HIGO: To Be Data Scientist Challenge

Muhammad Faurel Gema Augista

Faurel Gema

Requirement

- Buatlah sample data untuk bahan analisis
- Buatkan Analisis Data tersebut berdasarkan variable berikut ini: **Nama Lokasi, Jam Login, Nama, Email, No Telp, Tahun Lahir, Merek HP, Digital Interest, Location Type**

- Kembangkan variable diatas menjadi beberapa variable baru (Nilai Plus)

- Membuat System Confidence Interval (Nilai Plus)

Sample Data

I am using Faker, python library for generate data randomly

We union with new random column:

1. Login timestamp
2. Phone number
3. Type_of_device
4. Digital_interest
5. Location_type

And basically, there are more columns:

1. Job
2. company
3. Ssn
4. Residence
5. current_location
6. blood_group
7. Website
8. Username
9. Name
10. Sex
11. Address (unused, so we decided not to use this basic column)
12. Mail
13. birthdate

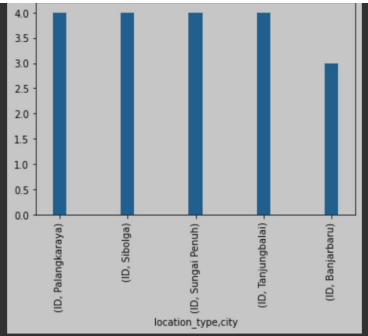
Total New Variable after development

Source Column	New Column
1. residence	1. address (New column and get from column residence before \n 2. city (get from residence after \n) 3. province (get from residence)
2. birthdate	4. age (generate age from birthdate)
3. login_timestamp	5. login (get from day of login_timestamp)
4. age	6. age_group (kelompok usia)
5. mail	7. email_group

Total 7 new variable

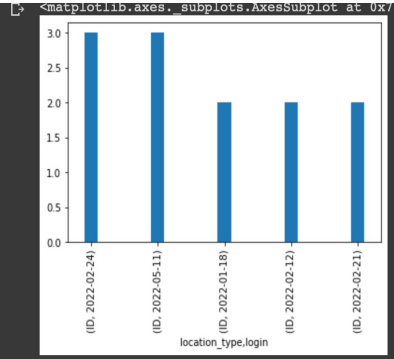
Variable Analysis

Name of location and location type (City with Most internet users)



first analysis 1: City with most internet users are Palangkaraya, Sibolga, Sungai penuh and tanjungbalai and they're have same values

Login timestamp



the most highest traffic internet is in 24 February 2022 and 11 May 2022, both have the same value

Variable Analysis

Merek hp (disini saya gunakan type_of_device)

```
[15] df.groupby(["type_of_device"])["id"].count().reset_index(name="total_device")
```

	type_of_device	total_device
0	android	16
1	ipad	15
2	iphone	23
3	mac	20
4	tablet	26

Tablet is the most device that people using for accessing internet

Digital Interest

```
[16] df.groupby(["digital_interest"])["id"].count().reset_index(name="total_people_love")
```

	digital_interest	total_people_love
0	ecommerce	16
1	music	20
2	science	25
3	social media	17
4	travelling	22

```
df.groupby(['digital_interest']).location_type.value_counts().nlargest(3)
```

```
digital_interest  location_type
science           ID              25
travelling        ID              22
music             ID              20
Name: location_type, dtype: int64
```

There is science that people love the most. aren't we all?

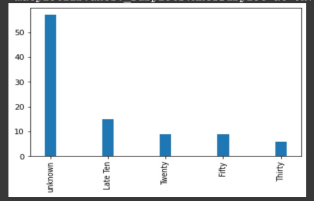
Variable Analysis

Tahun (saya kembangkan jadi age of group)

	age_group	total_people_by_age
0	Late Ten	15
1	Twenty	9
2	Thirty	6
3	Fourty	4
4	Fifty	9
5	unknown	57

```
df.age_group.value_counts().nlargest(5).plot(kind="bar", stacked=True, width=0.2)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fefc1ebbad0>
```



there is unknown group of age that most using internet. and followed by late ten and twenty

Analisa di column name (saya menemukan gelar yang paling banyak digunakan)

Analysis of Name

```
name = df['name']
new_name = []
for n in name:
  new_name.extend(n.split())
  total_name = {i:new_name.count(i) for i in new_name}
print(total_name)
```

```
{'Eva': 1, 'Maryati': 2, 'Kasiran': 1, 'Tarihoran,': 1, 'M.Kom.': 1, 'R.': 2, 'Irma': 1, 'Widodo,': 1, 'S.T
```

```
import heapq
heapq.nlargest(5, total_name, key=total_name.get)
```

```
['Drs.', 'S.Psi', 'KH.', 'S.Ked', 'M.Ak']
```

There are more people with Drs title, and the rest is S.Psi, KH, S.Ked and M.Ak

Variable Analysis

Email (most used mail provider)

```
[ ] df['email_group'] = [i.rsplit('@', 1)[1] for i in df['mail']]
```

```
df['email_group']
```

```
0       gmail.com
1       gmail.com
2       yahoo.com
3     hotmail.com
4       gmail.com
          ...
95      gmail.com
96      gmail.com
97      yahoo.com
98    hotmail.com
99    hotmail.com
Name: email_group, Length: 100, dtype: object
```

```
df.groupby(["email_group"])["id"].count().reset_index(name="total_email_protocol")
```

	email_group	total_email_protocol
0	gmail.com	34
1	hotmail.com	39
2	yahoo.com	27

Hmm... hotmail.com is most used email provider here.

System Confidence Interval

```
import numpy as np
import scipy.stats as st

data_test = df['age']

# create 90% confidence interval
st.t.interval(alpha=0.90, df=len(data_test)-1,
              loc=np.mean(data_test),
              scale=st.sem(data_test))
```

```
(55.63160638300809, 66.1883936169919)
```

So with this system confidence result, I can take the conclusion, i really 90% confidence that mean from these ages from dataset have a range between 55 to 66.

Conclusion

1. I can take the conclusion, i really 90% confidence that mean from these ages from dataset have a range between 55 to 66.

2. The most people with Drs, S.Psi, KH, S.Ked and M.Ak title, mostly with age older than 55 years old and they're really love science and travelling

3. They are moslty using tablet because have the eye problem. the wide kind of tablet screen helped them a lot to find information about science or seeing beautiful place or travelling memories

Thank you so much!

Click here to see the code at github