

جامعة طيبة

كلية علوم وهندسة الحاسبات

**(**قسم الطالبات)

KINGDOM OF SAUDI ARABIA

Ministry of Higher Education

Taibah University

College of Computer Science and

Engineering (Girls Section)

**A project for**

**Data Mining & Data Warehouse IS-372**

**"**Dates Export in Saudi Arabia**"**

|  |  |  |
| --- | --- | --- |
| **4150201** | **F10** | **Areej Hilal Alsehli** |
| **4258567** | **F 4** | **Rand Abdulaziz Al-subeeh** |

**Supervised BY dr.Alaa AlharbI**

**Introduction**

**Objective:**

This project aims to achieve several objectives, including studying the development and growth of the Saudi date export market and helping to understand the most influential countries in this sector's prosperity. From other perspectives, this project was selected for implementation to support and contribute to Vision 2030, which aims to strengthen Saudi Arabia's position as the world's largest exporter of dates and raise the value of non-oil exports

**The importance of data analysis and mining:**

The importance of data analysis and mining in the current era lies fundamentally in making accurate decisions and making predictions based on previous studies. It helps in understanding patterns and relationships, which improves the way data is handled and leads to things such as: increased competitiveness, improved innovations, and increased efficiency.

**Data Selection**

The data was retrieved from the Open Data Platform [1], a platform developed by the Saudi Data and Artificial Intelligence Authority (SDAIA). The official publisher of the data is the national center for palms and dates [2].

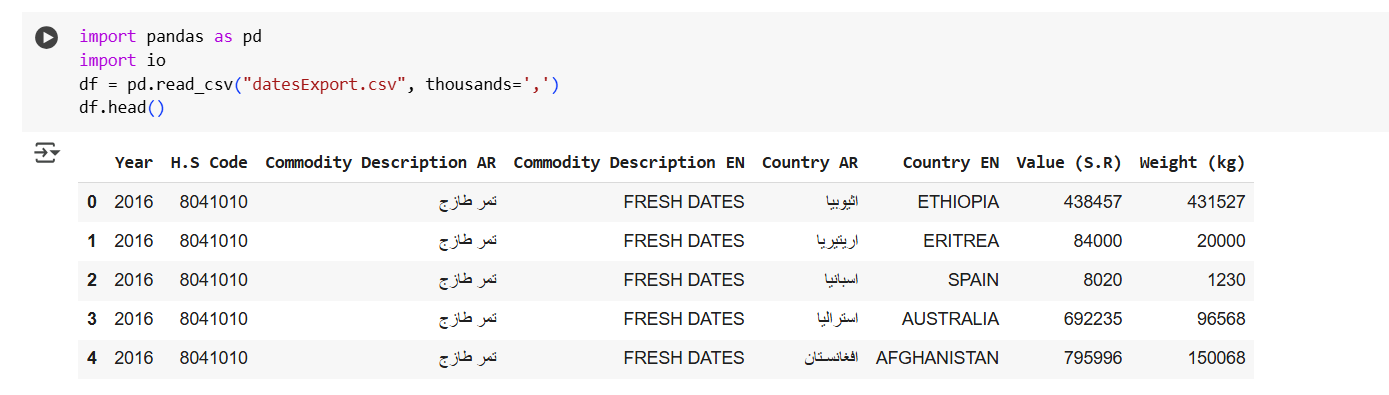
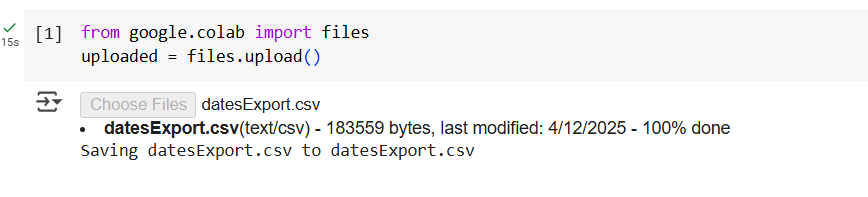
The data covers exports, including the year, product classification, countries, export value, and shipped quantity. Each row represents an export transaction to a specific destination, reflecting the global distribution of dates

The total number of the dataset is 1911 rows, from the period 2016-2023.

This data set was selected to understand the date trading sector and gain a broader perspective in the past and current years, to influence the efficiency of choices and discover patterns through analysis in order to improve decision making.

**Datasets Description**

|  |  |  |
| --- | --- | --- |
| **Description** | **Data type** | **Data** |
| Export year | numeric | Year |
| Weight of the exported shipment | numeric | Weight (kg) |
| A standardized numerical method of classifying traded products [3] | numeric | H.S code |
| Type of dates exported in Arabic | text | Commodity Description AR |
| Type of dates exported in English | text | Commodity Description EN |
| Shipment price in riyals | numeric | Value (S.R) |
| The country to which it was exported to, written in Arabic | text | Country AR |
| The country to which it was exported to, written in English | text | Country EN |

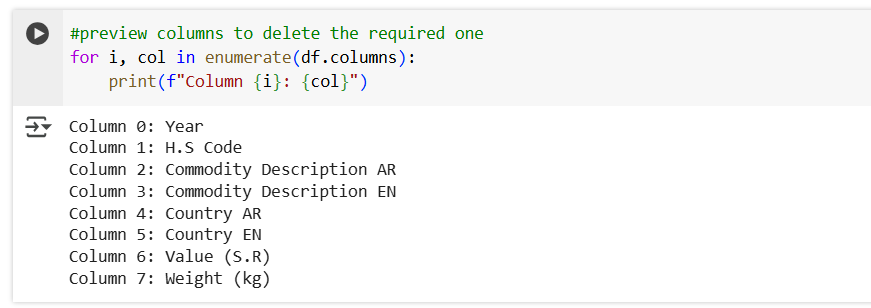
- Reading the file and printing the first 5 rows.

**Data** **Preprocessing**

In this section, the most important work necessary to prepare the dataset before starting the project was carried out, which included the following:

**Cleaning:**

It was noticed that the country and product description columns were in two different languages, and Arabic was removed to improve the data entered the model and analysis, and the English columns were renamed for uniqueness.



A screenshot of a computer

AI-generated content may be incorrect.



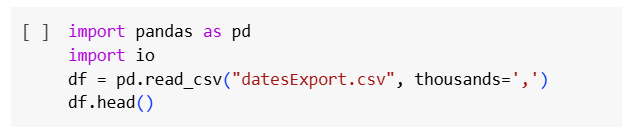
**handling missing and** **duplicates values:**

As part of the Preprocessing, a missing and duplicates value was checked, and it was found that there was none in the data**A screenshot of a computer code

AI-generated content may be incorrect.**

**transforming data into a suitable format for analysis:**

The commas were changed so that Python reads them as INT and not as STRING, using (thousands=',') which considered Data Parsing because It tells pandas how to interpret the data directly when it is read, without the need for subsequent manual cleanup

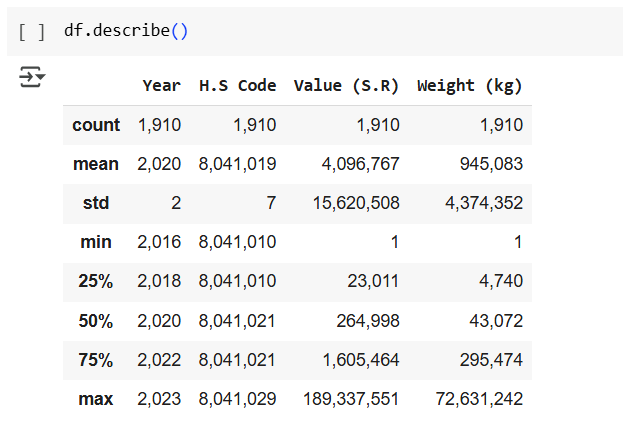
****

**Descriptive Statistics and Visualization**

In this section descriptive statistical analysis has been Conduct on the dataset to gain insights into its characteristics.

Descriptive analysis provides us with a broader understanding of data and identifies problems and patterns. It presents us with a numerical summary of the data, such as (median, mean, and standard deviation). It helps us understand the distribution and deal with extreme values. Then, we present it through graphical visualization.

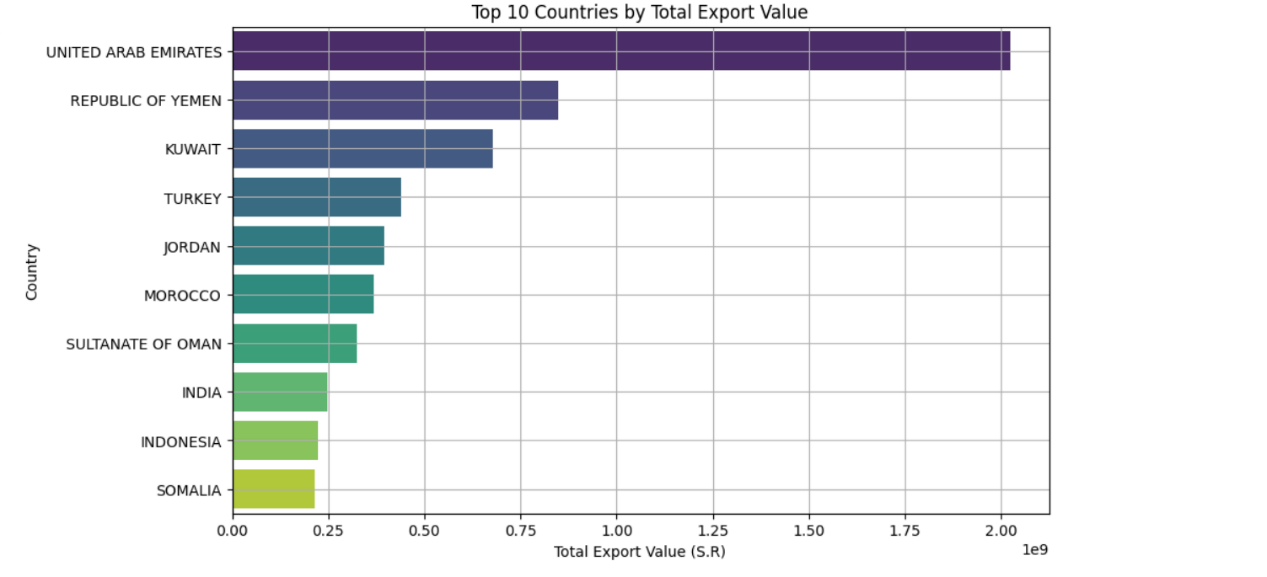
**descriptive statistical analysis**

****

From the statistical analysis shown above, it can be concluded that the data shows a concentration around a specific commodity imported or exported in very different quantities and values. It is possible that a few records have a significant impact on the overall averages due to their outliers, which will be Visualized beneath.

**Visualization**

**BAR PLOT**

This chart represents the top 10 countries in terms of export value in Saudi Riyals. It clearly shows that the United Arab Emirates comes in first place by a very large margin over the rest of the countries, followed by Yemen, then Kuwait. At the end of the list, Somalia and Indonesia are shown with the lowest export value, indicating the market's dependence mainly on UAE exports, which highlights the opportunities available to strengthen the market towards a specific country.

**LINE PLOT**

A graph with a line and a line

AI-generated content may be incorrect.In this graph, the study of the value of UAE exports during the years available in the data list was determined after noting that the UAE has the highest list of total exports, as the curve in the graph shows a continuous increase over the years starting from 2016 and reaching 2023, with a clear decrease in the year 2020 compared to 2019, and this may be due to the occurrence of the Covid-19 pandemic.

Multiple Line Chart

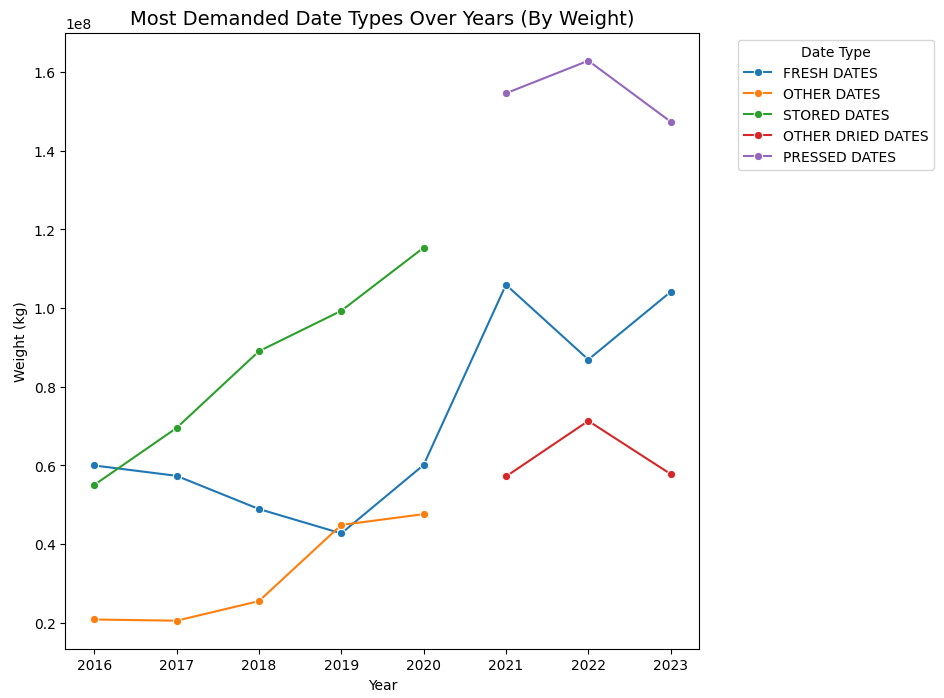
This graph was created to understand the most in-demand date varieties during the years 2016-2023, allowing for a greater understanding of market demand. As shown, the following is analyzed:

First: Highest Demand

* Pressed dates are the most in demand by weight.
* Fresh dates are the most in demand by timeframe, as they are the only variety that has been exported continuously from 2016 to 2023, while the remaining varieties vary over time.

Second: Conclusions

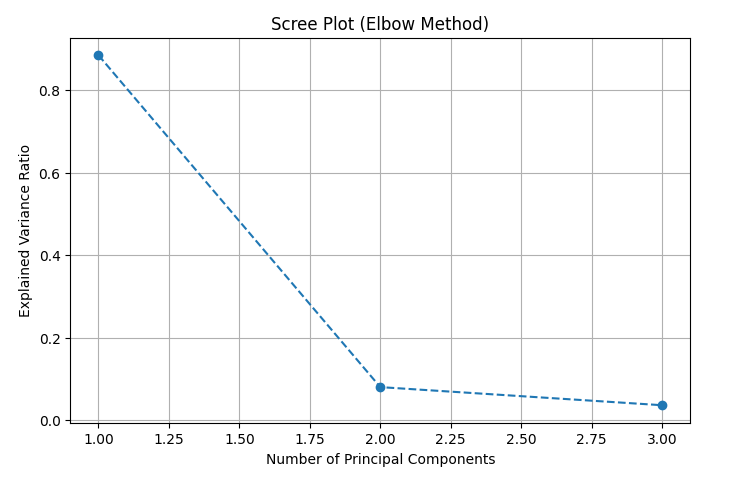
Pressed dates were the most in-demand product in the period from 2020 to 2023.

Fresh Dates showed a decline in 2019 but rebounded in 2020, indicating a revitalized market

**Principal Component Analysis (PCA) Implementation**

In this part of the project, PCA was implemented to reduce the dimensions of the data while preserving as much information as possible. This provides us with a simplification of complex data, improves algorithms, and reduces noise resulting from low-variance components. The elbow method was implemented and displayed using a scree plot, and then PCA SCATTER PLOT was implemented based on that.

**Scree Plot (Elbow Method)**

This diagram below represents the elbow method, where the elbow is the point after which the variance curve begins to flatten, i.e.: adding any additional components no longer adds important information. In this diagram, the elbow is located at point 2 (i.e.: when the number of principal components = 2) and represents the best balance where the error rate becomes more stable.

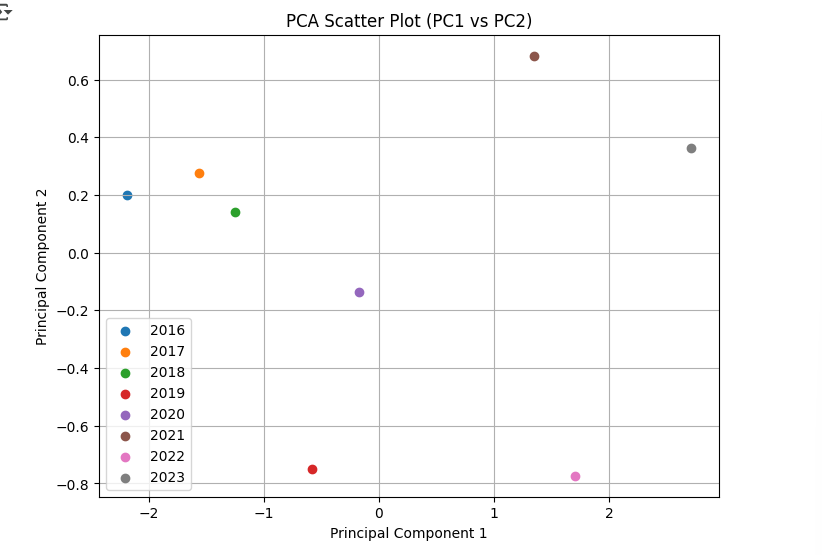
**PCA Scatter plot**

THIS SCATTER PLOT provides visual representation from 2016 to 2023 after dimension reduction to 2 main components where every year is represented as a dot, and Their positioning in the graph reflects the data pattern for that year based on principal components.

**Conclusion:**

Convergent data, such as 2016-2017, indicates similar characteristics.

Divergent data, such as 2016-2021, indicates significant differences in the data pattern.

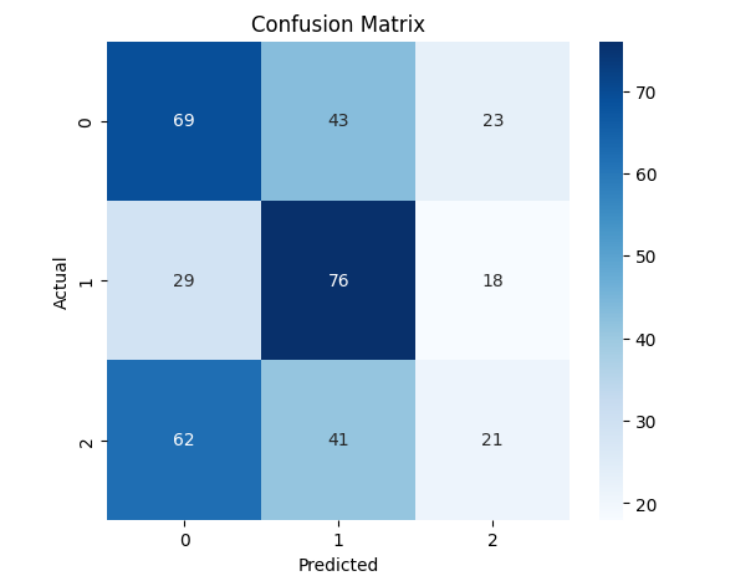
****

**Classification**

In this section we start training the Random Forest Classifier that aimed to Predict the 2024 Exports Based on Date Type, first by classify the export commodities into three **weight categories** (Low – Medium - High) based on two features: **Year** and **Commodity Description** (encoded as a numeric code)

And evaluating the classifier's performance using a **classification report**, **confusion matrix**, and **heatmap visualization**.

**Confusion Matrix (Heatmap Visualization)**

****

**Conclusion:**

1. The heatmap provides us with a visualization to the distribution of Actual vs predicted.
2. 76 correct "Low" classifications that are seen in the center is considered the strongest block.
3. There are misclassifications that are frequent and diverse.
4. Significant confusion between all three categories. For example:

- 62 actual "Medium" samples were classified as "High".

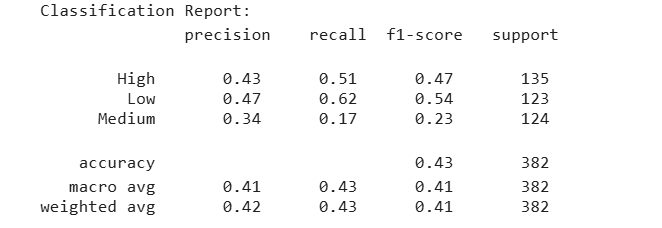
- 43 actual "High" were classified as "Medium"

**Commodity Description Predicted Weight Category 2024**

**A close-up of a white background

AI-generated content may be incorrect.**

**Evaluation and Results Interpretation**

****

**Classification Report Summary:**

* Overall accuracy: 43%
* Macro average: precision: 41%, recall: 43%, F1-score: 41%

**Conclusion**

This project covers a comprehensive analysis of date export data from Saudi Arabia during the period 2016–2023. The goal is to understand the market and identify the most prominent importing countries and the most in-demand date varieties, supporting the Kingdom's efforts to meet Vision 2030's goal of boosting exports. The data was processed from various aspects, including cleaning, sorting, and analysis using descriptive statistical techniques and graphs, providing a deeper understanding of the market.

Finally, the random forest classifier was used to predict 2024 exports. The results showed that the UAE represents the largest export destination, with the most in-demand date varieties highlighted. The confusion matrix also provided an accuracy of 43%, highlighting the challenges of accurate classification

**Recommendations**

- Improving data quality:

such as adding new features such as packaging type or storage method.

- Seasonal analysis:

such as analyzing specific months, such as religious seasons (Ramadan and Hajj), or a global issue such as the COVID-19 pandemic.

**Appendix:**

https://colab.research.google.com/drive/18IWiX7-2AaGr44Y\_L7LyPIZ9TJP51XRw?usp=sharing

References:

[1] data source: https://open.data.gov.sa/en/datasets/view/d0a9c6c5-ad58-40e7-be1e-772990d001ad

[2] publisher :https://ncpd.gov.sa/ar/reports-statistical

[3] https://www.trade.gov/harmonized-system-hs-codes