Data Engineer Test

By Muhammad Rizky Halawi

Content

Preparation of This Project

The Data:

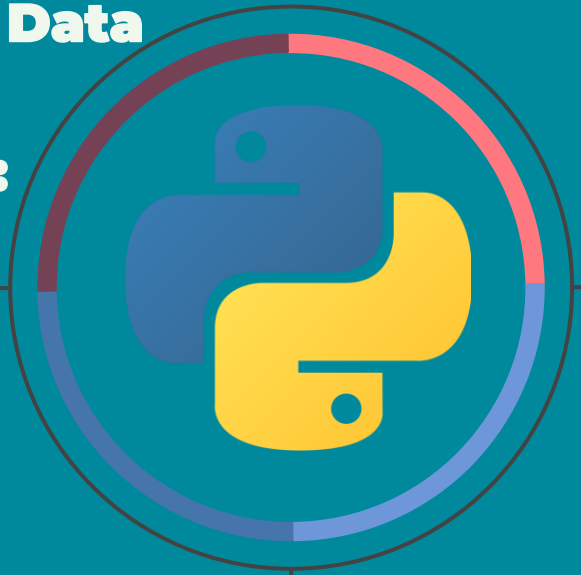
1. Master Data_Data Engineer
2. Scan Record Data

Important things:

1. Waybill No.: Unique ID
2. Origin Branch: Origins Area Packet
3. POD Branch : Destinations Area
4. Destination: Destination of district
5. Record Time: Time of Packet get pickup
6. POD Time: Time of Packet already Delivered
7. Arrival Scan: Time of Packet Arrive at Branch
8. Delivery Scan: Time of Packet Start to deliver at Destination

Tools

Pyton: for Analyze Data



Jupyter Notebook: Make a Script

Tableau : for Data Vizualization

Question

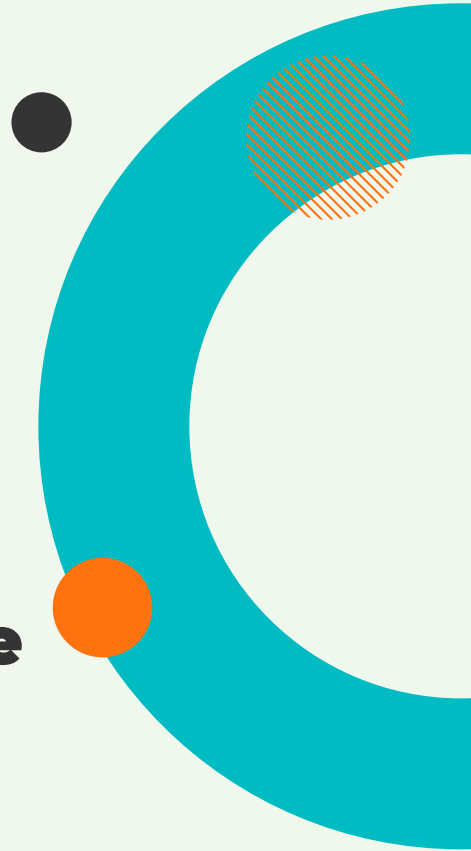
01 Data Cleansing

02 Make a New Data

03 Measure Time Interval

04 Measure mean of Interval time For Each Branch

05 Data Visualization



Processing

Data Cleansing

- Import data and read data
- Merge data and profilling
- Check missing data and delete
- Check Duplicates
- Delete missing and duplicate
- Last Data Profiling

Before

	Waybill No.	Shipping Date	Origin Branch	Pickup Courier	Destination
0	IDD040187141499	02/05/2020	CP BATU AJI	NaN	LUBUK BAJA
1	IDD040187141499	02/05/2020	CP BATU AJI	NaN	LUBUK BAJA
2	IDD041824111253	02/05/2020	CP BATAM KOTA	NaN	BATAM KOTA
3	IDD041824111253	02/05/2020	CP BATAM KOTA	NaN	BATAM KOTA
4	IDP195673872055	02/05/2020	CP TANDES	Benekditus Eko Margono	TAMBAKSARI
5	IDP195673872055	02/05/2020	CP TANDES	Benekditus Eko Margono	TAMBAKSARI
6	IDS006668957969	03/05/2020	CP VIP HQ	HQ POP	GENTENG

After

Waybill No.	Shipping Date	Origin Branch	Pickup Courier	Destination	Cacu
IDP195673872055	02/05/2020	CP TANDES	Benekditus Eko Margono	TAMBAKSARI	0.25(
IDP195673872055	02/05/2020	CP TANDES	Benekditus Eko Margono	TAMBAKSARI	0.25(
IDS006668957969	03/05/2020	CP VIP HQ	HQ POP	GENTENG	7.00(
IDS006668957969	03/05/2020	CP VIP HQ	HQ POP	GENTENG	7.00(
IDS008961448998	03/05/2020	CP VIP HQ	HQ POP	CIMAHI SELATAN	1.76(
IDS008961448998	03/05/2020	CP VIP HQ	HQ POP	CIMAHI SELATAN	1.76(
IDS005071754751	03/05/2020	CP VIP HQ	HQ POP	BOJONGLOA KALER	0.54(
IDS005071754751	03/05/2020	CP VIP HQ	HQ POP	BOJONGLOA KALER	0.54(
IDS001516566067	03/05/2020	CP VIP HQ	VIP HQ	GENUK	0.97(
IDS001516566067	03/05/2020	CP VIP HQ	VIP HQ	GENUK	0.97(

New Data

Data Checking

Make new Columns Depend on Request (Arrival and Delivery Scan)

Change from Object to Datetime Data

Make New Data Frame Depend on Request

Result:

Waybill No.	Origin Branch	POD Branch	Destination
IDP195673872055	CP TANDES	CP MULYOREJO	TAMBAKSARI
IDP195673872055	CP TANDES	CP MULYOREJO	TAMBAKSARI
IDS006668957969	CP VIP HQ	CP BUBUTAN	GENTENG
IDS006668957969	CP VIP HQ	CP BUBUTAN	GENTENG
IDS008961448998	CP VIP HQ	CP CICENDO	CIMAHI SELATAN
...	...	...	...
IDS006122998013	CP VIP HQ	CP CILEUNGSI	GUNUNG PUTRI
IDD036573031273	CP RUMBAI PESISIR	CP RUMBAI PESISIR	RUMBAI PESISIR
IDD036573031273	CP RUMBAI PESISIR	CP RUMBAI PESISIR	RUMBAI PESISIR
IDD036573031273	CP RUMBAI PESISIR	CP RUMBAI PESISIR	RUMBAI PESISIR
IDD036573031273	CP RUMBAI PESISIR	CP RUMBAI PESISIR	RUMBAI PESISIR

Pickup Scan	Arrival Scan	Delivery Scan	POD Scan
2020-02-05	2020-04-05 10:01:09	2020-04-05 10:02:36	2020-06-05 08:40:03
2020-02-05	2020-04-05 10:01:09	2020-04-05 10:02:36	2020-06-05 08:40:03
2020-03-05	2020-04-05 10:01:09	2020-04-05 10:02:36	2020-04-05 15:16:59
2020-03-05	2020-04-05 10:01:09	2020-04-05 10:02:36	2020-04-05 15:16:59
2020-03-05	2020-04-05 10:01:09	2020-04-05 10:02:36	2020-04-05 13:29:25
...	...	...	...
2020-05-28	2020-05-21 07:24:27	2020-05-21 07:24:56	2020-05-29 11:25:53
2020-05-28	2020-05-21 07:19:46	2020-05-21 07:20:23	2020-05-28 16:50:09
2020-05-28	2020-05-21 07:19:46	2020-05-21 07:20:23	2020-05-28 16:50:09
2020-05-28	2020-05-21 07:19:46	2020-05-21 07:20:23	2020-05-28 16:50:09
2020-05-28	2020-05-21 07:19:46	2020-05-21 07:20:23	2020-05-28 16:50:09

Measure Time Interval

Data Checking

↓

Make new Columns

↓

Make the Operation Depend on request

↓

Check the Result

Request Notes:

Must measure between:
Pickup Scan – Arrival Scan
Arrival Scan – Delivery Scan
Delivery Scan – POD Time
Pickup Scan – POD Time

Result:

Waybill No.	Rentang P-A	Rentang A-D	Rentang D-P	Rentang P-P
IDP195673872055-61	days +13:58:51	00:01:27	60 days 22:37:27	121 days 08:40:03
IDP195673872055-61	days +13:58:51	00:01:27	60 days 22:37:27	121 days 08:40:03
IDS006668957969-32	days +13:58:51	00:01:27	0 days 05:14:23	31 days 15:16:59
IDS006668957969-32	days +13:58:51	00:01:27	0 days 05:14:23	31 days 15:16:59
IDS008961448998-32	days +13:58:51	00:01:27	0 days 03:26:49	31 days 13:29:25
...	...	...	...	...
IDS006122998013	6 days 16:35:33	00:00:29	8 days 04:00:57	1 days 11:25:53
IDD036573031273	6 days 16:40:14	00:00:37	7 days 09:29:46	0 days 16:50:09
IDD036573031273	6 days 16:40:14	00:00:37	7 days 09:29:46	0 days 16:50:09
IDD036573031273	6 days 16:40:14	00:00:37	7 days 09:29:46	0 days 16:50:09
IDD036573031273	6 days 16:40:14	00:00:37	7 days 09:29:46	0 days 16:50:09

Mean interval time of each Branch

Data Checking

Import datetime

Measure Mean of Each Branch

Make to Data Frame

Result:

POD Branch	Rentang P-P	Rentang jumlah per-jam
CP AKCAYA	8 days 06:03:07.416666	198.052060
CP ALANG-ALANG LEBAR	1 days 10:52:37.481012	34.877078
CP BABELAN	-1 days +04:32:20.955752	-19.460846
CP BALIKPAPAN UTARA	0 days 14:01:18.600000	14.021833
CP BANJARBARU SELATAN	0 days 13:20:28.269841	13.341186
...	...	...
CP WONOAYU	2 days 11:49:20.230769	59.822286
CP WONOGIRI	-10 days +19:18:53.843750	-220.685043
CP WONOMULYO	0 days 13:00:50.583333	13.014051
CP WONOSARI	0 days 21:00:22.545454	21.006263
CP WONOSOBO	1 days 07:31:07.733333	31.518815

Code to Measure mean for each Branch:

```
data_CP= master_data.groupby('POD Branch')['Rentang P-P'].mean(numeric_only=False)
df = pd.DataFrame(data_CP)
print(df)
```

Data Visualization

Data Engineer Test

Pesebaran Pengiriman

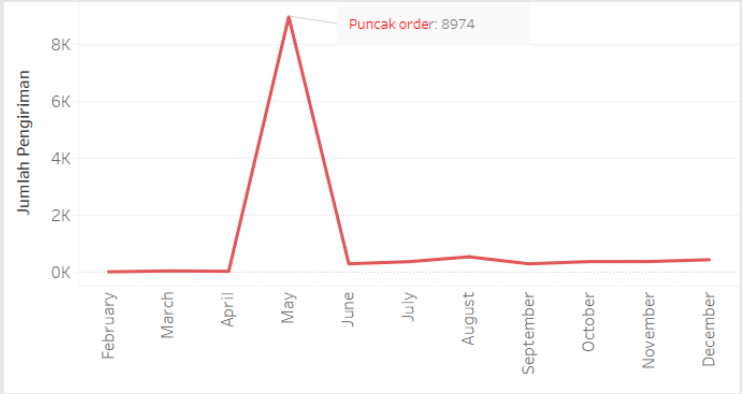
FILTER =

Destination
All

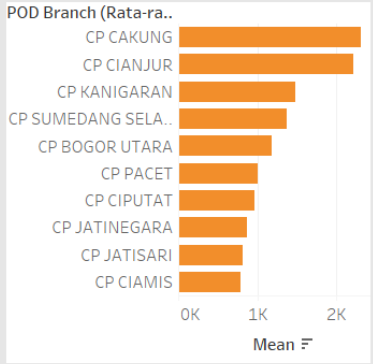
POD Branch1
All

Limit
Top 15 by COUNT([POD Branch1])

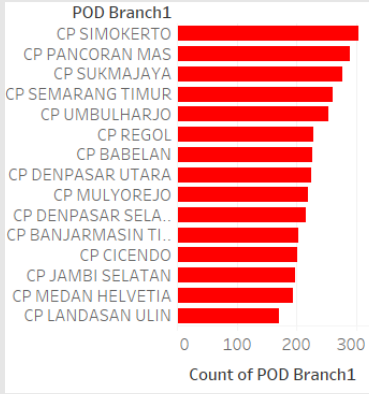
Order Per-Bulan(Month)



Rata2 Interval waktu Per Cabang(Hours)



Jumlah Order Per-Cabang



Problem While do this Project

Problem make and Analyze the data:

1. Confuse to merge and choose the data. (what a columns must I used)
2. have Problem in Indentify the data
3. sometimes need thinking more than twice for make simply data to next analyze
4. have problem to operation with different data types
5. Have bad internet
6. need to much time to decides what code I used.

Problem make a Visualization Data:

1. need more time to decides what visualization I used
2. Have problem to connect more than one data
3. to much data can't read by the maps Graphic (Longitude, Latitude)

THANK YOU

Contact Me

https://github.com/MRHalawi12

rizkyhalawi@gmail.com

DRAMAGA, BOGOR, INDONESIA