P0128 BRAND\_A 205000 Jakarta Timur DKI Jakarta **4** 1702573 2019-10-16 15696 P2968 BRAND\_R 4475000 **5** 1672906 2019-07-16 P0710 BRAND\_C 12748 Jakarta Utara DKI Jakarta 4 520000 DKI Jakarta **6** 1711399 2019-11-04 16791 Jakarta Barat P0860 BRAND C 1465000 P3342 BRAND S **7** 1695367 2019-09-26 13069 Surakarta Jawa Tengah 205000 **8** 1741846 2019-12-30 16873 Jakarta Barat DKI Jakarta P3203 BRAND S 450000 **9** 1720189 2019-11-24 14723 P1701 BRAND\_H 149000 Banten 2 Tangerang In [ ]: #Menampilkan info dari dataset (nama kolom, banyak row null-non null, tipe data) print('Info dataset') dataset.info() Info dataset <class 'pandas.core.frame.DataFrame'> RangeIndex: 5000 entries, 0 to 4999 Data columns (total 9 columns): # Column Non-Null Count Dtype 0 order\_id 5000 non-null int64 order\_date 5000 non-null object 1 2 customer\_id 5000 non-null int64 5000 non-null 3 city object 4 province 5000 non-null object product\_id 5000 non-null 5 object 5000 non-null 6 brand object 5000 non-null 7 quantity int64 item\_price 5000 non-null int64 dtypes: int64(4), object(5) memory usage: 351.7+ KB Penjelasan Dataset - Detail Kolom 1. order id: ID dari order/transaksi, 1 transaksi bisa terdiri dari beberapa produk, tetapi hanya dilakukan oleh 1 customer 2. order date: tanggal terjadinya transaksi 3. customer\_id: ID dari pembeli, bisa jadi dalam satu hari, 1 customer melakukan transaksi beberapa kali 4. city: kota tempat toko terjadinya transaksi 5. **province**: provinsi (berdasarkan city) 6. product\_id : ID dari suatu product yang dibeli 7. brand : brand/merk dari product. Suatu product yang sama pasti memiliki brand yang sama 8. quantity: Kuantitas / banyaknya product yang dibeli 9. item\_price: Harga dari 1 product (dalam Rupiah). Suatu product yang sama, bisa jadi memiliki harga yang berbeda saat dibeli 2. Penambahan Kolom Order Month pada Dataset In [ ]: #import library datetime import datetime #Untuk mendapatkan bulan dan tahun pada kolom order\_date (string/object -> datetime) dataset['order\_month'] = dataset['order\_date'].apply(lambda x: datetime.datetime.strptime(x, "%Y-%m-%d").strftime('%Y-%m')) #Tampilkan 5 hasil teratas print("Data Retail Raw dengan Order Month\n") dataset.head(10) Data Retail Raw dengan Order Month Out[ ]: province product\_id brand quantity item\_price order\_month order\_id order\_date customer\_id **0** 1703458 2019-10-17 P1910 BRAND\_J 740000 14004 Jakarta Selatan DKI Jakarta 2019-10 10 **1** 1706815 2019-10-24 17220 Jakarta Selatan DKI Jakarta P2934 BRAND\_R 604000 2019-10 **2** 1710718 2019-11-03 16518 Jakarta Utara DKI Jakarta P0908 BRAND\_C 1045000 2019-11 **3** 1683592 2019-08-19 16364 Jakarta Barat DKI Jakarta P0128 BRAND\_A 205000 2019-08 15696 **4** 1702573 2019-10-16 DKI Jakarta P2968 BRAND\_R 4475000 2019-10 Jakarta Timur P0710 BRAND\_C 520000 **5** 1672906 2019-07-16 12748 Jakarta Utara DKI Jakarta 2019-07 **6** 1711399 2019-11-04 16791 P0860 BRAND\_C 1465000 2019-11 Jakarta Barat DKI Jakarta **7** 1695367 2019-09-26 13069 Surakarta Jawa Tengah P3342 BRAND\_S 205000 2019-09 **8** 1741846 2019-12-30 16873 P3203 BRAND\_S 32 450000 2019-12 Jakarta Barat DKI Jakarta 149000 2019-11 **9** 1720189 2019-11-24 14723 Tangerang Banten P1701 BRAND\_H 3. Penambahan Kolom GMV (Gross Merchandise Value) pada **Dataset** In [ ]: #GMV = harga barang \* kuantitas dataset['gmv'] = dataset['item\_price'] \* dataset['quantity'] #Tampilkan ukuran data print('Ukuran dataset') print('%d baris dan %d kolom\n' %dataset.shape) #Tampilkan 5 data teratas print('Lima data teratas:') dataset.head(10) Ukuran dataset 5000 baris dan 11 kolom Lima data teratas: Out[ ]: order\_id order\_date customer\_id city province product\_id brand quantity item\_price order\_month gm Jakarta DKI 14004 P1910 BRAND\_J 740000 **0** 1703458 2019-10-17 10 740000 2019-10 Selatan Jakarta DKI Jakarta **1** 1706815 2019-10-24 17220 P2934 BRAND R 604000 120800 2019-10 Selatan Jakarta DKI Jakarta **2** 1710718 2019-11-03 16518 P0908 BRAND C 1045000 2019-11 836000 Utara Jakarta DKI Jakarta **3** 1683592 2019-08-19 16364 P0128 BRAND A 205000 2019-08 82000 Jakarta Barat DKI Jakarta **4** 1702573 2019-10-16 15696 P2968 BRAND\_R 4475000 2019-10 895000 Timur Jakarta Jakarta DKI **5** 1672906 2019-07-16 12748 P0710 BRAND C 520000 208000 2019-07 Utara Jakarta Jakarta DKI P0860 BRAND C **6** 1711399 2019-11-04 16791 1465000 2019-11 146500 Jakarta Barat Jawa **7** 1695367 2019-09-26 13069 Surakarta P3342 BRAND S 205000 2019-09 41000 Tengah Jakarta DKI **8** 1741846 2019-12-30 16873 P3203 BRAND S 450000 2019-12 1440000 Jakarta Barat **9** 1720189 2019-11-24 14723 Tangerang Banten P1701 BRAND\_H 149000 2019-11 29800 4. Membuat Data Agregat In [ ]: #Lakukan penjumlahan kolom 'gmv' lalu kelompokkan berdasarkan kolom order\_month monthly\_amount = dataset.groupby('order\_month')['gmv'].sum().reset\_index() #Tampilkan hasil print('Sum `GMV` Grouping By order\_month') monthly\_amount Sum `GMV` Grouping By order\_month Out[]: order\_month 2019-07 3524041000 1 2019-08 4452923000 2 2019-09 3947002000 2019-10 6719937000 3 2019-11 6182229000 2019-12 8148235000 5 5. Plot Pertama: Membuat Line Chart Trend Pertumbuhan **GMV** In [ ]: #import library matplotlib untuk visualisasi import matplotlib.pyplot as plt #Definisikan nilai x = monthly\_amount['order\_month'] dan y = monthly\_amount['gmv'] plt.plot(monthly\_amount['order\_month'], monthly\_amount['gmv']) #Tampilkan grafik plt.show() monthly\_amount['order\_month'] 2019-08 2019-09 2019-10 2019-11 2019-12 Out[ ]: 0 2019-07 2019-08 2019-09 2019-10 2019-11 2019-12 Name: order\_month, dtype: object 6. Cara Alternatif: Fungsi .plot() pada pandas Dataframe In [ ]: #import library matplotlib untuk visualisasi import matplotlib.pyplot as plt #Lakukan agregasi (sum gmv grouping by order\_month) lalu plot dataset.groupby(['order\_month'])['gmv'].sum().plot() #Tampilkan hasil plt.show() 2019-08 2019-09 2019-10 2019-11 2019-12 order\_month 7. Mengubah Figure Size In [ ]: #import library matplotlib untuk visualisasi import matplotlib.pyplot as plt **#Ubah ukuran figure** plt.figure(figsize=(15,5)) #Lakukan agregasi (sum gmv grouping by order\_month) lalu plot dataset.groupby(['order\_month'])['gmv'].sum().plot() #Tampilkan hasil plt.show() 2019-07 2019-08 2019-09 2019-10 2019-11 2019-12 order\_month 8. Menambahkan Title and Axis Labels In [ ]: #import library matplotlib untuk visualisasi import matplotlib.pyplot as plt #Ubah ukuran figure (panjang x lebar) plt.figure(figsize=(15,5)) #Lakukan agregasi (sum gmv grouping by order\_month) lalu plot dataset.groupby(['order\_month'])['gmv'].sum().plot() #Menambahkan judul pada grafik plt.title('Monthly GMV Year 2019') #Menambahkan label/keterangan pada x-axis plt.xlabel('Order Month') #Menambahkan label/keterangan pada y-axis plt.ylabel('Total GMV') #Tampilkan hasil plt.show() Monthly GMV Year 2019 Total GMV 2019-08 2019-11 2019-12 2019-07 2019-09 2019-10 Order Month 9. Kustomisasi Title and Axis Labels In [ ]: #import library matplotlib untuk visualisasi import matplotlib.pyplot as plt #Ubah ukuran figure (panjang x lebar) plt.figure(figsize=(15,5)) #Lakukan agregasi (sum gmv grouping by order\_month) lalu plot dataset.groupby(['order\_month'])['gmv'].sum().plot() #Menambahkan judul pada grafik plt.title('Monthly GMV Year 2019', loc='center', fontsize=20, color='blue', pad=20) #Menambahkan label/keterangan pada x-axis plt.xlabel('Order Month', fontsize=15) #Menambahkan label/keterangan pada y-axis plt.ylabel('Total Amount', fontsize=15) #Tampilkan hasil plt.show() Monthly GMV Year 2019 Total Amount 2019-07 2019-08 2019-09 2019-10 2019-11 2019-12 Order Month 10. Kustomasi Grafik Plot In [ ]: #import library matplotlib untuk visualisasi import matplotlib.pyplot as plt #Ubah ukuran figure (panjang x lebar) plt.figure(figsize=(15,5)) #Lakukan agregasi (sum gmv grouping by order\_month) lalu plot dengan warna hijau, penanda ti tik = 'o', ketebalan = 2 dan garis -.dataset.groupby(['order\_month'])['gmv'].sum().plot(color='green', marker='o', linestyle='-.' , linewidth=2) #Menambahkan judul pada grafik plt.title('Monthly GMV Year 2019', loc='center', pad=20, fontsize=20, color='blue') #Menambahkan label/keterangan pada x-axis plt.xlabel('Order Month', fontsize=15) #Menambahkan label/keterangan pada y-axis plt.ylabel('Total Amount', fontsize=15) #Tampilkan hasil plt.show() Monthly GMV Year 2019 Total 2019-08 2019-11 2019-07 2019-09 2019-10 2019-12 Order Month 11. Kustomasi Grid In [ ]: #import library matplotlib untuk visualisasi import matplotlib.pyplot as plt #Ubah ukuran figure (panjang x lebar) plt.figure(figsize=(15,5)) #Lakukan agregasi (sum gmv grouping by order\_month) lalu plot dengan warna hijau, penanda ti tik = 'o', ketebalan = 2 dan garis -.dataset.groupby(['order\_month'])['gmv'].sum().plot(color='green', marker='o', linestyle='-.' , linewidth=2) #Menambahkan judul pada grafik plt.title('Monthly GMV Year 2019', loc='center', pad=20, fontsize=20, color='blue') #Menambahkan label/keterangan pada x-axis plt.xlabel('Order Month', fontsize=15) #Menambahkan label/keterangan pada y-axis plt.ylabel('Total Amount', fontsize=15) #Tambahkan grid (garis bantu bayangan) dengan ketebalan 1.5 dan tipe garis = : plt.grid(color='darkgray', linestyle=':', linewidth=1.5) #Tampilkan hasil plt.show() Monthly GMV Year 2019 **Total Amount** 2019-07 2019-08 2019-11 2019-12 Order Month 12. Kustomasi Axis Ticks In [ ]: #import library matplotlib untuk visualisasi import matplotlib.pyplot as plt #Ubah ukuran figure (panjang x lebar) plt.figure(figsize=(15,5)) #Lakukan agregasi (sum gmv grouping by order\_month) lalu plot dengan warna hijau, penanda ti tik = 'o', ketebalan = 2 dan garis dataset.groupby(['order\_month'])['gmv'].sum().plot(color='green', marker='o', linestyle='-.' , linewidth=2) #Menambahkan judul pada grafik plt.title('Monthly GMV Year 2019', loc='center', pad=20, fontsize=20, color='blue') #Menambahkan label/keterangan pada x-axis plt.xlabel('Order Month', fontsize=15) #Menambahkan label/keterangan pada y-axis plt.ylabel('Total Amount (milyar)', fontsize=15) #Tambahkan grid (garis bantu bayangan) dengan ketebalan 1.5 dan tipe garis = : plt.grid(color='darkgray', linestyle=':', linewidth=1.5) #Dapatkan nilai y (variabel labels) dan ticklabel object (locations) labels, locations = plt.yticks() #Ubah nilai y menjadi lebih sederhana dengan membagi nilai y dengan 10^9 --> ubah tipedatany a menjadi int plt.yticks(labels, (labels/1000000000).astype(int)) #Tampilkan hasil plt.show() Monthly GMV Year 2019 Total Amount (milyar) 2019-08 2019-09 2019-10 2019-11 2019-12 2019-07 Order Month 13. Menentukan Batas Minimum dan Maksimum Axis Ticks In [ ]: #import library matplotlib untuk visualisasi import matplotlib.pyplot as plt #Ubah ukuran figure (panjang x lebar) plt.figure(figsize=(15,5)) #Lakukan agregasi (sum gmv grouping by order\_month) lalu plot dengan warna hijau, penanda ti tik = 'o', ketebalan = 2 dan garis -.dataset.groupby(['order\_month'])['gmv'].sum().plot(color='green', marker='o', linestyle='-.' , linewidth=2) #Menambahkan judul pada grafik plt.title('Monthly GMV Year 2019', loc='center', pad=20, fontsize=20, color='blue') #Menambahkan label/keterangan pada x-axis plt.xlabel('Order Month', fontsize=15) #Menambahkan label/keterangan pada y-axis plt.ylabel('Total Amount', fontsize=15) #Tambahkan grid (garis bantu bayangan) dengan ketebalan 1.5 dan tipe garis = : plt.grid(color='darkgray', linestyle=':', linewidth=1.5) #Seting plot grafik dengan nilai minimum pada y sebesar 0 plt.ylim(ymin=0) #Dapatkan nilai y (variabel labels) dan ticklabel object (locations) labels, locations = plt.yticks() #Ubah nilai y menjadi lebih sederhana dengan membagi nilai y dengan 10^9 --> ubah tipedatany a menjadi int plt.yticks(labels, (labels/1000000000).astype(int)) #Tampilkan hasil plt.show() Monthly GMV Year 2019 Total Amount 2019-07 2019-11 2019-12 Order Month 14. Menambahkan Informasi Pada Plot In [ ]: #import library matplotlib untuk visualisasi import matplotlib.pyplot as plt #Ubah ukuran figure (panjang x lebar) fig = plt.figure(figsize=(15,5)) #Lakukan agregasi (sum gmv grouping by order\_month) lalu plot dengan warna hijau, penanda ti tik = 'o', ketebalan = 2 dan garis -.dataset.groupby(['order\_month'])['gmv'].sum().plot(color='green', marker='o', linestyle='-.' , linewidth=2) #Menambahkan judul pada grafik plt.title('Monthly GMV Year 2019', loc='center', pad=20, fontsize=20, color='blue') #Menambahkan label/keterangan pada x-axis plt.xlabel('Order Month', fontsize=15) #Menambahkan label/keterangan pada y-axis plt.ylabel('Total Amount', fontsize=15) #Tambahkan grid (garis bantu bayangan) dengan ketebalan 1.5 dan tipe garis = : plt.grid(color='darkgray', linestyle=':', linewidth=1.5) #Seting plot grafik dengan nilai minimum pada y sebesar 0 plt.ylim(ymin=0) #Dapatkan nilai y (variabel labels) dan ticklabel object (locations) labels, locations = plt.yticks() #Ubah nilai y menjadi lebih sederhana dengan membagi nilai y dengan 10^9 --> ubah tipedatany plt.yticks(labels, (labels/1000000000).astype(int)) #Berikan text pada grafik berupa text penjelasan plt.text(0.45,0.70, 'The GMV increased significantly on October 2019', transform=fig.transFi gure, color='red') #Tampilkan hasil plt.show() Monthly GMV Year 2019 Total Amount 2019-07 2019-09 2019-11 2019-12 Order Month 16. Pengaturan Parameter untuk Menyimpan Gambar In [ ]: #import library matplotlib untuk visualisasi import matplotlib.pyplot as plt #Ubah ukuran figure (panjang x lebar) fig = plt.figure(figsize=(15,5)) #Lakukan agregasi (sum gmv grouping by order\_month) lalu plot dengan warna hijau, penanda ti tik = 'o', ketebalan = 2 dan garis -.dataset.groupby(['order\_month'])['gmv'].sum().plot(color='green', marker='v', linestyle='-.' , linewidth=2) #Menambahkan judul pada grafik plt.title('Monthly GMV Year 2019', loc='center', pad=20, fontsize=20, color='blue') #Menambahkan label/keterangan pada x-axis plt.xlabel('Order Month', fontsize=15) #Menambahkan label/keterangan pada y-axis plt.ylabel('Total Amount', fontsize=15) #Tambahkan grid (garis bantu bayangan) dengan ketebalan 1.5 dan tipe garis = : plt.grid(color='darkgray', linestyle=':', linewidth=1.5) #Seting plot grafik dengan nilai minimum pada y sebesar 0 plt.ylim(ymin=0) #Dapatkan nilai y (variabel labels) dan ticklabel object (locations) labels, locations = plt.yticks() #Ubah nilai y menjadi lebih sederhana dengan membagi nilai y dengan 10^9 --> ubah tipedatany a menjadi int plt.yticks(labels, (labels/1000000000).astype(int)) #Berikan text pada grafik berupa text penjelasan plt.text(0.45,0.72, 'The GMV increased significantly on October 2019', transform=fig.transFi gure, color='red') #save grafik dengan format png dan kualitas 95 (pada colab path default : /content/monthly\_g mv.png) plt.savefig('monthly\_gmv.jpg', quality=95) #Tampilkan hasil plt.show() Monthly GMV Year 2019 Total Amount 1 2019-08 2019-09 2019-10 2019-11 2019-12 2019-07 Order Month **Mini Project: Daily Number of Customers on Desember** In [ ]: # Import library yang dibutuhkan import datetime import pandas as pd import matplotlib.pyplot as plt # Baca dataset retail\_raw\_reduced.csv dataset = pd.read\_csv('https://storage.googleapis.com/dqlab-dataset/retail\_raw\_reduced.csv') # Buat kolom order\_month dataset['order\_month'] = dataset['order\_date'].apply(lambda x: datetime.datetime.strptime(x, "%Y-%m-%d").strftime('%Y-%m')) # Buat kolom gmv dataset['gmv'] = dataset['item\_price'] \* dataset['quantity'] # Plot grafik sesuai dengan instruksi plt.figure(figsize=(10, 5)) dataset[dataset['order\_month'] == '2019-12'].groupby(['order\_date'])['customer\_id'].nunique(). plot(color='red', marker='.', linewidth=2) plt.title('Daily Number of Customers - December 2019', pad=30, loc='left', fontsize=20, colo r='orange') plt.xlabel('Order Date', fontsize=15, color='blue') plt.ylabel('Number of Customers', fontsize=15, color='blue') plt.grid(color='darkgray', linestyle=':', linewidth=0.5) plt.ylim(ymin=0) plt.show() Daily Number of Customers - December 2019 40 35 of Customers 25 20 ē 15 10 2019-12-01 2019-12-06 2019-12-12 2019-12-18 2019-12-24 2019-12-30 **Order Date** 

Visualisasi Data dengan Matplotlib - Part I

In [3]: #import library pandas untuk membaca file csv (comma separated value)

dataset = pd.read\_csv('https://storage.googleapis.com/dqlab-dataset/retail\_raw\_reduced.csv')

province product\_id

P1910 BRAND J

P2934 BRAND\_R

P0908 BRAND C

DKI Jakarta

DKI Jakarta

DKI Jakarta

DKI Jakarta

city

14004 Jakarta Selatan

17220 Jakarta Selatan

Jakarta Utara

Jakarta Barat

16518

16364

brand quantity item\_price

2

740000

604000

1045000

1. Pengenalan Dataset

#import dataset csv dari API dqlab

print('%d baris dan %d kolom\n' %dataset.shape)

import pandas as pd

#Tampilkan ukuran data print('Ukuran dataset')

5000 baris dan 9 kolom

Sepuluh data teratas:

**0** 1703458 2019-10-17

**1** 1706815 2019-10-24

**2** 1710718 2019-11-03

**3** 1683592 2019-08-19

dataset.head(10)

Ukuran dataset

Out[3]:

#Tampilkan 10 data teratas print('Sepuluh data teratas:')

order\_id order\_date customer\_id