pd1

March 9, 2022

1 Praca domowa 1

1.1 EDA

1.1.1 Paulina Jaszczuk

1.1.2 Import pakietów

[48]:
```
import pandas as pd
import numpy as np
import matplotlib
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
np.random.seed(23)

# ustawia domyślną wielkość wykresów
plt.rcParams['figure.figsize'] = (8,6)
# to samo tylko dla tekstu
plt.rcParams['font.size'] = 16
# ustawia wielkość tekstów dla wykresów seaborn zależną od wielkości wykresu
sns.set_context('paper', font_scale=1.4)
```

1.1.3 Wczytanie danych i wstępne informacje

[8]:
```
data = pd.read_csv("C:\\Users\\pauli\\Downloads\\wb\\students.csv")
data.shape
```

[8]:
```
(1044, 34)
```

Zadaniem jest eksploracja zbioru dotyczącego uczniów [link]. Nasze dane zawierają 1044 obserwacje oraz 34 zmienne objaśniające - 31 z nich to różne dane charakteryzujące ucznia, między innymi relacje rodzinne, czas spędzany na nauce czy ilość spożywanego alkoholu. Pozostałe trzy to oceny - na pierwszy semestr, drugi i ocena całoroczna. Finalnym zadaniem jest predykcja właśnie tej ostatniej.

[11]:
```
data.head(5)
```

[11]:

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	…	\
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	…	
1	GP	F	17	U	GT3	T	1	1	at_home	other	…	
2	GP	F	15	U	LE3	T	1	1	at_home	other	…	
3	GP	F	15	U	GT3	T	4	2	health	services	…	
4	GP	F	16	U	GT3	T	3	3	other	other	…	

	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3	class
0	3	4	1	1	3	6	5	6	6	math
1	3	3	1	1	3	4	5	5	6	math
2	3	2	2	3	3	10	7	8	10	math
3	2	2	1	1	5	2	15	14	15	math
4	3	2	1	2	5	4	6	10	10	math

[5 rows x 34 columns]

[9]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1044 entries, 0 to 1043
Data columns (total 34 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   school      1044 non-null   object
 1   sex         1044 non-null   object
 2   age         1044 non-null   int64
 3   address     1044 non-null   object
 4   famsize     1044 non-null   object
 5   Pstatus     1044 non-null   object
 6   Medu        1044 non-null   int64
 7   Fedu        1044 non-null   int64
 8   Mjob        1044 non-null   object
 9   Fjob        1044 non-null   object
 10  reason      1044 non-null   object
 11  guardian    1044 non-null   object
 12  traveltime  1044 non-null   int64
 13  studytime   1044 non-null   int64
 14  failures    1044 non-null   int64
 15  schoolsup   1044 non-null   object
 16  famsup      1044 non-null   object
 17  paid        1044 non-null   object
 18  activities  1044 non-null   object
 19  nursery     1044 non-null   object
 20  higher      1044 non-null   object
 21  internet    1044 non-null   object
 22  romantic    1044 non-null   object
 23  famrel      1044 non-null   int64
```

```
 24  freetime    1044 non-null   int64
 25  goout       1044 non-null   int64
 26  Dalc        1044 non-null   int64
 27  Walc        1044 non-null   int64
 28  health      1044 non-null   int64
 29  absences    1044 non-null   int64
 30  G1          1044 non-null   int64
 31  G2          1044 non-null   int64
 32  G3          1044 non-null   int64
 33  class       1044 non-null   object
dtypes: int64(16), object(18)
memory usage: 277.4+ KB
```

Jak widać dane nie zawierają braków. Dane objaśniające można podzielić na zmienne binarne (jest ich 14, m.in. `sex` czy `famsum` - wsparcie rodziny), nominalne (4; np. `Mjob` - praca matki), uporządkowane (11; np. `Medu` - wykształcenie matki) oraz zliczenia (2; np. `Age`). Zmienne zawierające dane o ocenach są również uporządkowanymi kolumnami numerycznymi.

[10]:
```
data.describe()
```

[10]:

	age	Medu	Fedu	traveltime	studytime
count	1044.000000	1044.000000	1044.000000	1044.000000	1044.000000
mean	16.726054	2.603448	2.387931	1.522989	1.970307
std	1.239975	1.124907	1.099938	0.731727	0.834353
min	15.000000	0.000000	0.000000	1.000000	1.000000
25%	16.000000	2.000000	1.000000	1.000000	1.000000
50%	17.000000	3.000000	2.000000	1.000000	2.000000
75%	18.000000	4.000000	3.000000	2.000000	2.000000
max	22.000000	4.000000	4.000000	4.000000	4.000000

	failures	famrel	freetime	goout	Dalc
count	1044.000000	1044.000000	1044.000000	1044.000000	1044.000000
mean	0.264368	3.935824	3.201149	3.156130	1.494253
std	0.656142	0.933401	1.031507	1.152575	0.911714
min	0.000000	1.000000	1.000000	1.000000	1.000000
25%	0.000000	4.000000	3.000000	2.000000	1.000000
50%	0.000000	4.000000	3.000000	3.000000	1.000000
75%	0.000000	5.000000	4.000000	4.000000	2.000000
max	3.000000	5.000000	5.000000	5.000000	5.000000

	Walc	health	absences	G1	G2
count	1044.000000	1044.000000	1044.000000	1044.000000	1044.000000
mean	2.284483	3.543103	4.434866	11.213602	11.246169
std	1.285105	1.424703	6.210017	2.983394	3.285071
min	1.000000	1.000000	0.000000	0.000000	0.000000
25%	1.000000	3.000000	0.000000	9.000000	9.000000
50%	2.000000	4.000000	2.000000	11.000000	11.000000
75%	3.000000	5.000000	6.000000	13.000000	13.000000

```
max       5.000000     5.000000    75.000000    19.000000    19.000000
```

```
                G3
count  1044.000000
mean     11.341954
std       3.864796
min       0.000000
25%      10.000000
50%      11.000000
75%      14.000000
max      20.000000
```

1.1.4 Zmienna celu - G3

[14]:
```
data["G3"]
```

[14]:
```
0        6
1        6
2       10
3       15
4       10
        ..
1039    10
1040    16
1041     9
1042    10
1043    11
Name: G3, Length: 1044, dtype: int64
```

[58]:
```
data['G3'].sort_values().unique()
```

[58]:
```
array([ 0,  1,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13, 14, 15, 16, 17, 18,
       19, 20], dtype=int64)
```

Brak ocen 2 i 3.

[60]:
```
data.G3.value_counts()
```

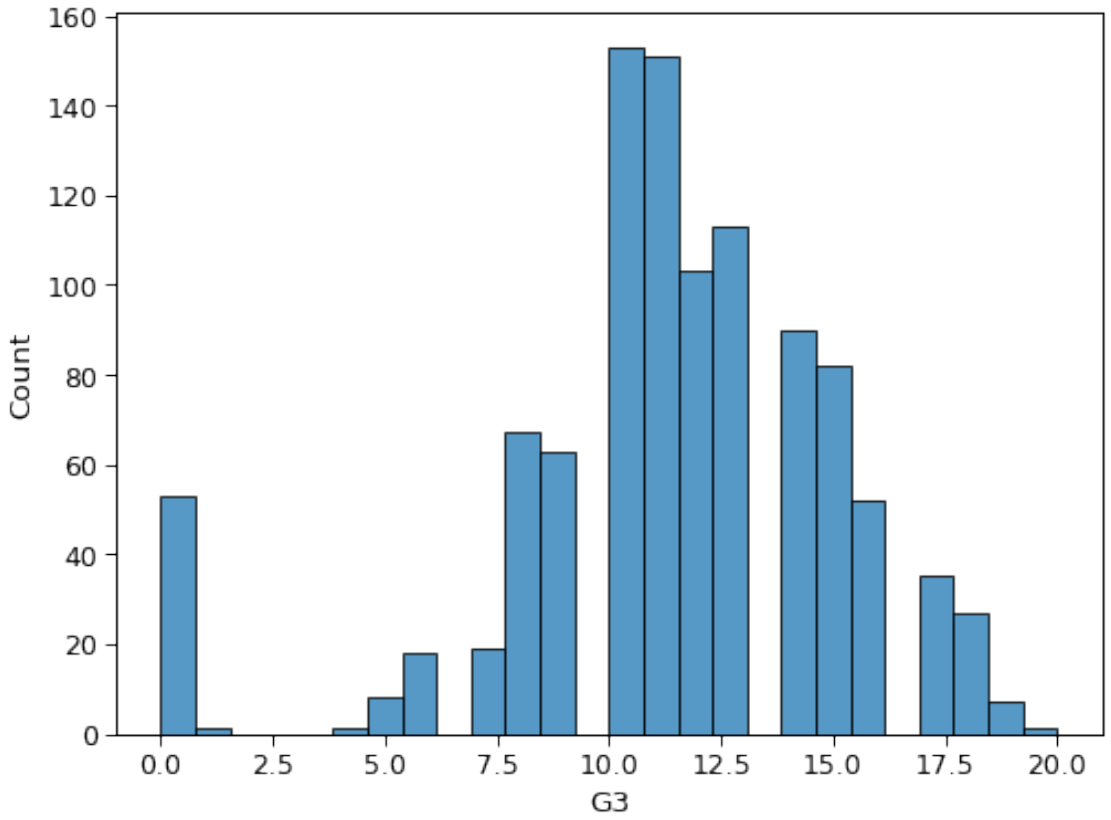
[60]:
```
10    153
11    151
13    113
12    103
14     90
15     82
8      67
9      63
0      53
```

```
16      52
17      35
18      27
7       19
6       18
5        8
19       7
4        1
1        1
20       1
Name: G3, dtype: int64
```

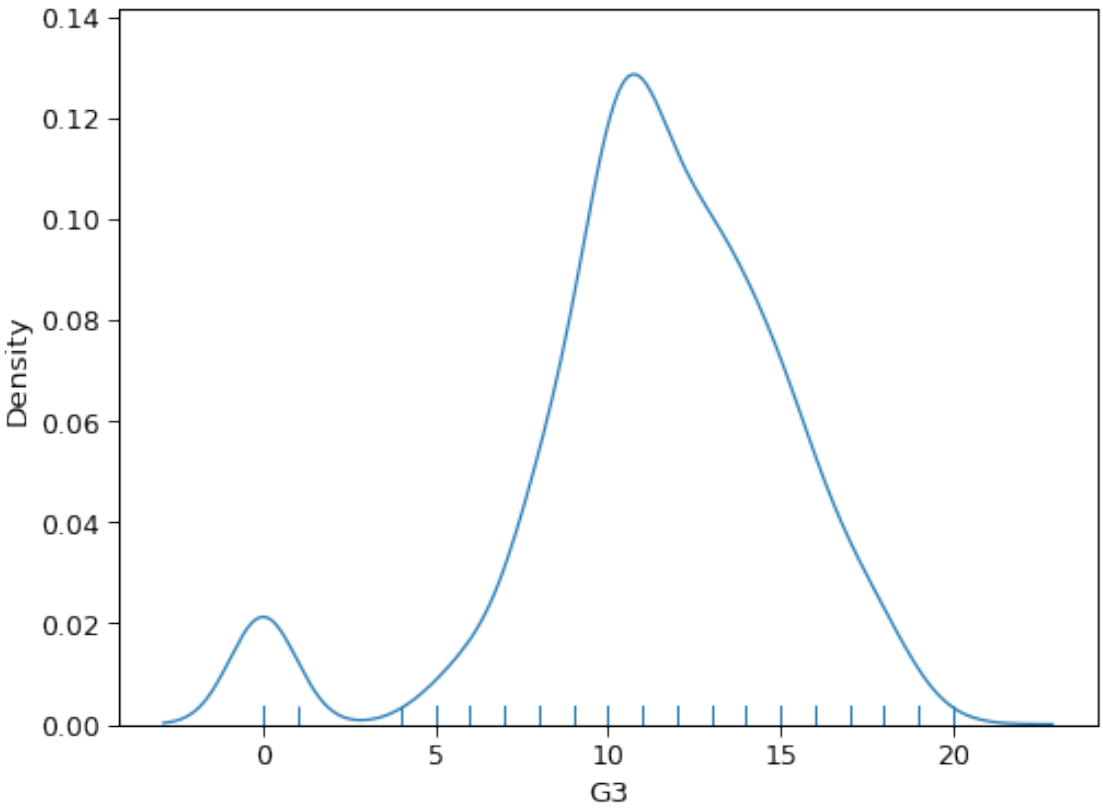
Najwięcej 10, najmniej 1 i 20.

[59]:
```
sns.histplot(data['G3'])
plt.plot()
```

[59]:
```
[]
```



[19]:
```
sns.distplot(data['G3'],rug=True, hist=False);
```



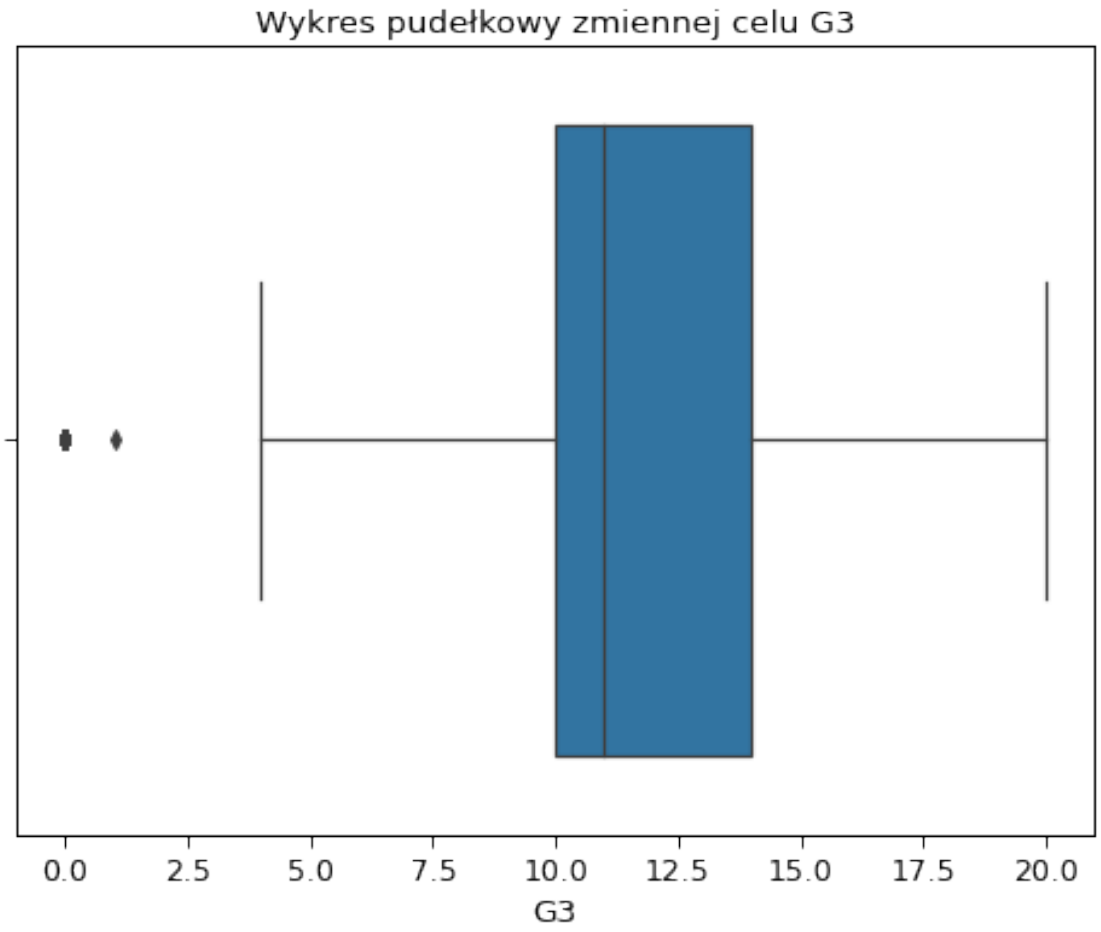
Rozkład zmiennej celu przypomina nieco rozkład normalny, ale ma pik w zerze (brak oceny - pzerwanie nauki?) i dłuższy lewy ogon.

[22]:
```
data['G3'].skew()
```

[22]:
```
-0.9859646596265084
```

Skośność potwierdza nam wrażenia wizualne dot. tego, że rozkład jest lewoskośny.

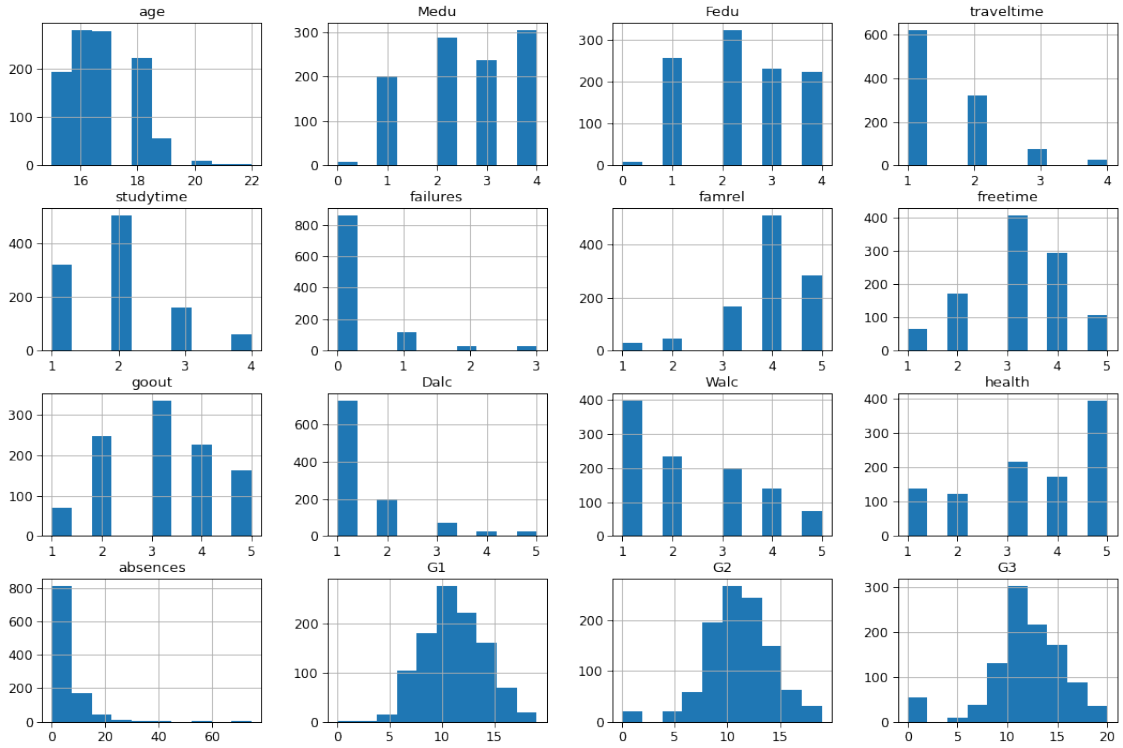
[25]:
```
box_plot = sns.boxplot(data['G3'])
box_plot.set_title('Wykres pudełkowy zmiennej celu G3')
plt.show()
```



Znowu potwierdzenie lewostronnej skośności - boxplot wskazuje outliery po lewej stronie (pik w zerze z wykresu rozkładu?). Nie mamy jednak pewności czy powinny być one usunięte - być może to ważne dane pomiarowe.

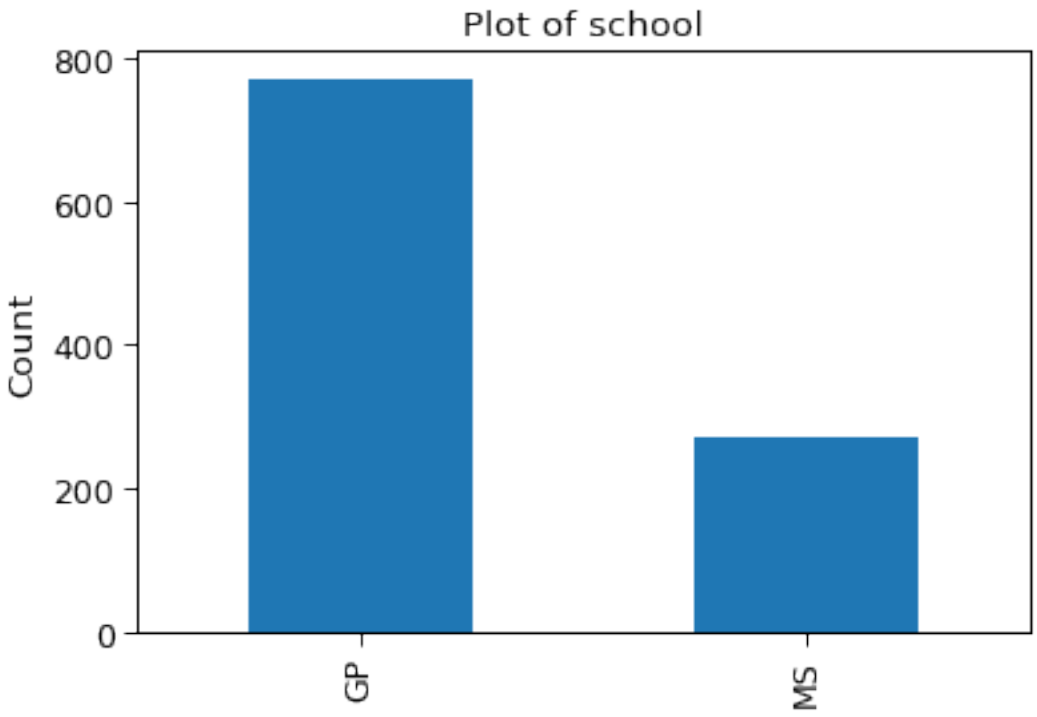
1.1.5 Rozkłady zmiennych objaśniających

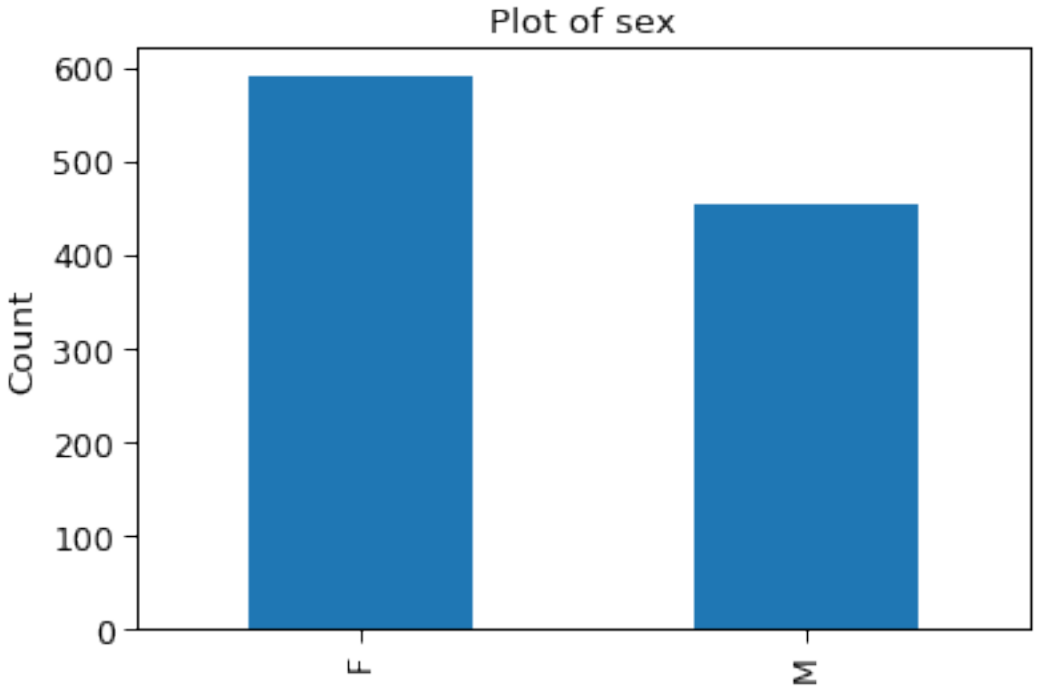
[27]:
```
data.hist(figsize=(18,12))
plt.show()
```

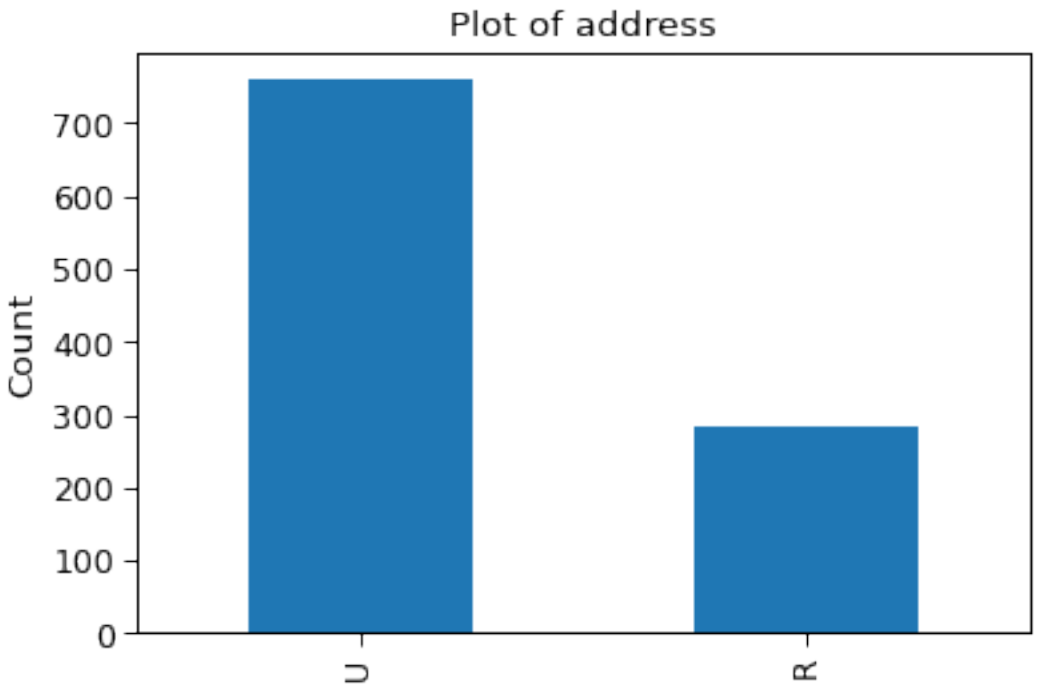


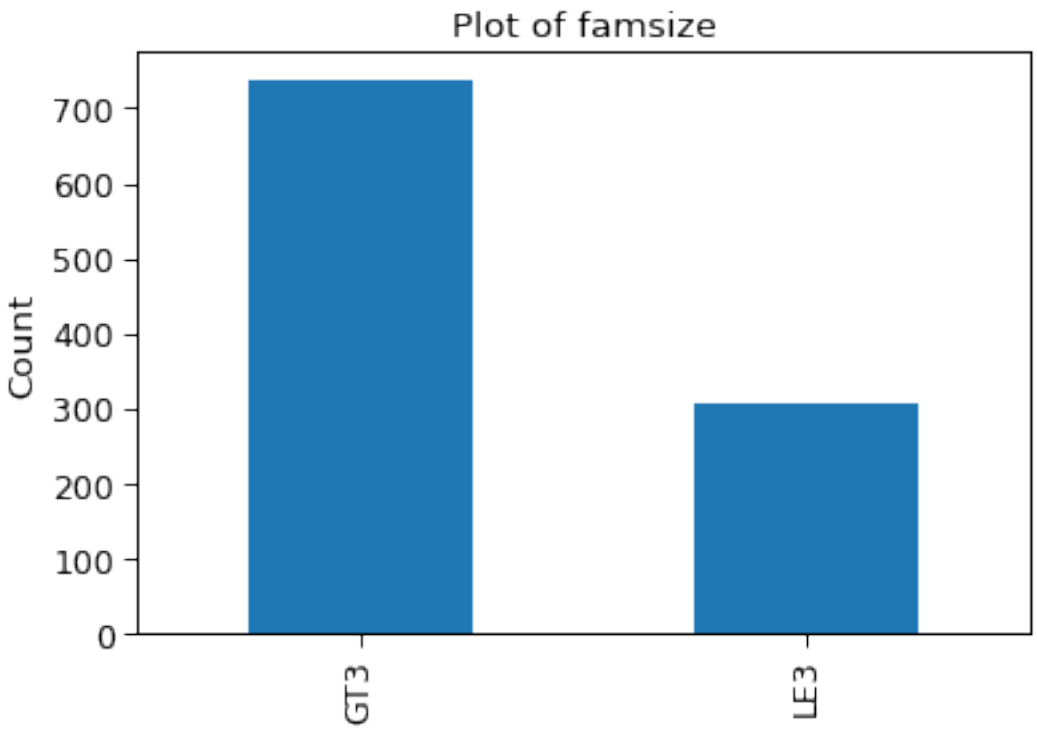
Mimo lewostronnej skośności, zmienne `G1`, `G2` i `G3` są nienaturalnie pochylone w prawo.

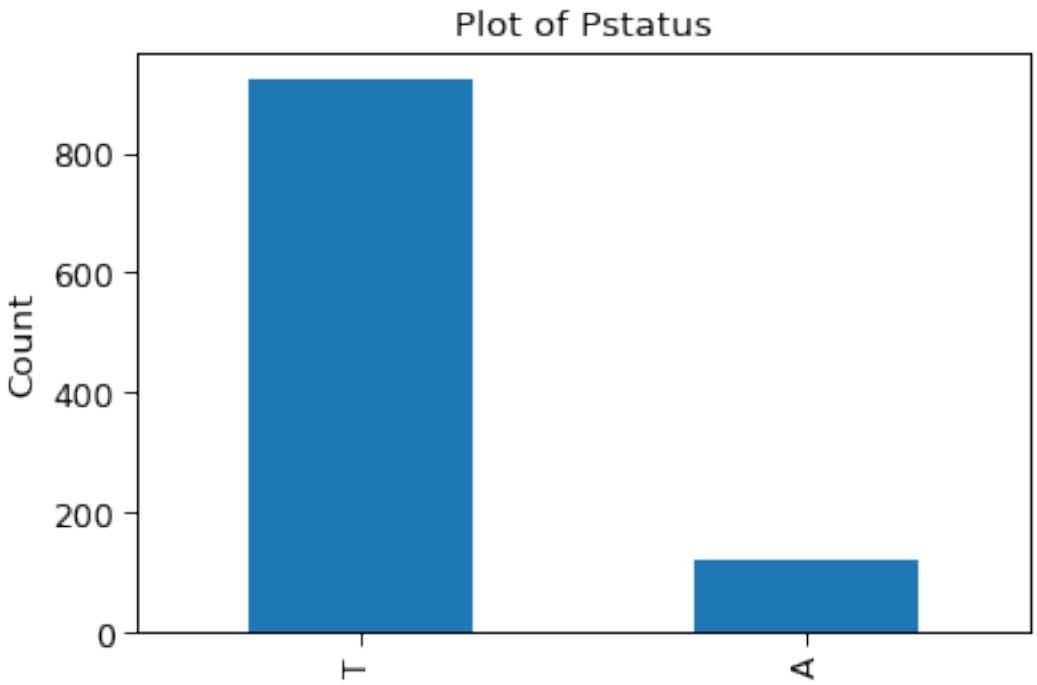
[35]:
```
categorical_data = data.iloc[:,np.r_[0:2, 3:6, 8:12, 15:23, 33]]
for column in categorical_data:
  categorical_data[column].value_counts().plot(kind = "bar", figsize = (6,4))
  plt.title("Plot of %s"%column)
  plt.ylabel("Count")
  plt.show()
matplotlib.rcParams.update({'font.size': 15})
```

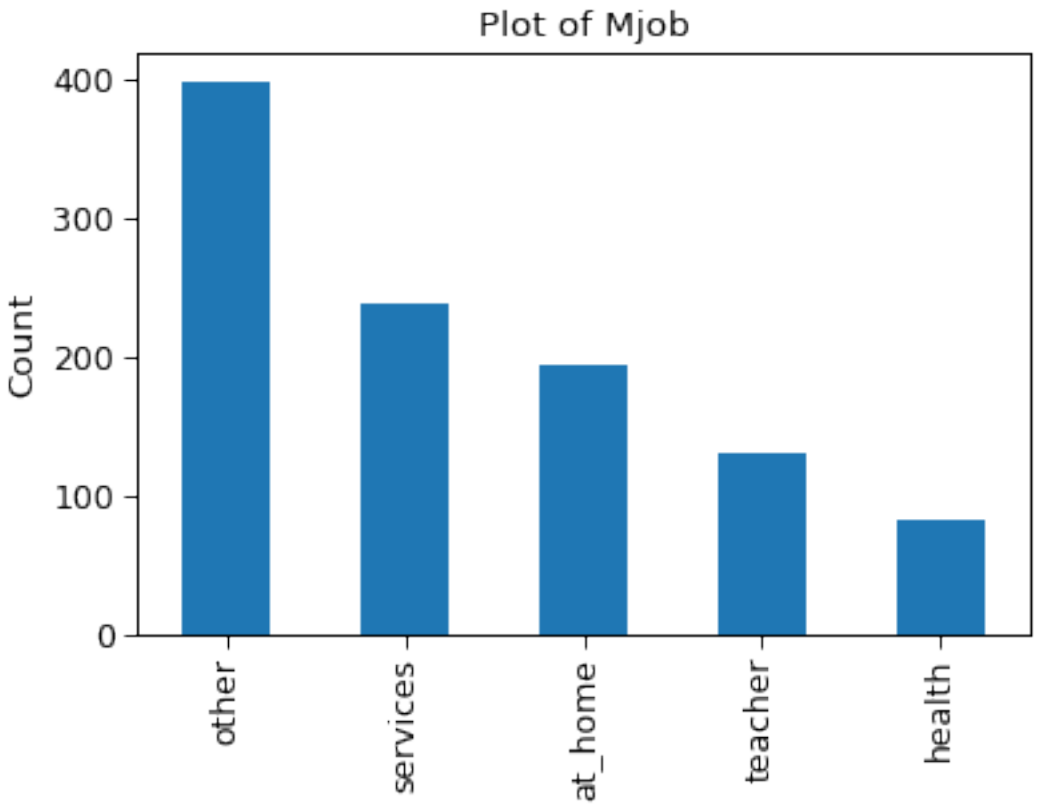

Plot of school
Count
800
600
400
200
0
GP
MS

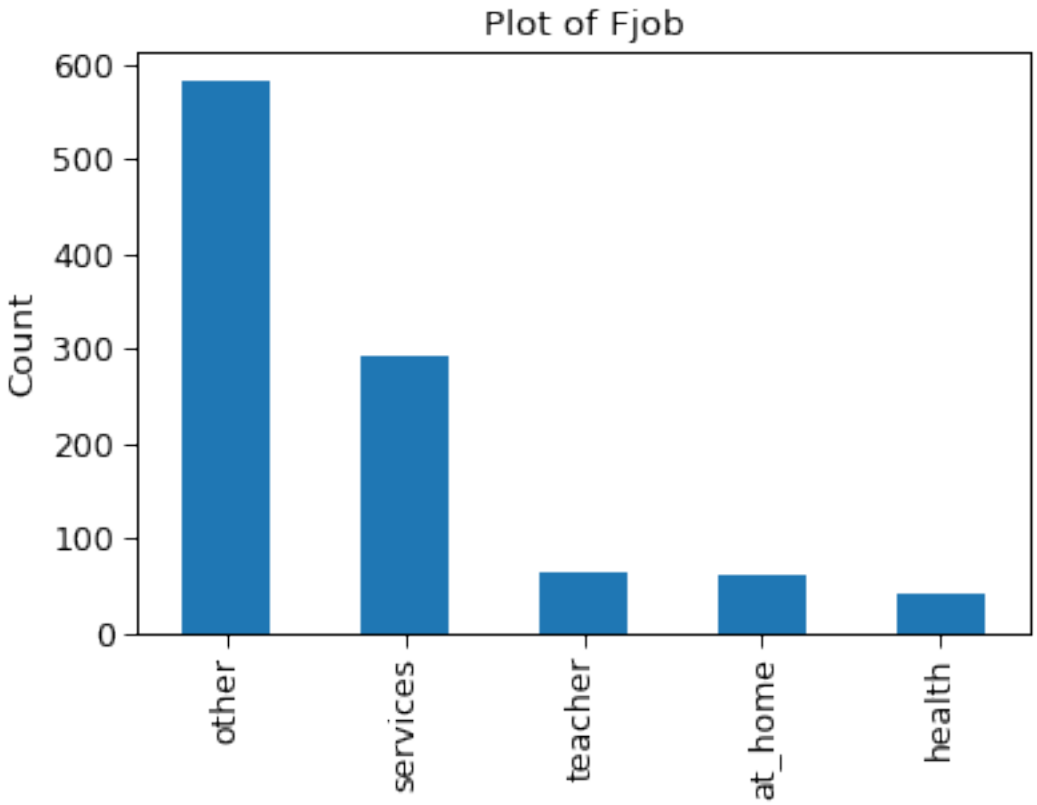

Plot of sex
Count
600
500
400
300
200
100
0
F
M

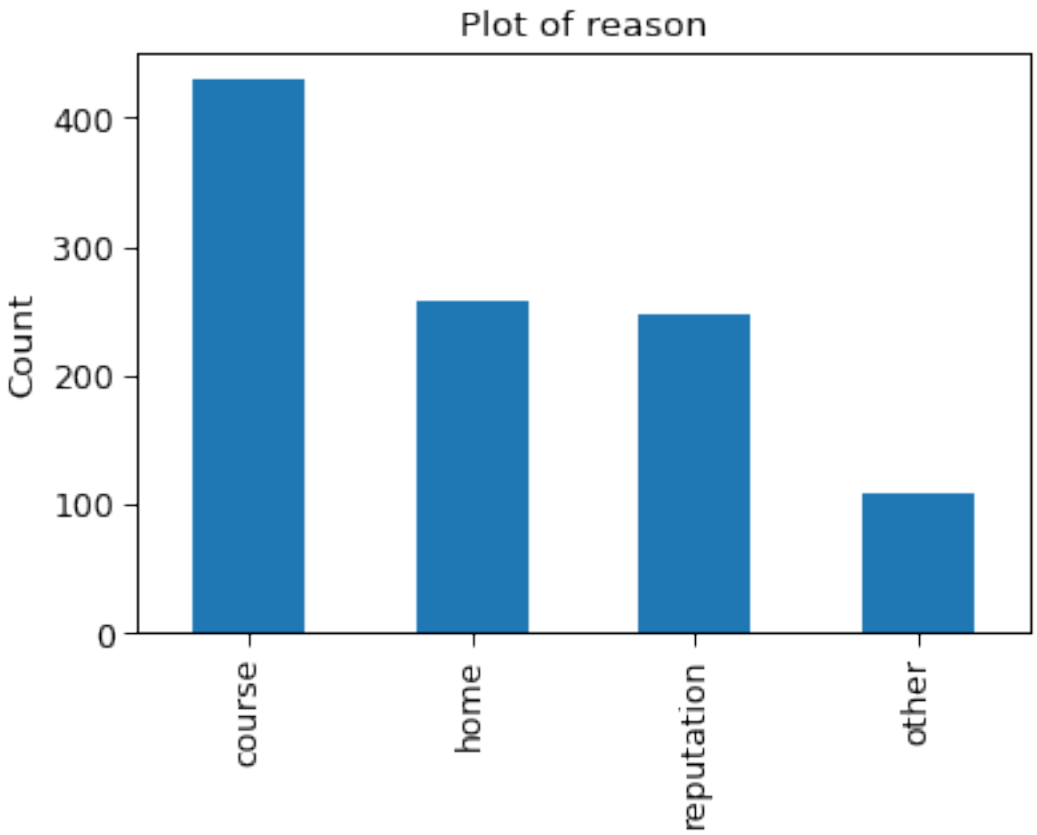

Plot of address
Count
0
100
200
300
400
500
600
700
U
R

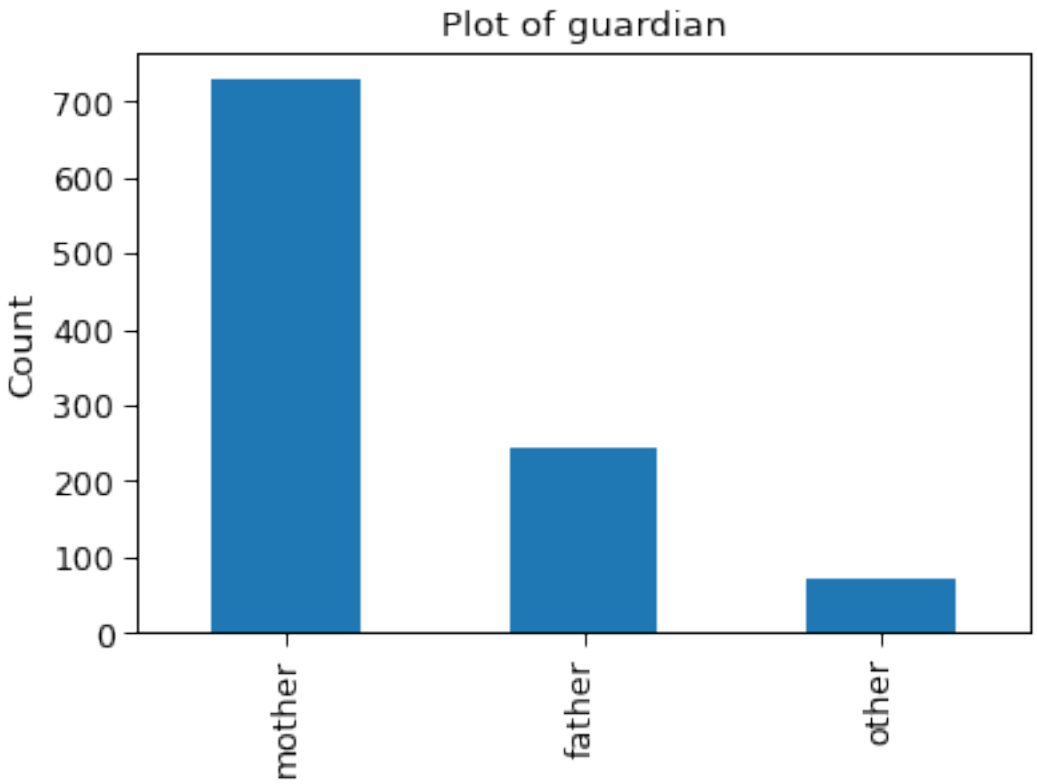

Plot of famsize
Count
0
100
200
300
400
500
600
700
GT3
LE3

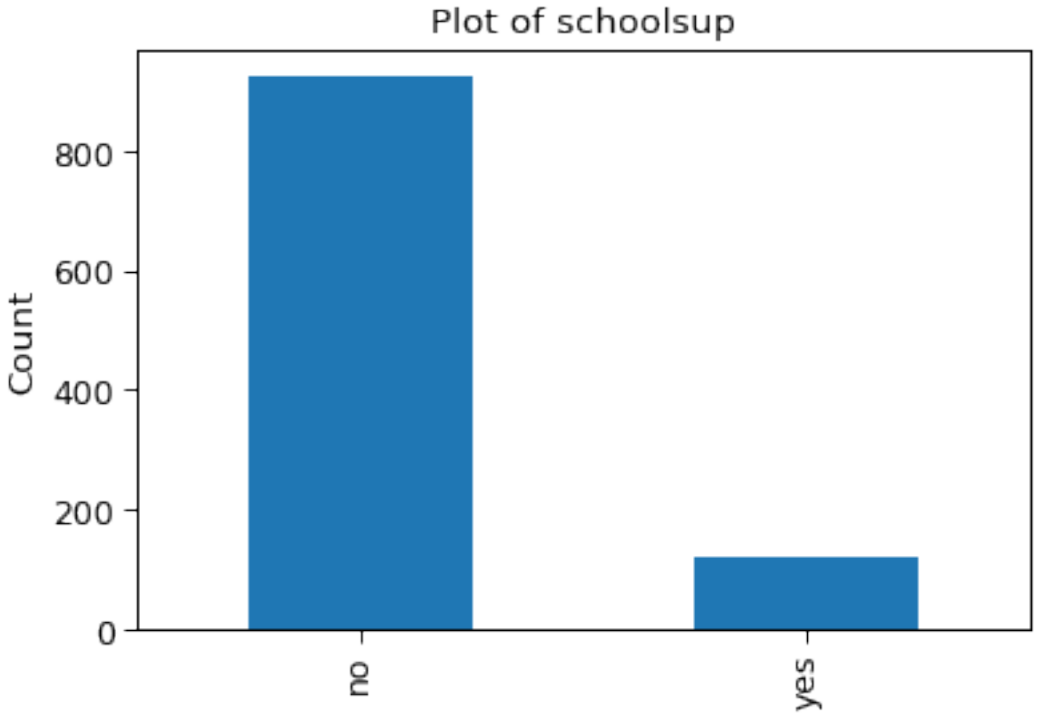

Plot of Pstatus
Count
800
600
400
200
0
T
A

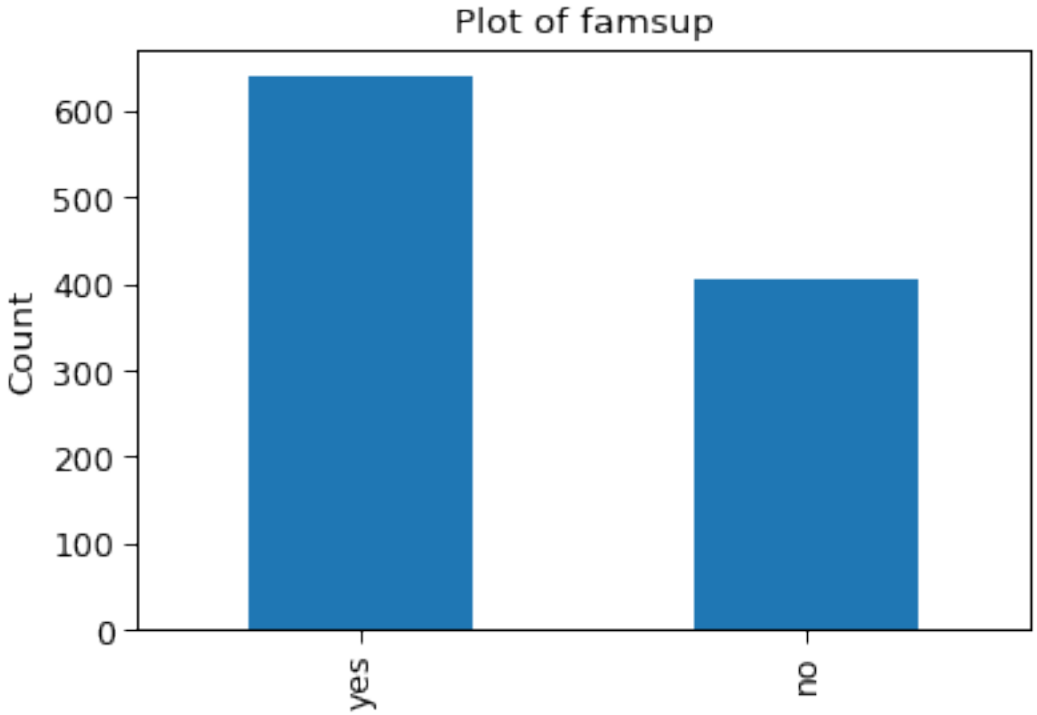

Plot of Mjob
Count
400
300
200
100
0
other
services
at_home
teacher
health

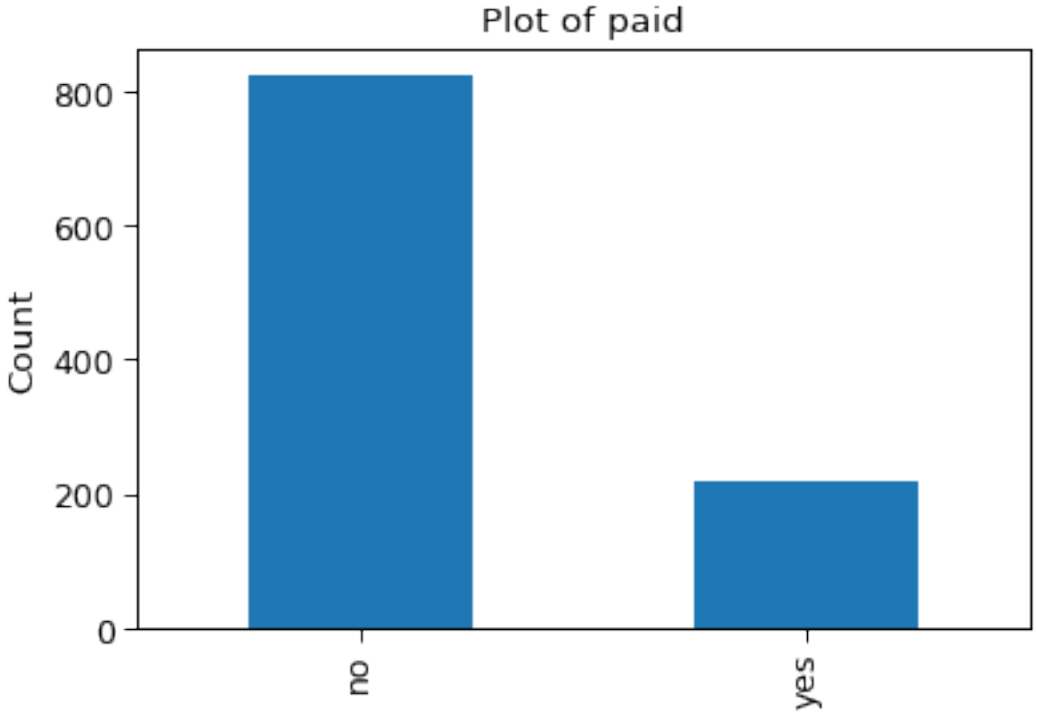

Plot of Fjob
Count
600
500
400
300
200
100
0
other
services
teacher
at_home
health

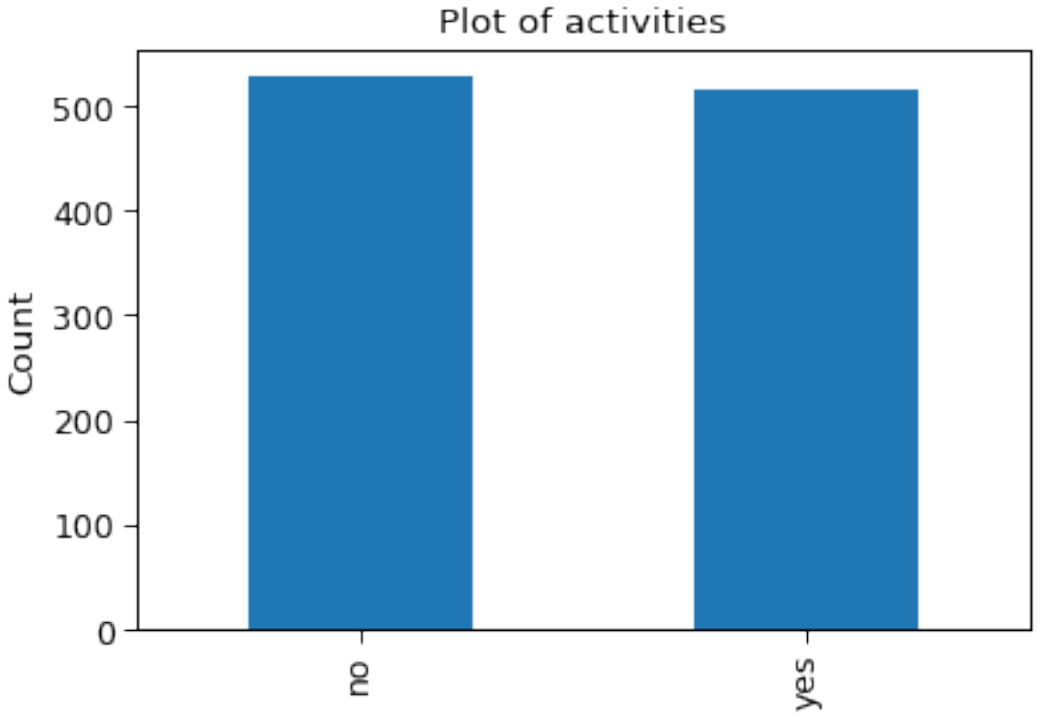

Plot of reason
Count
400
300
200
100
0
course
home
reputation
other

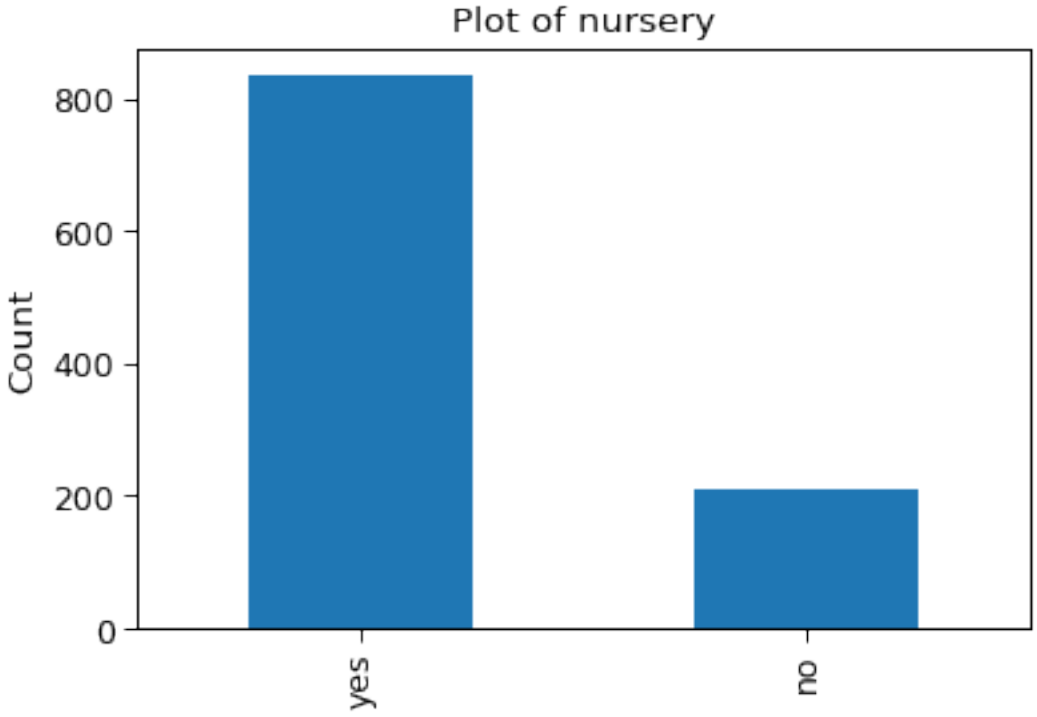

Plot of guardian
Count
0
100
200
300
400
500
600
700
mother
father
other

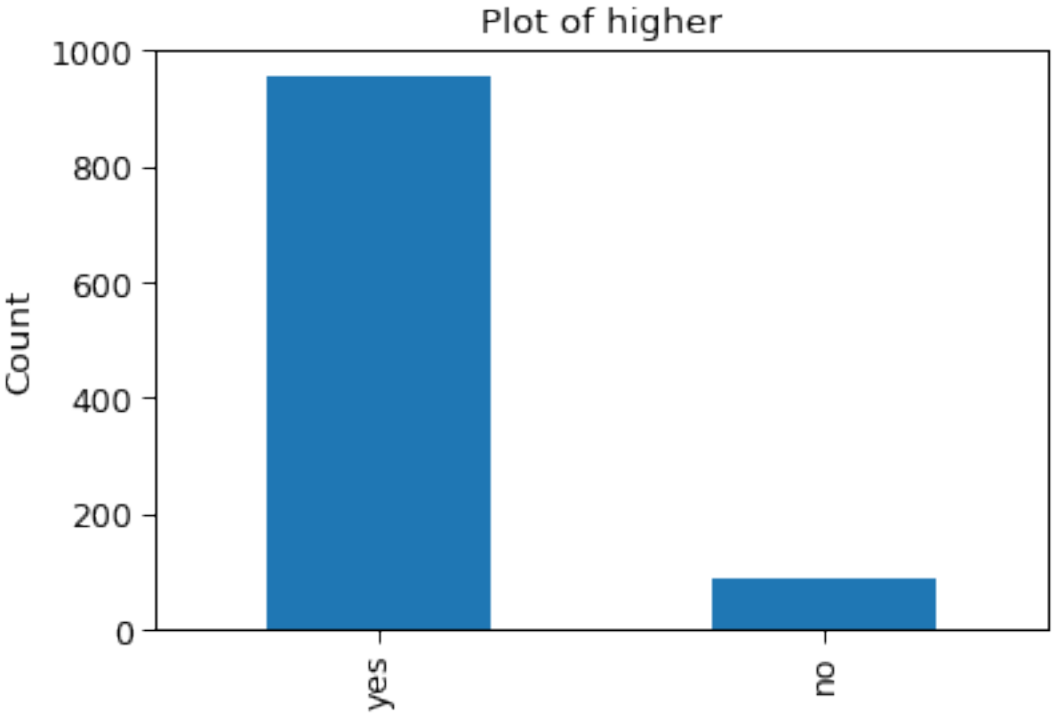

Plot of schoolsup
Count
0
200
400
600
800
no
yes

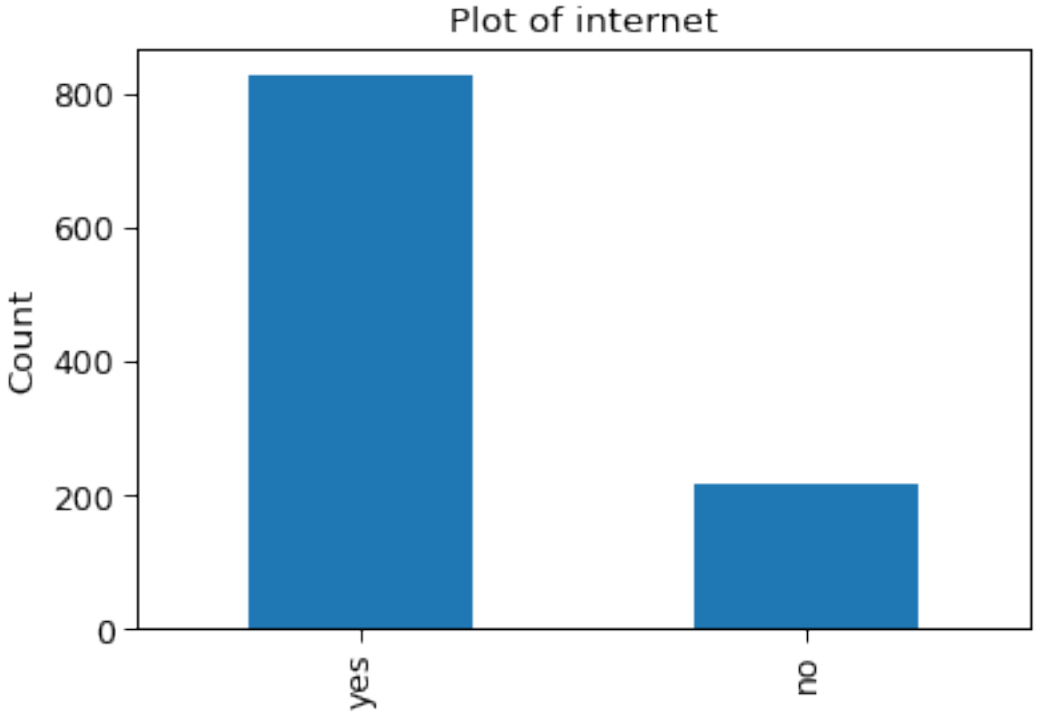

Plot of famsup
Count
600
500
400
300
200
100
0
yes
no

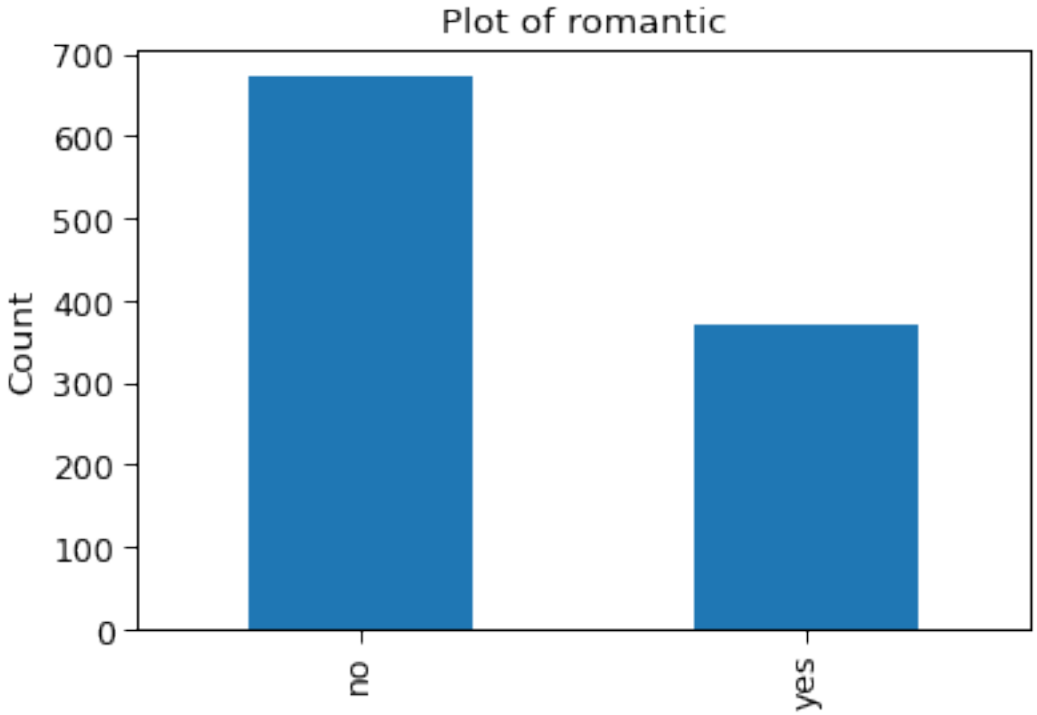

Plot of paid
Count
800
600
400
200
0
no
yes

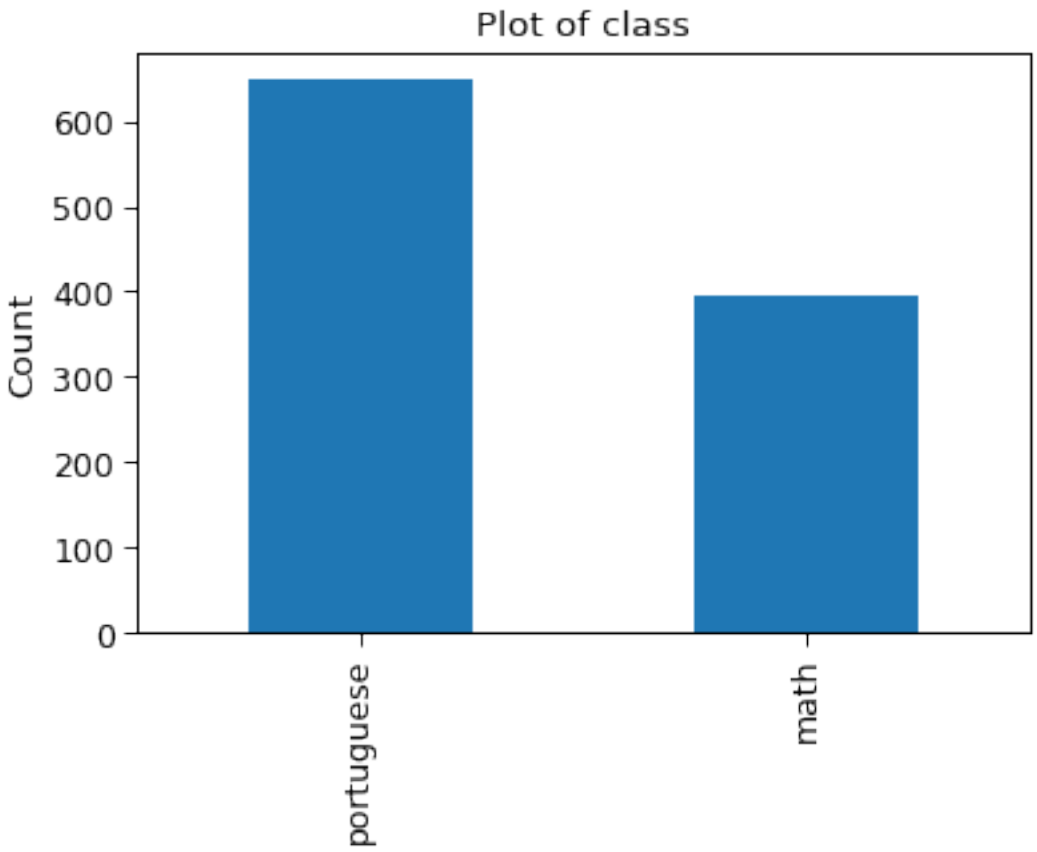

Plot of activities
Count
500
400
300
200
100
0
no
yes


Plot of nursery
Count
800
600
400
200
0
yes
no


Plot of higher
1000
800
600
400
200
0
Count
yes
no


Plot of internet
800
600
400
200
0
Count
yes
no


Plot of romantic
Count
700
600
500
400
300
200
100
0
no
yes

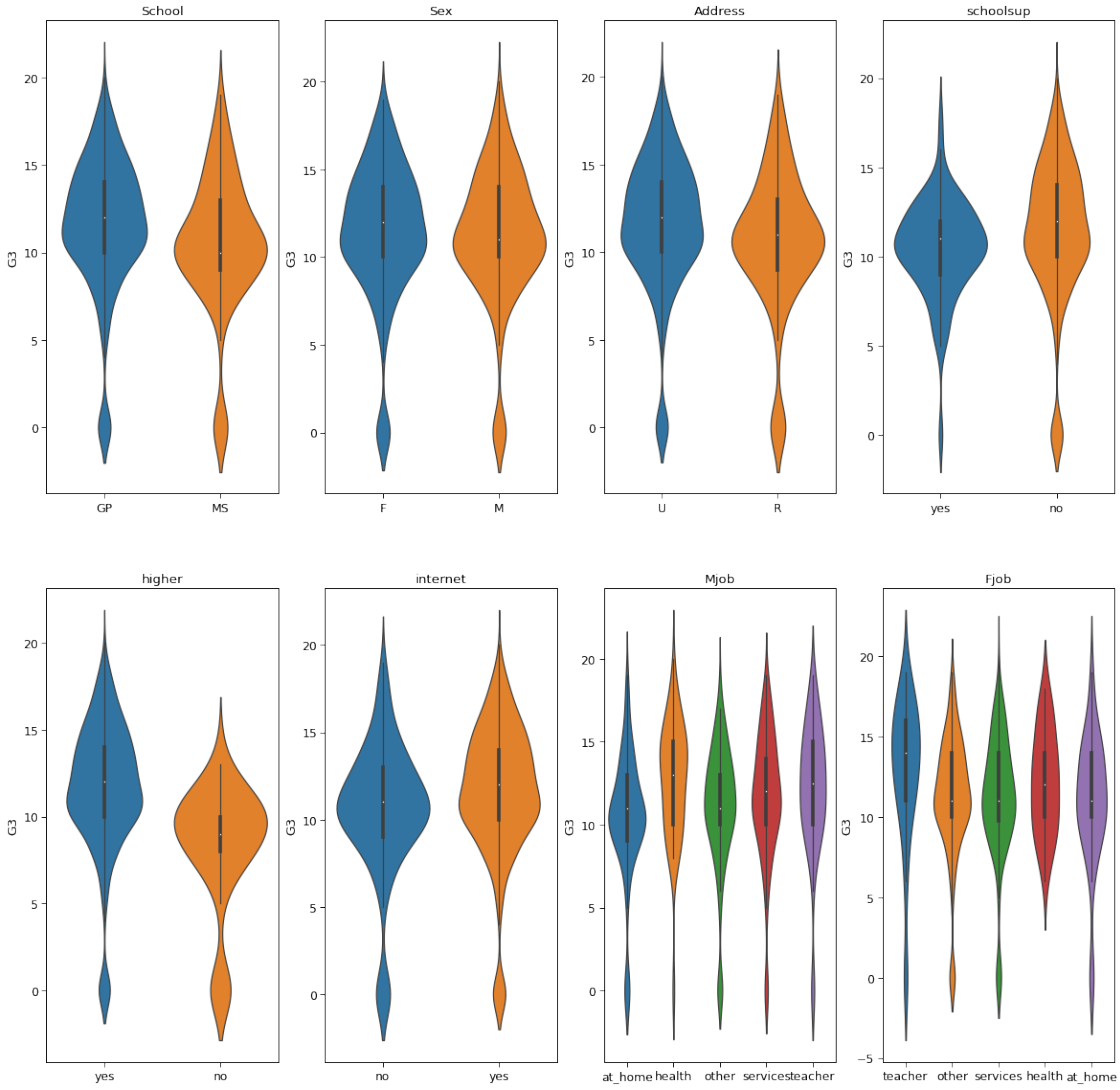


Największe dysproporcje występują w rozkładzie znmiennych `Pstatus` (rodzice żyjący razem/osobno), `schoolsup` (dodatkowe wsparcie edukacyjne) oraz `higher` (chęć pobierania dalszej edukacji). Na podstawie rozkładów można wywnioskować, że większość ankietowanych chodziło do szkoły Gabriel Pereira, mieszkało w mieście, pochodziło z większej rodziny, było wychowywanych przez matkę.

[40]:
```
fig, axs = plt.subplots(2, 4,figsize=(20, 20))
sns.violinplot(ax=axs[0, 0],x=data['school'],y= data['G3'])
axs[0, 0].set_title('School')
axs[0, 0].set_xlabel('')
sns.violinplot(ax=axs[0, 1],x=data['sex'], y=data['G3'])
axs[0, 1].set_title('Sex')
axs[0, 1].set_xlabel('')
sns.violinplot(ax=axs[0, 2],x=data['address'], y=data['G3'])
axs[0, 2].set_title('Address')
axs[0, 2].set_xlabel('')
sns.violinplot(ax=axs[0, 3],x=data['schoolsup'], y=data['G3'])
axs[0, 3].set_title('schoolsup')
axs[0, 3].set_xlabel('')
sns.violinplot(ax=axs[1, 0],x=data['higher'], y=data['G3'])
```

```
axs[1, 0].set_title('higher')
axs[1, 0].set_xlabel('')
sns.violinplot(ax=axs[1, 1],x=data['internet'], y=data['G3'])
axs[1, 1].set_title('internet')
axs[1, 1].set_xlabel('')
sns.violinplot(ax=axs[1, 2],x=data['Mjob'], y=data['G3'])
axs[1, 2].set_title('Mjob')
axs[1, 2].set_xlabel('')
sns.violinplot(ax=axs[1, 3],x=data['Fjob'], y=data['G3'])
axs[1, 3].set_title('Fjob')
axs[1, 3].set_xlabel('')
```

[40]: Text(0.5, 0, '')



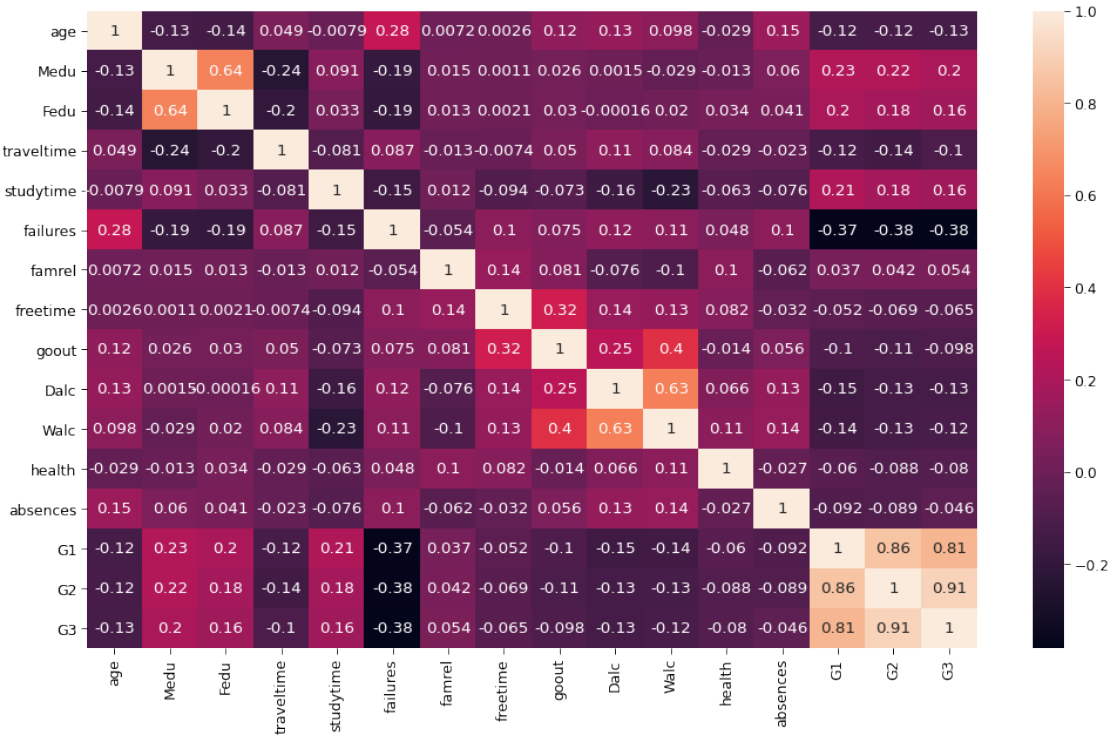
Powyżej przedstawiłam najciekawsze wykresy wiolinowe. Można z nich odczytać następujące fakty: - wyższe oceny końcowe uzyskiwali uczniowie ze szkoły Gabriel Pereira, zamieszkali w mieście, nie korzystający z dodatkowych pomocy (wynika z tego, że zajęcia dodatkowe były najczęściej brane w formie doszkalających korepetycji, nie takich, które jeszcze bardziej rozwijają), chcący kontynuować naukę oraz posiadający internet - średnio wyższe oceny dostawały dziewczyny, ale to wśród chłopców było więcej 'wybitnych jednostek' - dzieci rodziców pracujących w służbie zdrowia (zwłaszcza ojców) otrzymywały najwyższe oceny

1.1.6 Obserwacje odstające

Większość zmiennych jest kategorycznych - trudno stwierdzić, że jakiekolwiek zmienne są odstające

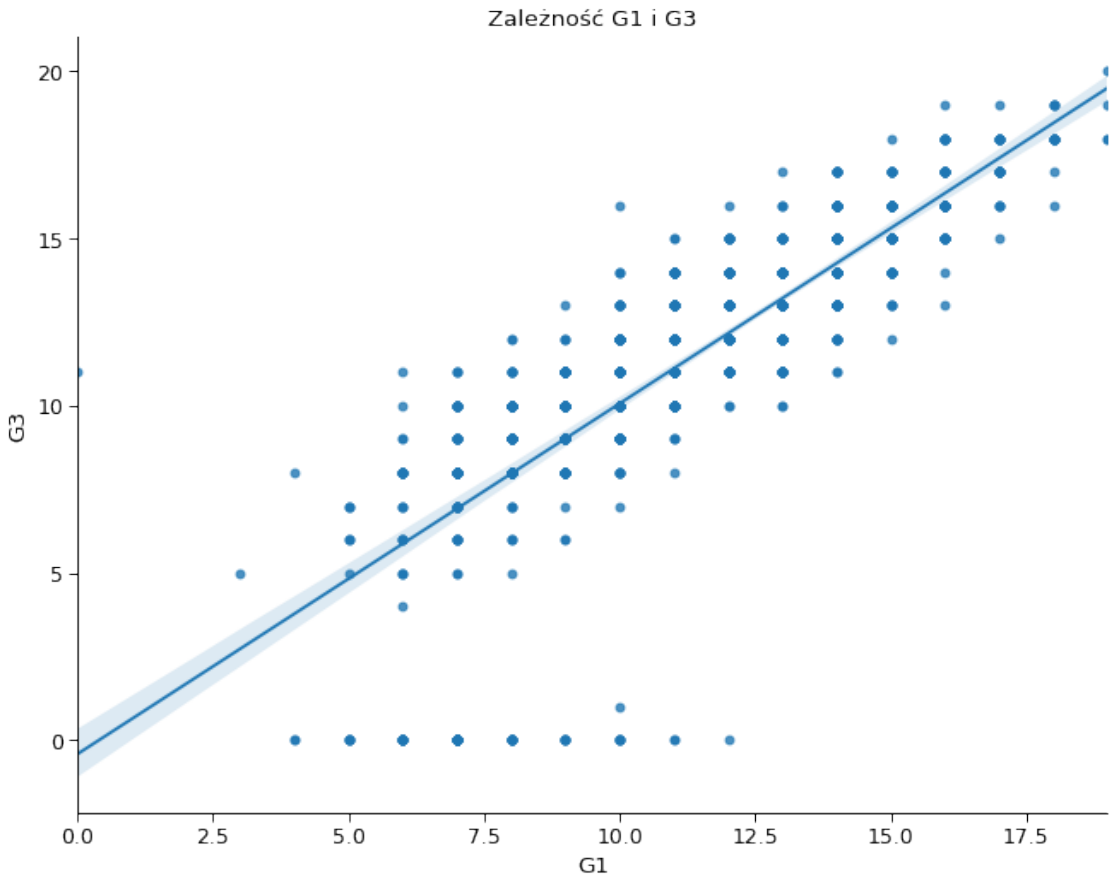
1.1.7 Korelacja zmiennych

[52]:
```
fig_dims = (17, 10)
fig, ax = plt.subplots(figsize=fig_dims)
sns.heatmap(data.corr(), ax=ax, annot=True)
plt.show()
```

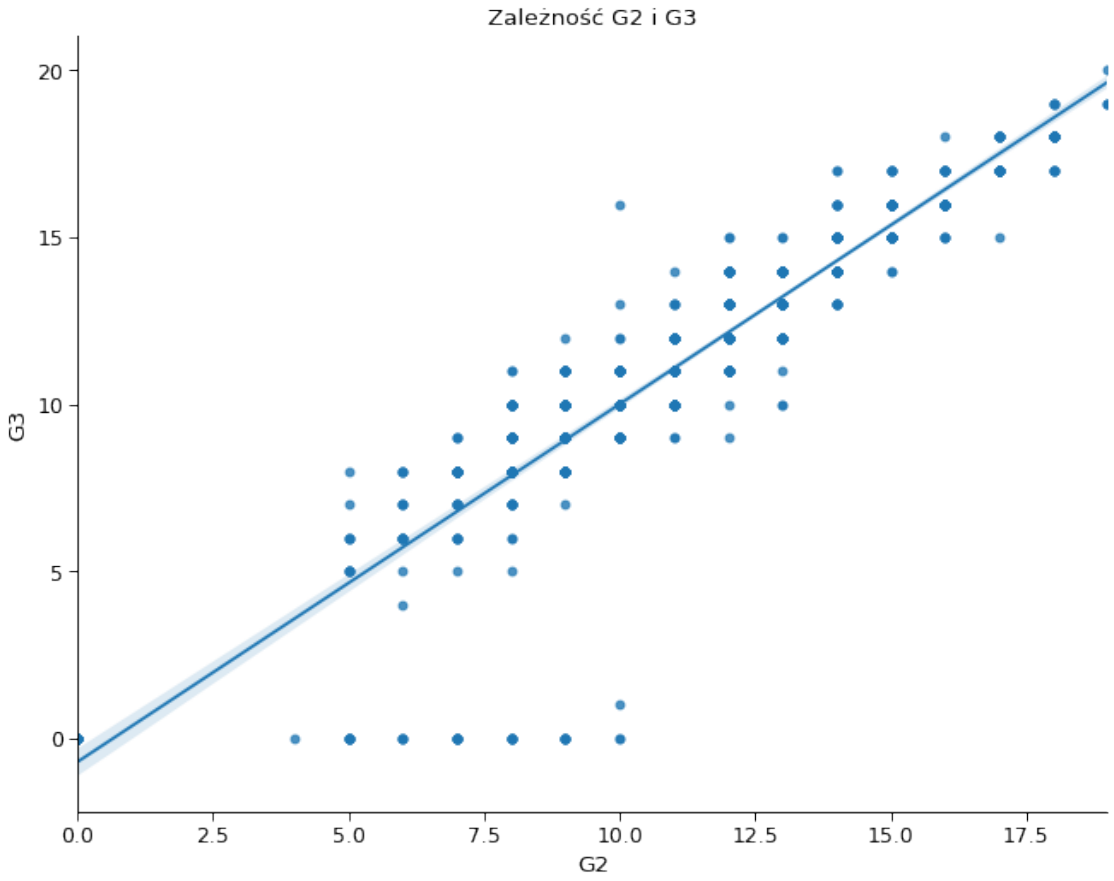


Widać sporą korelację pomiędzy wykształceniem matki i ojca (ludzię zazwyczaj wiążą się z osobami o podobnym stopniu wykształcenia?) oraz korelację pomiędzy zmiennymi `G1`, `G2` i `G3`. Nie dziwi nas to - ocena z pierwszego semestru wpływa znacząco na tę z drugiego oraz tę końcową. Jeśli chodzi o inne zmienne, `G3` jest najsiliej skorelowana ze zmienną `failure` - jest to korelacja negatywna, co też nie dziwi - im więcej poprawek, tym mniejsze prawdopodobieństwo słabej oceny.

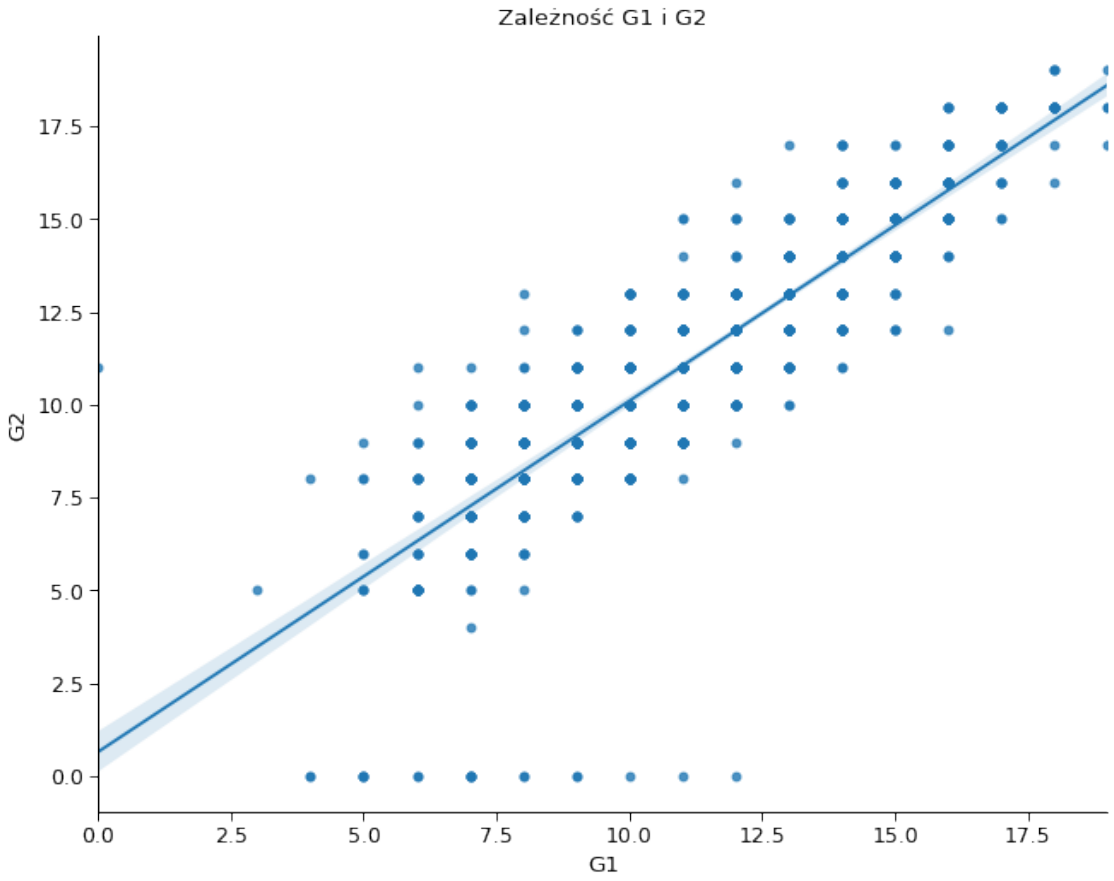
[54]:
```
sns.lmplot( y='G3', x='G1', data=data, height=7, aspect=1.3)
plt.title("Zależność G1 i G3")
plt.show()
```



[55]:
```
sns.lmplot( y='G3', x='G2', data=data, height=7, aspect=1.3)
plt.title("Zależność G2 i G3")
plt.show()
```



[56]:
```
sns.lmplot( y='G2', x='G1', data=data, height=7, aspect=1.3)
plt.title("Zależność G1 i G2")
plt.show()
```



Wnioski jak wyżej - zależność istnieje.

1.1.8 Wnioski

Naszą zmienną celu jest `G3` - ocena końcowa. Jest ona silnie skorelowana z `G1` i `G2` - ocenami kolejno na pierwszy i drugi semestr. Dane przedstawiają logiczne zależności - średnio wyższe oceny otrzymują uczniowie z dostępem do internetu oraz ci planujący dalszą edukacje. Widać również w nich pewne zależności społeczne - dzieci pracowników służby zdrowia otrzymują średnio wyższe oceny niż te, których rodzice pracują w innych sektorach. Trudno ocenić, czy w zbiorze występują obserwacje odstające - większość zmiennych jest kategoryczna.