R programming

POLLEN E MAAKANA

**PS2 PS2**

**POLLEN E MAAKANA**

**2437481**

**R PROGRAMMING R PROGRAMMING**

Table of Contents

[Part A Research 2](#_Toc181835366)

[Q1 Solutions 2](#_Toc181835367)

[Q2 solutions 3](#_Toc181835368)

[Q3 Solutions 3](#_Toc181835369)

[Q4 Solutions 4](#_Toc181835370)

[Part2 Report 5](#_Toc181835371)

[Introduction 5](#_Toc181835372)

[Code implementation: 5](#_Toc181835373)

[Data Loading and Cleaning: 5](#_Toc181835374)

[Preforming Descriptive Statistics and Summary: 7](#_Toc181835375)

[Analysis of Productive Crops and Regions: 8](#_Toc181835376)

[Visualization of Trends: 10](#_Toc181835377)

[Price Prediction: 11](#_Toc181835378)

[seasonal Trends Analysis: 13](#_Toc181835379)

[Summary gained from agricultural data analysis 14](#_Toc181835380)

[Performing crops and region 14](#_Toc181835381)

[Pricing and Production trends 16](#_Toc181835382)

[Recommendations for crop production 17](#_Toc181835383)

[Harvest Estimation 17](#_Toc181835384)

[Seasonal Recommendations 18](#_Toc181835385)

[Conclusion 18](#_Toc181835386)

[References 19](#_Toc181835387)

# Part A Research

## Q1 Solutions

1. The packages are ggplot2, dplyr, tidyr, readr and tibble.to use the packages in r programming install it first using this command install.packages(“tidyverse”) and call it using library(tidyverse) to access its features here it a breakdown of each feature of the package:

* data visualization and exploration with ggplot2 it is based on the grammar of graphics to construct beautiful informative plots. example aes() function it is used to define how data variable is mapped on visual properties.
* data manipulation with dplyr provides set of functions such as select() function which will pick variable based on their name.
* data transformation or tidying with tidyr helps in reshaping and cleaning data into tidy format example gather () function transform data from wide format to long format is useful when you have multiple column that represent variable of single value.
* data importing and cleaning with readr it is used for reading data from different files and importing the files example of the function is read\_csv().
* data handling with tibble it is designed to be user friendly by enhancing error handling and simplifies data manipulation of data frames. Example of its function tibble() which create new tibble

## Q2 solutions

1. 3 ways of visualizing data is by using Bar chart, histogram and box plot here is a breakdown of each visualizing data:

* Histogram shows the distribution of continuous value or variable. here is a code example of plotting histogram:
  + ggplot(mtcars, aes(x=mpg)) +geom\_histogram() +labs(title=”histogram”)
* Box plot compares the distribution of continuous data example of the code:
  + ggplot(mpg, aes(x=class,y=hwy))+geom\_boxplot()+labs(tittle=”mills”,y=”car class”,x=”high”)
* bar chart it compares categorical data example of code:
  + ggplot(diamonds, aes(x=cut))+geom\_bar()+labs(tittle=”cut”,x=”yes”,y=”count”)

## Q3 Solutions

1.The saying points out that a lot of statistical claims can be unreliable, often because of data quality issues like missing or incomplete information. If these issues are not addressed properly, they can lead to false conclusions, which can influence analysis (Reinhart, 2015). This emphasizes the importance of data cleaning in R programming since it makes sure that the data is correct before analyzing. Data cleaning assists in rectifying the problems, dealing with the missing information and ensuring uniformity all of which enhance and improve the reliability of the results. Packages such as dplyr and tidyr in R provide the ability to perform data cleaning in a quicker and more efficient manner. Because of this step, analysts can be more confident in their outcomes and make optimal choices (Miller, 2018).

## Q4 Solutions

1. characteristic of messy data:

* Variables are stored in both row and column: variables are found in both row and column. This can cause disturbance in analysing, to fix this you can use the pivot\_longer() function from the tidyverse package to convert wide format data into long-format data
* Column headers are values not variable names: this makes data difficult to analyze. To solve this you can rename the columns to reflect meaningful variable names using colnames()
* Multiple variables are stored in one column: to solve this you can use the separate () function from the tidyr package to split a single column into multiple column

# Part2 Report

## Introduction

The agricultural data analysis is important in understanding the performance of crops and regional productivity. By analyzing production quantities, price trends and regional productivity. This report points on revealing insight that can guide decision on crop selection and shape market strategies with accurate data, farmers can optimize their yields, meet market demand and boost their profitability. This report takes a closer look at various crops and regions, visualizing trends that tell a story about agriculture performance, providing ways on supporting crops production strategies, accurate Harvest estimation and aid in seasonal planning

## Code implementation:

***Note:*** *I will first explain the code and then provide a screenshot of the code along with its output.*

### Data Loading and Cleaning:

started by installing tidyverse package, which is designed for data science, this package includes data manipulation, data visualization and more. Data cleaning and loading was performed using the tidyverse package such as readr to read through the dataset. Data it is variable which will hold CSV file and load the dataset. Print function was used to view the output of data as it is shown below

A blue rectangle with black text

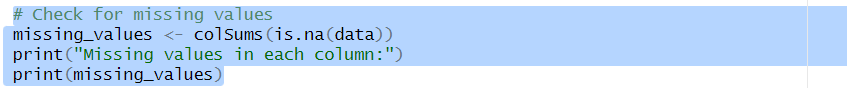
Description automatically generated

Output:

A screenshot of a computer

Description automatically generated

The code below checks for missing values in a column using colSum() and is.na() to identify incomplete data

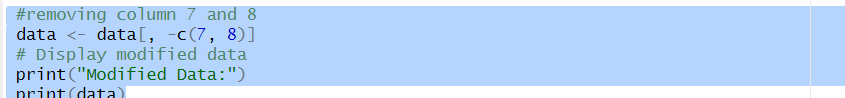


Output

:A computer code with text

Description automatically generated

columns 7 and 8 were removed since they were not having values to ensure the quality of data.



Output:

A screenshot of a computer

Description automatically generated

This code removes missing values in a row and provide clean dataset

A blue rectangle with black text

Description automatically generated

Output:

A screenshot of a computer

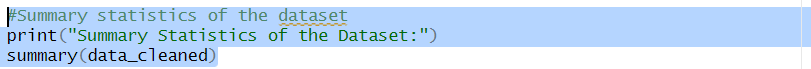
Description automatically generated

Cleaning data is important as it makes data to have quality and analysis to be accurate, after cleaning the data we proceeded to the following stage of statistical analysis since the data was accurate.

### Preforming Descriptive Statistics and Summary:

Descriptive statistics and summary were applied to understand the characteristics of dataset.

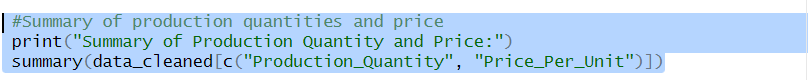
* This command displays the summary of statistics such as mean, median, and many more



Output:

A screenshot of a computer

Description automatically generated

The command summarizes the production quantity and price per unit, providing the understanding average market price based on the production quantity

Output:

A screenshot of a computer screen

Description automatically generated

### Analysis of Productive Crops and Regions:

Library (dplyr) was loaded to allow the manipulation of dataframes in a user-friendly way. Data was grouped by crop name and production quantity was summarized.

A blue screen with black text

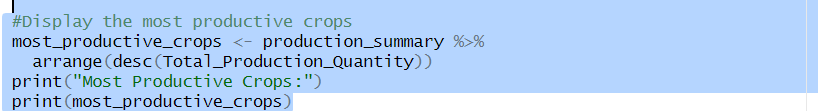
Description automatically generated

Output:

A screenshot of a computer

Description automatically generated

* After getting the production summary the crop name was arranged in descending order to find the highest production quantity with crop name show on the output.

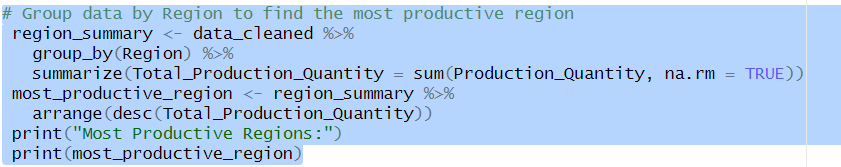


Output:

A screenshot of a computer

Description automatically generated

* Cleaned data was grouped by a region to find the most productive region



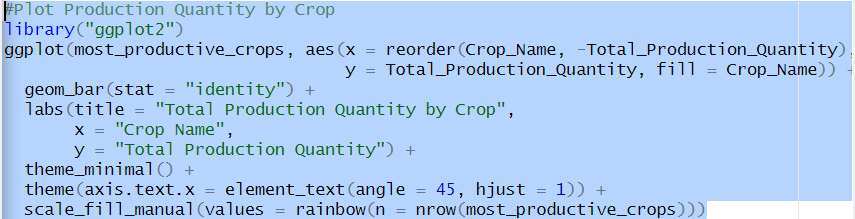
Output:

A screenshot of a computer code

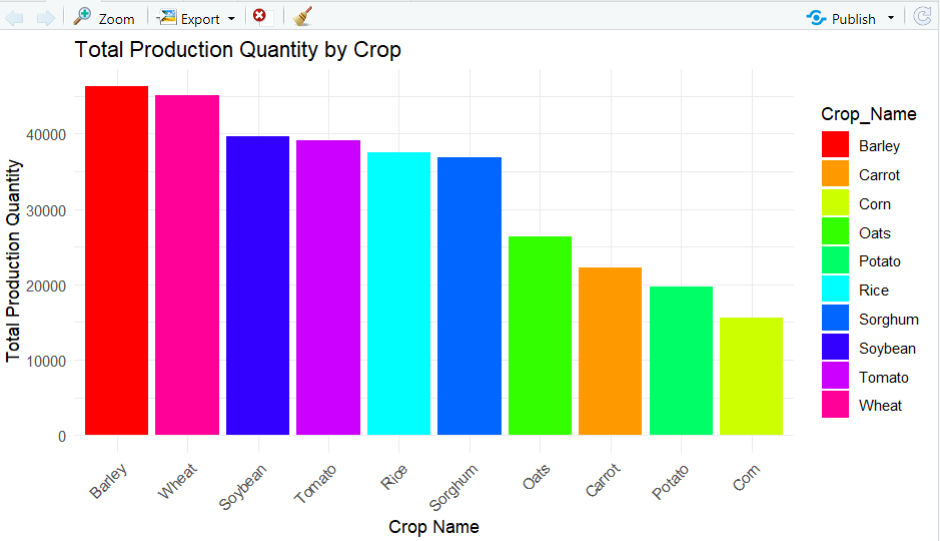
Description automatically generated

### Visualization of Trends:

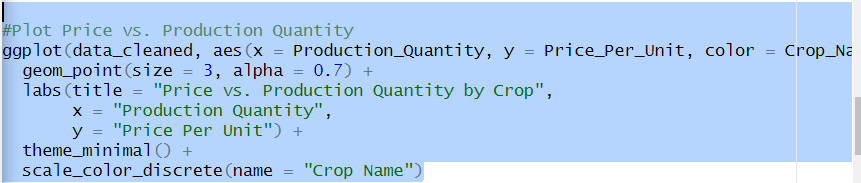
ggplot2 was used to show trends, the bar plot was created to show total production quantities by crops.

****

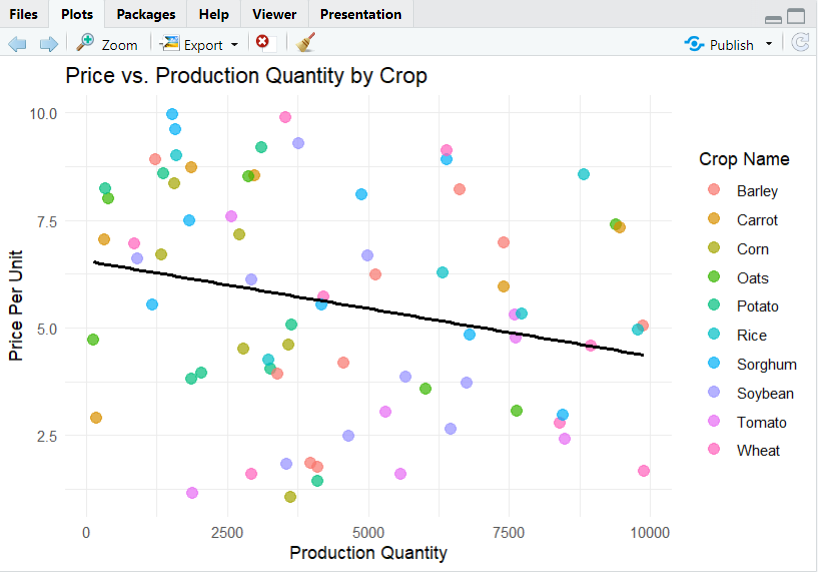
Output:



* Scatter plot was plotted to show the relationship between production quantity and price per unit price



Output:



* This code plots production quantity by region using bar plot

A computer screen with text

Description automatically generated

Output:

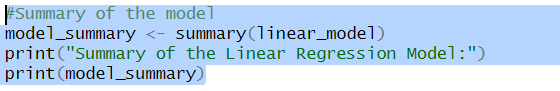
A graph of different colored rectangular shapes

Description automatically generated

### Price Prediction:

* Linear regression model was created to predict price per unit based on production quantity

****



Output:

A screenshot of a computer

Description automatically generated

* The command below is for printing predicted prices

A screen shot of a computer

Description automatically generated

* The below output shows the predicted prices for new production quantity

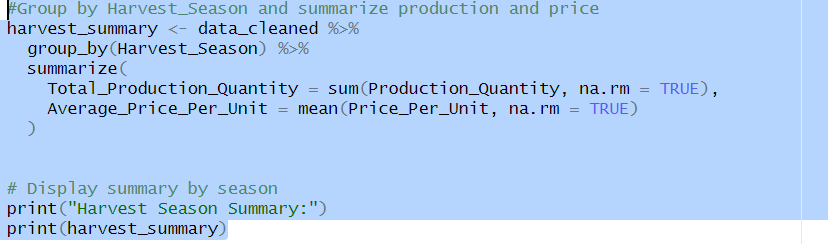
Output:

A screen shot of a computer

Description automatically generated

### seasonal Trends Analysis:

* The code shows seasonal trend analysis for summarized production and price which is group by harvest season by revealing clear patterns in production and pricing

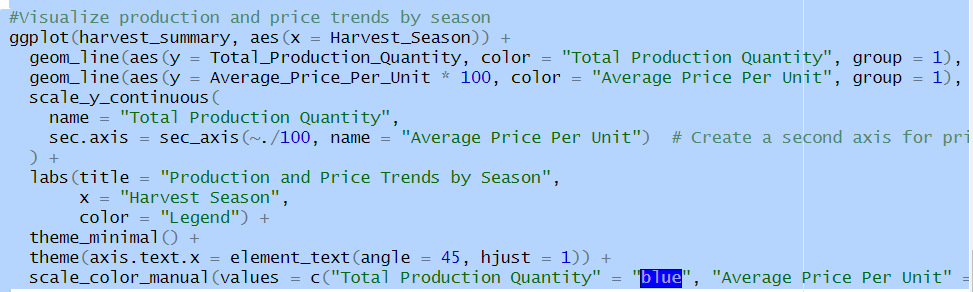
****

Output:

A screen shot of a computer code

Description automatically generated

* This code visualizes production and price trends per season



Output:

A graph with a line and numbers

Description automatically generated with medium confidence

## Summary gained from agricultural data analysis

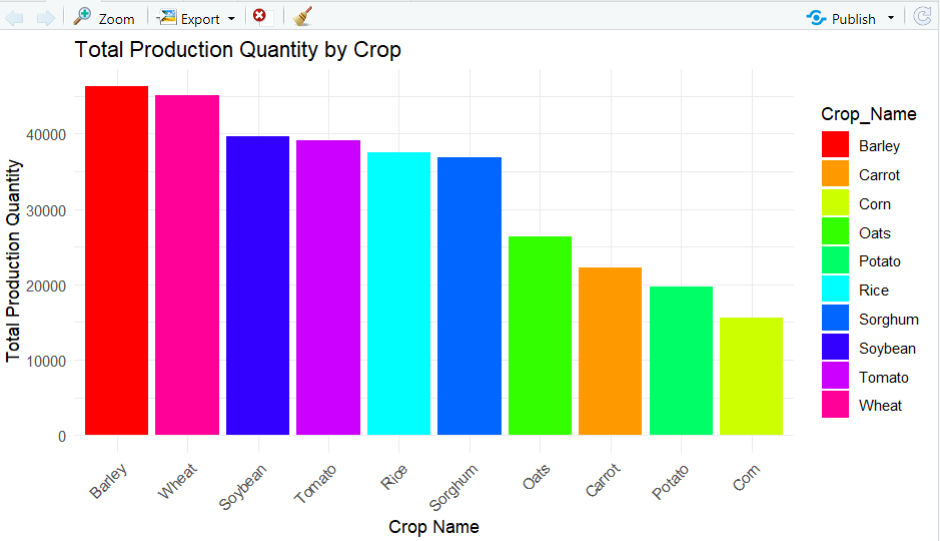
* This was a clean data that was used to analyze agricultural data

A screenshot of a computer

Description automatically generated

### Performing crops and region

#### Performing crops



* The finding shows that the most productive crop was Barley with the production of 46292, which makes it stronger priority for crop followed by wheat.

#### Performing region

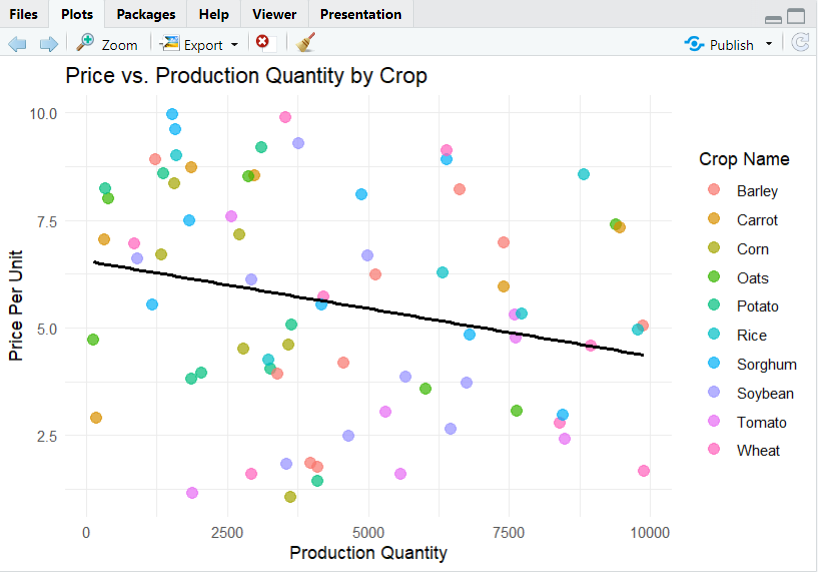
A graph of different colored rectangular shapes

Description automatically generated

* The finding shows that the Most productive region is the north region with the production quantity of 81985 this shows that north region it is important for investments

These findings show the top productive crops and regions, by knowing this you can decide where to send the crops in the regions which are productive like Barley in the north, this can boost agriculture output and economic returns.

### Pricing and Production trends



* The plot suggests a negative correlation between production quantity and the price per unit. As the production quantity increases prices per unit tends to decrease and the trend is not linear.

A graph with a line and numbers

Description automatically generated with medium confidence

The seasonal trends show that:

* Spring yielded the highest production quantity and had a competitive average price per unit. This is the best season to plant high yield crops to maximize production
* Fall is the best season for harvesting with its competitive price farmers can harvest during this season to match the demand and to profit from.

## Recommendations for crop production

Based on the insight from agricultural data analysis:

* High yield crops must be prioritized such as barley and wheat which shows strong production level. This crop must be allocated resource and planted on schedules aligning with the best season to optimize the output. On balancing the market plant soybean to avoid oversaturating of crops in the market which can cause price to drop and planting different soil it can improve soil health.
* Optimize resource allocation by region invest in regions which are productive such as the north by allocating resource that can enhance crops health. Engage with farmers to understand their challenges by doing so it can help on implementing a solution that is innovative
* Seasonal planning, prioritize on planting high yield crops in spring and harvest them on fall this aligns with competitive pricing patterns which can lead to profits

### Harvest Estimation

Predict yield based on data trends, analyze historical data by checking the data of the past yields. Considering patterns which were related to soil, weather and farming practices. Monitor demand to make a best decision on crop quantities.

### Seasonal Recommendations

1. Plant in spring to maximize yields, spring has the highest yield production quantity by planting high demand crops in this season it can boost the output
2. Harvest in Fall since fall this aligns with competitive pricing patterns by harvesting in this season it can be favorable for market conditions and it will lead to making a

profit.

## Conclusion

The analysis of agricultural production data has revealed important information that can help farmers boost their crop yields and improve harvesting strategies. By focusing on high-yield crops like Barley, Wheat, and Soybean, especially in the productive regions of the North and West, farmers have a great opportunity to increase both their harvests and profits.

Timing is everything and aligning with planting schedules in best seasons such as Spring and Fall can make a big difference in crop growth. It’s also important to keep an eye on the relationship between how much is produced and market prices to avoid flooding the market

# References

1. Miller, J. (2018). The importance of data quality in analysis: Avoiding false conclusions through data cleaning. *Journal of Data Science Methods, 12*(3), 123–135.
2. Reinhart, A. (2015). *Statistics done wrong: the woefully complete guide*. No Starch Press.
3. *GeeksforGeeks****.*** *(2020, June 7). What are the Tidyverse packages in R language? GeeksforGeeks.* [*https://www.geeksforgeeks.org/what-are-the-tidyverse-packages-in-r-language/*](https://www.geeksforgeeks.org/what-are-the-tidyverse-packages-in-r-language/)‌
4. Grolemund, G., & Wickham, H. (2017). *R for Data Science*. Had.co.nz. <https://r4ds.had.co.nz/>
5. 5. Grolemund, G. (2024). *Whole game – R for Data Science (2e)*. Hadley.nz. https://r4ds.hadley.nz/whole-game

‌