Big Data Analytics and Visualization

Instructors: Dr. Sinan Kamal, Eng. Waed Al-Sawareah

Prepared by: Saja Abdulazeez on June 20th, 2023

Updated EDA notebook and summarized report: March 27th, 2024

Agriculture in Jordan

From the Jordan valley to Irbid and Ghor Al-Safi, agriculture holds a great significance in Jordan, harvesting olives and tending to livestock are traditions deep rooted within our culture. This country's unique geographical location and climatic conditions is what helped to shape the agricultural landscape we know today, utilizing open lands, rain-fed, and irrigation farming methods. However, due to climate change, the weather has been unpredictable, and farmers are finding it difficult to improve their land use and are often disappointed with crop production, this paper covers weather-based crop prediction for the year 2022, in order to avoid losses and make better decisions on which crops to grow.

The Agricultural Market

Agriculture plays a fundamental role in providing food security and economic stability in the kingdom and outside of it, since 2010 we notice a surge in exports of fruits and vegetables to the European market, reaching a total of €651 million in 2015 with tomatoes accounting for 65% of all vegetable exports, additionally in that year, agricultural exports have shifted towards the gulf market with UAE being in the lead[1]

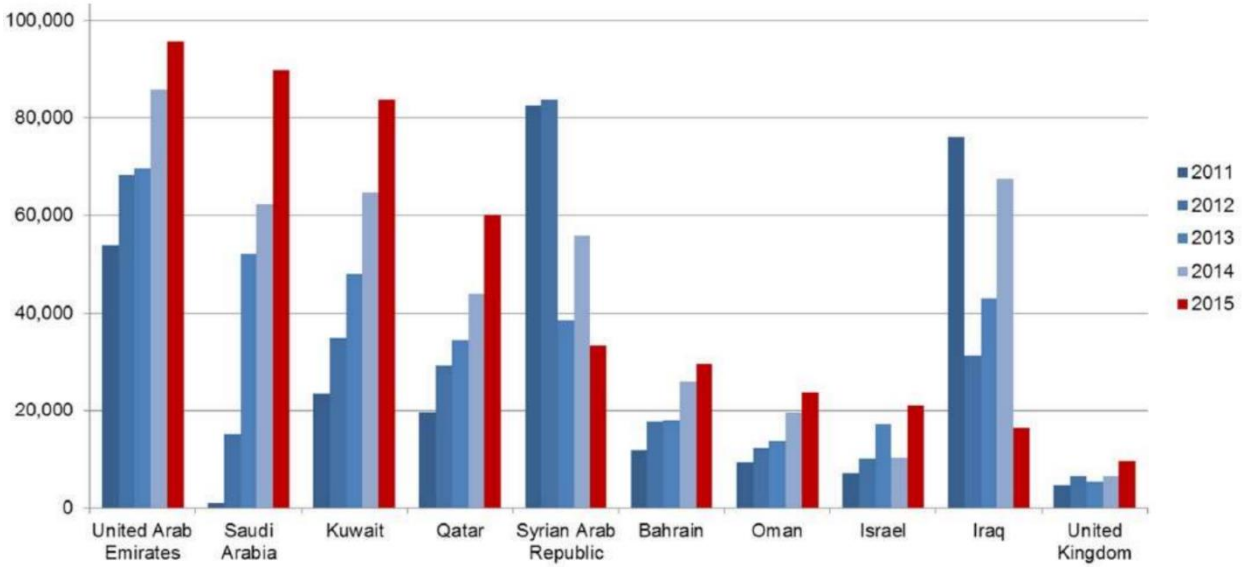


Figure 1 Jordanian vegetable exports (HS07): main markets (Leeters & Rikken, 2016)

Consider the following table for the main agricultural exports from Jordan in 2017[2]:

Commodity	,000 US$	MT unless otherwise noted
Tomatoes	223,054	282,271
Live sheep (number)	161,827	497,091 head
Peppers	56,068	47,970
Livestock forage	36,395	30,857
Cheese	28,034	6,436
Squash	23,372	27,693
Sweet melon	15,034	35,417
Cucumbers	11,545	19,024
Watermelons	10,424	19,095
Poultry meat	9,998	5,034
Cauliflower	9,717	14,414
Eggs (number)	7,187	34,055,400 eggs

Main Challenges

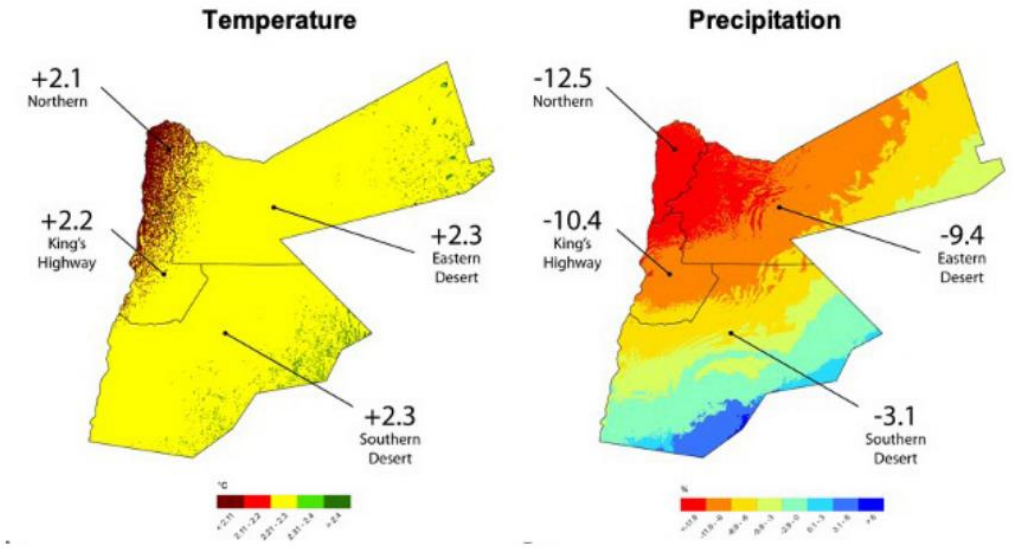
It's clear that Jordan plays a major role in agricultural exports, however, reaching those numbers is no easy task, agriculture in Jordan has faced (and currently is facing) some major challenges including but not limited to:

- The unpredictable weather due to climate change has negatively affected the agricultural sector, causing a drop in some crop yields and over-production in others. [3]
- Majority of crop production systems in Jordan are open field and depend on rainfall and stores of rainwater, which means crop production can be heavily influenced by the weather. 31% of Jordan Valley's vegetable production systems are open field[1] compared to only 11% which are in controlled greenhouse environments
- On august 31st 2022, the kingdom witnessed a 10-day heat wave rising +15° than normal, decreasing tomatoes in the central market by 50% in that year. [4]

- This increase in temperature has drastically affected precipitation levels causing drastic droughts across the kingdom and unfortunately this is expected to steadily increase by 2030.

Figure ES.2 Projected changes in annual mean temperature and total annual precipitation in Jordan

Projected changes by 2030 for Representative Concentration Pathway 8.5 (high emissions



This drastic shift in weather is out of our hands we need to find a way to adapt to this change, with that being said, can we find a way to predict the yield of crops based on historical weather data, to aid the decision-making process for agricultural exports in order to avoid losses in the long run?

Strategy

Our main business goal is to stay ahead of the market, with today's technologies and abundance of data we can indeed gain confidence in making decisions regarding what crops to focus on producing for the upcoming years (2022). This can be achieved by using big data tools to analyze historical multivariate timeseries weather data collected by Arabia Weather and crop yields across the years from the Department of Statistics data bank to discover a relationship between weather and crop yield. With this relationship we are able to predict the yield of crops in the upcoming years with machine learning in PySpark.

About the datasets

As mentioned briefly above, we have been given 3 data sets, 2 historical weather datasets from 2017-2023 for regions Irbid and Ghor El Safi, and a crop production data set for the respective regions from the years 2017-2021

- Crop production: the data bank contains data of 27 crops and their productions from the years 1994-2021 from different regions in Jordan
    - I will approach this by choosing the regions Ghor El Safi and Irbid , from the years 2017-2021 (most recent)
    - For each year this data set classifies production into two seasons: Summer and Winter
    - Based on my research I chose the crops Tomatoes, Okra and Onion dry for the following reasons:
        - Tomatoes are the most produced and exported outside of Jordan
        - Okra plants are drought-tolerant vegetables[5] and thrive in high summer temperatures which in my opinion is using the increasing temperatures in our favor. It can withstand a temperature up to 41^0 Celsius but are best grown between 29°-35 °. [6]
        - Onion on the other hand are cold-season crops[7] and grows in temperatures ranging from 12°-22° degrees Celsius and onion seeds can germinate in temperatures as low as 2°. [8]
        - In EDA I also explored the crops Cucumbers and Broad Beans
        - Broad beans are considered a cool season crop meaning they thrive in late spring – autumn and are quite self-sufficient[9]. Best temperatures are 15.5-18.3 and will not grow in temps under 4.4 or above 23.8[10]
        - Cucumbers grow best in temps between 24-29 [11]

- Irbid and Ghor Safi: these two datasets contain real time historical weather information that are collected every 5-6 hours daily from the years 2017-2023, the readings include:
    - Station: this could be Irbid or Ghor El Safi
    - Date/Time (year): these are the timestamps for when the reading was taken in the form *'2017-01-05 09:00:00'* indicating this reading was taken January 5th, 2017, at 9AM
    - Air Dew Point: The dew point represents the level of moisture, specifically the amount of water vapor present in the air and is measured in degrees Celsius. It's based on three factors, air temperature, humidity, and atmospheric pressure. [12]
    - Air Temperature: the temperature in degrees Celsius.
    - Humidity: measured in % is the percentage of humidity relative to the reading.
    - Manual Present Weather: a brief description of the weather in that day.
    - Cloud Type: classified in 6 main types to classify the cloud distribution they include altocumulus, cirrostratus, cirrus, cumulonimbus, cumulus, nimbostratus(only in Irbid dataset)
    - Cloud Cover (Okta): is the cloud cover in okta which is the measure of cloud cover ranging from 1-8; 1 being sparse clouds and 8 full cloud cover with no breaks.[13]
    - Cloud Cover %: percentage of cloud cover.
    - Wind Direction (Degrees): direction of the wind in degrees from 0-360.
    - Wind Speed MPS: speed of wind in meters per second.
    - Wind Type: describes the wind as either calm or normal.

The effects of weather on crop production

By this point we have established the effects of temperature on crop production and growth, so what about winds and cloud cover?

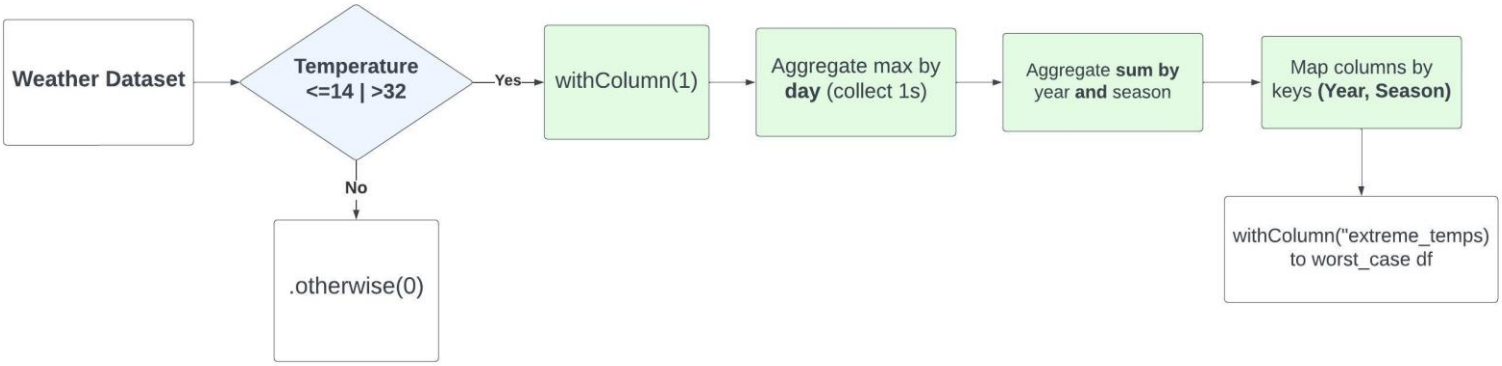
A light wind can actually promote the growth of plants since it increases the supplies of CO2 promoting photosynthesis however, as wind speeds increase, plant development slows down[16]; wind can alter the balance of hormones in plants and according to the Beaufort wind scale[17] , winds speeds of over 5.5m/s can cause breakage and harm the crops:

0 --- Calm	less than 1 mph (0 m/s)	Smoke rises vertically
1 --- Light air	1 - 3 mph 0.5-1.5 m/s	Smoke drifts with air, weather vanes inactive
2 --- Light breeze	4 - 7 mph 2-3 m/s	Weather vanes active, wind felt on face, leaves rustle
3 --- Gentle breeze	8 - 12 mph 3.5-5 m/s	Leaves & small twigs move, light flags extend
4 --- Moderate breeze	13 - 18 mph 5.5-8 m/s	Small branches sway, dust & loose paper blows about
5 --- Fresh breeze	19 - 24 mph 8.5-10.5 m/s	Small trees sway, waves break on inland waters
6 --- Strong breeze	25 - 31 mph 11-13.5 m/s	Large branches sway, umbrellas difficult to use
7 --- Moderate gale	32 - 38 mph 14-16.5 m/s	Whole trees sway, difficult to walk against wind
8 --- Fresh gale	39 - 46 mph 17-20 m/s	Twigs broken off trees, walking against wind very difficult

The correlations I found in data preprocessing (irbid_EDA_Viz.ipynb)

Feature Engineering

The strategy followed here was motivated from presenting findings in the worst case, and the best case; where the worst case contains weather data values that would affect the production negatively and the best case is where the weather would aid in better crop productions, take this flowchart as an example on worst case with example to temperature:



Applying similar to map reduce which is to assign the key values, in this case are season, year, and region; assign a value 1 to each instance that fulfills the requirement, collect the 1s by the day, then aggregate by season and year to get a data that can input into a function that maps the values retained to the crop data set. I will demonstrate with screen shots; I keep track of the count and order of rows by year (since its time series data) throughout the process

1. Make temporary df that adds a column which assigns 1 to extreme temps

```
temp1= irbid_weather.withColumn("extrm_temps",
    F.when((F.col("AirTemperature") <= 12) | (F.col("AirTemperature") > 30), 1)
     .otherwise(0)
)

temp1.show(3)
temp1.count()
```

```
+------+-------------------+-------------+-----------+---------------------+---------+--------+----+-----+---+------+-----------+
|Region|          Date/Time|AirTemperature|CloudsCover%|WindDirection(Degrees)|WindSpeed|WindType|year|month|day|Season|extrm_temps|
+------+-------------------+-------------+-----------+---------------------+---------+--------+----+-----+---+------+-----------+
| Irbid|2017-01-01 06:00:00|            5|         25|                    0|        0|    calm|2017|    1|  1|Winter|          1|
| Irbid|2017-01-01 09:00:00|            8|         25|                  300|        3|  normal|2017|    1|  1|Winter|          1|
| Irbid|2017-01-01 12:00:00|           11|         38|                  250|        3|  normal|2017|    1|  1|Winter|          1|
+------+-------------------+-------------+-----------+---------------------+---------+--------+----+-----+---+------+-----------+
only showing top 3 rows

8043
```

2. Collect the 1s with F.max and group by day, month, year, and season

```
#grouping by day
extrm_temps_days=temp1.groupBy('day',"month","year",'season').agg(F.max('extrm_temps').alias('extrm_temps')).
extrm_temps_days.show(10)
extrm_temps_days.count()
```

```
+---+-----+----+------+-----------+
|day|month|year|season|extrm_temps|
+---+-----+----+------+-----------+
|  1|    1|2017|Winter|          1|
|  2|    1|2017|Winter|          1|
|  3|    1|2017|Winter|          1|
|  4|    1|2017|Winter|          1|
|  5|    1|2017|Winter|          1|
|  6|    1|2017|Winter|          1|
|  7|    1|2017|Winter|          1|
|  8|    1|2017|Winter|          1|
|  9|    1|2017|Winter|          1|
| 10|    1|2017|Winter|          1|
+---+-----+----+------+-----------+
only showing top 10 rows

2040
```

3. Get the sum of the aggregate of key values year and season (notice count)

```
extrm_temps_years=extrm_temps_days.groupBy('year','season').agg(F.sum('extrm_temps').alias("sum_extrm_temps")).
extrm_temps_years.show(10)
extrm_temps_years.count()
```

```
+----+-------------+---------------+
|year|       season|sum_extrm_temps|
+----+-------------+---------------+
|2017|Autumn/Spring|             41|
|2017|       Summer|             43|
|2017|       Winter|             60|
|2018|Autumn/Spring|             36|
|2018|       Summer|             60|
|2018|       Winter|             92|
|2019|Autumn/Spring|             75|
|2019|       Summer|             56|
|2019|       Winter|             93|
|2020|Autumn/Spring|             54|
+----+-------------+---------------+
only showing top 10 rows

19
```

4. Defined a function that takes the parameter test which is a dataframe, then another function that is inner map which maps according to the keys, year and season. The test.get retrieves the values according to the keys and returns none if the key was not found, the inner map is is passed to F.udf which in PySpark is a user defined function to form the inner map function

```
[ ]  def map_values(test):
         def inner_map(year, season):
             return test.get((year, season), None)
         return F.udf(inner_map)
```

5. The temp_values function which goes through each row in the final grouped extreme_temps_year in a for loop. It assigns the key values that are year and season in the extreme_temps_years, it makes it in a form that I can put though the function mentioned in 4. This is also known as a lookup dictionary.

```
[ ]  temp_values = {(row['year'], row['season']): row['sum_extrm_temps'] for row in extrm_temps_years.collect()}
```

6. Finally I map the values of temperature to a new df named worst_case and the resulting data frame is as such

```
mapping = map_values(temp_values)
#adding new column in crop_worstacase
crop_worstcase = crop.withColumn("extreme_temperatures", mapping(F.col("year"), F.col("season")))
crop_worstcase.show()
```

```
+---------+------+----+------+----------+-------------------+
|     Crop|Region|Year|Season|Production|extreme_temperatures|
+---------+------+----+------+----------+-------------------+
| Tomatoes| Irbid|2017|Summer|   11124.5|                 43|
| Tomatoes| Irbid|2018|Summer|   15540.4|                 60|
| Tomatoes| Irbid|2019|Summer|    6292.4|                 56|
| Tomatoes| Irbid|2020|Summer|    5122.7|                 38|
| Tomatoes| Irbid|2021|Summer|    2968.9|                 70|
|     Okra| Irbid|2017|Summer|     319.0|                 43|
|     Okra| Irbid|2018|Summer|    1003.1|                 60|
|     Okra| Irbid|2019|Summer|    1808.7|                 56|
|     Okra| Irbid|2020|Summer|    1428.2|                 38|
|     Okra| Irbid|2021|Summer|     867.3|                 70|
|Onion dry| Irbid|2017|Summer|     294.0|                 43|
|Onion dry| Irbid|2018|Summer|     781.9|                 60|
|Onion dry| Irbid|2019|Summer|    5616.0|                 56|
|Onion dry| Irbid|2020|Summer|    6269.1|                 38|
|Onion dry| Irbid|2021|Summer|    7770.8|                 70|
```

The same rocedure was applied to the rest of the features selected which are:

Bad_winds: windspeeds >= 5mps

Too_cloudy: cloud cover > 30%

Harmful_wind_directoins: wind directions >=45

References

[1] Status report: SME Vegetable Farming in Jordan. (n.d.). Available at: https://www.hollandhortisupportjordan.com/wp-content/uploads/2019/09/Jordan-SME-Horti-Sector-2019.pdf.

[2]http://moenv.gov.jo/ebv4.0/root_storage/en/eb_list_page/climate_smart_agriculture_action_plan-jordan-4.pdf

[3] Tayseer, R. (2023). *Weather fluctuations play spoilsport for farmers*. [online] The Jordan Times. Available at: https://jordantimes.com/news/local/weather-fluctuations-play-spoilsport-farmers%C2%A0 [Accessed 12 Jun. 2023

[4] Mustafa, M.I. and Times, T.J. (n.d.). *Heatwave takes toll on crop yield: Jordan*. [online] www.zawya.com. Available at: https://www.zawya.com/en/world/middle-east/heatwave-takes-toll-on-crop-yield-jordan-fx2qaoup [Accessed 19 Jun. 2023].].

[5] Austin American-Statesman. (n.d.). *GARDENING. Plant okra in summer's heat*. [online] Available at: https://www.statesman.com/story/news/2017/05/08/gardening-plant-okra-in-summers-heat/10405904007/#:~:text=A%20drought%2Dtolerant%20vegetable%2C%20okra [Accessed 19 Jun. 2023].

[6] M, J. (n.d.). *How Big Does Okra Get?* [online] GreenUpSide. Available at: https://greenupside.com/how-big-does-okra-get/.

[7] Brad (2021). *When To Plant Onions – Planting Guide 2023*. [online] Northern Nester. Available at: https://northernnester.com/when-to-plant-onions/#:~:text=Onions%20are%20a%20cold%2Dseason [Accessed 19 Jun. 2023].

[8] www.agrifarming.in. (2020). *Onion Seed Germination, Time, Temperature, Procedure | Agri Farming*. [online] Available at: https://www.agrifarming.in/onion-seed-germination-time-temperature-procedure.

[9] Edwards, S. (2020). *How To Grow Broad Beans | A Guide By Seasol*. [online] Seasol. Available at: https://www.seasol.com.au/tomatoes-herbs-and-vegetables/how-to-grow-vegetables-broad-beans/#:~:text=Broad%20beans%20are%20a%20cool.

[10] Psu.edu. (2013). *Broad bean, dry | Diseases and Pests, Description, Uses, Propagation*. [online] Available at: https://plantvillage.psu.edu/topics/broad-bean-dry/infos.

[11] M, J. (n.d.). *Lowest Temperature Cucumber Plants Can Tolerate (3 Things To Know)*. [online] GreenUpSide. Available at: https://greenupside.com/what-is-the-lowest-temperature-cucumber-plants-can-tolerate/#:~:text=According%20to%20the%20University%20of [Accessed 19 Jun. 2023].

[12] Netatmo. (n.d.). *What is a dew point?* [online] Available at: https://www.netatmo.com/en-eu/weather-guide/dew-point [Accessed 19 Jun. 2023].

[13] Met Office. (n.d.). *How we measure cloud*. [online] Available at: https://www.metoffice.gov.uk/weather/guides/observations/how-we-measure-cloud#:~:text=1%20okta%20represents%20a%20cloud [Accessed 19 Jun. 2023].

[14] VEDANTU. (n.d.). *Relation Between Temperature and Humidity*. [online] Available at: https://www.vedantu.com/geography/relation-between-temperature-and-humidity.

[15] Quora. (n.d.). *What is the role of temperature in the formation of wind?* [online] Available at: https://www.quora.com/What-is-the-role-of-temperature-in-the-formation-of-wind [Accessed 19 Jun. 2023].

[16] Agriculture, A.-A.F.B.F. for (n.d.). *How Does Weather Affect Farming?* [online] www.agfoundation.org. Available at: https://www.agfoundation.org/news/how-does-weather-affect-farming#:~:text=Farmers%20rely%20on%20good%20weather.

[17] gyre.umeoce.maine.edu. (n.d.). *Variable Description*. [online] Available at: http://gyre.umeoce.maine.edu/data/gomoos/buoy/php/variable_description.php?variable=wind_2_speed.

[18] Collibra (2022). *The Importance of Data Governance*. [online] Collibra. Available at: https://www.collibra.com/us/en/blog/importance-of-data-governance.

[19] Lee, J.B., Xiangrui Meng, and Denny (2016). *Why you should use Spark for machine learning*. [online] InfoWorld. Available at: https://www.infoworld.com/article/3031690/why-you-should-use-spark-for-machine-learning.html.