LAPORAN HOMEWORK UNSUPERVISED LEARNING

ANGGOTA KELOMPOK

Celestial Randy

Sonia Epifany Sandah

Oky Hariawan

Risca Naquitasia

Mochamad Choiril Iman

Ahmad Reza

Yehezkiel Novianto A.

DATA PRE-PROCESSING (1)

- Handle Missing/Null Value
- Handle Duplicated Value
- Adjust Data Type
- Redefine Numerical & Categorical Data

HANDLE MISSING & NULL VALUE

Insight:

- Dari dataset ini kolom work_province, work_city, sum_yr_1, age, sum_yr_2, work_country, gender memiliki missing value
- Untuk kolom sum_yr_1, age, sum_yr_2, work_country, gender karena memiliki persentasi dibawah 1% maka akan di hapus
- Untuk kolom work_province, work_city walaupun memiliki persentase dibawah 6% karena memiliki jumlah yang besar maka akan di imputasi menggunakan nilai modus

	feature	missing_value	percentage
0	work_province	3248	5.157
1	work_city	2269	3.602
2	sum_yr_1	551	0.875
3	age	420	0.667
4	sum_yr_2	138	0.219
5	work_country	26	0.041
6	gender	3	0.005

ERROR HANDLING : INCORRECT VALUE

```
[ ]  print('Incorrect `last_flight_date` percentage: ', end='')
     print(str(round(df_clean[df_clean.last_flight_date.str.contains('2014/2/29')]['last_flight_date'].count()/len(df_clean), 3)), '%')

     Incorrect `last_flight_date` percentage: 0.007 %
```

Insight:

- Kami menemukan error saat proses Adjust data type sehingga akan dilakukan pengecekan incorrect value

"ParserError: day is out of range for month: 2014/2/29 0:00:00“

- Karena persentase incorrect value kolom last_flight_date hanya 0.007%, maka baris ini akan di hapus

HANDLE DUPLICATED VALUE

Insight:

- Tidak ditemukan adanya duplicated value dalam dataset ini

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

     0
```

ADJUST DATA TYPE

What We Do:

- Mengubah data type kolom `age` dari float menjadi int64
- Mengubah data type in the ffp_date, dirst_fligt_date, load_time, last_flight_date columns dari object menjadi datetime

```
df_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 61437 entries, 0 to 62986
Data columns (total 23 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   member_no          61437 non-null  int64
 1   ffp_date           61437 non-null  datetime64[ns]
 2   first_flight_date  61437 non-null  datetime64[ns]
 3   gender             61437 non-null  object
 4   ffp_tier           61437 non-null  int64
 5   work_city          61437 non-null  object
 6   work_province      61437 non-null  object
 7   work_country       61437 non-null  object
 8   age                61437 non-null  int64
 9   load_time          61437 non-null  datetime64[ns]
 10  flight_count       61437 non-null  int64
 11  bp_sum             61437 non-null  int64
 12  sum_yr_1           61437 non-null  float64
 13  sum_yr_2           61437 non-null  float64
 14  seg_km_sum         61437 non-null  int64
 15  last_flight_date   61437 non-null  datetime64[ns]
 16  last_to_end        61437 non-null  int64
 17  avg_interval       61437 non-null  float64
 18  max_interval       61437 non-null  int64
 19  exchange_count     61437 non-null  int64
 20  avg_discount       61437 non-null  float64
 21  points_sum         61437 non-null  int64
 22  point_notflight    61437 non-null  int64
dtypes: datetime64[ns](4), float64(4), int64(11), object(4)
memory usage: 11.2+ MB
```

REDEFINE NUMERICAL & CATEGORICAL DATA

What We Do:

- Membagi data menjadi kolom numerical dan kolom categorical untuk kemudahan processing data kedepannya

```
#Numeric
numerical = df_clean.loc[:, (df_clean.dtypes == int) | (df_clean.dtypes == float)].columns.tolist()
numerical
```

```
['member_no',
 'ffp_tier',
 'age',
 'flight_count',
 'bp_sum',
 'sum_yr_1',
 'sum_yr_2',
 'seg_km_sum',
 'last_to_end',
 'avg_interval',
 'max_interval',
 'exchange_count',
 'avg_discount',
 'points_sum',
 'point_notflight']
```

```
#Categorical
categorical = df_clean.loc[:, (df_clean.dtypes != int) & (df_clean.dtypes != float)].columns.tolist()
categorical
```

```
['ffp_date',
 'first_flight_date',
 'gender',
 'work_city',
 'work_province',
 'work_country',
 'load_time',
 'last_flight_date']
```

EDA

- Statistic Descriptive
- Univariate Analysis
- Multivariate Analysis
- Conclusion

STATISTIC DESCRIPTIVE (1)

```
#Statistical summary numerical
df.describe()
```

	member_no	ffp_tier	age	flight_count	bp_sum	sum_yr_1	sum_yr_2	seg_km_sum	last_to_end	avg_interval	max_interval	exchange_count	avg_discount	points_sum	point_n
count	62988.000000	62988.000000	62568.000000	62988.000000	62988.000000	62437.000000	62850.000000	62988.000000	62988.000000	62988.000000	62988.000000	62988.000000	62988.000000	62988.0000	6298
mean	31494.500000	4.102162	42.476346	11.839414	10925.081254	5355.376064	5604.026014	17123.878691	176.120102	67.749788	166.033895	0.319775	0.721558	12545.7771	
std	18183.213715	0.373856	9.885915	14.049471	16339.486151	8109.450147	8703.364247	20960.844623	183.822223	77.517866	123.397180	1.136004	0.185427	20507.8167	
min	1.000000	4.000000	6.000000	2.000000	0.000000	0.000000	0.000000	368.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.0000	
25%	15747.750000	4.000000	35.000000	3.000000	2518.000000	1003.000000	780.000000	4747.000000	29.000000	23.370370	79.000000	0.000000	0.611997	2775.0000	
50%	31494.500000	4.000000	41.000000	7.000000	5700.000000	2800.000000	2773.000000	9994.000000	108.000000	44.666667	143.000000	0.000000	0.711856	6328.5000	
75%	47241.250000	4.000000	48.000000	15.000000	12831.000000	6574.000000	6845.750000	21271.250000	268.000000	82.000000	228.000000	0.000000	0.809476	14302.5000	
max	62988.000000	6.000000	110.000000	213.000000	505308.000000	239560.000000	234188.000000	580717.000000	731.000000	728.000000	728.000000	46.000000	1.500000	985572.0000	14

Insight:

- Dataset ini memiliki 15 kolom numerical
- Kolom member_no, fpp_tier, age, avg_discount terlihat memiliki distribusi normal
- Sementara kolom lainnya tampak memiliki distribusi skewed (mean > median)
- Dalam kolom age ada customer yang berusia 110 tahun, dan tampak abnormal sehingga dapat dihapus di proses selanjutnya

STATISTIC DESCRIPTIVE (2)

```
#Statistical summary categorical
df.describe(include='object')
```

	ffp_date	first_flight_date	gender	work_city	work_province	work_country	load_time	last_flight_date
count	62988	62988	62985	60719	59740	62962	62988	62988
unique	3068	3406	2	3234	1165	118	1	731
top	1/13/2011	2/16/2013	Male	guangzhou	guangdong	CN	3/31/2014	3/31/2014
freq	184	96	48134	9386	17509	57748	62988	959

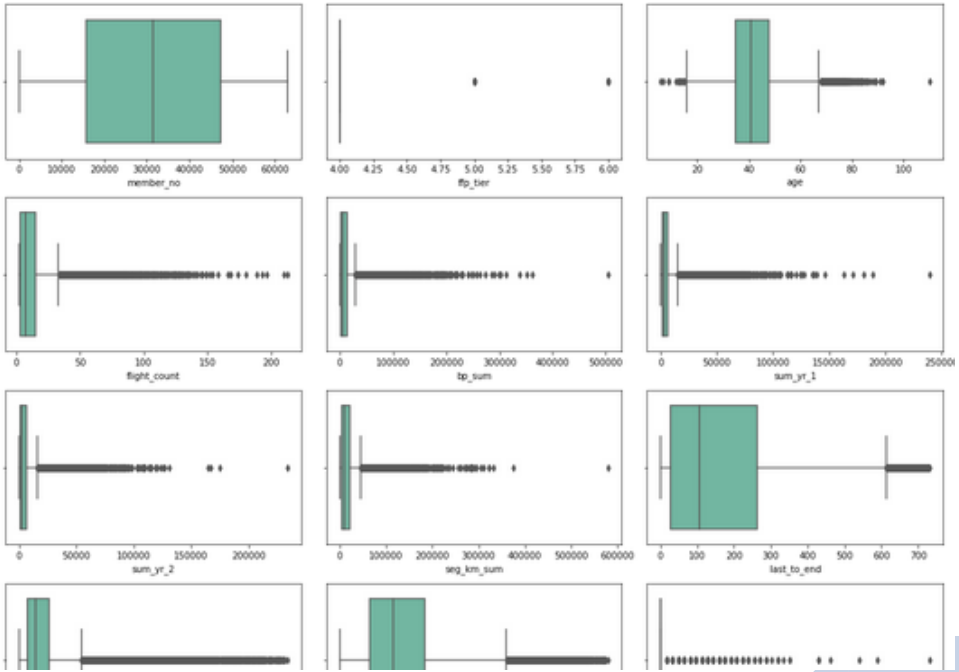
Insight:

- Dataset ini memiliki 8 kolom categorical
- Tampak kebanyakan kolom memiliki nilai unique yang tinggi kecuali gender & load_time

UNIVARIATE ANALYSIS (BOXPLOT)

Insight:

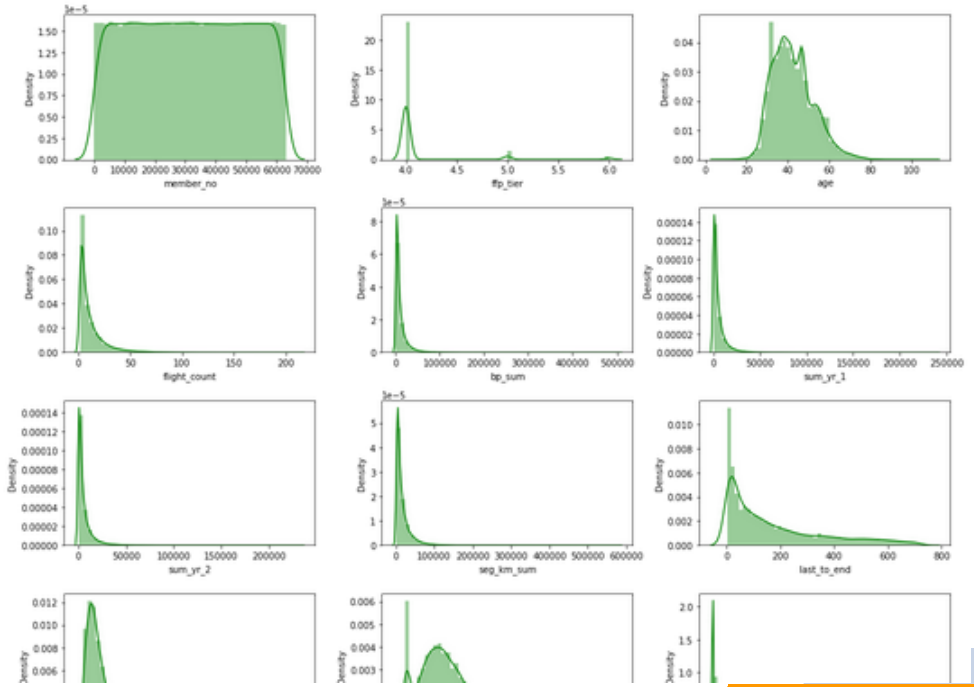
- Dari Boxplot dapat terlihat jika hampir seluruh feature memiliki outlier kecuali `member_no` & `ffp_tier`



UNIVARIATE ANALYSIS (DISTRIBUTION PLOT)

Insight:

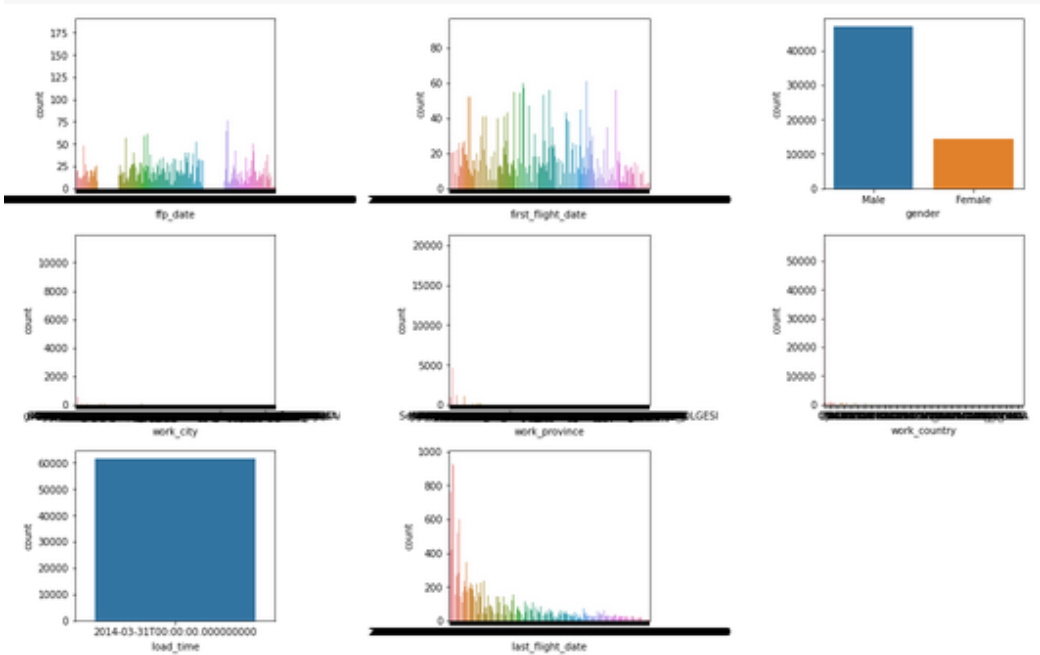
- Dari Distribution Plot dapat terlihat jika hampir seluruh feature memiliki distribusi positif skewed kecuali member_no & avg_discount



UNIVARIATE ANALYSIS (COUNT PLOT)

Insight:

- Kebanyakan customer adalah Laki-laki
- Tampak kebanyakan feature memiliki nilai unique yang besar
- Untuk load_time hanya memiliki 1 value 2014-03-31 yang dimana merupakan tangal data diambil



MULTIVARIATE ANALYSIS (HEAT MAP)

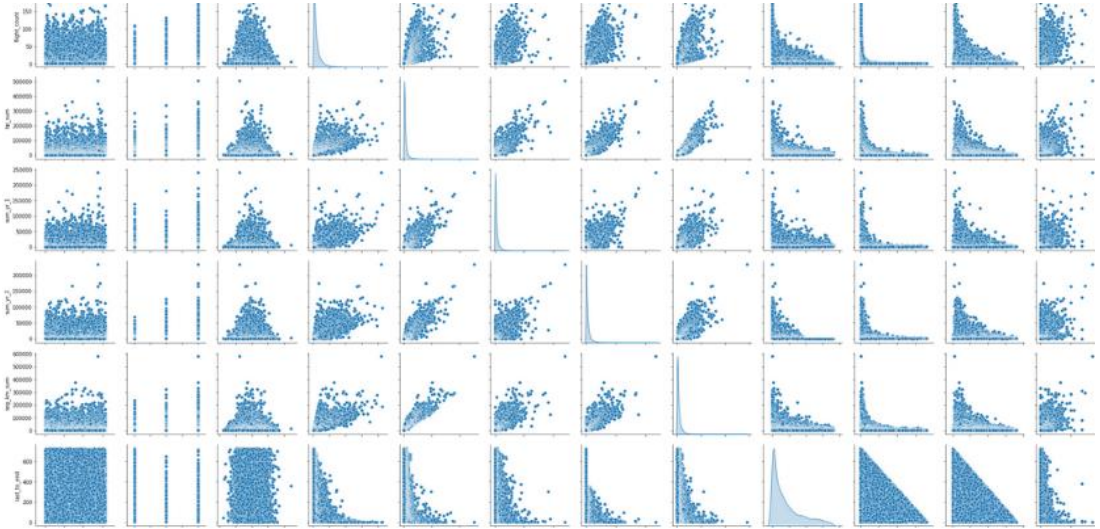
Insight:

- Dari Heatmap korelasi dapat terlihat jika banyak feture yang memiliki korelasi positif yang kuat > 0.7

MULTIVARIATE ANALYSIS (PAIR PLOT)

Insight:

- Dalam Pair Plot terdapat beberapa feature yang memiliki korelasi linear seperti last_to_end dengan avg_interval



EDA CONCLUSSION

Discussion:

- Hampir seluruh feature numerical memiliki outlier kecuali member_no & ffp_tier
- Hampir seluruh feature numerical memiliki distribusi positif skewed kecuali member_no & avg_discount
- Kebanyakan customer adalah Laki-laki
- Tampak kebanyakan feature catogorical memiliki nilai unique yang besar
- Untuk load_time hanya memiliki 1 value 2014-03-31 yang dimana merupakan tangal data diambil
- Banyak feature yang memiliki korelasi positif yang tinggi terhadap feature lainnya > 0.7

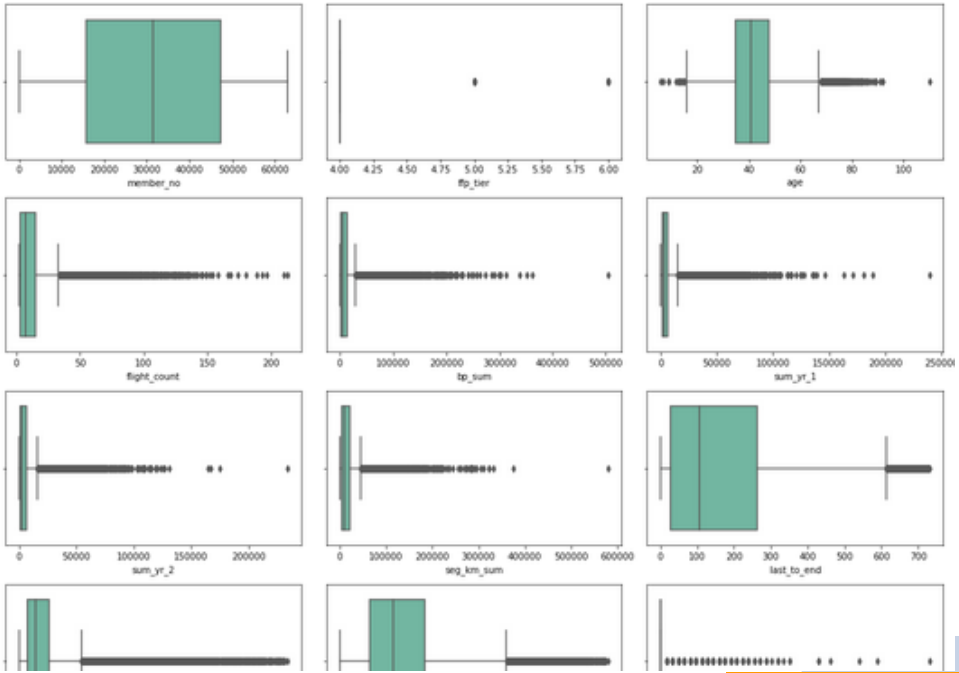
DATA PRE-PROCESSING (2)

- Handle Outliers
- Feature Selection
- Feature Engineering

HANDLE OUTLIERS (1)

Insight:

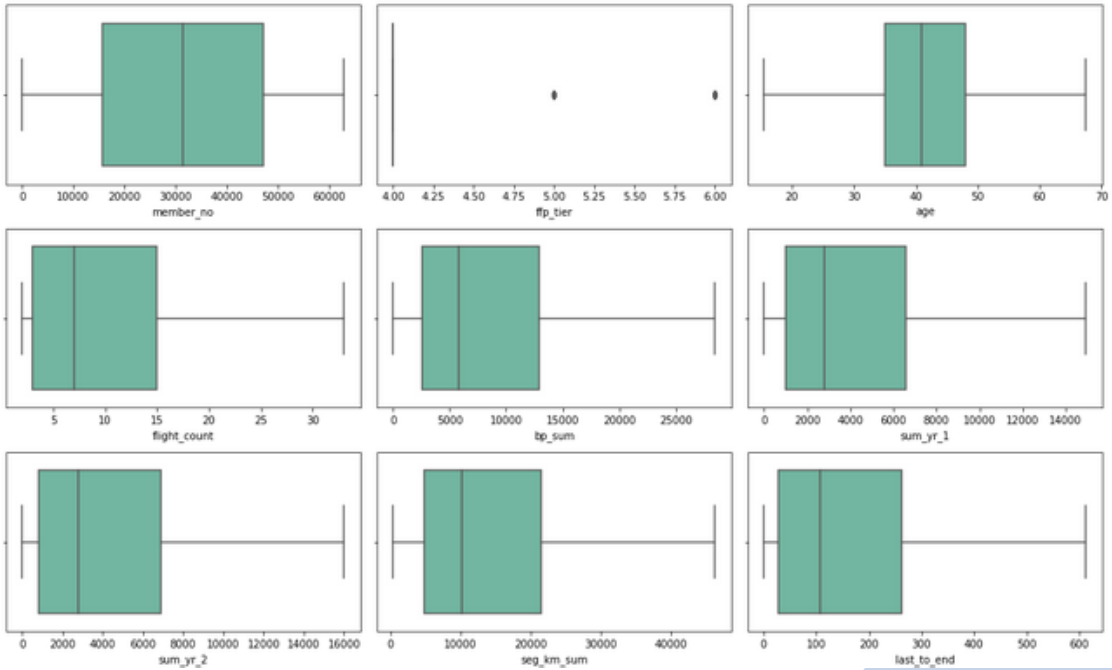
- Beberapa kolom masih ditemukan outliers yaitu age, flight_count, bp_sum, sum_yr_1, sum_yr_2, seg_km_sum, last_to_end, avg_interval, max_interval, exchange_count, avg_discount, point_sum, point_not_flight
- Kami menghapus outliers dengan menerapkan konsep IQR



HANDLE OUTLIERS (2)

Insight:

- Untuk Feature exchange_count karena data hanya terpusat pada satu value (High limit dan limit sama) maka outlier tidak dihilangkan untuk memberikan variasi data
- Untuk feature point_not_flight outlier tidak sepenuhnya dihilangkan akan tetapi diberikan batas maksimal



FEATURE ENGINEERING & FEATURE SELECTION

What We Do:

- Menambah feature ffp_time untuk mengetahui berapa bulan customer join program saat data diambil
- Menghapus feature yang tidak relevan: ffp_time, sum_yr_2,seg_km_sum, avg_discount,gender, work_city, work_province, work_country,last_flight_date, first_flight_date,avg_interval, max_interval,member_no, ffp_tier, age, bp_sum, exchange_count,points_sum,point_not_flight
- Menghapus kolom yang redundan: load_time, ffp_date

MODELING

- Standardization
- Modeling
- Business Recommendation

STANDARDIZATION

What We Do:

- Melakukan standarisasi pada data dengan bantuan StandardScaler()

```
] df_clean.skew()

flight_count    1.212919
sum_yr_1        1.121951
last_to_end     1.114797
dtype: float64
```

```
] from sklearn.cluster import KMeans
  from sklearn.decomposition import PCA
  from sklearn.metrics import silhouette_score
  from sklearn.cluster import AgglomerativeClustering
  from sklearn.preprocessing import StandardScaler, MinMaxScaler
```

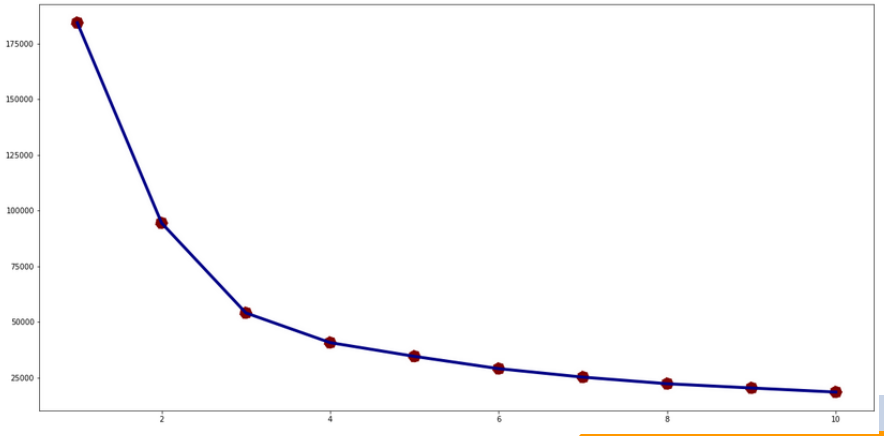
```
] sc_data = StandardScaler()
  data_std = sc_data.fit_transform(df_clean.astype(float))
```

MODELING (1)

What We Do:

Modeling dilakukan dengan clustering menggunakan k-means dimana berdasarkan elbow method digunakan jumlah cluster yaitu sebanyak 3 cluster.

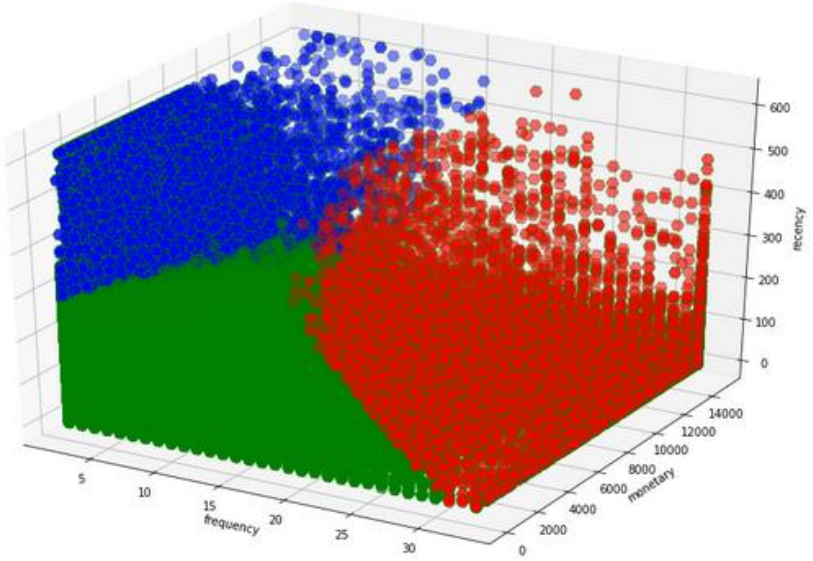
```
from sklearn.cluster import KMeans
inertia = []

for i in range(1, 11):
    kmeans = KMeans(n_clusters=i, random_state=0)
    kmeans.fit(data_std)
    inertia.append(kmeans.inertia_)



plt.figure(figsize=(20, 10))
# plt.plot(inertia)

sns.lineplot(x=range(1, 11), y=inertia, color='#000087', linewidth = 4)
sns.scatterplot(x=range(1, 11), y=inertia, s=300, color='#800000',  linestyle='--')
```



MODELING (2)

What We Do:

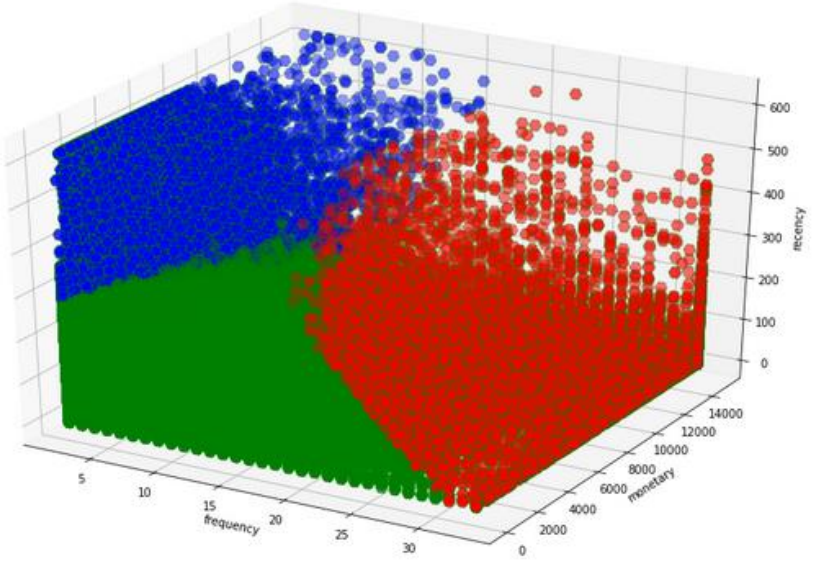
- Feature yang mempengaruhi dalam segmentasi pelanggan bedasarkan Recency, Frequency, Monetary (RFM) yaitu jarak waktu penerbangan terakhir ke pesanan penerbangan paling akhir, banyaknya penerbangan yang dilakukan konsumen, dan penghasilan dari pembelian tiket pesawat.



MODELING (2)

Insight:

- Cluster 0 merupakan pelanggan yang memiliki frekuensi terbang rendah (rata-rata 4x penerbangan), memberikan rata - rata revenue sedang (rata- rata 3134), dan memiliki selisih waktu paling lama sejak waktu penerbangan terakhir mereka terhitung saat data ini diambil (rata-rata 452 hari)
- Cluster 1 merupakan pelanggan yang memiliki frekuensi terbang tinggi (rata-rata 25x penerbangan), memberikan rata - rata revenue tinggi (rata- rata 11365), dan memiliki selisih waktu paling sebentar sejak waktu penerbangan terakhir mereka terhitung saat data ini diambil (rata-rata 58 hari)
- Cluster 2 merupakan pelanggan yang memiliki frekuensi terbang sedang (rata-rata 8x penerbangan), memberikan rata - rata revenue rendah (rata- rata 2238), dan memiliki selisih cukup lama sejak waktu penerbangan terakhir mereka terhitung saat data ini diambil (rata-rata 100 hari).



BUSINESS RECOMMENDATION

Discussion:

- Fokus untuk menggencarkan strategi marketing kepada pelanggan yang tergolong cluster 1, untuk meningkatkan jumlah penerbangan mereka yang masih rendah dan sudah lama tidak melakukan penerbangan. Strategi marketing dapat dilakukan dengan memberikan promo discount atau partnership dengan travel agent untuk memberikan promo paket liburan menarik di luar kota, agar konsumen tersebut tertarik untuk terbang kembali menggunakan maskapai ini.
- Fokus untuk mempertahankan pelanggan yang tergolong cluster 2 dengan menawarkan pembuatan membership premium maskapai.

TERIMA KASIH