INFO 6210

Data Management and Database Design

Gathering, Scraping, Munging and Cleaning Data

Assignment 1

Report

Abstract

In this assignment, we will gather real-world data. This process is often called data munging or data wrangling. All our database tables are populated with real-world data, which is focused on the most popular sneakers that are sold in recent years on a sneakers database website. Built up from three different data sources, namely the API, web scraper and csv file, our thematically related database successfully passed the auditing in validity, completeness and consistency, and is presented with the Entity-Relationship diagram.

Data Sources

1. Web Scraper

We scraped **thesneakerdatabase.com** using **BeautifulSoup**.

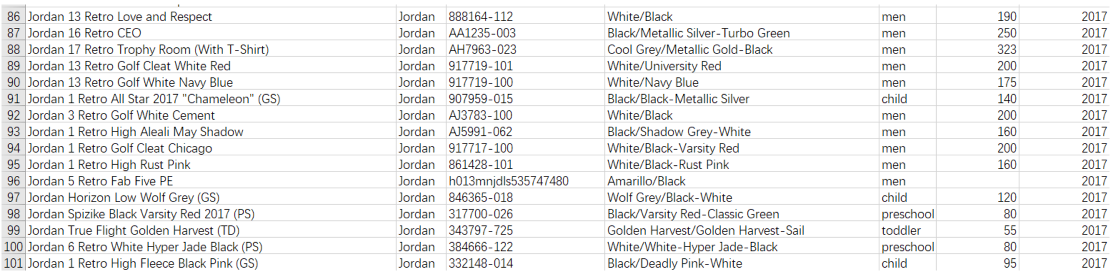
(URL=http://www.thesneakerdatabase.com)

This thesneakerdatabase.com is an online database that primarily stores information of sneakers. In this assignment, we choose to scrap its currently posted sneaker data, which is consisted of about 8k sneakers from brands such as Nike, Jordan, Converse, and Adidas. In this assignment, we successfully scraped information of popular sneakers on its market, in a total number of **100** products.

The file 'test1.csv' contains the data received from the web scraper with the following columns:

NAME, BRAND, STYLE ID, COLORWAY, GENDER, RETAIL PRICE, RELEASE YEAR

Here follows a screenshot of the data received.



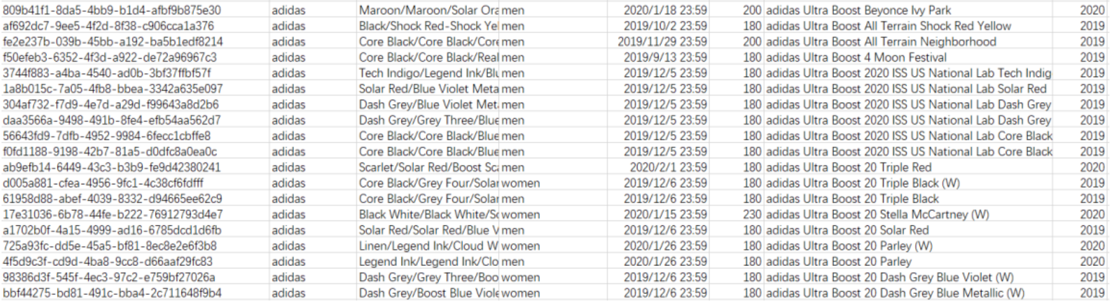
86	Jordan 13 Retro Love and Respect	Jordan	888164-112	White/Black	men	190	2017
87	Jordan 16 Retro CEO	Jordan	AA1235-003	Black/Metallic Silver-Turbo Green	men	250	2017
88	Jordan 17 Retro Trophy Room (With T-Shirt)	Jordan	AH7963-023	Cool Grey/Metallic Gold-Black	men	323	2017
89	Jordan 13 Retro Golf Cleat White Red	Jordan	917719-101	White/University Red	men	200	2017
90	Jordan 13 Retro Golf White Navy Blue	Jordan	917719-100	White/Navy Blue	men	175	2017
91	Jordan 1 Retro All Star 2017 "Chameleon" (GS)	Jordan	907959-015	Black/Black-Metallic Silver	child	140	2017
92	Jordan 3 Retro Golf White Cement	Jordan	AJ3783-100	White/Black	men	200	2017
93	Jordan 1 Retro High Aleali May Shadow	Jordan	AJ5991-062	Black/Shadow Grey-White	men	160	2017
94	Jordan 1 Retro Golf Cleat Chicago	Jordan	917717-100	White/Black-Varsity Red	men	200	2017
95	Jordan 1 Retro High Rust Pink	Jordan	861428-101	White/Black-Rust Pink	men	160	2017
96	Jordan 5 Retro Fab Five PE	Jordan	h013mnjdls535747480	Amarillo/Black	men		2017
97	Jordan Horizon Low Wolf Grey (GS)	Jordan	846365-018	Wolf Grey/Black-White	child	120	2017
98	Jordan Spizike Black Varsity Red 2017 (PS)	Jordan	317700-026	Black/Varsity Red-Classic Green	preschool	80	2017
99	Jordan True Flight Golden Harvest (TD)	Jordan	343797-725	Golden Harvest/Golden Harvest-Sail	toddler	55	2017
100	Jordan 6 Retro White Hyper Jade Black (PS)	Jordan	384666-122	White/White-Hyper Jade-Black	preschool	80	2017
101	Jordan 1 Retro High Fleece Black Pink (GS)	Jordan	332148-014	Black/Deadly Pink-White	child	95	2017

Figure 1-1-1 Screenshot of Data Gathered from Web Scraper

2.Web API

In this assignment, we choose to use the API called **Sneaker Database API.** This API is provided by thesneakerdatabase.com. It is one of the most popular websites that provides information for sneaker lovers with its well-covered sneakers data. After taking several times of tests, we finally received **2,000** feedback. The latest dataset we received has been stored as a Jason dataset, and will be used in further auditing process.

Here follows a screenshot of the dataset we got:



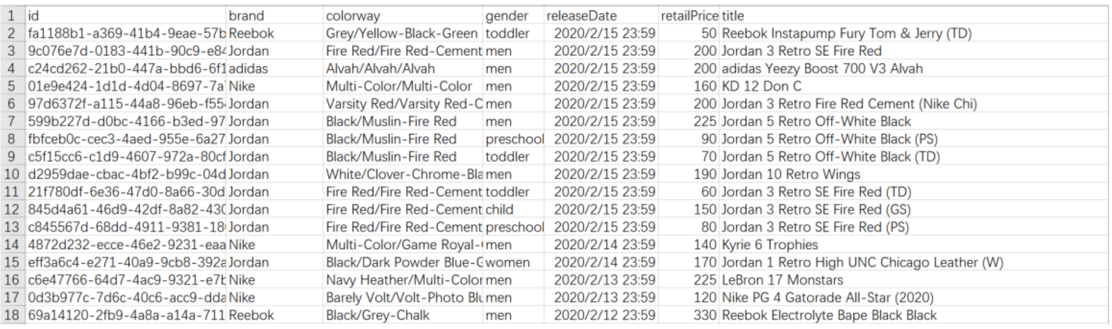
809b41f1-8da5-4bb9-b1d4-afbf9b875e30	adidas	Maroon/Maroon/Solar Ora	men	2020/1/18 23:59	200	adidas Ultra Boost Beyonce Ivy Park	2020
af692dc7-9ee5-4f2d-8f38-c906cca1a376	adidas	Black/Shock Red-Shock Ye	men	2019/10/2 23:59	180	adidas Ultra Boost All Terrain Shock Red Yellow	2019
fe2e237b-039b-45bb-a192-ba5b1edf8214	adidas	Core Black/Core Black/Core	men	2019/11/29 23:59	200	adidas Ultra Boost All Terrain Neighborhood	2019
f50efeb3-6352-4f3d-a922-de72a96967c3	adidas	Core Black/Core Black/Real	men	2019/9/13 23:59	180	adidas Ultra Boost 4 Moon Festival	2019
3744f883-a4ba-4540-ad0b-3bf37ffbf57f	adidas	Tech Indigo/Legend Ink/Blu	men	2019/12/5 23:59	180	adidas Ultra Boost 2020 ISS US National Lab Tech Indig	2019
1a8b015c-7a05-4fb8-bbea-3342a635e097	adidas	Solar Red/Blue Violet Meta	men	2019/12/5 23:59	180	adidas Ultra Boost 2020 ISS US National Lab Solar Red	2019
304af732-f7d9-4e7d-a29d-f99643a8d2b6	adidas	Dash Grey/Blue Violet Met.	men	2019/12/5 23:59	180	adidas Ultra Boost 2020 ISS US National Lab Dash Grey	2019
daa3566a-9498-491b-8fe4-efb54aa562d7	adidas	Dash Grey/Grey Three/Blue	men	2019/12/5 23:59	180	adidas Ultra Boost 2020 ISS US National Lab Dash Grey	2019
56643fd9-7dfb-4952-9984-6fecc1cbffe8	adidas	Core Black/Core Black/Blue	men	2019/12/5 23:59	180	adidas Ultra Boost 2020 ISS US National Lab Core Black	2019
f0fd1188-9198-42b7-81a5-d0dfc8a0ea0c	adidas	Core Black/Core Black/Blue	men	2019/12/5 23:59	180	adidas Ultra Boost 2020 ISS US National Lab Core Black	2019
ab9efb14-6449-43c3-b3b9-fe9d42380241	adidas	Scarlet/Solar Red/Boost Sc	men	2020/2/1 23:59	180	adidas Ultra Boost 20 Triple Red	2020
d005a881-cfea-4956-9fc1-4c38cf6fdfff	adidas	Core Black/Grey Four/Solar	women	2019/12/6 23:59	180	adidas Ultra Boost 20 Triple Black (W)	2019
61958d88-abef-4039-8332-d94665ee62c9	adidas	Core Black/Grey Four/Solar	men	2019/12/6 23:59	180	adidas Ultra Boost 20 Triple Black	2019
17e31036-6b78-44fe-b222-76912793d4e7	adidas	Black White/Black White/S	women	2020/1/15 23:59	230	adidas Ultra Boost 20 Stella McCartney (W)	2020
a1702b0f-4a15-4999-ad16-6785dcd1d6fb	adidas	Solar Red/Solar Red/Blue V	men	2019/12/6 23:59	180	adidas Ultra Boost 20 Solar Red	2019
725a93fc-dd5e-45a5-bf81-8ec8e2e6f3b8	adidas	Linen/Legend Ink/Cloud W	women	2020/1/26 23:59	180	adidas Ultra Boost 20 Parley (W)	2020
4f5d9c3f-cd9d-4ba8-9cc8-d66aaf29fc83	adidas	Legend Ink/Legend Ink/Clo	men	2020/1/26 23:59	180	adidas Ultra Boost 20 Parley	2020
98386d3f-545f-4ec3-97c2-e759bf27026a	adidas	Dash Grey/Grey Three/Boo	women	2019/12/6 23:59	180	adidas Ultra Boost 20 Dash Grey Blue Violet (W)	2019
bbf44275-bd81-491c-bba4-2c711648f9b4	adidas	Dash Grey/Boost Blue Viole	women	2019/12/6 23:59	180	adidas Ultra Boost 20 Dash Grey Blue Metallic (W)	2019

Figure 1-2-1 Screenshot of Data Gathered from Web API

3. Raw text, csv, xml or excel data.

In this part, we successfully generate csv. files by reforming the data we got. Each of the original csv files contains 2,000 rows of data that are filled in 3 or 8 columns.

Here follows a screenshot of csv data we gathered:



1	id	brand	colorway	gender	releaseDate	retailPrice	title
2	fa1188b1-a369-41b4-9eae-57b	Reebok	Grey/Yellow-Black-Green	toddler	2020/2/15 23:59	50	Reebok Instapump Fury Tom & Jerry (TD)
3	9c076e7d-0183-441b-90c9-e84	Jordan	Fire Red/Fire Red-Cement	men	2020/2/15 23:59	200	Jordan 3 Retro SE Fire Red
4	c24cd262-21b0-447a-bbd6-6f1	adidas	Alvah/Alvah/Alvah	men	2020/2/15 23:59	200	adidas Yeezy Boost 700 V3 Alvah
5	01e9e424-1d1d-4d04-8697-7a	Nike	Multi-Color/Multi-Color	men	2020/2/15 23:59	160	KD 12 Don C
6	97d6372f-a115-44a8-96eb-f55	Jordan	Varsity Red/Varsity Red-C	men	2020/2/15 23:59	200	Jordan 3 Retro Fire Red Cement (Nike Chi)
7	599b227d-d0bc-4166-b3ed-97	Jordan	Black/Muslin-Fire Red	men	2020/2/15 23:59	225	Jordan 5 Retro Off-White Black
8	fbfceb0c-cec3-4aed-955e-6a27	Jordan	Black/Muslin-Fire Red	preschool	2020/2/15 23:59	90	Jordan 5 Retro Off-White Black (PS)
9	c5f15cc6-c1d9-4607-972a-80cf	Jordan	Black/Muslin-Fire Red	toddler	2020/2/15 23:59	70	Jordan 5 Retro Off-White Black (TD)
10	d2959dae-cbac-4bf2-b99c-04d	Jordan	White/Clover-Chrome-Bla	men	2020/2/15 23:59	190	Jordan 10 Retro Wings
11	21f780df-6e36-47d0-8a66-30d	Jordan	Fire Red/Fire Red-Cement	toddler	2020/2/15 23:59	60	Jordan 3 Retro SE Fire Red (TD)
12	845d4a61-46d9-42df-8a82-43(	Jordan	Fire Red/Fire Red-Cement	child	2020/2/15 23:59	150	Jordan 3 Retro SE Fire Red (GS)
13	c845567d-68dd-4911-9381-18	Jordan	Fire Red/Fire Red-Cement	preschool	2020/2/15 23:59	80	Jordan 3 Retro SE Fire Red (PS)
14	4872d232-ecce-46e2-9231-eaa	Nike	Multi-Color/Game Royal-(	men	2020/2/14 23:59	140	Kyrie 6 Trophies
15	eff3a6c4-e271-40a9-9cb8-392a	Jordan	Black/Dark Powder Blue-G	women	2020/2/14 23:59	170	Jordan 1 Retro High UNC Chicago Leather (W)
16	c6e47766-64d7-4ac9-9321-e7b	Nike	Navy Heather/Multi-Color	men	2020/2/13 23:59	225	LeBron 17 Monstars
17	0d3b977c-7d6c-40c6-acc9-dda	Nike	Barely Volt/Volt-Photo Blu	men	2020/2/13 23:59	120	Nike PG 4 Gatorade All-Star (2020)
18	69a14120-2fb9-4a8a-a14a-711	Reebok	Black/Grey-Chalk	men	2020/2/12 23:59	330	Reebok Electrolyte Bape Black Black

Figure 1-3-1 Screenshot of Data Gathered from Raw Csv Files

Conceptual Data Modeling

Conceptual data modeling is a visual representation of an organization's data semantics. The things we need to store is called Entity. Correspondingly, these entities have characteristics. At the same time, all the entities should have a thematic relation. We will explain it using object model diagram.

From the following Entity Relationship graph, we could see three different tables, namely colorway, retail price and gender. The primary key of these three tables are the 'title' of the sneakers. Since the combined table contains all the fields of the three tables, we could consider it as the conceptual database schema.

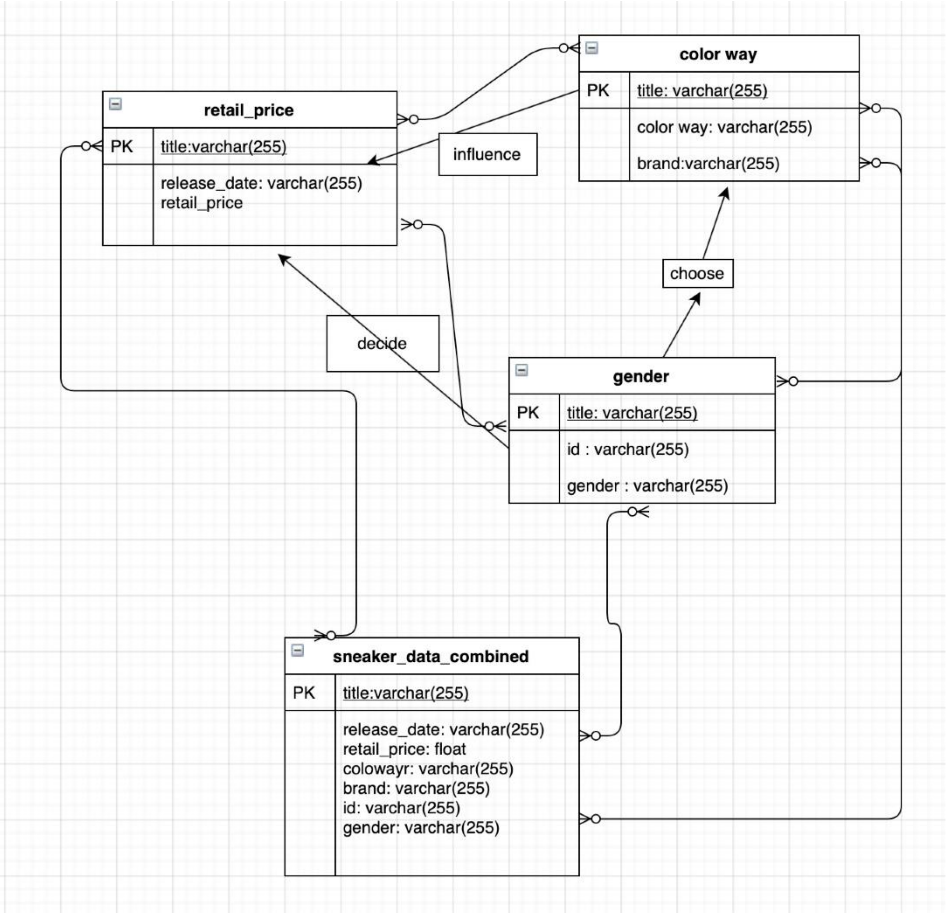


Figure 2-1-1 Entity Relationship Graph

From the relationship between the three tables, we could observe that the sneakers with the same color could have different retail prices, and the sneakers of the same price could have different colors. Thus, the relationship of color way and retail_price is Many-to-Many relationship. So are the retail_price - gender, colorway-gender relationship, in other words, three tables are all Many-to-Many relationship.

Data Auditing

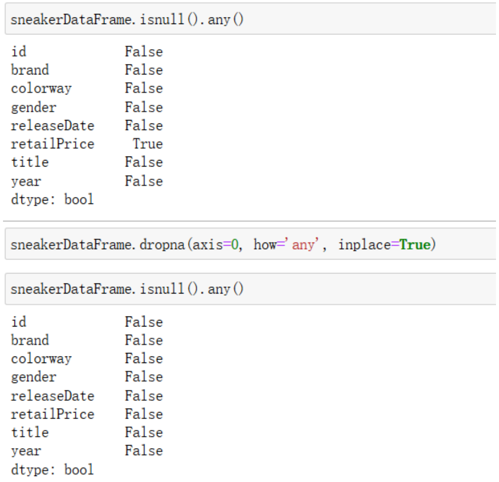
Audit validity/ accuracy

In this part, we will audit the validity of our data. The main process is as below:

1.Check whether there are any null or duplicates

2.Delete the null or duplicated data

In the first version of the rawSneakerJsonData table, we spotted several null values. Among those 2,000 rows of data, we cleaned all those data with any null values and at last, 1,694 rows of data remained. Because currently there are no duplicates and null values, the database is already cleaned.



```
sneakerDataFrame.isnull().any()
```

```
id             False
brand          False
colorway       False
gender         False
releaseDate    False
retailPrice     True
title          False
year           False
dtype: bool
```

```
sneakerDataFrame.dropna(axis=0, how='any', inplace=True)
```

```
sneakerDataFrame.isnull().any()
```

```
id             False
brand          False
colorway       False
gender         False
releaseDate    False
retailPrice    False
title          False
year           False
dtype: bool
```

Figure 3-1-1 screenshot of Data validity auditing process and result

Audit Completeness

There are 7 columns contained in the dataset, which are brand name, release date, retail price, color, brand, sneaker id and gender. These data cover all the important entities of sneakers which should have, and thus satisfy the completeness of database.

Audit Consistency and Uniformity

The datasets are inner related by a common entity, title, which relates all the datasets. Thus, the final combined database is considered uniformed and consistent.

Final Report

In this assignment, the following files are generated and used: sneaker.csv, gender.csv, Sneaker_title.csv, colorway.csv, retailPrice.csv,test1.csv. These data are gathered from three different source, web API, web scraping and raw data, which are then merged to form the final conceptual model.

Conclusion

In this assignment, firstly, we gathered data from different source, which are web API, web scraper and raw data. Then the data is cleaned, reformatted and combined. During the process, the null values are cleaned, and the structure and relationship of the data is much clearer. Finally, we formed the conceptual model and draw the Entity Relationship graph to further clarify the model.

Contribution

Original contribution: 40% By External source: 20% Provided by the TA documents : 40%.

Citations

https://app.swaggerhub.com/apis-docs/tg4solutions/the-sneaker-database/1.0.0#/sneakers/getSneakers

http://www.thesneakerdatabase.com/

https://towardsdatascience.com/web-scraping-using-selenium-and-beautifulsoup-

99195cd70a58

http://unclechen.github.io/2016/12/11/python%E5%88%A9%E7%94%A8beautifulsoup+selenium%E8%87%AA%E5%8A%A8%E7%BF%BB%E9%A1%B5%E6%8A%93%E5%8F%96%E7%BD%91%E9%A1%B5%E5%86%85%E5%AE%B9/

License