**Avaliação A1 - Nano Degree - Big Data Analytics e Processamento de Dados Massivos**

**Professor: Ramon Pereira**

**2020/2**

**Grupo:**

**Tatiana Resende**

**Nairah Matsuoka**

**Oswaldo Junior**

**Paulo Rossi**

**Carlos Henrique C Matos**

**Contexto**

O seu grupo foi selecionado para realizar uma análise de dados sobre Síndrome Respiratória Aguda no Brasil. Vocês deverão responder as questões abaixo relacionadas a doença utilizando uma base de dados histórica do ano de 2018.

O grupo deverá preparar uma apresentação descritiva da base de dados e com as respostas encontradas nas questões abaixo. A apresentação será realizada no final da aula.

Além da apresentação, o grupo deverá entregar esse documento preenchido com as respostas.

**Análise Descritiva**

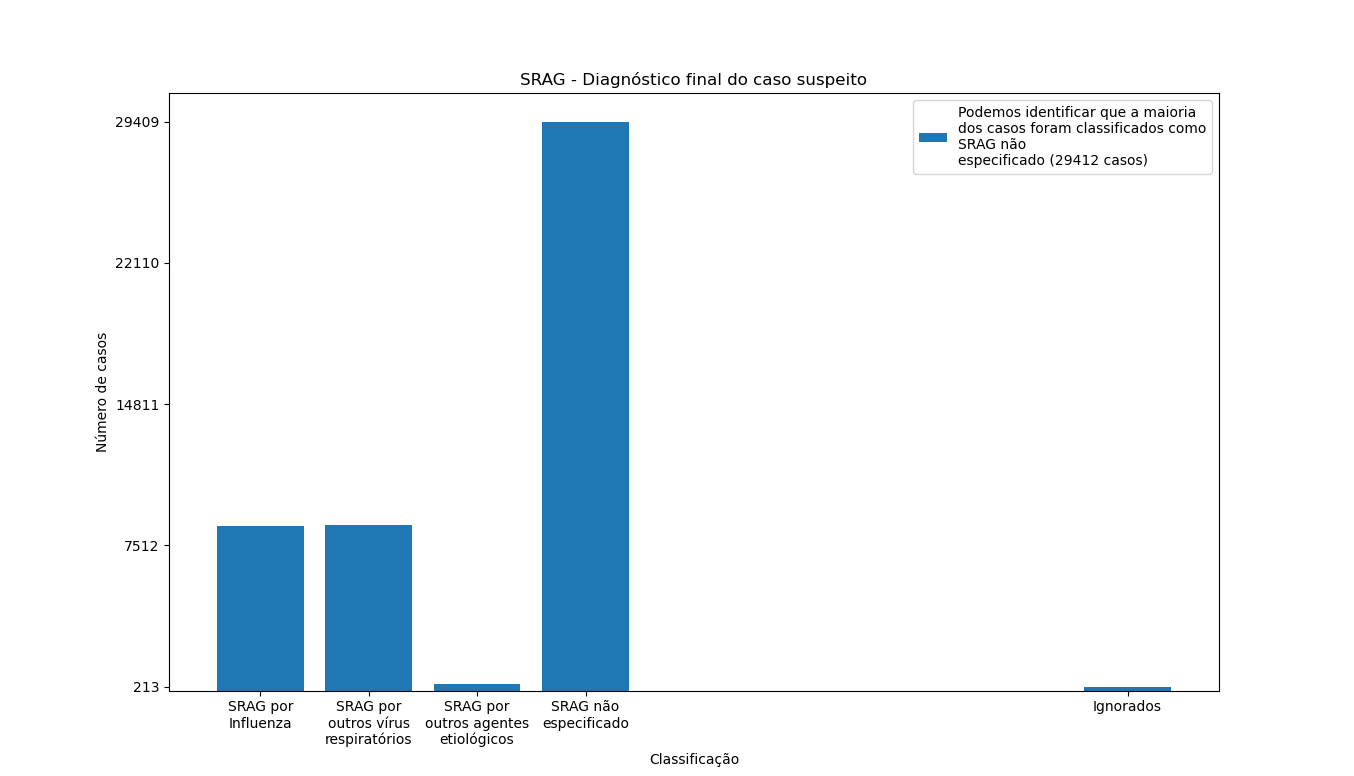
1. Assumindo o tamanho da base de dados atual, qual ferramenta e plataforma vocês recomendariam para o processamento destes dados e por quê?  
     
   *Pelo tamanho da base de dados (.CSV) a sugestão é o processamento no excel até mesmo pela facilidade e intuição da ferramenta. Porém pode ser facilmente implementada em qualquer linguagem indicada para análise de dados, tal como, python. Podendo assim, automatizar o processo de diagnostico futuro.*

1. Caso a quantidade de registros dessa base fosse aumentado em 100 vezes, a recomendação seria a mesma? Por que? Se não, qual seria a nova recomendação e por quê?  
     
   *Não. O processamento dos dados mesmo aumentado em 100 vezes, poderia ser feito em com phyton ou R utilizando um setup local. Vale a pena avaliação se no caso a base dados seria processada diretamente na máquina ou salva em um servidor na nuvem. Iria depender por exemplo do tempo disponível para a execução do trabalho.*
2. De posse da tabela, faça a identificação dos tipos das colunas dessa base de dados

1. Qual a quantidade de pessoas que estão acima do IMC ideal?

*1466 pessoa(s) estão acima do IMC ideal (de 24.9)*

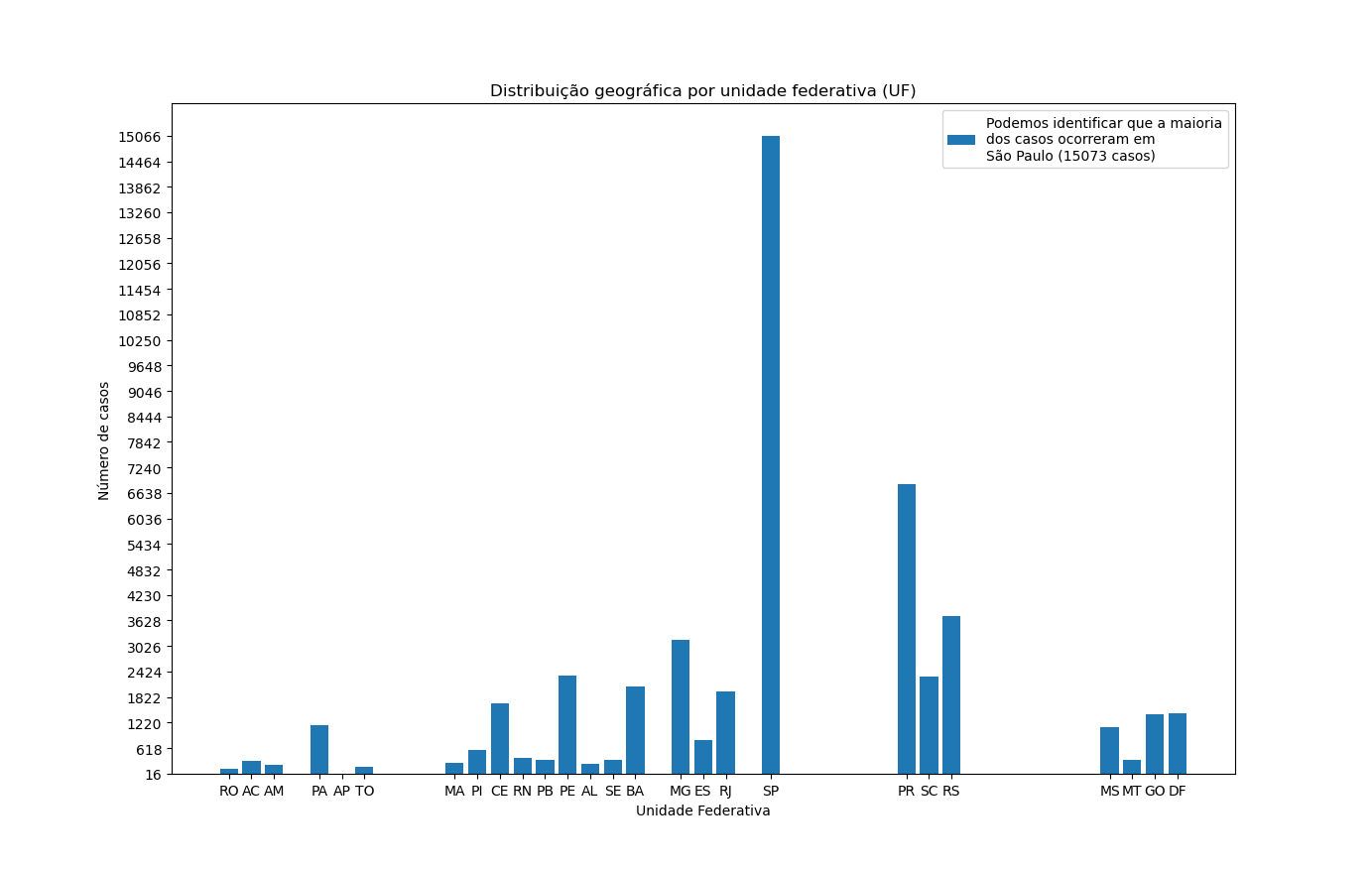
1. Faça um histograma da variável CLASSI\_FIN. O que esse histograma nos diz?



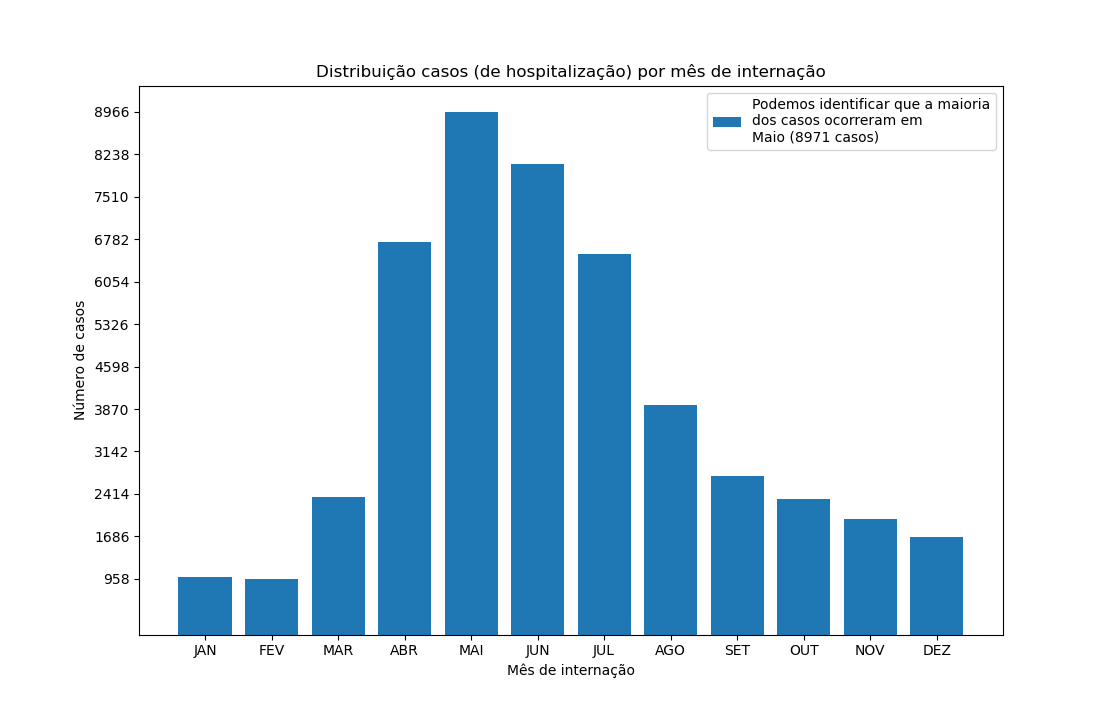
*Podemos identificar que a maioria dos casos foram classificados como SRAG não especificado (29412 casos)*

*- Existe uma categoria não identifica, sendo ela a categoria 9, representando um pequeno percentual, 0,4%.*

1. Quantas pessoas realizaram PCR do tipo convencional?  
     
   *Pela análise encontramos 712 registros*.
2. Faça o histograma da distribuição geográfica por unidade federativa (UF) dos pacientes com desta base de dados.



1. Faça um histograma por mês de internação dos pacientes



1. É possível identificar o número de mortes nesta base de dados? Se sim, coloque o número de mortos por SRAG em 2018 na sua frequência absoluta e na sua frequência relativa. Se não, qual estratégia o grupo utilizaria para conseguir estes números de forma real ou por uma aproximação boa?

*Sim, é possível identificar o número de óbitos através da evolução clinica. Identificamos o número de mortos pelos critérios de data de óbito e evolução, totalizando 5.191 mortes em 2018, sendo a frequência absoluta.*

*A frequência relativa em relação a cura e registros ignorados, totaliza em 35.239 registros em 2018.*

*12% de óbitos – frequência absoluta*

*87% de cura + ignorado – frequência relativa*

1. Qual a quantidade média de internações por SRAG no ano de 2018?

*No ano de 2018 ocorreram 46994 internações, aproximadamente 3916 por mês, 128 por dia.*

**Análise Preditiva**

1. Enumere os tipos de tarefas preditivas que podem ser realizadas com esta base de dados e dê exemplos com as colunas cada uma delas.

*Tarefas:*

*Prever índice de óbitos com base nas internações de uti;*

*Prever evolução clínica;*

*Alguns Campos:*

*33 - Ocorreu internação? (1 sim 2 não 9 ignorado) | HOSPITAL*

*41 - Foi internado em UTI? (1 sim 2 não 9 ignorado) | UTI*

*50 - Evolução clínica (1-Recebeu alta por cura 2-Evoluiu para óbito 9-Ignorado) | EVOLUCAO*

1. Seguindo os princípios da metodologia CRISP-DM, escolha uma das tarefas respondidas na questão número um e crie um ciclo de um modelo de análise preditiva de forma textual para ser explicado.
   1. *Levantamento do objetivo: Prever evolução clínica;*
   2. *ETL base dados*
   3. *Dicionário de dados | entendimento de cada variável*
   4. *Preparação: normalização e higienização dos dados*
   5. *Modelagem e Ferramenta: após os dados higienizados executar o processamento dos dados, carga de dados, redução de dimensionalidade e execução do modelo*
   6. *Avaliação do modelo: entender e avaliar os resultados gerados pelos modelos*
   7. *Implementação*