**TECH CHALLENGE - FASE 3**

**Análise do comportamento da População durante a Pandemia de COVID-19 utilizando PNAD-COVID 19 do IBGE**

--------------------------------------------------------------------------------------------------------

Carlos Santos

Fabia Bocayuva

Lucas Mathias

**APRESENTAÇÃO DO PROJETO**

O Hospital Hospitalize-se, com atuação em todo o território nacional, desempenhou um papel vital no combate à pandemia de Covid-19. Diante de desafios complexos, a equipe do hospital demonstrou alta capacidade de adaptação. Preocupado com futuros surtos, o Hospitalize-se contratou nossa consultoria de Data Analytics.

Com ampla experiência no setor de saúde, oferecemos soluções estratégicas baseadas em dados que capacitam instituições como o Hospitalize-se a tomar decisões mais assertivas e orientadas a dados. Desde análises exploratórias simples até abordagens mais sofisticadas, como o uso de machine learning, empregamos um leque de técnicas avançadas para apoiar nossos clientes.

Nossa missão é realizar uma análise aprofundada que revele padrões de comportamento, características da população e relevância dos sintomas apresentados durante a pandemia de Covid-19. Nosso objetivo é extrair indicadores para que o Hospitalize-se possa se preparar de forma eficaz para possíveis novos surtos — desde a gestão de medicamentos e equipamentos até o dimensionamento adequado de equipes médicas.

Após reuniões estratégicas com o cliente, decidimos utilizar a base de dados do IBGE, a PNAD-COVID19 considerando dados de **09/20** a **11/20**, que se tratam dos dados mais atualizados da base do PNAD. Essa base oferece uma visão ampla e estatisticamente representativa do impacto da pandemia em diferentes regiões e perfis populacionais no Brasil. A pesquisa utilizou procedimentos estatísticos rigorosos que garantem que a amostra represente de maneira aproximada a população brasileira.

Essa representatividade é fundamental para que o Hospital Hospitalize-se, que possui abrangência nacional, possa planejar suas ações e alocar recursos com eficiência em um cenário de incertezas. A análise dos dados nos permitirá identificar tendências, padrões de comportamento e fatores de risco que orientarão a formulação de estratégias adequadas, desde a aquisição de medicamentos até a alocação de equipamentos e a organização da equipe médica.

**PLANO DE PROJETO**

Dividimos o projeto em algumas etapas para melhor organização.

**1 - SOBRE OS DADOS**

**Características da Base:** Conhecendo a base de dados do PNAD-COVID do IBGE.

* Seleção das questões relevantes para desenvolvimento de indicadores para o planejamento em caso de uma nova pandemia de COVID-19.
* Definição de armazenamento e distribuição dos dados presentes na base do PNAD.
* Definição das características de análise essenciais para realização do planejamento.

**Recriação do ambiente de dados:** documentação necessária para replicar a estrutura de dados definida na etapa 1.

**2 - ANÁLISES**

* Análise exploratória de dados (EDA):
* Características gerais da população
* Infecção pelo Covid e Sintomas
* Análise de doenças crônicas
* Correlações.
* Providências para tratar os sintomas
* Análise Geográfica
* Distribuição de idade em relação à internação e intubação.
* Análise de cluster (K-means)

**3 - CONCLUSÃO**

**SOBRE OS DADOS**

**Características da Base**

Com a decisão de utilizar a base de dados PNAD-COVID do IBGE, tornou-se necessário definir o meio mais eficiente para acessá-la. Dado o grande volume de informações, optamos por utilizar um banco de dados em nuvem, o que nos garante escalabilidade, agilidade nas consultas e acesso global, além de aumentar a eficiência do processo.

Ao analisar a documentação da PNAD-COVID, identificamos a existência de uma versão da base já estruturada e disponível no BigQuery, plataforma existente do Google Cloud. Diante disso, decidimos utilizar essa base como referência para conduzir nossas análises, maximizando o desempenho e a acessibilidade aos dados necessários.

A base de dados do PNAD contém uma grande quantidade de informações resultantes de um questionário aplicado a uma amostra representativa da população. Para tornar a pesquisa e análise mais focadas e eficientes, optamos por selecionar um conjunto específico de variáveis, correspondentes a 20 questões do questionário original. Essa abordagem permite concentrar nossos esforços nas informações mais relevantes para os objetivos do projeto, facilitando a extração de insights estratégicos.

As questões utilizadas foram:

1. Qual sua idade?
2. Sexo
3. Cor ou raça
4. Qual sua escolaridade?
5. Na semana passada, o(a) Sr(a) teve quais sintomas
6. Por causa disso, o(a) Sr(a) foi a algum estabelecimento de saúde (na semana passada)?
7. Que providência o(a) Sr(a) tomou para se recuperar desses sintomas (na semana passada)?
8. Em que local(is) buscou o atendimento (na semana passada)?
9. Ao procurar o estabelecimento de saúde, teve que ficar internado(a) por um dia (24 horas) ou mais?
10. Durante esta internação, o(a) Sr(a) foi sedado(a), entubado(a) e colocado(a) em respiração artificial com ventilador?
11. O(A) Sr(a) tem algum plano de saúde médico, seja particular, de empresa ou de órgão público?
12. O(A) Sr(a) fez algum teste para saber se estava infectado(a) pelo coronavírus?
13. Qual tipo de teste utilizado (swab, sangue dedo ou veia).
14. Algum médico já lhe deu o diagnóstico de alguma dessas doenças?
15. Na semana passada, devido à pandemia do Coronavírus, em que medida o(a) Sr(a) restringiu o contato com as pessoas?
16. Na semana passada, por pelo menos uma hora, o(a) Sr(a) trabalhou ou fez algum bico?
17. Quanto recebe (ou retira) normalmente em seu trabalho?
18. Quanto recebeu (ou retirou) efetivamente em seu trabalho?

**Recriação do ambiente de dados**

A recriação da base de dados é uma etapa crucial para garantir a reprodutibilidade do experimento, pois um estudo analítico de qualidade precisa ser replicável, permitindo que outros pesquisadores ou equipes do Hospitalize-se possam validar os resultados obtidos no futuro.

Dessa maneira, este capítulo é dedicado a detalhar a estrutura e os procedimentos utilizados para replicar a base de dados na nuvem, especificamente no BigQuery. Explicaremos como foi configurada a infraestrutura, garantindo que o processo possa ser seguido e replicado com precisão em futuros estudos.

O primeiro passo é acessar a base PNAD-19 do IBGE pelo [LINK](https://basedosdados.org/dataset/c747a59f-b695-4d19-82e4-fef703e74c17?table=5dcaf8f0-6509-4dea-958b-4d23bc2a8695&utm_term=&utm_campaign=Conjuntos+de+dados+-+Gratuito&utm_source=adwords&utm_medium=ppc&hsa_acc=9488864076&hsa_cam=20482085189&hsa_grp=153440174097&hsa_ad=687156040280&hsa_src=g&hsa_tgt=dsa-2259102140321&hsa_kw=&hsa_mt=&hsa_net=adwords&hsa_ver=3&gad_source=1&gclid=CjwKCAjw0t63BhAUEiwA5xP54YkgeNMUq5j1cGiw_3chMspwy1CDOCLxMetzHwjMrUt1fqXoD80F4hoC6hAQAvD_BwE) e acessar a base de dados no ambiente do BigQuery: .

Interface gráfica do usuário, Texto

Descrição gerada automaticamente com confiança média

Como estamos lidando com um projeto já finalizado, sem previsões de alterações na estrutura dos dados, podemos utilizar diretamente a base de dados disponibilizada pelo IBGE. No entanto, caso estivéssemos operando em um ambiente dinâmico e sujeito a atualizações contínuas, seria uma boa prática criar uma réplica da base fornecida pelo IBGE, com atualizações periódicas. Isso garantiria que nossas aplicações continuassem funcionando sem interrupções, mesmo diante de possíveis mudanças na estrutura dos dados originais, assegurando a estabilidade e consistência do projeto ao longo do tempo.

O segundo passo se divide em dois caminhos, são eles:

* **Acesso via Python**: para viabilizar esse acesso, foi necessário criar uma chave de autenticação por meio do IAM - Administrator, configurada com permissões de leitura específicas para as bases do BigQuery. Após a geração da chave, é imprescindível recuperar as informações do projeto previamente criado no BigQuery, o que garante uma cobrança segmentada e controlada. Com a configuração finalizada, as credenciais são integradas ao script Python, o que permite realizar consultas dinâmicas diretamente na base ou acessar jobs previamente configurados (consultas predefinidas), otimizando o processo de análise. Em nosso projeto decidimos deixar a criação da consulta a cargo do notebook.
* **Acesso via PowerBI**:Para conectar o Power BI ao BigQuery, duas opções estão disponíveis:

1. **Importação:** Baixar uma cópia dos dados do BigQuery para o seu computador.
2. **Conexão direta:** Conectar o Power BI diretamente ao BigQuery, permitindo acesso aos dados em tempo real.

Em nosso trabalho, escolhemos a **importação dos dados**. Essa escolha foi justificada pelo baixo volume de dados obtido através da nossa consulta, devido ao uso de técnicas de agregação. Também possibilita que os usuários possam visualizar os resultados da nossa pesquisa, mesmo que não tenham acesso à plataforma.

Em nosso ambiente BigQuery, criamos um Projeto e um Dataset para armazenar tabelas com o resultado das consultas SQL que utilizamos. Depois selecionamos as tabelas que pretendemos utilizar no dashboard e importamos para o Power BI.

Tabela

Descrição gerada automaticamente

Sequência para importação no Power BI:

* + Abra o Power BI Desktop.
  + Clique em "Obter Dados" > "Mais" > "BigQuery".
  + Após a autenticação aparecerá uma tela com a lista dos seus projetos no BigQuery.
  + Selecione a tabela que deseja e clique em "Carregar".

Abaixo apresentamos o diagrama que representa o modelo que adotamos:

Diagrama

Descrição gerada automaticamente

**ANÁLISE DE DADOS**

Para auxiliar nas análises, foi desenvolvido um dashboard no Power BI, complementado por análises de correlação e distribuição de variáveis realizadas em Python, proporcionando uma visão mais detalhada e interativa dos dados. A seguir, descrevemos os principais pontos observados:

**Características gerais da população**

No período selecionado (09/2020 a 11/2020), havia 1.149.197 registros na base de dados. Pudemos notar que a população da amostra representa fortemente as características da população brasileira em categorias como idade, sexo, raça/cor, escolaridade.

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

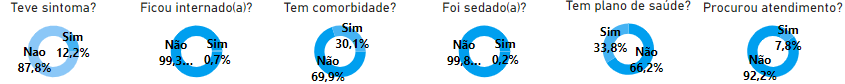
**Infecção pelo Covid e Sintomas**

Considerando apenas os casos em que o resultado foi positivo, não notamos diferenças significativas nas características como sexo, raça/cor ou escolaridade em relação à população geral. Com relação a idade, pessoas a partir dos 30 anos foram as que mais tiveram o resultado positivo, representando **73,5%** do total.

Diagrama

Descrição gerada automaticamente

**2,9%** da amostra testou positivo, representando aproximadamente **17 milhões** de pessoas e desses, **7,8%** buscaram atendimento em algum estabelecimento de saúde e a maioria com **66,2%**, não possui plano de saúde. Também notamos muitos resultados positivos que não apresentaram os sintomas considerados na pesquisa **87,8%.**



Também notamos um aumento de casos no decorrer do tempo, com **2,5%** casos em setembro, **2,9%** em outubro e **3,3%** em novembro.  Entre os que testaram **positivo**, os principais sintomas relatados foram:

Gráfico

Descrição gerada automaticamente

**Dor de cabeça**, foi o sintoma mais frequente com **12,07%** em geral. Ao analisar por faixa de idade, percebemos uma diferença nos sintomas apresentados em pessoas com mais de 60 anos e que contraíram a doença, onde, **dor muscular, fadiga e tosse** foram os sintomas mais frequentes.

Gráfico, Gráfico de funil

Descrição gerada automaticamente

**Análise de doenças crônicas**

**Hipertensão** foi a doença relatada com maior frequência considerando todas as idades e percebemos que a situação é diferente para pessoas abaixo dos 30 anos, que citam doenças pulmonares com maior frequência.

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

Para a faixa abaixo dos 30 anos, **doenças pulmonares** foram relatadas com maior frequência.

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

**Correlações**

A análise de correlação é fundamental para entender, com base em métodos estatísticos, como as variáveis se relacionam entre si.

Realizamos três análises principais nesse contexto: a primeira explorou a correlação entre os diferentes sintomas da COVID-19, a segunda investigou a relação entre esses sintomas e a presença de doenças crônicas, a última análise considerou a associação entre os sintomas e os casos de internação e intubação.

Para realizar as correlações, foi necessário transformar as variáveis categóricas em booleanas, onde respostas positivas para os sintomas foram consideradas como True e as demais opções como False.

Gráfico

Descrição gerada automaticamente

O heatmap evidencia a correlação entre os sintomas, destacando associações moderadas relevantes, como entre **dor de garganta** e **tosse**, **dor muscular** e **fadiga**, **dor de cabeça** e **dor muscular**, **febre** e **tosse**, **febre** e **dor de garganta**, **tosse** e **nariz entupido/escorrendo**, **dificuldade para respirar** e **dor no peito**, além de **dificuldade para respirar** e **fadiga**.

Essas correlações seguem uma lógica clínica, onde o agravamento de um sintoma está associado à intensificação de outro, o que faz sentido no contexto de evolução da doença.

Esses insights são fundamentais para o planejamento de cuidados futuros, pois ajudam a antecipar combinações de sintomas em pacientes com COVID-19, permitindo uma preparação clínica e logística mais eficiente.

Gráfico

Descrição gerada automaticamente

A segunda análise de correlação explora a relação entre os sintomas da COVID-19 e a presença de doenças crônicas, sendo fundamental para avaliar se essas condições podem influenciar a manifestação de sintomas.

Observamos que, embora doenças crônicas não causem necessariamente o surgimento de novos sintomas, elas podem intensificar os já existentes, aumentando a gravidade do quadro clínico.

Esse fator foi uma importante fonte de preocupação durante a pandemia, evidenciando a necessidade de atenção redobrada em pacientes com comorbidades, que apresentaram maior risco de complicações.

Gráfico, Gráfico de mapa de árvore

Descrição gerada automaticamente

Na correlação entre sintomas e casos de internação e intubação, não se evidenciou que um sintoma isolado seja diretamente responsável por internações ou intubações. No entanto, é importante destacar que a análise de correlação.

Entretanto, é essencial considerar os cenários clínicos observados nos hospitais durante a pandemia, que revelaram grande preocupação com pacientes que apresentavam doenças crônicas, comorbidades e idade avançada.

Embora esses fatores não estejam necessariamente associados a sintomas específicos, eles aumentam o risco de internação e intubação devido a fragilidades biológicas.

Esses dados destacam a vulnerabilidade dessas populações, que se mostraram mais propensas a desenvolver complicações graves da COVID-19.

**Providências para tratar os sintomas**

Entre os que testaram positivo, **7,8%** procuraram atendimento em algum estabelecimento de saúde e postos de saúde foram os locais mais procurados por essas pessoas.

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

Entre as pessoas que **não** buscaram atendimento, **ficar em casa**, foi a principal providência tomada para cuidar dos sintomas, conforme mostramos nos gráficos a seguir, refletindo a orientação disseminada pelas autoridades médicas na época.

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

**Análise Geográfica**

Em **novembro de 2020**, o estado que apresentou a maior proporção de casos positivos com relação à sua população foi **Roraima,** onde **7,9%** da população apresentou sintomas**. Minas Gerais** foi o estado com menor proporção, com **1,85%**.

Mapa

Descrição gerada automaticamente

Gráfico, Gráfico de barras

Descrição gerada automaticamente

**Análise de distribuição de idade em relação à internação e intubação**

Análise focada em identificar como grupos de risco, especialmente pessoas mais velhas, demonstram uma preocupação significativa quanto ao agravamento da COVID-19, especialmente nos casos de internação e intubação. A partir da análise do boxplot, observa-se de forma clara que a idade avançada é um fator de risco relevante, com maior incidência de internações e, de maneira ainda mais acentuada, nos casos de entubação. Esses resultados reforçam a vulnerabilidade desse grupo etário frente à gravidade da doença, e resulta em um indicador importante no tratamento da COVID-19.

Gráfico, Diagrama

Descrição gerada automaticamente

**Análise de cluster (K-means)**

O processo de análise de cluster k-means realizado nos dados do PNAD-Covid19, com o objetivo de identificar indicadores nos sintomas mais relevantes em pessoas testadas positivas para Covid-19, abaixo descrevemos todo o processo da análise:

**Separação dos casos positivos:** Inicialmente, é feita uma filtragem dos dados para selecionar apenas os casos positivos para Covid-19. Foram considerados casos positivos em um dos três testes presentes na base do PNAD (teste de cotonete na boca ou nariz, exame de sangue furo no dedo ou exame de sangue furo da veia do braço).

**Tratamento dos dados categóricos:** Todos os dados categóricos relacionados a sintomas são transformados em variáveis binárias, onde True indica a presença de um sintoma (resposta positiva sobre o sintoma) e False indica a ausência (demais respostas). Esse tratamento padroniza os dados para as próximas análises.

**Análise de componentes múltiplos (MCA):** Com os dados filtrados e tratados, aplica-se a técnica de Análise de Componentes Múltiplos (MCA), que permite reduzir a dimensionalidade dos dados categóricos e identificar as principais variáveis que explicam a variabilidade entre os casos, dividimos em duas dimensões.

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente

**Remoção de outliers:** Durante a análise MCA, outliers são identificados, utilizando a distância euclidiana, e removidos. Esses pontos fora do padrão podem distorcer a formação dos clusters, dificultando a obtenção de conclusões relevantes. A equipe do Hospitalize-se foi notificada da importância de observar esses outliers com atenção, pois podem representar características raras ou particulares que merecem uma análise mais detalhada.

**Verificação da quantidade de clusters:** Utilizamos os métodos elbow e silhouette, para identificar a quantidade ideal de clusters. Além dos métodos específicos para verificação de quantidade de cluster, observamos também a dispersão dos pontos presentes na análise MCA.

Gráfico, Gráfico de linhas

Descrição gerada automaticamenteGráfico, Gráfico de linhas

Descrição gerada automaticamente

**Aplicação do K-Means:** Analisando os resultados da etapa anterior, identificamos que a quantidade ideal de clusters para o nosso conjunto de dados é de 15. Com essa definição, aplicamos o algoritmo k-means aos dados transformados pela análise MCA. A partir desse processo, os casos foram agrupados de forma que padrões significativos pudessem ser observados e analisados dentro de cada cluster, facilitando a identificação de tendências e características comuns entre os grupos formados.



contagem das tuplas por cluster

**Agrupamento de clusters e cálculos estatísticos:** Em cada cluster formado, é calculada a mediana da relevância geral das variáveis MCA, bem como o desvio padrão de ambas dimensões resultantes do MCA. Isso permite identificar clusters mais homogêneos e com maior relevância para as variáveis estudadas.

Tela de computador com texto preto sobre fundo branco

Descrição gerada automaticamente

mediana e desvio padrão das dimensões MCA por cluster

Texto branco sobre fundo preto

Descrição gerada automaticamente

combinação entre dimensões x e y da análise MCA

**Análise gráfica**

A análise gráfica dos clusters é realizada, com foco nos clusters que apresentaram relevância percentil de 75% individual e na intersecção das duas dimensões da análise MCA. Realizamos um cálculo com base na média de desvios padrão para remover clusters com alto desvio, porém não foi encontrado nenhum cluster a ser removido.

Gráfico, Gráfico de barras

Descrição gerada automaticamente

Ao concluir a análise, é possível identificar os sintomas que se destacam com frequência acima da média nos clusters mais relevantes. Esses sintomas tornam-se indicadores importantes, oferecendo insights essenciais para orientar a preparação de recursos em um possível novo surto de Covid-19, permitindo uma alocação mais eficiente e estratégica. Além disso, contribuem para a triagem de casos suspeitos, aprimorando a capacidade de resposta e diagnóstico.

**CONCLUSÃO**

Durante as análises realizadas na etapa 2, foi possível a identificação de alguns indicadores, são eles:

* **Faixa etária da população:** Este indicador tem como objetivo compreender a distribuição proporcional da população por faixa etária. Dado que a idade foi identificada como um fator crucial no agravamento da COVID-19, entender essa distribuição torna-se essencial para o planejamento estratégico. A análise da proporção populacional por faixa etária oferece subsídios importantes para antecipar demandas de atendimento, especialmente em casos de internação e entubação.
* **Sintomas relevantes em casos positivos de COVID-19:** O indicador de sintomas é essencial para identificar possíveis casos positivos de COVID-19, com base em uma análise estatística avançada. Esse indicador busca otimizar a testagem de pacientes com suspeita da doença e mapear a necessidade de recursos de triagem, considerando a distribuição de sintomas relevantes nas diferentes regiões do Brasil. Os sintomas identificados como mais significativos incluem combinações de febre, tosse, dor de garganta, dor de cabeça, nariz entupido/escorrendo, fadiga, perda de olfato ou paladar e dor muscular. Esse mapeamento é crucial para melhorar a alocação de recursos e agilizar o diagnóstico nas unidades de saúde.
* **Escolaridade da população em função de comunicação eficaz:** A análise permitirá identificar as formas mais eficientes de comunicação com a população, otimizando as campanhas de conscientização e divulgação de informações sobre a COVID-19.

Com base nos indicadores gerados pelas análises, o Hospitalize-se estará mais capacitado para identificar grupos prioritários de pacientes, desenvolver campanhas de comunicação mais eficazes (adequando a mensagem e o meio de comunicação conforme o perfil da população) e otimizar a logística de estoques de medicamentos e equipamentos. Dessa forma, o hospital não apenas aprimorará a gestão de recursos e o planejamento estratégico, como também alcançará maior eficiência financeira, reduzindo desperdícios e alocando recursos de forma mais inteligente.

**LINKS**

**Base PNAD-COVID:**

<https://basedosdados.org/dataset/c747a59f-b695-4d19-82e4-fef703e74c17?table=5dcaf8f0-6509-4dea-958b-4d23bc2a8695&utm_term=&utm_campaign=Conjuntos+de+dados+-+Gratuito&utm_source=adwords&utm_medium=ppc&hsa_acc=9488864076&hsa_cam=20482085189&hsa_grp=153440174097&hsa_ad=687156040280&hsa_src=g&hsa_tgt=dsa-2259102140321&hsa_kw=&hsa_mt=&hsa_net=adwords&hsa_ver=3&gad_source=1&gclid=CjwKCAjw0t63BhAUEiwA5xP54YkgeNMUq5j1cGiw_3chMspwy1CDOCLxMetzHwjMrUt1fqXoD80F4hoC6hAQAvD_BwE>

**Github:** <https://github.com/magicblak/Analise_PNAD_COVID.git>