Dataset :

flight

Unsupervised Learning

Created by : Deep Learning 4.0

QuickDA

EDA using QuickDA library

Including Overview, Variables, Interactions, Correlations, Missing Values, and Samples

Overview

Overview

Overview | Alerts 60 | Reproduction

Dataset statistics

Number of variables	23
Number of observations	62988
Missing cells	6655
Missing cells (%)	0.5%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	11.1 MiB
Average record size in memory	184.0 B

Variable types

Numeric	14
Categorical	9

Overview

Overview | Alerts 60 | Reproduction

Alerts

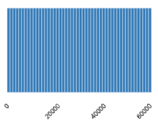
LOAD_TIME has constant value "3/31/2014"	Constant
FFP_DATE has a high cardinality: 3068 distinct values	High cardinality
FIRST_FLIGHT_DATE has a high cardinality: 3406 distinct values	High cardinality
WORK_CITY has a high cardinality: 3234 distinct values	High cardinality
WORK_PROVINCE has a high cardinality: 1165 distinct values	High cardinality
WORK_COUNTRY has a high cardinality: 118 distinct values	High cardinality
LAST_FLIGHT_DATE has a high cardinality: 731 distinct values	High cardinality
FLIGHT_COUNT is highly correlated with BP_SUM and 5 other fields	High correlation
BP_SUM is highly correlated with FLIGHT_COUNT and 5 other fields	High correlation
SUM_YR_1 is highly correlated with FLIGHT_COUNT and 3 other fields	High correlation
SUM_YR_2 is highly correlated with FLIGHT_COUNT and 4 other fields	High correlation
SEG_KM_SUM is highly correlated with FLIGHT_COUNT and 5 other fields	High correlation
LAST_TO_END is highly correlated with FLIGHT_COUNT and 4 other fields	High correlation
AVG_INTERVAL is highly correlated with MAX_INTERVAL	High correlation
MAX_INTERVAL is highly correlated with AVG_INTERVAL	High correlation
EXCHANGE_COUNT is highly correlated with Point_NotFlight	High correlation
Points_Sum is highly correlated with FLIGHT_COUNT and 5 other fields	High correlation
Point_NotFlight is highly correlated with EXCHANGE_COUNT	High correlation
FFP_TIER is highly correlated with FLIGHT_COUNT and 4 other fields	High correlation
FLIGHT_COUNT is highly correlated with FFP_TIER and 6 other fields	High correlation
BP_SUM is highly correlated with FFP_TIER and 6 other fields	High correlation
SUM_YR_1 is highly correlated with FLIGHT_COUNT and 4 other fields	High correlation
SUM_YR_2 is highly correlated with FFP_TIER and 5 other fields	High correlation
SEG_KM_SUM is highly correlated with FFP_TIER and 6 other fields	High correlation
AVG_INTERVAL is highly correlated with MAX_INTERVAL	High correlation
MAX_INTERVAL is highly correlated with AVG_INTERVAL	High correlation
EXCHANGE_COUNT is highly correlated with FLIGHT_COUNT and 3 other fields	High correlation
Points_Sum is highly correlated with FFP_TIER and 6 other fields	High correlation
FLIGHT_COUNT is highly correlated with BP_SUM and 4 other fields	High correlation
BP_SUM is highly correlated with FLIGHT_COUNT and 4 other fields	High correlation
SUM_YR_1 is highly correlated with FLIGHT_COUNT and 3 other fields	High correlation
SUM_YR_2 is highly correlated with FLIGHT_COUNT and 4 other fields	High correlation
SEG_KM_SUM is highly correlated with FLIGHT_COUNT and 4 other fields	High correlation
LAST_TO_END is highly correlated with SUM_YR_2	High correlation
AVG_INTERVAL is highly correlated with MAX_INTERVAL	High correlation
MAX_INTERVAL is highly correlated with AVG_INTERVAL	High correlation
EXCHANGE_COUNT is highly correlated with Point_NotFlight	High correlation
Points_Sum is highly correlated with FLIGHT_COUNT and 4 other fields	High correlation
Point_NotFlight is highly correlated with EXCHANGE_COUNT	High correlation
FFP_TIER is highly correlated with LOAD_TIME	High correlation
LOAD_TIME is highly correlated with FFP_TIER and 1 other fields	High correlation
GENDER is highly correlated with LOAD_TIME	High correlation
FFP_TIER is highly correlated with FLIGHT_COUNT and 4 other fields	High correlation
FLIGHT_COUNT is highly correlated with FFP_TIER and 6 other fields	High correlation
BP_SUM is highly correlated with FFP_TIER and 6 other fields	High correlation
SUM_YR_1 is highly correlated with FFP_TIER and 6 other fields	High correlation
SUM_YR_2 is highly correlated with FFP_TIER and 6 other fields	High correlation
SEG_KM_SUM is highly correlated with FLIGHT_COUNT and 5 other fields	High correlation
AVG_INTERVAL is highly correlated with MAX_INTERVAL	High correlation
MAX_INTERVAL is highly correlated with AVG_INTERVAL	High correlation
EXCHANGE_COUNT is highly correlated with FLIGHT_COUNT and 5 other fields	High correlation
Points_Sum is highly correlated with FFP_TIER and 6 other fields	High correlation
WORK_CITY has 2269 (3.6%) missing values	Missing
WORK_PROVINCE has 3248 (5.2%) missing values	Missing
MEMBER_NO is uniformly distributed	Uniform
MEMBER_NO has unique values	Unique
SUM_YR_1 has 9915 (15.7%) zeros	Zeros
SUM_YR_2 has 11812 (18.8%) zeros	Zeros
EXCHANGE_COUNT has 54254 (86.1%) zeros	Zeros
Point_NotFlight has 42480 (67.4%) zeros	Zeros

Variables

MEMBER_NO
Real number ($\mathbb{R}_{\geq 0}$)

UNIFORM
UNIQUE

Distinct	62988	Minimum	1
Distinct (%)	100.0%	Maximum	62988
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	31494.5	Memory size	492.2 KiB



Toggle details

FFP_DATE
Categorical

HIGH CARDINALITY

Distinct	3068
Distinct (%)	4.9%
Missing	0
Missing (%)	0.0%
Memory size	492.2 KiB

1/13/2011	184
1/1/2013	165
3/1/2013	100
11/17/2010	99
1/14/2011	95
Other values (3063)	62345

Toggle details

FIRST_FLIGHT_DATE
Categorical

HIGH CARDINALITY

Distinct	3406
Distinct (%)	5.4%
Missing	0
Missing (%)	0.0%
Memory size	492.2 KiB

2/16/2013	96
9/30/2012	85
2/15/2013	84
9/9/2005	78
4/21/2012	70
Other values (3401)	62575

Toggle details

GENDER
Categorical

HIGH CORRELATION

Distinct	2
Distinct (%)	< 0.1%
Missing	3
Missing (%)	< 0.1%
Memory size	492.2 KiB

Male	48134
Female	14851

Toggle details

FFP_TIER
Categorical

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION

Distinct	3
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	492.2 KiB

4	58066
5	3409
6	1513

Toggle details

WORK_CITY
Categorical

HIGH CARDINALITY
MISSING

Distinct	3234
Distinct (%)	5.3%
Missing	2269
Missing (%)	3.6%

guangzhou	9386
beijing	7845
shanghai	5001
shenzhen	3605
dalian	1979

WORK_PROVINCE
Categorical

HIGH CARDINALITY
MISSING

Distinct	1165
Distinct (%)	2.0%
Missing	3248
Missing (%)	5.2%
Memory size	492.2 KiB

guangdong	17509
beijing	8014
shanghai	4998
liaoning	4182
xinjiang	2512
Other values (1160)	22525

Toggle details

WORK_COUNTRY
Categorical

HIGH CARDINALITY

Distinct	118
Distinct (%)	0.2%
Missing	26
Missing (%)	< 0.1%
Memory size	492.2 KiB

CN	57748
HK	991
JP	875
KR	790
US	575
Other values (113)	1983

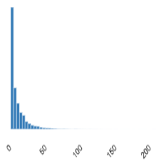
Toggle details

Variables

FLIGHT_COUNT
Real number ($\mathbb{R}_{\geq 0}$)

HIGH_CORRELATION
HIGH_CORRELATION
HIGH_CORRELATION
HIGH_CORRELATION

Distinct	153	Minimum	2
Distinct (%)	0.2%	Maximum	213
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	11.83941386	Memory size	492.2 KiB

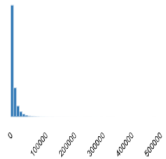


Toggle details

BP_SUM
Real number ($\mathbb{R}_{\geq 0}$)

HIGH_CORRELATION
HIGH_CORRELATION
HIGH_CORRELATION
HIGH_CORRELATION

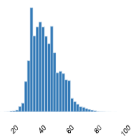
Distinct	23449	Minimum	0
Distinct (%)	37.2%	Maximum	505308
Missing	0	Zeros	565
Missing (%)	0.0%	Zeros (%)	0.9%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	10925.08125	Memory size	492.2 KiB



Toggle details

AGE
Real number ($\mathbb{R}_{\geq 0}$)

Distinct	84	Minimum	6
Distinct (%)	0.1%	Maximum	110
Missing	420	Zeros	0
Missing (%)	0.7%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	42.47634574	Memory size	492.2 KiB



Toggle details

~~LOAD_TIME~~
~~Categorical~~

CONSTANT
HIGH_CORRELATION
REJECTED

Distinct	1
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	492.2 KiB

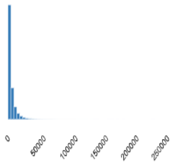
Toggle details

Variables

SUM_YR_1
Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
ZEROS

Distinct	15828	Minimum	0
Distinct (%)	25.4%	Maximum	239560
Missing	551	Zeros	9915
Missing (%)	0.9%	Zeros (%)	15.7%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	5355.376064	Memory size	492.2 KiB

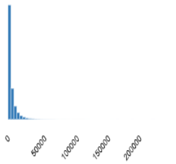


Toggle details

SUM_YR_2
Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
ZEROS

Distinct	16767	Minimum	0
Distinct (%)	26.7%	Maximum	234188
Missing	138	Zeros	11812
Missing (%)	0.2%	Zeros (%)	18.8%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	5604.026014	Memory size	492.2 KiB

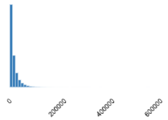


Toggle details

SEG_KM_SUM
Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION

Distinct	29081	Minimum	368
Distinct (%)	46.2%	Maximum	580717
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	17123.87869	Memory size	492.2 KiB



Toggle details

LAST_FLIGHT_DATE
Categorical

HIGH CARDINALITY

Distinct	731
Distinct (%)	1.2%
Missing	0
Missing (%)	0.0%
Memory size	492.2 KiB

3/31/2014	959
3/30/2014	933
3/28/2014	924
3/29/2014	779
3/27/2014	767
Other values (726)	58626

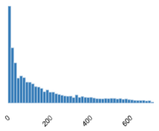
Toggle details

Variables

LAST_TO_END
Real number ($\mathbb{R}_{\geq 0}$)

HIGH_CORRELATION
HIGH_CORRELATION

Distinct	731	Minimum	1
Distinct (%)	1.2%	Maximum	731
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	176.1201022	Memory size	492.2 KiB

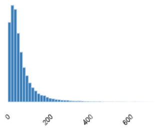


Toggle details

AVG_INTERVAL
Real number ($\mathbb{R}_{\geq 0}$)

HIGH_CORRELATION
HIGH_CORRELATION
HIGH_CORRELATION
HIGH_CORRELATION

Distinct	10706	Minimum	0
Distinct (%)	17.0%	Maximum	728
Missing	0	Zeros	421
Missing (%)	0.0%	Zeros (%)	0.7%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	67.74978791	Memory size	492.2 KiB

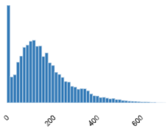


Toggle details

MAX_INTERVAL
Real number ($\mathbb{R}_{\geq 0}$)

HIGH_CORRELATION
HIGH_CORRELATION
HIGH_CORRELATION
HIGH_CORRELATION

Distinct	706	Minimum	0
Distinct (%)	1.1%	Maximum	728
Missing	0	Zeros	421
Missing (%)	0.0%	Zeros (%)	0.7%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	166.0338953	Memory size	492.2 KiB

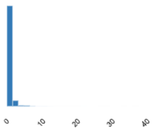


Toggle details

EXCHANGE_COUNT
Real number ($\mathbb{R}_{\geq 0}$)

HIGH_CORRELATION
HIGH_CORRELATION
HIGH_CORRELATION
HIGH_CORRELATION
ZEROS

Distinct	28	Minimum	0
Distinct (%)	< 0.1%	Maximum	46
Missing	0	Zeros	54254
Missing (%)	0.0%	Zeros (%)	86.1%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	0.3197751953	Memory size	492.2 KiB

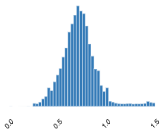


Toggle details

Variables

avg_discount
Real number ($\mathbb{R}_{\geq 0}$)

Distinct	54179	Minimum	0
Distinct (%)	86.0%	Maximum	1.5
Missing	0	Zeros	8
Missing (%)	0.0%	Zeros (%)	< 0.1%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	0.7215577706	Memory size	492.2 KiB

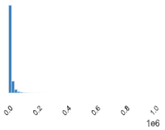


Toggle details

Points_Sum
Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION

Distinct	25062	Minimum	0
Distinct (%)	39.8%	Maximum	985572
Missing	0	Zeros	423
Missing (%)	0.0%	Zeros (%)	0.7%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	12545.7771	Memory size	492.2 KiB

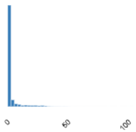


Toggle details

Point_NotFlight
Real number ($\mathbb{R}_{\geq 0}$)

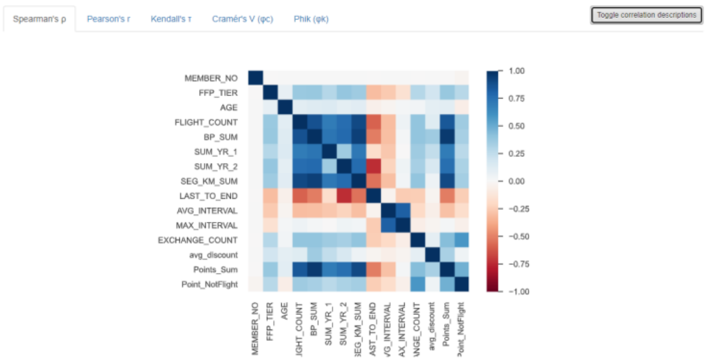
HIGH CORRELATION
HIGH CORRELATION
ZEROS

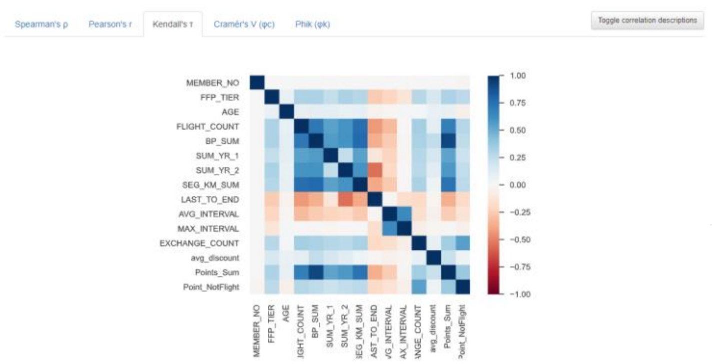
Distinct	99	Minimum	0
Distinct (%)	0.2%	Maximum	140
Missing	0	Zeros	42480
Missing (%)	0.0%	Zeros (%)	67.4%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	2.728154569	Memory size	492.2 KiB

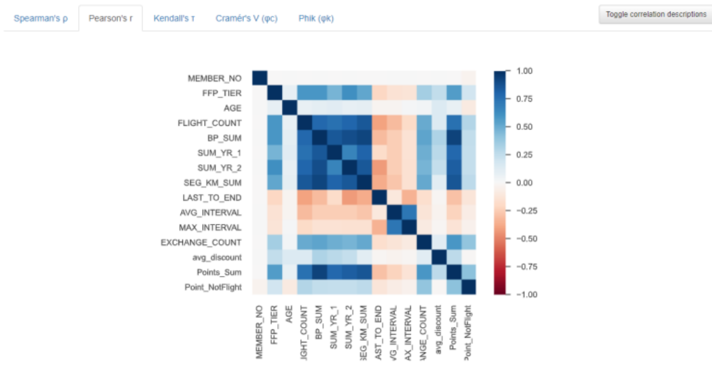


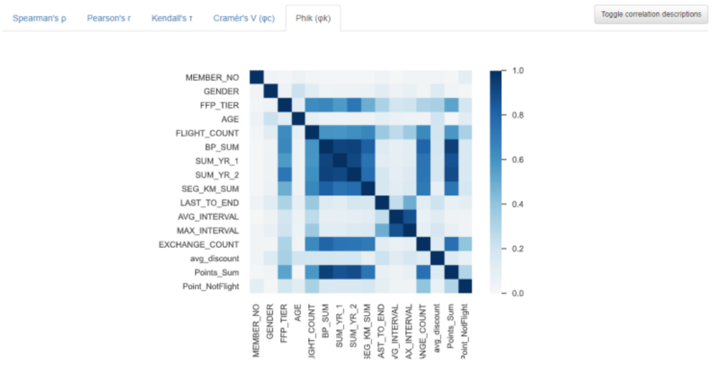
Toggle details

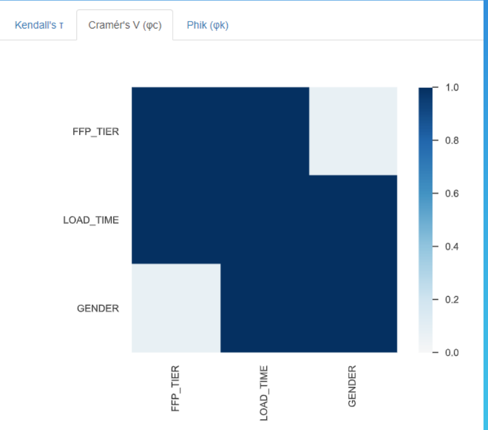
Correlations



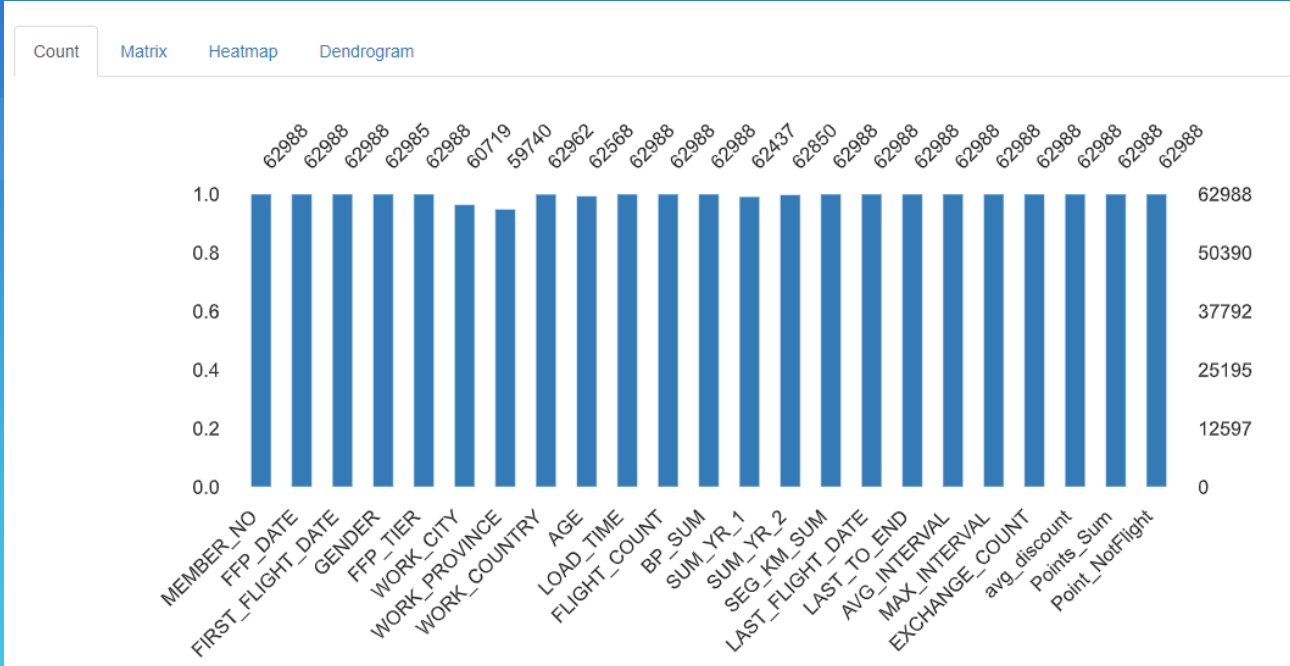








Missing Values



A simple visualization of nullity by column.

Sample

First rows

	MEMBER_NO	FFP_DATE	FIRST_FLIGHT_DATE	GENDER	FFP_TIER	WORK_CITY	WORK_PROVINCE	WORK_COUNTRY	AGE	LOAD_TIME	FLIGHT_COUNT
0	54993	11/2/2006	12/24/2008	Male	6	.	beijing	CN	31.0	3/31/2014	210
1	28065	2/19/2007	8/3/2007	Male	6	NaN	beijing	CN	42.0	3/31/2014	140
2	55106	2/1/2007	8/30/2007	Male	6	.	beijing	CN	40.0	3/31/2014	135
3	21189	8/22/2008	8/23/2008	Male	5	Los Angeles	CA	US	64.0	3/31/2014	23
4	39546	4/10/2009	4/15/2009	Male	6	guiyang	guizhou	CN	48.0	3/31/2014	152
5	56972	2/10/2008	9/29/2009	Male	6	guangzhou	guangdong	CN	64.0	3/31/2014	92
6	44924	3/22/2006	3/29/2006	Male	6	wulumuqishi	xinjiang	CN	46.0	3/31/2014	101
7	22631	4/9/2010	4/9/2010	Female	6	wenzhoushi	zhejiang	CN	50.0	3/31/2014	73
8	32197	6/7/2011	7/1/2011	Male	5	DRANCY	NaN	FR	50.0	3/31/2014	56
9	31645	7/5/2010	7/5/2010	Female	6	wenzhou	zhejiang	CN	43.0	3/31/2014	64

	BP_SUM	SUM_YR_1	SUM_YR_2	SEG_KM_SUM	LAST_FLIGHT_DATE	LAST_TO_END	AVG_INTERVAL	MAX_INTERVAL	EXCHANGE_COUNT	avg_discount	Points_Sum	Point_NotFlight
0	505308	239560.0	234188.0	580717	3/31/2014	1	3.483254	18	34	0.961639	619760	50
1	362480	171483.0	167434.0	293678	3/25/2014	7	5.194245	17	29	1.252314	415768	33
2	351159	163618.0	164982.0	283712	3/21/2014	11	5.298507	18	20	1.254676	406361	26
3	337314	116350.0	125500.0	281336	12/26/2013	97	27.863636	73	11	1.090870	372204	12
4	273844	124560.0	130702.0	309928	3/27/2014	5	4.788079	47	27	0.970658	338813	39
5	313338	112364.0	76946.0	294585	1/13/2014	79	7.043956	52	10	0.967692	343121	15
6	248864	120500.0	114469.0	287042	3/31/2014	1	7.190000	28	20	0.965347	298873	29
7	301864	82440.0	114971.0	287230	3/29/2014	3	10.111111	45	7	0.962070	351198	14
8	262958	72596.0	87401.0	321489	3/26/2014	6	13.054545	94	5	0.828478	295158	7
9	204855	85258.0	60267.0	375074	3/17/2014	15	11.333333	73	13	0.708010	251907	16

Sample

Last rows

	MEMBER_NO	FFP_DATE	FIRST_FLIGHT_DATE	GENDER	FFP_TIER	WORK_CITY	WORK_PROVINCE	WORK_COUNTRY	AGE	LOAD_TIME	FLIGHT_COUNT	BP_SUM	SUM_YR_1	SUM_YR_2	SEG_KM_SUM	LAST_FLIGHT_DATE	LAST_TO_END	AVG_INTERVAL	MAX_INTERVAL	EXCHANGE_COUNT	avg_discount	Points_Sum	Point_NotFlight
62978	22761	4/14/2011	4/14/2011	Male	4	shantou	guangdongsheng	CN	48.0	3/31/2014	2	0	0.0	370.0	760	6/24/2013	282	0.0	0	0	0.28	0	0
62979	34330	3/16/2013	3/17/2013	Male	4	wulumuqi	xinjiang	CN	41.0	3/31/2014	2	0	NaN	0.0	746	3/19/2013	379	2.0	2	0	0.25	0	0
62980	1761	8/7/2012	9/9/2012	Female	4	shenzhen	guangdong	CN	29.0	3/31/2014	2	0	0.0	0.0	6138	9/21/2012	558	12.0	12	0	0.00	0	0
62981	15206	12/2/2011	12/2/2011	Female	4	guangzhou	guangdong	CN	42.0	3/31/2014	2	0	0.0	0.0	2158	10/6/2013	178	3.0	3	0	0.00	0	0
62982	16415	1/20/2013	1/20/2013	Female	4	beijing	.	CN	35.0	3/31/2014	2	0	0.0	0.0	3848	1/20/2013	437	0.0	0	0	0.00	0	0
62983	18375	5/20/2011	6/5/2013	Female	4	guangzhou	guangdong	CN	25.0	3/31/2014	2	0	0.0	0.0	1134	6/9/2013	297	4.0	4	1	0.00	12318	22
62984	36041	3/8/2010	9/14/2013	Male	4	foshan	guangdong	CN	38.0	3/31/2014	4	0	0.0	0.0	8016	1/3/2014	89	37.0	60	14	0.00	106972	43
62985	45690	3/30/2006	12/2/2006	Female	4	guangzhou	guangdong	CN	43.0	3/31/2014	2	0	0.0	0.0	2594	3/3/2014	29	166.0	166	0	0.00	0	0
62986	61027	2/6/2013	2/14/2013	Female	4	guangzhou	guangdong	CN	36.0	3/31/2014	2	0	0.0	0.0	3934	2/26/2013	400	12.0	12	0	0.00	0	0
62987	61340	2/17/2013	2/17/2013	Female	4	shanghai	.	CN	29.0	3/31/2014	2	0	NaN	0.0	4222	2/23/2013	403	6.0	6	0	0.00	0	0

Berdasarkan EDA, terdapat 5 fitur yang akan diubah dtypenya, yakni:

1. 'FFP_DATE', 'FIRST_FLIGHT_DATE', 'LAST_FLIGHT_DATE', 'LOAD_TIME' menjadi datetime64[ns], karena value pada fitur tersebut mendeskripsikan waktu.

2. 'FFP_TIER' menjadi object, karena hanya terdapat 3 unique value.

```
pd.set_option('display.max_rows', None)
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 62988 entries, 0 to 62987
Data columns (total 23 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   MEMBER_NO          62988 non-null  int64
 1   FFP_DATE           62988 non-null  object
 2   FIRST_FLIGHT_DATE  62988 non-null  object
 3   GENDER             62985 non-null  object
 4   FFP_TIER           62988 non-null  int64
 5   WORK_CITY          60719 non-null  object
 6   WORK_PROVINCE      59740 non-null  object
 7   WORK_COUNTRY       62962 non-null  object
 8   AGE                62568 non-null  float64
 9   LOAD_TIME          62988 non-null  object
 10  FLIGHT_COUNT       62988 non-null  int64
 11  BP_SUM             62988 non-null  int64
 12  SUM_YR_1           62437 non-null  float64
 13  SUM_YR_2           62850 non-null  float64
 14  SEG_KM_SUM         62988 non-null  int64
 15  LAST_FLIGHT_DATE   62988 non-null  object
 16  LAST_TO_END        62988 non-null  int64
 17  AVG_INTERVAL       62988 non-null  float64
 18  MAX_INTERVAL       62988 non-null  int64
 19  EXCHANGE_COUNT     62988 non-null  int64
 20  avg_discount       62988 non-null  float64
 21  Points_Sum         62988 non-null  int64
 22  Point_NotFlight    62988 non-null  int64
dtypes: float64(5), int64(10), object(8)
memory usage: 11.1+ MB
```

```
df['FFP_DATE'] = pd.to_datetime(df['FFP_DATE'])
df['FIRST_FLIGHT_DATE'] = pd.to_datetime(df['FIRST_FLIGHT_DATE'])
df['LAST_FLIGHT_DATE'] = pd.to_datetime(df['LAST_FLIGHT_DATE'], errors='coerce')
df['LOAD_TIME'] = pd.to_datetime(df['LOAD_TIME'])

# mengubah data menjadi str
df['FFP_TIER'] = df['FFP_TIER'].astype(str)
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 62988 entries, 0 to 62987
Data columns (total 23 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   MEMBER_NO          62988 non-null  int64
 1   FFP_DATE           62988 non-null  datetime64[ns]
 2   FIRST_FLIGHT_DATE  62988 non-null  datetime64[ns]
 3   GENDER             62985 non-null  object
 4   FFP_TIER           62988 non-null  object
 5   WORK_CITY          60719 non-null  object
 6   WORK_PROVINCE      59740 non-null  object
 7   WORK_COUNTRY       62962 non-null  object
 8   AGE                62568 non-null  float64
 9   LOAD_TIME          62988 non-null  datetime64[ns]
 10  FLIGHT_COUNT       62988 non-null  int64
 11  BP_SUM             62988 non-null  int64
 12  SUM_YR_1           62437 non-null  float64
 13  SUM_YR_2           62850 non-null  float64
 14  SEG_KM_SUM         62988 non-null  int64
 15  LAST_FLIGHT_DATE   62567 non-null  datetime64[ns]
 16  LAST_TO_END        62988 non-null  int64
 17  AVG_INTERVAL       62988 non-null  float64
 18  MAX_INTERVAL       62988 non-null  int64
 19  EXCHANGE_COUNT     62988 non-null  int64
 20  avg_discount       62988 non-null  float64
 21  Points_Sum         62988 non-null  int64
 22  Point_NotFlight    62988 non-null  int64
dtypes: datetime64[ns](4), float64(5), int64(9), object(5)
memory usage: 11.1+ MB
```

Categorical & Numerical

Categorical

```
df[cat].describe()
```

	gender	ffp_tier	work_city	work_province	work_country
count	62985	62988	60719	59740	62962
unique	2	3	3234	1165	118
top	Male	4	guangzhou	guangdong	CN
freq	48134	58066	9386	17509	57748

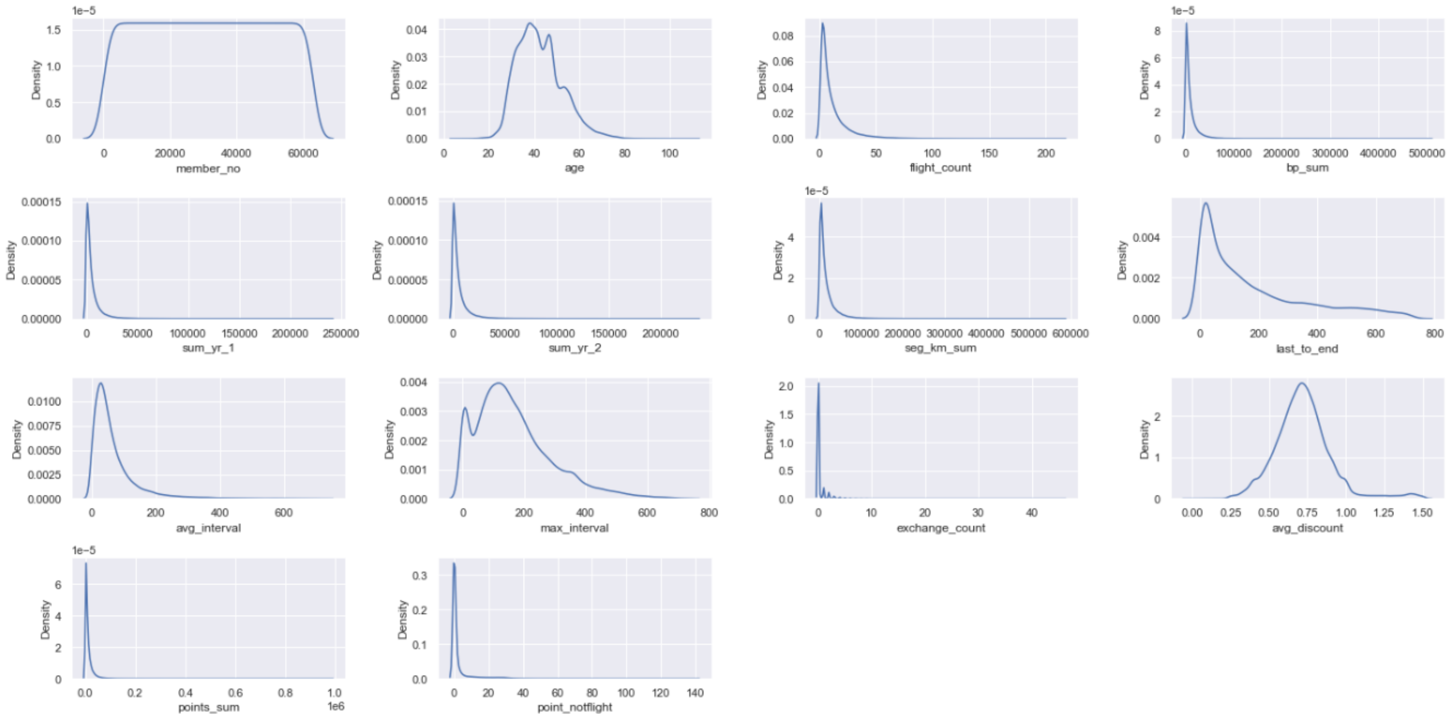
Numerical

```
df[num].describe()
```

	member_no	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
count	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	62988.000000
mean	31494.500000	42.476346	11.839414	10925.081254	5355.376064	5604.026014	17123.878691	176.120102	67.749788	166.033895	0.319775	0.721558	12545.7771	2.728155
std	18183.213715	9.885915	14.049471	16339.486151	8109.450147	8703.364247	20960.844623	183.822223	77.517866	123.397180	1.136004	0.185427	20507.8167	7.364164
min	1.000000	6.000000	2.000000	0.000000	0.000000	0.000000	368.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.0000	0.000000
25%	15747.750000	35.000000	3.000000	2518.000000	1003.000000	780.000000	4747.000000	29.000000	23.370370	79.000000	0.000000	0.611997	2775.0000	0.000000
50%	31494.500000	41.000000	7.000000	5700.000000	2800.000000	2773.000000	9994.000000	108.000000	44.666667	143.000000	0.000000	0.711856	6328.5000	0.000000
75%	47241.250000	48.000000	15.000000	12831.000000	6574.000000	6845.750000	21271.250000	268.000000	82.000000	228.000000	0.000000	0.809476	14302.5000	1.000000
max	62988.000000	110.000000	213.000000	505308.000000	239560.000000	234188.000000	580717.000000	731.000000	728.000000	728.000000	46.000000	1.500000	985572.0000	140.000000

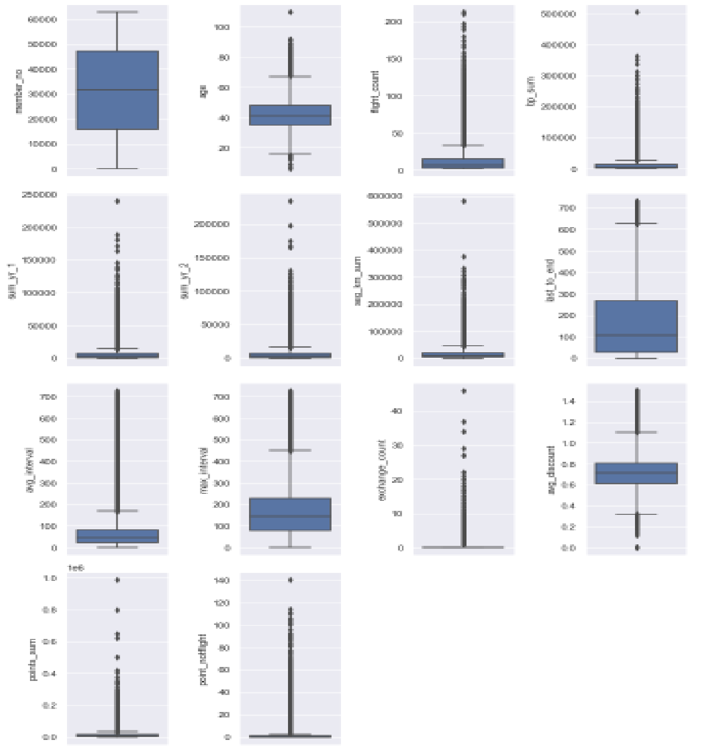
Numerical kdePlot

```
plt.figure(figsize= (20,10))
for i in range(len(num)):
    plt.subplot(4, 4, i+1)
    sns.kdeplot(x = df[num[i]])
    plt.tight_layout()
```



Numerical BoxPlot

```
plt.figure(figsize= (10,15))
for i in range(len(num)):
    plt.subplot(4, 4, i+1)
    sns.boxplot(y = df[num[i]], orient='v')
    plt.tight_layout()
```



Handling Missing Value

```
df.isna().sum()
```

```
member_no                0
ffp_date                 0
first_flight_date        0
gender                   3
ffp_tier                 0
work_city             2269
work_province         3248
work_country            26
age                    420
load_time                0
flight_count             0
bp_sum                   0
sum_yr_1               551
sum_yr_2               138
seg_km_sum               0
last_flight_date       421
last_to_end              0
avg_interval             0
max_interval             0
exchange_count           0
avg_discount             0
points_sum               0
point_notflight          0
dtype: int64
```

```
df_clean = df.copy()
```

```
df_clean['age'].fillna(df_clean['age'].mean(), inplace=True)
```

Imputasi fitur 'age' dengan mean, berdasarkan distribusinya yang cenderung normal maka lebih baik diimputasi dengan nilai mean.

```
df_clean['sum_yr_1'].fillna(df_clean['sum_yr_1'].median(), inplace=True)
df_clean['sum_yr_2'].fillna(df_clean['sum_yr_2'].median(), inplace=True)
```

Imputasi fitur 'sum_yr_1' dan 'sum_yr_2' dengan median, berdasarkan distribusinya yang skewed-right maka lebih baik diimputasi dengan nilai median.

```
df_clean['gender'] = df_clean['gender'].fillna(df_clean['gender'].mode()[0])
df_clean['work_city'] = df_clean['work_city'].fillna(df_clean['work_city'].mode()[0])
df_clean['work_province'] = df_clean['work_province'].fillna(df_clean['work_province'].mode()[0])
df_clean['work_country'] = df_clean['work_country'].fillna(df_clean['work_country'].mode()[0])
```

```
df_clean['last_flight_date'] = df_clean['last_flight_date'].fillna(df_clean['last_flight_date'].mode()[0])
```

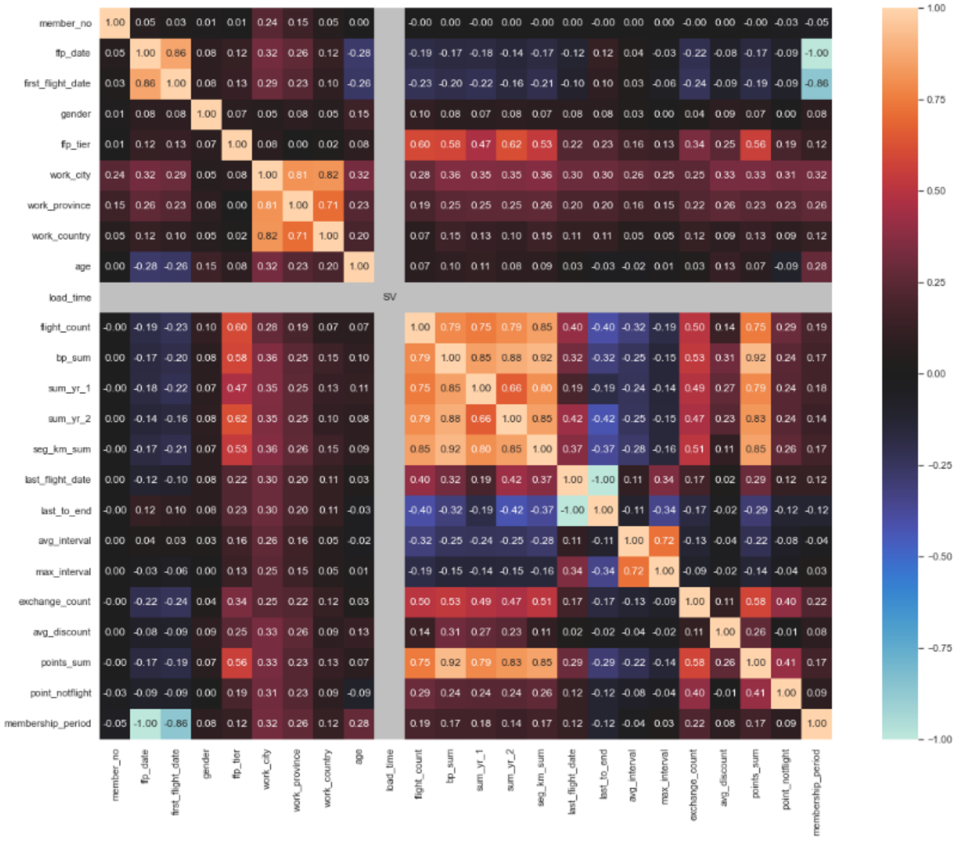
Imputasi fitur 'gender', 'work_city', 'work_province', 'work_country' dengan modus, karena fitur tersebut merupakan kategorikal maka lebih baik diisi dengan nilai modus. Sedangkan fitur 'last_fight_date' yang sebenarnya merupaka kategorikal hanya saja valuenya berupa waktu maka lebih baik diimputasi dengan nilai modus juga.

```
df_clean.isnull().sum()
```

```
member_no            0
ffp_date             0
first_flight_date    0
gender               0
ffp_tier             0
work_city            0
work_province        0
work_country         0
age                  0
load_time            0
flight_count         0
bp_sum               0
sum_yr_1             0
sum_yr_2             0
seg_km_sum           0
last_flight_date     0
last_to_end          0
avg_interval         0
max_interval         0
exchange_count       0
avg_discount         0
points_sum           0
point_notflight      0
dtype: int64
```

Feature Selection

```
fig, ax = plt.subplots(figsize=(20, 15))
associations(df_fs, ax=ax)
```



kolom yang akan dipilih berdasarkan RFM degan metode reduce dimensionality:

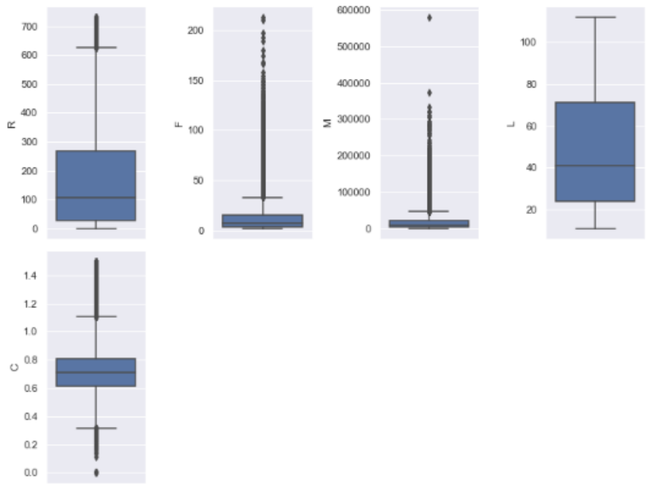
1. R: last_to_end
2. F: flight_count
3. M: seg_km_sum
4. L: membership_period
5. C: avg_discount

```
df_rd = df_fs.copy()
df_rd = df_rd[['last_to_end', 'flight_count', 'seg_km_sum', 'membership_period', 'avg_discount']]
df_rd.columns = ['R', 'F', 'M', 'L', 'C']
df_rd.describe()
```

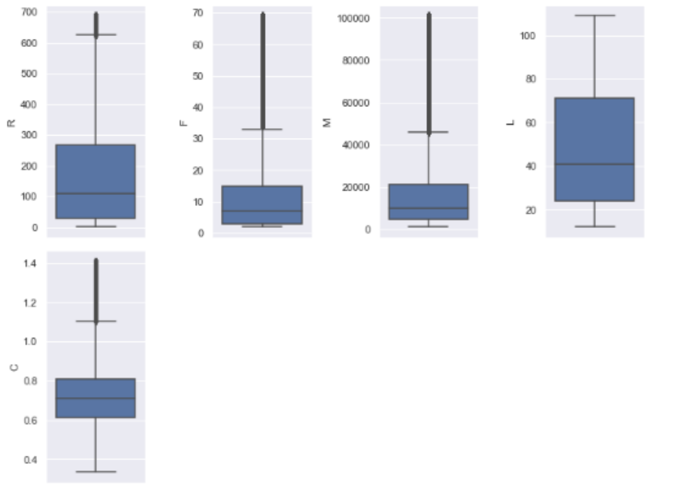
	R	F	M	L	C
count	62988.000000	62988.000000	62988.000000	62988.000000	62988.000000
mean	176.120102	11.839414	17123.878691	48.287499	0.721558
std	183.822223	14.049471	20960.844623	27.831879	0.185427
min	1.000000	2.000000	368.000000	11.000000	0.000000
25%	29.000000	3.000000	4747.000000	24.000000	0.611997
50%	108.000000	7.000000	9994.000000	41.000000	0.711856
75%	268.000000	15.000000	21271.250000	71.000000	0.809476
max	731.000000	213.000000	580717.000000	112.000000	1.500000

Handling Outliers

```
cols = df_ro.columns
plt.figure(figsize= (10,15))
for i in range(len(cols)):
    plt.subplot(4, 4, i+1)
    sns.boxplot(y = df_ro[cols[i]], orient='v')
    plt.tight_layout()
```




```
cols = df_ro.columns
plt.figure(figsize= (10,15))
for i in range(len(cols)):
    plt.subplot(4, 4, i+1)
    sns.boxplot(y = df_ro[cols[i]], orient='v')
    plt.tight_layout()
```



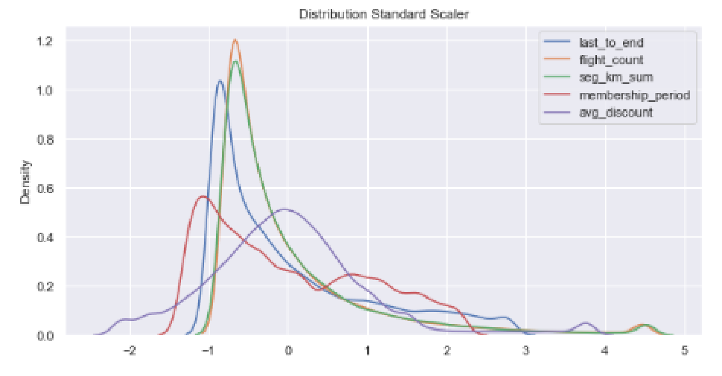

```
df_ro.describe()
```

	R	F	M	L	C
count	62988.000000	62988.000000	62988.000000	62988.000000	62988.000000
mean	175.932574	11.626945	16757.990417	48.272496	0.721600
std	183.287089	12.811748	18565.703613	27.796717	0.182497
min	1.000000	2.000000	1190.000000	12.000000	0.334762
25%	29.000000	3.000000	4747.000000	24.000000	0.611997
50%	108.000000	7.000000	9994.000000	41.000000	0.711856
75%	268.000000	15.000000	21271.250000	71.000000	0.809476
max	687.000000	69.000000	100841.280000	109.000000	1.410000

Feature Transformation
Standard Scaler & MinMax Scaler

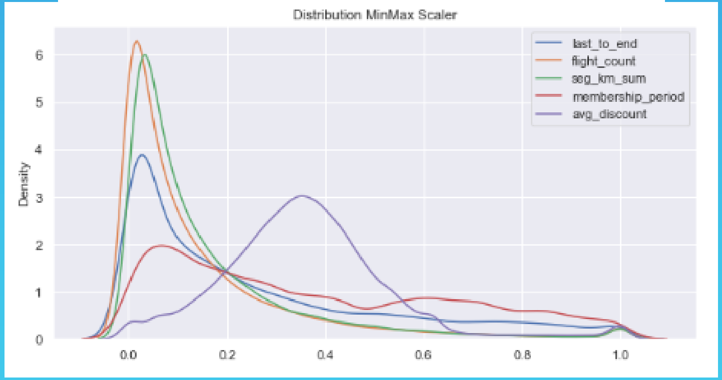
```
col_name = list(df_ro.columns)

sc = StandardScaler()
df_std_sc = sc.fit_transform(df_ro)
df_std_sc = pd.DataFrame(df_std_sc, columns=col_name)
df_std_sc.head()
```

	R	F	M	L	C
0	-0.954426	4.478195	4.528994	1.429227	1.315308
1	-0.921690	4.478195	4.528994	1.321300	2.908085
2	-0.899866	4.478195	4.528994	1.321300	2.921023
3	-0.430653	0.887712	4.528994	0.673736	2.023436
4	-0.932602	4.478195	4.528994	0.385930	1.364728

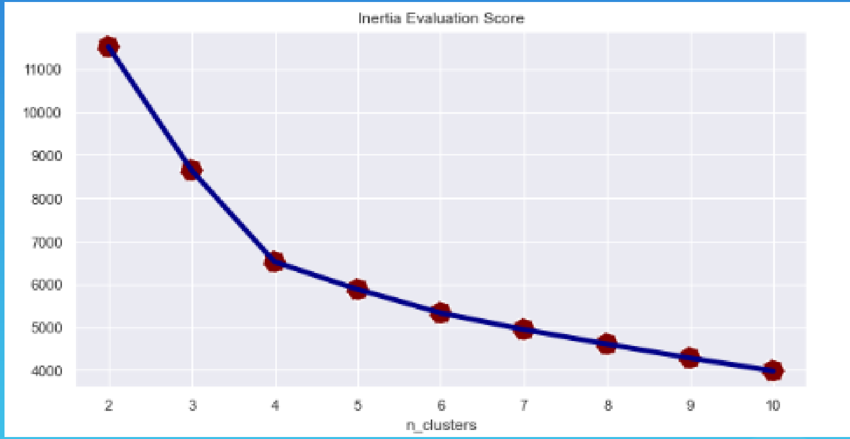


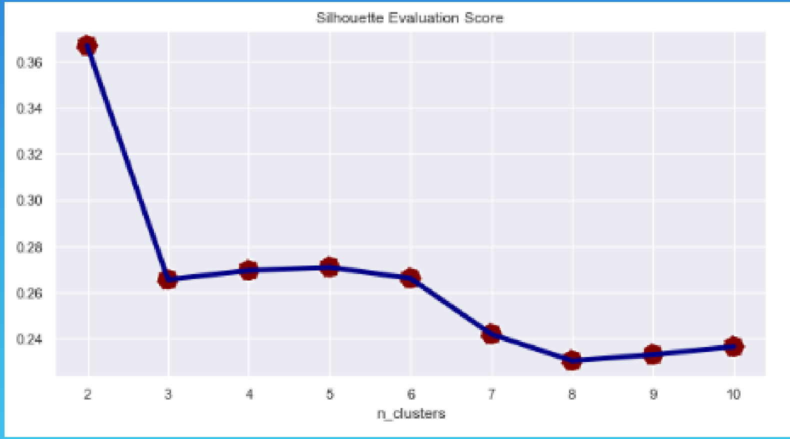

```
col_name = list(df_ro.columns)

mm = MinMaxScaler()
df_std_mm = mm.fit_transform(df_ro)
df_std_mm = pd.DataFrame(df_std_mm, columns=col_name)
df_std_mm.head()
```

	R	F	M	L	C
0	0.000000	1.000000	1.0	0.783505	0.583012
1	0.008746	1.000000	1.0	0.752577	0.853348
2	0.014577	1.000000	1.0	0.752577	0.855544
3	0.139942	0.313433	1.0	0.567010	0.703200
4	0.005831	1.000000	1.0	0.484536	0.591400



Modeling and evaluation

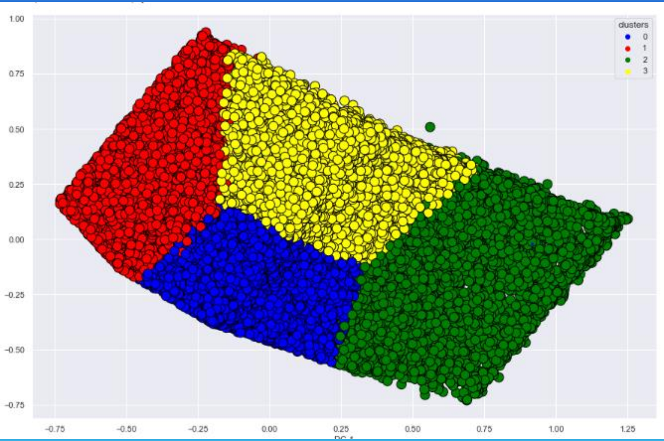




Business Insight

```
pca = PCA(n_components=2)

pca.fit(df_std_mm)
pcs = pca.transform(df_std_mm)

df_pca = pd.DataFrame(data = pcs, columns = ['PC 1', 'PC 2'])
df_pca['clusters'] = df_cluster['clusters']
df_pca.sample(10)
```

	PC 1	PC 2	clusters
46899	-0.216	0.099	1
27072	-0.039	-0.176	0
52257	-0.017	0.033	0
2410	0.868	0.186	2
3858	0.667	0.255	2
53000	-0.118	-0.278	0
21008	-0.054	-0.392	0
4304	0.485	0.186	2
14571	0.203	-0.076	0
8559	0.211	-0.167	0



	R		F		M		L		C	
	mean	median	mean	median	mean	median	mean	median	mean	median
clusters										
0	100.085	79.000	9.504	8.000	13622.337	10600.000	29.040	27.000	0.704	0.700
1	480.511	475.000	3.899	3.000	6067.381	4289.500	38.377	32.000	0.716	0.715
2	28.535	14.000	42.649	40.000	61278.551	56234.000	62.266	62.000	0.788	0.749
3	116.608	86.000	9.904	8.000	14123.915	11325.000	80.793	80.000	0.729	0.713

Berdasarkan hasil k-means clustering dan pca di atas, dapat disimpulkan:
1. cluster 0 (potential customer) merupakan kelompok pelanggan penerbangan yang tidak terlalu sering/tidak jarang juga melakukan penerbangan, serta merupakan pelanggan memiliki masa aktif sebagai member paling sebentar/pelanggan baru (L rendah), tingkat monetarynya terbilang sedang.
2. cluster 1 (low-valued customer) merupakan kelompok pelanggan penerbangan yang jarang melakukan penerbangan (frekuensi kecil), menghabiskan uang paling sedikit (monetery kecil)
3. cluster 2 (high-valued customer) merupakan kelompok pelanggan penerbangan yang sering melakukan penerbangan (frekuensi tinggi), menghabiskan uang paling banyak (monetary tinggi)
4. cluster 3 (retain-required customer) merupakan kelompok pelanggan penerbangan yang tidak terlalu sering/tidak jarang juga melakukan penerbangan, tetapi merupakan pelanggan yang memiliki masa aktif sebagai member paling lama/pelanggan lama (L paling tinggi), tingkat monetarynya terbilang sedang.

Business Recommendation

Promo yang akan diberikan dalam business recommendation berupa kupon dan diskon penerbangan.
Loyalty point berupa pemberian poin tambahan disetiap kali melakukan penerbangan sesuai dengan kelas pelanggan.
1. Untuk cluster 0 yang dapat dikategorikan sebagai pelanggan baru, kita bisa memberikan promo.
2. Untuk cluster 3 yang dikategorikan sebgai pelanggan loyal, kita dapat memberikan 'loyalty point + promo'
3. untuk cluster 2 yang dikategorikan sebagai kita dapat memberikan 'loyalty point + peningkatan kelas pelanggan(4,5,6) + promo' dengan benefit tertentu di setiap tingkatannya
4. untuk cluster 1 dapat diberikan promo yang sama dengan cluster 0 dengan tujuan meningkatkan frekuensi penerbangan.

	clusters	Total Customers
0	0	26820
1	1	12942
2	2	6015
3	3	17211

Thank You!!!