Peningkatan Keandalan Analisis Penjualan dengan Pendeteksian Pencilan Menggunakan Metode Z-score

Asa Do'a Uyi (122450005), Try Yani Rizki Nur Rohmah (122450020), Marleta Cornelia Leander (122450092), Sofyan Fauzi Dzaki Arif (122450116), Amalia Melani Putri (122450122)

Pendahuluan

Di zaman bisnis yang memiliki sifat yang dinamis, kita memerlukan tingkat yang mendalam untuk memahami perilaku penjualan. Hal ini merupakan sesuatu yang sangat penting dan berguna untuk kelangsungan dan kesuksesan sebuah perusahaan. Kegiatan menganalisis penjualan yang andal adalah dasar yang sangat penting untuk mengambil sebuat keputusan yang tepat dan strategis. Namun, dalam seiring waktu, sering terjadi anomali atau yang biasa disebut dengan pencilan(*Outlier*) di dalam proses analisis tersebut. Hal tersebut dapat membuat suatu kebingungan terhadap pemahaman kita akan data penjualan yang kita miliki. Untuk mengatasi tantangan ini, penggunaan metode Z-score telah menjadi pilihan yang tepat dalam mendeteksi dan menangani pencilan agar sebuah perusahaan dapat mengambil keputusan yang tepat, cerdas, dan terinformasi.

Kata Kunci: Pencilan, Metode Z-Score, Analisis Penjualan, dan Peningkatan.

1. Metode

Dalam pembahasan kali ini untuk mencari pencilan data

Rumus z-score:

$$Z = \frac{(x - \mu)}{\sigma}$$

Z:z-score
X: Nilai yang di amati
μ: rata-rata
σ: standar deviasi

Untuk mengimplementasikan metode Z-score agar dapat mendeteksi outlier pada data penjualan air, dapat dilakukan langkah-langkah berikut:

1. Mengunggah dataset. Dataset berjenis csv dan memanfaatkan modul Pandas, Numpy, and Mathplotlib.
2. Menggunakan fungsi **'detect_outliers_zscore'** untuk mengidentifikasi nilai pencilan.
3. Hasil deteksi menampilkan indeks data outlier seta rincian data yang dianggap aneh. Proses ini membantu pengidentifikasian data yang mungkin menganggu penelitian.
4. Code yang disajikan menghasilkan output yang menyajikan informasi tentang dataset, statistik deskriptif, deteksi outlier, distribusi data, dan visualisasi data dalam bentuk *boxplot.*

2. Pembahasan

Pengimplementasianya dalam kode python dengan menggunakan colab untuk mencari data pencilan dengan menggunakan z-score dilakukan sebagai berikut:

a. Mengimport datanya yang sudah di cari dengan menggunakan import files yang sudah di taruh pada google.colab dan data yang di gunakan adalah csv yang sebelumya data yang dapat di akses melalui link: https://katalog.data.go.id/dataset/nilai-penjualan-air-minum-berdasarkan-kategori-pelanggan-perumda-air-minum-tirta-giri-nata-di-w

```python
from google.colab import files
import pandas as pd

# Mengunggah file dataset
uploaded = files.upload()

# Membaca dataset yang diunggah
for file_name in uploaded.keys():
    if file_name.endswith('.csv'):
        df = pd.read_csv(file_name)
        print("Dataset", file_name, "telah berhasil diunggah.")

# Tampilkan dataset yang berhasil diunggah
print("Isi dataset:")
print(df.head())

```

Output yang di hasilkan:

```
Dataset data_penjualan-air (3).csv telah berhasil diunggah.
Isi dataset:
   id  kode_provinsi nama_provinsi  kode_kabupaten_kota nama_kabupaten_kota  \
0   1             32    JAWA BARAT                 3274        KOTA CIREBON
1   2             32    JAWA BARAT                 3274        KOTA CIREBON
2   3             32    JAWA BARAT                 3274        KOTA CIREBON
3   4             32    JAWA BARAT                 3274        KOTA CIREBON
4   5             32    JAWA BARAT                 3274        KOTA CIREBON

                  kategori_pelanggan  nilai_penjualan  satuan  tahun
0                        RUMAH TANGGA     1.000000e+13  RUPIAH   2022
1  HOTEL, TOKO, INDUSTRI, PERUSAHAAN     1.701368e+10  RUPIAH   2022
2        RUMAH SAKIT NON PEMERINTAH      4.481373e+08  RUPIAH   2022
3                TEMPAT PERIBADATAN      2.805502e+08  RUPIAH   2022
4           SARANA / FASILITAS UMUM      1.971827e+08  RUPIAH   2022
```

Data data yang dihasilkan sudah diurutkan sesuai dengan baris masing masing dan baris yang tersedia adalah id, kode_provinsi, nama_provinsi, kode_kabupaten_kota, nama_kabupaten_kota, kategori_pelanggan, nilai_penjualan, satuan, tahun. Dengan jumlah kolom yaitu ada 40 kolom.

b. Dengan memanfaatkan fungsi dari outlier detector dari pemrograman berbasis fungsi dalam mencari nilai statistik dengan menggunakan metrik statistik seperti mean ataupun standar deviasi untuk menentukan apakah suatu titik data dianggap sebagai outlier. Pemanfaatan metode statistik ini kita bisa mencari nilai pencilan dalam suatu data yaitu z-score dengan rumus $\frac{(x-\mu)}{\sigma}$ jika telah diketahui nilai mean dan juga standar deviasinya.

```python
import pandas as pd
import numpy as np

def detect_outliers_zscore(data, threshold=3):
    mean = np.mean(data)
    std_dev = np.std(data)
    z_scores = [(x - mean) / std_dev for x in data]
    outliers = np.where(np.abs(z_scores) > threshold)
    return outliers[0]

# Membaca dataset dari file CSV
data = pd.read_csv('data_penjualan-air.csv')

# Menentukan kolom yang akan dianalisis (misalnya: 'nilai_penjualan')
column_to_analyze = 'nilai_penjualan'

# Mendeteksi pencilan
outliers_indices = detect_outliers_zscore(data[column_to_analyze])

# Menampilkan keterangan
print("Data yang mengalami pencilan berdasarkan Z-score method dengan threshold 3:")
print("-"*70)
print("Kolom yang dianalisis:", column_to_analyze)
print("Jumlah data:", len(outliers_indices))
print("Indeks data yang mengalami pencilan:", outliers_indices)
print("-"*70)
print("Detail transaksi yang mencurigakan:")
print(data.loc[outliers_indices])
```

output yang dikeluarkan merupakan nilai ataupun data yang dimana nilai_penjualanya cukup jauh dibandingkan dengan nilai penjualan di katagori_pelanggan yang lain, dan didapatkan bahwasannya nilai pada kategori rumah tangga, dan tempat

peribadatan memiliki nilai_penjualan yang sangat tinggi, bisa kita sebut juga sebagai nilai pencilan. nilai _penjualan pada kategori rumah tangga adalah 1.000000e+13 sedangkan untuk tempat peribadatan adalah 8.888890e+12.

```
Data yang mengalami pencilan berdasarkan Z-score method dengan threshold 3:
----------------------------------------------------------------------
Kolom yang dianalisis: nilai_penjualan
Jumlah data: 2
Indeks data yang mengalami pencilan: [ 0 11]
----------------------------------------------------------------------
Detail transaksi yang mencurigakan:
    id  kode_provinsi nama_provinsi  kode_kabupaten_kota nama_kabupaten_kota  \
0    1             32    JAWA BARAT                 3274        KOTA CIREBON
11  12             32    JAWA BARAT                 3274        KOTA CIREBON

    kategori_pelanggan  nilai_penjualan  satuan  tahun
0          RUMAH TANGGA     1.000000e+13  RUPIAH   2022
11  TEMPAT PERIBADATAN     8.888890e+12  RUPIAH   2021
```

c. Mencari mean dan standar deviasi menggunakan fungsi yang tersedia di dalam pandas dan juga numpy dan rumus dari z-score bisa menginputkan sendiri dengan cara memasuki rumus yang telah ditentukan dalam membuat rumus dari z-score.

```
[ ]
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

def detect_outliers_zscore(data, threshold=3):
    mean = np.mean(data)
    std_dev = np.std(data)
    z_scores = [(x - mean) / std_dev for x in data]
    outliers = np.where(np.abs(z_scores) > threshold)
    return outliers[0]

# Membaca dataset dari file CSV
df = pd.read_csv("data_penjualan-air.csv")

# Menampilkan informasi dataset
print("Informasi dataset:")
print(df.info())

# Menampilkan statistik deskriptif
print("\nStatistik Deskriptif:")
print(df.describe())

# Menghitung mean (rata-rata) dan standar deviasi dari kolom 'nilai_penjualan'
mean = df['nilai_penjualan'].mean()
std_dev = df['nilai_penjualan'].std()

print("\nRata-rata (mean) kolom 'nilai_penjualan':", mean)
print("Standar Deviasi (std) kolom 'nilai_penjualan':", std_dev)

# Menampilkan boxplot
plt.figure(figsize=(8,6))
plt.boxplot(df['nilai_penjualan'])
plt.title('Boxplot Kolom "nilai_penjualan"')
plt.ylabel('Nilai')
plt.show()

# Mendeteksi pencilan
outliers_indices = detect_outliers_zscore(df['nilai_penjualan'])

# Menampilkan keterangan
print("\nData yang mengalami pencilan berdasarkan Z-score method dengan threshold 3:")
print("-"*70)
print("Kolom yang dianalisis: nilai_penjualan")
print("Jumlah data:", len(outliers_indices))
print("Indeks data yang mengalami pencilan:", outliers_indices)
print("-"*70)
print("Detail transaksi yang mencurigakan:")
print(df.loc[outliers_indices])

```

Output yang dihasilkan informasi yang diberikan merupakan informasi mengenai datanya yang tersedia, dan juga beberapa informasi mengenai statistik deskripsi yang diberikan dalam setiap kolom secara lengkap.

Statistik Deskriptif	
Count	4.00E+01
Mean	4.79E+11
Std	2.09E+12
Min	5.00E+02
25%	2.70E+08
50%	1.51E+09
75%	1.57E+10
Max	1.00E+13

Dengan nilai rata rata 478979970487.925 dan standar deviasi-nya adalah 2086851511961.632

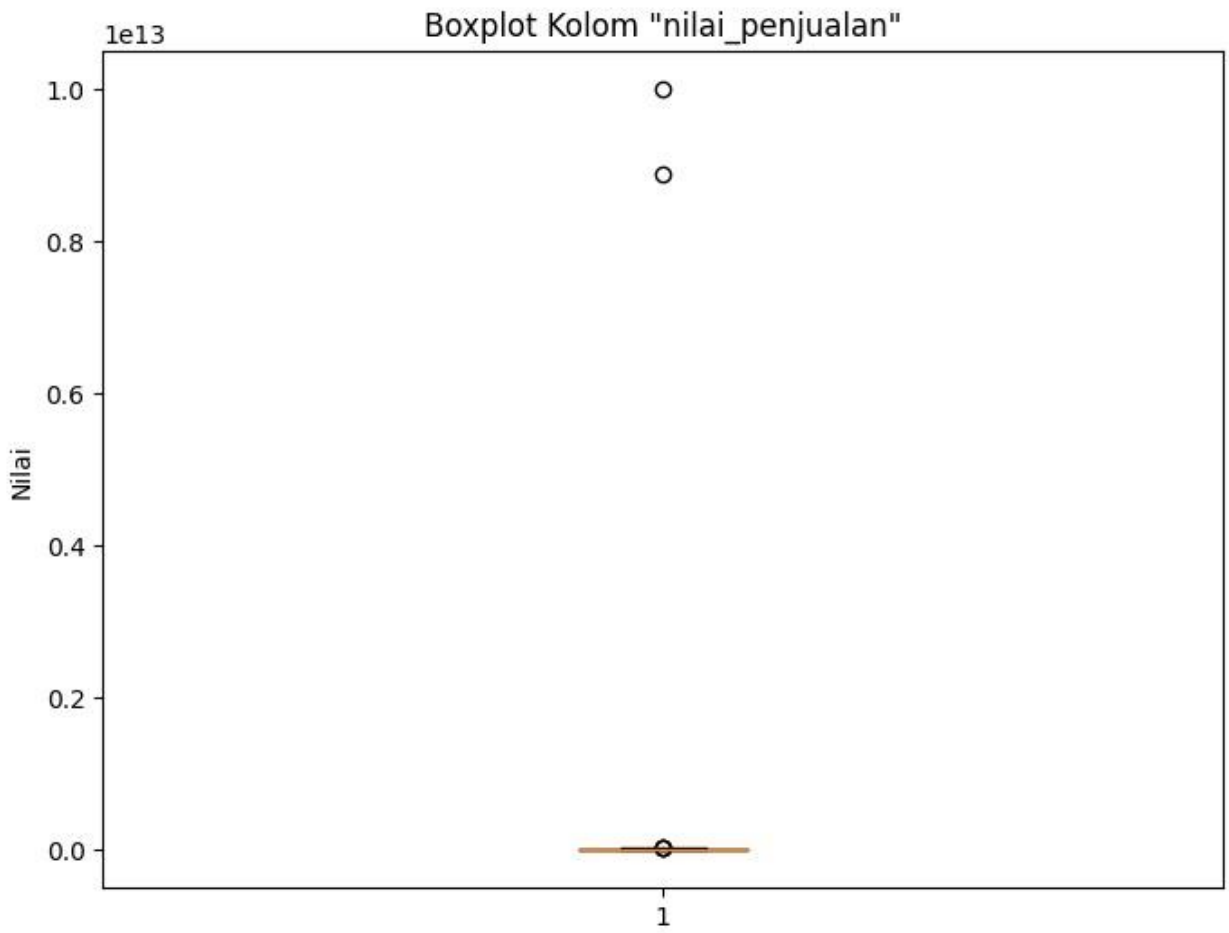
```
Informasi dataset:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40 entries, 0 to 39
Data columns (total 9 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   id                   40 non-null     int64
 1   kode_provinsi        40 non-null     int64
 2   nama_provinsi        40 non-null     object
 3   kode_kabupaten_kota  40 non-null     int64
 4   nama_kabupaten_kota  40 non-null     object
 5   kategori_pelanggan   40 non-null     object
 6   nilai_penjualan      40 non-null     float64
 7   satuan               40 non-null     object
 8   tahun                40 non-null     int64
dtypes: float64(1), int64(4), object(4)
memory usage: 2.9+ KB
None

Statistik Deskriptif:
              id  kode_provinsi  kode_kabupaten_kota  nilai_penjualan
count  40.000000           40.0                 40.0     4.000000e+01
mean   20.500000           32.0               3274.0     4.789800e+11
std    11.690452            0.0                  0.0     2.086852e+12
min     1.000000           32.0               3274.0     5.000000e+02
25%    10.750000           32.0               3274.0     2.703675e+08
50%    20.500000           32.0               3274.0     1.514463e+09
75%    30.250000           32.0               3274.0     1.567013e+10
max    40.000000           32.0               3274.0     1.000000e+13

            tahun
count    40.00000
mean   2020.00000
std       1.43223
min    2018.00000
25%    2019.00000
50%    2020.00000
75%    2021.00000
max    2022.00000

Rata-rata (mean) kolom 'nilai_penjualan': 478979970487.925
Standar Deviasi (std) kolom 'nilai_penjualan': 2086851511961.632
```

Selanjutnya adalah memvisualisasikan datanya dengan bentuk plot jika kita lihat bahwasanya ada dua pencilanya yang dimana nilai nya jauh dari data atau rata rata yang lainya yang berkisar 4.789800e+11 sedangkan nilai yang muncul pada pencilan yaitu di kategori rumah tangga yaitu 1.000000e+13 yang jauh diatas rata rata.




```
Data yang mengalami pencilan berdasarkan Z-score method dengan threshold 3:
--------------------------------------------------------------------------
Kolom yang dianalisis: nilai_penjualan
Jumlah data: 2
Indeks data yang mengalami pencilan: [ 0 11]
--------------------------------------------------------------------------
Detail transaksi yang mencurigakan:
    id  kode_provinsi nama_provinsi  kode_kabupaten_kota nama_kabupaten_kota  \
0    1             32    JAWA BARAT                 3274        KOTA CIREBON
11  12             32    JAWA BARAT                 3274        KOTA CIREBON

    kategori_pelanggan  nilai_penjualan  satuan  tahun
0         RUMAH TANGGA     1.000000e+13  RUPIAH   2022
11  TEMPAT PERIBADATAN     8.888890e+12  RUPIAH   2021
```

Link Colab : https://colab.research.google.com/drive/1BudA91wRcghQqmVoJp2VyTqzQEu9p5-H?usp=sharing

3. Kesimpulan

Dalam upaya untuk meningkatkan keandalan analisis penjualan, penggunaan metode Z-score sudah terbukti jika pendekatan efektif untuk mendeteksi data pencilan atau outlier adalah dengan metode Z-score. Dengan menerapkan metode Z-score dalam dataset penjualan air, didapatkan nilai-nilai yang signifikan secara statistik dalam pengidentifikasian outlier dalam data penjualan.

Dalam hasil analisis ditunjukkan jika ada dua nilai penjualan yang signifikan, yaitu pada kategori rumah tangga dan tempat peribadatan, yang dapat dianggap sebagai nilai pencilan. Nilai penjualan yang sangat tinggi pada kategori-kategori ini secara jelas ditunjukkan jika dibandingkan dengan nilai rata-rata yang lebih rendah dari nilai penjualan tersebut.

Dengan demikian, melalui penggunaan metode Z-score ini dikatakan berhasil dalam mengidentifikasi dan menggambarkan outlier di dataset penjualan air ini dan dapat membantu perusahaan untuk memahami pola penjualan yang sebenarnya dan mengambil keputusan yang lebih tepat berdasarkan analisis data yang lebih akurat dan andal.

Daftar Pustaka

[1] Y. &. Z. Q. Zhang, "Application of Z-Score Method in Detecting Outliers of Financial Data.," *In 2020 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE)* , pp. 1-5, 2020.

[2] W. &. o. McKinney, "Data Structures for Statistical Computing in Python.," *Proceedings of the 9th Python in Science Conference,* pp. 56-61, 2010.