Capstone Project

Sprint 3

Laura Ansari
12 April

What are we going to cover today?

Project intro

Data exploration

Dataset & pre-processing

Modeling

Next steps

Project intro.

Why does it matter?

90% of the world's population is exposed to bad air quality from time to time.

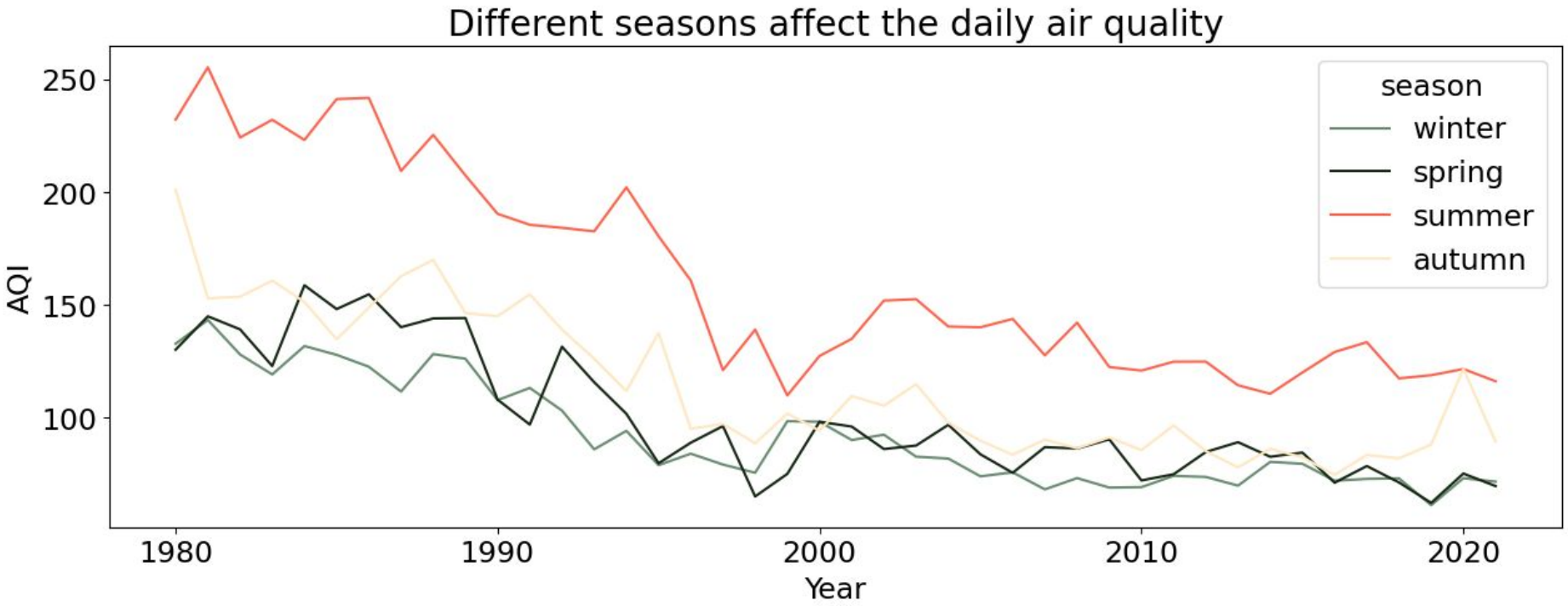
Some individuals are at higher risks: children, elderly, pregnant people, and people with asthma.

How can these high-risk groups plan their outdoor activities so they minimise risk?

How can we use machine learning to predict air quality accurately so that high-risk individuals can minimise their exposure to bad air quality?

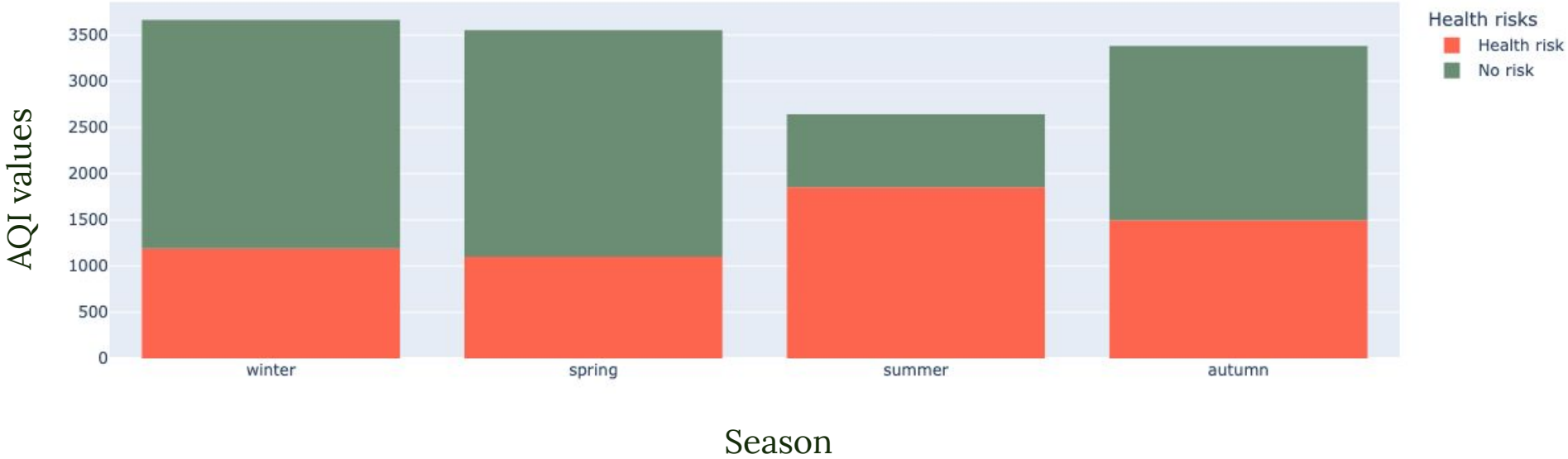
Data exploration.

Different times of the year have different levels of risk



High-risk groups are more exposed during the summer

Health risks for high-risk groups



Dataset & preprocessing.

Cleaning and pre-processing

Duplicates

No duplicates found in the dataset

Missing values

No missing values

Feature engineering

Lag features and time features for regression and decision trees

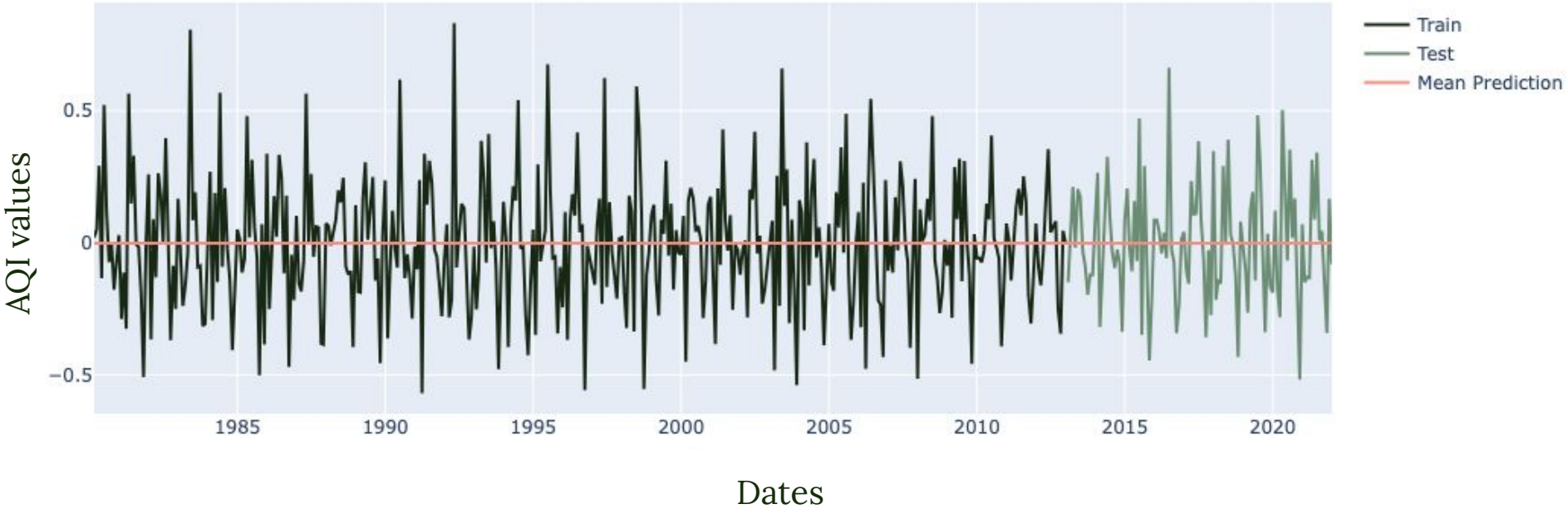
Standardization

Standard scaler (where applicable), log transformation

Modeling.

Baseline model: Predicting mean over training set

Predictions over the train and test set



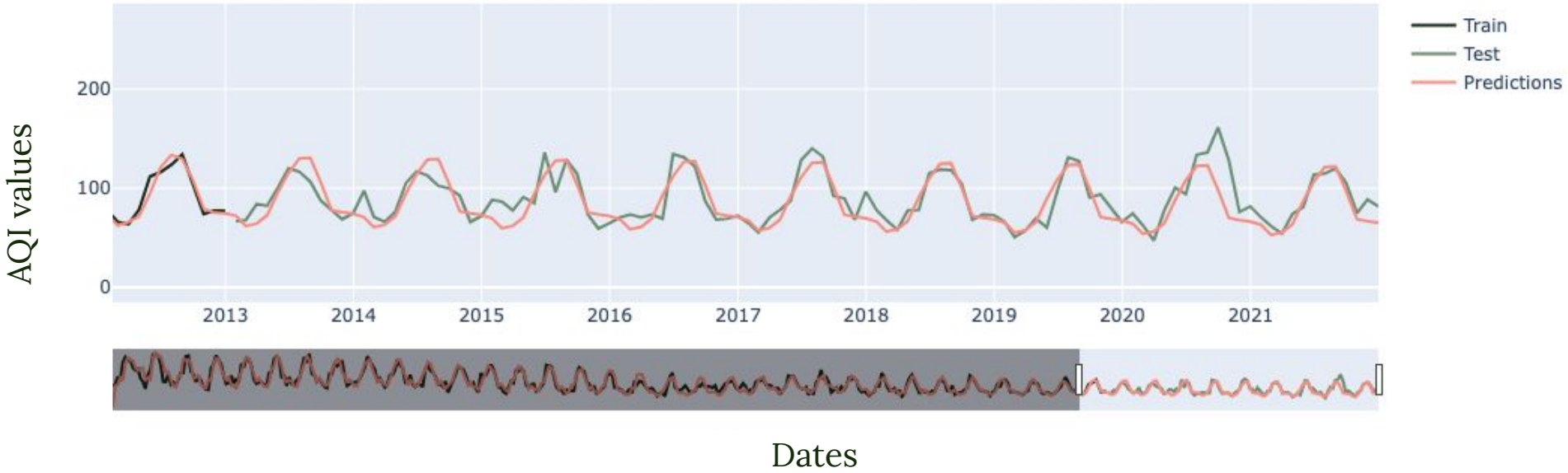
	Test score (MAPE)	Test score (MAPE)
Baseline model	99 %	100 %
Linear regression	? %	? %
XGBoost	? %	? %
SARIMA	? %	? %

	Test score (MAPE)	Test score (MAPE)
Baseline model	**99 %**	**100 %**
Linear regression	**25 %**	**20 %**
XGBoost	**? %**	**? %**
SARIMA	**? %**	**? %**

	Test score (MAPE)	Test score (MAPE)
Baseline model	**99 %**	**100 %**
Linear regression	**25 %**	**20 %**
XGBoost	**24 %**	**21 %**
SARIMA	**? %**	**? %**

SARIMA model

Predictions over the test set



	Test score (MAPE)	Test score (MAPE)
Baseline model	99 %	100 %
Linear regression	25 %	20 %
XGBoost	24 %	21 %
SARIMA	13 %	12 %

Next steps.

Interacting with the model

Choose the location

Enter the date

Check if the air quality is suitable

Future vision