | 9.0 | 4 days 01:18:00 | 97.3 |

[8]:
```
magic_df.to_csv("Magic_Checkouts_Duration_Inraw_2012-2021.csv", index=False)
```