

Grupo:

Andreza Aparecida dos Santos (164213)

Leonardo Marçal (225240)

Ligia Vasconcellos (081938)

Mariana Amaral (262866)

Análise Exploratória: Bases Estudadas e Adotadas

Para as bases adotadas neste projeto, realizamos uma centralização dos dados em um único arquivo para que as análises exploratórias das variáveis em relação a nossa variável de resposta pudesse ser feita de forma mais direta. Tivemos um grande desafio inicial para centralizar as bases, devido ao tamanho e alto custo operacional. Foram centralizadas as seguintes bases:

Base de dados	Endereço Web	Resumo Descritivo
1. Covid Saude Gov	https://covid.saude.gov.br/	Base com informações por indivíduos contaminados com COVID-19
2. Registro de ocupação hospitalar COVID-19	https://opendatasus.saude.gov.br/dataset/registro-de-ocupacao-hospitalar/resource/f9391f7c-9775-4fac-a3ce-bf384e2674c2?view_id=04f2877a-2ea0-4b59-b630-5c530d8db3f2	Base com informações de disponibilidade de leitos de UTI e Clínico.
3. Tabelas por municípios - UIT, respiradores, médicos e enfermeiros	https://agenciadenoticias.ibge.gov.br/agencia-detalle-e-midia.html?view=mediaibge&catid=2103&id=3702	Base com informações sobre disponibilidade de respiradores, leitos, médicos e enfermeiros.
4. PIB		Base com informações de distribuição do PIB pelo território nacional
5. Trabalho Informal		Base com informações sobre quantidade de trabalhadores informais por município e estado da federação.

Após todo o trabalho de centralização, obtivemos um banco de dados de 1.189.743 linhas e 194 colunas. Foi realizada uma primeira etapa de retirada de variáveis redundantes, a exemplo de códigos que representam os municípios, nomes e códigos de hospitais e outros, de modo que ao final permaneceram 159 colunas no banco de dados.

Variável Resposta: Evolução

Esta variável está presente na base Covid Saude Gov, nossa base principal, cujo dicionário pode ser encontrado neste [link](#). No dicionário ela está definida como na tabela a seguir:

Nome do Campo	Tipo	Categoria	Descrição	Características	DBF
77- Evolução do caso	Varchar2(1)	1 - Cura 2 - Óbito 3 - Óbito por outras causas 9 - Ignorado	Evolução do caso	Campo Essencial	EVOLUCAO

As quantidades de cada categoria presente na base de dados e suas proporções em relação ao total estão mostradas na tabela a seguir.

Código	Legenda	Qtd	Proporção
1	Cura	696.675	66,85%
2	Óbito	300.504	28,84%
3	Óbito por outras	13.939	1,34%
9	Ignorado	31.028	2,98%

A variável *target* da nossa análise será morte ou não morte. Portanto, 28,84% do banco de dados corresponde a “1” e o restante “0” para efeito da modelagem de dados.

Correlações

Variáveis	Correlação com a resposta	Correlação com as respostas módulo	Ranking Correlação
resposta	1	1	
NU_IDADE_N	0,3082483341	0,3082483341	1º
EVOLUCAO	0,2040893051	0,2040893051	2º
PCR_FLUBLI	0,1634213517	0,1634213517	3º
SUPPORT_VEN	-0,1454093278	0,1454093278	4º

CLASSI_FIN	0,1374140431	0,1374140431	5°
UTI	-0,1072207522	0,1072207522	6°
PCR_RESUL	-0,09570849905	0,09570849905	7°
VACINA	0,08607407723	0,08607407723	8°
GARGANTA	0,07260183595	0,07260183595	9°
TOSSE	0,0721592907	0,0721592907	10°
TP_FLU_AN	-0,07072355424	0,07072355424	11°
Trabalhadores informais no Interior (julho 2019)	-0,06812237426	0,06812237426	12°
DIARREIA	0,06677594291	0,06677594291	13°
PERD_PALA	0,06417520664	0,06417520664	14°
VOMITO	0,06341441284	0,06341441284	15°
PERD_OLFT	0,06339685174	0,06339685174	16°
CO_MUN_RES	-0,06306884822	0,06306884822	17°
Trabalhadores informais na unidade da Federação (julho 2019)	-0,06101799955	0,06101799955	18°
FEBRE	0,05975408567	0,05975408567	19°
OUTRO_SIN	0,0594066872	0,0594066872	20°
POS_AN_FLU	0,0518077837	0,0518077837	21°
DOR_ABD	0,0517385318	0,0517385318	22°
ASMA	0,05069335218	0,05069335218	23°
Trabalhadores informais Capital (julho 2019)	-0,04978667898	0,04978667898	24°
ANTIVIRAL	0,04867723188	0,04867723188	25°
Trabalhadores informais nas Regiões Metropolitanas (julho 2019)	-0,04844008562	0,04844008562	26°
TP_FLU_PCR	-0,04600938569	0,04600938569	27°
AVE_SUINO	0,04571294026	0,04571294026	28°
FADIGA	0,0424554866	0,0424554866	29°
M_AMAMENTA	0,04236382046	0,04236382046	30°
TP_SOR	-0,04203243736	0,04203243736	31°
SATURACAO	-0,04136862013	0,04136862013	32°
MAE_VAC	0,03745943461	0,03745943461	33°
CS_RACA	-0,03420049462	0,03420049462	34°
PCR_FLUASU	0,03400286121	0,03400286121	35°
RAIOX_RES	0,03173379612	0,03173379612	36°
RES_AN	-0,03139745334	0,03139745334	37°
CO_PS_VGM	-0,03120752121	0,03120752121	38°
RES_IGM	-0,03072759831	0,03072759831	39°
TP_AM_SOR	-0,03051759515	0,03051759515	40°
TP_ANTIVIR	-0,03042839646	0,03042839646	41°
DISPNEIA	-0,029501274	0,029501274	42°
Médicos a cada 100 mil habitantes	-0,0293795968	0,0293795968	43°

RES_IGG	-0,02921669868	0,02921669868	44°
CARDIOPATI	-0,02816408404	0,02816408404	45°
DESC_RESP	-0,02812972104	0,02812972104	46°
OBESIDADE	0,02746086569	0,02746086569	47°
CS_ESCOL_N	-0,02739727313	0,02739727313	48°
CRITERIO	0,0269290762	0,0269290762	49°
Leitos de UTI do SUS	-0,02629418404	0,02629418404	50°
DIABETES	-0,02532953846	0,02532953846	51°
saidaSuspeitaObitos	0,02532524173	0,02532524173	52°
HOSPITAL	0,02454511517	0,02454511517	53°
Total de médicos	-0,0239384145	0,0239384145	54°
HISTO_VGM	0,02294157273	0,02294157273	55°
NOSOCOMIAL	0,02199283035	0,02199283035	56°
Médicos do SUS	-0,021808481	0,021808481	57°
HEMATOLOGI	0,02157159506	0,02157159506	58°
Enfermeiros a cada 100 mil habitantes	-0,02136154593	0,02136154593	59°
TP_AMOSTRA	0,02060659097	0,02060659097	60°
SIND_DOWN	0,01962484587	0,01962484587	61°
Total de Enfermeiros	-0,01941945497	0,01941945497	62°
Médicos do SUS a cada 100 mil habitantes	-0,01941889783	0,01941889783	63°
Total de Respiradores	-0,01712575613	0,01712575613	64°
HEPATICA	0,01692491634	0,01692491634	65°
OUT_MORBI	-0,01637837143	0,01637837143	66°
População Estimada (2019)	-0,01637212932	0,01637212932	67°
SURTO_SG	0,01623339773	0,01623339773	68°
PUERPERA	0,0158402385	0,0158402385	69°
Respiradores a cada 100 mil habitantes	-0,01533942096	0,01533942096	70°
Total de Respiradores do SUS	-0,01531217304	0,01531217304	71°
RES_IGA	0,01473156065	0,01473156065	72°
Enfermeiros do SUS	-0,01391011142	0,01391011142	73°
IMUNODEPRE	0,01384490237	0,01384490237	74°
Leitos do SUS a cada 100 mil habitantes	-0,01373816751	0,01373816751	75°
ocupacaoConfirmadoUti	0,01243504431	0,01243504431	76°
AMOSTRA	-0,01146864826	0,01146864826	77°
Total de Leitos de UTI	-0,01055184792	0,01055184792	78°
ocupacaoSuspeitoUti	0,01004563541	0,01004563541	79°
Respiradores do SUS a cada 100 mil habitantes	-0,009826774379	0,009826774379	80°
POS_AN_OUT	0,008775331278	0,008775331278	81°
Enfermeiros do SUS a cada 100 mil habitantes	-0,007196245563	0,007196245563	82°

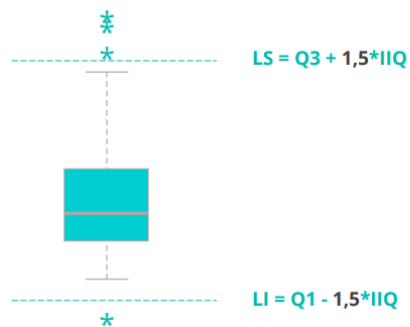
ocupacaoSuspeitoCli	-0,007158112347	0,007158112347	83°
PNEUMOPATI	0,00703267662	0,00703267662	84°
RENAL	-0,006201024736	0,006201024736	85°
Leitos de UTI a cada 100 mil habitantes	-0,00609074339	0,00609074339	86°
POS_PCRFLU	0,005661185767	0,005661185767	87°
TOMO_RES	0,005589002653	0,005589002653	88°
CS_GESTANT	0,005349243581	0,005349243581	89°
TP_TES_AN	-0,004865579655	0,004865579655	90°
ocupacaoConfirmadoCli	0,004813526427	0,004813526427	91°
POS_PCROUT	0,003335672594	0,003335672594	92°
saidaSuspeitaAltas	-0,002900310351	0,002900310351	93°
CS_ZONA	-0,00271891382	0,00271891382	94°
saidaConfirmadaAltas	-0,002208295127	0,002208295127	95°
saidaConfirmadaObitos	0,001573690611	0,001573690611	96°
NEUROLOGIC	-0,0004460987556	0,0004460987556	97°

Covariáveis

Explicitaremos no relatório apenas as análises feitas para as variáveis com maior correlação com a resposta e mais emblemáticas sobre nossa óptica de interpretação, a exemplo de comorbidades pré existentes. Para todas as demais foram obtidas as mesmas métricas.

Para as variáveis contínuas, foram levantadas, além das volumetrias cruzadas, medidas de tendência central: média, mediana e quartis. Para entendimentos dos pontos outliers, estabelecemos limites inferiores e superiores por meio do intervalo interquartil (intervalo entre Q1 e Q3), conforme figura abaixo. Posteriormente, trataremos estes pontos atípicos os substituindo por tais limites encontrados. Mapeamos o % de não preenchimento das variáveis e trataremos esta característica caso a caso.

Outlier ou dado discrepante são as observações muito diferentes das demais, consideradas ponto fora da curva. Geralmente, são classificados como *outlier* quando estão acima do Limite Superior (LS) ou abaixo do Limite Inferior (LI).



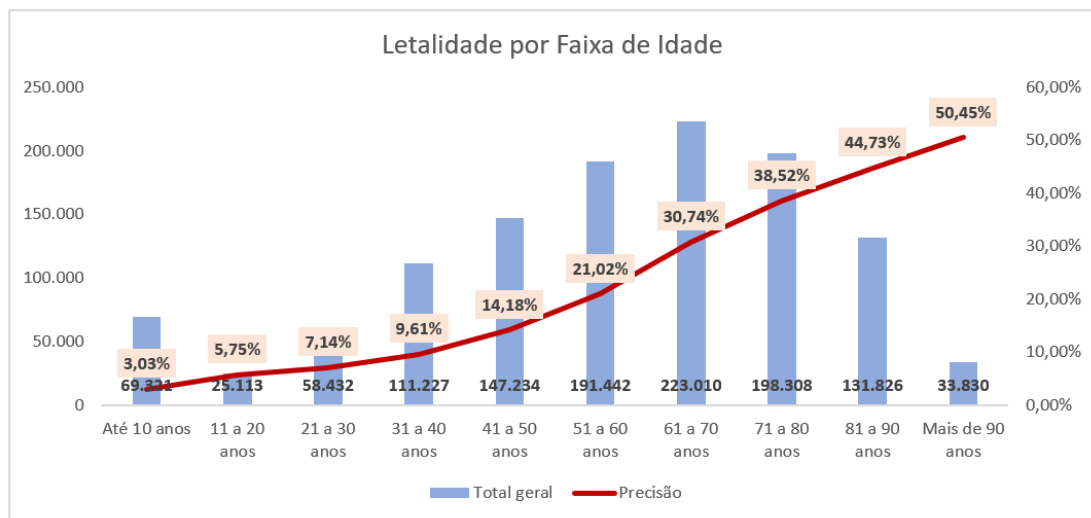
O valor 1,5 é comumente sugerido na literatura, porém na prática ele pode ser parametrizado de acordo com a severidade desejada na detecção do *outlier*.

Também será entregue, junto ao trabalho, um dicionário com o significado das variáveis que estão sendo utilizadas.

1. Idade

Essa é a variável mais relevante nos estudos feitos até o momento, sendo o principal critério para a vacinação atual. De fato, foi a variável com maior correlação com a resposta. À medida que a idade aumenta, a proporção de mortes por faixa de idade também aumenta consideravelmente.

Faixa de Idade	Não Óbito por COVID	Óbitos por COVID	Total geral	Precisão
Até 10 anos	67.224	2.097	69.321	3,03%
11 a 20 anos	23.668	1.445	25.113	5,75%
21 a 30 anos	54.259	4.173	58.432	7,14%
31 a 40 anos	100.540	10.687	111.227	9,61%
41 a 50 anos	126.357	20.877	147.234	14,18%
51 a 60 anos	151.199	40.243	191.442	21,02%
61 a 70 anos	154.463	68.547	223.010	30,74%
71 a 80 anos	121.913	76.395	198.308	38,52%
81 a 90 anos	72.854	58.972	131.826	44,73%
Mais de 90 anos	16.762	17.068	33.830	50,45%

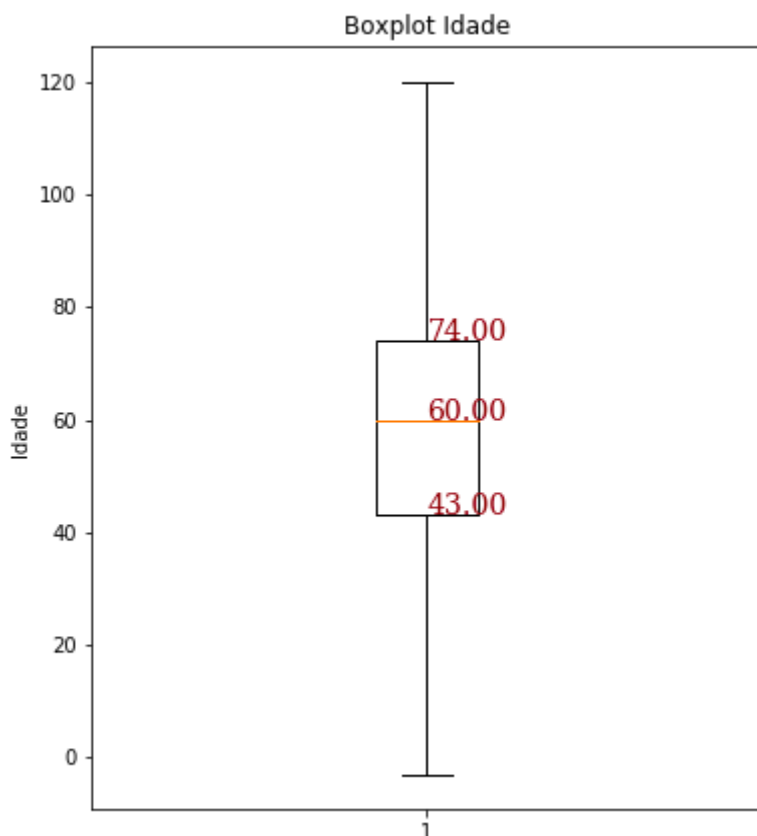


MEDIA = 56.744226274077675
 MEDIANA = 60.0
 MODA = 66
 DESVIO_PADRAO = 22.49598017517659

Q1 = 43.0
 Q2 = 60.0
 Q3 = 74.0
 Q4 = 148

Limite Superior = 120.5
 Limite Inferior = -3.5

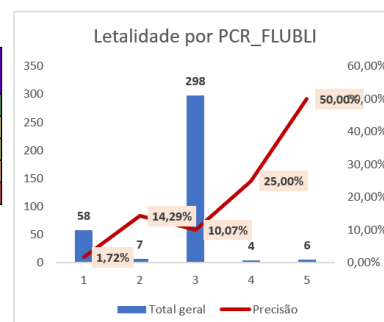
Correlacao = 0.3082483341157743
 Quantidade de outliers 10
 Quantidade valores faltantes 0



2. PCR_FLUBLI

Apesar de ser a segunda maior correlação com a resposta, apresenta um número muito grande de missing, tendo apenas 0,03% dos pacientes com este campo preenchido, podendo ser pouco relevante como discriminante no algoritmo em comparação às demais variáveis.

PCR_FLUBLI	Não Óbito por COVID	Óbitos por COVID	Total geral	Precisão
1	57	1	58	1,72%
2	6	1	7	14,29%
3	268	30	298	10,07%
4	3	1	4	25,00%
5	3	3	6	50,00%

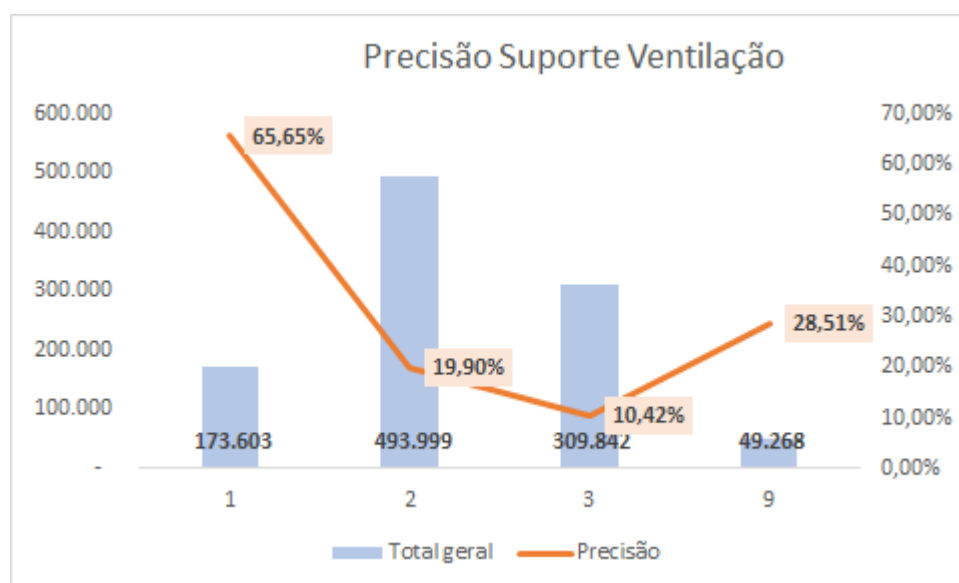


3. Suporte Ventilação

- 14% dos valores não preenchidos
- Entendemos que esta variável não tem potencial para o modelo, pois apesar de ser uma variável dependente com alto potencial para explicar a letalidade, a necessidade de suporte para respirar ocorre quando o

indivíduo já contraiu a doença, não sendo relevante na predição de grupos que deveriam ser priorizados na vacinação.

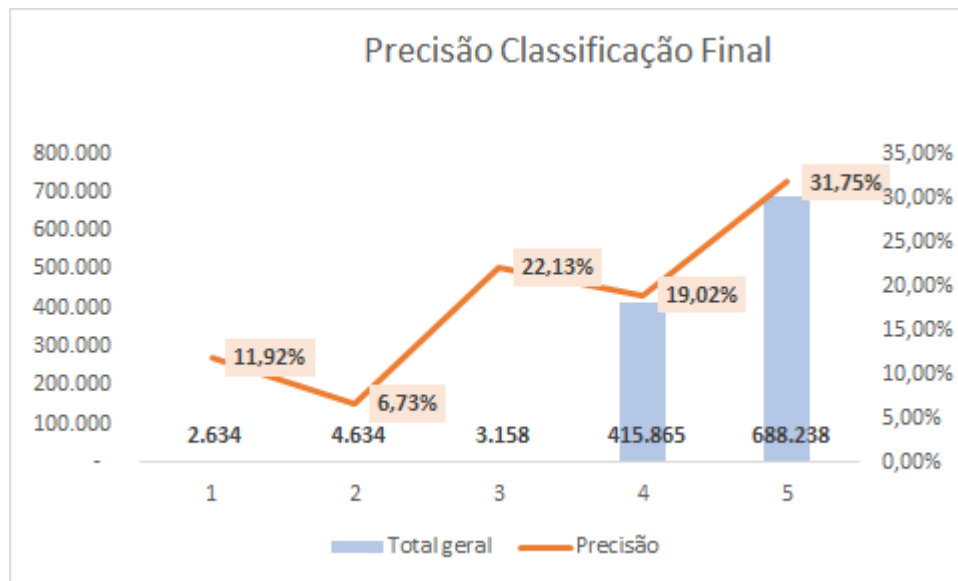
SUPPORT_VEN	Não Óbito por COVID	Óbitos por COVID	Total geral	Precisão
1	59.633	113.970	173.603	65,65%
2	395.676	98.323	493.999	19,90%
3	277.546	32.296	309.842	10,42%
9	35.223	14.045	49.268	28,51%



4. Classificação

- 6% dos valores não preenchidos.
- Entendemos que esta variável não tem potencial para o modelo, pois apesar de ser uma variável dependente com alto potencial para explicar a letalidade, a classificação é uma condição do paciente já contaminado.

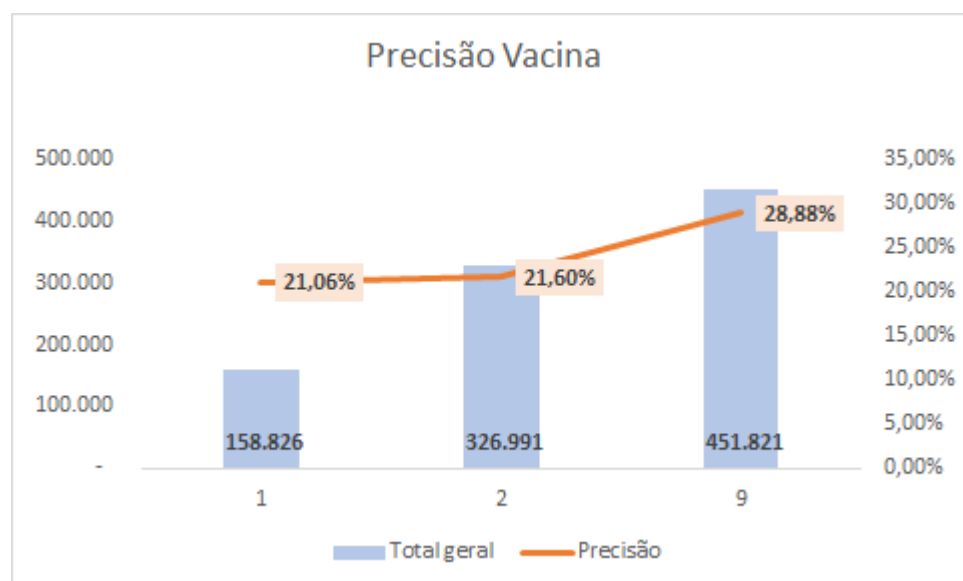
CLASSI_FIN	Não Óbito por COVID	Óbitos por COVID	Total geral	Precisão
1	2.320	314	2.634	11,92%
2	4.322	312	4.634	6,73%
3	2.459	699	3.158	22,13%
4	336.782	79.083	415.865	19,02%
5	469.735	218.503	688.238	31,75%



5. Vacina

- 21% dos valores não preenchidos.
- Entendemos que esta variável não tem potencial para o modelo, pois apesar de ser uma variável dependente com alto potencial para explicar a letalidade, a vacina não é relevante para entendermos qual grupo deve ser priorizado.

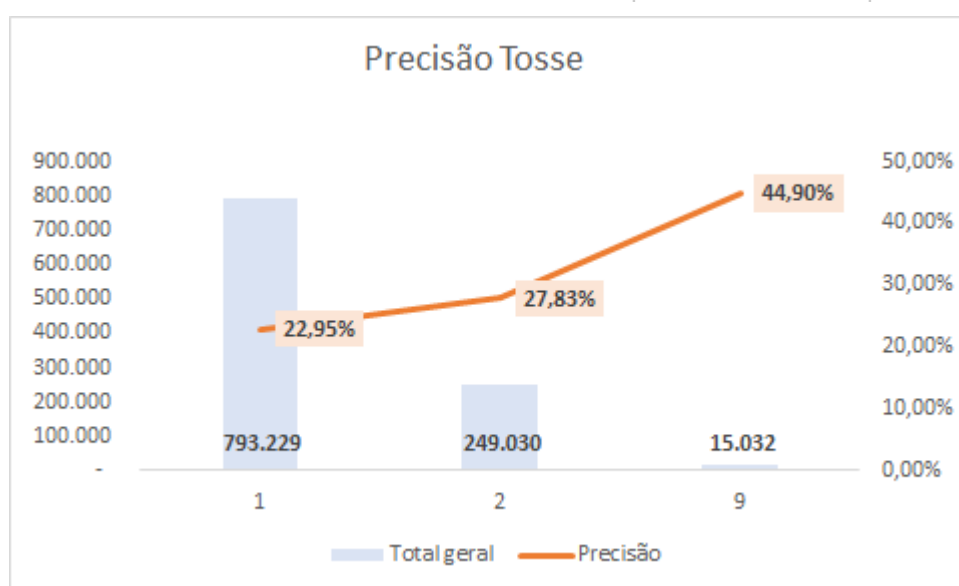
VACINA	Não Óbito por COVID	Óbitos por COVID	Total geral	Precisão
1	125.378	33.448	158.826	21,06%
2	256.351	70.640	326.991	21,60%
9	321.316	130.505	451.821	28,88%



6. TOSSE

- 11% dos valores não preenchidos.
- Entendemos que esta variável não tem potencial para o modelo, pois apesar de ser uma variável dependente com alto potencial para explicar a letalidade, a presença do sintoma tosse ocorre quando o indivíduo já contraiu a doença, não sendo relevante na predição de grupos que deveriam ser priorizados na vacinação.

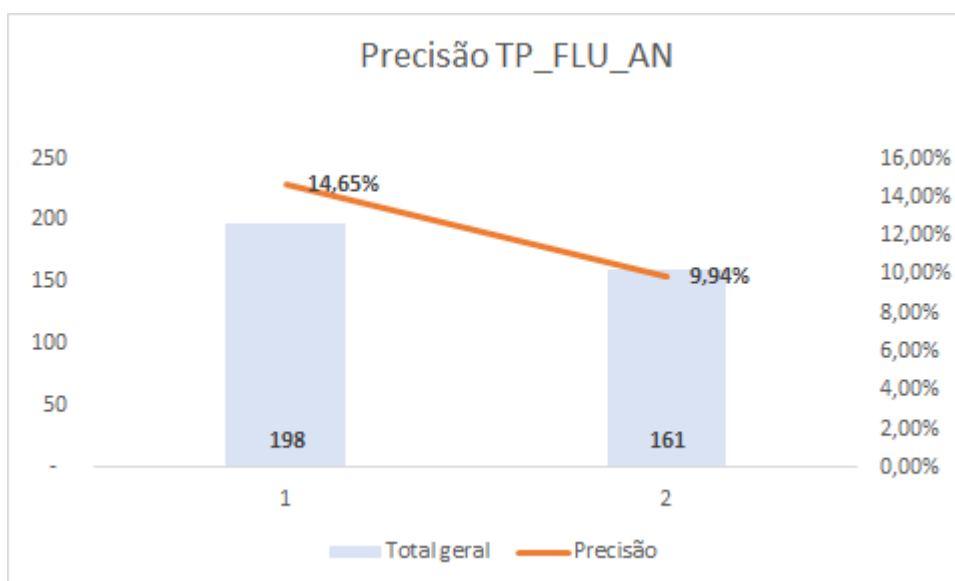
TOSSE	Não Óbito por COVID	Óbitos por COVID	Total geral	Precisão
1	611.174	182.055	793.229	22,95%
2	179.722	69.308	249.030	27,83%
9	8.282	6.750	15.032	44,90%



7. TP_FLU_AN

- Menos de 1% de preenchimento.
- Entendemos que esta variável não tem potencial de colaborar na inferência.

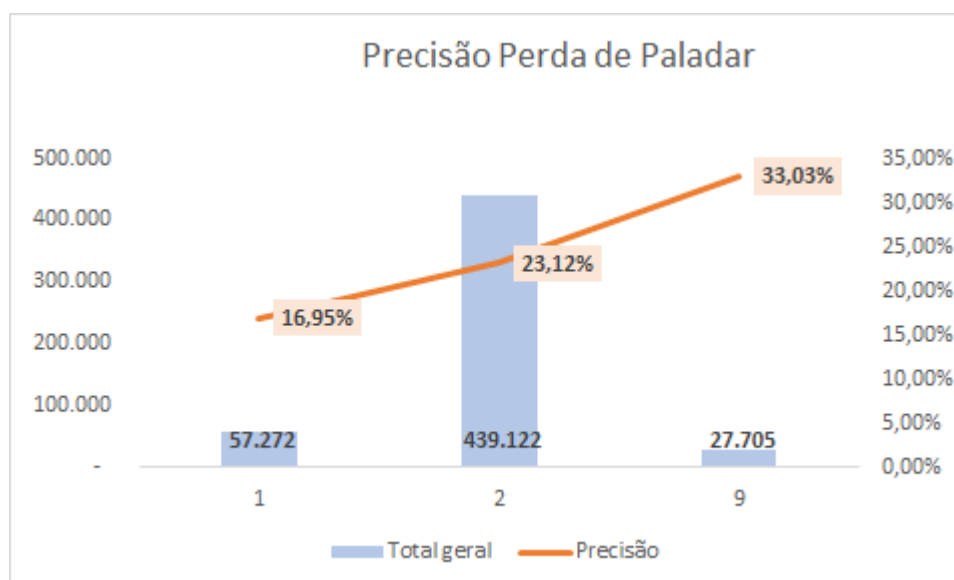
TP_FLU_AN	Não Óbito por COVID	Óbitos por COVID	Total geral	Precisão
1	169	29	198	14,65%
2	145	16	161	9,94%



8. Perda de Paladar

- Apenas 39% de preenchimento.
- Entendemos que esta variável não tem potencial de colaborar na inferência por ser um sintoma típico de quando o paciente já está contaminado.

PERD_PALADAR	Não Óbito por COVID	Óbitos por COVID	Total geral	Precisão
1	47.562	9.710	57.272	16,95%
2	337.603	101.519	439.122	23,12%
9	18.555	9.150	27.705	33,03%

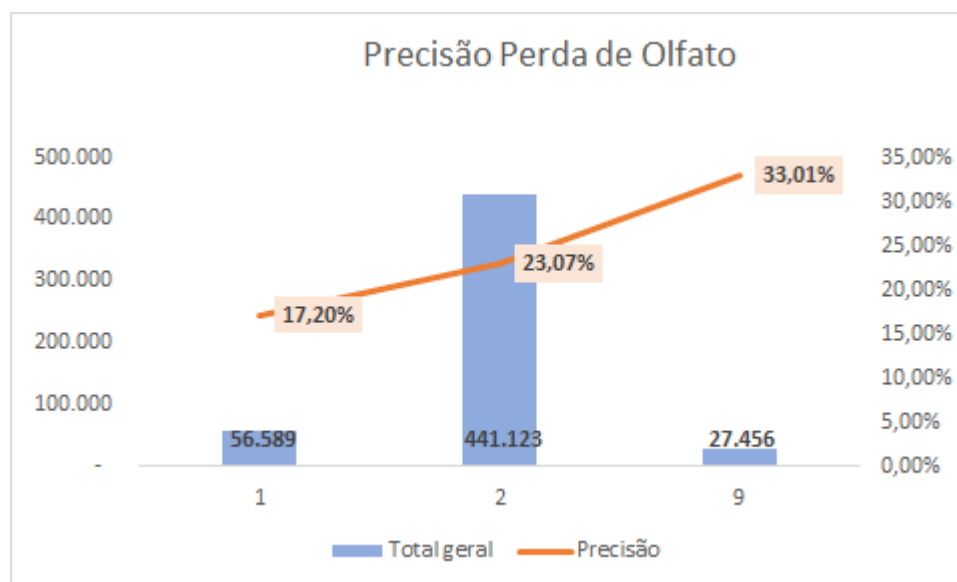


9. Perda de Olfato

- Apenas 39% de preenchimento.

- Entendemos que esta variável não tem potencial de colaborar na inferência por ser um sintoma típico de quando o paciente já está contaminado.

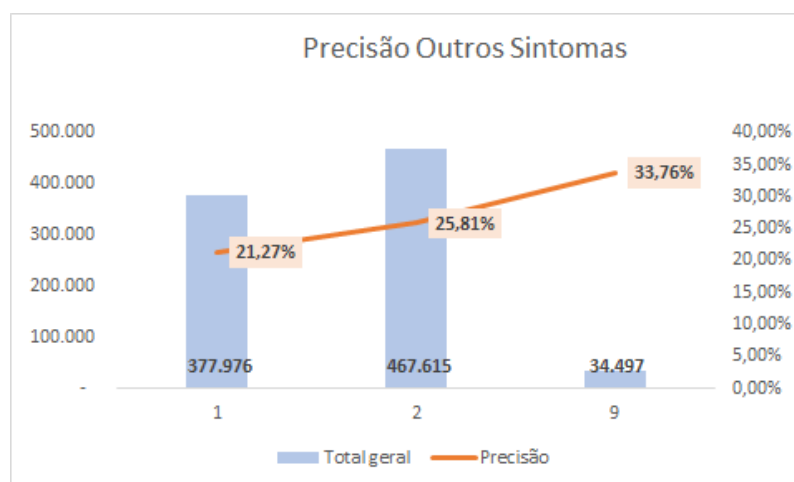
PERD_OLFATO	Não Óbito por COVID	Óbitos por COVID	Total geral	Precisão
1	46.858	9.731	56.589	17,20%
2	339.365	101.758	441.123	23,07%
9	18.392	9.064	27.456	33,01%



10. Outros Sintomas

- Apenas 42% de preenchimento.
- Entendemos que esta variável não tem potencial de colaborar na inferência por ser um sintoma de quando o paciente já está contaminado.

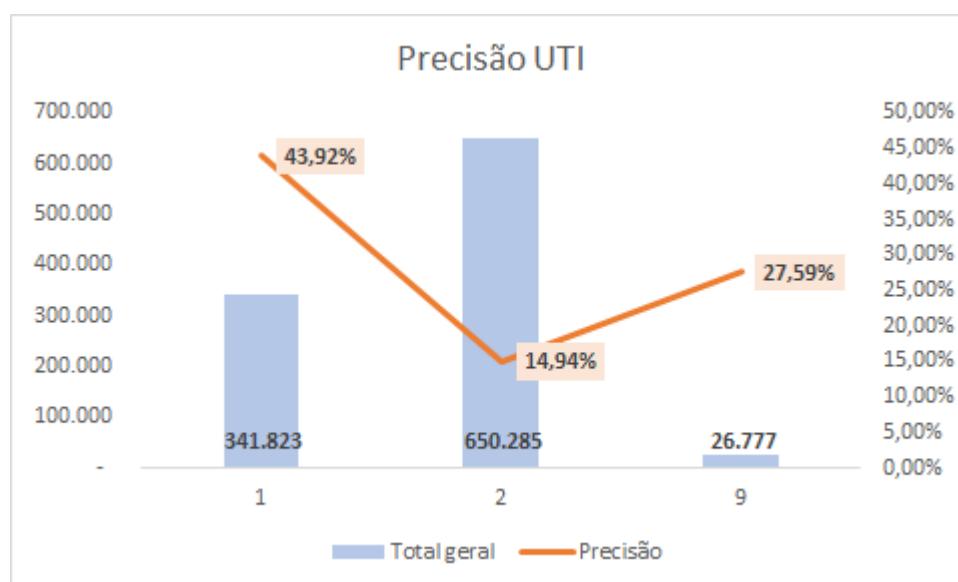
OUTRSO_SINTOMAS	Não Óbito por COVID	Óbitos por COVID	Total geral	Precisão
1	297.599	80.377	377.976	21,27%
2	346.908	120.707	467.615	25,81%
9	22.851	11.646	34.497	33,76%



11. UTI

- 14% de não preenchimento.
- Analogamente ao suporte respiratório, esta variável tem forte correlação com letalidade, mas não faz sentido para a predição.

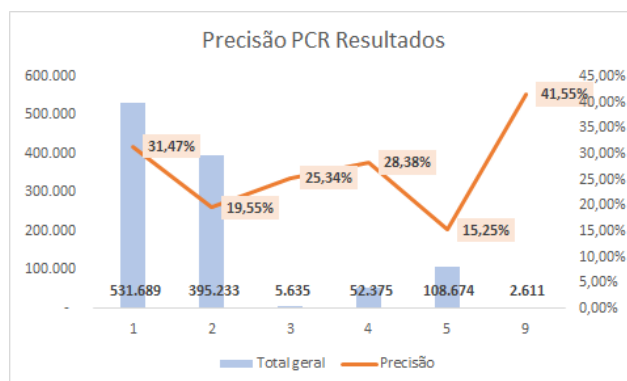
UTI	Não Óbito por COVID	Óbitos por COVID	Total geral	Precisão
1	191.685	150.138	341.823	43,92%
2	553.126	97.159	650.285	14,94%
9	19.390	7.387	26.777	27,59%



12. PCR_RESULTADOS

- 8% de não preenchimento.
- Resultado de exames relacionados após suspeita de COVID. Inválido para predição em questão.

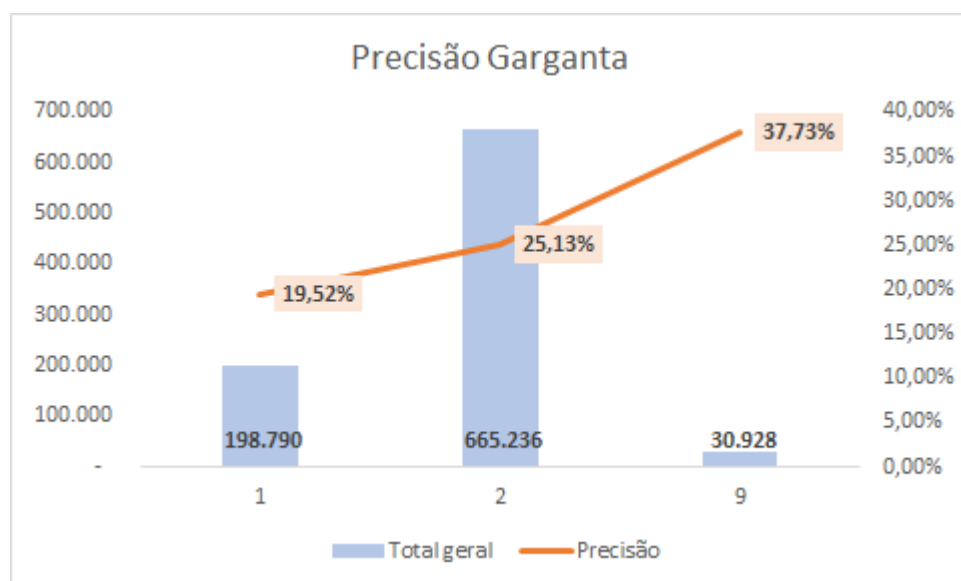
PCR_RESUL	Não Óbito por COVID	Obitos por COVID	Total geral	Precisão
1	364.374	167.315	531.689	31,47%
2	317.949	77.284	395.233	19,55%
3	4.207	1.428	5.635	25,34%
4	37.512	14.863	52.375	28,38%
5	92.106	16.568	108.674	15,25%
9	1.526	1.085	2.611	41,55%



13. Sintoma Garganta

- 25% de não preenchimento.
- Por se tratar de um sintoma, ocorre após a contração da doença, sendo inválido para inferência.

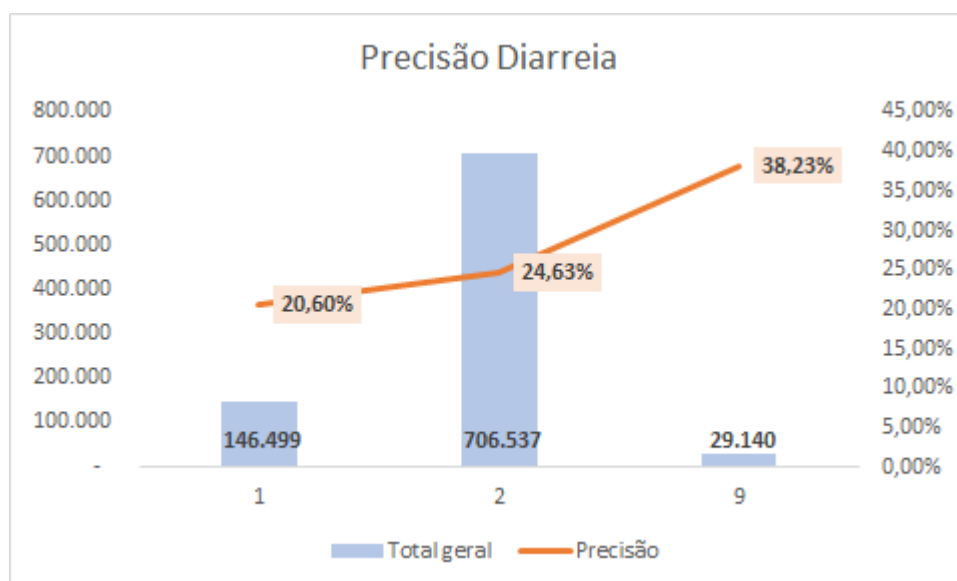
GARGANTA	Não Óbito por COVID	Obitos por COVID	Total geral	Precisão
1	159.993	38.797	198.790	19,52%
2	498.084	167.152	665.236	25,13%
9	19.260	11.668	30.928	37,73%



14. Sintoma Diarréia

- 26% de não preenchimento.
- Por se tratar de um sintoma, ocorre após a contração da doença, sendo inválido para inferência.

DIARREIA	Não Óbito por COVID	Obitos por COVID	Total geral	Precisão
1	116.317	30.182	146.499	20,60%
2	532.550	173.987	706.537	24,63%
9	18.001	11.139	29.140	38,23%

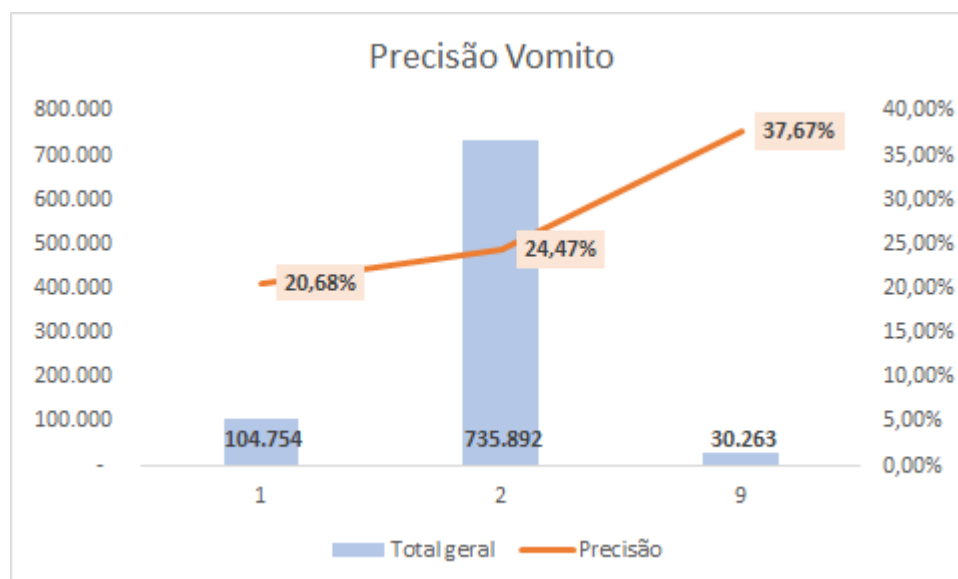


15. Sintoma Vômito

- 27% de não preenchimento.

- Por se tratar de um sintoma, ocorre após a contração da doença, sendo inválido para inferência.

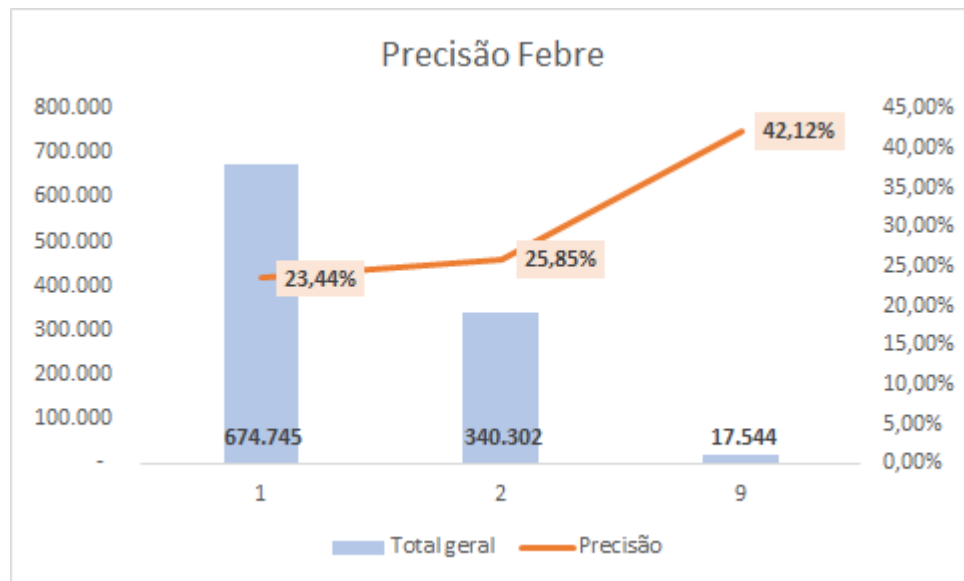
VOMITO	Não Óbito por COVID	Obitos por COVID	Total geral	Precisão
1	83.092	21.662	104.754	20,68%
2	555.826	180.066	735.892	24,47%
9	18.864	11.399	30.263	37,67%



16. Sintoma Febre

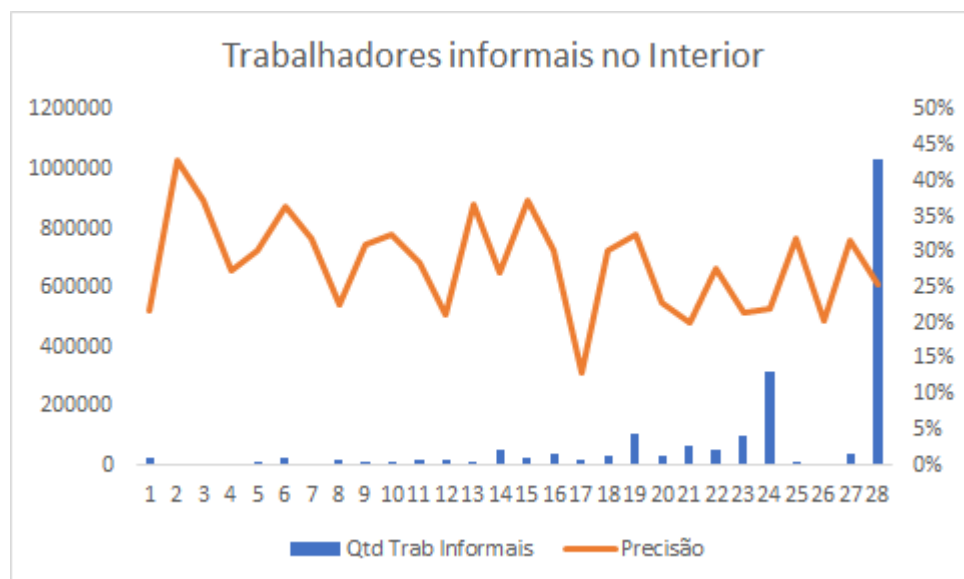
- 13% de não preenchimento.
- Por se tratar de um sintoma, ocorre após a contração da doença, sendo inválido para inferência.

FEBRE	Não Óbito por COVID	Obitos por COVID	Total geral	Precisão
1	516.618	158.127	674.745	23,44%
2	252.339	87.963	340.302	25,85%
9	10.154	7.390	17.544	42,12%



17. Quantidade de Trabalhadores Informais

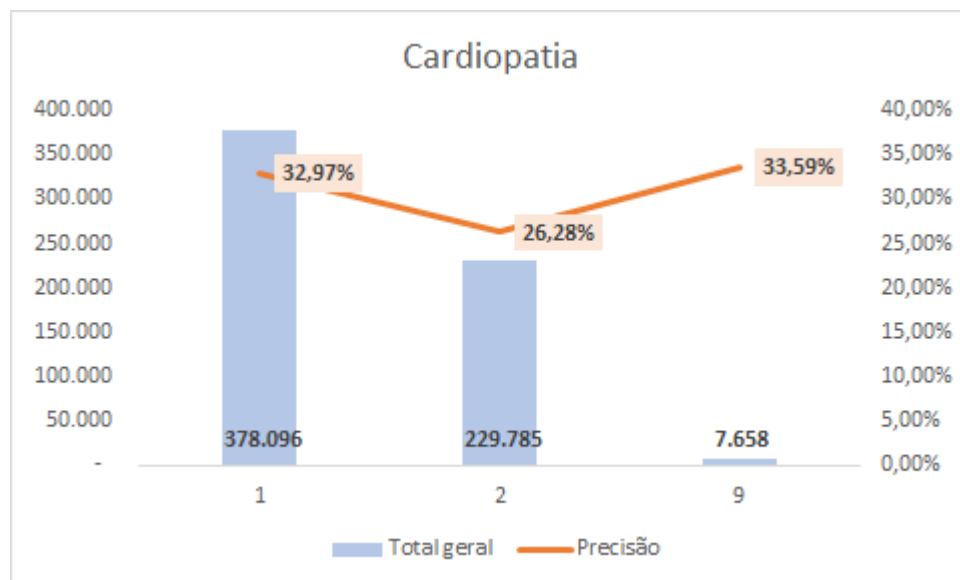
- 13% de não preenchimento.
- Não há uma relação clara de crescimento ou decréscimo à medida que a quantidade de trabalhadores informais cresce.



18. Cardiopatias

- 52% de preenchimento.
- É uma comorbidade prévia à contaminação, sendo um tipo de variável com alto potencial para explicar a prioridade ou não de uma vacina.

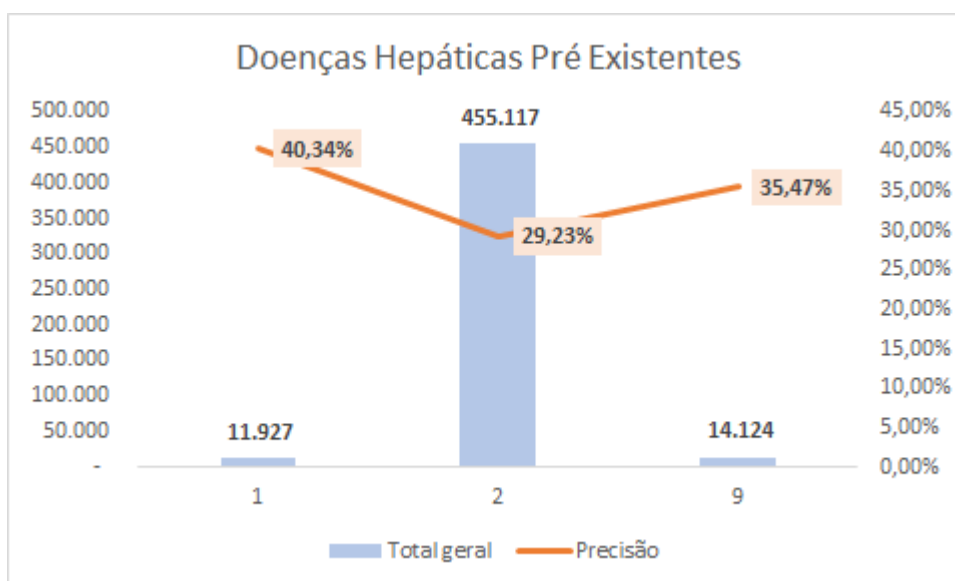
CARDIOPATIA	Não Óbito por COVID	Óbitos por COVID	Total geral	Precisão
1	253.419	124.677	378.096	32,97%
2	169.405	60.380	229.785	26,28%
9	5.086	2.572	7.658	33,59%



19. Doenças Hepáticas

- 40% de preenchimento.
- É uma comorbidade prévia à contaminação, sendo um tipo de variável com alto potencial para explicar a prioridade ou não de uma vacina.

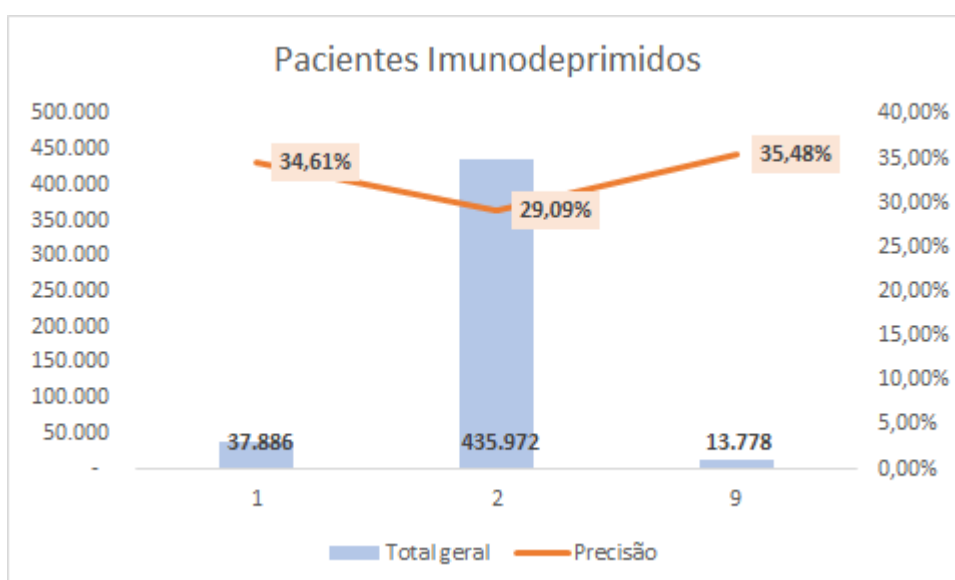
DOENÇAS HEPATICAS	Não Óbito por COVID	Óbitos por COVID	Total geral	Precisão
1	7.116	4.811	11.927	40,34%
2	322.100	133.017	455.117	29,23%
9	9.114	5.010	14.124	35,47%



20. Pacientes Imunodeprimidos

- 41% de preenchimento.
- É uma comorbidade prévia à contaminação, sendo um tipo de variável com alto potencial para explicar a prioridade ou não de uma vacina.

PACIENTES IMUNODEPRIMIDOS	Não Óbito por COVID	Óbitos por COVID	Total geral	Precisão
1	24.772	13.114	37.886	34,61%
2	309.163	126.809	435.972	29,09%
9	8.890	4.888	13.778	35,48%

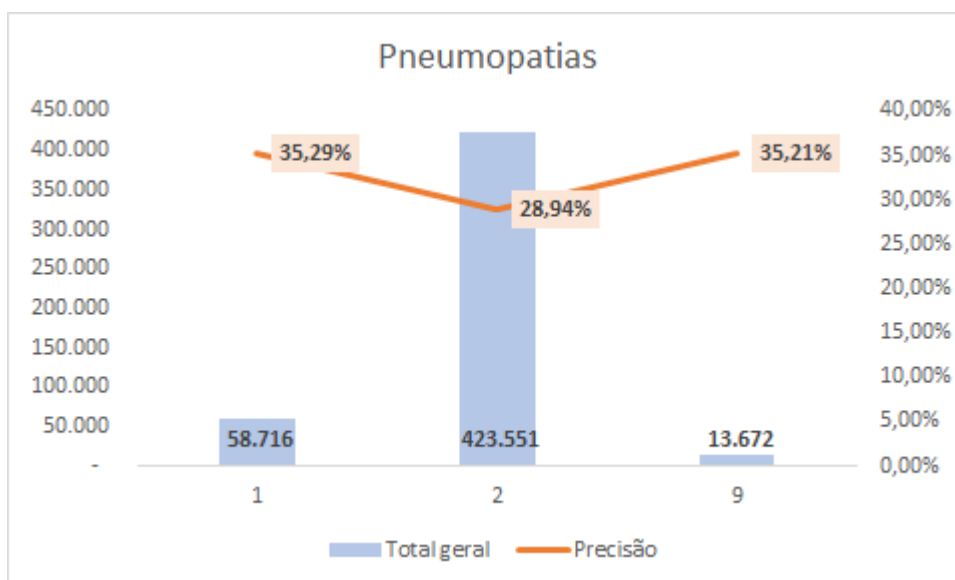


21. Pneumopatias

- 42% de preenchimento.

- É uma comorbidade prévia à contaminação, sendo um tipo de variável com alto potencial para explicar a prioridade ou não de uma vacina.

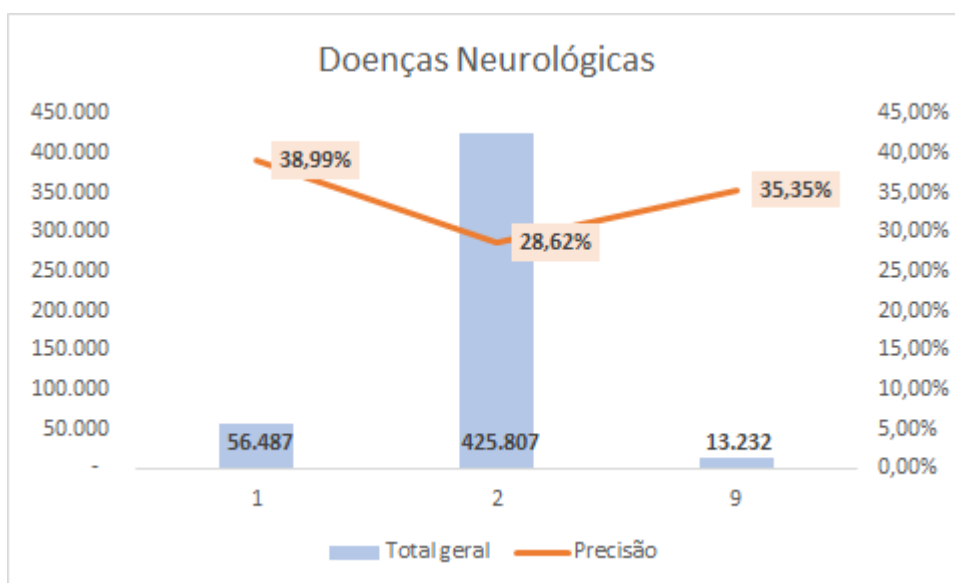
PNEUMOPATIAS	Não Óbito por COVID	Óbitos por COVID	Total geral	Precisão
1	37.996	20.720	58.716	35,29%
2	300.968	122.583	423.551	28,94%
9	8.858	4.814	13.672	35,21%



22. Doenças Neurológicas

- 42% de preenchimento.
- É uma comorbidade prévia à contaminação, sendo um tipo de variável com alto potencial para explicar a prioridade ou não de uma vacina.

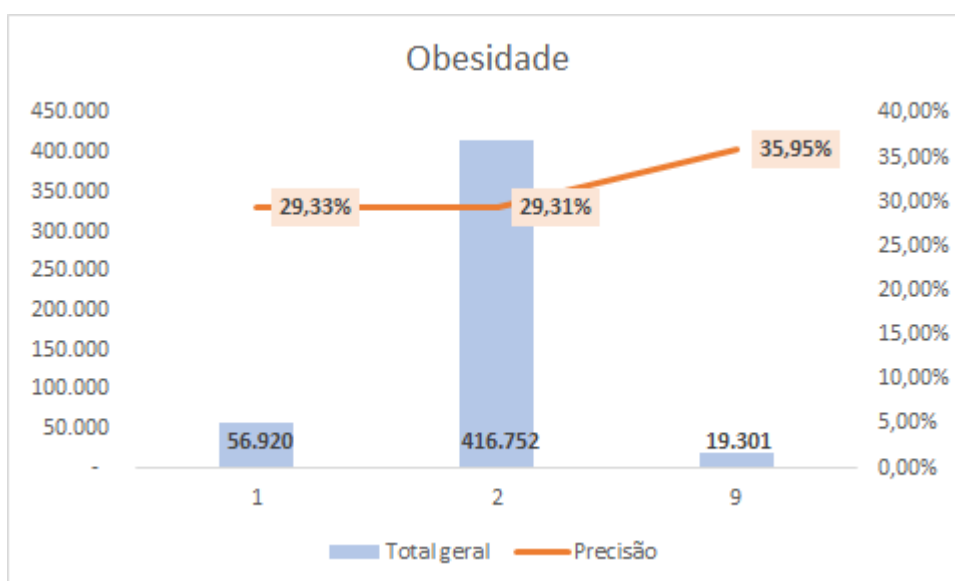
DOENÇAS NEUROLÓGICAS	Não Óbito por COVID	Óbitos por COVID	Total geral	Precisão
1	34.463	22.024	56.487	38,99%
2	303.926	121.881	425.807	28,62%
9	8.555	4.677	13.232	35,35%



23. Obesidade

- 41% de preenchimento
- É uma comorbidade prévia à contaminação, sendo um tipo de variável com alto potencial para explicar a prioridade ou não de uma vacina.

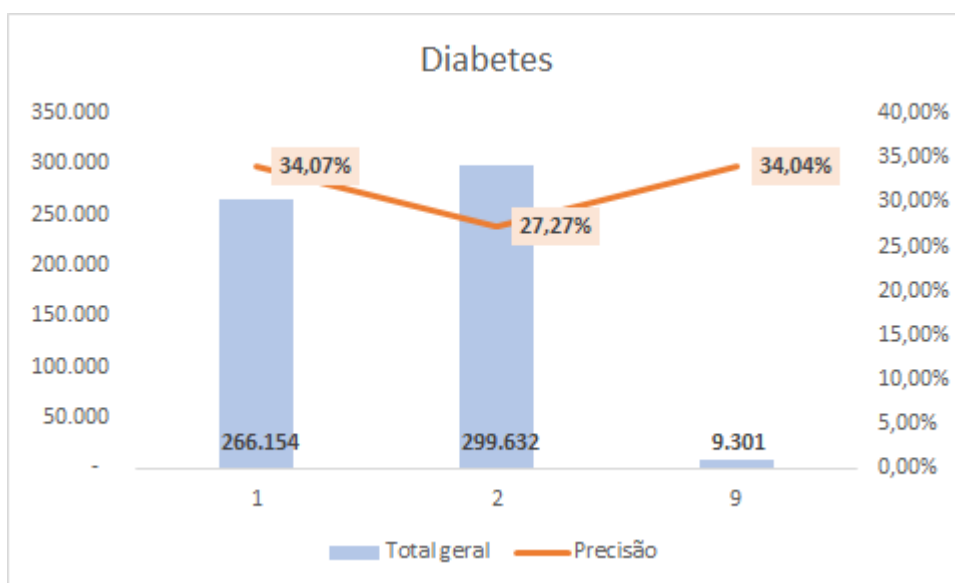
OBESIDADE	Não Óbito por COVID	Óbitos por COVID	Total geral	Precisão
1	40.228	16.692	56.920	29,33%
2	294.582	122.170	416.752	29,31%
9	12.362	6.939	19.301	35,95%



24. Diabetes

- 48% de preenchimento.
- É uma comorbidade prévia à contaminação, sendo um tipo de variável com alto potencial para explicar a prioridade ou não de uma vacina.

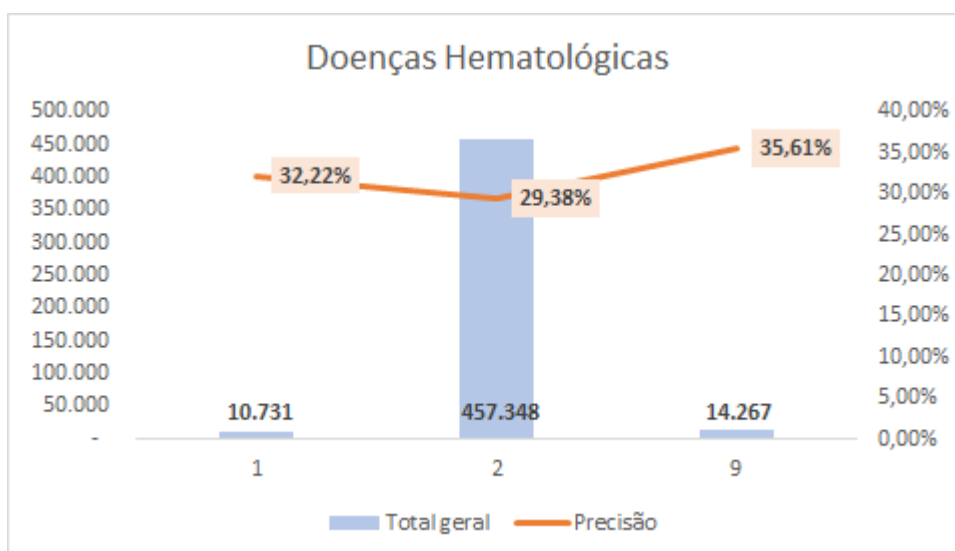
DIABETES	Não Óbito por COVID	Óbitos por COVID	Total geral	Precisão
1	175.487	90.667	266.154	34,07%
2	217.934	81.698	299.632	27,27%
9	6.135	3.166	9.301	34,04%



25. Doenças Hematológicas

- 41% de preenchimento.
- É uma comorbidade prévia à contaminação, sendo um tipo de variável com alto potencial para explicar a prioridade ou não de uma vacina.

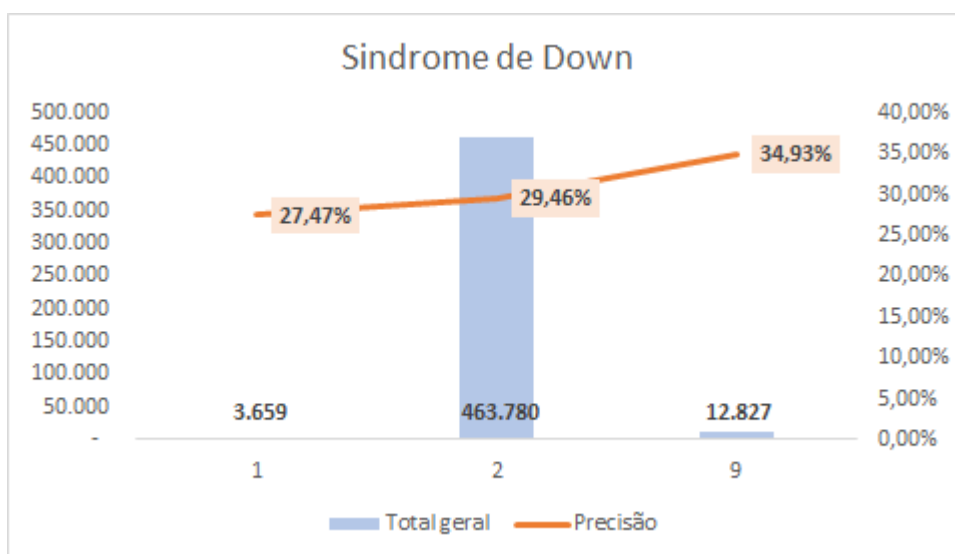
DOENÇAS HEMATOLOGICAS	Não Óbito por COVID	Óbitos por COVID	Total geral	Precisão
1	7.273	3.458	10.731	32,22%
2	322.997	134.351	457.348	29,38%
9	9.187	5.080	14.267	35,61%



26. Síndrome de Down

- 40% de preenchimento.
- É uma comorbidade prévia à contaminação, sendo um tipo de variável com alto potencial para explicar a prioridade ou não de uma vacina.

Síndrome de down	Não Óbito por COVID	Óbitos por COVID	Total geral	Precisão
1	2.654	1.005	3.659	27,47%
2	327.150	136.630	463.780	29,46%
9	8.346	4.481	12.827	34,93%

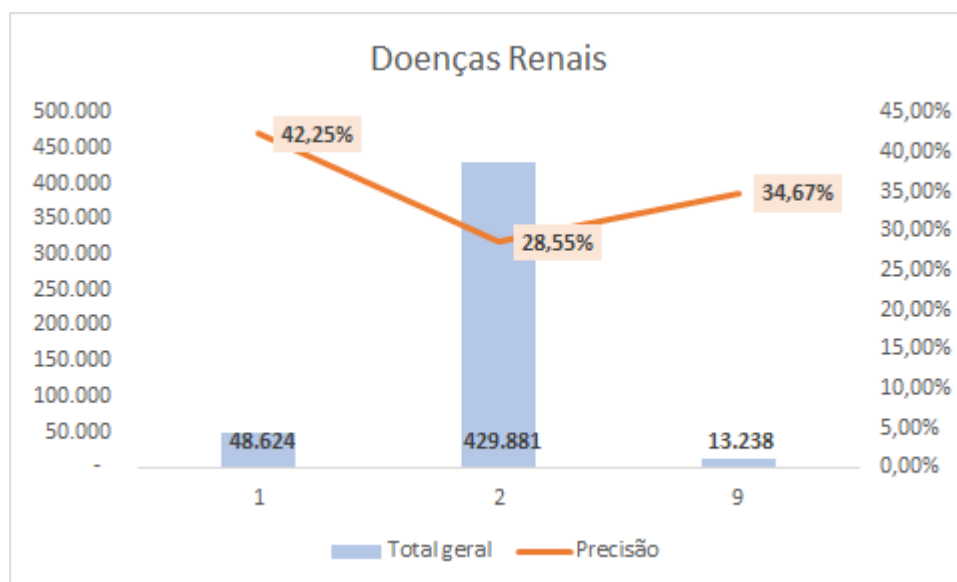


27. Doenças Renais

- 41% de preenchimento.

- É uma comorbidade prévia à contaminação, sendo um tipo de variável com alto potencial para explicar a prioridade ou não de uma vacina.

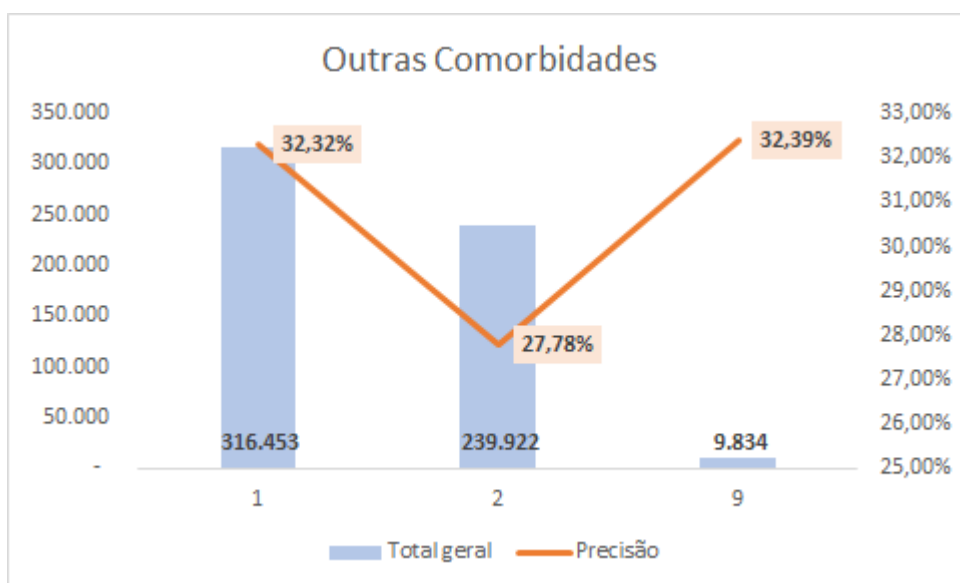
DOENÇAS RENAIS	Não Óbito por COVID	Óbitos por COVID	Total geral	Precisão
1	28.078	20.546	48.624	42,25%
2	307.139	122.742	429.881	28,55%
9	8.648	4.590	13.238	34,67%



28. Outras Comorbidades

- 48% de preenchimento.
- É uma comorbidade prévia à contaminação, sendo um tipo de variável com alto potencial para explicar a prioridade ou não de uma vacina.

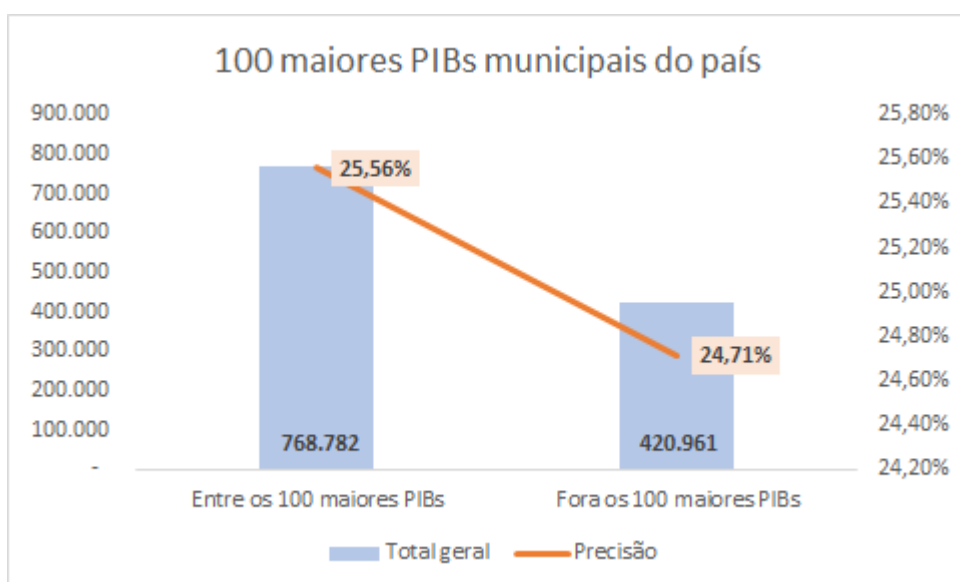
OUTRAS COMORBIDADES	Não Óbito por COVID	Óbitos por COVID	Total geral	Precisão
1	214.184	102.269	316.453	32,32%
2	173.271	66.651	239.922	27,78%
9	6.649	3.185	9.834	32,39%



29. 100 maiores PIBs do país

- Paciente pertence ou não a um município entre os 100 maiores PIBs do Brasil.
- Esperávamos uma relação inversa nessa variável, imaginando que municípios mais ricos conseguiriam letalidades menores devido a recursos.

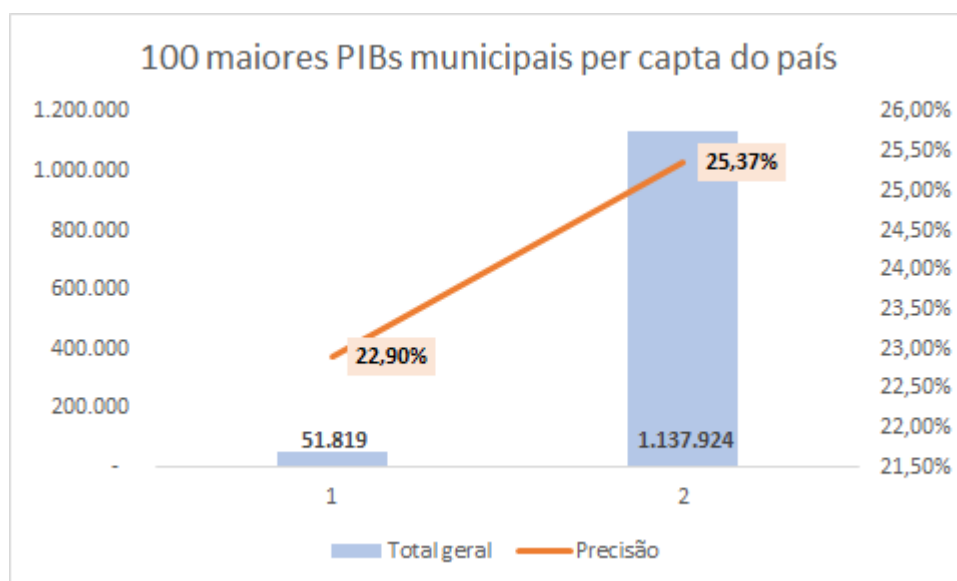
Municípios com 100 maiores PIBs do país	Não Óbito por COVID	Óbitos por COVID	Total geral	Precisão
Entre os 100 maiores PIBs	572.302	196.480	768.782	25,56%
Fora os 100 maiores PIBs	316.937	104.024	420.961	24,71%



30. 100 maiores PIBs per capita do país

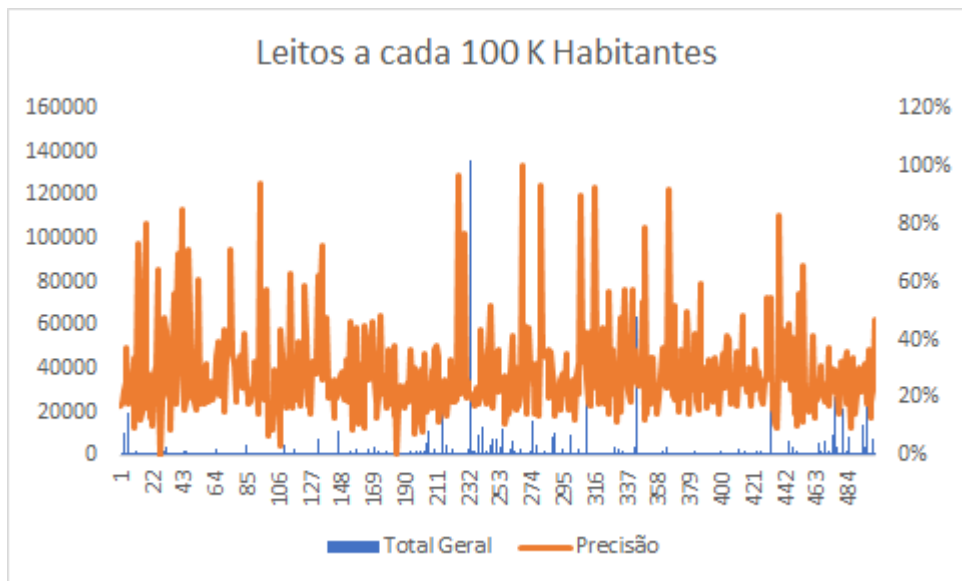
- Paciente pertence ou não a um município entre os 100 maiores PIBs per capita do Brasil.
- Para este conceito, a precisão se inverte e é bem interessante como faz sentido, pois para municípios com PIBs melhor distribuídos, a taxa de letalidade é menor.

Municípios com 100 maiores PIBs Per Capta do país	Não Óbito por COVID	Obitos por COVID	Total geral	Precisão
1	39.955	11.864	51.819	22,90%
2	849.284	288.640	1.137.924	25,37%



31. Leitos a cada 100K habitantes

- Não há uma relação inversa clara de que quanto mais leitos menor a letalidade.
- Variável com ampla variabilidade e muitos outliers. Após o tratamento, pode se tornar mais relevante.

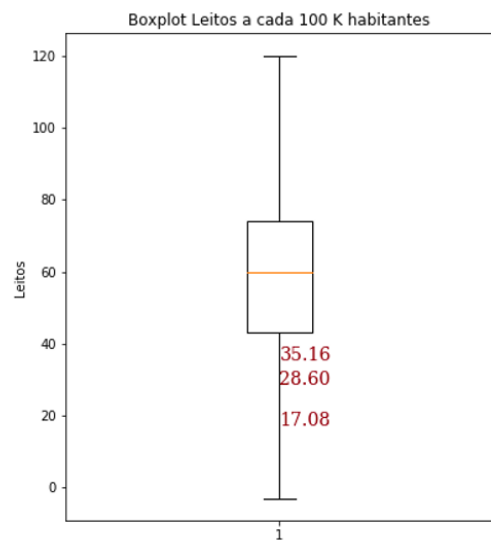


MEDIA = 28.633024816529346
 MEDIANA = 28.5993586528526
 MODA = 28.5993586528526
 DESVIO_PADRAO = 20.929047766253667

 Q1 = 17.9478944285957
 Q2 = 28.5993586528526
 Q3 = 36.6051699810984
 Q4 = 228.026848078059

 Limite Superior = 64.59108330985245
 Limite Inferior = -10.038018900158352

 Correlacao = -0.006090743389868601
 Quantidade de outliers 9759
 Quantidade valores faltantes 157502



32. Leitos a cada 100K habitantes

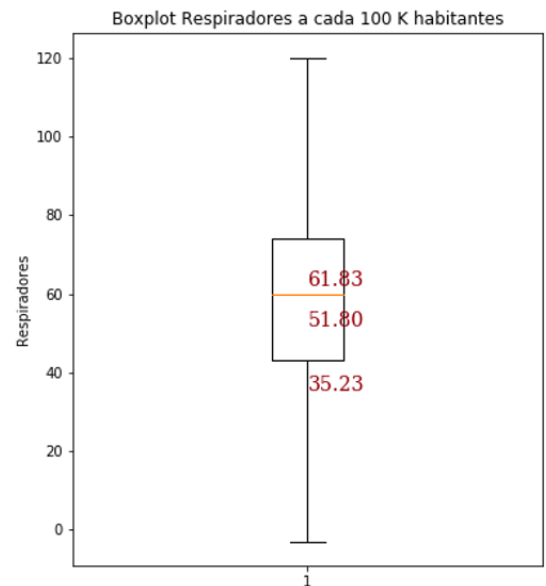
- Não há uma relação inversa clara de que quanto mais leitos menor a letalidade.
- Variável com ampla variabilidade e muitos outliers. Após o tratamento, pode se tornar mais relevante.

MEDIA = 51.34953771755825
MEDIANA = 54.679193698873796
MODA = 58.5454336806256
DESVIO_PADRAO = 24.84082172663392

Q1 = 36.0094064266253
Q2 = 54.679193698873796
Q3 = 62.7473246159213
Q4 = 175.754912804351

Limite Superior = 102.8542018998653
Limite Inferior = -4.097470857318704

Correlacao = -0.015339420960492118
Quantidade de outliers 18052
Quantidade valores faltantes 157502



Conclusões e Próximos Passos

Interessante notar como as variáveis que imaginávamos, devido a interpretação de seu significado, que serão muito correlacionadas com a letalidade não o são. Outras variáveis que não imaginávamos se fazem muito relevantes com a incidência de morte entre os pacientes contaminados. Outro ponto importante é que de fato a variável idade, filtro utilizado hoje para a priorização da vacina, é altamente discriminante para a nossa variável dependente.

Durante a análise, foi possível identificar também variáveis com alto potencial de discriminação, mas que, por se tratarem de sintomas da doença, não poderão ser consideradas. Isso porque pela cronologia lógica, os sintomas só aparecem após a contaminação e o que buscamos com esse trabalho é responder quais perfis de grupos deveriam ter suas prevenções priorizadas para que não fiquem doentes por apresentarem maior risco de morte.

Como próximos passos, faremos os processamentos de ETL, a fim de tratarmos e modelarmos cada uma das variáveis antes da aplicação de algoritmos preditivos.