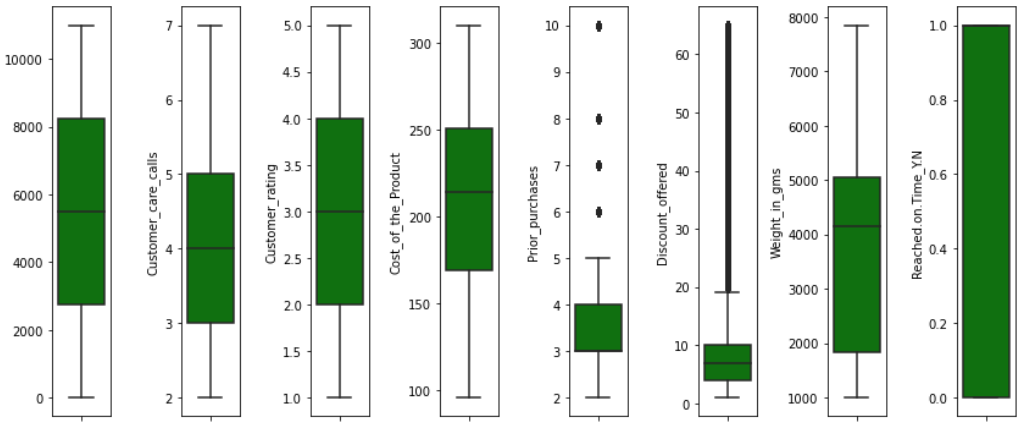
EDA

1. Descriptive Statistics
   a. Tidak ditemukan. Tipe data dan nama kolom dan isinya sudah sesuai. Menggunakan df.info
   b. Sudah sesuai, tidak ditemukan kolom yang kosong
   c. Nilai mean dan median discount_offered, weight_in_gms dan Reached.on.Time_Y.N

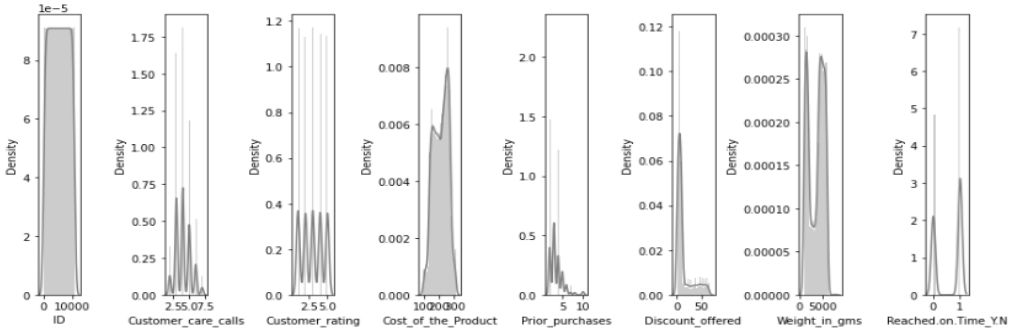
2. Univariate Analysis
   - Pada kolom Mode_of_Shipment, jumlah mode Ship lebih tinggi dibandingkan yang lain

```
Ship      7462
Flight    1777
Road      1760
Name: Mode_of_Shipment, dtype: int64
```

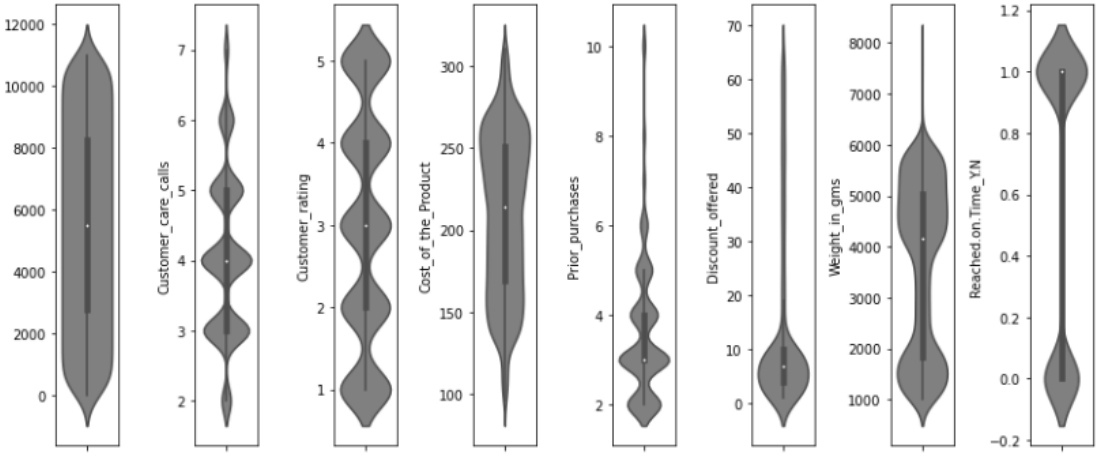
   - Ditemukan outlier pada kolom prior_purchases dan discount_offered. Untuk kolom2 lain tidak ditemukan outlier. Berikut visualisasi:



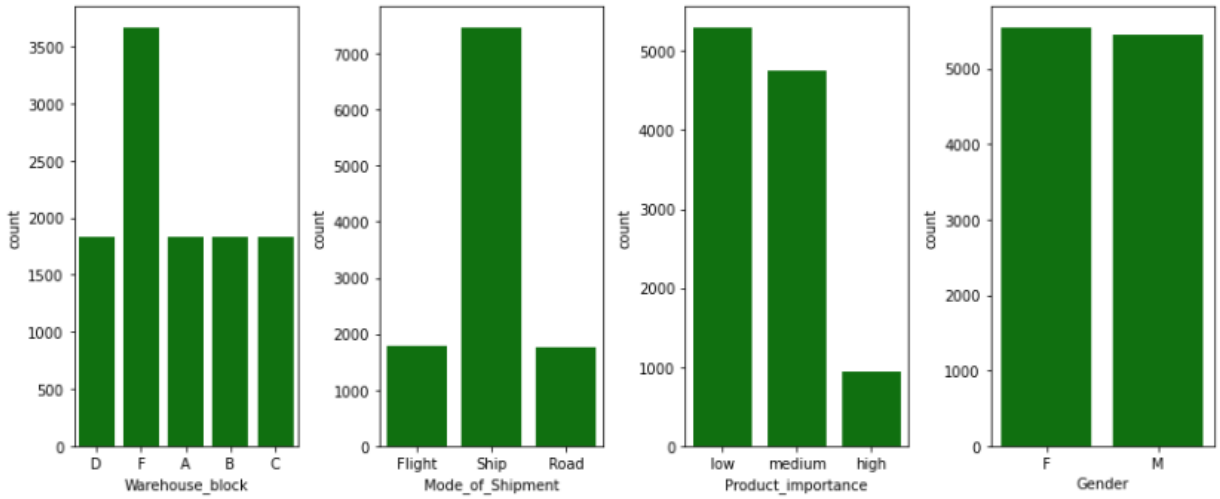
   Untuk data outlier pada data preprocessing harus dianalisa terlebih dulu misalkan data harus dibersihkan terlebih dulu.
   - Kolom customer care call distribusnya multi modal dan banyak data lonjakan
   - Kolom customer rating distribusinya multi modal dan banyak data lonjakan
   - Kolom cost of the product mendekati normal, namun ada lonjakan
   - Kolom prior purchase mendekati normal namun memiliki tail da nada lonjakan
   - Kolom discount_offered distribusinya positif skewed, dan ada lonjakan yang cukup jauh
   - Kolom weight in gms distribusi nya multi modal da nada lonjakan di beberapa titik
   - Kolom reached on time bimodal



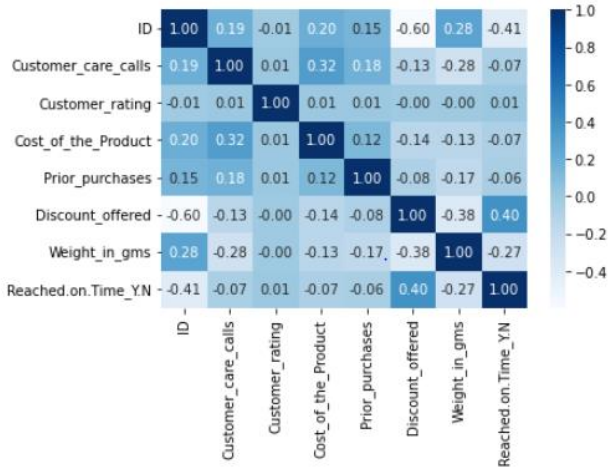
- Distribusi customer care calls terbanyak ada pada 3 kali telepon dan 4 kali telepon
- Untuk customer rating 1-5 distribusinya rata
- Untuk Cost of the product nilai distribusi terbanyak ada di angka 260
- Untuk prior purchase distribusi terbanyak ada di nilai ke3. Outlier banyak ditemukan pada nilai 6 ke atas
- Untuk discount offered terbanyak ada di 2-10. Outlier banyak ditemukan di nilai 10 ke atas
- Untuk weight in gms nilai terbanyak ada di 4500 – 5800 dan 1500. Outlier ditemukan di nilai 6000 ke atas



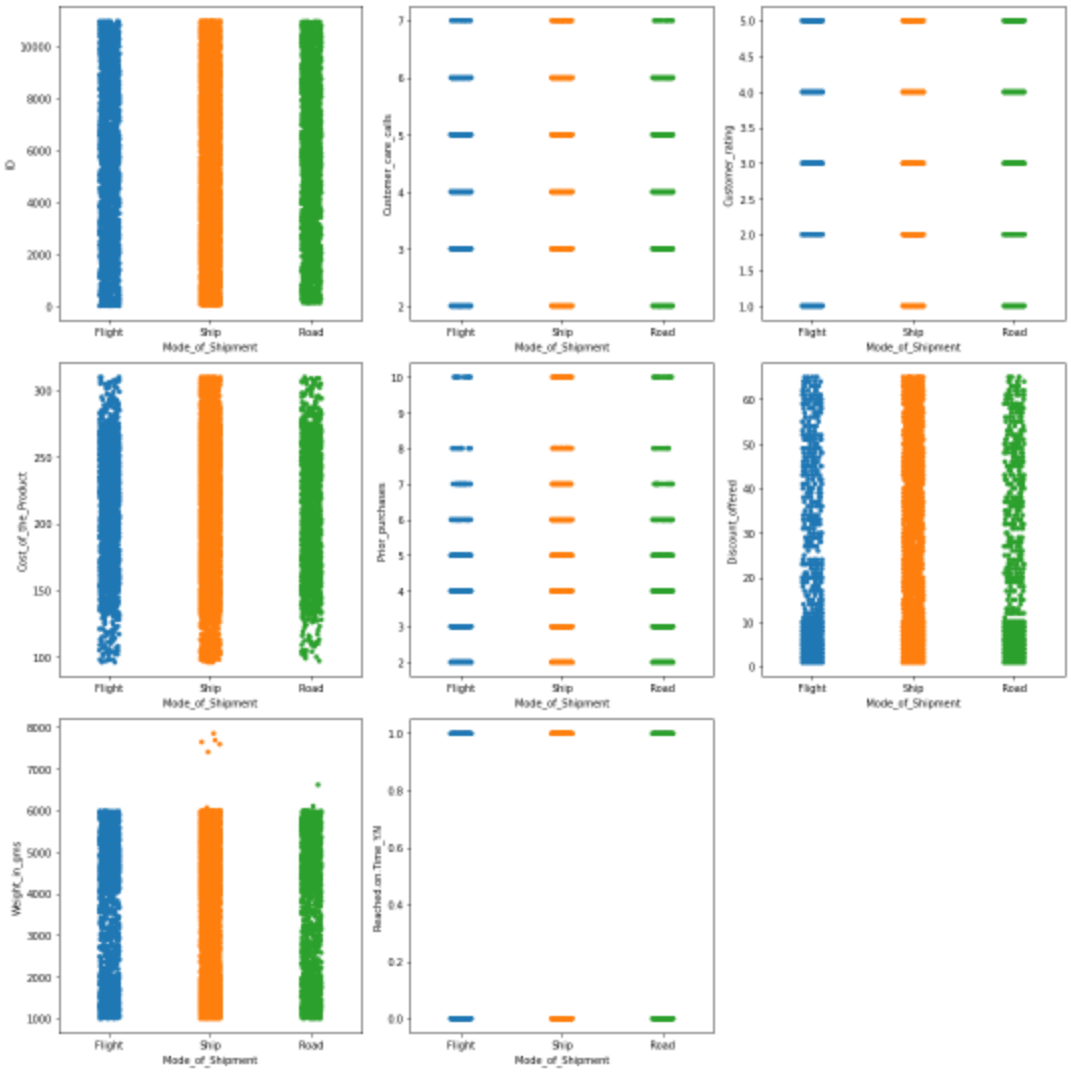
- Distribusi warehouse pada kolom F timpang. Untuk blok lain sudah seimbang
- Distribusi mode shipment “ship” timpang dibandingkan dengan mode flight dan road
- Distribusi product importance “high” timpang
- Distribusi gender dan female sudah seimbang



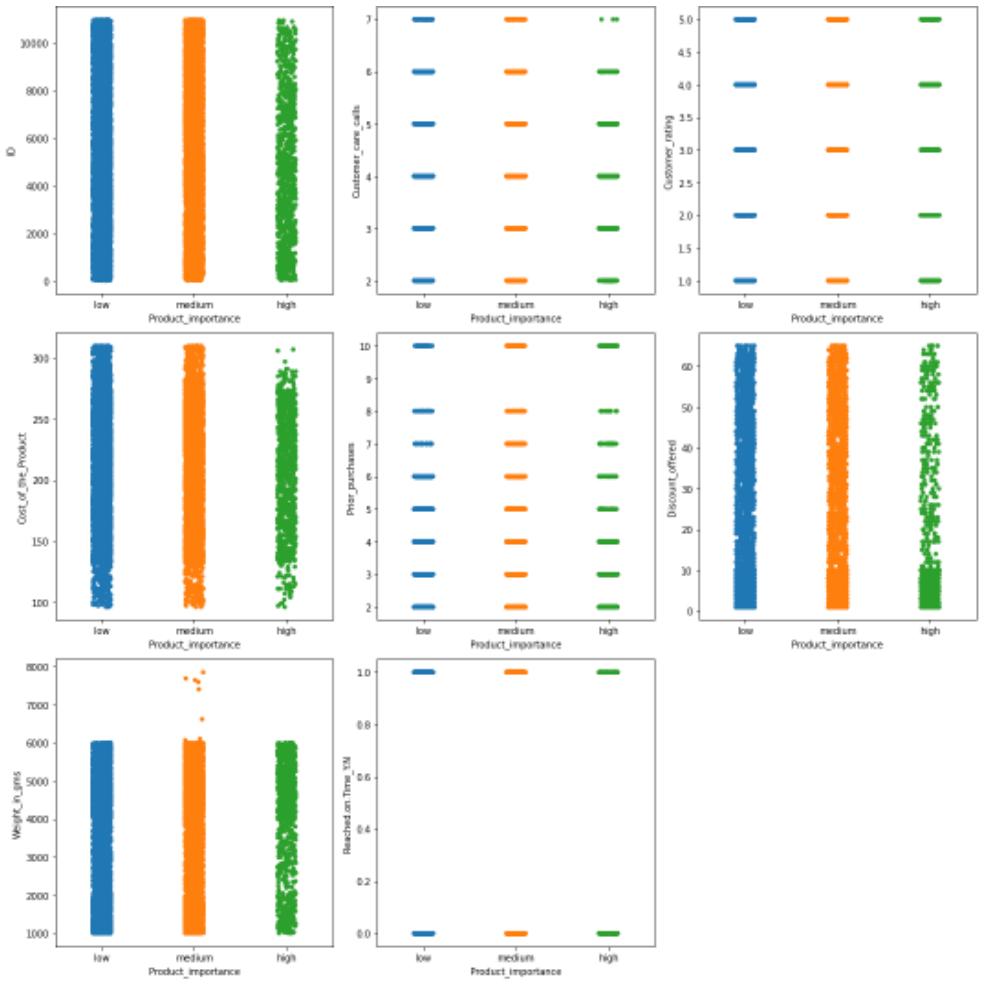
3. Multivariate Analysis
   - Korelasi antar variable lemah nilainya < 70



   - Sulit untuk membaca scatterplot
   - Mode shipment “ship” lebih padat
   - Cost of the product 0-300 rata distribusninya di mode shipment “ship”, untuk road dan flight 140-270
   - Discound offered lebih banyak diberikan ke mode shipment “ship”



- Product importance high cost of the product nya kebanyakan ada dikisaran130-260
- Product importance high lebih sedikit diberi discount offered



4. Business Insight
   - Masih banyak jumlah pengiriman yang tidak on time (1)

     ```
     1    6563
     0    4436
     Name: Reached.on.Time_Y.N, dtype: int64
     ```

   - Warehouse F jumlahnya mendominasi
   - Mode of Shipment Ship jumlah nya terbanyak
   - Product importance high jumlahnya jauh lebih sedikit dibandingkan dengan low dan medium