Wykład 6. Przygotowanie raportu

Michał Sochański

Skąd wziąć dane?

- www.kaggle.com
- www.dataquest.io/blog/free-datasets-for-projects/
- www.freecodecamp.org/news/https-medium-freecodecamp-org-best-free-open-data-sources-anyone-can-use-a65b514b0f2d/

- Najlepiej znaleźć zbiory o formacie tekstowym (txt, csv) bądź Excelowym (aby następnie wczytać je w Pythonie).

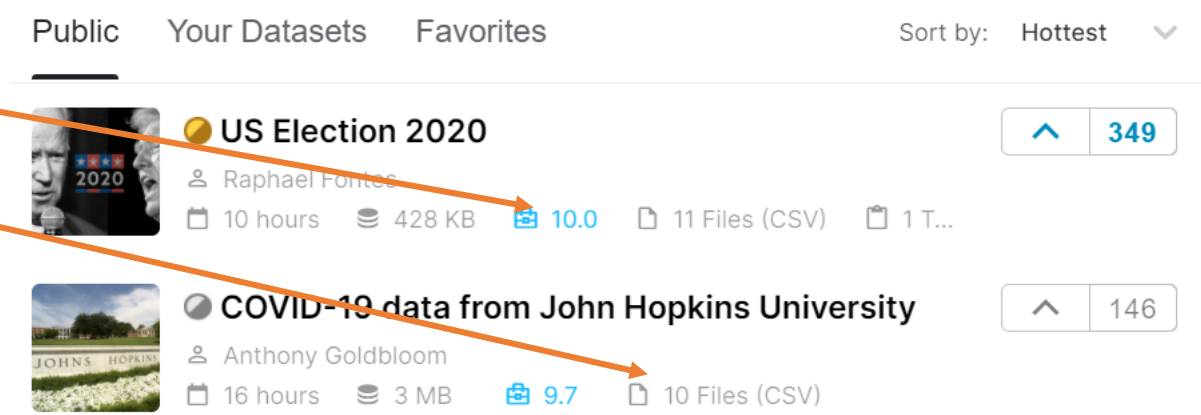
Zbiór z www.kaggle.com

https://www.kaggle.com/docs/datasets - ogólne informacje o sposobie korzystania ze zbiorów na stronie kaggle

https://www.kaggle.com/datasets - lista zbiorów

Przy wyborze zbioru warto zwrócić uwagę na:

- Parametr usability
- Liczbę plików i format



Po kliknięciu na dany zbiór warto zajrzeć do zakładek (u góry strony):

- **Data** – ogólne informacje o zbiorze (tam powinna się znajdować informacja o źródle danych, o zmiennych itd.).
- **Tasks** – zawiera sugestie co do potencjalnie interesujących aspektów zbioru, które warto zbadać (uwaga – niektóre sugestie mogą się wiązać ze stosowaniem metod uczenia maszynowego, którego nie przerabialiśmy).
- **Discussion** – tu pojawiają się często pytania co do zbioru (np. opisu zmiennych itd.), czasem z odpowiedziami.
- **Metadata** – istotne informacje dodatkowe o zbiorze

Jakie warunki powinny spełniać dane?

1. Przynajmniej 50 obserwacji.
2. Przynajmniej trzy zmienne numeryczne i przynajmniej dwie zmienne kategorialne (nominalne lub porządkowe).
3. Łącznie w danych powinno być minimum 300-400 pól.
4. Dane powinny być dobrze opisane, tzn. powinniśmy wiedzieć co oznaczają poszczególne zmienne.

Uwaga - zbiór nie musi być oparty na faktycznych danych.

Podpowiedź – wstępne informacje o zbiorze

1. Jakie mamy zmienne w danych, jakiego typu?
2. Czy zmienne zawierają braki danych? Jeśli tak, to co może być tego powodem?
3. Czy należy odfiltrować pewne dane w dalszych analizach? Jeśli tak, dlaczego?
4. Czy dane są w formacie „tidy” (czy każda kolumna w danych odpowiada faktycznie jednej zmiennej, a każdy rząd obserwacji)?
5. Należy dokonać wyboru zmiennych, których będziemy używać w analizach.

Podpowiedź – wstępne analizy

1. Dokonujemy wstępnej eksploracji danych – sprawdzamy strukturę poszczególnych zmiennych – liczbę kategorii, rozkłady, braki danych. Próbujemy wizualizować poszczególne zmienne i związki między nimi.
2. Należy zadać sobie pytanie, które zmienne mogą być powiązane, bądź które powiązania chcemy zbadać. Możemy też postawić wstępną hipotezę co do takich powiązań.
3. W razie potrzeby należy dodać nowe zmienne, będące transformacją zmiennych obecnych w zbiorze.
4. Dobieramy metody statystyczne oraz tabele i wizualizacje do typu zmiennych oraz przeprowadzanych analiz.

Co powinno znaleźć się w raporcie?

1. Strona tytułowa: przynajmniej z informacją o osobach przygotowujących raport i z tytułem pracy.
2. Spis treści (opcjonalnie).
3. Opis źródła danych oraz przedstawienie wszelkich dodatkowych informacji o danych (np. tych zawartych w plikach czy podstronach „metadata”). Z jakiej strony internetowej pobraliśmy dane? Z jakiego roku pochodzą? Czy są to dane typu open-source? Itd.
4. Skrótowy opis zawartości danych (wyjaśnienie znaczenia zmiennych, które będą poddane analizie, ew. również wymienienie zmiennych, które zdecydowaliśmy się pominąć).
5. Skrótowy opis przeprowadzonego procesu czyszczenia danych – o ile był konieczny (czy występowały braki danych, błędy bądź wartości odstające?).

Co powinno znaleźć się w raporcie?

6. Opis celu i kierunku przeprowadzonych analiz. Jakie problemy będziemy badać oraz na jakie pytania będziemy chcieli odpowiedzieć (albo spróbować odpowiedzieć) poprzez analizę danych? Jakie zależności chcemy ogólnie opisać w pracy? Pomiędzy którymi zmiennymi?
7. Przedstawiając cel analiz można przygotować kilka pytań, na które będziemy starać się odpowiedzieć w raporcie, np.: „jaki kierunek studiów jest najczęściej wybierany przez astronautów?”, „czy płeć różnicuje liczbę podróży kosmicznych?”, „jak zarobki są powiązane z wiekiem?”. Ogólnie – próbujemy odpowiedzieć na pytanie, jakie są charakterystyki danej zmiennej oraz powiązania pomiędzy nią i innymi zmiennymi.

Co powinno znaleźć się w raporcie? Zasadnicza część pracy.

8. Charakterystyka rozkładu wybranych pojedynczych zmiennych za pomocą tabel bądź odpowiednich wykresów.
9. Charakterystyka związków pomiędzy wybranymi zmiennymi za pomocą wykresów, tabel jak również współczynników korelacji czy testów statystycznych.
10. Wnioski z przeprowadzonych analiz. Jakie wyniki uzyskano? Jak można je interpretować? Jak można odpowiedzieć na zadane na początku pytania?

Co powinno znaleźć się w raporcie? Dodatkowe uwagi

8. W raporcie powinny znaleźć się **minimum cztery** tabele oraz minimum **cztery typy wykresów** (co oznacza, że powinny się pojawić przynejmniej cztery wykresy). Co najmniej dwa wykresy powinny przedstawić porównanie dwóch (lub więcej) zmiennych.
9. Do raportu dołączone powinny być:
   - Kod źródłowy Python (służący do wczytania i (ew.) czyszczenia danych jak również do analiz oraz wykresów – nie musi zawierać etapu eksploracji danych)
   - Plik z danymi albo link do pliku

Wymogi formalne odnośnie raportu

1. Przygotowujemy raport w TeXu bądź MS Office.
2. Rozmiar czcionki 12.
3. Czcionka Times New Roman.
4. Interlinia 1,15 do 1,5.
5. Tekst wyjustowany.
6. Cytowania wg standardu APA.

TeX – klasa „kmgr”

- Dla raportów przygotowanych w TeXu dysponujemy gotową klasą ’kmgr’, która została przygotowana specjalnie na potrzeby pisania prac dyplomowych na kognitywistyce.
- Klasa ’kmgr’:

http://uamkogni.home.amu.edu.pl/tex/kogni-dyplom.zip

- Opis klasy i zasady przygotowania pracy dyplomowej na kognitywistyce:

https://plupkowski.files.wordpress.com/2020/02/sem-dypl-technikalia-2020.pdf

- Warto również zajrzeć na stronę

https://marcinjukiewicz.wordpress.com/dydaktyka/analiza-i-wizualizacja-danych/

Obok wytycznych dotyczących raportów (pokrywają się w większości z tutaj podanymi) są tam materiały dotyczące np. NumPy i pandas.

Przygotowanie tabel

Główne wytyczne:

1. Mniejsza czcionka podpisu tabeli i jej zawartości 10-11.
2. Podpis wyśrodkowany.
3. Tabela wyśrodkowana.
4. Tabela powinna mieć swój numer (albo liczony od początku pracy albo dla każdego rozdziału osobno), do którego powinno się odwoływać w tekście.

Przygotowanie tabel

Główne wytyczne:

5. Jednostka danych powinna być umieszczona wraz z opisem kolumny (lub wiersza) w kwadratowym nawiasie.

6. Dane tego samego typu powinny być przedstawione w spójny sposób (jeśli chodzi np. o liczbę miejsc po przecinku).

7. Numer tabeli i opis jej zawartości powinien znajdować się **nad** tabelą.

Uwaga – w przypadku pracy w TeXu, tabele można wykonać np. przez standardowe funkcje dostępne w TeXu. Jeśli wykonujemy raport w Wordzie, można stworzyć tabelkę w Wordzie (wpisując odpowiednie wartości), albo przygotować ją w Excelu i następnie odpowiednio przenieść do Worda.

Przykład poprawnie wykonanej tabeli

Tabela 5.1: Zakresy wartości światłości związanych z określoną barwą wykorzystanych diod LED [4]

Barwa emitowanego światła	Zakres światłości [mcd]
Czerwona	350-600
Zielona	450-1000
Niebieska	250-500

Przykład poprawnie i niepoprawnie wykonanej tabeli (zaokrąglenia)

Poprawna i niepoprawna tabela:

Bodziec 1 [mV]	Bodziec 2 [mV]	Bodziec 3 [mV]
3,58	3,30	4,10
1,02	1,51	1,08
0,84	0,85	0,78

Bodziec 1 [mV]	Bodziec 2 [mV]	Bodziec 3 [mV]
3,57721	3,3	4,1
1,019998234	1,511132	1,0848327194764
0,83956	0,85	0,775

Przygotowanie wykresu lub rysunku

Główne wytyczne:

1. Typ wykresu powinien być odpowiednio dobrany do danych (w szczególności – ilościowych bądź jakościowych).
2. Czcionka podpisu 10-11.
3. Jeśli wykres zawiera tekst, powinien być on czytelny.
4. Podpis i wykres powinny być wyśrodkowane.
5. Wykres powinien mieć swój numer (albo liczony od początku pracy albo dla każdego rozdziału osobno), do którego powinno się odwoływać w tekście. Numeracja powinna być wspólna dla rysunków i wykresów.

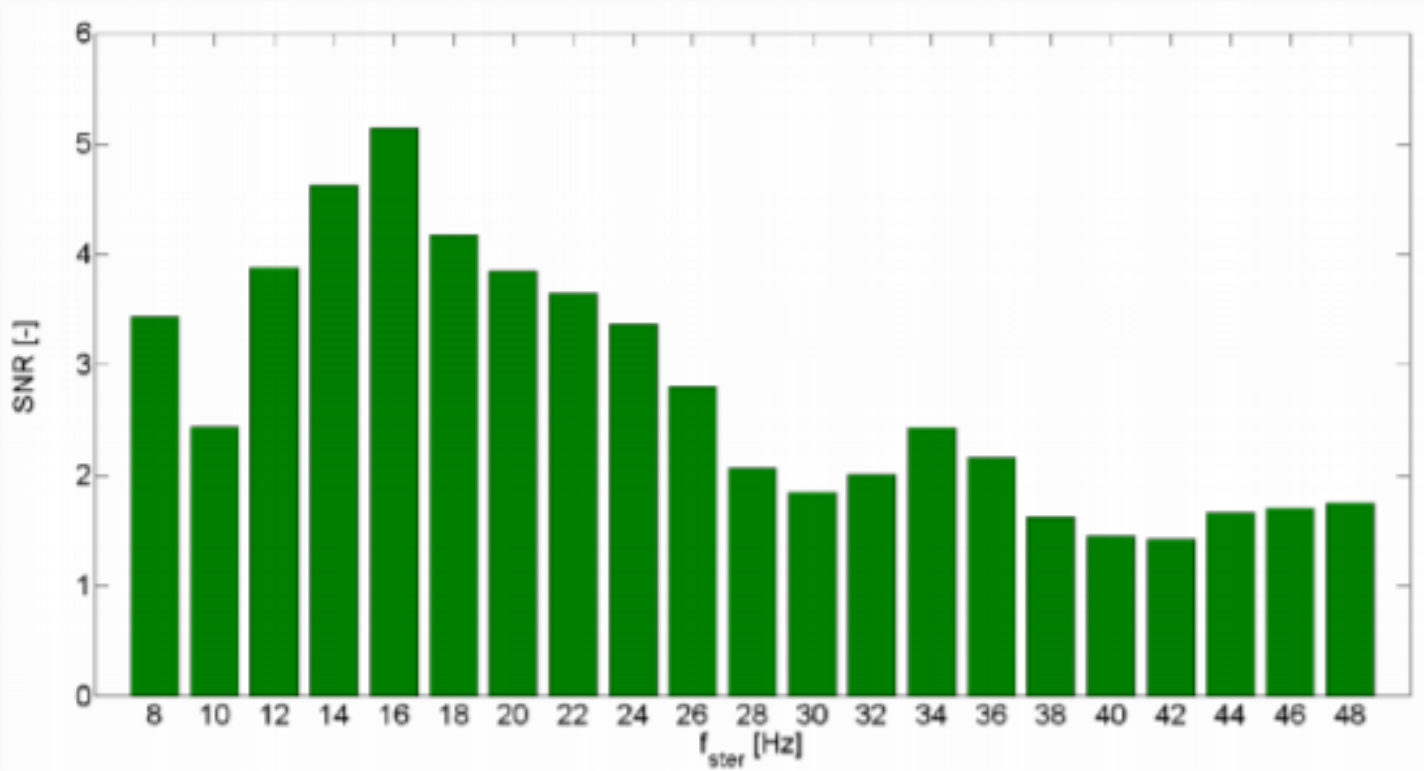
Przygotowanie wykresu lub rysunku

Główne wytyczne:

6. Numer wykresu i opis jego zawartości powinien znajdować się **pod** wykresem.
7. **Nie załączamy** (górnych) tytułów wykresów, z wyjątkiem sytuacji, w której na jednym rysunku mamy kilka wykresów – wtedy każdy z nich może być dodatkowo opisany za pomocą tytułu.
8. Osie powinny być opisane (tytuł + jednostka w nawiasie kwadratowym).
9. Legenda nie zasłania wykresów.
10. Unikanie błędów, o których mówiliśmy na wykładzie dotyczącym wizualizacji.

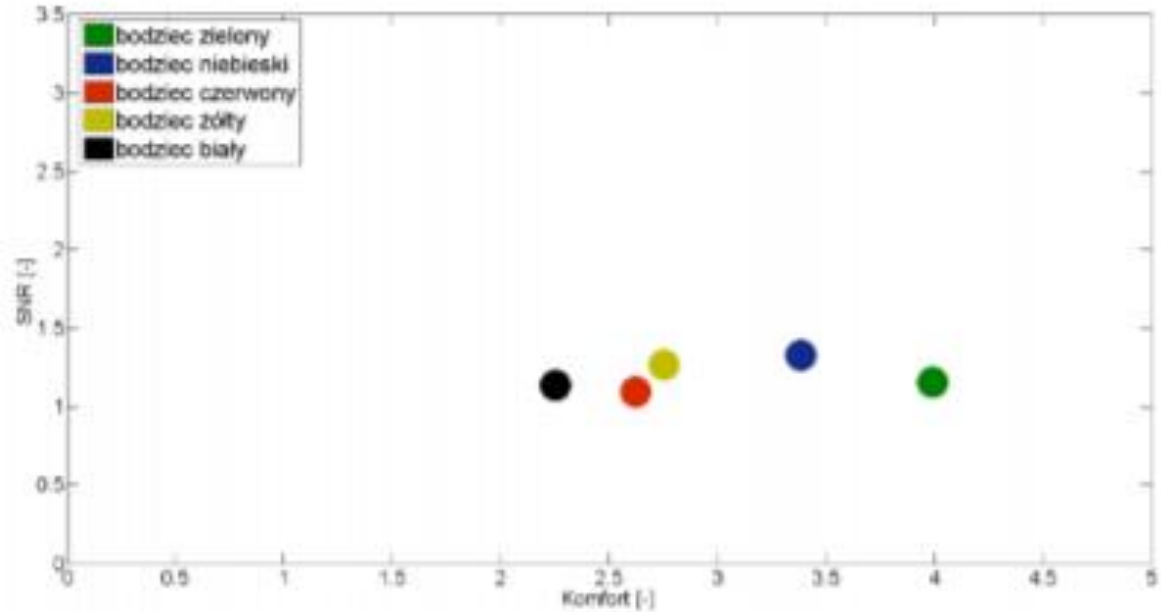
Uwaga – każda wizulizacja (jak i tabelka) powinna być w jakiś sposób skomentowana i zinterpretowana w tekście raportu.

Przykład poprawnie przygotowanego wykresu:



Rysunek 5.10: Wartości współczynnika SNR uzyskane w zależności od f_{ster} bodźca zielonego [53]

Przykład poprawnie przygotowanego wykresu:



Rysunek 5.8: Zależność między współczynnikiem SNR a oceną komfortu barwy bodźca i f_{ster} [53]

Zaliczanie raportu

W trakcie zaliczenia mogę Państwa prosić o:

1. Przedstawienie źródeł i struktury danych – w tym opis zmiennych zawartych w zbiorze i tych, które zostały wybrane do analiz (z uzasadnieniem).

2. Opis procesu czyszczenia danych (jeśli był konieczny).

3. Omówienie sposobu wykonywania analiz (na etapie eksploracji danych).

4. Omówienie kodu (Python).

Zaliczanie raportu

5. Opisanie sposobu sporządzenia wykresów.
6. Opisanie sposobu sporządzenia tabel.
7. Omówienie wniosków (ten element pojawi się na pewno!).

Uwaga – w trakcie zaliczenia każdy z członków zespołu powinien omówić część aspektów pracy/ odpowiedzieć na część pytań. Zakładam, że praca nad raportem będzie podzielona pomiędzy członków zespołu, stąd każda osoba będzie mogła w szczególności omówić ten aspekt pracy, za który była odpowiedzialna.

Jak będzie oceniany raport?

Oceniane elementy raportu wraz z **udziałem procentowym w finalnej ocenie raportu**:

1. Część wprowadzająca (opis zmiennych, danych itd. – patrz slajdy 8-9). 15%
2. Styl i poprawność językowa, oraz przejrzystość pracy (w tym wymogi dotyczące czcionki itd.). 15%
3. Techniczna strona wykonania wizualizacji i tabelek: czytelność, zgodność z wytycznymi (patrz slajdy 14-21); odpowiednia liczba wizualizacji i tabel (patrz slajd 11). 15%
4. Analizy rozkładu pojedynczych zmiennych (od strony merytorycznej): adekwatny dobór metod (wizualizacji, statystyk podsumowujących itd.) do charakteru zmiennych, poprawna interpretacja wyników. 15%
5. Analizy powiązań pomiędzy zmiennymi (od strony merytorycznej): adekwatny dobór metod (wizualizacji, testów itd.) do charakteru zmiennych, oraz ich poprawne zastosowanie; poprawna interpretacja wyników; jakość i trafność analiz, jak również ich odniesienie do pytań i problemów zarysowanych we wstępie pracy. 30%
6. Kod programu (Python). Patrz slajd 11. 10%

Ostateczna ocena wyznaczana jest następująco:

- poniżej 60% punktów – **ndst**
- od 60%, poniżej 67% – **dst**
- od 68%, poniżej 74% – **dst+**
- od 75%, poniżej 82% – **db**
- od 83%, poniżej 90% – **db+**
- od 91% punktów – **bdb**