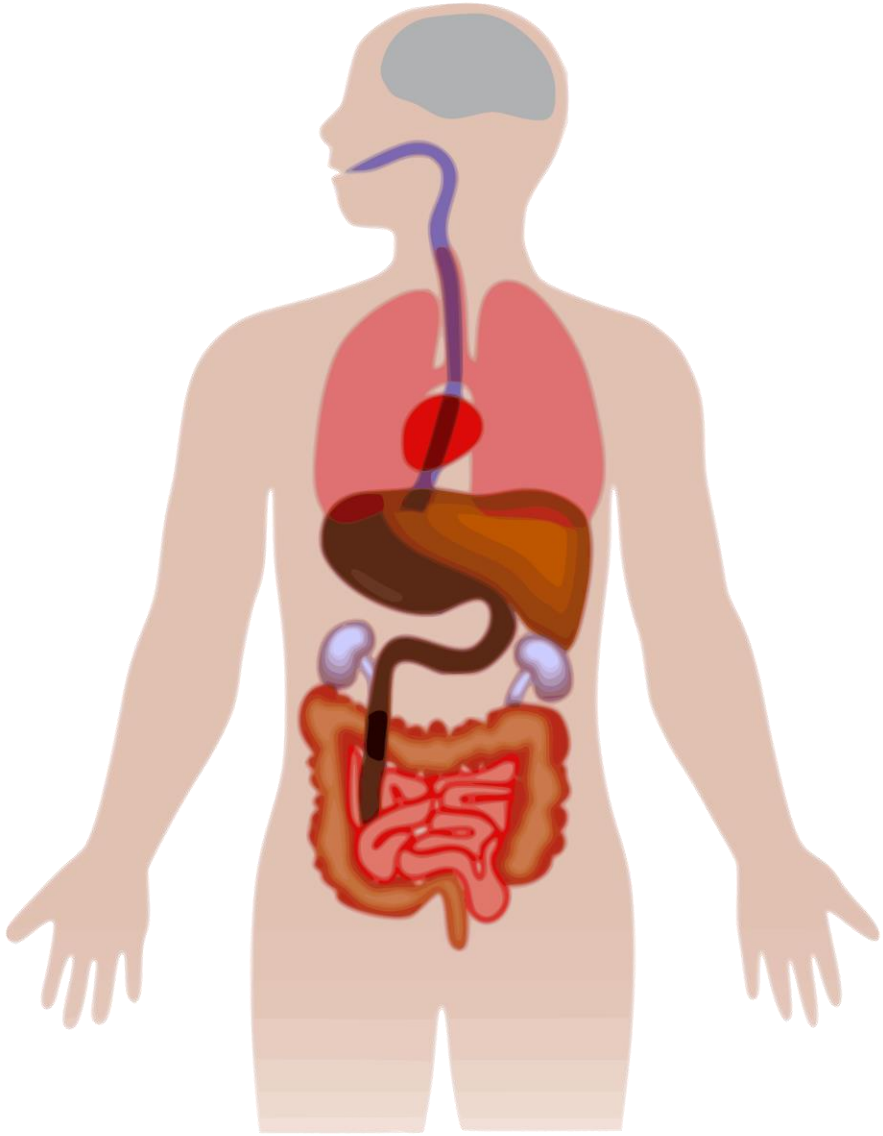
STT 303 Final Project

Chris Day, Samuel Bryan, Enleo Dahal

Agenda

- Data Explanation
- Cleaning
- Modeling
  - Regression
  - Classification
  - Clustering
- Interpretation
- Conclusion

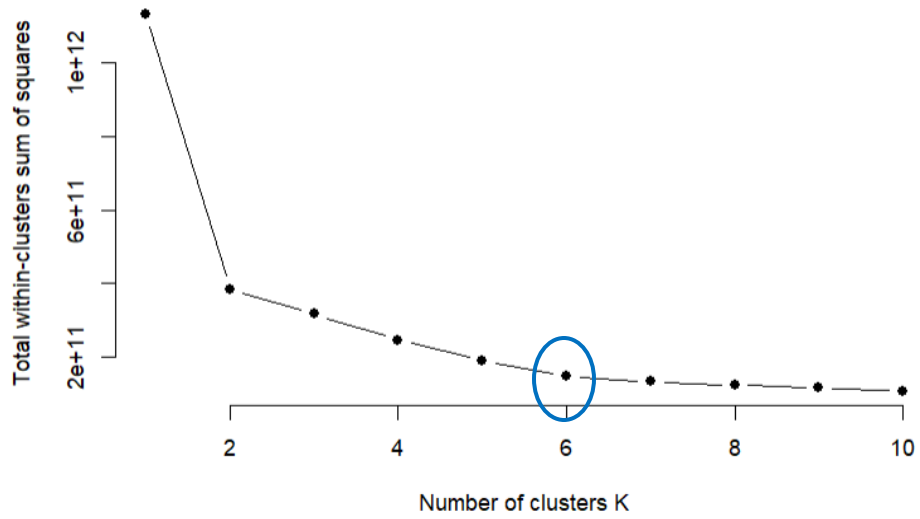


Dataset

- Public study of Genotype Tissue Expression
- Sample of gene expression levels
- Principal Component Analysis
  - Summarize Variation
- Median Gene Level Dataset

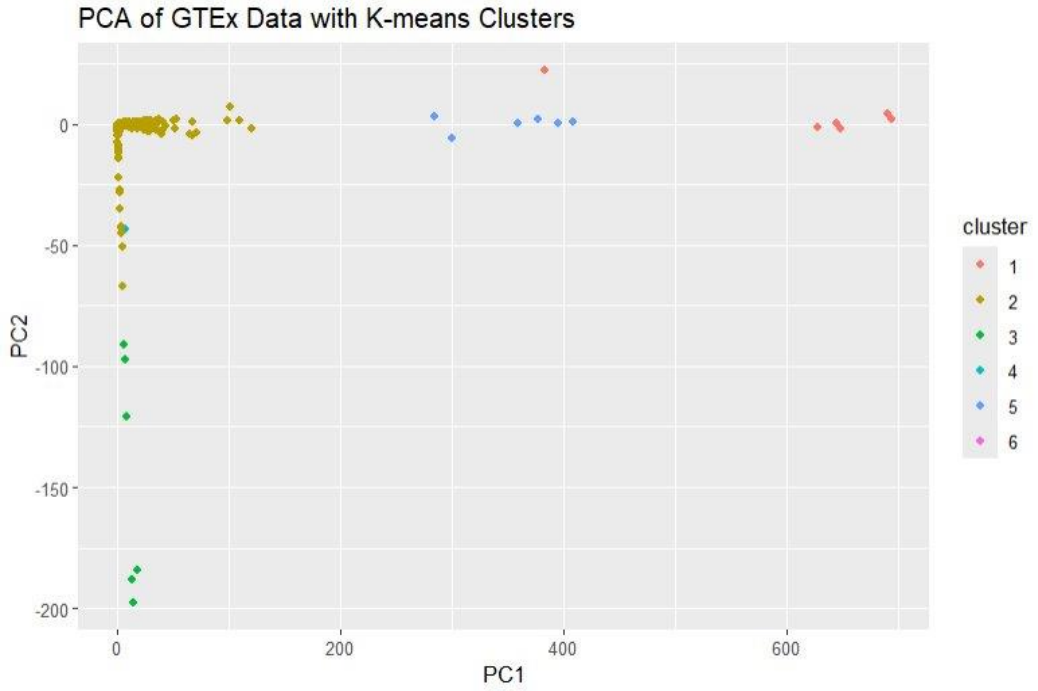
K-means Clustering

- To find the K-means of the data, we needed to determine how many clusters were needed.
- This was done by performing the elbow method above by seeing where the error was at its lowest point without overfitting.
- This is what is referred to as the “elbow point”, representing the best amount of clusters to start with for our data.



Cluster Visualization

- The result of the K-means clusters was created by comparing PC1 and PC2 variables of the GTEX data to one another.
- With these clusters, we are able to tell where part of the data aggregates.
- Data demonstrates that the clusters 1 and 6 represent outliers in the clustering data.

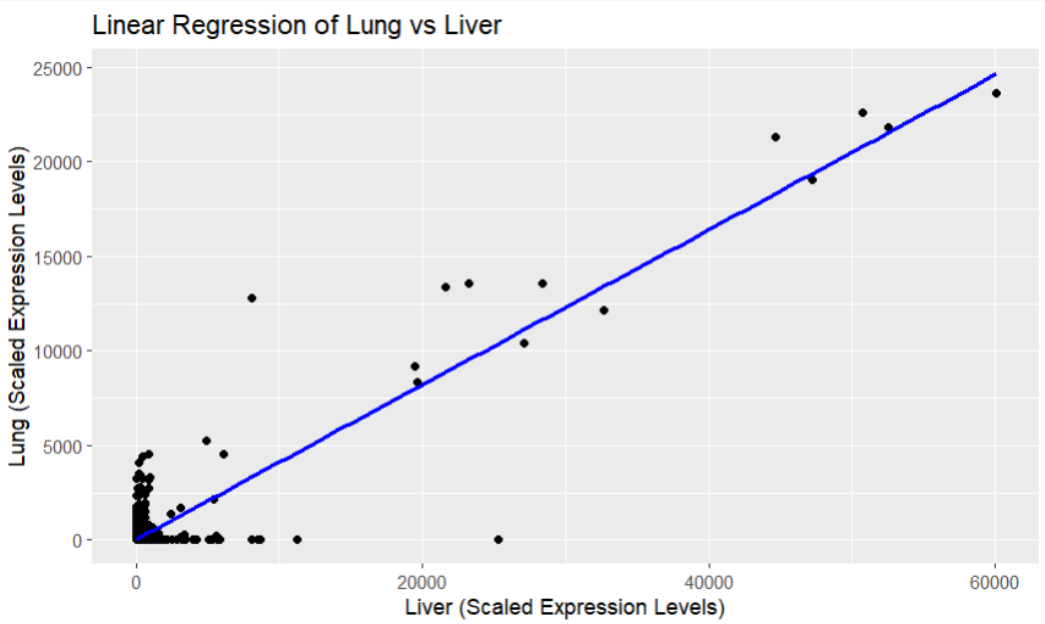


Regression: Lung vs Liver

-Strong positive linear relationship at r = 0.91

-Numerous symmetrical points near 0 and few outliers factor into this strong coefficient.

-The RMSE (Root Mean Squared Error) measures the average magnitude of differences between the predicted and actual values. The RMSE here is 108 units.

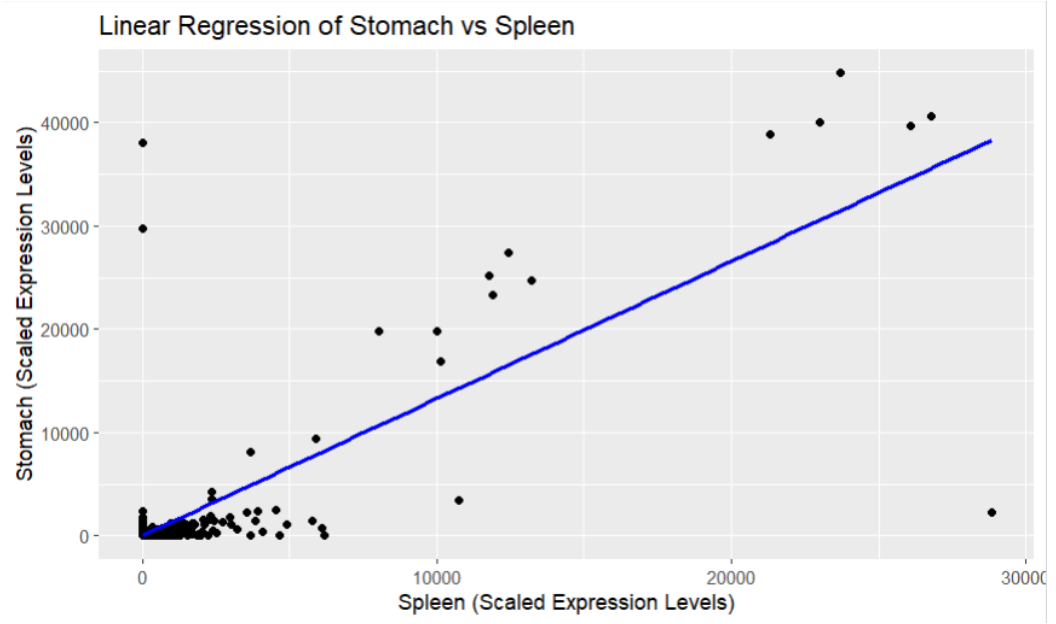


Regression: Stomach vs Spleen

-Strong positive linear relationship at r = 0.8.

-An unsymmetrical cluster around 0 and the presence of more outliers drag this correlation coefficient down.

-The RMSE here is 297 units. This is nearly triple that of Lung vs Liver.

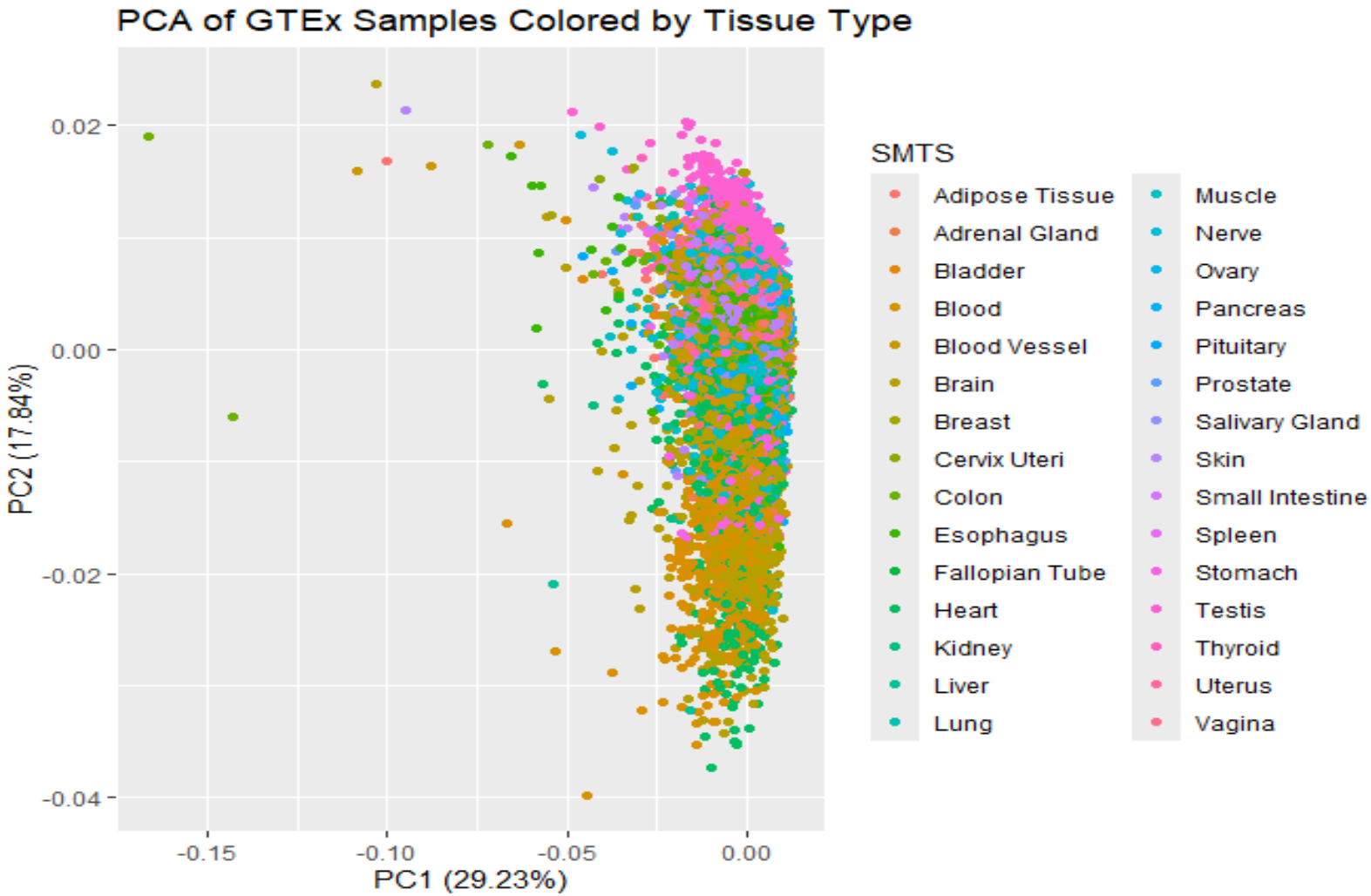


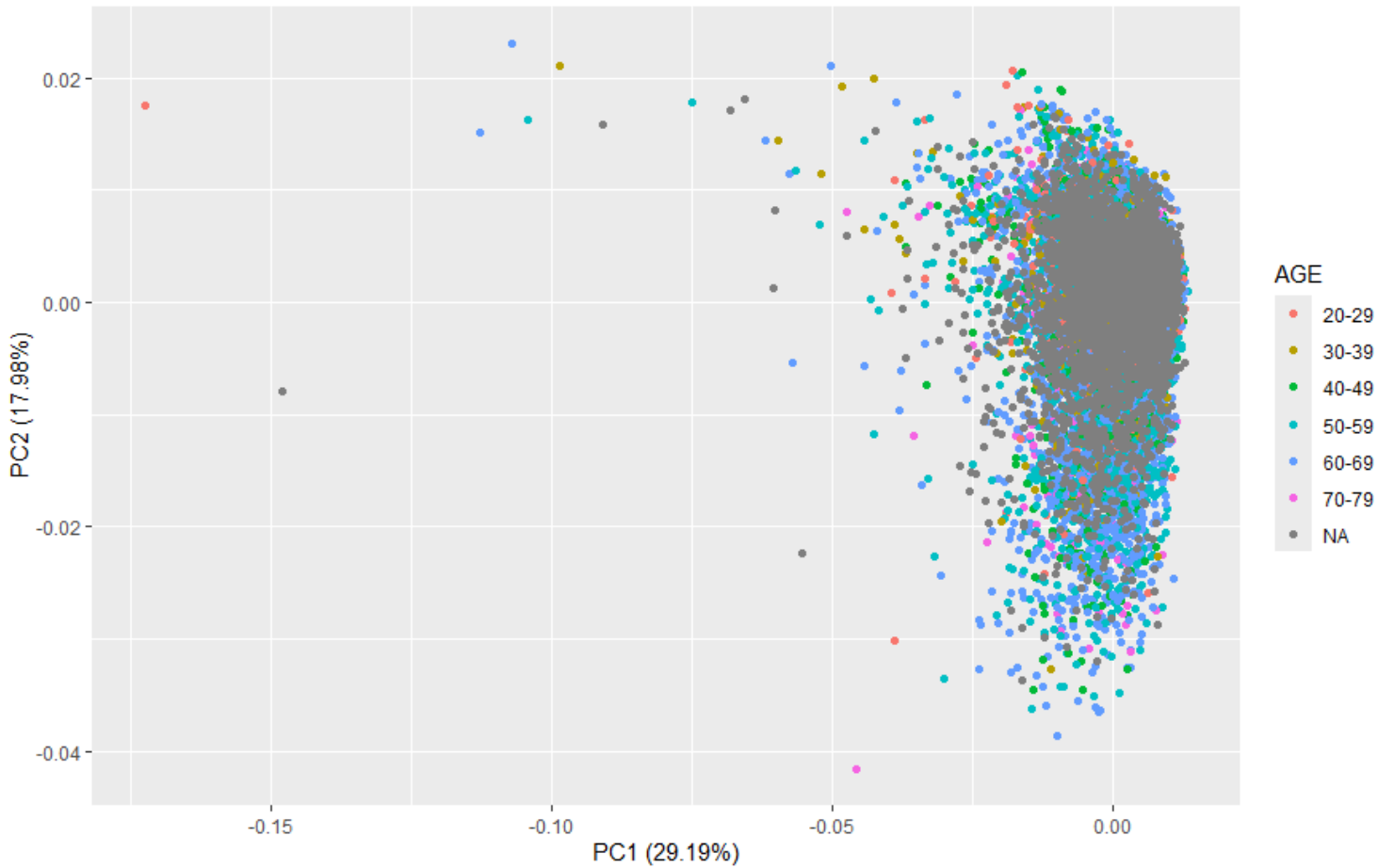
Regression Final Thoughts

-While the regression models imply that there is a significant relationship between the sets of tissues, the high RSME's show that there is more to this relationship beyond the regression.

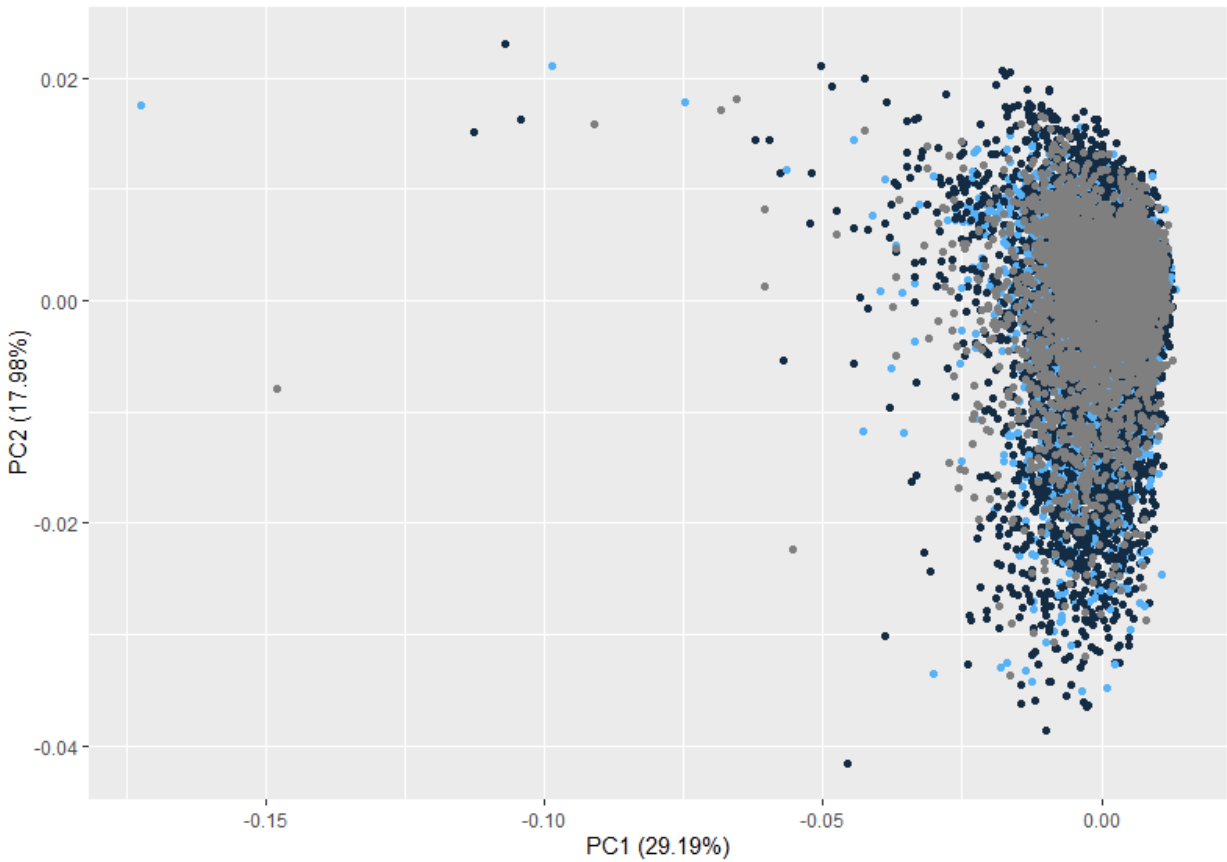
-The high scores of 108 and 297 units for the respective models are evidence that the regressions do not tell the whole stories.

	Adipose Tissue	Adrenal Gland	Bladder	Blood	Blood Vessel	Brain	Breast	Cervix Uteri	Colon	Esophagus	Fallopian Tube
1	331	73	42	272	314	704	126	18	241	328	14
2	11	37	0	511	14	1478	7	1	14	61	0
3	839	159	35	250	982	874	332	27	563	1057	15






```
  20-29 30-39 40-49 50-59 60-69 70-79
1   306   329   463  1096  1256   151
2    61    96   221   703  1177   122
3   607   707  1056  2537  3032   359
> 
```



	1	2
1	2494	1107
2	1752	628
3	5650	2648

```
> table(full_data$Cluster, full_data$SMTS)

  Adipose Tissue Adrenal Gland Bladder Blood Blood Vessel Brain Breast Cervix Uteri Colon Esophagus Fallopian Tube
1            331            73      42   272          314   704    126           18   241       328            14
2             11            37       0   511           14  1478      7            1    14        61             0
3            839           159      35   250          982   874    332           27   563      1057            15

  Heart Kidney Liver Lung Muscle Nerve Ovary Pancreas Pituitary Prostate Salivary Gland Skin Small Intestine Spleen
1   209      25    36  166    285   169    42       78        75       57             47  542              49     50
2   311      49   101    6     37     3     0       34         5        9              0   14               2      3
3   313      33   108  383    430   451   137      214       223      198            118 1335             123    200

  Stomach Testis Thyroid Uterus Vagina
1      68    173     159     31     32
2     111      0       7      0      1
3     189    210     464    109    118
>
```