

```
In [3]: import os
import pandas as pd
pd.options.display.float_format = '{:20,.2f}'.format
pd.set_option('display.max_rows', 5000)
pd.set_option('display.max_columns', 5000)
pd.set_option('display.width', 1000)
pd.set_option('display.max_colwidth', -1)
```

## Ebooks 1 - 784 Datasets ID 12

#<https://sites.google.com/site/anhaidgroup/useful-stuff/data>  
(<https://sites.google.com/site/anhaidgroup/useful-stuff/data>)

```
In [6]: ebooks1_path = (r'/home/ubuntu/jupyter/ServerX/1_Standard Data Integrat
r'/Unprocessed Data/product_samples/ebooks1')
ebooks1_fields = ['record_id', 'title', 'description', 'publisher', 'price']
```

### Dataset 1 - Ebooks

```
In [127]: raw_ebooks = pd.read_csv(os.path.join(ebooks1_path, 'ebooks.csv'))
```

```
In [128]: raw_ebooks.columns
```

```
Out[128]: Index(['record_id', 'ISBN', 'description', 'price', 'date', 'publishe
r', 'review_count', 'title', 'rating_value', 'author', 'length', 'shor
t_description'], dtype='object')
```

```
In [78]: ebooks_data = pd.read_csv(
    os.path.join(ebooks1_path, 'ebooks.csv'),
    sep = ',',
    usecols = ebooks1_fields,
    index_col = None,
)#[ebooks1_fields]
```

```
In [79]: ebooks_data.rename(
    columns = {
        'title' : 'name',
        'publisher': 'producer',
    },
    inplace = True
)
```

```
In [80]: ebooks_data.columns
```

```
Out[80]: Index(['record_id', 'name', 'description', 'producer', 'price'], dtype
='object')
```

```
In [37]: ebooks_data.head()
```

```
Out[37]:
```

	record_id	name	description	producer
0	1	A Healthy Guide to Sport	This book provides a comprehensive plan that helps parents guide their children toward a healthy love of sports. It will show parents, and their kids, how to get involved in triathlon and other "lifestyle" sports that can be pursued for a lifetime. Triathlon is a fantastic endeavor for children because it promotes involvement in three "lifestyle" sports. No matter what the age, children who develop some skills in swimming, biking and running will be able to pursue any sport they would like when they are older. Sport should be the most enjoyable (and educational) activity our children get to do. It should provide them with a positive feeling about themselves. It should teach them that they can improve themselves in all endeavors of life by working hard and persevering. It should teach them about the benefits of working as part of a team or training group. It should show them how much they gain in life by striving to achieve their goals.	Meyer & Meyer

## Dataset 2 - iTunes

```
In [4]: import pandas as pd
import os
```

```
In [7]: itunes_data = pd.read_csv(
    os.path.join(ebooks1_path, 'itunes.csv'),
    sep = ',',
    usecols = ebooks1_fields,
)#[ebooks1_fields]
```

```
In [8]: itunes_data.columns
```

```
Out[8]: Index(['record_id', 'publisher', 'description', 'title', 'price'], dtype='object')
```

```
In [39]: itunes_data.rename(
    columns = {
        'title' : 'name',
        'publisher': 'producer',
    },
    inplace = True
)
```

```
In [40]: itunes_data.head()
```

```
Out[40]:
```

	record_id	name	description	producer	price
0	1	Game of My Life Virginia Tech Hokies	Virginia Tech's Shayne Graham trots onto the field at West Virginia on November 6, 1999, with two thoughts in his mind. One is a missed field goal that would have beaten Miami a year earlier. The other is the 44-yard field goal he is about to try against the Mountaineers, a kick he must make if the Hokies are to stay unbeaten and on track for a national championship. Head down, he focuses on his mark as the ball is snapped. He steps forward, the dream of an entire team resting with his leg. Now, hear Graham's memory of that kick in his own words, for the first time. Game of My Life: Virginia Tech Hokies, first published in 2006, celebrates the extraordinary football and basketball moments that have shaped the college's rich athletic heritage. Through interviews with some of the school's most prestigious athletes, Hokies fans can relive the big games that defined the school's winning tradition. Carroll Dale, later a fixture with the Green Bay Packers, dove arms outstretched to haul in a crucial two-point conversion in a 1957 game against the University of Richmond. Les Henson shot from the baseline the other baseline as the clock neared zero against Florida State in 1980. Chris Smith went well beyond the "double-double" standard for points and rebounds. How about 30 and 31 against Marshall in 1959? Corey Moore made life miserable for Clemson quarterback Brandon Streeter one night in 1999. Bruce Smith did the same for Duke quarterback Ben Bennett in 1983. The Hokies' Jim Pyne, meanwhile, made sure Syracuse's Kevin Mitchell didn't do the same to Tech quarterback Maurice DeShazo in 1993. Carlos Dixon, Mike Imoh, Andre Davis, Dell Curry, Bryan Still, Don Strock, Bryan Randall all the Tech greats from the gridiron and hardwood are in these pages, including coach Frank Beamer. Join thousands of Virginia Tech fans in remembering these cherished stories. For the athletes within, these truly were the games of their lives.	Sports Publishing	16.99
1	2	Dale Brown's Basketball Coaches Organizational Handbook (2nd Edition)	In this practical, easy-to-read and apply book, LSU's coach of 25 years shows coaches at all competitive levels how they can become more productive and effective with less time through better organizational skills. Coach Brown covers practically every aspect of the coaching profession as he allows you to learn from his mistakes and successes. Includes chapters on managers' duties and guidelines, team meetings, motivation, practice organization, scouting, road trips, media relations and recruiting. Even includes a chapter on organizing summer youth basketball camps! A must have for coaches at any level of play.	Coaches Choice	19.99

record_id		name	description	producer	price
2	3	Ya Gotta Believe!	<p>Are you a true Mets fan? Were you there when they won the 1986 World Series in the seventh game? Did you stand and cheer as the Mets demolished the St. Louis Cardinals to become the National League Champions in 2000? Do you know why the original team colors were orange and blue? How much do you really know about those lovable heroes who have brought fortune, glory, and two World Championship trophies to New York? Are you a true believer? Do you know: *Who the Hall of Fame outfielder was who played for the Mets in their inaugural season and went on to become a broadcaster for the Philadelphia Phillies? *Which Mets outfielder ran the bases backward after hitting his 100th career home run in a game in 1963? *Which rookie outfielder swiped 24 bases in 1981 and became one of the most popular players ever to play for the Mets? *When Tom Seaver's rookie year was? *Who holds the single-season Mets record for home runs? It's all here, with highlights of the team's exciting history, from the club's beginnings in 1962 to today, including postseason play. From Casey Stengel to Tom Seaver; from Doc to Mookie-- to Mike and Fonzie-- questions and answers, sidebars, fascinating bios and photos gathered by lifelong Mets fan Michael Lichtenstein. Much more than just facts and trivia, Ya Gotta Believe! is something no Mets fan can do without.</p>	St. Martin's Press	7.99
3	4	Canada's Other Game	<p>The story of Canada's other game from its invention by a Canadian to its current struggle for popularity. Basketball, the only major world sport undeniably invented by a Canadian, has ironically failed to win Canadians' hearts more than a century after its creation. James Naismith's brainchild is a popular recreational pastime in his homeland, but players with bigger dreams had better take their talents south of the border. Canadian hoops has languished in the seemingly eternal shadow of hockey, with its cannibalization of air time, advertising dollars, and corporate capital. Faced with limited opportunities at home, as many as 50 teenagers flock to U.S. prep schools and colleges every year to chase their dreams of college stardom and, much less likely, a shot at glory in the NBA. Against all odds, a skinny kid from Victoria named Steve Nash managed to reach the pinnacle of the sport, with a whirling-dervish style that earned him two MVP awards in the world's greatest league. Today, a new generation of Canadians stand poised to follow in Nash's path. But will their success spark a renaissance back home? This book chronicles basketball's struggle to overcome its history as a poor cousin in a hockey-mad nation.</p>	Dundurn	12.99

record_id	name	description	producer	price
4	"Double Duty" Radcliffe: 36 Years of Pitching & Catching In the Negro Leagues	"Double Duty" Radcliffe is the biography of one of the most unique baseball stars in history. Nicknamed "Double Duty" because he was an all-star pitcher and catcher, Radcliffe played from 1919-1954 in the Negro Leagues with teams such as the Chicago American Giants, Homestead Grays, Kansas City Monarchs, Birmingham Black Barons, New York Black Yankees and Memphis Red Sox. In this book, called "amazing" by Sports Illustrated, Radcliffe's own words are intertwined with background information and interviews with more than 30 ex-teammates. Radcliffe won more than 300 games on the mound, belted more than 300 homers, pitched and caught multiple no-hitters, was named league MVP at age 41, threw a complete game shutout past age 50, and played with and against every black and white player you can think of, including Satchel Paige, Josh Gibson, Jackie Robinson, Honus Wagner, Jimmie Foxx and Bob Feller! The braggadocious Radcliffe was known almost as much for his talking as for his playing, but, his opponents said, "he could back his words up." Radcliffe was a wonderful baseball player, but may have made more contributions to the game after retirement as he lived to age 103 and helped educated the public on the Negro Leagues and its players.	Kyle McNary	6.99

## Candset for ebooks

```
In [41]: ebook_candset_fields = ['_id', 'ltable.record_id', 'rtable.record_id']
```

```
In [42]: ebooks1_candset = pd.read_csv(
    os.path.join(ebooks1_path, 'candset.csv'),
    comment = '#',
    sep = ',',
)
```

```
In [43]: ebooks1_candset.head()
```

```
Out[43]:
```

	_id	ltable.record_id	rtable.record_id	ltable.author	ltable.date	ltable.length	ltable.price	ltable.gold
0	0	1	35	mike harris & Frank Beamer	Jul 07, 2015	264.00	16.99	
1	1	1	1222	mike harris & Frank Beamer	Jul 07, 2015	264.00	16.99	
2	2	1	3143	mike harris & Frank Beamer	Jul 07, 2015	264.00	16.99	
3	3	1	3582	mike harris & Frank	Jul 07,	264.00	16.99	

## Labeled data for ebooks

```
In [44]: ebooks1_labeled_data_fields = ['_id', 'ltable.record_id', 'rtable.record_id', 'gold']
```

```
In [45]: ebooks1_labeled_data = pd.read_csv(
    os.path.join(ebooks1_path, 'labeled_data.csv'),
    comment = '#',
    sep = ',',
    usecols = ebooks1_labeled_data_fields,
)[ebooks1_labeled_data_fields]
```

```
In [46]: ebooks1_labeled_data.rename(
    columns = {
        'ltable.record_id': 'itunes_id',
        'rtable.record_id': 'ebooks_id',
        'gold': 'duplicate',
    },
    inplace=True
)
```

```
In [47]: ebooks1_labeled_data.head()
```

```
Out[47]:
```

	_id	itunes_id	ebooks_id	duplicate
0	44	44	9650	0
1	60	66	3496	0
2	68	66	7326	0
3	107	92	6590	1
4	115	112	5954	0

```
In [48]: itunes_data[itunes_data['record_id'] == 92]['name'].to_string(index=False)
```

```
Out[48]: ' 100 Things Rockies Fans Should Know & Do Before They Die'
```

```
In [49]: itunes_data[itunes_data['record_id'] == 92]['description'].to_string(index=False)
```

```
Out[49]: ' This series will help baseball lovers get the most out of being a fan. It takes years of franchise history and distills it to the absolute best and most compelling, identifying in an informative, lively, and illuminating way the personalities, events, and facts every fan should know without hesitation. Numbers, nicknames, memorable moments, singular achievements, and signature plays all highlight the list of 100.'
```

```
In [50]: ebooks_data[ebooks_data['record_id'] == 6590]['name'].to_string(index=False)
```

```
Out[50]: ' 100 Things Rockies Fans Should Know & Do Before They Die'
```

```
In [51]: ebooks_data[ebooks_data['record_id'] == 6590]['description'].to_string(index=False)
```

```
Out[51]: ' All baseball enthusiasts want to see their team win the World Series in their lifetime. But being a fan is about more than watching your team win it all. This series will help baseball lovers get the most out of being a fan. It takes years of franchise history and distills it to the absolute best and most compelling, identifying in an informative, lively, and illuminating way the personalities, events, and facts every fan should know without hesitation. Numbers, nicknames, memorable moments, singular achievements, and signature plays all highlight the list of 100. 100 Things also includes things baseball fans should see and do before they join their heroes at the Pearly Gates. This book contains numerous tips and suggestions for enjoying a team on a different, more involved, level.'
```


## Candset

```
In [81]: ebooks1_candset_data = pd.read_csv(
    os.path.join(ebooks1_path, 'candset.csv'),
    comment = '#',
    sep = ',',
    usecols = ebooks1_labeled_data_fields,
)[ebooks1_labeled_data_fields]
```

```
In [87]: ebooks1_candset_data[ebooks1_candset_data['_id'] == 107]
```

```
Out[87]:
```

	_id	ltable.record_id	rtable.record_id	ltable.author	ltable.date	ltable.length	ltable.price	ltable...
107	107	36	4437	Susan Slusser	Jun 01, 2015	336.00	9.99	

◀  ▶

```
In [88]: ebooks1_labeled_data[ebooks1_labeled_data['itunes_id'] == 36]
```

```
Out[88]:
```

	_id	itunes_id	ebooks_id	duplicate
--	-----	-----------	-----------	-----------

```
In [86]: ebooks1_labeled_data.head()
```

```
Out[86]:
```

	_id	itunes_id	ebooks_id	duplicate
0	44	44	9650	0
1	60	66	3496	0
2	68	66	7326	0
3	107	92	6590	1
4	115	112	5954	0

```
In [83]: len(ebooks1_candset_data)
```

```
Out[83]: 18383
```

```
In [84]: len(ebooks_data)
```

```
Out[84]: 14110
```

```
In [85]: len(itunes_data)
```

```
Out[85]: 6500
```

## Candset - Labeled data comparison

```
In [89]: raw_ebooks1_candset_data = pd.read_csv(  
    os.path.join(ebooks1_path, 'candset.csv'),  
    comment = '#',  
    sep = ',',  
    #usecols = ebooks1_labeled_data_fields,  
    )[ebooks1_labeled_data_fields]
```

```
In [90]: raw_ebooks1_labeled_data = pd.read_csv(  
    os.path.join(ebooks1_path, 'labeled_data.csv'),  
    comment = '#',  
    sep = ',',  
    #usecols = ebooks1_labeled_data_fields,  
    )[ebooks1_labeled_data_fields]
```



```
In [95]: raw_ebooks1_candset_data[raw_ebooks1_candset_data['_id'] == 44]
```

```
Out[95]:
```

	_id	ltable.record_id	rtable.record_id	ltable.author	ltable.date	ltable.length	ltable.price	ltable
44	44	10	6609	Phil Pepe	Apr 07, 2015	224.00	16.99	

```
In [96]: len(raw_ebooks1_candset_data)
```

```
Out[96]: 18383
```

```
In [108]: len(raw_ebooks1_labeled_data)
```

```
Out[108]: 400
```

```
In [107]: len(raw_ebooks1_labeled_data[raw_ebooks1_labeled_data['gold'] == 1])
```

```
Out[107]: 131
```

131 duplicates in labeled data, warum?

```
In [105]: raw_ebooks1_labeled_data[raw_ebooks1_labeled_data.duplicated()]
```

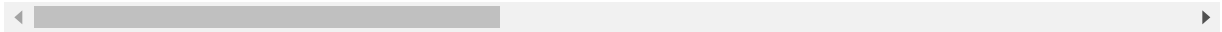
```
Out[105]:
```

	_id	ltable.record_id	rtable.record_id	ltable.author	ltable.date	ltable.description	ltable.length	ltable
--	-----	------------------	------------------	---------------	-------------	--------------------	---------------	--------

```
In [102]: raw_ebooks1_labeled_data[raw_ebooks1_labeled_data['ltable.record_id'] ==
```

```
Out[102]:      _id  ltable.record_id  rtable.record_id  ltable.author  ltable.date  ltable.description  ltable.length
```

						This series will help baseball lovers get the most out of being a fan. It takes years of franchise history and distills it to the absolute best and most compelling, identifying in an informative, lively, and illuminating way the personalities, events, and facts every fan should know without hesitation. Numbers, nicknames, memorable moments, singular achievements, and signature plays all highlight the list of 100.	
3	107	92	6590	Adrian Dater	1-Apr-09		256



```
In [110]: itunes_data.head(1)
```

```
Out[110]:
```

	record_id	name	description	producer	price
0	1	Game of My Life Virginia Tech Hokies	Virginia Tech's Shayne Graham trots onto the field at West Virginia on November 6, 1999, with two thoughts in his mind. One is a missed field goal that would have beaten Miami a year earlier. The other is the 44-yard field goal he is about to try against the Mountaineers, a kick he must make if the Hokies are to stay unbeaten and on track for a national championship. Head down, he focuses on his mark as the ball is snapped. He steps forward, the dream of an entire team resting with his leg. Now, hear Graham's memory of that kick in his own words, for the first time. Game of My Life: Virginia Tech Hokies, first published in 2006, celebrates the extraordinary football and basketball moments that have shaped the college's rich athletic heritage. Through interviews with some of the school's most prestigious athletes, Hokies fans can relive the big games that defined the school's winning tradition. Carroll Dale, later a fixture with the Green Bay Packers, dove arms outstretched to haul in a crucial two-point conversion in a 1957 game against the University of Richmond. Les Henson shot from the baseline the other baseline as the clock neared zero against Florida State in 1980. Chris Smith went well beyond the "double-double" standard for points and rebounds. How about 30 and 31 against Marshall in 1959? Corey Moore made life miserable for Clemson quarterback Brandon Streeter one night in 1999. Bruce Smith did the same for Duke quarterback Ben Bennett in 1983. The Hokies' Jim Pyne, meanwhile, made sure Syracuse's Kevin Mitchell didn't do the same to Tech quarterback Maurice DeShazo in 1993. Carlos Dixon, Mike Imoh, Andre Davis, Dell Curry, Bryan Still, Don Strock, Bryan Randall all the Tech greats from the gridiron and hardwood are in these pages, including coach Frank Beamer. Join thousands of Virginia Tech fans in remembering these cherished stories. For the athletes within, these truly were the games of their lives.	Sports Publishing	16.99

```
In [111]: itunes_data['source'] = 'itunes'
```

```
In [112]: ebooks_data['source'] = 'ebooks'
```

```
In [113]: ebooks1_all = pd.concat([itunes_data, ebooks_data])
```

```
In [114]: len(ebooks1_all) == len(itunes_data) + len(ebooks_data)
```

```
Out[114]: True
```

```
In [2]: ebooks1_all_path = (r'/home/ubuntu/jupyter/ServerX/1_Standard Data Inter  
r'/Processed Data/product_samples/')
```

```
In [119]: ebooks1_all.rename(columns={'record_id': 'Id'}, inplace=True)
```

```
In [120]: ebooks1_all.to_csv(ebooks1_all_path + 'ebooks1_all.csv')
```

```
In [121]: ebooks1_all[ebooks1_all['description'] == None]
```

```
Out[121]:
```

	Id	name	description	producer	price	source
--	----	------	-------------	----------	-------	--------

```
In [123]: ebooks1_all[ebooks1_all['producer'] == None]
```

```
Out[123]:
```

Id	name	description	producer	price	source
----	------	-------------	----------	-------	--------

```
In [124]: ebooks1_all[ebooks1_all['price'] == None]
```

```
Out[124]:
```

Id	name	description	producer	price	source
----	------	-------------	----------	-------	--------

```
In [125]: len(ebooks1_all)
```

```
Out[125]: 20610
```

```
In [4]: ebooks1_all = pd.read_csv(ebooks1_all_path + 'ebooks1_all.csv')
```

```
In [5]: ebooks1_all.head(2)
```

```
Out[5]:
```

Unnamed: 0	Id	name	description	producer	price	source
0	0	1	Game of My Life Virginia Tech Hokies	Sports Publishing	16.99	itunes

Unnamed: 0	Id	name	description	producer	price	source
1	1	2	<p>Dale Brown's Basketball Coaches Organizational Handbook (2nd Edition)</p> <p>In this practical, easy-to-read and apply book, LSU's coach of 25 years shows coaches at all competitive levels how they can become more productive and effective with less time through better organizational skills. Coach Brown covers practically every aspect of the coaching profession as he allows you to learn from his mistakes and successes. Includes chapters on managers' duties and guidelines, team meetings, motivation, practice organization, scouting, road trips, media relations and recruiting. Even includes a chapter on organizing summer youth basketball camps! A must have for coaches at any level of play.</p>	Coaches Choice	19.99	itunes