Here's your **Cancer Cell Invasion Analysis Project Documentation** rewritten in a **clean, structured, and professional format**, keeping **all your details intact** but with improved organization, readability, and scientific clarity:

Cancer Cell Invasion Analysis Project Documentation

Project Overview

- **Objective:** Analyze time-lapse cancer cell invasion behavior using quantitative imaging metrics and machine learning approaches.
- **Duration:** October 19–25, 2025
- **Analysis Type:** Time-series imaging data analysis with Principal Component Analysis (PCA) and clustering methods.

---

1. Data Preparation and Feature Selection

1.1 Dataset Summary

- **Samples:** 4 experimental conditions
    - `test_S1`, `test_S2`, `train_S1`, `train_S2`
- **Timepoints:** 12 per sample (`t0`–`t11`)
- **Total Observations:** 48 data points

1.2 Feature Engineering

Selected **16 quantitative invasion metrics** from cellular imaging data:

Category	Features
Spatial Metrics	mean_radius, max_radius, std_radius
Leader Cell Behavior	n_leader_cells, mean_leader_radius
Dispersion Measures	radius_skewness, radius_kurtosis, dispersion_index
Morphology	mean_volume, mean_compactness, mean_extent
Population Measures	cell_count, cell_density
Distance Metrics	mean_m_dist, median_m_dist

---

2. Principal Component Analysis (PCA)

2.1 PCA Results

Component	Variance Explained
PC1	35.99%
PC2	22.91%
PC1 + PC2 (Combined)	**58.90%**

Interpretation:
The first two principal components capture the dominant invasion behavior patterns across samples.

2.2 Statistical Validation

- **One-way ANOVA Results:**
    - PC1 by condition: **$p = 1.10 \times 10^{-10}$**
    - PC2 by condition: **$p = 6.14 \times 10^{-25}$**
- **Conclusion:** Strong, statistically significant separation between experimental conditions.

---

3. Clustering Analysis

3.1 K-Means Clustering (k = 4)

Each cluster aligned perfectly with one experimental condition:

Cluster	Sample	Observations	Mean Invasion (μm)
0	test_S2	12	163.1
1	train_S2	12	180.8
2	test_S1	12	203.5
3	train_S1	12	185.4

Key Finding:
Perfect sample separation — no overlap between conditions in the PCA feature space.

Invasion Ranking:
test_S1 > train_S1 > train_S2 > test_S2

---

4. Sample-Specific Temporal Dynamics

4.1 PCA Trajectories

test_S1 (Most Invasive)

- **PC1:** Fluctuates around 0–1, dips at *t8*, recovers later
- **PC2:** Starts high (≈3.5), gradually declines to 1.2
- **Pattern:** Active, early invasion with later consolidation.

train_S1 (Moderately Invasive)

- **PC1:** Negative (−6 to −1), sharp increase at *t4*
- **PC2:** Stable near 0

- **Pattern:** Low initial invasion, activation mid-experiment.

train_S2 (Dynamic Invasion)

- **PC1:** Rising trend (0.8 → 4.0), peaks at *t7*
- **PC2:** Fluctuates (2.0 → 0.2 → 1.8)
- **Pattern:** Progressive invasion with dynamic behavioral changes.

test_S2 (Late-Activated)

- **PC1:** Negative start, sharp rise to 3.6 at *t8*
- **PC2:** Consistently negative (−2 to −4)
- **Pattern:** Delayed activation, distinct behavioral profile.

---

5. Leader Cell Analysis

5.1 Cluster Characteristics: Leader Cell Distribution

Sample	Approx. Leader Cells	Invasion Behavior
train_S2	~8 (highest)	Most aggressive, leader-driven invasion
test_S1	~5.5	High invasion with moderate leader activity
train_S1	~5.0	Moderate leader cell presence
test_S2	~2 (lowest)	Least leader-driven, delayed invasion

5.2 Biological Significance

- **Leader Cells:** Pioneer cells that initiate and guide collective invasion fronts.
- **Implications from Analysis:**

- `train_S2`'s high leader cell count explains its dynamic, progressive invasion pattern.
- `test_S2`'s low leader cell count aligns with its delayed activation profile.
- **Leader cells are likely major contributors to PC1**, representing invasion spread and dispersion.

5.3 Leader Cell–PCA Correlation

Observation	Correlation
High leader cell count (train_S2)	Rising PC1 values, dynamic invasion
Low leader cell count (test_S2)	Negative PC2, delayed activation

Conclusion:
Leader cell activity is a **key mechanistic driver** of sample differentiation in PCA space, validating the biological relevance of the extracted features.

---

6. Feature Loading Insights

- **High positive loadings on PC1:**
  `n_leader_cells`, `mean_leader_radius`, `mean_radius`, `dispersion_index`
  → Represent overall invasion spread and leader-driven activity.
- **High loadings on PC2:**
  `mean_compactness`, `mean_extent`, `radius_skewness`
  → Represent morphological and structural invasion traits.

---

7. Temporal and Biological Interpretation

7.1 Temporal Phases

Phase	Timepoints	Description

Early Invasion	t0–t3	Initial spreading activity
Mid-phase Transition	t4–t7	Critical shift in invasion dynamics
Late Consolidation	t8–t11	Stabilization and structure formation

7.2 Biological Insights

- **test_S1:** Early peak and strong invasive phenotype.
- **train_S1:** Moderate invasion, activation mid-experiment.
- **train_S2:** Leader-driven, dynamically progressive invasion.
- **test_S2:** Unique, leader-independent delayed activation.

---

8. Technical Implementation

8.1 Data Processing Pipeline

1. Feature standardization (mean = 0, variance = 1)
2. PCA transformation using **scikit-learn**
3. **K-means clustering (k=4)**
4. Hierarchical clustering validation
5. Time-series trajectory analysis

8.2 Visualization Methods

- PCA scatter plots (by condition & cluster)
- Individual sample trajectories (PC1/PC2 vs. time)
- Feature loading heatmaps

- Scree plots (component selection)
- 2×2 comparative trajectory grids

8.3 Generated Output Files

File	Description
`pca_clustering_analysis.png`	Main PCA & clustering visualization
`pca_clustering_results.csv`	Numerical PCA and clustering results
`all_samples_pca_4grids.png`	Comparative PCA trajectory plots

---

9. Key Biological Conclusions

- Leader cells are central to invasion mode differentiation.
- Distinct invasion phenotypes correspond to specific PCA–cluster profiles.
- Dynamic and morphological metrics capture meaningful biological behavior.
- PCA and clustering confirm **clear separation** of experimental conditions.

---

10. Next Steps

1. **Feature-level analysis:** Plot raw invasion metrics for all samples.
2. **Biological correlation:** Link PCA patterns with experimental treatments.
3. **Validation:** Cross-check raw imaging data at key timepoints.
4. **Predictive modeling:** Build classifiers for invasion phenotypes.

5. **Temporal testing:** Perform within-sample statistical comparisons.

---

11. Technical Notes

- **Software:** Python (`pandas`, `matplotlib`, `scikit-learn`)
- **Approach:** Unsupervised learning with temporal component
- **Reproducibility:** All code and parameters logged
- **Validation:** Multiple clustering methods confirmed consistency

---

Summary

This project identified **distinct invasion phenotypes and temporal dynamics** across cancer cell samples.
Leader cell behavior emerged as a **major mechanistic factor** driving invasion variability, validated through PCA and clustering.
The analysis framework provides a **quantitative and biologically meaningful foundation** for deeper mechanistic exploration and therapeutic hypothesis generation.

Would you like me to format this into a **professional PDF report** (with sections, tables, and figure placeholders) or a **Word document** next?