PETER GLENN .COM

a scraping project
by: Ido Farhi

CUSTOMER LOGIN | ORDER STATUS | STORE LOCATIONS | HELP & SUPPORT | CART

Peter Glenn
OVER 50 YEARS OUTDOORS

SEARCH by keyword, product name or item #

1-800-818-0946

ORDERS OVER $49 SHIP FREE

MEN | WOMEN | KIDS | SKI | SNOWBOARD | RUN | TRAVEL | WATER | SKATES | BRANDS | SALE

CHECK OUT THE NEW SPRING ARRIVALS

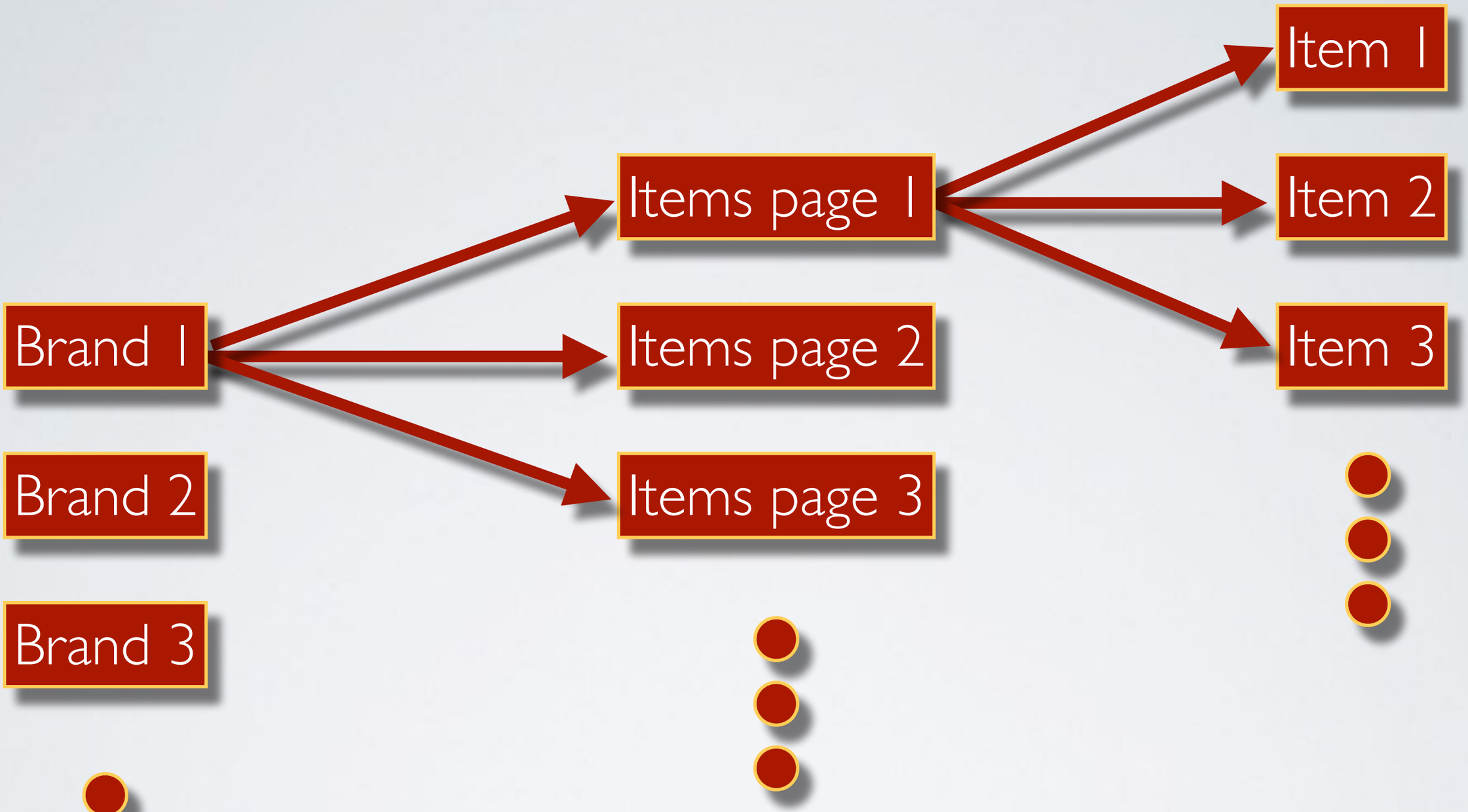
Peter Glenn
eGIFT CARD
THE PERFECT GIFT.
Delivered fast and easy.
BUY NOW

FFERS +

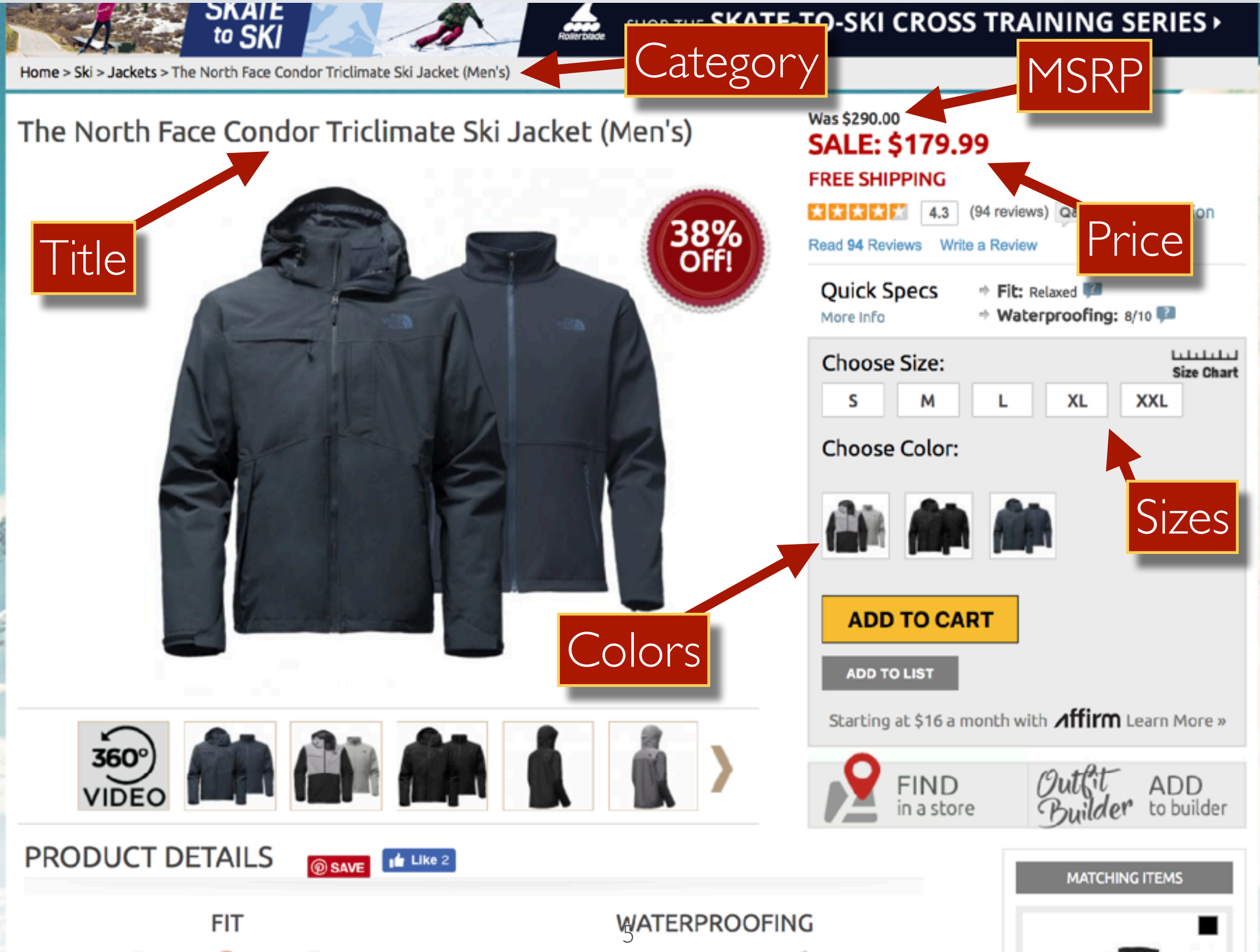
BRANDS

A	B	C	D
Abrams Books	Balega	Camaro	DAKINE
Adidas	Bearpaw	Camas Designs	Dalbello
AFRC	Bent Metal	Camelbak	DC Shoes
Ahnu	Bergans of Norway	Capita	Descente
Airblaster	Black Diamond Equipment	CEP	DOUBLE DIAMOND
Airhole	Blizzard	Coal	Dragon Goggles
Allsport	Blundstone	Colmar	Dynastar
Alp-n-Rock Clothing	Bogner	Columbia	
Alpinestars	Bomber Ski	CP	
Altra	BOS. & CO.	Crab Grab	
Altro Optics	Boulder Gear	crazeeHeads	
Amphipod	Brooks	CWB	
Anita	Brunton	CWX	
Anon	Bugatchi		
Apex Ski Boots	Bula		
ARBOR	Burton		
Arc'teryx			
Arcade			
Armada			
Astis			
Astral			
Atomic			

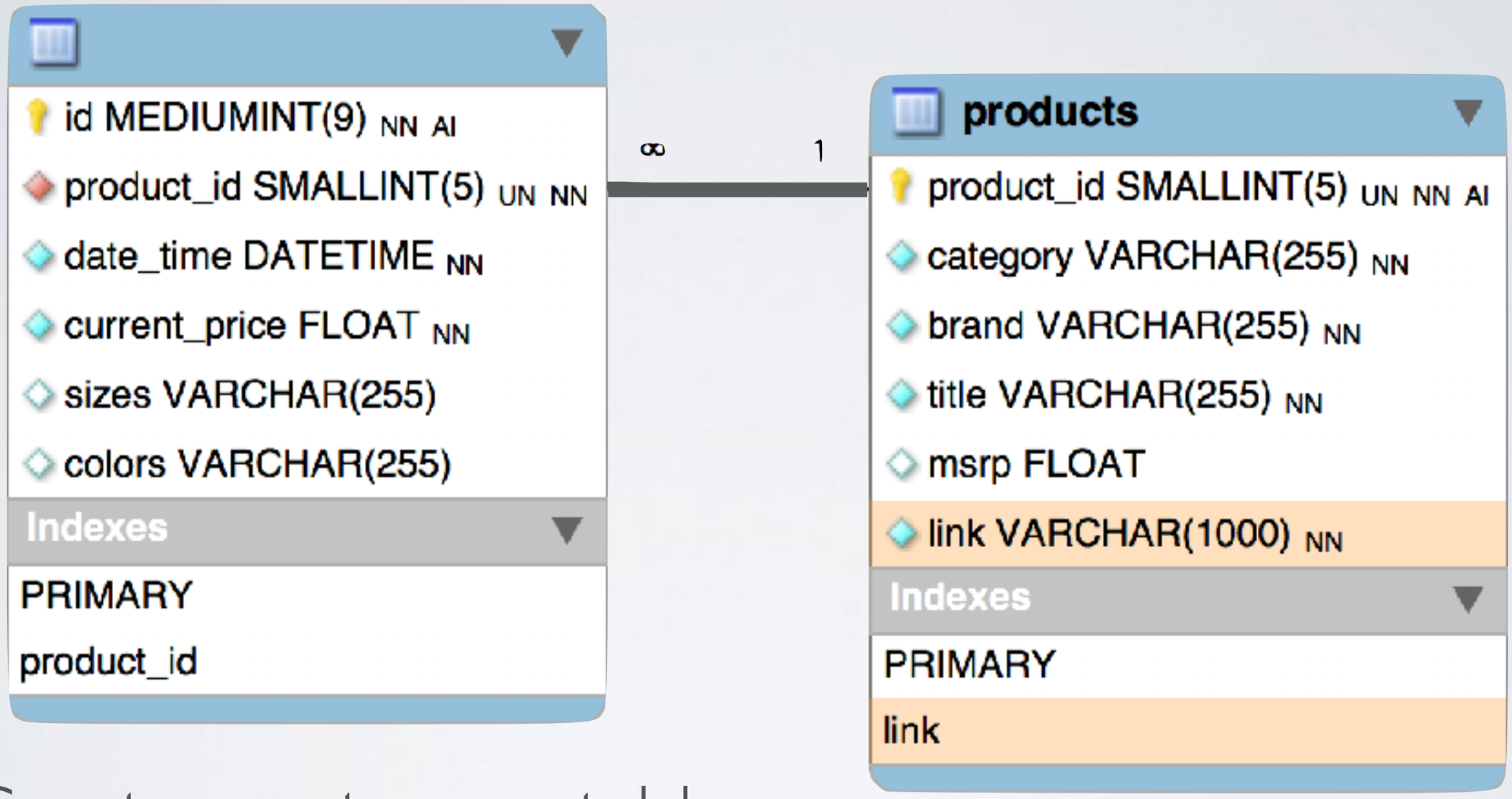
E	F	G	H
Eisbar	Feetures!	GIRO	Harmony
Elan	FERA	Globe	Head
Elan Blanc	Fire + Ice	GNU Snowboards	Headsweats
Electric	FISCHER	GOLDBERGH	Helly Hansen
Erin Snow	Fitletic	Goldwin	Hi-Tec
Euro Socks	Fjallraven	GoPole	High Range Gear


Brand 1
Brand 2
Brand 3
Items page 1
Items page 2
Items page 3
Item 1
Item 2
Item 3

Information scraped from each page



DATABASE (mySQL)



- Create a category table
- Create a brand table

SOME NUMBERS

- Distinct items scraped: ~5,500
- Rows in database: ~100,000
- Time to scrape each item: 2 seconds
- Total run time: ~3 hours … not reasonable

MULTI***PROCESSING***
VS.
MULTI***THREADING***

- **Process**: running code.
- A process can have multiple threads.
  They run the same code belonging to the parent process in parallel.
- Threads require fewer resources and are "cheaper" to create and run

MULTITHREADING

- Based on `concurrent.futures`
- Part of the standard library that provides a higher-level control over threads.
- The threads are modelled as asynchronous tasks.
- `concurrent.futures => threading => _thread`
- Each thread collects one brand at a time
- Total run time: ~20 minutes (~10 on AWS)

THE “ $ > 1,000 ” BUG

Home > Ski > Skis > Blizzard Quattro 8.4 TI Ski System with Bindings (Men's)

Blizzard Quattro 8.4 TI Ski System with Bindings (Men's)

33% OFF!

Was $1,080.00

SALE: $719.99

FREE SHIPPING

0.0 (No reviews) Q&A Ask a Question

Be the first to Write a Review

Quick Specs

More Info

- **Profile:** Rocker
- **Waist Width:** 84

Choose Size:

167 | 174 | 181

ADD TO CART

ADD TO LIST

Starting at $64 a month with Affirm Learn More »

FIND in a store

Outfit Builder ADD to builder

```
295         try:
296             # Catch exception in case no msrp found
297             msrp = float(soup.find('span', itemprop='price_msrp').text[1:])
298         except AttributeError:
299             # no different price & msrp. They are equal.
300             msrp = current_price
```

- Search for item msrp, if not found, set it equal to the current price.
- Can you guess what happens *when msrp is $1,000* or more ?
- How is multithreading related ?
- Then how was this bug caught ???
- Moral of the story…

CHALLENGES

- How to scrape the whole site ?
- Long I/O times
- Server uptime
- Uniqueness of data rows
-

FUTURE DEVELOPMENT

```
TODO: add user-definable options for scope of download
```

```
TODO: if title starts with brand name, remove it from title.
```

```
TODO: Sort companies by number items when scraping (shorter total time to scrape)
```

```
TODO: add categories table and brands table
```

```
TODO: Update products MSRP/TITLE/etc with new changes (provided same link)
```

```
TODO: Improve logging and usage of logging library
```

```
TODO: Add proper exception handling for concurrent futures library
```

```
TODO: Currently only takes lowest price. Add ability to pull price by color/size
```

USEFUL INSIGHTS

- Always comment !
  Explain WHY you wrote the code, not what it does
- Methods should be ~10 lines or less
- Use shortcuts (ctrl + B, cmd + [/] , …)
- Write good asserts - error report!

Dashboards Queries Alerts Create Search queries... Idodo

main

Refresh

Items for sale right now

items for sale

5,496
items for sale

3 days ago

Items added today

Items added today

6
Items added today

2 days ago

Items removed today

Items removed today

15
Items removed today

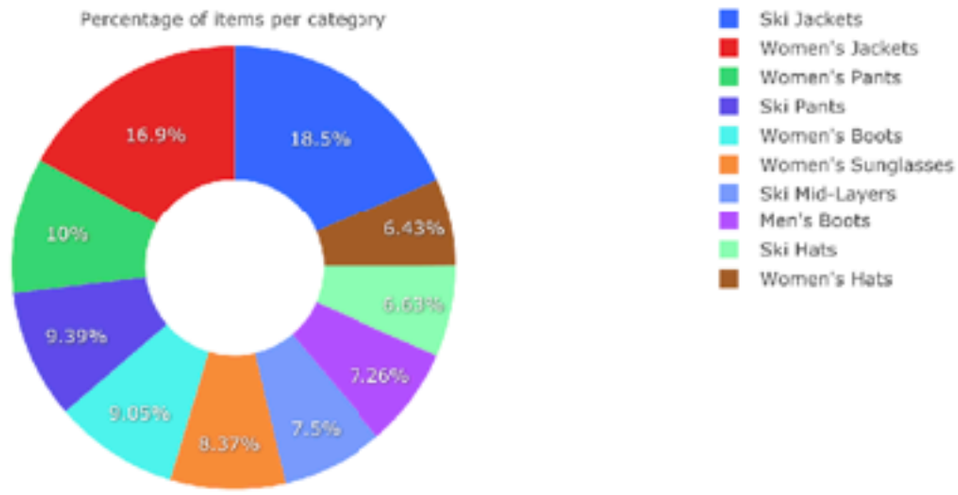
2 days ago

percentage products by category

Products by category

Number of categories

10



2 days ago

New items today

New items for sale today

title	brand	current_price	sizes	colors
anit				
Anita Air Control Padded Sports Bra (Women's)	Anita	$79.00	30,32,34,36,38,40,42	White,Anthracite,Pink
Anita Air Control Padded Sports Bra (Women's)	Anita	$79.00	30,32,34,36,38,40,42	White,Anthracite,Pink
Anita Dynamix Star Control Sports Bra (Women's)	Anita	$76.00	32,34,36,38,40	Peacock,White
Anita Dynamix Star Control Sports Bra (Women's)	Anita	$76.00	32,34,36,38,40	Peacock,White

2 days ago

Most discounted items

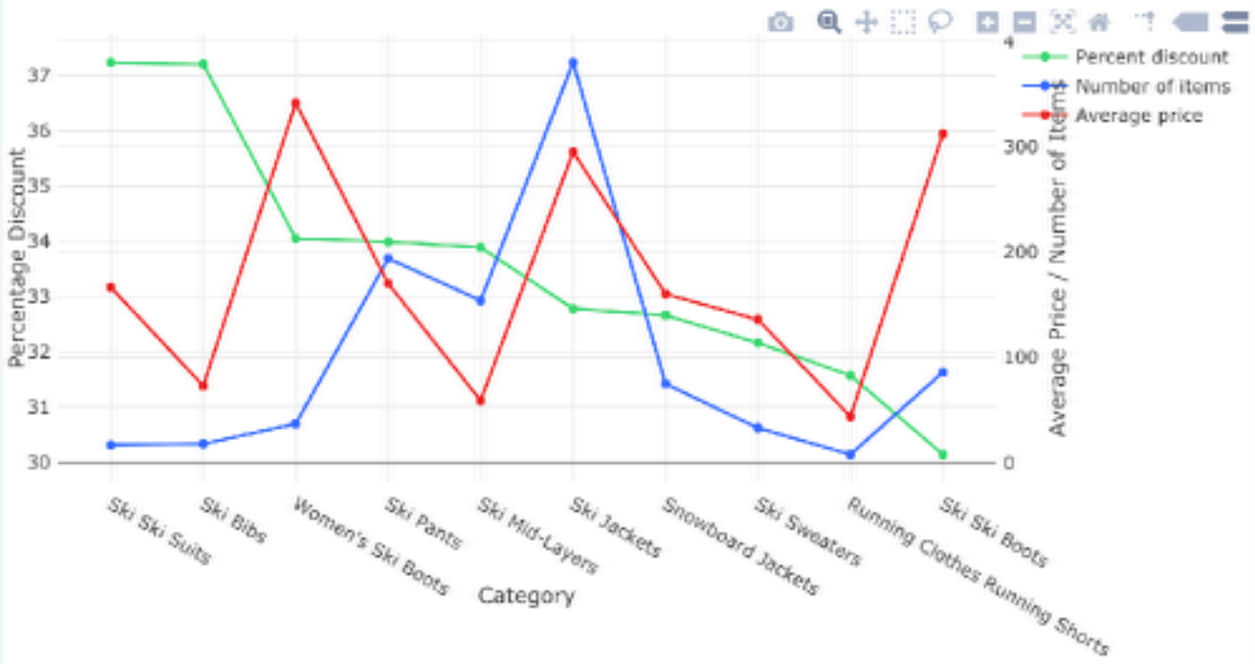
Highest discounts

2 days ago

Percent discount per category

Products by category

Sort by: Percent discount | Sort order: desc | Number OF categories: 10



2 days ago

Most discounted items

Highest discounts

discount: 70

Percent Discount	Title	Link
76.25	Obermeyer Sassy Knit Hat (Little Kids')	https://www.peterglenn.com/product/obermeyer-sassy-knit-hat-little-kids
75.02	Alpinestars Arubix Boardshorts (Men's)	https://www.peterglenn.com/product/alpinestars-arubix-boardshorts-mens
74.56	Marker Gillett Bib (Toddlers')	https://www.peterglenn.com/product/marker-gillett-bib-toddlers
74.28	Bugatchi 3/4 Coat (Men's)	https://www.peterglenn.com/product/bugatchi-34-coat-mens-0
73.34	Liquid Infinity Shell Snowboard Pant (Men's)	https://www.peterglenn.com/product/liquid-infinity-shell-snowboard-pant-mens
72.47	Montanaco Faux Lamb Ruffled Vest (Women's)	https://www.peterglenn.com/product/montanaco-faux-lamb-ruffled-vest-womens
71.42	Burton Base Camp Hybrid Short (Men's)	https://www.peterglenn.com/product/burton-base-camp-hybrid-short-mens
71.40	Mountain Force Rider III Insulated Ski Jacket (Women's)	https://www.peterglenn.com/product/mountain-force-rider-iii-insulated-ski-jacket-womens
71.32	Obermeyer Kismet Ski Jacket (Little Girls')	https://www.peterglenn.com/product/obermeyer-kismet-ski-jacket-little-girls

2 days ago

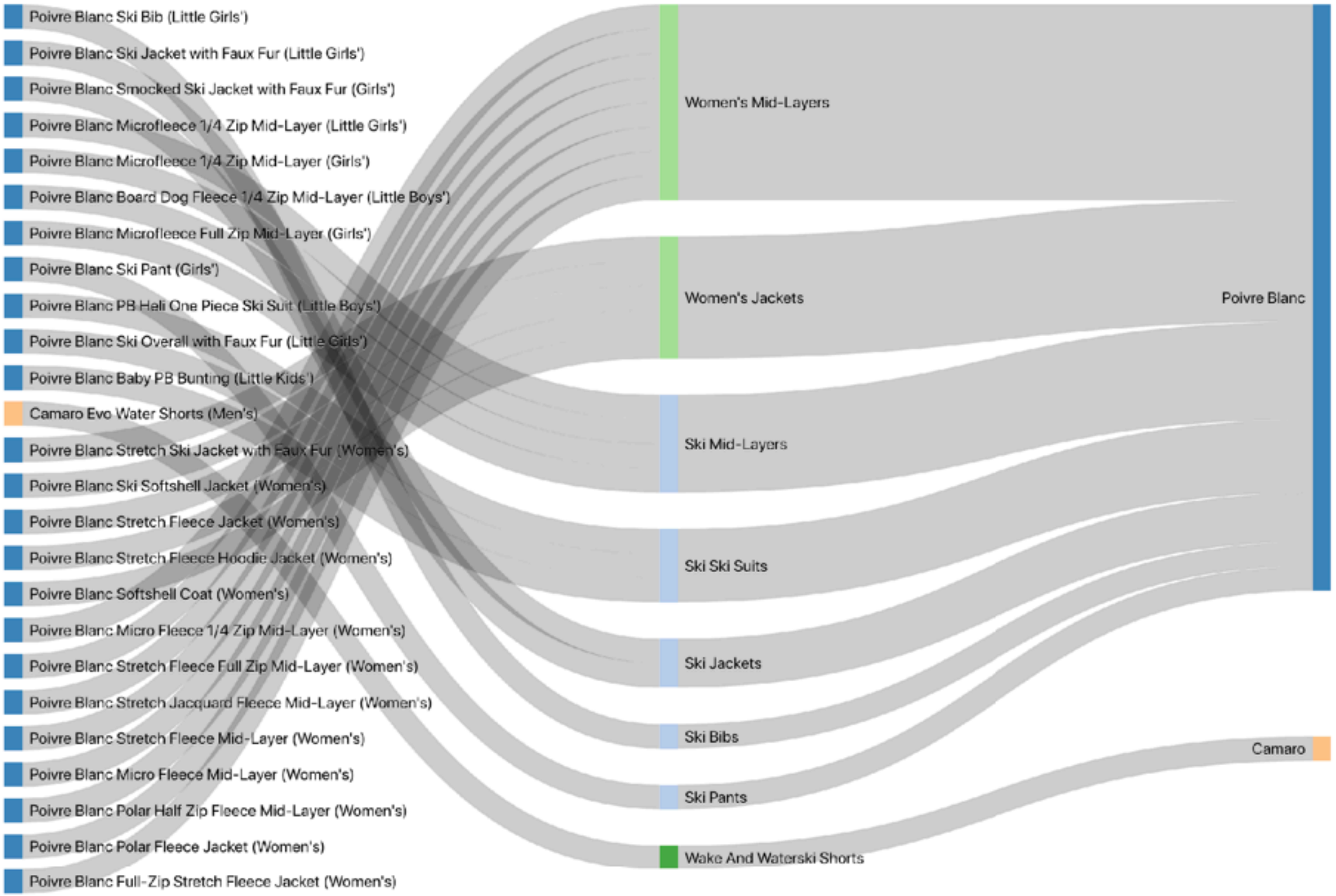
sankey visualization

Brand 1: Camaro

Brand 2: Poivre Blanc

Order by: stage2 asc

Number of items to show: 25



All items for sale right now

Items for sale today

category	brand	title	current_price	msrp	sizes	colors
north face						
Men's Jackets	Vans	Vans X The North Face Torrey Coach Jacket	$124.99	$179.00	S,M,L,XL	TNF Black,TNF Red TNF Black,TNF Yellow TNF Black
Kids' Boots	The North Face	The North Face Shellista Lace III Boot (Kids')	$64.99	$75.00	10,12,13	Tagumi Brown Barolo Red
Ski Pants	The North Face	The North Face Freedom Insulated Ski Pant (Girls')				
Kids' Boots	The North Face	The North Face Shellista Extreme Boot (Girls')	$59.99	$69.95	1,3	Dachsund Brown Moonlight Ivory,Dark Gull Grey Cerise Pink
Women's Baselayers	The North Face	The North Face Warm Long Sleeve Crew Neck (Women's)	$39.99	$50.00	M	TNF Black
Ski Pants	The North Face	The North Face Snowquest Triclimate Ski Pant (Boys')	$89.99	$119.95	M	Cosmic Blue
Ski Pants	The North Face	The North Face Mossbud Freedom Insulated Ski Pant (Girls')	$89.99	$109.95	S,XL	TNF Black
Ski Mid-Layers	The North Face	The North Face Sherparazo Fleece Jacket (Boys')	$69.99	$84.95	XL	TNF Black
Kids' Boots	The North Face	The North Face Thermoball Shellista Boot (Girls')	$69.99	$80.00	1,2,3,4,6,7	TNF Black
Kids' Boots	The North Face	The North Face Thermoball Shellista Boot (Little Girls')	$69.99	$80.00	11,12,13	TNF Black
Kids' Jackets	The North Face	The North Face Glacier Track Fleece Jacket (Girls')	$49.99	$54.95	S,M,L,XL	TNF Medium Grey Heather Collar Blue,Algiers Blue
Travel And Trail Backpacks	The North Face	The North Face Pinyon Backpack (Women's)	$59.99	$78.95	O/S	Magic Magenta
Ski Ski Suits	The North Face	The North Face Glacier One Piece Ski Suit (Little Kids')	$44.99	$49.95	12M,18M,24M,6M	Sky Blue Classic Camo Print,TNF Medium Grey Heather lilac Sachet
Ski Ski Suits	The North Face	The North Face Tailout One Piece Ski Suit (Little Kids')	$79.99	$98.95	0_3MO,3_6MO,6_12MO	Bright Cobalt Blue
Ski Ski Suits	The North Face	The North Face Tailout One Piece Ski Suit (Little Kids')	$79.99	$98.95	0_3MO,3_6MO	Bright Cobalt Blue

← 1 2 3 4 5 6 ... 23 →

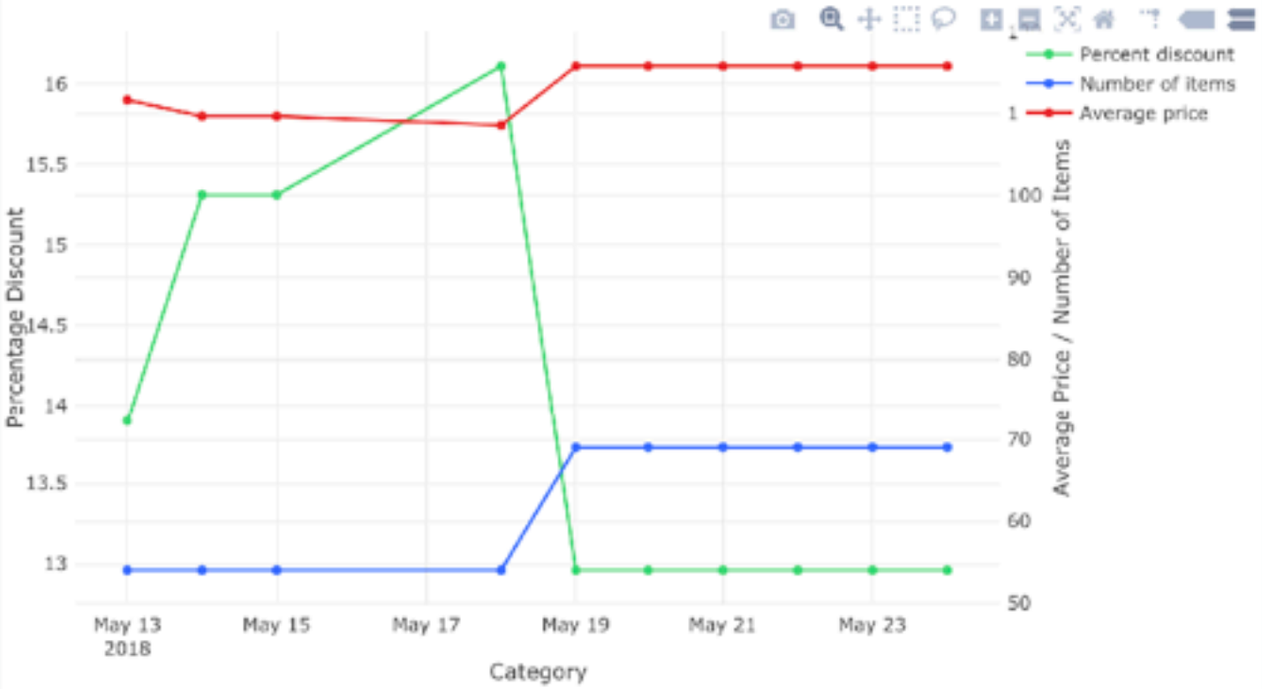
3 days ago

Category discount by date

Products by category

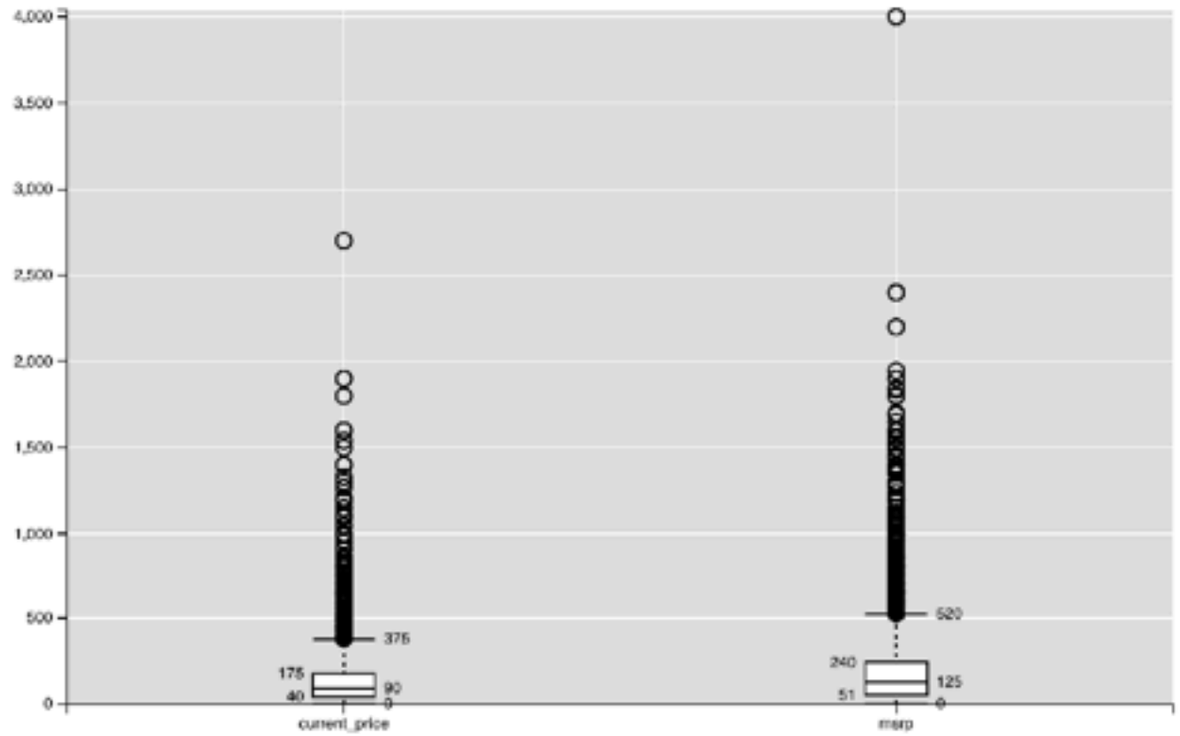
Category

Women's Running Shoes



2 days ago

Box plot distribution of price and MSRP right now



3 days ago