

# VERSLAG R0855381

Title	Verslag r0855381		
Course	Business Analytics Major		
OPO Code		Author	Amine Moussaif
Version	1.0	Review	
Created on		Last update	31/08/2023

## Roken verslag - Amine Moussaif

### 1. INTRODUCTIE:

Dit verslag gaat rond roken en de verschillende subonderwerpen hierover. Eerst begon ik met het zoeken van drie bruikbare datasets van verschillende formaat. Na een lange tijd zoeken vond ik alleen csv-datasets die bruikbaar waren. Hierdoor heb ik besloten 2 datasets te converteren naar een xlsx-dataset en naar een txt-dataset. Dit heb ik gedaan met de <https://convertio.co/> tool. Alle files die tot dit verslag behoren zijn te vinden op projectwerkt onder mijn persoonlijke repository

### 2. DATABRONNEN:

1. Body signals of smoking (csv file)

<https://www.kaggle.com/datasets/kukuroo3/body-signal-of-smoking>

2. Smoking related lung cancer (csv => converted to => xlsx)

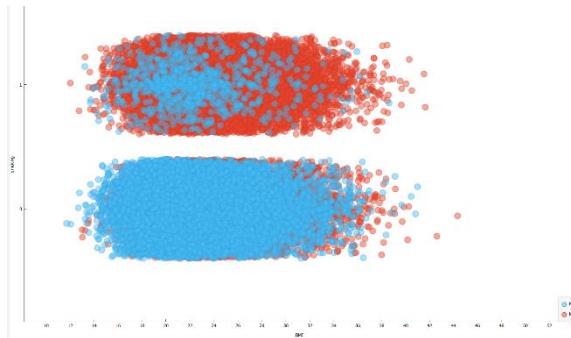
<https://www.kaggle.com/datasets/raddar/smoking-related-lung-cancers?resource=download>

3. Smoking dataset from the UK (csv => converted to => txt)

<https://www.openintro.org/data/index.php?data=smoking>

### 3. VRAGEN

1) Is er een verband tussen roken en een lage bmi hebben?



Aan de hand van deze scatter plot kunnen we zien dat het bij zowel de rokers en niet rokers gelijk verdeeld is. Dit wilt zeggen dat er geen directe verband is. Ik merkte hier ook op dat de meerderheid van de rokers mannen zijn. ( Aan de hand van model in 4.2)

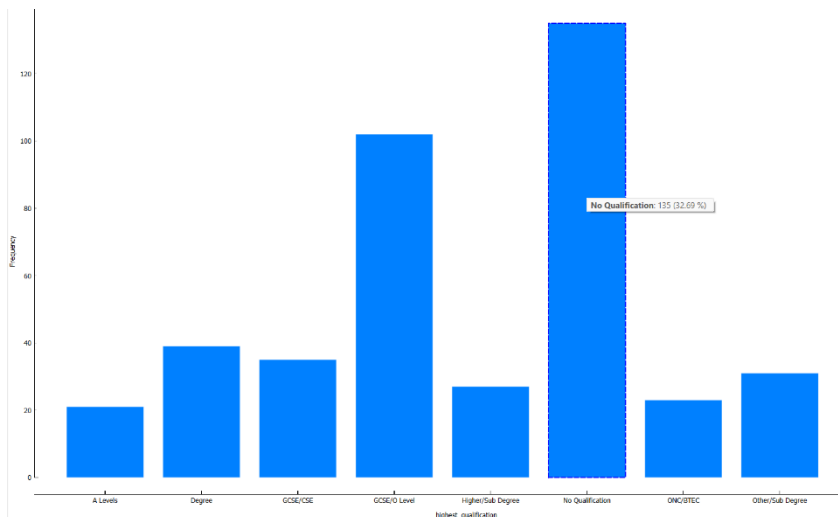
2) Kunnen we aan de hand van lichaam signalen voorspellen of iemand rookt ?

Model	AUC	CA	F1	Precision	Recall
Neural Network	0.800	0.717	0.720	0.727	0.717
Naive Bayes	0.756	0.689	0.694	0.729	0.689
Logistic Regression	0.754	0.675	0.669	0.667	0.675
5NN	0.604	0.612	0.602	0.598	0.612

De classification accuracy die de accuraatheid geeft van de predicties is niet vrij hoog maar ook niet laag. Van alle modellen zien we dat neural network het beste presteert met een CA van 71.7. Dit wilt zeggen dat het model 71.7 % van de tijd heeft kunnen voorspellen of

iemand rookt of niet. In deze context zou ik zeggen dat dit niet zeer accuraat is maar wel goed. (Aan de hand van model in 4.2)

3) Roken mensen zonder diploma/opleiding meer dan mensen met diploma/opleiding zoals de stereotype vertelt?



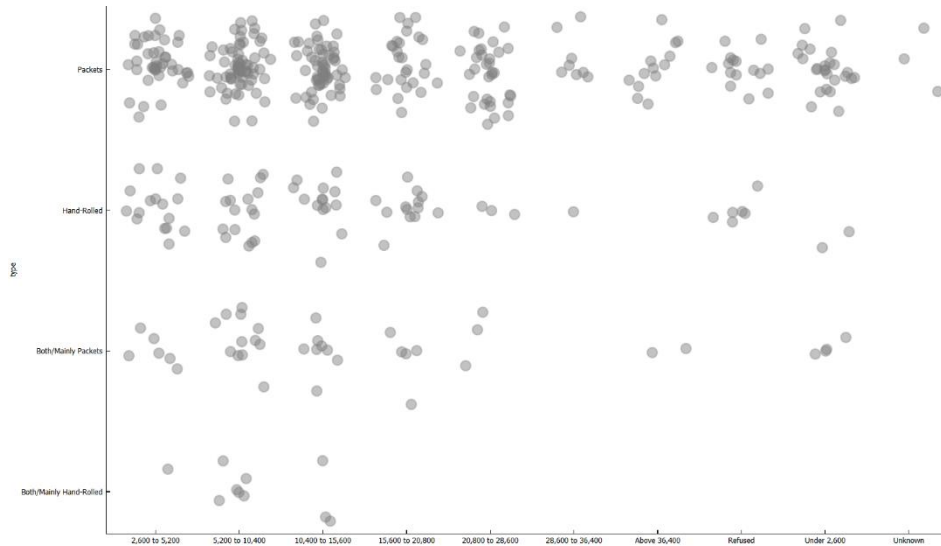
Op eerste zicht dacht ik dat het inderdaad zo was, omdat je duidelijk kan zien dat 32.69% van de rokers geen diploma/opleiding heeft. Daarna realiseerde ik me dat als je alle opleidingsniveaus samen optelt, je een percentage bekomt van 67.31% wat veel hoger is dan de mensen zonder opleiding. Dit bewijst dus dat in deze studie mensen met een opleiding/qualificatie meer roken dan de mensen zonder. (kijk 4.1)

4) Kan kNN voorspellen wat een person rookt(packets, hand-rolled or both) aan de hand van inkomen, opleidingsniveua, enz..?

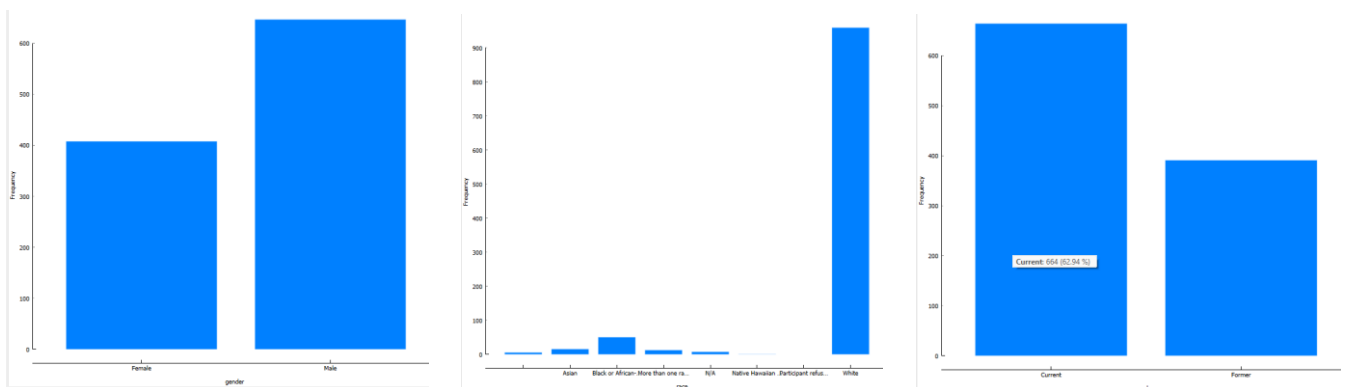
Model	AUC	CA	F1	Precision	Recall
kNN	0.572	0.713	0.593	0.508	0.713
Neural Network	0.566	0.644	0.566	0.505	0.644
Naive Bayes (1)	0.529	0.299	0.378	0.519	0.299
Logistic Regression (1)	0.533	0.655	0.564	0.495	0.655

kNN bekomt een percentage van 71.3% wat in deze context vrij goed is. (kijk 4.1)

Ik merkte hier ook op dat mensen met een lager inkomen meer rollen met de hand tegenover de mensen die een hoger inkomen hebben. Deze roken vooral pakjes. (kijk 4.1)



## 5) Welke soort mensen heeft stage 2 kanker?

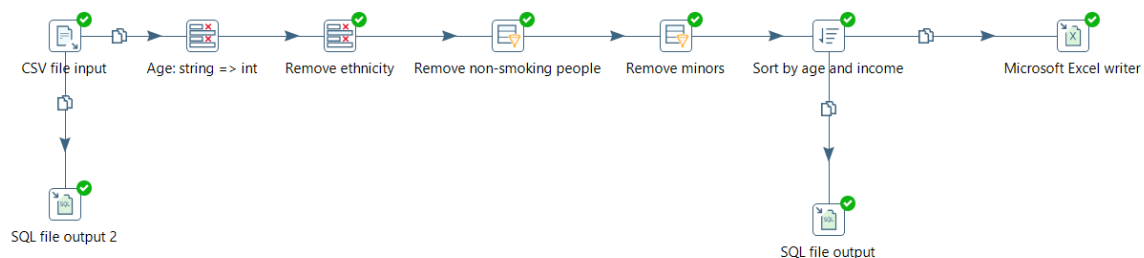


Hoge stage 2 longkanker wordt meer gevonden bij blanke mannen die nog steeds roken. Een belangrijke factor die hier natuurlijk speelt is dat deze data is verzameld in Amerika en daar is de meerderheid natuurlijk blank dus is het ook logisch dat dit in de distributie blijkt.

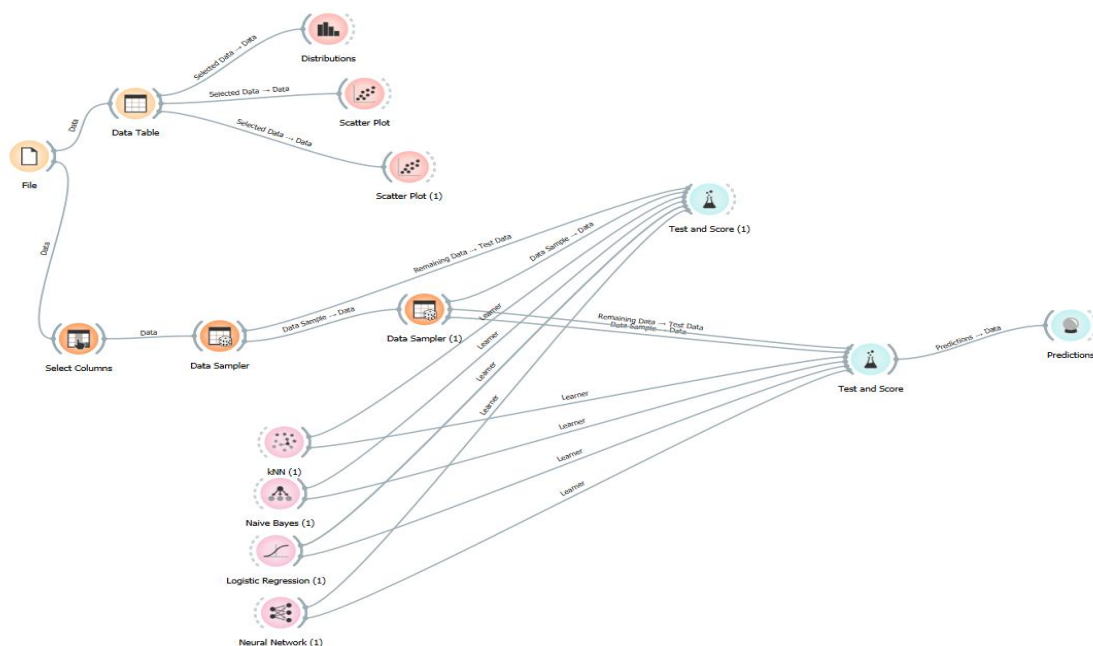
(Kijk naar model 4.3)

## 4.1 ETL VOOR DE CSV FILE:

We beginnen met het inladen van de csv file. Eerst begin ik met het converteren van de age field van string naar integer. Daarna verwijderen ik de ethnicity kolom omdat deze geen meerwaarde had aan de analyse. Ook heb ik de niet-rokers verwijderd van de dataset omdat ik in mijn onderzoek alleen de rokers wou. Ook verwijder ik de minderjarige omdat ik het verband van inkomen wil vergelijken met wat er gerookt wordt en minderjarigen hebben geen echte inkomen.

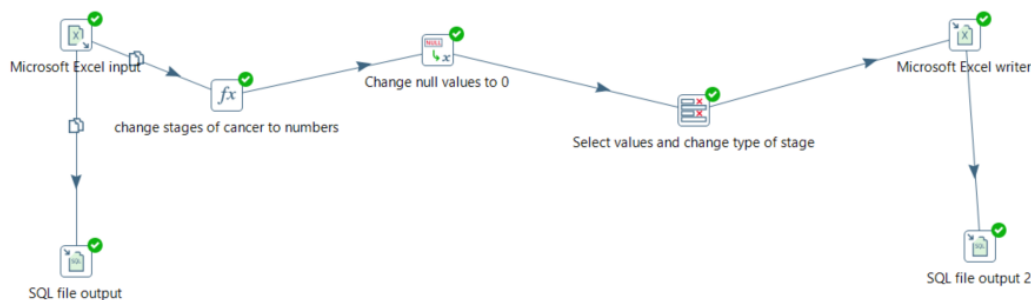


Daarna sorteren we de rijen op leeftijd en inkomen. Daaruit creëren we een SQL file output die we daarna gebruiken de data in de database te zetten, en een xml file die we kunnen gebruiken in orange om de data te analyseren

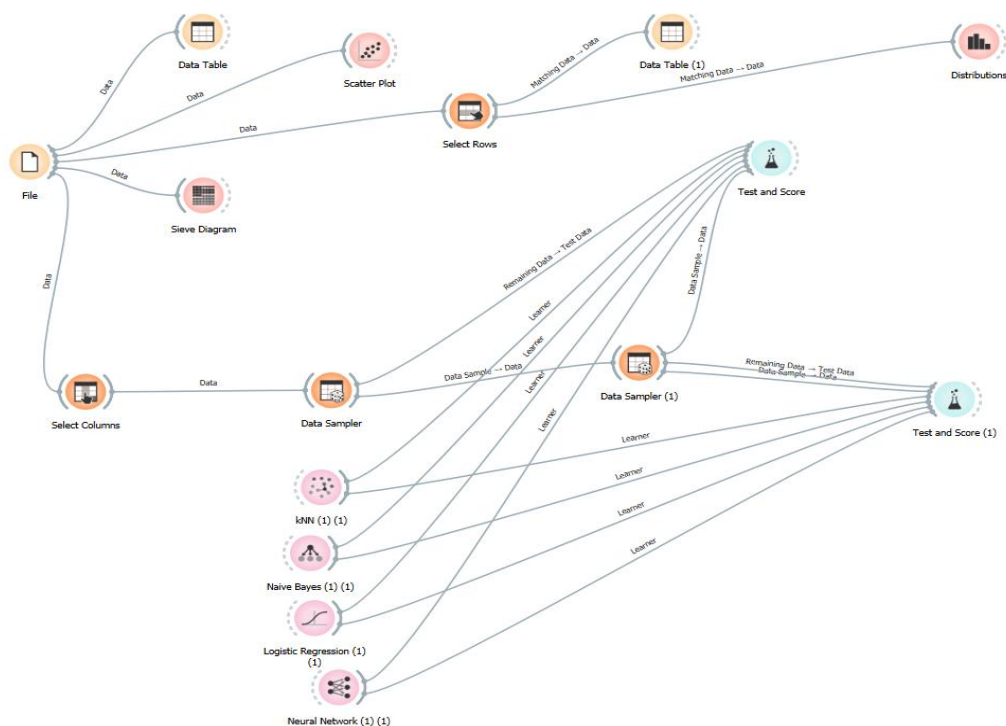




## 4.3 ETL VOOR DE XLSX FILE:

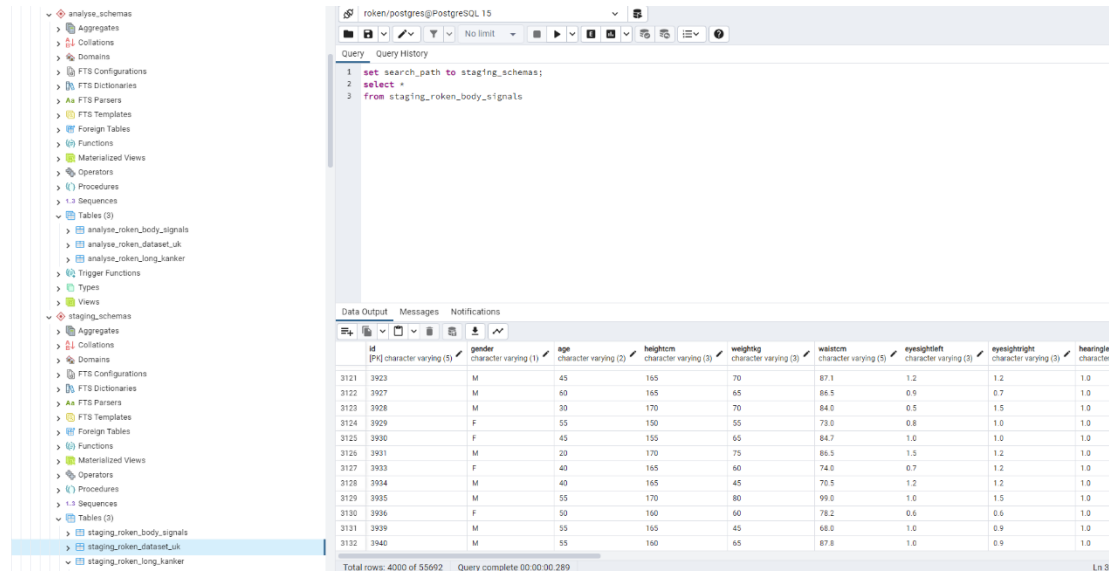


We beginnen met de microsoft excel input. Daarna heb ik een functie geschreven die de stages van kanker verandert in 0, 1 of 2. Dit is gemakkelijker om straks te analyseren. Ook waren er veel null waardes die ik veranderd heb in een 0. De null values van race heb ik veranderd naar "Other". Hierna heb ik de kolommen gekozen die ik nodig had en heb ik ook de type van stage\_of\_cancer veranderd naar int omdat deze na de change van null values een string was geworden. Als laatste heb ik een excel writer en een sql output toegevoegd om de data in orange te analyseren en om de data op de analyse database te gooien.



## 5.DATABANK(DS)

Local:



Query:

```
1 set search_path to staging_schemas;
2 select *
3 from staging_roken_body_signals
```

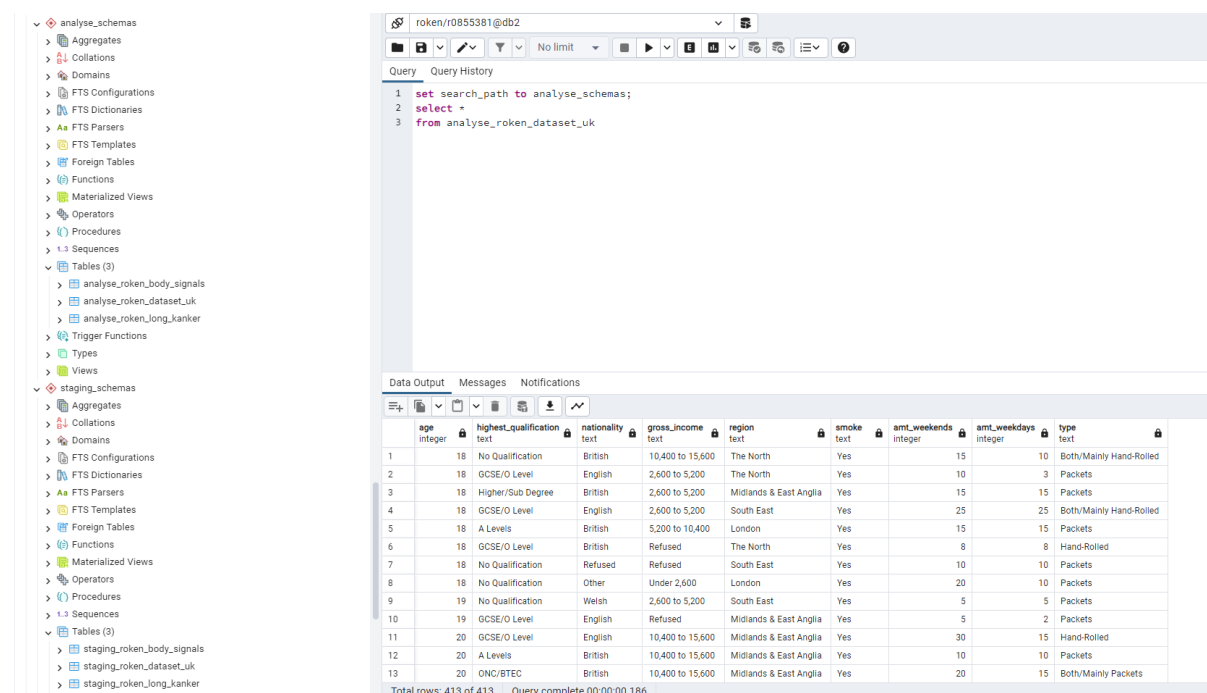
Data Output:

id	gender	age	height	weight	walston	eyesightleft	eyesightright	hearingleft	hearingright
3121	M	45	165	70	87.1	1.2	1.2	1.0	1.0
3122	M	60	165	65	86.5	0.9	0.7	1.0	1.0
3123	M	30	170	70	84.0	0.5	1.5	1.0	1.0
3124	F	55	150	55	73.0	0.8	1.0	1.0	1.0
3125	F	45	155	65	84.7	1.0	1.0	1.0	1.0
3126	M	20	170	75	86.5	1.5	1.2	1.0	1.0
3127	F	40	165	60	74.0	0.7	1.2	1.0	1.0
3128	M	40	165	45	70.5	1.2	1.2	1.0	1.0
3129	M	55	170	80	99.0	1.0	1.5	1.0	1.0
3130	F	50	160	60	78.2	0.6	0.6	1.0	1.0
3131	M	55	165	45	68.0	1.0	0.9	1.0	1.0
3132	M	55	160	65	87.8	1.0	0.9	1.0	1.0

Total rows: 4000 of 55692 Query complete 00:00:00.289

Aan de hand van de sql-file outputs heb ik de data op mijn local database kunnen zetten en op de school-database .

52223:



Query:

```
1 set search_path to analyse_schemas;
2 select *
3 from analyse_roken_dataset_uk
```

Data Output:

age	highest_qualification	nationality	gross_income	region	smoke	amt_weekends	amt_weekdays	type
18	No Qualification	British	10,400 to 15,600	The North	Yes	15	10	Both/Mainly Hand-Rolled
18	GCSE/O Level	English	2,600 to 5,200	The North	Yes	10	3	Packets
18	Higher/Sub Degree	British	2,600 to 5,200	Midlands & East Anglia	Yes	15	15	Packets
18	GCSE/O Level	English	2,600 to 5,200	South East	Yes	25	25	Both/Mainly Hand-Rolled
18	A Levels	British	5,200 to 10,400	London	Yes	15	15	Packets
18	GCSE/O Level	British	Refused	The North	Yes	8	8	Hand-Rolled
18	No Qualification	Refused	Refused	South East	Yes	10	10	Packets
18	No Qualification	Other	Under 2,600	London	Yes	20	10	Packets
19	No Qualification	Welsh	2,600 to 5,200	South East	Yes	5	5	Packets
19	GCSE/O Level	English	Refused	Midlands & East Anglia	Yes	5	2	Packets
20	GCSE/O Level	English	10,400 to 15,600	Midlands & East Anglia	Yes	30	15	Hand-Rolled
20	A Levels	British	10,400 to 15,600	Midlands & East Anglia	Yes	10	10	Packets
20	ONC/BTEC	British	10,400 to 15,600	Midlands & East Anglia	Yes	20	15	Both/Mainly Packets

Total rows: 413 of 413 Query complete 00:00:00.186



## 6. CONCLUSIE

In dit verslag keken we naar verschillende data uit verschillende landen. De dataset van body signals kwam bijvoorbeeld uit Zuid-Korea en die van dataset\_uk kwam van Engeland. Dit was interessant omdat elke land natuurlijk andere statistieken heeft.

Om de ruwe data te optimaliseren en te verfijnen, hebben we gebruik gemaakt van Pentaho. Binnen dit proces hebben we ETL-methodologie (Extract, Transform, Load) toegepast, wat ons in staat stelde om de ruwe data te transformeren naar verfijnde en schone informatie. De gegenereerde output van Pentaho hebben we vervolgens ingezet voor data-analyse met behulp van het Orange platform. Ik heb ontdekt dat Pentaho een heel handige tool is voor het transformeren van data en het ook de data-analyse vergemakkelijkt.

Aan de hand van de geanalyseerde data heb ik al mijn vragen kunnen beantwoorden en heb ik veel interessante dingen geleerd over roken op veel verschillende manieren. Eerst leek de opdracht moeilijk en had ik veel moeite met het vinden van nuttige datasets, maar na het vinden van de datasets en het leren werken met Pentaho, werd de opdracht heel leuk en interessant. Deze ervaring heeft me sterk gemotiveerd om in de toekomst verder te gaan met data-analyse.