Rural Municipality Crop Yield Prediction

Ethan Lam

Agenda

Introduction

The Problem

Solution

The problems

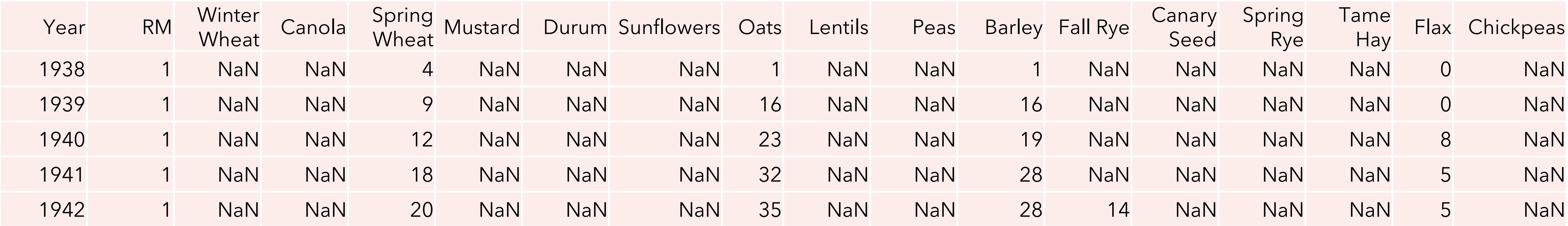
I want to find the best place in Saskatchewan to grow Barley

I want to know if my investment is good in terms of the ROI or I want to forecast the yields in a specific Rural Municipality of a crop

Data Collection and Preprocessing

- Crop yield data:
  - Crop yields by Rural Municipality (RM) are produced annually from the Ministry of Saskatchewan Crop Report and Saskatchewan Crop Insurance Corporation
  - Data provided from 1938 to 2022.
- Geospatial:

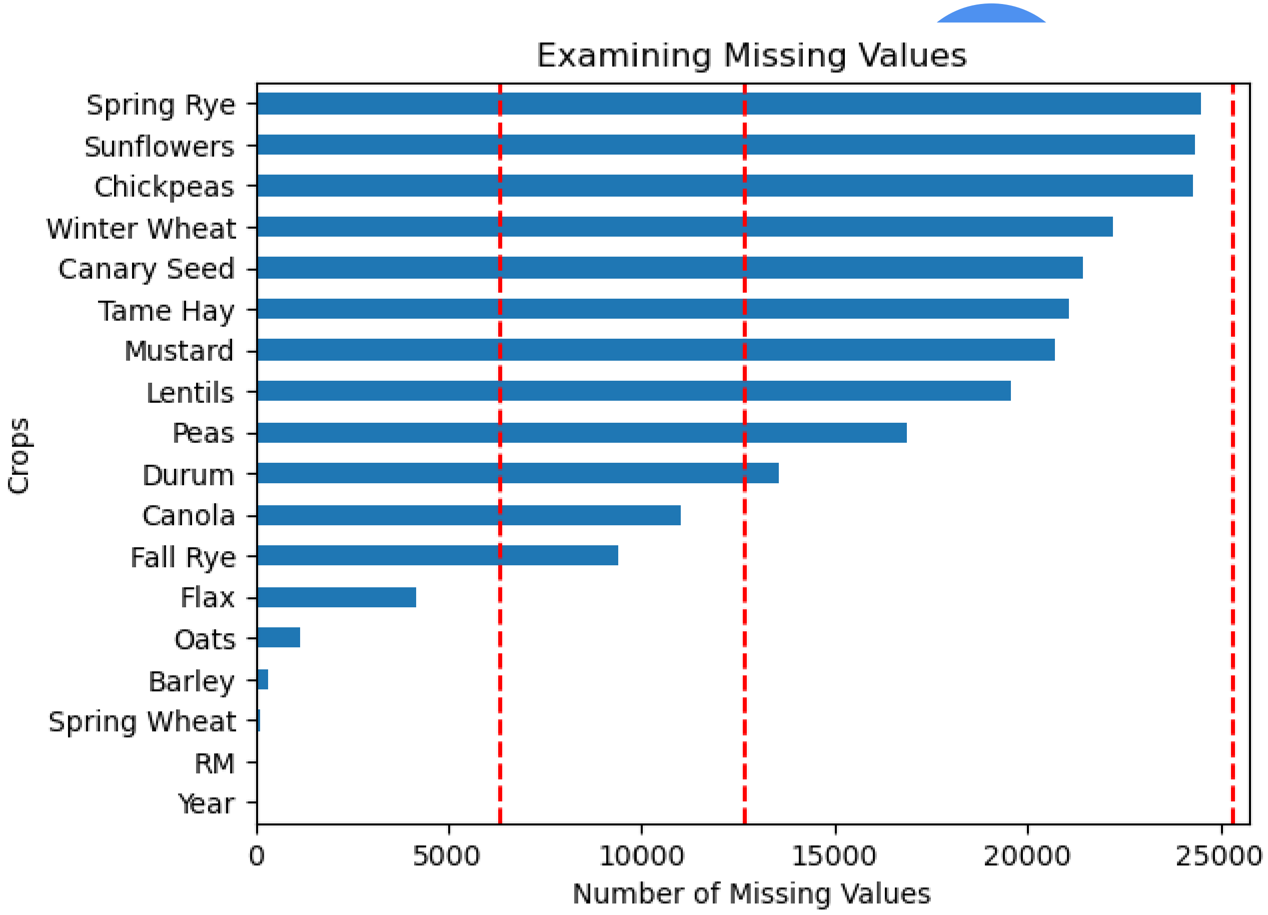
  The shapefile from Government of Saskatchewan
- First look at the data



Year	RM	Winter Wheat	Canola	Spring Wheat	Mustard	Durum	Sunflowers	Oats	Lentils	Peas	Barley	Fall Rye	Canary Seed	Spring Rye	Tame Hay	Flax	Chickpeas
1938	1	NaN	NaN	4	NaN	NaN	NaN	1	NaN	NaN	1	NaN	NaN	NaN	NaN	0	NaN
1939	1	NaN	NaN	9	NaN	NaN	NaN	16	NaN	NaN	16	NaN	NaN	NaN	NaN	0	NaN
1940	1	NaN	NaN	12	NaN	NaN	NaN	23	NaN	NaN	19	NaN	NaN	NaN	NaN	8	NaN
1941	1	NaN	NaN	18	NaN	NaN	NaN	32	NaN	NaN	28	NaN	NaN	NaN	NaN	5	NaN
1942	1	NaN	NaN	20	NaN	NaN	NaN	35	NaN	NaN	28	14	NaN	NaN	NaN	5	NaN

Data Quality Check

- Check for NULL/Missing values
  - Many missing values
- Check for duplicate
  - No duplicate was found

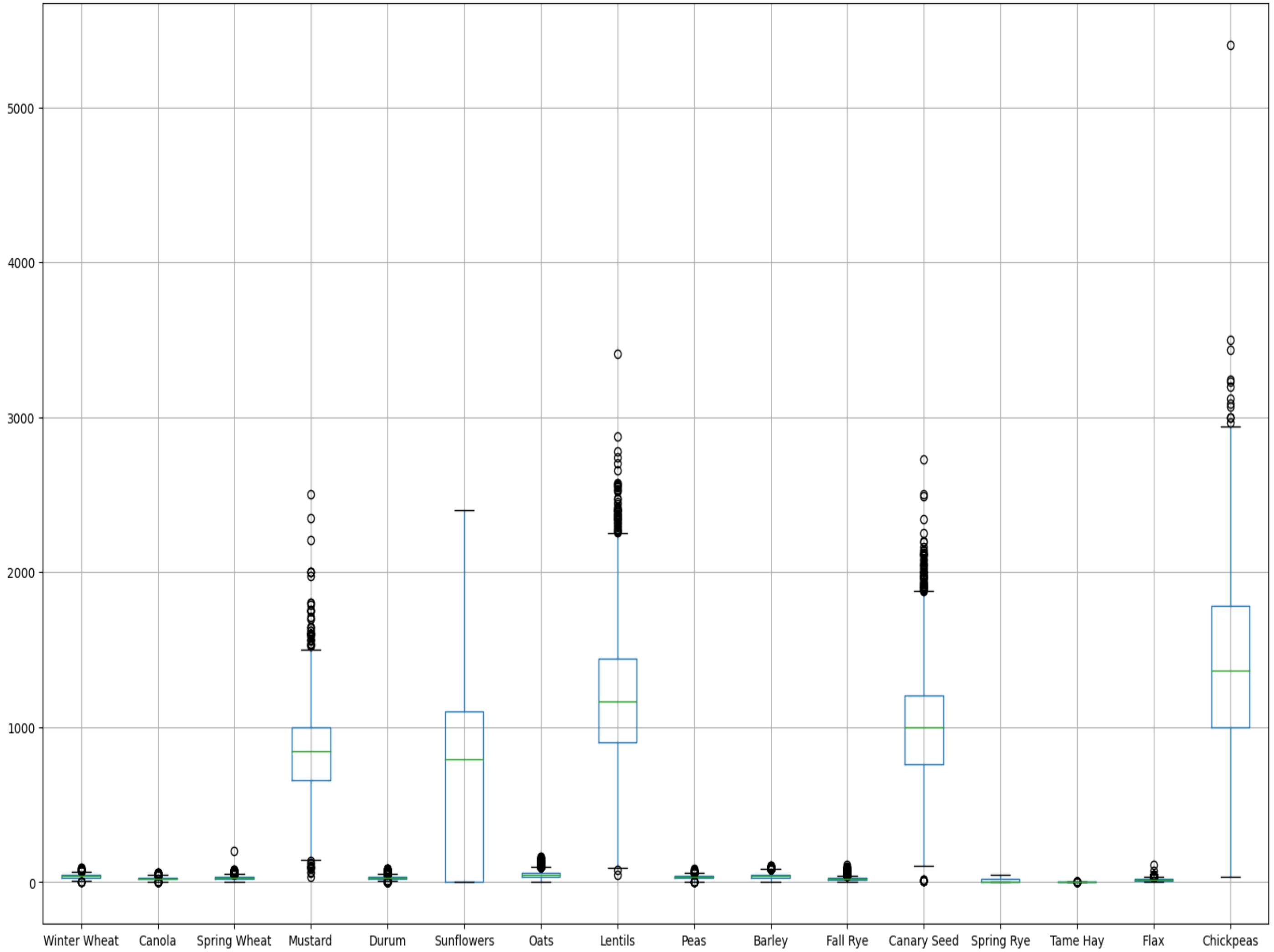


EXPLORATORY DATA ANALYSIS

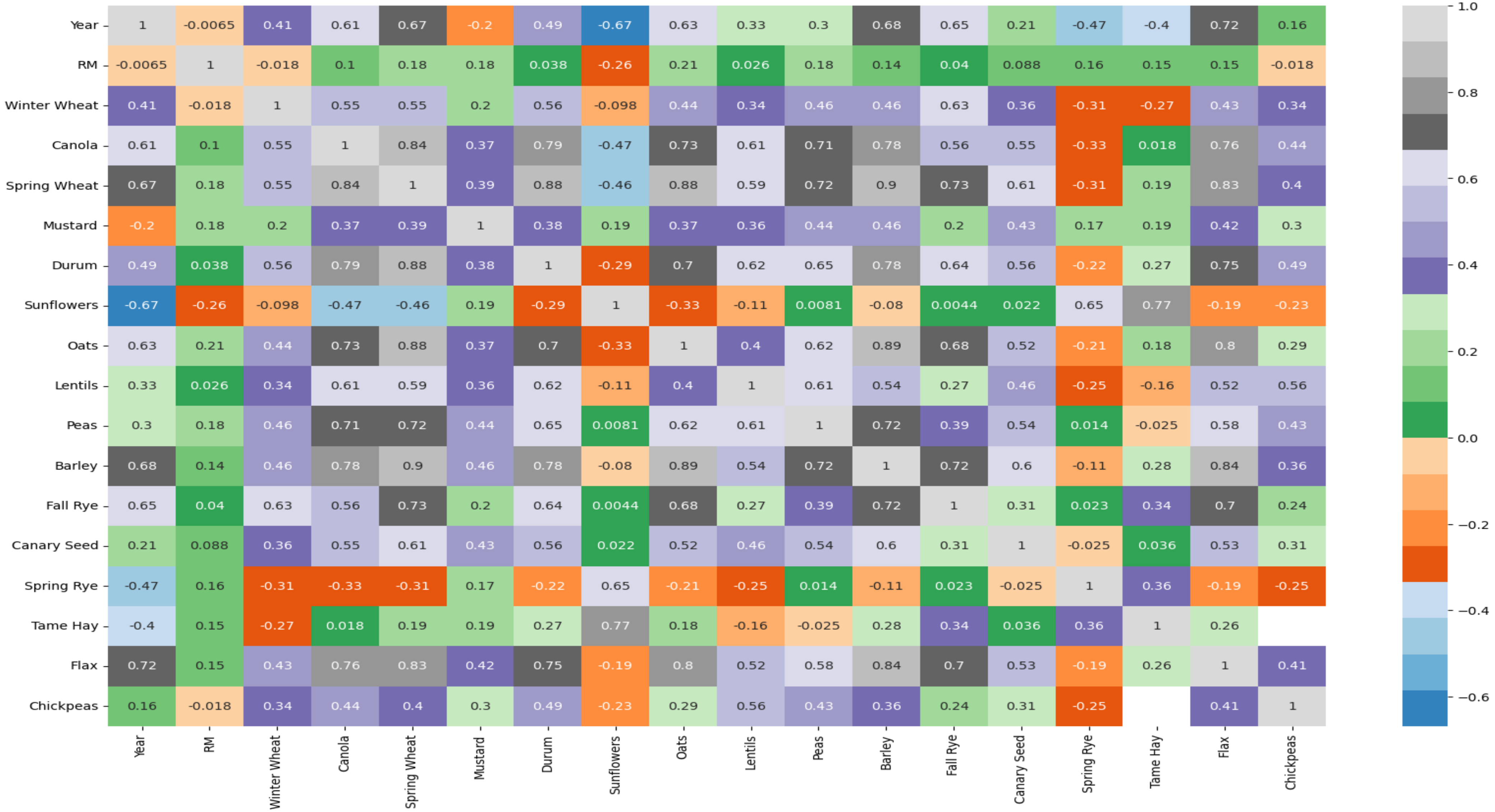
- Understand that there are 25312 entries in the dataset
- There are 18 columns in the dataset
- There are 299 RMs

Check Outliers

There are many outliers in the data of Mustard, Lentils, Canary Seeds, and Chickpeas



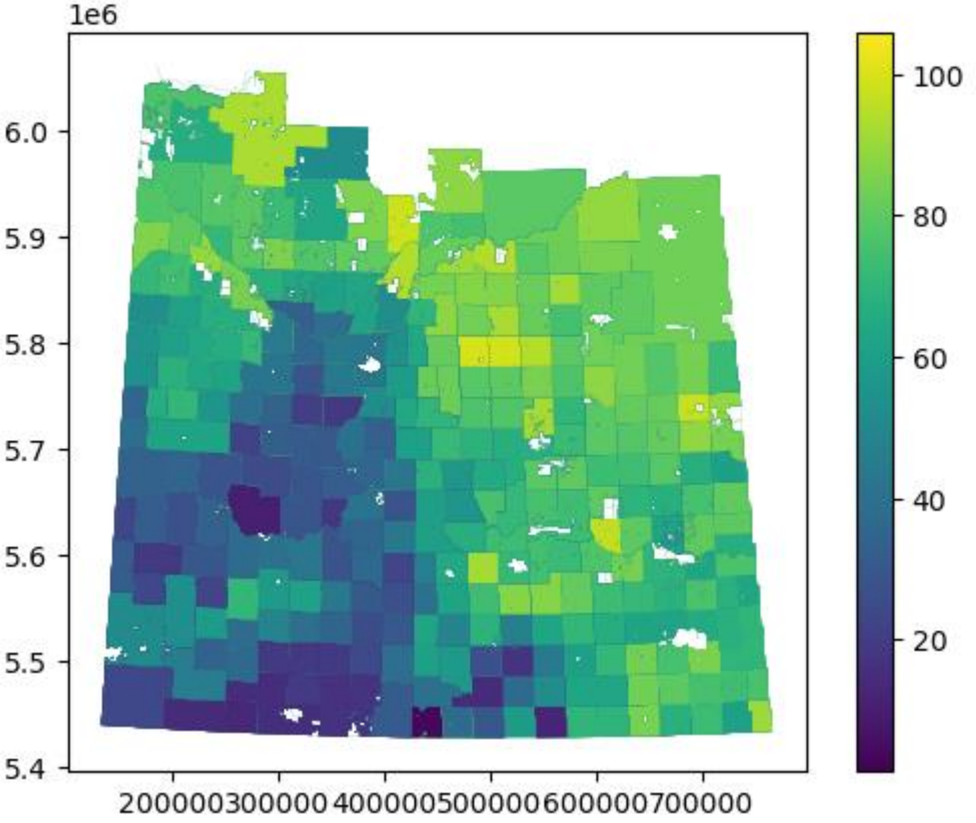
Correlation Matrix

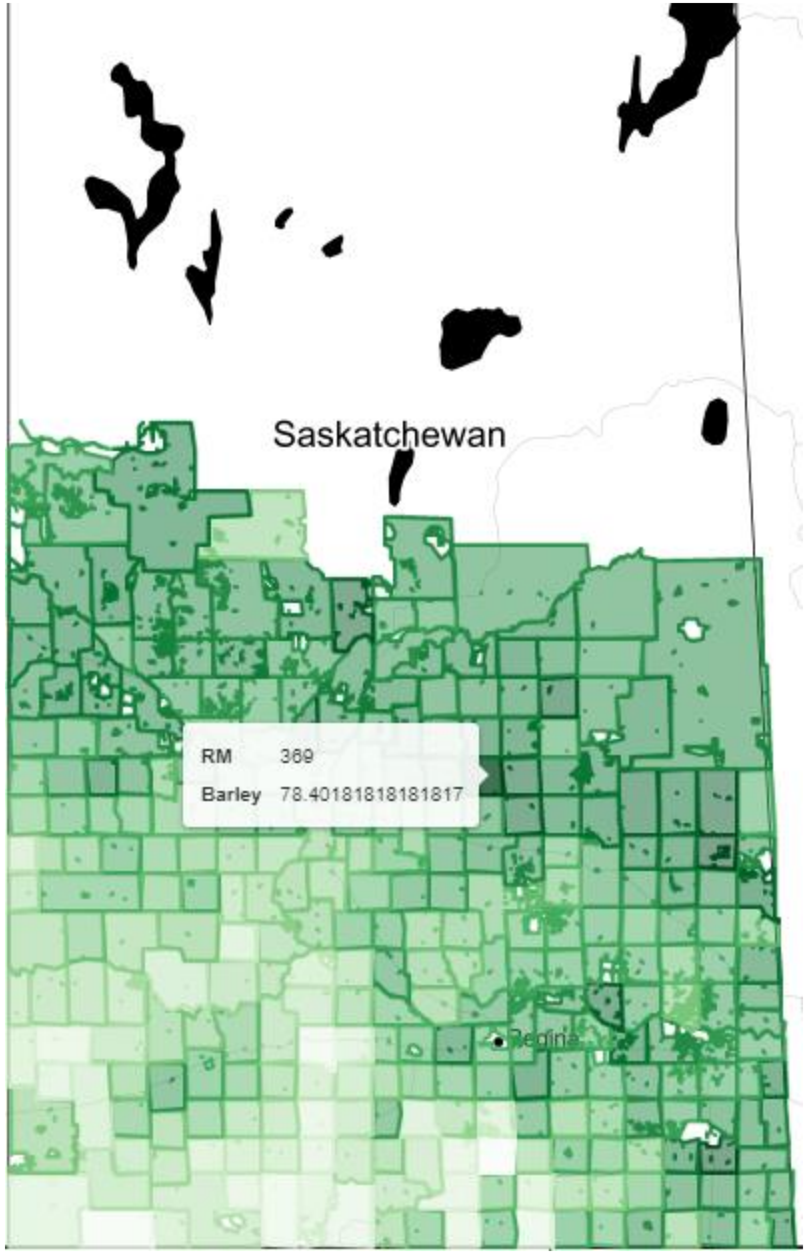


	Year	RM	Winter Wheat	Canola	Spring Wheat	Mustard	Durum	Sunflowers	Oats	Lentils	Peas	Barley	Fall Rye	Canary Seed	Spring Rye	Tame Hay	Flax	Chickpeas
Year	1	-0.0065	0.41	0.61	0.67	-0.2	0.49	-0.67	0.63	0.33	0.3	0.68	0.65	0.21	-0.47	-0.4	0.72	0.16
RM	-0.0065	1	-0.018	0.1	0.18	0.18	0.038	-0.26	0.21	0.026	0.18	0.14	0.04	0.088	0.16	0.15	0.15	-0.018
Winter Wheat	0.41	-0.018	1	0.55	0.55	0.2	0.56	-0.098	0.44	0.34	0.46	0.46	0.63	0.36	-0.31	-0.27	0.43	0.34
Canola	0.61	0.1	0.55	1	0.84	0.37	0.79	-0.47	0.73	0.61	0.71	0.78	0.56	0.55	-0.33	0.018	0.76	0.44
Spring Wheat	0.67	0.18	0.55	0.84	1	0.39	0.88	-0.46	0.88	0.59	0.72	0.9	0.73	0.61	-0.31	0.19	0.83	0.4
Mustard	-0.2	0.18	0.2	0.37	0.39	1	0.38	0.19	0.37	0.36	0.44	0.46	0.2	0.43	0.17	0.19	0.42	0.3
Durum	0.49	0.038	0.56	0.79	0.88	0.38	1	-0.29	0.7	0.62	0.65	0.78	0.64	0.56	-0.22	0.27	0.75	0.49
Sunflowers	-0.67	-0.26	-0.098	-0.47	-0.46	0.19	-0.29	1	-0.33	-0.11	0.0081	-0.08	0.0044	0.022	0.65	0.77	-0.19	-0.23
Oats	0.63	0.21	0.44	0.73	0.88	0.37	0.7	-0.33	1	0.4	0.62	0.89	0.68	0.52	-0.21	0.18	0.8	0.29
Lentils	0.33	0.026	0.34	0.61	0.59	0.36	0.62	-0.11	0.4	1	0.61	0.54	0.27	0.46	-0.25	-0.16	0.52	0.56
Peas	0.3	0.18	0.46	0.71	0.72	0.44	0.65	0.0081	0.62	0.61	1	0.72	0.39	0.54	0.014	-0.025	0.58	0.43
Barley	0.68	0.14	0.46	0.78	0.9	0.46	0.78	-0.08	0.89	0.54	0.72	1	0.72	0.6	-0.11	0.28	0.84	0.36
Fall Rye	0.65	0.04	0.63	0.56	0.73	0.2	0.64	0.0044	0.68	0.27	0.39	0.72	1	0.31	0.023	0.34	0.7	0.24
Canary Seed	0.21	0.088	0.36	0.55	0.61	0.43	0.56	0.022	0.52	0.46	0.54	0.6	0.31	1	-0.025	0.036	0.53	0.31
Spring Rye	-0.47	0.16	-0.31	-0.33	-0.31	0.17	-0.22	0.65	-0.21	-0.25	0.014	-0.11	0.023	-0.025	1	0.36	-0.19	-0.25
Tame Hay	-0.4	0.15	-0.27	0.018	0.19	0.19	0.27	0.77	0.18	-0.16	-0.025	0.28	0.34	0.036	0.36	1	0.26	
Flax	0.72	0.15	0.43	0.76	0.83	0.42	0.75	-0.19	0.8	0.52	0.58	0.84	0.7	0.53	-0.19	0.26	1	0.41
Chickpeas	0.16	-0.018	0.34	0.44	0.4	0.3	0.49	-0.23	0.29	0.56	0.43	0.36	0.24	0.31	-0.25		0.41	1

GIS Analysis

The place has the highest average Barley yield in the last 10 year is RM: 369



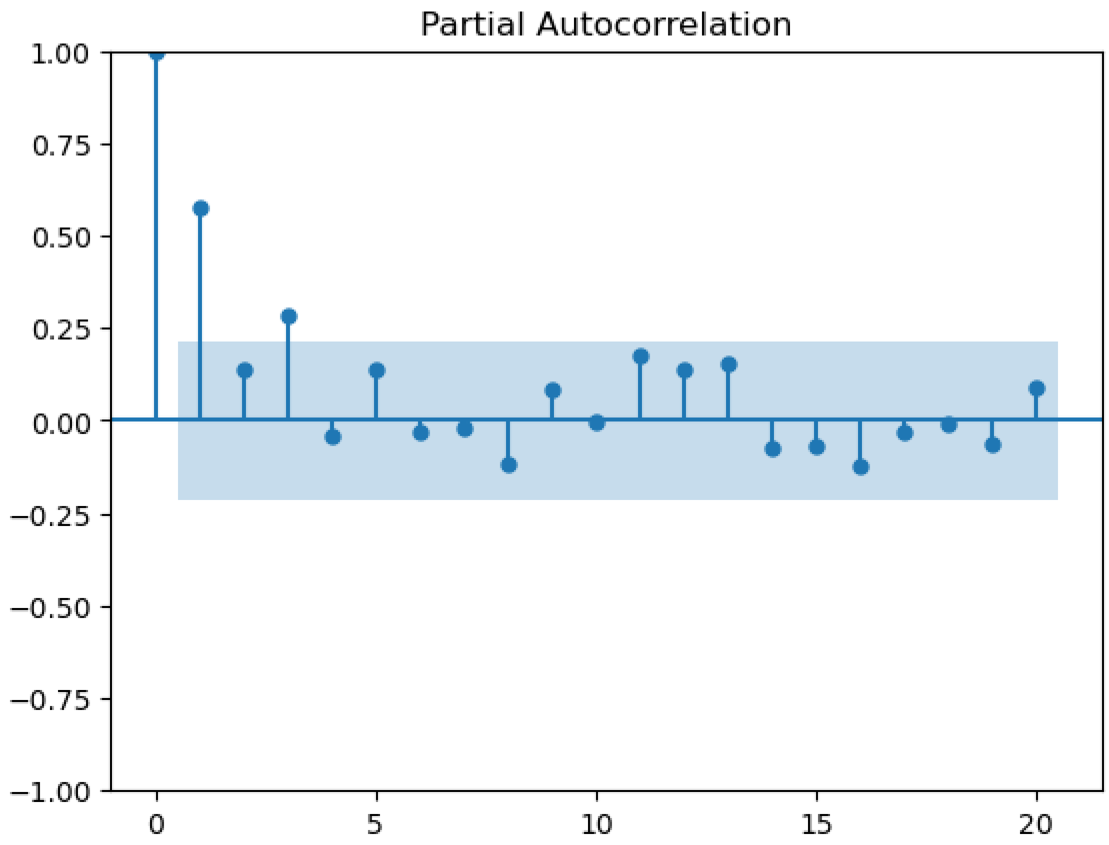


RM	369
Barley	78.40181818181817

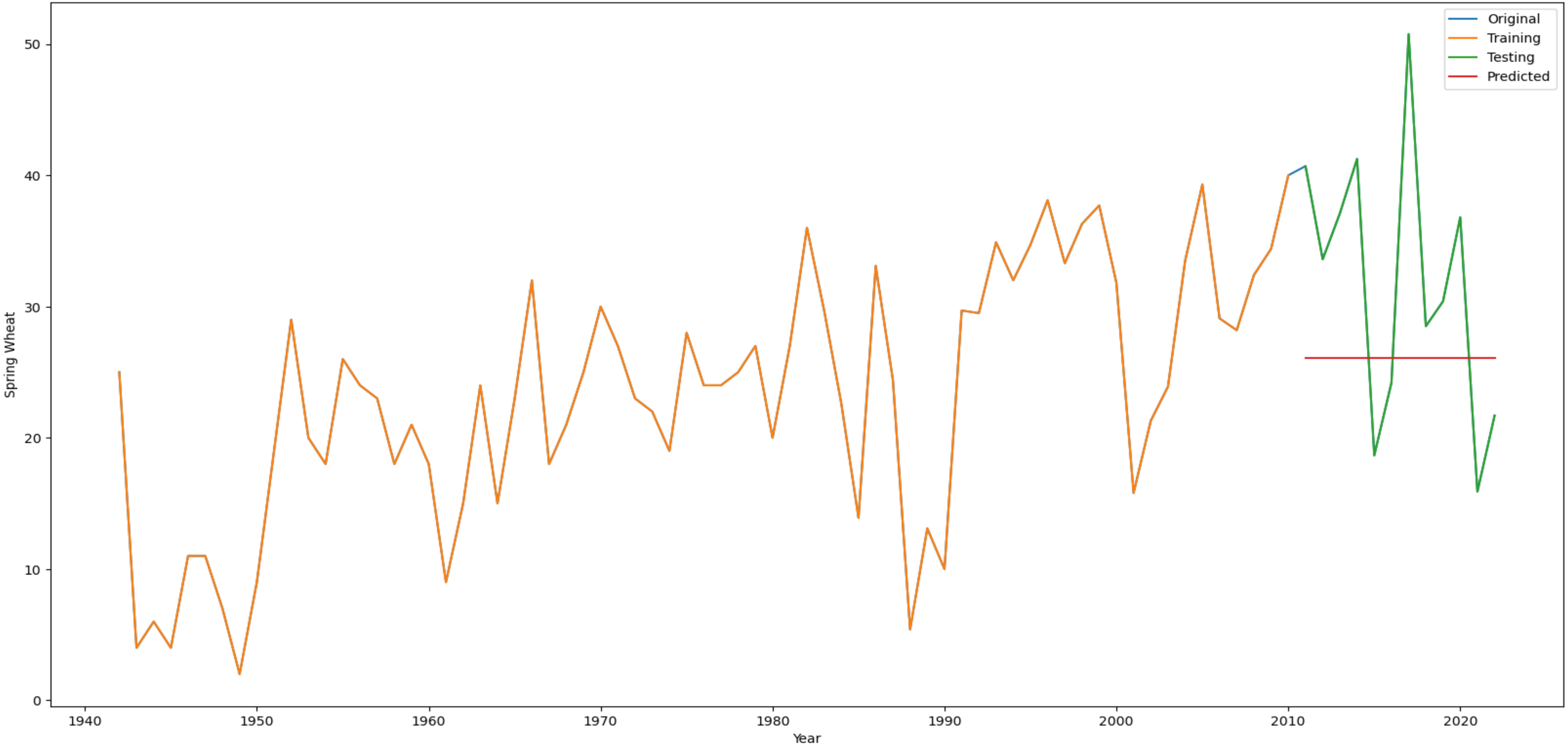
Time Series: Partial Autocorrelation (Spring Wheat)

- When analyzing the plot, we can see that the first lag has a very strong correlation to our future value.

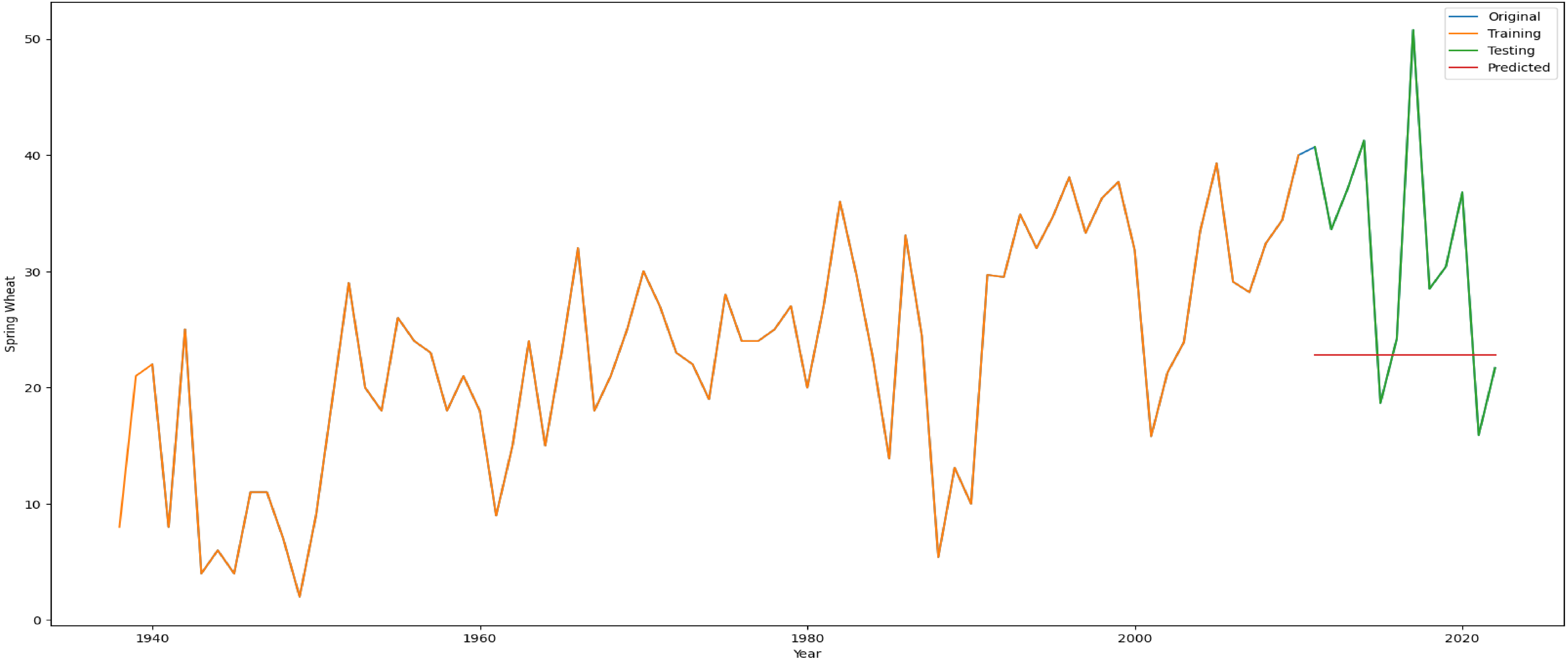
- Lag 5 is the last lag the clearly goes above the green threshold line. As such, we now know to use 5 lags to create our auto regression model



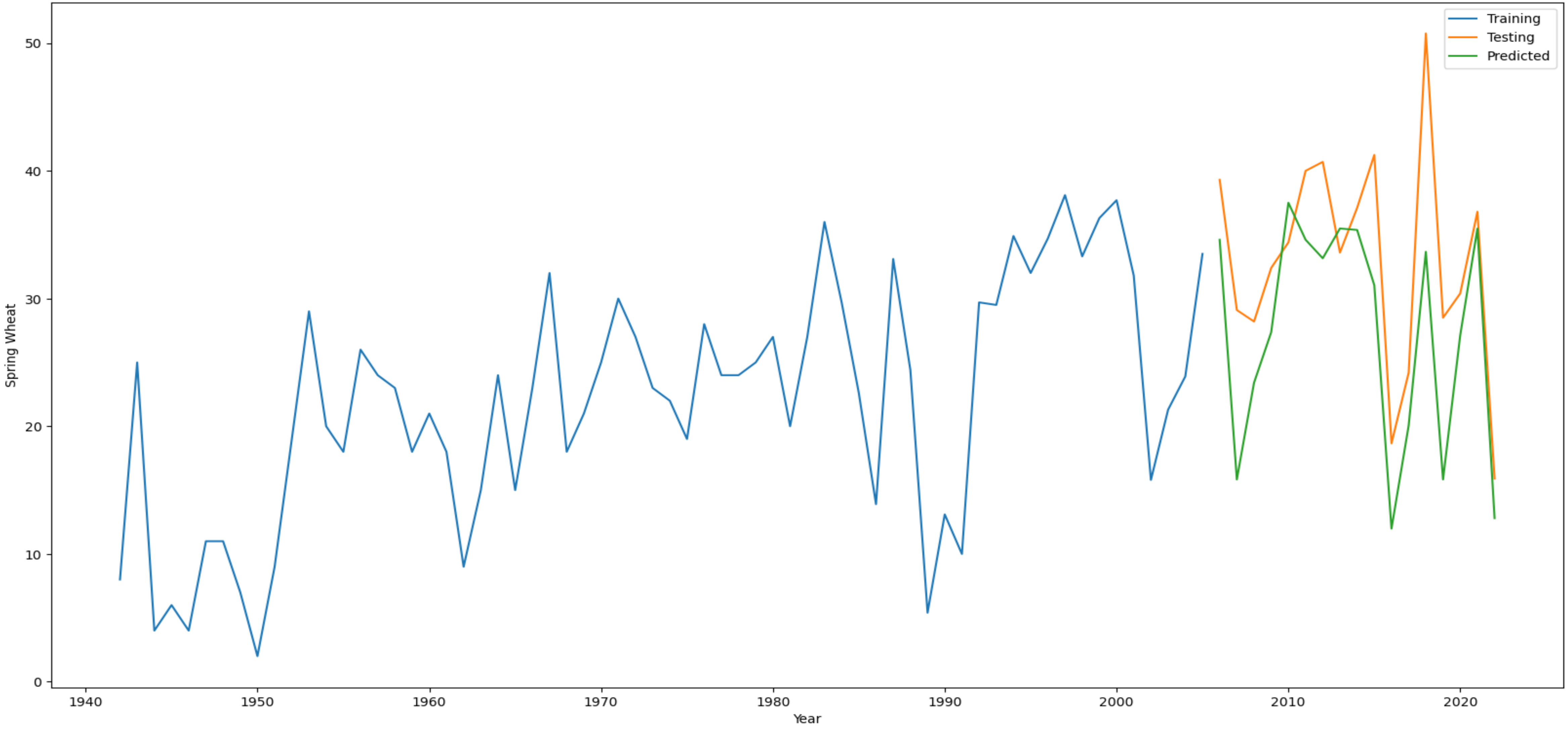
Time Series (Autoregressive model)



Time Series: ARIMA

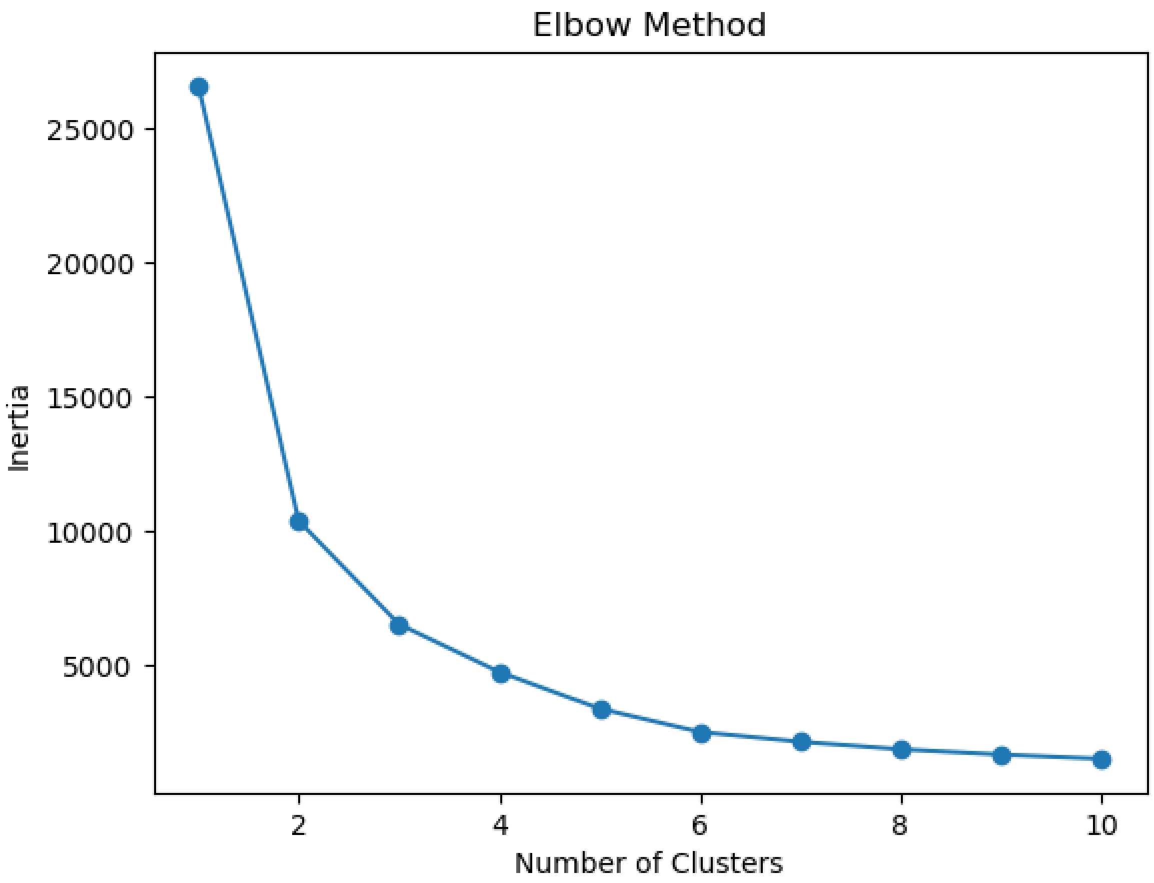


Time Series: XGBoost



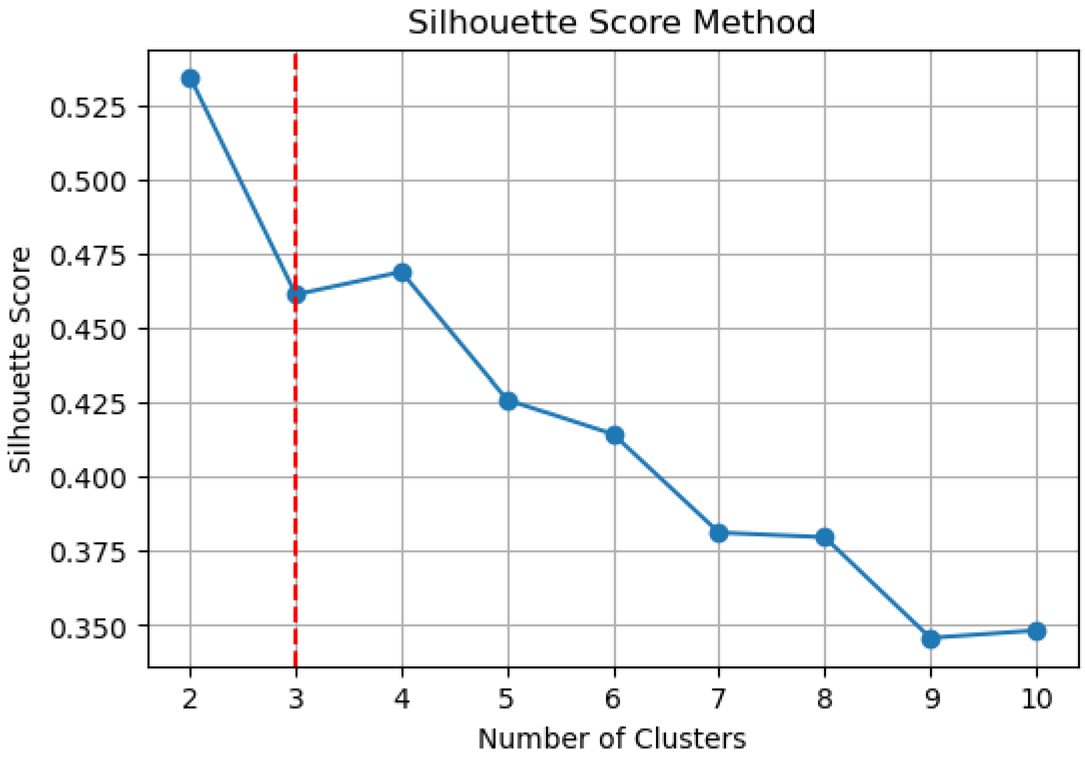
Unsupervised Learning – k number

- The image show that the k=5 is not a bad choice



Silhouette Analysis

- Used to determine the degree of separation between clusters



Thank you