GROUP 4: CAMY AND RENA

THE IMPACT OF CLUSTERING ON REGRESSION MODELS

AGENDA

(1) **Problem Statement**

(2) **Data Processing and Cleaning**

(3) **Exploratory Data Analysis (EDA):** insights gained from our data exploration.

(4) **Modeling and Analysis**

(5) **Challenges and Future Directions**

PROBLEM STATEMENT

- **Observation**: potential correlations exist between higher PEFR values and demographic factors (e.G., Age, weight, height, smoking status, occupation, and living environment).
- **Goal**: identify and determine these correlations.
- **Approach**: analyze summary PERF statistics for each patient, combining them with demographic data to uncover trends.

DATA PROCESSING AND CLEANING

Data Processing and Imputation Steps:

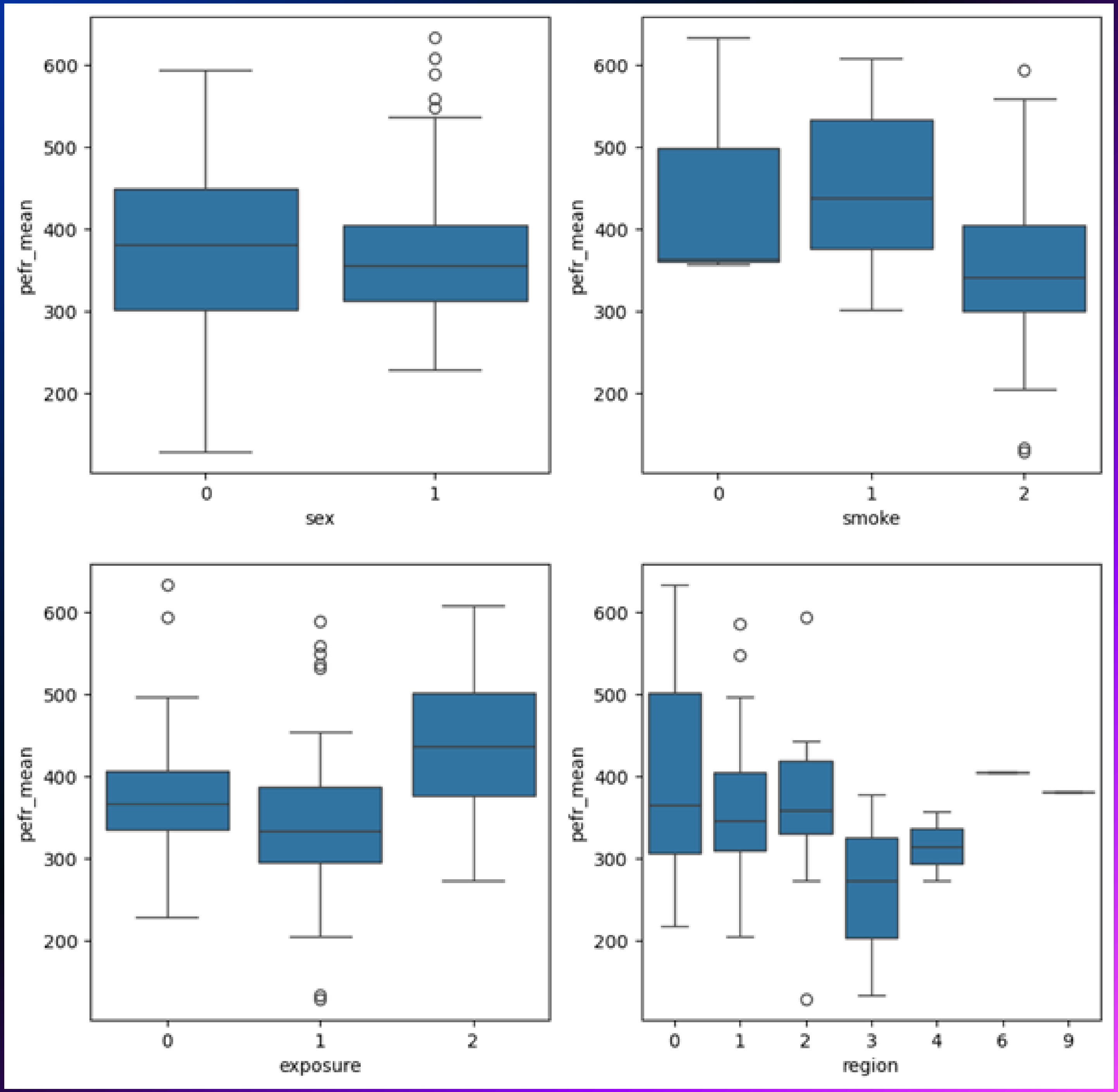
1. **Translation and Cleanup**:
    - Standardized and cleaned data for consistency and compatibility.
2. **Filtering Continuous Data**:
    - Retained only patients with 12+ months of continuous data, reducing the dataset to 82 patients.
3. **Handling Missing Data**:
    - Short gaps (≤14 days): Forward-fill imputation.
    - Long gaps (>14 days): Linear interpolation.
    - Excluded columns with >30% missing data to avoid bias.
4. **PEFR Imputation**:
    - Used **perf_am** values for statistical modeling due to their higher average performance.
    - For missing **perf_am** values, substituted the maximum value from **perf_pm** and **perf_others**.

EXPLORATORY DATA ANALYSIS (EDA)

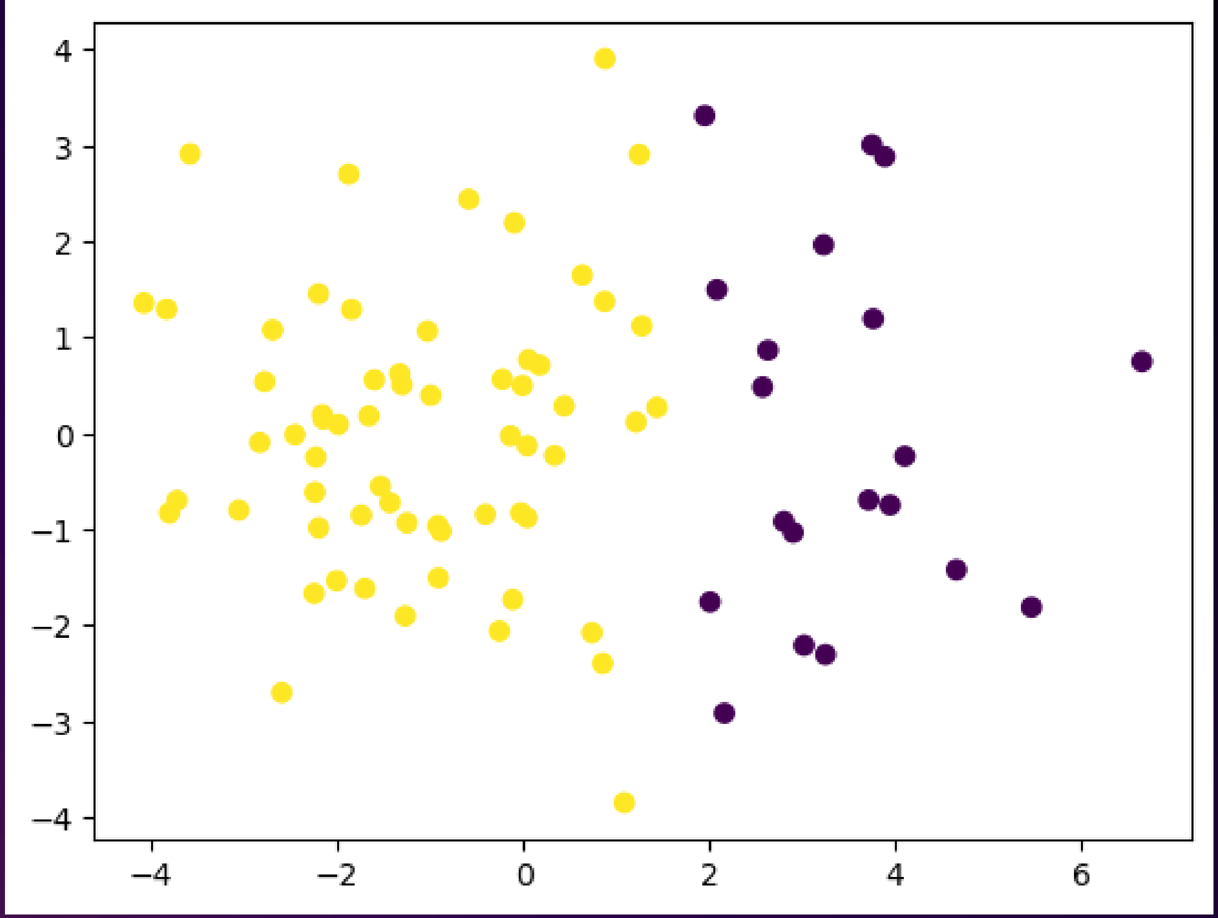
Categorical Features

- sex
- smoke
- exposure <= occupation
- region <= address

pefr_mean



CLUSTERING MODEL



Principle Component Analysis (PCA)
n_components=12

KMeans Model
n_clusters=2

Scores	Averages
Silhouette	0.2533753916
Calinski Harabasz	22.7460078588
Davies Bouldin	1.7472630037

REGRESSION MODEL

LinearRegression Model

Target Feature: pefr_mean
Attribute Features: age, sex, smoke, smoke_amount, height, weight, BMI, BSA, exposure, region, pefr_count, pefr_min, pefr_max, pefr_sum, pefr_std, pefr_skew

Fold	Mean Squared Error	R-Squared	Relative Error
1	2365.391188	0.633512	1.657084
2	1999.509390	0.817059	1.598833
3	4175.302437	0.577034	0.994056
4	2589.772276	0.817179	2.841265
5	1430.315555	0.836624	1.168266
Avg.	**2512.0581692**	**0.7362816**	**1.6519008**

CLUSTER 0

Fold	Mean Squared Error	R-Squared	Relative Error
1	-1825.177999	0.234909	-1.441919
2	-1727.024403	0.140599	-1.437139
3	-2632.789441	0.195195	-0.610284
4	-2392.675366	0.159998	-2.651978
5	-930.100458	0.103801	-0.967719
Avg.	**-1901.5535334**	**0.166904**	**-1.4217878**

CLUSTER 1

Fold	Mean Squared Error	R-Squared	Relative Error
1	-843.816753	-0.072713	-0.603564
2	10.068633	-0.102984	-0.424392
3	-3846.560593	0.321985	-0.547542
4	-1439.655907	-0.217421	-1.988669
5	-279.452434	-0.070752	-0.347041
Avg.	**-1279.8834108**	**-0.0283306**	**-0.7822416**

CHALLENGES & FUTURE DIRECTIONS

Modeling

Explore and use different Regression models
Explore and use different Clustering models

Features and Parameters

Different attribute features of the regression model
Generating different numbers of clusters

THANK YOU

[1] *2.3. clustering*. scikit. (n.d.). https://scikit-learn.org/1.5/modules/clustering.html

[2] Better Health Channel. (2021, April 6). *Asthma and your workplace*. https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/asthma-and-your-workplace