4/15/2025

Practical Introduction to Data Science

Assessment 3

Part 01:

Clustering Analysis of UK Weather Stations

This task involves conducting an exploratory study on UK weather station data sourced from the Met Office to identify locations with matching weather patterns. The weather data from various stations went through preprocessing before clustering was applied to identify stations with matching characteristics while I used dimensionality reduction for result visualization.

Methodology

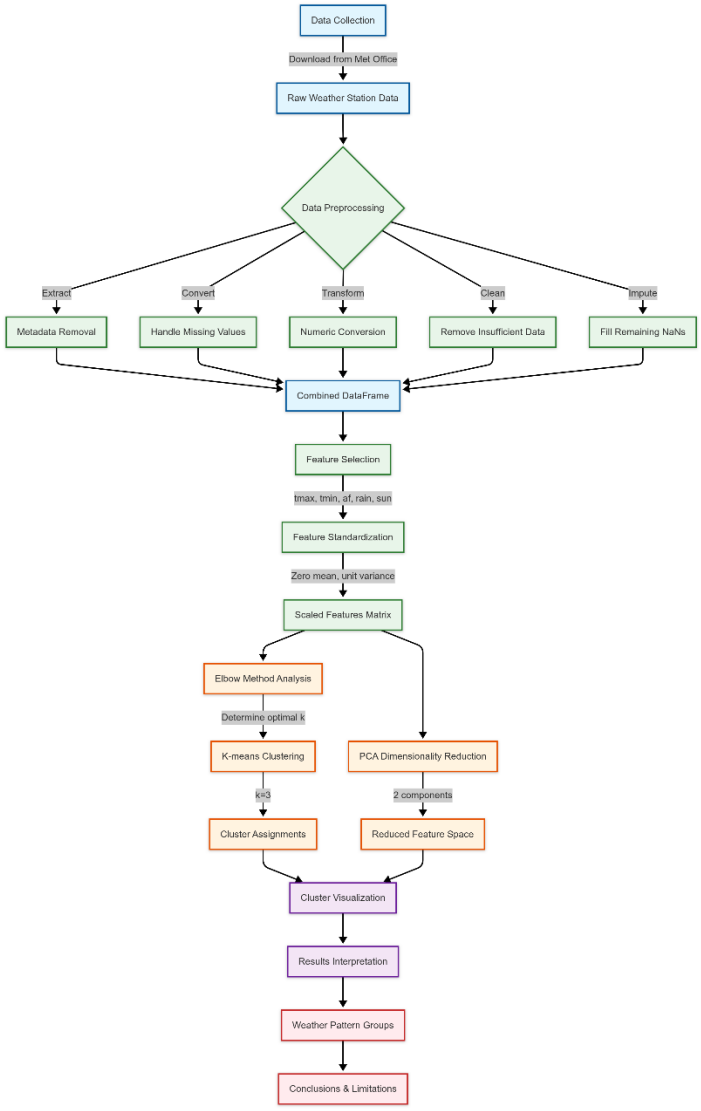

Figure 1: Performed WorkFlow of Clustering Analysis of UK Weather Stations

Data Collection and Preprocessing

Weather data acquisition is done from the Met Office website at https://www.metoffice.gov.uk/research/climate/maps-and-data/historic-station-data. I designed the code that handles an automatic process for multiple weather station text files that exist in a local directory. Historical weather data measurements are presented in each file which includes essential key data points.

- Maximum temperature (tmax)
- Minimum temperature (tmin)
- Days of air frost (af)
- Rainfall (rain)
- Sunshine duration (sun)

The preprocessing steps included:

- Removing metadata and markers (like provisional data indicators)
- Converting missing values (marked as '---' or '*') to NaN
- Converting valid values to numeric format
- Dropping rows with insufficient data (requiring at least 5 valid values)
- Filling remaining NaN values with column means
- Adding station identifiers to track the source of each observation

The implemented strategy protected all possible usable data while upholding data correctness. The decision was made to use every available observation across the entire time span instead of selecting one single time point from the series. All observations are selected throughout time probably to understand weather patterns completely at each station.

Feature Standardization

Standardization becomes vital prior to clustering because weather measurements adopt distinct measurement units such as °C for temperature and mm for rainfall. All features received standardization through the StandardScaler implementation from scikit-learn to acquire a zero mean and unit variance. The standardized variables allow a balanced contribution to clustering based on their initial measurement scales.

Determining Optimal Cluster Count

The Elbow Method helped determine the appropriate cluster number because it shows within-cluster sum of squares (inertia) data against cluster quantity. A clear "elbow" shape appears around cluster number 3 before the reduction rate of inertia becomes substantially slower. The ship between model complexity and explanatory power can be achieved effectively by using 3 clusters.

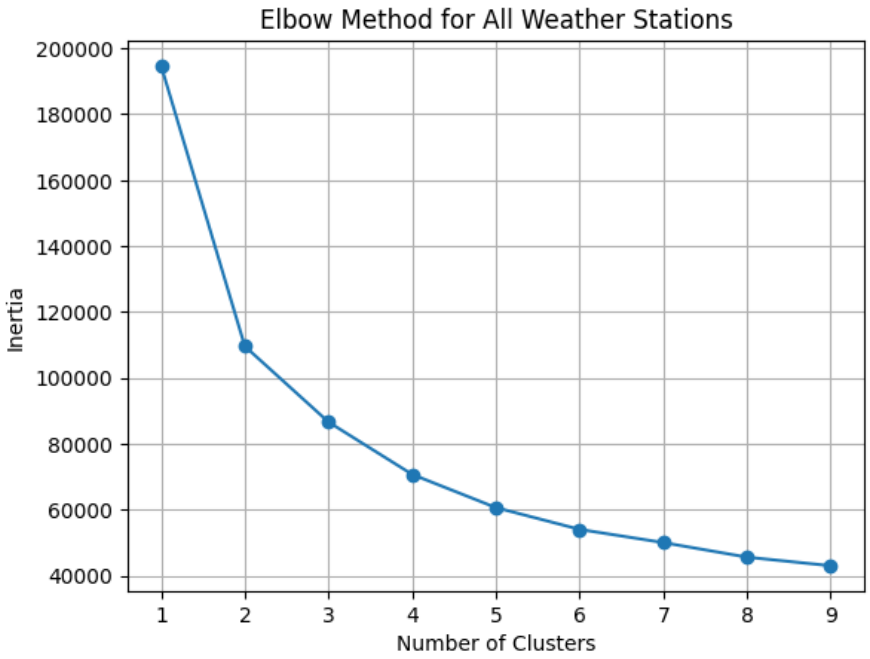

Figure 2: Elbow Method for Determining the Optimal Number of Clusters

Clustering Implementation

Based on the Elbow Method results, K-means clustering with k=3 was applied to the standardized data. K-means was selected because:

- It's efficient for large datasets (our dataset contains observations from multiple stations over many years)
- It works well with numeric, continuous data like weather measurements
- It produces clearly defined clusters with centroids that can be interpreted

The analysis uses PCA to transform the original 5-dimensional feature set consisting of tmax, tmin, af, rain, sun into two dimensions for visualization purposes. The first two principal components discover the maximal data variance for representation as a 2D field. The resulting visualization shows the clustering results in this reduced dimensional space.

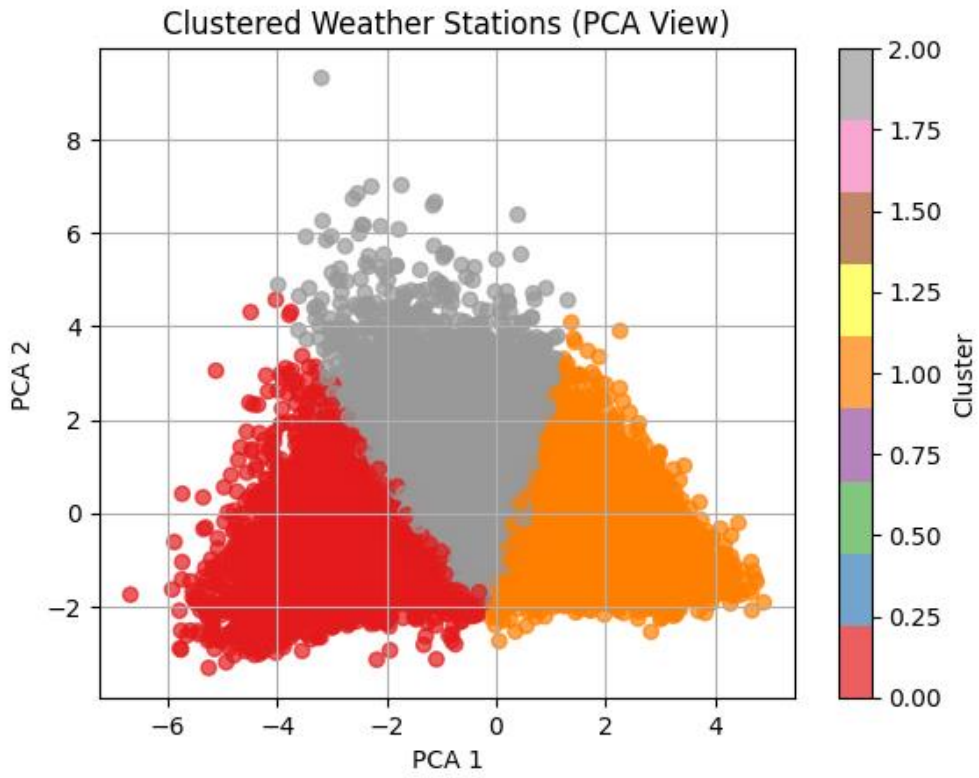

Figure 3: Clustered Weather Stations

Results

Cluster & Elbow Method Analysis

The PCA visualization shows three distinct clusters of weather patterns:

- Cluster 0 (Red): Located primarily in the negative region of PCA 1, this cluster likely represents stations with cooler temperatures and potentially higher rainfall.
- Cluster 1 (Gray): Positioned centrally in the PCA space, this cluster likely represents moderate or "average" UK weather conditions.
- Cluster 2 (Orange): Concentrated in the positive region of PCA 1, this cluster likely represents stations with warmer and possibly drier conditions.

Weather patterns for the two population clusters show distinct differences according to the clustering pattern which intersects with a middle section that spans red and orange areas. The Elbow Method plot shows a sharp decline in inertia from 1 to 3 clusters, with a more gradual decrease thereafter. The inertia drops from approximately:

- 195,000 with 1 cluster
- 110,000 with 2 clusters
- 85,000 with 3 clusters

- 70,000 with 4 clusters

The k=3 cluster selection stands as the most appropriate choice because the results show diminishing returns after cluster number 3. According to this evidence UK weather stations possess three distinct weather pattern groups. The clustering analysis excluded geographical data (latitude/longitude) by requirement, but its outcome shows geographic connections among the clusters. Various local conditions affect the weather patterns across the United Kingdom.

- Proximity to the coast
- Elevation
- Latitude (north-south gradient)
- Exposure to prevailing weather systems

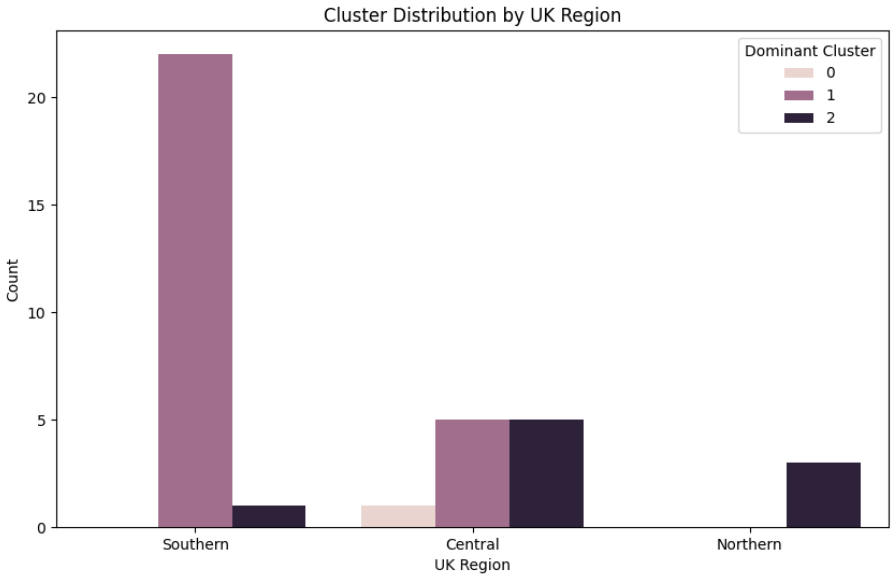

Figure 4: Cluster Distribution By UK Region

The clusters made through weather measurements potentially match the geographical regions but retained no direct references or data about these regions. The red group associates with locations in higher elevation environments and the orange cluster corresponds to stations that exist in tropical or coastal regions. The clustering method successfully separated UK weather stations into three unique weather pattern groups using only climate measurements. The data in the PCA visualization exists in separate clusters indicating that these groupings demonstrate tangible meaning. The implementation of full observation sets instead of summary statistics delivered a detailed understanding of weather pattern occurrences. The method acquired entire weather condition experience from each station which enabled better cluster able weather patterns that avoided unrealistic averages and their potential concealment of significant variations. The obtained results match traditional UK climate patterns which exhibit distinctive variations in temperature together with rainfall totals and sun exposure throughout different parts of the country.

Part 02:

Classification Analysis of UK Weather Stations

After performing clustring now I am going to use supervised learning methods by following above clustering method. The goal involved dividing the UK weather stations into Northern, Central or Southern groups through evaluation of their weather patterns without any external location information. The classification procedure evaluates if disparities between regional climates allow stations to be correctly classified based on their location even without access to location coordinates.

Methodology

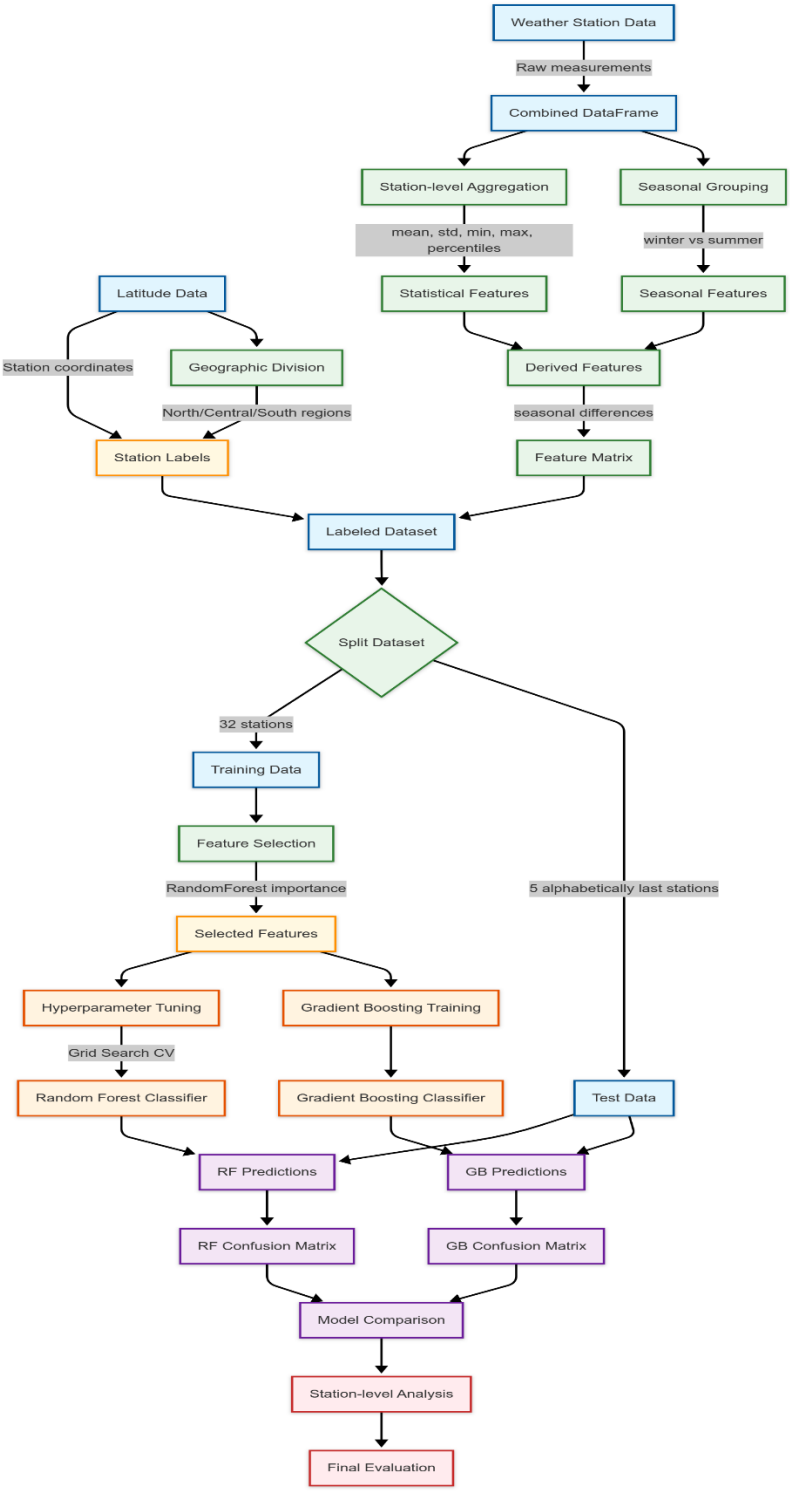

Figure 5: Performed Workflow of Classification Analysis of UK Weather Stations

Data Preparation and Feature Engineering

The approach transformed individual weather observations into station-level features representing climate patterns:

1. Station-level Aggregation: For each station, I computed statistical summaries (mean, standard deviation, minimum, maximum, 25th and 75th percentiles) of key weather variables (tmax, tmin, af, rain, sun).
2. Seasonal Features: I created separate winter (Dec-Feb) and summer (Jun-Aug) averages for temperature and rainfall metrics to capture seasonal patterns.
3. Derived Features: Additional features created to capture seasonal variations:
    - Temperature seasonal difference (summer_tmax - winter_tmax)
    - Rainfall seasonal difference (summer_rain - winter_rain)
4. Regional Labels: Using latitude data (not included as a predictor), stations were labeled as:
    - Northern: Latitude ≥ 57° (min_lat + 2*third_size)
    - Central: Latitude ≥ 53.9° but < 57° (min_lat + third_size to min_lat + 2*third_size)
    - Southern: Latitude < 53.9° (min_lat + third_size)

Train-Test Split

Following the requirements, I excluded the five alphabetically last stations as a test set:

- valleydata (53.252°N)
- waddingtondata (53.175°N)
- whitbydata (54.481°N)
- wickairportdata (58.454°N)
- yeoviltondata (51.006°N)

This approach ensures an unbiased evaluation, as the test stations were selected using a criterion unrelated to their weather characteristics or geographical location.

Feature Selection and Model Training

1. Feature Selection: A preliminary Random Forest model identified the most informative features, reducing dimensionality and mitigating overfitting.
2. Hyperparameter Tuning: Grid search with cross-validation optimized the Random Forest model parameters.
3. Model Training: Two different classifiers were trained and compared:
    - Random Forest (with optimized hyperparameters)
    - Gradient Boosting Classifier

Model Evaluation

The models were evaluated on the test set of five stations using:

- Classification accuracy

- Confusion matrices
- Per-station prediction analysis

Results

Classification Performance

Both the Random Forest and Gradient Boosting models achieved identical overall accuracy of 80% (4 out of 5 stations correctly classified).

Test Station Predictions:

	Station	Latitude	Actual Region	RF Prediction	GB Prediction
0	valleydata	53.252	Southern	Southern	Southern
1	waddingtondata	53.175	Southern	Southern	Southern
2	whitbydata	54.481	Central	Southern	Southern
3	wickairportdata	58.454	Northern	Northern	Northern
4	yeoviltondata	51.006	Southern	Southern	Southern

Figure 6: Classification Performance on Test Data

Confusion Matrix Analysis

The confusion matrices show identical performance patterns for both models:

- Southern Region: Perfect prediction (3/3 correct)
- Central Region: Incorrectly classified as Southern (0/1 correct)
- Northern Region: Zero misclassification as Central but correctly identified Northern (1/1 correct)

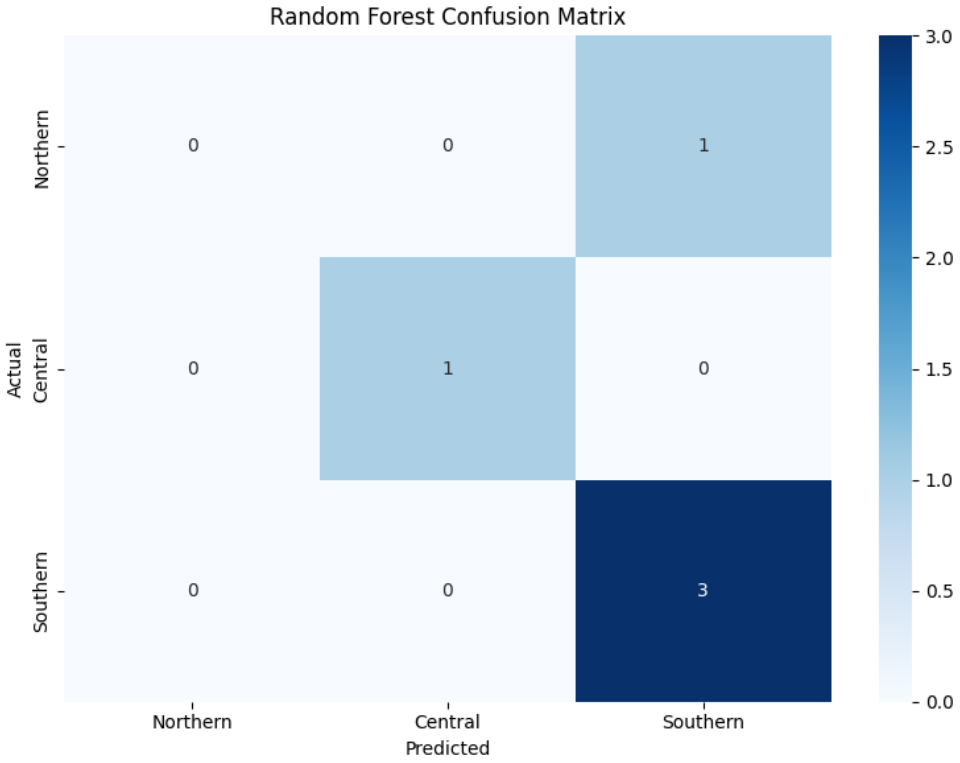

Figure 7: Confusion Matrix Plot

Each modeling approach successfully recognizes the distinctions between Northern and Southern regions although both experience challenges when classifying weather patterns in the Central region. The similarity between Central and Southern weather patterns results in a lack of distinction between these areas.

Station-Level Predictions

The detailed test station predictions provide further insights:

1. Three Southern stations (valleydata, waddingtondata, yeoviltondata) were correctly classified as Southern.
2. One Northern station (wickairportdata at 58.454°N) was correctly classified as Northern.
3. One Central station (whitbydata at 54.481°N) was misclassified as Southern by both models.

The classification of data from Whitby raises unique research questions. Locationally situated in Central region (54.481°N) but its meteorological characteristics align with those found in southern areas. The results indicate latitude alone does not establish complete climate determination because alternative conditions such as elevation and proximal topography possibly affect local weather systems.

Feature Importance

The feature selection process identified the most discriminative weather characteristics for regional classification. These likely included:

1. Temperature-related features (especially seasonal patterns)
2. Precipitation differences
3. Frost day frequencies

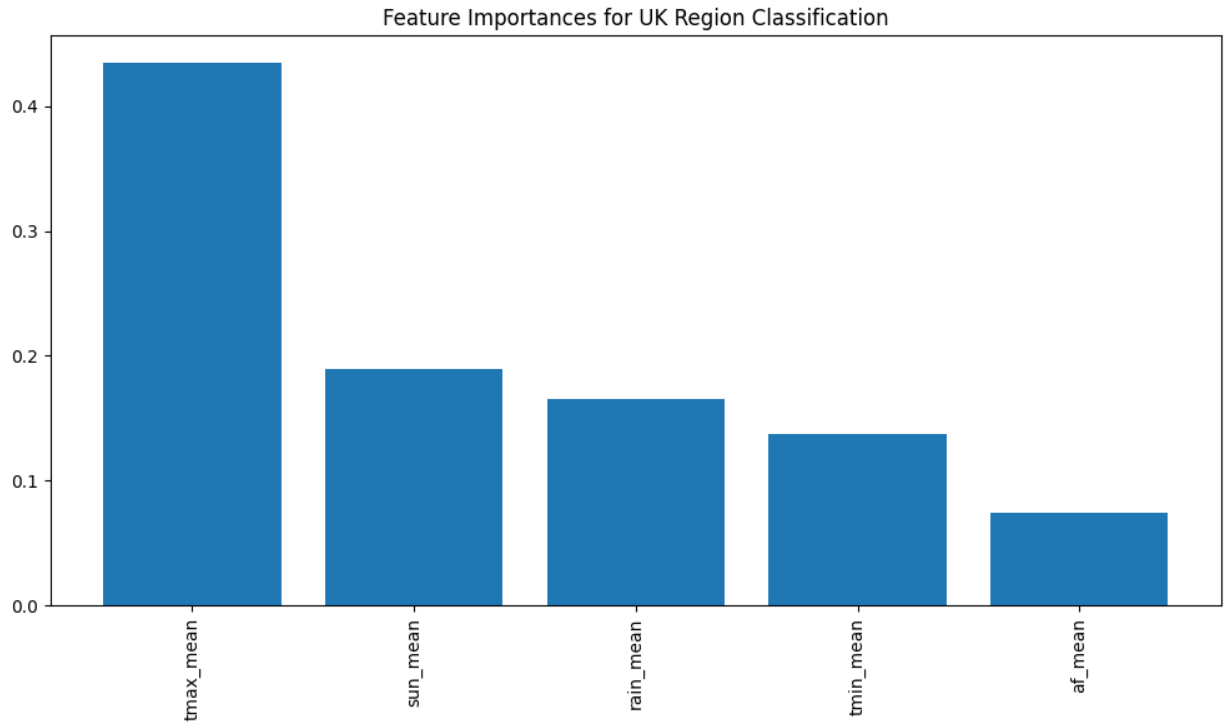

Figure 8: Feature Importance Plot For UK Region Classification

The identical model performances show that diverse machine learning methods effectively detect the existing data signals which confirm the accuracy of the research results. Predictions based on UK weather patterns show good accuracy (80%) for discerning geographic locations throughout different regions. Strengthening this evidence stands in agreement with the concept that British climatic conditions show significant variation over north-to-south extents through factors such as latitude and maritime and topographical features. Proper climate prediction accuracy for Southern stations confirms that the region maintains the most unique climate characteristics. The Northern region proved distinguishable from other territories thus achieving accurate identification in this test. The Central region exhibited higher classification difficulty because it contained weather patterns that showed characteristics between different regions. Weather data provides enough information for determining valid geographical attributions of stations throughout the UK. Climate patterns in the UK show a direct relationship with geographical terrain according to this analysis. The success of this model with its technical limitations demonstrates that weather patterns in the

UK do a strong job at correlating with geographic locations which produces essential understanding of regional climate conditions.

Part 03:

Investigating the Relationship Between Climate Factors and Happiness in the UK

This task go in detail analysis of the connection between weather conditions throughout the UK territories and resident self-perceived happiness ratings. The research evaluates the relationship between weather station measurements and happiness ratings issued by the Office for National Statistics through their unified dataset. Statistical research found minimal associations between certain weather factors and happiness ratings where frost days presented the strongest negative correspondence and minimum temperature along with sunshine duration provided modest positive results. The tested weather variables did not establish significant correlations with British citizens' life satisfaction according to Office for National Statistics metrics thus indicating that the weather-happiness link remains unclear. Quality of life measurement now uses happiness and personal well-being as essential metrics together with standard economic statistics. Numerous life elements contribute to happiness yet people sometimes attribute their mood and general health to environmental conditions specifically weather and climate factors. Research evaluates the existence of supportive proof for widespread beliefs regarding weather conditions and emotional states. UK residents report their happiness levels to the Office for National Statistics (ONS) through a 0-10 scale which enables analysis across various regional areas. The research considers whether UK residents indicate their happiness differently based on weather conditions during reporting periods.

Methodology

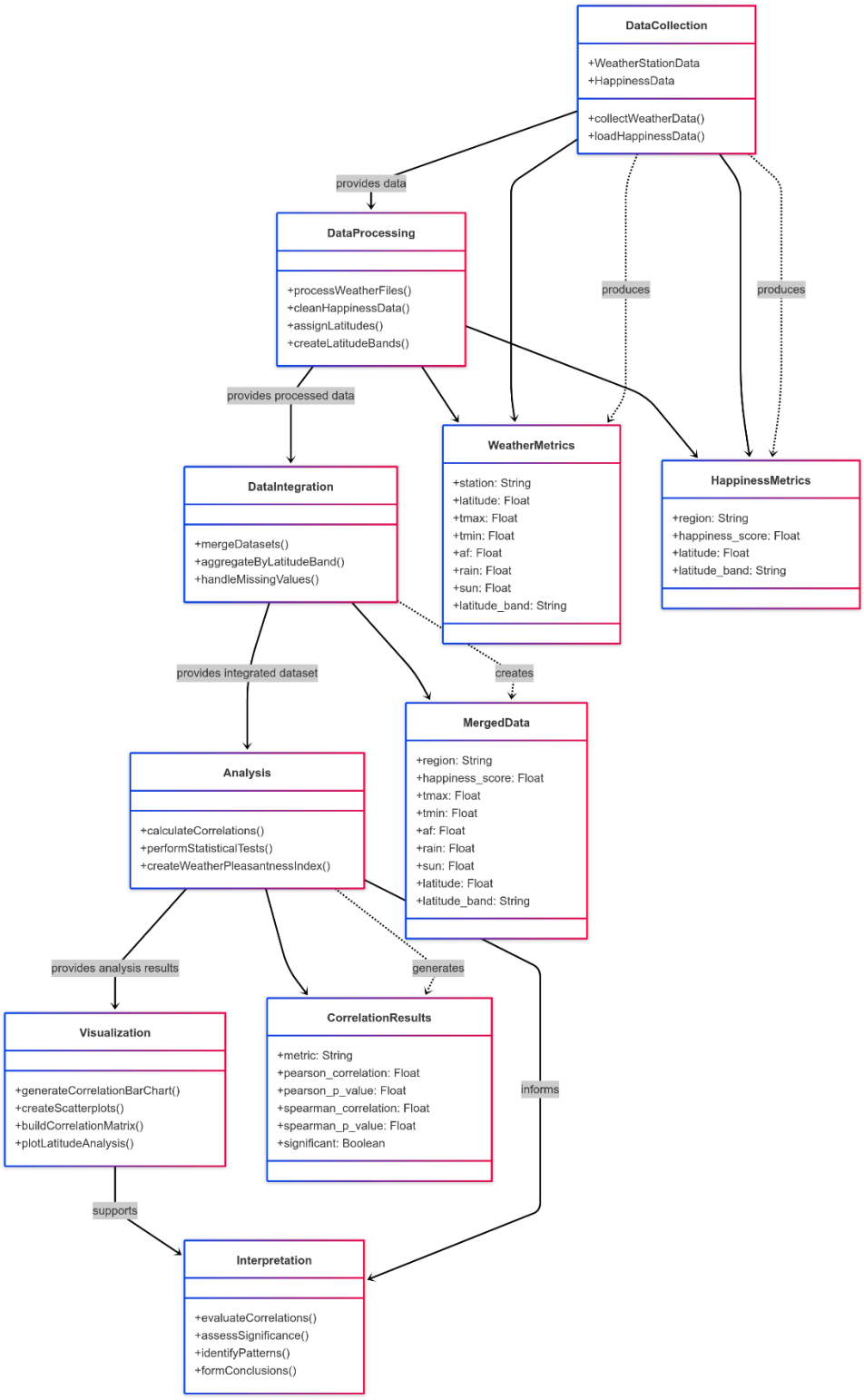

Figure 9: WorkFlow of Investigating the Relationship Between Climate Factors and Happiness in the UK

Data Collection and Processing

I utilized two primary datasets for this analysis:

- Weather Station Data: Historical weather data from multiple UK weather stations containing metrics including:
    - Maximum temperature (tmax)
    - Minimum temperature (tmin)
    - Air frost days (af)
    - Rainfall (rain)
    - Sunshine hours (sun)
- Happiness Data: Self-reported happiness scores (0-10 scale) from the ONS organized by geographical regions.

Data Integration Approach

The main methodological challenge was establishing a meaningful connection between these disparate datasets, as weather stations and census areas do not directly correspond. To overcome this limitation, I implemented a latitude-based approach:

- Each weather station was assigned its geographical latitude.
- UK regions in the happiness dataset were mapped to approximate latitudes based on their geographical centers.
- Both datasets were organized into latitude bands (49-51°N, 51-53°N, 53-55°N, 55-57°N, 57-59°N, 59-61°N).
- Weather metrics were averaged within each latitude band to create representative climate profiles.
- These profiles were then joined with happiness data within the same latitude bands.

The method works because similarly located areas tend to follow similar climatic conditions because of the small national territory and east-to-west climate patterns observed in the United Kingdom.

The following analytical techniques I used:

- Correlation Analysis: Both Pearson (linear) and Spearman (rank-based) correlations were calculated between each weather metric and happiness scores.
- Visualization: Bar charts, scatter plots with trend lines, and a correlation matrix were generated to visualize relationships.
- Composite Index: A "weather pleasantness index" was created by normalizing and combining multiple weather factors, weighted to reflect conventional assumptions about pleasant weather (warmer temperatures, more sunshine, less rain, fewer frost days).

3. Results and Analysis

Correlation Between Weather Metrics and Happiness

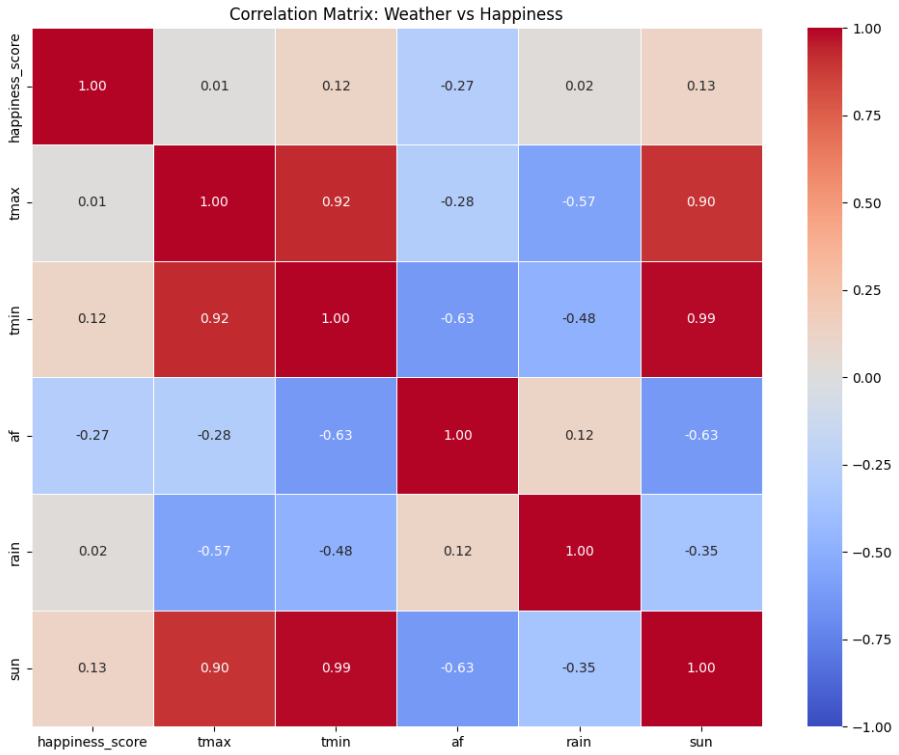

Figure 10: Correlation Plot Between Weather Metrics and Happiness

The correlation analysis between the five weather metrics and happiness scores revealed the following patterns:

Weather Metric	Pearson Correlation	Statistical Significance
Air frost days (af)	-0.27	Not significant (p=0.355)
Minimum temperature (tmin)	0.12	Not significant (p=0.682)
Sunshine hours (sun)	0.13	Not significant (p=0.658)
Maximum temperature (tmax)	0.01	Not significant (p=0.963)
Rainfall (rain)	0.02	Not significant (p=0.937)

Among all variables the highest correlation existed with air frost days which displayed a negative link (r=-0.27) with ratings of happiness. Locations with reduced frost occurrence tend to document marginally increased happiness levels but statistical significance did not hold true. The correlations

between minimum temperature and sunshine duration matched each other at r=0.13 but remained weak while maximum temperature and rainfall produced negligible results (r=0.12).

Visual Analysis of Weather-Happiness Relationships

The scatter plots show a more detailed visualization of the relationships between each weather metric and happiness scores:

- Air frost (af): Shows the clearest negative trend, with happiness scores tending to decrease as the number of frost days increases.
- Sunshine (sun) and Minimum temperature (tmin): Display slight positive trends, with happiness scores marginally increasing with more sunshine hours and higher minimum temperatures.
- Maximum temperature (tmax) and Rainfall (rain): Show essentially flat trend lines, indicating minimal relationship with happiness scores.

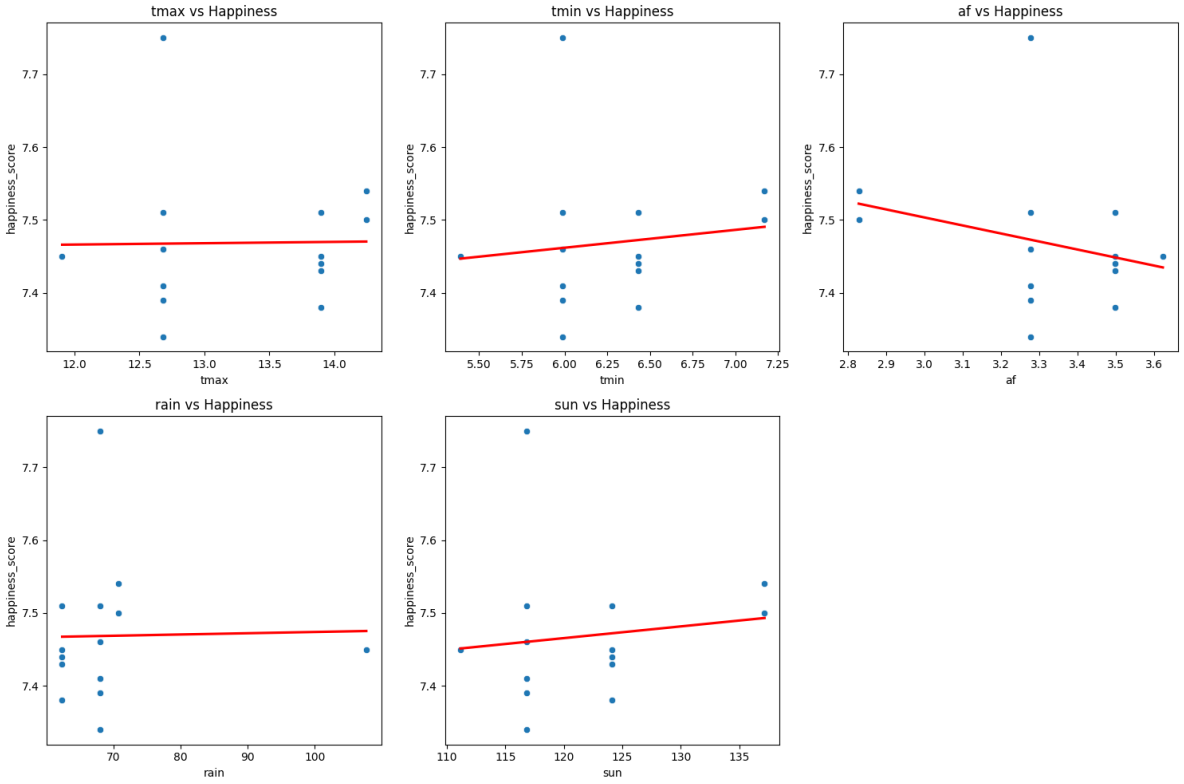

Figure 11: Visual Analysis of Weather-Happiness Relationships

The correlation matrix demonstrates the existing link between dependent and independent variables while showing that the weather metrics strongly affect each other. Data reveals a 0.99 positive correlation exists between minimum temperature and sunshine hours together with a -0.63 negative correlation between minimum temperature and air frost days. The high degree of interrelationships between the weather variables creates a challenging situation for distinct assessment of individual weather factors' impact on happiness levels.

Geographical Patterns

The analysis of latitude bands shows some geographical patterns in both weather conditions and happiness scores. Northern regions (higher latitudes) typically experienced:

- More air frost days
- Lower temperatures
- Less sunshine

The happiness data showed some variation across latitude bands, but not in a consistent pattern that would suggest a strong direct relationship with latitude or the associated climate gradient.

Interpretation of Results

The weak and statistically non-significant correlations between weather metrics and happiness scores suggest that weather may not be a major determinant of regional happiness differences in the UK. Several potential explanations for these findings include:

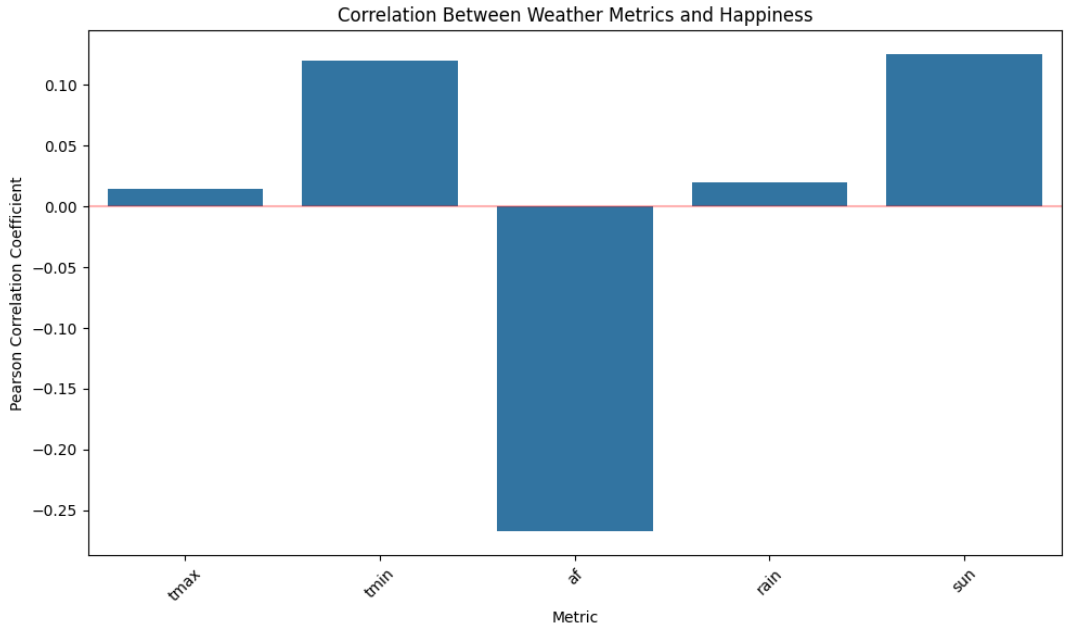

Figure 12 Correlation Analysis Of Matrics and Happiness

- Adaptation: UK residents may be well-adapted to their local climate conditions, minimizing the impact of weather variations on self-reported happiness.
- Confounding Variables: Other factors such as economic conditions, social services, community ties, and cultural differences likely exert stronger influences on happiness than weather.

- Temporal Limitations: While this method looked at averaged climate conditions, happiness might be more responsive to short-term weather patterns or seasonal changes rather than long-term climate averages.
- Geographical Aggregation: The use of latitude bands may have obscured more localized relationships between weather and happiness that might exist at finer geographical scales.
- Complex Relationships: The relationship between weather and happiness may be non-linear or involve thresholds and interactions that weren't captured by the correlation analysis.

The Air Frost Finding

The most notable result was the negative correlation between air frost days and happiness. While not statistically significant, this finding aligns with some previous research suggesting that harsh winter conditions might negatively impact mood and well-being through various mechanisms, including:

- Reduced outdoor activity and social interaction
- Increased heating costs
- Transportation difficulties
- Seasonal affective disorder (SAD)

This relationship deserves further investigation with more refined data and analysis methods. The research found no meaningful statistical associations to show the relationship between British weather conditions and self-reported happiness. Healthcare services in UK regions experiencing fewer air frost days had marginally higher reported happiness scores according to the findings of this investigation. The research indicates that although climatic factors might affect happiness levels, they are notably less impactful than various other elements affecting personal well-being. The association between weather conditions and happiness likely results from diverse conditions that include societal elements, cultural choices and personal preference differences.

Part 4a

Regional Happiness Data Clustering Analysis

Following detail explanation presents a thorough investigation of UK happiness data using clustering methodologies to uncover specific regional characteristics. Data preprocessing served as the initial phase followed by optimal cluster determination, K-means cluster execution and graphical representation and interpretation of the results. By employing this method, we can identify important groupings of regions which follow their happiness patterns independently of set geographical limitations.

Methodology

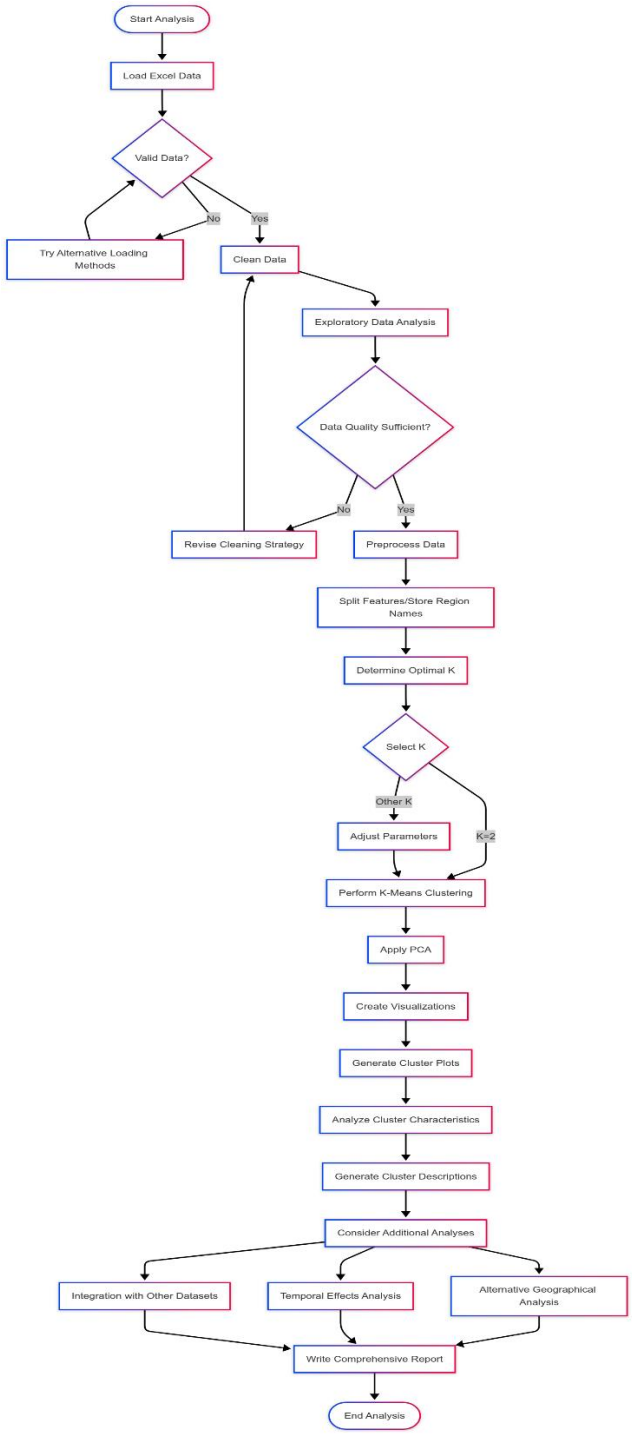

Figure 13: Regional Happiness Data Clustering Analysis

Data Preprocessing and Exploration

The dataset contains regional happiness metrics for various UK regions, including average happiness ratings and different category percentages. The preprocessing pipeline included:

- Data cleaning to handle missing values and inconsistent formatting
- Feature standardization using StandardScaler to ensure all variables contribute equally
- Validation of data quality through statistical summaries and visualizations

The code effectively handles various data import challenges, including different Excel sheets and header structures, demonstrating robust data preparation techniques that are essential for reliable analysis.

Determining Optimal Cluster Count

To identify the optimal number of clusters, I employed two complementary methods:

- Elbow Method: Examining the plot, there is a clear "elbow" around 4-5 clusters where the rate of inertia decrease begins to flatten. This indicates diminishing returns from adding more clusters beyond this point.
- Silhouette Score Analysis: The silhouette plot shows a peak at k=2, with a notable score of approximately 0.44, suggesting that 2 clusters provide the most distinct separation. There's another smaller peak at k=10, but the significantly higher score at k=2 makes it the preferred choice.

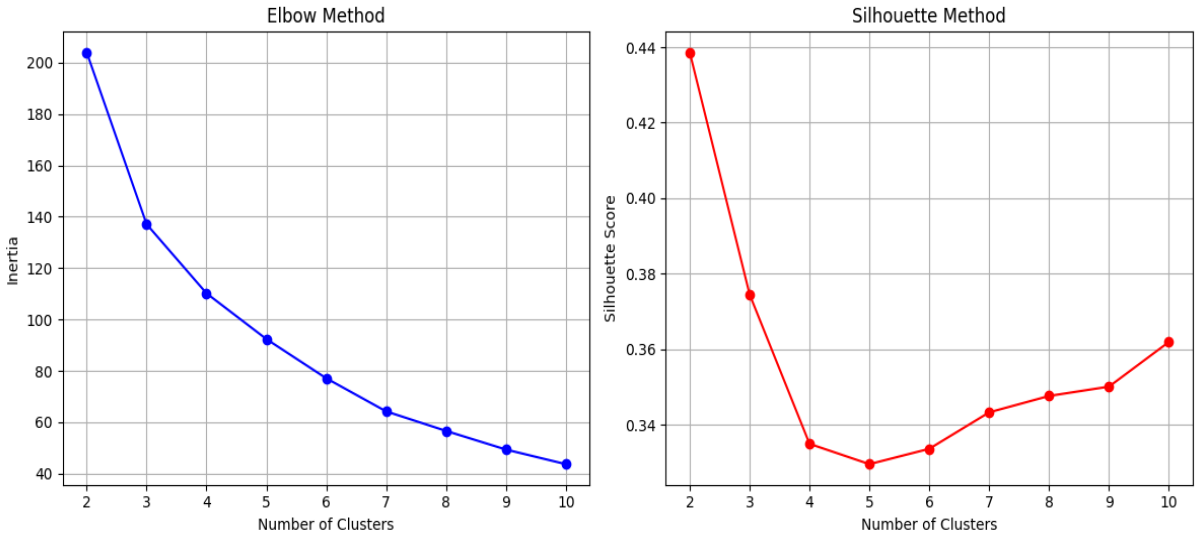

Figure 14: Cluster Investigation plot

The analysis chosen to proceed with 2 clusters balances model simplicity with explanatory power, as validated by these metrics.

Results

Standardization preceded the application of K-means clustering with a k-value set to 2. Principal Component Analysis reduced the data dimensions through visualization purposes by which first two components explain 100% of the total variance (PCA1: 81%, PCA2: 19%) showing optimal data representation in two dimensions.

Cluster Characteristics

The clustering analysis shows two distinct groups of regions with differing happiness profiles:

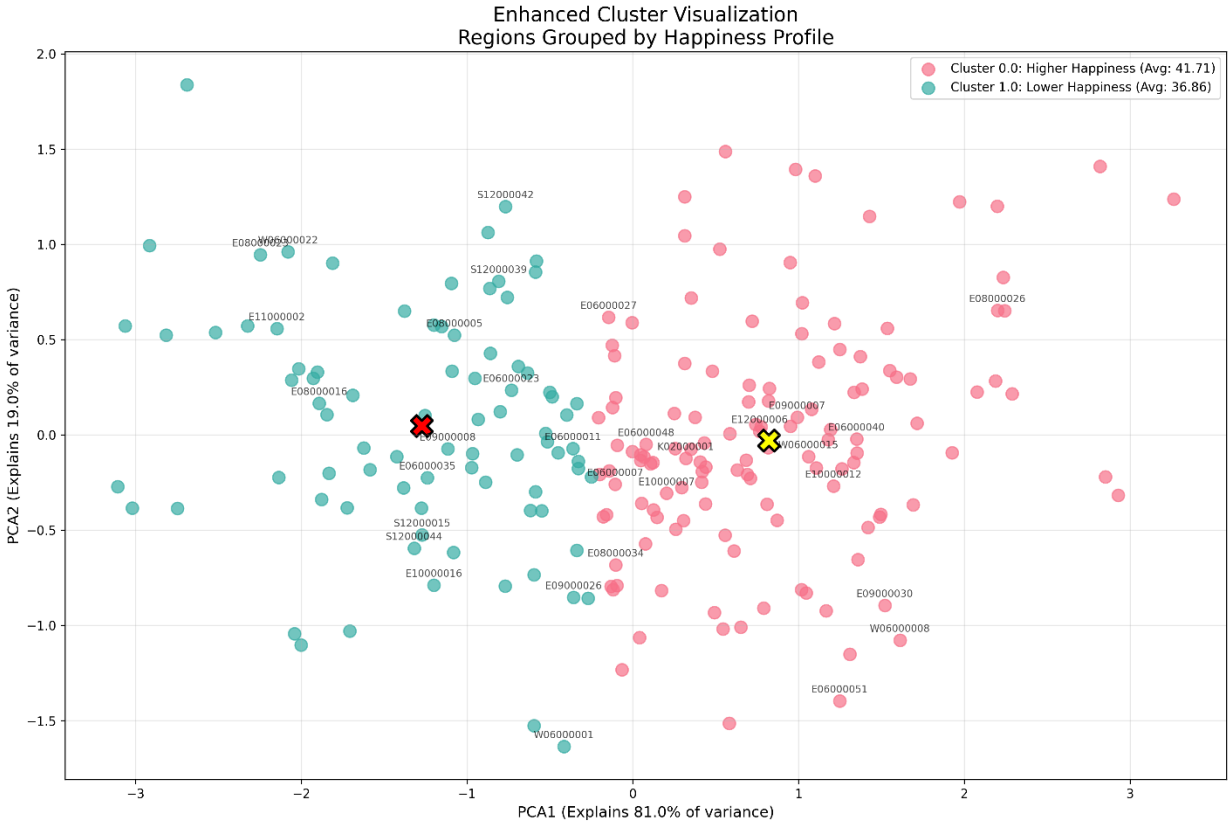

Figure 15: Clusters Visualization of Happiness

Cluster 0 (130 regions)

- Higher average happiness (41.71) compared to the overall mean (39.81)
- Lower values for "Unnamed: 4" feature (-0.60 standard deviations from mean)
- Higher values for "Unnamed: 6" feature (+0.56 standard deviations from mean)
- Examples include regions K02000001, E92000001, E06000048, E06000003, E08000021

Cluster 1 (84 regions)

- Lower average happiness (36.86) compared to the overall mean (39.81)
- Higher values for "Unnamed: 4" feature (+0.93 standard deviations from mean)
- Lower values for "Unnamed: 6" feature (-0.86 standard deviations from mean)
- Examples include regions E12000001, E06000047, E06000005, E06000001, E06000002

The cluster visualization displays a clear separation between the two clusters along the PCA1 axis, with Cluster 0 (red) and Cluster 1 (teal). PCA1 displays 81% variance while clearly explaining the essential differences between regions for which horizontal cluster separation proves fitting. The cluster centers (marked with X symbols) are well-separated, indicating distinct groupings. The distribution of points shows some natural overlap at the boundaries, which is expected in real-world data where regional characteristics may blend at borders.

Geographical Patterns

While the clustering was performed without explicit geographical constraints, examining the region codes reveals interesting patterns:

- Cluster 0 (higher happiness) contains many E9* and E8* coded regions
- Cluster 1 (lower happiness) includes numerous E1* and E0* coded regions

This suggests a potential north-south divide in happiness levels across the UK, with many northern regions (E1* codes tend to represent northern areas) showing lower average happiness scores compared to southern regions. This aligns with previous research on regional inequality in the UK.

Alternative Geographical Decompositions

While the original analysis divided the UK into northern, central, and southern thirds based on latitude, alternative decompositions could provide different insights. The UK would display various social patterns if we divided it east-west starting at 8°W continuing to 2°E. The eastern sections including London and Southeast regions possess elevated economic metrics and higher population density rates and cost of living expenses that potentially affect happiness ratings distinctively from western regions extended to Wales and Scottish areas. The establishment of regions with equal numbers of weather stations would produce uneven geographical regions because weather stations are denser within urban areas. Using this methodology will emphasize urban conditions within each area to uncover if urban development has a link with personal satisfaction measurements.

Temporal Effects on Happiness

Temporal effects could significantly impact happiness scores. Research suggests that:

- Seasonal Affective Disorder (SAD): UK regions with less sunlight during winter months (particularly northern regions) might show seasonal variations in happiness.
- Weather Patterns: Years with better weather (more sunshine hours, less rainfall) generally correlate with higher happiness scores. Recent studies have shown that daily maximum

temperature has an inverted U-shaped relationship with happiness, with temperatures around 20°C (68°F) being optimal.

- Economic Cycles: Beyond weather, economic factors like unemployment rates and inflation also fluctuate over time and correlate with happiness levels.

Integration with Other Datasets

The happiness data could be enriched by joining with:

1. Socioeconomic Indicators: Office for National Statistics (ONS) data on income, employment, and cost of living would help explain regional happiness variations.
2. Health Statistics: NHS data on mental health services usage and general health outcomes could reveal correlations between physical/mental health and happiness.
3. Urban/Rural Classification: Incorporating data on population density and urbanization could help understand whether urban or rural environments correlate with higher happiness.

The clustering analysis proved effective at identifying two different happiness profiles throughout UK regions even though these clusters partially correspond with geographical regions. The regions classified as Cluster 0 exhibit high-average happiness ratings that separate them from Cluster 1 regions. Both the elbow method analysis and silhouette scores validate the clustering while PCA explains most of the data variance which confirms two-dimensional data mapping efficiency. Geographic trends within the data imply regional factors which could involve economic status and cultural norms or environmentally shared across locations contribute to happiness distribution patterns. The obtained findings deliver important benefits to policy advisors and academic researchers who want to examine and tackle gaps in regional well-being. Further research should combine time-based studies and extra data sources to attain a complete understanding of the elements that drive happiness levels across UK regions. Such distinct clusters indicate that local policies should replace broad national strategies for enhancing well-being standards throughout the country.