Amazon Fashion Discovery Engine.
Case study.
Used 183k results for women's tops
Results obtained using Amazon's API, not web scraping. → Against Amazon's policies.
Stuff is stored in "tops_fashion.json"
↳ just a format of storing data.

→ Descriptions are not used because they're lengthy, longer to process.
→ Didn't want students wait for long processing times.
→ Titles provide most of the useful value.
→ pandas syntax `x.head()`
↓ dataframe.
↳ gives first couple of values in df.
↳ 5 by default

→ out of the 19 features provided, we only use 6 features → ASIN (id)
→ brand
→ color
→ product_type_name
→ url of image of product
→ title of product → short & informative
↳ because of the way Amazon increments search results

→ There are some missing values due to human error → totally natural.
⇒ Data Cleaning & Understanding
→ overlooked & underappreciated.
→ helps understanding and clarify
→ takes up majority of time. during full time job.

→ `x["column"].describe()`
↓ df
↳ provides some quick metrics for that column.
↳ any column in df

```
count      183138
unique         72
top         SHIRT
freq       167794
Name: product_type_name, dtype: o
```
↳ that column name.
↳ the no. of values
↳ the freq. of top value.

→ `x["some_column"].unique()`
→ frequency of each product type.
`x.most_common()`

```
[('SHIRT', 167794),
 ('APPAREL', 3549),
 ('BOOKS_1973_AND_LATER', 3336),
 ('DRESS', 1584),
 ('SPORTING_GOODS', 1281),
 ('SWEATER', 837),
 ('OUTERWEAR', 796),
 ('OUTDOOR_RECREATION_PRODUCT', 729),
 ('ACCESSORY', 636),
 ('UNDERWEAR', 425)]
```
↳ big disparity btw 1st & 2nd
↳ curious how this turned up in women's tops.

Ask a question.
Data Understanding
figure out the ans.