Instituto de Educação Superior de Brasília - IESB
Ciência de Dados e Inteligência Artificial

Análise exploratória e visualização dos dados sobre COVID no Brasil e Distrito Federal

por

Victor Augusto Souza Resende

1922120027

Brasília - DF, 30 de novembro de 2020

Conteúdo

Lista de Figuras

1 COVID-19 - Novo Coronavírus

De acordo com o Ministério da Saúde Brasileiro[1], o Coronavírus é uma família de vírus que causam infecções respiratórias. Os primeiros coronavírus humanos foram identificados em meados da década de 1960. A maioria das pessoas se infecta com os coronavírus comuns ao longo da vida, sendo as crianças pequenas mais propensas a se infectarem com o tipo mais comum do vírus.

Porém, em dezembro de 2019, houve a transmissão de um novo coronavírus, denominado SARS-CoV-2, o qual foi identificado em Wuhan na China e causou a COVID-19, em seguida sendo transmitida e disseminada pelo mundo através dos humanos. De acordo com a organização Mundial da Saúde (OMS) o vírus possui sintomas parecidos com o de uma gripe normal, do qual incluem: Dor de cabeça, febre, tosse, dor de garganta e também perda de olfato ou paladar. Dessa maneira, caso o paciente não seja tratado da maneira correta pode haver risco de morte do mesmo, principalmente pacientes que já possui estado imunológico enfraquecido defino a outros fatores.

O vírus que causa a COVID-19 é transmitido principalmente por meio de gotículas geradas quando um indivíduo infectada tosse, espirra, secreta ou exala. Tais gotículas são pesadas para pairar sobre o ar, sendo então rapidamente depositadas na superfície por meio da ação da gravidade. Um indivíduo pode ser infectado ao inalar o vírus se estiver próximo de outro do qual é hospedeiro do vírus COVID-19 ou ao tocar em uma superfície contaminada e, em seguida, passar as mãos nos olhos, no nariz ou na boca.

O Brasil registrou o primeiro caso do novo coronavírus SARS-CoV-2, causador da doença COVID-19, no dia 26 de fevereiro, em São Paulo. Um homem de 61 anos, cuja identidade não foi revelada, que havia passado férias no continente europeu, mais especificamente no país da Itália de 9 a 21 de fevereiro, na região da Lombardia, um dos epicentros da crise naquele país. Desde então, a infecção se alastrou por todos os estados por meio de um tipo de transmissão chamada de comunitária, que não permite se saber onde, exatamente, uma pessoa contraiu o vírus.

Entre as nações das quais apresentaram pacientes infectados com o vírus, o Brasil certamente é uma das nações mais impactada. Atualmente o país brasileiro conta **158.456 mortes** e **5.468.270 casos de coronavírus**. Dados do Ministério da Saúde até **28 de outubro**. A figura 1, extraída do site "CORONAVÍRUS BRASIL"detalham esses números divulgados pelo Ministério da Saúde até o dia 28 de Outubro de 2020.

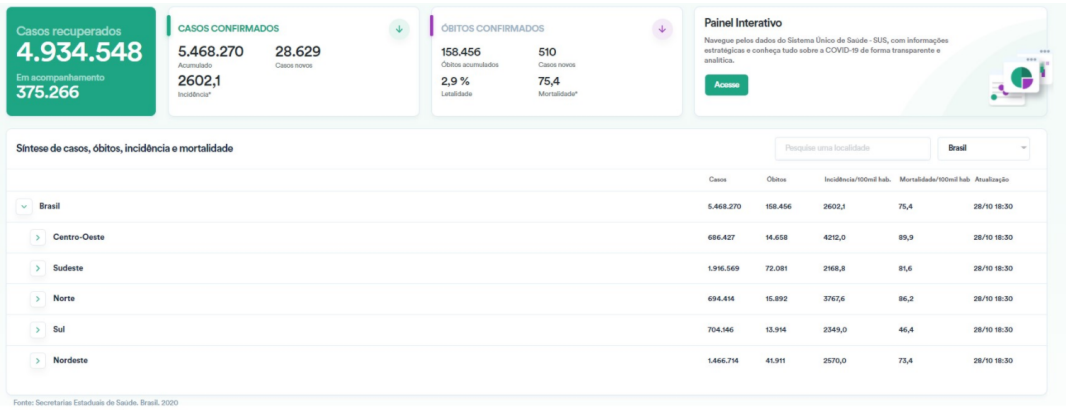


Figura 1: Casos de cononavírus no Brasil - Ministério da Saúde

[1]Ministério da Saúde Brasileiro: https://coronavirus.saude.gov.br/

1.1 Consequências COVID-19

Em 2019 havia grande apreensão sobre o ano de 2020, tal ano poderia conter o desdobramento de questões como: Guerra comercial entre EUA e China, eleições presidenciais dos EUA e o impacto do Brexit na economia mundial. Por conta disso o Fundo Monetário Internacional (FMI) previu crescimento global de apenas 3.4%, porém no fim de 2019 houve a primeira disseminação da COVID-19 em Wuhan, na China, do qual se espalhou em praticamente todo o mundo, tornando assim o ano de 2020 um cenário totalmente diferente diante às previsões feitas.

Com a chegada do novo Coronavírus, diversas nações decretaram isolamento social, fecharam fronteiras aéreas e terrestres e decretaram quarentena visando conter o vírus que ainda não se sabia muito bem como tratar. A primeira grande potencia mundial, a China, foi o primeiro país a ser afetada por longos dias de lockdown e fechamento de fábricas. Sendo assim, houve diminuição no consumo e na produção, interrompendo a cadeia de abastecimento global, afetando empresas e governo em todo o mundo. Quando o vírus já estava instalado mundialmente, a economia global mergulhou em recessão em 2020, principalmente no primeiro semestre. Os consumidores também mudaram seus padrões de consumo, resultando na escassez de muitos produtos nos supermercados em todo o mundo. Os mercados financeiros globais registraram quedas acentuadas em níveis semelhantes, ou acima, da crise financeira de 2008/2009. Setores como turismo e comércio agora iniciavam previsões sobre déficits econômicos.

De acordo com o artigo Economic effects of coronavirus outbreak (COVID-19) on the world economy publicado por Nuno Fernandes comparações da crise do novo Coronavírus com o surto de SARS registrado em 2002/2003 não é válido, uma vez que em 2003 a china representava apenas 3% da população mundial, em 2020 esse número já é equivalente à 16%, sendo qualquer choque ocorrido na China impactante em diversas economias globais. O autor pontua também que a China é o principal comprador de bens e serviços globais, e que também, a economia global está mais integrada nos dias atuais se comparado há 15 anos.

De acordo com a Associação Internacional de Transportes Aéreos (IATA) a previsão era um rombo de até 113 bilhões de dólares[2] para tal setor. Houve grande cancelamento de passagens de avião, reservas de hotéis e eventos, do qual gerou grandes perdas econômicas e alta taxa de desemprego para diversos países em diferentes setores da economia em decorrência da quarentena e o distanciamento social. Dessa maneira o Fundo Monetário Internacional (FMI) decidiu revisar o crescimento global e para cada país, sendo de maneira majoritária, grande queda nas nações analisadas.

Para tentar conter a pandemia de maneira mais rápida possível, foram iniciados estudos sobre uma possível vacina de imunização. Diversos países entraram nessa corrida pela vacina porém as previsões para vacinação são apenas para o primeiro trimestre de 2021. Dessa maneira, diversos países tiveram que transferir em massa verbas planejadas em outros setores para o âmbito da saúde, tentando conter o número de infecções e mortes em determinada população do território.

Com a pandemia gerada pelo novo Coronavírus, foram expostas o quão o setor da saúde estava carente de tecnologia, principalmente centros de atendimentos disponibilizados pelo governo brasileiro. Muitos hospitais não possuíam algum sistema de integração, outros usavam cadernos para anotar dados importantes. Dessa maneira, o Brasil se tornou um dos países mais afetados, e, consequentemente a América Latina tornou-se epicentro do vírus em meados do segundo semestre de 2020. Até o dia 28 de Outubro, o Brasil está entre os três países com mais infecções e mortes, sendo o número acumulado de mortes igual à 158 mil mortos e 5 milhões de brasileiros infectados, de acordo com os dados do Ministério da Saúde Brasileiro.

[2] Associação Internacional de Transportes Aéreos: https://valor.globo.com/empresas/noticia/2020/03/05/aviacao-comercial-deve-perder-ate-us-113-bi-de-receita-devido-ao-coronavirus-diz-iata.ghtml

1.2 Importância da análise de dados contra à COVID-19

Com a pandemia gerada pelo novo coronavírus, diversas áreas necessitaram trabalhar em conjunto, de biologia, medicina, comportamento humano, à economia. Dessa maneira, grandes massas de dados são geradas diariamente para verificar o comportamento causado pelo vírus nas áreas citadas.

Diversas projeções foram feitas para os mais variados âmbitos, projeções essas que assustavam até os mais experientes em suas áreas. Como estava sendo previsto, malefícios biológicos e econômicos de tal magnitude não eram vivenciados faziam-se anos. Tais projeções visavam ajudar na condução das consequências e dos desdobramentos da pandemia, porém algumas foram bastante criticadas, como por exemplo o isolamento social. Por meio da mídia e grandes meio de comunicação, foi possível destacar a validade dos meios científicos e o poder de tais dados/projeções.

Como citado pela revista *Health Analytics*[3]: "Embora o coronavírus tenha acentuado a promessa de ferramentas avançadas de análise, a pandemia também revelou a relativa imaturidade da tecnologia. Problemas relacionados ao acesso, compartilhamento e qualidade de dados ainda afetam a precisão dos algoritmos, bem como a capacidade de desenvolver algoritmos em primeiro lugar. Para um setor que passou anos com o EHR (electronic health records) e os dados digitais, é preciso se perguntar: por que tantos cuidados de saúde ainda são reativos em vez de proativos?".

No inicio do surto do novo coronavírus, grandes organizações de saúde, sendo essas hospitais ou mesmos órgãos de saúde, do mundo se uniram para que houvesse uma grande coleta de dados. Dessa maneira, chefes de Estado (Orientados por grandes equipes de virologistas e infeciologistas) puderam criar estratégias para o combate e a contenção das consequências do vírus. Consequentemente, quanto mais dados fossem coletados mais precisa poderiam tornar-se projeções e estratégias para gerenciar a pandemia.

De acordo com a revista *Orange Business*[4] a utilização de artifícios de ciência de dados e inteligência artificial torna-se necessária em todos os estágios da pandemia, detecção, propagação do vírus, gerenciamento da pandemia e recuperação. A aplicação de ciência de dados na pandemia se iniciou no final de 2019, quando uma Startup de Torotono chamado *BlueDot*[5], detectou anomalias respiratórias em Wuhan por meio de cruzamento de dados de passagens aéreas, avisos de governos locais e relatórios de saúde. Tal detecção foi efetuada antes do pronunciamento oficial da Organização Mundial da Saúde (OMS) emitir alertas sobre um possível surto para esse vírus.

No gerenciamento contra a COVID-19, diversos sensores de calor, sejam esses por exemplo câmeras, que possuem inteligência artificial para devolver um relatório sobre a temperatura de uma determinada pessoa, quantos indivíduos foram medidos e quantos desses possuíam temperatura elevada (Um dos sintomas da COVID-19). Cruzamento de dados de celulares podem formar relatórios sobre o nível de isolamento social de indivíduos de determinado bairro ou cidade[6] ajudando a mover estratégias de contenção do vírus, como foi aplicado na Coreia do Sul, onde o governo entrava em contato com indivíduos que estivera próximo à outro que testou positivo em algum momento.

[3]Health Analytics: healthitanalytics.com/features/could-COVID-19-help-refine-ai-data-analytics-in-healthcare
[4]Orange Business: orange-business.com/en/blogs/ai-and-data-science-tool-battle-COVID-19
[5]BlueDot: cnbc.com/2020/03/03/bluedot-used-artificial-intelligence-to-predict-coronavirus-spread.html
[6]Navita: navita.com.br/blog/tecnologia-e-gestao-de-dados-celulares-ajudam-no-combate-ao-COVID-19/

2 Plano de análise

A COVID-19 desde que chegou ao território brasileiro mostrou-se um grande problema para a agravar ainda mais a situação precária da qual os hospitais se encontravam. Dessa maneira, foi-se necessário coletar os dados de cada enfermo para que pudesse serem feitas estatísticas sobre óbitos, casos, índices de incidência e mortalidade e assim tomar decisões sobre tais.

Consequentemente, com tantos dados sendo coletados, análises, previsões e projeções começaram a serem implantadas com o objetivo de tentar, com alguma estratégia, frear o avanço da doença no território brasileiro. O governo brasileiro, por meio do Ministério da Saúde, disponibiliza análises dos dados diariamente, porém análises são efetuadas também pelo consórcio de veículos de imprensa.

Para a confecção deste trabalho, foram coletados todos os dados iniciais sobre a doença no território brasileiro até o dia 28 de Outubro de 2020. Dessa maneira, o objetivo desse trabalho é chegar em números iguais ou próximos dos que foram notificados no dia 28 de Outubro por meio de ferramentas de software e estatística R, RStudio e a confecção do relatório por meio do software LaTeX.

A análise nesse trabalho se dividirá em três etapas (Sendo essas aplicadas tanto para os dados sobre o Brasil de forma geral, quanto para o Distrito Federal):

- **Etapa 1:**
  Análise da evolução da doença por mês, abordando número de casos novos, casos acumulados, óbitos novos e óbitos acumulados (até o dia 28 de Outubro). No caso da análise sobre o Brasil, será abordado de forma rasa estatísticas descritivas e analises sobre as regiões.

- **Etapa 2:**
  Análise da evolução da doença por mês, abordando número de casos novos, casos acumulados, óbitos novos e óbitos acumulados (até o dia 28 de Outubro), porém usando dados referentes ao Distrito Federal.

- **Etapa 3:**
  Análises sobre coeficientes de incidência, mortalidade e taxa de mortalidade obtidos até o dia 28 de Outubro sobre o Brasil e o Distrito Federal.

Ao fim, poderá concluir-se que, possivelmente, até o dia 28 de Outubro de 2020, tanto o Brasil quanto o Distrito Federal estão em desaceleração em relação ao aumento de números de casos novos/óbitos e como se desenvolveu os coeficientes, seja esses de incidência, mortalidade e taxa de mortalidade.

A meta desse plano de análise é tentar tornar possível a um leitor leigo o entendimento dos dados disponibilizados e demonstrar o quão impactante é cada número atribuído à determinada variável visando informar de maneira clara, lúcida e objetiva cada variável.

3 Descrição da base de dados

A base de dados foi coletada por meio do site do Ministério da Saúde do Brasil[7] porém foi acessada no dia 28 de Outubro de 2020, dos quais geraram as análises presentes neste trabalho. Da mesma maneira, o professor Sérgio Côrtes disponibilizou arquivos SAS dos quais os dados são referentes aos arquivos encontrados no site do Ministério da Saúde, porém tratados em SAS para tornar a confecção deste trabalho mais dinâmica e acessível.

3.1 Dicionário de Dados

Sendo assim, para tornar o entendimentos mais claro sobre as variáveis presentes e que serão avaliadas nesse trabalho, faz-se necessário a documentação do dicionário de tais dados e seus respectivos arquivos:

- **COVID19_CADASTRO-2020-10-28:** Arquivo original presente no Ministério da Saúde.
  - **Regiao:** Regiões presentes no território brasileiro.
  - **Estado:** Estados presentes no território brasileiro.
  - **Coduf:** Código referente ao determinado estado presente no território brasileiro.
  - **PopulacaoTCU2019:** População de determinado estado referente à determado estado brasileiro.
- **COVID19_SERIE-2019-10-18:** Arquivo original presente no Ministério da Saúde.
  - **Regiao:** Regiões presentes no território brasileiro.
  - **Estado:** Estados presentes no território brasileiro.
  - **Municipio:** Municípios presentes no território brasileiro.
  - **Coduf:** Código referente ao determinado estado presente no território brasileiro.
  - **Codmun:** Código referente ao determinado município presente no território brasileiro.
  - **CodRegiaoSaude:** Código referente à região de atuação da área de saúde.
  - **NomeRegiaoSaude:** Nome referente à região de atuação da área de saúde
  - **Data:** Data da coleta do dado (ano(aaaa) - mes(mm) - dia(dd))
  - **SemanaEpi:** Referente à semana do ano desde o início da coleta de dados epidemiológicos no território brasileiro. (Por exemplo, semana 9, nona semana do ano de 2020)
  - **CasosAcumulados:** Número total de casos confirmados por COVID-19 que foram registrados pelas Secretarias Municipais e Estaduais de Saúde no período considerado.
  - **CasosNovos:** Número de casos novos confirmados por COVID-19 que foram registrados pelas Secretarias Municipais e Estaduais de Saúde em relação ao dia anterior.
  - **ObitosAcumulados:** Número total de óbitos confirmados por COVID-19 que foram registrados pelas Secretarias Municipais e Estaduais de Saúde no período considerado.
  - **ObitosNovos:** Número de óbitos novos confirmados por COVID-19 que foram registrados pelas Secretarias Municipais e Estaduais de Saúde em relação ao dia anterior.
  - **RecuperadosNovos:** Número de pacientes recuperados de COVID-19 que foram registrados pelas Secretarias Municipais e Estaduais de Saúde em relação ao dia anterior.
  - **EmAcompanhamento:** Número de pacientes em acompanhamento em decorrência da COVID-19 que foram registrados pelas Secretarias Municipais e Estaduais de Saúde em relação ao dia anterior.
  - **Interior/metropolitana:** Tipo de moradia do paciente infectado.

[7] Ministério da Saúde - Covid, disponível em: covid.saude.gov.br

Os arquivos em formato SAS presentes na pasta disponibilizada pelo professor Sérgio Côrtes, foram tratados de maneira com que inibisse dados faltantes ou que possuíam ruídos. Dessa maneira, os seguintes arquivos são:

- **COVID19:** Dados sobre a pandemia do novo coronavírus em território brasileiro. Modificado em SAS.
  - **Regiao:** Regiões presentes no território brasileiro.
  - **Estado:** Estados presentes no território brasileiro.
  - **Municipio:** Municípios presentes no território brasileiro.
  - **Coduf:** Código referente ao determinado estado presente no território brasileiro.
  - **Codmun:** Código referente ao determinado município presente no território brasileiro.
  - **CodRegiaoSaude:** Código referente à região de atuação da área de saúde.
  - **NomeRegiaoSaude:** Nome referente à região de atuação da área de saúde
  - **Data:** Data da coleta do dado (ano(aaaa) - mes(mm) - dia(dd))
  - **SemanaEpi:** Referente à semana do ano desde o início da coleta de dados epidemiológicos no território brasileiro. (Por exemplo, semana 9, nona semana do ano de 2020)
  - **CasosAcumulados:** Número total de casos confirmados por COVID-19 que foram registrados pelas Secretarias Municipais e Estaduais de Saúde no período considerado.
  - **CasosNovos:** Número de casos novos confirmados por COVID-19 que foram registrados pelas Secretarias Municipais e Estaduais de Saúde em relação ao dia anterior.
  - **ObitosAcumulados:** Número total de óbitos confirmados por COVID-19 que foram registrados pelas Secretarias Municipais e Estaduais de Saúde no período considerado.
  - **ObitosNovos:** Número de óbitos novos confirmados por COVID-19 que foram registrados pelas Secretarias Municipais e Estaduais de Saúde em relação ao dia anterior.
  - **RecuperadosNovos:** Número de pacientes recuperados de COVID-19 que foram registrados pelas Secretarias Municipais e Estaduais de Saúde em relação ao dia anterior.
  - **EmAcompanhamento:** Número de pacientes em acompanhamento em decorrência da COVID-19 que foram registrados pelas Secretarias Municipais e Estaduais de Saúde em relação ao dia anterior.
  - **Interior/metropolitana:** Tipo de moradia do paciente infectado.
- **COVID19_TAXAS_BRASIL:** Dados sobre taxas, que, de forma geral, abordam sobre o país Brasil. Modificado em SAS.
  - **Populacao_Brasil:** População brasileira brasileira referente ao ano de 2019.
  - **Regiao:** Regiões presentes no território brasileiro.
  - **Estado:** Estados presentes no território brasileiro.
  - **Casos_Confirmados_Brasil:** Casos confirmados até o dia 28 de Outubro de 2020 por COVID-19 que foram registrados pelas Secretarias Municipais e Estaduais de Saúde.
  - **Obitos_Confirmados_Brasil:** Casos confirmados até o dia 28 de Outubro de 2020 por COVID-19 que foram registrados pelas Secretarias Municipais e Estaduais de Saúde.
  - **Brasil_Incidencia:** Coeficiente de casos por COVID-19 no território brasileiro no dia 28 de Outubro de 2020.
  - **Brasil_Mortalidade:** Coeficiente de mortalidade por COVID-19 no território brasileiro no dia 28 de Outubro de 2020.
  - **Brasil_Letalidade:** Taxa de letalidade por COVID-19 no território brasileiro no dia 28 de Outubro de 2020.

- **COVID_TAXAS_REGIAO:** Dados sobre taxas, que, de forma geral, abordam sobre as regiões presentes no território brasileiro. Modificado em SAS.
  - **Região:** Regioes presentes no território brasileiro.
  - **Casos_Confirmados_Regiao:** Casos confirmados até o dia 28 de Outubro de 2020 por COVID-19 que foram registrados pelas Secretarias Municipais e Estaduais de Saúde por região.
  - **Obitos_Confirmados_Regiao:** Casos confirmados até o dia 28 de Outubro de 2020 por COVID-19 que foram registrados pelas Secretarias Municipais e Estaduais de Saúde por região.
  - **Regiao_Incidencia:** Coeficiente de casos por COVID-19 em cada região brasileira no dia 28 de Outubro de 2020.
  - **Regiao_Mortalidade:** Coeficiente de mortalidade por COVID-19 em cada região brasileira no dia 28 de Outubro de 2020.
  - **Regiao_Letalidade:** Taxa de letalidade por COVID-19 em cada região brasileira no dia 28 de Outubro de 2020.
- **COVID_TAXAS_ESTADO:** Dados sobre taxas, que, de forma geral, abordam sobre os estados presentes no território brasileiro. Modificado em SAS.
  - **Regiao:** Regiões presentes no território brasileiro.
  - **Estado:** Estados presentes no território brasileiro.
  - **Coduf:** Código referente ao determinado estado presente no território brasileiro.
  - **PopulacaoTCU2019:** População de determinado estado referente à determinado estado brasileiro.
  - **Casos_Confirmados_Estado:** Casos confirmados até o dia 28 de Outubro de 2020 por COVID-19 que foram registrados pelas Secretarias Municipais e Estaduais de Saúde por estado.
  - **Obitos_Confirmados_Estado:** Casos confirmados até o dia 28 de Outubro de 2020 por COVID-19 que foram registrados pelas Secretarias Municipais e Estaduais de Saúde por estado.
  - **Estado_Incidencia:** Coeficiente de casos por COVID-19 em cada estado brasileira no dia 28 de Outubro de 2020.
  - **Estado_Mortalidade:** Coeficiente de mortalidade por COVID-19 em cada estado brasileira no dia 28 de Outubro de 2020.
  - **Estado_Letalidade:** Taxa de letalidade por COVID-19 em cada estado brasileira no dia 28 de Outubro de 2020.

Este trabalho foi elaborado por meio dos dados presentes na base de dados **COVID19** do qual foi modificado em SAS pelo professor Sérgio Côrtes. O motivo de tal escolha se deu pois, ao filtrar a variável **estado** para apenas o Brasil geral, os dados tornariam-se menores e mais rápida em relação à compilação destes em análises, obtendo **o mesmo resultado** do que caso escolhido a base de dados toda sem filtro algum. Como pode ser visto na seguinte imagem da qual representa uma amostra dos dados referentes ao estado Brasil:

	regiao	estado	municipio	coduf	codmun	codRegiaoSaude	nomeRegiaoSaude
1	Brasil			76	NA	NA	
2	Brasil			76	NA	NA	
3	Brasil			76	NA	NA	
4	Brasil			76	NA	NA	
5	Brasil			76	NA	NA	

Figura 2: Demonstração dados filtro Brasil

3.2 Conceitos básicos

Neste tópico será abordado conceitos por trás dos cálculos e sintaxes usados pelo Ministério da Saúde, onde encontram-se mais informações aprofundadas acessando este órgão de Saúde. Vale ressaltar que os dados usados referente à população[8] usados nos cálculos são estimativas de 2019 utilizadas pelo TCU para determinação das cotas do FPM (sem sexo e faixa etária).

Coeficiente de Incidência de COVID-19: Número de casos confirmados de COVID-19 por 100.000 habitantes, na população residente em determinado espaço geográfico, no período considerado. Estima o risco de ocorrência de casos de COVID-19 numa determinada população num período considerado. Número alcançado por meio do seguinte cálculo:

$$\frac{Número de casos confirmados de COVID-19 em residentes X 100.000}{População total residente no período determinado}$$

Coeficiente de Mortalidade por COVID-19: Número de óbitos por doenças COVID-19, por 100 mil habitantes, na população residente em determinado espaço geográfico, no ano considerado. Estima o risco de morte pela COVID-19 consideradas e dimensiona a sua magnitude como problema de saúde pública. Número alcançado por meio do seguinte cálculo:

$$\frac{Número de óbitos confirmados de COVID-19 em residentes X 100.000}{População total residente no período determinado}$$

Taxa de Letalidade por COVID-19: Número de óbitos confirmados de COVID-19 em relação ao total de casos confirmados, na população residente em determinado espaço geográfico, no período considerado. Esta taxa dá a ideia de gravidade da doença, pois indica o percentual de pessoas que morreram dentre os casos confirmados da doença. Número alcançado por meio do seguinte cálculo:

$$\frac{Número de óbitos confirmados de COVID-19 em determinada área e período X 100}{Número de casos confirmados de COVID-19 em determinada área e período}$$

Casos recuperados: Segundo a Organização Mundial da Saúde, para os casos de COVID-19 confirmados por critério laboratorial, considera-se como recuperados aqueles que tiveram dois resultados negativos para SARS-CoV-2 com pelo menos 1 dia de intervalo. Para os casos leves, a OMS estima que tempo entre o início da infecção e a recuperação dure até 14 dias.

[8] Populacao: datasus.saude.gov.br/populacao-residente

4 Análise exploratória e visualização dos dados do COVID-19 no Brasil

A análise exploratória tem como objetivo final explorar os dados presentes na base de dados, como por exemplo: explorar os tipos de dados presentes, a distribuição dos dados, quantificar os dados faltantes e diversas outras diretrizes que podem ser seguidas. Dessa maneira, será possível entender os dados e as relações que existem entre as diversas variáveis, consequentemente tornando possível a aplicação de técnicas estatísticas e/ou de modelagem de dados.

A biblioteca *Skim*[9] foi utilizada para efetuar, de maneira rápida, a exploração dos dados. Com tal biblioteca (por meio da ferramenta R) poderá concluir-se diversas características que a base de dados possui, ou seja, a biblioteca em questão apresenta um tipo de sumário sobre a base de dados que será avaliada. Vale ressaltar que a biblioteca Skim faz sumários separados de acordo com o tipo de dado presente na base de dados (por exemplo, sumário para variáveis numéricas e outro sumário para variáveis categóricas). Dessa maneira, é demonstrado algumas características encontradas em cada tipo de dado.

A análise exploratória da base de dados presente nesse trabalho se dividirá em duas partes, como explicado na seção 2. Primeiramente a análise feita para o país Brasil considerando os dados até o dia 28 de Outubro de 2020. Posteriormente, será analisado os dados presentes no Distrito Federal, dados esses referentes até o período do dia 28 de Outubro de 2020. É válido relembrar que, os dados aqui usados são oriundos de um filtro para o estado "Brasil", sendo assim tornou mais rápida as estratégias em cima dos dados e foi-se obtido o mesmo resultado do que caso não houvesse filtro, como evidenciado na figura 2.

4.1 Propriedades dos dados referente ao país Brasil

Por meio da biblioteca *Skim*, da qual foi explicada anteriormente, será avaliada as propriedades presentes nos dados, essas das quais não foram especificadas na 3.

Ao aplicarmos a função skim na base de dados é retornado uma tabela com diversas características. Primeiramente pode ser visto a seguinte imagem:

```
-- Data Summary ------------------------
                           Values
Name                       covid_Br
Number of rows             247
Number of columns          17
_________________________
Column type frequency:
  character                4
  Date                     1
  numeric                  12
_________________________
Group variables            None
```

Figura 3: Propriedades gerais dados Brasil

Na figura 3 pode-se perceber que o arquivo filtrado para **apenas dados referente à região Brasil** possui 247 linhas e 17 colunas. Dessas 17 colunas (ou pode-se chamar de variáveis), 12 variáveis são do tipo numéricas, 4 variáveis são do tipo carácter e apenas 1 variável é do tipo data. Na base de dados não foram encontrados grupos de variáveis.

[9]Skim: rdocumentation.org/packages/skimr/versions/1.0.3/topics/skim

4.2 Variáveis Numéricas

Como avaliado de maneira rasa na seção 3, variáveis do tipo numéricos são mais frequentes nessa base de dados. Dessa maneira, ainda no resultado retornado após a função skim, é possível visualizar as características presentes em cada variável do tipo numérico, como é visto na imagem posterior:

```
-- Variable type: numeric -------------------------------------------------------------------------------------------------------------------
# A tibble: 12 x 11
   skim_variable          n_missing complete_rate        mean        sd        p0       p25       p50       p75      p100 hist
 * <chr>                      <int>         <dbl>       <dbl>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl> <chr>
 1 coduf                          0             1          76         0        76        76        76        76        76 "▁▁▁▇▁▁▁▁"
 2 codmun                       247             0         NaN        NA        NA        NA        NA        NA        NA " "
 3 codRegiaoSaude               247             0         NaN        NA        NA        NA        NA        NA        NA " "
 4 semanaEpi                      0             1        26.4      10.2         9        18        26        35        44 "▇▇▇▇▇▇▇▇"
 5 populacaoTCU2019               0             1 210147125         0 210147125 210147125 210147125 210147125 210147125 "▁▁▇▁▁▁▁▁"
 6 casosAcumulado                 0             1     1940467.  1920247.         0     64194.   1313667   3783097   5468270 "▇▂▁▁▁▁▁▁"
 7 casosNovos                     0             1       22139.    17702.         0      4682     20647     34792     69074 "▇▅▅▃▂▁▁▁"
 8 obitosAcumulado                0             1       64393.    56302.         0      4374     57070    119076.   158456 "▇▂▂▂▂▂▂▂"
 9 obitosNovos                    0             1         642.      439.         0       247       641      1026.     1595 "▇▇▇▇▅▃▂▁"
10 Recuperadosnovos              54         0.781     1928989.  1666130.     22130    277149   1592281   3497337   4934548 "▇▂▂▂▂▂▂▂"
11 emAcompanhamentoNovos         54         0.781      470385.   231495.     14062    359767    499513    657297    817642 "▂▇▇▇▇▇▇▇"
12 interior/metropolitana       247             0         NaN        NA        NA        NA        NA        NA        NA " "
```

Figura 4: Propriedades variáveis numéricas dos dados referentes ao Brasil

Na figura 4 pode-se perceber que nas linhas horizontais estão as variáveis, já nas colunas verticais estão diversas estatísticas sobre as variáveis numéricas. Variáveis como **coduf, codmun, codRegiaoSaude, semanaEpi e interior/metropolitana** são variáveis representadas por números, porém nesse contexto não fazem sentido serem numéricas, pois caso haja aplicação de estatísticas, apenas a moda, isto é, a mais frequente, serviria para análises. As variáveis **casosAcumulado**, **casosNovos**, **obitosAcumulado** e **obitosNovos** não possuem dados faltantes. Entretanto, dados numéricos como **recuperado** e **emAcompanhamento** possuem 54 dados faltantes, porém estão 0.781% preenchidas (de uma escala de 0 a 1), o que pode ser considerado algo positivo ainda pois possuem grande gama de dados. Dessa maneira, as variáveis que serão analisadas nesse tópico serão: **casosNovos, casosAcumulado, obitosNovos, obitosAcumulado, emAcompanhamentoNovos e recuperadosNovos**.

4.2.1 Casos Novos

Com o inicio da disseminação do novo coronavírus em território brasileiro, foi-se necessária a coleta dos dados referentes aos novos enfermos, isto é, pessoas que foram infectadas. No Brasil tais dados são colhidos um dia anterior e expostos um dia após a coleta, sendo assim, temos as seguintes estatísticas gerais em relação ao novos enfermiços ocasionados pela COVID-19:

```
> summary(covid_Br$casosNovos)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0    4682   20647   22139   34792   69074
```

Figura 5: Estatísticas casos novos

De acordo com a figura 4 e 5, a variável **casosNovos** possuiu média geral de 22.139 pessoas com diagnóstico positivo para COVID-19 por dia (levando em consideração todos os meses de coleta de dados), com desvio padrão alto, do qual foi igual à 17.702 pessoas, para cima ou para baixo, com diagnóstico positivo para COVID-19 por dia, isso mostra a grande variação nos dados. Na distribuição da coleta dos dados, até 50% dessas datas diárias obtiveram número de casos novos menores ou igual à 20.647 casos por dia. Como evidenciado na figura 5, o máximo de casos novos por dia foi igual à 69.074 diagnósticos positivos, tal dia é referente à data 29 de Julho de 2020, como é evidenciado na seguinte imagem:

```
> covid_Br$data[covid_Br$casosNovos == 69074]
[1] "2020-07-29"
```

Figura 6: Data referente ao maior número de casos novos por dia no Brasil

Avaliar a distribuição dos dados pode evidenciar características que ainda são ocultas sobre como estão distribuídos tais dados. Sendo assim, criou-se o seguinte gráfico de densidade dos dados para a variável casosNovos:

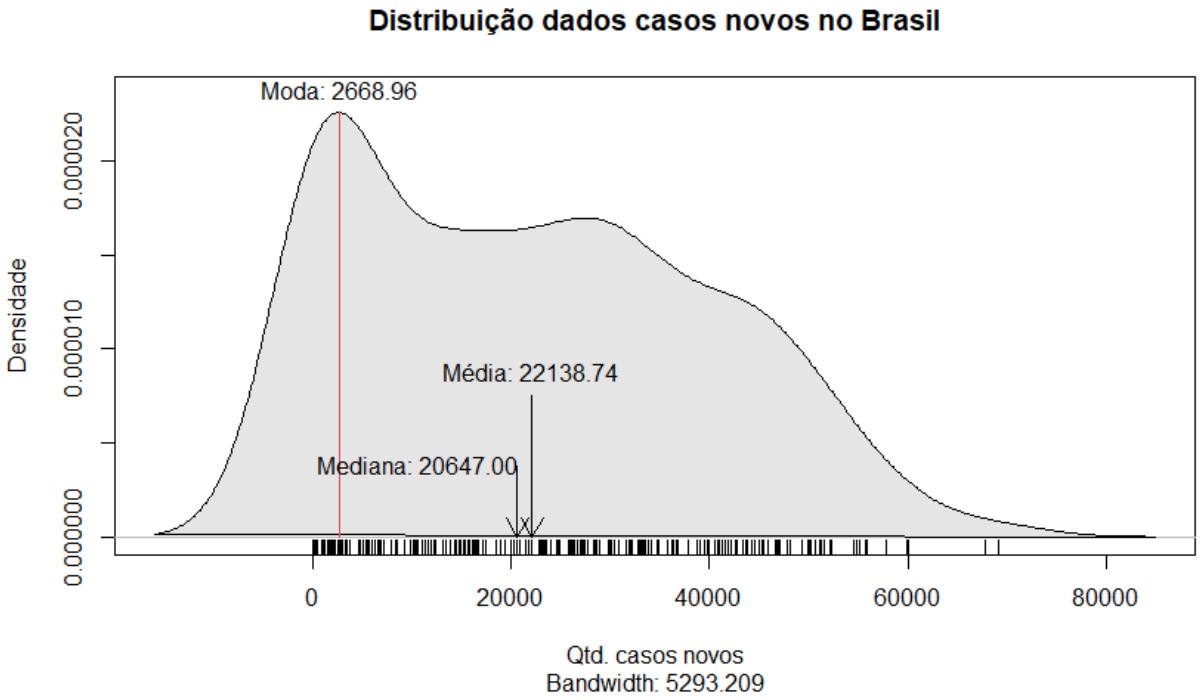


Figura 7: Distribuição dos dados sobre casos novos por quantidade de enfermos novos

Na figura 7 entende-se que a quantidade de casos novos é mais densa quando há menos de 20.000 novos diagnósticos positivos por dia. Aplicando as funções skewness() e kurtosis() presente no software R, será possível avaliar o índice de assimetria e de curtose de tal distribuição. Sendo assim, o índice de assimetria de casos novos no Brasil é de 0.3657, já a curtose foi igual à -0.9358. De acordo com a assimetria encontrada, tem-se uma assimetria moderada dos dados e possui curtose de natureza platicúrtica. Avaliando a moda, média e mediana, é possível verificar uma assimetria à direita dos dados.

Portanto, os dados referentes à casos novos no Brasil possuem uma distribuição assimétrica moderada para a direita com curtose de natureza platicúrtica. Tal cenário é ideal quando se trata sobre a densidade dos dados por quantidade de enfermos que testara positivo ao novo coronavírus, pois as menores quantidades de casos novos foram mais frequentes/ estão densas.

Entretanto o governo deve levar em consideração a assimetria moderada dos dados, pois demonstra que, como é visto na figura 7 a média e mediana dos casos ultrapassaram 20.000 infectados, medidas essas muito distantes da moda encontrada, média essa que é afetada por valores extremos. Então entende-se que, para frear o avanço de indivíduos infectados seria necessário que a distribuição dos dados possuísse uma curva de assimetria forte para a direita.

Para verificar, concluir e finalizar tal análise sobre os dados referentes à casos novos no Brasil, pode-se fazer o uso do gráfico de histograma, que mostrará a frequência das quantidades de casos novos.

Então, com o uso de um histograma na variável casos novos, será possível analisar e entender como se arranjou a distribuição de frequência dos dados de tal variável após a coleta de dados que teve fim em 28 de Outubro de 2020:

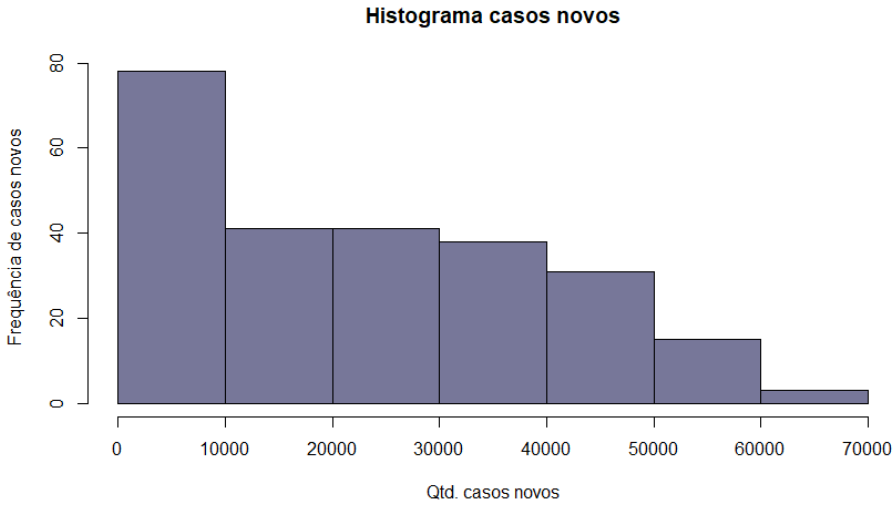


Figura 8: Histograma da frequência de casos novos por quantidade de enfermos novos

De acordo com a figura 8 é possível concluir que a distribuição está assimétrica à direita, que é ideal. Percebe-se que, levando em consideração todos os dias desde o inicio da coleta dos dados, que, a maior frequência de casos novos ocorreu em dias que tiveram números menos ou iguais à 10.000 casos novos, sendo o máximo, como visto na figura 5 pouco frequente. Porém vale chamar a atenção das autoridades de saúde sobre quão constante estão dias que arrecadam até 30.000 casos novos.

Portanto, para a melhor compreensão do leitor, é possível mostrar a distribuição de casos novos no território brasileiro em um gráfico de linhas do quão mostrará a quantidade de casos novos por determinado dia/mês:

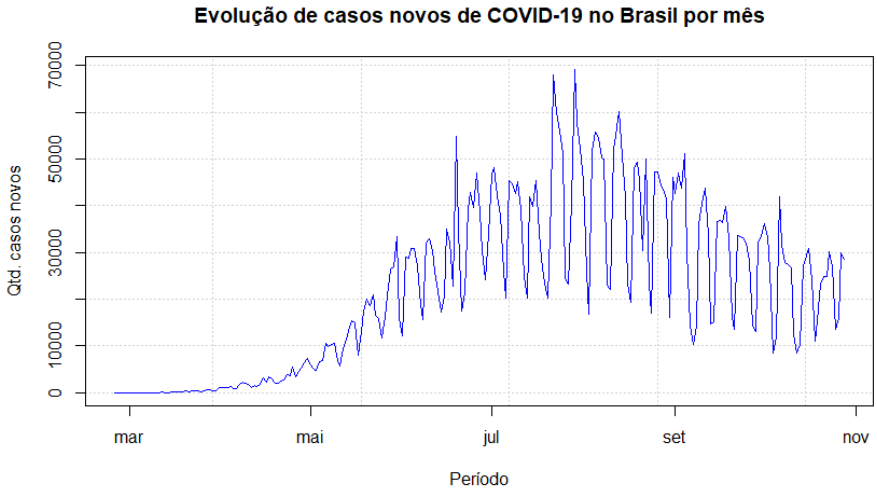


Figura 9: Casos novos de COVID-19 por mês no Brasil

Avaliando o gráfico da figura 9 pode-se perceber que a distribuição dos dados em relação ao mês possui muitas sazonalidades, sazonalidades essas ocorridas porquê os hospitais muitas vezes faziam diversas reavaliações sobre diagnósticos passados, computando casos novos reavaliados de uma maneira muito brusca no outro dia. Verifica-se que o pico da pandemia no território brasileiro foi entre os meses de Julho e Setembro. Após tal data os números começaram a entrar em redução, chegando à mesma quantidade de casos novos obtidas no mês de Abril.

Com o uso do BoxPlot, um tipo de visualização estatística dos dados, torna-se possível verificar a mediana e os quartis da variável casos novos em cada mês referente à coleta dos dados:

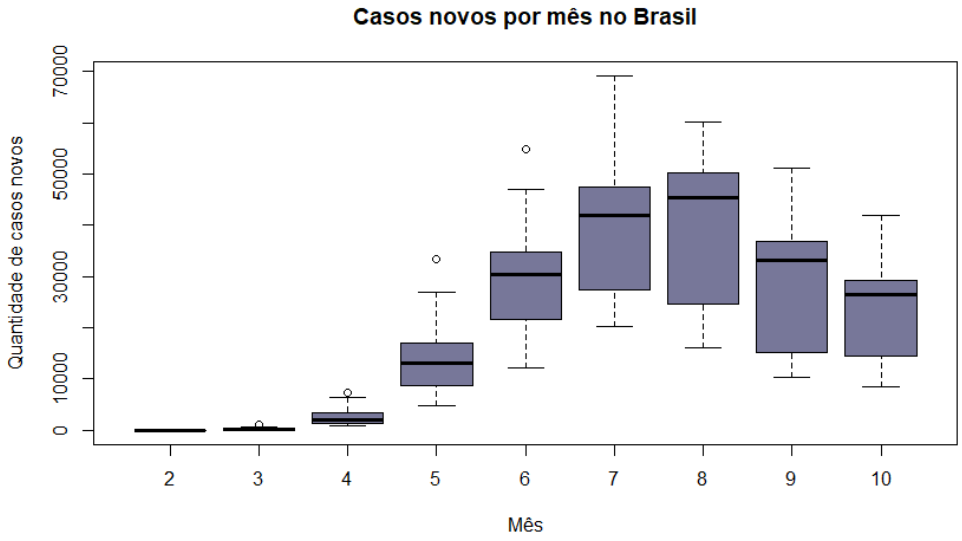


Figura 10: Boxplot - Casos novos por mês em numeral no Brasil

```
 Descriptive statistics by group
group: 2
   vars n mean   sd median trimmed mad min max range skew kurtosis   se
X1    1 5  0.4 0.55      0     0.4   0   0   1     1 0.29    -2.25 0.24
------------------------------------------------------------------------------------------
group: 3
   vars  n  mean    sd median trimmed   mad min  max range skew kurtosis   se
X1    1 31 184.3 247.2     57   143.6 84.51   0 1138  1138 1.92     4.54 44.4
------------------------------------------------------------------------------------------
group: 4
   vars  n mean   sd median trimmed  mad min  max range skew kurtosis    se
X1    1 30 2655 1682   2080    2408 1406 852 7218  6366 1.12     0.33 307.1
------------------------------------------------------------------------------------------
group: 5
   vars  n  mean   sd median trimmed  mad  min   max range skew kurtosis   se
X1    1 31 13833 6961  13140   13069 6328 4751 33274 28523 0.83     0.23 1250
------------------------------------------------------------------------------------------
group: 6
   vars  n  mean   sd median trimmed   mad   min   max range skew kurtosis   se
X1    1 30 29595 9653  30444   29103 10431 12247 54771 42524 0.42    -0.11 1762
------------------------------------------------------------------------------------------
group: 7
   vars  n  mean    sd median trimmed   mad   min   max range skew kurtosis   se
X1    1 31 40659 13711  41857   40111 13773 20229 69074 48845 0.17     -0.8 2462
------------------------------------------------------------------------------------------
group: 8
   vars  n  mean    sd median trimmed  mad   min   max range  skew kurtosis   se
X1    1 31 40187 13783  45392   41000 9208 16158 60091 43933 -0.56    -1.28 2476
------------------------------------------------------------------------------------------
group: 9
   vars  n  mean    sd median trimmed   mad   min   max range  skew kurtosis   se
X1    1 30 30089 11800  33169   30162 10390 10273 51194 40921 -0.28    -1.22 2154
------------------------------------------------------------------------------------------
group: 10
   vars  n  mean   sd median trimmed  mad  min   max range  skew kurtosis   se
X1    1 28 23476 9035  26530   23433 5842 8429 41906 33477 -0.23       -1 1707
```

Figura 11: Estatisticas descritivas sobre casos novos por mês em numeral (group)

Como evidenciado na figura 9 e agora na figura 10 o mês de Agosto, que desta vez fora representado pelo numeral respectivo, foi o pico da pandemia do novo coronavírus no território brasileiro, porém o valor máximo está presente no mês anterior, ou seja, o mês de Julho. Percebe-se também outro confirmação, o número de casos novos está descendo de acordo com os meses (após o pico em Julho e Agosto), porém as medianas, ou seja, 50% dos dados referentes à coletas diárias, estão próximas do terceiro quartil, demonstrando que a mediana dos casos continua bastante alta, algo que deve ser avaliado.

Avaliando a figura 11 o mês que possuiu maior média diária dentre tais meses, foi o mês de Julho com média de 40.659 pessoas com diagnóstico positivo para o vírus por dia, número esse que continuou na faixa dos quarenta mil em Agosto, perpetuando assim o pico da pandemia nesses dois meses. Porém como é evidenciado nas medidas que precedem o mês de Agosto, existe a queda dos casos novos positivos para COVID-19, o que é possível confirmar verificando as medidas encontradas no mês de Outubro. Nesse mês a média diária e a mediana foram iguais à 23.476 e 26.529 pessoas com diagnóstico positivo para o COVID-19 por dia, respectivamente. Os dados coletados são colhidos de forma diária, e, no mês de outubro, houveram dias em que o mínimo e máximo de casos novos foram iguais à 8.429 e 41.906, respectivamente, números esses que não eram tão baixos desde o mês 5, ou seja, o mês de Maio.

Para confirmar a tendência da queda de casos é possível criar uma linha de tendência que tentará seguir o sentido dos dados, dados esses referentes à cada dia da coleta do mesmo em determinado mês, como é visto na seguinte imagem:

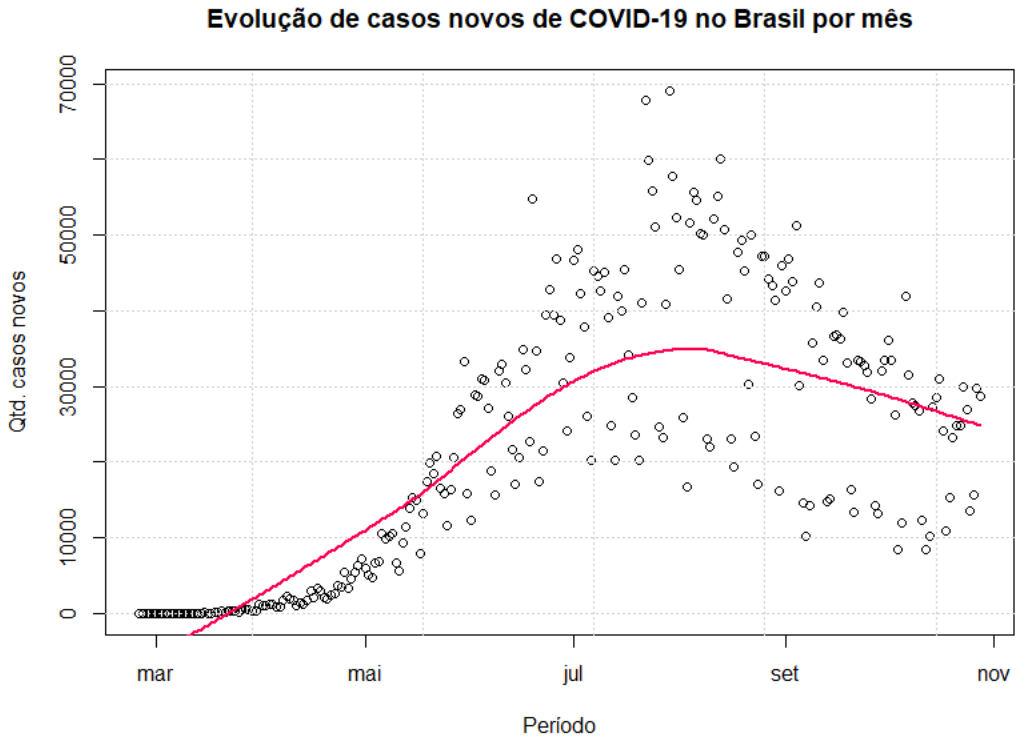


Figura 12: Linha de tendência casos novos por mês

4.2.2 Casos Acumulados

A variável **casosAcumulado** é uma variável que acumula (soma) todos os *casos novos* de COVID-19 desde o inicio da coleta dos dados. Dessa maneira, demonstrar as medidas de estatísticas descritivas não fazem sentido, uma vez que é a soma dos casos novos, portanto, casos acumulados. Sendo assim, o que nos importa está no número presente ao percentil 100, ou seja, o total acumulado. Como é possível verificar na imagem 4 o percentil 100 (p100) é igual à 5.468.270, isto é, desde o inicio da coleta dos dados até o dia 28 de Outubro de 2020, existiram 5.468.270 brasileiros que testaram positivo para o novo coronavírus. Como foi evidenciado no portal da Fio Cruz[10] os aproximadamente 5 milhões de brasileiros afetados são brasileiros residentes em regiões de periferia, onde a assistência médica é ainda uma questão a ser discutida.

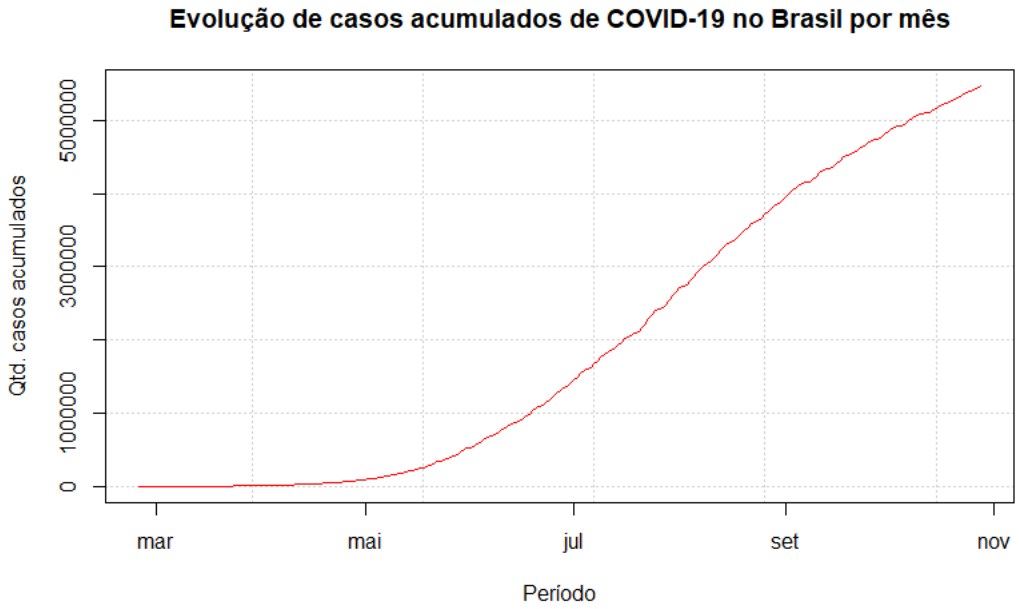


Figura 13: Casos acumulados de COVID-19 por mês no Brasil

Uma variável que trabalha com valores acumulados gera um gráfico como é visto na figura 13, um gráfico de natureza contínua. Gráficos sobre valores acumulados nunca descerá, pois não existem maneira de descontar os valores que já foram acumulados no mesmo. Dessa maneira, se o gráfico se tornar paralelo ao eixo X significa que casos novos não estão sendo acumulados, ou seja, as pessoas não estão testando positivo. Porém, como é possível ver na figura 13, até o dia 28 de Outubro de 2020 os casos acumulados continuaram a crescer, ou seja, as pessoas continuam testando positivo, porém em ritmo desacelerado como é possível perceber comparando os intervalos dos meses de Agosto com Setembro e Setembro com Outubro, o que é evidenciado na figura 10. Porém, sem a presença de autoridades de saúde em locais de grande incidência do vírus gerenciando a pandemia, essa curva pode voltar a crescer em ritmo acelerado como é visto nos meses entre Maio e Julho.

[10]Fio Cruz: portal.fiocruz.br/noticia/desigualdade-social-e-economica-em-tempos-de-covid-19

4.2.3 Óbitos Novos

O Estado brasileiro rapidamente tornou-se um dos que mais possuíam mortes diárias em decorrência da COVID-19. Antes pensasse que o vírus fosse fatal apenas para pessoas de idade, porém, descobriu-se que pessoas com histórico enfermiço também podiam correr este risco fatal. Quando houve a descoberta do vírus não se sabia a forma mais eficaz de tratamento, em decorrência disso muitos óbitos foram acontecendo durante o passar dos dias.

```
summary(covid_Br$obitosNovos)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0     247     641     642    1026    1595
```

Figura 14: Estatísticas óbitos novos

Dessa maneira, pode-se identificar na figura 3 e na figura 18 que, desde o inicio da pandemia e da coleta de dados até o dia 28 de Outubro de 2020, a média de óbitos novos diários foi de 642 óbitos. 50% de todos esses dias tiveram número novos de óbitos menores ou iguais à 641 novos óbitos. Na figura 18 torna-se possível verificar também que o máximo de óbitos novos em um único dia foi igual à 1.595 vítimas. O número máximo de vítimas foi identificado no dia 29 de Julho de 2020, justamente no mesmo dia em que houve o maior número notificado de casos novos por COVID-19.

```
> covid_Br$data[covid_Br$obitosNovos == 1595]
[1] "2020-07-29"
```

Figura 15: Data referente ao maior número de óbitos novos

A fim de avaliar a distribuição de dados referente à óbitos novos, tem-se o seguinte gráfico de densidade para a variável referida:

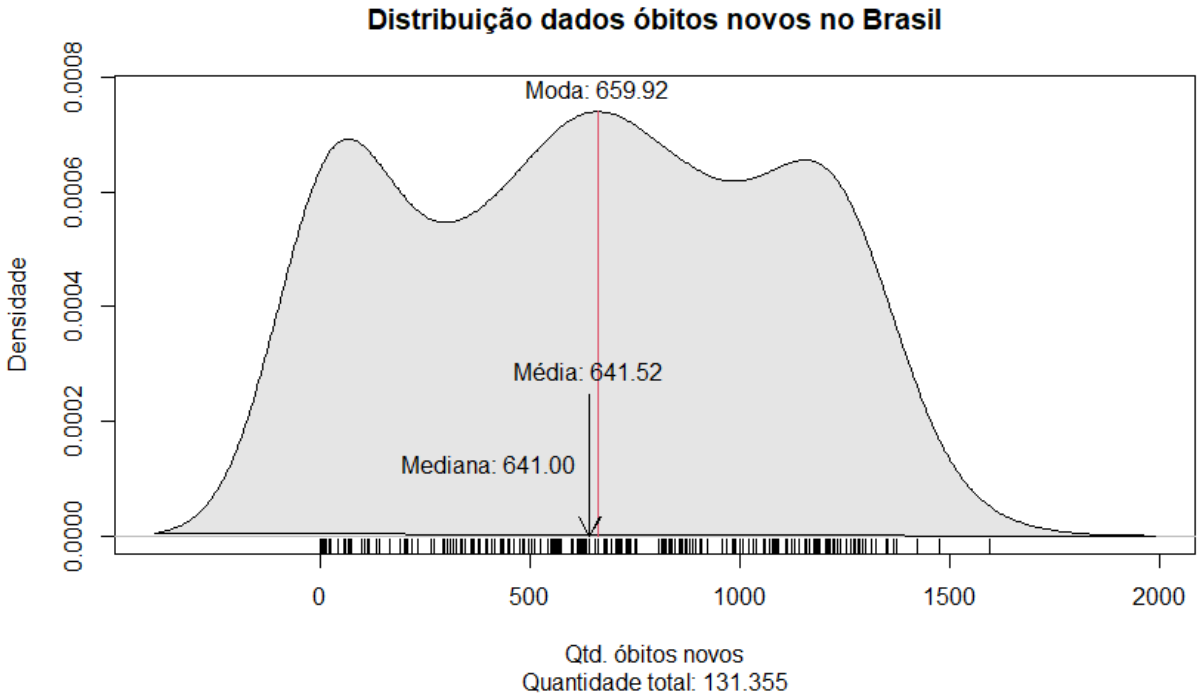


Figura 16: Distribuição dos dados sobre óbitos novos por quantidade de óbitos novos

Na figura 16 entende-se que a quantidade de óbitos novos está bem distribuída. Consequentemente, aplicando as funções skweness() e kurtosis() presente no software R, será possível avaliar o índice de assimetria e de curtose de tal distribuição. Sendo assim, o índice de assimetria de óbitos novos no Brasil é de 0.008811, já a curtose é igual à -1.216. De acordo com a assimetria encontrada, tem-se uma distribuição praticamente simétrica e possui curtose de natureza platicúrtica. Avaliando a moda, média e mediana, é possível verificar simetria nos dados.

Portanto, os dados referentes à óbitos novos no Brasil possuem uma distribuição simétrica e curtose de natureza platicúrtica. Tal cenário não é o ideal quando se trata sobre densidade de tais dados por quantidade de vítimas, mostrando então que, todas as quantidades, da mais baixa às mais altas foram densas (mesmo tais densidades possuindo eixo Y baixo).

Sendo assim, o governo deve reavaliar as causas de tal distribuição em relação à óbitos novos, pois, fica evidenciado um mal gerenciamento em relação à frear a quantidade de óbitos novos por parte de planos de saúde aplicados pelo governo e Ministério da Saúde.

Para confirmar e concluir tal análise, é possível fazer o uso do gráfico de histograma, que mostrará a frequência de tais óbitos novos por quantidade. Portanto, por meio do uso de histograma poderá ser possível entender como se deu a distribuição de frequência dos óbitos novos na coleta dos dados desde o inicio da pandemia até o dia 28 de Outubro:

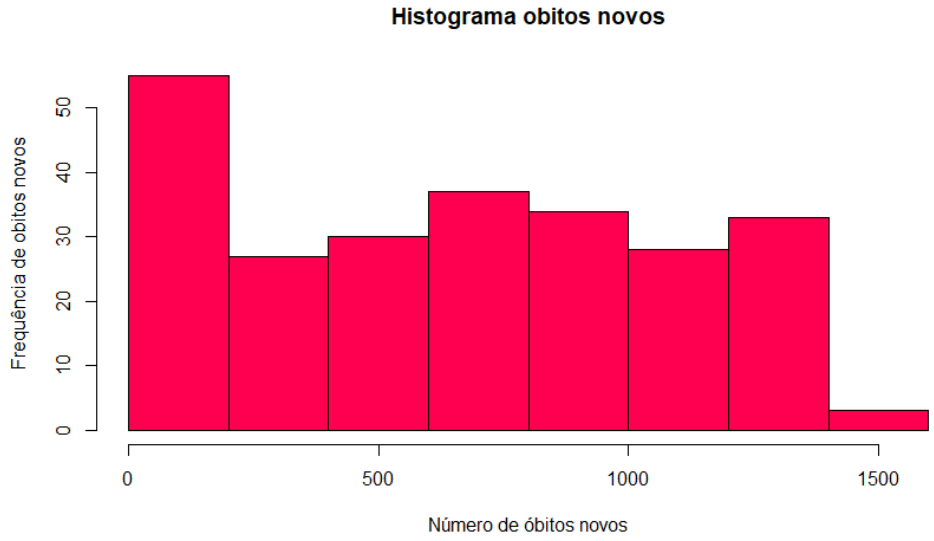


Figura 17: Histograma da frequência de óbitos novos por número de vítimas novas

Avaliando o histograma da figura 17 perceber-se que os dados estão simétricos, menos nas extremidades, onde isso pode ser representado como um fator positivo. A extremidade do inicio demonstra que muitos dias tiveram menos de 250 óbitos (tais dias que devem ser referentes ao inicio da pandemia), porém onde há o valor máximo, isto é, onde houveram mais de 1500 óbitos novos, a frequência demonstrou-se baixa, portanto poucos dias obtiveram aproximadamente, 1500 vítimas fatais novas.

Portanto, faz-se necessária a análise sobre o número de óbitos novos decorridos da COVID-19 por dia/mês, como é representado no seguinte gráfico de linha:

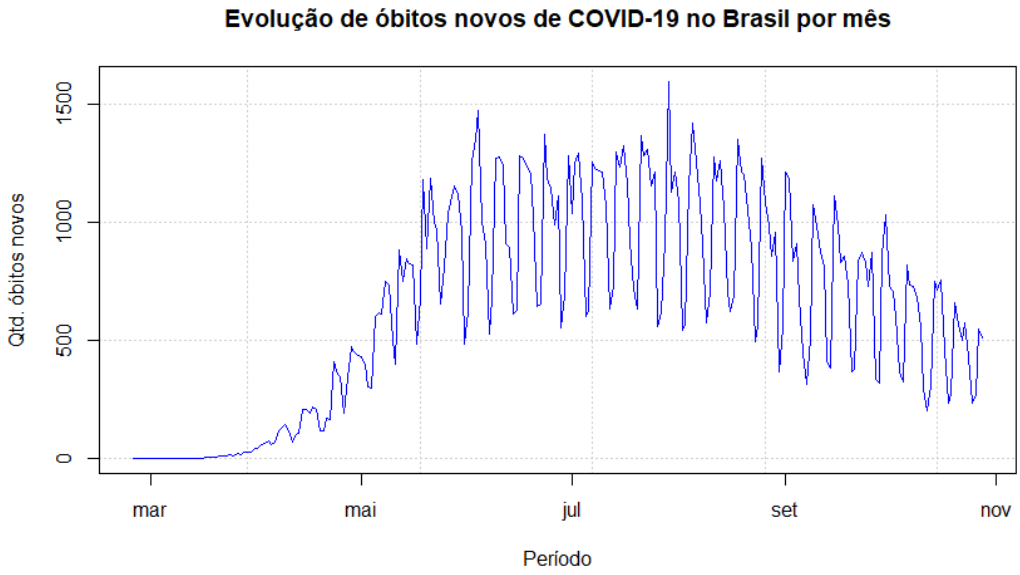


Figura 18: óbitos novos decorrente de COVID-19 por mês

No gráfico gerado referente ao número de óbitos novos por mês é possível verificar grande similaridade com o gráfico de casos novos por mês na figura 9, nos mostrando uma possível relação entre as duas variáveis. Sendo assim, torna-se justa a análise de correlação entre as variáveis Casos Novos e Óbitos Novos por meio de uma matriz de correlação:

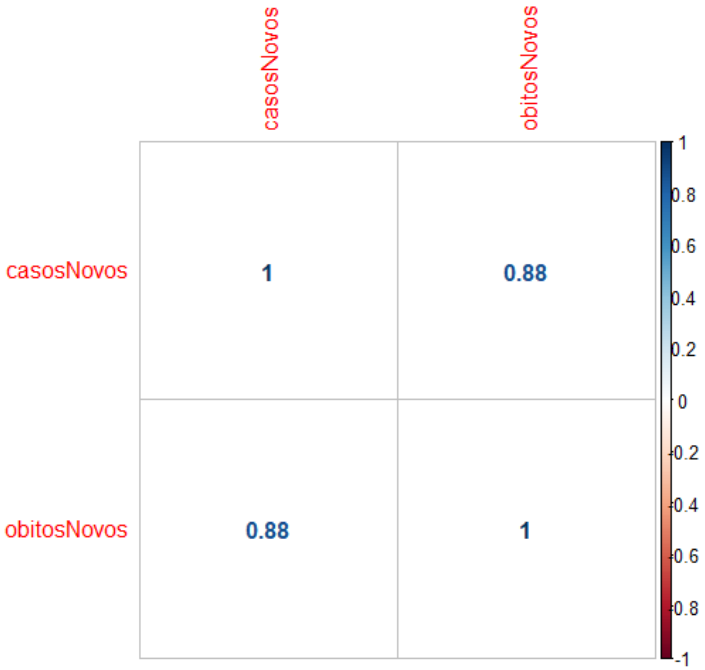


Figura 19: Matriz de correlação - óbitos novos e casos novos

Na matriz de correlação gerada, presente na figura 19, é evidente que as variáveis Casos Novos e Óbitos Novos possuem correlação positiva forte igual à 0.88, quando, de acordo com o estatístico Karl Pearson o máximo alcançado positivamente é 1. Sendo assim, pode-se analisar que, quando há o aumento do número de casos novos, possivelmente haverá o aumento de óbitos novos. Consequentemente, tal análise pode explicar a semelhança dos gráficos gerados para ambas variáveis.

Dessa maneira, é possível confirmar tal relação das duas variáveis por meio de um gráfico de dispersão da qual irá conter dados das duas variáveis cruzadas, e, posteriormente, a inserção de uma linha de regressão no gráfico referido, da qual tentará, por meio de uma linha linear, passar por todos os pontos confirmando a correlação entre as variáveis:

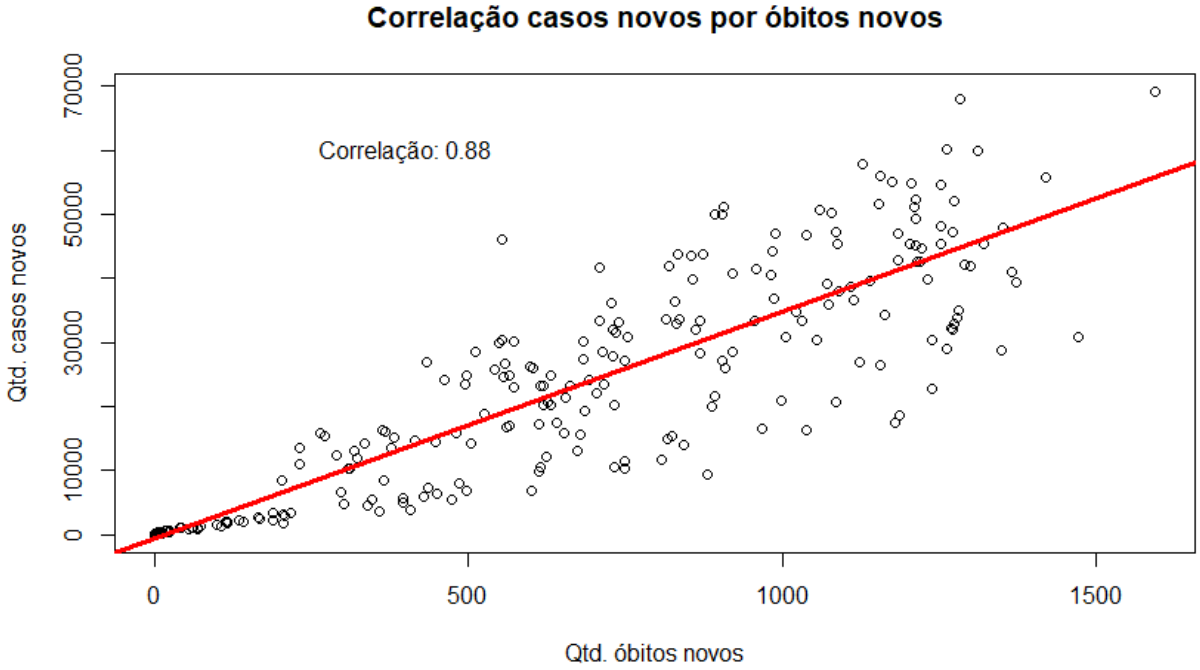


Figura 20: Regressão linear simples - óbitos novos e casos novos

Portanto, de acordo com o gráfico da figura 20 percebe-se que as variáveis Casos Novos e Óbitos novos possuem correlação linear forte positiva equivalente à 0.88, concluindo a ideia de que, quando há o aumento de diagnósticos positivos para COVID-19, aumentam-se as fatalidades novas decorrentes da mesma. Portanto, o governo brasileiro pode trabalhar para conter a variável mais maleável, isto é, os casos novos, podendo futuramente ocasionar com grande impacto o número de fatalidades novas decorrentes de doentes que possuem COVID-19.

Sendo assim, a continuação da análise em cima dos números obtidos pela variável **óbitos novos** torna-se necessária pois existem diversos fatores que podem ser explorados e explicados em estatísticas encontradas após a efetuação da exploração de dados com uso principalmente de gráficos de diversas naturezas.

Dessa maneira, visando entender os dados presenta em tal variável de forma profunda com uso de estatísticas descritivas e gráficos em cada mês de coleta dos dados. Assim, foi-se gerado o seguinte gráfico BoxPlot da variável óbitos novos por mês em numeral. Posteriormente estão as estatísticas referentes à cada mês representadas pela nomenclatura group:

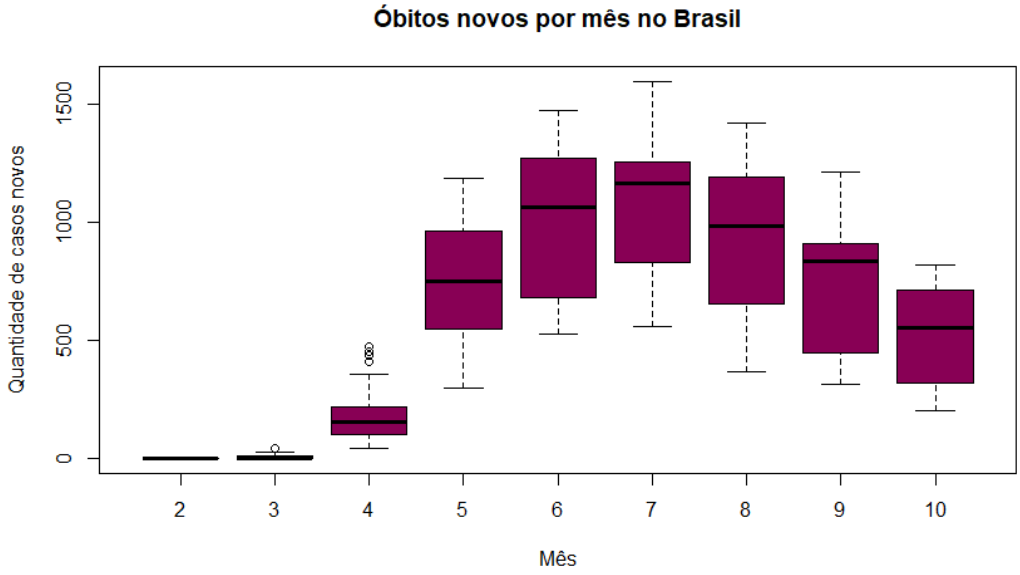


Figura 21: Boxplot - Óbitos novos por mês em numeral no Brasil

```
 Descriptive statistics by group
group: 2
   vars n mean sd median trimmed mad min max range skew kurtosis se
X1    1 5    0  0      0       0   0   0   0     0  NaN      NaN  0
------------------------------------------------------------------------------------------
group: 3
   vars  n mean    sd median trimmed mad min max range skew kurtosis   se
X1    1 31 6.48 10.11      0    4.56   0   0  42    42 1.73     2.72 1.82
------------------------------------------------------------------------------------------
group: 4
   vars  n mean    sd median trimmed   mad min max range skew kurtosis    se
X1    1 30  190 130.6    153   174.6 87.47  40 474   434 0.87    -0.59 23.85
------------------------------------------------------------------------------------------
group: 5
   vars  n  mean    sd median trimmed   mad min  max range  skew kurtosis    se
X1    1 31 755.3 266.6    749   755.9 323.2 296 1188   892 -0.03    -1.17 47.88
------------------------------------------------------------------------------------------
group: 6
   vars  n mean    sd median trimmed   mad min  max range  skew kurtosis    se
X1    1 30 1009 292.9   1066    1016 307.6 525 1473   948 -0.28    -1.46 53.47
------------------------------------------------------------------------------------------
group: 7
   vars  n mean    sd median trimmed   mad min  max range  skew kurtosis    se
X1    1 31 1061 283.6   1163    1073 188.3 555 1595  1040 -0.49    -1.03 50.94
------------------------------------------------------------------------------------------
group: 8
   vars  n  mean    sd median trimmed   mad min  max range  skew kurtosis    se
X1    1 31 932.5 300.2    984   938.3 407.7 366 1421  1055 -0.21    -1.36 53.92
------------------------------------------------------------------------------------------
group: 9
   vars  n  mean    sd median trimmed   mad min  max range  skew kurtosis    se
X1    1 30 752.4 275.6  832.5     754 226.1 310 1215   905 -0.27    -1.19 50.33
------------------------------------------------------------------------------------------
group: 10
   vars  n mean    sd median trimmed   mad min max range  skew kurtosis    se
X1    1 28  518 195.4    554   520.8 258.7 201 819   618 -0.19    -1.47 36.93
```

Figura 22: Estatísticas descritivas sobre óbitos novos por mês em numeral (group)

Novamente é possível verificar no gráfico gerado na figura 21 **comportamento** extremamente simular em relação ao gráfico contido na figura 10 (onde foram avaliado estatísticas sobre casos novos), onde os meses de picos se deram entre Julho e Agosto, cabendo ao mês de Julho o maior número de casos novos e agora, como avaliado na figura 15, o mês em que houve o maior número de óbitos novos.

Avaliando a figura 22 tem-se que a maior média de óbitos novos ocorreu no mês de Julho, com o equivalente à 1.061 vítimas novas fatais decorrente da COVID-19 por dia. Porém, fica evidente que a partir do mês de Agosto o número de óbitos novos inicia um estágio de declínio, sendo os três meses posteriores a Agosto com queda significativa de redução de 100 mortes por mês na média. Em Outubro, o mês que possuiu menor média de vítimas novas, a média de óbitos foi igual à 518 óbitos novos. Nesse mês houveram dias em que o máximo e o mínimo de óbitos foram iguais à 819 e 201 novas vitimas fatais, números que não eram tão baixos desde o mês de Maio. Porém a mediana referente ao mês de Outubro foi de 554 óbitos novos, ou seja, em 50% dos 28 dias do mês de Outubro verificaram números menores ou iguais à 554 óbitos novos. Tal medida se encontra muito próxima da média, o que não é recomendado a longo prazo (mesmo em um mês de redução).

Por meio do uso de um gráfico que nos gera uma linha de tendência, torna-se visível de maneira simplória a tendência dos dados referentes à cada dia da coleta da variável óbitos novos em determinado mês:

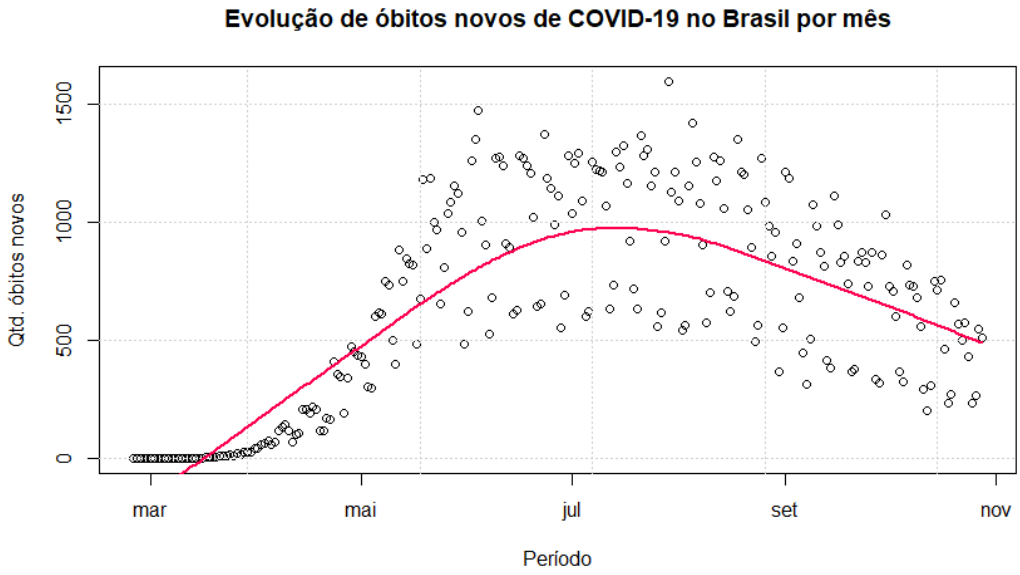


Figura 23: Linha de tendência óbitos por mês

Avaliando a linha de tendência gerada, do qual está referente na figura 23, percebe-se uma tendência de queda nos dados da variável casos novos ao passar do meses, como foi avaliado anteriormente na figura 21. Tal queda se inicia no final do mês de Agosto e continua em declínio até o fim do período presente na base de dados, isto é, 28 de Outubro.

4.2.4 Óbitos Acumulados

A variável **óbitos acumulado** é uma variável acumulativa (soma), isto é, soma todos os dados referentes à óbitos novos decorrentes de COVID-19 desde o inicio da coleta de tais dados. Sendo assim, como este trabalho visa explicar estatísticas sobre a pandemia de maneira clara ao leitor leigo, não torna-se necessária a análise de medidas de tendencia central, apenas o que importa é o valor do 100º percentil, ou seja, o valor total acumulado até então. Dessa maneira, é possível verificar na imagem 4 que o número total de óbitos até o dia 28 de Outubro de 2020, no território brasileiro, é equivalente à 158.456 óbitos. Por meio da variável óbitos acumulados pode-se perceber o quão rápida foi a soma de mortes:

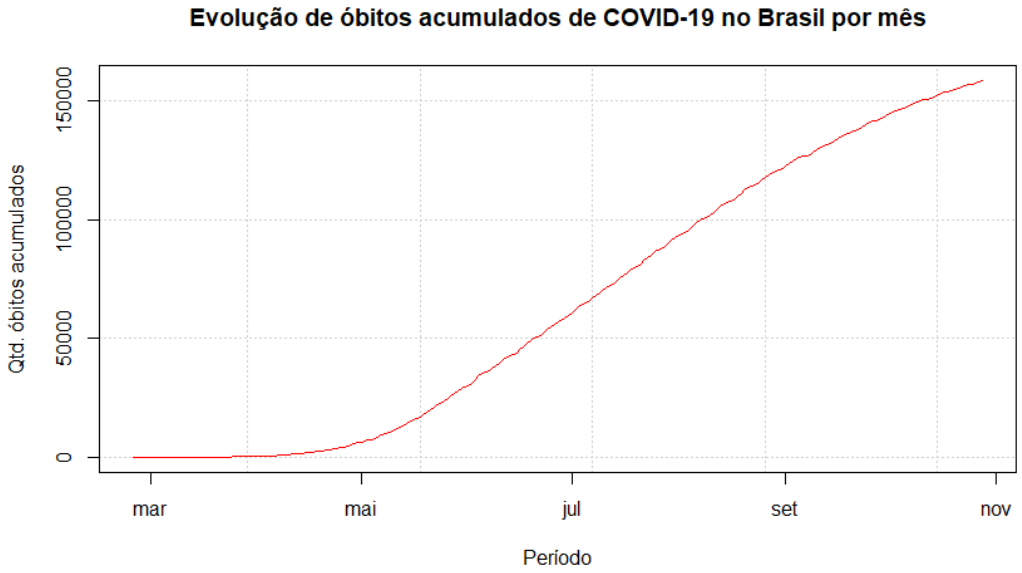


Figura 24: Óbitos acumulados de COVID-19 por mês no Brasil

Pode-se enxergar na figura 24 que o número de óbitos cresceu rapidamente entre os meses de Maio e Julho se comparado ao período anterior, de Março à Maio. Esse crescimento pode ter sido desencadeado por falta de fiscalização e reavaliação da causa de morte dos pacientes de meses anteriores, uma vez que o vírus era novo e poucas ferramentas existiam para a sua detecção. Um fator negativo presente no gráfico se da por ele ainda continuar em contínuo crescimento, pois, por ser um gráfico representado por uma variável acumulativa ele nunca entrará em declínio, apenas, nos melhores dos casos, ficará paralelo ao eixo X, o que significa que não está havendo óbitos novos decorrentes de COVID-19 pois novos números sobre tal não estão sendo somados/acumulados.

4.2.5 Em Acompanhamento

Com a disseminação em massa do vírus diante a população brasileira, houveram muitas internações e atendimentos médicos visando o melhor tratamento para que o individuo fosse o quanto antes curado. Dessa maneira, o governo e diversos planos de saúde lançaram campanhas de telemedicina[11] com o intuito de possíveis contaminados conseguirem ter contato com determinada equipe hospitalar e serem acompanhados ao longo do enfrentamento contra vírus.

Sendo assim, o governo brasileiro por meio do Ministério da Saúde iniciou a coleta dos dados sobre pacientes em acompanhamentos e divulgação dos mesmos apenas a partir da data de 19 de Abril de 2020, portanto, os 54 dias anteriores (em relação ao inicio da coleta e divulgação dos dados) foram nulos, pois não era aplicado tal procedimento ate então.

De acordo com a figura 4 os dados referentes ao Brasil e pacientes em acompanhamentos possuíram 54 dados faltantes. A média de pacientes em acompanhamento por dia desde o inicio da pandemia foi de 47.0385 pacientes, porém com desvio padrão alto igual à 23.1495 pacientes (dependendo do dia de avaliação no período da pandemia), pois, certamente tal número obteve grande variação por causa dos meses de pico, Julho e Agosto, onde houve o pico de infectados e então pessoas procuravam acompanhamento. A mediana, isto é, metade dos dias que houveram a coleta e divulgação de tais dados possuíram números menores ou iguais à 49.9513 pacientes em acompanhamento. Agora visualizando a quantidade de pacientes em acompanhamento por mês, é gerado o seguinte gráfico:

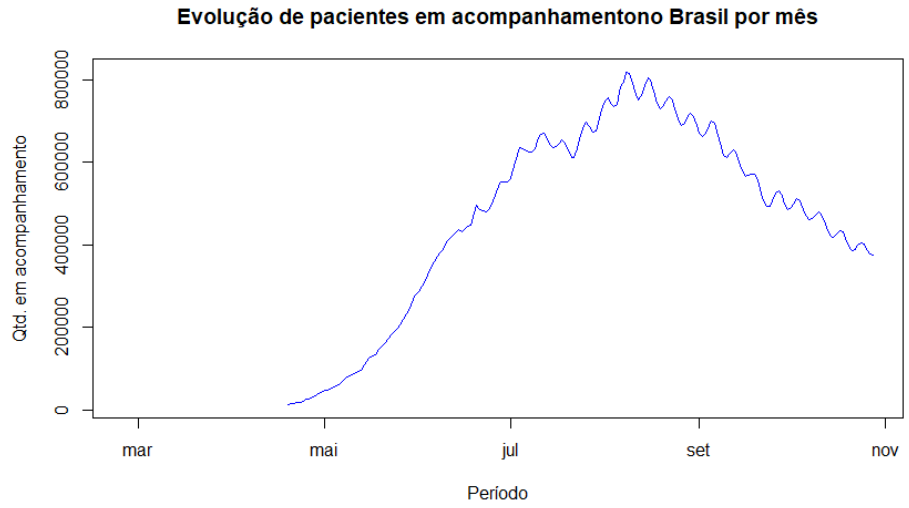


Figura 25: Pacientes em acompanhamento por COVID-19 no Brasil por mês

Analisando a figura 26 percebe-se um crescimento continuo no inicio do mês de Abril (onde houve o inicio da coleta de tais dados) até meados do mês de Julho, após tal período inicia-se grandes sazonalidades nos dados diários, acabando com o crescimento continuo. Esse crescimento continuo visto no mês de Abril até meados de Julho era um fator extremamente positivo, pois demonstrava que dia após dia, novos pacientes estavam procurando tratamento médico contra o novo coronavírus por meio de algum tipo de equipe médica.

É possível perceber similaridade do gráfico presente na figura 26 que trata sobre pacientes em acompanhamento com o gráfico da figura 10 que se refere ao casos novos por período. Pode-se interpretar que, com o crescimento de pacientes novos infectados pela doença, houve também o crescimento da procura de tais pacientes em tratamentos médicos contra o vírus. E, com a redução de novos diagnósticos positivos, houve também a redução obvia de pacientes em tratamento.

[11] Telemedicina: secad.artmed.com.br/blog/medicina/telemedicina-na-pandemia-do-novo-coronavirus/

4.2.6 Recuperados

De acordo com especialistas, o termo "curado"é prematuro, uma vez que diversos pacientes apresentam sequelas após tratamentos contra o novo coronavírus. Os pacientes que testaram positivo ao vírus podem reagir de diversas formas, tudo depende de seu sistema imunológico, alguns necessitam de UTI, outros apenas repouso e alguns medicamentos, tudo deve ser bem acompanhado por meio de determinado médico. É válido também dizer que diversos indivíduos que já receberam alta hospitalar após o tratamento contra a COVID-19 se reinfectaram novamente, corroborando a teoria de imunização após contrair o vírus.

Sendo assim, como a variável emAcompanhamento, os dados sobre pacientes recuperados apenas começaram a ser coletados e publicados a partir do dia 19 de Abril de 2020. Portanto, os 54 dias anteriores (em relação ao inicio da coleta e da divulgação dos dados) foram nulos, pois não eram coletados e divulgados dados sobre pacientes recuperados.

De acordo com a figura 4 e explicado no parágrafo anterior, a variável recuperados possuí 54 dados faltantes. A média de pacientes recuperados foi de 1.928.989 com desvio padrão alto igual à 1.666.130, tal variação pode ter se dado pois a cada dia que passava o número de pacientes recuperados ia crescendo cada vez mais, e, a média é uma medida de tendência central que sofre alteração com dados discrepantes. Porém ao avaliar o seguinte gráfico percebemos que a variável não se trata de pacientes **novos** recuperados, e sim, acumulados:

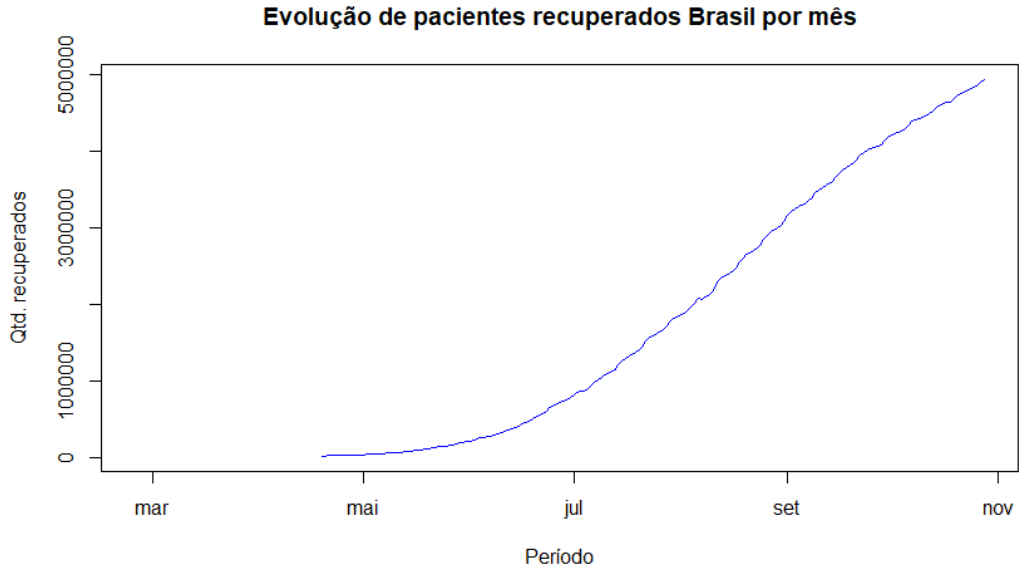


Figura 26: Pacientes recuperados de COVID-19 no Brasil por mês

Sendo assim, a grandiosidade da média pode ser explicada por tal motivo (pois na verdade, eram dados acumulados), e é perceptível que o gráfico nos refere à uma variável que possui valores acumulados, o que inicialmente não tinha sido explicado pelo portal do Ministério da Saúde. Portanto, até o dia 28 de Outubro de 2020 houveram 4.934.548 pacientes recuperados.

O número de 4.934.548 pacientes recuperados é um ótimo sinal, pois, está próximo do número de casos novos acumulados, que é de 5.468.270. Portanto, pode-se interpretar que, dos 5.468.270 pacientes que obtiveram diagnóstico positivo para COVID-19, aproximadamente 90%, isto é, 4.934.548 pacientes conseguiram se recuperar.

4.3 Variáveis categóricas

O método skim presente em uma das bibliotecas do software R torna possível a exploração dos dados presentes da base de dados, como foi dito anteriormente na seção 4. Foram avaliados até então as principais variáveis numéricas presentes na base de dados referente ao país Brasil, sendo essas: Casos Novos, Casos Acumulados, Óbitos Novos e Óbitos Acumulados. Dessa maneira, é justa a análise dos dados categóricos presentes na base de dados, e, após usar o método skim, as informações sobre os dados categóricos é retornado da seguinte maneira (Dados retornados após o filtro, na base de dados COVID19.SAS, para estado igual à "Brasil", como explicado na seção 3):

```
-- Variable type: character -----------------------------------------------------
# A tibble: 4 x 8
  skim_variable   n_missing complete_rate   min   max empty n_unique whitespace
* <chr>               <int>         <dbl> <int> <int> <int>    <int>      <int>
1 regiao                  0             1     6     6     0        1          0
2 estado                  0             1     0     0   247        1          0
3 municipio               0             1     0     0   247        1          0
4 nomeRegiaoSaude         0             1     0     0   247        1          0
```

Figura 27: Propriedades variáveis categóricas dos dados referentes ao Brasil

Portanto, analisando a figura 27, a base dos dados sobre COVID-19 referentes ao Brasil, possui 4 variáveis categóricas, sendo essas: regiao, estado, municipio e nomeRegSaude, variáveis essas das quais já foram explicadas anteriormente na seção 3. Porém, é válido ressaltar, como explicado anteriormente na seção 4.2, que as variáveis numéricas presente na base dados, da qual foram entendidas como numéricas na verdade possuíam caráter categórico, sendo essas: coduf, codmun, codRegiaoSaude, semanaEpi e interior/metropolitana. A abordagem usada no entendimento do tipo da variável leva em consideração possíveis e futuros estudos e aplicações de estatísticas. Sendo assim, as variáveis categóricas referentes ao Brasil são: **regiao, estado, municipio, nomeRegSaude, coduf, codmun, codRegiaoSaude, semanaEpi e interior/metropolitana**.

Entretanto, como foi avaliado, apenas as variáveis semanaEpi e codmun possui campos preenchidos, sendo tais campos preenchidos com o código referente à semana Epidemologica e à região referente ao Brasil, respectivamente. Porém, como entende-se na figura 27 os valores presentes na variável coduf possuem apenas um único valor, que é referente ao código da região Brasil, ou seja, 76, como é possível visualizar na seguinte imagem da base de dados na IDE RStudio:

	regiao	estado	municipio	coduf	codmun	codRegiaoSaude	nomeRegiaoSaude	semanaEpi	interior/metropolitana
1	Brasil			76	NA	NA		9	NA
2	Brasil			76	NA	NA		9	NA
3	Brasil			76	NA	NA		9	NA
4	Brasil			76	NA	NA		9	NA
5	Brasil			76	NA	NA		9	NA
6	Brasil			76	NA	NA		10	NA
7	Brasil			76	NA	NA		10	NA
8	Brasil			76	NA	NA		10	NA
9	Brasil			76	NA	NA		10	NA
10	Brasil			76	NA	NA		10	NA
11	Brasil			76	NA	NA		10	NA
12	Brasil			76	NA	NA		10	NA
13	Brasil			76	NA	NA		11	NA
14	Brasil			76	NA	NA		11	NA
15	Brasil			76	NA	NA		11	NA
16	Brasil			76	NA	NA		11	NA
17	Brasil			76	NA	NA		11	NA
18	Brasil			76	NA	NA		11	NA
19	Brasil			76	NA	NA		11	NA

Figura 28: Visualização variáveis categóricas Brasil

Em relação às propriedades presentes em tais dados categóricos, como é visto nas figuras 27 e 28, percebe-se que nenhuma das quatro variáveis categóricas apresenta dados faltantes, porém três variáveis: estado, município e nomeRegSaude estão vazias, ou seja, não possuem dados faltantes pois nem se quer foram preenchidas nas periódicas coletas de dados das quais foram efetuadas. A única variáveis categórica que apresenta dados preenchidos nas 247 linhas é a variável Região, da qual possui valor único referente ao filtro utilizado na base, ou seja, apenas a região que era igual ao campo Brasil.

Portanto, após avaliar a situação das variáveis categóricas presentes na base de dados para o filtro da região Brasil, conclui-se que não é possível realizar análises sobre as variáveis categóricas. Porém, será executado no próximo tópico, um filtro referente aos estados e especificamente ao Distrito Federal, do qual haverá a realizações de análises das variáveis da mesma maneira da qual foi concluída nesse tópico.

4.4 Breve análise exploratória e visualização dos dados sobre COVID-19 regiões do Brasil

Visando agora explorar a base de dados toda sem filtro para estado igual à "Brasil", foi-se retirado tal filtro para então explorar dados referentes aos estados brasileiros. Também, para tornar a criação de gráficos mais dinâmica e atrativa ao leitor, foi-se usado a biblioteca ggplot2 presente no software R, com ela a manipulação de gráficos torna-se mais rica e maleável por meio do autor.

4.4.1 Casos COVID-19 por região

Sendo assim, para tentar descobrir como estão distribuídos os 5.458.270 pacientes infectados com o novo coronavírus pelas regiões brasileiras, gerou-se o seguinte gráfico:

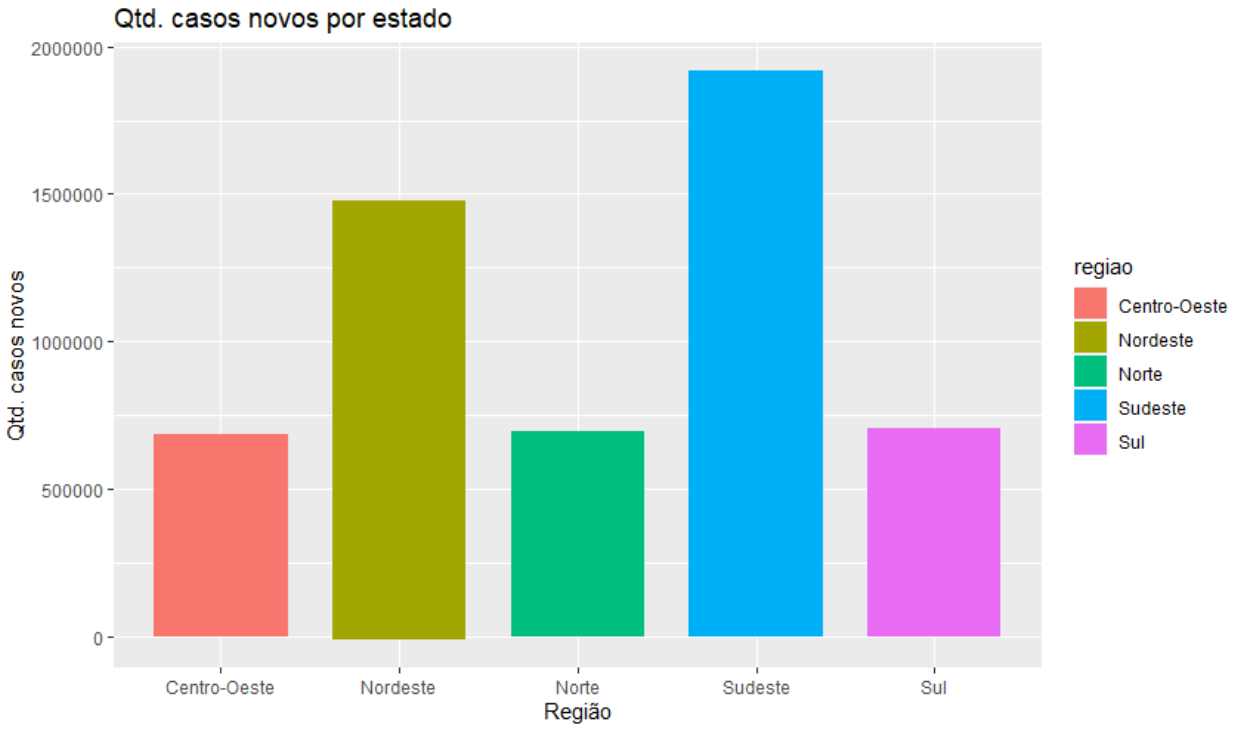


Figura 29: Quantidade de casos novos por região brasileira

Podemos perceber que, dos aproximadamente 5 milhões e meio de pacientes infectados, a maior parte está na região Sudeste e Nordeste, já as demais regiões, Sul, Centro-Oeste e Norte apresentam dados bastante próximos entre si.

Os primeiros casos relatados de brasileiros infectados com o novo coronavírus em território brasileiro vieram da capital São Paulo[12], no Sudeste, não demorou muito para o vírus se espalhar por todas regiões brasileiras. A região Sudeste por ser a mais povoada e populosa do Brasil rapidamente tornou-se o epicentro da disseminação do vírus em território nacional. Porém, ao chegar em na região Nordeste, o vírus se alastrou tão rápido quanto ocorrido no Sudeste pois em diversos estados presente na região Nordeste não haviam assistência médica de qualidade para a população mais carente, seja essa assistência referente ao tratamento quanto à divulgação dos cuidados necessários contra o vírus. De acordo com o IBGE[13].

[12]G1 Globo: g1.globo.com/sp/sao-paulo/noticia/2020/08/26/primeiro-caso-confirmado-de-covid-19-no-brasil-ocorreu-em-sp-e-completa-seis-meses-nesta-quarta.ghtml
[13]IBGE: www.ibge.gov.br/estatisticas/sociais/populacao.html

4.4.2 Casos COVID-19 por estado

Dessa maneira, visando a visualização de pacientes diagnósticas positivamente com COVID-19 por estado, tem-se o seguinte gráfico de barras:

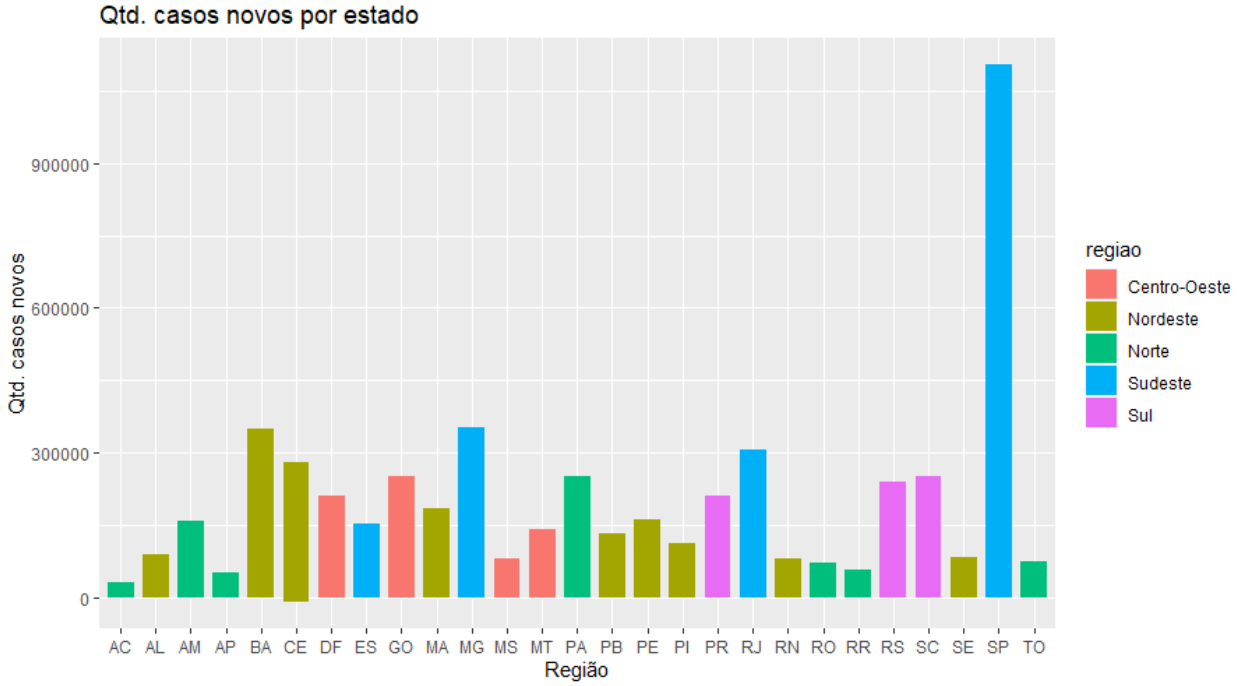


Figura 30: Quantidade de casos novos por estado brasileira

No gráfico da figura 30 é evidente que o estado de São Paulo (SP) foi o que obteve maior número de pacientes infectados por COVID-19, tal fato pode ter se dado por São Paulo ser o estado mais populoso do território brasileiro, ajudando na disseminação rápida do vírus. Os outros estados do Sudeste não obtiveram números tão discrepantes como o estado de São Paulo. Na região Centro-Oeste o estado de Goiás foi o que obteve maior número de pacientes infectados. Já na região Sul, os três estados que a compõe (Santa Catarina, Paraná e Rio grande do Sul) possuíram números similares entre si. Por fim, na região Norte os números de pacientes infectados por estado foi bastante variante, porém com destaque no estado do Pará (PA) do qual foi o estado com maior número de diagnósticos positivo para COVID-19 na região Norte.

4.4.3 Óbitos COVID-19 por região

Das nações mundiais, o Brasil é um dos principais países em relação à quantidade de óbitos decorrentes de COVID-19. Sendo assim, visando visualizar e analisar como se distribuiu os números de óbitos pelo novo coronavírus, gerou-se o seguinte gráfico de barras separado por regiões:

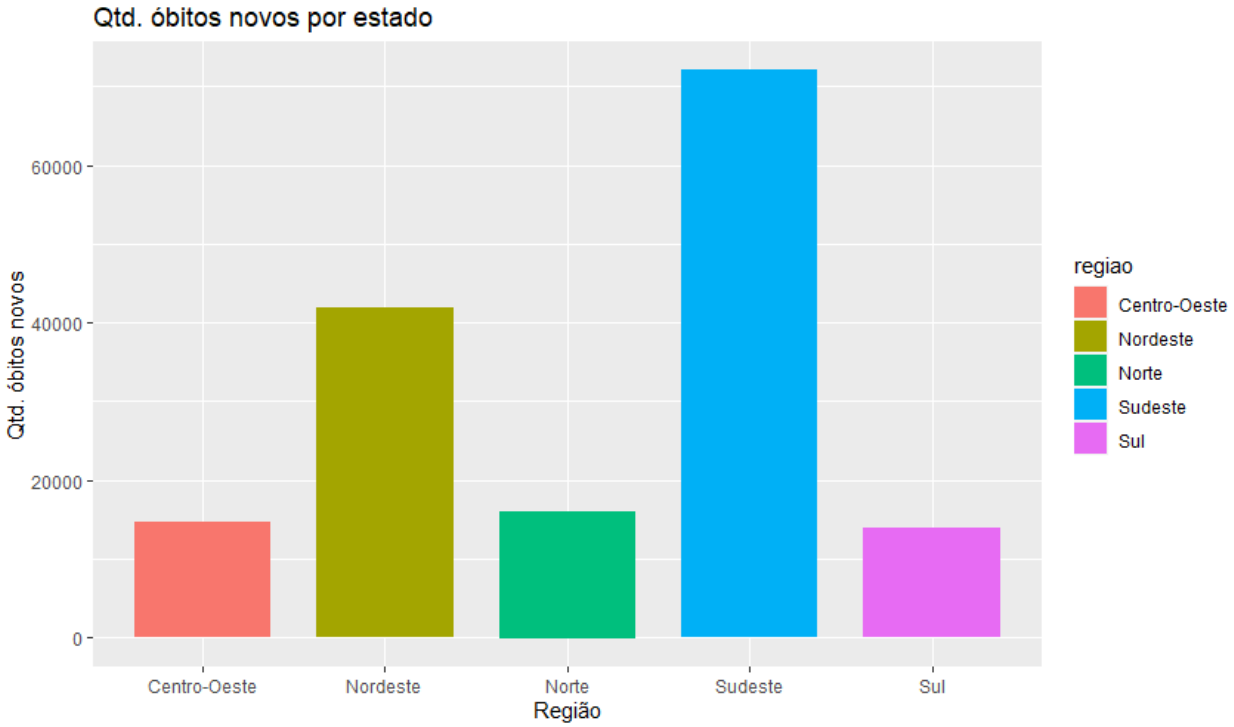


Figura 31: Quantidade de óbitos novos por região brasileira

De acordo com o gráfico de barras apresentado na figura 31 e pela tabela disponibilizada pelo Ministério da Saúde na figura 1, percebe-se que novamente, como visto na figura 29, a região Sudeste é a que possuí maior número de óbitos decorrentes de COVID-19, número esse equivalente à soma dos números da região Norte, Centro-Oeste e Sul.

Já a região Nordeste foi a segunda que mais possuiu óbitos por COVID-19 no território brasileiro. As regiões Norte, Sul e Centro-Oeste possuíram número parecidos em relação ao número de vitimas, o mesmo fenômeno encontrado no gráfico da figura 29 sobre casos novos. Portanto, como dito em análises anteriores, pode-se concluir que as regiões que possuíram grande número de diagnósticos positivos para COVID-19, também possuíram maior número de óbitos decorrentes do novo coronavírus.

Concluí-se que, a região que possui maior número de casos também é a região que mais possui óbitos decorrentes de COVID-19, isto é, a região Sudeste, que é representante da parte do território brasileiro que mais possui brasileiros residentes.

4.4.4 Óbitos COVID-19 por estado

Consequentemente, torna-se justa a análise da distribuição do número de óbitos por estado para tentar visualizar e analisar em que sentido se deu tal distribuição dos óbitos. Então, por meio do gráfico de barras para essa variável, tem-se:

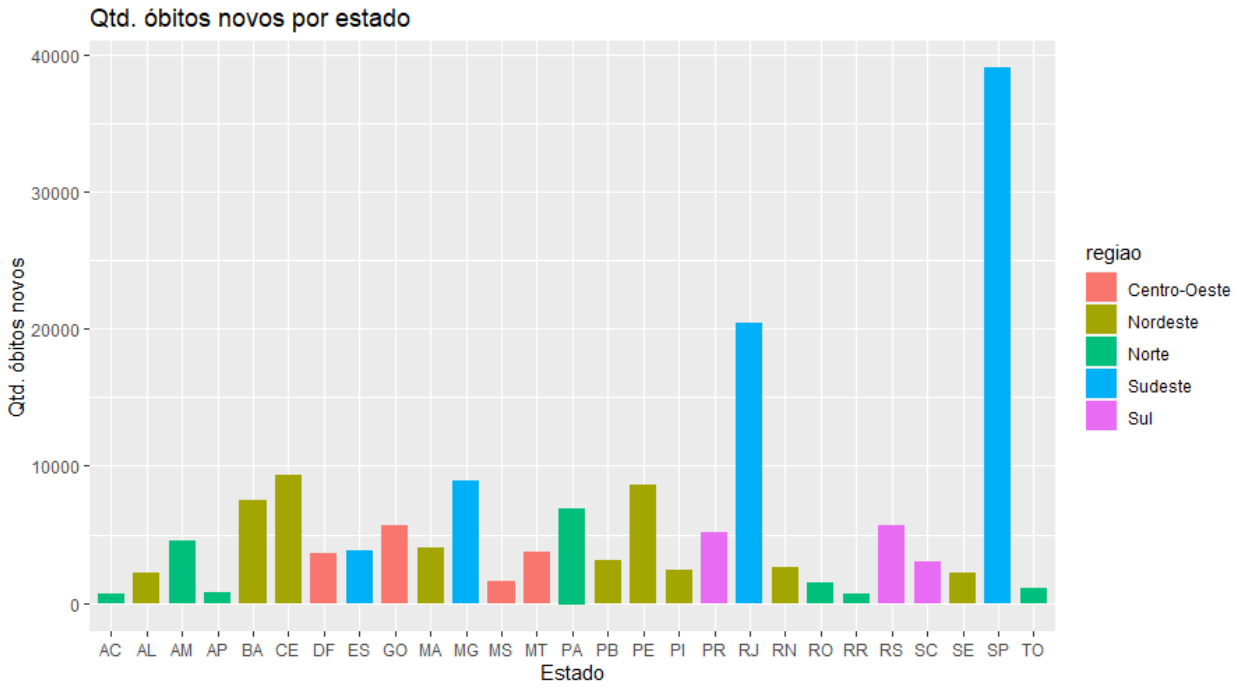


Figura 32: Quantidade de óbitos novos por estado brasileira

Analisando o gráfico da figura 32, é evidente que o estados da região do Sudeste são os que mais possuem vítimas fatais decorrentes da COVID-19. O maior número de óbitos decorrentes do novo coronavírus é referente ao estado de São Paulo (SP) seguido do estado do Rio de Janeiro (RJ), estados esses que compõe a região Sudeste.

Na região Centro-Oeste, o estado de Goiás foi o estado que possuiu o maior número de óbitos, estado esse que também possuiu o maior número de diagnósticos positivos na região Centro-Oeste. Já na região Norte, o estado que mais possuiu número óbitos foi o estado do Pará (PA), estado que também foi o que obteve maior número de diagnósticos positivos para COVID-19 na região Norte. Já os três estados que compõe a região Sul possuíram número de óbitos similares, da mesma maneira se comparado com o gráfico da figura 30. Por fim, a região Nordeste obteve vasta variância entre os números de vitimas fatais decorrentes do vírus, porém os destaques dos estados que possuíram mais óbitos foram Bahia (BA), Ceará (CE) e Pernambuco (PE).

5 Análise exploratória e visualização dos dados do COVID-19 no Distrito Federal

Como explicado no plano de análise na seção 2, agora será efetuado a mesma análise porém para a região autônoma do Distrito Federal, do qual possui em seu território a capital do país Brasil.

Na análise dos dados gerados pelo COVID-19 para o Distrito Federal, será usado o método skim() presente na biblioteca Skim por meio do software R e RStudio. Sendo assim, as análises serão da mesma maneira e com o mesmo padrão como efetuado nos dados referentes ao Brasil.

É valido ressaltar que, foi-se utilizado um filtro na base de dados **covid19** para o estado do **DF** . Dados esses coletados até o dia 28 de Outubro de 2020 referentes à pandemia deste estado após a aplicação do filtro dito.

5.1 Propriedades dos dados referente ao estado Distrito Federal

Por meio da biblioteca Skim, será avaliada as propriedades gerais presentes nos dados coletados, propriedades essas que não foram explicadas na seção 3.

Ao aplicarmos a função skim na base de dados é retornado uma tabela sinalizando as diversas características que serão encontrados nos dados das futuras análises:

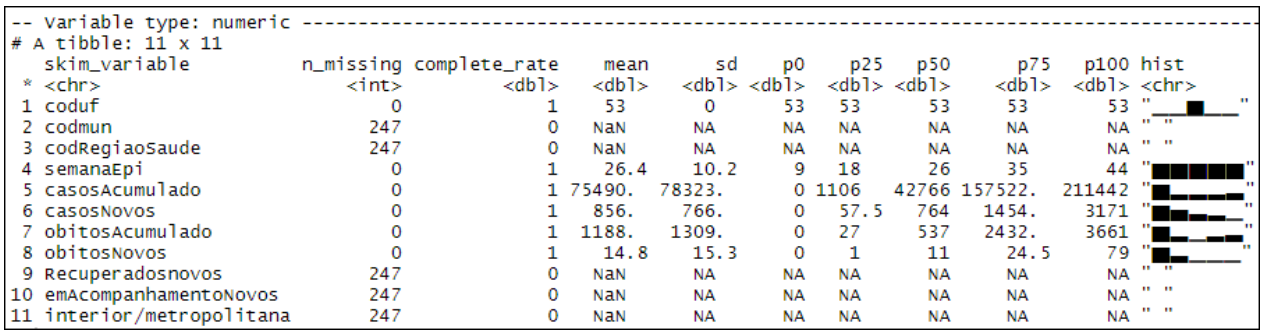
```
> skim(covid_DF)
-- Data Summary ------------------------
                           Values
Name                       covid_DF
Number of rows             247
Number of columns          16
_________________________
Column type frequency:
  character                4
  Date                     1
  numeric                  11
_________________________
Group variables            None
```

Figura 33: Propriedades gerais dados Distrito Federal

Sendo assim, após aplicado o método skim para análise das propriedades dos dados referentes ao Distrito Federal, representado na figura 33, tem-se que existem 247 linhas e 16 colunas na base de dados filtrada para esse estado. A frequência de variáveis tipo carácter (ou categóricos) é de 4, existem 1 variável do tipo referentes à data e 11 variáveis do tipo numéricos. É evidenciado também que não existem grupos de variáveis na base de dados filtrada.

5.2 Variáveis Numéricas

Como citado anteriormente, existem 11 variáveis numéricas na base de dados filtrada para a região autônoma do Distrito Federal, porém, aplicando novamente o método skim, gera-se a seguinte tabela:



```
-- Variable type: numeric -------------------------------------------------------------------------------
# A tibble: 11 x 11
   skim_variable          n_missing complete_rate    mean      sd    p0    p25   p50      p75   p100 hist
 * <chr>                      <int>         <dbl>   <dbl>   <dbl> <dbl>  <dbl> <dbl>    <dbl>  <dbl> <chr>
 1 coduf                          0             1    53       0      53   53       53      53       53 "▁▁▁▁▇▁▁▁"
 2 codmun                       247             0   NaN      NA      NA   NA       NA      NA       NA " "
 3 codRegiaoSaude               247             0   NaN      NA      NA   NA       NA      NA       NA " "
 4 semanaEpi                      0             1    26.4    10.2    9   18       26      35       44 "▇▇▇▇▇▇"
 5 casosAcumulado                 0             1 75490.  78323.      0 1106   42766 157522.  211442 "▇▂▂▂▂▃"
 6 casosNovos                     0             1   856.    766.      0   57.5   764   1454.    3171 "▇▅▂▃▂▁"
 7 obitosAcumulado                0             1  1188.   1309.      0   27     537   2432.    3661 "▇▂▁▂▂▃"
 8 obitosNovos                    0             1    14.8    15.3     0    1      11     24.5     79 "▇▃▂▁▁▁"
 9 Recuperadosnovos             247             0   NaN      NA      NA   NA       NA      NA       NA " "
10 emAcompanhamentoNovos        247             0   NaN      NA      NA   NA       NA      NA       NA " "
11 interior/metropolitana       247             0   NaN      NA      NA   NA       NA      NA       NA " "
```

Figura 34: Propriedades variáveis numéricas dos dados referentes ao Distrito Federal

De acordo com a figura 34 as linhas são representadas pelas variáveis presentes na base de dados filtrada para o estado do DF e as colunas são medidas de estatística descritiva sobre os dados numéricos. Variáveis como **coduf, codmun, codRegiaoSaude, semanaEpi e interior/metropolitana** são variáveis representadas por números, porém nesse contexto que não fazem sentido pois futuramente haverá a aplicação de estatísticas sobre tais. Entretanto, dessas, codmun, codRegiaoSaude e interior/metropolitana possuem 100% dos seus dados faltantes.

As variáveis numéricas **Recuperadosnovos** e **emAcompanhamentoNovos** são variáveis numéricas autênticas, porém todos os seus respectivos dados são faltantes, pois, por algum motivo, não foram colhidos ou divulgados dados de tais variáveis referentes ao Distrito Federal.

Sendo assim, as variáveis numéricas que possuem dados preenchidos em relação ao Distrito Federal, são: **casosNovos**, **casosAcumulado**, **obitosNovos** e **obitosAcumulado**. Em relação às essas variáveis, entende-se que todas estão preenchidas totalmente. Portanto, nos próximos tópicos serão analisados, por meio de gráficos, tais variáveis numéricas, visando a explicação e conclusões sobre os dados e suas respectivas estatísticas e distribuições.

5.2.1 Casos Novos

O Distrito Federal é representante da capital política do país, tal região autônoma possui grande infraestrutura porém vasta diversidade. Dessa maneira, será analisado o numero de indivíduos que testaram positivo para o novo coronavírus nesse determinado território que compõe o Brasil. Sendo assim, na figura 34 é possível verificar as estatísticas descritivas referentes ao número de casos novos de COVID-19 no DF, porém, de maneira à complementar os dados das estatísticas, foi-se gerada a seguinte tabela:

```
> summary(covid_DF$casosNovos)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    0.0    57.5   764.0   856.0  1454.5  3171.0 
```

Figura 35: Estatísticas casos novos

De acordo com as figuras 34 e 35, a média de diagnósticos positivos para o novo coronavírus foi de 856 casos por dia no Distrito Federal, porém com desvio padrão igual à 766 novos casos para mais ou para menos, dependendo do dia da avaliação no período da pandemia. Pode-se verificar que 50% dos dias que houve coleta dos dados obtiveram quantidade igual ou menor à 764 diagnósticos positivos para indivíduos residentes do Distrito Federal.

O maior número de casos em determinado dia foi igual à quantidade de 3.171 diagnósticos positivos, tal fenômeno que ocorreu no dia 25 de Agosto de 2020, como é possível perceber na seguinte imagem:

```
> covid_DF$data[covid_DF$casosNovos == 3171]
[1] "2020-08-25"
```

Figura 36: Data referente ao maior número de casos novos por dia no DF

Então, para avaliar a distribuição dos dados referentes à casos novos no Distrito Federal, tem-se o seguinte gráfico de densidade:

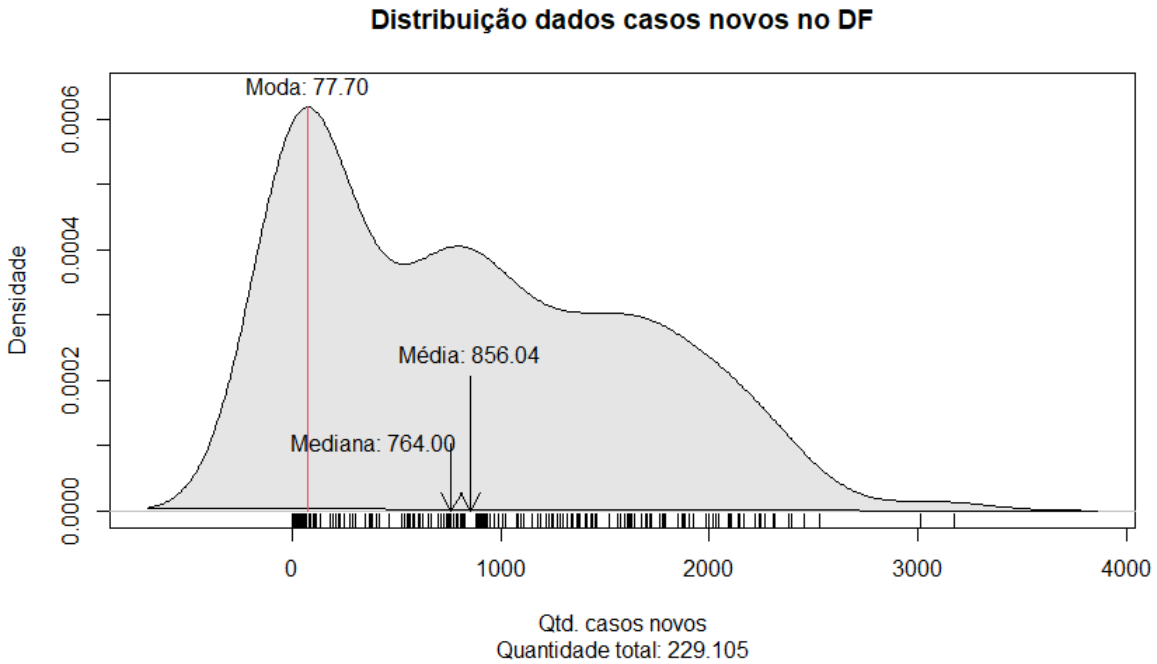


Figura 37: Data referente ao maior número de casos novos por dia no DF

Na figura 37 percebe-se que a quantidade de casos novos é mais densa quando há menos de 1.000 novos diagnósticos positivos por dia. Aplicando as funções skwness() e kurtosis() presente no software R, será possível avaliar o índice de assimetria e grau de curtose da distribuição. Então, o índice de assimetria de casos novos no Distrito Federal é de 0.537, já o grau de curtose é igual à -0.7527. De acordo com a assimetria encontrada, tem-se uma assimetria moderada na distribuição dos dados e possui curtose de natureza platicúrtica. Avaliando a moda, média e mediana, verifica-se uma assimetria à direita dos dados.

Portanto, os dados referentes à casos novos no Distrito Federal possuem uma distribuição assimétrica moderada para a direita com curtose de natureza platicúrtica. Tal cenário reflete de maneira sucinta a análise encontrada na distribuição dos dados referentes ao Brasil.

A fim de confirmar e concluir a análise feita para essa variável, fez-se uso do gráfico de histograma demonstra a frequência de cada quantidade de casos novos no DF. Portanto, com o uso do gráfico de histograma é possível analisar e entender a distribuição por frequência dos dados sobre casos novos após a coleta dos dados que obteve fim (neste trabalho) no dia 28 de Outubro de 2020:

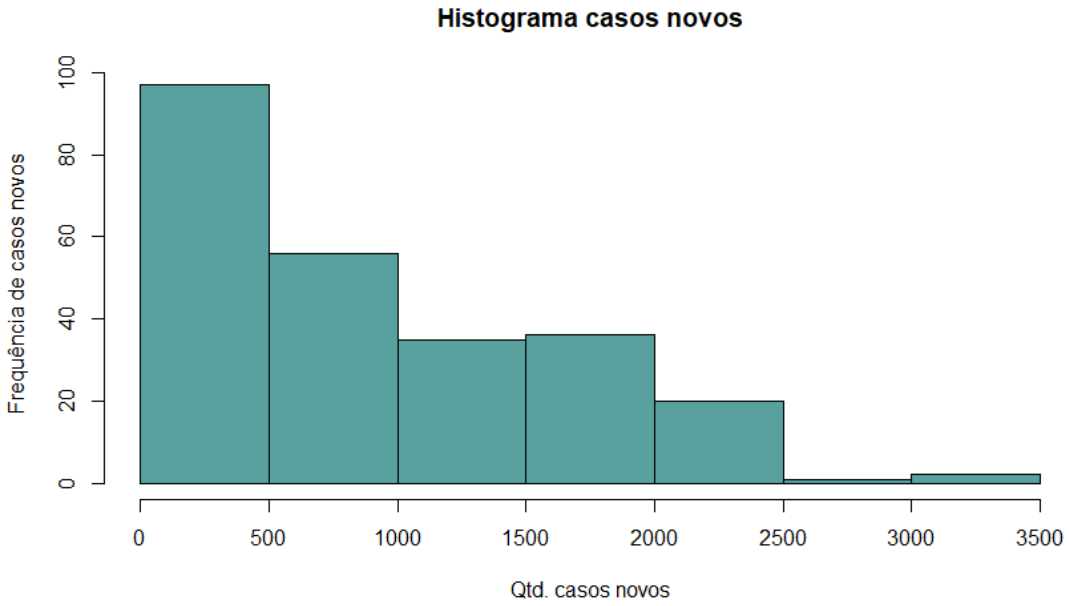


Figura 38: Histograma da frequência de casos novos por quantidade de enfermos novos

Sendo assim, pode-se considerar no território da capital brasileira um cenário ideal ao que diz sobre o gerenciamento contra o novo coronavírus. No período de coleta dos dados referentes à variável casos novos até o fim em 28 de Outubro de 2020, percebe-se que, ao longo de tais dias, a quantidade de 500 diagnósticos positivos para COVID-19 foi a mais frequente, e, quanto maior era a quantidade de casos novos em indivíduos residentes do DF, menor foi frequente de tal número.

Então, para a melhor compreensão do leitor, é possível, por meio de um gráfico de linhas, apresentar a distribuição de casos novos mostrando então a quantidade de casos novos por dia/mês:

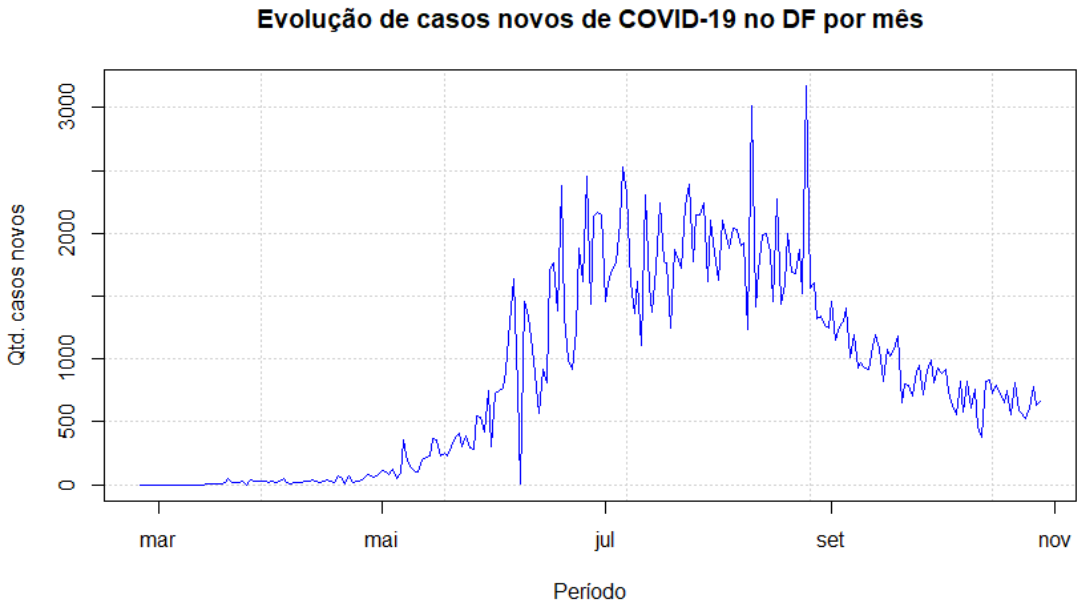


Figura 39: Casos novos de COVID-19 por mês no Distrito Federal

Avaliando o gráfico de linhas gerado e apresentado na figura 39 pode-se perceber que houvera grande sazonalidade nos dados a partir do mês de Maio. A partir de tal data então percebe-se uma grande crescente nos diagnósticos positivos para COVID-19, permanecendo entre 1.000 e 3.000 casos por dia entre os meses de Maio e o fim de Agosto, período esse que é referente ao pico da pandemia no Distrito Federal. Após o final do mês de Agosto é possível visualizar a redução de novos casos do vírus até o fim da coleta dos dados desse trabalho, que terminou em 28 de Outubro de 2020.

Com o uso do gráfico BoxPlot, será possível avaliar as estatísticas presentes em cada mês e como foi a distribuição de tais dados em determinado mês referente ao casos novos para a região autônoma Distrito Federal:

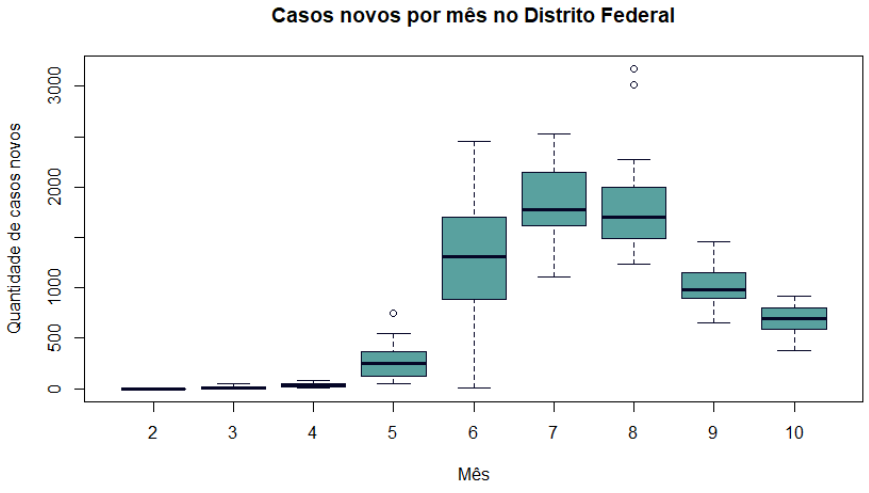


Figura 40: Boxplot - Casos novos por mês em numeral no Distrito Federal

```
> describeBy(covid_DF$casosNovos, month(covid_DF$data))

 Descriptive statistics by group
group: 2
   vars n mean sd median trimmed mad min max range skew kurtosis se
X1    1 5    0  0      0       0   0   0   0     0  NaN      NaN  0
------------------------------------------------------------------------------------------
group: 3
   vars  n  mean    sd median trimmed  mad min max range skew kurtosis   se
X1    1 31 10.71 13.52      4    8.68 5.93   0  45    45 0.99     -0.3 2.43
------------------------------------------------------------------------------------------
group: 4
   vars  n  mean   sd median trimmed   mad min max range skew kurtosis   se
X1    1 30 34.13 21.6     27   32.04 15.57   5  81    76 0.78    -0.68 3.94
------------------------------------------------------------------------------------------
group: 5
   vars  n  mean    sd median trimmed   mad min max range skew kurtosis    se
X1    1 31 271.7 160.1    251   254.4 177.9  50 752   702 0.88     0.67 28.75
------------------------------------------------------------------------------------------
group: 6
   vars  n mean    sd median trimmed   mad min  max range skew kurtosis    se
X1    1 30 1315 589.5   1310    1297 614.5   6 2455  2449 0.17    -0.66 107.6
------------------------------------------------------------------------------------------
group: 7
   vars  n mean    sd median trimmed mad  min  max range skew kurtosis    se
X1    1 31 1841 357.2   1773    1845 467 1109 2529  1420 0.02     -0.9 64.15
------------------------------------------------------------------------------------------
group: 8
   vars  n mean    sd median trimmed   mad  min  max range skew kurtosis    se
X1    1 31 1798 447.3   1700    1742 422.5 1233 3171  1938 1.32     2.02 80.33
------------------------------------------------------------------------------------------
group: 9
   vars  n mean    sd median trimmed   mad min  max range skew kurtosis    se
X1    1 30 1007 200.1  980.5     999 200.9 653 1456   803 0.34    -0.53 36.53
------------------------------------------------------------------------------------------
group: 10
   vars  n  mean    sd median trimmed   mad min max range  skew kurtosis    se
X1    1 28 685.6 132.6    698   689.9 166.1 377 915   538 -0.28    -0.77 25.06
```

Figura 41: Estatísticas descritivas sobre casos novos por mês em numeral (group)

Como evidenciado inicialmente na figura 39 e agora na figura 40 o período de pico da pandemia no Distrito Federal se estendeu de Maio à Agosto, a redução de novos casos se iniciou em Agosto, porém de forma mais conclusiva apenas a partir do mês de Setembro. Diferente da análise feita para o Brasil, em que mesmo com a redução dos casos por mês a mediana ainda era próxima do terceiro quartil, no Distrito Federal a mediana esteve sempre abaixo consideravelmente do terceiro quartil, evidenciando um bom gerenciamento da pandemia nesse distrito. Os picos de casos por dia aconteceram em Agosto, como foi explicado na figura 36, porém, ao visualizar a figura 40 percebe-se que foram dias excêntricos, pois, no mês de Agosto já se iniciava o fenômeno de redução dos casos.

Na figura 41 tem-se representado as estatísticas referente à cada mês em numeral do gráfico BoxPlot gerado na figura 40 para a distribuição dos dados de casos novos no Distrito Federal. Sendo assim, percebe-se que a partir do mês 5, isto é, o mês de Maio, a média de casos por dia vem crescendo por mês até o pico em Julho e Agosto, quando então, no mês de Setembro tal medida começa a reduzir ao longo dos próximos meses. No mês de Outubro percebe-se que o alcance, isto é, o número máximo de casos menos o número mínimo de casos foi o menor desde o mês de Maio, fenômeno parecido encontrado na medida de mediana desse mês. Ou seja, desde o mês de Maio 50% dos dias não obtiveram número baixos de diagnósticos positivos para o novo coronavírus como é visto no mês de Outubro, representando um fator extremamente positivo no combate ao vírus presente no Distrito Federal.

Os meses em que houveram maior número de casos foram Julho e Agosto com médias iguais à 1.841 e 1.798 casos novos, respectivamente. No mês de Julho o desvio padrão não foi tão alto se comparado à media, sendo equivalente à 352 casos para mais ou para menos dependendo do dia avaliado nesse mês. Em Julho, até metade dos dias desse mês contabilizaram números de diagnósticos para o novo coronavírus iguais ou menores à 1.773. Já o mês de agosto a alta na média de novos casos continuou, porém o desvio padrão

aumento para 447 casos positivos para mais ou para menos dependendo do dia de avaliação desse mês. Em Agosto até 50% dos dias obtiveram quantidade menores ou iguais à 1700 casos novos.

A fim de confirmar a tendência dos dados, do qual como explicado apresenta uma tendencia de queda, ou seja, um fato extremamente positivo em relação ao combate do vírus no Distrito Federal, tem-se a criação do seguinte gráfico do qual possui uma linha de tendência que tenta seguir a tendência da distribuição dos dados no período em que foi dado:

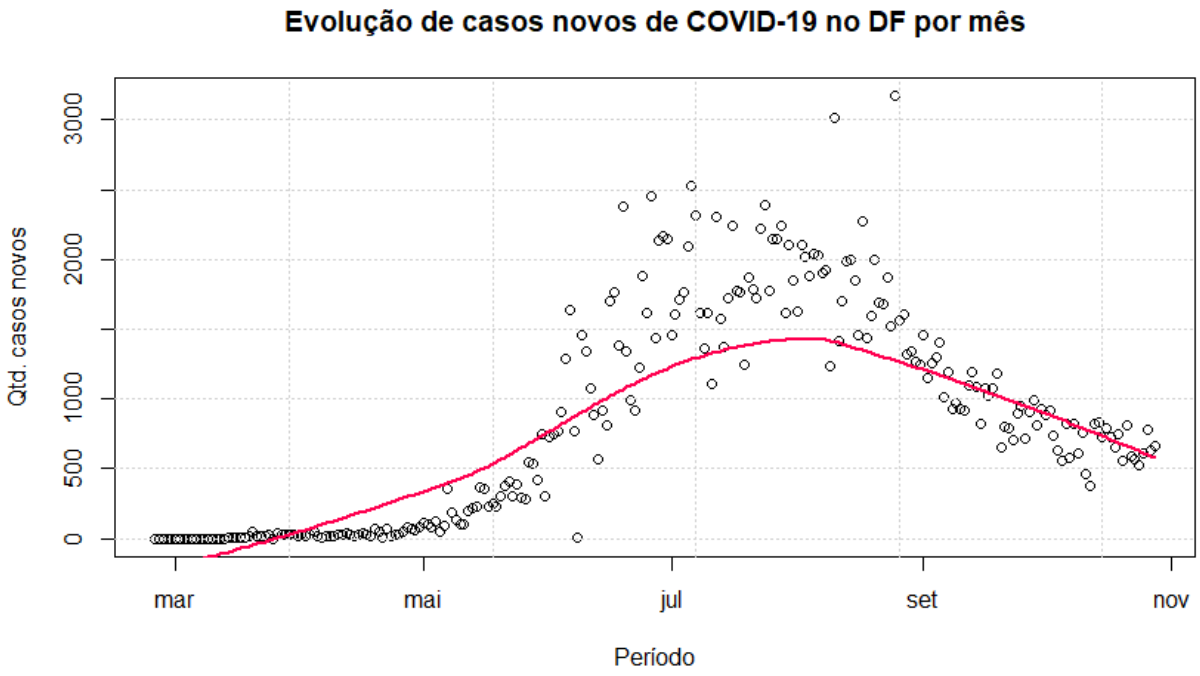


Figura 42: Linha de tendência casos novos por mês

Portanto, percebe-se no gráfico apresentado na figura 42 que a tendência dos dados referentes à quantidade de casos novos até o fim da coleta de tais dados em 28 de Outubro de 2020 no Distrito Federal, é de que realmente haja a redução com o passar dos meses, pois onde houve tendência de crescimento em relação à quantidade de diagnósticos positivos foram nos meses avaliados anteriormente, em Julho e Agosto.

5.2.2 Casos Acumulados

A variável **casosAcumulado** é uma variável apresentada por caráter acumulativo, isto é, com o passar dos dias foram-se somando os números anteriores, números esses referentes à quantidade de casos novos de COVID-19 no Distrito Federal até o dia 28 de Outubro de 2020. Sendo assim, a apresentação de estatísticas descritivas sobre os números encontrados aqui não se fazem justa, pois a única medida que importa é a quantidade presente no 100º quartil, isto é, a quantidade final somada desde o inicio da coleta dos dados no DF.

Dessa maneira, no Distrito Federal houveram ao todo 211.442 de indivíduos que testaram positivo para COVID-19, o que equivale dizer que, infelizmente 7% da população brasiliense contraiu o vírus até o dia 28 de Outubro de 2020, pois, de acordo com o levantamento do TCU, a população residente no DF é equivalente à 3.015.268 de habitantes. Por meio do gráfico de casos acumulados pode-se representar tal medida encontrada:

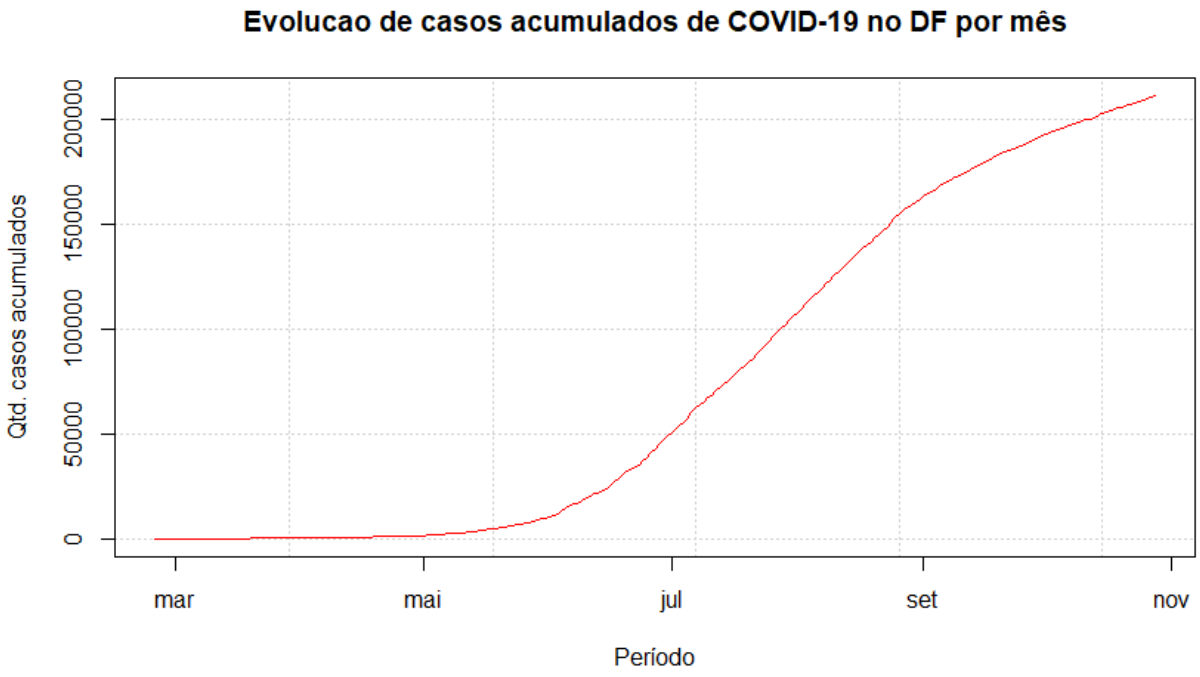


Figura 43: Casos acumulados de COVID-19 por mês no Distrito Federal

Ao se refletir sobre a representação de uma variável de natureza acumulativa, percebe-se que, sempre haverá o crescimento, e então para ser evidenciado uma redução dos casos, seria necessário que a linha do gráfico torna-se paralela ao eixo X, pois assim seria interpretado que os indivíduos não estão mais testando positivo para o novo coronavírus. Tal gráfico é representado na figura 43 e é de grande importância pois pode ajudar no gerenciamento da pandemia por parte dos gestores de saúde representantes do Distrito Federal.

5.2.3 Óbitos Novos

O Distrito Federal é o território onde se localiza a capital política do país Brasil, consequentemente, a renda per capita é alta, pois grande parte dos políticos decidem morar nesse distrito, elevando assim a renda per capita do local. Sendo assim, existem diversos hospitais que são referencias no Brasil, principalmente em Brasília, cominando em ajuda intensiva caso algum cidadão teste positivo ao vírus. Entretanto a maioria da população não é composta por políticos, e muitas vezes é residente da periferia, onde hospitais estão sucateados/lotados. Sendo assim, grande parte dos óbitos ocorridos nesse distrito são de pessoas dessas características.

```
> summary(covid_DF$obitosNovos)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0     1.0    11.0    14.8    24.5    79.0
```

Figura 44: Estatísticas óbitos novos

Assim, como está representado na tabela da figura 48 a média de óbitos novos por dia até o fim da coleta dos dados em 28 de Outubro de 2020 foi de 14 vitimas fatais diárias decorrentes do novo coronavírus, quantidade baixa caso comparado à outros estados. O máximo de óbitos por dia foi de 79 vítimas, quantidade essa que ocorreu no dia 6 de Agosto de 2020, como é representado na seguinte imagem:

```
> covid_DF$data[covid_DF$obitosNovos == 79]
[1] "2020-08-06"
```

Figura 45: Data referente ao maior número de óbitos novos

Avaliando a distribuição de densidade dos dados, tem-se o seguinte gráfico:

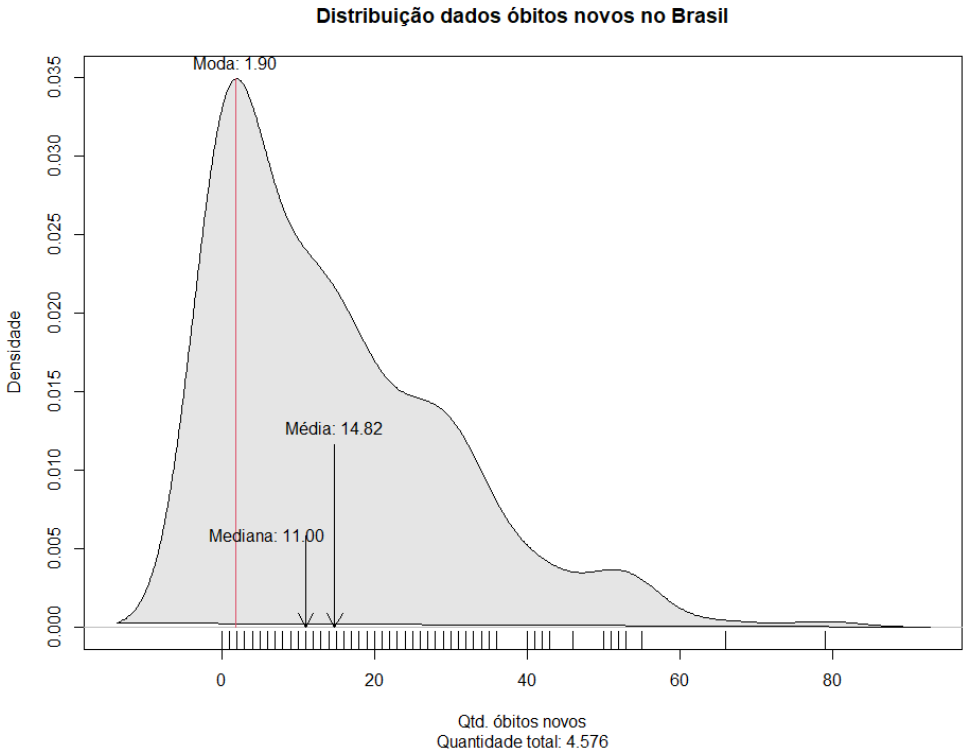


Figura 46: Data referente ao maior número de óbitos novos

Na figura 46 entende-se que a quantidade de óbitos novos é muito densa para quantidade pequena. Aplicando as funções skwness e kurtosis() presente no software R, será possível avaliar o índice de assimetria e grau de curtose da distribuição de tais dados. Portanto, o índice de assimetria de óbitos novos no Distrito Federal foi de 1.163, já a curtose foi igual à 1.133. De acordo com a assimetria encontrada, tem-se uma assimetria forte de acordo com a distribuição dos dados e possui curtose de natureza platicúrtica. Avaliando a moda, média e mediana, verifica-se uma assimetria à direita dos dados.

Portanto, os dados referentes à óbitos novos no Distrito Federal possuem uma distribuição assimétrica forte para a direita com curtose de natureza platicúrtica. Cenário ideal, pois a maior densidade de óbitos novos por dia se concentra em quantidades pequenas.

A fim de confirmar e concluir a análise feita, será avaliado por meio de um gráfico de histograma que mostra a frequência de quantidade de óbitos novos no DF. Por meio do histograma torna-se possível a análise da distribuição de frequências dos dados sobre óbitos novos no Distrito Federal até o fim da coleta dos dados em 28 de Outubro de 2020:

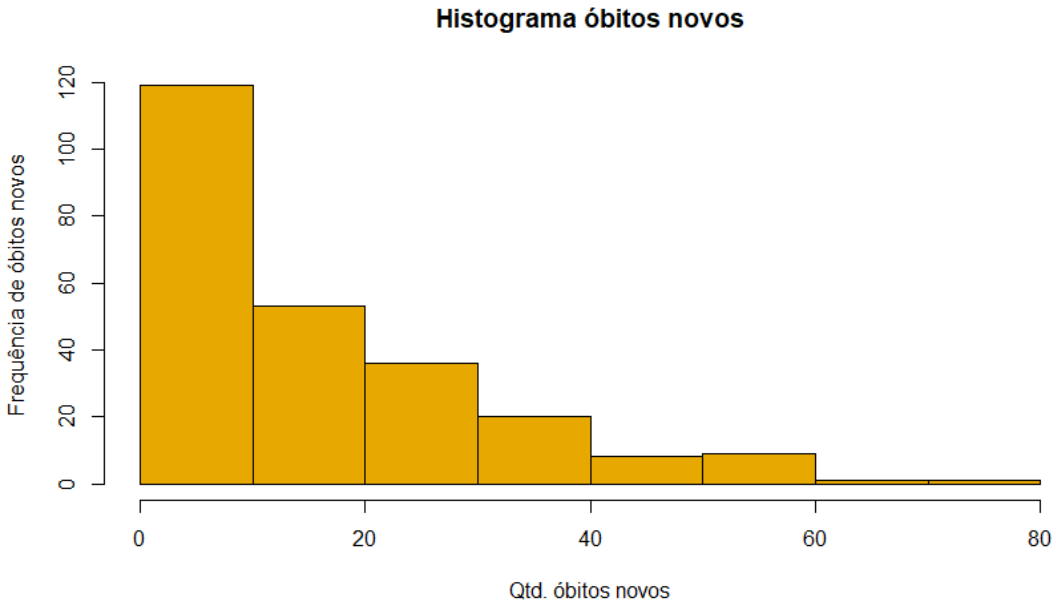


Figura 47: Histograma da frequência de óbitos novos por número de vítimas novas

De acordo com o histograma gerado e representado na figura 47, durante todos os dias da coleta de dados referentes à variável óbitos novos no DF, a frequência mais comum foi de 10 óbitos por dia, e, quanto maior foi o número de óbitos, a frequência de ocorrer durante o período da coleta dos dados foi menor, um fator positivo pois demonstra que a frequência de vítimas durante os dias da pandemia no DF não foram altos.

Portanto, a representação da quantidade de óbitos novos por COVID-19 em relação ao dia/mês pode ser representada por meio de um gráfico de linhas, como na seguinte maneira:

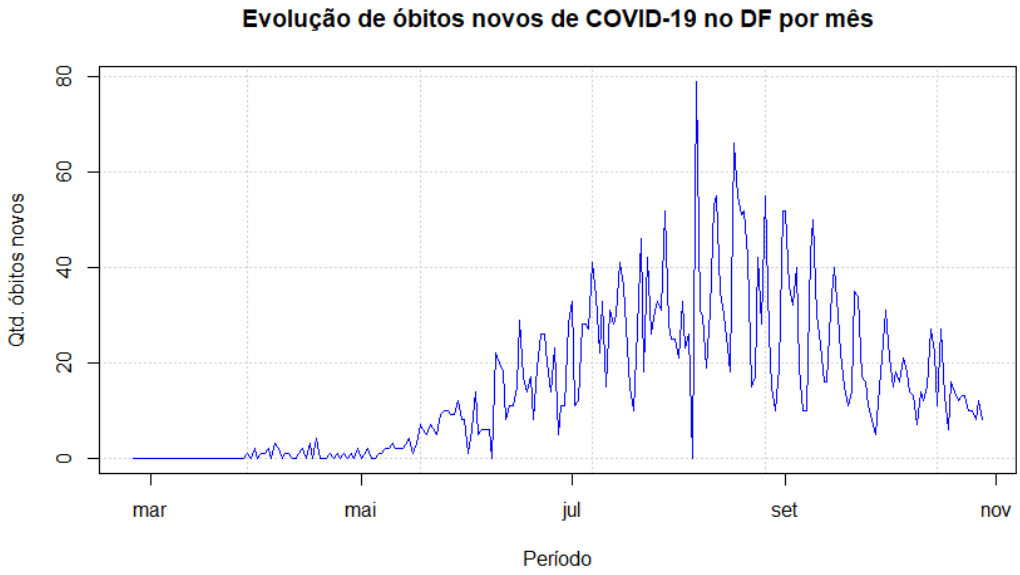


Figura 48: óbitos novos decorrente de COVID-19 por mês

No gráfico da figura 48 pode-se perceber um comportamento parecido com o gráfico da figura 39 do qual trata sobre a quantidade de casos novos por mês. O pico de mortes aconteceu entre o meses de Julho e Agosto, justamente onde ocorreu o pico de diagnósticos positivos para residentes do Distrito Federal. Sendo assim, pode-se refletir que há a possibilidade da variável Casos Novos ter correlação com a variável Óbitos Novos. Dessa maneira, por meio de uma matriz de correlação tem-se a correlação entre as duas variáveis referentes aos dados do Distrito Federal:

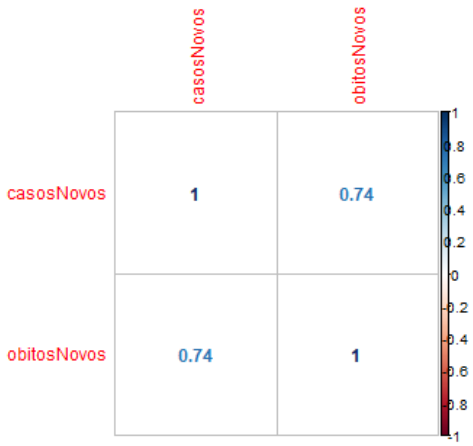


Figura 49: Matriz de correlação - óbitos novos e casos novos

Na matriz de correlação gerada apresentada na figura 49, conclui-se que as variáveis Casos Novos e Óbitos Novos possuem forte correlação positiva equivalente à 0.74, pois, de acordo com Karl Pearson, quanto mais próximo de 1 mais correlacionadas estão as variáveis. Sendo assim, entende-se que, quando há o aumento da quantidade de casos novos no DF, também aumentam-se a quantidade de óbitos novos. Então, tal análise

pode explicar a semelhança dos gráficos citados anteriormente para ambas variáveis.

Portanto, por meio de uma regressão linear simples pode-se concluir a análise de correlação feita anteriormente entre as variáveis casos novos e óbitos novos no DF. Então, a regressão será apresentada por meio de um gráfico de dispersão das variáveise uma linha de regressão:

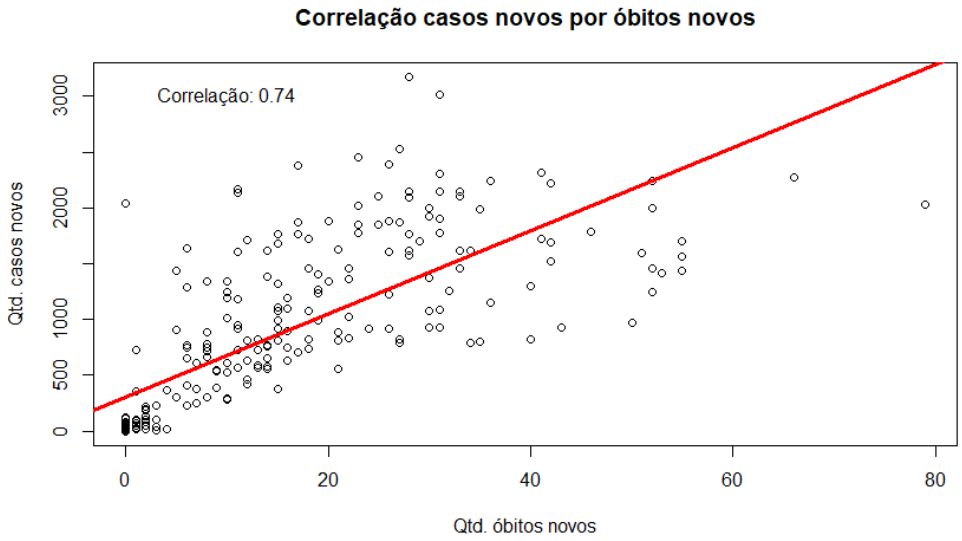


Figura 50: Regressão linear simples - óbitos novos e casos novos

Portanto, com o gráfico de dispersão e regressão linear simples presente na figura 50 percebe-se uma forte correlação entre as variáveis, da qual resultou em uma linha de regressão crescente positiva, concluindo assim a análise feita anteriormente que, quando há o aumento de indivíduos diagnosticas com COVID-19 no Distrito Federal, aumentam-se o número de vítimas novas decorrente do vírus.

Visando entender como os dados se distribuíram nos meses de coleta dos dados até o dia 28 de Outubro de 2020 no Distrito federal, gerou-se o seguinte gráfico BoxPlot:

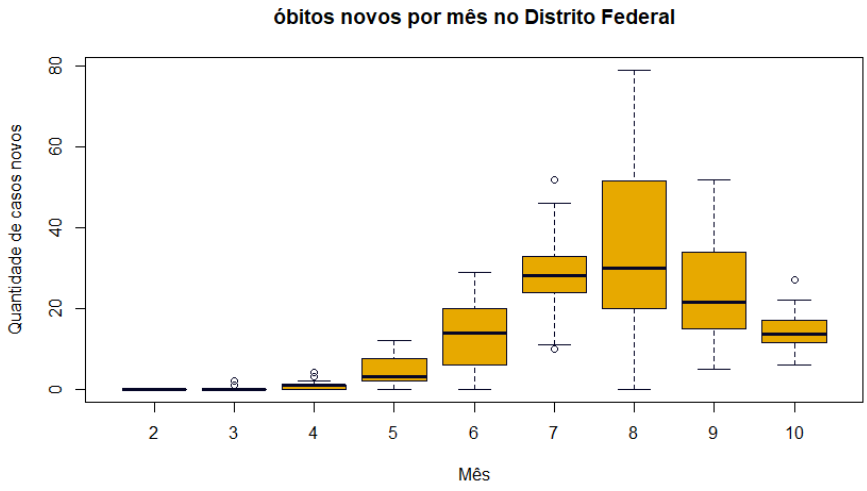


Figura 51: Boxplot - Óbitos novos por mês em numeral no Distrito Federal

```
> describeBy(covid_DF$obitosNovos, month(covid_DF$data))

 Descriptive statistics by group
group: 2
   vars n mean sd median trimmed mad min max range skew kurtosis se
X1    1 5    0  0      0       0   0   0   0     0  NaN      NaN  0
------------------------------------------------------------------------------------------
group: 3
   vars  n mean  sd median trimmed mad min max range skew kurtosis   se
X1    1 31  0.1 0.4      0       0   0   0   2     2 3.94    15.05 0.07
------------------------------------------------------------------------------------------
group: 4
   vars  n mean   sd median trimmed  mad min max range skew kurtosis  se
X1    1 30  0.9 1.09      1    0.71 1.48   0   4     4 1.11     0.41 0.2
------------------------------------------------------------------------------------------
group: 5
   vars  n mean   sd median trimmed  mad min max range skew kurtosis   se
X1    1 31 4.52 3.53      3    4.32 2.97   0  12    12 0.43     -1.2 0.63
------------------------------------------------------------------------------------------
group: 6
   vars  n mean   sd median trimmed mad min max range skew kurtosis   se
X1    1 30 13.9 8.04     14   13.67 8.9   0  29    29  0.2    -1.05 1.47
------------------------------------------------------------------------------------------
group: 7
   vars  n  mean    sd median trimmed  mad min max range skew kurtosis  se
X1    1 31 28.45 10.04     28   28.36 7.41  10  52    42  0.1    -0.29 1.8
------------------------------------------------------------------------------------------
group: 8
   vars  n  mean    sd median trimmed   mad min max range skew kurtosis   se
X1    1 31 33.94 18.14     30   33.08 17.79   0  79    79 0.48    -0.49 3.26
------------------------------------------------------------------------------------------
group: 9
   vars  n  mean    sd median trimmed   mad min max range skew kurtosis   se
X1    1 30 24.47 12.88   21.5   23.58 14.83   5  52    47 0.44    -0.94 2.35
------------------------------------------------------------------------------------------
group: 10
   vars  n mean   sd median trimmed  mad min max range skew kurtosis   se
X1    1 28 14.5 5.36   13.5   14.12 3.71   6  27    21 0.72    -0.03 1.01
```

Figura 52: Estatísticas descritivas sobre óbitos novos por mês em numeral (group)

De acordo com o BoxPlot gerado e representado na figura 51 os meses que houveram picos de novas vítimas fatais decorrentes do novo coronavírus foram realmente em Julho e Agosto, sendo Agosto um mês com dados extremamente péssimos em relação ao combate contra o vírus no Distrito Federal. A redução do número de mortes ocorre apenas a partir do mês de Setembro e em Outubro a queda já é bastante significativa, fator positivo.

Na figura 52 tem-se as estatísticas sobre cada mês referente à variável óbitos novos no Distrito Federal. Sendo assim, é possível confirmar que, os meses de Julho e Agosto foram os que obtiveram maior média de mortes quando levado em consideração os dados referente aos dias de tais meses. No dias que compõe o mês de Julho, até 50% desses obtiveram a quantidade menor ou igual à 28 óbitos por dia, já Agosto obteve metade de seus dias com a quantidade de vítimas fatais menor ou igual à 30 óbitos. Nos dias de Julho a menor quantidade de óbitos foi de 10 óbitos e o máximo foi igual à 52 óbitos, porém em agosto houve um dia que o menor número de óbitos foi igual à 0, número que não faz sentido após as análises efetuadas sendo possível considerar tal número como uma falha no sistema de coleta ou divulgação dos dados.

A queda da quantidade de óbitos por mês só se confirma em Outubro, quando a média de óbitos novos foi apenas de 14 vítimas, média essa que não era contabilizada desde o mês de Junho. Em Outubro houvera um dia que o mínimo de vitimas foi igual à apenas 6 e o máximo igual à 27 óbitos, número máximo que novamente não era evidenciado desde o mês de Junho.

Portanto, para confirmar a tendência dos dados sobre óbitos novos referente ao Distrito Federal, gerou-se um gráfico de dispersão contendo uma linha de tendência que tentar seguir o sentido dos dados, como é visto na seguinte imagem:

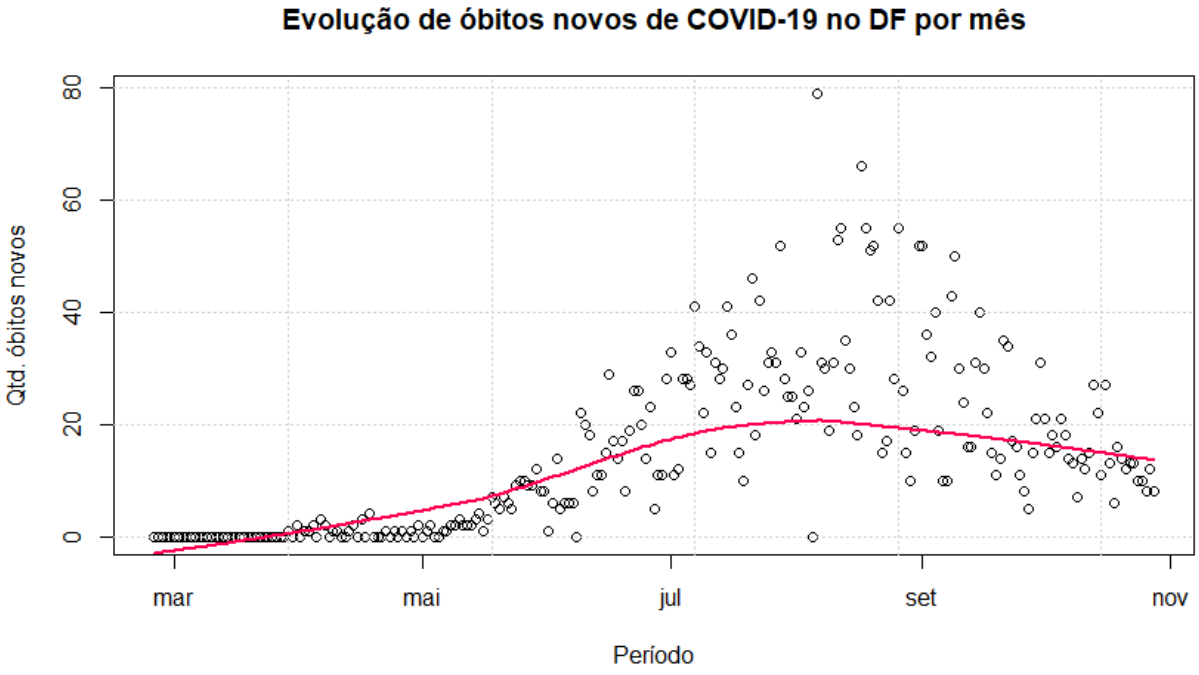


Figura 53: Linha de tendência óbitos por mês

Então, de acordo com o gráfico da figura 53 confirma-se que, existe a tendência da redução da quantidade de óbitos ocasionados por COVID-19 ao longo dos meses porém não de forma acentuada, fenômeno que se explica pois, o Distrito Federal não possuiu número elevado de mortes como em outros estados, e quando possuiu foram apenas em dois meses.

5.2.4 Óbitos Acumulados

A variável **óbitos acumulados** é uma variável que se desenvolve por meio da soma das quantidades dos dias anteriores observados em óbitos novos, sendo assim, é uma variável acumulativa dos dados referentes ao Distrito Federal dos quais foram coletados até o dia 28 de Outubro de 2020. Sendo assim, não é válida a aplicação e exploração de medidas de estatística descritiva, importando apenas o valor final da soma do qual foi acumulado, ou seja, o valor do 100º percentil.

Sendo assim, o valor total de óbitos no Distrito Federal até o dia 28 de Outubro de 2020 foi de 3.661 vítimas fatais, o que, por meio do gráfico pode-se ter ideia do crescimento da curva ao longo dos meses da coleta de tais dados:

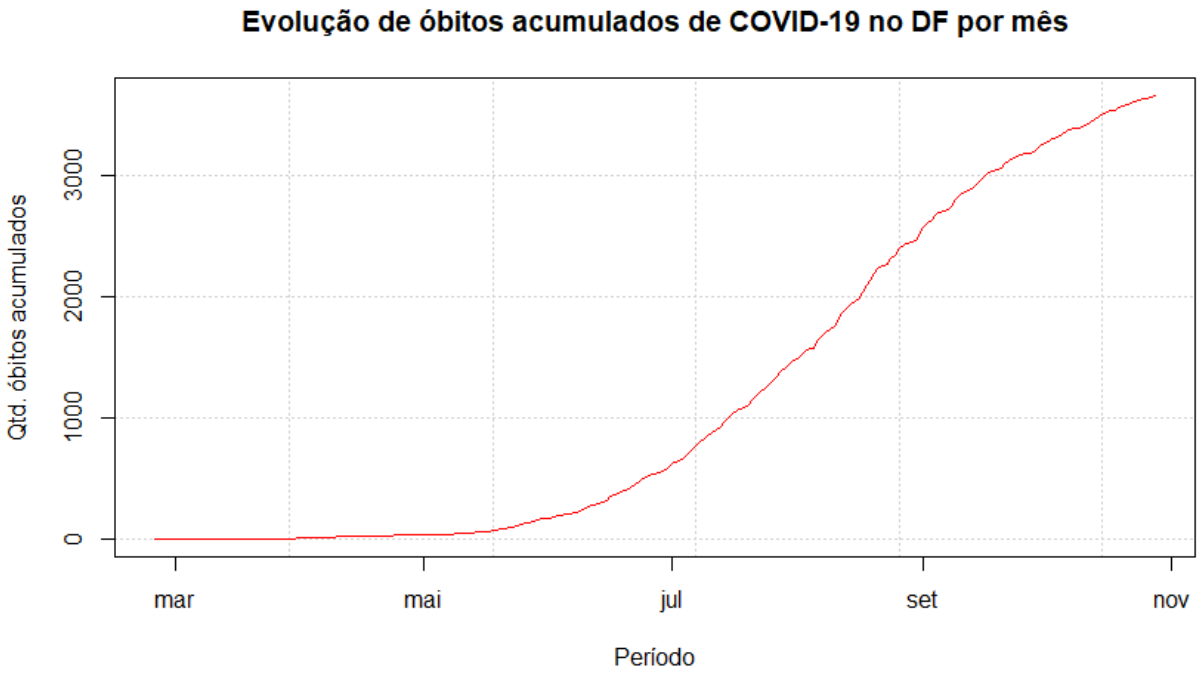


Figura 54: Óbitos acumulados de COVID-19 por mês no Distrito Federal

Como pode ser interpretado do gráfico gerado e presenta na figura 54, a curva de óbitos acumulados sempre será crescente ou continua, pois se trata da soma de óbitos novos, sendo assim, para concluir uma queda do número de óbitos, basta a linha estar paralela ao eixo X pois não seriam somados novos valores a serem acumulados.

5.3 Variáveis categóricas

O método skim presenta na biblioteca R nos retorna estatísticas sobre cada tipo de variável. Dessa maneira, até então foram exploradas as variáveis numéricas: Casos Novos, Casos Acumulados, Óbitos Novos e Óbitos Acumulados. Dendo assim, aplicou-se tal método para encontrar estatísticas sobre os dados categóricos em relação ao Distrito Federal (Dados retornados após o filtro, na base de dados COVID19.SAS, para o estado igual à "DF"como explicado na seção 2:

```
-- Variable type: character -----------------------------------------------------
# A tibble: 4 x 8
  skim_variable   n_missing complete_rate   min   max empty n_unique whitespace
* <chr>               <int>         <dbl> <int> <int> <int>    <int>      <int>
1 regiao                  0             1    12    12     0        1          0
2 estado                  0             1     2     2     0        1          0
3 municipio               0             1     0     0   247        1          0
4 nomeRegiaoSaude         0             1     0     0   247        1          0
```

Figura 55: Propriedades variáveis categóricas dos dados referentes ao Distrito Federal

Portanto, analisando a figura **??**, a base dos dados sobre COVID-19 referentes ao Brasil, possui 4 variáveis categóricas, sendo essas: regiao, estado, municipio e nomeRegSaude, variáveis essas das quais já foram explicadas anteriormente na seção 34. Porém, é válido ressaltar, como explicado anteriormente há a existência de variáveis numéricas presente na base dados da qual foram entendidas como numéricas, porém, possuem caráter categórico, sendo essas: coduf, codmun, codRegiaoSaude, semanaEpi e interior/metropolitana. Sendo assim, as variáveis categóricas referentes ao Brasil são: **regiao, estado, municipio, nomeRegSaude, coduf, codmun, codRegiaoSaude, semanaEpi e interior/metropolitana**.

Portanto, na figura 56 percebe-se que não existem nenhum dado faltante nas variáveis de natureza categórica presente na base de dados filtrada para o Distrito Federal. O número de região foi único, sendo representado obviamente pelo número referente ao Centro-Oeste. Entretanto percebe-se que as variáveis municipio e nomeRegiaoSaude não possuíram dados, isto é, estão vazias nas 247 linhas, fato esse que pode ser explicado pois o DF não possui municípios.

Em relação às demais variáveis que foram entendidas como numéricas porém possuem caráter categórico puderam ser analisadas na seção 5.2. Entretanto as variáveis coduf, codmun e codRegiaoSaude possuíram valores constantes ou não foram aplicados valores à tais, como em codmun e codRegiaoSaude. Já em relação á semanaEpi, entende-se que é a semana do ano porém é uma variável atrativas para a efetuação de análises de séries temporais, o que não será abordado neste trabalho. Dessa maneira, para tornar a visualização de tais variáveis categóricas mais simplória, tem-se a representação da tabela aplicada no software RStudio:

	regiao	estado	municipio	coduf	codmun	codRegiaoSaude	nomeRegiaoSaude	semanaEpi	interior/metropolitana
1	Centro-Oeste	DF		53	NA	NA		9	NA
2	Centro-Oeste	DF		53	NA	NA		9	NA
3	Centro-Oeste	DF		53	NA	NA		9	NA
4	Centro-Oeste	DF		53	NA	NA		9	NA
5	Centro-Oeste	DF		53	NA	NA		9	NA
6	Centro-Oeste	DF		53	NA	NA		10	NA
7	Centro-Oeste	DF		53	NA	NA		10	NA
8	Centro-Oeste	DF		53	NA	NA		10	NA
9	Centro-Oeste	DF		53	NA	NA		10	NA
10	Centro-Oeste	DF		53	NA	NA		10	NA
11	Centro-Oeste	DF		53	NA	NA		10	NA
12	Centro-Oeste	DF		53	NA	NA		10	NA
13	Centro-Oeste	DF		53	NA	NA		11	NA
14	Centro-Oeste	DF		53	NA	NA		11	NA
15	Centro-Oeste	DF		53	NA	NA		11	NA
16	Centro-Oeste	DF		53	NA	NA		11	NA
17	Centro-Oeste	DF		53	NA	NA		11	NA
18	Centro-Oeste	DF		53	NA	NA		11	NA

Figura 56: Visualização variáveis categóricas Distrito Federal

6 Taxas e Coeficientes

Neste tópico será apresentado os conceitos explicados anteriormente na seção 3.2 referentes ao país Brasil e ao território autônomo do Distrito Federal.

6.1 Brasil

Por meio das taxas e coeficiente calculados sobre os dados do Brasil até o dia 28 de Outubro de 2020, pode-se avaliar e analisar diversos fatores que podem acender alertas sobre o gerenciamento da pandemia no país.

6.1.1 Coeficiente de Incidência de COVID-19

De acordo com os dados coletados e divulgados referentes ao território do país Brasil, a taxa de incidência do novo coronavírus até o dia 28 de Outubro de 2020 foi igual à 2.602 por 100 mil habitantes. A incidência mede o risco ou probabilidade de ocorrer a doença na população exposta. De acordo com o calculo efetuado, até o dia 28 de Outubro de 2020, a cada 100 mil habitantes existe a probabilidade de que 2.602 deles contraíam a COVID-19.

6.1.2 Coeficiente de Mortalidade de COVID-19

O coeficiente de mortalidade apresentado no território brasileiro até o dia 28 de Outubro de 2020 foi igual à 75.4. Dessa maneira, em uma população de 100 mil habitantes há o risco de 75 óbitos decorrentes de COVID-19 no Brasil.

6.1.3 Taxa de letalidade de COVID-19

A taxa de letalidade do novo coronavírus no Brasil em relação aos dados coletados até o dia 28 de Outubro de 2020 foi de 2.898% óbitos a cada 100 diagnósticos novos positivos de brasileiros.

6.2 Distrito Federal

A fim de analisar os coeficientes e taxas decorrentes da pandemia do novo coronavírus no Distrito Federal, onde está localizada a capital do Brasil, fez-se este levantamento.

6.2.1 Coeficiente de Incidência de COVID-19

De acordo com os dados coletados e divulgados referentes ao território do país Brasil, a taxa de incidência do novo coronavírus até o dia 28 de Outubro de 2020 foi igual à 7.012 por 100 mil habitantes.A incidência mede o risco ou probabilidade de ocorrer a doença na população exposta. De acordo com o calculo efetuado, até o dia 28 de Outubro de 2020, a cada 100 mil habitantes existe a probabilidade de que 7.012 deles contraíam a COVID-19. Número três vezes mais elevado do que se comparado com o coeficiente de incidência geral do Brasil.

6.2.2 Coeficiente de Mortalidade de COVID-19

O coeficiente de mortalidade apresentado no território brasileiro até o dia 28 de Outubro de 2020 foi igual à 121.42. Dessa maneira, em uma população de 100 mil habitantes há o risco de 121 óbitos decorrentes de COVID-19 no Brasil. Percebe-se que no estado do Distrito Federal os enfermos morrem mais do que o coeficiente de mortalidade geral do Brasil.

6.2.3 Taxa de letalidade de COVID-19

A taxa de letalidade do novo coronavírus no Brasil em relação aos dados coletados até o dia 28 de Outubro de 2020 foi de 1.731% óbitos a cada 100 diagnósticos novos positivos de brasileiros.

7 Conclusão

Então, conclui-se após a análise feita sobre os dados gerados pelo novo coronavírus no Brasil, da qual discorreu na seção 4, que o Brasil alcançou o pico da pandemia nos meses de Julho e Agosto, diminuindo o número de casos e óbitos após o mês de Agosto porém ainda com mediana elevada para as variáveis impactantes na área de saúde.

No Brasil a procura por atendimento médico do individuo contaminada com o vírus se mostrou um fator muito importante, pois, aproximadamente 90% dos brasileiros infectados conseguiu vencer o vírus e tornar-se saudável novamente. Entretanto no período da coleta dos dados, o número de casos acumulado e óbitos acumulado foi extremamente grande, situação essa que têm ligação com a má gestão hospitalar para periferias e para a população carente. Dos dados coletados até o dia 28 de Outubro, foi apresentado que a pandemia em território brasileiro está menos intensa, porém tal fator deve ser levado em conta pois a população pode acabar relaxando em relação aos cuidados que antes tivera, acarretando em uma segunda onda.

No Brasil as taxas e coeficientes que se desenvolveram até o dia 28 de Outubro como preocupantes porém controladas se for levado em consideração a totalidade da população brasileira, entretanto deve-se ser dada a devida atenção pelas autoridades de saúde.

Já o Distrito Federal se mostrou um dos territórios que compõe o país Brasil que foi bastante impactado pelo vírus, principalmente na região do Centro-Oeste. Os dados sobre indivíduos novos infectados no DF se mostraram em uma crescente intensa, que obteve picos no mesmos meses entre Junho e Julho.

No território autônomo Distrito Federal, a pandemia parecia estar menos intensa após o mês de Agosto, porém avaliando os dados diários sobre a quantidade casos novos, percebeu-se que, no final do mês de Outubro começaram a crescer novamente, mesmo que de forma mais tímida, fator esse que novamente pode ser explicado pelo relaxamento da população em relação aos cuidados contra o vírus, o que deve ser reavaliada pela equipe de gestão de saúde para conter tal crescimento, pois como demonstrado está correlacionado com o crescimento de óbitos novos.

O Distrito Federal possui um território bastante povoado, aproximadamente 3.015.268 de habitantes de acordo com o TCU. Sendo assim, as taxas e coeficientes em relação ao vírus se mostraram bastantes preocupantes, o que deve ser levado em consideração no planejamento de gerenciamento da saúde contra o novo coronavírus.

Não foi possível efetuar análises sobre dados referentes à pacientes em tratamento ou recuperados no território do Distrito Federal, dados esses que não foram disponibilizados ou que até mesmo podem não ter sido coletados, um fator extremamente negativa uma vez que são números extremamente vantajosos em planos contra o vírus no Distrito Federal.

8 Anexo código em R utilizado

```
rm(list=ls())
#Bibliotecas necessarias
pacman::p_load(pacman, dplyr, tidyr, tidyverse, haven, lubridate,
               psych, skimr, corrplot, ggplot2, e1071)

#Lendo arquivos em formato SAS
covid19 <- read_sas("~/Dados/Covid-19/covid19.sas7bdat", NULL)
covid_serie <- read_sas("~/Dados/Covid-19/covid19_serie.sas7bdat", NULL)

#Filtrando os arquivos para: Brasil e DF
covid_DF <- covid_serie[covid_serie$estado == "DF" & !is.na(covid_serie$estado),]
covid_Br <- covid19[covid19$regiao == "Brasil" & !is.na(covid19$estado),]


#Arrumando Erro de notacao cientifica nos graficos
options("scipen"=100, "digits"=4)


############################ MES ############################
####################### BRASIL ##########################

#Analise Exploratoria rapida Brasil
skim(covid_Br)
summary(covid_Br)

#(Conferindo se o filtro para o Brasil possui as mesmas analises caso usasse a base toda)
skim(covid19)
skim(covid_serie)

############################ DISTRITO FEDERAL ##############################
#Analise Exploratoria rapida Distrito Federal
skim(covid_DF)
summary(covid_DF)

############################ BRASIL ##############################
summary(covid_Br$casosNovos)
covid_Br$data[covid_Br$casosNovos == 69074]
covid_Br$data[covid_Br$obitosNovos == 1595]
View(covid_Br[,c(1,2,3,4,5,6,7,9,17)])


#Conjunto de graficos Brasil
#Histograma BRASIL CASOS
hist(covid_Br$casosNovos, col="#777799",
     main = "Histograma casos novos", xlab = "Qtd. casos novos",
     ylab = "Frequência de casos novos")

#indice de assimetria e curtose
skewness(covid_Br$casosNovos)
kurtosis(covid_Br$casosNovos)
```

```
#LinePlot BRASIL CASOS
par(mfrow = c(1, 1))
plot(covid_Br$data, covid_Br$casosNovos,
     main = "Evolução de casos novos de COVID-19 no Brasil por mês",
     ylab = "Qtd. casos novos", xlab = "Período",
     type = "l", col = "blue", panel.first = grid())

plot(covid_Br$data, covid_Br$casosAcumulado,
     main = "Evolução de casos acumulados de COVID-19 no Brasil por mês",
     ylab = "Qtd. casos acumulados", xlab = "Período",
     type = "l", col = "red", panel.first = grid())


#Linha tendencia suave BRASIL CASOS
plot(covid_Br$casosNovos ~ covid_Br$data,
     data = covid_Br,
     xlab = "Período",
     ylab = "Qtd. casos novos",
     main = "Evolução de casos novos de COVID-19 no Brasil por mês")
with(covid_Br, {
  lines(lowess(x = covid_Br$data, y = covid_Br$casosNovos),
        lwd = 2,
        col = "#ff0050", panel.first = grid())})

plot(covid_Br$casosAcumulado ~ covid_Br$data,
     data = covid_Br,
     xlab = "Período",
     ylab = "Qtd. casos acumulados",
     main = "Evolução de casos acumulados de COVID-19 no Brasil por mês")
with(covid_Br, {
  lines(lowess(x = covid_Br$data, y = covid_Br$casosAcumulado),
        lwd = 2,
        col = "#ff0050", panel.first = grid())})
par(mfrow = c(1, 1))

#LinePlot BRASIL OBITOS
skewness(covid_Br$obitosNovos)
kurtosis(covid_Br$obitosNovos)

hist(covid_Br$obitosNovos, col = "#ff0050", ylab = "Frequência de obitos novos",
     xlab = "Número de óbitos novos", main = "Histograma obitos novos")


plot(covid_Br$data, covid_Br$obitosNovos,
     main = "Evolução de óbitos novos de COVID-19 no Brasil por mês",
     ylab = "Qtd. óbitos novos", xlab = "Período",
     type = "l", col = "blue", panel.first = grid())

plot(covid_Br$data, covid_Br$obitosAcumulado,
     main = "Evolução de óbitos acumulados de COVID-19 no Brasil por mês",
     ylab = "Qtd. óbitos acumulados", xlab = "Período",
     type = "l", col = "red", panel.first = grid())
par(mfrow=c(1, 1))
```

```
#Linha tendencia suave BRASIL OBITOS
plot(covid_Br$obitosNovos ~ covid_Br$data,
     data = covid_Br,
     xlab = "Período",
     ylab = "Qtd. óbitos novos",
     main = "Evolução de óbitos novos de COVID-19 no Brasil por mês")
with(covid_Br, {
  lines(lowess(x = covid_Br$data, y = covid_Br$obitosNovos),
        lwd = 2,
        col = "#ff0050", panel.first = grid())})

plot(covid_Br$obitosAcumulado ~ covid_Br$data,
     data = covid_Br,
     xlab = "Período",
     ylab = "Qtd. óbitos acumulados",
     main = "Evolução de óbitos acumulados de COVID-19 no Brasil por mês")
with(covid_Br, {
  lines(lowess(x = covid_Br$data, y = covid_Br$obitosAcumulado),
        lwd = 2,
        col = "#ff0050", panel.first = grid())})
par(mfrow = c(1, 1))

#Em acompanhamento BR
skim(covid_Br$emAcompanhamentoNovos)
summary(covid_Br$emAcompanhamentoNovos)
plot(covid_Br$emAcompanhamentoNovos ~ covid_Br$data,
     main = "Evolução de pacientes em acompanhamentono Brasil por mês",
     ylab = "Qtd. em acompanhamento", xlab = "Período",
     type = "l", col = "blue", panel.first = grid())

#Recuperado BR
skim(covid_Br$Recuperadosnovos)
plot(covid_Br$Recuperadosnovos ~ covid_Br$data,
     main = "Evolução de pacientes recuperados Brasil por mês",
     ylab = "Qtd. recuperados", xlab = "Período",
     type = "l", col = "blue", panel.first = grid())


############################ DF ##############################
#Conjunto de graficos DF
View(covid_DF[,c(1,2,3,4,5,6,7,9, 16)])
summary(covid_DF$obitosNovos)
summary(covid_DF$casosNovos)
covid_DF$data[covid_DF$casosNovos == 3171]
covid_DF$data[covid_DF$obitosNovos == 79]

#Indice de assimetria e curtose
skewness(covid_DF$casosNovos)
kurtosis(covid_DF$casosNovos)

skewness(covid_DF$obitosNovos)
kurtosis(covid_DF$obitosNovos)

#Histogramas DF
```

```
hist(covid_DF$casosNovos, col = "#59a19f",
     main = "Histograma casos novos", xlab = "Qtd. casos novos",
     ylab = "Frequência de casos novos")

hist(covid_DF$obitosNovos, col = "#e7a900",
     main = "Histograma óbitos novos", xlab = "Qtd. óbitos novos",
     ylab = "Frequência de óbitos novos")

#LinePlot DF CASOS
plot(covid_DF$data, covid_DF$casosNovos,
     main = "Evolução de casos novos de COVID-19 no DF por mês",
     ylab = "Qtd. casos novos", xlab = "Período",
     type = "l", col = "blue", panel.first = grid())

plot(covid_DF$data, covid_DF$casosAcumulado,
     main = "Evolucao de casos acumulados de COVID-19 no DF por mês",
     ylab = "Qtd. casos acumulados", xlab = "Período",
     type = "l", col = "red", panel.first = grid())
par(mfrow=c(1, 1))

#Linha tendencia suave DF CASOS
plot(covid_DF$casosNovos ~ covid_DF$data,
     data = covid_DF,
     xlab = "Período",
     ylab = "Qtd. casos novos",
     main = "Evolução de casos novos de COVID-19 no DF por mês")
with(covid_DF, {
  lines(lowess(x = covid_DF$data, y = covid_DF$casosNovos),
        lwd = 2,
        col = "#ff0050", panel.first = grid())})

plot(covid_DF$casosAcumulado ~ covid_DF$data,
     data = covid_DF,
     xlab = "Período",
     ylab = "Qtd. casos acumulados",
     main = "Evolução de casos acumulados de COVID-19 no DF por mês")
with(covid_DF, {
  lines(lowess(x = covid_DF$data, y = covid_DF$casosAcumulado),
        lwd = 2,
        col = "#ff0050", panel.first = grid())})
par(mfrow = c(1, 1))


#LinePlot DF OBITOS
skewness(covid_DF$obitosNovos)
kurtosis(covid_DF$obitosNovos)

plot(covid_DF$data, covid_DF$obitosNovos,
     main = "Evolução de óbitos novos de COVID-19 no DF por mês",
     ylab = "Qtd. óbitos novos", xlab = "Período",
     type = "l", col = "blue", panel.first = grid())

plot(covid_DF$data, covid_DF$obitosAcumulado,
     main = "Evolução de óbitos acumulados de COVID-19 no DF por mês",
```

```
     ylab = "Qtd. óbitos acumulados", xlab = "Período",
     type = "l", col = "red", panel.first = grid())
par(mfrow=c(1, 1))

#Linha tendencia suave DF OBITOS
par(mfrow = c(2, 1))
plot(covid_DF$obitosNovos ~ covid_DF$data,
     data = covid_DF,
     xlab = "Período",
     ylab = "Qtd. óbitos novos",
     main = "Evolução de óbitos novos de COVID-19 no DF por mês")
with(covid_DF, {
  lines(lowess(x = covid_DF$data, y = covid_DF$obitosNovos),
        lwd = 2,
        col = "#ff0050", panel.first = grid())})

plot(covid_DF$obitosAcumulado ~ covid_DF$data,
     data = covid_DF,
     xlab = "Período",
     ylab = "Qtd. óbitos acumulados",
     main = "Evolução de óbitos acumulados de COVID-19 no DF por mês")
with(covid_DF, {
  lines(lowess(x = covid_DF$data, y = covid_DF$obitosAcumulado),
        lwd = 2,
        col = "#ff0050", panel.first = grid())})
par(mfrow = c(1, 1))



############################################################
#Boxplot Casos novos meses Brasil
tapply(covid_Br$casosNovos, month(covid_Br$data), summary)
describeBy(covid_Br$casosNovos, month(covid_Br$data))

boxplot(covid_Br$casosNovos ~ month(covid_Br$data),
        main="Casos novos por mês no Brasil",
        xlab="Mês",
        ylab="Quantidade de casos novos",
        col="#777799",
)

#Boxplot Obitos novos meses Brasil
tapply(covid_Br$obitosNovos, month(covid_Br$data), summary)
describeBy(covid_Br$obitosNovos, month(covid_Br$data))

boxplot(covid_Br$obitosNovos ~ month(covid_Br$data),
        main="Óbitos novos por mês no Brasil",
        xlab="Mês",
        ylab="Quantidade de casos novos",
        col="#880055",
)

#Boxplot Casos novos meses DF
tapply(covid_DF$casosNovos, month(covid_DF$data), summary)
```

```
describeBy(covid_DF$casosNovos, month(covid_DF$data))

boxplot(covid_DF$casosNovos ~ month(covid_DF$data),
        main="Casos novos por mês no Distrito Federal",
        xlab="Mês",
        ylab="Quantidade de casos novos",
        col="#59a19f",
        border="#050627",
)

#Boxplot Obitos novos meses DF
tapply(covid_DF$obitosNovos, month(covid_DF$data), summary)
describeBy(covid_DF$obitosNovos, month(covid_DF$data))

boxplot(covid_DF$obitosNovos ~ month(covid_DF$data),
        main="óbitos novos por mês no Distrito Federal",
        xlab="Mês",
        ylab="Quantidade de casos novos",
        col="#e7a900",
        border="#050627",
)

########## Regressoes e correlacoes ##########
#Regressao BR
plot(covid_Br$casosNovos ~ covid_Br$obitosNovos, xlab = "Qtd. óbitos novos",
     ylab = "Qtd. casos novos", main = "Correlação casos novos por óbitos novos")
abline(lm(covid_Br$casosNovos ~ covid_Br$obitosNovos), col = "red", lwd = 3)
text(paste("Correlação:", round(cor(covid_Br$casosNovos, covid_Br$obitosNovos), 2)),
     x = 400, y = 60000)

#Regressao DF
plot(covid_DF$casosNovos ~ covid_DF$obitosNovos, xlab = "Qtd. óbitos novos",
     ylab = "Qtd. casos novos", main = "Correlação casos novos por óbitos novos")
abline(lm(covid_DF$casosNovos ~ covid_DF$obitosNovos), col = "red", lwd = 3)
text(paste("Correlação:", round(cor(covid_DF$casosNovos, covid_DF$obitosNovos), 2)),
     x = 10, y = 3000)

#Correlacao BR
coree <- covid_Br[,c(12,14)]
N<-cor(coree)
head(round(N,2))
png("correlacao_numeroBR.PNG")
corrplot(N, method= "number")
corrplot

#Correlacao DF
coree1 <- covid_DF[,c(11,13)]
M<-cor(coree1)
head(round(M,2))
png("correlacao_numeroDF.PNG")
corrplot(M, method= "number")
corrplot

############################REGIAO########################################
```

```
ggplot(covid_serie, aes(y = casosNovos, x = regiao, fill = regiao)) +
  geom_bar(stat = "identity", width = .75) +
  ggtitle("Qtd. casos novos por região") +
  labs(x = "Região", y = "Qtd. casos novos")

ggplot(covid_serie, aes(y = casosNovos, x = estado, fill = regiao)) +
  geom_bar(stat = "identity", width = .75) +
  ggtitle("Qtd. casos novos por estado") +
  labs(x = "Estado", y = "Qtd. casos novos")


ggplot(covid_serie, aes(y = obitosNovos, x = regiao, fill = regiao)) +
  geom_bar(stat = "identity", width = .75) +
  ggtitle("Qtd. óbitos novos por região") +
  labs(x = "Região", y = "Qtd. óbitos novos")

ggplot(covid_serie, aes(y = obitosNovos, x = estado, fill = regiao)) +
  geom_bar(stat = "identity", width = .75) +
  ggtitle("Qtd. óbitos novos por estado") +
  labs(x = "Estado", y = "Qtd. óbitos novos")



########################################
#Distribuicao dados casos novos Brasil
den <- density(covid_Br$casosNovos)
str(den)

x <- eval(parse(text = den$data.name))
ma <- mean(x)             # Média da amostra.
md <- median(x)           # Mediana da amostra.

modal <- which.max(den$y)
modal <- list(x = den$x[modal], y = den$y[modal])

plot(den,
     type = "n",
     xlab = "Qtd. casos novos",
     ylab = "Densidade",
     ylim = c(0, modal$y + strheight("1")),
     main = "Distribuição dados casos novos no Brasil",
     sub = paste("Quantidade total:", round(den$bw,3)))
with(den, polygon(x, y, col = "gray90"))
with(modal, {
  segments(x, 0, x, y, col = 2)
  text(x, y, labels = sprintf("Moda: %0.2f", x), pos = 3)
})
arrows(ma, 0, ma, modal$y/3, code = 1, length = 0.15)
text(ma, modal$y/3, labels = sprintf("Média: %0.2f", ma), pos = 3)
arrows(md, 0, md, modal$y/6, code = 1, length = 0.15)
text(ma, modal$y/6, labels = sprintf("Mediana: %0.2f", md),
     pos = ifelse(md<ma, 2, 4))
rug(covid_Br$casosNovos)
```

```
##########Distribuicao dados obitos novos Brasil
den <- density(covid_Br$obitosNovos)
str(den)

x <- eval(parse(text = den$data.name))
ma <- mean(x)             # Média da amostra.
md <- median(x)           # Mediana da amostra.

modal <- which.max(den$y)
modal <- list(x = den$x[modal], y = den$y[modal])

plot(den,
     type = "n",
     xlab = "Qtd. óbitos novos",
     ylab = "Densidade",
     ylim = c(0, modal$y + strheight("1")),
     main = "Distribuição dados óbitos novos no Brasil",
     sub = paste("Quantidade total:", round(den$bw,3)))
with(den, polygon(x, y, col = "gray90"))
with(modal, {
  segments(x, 0, x, y, col = 2)
  text(x, y, labels = sprintf("Moda: %0.2f", x), pos = 3)
})
arrows(ma, 0, ma, modal$y/3, code = 1, length = 0.15)
text(ma, modal$y/3, labels = sprintf("Média: %0.2f", ma), pos = 3)
arrows(md, 0, md, modal$y/6, code = 1, length = 0.15)
text(ma, modal$y/6, labels = sprintf("Mediana: %0.2f", md),
     pos = ifelse(md<ma, 2, 4))
rug(covid_Br$obitosNovos)


#######################################
#Distribuicao dados casos novos DF
den <- density(covid_DF$casosNovos)
str(den)

x <- eval(parse(text = den$data.name))
ma <- mean(x)             # Média da amostra.
md <- median(x)           # Mediana da amostra.

modal <- which.max(den$y)
modal <- list(x = den$x[modal], y = den$y[modal])

plot(den,
     type = "n",
     xlab = "Qtd. casos novos",
     ylab = "Densidade",
     ylim = c(0, modal$y + strheight("1")),
     main = "Distribuição dados casos novos no DF",
     sub = paste("Quantidade total:", round(den$bw,3)))
with(den, polygon(x, y, col = "gray90"))
with(modal, {
  segments(x, 0, x, y, col = 2)
```

```
  text(x, y, labels = sprintf("Moda: %0.2f", x), pos = 3)
})
arrows(ma, 0, ma, modal$y/3, code = 1, length = 0.15)
text(ma, modal$y/3, labels = sprintf("Média: %0.2f", ma), pos = 3)
arrows(md, 0, md, modal$y/6, code = 1, length = 0.15)
text(ma, modal$y/6, labels = sprintf("Mediana: %0.2f", md),
     pos = ifelse(md<ma, 2, 4))
rug(covid_DF$casosNovos)

#Distribuicao dados obitos novos DF
den <- density(covid_DF$obitosNovos)
str(den)

x <- eval(parse(text = den$data.name))
ma <- mean(x)             # Média da amostra.
md <- median(x)           # Mediana da amostra.

modal <- which.max(den$y)
modal <- list(x = den$x[modal], y = den$y[modal])

plot(den,
     type = "n",
     xlab = "Qtd. óbitos novos",
     ylab = "Densidade",
     ylim = c(0, modal$y + strheight("1")),
     main = "Distribuição dados óbitos novos no Brasil",
     sub = paste("Quantidade total:", round(den$bw,3)))
with(den, polygon(x, y, col = "gray90"))
with(modal, {
  segments(x, 0, x, y, col = 2)
  text(x, y, labels = sprintf("Moda: %0.2f", x), pos = 3)
})
arrows(ma, 0, ma, modal$y/3, code = 1, length = 0.15)
text(ma, modal$y/3, labels = sprintf("Média: %0.2f", ma), pos = 3)
arrows(md, 0, md, modal$y/6, code = 1, length = 0.15)
text(ma, modal$y/6, labels = sprintf("Mediana: %0.2f", md),
     pos = ifelse(md<ma, 2, 4))
rug(covid_DF$obitosNovos)
```