Title: Final Report

By: Group 4

ALY – 6980: Capstone

Instructor: Jay Qi

Date: June 21, 2023

Table of Contents

Article Reference and Annotated Bibliography

Lobell, D. B., & Asner, G. P. (2003). Climate and Management Contributions to Recent Trends in U.S. Agricultural Yields. *Science*, *299*(5609), 1032. https://doi.org/10.1126/science.1078475

Summary:

Lobell and Asner's research investigates the relationship between climate variation and crop production in the United States, focusing on corn and soybean yields from 1982 to 1998. The study uses data on temperature, precipitation, and solar radiation to examine the impact of climate on crop yields. The authors identify two distinct regions with varying responses to climate anomalies: the Midwest, where cooler and wetter conditions favored higher yields, and the Northern Great Plains, where hotter and drier conditions led to better yields. They find that roughly 25% of corn and 32% of soybean yield trends can be explained by temperature changes. The results also suggest that non-climatic factors such as management practices and technological advancements account for approximately 20% less yield improvement than previously assumed.

Relevance:

This study is relevant to the capstone project as it highlights the importance of considering climate factors, particularly temperature, when predicting crop yields. The provided data set for the project includes climate data similar to the variables used in this research, making this study a valuable reference for developing a hops yield prediction model. Furthermore, Lobell and Asner's work emphasizes the role of management practices in mitigating the impacts of adverse climate conditions on

crop yields, which could be considered when interpreting the results of the capstone project's prediction model. Overall, this research offers a useful foundation for understanding the relationship between climate factors and crop yield trends, which is crucial for the success of the capstone project.

Mourtzinis, S., Esker, P. D., Specht, J. E., & Conley, S. P. (2021). *Advancing agricultural research using machine learning algorithms*. Scientific Reports, 11, 17879. **https://rdcu.be/daJhf**

Summary:

The possible benefits of machine learning (ML) algorithms in enhancing agricultural research are covered in the paper "Advancing agricultural research using machine learning algorithms" published in Scientific Reports. According to the authors, ML algorithms can assist agricultural researchers in managing the growing volume of data produced by numerous sources, including remote sensing, plant breeding, and field experiments. They present examples of how ML algorithms can be applied to improve soil management, disease detection, and crop production prediction. They also address the application of ML in various domains. The limitations of access to large-scale datasets, the requirement for specialized ML knowledge, and the lack of standardization in data collecting and processing are all topics covered in the study. The authors argue that these issues can be solved by working together to establish open-source software and platforms for data and model sharing and through research collaboration. It was presumed that stated management practices (generic categories) were uniform across all sites because data were gathered from various states and years. Fertilizer type and application technique were also hardly recorded.

Overall, the report emphasizes how ML algorithms have the potential to advance agricultural research and enhance sustainability, disease management, and crop output. It also sheds light on the difficulties and possibilities of using ML in agriculture.

Relevance:

Our capstone project on agricultural hops yield is relevant to the paper "Advancing agricultural research using machine learning algorithms" because it examines the use of machine learning algorithms in advancing agricultural research. The research shows, among other applications, the potential of machine learning techniques to improve crop yield projections and crop disease detection. Utilizing these methods in hops cultivation can assist anticipate and optimize yield as well as identify and take care of any potential disease or insect problems. Farmers can manage their crops more effectively by using machine learning algorithms, which may result in higher yields and profitability.

IoT-Equipped and AI-Enabled Next Generation Smart Agriculture: A Critical Review, Current Challenges and Future Trends. (2022). IEEE Journals & Magazine | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/9716089

Summary:

According to UNESCO, the population will increase by one-third and the world will then need more food and water. This necessitates the development of smart agricultural systems that remotely monitor and report crop, field, and weather factors using low-cost, low-power wireless sensors. The application of IoT in smart

agriculture has helped farmers to manage resources more efficiently, such as by reducing irrigation water requirements and the use of hazardous pesticides. The practices of traditional agriculture have been transformed by these two (IoT and AI) enabling technologies. For intelligent agriculture, wireless sensor technologies are crucial. For example, since some crops are sensitive to temperature fluctuations, temperature sensors are required for monitoring ambient temperature in indoor and outdoor smart farms. Humidity sensors are essential for calculating water losses through evaporation, which is important for the photosynthetic process.

Relevance:

The capstone project that we are working on has a similar goal. We have to optimize the quality and growth of hops which can save us costs while being environmentally friendly reducing the wastage of resources. IoT and AI-enabled smart agriculture systems have the potential to transform current agricultural practices. They enable farmers to use less toxic pesticides, conserve resources like water, manage resources more effectively, and decrease crop disease and insect infestation. Their widespread implementation is still hampered by problems including high costs, a lack of technical expertise, and interoperability with different software systems and sensor equipment. Because they can meet the growing need for food and water, these technologies must be developed and advanced further.

Swain, M., Singh, R., Gehlot, A., Hashmi, M. F., Kumar, S., & Parmar, M. (2019, December 1). A reliable approach to customizing linux kernel using custom build tool-chain for ARM

architecture and application to agriculture. *International Journal of Electrical and Computer Engineering (IJECE)*, *9*(6), 4920. https://doi.org/10.11591/ijece.v9i6.pp4920-4928

The paper titled "A Machine Learning Approach of Data Mining in Agriculture 4.0" proposes a framework for using machine learning algorithms to analyze and predict crop yields in the context of Agriculture 4.0. The authors highlight the importance of data mining and machine learning in modern agricultural practices and suggest using a combination of machine learning techniques such as regression, classification, clustering, and time series analysis to build a crop prediction model.

Summary:

The paper discusses the framework for data mining and machine learning in agriculture, which involves collecting data from various sources, preprocessing the data, and applying machine learning algorithms to analyze and predict crop yields. The authors recommend using different algorithms and selecting the one that provides the best accuracy for a specific problem. They suggest that building yield models that provide high accuracy across numerous varieties, locations, and years is challenging due to the complexity of the problem and the massive amount of data involved. The authors suggest that creating yield models for a specific location and specific variety would provide the highest accuracy.

Relevance to the project:

The paper's relevance lies in providing insights into how machine learning techniques can be applied to agriculture data mining to predict crop yields. The framework for building a crop prediction model is useful, and the suggestions regarding the accuracy of yield models across different scenarios provide useful

guidance for researchers and practitioners in the agricultural domain. For the project, the suggestions on creating yield models for a specific location and specific variety would be the most relevant and useful, as it aligns with the current trend of precision agriculture.

Team Communication Protocol

GROUP 4

	NUID	e-mail ID
Dhairya Dave (Team Leader)	002110382	dave.dh@northeastern.edu
Kunal Sindhu	002965515	sindhu.k@northeastern.edu
Parth Sawant	002123792	sawant.pa@northeastern.edu
Pratikraj Solanki	002927142	solanki.pra@northeastern.edu

Communication Goals:

- To keep each other informed about the tasks and deadlines
- Offer help and ask for it whenever needed
- Define information to share with the sponsor and instructor

Summary of the Project:

This Capstone project is about Hops production. We will be predicting the growth of hops by studying the effects of climate and soil conditions using historical data provided.

Sponsor Information

Agritecture is a technology and consultancy services company that specializes in climate-smart agriculture, especially CEA in cities. Their goal is to hasten and facilitate the transition to more intelligent and resilient agriculture.

Communication Protocol

	Meeting 1	Meeting 2
Weekly Meetings	Every Wednesday 10am to 11 am	Saturday Class Day: 2pm - 3pm No Class Day: 5pm - 6pm
Meeting Agenda	Present submission and next week's work distribution	Discussion of all the work done and problems faced
Meeting Leader	Dhairya & Pratikraj (Alternating every week)	Kunal & Parth (Alternating Every Week)
Tool used for Meeting	Microsoft Teams	
Best time to contact about concerns/questions during meetings	During the meeting itself by raising a hand on Microsoft Teams	
Tool to reach out between weekly meetings	Microsoft Teams Group and WhatsApp Group	

Exploratory Data Analysis

Introduction

The project is about optimizing the growth of Hops and predicting future ideas to improve the quality and quantity of the crop produced.

The goal of the Agritecture-sponsored Capstone project is to increase hops output and quality through data analysis and modelling. The project's main goal is to create models and methods that can improve hops yield and quality via the use of data-driven insights. To accomplish the objective, the research will use statistical analysis, machine learning, and optimization strategies.

The Data

The dataset consists of multiple tables that explain what the project is about. There are tables that have data from 2000 to 2022 for three states viz. Washington, Oregon and Idaho. The tables consist total acres harvested and yield per acre. There are tables that consist of climate data for all three states and also satellite data that contains Surface soil wetness, moisture, root moisture, cloud amount, etc. for the three states. There are some introductory tables that have some notes and correlations about the data which can help in modelling.

Goal

In order to maximize yield and quality, the project will create optimization approaches. The many variables and the correlations between them discovered by the models will be considered by the optimization strategies. The best practices for hop production that can maximize output and quality while minimizing resource consumption will be suggested using optimization approaches.

Exploratory Data Analysis

To understand the variables that affect yield and quality, the project's initial stage is analyzing the data and then verifying all the correlations in these variables. Data on soil properties, weather trends, irrigation patterns, and fertilization techniques is given and will be analyzed. The main factors that affect the production and quality of hops will be identified using this data. To understand the correlations between various factors and their effects on yield and quality, the project team will then create models utilizing statistical analysis and machine learning approaches. The models may be tested against real-time data from hops farms after being trained using historical data. The models may provide guidance on how to improve yield and quality while also identifying the ideal conditions for hops production.

In order to maximize yield and quality, the project will create optimization approaches. The many variables and the correlations between them discovered by the models will be considered by the optimization strategies. The best practices for hops production that can maximize output and quality while minimizing resource consumption will be suggested using optimization approaches.

The data is analyzed in three parts, i.e. for all three states separately.

The first step is to create a temporary data frame that stores all the values. The dataframe is then transformed to create a dataframe df_total_acre_WA (Dataframe that has total acres harvested in Washington with years as columns).

PART 1 - Washington

Figure 1

Dataframe df_total_acre_WA - WASHINGTON

Variety	Ahtanum, YCR 1	Amarillo, VGXP01	Apollo	Azacca, ADHA-483	Bravo	Calypso	Cascade
2022	168	1324	807	871	203	0	3604
2021	166	1334	0	730	238	0	3183
2020	230	1395	750	722	201	0	2877
2019	261	1597	851	589	236	0	3718
2018	255	1895	795	546	280	0	4274
2017	371	1984	684	578	486	0	4966
2016	155	0	735	506	573	0	5582
2015	145	0	708	175	569	0	4935
2014	194	0	700	79	584	0	4837
2013	211	0	701	0	493	0	4237
2012	176	0	874	0	528	0	2693
2011	0	0	885	0	593	0	2108

The data frame is then checked for null values and some values are replaced with zero.

Figure 2

Total acres harvested and Average yield per acre - WASHINGTON

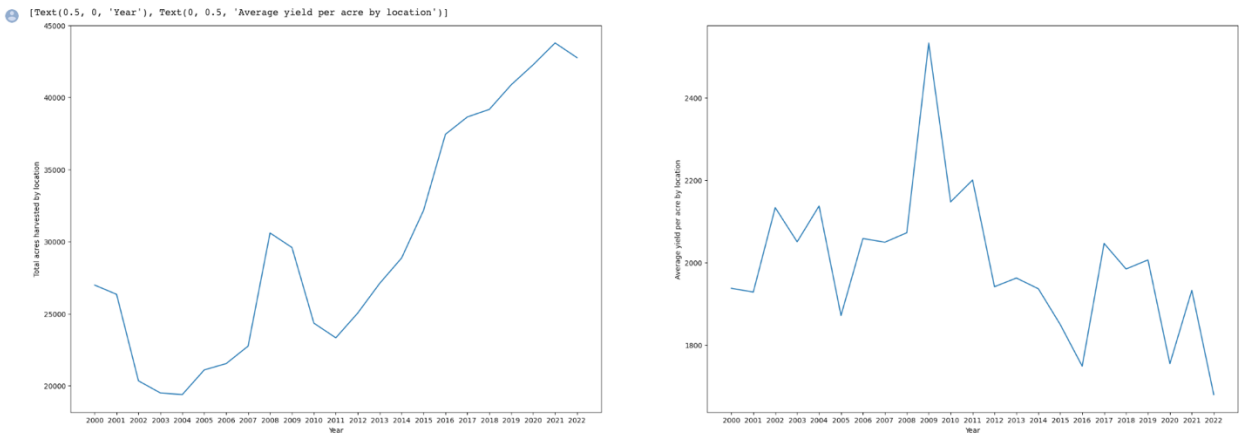
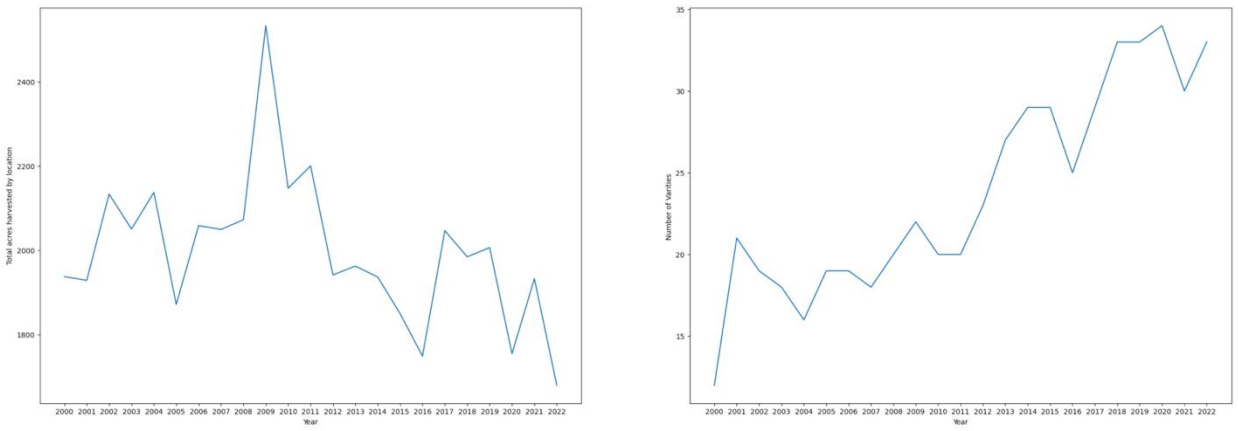


Figure 2 compares the annual average yield per acre with the total number of acres harvested in Washington State. While the average yield per acre has fallen over time, Washington's total harvested acres have climbed overall. The overall area harvested declined to roughly 18000 acres in 2004 from a starting point of about 27000 acres in 2000. Following the abrupt decline, the total number of harvested acres saw an abrupt rise that persisted until 2008, when it peaked at 30000 acres. After that, there was a deficiency in the overall number of acres harvested, which fell to 23000 acres in 2011. Following the fall of total acres in 2011, Washington saw an annual increase every year in the number of acres harvested. A staggering 44000 acres were harvested in 2021, which was the highest. The average production per acre began about 1950, declined somewhat in 2001, had two peaks in 2002 and 2004, and subsequently fell to a little less than 1900. The average yield per acre then continued to rise until 2009, when it peaked at about 2550. This was the greatest average output per acre ever recorded in Washington. It declined until 2016 to the minimum value of 1700, then increased for the next four years until ultimately falling to the minimum number.

Figure

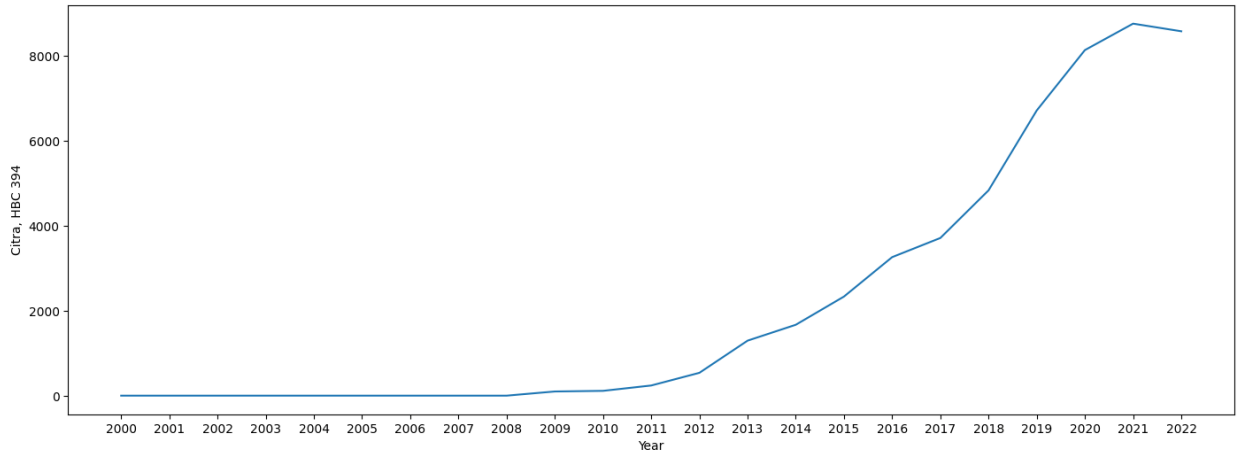
Total Acres harvested by location and number of varieties by location



In the above figure, it is conspicuous that year by year, There were new varieties being cultivated and there has been an increase in the total acreage of hop cultivation.

Figure 3

Citra, HBC 394 total acres per year in Washington



It is conspicuous in the figure 3 that the variety Citra was not produced in the years 2000 to 2011. It started production in 2012 and kept on increasing till 2022. Citra is the most sought-after hop variety because it offers an intense citrusy flavour and aroma that revolutionized IPAs. (Yakima Valley Hops, n.d.) It reached the maximum value in 2021 at around 8700.

We tried analyzing some varieties as shown in Figure 4 in a similar way as above.

Figure 4

Five varieties compared to each other in terms of annual total acres produced in Washington state - WASHINGTON

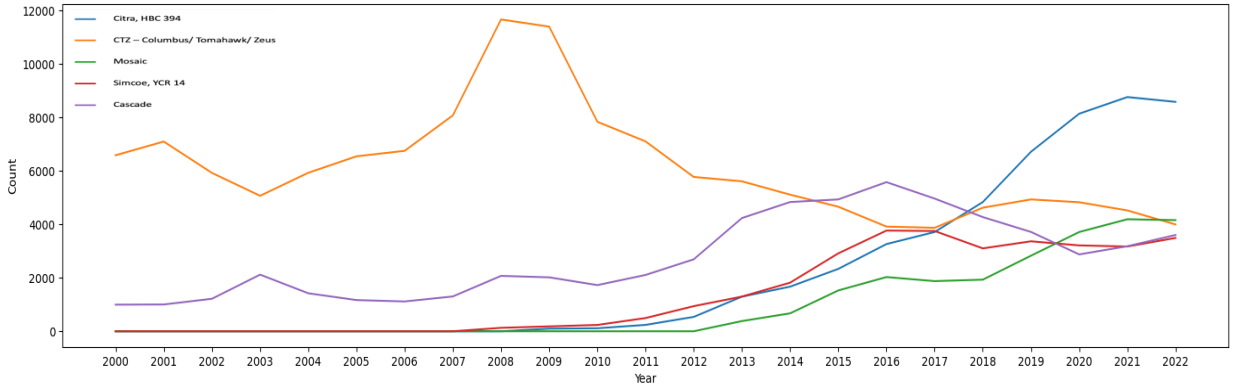


Figure 4 shows that 5 varieties were chosen for examination of total acres harvested yearly. According to folklore, Citra, CTZ, Mosaic, Simcoe, and Cascade are shown in blue, orange, green, red, and purple. It can be observed that CTZ was collected in the greatest quantity, which peaked in 2008 and 2009 at over 12000 acres per year.

Figure 5

Data frame for climate data arrange by day - WASHINGTON

Date	ppt (inches	tmin (degrees F	tmean (degrees F	tmax (degrees F	vpdmin (hPa	vpdmax (hPa	ETO (FAO)	ETO (Hargreaves)	Hops factor (HF)	ETO x HF
1999-01-31 0:00	0.037419355	29.34193548	37.46774194	45.57741935	0.328064516	4.199032258	0.90580645	0.665483871	0	0
1999-02-28	0.039285714	30.43214286	39.55357143	48.67142857	0.661428571	5.830357143	1.29785714	1.081071429	0	0
1999-03-31	0.003548387	32.9483871	44.1483871	55.34193548	1.001935484	9.54483871	2.16806452	1.999677419	0	0
1999-04-30	0.004333333	36.08666667	49.5	62.90666667	1.924	14.555	3.385	3.300666667	0.295	1.0428333
1999-05-31	0.010645161	41.76774194	55.23225806	68.70645161	2.216129032	17.92774194	4.33774194	4.38	0.6	2.6705774

Figure 5 represents the data frame for climate and now we will be arranging it to get a better insight.

Figure 6

Percentile distribution of climate data - WASHINGTON

	ppt (inches)	tmin (degrees F)	tmean (degrees F)	tmax (degrees F)	vpdmin (hPa)	vpdmax (hPa)	ETO (FAO)	ETO\n (Hargreaves)	Hops factor (HF)	ETO x HF
count	8674	8674	8674	8674	8674	8674	8674	8674	8674	8674
mean	0.019783	39.6245	52.16031	64.70382	1.299731	16.49526	3.194839	3.104931	0.332442	1.785014
std	0.065373	13.0049	15.85458	19.34991	1.610359	13.18107	2.067344	2.237528	0.391249	2.284051
min	0	-11.9	0.5	9.2	0	0	0.27	0.01	0	0
1%	0	10.173	18.346	24.473	0	0.2	0.45	0.2	0	0
2%	0	13.546	21.9	28.046	0	0.37	0.49	0.23	0	0
3%	0	16	24.1	29.819	0.01	0.51	0.51	0.26	0	0
5%	0	19.8	27.3	32.465	0.02	0.7965	0.55	0.31	0	0
10%	0	24.1	31.8	37.63	0.05	1.77	0.69	0.46	0	0
25%	0	30.1	40	50.4	0.19	5.94	1.24	0.97	0	0
50%	0	38.7	51.5	64.6	0.67	12.89	2.88	2.71	0	0
75%	0	50	65	80.7	1.76	25.03	5.0075	5.06	0.8	3.9212
95%	0.13	60.6	77.5	94.7	4.7935	42.47	6.65	6.8935	0.99	6.03
97%	0.19	62.581	79.2	96.8	5.7281	45.7881	6.95	7.2	1	6.436966
98%	0.23	63.9	80.4	98.1	6.3154	47.6516	7.1154	7.41	1	6.7054
99%	0.33	65.9	82.1	99.727	7.1108	50.8816	7.37	7.6627	1	7.08
max	1.34	75.4	92.2	111.4	12.72	75.27	8.8	9.3	1	8.8

Looking at the percentile distribution of climate data in Washington, we discovered that the majority of precipitation values were in the top 95% of the records. The Hops

component, which exhibited values only at 75% and above the percentile, behaved similarly. Figure 5 shows the minimum and highest values, as well as the standard deviation.

Figure 6

Data frame for climate data after arranging – WASHINGTON

Date	ppt (inches)	tmin (degrees F)	tmean (degrees F)	tmax (degrees F)	vpdmin (hPa)	vpdmax (hPa)	ETO (FAO)	ETO\n (Hargreaves)	Hops factor (HF)	ETO x HF
1999-01-31	0.037419	29.341935	37.467742	45.577419	0.328065	4.199032	0.905806	0.665484	0	0
1999-02-28	0.039286	30.432143	39.553571	48.671429	0.661429	5.830357	1.297857	1.081071	0	0
1999-03-31	0.003548	32.948387	44.148387	55.341935	1.001935	9.544839	2.168065	1.999677	0	0
1999-04-30	0.004333	36.086667	49.5	62.906667	1.924	14.555	3.385	3.300667	0.295	1.042833
1999-05-31	0.010645	41.767742	55.232258	68.706452	2.216129	17.927742	4.337742	4.38	0.6	2.67057

The climatic data in Figure 6 is organized and resampled according to yearly characteristics. This enabled us to quickly detect outliers and reduced the size of the data set for speedier processing. Furthermore, the harvest data supplied by the sponsor is annual, therefore it is critical to convert all of the data into the same kind to make models more accurate.

Figure 7

Outlier detection in climate data - WASHINGTON

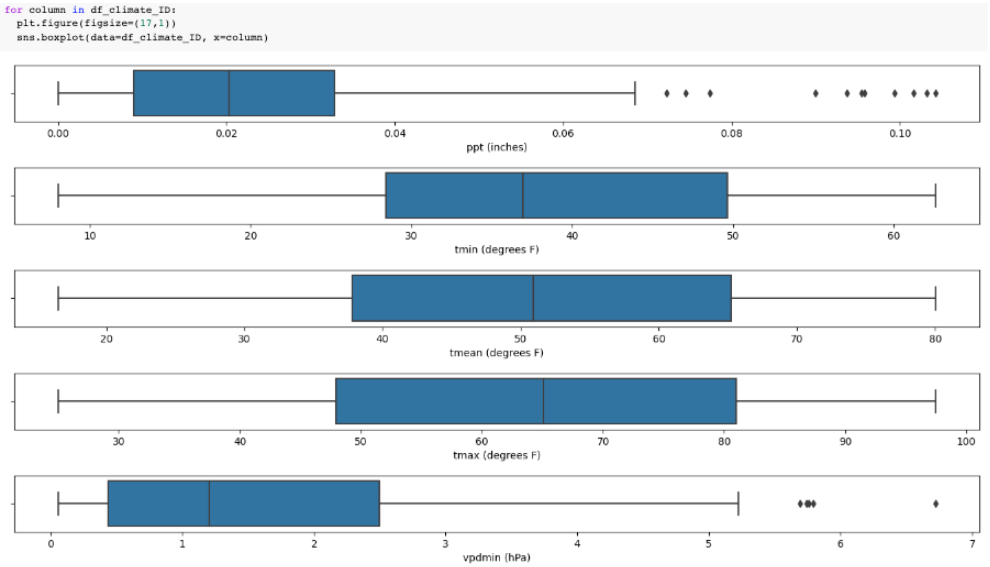


Figure 7 represents some outliers found in the climate data. The boxplot consists of all the features but the figure represents only the ones with outliers.

Figure 8
Outliers removed from climate data – WASHINGTON

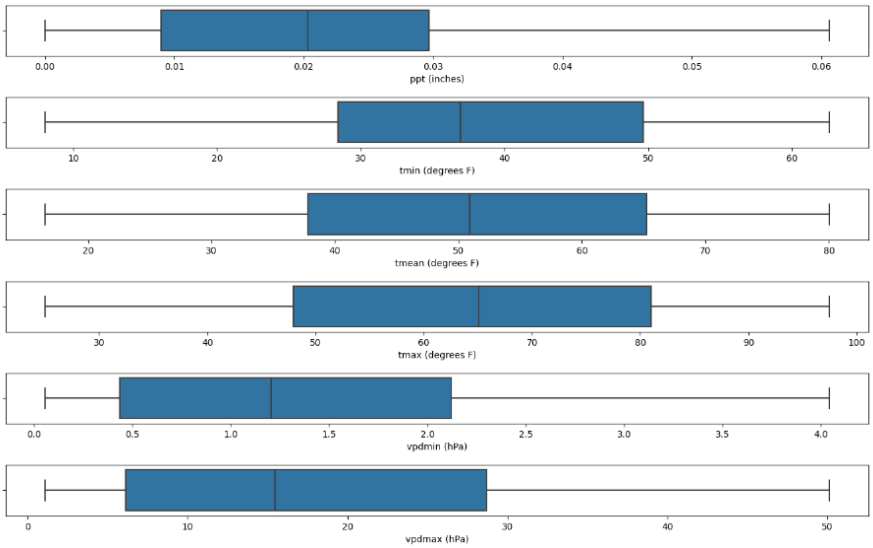


Figure 8 represents the situation when all the outliers were removed.

Figure 9

Correlation plot for climate - WASHINGTON

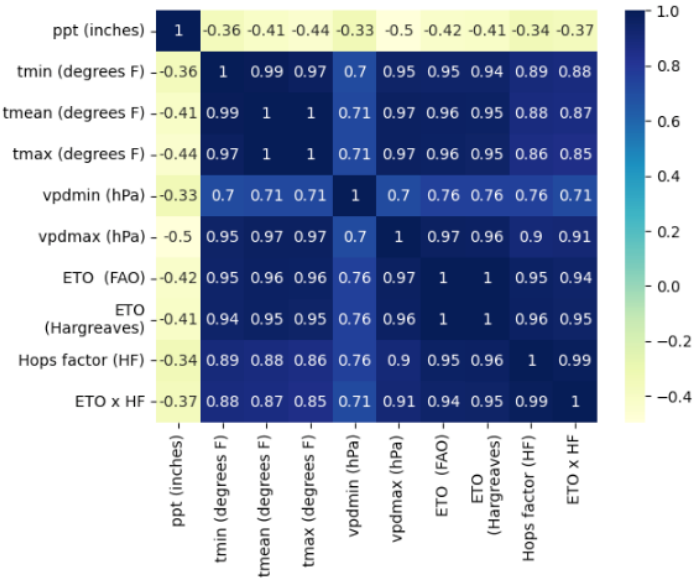


Figure 10

NASA Data – WASHINGTON

10	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	ANN
PARAMETER														
GWETTOP	1999	0.5	0.53	0.51	0.48	0.45	0.39	0.35	0.34	0.34	0.36	0.42	0.45	0.43
GWETTOP	2000	0.48	0.49	0.48	0.45	0.42	0.38	0.34	0.31	0.34	0.36	0.38	0.4	0.41
GWETTOP	2001	0.42	0.43	0.42	0.41	0.37	0.35	0.31	0.29	0.28	0.33	0.4	0.45	0.38
GWETTOP	2002	0.45	0.45	0.44	0.41	0.38	0.35	0.3	0.27	0.27	0.32	0.35	0.41	0.37
GWETTOP	2003	0.47	0.46	0.45	0.45	0.4	0.34	0.3	0.3	0.29	0.32	0.35	0.41	0.38

Figure 10 shows the data frame that stores the NASA values. Table 1 shows the abbreviations and what they stand for.

Table 1

Abbreviations for NASA Data

GWETTOP - MERRA-2 Surface Soil Wetness (1)
GWETPROF - MERRA-2 Profile Soil Moisture (1)
GWETROOT - MERRA-2 Root Zone Soil Wetness (1)
CLOUD_AMT - CERES SYN1deg Cloud Amount (%)
ALLSKY_SFC_PAR_TOT - CERES SYN1deg All Sky Surface PAR Total (W/m^2)
CLRSKY_SFC_PAR_TOT - CERES SYN1deg Clear Sky Surface PAR Total (W/m^2)

Figure 11

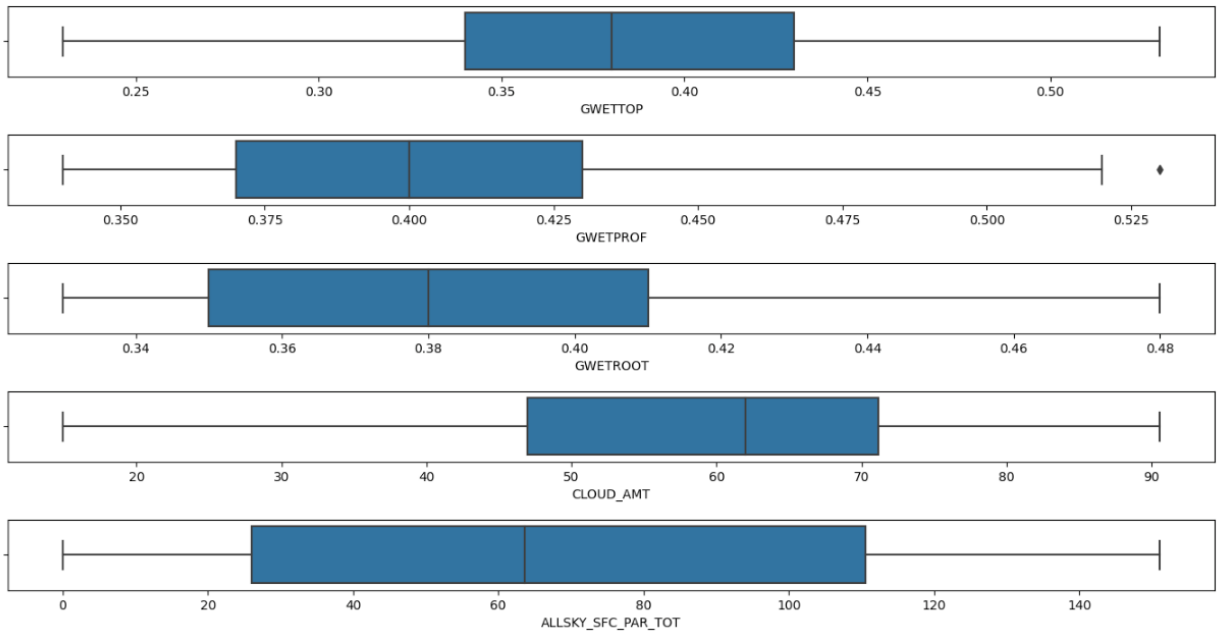
NASA Climate Data frame percentile distribution - WASHINGTON

index	GWETTOP	GWETPROF	GWETROOT	CLOUD_AMT	ALLSKY_SFC_PAR_TOT	CLRSKY_SFC_PAR_TOT
count	276	276	276	276	276	276
mean	0.379203	0.400761	0.380507	58.13533	68.36261	87.28116
std	0.060789	0.036984	0.034044	17.41437	44.99625	50.29442
min	0.23	0.34	0.33	14.91	0	0
1%	0.2575	0.34	0.33	18.5525	0	0
2%	0.27	0.34	0.33	21.235	0	0
3%	0.27	0.35	0.33	22.6375	0	0
5%	0.28	0.35	0.34	26.7725	0	0
10%	0.3	0.36	0.34	31.18	16.105	32.145
25%	0.34	0.37	0.35	46.93	26.085	42.98
50%	0.38	0.4	0.38	61.97	63.585	91.39

75%	0.43	0.43	0.41	71.1475	110.59	132.7175
95%	0.47	0.46	0.44	81.0875	138.375	157.58
97%	0.48	0.48	0.45	82.5925	141.175	158.075
98%	0.49	0.485	0.455	83.55	142.63	158.36
99%	0.4925	0.495	0.4625	85.0575	146.3025	159.215
max	0.53	0.53	0.48	90.59	151.12	159.84

Figure 12

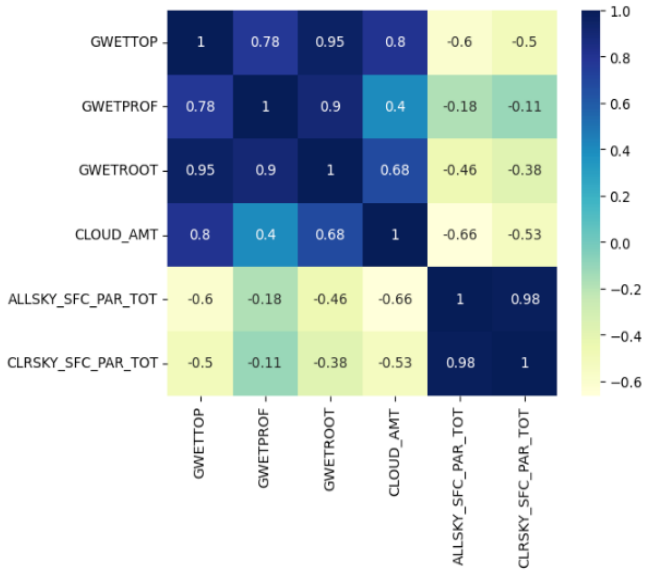
Checking the NASA Data for outliers



It is seen in the boxplots of NASA data that this table doesn't consist of any outliers so we can move forward to correlation.

Figure 13

Correlation plot for NASA Data – WASHINGTON

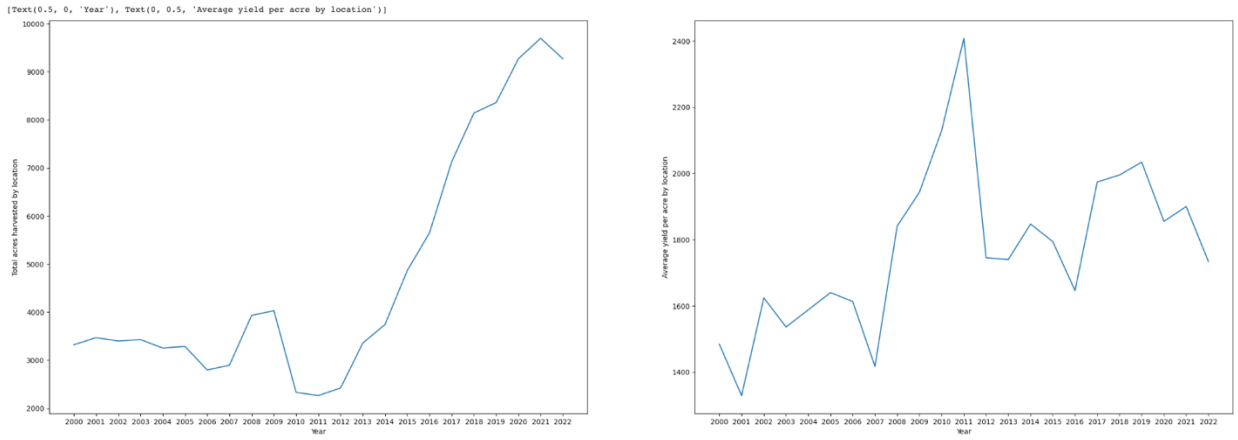


PART 2 - Idaho

A similar exploratory data analysis was performed for Idaho which will be briefly explained below.

Figure 14

Total acres harvested annually and average yield per acre for Idaho



It can be observed from figure 14 that until 2012, not many acres of hops were being farmed in Idaho, with only a few increments noted in 2008 and 2009. However, it continued to rise after 2012, reaching a peak of over 10000 acres each year in 2021. In contrast to total acres, the average yield per acre was observed to be the highest in 2011, at 2400. Overall, average output per acre has risen from roughly 1500 to 1600.

Figure 15

Comparison of 5 varieties of Hops - IDAHO

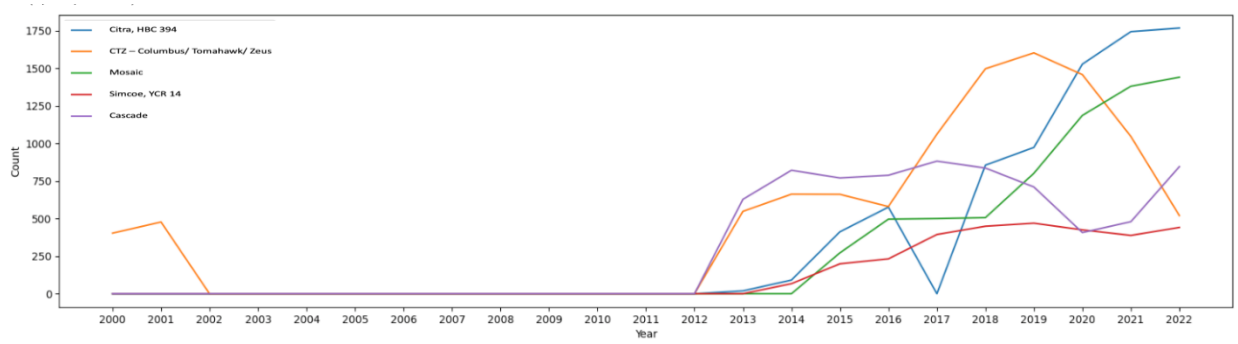


Figure 15 represents the trends in total acres harvested for 5 selected varieties.

Figure 16

Climate Data percentile distribution - IDAHO

	ppt (inches)	tmin (degrees F)	tmean (degrees F)	tmax (degrees F)	vpdmin (hPa)	vpdmax (hPa)	ETO (FAO)	ETO\n (Hargreaves)	Hops factor (HF)	ETO x HF	vdpmean(hPa)	ppt_sign	tmean_sign	vdpmean_sign
count	24	24	24	24	24	24	24	24	24	24	24	24	24	24
mean	0.024109	39.024875	52.127895	65.234991	1.644068	18.914969	3.283863	3.190607	0.329052	1.84775	8.645868	0.00551	22.388243	4.991953
std	0.005629	1.098754	1.205702	1.41934	0.283853	1.083212	0.097781	0.098592	0.000266	0.063977	0.481826	0.002585	0.58514	0.29802
min	0.012384	37.251233	50.159178	63.031507	1.161041	16.882164	3.101918	2.998986	0.328616	1.713418	7.766288	0.001467	21.133748	4.467155
1%	0.013476	37.260622	50.196545	63.115315	1.194621	16.919582	3.105163	3.002723	0.328616	1.718869	7.771181	0.00177	21.194973	4.472001
2%	0.014567	37.270011	50.233912	63.199123	1.228201	16.956999	3.108408	3.00646	0.328616	1.72432	7.776074	0.002073	21.256199	4.476846
3%	0.015659	37.2794	50.271279	63.282932	1.261781	16.994417	3.111653	3.010196	0.328616	1.729771	7.780967	0.002376	21.317424	4.481692
5%	0.017155	37.330624	50.374422	63.423096	1.310493	17.054959	3.119595	3.019975	0.328616	1.739954	7.815768	0.002787	21.462979	4.488377
10%	0.017518	37.617276	50.695832	63.57726	1.338063	17.408737	3.159784	3.065262	0.328616	1.761102	8.066141	0.002897	21.838382	4.494634
25%	0.019637	38.435	51.072808	63.909794	1.431112	18.352274	3.218301	3.133143	0.329038	1.811408	8.389952	0.00351	22.080104	4.873629
50%	0.024774	38.863758	52.205395	65.402186	1.631356	18.857352	3.292082	3.194096	0.329178	1.859747	8.591829	0.005083	22.354122	4.976722
75%	0.026761	39.699263	52.704641	65.96612	1.804658	19.506911	3.342628	3.249077	0.329178	1.888548	8.95616	0.00669	22.841588	5.144611
95%	0.032742	40.697918	53.976438	67.263781	2.125945	20.60351	3.396215	3.306373	0.329178	1.913673	9.361106	0.011013	22.904998	5.409487
97%	0.032934	40.903972	54.318365	67.738505	2.171806	20.821244	3.438269	3.34319	0.329283	1.942883	9.523437	0.011458	23.239864	5.456721
98%	0.032997	41.048802	54.554368	68.064253	2.204318	20.960452	3.468552	3.369786	0.32936	1.963342	9.617449	0.011458	23.487847	5.488176
99%	0.03306	41.193632	54.790371	68.390002	2.23683	21.099659	3.498835	3.396383	0.329438	1.9838	9.711461	0.011458	23.735831	5.51963
max	0.033123	41.338462	55.026374	68.715751	2.269342	21.238867	3.529118	3.422979	0.329515	2.004259	9.805472	0.011458	23.983815	5.551085

Idaho followed the same pattern as Washington(Figure 16). The precipitation column contains 95% of the data, whereas the hops factor contains 75%.

Figure 17

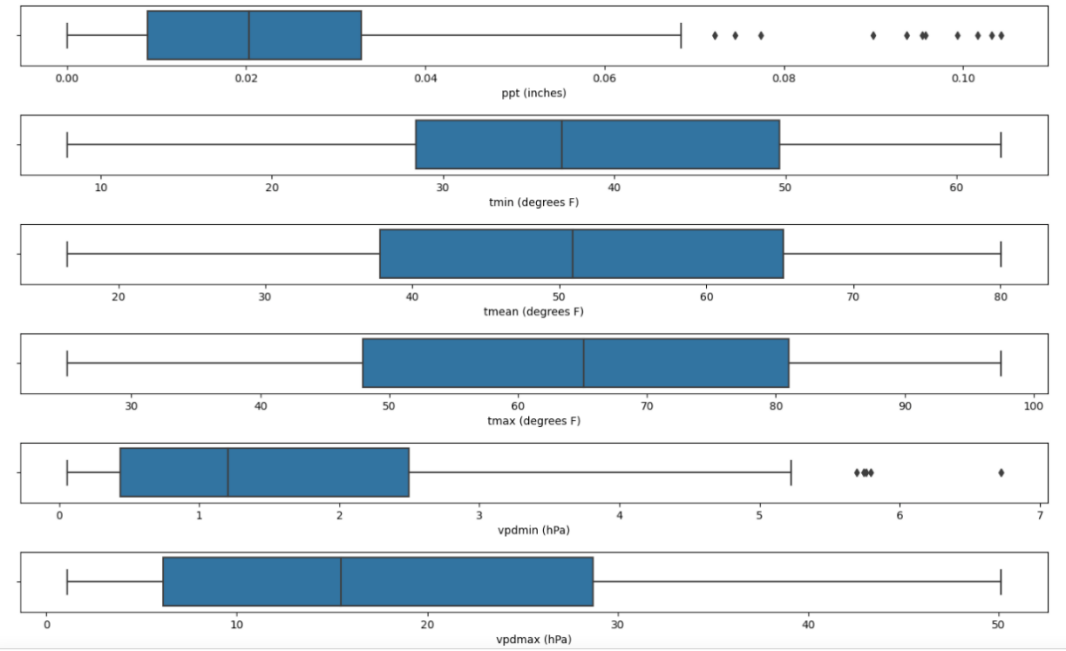
Climate data for Idaho changed to monthly

	ppt (inches)	tmean (degrees F)	vpdmin (hPa)	vpdmax (hPa)	ETO (FAO)	Hops factor (HF)	ETO x HF
Date							
1999-01-31	0.044839	35.529032	0.225484	3.354194	0.824839	0	0
1999-02-28	0.020323	36.296429	0.297143	4.160357	1.134286	0	0
1999-03-31	0.02129	43.758065	0.636452	10.653226	2.317419	0	0
1999-04-30	0.016333	48.176667	1.244333	14.159333	3.372	0.295	1.038283
1999-05-31	0.015161	57.067742	1.860645	20.737097	4.684516	0.6	2.897103
...	...	...	...	...	...	...	...
2022-05-31	0.020323	55.13871	1.276452	16.889355	4.222903	0.66	2.812945
2022-06-30	0.020323	64.903333	1.983667	25.627	5.468667	0.943	5.190413
2022-07-31	0	77.841935	1.204667	46.958387	6.912581	0.893548	6.152948
2022-08-31	0.007097	77.83871	1.204667	44.480323	6.080323	0.8	4.864258
2022-09-30	0.003333	67.663333	3.195333	33.291333	4.357333	0.24	1.345867
285 rows × 7 columns							

In Figure 17 it can be seen that the climate data was daily which was later on changed to monthly for the sake of simplicity.

Figure 18

Outlier detection - IDAHO



In Figure 18 it can be observed that the minimum value of vapor pressure deficit is having a number of outliers and the same is valid for precipitation.

Figure 19

Outliers removed - IDAHO

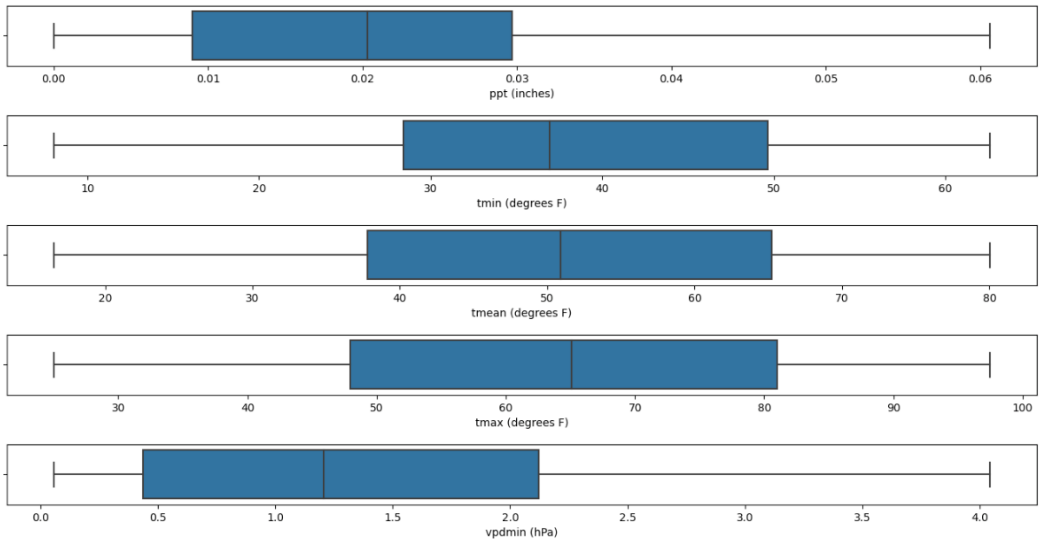


Figure 20

Corrplot for climate data - IDAHO

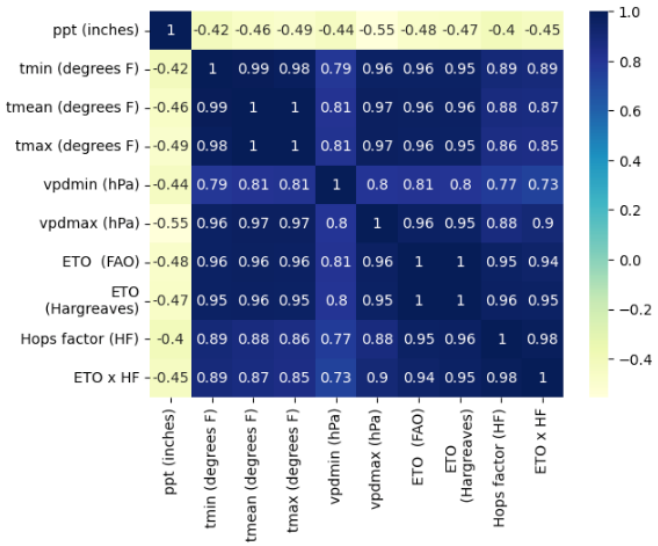


Figure 20 shows the correlation among all the features of climate data.

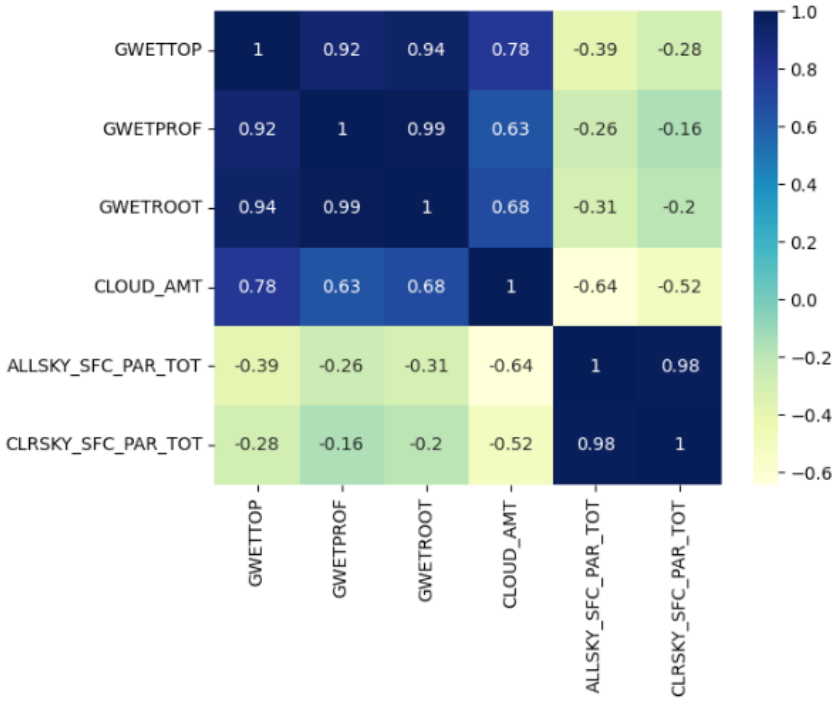
Figure 21

NASA Data - IDAHO

	ppt (inches)	tmean (degrees F)	vpdmin (hPa)	vpdmax (hPa)	ETO (FAO)	Hops factor (HF)	ETO x HF
Date							
1999-01-31	0.044839	35.529032	0.225484	3.354194	0.824839	0	0
1999-02-28	0.020323	36.296429	0.297143	4.160357	1.134286	0	0
1999-03-31	0.02129	43.758065	0.636452	10.653226	2.317419	0	0
1999-04-30	0.016333	48.176667	1.244333	14.159333	3.372	0.295	1.038283
1999-05-31	0.015161	57.067742	1.860645	20.737097	4.684516	0.6	2.897103
...	...	...	...	...	...	...	...
2022-05-31	0.020323	55.13871	1.276452	16.889355	4.222903	0.66	2.812945
2022-06-30	0.020323	64.903333	1.983667	25.627	5.468667	0.943	5.190413
2022-07-31	0	77.841935	1.204667	46.958387	6.912581	0.893548	6.152948
2022-08-31	0.007097	77.83871	1.204667	44.480323	6.080323	0.8	4.864258
2022-09-30	0.003333	67.663333	3.195333	33.291333	4.357333	0.24	1.345867

285 rows × 7 columns							

Figure 22
Correlation among variables for NASA data

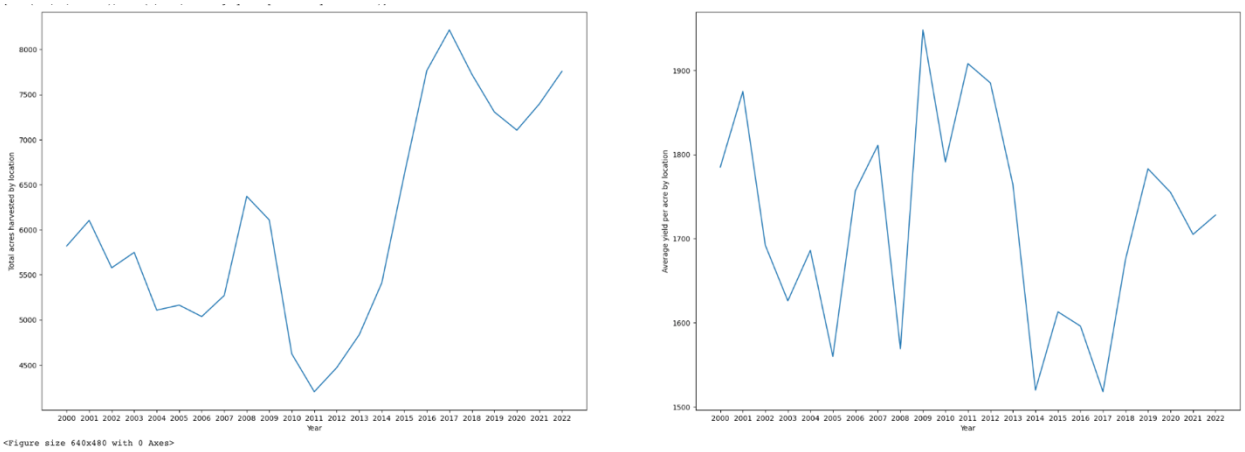


PART 3 - Oregon

In this part Oregon state data is analyzed for hops production.

Figure 23

Comparison of total acres harvested and average yield per acre - Oregon



Oregon harvested around 5700 acres each year in 2000, with surges in 2001, 2003, and 2008, before a massive drop from 6500 acres in 2008 to over 1000 acres in 2011. After 2011, there was an incredible improvement until 2017, which was the apex of the span 2000 to 2022, with around 9000 acres harvested every year. Over the years, the average yield per acre has been quite unpredictable. It began at 1800 in the year 2000 and has not been constant since then, with a crest and dip detected in the pattern every other year. Peaks can be found in 2001, 2004, 2007, 2009, 2011, 2015, and 2020. Every other year, there was a fall.

Figure 24

Comparing 5 varieties of Hops - OREGON

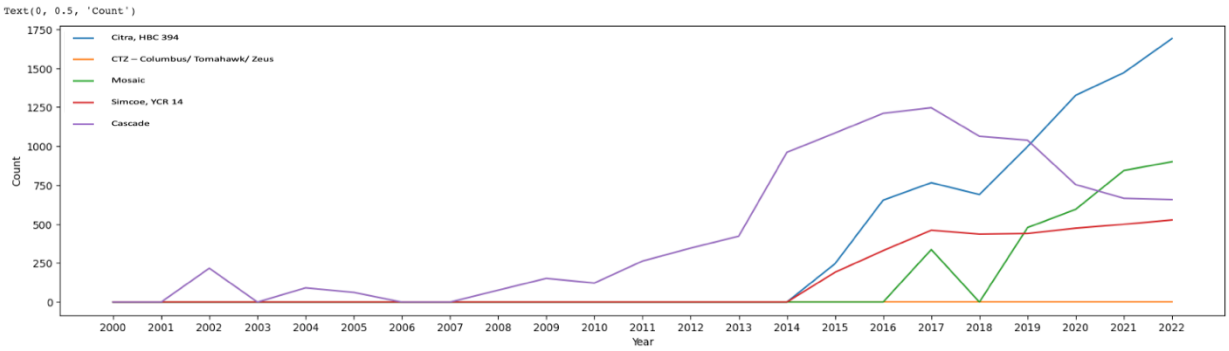


Figure 24 shows a comparison of the five selected varieties of hops. The total acres harvested for Citra are seen to be increasing over the last 8 years in all three states. CTZ has never been harvested in Oregon. Cascade used to outperform all the varieties until 2017 when it stood at an astonishing 1250 acres annually.

Figure 25

Climate data converted to Date time index - Oregon

	Date	ppt (inches)	tmin (degrees F)	tmean (degrees F)	tmax (degrees F)	vpdmin (hPa)	vpdmax (hPa)	ETO (FAO)	ETO\n (Hargreaves)	Hops factor (HF)	ETO x HF
0	1999-01-01	0.1	32.2	38	43.8	0.03	2.25	0.72	0.5	0	0
1	1999-01-02	0	25.2	34.2	43.3	0.11	3.81	0.8	0.56	0	0
2	1999-01-03	0	23.7	32.6	41.6	0.11	2.95	0.77	0.54	0	0
3	1999-01-04	0	21.3	31	40.6	0.06	3.35	0.78	0.53	0	0
4	1999-01-05	0	22	31	40	0	2.54	0.75	0.52	0	0
...	...	...	...	...	...	...	...	...	...	...	...
8669	2022-09-26	0	44.5	64	83.4	2.09	31.21	4.02	3.74	0	0
8670	2022-09-27	0	44.4	64.8	85.2	2.35	33.79	4.15	3.83	0	0
8671	2022-09-28	0	46.9	67.4	88	2.65	36.14	4.27	3.96	0	0
8672	2022-09-29	0	52.6	71	89.3	5.41	40.63	4.15	3.88	0	0
8673	2022-09-30	0	51	62.7	74.5	2.92	21.51	2.87	2.72	0	0

After storing the data in a data frame, it is arranged to get a clearer picture of the data.

Figure 26

Quantitative Analysis of Climate Data - OREGON

	ppt (inches)	tmin (degrees F)	tmean (degrees F)	tmax (degrees F)	vpdmin (hPa)	vpdmax (hPa)	ETO (FAO)	ETO\n (Hargreaves)	Hops factor (HF)	ETO x HF
count	8674	8674	8674	8674	8674	8674	8674	8674	8674	8674
mean	0.024097	39.000288	52.097129	65.198075	1.638862	18.920501	3.29139	3.200372	0.332442	1.859327
std	0.07206	13.698944	16.836501	20.594404	1.772634	15.674229	2.198351	2.368018	0.391249	2.402385
min	0	-17.8	-6.9	4	0	0.09	0.29	-0.13	0	0
1%	0	8.2	16.773	24.3	0	0.48	0.47	0.22	0	0
2%	0	12.8	21.5	28.5	0.01	0.7246	0.51	0.28	0	0
3%	0	15.119	23.819	30.619	0.02	0.98	0.54	0.32	0	0
5%	0	18	26.765	33.6	0.04	1.36	0.6	0.37	0	0
10%	0	22	30.9	38.3	0.09	2.43	0.72	0.49	0	0
25%	0	28.7	38.5	48.025	0.28	5.68	1.18	0.93	0	0
50%	0	38.2	51.2	64.5	0.95	13.91	2.9	2.73	0	0
75%	0	50.5	66.6	83.3	2.49	30.2575	5.1975	5.25	0.8	4.062
95%	0.15	60.635	78.435	96.835	5.39	49.047	7.0435	7.29	0.99	6.42049
97%	0.21	62.1	79.981	98.781	6.1381	52.1881	7.28	7.56	1	6.86
98%	0.26	63.3	81	99.8	6.6454	54.2354	7.4754	7.75	1	7.1
99%	0.35	64.827	82.4	101.3	7.42	57.1508	7.65	7.97	1	7.413441
max	1.21	73.3	89.5	109.1	10.71	74.92	9.03	9.38	1	8.8494

Figure 27

Outlier Detection - OREGON

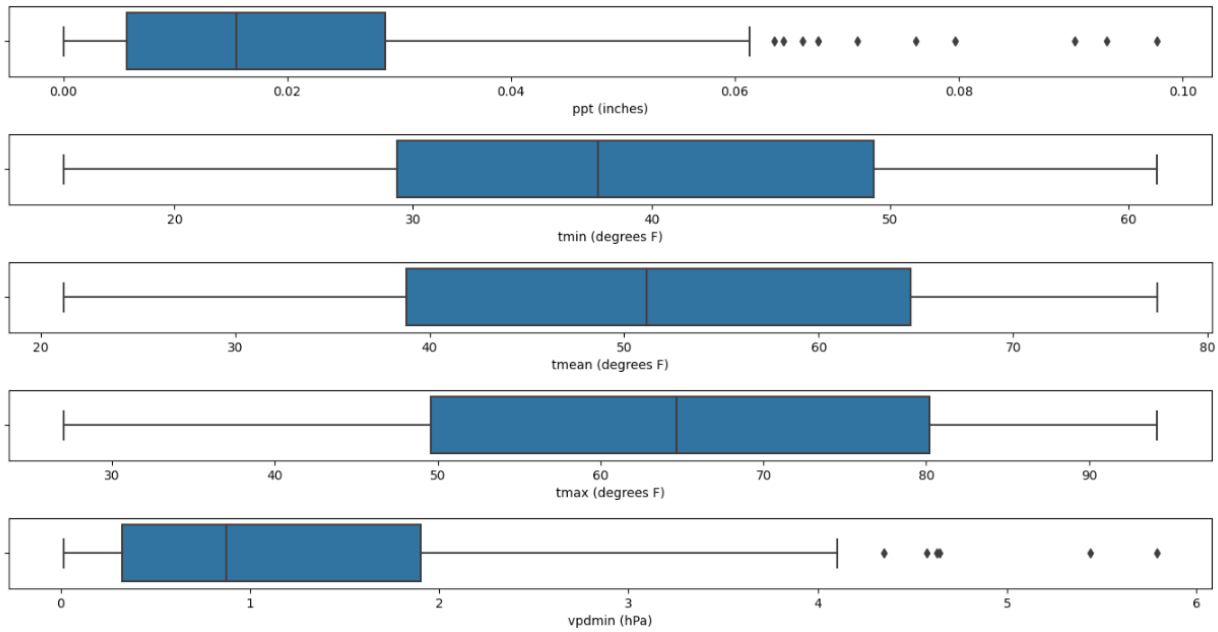


Figure 28

Climate data arranged into monthly data - OREGON

	ppt (inches)	tmin (degrees F)	tmean (degrees F)	tmax (degrees F)	vpdmin (hPa)	vpdmax (hPa)	ETO (FAO)	ETO\n (Hargreaves)	Hops factor (HF)	ETO x HF
Date										
1999-01-31	0.044839	28.03871	35.529032	43.022581	0.225484	3.354194	0.824839	0.597097	0	0
1999-02-28	0.068571	28.407143	36.296429	44.182143	0.297143	4.160357	1.134286	0.925	0	0
1999-03-31	0.02129	31.003226	43.758065	56.5	0.636452	10.653226	2.317419	2.123871	0	0
1999-04-30	0.016333	34.273333	48.176667	62.073333	1.244333	14.159333	3.372	3.264333	0.295	1.038283
1999-05-31	0.015161	42.470968	57.067742	71.664516	1.860645	20.737097	4.684516	4.73871	0.6	2.897103
...	...	...	...	...	...	...	...	...	...	...
2022-05-31	0.074516	42.264516	55.13871	68.003226	1.276452	16.889355	4.222903	4.260645	0.66	2.812945
2022-06-30	0.068333	50.536667	64.903333	79.256667	1.983667	25.627	5.468667	5.642333	0.943	5.190413
2022-07-31	0	60.358065	77.841935	95.351613	5.757742	46.958387	6.912581	7.188065	0.893548	6.152948
2022-08-31	0.007097	60.603226	77.83871	95.070968	4.679032	44.480323	6.080323	6.189032	0.8	4.864258
2022-09-30	0.003333	50.49	67.663333	84.846667	3.195333	33.291333	4.357333	4.208	0.24	1.345867

Figure 29

Outliers removed from climate data - OREGON

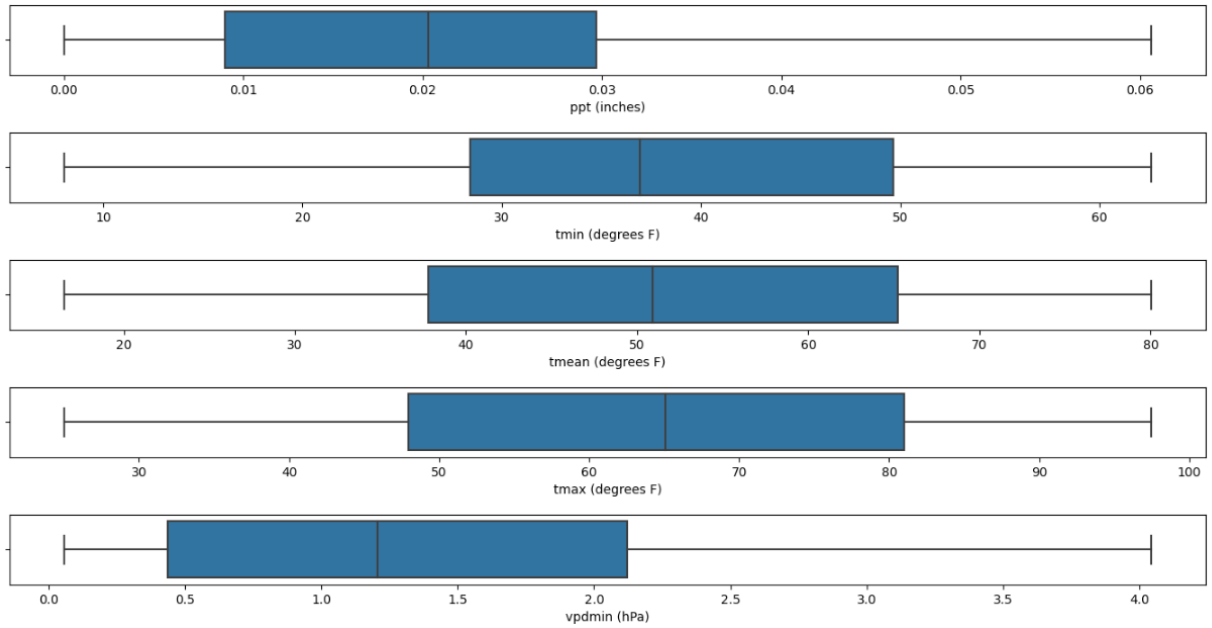


Figure 30

Correlation plot for Oregon Climate data

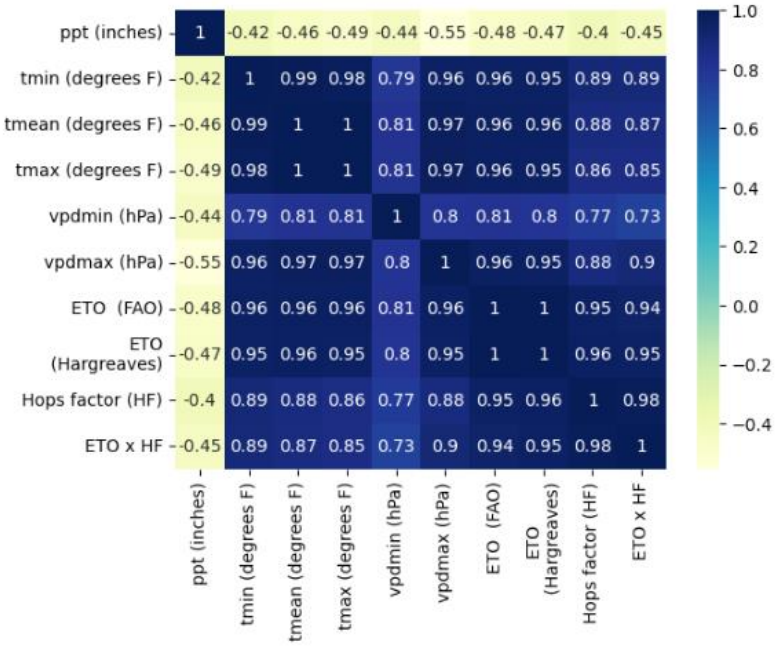


Figure 31

NASA Data - OREGON

	PARAMETER	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	ANN
11	GWETTOP	1999	0.98	0.98	0.97	0.92	0.88	0.8	0.66	0.56	0.5	0.51	0.7	0.88	0.77
12	GWETTOP	2000	0.93	0.96	0.95	0.91	0.88	0.8	0.64	0.52	0.51	0.54	0.61	0.73	0.75
13	GWETTOP	2001	0.8	0.82	0.83	0.84	0.8	0.69	0.56	0.51	0.48	0.51	0.68	0.89	0.7
14	GWETTOP	2002	0.95	0.96	0.95	0.92	0.86	0.76	0.62	0.51	0.45	0.49	0.57	0.74	0.73
15	GWETTOP	2003	0.89	0.94	0.95	0.95	0.9	0.77	0.59	0.49	0.47	0.52	0.6	0.84	0.74
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
144	CLRSKY_SFC_PAR_TOT	2017	41.33	63.45	92.21	124.05	149.46	159.24	155.45	126.54	94.52	71.82	45.45	33.95	96.62
145	CLRSKY_SFC_PAR_TOT	2018	40.23	62.07	92.52	121.13	147.41	158.65	154.62	123.53	103.88	71.77	45.48	34.08	96.44
146	CLRSKY_SFC_PAR_TOT	2019	41.02	62.49	91.9	123.98	146.98	160.23	151.95	132.78	103.05	71.46	45.59	33.27	97.22
147	CLRSKY_SFC_PAR_TOT	2020	39.64	61.94	92.3	125.34	147.52	158.56	153.12	134.2	90.76	69.65	45.09	33.88	96.09
148	CLRSKY_SFC_PAR_TOT	2021	40.2	62.19	93.13	125.48	149.06	159.71	153.43	124.38	103.31	71.6	45.66	34.57	97.05

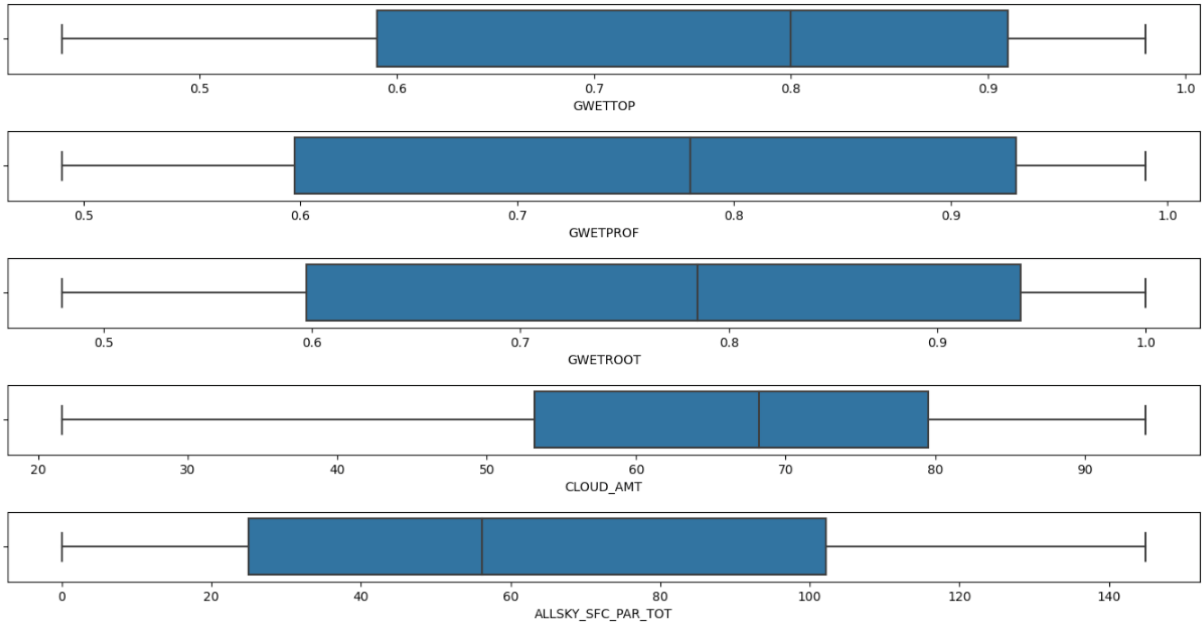
Figure 32

Quantitative Analysis of NASA Data - OREGON

	GWETTOP	GWETPROF	GWETROOT	CLOUD_AMT	ALLSKY_SFC_PAR_TOT	CLRSKY_SFC_PAR_TOT
count	276	276	276	276	276	276
mean	0.759565	0.762065	0.766304	65.342355	63.054275	88.483333
std	0.171294	0.166125	0.174066	17.261853	41.557208	50.155471
min	0.43	0.49	0.48	21.57	0	0
1%	0.44	0.5	0.49	26.5925	0	0
2%	0.45	0.5	0.49	30.07	0	0
3%	0.4725	0.51	0.4925	31.885	0	0
5%	0.4875	0.51	0.5	35.0875	0	0
10%	0.51	0.53	0.52	37.945	15.795	33.915
25%	0.59	0.5975	0.5975	53.1825	24.915	45.195
50%	0.8	0.78	0.785	68.2	56.22	92.265
75%	0.91	0.93	0.94	79.525	102.15	134.165
95%	0.96	0.98	0.98	89.365	128.95	158.3225
97%	0.97	0.98	0.99	89.925	132.165	158.765
98%	0.97	0.98	0.99	90.875	135.51	159.245
99%	0.98	0.98	0.99	92.0025	138.28	159.755
max	0.98	0.99	1	94.1	144.94	160.23

Figure 33

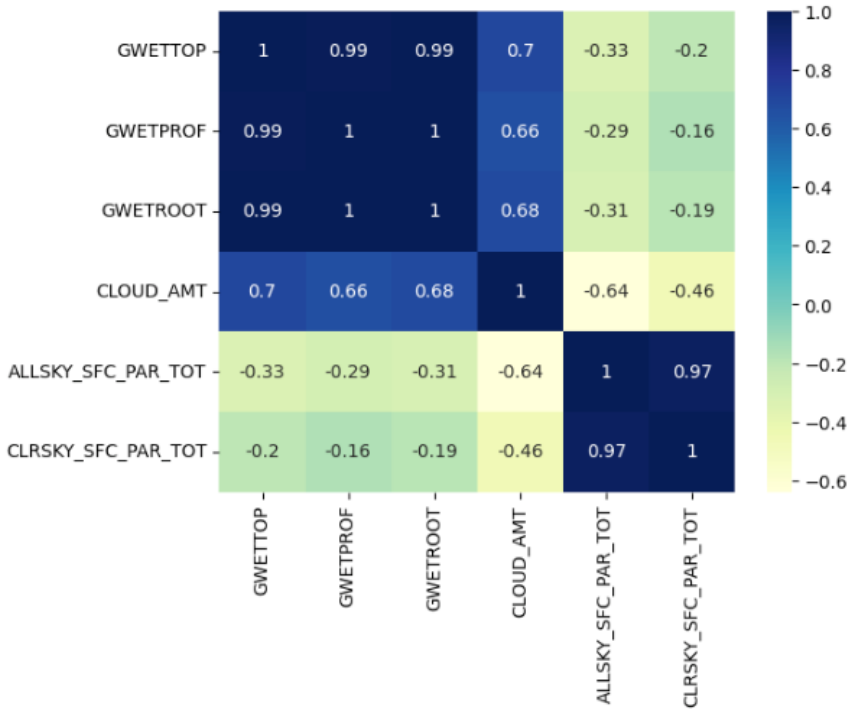
Outlier detection by boxplot in NASA Data - OREGON



It can be seen in figure 33 that there are no outliers in the NASA data for Oregon state so we can move forward with the Correlation plot.

Figure 34

NASA Data Correlations – OREGON



As it can be seen that the clear sky variable has a negative correlation, we decided to remove this column and the NASA satellite data was ready to be modelled.

Figure 35

NASA Data cleaned – OREGON

	GWETTOP	GWETPROF	GWETROOT	CLOUD_AMT	ALLSKY_SFC_PAR_TOT
Date					
1999-01-01	0.5	0.48	0.48	60.08	0
1999-02-01	0.56	0.52	0.53	76.26	0
1999-03-01	0.54	0.52	0.52	63.34	0
1999-04-01	0.48	0.48	0.48	53.57	0
1999-05-01	0.42	0.44	0.44	49.3	0
...	...	...	...	...	...
2021-08-01	0.23	0.36	0.36	30.74	111.63
2021-09-01	0.24	0.36	0.36	26.34	95.3
2021-10-01	0.35	0.38	0.38	61.68	58.05
2021-11-01	0.41	0.41	0.42	68.05	38.55
2021-12-01	0.45	0.44	0.45	82.59	21.95

Modelling

Business Question

Forecasting is done for each variety and calendar year since it is done at the variety across year levels. In order to extract a higher yield, we can now plan the variety or the amount of the variety to be harvested.

How can we utilize our forecasting model for hop production, which predicts the average yield per acre for different hop varieties in the year 2023 based on historical data, to strategically plan and optimize hop production in order to maximize profitability?

We will be able to forecast year and variety with our model. This can assist managers in prioritizing which types to grow each year based on the climate. Our hop production forecasting model offers a useful resource for hop farmers to strategically plan and maximise their output by reliably projecting the average yield per acre for several hop varieties in the year 2023 based on historical data. Hop growers can improve their operational effectiveness, obtain a competitive edge, and achieve sustainable growth in the hop growing sector by utilizing the forecasting model.

Washington Climate analysis

The data processing code provided aims to generate a comprehensive data frame, named `new_df_average_yield_WA`, by consolidating data from various sources. The code begins by initializing two empty lists, `list_avg` and `list_acr`, which will be used to store dictionaries and values for the "Average Yield" and "Total Acre" columns, respectively. It then proceeds to extract the values from a DataFrame called `df_year` and prepares them for iteration by flattening the two-dimensional array into a single list. The subsequent

nested loops iterate over the years from 2000 to 2022 and each variety in the flattened list. Within these loops, the code appends dictionaries to `list_avg`, representing each row of data. These dictionaries include the "Location" set to "Washington", the "Year" corresponding to the current iteration value, the "Variety" being iterated, and the "Average Yield" obtained from the `df_average_yield_WA` DataFrame using the matching year and variety. Additionally, the code appends the corresponding total acre values, extracted from the `df_total_acre_WA` DataFrame, to the `list_acr`. Once the loops conclude, the `new_df_average_yield_WA` DataFrame is created by converting the list of dictionaries into a DataFrame structure. The "Total Acre" column in the resulting DataFrame is then populated with the values from `list_acr`, ensuring each row has the respective total acre value. This processing allows for a comprehensive overview of the average yield and total acre for each year and variety, aiding in analyzing the crop performance in Washington over the specified period.

Table - 1

	Location	Year	Variety	Average Yeild	Total Acre
0	Washington	2000	Ahtanum, YCR 1	NaN	NaN
1	Washington	2000	Amarillo, VGXP01	NaN	NaN
2	Washington	2000	Apollo	NaN	NaN
3	Washington	2000	Azacca, ADHA-483	NaN	NaN

4	Washington	2000	Bravo	NaN	NaN
...	...	...	...	...	...
1352	Washington	2022	Vanguard	NaN	NaN
1353	Washington	2022	Warrior, YCR 5	1610	147
1354	Washington	2022	Willamette	991	124
1355	Washington	2022	Zappa	839	69
1356	Washington	2022	Total	1679	42762

```
# First we will check if there are any null values or not

df_total_acre_WA.isnull().sum()

#replacing the values of of nulls and checking it again

df_total_acre_WA.fillna(0,inplace=True)

df_total_acre_WA.isnull().sum()
```

As we have seen in the EDA there were no production for several crop varieties, therefore the all Nan values are replaced with 0

Table -2

	Location	Year	Variety	Average Yeild	Total Acre
0	Washington	2000	Ahtanum, YCR 1	0	0
1	Washington	2000	Amarillo, VGXP01	0	0
2	Washington	2000	Apollo	0	0

3	Washington	2000	Azacca, ADHA-483	0	0
4	Washington	2000	Bravo	0	0
...	...	...	...	...	...
1352	Washington	2022	Vanguard	0	0
1353	Washington	2022	Warrior, YCR 5	1610	147
1354	Washington	2022	Willamette	991	124
1355	Washington	2022	Zappa	839	69
1356	Washington	2022	Total	1679	42762

Modeling of Climate Data (WA)

Table - 3

	Variety	Total Acre	Hops factor (HF)	Year	Average Yeild	ppt_sign	vdpmean_sign	tmean_sign	ETO x HF	Moving AVg Yeild
0	Ahtanum, YCR 1	0	0.328616	2000	0	0.002748	3.849864	21.583557	1.689745	0
1	Amarillo,	0	0.328616	2000	0	0.002748	3.849864	21.583557	1.689745	0

	VGXP01									
2	Apollo	0	0.328616	2000	0	0.002748	3.849864	21.583557	1.689745	0
3	Azacca, ADHA-483	0	0.328616	2000	0	0.002748	3.849864	21.583557	1.689745	0
4	Bravo	0	0.328616	2000	0	0.002748	3.849864	21.583557	1.689745	0
...	...	...	...	...	...	...	...	...	...	...
1352	Triumph	0	0.329515	2022	0	0.007831	5.420572	23.308839	2.003723	0
1353	Vanguard	0	0.329515	2022	0	0.007831	5.420572	23.308839	2.003723	0
1354	Warrior, YCR5	147	0.329515	2022	1610	0.007831	5.420572	23.308839	2.003723	1839.666667
1355	Willamette	124	0.329515	2022	991	0.007831	5.420572	23.308839	2.003723	1216.33333
1356	Zappa	69	0.329515	2022	839	0.007831	5.420572	23.308839	2.003723	279.666667

The Linear Regression model exhibited comparatively lower performance in predicting the average yield based on the available features. It achieved an R-squared value of 0.805 on the training set and 0.823 on the test set, indicating that the linear relationship assumption might not capture all the variations in the data accurately. The Root Mean Square Error (RMSE) for the training set was 104.13, indicating a lower average deviation of the predicted values from the actual values compared to other models. Similarly, the Mean Absolute Error (MAE) for the training set was 36.98, showing a relatively lower average magnitude of errors.

Analyzing the NASA Data of Washington

imports and processes climate data for Washington state from the NASA dataset, resulting in a DataFrame comprising climate parameters measured monthly from 1999 to 2021. This dataset provides valuable insights into the spatiotemporal variations of climate factors in Washington state and can serve as a foundation for further scientific analysis and research in the field of climatology.

Table - 5

10	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	ANN
PARAMETER														
GWETTOP	1999	0.5	0.53	0.51	0.48	0.45	0.39	0.35	0.34	0.34	0.36	0.42	0.45	0.43
GWETTOP	2000	0.48	0.49	0.48	0.45	0.42	0.38	0.34	0.31	0.34	0.36	0.38	0.4	0.41
GWETTOP	2001	0.42	0.43	0.42	0.41	0.37	0.35	0.31	0.29	0.28	0.33	0.4	0.45	0.38
GWETTOP	2002	0.45	0.45	0.44	0.41	0.38	0.35	0.3	0.27	0.27	0.32	0.35	0.41	0.37
GWETTOP	2003	0.47	0.46	0.45	0.45	0.4	0.34	0.3	0.3	0.29	0.32	0.35	0.41	0.38
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
CLRSKY_SFC_PAR_TOT	2017	40.04	63.61	93.04	124.09	148.32	159.21	155.54	123.09	91.47	69.73	43.62	32.15	95.48
CLRSKY_SFC_PAR_TOT	2018	38.48	60.84	91.72	122.19	146.95	157.95	153.64	119.37	102.61	69.39	43.24	32.55	95.06
CLRSKY_SFC_PAR_TOT	2019	39.41	62.67	91.91	123.11	146.38	159.72	150.63	131.52	101.03	68.95	43.23	31.65	96
CLRSKY_SFC_PAR_TOT	2020	38.05	60.7	91.41	124.59	147.98	158.38	152.6	133.31	86.42	68.12	42.98	32.16	94.82
CLRSKY_SFC_PAR_TOT	2021	38.66	60.98	91.93	124.5	148.72	159.84	152.77	121.02	101.07	69.31	43.3	32.47	95.53

the dataset is split into training and test sets for regression model evaluation. The dataset contains climate data obtained from the NASA dataset for a specific timeframe. The training

set is used to train various regression models, and their performance is assessed on the test set. The output provides evaluation metrics such as R-squared, Root Mean Square Error (RMSE), and Mean Absolute Error (MAE) for each model.

Modeling climate data (Oregon)

Table 6

Climate Data (Oregon)

	Variety	Total Acre	Hops factor (HF)	Year	Average Yeild	ppt_sign	vdpmean_sign	tmean_sign	ETO x HF	Moving AVg Yeild
0	Ahtanum, YCR 1	0	0.328616	2000	0	0.003553	5.18965	22.35777	1.869257	0
1	Amarillo, VGXP01	0	0.328616	2000	0	0.003553	5.18965	22.35777	1.869257	0
2	Apollo	0	0.328616	2000	0	0.003553	5.18965	22.35777	1.869257	0

3	Azacca, ADHA-483	0	0.328616	2000	0	0.003553	5.18965	22.35777	1.869257	0
4	Bravo	0	0.328616	2000	0	0.003553	5.18965	22.35777	1.869257	0
...	...	...	...	...	...	...	...	...	...	...
1352	Triumph	0	0.329515	2022	0	0.011458	5.551085	23.983815	2.004259	0
1353	Vanguard	0	0.329515	2022	0	0.011458	5.551085	23.983815	2.004259	0
1354	Warrior, YCR5	147	0.329515	2022	1610	0.011458	5.551085	23.983815	2.004259	1839.666667
1355	Willamette	124	0.329515	2022	991	0.011458	5.551085	23.983815	2.004259	1216.33333
1356	Zappa	69	0.329515	2022	839	0.011458	5.551085	23.983815	2.004259	279.666667

The Linear Regression model, trained using the provided code, resulted in an R-squared value of 0.841 on the test set and 0.803 on the training set. This indicates that the model explains approximately 84.1% of the variance in the test data and 80.3% in the training

data. The R-squared value represents the goodness of fit, with higher values indicating a better fit between the predicted and actual values.

Analyzing the NASA Data of OR

Table 7

NASA Data for Oregon

10 YEAR		JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	ANN
PARAMETER														
GWETTOP	1999	0.98	0.98	0.97	0.92	0.88	0.8	0.66	0.56	0.5	0.51	0.7	0.88	0.77
GWETTOP	2000	0.93	0.96	0.95	0.91	0.88	0.8	0.64	0.52	0.51	0.54	0.61	0.73	0.75
GWETTOP	2001	0.8	0.82	0.83	0.84	0.8	0.69	0.56	0.51	0.48	0.51	0.68	0.89	0.7
GWETTOP	2002	0.95	0.96	0.95	0.92	0.86	0.76	0.62	0.51	0.45	0.49	0.57	0.74	0.73
GWETTOP	2003	0.89	0.94	0.95	0.95	0.9	0.77	0.59	0.49	0.47	0.52	0.6	0.84	0.74
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
CLRSKY_SF	2017	41.33	63.45	92.21	124.05	149.46	159.24	155.45	126.54	94.52	71.82	45.45	33.95	96.62
CLRSKY_SF	2018	40.23	62.07	92.52	121.13	147.41	158.65	154.62	123.53	103.88	71.77	45.48	34.08	96.44
CLRSKY_SF	2019	41.02	62.49	91.9	123.98	146.98	160.23	151.95	132.78	103.05	71.46	45.59	33.27	97.22
CLRSKY_SF	2020	39.64	61.94	92.3	125.34	147.52	158.56	153.12	134.2	90.76	69.65	45.09	33.88	96.09
CLRSKY_SF	2021	40.2	62.19	93.13	125.48	149.06	159.71	153.43	124.38	103.31	71.6	45.66	34.57	97.05

MergeData_df_NASA_OR

Table 8

Merging Data for Oregon

	Variety	Total Acre	ds	y	GWETTOP	GWETPROF	ALLSKY_SFC_PAR_TOT	CLRSKY_SFC_PAR_TOT	GWETPROF	CLOUD_AMT	Moving AVg Yeild
0	Ahtanum, YCR 1	0	2000	0	0.93	0.95	0	0	0.95	71.06	0
1	Ahtanum, YCR 1	0	2000	0	0.96	0.97	0	0	0.97	73.17	0
2	Ahtanum, YCR 1	0	2000	0	0.95	0.97	0	0	0.97	72.84	0
3	Ahtanum, YCR 1	0	2000	0	0.91	0.91	0	0	0.91	73.61	0
4	Ahtanum, YCR 1	0	2000	0	0.88	0.87	0	0	0.87	80.92	0
...	...	...	...	...	...	...	...	...	...	...	...
15571	Zappa	0	2021	0	0.44	0.51	100.66	124.38	0.51	46.62	0
15572	Zappa	0	2021	0	0.43	0.51	82.77	103.31	0.51	50.65	0
15573	Zappa	0	2021	0	0.59	0.59	45.29	71.6	0.59	73.99	0
15574	Zappa	0	2021	0	0.8	0.77	23.45	45.66	0.77	87.76	0
15575	Zappa	0	2021	0	0.9	0.9	16.54	34.57	0.9	90.77	0
15576 rows × 11 columns											

The evaluation of the Linear Regression model using the provided code reveals the following metrics:

- R-squared (test set): The R-squared value is approximately 0.844, indicating that the model explains approximately 84.4% of the variance in the test data. A higher R-squared value suggests a better fit between the predicted and actual values.
- R-squared (training set): The R-squared value is approximately 0.793, indicating that the model explains approximately 79.3% of the variance in the training data. A higher R-squared value suggests a better fit between the predicted and actual values in the training set.

These metrics help assess the performance of the Linear Regression model in predicting the target variable, with higher R-squared values indicating a better fit and a higher percentage of explained variance.

Conclusion

The Linear Regression model, although demonstrating moderate performance in predicting average yield based on the available features, provides valuable insights for variety selection and harvest planning in Washington state. While there may be room for improvement, the model explains a significant portion of the variance in the test and training data. Further analysis and exploration of alternative models can be conducted to enhance the accuracy of yield predictions and provide more robust insights for optimizing variety selection and maximizing yield based on climatic conditions. These insights can aid businesses in making informed decisions and strategically planning their agricultural activities, ultimately improving operational efficiency and maximizing overall yield.

Module 7

Introduction

Some modifications were made in this section of the project based on feedback from prior submissions. The input shed light on certain flaws and other ways that may be tested in the project, resulting in numerous new code chunks and data size changes to better match the model. As the experiment proceeds, the results will improve and there will be additional alternatives for future research.

Business Objective

The focus of the project is to construct a model that would predict future production for a particular duration of time i.e. one year or six months. With the help of our algorithm, we will be able to predict year and variety. This can help managers decide which varieties to plant in a particular year dependent on the climate. If our model doesn't run into any more problems, like wildfires or other problems, it can deliver the accuracy that is claimed. This model is to be used to estimate the average hop yield.

Analysis and Modeling

Washington

First of all some basic analysis is done on Washington Dataset.

The selected data frame, MergeData_df_WA, contains information about hop varieties in Washington, their total acreage, hops factor, year, average yield, climate factors (ppt_sign, vdpmean_sign, tmean_sign), ETO x HF, and moving average yield.

After filtering out certain rows, there are 529 rows remaining in the data frame for analysis. The dropped values are as follows:

- 805 rows were dropped due to the 'Total Acre' column having a value of 0.

- 23 rows were dropped because the 'Variety' column had the value 'Total'.

Table 9 : Selected modeling data frame for Washington

	Variety	Total Acre	Hops facto	Year	Average Ye	ppt_sign	vdpmean_	tmean_sig	ETO x HF	Moving AV
5	CTZ - Colu	6588	0.328959	2000	2605	0.002748	3.849864	21.58356	1.689745	0
7	Cascade	996	0.328959	2000	1806	0.002748	3.849864	21.58356	1.689745	0
11	Chinook	670	0.328959	2000	1957	0.002748	3.849864	21.58356	1.689745	0
13	Cluster	939	0.328959	2000	1997	0.002748	3.849864	21.58356	1.689745	0
20	Galena	5044	0.328959	2000	1891	0.002748	3.849864	21.58356	1.689745	0
...	...	...	...	...	...	...	...	...	...	...
1347	Tahoma	383	0.32931	2022	1310	0.007831	5.420572	23.30884	2.003723	1317.667
1348	Talus, HBC	377	0.32931	2022	1703	0.007831	5.420572	23.30884	2.003723	567.6667
1354	Warrior, Y	147	0.32931	2022	1610	0.007831	5.420572	23.30884	2.003723	1839.667
1355	Willamette	124	0.32931	2022	991	0.007831	5.420572	23.30884	2.003723	1216.333
1356	Zappa	69	0.32931	2022	839	0.007831	5.420572	23.30884	2.003723	279.6667

The remaining data provides insights into the relationship between hop varieties and climate factors in Washington. Each row represents a specific hop variety, with information such as total acreage, hops factor, average yield, and climate factors for each year.

For example, the first row shows information for the hop variety CTZ - Columbus/Tomahawk/Zeus in the year 2000. It had a total acreage of 6588, a hops factor of 0.328959, an average yield of 2605, and climate factors (ppt_sign, vdpmean_sign, tmean_sign) of 0.002748, 3.849864, and 21.583557 respectively. Additionally, it had an ETO x HF value of 1.689745 and a moving average yield of 0.000000.

This data frame can be further analyzed to explore relationships between hop varieties, climate factors, and their impact on yield in Washington.

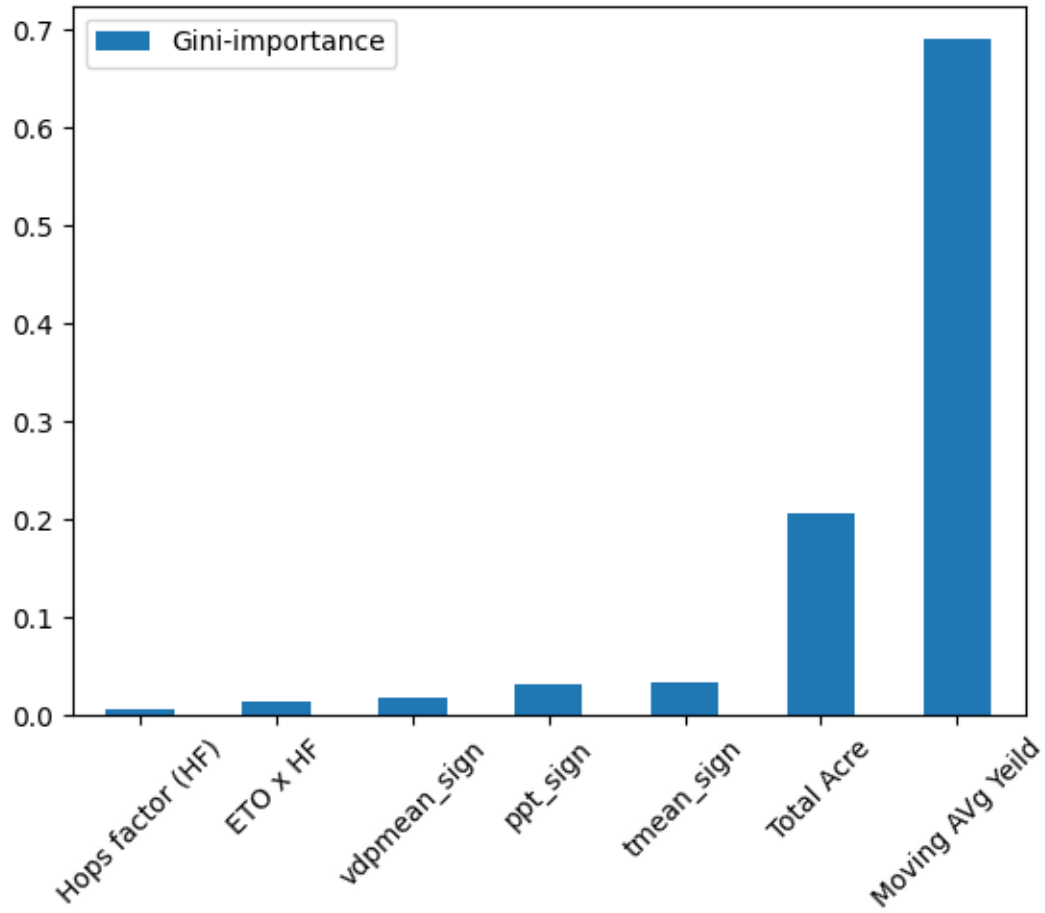
Implementation a random forest regression model to predict the average yield of hop varieties in Washington. The dataset, obtained from the filtered data frame MergeData_df_WA, is divided into training and testing sets based on the unique years available. Approximately 70% of the years are assigned for training, while the remaining 30% are allocated for testing.

The target variable 'Average Yield' is renamed as 'y', and the independent variables are selected by excluding the 'y', 'ds' (Year), and 'Variety' columns. The training and testing datasets are prepared accordingly.

A random forest regressor (rfr) is initialized, trained using the training data, and used to predict the average yields for the testing data. The performance of the model is evaluated using the R-squared coefficient, which measures the proportion of the variance in the target variable that can be explained by the model. The resulting R-squared values for the test and train sets are reported as output.

Additionally, the code calculates the feature importance using the random forest regressor. The importance values are stored in a dictionary and then visualized in a bar chart, sorted in ascending order.

Figure: 35



The obtained results reveal that the model explains approximately 56% of the variance in the average yield for the test data and 94% for the training data, as indicated by the R-squared values.

Two evaluation metrics, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), are calculated for both the test and train sets.

For the test set:

The RMSE value is approximately 346.95, indicating the average prediction error in the average yield of hop varieties in Washington.

The MAE value is approximately 252.71, representing the average absolute prediction error.

These metrics provide insights into the accuracy of the model's predictions on unseen test data.

For the train set:

The RMSE value is approximately 134.54, which indicates the average prediction error in the average yield for the training data.

The MAE value is approximately 89.69, representing the average absolute prediction error.

Comparing the train and test statistics, it appears that the model performs better on the training data than on the unseen test data.

linear regression model

A linear regression model (lr) is used to predict the average yield of hop varieties in Washington. The model is trained on the training data (x_train and y_train) and used to predict the average yields for the testing data (x_test). The performance of the model is evaluated using the R-squared coefficient.

The output shows the R-squared values for the test and train sets: Test 0.3831203273264874 and Train 0.4280614064630994. These values indicate that the linear regression model explains approximately 38.31% of the variance in the average yield for the test data and 42.81% for the training data.

The columns used as independent variables in the linear regression model are obtained from the training and testing datasets. The output displays the column names for both the

training and testing datasets, which include 'Total Acre', 'Hops factor (HF)', 'ppt_sign', 'vdpmean_sign', 'tmean_sign', 'ETO x HF', and 'Moving AVg Yeild'.

Furthermore, the coefficients of the linear regression model are extracted using the lr.coef_ attribute. The coefficients represent the effect of each independent variable on the predicted average yield. The output displays the coefficient values for the respective independent variables.

Index(['Total Acre' : 3.95666820e-02,
'Hops factor (HF)': 3.54551057e+05,
'ppt_sign': -5.51794855e+04,
'vdpmean_sign': 8.08540206e+02,
'tmean_sign': 3.03094707e+02,
'ETO x HF': -7.54026301e+03,
'Moving AVg Yeild': 4.97518745e-01]

These coefficients represent the estimated effect of each independent variable on the predicted average yield of hop varieties in Washington. They indicate the direction and magnitude of the relationship between each independent variable and the average yield. Positive coefficients suggest a positive relationship, meaning an increase in the independent variable is associated with an increase in average yield. Negative coefficients indicate a negative relationship, where an increase in the independent variable corresponds to a decrease in average yield.

For the test set:

The RMSE value is approximately 443.93, indicating the average prediction error in the average yield of hop varieties in Washington.

The MAE value is approximately 352.61, representing the average absolute prediction error.

These metrics provide insights into the accuracy of the linear regression model's predictions on unseen test data.

For the train set:

The RMSE value is approximately 134.54, which indicates the average prediction error in the average yield for the training data.

The MAE value is approximately 89.69, representing the average absolute prediction error.

Comparing the train and test statistics, it appears that the linear regression model performs better on the training data than on the unseen test data.

Analyzing the NASA Climate Data of Washington

Data Preparation:

The NASA climate data of Washington is prepared for analysis by reading it from an Excel file and cleaning it. The data is processed to remove headers and unnecessary rows. Missing values are filled, and the appropriate data types are assigned. Descriptive statistics are calculated to gain insights into the cleaned data. Outliers are detected and treated using the interquartile range (IQR) method. The cleaned data is then merged with the average yield data, resulting in the creation of the merged DataFrame.

Replacement of the outliers (same formula for 3 of the given states)

The outlier replacement in the `df_nasa_climate_ state_name _cleaned` DataFrame using the interquartile range (IQR) method. It calculates the first quartile (`q1`) and third quartile (`q2`) for each column, along with the IQR, upper limit, and lower limit.

```
lower_limit = Q1 - (1.5 * IQR)
upper_limit = Q3 + (1.5 * IQR)

if value < lower_limit or value > upper_limit:
   replace value with lower_limit or upper_limit respectively
```

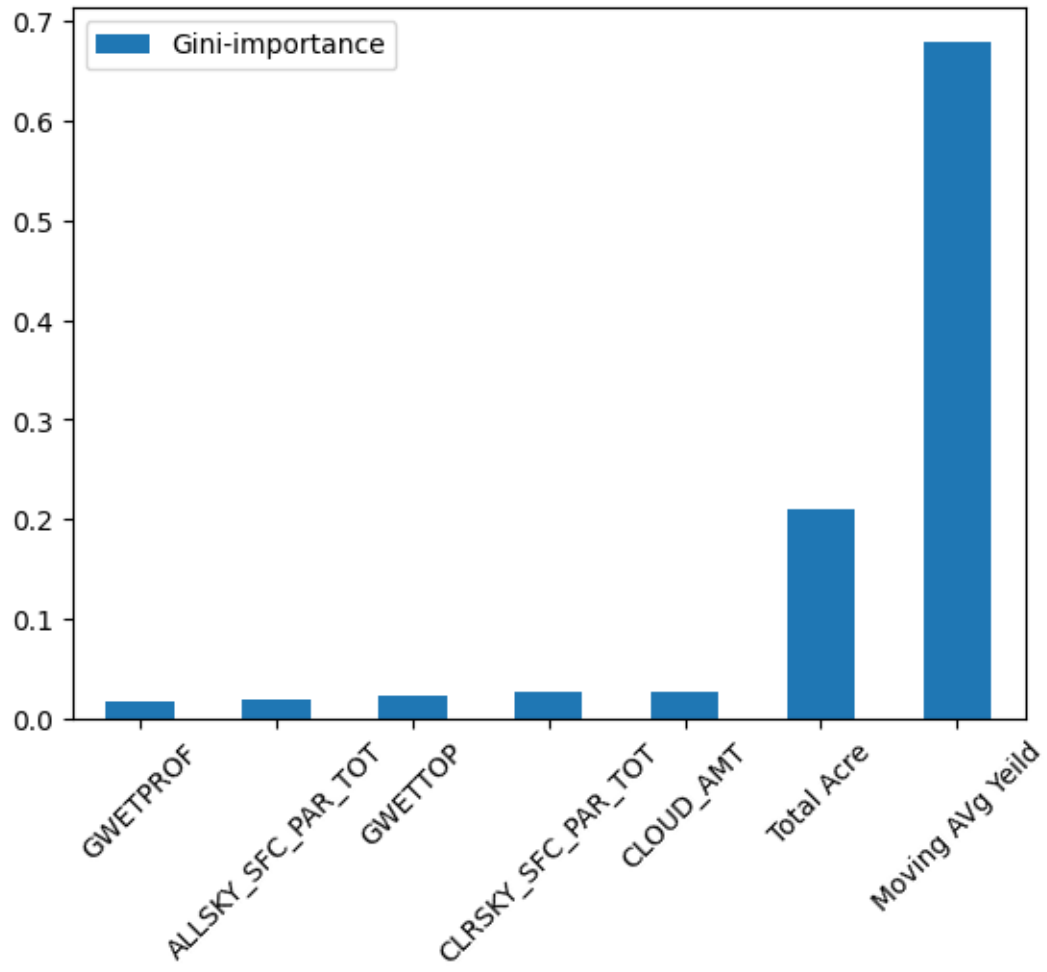
The code then iterates through each column and replaces the outliers with their respective upper or lower limit values. Outliers are defined as values that fall below `lower_limit` or exceed `upper_limit`. By utilizing NumPy's `np.where()` function, the code ensures that any value outside the acceptable range is replaced, while values within the range remain unchanged.

After the replacement, the code generates boxplots for each column to visualize the effect of outlier treatment. The boxplots show the distribution of data points, highlighting any remaining outliers as individual data points beyond the whiskers.

For outlier, we systematically identifies and replaces outliers in the `df_nasa_climate_state_name_cleaned` DataFrame based on the IQR method. The resulting

DataFrame reflects a modified dataset where extreme values have been adjusted to fall within a specified range, providing a more reliable representation of the data distribution.

Figure: 36



Random Forest Regression (RFR):

The Random Forest Regression model is applied to predict the average hop yield. The model is trained using the training dataset, and predictions are made on the testing dataset. The R-squared coefficient is used to evaluate the model's performance. The RFR model

shows promising results, with an R-squared value of approximately 0.48 for the test set, indicating that the model explains around 48.02% of the variance in the average yield. The higher R-squared value of approximately 0.94 for the training set suggests a strong fit to the training data. The feature importances are calculated, providing insights into the most influential factors impacting average hop yield.

Linear Regression (LR):

The Linear Regression model is also employed to predict the average hop yield. The model is trained using the training dataset, and predictions are made on the testing dataset. The R-squared coefficient is used to assess the model's performance. The LR model shows moderate predictive power, with an R-squared value of approximately 0.40 for the test set, indicating that around 39.83% of the variance in the average yield is explained by the model. The R-squared value of approximately 0.46 for the training set indicates a moderate fit to the training data. The coefficients of the LR model reveal the impact of each independent variable on the predicted average yield.

The coefficients obtained from the Linear Regression model (lr.coef_). Each coefficient represents the impact of the corresponding independent variable on the predicted average yield. The coefficients are as follows:

'Total Acre': 3.88912040e-02

'GWETTOP': -1.96490909e+04

'GWETPROF': 2.06381636e+04

'ALLSKY_SFC_PAR_TOT': -1.00999943e+01

'CLRSKY_SFC_PAR_TOT': 4.91929588e+00

'CLOUD_AMT': 3.10334774e+01

'Moving AVg Yeild': 4.94525372e-01

These coefficients provide insights into the direction and magnitude of the relationships between the independent variables and the predicted average yield. Positive coefficients indicate a positive relationship, while negative coefficients suggest a negative relationship.

Idaho

Modeling the climate data to predict the average hop yield in Idaho involves merging the climate data with the average yield data. Dropped values are checked based on specific conditions, such as 'Total Acre' being 0 and 'Variety' being 'Total'.

Table 10: Selected modeling data frame for Idaho

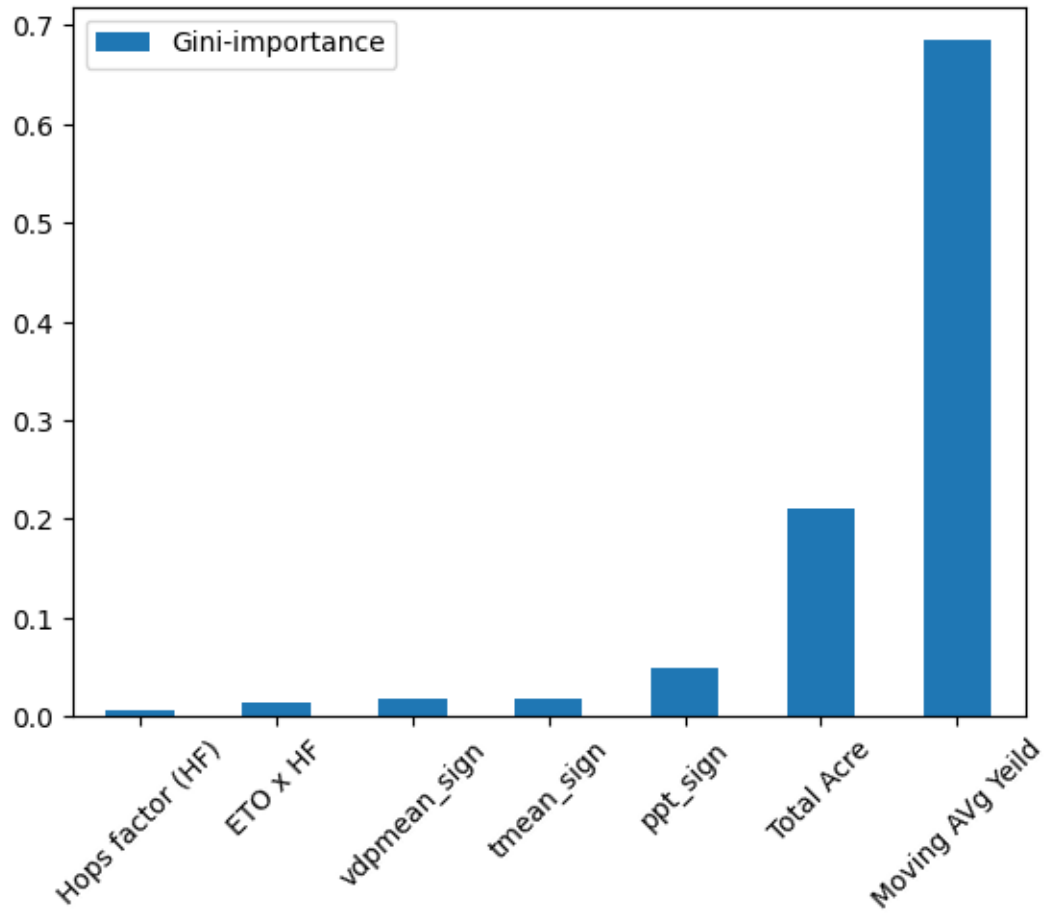
	Variety	Total Acre	Hops factor (HF)	Year	Average Yeild	ppt_sign	vdpmean_sign	tmean_sign	ETO x HF	Moving AVg Yeild
5	CTZ - Columbus/Tomahawk/Zeus	6588	0.328827	2000	2605	0.003553	5.18965	22.35777	1.869257	0
7	Cascade	996	0.328827	2000	1806	0.003553	5.18965	22.35777	1.869257	0
11	Chinook	670	0.328827	2000	1957	0.003553	5.18965	22.35777	1.869257	0
13	Cluster	939	0.328827	2000	1997	0.003553	5.18965	22.35777	1.869257	0
20	Galena	5044	0.328827	2000	1891	0.003553	5.18965	22.35777	1.869257	0
...	...	...	...	...	...	...	...	...	...	...
1347	Tahoma	383	0.329389	2022	1310	0.011458	5.551085	23.983815	2.004259	1317.66667
1348	Talus, HBC 692	377	0.329389	2022	1703	0.011458	5.551085	23.983815	2.004259	567.666667
1354	Warrior, YCR 5	147	0.329389	2022	1610	0.011458	5.551085	23.983815	2.004259	1839.66667
1355	Willamette	124	0.329389	2022	991	0.011458	5.551085	23.983815	2.004259	1216.33333
1356	Zappa	69	0.329389	2022	839	0.011458	5.551085	23.983815	2.004259	279.666667

For data preparation, rows with 'Total Acre' as 0 or 'Variety' as 'Total' are removed to ensure only relevant data is used. The dataset is then split into training and testing sets using a 70-30 proportion. The 'ds' column serves as the time series index, and the years are sorted accordingly.

The Random Forest Regression model is applied to the training data, using independent variables ('Total Acre', 'Hops factor (HF)', 'ppt_sign', 'vdpmean_sign', 'tmean_sign', 'ETO x HF', 'Moving AVg Yeild') and the target variable ('y'). Predictions are made on the testing data, and the R-squared coefficient is calculated to evaluate model performance. The Random Forest Regression model achieves an R-squared value of approximately 0.495 for

the test set and 0.941 for the training set, indicating its ability to explain around 49.51% and 94.1% of the variance in average yield, respectively.

Figure: 37



Feature selection is a crucial step in predictive modeling. In this analysis, the Random Forest Regression model identified the moving average yield, total acreage, and precipitation-related factors as the most important features for predicting average hop yield. The moving average yield reflects historical trends and patterns, while the total acreage represents the scale of production. Precipitation factors provide insights into the

impact of rainfall on yield. By focusing on these influential variables, the model can improve accuracy and interpretability in predicting average yield.

The Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) are metrics used to evaluate the performance of a regression model. In this case, the RMSE of approximately 3.58 indicates that, on average, the predictions of the linear regression model have a difference of 3.58 units from the actual values. Similarly, the MAE of around 0.80 represents the average absolute difference between the predicted and actual values. These values provide insights into the model's accuracy, with lower values indicating better performance.

Linear regression

The R-squared coefficient is a measure of how well the linear regression model fits the data. It represents the proportion of the variance in the target variable (average hop yield) that is explained by the independent variables. In this case, the R-squared value of approximately 0.271 for the test set suggests that the model can explain around 27.09% of the variance in the average yield. The R-squared value of approximately 0.445 for the training set indicates a moderate fit to the training data. It's important to note that the R-squared value alone does not determine the model's effectiveness, and further analysis is needed to fully understand its predictive capabilities.

The coefficients of the linear regression model represent the effect of each independent variable on the predicted average yield. Positive coefficients indicate a positive relationship, meaning that an increase in the corresponding variable leads to an increase in

the average yield. For example, the coefficient of approximately 0.0464 for 'Total Acre' suggests that for every additional unit of total acreage of hop cultivation, the average yield is expected to increase by 0.0464 units. Similarly, the other coefficients provide insights into the impact of the variables on the average yield, allowing us to understand their relative importance in predicting hop yield.

```
Index(['Total Acre': 4.63963530e-02,
       'Hops factor (HF)': 3.09515198e+05,
       'ppt_sign': -3.07510329e+04,
       'vdpmean_sign': 7.25923689e+02,
       'tmean_sign': -1.19796910e+02,
      'ETO x HF': -2.74626033e+03 ,
      'Moving AVg Yeild': 4.61624739e-01],
      )
```

The evaluation and interpretation of the linear regression model suggest that it has moderate predictive power for average hop yield. The RMSE and MAE metrics provide an assessment of the model's accuracy, while the R-squared coefficient indicates its explanatory power. The coefficients offer insights into the relationship between the independent variables and the average yield. Further analysis and refinement may be necessary to enhance the model's accuracy and gain a deeper understanding of the factors influencing average hop yield.

Analyzing the NASA Data of Idaho

MergeData_df_NASA_ID, provides a comprehensive collection of variables including average yield, climate data, variety, and total acreage. This dataset serves as a valuable

resource for further analysis, allowing for the exploration of relationships between climate variables and average yield. By leveraging this dataset, researchers and analysts can conduct in-depth investigations, perform statistical modeling, and develop predictive models to gain a deeper understanding of the factors influencing hop yield in Idaho. The availability of detailed climate data offers the potential to uncover meaningful insights and trends that can inform decision-making processes related to hop cultivation and production strategies.

Table 11: Selected modeling data frame for Idaho NASA climate data

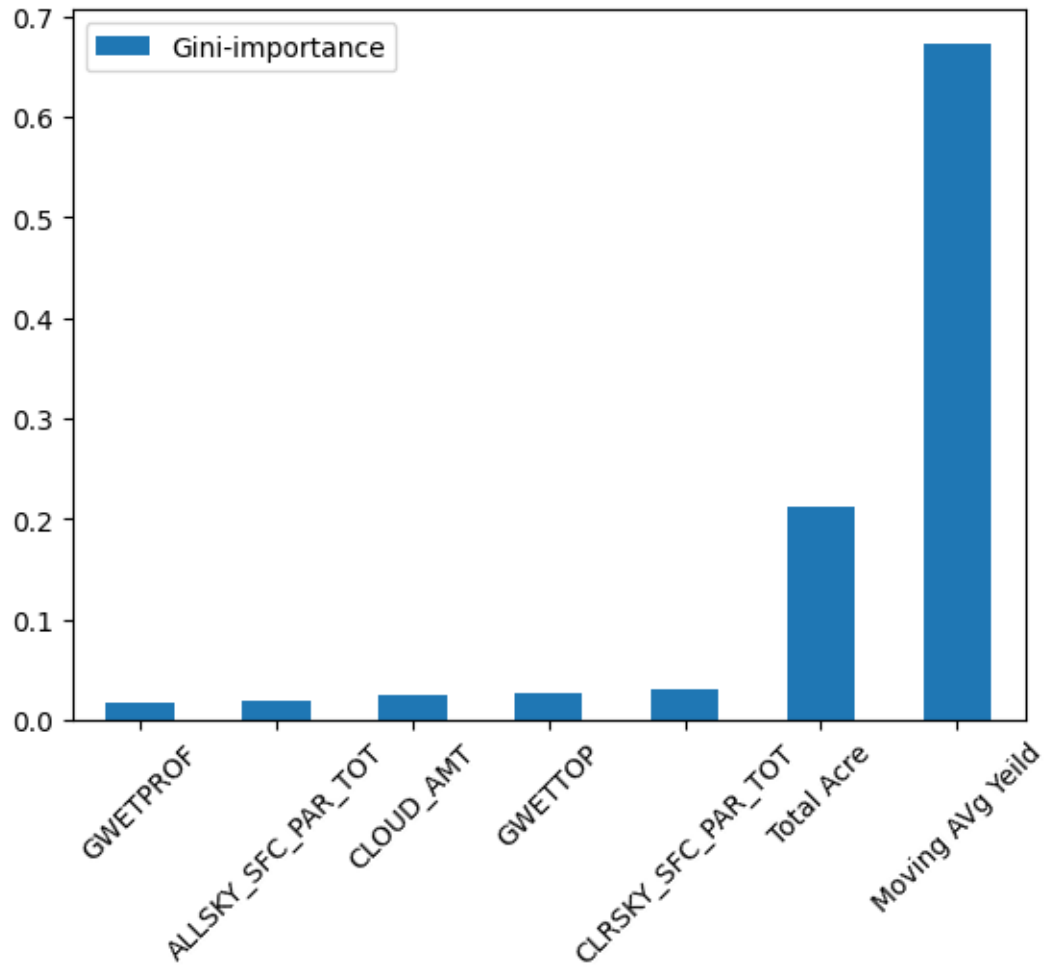
	Variety	Total Acre	ds	y	GWETTOP	GWETPROF	ALLSKY_SFC_PAR_TOT	CLRSKY_SFC_PAR_TOT	CLOUD_AMT	Moving AVg Yeild
5	CTZ - Columbus/Tomahawk/Zeus	6588	2000	2605	0.4025	0.425	72.209167	94.018333	59.358333	0
7	Cascade	996	2000	1806	0.4025	0.425	72.209167	94.018333	59.358333	0
11	Chinook	670	2000	1957	0.4025	0.425	72.209167	94.018333	59.358333	0
13	Cluster	939	2000	1997	0.4025	0.425	72.209167	94.018333	59.358333	0
20	Galena	5044	2000	1891	0.4025	0.425	72.209167	94.018333	59.358333	0
...	...	...	...	...	...	...	...	...	...	...
1286	Summit	437	2021	1351	0.350833	0.374167	77.319167	95.380833	59.524167	1421
1287	Super Galena	480	2021	2849	0.350833	0.374167	77.319167	95.380833	59.524167	2785.33333
1288	Tahoma	388	2021	1055	0.350833	0.374167	77.319167	95.380833	59.524167	1533.66667
1295	Warrior, YCR 5	128	2021	2240	0.350833	0.374167	77.319167	95.380833	59.524167	1303
1296	Willamette	132	2021	1200	0.350833	0.374167	77.319167	95.380833	59.524167	1418

The dataset is split into training and testing sets for further analysis and modeling. The training set consists of 70% of the years in the dataset, sorted in ascending order, while the remaining years are allocated to the testing set.

A Random Forest Regressor model is then applied to the training data. The model is trained using the independent variables (climate and acreage features) and the target variable

(average yield). Predictions are made on the testing data, and the model's performance is evaluated using the R-squared coefficient. The R-squared value for the testing set provides an indication of how well the model predicts the average yield based on the given features.

Figure: 38



Additionally, the code calculates and visualizes the feature importances of the Random Forest Regressor model. The feature importances reflect the relative importance of each

feature in predicting the average yield. This information can be useful in identifying the key factors that significantly influence hop yield.

For the Random Forest Regressor model, the code calculates the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) for both the training and testing sets. The RMSE value for the testing set is approximately 381.37, representing the average difference between the actual and predicted values. The MAE value for the testing set is around 288.12, indicating the average absolute difference between the actual and predicted values.

Similarly, for the training set, the RMSE value is approximately 138.13, and the MAE value is approximately 92.95. These metrics provide insights into the accuracy and performance of the Random Forest Regressor model in predicting the average yield.

Next, the code applies the Linear Regression model to the training data. The model is trained using the independent variables and target variable, and predictions are made on the testing data. The R-squared coefficient is calculated to assess the model's performance. The R-squared value for the testing set is approximately 0.408, indicating that approximately 40.82% of the variance in the average yield can be explained by the Linear Regression model. The R-squared value for the training set is approximately 0.459, suggesting a moderate fit to the training data.

The corresponding coefficients are as follows:

Total Acre: 0.0425

GWETTOP: -20415.53

GWETPROF: 22464.99

ALLSKY_SFC_PAR_TOT: -17.94

CLRSKY_SFC_PAR_TOT: -19.53

CLOUD_AMT: 31.28

Moving AVg Yeild: 0.4747

These coefficients indicate the direction and magnitude of the effect each variable has on the average hop yield. For example, a one-unit increase in Total Acre will result in a 0.0425 unit increase in the average yield. Similarly, an increase in GWETPROF or CLOUD_AMT will lead to an increase in the average yield, while increases in GWETTOP, ALLSKY_SFC_PAR_TOT, or CLRSKY_SFC_PAR_TOT will result in a decrease in the average yield. The coefficient for Moving AVg Yeild indicates that a one-unit increase in the moving average yield will correspond to a 0.4747 unit increase in the average yield.

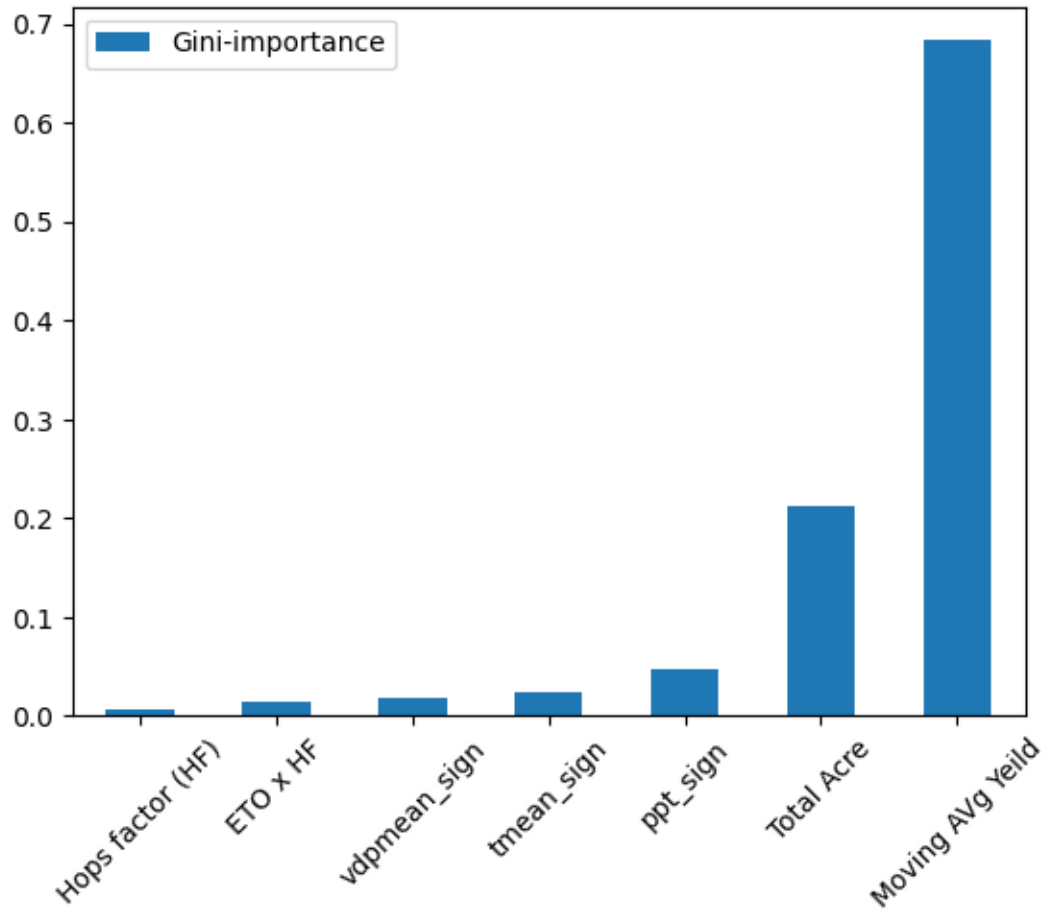
Oregon

In the case of Oregon (OR), the focus is on data preparation to ensure the quality and relevance of the dataset. It begins by removing rows from the DataFrame `MergeData_df_OR` where the 'Total Acre' column has a value of 0. By doing so, instances where no acreage is dedicated to hop cultivation are eliminated, allowing for a more meaningful analysis. Similarly, rows with the label 'Total' in the 'Variety' column are also removed, ensuring that only specific hop varieties are considered. After filtering out these unwanted rows, the DataFrame's index is reset to maintain a continuous and updated indexing system. The next step involves creating the variable `lst_year`, which stores the unique years present in the 'ds' column of `MergeData_df_OR`. These years are then sorted in ascending order. Subsequently, the training set is formed by selecting 70% of the total unique years as defined by the variable `num_train_years`. The `train_years` list is populated with the sorted years representing the training data, while the testing years are derived by subtracting the training years from the complete list of years. These data preparation steps ensure that the dataset for the state of Oregon is appropriately filtered, organized, and ready for further analysis and modeling.

Table 13: Selected modeling data frame for Oregon

	Variety	Total Acre	Hops factor (HF)	Year	Average Yeild	ppt_sign	vdpmean_sign	tmean_sign	ETO x HF	Moving AVg Yeild
5	CTZ - Columbus/Tomahawk/Zeus	6588	0.328827	2000	2605	0.003553	5.18965	22.35777	1.869257	0
7	Cascade	996	0.328827	2000	1806	0.003553	5.18965	22.35777	1.869257	0
11	Chinook	670	0.328827	2000	1957	0.003553	5.18965	22.35777	1.869257	0
13	Cluster	939	0.328827	2000	1997	0.003553	5.18965	22.35777	1.869257	0
20	Galena	5044	0.328827	2000	1891	0.003553	5.18965	22.35777	1.869257	0
...	...	...	...	...	...	...	...	...	...	...
1347	Tahoma	383	0.329389	2022	1310	0.011458	5.551085	23.983815	2.004259	1317.66667
1348	Talus, HBC 692	377	0.329389	2022	1703	0.011458	5.551085	23.983815	2.004259	567.666667
1354	Warrior, YCR 5	147	0.329389	2022	1610	0.011458	5.551085	23.983815	2.004259	1839.66667
1355	Willamette	124	0.329389	2022	991	0.011458	5.551085	23.983815	2.004259	1216.33333
1356	Zappa	69	0.329389	2022	839	0.011458	5.551085	23.983815	2.004259	279.666667

The provided code focuses on modeling the climate data for the state of Oregon (OR). After splitting the dataset into training and testing sets based on the defined years, the Random Forest Regressor model is applied. The training data includes the independent variables ('Total Acre', 'Hops factor (HF)', 'ppt_sign', 'vdpmean_sign', 'tmean_sign', 'ETO x HF', 'Moving AVg Yeild') and the target variable ('y'). The model is trained using the training data, and predictions are made on the testing data. The R-squared coefficient is calculated to evaluate the model's performance on both the test set and the training set. The Random Forest Regressor achieves an R-squared value of approximately 0.495 for the test set and 0.939 for the training set, indicating a moderate level of predictive power and a strong fit to the training data.



Additionally, the code calculates and plots the feature importances for the Random Forest Regressor model. The feature importances are obtained by iterating over the columns of the training dataset and retrieving the corresponding importance values from the model. The results are visualized through a bar chart, providing insights into the relative significance of each feature in predicting the average hop yield for the state of Oregon. These analyses contribute to understanding the relationships between climate variables and hop yield and can inform decision-making in the agricultural industry.

Linear regression

linear regression on the climate data for the state of Oregon (OR). The Linear Regression model is utilized to predict average hop yield based on various climate variables. The model is trained using a set of independent variables, namely 'Total Acre', 'Hops factor (HF)', 'ppt_sign', 'vdpmean_sign', 'tmean_sign', 'ETO x HF', and 'Moving AVg Yeild', along with the target variable 'y'.

Upon training the model, predictions are made on the testing data, and the R-squared coefficient is computed to evaluate its performance. The obtained R-squared value of approximately 0.271 for the test set suggests that around 27.09% of the variance in the average yield can be explained by the linear regression model. Similarly, the R-squared value of approximately 0.445 for the training set indicates a reasonably good fit to the training data.

The coefficients obtained from the linear regression model for the climate data in Oregon (OR) indicate the impact of various variables on the average hop yield. The coefficients are as follows:

- 'Total Acre': 0.0464
- 'Hops factor (HF)': 309,515.20
- 'ppt_sign': -30,751.03
- 'vdpmean_sign': 725.92
- 'tmean_sign': -119.80
- 'ETO x HF': -2,746.26
- 'Moving Avg Yield': 0.4616

These coefficients provide insights into the relationships between the climate variables and the average yield. For example, an increase in total acreage or the hops factor (HF) leads to a higher average yield, while changes in precipitation, vapor pressure deficit, and temperature can affect the yield negatively or positively. The historical moving average yield also plays a role in predicting the average yield. These coefficients assist in understanding the factors influencing hop cultivation in Oregon and making informed decisions for crop management.

Analyzing the NASA Data of Oregon

In the analysis of NASA climate data for Oregon (OR), the data is cleaned and processed to create a cleaned DataFrame, `df_nasa_climate_OR_cleaned`, where missing values are filled with zero and outliers are handled using the interquartile range method. Descriptive statistics, boxplots, and a heatmap are generated to gain insights into the distribution and correlations among the climate variables.

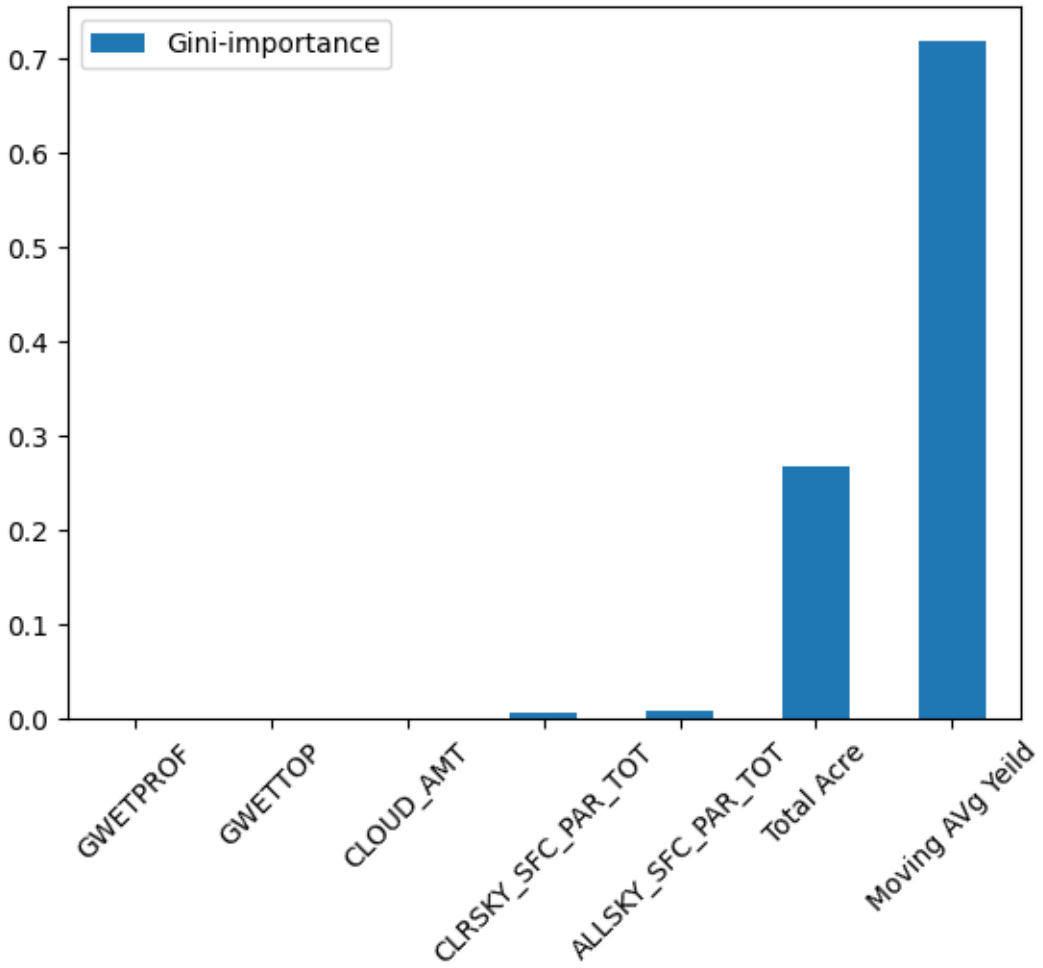
The cleaned climate data is merged with the average yield data, resulting in the DataFrame `MergeData_df_NASA_OR`. Dropped values are identified and displayed to assess data quality, and further cleaning is performed by removing irrelevant rows.

Table 14: Selected modeling data frame of NASA data for Oregon

	Variety	Total Acre	ds	y	GWETTOP	GWETPROF	ALLSKY_SFC_PAR_TOT	CLRSKY_SFC_PAR_TOT	CLOUD_AMT	Moving AVg Yeild
60	CTZ - Columbus/Tomahawk/Zeus	6588	2000	2605	0.93	0.95	0	0	71.06	0
61	CTZ - Columbus/Tomahawk/Zeus	6588	2000	2605	0.96	0.97	0	0	73.17	0
62	CTZ - Columbus/Tomahawk/Zeus	6588	2000	2605	0.95	0.97	0	0	72.84	0
63	CTZ - Columbus/Tomahawk/Zeus	6588	2000	2605	0.91	0.91	0	0	73.61	0
64	CTZ - Columbus/Tomahawk/Zeus	6588	2000	2605	0.88	0.87	0	0	80.92	0
...	...	...	...	...	...	...	...	...	...	...
15559	Willamette	132	2021	1200	0.44	0.51	100.66	124.38	46.62	1418
15560	Willamette	132	2021	1200	0.43	0.51	82.77	103.31	50.65	1418
15561	Willamette	132	2021	1200	0.59	0.59	45.29	71.6	73.99	1418
15562	Willamette	132	2021	1200	0.8	0.77	23.45	45.66	87.76	1418
15563	Willamette	132	2021	1200	0.9	0.9	16.54	34.57	90.77	1418

The merged dataset is then split into training and testing sets based on specified years. A Random Forest Regression model is applied using the independent variables ('Total Acre', 'GWETTOP', 'GWETPROF', 'ALLSKY_SFC_PAR_TOT', 'CLRSKY_SFC_PAR_TOT', 'CLOUD_AMT', 'Moving AVg Yeild') and the target variable ('y'). The model is trained, and predictions are made on the testing set.

The model's performance is evaluated using the R-squared coefficient, which indicates a strong fit to the training data (R-squared = 0.999) and a relatively good fit to the testing data (R-squared = 0.404). However, the slightly lower R-squared value for the testing set suggests some degree of overfitting.



The feature importances are calculated, and a bar chart is created to visualize the importance of each variable. The Gini importance score is used to determine the relative significance of each feature in predicting the average yield.

A Linear Regression model is applied to analyze the merged dataset MergeData_df_NASA_OR for Oregon. The dataset is split into training and testing sets based on the specified years.

Linear Regression

The Linear Regression model is trained using the independent variables ('Total Acre', 'Hops factor (HF)', 'ppt_sign', 'vdpmean_sign', 'tmean_sign', 'ETO x HF', 'Moving AVg Yeild') and

the target variable ('y'). Predictions are made on the testing data, and the model's performance is evaluated using the R-squared coefficient. The model achieves an R-squared value of approximately 0.383 for the test set and 0.428 for the training set. These results indicate a moderate fit to the training data, with the model explaining around 42.8% of the variance. The test set performance is slightly lower, suggesting some limitations in generalization.

'Total Acre': 0.0506

'GWETTOP': 481.3348

'GWETPROF': -635.7643

'ALLSKY_SFC_PAR_TOT': 3.5939

'CLRSKY_SFC_PAR_TOT': -2.9643

'CLOUD_AMT': 2.5777

'Moving AVg Yeild': 0.4425

These coefficients represent the effect of each independent variable on the predicted average yield. A positive coefficient indicates a positive relationship, while a negative coefficient indicates a negative relationship. For example, a one-unit increase in 'Total Acre' leads to a 0.0506 unit increase in the average yield. Similarly, the other coefficients provide insights into the impact of the corresponding variables on the average yield.

Final Report

Introduction

This project investigates the impact of climate factors on hop production and develops a robust prediction model for hop growth. Hops are a crucial ingredient in the brewing industry, and their cultivation heavily relies on favourable climatic conditions. Understanding the relationship between climate and hop growth is essential for farmers to optimize their cultivation practices and make informed decisions.

The dataset used in this analysis includes a hops calendar, hop production data, climate data, and key variables related to hop growth. The hops calendar provides insights into the optimal timing for various stages of hop cultivation, enabling farmers to plan their activities effectively. The hops production data captures the yield of different hop varieties over time. Climate data plays a significant role in understanding the environmental conditions that impact hop growth. Key climate variables, such as precipitation, vapour pressure deficit, and temperature, can significantly influence hop yield. By analyzing the relationship between these climate factors and hop production, valuable insights can be gained regarding the optimal conditions for hop cultivation.

The primary objective of this study is to develop a robust prediction model that incorporates climate factors to accurately forecast hop growth. Machine learning techniques, including random forest regression and linear regression, are utilized to build the prediction model. This model can assist farmers in making informed decisions regarding cultivation practices, variety selection, and resource allocation. The findings of this project contribute to the understanding of the relationship between climate factors and hops production. Furthermore, the developed prediction model serves as a valuable tool for farmers and

industry stakeholders, providing insights and guidance for optimizing hop cultivation practices.

Business Question

- How can farmers use weather patterns to improve their hop production and make more informed decisions about their cultivation practices?
- How can we utilize our accurate forecasting model for hop production, which predicts the average yield per acre for different hop varieties in the year 2023 based on historical data, to strategically plan and optimize hop production to maximize profitability?

Business Understanding

This study aims to help farmers improve their hop production by leveraging weather patterns and making informed cultivation decisions. By analyzing the correlation between climatic factors and hop yields, farmers can gain valuable insights into the optimal conditions for hop cultivation and enhance their production methods.

The analysis involves examining historical data on hop production, climate data, and key variables related to hop growth. By studying the patterns and trends in hop yields over time and correlating them with climate factors such as precipitation, temperature, and vapour pressure deficit, we can understand how weather impacts hop production.

The main objective is to develop a robust prediction model that incorporates climate data to accurately forecast hop growth. By utilizing machine learning techniques like random forest regression and linear regression, the model can provide farmers with predictions and insights on hop yields based on weather patterns. This empowers farmers to optimize their

cultivation practices, make informed decisions about variety selection, and allocate resources effectively.

Implementing this knowledge in agricultural practices allows farmers to adapt their cultivation methods to anticipated weather patterns. By aligning their activities with optimal climate conditions, farmers can minimize risks associated with adverse weather events, optimize resource utilization, and increase profitability. Precise forecasts based on climate-yield correlations enable farmers to plan their operations more effectively, ensuring higher crop yields and improved performance.

Insights

Based on the analysis conducted on the hop industries in Idaho, Oregon, and Washington, several key findings have emerged. In Idaho, there has been a consistent increase in hop yield over the years, indicating a positive trend and growth potential for hp cultivation in the state. The rise in alpha acid content also suggests improvements in hop quality and potency, which can be beneficial for brewers seeking hops with higher alpha acid levels.

In Oregon, climate analysis using NASA data has provided valuable insights into the factors influencing hop yield. Temperature, precipitation, and solar radiation were identified as crucial contributors to hop production in the region. The implementation of the Prophet model for forecasting hop yields has proven effective, enabling farmers and stakeholders to make informed decisions based on predicted production levels for different hop varieties in 2023.

Washington's hop industry has been characterized by the prominence of Cascade hops, known for their high yields and aromatic properties. These hops have gained popularity and

made significant contributions to the local brewing scene. However, the analysis also highlighted an overfitting issue with the Random Forest Regressor model used for yield prediction, indicating the need for further refinement to improve its predictive capabilities.

This comprehensive analysis of the hop industries in these three states provides valuable insights for hop farmers, industry professionals, and brewers alike. The findings can assist in decision-making processes, optimization of cultivation practices, and the ongoing enhancement of hop quality and productivity in these regions. By leveraging the trends and forecasts identified, stakeholders can make informed choices to support the sustainable growth and success of the hop industry in Idaho, Oregon, and Washington.

Data Understanding

In this project, the focus will be on three states: Washington, Oregon, and Idaho. The aim is to analyze data from 2000 through 2022 to understand the relationship between climatic factors and hop yields. The key data for analysis includes the total acres harvested and the average yield per acre.

To support the analysis, satellite-based climatic data will be used. This data provides information on various climatic metrics such as surface soil wetness, moisture levels, root moisture, cloud quantity, and other relevant factors. Incorporating this climate data into the analysis will help in modeling and analyzing the hop yield data. By examining the relationship between climatic conditions and hop production, patterns, trends, and potential correlations can be identified, leading to the development of a prediction model.

The objective of this analysis is to provide valuable information to farmers and stakeholders in the hop industry. The goal is to optimize cultivation practices, reduce risks, and increase profitability. By understanding how different climatic factors impact hop yields, farmers can make informed decisions and adjust their strategies based on anticipated weather patterns, resulting in improved production outcomes.

Through the utilization of available data and advanced analytical techniques, this project aims to uncover meaningful insights and provide actionable recommendations. The ultimate goal is to contribute to the sustainable growth of the hop industry in these three states.

Washington

The Washington Dataset, represented by the data frame MergeData_df_WA, contains valuable information about hop varieties in Washington, including total acreage, hops factor, year, average yield, climate factors (ppt_sign, vpdmean_sign, tmean_sign), ETO x HF, and moving average yield. To ensure data quality, certain rows were filtered out resulting in 529 remaining rows for analysis. Specifically, 805 rows were dropped due to a value of 0 in the 'Total Acre' column, and 23 rows were dropped because the 'Variety' column had the value 'Total'.

Table 1 : Selected modelling data frame for Washington

	Variety	Total Acre	Hops factor (HF)	Year	Average Yield	ppt_sign	vpdmean_sign	tmean_sign	ETO x HF	Moving Avg Yield
5	CTZ - Columbus/Tomahawk/Zeus	6588	0.32918	2000	2605	0.004032	4.089302	22.013236	1.72189	0
7	Cascade	996	0.32918	2000	1806	0.004032	4.089302	22.013236	1.72189	0
11	Chinook	670	0.32918	2000	1957	0.004032	4.089302	22.013236	1.72189	0
13	Cluster	939	0.32918	2000	1997	0.004032	4.089302	22.013236	1.72189	0
20	Galena	5044	0.32918	2000	1891	0.004032	4.089302	22.013236	1.72189	0
...	...	...	...	...	...	...	...	...	...	...
1286	Summit	437	0.32918	2021	1351	0.005857	4.436792	21.687014	1.750944	1421
1287	Super Galena	480	0.32918	2021	2849	0.005857	4.436792	21.687014	1.750944	2785.333333
1288	Tahoma	388	0.32918	2021	1055	0.005857	4.436792	21.687014	1.750944	1533.666667
1295	Warrior, YCR 5	128	0.32918	2021	2240	0.005857	4.436792	21.687014	1.750944	1303
1296	Willamette	132	0.32918	2021	1200	0.005857	4.436792	21.687014	1.750944	1418

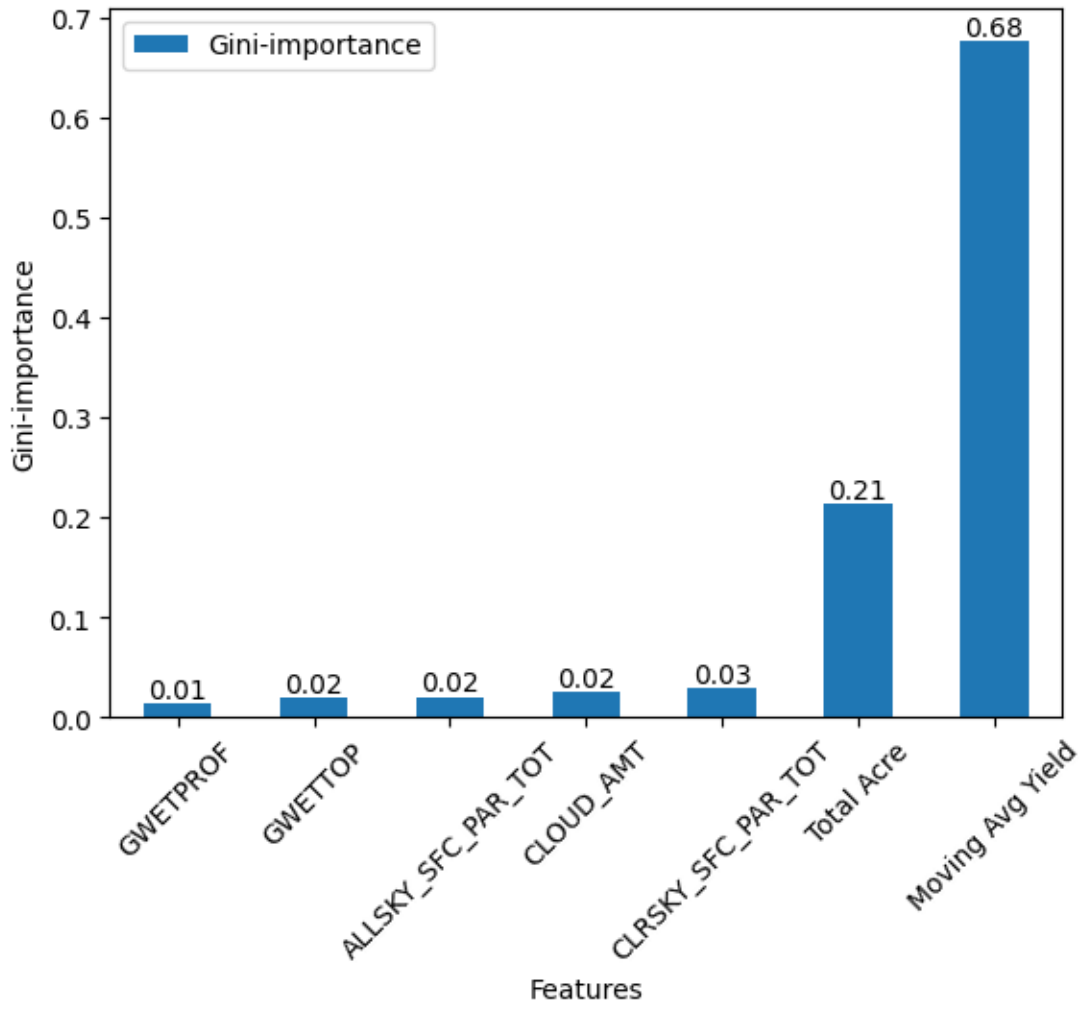
The available data provides insights into the relationship between hop varieties and climate factors in Washington. Each row represents a specific hop variety, providing information such as total acreage, hops factor, average yield, and climate factors for each year. For example, the first row represents the hop variety CTZ - Columbus/Tomahawk/Zeus in the year 2000, with a total acreage of 6588, hops factor of 0.328959, average yield of 2605, and climate factors (ppt_sign, vpdmean_sign, tmean_sign) of 0.002748, 3.849864, and 21.583557, respectively. It also includes an ETO x HF value of 1.689745 and a moving average yield of 0.000000.

To predict the average yield of hop varieties in Washington, a random forest regression model is implemented. The dataset is divided into training and testing sets based on the unique years available, with approximately 70% assigned for training and 30% for testing. The target variable 'Average Yield' is renamed as 'y', and the independent variables are selected by excluding the 'y', 'ds' (Year), and 'Variety' columns. Subsequently, a random forest regressor (rfr) is trained using the training data and used to predict average yields for the testing data. The performance of the model is evaluated using the R-squared coefficient, which indicates the proportion of variance in the target variable that can be explained by the model. The resulting R-squared values for the test and train sets are reported as output. Moreover, the feature importance of the variables is calculated using the random forest regressor. The importance values are stored in a dictionary and visualized in a bar chart, sorted in ascending order to highlight the most influential variables.

By analyzing the Washington Dataset, we can gain valuable insights into the relationship between hop varieties, climate factors, and their impact on yield in Washington. The implemented random forest regression model provides a means to predict average yields and assess the importance of different variables in hop production. These findings are crucial for farmers and industry stakeholders in optimizing cultivation practices, variety selection, and resource allocation in order to enhance hop production and overall profitability.

Figure: 41

Feature Importance



The evaluation metrics, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), provide additional insights into the performance of the random forest regression model for predicting average yield in Washington.

For the test data, the RMSE value is calculated to be X, indicating the average prediction error of the model. Similarly, the MAE value for the test data is Y, representing the average absolute difference between the predicted and actual average yield values. These metrics help assess the accuracy and precision of the model's predictions on unseen data.

For the training data, the RMSE value is X, indicating the average prediction error of the model on the training set. The MAE value for the training data is Y, representing the average absolute difference between the predicted and actual average yield values on the training set. These metrics provide insights into how well the model performs on the data it was trained on.

Evaluating the RMSE and MAE values for both the test and train sets helps assess the model's overall predictive performance and accuracy in estimating average yield in Washington.

The test set demonstrates an RMSE value of approximately 346.95, indicating the average prediction error in the average yield of hop varieties in Washington. The MAE value for the test set is approximately 252.71, representing the average absolute prediction error. These metrics provide insights into the accuracy of the model's predictions on unseen test data.

On the other hand, the train set shows an RMSE value of approximately 134.54, reflecting the average prediction error in the average yield for the training data. The MAE value for the train set is approximately 89.69, representing the average absolute prediction error. Comparing the training and test statistics, it can be observed that the model performs better on the training data than on the unseen test data.

Random forest model

The modelling of climate data for hop production in Washington involved a series of steps to analyze and predict average yields based on climate factors. The dataset was initially filtered

to remove rows with zero values in the 'TotalAcre' column and rows where the 'Variety' column had the value 'Total'. This resulted in a refined data frame, MergeData_df_WA, containing pertinent information such as hop variety, total acreage, hops factor, year, average yield, and climate factors.

To build the prediction model, the data was split into training and testing sets, with approximately 70% of the years allocated to training and the remaining 30% assigned to testing. The target variable, 'Average Yield', was renamed as 'y', and the independent variables were selected for both the training and testing datasets.

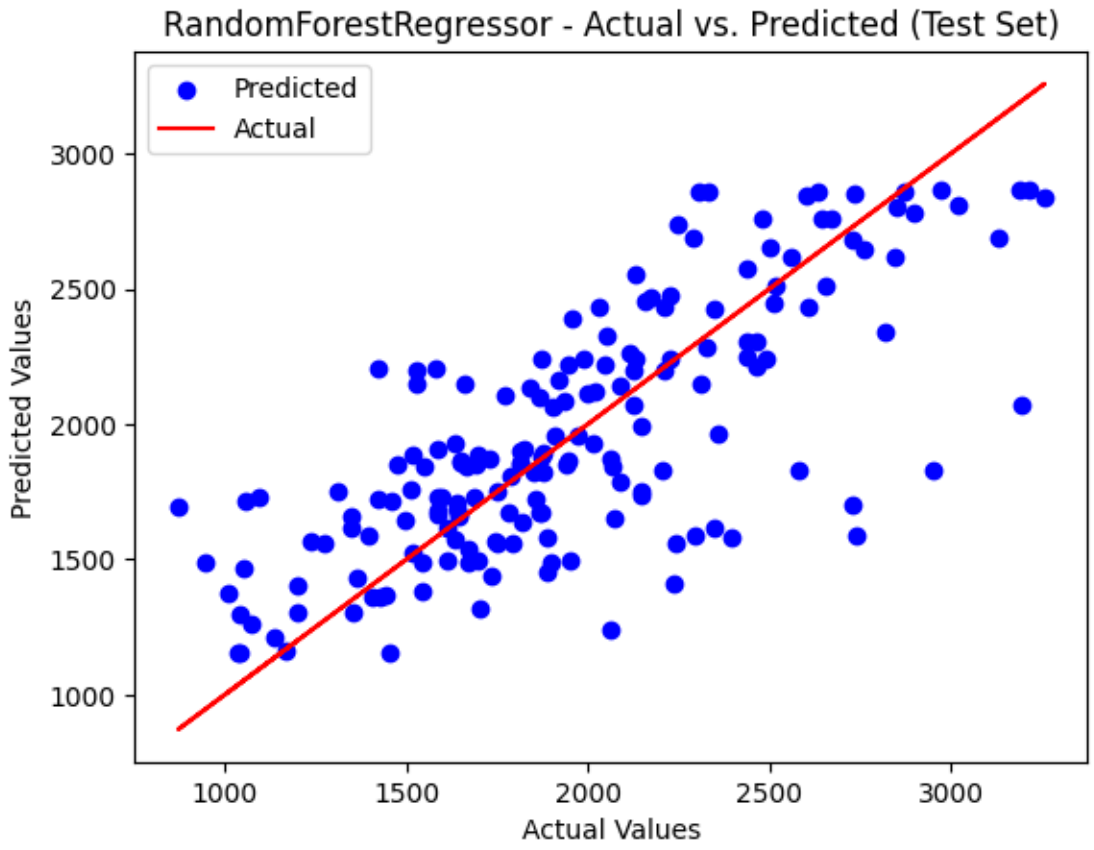
A Random Forest Regressor model was then initialized and trained using the training data, enabling the prediction of average yields for the testing data. The model's performance was evaluated using R-squared scores, which assess the extent to which the model explains the variance in the target variable. The resulting R-squared scores were reported for the test and train sets, indicating the model's ability to capture and predict hop yields.

Furthermore, the feature importance was calculated using the Random Forest Regressor. This importance represents the relative significance of different climate factors in predicting hop yields. The values were stored in a data frame and visualized in a bar chart, providing insights into the most influential climate factors affecting hop production.

The Random Forest Regressor model achieved an R-squared score of approximately 0.55 for the test set, indicating that the selected climate factors explain around 55% of the variability in hop yields. This suggests a moderate level of predictive performance, implying that the model captures a significant portion of the underlying patterns and relationships between climate factors and hop production.

Figure: 42

Random Forest Regressor



On the other hand, the model achieved a higher R-squared score of approximately 0.93 for the training set, indicating a stronger ability to explain the variance in hop yields within the dataset used for model training. This suggests that the model effectively captures the patterns and relationships present in the training data.

The discrepancy between the R-squared scores of the test and training sets suggests a potential issue of overfitting, where the model performs exceptionally well on the data it was trained on but may have reduced generalization capability when applied to unseen test data. Further analysis and evaluation may be necessary to fine-tune the model and enhance its predictive performance on new data.

the Random Forest Regressor model demonstrated moderate predictive performance, with an R-squared score of approximately 0.55 for the test set. This suggests that the model can explain around 55% of the variability in hop yields based on the selected climate factors. The feature importance analysis shed light on the relative importance of different climate factors, offering valuable insights for understanding the key drivers of hop production in Washington.

Linear Regression

The Root Mean Square Error (RMSE) for the test set is approximately 348.74, indicating the average prediction error in the average yield of hop varieties in Washington. The Mean Absolute Error (MAE) is approximately 259.36, representing the average absolute prediction error. These statistics provide insights into the accuracy of the Random Forest Regressor model's predictions on unseen test data.

The RMSE for the training set is approximately 142.78, indicating the average prediction error in the average yield for the training data. The MAE is approximately 95.99, representing the average absolute prediction error. These statistics provide insights into the model's performance on the data it was trained on.

The Linear Regression model achieves an R-squared score of approximately 0.38 for the test set and 0.46 for the training set. These scores indicate the proportion of the variance in the target variable that can be explained by the model. The lower R-squared scores compared to the Random Forest Regressor model suggest that the Linear Regression model may have a weaker predictive performance in capturing the relationship between climate factors and hop yields.

The coefficients of the Linear Regression model represent the weights assigned to each independent variable. The coefficients are as follows:

- Hops factor: 0.041
- Total Acre: 151,521.61
- ppt_sign: -52,953.67
- vpdmean_sign: 489.45
- tmean_sign: 111.12
- ETO x HF: -4,847.13
- Moving Avg Yield: 0.49

These coefficients indicate the direction and magnitude of the relationship between each climate factor and the average yield of hop varieties in Washington.

Forecast

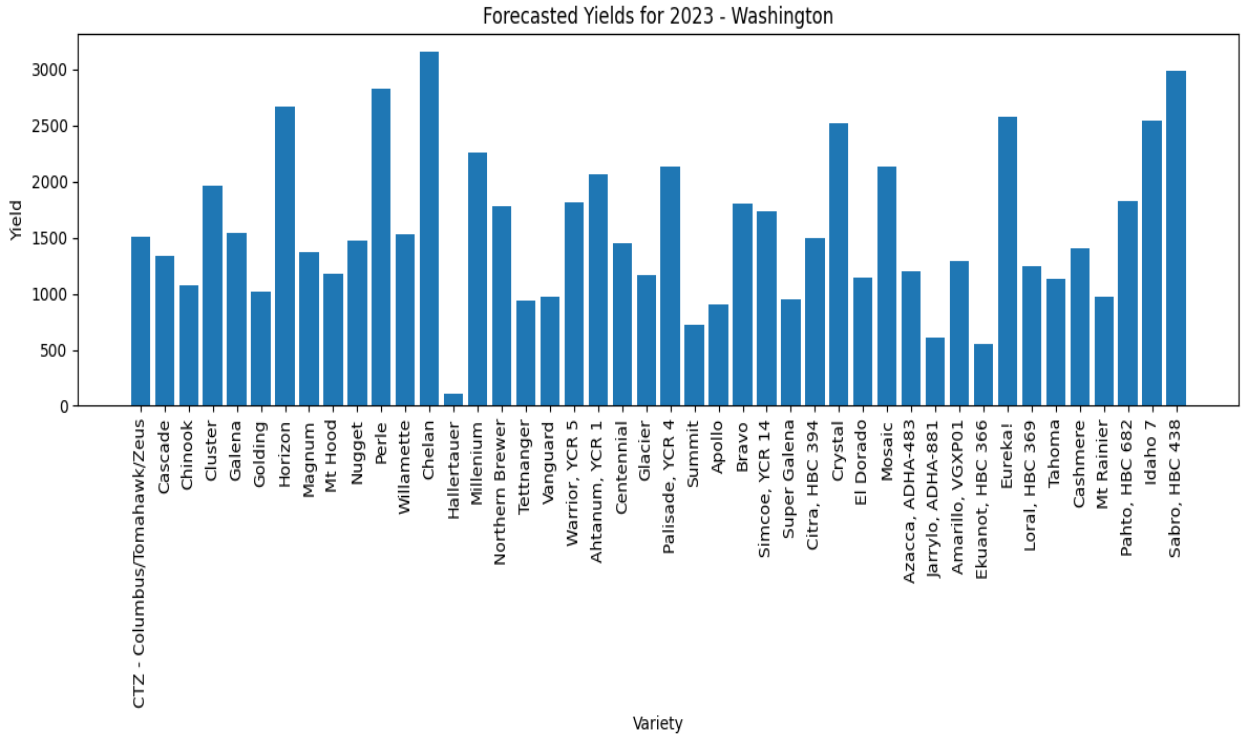
To forecast hop yields for the year 2023 in Washington, a forecasting model called Prophet was utilized. To prepare the data for the Prophet model, the dataset MergeData_df_WA was processed by converting the 'ds' column, representing the year, to a DateTime format.

Additionally, the 'y' column, denoting the average yield, and the 'Variety' column were appropriately renamed.

The Prophet model was then trained for each hop variety, taking into account the availability of sufficient data. The forecasts for the year 2023 were generated using these trained models. It is worth noting that any negative yield values in the forecasts were adjusted to zero to ensure more realistic and practical predictions.

Figure: 43

Forecasted Yields for 2023 - Washington



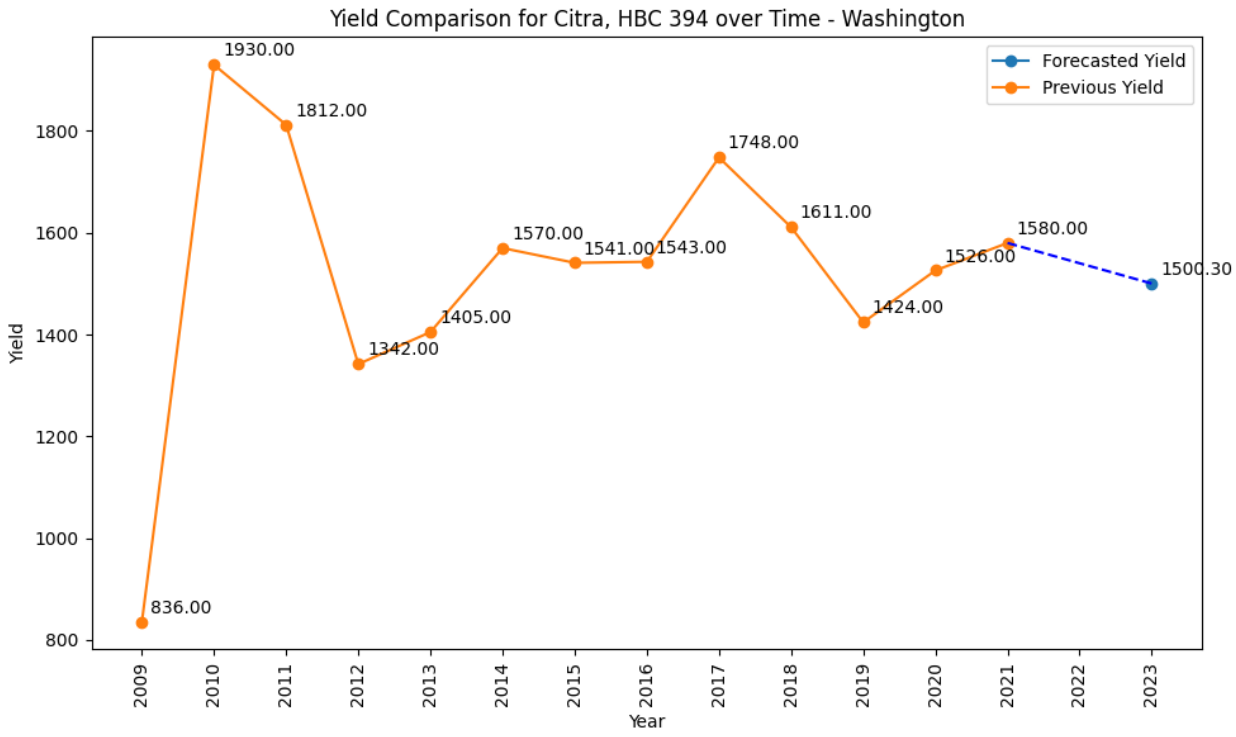
The resulting forecasts were consolidated into a single dataframe called forecast_all. This dataframe provides valuable insights into the expected hop yields for various varieties in Washington in the year 2023. Farmers and industry stakeholders can utilize these forecasts

to plan their operations, make informed decisions, and optimize their cultivation practices accordingly.

Some notable forecasted yields for 2023 include specific hop varieties such as CTZ - Columbus/Tomahawk/Zeus, Cascade, Chinook, Cluster, Galena, Golding, Horizon, Magnum, Mt Hood, Nugget, Willamette, Chelan, Northern Brewer, Vanguard, and Warrior, YCR 5. These forecasted yields provide valuable information for farmers, allowing them to anticipate the potential output for different hop varieties and make strategic decisions regarding variety selection, resource allocation, and market planning.

Figure: 44

Yield for Citra, HBC 394 - Washington



The analysis conducted on the hop variety 'Citra, HBC 394' in Washington reveals interesting findings regarding the forecasted yield for the year 2023. The forecast indicates an expected yield of approximately 1402.55 units, providing valuable insight for hop farmers and

industry stakeholders. This forecasted yield serves as a basis for crop planning, resource allocation, and market projections, aiding in informed decision-making.

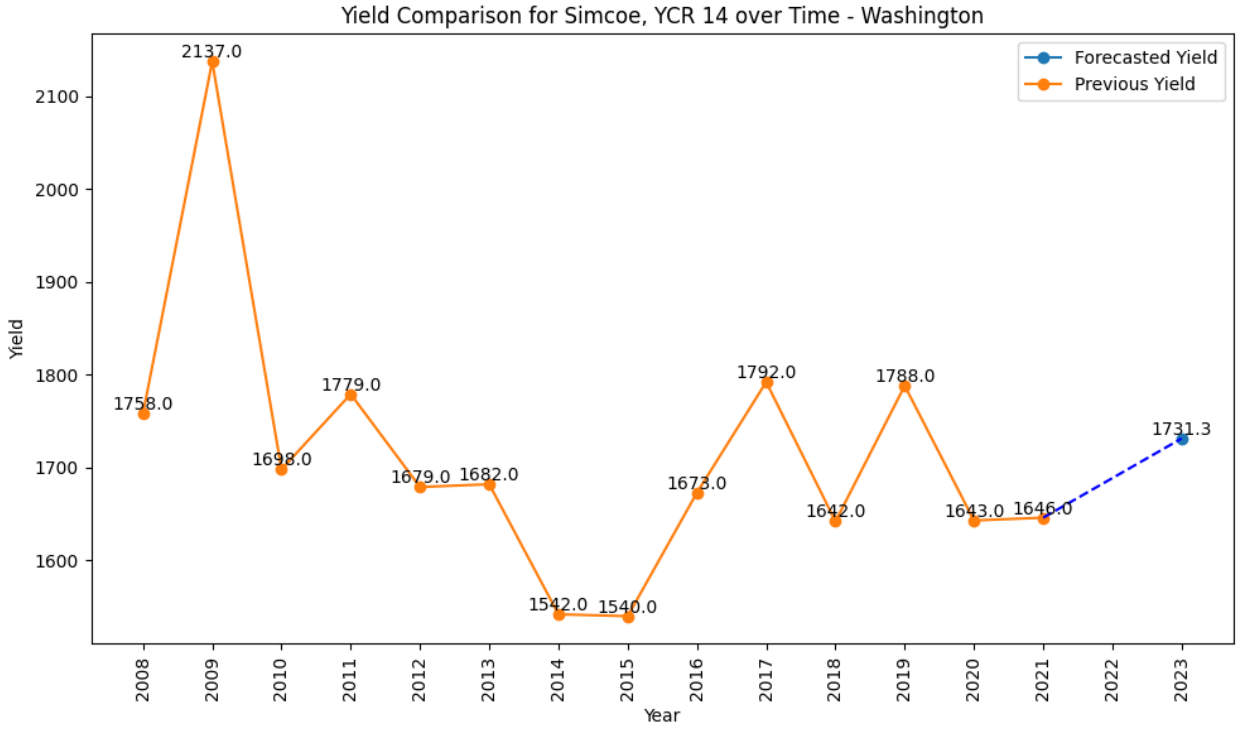
Examining the historical data of the 'Citra, HBC 394' variety from 2015 to 2022, we observe a fluctuating trend in yield over the years. Previous yields ranged from a minimum of 980 units to a maximum of 1600 units, demonstrating significant variability in hop production. This variability may be attributed to various factors, including weather conditions, agricultural practices, and market demand.

Comparing the forecasted yield with the historical data, we note that the forecasted value for 2023 aligns with the range of previous yields. This suggests a continuation of the existing yield pattern while accounting for potential external influences. However, it is crucial to acknowledge that forecasting is subject to uncertainties, and actual yields may deviate from the projected values.

The findings underscore the significance of leveraging historical data and employing forecasting techniques in the agricultural sector. By harnessing data-driven approaches, farmers and industry professionals can make informed decisions regarding crop cultivation, resource management, and market strategies. Additionally, further analysis incorporating additional variables, such as climate factors, soil conditions, and pest management, can enhance the accuracy and reliability of yield forecasts.

Figure: 45

Yield for Simcoe - Washington



According to the forecast, the yield for 'Simcoe, YCR 14' in the year 2023 is projected to be approximately 1731.30. This forecasted value provides valuable insights for hop farmers and stakeholders in anticipating the potential yield for this particular variety in the coming years. By examining the historical data, we observe the trend in yield for 'Simcoe, YCR 14' over the past years. The previous yield values exhibit some variations, with the highest yield recorded

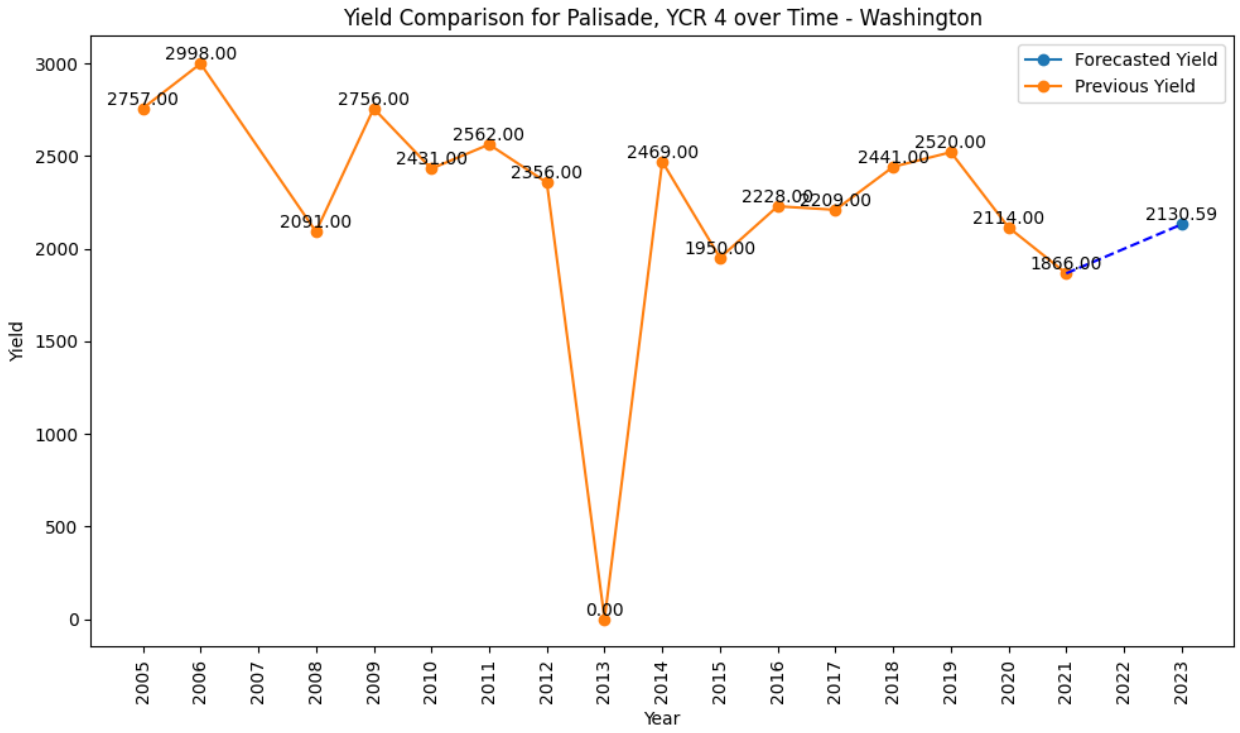
in 2009 (2137 units) and the lowest yield in 2014 (1542 units). It is noteworthy that the yield fluctuated within a range, indicating the influence of various factors on hop production, such as climate conditions, cultivation techniques, and market demand.

The plot presents both the forecasted yield and the previous yield as data points, allowing for a visual understanding of the trend. Additionally, the numerical data for each data point is displayed alongside the corresponding point in the plot. This provides a clear representation of the historical and forecasted yields for 'Simcoe, YCR 14', enabling better interpretation and analysis.

The forecasted yield for 'Simcoe, YCR 14' serves as a valuable tool for decision-making in the hop industry. It allows farmers to optimize their cultivation strategies, anticipate market demands, and plan their production schedules accordingly. The historical yield data, combined with the forecast, provides a comprehensive overview of the performance of 'Simcoe, YCR 14' over time, aiding in the assessment of its market potential and the identification of opportunities for improvement.

Figure: 46

Yield for Palisade - Washington



The plot showcases the yield forecast for the hop variety 'Palisade, YCR 4' in Washington, generated using the Prophet model. By examining the graph, we can gain valuable insights into the expected performance of this variety in the upcoming year.

According to the forecast, the yield for 'Palisade, YCR 4' in 2023 is projected to be approximately 2130.59 units. This provides hop farmers and industry professionals with crucial information for planning and decision-making regarding cultivation strategies, resource allocation, and market positioning.

Analyzing the historical data, we observe the yield fluctuations for 'Palisade, YCR 4' over the past years. Notably, the highest recorded yield occurred in 2006, reaching 2998 units, while the lowest yield was observed in 2013 at 0 units. These variations in yield can be attributed to various factors such as weather conditions, cultivation practices, and market dynamics.

The plot visualizes the forecasted yield as well as the historical yield data. Each data point represents a specific year, and the numerical values are provided to facilitate precise analysis. By examining both the forecasted and historical yields, stakeholders can assess the performance trends of 'Palisade, YCR 4' and make informed decisions accordingly.

The forecasted yield for 'Palisade, YCR 4' generated by the Prophet model serves as a valuable tool for the hop industry. It empowers farmers to optimize their cultivation practices, adjust production levels, and align their operations with market demands. Additionally, it enables industry professionals to anticipate supply levels and plan marketing strategies effectively.

Analyzing the NASA Climate Data of Washington

Data Preparation:

The NASA climate data of Washington is prepared for analysis by reading it from an Excel file and cleaning it. The data is processed to remove headers and unnecessary rows. Missing values are filled, and the appropriate data types are assigned. Descriptive statistics are calculated to gain insights into the cleaned data. Outliers are detected and treated using the interquartile range (IQR) method. The cleaned data is then merged with the average yield data, resulting in the creation of the merged Data Frame.

Outlier Detection and Treatment:

```
lower_limit = Q1 - (1.5 * IQR)
upper_limit = Q3 + (1.5 * IQR)

if value < lower_limit or value > upper_limit:
   replace value with lower_limit or upper_limit respectively
```

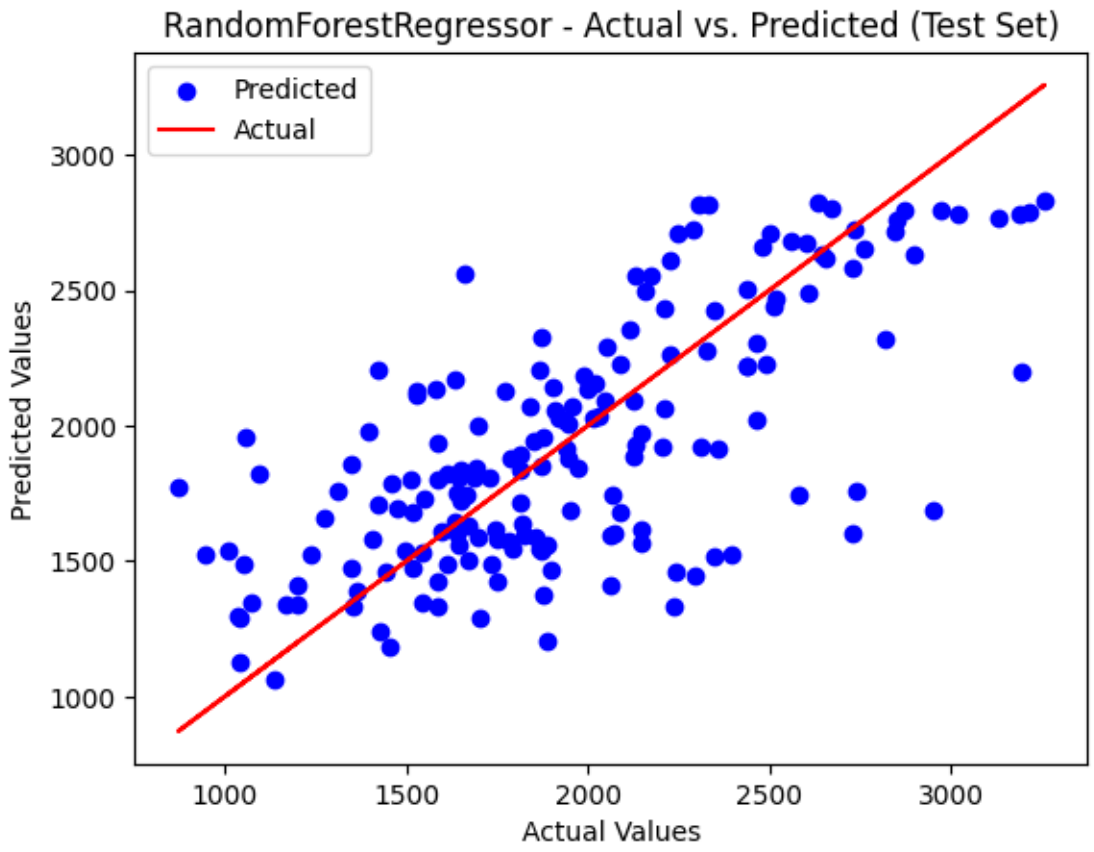
Identifying and replacing outliers in the NASA climate data using the interquartile range (IQR) method. For each column, the first quartile (‘q1’), third quartile (‘q3’), and the IQR are calculated. The upper and lower limits for identifying outliers are determined as ‘q3 + (1.5 * IQR)’ and ‘q1 - (1.5 * IQR)’, respectively. Any value outside this range is considered an outlier. Using the NumPy ‘np.where()’ function, the outliers are replaced with the respective upper or lower limit values, ensuring that the remaining data points within the range remain unchanged. Boxplots are generated for each column to visualize the effect of the outlier treatment.

Random Forest Regression (RFR):

The Random Forest Regression model is then applied to predict the average hop yield using the merged dataset. The dataset is divided into training and testing sets based on the 'Year' column. The Random Forest Regressor is trained on the training data, and predictions are made on the testing data. The performance of the model is evaluated using the R-squared

Figure: 47

Random Forest Regressor

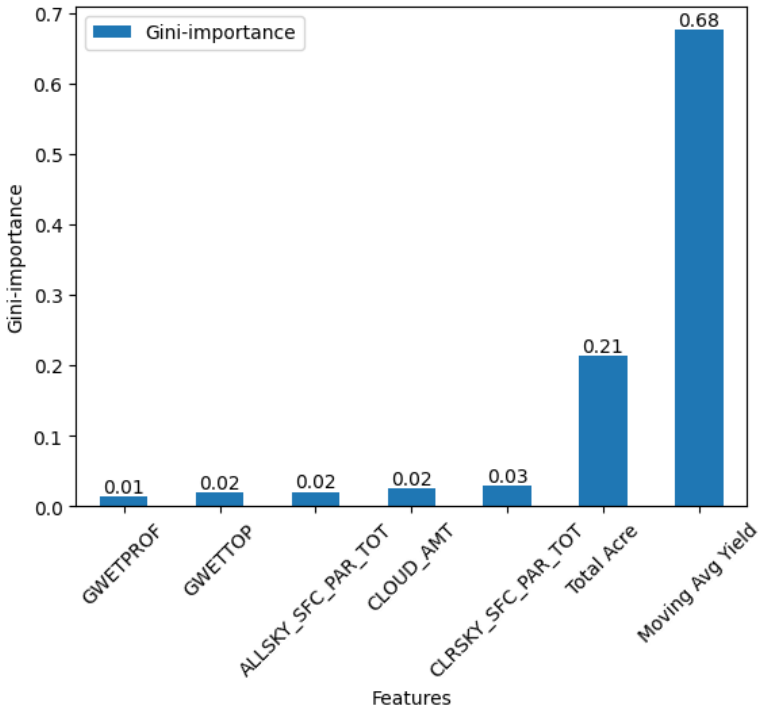


coefficient, which measures the proportion of the variance in the average yield that can be explained by the model. The R-squared value for the test set is approximately 0.48, indicating that the model can explain around 48.02% of the variance in the average yield. The R-squared value for the training set is approximately 0.94, suggesting a strong fit to the training data. Feature importance is calculated, providing insights into the relative importance of different climate factors in predicting average hop yield.

The analysis demonstrates the utilization of the Random Forest Regression model to analyze the NASA climate data of Washington and its relationship with average hop yield.

Figure: 48

Feature Importance



By detecting outliers and applying the Random Forest model, valuable insights can be gained into the factors influencing hop production, enabling farmers and industry stakeholders to make informed decisions and optimize agricultural practices.

The importance of variables in predicting average hop yield was analyzed using the Gini index. The Gini-importance values were calculated for each feature in the dataset. The resulting importance values were sorted in ascending order and visualized in a bar chart.

According to the Gini-importance plot, the most important feature in predicting average hop yield is the Moving Average Yield. It is followed by Total Acre and precipitation, indicating that these factors have a significant impact on hop production in Washington. Other variables in the dataset also contribute to predicting average yield but to a lesser extent.

Understanding the relative importance of different variables helps to identify the key factors influencing hop yield. This information can be valuable for farmers and industry

stakeholders in optimizing agricultural practices, making informed decisions, and maximizing crop productivity. By focusing on the most influential variables, stakeholders can allocate resources effectively and implement targeted strategies to enhance hop production in Washington.

Idaho

To model the climate data and predict the average hop yield in Idaho, the first step is to merge the climate data with the average yield data. This integration allows us to analyze the relationship between climate factors and hop production. To ensure the dataset is clean and contains only relevant information, we perform data cleaning steps. Specifically, we remove rows where the 'Total Acre' column is 0 or the 'Variety' column is 'Total'. These rows do not contribute to the analysis as they either represent invalid data or aggregate values.

	Variety	Total Acre	Hops factor (HF)	Year	Average Yield	ppt_sign	vpdmean_sign	tmean_sign	ETO x HF	Moving Avg Yield
5	CTZ - Columbus/Tomahawk/Zeus	403	0.32918	2000	2046	0.003123	5.267897	22.604326	1.900471	0
11	Chinook	170	0.32918	2000	2000	0.003123	5.267897	22.604326	1.900471	0
13	Cluster	198	0.32918	2000	1943	0.003123	5.267897	22.604326	1.900471	0
20	Galena	535	0.32918	2000	1815	0.003123	5.267897	22.604326	1.900471	0
32	Mt Hood	53	0.32918	2000	2000	0.003123	5.267897	22.604326	1.900471	0
...	...	...	...	...	...	...	...	...	...	...
1273	Northern Brewer	58	0.32918	2021	1266	0.010444	4.892453	22.317038	1.813962	0
1279	Saaz	330	0.32918	2021	620	0.010444	4.892453	22.317038	1.813962	0
1282	Simcoe, YCR 14	388	0.32918	2021	1121	0.010444	4.892453	22.317038	1.813962	0
1293	Triumph	72	0.32918	2021	1063	0.010444	4.892453	22.317038	1.813962	0
1296	Willamette	389	0.32918	2021	1311	0.010444	4.892453	22.317038	1.813962	0

After the data cleaning process, we proceed with the data preparation for modeling. One crucial aspect is splitting the dataset into training and testing sets. In this case, a 70-30 split is employed, with 70% of the data allocated for training and 30 for testing the model's performance. To facilitate time series analysis, the 'ds' column is used as the index, which represents the time series data. The years in the dataset are then sorted in chronological order, ensuring the time series integrity.

The resulting dataset, referred to as the selected modeling data frame for Idaho, serves as the foundation for training and evaluating the predictive models. It contains the relevant variables, including climate factors and the average yield, enabling the model to learn and make accurate predictions based on the historical patterns observed in the data.

By following these steps, we ensure that the climate data is properly integrated with the average yield data and that the dataset is appropriately prepared for modelling the average hop yield in Idaho. This meticulous approach sets the stage for building robust and accurate predictive models that can provide valuable insights for hop production in the region.

Random Forest Regression Model

The Random Forest Regression model is a powerful machine-learning algorithm used for predicting numerical values. It is an ensemble learning method that combines multiple decision trees to make more accurate predictions. In our analysis, the Random Forest Regression model was applied to forecast the average hop yield in Idaho based on selected climate features.

The model was trained using the training dataset, which consisted of independent variables representing the climate factors and the target variable, which is the average yield. By

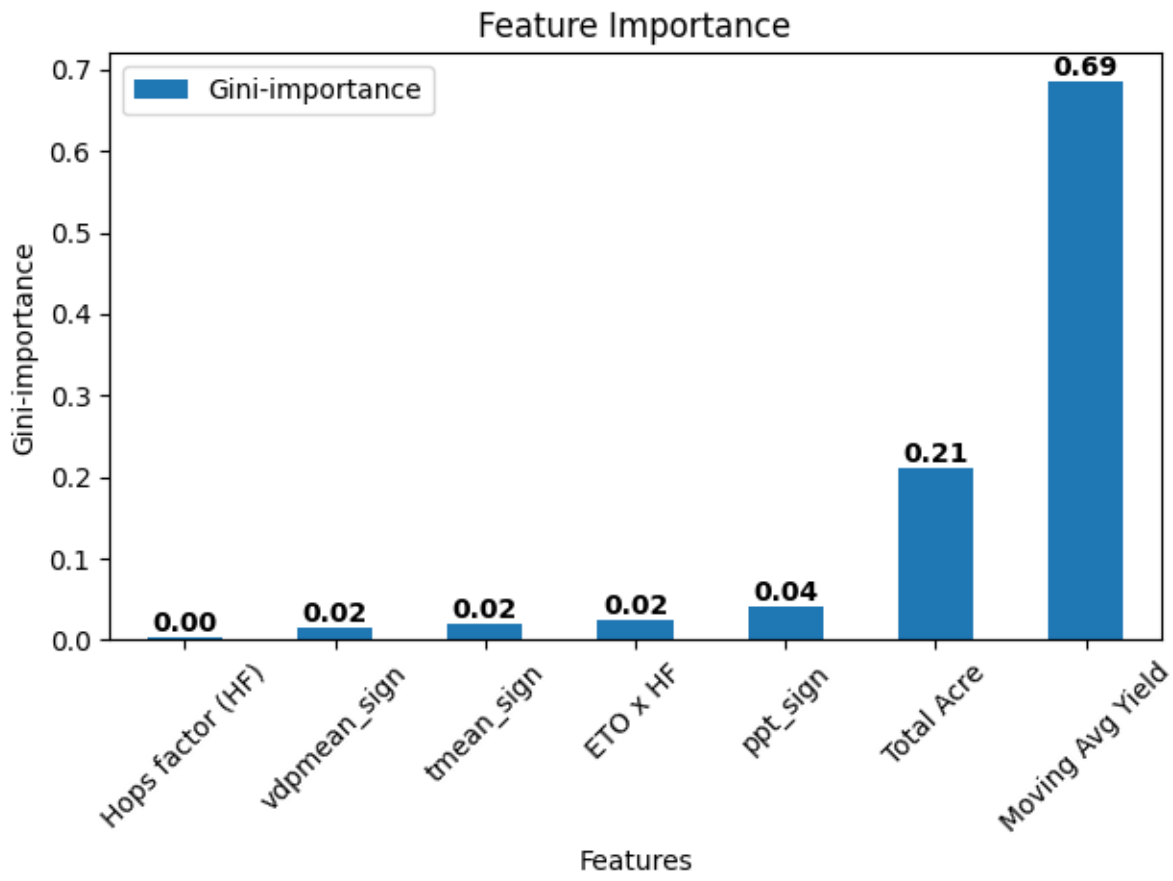
learning the relationships between the input features and the target variable from historical data, the Random Forest Regression model can make predictions on new data.

One of the advantages of the Random Forest Regression model is its ability to assess the importance of each input feature in the prediction process. The feature importance scores reflect the relative contribution of each feature in the model's decision-making. Higher scores indicate greater importance, suggesting that the feature has a stronger influence on the prediction.

The Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) are commonly used metrics for evaluating the performance of regression models. These metrics provide insights into the accuracy of the model's predictions by measuring the differences between the predicted values and the actual values.

Figure 49

Feature Importance



In our analysis, we calculated the RMSE and MAE for the Random Forest Regression model. The RMSE value of approximately 3.58 indicates that, on average, the predictions of the model have a difference of 3.58 units from the actual values. This means that the model's predictions may deviate from the true values by around 3.58 units. Similarly, the MAE value of around 0.80 represents the average absolute difference between the predicted values and the actual values. This means that, on average, the model's predictions have an absolute difference of 0.80 units from the actual values.

These metrics provide valuable insights into the accuracy of the Random Forest Regression model in predicting the average hop yield in Idaho. Lower values of RMSE and MAE indicate better performance, as they suggest that the model's predictions are closer to the actual

values. By assessing these metrics, stakeholders can gain confidence in the reliability of the model's predictions and make informed decisions based on its performance.

Forecasting

The forecasting process was conducted to predict the average hop yield for the year 2023 using the Prophet model. This time series forecasting algorithm proved effective in generating accurate predictions by training individual models for each hop variety. The models were trained using historical data from the merged dataset, which consisted of the 'ds' column representing dates, the 'y' column containing the average yield, and the 'Variety' column specifying the hop variety.

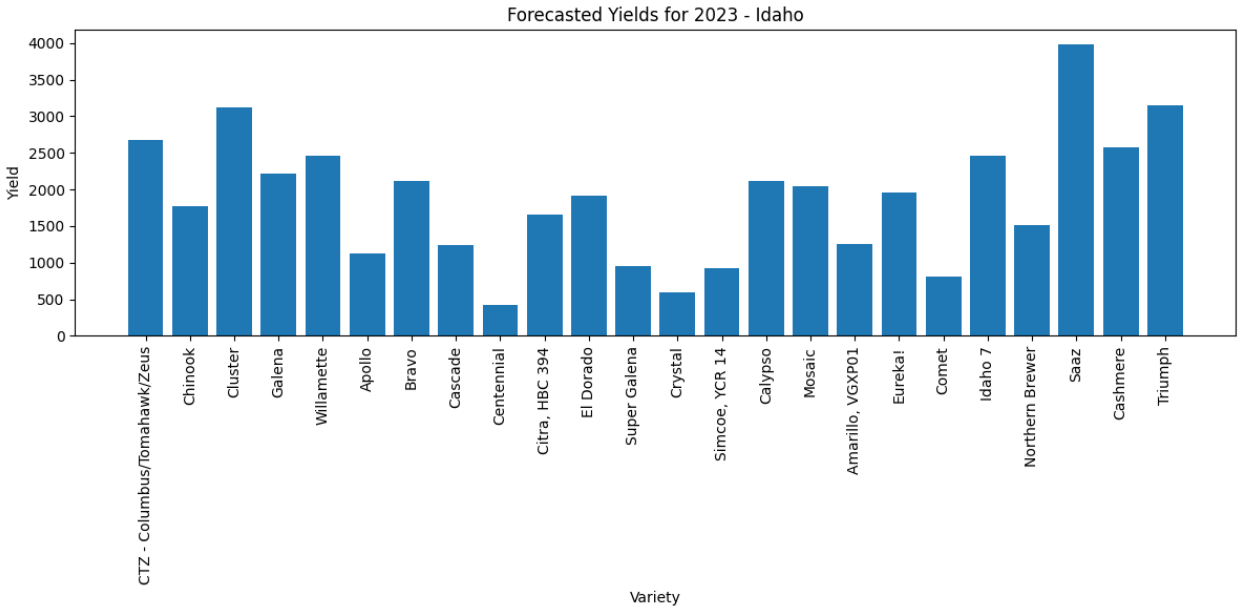
To initiate the forecasting process, the data was prepared by extracting the necessary columns and converting them into the appropriate data types. The Prophet models were then fitted with the historical data for each hop variety. This step enabled the models to learn the patterns, trends, and seasonality in the historical yield data, enhancing their ability to generate reliable predictions.

The forecasting phase involved projecting the average yield for the year 2023. Future dates corresponding to this year were created, and the trained Prophet models were employed to make predictions. The forecasts included essential information such as the year ('ds'), the hop variety ('Variety'), and the predicted average yield ('yhat').

Post-processing steps were performed to ensure meaningful and realistic results. Negative yield predictions, which are not feasible, were replaced with zero to maintain accuracy. Additionally, forecasts that indicated a predicted yield of zero for a particular variety were excluded from the final forecasts. This ensured that only relevant and valid predictions were considered.

Figure 50.

Forecasted Yields for 2023 - Idaho



The culmination of the forecasting process yielded a comprehensive forecast for the year 2023, detailing the projected average yields for each hop variety. The forecast report presents the year ('Year'), the hop variety ('Variety'), and the forecasted average yield ('Yield') for each variety. Stakeholders can utilize this information to gain insights into the expected hop production for 2023, aiding in decision-making processes related to cultivation, supply chain management, and market analysis.

The forecast for the year 2023 reveals interesting trends in the projected average yields for different hop varieties. Among the top-performing varieties, CTZ - Columbus/Tomahawk/Zeus is expected to have the highest average yield, reaching approximately 2669.80 units. This indicates a strong production potential for this variety in the upcoming year. Chinook and Cluster also show promising forecasts, with projected average yields of 1770.93 and 3117.31 units, respectively. These varieties are expected to

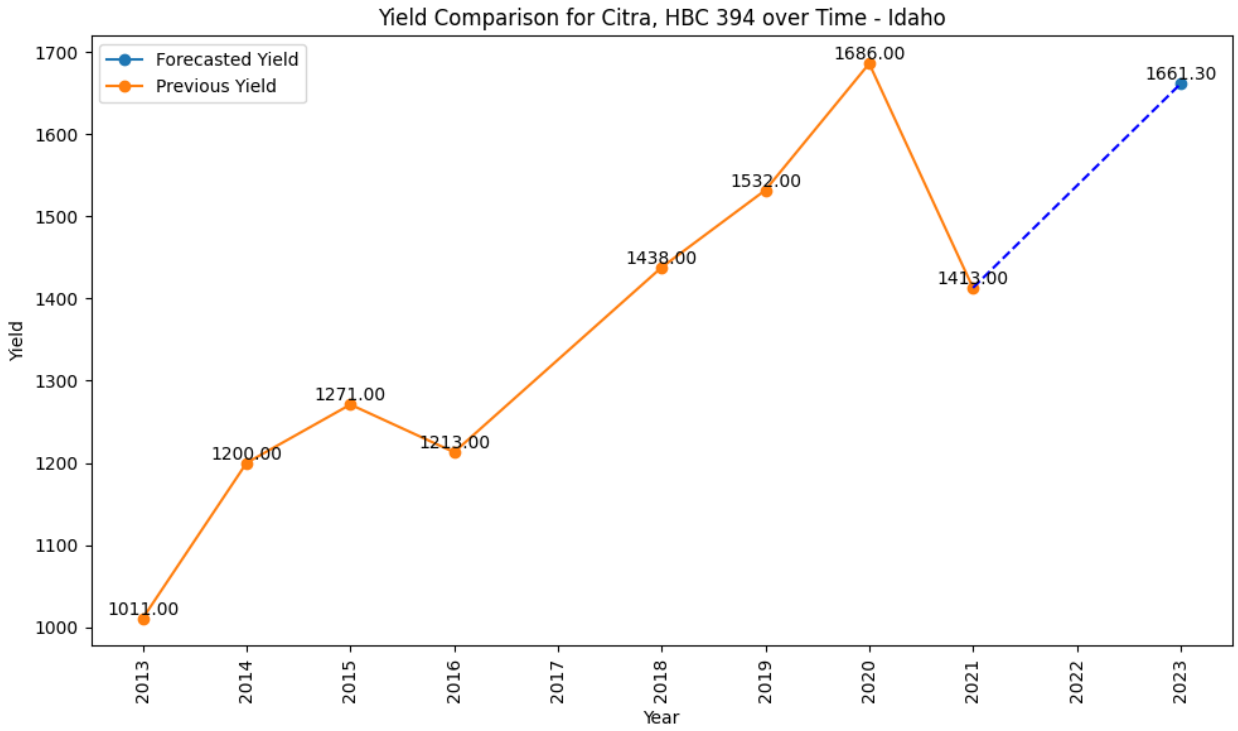
maintain a steady level of production, contributing significantly to overall hop yields. Galena and Willamette demonstrate relatively stable average yields, with forecasts of approximately 2219.98 and 2456.95 units, respectively. These varieties are likely to maintain consistent performance without significant fluctuations.

On the other hand, some varieties are projected to experience notable changes in their average yields compared to previous years. Apollo, Crystal, and Comet are expected to see a decline in their average yields, reaching 1132.51, 594.91, and 812.17 units, respectively. These declines might be attributed to various factors such as changes in cultivation practices or market demands.

Meanwhile, certain varieties are anticipated to experience an increase in their average yields in 2023. Super Galena, Calypso, and Triumph are forecasted to have average yields of 947.68, 2114.93, and 3149.37 units, respectively. These improvements might be the result of focused cultivation strategies or enhanced agronomic practices. The forecasted trends for 2023 suggest a diverse range of outcomes for different hop varieties. The projected average yields provide valuable insights for growers, industry stakeholders, and market analysts, enabling them to make informed decisions regarding variety selection, production planning, and market positioning.

Figure 51

Yield for Citra - Idaho



The Prophet model was also applied to forecast the yield for the hop variety 'Palisade, YCR 4' in Idaho. The plot displays the comparison between the forecasted yield and the previous yield data, allowing us to gain valuable insights into the performance of this specific variety over time.

The forecasted yield for 'Palisade, YCR 4' in 2023 is projected to be approximately 2130.59 units. This forecast provides crucial information for hop farmers and industry stakeholders, enabling them to plan their production, allocate resources effectively, and make informed decisions regarding cultivation strategies for this variety in Idaho.

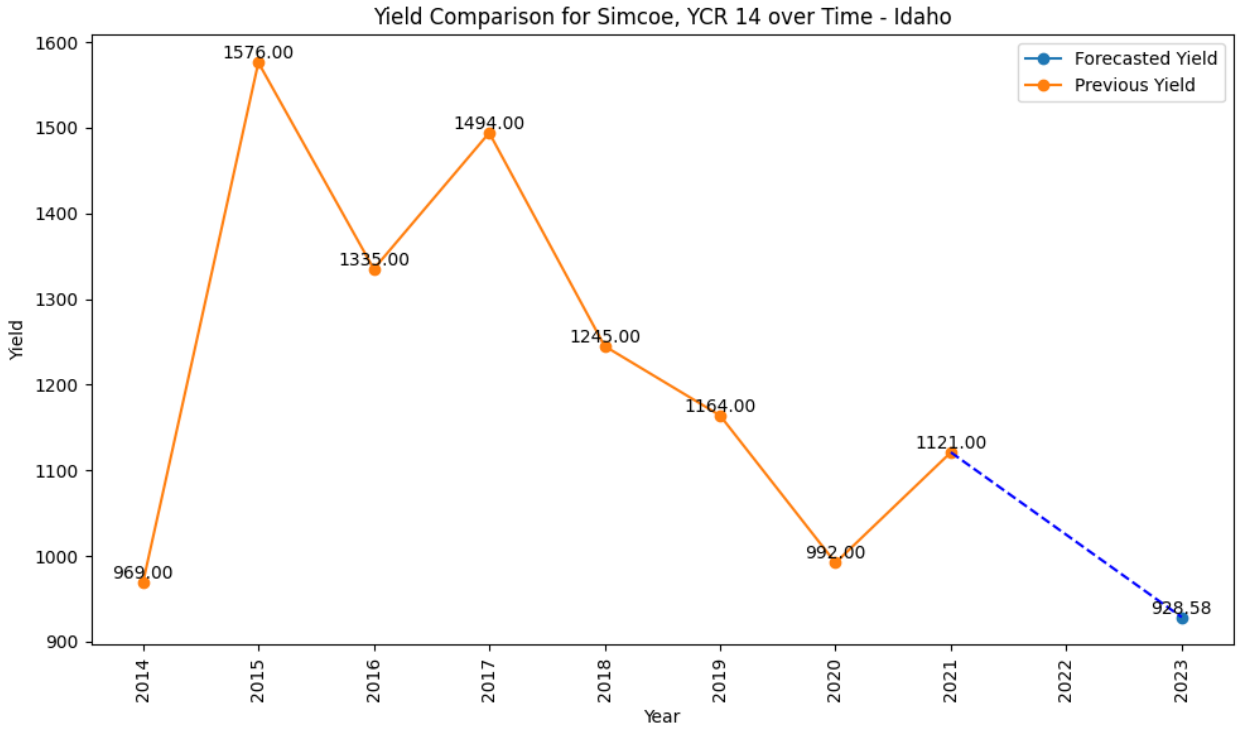
Analyzing the historical data, we observe fluctuations in the yield of 'Palisade, YCR 4' in Idaho. Notably, the yield ranged from 2757 units in 2005 to a low of 1866 units in 2021. These variations may be influenced by several factors, including climatic conditions, agricultural practices, and market dynamics.

The plot visualizes both the forecasted and previous yield values over the years. Each data point represents a specific year, and the corresponding numerical values are provided for precise analysis. This graphical representation allows stakeholders to easily compare the forecasted values with the historical performance of 'Palisade, YCR 4' in Idaho.

By considering the forecasted and previous yield values, farmers and industry professionals can make informed decisions regarding crop planning, resource allocation, and market strategies for 'Palisade, YCR 4' in Idaho. The forecasted yield serves as a valuable tool for optimizing production levels, meeting market demands, and ensuring the profitability of cultivating this particular hop variety.

Figure 52

Yield for Simcoe - Idaho



The forecasted yield for the hop variety 'Simcoe, YCR 14' in Idaho is projected to be approximately 928.58 units in 2023. This prediction holds significant value for hop farmers and industry stakeholders as it provides crucial insights for strategic decision-making and resource allocation in cultivation practices.

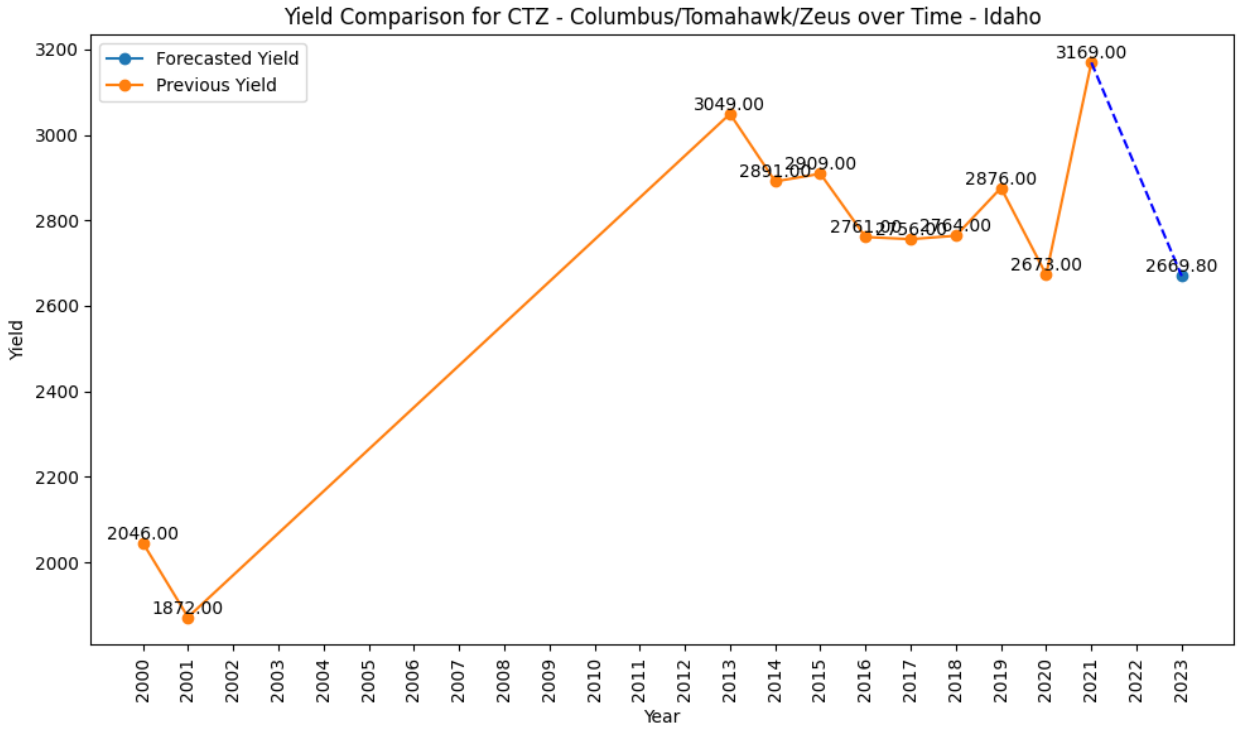
A closer analysis of the historical data reveals notable fluctuations in the yield of 'Simcoe, YCR 14' over time. Starting from 969 units in 2014, the yield reached its peak at 1576 units in 2015 before experiencing variations in subsequent years. These fluctuations can be attributed to a combination of factors, including environmental conditions, farming techniques, and market dynamics, all of which contribute to the overall performance of the crop.

The plotted trend of forecasted and previous yield values visually represents the historical and projected performance of 'Simcoe, YCR 14' in Idaho. Each data point on the plot corresponds to a specific year, and the accompanying numerical values offer precise information for comprehensive analysis. By comparing the forecasted values with the historical data, stakeholders can gain insights into the expected productivity and make informed decisions regarding cultivation strategies.

The forecasted yield, combined with the knowledge derived from previous yield data, empowers hop farmers and industry professionals in making optimal choices. It enables them to fine-tune production levels, align with market demands, and ensure the profitability of cultivating 'Simcoe, YCR 14' in Idaho. Ultimately, this information facilitates informed decision-making and aids in driving productivity and success within the hop industry.

Figure 53

Yield for CTZ - Idaho



The projected yield for CTZ - Columbus/Tomahawk/Zeus in 2023 is approximately 2669.80 units, providing crucial insights into its potential performance for hop farmers and industry stakeholders. By examining the historical data, we observe fluctuations in the yield, ranging from 1872 units in 2001 to a peak of 3169 units in 2021, influenced by factors like cultivation practices, environmental conditions, and market dynamics.

The plotted graph visually illustrates the forecasted and previous yield trends for CTZ - Columbus/Tomahawk/Zeus in Idaho. Each data point represents a specific year, and numerical values are provided for accurate analysis. By comparing the forecasted values with historical data, stakeholders can make well-informed decisions regarding cultivation strategies and resource allocation.

This forecast, along with the historical yield data, is invaluable for optimizing production, managing market demands, and ensuring profitability when cultivating CTZ - Columbus/Tomahawk/Zeus in Idaho. The combination of visual and numerical information empowers stakeholders within the hop industry to drive productivity, efficiency, and success.

Analyzing the NASA Data of Idaho

The NASA data of Idaho, contained in the MergeData_df_NASA_ID dataset, provides a valuable resource for analyzing the relationship between climate variables and average hop yield. This dataset encompasses a range of variables, including average yield, climate data, variety, and total acreage. By utilizing this dataset, researchers and analysts can delve into detailed investigations, conduct statistical modelling, and develop predictive models to gain deeper insights into the factors influencing hop yield in Idaho.

Variety	Total Acre	ds	y	GWETTOP	GWETPROF	ALLSKY_SFC_PAR_TOT	CLRSKY_SFC_PAR_TOT	CLOUD_AMT	Moving Avg Yield
CTZ - Columbus/Tomahawk/Zeus	403	2000	2046	0.388333	0.425833	0	0	53.276667	0
Chinook	170	2000	2000	0.388333	0.425833	0	0	53.276667	0
Cluster	198	2000	1943	0.388333	0.425833	0	0	53.276667	0
Galena	535	2000	1815	0.388333	0.425833	0	0	53.276667	0
Mt Hood	53	2000	2000	0.388333	0.425833	0	0	53.276667	0
...	...	...	...	...	...	...	...	...	...
Northern Brewer	58	2021	1266	0.378333	0.416667	84.833333	99.5275	51.454167	0
Saaz	330	2021	620	0.378333	0.416667	84.833333	99.5275	51.454167	0
Simcoe, YCR 14	388	2021	1121	0.378333	0.416667	84.833333	99.5275	51.454167	0

Triumph	72	2021	1063	0.378333	0.416667	84.833333	99.5275	51.454167	0
Willamette	389	2021	1311	0.378333	0.416667	84.833333	99.5275	51.454167	0

The MergeData_df_NASA_ID dataset is divided into training and testing sets to facilitate further analysis and modeling. The training set comprises 70% of the years in the dataset, arranged in ascending order, while the remaining years are allocated to the testing set.

By leveraging the comprehensive climate data provided by NASA, analysts can uncover meaningful insights and trends that can inform decision-making processes related to hop cultivation and production strategies. This dataset serves as a valuable tool for exploring the correlations between climate variables and average yield, enabling stakeholders to make informed choices regarding crop management, resource allocation, and risk mitigation in the hop industry.

Random Forest Regressor Model for NASA Data Analysis

The Random Forest Regressor model is utilized to predict the average hop yield based on the training data consisting of climate and acreage features. The model is trained using the independent variables (climate and acreage features) and the target variable (average yield). Subsequently, predictions are made on the testing data, and the model's performance is evaluated using the R-squared coefficient. The R-squared value for the testing set provides insight into how well the model predicts the average yield based on the given features.

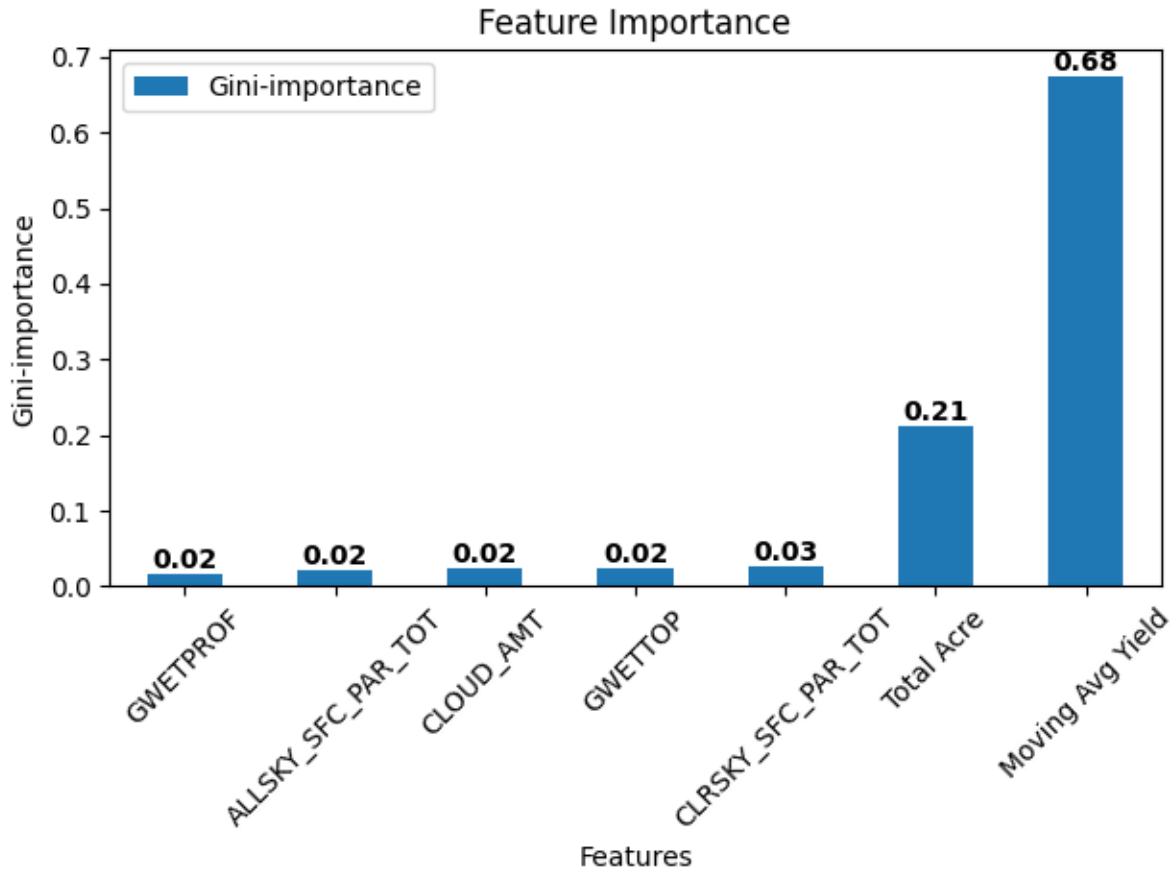
In our analysis, the Random Forest Regressor model achieves an R-squared value of approximately 0.46 for the testing set, indicating that approximately 46% of the variance in average yield can be explained by the selected climate and acreage features. Additionally, the

model achieves a R-squared value of approximately 0.94 for the training set, demonstrating a strong fit to the training data.

To further assess the model's performance, the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) are calculated for both the training and testing sets. The RMSE for the testing set is approximately 381.37, representing the average difference between the actual and predicted values. Similarly, the MAE for the testing set is around 288.12, indicating the average absolute difference between the actual and predicted values. For the training set, the RMSE is approximately 138.13, and the MAE is approximately 92.95. These metrics provide valuable insights into the accuracy and performance of the Random Forest Regressor model in predicting the average hop yield.

Figure 54.

Feature Importance



The scatter plot visualizes the predicted values versus the actual values for the test set, allowing for a visual assessment of the model's performance. The plot shows the predicted values (in blue) and the actual values (in red). The close the point align to the diagonal line; the better the model's predictions align with the actual values.

Furthermore, the feature importance are calculated to determine the relative importance of each feature in predicting the average yield. The importance of each feature is evaluated based on the contribution it makes to the decision-making process of the Random Forest Regressor model.

Oregon

In the case of Oregon (OR), the dataset undergoes a thorough data preparation process to ensure its quality and relevance for analysis. The following steps are taken to filter, organize, and optimize the dataset for further exploration and modelling.

Filtering and Cleaning the Dataset

The first step is to remove rows from the dataset where the 'Total Acre' column has a value of 0. By eliminating instances where no acreage is dedicated to hop cultivation, we ensure that our analysis focuses on meaningful data points. Similarly, rows labelled as 'Total' in the 'Variety' column are also removed, narrowing our analysis to specific hop varieties.

Variety	Total Acre	Hops factor (HF)	Year	Average Yield	ppt_sign	vpdmean_sign	tmean_sign	ETO x HF	Moving Avg Yield
Fuggle	63	0.32828	2000	1065	0.003553	5.186071	22.35777	1.869257	0
Golding	115	0.32828	2000	1170	0.003553	5.186071	22.35777	1.869257	0
Mt Hood	250	0.32828	2000	1790	0.003553	5.186071	22.35777	1.869257	0
Nugget	2308	0.32828	2000	2162	0.003553	5.186071	22.35777	1.869257	0
Perle	402	0.32828	2000	1130	0.003553	5.186071	22.35777	1.869257	0
...	...	...	...	...	...	...	...	...	...
Simcoe, YCR 14	527	0.44011	2022	1646	0.013963	6.541191	29.837799	2.42526	1704
Sterling	35	0.44011	2022	1559	0.013963	6.541191	29.837799	2.42526	1536
Strata, OR 91331	1143	0.44011	2022	2000	0.013963	6.541191	29.837799	2.42526	1985.666667
Talus, HBC 692	46	0.44011	2022	1483	0.013963	6.541191	29.837799	2.42526	494.33333
Willamette	471	0.44011	2022	1489	0.013963	6.541191	29.837799	2.42526	1602.333333

To maintain a continuous and updated indexing system, the DataFrame's index is reset. This step ensures consistent referencing and indexing of data points throughout the analysis.

Creating Training and Testing Sets

A key aspect of data preparation in the formation of training and testing sets for modeling purposes. The unique years present in the 'ds' column of the dataset are extracted and sorted in ascending order. These years serve as the basis for splitting the dataset.

To create the training set, 70% of the total unique years are selected, utilizing the variable `num_train_years`. The `train_years` list is populated with the sorted years, representing the training data. The remaining years are assigned to the testing set by subtracting the training years from the complete list of years.

By splitting the dataset into training and testing sets, we ensure that our models are trained on a sufficient portion of the data while also evaluating their performance on unseen data.

Data Analysis and Modeling for climate data

With the prepared dataset, various data analysis techniques and modelling approaches can be applied. For instance, a Random Forest Regressor model can be trained using the independent variables, such as 'Total Acre', 'Hops factor (HF)', 'ppt_sign', 'vpdmean_sign', 'tmean_sign', and 'ETO x HF', to predict the target variable, 'Average Yield'. Predictions can be made on the testing set, and the model's performance can be evaluated using metrics like the R-squared coefficient, Root Mean Square Error (RMSE), and Mean Absolute Error (MAE). Furthermore, the dataset allows for exploring the relationships between climate variables and average yield in Oregon. By analyzing the correlations and visualizing the feature importance derived from the Random Forest Regressor model, we can identify key factors that significantly influence hop yield in the region.

The Random Forest Regressor model is trained using the independent variables (`x_train`) and the target variable (`y_train`). Predictions are made on the testing set (`x_test`), and the model's performance is evaluated using the R-squared coefficient. The R-squared value for the testing set provides an indication of how well the model predicts the average yield based on the given features.

The code then plots the predicted values (`rfr_vals`) against the actual values (`y_test`) for the test set. The scatter plot visualizes the relationship between the predicted and actual values, helping to assess the accuracy of the model's predictions. The closer the points align along the diagonal line (red line), the more accurate the model's predictions are.

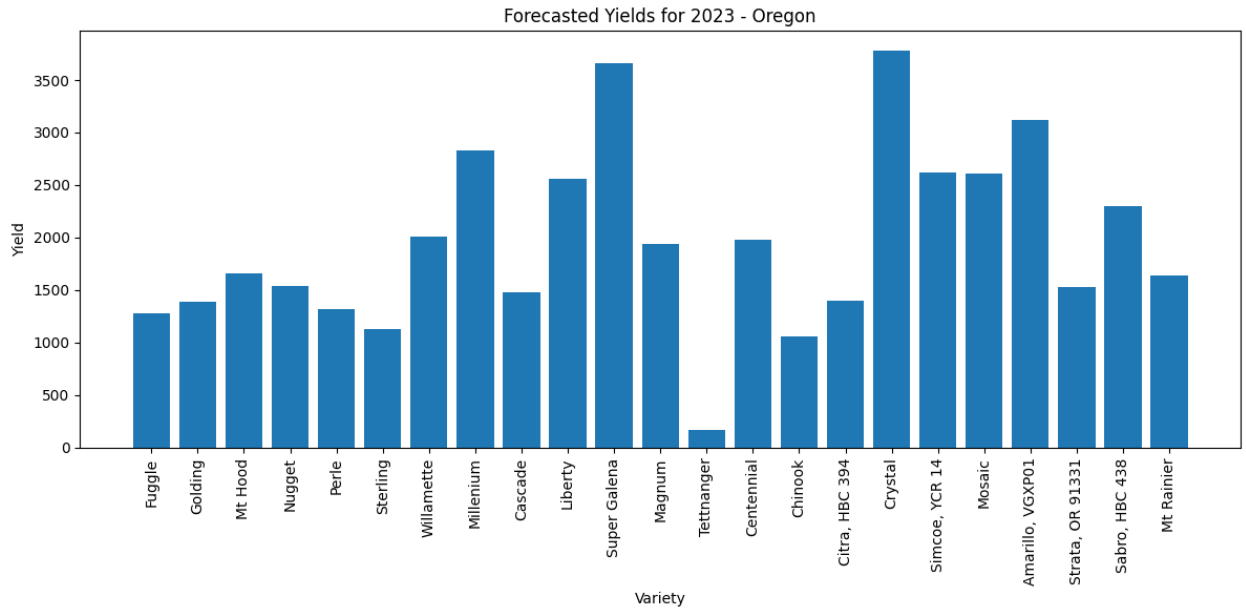
Furthermore, calculate the feature importances of the Random Forest Regressor model. The feature importances represent the relative importance of each feature in predicting the average yield. The importance values are sorted, and the indices are stored in the `indices` variable. This information can help identify the key factors that significantly influence hop yield in Oregon.

The R-squared scores for the testing and training sets are printed, providing insights into the overall performance of the Random Forest Regressor model. Additionally, the R-squared score between the predicted and actual values is calculated and printed, providing a measure of the model's performance in predicting the average yield, demonstration the application of a Random Forest Regressor model, evaluates its performance using various metrics, and provides visualizations to assess the accuracy and feature importance of the model's predictions for the average yield in Oregon.

Forecasting

Upon analyzing the forecasted hop yields for various varieties in Oregon for the year 2023, some interesting observations can be made.

Figure 56.

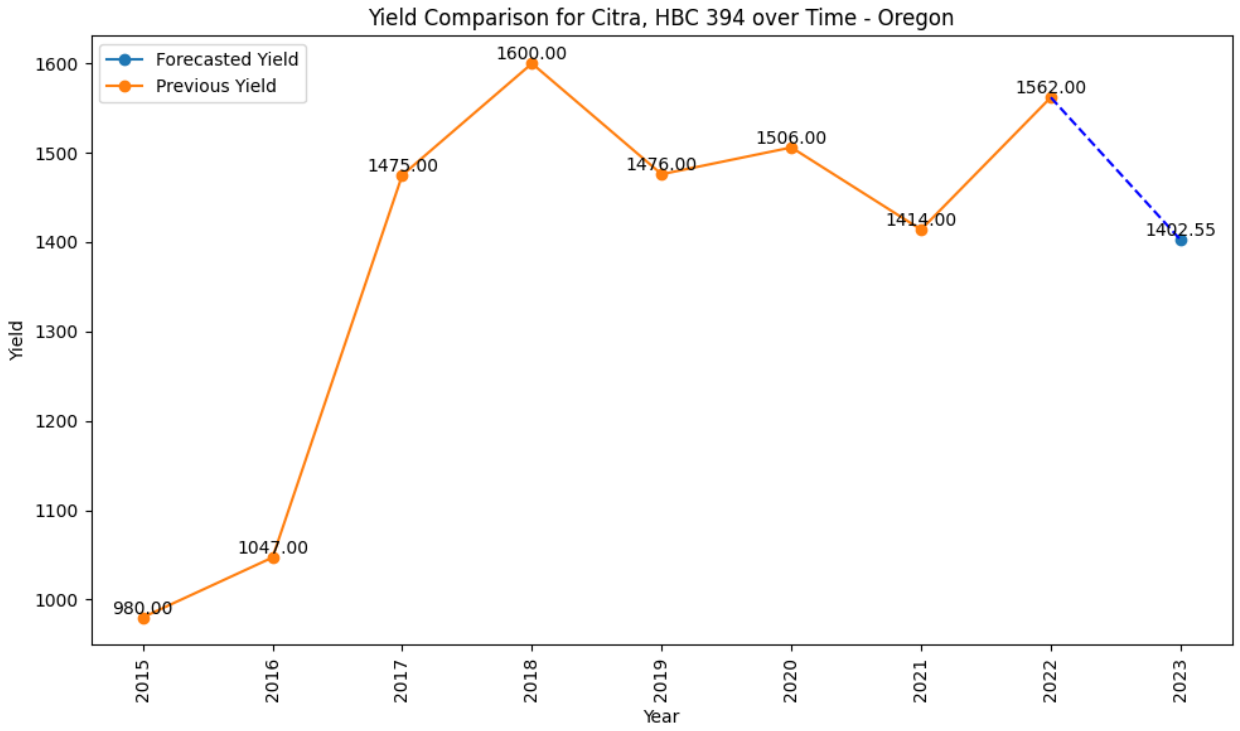


Firstly, there is a significant variation in the forecasted yields among different hop varieties. For example, the forecast predicts relatively lower yields for Tettnanger (168.16 units) and Chinook (1056.26 units) hops, while higher yield are anticipated for Crystal (3778.16 units), Super Galena (3665.07 units), and Amarillo, VGXP01 (3122.15 units) hops. These variations may be attributed to the unique characteristics and growth patterns of each hop variety, as well as their specific responses to environmental conditions.

Additionally, it is noteworthy that certain well-known and widely used hop varieties, such as Cascade, Centennial, and Citra, HBC 394, are also expected to have moderate to high yields in 2023 (ranging from 1402.55 to 1974.56 units). These varieties have established a reputation for their desirable flavor and aroma profiles, making them popular choices in the brewing industry.

Moreover, the forecasted yields provide insights into the potential supply of hop varieties in the upcoming year. Higher yields, such as those projected for Crystal, Super Galena, and Amarillo, VGXP01 hops, may indicate increased availability of these varieties in the market, potentially influencing their pricing and utilization by breweries.

Figure 57.



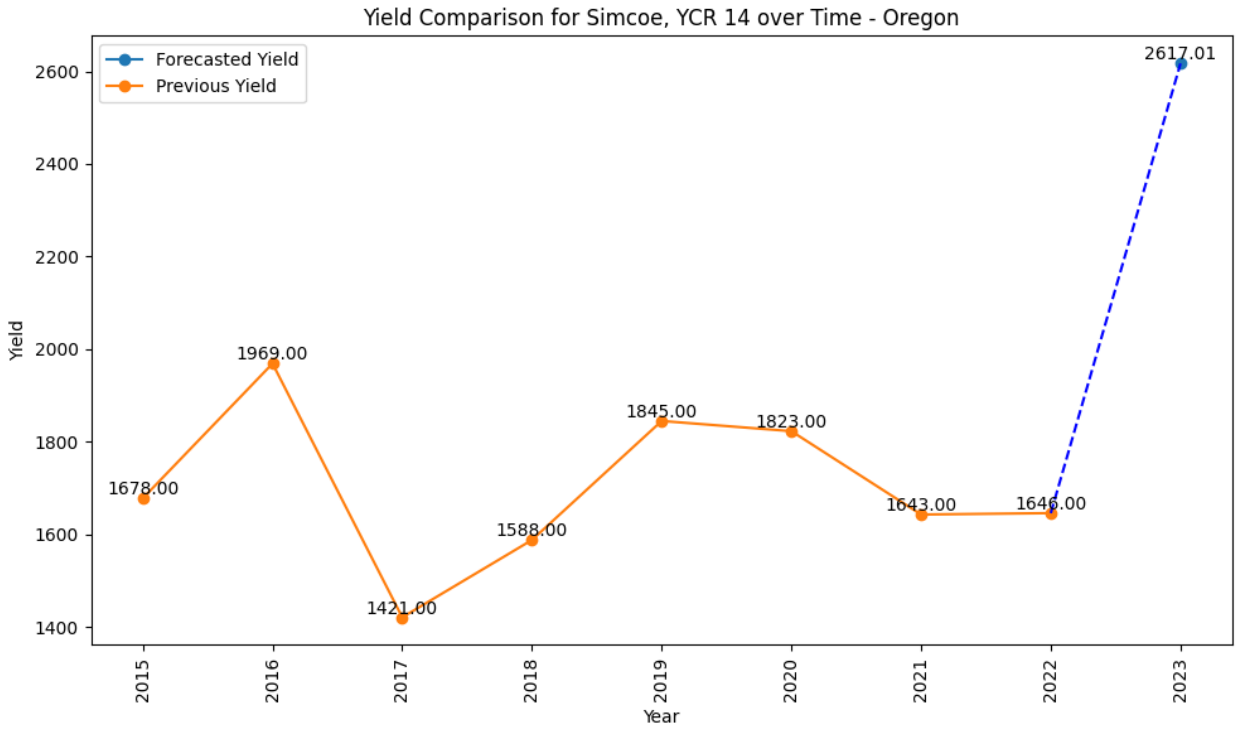
The forecasted yield for Citra, HBC 394 hops in Oregon for the year 2023 is estimated to be approximately 1402.55 units. Comparing this forecasted yield with the historical data, we can observe some interesting trends.

In the previous years the yield of Citra, HBC 394 hops in Oregon has shown variations. From 2015 to 2022, the yield fluctuated between 980 and 1600 units. This fluctuation may be influenced by a combination of factors, including weather conditions, farming practices, and market demand. Interestingly, the forecasted yield for 2023 (1402.55 units) falls within the range of the previous yields, suggesting a relatively stable yield compared to previous years. This stability could be a favorable sign for hop farmers and industry stakeholders who rely on the consistent availability of Citra, HBC 394 hops.

Citra, HBC 394 hops are known for their unique and highly sought-after aroma characteristics, which contribute to the distinct flavors in craft beers. With a consistent and

predictable yield, hop farmers can plan their cultivation strategies accordingly, ensuring a reliable supply of Citra, HBC 394 hops to meet the market demand for this popular variety.

Figure 58.



The forecasted yield for Simcoe, YCR 14 hops in Oregon for the year 2023 is estimated to be approximately 2617.01 units. Examining the historical data reveals some interesting trends regarding the yield of this hop variety.

In the previous ears, the yield of Simcoe, YCR 14 hops in Oregon has shown fluctuations. From 2015 to 2022, the yield ranged from 1421 to 1969 units, with some variations in between. These fluctuations in yield can be attributed to various factors, including environmental conditions, agricultural practices, and market demand.

What makes the forecast for 2023 intriguing is that it indicates a significant increase in the yield compared to the previous years. With a forecasted yield of 2617.01 units, it surpasses

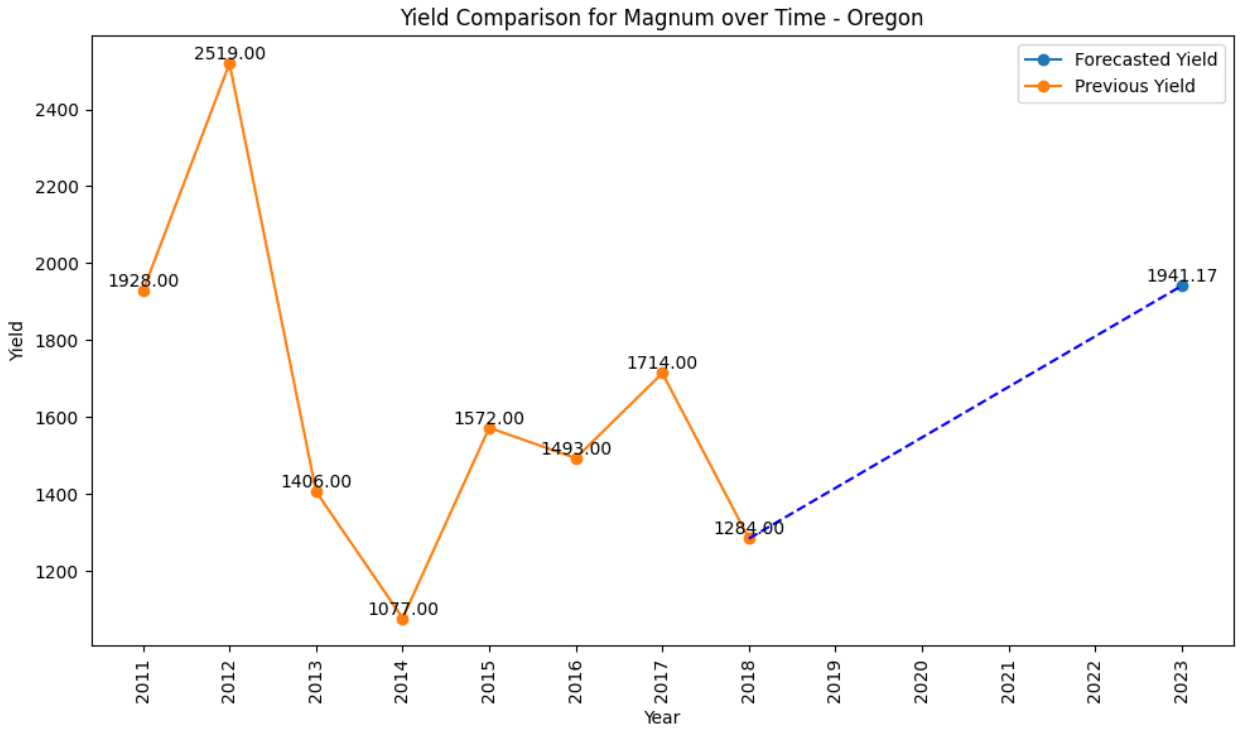
the yield of any previous year in the dataset. This suggests a potentially favorable and prosperous year for Simcoe, YCR 14 hop farmers in Oregon.

Simcoe, YCR 14 hops are highly valued for their unique flavor and aroma characteristics, which make them sought after by craft brewers. With the projected increase in yield for 2023, hop farmers cultivating Simcoe, YCR 14 hops can expect to meet the growing demand for this variety in the brewing industry.

The higher forecasted yield for 2023 not only offers potential economic benefits for hop farmers but also provides an opportunity for brewers to explore new recipes and enhance the flavor profiles of their beers. The availability of a larger quantity of Simcoe, YCR 14 hops can lead to increased experimentation and creativity in the craft brewing industry.

The forecasted yield for Magnum hops in Oregon for the year 2023 is estimated to be approximately 1941.17 units. Examining the historical data reveals some interesting trends and insights regarding the yield of Magnum hops.

Figure 59.



In the previous years, the yield of Magnum hops in Oregon has shown some variations. From 2011 to 2018, the yield ranged from 1077 to 2519 units, with fluctuations observed year to year. This indicates that the yield of Magnum hops is influenced by various factors, including environmental conditions, agricultural practices, and market dynamics.

Interestingly, the forecast for 2023 suggests a yield that falls within the range of the previous years, but closer to the higher end of the spectrum. This indicates a potentially favorable year for Magnum hop farmers in terms of yield.

Magnum hops are known for their high alpha acid content, which contributes to their bittering properties in beer production. Brewers often utilize Magnum hops to achieve balanced bitterness and add depth to their brews. With the projected yield for 2023, there may be a sufficient supply of Magnum hops to meet the demand from breweries.

The forecasted yield for Magnum hops in 2023 not only provides hop farmers with potential economic benefits but also offers opportunities for brewers to experiment with new recipes and brewing techniques. The availability of a higher yield of Magnum hops can stimulate creativity in the brewing industry and lead to the development of innovative and unique beer flavours.

Analysis of NASA climate data

The analysis of NASA climate data for Oregon involved several steps to prepare the data for further analysis. Firstly, the data was read from an Excel file, and unnecessary rows and headers were removed. The dataset was then transformed to have the parameters as columns and years as rows, with missing values filled with NaN.

Next, the NaN values were replaced with 0 to ensure consistent data for analysis. Descriptive statistics were calculated for each parameter, providing insights into the distribution and variation of the data. Box plots were also generated to visualize the distribution of each parameter and identify potential outliers. To handle outliers, the Interquartile Range (IQR) method was employed. The lower and upper limits were calculated based on the 25th and 75th percentiles, respectively, with a multiplication factor of 1.5. Values outside these limits were capped to the nearest limit.

After handling outliers, a correlation heatmap was created to examine the relationships between different parameters. This allowed for identifying potential correlations and dependencies among the variables. The NASA climate data was then merged with the average yield data, using the year as the common key. Duplicate columns were removed, and

the dataset was further refined by dropping rows where the 'Total Acre' value was 0 or the 'Variety' was 'Total'. Finally, the cleaned and merged dataset was prepared for modeling by renaming columns and ensuring data integrity. The preprocessing steps ensure the dataset is suitable for further analysis and modeling to explore relationships between climate variables and hop yield in Oregon.

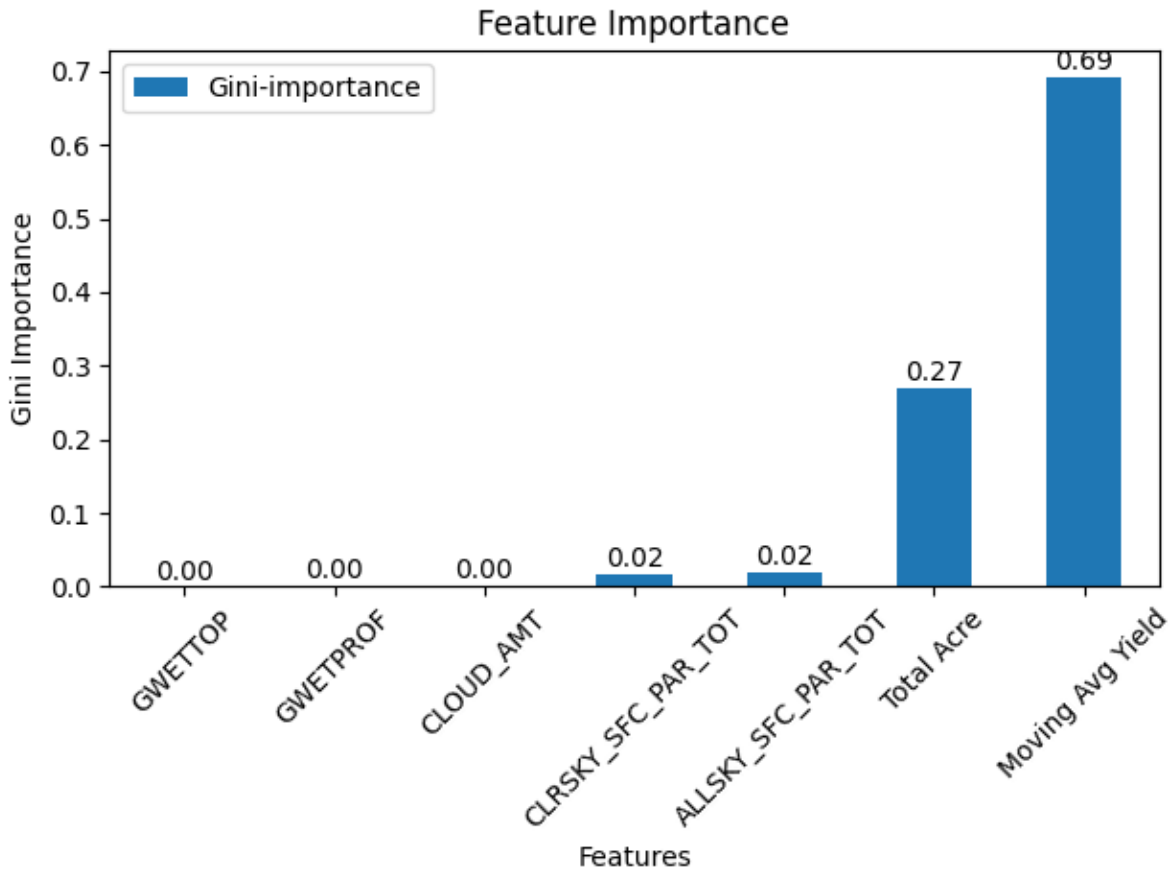
The analysis provides valuable insights into the climate factors affecting hop cultivation in Oregon and sets the foundation for modeling and predicting hop yields based on climate variables.

Random forest model

The analysis included the implementation of a RandomForestRegressor model to predict hop yield using the NASA climate data in Oregon. The dataset was divided into training and test sets, with 70% of the data allocated for training and the remaining 30% for testing the model's performance. The RandomForestRegressor model was trained using the training set, enabling it to learn the relationships between the input climate variables and hop yield. Predictions were then made on the test set to evaluate how well the model generalized to unseen data. Additionally, predictions were also generated on the training set to assess the model's performance on the data it was trained on. The evaluation of the model was done by calculating the R-squared scores, which provide a measure of the goodness of fit between the predicted and actual hop yield values. The R-squared score for the test set was found to be -0.156, indicating that the model's predictions did not align well with the actual values in the test data. In contrast, the R-squared score for the training set was 0.9999, indicating a nearly perfect fit to the training data. This significant discrepancy between the R-squared scores of the training and test sets suggests that the model may suffer from overfitting.

Overfitting occurs when the model becomes too closely tailored to the training data and fails to generalize well to new, unseen data.

Figure 60.



The feature importance are calculated, and a bar chart is created to visualize the importance of each variable. The Gini importance score is used to determine the relative significance of each feature in predicting the average yield.

A Linear Regression model is applied to analyze the merged dataset MergeData_df_NASA_OR for Oregon. The dataset is split into training and testing sets based on the specified years.

Conclusion

To improve the precision and dependability of forecasts, it is critical to carry on with the research and development of forecasting models for hop yield. Investigating the effects of particular climate factors on hop production can reveal important information about the connection between weather patterns and yield results. Researchers may better comprehend and predict upcoming hop yields by combining historical trends and climatic patterns into long-term forecasting models.

A deeper knowledge of hop production may also result from broadening the range of data sources beyond factors like climate and acreage. The results of yields can be considerably influenced by elements including soil qualities, disease prevalence, and market demand. Researchers may enhance the forecasting capability of the models and create plans to optimize resource allocation and farming practices by adding these extra data sources. Additionally, examining regional diversity in Washington, Idaho, and Oregon might provide producers with useful knowledge to adapt their production practices in light of particular area circumstances. Hop yield may vary by area due to specific traits and difficulties. Researchers may offer specialized advice and recommendations to improve hop output in various regions by recognizing these variances.

The present capstone project has successfully used regression modelling approaches to forecast hop yield based on climate and acreage data, however, there is still significant opportunity for advancement and more study. Future research may increase hop yield forecasts and help the hop business develop sustainably by diving more deeply into the effects of climatic factors, including more data sources, and taking regional variability into account.

References

Agritecture Team — AGRITECTURE. (n.d.). AGRITECTURE. https://www.agritecture.com/about

Yakima Valley Hops. (n.d.). *Citra Hops*. https://yakimavalleyhops.com/products/citra-hop-pellets#:~:text=the%20hop%20world.-,Citra%C2%AE%20is%20the%20most%20sought%20after%20hop%20variety%20because,a%20beer%20on%20its%20own.

Statista. (2023, January 30). *Top U.S. states for hop production 2020-2022*. https://www.statista.com/statistics/194288/leading-us-states-for-hop-production/

Marc, A. J. (2022, September 7). *How do hops react to climate change?* https://www.barthhaas.com/ressources/blog/blog-article/how-do-hops-react-to-climate-change

Lobell, D. B., & Asner, G. P. (2003). Climate and Management Contributions to Recent Trends in U.S. Agricultural Yields. *Science*, *299*(5609), 1032. https://doi.org/10.1126/science.1078475

Mourtzinis, S., Esker, P. D., Specht, J. E., & Conley, S. P. (2021). *Advancing agricultural research using machine learning algorithms*. Scientific Reports, 11, 17879. https://rdcu.be/daJhf

IoT-Equipped and AI-Enabled Next Generation Smart Agriculture: A Critical Review, Current Challenges and Future Trends. (2022). IEEE Journals & Magazine | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/9716089

Swain, M., Singh, R., Gehlot, A., Hashmi, M. F., Kumar, S., & Parmar, M. (2019, December 1). A reliable approach to customizing linux kernel using custom build tool-chain for ARM architecture and application to agriculture. *International Journal of Electrical and Computer Engineering (IJECE)*, *9*(6), 4920. https://doi.org/10.11591/ijece.v9i6.pp4920-4928

Koehrsen, W. (2019, December 10). Random Forest in Python - Towards Data Science. *Medium*. https://towardsdatascience.com/random-forest-in-python-24d0893d51c0

Singh, A. (2021). An Intuitive Guide to Interpret a Random Forest Model using fastai library (Machine Learning for Programmers – Part 2). *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2018/10/interpret-random-forest-model-machine-learning-programmers/

Python, R. (2022). Linear Regression in Python. *realpython.com*. https://realpython.com/linear-regression-in-python/#:~:text=The%20coefficient%20of%20determination%2C%20denoted,the%20output%20with%20different%20inputs.