

Gene Expression Analysis with R

presented by Amina Aboulhana

Project Overview



Data

Used GTEx RNA-seq data across human tissues

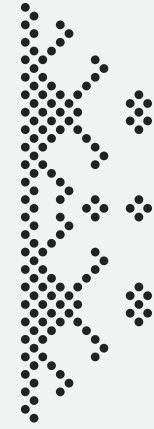
Goal

Explore gene expression by tissue, sex and age

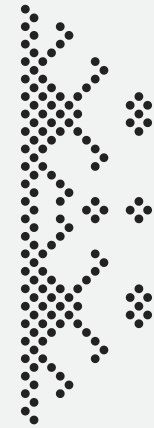
Applied R-based methods

PCA, clustering, and regression

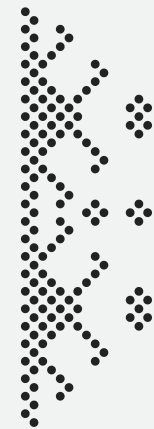
Why GTEx?



Public dataset with rich metadata

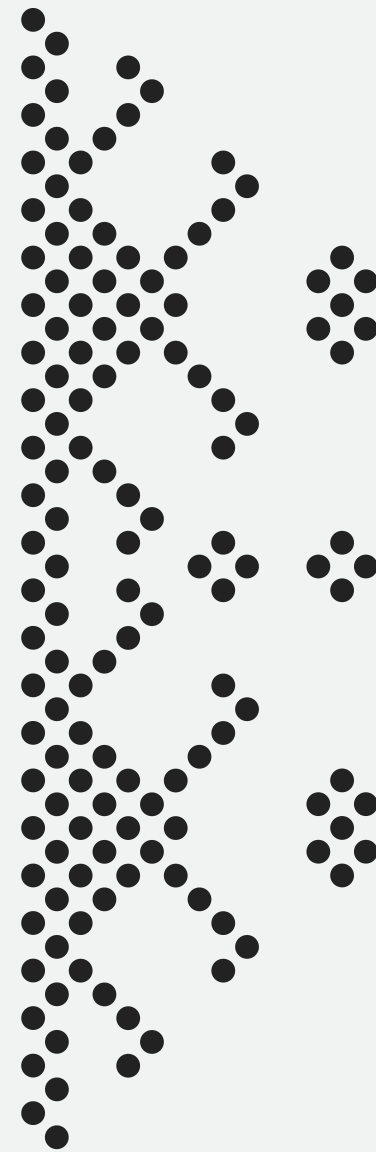


Covers a wide range of tissues and conditions



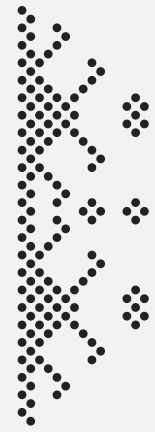
Great opportunity to apply statistical learning

Initial Plan



Step by Step

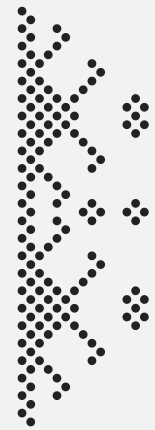
- Load and clean data
- Reduce dimensionality using PCA
- Use random forest to predict age
- Explore gene clusters with heatmaps



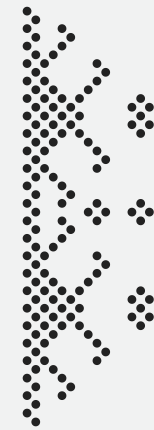
Selected 10,000 genes



**Converted characters
to numeric values**



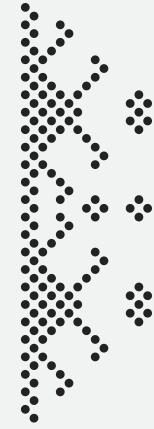
**Applied log transformation
(log1p)**



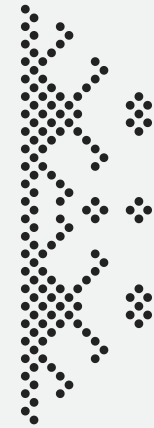
**Filtered low-variance
genes**

Data Setup

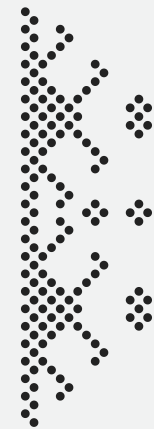
Exploratory Goals



**Visualize expression variation by
sex/tissue (PCA)**

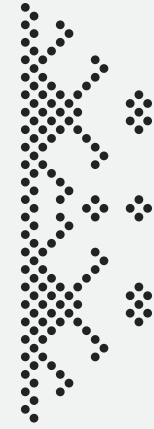


**Identify co-expressed gene
clusters (heatmaps)**

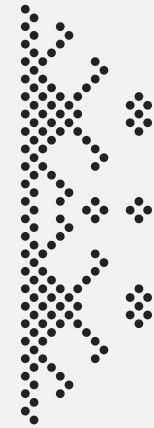


**Predict age using gene expression
(regression)**

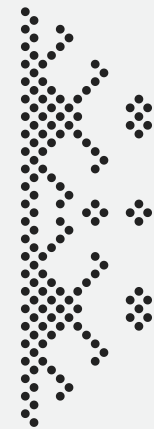
What Worked



Successfully merged gene and metadata

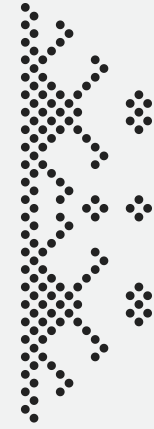


Cleaned, filtered, and transformed expression data

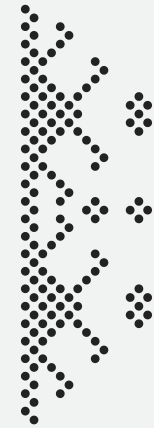


Built code to safely handle missing/incomplete values

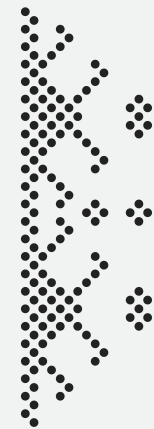
What Went Wrong



PCA failed due to too few variable genes

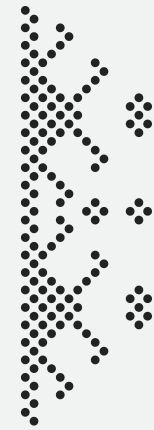


Increasing genes led to memory crashes

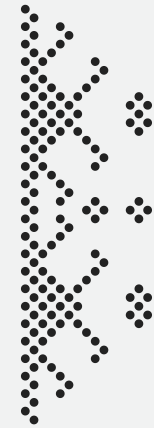


Regression failed due to insufficient valid predictors

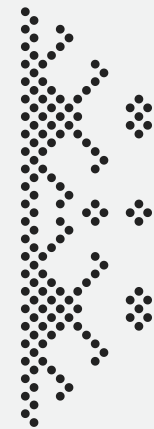
Technical Barriers



Memory overload from high-dimensional data

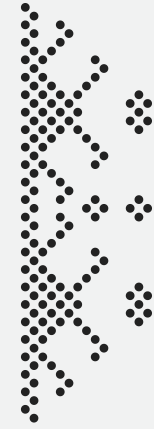


Hidden formatting issues in numeric fields

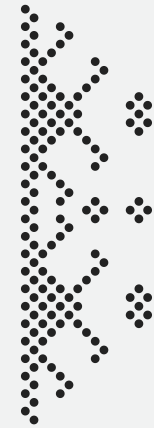


Functions like `prcomp()` and `randomForest()` broke

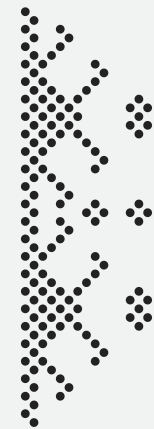
Outside Help



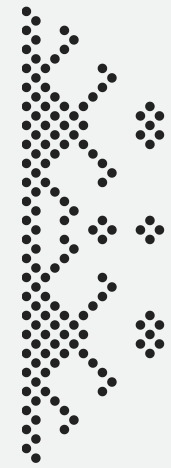
**Worked with programming
professional**



**Debugged code and reviewed
data structure**



**Concluded system couldn't
retain enough significant
variables**



Narrow scope early



**Use DelayedArray
or cloud tools**



**Validate structure and
values before modeling**



**Modular, flexible
code is critical**

Lessons Learned: Looking Forward

Final Takeaway



While the analysis did not meet my original goals, the experience was valuable. i learned to adapt, troubleshoot real-world issues, and better understand the demands of biological data science.