OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG

SAP UCC University Competence Center

Data Generator for SAP Solutions using Benerator Tool

Supervised by **Tim Böttcher**

Baba Pakruddin Tailor
Julian Reddy Allam
Naga Sai Krishna Ayinampudi
Sai Rajesh Vanimireddy

Agenda:

1) Aim for data generation
2) Global bike INC.
3) Literature research
4) Benerator configuration
5) Generating data and Analysis
6) Conclusion
7) References

- **AIM FOR DATA GENERATION**

Aim For Data Generation:

•Real-world data is often subject to several privacy constraints.

• Under these constraints, researchers often resort to generate data to verify the efficacy.

• The generated data must be realistic and correct in terms of size and distributions.

• Methods of generating datasets for different purposes can be quite different.

• Our work concentrates on generation of test instances to analyze business process.

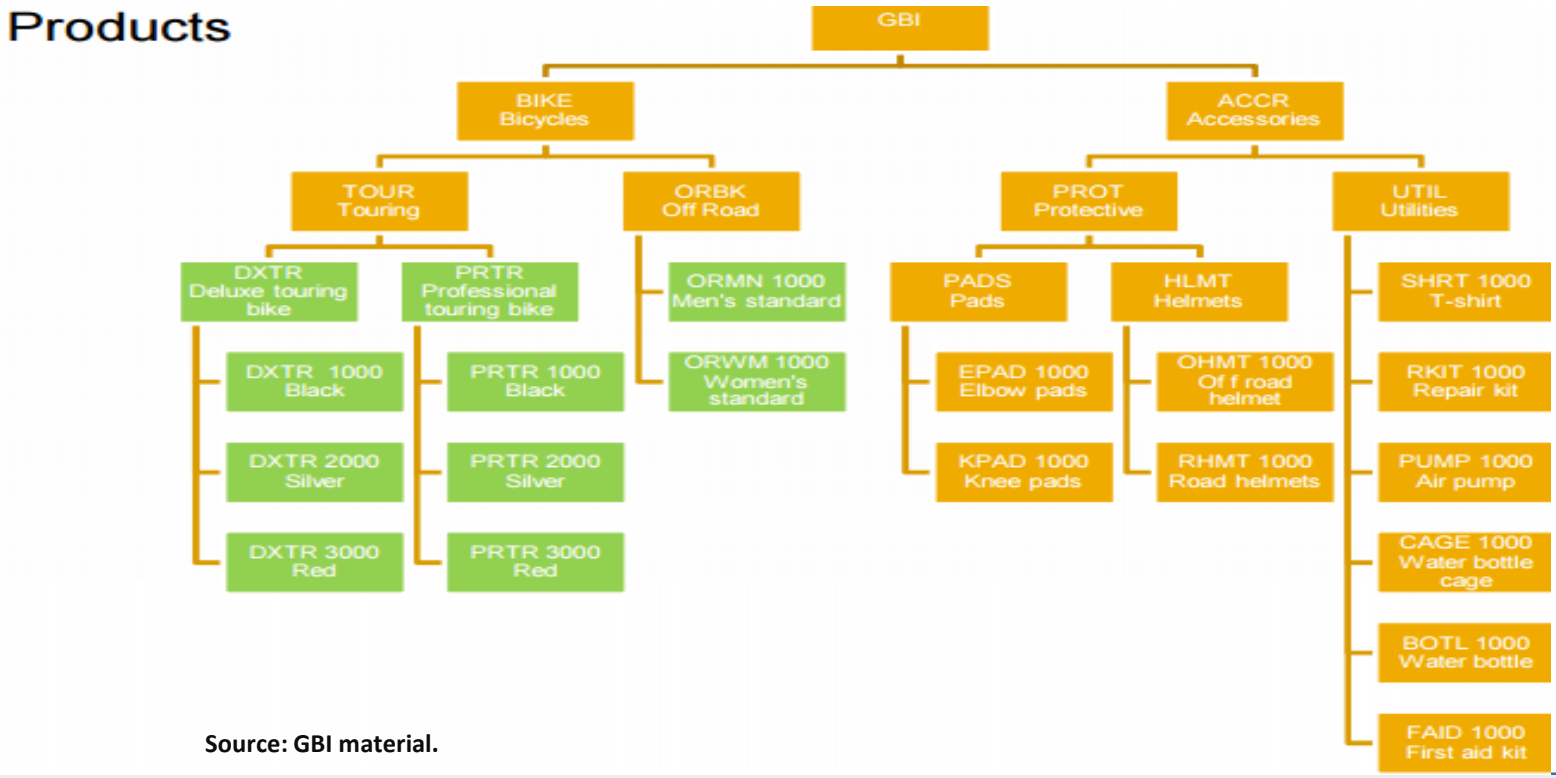
Aim For Data Generation:

- Realistic represents things in a way that is accurate and true to life.
- Synthetic (of a proposition) having truth or falsity determinable by recourse to experience.
- Synthetic data generators allow us to generate large volumes of data with well-understood characteristics.
- We can easily vary the characteristics of the generated data by varying the input parameters of the data generator.

- **Global Bike INC.**

Global Bike INC. :

• GBI is an eminent bicycle company producing bikes and accessories for both touring and off-road racing.



Source: GBI material.

Global Bike INC. :

- **CUSTOMERS**

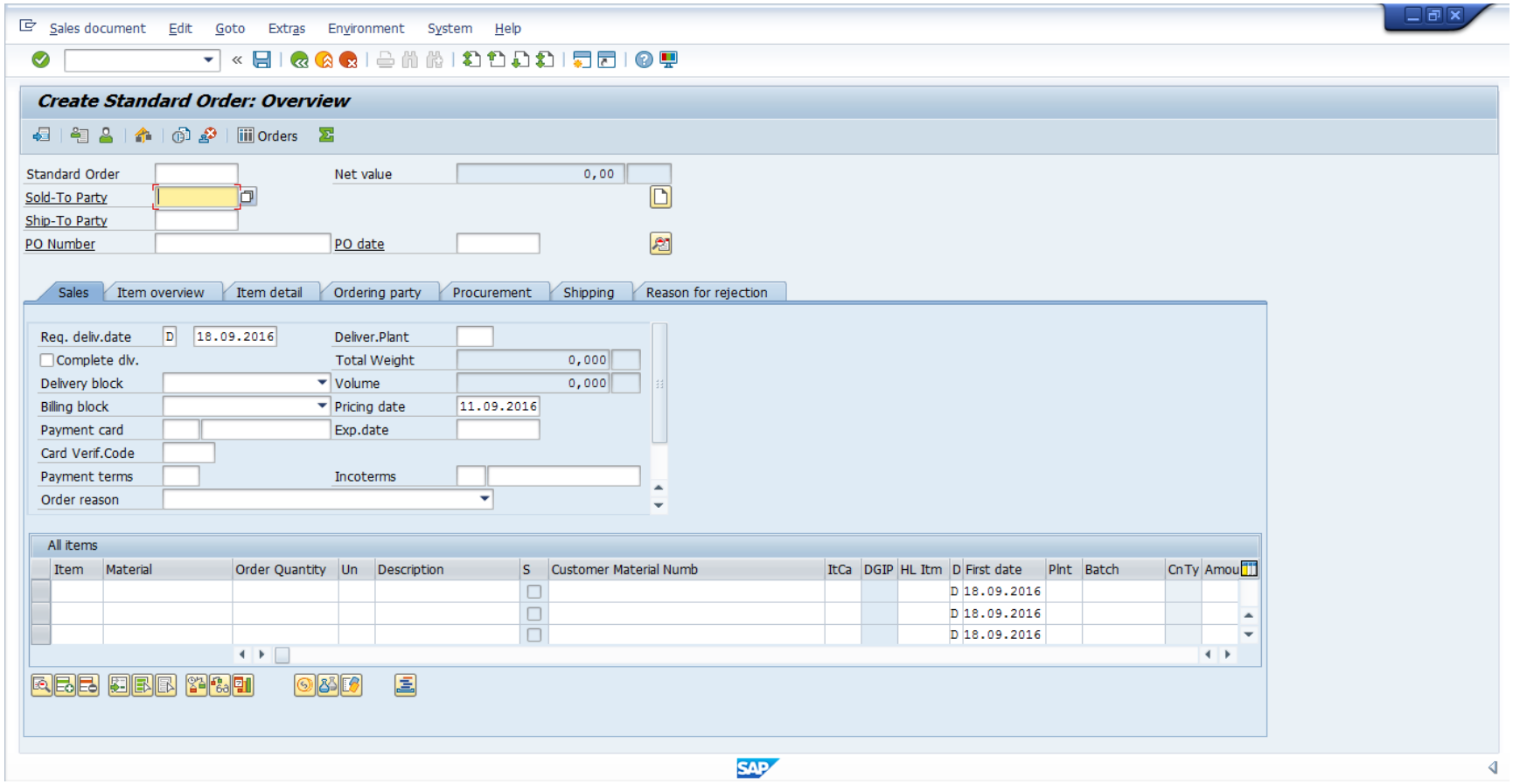
10014	NEW YORK CITY	BIG APPLE BIKES	2000	US00
18033	BOSTON	BEANTOWN BIKES	5000	US00
19073	PHILADELPHIA	PHILLY BIKES	3000	US00
20004	WASHINGTON DC	DC BIKES	11000	US00
30319	ATLANTA	PEACHTREE BIKES	4000	US00
32804	ORLANDO	THE BIKE ZONE	25011	US00
48076	DETROIT	MOTOWN BIKES	8000	
49504	GRAND RAPIDS	FURNITURE CITY BIKES	7000	
60515	CHICAGO	WINDY CITY BIKES	6000	
80111	DENVER	ROCKY MOUNTAIN BIKES	1000	
92612	IRVINE	SOCAL BIKES	9000	
94304	PALO ALTO	SILICON VALLEY BIKES	10000	
98004	SEATTLE	NORTHWEST BIKES	12000	

04227	LEIPZIG	DRAHTESEL	18000	DE00
16341	BERLIN	CAPITAL BIKES	16000	DE00
17389	ANKLAM	OSTSEERAD	21000	DE00
22760	HAMBURG	ALSTER CYCLING	14000	DE00
22767	HAMBURG	RED LIGHT BIKES	23000	DE00
30627	HANNOVER	CRUISER BIKES	17000	DE00
39130	MAGDEBURG	VELODOM	24000	DE00
44784	BOCHUM	FAHRPOTT	19000	DE00
60549	FRANKFURT	AIRPORT BIKES	13000	DE00
69115	HEIDELBERG	NECKARAD	20000	DE00
70825	STUTTGART	RÄDLELAND	22000	DE00
92275	MÜNCHEN	BAVARIA BIKES	15000	DE00

Source: GBI material.

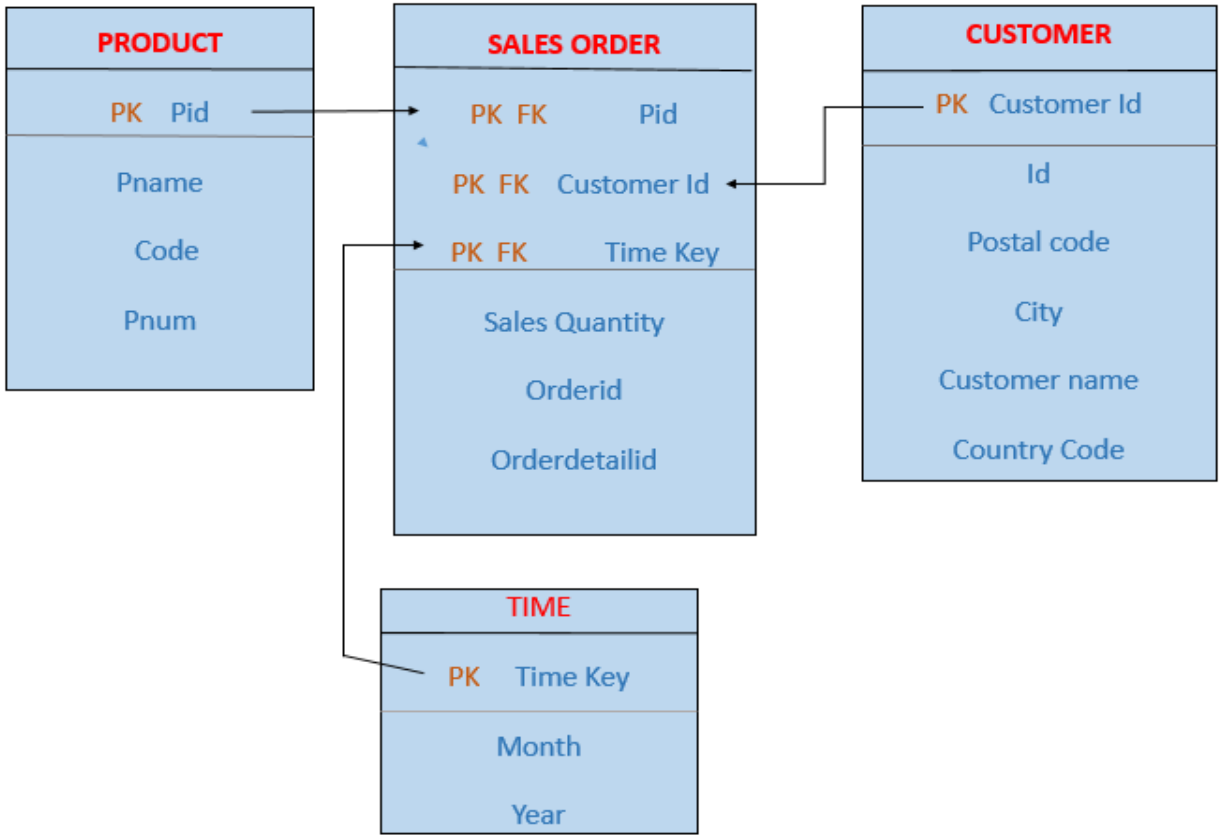
Global Bike INC. :

- **SALES ORDER CREATION IN ERP :**



Global Bike INC. :

- **DATA MODEL:**

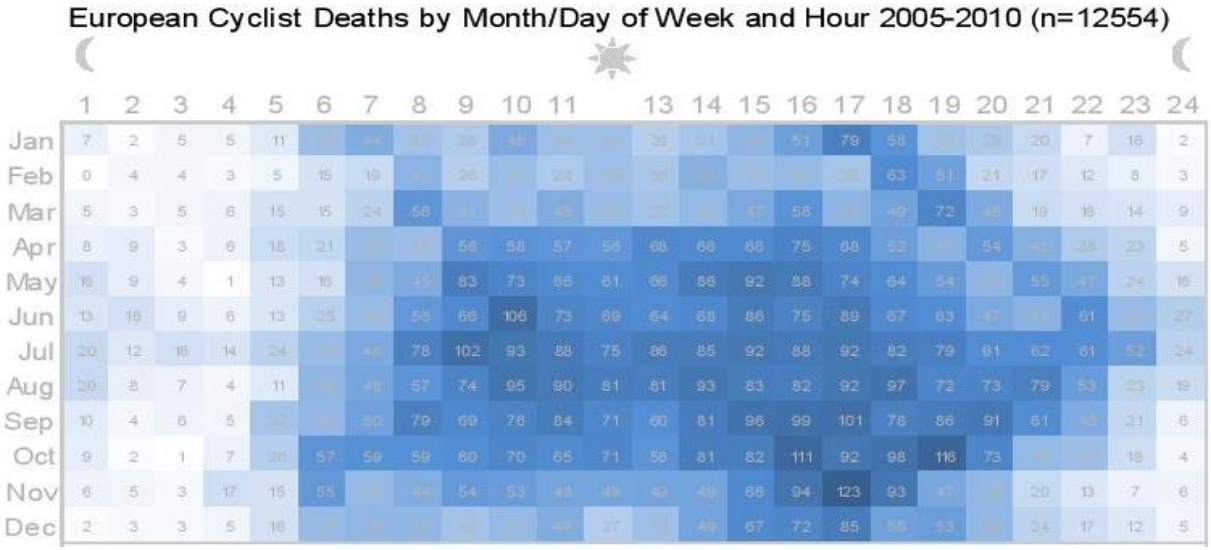


- **Literature Research**

Literature Research:

- In order to generate realistic data we have researched for certain rules which affects the sales.
- **MONTHLY DISTRIBUTION:**

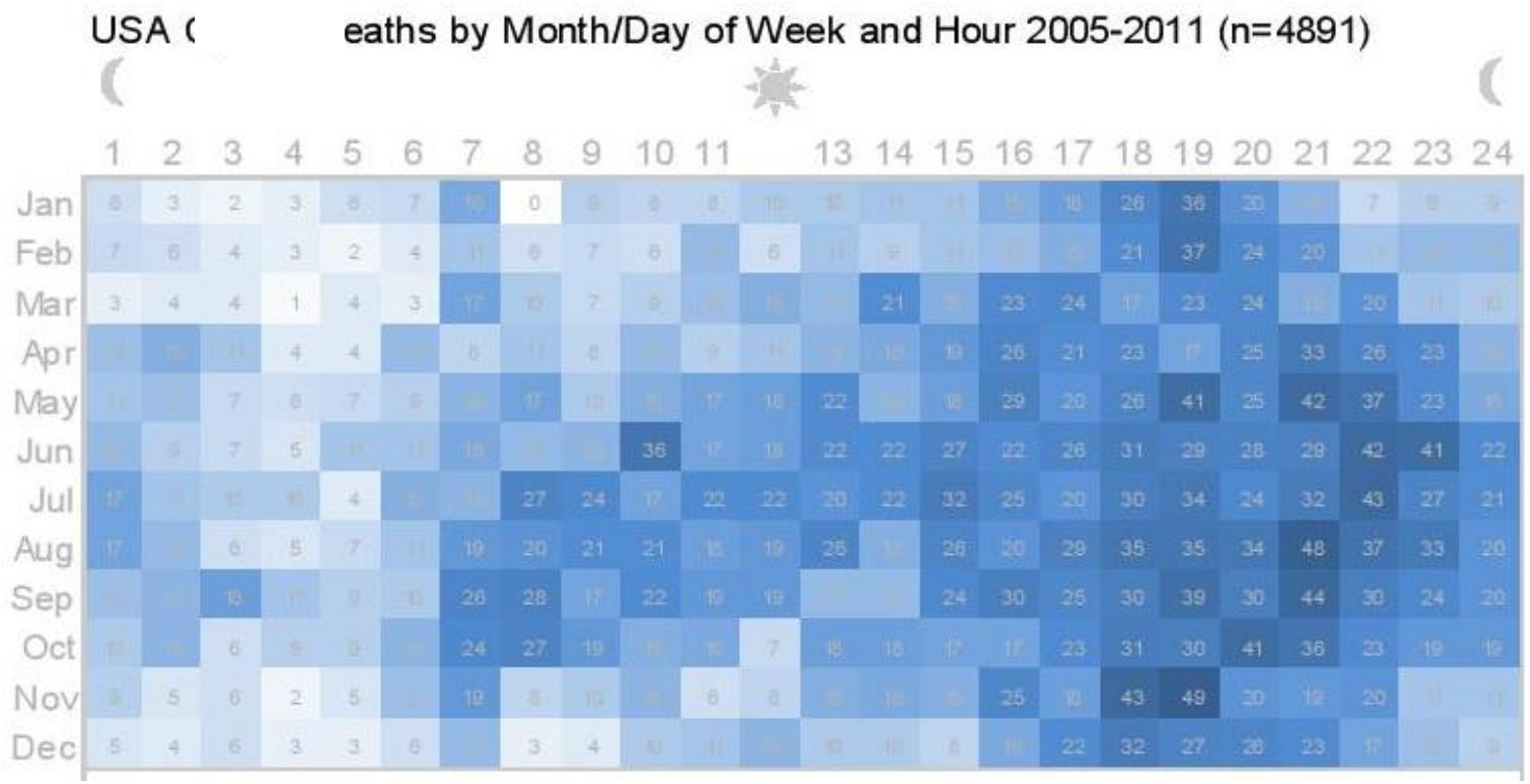
Cyclist traffic fatalities by month or day of week and by time of day, EU, 2005-2010



Source: OECD/International transport forum.

- **MONTHLY DISTRIBUTION:**

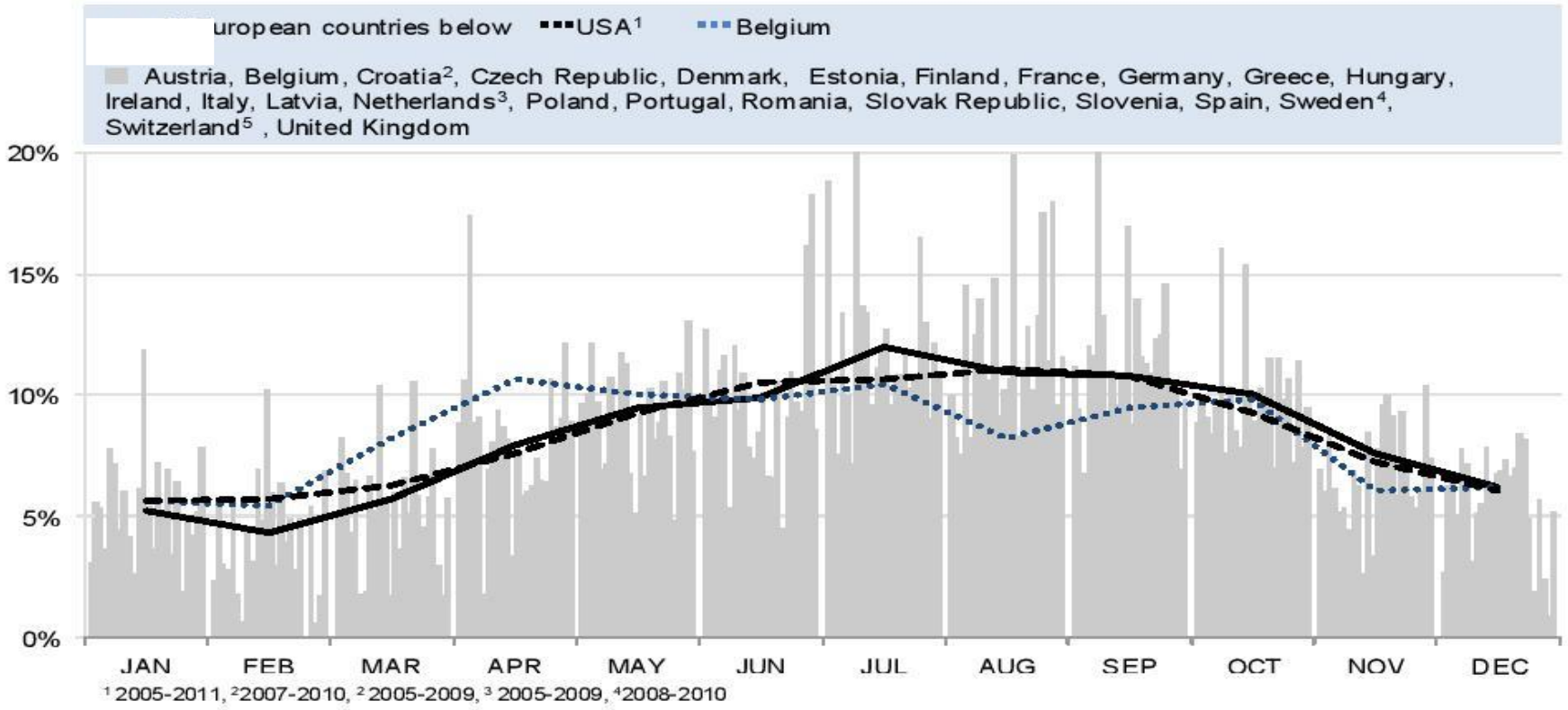
Figure 4.2 **Cyclist traffic fatalities by month or day of week and by time of day, USA, 2005-2011**



Source: OECD/International transport forum.

- **MONTHLY DISTRIBUTION:**

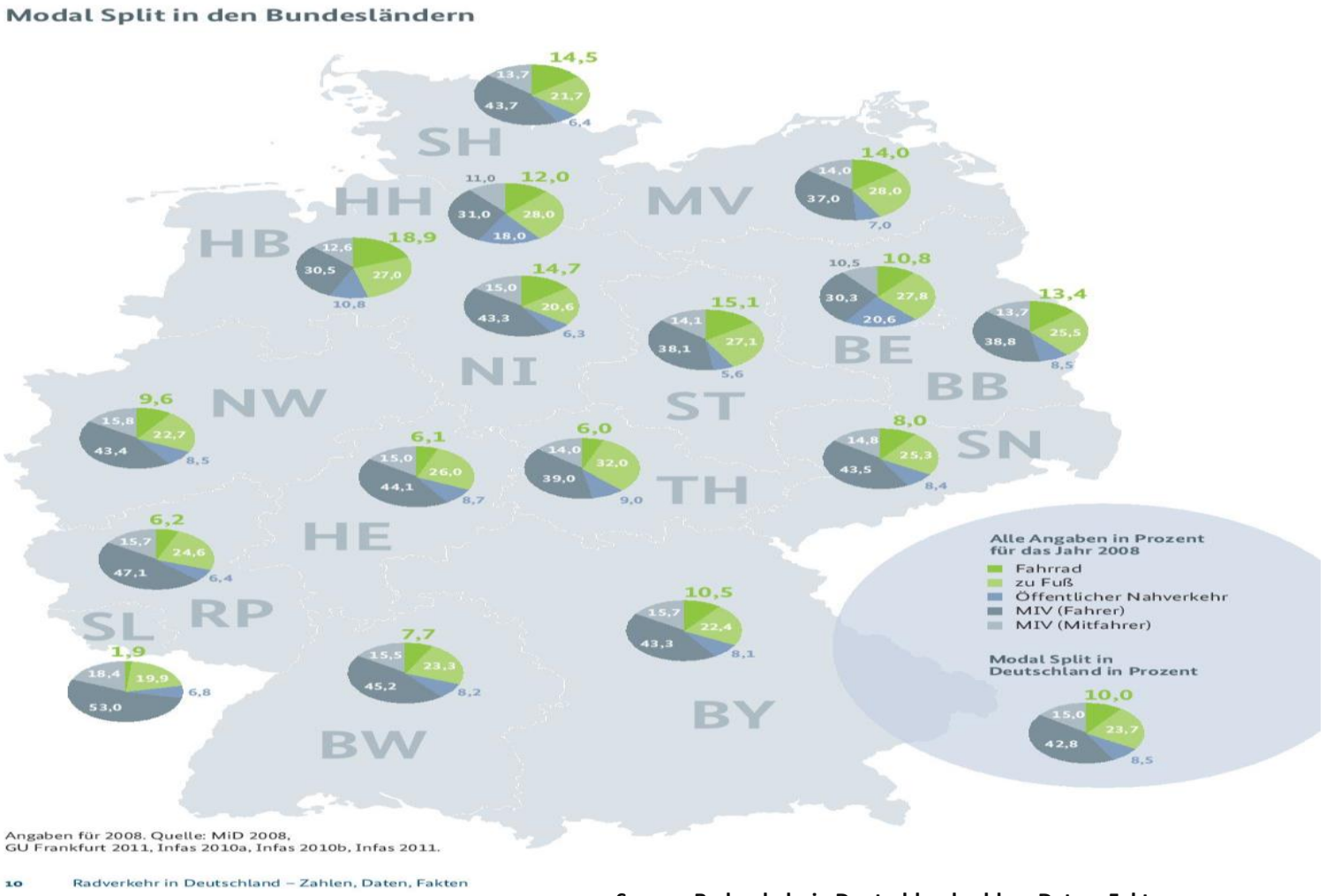
Figure 4.3 **Percentage of all reported fatal bicycle crashes occurring by month, selected European countries.**



Source: EU CARE database, 2005-2010 and USA FARS database 2005-2011

Source: OECD/International transport forum.

- **DISTRIBUTION BY CITIES:**

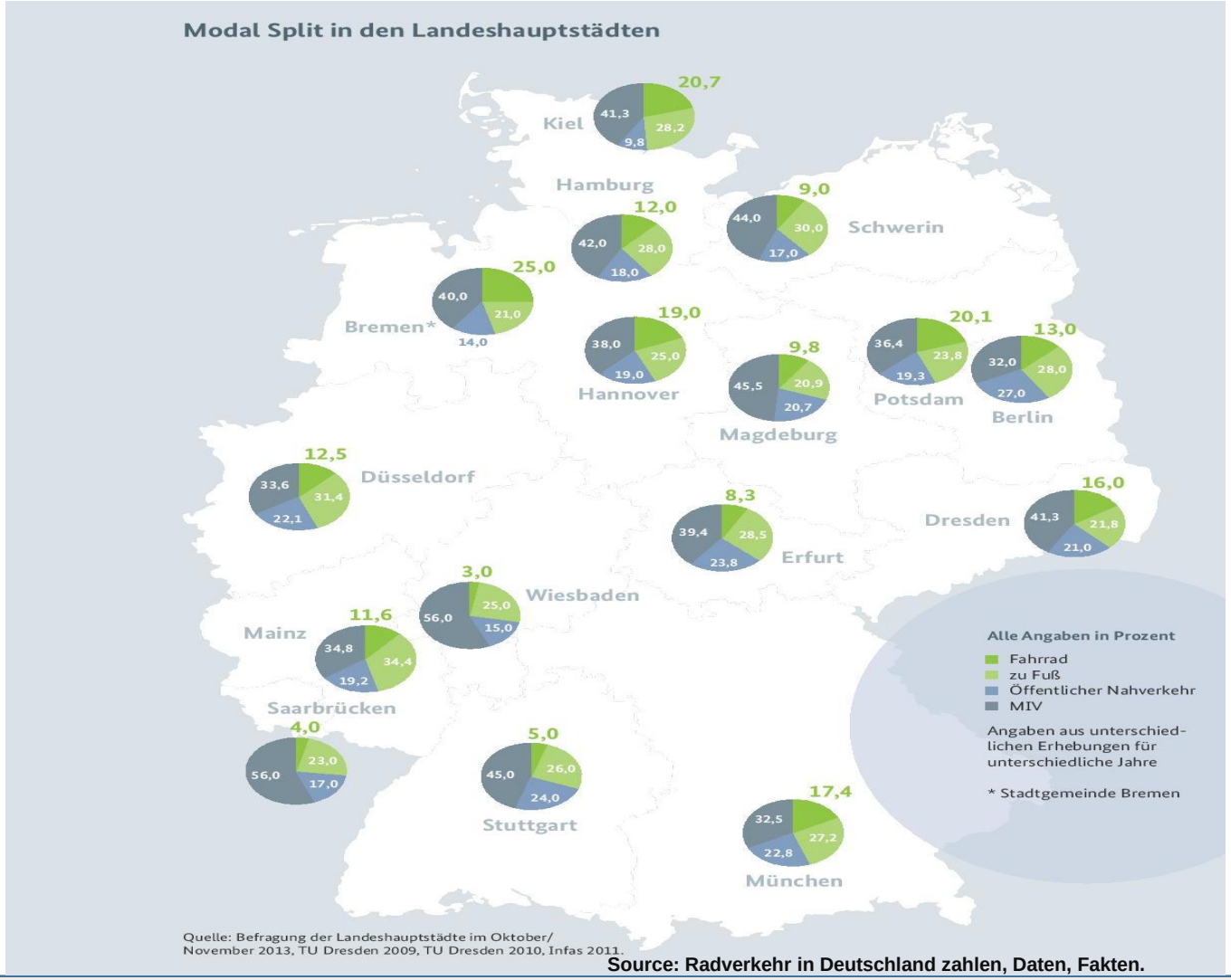


Angaben für 2008. Quelle: MiD 2008,
GU Frankfurt 2011, Infas 2010a, Infas 2010b, Infas 2011.

10 Radverkehr in Deutschland – Zahlen, Daten, Fakten

Source: Radverkehr in Deutschland zahlen, Daten, Fakten.

- **DISTRIBUTION BY CITIES:**



Source: Radverkehr in Deutschland zahlen, Daten, Fakten.

Literature Research:

- **DISTRIBUTION BY CITY:**

CITIES WITH THE MOST BICYCLISTS

THESE cities have the largest number of bicyclists riding on their streets.

CITY	POPULATION	NUMBER OF BIKE COMMUTERS	% OF BIKE COMMUTERS
NEW YORK, NY	8,336,697	36,496	1%
CHICAGO, IL	2,714,844	19,147	1.6%
PORTLAND, OR	603,650	18,912	6.1%
LOS ANGELES, CA	3,857,786	17,223	1%
SAN FRANCISCO CITY, CA	825,863	16,864	3.8%
SEATTLE CITY, WA	634,541	15,007	4.1%
PHILADELPHIA, PA	1,547,607	13,726	2.3%
WASHINGTON, D.C.	632,323	13,493	4.1%
MINNEAPOLIS, MN	392,871	9,688	4.5%
DENVER, CO	634,265	9,416	2.9%
MADISON, WI	240,315	8,375	6.2%
AUSTIN, TX	842,595	6,999	1.6%
SAN DIEGO, CA	1,338,354	6,929	1.1%
BOULDER, CO	101,812	6,560	12.1%
BOSTON, MA	637,516	6,536	2%
FORT COLLINS, CO	148,634	6,190	7.9%
TUCSON, AZ	524,278	6,189	2.8%
EUGENE, OR	157,984	6,121	8.7%
DAVIS, CA	66,009	5,830	19.1%
CAMBRIDGE, MA	106,456	5,067	8.5%
SACRAMENTO, CA	475,524	5,016	2.6%
OAKLAND, CA	400,740	5,012	2.7%
PHOENIX, AZ	1,488,759	4,784	0.7%
BERKELEY, CA	115,417	4,290	7.6%
TEMPE, AZ	166,862	3,966	4.5%

Source: 2013 American Community Survey data report.

Literature Research:

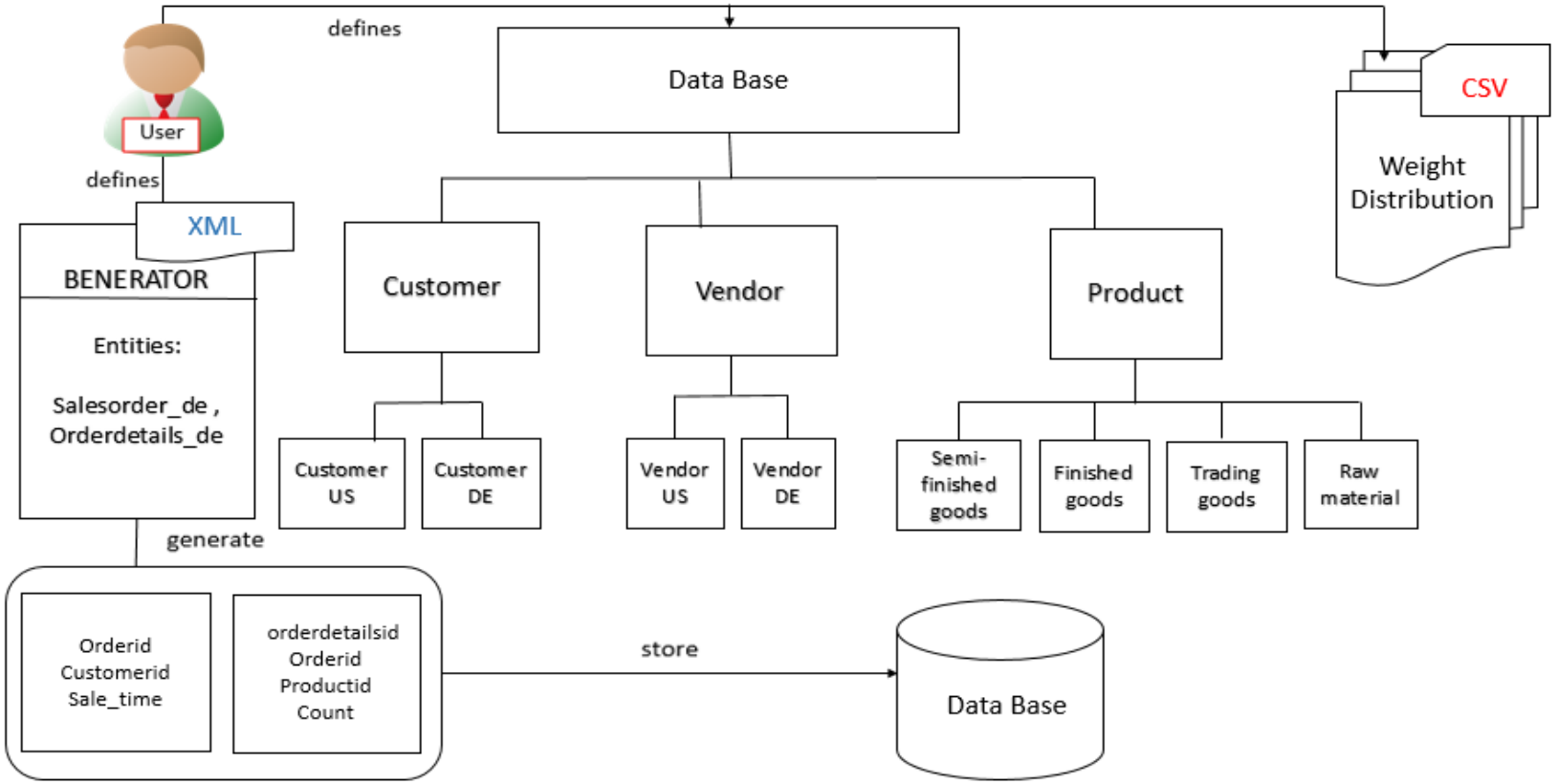
- **PRODUCT DISTRIBUTION:**
- Sales distribution per gender.
- Colour preferences.
- Percentage of sales by bikes, parts and accessories.

- **YEARLY DISTRIBUTION**:
- Population variation from 2009-2011.
- Number of bike users in particular year.

- **Benerator Configuration**

Benerator Configuration:

- **OVERVIEW:**



Benerator Configuration:

- **DESCRIPTOR.XML FILE:**

Populating Database

Creating Tables

Date Specification

Generating Entities and Analysing

Benerator Configuration:

- **POPULATING DATABASE:** Establishing connection for data base.

<database id="db"

url="jdbc:mysql://localhost:3306/datagenerator"

Driver="com.mysql.jdbc.Driver"

schema="datagenerator"

catalog="datagenerator"

user="***"**

password="*****"/>**

Benerator Configuration:

• **CREATING TABLES:** Tables are created to assign a path for generating entities.

Salesorders Table:

```
<execute target="db" >
create table salesorders(
orderid int AUTO_INCREMENT,
customerid int,
sale_time varchar(100),
PRIMARY KEY(orderid))
</execute>
```

Orderdetails Table:

```
<execute target="db" >
create table ordersdetails(
 orderdetailid int unique AUTO_INCREMENT,
 productid int,
orderid int,
 count int)
</execute>
```

Benerator Configuration:

- **Date Specification :** Each month is specified with unique identity using bean classes to define them globally.
- **Example: Date specification for January 2009**

```
<bean id="dtGen0901" class="DateTimeGenerator">
<property name='minDate‘ value='2009-01-01'/>
<property name='maxDate' value='2009-01-31'/>
<property name='dateGranularity'  value='00-00-01'  />
</bean>
```

Benerator Configuration:

- **Generating Entities :** Entities can be generated as per user requirement.

```
<generate name="salesorders_us" type="salesorders_us" count="67" consumer="db,ConsoleExporter">

<id name="orderid" type="int" min="1" max="67"  />

<variable name="weightings" source="weightings01.wgt.csv" distribution="weighted"/>

 <reference name="customerid" type="int" targetType="salesorders_us" source="db" selector="select id from
customer_us" nullable="false" cyclic="true" script="{weightings}"/>

<attribute name="sale_time" type="datetime" nullable="false" generator="dtGen0901"/>

<generate name="ordersdetails_us" type="ordersdetails_us"  minCount="1" maxCount="100"
consumer="db,ConsoleExporter">

<id name="orderdetailid"  generator="newIncrementalIdGenerator"  mode="ignored" />

<reference name="orderid" script="salesorders_us.orderid"/>

<variable name="weightings01" source="Hproduct_us.wgt.csv" distribution="weighted"/>

<reference name="productid" type="int" targetType="salesorders_us" source="db" selector="select pid from
product_us" nullable="false" cyclic="true" script="{weightings01}"/>

<attribute name="count" type="int" min="1" max="20" />

 </generate>

</generate>
```

- **Generating Entities and Analysis**

Generating Entities and Analysis:

- **TABLES CREATED:**

Salesorders Table:

orderid	customerid	sale_time
1	1	2009-01-22
2	4	2009-01-31
3	12	2009-01-29
4	11	2009-01-07
5	12	2009-01-04
6	7	2009-01-05
7	7	2009-01-13

Orderdetails Table:

orderdetailid	productid	orderid	count
1	15	1	17
2	11	1	6
3	6	1	20
4	9	1	20
5	12	1	5
6	18	1	12
7	13	1	20

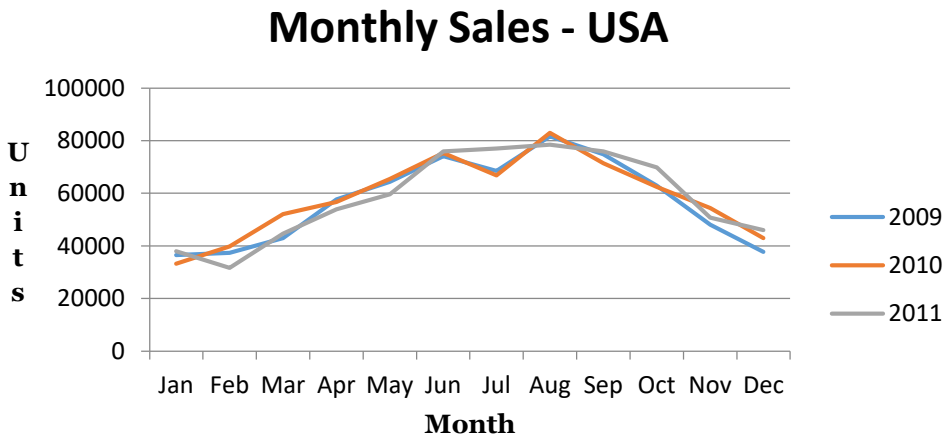
Generating Entities and Analysis:

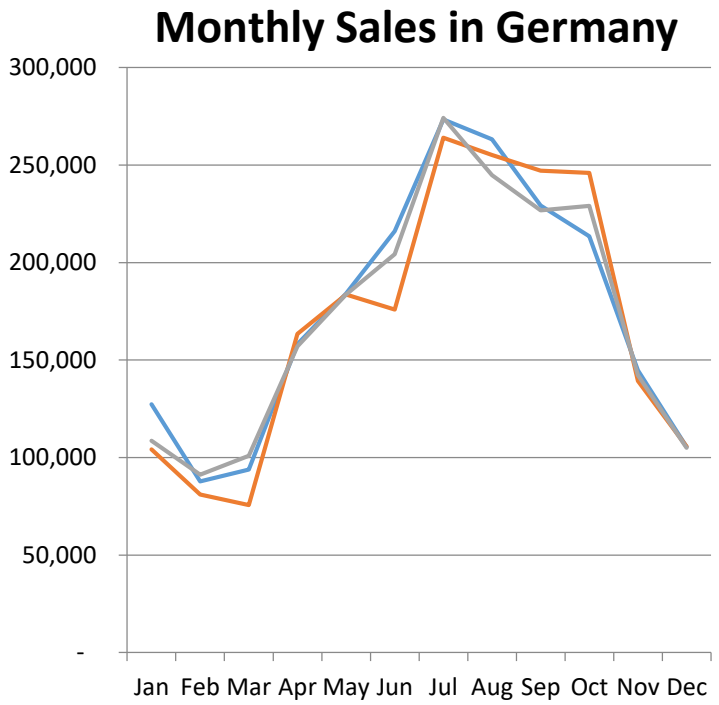
• JOIN clause is used to combine Product, Customer, Salesorders, Orderdetails tables to return required rows in sales-order table.

1	orderid	orderde	customerid	customername	pid	pname	sale_time	city	postalcode	countrycode
2	1	1	2	BIG APPLE BIKES	15	PROFESSIONAL TOURING BIKE (BLACK)	2009-01-08	NEW YORK CITY	10014	US00
3	1	2	2	BIG APPLE BIKES	6	REPAIR KIT	2009-01-08	NEW YORK CITY	10014	US00
4	1	3	2	BIG APPLE BIKES	7	ROAD HELMET	2009-01-08	NEW YORK CITY	10014	US00
5	1	4	2	BIG APPLE BIKES	14	MEN'S OFF ROAD BIKE	2009-01-08	NEW YORK CITY	10014	US00
6	1	5	2	BIG APPLE BIKES	15	PROFESSIONAL TOURING BIKE (BLACK)	2009-01-08	NEW YORK CITY	10014	US00
7	1	6	2	BIG APPLE BIKES	13	DELUXE TOURING BIKE (SILVER)	2009-01-08	NEW YORK CITY	10014	US00
8	1	7	2	BIG APPLE BIKES	10	WATER BOTTLE CAGE	2009-01-08	NEW YORK CITY	10014	US00
9	1	8	2	BIG APPLE BIKES	14	MEN'S OFF ROAD BIKE	2009-01-08	NEW YORK CITY	10014	US00
10	1	9	2	BIG APPLE BIKES	10	WATER BOTTLE CAGE	2009-01-08	NEW YORK CITY	10014	US00
11	1	10	2	BIG APPLE BIKES	13	DELUXE TOURING BIKE (SILVER)	2009-01-08	NEW YORK CITY	10014	US00
12	1	11	2	BIG APPLE BIKES	15	PROFESSIONAL TOURING BIKE (BLACK)	2009-01-08	NEW YORK CITY	10014	US00
13	1	12	2	BIG APPLE BIKES	13	DELUXE TOURING BIKE (SILVER)	2009-01-08	NEW YORK CITY	10014	US00
14	1	13	2	BIG APPLE BIKES	17	PROFESSIONAL TOURING BIKE (SILVER)	2009-01-08	NEW YORK CITY	10014	US00
15	1	14	2	BIG APPLE BIKES	14	MEN'S OFF ROAD BIKE	2009-01-08	NEW YORK CITY	10014	US00
16	2	15	1	ROCKY MOUNTAIN BIKES	15	PROFESSIONAL TOURING BIKE (BLACK)	2009-01-07	DENVER	80111	US00
17	2	16	1	ROCKY MOUNTAIN BIKES	13	DELUXE TOURING BIKE (SILVER)	2009-01-07	DENVER	80111	US00
18	2	17	1	ROCKY MOUNTAIN BIKES	3	FIRST AID KIT	2009-01-07	DENVER	80111	US00
19	2	18	1	ROCKY MOUNTAIN BIKES	7	ROAD HELMET	2009-01-07	DENVER	80111	US00
20	2	19	1	ROCKY MOUNTAIN BIKES	18	WOMEN'S OFF ROAD BIKE EN	2009-01-07	DENVER	80111	US00
21	2	20	1	ROCKY MOUNTAIN BIKES	13	DELUXE TOURING BIKE (SILVER)	2009-01-07	DENVER	80111	US00
22	2	21	1	ROCKY MOUNTAIN BIKES	16	PROFESSIONAL TOURING BIKE (RED)	2009-01-07	DENVER	80111	US00
23	2	22	1	ROCKY MOUNTAIN BIKES	9	WATER BOTTLE	2009-01-07	DENVER	80111	US00
24	2	23	1	ROCKY MOUNTAIN BIKES	15	PROFESSIONAL TOURING BIKE (BLACK)	2009-01-07	DENVER	80111	US00
25	2	24	1	ROCKY MOUNTAIN BIKES	12	DELUXE TOURING BIKE (RED)	2009-01-07	DENVER	80111	US00

Generating Entities and Analysis:

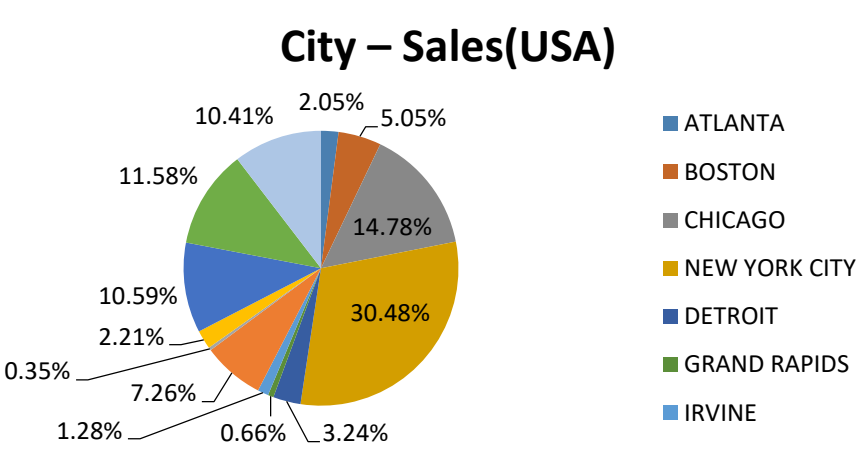
- Analysis were done as per generated data.

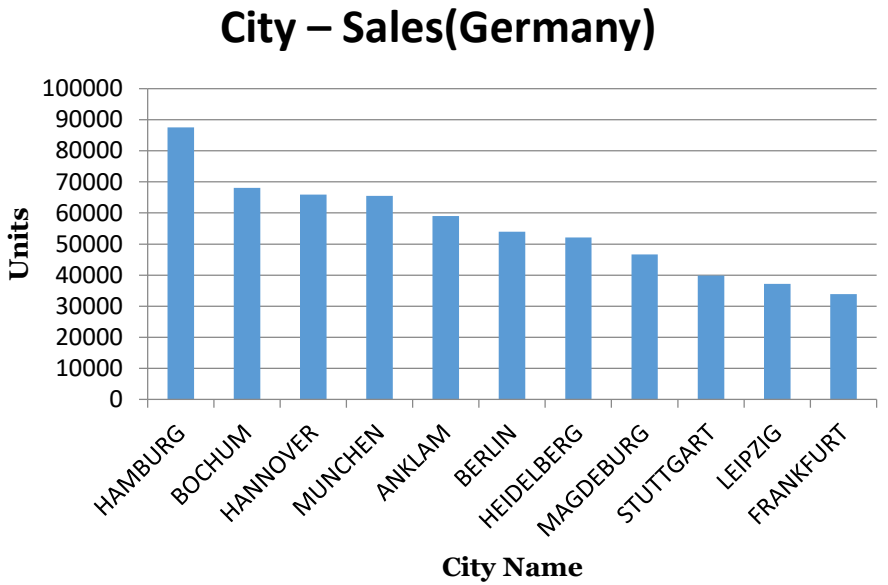




Generating Entities and Analysis:

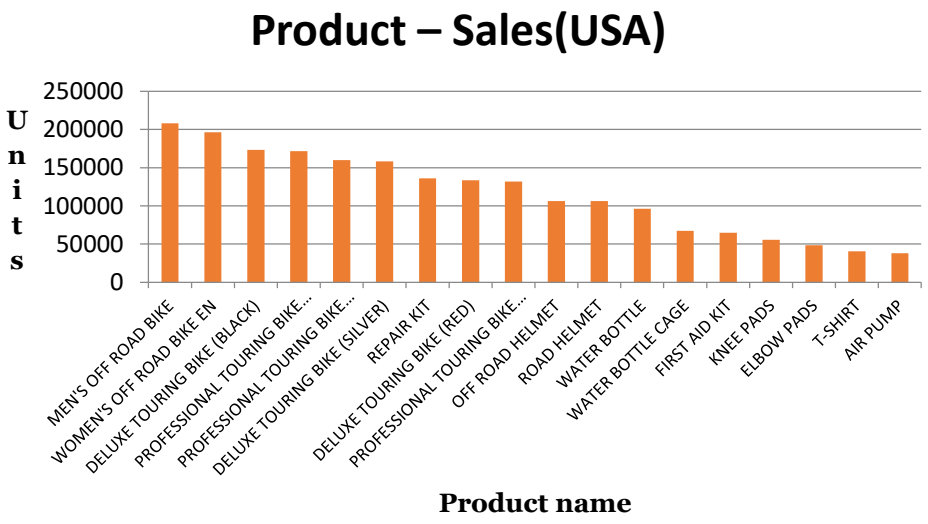
- Analysis were done as per generated data.

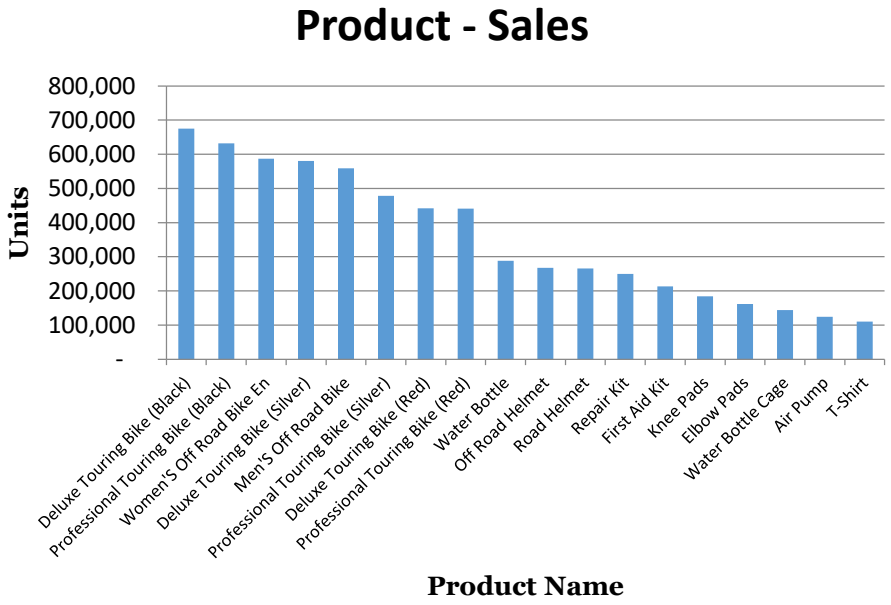




Generating Entities and Analysis:

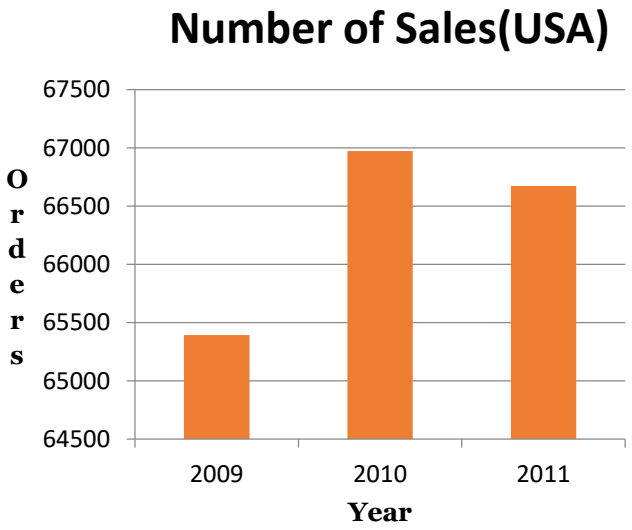
- Analysis were done as per generated data.

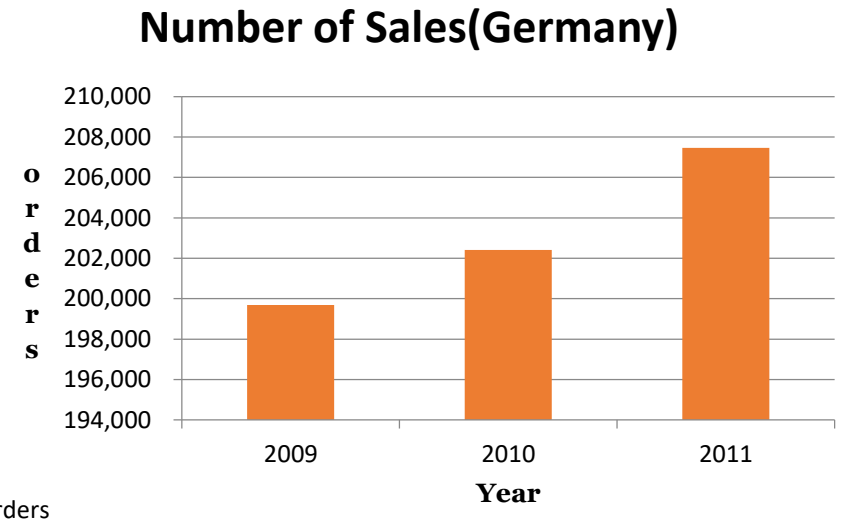




Generating Entities and Analysis:

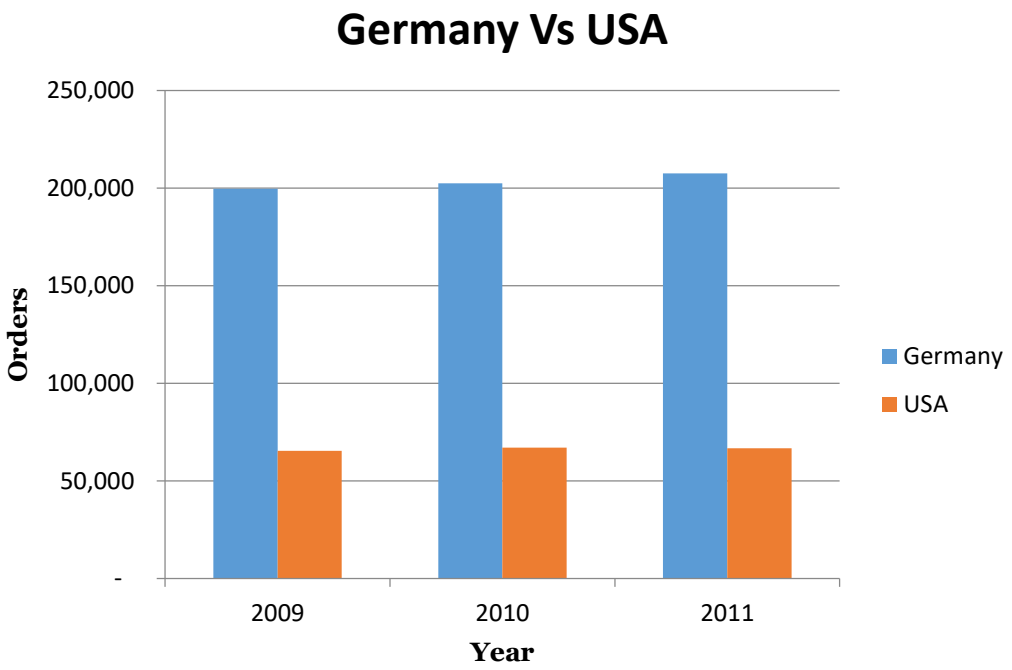
- Analysis were done as per generated data.





Generating Entities and Analysis:

- Analysis were done as per generated data.



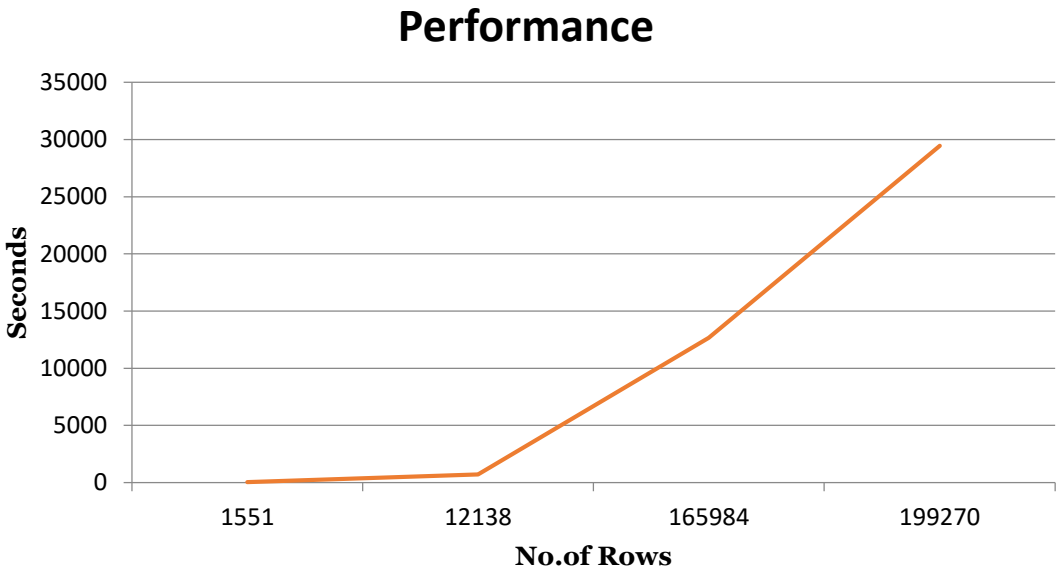
- **Performance and Limitations**

Performance and Limitations:

- **PERFORMANCE:**

- Performance of benerator was tested for different number of datasets generated.

- Generating time increases as the number of entities to be generated increases.



Performance and Limitations:

- **LIMITATIONS:**

- Development for the tool ended in 2009, with release v 0.9.8.
- The online forum is inactive and no longer accepts registrations.
- The documentation is not exhaustive enough to cover all use cases.

- **Conclusion**

Conclusion:

• We have generated systematic data sets, which were organized in format using benerator tool, and data is extracted in Excel sheet. Later we have analyzed the data of increment and decrement in sales as per seasonal conditions, city wise distribution of bikes sales. Also found the differences of sales in Germany and in the USA on different time scales. These output realistic datasets can be used for analysis purposes.

- **References**

References:

[1] V. Ayala-Rivera, P. McDonagh, T. Cerqueus, L. Murphy, Synthetic data generation using benerator tool, arXiv preprint arXiv:1311.3312 (2013).

[2] P. K ”u nzel, The radschnellweg regio velo 01 nsterland in westm ”u. route, use, planning and financing .

[3] K. McLeod, D. Flusche, A. Clarke, Where we ride: Analysis of bicycling in American cities (2013).

[4] B. G. G. B. F. F. H. I. T. N. P. S. S. Members of national Bicycle Industry Associations in 14 different countries: Austria, Belgium, Turkey., European bicycle market (2015) 10–75.

[5] C. Emond, W. Tang, S. Handy, Explaining gender differences in bicycle behavior, in: Active Living Research Conference February, volume 19.

References:

[6] M. Nezhad, K. Kavehnezhad, et al., Choosing the right color: a way to increase sales, International Journal of Asian Social Science 3 (2013) 1442–1457.

[7] J. A. Bellizzi, R. E. Hite, Environmental color, consumer feelings, and purchase likelihood, Psychology & marketing 9 (1992) 347–363.

[8] N. Torslov, Cycling,health and safety, OECD/International Transport Forum(2013),Cycling,Health and Safety OECD Publishing/ITF.

[9] Houkjr, Kenneth, Kristian Torp, and Rico Wind. ”Simple and realistic data generation.” Proceedings of the 32nd international conference on Very large data bases. VLDB Endowment, 2006.

References:

[10] Gabriel Ghinita, Yufei Tao, and Panos Kalnis. On the Anonymization of Sparse High-Dimensional Data. In ICDE, pages 715724. Ieee, April 2008.

[11] Lu, Qing Chang, et al. ”Inter-city travel behavior adaptation to extreme weather events.” Journal of Transport Geography 41 (2014): 148-153.

[12] Cui, Yuchen, Sabyasachee Mishra, and Timothy F. Welch. ”Land use effects on bicycle ridership: a framework for state planning agencies.” Journal of Transport Geography 41 (2014): 220-228.

[13] Phung, Justin, and Geoff Rose. ”Temporal variations in usage of Melbourne’s bike paths.” Proceedings of 30th Australasian Transport Research Forum, Melbourne . Of 2007.

References:

[14] Brandenburg, Christiane, Andreas Matzarakis, and Arne Arnberger. “Weather and cycling-a first approach to the effects of weather conditions on cycling”Meteorological applications (2007): 61-67.

[15] Whiting, Mark A., Jereme Haack, and Carrie Varley. ”Creating realistic, scenario-based synthetic data for test and evaluation of information analytics software.” Proceedings of the 2008 Workshop on BEyond time and errors: novel evaLuation methods for Information Visualization. ACM, 2008.

[16] T Varga and Horst Bunke. Generation of synthetic training data for an HMM-based handwriting recognition system. Document Analysis and Recognition, 2003.


thank you for your attention