FINAL PROJECT:

PROPOSAL FOR DATASET

Team Members:

Dhairya Dave (NUID: 002110382)

Manan Soni (NUID:002982645)

Parth Savaliya (NUID: 002982302)

Proposal of a Dataset

CONTEXT

This dataset is about **Product/Fashion recommendation** of H&M. It have 105542 rows and 25 variables about product, sections, description, color, etc. Another dataset related to this is customer like if customer has membership, what is age, etc. This dataset is officially given by H&M and it requires approx. 300 mb storage space. This is for proposal and understanding the research objectives. In nutshell, what we can do to H&M using this data. This is just short proposal for selection of dataset. As a perspective of analyst, this kind of data we will have to sort in future for company. Provided data is of **Sept 2018 to Sept 2020**. To add to this, Dataset can be used for fashion recommendation or to determine the taste of purchases.

CONTENT

- This dataset has these features for recommendation and understanding the products. This also includes different wears photos for business possibilities. Skewness is good according to number of samples.
- It has another dataset in relation of customer but it won't be that useful like we can get gender, age and membership to understand shopping taste from that of customers.

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
article_id*	1	105542	52771.50	30467.50	52771.5	52771.50	39119.14	1	105542	105541	0.00	-1.20	93.78
product_code*	2	105542	21933.44	13523.09	21426.0	21732.95	17044.71	1	47224	47223	0.09	-1.16	41.63
prod_name*	3	105542	23239.19	13373.17	23484.0	23305.22	17112.17	1	45875	45874	-0.05	-1.20	41.16
product_type_no	4	105542	234.86	75.05	259.0	246.63	19.27	-1	762	763	-1.42	1.17	0.23
product_type_name*	5	105542	74.47	36.38	88.0	77.04	37.06	1	131	130	-0.48	-1.22	0.11
product_group_name*	6	105542	8.72	3.73	9.0	8.75	1.48	1	19	18	0.02	0.93	0.01
graphical_appearance_no	7	105542	1009515.08	22413.59	1010016.0	1010012.91	1.48	-1	1010029	1010030	-45.02	2024.62	68.99
graphical_appearance_name*	8	105542	17.97	10.08	26.0	19.00	1.48	1	30	29	-0.69	-1.21	0.03
colour_group_code*	9	105542	22.14	13.84	15.0	21.13	10.38	1	50	49	0.53	-1.29	0.04
colour_group_name*	10	105542	18.09	16.14	15.0	16.34	17.79	1	50	49	0.72	-0.81	0.05
perceived_colour_value_id	11	105542	3.21	1.56	4.0	3.12	1.48	-1	7	8	0.27	-0.09	0.00
perceived_colour_value_name*	12	105542	3.09	1.45	3.0	2.94	1.48	1	8	7	0.79	-0.41	0.00
perceived_colour_master_id	13	105542	7.81	5.38	5.0	7.18	4.45	-1	20	21	0.80	-0.36	0.02
perceived_colour_master_name*	14	105542	7.93	6.08	7.0	7.45	7.41	1	20	19	0.52	-1.24	0.02
department_no	15	105542	4532.78	2712.69	4222.0	4387.86	3789.53	1201	9989	8788	0.27	-1.40	8.35
department_name*	16	105542	122.39	70.50	115.0	121.01	87.47	1	250	249	0.18	-1.12	0.22
index_code*	17	105542	4.46	2.73	4.0	4.31	4.45	1	10	9	0.21	-1.10	0.01
index_name*	18	105542	5.63	2.48	6.0	5.73	2.97	1	10	9	-0.25	-0.84	0.01
index_group_no	19	105542	3.17	4.35	2.0	2.46	1.48	1	26	25	4.59	21.33	0.01
index_group_name*	20	105542	2.38	1.15	3.0	2.31	1.48	1	5	4	0.19	-0.96	0.00
section_no	21	105542	42.66	23.26	46.0	42.72	29.65	2	97	95	-0.08	-1.10	0.07
section_name*	22	105542	29.71	17.63	28.0	29.98	23.72	1	56	55	-0.04	-1.40	0.05
garment_group_no	23	105542	1010.44	6.73	1009.0	1010.19	8.90	1001	1025	24	0.32	-1.29	0.02
garment_group_name*	24	105542	9.36	6.02	7.0	9.11	5.93	1	21	20	0.44	-0.98	0.02
detail_desc*	25	105126	22216.17	12655.24	22240.5	22344.77	16460.57	1	43404	43403	-0.05	-1.22	39.03

RESEARCH OBJECTIVES

1. Determination of customer and their status of membership with age.
2. Each items has department plus section with color options we can get which product is most repetitive with respect to other variable like product type?
3. How much men products and women products ration in this data?
4. To be specific, in women how many percentages of different product type? Which product has significant numbers.
5. This is overview so if we get transactions, we can go much deeper for great business solution.
6. We can get the production relative frequencies about colors and type as we have different variables.

SOURCE

Kaggle: https://www.kaggle.com/competitions/h-and-m-personalized-fashion-recommendations/data