LAPORAN HOMEWORK DATA PROCESSING

ANGGOTA KELOMPOK

Celestial Randy

Sonia Epifany Sandah

Oky Hariawan

Risca Naquitasia

Mochamad Choiril Iman

Ahmad Reza

Yehezkiel Novianto A.

DATA CLEANSING

- Handle Missing Values
- Handle Duplicated Data
- Handle Outliers
- Feature Transformation
- Feature Encoding
- Handle Class Imbalance

HANDLE MISSING VALUE

Insight:

- Ditemukan 4 kolom yang memiliki missing values yaitu : Company, Agent, Country, Children

	feature	missing_value	percentage
0	company	112593	94.307
1	agent	16340	13.686
2	country	488	0.409
3	children	4	0.003

What We Do:

- Kolom children karena hanya memiliki 4 baris data yang kosong maka missing value tersebut akan di hapus
- Kolom agent karena merupakan ID maka nilai akan di set 0 untuk data yang tidak memiliki ID dan 1 untuk data yang memiliki ID.
- Kolom company karena banyak missing value (94%) maka feature company dapat di drop
- Kolom country karena sebagian besar nilainya adalah PRT dan missing value rationya kecil, maka missing value tersebut akan diisi dengan nilai PRT

HANDLE DUPLICATED VALUE

Insight:

- Kami tidak menemukan adanya indikasi data ganda atau duplicate.

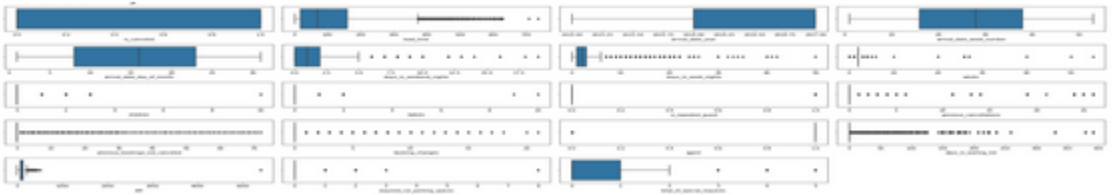
```
[ ] df_clean.duplicated().sum()

    0
```

```
[ ] df_clean.duplicated(subset=['name', 'email', 'phone-number', 'credit_card']).sum()

    0
```

HANDLE OUTLIERS

Insight:

- Dari visualisasi disamping dapat terlihat bahwa dalam data ini memiliki banyak outlier. Untuk mengatasi hal tersebut kami melakukan beberapa treatment



What We Do:

- Mengubah outlier menjadi batas atas (high limit)
- Ganti Nilai Outlier Menjadi Nilai Terdekat
- Ubah Nilai Outlier Menjadi 2 Variasi

HANDLE OUTLIERS
Mengubah outlier menjadi batas atas (high limit)

What We Do:

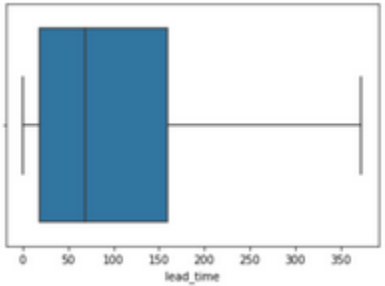
- Mengubah outlier menjadi batas atas dari feature. Tujuannya adalah membatasi value feature agar persebarannya tidak terlalu luas. Hal tersebut digunakan untuk feature: Lead time dan ADR

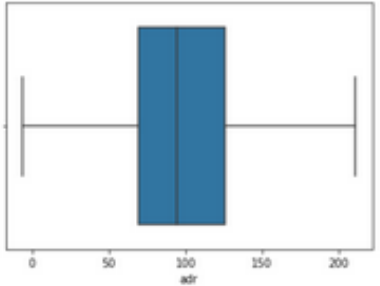
```
[ ] #Cek nilai outlier
    Q1 = df_clean['lead_time'].quantile(0.25)
    Q3 = df_clean['lead_time'].quantile(0.75)
    IQR = Q3 - Q1
    low_limit = Q1 - (1.5*IQR)
    high_limit = Q3 + (1.5*IQR)
```

```
[ ] low_limit,high_limit,Q1,Q3

    (-195.0, 373.0, 18.0, 160.0)
```

```
[ ] df_clean.loc[(df_clean.lead_time > high_limit,'lead_time')] = high_limit
    #Karena lead time dibatasi hanya 1 tahun
```

Result:





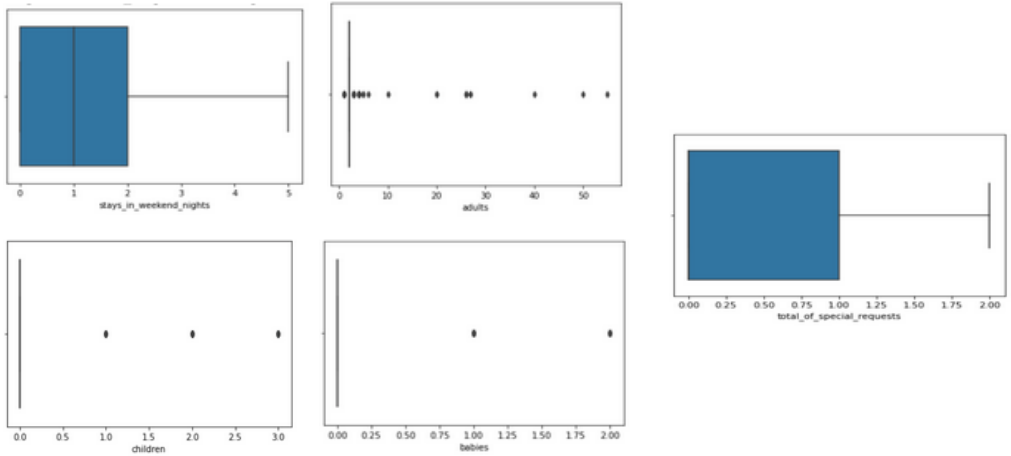
HANDLE OUTLIERS
Ganti Nilai Outlier Menjadi Nilai Terdekat

What We Do:

- Untuk handling outlier yang ke 2 adalah outlier dibiarkan. Tujuannya adalah karena untuk memberikan variasi data dikarenakan data hanya bertumpuk pada 1 value. Hal tersebut digunakan untuk feature: Stays in Weekend Nights, Adults, Children, Babies, Total Special Request

```
[ ] #Cek nilai Outlier
    Q1 = df_clean['stays_in_weekend_nights'].quantile(0.25)
    Q3 = df_clean['stays_in_weekend_nights'].quantile(0.75)
    IQR = Q3 - Q1
    low_limit = Q1 - (1.5*IQR)
    high_limit = Q3 + (1.5*IQR)

[ ] low_limit,high_limit,Q1,Q3

    (-3.0, 5.0, 0.0, 2.0)

[ ] df_clean.stays_in_weekend_nights.value_counts()

    0    51996
    2    33387
    1    30625
    4     1855
    3     1259
    6      153
    5       79
    8       60
    7       16
    9       11
    10       7
    12       5
    13       3
    16       3
    14       2
    18       1
    19       1
    Name: stays_in_weekend_nights, dtype: int64

[ ] df_clean.loc[(df_clean.stays_in_weekend_nights > high_limit,'stays_in_weekend_nights')] = 2.0
    #Karena kita anggap customer paling banyak stay 2 hari (modus = 0 diabaikan)
```

Result:

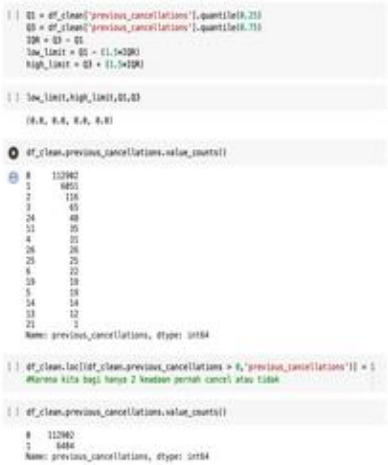


HANDLE OUTLIERS
Ubah Nilai Outlier Menjadi 2 Variasi

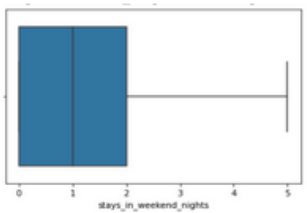
What We Do:

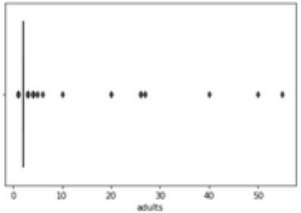
Mengubah nilai outlier menjadi hanya 2 variasi karena persebaran hanya tertumpuk pada 1 nilai dan secara logika feature tersebut dapat disederhanakan menjadi 2 nilai.

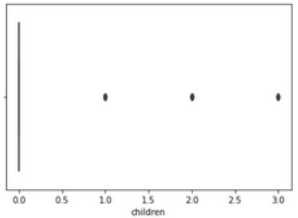
Hal tersebut digunakan untuk feature: Previous Cancellations, Previous Bookings Not Cancelled, Booking Changes, Days in Waiting List, Required Car Parking Spaces

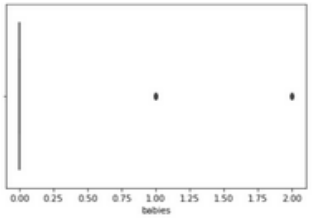


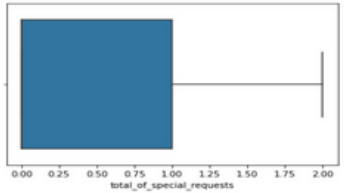
Result:





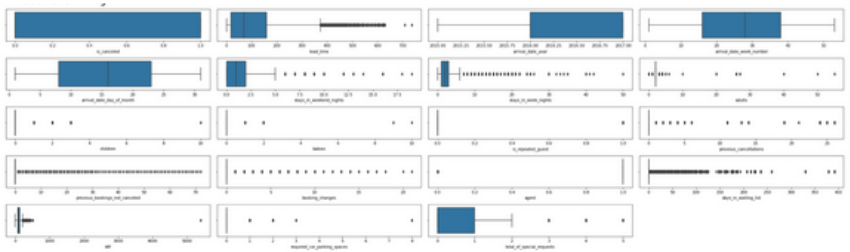




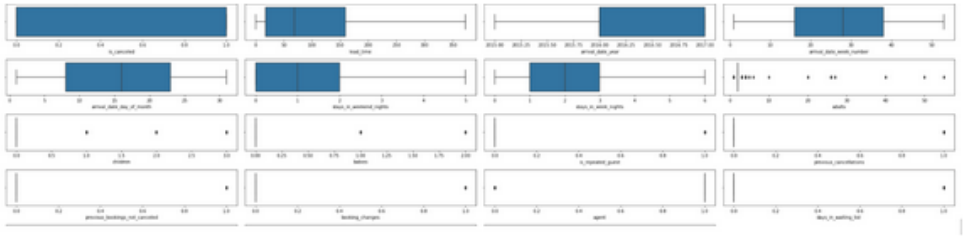


HANDLE OUTLIERS
Hasil Handling Outlier

Before:



After:



FEATURE TRANSFORMATION

What We Do:

- Kami melakukan normalized / re-scale kolom numerical yang memiliki angka variance tinggi guna mempermudah interpretasi beberapa model Machine Learning.

Result:

	lead_time_norm	arrival_date_week_number_norm	arrival_date_day_of_month_norm	days_in_waiting_list_norm	adr_norm
count	119386.000000	119386.000000	119386.000000	119386.000000	119386.000000
mean	0.274082	0.503173	0.493285	0.030975	0.492270
std	0.271204	0.261641	0.292693	0.173251	0.205512
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.048257	0.288462	0.233333	0.000000	0.347996
50%	0.184987	0.519231	0.500000	0.000000	0.464347
75%	0.428954	0.711538	0.733333	0.000000	0.608798
max	1.000000	1.000000	1.000000	1.000000	1.000000

FEATURE ENCODING

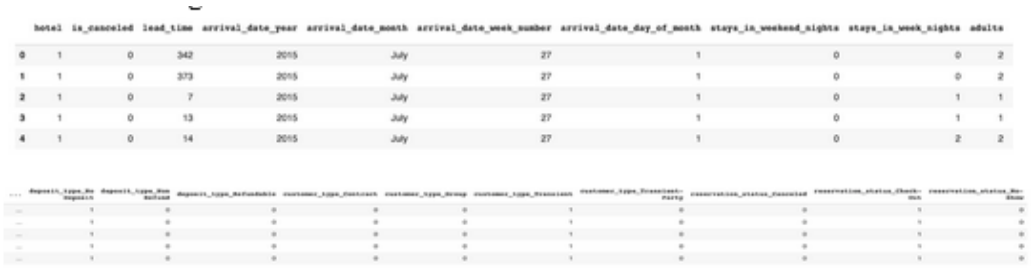
What We Do:

- Label Encoding untuk fitur hotel (karena ada 2 unique), Reserved Room Type dan Assigned Room Type (karena unique data nya ordinal).

- One Hot Encoding untuk fitur Distribution Channel, Deposit Type, Customer Type, Reservation Status dimana fitur tersebut memiliki nilai unique lebih dari 2 dan tidak bersifat ordinal.

Result:

	hotel	reserved_room_type	assigned_room_type
0	1	2.0	2.0
1	1	2.0	2.0
2	1	0.0	2.0
3	1	0.0	0.0
4	1	0.0	0.0



FEATURE TRANSFORMATION

What We Do:

- Memisahkan kolom y untuk target, x untuk kolom selain target
- Melakukan split data menjadi train dan test
- RUS dan ROS dengan package Imblearn

Result:

```
Original data train:
is_canceled
0               52608
1               30962
dtype: int64
Random under-sampling:
is_canceled
0               30962
1               30962
dtype: int64
Random over-sampling:
is_canceled
0               52608
1               52608
dtype: int64
```

FEATURE ENGINEERING

- Feature Selection
- Feature Extraction
- Additional Features
- Chosen Features

FEATURE SELECTION

What We Do:

- Menghapus kolom Name, Email, Phone-number, Credit Card karena tidak memiliki pola

Result:

```
df_clean[['name', 'email', 'phone-number', 'credit_card']].head()
```

	name	email	phone-number	credit_card
0	Ernest Barnes	Ernest.Barnes31@outlook.com	669-792-1661	••••••••••4322
1	Andrea Baker	Andrea_Baker94@aol.com	858-637-6955	••••••••••9157
2	Rebecca Parker	Rebecca_Parker@comcast.net	652-885-2745	••••••••••3734
3	Laura Murray	Laura_M@gmail.com	364-656-8427	••••••••••5677
4	Linda Hines	LHines@verizon.com	713-226-5883	••••••••••5498

```
#Delete 'name', 'email', 'phone-number', 'credit_card' attributes
df_clean.drop(['name', 'email', 'phone-number', 'credit_card'], axis = 1, inplace=True)
```

- Menghapus kolom Children dan Babies karena sudah ada kolom Kids serta kolom Stays in Weekend Nights dan Stays in Week Nights karena sudah ada kolom Total Stays

```
#Delete 'stays_in_weekend_nights', 'stays_in_week_nights', 'children', & 'babies'
df_clean.drop(['stays_in_weekend_nights', 'stays_in_week_nights', 'children', 'babies'], axis = 1, inplace=True)
```

FEATURE EXTRACTION (1)

Discussion:

- **Total Stays**, dari kolom stays_in_weekend_nights dan stays_in_week_nights, kita bisa mendapatkan total jumlah hari menginap. Terlihat data total stays bernilai 0, hal ini memungkinkan sebab konsumen dapat check in dan checkout dihari yang sama.
- **Total Guest**, dari kolom `adults, children and babies` kita bisa mendapatkan total jumlah orang yang menginap.
- **Kids**, Dari kolom children and babies kita bisa mendapatkan total jumlah anak yang menginap. Karena biasanya hanya ada kategori tamu dewasa dan anak (dibawah 17 tahun) dalam pemesanan kamar hotel.
- **Guest locations**, karena kolom country memiliki data yang sangat besar, maka akan dikategorikan menjadi local & international. Pengkategorian mengacu pada lokasi hotel pada dataset yang berada di negara Portugal untuk penduduk local dan selain itu akan di kategorikan sebagai international.

FEATURE EXTRACTION (2)

Discussion:

- **Arrival columns**, dari kolom ini akan digenerate kolom baru berdasarkan bulan dan hari.
- **Meal Columns**, dari kolom ini didapat data bernilai 'undentified'. Dari source dataset ini menyebutkan bahwa undentified sama dengan SC (no meal package), untuk itu kami akan mengganti nilai undentified menjadi SC.
- **Distribution Channel**, dari kolom ini didapat data bernilai Undefined dan akan di ganti menjadi TA/TO yang merupakan nilai modus.
- **Market Segment**, dari kolom ini didapat data bernilai Undefined dan akan di ganti menjadi online TA yang merupakan modus.

ADDITIONAL FEATURES (1)

Discussion:

- **Customer Satisfaction Rate (Rating):** Fitur ini perlu ditambahkan karena dapat digunakan untuk menilai kinerja pelayanan hotel dan mungkin saja berpengaruh terhadap cancellation rate. Hipotesis dari kami adalah semakin tinggi rating yang diberikan kepada hotel, maka akan semakin rendah cancellation rate.
- **Total Revenue:** Pada Dataset hanya disajikan data mengenai Average Daily Rate (ADR), yang didapatkan dari pendapatan kamar/kamar terisi. Fitur ini belum belum menggambarkan pengeluaran tiap customer per order. Total Revenue dapat digunakan untuk menganalisis efektifivitas pricing yang telah ditetapkan dan mungkin mungkin saja berpengaruh terhadap cancellation rate.
- **Reason For Staying:** Alasan seseorang melakukan pemesanan hotel diantaranya adalah Business Trip atau Holiday. Dengan mengetahui alasan tersebut, kedepannya dapat dibuat Marketing Campaign dengan bekerja sama dengan sebuah perusahaan atau tempat wisata di daerah setempat. Selain itu, kita juga akan mengetahui hubungan antara Reason For Staying dan Cancellation rate di hotel tersebut.

ADDITIONAL FEATURES (2)

Discussion:

- **Reason For Cancelled:** Alasan seseorang melakukan pembatalan reservation hotel diantaranya adalah Change Mind, Find Better Hotel, Weather Issue, dan lain sebagainya. Faktor alasan pembatalan yang dikarenakan alasan pribadi customer dapat dimaklumi, namun jika alasan pembatalan karena find better hotel, diperlukan evaluasi kinerja ataupun fasilitas dari hotel tersebut dibandingkan dengan hotel competitor sekitar.
- **Promotion Apply:** Dengan adanya fitur ini, dapat diketahui efektifitas promosi yang sedang dijalankan, sekaligus mengetahui apakah customer yang order dengan menggunakan promo, kemungkinan cancel nya akan lebih tinggi atau rendah.

CHOSEN FEATURES

Discussion:

Berdasarkan correlation heatmap pada multivariate analysis:

- total_of_special_request
- required_car_parking_spaces
- booking_change
- previous_bookings_not_canceled
- is_repeated guest
- agent
- previous_cancellation
- lead_time_norm
- adr
- total_stay
- total_guest

Berdasarkan hasil analisis kami pada stage sebelumnya:

- hotel
- lead_time
- arrival_date_year
- arrival_date_month
- is_repeated_guest
- previous_cancellations
- previous_bookings_not_canceled
- booking_change
- adr
- assigned_room_type
- deposit_type
- days_in_waiting_list
- customer_type
- total_stays

TERIMA KASIH