JUSIKOM PRIMA (Jurnal Sistem Informasi dan Ilmu Komputer Prima)
Vol. 4 No. 2, Februari 2021 E-ISSN : 2580-2879

ANALISIS BIG DATA DENGAN METODE *EXPLORATORY DATA ANALYSIS* (EDA) DAN METODE VISUALISASI MENGGUNAKAN JUPYTER NOTEBOOK

Muhammad Radhi, Amalia, Daniel Ryan Hamonangan Sitompul, Stiven Hamonangan Sinurat, Evta Indra
Program Studi Sistem Informasi, Fakultas Teknologi dan Ilmu Komputer, Universitas Prima Indonesia
Jalan Sampul
E-mail : muhammad_radhi05@yahoo.com

ABSTRAK- Dengan adanya kebutuhan akan informasi dari data yang digunakan dalam kegiatan bisnis, maka perlu dilakukan eksplorasi data untuk mengetahui informasi tersebut. Dalam proses eksplorasi data dapat dilakukan dengan menggunakan grafik, sedangkan dalam penggunaan grafik tersebut dapat berguna untuk mengidentifikasi pola-pola yang ada pada data. Penelitian ini dilakukan dengan tujuan untuk memperoleh informasi dari hasil eksplorasi data dengan menggunakan data penjualan barang elektronik dalam memperoleh informasi sebagai bahan pertimbangan dalam merencanakan strategi peningkatan usaha. Bahwa dalam penjualan barang elektronik dari bulan januari 2019 – desember 2019 yang paling banyak terjual adalah baterai AAA(4 packs) sedangkan barang yang paling sedikit terjual adalah LG Dryer. Dan untuk produk elektronik yang paling mahal terjual sepanjang bulan januari – desember adalah Macbook Pro Laptop dengan kisaran harga sebesar 1750 dolar.

Kata kunci : *Exploratory Data Analysis, Big Data, Elektronik, Sales Prediction*

1. PENDAHULUAN

Big Data merupakan teknologi baru dalam teknologi informasi yang memungkinkan pengelolaan, penyimpanan, dan analisis data dalam berbagai bentuk dalam jumlah besar[1]. Disebut *Big Data* karena data yang sangat besar semakin sulit untuk digunakan sebagai informasi. Munculnya *Big Data* telah menimbulkan dua hal, yaitu kelebihan dan kekurangan. Manfaat yang disebutkan sebelumnya adalah memudahkan manusia. Sedangkan kekurangannya tidak semuanya benar, beberapa pihak yang tidak bertanggung jawab memanipulasi data. Oleh karena itu, diperlukan pengetahuan yang mendalam tentang *Big Data* agar data dapat digunakan secara optimal.

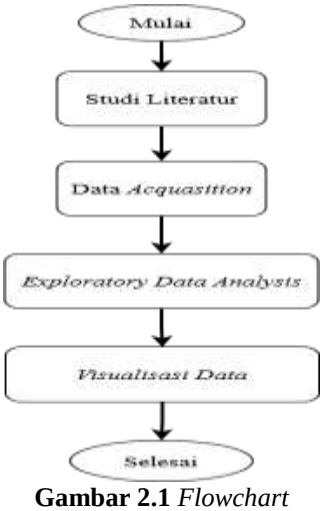
Dengan adanya kebutuhan akan informasi dari data yang digunakan dalam kegiatan bisnis, maka perlu dilakukan eksplorasi data untuk mengetahui informasi tersebut. Dalam proses eksplorasi data dapat dilakukan dengan menggunakan grafik, sedangkan dalam penggunaan grafik tersebut dapat berguna untuk mengidentifikasi pola-pola yang ada pada data.[2] Analisis yang akan dilakukan pada proses eksplorasi menggunakan metode visual data (VD) dan *exploratory data analysis (EDA)*. Dalam pembuatan visualisasinya akan digunakan Notebook Jupyter.

Berdasarkan latar belakang masalah yang ada maka diperoleh beberapa rumusan masalah yaitu mengetahui bagaimana cara mengimplementasikan metode *Exploratory Data Analysis (EDA)* dan metode *Visual Data (VD)* ke dalam pengolahan data penjualan yang dihasilkan dari eksplorasi apa yang didapat setelah menerapkan metode ini.

Penelitian ini dilakukan dengan tujuan untuk memperoleh informasi dari hasil eksplorasi data dengan menggunakan data penjualan barang elektronik dalam memperoleh informasi sebagai bahan pertimbangan dalam merencanakan strategi peningkatan usaha

2. METODE PENELITIAN

Pada penelitian ini menggunakan metode *Exploratory Data Analysis* (EDA) dan *Visual Data(VD).*



Gambar 2.1 *Flowchart*

2.1 Studi Literatur

Metode studi literatur adalah serangkaian kegiatan yang berkenaan dengan metode pengumpulan data pustaka, membaca dan mencatat, serta mengelolah bahan penelitian[3].

2.2 *Data Acqusition*

Data Acqusition adalah alat atau proses yang digunakan untuk mengumpulkan informasi untuk tujuan menganalisa suatu gejala fisik atau elektrik seperti voltase, arus, suhu, tekanan, atau suara menggunakan komputer.[4] Data yang diakusisi pada penelitian ini adalah data penjualan elektronik selama 1 tahun.

2.3 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) adalah proses menganalisis dan menampilkan data bertujuan mendapatkan pemahaman yang lebih baik tentang wawasan dari data[5]. Ada berbagai langkah yang dilakukan saat melakukan EDA, berikut ini adalah langkah - langkah umum yang dapat diambil dalam melakukan analisis EDA data:

a. Memaksimalkan wawasan ke dalam kumpulan data.
b. Mengungkap struktur data.
c. Ekstrak variabel yang penting.
d. Mendeteksi outlier dan anomali.
e. Melakukan uji asumsi.
f. Mengembangkan model.
g. Menentukan faktor yang optimal.

Kebanyakan teknik *EDA* adalah berbentuk grafis dengan beberapa teknik kuantitatif. Peran utama EDA adalah untuk mengeksplorasi data secara terbuka, dan grafik bertujuan memperkuat analisis yang dilakukan.[6] Berikut adalah beberapa jenis teknik grafis sederhana yang banyak digunakan

1. Plotting data mentah seperti data traces, histograms, bihistograms, probability plots, lag plots, block plots, dan Youden plots.
2. Plotting statistik sederhana seperti mean plots, standard deviation plots, box plots.

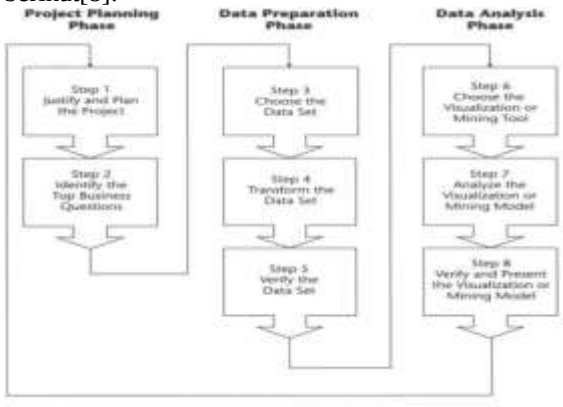
Exploratory Data Analysis (EDA) memiliki 5 tahapan yaitu[7]

Gambar 1. *Exploratory Data Analysis (EDA)*

Pada metode EDA, tidak diikuti dengan penerapan model, melainkan diikuti dengan tujuan yang menyimpulkan model apa yang cocok untuk digunakan.

2.4 Visual Data Mining

Visual data mining merupakan metode yang akan digunakan dalam penelitian ini. Metode visual data mining ini terdapat 3 tahapan yaitu sebagai berikut[8]:



Gambar 2 Tahapan Visual Data[8]

Pada metode ini, terdapat tiga tahapan yaitu [8]:

1. Project Planning Phase.
   Tahap ini terdiri dari justify and plan project yang bertujuan untuk melakukan identifikasi tujuan dan melakukan penentuan scope, dan identify the top business question bertujuan untuk mengetahui kebutuhan yang diperlukan dalam pembuatan visualisasi dengan dilakukan diskusi terhadap narasumber.
2. Data Preparation Phase
   Tahap ini terdiri dari choose the data set bertujuan untuk melakukan pemilihan data yang akan digunakan dalam penelitian, transform the data set bertujuan untuk melakukan transformasi data seperti cleansing dan filtering. dan verify the data set bertujuan untuk melakukan cross check untuk dilihat kembali apakah terdapat kesalahan pada data.
3. Data Analysis Phase.
   Tahap ini terdiri darichoose the visualization or mining toolsbertujuan untuk memilih tools dan model chart yang akan digunakan, analyze the Visualization or Mining Model bertujuan untuk menganalisis hasil dari visualisasi yang sudah dibuat sebelumnya, dan verify and present the visualization or mining model bertujuan untuk memverifikasi dan mempresentasikan hasil visualisasi kepada pengguna untuk menunjukkan informasi apa yang didapatkan dari visualisasi tersebut.

3. HASIL DAN PEMBAHASAN

3.1 Analisis Masalah

Disebut *Big Data* karena data yang sangat besar semakin sulit untuk digunakan sebagai informasi. Munculnya *Big Data* telah menimbulkan dua hal, yaitu kelebihan dan kekurangan. Manfaat yang disebutkan sebelumnya adalah memudahkan manusia

3.2 Data Akusisi

Data yang digunakan yaitu terkait dengan data penjualan elektronik tahun 2019 dimulai dari bulan januari hingga desember, data bersumber dari[9]. Dimana dataset tersebut memiliki 217244 data didalam 6 kolom.

```
<bound method NDFrame.head of       Order ID                    Product  Quantity
0       176558          USB-C Charging Cable         2     11.95
1          NaN                           NaN       NaN       NaN
2       176559  Bose SoundSport Headphones         1     99.99
3       176560                Google Phone         1    600
4       176560            Wired Headphones         1     11.99
...        ...                           ...       ...       ...
217239  248166  Bose SoundSport Headphones         1     99.99
217240  248167      AAA Batteries (4-pack)         3      2.99
217241  248168       AA Batteries (4-pack)         1      3.84
217242  248169          USB-C Charging Cable       1     11.95
217243  248170       AA Batteries (4-pack)         1      3.84

            Order Date                  Purchase Address
0       04/19/19 08:46          917 1st St, Dallas, TX 75001
1                  NaN                                   NaN
2       04/07/19 22:30      682 Chestnut St, Boston, MA 02215
3       04/12/19 14:38     669 Spruce St, Los Angeles, CA 90001
4       04/12/19 14:38     669 Spruce St, Los Angeles, CA 90001
...               ...                                   ...
217239  09/29/19 12:19  868 Hickory St, San Francisco, CA 94016
217240  09/31/19 16:26       208 Lakeview St, Boston, MA 02215
217241  09/02/19 07:25          568 12th St, Seattle, WA 98101
217242  09/08/19 12:10  806 Walnut St, San Francisco, CA 94016
217243  09/16/19 08:13       738 Park St, Los Angeles, CA 90001

[217244 rows x 6 columns]>
```

Gambar 3 Dataset

3.3 *Exploratory Data Analysis (EDA)*

Sistem yang mampu memvisualisasikan data menggunakan metode *EDA* dan Visualisasi data untuk membantu dalam menentukan penjualan barang elektronik.

1. *Import Library*

Memasukkan library yang dibutuhkan untuk pengelolaan data

Source Code:

```
import os
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

2. *Merge* Dataset

Dataset penjualan barang elektronik dari bulan januari sampai desember dijadikan menjadi 1 dataset untuk dilakukan pengolahan data.

Source Code:

```
files=[file for file in os.listdir
("F:/EDA_projects/Sales_Analysis/SalesAna
lysis/Sales_Data")]
for file in files:
    print(file)
all_data.to_csv('F:/EDA_projects/Sales_An
alysis/SalesAnalysis/Sales_Data/all_data.
csv',index=False)
```

3. *Assesing* Dataset

Pengecekan data kosong serta dilakukan penghapusan terhadap data yang kosong tersebut

Source Code:

```
all_data.isnull().sum()
```

```
Order ID              545
Product               545
Quantity Ordered      545
Price Each            545
Order Date            545
Purchase Address      545
Month              186850
sales              186850
city               186850
Hour               186850
dtype: int64
```

Gambar 4 Jumlah Data Kosong

```
all_data = all_data.dropna(how='all')
all_data.shape
```

```
(558205, 10)
```

Gambar 5 Jumlah Data Dihapus

4. *Filtering* Dataset

Filtering dilakukan menurut month dan order date.

Source Code:

```
all_data['Month']=all_data['Order
Date'].apply(month)
all_data['Month'].unique()
filter=all_data['Month']=='Order Date'
len(all_data[~filter])
all_data=all_data[~filter]
all_data.head()
```

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Month
0	176558	USB-C Charging Cable	2	11.95	04/19/19 08:46	917 1st St, Dallas, TX 75001	04
2	176559	Bose SoundSport Headphones	1	99.99	04/07/19 22:30	682 Chestnut St, Boston, MA 02215	04
3	176560	Google Phone	1	600	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001	04
4	176560	Wired Headphones	1	11.99	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001	04
5	176561	Wired Headphones	1	11.99	04/30/19 09:27	333 8th St, Los Angeles, CA 90001	04

Gambar 6 *Filtering Dataset*

5. Data *Casting*

Data *Casting* adalah mengubah nilai yang awalnya dari tipe data a, menjadi tipe data b tipe untuk dilakukan visualiasi data. Data *Casting* dataset pada kolom *Month* menjadi *integer*, *Price Each* menjadi *float*, *Quantity Ordered* Menjadi *integer*

Source Code:

```
all_data['Month']=all_data['Month'].astyp
e(int)
all_data['Price     Each']=all_data['Price
Each'].astype(float)
all_data['Quantity
Ordered']=all_data['Quantity
Ordered'].astype(int)
all_data['sales']=all_data['Quantity
Ordered']*all_data['Price Each']
all_data.head(5)
```

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Month	sales
0	176558	USB-C Charging Cable	2	11.95	04/19/19 08:46	917 1st St, Dallas, TX 75001	4	23.90
2	176559	Bose SoundSport Headphones	1	99.99	04/07/19 22:30	682 Chestnut St, Boston, MA 02215	4	99.99
3	176560	Google Phone	1	600.00	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001	4	600.00
4	176560	Wired Headphones	1	11.99	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001	4	11.99
5	176561	Wired Headphones	1	11.99	04/30/19 09:27	333 8th St, Los Angeles, CA 90001	4	11.99

Gambar 7 Data *Casting*

3.4 Visualisasi Data

Visualisasi adalah rekayasa dalam pembuatan gambar, diagram atau animasi untuk penampilan suatu informasi dalam penjelasan lain visualisasi adalah konversi data ke dalam format visual atau tabel sehingga karakteristik dari data dan relasi diantara item data atau atribut dapat di analisis atau dilaporkan, dan visualisasi data adalah satu dari yang teknik paling baik dan menarik untuk eksplorasi data.[10]

1. *Grouping* Data *Month*

Membuat group data total penjualan setiap bulan untuk mempermudah dalam menyajikan visualisasi data.

Source Code:

```
all_data.groupby('Month')['sales'].sum()
```

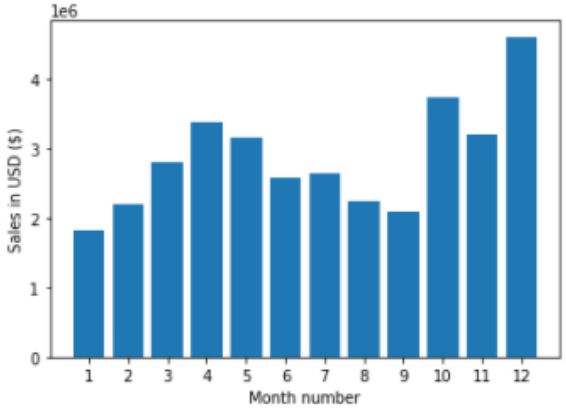
```
Month
1     1.822257e+06
2     2.202022e+06
3     2.807100e+06
4     3.390670e+06
5     3.152607e+06
6     2.577802e+06
7     2.647776e+06
8     2.244468e+06
9     2.097560e+06
10    3.736727e+06
11    3.199603e+06
12    4.613443e+06
Name: sales, dtype: float64
```

Gambar 8 *Grouping* Data *Month*

Dari data gambar 8 ditampilkan kedalam diagram batang untuk mempermudah membaca data.

Source Code:

```
months=range(1,13)
plt.bar(months,all_data.groupby('Month')[
'sales'].sum())
plt.xticks(months)
plt.ylabel('Sales in USD ($)')
plt.xlabel('Month number')
plt.show()
```



Gambar 9 *Diagram* Batang

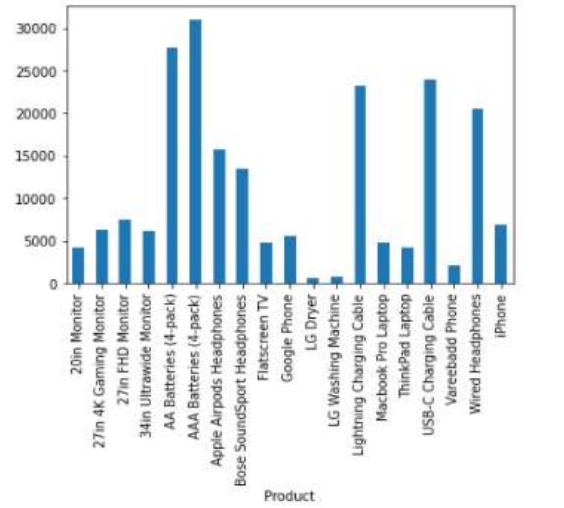
2. *Grouping* Data *Product*

Setelah membuat group data total penjualan setiap bulan untuk mempermudah dalam menyajikan visualisasi data, tahap selanjutnya membuat group data dari *product* dan *quantity ordered.*

Source Code:

```
all_data.groupby('Product')['Quantity
Ordered'].sum().plot(kind='bar')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0xe60459c48>
```



Gambar 10 *Grouping* Data *Product*\

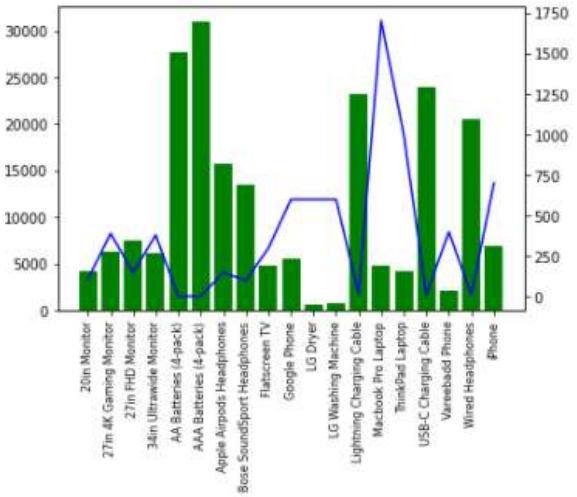
3. *Grouping* Data *Product* dan *Price Each*

Untuk melihat apakah ada pengaruh harga terhadap barang yang paling banyak terjual setiap bulannya maka dapat lihat pada gambar 11 berikut:

Source Code:

```
products=all_data.groupby('Product')['Qua
ntity Ordered'].sum().index
quantity=all_data.groupby('Product')['Qua
ntity Ordered'].sum()
prices=all_data.groupby('Product')['Price
Each'].mean()
plt.figure(figsize=(40,24))
fig,ax1 = plt.subplots()
ax2=ax1.twinx()
ax1.bar(products, quantity, color='g')
ax2.plot(products, prices, 'b-')
ax1.set_xticklabels(products,
rotation='vertical', size=8)
```

```
[Text(0, 0, '20in Monitor'),
 Text(0, 0, '27in 4K Gaming Monitor'),
 Text(0, 0, '27in FHD Monitor'),
 Text(0, 0, '34in Ultrawide Monitor'),
 Text(0, 0, 'AA Batteries (4-pack)'),
 Text(0, 0, 'AAA Batteries (4-pack)'),
 Text(0, 0, 'Apple Airpods Headphones'),
 Text(0, 0, 'Bose SoundSport Headphones'),
 Text(0, 0, 'Flatscreen TV'),
 Text(0, 0, 'Google Phone'),
 Text(0, 0, 'LG Dryer'),
 Text(0, 0, 'LG Washing Machine'),
 Text(0, 0, 'Lightning Charging Cable'),
 Text(0, 0, 'Macbook Pro Laptop'),
 Text(0, 0, 'ThinkPad Laptop'),
 Text(0, 0, 'USB-C Charging Cable'),
 Text(0, 0, 'Vareebadd Phone'),
 Text(0, 0, 'Wired Headphones'),
 Text(0, 0, 'iPhone')]
<Figure size 2880x1728 with 0 Axes>
```



Gambar 11 *Product* dan *Price Each*

4. KESIMPULAN DAN SARAN

Pada penlitian ini dapat diarik kesimpulan bahwa dalam penjualan barang elektronik dari bulan januari 2019 – desembar 2019 yang paling banyak terjual adalah baterai AAA(4 packs) sedangkan barang yang paling sedikit terjual adalah LG Dryer. Dan untuk produk elektronik yang paling mahal terjual sepanjang bulan januari – desember adalah Macbook Pro Laptop dengan kisaran harga sebesar 1750 dolar.

Penelitian ini hanya menggunakan diagram bar sehingga diharapkan kepada peneliti selanjutnya dapat mengembangkan model visualisasi data seperti *pie chart, scatter plot*, dan *box plot*

DAFTAR PUSTAKA

[1] O. Media, *Change the world with data . We ' ll show you how* . 2012.

[2] A. Rijali, "Analisis Data Kualitatif Ahmad Rijali UIN Antasari Banjarmasin," vol. 17, no. 33, pp. 81–95, 2018.

[3] E. D. Kartiningrum, "Panduan Penyusunan Studi Literatur," *Lemb. Penelit. dan Pengabdi. Masy. Politek. Kesehat. Majapahit, Mojokerto*, pp. 1–9, 2015.

[4] E. Koutroulis and K. Kalaitzakis, "Development of an integrated data-acquisition system for renewable energy sources systems monitoring," *Fuel Energy Abstr.*, vol. 44, no. 3, p. 163, 2003, doi: 10.1016/s0140-6701(03)81847-7.

[5] A. Chandra, "Memahami Data Dengan Exploratory Data Analysis," 2019. https://medium.com/data-folks-indonesia/memahami-data-dengan-exploratory-data-analysis-a53b230cce84.

[6] Izza Aditya R, "Exploratory Data Analysis." https://medium.com/@izza.aditya.12.ia/exploratory-data-analysis-c0b6bf5ae706.

[7] D. Aryanti and J. Setiawan, "Visualisasi Data Penjualan dan Produksi PT Nitto Alam Indonesia Periode 2014-2018," *Ultim. InfoSys*, vol. 9, no. 2, pp. 86–91, 2019, doi: 10.31937/si.v9i2.991.

[8] M. Biba, N. R. Vajjhala, and L. Nishani, "Visual data mining for collaborative filtering: A state-of-the-art survey," *Decis. Manag. Concepts, Methodol. Tools, Appl.*, vol. 3–4, no. January 2021, pp. 1274–1292, 2017, doi: 10.4018/978-1-5225-1837-2.ch059.

[9] "Dataset Sales," 2019. https://www.kaggle.com/.

[10] E. Lestariningsih, E. Ardhianto, W. T. Handoko, E. Supriyanto, and S. L. R. A, "Visualisasi Data Penduduk Berbasis Web di Kelurahan Mranggen Kecamatan Mranggen Kabupaten Demak menggunakan Highcart 5.0.6," *J. Teknol. Inf. Din.*, vol. 21, no. 2, pp. 146–153, 2016.