H&M Database System

Group 3:
Zhikun Chen, Jinglin He, Xinlei He, Wan Ting Hsu, Sepideh Kiumarsi, Yuexin Li, Yuting Wang

Background

H&M is the world's second-largest clothing retailer

- Operates in more than 74 countries with more than 5,000 stores worldwide
- Closed 5% of stores in 2021 because of COVID

Shift from traditional retail settings to eCommerce

Problem

Lots of competition within the online retail space

- H&M needs to differentiate itself

Different consumer expectations in eCommerce that H&M needs to adapt to

Create the best online shopping experience for shoppers

Our Proposed Solution

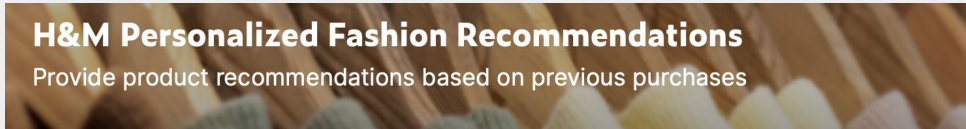
To help H&M more easily access consumer data we have created an organized database that can be used to understand their customers and ultimately develop product recommendation systems

Original Data Description

- **Data Source**: Kaggle
- **Data Name**: H&M Personalized Fashion Recommendation datasets
  - Article csv file
  - Customer csv file
  - Transaction csv file
- **Data Provider**: H&M Group
- **Data Access Link**: https://www.kaggle.com/competitions/h-and-m-personalized-fashion-recommendations/data

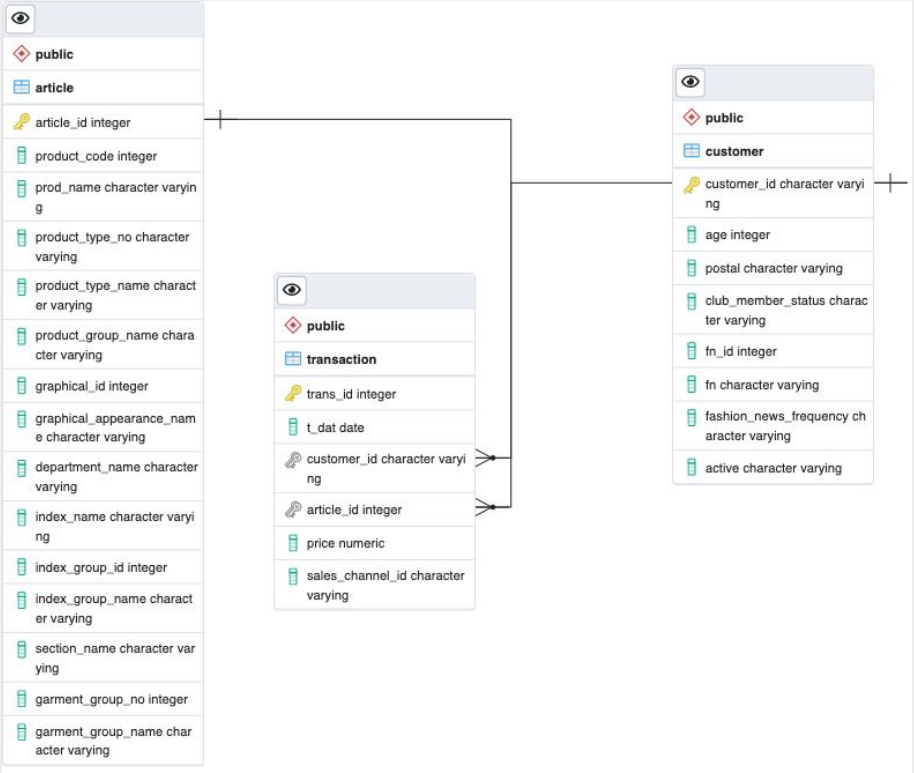
Key data information:

- **Article file**
  - Article id, product name, product type, graphical appearance, color, department etc.
- **Customer file**
  - Customer id, customer active status, customer membership status, age.
- **Transaction file**
  - Transaction date, customer id, article article (products purchased), price, sales channel



Normalization - 1NF

- Each table cell should contain a single value.
- Each record needs to be unique.



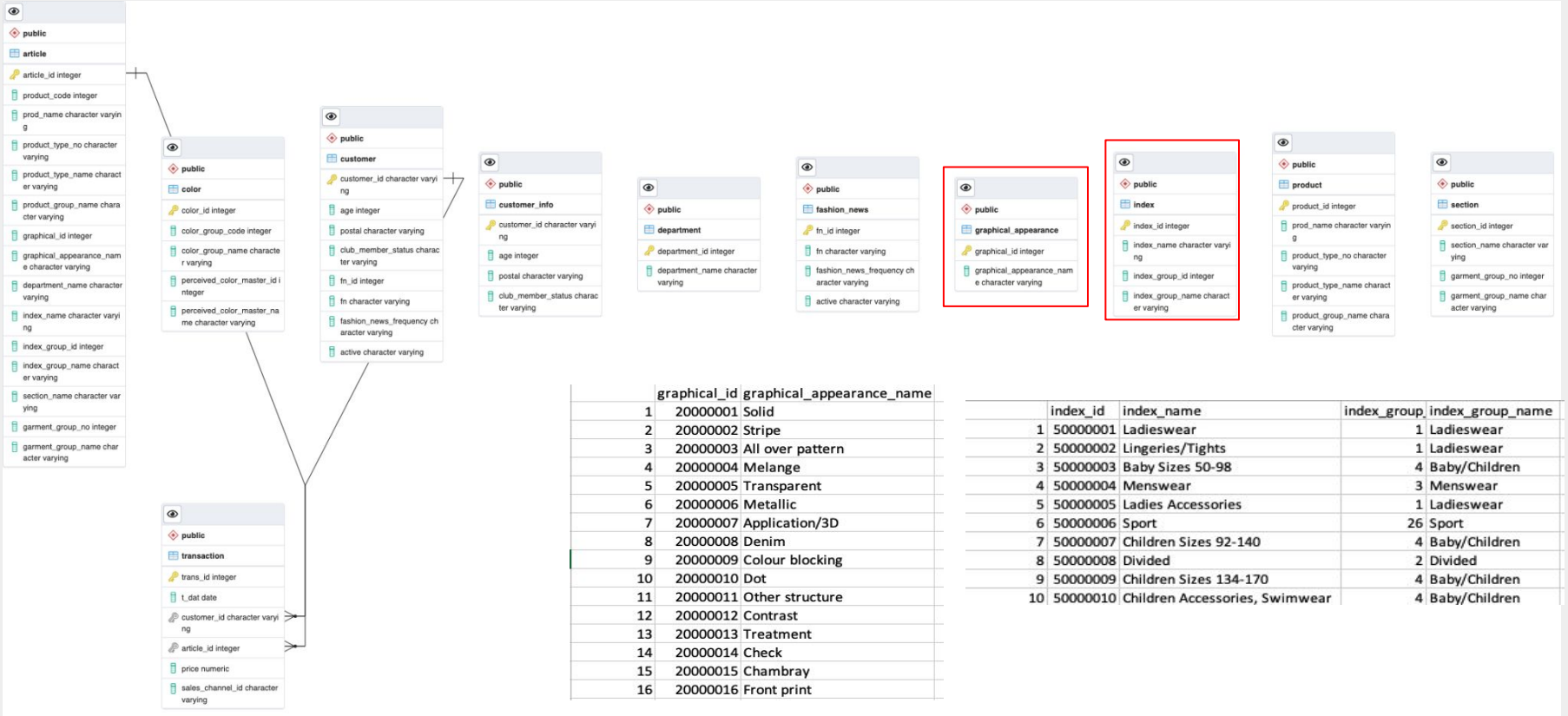
article_id	product_code	prod_name	product_type_no	product_type_name	product_group_name	graphical_appearance_no	graphical_appearance_name	colour_group_code	colour_group_name	perceived_colour_value_id	perceived_colour
108775015	108775	Strap top	253	Vest top	Garment Upper body	1010016	Solid	9	Black	4	Dark
108775044	108775	Strap top	253	Vest top	Garment Upper body	1010016	Solid	10	White	3	Light
108775051	108775	Strap top (1)	253	Vest top	Garment Upper body	1010017	Stripe	11	Off White	1	Dusty Light
110065001	110065	OP T-shirt (Idro)	306	Bra	Underwear	1010016	Solid	9	Black	4	Dark
110065002	110065	OP T-shirt (Idro)	306	Bra	Underwear	1010016	Solid	10	White	3	Light
110065011	110065	OP T-shirt (Idro)	306	Bra	Underwear	1010016	Solid	12	Light Beige	1	Dusty Light
111565001	111565	20 den 1p Stockings	304	Underwear Tights	Socks & Tights	1010016	Solid	9	Black	4	Dark
111565003	111565	20 den 1p Stockings	302	Socks	Socks & Tights	1010016	Solid	13	Beige	2	Medium Dusty
111586001	111586	Shape Up 30 den 1p Tights	273	Leggings/Tights	Garment Lower body	1010016	Solid	9	Black	4	Dark
111593001	111593	Support 40 den 1p Tights	304	Underwear Tights	Socks & Tights	1010016	Solid	9	Black	4	Dark
111609001	111609	200 den 1p Tights	304	Underwear Tights	Socks & Tights	1010016	Solid	9	Black	4	Dark
112679048	112679	SWEATSHIRT OC	252	Sweater	Garment Upper body	1010001	All over pattern	7	Grey	1	Dusty Light
112679052	112679	SWEATSHIRT OC	252	Sweater	Garment Upper body	1010001	All over pattern	71	Light Blue	1	Dusty Light
114428026	114428	Alice BANDEAU 2-p	306	Bra	Underwear	1010017	Stripe	10	White	3	Light

customer_id	FN	Active	club_member_status	fashion_news_frequency	age
00000dbacae5abe5e23885899a1fa44253a17956c6d1c3d25f88aa139fdfc657			ACTIVE	NONE	49
0000423b00ade91418cceaf3b26c6af3dd342b51fd051eec9c12fb36984420fa			ACTIVE	NONE	25
000058a12d5b43e67d225668fa1f8d618c13dc232df0cad8ffe7ad4a1091e318			ACTIVE	NONE	24
00005ca1c9ed5f5146b52ac8639a40ca9d57aeff4d1bd2c5feb1ca5dff07c43e			ACTIVE	NONE	54
00006413d8573cd20ed7128e53b7b13819fe5cfc2d801fe7fc0f26dd8d65a85a	1.0	1.0	ACTIVE	Regularly	52
000064249685c11552da43ef22a5030f35a147f723d5b02ddd9fd22452b1f5a6					
0000757967448a6cb83efb3ea7a3fb9d418ac7adf2379d8cd0c725276a467a2a			ACTIVE	NONE	20
00007d2de826758b65a93dd24ce629ed66842531df6699338c5570910a014cc2	1.0	1.0	ACTIVE	Regularly	32

t_dat	customer_id	article_id	price	sales_channel_id
9/20/18	000058a12d5b43e67d225668fa1f8d618c13dc232df0cad8ffe7ad4a1091e318	663713001	0.050830508	2
9/20/18	000058a12d5b43e67d225668fa1f8d618c13dc232df0cad8ffe7ad4a1091e318	541518023	0.030491525	2
9/20/18	00007d2de826758b65a93dd24ce629ed66842531df6699338c5570910a014cc2	505221004	0.015237288	2
9/20/18	00007d2de826758b65a93dd24ce629ed66842531df6699338c5570910a014cc2	685687003	0.016932203	2
9/20/18	00007d2de826758b65a93dd24ce629ed66842531df6699338c5570910a014cc2	685687004	0.016932203	2
9/20/18	00007d2de826758b65a93dd24ce629ed66842531df6699338c5570910a014cc2	685687001	0.016932203	2
9/20/18	00007d2de826758b65a93dd24ce629ed66842531df6699338c5570910a014cc2	505221001	0.020322034	2
9/20/18	00083cda041544b2fbb0e0d2905ad17da7cf1007526fb4c73235dccbbc132280	688873012	0.030491525	1
9/20/18	00083cda041544b2fbb0e0d2905ad17da7cf1007526fb4c73235dccbbc132280	501323011	0.053372881	1

Normalization - 2NF

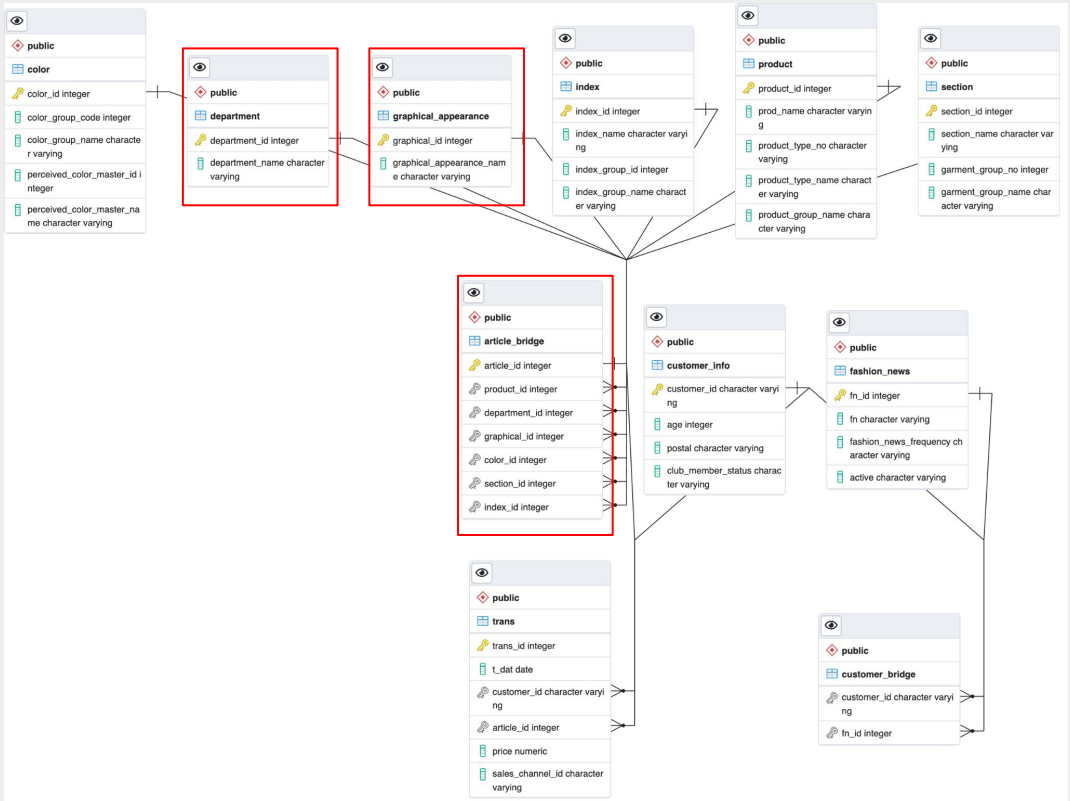
- The table must be already in 1 NF and all non-key columns of the tables must depend on the PRIMARY KEY
- The partial dependencies are removed and placed in a separate table



	graphical_id	graphical_appearance_name
1	20000001	Solid
2	20000002	Stripe
3	20000003	All over pattern
4	20000004	Melange
5	20000005	Transparent
6	20000006	Metallic
7	20000007	Application/3D
8	20000008	Denim
9	20000009	Colour blocking
10	20000010	Dot
11	20000011	Other structure
12	20000012	Contrast
13	20000013	Treatment
14	20000014	Check
15	20000015	Chambray
16	20000016	Front print

	index_id	index_name	index_group_	index_group_name
1	50000001	Ladieswear	1	Ladieswear
2	50000002	Lingeries/Tights	1	Ladieswear
3	50000003	Baby Sizes 50-98	4	Baby/Children
4	50000004	Menswear	3	Menswear
5	50000005	Ladies Accessories	1	Ladieswear
6	50000006	Sport	26	Sport
7	50000007	Children Sizes 92-140	4	Baby/Children
8	50000008	Divided	2	Divided
9	50000009	Children Sizes 134-170	4	Baby/Children
10	50000010	Children Accessories, Swimwear	4	Baby/Children

Normalization - 3NF
Finalized ERD



- Non-Primary key columns shouldn't depend on the other non-Primary key columns
- There is no transitive functional dependency

article_id	colour_id	department_	graphical_id	index_id	product_id	section_id	articlebridge
108775015	41000001	31000001	20000001	50000001	10000001	60000001	108775015
108775044	41000002	31000001	20000001	50000001	10000001	60000001	108775044
108775051	41000003	31000001	20000002	50000001	10000002	60000001	108775051
110065001	41000001	31000002	20000001	50000002	10000003	60000002	110065001
110065002	41000002	31000002	20000001	50000002	10000003	60000002	110065002
110065011	41000004	31000002	20000001	50000002	10000003	60000002	110065011
111565001	41000001	31000003	20000001	50000002	10000004	60000003	111565001
111586001	41000001	31000003	20000001	50000002	10000006	60000003	111586001
111593001	41000001	31000003	20000001	50000002	10000007	60000003	111593001
111609001	41000001	31000003	20000001	50000002	10000008	60000003	111609001
116379047	41000009	31000001	20000001	50000001	10000011	60000001	116379047
118458004	41000011	31000001	20000004	50000004	10000012	60000005	118458004
118458029	41000010	31000001	20000004	50000004	10000012	60000005	118458029
118458038	41000013	31000001	20000004	50000004	10000012	60000005	118458038

department_id	department_name
31000001	Jersey Basic
31000002	Clean Lingerie
31000003	Tights basic
31000004	Baby basics
31000005	Casual Lingerie
31000007	Jersey
31000008	EQ & Special Collections
31000009	Hair Accessories

graphical_id	graphical_appearance_name
20000001	Solid
20000002	Stripe
20000003	All over pattern
20000004	Melange
20000005	Transparent
20000006	Metallic
20000007	Application/3D
20000008	Denim
20000009	Colour blocking

ETL Process

1. Clean tables by manipulating the data to only the attributes we want for the table. As you can see there are 105,542 rows for the color table.

2. Find only the distinct rows. We eliminated the duplicate rows. Now we have 296 distinct color combination rows.

3. Create a primary key (ID) for each table.

4. Generate bridge tables with primary keys from related tables.

1

```r
colour <- articles %>% select(colour_group_code, colour_group_name, perceived_colour_value_id, perceived_colour_value_name, perceived_colour_master_id,
perceived_colour_master_name)
colour
```

Description: df [105,542 × 6]

colour_group_code <int>	colour_group_name <chr>	perceived_colour_value_id <int>	perceived_colour_value_name <chr>	perceived_colour_master_id <int>	perceived_colour_master_name <chr>
9	Black	4	Dark	5	Black
10	White	3	Light	9	White
11	Off White	1	Dusty Light	9	White
9	Black	4	Dark	5	Black
10	White	3	Light	9	White
12	Light Beige	1	Dusty Light	11	Beige
9	Black	4	Dark	5	Black
13	Beige	2	Medium Dusty	11	Beige
9	Black	4	Dark	5	Black
9	Black	4	Dark	5	Black

1–10 of 105,542 rows

Previous 1 2 3 4 5 6 … 100 Next

2

```r
colour <- unique(colour)
colour
```

Description: df [296 × 6]

	colour_group_code <int>	colour_group_name <chr>	perceived_colour_value_id <int>	perceived_colour_value_name <chr>	perceived_colour_master_id <int>
1	9	Black	4	Dark	5
2	10	White	3	Light	9
3	11	Off White	1	Dusty Light	9
6	12	Light Beige	1	Dusty Light	11
8	13	Beige	2	Medium Dusty	11
12	7	Grey	1	Dusty Light	12
13	71	Light Blue	1	Dusty Light	2
15	6	Light Grey	1	Dusty Light	12
16	73	Dark Blue	4	Dark	2
17	8	Dark Grey	4	Dark	12

1–10 of 296 rows | 1–6 of 6 columns

Previous 1 2 3 4 5 6 … 30 Next

3

```r
colour = colour %>% mutate(colour_id = sub('^', '4', row_number()+ 1000000))
colour
```

Description: df [296 × 7]

	colour_group_code <int>	colour_group_name <chr>	perceived_colour_value_id <int>	perceived_colour_value_name <chr>	perceived_colour_master_id <int>	perceived_colour_master_name <chr>	colour_id <chr>
1	9	Black	4	Dark	5	Black	41000001
2	10	White	3	Light	9	White	41000002
3	11	Off White	1	Dusty Light	9	White	41000003
6	12	Light Beige	1	Dusty Light	11	Beige	41000004
8	13	Beige	2	Medium Dusty	11	Beige	41000005
12	7	Grey	1	Dusty Light	12	Grey	41000006
13	71	Light Blue	1	Dusty Light	2	Blue	41000007
15	6	Light Grey	1	Dusty Light	12	Grey	41000008
16	73	Dark Blue	4	Dark	2	Blue	41000009
17	8	Dark Grey	4	Dark	12	Grey	41000010

1–10 of 296 rows

Previous 1 2 3 4 5 6 … 30 Next

Analytical Procedure

Why do we interact with data?

- Data is not always in the format as wish
- Employees who have different job responsibilities have different requirements for data analysis results

Customer Needs

- Fast, easy and convenient
- Getting a better understanding of the customers profile and demand for marketing campaign
- Different requirements for data

Results and Benefits

- Ops team: To know better about the fashion and trends in the market and make appropriate designing and manufacturing decisions
- C-level: To obtain business insights and support decision-making procedure through data analysis.

Tools & Techniques

We have multiple front-end and back-end interaction tools for our SQL database, such as Python, PGAdmin, and Metabase. Each tool has its own feature.

- **Python** is a front end programming language to help with data visualization, ETL process before data ingestion, and data analytics. Psycopg2 is the library we used for connection to our SQL server. We can use Hadoop for HDFS and Spark for ML analytics.

- **pgAdmin** is the platform for our SQL server. Our clients or our product manager can directly interact with our database using query language on pgAdmin.

- **Metabase** is the automated data visualization tool for our SQL database. Our data security is ensured with limited views for different levels of analyst and manager. Different positions within the company can view different visualizations and it is always updated.

Database Interaction Demo

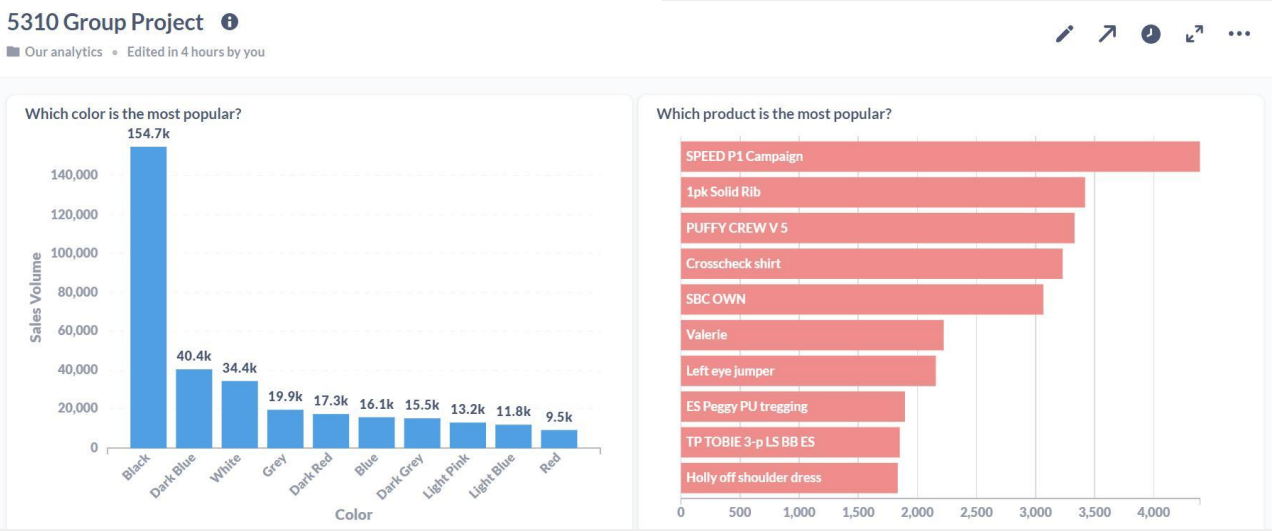
Query Editor | Query History

```
28  Question4: Which color is the most popular?
29  SELECT c.color_group_name, c.perceived_color_master_name, COUNT(c.color_group_name) AS number_of_color
30  FROM article_bridge AS a
31  INNER JOIN color AS c
32  ON a.color_id = c.color_id
33  INNER JOIN trans AS t
34  ON a.article_id = t.article_id
35  GROUP BY c.color_group_name, c.perceived_color_master_name
36  ORDER BY number_of_color DESC;
37
```

Data Output | Explain | Messages | Notifications

	color_group_name character varying	perceived_color_master_name character varying	number_of_color bigint
1	Black	Black	154721
2	Dark Blue	Blue	40383
3	White	White	34417
4	Grey	Grey	19897
5	Dark Red	Red	17283
6	Blue	Blue	16104
7	Dark Grey	Grey	15549
8	Light Pink	Pink	13205

http://160.39.250.97:3000/public/dashboard/e1077600-e5b2-485d-b939-e730658f1e3f



Conclusion

- Gain deeper business insights from the results generated by data analytics
  - more targeted marketing strategy and campaign based on the data analytical results.

- Gain better insight of customer behaviors, and helps H&M decision making process
  - eg. most popular color and graphical print among customers

- Provide information and experience for ETL procedure implementation in other department (e.g accounting, HR, public relations), improving company operation efficiency.

Thank you!