Take-Home Exam: Multi-City Airbnb Market Analysis

EAS 510 - Basics of AI
Fall 2025

Due: Monday, November 17, 2025 at 11:59 PM

1 Assignment Overview

- **Weight:** 20% of Exam 2 grade
- **Format:** Jupyter Notebook with written analysis
- **Data:** Airbnb listings from multiple cities (provided on UB Learns)
- **Due:** Monday, November 17, 2025 at 11:59 PM

2 Learning Objectives

By completing this assignment, you will:

1. Apply PCA (Principal Component Analysis) to real-world datasets
2. Implement K-means clustering with appropriate data preprocessing

3 Assignment Task

You will apply PCA and K-means clustering to **ALL FOUR** Airbnb datasets:

- **Portland, OR** - Mid-sized Pacific Northwest city
- **Pacific Grove, CA** - Coastal California town
- **Albany, NY** - State capital and university town
- **Bozeman, MT** - Mountain town with university

3.1 Core Task

Compare how PCA components and clustering results differ across all four cities.

4 Technical Requirements

4.1 1. Data Preprocessing

- Apply StandardScaler to numeric features before PCA
- Handle missing values appropriately
- Remove obvious outliers (document your approach)
- Set a fixed `random_state` for reproducibility
- Ensure consistent file paths (datasets should be in the same directory as your notebook)

4.2 2. PCA Analysis

Required numeric features (exactly these):

```
numeric_columns = ['price', 'accommodates', 'bedrooms', 'beds', '
    bathrooms_text',
                   'review_scores_rating', 'review_scores_accuracy',
                   'review_scores_cleanliness', 'review_scores_checkin',
                   'review_scores_communication', 'review_scores_location',
                   'review_scores_value', 'number_of_reviews',
                   'availability_365', 'minimum_nights', 'maximum_nights']
```

- Determine components needed for 90%+ variance explained
- Interpret top 2 principal components for each city
- Create and include a scree plot (variance explained)

4.3 3. K-Means Clustering

- Use first 2 principal components
- Apply elbow method to find optimal k
- Calculate silhouette score
- Describe cluster characteristics
- Include a scatter plot showing clusters by city

4.4 4. Comparison

- Compare PCA components between your cities (discuss loadings and key variable contributions)
- Compare clustering performance (silhouette scores)
- Identify key differences in structure or interpretability

5 Expected Deliverables

5.1 Jupyter Notebook (.ipynb)

Your notebook should include:

5.1.1 Section 1: Data Loading & Preprocessing

- Load datasets for all 4 cities
- Apply StandardScaler and handle missing values/outliers
- Show basic statistics

5.1.2 Section 2: PCA Analysis

- Run PCA on all 4 cities
- Interpret top 2 components for each city
- Show variance explained (include scree plot)

5.1.3 Section 3: K-Means Clustering

- Apply clustering using first 2 PC components for each city
- Use elbow method and calculate silhouette scores
- Describe clusters found in each city
- Include scatter plots of clusters by city

5.1.4 Section 4: Cross-City Comparison & Conclusions

- Compare PCA results across all cities
- Compare clustering performance
- Summarize key patterns and differences (2–3 paragraphs)

5.2 File Naming and Submission Format

- Submit a single Jupyter Notebook named: `lastname_firstname_takehome.ipynb`
- Ensure all code runs without errors before submission
- All plots and tables should render inline

5.3 Environment Requirements

Use Python 3.10+ and the following packages:

- pandas, numpy, scikit-learn, matplotlib, seaborn

Component	Excellent (90–100%)	Good (80–89%)	Satisfactory (70–79%)	Below (<70%)
Data Preprocessing (30%)	Proper scaling, outlier handling, well-documented	Good preprocessing with minor issues	Basic preprocessing completed	Incomplete or incorrect preprocessing
PCA Analysis (35%)	Correct implementation, clear interpretation, scree plot included	Good PCA with mostly correct interpretation	Basic PCA with some interpretation	Misapplied or unclear PCA
K-Means Clustering (25%)	Optimal k, good metrics, cluster visualization	Good clustering with appropriate methods	Basic clustering analysis	Major issues or missing elements
Comparison & Code Quality (10%)	Clear comparison, clean code, well-commented notebook	Adequate comparison and code quality	Basic comparison and acceptable code	Poor documentation or disorganized code

6 Evaluation Rubric

7 Data Files

The following datasets are provided as attachments on UB Learns:

- `portland_listings.csv` - Portland, OR
- `pacific_grove_listings.csv` - Pacific Grove, CA
- `albany_listings.csv` - Albany, NY
- `bozeman_listings.csv` - Bozeman, MT

Example notebook: See `take_home_exam_example-2.ipynb` attached to this assignment on UB Learns for a complete Portland analysis example!

8 Submission Instructions

1. Complete your analysis in a Jupyter Notebook
2. Ensure all code runs without errors
3. Include clear markdown explanations for each section
4. Submit your `.ipynb` file via UB Learns
5. **Deadline:** Monday, November 17, 2025 at 11:59 PM

9 Academic Integrity

This is an individual assignment. You may:

- Use course materials, textbooks, and documentation
- Discuss general concepts with classmates
- Use online resources for technical implementation help

You may **NOT**:

- Share code or analysis with other students
- Copy solutions from online sources
- Submit work that is not substantially your own

Good luck with your analysis!