TUGAS

SUB-CPMK KE-1

MATA KULIAH DATA MINING (B)

"JENIS-JENIS DATA DAN MELAKUKAN ANALISIS DESKRIPTIF SERTA VISUAL UNTUK MENGETAHUI KARAKTERISTIK DATA"

DISUSUN OLEH:

Reza Putri Angga (22083010006)

DOSEN PENGAMPU:

Trimono, S.Si., M.Si. (21119950908269)

PROGRAM STUDI SAINS DATA

FAKULTAS ILMU KOMPUTER

UNIVERSITAS PEMBANGUNAN NASIONAL "VETERAN" JAWA TIMUR

2024

STUDI KASUS DAN PEMBAHASAN

Berdasarkan penjelasan dan kosep awal data mining, jawablah pertanyaan berikut.

1. Jelaskan dan berikan contoh (minimal lima) jenis-jenis data yang anda ketahui berdasarkan skala pengukuran, sumber data, serta jumlah variabel dan periode waktu.

   Jawaban dan Pembahasan :

   Menurut Kamus Besar Bahasa Indonesia (KBBI), data merupakan kumpulan fakta atau informasi yang diperoleh melalui pengamatan, pengukuran, dan penelitian yang dapat direpresentasikan dalam bentuk angka, teks, maupun gambar. Dalam konteksnya, data terdiri dari berbagai jenis yang dapat dikategorikan berdasarkan karakteristik tertentu. Meliputi skala pengukuran, sumber data, dan jumlah variabel serta periode waktu yang akan dijelaskan lebih lanjut sebagai berikut.

   A. Skala Pengukuran

   Skala pengukuran merupakan sebuah metode atau cara untuk menetapkan tingkatan atau ukuran suatu variabel data berdasarkan jenis data yang dikandungnya. Hal ini dilakukan dengan memberikan angka atau simbol untuk merepresentasikan karakteristik atau atribut yang dimaksud dalam variabel. Skala pengukuran berfungsi untuk menjadi pedoman menentukan alat ukur data kuantitatif, memfasilitasi analisis, dan menentukan interpretasi lebih lanjut. Terdapat beberapa jenis skala pengukuran untuk menentukan jenis data, di antaranya yakni.

   1. Skala Nominal

   Skala nominal merupakan jenis skala yang membedakan kategori berdasarkan jenis atau macamnya. Bertujuan untuk mengklasifikasikan objek, individu, atau kelompok menjadi kategori tertentu tanpa adanya tingkatan. Hal ini dilakukan dengan memberikan angka atau simbol untuk menunjukkan adanya perbedaan antara satu karakteristik dengan karakteristik lainnya dan untuk mengetahui ada atau tidak adanya karakteristik pada objek yang diukur. Terdapat beberapa contoh data yang memiliki skala nominal, meliputi.

   - Data golongan darah yang terdiri atas A, B, AB, dan O.
   - Data jenis kelamin yang terdiri atas laki-laki dan perempuan.
   - Data jenis kendaraan pribadi yang terdiri atas motor dan mobil.

- Data metode belajar yang terdiri atas visual, auditori, dan kinestetik.
- Data lokasi geografis yang terdiri atas dataran tinggi dan dataran rendah.
- Data status pernikahan yang terdiri atas menikah, belum menikah, dan cerai.
- Data warna rambut yang terdiri atas warna rambut hitam dan warna rambut bukan hitam.

2. Skala Ordinal

Skala ordinal merupakan jenis skala yang membedakan kategori berdasarkan tingkatan atau urutannya. Dimulai dari tingkatan yang paling rendah hinga tingkatan yang paling tinggi berdasarkan karakteristik atau atribut tertentu. Pada skala ini, setiap objek memiliki posisi relatif terhadap yang lain, namun jarak antar posisi tidak selalu sama. Hal ini dilakukan dengan memberikan angka atau simbol untuk menentukan dan menunjukkan posisi relatif suatu objek. Terdapat beberapa contoh data yang memiliki skala ordinal, meliputi.

- Data tingkat suhu panas yang terdiri atas hangat, panas, dan sangat panas.
- Data tingkat kasta di Bali yang terdiri atas brahmana, satria, waisya, dan sudra.
- Data peringkat kelas yang terdiri atas peringkat 1, peringkat 2, dan peringkat 3.
- Data tingkat kesepakatan yang terdiri atas kurang setuju, setuju, dan sangat setuju.
- Data tingkat kepuasan pelanggan yang terdiri atas tidak puas, puas, dan sangat puas.
- Data tingkatan pendidikan yang terdiri atas tidak sekolah, SD, SMP, SMA, S1, S2, dan S3.
- Data tingkat pangkat militer yang terdiri atas prajurit, sersan, letnan, kapten, mayor, letnan kolonel, kolonel, dan jendral.

3. Skala Interval

Skala Interval merupakan jenis skala yang memiliki karakteristik dari skala nominal dan skala ordinal. Kemudian, ditambahkan dengan karakteristik lain, yakni adanya interval yang tetap antara kategori atau nilai yang diukur. Dalam skala ini, perbedaan antara kategori atau nilai dengan selang atau jarak tertentu memiliki nilai yang sama. Dengan tidak memiliki nilai nol mutlak yang

menunjukkan ketiadaan variabel yang diukur. Terdapat beberapa contoh data yang memiliki skala interval, meliputi.

- Data rentang skala IQ mahasiswa, misalnya terdiri atas 100-110, 111-120, dan seterusnya.
- Data rentang skor ujian mahasiswa, misalnya yang terdiri atas 70-80, 81-90, dan seterusnya.
- Data rentang kadar kolestrol, misalnya terdiri atas 180-190 mg/dL, 191-200 mg/dL, dan seterusnya.
- Data rentang suhu di suatu daerah dalam celcius, misalnya terdiri atas 30-40C, 41-50C, dan seterusnya.
- Data rentang usia pekerja di suatu perusahaan, misalnya terdiri atas 20-30 tahun, 31-40 tahun, dan seterusnya.
- Data rentang durasi waktu yang terdiri atas jam, misalnya terdiri atas 10-20 menit, 21-30 menit, dan seterusnya.
- Data rentang pengukuran tinggi badan mahasiswa, misalnya terdiri atas 140-150 cm, 151-160 cm, dan seterusnya.

4. Skala Rasio

Skala rasio merupakan jenis skala yang memiliki karakteristik skala nominal, skala ordinal, dan skala interval. Kemudian ditambahkan dengan karakteristik lain, yakni adanya nilai nol yang bersifat mutlak dan tidak dapat diubah meskipun menggunakan skala yang lain. Nilai nol mutlak ini, memungkinkan adanya nilai perbandingan proposional yang tepat antar data yang ada. Sehingga, dapat dilakukan operasi matematika dasar seperti pembagian. Terdapat beberapa contoh data yang memiliki skala rasio, meliputi.

- Data suhu dalam kelvin, dengan suhu ruangan A sebesar 150 kelvin dan suhu ruangan B sebesar 300 kelvin. Dengan demikian, suhu di ruangan B merupakan dua kali lipat suhu di ruangan A.
- Data tinggi badan, dengan orang A memiliki tinggi sebesar 200 cm dan orang B memiliki tinggi sebesar 100 cm. Dengan demikian, tinggi badan orang A merupakan dua kali lipat tinggi badan orang B.

- Data berat badan, dengan orang A memiliki berat badan sebesar 40 kg dan orang B memiliki berat badan sebesar 80 kg. Dengan demikian, berat badan orang A merupakan dua kali lipat berat badan orang B.
- Data output produksi, dengan pabrik A menghasilkan sebesar 500 unit produk per-hari dan pabrik B menghasilkan sebesar 1000 unit produk per-hari. Dengan demikian, produksi pabrik B perhari merupakan dua kali lipat produksi pabrik A perhari.
- Data pendapatan, dengan orang A memiliki pendapatan sebesar Rp.5.000.00,00 per-bulan dan orang B memiliki pendapatan sebesar Rp.10.000.000,00 per-bulan. Dengan demikian, pendapatan orang B perbulan merupakan dua kali pendapatan orang A perbulan.
- Data luas tanah, dengan kawasan A memiliki rata-rata luas tanah sebesar 300 meter persegi dan kawasan B memiliki rata-rata luas tanah sebesar 200 meter persegi. Dengan demikian, rata-rata luas tanah di kawasan A merupakan 1,5 kali lipat rata-rata luas tanah di kawasan B.
- Data jarak tempuh, dengan mobil A menempuh jarak sebesar 10 km dengan satu liter bahan bakar dan mobil B menempuh jarak sebesar 20 km dengan satu liter bahan bakar. Dengan demikian, mobil B dapat menempuh dua kali lipat jarak mobil A dengan satu liter bahan bakar.

B. Sumber Data

Sumber data merupakan cara atau proses perolehan data yang dilakukan oleh peneliti. Meliputi cara untuk identifikasi, pengumpulan, dan penganalisisan informasi yang digunakan dalam sebuah penelitian. Terdapat dua jenis kategori data menurut sumbernya, di antaranya yakni.

1. Data Primer

Data primer merupakan data yang diperoleh oleh individu ataupun kelompok secara langsung dari sumbernya atau sumber pertama (tidak melalui perantara). Data primer sering disebut dengan data asli dengan sifat *up to date* untuk menjawab pertanyaan penelitian. Metode yang dapat dilakukan untuk mengumpulkan data primer, meliputi *survei* menggunakan pertanyaan lisan dan

tertulis, metode observasi dengan pengamatan langsung terhadap aktivitas atau kejadian tertentu dan berfokus pada observasi.

Diskusi terfokus ataupun FGD, wawancara, serta penyebaran kuesioner. Terdapat beberapa contoh data primer, meliputi data hasil survei kuesioner kepuasan pelanggan terhadap layanan sebuah restoran, data hasil catatan observasi perilaku migrasi burung-burung di suatu daerah selama musim dingin, data hasil wawancara terhadap wawasan mahasiswa terhadap pengetahuan nasional, data hasil wawancara terhadap pengusaha lokal mengenai dampak regulasi pemerintah terhadap bisnis mereka.

Data sensus penduduk yang biasnaya dilakukan oleh badan pusat statistik, data hasil diskusi terfokus mengenai cara pengembangan aktivitas belajar mengajar di sebuah kelas selama satu bulan, dan data hasil percobaan eksperimen reaksi kimia di laboratorium secara langsung.

2. Data Sekunder

Data sekunder merupakan data yang diperoleh oleh individu ataupun kelompok dari semua sumber yang sudah tersedia sebelumnya melalui media perantara. Data sekunder dapat berupa beragam jenis informasi bukti, catatan, ataupun laporan historis yang diperoleh dari berbagai sumber, seperti buku, materi, laporan, dan *website*. Terdapat beberapa contoh data sekunder, meliputi data demografi dari sensus penduduk yang dipublikasikan badan pusat statistik.

Laporan keuangan tahunan yang dipublikasikan suatu perusahaan, hasil survei konsumen yang dipublikasikan suatu restoran, data cuaca histrois yang dipublikasikan lembaga meteorologi, data kependudukan yang dipublikasikan badan pusat statistik, data impor ekspor yang dipublikasikan bea cukai, data isu suatu daerah dari media sosial, data gambar kondisi fisik suatu wilayah dari peta, dan data spesifik mengenai riset tetentu dari jurnal ilmiah.

C. Jumlah Variabel dan Periode Waktu

Jumlah variabel dan periode waktu merupakan seberapa banyak jumlah dimensi waktu dan aspek yang diamati dan dipelajari dalam suatu periode waktu tertentu. Mencakup jumlah variabel atau dimensi yang menjadi fokus analisis serta durasi waktu

yang ditentukan untuk melakukan pengumpulan data. Terdapat tiga jenis kategori data berdasarkan jumlah variabel dan periode waktu, di antaranya yakni.

1. **Data Cross Section**

   Data cross section merupakan jenis data yang terdiri atas lebih dari satu objek dan lebih dari satu variabel yang diperoleh oleh individu ataupun kelompook pada suatu titik waktu tertentu. Sehingga memberikan gambaran karakteristik yang dimiliki oleh beragam individua tau kelompok pada titik waktu tertentu. Terdapat beberapa contoh data cross section, meliputi.

   - Data sensus penduduk pada 2020 dengan objek provinsi dan kota, serta variabel usia, pekerjaan, pendapatan.
   - Data emisi gas rumah kaca diberbagai negara pada tahun 2023 dengan objek semua negara, serta variabel pengamatan emisi CO2, metana.
   - Data demografi berbagai negara pada tahun 2020 dengan objek semua negara dan wilayahnya, serta variabel tingkat kelahiran, angka harapan hidup.
   - Data survei prefensi politik di Indonesia pada tahun 2024 dengan objek responden dari berbagai daerah, serta variabel pilihan presiden, partai politik.
   - Data survei kepuasan pelanggan berbagai restoran diSurabaya pada bulan Januari dengan objek semua restoran dan pelanggan, serta variabel jenis restoran, harga, rasa makanan.
   - Data penjualan retail dari beberapa *merk* pakaian pada tahun 2020 dengan objek berbagai *merk* pakaian dan konsumen serta variabel jumlah penjualan, tanggal penjualan.
   - Data harga saham diperusahaan A, B, dan C dengan objek harga saham disemua perusahaan tersebut serta variabel harga saham dalam mata uang tertentu, waktu dari berbagai hari dalam satu tahun.

2. **Data Time Series**

   Data time series merupakan data yang terdiri atas satu objek dan lebih dari atau sama dengan satu variabel yang diperoleh oleh individu ataupun kelompok dalam rentang waktu tertentu yang diukur secara berkala, seperti harian, mingguan, bulanan, ataupun tahunan. Terdapat beberapa contoh data time series, meliputi.

- Data covid pada rentang tahun 2020-2021 dengan objek covid dan variabel angka pertumbuhan, angka kematian.
- Data cuaca harian di kota Surabaya dengan objek cuaca di kota Surabaya dan variabel suhu, kelembapan, tekanan udara.
- Data perubahan kinerja sebuah aplikasi B di 2023-2024 dengan objek aplikasi B dan variabel waktu respon aplikasi, jumlah pengguna aktif.
- Data penjualan mobil pada *dealer* A pada rentang tahun 2021-2022 dengan objek *dealer* A dan variabel jumlah pembelian, harga pembelian.
- Data produksi pada di Jawa Barat pada tahun 2019-2020 dengan objek padi di Jawa Barat dan variabel luas panen, hasil panen, jenis varietas padi.
- Data pergerakan harga emas di pasar global pada tahun 2021-2023 dengan objek harga emas dan variabel nilai tukar mata uang, suku bunga, inflansi.
- Data tingkat pengangguran pada lima tahun terakhir di Indonesia dengan objek tingkat pengangguran dan variabel jumlah penduduk usia kerja, kebijakan pemerintah.

3. Data Pooled (Panel)

Data pooled atau yang sering dikenal dengan istilah data panel merupakan gabungan antara data cross section dan data time series. Data panel memiliki satu objek dengan lebih dari satu variabel yang diperoleh oleh individu ataupun kelompok dalam beberapa periode waktu tertentu. Terdapat beberapa contoh data panel, meliputi.

- Data nasabah dan jumlah tabungan yang diamati selama tiga tahun terakhir dengan objek nasabah bank dan variabel jumlah tabungan, tanggal waktu menabung.
- Data kinerja akademis dari sekolah yang diamati setiap semester selama tujuh tahun terakhir dengan objek sekolah dan variabel jam kerja, jumlah capaian.
- Data kemiskinan di Indonesia yang diamati setiap tahun sejak tahun 2010 dengan objek wilayah Indonesia dan variabel pendapatan, tingkat kemiskinan.
- Data penjualan *online* sebuah toko dalam satu tahun dengan objek toko *online* dan variabel jumlah pengunjung *website*, jumlah produk terjual, pendapatan penjualan.

- Data keuntungan bulanan dari perusahaan tekstil yang diamati selama dua tahun terakhir dengan objek perusahaan tekstil dan variabel jumlah produksi, keuntungan.
- Data pendapatan dan konsumsi rumah tangga di Surabaya yang diamati selama lima tahun terakhir dengan objek rumah tangga di Surabaya dan variabel pendapatan, konsumsi.
- Data kesehatan pasien rumah sakit yang diamati setiap bulan dalam empat tahun terakhir dengan objek data kesehatan dan variabel berbagai faktor di dalam data kesehatan.

2. Jelaskan perbedaan antara data cross section dan data panel.

Jawaban dan Pembahasan :

Data cross section dan data panel memiliki beberapa perbedaan jika dilihat berdasarkan cara pengumpulan dan struktur yang dimilikinya, di antaranya yakni.

A. Pengumpulan Data

Pada data cross section dilakukan pengumpulan data dari satu dimensi (entitas) pada waktu tertentu dan mencari perbedaan antar individu dalam entitas yang diamati. Sedangkan, pada data panel dilakukan pengumpulan data dari dua dimensi (entitas dan waktu) dalam beberapa periode waktu dan melakukan analisis perubahan pada objek individu (entitas) di antar waktu.

B. Dimensi Waktu

Pada data cross section dilakukan pengumpulan data pada waktu tertentu seperti di satu tahun. Sedangkan, pada data panel pengumpulan data dilakukan pada beberapa waktu tertentu yang berbeda seperti di tiga tahun yang terdiri atas tahun pertama, tahun kedua, dan tahun ketiga.

C. Struktur Data

Pada data cross section struktur data terdiri atas setiap baris yang mewakili satu unit objek (entitas) di waktu tertentu. Sedangkan, pada data panel struktur data terdiri atas setiap baris yang mewakili satu unit objek (entitas) dengan beberapa waktu tertentu yang berbeda.

D. Analisis Longitudinal (Perubahan Perkembangan Fenomona dari Waktu ke Waktu)

Pada data cross section dikhususkan untuk analisis fenomena pada satu titik waktu tertentu yang mudah untuk dikumpulkan dan dianalisis. Sedangkan, pada data panel memungkinkan analisis longitudional dengan perubahan dari waktu ke waktu untuk melihat perubahan waktu ke waktu dan identifikasi adanya efek yang diperoleh.

E. Informasi yang di Peroleh

Pada data cross section tidak memberikan informasi mengenai *tren* atau perubahan dari waktu ke waktu dan tidak dapat mengidentifikasi efek yang diperoleh. Sedangkan, pada data panel memberikan informasi mengenai *tren* atau perubahan dari waktu ke waktu, namun lebih sulit untuk dikumpulkan dan dianalisis.

3. Carilah data panel yang terdiri dari minimal 3 variabel dengan periode pencacatan waktu minimal 24 periode. Dari data tersebut, lakukan analisis dekstriptif dan analisis secara visual untuk memperoleh karakteristik dari data tersebut. (Lampirkan sumber dan data yang digunakan).

Jawaban dan Pembahasan :

Pada penugasan ini diperguankan dataset “Walmart Sales Dataset Of 45 Stores”. Walmart merupakan perusahaan ritel multinasional di Amerika yang berdiri sejak tahun 1962. Dalam dataset ini terdapat 8 kolom dan 6435 baris dengan periode pencatatan dimulai pada tahun 2010 sampai dengan tahun 2012 yang dapat diakses dari kaggle menggunakan *link* berikut https://www.kaggle.com/datasets/varsharam/walmart-sales-dataset-of-45stores.

Namun, untuk melakukan analisis deskriptif dan analisis visual untuk tugas ini, terfokus pada analisis di *store* kode 1 menggunakan kolom *store*, *date*, *weekly_sales*, *holiday_flag*, dan *temperature* dengan periode pencatatan dimulai pada tahun 2010 sampai dengan 2012. Mengenai kolom yang dipergunakan untuk analisis dapat dijelaskan lebih lanjut, sebagai berikut.

Nama Kolom	Keterangan
Date (Tanggal Periode)	Berisi tanggal periode pencatatan penjualan yang dilakukan per-minggu mulai tahun 2010-2012.
Store (Toko)	Berisi kode *store* yang dimulai dari 1-45. Namun, untuk penugasan ini digunakan hanya *store* 1.
Weekly Sales (Penjualan Mingguan)	Berisi pendapatan mingguan dari *store* 1.
Holiday Flag (Hari Libur)	Berisi iya atau tidaknya hari libur untuk memberikan informasi tambahan mengenai bagaimana penjualan berfluktuasi selama hari libur.
Temperature (Suhu)	Berisi nilai suhu untuk memberikan informasi tambahan mengenai bagaimana penjualan berfluktuasi selama terjadinya penurunan atau peningkatan suhu.

Setelah mengetahui nama-nama kolom, selanjutnya akan dilakukan proses *pre-processing data*, analisis deskriptif dan analisis visual untuk mengetahui karakteristik data yang akan dijelaskan lebih lanjut sebagai berikut.

A. Pre-Processing Data

Pre-processing data dilakukan untuk *load* dataset awal, melakukan pengubahan nama kolom agar lebih mudah dipahami, dan melakukan pemilihan pengambilan data yang akan dianlisis lebih lanjut.

1. Melakukan Load Dataset Awal

Dilakukan proses load dataset "Walmart Sales Dataset Of 45 Stores" yang tersimpan dalam *file* "walmart-sales-dataset-of-45stores.csv" dengan format csv. Dimana, data tersebut disimpan dalam variabel data dengan kolom *Store, Date, Weekly_Sales, Holiday_Flag, Temperature, Fuel_Price, CPI,* dan *Unemployment*.

1. Melakukan Load Dataset

```
import pandas as pd

data = pd.read_csv("walmart-sales-dataset-of-45stores.csv")
data
```

	Store	Date	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment
0	1	05-02-2010	1643690.90	0	42.31	2.572	211.096358	8.106
1	1	12-02-2010	1641957.44	1	38.51	2.548	211.242170	8.106
2	1	19-02-2010	1611968.17	0	39.93	2.514	211.289143	8.106
3	1	26-02-2010	1409727.59	0	46.63	2.561	211.319643	8.106
4	1	05-03-2010	1554806.68	0	46.50	2.625	211.350143	8.106
...	...	...	...	...	...	...	...	...
6430	45	28-09-2012	713173.95	0	64.88	3.997	192.013558	8.684
6431	45	05-10-2012	733455.07	0	64.89	3.985	192.170412	8.667
6432	45	12-10-2012	734464.36	0	54.47	4.000	192.327265	8.667
6433	45	19-10-2012	718125.53	0	56.47	3.969	192.330854	8.667
6434	45	26-10-2012	760281.43	0	58.85	3.882	192.308899	8.667

6435 rows × 8 columns

tampilan dataset awal Walmart

2. Melakukan Pengubahan Nama (Rename) Kolom Dan Mengurutkan (Sorting) Tanggal Periode Pencatatan Awal Hingga Akhir

2. Melakukan Rename Nama Kolom Dan Sorting Tanggal Periode Pencatatan Awal Hingga Akhir

```
#merename kolom agar lebih mudah dipahami
data.rename(columns={"Store": "Toko", "Date": "Tanggal Periode", "Weekly_Sales": "Penjualan Mingguan",
                     "Holiday_Flag": "Hari Libur", "Temperature": "Suhu", "Fuel_Price": "Harga Bahan Bakar",
                     "CPI": "CPI", "Unemployment": "Tingkat Pengangguran Toko"}, inplace=True)

#mengubah format kolom 'Tanggal Periode' menjadi tipe datetime dan mengurutkannya
data["Tanggal Periode"] = pd.to_datetime(data["Tanggal Periode"], format="%d-%m-%Y")
data["Toko"] = pd.to_numeric(data["Toko"], errors="coerce")
data.sort_values(by=["Tanggal Periode", "Toko"], inplace=True, ignore_index=True)

data
```

	Toko	Tanggal Periode	Penjualan Mingguan	Hari Libur	Suhu	Harga Bahan Bakar	CPI	Tingkat Pengangguran Toko
0	1	2010-02-05	1643690.90	0	42.31	2.572	211.096358	8.106
1	2	2010-02-05	2136989.46	0	40.19	2.572	210.752605	8.324
2	3	2010-02-05	461622.22	0	45.71	2.572	214.424881	7.368
3	4	2010-02-05	2135143.87	0	43.76	2.598	126.442065	8.623
4	5	2010-02-05	317173.10	0	39.70	2.572	211.653972	6.566
...	...	...	...	...	...	...	...	...
6430	41	2012-10-26	1316542.59	0	41.80	3.686	199.219532	6.195
6431	42	2012-10-26	514756.08	0	70.50	4.301	131.193097	6.943
6432	43	2012-10-26	587603.55	0	69.17	3.506	214.741539	8.839
6433	44	2012-10-26	361067.07	0	46.97	3.755	131.193097	5.217
6434	45	2012-10-26	760281.43	0	58.85	3.882	192.308899	8.667

6435 rows × 8 columns

tampilan pengubahan nama kolom dan sorting tanggal

Dilakukan proses pengubahan nama kolom menjadi nama yang lebih mudah dipahami secara urut mengguanakn fungsi *rename* dengan *key:value*, meliputi Toko, Tanggal Periode, Penjualan Mingguan, Hari Libur, Suhu, Harga

Bahan Bakar, CPI, dan Tingkat Pengangguran Toko. Kemudian dilakukan pengubahan tipe data di “Tanggal Periode” menjadi *datetime* agar dapat di*sorting* dengan tanggal periode pencatatan mulai tahun 2010 sampai dengan 2012.

3. Melakukan Pemilihan Pengambilan Data Yang Akan Di Analisis Lebih Lanjut

3. Melakukan Pemilihan Pengambilan Data Yang Akan Di Analisis Lebih Lanjut

```
import numpy as np

#filter toko 1 untuk analisis lebih lanjut
toko_terpilih = [1]
data_new = data[data["Toko"].isin(toko_terpilih)]

#filter kolom yang ingin dianalisis
kolom_terpilih = ["Tanggal Periode", "Toko", "Penjualan Mingguan", "Hari Libur", "Suhu"]
data_new = data_new[kolom_terpilih]

data_new
```

	Tanggal Periode	Toko	Penjualan Mingguan	Hari Libur	Suhu
0	2010-02-05	1	1643690.90	0	42.31
45	2010-02-12	1	1641957.44	1	38.51
90	2010-02-19	1	1611968.17	0	39.93
135	2010-02-26	1	1409727.59	0	46.63
180	2010-03-05	1	1554806.68	0	46.50
...	...	...	...	...	...
6210	2012-09-28	1	1437059.26	0	76.08
6255	2012-10-05	1	1670785.97	0	68.55
6300	2012-10-12	1	1573072.81	0	62.99
6345	2012-10-19	1	1508068.77	0	67.97
6390	2012-10-26	1	1493659.74	0	69.16

143 rows × 5 columns

tampilan pemilihan data yang akan di analisis lebih lanjut

```
#mengecek nilai yang hilang
missing_values = data_new.isnull().sum()
print("Jumlah Missing Value Setiap Kolom : ")
(missing_values)
```

```
Jumlah Missing Value Setiap Kolom :

Tanggal Periode       0
Toko                  0
Penjualan Mingguan    0
Hari Libur            0
Suhu                  0
dtype: int64
```

tampilan pengecekan missing value

```
data_new.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 143 entries, 0 to 6390
Data columns (total 5 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Tanggal Periode     143 non-null    datetime64[ns]
 1   Toko                143 non-null    int64
 2   Penjualan Mingguan  143 non-null    float64
 3   Hari Libur          143 non-null    int64
 4   Suhu                143 non-null    float64
dtypes: datetime64[ns](1), float64(2), int64(2)
memory usage: 6.7 KB
```

tampilan tipe data yang akan dianalisis

Dilakukan proses pemilihan data yang akan dianalisis lebih lanjut, yakni toko dengan kode 1 dan memilih nama kolom, meliputi Tanggal Periode, Toko,

Penjualan Mingguan, Hari Libur, dan Suhu. Dimana jumlah data yang akan dianalisis lebih lanjut tersimpan dalam variabel “data_new” dengan 143 baris dan 5 kolom.

B. Analisis Deskriptif Untuk Memperoleh Karakteristik Data

Analisis deskriptif dilakukan untuk memberikan pemahaman mengenai karakteristik dasar dari suatu dataset atau kumpulan data. Dilakukan beberapa analisis deskriptif, meliputi perhitungan statistik dasar baik secara kumulatif maupun disetiap tahun, perhitungan skewness dan kurtosis, serta pengecekan keterkaitan antar variabel.

1. Perhitungan Statistik Deskriptif

1.1 Di Rentang Tahun 2010-2012 Secara Kumulatif

1. Perhitungan Statistik Deskriptif

```
#melakukan perhitungan mean, median, modus, standar deviasi, kuartil, minimum, maksimum
#disemua kolom, kecuali "Tanggal Periode" karena format datetime ditahun 2010 dan 2011 dan Store karena hanya 1

import numpy as np

#define function untuk perhitungan menggunakan library numpy
def print_stats(label, data):
    print(label)

    #perhitungan
    mean = np.mean(data)
    median = np.median(data)
    counts = np.bincount(data.astype(int))
    mode = np.argmax(counts)
    std = np.std(data)
    kuartil = np.percentile(data, [25, 50, 75])
    maksimum = np.max(data)
    minimum = np.min(data)

    #penampilan hasil
    print("Mean (Rata-Rata): ", mean)
    print("Median (Nilai Tengah) : ", median)
    print("Modus (Nilai Sering Muncul) : ", mode)
    print("Standar Deviasi: ", std)
    print("Kuartil : ", kuartil)
    print("Maksimum : ", maksimum)
    print("Minimum : ", minimum)
    print()

#penampilan ouput tiap kolom
print("'Statistik Dasar Di Masing-Masing Kolom Pada Tahun 2010-2012'")
columns = ["Penjualan Mingguan", "Hari Libur", "Suhu"]
for column in columns:
    print_stats(f"Statistik Deskriptif Untuk Kolom '{column}' : ", data_new[column].values)
```

```
'Statistik Dasar Di Masing-Masing Kolom Pada Tahun 2010-2012'
Statistik Deskriptif Untuk Kolom 'Penjualan Mingguan' :
Mean (Rata-Rata):  1555264.3975524476
Median (Nilai Tengah) :  1534849.64
Modus (Nilai Sering Muncul) :  1316899
Standar Deviasi:  155434.42363856622
Kuartil :  [1458104.69 1534849.64 1614892.03]
Maksimum :  2387950.2
Minimum :  1316899.31

Statistik Deskriptif Untuk Kolom 'Hari Libur' :
Mean (Rata-Rata):  0.06993006993006994
Median (Nilai Tengah) :  0.0
Modus (Nilai Sering Muncul) :  0
Standar Deviasi:  0.2550291262770695
Kuartil :  [0. 0. 0.]
Maksimum :  1
Minimum :  0

Statistik Deskriptif Untuk Kolom 'Suhu' :
Mean (Rata-Rata):  68.30678321678322
Median (Nilai Tengah) :  69.64
Modus (Nilai Sering Muncul) :  80
Standar Deviasi:  14.200572115008763
Kuartil :  [58.265 69.64  80.485]
Maksimum :  91.65
Minimum :  35.4
```

tampilan statistika deskriptif setiap kolom secara kumulatif tahun 2010-2012

Dilakukan perhitungan statistik deskriptif disetiap kolom dalam rentang tahun secara kumulatif, yakni mulai tahun 2010 sampai dengan 2012. Bertujuan untuk memberikan deskripsi data dan pemahaman distribusi. Dengan menggunakan fungsi bawaan *numpy* dengan np.*nilai statistik yang ingin dicari*. Dilakukan perhitungan statistik dasar meliputi *mean, median, modus, standar deviasi,* kuartil, maksimum, dan minimum disetiap kolom kecuali kolom "Tanggal Periode".

Dikarenakan memiliki tipe *datetime* yang kurang representatif jika dihitung statistik deskriptifnya dan kolom "Toko" karena hanya ada 1. Diperoleh hasil bahwa kolom "Penjualan Mingguan" memiliki variasi yang cukup signifikan, dengan *mean* sekitar 1.555.264 dan *standar deviasi* sekitar 155.434 yang menunjukkan adanya penjualan berkisar diantara nilai yang cukup beragam.

Kolom "Hari Libur" menunjukkan bahwa tidak ada mayoritas adanya hari libur dengan nilai *mean* mendekati 0 dan sekitar 6,99% dari setiap data tanggal periode tersebut adalah hari libur.

1.2 Di Rentang Tahun 2010, 2011, Dan 2012

```
#melakukan perhitungan mean, median, modus, standar deviasi, kuartil, minimum, maksimum
#disemua kolom, kecuali "Tanggal Periode" karena format datetime di masing-masing tahun dan Store karena hanya 1

import numpy as np

#define function untuk perhitungan menggunakan library numpy
def print_stats(label, data):
    print(label)

    #perhitungan
    mean = np.mean(data)
    median = np.median(data)
    mode = np.argmax(np.bincount(data.astype(int)))
    std = np.std(data)
    kuartil = np.percentile(data, [25, 50, 75])
    maksimum = np.max(data)
    minimum = np.min(data)

    #penampilan hasil
    print("Mean (Rata-Rata): ", mean)
    print("Median (Nilai Tengah) : ", median)
    print("Modus (Nilai Sering Muncul) : ", mode)
    print("Standar Deviasi: ", std)
    print("Kuartil : ", kuartil)
    print("Maksimum : ", maksimum)
    print("Minimum : ", minimum)
    print()

#melakukan perhitungan statistik untuk tahun tertentu
def calculate_statistics(data, year):
    print(f"'Statistik Dasar Di Masing-Masing Kolom Pada Tahun {year}'")
    for column in ["Penjualan Mingguan", "Hari Libur", "Suhu"]:
        print_stats(f"Statistik Deskriptif Untuk Kolom '{column}' : ", data[column].values)

#filter data untuk setiap tahun 2010-2012 dan hitung statistiknya
for year in range(2010, 2013):
    data_year = data_new[data_new["Tanggal Periode"].dt.year == year]
    calculate_statistics(data_year, year)
```

```
'Statistik Dasar Di Masing-Masing Kolom Pada Tahun 2010'
Statistik Deskriptif Untuk Kolom 'Penjualan Mingguan' :
Mean (Rata-Rata):  1526642.3333333333
Median (Nilai Tengah) :  1494365.495
Modus (Nilai Sering Muncul) :  1345454
Standar Deviasi:  173414.20379141424
Kuartil :  [1429059.18  1494365.495 1552446.13 ]
Maksimum :  2387950.2
Minimum :  1345454.0

Statistik Deskriptif Untuk Kolom 'Hari Libur' :
Mean (Rata-Rata):  0.08333333333333333
Median (Nilai Tengah) :  0.0
Modus (Nilai Sering Muncul) :  0
Standar Deviasi:  0.2763853991962833
Kuartil :  [0. 0. 0.]
Maksimum :  1
Minimum :  0

Statistik Deskriptif Untuk Kolom 'Suhu' :
Mean (Rata-Rata):  67.4975
Median (Nilai Tengah) :  68.525
Modus (Nilai Sering Muncul) :  80
Standar Deviasi:  14.602948261110381
Kuartil :  [54.0175 68.525  80.9175]
Maksimum :  87.16
Minimum :  38.51
```

tampilan statistika deskriptif setiap kolom tahun 2010

```
'Statistik Dasar Di Masing-Masing Kolom Pada Tahun 2011'
Statistik Deskriptif Untuk Kolom 'Penjualan Mingguan' :
Mean (Rata-Rata):  1556190.7467307688
Median (Nilai Tengah) :  1537166.67
Modus (Nilai Sering Muncul) :  1316899
Standar Deviasi:  162604.905833849
Kuartil :  [1456380.2025 1537166.67   1608537.0225]
Maksimum :  2270188.99
Minimum :  1316899.31

Statistik Deskriptif Untuk Kolom 'Hari Libur' :
Mean (Rata-Rata):  0.07692307692307693
Median (Nilai Tengah) :  0.0
Modus (Nilai Sering Muncul) :  0
Standar Deviasi:  0.2664693550105965
Kuartil :  [0. 0. 0.]
Maksimum :  1
Minimum :  0

Statistik Deskriptif Untuk Kolom 'Suhu' :
Mean (Rata-Rata):  67.65826923076924
Median (Nilai Tengah) :  68.575
Modus (Nilai Sering Muncul) :  59
Standar Deviasi:  15.848241072161949
Kuartil :  [56.765  68.575  83.0325]
Maksimum :  91.65
Minimum :  35.4
```

tampilan statistika deskriptif setiap kolom tahun 2011

```
'Statistik Dasar Di Masing-Masing Kolom Pada Tahun 2012'
Statistik Deskriptif Untuk Kolom 'Penjualan Mingguan' :
Mean (Rata-Rata):  1586094.3725581397
Median (Nilai Tengah) :  1582083.4
Modus (Nilai Sering Muncul) :  1319325
Standar Deviasi:  113736.1198451455
Kuartil :  [1509568.42 1582083.4  1655685.98]
Maksimum :  1899676.88
Minimum :  1319325.59

Statistik Deskriptif Untuk Kolom 'Hari Libur' :
Mean (Rata-Rata):  0.046511627906976744
Median (Nilai Tengah) :  0.0
Modus (Nilai Sering Muncul) :  0
Standar Deviasi:  0.21059035204970736
Kuartil :  [0. 0. 0.]
Maksimum :  1
Minimum :  0

Statistik Deskriptif Untuk Kolom 'Suhu' :
Mean (Rata-Rata):  69.99441860465114
Median (Nilai Tengah) :  70.33
Modus (Nilai Sering Muncul) :  77
Standar Deviasi:  11.169502144247591
Kuartil :  [63.865 70.33  78.345]
Maksimum :  86.11
Minimum :  45.32
```

tampilan statistika deskriptif setiap kolom tahun 2012

Dilakukan perhitungan statistik deskriptif disetiap kolom dalam rentang tahun dimulai pada 2010, 2011, dan 2012. Dengan tujuan memberikan deskripsi data dan pemahaman distribusi. Dengan menggunakan fungsi bawaan *numpy* dengan np.*nilai statistik yang ingin dicari*. Dilakukan perhitungan yang sama seperti kumulatif namun disetiap tahun. Diperoleh hasil bahwa kolom "Penjualan Mingguan" memiliki variasi yang signifikan dalam nilai penjualan mingguan disetiap tahun dengan *mean* yang bervariasi.

Distribusi cenderung luas dengan *standar deviasi* yang tinggi. Kolom "Hari Libur" memiliki keberadaan hari libur yang relatif sedikit ditiap tahunnya dengan distribusi data menunjukkan mayoritas 0. Kolom "Suhu" memiliki variasi dengan nilai yang berbeda-beda disetiap tahunnya dengan distribusi yang cenderung bervariasi dan memiliki kuartil relatif serupa. Sehingga, menunjukkan adanya konsistensi dalam distribusi suhu.

```
#perhitungan menggunakan describe

data_describe = data_new.describe()
data_describe
```

	Tanggal Periode	Toko	Penjualan Mingguan	Hari Libur	Suhu
count	143	143.0	1.430000e+02	143.000000	143.000000
mean	2011-06-17 00:00:00	1.0	1.555264e+06	0.069930	68.306783
min	2010-02-05 00:00:00	1.0	1.316899e+06	0.000000	35.400000
25%	2010-10-11 12:00:00	1.0	1.458105e+06	0.000000	58.265000
50%	2011-06-17 00:00:00	1.0	1.534850e+06	0.000000	69.640000
75%	2012-02-20 12:00:00	1.0	1.614892e+06	0.000000	80.485000
max	2012-10-26 00:00:00	1.0	2.387950e+06	1.000000	91.650000
std	NaN	0.0	1.559808e+05	0.255926	14.250486

tampilan statistika deskriptif menggunaka describe

Atau pengecekan nilai statistika deskriptif dapat digunakan fungsi *describe()* yang akan memberikan ringkasan statistik, meliputi *count, mean, standar deviasi, percentil 25 50 75,* dan *maksimum*.

2. **Pengecekan Data Terdistribusi Dengan Nilai Rata-Rata Menggunakan Skewness Dan Pengecekan Data Mengandung Banyak Nilai Ekstrem Atau Tidak Menggunakan Kurtosis**

2. Pengecekan Data Terdistribusi Dengan Nilai Rata-Rata Menggunakan Skewness Dan Pengecekan Data Mengandung Banyak Nilai Ekstrem Atau Tidak Menggunakan Kurtosis

```
#perhitungan dan visualisasi skewness dan kurtosis
#untuk mengecek apakah data terdistribusi dengan nilai rata-rata dan apakah data mengandung banyak nilai ekstrem

#menghitung skewness dan kurtosis dari nilai yang bisa diperhitungkan
numerical_columns = data_new.select_dtypes(include=["number"]).columns
skewness = data_new[numerical_columns].skew()
kurtosis = data_new[numerical_columns].kurtosis()

print("'Skewness'")
print(skewness)
print("\n'Kurtosis'")
print(kurtosis)
```

```
'Skewness'
Toko                0.000000
Penjualan Mingguan  2.180276
Hari Libur          3.408571
Suhu               -0.406924
dtype: float64

'Kurtosis'
Toko                0.000000
Penjualan Mingguan  8.151506
Hari Libur          9.754602
Suhu               -0.873924
dtype: float64
```

tampilan skewness dan kurtosis

Dilakukan perhitungan skewness untuk pengecekan apakah data banyak mengandung nilai ekstrem atau tidak dan kurtosis untuk pengecekan apakah distribusi data condong ke satu sisi (skewness positif) atau sisi lainnya (skewness negatif) dari nilai rata-rata. Perhitungan skewness dan kurtosis ini dilakukan pada semua kolom yang memiliki tipe data *number* atau numerik, meskipun sebenarnya pada kolom “Toko” tidak perlu diperhitungkan karena hanya terdapat 1 toko.

Dari perhitungan skewness, kolom “Penjualan Mingguan” distribusi cenderung miring ke kanan (2.18), kolom “Hari Libur” distribusi cenderung miring ke kanan (3.41), sementara kolom “Suhu” distribusi cenderung simetris (-0.41). Dari kurtosis, penjualan mingguan sebesar (8.15) menunjukkan adanya banyak nilai ekstrem didalam distribusi. Kolom “Hari Libur” sebesar (9.75) menunjukkan adanya banyak nilai ekstrem didalam distribusi data.

Sementara kolom “Suhu” sebesar (-0.87) menunjukkan distribusi yang lebih datar daripada distribusi normal.

3. Pengecekan Keterkaitan Antar Variabel (2 Variabel Dan 3 Variabel)

Dilakukan pengecekan keterkaitan antar dua maupun tiga variabel untuk *tambahan informasi* menngenai bagaimana suatu variabel mempengaruhi variabel yang lain. Fokusnya dilakukan untuk mengetahui bagaimana penjualan mingguan, bulanan, dan tahunan didalam tanggal periode dan mengetahui hubungan antar kolom atau variabel menggunakan heatmap.

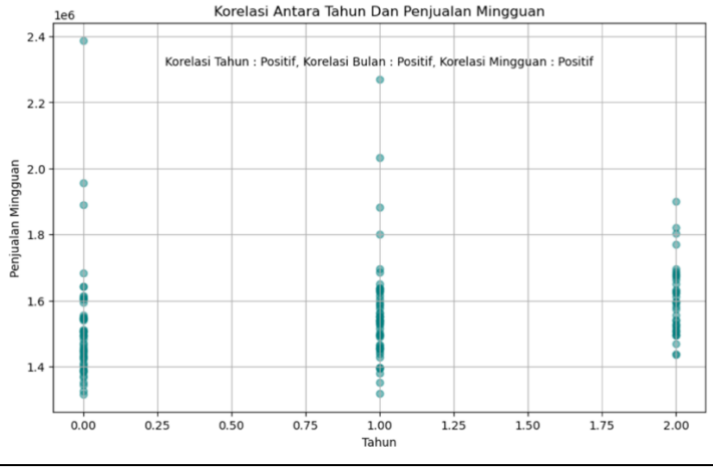
3.1 Analisis Korelasi 2 Variabel Dengan Tanggal Periode Terhadap Penjualan Mingguan Dalam Rentang Mingguan, Bulanan, Dan Tahunan

3. Pengecekan Keterkaitan Antar Variabel (2 Variabel Dan 3 Variabel)

Analisis Korelasi 2 Variabel Dengan Tanggal Periode Terhadap Penjualan Mingguan Dalam Rentang Mingguan, Bulanan, Dan Tahunan

```
#analisis korelasi tanggal periode dengan penjualan mingguan
import numpy as np
import matplotlib.pyplot as plt

#memecah Tanggal Periode menjadi fitur tambahan
start_date = data_new["Tanggal Periode"].min()
data_new["Year"] = (data_new["Tanggal Periode"] - start_date).dt.days // 365
data_new["Month"] = (data_new["Tanggal Periode"] - start_date).dt.days // 30
data_new["Week"] = (data_new["Tanggal Periode"] - start_date).dt.days // 7

#menghitung korelasi antara fitur tambahan dan penjualan mingguan
correlation_year = data_new["Year"].corr(data_new["Penjualan Mingguan"])
correlation_month = data_new["Month"].corr(data_new["Penjualan Mingguan"])
correlation_week = data_new["Week"].corr(data_new["Penjualan Mingguan"])

#hasil korelasi
print(f"Korelasi Antara Tahun Dan Penjualan Mingguan: {correlation_year}")
print(f"Korelasi Antara Bulan Dan Penjualan Mingguan: {correlation_month}")
print(f"Korelasi Antara Minggu Dalam Tahun Dan Penjualan Mingguan: {correlation_week}")

#hasil perhitungan korelasi
correlation_year_desc = "Positif" if correlation_year > 0 else "Negatif" if correlation_year < 0 else
"Tidak Berkorelasi"
correlation_month_desc = "Positif" if correlation_month > 0 else "Negatif" if correlation_month < 0 else
"Tidak Berkorelasi"
correlation_week_desc = "Positif" if correlation_week > 0 else "Negatif" if correlation_week < 0 else
"Tidak Berkorelasi"

#visualisasi scatter plot dengan garis lurus korelasi
plt.figure(figsize=(10, 6))
plt.scatter(data_new["Year"], data_new["Penjualan Mingguan"], color="teal", alpha=0.5)
plt.title("Korelasi Antara Tahun Dan Penjualan Mingguan")
plt.xlabel("Tahun")
plt.ylabel("Penjualan Mingguan")
plt.text(0.5, 0.9, f"Korelasi Tahun : {correlation_year_desc}, Korelasi Bulan : {correlation_month_desc},
        Korelasi Mingguan : {correlation_week_desc}",
        horizontalalignment="center", verticalalignment="center", transform=plt.gca().transAxes)
plt.grid(True)

plt.show()
```

```
Korelasi Antara Tahun Dan Penjualan Mingguan: 0.21650157846810206
Korelasi Antara Bulan Dan Penjualan Mingguan: 0.215425046303892
Korelasi Antara Minggu Dalam Tahun Dan Penjualan Mingguan: 0.21453921574872156
```



tampilan korelasi tanggal periode terhadap mingguan, bulanan, tahunan

Dilakukan analisis keterkaitan kolom "Tanggal Periode" dengan "Penjualan Mingguan" dilakukan dengan membagi tanggal periode menjadi fitur tambahan seperti tahun, bulan, dan minggu dalam tahun. Kemudian, korelasi antara fitur-fitur tambahan tersebut dan penjualan mingguan dihitung. Korelasi ini memberikan gambaran tentang seberapa erat hubungan antara waktu (tahun, bulan, minggu) terhadap penjualan mingguan menggunakan fungsi *corr()*.

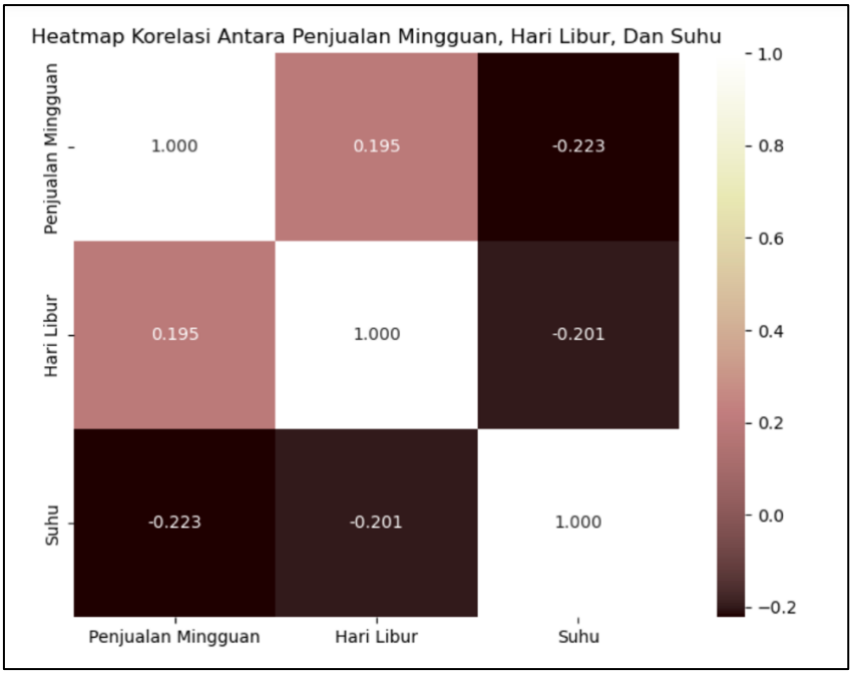
Diperoleh hasil, terdapat hubungan positif yang lemah antara waktu pencatatan, diantaranya tahun dan penjualan mingguan sebesar 0,217, bulan dan penjualan mingguan sebesar 0,215, serta minggu dan penjualan mingguan sebesar 0,21. Sehingga, terdapat kecendrungan bahwa penjualan mingguan meningkat seiring dengan berjalannya waktu. Namun, terdapat pula faktor-faktor lain yang memengaruhi.

3.2 Analisis Korelasi 3 Variabel Di Setiap Kolom Data

Analisis Korelasi 3 Variabel Di Setiap Kolom Data

```
#korelasi penjualan mingguan, hari libur, suhu
#dengan visualisasi heatmap
import seaborn as sns
import matplotlib.pyplot as plt

# Menghitung matriks korelasi
correlation_matrix = data_new[["Penjualan Mingguan", "Hari Libur", "Suhu"]].corr()
print("Matriks Korelasi : ", correlation_matrix)

#visualisasi heatmap
plt.figure(figsize=(8, 6))
sns.heatmap(correlation_matrix, annot=True, cmap="pink", fmt=".3f")
plt.title("Heatmap Korelasi Antara Penjualan Mingguan, Hari Libur, Dan Suhu")

plt.show()
```

```
Matriks Korelasi :                    Penjualan Mingguan  Hari Libur      Suhu
Penjualan Mingguan            1.000000    0.194905 -0.222701
Hari Libur                    0.194905    1.000000 -0.200543
Suhu                         -0.222701   -0.200543  1.000000
```



tampilan korelasi tiga variabel dari setiap kolom

Dilakukan analisis keterkaitan antar kolom dilakukan agar mengetahui hubungan keterkaitan antar kolom. Dengan melakukan perhitungan matriks korelasi antara ketiga variabel atau kolom “Penjualan Mingguan”, “Hari Libur”,

dan “Suhu”. Kemudian dilakukan visualisasi heatmap dengan semakin dekat ke angka 1 semakin tinggi korelasinya dan semakin dekat ke angka 0 semakin rendah korelasinya. Dilakukan contoh pembacaan bahwa “Penjualan Mingguan” dan “Hari Libur” memiliki korelasi sebesar 0,195.

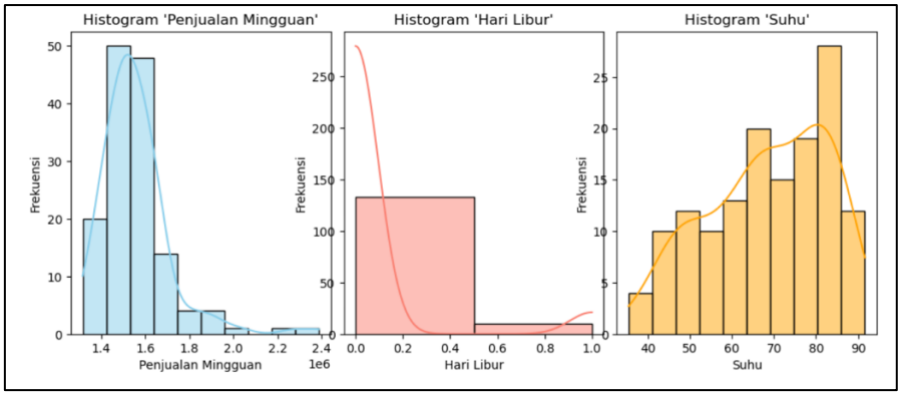
C. Analisis Visual Untuk Memperoleh Karakteristik Data

Analisis visual dilakukan untuk memperoleh karakteristik data dengan cara memvisualisasikan data secara grafis atau diagramatik. Bertujuan untuk memahami pola, hubungan, dan distribusi data. Dilakukan beberapa analsisi visual, meliputi pengecekan data berdistribusi normal menggunakan histogram, pengecekan proporsi data binner, pengecekan nilai letak kuartil, dan tren visualisasi penjualan mingguan.

1. Pengecekan Data Berdistribusi Normal Atau Tidak Menggunakan Histogram

1. Pengecekan Data Berdistribusi Normal Atau Tidak Menggunakan Histogram

```
#visualisasi histogram untuk melakukan pengecekan apakah suatu data berdistribusi normal atau tidak
#pada semua kolom kecuali "Tanggal Periode" dan "Toko" karena memiliki tipe datetime dan hanya 1

import matplotlib.pyplot as plt
import seaborn as sns

#plot histogram semua kolom,
plt.figure(figsize=(10, 8))

#penjualan mingguan
plt.subplot(2, 3, 1)
sns.histplot(data_new["Penjualan Mingguan"], bins=10, color="skyblue", kde=True)
plt.title("Histogram 'Penjualan Mingguan'")
plt.xlabel("Penjualan Mingguan")
plt.ylabel("Frekuensi")
plt.tight_layout()

#hari libur
plt.subplot(2, 3, 2)
sns.histplot(data_new["Hari Libur"], bins=2, color="salmon", kde=True)
plt.title("Histogram 'Hari Libur'")
plt.xlabel("Hari Libur")
plt.ylabel("Frekuensi")

#suhu
plt.subplot(2, 3, 3)
sns.histplot(data_new["Suhu"], bins=10, color="orange", kde=True)
plt.title("Histogram 'Suhu'")
plt.xlabel("Suhu")
plt.ylabel("Frekuensi")


plt.show()
```



tampilan visualisasi histogram untuk pengecekan distribusi setiap data

Dilakukan analisis visualisasi histogram untuk mengetahui apakah data dalam kolom “Penjualan Mingguan”, “Hari Libur”, dan “Suhu” memiliki distribusi normal atau tidak. Dengan menggunakan *library* matplotlib dan seaborn untuk visualiasi dan dilakukan pengambilan nilai dari setiap kolom yang ditentukan. Diperoleh hasil, bahwa sumbu x menunjukkan rentang nilai dan y menunjukkan frekuensi.

Dari hasil visualisasi tersebut, dapat diperlihatkan bahnya hanya kolom “Suhu” dan “Penjualan Mingguan” yang hampir memiliki visualisasi mirip dengan distribusi normal. Dengan catatan jika di visualisasikan distribusi normal akan berbentuk lonceng simetris dengan puncak ditengah dan sisi yang landai kearah sumbu x. Kemudian, untuk membuktikan apakah data tersebut berdistribusi normal dapat dilakukan pengecekan menggunakan uji *shapiro-wilk* sebagai *informasi tambahan.*

```
#pengecekan apakah berdistribusi normal
from scipy.stats import shapiro

stat_penjualan, p_penjualan = shapiro(data_new["Penjualan Mingguan"])
alpha = 0.05

#untuk penjualan mingguan
print("'Uji Normalitas Untuk Data Penjualan Mingguan'")
print("Statistik Uji :", stat_penjualan)
print("P-Value:", p_penjualan)
if p_penjualan > alpha:
    print("Data Berdistribusi Normal")
else:
    print("Data Tidak Berdistribusi Normal")


#untuk data suhu
stat_suhu, p_suhu = shapiro(data_new["Suhu"])
print("'Uji Normalitas Untuk Data Suhu'")
print("Statistik Uji : ", stat_suhu)
print("p-Value:", p_suhu)
if p_penjualan > alpha:
    print("Data Berdistribusi Normal")
else:
    print("Data Tidak Berdistribusi Normal")
```

```
'Uji Normalitas Untuk Data Penjualan Mingguan'
Statistik Uji : 0.8375532627105713
P-Value: 2.795954266721079e-11
Data Tidak Berdistribusi Normal
'Uji Normalitas Untuk Data Suhu'
Statistik Uji :  0.9539842009544373
p-Value: 0.00010702353756641969
Data Tidak Berdistribusi Normal
```

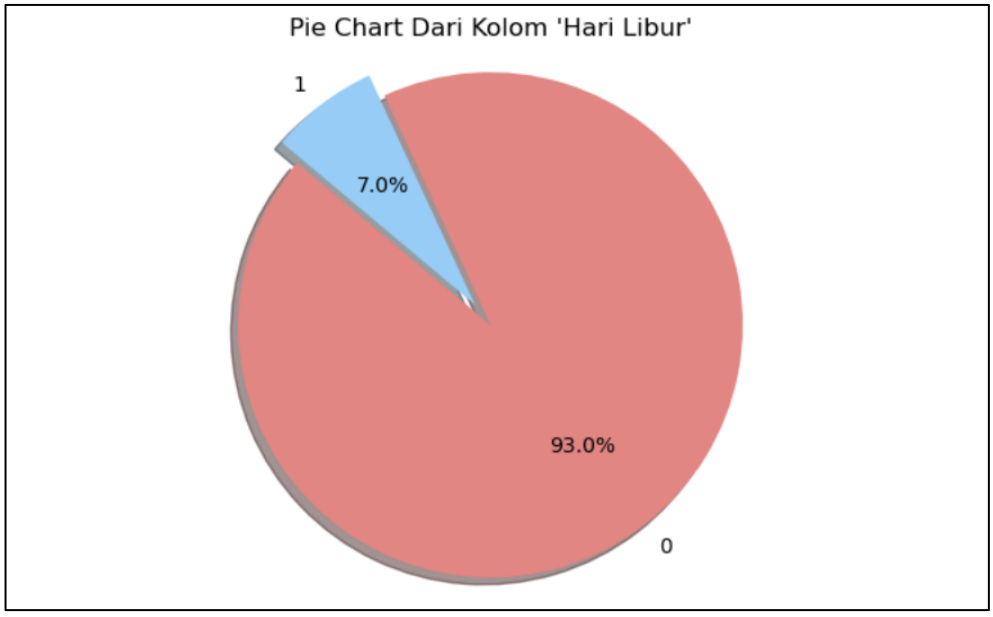
tampilan uji untuk distribusi normal atau tidak

Diperoleh hasil bahwa kedua data dalam kolom “Penjualan Mingguan” dan “Suhu” tidak berdistribusi normal. Dan, pada kolom “Hari Libur” lebih baik divisualisasikan menggunakan bar chart karena bersifat biner.

2. Pengecekan Proporsi Data Hari Libur (Karena Biner) Di Gunakan Pie Chart

2. Pengecekan Proporsi Data Hari Libur (Karena Biner) Di Gunakan Pie Chart

```
#visualisasi bar chart untuk mengetahui proporsi hari libur yang memiliki nilai biner 0 dan 1
import matplotlib.pyplot as plt

#jumlah nilai 0 dan 1 dalam kolom "Hari Libur"
count_0 = (data_new["Hari Libur"] == 0).sum()
count_1 = (data_new["Hari Libur"] == 1).sum()

#label dan nilai untuk pie chart
labels = ["0", "1"]
sizes = [count_0, count_1]
colors = ["lightcoral", "lightskyblue"]
explode = (0, 0.1)  #menekankan potongan '1'


plt.pie(sizes, explode=explode, labels=labels, colors=colors, autopct='%1.1f%%', shadow=True, startangle=140)
plt.title("Pie Chart Dari Kolom 'Hari Libur'")
plt.axis("equal")

plt.show()
```



tampilan visualisasi pie chart untuk proporsi angka biner di kolom "Hari Libur"

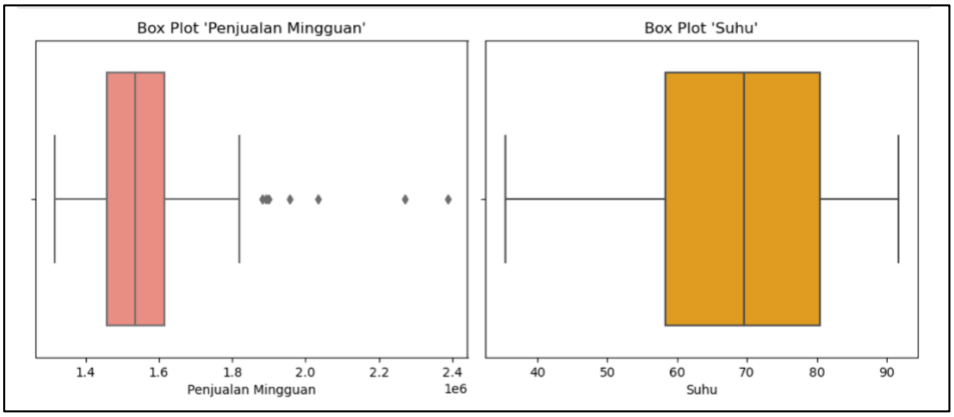
Dilakukan visualisasi bar chart untuk mengetahui proporsi banyaknya hari libur yang dilambangkan dengan angka 1 dan tidak hari libur yang dilambangkan dengan angka 0. Dari kolom "Hari Libur". Hal ini dilakukan bar chart coock digunakan untuk mengetahui karakteristik banyaknya proporsi data biner. Dengan melakukan perhitungan angka 0 dan 1 kemudian direpresentasikan dalam persen.

Diperoleh bahwa selama periode pencatatan mingguan penjualan, terdapat sebesar 7% hari libur dan 93% bukan hari libur.

3. Pengecekan Nilai Letak Kuartil Menggunakan Diagram Kotak (Box Plot)

3. Pengecekan Nilai Letak Kuartil Menggunakan Box Plot

```
#nilai letak kuartil hanya pada kolom "Penjualan Mingguan" dan "Suhu"
#karena kolom "Tanggal Periode" memiliki tipe data datetime dan "Hari Libur" memiliki tipe data biner (0,1)

import matplotlib.pyplot as plt
import seaborn as sns

#fungsi untuk menghitung batas atas dan batas bawah berdasarkan IQR
def calculate_bounds(data):
    Q1 = data.quantile(0.25)
    Q3 = data.quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    return lower_bound, upper_bound

#tata letak subplot
plt.figure(figsize=(10, 8))

#box plot untuk Penjualan Mingguan
plt.subplot(2, 2, 1)
sns.boxplot(x=data_new["Penjualan Mingguan"], color="salmon")
plt.title("Box Plot 'Penjualan Mingguan'")
plt.xlabel("Penjualan Mingguan")

#box plot untuk Suhu
plt.subplot(2, 2, 2)
sns.boxplot(x=data_new["Suhu"], color="orange")
plt.title("Box Plot 'Suhu'")
plt.xlabel("Suhu")

plt.tight_layout()
plt.show()
```



tampilan visualisasi letak kuartil dan outlier dari kolom "Penjualan Mingguan" dan "Suhu"

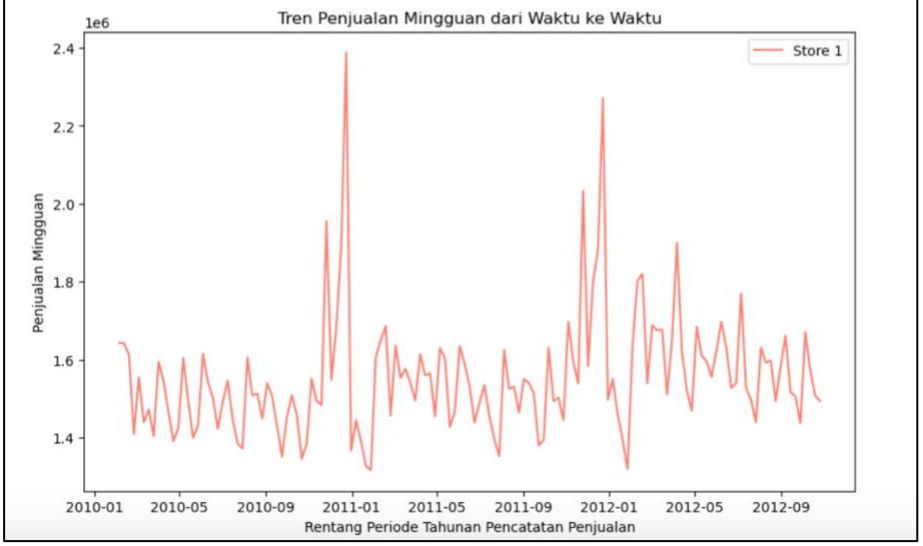
Dilakukan proses visualisasi box plot pada kolom "Penjualan Mingguan" dan "Suhu" yang bertujuan untuk mengidentikasi distribusi data, tren, dan kuartil dari distribusi data. Dengan menggunakan *IQR* untuk menampilkan box plot memahami pola distribusi data. Diperoleh hasil bahwa "Penjualan Mingguan" memiliki beberapa *outlier*, sedangkan di "Suhu" tidak memiliki *outlier*.

4. Tren Penjualan Mingguan, Bulanan, Dan Tahunan

4.1.1 Tren Penjualan Mingguan Dari Waktu Ke Waktu 2010-2012

4. Visualisasi Tren Penjualan Mingguan, Bulanan, Dan Tahunan

```
#visualisasi tren penjualan mingguan dari rentang tahun 2010-2012
import matplotlib.pyplot as plt

# tren penjualan
data_new["Tanggal Periode"] = pd.to_datetime(data_new["Tanggal Periode"])
plt.figure(figsize=(10, 6))
toko_data = data_new[data_new["Toko"] == 1]
plt.plot(toko_data["Tanggal Periode"], toko_data["Penjualan Mingguan"], label="Store 1", color="salmon")
plt.title("Tren Penjualan Mingguan dari Waktu ke Waktu")
plt.xlabel("Rentang Periode Tahunan Pencatatan Penjualan")
plt.ylabel("Penjualan Mingguan")
plt.legend()

plt.show()
```



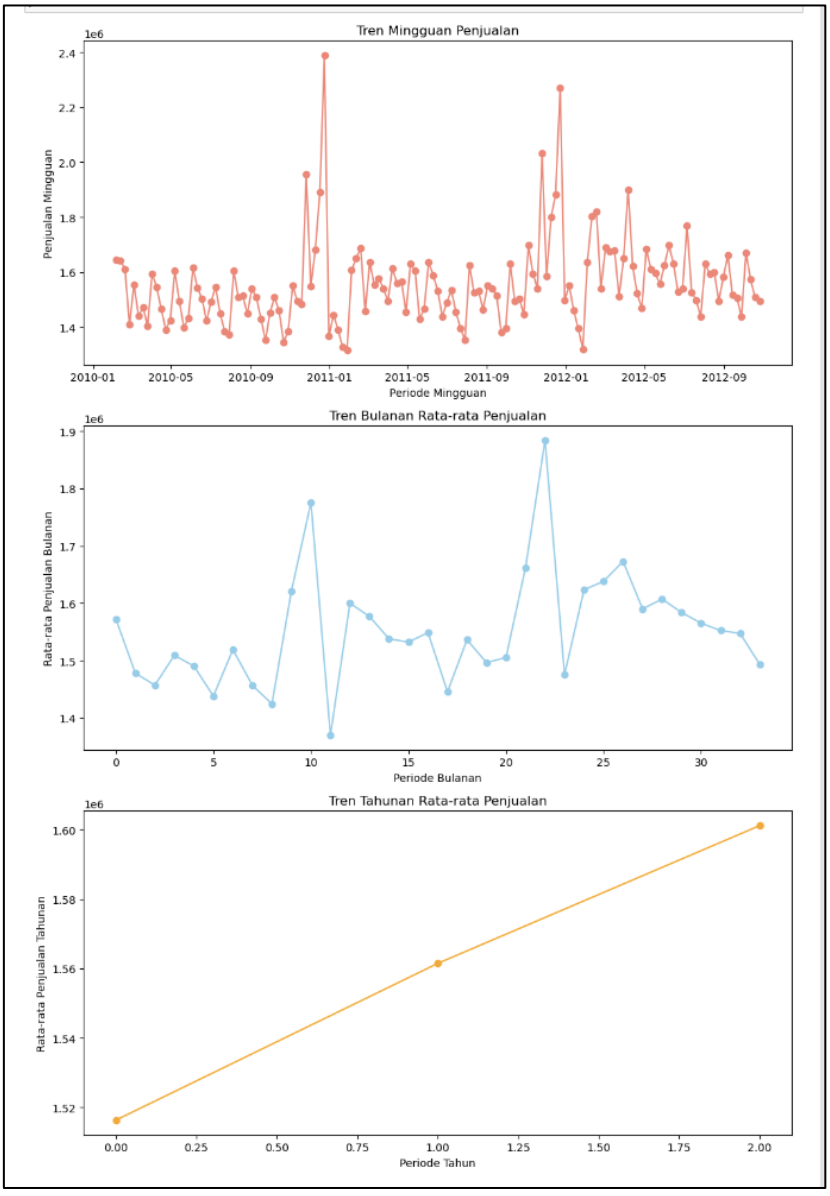
tampilan visualisasi tren penjualan mingguan dari waktu ke waktu 2010-2012

Dilakukan visualisasi untuk mengetahui bagaimana karakteristik data dalam tren dikolom “Penjualan Minggun” dalam rentang waktu 2010-2012. Untuk memperlihatkan apakah terdapat kecendrungan penurunan atau kenaikan penjualan mingguan di waktu tertentu. Dengan menggunakan line chart menggunakan sumbu x rentang periode tahunan pencatatan penjualan mingguan dan sumbu y merupakan rentang frekuensi penjualan mingguan.

Diperoleh hasil bahwa terjadi peningkatan dan penurunan fluktuatif yang tidak teratur dan tidak stabil, namun berdasarkan visualisasi tersebut “Penjualan Mingguan” cenderung meningkat. Puncaknya adalah dibulan januari 2011 dan januari 2012.

4.1.2 Tren Penjualan Mingguan Dari Waktu Mingguan, Bulanan, Dan Tahunan

```python
#visualisasi rata-rata tren penjualan mingguan, bulanan, dan tahunan

import matplotlib.pyplot as plt

#membuat figure dan axis
fig, axs = plt.subplots(3, 1, figsize=(10, 15))

#plot tren mingguan
axs[0].plot(data_new["Tanggal Periode"], data_new["Penjualan Mingguan"], color="salmon", marker="o")
axs[0].set_title("Tren Mingguan Penjualan")
axs[0].set_xlabel("Periode Mingguan")
axs[0].set_ylabel("Penjualan Mingguan")

#plot tren bulanan
monthly_sales = data_new.groupby("Month")["Penjualan Mingguan"].mean()
axs[1].plot(monthly_sales.index, monthly_sales.values, color="skyblue", marker="o")
axs[1].set_title("Tren Bulanan Rata-rata Penjualan")
axs[1].set_xlabel("Periode Bulanan")
axs[1].set_ylabel("Rata-rata Penjualan Bulanan")

#plot tren tahunan
yearly_sales = data_new.groupby("Year")["Penjualan Mingguan"].mean()
axs[2].plot(yearly_sales.index, yearly_sales.values, color="orange", marker="o")
axs[2].set_title("Tren Tahunan Rata-rata Penjualan")
axs[2].set_xlabel("Periode Tahun")
axs[2].set_ylabel("Rata-rata Penjualan Tahunan")


plt.tight_layout()
plt.show()
```



tampilan visualisasi tren penjualan mingguan dari waktu mingguan, bulanan, dan tahunan

Dilakukan visualisasi untuk menampilkan rata-rata tren penjualan mingguan, bulanan, dan tahunan. Visualisasi tersebut terdiri dari tiga plot yang masing-masing menunjukkan pola penjualan dari perspektif waktu yang berbeda. Pada plot pertama, divisualisasikan tren penjualan mingguan dari rentang tahun 2010-2012 yang menunjukkan adanya penjualan tahunan yang meningkat dan menurun secara fluktuatif, namun cenderung meningkat.

Pada plot kedua, divisualisasikan tren rata-rata penjualan bulanan dari data yang dikelompokkan berdasarkan bulan yang meningkat dan menurun secara fluktuatif. Dan, pada plot ketiga divisualisasikan tren rata-rata penjualan tahunan dari data yang dikelompokkan berdasarkan tahun yang menunjukkan adanya peningkatan secara signifikan disetiap tahun selama 2010-2012 yang dibagi menjadi rentang tertentu.

Kesimpulan :

Berdasarkan proses analisis deskriptif yang telah dilakukan untuk mengetahui karakteristik data. Diketahui beberapa hal, diantaranya yakni adanya data pada “Penjualan Mingguan” memiliki variasi signifikan dari waktu ke waktu berdasarkan rata-rata penjualannya disertai dengan nilai standar deviasi yang tinggi. Pada “Hari Libur” memiliki karakteristik jenis data biner dengan mayoritas memiliki nilai 0 yang berarti adanya mayoritas hari kerja. Pada “Suhu” memiliki variasi disetiap tahunnya, namun cenderung konsisten yang dibuktikan dengan adanya kuartil yang relatif sama dari tahun ke tahun.

Berdasarkan proses perhitugan kurtosis dan skewness diketahui adanya kecendrungan distribusi yang miring ke kanan pada “Penjualan Mingguan” dan “Hari Libur” serta distribusi lebih simetris namun datar pada “Suhu”. Terdapat pula banyak nilai ekstrem dari “Penjualan Mingguan” dan “Hari Libur”. Berdasarkan proes perhitungan analisis korelasi sebagai informasi tambahan diketahui bahwa adanya korelasi positif lemah dari skala tahunan, bulanan, dan mingguan terhadap “Penjualan Mingguan”. Adanya korelasi positif lemah dari “Hari Libur” dan “Penjualan Mingguan”.

Kemudian, adanya korelasi negatif dari “Suhu” dan “Penjualan Mingguan”, sehingga semakin tinggi suhu maka semakin menurun penjualan. Berdasarkan proses

analisis visual yang telah dilakukan untuk mengetahui karakteristik data. Diketahui beberapa hal, di antaranya yakni berdasarkan visualisasi histogram dan uji normalitas diketahui tidak ada data yang berdistribusi normal, mekipun dalam visualisasinya pada "Penjualan Mingguan" dan "Suhu" visualisasinya mirip dengan distribusi normal. Kemudian, dilakukan visualisasi untuk mengetahui proporsi "Hari Libur" menggunakan pie chart.

Lalu, berdasarkan visualisasi dan tren "Penjualan Mingguan" diseluruh rentang tahun 2010-2011 memiliki visualisasi fluktiatuf yang cenderung mengalami peningkatan. Kemudian, untuk visualisasi tren rata-rata direntang mingguan dan bulanan juga cenderung mengalami visualisasi fluktuatif yang cenderung mengalami peningkatan. Sedangkan, pada visualisasi tahunan diperlihatkan adanya peningkatan penjualan signifikan di setiap tahunnya. Dengan demikian, diperlihatkan bahwa data pada toko 1 memiliki variasi yang signifikan dari waktu ke waktu, terdapat fluktuasi yang cenderung meningkat dalam penjualan, serta adanya pengaruh faktor-faktor seperti hari libur dan suhu terhadap penjualan mingguan.

Refrensi :

Link Sumber Dataset : https://www.kaggle.com/

Link Dataset : https://www.kaggle.com/datasets/varsharam/walmart-sales-dataset-of-45stores