**Relatório do Desafio**

**Pergunta 1**

Quantas instâncias e características existem, respectivamente, no dataset?

( ) (12, 5110)

( ) (7, 5000)

( ) (5000, 7)

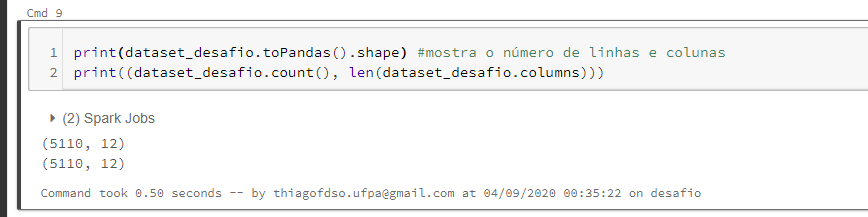
(X) (5110, 12)

**R.: Executar o comando**

**print(dataset\_desafio.toPandas().shape),**

**ou**

**print((dataset\_desafio.count(), len(dataset\_desafio.columns)))**



**Pergunta 2**

Quantas variáveis do tipo “string” estão presentes no dataset?

( ) 3

(X) 6

( ) 4

( ) 2

**R.: Executar o comando**

**len([entry[1] for entry in dataset\_desafio.dtypes if entry[1]=='string'])**



**Pergunta 3**

Qual é a idade (age) média dos entrevistados?

( ) 55,12 anos

( ) 22,61 anos

( ) 45,28 anos

(X) 43,22 anos

**R.: Executar o comando**

**dataset\_desafio.agg({"age": "avg"}).show()**



**Pergunta 4**

Sobre a distribuição de AVC em relação ao sexo (gender) dos entrevistados,

é **CORRETO** afirmar:

( ) Não podem ser identificadas diferenças entre os gêneros, pois o dataset está equilibrado (mulheres=homens)

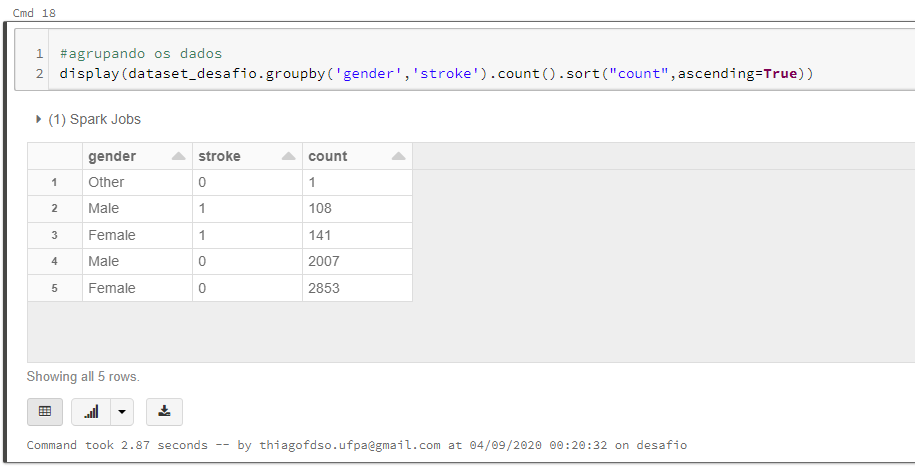
( ) Existe no dataset uma maior quantidade de homens que sofreram AVC

(X) Apesar da pouca diferença, existe uma maior quantidade de mulheres que sofreram AVC

( ) Existe no dataset apenas dois tipos de gêneros, homens e mulheres

**R.: Executar o comando**

**display(dataset\_desafio.groupby('gender','stroke').count().sort("count",ascending=True))**



**Pergunta 5**

Sobre o dataset é correto afirmar, **EXCETO**:

( ) Existem dois tipos diferentes de classes de residências (“Residence\_type”) presentes nesse dataset

(X) O dataset está balanceado. Existem quantidades similares de instâncias de indivíduos que sofreram AVC e que não sofreram dessa enfermidade

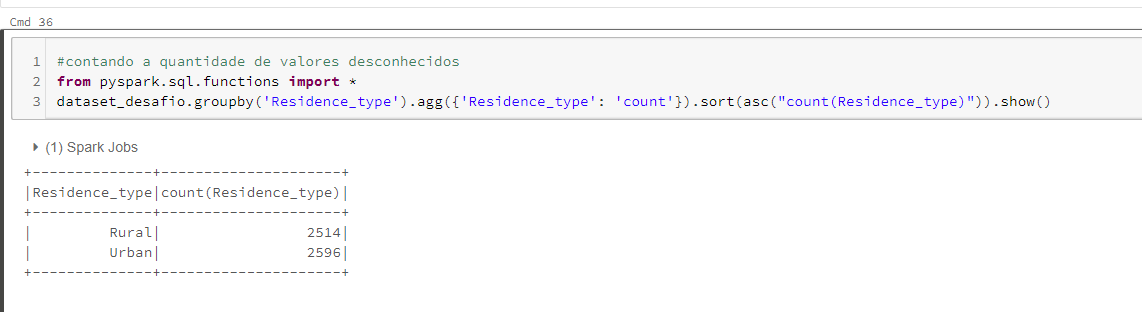
( ) Existem dados categóricos e numéricos presentes neste dataset. Um exemplo de dados categóricos é o “Residence\_type”

( ) A variável bmi possui valores não numéricos

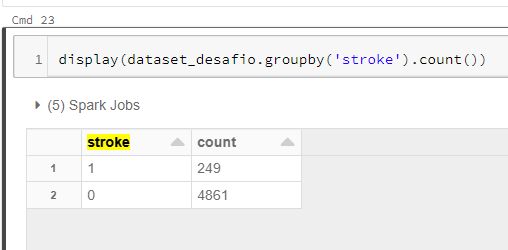
**R.: Executar os comandos**

**from pyspark.sql.functions import \***

**dataset\_desafio.groupby('Residence\_type').agg({'Residence\_type': 'count'}).sort(asc("count(Residence\_type)")).show()**

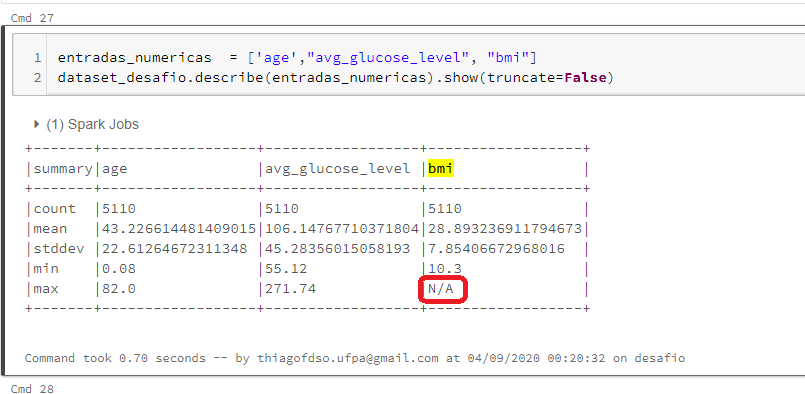


**display(dataset\_desafio.groupby('stroke').count())**



**entradas\_numericas = ['age',"avg\_glucose\_level", "bmi"]**

**dataset\_desafio.describe(entradas\_numericas).show(truncate=False)**

****

**Pergunta 6**

Qual é o valor da mediana para a variável do nível médio de glicose do

entrevistado (“avg\_glucose\_level”)?

(X) 95

( ) 120

( ) 78

( ) 271,74

**R.: Executar o comando e selecionar o boxplot para visualizar**

**display(dataset\_desafio.select('avg\_glucose\_level'))**

**Pergunta 7**

Analisando o padrão de dispersão da variável do nível médio de glicose do

entrevistado (“avg\_glucose\_level”) é correto afirmar, **EXCETO**:

( ) O terceiro quartil corresponde ao valor de 120. Desse modo, podemos dizer que 75% dos dados estão abaixo desse valor

(X) Existem valores que correspondem a possíveis outliers. Esses valores **certamente devem ser eliminados** do dataset, pois sempre causam problemas

( ) A dispersão dos dados no terceiro quartil é maior que no segundo, pois existe uma maior quantidade de valores diferentes no terceiro quartil

( ) Os possíveis outliers estão localizados após o limite superior do boxplot

**R.: Executar o comando e selecionar o boxplot para visualizar**

**display(dataset\_desafio.select('avg\_glucose\_level'))**

**Pergunta 8**

Analisando a dispersão dos dados para a variável idade (“age”), é correto afirmar, **EXCETO**:

( ) O maior existente para a idade dos entrevistados corresponde a 82 anos

(X) A mediana para essa variável corresponde ao valor de 68 anos

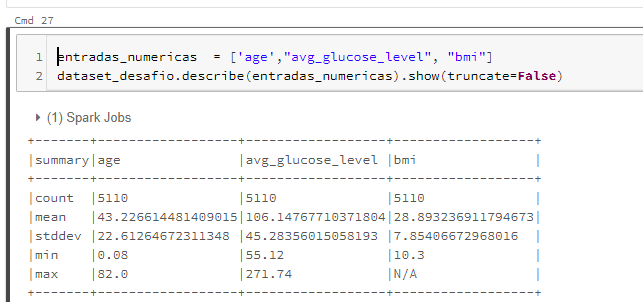
( ) Pelo Boxplot não é possível identificar possíveis outliers

( ) O primeiro quartil indica que 25% dos dados estão abaixo de 30 anos

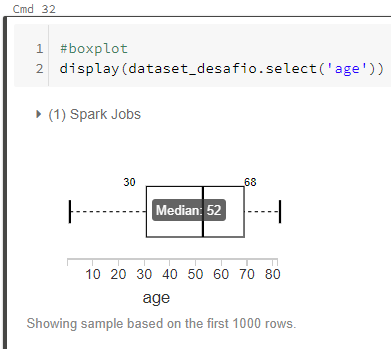
**R.: Executar os comando e selecionar o boxplot para visualizer**

**entradas\_numericas = ['age',"avg\_glucose\_level", "bmi"]**

**dataset\_desafio.describe(entradas\_numericas).show(truncate=False)**



**display(dataset\_desafio.select('age'))**



**Pergunta 9**

Quantas classes diferentes para a variável “work\_type” existem no dataset?

( ) 6

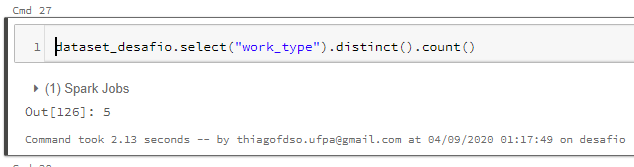
( ) 2

( ) 4

(X) 5

**R.: Executar o comando**

**dataset\_desafio.select("work\_type").distinct().count()**



**Pergunta 10**

Dentre as classes de tipos de trabalhos existentes (work\_type), qual é aquela que possui uma maior quantidade de instâncias?

(X) Private

( ) Never\_worked

( ) Govt\_job

( ) Self-employed

**R.: Executar o comando**

**dataset\_desafio.groupBy("work\_type").agg({'work\_type': 'count'}).sort("count(work\_type)",ascending=False).select("work\_type").show(1)**



**Pergunta 11**

Qual foi, respectivamente, o percentual de dados utilizados para o treinamento e teste do modelo?

( ) (20%, 80%)

(X) (80%, 20%)

( ) (30%, 70%)

( ) (70%, 30%)

**R.: Comando usado para divisão**

**train\_data, test\_data = model.randomSplit([.8,.2],seed=1)**

**Pergunta 12**

Analisando as variáveis “bmi” e “smoking\_status”, é **CORRETO** afirmar:

( ) Ambas são variáveis numéricas

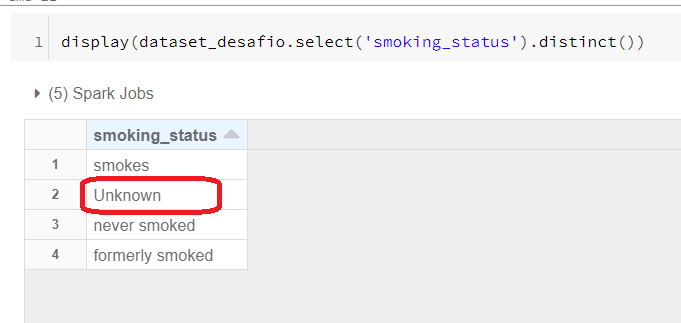
(X) Ambas possuem instâncias com valores desconhecidos

( ) Existem oito classes distintas de “smoking\_status”

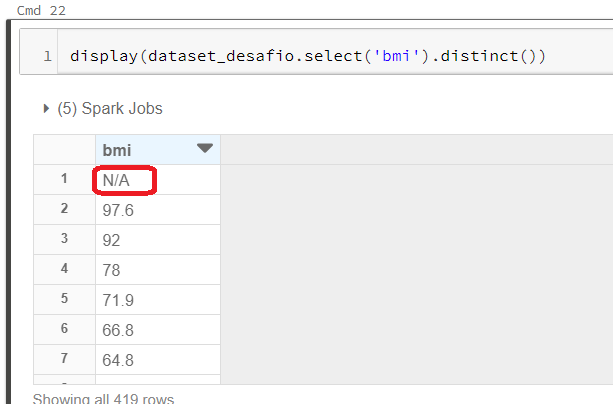
( ) A variável “bmi” possui apenas valores numéricos

**R.: Executar os comandos**

**display(dataset\_desafio.select('smoking\_status').distinct())**

****

**display(dataset\_desafio.select('bmi').distinct())**



**Pergunta 13**

Após o agrupamento dos dados de ‘smoking\_status’ e ‘stroke’, é **CORRETO**

afirmar que:

(X) Dentre os entrevistados que sofreram AVC, existem uma maior quantidade de indivíduos da classe que nunca fumaram (never smoked)

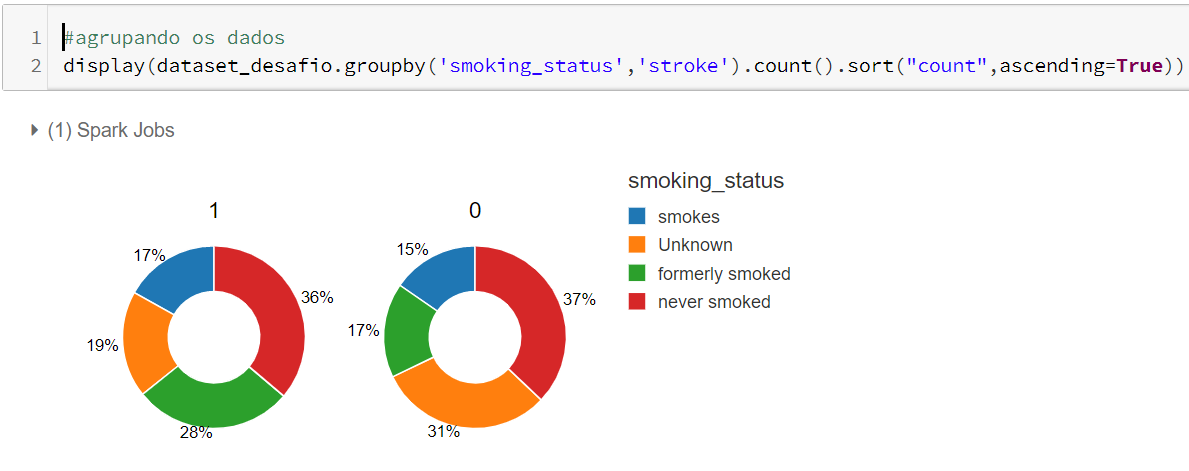
( ) Neste dataset existe uma maior quantidade de indivíduos que sofreram AVC

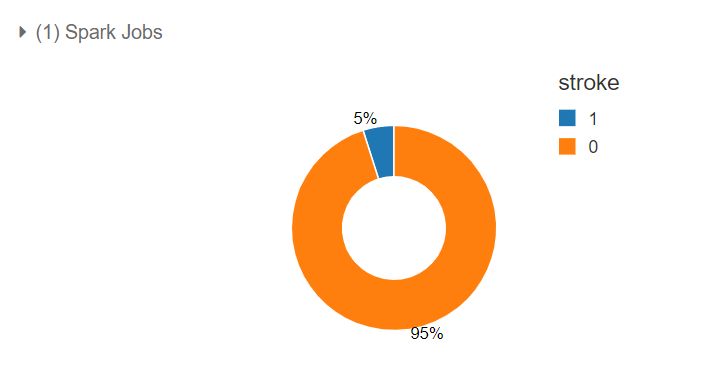
( ) Existem seis classes diferentes de “smoking\_status”

( ) Não é possível realizar o agrupamento, pois os dados possuem dimensões diferentes

**R.: Executar o comando e escolher o gráfico de pizza**

**display(dataset\_desafio.groupby('smoking\_status','stroke').count().sort("count",ascending=True))**





**Pergunta 14**

Sobre a relação entre a hipertensão (hypertension) e o AVC (stroke) presente

neste dataset, é **CORRETO** afirmar:

(X) A proporção de incidência de AVC é maior nos indivíduos que sofrem de hipertensão

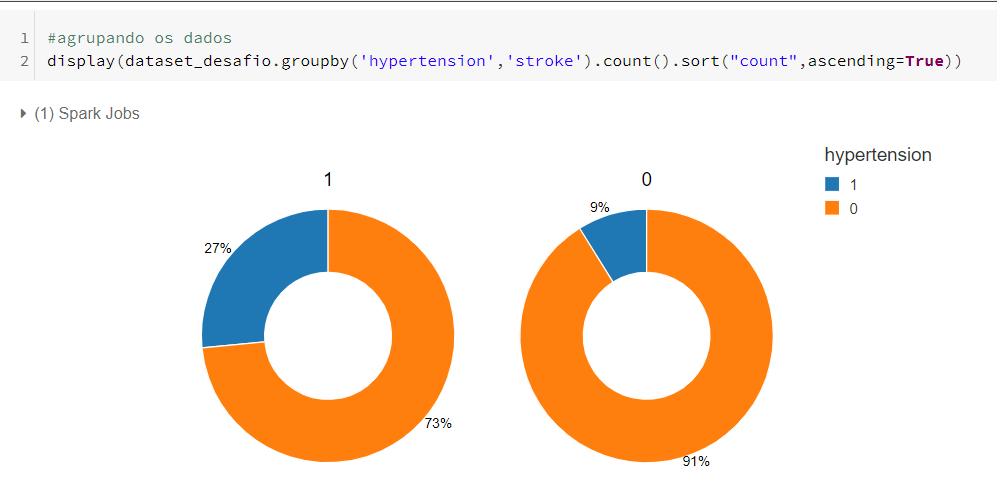
( ) Existe uma maior quantidade de dados de indivíduos não hipertensos

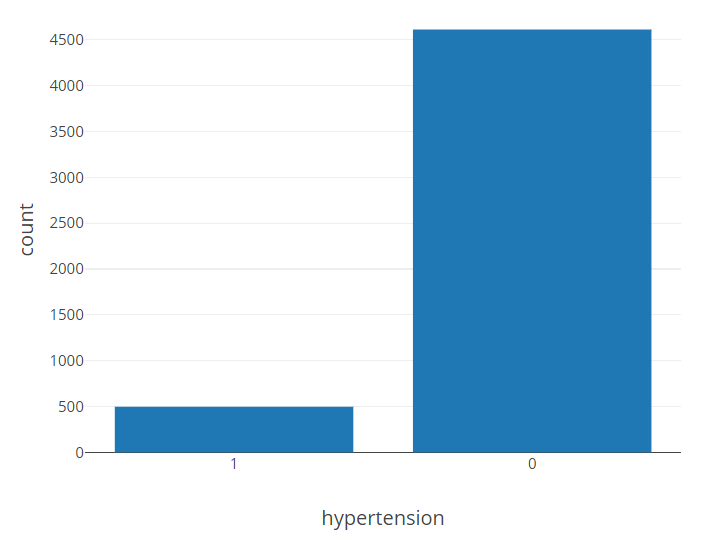
( ) A proporção entre indivíduos hipertensos e não hipertensos no dataset é a mesma

( ) Os dados mostram que este dataset está balanceado

**R.: Executar o comando e utilize os gráficos de pizza e barra**

**display(dataset\_desafio.groupby('hypertension','stroke').count().sort("count",ascending=True))**





**Pergunta 15**

Sobre o algoritmo de regressão logística aplicado para a previsão da ocorrência de AVC, é correto afirmar, **EXCETO**:

( ) A acurácia do modelo é superior a 90%

( ) Como o dataset está desbalanceado, a acurácia (accuracy) resultante pode estar enviesada

( ) A árvore de decisão também poderia ser aplicada para esse modelo de classificação.

(X) A regressão logística não deveria ser aplicada ao problema, pois ela trabalha apenas com dados categóricos.

**R.: Executar os comandos**

**#define o validador do modelo de classificação**

**from pyspark.ml.evaluation import MulticlassClassificationEvaluator**

**#avaliando a classificação realizada pela regressão logística**

**acc\_evaluator = MulticlassClassificationEvaluator(labelCol="stroke", predictionCol="prediction", metricName="accuracy")**

**acuracia\_regressao = acc\_evaluator.evaluate(previsao\_regressao)**

**print('Regressão Logística: {0:2.2f}%'.format(acuracia\_regressao\*100))**

