Module_3:

Team Members:

Molly Kessenich and Luke Shee

Project Title:

How does patient age affect the expression of genes associated with skin cutaneous melanoma metastasis?

Project Goal:

The goal of this project is to investigate whether patient age influences the risk of metastasis in skin cutaneous melanoma by examining patterns in clinical data from The Cancer Genome Atlas (TCGA) GSE62944 dataset, including its RNA Sequencing data and patient metadata. Specifically, the projects aims to answer "Is patient age associated with patterns of expression in metastasis-related genes in cutaneous melanoma of the skin (SKCM)? Do younger and older patients cluster differently based on the expression of sixteen genes associated with tissue invasion and metastasis?"

Disease Background:

Pick a hallmark to focus on, and figure out what genes you are interested in researching based on that decision. Then fill out the information below.

- Cancer hallmark focus:
  - Tissue invasion and metastasis
- Overview of hallmark:
  - Metastasis or cancer cell invasion/spread to surrounding tissues is the leading cause of cancer deaths and allows cancer to spread beyond its source. This ability of cancer cells to detach from neighboring cells and travel to different parts of the body and develop malignant tumors elsewhere in the body is deadly.
- Genes associated with hallmark to be studied (describe the role of each gene, signaling pathway, or gene set you are going to investigate):
  - Loss of E-cadherin (CDH1) weakens cell-cell junctions and allows melanoma cells to detach from the primary tumor, a key step in metastasis.
  - Increased expression of integrins (such as ITGA5 and ITGB1) enhances cell–ECM adhesion and migration, enabling tumor cells to move through surrounding tissues.
  - Matrix metalloproteinases (MMPs), including MMP2, MMP9, and MMP14, degrade the basement membrane and extracellular matrix components, creating physical pathways for invasion. Together, these genes are regulated by signaling pathways such as TGF-β, Wnt/β-catenin, and EMT transcription factors (SNAI1, TWIST1, ZEB1), which collectively promote the transition from a stationary to an invasive cell phenotype in melanoma progression.

Focus on one cancer type: Skin Cutaneous Melanoma

- Prevalence & incidence:
  - Fifth most common cancer type in the US
  - About 1 in 27 men and 1 in 40 women in the U.S. will be diagnosed with melanoma during their lifetime
  - About 100,000 new cases and about 8,000 deaths annually
  - Incidence has been rising in the past few decades (particularly in non-Hispanic white populations) however rates are stabilizing in younger groups due to improved sun protection awareness.
  - From ~13.8 per 100,000 in 1990 to ~22.6 per 100,000 in 2018.
  - Higher prevalent rates in countries with higher UV exposure (Australia, New Zealand, southern U.S.)
  - https://www.cancer.org/cancer/types/melanoma-skin-cancer/about/key-statistics.html?utm
- Risk factors (genetic, lifestyle) & Societal determinants:
  - Ultraviolet (UV) light exposure increases risk greatly
  - Individuals with lighter complexion/fair skin are at a higher risk
  - History of sun burns increases risk
  - A high number of melanocytic nevi (moles) and presence of atypical/dysplastic moles is a risk factor
  - Family history and germline mutations such as gene mutations CDKN2A increase risk
  - Incidence rises with age
  - Increased risk for immunosuppressed individuals
  - Ethnicity/race: less common in non-white populations, but worse outcomes and later diagnosis in non-white populations
  - https://pmc.ncbi.nlm.nih.gov/articles/PMC8366310/?utm
- Standard of care treatments (& reimbursement):
  - The best treatment for cutaneous melanoma is surgery, which includes primarily Mohs micrographic surgery as well as tissue sparing and other excisions.
  - Non-surgical melanoma treatments, such as imiquimod and some radiation therapies, can also be applied to less potent cases, like with lentigo maligna melanoma (caused by sun-damaged skin) or melanoma that hasn't spread outside the original location (in situ).
  - Different stages of melanoma warrant different treatments, making it important to diagnose the current stage a patient's melanoma is.
    - Stage 0: If the cancer has not gone below the skin surface, then melanoma can be removed through one or multiple wide excision surgeries or any of the aforementioned non-surgical methods.
    - Stage 1: When melanoma gets deeper in the skin, wide excision surgeries can still be used to treat it. If the melanoma get is nearby a lymph node, then patients may get a sentinel lymph node biopsy (SLNB) to check for cancer cells; if cancer cells are present, then a medical team may suggest a lymph node dissection and other treatments to monitor the lymph node (e.g. occasional ultrasound imaging) and/or to help prevent the cancer from becoming recurrent (e.g. immune checkpoint inhibitors or therapy drugs).
    - Stage 2: The cancer has deepened further since stage 1 but still hasn't grown beyond the skin. The treatments are the same as stage 1 with heavier emphasis on SLNB. Doctors may also recommend radiation therapy to reduce chance of recurrence.
    - Stage 3: The melanoma has infested further into nearby skin areas and lymph nodes, now necessitating wide excision surgeries and lymph nodes dissections, alongside the previous recurrence treatments. Medical professionals may also inject a vaccine straight

into the melanoma, like the T-VEC, interleukin-2, or Bacille Calmette-Guerin (BCG) vaccines.

- Stage 4: This stage marks when the melanoma cancer has metastasized to other body parts. Treatments include the T-VEC vaccine, radiation therapy, and certain surgeries if there are only a few metastases. For more widespread cancer cases, recent findings suggest immunotherapy and more targeted drug therapies to treat the melanoma, though chemotherapy can still be employed to help out.

- In America, reimbursement for skin cutaneous melanoma can be partially covered by Medicare and private insurance plans, though what medical expenses they cover vary. Other government programs, patient assistance programs, and possibly Social Security disability benefits can also help reimburse melanoma patients (see this link for a list of those options: https://melanoma.org/patients-caregivers/find-support/financial-assistance/)
- Sources:

https://www.aad.org/member/clinical-quality/guidelines/melanoma
https://www.cancer.org/cancer/types/melanoma-skin-cancer/treating/by-stage.html Gemini AI

- Biological mechanisms (anatomy, organ physiology, cell & molecular physiology):
  - Melanoma typically starts when skin cells experience DNA changes or mutations, due to ultraviolet (UV) radiation or other unknown causes.
  - Skin pigmentation can play a role in melanoma development, since lighter skin pigmentations cannot block out DNA-damaging UV light, from the sun or other sources, as greatly as darker pigmentations.

When melanoma is genetically induced, certain genes like the p53 tumor-suppressor or BRAF are disrupted in some way.

- The p53 gene is responsible for ordering pathway repairs or apoptosis to cells with DNA damage, so mutations to that p53 gene and other tumor suppressors can disable this function and let unregulated cell proliferation occur.
- BRAF is a gene whose RNA is transcribed and translated into a growth signal protein that, when bound to a membrane receptor on another cell, activates a transduction pathway that ultimately sends a protein into the cell nucleus to turns on genes coding for cell division. For melanoma, BRAF can be mutated to always be activated, thus making skin cells constantly send out growth signals that generate excess cancer cells.
- Other mutated genes, or oncogenes, causing melanoma typically trigger signaling pathways (e.g. MAPK, PI3K/AKT, c-KIT pathways) that counteract regulation to cell growth or increase the survivability of cancer cells.
- When left without treatment, cancer cells may start sending out pioneer cells, which tend to have altered cell-cell adhesion molecules that help in tethering to other tissue cells, that invade and attach to other parts of the human body, initiating deadly metastases outside the original melanoma location.
- Sources:

https://www.mayoclinic.org/diseases-conditions/skin-cancer/symptoms-causes/syc-20377605
https://www.healthwellfoundation.org/realworldhealthcare/the-mechanics-of-melanoma/
https://www.cell.com/fulltext/S0092-8674(00)81683-9

Data-Set:

How was the data collected?

- The data was collected by The Cancer Genome Atlas (TCGA) which collected tumor tissue samples and clinical information across thousands of cancer patients across the United States.
  - Tissue Source Sites find cancer patients that donate their tumor biospecimens, including blood and/or tissue parts, to them. They then give the biospecimens to the Biospecimen Core Resource, a facility that checks and acquires clinical data and metadata from the biospecimens. Other groups continue to further process and polish the data before being submitted to TCGA.
- For this specific study, researchers reprocessed RNA-sequencing data for 9,264 tumor samples (including 472 Skin Cutaneous Melanoma (SKCM) samples) using a new computational pipeline.
  - The RNA-sequencing data was further subsetted into a filtered csv sheet that includes recorded values for only the top 3000 variance genes.
- Clinical information/metadata such as patient age, tumor stage, gender, and sample type (primary vs metastatic) was downloaded directly from the TCGA clinical files.
  - Source of processed Data: Rahman, M., Jackson, L. K., Johnson, W. E., Li, D. Y., Bild, A. H., & Piccolo, S. R. (2015). Alternative preprocessing of RNA-Sequencing data in The Cancer Genome Atlas leads to improved analysis results. Bioinformatics (Oxford, England), 31(22), 3666–3672. https://doi.org/10.1093/bioinformatics/btv377
  - Source of original TCGA Data: Wilks, C., Cline, M. S., Weiler, E., Diehkans, M., Craft, B., Martin, C., Murphy, D., Pierce, H., Black, J., Nelson, D., Litzinger, B., Hatton, T., Maltbie, L., Ainsworth, M., Allen, P., Rosewood, L., Mitchell, E., Smith, B., Warner, J., Groboske, J., … Maltbie, D. (2014). The Cancer Genomics Hub (CGHub): overcoming cancer through the power of torrential data. Database : the journal of biological databases and curation, 2014, bau093. https://doi.org/10.1093/database/bau093

What techniques were used?

- RNA-sequencing techniques were done on the biospecimens (blood and tissue samples) taken from the patients' tumors of varying cancer types. According to Gemini AI, the procedure to do RNA sequencing involves extracting RNA from the tumor sample, fragmenting the RNA, applying reverse transcription to convert the RNA fragments to a cDNA sequence, and processing that cDNA in a high-throughput sequencer to attain the RNA-Seq data.
- The researchers who re-processed the TCGA data also applied single-sample normalization techniques onto the datasets

What relevant clinical features are in the dataset?

- Cancer type: Skin Cutaneous Melanoma (SKCM)
- Age at diagnosis: independent variable
- Sample type: primary vs metastatic
- Pathologic stage

What units are the data measured in?

- Age: measured in years
- Sample type: primary tumor, metastatic tumor, solid tissue normal (categorical)
- Pathologic stage: stages I-IV represents disease severity

When was the data collected and by who?

- The data was collected by multiple medical centers across the U.S. in contribution to the National Institute of Health's "The Cancer Genome Atlas" project, which began in late 2005. Since the start, the collection of cancer genome data has been continually added to throughout the decades since.

Are there potential areas of bias?

- The researchers, Rahmen et al., who re-processed TCGA data declared no conflict of interest in their study, and they were funded by reputable health organizations like the National Library of Medicine training fellowship and the National Institutes of Health. Given how the data from TCGA was collected and likely reviewed on a national level, the influence of bias on the current dataset is likely trivial.

Data Analyis:

Methods

Research Question: *Is patient age associated with patterns of expression in metastasis-related genes in cutaneous melanoma of the skin (SKCM)? Do younger and older patients cluster differently based on the expression of sixteen genes associated with tissue invasion and metastasis?*

To explore this question, we used K-means clustering, an unsupervised machine learning method that groups samples based on feature similarity. The algorithm iteratively assigns each sample to the cluster with the nearest centroid and updates centroid positions to minimize the within-cluster sum of squares (inertia). This optimization creates clusters that are as compact and distinct as possible, capturing natural structure in the data without predefined labels.

Clinical metadata and RNA-seq expression values were filtered to include only SKCM samples, then merged into a single dataset containing sample_id, age_at_initial_pathologic_diagnosis, sample_type, and 16 metastasis-related genes. We visualized the age distribution using a histogram and, because the distribution was unimodal, applied clustering to all patients rather than splitting by age groups.

Before clustering, gene expression values were standardized using z-score scaling (StandardScaler) to ensure equal weighting across genes. We ran K-means for k = 2–5 and evaluated cluster quality using the Elbow Method (inertia) and Silhouette Score. To visualize the resulting cluster structure, we used Principal Component Analysis (PCA) to project the 16-dimensional gene space into two components.

To assess how well the clusters generalized, we performed a train–validation split (70%/30%), fit K-means on the training set, and computed validation metrics including Silhouette Score, Davies–Bouldin Index, and Calinski–Harabasz Index. Finally, to test whether clusters differed by age, we used a Kruskal–Wallis test, a nonparametric method for comparing distributions across multiple groups.

Analysis

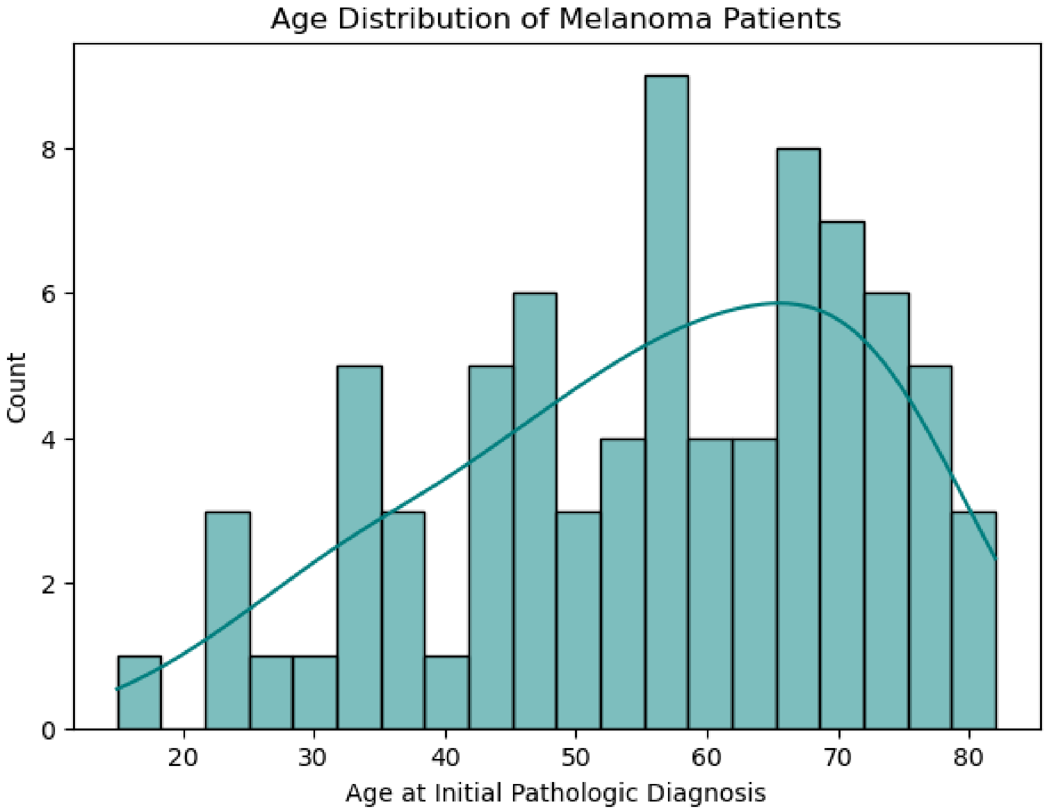
In [1]:

```
# (1.) Filtering relevant data from datasets and merging them into merged_data DataFram


import pandas as pd

metadata = pd.read_csv('GSE62944_metadata.csv') # clinical metadata
rnadata = pd.read_csv('GSE62944_subsample_log2TPM.csv') # RNA-seq log2(TPM) data

# Normalize rnadata so genes are the index (file has genes in 'Unnamed: 0')
if 'Unnamed: 0' in rnadata.columns:
    rnadata = rnadata.rename(columns={'Unnamed: 0': 'gene'}).set_index('gene')
elif 'gene' in rnadata.columns and rnadata.index.name != 'gene':
    rnadata = rnadata.set_index('gene')

# Filter for just melanoma (SKCM) samples
skcm_metadata = metadata[metadata['cancer_type'] == 'SKCM'].copy()

# Prefer age_at_diagnosis if age_at_initial_pathologic_diagnosis is missing
if ('age_at_initial_pathologic_diagnosis' not in skcm_metadata.columns
        or skcm_metadata['age_at_initial_pathologic_diagnosis'].isna().all()):
    if 'age_at_diagnosis' in skcm_metadata.columns:
        skcm_metadata['age_at_initial_pathologic_diagnosis'] = skcm_metadata['age_at_di

# Standardize sample id and sample type column names used downstream
if 'sample' in skcm_metadata.columns:
    skcm_metadata = skcm_metadata.rename(columns={'sample': 'sample_id'})
if 'submitted_tumor_site' in skcm_metadata.columns:
    skcm_metadata = skcm_metadata.rename(columns={'submitted_tumor_site': 'sample_type'

# Filter for the clinical features we need (now using standardized names)
skcm_metadata = skcm_metadata[['sample_id', 'age_at_initial_pathologic_diagnosis', 'sam

# Filter for relevant genes and check if they are in the rnadata index
genes = [
    'BRAF','NRAS','AXL','MITF','MMP2','MMP9','MMP14','FN1','VIM','TGFB1',
    'CASP8','KISS1','PTEN','TERT','CDKN2A','NEDD9'
]
present_genes = [gene for gene in genes if gene in rnadata.index]
print(f"Genes present in RNA-seq data: {present_genes}")
missing_genes = [gene for gene in genes if gene not in rnadata.index]
print(f"Genes missing in RNA-seq data: {missing_genes}")

# Adjust columns and rows of rnadata to merge the metadata with the RNA-seq data
common_samples = sorted(set(skcm_metadata['sample_id']) & set(rnadata.columns))
# Use only present_genes (avoids KeyError if some genes missing)
rnadata_subset = rnadata.loc[present_genes, common_samples].T.reset_index().rename(colu
merged_data = pd.merge(skcm_metadata, rnadata_subset, on='sample_id', how='inner')

print() # For spacing
print(merged_data.head())
```

```
Genes present in RNA-seq data: ['BRAF', 'NRAS', 'AXL', 'MITF', 'MMP2', 'MMP9', 'MMP14',
'FN1', 'VIM', 'TGFB1', 'CASP8', 'KISS1', 'PTEN', 'TERT', 'CDKN2A', 'NEDD9']
Genes missing in RNA-seq data: []

                      sample_id  age_at_initial_pathologic_diagnosis  \
0  TCGA-W3-A828-06A-11R-A352-07                                 66.0
1  TCGA-FW-A5DX-01A-11R-A27Q-07                                 71.0
2  TCGA-D3-A5GU-06A-11R-A27Q-07                                 36.0
3  TCGA-BF-AAP7-01A-11R-A40A-07                                 76.0
4  TCGA-IH-A3EA-01A-11R-A20F-07                                 61.0

     sample_type      BRAF      NRAS       AXL      MITF       MMP2      MMP9  \
0  Head and Neck  1.936795  5.872184  6.833947  3.452778  10.551046  5.024924
1          Trunk  3.793577  6.729344  3.386108  7.629105   9.304800  2.958118
2    Extremities  2.595629  6.365169  3.862842  6.533163   4.979634  3.861427
3    Extremities  1.553574  4.519536  4.101128  5.656342   8.474425  5.514855
4  Head and Neck  1.312264  4.209623  3.739283  5.256164   4.388166  4.999968

       MMP14        FN1        VIM     TGFB1     CASP8     KISS1      PTEN  \
0  10.351109  11.317356  11.806829  7.223138  4.178757  0.328334  5.779152
1   8.836577   8.198240  11.682906  3.464710  4.980035  0.000000  4.139859
2   8.661882   6.914597  12.018090  5.402041  3.262768  0.000000  5.035510
3   9.566007   9.310747  12.240442  6.075530  4.246337  0.000000  4.050369
4   9.808056  10.399576  13.882052  6.051691  3.018274  0.000000  3.175137

       TERT    CDKN2A     NEDD9
0  0.540724  5.886559  5.581397
1  4.206887  7.839580  5.540719
2  2.159369  2.251276  5.776669
3  1.067104  6.637780  3.255752
4  1.632295  5.553466  5.007947
```

In [2]:

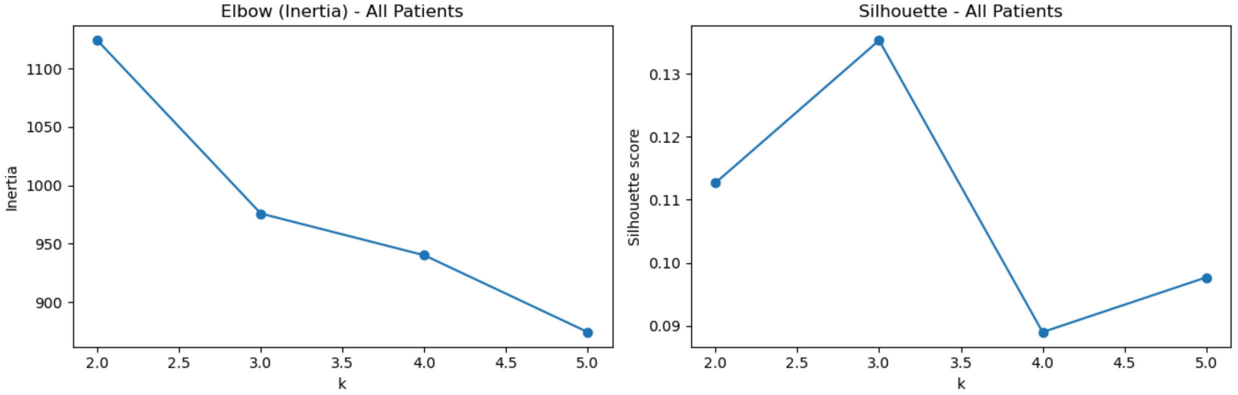
```python
# (2.) Explore age distribution to decide how to handle age
import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(7,5))
sns.histplot(merged_data['age_at_initial_pathologic_diagnosis'], bins=20, kde=True, col
plt.title("Age Distribution of Melanoma Patients")
plt.xlabel("Age at Initial Pathologic Diagnosis")
plt.ylabel("Count")
plt.show()
```



In [6]:

```
# Make elbow and silhouette plots for all patients together
import numpy as np, pandas as pd, seaborn as sns, matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
from sklearn.metrics import silhouette_score
from scipy.stats import kruskal

# 1) Features = 16 genes metastatic related genes
non_gene = ["sample_id","age_at_initial_pathologic_diagnosis","sample_type"]
gene_cols = [c for c in merged_data.columns if c not in non_gene]

all_patients = merged_data.dropna(subset=gene_cols).copy()
X = all_patients[gene_cols].values

# 2) Scale
Xs = StandardScaler().fit_transform(X)

# 3) Elbow and silhouette plots
Ks = range(2, 6)
inertias, sils = [], []
lbls_cache = {}  # store labels per k so we can reuse later

for k in Ks:
    km = KMeans(n_clusters=k, random_state=42, n_init='auto', max_iter=1000)
    labels = km.fit_predict(Xs)
    lbls_cache[k] = (km, labels)
    inertias.append(km.inertia_)
    sils.append(silhouette_score(Xs, labels))

sil_max = max(sils)

fig, ax = plt.subplots(1,2, figsize=(12,4))
ax[0].plot(list(Ks), inertias, 'o-')
```

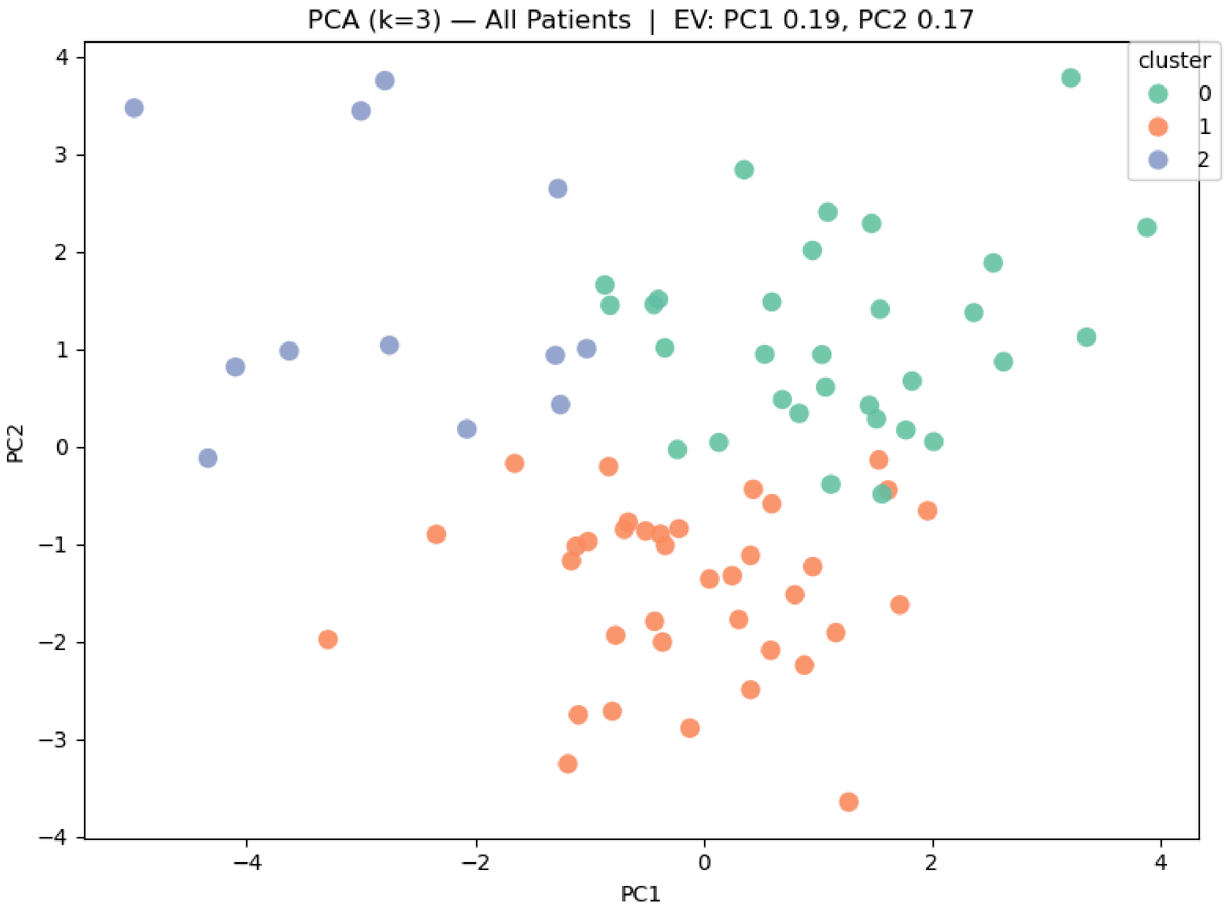
```
ax[0].set_title('Elbow (Inertia) - All Patients')
ax[0].set_xlabel('k')
ax[0].set_ylabel('Inertia')

ax[1].plot(list(Ks), sils, 'o-')
ax[1].set_title('Silhouette - All Patients')
ax[1].set_xlabel('k')
ax[1].set_ylabel('Silhouette score')
plt.tight_layout()
plt.show()
print("Silhouette score: ", sil_max)
```



```
Silhouette score:  0.1352877861758583
```

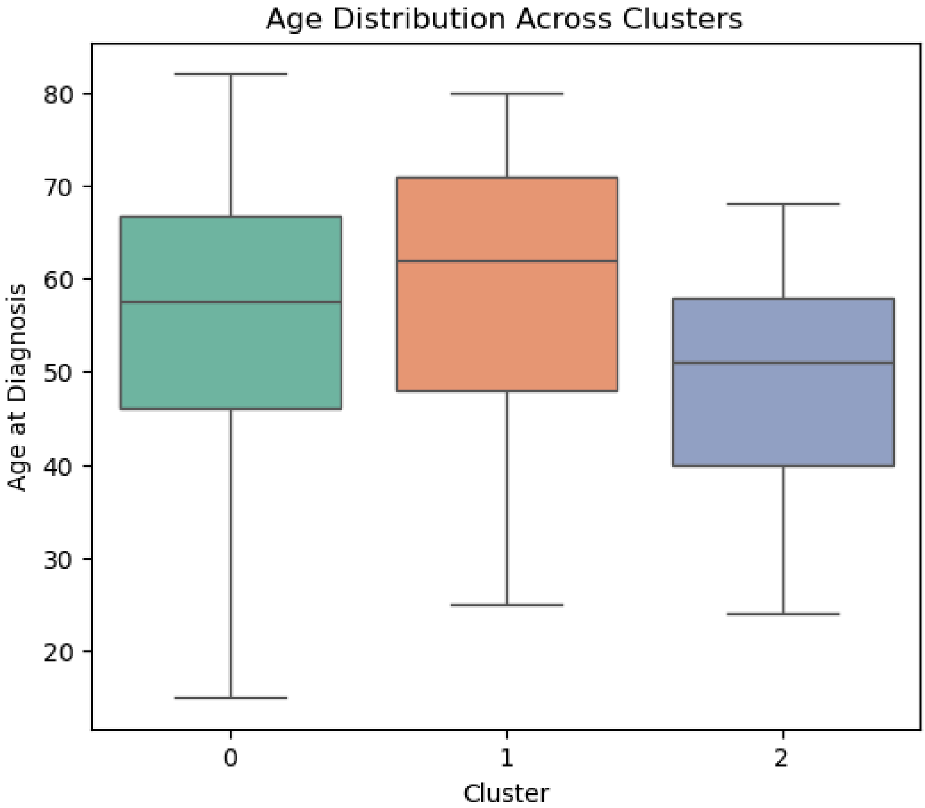
In [7]:

```
# Select features
non_gene = ["sample_id", "age_at_initial_pathologic_diagnosis", "sample_type"]
gene_cols = [c for c in merged_data.columns if c not in non_gene]

# Clean copy with no missing gene values
all_patients = merged_data.dropna(subset=gene_cols).copy()

# Builds the feature matrix and scales it
X = all_patients[gene_cols].values
scaler = StandardScaler()
Xs = scaler.fit_transform(X)

# Choose k based on elbow and silhouette plots
# The “elbow” appears around k=3, where inertia improvement slows down suggesting 3 clu
# score is also at k=3.
k = 3  # <-- change after looking at plots

# Fit K-means and attach cluster labels to the dataframe
kmeans = KMeans(n_clusters=k, random_state=42, n_init='auto', max_iter=1000)
labels = kmeans.fit_predict(Xs)
all_patients['cluster'] = labels

pca = PCA(n_components=2)

# PCA for visualization
pcs = pca.fit_transform(Xs)

# Add PCs to a plotting copy
plot_df = all_patients.copy()
plot_df['PC1'] = pcs[:, 0]
plot_df['PC2'] = pcs[:, 1]

ev = pca.explained_variance_ratio_  # fraction of variance by PC1, PC2

plt.figure(figsize=(8,6))
sns.scatterplot(
    data=plot_df,
    x='PC1', y='PC2',
    hue='cluster',
    palette='Set2',
    s=80, alpha=0.9, edgecolor='none'
)
plt.title(f"PCA (k={k}) – All Patients  |  EV: PC1 {ev[0]:.2f}, PC2 {ev[1]:.2f}")
plt.xlabel("PC1"); plt.ylabel("PC2")
plt.legend(title='cluster', bbox_to_anchor=(1.02, 1), borderaxespad=0)
plt.tight_layout()
plt.show()

# Do clusters align with age? Prints descriptive stats and runs Kruskal-Wallis test
print("\nSamples per cluster:")
print(all_patients['cluster'].value_counts().sort_index())

print("\nMean age per cluster:")
print(all_patients.groupby('cluster')['age_at_initial_pathologic_diagnosis'].mean())

# Visual comparison of age across clusters using a box plot
plt.figure(figsize=(6,5))
sns.boxplot(data=all_patients, x='cluster', y='age_at_initial_pathologic_diagnosis', pa
plt.title("Age Distribution Across Clusters")
plt.xlabel("Cluster"); plt.ylabel("Age at Diagnosis")
plt.show()

from scipy.stats import kruskal

# Drop missing ages
clean = all_patients.dropna(subset=['age_at_initial_pathologic_diagnosis']).copy()

# Make groups only for clusters that have 2+ samples
groups = [g['age_at_initial_pathologic_diagnosis'].values
          for _, g in clean.groupby('cluster')
          if len(g) > 1]

# Only run if there are at least 2 non-empty groups
if len(groups) >= 2:
    H, p = kruskal(*groups)
    print(f"Kruskal-Wallis test (age across clusters): H={H:.2f}, p={p:.4f}")
    if (p < 0.05):
        print(f"There is a significant difference in ages at diagnosis between clusters
    else:
        print(f"There is NOT a significant difference in ages at diagnosis between clus
else:
    print("Not enough clusters with data to run Kruskal-Wallis test.")
```



```
Samples per cluster:
cluster
0    31
1    37
2    12
Name: count, dtype: int64

Mean age per cluster:
cluster
0    55.200000
1    58.972973
2    48.750000
Name: age_at_initial_pathologic_diagnosis, dtype: float64
```

```
/var/folders/wm/rgqw3cq942b_5gxs1xx22wsh0000gn/T/ipykernel_20776/1136874448.py:58: Futur
eWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0.
Assign the `x` variable to `hue` and set `legend=False` for the same effect.

  sns.boxplot(data=all_patients, x='cluster', y='age_at_initial_pathologic_diagnosis', p
alette='Set2')
```



```
Kruskal–Wallis test (age across clusters): H=4.71, p=0.0951
There is NOT a significant difference in ages at diagnosis between clusters (p=0.0951 >=
0.05)
```

In [8]:
```
# Our attempt to apply train_test_split to our clustering model (we didn't know exactly

from sklearn.model_selection import train_test_split
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
import numpy as np


# Split data into training and validation sets
X_train, X_val = train_test_split(X, test_size=0.3, random_state=42)

# Iterate through different numbers of clusters (hyperparameter tuning)
for n_clusters in range(2, 6):
    kmeans = KMeans(n_clusters=n_clusters, random_state=42, n_init='auto')
    kmeans.fit(X_train)  # Train on the training set

    # Predict clusters for the validation set
    val_labels = kmeans.predict(X_val)

    # Evaluate performance on the validation set using a clustering metric
    if len(np.unique(val_labels)) > 1:  # Silhouette score requires at least 2 clusters
        score = silhouette_score(X_val, val_labels)
        print(f"Number of clusters: {n_clusters}, Silhouette Score on Validation Set: {
    else:
        print(f"Number of clusters: {n_clusters}, Not enough clusters for Silhouette Sc
```

```
Number of clusters: 2, Silhouette Score on Validation Set: 0.1575
Number of clusters: 3, Silhouette Score on Validation Set: 0.1168
Number of clusters: 4, Silhouette Score on Validation Set: 0.0971
Number of clusters: 5, Silhouette Score on Validation Set: 0.0670
```

In [9]:
```python
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score, davies_bouldin_score, calinski_harabasz_s
import numpy as np

# Split data into training and validation sets
X_train, X_val = train_test_split(X, test_size=0.30, random_state=42, shuffle=True)

# Scale using train fit only
scaler = StandardScaler().fit(X_train)
Xs_train = scaler.transform(X_train)
Xs_val   = scaler.transform(X_val)

# Try k = 2, 3, 4, 5; evaluate on train and validation sets; check stability on validat
for n_clusters in range(2, 6):
    km = KMeans(n_clusters=n_clusters, random_state=42, n_init='auto', max_iter=1000)
    # Fit on train
    train_labels = km.fit_predict(Xs_train)
    # Predict assignments on validation using train centroids
    val_labels_pred = km.predict(Xs_val)

    # Internal metrics (need >=2 clusters in labels)
    if len(np.unique(train_labels)) > 1 and len(np.unique(val_labels_pred)) > 1:
        sil_tr = silhouette_score(Xs_train, train_labels)
        sil_va = silhouette_score(Xs_val,   val_labels_pred)
        db_tr  = davies_bouldin_score(Xs_train, train_labels)
        db_va  = davies_bouldin_score(Xs_val,   val_labels_pred)
        ch_tr  = calinski_harabasz_score(Xs_train, train_labels)
        ch_va  = calinski_harabasz_score(Xs_val,   val_labels_pred)
    else:
        sil_tr = sil_va = db_tr = db_va = ch_tr = ch_va = np.nan

    print(
        f"k={n_clusters} | "
        f"Sil(TR)={sil_tr:.3f}, Sil(VAL)={sil_va:.3f} | "
        f"DB(TR)={db_tr:.2f}, DB(VAL)={db_va:.2f} | "
        f"CH(TR)={ch_tr:.1f}, CH(VAL)={ch_va:.1f} | "
    )
```

```
k=2 | Sil(TR)=0.159, Sil(VAL)=0.082 | DB(TR)=1.82, DB(VAL)=1.95 | CH(TR)=3.7, CH(VAL)=2.
0 |
k=3 | Sil(TR)=0.117, Sil(VAL)=0.080 | DB(TR)=2.08, DB(VAL)=2.29 | CH(TR)=6.1, CH(VAL)=3.
0 |
k=4 | Sil(TR)=0.094, Sil(VAL)=0.066 | DB(TR)=2.21, DB(VAL)=1.87 | CH(TR)=6.0, CH(VAL)=3.
0 |
k=5 | Sil(TR)=0.112, Sil(VAL)=0.082 | DB(TR)=1.87, DB(VAL)=1.75 | CH(TR)=6.3, CH(VAL)=3.
5 |
```

Verify and validate your analysis:

Verification

Since we used K-means clustering for our machine learning model, we decided to find the silhouette score using the sklearn.metrics library to determine how well the model assigned patients to distinct, non-overlapping clusters. After finding which number of clusters produced the best silhouette score, our K-means model ended up with a max silhouette score of 0.1352877861758583. This score is around 0, meaning that the clusters were defined poorly and are in close proximity with one another, if not partially overlapping. The low silhouette score also implies that the differences in expression of our list of melanoma metastasis genes did not form any clear-cut trends amongst the patients. Given the high number of genes put in as attributes for our clustering model, it's possible that one or more of those genes we picked had very little expression correlation with the others, diversifying the patient gene expression data in our subset to the point in which no strong, holistic patterns were found. Thus, we find it believable that our data didn't form any strongly-defined clusters across melanoma patients.

To see if the patients' age of diagnosis were significantly different across clusters, we applied a Kruskal–Wallis test, via a function from the scipy.stats library, that measures the statistical significance of the median diagnosis ages between those clusters. The Kruskal–Wallis test ended up with a p-value of 0.0951. Assuming that the significance level (alpha value) is set to 0.05, the 0.0951 p-value being over 0.05 means that there are no significant differences in median diagnosis ages between the clusters, suggesting that patient age of diagnosis is not associated with expression of the melanoma metastasis-related genes we picked. Ultimately, we found this result to be believable because the clustering quality of our model was already not the best, as shown by our low max silhouette score; this bad clustering suggests that the characteristics of patients, such as diagnosis age, between different clusters are more likely to have some overlap with each other.

Finally, we employed sklearn’s train_test_split function to determine how consistently our model clusters unlabeled data by comparing cluster scores from the training dataset and a separate, newly-generated test dataset. In addition to silhouette scores, we decided to also compare Davies-Bouldin and Calinski-Harabasz scores for the training and test dataset clustering at different quantities of clusters, and the result trends are described below:

1. Silhouette score (Sil):
   - Training: scores slightly decrease as k increases (generally)
   - Validation: scores are consistent across k values (about 0.08)
   - The validation scores are lower and show more variance across clusters

Interpretation: Our data likely form overlapping clusters, and the highest overall silhouette was at k = 2 (train = 0.159, val = 0.082), suggesting that simpler clustering may generalize slightly better. 2. Davies-Bouldin (DB):

- Training vs validation: all around 1.8-2.3, showing slight variance

Interpretation: cluster compactness and separation are similar across all k values, with no major improvement beyond k=2 to k=3 3. Calinski-Harabasz (CH):

- Training CH improves with more clusters (3.7 at k = 2 to 6.3 at k = 5), but validation CH doesn’t increase the same way, and values are lower (2.0 to 3.5).

Interpretation: for the training data, as the clusters are more compact (increasing CH), the silhouette score decreases, meaning they also become less distinct Generally, the model fits slightly better on the training samples than on unseen validation samples; however, the scores remain in the same/similar range, leading us to believe that the cluster structure is relatively stable but weakly defined. Due to similar trends across the training and validation data sets, but slightly higher values for the training data set, our model is somewhat overfitting, but generally predicts fairly well.

Validation

Contrary to our findings, many external studies on the effect of age on melanoma metastasis, like the ones listed below, support that older melanoma patients have a higher risk of metastasis. This lack of literature validation suggests that our study may be limited or flawed in some way. Albeit, most of these outside sources heavily focus on HAPLN1, a gene not included in the list of melanoma metastasis-related genes utilized for training our model, in letting age influence melanoma metastasis (Kaur et al., 2019; Marino-Bravante et al., 2024; Weeraratna, 2024). Another study conducted by Kretschmer et al. even found that for melanoma patients, there's "an increased risk of SLN-metastasis for individuals 40 years of age or younger" compared to older people, challenging the researched notion that melanoma metastasis probability increases with older age (2010). Given the mixed results from our study and the outside literature, it is possible that age has a more complex effect on melanoma metastasis, along with the genes associated with it, than we initially expected.

Sources:

- Kaur, A., Ecker, B. L., Douglass, S. M., Kugel, C. H., 3rd, Webster, M. R., Almeida, F. V., Somasundaram, R., Hayden, J., Ban, E., Ahmadzadeh, H., Franco-Barraza, J., Shah, N., Mellis, I. A., Keeney, F., Kossenkov, A., Tang, H. Y., Yin, X., Liu, Q., Xu, X., Fane, M., … Weeraratna, A. T. (2019). Remodeling of the Collagen Matrix in Aging Skin Promotes Melanoma Metastasis and Affects Immune Cell Motility. Cancer discovery, 9(1), 64–81. https://doi.org/10.1158/2159-8290.CD-18-0193
- Kretschmer, L., Starz, H., Thoms, K. M., Satzger, I., Völker, B., Jung, K., Mitteldorf, C., Bader, C., Siedlecki, K., Kapp, A., Bertsch, H. P., & Gutzmer, R. (2011). Age as a key factor influencing metastasizing patterns and disease-specific survival after sentinel lymph node biopsy for cutaneous melanoma. International journal of cancer, 129(6), 1435–1442. https://doi.org/10.1002/ijc.25747
- Marino-Bravante, G. E., Carey, A. E., Hüser, L., Dixit, A., Wang, V., Kaur, A., Liu, Y., Ding, S., Schnellmann, R., Gerecht, S., Gu, L., Eisinger-Mathason, T. S. K., Chhabra, Y., & Weeraratna, A. T. (2024). Age-dependent loss of HAPLN1 erodes vascular integrity via indirect upregulation of endothelial ICAM1 in melanoma. Nature Aging, 4(3), 350–363. https://doi.org/10.1038/s43587-024-00581-8
- Weeraratna, A. (2024, March 12). Age-related changes in skin may contribute to melanoma metastases | Johns Hopkins Medicine. Johns Hopkins Medicine. https://www.hopkinsmedicine.org/news/newsroom/news-releases/2024/03/age-related-changes-in-skin-may-contribute-to-melanoma-metastases

Another possible reason for our model's inability to establish a significant relationship between age and melanoma metastasis gene expression is the lack of age-related genes, such as the aforementioned HAPLN1 gene, contributing to the machine learning model. A study by Tian et al. (2023) elaborates on the existence of specific age-related genes linked to melanoma progression, none of which were the genes we listed and used as attributes for our clustering model. The existence of age-related genes for melanoma suggests that our selected list of genes, although supposedly related to melanoma metastasis based on prior research, may not be strongly linked to age, leading to a lack of significant diagnosis-age-related differences between clusters generated from our model.

Sources:

- Tian, C., Liu, S., & Huo, R. Identification of the ageing-related prognostic gene signature, and the associated regulation axis in skin cutaneous melanoma. Sci Rep 13, 24 (2023). https://doi.org/10.1038/s41598-022-22259-0

Conclusions and Ethical Implications:

- Overall, our study on the effect of the expression of certain melanoma metastasis-related genes and age of diagnosis found no significant age-related correlation across the 3 clusters formed from 16 metastasis-related genes. Although the train-test split confirmed that our clustering model predicts well enough despite its overfitting aspects, our machine learning model of K-means clustering generated clusters, based on selected gene expression data, that were not very distinct from one another, resulting in a low maximum silhouette score of ~0.135. Furthermore, a Kruskal–Wallis test revealed no significant differences in diagnosis age medians between our generated clusters, aligning with our conclusion that age does not correlate with melanoma metastasis-related gene expression based on our results. This conclusion contrasts with the findings of outside literature; however, this suggests that limitations in the data or model design may have influenced our findings to some degree.

- Ethically, it is essential to consider that when drawing conclusions from data based on gene expression alone and not incorporating other clinical features (e.g. tumor stage, other patient demographics, and sampling bias), those conclusions do not account for all of the determinants of metastasis in patients. Generally, when drawing conclusions from our clustering model and clustering models in general, it is essential not to forget that these conclusions can help generate hypotheses; however, they should not impact clinical decision-making.

    - For instance, if a melanoma patient or their support team were to only use the results from our study in guiding their medical decisions, they may be led to believe age is not a major factor in driving melanoma metastasis, even though outside literature would argue otherwise. This ill-informed perception may make them less worried about patient age than they should be, putting older melanoma patients at risk of choosing not to get checked up on often and other bad health decisions. To avoid this scenario, the limitations of our study and results should be clearly stated and emphasized.
- Considering the research field, the results about age not showing a relation to several melanoma metastasis genes may spur researchers to look more into this subject and scrutinize other studies that found impacts of age on melanoma metastasis. This can inspire more experimental exploration in the matter of melanoma metastasis-related genes to expand the realm of knowledge in it, but it can also divert attention away from more credible studies finding significant relationships between age and melanoma metastasis and lead organizations to invest more funds into following up on our findings, which may not be successful given the limitations of our study.

Limitations and Future Work:

- Our analysis did not generate a significant answer to our question as we did not find significant differences based on age across our clusters. This is most likely due to various factors we did not consider. We analyzed the gene expression of metastasis-related genes; however, we did not take into account whether the tumor in the sample was metastatic or not, or its stage, which are important factors to consider. Just because there is increased expression of metastasis-related genes does not directly mean that there is increased metastasis, which was a limitation in our study.

- It’s also possible that the melanoma metastasis-related genes featured in the model do not experience expression changes due to age, as well as other genes that were not included, which could deviate our results from those of other studies on the topic. Another potential future direction is to examine different and/or additional genes based on further research.

- In the future, increasing the amount of data incorporated into our model using different features from the clinical data set could help us draw conclusions related to our question and potentially see differences based on age in our clusters. If we incorporated features such as pathologic tumor stage and clinical metastasis, and then formed clusters based on these factors as well, there is a higher chance that we could find differences in age across clusters, as the literature shows a correlation between metastasis and age. We originally had this factor as a feature in our question/project goal; however, due to the complexity and time limitations, we decided against incorporating that additional feature in the data analysis and clustering.
- To incorporate this in our data, we could use more features in our clustering and perform hierarchical clustering on two separate groups and then compare these statistically/graphically. For example, we could split the data into samples with primary tumors and samples with metastatic tumors and perform the same clustering on each, then compare and use the same means of statistical verification using age (box plot and Kruskal–Wallis test) to incorporate age. The primary tumor group would act as a control group.
- Additionally, incorporating more clinically relevant features (such as tumor stage or metastatic status) may help improve cluster separation and reduce noise, which can help the model generalize better, even though increasing dimensionality does not inherently reduce overfitting in K-means.

NOTES FROM YOUR TEAM:

This is where our team is taking notes and recording activity.

Progress Notes

- 10/23/2025: Had Comp BME class that further introduced the Module 3 project. We discussed and decided to narrow down our project scope to skin cutaneous melanoma and the cancer hallmark “Tissue Invasion and Metastasis.”
- 10/25/2025: Completed work for Module 3 first check-in, including filling out the disease background questions and data-set sections, starting our team notes and TA questions, and figuring out which question we want to answer for our project.
- 10/30/2025: Received and reviewed TA feedback and added a gene element to our question, creating a research list of genes that are related to melanoma metastasis. Decided on which method to use and began working on the code over the weekend.
  - Luke was also sick with the flu this week, so only Molly attended classes this week. Both still gave time to work on the things mentioned above.
- 11/3/2025: Both worked remotely on narrowing down our method and beginning the code for the second check-in, including sorting relevant data into a new DataFrame table and implement sklearn Kmeans to find clusters to be displayed.
- 11/6/2025: Got feedback about our code and overall method/analysis section, and discussed more about how to revise and validate our method and results for next check-in.
- 11/8/2025: Revised and changed our model to form clusters based on patients of all ages and used statistical verification to see if there were differences between clusters based on patient age. Applied train test split to validate our method
- 11/11/2025: Worked on drawing conclusions from out train test split and our verification/validation and determined limitations and future work to finalize the project and tie everything together.

- 11/12/2025: Final edits and submission based on feedback.

Extra Gene Background Notes

Original 10 (+ additional MMPs) found from background research

1. BRAF

- Gene that can be mutated to always be activated, thus making skin cells constantly send out growth signals that generate excess cancer cells.
- Involved in many cancer things, including metastasis

2. NRAS

- “there are higher rates of lymph node metastasis which may be related to increased cell motility and higher rates of epithelial mesenchymal transition (EMT) in NRAS mutated tumor cells.”

https://pmc.ncbi.nlm.nih.gov/articles/PMC9485270/ 3. AXL

- “Metastases in the sentinel lymph nodes were detected in 14 out of 61 patients, and these were associated with AXL-positive immunoreactivity in the primary tumor ($p < 0.0001$)”

https://pmc.ncbi.nlm.nih.gov/articles/PMC8566987/ 4. MITF

- “MITF is a transcription factor central to melanocyte survival, proliferation, and melanin-pigment production40. MIFT amplification has been found in 21% of metastases compared to 10% of primary melanomas41.”

https://pmc.ncbi.nlm.nih.gov/articles/PMC9485270/ 5. MMP1 6. MMP9 7. MMP13

- MMPs degrade the basement membrane and extracellular matrix components, creating physical pathways for invasion
- MMP 1, 9, and 13 shown to correlate to malignant melanoma metastasis

https://pmc.ncbi.nlm.nih.gov/articles/PMC11146457/ 8. MMP2

- “Furthermore, increased MMP-2 expression correlated with hematogeneous metastasis (Väisänen et al. 1996)”

https://www.sciencedirect.com/science/article/pii/S0022202X15409777 9. MMP3 10. MMP14

- MMP3 and MMP14, alongside the other mentioned MMPs, are also found to contribute to melanoma metastasis according to this source:

https://pmc.ncbi.nlm.nih.gov/articles/PMC11676509/#sec5-ijms-25-13558 11. FN1

- The study below found that Fibronectin 1 (FN1) demonstrated higher expression in metastatic melanoma cells compared to normal ones, also proposing that FN1 promotes metastasis by “inhibiting apoptosis and regulating EMT.”

https://pmc.ncbi.nlm.nih.gov/articles/PMC6503329/ 12. VIM

- Found to predict hematogenous metastasis (spreading cancer cells through the blood) for melanomas: “the results showed that over-expression of vimentin was frequently observed in primary melanoma patients with hematogenous metastasis (P < 0.05), not associated with lymph node metastasis (P > 0.05)."

https://pmc.ncbi.nlm.nih.gov/articles/PMC2928198/

- According to Gemini AI, overexpressed VIM genes enhances the motility and invasion rate of cancer cells, possibly by mediating EMT that leads to cells losing cell-to-cell junctions and becoming mesenchymal.

13. TGFβ1

- TGFβ1, or the TGFβ family in general, promotes melanoma tumors and metastasis by managing an environment that protects tumor growth and exerting tumor growth functions to increase cancer cell motility and invasiveness.

https://pubmed.ncbi.nlm.nih.gov/21619542/ https://www.nature.com/articles/s41419-024-07305-1

- According to Gemini AI, TGFβ1 can also activate MMP-2 and inhibit primary melanoma tumor growth.

Other potential genes related to metastasis from additional sources: 14. CASP8

- Metastasis initiation gene — "the loss of function of which favours dissemination by protecting tumour cells from programmed death upon release of integrin-mediated anchoring at the invasive front35.

https://www.nature.com/articles/nrg2101 15. Kiss-1

- "Knockout of Kiss-1 resulted in neoplastic proliferation after the tumor cells seeded themselves in the metastatic niche34" (area other than primary site)

https://pmc.ncbi.nlm.nih.gov/articles/PMC9485270/ 16. PTEN

- Tumor suppressor — "With the silencing of Pten, melanocytes undergo transformation and subsequent metastasis; thereby making Pten silencing a molecular marker for metastatic melanoma35"

https://pmc.ncbi.nlm.nih.gov/articles/PMC9485270/ 17. TERT

- "there is an increase in the TERT promoter mutations in metastases compared to primary melanomas, which represents an independent prognostic factor21,37,38"

https://pmc.ncbi.nlm.nih.gov/articles/PMC9485270/

- Possibly considered a secondary or phenotypic genetic driver for melanoma metastasis

18. CDKN2A

- "Other genetic changes that have been more commonly identified in metastases compared to primary tumors include loss-of-function of CDKN2A mutations"

https://pmc.ncbi.nlm.nih.gov/articles/PMC9485270/ 19. NEDD9

- Concluded from the study below to be a metastasis gene that has "enhanced invasion in vitro and metastasis in vivo of both normal and transformed melanocytes"

https://www.cell.com/cell/fulltext/S0092-8674(06)00718-5?
_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS0092867406007185%3Fs

Genes related to the Wnt/β-catenin pathway Genes related to the PI3K/AKT pathway MET, ASPM, AKAP9, IMP3, PRKCA, RPA3, and SCAP2

- “Functional assays have confirmed the pro-invasive properties of MET, ASPM, AKAP9, IMP3, PRKCA, RPA3, and SCAP2.61,62”

https://pmc.ncbi.nlm.nih.gov/articles/PMC9485270/ Possibly RIPK3
https://pmc.ncbi.nlm.nih.gov/articles/PMC8566987/ https://www.nature.com/articles/cddis2015240

QUESTIONS FOR YOUR TA:

These are questions we have for our TA.

First Check-in Questions

- Do you think our question is too simple, and if so, how do you suggest we revise it?
- Is there any way to measure metastasis via the Metadata dataset even if our cancer type doesn’t have much data in its directly metastasis-related columns? (otherwise, we’ll probably keep to the gene sequence data to study metastasis)
- Does our background bullets look good, or do we still need to flesh out a few topics? Is there any background bullets that we need to relate more to our cancer hallmark, tissue invasion and metastasis?
- Should we include more specifically relating to associated genes?
- Is it ok to use both the RNA-sequencing csv and Metadata csv for answering our Jupyter research question?

Second Check-in Questions

- Is clustering a good match for our project objective, or would another machine learning method like classification be better in your opinion?
- Should we do a dimensionality reduction before applying K-means clustering?
- Does the current code and method of filtering/merging the data sets make sense given our proposed methods?
- How would you suggest we best display our cluster results graphically (like through heatmaps, SCA plotting, something else entirely)?

Third Check-in Questions

- We did not finish the test-train split function in time. Do you have any advice how we should apply it to our code?
- Can you check the validation section to see if it's good? Most of the sources did not support our findings?
    - Should we possibly rethink the list of melanoma metastasis-related genes we have?