```
# eda_markets_complete.py

import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from google.colab import files

# Upload dataset
uploaded = files.upload()

# Load dataset (ganti nama file jika beda)
df = pd.read_csv('markets_cleaned.csv')

# Set style visualisasi
sns.set(style="whitegrid")

# 1. Informasi Umum Dataset
print("=== Informasi Dataset ===")
print(df.info())
print(df.describe(include='all'))
print("Nilai kosong per kolom:\n", df.isnull().sum())

# 2. Jumlah Pasar per Negara Bagian
plt.figure(figsize=(12, 6))
state_counts = df['state'].value_counts()
sns.barplot(x=state_counts.index, y=state_counts.values, palette="viridis")
plt.title("Jumlah Pasar per Negara Bagian")
plt.xlabel("Negara Bagian")
plt.ylabel("Jumlah Pasar")
plt.xticks(rotation=90)
plt.tight_layout()
plt.show()

# 3. Produk Terpopuler
product_columns = [
    'Bakedgoods', 'Cheese', 'Crafts', 'Flowers', 'Eggs', 'Seafood',
    'Herbs', 'Vegetables', 'Honey', 'Jams', 'Maple', 'Meat', 'Nursery',
    'Nuts', 'Plants', 'Poultry', 'Prepared', 'Soap', 'Trees', 'Wine', 'Fruits'
]

product_counts = df[product_columns].sum().sort_values(ascending=False)
plt.figure(figsize=(12, 6))
sns.barplot(x=product_counts.index, y=product_counts.values, palette="magma")
plt.title("Produk Terpopuler")
plt.xlabel("Produk")
plt.ylabel("Jumlah Pasar yang Menjual")
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()

# 4. Lama Pasar Buka vs Jumlah Produk Dijual
plt.figure(figsize=(10, 6))
sns.scatterplot(data=df, x='months_open', y='num_items_sold')
plt.title("Lama Pasar Buka vs Jumlah Produk Dijual")
plt.xlabel("Lama Pasar Buka (bulan)")
plt.ylabel("Jumlah Produk Dijual")
plt.tight_layout()
plt.show()

# 5. Sebaran Lokasi Pasar
plt.figure(figsize=(10, 6))
sns.scatterplot(data=df, x='lon', y='lat', hue='state', legend=False, alpha=0.6)
plt.title("Sebaran Lokasi Pasar")
plt.xlabel("Longitude")
plt.ylabel("Latitude")
plt.tight_layout()
plt.show()

# 6. Korelasi Antar Fitur Numerik (hanya kolom numerik)
plt.figure(figsize=(10, 8))
numeric_cols = df.select_dtypes(include=['number'])  # pilih hanya kolom numerik
sns.heatmap(numeric_cols.corr(), annot=True, cmap='coolwarm', fmt=".2f", linewidths=0.5)
plt.title("Korelasi Fitur Numerik")
plt.tight_layout()
plt.show()
```

```
# 7. Jumlah Pasar Berdasarkan Status Pembayaran (Jika ada kolom 'payment_status')
if 'payment_status' in df.columns:
    plt.figure(figsize=(8, 5))
    sns.countplot(data=df, x='payment_status')
    plt.title("Jumlah Pasar berdasarkan Status Pembayaran")
    plt.tight_layout()
    plt.show()
else:
    print("Kolom 'payment_status' tidak ditemukan, analisis dilewati.")

# 8. Rata-Rata Harga Produk per Negara Bagian (Jika ada kolom 'average_price')
if 'average_price' in df.columns:
    plt.figure(figsize=(12, 6))
    sns.boxplot(data=df, x='state', y='average_price')
    plt.title("Distribusi Harga Produk per Negara Bagian")
    plt.xticks(rotation=90)
    plt.tight_layout()
    plt.show()
else:
    print("Kolom 'average_price' tidak ditemukan, analisis dilewati.")

# 9. Perbandingan Produk Organik vs Non-Organik (Jika ada kolom 'organic')
if 'organic' in df.columns:
    organic_counts = df['organic'].value_counts()
    plt.figure(figsize=(6, 4))
    sns.barplot(x=organic_counts.index, y=organic_counts.values)
    plt.title("Perbandingan Produk Organik vs Non-Organik")
    plt.tight_layout()
    plt.show()
else:
    print("Kolom 'organic' tidak ditemukan, analisis dilewati.")

# 10. Trend Pembukaan Pasar per Tahun (Jika ada kolom 'open_date')
if 'open_date' in df.columns:
    df['year_open'] = pd.to_datetime(df['open_date'], errors='coerce').dt.year
    plt.figure(figsize=(12, 6))
    sns.countplot(data=df, x='year_open')
    plt.title("Trend Pembukaan Pasar per Tahun")
    plt.xticks(rotation=90)
    plt.tight_layout()
    plt.show()
else:
    print("Kolom 'open_date' tidak ditemukan, analisis dilewati.")

# 11. Hubungan Jumlah Vendor dengan Jumlah Produk Dijual (Jika ada kolom 'num_vendors')
if 'num_vendors' in df.columns:
    plt.figure(figsize=(10, 6))
    sns.scatterplot(data=df, x='num_vendors', y='num_items_sold')
    plt.title("Jumlah Vendor vs Jumlah Produk Dijual")
    plt.xlabel("Jumlah Vendor")
    plt.ylabel("Jumlah Produk Dijual")
    plt.tight_layout()
    plt.show()
else:
    print("Kolom 'num_vendors' tidak ditemukan, analisis dilewati.")

# 12. Distribusi Lama Pasar Buka
plt.figure(figsize=(8, 5))
sns.histplot(df['months_open'], bins=30, kde=True)
plt.title("Distribusi Lama Pasar Buka (bulan)")
plt.xlabel("Bulan")
plt.tight_layout()
plt.show()

# 13. Perbandingan Produk Sayuran dan Buah
veg_sum = df['Vegetables'].sum()
fruit_sum = df['Fruits'].sum()
plt.figure(figsize=(6, 4))
plt.bar(['Vegetables', 'Fruits'], [veg_sum, fruit_sum], color=['green', 'orange'])
plt.title("Perbandingan Produk Sayuran dan Buah")
plt.ylabel("Jumlah Pasar yang Menjual")
plt.tight_layout()
plt.show()

# 14. Boxplot Jumlah Produk Dijual per Negara Bagian
plt.figure(figsize=(12, 6))
sns.boxplot(data=df, x='state', y='num_items_sold')
sns.boxplot(data=df, x="state", y="num_items_sold")
plt.title("Sebaran Jumlah Produk Dijual per Negara Bagian")
plt.xlabel("Negara Bagian")
plt.ylabel("Jumlah Produk Dijual")
plt.xticks(rotation=90)
plt.tight_layout()
plt.show()

# 15. Produk dengan Penjualan Terendah
product_counts_asc = df[product_columns].sum().sort_values()
plt.figure(figsize=(12, 6))
sns.barplot(x=product_counts_asc.index, y=product_counts_asc.values, palette="coolwarm")
plt.title("Produk dengan Penjualan Terendah")
plt.xlabel("Produk")
plt.ylabel("Jumlah Pasar yang Menjual")
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

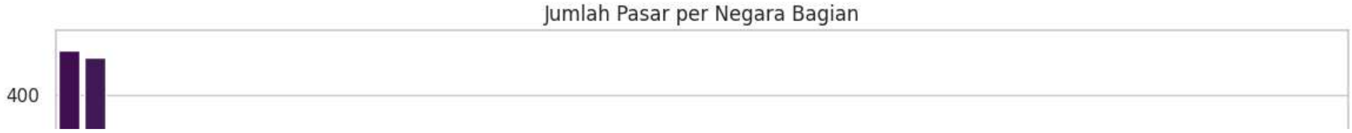
Choose Files markets_cleaned.csv

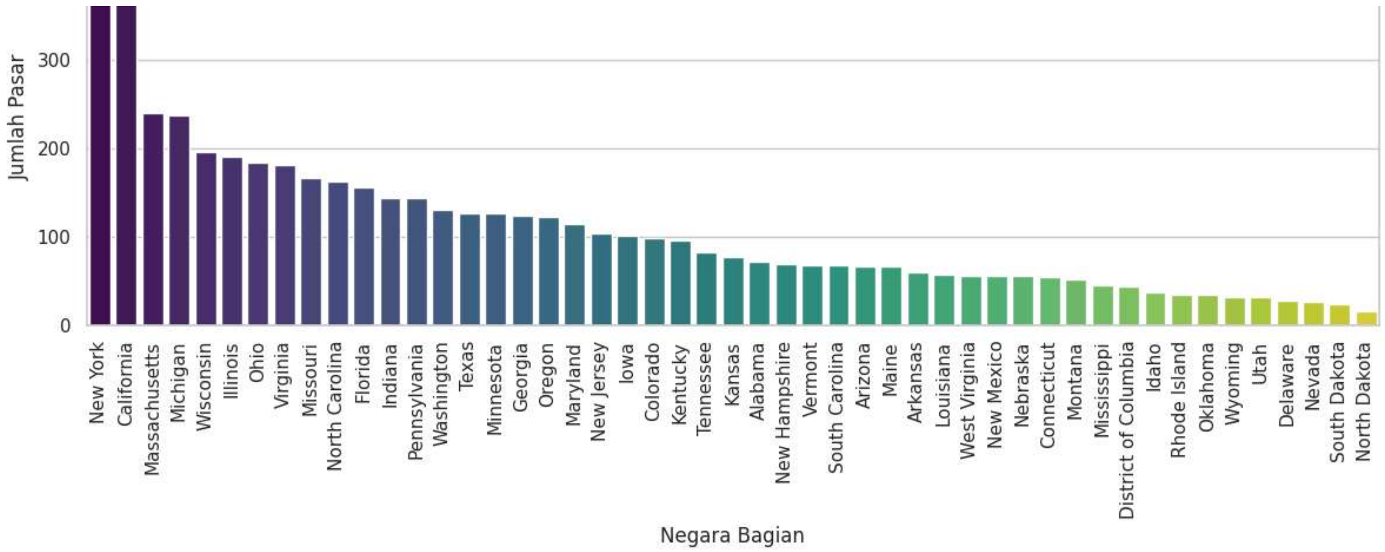
- **markets_cleaned.csv**(text/csv) - 853696 bytes, last modified: 5/17/2025 - 100% done

```
Saving markets_cleaned.csv to markets_cleaned (2).csv
=== Informasi Dataset ===
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5343 entries, 0 to 5342
Data columns (total 39 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   Unnamed: 0      5343 non-null   int64  
 1   name            5343 non-null   object 
 2   city            5340 non-null   object 
 3   county          5341 non-null   object 
 4   state           5343 non-null   object 
 5   lat             5339 non-null   float64
 6   lon             5339 non-null   float64
 7   months_open     5343 non-null   int64  
 8   Bakedgoods      5343 non-null   int64  
 9   Beans           5343 non-null   int64  
 10  Cheese          5343 non-null   int64  
 11  Coffee          5343 non-null   int64  
 12  Crafts          5343 non-null   int64  
 13  Eggs            5343 non-null   int64  
 14  Flowers         5343 non-null   int64  
 15  Fruits          5343 non-null   int64  
 16  Grains          5343 non-null   int64  
 17  Herbs           5343 non-null   int64  
 18  Honey           5343 non-null   int64  
 19  Jams            5343 non-null   int64  
 20  Juices          5343 non-null   int64  
 21  Maple           5343 non-null   int64  
 22  Meat            5343 non-null   int64  
 23  Mushrooms       5343 non-null   int64  
 24  Nursery         5343 non-null   int64  
 25  Nuts            5343 non-null   int64  
 26  PetFood         5343 non-null   int64  
 27  Plants          5343 non-null   int64  
 28  Poultry         5343 non-null   int64  
 29  Prepared        5343 non-null   int64  
 30  Seafood         5343 non-null   int64  
 31  Soap            5343 non-null   int64  
 32  Tofu            5343 non-null   int64  
 33  Trees           5343 non-null   int64  
 34  Vegetables      5343 non-null   int64  
 35  WildHarvested   5343 non-null   int64  
 36  Wine            5343 non-null   int64  
 37  num_items_sold  5343 non-null   int64  
 38  state_pop       5343 non-null   float64
dtypes: float64(3), int64(32), object(4)
memory usage: 1.6+ MB
None
         Unnamed: 0                        name         city      county  \
count   5343.000000                        5343         5340        5341   
unique          NaN                        5075         3177        1122   
top             NaN  Main Street Farmers Market  Philadelphia  Washington   
freq            NaN                           8           39          64   
mean    2671.000000                         NaN          NaN         NaN   
std     1542.535575                         NaN          NaN         NaN   
min        0.000000                         NaN          NaN         NaN   
25%     1335.500000                         NaN          NaN         NaN   
50%     2671.000000                         NaN          NaN         NaN   
75%     4006.500000                         NaN          NaN         NaN   
max     5342.000000                         NaN          NaN         NaN   

           state          lat          lon  months_open   Bakedgoods  \
count       5343  5339.000000  5339.000000  5343.000000  5343.000000   
unique        49          NaN          NaN          NaN          NaN   
top     New York          NaN          NaN          NaN          NaN   
freq         450          NaN          NaN          NaN          NaN   
mean         NaN   -89.888501    39.453910     6.376567     0.885458   
std          NaN    15.750410     4.483651     2.674895     0.318499   
min          NaN  -124.416226    25.109214     1.000000     0.000000   
25%          NaN   -96.150590    36.857087     5.000000     1.000000   
50%          NaN   -85.701673    40.056583     6.000000     1.000000   
75%          NaN   -77.227226    42.517589     7.000000     1.000000   
max          NaN   -67.277359    48.943331    12.000000     1.000000   

              Beans  ...     Prepared      Seafood         Soap         Tofu  \
count   5343.000000  ...  5343.000000  5343.000000  5343.000000  5343.000000   
unique          NaN  ...          NaN          NaN          NaN          NaN   
top             NaN  ...          NaN          NaN          NaN          NaN   
freq            NaN  ...          NaN          NaN          NaN          NaN   
mean       0.144862  ...     0.620438     0.248362     0.690249     0.040240   
std        0.351995  ...     0.485323     0.432104     0.462434     0.196539   
```

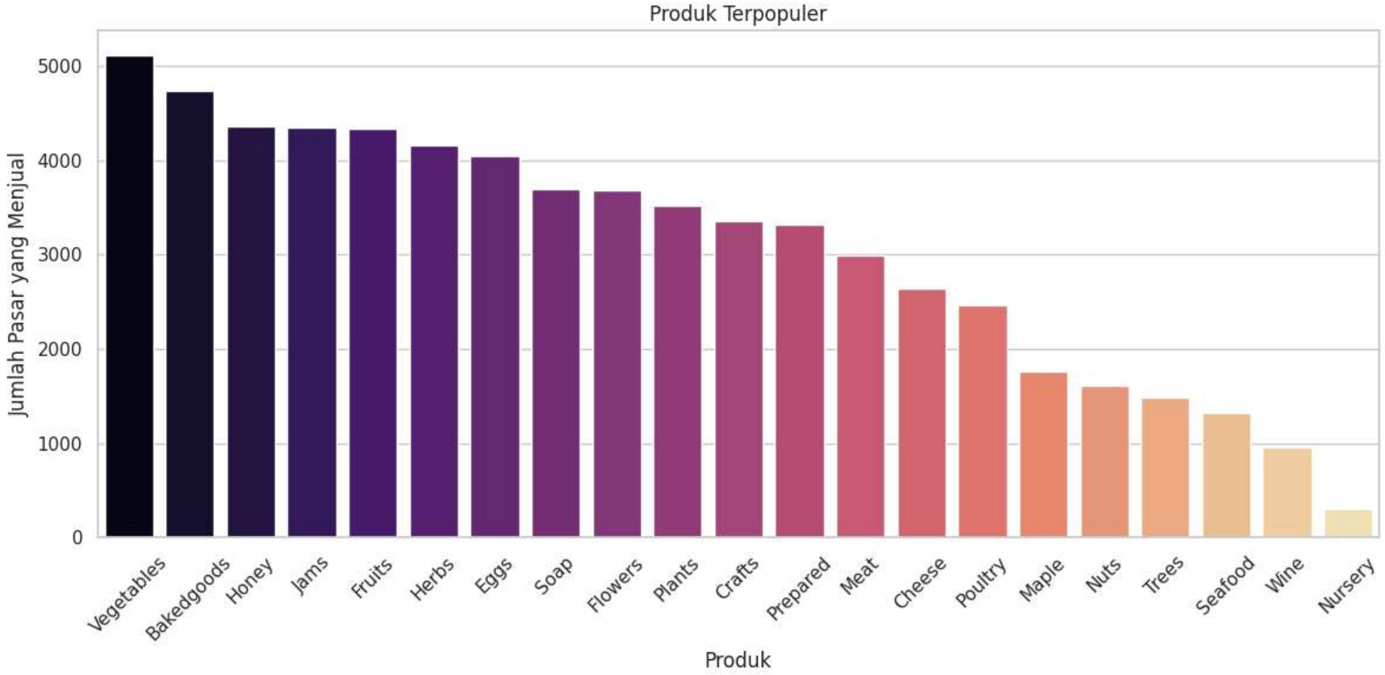
```
min        0.000000  ...     0.000000     0.000000     0.000000     0.000000
25%        0.000000  ...     0.000000     0.000000     0.000000     0.000000
50%        0.000000  ...     1.000000     0.000000     1.000000     0.000000
75%        0.000000  ...     1.000000     0.000000     1.000000     0.000000
max        1.000000  ...     1.000000     1.000000     1.000000     1.000000

              Trees  Vegetables  WildHarvested         Wine  num_items_sold  \
count   5343.000000  5343.00000    5343.000000  5343.000000     5343.000000   
unique          NaN         NaN            NaN          NaN             NaN   
top             NaN         NaN            NaN          NaN             NaN   
freq            NaN         NaN            NaN          NaN             NaN   
mean       0.279057     0.95714       0.148231     0.178551       13.544076   
std        0.448577     0.20256       0.355362     0.383012        5.791125   
min        0.000000     0.00000       0.000000     0.000000        0.000000   
25%        0.000000     1.00000       0.000000     0.000000       10.000000   
50%        0.000000     1.00000       0.000000     0.000000       14.000000   
75%        1.000000     1.00000       0.000000     0.000000       18.000000   
max        1.000000     1.00000       1.000000     1.000000       28.000000   

           state_pop  
count   5.343000e+03  
unique           NaN  
top              NaN  
freq             NaN  
mean    1.107189e+07  
std     1.023976e+07  
min     5.841530e+05  
25%     4.741079e+06  
50%     6.745408e+06  
75%     1.288058e+07  
max     3.880250e+07  

[11 rows x 39 columns]
Nilai kosong per kolom:
 Unnamed: 0        0
name              0
city              3
county            2
state             0
lat               4
lon               4
months_open       0
Bakedgoods        0
Beans             0
Cheese            0
Coffee            0
Crafts            0
Eggs              0
Flowers           0
Fruits            0
Grains            0
Herbs             0
Honey             0
Jams              0
Juices            0
Maple             0
Meat              0
Mushrooms         0
Nursery           0
Nuts              0
PetFood           0
Plants            0
Poultry           0
Prepared          0
Seafood           0
Soap              0
Tofu              0
Trees             0
Vegetables        0
WildHarvested     0
Wine              0
num_items_sold    0
state_pop         0
dtype: int64
<ipython-input-14-37e7cba55788>:26: FutureWarning: 

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legenc

  sns.barplot(x=state_counts.index, y=state_counts.values, palette="viridis")
```

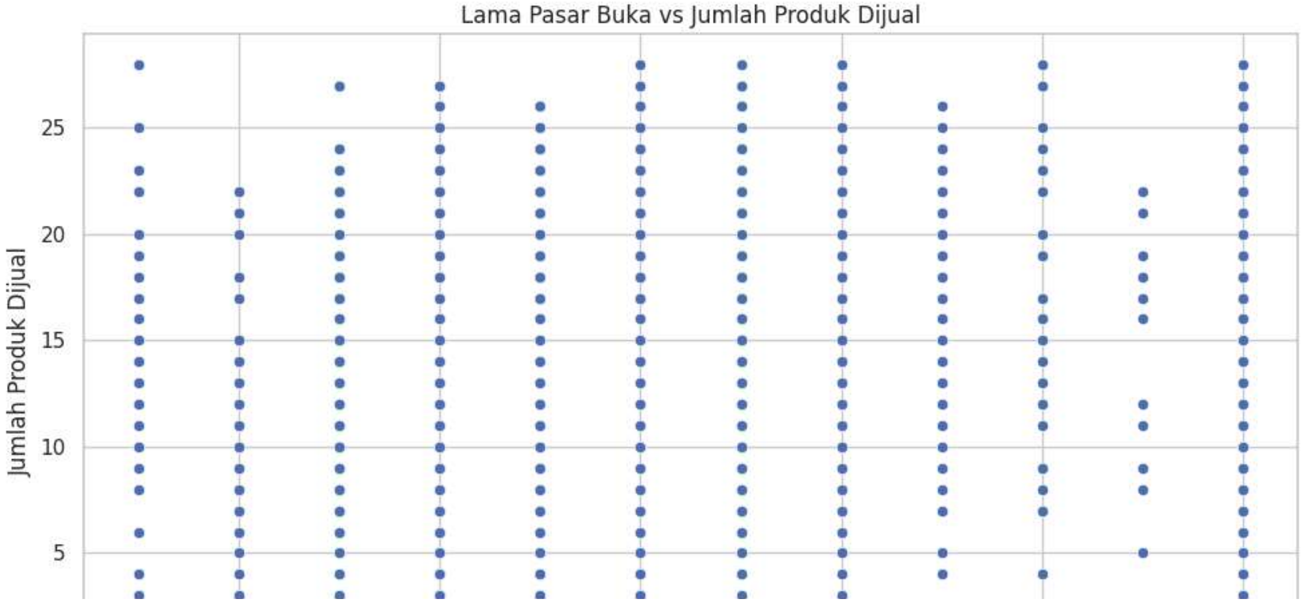
Jumlah Pasar per Negara Bagian



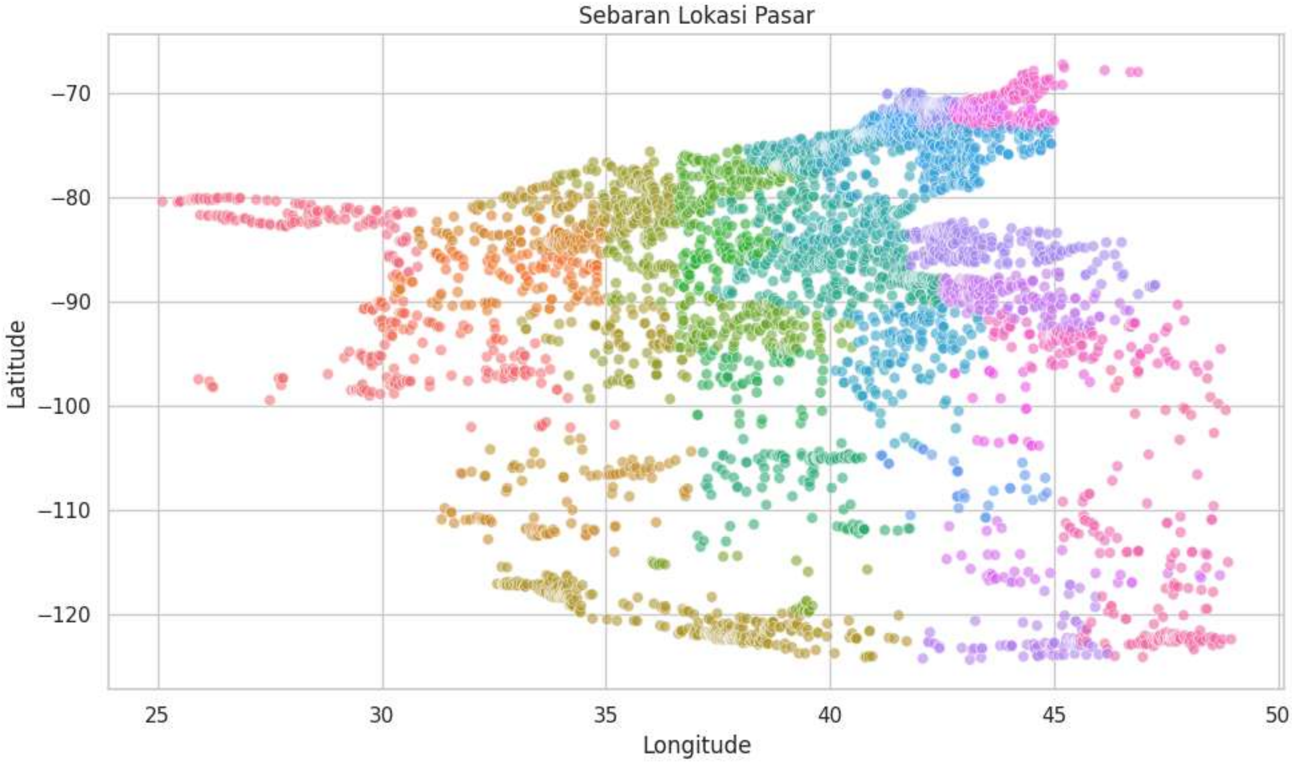


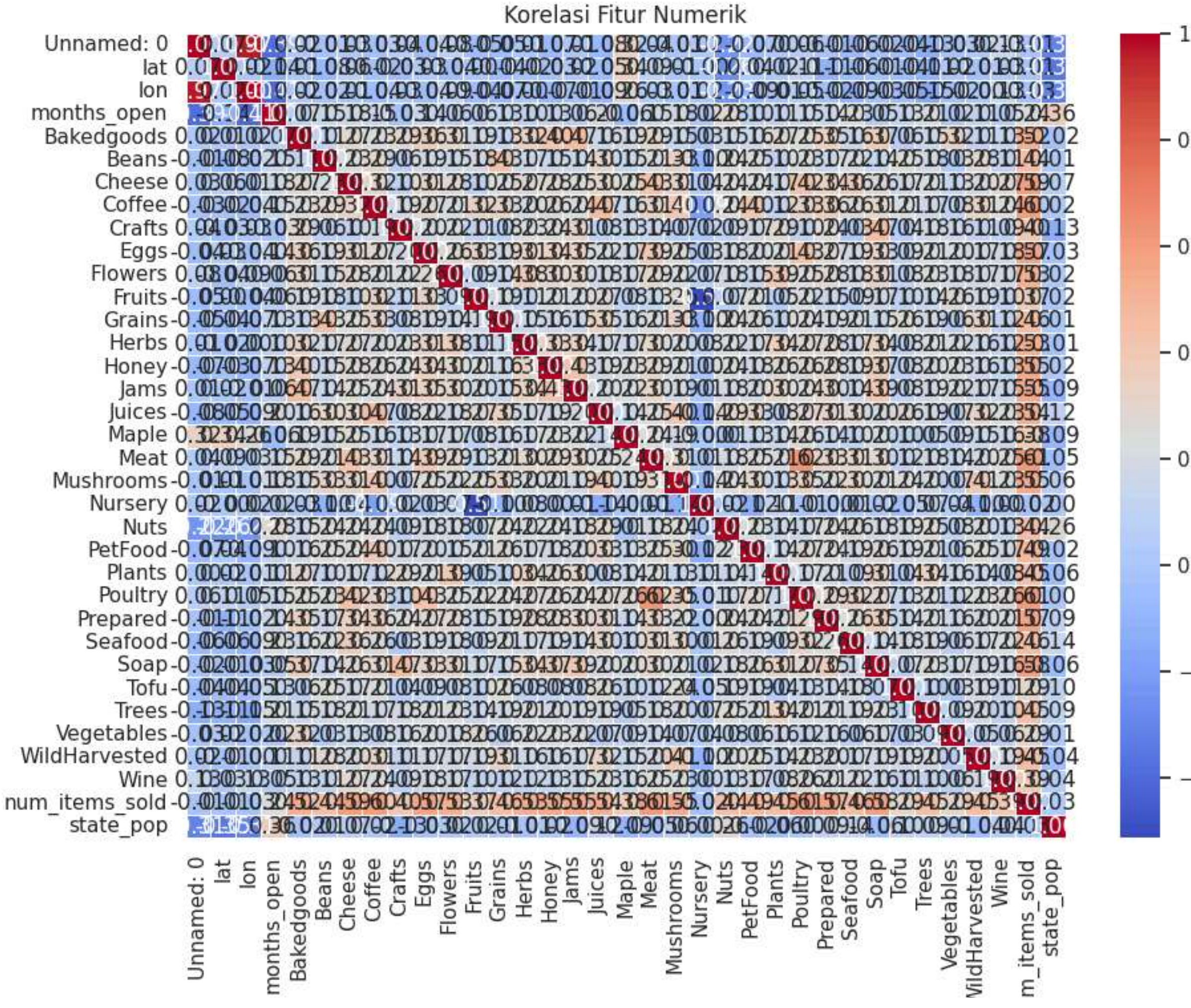
```
<ipython-input-14-37e7cba55788>:43: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legenc

  sns.barplot(x=product_counts.index, y=product_counts.values, palette="magma")
```



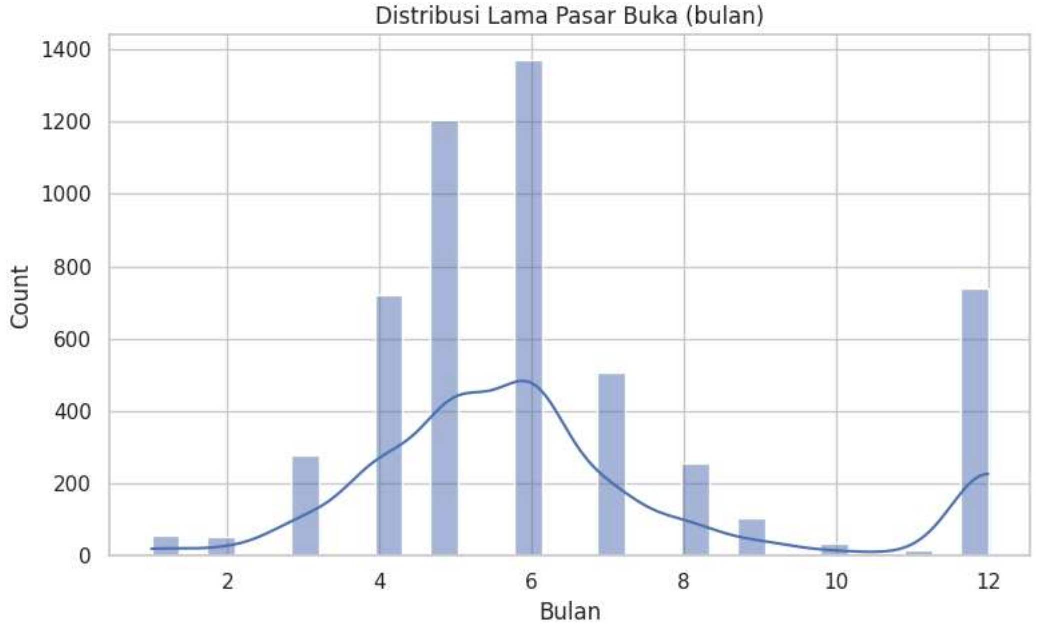


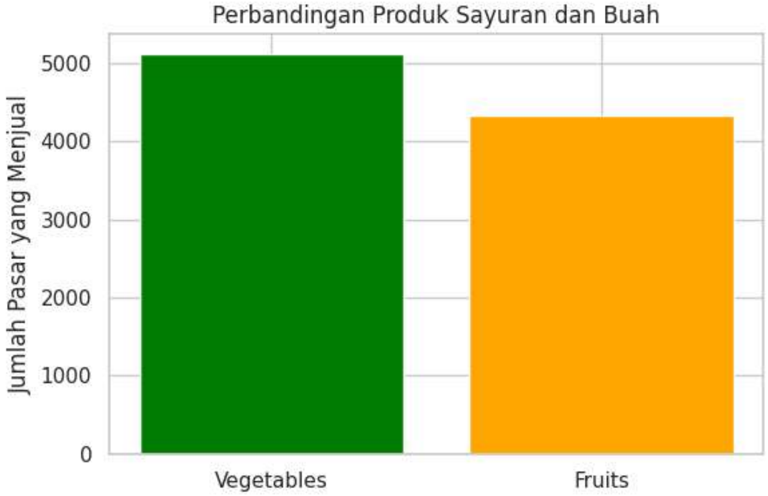
0
2
4
6
8
10
12
Lama Pasar Buka (bulan)

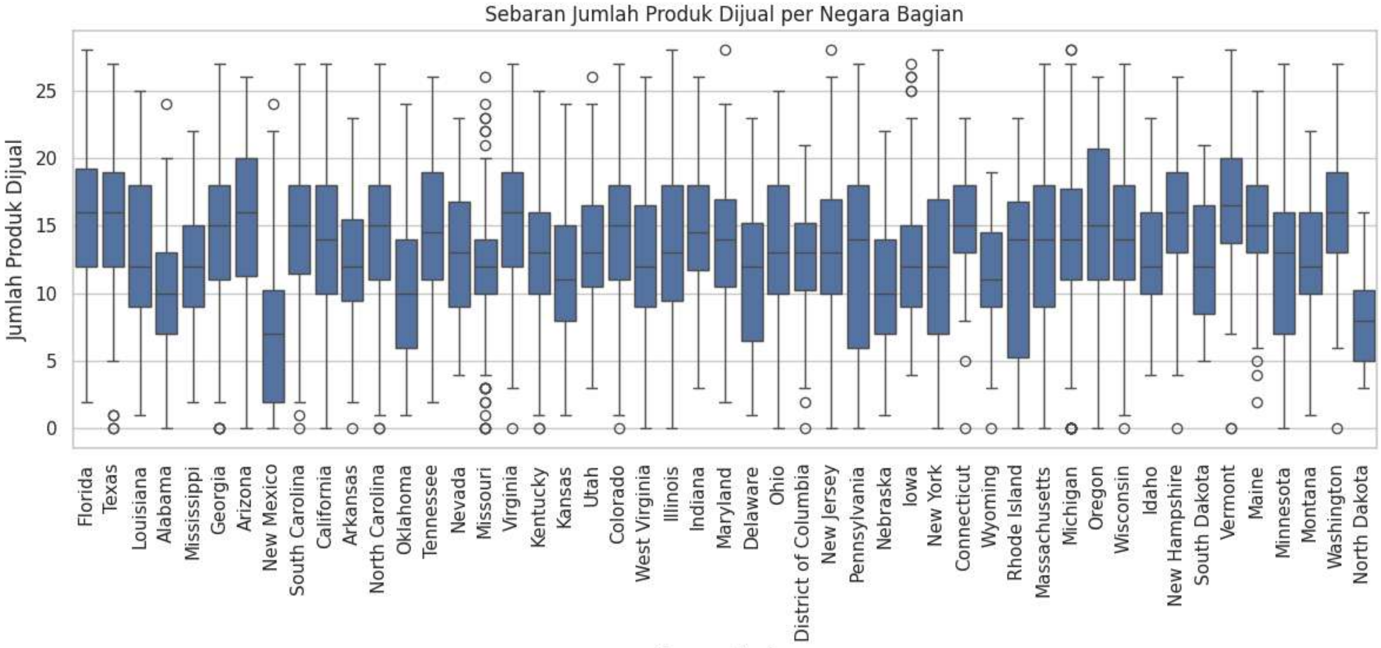

Sebaran Lokasi Pasar
−70
−80
−90
−100
−110
−120
Latitude
25
30
35
40
45
50
Longitude


Korelasi Fitur Numerik
Unnamed: 0
lat
lon
months_open
Bakedgoods
Beans
Cheese
Coffee
Crafts
Eggs
Flowers
Fruits
Grains
Herbs
Honey
Jams
Juices
Maple
Meat
Mushrooms
Nursery
Nuts
PetFood
Plants
Poultry
Prepared
Seafood
Soap
Tofu
Trees
Vegetables
WildHarvested
Wine
num_items_sold
state_pop
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

```
Kolom 'payment_status' tidak ditemukan, analisis dilewati.
Kolom 'average_price' tidak ditemukan, analisis dilewati.
Kolom 'organic' tidak ditemukan, analisis dilewati.
Kolom 'open_date' tidak ditemukan, analisis dilewati.
Kolom 'num_vendors' tidak ditemukan, analisis dilewati.
```

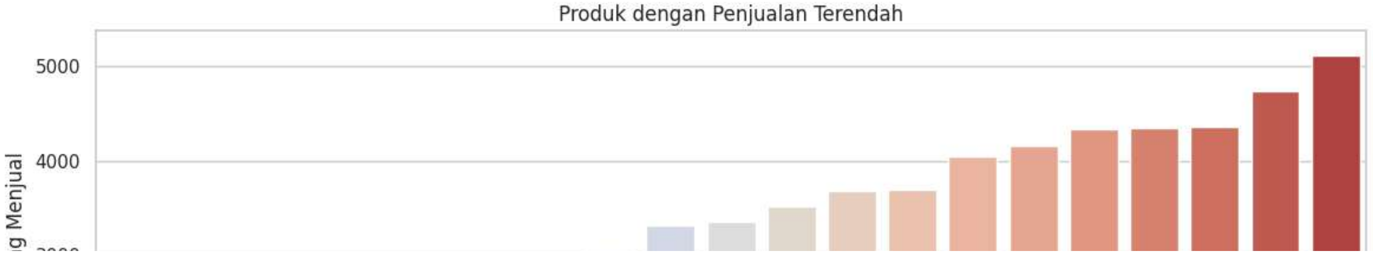






Negara Bagian

```
<ipython-input-14-37e7cba55788>:165: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legenc

  sns.barplot(x=product_counts_asc.index, y=product_counts_asc.values, palette="coolwarm")
```


Produk dengan Penjualan Terendah
g Menjual