Visualizing distributions

INTRODUCTION TO DATA VISUALIZATION WITH JULIA

Gustavo Vieira Suñe
Data Analyst

datacamp

Onion and wheat prices in Kerala, India

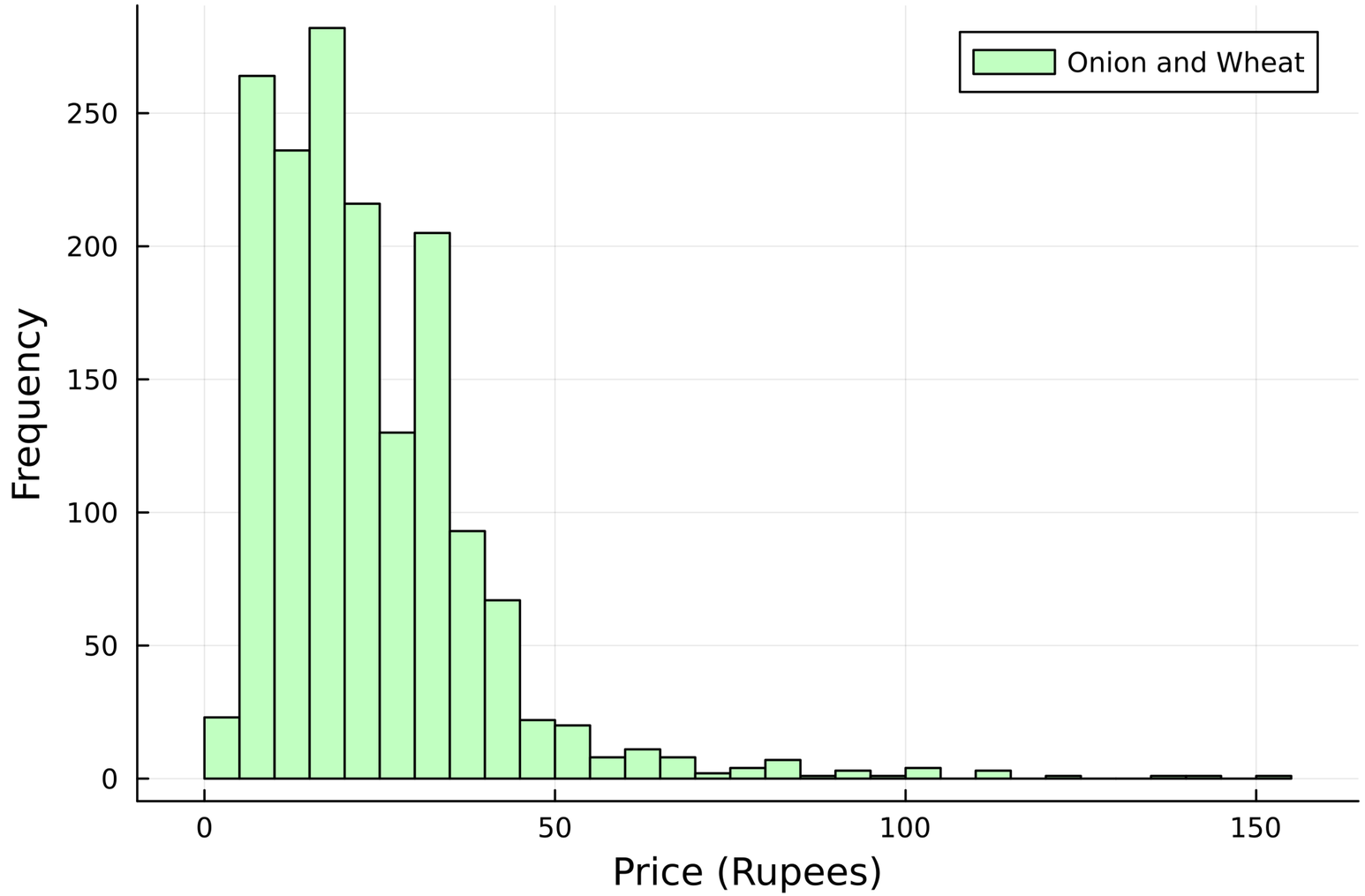
- `kerala` DataFrame:

Date	Centre	Commodity	Price
JAN-2001	Ernakulam	Onion	10.0
JAN-2001	Ernakulam	Wheat	12.5
JAN-2001	Khozhikode	Onion	9.0
JAN-2001	Khozhikode	Wheat	14.0
...	...	...	...
MAR-2021	Trivandrum	Onion	45.0
MAR-2021	Trivandrum	Wheat	34.0

- How much did the prices of onions and wheat differ?

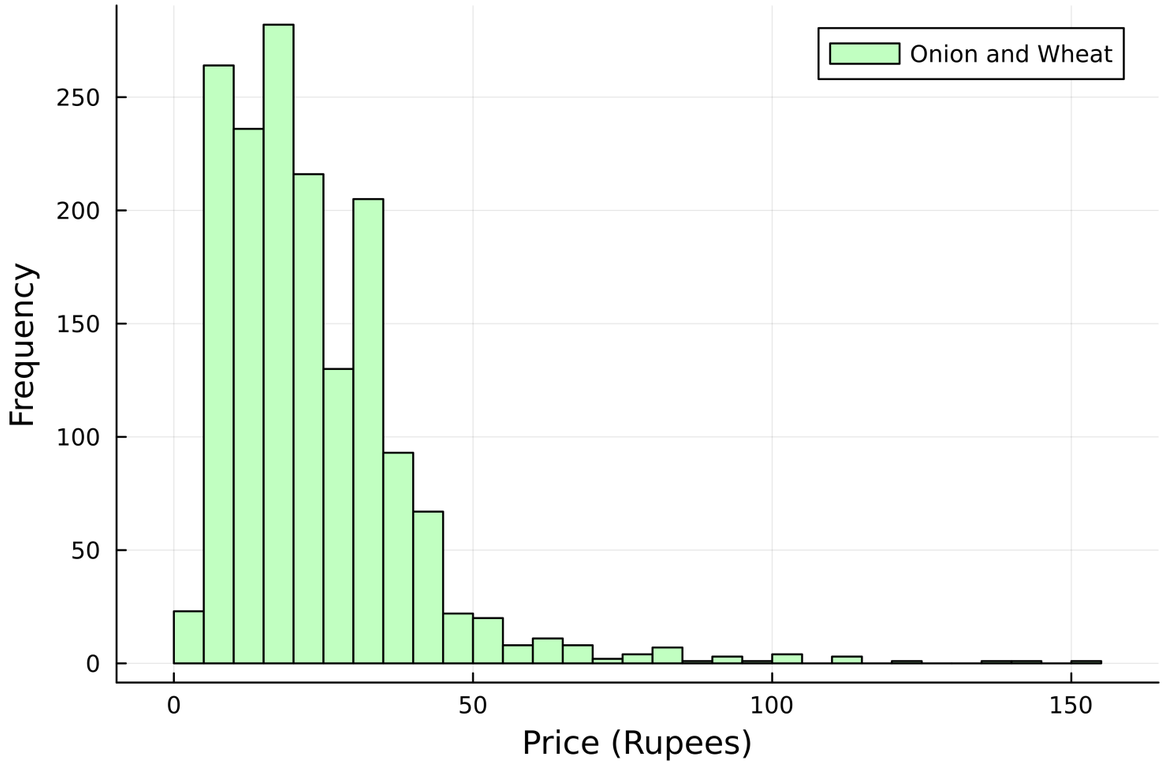
Commodity	Mean Price
Onion	25.7442
Wheat	20.6261

Visualizing distributions with histograms



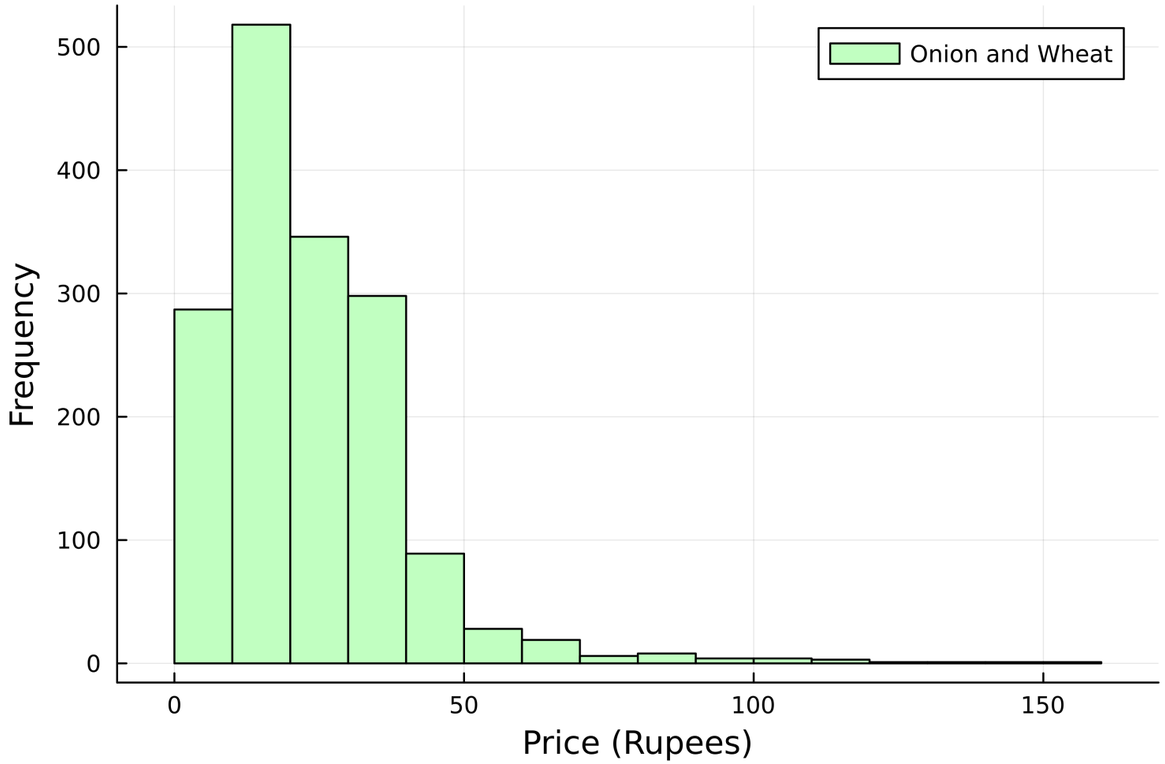
Distribution of onion and wheat prices

```
# Plot a histogram
histogram(
    kerala[:, :Price],
    # Add a label
    label="Onion and Wheat",
    # Choose bar color
    color=:darkseagreen1,
)
# Add axis labels
xlabel!("Price (Rupees)")
ylabel!("Frequency")
```



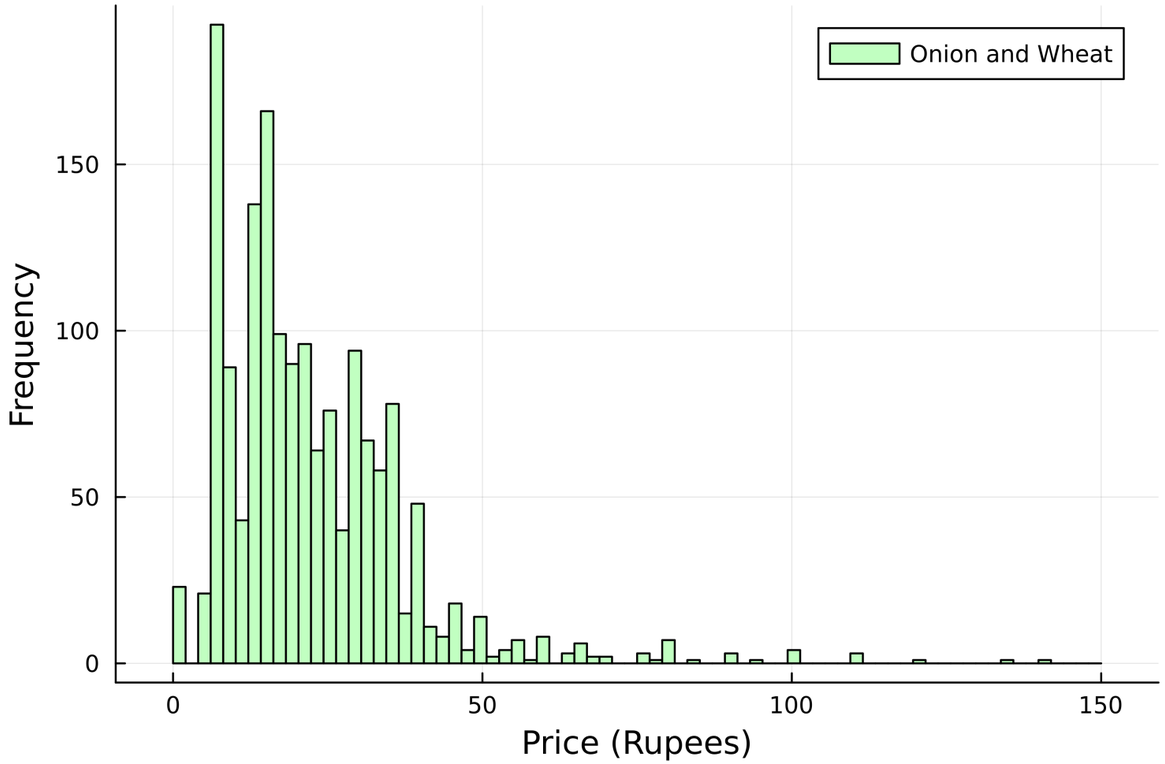
Number of bins

```
# Plot a histogram
histogram(
    kerala[:, :Price],
    # Add a label
    label="Onion and Wheat",
    # Choose bar color
    color=:darkseagreen1,
    # Number of bins
    bins=20,
)
# Add axis labels
xlabel!("Price (Rupees)")
ylabel!("Frequency")
```



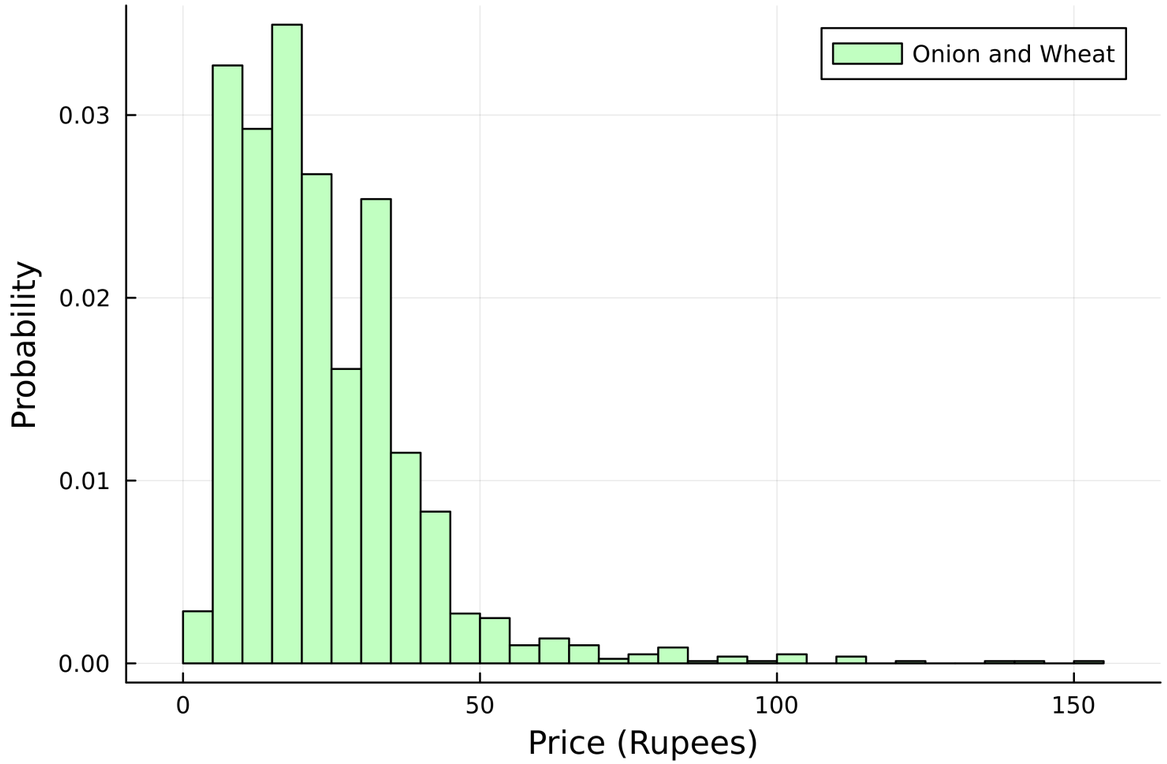
Number of bins

```
# Plot a histogram
histogram(
    kerala[:, :Price],
    # Add a label
    label="Onion and Wheat",
    # Choose bar color
    color=:darkseagreen1,
    # Number of bins
    bins=range(0, 150, 75),
)
# Add axis labels
xlabel!("Price (Rupees)")
ylabel!("Frequency")
```



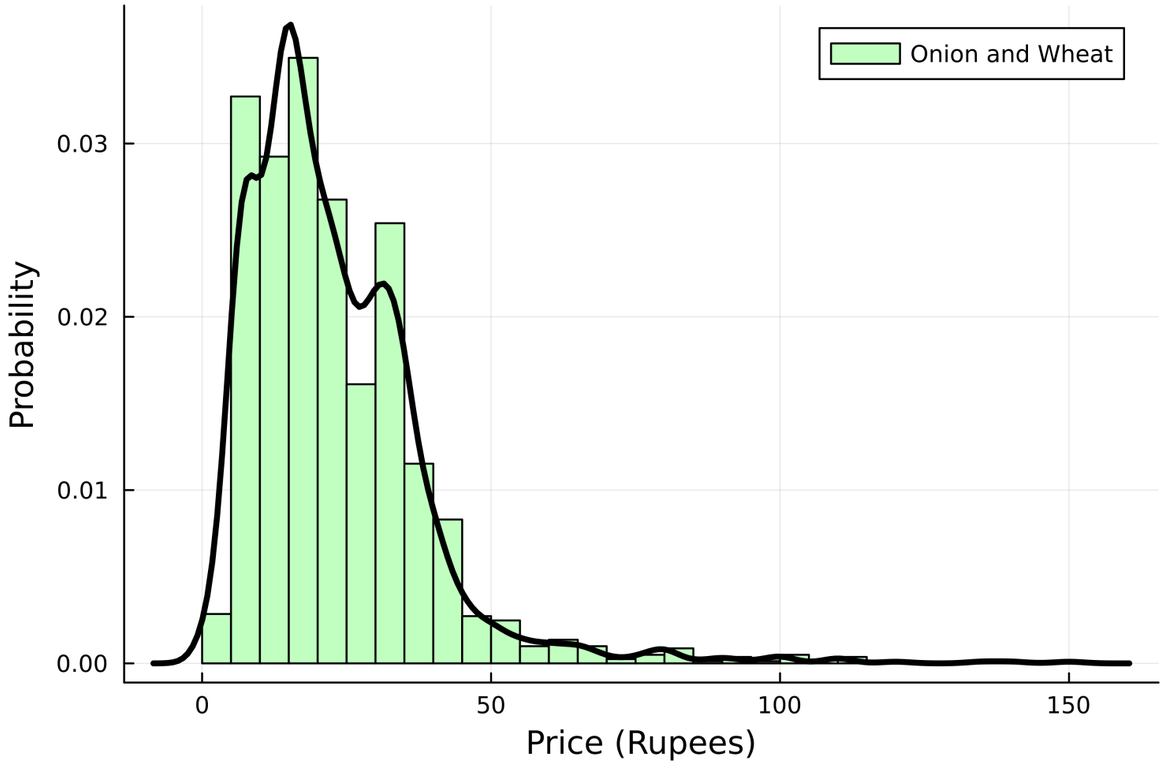
Normalized histogram

```
# Plot a normalized histogram
histogram(
    kerala[:, :Price],
    # Add a label
    label="Onion and Wheat",
    # Choose bar color
    color=:darkseagreen1,
    # Normalize it
    normalize=true,
)
# Add axis labels
xlabel!("Price (Rupees)")
ylabel!("Probability")
```



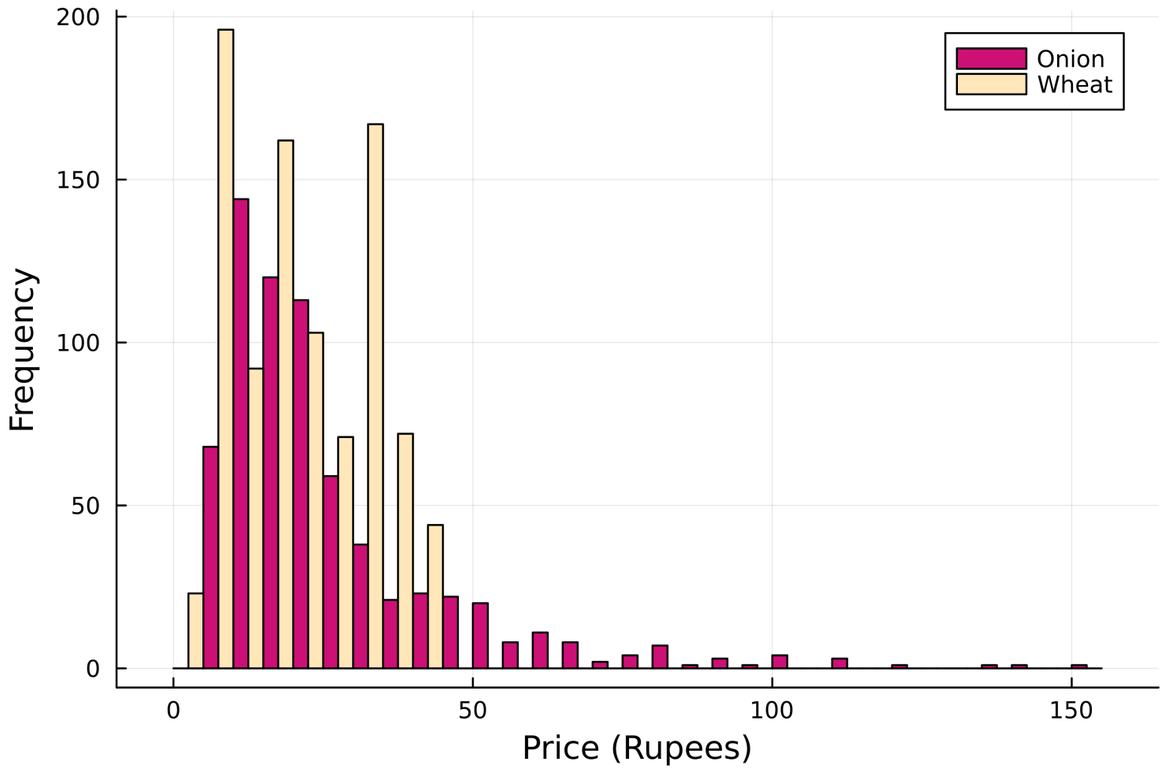
Probability distribution

```
using StatsPlots

density!(
    kerala[:, :Price],
    color=:black,
    linewidth=3,
    label=false
)
```



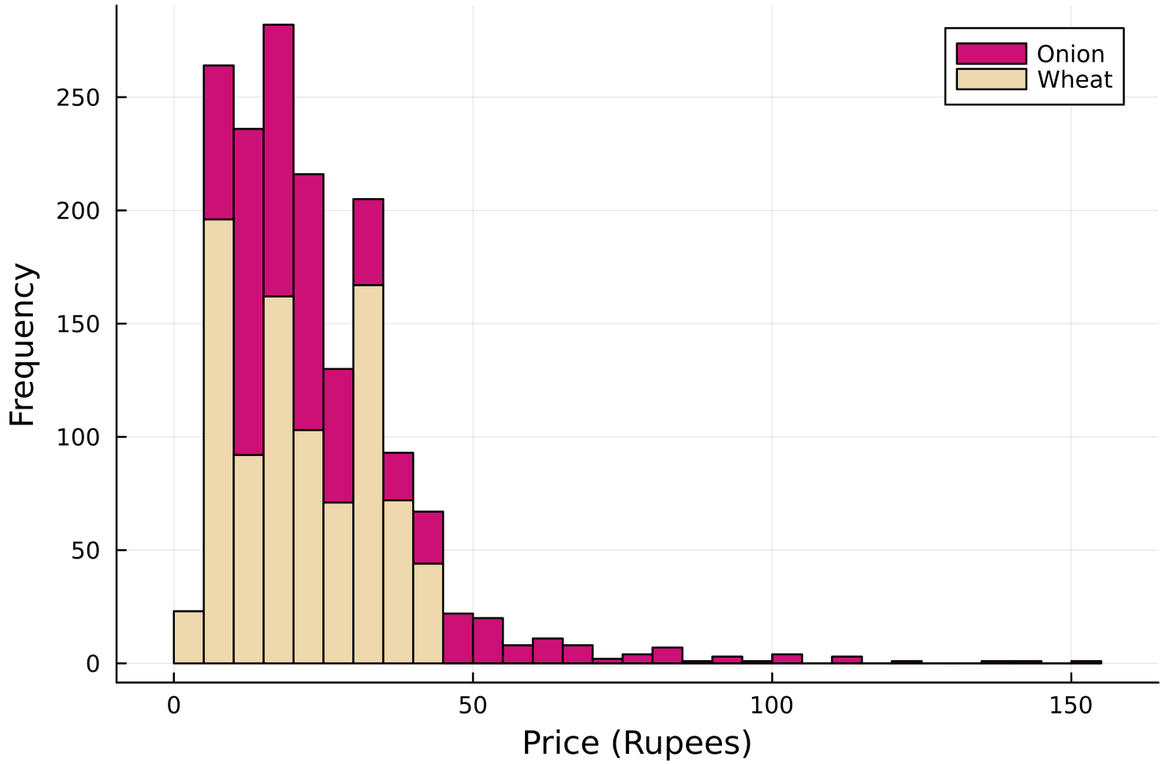
Prices per commodity

```
using StatsPlots

# Grouped histogram
groupedhist(
    kerala[:, :Price],
    # Group by commodity
    group=kerala[:, "Commodity"],
    # Select colors
    color=[:deeppink3 :wheat2]
)
xlabel!("Price (Rupees)")
ylabel!("Frequency")
```

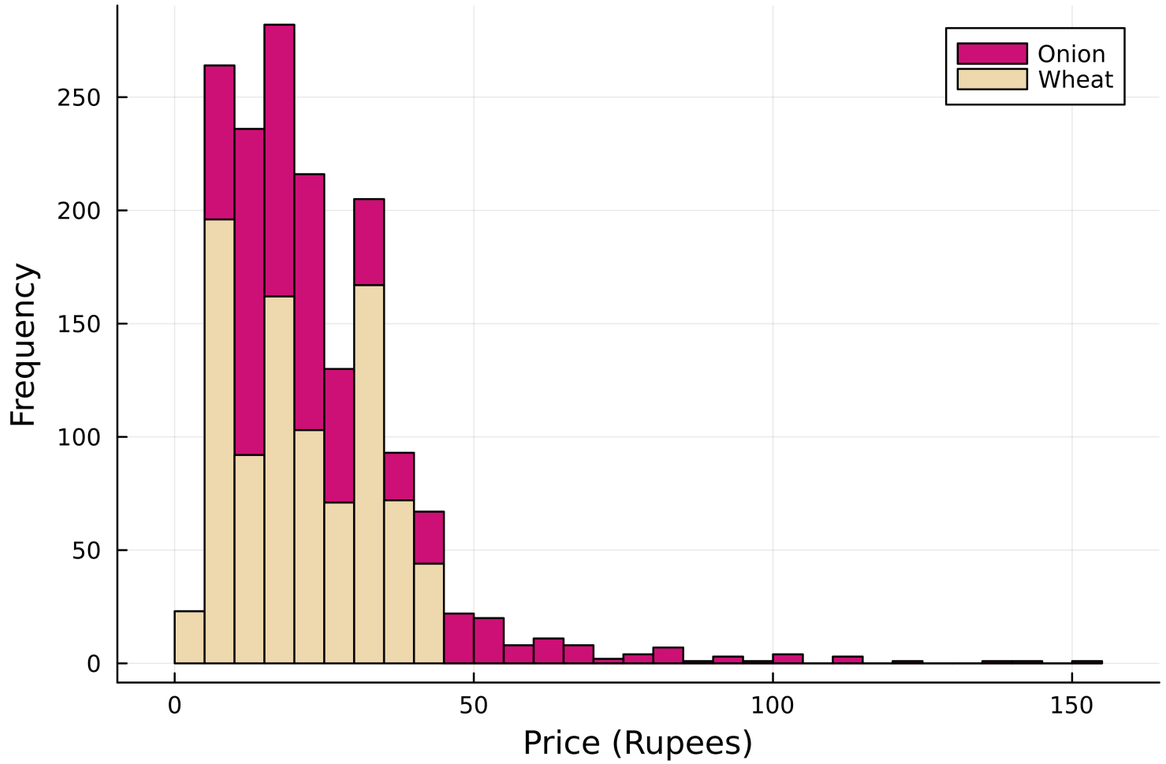


Stacked histogram

```
using StatsPlots

# Stacked histogram
groupedhist(
    kerala[:, :Price],
    # Group by commodity
    group=kerala[:, "Commodity"],
    # Select colors
    color=[:deeppink3 :wheat2]
    # Stack the bars
    bar_position=:stack,
)
xlabel!("Price (Rupees)")
ylabel!("Frequency")
```



A subtle difference



- The peak prices appear to be very similar.

Commodity	Mean Price
Onion	25.7442
Wheat	20.6261

- Onion prices exhibit a long tail.
- Median prices are almost the same.

Commodity	Median Price
Onion	20.0
Wheat	19.5

- Difference in means is cause by the tail!

Let's practice!

INTRODUCTION TO DATA VISUALIZATION WITH JULIA

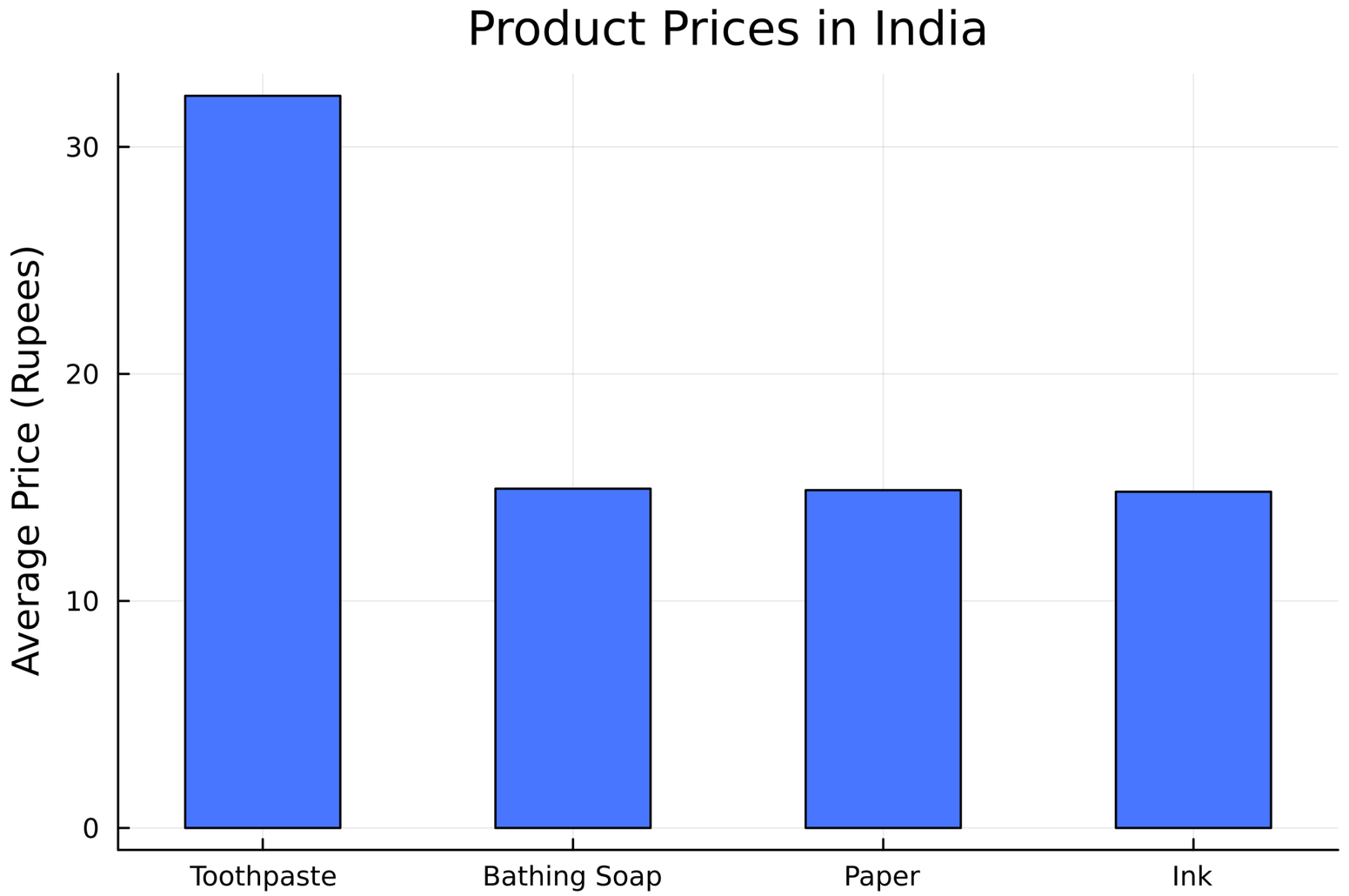
Dropping Bars

INTRODUCTION TO DATA VISUALIZATION WITH JULIA

Gustavo Vieira Suñe
Data Analyst

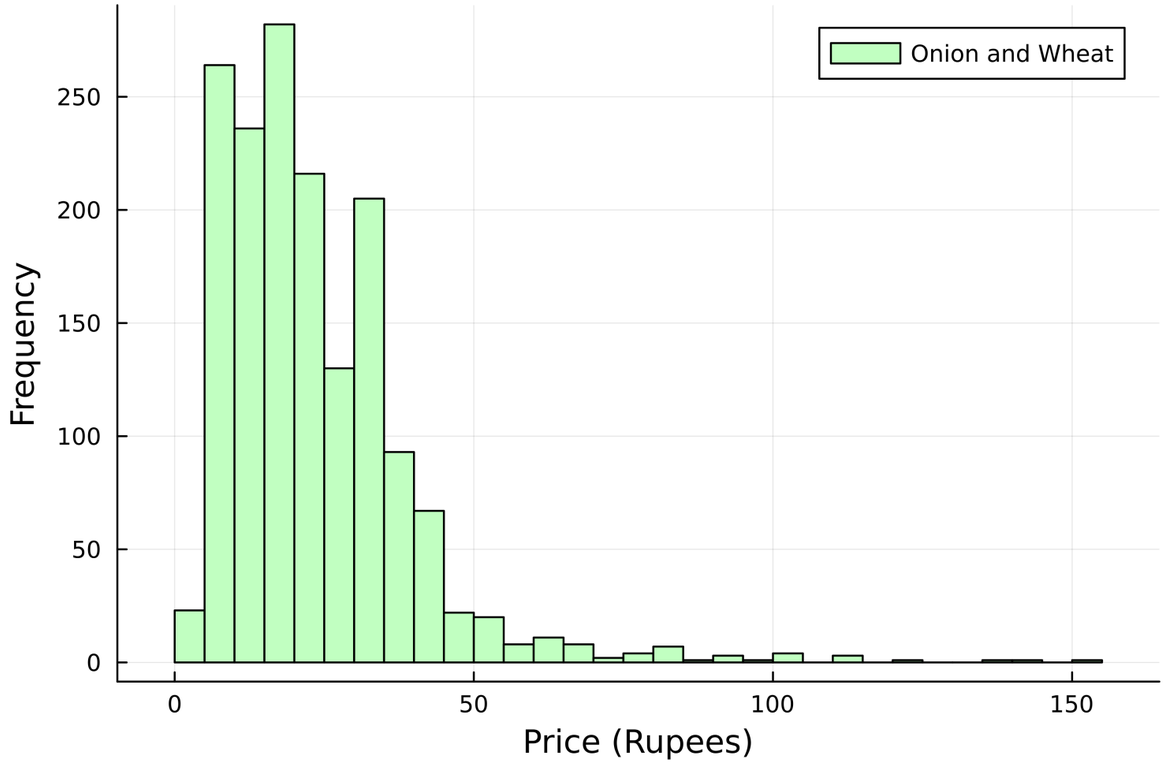
datacamp

Bar charts

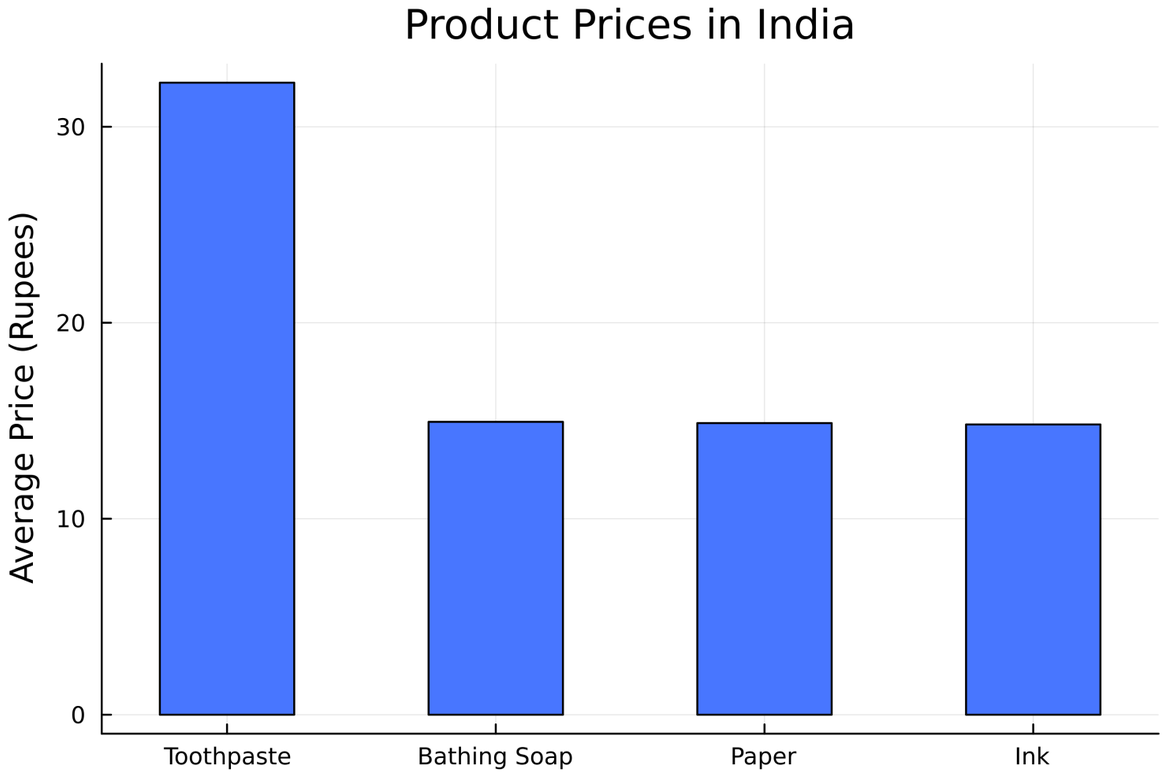

Product Prices in India
Average Price (Rupees)
0
10
20
30
Toothpaste
Bathing Soap
Paper
Ink

Bar charts vs histograms

- Distributions of numerical data



- Comparison of categories



Our dataset

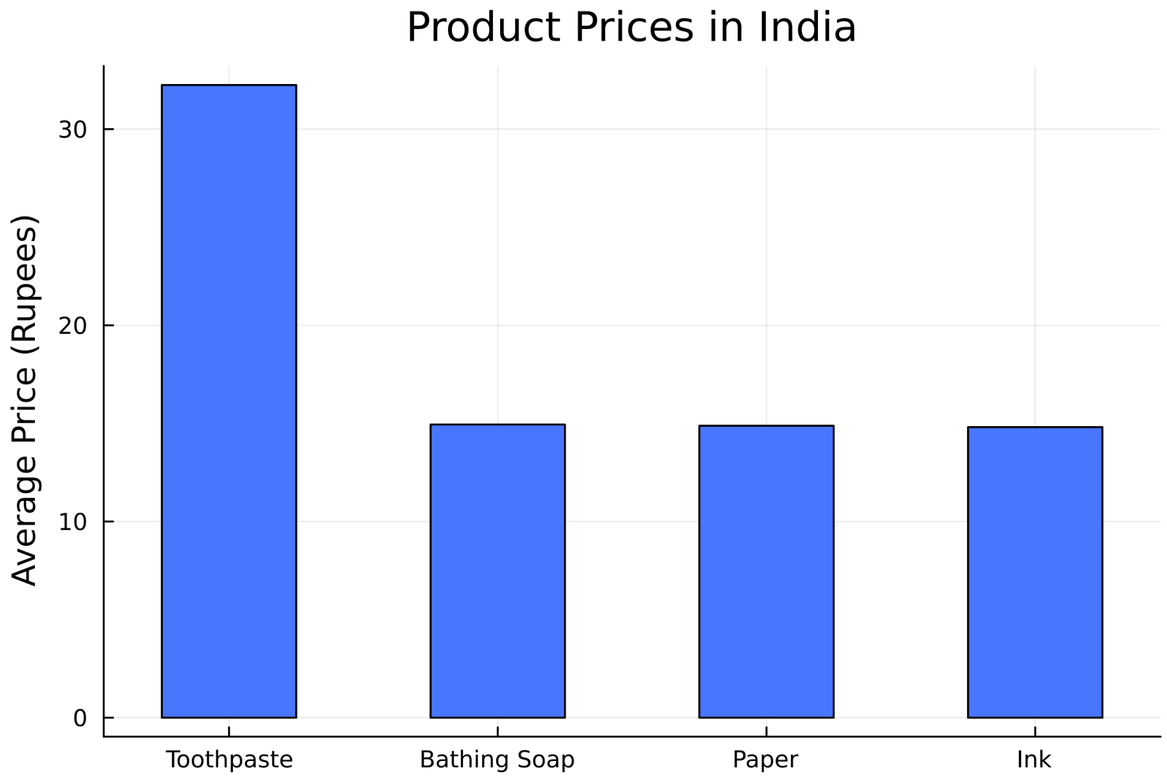
Date	State	Centre	Product	Price
APR-2001	Andhra Pradesh	Chittoor	Bathing Soap	10.0
APR-2001	Andhra Pradesh	Chittoor	Ink	10.0
...	...	...	...	...
SEP-2016	Bihar	Patna	Paper	38.0
SEP-2016	Bihar	Patna	Toothpaste	50.0

- Mean prices (`using Statistics`)

```
# Group by each product
grouped = groupby(product, :Product)
# Calculate average prices
mean_prices = combine(
    grouped,
    :Price => mean
)
# Sort from higest to lowest
sorted_mean_prices = sort(
    mean_prices,
    :Price_mean,
    rev=true
)
```

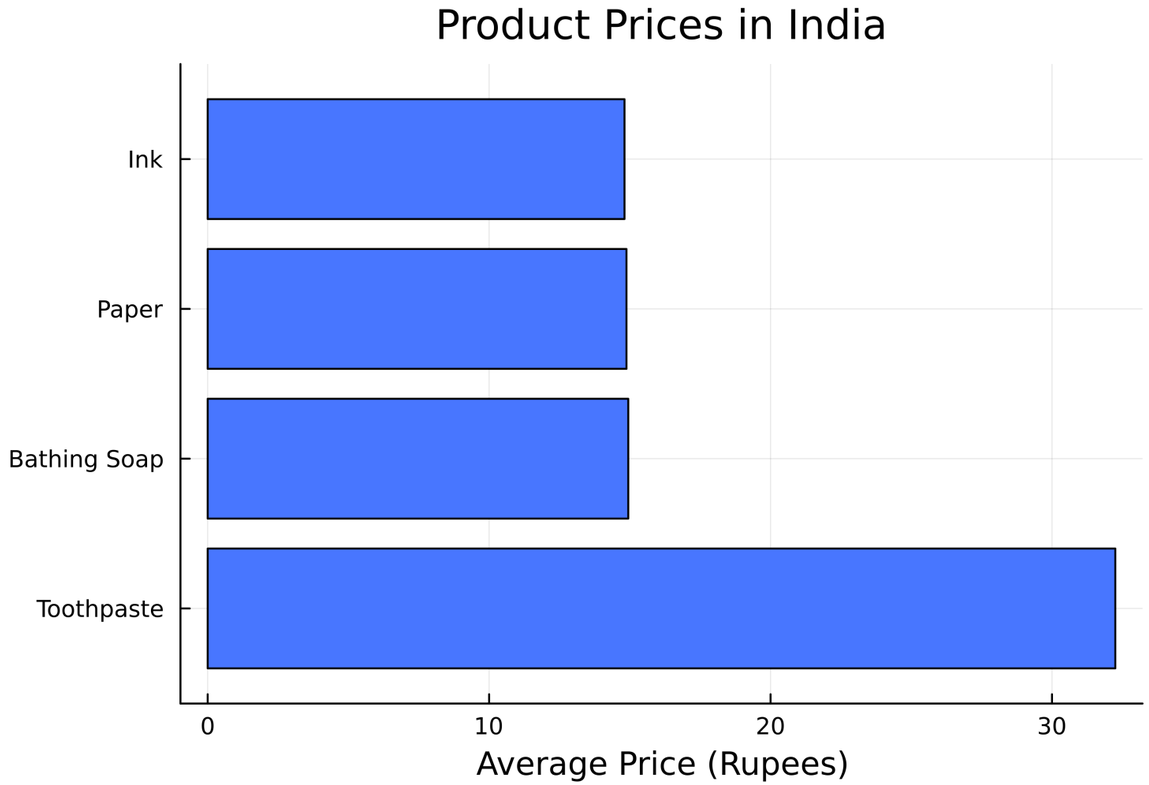
Creating bar charts

```
# Create bar chart
bar(
    # Categories in x-axis
    sorted_mean_prices.Product,
    # Values in y-axis
    sorted_mean_prices.Price_mean,
    # Set bar width
    bar_width=0.5
    label=false,
    color=:royalblue1,
 )
title!("Product Prices in India")
ylabel!("Average Price (Rupees)")
```



Horizontal bar charts

```
# Create bar chart
bar(
    # Values in the y-axis
    sorted_mean_prices.Product,
    # Values in x-axis
    sorted_mean_prices.Price_mean,
    # Set orientation to horizontal
    permute=(:x, :y),
    label=false,
    color=:royalblue1,
)
title!("Product Prices in India")
xlabel!("Average Price (Rupees)")
```



Products by state

- Filter for selected states

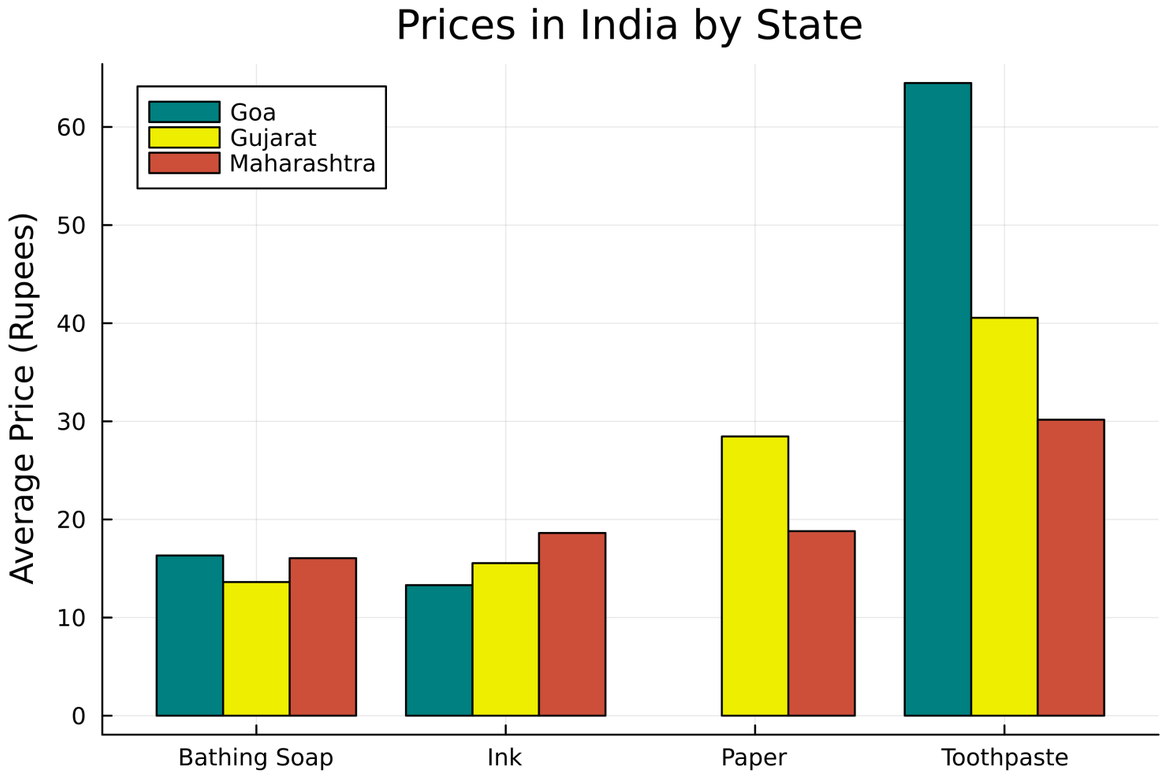
```
filtered_product = filter(
    row -> row.State in ["Maharashtra", "Goa", "Gujarat"],
    product)
```

- Average prices of each product by state

```
# Group by state and product
filtered_grouped = groupby(filtered_product, [:State, :Product])
# Calculate average prices
filtered_mean_prices = combine(
    filtered_grouped, :Price => mean
)
```

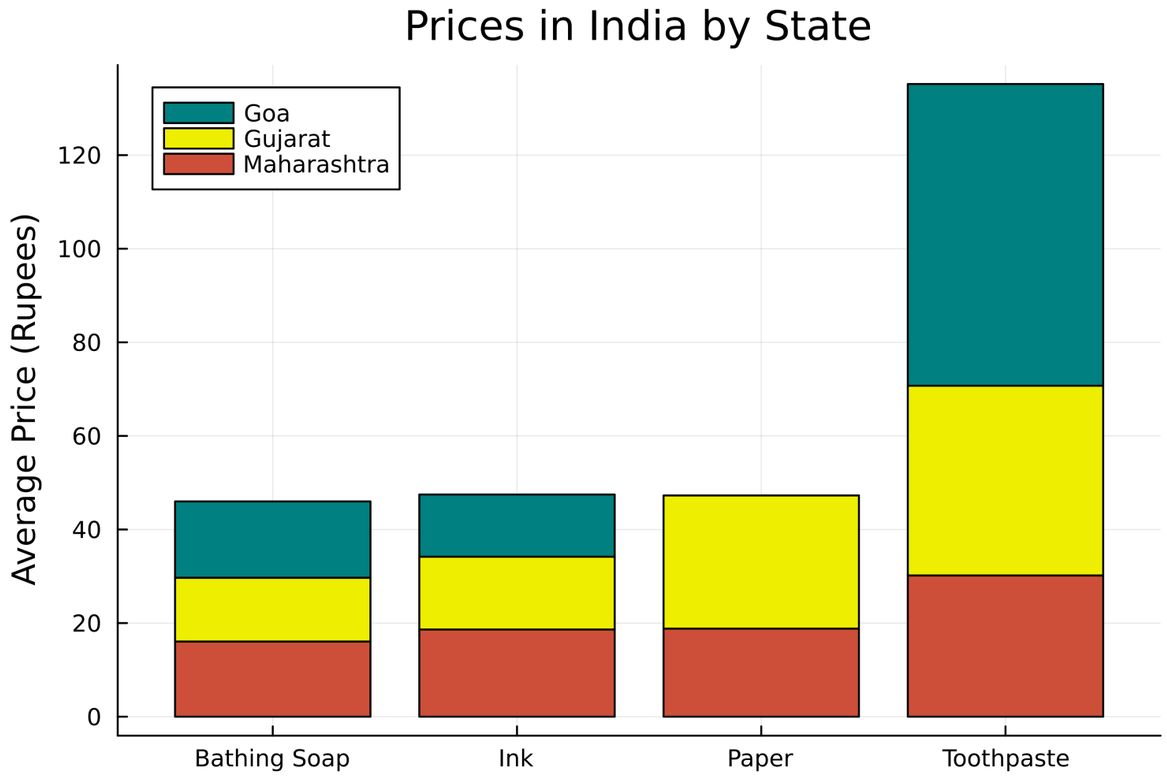
Grouped bar charts

```
# Create grouped bar chart
groupedbar(
    # Categories in x-axis
    filtered_mean_prices.Commodity,
    # Values in y-axis
    filtered_mean_prices.Price_mean,
    # Define groups
    group=filtered_mean_prices.State,
    color=[:teal :yellow2 :tomato3]
)
title!("Prices in India by State")
ylabel!("Average Price (Rupees)")
```



Stack the bars

```
# Create grouped bar chart
groupedbar(
    # Categories in x-axis
    filtered_mean_prices.Commodity,
    # Values in y-axis
    filtered_mean_prices.Price_mean,
    # Define groups
    group=filtered_mean_prices.State,
    color=[:teal :snow2 :tomato3],
    # Stack the bars
    bar_position=:stack
)
title!("Product Prices in India by State")
ylabel!("Average Price (Rupees)")
```



Let's practice!

INTRODUCTION TO DATA VISUALIZATION WITH JULIA

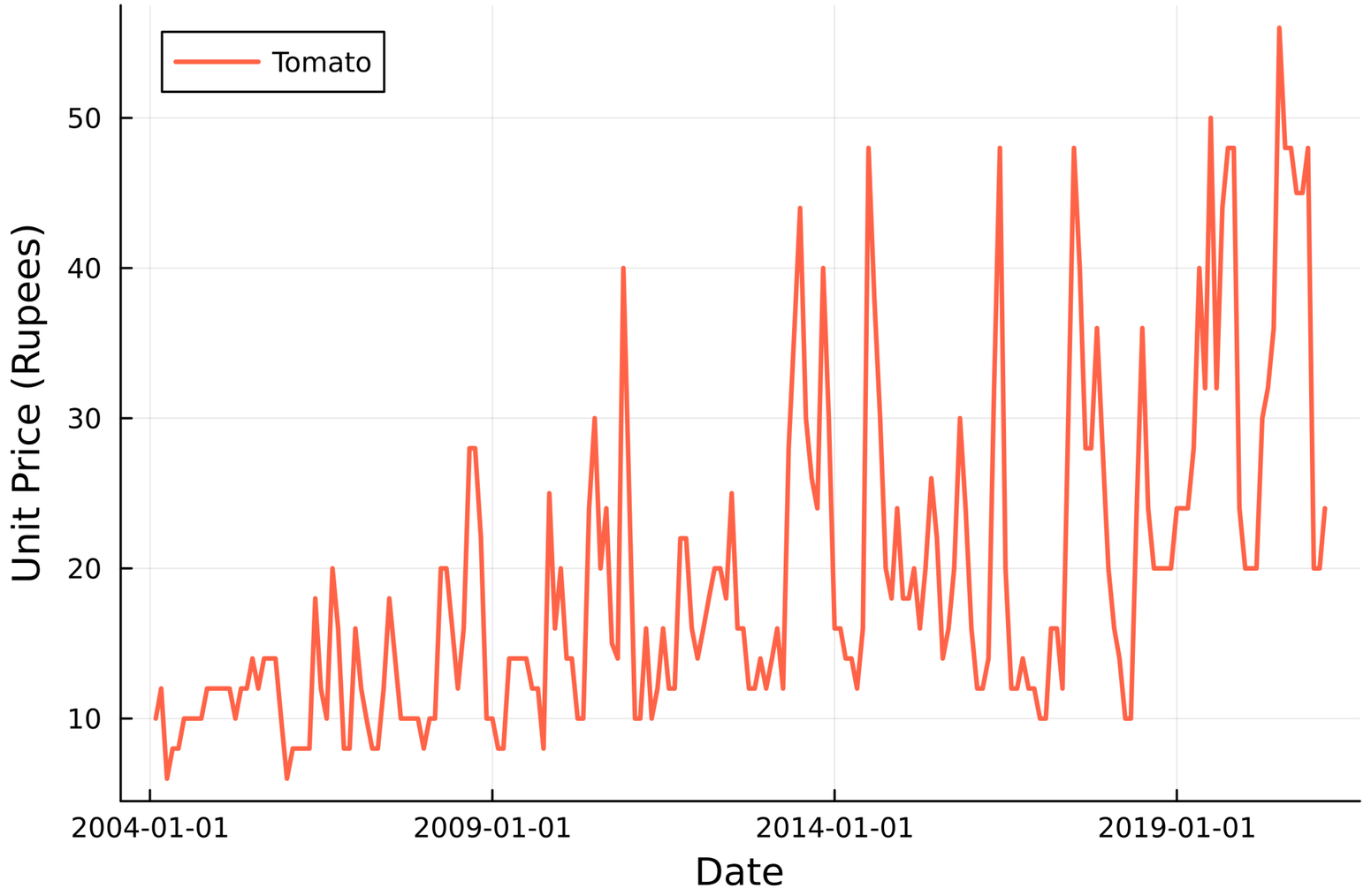
Visualizing time series

INTRODUCTION TO DATA VISUALIZATION WITH JULIA

Gustavo Vieira Suñe
Data Analyst

datacamp

Time series


Tomato
Unit Price (Rupees)
10
20
30
40
50
2004-01-01
2009-01-01
2014-01-01
2019-01-01
Date

Tomato prices

- `tomato` DataFrame

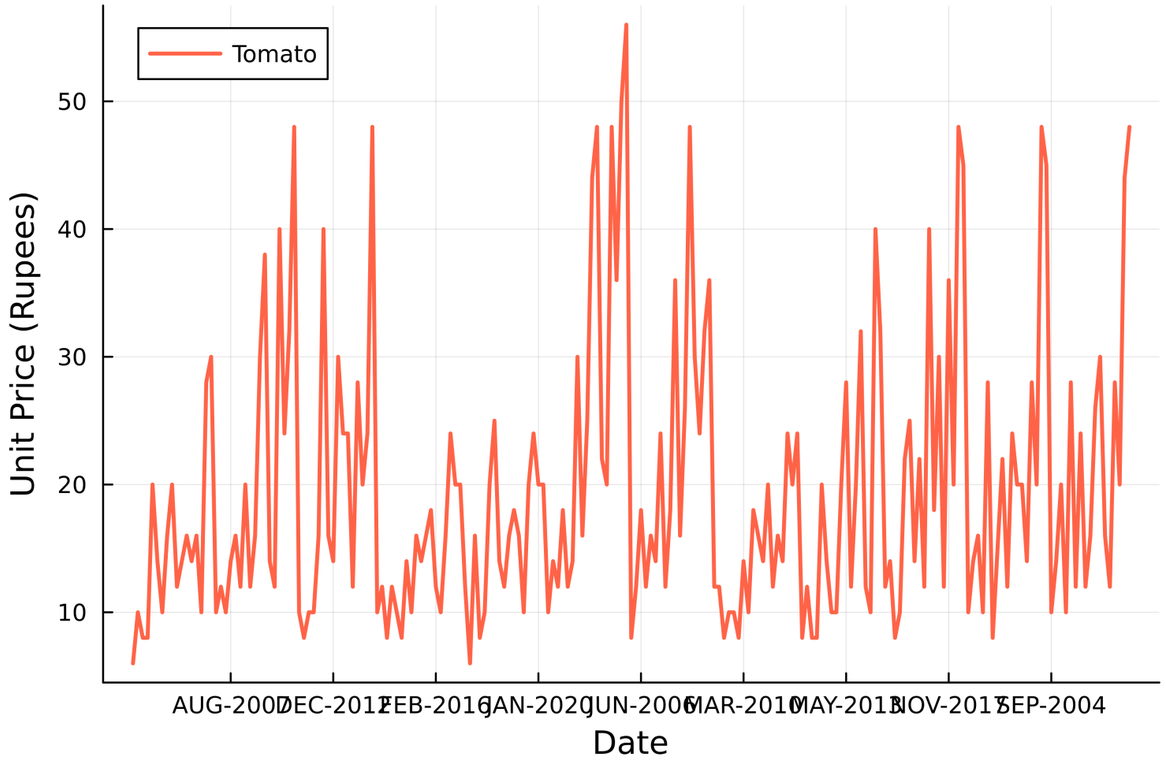
Date	Price (Rupees)
FEB-2004	10.0
MAR-2004	12.0
APR-2004	6.0
MAY-2004	8.0
...	...

- Sort dates

```
tomato.Date = sort(
  tomato.Date, :Date
)
```

```
# Plot time series
plot(
    tomato.Date,
    tomato.Price,
    # Customize the plot
    linewidth=2,
    linecolor=:tomato,
    label="Tomato"
)
xlabel!("Date")
ylabel!("Unit Price (Rupees)")
```

Tomato prices



- `tomato` DataFrame

Date	Price (Rupees)
APR-2004	6.0
APR-2005	10.0
APR-2006	8.0
APR-2007	8.0
...	...

- Date column has strings!

Dates with Julia

```
using Dates

birthday = Date("1989-12-04")
```

- Other formats

```
birthday = Date(
    "1989/DEC/04", dateformat"y/u/d"
)
birthday = Date(
    "Dec 4, 1989", dateformat"u d, y"
)
```

Code	Match	Examples
Y/y	Year (YYYY)	1989, 2023
m	Month (MM)	1, 10
u	Abbreviated Month	Jan, DEC
U	Month Name	January, DECEMBER
d	Day (DD)	4, 28
H	Hour (HH)	12, 22
M	Minute (MM)	05, 25
S	Second (SS)	10, 59

[1] https://docs.julialang.org/en/v1/stdlib/Dates/#Period-Types

Tomato prices with Dates

- Convert strings to Dates

```
tomato.Date = Date.(
    tomato.Date, dateformat"u-y"
)
```

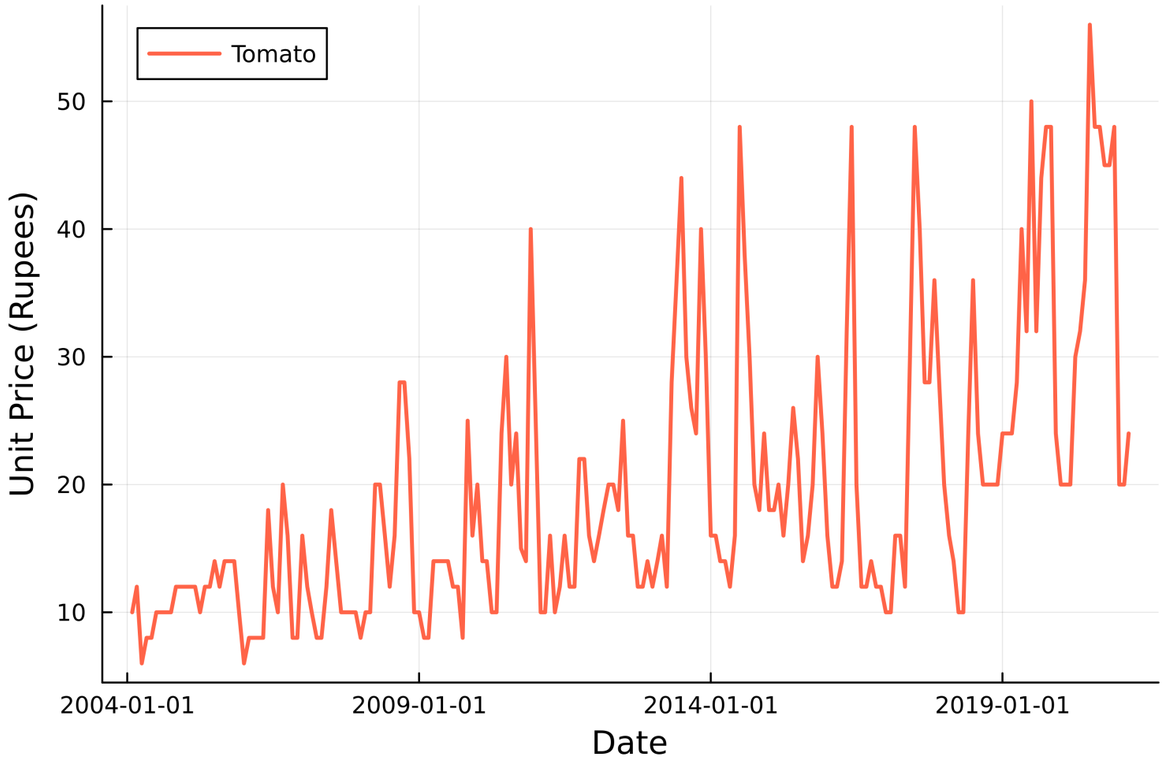
- Sort by date

```
tomato.Date = sort(
    tomato.Date, :Date
)
```

Date	Price (Rupees)
2004-02-01	10.0
2004-03-01	12.0
2004-04-01	6.0
2004-05-01	8.0
...	...

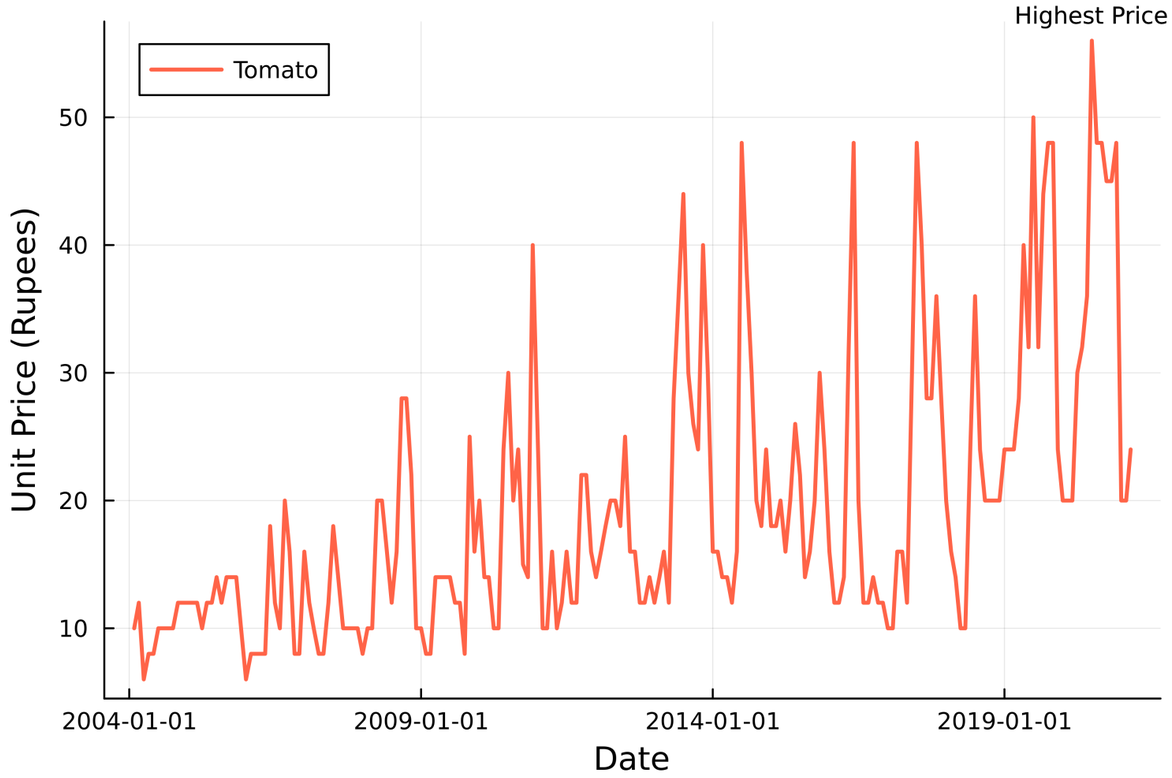
Tomato price time series

```
# Plot time series
plot(
    tomato.Date,
    tomato.Price,
    # Customize the plot
    linewidth=2,
    linecolor=:tomato,
    label="Tomato"
)
xlabel!("Date")
ylabel!("Unit Price (Rupees)")
```



Annotating a plot

```
# Row with highest price
maximum_price = tomato[
    argmax(tomato.Price), :]
# Annotate the plot
annotate!(
    # Coordinates
    maximum_price.Date,
    maximum_price.Price + 2,
    # Annotation text
    "Highest Price",
    annotationfontsize=8
)
```



Let's practice!

INTRODUCTION TO DATA VISUALIZATION WITH JULIA

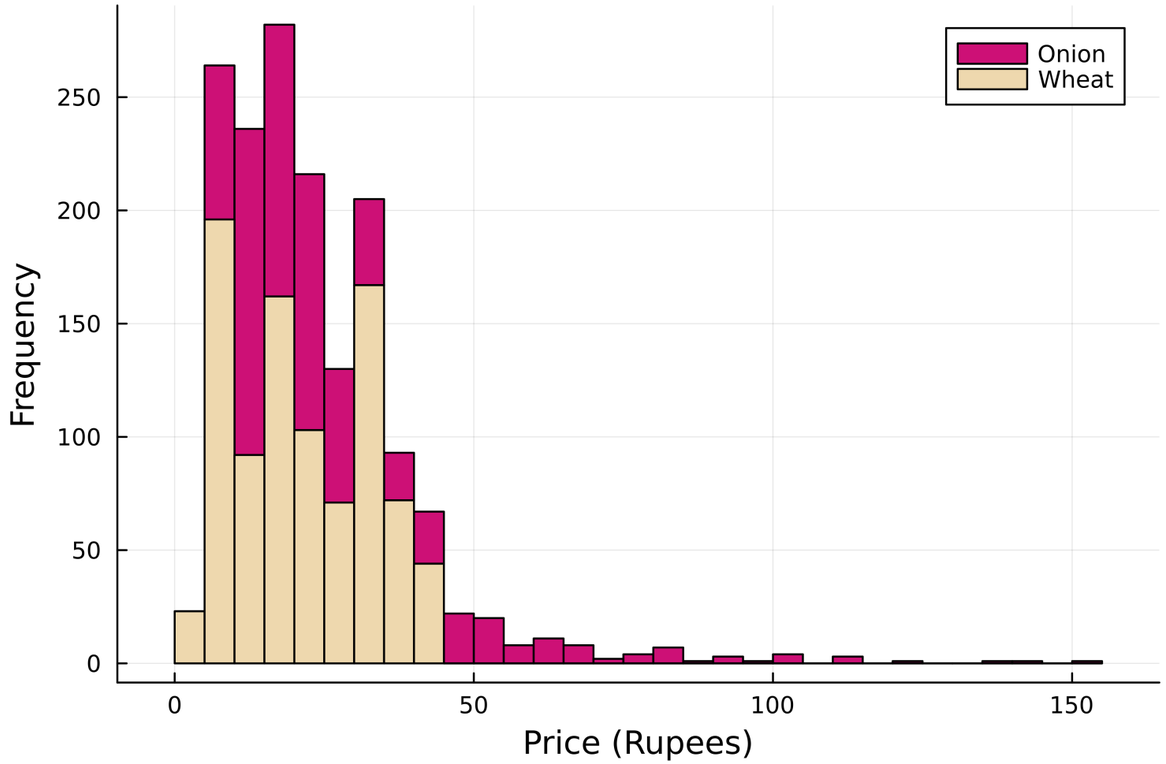
Boxes and violins

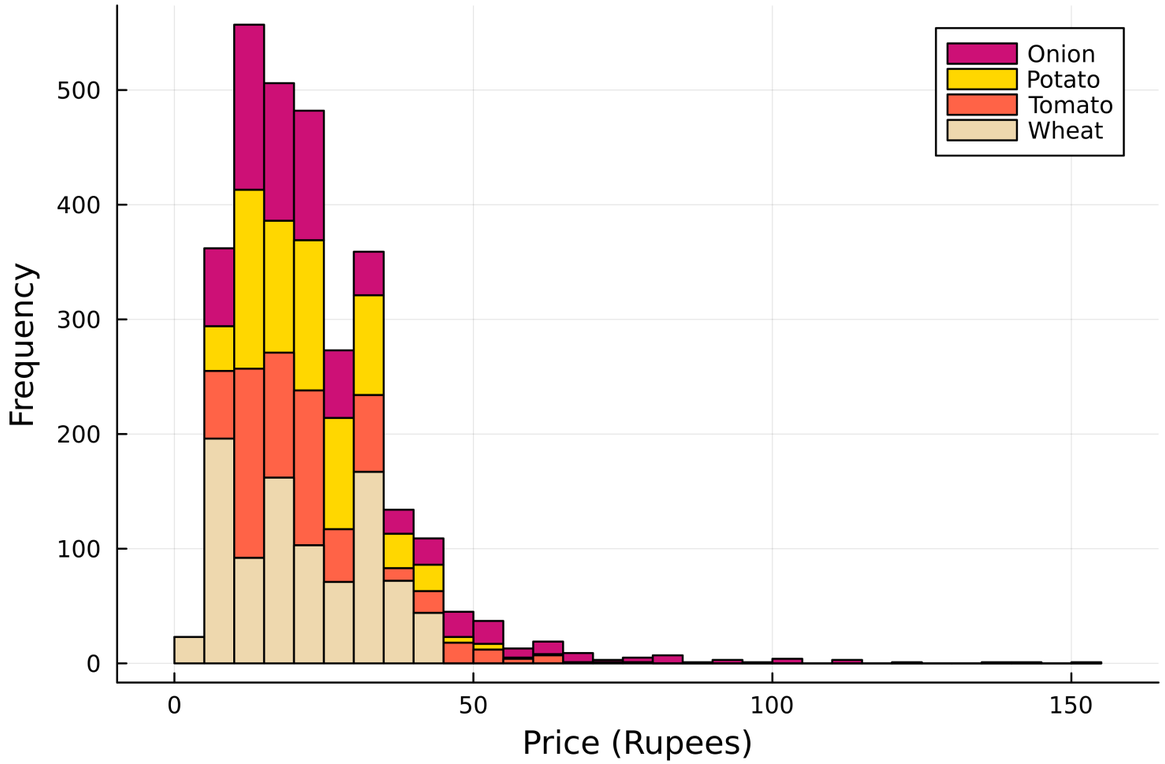
INTRODUCTION TO DATA VISUALIZATION WITH JULIA

Gustavo Vieira Suñe
Data Analyst

datacamp

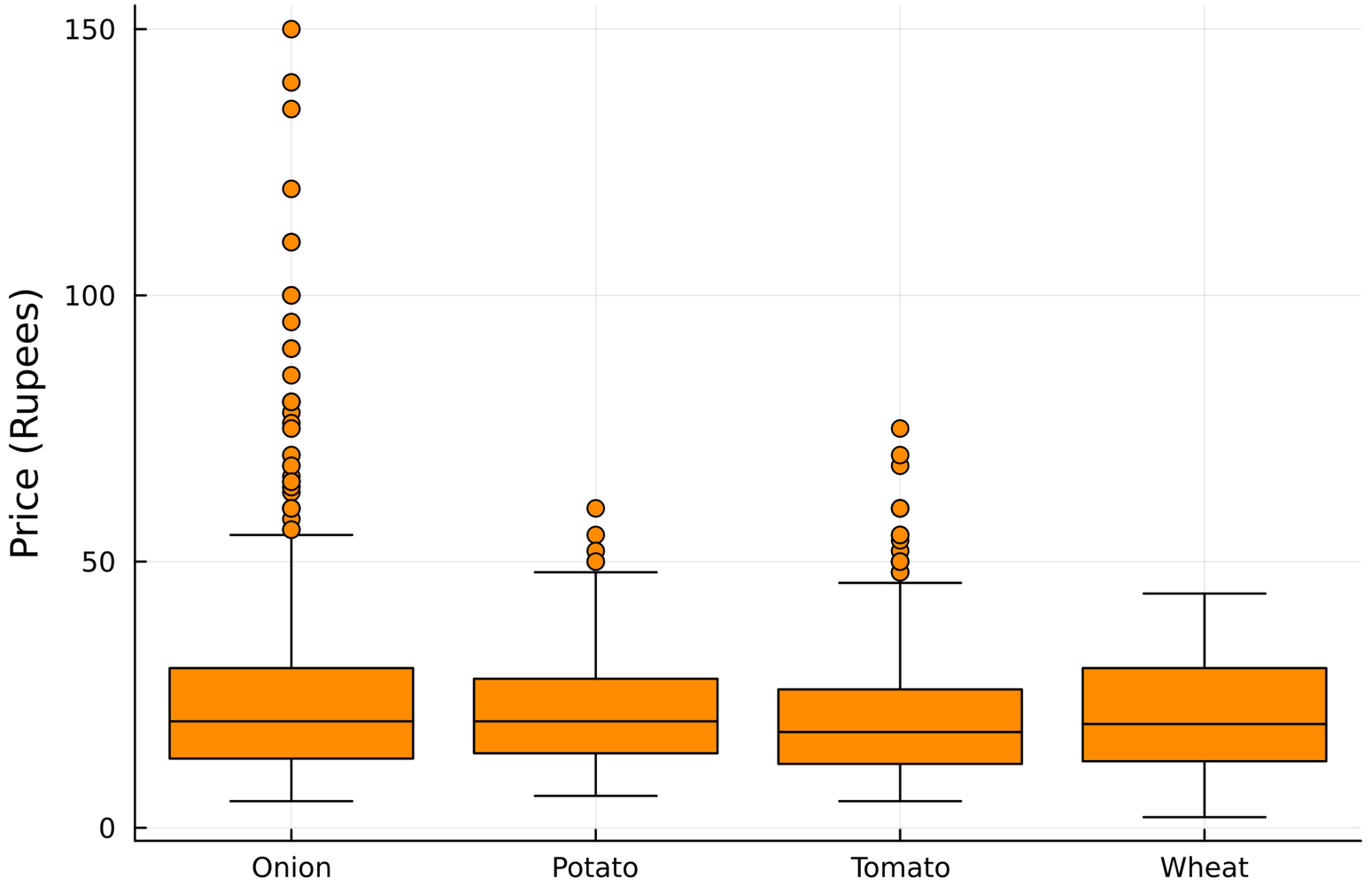
Histograms with many categories





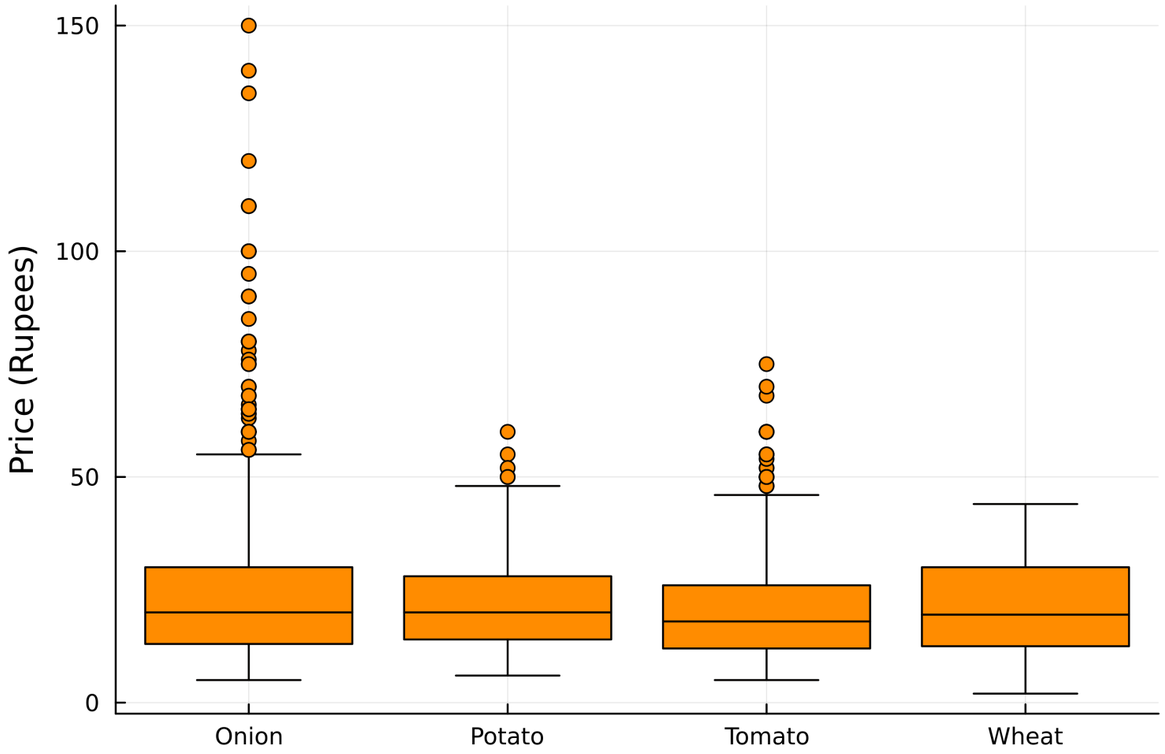
- Hard to compare many categories

Box plots



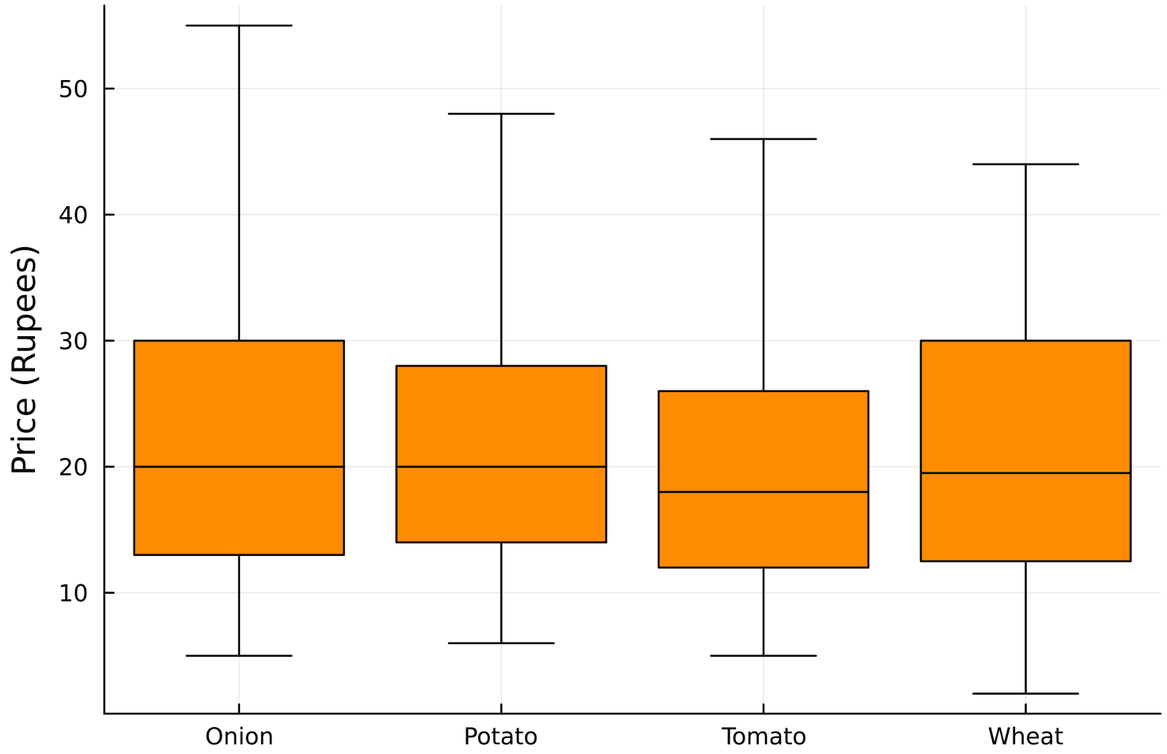
Boxes of product

```
using StatsPlots


boxplot(
    # Categories in x-axis
    kerala[:, :Commodity],
    # Values in y-axis
    kerala[:, :Price],
    color=:darkorange,
    label=false,
)
ylabel!("Price (Rupees)")
```

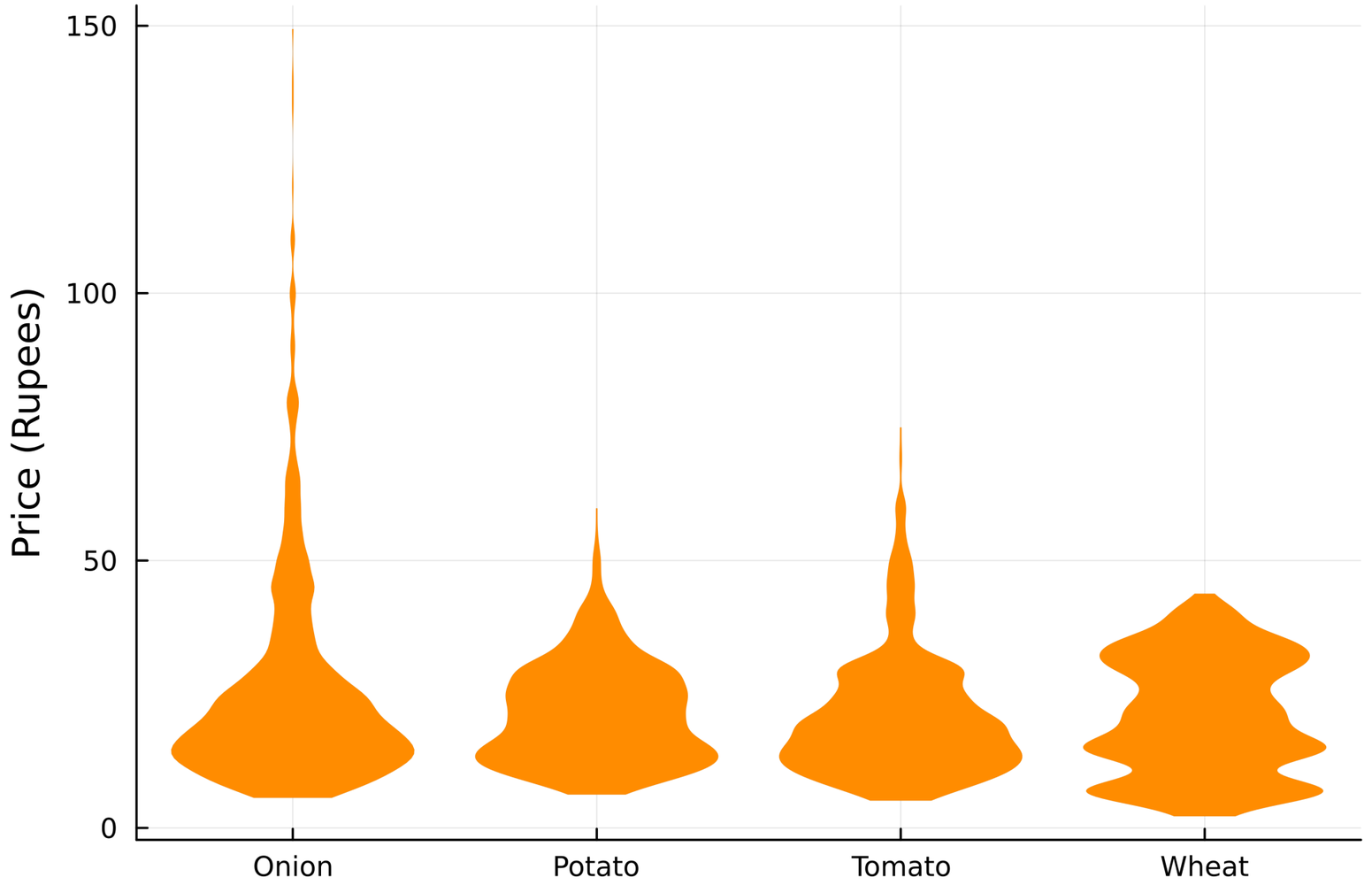


Hiding outliers

```
using StatsPlots

boxplot(
    # Categories in x-axis
    kerala[:, "Commodity"],
    # Values in y-axis
    kerala[:, :Price],
    color=:darkorange,
    label=false,
    # Hide outliers
    outliers=false,
)
ylabel!("Price (Rupees)")
```

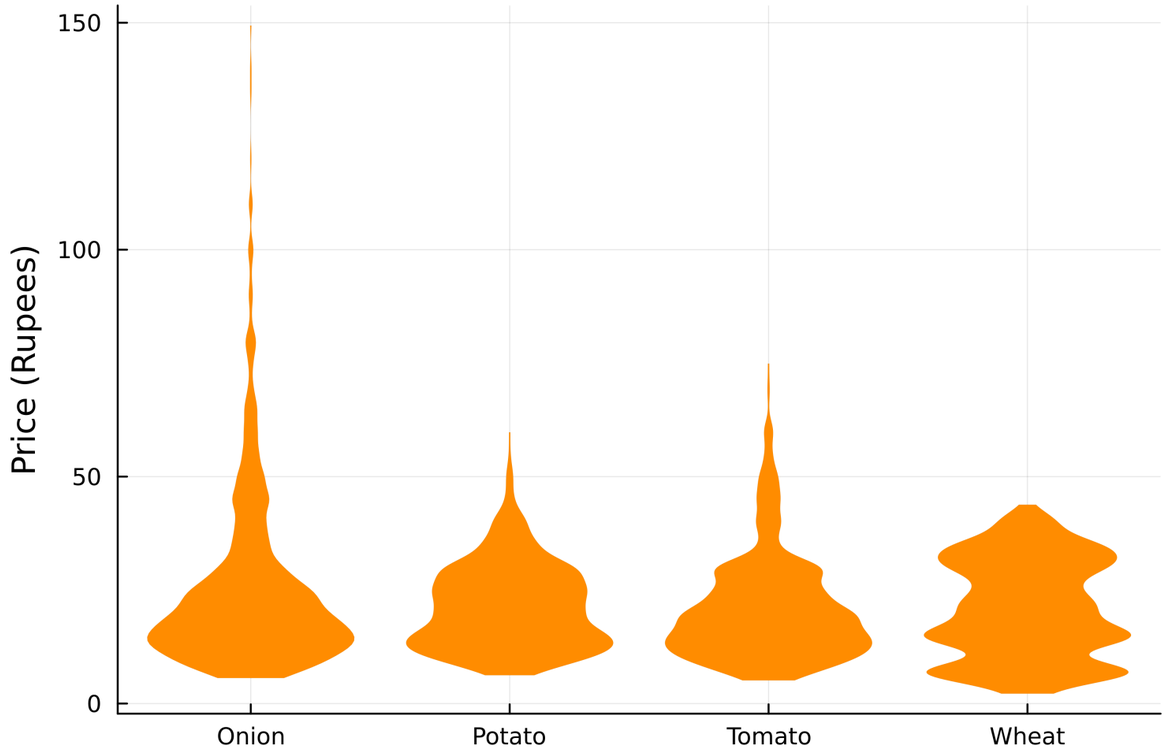


Violin plots



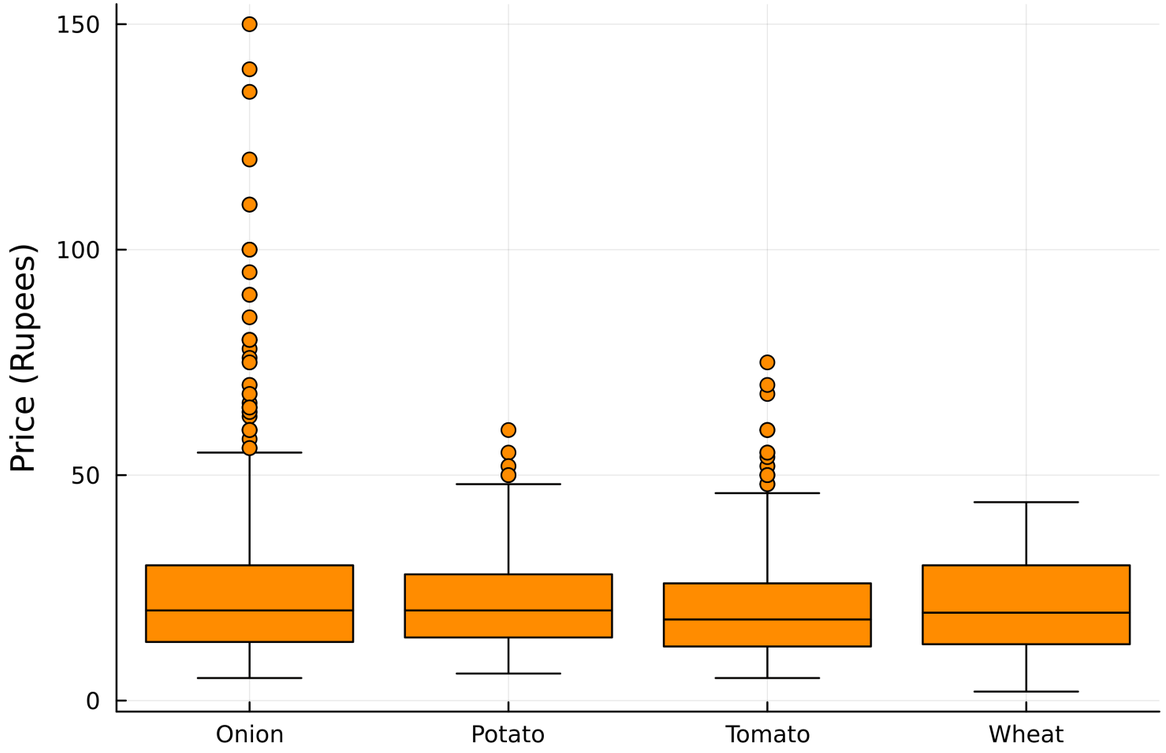
Distributions with violin plot

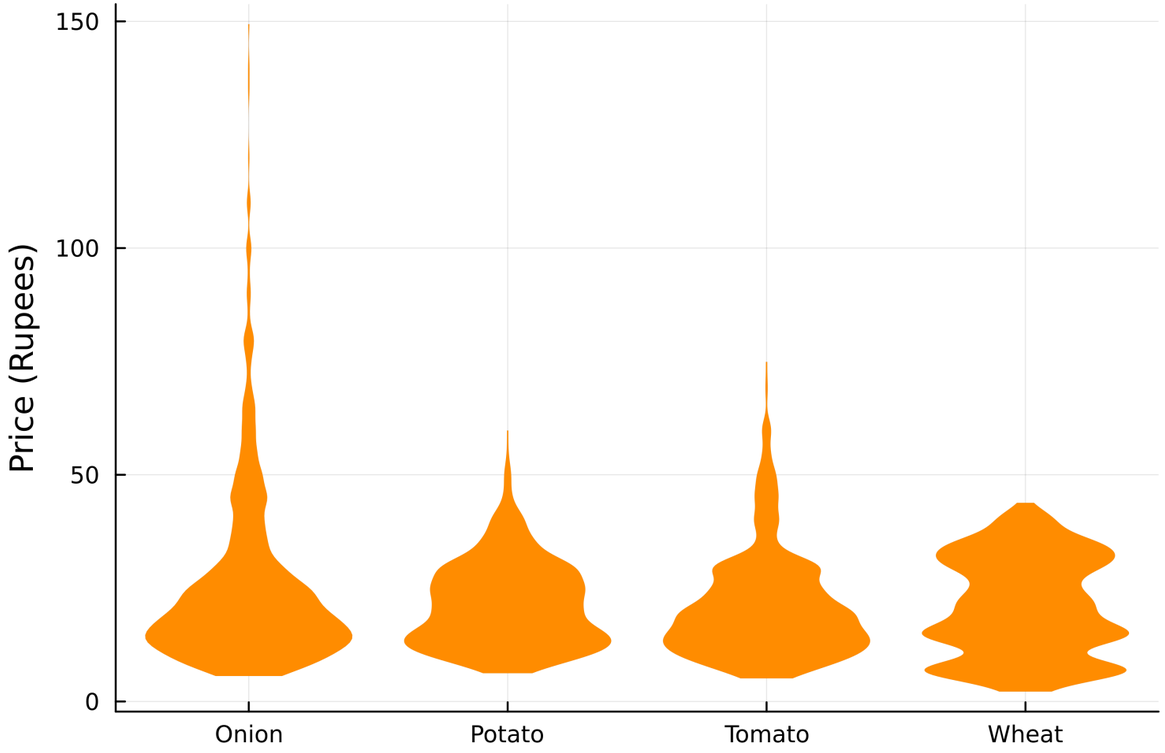
```
# Create violin plot
violin(
    # Categories in x-axis
    kerala[:, :Commodity],
    # Values in y-axis
    kerala[:, :Price],
    # Remove lines
    linewidth=0,
    color=:darkorange,
    label=false,
)
ylabel!("Price (Rupees)")
```



Boxes or violins?

- Box plots
  - Central tendency, spread & skewness
  - Emphasize outliers
- Violin plots
  - Details of distribution
  - Shape is relevant (e.g., multimodal)





Let's practice!

INTRODUCTION TO DATA VISUALIZATION WITH JULIA