Raport hurtowni danych o tematyce zdrowia psychicznego

Wstęp

Zdrowie psychiczne od zawsze było bardzo interesującym obszarem badawczym, ale wraz z zachodzącymi zmianami w społeczeństwie jego wartość stale wzrasta, osiągając niebywałą liczbę zainteresowanych. Problemy, z którymi zmierzają się ludzie w obecnych czasach, są przeróżne, aczkolwiek wiele z nich jest silnie zakorzenionych w psychice, dlatego kluczowe jest przyjrzenie się tej tematyce bliżej.

Celem projektu – hurtowni danych – było stworzenie bazy, która odpowie na część nurtujących nas pytań dziedzinie zdrowia psychicznego.

Analiza problemu, przegląd danych

Do tematyki postawionego przez nas problemu można podejść na wiele sposobów. Na podstawie wybranej ścieżki, będzie można odpowiedzieć na odmienne pytania badawcze, ponieważ w zależności od dokonanego wyboru zostaną wyselekcjonowane inne dane. Istnieje wiele wartościowych źródeł o tematyce zdrowia psychicznego, takie jak: Eurostat, World Health Organization, Office of National Statistics (Wielka Brytania), oficjalna strona rządowa (GOV), Główny Urząd Statystyczny, Narodowy Fundusz Zdrowia i inne. Dane gromadzone są przez organizacje działające w konkretnych państwach, co powoduje, że ich zasięg jest ograniczony do danego kraju, a także przez międzynarodowe organizacje.

Nasz zespół, po dogłębnej analizie tematu oraz eksploracji dostępnych źródeł i danych doszedł do wniosku, że warto pochylić się nad problemem zaburzeń i ich wpływu na funkcjonowanie jednostek w codziennym życiu, a także działaniem, jakie podejmują szpitale w celu niesienia pomocy pozytywnie zdiagnozowanym jednostkom.

W związku z tym postawiono kilka problemów badawczych:

- rodzaj schorzenia psychicznego uwarunkowany jest miejscem zamieszkania (województwo, mniejsze miasta ?=? mniejsza świadomość problemu);
- liczba szpitali psychiatrycznych w danym województwie, a liczba zgonów (im mniej szpitali, tym więcej osób nieleczonych, zatem powinien zostać zaobserwowany zwiększony współczynnik umieralności);
- liczba samobójstw, a liczba zdiagnozowanych, poddanych leczenia osób (wzrost samobójstw obserwowalny jest w województwach o mniejszej, całkowitej liczbie przypadków;
- typ szpitalu ma wpływ na liczbę poddanych leczeniu osób (zatem szpital psychiatryczny powinien mieć większą liczbę zdiagnozowanych chorych i cechować się mniejszą liczbą zgonów);
- rodzaj schorzenia a liczba zaświadczeń (niektóre schorzenia odznaczają się bardziej na tle pozostałych → ich występowanie jest częstsze);

- w jednym z województw silnie przeważa liczba zgonów (większa liczba osób chorych);
- okres pandemiczny przyczynił się do wzrostu liczby osób z zaburzeniami psychicznymi;
- zbyt mała liczba łóżek w szpitalu psychiatrycznym w porównaniu do liczebności ludności na jedno łóżko powoduje zwiększoną liczbę samobójstw;
- wpływ płci na liczbę zachorowań;
- zależność wieku od liczby osób cierpiących na zaburzenia na tle psychicznym.

Dane w tym zakresie po analizie wielu źródeł udało się wydobyć ze stron:

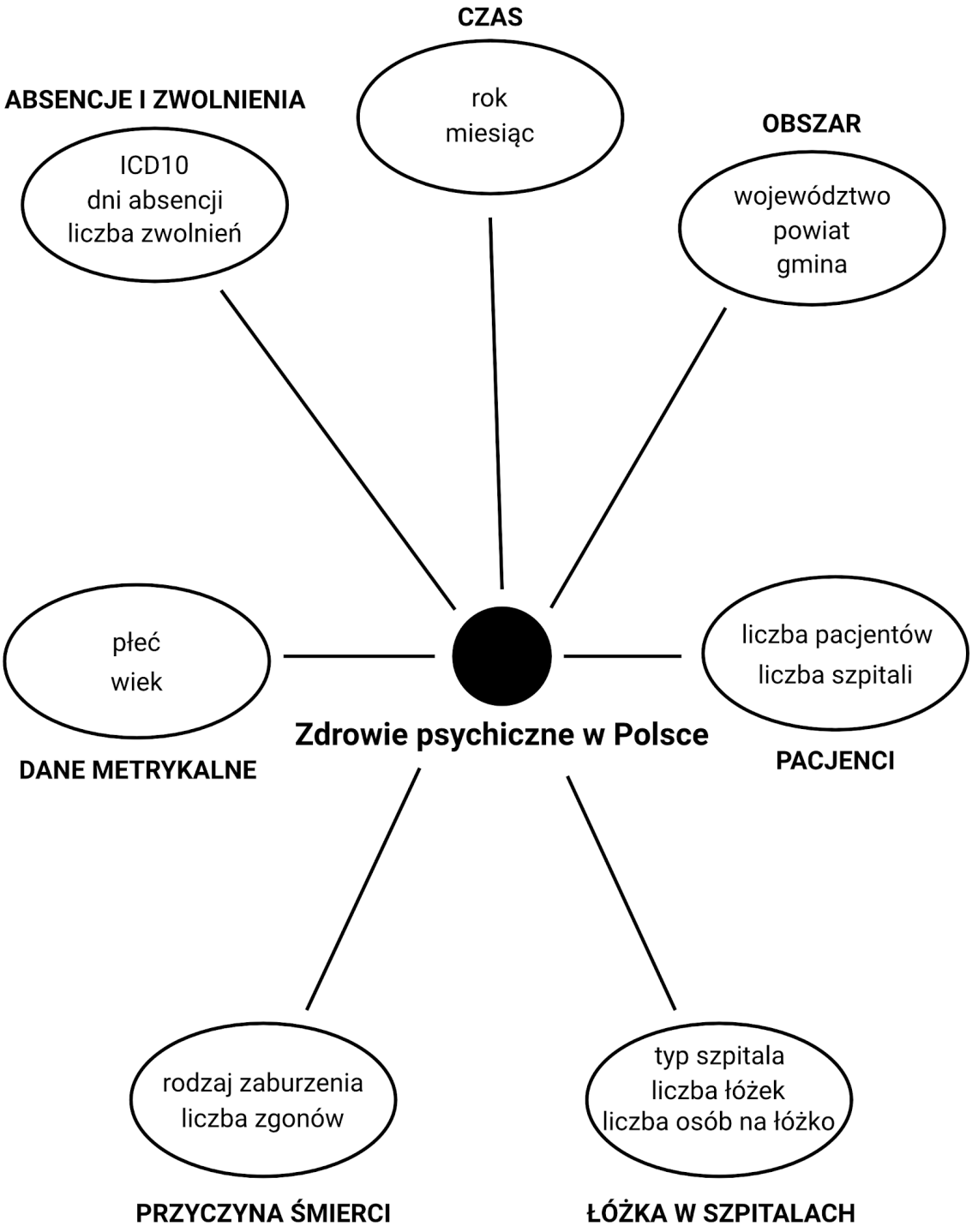
- Eurostatu (https://tinyurl.com/4hsu87pe),
- Generalnego Urzędu Statystycznego (https://bdl.stat.gov.pl/bdl/dane/podgrup/temat)
- oficjalnej strony rządowej, z Bazy Analiz Systemowych i Wdrożeniowych (https://basiw.mz.gov.pl/mapy-informacje/mapa-2022-2026/analizy/).

Etapy projektowania bazy danych

Proces projektowania bazy danych jest wieloetapowym przedsięwzięciem, jednak można wyróżnić w nim trzy główne etapy:

1. **Model koncepcyjny**

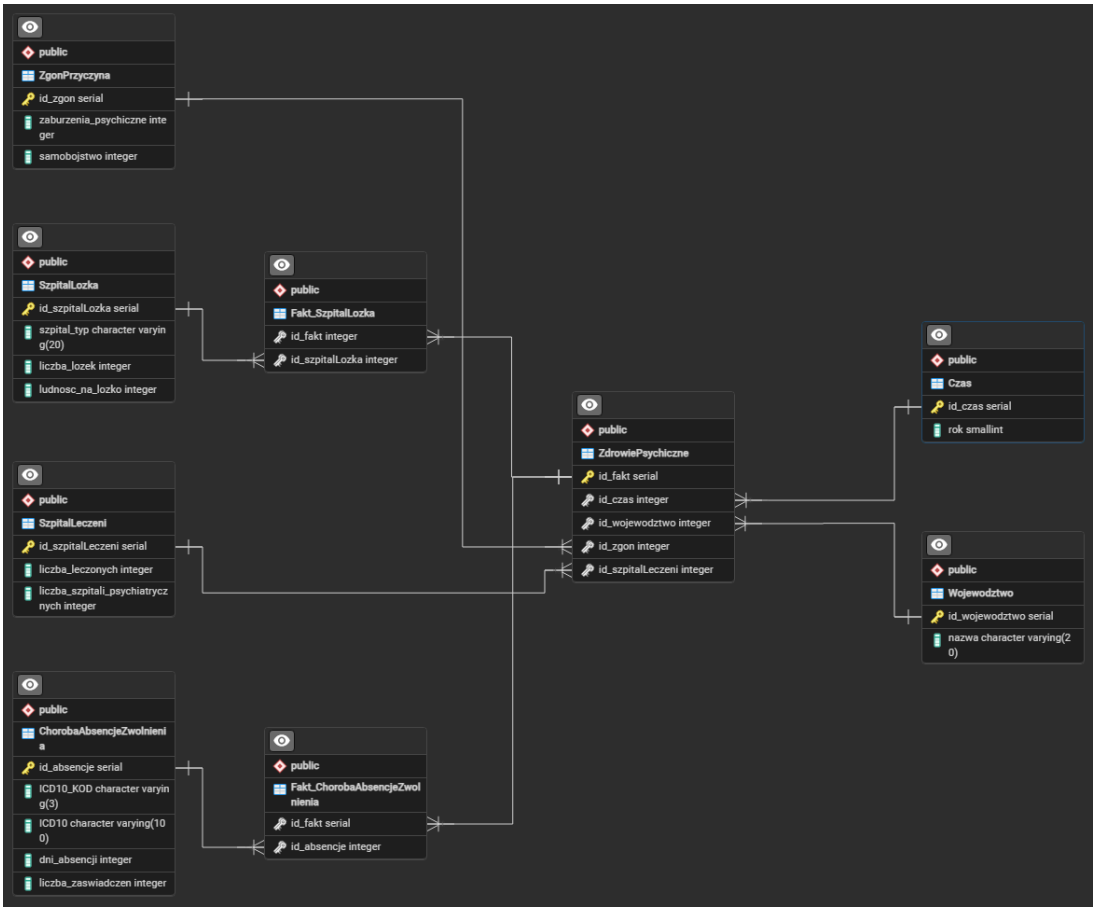
   Model skupia się na zebraniu niezbędnych wymagań, które będzie miała za zadanie spełniać hurtownia danych. W naszym przypadku, w oparciu o wstępną analizę wybranego zagadnienia, skupiłyśmy się głównie na zbiorach ogólnodostępnych. Można je odnaleźć na rzetelnych stronach internetowych. Do wstępnego zarysu problemu wykorzystałyśmy metodę kropki, która służy do wizualnego przedstawienia głównego punktu badań – zdrowia psychicznego – w celu wykazania zależności między innymi powiązaniami.


CZAS
rok
miesiąc
ABSENCJE I ZWOLNIENIA
ICD10
dni absencji
liczba zwolnień
OBSZAR
województwo
powiat
gmina
płeć
wiek
DANE METRYKALNE
Zdrowie psychiczne w Polsce
liczba pacjentów
liczba szpitali
PACJENCI
rodzaj zaburzenia
liczba zgonów
PRZYCZYNA ŚMIERCI
typ szpitala
liczba łóżek
liczba osób na łóżko
ŁÓŻKA W SZPITALACH

W ramach arkusza obiektów dobrym rozwiązaniem okazał się po prostu Excel, gdzie każda kolumna stanowi reprezentację atrybutu, a każdy wiersz odzwierciedla jednostkę danych. Arkusz obiektów aktualizowano na kolejnych etapach projektowania hurtowni, a rezultaty, w wersji roboczej, zostały umieszczone w załączonym pliku *arkusz_obiektów.xlsx*.
Zakres danych został okrojony do lat od 2010 do 2022, dzięki czemu można łatwo uchwycić (brak) wpływ(u) na liczbę zdiagnozowanych zaburzeń psychicznych. Uwzględnione zostały dane z obszaru Polski, ze wszystkich województw. Z pozyskanych zbiorów danych wyodrębniono najciekawsze atrybuty, uwzględniając specyficzny charakter dostępnych danych (w tym ich dużą ogólność) oraz ograniczenia związane z późniejszą potrzebą połączenia danych w spójne, logiczne relacje. Wśród dokonanych wyborów można wyróżnić: liczbę zgonów w stosunku do rodzaju zaburzeń psychicznych, liczbę samobójstw, typ szpitala, liczbę łóżek a ich zapotrzebowanie, rodzaj choroby, liczbę zaświadczeń oraz absencji związanych ze zdrowiem psychicznym, liczbę porad oraz hospitalizacji, a to wszystko względem województwa i daty (roku).

2. Model logiczny

Założeniem modelu logicznego jest przekształcenie zdefiniowanej wcześniej koncepcji w formalny model danych. Wobec tego, na tym etapie projektowania najważniejsza okazuje się architektura implementowanej hurtowni. W ramach jej tworzenia przygotowano diagram ERD przedstawiający strukturę hurtowni.



Zdecydowano się na strukturę gwiazdy, która pozwoli na szybsze tworzenie zapytań oraz mniejsze skomplikowanie. Hurtownia z założenia nie miała być bardzo dużych rozmiarów, więc wczytywanie danych nie powinno stanowić tu problemu. Przy modelowaniu relacji zastosowano notację Barkera (tzw. "kruczej stopki"). Dla relacji krotności jeden-do-wielu klucz obcy z encji *ZgonPrzyczyna* oraz *SzpitalLeczeni* przechowywany jest w głównej encji faktów: *ZdrowiePsychiczne*. Dla części danych (encje *SzpitalLozka* oraz *ChorobaAbsencjeZwolnienia*) zdecydowano się na połączenie ich za pomocą encji złączeń po ich kluczach głównych, w związku z relacjami wiele-do-wielu z encją *ZdrowiePsychiczne*, której klucz główny *id_fakt* także znajduje się w encjach o tej funkcji.
Diagram przedstawia w sposób czytelniejszy związki zachodzące pomiędzy encjami oraz ich atrybutami. Pozwolił także na spostrzeżenie wcześniej niezauważonych niespójności, które pojawiły się na etapie modelowania koncepcyjnego i ich modyfikację bądź wyeliminowanie. Tak też stało się z dodatkową encją zawierającą informacje metrykalne dla danych (płeć, zakres wieku) – patrząc na schemat oraz dostępne dane, nasz zespół uznał, że najlepszą decyzją (przynajmniej w obecnym momencie) jest wyeliminowanie tej tabeli, aby zachować spójność w obrębie całej struktury bazy.

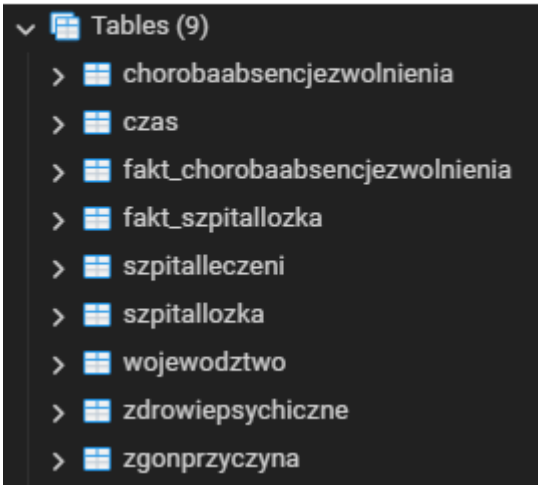
3. Model fizyczny

Model fizyczny to już ostatni etap projektowania hurtowni, który polega na przekształceniu modelu logicznego w rzeczywistą bazę danych. W tym kroku nastąpiła implementacja schematu, a więc przekształceniu uległy encje oraz ich atrybuty na tabele i kolumny. Do implementacji wykorzystano system zarządzania bazami danych PostgreSQL. Zadbano, aby w strukturze tabel kolumny nie przechowywały wartości *null*. W celu zachowania spójności w nazewnictwie, nazwy tabel zaczynają się od dużych liter – w konwencji *PascalCase*. Poszczególne kolumny nazwano z zastosowaniem notacji *snake_case*. Klucze poszczególnych tabel zaczynają się od *id_*. Takie nazewnictwo zostało zachowane w obszarze całego projektu, również na diagramie ERD i w plikach z danymi.

Fizyczna implementacja modelu została załączona w pliku *hurtownia.sql*. Skrypt definiuje strukturę bazy oraz zapełnia poszczególne tabele wcześniej przygotowanymi danymi.

Dane wczytywane do tabel odpowiednio przygotowano: na początkowym etapie połączono pliki z poszczególnych lat oraz kolumny o zbliżonej tematyce, dostępne w osobnych plikach. Dla części danych należało zsumować wartości z poszczególnych obszarów, aby ujednolicić dane do poziomu województwa, wybranego w implementowanej bazie. Następnie ujednolicono nazewnictwo kolumn i wartości oraz sformatowano dane do odpowiednich typów danych. Dane z plików Excel ostatecznie zostały przekonwertowane do plików tekstowych. Celem było uniknięcie problemów z wgrywaniem danych do bazy, które pojawiły się przy plikach innego typu. Wszystko to z użyciem języka R, który wykorzystano także na etapie analizy danych.

Tabele w hurtowni danych



Złączenie tabel z tabelą faktów

ZdrowiePsychiczne | Czas | Województwo | ZgonPrzyczyna | SzpitalLeczeni

	id_fakt integer	id_czas integer	id_wojewodztwo integer	id_zgon integer	id_szpitalleczeni integer	id_czas integer	rok smallint	id_wojewodztwo integer	nazwa character varying (20)	id_zgon integer	zaburzenia_psychiczne integer	samobojstwo integer	id_szpitalleczeni integer	liczba_leczonych integer	liczba_szpitali_psychiatrycznych integer
1	1	1	1	1	1	1	2010	1	dolnośląskie	1	205	529	1	11589	5
2	2	2	1	2	2	2	2011	1	dolnośląskie	2	207	520	2	12550	5
3	3	3	1	3	3	3	2012	1	dolnośląskie	3	286	543	3	14053	6
4	4	4	1	4	4	4	2013	1	dolnośląskie	4	41	656	4	13275	6
5	5	5	1	5	5	5	2014	1	dolnośląskie	5	89	563	5	13294	6
6	6	6	1	6	6	6	2015	1	dolnośląskie	6	248	477	6	13412	5
7	7	7	1	7	7	7	2016	1	dolnośląskie	7	249	405	7	13517	5
8	8	8	1	8	8	8	2017	1	dolnośląskie	8	278	371	8	12957	5
9	9	9	1	9	9	9	2018	1	dolnośląskie	9	340	381	9	12614	5

Total rows: 208 of 208 | Query complete 00:00:00.054

Fakt_ChorobaAbsencjeZwolnienia

	id_fakt [PK] integer	id_absencje [PK] integer
1	10	1014
2	10	1008
3	10	1009
4	10	1010
5	10	1011
6	10	1012
7	10	1013
8	10	1000
9	10	1015

Total rows: 1000 of 2302 | Query complete 00:00:00.056

Fakt_SzpitalLozka

	id_fakt [PK] integer	id_szpitallozka [PK] integer
1	11	1
2	11	4
3	128	55
4	128	58
5	141	61
6	141	64
7	154	67
8	154	70
9	167	73

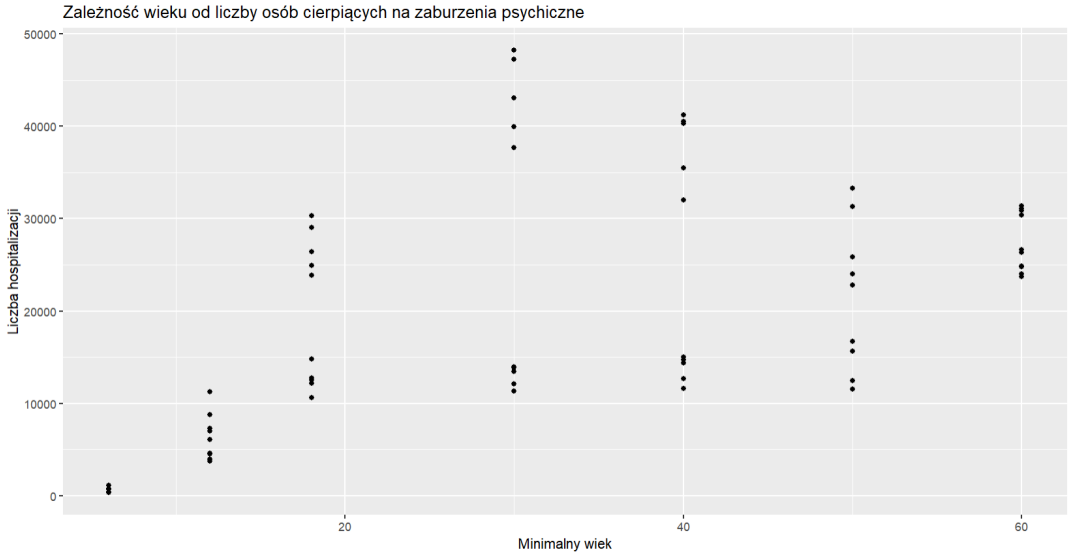
Total rows: 96 of 96 Query complete 00:00:00.062

Analiza danych

Do przeprowadzenia analizy danych został wykorzystany język R, który umożliwił utworzenie dość przyzwoitych wykresów. Głównie skupiłyśmy się na próbie uchwycenia zależności zawartych w zagadnieniach badawczych, które wcześniej zdefiniowałyśmy. Poniżej przedstawione zostały wykresy wraz z krótkim ich omówieniem oraz płynącymi z nich wnioskami.

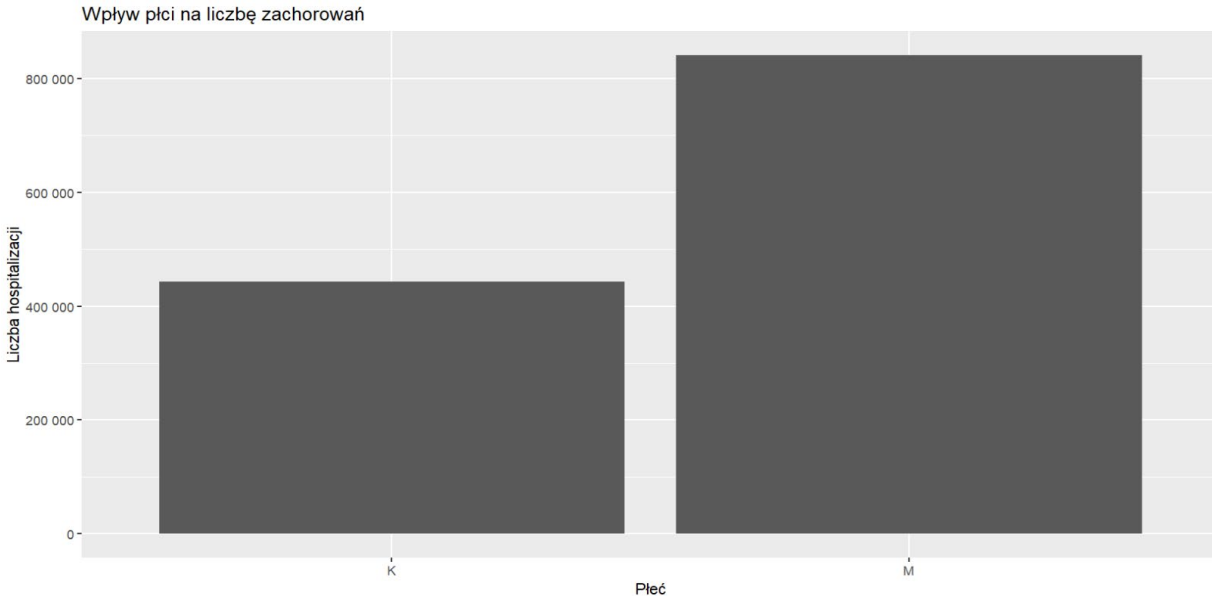
Warto dodać, że wykresy nr 1. oraz 2. zostały przygotowane z danych niezrealizowanych w hurtowni. Dane te zostały odrzucone praktycznie na końcu projektu w związku z niespójnościami między danymi już zaakceptowanymi do hurtowni. Rezultaty jednak uważamy za ciekawe, więc zdecydowałyśmy dodatkowo je załączyć, gdyż jednocześnie zostały wcześniej odpowiednio przetworzone z dostępnych plików.

1. Zależność wieku od liczby osób cierpiących na zaburzenia psychiczne.



Jak można zauważyć, największa liczba osób hospitalizowanych jest w wieku dojrzałym, a więc około 30-40 lat. Warto jednak dodać, że niekoniecznie wiąże się to z taką samą liczbą osób cierpiących na zaburzenia psychiczne. Należy wziąć pod uwagę wystąpienie zjawiska zakłamania danych, które spowodowane przez niezgłaszanie się osób chorych do specjalistycznych jednostek, a decydowanie się na samodzielną poprawę swojego stanu zdrowia.

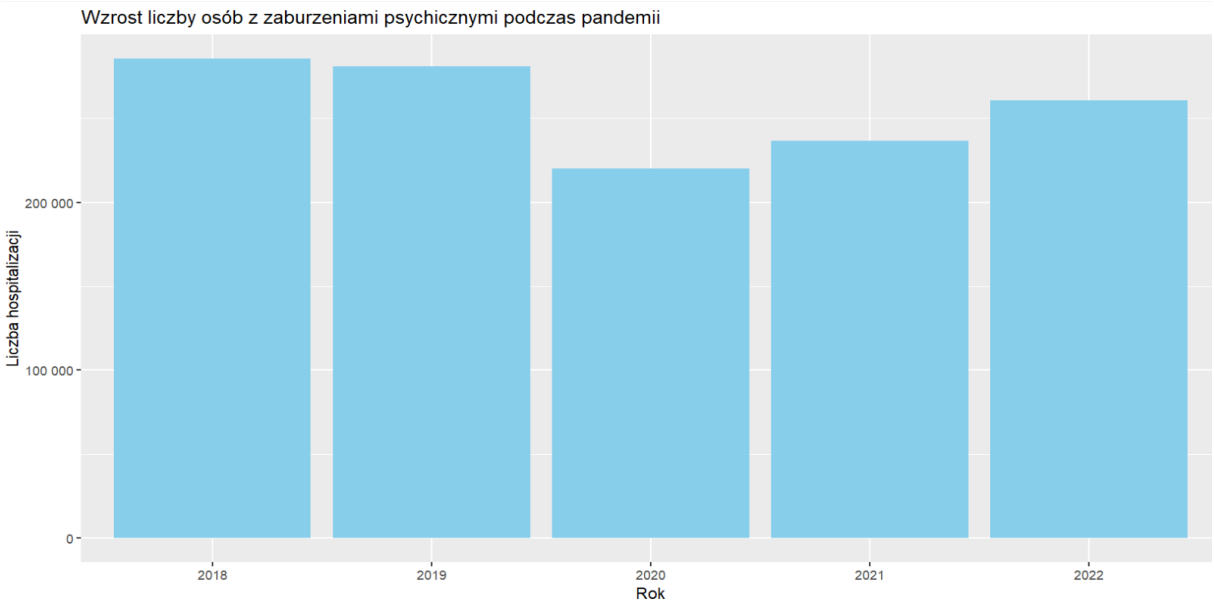
2. Wpływ płci na liczbę zachorowań.



Z powyższego wykresu jednoznacznie wynika, że wśród zdiagnozowanych przypadków, w zestawieniu podjętych hospitalizacjach przeważają mężczyźni i to prawie dwukrotnie względem kobiet. Może być to spowodowane mniejszym zamiarem dzielenia się swoimi problemami z innymi osobami ze strony mężczyzn. Możliwym skutkiem jest sytuacja, w której nie będą sobie w stanie poradzić samemu z emocjami. Jednak z drugiej strony, może

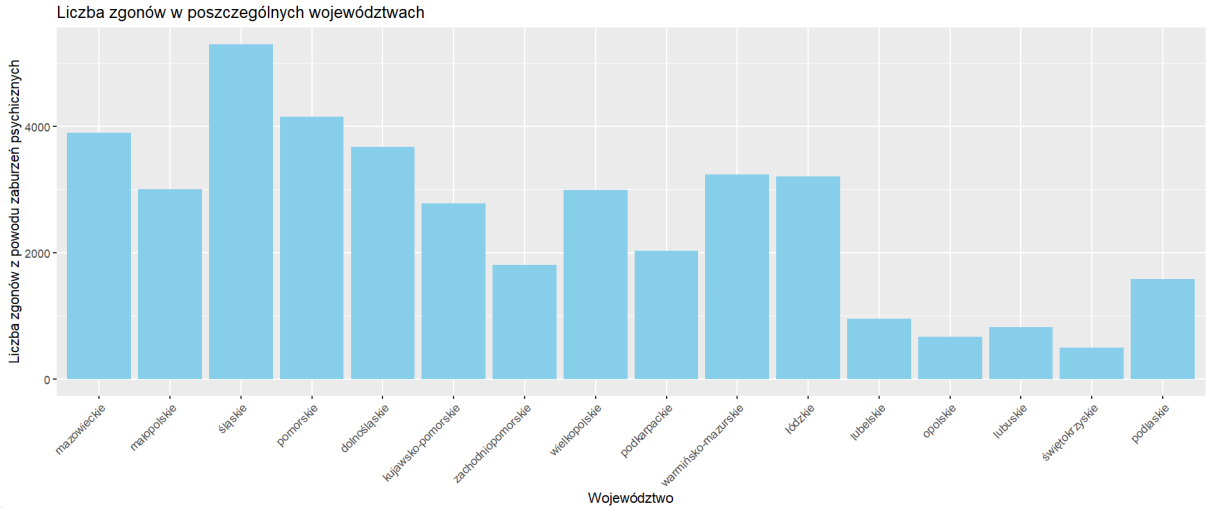
pozwolić to na szybsze dostrzeżenie problemu i natychmiastowe zgłoszenie się do specjalisty w celu podjęcia leczenia.

3. Wzrost liczby osób z zaburzeniami psychicznymi podczas pandemii.



Na powyższym wykresie można zauważyć brak zależności pomiędzy pandemią w roku 2020 a późniejszym wzrostem liczby hospitalizowanych. Z wykresu jasno wynika, że więcej osób z zaburzeniami psychicznymi leczyło się w okresie przed pandemią COVID-19. Warto dodać, że nie jest to jednak równoznaczny wniosek z liczbą chorych, a jedynie z liczbą osób hospitalizowanych. Osoby chore, nie w stanie bezpośredniego zagrożenia życia, mogły zgłaszać się i zostać zarejestrowane rzadziej ze względu na panujące w tym okresie obostrzenia.

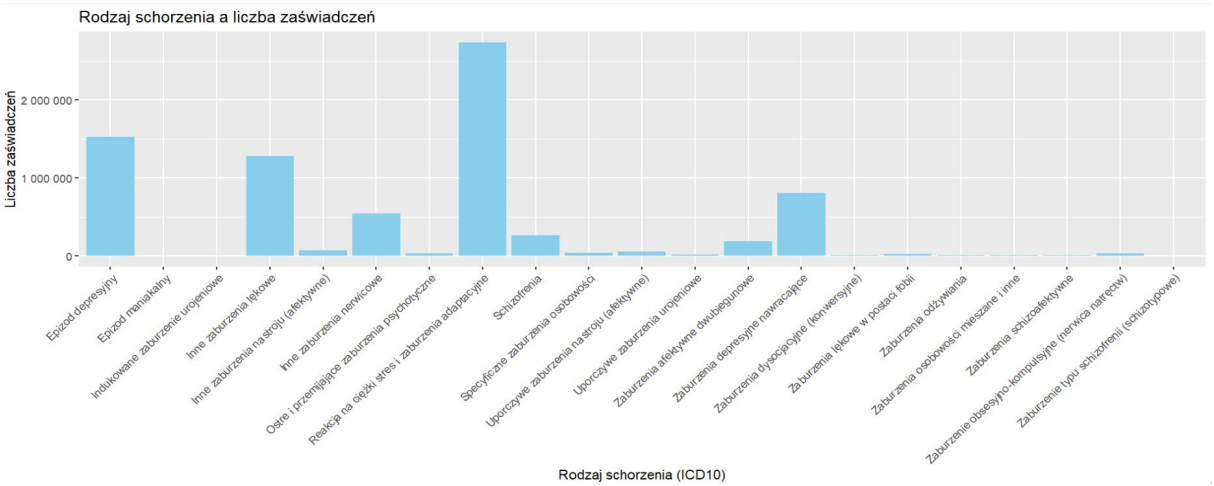
4. Liczba zgonów w poszczególnych województwach.



Wcześniej założyłyśmy, że występuje zależność pomiędzy liczbą zgonów a miejscem zamieszkania. Z powyższego wykresu można wnioskować, że do województw o największej

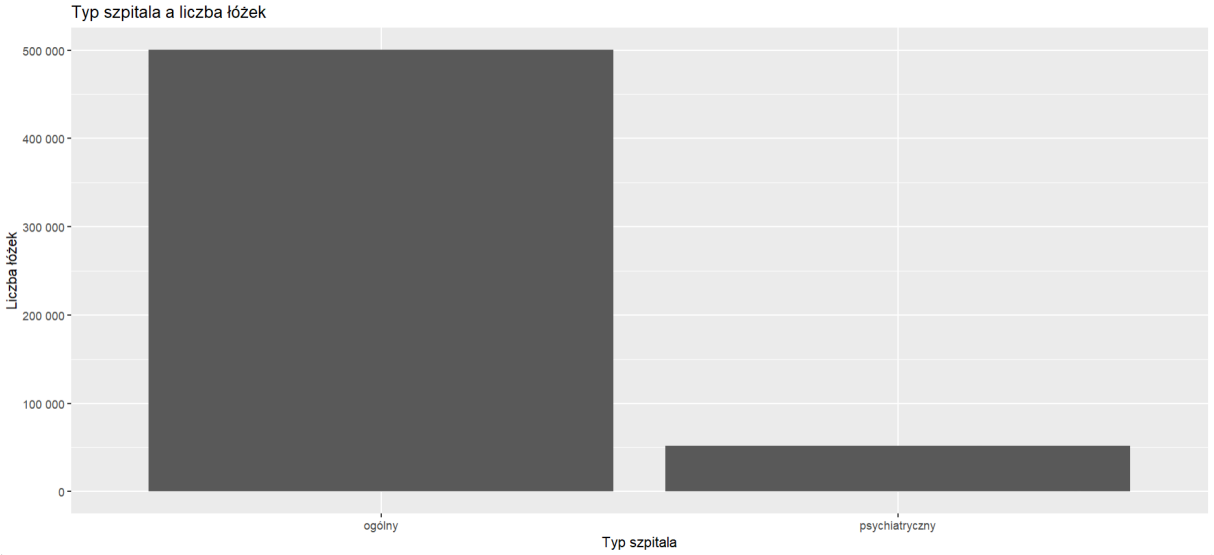
liczebności zgonów można zaliczyć: woj. śląskie, woj. pomorskie oraz woj. dolnośląskie. Zatem niekoniecznie jest to powiązane z większymi miastami, a być może z charakterystyką społeczności zamieszkującej te tereny.

5. Rodzaj schorzenia a liczba zaświadczeń



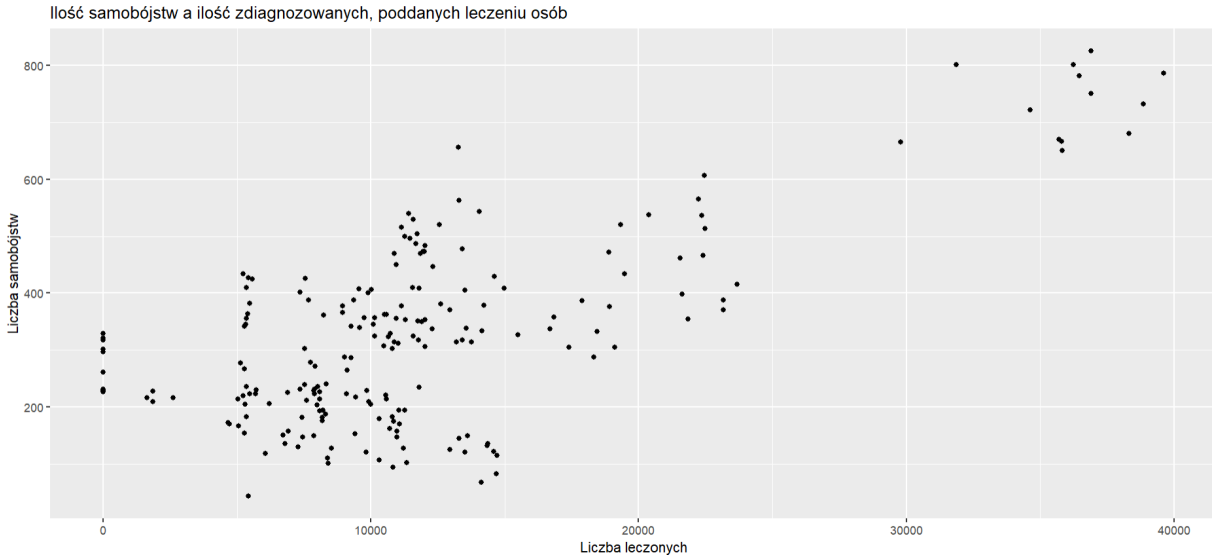
Powyższy wykres pokazuje, że najwięcej zaświadczeń wydawanych jest w kontekście ciężkiego stresu związanego z zaburzeniami adaptacyjnymi, epizodów depresyjnych, zaburzeń lękowych. A więc problemów, które zostają wywołane m.in. przez negatywne doświadczenia u osób chorych.

6. Typ szpitala a liczba łóżek.



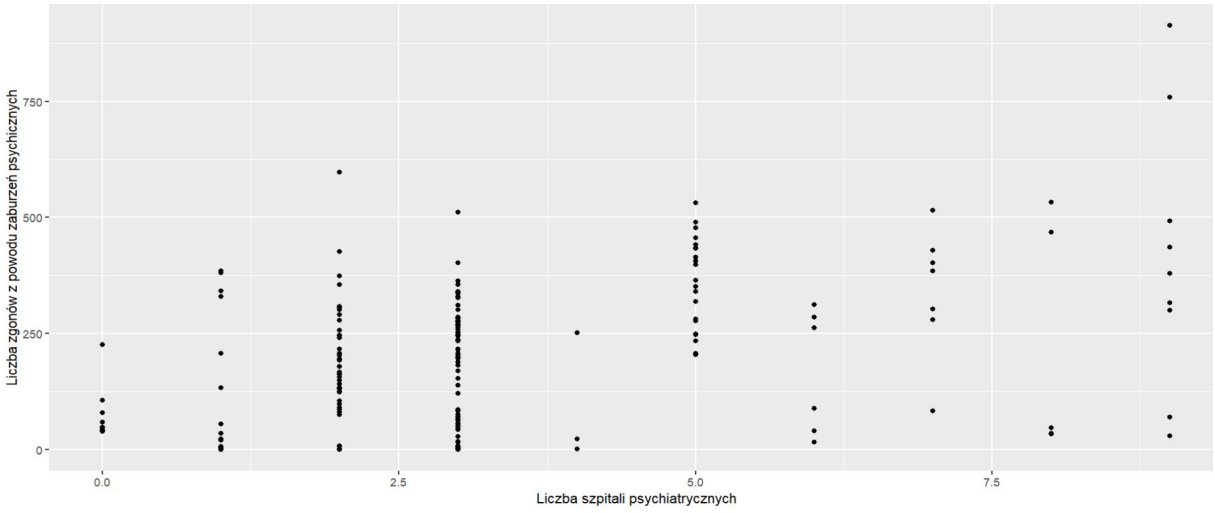
Powyższy wykres ma jednoznaczny wydźwięk – szpitale ogólne charakteryzuje większa liczba dostępnych łóżek. Ma to zapewne związek z tym, że są to placówki dużo większe niż szpitale ukierunkowane na leczenie chorób o jednym podłożu. Zapewne, gdyby wydzielić liczbę łóżek w szpitalu ogólnym na oddziale psychiatrycznym, wartości byłyby bardziej zbliżone.

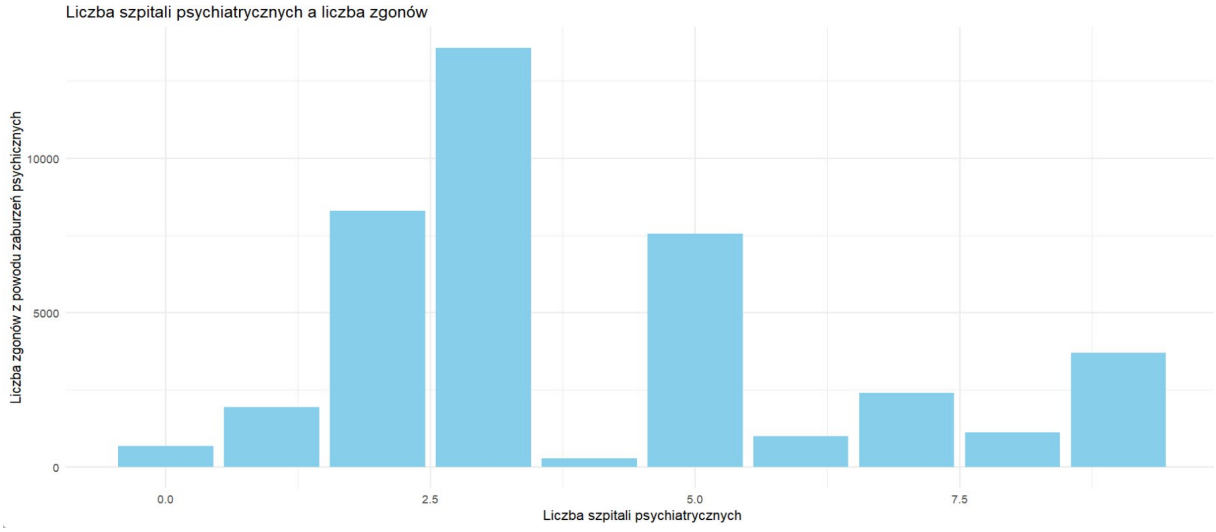
7. Liczba samobójstw a liczba zdiagnozowanych, poddanych leczeniu osób.



Analiza danych wskazuje, że największa liczba samobójstw względem całkowitej liczby osób leczonych występuje w przypadku +-1000 osób chorych- niejako połowa z nich odbiera sobie życie.

8. Liczba zgonów z powodu zaburzeń psychicznych a liczba szpitali psychiatrycznych.





Jak można zauważyć, największa liczba zgonów jest zlokalizowana w obrębie szpitali psychiatrycznych, gdzie ich liczba waha się w granicach 2000-2500. Zatem nie występuje tutaj zależność, która mówi o tym, że im więcej szpitali psychiatrycznych, tym większa liczba zgonów z powodu zaburzeń psychicznych.

Wnioski

Stworzona hurtownia danych oraz przeprowadzone analizy pozwoliły nam jednoznacznie stwierdzić, że:

- największa liczba osób hospitalizowanych mieści się w przedziale wiekowym 30-40 lat;
- mężczyźni są hospitalizowani znacznie częściej niż kobiety;
- pandemia COVID-19 nie wpłynęła znacząco na wzrost osób z zaburzeniami psychicznymi, w porównaniu do sytuacji przed tym okresem;
- wśród regionów o największym odsetku osób leczonych na zaburzenia psychiczne znalazły się: woj. śląskie, woj. dolnośląskie oraz woj. pomorskie;
- najwięcej zaświadczeń lekarskich wydawano w przypadkach ciężkiego stresu, epizodów depresyjnych oraz zaburzeń lękowych;
- szpitale ogólne mają więcej łóżek w porównaniu do szpitali psychiatrycznych, co wynika z ich większej skali i zakresu działania;
- najwyższa liczba samobójstw przypada na grupę około 1000 osób chorych, co sugeruje, że blisko połowa z nich może odbierać sobie życie;
- nie występuje zależność między liczbą szpitali psychiatrycznych a liczbą zgonów osób cierpiących na dolegliwości o podłożu psychicznym.

Przejście przez kolejne etapy implementacji hurtowni – od koncepcji po implementację – okazały się wielopoziomowym oraz długotrwałym przedsięwzięciem, wymagającym skrupulatności oraz dogłębnej analizy problemu, oraz dostępnych, a także pozyskanych danych.

Rezultatem projektu jest hurtownia danych o zdrowiu psychicznym w Polsce, dzięki której możemy odpowiedzieć na wiele interesujących pytań, w zależności od roku i województwa. Z pewnością hurtownię można poszerzać o więcej interesujących danych, jednak ze względu na ograniczenia czasowe zdecydowałyśmy się nie zwiększać zakresu projektu.