

Tempo: 10 - 20 min

Dificuldade: █ █ █ █ █

Fácil

Ponto principal do Snack:

Trabalhar com o Orange permitirá que você, no início, foque em entender os conceitos e aplicações de Data Science e não no ferramental.

criando o primeiro fluxo de trabalho orange

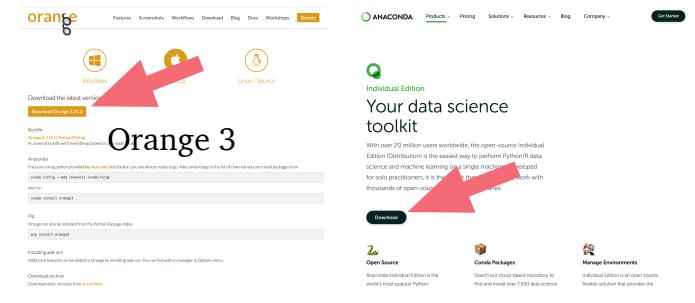
A escolha das ferramentas para se trabalhar com dados é de extrema importância para a produtividade, confiabilidade nos resultados e possibilidade de utilizar essas ferramentas no futuro em ambientes reais com dados reais. Nossa jornada começa com uma ferramenta cujo o propósito é "Foque no aprendizado!". Obviamente, para aquele que quiserem ir além, apresentaremos uma ferramenta que você provavelmente usará por toda sua vida profissional.

Objetivos de Aprendizagem:

- Possuir um ambiente pronto para o início do trabalho com dados e aprendizagem de máquina.
- Entender um fluxo de trabalho Orange.
- Reconhecer e executar um script em Python no Spyder.

ANTES DE COMEÇAR

- Baixar o **Orange 3** em <https://orange.biolab.si/download/>.
- Baixar o Anaconda em <https://www.anaconda.com/products/individual>.
 - Depois de instalar o anaconda, inicie ele e no Anaconda Navigator pedir para instalar o Spyder IDE.
- Baixa os arquivos de dados da competição Titanic do Kaggle em <https://www.kaggle.com/c/3136/download-all>.

**Fontes de Dados Titanic**

Nesse exemplo usaremos um arquivo de dados (ou dataset) popular chamado Titanic. Esse arquivo possui informações sobre as pessoas que embarcaram no Navio Titanic e marcações que informam se foi um sobrevivente ou não. Chamamos cada campo (coluna) do dataset de **feature** e entender as features, seus valores e como se distribuem é de imensa importância para a área de dados. Faça o download do dataset no link da sessão "Antes de começar" para iniciarmos o trabalho.

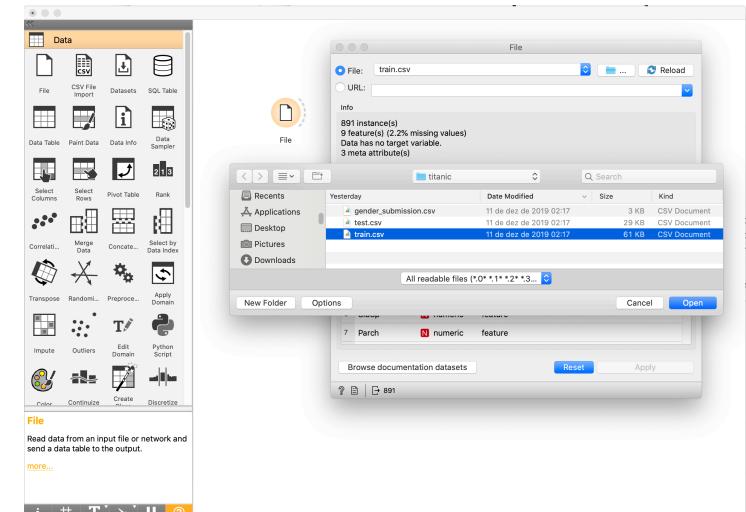
FEATURE	DEFINIÇÃO	DICIONÁRIO
Survived	Sobrevivente	0 = Não, 1 = Sim
Pclass	Classe do Ticket	1 = 1 ^a , 2 = 2 ^a , 3 = 3 ^a
Sex	Sexo	Numérico
Age	Idade	Numérico
Sibsp	Número de irmãos ou esposa a bordo.	Numérico
parch	Número de filhos ou pais a bordo.	Numérico
ticket	Número do Ticket	Numérico
fare	Preço da passagem	Numérico
cabin	Número da cabine	Alfanumérico
embarked	Porto onde embarcou.	C = Cherbourg, Q = Queenstown, S = Southampton

Conhecendo as features

Inicie o Orange e adicione um componente File na área de trabalho. Você acessa as propriedades dos componentes clicando 2 vezes nele.

Vá até o diretório onde você fez o download do dataset e carregue o arquivo train.csv.

Discutiremos mais tarde o que são cada um dos arquivos desse diretório.



Conhecendo (Cont.)

Agora que temos o **dataset** configurado podemos explorar seus dados. Vamos adicionar um componente chamado **Feature Statistics** (Estatística das Features). Uma forma de fazer isso é clicando no arco cinza a direita do componente File e arrastando ao soltar surgirá uma lista de busca de componentes, digite **Feature Statistics** e selecione ele na lista. Após isso, na caixa de seleção de cores, selecione a feature **Sex**. Você verá todos os histogramas agora com as barras coloridas de azul (Female ou Mulher) e vermelho (Male ou homem). Olhando para feature **Sex** sabemos percebemos facilmente que existiam mais homens que mulheres no Titanic. Ainda, uma informação importante é que não existem **missing values** (valores faltantes ou ausentes) para a feature **Sex** (Última coluna da tabela).



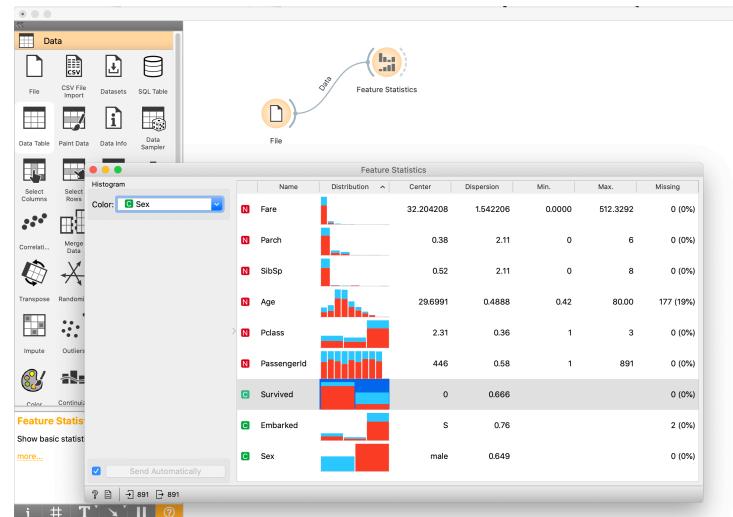
DESAFIOS: Usando apenas a visão de Feature Statistics identificar:

1. O sexo influenciou na sobrevivência dos passageiros.
2. Qual a classe de tickets com maior número de não sobreviventes.
3. Qual o porto de embarque com maior número de não sobreviventes.
4. A idade influenciou na sobrevivência dos passageiros?

Após carregado a lista de features, seus tipos, papéis e valores são listados.

Podemos ver 3, dos 5 tipos possíveis: **Text** - Cadeias de caracteres, ou strings; **Numeric** - Números; **Categorical** - Categorias ou enumeráveis. Temos nesse grupo campos com valores contáveis e de domínio finito como Sexo (Feminino ou Masculino) e Dias da semana (Seg, Terc, Qua, Qui, Sex, Sab, Dom) por exemplo.

Nessa etapa devemos configurar o papel (**Role**) da feature **Survived** para **target**. Ao fazer isso nossos classificadores saberão que essa é a feature a ser usada como classe.



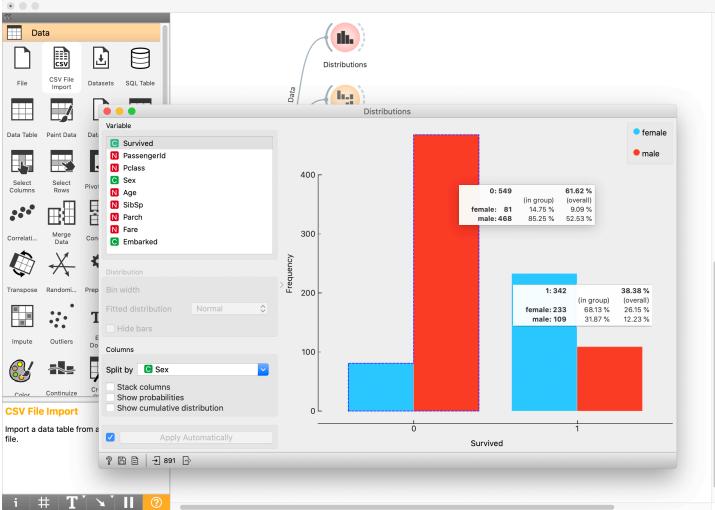
Uma forma outra forma de explorar os dados do **dataset** é com o componente **Data Table** (Tabela de Dados). Com ele podemos ter acesso a todas as linhas do dataset e suas informações. Muitas vezes é interessante navegarmos um pouco pelos dados em busca de informações que talvez nos ajudem na etapa de **Engenharia de Features** (Veremos isso em breve).

Uma dica MUITO valiosa é observar que todos os nomes possuem um pronome de tratamento. Essa informação aparentemente sem valor vai ser imprescindível para reconstrução de uma feature no futuro.

Apesar de ser muito difícil saber qual é essa feature agora e sem estudar mais a fundo o dataset, você consegue dar um

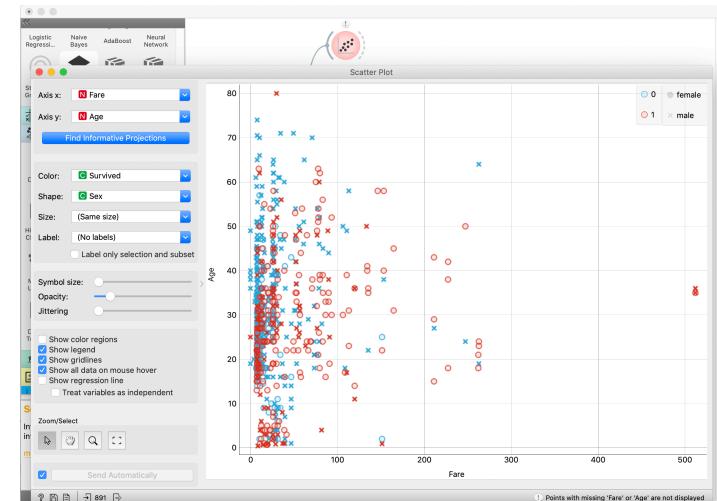
O entendimento completo do funcionamento de um Histograma é imprescindível para quem trabalha com dados. Invista algum tempo estudando histogramas. Sugestão de ponto de partida: [Wikipedia](#); Aula super didática: [Khan Academy](#).

Conhecendo as features



Agora vamos conhecer o componente **Distributions** (Distribuições). Ele é uma versão melhorada do **Feature Statistics**, porém só permite a visualização de uma feature por vez e um agrupamento. Na imagem ao lado respondemos, em detalhes, o desafio 1, o sexo influenciou sim a sobrevivência dos passageiros. Ao lado vemos que 85,25% dos passageiros mortos eram do sexo masculino. Esse comportamento também pode ser percebido nos sobreviventes. Vemos que 68,13% dos sobreviventes são do sexo feminino. Logo, o sexo teve uma grande influência na sobrevivência do passageiro.

Para esse Snack, o último componente que veremos é o **Scatter Plot** (Diagrama de Dispersão). No **Scatter Plot** podemos combinar várias features em busca de correlações ou comportamentos. Para nosso exemplo eu escolhi colocar o **Fare** (valor da tarifa do passageiro) no eixo X e **Age** (idade) no eixo Y. Ainda usei cores para representar se o passageiro sobreviveu (azul para não sobrevivente e vermelho para sobrevivente) e formas para representar o sexo do passageiro (círculos para mulheres e x para homens). Aqui temos dois comportamentos claros: o número de sobreviventes aumenta quando seu ticket é mais caro e pouquíssimas mulheres com tickets caros morreram.



DESAFIOS: Agora com ferramental maior que outros insights (ideias) poderíamos sugerir?

1. Em relação a classe a qual o passageiro pertence.
2. Em relação ao passageiro viajar sozinho ou não.
3. Outras?



GOSTOU DO MATERIAL?

Esse é um material de formação e também divulgação dos serviços em Data Science e Machine Learning prestados pela Leanworks. Foi produzido por um dos fundadores da empresa o [Prof. Antonio Luiz Cavalcanti](#) (Link do perfil no Linkedin para treinamentos e consultoria).

A Leanworks é uma empresa de inovação e transformação digital que foca em entrega de valor ao cliente. Na Leanworks desenvolvemos o fluxo completo de entrega de valor. Trabalhamos desde o mapeamento das oportunidades de transformação digital dentro da organização, a devida priorização de acordo com o impacto/retorno/investimento, o planejamento e execução dos projetos priorizados, o desenho de OKRs para medição dos resultados dos projetos e acompanhamento pós implantação para garantia dos resultados planejados.

Temos times prontos para:

- Desenho de negócios baseados em plataformas.
- Concepção, desenho e desenvolvimento de software em geral.
- Estabelecimento de Data Driven Company, com Data Science e Machine Learning.
- Estabelecimento de núcleos de inovação e endoempreendedorismo.

Se você possui uma demanda digital específica, nos consulte. Caso não possamos tratá-la conectaremos você a quem possa.