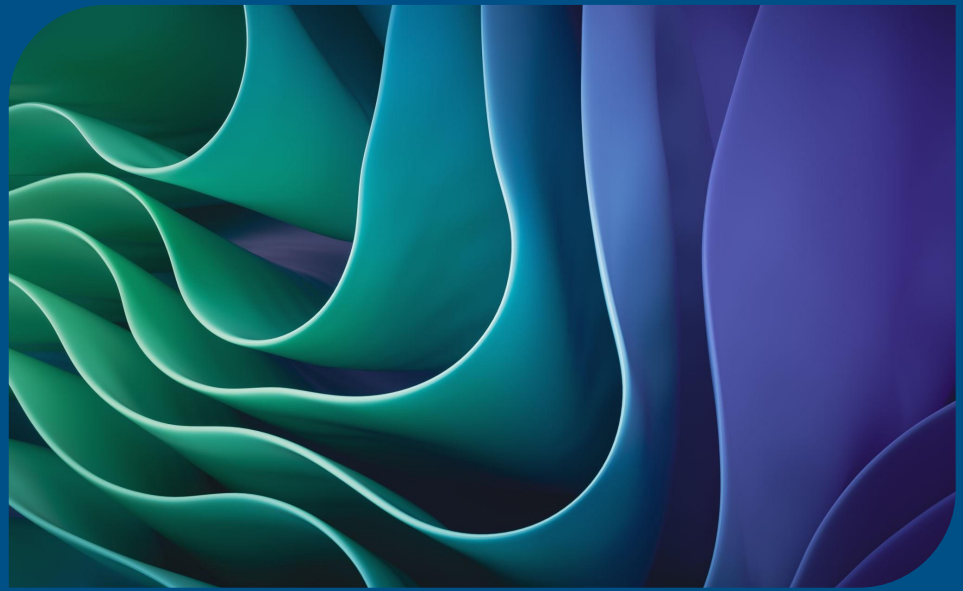
Spring, 2025

Analyzing tissue-specific gene expression from GTEx through linear regression, logistic regression, and PCA k-means clustering

Henry Lock



STT 303 Final Project

1. **Project Overview**
2. Data sources and preparation
3. Linear Regression
4. Logistic Regression Classification
5. PCA and k-means clustering
6. Biological relevance
7. Conclusions

1. Project Overview
2. **Data sources and preparation**
3. Linear Regression
4. Logistic Regression Classification
5. PCA and k-means clustering
6. Biological relevance
7. Conclusions

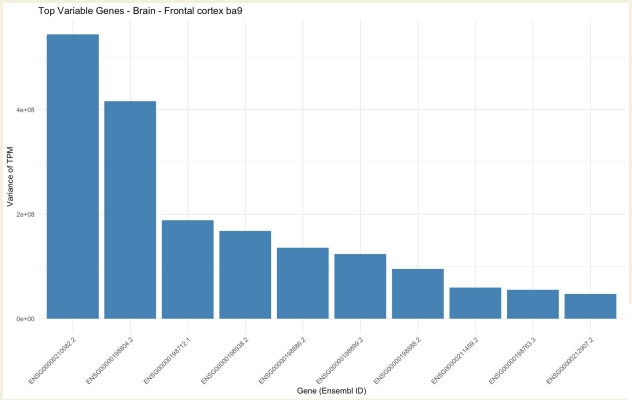
dGTEx

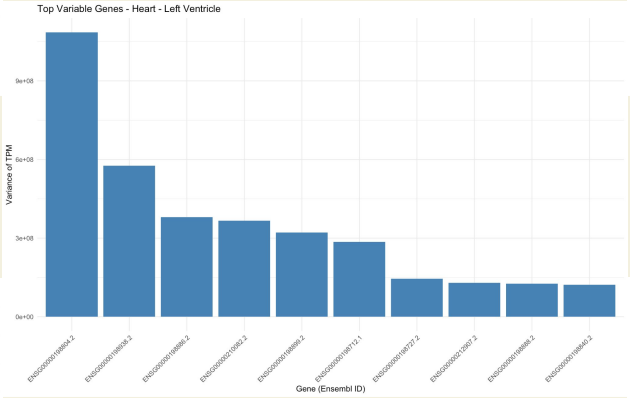
- The Adult Genotype-Tissue Expression (GTEx) project

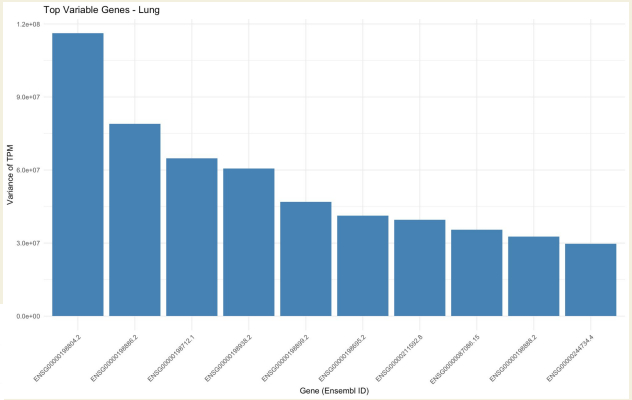
- Gene TPMs by tissue:
  - Brain - Frontal Cortex (BA9)
  - Heart - Left Ventricle
  - Lung
  - Muscle - Skeletal
  - Pituitary

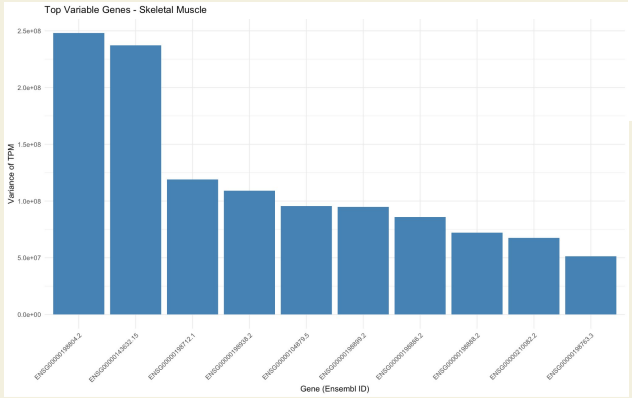
Data sources

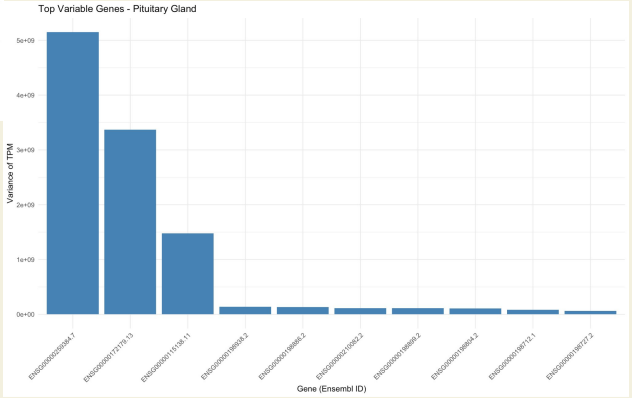
Top 10 variable genes by tissue type







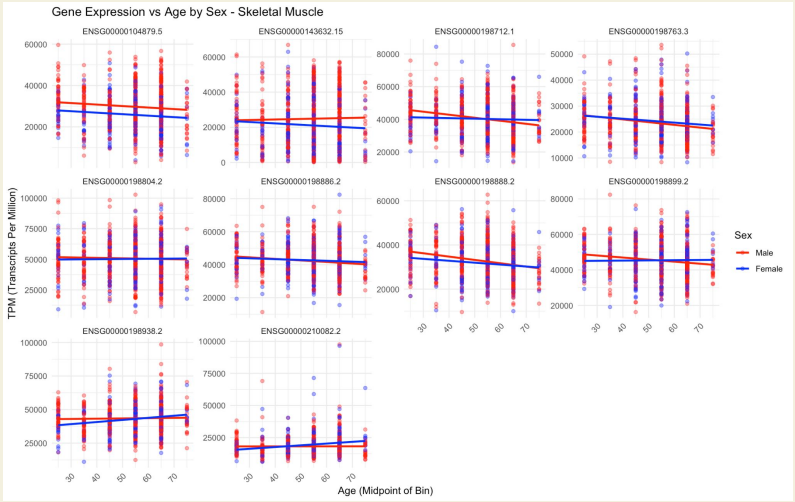


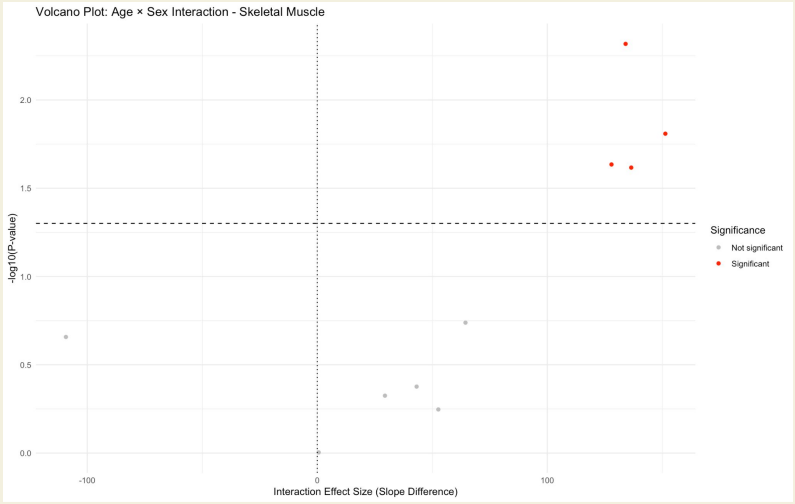


1. Project Overview
2. Data sources and preparation
3. **Linear Regression**
4. Logistic Regression Classification
5. PCA and k-means clustering
6. Biological relevance
7. Conclusions

Linear Regression

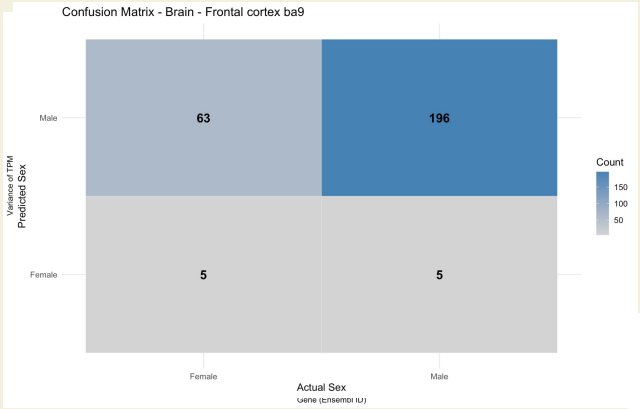
- Gene expression v age by sex
- Compared significance of slope differences using t-test
- Found four significant genes in skeletal muscle tissue

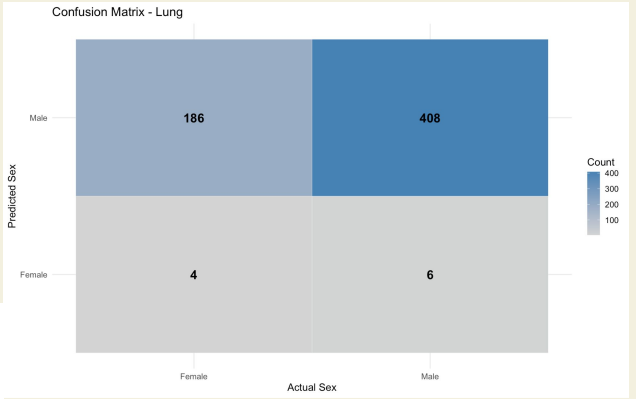


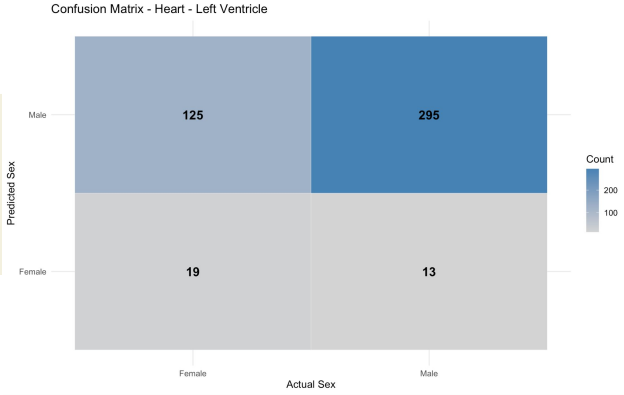


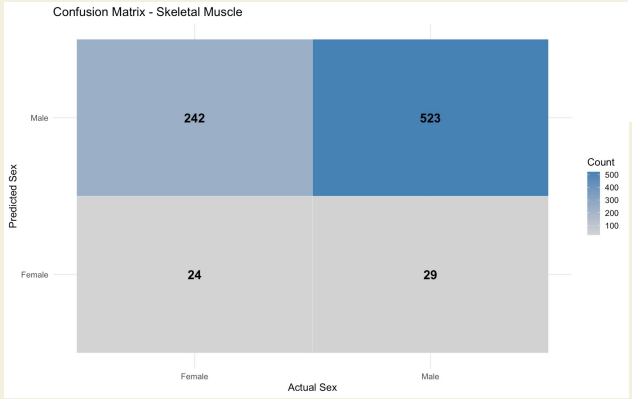
1. Project Overview
2. Data sources and preparation
3. Linear Regression
4. **Logistic Regression Classification**
5. PCA and k-means clustering
6. Biological relevance
7. Conclusions

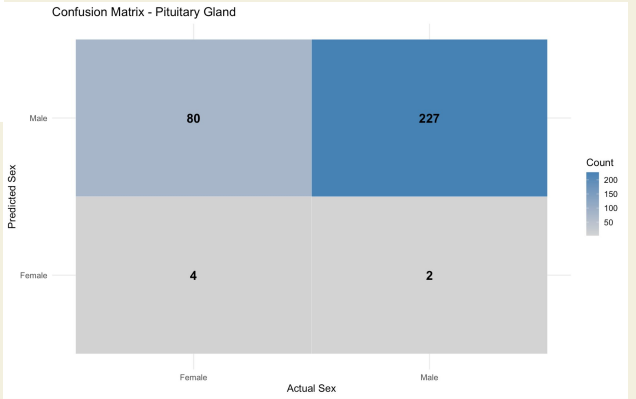
Accuracy of logistic regression classification







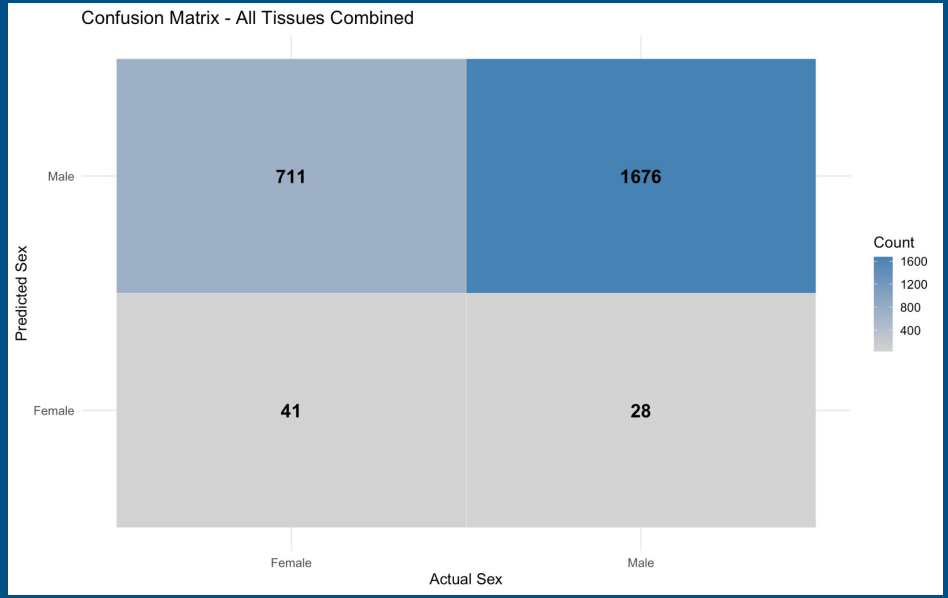




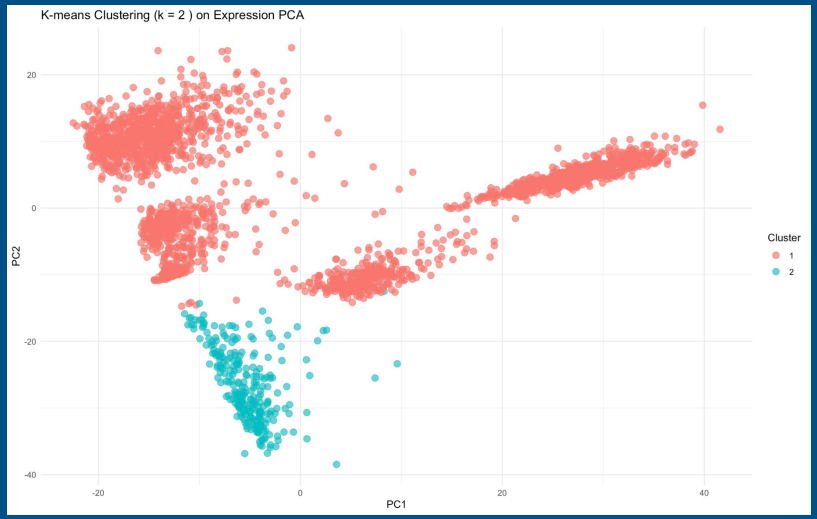
All tissue sample classification

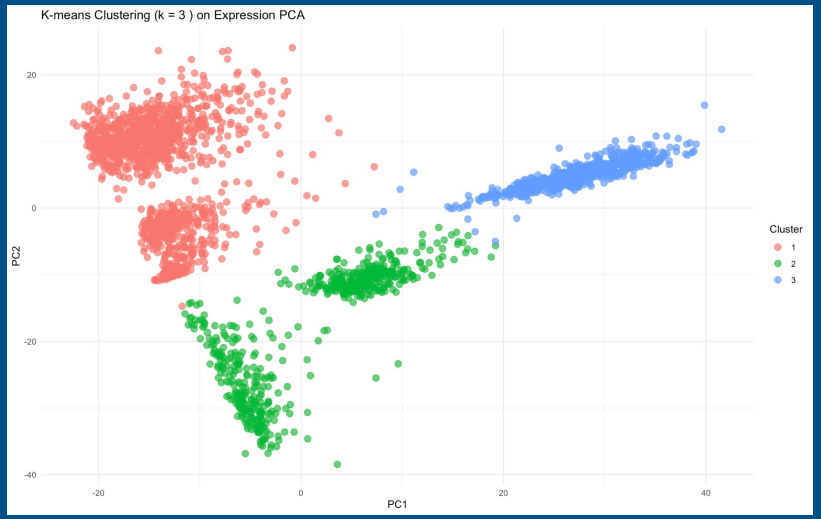
- 1676/2387 male predictions
  - 70.21%

- 41/69 female predictions
  - 59.42%



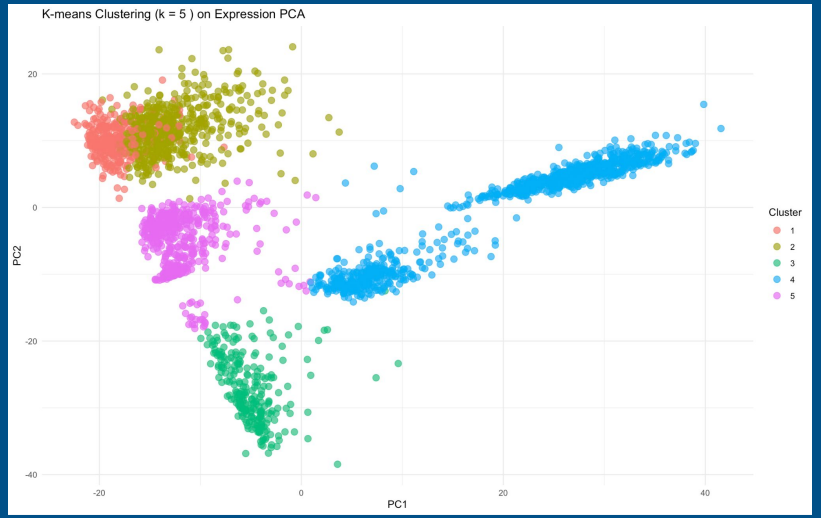
1. Project Overview
2. Data sources and preparation
3. Linear Regression
4. Logistic Regression Classification
5. **PCA and k-means clustering**
6. Biological relevance
7. Conclusions

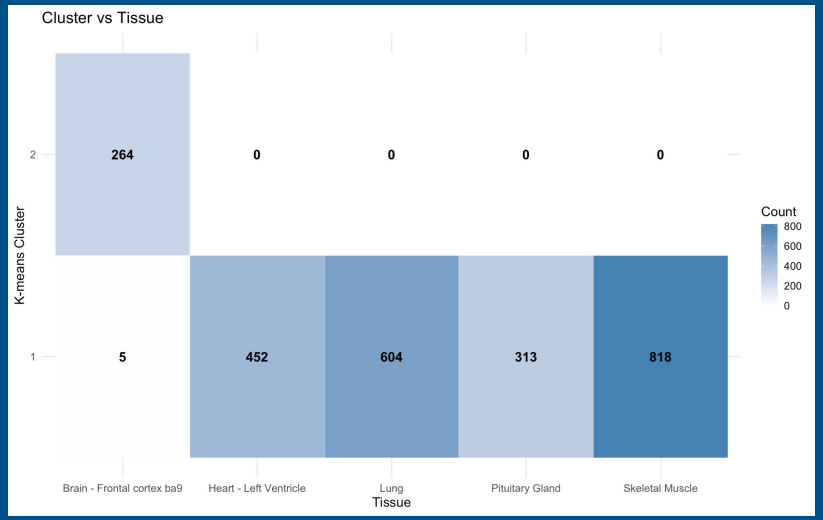


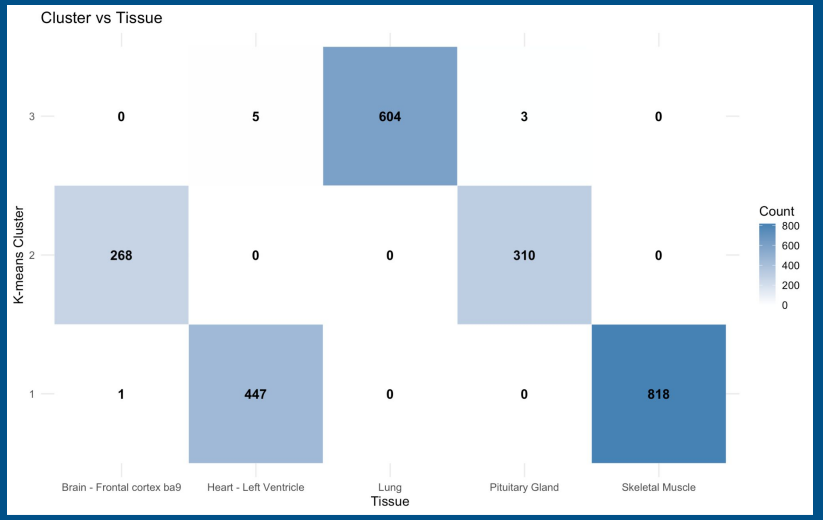


PCA and k-means cluster

- Top 1000 variable genes from all tissues
- Performed for 2, 3, 5 clusters
- Tissue type was the best clustered, age and sex not so much

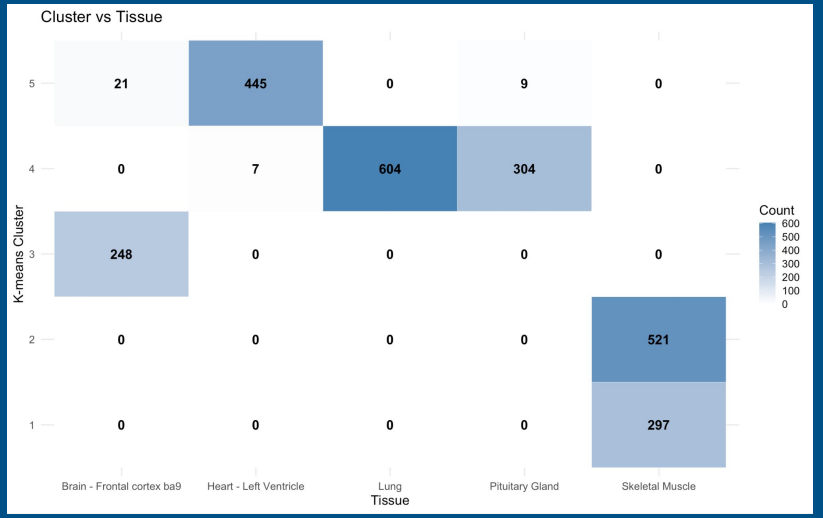






PCA and k-means cluster

- Top 1000 variable genes from all tissues
- Performed for 2, 3, 5 clusters
- Tissue type was the best clustered, age and sex not so much



1. Project Overview
2. Data sources and preparation
3. Linear Regression
4. Logistic Regression Classification
5. PCA and k-means clustering
6. **Biological relevance**
7. Conclusions

Specific gene example for tissue types

- Brain - Frontal Cortex BA9: MT-CO2; Key in ATP production
  - MELAS(Mitochondrial Encephalomyopathy, Lactic Acidosis, and Stroke-like episodes
- MT-ND1 and MT-ND2: Subunits for the NADH dehydrogenase protein (ATP)
  - MELAS
  - Leber Hereditary Optic Neuropathy (LHON)
- Skeletal Muscle: ACTA1; Alpha-skeletal actin. Helps muscles fibers contract
  - Various myopathies
- Pituitary Gland: PRL; Encodes prolactin hormone, lactation signaling and reproduction functions
  - Hyperprolactinemia

1. Project Overview
2. Data sources and preparation
3. Linear Regression
4. Logistic Regression Classification
5. PCA and k-means clustering
6. Biological relevance
7. **Conclusions**

Conclusions

- Linear regression: Sex-specific aging effects for treatment plans and testing
  - Different tissue types

- Logistic regression: Sex-predicting genes linked to disease could reveal mechanism
  - More female samples

- Clustering: Sub-clusters of genes within tissue types could reveal genes at risk to organ-specific diseases
  - Clustering genes within tissue type