WINE NOT?

A MERLOT ANALYSIS
FROM GEOSPATIAL DATA

CHRISTIAN BUDA
EDOARDO DI MARTINO
GIULIO PECILE
LEONARDO PLINI

THE GOAL

WHAT IF WE COULD HAVE AN IDEA OF THE PUBLIC RECEPTION A WINE COULD HAVE BEFORE EVEN PUTTING IT ON THE MARKET?

BASED ON THE ORGANOLEPTIC PROPERTIES, THE QUALITIES OF THE SOIL WERE THE GRAPES WERE CULTIVATED AND THE WEATHER ASSOCIATED TO THE GROWTH YEAR, WE PROPOSE TO PREDICT HOW THE CONSUMER WILL WELCOME THE PRODUCT

HOW?

CREATE OUR VERY OWN DATASET:

We needed to build a dataset comprised of organoleptic properties of different Merlot wines, together with weather and soil information regarding the area in which the grapes were cultivated

PREDICT THE WINE'S RECEPTION:

Users and experts gather on the internet to rate wines: are we able to predict if a wine will be appreciated or not based on its organolectic properties and its area of production?

UNDERSTAND OUR RESULTS:

We decided to use different techniques to better explain the results of our models

OUR WORK:

OBTAINING THE DATA

PREPROCESSING AND DATA EXPLORATION

BUILDING MODELS

INTERPRETING MODELS

CHAPTER 01:

The gruesome process of collecting data

Restricting to a single variety of grape, we needed data about:

- how good the wine is
- cultivation site data
- weather data

HARD TO FIND!!

We got in touch with some companies involved in the field, but none of them had the data we needed

HORT@ From research to field

TERROIR FROM SPACE

Solution: collect data from different sources

Five (and a half) painful steps

- Organoleptic properties
- Geographical coordinates
- Soil data
- Weather data
- Wine scores

Kaggle

Aziende Vinicole + Google

ISRIC

Visual Crossing

vivino

Downloading data from Kaggle is easy, what about the other parts?

First step: wines' chemical properties

Kaggle dataset:

- Contains information about different wines and their properties

name	producer	nation	abv	degree	sweet	acidity	body	tannin	price	year	ml
Citra, Merlot	Citra	Italy	13.00	17.0	1	3	4	3	0	2017	750
Geografico, Pulleraia	Geografico	Italy	13.50	17.0	1	3	4	4	108000	2012	750
Castello di Ama, L'Apparita	Castello di Ama	Italy	14.50	17.0	1	4	5	5	500000	2014	750
Masseto	Ornellaia	Italy	15.00	0.0	2	4	5	5	1900000	2016	750

- Filtered for italian Merlots

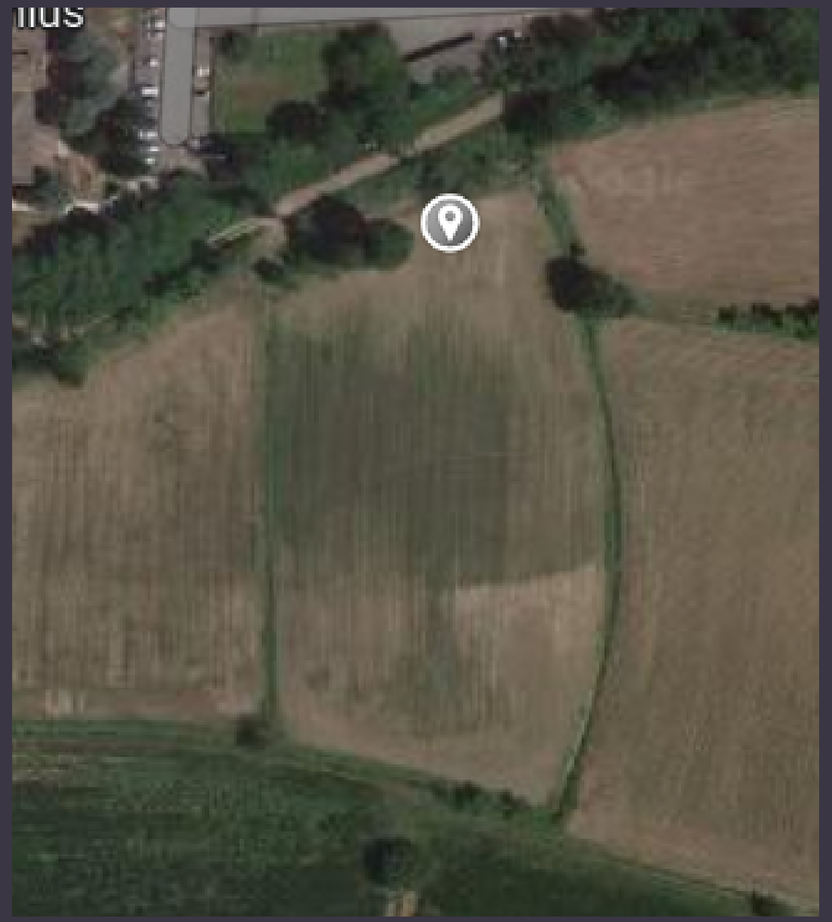
Related work:

- Study conducted by Cortez et al., 2009, on Vinho Verde
- This data is not suited for our purposes

www.kaggle.com/datasets/dev7halo/wine-information

Second step: retrieve latitude and longitude

- We needed a couple of coordinates for each observation. This is why we reduced the analysis to just Italy
- We made http requests to retrieve a unique street id and CAP code from www.*aziendevinicole.it*
- We queried Google Earth to obtain latitude and longitude data
- We double checked on the satellite images that these areas had a wineyard
- We dealt manually with the cases for which we had bad results



Third and fourth step: soil and weather data

Soil data:

- ISRIC data
- *www.isric.org*
- *www.soilgrids.org*
- It grants access to soil data after specifying latitude and longitude
- They provide an API interface for requests

Climate database:

- *www.visualcrossing.com*
- Used a Python pipeline to query weather data for each location
- Collected data spanning from March to September of the year of production for each wine
- Computed average statistics for each wine according to scientific knowledge on grape vine growth

Fifth step: wines' scores

- We used the world's largest wine marketplace: Vivino
- There were more exceptions than rules, so, due to the small number of data points, we performed this operation by hand
- **NOTE**: there are some better review sites from professional sommeliers with a proper API, but they are not as cheap as Vivino
- We retrieved the scores for each wine and the organoleptic properties for the test dataset

CHAPTER 02:

The slightly less gruesome process of preprocessing data

THE GOOD:

- General cleaning of the data
- One-hot encoding and label-encoding of categorical features
- Dealing with some weather NaNs by substituting them with the average values of three rows with similar coordinates

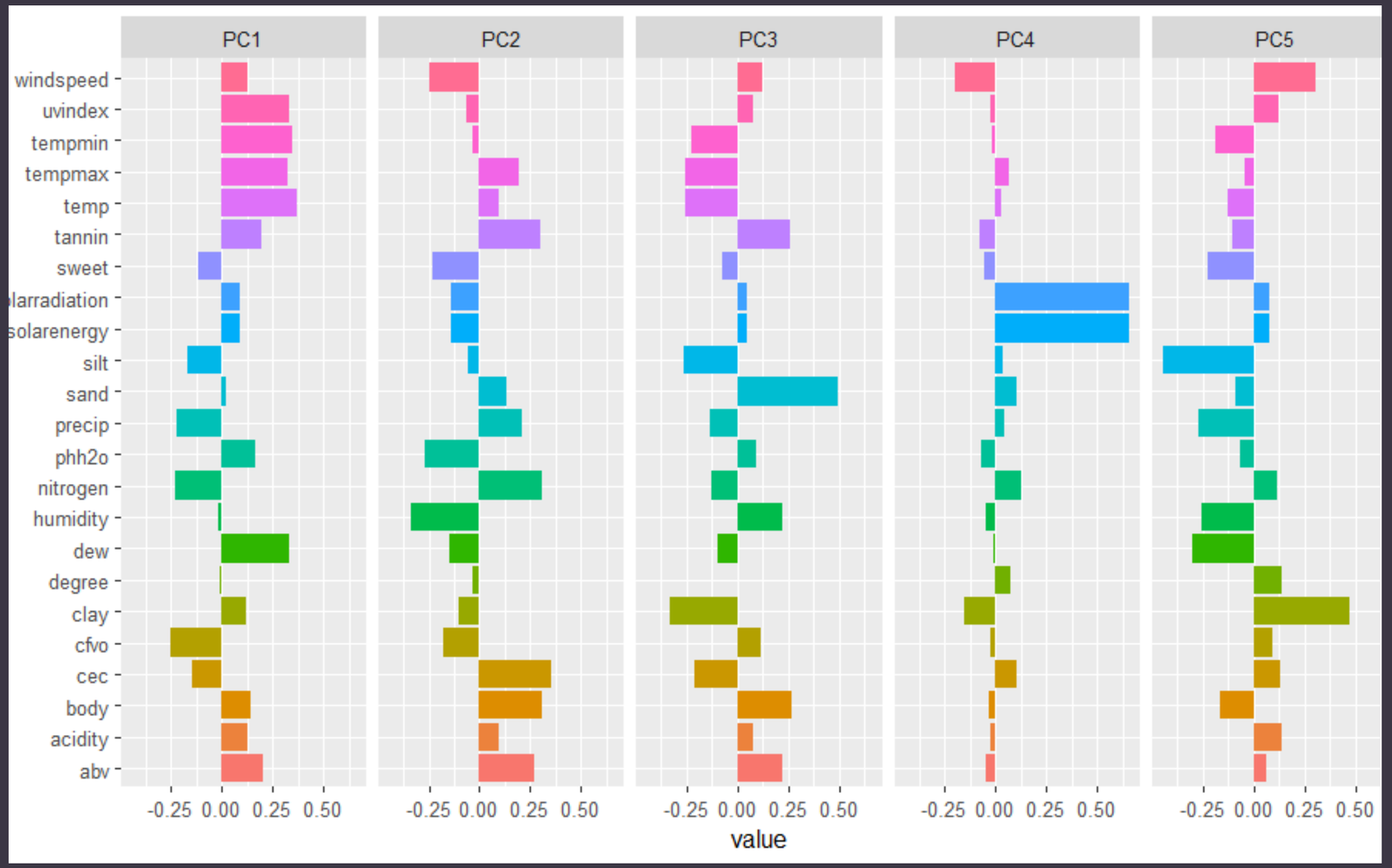
THE BAD & THE UGLY:

- Adding some missing years by hand
- Re-query the climate database to account for lost information due to wrong year
- Manually searching for and adding missing infos about a plethora of wines

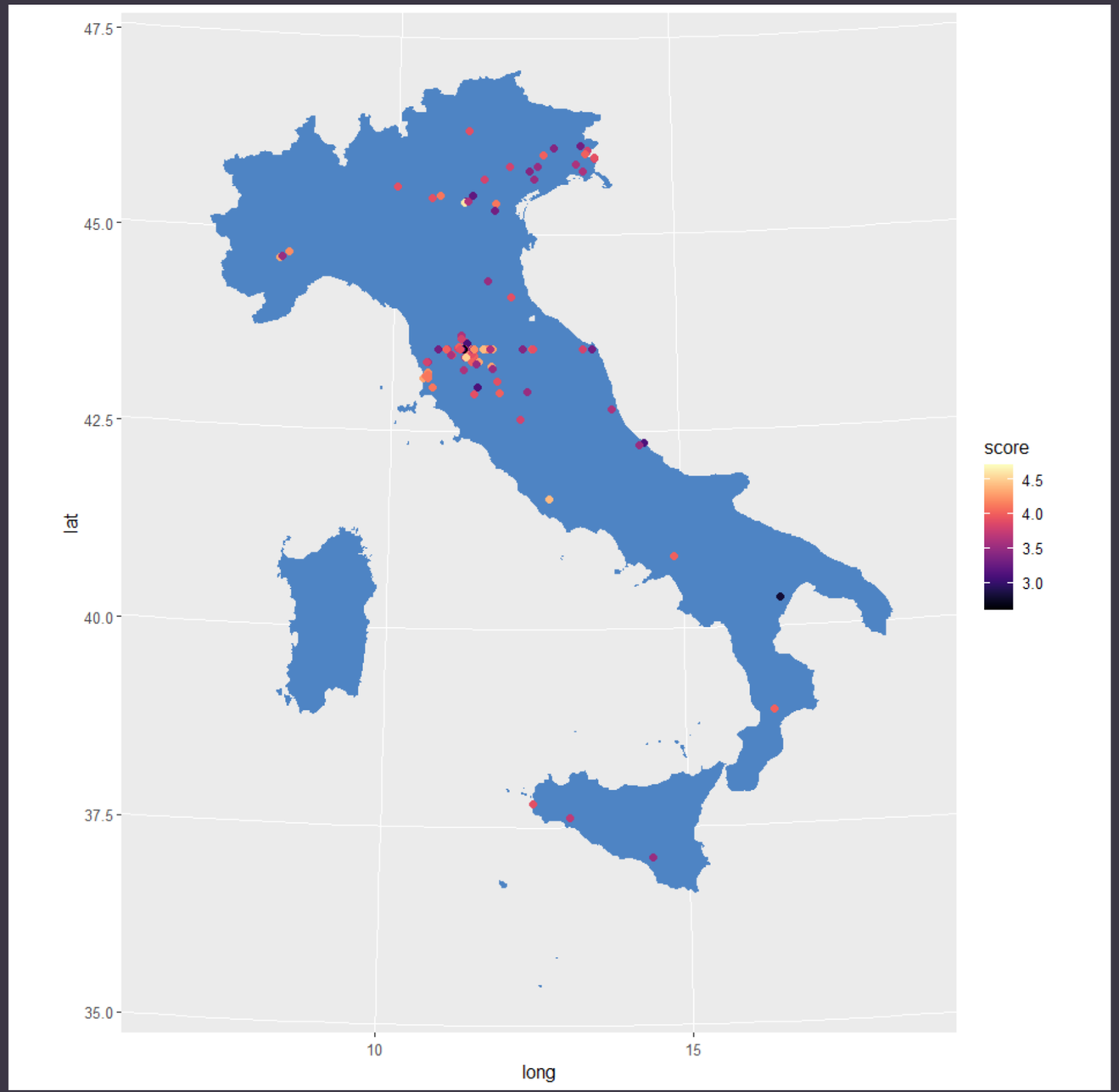
CHAPTER 03:

Data visualization

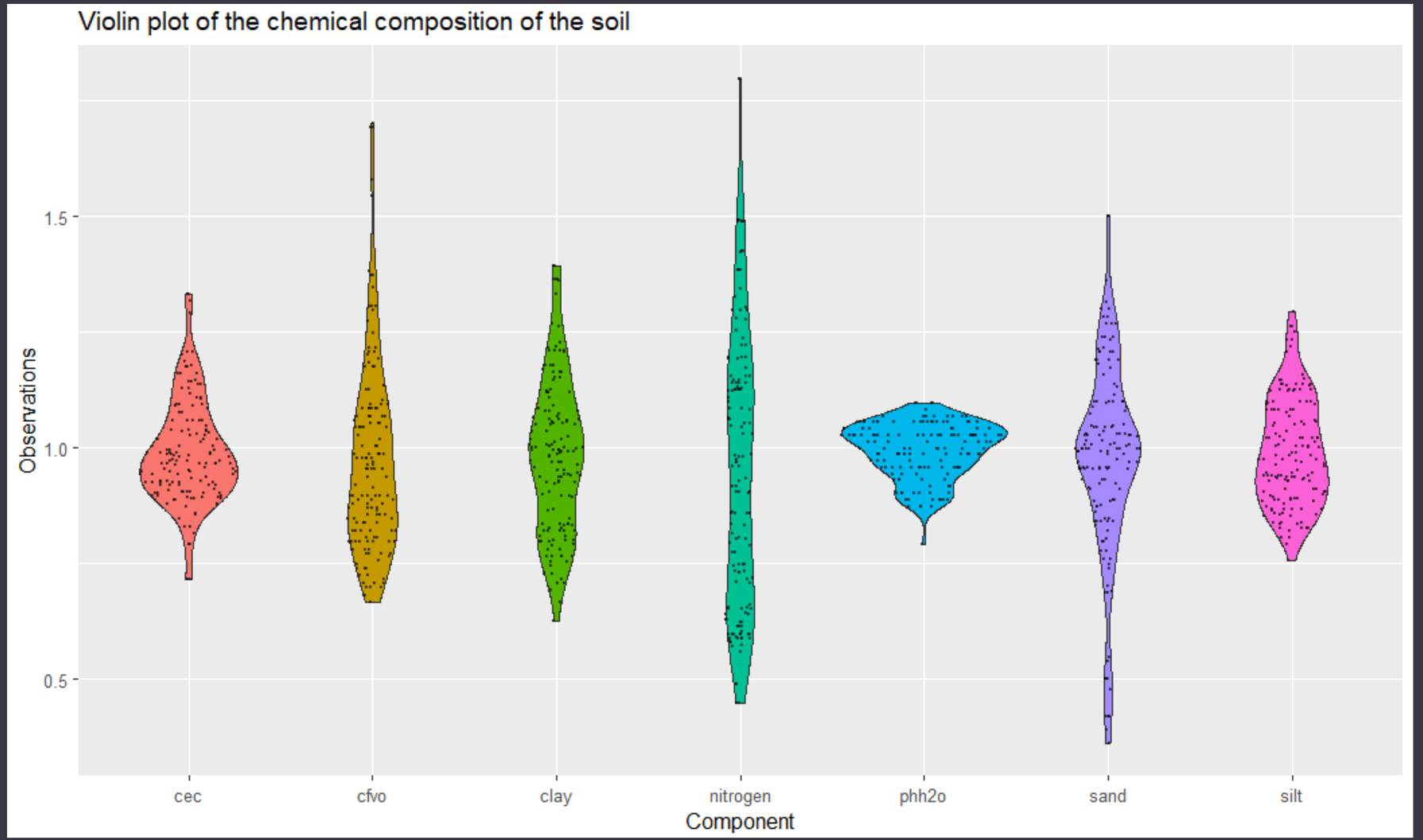
LET'S START WITH A PCA, TO IDENTIFY THE FIVE MAIN PROFILES OF WINES



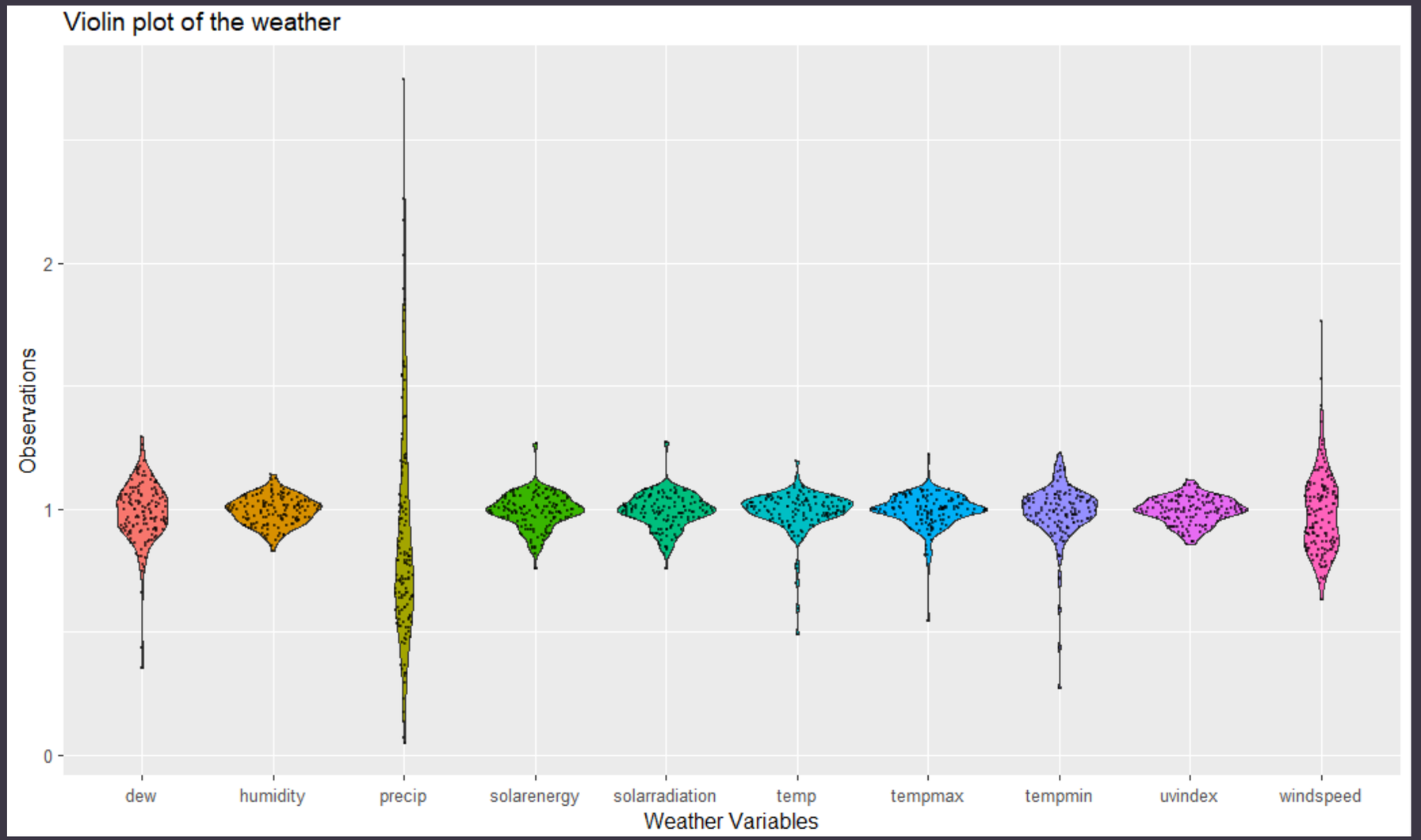
THE SPATIAL COORDINATES CAN OFFER US A CLEARER VIEW OF WHAT IS HAPPENING



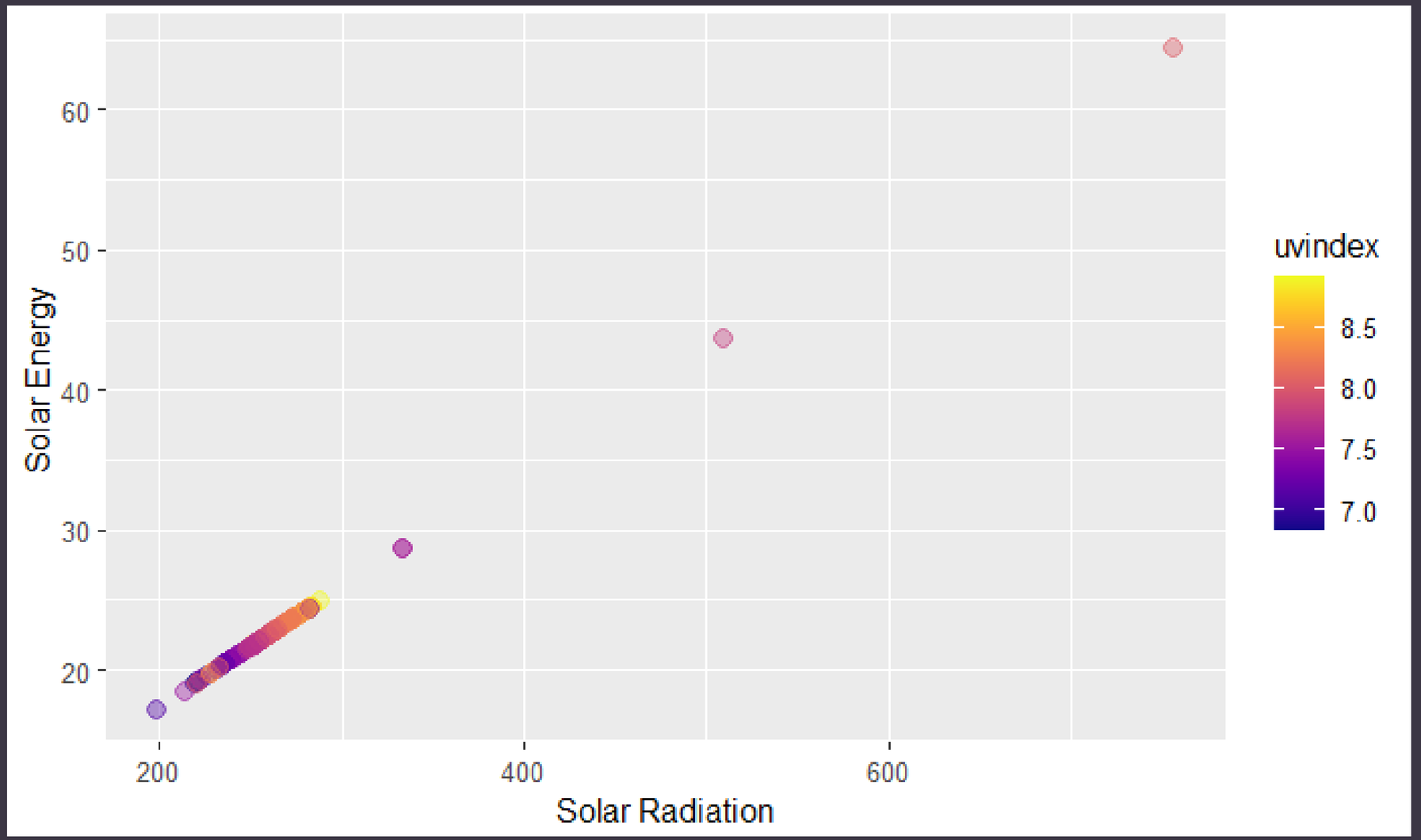
VIOLIN PLOT OF THE (SCALED) SOIL PROPERTIES



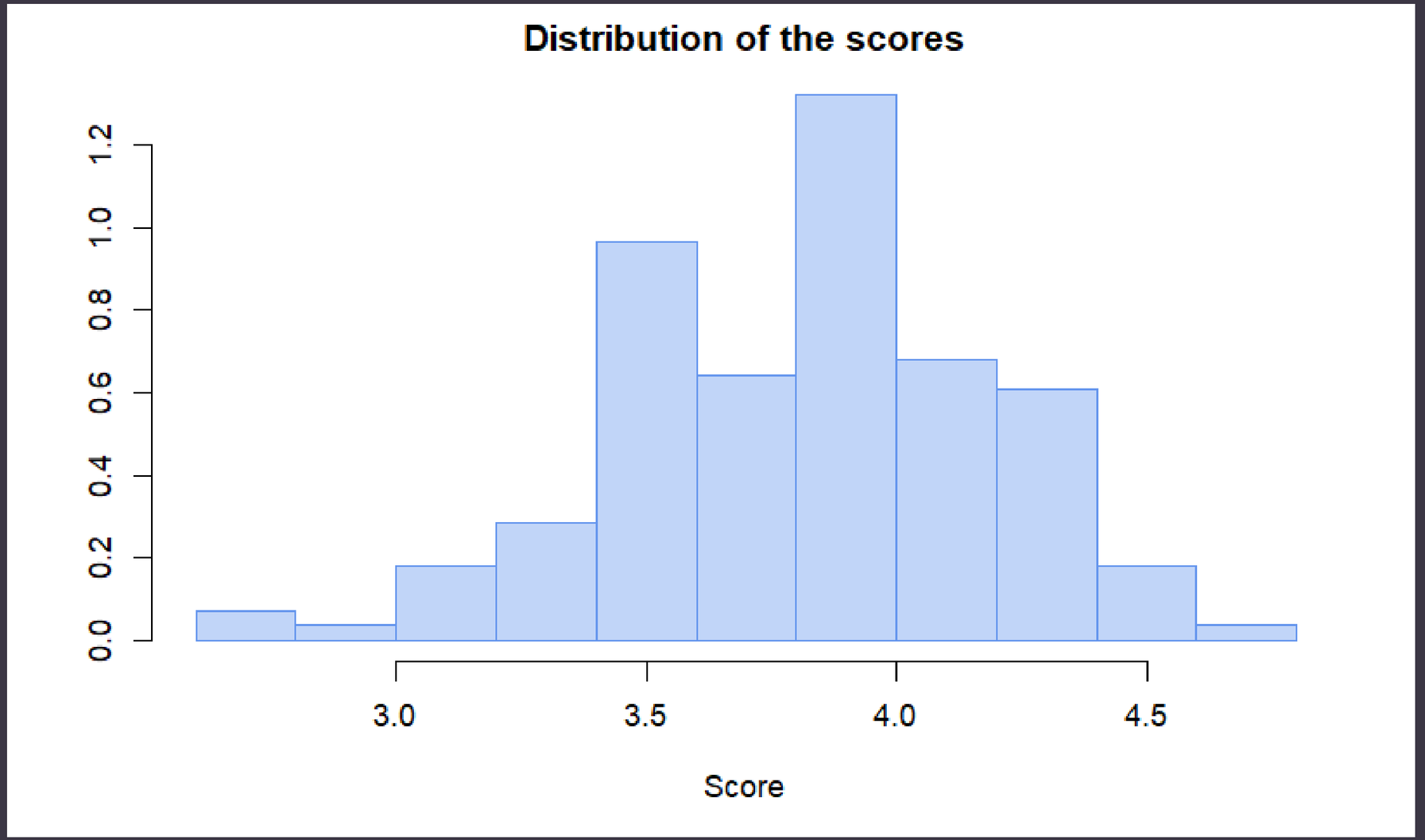
VIOLIN PLOT OF THE (SCALED) WEATHER CONDITIONS



CORRELATION BETWEEN SOLAR RADIATION AND SOLAR ENERGY



SCORES DISTRIBUTION OVER THE DATASET



Mean = 3.9, SE = 0.4

Bonus chapter:

A new, teeny tiny dataset

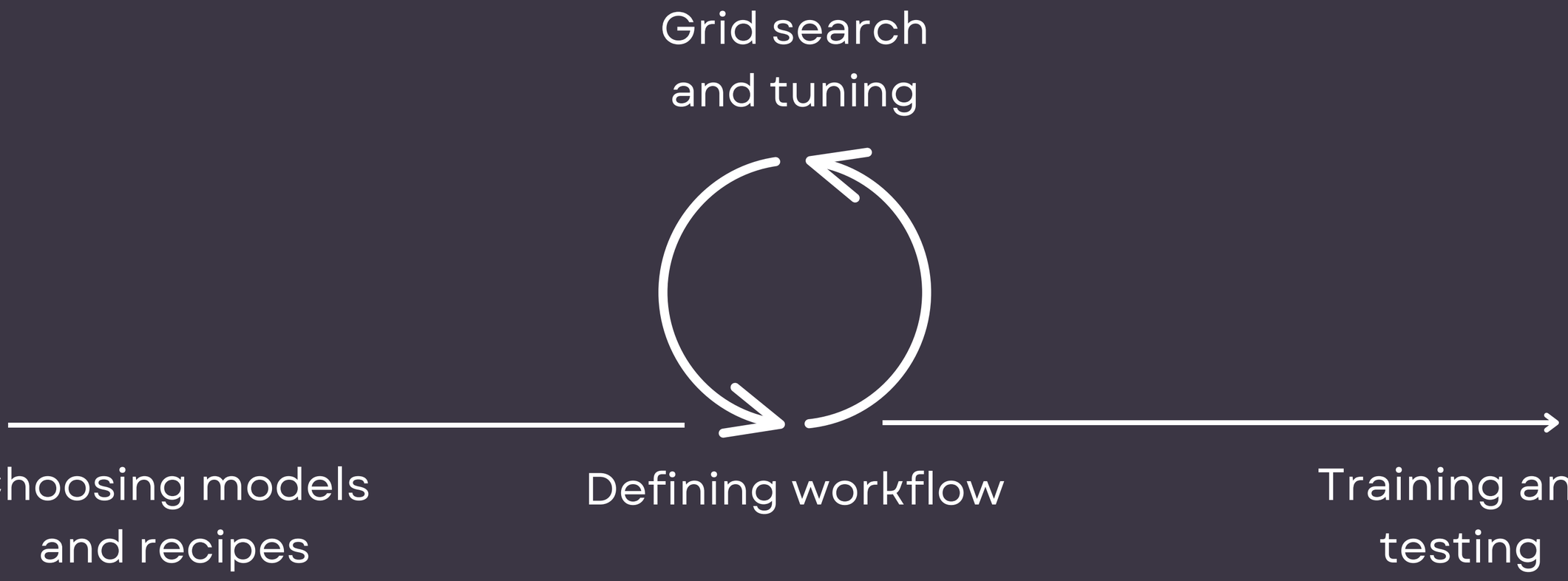
Five new Merlots, not present in the original dataset

- Data collected by hand
- Later used to test our models

Name	Producer	Nation	abv	year	local1	score	latitudine	longitudine	cec	cfvo	clay	nitrogen	phh2o	sand	silt	tempmax	tempmin	temp	dew	humidity	precip	windspeed	solarradiation	solarenergy	uvindex
Sodale	Cotarella	Italy	13.8	2015	Lazio	3.9	42.652656	12.244379	266.0	104.0	292.0	290.0	74.0	248.0	460.0	25.0720	13.5315	19.3305	12.3155	66.3775	391.95	23.3115	264.5715	22.8455	8.070
Varneri Collio	Marco Felluga	Italy	13.0	2017	Friuli-Venezia Giulia	3.7	46.079854	13.521687	371.0	141.0	307.0	511.0	54.0	269.0	424.0	21.4415	10.4785	15.6210	9.9165	71.4295	693.19	17.2485	234.8435	20.2690	7.225
Decima Aurea	Tenuta Santa Maria di Gaetano Bertani	Italy	14.5	2010	Veneto	4.3	45.504583	10.945944	261.0	128.0	291.0	529.0	69.0	256.0	453.0	24.4570	13.8630	19.3400	13.0270	72.0320	656.68	18.1935	231.3635	19.9860	7.110
Manero Rosso	Fattoria del Cerro	Italy	13.5	2018	Toscana	3.4	43.089941	11.860604	265.0	110.0	271.0	421.0	75.0	265.0	464.0	24.3605	13.7280	19.1660	12.3960	68.1365	395.10	17.4135	217.2285	18.7460	7.770
Bolgheri Rosso	Tenuta Campo al Mare	Italy	15.0	2019	Toscana	4.0	43.175386	10.598261	251.0	113.0	275.0	267.0	76.0	358.0	367.0	23.6510	13.7565	18.8800	13.7300	74.1290	304.51	18.8530	756.3405	64.3930	8.020

CHAPTER 04:

Models



Models' performance

	Linear Regression	Random Forest	XGBoost	kSVM	Real Score
1	3.79	3.75	3.81	3.86	3.90
2	4.15	3.70	3.81	3.89	3.70
3	4.05	3.98	4.46	3.89	4.30
4	3.72	3.63	3.48	3.86	3.40
5	4.38	4.34	4.52	3.94	4.00
RMSE	0.32	0.24	0.25	0.29	

- Random forest and XGBoost perform the best
- Linear regression has the highest RMSE
- The SVM's predictions are all similar between them

- A bigger test dataset would be needed to have a deeper analysis

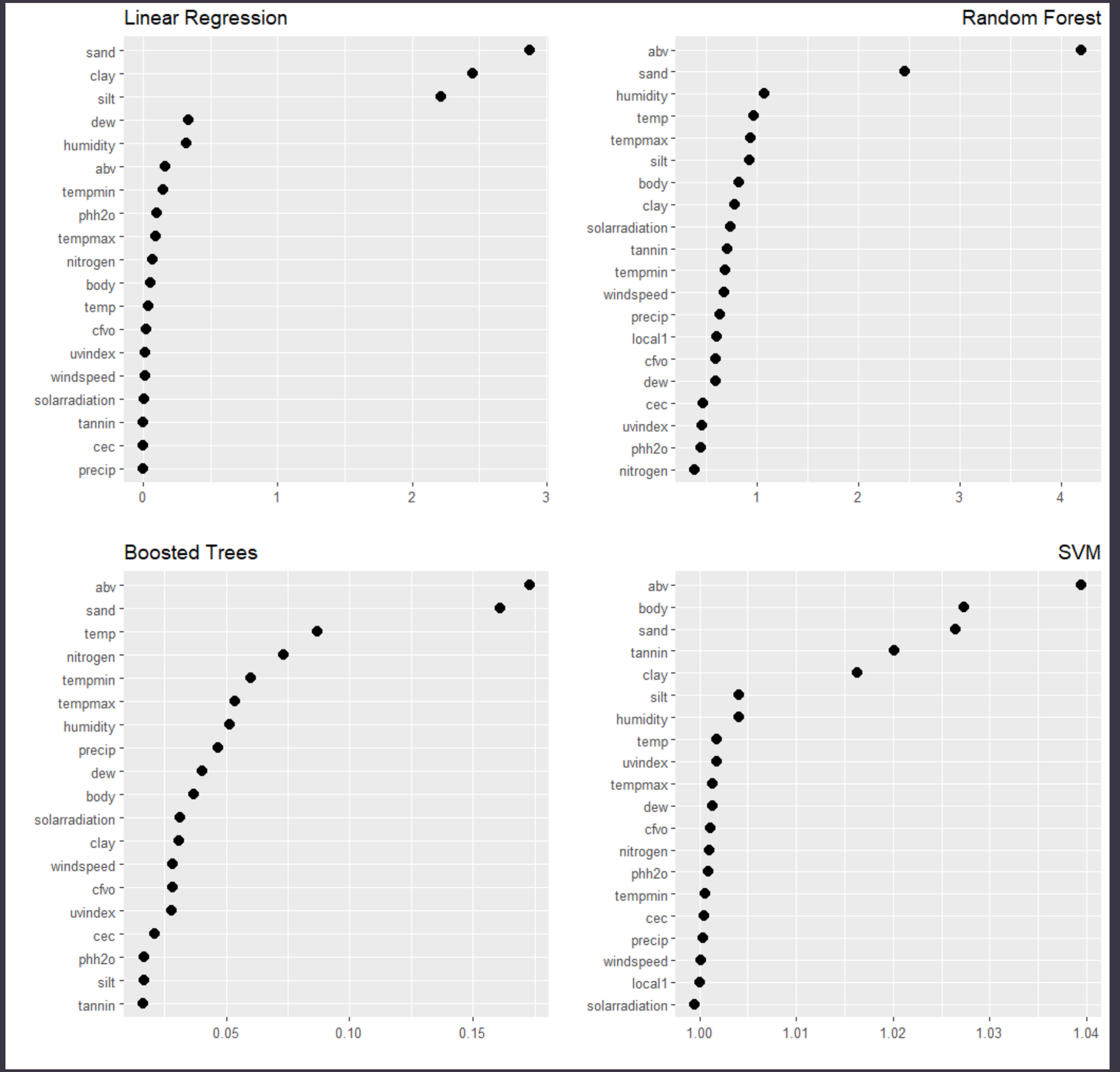
CHAPTER 04:

Interpreting our models

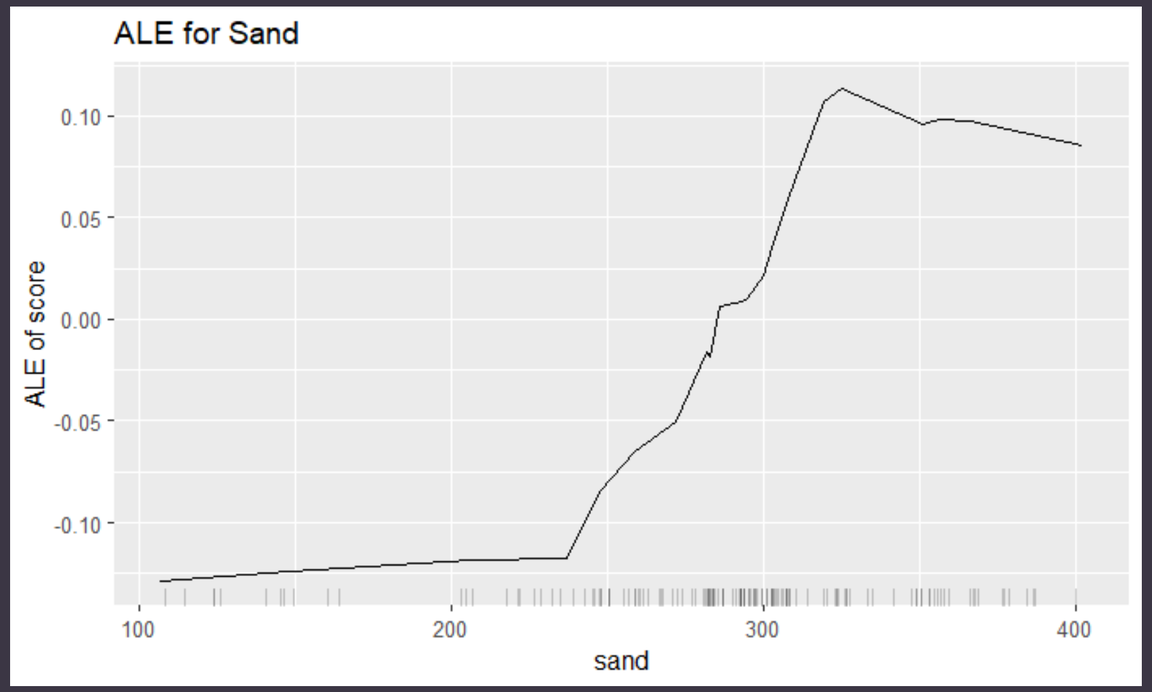
We used some IML techniques to retrieve information about our models, in order to better understand the predictions.

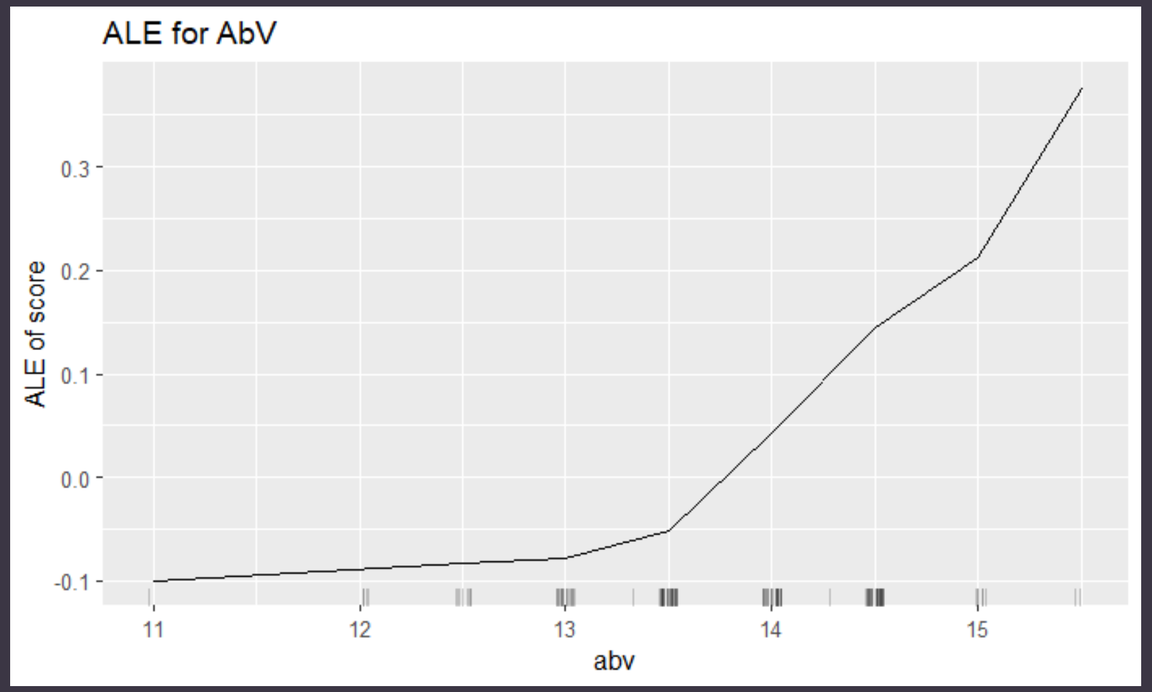
Following are some plots regarding our findings

FEATURE IMPORTANCE FOR THE DIFFERENT MODELS

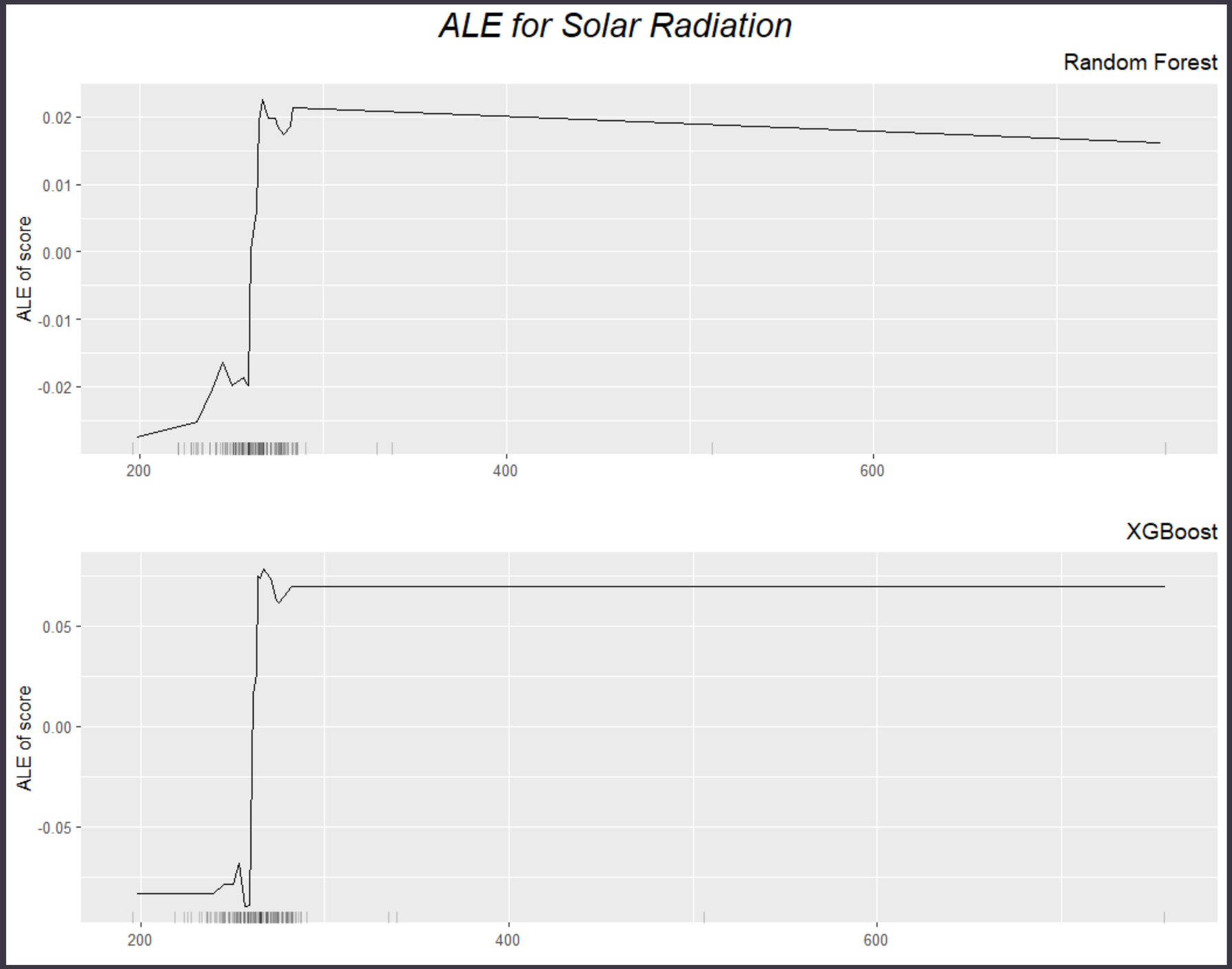


ACCUMULATED LOCAL EFFECT FOR SAND AND ABV IN RANDOM FOREST

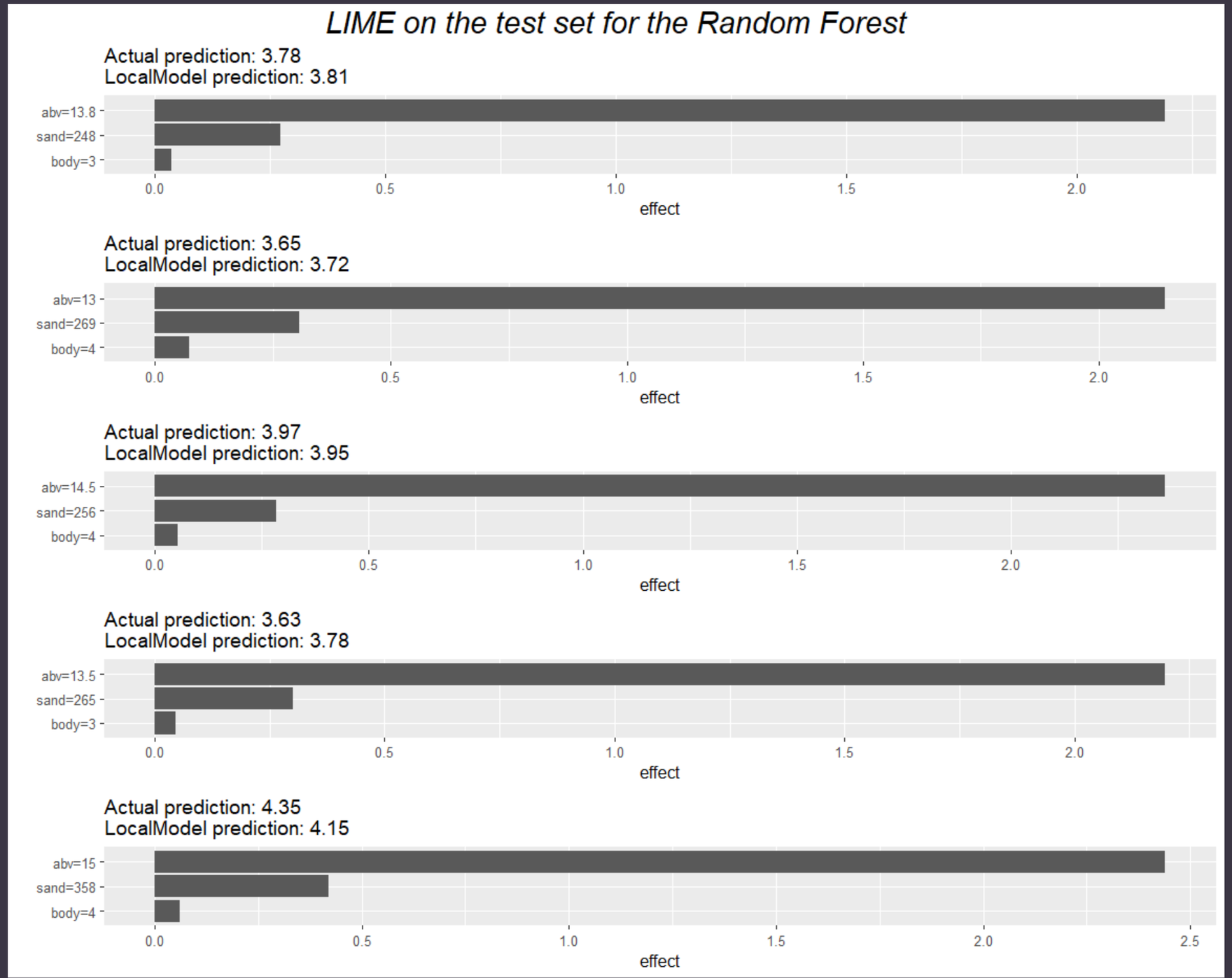




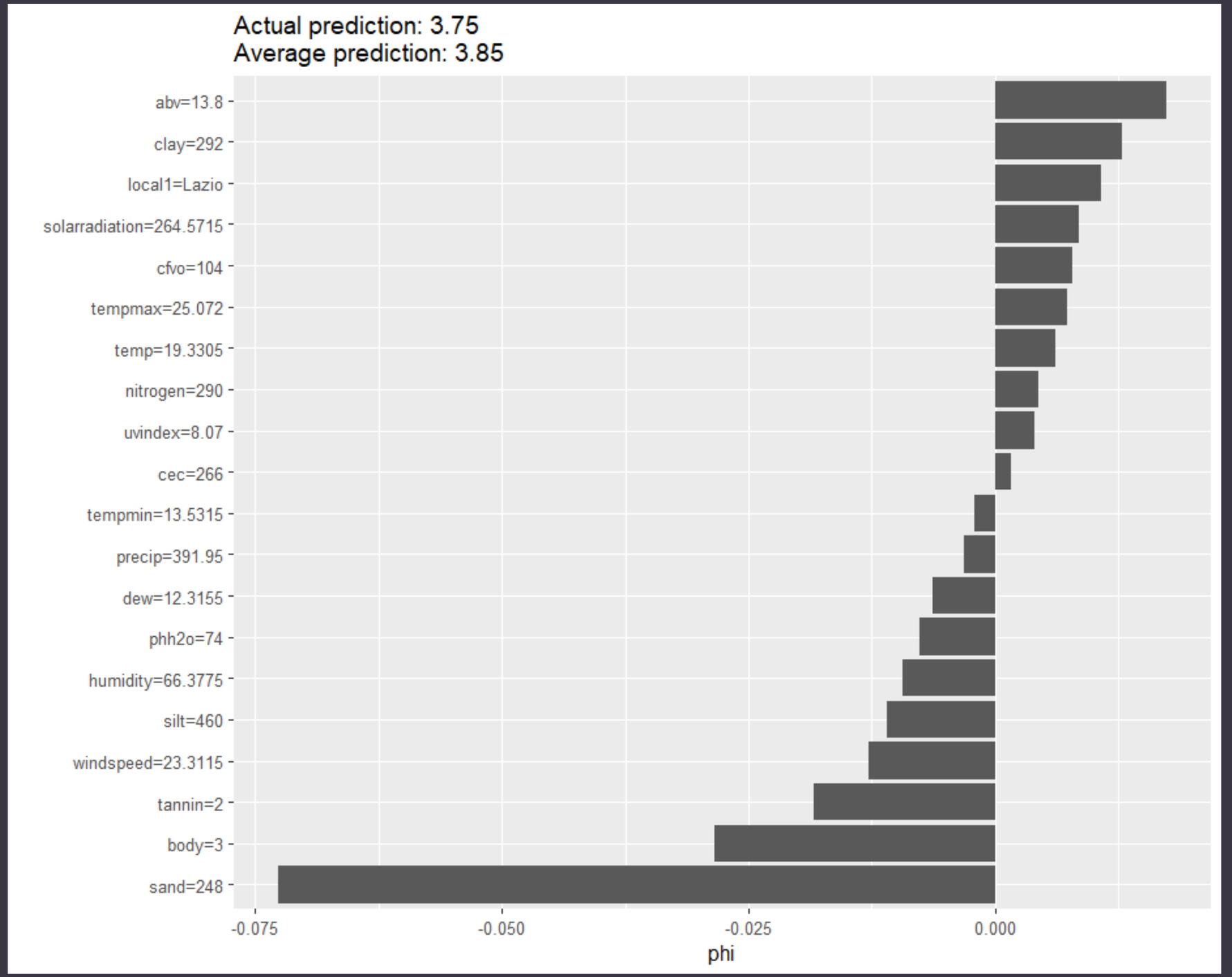
ALE FOR SOLAR RADIATION IN RANDOM FOREST AND XGBOOST



LIME FOR THE RANDOM FOREST PREDICTIONS



SHAPLEY VALUES FOR THE FIRST OBSERVATION OF THE TEST SET (RANDOM FOREST)



CHAPTER 05:

Endgame

CONSIDERATIONS ON OUR WORK

- Interesting (although not accurate) results on the test dataset
- We can hope to understand the range of appreciation of a new midrange wine

WHAT MORE CAN BE DONE

- A bigger dataset is needed to be able to tell more about the actual performances, especially for wines that are exceptionally good or exceptionally bad
- The next step would be to repeat this analysis with a different type of wine

The winEnd

https://www.youtube.com/watch?v=SQWeMj48arw