Analysis Clustering Customers

Ersa Zhafrina Febriani

This project consists of data cleaning, simple analysis, EDA, to clustering analysis. From the data given, there are fields as follows :

Fields

Field	Description
taskId	Unique identifier for the task that generated by system.
taskCreatedTime	Time at when the task was created.
taskCompletedTime	Time at when the task was completed.
taskAssignedTo	Worker that doing the task.
taskLocationDone	Coordinate of where the task was completed.
flow	Flow or type of the task.
cod	Contains data for the COD system.
cod.amount	Amount of money from COD.
cod.received	COD has been received or not.
UserVar	Contains more specified data, in this case the 'UserVar' is about delivery task data.
UserVar.taskStatus	Delivery status code.
UserVar.taskStatusLabel	Delivery status label.
UserVar.taskDetailStatus	Detailed delivery status code.
UserVar.taskDetailStatusLabel	Detailed delivery status label.
UserVar.branch_origin	Branch code of the origin.
UserVar.branch_dest	Branch code of the destination.
UserVar.weight	Weight of the package.

But from some of the fields above I only use the fields that are needed for this analysis project.

A. Import Data

Because the dataset is provided in .json, it takes several steps to load the dataset. First I separate the existing nested lists so that they form a good dataset, then export to excel to tidy up a bit and reload the data in the form of .xlsx. This method may seem complicated and long, but I have to use this method to save time, because I started this project after the Eid holiday which is May 2, 2023. This is the General Information of this Dataset.

General Information

```
[365] df_sample.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8334 entries, 0 to 8333
Data columns (total 17 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   taskId                       8334 non-null   object
 1   taskCreatedTime              8334 non-null   object
 2   taskCompletedTime            7566 non-null   object
 3   taskAssignedTo               8333 non-null   object
 4   taskLocationDone             7566 non-null   object
 5   flow                         8334 non-null   object
 6   cod.amount                   2358 non-null   float64
 7   cod.received                 2358 non-null   float64
 8   UserVar.taskStatus           7572 non-null   object
 9   UserVar.taskStatusLabel      7572 non-null   object
 10  UserVar.taskDetailStatus     7572 non-null   object
 11  UserVar.taskDetailStatusLabel 7572 non-null  object
 12  UserVar.branch_origin        8041 non-null   object
 13  UserVar.branch_dest          8334 non-null   object
 14  UserVar.weight               8334 non-null   float64
 15  receiver_city                8282 non-null   object
 16  taskStatus                   8334 non-null   object
dtypes: float64(3), object(14)
memory usage: 1.1+ MB
```

B. Data Understanding

There are 8334 rows of data with 17 features, for this case the dataset is considered to have no target/output/label so in this segmentation case we will use Machine Learning Unsupervised - Clustering type because I think when using Machine Learning Supervised type, the column features provided are lacking so do not have significant information. Because taskStatus and taskStatusLabel have the same function, one of the two must be dropped. The same condition is also owned by taskDetailStatus and taskDetailStatusLabel, so that the taskStatus and taskDetailStatus columns will be dropped. taskId will also be dropped. A new variable is created where the variable is how long the task has been completed (in hours).

timeLoad
0.48
3.88
5.01
9.62
2.18

The dataset is divided into 2, namely numerical data and categorical data. There are 4 Numerical feature columns and 9 Categorical feature columns.

```
[372] numerics = ['int8','int16', 'int32', 'int64', 'float16', 'float32', 'float64']
      display(df_sample.select_dtypes(include=numerics).columns)
      print(df_sample.select_dtypes(include=numerics).shape)
      numericals = df_sample.select_dtypes(include=numerics)
      numericals.head(3)

      Index(['cod.amount', 'cod.received', 'UserVar.weight', 'timeLoad'], dtype='object')
      (8334, 4)
```

	cod.amount	cod.received	UserVar.weight	timeLoad
0	685000.0	1.0	13.0	0.48
1	53500.0	1.0	1.3	3.88
2	179500.0	1.0	3.0	5.01

Ada 4 kolom fitur Numerical

Non-Numerical / Categorical

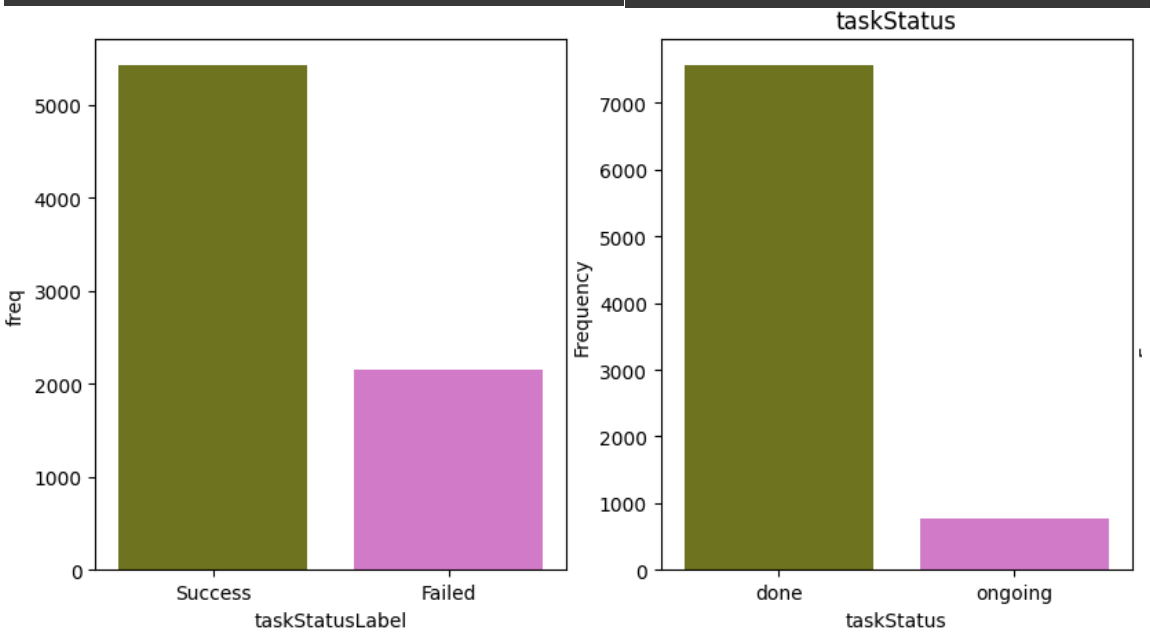
```
[373] display(df_sample.select_dtypes(include=['object']).columns)
      print(df_sample.select_dtypes(include=object).shape)
      categoricals = df_sample.select_dtypes(include=['object'])
      categoricals.head(3)

      Index(['taskAssignedTo', 'taskLocationDone', 'flow', 'UserVar.taskStatusLabel',
             'UserVar.taskDetailStatusLabel', 'UserVar.branch_origin',
             'UserVar.branch_dest', 'receiver_city', 'taskStatus'],
            dtype='object')
      (8334, 9)
```

C. Simple Analysis and EDA

The dataset used during the analysis is a dataset where the missing value has not been handled, so there is still empty data in the dataset. This is done so that no information is lost from the dataset.

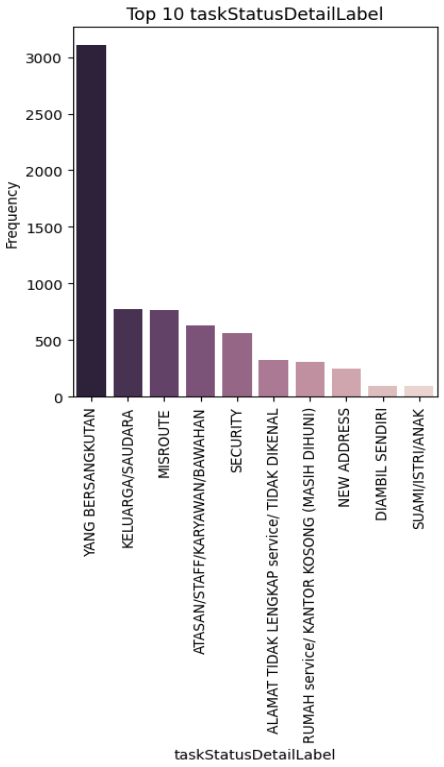
	taskStatusLabel	freq	percentage
0	Success	5427	71.67
1	Failed	2145	28.33

	taskStatus	freq	percentage
0	done	7572	90.86
1	ongoing	762	9.14



There are 2 status tasks, the first is the status task for whether the item has been successfully received or not, the second is whether the item has been sent successfully (successfully or failed) or is still in the shipping process.

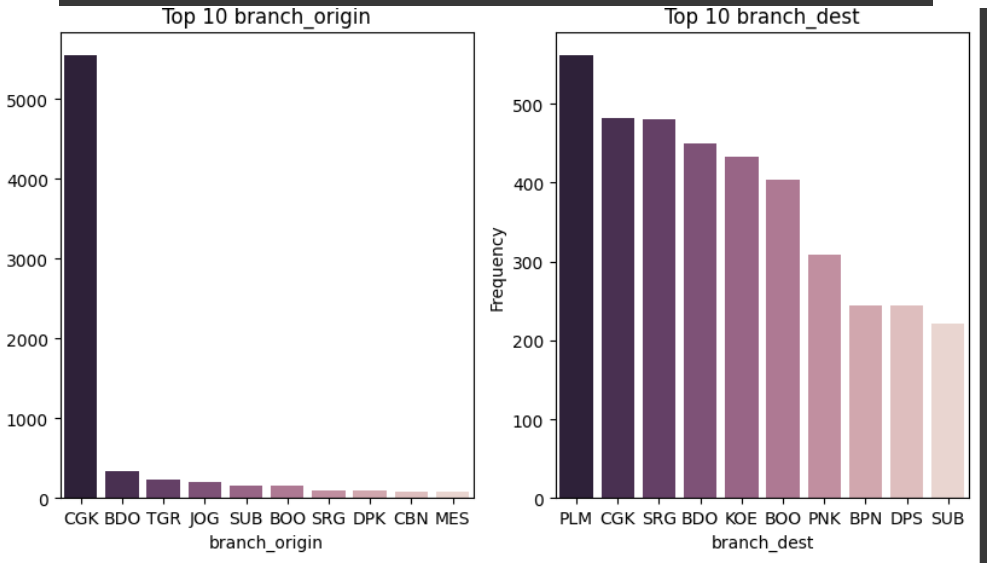
	taskStatusDetailLabel	freq	percentage
0	YANG BERSANGKUTAN	3109	41.06
1	KELUARGA/SAUDARA	774	10.22
2	MISROUTE	763	10.08
3	ATASAN/STAFF/KARYAWAN/BAWAHAN	634	8.37
4	SECURITY	564	7.45
5	ALAMAT TIDAK LENGKAP service/ TIDAK DIKENAL	322	4.25
6	RUMAH service/ KANTOR KOSONG (MASIH DIHUNI)	304	4.01
7	NEW ADDRESS	247	3.26
8	DIAMBIL SENDIRI	100	1.32
9	SUAMI/ISTRI/ANAK	94	1.24



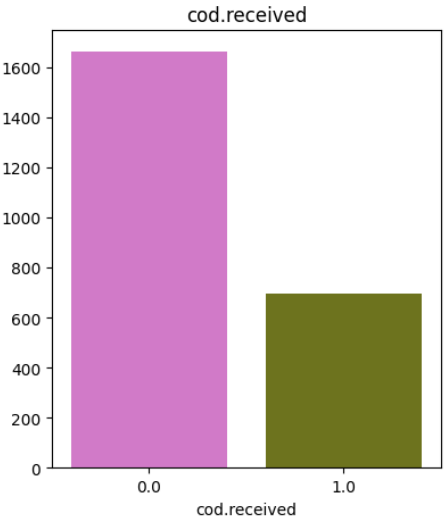
Packets received by "YANG BERSANGKUTAN" are the most packet status.

	branch_origin	freq	percentage
0	CGK	5550	69.02
1	BDO	341	4.24
2	TGR	226	2.81
3	JOG	206	2.56
4	SUB	164	2.04
5	BOO	158	1.96
6	SRG	95	1.18
7	DPK	89	1.11
8	CBN	85	1.06
9	MES	81	1.01

	branch_dest	freq	percentage
0	PLM	562	6.74
1	CGK	482	5.78
2	SRG	480	5.76
3	BDO	450	5.40
4	KOE	432	5.18
5	BOO	403	4.84
6	PNK	309	3.71
7	BPN	245	2.94
8	DPS	244	2.93
9	SUB	221	2.65

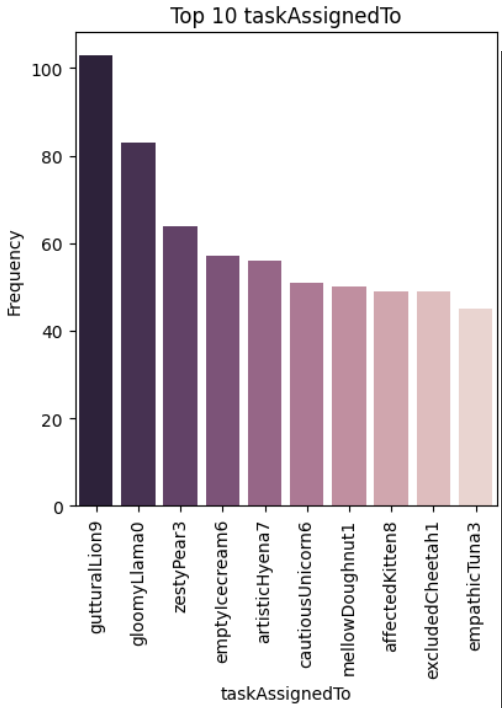


While the most origin branch is "CGK" and the most destination branch is "PLM".



	cod.received	freq	percentage
0	0.0	1663	70.53
1	1.0	695	29.47

COD packages that have been received are 1.0 status of 70.53%.



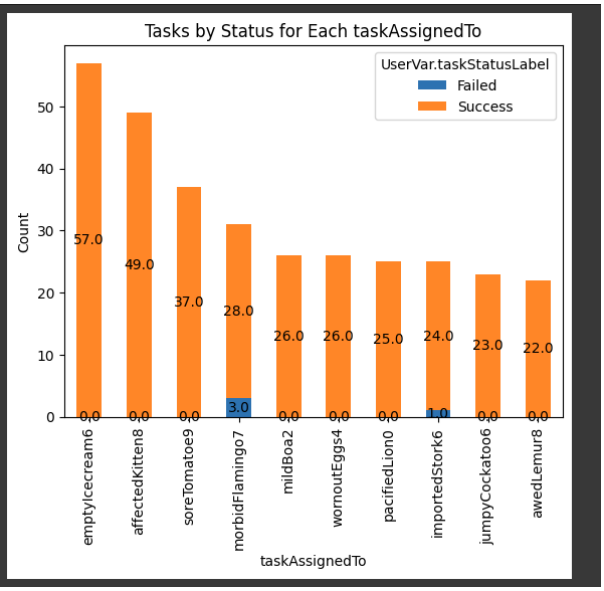
	taskAssignedTo	freq	percentage
0	gutturalLion9	103	1.24
1	gloomyLlama0	83	1.00
2	zestyPear3	64	0.77
3	emptyIcecream6	57	0.68
4	artisticHyena7	56	0.67
5	cautiousUnicorn6	51	0.61
6	mellowDoughnut1	50	0.60
7	affectedKitten8	49	0.59
8	excludedCheetah1	49	0.59
9	empathicTuna3	45	0.54

The courier who received the most tasks was gutturalLion9.

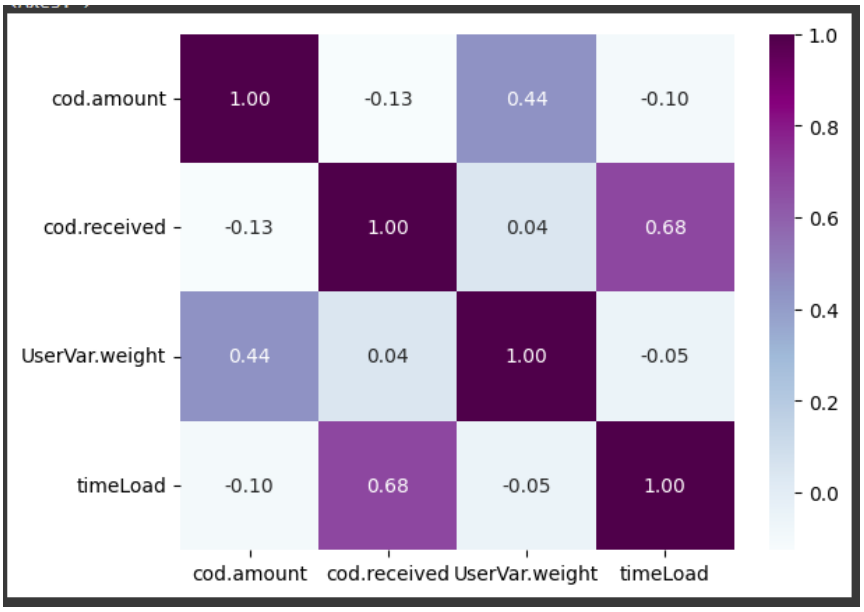
UserVar.taskStatusLabel	Failed	Success
taskAssignedTo		
affectedKitten8	0.00	274.76
importedStork6	0.03	179.82
emptyIcecream6	0.00	171.30
soreTomatoe9	0.00	167.98
wornoutEggs4	0.00	154.03
pacifiedLion0	0.00	142.30
peacefulVenison0	0.00	141.05
enragedCaribou6	0.00	81.75
crummyCrane8	0.00	70.99
guiltyBurritos6	0.00	63.10

The courier with the longest load time with success status is affectedKitten8.

UserVar.taskStatusLabel	Failed	Success
taskAssignedTo		
emptyIcecream6	0	57
affectedKitten8	0	49
soreTomatoe9	0	37
morbidFlamingo7	3	28
mildBoa2	0	26
wornoutEggs4	0	26
pacifiedLion0	0	25
importedStork6	1	24
jumpyCockatoo6	0	23
awedLemur8	0	22



But the courier with the highest number of successes is emptyIcecream6



Received and time have a high correlation

D. Modeling

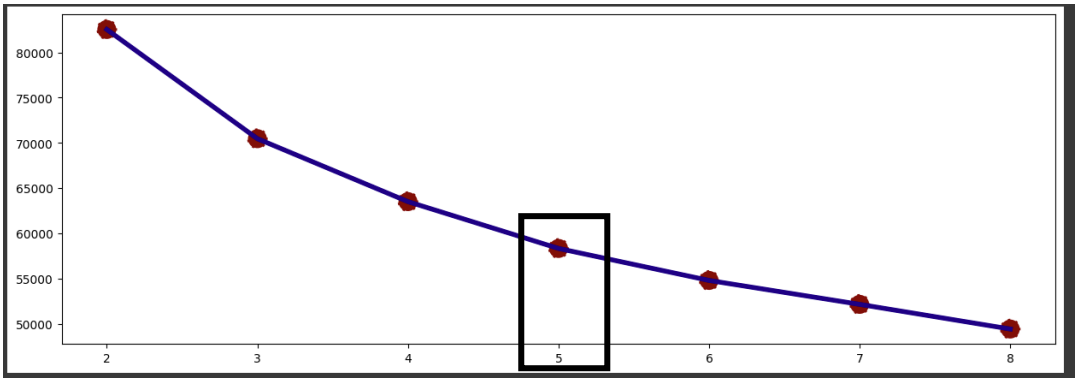
Before doing modeling, data pre-processing must be done first.

1. Data Cleaning

   At this stage, handling missing values, handling duplicate data, Encoding, and handling outliers are carried out
2. Scalling

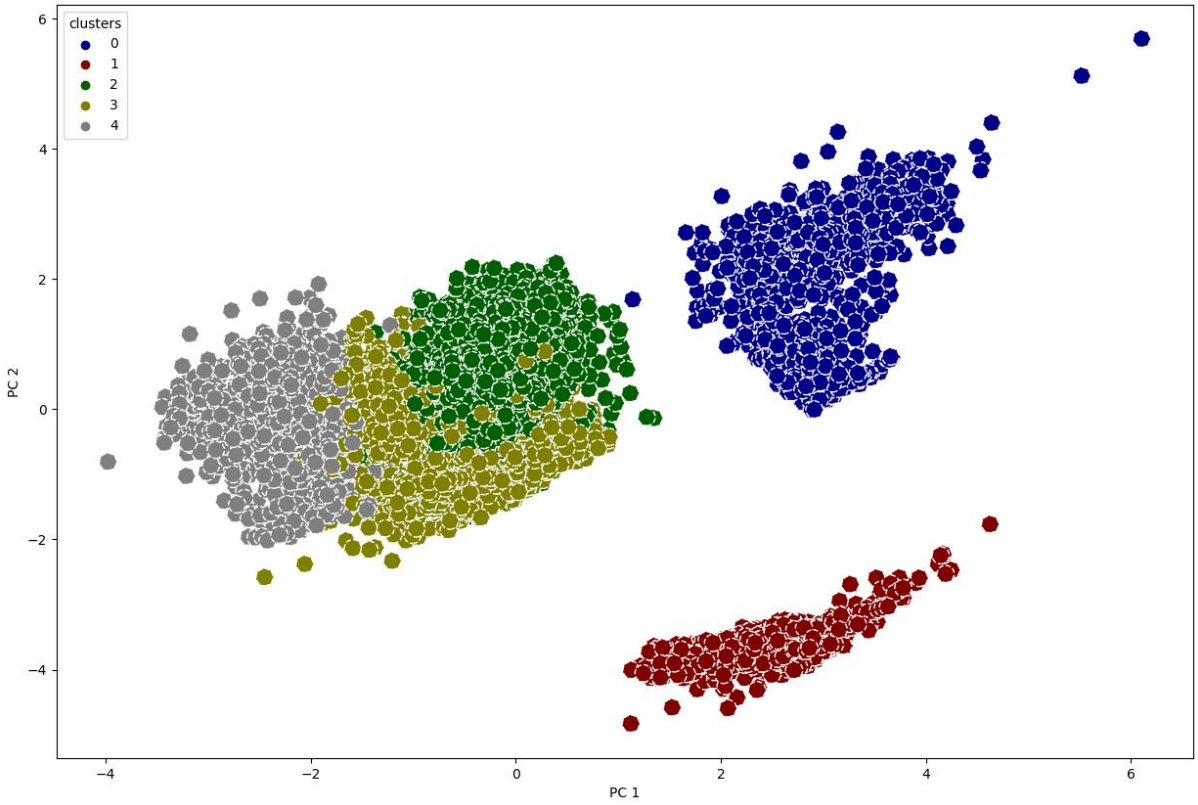
After Data Pre processing was done, then we can do the modelling

1. Find the best K



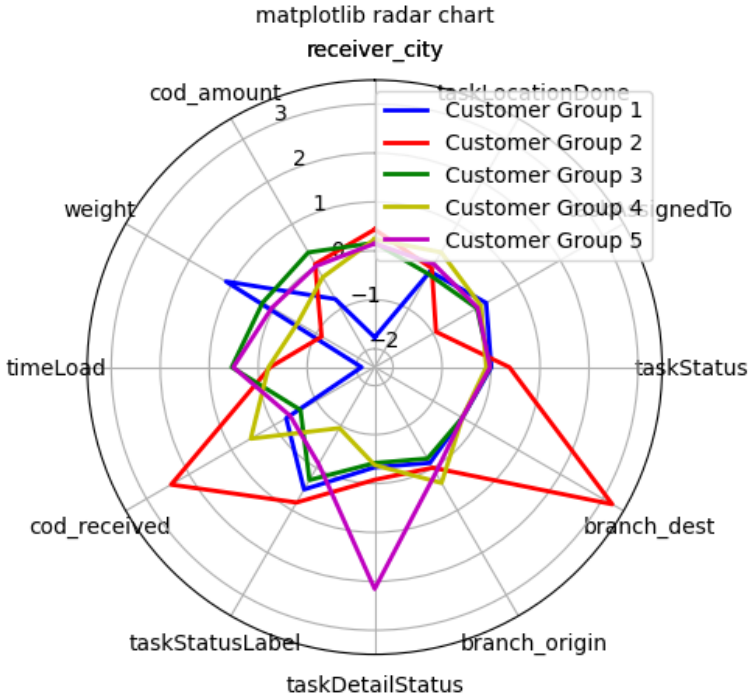
Best K is 5

2. Clustering
3. Clustering Visualization



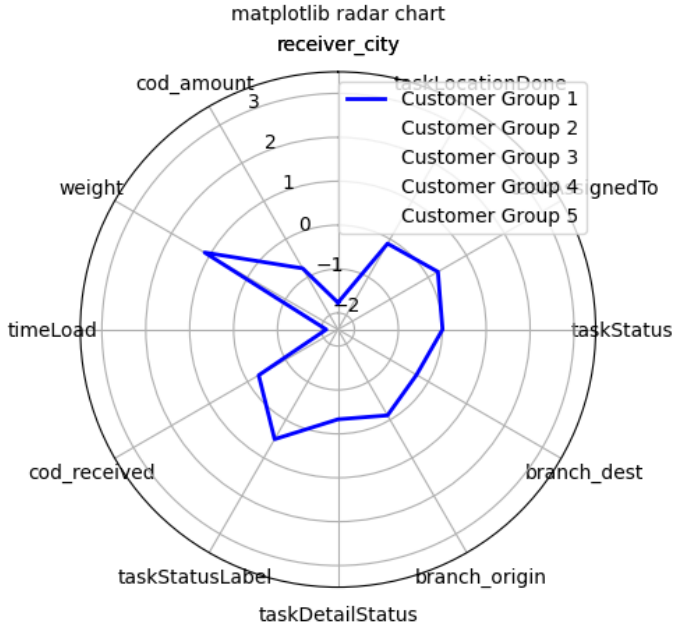
E. Insight – Analysis Clustering

clusters	cod.amount	weight	timeLoad	cod_received	taskStatus Label	taskDetail StatusLabel_ mean	taskDetail StatusLabel_med ian	branch_ origin	branch _dest	taskStatus	receiver_city
0	0.759136	0.072936	0.177412	TRUE	Success	SUPIR	YANG BERSANGKUTAN	CKR	KOE	done	KUTA SELATAN,BADUNG
1	0.796488	0.138876	0	FALSE	Unknown	UNKNOWN	UNKNOWN	DJJ	MDC	ongoing	LEGOK,TIGARAKSA
2	0.791863	0.161027	0.109375	UNKNOWN	Success	SECURITY	YANG BERSANGKUTAN	CGK	KDI	done	KOTA UTARA,GORONTALO
3	0.793139		0.04434	FALSE	Failed	MENUNGGU KONFIRMASI NILAI COD	MISROUTE	CKR	PBL	done	NGANJUK
4	0.791589	0.13535	0.092436	UNKNOWN	Success	PENJAGA KOS	SECURITY	SUB	MES	done	MAKALE,KAB.TANA TORA



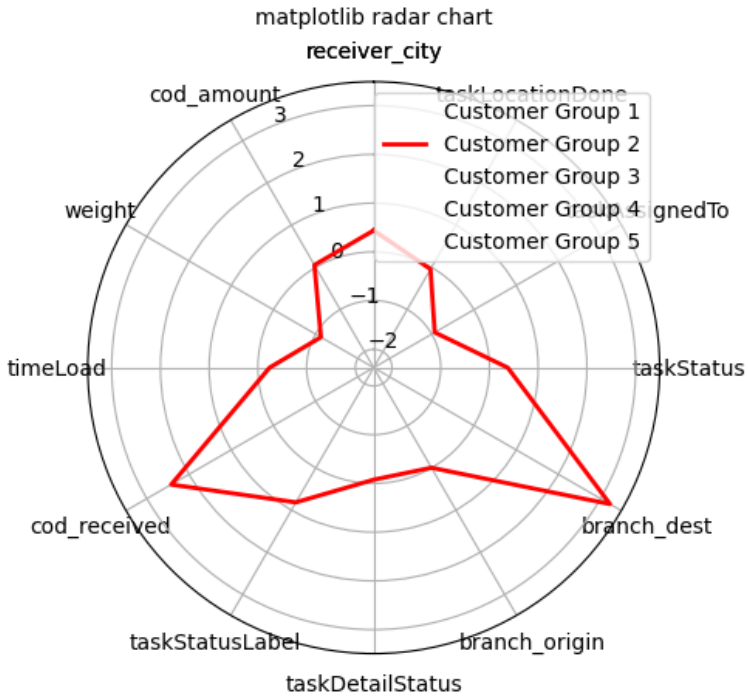
There are 5 Customer Cluster that has been made.

1. Cluster 0 – Customer Group 1



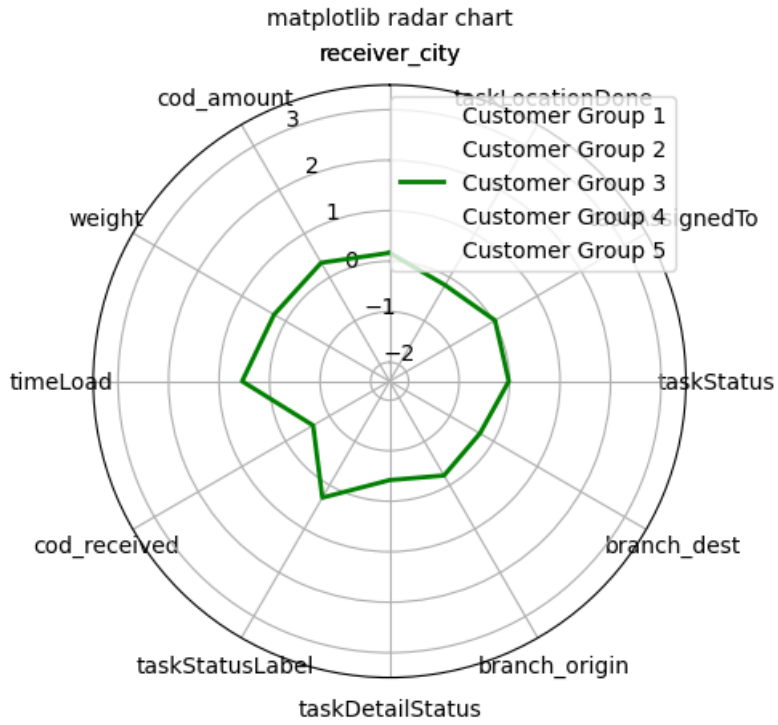
This customer group has a lower average cod amount and weight than the others, and has the highest average timeload or task completion time among the other clusters. In this cluster, on average, the item has been sent to the customer and received by the driver or person concerned. It can be concluded that customers in this cluster on average have a high level of success in completing their tasks.

2. Cluster 1 – Customer Group 2



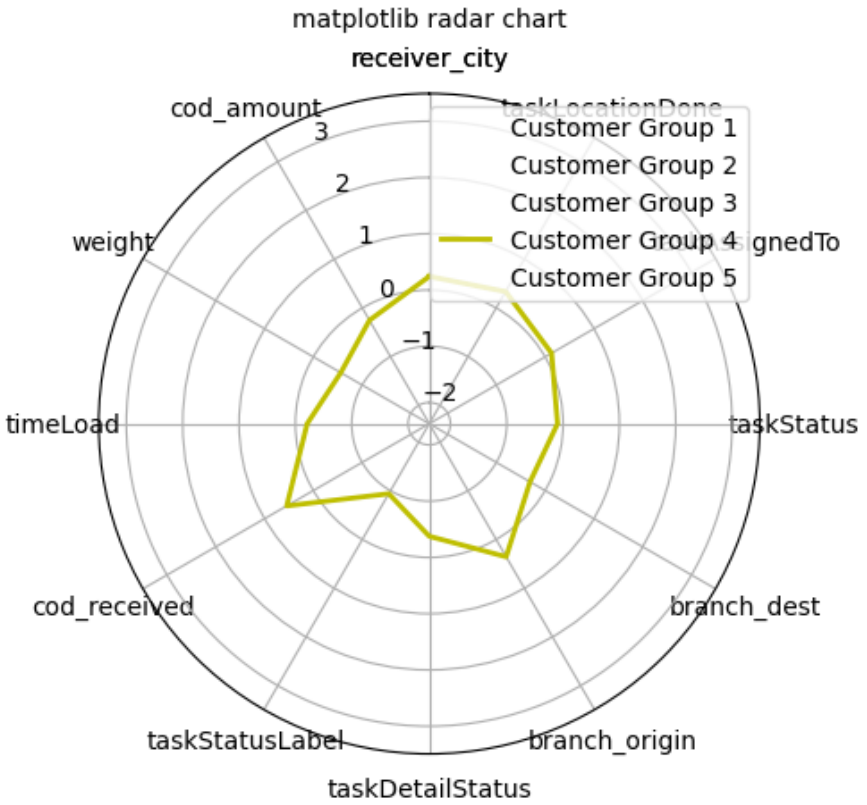
This customer group has the highest average cod amount than the others, while the weight is at the middle rate, and has an average time load or task completion time of 0. In this cluster, on average, goods have not been sent or have not been received by the customer because the task status is still ongoing, the goods are in the process of being shipped. It can be concluded that customers in this cluster are on average customers whose task status is still on going.

3. Cluster 2 – Customer Group 3



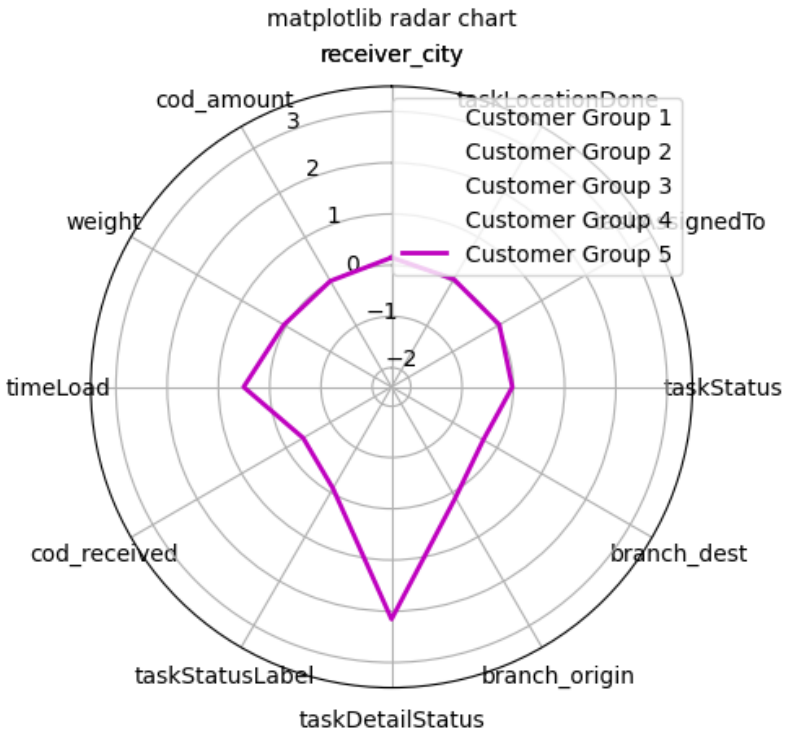
This customer group has an average cod amount, weight and time load at the middle rate. In this cluster, on average, the goods have been done or the task status has been done and successful, but it is not clear whether the goods have been received by the customer or not. The average packet recipient in the cluster is Security. It can be concluded that the average customer in this cluster is a customer whose task status has been done and successful but does not know whether the customer has received the goods or not because the average packet recipient in the cluster is Security and has an average cod amount, weight and time load at the middle rate.

4. Cluster 3 – Customer Group 4



This customer group has an average cod amount, weight and time load at the middle rate. In this cluster, on average, the task status is Failed, so the package fails to be received by the customer. The average cause of failure to receive packets is a misroute or waiting for confirmation of the COD value. It can be concluded that the average customer in this cluster is a customer whose task status is Failed or fails in delivery.

5. Cluster 4 – Customer Group 5



This customer group has an average cod amount and time load at the middle rate while the average weight is the lowest among the others. In this cluster, the average item has been completed or the status of the task has been completed and successful, but it is not clear whether the goods have been received by the customer or not. The average packet recipient in the cluster is the guard of the house or boarding house. It can be concluded that the average customer in this cluster is a customer whose task status has been completed and successful but does not know whether the customer has received the goods or not because the average package recipient in the cluster is a house keeper or boarding house keeper. This cluster is almost the same as Cluster 2 or Customer Group 3, but the difference is that this cluster has the lowest weight among the others.

F. Conclusion

From the clustering that has been done, there are 5 clustering with different characteristics. Clustering for this dataset will be very useful for the Risk Management team because we can find out 5 different task flow characteristics and can analyse which clusters will become a loophole for fraud and other analysis. Usually this clustering is done to determine the type of customer and can be used to increase customer satisfaction and measure the churning rate of existing datasets. but for this case, clustering can be used to determine the types or characteristics of delivery flows which can later be used for risk management analysis.

G. Recommendation

After knowing the characteristics of each existing cluster, my recommendation is to discuss with the risk management team to carry out further analysis and make decisions together. In my opinion, cluster 2 and cluster 4 (which is Customer Group 3 and 5) are clusters that can trigger loopholes for fraud, so there is a need for improvement both in terms of application security or SOP security from field staff or others terms and conditions. To carry out further and in-depth analysis, data relevant to this dataset is needed and also some people who understand in risk management.