Università di Pisa	A.A. 2012-2013
Data Mining II	

Project assignments / Part 1

Software used

- Postgres
  - More complex to import the data
  - Easier to compute some aggregations
  - No plotting functionalities → export aggregates to make plot with some spreadsheet

Import (example)

```
create TABLE dm2_progetti."vendite" (
  scontrino bigint,
  data_id integer,
  cliente_id integer,
  articolo_id integer,
  tipologia_id integer,
  importo float,
  qta_pezzi float,
  qta_peso float,
  promozione char(1)
  )
WITH (   OIDS=FALSE );
ALTER TABLE dm2_progetti.vendite  OWNER TO postgres;


copy dm2_progetti."vendite" from '/home/SSET_VEN_CORSODM_dots.csv'
using delimiters ';'   CSV HEADER
```

Import (issues)

- Some text files are in UTF-8 format, others in ISO-8859
  - Instruct the tool used to import, or
  - Convert, e.g.: iconv --from-code=ISO-8859-1 --to-code=UTF-8 SSET_MKT_CORSODM.csv > SSET_MKT_CORSODM_utf8.csv
- Some strings contain single quotes: '
  - Remove them if troublesome (not for Postgres)
- Decimal separators for numbers is the comma
  - Instruct the tool used to import, or replace them in the input file

Data exploration

- How many products in each Settore?

"MULTIMEDIA";505

"GROCERY ALIMENTARI";855

"INIZIATIVE SPECIALI";44

"FRESCHISSIMI";889

"FRESCHI";661

"EROGAZIONE CARBURANTI";3

"SALUTE";585

"PERSONA";926

"NON DEFINITO";7

"STAGIONALI E BRICO";1404

"CONFEZIONATO PER VENDITA";63

"CASA";664

"CHIMICA";397

```
SELECT   marketing.settore, count(*)
FROM   dm2_progetti.marketing
group by 1
```

Data exploration

- How many single items of each Settore are sold?

(Name, #items, #distinct items)

"CASA";2527;118

"CHIMICA";32512;1000

"CONFEZIONATO PER VENDITA";28621;3

"FRESCHI";105847;1076

"FRESCHISSIMI";129107;857

"GROCERY ALIMENTARI";125964;2165

"INIZIATIVE SPECIALI";1052;5

"MULTIMEDIA";1135;49

"PERSONA";632;104

"STAGIONALI E BRICO";1054;125

```
SELECT   c.settore, count(*)

FROM   dm2_progetti.vendite a,
dm2_progetti.articoli b,
dm2_progetti.marketing c

where a.articolo_id=b.articolo_id and
b.cod_mkt_id=c.cod_marketing_id
group by 1
```

Profiling

- For each customer, compute volume and number of single purchases, and # visits

(cliente_id, volume, purchases, visits)

1146;21.17;15;1

1207;666.76;328;83

3563;427.81;175;28

5379;204.4;86;10

5801;14.56;7;1

5892;258.99;139;30

6228;424.23;183;27

7191;107.75;39;10

7409;120.92;79;14

7758;139.21;81;12

```
SELECT   a.cliente_id, sum(a.importo), count(*),
count(distinct a.scontrino)

FROM   dm2_progetti.vendite a, dm2_progetti.articoli b,
dm2_progetti.marketing c

where a.articolo_id=b.articolo_id and
b.cod_mkt_id=c.cod_marketing_id

[and c.settore = "X"]

group by 1
limit 10
```

Individual events detection

- Basic building block: compute visits and products sold for each customer and month

(cliente_id, month, #products, visits)

1146;5;15;1

1207;4;44;12

1207;5;36;11

1207;6;32;9

1207;7;90;22

1207;8;69;15

1207;9;57;14

3563;4;22;3

3563;5;19;4

3563;6;39;6

3563;7;35;5

3563;8;46;8

3563;9;14;2

```
SELECT   a.cliente_id, d.mese_n, count(*), count(distinct a.scontrino)
FROM   dm2_progetti.vendite a, dm2_progetti.data d
where a.data_id=d.data_id
group by 1,2
order by 1,2
```

Individual events detection / churn

- Number of visits in the last n months (here n=2)

1146;0

1207;29

3563;10

5379;6

5801;1

5892;10

6228;5

7191;5

7409;4

7758;9

```
select agg.cliente_id, sum(case agg.mese_n>7 when true then
agg.visits else 0 end)
from (
	SELECT	a.cliente_id, d.mese_n, count(*) as products,
count(distinct a.scontrino) as visits
	FROM	dm2_progetti.vendite a, dm2_progetti.data d
	where a.data_id=d.data_id
	group by 1,2
) agg
group by 1
```

Individual events detection / focus

- Products sold in each reparto in each 3-month slot

989269;"LIBERO SERVIZIO";0;8

938840;"GASTRONOMIA";0;9

207653;"SURGELATI";30;32

345877;"LIBERO SERVIZIO";0;2

```
create table dm2_progetti.vendite_reparti as
( select agg.cliente_id, agg.reparto, sum(case agg.mese_n<7 when true then agg.products
else 0 end) as p1,  sum(case agg.mese_n>6 when true then agg.products else 0 end) as p2
from (
      SELECT   ven.cliente_id, dat.mese_n, mkt.reparto, count(*) as products
      FROM   dm2_progetti.vendite ven, dm2_progetti.data dat, dm2_progetti.articoli art,
dm2_progetti.marketing mkt
      where ven.data_id=dat.data_id and ven.articolo_id=art.articolo_id and
art.cod_mkt_id=mkt.cod_marketing_id and mkt.settore='FRESCHI'
      group by 1,2,3
) agg group by 1,2 )
```

Individual events detection / focus

- Customers that are focusing
  - (REP.p2/ALL.p2)/(REP.p1/ALL.p1) >= 2 rewritten as REP.p2 * ALL.p1 >= 2 * ALL.p2 * REP.p1

989269;"no"
938840;"no"
207653;"no"
345877;"no"
273928;"no"
572747;"yes"
746382;"yes"
54262;"no"
400723;"no"
552220;"no"
680352;"no"
740049;"no"
473272;"yes"

```
select REP.cliente_id, (case (REP.p2 * ALLR.p1 > 2.0 * ALLR.p2 * REP.p1)
when true then 'yes' else 'no' end) as focus
from dm2_progetti.vendite_reparti REP,
     ( select a.cliente_id, sum(a.p1) as p1, sum(a.p2) as p2 from
dm2_progetti.vendite_reparti a group by 1) ALLR
where REP.cliente_id=ALLR.cliente_id
```

"yes";710
"no";7178

Time series

- Which product to follow?
- Our choice: level of “categoria” → Choose the most promising one

"LATTE";9505
"SALUMI A SERVIZIO ASSISTITO";7632
"FORMAGGI FRESCHI";6902
"YOGURT";4876
"GELATI";4730
"SALUMI LIBERO SI
"GRASSI";3344
"UOVA";3142
"FORMAGGI STAGIONATI SERVIZIO ASSISTITO";3138
"FORMAGGI FRESCHI SERVIZIO ASSISTITO";2824

```
SELECT   mkt.categoria, count(distinct (ven.cliente_id,
dat.settimana_anno))
FROM   dm2_progetti.vendite ven, dm2_progetti.data dat,
dm2_progetti.articoli art, dm2_progetti.marketing mkt
where ven.data_id=dat.data_id and ven.articolo_id=art.articolo_id
and art.cod_mkt_id=mkt.cod_marketing_id
      and mkt.settore='FRESCHI'
group by 1
order by 2 desc
```

Time series

- Compute #products in each week of each customer

5892;18;1	1207;26;1	3563;17;1
5892;19;2	1207;27;4	3563;25;1
5892;20;1	1207;30;2	3563;28;1
5892;21;2	1207;31;2	3563;29;1
5892;23;2	1207;32;2	3563;34;1
5892;26;1	1207;33;1	
5892;27;3	1207;34;2	
5892;28;1	1207;35;1	
5892;29;2	1207;36;1	
5892;30;1	1207;37;2	
5892;31;1	1207;38;2	
5892;32;2	1207;39;3	
5892;33;1	1207;40;1	
5892;34;2		
5892;36;1		
5892;38;2		

```
SELECT   cli.cliente_id, dat.settimana_anno, count(*)
FROM   dm2_progetti.vendite ven, dm2_progetti.data dat,
       dm2_progetti.articoli art, dm2_progetti.marketing mkt,
       dm2_progetti.clienti cli
where ven.data_id=dat.data_id and ven.articolo_id=art.articolo_id
    and art.cod_mkt_id=mkt.cod_marketing_id  and
    mkt.settore='FRESCHI' and mkt.categoria='LATTE' and
    cli.cliente_id=ven.cliente_id
group by 1,2
order by 1,2
```