In [1]:
```
# H & M Sales Insights
# The following project aims to analyze and draw insights from the data set.
# The dataset contains information about sales and customer information.
# we will clean and transform necessary data in python and visualize using Power Bi
```

In [2]:
```
# The following are the keys problems statements to be answered...
# 1.Overview of sales during 2018-2020
# 2.Which age customers buys the most?
# 3.What are the top 5 selling products ?
# 4.Does membership impacts the sales of the company?
# 4.What are the products that brings more revenue?
# 5.What age group of people is interested in purchasing?
# 6.What age customers are interested in club memberships?
# 7.Does Fashion news delivered to customers converts into sales?
# 8.What products contribute the least sales?
# 9.How can we improve the sales of least revenue generated products?
# 10.Provide recommendations to improve sales from least purchased age group?
```

In [3]:
```
# To begin with the analysis let's start to understand the dataset..
```

In [4]:
```
# There are 3 datasets available for the analysis..
# 1.Articles    - contains data about product information
# 2.customers   - contains data about customer information
# 3.transaction - contains data about purchase information
```

In [5]:
```
#importing necessary packages and libraries

import pandas as pd
from pandasql import sqldf
```

In [6]:
```
# importing datasets

df_articles = pd.read_csv("D:/Data Analyst Projects/H & M sales insights/articles/articles.csv")
df_customers = pd.read_csv("D:/Data Analyst Projects/H & M sales insights/customers/customers.csv")
df_transactions = pd.read_csv("D:/Data Analyst Projects/H & M sales insights/transactions/transactions.
```

In [7]:
```
# viewing sample data
df_articles.head()
```

Out[7]:

	article_id	product_code	prod_name	product_type_no	product_type_name	product_group_name	graphical_appearance_no
0	108775015	108775	Strap top	253	Vest top	Garment Upper body	1010016
1	108775044	108775	Strap top	253	Vest top	Garment Upper body	1010016
2	108775051	108775	Strap top (1)	253	Vest top	Garment Upper body	1010017
3	110065001	110065	OP T-shirt (Idro)	306	Bra	Underwear	1010016
4	110065002	110065	OP T-shirt (Idro)	306	Bra	Underwear	1010016

5 rows × 25 columns

In [11]:
```
# viewing sample data
df_customers.head()
```

Out[11]:

	customer_id	FN	Active	club_member_status	fashion_news_frequency	age	
0	00000dbacae5abe5e23885899a1fa44253a17956c6d1c3...	NaN	NaN	ACTIVE	NONE	49.0	52043
1	0000423b00ade91418cceaf3b26c6af3dd342b51fd051e...	NaN	NaN	ACTIVE	NONE	25.0	2973a
2	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	NaN	NaN	ACTIVE	NONE	24.0	64f17
3	00005ca1c9ed5f5146b52ac8639a40ca9d57aeff4d1bd2...	NaN	NaN	ACTIVE	NONE	54.0	5d365
4	00006413d8573cd20ed7128e53b7b13819fe5cfc2d801f...	1.0	1.0	ACTIVE	Regularly	52.0	25fa5d

In [12]:

```
# viewing sample data
df_transactions.head()
```

Out[12]:

	t_dat	customer_id	article_id	price	sales_channel_id
0	2018-09-20	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	663713001	0.050831	2
1	2018-09-20	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	541518023	0.030492	2
2	2018-09-20	00007d2de826758b65a93dd24ce629ed66842531df6699...	505221004	0.015237	2
3	2018-09-20	00007d2de826758b65a93dd24ce629ed66842531df6699...	685687003	0.016932	2
4	2018-09-20	00007d2de826758b65a93dd24ce629ed66842531df6699...	685687004	0.016932	2

In [13]:

```
# Preparing the data
# we will filter out the columns which are necessary for analysis in articles dataframe

df_articles = sqldf("""
Select
 article_id
,prod_name
,product_type_name
,product_group_name
,colour_group_name
,index_name
from df_articles
""")
```

In [14]:

```
# No null values

df_articles.isna().sum()
```

Out[14]:

```
article_id            0
prod_name             0
product_type_name     0
product_group_name    0
colour_group_name     0
index_name            0
dtype: int64
```

In [15]:

```
# checking for duplicate rows

df_articles.duplicated().sum()
```

Out[15]:

```
0
```

In [16]:

```
df_articles.to_csv("D:/Data Analyst Projects/H & M sales insights/df_articles.csv")
```

In [17]:

```
# we will filter out the columns which are necessary for analysis in articles dataframe

df_customers = sqldf ("""
select
customer_id
,club_member_status
,fashion_news_frequency
,age
from df_customers
""")
```

In [18]:

```
# Cheking null values

df_customers.isna().sum()
```

Out[18]:

```
customer_id                   0
club_member_status         6062
fashion_news_frequency    16011
age                       15861
dtype: int64
```

In [19]:

```
# We will find the no of records present and decide to drop or fill null values..

len(df_customers)
```

Out[19]:

```
1371980
```

In [20]:
```
#We will drop the records since the records count is high..

df_customers.dropna()
```

Out[20]:

	customer_id	club_member_status	fashion_news_frequency	age
0	00000dbacae5abe5e23885899a1fa44253a17956c6d1c3...	ACTIVE	NONE	49.0
1	0000423b00ade91418cceaf3b26c6af3dd342b51fd051e...	ACTIVE	NONE	25.0
2	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	ACTIVE	NONE	24.0
3	00005ca1c9ed5f5146b52ac8639a40ca9d57aeff4d1bd2...	ACTIVE	NONE	54.0
4	00006413d8573cd20ed7128e53b7b13819fe5cfc2d801f...	ACTIVE	Regularly	52.0
...	...	...	...	...
1371975	ffffbbf78b6eaac697a8a5dfbfd2bfa8113ee5b403e474...	ACTIVE	NONE	24.0
1371976	ffffcd5046a6143d29a04fb8c424ce494a76e5cdf4fab5...	ACTIVE	NONE	21.0
1371977	ffffcf35913a0bee60e8741cb2b4e78b8a98ee5ff2e6a1...	ACTIVE	Regularly	21.0
1371978	ffffd7744cebcf3aca44ae7049d2a94b87074c3d4ffe38...	ACTIVE	Regularly	18.0
1371979	ffffd9ac14e89946416d80e791d064701994755c3ab686...	PRE-CREATE	NONE	65.0

In [21]:
```
# checking for duplicate rows

df_customers.duplicated().sum()
```

Out[21]:
```
0
```

In [22]:
```
# No null values

df_transactions.isna().sum()
```

Out[22]:
```
t_dat               0
customer_id         0
article_id          0
price               0
sales_channel_id    0
dtype: int64
```

In [23]:
```
# checking for duplicate rows

df_transactions.duplicated().sum()
```

Out[23]:
```
2974905
```

In [24]:
```
# we will drop the duplicate rows

df_transactions.drop_duplicates()
```

Out[24]:

	t_dat	customer_id	article_id	price	sales_channel_id
0	2018-09-20	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	663713001	0.050831	2
1	2018-09-20	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	541518023	0.030492	2
2	2018-09-20	00007d2de826758b65a93dd24ce629ed66842531df6699...	505221004	0.015237	2
3	2018-09-20	00007d2de826758b65a93dd24ce629ed66842531df6699...	685687003	0.016932	2
4	2018-09-20	00007d2de826758b65a93dd24ce629ed66842531df6699...	685687004	0.016932	2
...	...	...	...	...	...
31788319	2020-09-22	fff2282977442e327b45d8c89afde25617d00124d0f999...	929511001	0.059305	2
31788320	2020-09-22	fff2282977442e327b45d8c89afde25617d00124d0f999...	891322004	0.042356	2
31788321	2020-09-22	fff380805474b287b05cb2a7507b9a013482f7dd0bce0e...	918325001	0.043203	1
31788322	2020-09-22	fff4d3a8b1f3b60af93e78c30a7cb4cf75edaf2590d3e5...	833459002	0.006763	1
31788323	2020-09-22	fffef3b6b73545df065b521e19f64bf6fe93bfd450ab20...	898573003	0.033881	2

28813419 rows × 5 columns

In [25]:
```
#filtering out necessary data

df_transactions = sqldf("""
select
customer_id
,article_id
,price
,t_dat
from df_transactions
""")
```

In [26]:

```
# Export to csv for viewing sales overview of data

df_transactions.to_csv("D:/Data Analyst Projects/H & M sales insights/transactions/df_transactions.csv"
```

In [27]:

```
df_transactions.head
```

Out[27]:

```
<bound method NDFrame.head of                                                customer_id  article_
id  \
0         000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...   663713001
1         000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...   541518023
2         00007d2de826758b65a93dd24ce629ed66842531df6699...   505221004
3         00007d2de826758b65a93dd24ce629ed66842531df6699...   685687003
4         00007d2de826758b65a93dd24ce629ed66842531df6699...   685687004
...                                                     ...         ...
31788319  fff2282977442e327b45d8c89afde25617d00124d0f999...   929511001
31788320  fff2282977442e327b45d8c89afde25617d00124d0f999...   891322004
31788321  fff380805474b287b05cb2a7507b9a013482f7dd0bce0e...   918325001
31788322  fff4d3a8b1f3b60af93e78c30a7cb4cf75edaf2590d3e5...   833459002
31788323  fffef3b6b73545df065b521e19f64bf6fe93bfd450ab20...   898573003

             price       t_dat
0         0.050831  2018-09-20
1         0.030492  2018-09-20
2         0.015237  2018-09-20
3         0.016932  2018-09-20
```

In [28]:

```
df_trans_dtl = df_transactions[['customer_id','price']].copy()
```

In [29]:

```
# Calculating purchase value for each customer
df_trans_dtl = df_trans_dtl.groupby('customer_id').sum()
```

In [30]:

```
df_customer_details = df_customers.merge(df_trans_dtl,on='customer_id',how='inner')
```

In [31]:

```
df_customer_details.head()
```

Out[31]:

	customer_id	club_member_status	fashion_news_frequency	age	price
0	00000dbacae5abe5e23885899a1fa44253a17956c6d1c3...	ACTIVE	NONE	49.0	0.648983
1	0000423b00ade91418cceaf3b26c6af3dd342b51fd051e...	ACTIVE	NONE	25.0	2.601932
2	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	ACTIVE	NONE	24.0	0.704780
3	00005ca1c9ed5f5146b52ac8639a40ca9d57aeff4d1bd2...	ACTIVE	NONE	54.0	0.060983
4	00006413d8573cd20ed7128e53b7b13819fe5cfc2d801f...	ACTIVE	Regularly	52.0	0.469695

In [32]:

```
#exporting csv for analysing purchasing behaviour of customers
df_customer_details.to_csv("D:/Data Analyst Projects/H & M sales insights/customer_details.csv")
```

In [35]:

```
# Top selling Products Characteristics
# lets take for top 50 most sold products
topsold = df_transactions["article_id"].value_counts()
top_50 = topsold.iloc[:50]
top_50 = top_50.reset_index()
top_50.rename(columns= {"count":"quantity"},inplace=True)
top_50
```

Out[35]:

	article_id	quantity
0	706016001	50287
1	706016002	35043
2	372860001	31718
3	610776002	30199
4	759871002	26329
5	464297007	25025
6	372860002	24458
7	610776001	22451
8	399223001	22236
9	706016003	21241
10	720125001	21063

In [36]:

```
# collecting details for top 50 products
top_50_details = sqldf("""
select * from
top_50 p inner join
df_articles a
on a.article_id=p.article_id
""")
```

In [37]:
```
top_50_details
```

Out[37]:

	article_id	quantity	article_id	prod_name	product_type_name	product_group_name	colour_group_name	index_nai
0	706016001	50287	706016001	Jade HW Skinny Denim TRS	Trousers	Garment Lower body	Black	Divid
1	706016002	35043	706016002	Jade HW Skinny Denim TRS	Trousers	Garment Lower body	Light Blue	Divid
2	372860001	31718	372860001	7p Basic Shaftless	Socks	Socks & Tights	Black	Lingeries/Tig
3	610776002	30199	610776002	Tilly (1)	T-shirt	Garment Upper body	Black	Ladieswe
4	759871002	26329	759871002	Tilda tank	Vest top	Garment Upper body	Black	Divid
5	464297007	25025	464297007	Greta Thong Mynta Low 3p	Underwear bottom	Underwear	Black	Lingeries/Tig

In [40]:
```
#exporting csv for analysing top_50_details
top_50_details.to_csv("D:/Data Analyst Projects/H & M sales insights/top_50_details.csv")
```