Predict and analyze patient outcome from medical record

1

Looking at data

Data

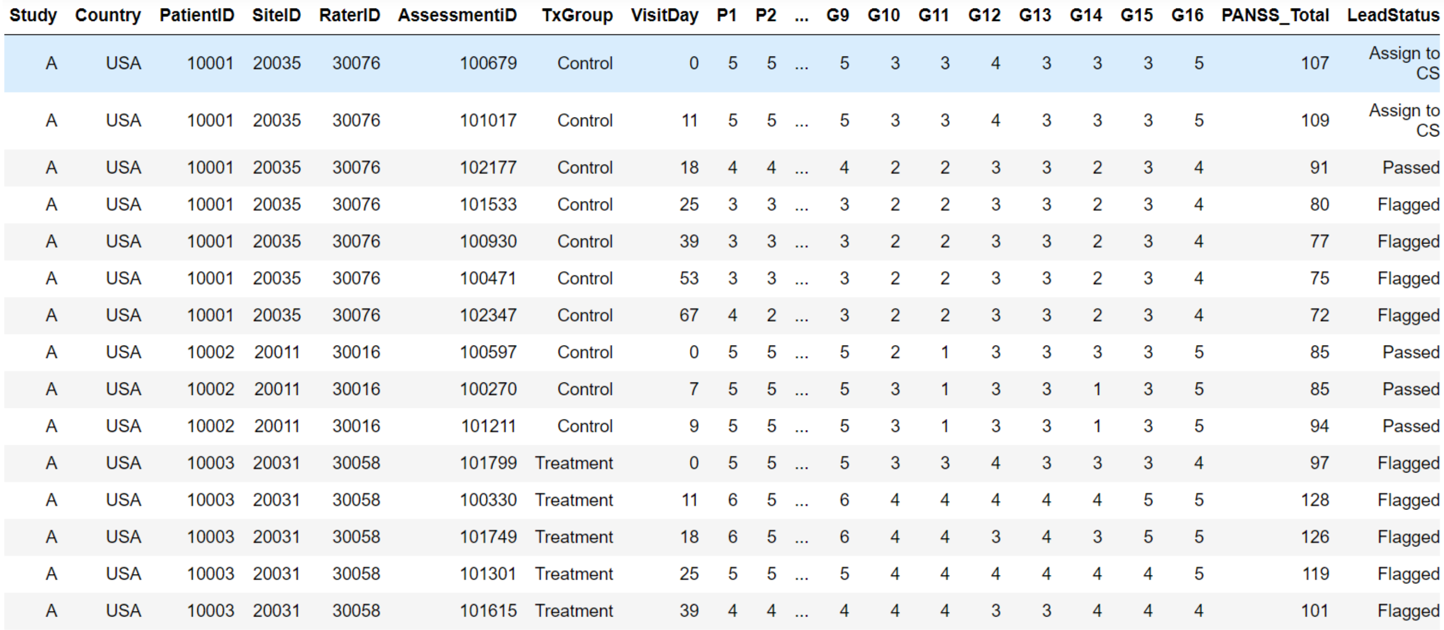
5 datasets with 5 trialss (studyA-E)

Variable:

1. Study - A character indicating which of the five studies the data represents.
2. Country - The country where the assessment was conducted.
3. PatientID - An identification number given to each unique patient.
4. SiteID - An identification number given to each unique assessment site.
5. RaterID - An identification number given to each unique rater.
6. AssessmentiD - An idenfication number given to each unique assessment conducted.
7. TxGroup - A string corresponding to the patient’s (randomly) assigned treatment group.
8. VisitDay - An integer corresponding to the number of days that have passed since the baseline assessment.
9. P1-P7 - The scores corresponding to each of the 7 positive symptoms of the assessment.
10. N1-N7 - The scores corresponding to each of the 7 negative symptoms of the assessment.
11. G1-G16 - The scores corresponding to each of the 16 general psychopathology symptoms of the assessment.
12. PANSS_Total - The sum of of the ratings across the 30 PANSS items.
13. LeadStatus - A string indicating whether the assessment’s audit passed, was flagged, or was assigned to a CS (i.e. clinical specialist).

Data

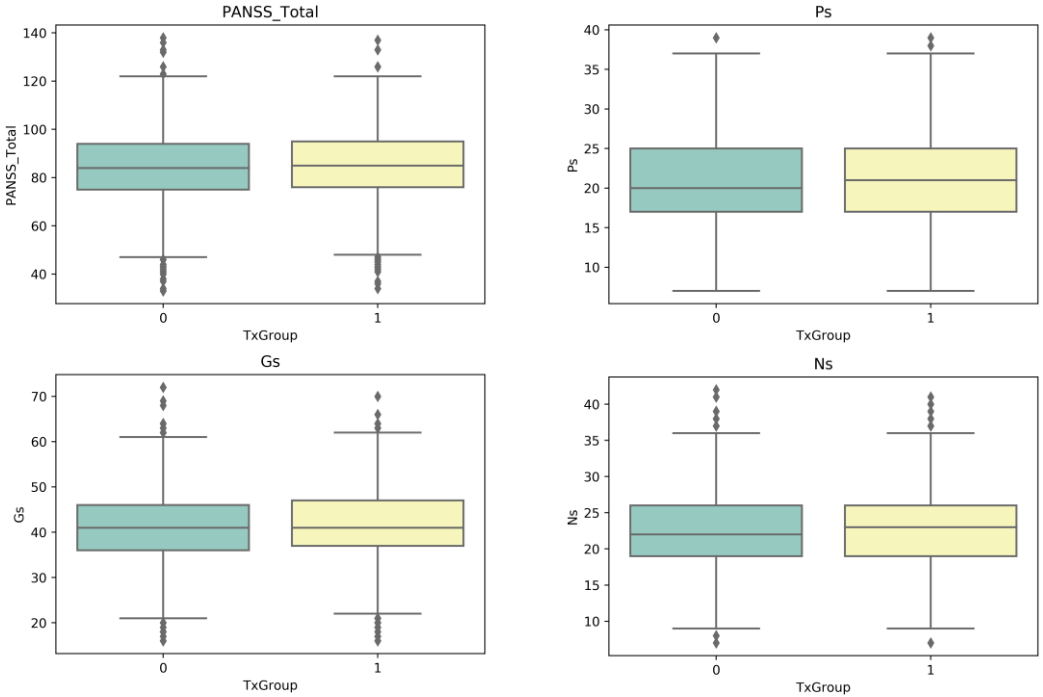
5 datasets with 5 trialss (studyA-E)



Study	Country	PatientID	SiteID	RaterID	AssessmentiD	TxGroup	VisitDay	P1	P2	...	G9	G10	G11	G12	G13	G14	G15	G16	PANSS_Total	LeadStatus
A	USA	10001	20035	30076	100679	Control	0	5	5	...	5	3	3	4	3	3	3	5	107	Assign to CS
A	USA	10001	20035	30076	101017	Control	11	5	5	...	5	3	3	4	3	3	3	5	109	Assign to CS
A	USA	10001	20035	30076	102177	Control	18	4	4	...	4	2	2	3	3	2	3	4	91	Passed
A	USA	10001	20035	30076	101533	Control	25	3	3	...	3	2	2	3	3	2	3	4	80	Flagged
A	USA	10001	20035	30076	100930	Control	39	3	3	...	3	2	2	3	3	2	3	4	77	Flagged
A	USA	10001	20035	30076	100471	Control	53	3	3	...	3	2	2	3	3	2	3	4	75	Flagged
A	USA	10001	20035	30076	102347	Control	67	4	2	...	3	2	2	3	3	2	3	4	72	Flagged
A	USA	10002	20011	30016	100597	Control	0	5	5	...	5	2	1	3	3	3	3	5	85	Passed
A	USA	10002	20011	30016	100270	Control	7	5	5	...	5	3	1	3	3	1	3	5	85	Passed
A	USA	10002	20011	30016	101211	Control	9	5	5	...	5	3	1	3	3	1	3	5	94	Passed
A	USA	10003	20031	30058	101799	Treatment	0	5	5	...	5	3	3	4	3	3	3	4	97	Flagged
A	USA	10003	20031	30058	100330	Treatment	11	6	5	...	6	4	4	4	4	4	5	5	128	Flagged
A	USA	10003	20031	30058	101749	Treatment	18	6	5	...	6	4	4	3	4	3	5	5	126	Flagged
A	USA	10003	20031	30058	101301	Treatment	25	5	5	...	5	4	4	4	4	4	4	5	119	Flagged
A	USA	10003	20031	30058	101615	Treatment	39	4	4	...	4	4	4	3	3	4	4	4	101	Flagged

Stakeholder would like to know –
Does the treatment has an effect on the disease?

Before the analysis : Ensure the treatment and control groups are substantially the same before the study begin – look at the summary statistic for the treatment and control groups

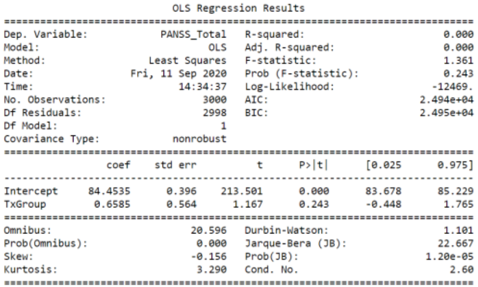


The median value is slightly different for Ps, Gs, Ns

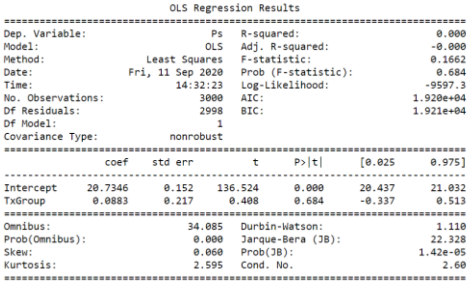
→ the difference between the current score and the previous one for the same patient)

Stakeholder would like to know –
Does the treatment has an effect on the disease?

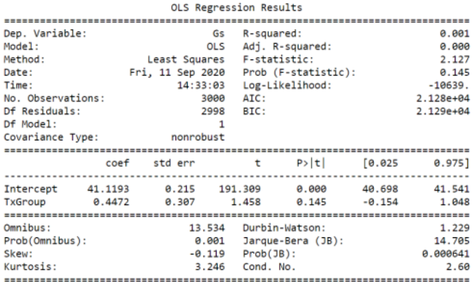
Before the analysis :Ensure the treatment and control groups are substantially the same before the study begin – look at the summary statistic for the treatment and control groups



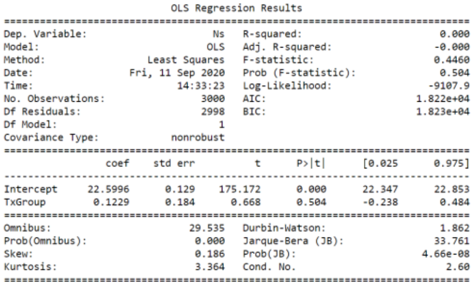
```
                            OLS Regression Results
==============================================================================
Dep. Variable:            PANSS_Total   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                  0.000
Method:                 Least Squares   F-statistic:                     1.361
Date:                Fri, 11 Sep 2020   Prob (F-statistic):              0.243
Time:                        14:34:37   Log-Likelihood:                -12469.
No. Observations:                3000   AIC:                         2.494e+04
Df Residuals:                    2998   BIC:                         2.495e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     84.4535      0.396    213.501      0.000      83.678      85.229
TxGroup        0.6585      0.564      1.167      0.243      -0.448       1.765
==============================================================================
Omnibus:                       20.596   Durbin-Watson:                   1.101
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               22.667
Skew:                          -0.156   Prob(JB):                     1.20e-05
Kurtosis:                       3.290   Cond. No.                         2.60
==============================================================================
```



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     Ps   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                 -0.000
Method:                 Least Squares   F-statistic:                    0.1662
Date:                Fri, 11 Sep 2020   Prob (F-statistic):              0.684
Time:                        14:32:23   Log-Likelihood:                -9597.3
No. Observations:                3000   AIC:                         1.920e+04
Df Residuals:                    2998   BIC:                         1.921e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     20.7346      0.152    136.524      0.000      20.437      21.032
TxGroup        0.0883      0.217      0.408      0.684      -0.337       0.513
==============================================================================
Omnibus:                       34.085   Durbin-Watson:                   1.110
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               22.328
Skew:                           0.060   Prob(JB):                     1.42e-05
Kurtosis:                       2.595   Cond. No.                         2.60
==============================================================================
```



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     Gs   R-squared:                       0.001
Model:                            OLS   Adj. R-squared:                  0.000
Method:                 Least Squares   F-statistic:                     2.127
Date:                Fri, 11 Sep 2020   Prob (F-statistic):              0.145
Time:                        14:33:03   Log-Likelihood:                -10639.
No. Observations:                3000   AIC:                         2.128e+04
Df Residuals:                    2998   BIC:                         2.129e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     41.1193      0.215    191.309      0.000      40.698      41.541
TxGroup        0.4472      0.307      1.458      0.145      -0.154       1.048
==============================================================================
Omnibus:                       13.534   Durbin-Watson:                   1.229
Prob(Omnibus):                  0.001   Jarque-Bera (JB):               14.705
Skew:                          -0.119   Prob(JB):                     0.000641
Kurtosis:                       3.246   Cond. No.                         2.60
==============================================================================
```

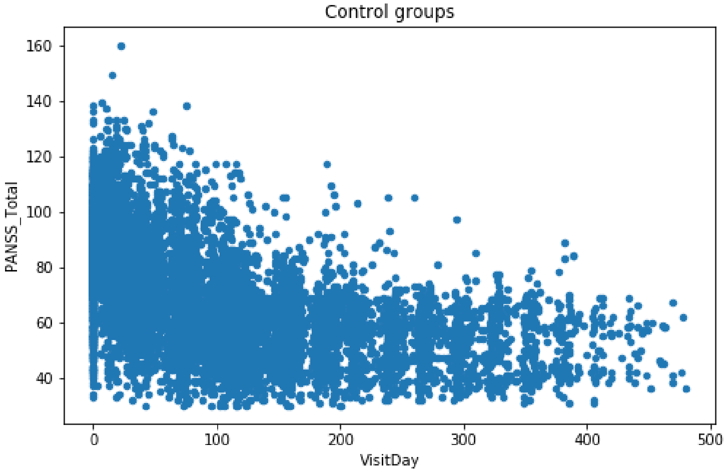


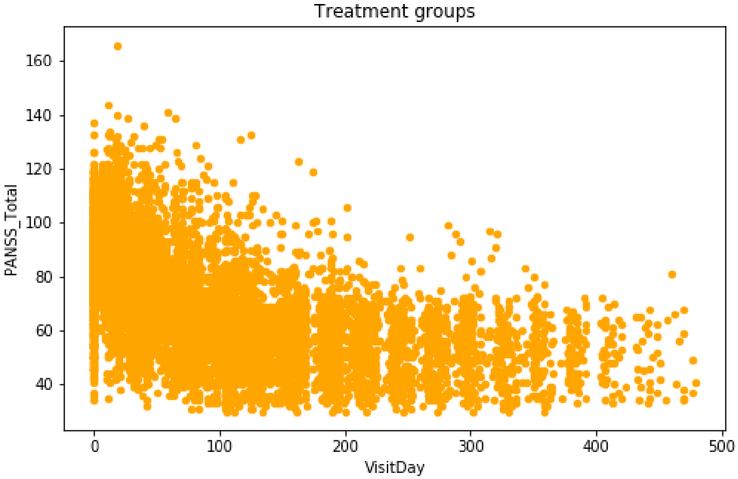
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     Ns   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                 -0.000
Method:                 Least Squares   F-statistic:                    0.4460
Date:                Fri, 11 Sep 2020   Prob (F-statistic):              0.504
Time:                        14:33:23   Log-Likelihood:                -9107.9
No. Observations:                3000   AIC:                         1.822e+04
Df Residuals:                    2998   BIC:                         1.823e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     22.5996      0.129    175.172      0.000      22.347      22.853
TxGroup        0.1229      0.184      0.668      0.504      -0.238       0.484
==============================================================================
Omnibus:                       29.535   Durbin-Watson:                   1.862
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               33.761
Skew:                           0.186   Prob(JB):                     4.66e-08
Kurtosis:                       3.364   Cond. No.                         2.60
==============================================================================
```

Null hypothesis test:
the difference between the control and treatment groups is not statistically significant

Stakeholder would like to know –
Dose the treatment has an effect on the disease?

Data visualization: Control groups v.s treatment groups: The plot looks similar





Stakeholder would like to know – Dose the treatment has an effect on the disease?

Original analysis : Used PANSS_Total_diff (the difference between the current PANSS_total score and the previous one for the same patient) to form regression against Visitday and treatment

Because the PANSS_total score varied from patient to patient, this ensures that we are evaluating the effects of the treatment, on the changes in the PANSS core.

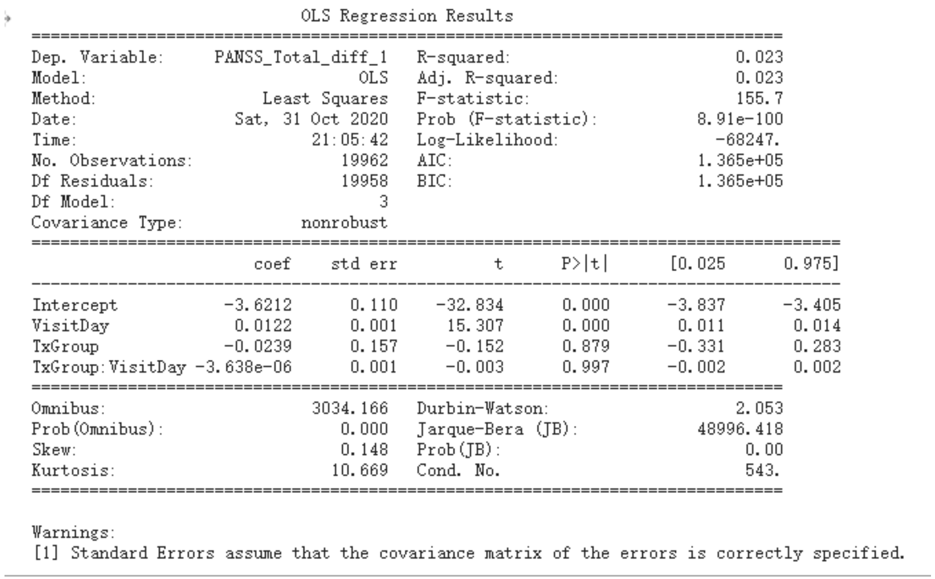
The linear regression is:
PANSS_Total_diff=$\beta 0+\beta 1 \textit{TxGroup}+\beta 2 \textit{VisitDay} + \beta 3 \textit{TxGroup}*\textit{VisitDay} + \epsilon$

TxGroup =1 (Treatment) PANSS_Total_diff =$\beta 0+\beta 1+\beta 2 \textit{VisitDay} + \beta 3 \textit{VisitDay} + \epsilon$
$= \beta 0+\beta 1 + (\beta 2+ \beta 3)\textit{VisitDay} + \epsilon$
TxGroup =0 (Control) PANSS_Total_diff =$\beta 0+\beta 2 \textit{VisitDay} + \epsilon$

Null-hypothesis: $\beta 3=0$

If the p-value for *TxGroup***VisitDay* is < 0.05: statistical significance

Stakeholder would like to know –
Does the treatment has an effect on the disease?



```
                            OLS Regression Results
==============================================================================
Dep. Variable:      PANSS_Total_diff_1   R-squared:                       0.023
Model:                             OLS   Adj. R-squared:                  0.023
Method:                  Least Squares   F-statistic:                     155.7
Date:                 Sat, 31 Oct 2020   Prob (F-statistic):          8.91e-100
Time:                         21:05:42   Log-Likelihood:                -68247.
No. Observations:                19962   AIC:                         1.365e+05
Df Residuals:                    19958   BIC:                         1.365e+05
Df Model:                            3
Covariance Type:             nonrobust
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
Intercept           -3.6212      0.110    -32.834      0.000      -3.837      -3.405
VisitDay             0.0122      0.001     15.307      0.000       0.011       0.014
TxGroup             -0.0239      0.157     -0.152      0.879      -0.331       0.283
TxGroup:VisitDay -3.638e-06      0.001     -0.003      0.997      -0.002       0.002
==============================================================================
Omnibus:                     3034.166   Durbin-Watson:                   2.053
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            48996.418
Skew:                           0.148   Prob(JB):                         0.00
Kurtosis:                      10.669   Cond. No.                         543.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

p-value for *TxGroup*VisitDay* = 0.997:

not statistical significance

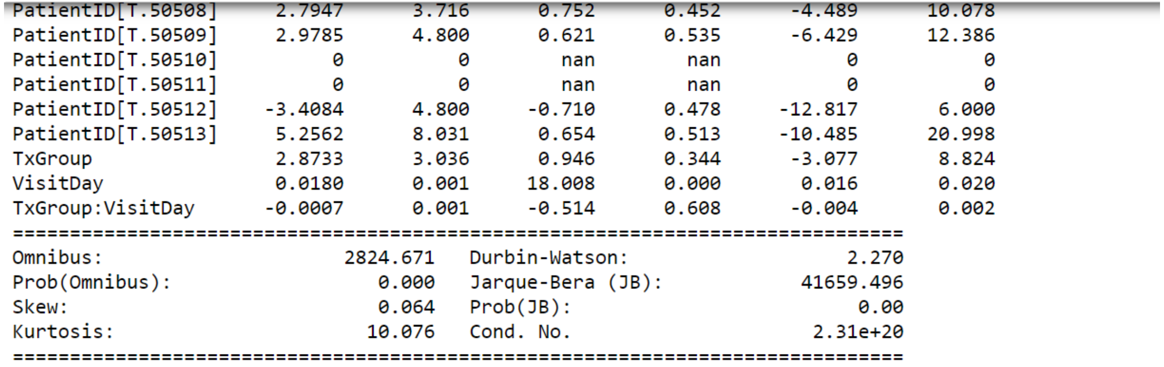
Stakeholder would like to know –
Does the treatment has an effect on the disease?

Analysis1. analysis 1 + added categorical variables for different studies to the regression

- In analysis 1, Different studies (trials) have different patient population. For example, the patient from study A is all from USA and the patient from study E is from China → Different studies may have bias result to the analysis

The linear regression become:

$PANSS\_Total\_diff = \beta0+\beta1 \mathit{TxGroup}+\beta2 \mathit{VisitDay} +\beta3 \mathit{TxGroup}*\mathit{VisitDay} + \beta4 \mathit{PatientID} +\epsilon$



```
PatientID[T.50508]      2.7947      3.716      0.752      0.452      -4.489      10.078
PatientID[T.50509]      2.9785      4.800      0.621      0.535      -6.429      12.386
PatientID[T.50510]           0          0        nan        nan           0           0
PatientID[T.50511]           0          0        nan        nan           0           0
PatientID[T.50512]     -3.4084      4.800     -0.710      0.478     -12.817       6.000
PatientID[T.50513]      5.2562      8.031      0.654      0.513     -10.485      20.998
TxGroup                 2.8733      3.036      0.946      0.344      -3.077       8.824
VisitDay                0.0180      0.001     18.008      0.000       0.016       0.020
TxGroup:VisitDay       -0.0007      0.001     -0.514      0.608      -0.004       0.002
==============================================================================
Omnibus:                     2824.671   Durbin-Watson:                   2.270
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            41659.496
Skew:                           0.064   Prob(JB):                         0.00
Kurtosis:                      10.076   Cond. No.                     2.31e+20
==============================================================================
```

The p value for TxGroup variable is 0.351 → Fail to reject the null hypothesis

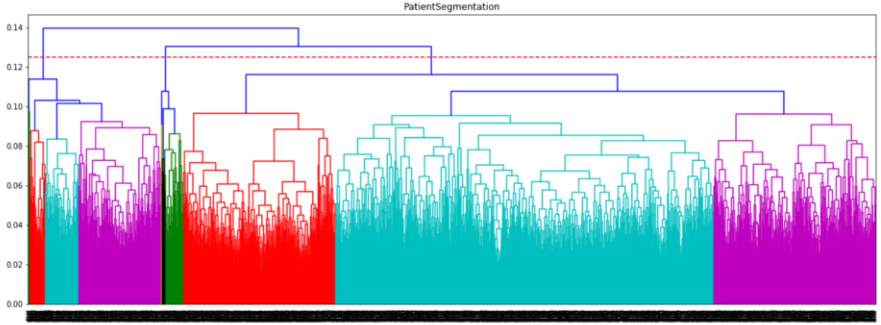
Patient segmentation – understand each groups oh patients for their mental health status

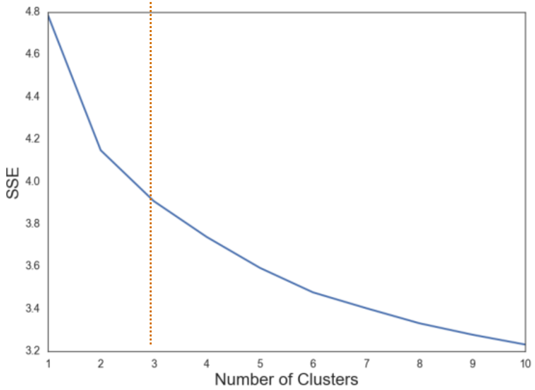
Data preprocess : data is first normalized (The distance measures are affected by the scale of the variables)

Data mining/evaluation: complete linkage clustering/ K-means clustering

- The red dotted line indicates the level at which the data is segmented. The dissimilarity between the three clusters created by seat the dotted line is greater than the ones that would have been created by cutting at a higher level (for two segments). Sub-segments below that are relatively similar with the other segments in their respective groups.

- Interpretable



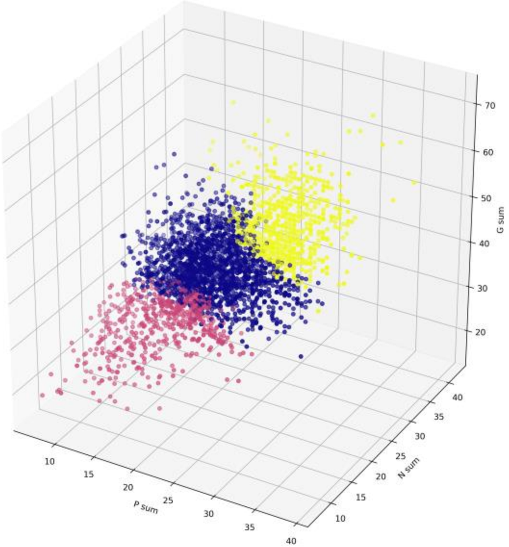


Patient segmentation – understand each groups oh patients for their mental health status

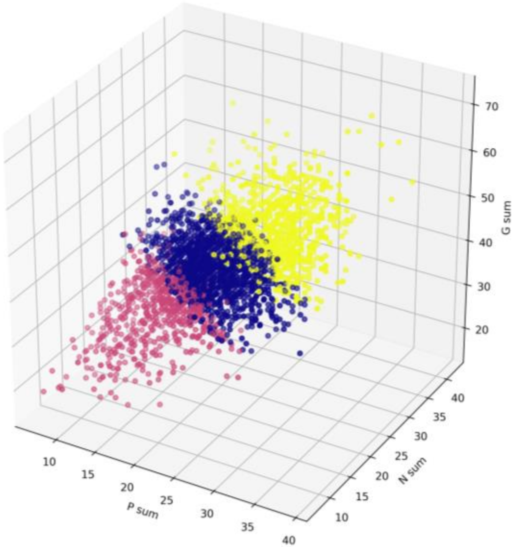
Data preprocess : data is first normalized (The distance measures are affected by the scale of the variables)

Data mining/evaluation: complete linkage clustering and K-means clustering (k=3)

complete linkage clustering



K-means clustering



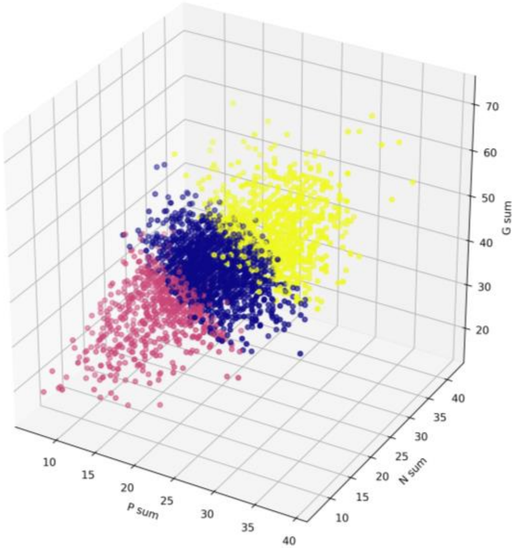
K-means clustering shows better separation of the individual clusters from each other.

Patient segmentation – understand each groups oh patients for their mental health status

Data preprocess : data is first normalized (The distance measures are affected by the scale of the variables)

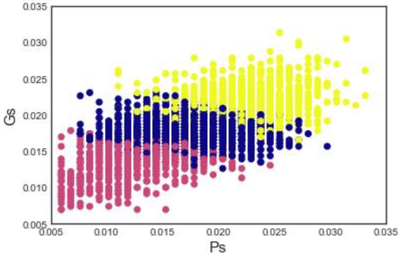
Data mining/evaluation: complete linkage clustering and K-means clustering (k=3)

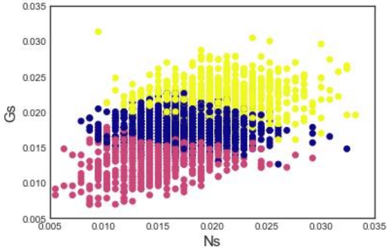
K-means clustering



1. Each cluster indicates a different combination of the sub-group scores. Moving from he ‘Pink’ cluster to the ‘Blue’ cluster requires an increase in all three sub-group scores.

2. To move from ‘blue’ cluster to ‘yellow’ cluster only requires a relatively smaller increase in the ‘G’ subgroup sum, whereas relatively similar changes in ‘Ps’ sum or ‘Ns’ sum doesn’t result in crossing over to the next cluster.


Gs
Ps


Gs
Ns