Capstone Project 2 Presentation

Sarifah bte Sapuan

Topics

- Data collection
- Creating ER Diagrams
- Creating tables and Loading data into SQL databases
- Data Preparation
- Data Analysis using SQL
- Creating Interactive Dashboard and Visualization using Excel Dashboards
- Conclusion

Data Collection

SOURCE Citation:

Łapczyński M., Białowąs S. (2013) Discovering Patterns of Users' Behaviour in an E-shop - Comparison of Consumer Buying Behaviours in Poland and Other European Countries,

“Studia Ekonomiczne”, nr 151, “La société de l'information : perspective européenne et globale : les usages et les risques d'Internet pour les citoyens et les consommateurs”, p. 144-153

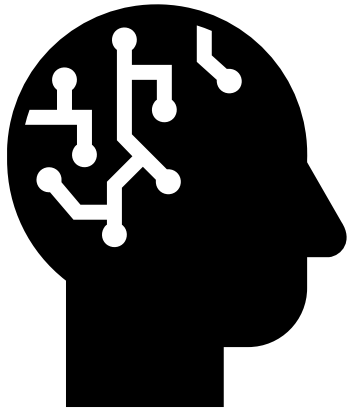
Clickstream Data for Online Shopping

https://archive.ics.uci.edu/ml/datasets/clickstream+data+for+online+shopping#

Data Collection

Method used by Study

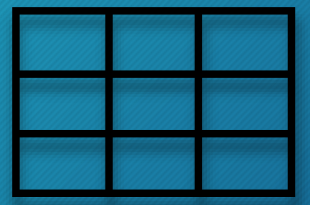
- Web usage mining
- Market basket analysis



Data Collection

- **Data Set Information:**

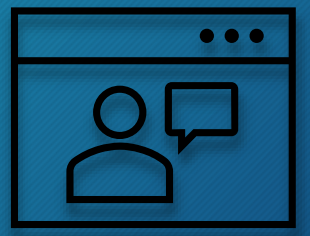
The dataset contains information on clickstream from an online store offering maternity wear



- **Attribute Information:**

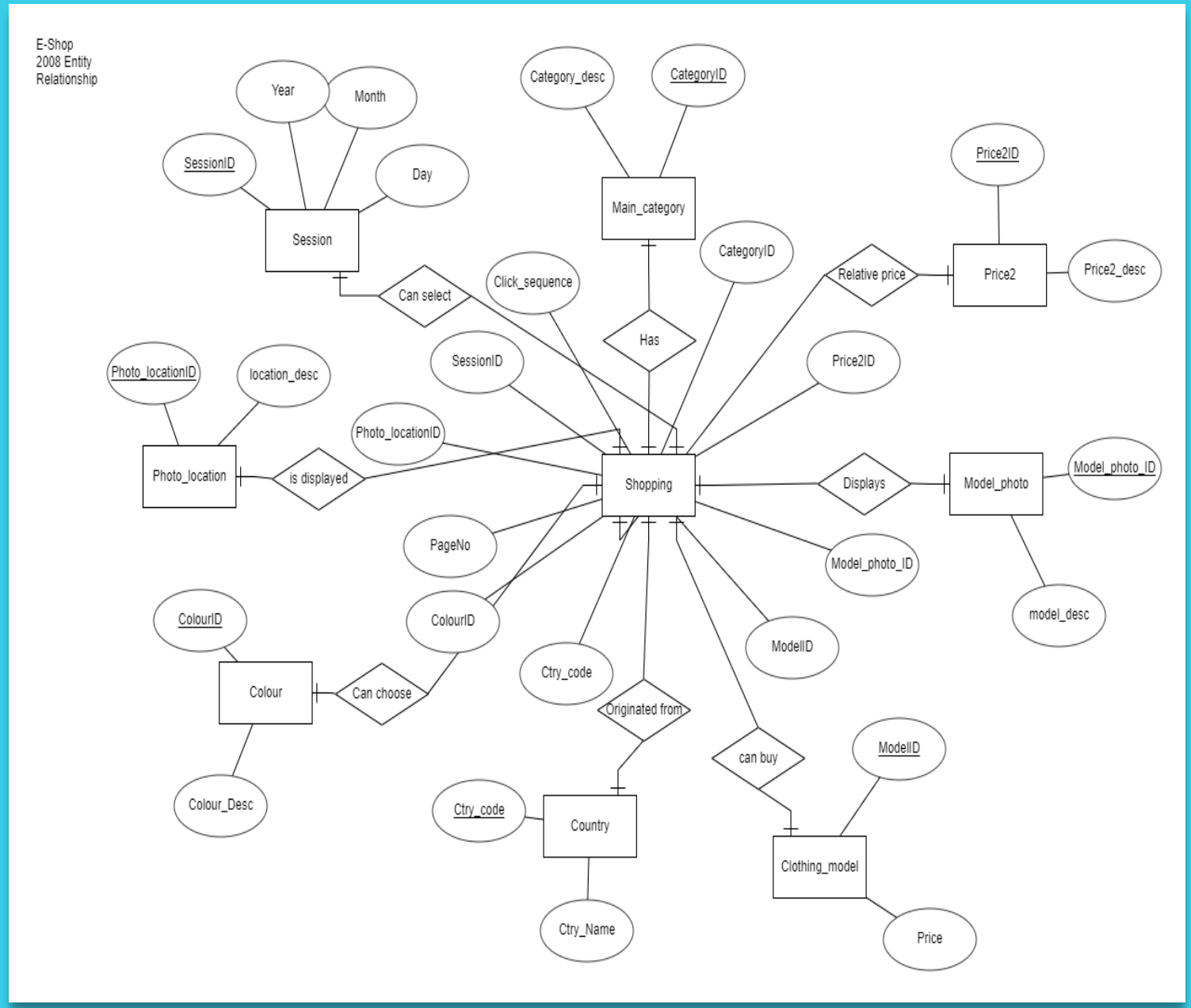
The dataset contains 14 variables from five months of 2008 data

Including country of origin of the IP address



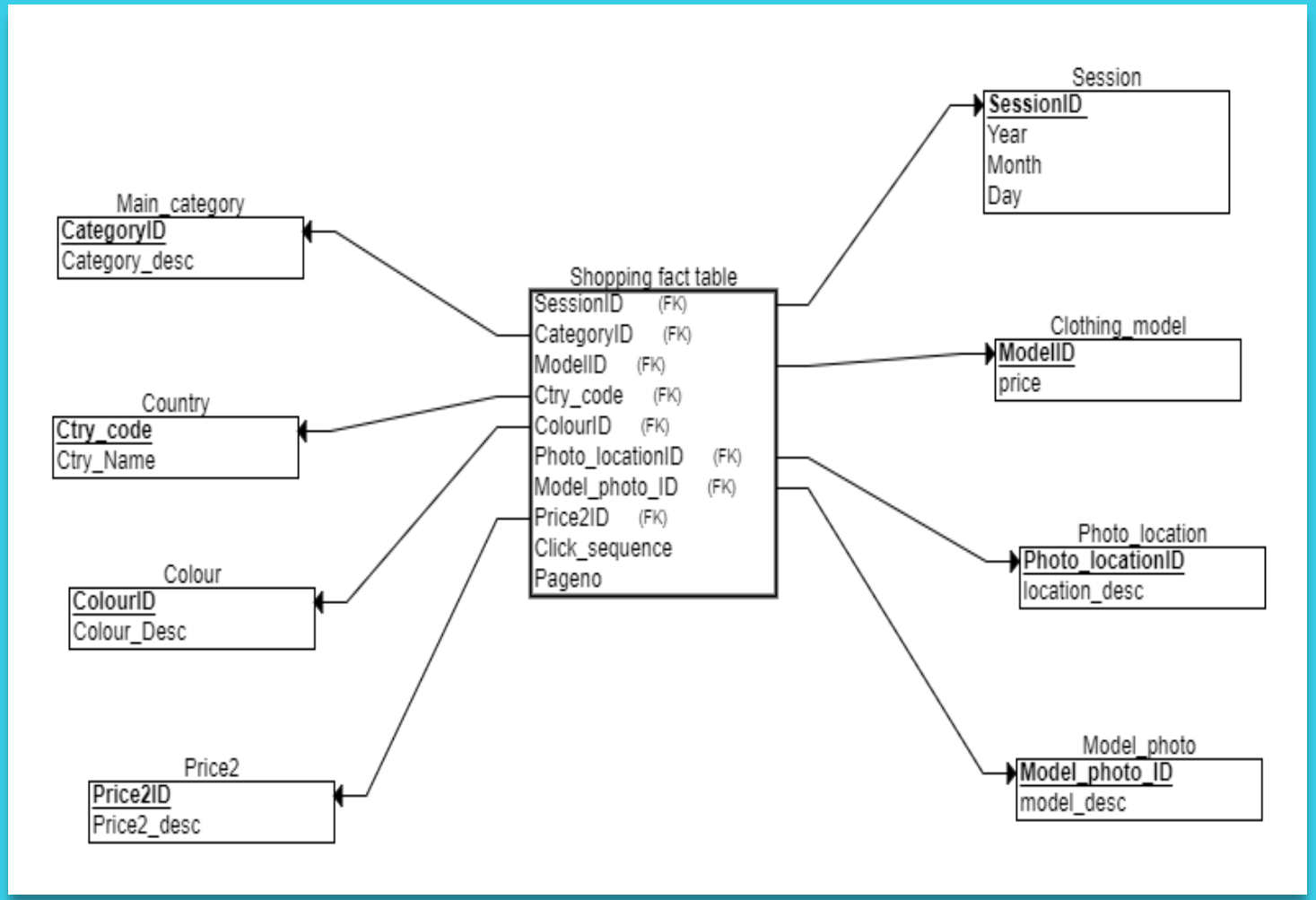
Creating ER Diagrams

- E Shop 2008 Entity Relation Diagram
- Star schema



Creating ER Diagrams

- E Shop 2008 Relationship Diagram
- Star schema



Creating tables and Loading data into SQL databases

1. Created DDL statements to create master and fact tables

2. Loaded the raw data into the tables

Creating tables and Loading data into SQL databases

Created DDL statements

1. DDL Statements for master tables

```sql
CREATE TABLE dbo.Main_category
(
  CategoryID INT NOT NULL,
  Category_desc VARCHAR(15) NOT NULL,
  PRIMARY KEY (CategoryID)
)

INSERT INTO dbo.Main_category
([CategoryID],[Category_desc])
VALUES
(1,'trousers'),
(2,'skirts'),
(3,'blouses'),
(4,'sale');

CREATE TABLE dbo.model_photo
(
  Model_photo_ID INT NOT NULL,
  Model_desc VARCHAR(10) NOT NULL,
  PRIMARY KEY (Model_photo_ID)
)

INSERT INTO dbo.model_photo
VALUES
(1,'en face'),
(2,'profile');
```

```sql
CREATE TABLE dbo.Colour
(
  ColourID INT NOT NULL,
  Colour_Desc VARCHAR(20) NOT NULL,
  PRIMARY KEY (ColourID)
)

INSERT INTO dbo.Colour
VALUES
(1,'beige'),
(2,'black'),
(3,'blue'),
(4,'brown'),
(5,'burgundy'),
(6,'gray'),
(7,'green'),
(8,'navy blue'),
(9,'of many colors'),
(10,'olive'),
(11,'pink'),
(12,'red'),
(13,'violet'),
(14,'white');
```

```sql
CREATE TABLE dbo.Photo_location
(
  Photo_locationID INT NOT NULL,
  location_desc VARCHAR(25) NOT NULL,
  PRIMARY KEY (Photo_locationID)
)

INSERT INTO dbo.Photo_location
VALUES
(1,'top left'),
(2,'top in the middle'),
(3,'top right'),
(4,'bottom left'),
(5,'bottom in the middle'),
(6,'bottom right');

CREATE TABLE Clothing_model
(
  ModelID VARCHAR(3) NOT NULL,
  price INT NOT NULL,
  PRIMARY KEY (ModelID)
);

CREATE TABLE Session
(
  SessionID_ INT NOT NULL,
  Year INT NOT NULL,
  Month INT NOT NULL,
  Day INT NOT NULL,
  PRIMARY KEY (SessionID_)
);
```

Creating tables and Loading data into SQL databases

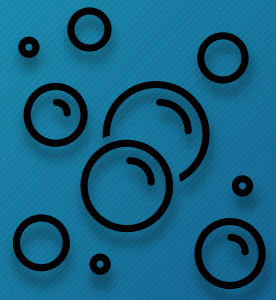
2. DDL Statements for fact table

```
CREATE TABLE dbo.Shopping
(
  SessionID INT NOT NULL,
  Ctry_code INT NOT NULL,
  CategoryID INT NOT NULL,
  ModelID VARCHAR(3) NOT NULL,
  ColourID INT NOT NULL,
  Photo_locationID INT NOT NULL,
  Model_photo_ID INT NOT NULL,
  Price2ID INT NOT NULL,
  Click_sequence INT NOT NULL,
  Pageno INT NOT NULL,
  FOREIGN KEY (CategoryID) REFERENCES Main_category(CategoryID),
  FOREIGN KEY (ModelID) REFERENCES Clothing_model(ModelID),
  FOREIGN KEY (SessionID) REFERENCES Session(SessionID),
  FOREIGN KEY (Ctry_code) REFERENCES Country(Ctry_code),
  FOREIGN KEY (ColourID) REFERENCES Colour(ColourID),
  FOREIGN KEY (Photo_locationID) REFERENCES Photo_location(Photo_locationID),
  FOREIGN KEY (Model_photo_ID) REFERENCES Model_photo(Model_photo_ID),
  FOREIGN KEY (Price2ID) REFERENCES Price2(Price2ID)
)

INSERT INTO dbo.Shopping
SELECT
[session_ID]
,[country]
,[page_1_main_category]
,[page_2_clothing_model]
,[colour]
,[location]
,[model_photography]
,[price_2]
,[order]
,[page]
from dbo.eshop2008;
```

Data Preparation

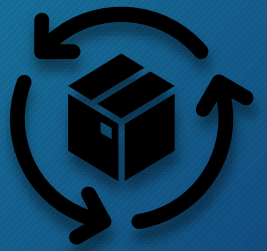
1. Data Cleaning



2. Data transformation

- Added derived column Country type in Country table

3. Normalization



Data Preparation - Normalization

First Normal Form

First Normal Form (1NF)

For a table to be in the 1 NF, it should follow the following 4 rules:

1.It should only have single(atomic) valued attributes/columns.

2.Values stored in a column should be of the same domain

3.All the columns in a table should have unique names.

4.And the order in which data is stored, does not matter.

session ID (PK)		**main category**
year		clothing model
month		colour
day		location
country		model photography
order		price
page		price 2

Data Preparation - Normalization

Second Normal Form

Second Normal Form (2NF)
For a table to be in the Second Normal Form,
1.It should be in the First Normal form.
2.It should not have Partial Dependency.

Session
Session_ID
Order
day
year
month
day
Ctry_code (FK)
page

Country
Ctry_code (PK)
Ctry_Name

Clothing_model
ModelID (PK)
price
price 2
CategoryID (FK)
ColourID (FK)
Photo_location (FK)
Model_photo_ID (FK)

Main_category
CategoryID (PK)
Category_desc

Colour
ColourID (PK)
Colour_Desc

Photo_location
photo_location (PK)
position_desc

Model_photo
Model_photo_ID (PK)
model_desc

Data Preparation - Normalization

Third Normal Form

Third Normal Form (3NF)
A table is said to be in the Third Normal Form when,
1.It is in the Second Normal form.
2.And, it doesn't have Transitive Dependency.

Shopping	Attributes
SessionID (FK)	Int
Ctry_code (FK)	Int
CategoryID (FK)	Int
ModelID (FK)	Varchar (3)
ColourID (FK)	Int
Photo_location (FK)	Int
Model_photo_ID (FK)	Int
price_2ID (FK)	Int
Click_sequence	Int
PageNo	Int

Session	Attributes
SessionID (PK)	Int
Year	Int
Month	Int
Day	Int

Clothing_model	Attributes
ModelID (PK)	Varchar (3)
price	money

Main_category	Attributes
CategoryID (PK)	Int
Category_desc	Varchar (15)

Country	Attributes
Ctry_code (PK)	Int
Ctry_Name	Varchar (30)

Colour	Attributes
ColourID (PK)	Int
Colour_Desc	Varchar (20)

Model_photo	Attributes
Model_photo_ID (PK)	Int
model_desc	Varchar (10)

Photo_location	Attributes
Photo_locationID (PK)	Int
location_desc	Varchar (25)

Price_2	Attributes
price_2ID (PK)	Int
Price_2_desc	Varchar (3)

Data Analysis using SQL

1. Main Query to retrieve data from fact table

```sql
-- select all related fields for each session

select sh.SessionID,sh.Click_sequence, sh.Pageno, ss.year, ss.month, ss.day,
mc.Category_desc, sh.ModelID, cl.Colour_Desc,
pl.location_desc,mp.Model_desc, pr.Price2_desc, c.Ctry_type, c. Ctry_Name
from Shopping sh
full outer join Session ss
on sh.SessionID=ss.SessionID
left join Main_category mc
on sh.CategoryID = mc.CategoryID
left join Colour cl
on sh.ColourID = cl.ColourID
left join [Photo_location] pl
on sh.[Photo_locationID] = pl.Photo_locationID
left join model_photo mp
on sh.Model_photo_ID = mp.Model_photo_ID
left join Price2 pr
on sh.Price2ID = pr.Price2ID
join dbo.Country c
on sh.Ctry_code=c.Ctry_code
order by sh.SessionID,sh.Click_sequence;
```

	SessionID	Click_sequence	Pageno	year	month	day	Category_desc	ModelID	Colour_Desc	location_desc	Model_desc	Price2_desc	Ctry_type	Ctry_Name
1	1	1	1	2008	4	1	trousers	A13	beige	bottom in the middle	en face	No	Poland	Poland
2	1	2	1	2008	4	1	trousers	A16	beige	bottom right	en face	No	Poland	Poland
3	1	3	1	2008	4	1	skirts	B4	olive	top in the middle	en face	Yes	Poland	Poland
4	1	4	1	2008	4	1	skirts	B17	gray	bottom right	profile	No	Poland	Poland
5	1	5	1	2008	4	1	skirts	B8	brown	top right	profile	Yes	Poland	Poland
6	1	6	4	2008	4	1	blouses	C56	gray	top left	profile	Yes	Poland	Poland
7	1	7	4	2008	4	1	blouses	C57	burgundy	top left	profile	No	Poland	Poland
8	1	8	4	2008	4	1	sale	P67	of many colors	bottom in the middle	en face	Yes	Poland	Poland
9	1	9	5	2008	4	1	sale	P82	gray	bottom left	profile	Yes	Poland	Poland
10	2	1	2	2008	4	1	skirts	B31	of many colors	bottom in the middle	en face	Yes	Poland	Poland
11	2	2	2	2008	4	1	skirts	B21	red	top left	en face	Yes	Poland	Poland
12	2	3	2	2008	4	1	skirts	B24	pink	top in the middle	en face	Yes	Poland	Poland

Query executed successfully. | DESKTOP-082V827 (15.0 RTM) | DESKTOP-082V827\sarif ... | Capstone2 | 00:00:01 | 165,474 rows

Data Analysis using SQL

```sql
-- select total ModelID by country
select sh.SessionID, cm.[ModelID], COUNT(*) as "Total by ModelID",
c.Ctry_Name, c.Ctry_type, cl.Colour_Desc
from Shopping sh join [dbo].[clothing_model] cm
on sh.[ModelID] = cm.[ModelID]
join Country c
on sh.ctry_code = c.Ctry_code
join Colour cl
on sh.ColourID = cl.ColourID
group by sh.SessionID,cm.[ModelID], c.Ctry_Name, c.Ctry_type, cl.Colour_Desc
order by sh.SessionID, COUNT(*) desc, c.Ctry_Name;
```

	SessionID	ModelID	Total by ModelID	Ctry_Name	Ctry_type	Colour_Desc
1	1	C56	1	Poland	Poland	gray
2	1	C57	1	Poland	Poland	burgundy
3	1	B4	1	Poland	Poland	olive
4	1	A16	1	Poland	Poland	beige
5	1	B8	1	Poland	Poland	brown
6	1	P67	1	Poland	Poland	of many colors
7	1	A13	1	Poland	Poland	beige
8	1	P82	1	Poland	Poland	gray
9	1	B17	1	Poland	Poland	gray
10	2	B27	2	Poland	Poland	black
11	2	A10	2	Poland	Poland	blue

```sql
--  select total for each colour
select cl.Colour_Desc, COUNT(cl.Colour_Desc) as Total , c.Ctry_Name, c.Ctry_type
from Shopping sh
join Colour cl
on sh.ColourID = cl.ColourID
join Country c
on sh.ctry_code = c.Ctry_code
group by cl.Colour_Desc, c.Ctry_Name, c.Ctry_type
order by Total desc;
```

	Colour_Desc	Total	Ctry_Name	Ctry_type
1	black	24593	Poland	Poland
2	blue	22457	Poland	Poland
3	gray	14203	Poland	Poland
4	brown	13857	Poland	Poland
5	white	13200	Poland	Poland
6	of many colors	11259	Poland	Poland
7	red	7067	Poland	Poland
8	beige	5959	Poland	Poland
9	green	5802	Poland	Poland
10	violet	4933	Poland	Poland
11	pink	4288	Poland	Poland

Additional SQL Queries for analysis

Creating Interactive Dashboard and Visualization using Excel Dashboards

1. Imported data into Excel from SQL
2. Created Pivot tables and charts
3. Combined in a dashboard

Creating Interactive Dashboard and Visualization using Excel Dashboards

OBJECTIVE:

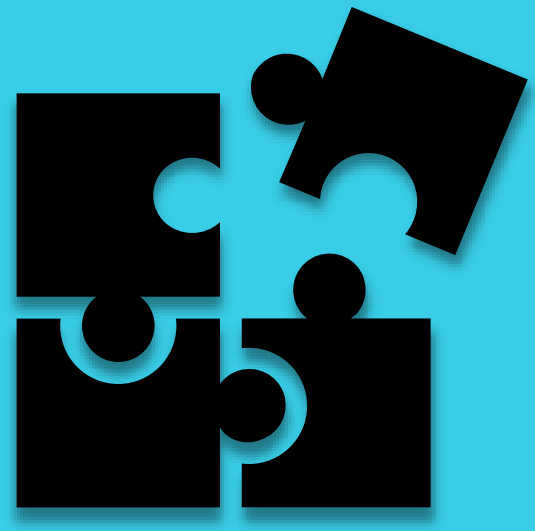
Discovering Patterns of Users' Behaviour in an Online shop – Comparison of Consumer Buying Behaviours in Poland and Other European Countries

Creating Interactive Dashboard and Visualization using Excel Dashboards

Pre-assumptions:

1. Website had multiple languages
- Potential customers within and outside Poland

2. Country was assigned based on the geolocation related to the IP addresses in the web logs and some maybe inaccurate



Creating Interactive Dashboard and Visualization using Excel Dashboards

RESULTS: Based on Blouses

1. 3 Most popular models
2. 3 Most popular colour
3. Price choice if higher than average product price
4. Average number of clicks per session
5. Highest number of clicks per session

Creating Interactive Dashboard and Visualization using Excel Dashboards

EXCEL Dashboard

1. 3 Most popular models
2. 3 Most popular colour
3. Price choice if higher than average product price
4. Average number of clicks per session
5. Highest number of clicks per session

Conclusion

Web usage mining

- enables insight into behaviour patterns and transactional patterns of online shoppers by using the statistical and data mining models

- Assist in Marketing and Sales strategies to increase the sales volume

Q & A