FACULDADE DE INFORMAÇÃO E ADMINISTRAÇÃO PAULISTA (FIAP)

FELIPE DE CAMPOS, JOÃO PEDRO VIEIRA, JOÃO PEDRO CHAMBRONE, LETICIA RESINA E VITOR HUGO RODRIGUES

CHALLENGE - LEPIC
DOCUMENTO BASE

São Paulo

2024

FELIPE DE CAMPOS MELLO ARNUS, JOÃO PEDRO DE SOUZA VIEIRA, JOÃO PEDRO OLIVEIRA CHAMBRONE, LETICIA CRISTINA GANDAREZ RESINA E VITOR HUGO GONÇALVES RODRIGUES

CHALLENGE - LEPIC;
DOCUMENTO BASE

Documento descritivo para entrega da Challenge do terceiro semestre de Engenharia de Software na FIAP.
Orientador: Paulo Sergio Sampaio

SÃO PAULO

2024

Descrição do Projeto e das variáveis

Projeto

Este projeto de **Data Science** tem como objetivo analisar um conjunto de dados relacionados à saúde, focando na identificação de fatores associados à diabetes. O conjunto de dados inclui informações detalhadas sobre o histórico médico e estilo de vida de vários indivíduos, bem como variáveis associadas a doenças crônicas. A análise busca entender padrões que podem ajudar na prevenção e controle do diabetes, examinando variáveis como hipertensão, colesterol, hábitos de consumo, atividade física e condições de saúde mental.

Para esta análise será utilizada a base ***"Sprint3 Diabetes.csv"***. Além disso, para uma melhor visualização das informações apresentadas, serão incluídos gráficos e histogramas.

Vale ressaltar que alguns valores serão aproximados, usando as casas decimais como referência, onde aqueles que tiverem a primeira casa sendo maior ou igual a 5, o valor do número da unidade será arredondado para 1 valor acima.Essa aproximação será feita apenas a título de explicação, todos os cálculos foram feitos com os valores reais para maior precisão.

Variáveis

O conjunto de dados contém diversas variáveis relacionadas à saúde e aos fatores de risco associados ao diabetes. Abaixo, descrevemos cada uma das variáveis presentes no banco de dados, incluindo o seu tipo, a presença de valores nulos e uma breve descrição dos atributos analisados.

1. Diabetes_012:

- Tipo: Float64
- Distribuição: Classificação da condição de diabetes(0: não tem; 1: pré diabetes; 2: diabetes).
- Média: 0.29.

2. HighBP:

- Tipo: Float64
- Distribuição: Indicador de hipertensão (1: sim; 0: não).
- Média: 0.42.

3. HighChol:

- Tipo: Float64
- Distribuição: Nível de colesterol alto(1: sim; 0: não).
- Média: 0.42.

4. CholCheck:

- Tipo: Float64
- Distribuição: Verificação recente de colesterol (1:sim; 0: não).
- Média: 0.96.

5. BMI:

- Tipo: Float64
- Distribuição: Índice de Massa Corporal.
- Média: 28.38.

6. Smoker:

- Tipo: Float64
- Distribuição: Indicador de tabagismo (1: fumante; 0: não fumante).
- Média: 0.44.

7. Stroker:

- Tipo: Float64
- Distribuição: Histórico de acidente vascular cerebral (AVC) (1: sim; 0: não).
- Média: 0.04.

8. HeartDiseaseorAttack:

- Tipo: Float64
- Distribuição: Histórico de ataque cardíaco ou doenças cardíacas (1: sim; 0: não).
- Média: 0.09.

9. PhysActivity:

- Tipo: Float64
- Distribuição: Prática de atividade física regular (1: sim; 0: não).
- Média: 0.75.

10. Fruits:

- Tipo: Float64

- Distribuição: Consumo regular de frutas (1: sim; 0: não).
- Média: 0.63.

11. Veggies:

- Tipo: Float64
- Distribuição: Consumo regular de vegetais (1: sim; 0: não).
- Média: 0.81.

12. HvyAlcoholConsump:

- Tipo: Float64
- Distribuição:Consumo pesado de álcool (1: sim; 0:não).
- Média: 0.05.

13. AnyHealthcare:

- Tipo: Float64
- Distribuição: Acesso a qualquer tipo de serviço de saúde (1: sim; 0: não).
- Média: 0.95.

14. NoDocbcCost:

- Tipo: Float64
- Distribuição: Incapacidade de consultar um médico devido a custos (1: sim; 0: não).
- Média: 0.08.

15. GenHlth:

- Tipo: Float64
- Distribuição: Autoavaliação da saúde geral (escala de 1 a 5, onde 1 é excelente e 5 é ruim).
- Média: 2.51.

16. MentHlth:

- Tipo: Float64
- Distribuição: Número de dias em que a saúde mental foi ruim nos últimos 30 dias.
- Média: 3.18.

17. PhysHlth:

- Tipo: Float64

- Distribuição: Número de dias em que a saúde física foi ruim nos últimos 30 dias.
- Média: 4.24

18. DiffWalk:

- Tipo: Float64
- Distribuição: Dificuldade de caminhar ou subir as escadas (1: sim; 0: não).
- Média: 0.16.

19. Sex:

- Tipo: Float64
- Distribuição: Sexo biologico (0: feminino; 1: masculino).
- Média: 0.44.

20. Age:

- Tipo: Float64
- Distribuição: Faixa etária categorizada (faixa etária reunida por valores numéricos).
- Média: 8.03.

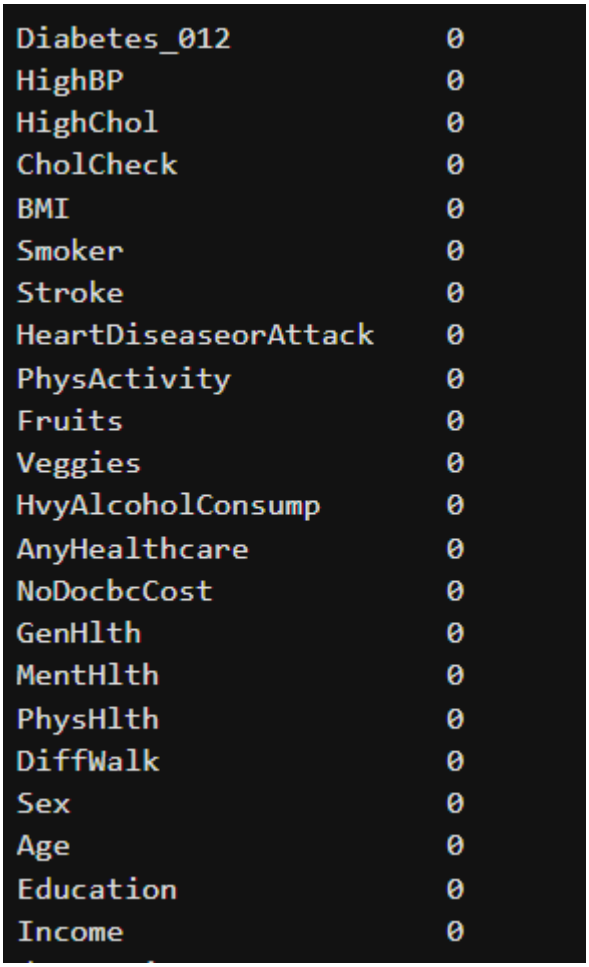
21. Education:

- Tipo: Float64
- Distribuição: Nível de escolaridade (de 1 a 6, onde 1 é sem ensino fundamental e 6 é pós-graduação).
- Média: 5.05.

22. Income:

- Tipo: Float64
- Distribuição:Faixa de renda anual (de 1 a 8, onde 1 representa renda baixa e 8 renda alta).
- Média: 6.05.

Valores Nulos



```
Diabetes_012            0
HighBP                  0
HighChol                0
CholCheck               0
BMI                     0
Smoker                  0
Stroke                  0
HeartDiseaseorAttack    0
PhysActivity            0
Fruits                  0
Veggies                 0
HvyAlcoholConsump       0
AnyHealthcare           0
NoDocbcCost             0
GenHlth                 0
MentHlth                0
PhysHlth                0
DiffWalk                0
Sex                     0
Age                     0
Education               0
Income                  0
```

Com base no retorno do comando executado, é possível verificar que não há valores nulos em nenhuma das variáveis, sendo assim, não há a necessidade de avaliar uma forma de contornar a existência dos mesmos.

Considerações sobre as variáveis

Após a verificação de que não existem valores nulos, foi executado do comando .describe(), para obter mais métricas. Uma vez que todas as colunas estão preenchidas com valores numéricos do tipo float, esse comando trouxe várias informações valiosas.

Por exemplo, observando algumas variáveis é possível perceber que seus valores alternam entre 1 e 0, tal qual valores booleanos, sendo assim, o retorno da média dessas variáveis já evidencia se há por exemplo se há um resultado positivo ou negativo para alguns mediadores, como a variável ***"Smoker"***, relativa a se o indivíduo fuma ou não, onde a média foi 0.44, e considerando que 0 é o indivíduo que não fuma e 1 o indivíduo que fuma, observa-se que das pessoas cuja as informações foram retiradas, a maioria não fuma.

	Diabetes_012	HighBP	HighChol	CholCheck	\
count	253680.000000	253680.000000	253680.000000	253680.000000	
mean	0.296921	0.429001	0.424121	0.962670	
std	0.698160	0.494934	0.494210	0.189571	
min	0.000000	0.000000	0.000000	0.000000	
25%	0.000000	0.000000	0.000000	1.000000	
50%	0.000000	0.000000	0.000000	1.000000	
75%	0.000000	1.000000	1.000000	1.000000	
max	2.000000	1.000000	1.000000	1.000000	

	BMI	Smoker	Stroke	HeartDiseaseorAttack	\
count	253680.000000	253680.000000	253680.000000	253680.000000	
mean	28.382364	0.443169	0.040571	0.094186	
std	6.608694	0.496761	0.197294	0.292087	
min	12.000000	0.000000	0.000000	0.000000	
25%	24.000000	0.000000	0.000000	0.000000	
50%	27.000000	0.000000	0.000000	0.000000	
75%	31.000000	1.000000	0.000000	0.000000	
max	98.000000	1.000000	1.000000	1.000000	

	PhysActivity	Fruits	...	AnyHealthcare	NoDocbcCost	\
count	253680.000000	253680.000000	...	253680.000000	253680.000000	
mean	0.756544	0.634256	...	0.951053	0.084177	
std	0.429169	0.481639	...	0.215759	0.277654	
min	0.000000	0.000000	...	0.000000	0.000000	
25%	1.000000	0.000000	...	1.000000	0.000000	
50%	1.000000	1.000000	...	1.000000	0.000000	
75%	1.000000	1.000000	...	1.000000	0.000000	
max	1.000000	1.000000	...	1.000000	1.000000	

	GenHlth	MentHlth	PhysHlth	DiffWalk	\
count	253680.000000	253680.000000	253680.000000	253680.000000	
mean	2.511392	3.184772	4.242081	0.168224	
std	1.068477	7.412847	8.717951	0.374066	
min	1.000000	0.000000	0.000000	0.000000	
25%	2.000000	0.000000	0.000000	0.000000	
50%	2.000000	0.000000	0.000000	0.000000	
75%	3.000000	2.000000	3.000000	0.000000	
max	5.000000	30.000000	30.000000	1.000000	

Estatísticas Descritivas

Usando métodos como média e mmediana, conseguimos algumas informações importantes para a nossa análise dos casos de diabetes e fatores associados a isso:

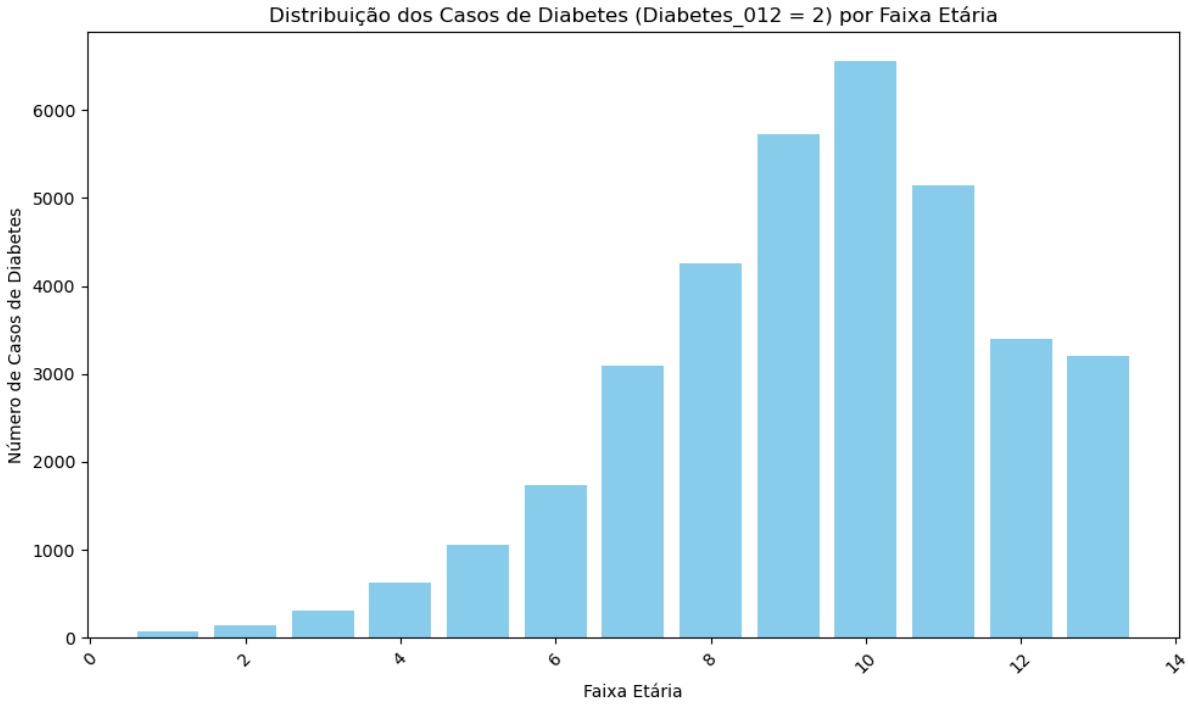
Incidência Geral de Diabetes por Faixa Etária

Calculando a soma dos casos de diabetes entre todos os indivíduos analisados, considerando que na variável ***"Diabetes_012"*** os números apresentados significam, respectivamente:

- 0 : Não possui diabetes
- 1 : Pré-diabetes
- 2 : Diabetes

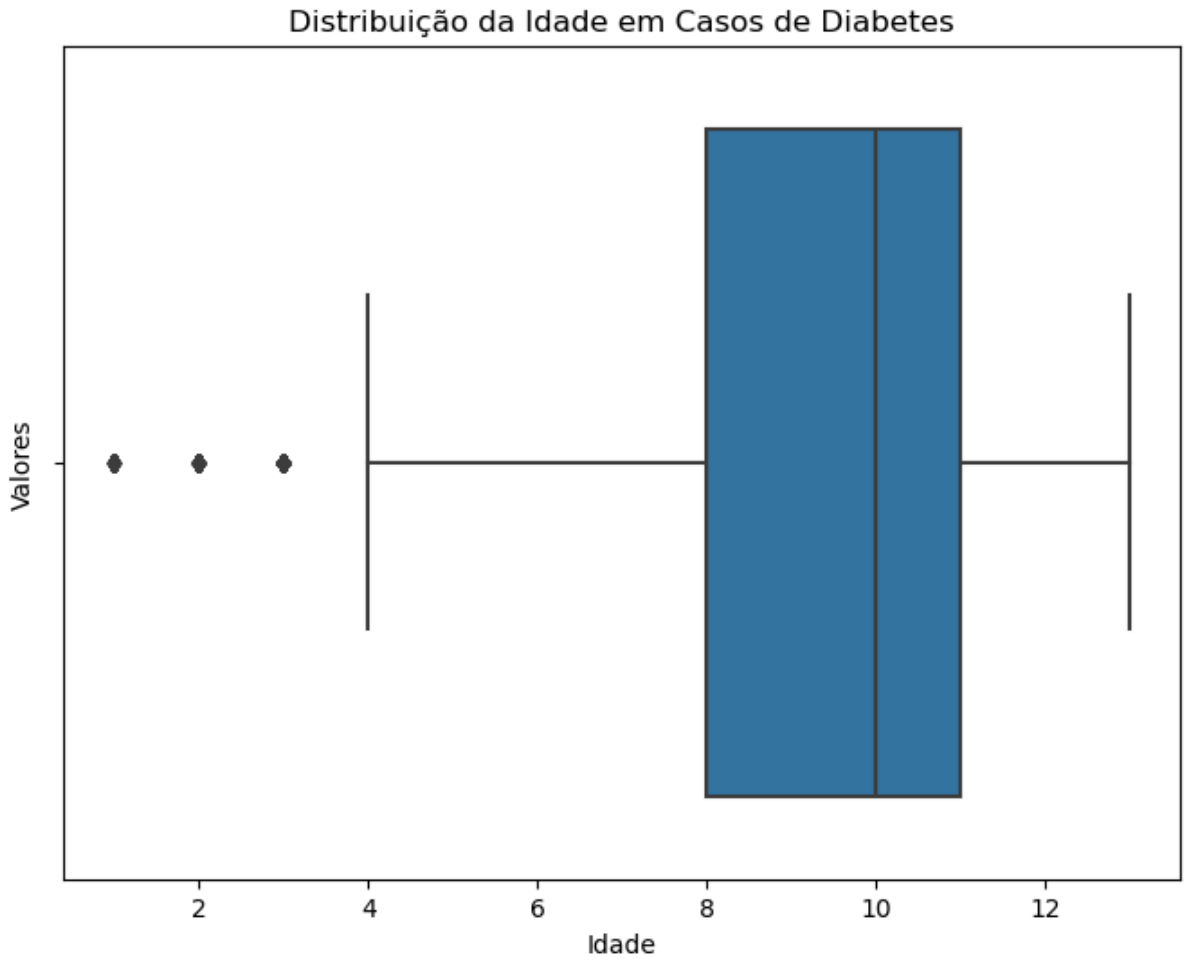
Sendo assim, está sendo levado em conta somente os casos onde na coluna ***"Diabetes_012"*** o número apresentado seja 2, e esses por sua vez são separados de acordo com a sua faixa etária, presente na coluna ***"Age"***.

Chegamos à seguinte distribuição da incidência de diabetes por faixa etária.



Com base neste gráfico, percebe-se que há uma maior incidência de diabetes em indivíduos que estão mais próximos da faixa etária 10. Vale ressaltar que o número da faixa etária não é o mesmo que a idade, uma vez que o valor máximo contido na variável ***"Age"*** é 13, mas ainda sim há variáveis que se tornaram inconsistentes se a maioridade presente fosse 13, como renda anual, e o consumo de álcool e cigarro.

Além disso, é possível ver que embora a média da variável idade seja 8, há uma maior ocorrência da faixa etária 10, o que pode indicar a existência de outliers. Mas para uma melhor verificação da existência desses outliers foi feita a geração de um gráfico boxplot, onde já foi apontado a existência de valores além dos limites indicados para esta variável.



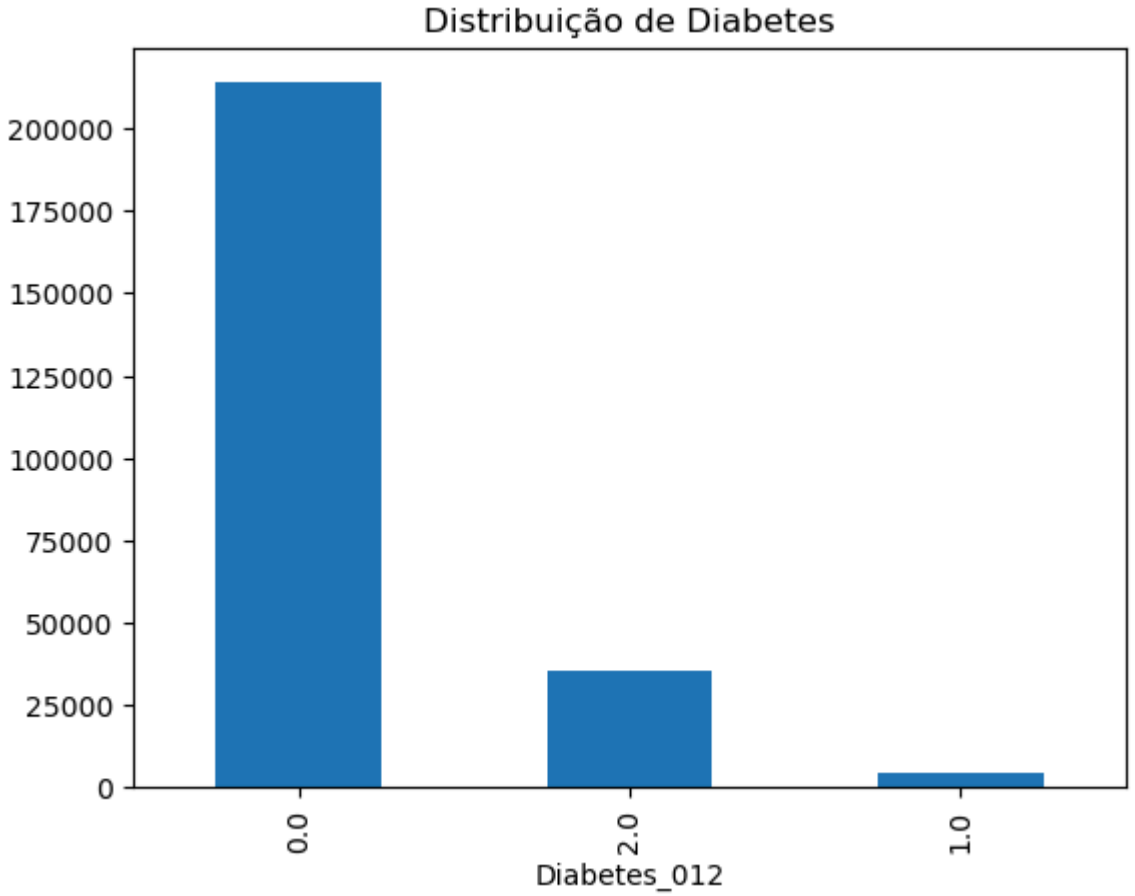
Situação do individual em relação a diabetes

Levando em consideração dessa vez somente a coluna ***"Diabetes_012"***, agora será verificada a distribuição da situação do indivíduo, ou seja se ele possui diabetes, pré-diabetes ou nenhuma das duas.

	Diabetes_012
count	253680.000000
mean	0.296921
std	0.698160
min	0.000000
25%	0.000000
50%	0.000000
75%	0.000000
max	2.000000

Analisando somente o retorno do comando .describe(), citado anteriormente, percebe se que a média da coluna é 0.29, e como já pré-estabelecido essa coluna possui apenas 3 valores possíveis, dessa forma, como a média está mais próxima de 0, a maioria dos indivíduos presentes no banco de dados não possui nem diabetes e nem pré-diabetes.

Para uma melhor visualização desse dado, segue um gráfico com a distribuição dos casos contidos na coluna.



Com esse, foi evidenciado o que já havia sido contestado por meio do cálculo da média, que a maioria dos indivíduos não possuem nada. Mas uma nova informação que o gráfico revelou foi que a segunda maior situação é de pessoas que possuem diabetes, e a menor é de pessoas que possuem pré-diabetes.

IMC

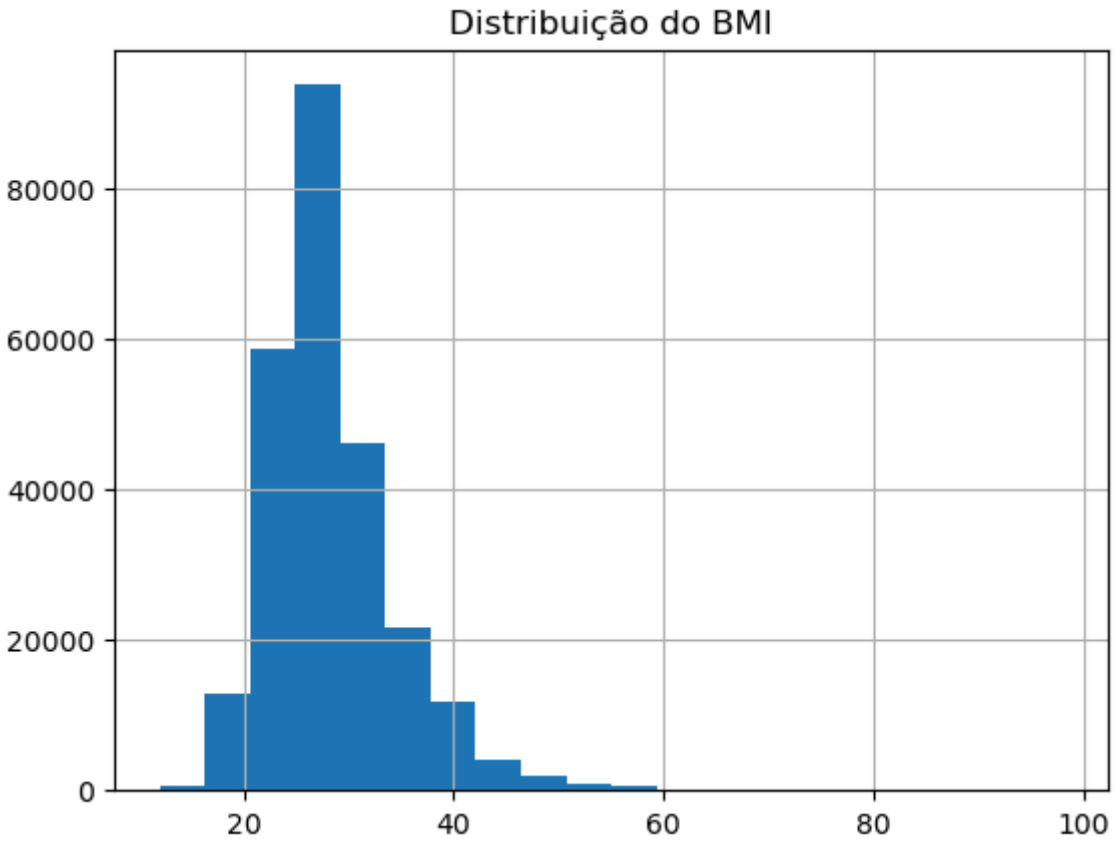
Para essa parte da análise será usada como base a coluna ***"BMI"***, onde estão contidos os valores do Índice de Massa Corporal (IMC) de cada indivíduo desse banco de dados.

Inicialmente, analisamos as medidas que nos foram fornecidas pelo comando .describe(), onde vemos que a média do IMC dos indivíduos é de 28.38, o que representa uma faixa já considerada de sobrepeso, além disso, também nos é mostrado que os valores têm uma grande variação, uma vez que o mínimo é de 12 (um valor considerado de magreza) e o máximo é de 98 (um valor considerado de obesidade grave).

	BMI
count	253680.000000
mean	28.382364
std	6.608694
min	12.000000
25%	24.000000
50%	27.000000
75%	31.000000
max	98.000000

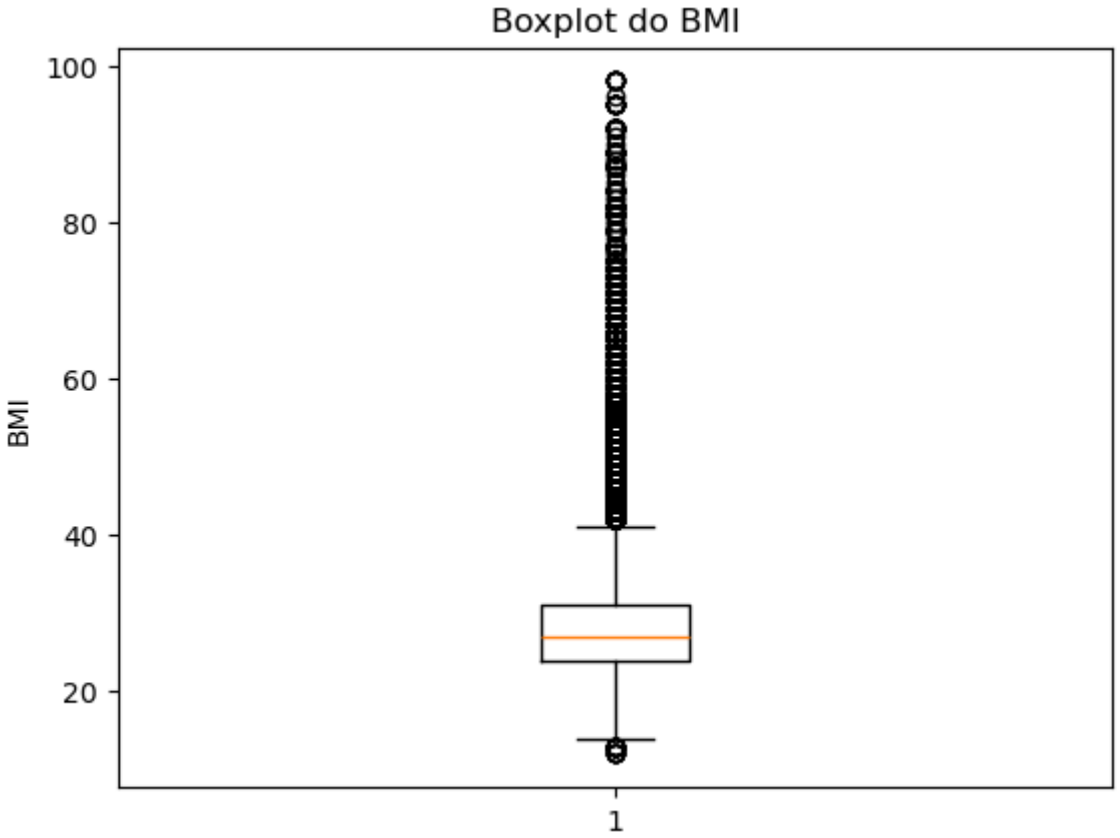
Com isso, percebemos que além de haver uma grande variação dos dados, a média dos indivíduos presentes nesse banco estão em uma situação de sobrepeso. Para confirmar essa variação, foi feito o cálculo do coeficiente de variação, o qual retornou um resultado de 23,28%, o que é um valor de variação considerado moderado, e que pode indicar a presença de outliers.

Para confirmar a presença desses outliers é preciso primeiro olhar para a distribuição dos valores em um histograma e em seguida se há a presença de valores além dos limites superiores ou inferiores em um boxplot.



Observando o histograma, percebesse que há uma grande concentração de valores entre 20 e 40, o que já era esperado, uma vez que é nesse trecho em que se encontra a média, mas além disso é perceptível que há valores que se distanciam bastante desse intervalo, uma vez que quando voltamos a olhar para o máximo retornado pelo .describe(),

vemos que há um valor que se aproxima de 100. Sendo isso, isso pode indicar a presença de outliers, mas para uma confirmação, recorreremos ao método de boxplot.



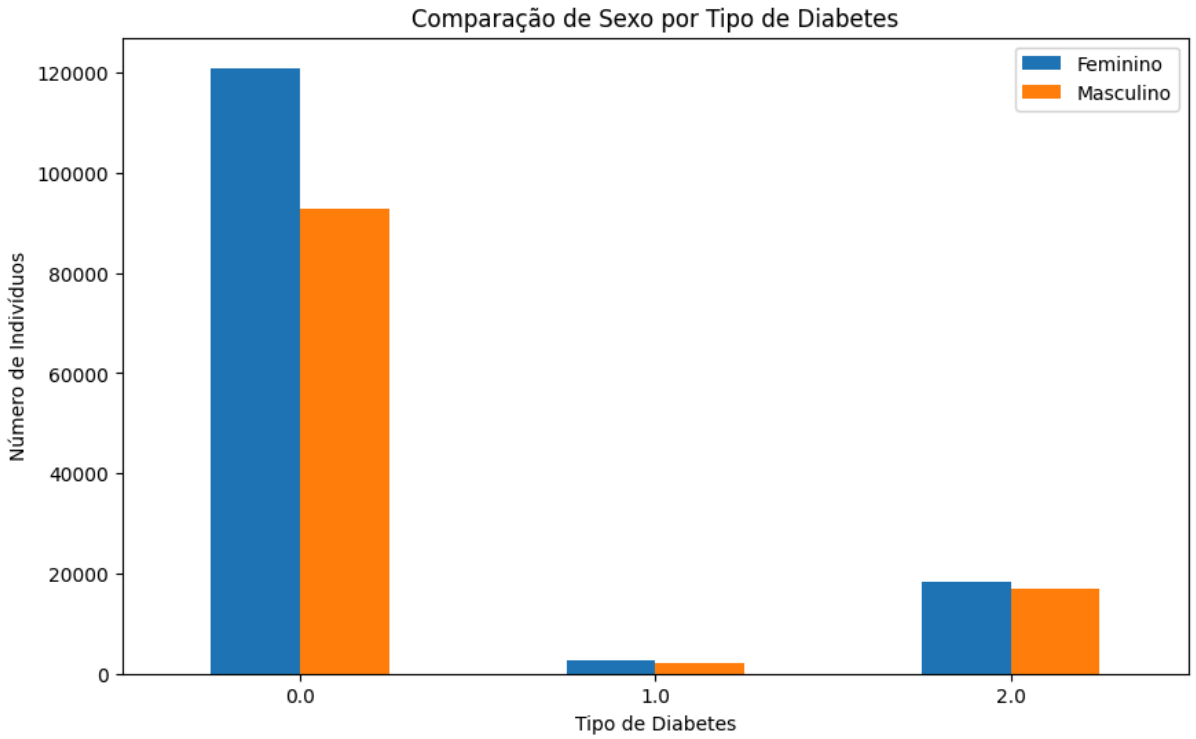
Por fim, com esse gráfico, pode-se ver a presença de outliers na distribuição desses valores relativos ao BMI,uma vez que há vários pontos além dos limites do boxplot, tanto os limites superiores, quanto inferiores.

Sexo

Para essa análise, é levado em consideração a variável ***"Sex"***, além do que já foi pré-definido sobre ela, que 0 é para feminino e 1 é para masculino.

Utilizando ainda dos retorno do comando .describe() e considerando também o que foi pré-estabelecido sobre essa variável, sendo o 0 sendo considerado feminino e 1 sendo masculino, é possível perceber que a média é de 0.44, evidenciando assim que há uma maior quantidade de mulheres do que de homens presentes nesse banco de dados.

Para uma melhor visualização da distribuição de cada sexo para cenario de diabetes, que como já falado anteriormente são não possui, possui pré-diabetes e possui diabetes, para respectivamente 0, 1 e 2 da variavavel ***"Diabetes_012"***, foi gerado um grafico de barras onde temos no eixo y o numero de individuos, e no eixo x o tipo de diabetes, e para a representação dos dois sexos foram geradas duas barras lado a lado de cores diferentes.



Com esse gráfico, percebe-se que houve uma maior incidência de mulheres nos 3 cenários possíveis.

Análises de Correlação e Testes de Hipóteses:

Essa seção é voltada para a análise da correlação e testes de hipóteses com o objetivo de entender as relações entre as variáveis presentes no dataset e investigar suposições específicas sobre o comportamento dos dados.

A análise de correlação é uma técnica estatística que mede o grau de associação entre duas ou mais variáveis. A correlação é representada por um coeficiente que varia de -1 a 1, onde 1 indica uma correlação positiva perfeita (quando uma variável aumenta, a outra também aumenta proporcionalmente), -1 indica uma correlação negativa perfeita (quando uma variável aumenta, a outra diminui proporcionalmente), e 0 indica que não há correlação linear entre as variáveis.

Essa técnica é especialmente útil para identificar relações entre variáveis que podem estar associadas. No contexto deste projeto, calcularemos a correlação entre vários índices, comparando com a variável ***"Diabetes_012"***. Através dessa análise, podemos identificar variáveis que se associam de forma mais forte com o desenvolvimento de diabetes.

Já os testes de hipóteses são procedimentos estatísticos utilizados para verificar se uma suposição sobre um determinado parâmetro populacional pode ser confirmada ou rejeitada com base nos dados da amostra. O teste de hipóteses segue um processo que inclui:

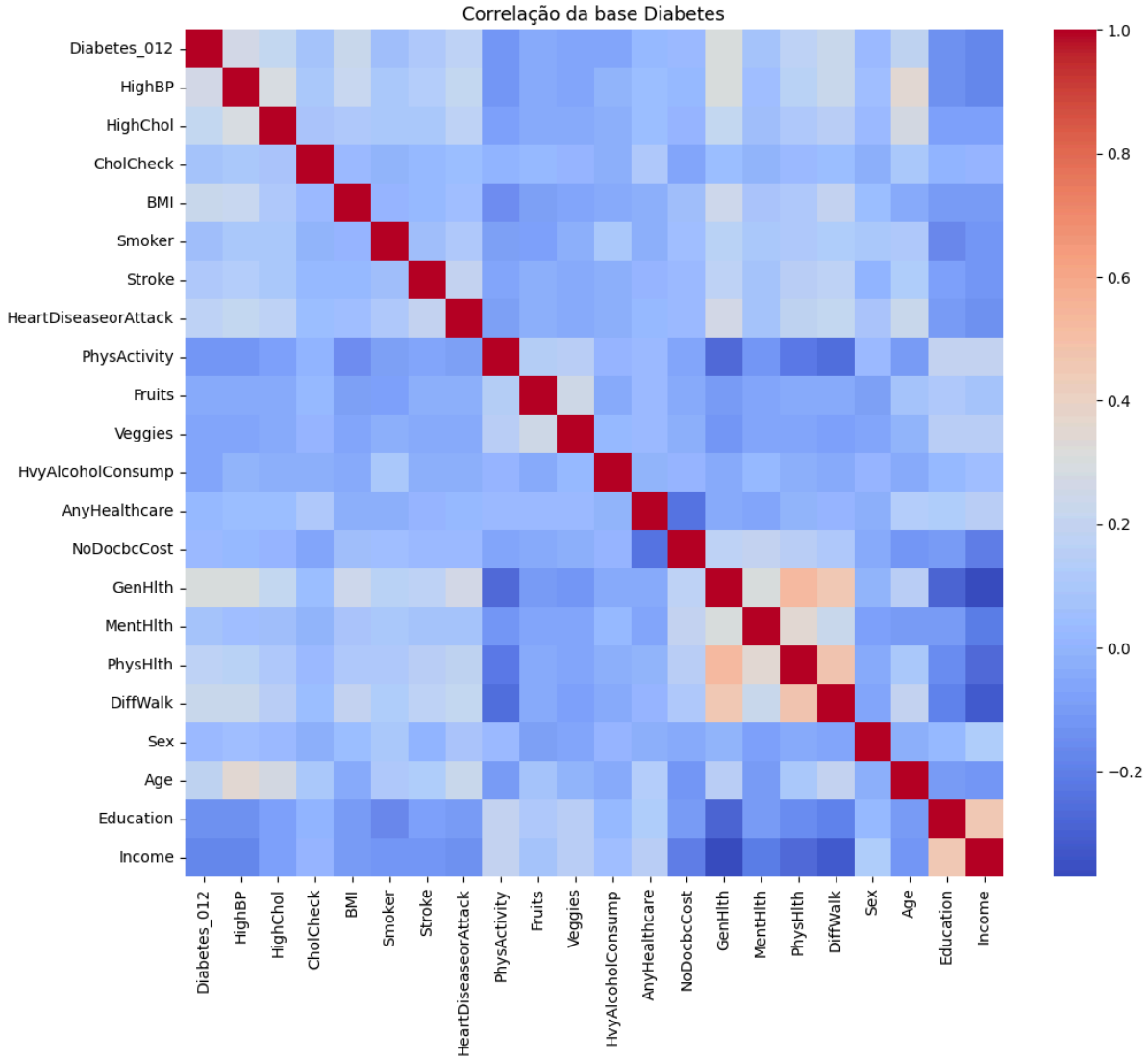
1. Formulação de uma hipótese nula (H_0): uma afirmação inicial que se assume verdadeira (por exemplo, "não há relação entre o BMI e o diabetes").

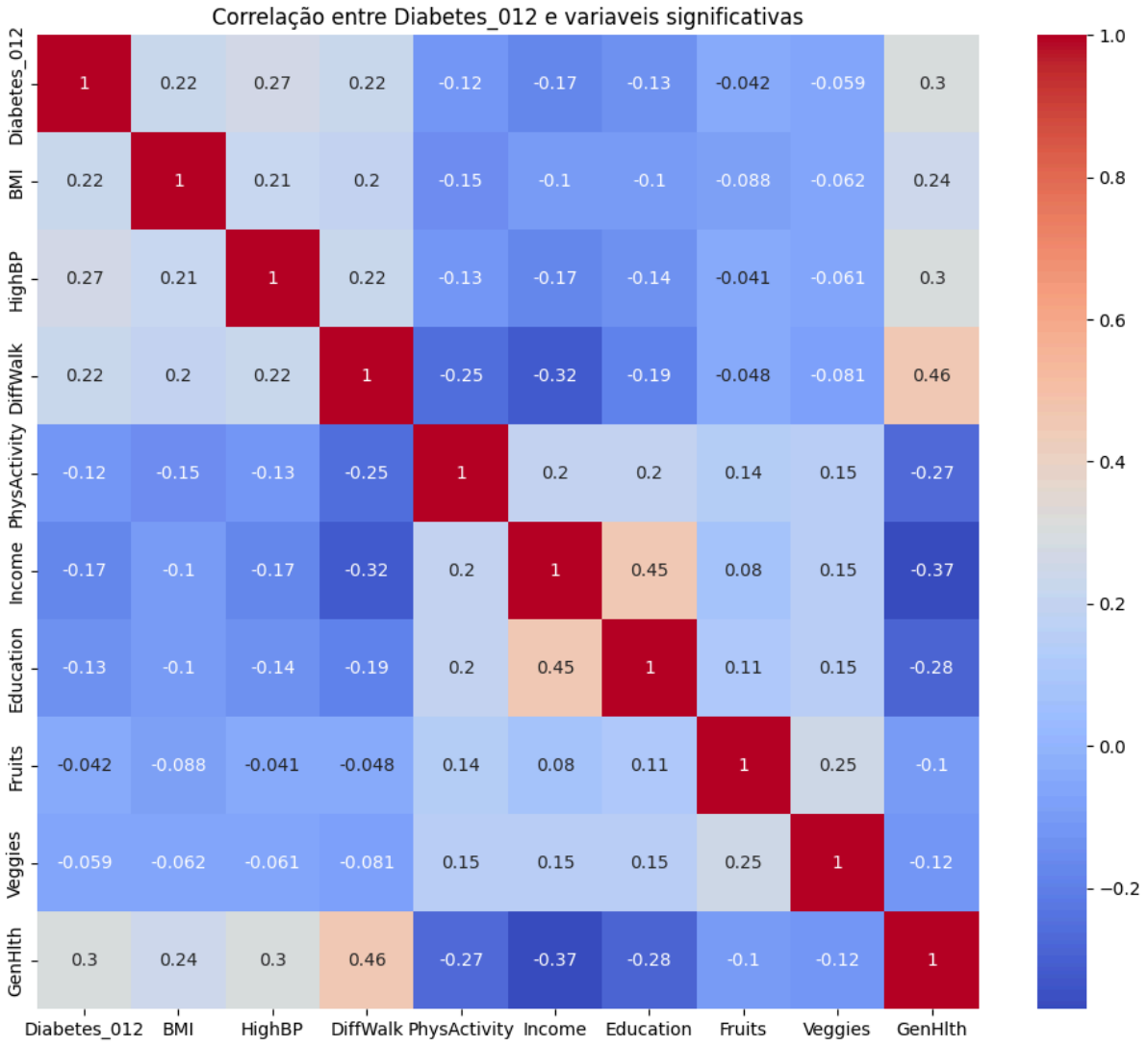
2. Hipótese alternativa (H_1): uma afirmação contrária à hipótese nula (por exemplo, "existe uma relação entre o BMI e o diabetes").
3. Cálculo de um valor estatístico para testar a validade da hipótese nula.
4. Decisão: Com base no valor-p (p-value) obtido, rejeitamos ou não a hipótese nula. Um p-value baixo (geralmente menor que 0,05) indica que há evidências suficientes para rejeitar a hipótese nula e aceitar a hipótese alternativa.

Esses testes serão utilizados para avaliar se variáveis como o BMI ou outros hábitos, como consumo de álcool ou prática de atividades físicas, têm impacto significativo no desenvolvimento de diabetes.

Correlação

Em primeira instância, para a análise da correlação entre as variáveis, foi criado um gráfico para uma comparação geral, e após isso foi criado um outro gráfico correlacional, mas dessa vez filtrando para usar apenas as variáveis mais significativas quando comparadas à variável ***"Diabetes_012"***.





Como no primeiro gráfico havia muitas informações e correlações muito fracas entre a maioria das variáveis, foi feita uma filtragem para as variáveis com maior possibilidade de ter uma correlação mais forte com a variável ***"Diabetes_012"***.

Observando a matriz de correlação apresentada, é preciso primeiro entender como as correlações são medidas. Se houver uma correlação, ela pode se apresentar de dois jeitos, uma correlação positiva ou negativa, tendo como indicadores a proximidade ao número 1 e -1, respectivamente. Porém, observando a matriz de correlação, percebemos, tanto pelos números quanto pelas cores, que não há nenhuma variável com uma correlação muito forte com a variável que estabelecemos como a principal para a análise.

Sendo que as correlações mais fortes observadas na matriz filtrada são das variáveis ***"BMI"***, ***"HighBP"***, ***"DiffWalk"*** e ***"GenHlth"*** , mesmo assim todas são correlação de nível moderado com o diagnóstico de diabetes. Além disso, também observa-se que variáveis como ***"PhysActivity"***, "***Income"*** e ***"Education"*** são as principais correlações inversas, mesmo que com um baixo nível de correlação para com o diagnóstico de diabetes.

Teste de Hipóteses

O teste de hipóteses é uma ferramenta fundamental na análise estatística, especialmente em projetos de Data Science, pois permite verificar a validade de suposições sobre dados com base em evidências amostrais. A técnica envolve formular duas hipóteses: a hipótese nula (H_0), que representa a ausência de efeito ou associação entre variáveis, e a hipótese alternativa (H_1), que sugere a existência de uma relação significativa. Ao comparar os resultados observados com os esperados sob a hipótese nula, podemos determinar se há evidências suficientes para rejeitá-la, aceitando a hipótese alternativa.

Neste trabalho, utilizamos diferentes testes de hipóteses para avaliar a relação entre diversas variáveis no contexto de um conjunto de dados de saúde. Em particular, investigamos a associação entre ***"BMI"*** (IMC), ***"Age"***, ***"PhyscActivity"***, e ***"HightChol"*** com o status de diabetes. Os testes escolhidos foram o ANOVA, para comparar médias de grupos, e o teste qui-quadrado, utilizado para verificar associações entre variáveis categóricas.

A seguir, exploramos a aplicação desses testes em diferentes variáveis do dataset, fornecendo interpretações dos resultados obtidos e explicando os conceitos estatísticos por trás dos cálculos, como o grau de liberdade e a estatística qui-quadrado. Este processo ajuda a esclarecer se fatores como a idade ou o colesterol alto estão significativamente associados ao desenvolvimento de diabetes, contribuindo para uma compreensão mais ampla da relação entre variáveis de saúde.

Diabetes e IMC

1. Hipóteses:

- **Hipótese nula (H_0)**: Não há diferença significativa entre as médias de IMC (BMI) nos três grupos de diabetes (Diabetes_012 = 0, 1, 2). Ou seja, a média do IMC é aproximadamente a mesma para pessoas sem diabetes, com pré-diabetes, e com diabetes.
- **Hipótese alternativa (H_1)**: Pelo menos uma das médias de IMC nos três grupos de diabetes é significativamente diferente.

2. O que o teste ANOVA faz:

O **ANOVA (Análise de Variância)** de uma via é utilizado para comparar as médias de três ou mais grupos. Neste caso, estamos comparando os valores médios de IMC entre as três categorias de diabetes (Diabetes_012):

- 0: Sem diabetes
- 1: Pré-diabetes
- 2: Diabetes

O ANOVA testa se as médias desses grupos são estatisticamente iguais ou se, pelo menos, uma das médias é significativamente diferente.

3. Interpretação dos resultados:

```
Estatística F: 6768.361066999288, p-valor: 0.0
Rejeitamos a hipótese nula. Existem diferenças significativas no BMI entre os grupos de diabetes.
```

- **Estatística F: 6768.36**: A estatística F é o valor calculado pelo ANOVA. Um valor de F muito alto indica que as médias entre os grupos podem ser significativamente diferentes. Neste caso, o valor de F é muito alto, sugerindo que há diferenças substanciais entre os grupos.
- **p-valor: 0.0**: O p-valor é a probabilidade de observar um resultado tão extremo (ou mais extremo) que os dados atuais, assumindo que a hipótese nula seja verdadeira. Um p-valor de 0.0 indica que a chance de observar essas diferenças entre as médias dos grupos de IMC, se não houvesse realmente diferença (isto é, se a hipótese nula fosse verdadeira), é praticamente inexistente.
- **Decisão**: Como o p-valor é muito menor que 0.05 (o nível de significância padrão), rejeitamos a hipótese nula. Isso significa que **existem diferenças significativas nas médias de IMC entre os três grupos de diabetes**. Portanto, há uma relação entre o status de diabetes e o valor do IMC.

4. Conclusão:

Com base nos resultados do teste ANOVA, concluímos que o IMC médio varia significativamente entre pessoas sem diabetes, pessoas com pré-diabetes e pessoas com diabetes. Este resultado sugere que o IMC pode estar relacionado ao risco de desenvolver diabetes, com diferentes médias de IMC entre os grupos.

Diabetes e Idade

1. Hipóteses:

- **Hipótese nula (H_0)**: Não há associação entre a faixa etária (Age) e o status de diabetes (Diabtes_012). Ou seja, as distribuições das faixas etárias são independentes da condição de diabetes.
- **Hipótese alternativa (H_1)**: Existe uma associação entre a faixa etária e o status de diabetes, o que significa que a condição de diabetes pode variar conforme as faixas etárias.

2. O que o teste qui-quadrado faz:

O **teste qui-quadrado** é usado para verificar se há uma associação entre duas variáveis categóricas. Nesse caso, as variáveis são:

- **Diabetes_012** (com três categorias: 0 para sem diabetes, 1 para pré-diabetes, 2 para diabetes)
- **Age** (categorias representando faixas etárias).

O teste compara a distribuição observada de combinações entre as categorias de ***"Diabetes_012"*** e ***"Age"*** com uma distribuição esperada sob a hipótese nula de que não há associação entre as variáveis.

3. Interpretação dos resultados:

```
Qui-quadrado: 9641.376530679843, p-valor: 0.0, Graus de liberdade: 24
Rejeitamos a hipótese nula. Existe associação entre Faixa Etária e Diabetes.
```

- **Qui-quadrado: 9641.38**: A estatística qui-quadrado é uma medida de quão diferentes as distribuições observadas são em relação às esperadas. Um valor muito alto de qui-quadrado sugere que as duas variáveis estão associadas, ou seja, as distribuições observadas se afastam muito do esperado sob a hipótese nula.
- **p-valor: 0.0**: O p-valor é extremamente baixo, próximo de zero. Isso significa que a probabilidade de obter as diferenças observadas nas distribuições das faixas etárias em relação ao status de diabetes, se não houvesse associação entre elas, é praticamente nula.
- **Graus de liberdade (dof): 24**: O número de graus de liberdade depende do número de categorias em cada variável. No caso, para três categorias de ***"Diabetes_012"*** e várias faixas etárias em ***"Age"***, o grau de liberdade é 24. Este valor indica a quantidade de combinações possíveis nas distribuições.

4. Conclusão:

- **Decisão**: Como o p-valor é muito menor do que 0.05 (o nível de significância típico), rejeitamos a hipótese nula.
- **Conclusão final**: Isso significa que existe uma associação significativa entre a faixa etária e o status de diabetes. Ou seja, as faixas etárias têm alguma relação com a presença ou ausência de diabetes, o que pode indicar que certas faixas etárias têm maior ou menor risco de desenvolver diabetes.

Diabetes e Colesterol Alto

1. Hipóteses:

- **Hipótese nula (H_0)**: Não há associação entre o status de colesterol alto (HighChol) e o status de diabetes (Diabetes_012). Ou seja, as pessoas que têm ou não colesterol alto estão distribuídas igualmente entre as categorias de diabetes.
- **Hipótese alternativa (H_1)**: Existe uma associação entre o status de colesterol alto e o status de diabetes, o que significa que a distribuição de colesterol alto varia conforme a presença ou ausência de diabetes.

2. O que o teste qui-quadrado faz:

O **teste qui-quadrado** é usado para verificar se há uma associação entre duas variáveis categóricas. Nesse caso, as variáveis são:

- **Diabetes_012** (com três categorias: 0 para sem diabetes, 1 para pré-diabetes, 2 para diabetes)
- **HighChol** (com duas categorias: 0 para não tem colesterol alto, 1 para tem colesterol alto).

O teste compara a distribuição observada de combinações entre ***"Diabetes_012"***e ***"HighChol"*** com a distribuição esperada, sob a hipótese nula de que as variáveis são independentes.

3. Interpretação dos resultados:

```
Qui-quadrado: 11258.920399414841, p-valor: 0.0, Graus de liberdade: 2
Rejeitamos a hipótese nula. Existe associação entre Diabetes e Colesterol Alto.
```

- **Qui-quadrado: 11258.92**: A estatística qui-quadrado é uma medida de quão diferentes as distribuições observadas são em relação às esperadas. Um valor muito alto de qui-quadrado indica que há uma forte evidência de que as variáveis ***"Diabetes_012"*** e ***"HighChol"*** estão associadas.
- **p-valor: 0.0**: O p-valor é muito pequeno (próximo de 0), o que indica que a probabilidade de observar uma diferença tão grande entre as distribuições, se não houvesse uma associação real, é extremamente baixa.
- **Graus de liberdade (dof): 2**: O número de graus de liberdade. No caso, temos três categorias para ***"Diabetes_012"*** e duas para ***"HighChol"***, resultando em 2 graus de liberdade.

4. Conclusão:

- **Decisão**: Como o p-valor é muito menor do que 0.05 (o nível de significância típico), rejeitamos a hipótese nula.
- **Conclusão final**: Isso significa que há uma associação significativa entre o status de diabetes e o colesterol alto. Em outras palavras, pessoas com diferentes condições de diabetes têm diferentes probabilidades de apresentar colesterol alto.

Conclusão

Com todos os dados mostrados até agora, pode-se perceber o impacto do diabetes e os fatores que podem influenciar em seus diagnósticos. Ao analisar variáveis que representam valores relativos a faixa etária, IMC, sexo, e outros fatores de saúde, identificamos padrões e correlações importantes que ajudam a compreender o perfil dos indivíduos com diabetes e os possíveis fatores de risco. As análises estatísticas realizadas, como a correlação e os testes de hipótese, trouxeram insights valiosos sobre as relações entre essas variáveis e o diabetes.

Diabetes por Faixa Etária

A análise da incidência de diabetes por faixa etária revelou que os casos de diabetes (Diabetes_012 = 2) estão mais concentrados em indivíduos nas faixas etárias mais avançadas, especificamente próximas à faixa etária 10. No entanto, é importante destacar que o valor na variável ***"Age"*** não representa diretamente a idade dos indivíduos, mas sim uma categorização por faixas etárias, sendo que o valor máximo é 13. A observação de uma alta incidência na faixa etária 10, que se encontra acima da média de idade (8), sugere a possível presença de outliers.

Essa hipótese foi confirmada pela análise de um gráfico boxplot da variável idade, que indicou a existência de valores além dos limites esperados, sugerindo que certos grupos etários têm uma incidência atípica de diabetes. Isso reforça a necessidade de um estudo mais aprofundado para identificar fatores específicos que possam estar influenciando o aumento de casos de diabetes em determinadas faixas etárias. Além disso, vale ressaltar que variáveis como renda anual, consumo de álcool e cigarro, que não são geralmente associadas a idades muito jovens, também apontam para a complexidade dessa categorização.

Situação Geral dos Indivíduos com Diabetes

A análise da situação geral dos indivíduos em relação à diabetes, baseada na variável ***"Diabetes_012"***, indica que a maioria dos indivíduos presentes no banco de dados não possui diabetes nem pré-diabetes. Isso foi confirmado ao verificar a média da coluna ***"Diabetes_012"***, que é 0.29, o que está mais próximo de 0, indicando que a maioria dos registros correspondem a indivíduos sem diabetes (Diabetes_012 = 0), ou seja, pessoas sem diabetes ou pré-diabetes.

A visualização gráfica dessa distribuição reforça o resultado: a maior parte da amostra é composta por indivíduos saudáveis em relação ao diabetes, seguidos por aqueles que têm diabetes (Diabetes_012 = 2), enquanto a menor parcela é de indivíduos com pré-diabetes (Diabetes_012 = 1). Este achado é interessante, pois demonstra que, embora a maioria dos indivíduos não apresente diabetes, os casos de diabetes confirmados ainda são mais prevalentes do que os de pré-diabetes. Isso pode indicar que muitos indivíduos só são diagnosticados quando a doença já está em estágio avançado.

IMC

A análise da coluna ***"BMI"***, que contém os valores do Índice de Massa Corporal (IMC) dos indivíduos, revela que a média de IMC é de 28.38, o que se encontra na faixa de sobrepeso. Isso já indica uma preocupação com o peso médio da população analisada. Além disso, a variabilidade nos dados é significativa, com valores que vão de 12 (classificado como magreza) até 98 (obesidade grave).

Ao calcular o coeficiente de variação, encontramos um valor de 23,28, indicando uma variação moderada. Isso pode sugerir a presença de outliers. Para visualizar melhor essa variação, foi gerado um histograma, que mostrou uma concentração de valores entre 20 e 40, o que é coerente com a média observada. Entretanto, alguns valores se distanciam bastante desse intervalo, como o valor máximo de quase 100, indicando a possibilidade de outliers.

A confirmação dessa hipótese veio com a análise do boxplot, onde vários pontos se situam além dos limites superior e inferior do gráfico, confirmando a existência de outliers na

distribuição do BMI. Esses outliers podem representar casos atípicos na população estudada, sugerindo uma maior investigação.

Sexo

Na análise da variável ***"Sex"***, onde 0 representa o sexo feminino e 1 o sexo masculino, observamos que a média de 0.44 indica uma predominância de mulheres na amostra de dados. Esse resultado sugere que a população estudada é composta por uma quantidade maior de mulheres em comparação aos homens.

Para facilitar a visualização da distribuição de indivíduos em relação à condição de diabetes (não possui, pré-diabetes e diabetes), foi gerado um gráfico de barras. Nesse gráfico, o eixo y representa o número de indivíduos e o eixo x mostra as diferentes categorias da variável ***"Diabetes_012"***. As barras para cada sexo foram representadas lado a lado, utilizando cores distintas.

A análise do gráfico revela uma maior incidência de mulheres em todos os três cenários de diabetes. Isso sugere que, nesta população, as mulheres estão mais representadas em todas as categorias relacionadas à condição de diabetes, o que pode ter implicações importantes para futuras investigações e intervenções voltadas à saúde.

Correlação

Na análise de correlação, inicialmente foi elaborado um gráfico que apresentava uma comparação geral entre as variáveis do dataset. Contudo, devido à quantidade de informações e à presença de correlações fracas entre a maioria das variáveis, uma segunda análise foi realizada, focando apenas nas variáveis consideradas mais significativas em relação à variável ***"Diabetes_012"***.

A matriz de correlação filtrada foi então gerada para proporcionar uma visão mais clara das relações entre as variáveis mais relevantes. As correlações são medidas em um intervalo que vai de -1 a 1, onde valores próximos de 1 indicam uma correlação positiva forte e valores próximos de -1 refletem uma correlação negativa forte. Ao examinar a matriz, observa-se que não existem variáveis com correlações muito fortes em relação à variável principal, ***"Diabetes_012"***.

As correlações positivas mais notáveis foram identificadas nas variáveis ***"BMI"***, ***"HighBP"***, ***"DiffWalk"*** e ***"GenHlth"***, todas apresentando correlações de nível moderado com o diagnóstico de diabetes. Além disso, variáveis como ***"PhysActivity"***, ***"Income"*** e ***"Education"*** mostraram-se como as principais correlações negativas, ainda que com um nível baixo de correlação em relação ao diagnóstico de diabetes. Esses achados sugerem que, embora algumas variáveis possam ter uma associação moderada com a diabetes, não há evidências claras de correlações fortes, indicando a complexidade da condição e a necessidade de considerar múltiplos fatores em análises futuras.

Teste de Hipótese

Os testes de hipóteses são ferramentas estatísticas fundamentais que permitem avaliar suposições sobre populações com base em amostras. Através do processo de formular uma hipótese nula (H_0), e uma hipótese alternativa (H_1), os pesquisadores podem determinar se as evidências observadas nos dados são suficientemente fortes para rejeitar a hipótese nula. A interpretação dos resultados, que geralmente envolve o cálculo de um valor p, ajuda a entender a significância estatística das associações observadas. Neste estudo, utilizamos diversos testes, como ANOVA e qui-quadrado, para investigar as relações entre o diabetes e variáveis como IMC, faixa etária e colesterol alto. Esses testes nos proporcionaram uma compreensão mais profunda das associações e dos possíveis fatores de risco relacionados à diabetes.

Diabetes e IMC

A aplicação do teste ANOVA revelou diferenças significativas nos valores de IMC entre os grupos de diabetes. Essa associação sugere que o IMC pode ser um fator de risco importante para o desenvolvimento de diabetes. O aumento do IMC está relacionado a um maior risco de transição para as categorias de pré-diabetes e diabetes, reforçando a importância de intervenções focadas na manutenção de um peso saudável como estratégia de prevenção.

Diabetes e Idade

A aplicação do teste qui-quadrado mostrou uma associação significativa entre a faixa etária e o status de diabetes. Com um valor de qui-quadrado elevado e um p-valor praticamente nulo, os resultados indicam que as diferentes faixas etárias têm uma relação direta com o risco de desenvolvimento de diabetes. Isso sugere que a idade desempenha um papel importante na progressão da condição, com certas faixas etárias apresentando maior ou menor probabilidade de diagnóstico de diabetes.

Diabetes e Colesterol Alto

Além disso, os testes qui-quadrado mostraram uma associação forte entre diabetes e colesterol alto. A elevada estatística do teste indica que os indivíduos com colesterol elevado têm maior probabilidade de serem diagnosticados com diabetes. Essa informação é crucial para a implementação de estratégias de saúde pública que promovam a monitorização dos níveis de colesterol como parte da abordagem geral para o controlo da diabetes.