```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Go to this URL in a browser: https://accounts.google.com/o/oauth2/auth?client_id=

Enter your authorization code:
··········
Mounted at /content/drive
```

```
import os
os.chdir('/content/drive/My Drive/subcription_project/')
!ls
```

```
public_sharing_collect_data_for_subcription.csv  subcription.ipynb
```

```
data = pd.read_csv("./public_sharing_collect_data_for_subcription.csv")
data.head()
```

	id	order_time	service_id	distance	total_cod	total_fee	total
0	NNFNT7	2019-02-04 14:43:10 UTC	SGN-BIKE	6.302000	2400000.0	83000	83
1	L8HYZC	2019-02-04 13:19:44 UTC	SGN-BIKE	5.818000	3600000.0	93000	93
2	M0PC9M	2019-02-04 12:27:04 UTC	SGN-BIKE	9.673000	3400000.0	120000	120
3	NUE89W	2019-02-04 18:27:53 UTC	SGN-BIKE	17.066999	450000.0	134000	134
4	4I19P6	2019-02-04 15:44:46 UTC	SGN-BIKE	14.135000	280000.0	121000	121

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3024981 entries, 0 to 3024980
Data columns (total 7 columns):
id            object
order_time    object
service_id    object
distance      float64
total_cod     float64
total_fee     int64
total_pay     float64
dtypes: float64(3), int64(1), object(3)
memory usage: 161.6+ MB
```

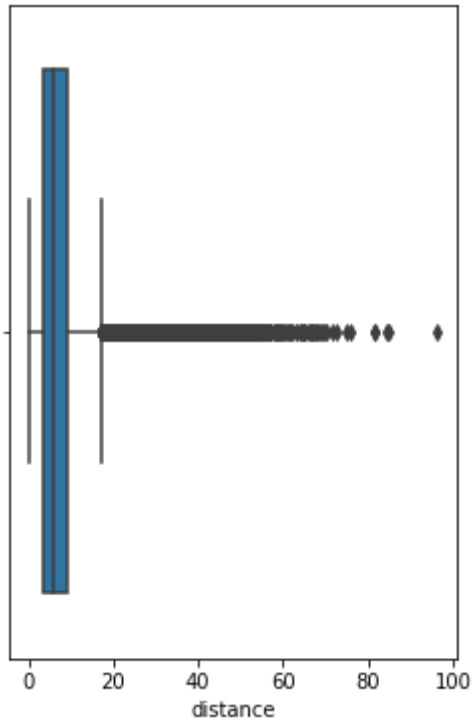
```
data['order_time'] = data['order_time'].astype('datetime64[ns]')


# import datatime as dt
# data['order_time'] =
# data['order_time'].dt.date

# pd.to_datetime(data['order_time'], format='%Y%m%d')
# data['order_date'] =  pd.to_datetime(data['order_time'])


data.info()
```

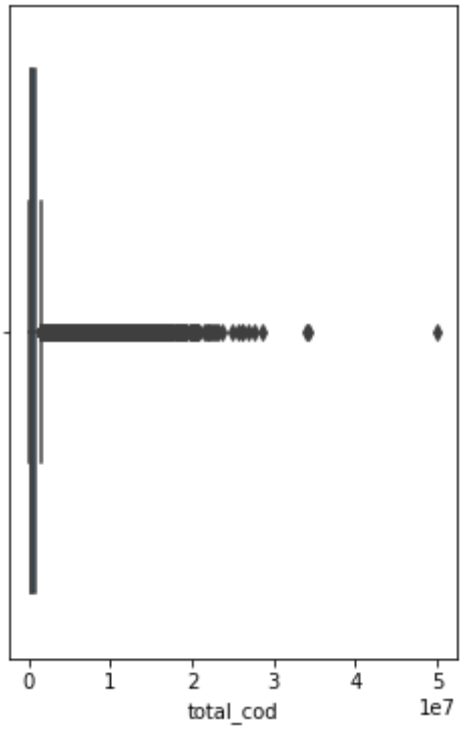
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3024981 entries, 0 to 3024980
Data columns (total 7 columns):
id            object
order_time    datetime64[ns]
service_id    object
distance      float64
total_cod     float64
total_fee     int64
total_pay     float64
dtypes: datetime64[ns](1), float64(3), int64(1), object(2)
memory usage: 161.6+ MB
```

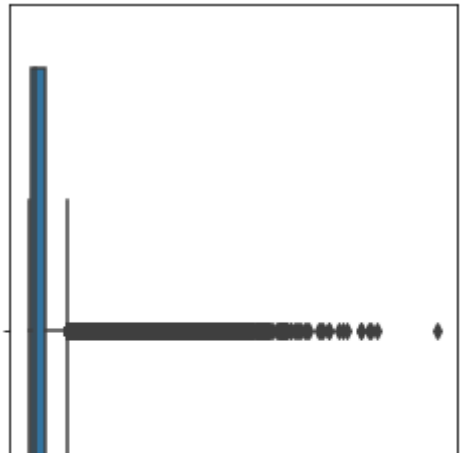
```
# data[data['total_cod']!=0]
```

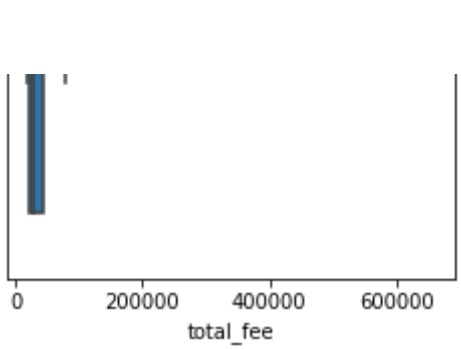
EDA

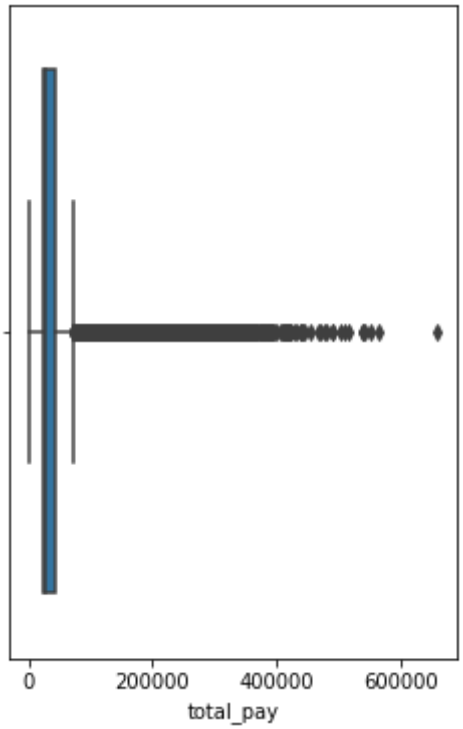
```
for col in data.iloc[:,3:7].columns:
  plt.figure(figsize=(4,6))
  sns.boxplot(data[col])
  plt.show()
```


0
20
40
60
80
100
distance

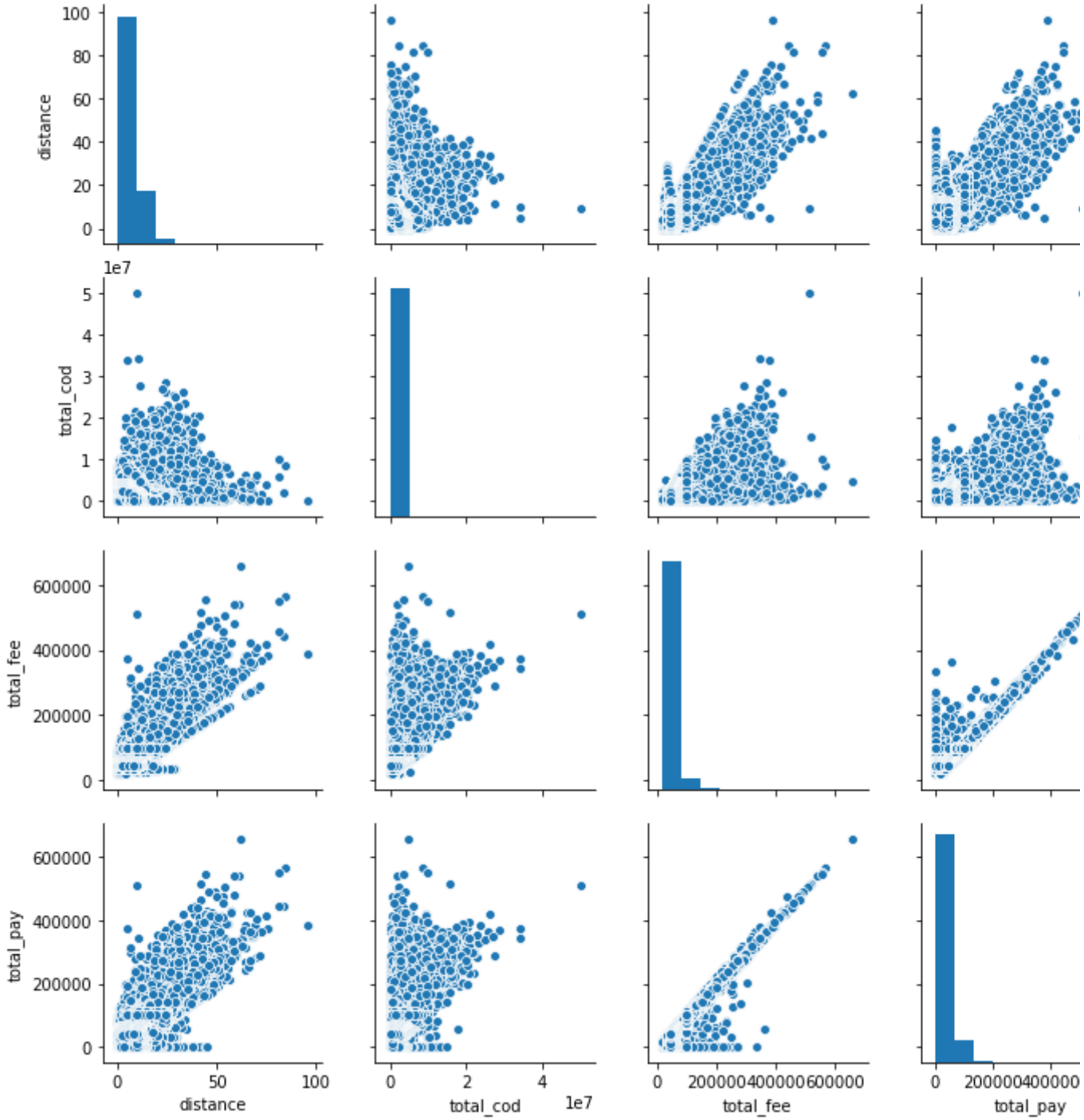

0
1
2
3
4
5
1e7
total_cod








```
sns.pairplot(data.iloc[:,3:7])
plt.show()
```




```
pd.set_option('display.float_format', lambda a: '%.2f' % a)
data.iloc[:,3:7].describe()
```

	distance	total_cod	total_fee	total_pay
count	3024981.00	3024981.00	3024981.00	3024981.00
mean	6.70	608087.43	38077.34	36553.83
std	4.63	752927.51	23506.99	23064.73
min	0.00	0.00	18000.00	0.00
25%	3.45	190000.00	23000.00	23000.00
50%	5.76	379000.00	29000.00	28000.00
75%	8.87	740000.00	45000.00	43000.00
max	96.13	49932000.00	659000.00	659000.00

```
data['ratio_fee_cod'] = data['total_fee']/data['total_cod']
data['ratio_pay_cod'] = data['total_pay']/data['total_cod']
```

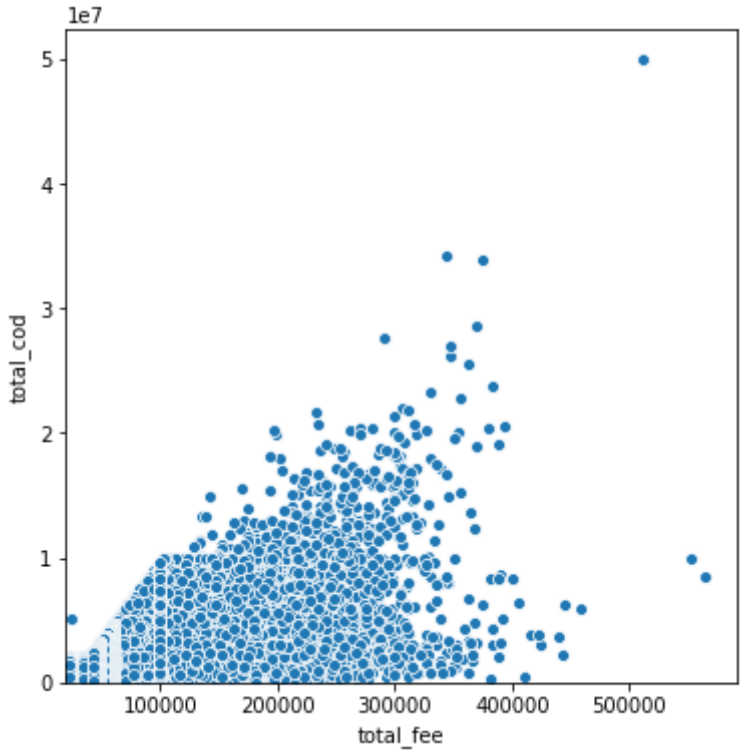
```
data.head()
```

	id	order_time	service_id	distance	total_cod	total_fee	total_pay
0	NNFNT7	2019-02-04 14:43:10	SGN-BIKE	6.30	2400000.00	83000	83000.00
1	L8HYZC	2019-02-04 13:19:44	SGN-BIKE	5.82	3600000.00	93000	93000.00
2	M0PC9M	2019-02-04 12:27:04	SGN-BIKE	9.67	3400000.00	120000	120000.00
3	NUE89W	2019-02-04 18:27:53	SGN-BIKE	17.07	450000.00	134000	134000.00
4	4I19P6	2019-02-04 15:44:46	SGN-BIKE	14.14	280000.00	121000	121000.00

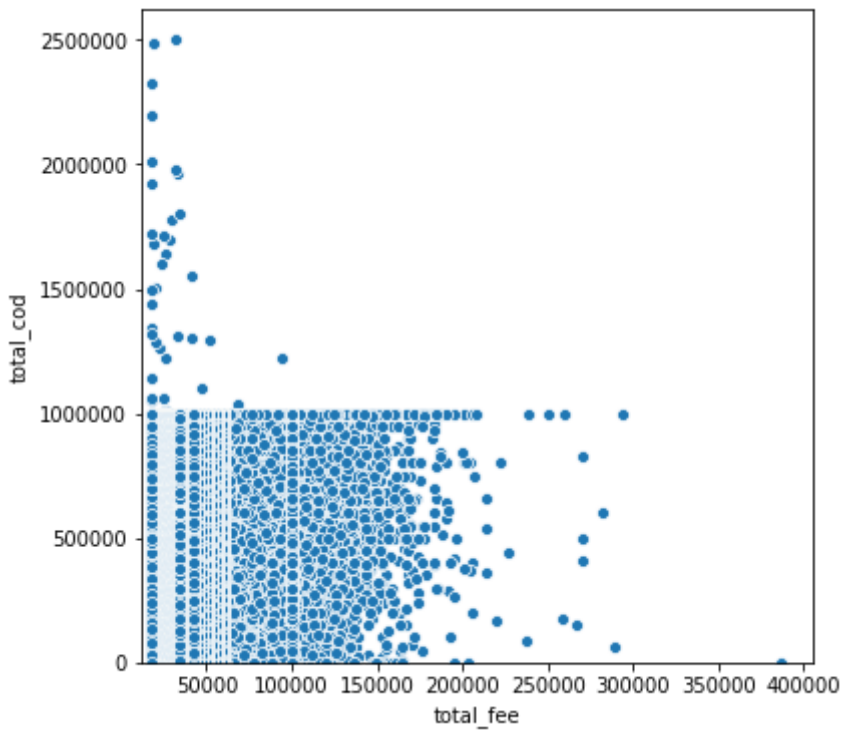
```
# data[data['ratio_fee_cod']>1]
data.service_id.value_counts()
# data[(data['service_id'] == 'SGN-BIKE')|(data['service_id'] == 'SGN-LUX')]
```

```
SGN-BIKE       1863612
SGN-POOL        975863
SGN-SAMEDAY     150813
SGN-DG           29255
SGN-LUX           5438
Name: service_id, dtype: int64
```

```
# sns.set(style='whitegrid')
plt.figure(figsize=(6,6))
ax = sns.scatterplot(x="total_fee", y="total_cod",palette='muted',data= data[(data['se
ax.set(xlim=(20000,None),ylim=(0,None))
ax.grid(False)
plt.show()
```

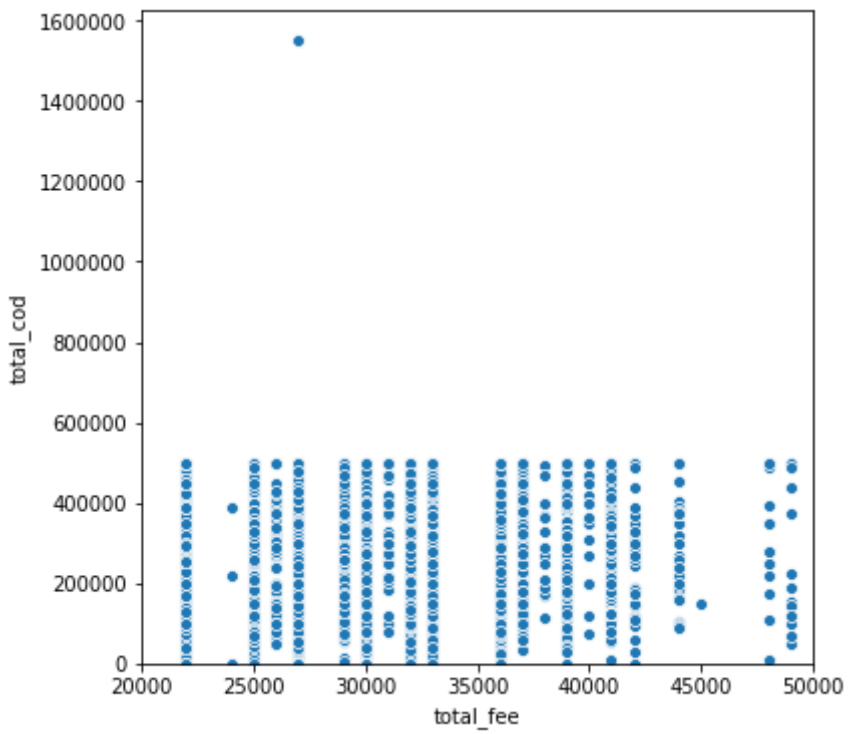



```
plt.figure(figsize=(6,6))
ax = sns.scatterplot(x="total_fee", y="total_cod", data=data[data['service_id'] == 'SC
ax.set(xlim=(12000,None),ylim=(0,None))
ax.grid(False)
plt.show()
```

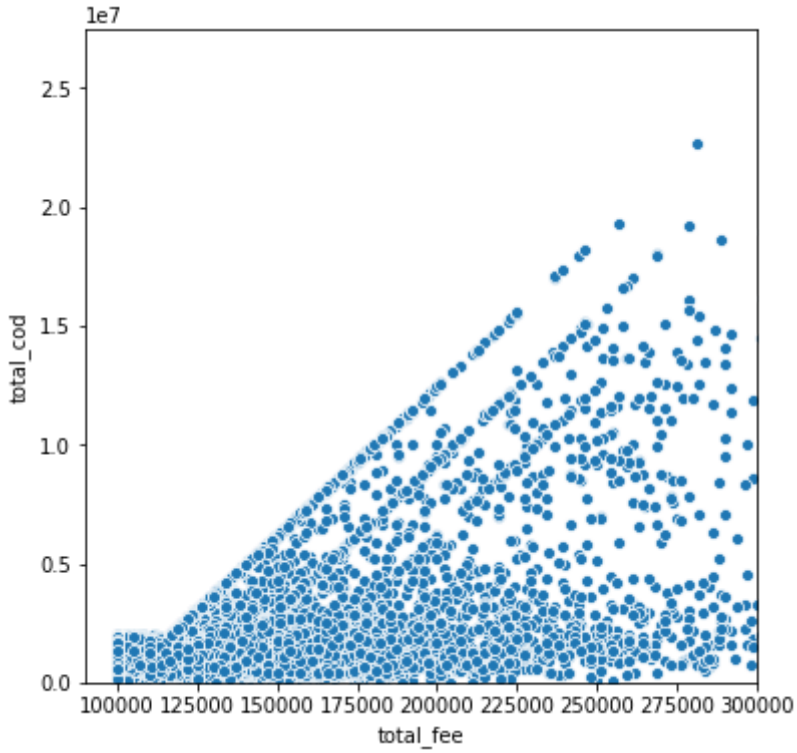



```
plt.figure(figsize=(6,6))
ax = sns.scatterplot(x="total_fee", y="total_cod", data=data[data['service_id'] == 'SC
```

```
ax.set(xlim=(20000,50000),ylim=(0,None))
ax.grid(False)
plt.show()
```

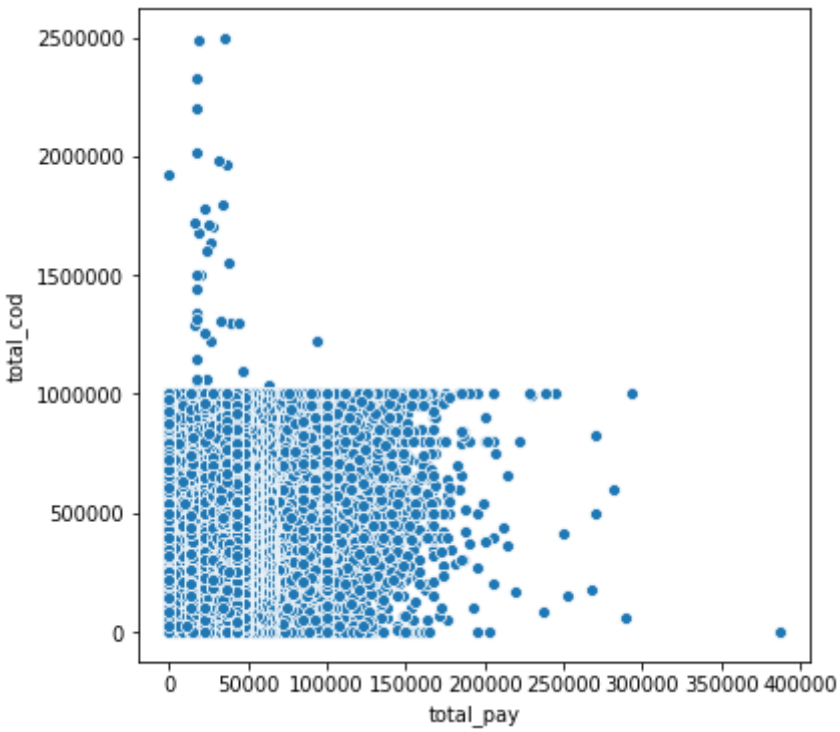



```
plt.figure(figsize=(6,6))
ax = sns.scatterplot(x="total_fee", y="total_cod", data=data[data['service_id'] == 'SC
ax.set(xlim=(90000,300000),ylim=(0,None))
ax.grid(False)

# plt.ylim(1000000,5000000)
# plt.xlim(20000,50000)
# ax.set(ylim=(10, 40))
plt.show()
```

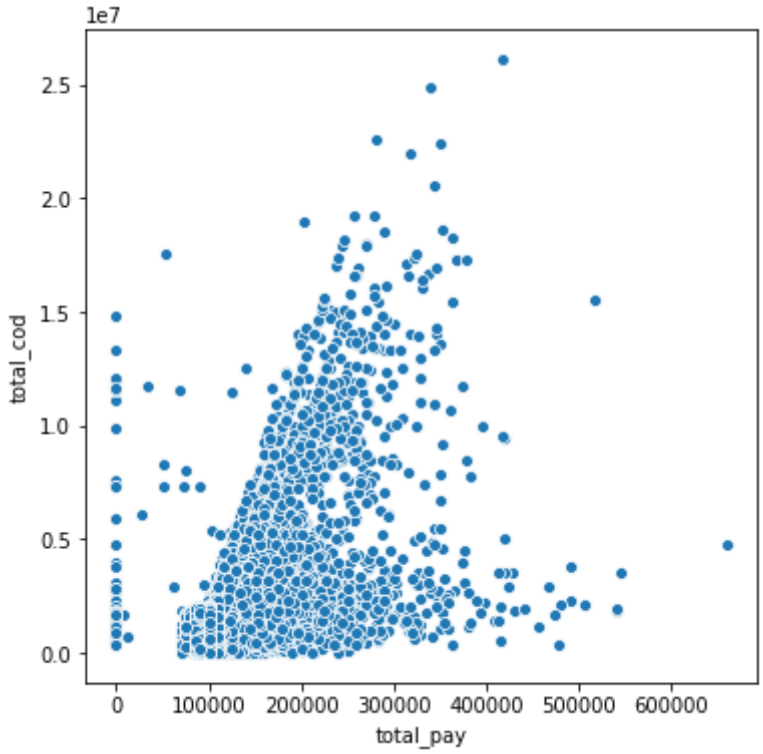



```
## total pay ??


cmap = sns.cubehelix_palette(dark=.3, light=.8, as_cmap=True)
plt.figure(figsize=(6,6))
ax = sns.scatterplot(x="total_pay", y="total_cod", data= data[(data['service_id'] == '
ax.grid(False)
plt.show()
```

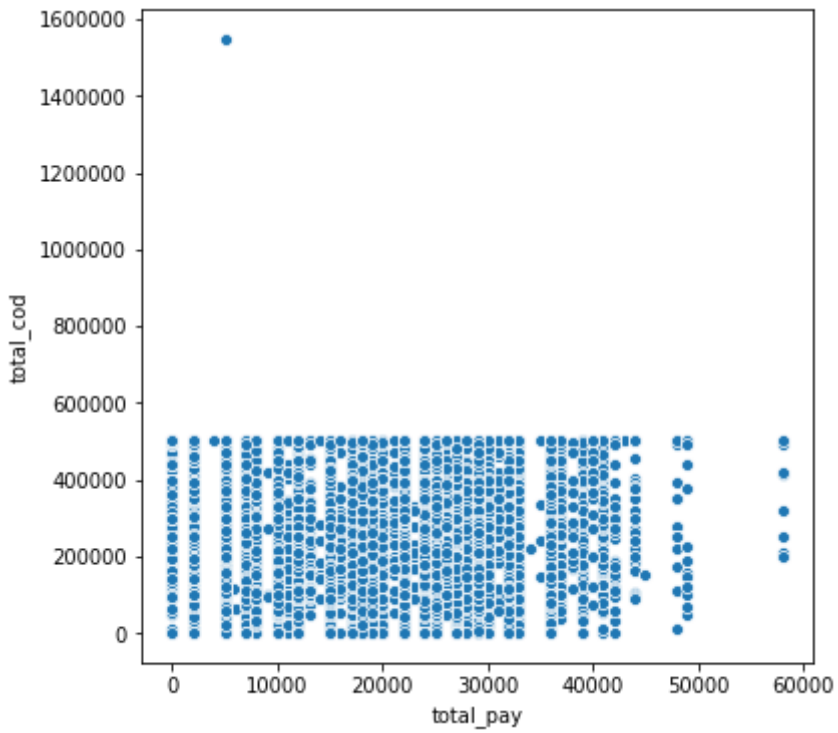
```
plt.figure(figsize=(6,6))
ax = sns.scatterplot(x="total_pay", y="total_cod", data= data[(data['service_id'] ==
ax.grid(False)
plt.show()
```




```
plt.figure(figsize=(6,6))
ax = sns.scatterplot(x="total_pay", y="total_cod", data= data[(data['service_id'] ==
ax.grid(False)
plt.show()
```




```
plt.figure(figsize=(6,6))
ax = sns.scatterplot(x="total_pay", y="total_cod", data= data[(data['service_id'] == '
ax.grid(False)
plt.show()
```



Ket luan1:

Các scatter của total_cod & total_pay cũng như total_cod & total_fee trên các dịch vụ là rất khác biệt. có đặc trưng về COD như sau BIKE&LUX, POOL, DG, SD .

```
##
bins = [0,0.05,0.1,0.15,0.2,np.inf]
labels = ['0-0.05','0.05-0.1','0.1-0.15','0.15-0.2','>0.2']
data['range_feeCod_BL'] = pd.cut(data[(data['service_id'] == 'SGN-BIKE')|(data['servic

data['range_feeCod_PL'] = pd.cut(data[(data['service_id'] == 'SGN-POOL')]['ratio_fee_c

data['range_feeCod_DG'] = pd.cut(data[(data['service_id'] == 'SGN-DG')]['ratio_fee_coc

data['range_feeCod_SD'] = pd.cut(data[(data['service_id'] == 'SGN-SAMEDAY')]['ratio_fe

# data[['range_BL','range_PL','range_DG','range_SD']].astype(float)
# data.info()

data[['range_feeCod_BL','range_feeCod_PL','range_feeCod_DG','range_feeCod_SD']].isna()
```

```
range_feeCod_BL    False
range_feeCod_PL    False
range_feeCod_DG    False
range_feeCod_SD    False
dtype: bool
```

```
range_BL = data[(data['service_id'] == 'SGN-BIKE')|(data['service_id'] == 'SGN-LUX')]|
# .groupby('range_BL')
print(range_BL.values/range_BL.sum()*100)
range_BL = pd.DataFrame(range_BL)
range_BL
```

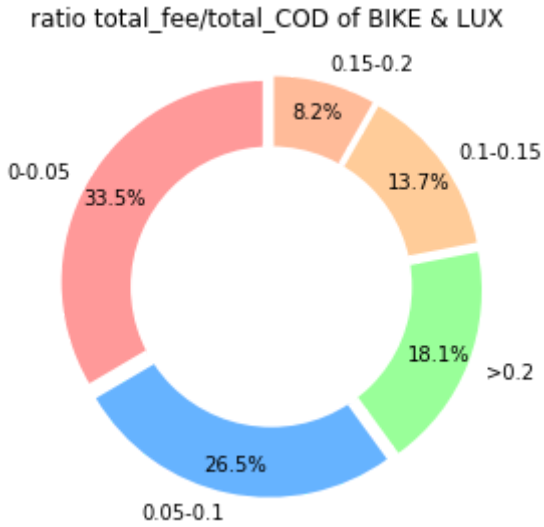
```
[33.45999304 26.5029293  18.103047   13.73927931  8.19475134]
```

	range_feeCod_BL
0-0.05	625384
0.05-0.1	495353
>0.2	338355
0.1-0.15	256794
0.15-0.2	153164

```
# Pie chart
labels = range_BL.index.values
```

```
sizes = range_BL.values
#colors
colors = ['#ff9999','#66b3ff','#99ff99','#ffcc99','#ffbb99']
#explsion
explode = (0.05,0.05,0.05,0.05,0.05)

ax1 = plt.pie(sizes, colors = colors, labels=labels, autopct='%1.1f%%', startangle=90,
#draw circle
centre_circle = plt.Circle((0,0),0.70,fc='white')
fig = plt.gcf()
fig.gca().add_artist(centre_circle)
plt.title("ratio total_fee/total_COD of BIKE & LUX ")
# Equal aspect ratio ensures that pie is drawn as a circle
# ax1.axis('equal')
plt.tight_layout()
plt.show()
```

```
/usr/local/lib/python3.6/dist-packages/ipykernel_launcher.py:8: MatplotlibDepreca
```

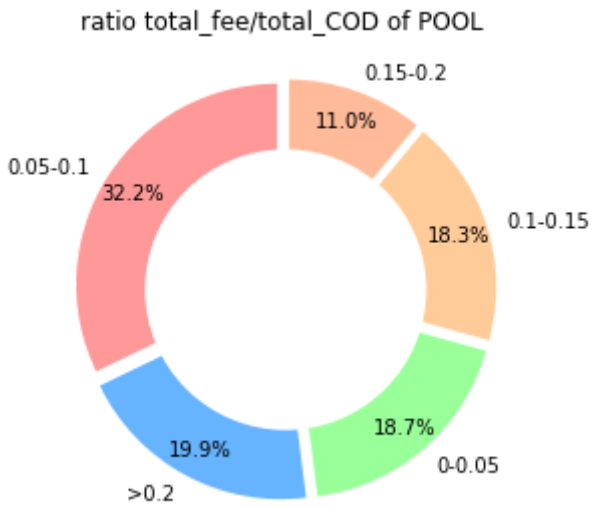



```
range_PL = data[(data['service_id'] == 'SGN-POOL')]['range_feeCod_PL'].value_counts()
print(range_PL.values/range_PL.sum()*100)
range_PL = pd.DataFrame(range_PL)
range_PL
```

```
[32.15277144 19.85135209 18.69616944 18.33904964 10.96065739]
```

```
# Pie chart
labels = range_PL.index.values
sizes = range_PL.values
#colors
colors = ['#ff9999','#66b3ff','#99ff99','#ffcc99','#ffbb99']
#explsion
explode = (0.05,0.05,0.05,0.05,0.05)

ax1 = plt.pie(sizes, colors = colors, labels=labels, autopct='%1.1f%%', startangle=90,
#draw circle
centre_circle = plt.Circle((0,0),0.70,fc='white')
fig = plt.gcf()
fig.gca().add_artist(centre_circle)
plt.title("ratio total_fee/total_COD of POOL ")
# Equal aspect ratio ensures that pie is drawn as a circle
# ax1.axis('equal')
plt.tight_layout()
plt.show()
```

```
/usr/local/lib/python3.6/dist-packages/ipykernel_launcher.py:8: MatplotlibDepreca
```



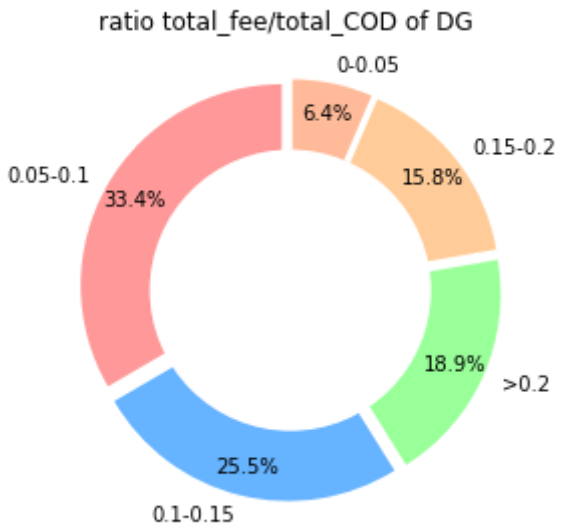

```
range_DG = data[(data['service_id'] == 'SGN-DG')]['range_feeCod_DG'].value_counts()
print(range_DG.values/range_DG.sum()*100)
range_DG = pd.DataFrame(range_DG)
range_DG
```

```
[33.39258246 25.49649633 18.86515126 15.79900872  6.44676124]
```

	range_feeCod_DG
0.05-0.1	9769
0.1-0.15	7459
>0.2	5519
0.15-0.2	4622
0-0.05	1886

```
# Pie chart
labels = range_DG.index.values
sizes = range_DG.values
#colors
colors = ['#ff9999','#66b3ff','#99ff99','#ffcc99','#ffbb99']
#explsion
explode = (0.05,0.05,0.05,0.05,0.05)

ax1 = plt.pie(sizes, colors = colors, labels=labels, autopct='%1.1f%%', startangle=90,
#draw circle
centre_circle = plt.Circle((0,0),0.70,fc='white')
fig = plt.gcf()
fig.gca().add_artist(centre_circle)
plt.title("ratio total_fee/total_COD of DG ")
# Equal aspect ratio ensures that pie is drawn as a circle
# ax1.axis('equal')
plt.tight_layout()
plt.show()
```

```
/usr/local/lib/python3.6/dist-packages/ipykernel_launcher.py:8: MatplotlibDepreca
```



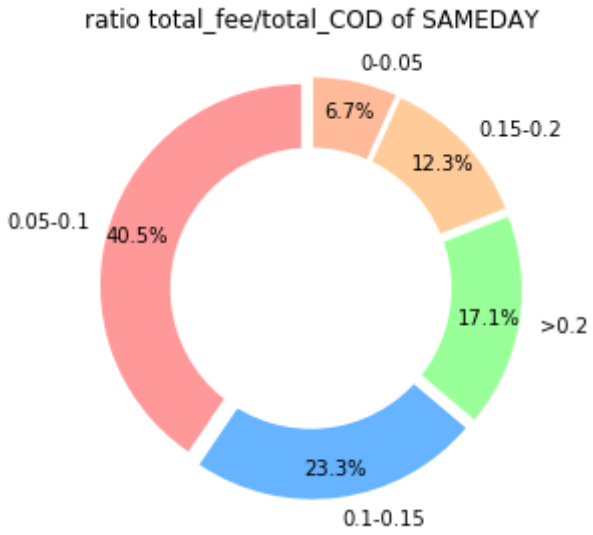
```
range_SD = data[(data['service_id'] == 'SGN-SAMEDAY')]['range_feeCod_SD'].value_counts
print(range_SD.values/range_SD.sum()*100)
range_SD = pd.DataFrame(range_SD)
range_SD
```

```
[40.50844423 23.32955382 17.08075564 12.33381738  6.74742894]
```

	range_feeCod_SD
0.05-0.1	61092
0.1-0.15	35184
>0.2	25760
0.15-0.2	18601
0-0.05	10176

```
# Pie chart
labels = range_SD.index.values
sizes = range_SD.values
#colors
colors = ['#ff9999','#66b3ff','#99ff99','#ffcc99','#ffbb99']
#explsion
explode = (0.05,0.05,0.05,0.05,0.05)

ax1 = plt.pie(sizes, colors = colors, labels=labels, autopct='%1.1f%%', startangle=90,
#draw circle
centre_circle = plt.Circle((0,0),0.70,fc='white')
fig = plt.gcf()
fig.gca().add_artist(centre_circle)
plt.title("ratio total_fee/total_COD of SAMEDAY")
# Equal aspect ratio ensures that pie is drawn as a circle
# ax1.axis('equal')
plt.tight_layout()
plt.show()
```

```
/usr/local/lib/python3.6/dist-packages/ipykernel_launcher.py:8: MatplotlibDepreca
```



Ket Luan 2: ratio of total_fee/ total_cod

`BIKE & LUX` : khoảng 0 -5% là phổ biến nhất với 32%, 5% -10%: chiếm 26%

`SAMEDAY` : khoảng 5%-10% thì chiếm 40%, khoảng 10%-15% chiếm 23.3 %, khoảng lớn hơn 0.2% chiếm

=> Bốc riêng sameday để phân tích xem khoảng mean của total_pay and total_fee

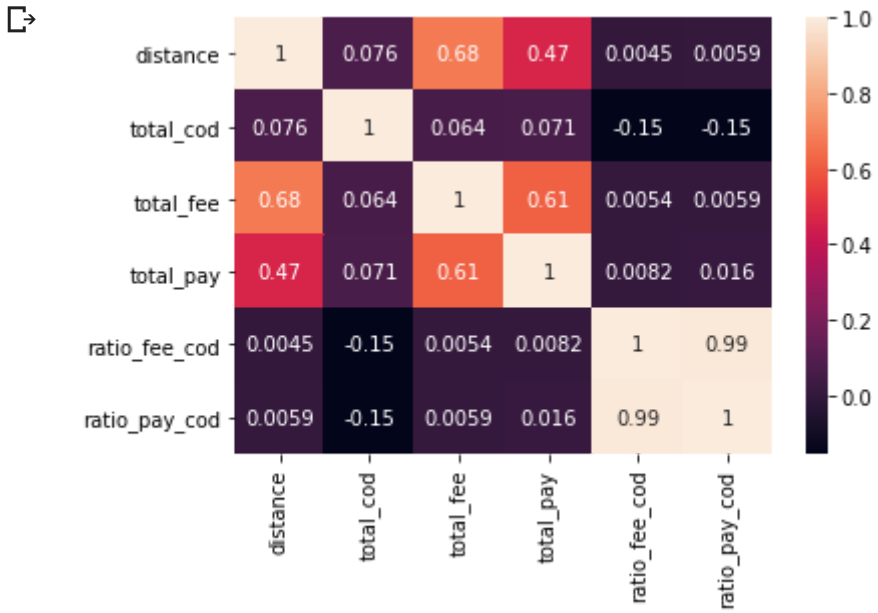
```
data[(data['service_id'] == 'SGN-SAMEDAY')].groupby(['range_feeCod_SD'])[['total_fee',
```

range_feeCod_SD	total_fee	total_pay	total_cod
0-0.05	22012.58	20618.42	481778.46
0.05-0.1	24078.75	22708.03	338841.83
0.1-0.15	24233.86	22749.57	197557.19
0.15-0.2	24483.74	22771.03	141364.24
>0.2	24774.77	22869.79	81993.03

```
corr = data[(data['service_id'] == 'SGN-SAMEDAY')].corr()
corr
```

	distance	total_cod	total_fee	total_pay	ratio_fee_cod	ratio_pay_
distance	1.00	0.08	0.68	0.47	0.00	0
total_cod	0.08	1.00	0.06	0.07	-0.15	-0
total_fee	0.68	0.06	1.00	0.61	0.01	0
total_pay	0.47	0.07	0.61	1.00	0.01	0
ratio_fee_cod	0.00	-0.15	0.01	0.01	1.00	0
ratio_pay_cod	0.01	-0.15	0.01	0.02	0.99	1

```
ax = sns.heatmap(corr,annot=True,fmt='.2g')
bottom, top = ax.get_ylim()
ax.set_ylim(bottom + .5, top - .5)
plt.show()
```




```
from IPython.display import HTML
HTML('''<script>
code_show=true;
function code_toggle() {
 if (code_show){
 $('div.input').hide();
 } else {
```

```
 $('div.input').show();
 }
 code_show = !code_show
}
$( document ).ready(code_toggle);
</script>
The raw code for this IPython notebook is by default hidden for easier reading.
To toggle on/off the raw code, click <a href="javascript:code_toggle()">here</a>.''')
```

The raw code for this IPython notebook is by default hidden for easier reading. To toggle on/off the raw code, c