Drughunters
Ottiliavej 9
2500 Valby
Tel +45 3630 1311
Fax +45 3630 1940
E-mail drughunters@lundbeck.com
www.drughunters.dk

Drughunters 2024 – Matematikopgave

Drughunters Medicines Corp. er en (fiktiv) medicinalvirksomhed, der forsøger at udvikle nye typer medicin, der er mere effektive end de eksisterende. For at hjælpe med dette er forskere ved Drughunters Medicines Corp interesserede i at lære, hvordan mennesker med samme sygdom kan opleve forskellige problemer.

Nu hvor vi lever i en stadig mere digital verden, bliver det nemmere og nemmere at indsamle og gemme data fra den virkelige verden, for eksempel data indsamlet fra læger på hospitaler. Ud fra disse data har vi en ny måde at forsøge at forstå mennesker, der lider af en bestemt sygdom.

I denne opgave kommer I til at arbejde som forskere hos Drughunters Medicines Corp., og I kommer til at anvende matematik og kunstig intelligens på et datasæt fra den virkelige verden for at få viden om 783 universitetsstuderende med depression og angst (disse data er ikke personhenførbare data, dvs. personerne kan ikke identificeres ud fra datasættet).

Efter selve opgaveformuleringen er der links til steder, hvor I kan få hjælp til at finde ud af, hvordan I udfører analyserne i f.eks. EXCEL. Der er også hjælp til at downloade og gemme filer i de rigtige formater etc.

Hjernesygdommen

1) I eksempel-datasættet (se **Bilag 1** i slutningen af dokumentet) er data fra studerende med depression og angst. Beskriv kort sygdommene i nogle få sætninger (symptomer, behandlinger, prognoser osv.) og beskriv kort PHQ score og GAD score.

Statistisk analyse af data ved hjælp af modellering og simulering

Når vi som forskere taler om, at vi skal analysere data, betyder det, at vi ønsker at få et overblik over data og at kunne drage nogle konklusioner vedrørende sammenhænge i data (hvis der overhovedet er nogle). Før vi begynder de mere avancerede matematiske analyser, er det meget vigtigt at bygge noget basisviden op omkring vores data.

2) Lav tabeller med beskrivende/deskriptiv statistik samt grafiske præsentationer af data (gennemsnit, optællinger, figurer osv.) af forskellige variable fra datasættet for at få et overblik over data.

Drughunters
Ottiliavej 9
2500 Valby
Tel +45 3630 1311
Fax +45 3630 1940
E-mail drughunters@lundbeck.com
www.drughunters.dk

Diskutér om der allerede nu kan etableres nogle konklusioner baseret på de tabeller og grafiske repræsentationer, I har lavet.

I bestemmer selv, hvor meget beskrivende statistik og figurer I vil præsentere for at give et overblik over de vigtige variable i jeres datasæt, men det vil selvfølgelig være oplagt at inkludere variable, der indeholder oplysninger om depression og angst.

Variable kan antage *kontinuerte* værdier (alder, højde, vægt etc.), eller de kan være angivet i kategorier (køn, aldersgruppe, blodtype etc.). I det næste spørgsmål vil vi se på sammenhængen mellem to kontinuerte variable.

3) Betragt de to kontinuerte variable, der omhandler depression og angst (phq_score og gad_score).

   Diskutér hvorfor det er interessant at sammenligne disse to variable.

Nu skal I lave analyser på phq_score og gad_score fra forrige spørgsmål.

4) For de to kontinuerte variable skal I lave et såkaldt punktdiagram (scatterplot), hvis I ikke allerede har gjort det i 2). Et punktdiagram er et plot, hvor I indsætter punkterne fra de to kontinuerte variable i et koordinatsystem.

   Tilføj en tendenslinje (regressionslinje eller trendlinje) og dens ligning. Forklar betydningen af skæringen med y-aksen og betydningen af konstanten a (hældning)

Hvis I betragter de resultater, I foreløbig er nået frem til, kan I begynde at overveje, hvad der betinger de sammenhænge, I måtte have fundet. Her er to begreber meget vigtige: *kausalitet* og *korrelation*.

5) Forklar med jeres egne ord, hvad disse to begreber betyder. Læg specielt vægt på forskellen mellem de to begreber.

   Hvis I har fundet nogle sammenhænge i spørgsmålene oven for, skal I diskutere, om sammenhængen kan skyldes korrelation eller kausalitet.

Drughunters
Ottiliavej 9
2500 Valby
Tel +45 3630 1311
Fax +45 3630 1940
E-mail drughunters@lundbeck.com
www.drughunters.dk

I mange skalaer inden for klinisk forskning arbejder man med såkaldte cut-offs, dvs. scorer, der definerer en bestemt tilstand. Lad os antage, at der for PHQ-scoren gælder følgende (I kan i excel-arket for hver patient se vurderingen af Depression Severity):

PHQ score	Depression Severity
0-4	None-minimal
5-9	Mild
10-14	Moderate
15-19	Moderately Severe
20-27	Severe

6)

a. Hvis ikke I allerede har gjort dette, skal I lave et histogram over PHQ-scoren for de 783 patienter og angive gennemsnit og standardafvigelse for fordelingen af PHQ-scoren og angiv ved optælling, hvor stor en procentdel af patienterne, der har Depression Severity: Severe
b. Synes I, det ser ud som om, at PHQ-scoren er normalfordelt? Argumentér for jeres svar.

Nu *antager* vi imidlertid, at PHQ-scoren er normalfordelt.

c. Angiv sandsynligheden for, at PHQ-scoren er >= 20 ved at 'slå op' i en normalfordeling med det gennemsnit og den standardafvigelse, I fandt i spørgsmål 6.a. Forklar hvorfor denne sandsynlighed ikke er identisk med procentdelen fra spørgsmål 6.a.

I de næste spørgsmål skal I arbejde med såkaldte simuleringer af data, dvs. at I *konstruerer* nye data, der *ligner* de data, I startede med i excel-arket. Dette gør man bl.a. for at få nogle mere robuste analyser og resultater. Det kan lade sig gøre at opnå dette, da man nu har meget mere data at arbejde med end oprindeligt (hvor man kun havde det ene datasæt, I sidder med nu).

I dette tilfælde er det PHQ-scoren, I skal arbejde med.

Drughunters
Ottiliavej 9
2500 Valby
Tel +45 3630 1311
Fax +45 3630 1940
E-mail drughunters@lundbeck.com
www.drughunters.dk

7)

a. Ved brug af det gennemsnit og den standardafvigelse I fandt i spørgsmål 6.a, skal I nu simulere PHQ-score variable, 25 gange hver med 783 studerende. Tegn et histogram over en tilfældig udvalgt af disse variable. Hvordan ser dette histogram ud? Ligner det histogrammet fra spørgsmål 6.a? Er det en korrekt måde at simulere PHQ-scoren på?
b. Simulér én variabel af længden 25*783=19575. Forklar hvorfor det er det samme som at simulere 25 variable hver af længden 783?
c. Hvad er sandsynligheden for at have Depression Severity = 'Severe' i dette simulerede datasæt af længden 19575 (angiv gennemsnit og standardafvigelse og bestem sandsynligheden ved hjælp af disse)? Diskutér om denne sandsynlighed stemmer overens med de fundne tal i spørgsmål 6.a. og i spørgsmål 6.c.?
d. Kan man sige noget om, hvor stor en del af befolkningen, der kan have severe depression? Skal man stole på tallet i spørgsmål 6.a., spørgsmål 6.c. eller spørgsmål 7.c.? Hvor mange mennesker svarer det til i Danmark? Argumentér for jeres svar.
e. Er det muligt på det nuværende grundlag at komme med nogle konklusioner vedrørende depression og angst? I så fald, hvad vil I anbefale til sundhedsmyndighederne eller Drughunters Medicines Corp.? Argumentér for jeres svar.

Hvordan man finder studerende med fællestræk i et datasæt ved hjælp af kunstig intelligens?

Kunstig intelligens refererer til at skabe computersystemer, der kan udføre "intelligent adfærd" som mennesker. *Machine learning* er et specialiseret område inden for kunstig intelligens, hvor computere bruger matematik til at lære ny information fra data uden at være specifikt programmeret til, hvad de skal gøre.

I machine learning er der to kategorier af metoder: *Supervised learning*-metoder og *Unsupervised learning*-metoder. En Supervised learning-metode trænes på et dataset, hvor vi ved, hvad "det rigtige svar" er. Hvis man f.eks. både har data på personer som udviklede en sygdom, og på personer som ikke udviklede sygdommen, kan man træne en algoritme til at forudsige, om en person kommer til at udvikle en sygdom eller ej. For Unsupervised learning-metoder – som f.eks. K-means clustering – er ikke noget "korrekt svar"; Her opdager algoritmerne mønstre i datasættene af sig selv.

Drughunters
Ottiliavej 9
2500 Valby
Tel +45 3630 1311
Fax +45 3630 1940
E-mail drughunters@lundbeck.com
www.drughunters.dk

I dette afsnit af opgaven skal I lære at anvende en Unsupervised machine learning-metode kaldet *clustering*, hvis mål er at organisere data, der har samme karakteristika, i undergrupper.

I dette tilfælde vil vi gerne vide: er der grupper af studerende ud af de 783, der i en eller anden forstand ligner hinanden? Hvis ja, hvad kan vi så lære om de forskellige grupper?

Til dette vil vi bruge metoden *K-means clustering*. Helt enkelt forklaret fungerer K-means ved følgende trin:

1. Definér antallet af grupper, som vi kalder *clusters*. Disse vil være antallet af clusters, som vi skal organisere datapunkter med fællestræk i.
2. Initialisér *centroids*, som er et sæt koordinater, der tjener som midtpunktet for hver gruppe eller cluster.
3. Mål den euklidiske (pythagoræiske) afstand mellem hvert datapunkt og centroids i hvert cluster.
4. Tildel hvert datapunkt til det cluster, hvis centroid det er tættest på.

Efter at datapunkterne er organiseret i deres første clusters, opdateres centroids-koordinaterne til at være gennemsnitskoordinaterne for alle datapunkterne i hvert cluster. Afstande genberegnes mellem hvert datapunkt og hvert nyt centroid for opdateret cluster-tildeling. Disse faser ("centroid-opdatering" og "cluster-tildeling") fortsætter i et bestemt antal iterationer, eller indtil datapunkterne ikke længere ændrer sig.

Nu vil vi prøve dette på datasættet for depression og angst!

Forbered en delmængde af jeres data

I et nyt Excel-ark skal I kopiere studerende med id-nummer 43, 136 og 289 for at tjene som cluster centroids (deres dataværdier vil tjene som cluster centroids for begyndelsen af clusters 1, 2 og 3). Disse tre studerende vil også være tre ud af 30 tilfældigt valgte studerende, der vil blive grupperet efter trinene i k-means. Kopiér 27 tilfældige studerende mere for at opnå i alt 30 studerende. Vi vil gruppere studerende baseret på deres **phq_score** og **gad_score**.

Drughunters
Ottiliavej 9
2500 Valby
Tel +45 3630 1311
Fax +45 3630 1940
E-mail drughunters@lundbeck.com
www.drughunters.dk

Beregn k-means ved at opdele studerende i 3 clusters

For hver studerende skal I måle den euklidiske afstand mellem denne studerende og hver centroid for de to variable. Det kan hjælpe at ”låse” en celle med ”$” symbolet, hvis I laver en formel til at lave beregningen. Evaluér derefter afstandene for hver studerende og indtast det cluster-nummer, som de har den korteste euklidiske afstand til i kolonnen ”Cluster assignment.” Det kan hjælpe at strukturere data i Excel sådan:

id	phq_score	gad_score	afstand til cluster 1	afstand til cluster 2	afstand til cluster 3	cluster tildeling
Centroid 1 - 43						
Centroid 2 - 136						
Centroid 3 - 289						
Studerende 1 - 43						
Studerende 2 - 136						
Studerende 3 - 289						
Tilfældig studerende 4						
Tilfældig studerende 5						
Tilfældig studerende 6						
...						
Tilfældig studerende 28						
Tilfældig studerende 29						
Tilfældig studerende 30						

I skal i tabellen oven for selv udfylde de første to kolonner. De første tre rækker skal være med værdier for de tre studerende, hvis værdier er de første centroids. De vises også i de næste tre rækker, som studerende, der skal grupperes efter k-means, efterfulgt af rækker af tilfældigt udvalgte studerende, som også vil bliv grupperet.

8) Skriv formlen I vil bruge til at måle afstanden mellem hvert af datapunkterne og hvert cluster centroid.

Visualisér cluster-tildeling

I har gennemført den første iteration af k-means cluster-tildeling, og hvert datapunkt skal nu have et clusternummer tildelt, som er det cluster, hvis centroid datapunktet har kortest afstand til.

9)

a. Opret et scatterplot af datapunkterne for at visualisere de resulterende clusters, så datapunkterne i hvert cluster har samme farve, og visualisér centroiderne. I kan følge guiden neden for (se understøttende links for at få hjælp).

Drughunters
Ottiliavej 9
2500 Valby
Tel +45 3630 1311
Fax +45 3630 1940
E-mail drughunters@lundbeck.com
www.drughunters.dk

Opdatér centroiderne til at være gennemsnittet af værdierne af datapunkterne i hvert cluster. I har nu nye centroids at måle afstanden til for hvert datapunkt. Mål afstanden mellem hvert datapunkt og hvert nyt cluster centroid, og tildel hvert datapunkt det nærmeste cluster centroid.

b. Opret et nyt scatterplot, der visualiserer den nye cluster-tildeling og centroids efter denne anden iteration.
c. Gentag denne ”centroid-opdatering” og ”cluster-tildeling” en gang mere i 3 samlede iterationer eller indtil datapunkterne ikke længere skifter deres cluster-tildeling, så I har op til tre samlede scatterplots, der visualiserer clusterne og centroiderne efter hver cluster-tildeling.

Tillykke! I har netop udført AI med k-means clustering på en delmængde af dette datasæt.

I det sidste afsnit skal I arbejde med Online notebook med allerede forberedt Python-kode til at udføre k-means clustering på 3 variable med all dataeksemplerne i dette datasæt. I tillæg til at analysere gad_score og phq_score, skal I også analysere **bmi**.

10)

a. Beskriv kort bmi som variabel som står for ”body mass index”. Forklar hvorfor bmi er en interessant variabel at analysere sammen med gad_score og phq_score, når man undersøger mennesker med depression og angst?
b. Følg trinene i notesbogen for at gruppere alle studerende baseret på gad_score, phq_score og bmi. Inkludér den endelige visualisering af clusterne inklusive en fortolkning af de resulterende clusters.
c. Er det muligt at finde andre clusters i dette datasæt med forskellige variable eller forskelligt antal clusters? I kan vælge op til 3 kontinuerte variable og ændre antallet af clusters ved at følge instruktionerne ved siden af de gule stjerner i notesbogen. Hvis I har fundet nogle clusters, bedes I inkludere en visualisering og en fortolkning af clusterne. Hvis I har fundet nogle clusters, hvad er da de næste trin, I vil tage for at verificere eller validere det, I har fundet? Bemærk: Nogle gange er der ingen resulterende clusters. Dette sker ofte i virkeligheden.

Generel opgavevejledning

Overordnet set er opgaven opbygget efter følgende model:

- **Spørgsmål 1** omhandler en valgt hjernesygdom og et valgt datasæt. Her handler det primært om at vise, at man er i stand til at udvælge hovedtrækkene og give en så kort og præcis beskrivelse som muligt.

Drughunters
Ottiliavej 9
2500 Valby
Tel +45 3630 1311
Fax +45 3630 1940
E-mail drughunters@lundbeck.com
www.drughunters.dk

- **Spørgsmål 2-7** omhandler matematiske metoder til grafisk og analytisk at få overblik over og analysere data. Bemærk, at der generelt ikke er noget korrekt facit med to streger under. Alle besvarelser afhænger af de valg, der hele tiden foretages fra jeres side. Det er det, en forsker gør i sit daglige arbejde.
- **Spørgsmål 8-10** omhandler kunstig intelligens til analyser af data og grafisk visualisering og fortolkning af resultaterne.
- Vi vil rigtig gerne se jeres posterpræsentation uanset om I har svaret på alle dele af opgaven eller ej.

Til eleverne

Som forsker må man leve med, at der ikke findes endegyldige og korrekte svar. Man må opsøge viden, som andre har skabt eller ved at lave sine egne forsøg. Og så må man med åbent sind holde den viden op imod sin egen videnskabelige hypotese, som derved be- eller afkræftes – eller som oftest kræver yderligere viden for at kunne drage en konklusion. Det kan være en lang og frustrerende proces selv for garvede forskere. Derfor forventer vi selvfølgelig ikke endegyldige løsninger fra jer, men gode forslag hvor der er tænkt over usikkerheder og begrænsninger.

Vi har forsøgt at hjælpe ved at give nogle links nedenfor og på vores hjemmeside Drughunters. Men det er ikke en udtømmende liste, så I kan sikkert sagtens finde mere og anden information selv. At kunne opsøge information og have en kritisk tilgang til sine kilder er en meget vigtig kompetence som forsker.

Til finaledagen vil bedømmelseskriterierne være 1/3 formidling og 2/3 faglighed. Det betyder, at det ikke gælder om at have så meget tekst som muligt, men at der skal være et naturligt flow i fortællingen, så læseren/tilhøreren kan forstå jeres vigtigste pointer. Omvendt er det selvfølgelig heller ikke nok at have en superflot poster, hvis man ikke har svaret på spørgsmålene. Husk at til den mundtlige præsentation behøver I ikke at gennemgå posteren slavisk. Her skal I fokusere på at fremhæve de pointer, som er særligt vigtige for jeres besvarelse. Dommerne har læst posteren på forhånd, men gemmer den endelige bedømmelse til de har set jeres præsentation, hvor de både vil inddrage jeres evne til at fortælle en sammenhængende historie og jeres besvarelse af opfølgende faglige spørgsmål.

Posteren skal ikke nødvendigvis være opdelt således, at I skal gennemgå alle 10 spørgsmål, men kan også være et udvalg af de figurer, tabeller og analyser, I har lavet.

Den skriftlige vurdering er selvfølgelig kun lavet på baggrund af posteren og skal ses som en kort tilbagemelding, ikke en dybtgående analyse af jeres poster.

Drughunters
Ottiliavej 9
2500 Valby
Tel +45 3630 1311
Fax +45 3630 1940
E-mail drughunters@lundbeck.com
www.drughunters.dk

Rent praktisk skal posteren indsendes som pdf i størrelsen 142x83 cm landskabsformat. Se kalenderen nedenfor.

Til lærerne

Brug gerne tid i klassen på at snakke om, hvordan hvert enkelt spørgsmål skal forstås, inden I kaster jer over besvarelsen.

Der kan hentes inspiration til, hvordan man kan arbejde med opgaverne på vores hjemmeside Drughunters.

Eksempler på referencer og links (find gerne flere selv)

Deskriptiv statistik, optællinger, lineær regression, sandsynligheder i normalfordeling og simuleringer i Excel:

- Excel - statistisk bearbejdning af stort datasæt - YouTube
- Statistik i excel - YouTube
- Adding The Trendline, Equation And R2 In Excel - YouTube
- How to Count Cells in Microsoft Excel (COUNT, COUNTA, COUNTIF, COUNTIFS Functions) - YouTube
- Excel Histogram with Normal Distribution Curve - YouTube
- Calculating Probabilities Using the Normal Distribution Function in Excel - YouTube
- Data simulation in Excel - YouTube
- How to Lock Cell References in Excel - YouTube
- Distance formula - Excel formula | Exceljet
- How to Calculate Euclidean Distance in Excel - Statology
- How to Create Multi-Color Scatter Plot Chart in Excel - YouTube

AI clustering:

- K-Means Clustering with Math. Common Unsupervised learning technique… | by sampath kumar gajawada | Towards Data Science

Drughunters
Ottiliavej 9
2500 Valby

Tel +45 3630 1311
Fax +45 3630 1940

E-mail drughunters@lundbeck.com
www.drughunters.dk

Kalender for Drughunters 2024

Med venlig hilsen
Drughunters 2024

Drughunters
Ottiliavej 9
2500 Valby
Tel +45 3630 1311
Fax +45 3630 1940
E-mail drughunters@lundbeck.com
www.drughunters.dk

Bilag 1

Nedenfor finder I et datasæt relateret til hjernesygdom. For at få adgang til data skal I oprette en konto på Kaggle og logge ind.

1. Depression og angst datasæt

Datasæt: Depression and anxiety data | Kaggle

Datasætbeskrivelse: Dette er et datasæt, der indeholder 783 rækker med information om patienter med/uden depressive og angst tendenser.

Variabel-information (for kategoriske variable er 0=nej, 1=ja):

1) **id:** nyt nummer for hver række viser at oplysningerne for hver patient står i samme række
2) **school_year:** en kategorisk variabel, hvor et lavt tal betyder kort skolegang, og et højt tal betyder lang skolegang
3) **age:** alder
4) **gender:** køn
5) **bmi:** BMI
6) **who_bmi:** kategorisk variabel, der viser inddelingen af BMI i vægt-kategorier
7) **phq_score:** variabel der viser depressionsniveau for den enkelte patient (lav score: mild depression, høj score: slem depression)
8) **depression_severity:** kategorisk variabel, der viser inddelingen af phq_score i depressionskategorier
9) **depressiveness:** indikation af, om patienten er depressiv eller ej
10) **suicidal:** selvmordstruet eller ej
11) **depression_diagnosis:** faktisk diagnose af depression
12) **depression_treatment:** behandlet for depression eller ej
13) **gad_score:** variabel der viser angstniveau for den enkelte patient (lav score: mild angst, høj score: slem angst
14) **anxiety_severity:** kategorisk variabel, der viser inddelingen af gad_score i angstkategorier
15) **anxiousness:** indikation af om patienten er angst eller ej
16) **anxiety_diagnosis:** faktisk diagnose af angst
17) **anxiety_treatment:** behandlet for angst eller ej
18) **epworth_score:** en 'søvnigheds'-score (lav score: ikke søvnig, høj score: meget søvnig)
19) **sleepiness:** indikation af, om patienten kan sove eller ej

Drughunters
Ottiliavej 9
2500 Valby
Tel +45 3630 1311
Fax +45 3630 1940
E-mail drughunters@lundbeck.com
www.drughunters.dk

Hints til klargøring af datasæt:

Klargøring af datasæt:

I kan overveje følgende for at jeres datasæt er nemmere at arbejde med:

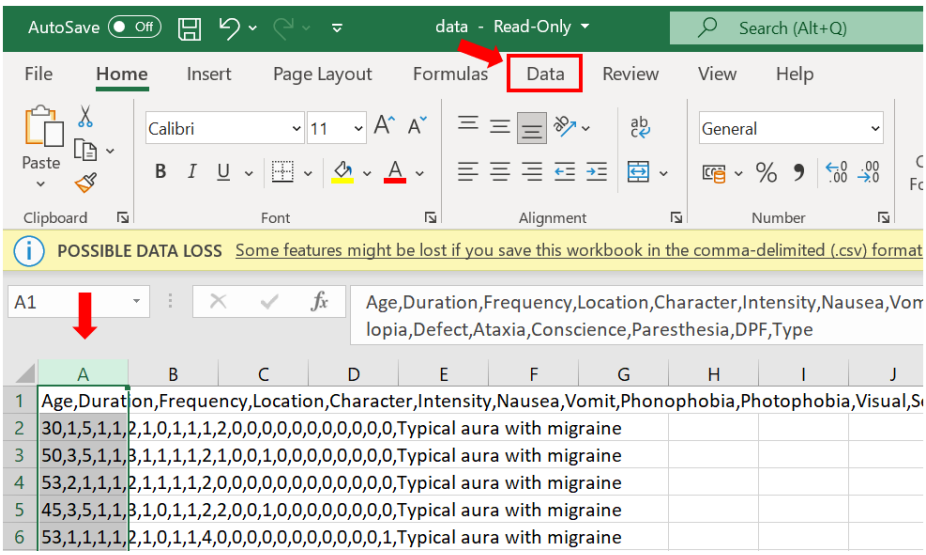
- Er der rækker med manglende værdier? I så fald kan I løse dette problem på flere måder: I kan beregne gennemsnittet eller medianen af de eksisterende værdier og sætte dette tal ind, hvor der er manglende værdier. I kan også beslutte jer til at slette disse rækker, og I kan faktisk også vælge at beholde dem.

.csv to excel file

Når I downloader et datasæt, kan det komme som en .csv-fil.

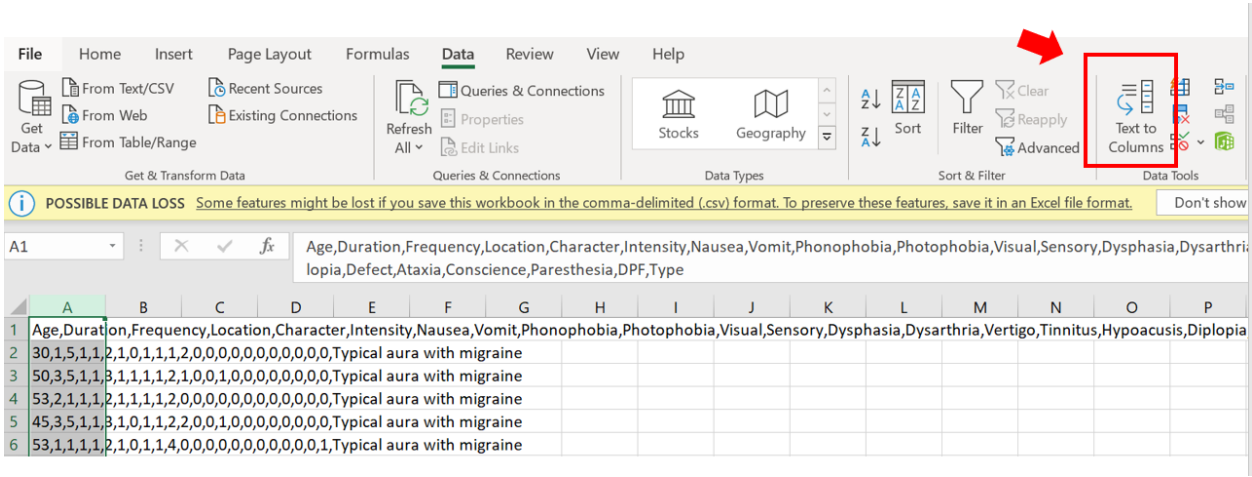
Følg disse nemme trin for at gøre det til et Excel-sheet.

1. Fremhæv kolonnen, der indeholder alle data, og klik på fanen Data.

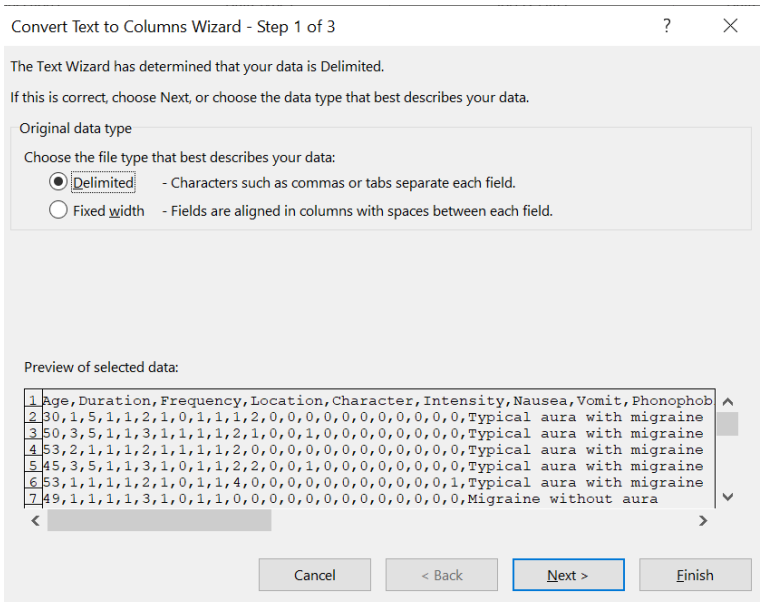


Drughunters
Ottiliavej 9
2500 Valby
Tel +45 3630 1311
Fax +45 3630 1940
E-mail drughunters@lundbeck.com
www.drughunters.dk

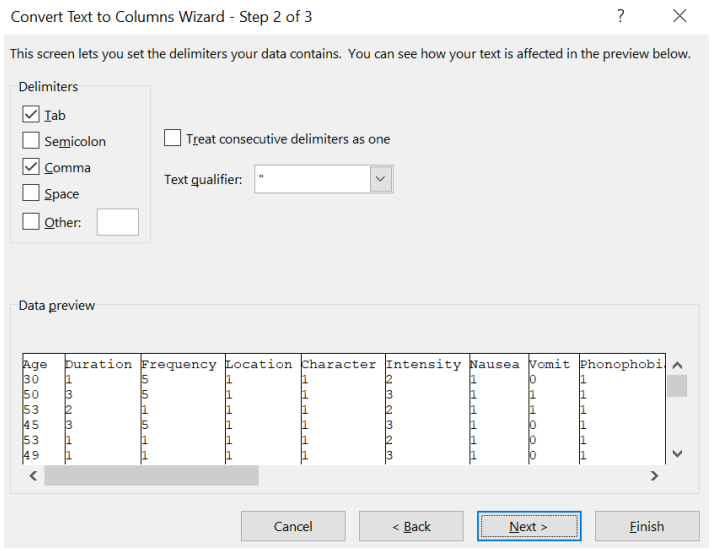
2. Klik på ”Text to Columns.”

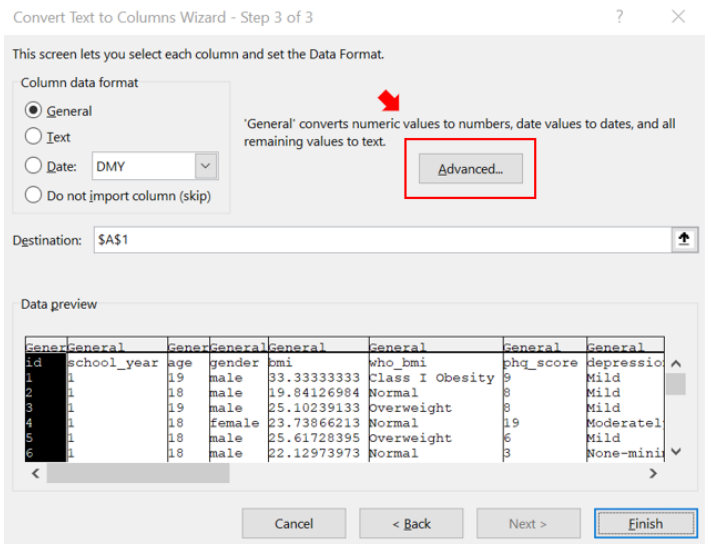


3. Fuldfør guidens trin med følgende indstillinger:

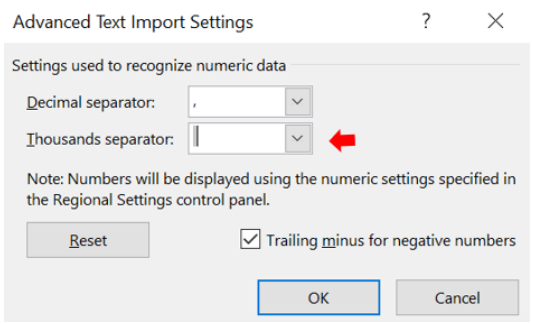


Drughunters
Ottiliavej 9
2500 Valby
Tel +45 3630 1311
Fax +45 3630 1940
E-mail drughunters@lundbeck.com
www.drughunters.dk



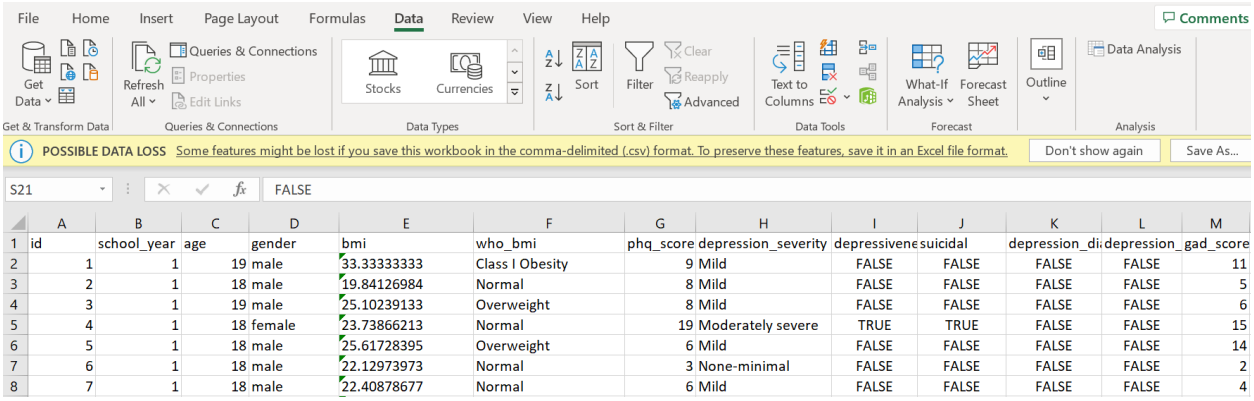


Indstil “tusinder separatoren” til et tomt mellemrum, ” ”.



4. Færdig. Gem gerne resultatet af disse trin som en Excel-version.

Drughunters
Ottiliavej 9
2500 Valby
Tel +45 3630 1311
Fax +45 3630 1940
E-mail drughunters@lundbeck.com
www.drughunters.dk



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	id	school_year	age	gender	bmi	who_bmi	phq_score	depression_severity	depressivene	suicidal	depression_di	depression_	gad_score
2	1	1	19	male	33.33333333	Class I Obesity	9	Mild	FALSE	FALSE	FALSE	FALSE	11
3	2	1	18	male	19.84126984	Normal	8	Mild	FALSE	FALSE	FALSE	FALSE	5
4	3	1	19	male	25.10239133	Overweight	8	Mild	FALSE	FALSE	FALSE	FALSE	6
5	4	1	18	female	23.73866213	Normal	19	Moderately severe	TRUE	TRUE	FALSE	FALSE	15
6	5	1	18	male	25.61728395	Overweight	6	Mild	FALSE	FALSE	FALSE	FALSE	14
7	6	1	18	male	22.12973973	Normal	3	None-minimal	FALSE	FALSE	FALSE	FALSE	2
8	7	1	18	male	22.40878677	Normal	6	Mild	FALSE	FALSE	FALSE	FALSE	4

5. I skal undervejs benytte EXCEL-funktionen 'Data Analysis'. Hvis I ikke finder den ude til højre under 'Data', kan I gøre følgende:
   Tryk File -> Options -> Add-ins -> Analysis ToolPak (VBA) og installér dette

Upload datafil til Google Colab Notebook

Notesbogen er allerede forberedt og klar til brug. Det eneste I skal gøre er at uploade den originale depression_anxiety_data.csv-fil til den. I kan følge disse trin for at gøre det.

1. Klik på mappeikonet.
2. Klik på upload filikonet for at vælge den datafil, der er gemt på jeres lokale computer.