

Dalam pengerjaan tugas penulis menggunakan 3 tools, yaitu:

- PostgreSQL  
PostgreSQL digunakan untuk memuat dataset dan sebagai syarat wajib untuk tugas ini.
- Python  
Python digunakan untuk analisis data. Python dipilih karena penulis lebih dikuasai dengan Python dan lebih fleksibel untuk mengolah data.
- Power BI  
Power BI digunakan untuk visualisasi data. Power BI dipilih karena penulis juga lebih dikuasai dan lebih banyak opsi untuk melakukan visualisasi.

Setelah menentukan tools penulis langsung mendownload data dari kaggle kemudian memuat dan melakukan data understanding melalui PostgreSQL.

## PostgreSQL

1. Mengganti penamaan titik “.” Menjadi “\_” pada kolom dataset agar tidak terjadi kesalahan pada query PostgreSQL dan tetap mempertahankan data mentah. Seperti “nama.barang” menjadi “nama\_barang”.
2. Mengisi nama kolom kosong pada kolom urutan nomor dengan nama “nomor” agar tidak ada nama kolom yang tidak berlabel.
3. Mengganti format cell pada kolom “kuantum” dan “nominal” pada dataset menjadi “number”. Karena terdapat missing value yang nantinya bisa menjadi error pada query PostgreSQL. Seperti “7.00E+06” menjadi “7000000”.
4. Membuat tabel di PostgreSQL



```
Query  Query History
1 CREATE TABLE public.penjualan_barang
2 (
3     nomor INT,
4     tanggal DATE,
5     nama_pembeli TEXT,
6     nama_barang TEXT,
7     kuantum INT,
8     nominal INT
9 )
```

Data Output Messages Notifications

ERROR: relation "penjualan\_barang" already exists

SQL state: 42P07

5. Melakukan import csv dataset ke dalam PostgreSQL

Query

Query History

Scratch Pad

1

SELECT \* from public."penjualan\_barang"

Data Output

Messages

Notifications

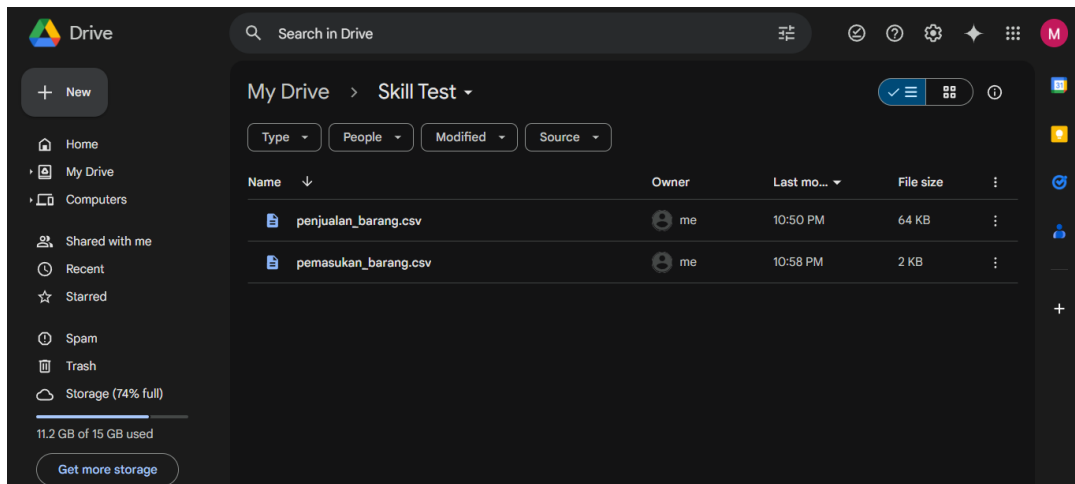
- Karena untuk proses selanjutnya penulis menggunakan Python. Maka untuk dataset di PostgreSQL ini penulis download kembali untuk kemudian penulis upload ke google colab. Hal ini penulis lakukan karena penulis meragukan kemampuan device penulis jika menjalankan Python di local device penulis. Dan jika penulis memaksakan untuk mengakses PostgreSQL local penulis untuk bisa public menggunakan Google Colab akan beresiko untuk device penulis karena perlu membuka akses public PostgreSQL.

Link Google Colab penulis

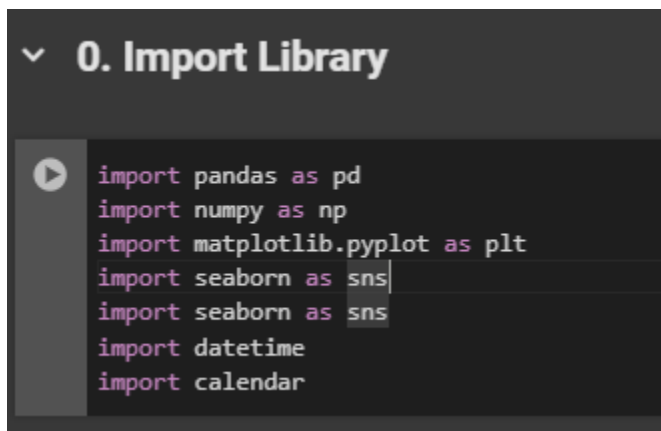
<https://colab.research.google.com/drive/1uLM2Nx9M8B0us7YzL12TROY-KBVW6ANx?usp=sharing>

## Python

- Untuk dataset penulis upload pada google drive untuk memudahkan dalam menjalankan ulang kode python.



2. Tahap awal dalam notebook python adalah melakukan import library



3. Melakukan import dataset dari google drive

### 1. Import Dataset

```
[5] #connect to drive
from google.colab import drive
drive.mount('/content/drive', force_remount=True)
```

Mounted at /content/drive

Menyediakan label kolom sesuai penamaan kolom pada dataset

```
[6] column_pemasukan = ["nomor", "tanggal", "nama_barang", "kuantum"]
column_penjualan = ["nomor", "tanggal", "nama_pembeli", "nama_barang", "kuantum", "nominal"]
```

```
[7] pemasukan = 'drive/My Drive/ Skill Test/pemasukan_barang.csv'
raw_pemasukan = pd.read_csv(pemasukan, sep = ",", encoding = "utf-8")
raw_pemasukan.columns=column_pemasukan
raw_pemasukan.head()
```

	nomor	tanggal	nama_barang	kuantum
0	1	5/29/2020	GULA	100000
1	2	7/7/2020	GULA	97950
2	3	7/16/2020	GULA	2050
3	4	11/12/2020	GULA	14000
4	5	12/24/2020	GULA	8000

Next steps: [Generate code with raw\\_pemasukan](#) [View recommended plots](#) [New interactive sheet](#)

```
[8] penjualan = 'drive/My Drive/ Skill Test/penjualan_barang.csv'
raw_penjualan = pd.read_csv(penjualan, sep = ",", encoding = "utf-8")
raw_penjualan.columns=column_penjualan
```

4. Untuk memastikan bahwa data tanggal dapat diolah dengan lancar, penulis menyesuaikan kembali data “tanggal” meskipun sebelumnya pada PostgreSQL sudah disesuaikan.

### 2. Date Correction

```
##Mengubah Format tanggal agar mudah dianalisis
raw_pemasukan['tanggal'] = raw_pemasukan['tanggal'].astype('datetime64[ns]')
raw_penjualan['tanggal'] = raw_penjualan['tanggal'].astype('datetime64[ns]')
```

5. Missing value checking, meskipun jika dilihat sekilas pada PostgreSQL sudah terlihat bahwa data sudah bersih. Akan tetapi di sini kembali dilakukan pengecekan jika terjadi kesalahan ketika import data.

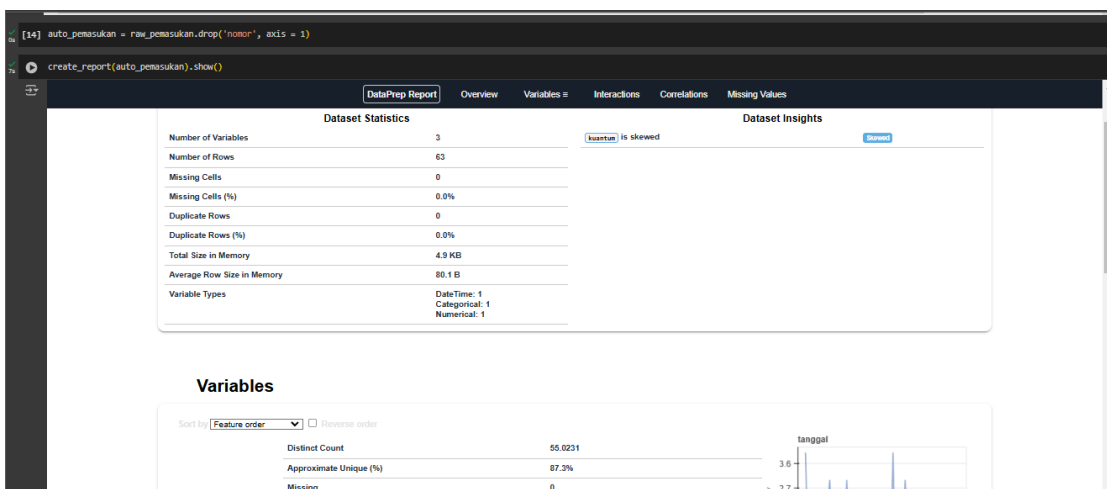
### 3. Missing Value Checking

```
[10] ##Mengecek ada tidaknya missing value
datasets = [raw_pemasukan, raw_penjualan]
titles = ['pemasukan', 'penjualan']

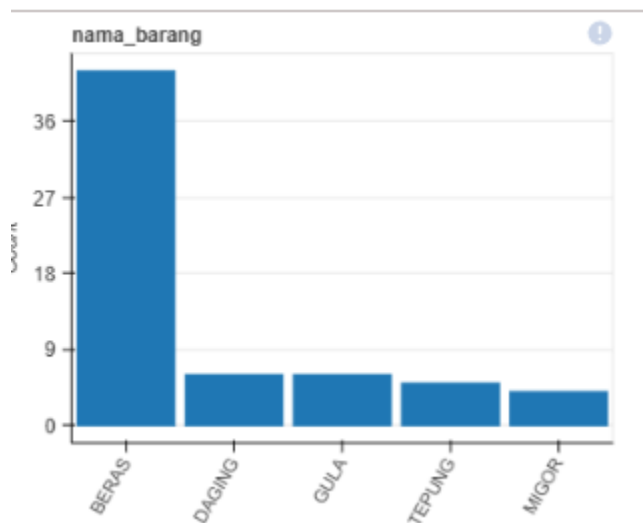
missing_check = pd.DataFrame({}, )
missing_check['datasets'] = titles
missing_check['features'] = [' ', '.join([col for col, null in df.isnull().sum().items()] for df in datasets)]
missing_check['null_amount'] = [df.isnull().sum().sum() for df in datasets]
missing_check['null_features_amount'] = [len([col for col, null in df.isnull().sum().items() if null > 0]) for df in datasets]
missing_check['null_features'] = [' ', '.join([col for col, null in df.isnull().sum().items() if null > 0]) for df in datasets]
missing_check.style.background_gradient(cmap='Blues')
```

	datasets	features	null_amount	null_features_amount	null_features
0	pemasukan	nomor, tanggal, nama_barang, kuantum	0	0	
1	penjualan	nomor, tanggal, nama_pembeli, nama_barang, kuantum, nominal	0	0	

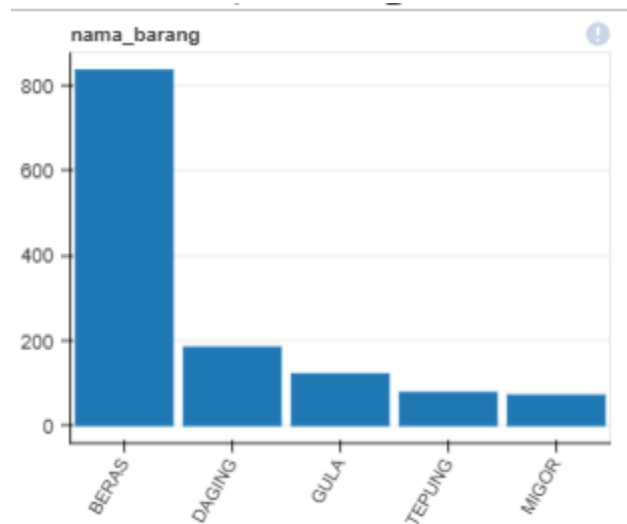
6. Dari sini masuk pada tahap analisis data. Sebelumnya penulis sudah menyiapkan beberapa pertanyaan untuk penulis temukan jawabannya pada tahap analisis data ini. Daftar pertanyaan penulis sebagai berikut:
  - Bagaimana tren barang masuk dan jual per bulan?
  - Tren penjualan, dan apa yang memengaruhinya?
  - Siapa saja pembeli dengan jumlah barang tertinggi selama 2020-2021?
  - Produk apa saja dengan pembelian tertinggi?
  - Berapa total pendapatan tiap bulan?
  - Pada bulan apa dengan transaksi terbanyak?
  - Pada hari apa dengan transaksi terbanyak?
7. Tahap awal analisis data penulis menggunakan Automated EDA, hal ini untuk memudahkan tahap awal analisis data dan menemukan insight.



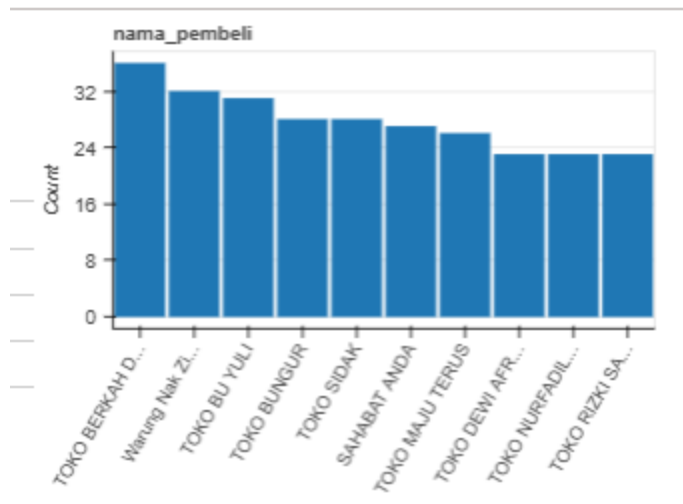
8. Dari tahap ini penulis menemukan bahwa dari dataset “pemasukan” barang dengan jumlah pemasukan terbesar adalah “BERAS”. Hal ini bisa terjadi dikarenakan beras menjadi bahan pokok makanan di Indonesia. Kemudian disusul “DAGING” dan “GULA”.



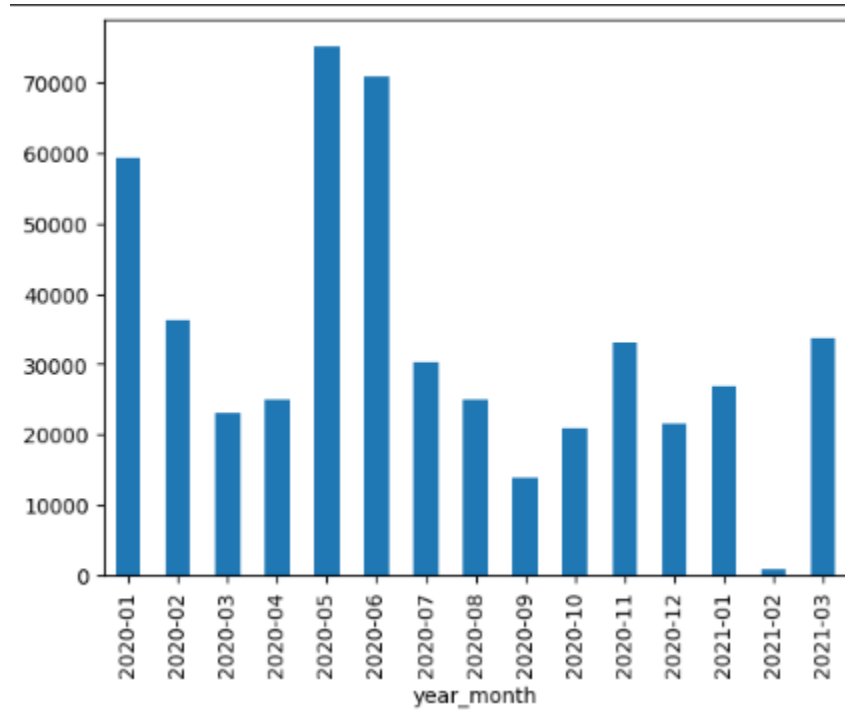
9. Sedangkan dari dataset “penjualan” juga memiliki tren yang sama. Sehingga untuk pemilihan stok pemasukan sudah sejalan dengan kebutuhan untuk penjualan.



Sedangkan secara total pembeli dengan pembelian terbanyak adalah TOKO BERKAH DUA PUTRA, Warung Nak Zihan, dan TOKO BU YULI. Meskipun begitu, secara keseluruhan pembeli membeli barang dengan jumlah yang banyak mendekati pembelian tertinggi.



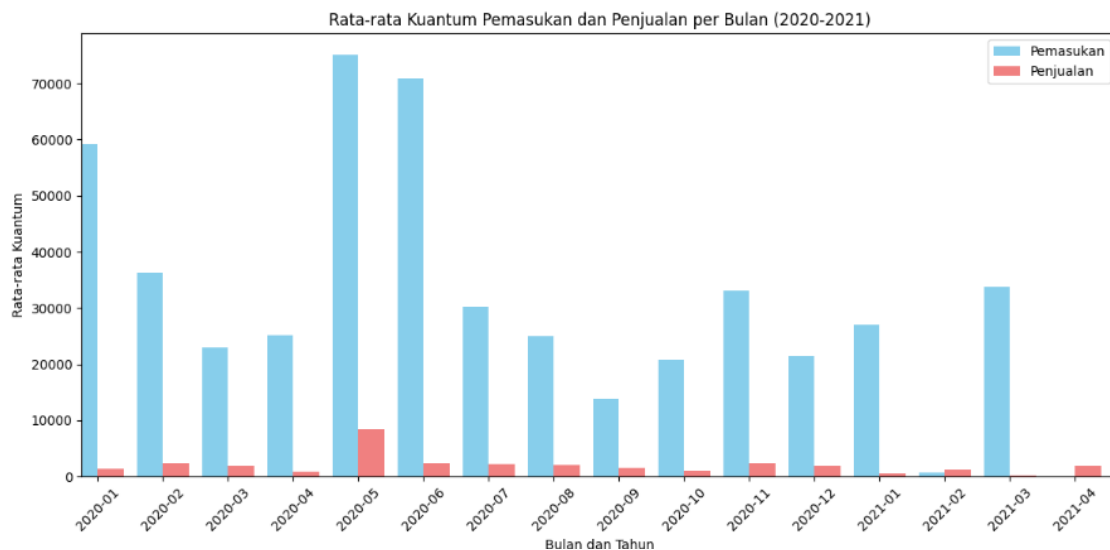
10. Pada tahap ini dilakukan analisis data secara manual. Dimulai dengan dataset “pemasukan”. Terlihat bahwa Pemasukan barang tertinggi terjadi pada bulan januari, Mei, dan Juni. Hal ini terjadi karena Januari menjadi awal bulan pada tahun baru sehingga stok barang akan ditambah untuk bulan ini. Sedangkan untuk bulan Mei 2020 sudah masuk ke dalam hari besar Ramadhan dimana penjualan seharusnya meningkat sampai pekan hari raya idul Fitri yaitu di bulan Juni. Nantinya akan dilihat apakah pada dataset penjualan akan sejalan dengan tren pemasukan barang ini



Dari data Penjualan ternyata memiliki perbedaan tren dengan data pemasukan. pada data pemasukan barang tinggi pada bulan januari, padahal dari data penjualan januari, februari,

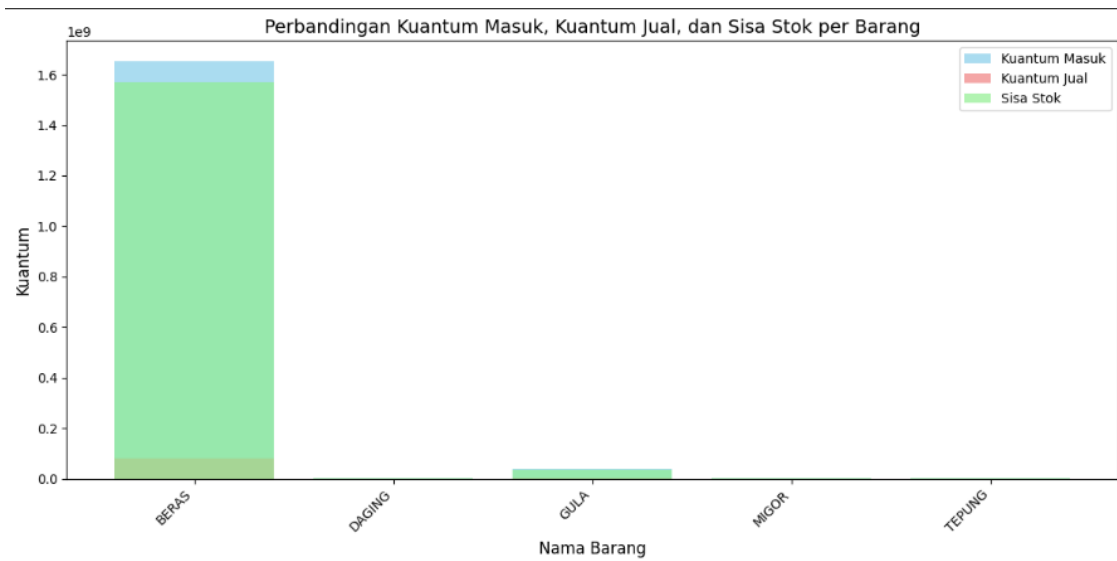
maret 2020 penjualan masih standar dan tidak ada peningkatan signifikan. Hal ini bisa menjadi wawasan baru bahwa awal bulan ditahun baru masih memiliki tren penjualan yang sama jadi tidak perlu melebihi pemasukan seperti bulan lain, kecuali akan diadakan promo yang memungkinkan untuk menarik lebih banyak pelanggan.

Sedangkan untuk bulan Mei 2020 pemasukan dan penjualan sejalan. Namun tidak dengan bulan Juni dengan pemasukan yang masih tinggi sedangkan penjualan sudah turun ke angka standar bulan lain. Hal ini memungkinkan terjadi overstok di gudang dan bisa menurunkan kualitas barang karena terlalu lama di simpan di gudang dan tidak efisien dalam strategi stok barang. Hal ini terjadi karena mungkin toko mengira bahwa perayaan Hari Raya Idul Fitri bisa menembus sampai bulan Juni, padahal pada kenyataannya pada tahun 2020 tersebut Idul Fitri jatuh pada tanggal 24-25 dan biasanya perayaan hanya sampai pada akhir pekan saja di tanggal tersebut. Sehingga perayaan tidak sampai pada bulan Juni.



- Selanjutnya dilakukan pengecekan untuk mengetahui Perbandingan Kuantum Masuk, Kuantum Jual, dan Sisa Stok per Barang.

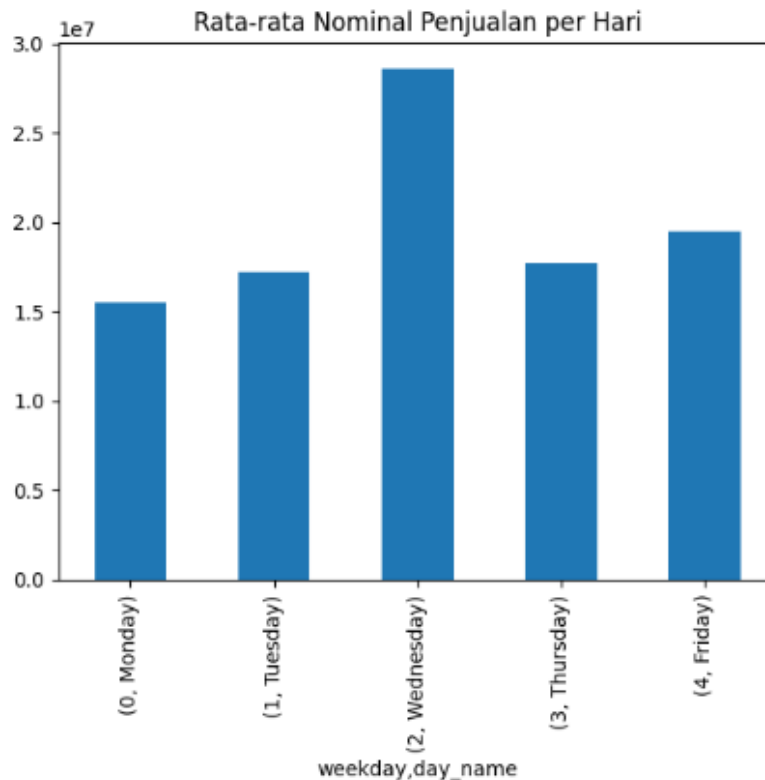
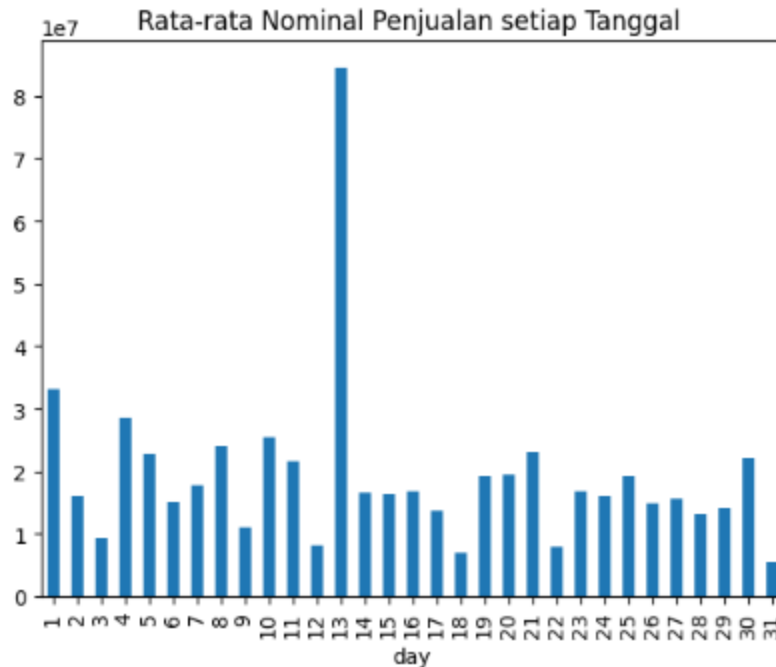




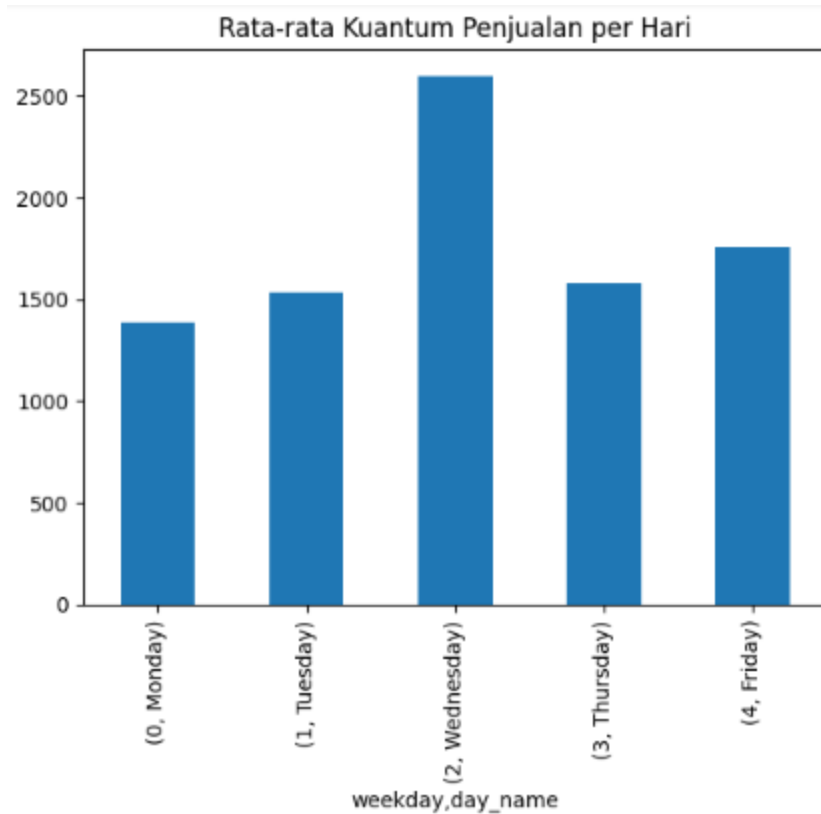
nama_barang	kuantum_masuk	kuantum_jual	sisa Stok
BERAS	1651315688	82540836	1568774852
DAGING	3036000	93810	2942190
GULA	38962000	1340616	37621384
MIGOR	1821008	96592	1724416
TEPUNG	654500	40100	614400

Dari hasil visualisasi terlihat bahwa barang terbanyak adalah BERAS seauai dengan hasil Automated EDA sebelumnya. Akan tetapi permasalahan di sini adalah karena antara pemasukan dan penjualan memiliki selisih yang jauh menyebabkan adanya overstok khususnya untuk barang DAGING dan GULA

12. Selanjutnya dilakuka pengecekan nominal masuk tertinggi berdasarkan tanggal dan hari



Terlihat bahwa nominal masuk tertinggi terjadi pada tanggal muda dan pertengahan bulan. Hal ini mengindikasikan daya beli lebih masih sering terjadi pada pertengahan bulan dengan hari Rabu. Hal ini juga sejalan dengan jumlah transaksi tertinggi juga terjadi pada hari Rabu, seperti terlihat pada grafik rata-rata kuantum penjualan per hari berikut



13. Save data untuk selanjutnya dilakukan visualisasi report menggunakan Power BI

```
5. Download Data untuk Visualisasi Power BI
```

```
from google.colab import files
gabungan.to_csv('datavis.csv', encoding = 'utf-8-sig', index = False)

files.download('datavis.csv')
```

Visualisasi Data dengan Power BI

Link Power BI: [https://app.powerbi.com/groups/me/reports/61a642c1-31a3-4e96-95e4-5f8e37ce97ad?ctid=2eee0858-323f-4952-a7ec-cf2846db9115&pbi\\_source=linkShare](https://app.powerbi.com/groups/me/reports/61a642c1-31a3-4e96-95e4-5f8e37ce97ad?ctid=2eee0858-323f-4952-a7ec-cf2846db9115&pbi_source=linkShare)