Análise Exploratória dos Dados

Iremos utilizar o arquivo que contém as estimativas de taxa de contaminação já preenchidas.

Descrição do problema

Você é um funcionário da OMS que deve avaliar os níveis de contaminação de um vírus em um determinado país. As pessoas dentro de uma sociedade podem estar conectadas de alguma maneira (familia, amizade ou trabalho) e cada pessoa possui um conjunto de atributos. Este vírus afeta esta sociedade como descrito a seguir:

- a taxa de contaminação varia de pessoa para pessoa;
- a taxa de contaminação de uma pessoa A para B é diferente de B para A e depende das características de ambas as pessoas (A e B);
- a contaminação só passa através de indivíduos conectados;
- não existe cura para essa doença;

O desafio

Foram coletados os dados de contaminação (ou seja, as taxas de contaminação) para metade desta sociedade. Neste problema, você deverá estimar a taxa para o restante dessa sociedade e decidir políticas de saúde com base nos resultados obtidos. Observação: Para determinar as taxas de contaminação, devem ser levados em consideração tanto as características dos infectados quanto dos infectantes.

Importando bibliotecas

In [1]:
```
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import plotly.express as px
import scipy.stats as stats
import seaborn as sns
```

Carregando os dados

In [2]:
```
df_full = pd.read_csv('../data/df_full_denorm.csv')
```

Visualizando os dados

In [3]:
```
df_full.head()
```

Out[3]:

	idade_V1	qt_filhos_V1	estuda_V1	trabalha_V1	pratica_esportes_V1	IMC_V1	idade_V2	qt_filhos_V2
0	44.0	1.0	1.0	0.0	1.0	22.200956	24.0	0.0
1	44.0	1.0	1.0	0.0	1.0	22.200956	35.0	1.0
2	24.0	0.0	0.0	0.0	1.0	25.378720	50.0	1.0
3	24.0	0.0	0.0	0.0	1.0	25.378720	30.0	2.0
4	35.0	1.0	0.0	0.0	1.0	19.952393	20.0	1.0

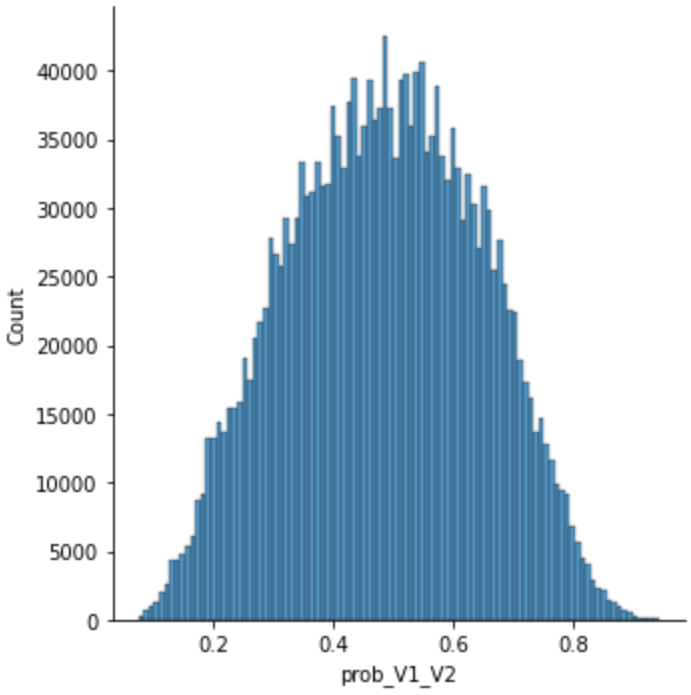
Estatísticas das probabilidades de contaminação

In [4]:
```
# taxas de contaminação

print("Medidas de tendência central das taxas de contaminação:")
print("Média das probabilidades = %.2f" % df_full['prob_V1_V2'].mean())
print("Desvio Padrão das probabilidades = %.2f" % df_full['prob_V1_V2'].std())
print("Mediana das probabilidades = %.2f" % df_full['prob_V1_V2'].median())
```

```
Medidas de tendência central das taxas de contaminação:
Média das probabilidades = 0.48
Desvio Padrão das probabilidades = 0.16
Mediana das probabilidades = 0.49
```

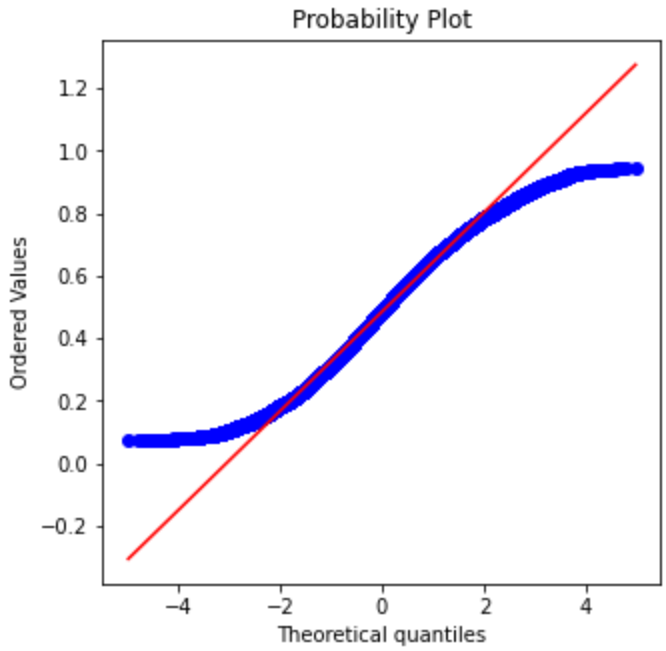
In [5]:
```
# histograma das taxas de contaminação

# full_df['prob_V1_V2'].plot(kind='hist', bins=100)
sns.displot(df_full['prob_V1_V2'], bins=100)
```

Out[5]:
```
<seaborn.axisgrid.FacetGrid at 0x7f793c071340>
```



In [6]:
```
# cálculo de quantis/gráfico de probabilidade
# distribuição normal usada como default

fig, ax = plt.subplots(figsize=(5, 5))
stats.probplot(df_full['prob_V1_V2'], plot=ax)
```

Out[6]:
```
((array([-4.96327375, -4.78824027, -4.6937436 , ...,  4.6937436 ,
          4.78824027,  4.96327375]),
  array([0.07446164, 0.0749127 , 0.07510291, ..., 0.94201946, 0.94211638,
         0.94224519])),
 (0.15913161654036384, 0.48491429162614047, 0.9959495274526712))
```



É possivel constatar, pelos gráficos acima, que a distribuição das probabilidades das taxas de contaminação se aproxima de uma distribuição normal.

Seguiremos com a análise por alguns dos fatores presentes nos dados. Sendo eles:

- Idade
- IMC
- Transportes mais utilizado

Análise por idade

Taxa média de transmissão por determinada faixa de idade, considerando `idade_V1` :

In [7]:
```
age_df_V1 = df_full.groupby(by=['idade_V1']).agg({'prob_V1_V2': 'mean',
                                                  'IMC_V1': 'count',
                                                  'pratica_esportes_V1': 'mean'})
age_df_V1.reset_index(inplace=True)
age_df_V1.sort_values('prob_V1_V2', ascending=False, inplace=True)
print(age_df_V1.shape)
age_df_V1.head()
```

```
(102, 4)
```

Out[7]:

	idade_V1	prob_V1_V2	IMC_V1	pratica_esportes_V1
100	110.0	0.610104	4	1.000
92	92.0	0.604383	8	1.000
99	106.0	0.590779	4	0.000
0	0.0	0.569827	4	1.000
1	1.0	0.557648	32	0.625

In [8]:
```
age_df_V1.describe()
```

Out[8]:

	idade_V1	prob_V1_V2	IMC_V1	pratica_esportes_V1

	idade_V1	prob_V1_V2	IMC_V1	pratica_esportes_V1
count	102.000000	102.000000	102.000000	102.000000
mean	50.872549	0.485815	19607.823529	0.651594
std	30.246077	0.053319	25883.221435	0.155019
min	0.000000	0.171588	4.000000	0.000000
25%	25.250000	0.476377	135.000000	0.647707
50%	50.500000	0.480497	4206.000000	0.659178
75%	75.750000	0.507408	35442.000000	0.666829
max	111.000000	0.610104	77504.000000	1.000000

Iremos analisar as faixas de idade que ocorrem pelo menos 4000 vezes. As idades que não serão analisadas, dada a condição anterior, são:

In [9]:
```
print(age_df_V1[age_df_V1['IMC_V1'] < 4000].sort_values('idade_V1')['idade_V1'].unique())
```

```
[  0.   1.   2.   3.   4.   5.   6.   7.  59.  60.  61.  62.  63.  64.
  65.  66.  67.  68.  69.  70.  71.  72.  73.  74.  75.  76.  77.  78.
  79.  80.  81.  82.  83.  84.  85.  86.  87.  88.  89.  90.  91.  92.
  93.  95.  96.  97. 100. 103. 106. 110. 111.]
```

In [10]:
```
age_df_V1 = age_df_V1[age_df_V1['IMC_V1'] >= 4000]
print(age_df_V1.shape)
age_df_V1.head()
```

```
(51, 4)
```

Out[10]:

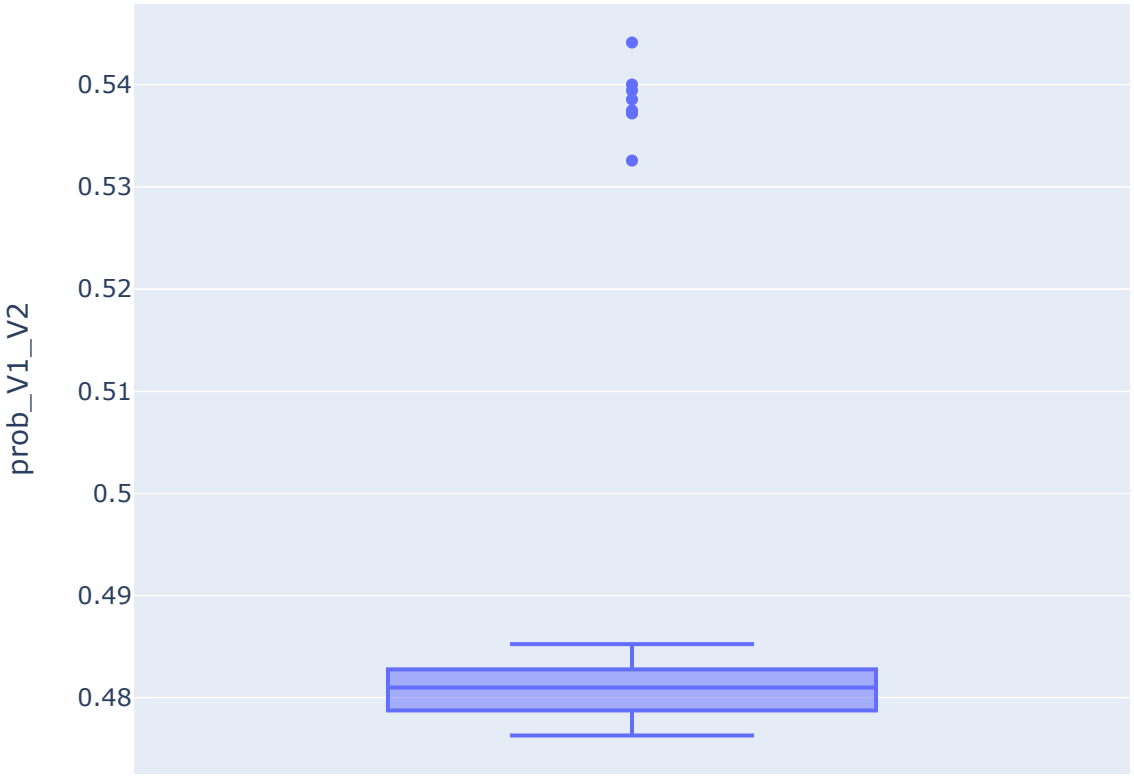
	idade_V1	prob_V1_V2	IMC_V1	pratica_esportes_V1
8	8.0	0.544148	4696	0.666951
10	10.0	0.540032	10396	0.664102
14	14.0	0.539454	28116	0.659696
11	11.0	0.538578	13752	0.653578
12	12.0	0.537494	17962	0.659949

In [11]:
```
age_df_V1.describe()
```

Out[11]:

	idade_V1	prob_V1_V2	IMC_V1	pratica_esportes_V1
count	51.000000	51.000000	51.000000	51.000000
mean	33.000000	0.489352	38560.431373	0.660007
std	14.866069	0.021462	24889.796841	0.005648
min	8.000000	0.476268	4612.000000	0.645353
25%	20.500000	0.478722	14634.000000	0.656457
50%	33.000000	0.480968	36004.000000	0.659485
75%	45.500000	0.482719	61992.000000	0.663611
max	58.000000	0.544148	77504.000000	0.673551

In [12]:
```
fig = px.box(age_df_V1, y='prob_V1_V2', hover_data=['idade_V1'])
fig.show(figsize=(5, 5))
```



A idade de 8 anos se mostra sendo outlier.

Identificando as 10 idades que possuem potencial de contaminação maiores:

In [13]:
```
age_df_V1[:10]
```

Out[13]:

	idade_V1	prob_V1_V2	IMC_V1	pratica_esportes_V1
8	8.0	0.544148	4696	0.666951
10	10.0	0.540032	10396	0.664102
14	14.0	0.539454	28116	0.659696
11	11.0	0.538578	13752	0.653578
12	12.0	0.537494	17962	0.659949
13	13.0	0.537255	23072	0.656553
15	15.0	0.537200	33756	0.661216
9	9.0	0.532580	7076	0.659129
17	17.0	0.485217	44272	0.665793
20	20.0	0.483226	60732	0.669038

As idade que mais tem chance de contaminação estão nas faixas de 8 e 15 anos e 17 e 20 anos. Também é

possível observar que as taxas de contaminação estão muito próximas umas das outras.

Taxa média de transmissão por determinada faixa de idade, considerando `idade_V2` :

In [14]:
```
age_df_V2 = df_full.groupby(by=['idade_V2']).agg({'prob_V1_V2': 'mean',
                                                  'IMC_V2': 'count',
                                                  'pratica_esportes_V2': 'mean'})
age_df_V2.reset_index(inplace=True)
age_df_V2.sort_values('prob_V1_V2', ascending=False, inplace=True)
age_df_V2.head()
```

Out[14]:

	idade_V2	prob_V1_V2	IMC_V2	pratica_esportes_V2
101	106.0	0.626347	2	0.00
104	124.0	0.613954	2	1.00
93	93.0	0.610039	8	0.75
98	98.0	0.594575	2	1.00
92	92.0	0.592729	10	0.60

In [15]:
```
age_df_V2.describe()
```

Out[15]:

	idade_V2	prob_V1_V2	IMC_V2	pratica_esportes_V2
count	105.000000	105.000000	105.000000	105.000000
mean	52.428571	0.487289	19047.600000	0.665699
std	31.251418	0.041729	25718.516821	0.138507
min	0.000000	0.286210	2.000000	0.000000
25%	26.000000	0.479460	112.000000	0.652902
50%	52.000000	0.484642	3352.000000	0.659173
75%	78.000000	0.488463	33172.000000	0.664480
max	124.000000	0.626347	76524.000000	1.000000

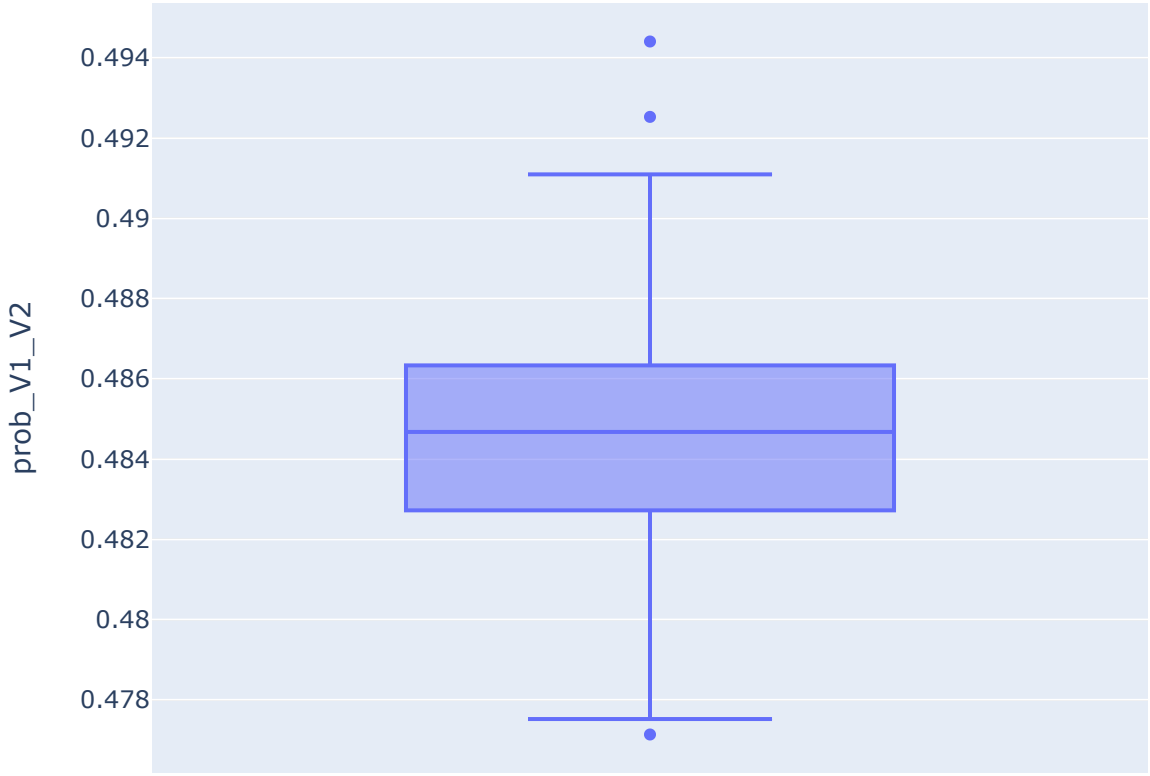
Iremos analisar apenas faixas etárias que ocorrem pelo menos 3000 vezes.

In [16]:
```
age_df_V2 = age_df_V2[age_df_V2['IMC_V2'] >= 3000]

age_df_V2.head()
```

Out[16]:

	idade_V2	prob_V1_V2	IMC_V2	pratica_esportes_V2
9	9.0	0.494408	7178	0.663695
11	11.0	0.492529	13902	0.656021
8	8.0	0.491096	4816	0.655316
13	13.0	0.490731	23258	0.657494
10	10.0	0.489475	10304	0.656832

In [17]:
```
fig = px.box(age_df_V2, y='prob_V1_V2', hover_data=['idade_V2'])
fig.show()
```



As idades de 9 e 11 anos se mostram sendo outliers.

Vamos analisar as top 10 idades com as maiores taxas de serem contaminadas.

In [18]:
```
age_df_V2[:10]
```

Out[18]:

	idade_V2	prob_V1_V2	IMC_V2	pratica_esportes_V2
9	9.0	0.494408	7178	0.663695
11	11.0	0.492529	13902	0.656021
8	8.0	0.491096	4816	0.655316
13	13.0	0.490731	23258	0.657494
10	10.0	0.489475	10304	0.656832
15	15.0	0.488713	33172	0.660316
12	12.0	0.488195	17640	0.651587
14	14.0	0.487925	27924	0.657785
19	19.0	0.487689	55514	0.657168
21	21.0	0.487270	64428	0.663314

As 10 idades que mais são contaminadas são de 8 a 15 anos e 19 a 21 anos.

Dados as 10 idades que mais podem ser contaminadas em `idade_V1` e `idade_V2` , vamos analisar as relações entre elas.

In [19]:
```
ages_V1 = [8, 9, 10, 11, 12, 13, 14, 15]
ages_v2 = [9, 10, 11, 12, 13, 15, 19, 20]

ages_df = df_full[df_full['idade_V1'].isin(ages_V1) & df_full['idade_V2'].isin(ages_v2)]
```

In [20]:
```
grau_df = ages_df.groupby(by=['grau']).agg({'prob_V1_V2': 'mean',
                                            'IMC_V2': 'count'})
grau_df.reset_index(inplace=True)
grau_df.sort_values('prob_V1_V2', ascending=False, inplace=True)
grau_df
```

Out[20]:

	grau	prob_V1_V2	IMC_V2
0	amigos	0.573397	5076
2	trabalho	0.569692	5174
1	familia	0.491666	5030

Nas faixas de idade que mais se contaminam, temos o grau/relação de amigos e ambiente de trabalho com maior frequência de contaminação e praticamente com as mesmas probabilidades. Já o ambiente familiar se mostra com menor probabilidade de contaminação.

Análise por IMC

Referência sobre classificação do IMC

In [21]:
```
def imc_status(value):
    if value <= 18.50:
        return 'abaixo do peso'
    elif value > 18.50 and value <= 24.99:
        return 'peso normal'
    elif value > 25.00 and value <= 29.99:
        return 'sobrepeso'
    else:
        return 'obesidade'

def fill_imc(df):
    df['IMC_status_V1'] = df['IMC_V1'].apply(imc_status)
    df['IMC_status_V2'] = df['IMC_V2'].apply(imc_status)

fill_imc(df_full)
df_full.head()
```

Out[21]:

	idade_V1	qt_filhos_V1	estuda_V1	trabalha_V1	pratica_esportes_V1	IMC_V1	idade_V2	qt_filhos_V2
0	44.0	1.0	1.0	0.0	1.0	22.200956	24.0	0.0
1	44.0	1.0	1.0	0.0	1.0	22.200956	35.0	1.0
2	24.0	0.0	0.0	0.0	1.0	25.378720	50.0	1.0
3	24.0	0.0	0.0	0.0	1.0	25.378720	30.0	2.0
4	35.0	1.0	0.0	0.0	1.0	19.952393	20.0	1.0

5 rows × 21 columns

Taxa média de transmissão por classificações de IMC, considerando `IMC_V1` :

In [22]:
```
imc_df_V1 = df_full.groupby(by=['IMC_status_V1']).agg({'prob_V1_V2': 'mean',
```

```
                                                        'IMC_V1': 'count'})
imc_df_V1.reset_index(inplace=True)
imc_df_V1.sort_values('prob_V1_V2', ascending=False, inplace=True)
imc_df_V1
```

Out[22]:

	IMC_status_V1	prob_V1_V2	IMC_V1
3	sobrepeso	0.485417	342336
1	obesidade	0.485087	275780
2	peso normal	0.484875	747818
0	abaixo do peso	0.484614	634064

Taxa média de transmissão por classificações de IMC, considerando `IMC_V2` :

In [23]:

```
imc_df_V2 = df_full.groupby(by=['IMC_status_V2']).agg({'prob_V1_V2': 'mean',
                                                       'IMC_V2': 'count'})
imc_df_V2.reset_index(inplace=True)
imc_df_V2.sort_values('prob_V1_V2', ascending=False, inplace=True)
imc_df_V2
```

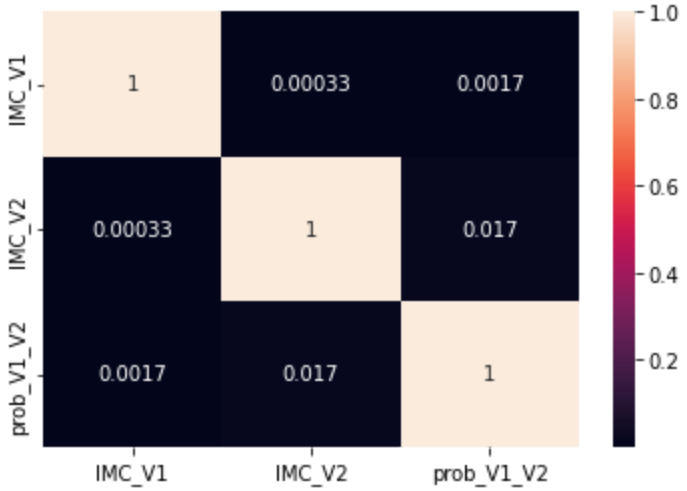
Out[23]:

	IMC_status_V2	prob_V1_V2	IMC_V2
1	obesidade	0.489831	275952
3	sobrepeso	0.486293	342666
2	peso normal	0.484680	748758
0	abaixo do peso	0.482300	632622

Nota-se que tanto considerando o `IMC_V1` quanto o `IMC_V2` , as probabilidades permaneceram muito próximas e afim de contornar isso e extrair maiores informações, vamos observar a correlação, se existente, entre o IMC a probabilidade de contaminação.

In [24]:

```
imc_df = df_full[['IMC_V1', 'IMC_V2', 'prob_V1_V2']]
corr_matrix = imc_df.corr()
sns.heatmap(corr_matrix, annot=True)
plt.show()
```



Concluimos, com o gráfico acima,que não há correlação entre o IMC a probabilidade de contaminação, pois os valores das correlações são próximas a zero.

Análise de transporte mais utilizado

Taxa média de transmissão por transportes mais utilizados pelos indivíduos, considerando `transporte_mais_utilizado_V1` :

In [25]:
```
transp_df_V1 = df_full.groupby(by=['transporte_mais_utilizado_V1']).agg({'prob_V1_V2': 'me
                                                                         'IMC_V1': 'count'
transp_df_V1.reset_index(inplace=True)
transp_df_V1.sort_values('prob_V1_V2', ascending=False, inplace=True)
transp_df_V1
```

Out[25]:

	transporte_mais_utilizado_V1	prob_V1_V2	IMC_V1
2	publico	0.495128	1148170
0	outros	0.474665	86784
1	particular	0.474458	669568
3	taxi	0.444741	95476

Nota-se que usuários de transporte público possuem maiores taxas e os usuários de taxi menores taxas.

Taxa média de transmissão por transportes mais utilizados pelos indivíduos, considerando `transporte_mais_utilizado_V2` :

In [26]:
```
transp_df_V2 = df_full.groupby(by=['transporte_mais_utilizado_V2']).agg({'prob_V1_V2': 'me
                                                                         'IMC_V2': 'count'
transp_df_V2.reset_index(inplace=True)
transp_df_V2.sort_values('prob_V1_V2', ascending=False, inplace=True)
transp_df_V2
```

Out[26]:

	transporte_mais_utilizado_V2	prob_V1_V2	IMC_V2
1	particular	0.551861	668176
0	outros	0.486670	86066
3	taxi	0.477687	96200
2	publico	0.446475	1149556

Aqui, o fato de usuários de transporte particular possuirem alta taxa de contração da doença se mostra curioso, já que naturalmente inferimos que o transporte público possui maior possibilidade de aglomeração, ocasionando contato, e assim a contaminação.

Como o uso do transporte particular se relaciona com os graus de relação?

In [27]:
```
transp_df = df_full[df_full['transporte_mais_utilizado_V2'] == 'particular']
```

In [28]:
```
grau_df = transp_df.groupby(by=['grau']).agg({'prob_V1_V2': 'mean',
                                              'IMC_V2': 'count'})
grau_df.reset_index(inplace=True)
grau_df.sort_values('prob_V1_V2', ascending=False, inplace=True)
grau_df
```

Out[28]:

	grau	prob_V1_V2	IMC_V2
2	trabalho	0.580497	222342

	grau	prob_V1_V2	IMC_V2
0	amigos	0.580361	222754
1	familia	0.494860	223080

Assim como observado na análise anterior do grau/relação, aqui também identificamos amigos e ambientem trabalho como os mais propensos a se contaminar.

Como o uso do transporte particular se relaciona com outros fatores?

In [29]:
```
prox_df = transp_df.groupby(by=['proximidade']).agg({'prob_V1_V2': 'mean',
                                                     'IMC_V2': 'count'})
prox_df.reset_index(inplace=True)
prox_df.sort_values('prob_V1_V2', ascending=False, inplace=True)
prox_df
```

Out[29]:

	proximidade	prob_V1_V2	IMC_V2
1	visita_casual	0.639747	200096
2	visita_frequente	0.552408	133726
3	visita_rara	0.512283	267216
0	mora_junto	0.446362	67138

In [30]:
```
df = df_full.groupby(by=['transporte_mais_utilizado_V2', 'proximidade', 'estado_civil_V2']

df.reset_index(inplace=True)
df.sort_values('prob_V1_V2', ascending=False, inplace=True)
df = df[df['IMC_V2'] >= 3000]

df[:10]
```

Out[30]:

	transporte_mais_utilizado_V2	proximidade	estado_civil_V2	prob_V1_V2	IMC_V2
22	particular	visita_casual	solteiro	0.640698	93980
20	particular	visita_casual	casado	0.640281	52776
21	particular	visita_casual	divorciado	0.638088	35386
23	particular	visita_casual	viuvo	0.636464	17954
6	outros	visita_casual	solteiro	0.575168	12116
4	outros	visita_casual	casado	0.574066	6902
5	outros	visita_casual	divorciado	0.573861	4656
53	taxi	visita_casual	divorciado	0.564090	5200
54	taxi	visita_casual	solteiro	0.564079	13618
52	taxi	visita_casual	casado	0.562780	7536

Usuários de transporte particular com visitas casuais tendem se contaminar com maior frequência, sejam eles solteiros, casados, divorciados ou viúvos. É possível notar que os 4 perfis com maior taxa de contaminação estão dentro das características citadas e com taxas de aquisição da doença acima de 60%.

Possíveis recomendações de políticas de saúde

Dado os fatores escolhidos para serem analisados com maior profundidade, foi observado:

- Idade: a faixa etária de pessoas que contaminam e podem ser contaminadas com maior probabilidade, se mostrou parecido, com idades variando de 8 a 15 anos até 19 e 21 anos; Também se observou, que dado as faixas etárias citadas anteriormente, os graus de amigos e trabalho mostraram as maiores frequências de contaminação e o ambiente familiar se mostrando com as menores frequências.

- IMC: já nessa análise não observamos correlações entre a classificação de IMC de uma pessoa e as probabilidades dela ser contaminada ou contaminar.

- Transporte mais utilizado: foi observado, por parte dos possíveis contaminadores, que o uso de transporte público possui as maiores taxas e o uso de taxi as menores taxas de contaminação; já em relação aqueles que podem contrair a doença, o uso de transporte particular se mostrou com as maiores taxas, o que pode ser considerado não intuitivo. ao analisar de forma um pouco mais aprofundada, foi percebido que os usuários de transporte público particular que se praticavam proximidade como visita casual, tinham as maiores chances de se contaminar.

Feitas as observações acima, possíveis recomendações surgem, sendo elas:

- Suspensão de interaçoes fora de ambientes familiares, com foco, mas não exclusivo, a faixa etária de 8 a 20 anos. Exemplos: adoção de home office para serviços não essenciais, realização de atividades como consultas e determinadas atividades físicas dentro de ambiente familiar, alternâncias de horários de entrada em shoppings, farmácias e supermercados, etc.

- Conscientização da população quanto a medidas de cuidados preventivos e o quanto a consciência desse cuidado pode ter importância em não pegar/passar a doença.

- Conscientização da população, para que mesmo em contatos curtos (casuais) não se relaxem as medidas, dado que mesmo que rápido um contato pode significar um contágio e por fim, a também conscientização que mesmo utilizando transporte particular, livre da aglomeração do transporte público, caso haja contato casual com um infectado, o risco de contaminação existe.