Hurtownie danych – Spr. 5.

PWr. WIZ, Data: 11-12.04.2022

Student	----------	Ocena
Indeks	256305	
Imię	Grzegorz	
Nazwisko	Dzikowski	

Zestaw składa się z 1 zadania. Pamiętaj o podaniu nr. indeksu oraz imienia i nazwiska.

Zad. 1. (Prezentacja wszystkich punktów tego zadania na zajęciach 11-12.04.2022)

Proces tworzenia hurtowni danych powinien być poprzedzony zrozumieniem „potrzeb biznesu" oraz rzeczywistości (dziedziny problemowej) reprezentowanej przez dostępne zasoby danych. Realizacja poniższego zadania ma uzmysłowić występujące problemy w określonym (wybranym) wycinku rzeczywistości, a następnie umożliwić zidentyfikowanie (określenie) potrzeb, celu i możliwości analiz biznesowych, by wspierać procesy decyzyjne (podejmowanie właściwych decyzji biznesowych).

Projekt HD – propozycja tematu

Proszę przygotować zakres realizacji projektu zgodnie z poniższą specyfikacją oraz przedyskutować propozycję projektu z osobą prowadzącą zajęcia. Poczynione uzgodnienia zarejestrować w formie wniosków. **Na zajęciach laboratoryjnych należy przedstawić na forum grupy swoją propozycję tematu projektu (uzasadniając celowość i jego główne elementy 1.1 – 1.6) wykorzystując PowerPoint**.

Zakres opracowania projektu HD – cz. 1.

1.1. Tytuł projektu Zanieczyszczenia wód w państwach Europy Środkowej w latach 2016 - 2019

1.2. Charakterystyka dziedziny problemowej

System wodny składa się z wielu części: z jezior, wód gruntowych, rzek, sztucznych zbiorników, brzegów morskich. Każdy z elementów stanowi istotną część działania państw. Od wody pitnej, przez transport po łowiectwo i rybactwo. Zanieczyszczenie którejkolwiek z tych części może mieć ogromne konsekwencje finansowe czy społeczne dla państw. Także zanieczyszczenie każdej z części w długim okresie czasu może spowodować przeniesienie zanieczyszczeń na inne części tego systemu.

1.3. Opis obszaru analizy wraz z uzasadnieniem (wybrany fragment dziedziny, przeznaczony do szczegółowej analizy i opracowania hurtowni danych)

Analiza będzie dotyczyła Polski oraz krajów sąsiadujących, należących do UE: Niemiec, Czech, Słowacji i Litwy. Identyfikacja składu zanieczyszczeń dostarczy ważnych informacji regulatorom

unijnym. Identyfikacja lokalizacji zanieczyszczeń pozwoli precyzyjnie określić, który kraj wymaga działania. Ponadto identyfikacja typów zanieczyszczonych wód pozwoli na określenie potencjalnego ich wpływu na społeczeństwo oraz kraj, oraz dostarczy przydatnych danych miejscom je oczyszczające.

1.4. Problemy

P1 – Problemy zdrowotne związane z brudną wodą

P2 – Wzrost kosztów oczyszczania wody pitnej

P3 – Spadek połowów ryb

P4 – Rozwój niebezpiecznych pasożytów i roślin

P5 – Zniszczenie fauny i flory rzek i jezior

P6 – Obniżenie walorów turystyczno-krajoznawczych

1.5. Cel przedsięwzięcia

1. Wykrycie głównych lokalizacji źródeł zanieczyszczeń
2. Określenie składu zanieczyszczeń w regionach
3. Zbadanie czynników czasowych: miesiąc, dzień

1.5.1. Oczekiwania i potrzeby w zakresie wsparcia podejmowania decyzji

Z perspektywy samorządów oraz władz państwa – analiza dostarczy informacji na temat:

- Głównych źródeł i składu zanieczyszczeń
- Rejonów, w których zanieczyszczenia występują
- Porównanie stopnia zanieczyszczeń w stosunku do innych krajów sąsiadujących

Z perspektywy oczyszczalni:

- Składu zanieczyszczeń, oraz pespektywa na przyszłość

Z perspektywy turystów:

- Które wody są czyste, a które nie

1.5.2. Zakres analizy – badane aspekty

Analiza dostarczając odpowiedzi na te pytania powinna dostarczyć informacji, które będą pomocne w zidentyfikowaniu składu oraz lokalizacji zanieczyszczeń, które regiony są ich największym producentem, oraz pomóc dostosować regulację prawne dotyczące emisji zanieczyszczeń. Na powyższe dane można nałożyć czas emisji zanieczyszczeń, który pozwoli jeszcze dodatkowo analizować czynniki czasowe emisji.

1.6. Źródła danych

1.6.1. Lokalizacja, format, dostępność

Dane dostępne są na stronie https://www.eea.europa.eu/data-and-maps/data/waterbase-water-quality-icm-1 w formacie CSV lub SQLITE. Dane są dostępne do pobrania za darmo, bez żadnych ograniczeń. Dodatkowo, są dostępne tak zwane dane obiektów przestrzennych (spatialobject) (abstrakcyjną reprezentację zjawiska świata rzeczywistego związaną z określonym położeniem lub obszarem geograficznym)[1]. Dodatkowo, dostępne są też pomocnicze bazy związane z obiektami wodnymi: Ilość wody w zbiornikach, biologia w zbiornikach wodnych, emisji substancji pomocniczych i niebezpiecznych przez kraje, oraz dane na temat przepływu wody wraz z ich stanem chemicznym.

1.6.2. Wstępna ocena (liczba rekordów, zakres czasowy danych - faktów)

Główna tabela (Waterbase – water quality ICM) zawiera 51 321 704 rekordów (13GB CSV) ze wszystkich krajów Unii Europejskej. Zakres czasowy to od 1900 do 2020 roku, jednak według mojej opinii – sensowne dane są w latach 2006 – 2019. Ze względu na liczbę danych zdecydowałbym się ograniczyć je jeszcze bardziej, do lat 2016-2019[2] Dodatkowo, dane zawierają 39 unikalnych krajów. Jednak ograniczyłbym je jeszcze bardziej, do tych krajów, które poddajemy analizie: Polsce, Niemiec, Czech, Słowacji, Litwy

1.6.3. Fakty

Lp.	Fakt	Miary
1.	Pomiar Jakości Wody	Czas, Lokalizacja, Skład Chemiczny, Kategoria zbiornika wodnego

1.6.4. Kontekst analizy faktów np. czas (ziarnistość), lokalizacja, warunki pogodowe, itd.

Lp.	Kontekst analizy - wymiary	Własności
1.	Czas	Ziarnistość: 1 dzień, dane lat 2016-2019. Dane przydatne ze względu na zmiany czasowe oraz trendy
2.	Lokalizacja	Ziarnistość: Lokalizacja GPS stacji. Dane przydatne na potrzeby porównania z innymi krajami, oraz zlokalizowania źródeł zanieczyszczeń

[1] https://inspire.ec.europa.eu/glossary/SpatialObject
[2] Tabela 1 Liczba danych na każdy rok

3.	Skład chemiczny	981 różnych składników. Dane przydatne na potrzeby analizy składu chemicznego oraz głównych regulacji emisji

ZAŁĄCZNIKI:

ROK	Liczba danych
1900	9
1941	1
1944	9
1960	75
1961	229
1962	378
1963	458
1964	566
1965	956
1966	891
1967	526
1968	807
1969	869
1970	780
1971	857
1972	943
1973	2219
1974	3165
1975	4297
1976	7509
1977	6942
1978	8231
1979	8496
1980	8908
1981	9286
1982	11413
1983	12537
1984	10891
1985	11315
1986	12541
1987	18616
1988	20320
1989	36437
1990	63312
1991	65560
1992	77490
1993	88228
1994	98415
1995	105945
1996	118109
1997	122139
1998	126046
1999	156391

2000	210833
2001	187604
2002	222405
2003	336500
2004	481214
2005	640014
2006	1307074
2007	2785007
2008	2179022
2009	1922818
2010	2137075
2011	2326176
2012	2818056
2013	2788291
2014	2943605
2015	4507432
2016	4354989
2017	4725395
2018	5452797
2019	7848476
2020	30331

Tabela 1 Liczba danych na każdy rok

Kod Kraju	Liczba danych
AL	4082
AT	1116738
BA	24113
BE	1052810
BG	277397
CH	234201
CY	189729
CZ	1449098
DE	242484
DK	1193548
EE	216919

EL	259001
ES	128471
FI	658610
FR	21384297
HR	385592
HU	278784
IE	1332945
IS	4675
IT	7910322
LI	3993
LT	301334
LU	4321
LV	130638
ME	591
MK	50294
MT	9792
NL	774848
NO	328088
PL	3418638
PT	98075
RO	33693
RS	430933
SE	540472
SI	190210
SK	778269
TR	1315
UK	5969727

XK	19179

Tabela 2 Liczba danych per kraj