Logboek

04-06-2018:

Voor de meeting om 15:40 hebben we eerst samen alle voor en nadelen van iedere database besproken en onze voorkeuren geuit, daarna hebben we ons voorbereid op het eerste gesprek door een paar vragen te formuleren. Deze vragen gingen voornamelijk over de indeling van het vak en de hoeveelheid keuzevrijheid we krijgen bij het kiezen van de deelvragen.
Verder hebben we contactinformatie gedeeld, een drive aangemaakt, een github aangemaakt en besloten woensdagochtend(de 6e) definitief de dataset te kiezen.
Tenslotte hebben we besloten om elke dag om 11:00 op Science Park te zijn om samen aan het project te werken. Behalve de volgende dag. Dat was een vrije dag.

05-06-2018:

Rustig aan gedaan, wel heeft Luc nog een gebruikershandleiding voor github gemaakt om het aantal *merge failures* die er nog aan zaten te komen zo veel mogelijk te beperken.

06-06-2018:

Groepsmeeting 12:00-14:00;
Definitieve keuze gemaakt voor de *food prices* dataset. De eerder opgestelde plus en minpunten van deze dataset staan hieronder.

Global Food Prices Database (WFP):

Pro's:

1. Leuk/interessant
2. Veel creatieve mogelijkheden
3. Gigantisch

Cons:

1. Gigantisch
2. For this dataset it is Mandatory to include additional numeric data (think of temperature, amount of rain, currency prices, etc.)

Tenslotte hebben we een aantal mogelijke extra deelvragen bedacht.

Extra vragen

- Weer

Prijzen van producten worden beïnvloed door het weer en daar zijn datasets over te vinden

- Oorlog

Er zijn conflicten, zoals oorlog en politieke gebeurtenissen die prijzen van producten beïnvloeden. Het is niet zeker of daar datasets voor te vinden zijn.

- Wages en brood

Het is interessant om te zien of de salarissen van werkers correlatie vertonen met de prijzen van brood in landen. Er is gekozen voor brood, want we wilden de focus op een bepaald product leggen.

Voor het gesprek met de Willemijn hebben we eerst nog een vragenlijst opgesteld met vragen over alle problemen waar we eventueel tegenaan dachten te gaan lopen.

Gesprek met Willemijn
14:00
Willemijn gaf ons de volgende tips mee:.
-Letten op veranderingen in USD (USD van 1 jaar kiezen)
-Dataset vinden over de conflicten (maar niet teveel tijd aan besteden)
-Dataset verklaren op het allerlaatste
-Tijd is geen probleem meer, want we hebben een slider
-Dataset zo min mogelijk wijzigen (outliers)
-Python script in git

07-06-2018

11:00 op Science Park.

Na de dataset met Pandas te hebben bestudeerd kwamen we de volgende observaties naar boven.
Besluiten/observaties:

-Er zijn geen missing values ontdekt.

-Brood bleek niet overal verkocht te worden, dus is er besloten om naar de meest verkochte producten per land te kijken.

-Er is besloten om verschillende datasets te maken voor verschillende producten

-Producten die hetzelfde zijn, zijn in één groep gezet, vb. 'rice (low quality) → rice' en alle 'wheatsoorten' zijn ook in één groep gezet

-Er is besloten de id's allemaal weg te laten, aangezien deze niet relevant zijn voor de data-analyse

Gewijzigd:
Dataset is opgesplitst in top 10 meest voorkomende producten + brood

Adm0_name → adm_name (Omdat Pandas het anders niet kon verwerken (dit bleek later nutteloos te zijn))

Mp_price and cur_name komen naast elkaar te staan

Te verwijderen:

Adm_id, adm1_id, market_id, cm_id, cur_id, pt_id, um_id, mp_commoditysource, pt_name
Deze dingen zijn weggehaald, omdat ze niet relevant zijn.

Te Wijzigen

- Outliers

- De prijzen van de producten moeten over dezelfde hoeveelheden gaan (normaliseren)
  Pounds omzetten naar kg
  (worden afgerond op 2 decimalen voor de consistentie)

- Gemiddelde prijs berekenen van elke markt per product (National average)

Markten mergen

- Prijzen terugrekenen naar een vaste dollar eenheid

Huidige waarde van de dollar tegen de currencies van elk land zetten (we gaan ervan uit dat dit niet verandert over de jaren heen)

Fouten in de dataset
-Willekeurige punten in dataset
-Currency ‘Somailand Shilling’ staat erin (afkorting is SOS)
Maar ook SOS staat erin

Alle vreemde meeteenheden die sommige landen gebruikten hebben we opgezocht en naar kilogrammen gelinkt, sommige rare eenheden zoals Marmite waren nog wat werk op uit te pluizen. Hieronder een voorbeeld:

Rijst
Kg
Pound → 0.45359237 KG
Cuartilla → 2,875575 kilogramos
90 KG
MT → 1000000000 KG (Miljard kilo)
45 KG
500 G → 0.5 KG
50 KG
Marmite → 2.72155422 KG
100KG
25KG

Bread
Unit → ??
500 G → 0.5 KG
Pound → 0.45359237 KG
400 G → 0.4 KG
150 G → 0.15 KG
Loaf → 1.81436948 KG ??
Bij brood stonden de eenheden Unit en Loaf er regelmatig in, beiden werden na een hoop gespit op internet op 0,6kg ingeschat.

08-06-2018

11:00 op SP.
Vraag voor Willemijn:
Database blijkt veranderd te zijn, moet daar iets mee gebeuren?

-Algoritmes om de databases te cleanen, zijn geschreven.
-Hoeveelheden van producten zijn omgezet
-Library van valuta's en USD is gecreeerd
-National Average's zijn berekend
-Probleem met komma's in database is gefixt
-Klimaat datasets zijn opgezocht

Week 2

11-06-2018

Meeting met Willemijn, we hebben het erover gehad eventueel te BigMac index te gebruiken om te compenseren voor inflatie (niet gedaan want onhandig).
-De speurtocht naar een goede dataset voor het klimaat had nog te weinig opgeleverd dus daar moest wat tempo in.

-Misschien leuk om een gezondheidsdataset te zoeken?

12-06-2018

Data wordt gesorteerd, want het staat in niet allemaal op chronologische volgorde.
Goede (redelijke) dataset gevonden voor het klimaat, helaas per werelddeel ipv per land maar het moet maar.
Deze temperatuurdataset is *gecleaned*.
Daarnaast was er een probleem met de Rice data, want er was een foutieve extra kolom gevormd die eruit gehaald moest worden. Dat is gefixt.

Verder zijn er allerlei verschillende data-analyses gedaan, zoals tabulation, crosstabulation, bar plots, histogrammen, boxplots, density plots en scatter plots.

13-06-2018

Meeting met Willemijn:
-We moeten hypotheses opstellen over de deelvragen en daar intelligentie dingen over zeggen.
-Meer plotjes maken.

Hypotheses in grote lijnen:
Klimaat beinvloedt prijzen, de prijs van ingrediënten correleert met van hun producten producten, regio's hebben vergelijkbare prijzen.
Vervolgens hebben we afgesproken thuis zelf deze hypotheses te verfijnen en morgen de bevindingen te delen.

14-06-2018

Er moeten verschillende voedsels uit verschillende datasets vergeleken worden, maar het probleem is dat deze datasets niet gelijk zijn qua grootte en index. Hierdoor kunnen we de relaties tussen de data niet visualiseren.
Er moet een manier gevonden worden om verschillende producten tegen elkaar uit te zetten, zodat de data, landen en index overeenkomen. Het is namelijk niet zo dat alle de data van alle landen op hetzelfde moment begint. Dit bleek echt 10x lastiger dan gedacht dus daar zijn we bijna de hele dag mee bezig geweest. Met een beetje resultaat.

15-06-2018

We hadden al eerder afgesproken om deze dag vrij te nemen en onze voortgang liet dat toe.

Week 3

18-06-2018

Alvast vragen opgesteld voor het gesprek met Willemijn morgen.
Vragen voor Willemijn:

Hoe verander je de index van een lijngrafiek?
Wat is de beste manier om data met 3 variabelen te plotten; de landen, het tijdstip en de prijs?

Verschillende datasets met elkaar vergelijken, ze hebben niet dezelfde hoeveelheid entries.

Klimaatdata, geen data van alle landen, hoe moet dat vergeleken worden.

Is het mogelijk om alleen naar de data vanaf 2014 te kijken, aangezien we van alle landen data hebben vanaf dat jaar

Welke waarden kunnen we gebruiken voor clustering, aangezien we niet veel getallenwaarden hebben.

Heeft het zin om te clusteren?

19-06-2018

Gesprek met Willemijn:

Clustering: Met temperatuur
Of landen naar getallen omzetten (vertekening van data)
Regio's als cluster

Regressie om nieuwe patronen te ontdekken

20-06-2018

Veel bezig geweest met het begrijpen van de werking en het nut van regressie en clusteren op onze database. Verder doorgewerkt aan grafieken maken en methodes bedacht om het te visualiseren.

-Clustering met de gemiddelde productprijzen per land per jaar?

-Bedacht welke plotten er op de website moeten.

21-06-2018

Vragen voor Willemijn opgesteld voor het gesprek.
Vooral gehad over manieren om de verplichte machine learning technieken toe te passen op onze data.
Daarna een begin gemaakt met het schrijven van het verslag en het klaarmaken van de data voor de website die volgende week gemaakt moet worden.

22-06-2018

Nog meer data plotten
Clustering met one-hot encoding en werelddelen ipv landen. Hiermee plots gemaakt van clusters. Kwam weinig nuttigs van naar voren.
Inleiding en methode geschreven voor verslag.

Week 4

25-06-2018

Luc: Github schoongemaakt, website begonnen, plotly uitgevogeld.
Chandni: Technisch rapport schrijven, logboek netjes maken, plotly uitgevogeld.
Stijn: Technisch rapport schrijven, geholpen met de website.
Emiel: Technisch rapport schrijven, css begonnen.
-Werken aan technical rapport en logboek
-Website opzetten
-Website op GitHub
-Kaart van plotly werkend krijgen
-Technisch rapport schrijven

Vragen voor Willemijn voor gesprek morgen
-Let ze op taal, vertalingen van Engels naar Nederland (EDA, in depth analysis, productnamen)
-Discussie en resultaten verschil?
-Wat voor plaatjes moeten in de docs
-Logboek
-Wanneer is volgende afspraak

Verslag zo ver mogelijk af voor morgen, dus vandaag al resultaten en discussie schrijven.

26-06-2018

Luc: Alghoritme geschreven waarmee plotly plot gemaakt zou worden (helaas moeten we ervoor betalen…..), technisch rapport geschreven
Chandni: Technisch rapport geschreven, plotjes uitgezocht voor de website
Stijn: Technisch rapport geschreven, gewerkt aan bronnen
Emiel: Technisch rapport geschreven, plotjes uitgezocht voor de website
-Plotly wereldkaart is werkend! (maar niet zo als we willen)
Technisch rapport.

Gesprek met willemijn;
-12 uur afspraak Willemijn donderdag
-vanavond/morgenochtend eerste versie verslag af.

Rest van de dag aan verslag gewerkt om het zo af mogelijk te krijgen.
Daarnaast gewerkt aan de website die al een beetje vorm begint te krijgen.

27-06-2018

Luc: De plotly plots zitten maken, en website infrastructuur omgegooid.
Chandni: Expo, plotten op plotly.
Stijn: Logboek netjes gemaakt, plotten op plotly.
Emiel: plotten op plotly en de css werkend gekregen.
-Er is aan de website gewerkt.
-Er is een manier gevonden om de plotly kaarten op de website te krijgen.
-Het is gelukt om de css op de site toe te passen
-Dit logboek is netjes gemaakt

28-06-2018

Luc: Website afgemaakt, Technisch rapport laatste controle.
Chandni: Feedback van technisch rapport toegepast
Stijn: Details aan technisch rapport toegevoegd en aan website gewerkt.
Emiel:Website afgemaakt Technisch rapport laatste controle.
Gesprek met Willemijn
Feedback gekregen op technisch rapport en de website laten zien

De feedback die op het technisch rapport is gegeven wordt verwerkt.
De plotly kaarten zijn per jaar op de website gezet. Er is voor gekozen om de bezoeker van de site een product te laten kiezen en daarna een jaar. Daarvan komt er een kaart tevoorschijn. De gebruiker kan over de landen heen scrollen en zien hoeveel het product kost in US dollar. Ook krijgen landen met een hogere prijs, een donkerdere kleur op de kaart en landen met een lagere prijs een lichtere kleur.

-Het verslag veel verbeterd
-Kaarten gemaakt van alle relevante data en hier scriptjes van gemaakt die op de website kunnen
-Alle plots die we willen laten zien op de site verzameld en op de site gezet
-De website afgemaakt
-Voorbereid op de presentatie
-Alles ingeleverd