VERSLAG PROJECT DATAVISUALISATIE

ANNETTE DE MEESTER, EMILIE DE WITTE EN ELINE VANDERPIJPEN

DATASET

Als mogelijke datasets, hadden we de volgende keuzes opgegeven:

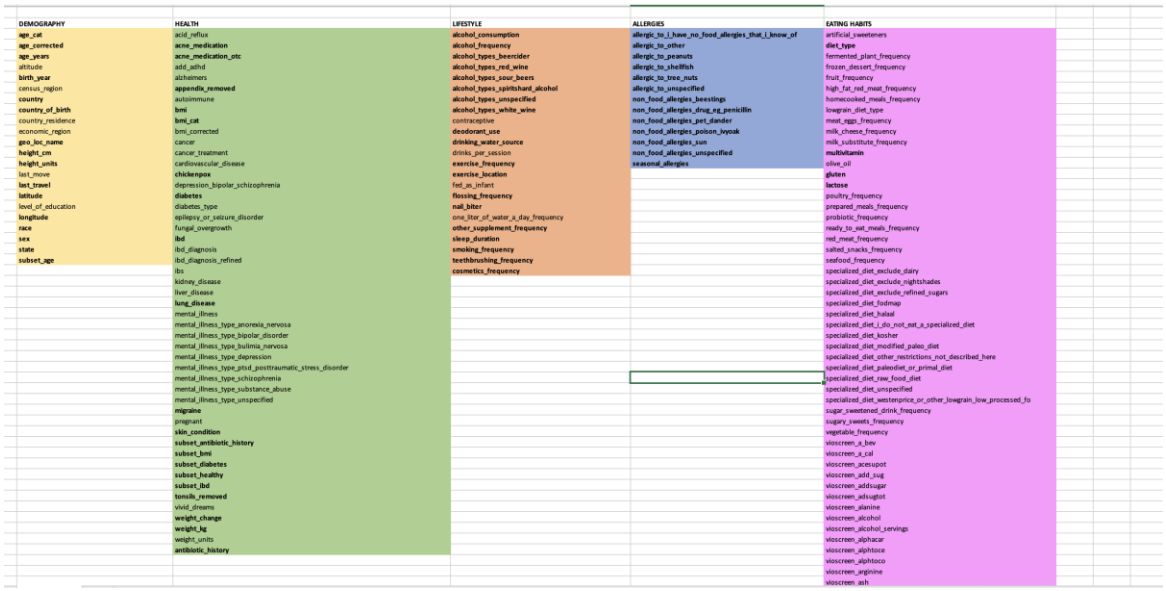
- Infrabel: stiptheid van treinen
- American gut project
- European social survey over klimaatverandering

Aangezien de datasets van Infrabel en de European Social Surveys in de voorbije jaren al door meerdere groepjes bestudeerd zijn, kozen wij ervoor om ons te verdiepen in de dataset van het **American Gut Project**. Deze dataset werd opgesteld met als doel het microbioom van diverse populaties in kaart te brengen en de verbanden tussen levensstijl, gezondheid en dieetgewoonten enerzijds, en het microbioom anderzijds, te onderzoeken.

Het betreft een publiek toegankelijke dataset. Niet alleen is de data vrij beschikbaar, maar iedereen die dat wil, kan zich ook registreren om deel te nemen aan de studie. De specifieke dataset die wij hebben geanalyseerd, is te vinden via de volgende link:

https://figshare.com/articles/dataset/Full_American_Gut_Project_mapping_file/6137315?file=11070449

Ze bevat gegevens over de gezondheid, levensstijl en dieetgewoonten van 17.855 personen uit 49 landen.



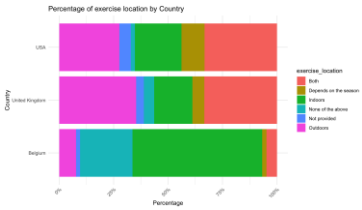
Hierbij hebben we alle variabelen waarbij minstens 70% ingevuld was aangeduid in het vet.

EERSTE IDEEËN

Als eerste idee dachten we om een observationele analyse te doen van de volledige dataset waarbij we ons zouden focussen op de volgende vragen:

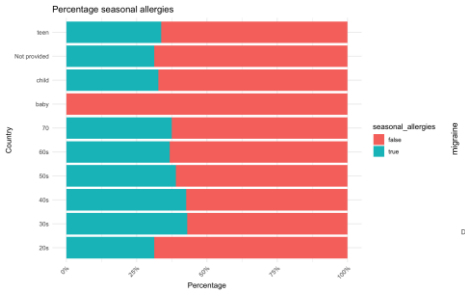
- Wat zijn de verschillen in levensstijl tussen landen/geslacht/leeftijd?
- Wat voor mensen hebben een bepaalde ziekte? We kijken hiervoor naar gewicht, leeftijd, onderwijsniveau en woonplaats om reverse causality tussen levensstijl en ziekte te vermijden.

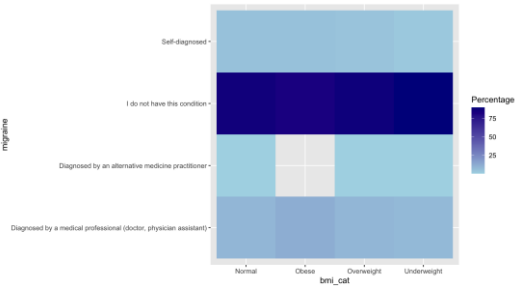
Voor de verschillen in levensstijl dachten we om per categorie een wereldkaart te tonen waarin landen gekleurd zouden zijn volgens het percentage van de bevolking met deze levensstijl of bij numerieke variabelen volgens de gemiddelde waarde voor deze levensstijl. Voor categorische variabelen, zouden we ook een stacked bar chart toevoegen om de verdeling over de categorieën in de verschillende landen te bestuderen. Bij het opstellen van de stacked bar charts zagen we dat, door de hoeveelheid van categorieën, deze voorstelling misschien niet ideaal was en hebben we in ons uiteindelijk project gekozen voor andere visualisaties.

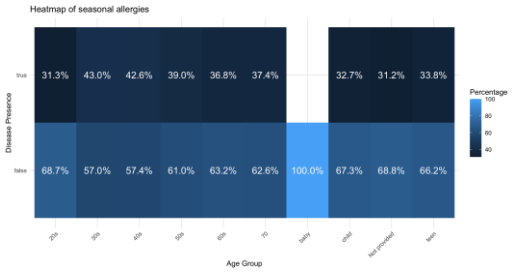


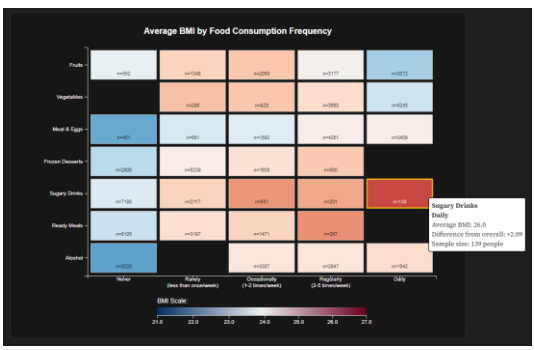
Echter, omdat we in de dataset zagen dat er toch wel wat landen waren met maar enkele metingen en dit misschien een verkeerd beeld zou geven, hebben we ervoor gekozen om ons enkel te focussen op de Verenigde staten, aangezien de meeste samples uit de dataset uit de Verenigde staten komen. In plaats van een wereldkaart te kleuren, hebben we in onze uiteindelijke visualisatie de staten van Amerika gekleurd.

Om de ziektes en allergieën over de populaties te analyseren zouden we met stacked bar charts en heatmaps werken om de verschillen in distributie tussen geslachten, leeftijdsgroepen, educatieniveaus en BMI categorieën te bekijken.









Omdat we toch veel variabelen in onze dataset hebben en het wat ongestructureerd of onoverzichtelijk zou kunnen overkomen om deze allemaal in onze visualisatie op te nemen, beslisten we om onze focus te beperken tot de consumptie van alcohol, zie volgende sectie.

UITEINDELIJKE PLAN

Als uiteindelijke visualisatie kozen we ervoor alcoholconsumptie binnen de Verenigde staten te bestuderen. We focusten hierbij op de volgende drie thema's:

- **Demografie**: Hoe verschilt de consumptiehoeveelheid tussen staten, geslacht en verschillende leeftijdsgroepen?
- **Gezondheid**: Hoe verschilt de gezondheid tussen personen met verschillende consumptiehoeveelheden?
- **Levensstijl**: Hoe verschilt de levensstijl tussen personen met verschillende consumptiehoeveelheden?

Omdat de dataset heel wat missing values bevatte, hebben we onze analyse uitgevoerd op een gefilterde dataset. Deze filtering gebeurde via de volgende stappen:

1) Neem enkel de samples uit de Verenigde Staten.
2) Selecteer de variabelen uit de dataset die voor deze geselecteerde samples minstens voor 70% ingevuld waren.
3) Selecteer uit deze nieuwe dataset enkel de volledige samples.

De uiteindelijke visualisatie is te vinden via de volgende link:

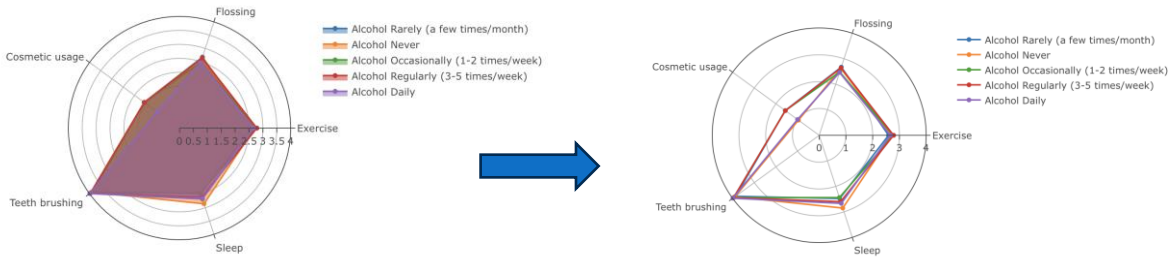
https://datavisualiatie-ugent.github.io/project-dv25-5/

MOEILIJKHEDEN IN UITEINDELIJKE VISUALISATIE

Omdat we elk een thema binnen de visualisatie hadden uitgewerkt, miste er wat eenheid in de uiteindelijke visualisatie. Zo werden zelfde categorieën over de visualisaties heen door verschillende kleuren aangeduid en werd er geen eenzelfde taal gebruikt in de categorieën/titels/teksten. Onze dataset stond in het Engels, waardoor we gekozen hebben om onze hele visualisatie in het Engels op te stellen. Dit omdat we anders alle categorieën in de dataset zouden moeten vertalen en op die manier ook aanpassen in de dataset, wat ietwat omslachtig zou zijn.

Levensstijl:

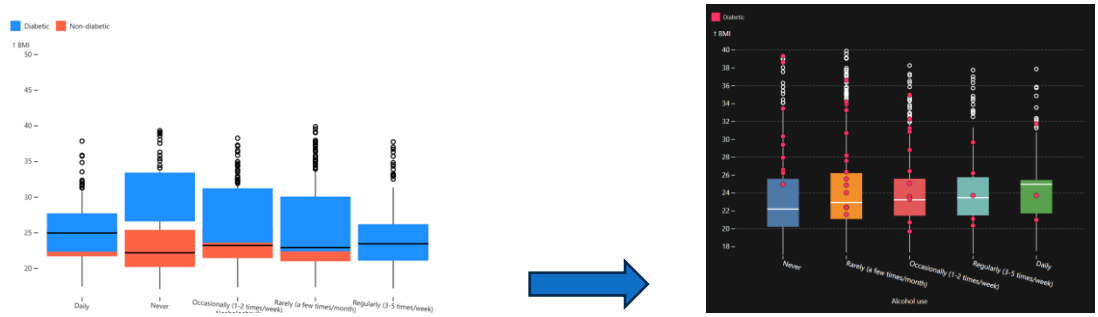
Voor de radar chart hadden we oorspronkelijk de verschillende categorieën ingekleurd, zoals gebruikelijk is bij dit type visualisatie. Toen we dit in onze eigen weergave toepasten, merkten we echter dat de leesbaarheid sterk afnam bij meer dan twee categorieën. Daarom hebben we ervoor gekozen om de chart niet langer in te kleuren, zodat het overzicht behouden blijft. Daarnaast toonde de as aanvankelijk kommagetallen, wat het aflezen bemoeilijkte. Om dit te verhelpen, hebben we de as vereenvoudigd en enkel de waarden 0, 1, 2, 3 en 4 weergegeven.



Gezondheid:

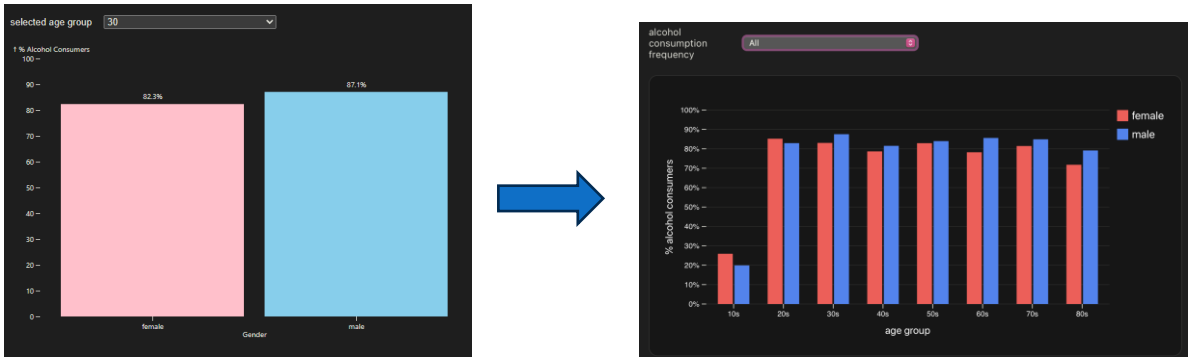
Het was initieel de bedoeling om met bar charts de correlatie tussen alcoholconsumptie en verschillende gezondheidsfactoren te modelleren, zoals IBD, migraines, stress (die zou gemodelleerd worden dmv 'nail biter') etc. Maar omdat er weinig samples waren die positief waren voor de meeste aandoeningen bleven we enkel met 'migraine' over, waardoor het interactieve aspect van de grafiek wegviel en er een simpele barchart overbleef.

Om de correlatie met diabetes te visualiseren was het initiële plan om in de boxplot het deel dat diabeet was “in te kleuren” door het te overlappen met de boxplot van enkel diabetici, maar dat leverde moeilijkheden op. Dit kwam omdat er weinig diabeten-samples waren en veel van deze samples waren outliers, wat resulteerde in eigenaardige boxplots. Daardoor hebben we de boxplots laten vallen en enkel de samples geprojecteerd, wat weinig statistische draagkracht heeft omdat er zo weinig zijn, maar wel ruimte open laat voor verdere hypothesestellingen en analyse.



Demografie:

Om het verband tussen alcoholconsumptie en geslacht/leeftijdsgroep te onderzoeken, begonnen we met het gebruik van bar charts waarbij de gebruiker een specifieke leeftijdsgroep kon selecteren. Op die manier konden de percentages alcoholgebruikers binnen die leeftijdsgroep tussen de geslachten vergeleken worden. Deze visualisatie maakte het echter lastig om verschillende leeftijdsgroepen onderling te vergelijken. Daarom zijn we overgestapt op één grouped bar chart per alcoholconsumptiecategorie, met geslacht als categorie, waarop alle leeftijdsgroepen samen worden weergegeven.



LOGBOEK

Taakverdeling:

Opzoeken datasets	Annette, Emilie, Eline (elk eentje)
Eerste visualisatie ideeën maken en presentatie opstellen hierover	Eline
Dataset filteren	Emilie
Visualisaties demografie	Annette
Visualisaties levensstijl	Eline
Visualisaties ziektes	Emilie
Plaatsen van visualisaties op Github (alle visualisaties uit notebooks overzetten naar Observable framework) + deploy via github pages	Annette
Tekst bij visualisaties zetten en hier een geheel van maken	Emilie
Eindpresentatie en verslag	Eline

Tijdsbesteding:

28/02	Opzoeken datasets
03/03	Meeting met prof: keuze dataset
28/03	Meeting om eerste ideeën te bespreken
28/03	Eerste filtering dataset
28/03	Eerste visualisaties maken en presentatie over eerste ideeën opstellen
31/03	Meeting met prof: eerste ideeën voorstellen
17/04	Meeting over volgende aanpak
17/04	Finale filtering dataset
24/04	Eerste versie visualisaties opstellen
25/04	Meeting om visualisaties op elkaar af te stemmen
26/04	Verder werken aan visualisaties
28/04	Meeting met prof: voorstellen van visualisaties
10/05	Afwerken van de visualisaties en opstellen van de teksten tussen visualisaties
11/05	Opstellen van presentatie en werken aan verslag
18/05	Afwerken verslag