UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

DEPARTAMENTO DE COMPUTAÇÃO

GABRIEL TEIXEIRA CASCHERA

**CRIAÇÃO DE FERRAMENTA PARA AUXÍLIO NA VISUALIZAÇÃO E ANÁLISE EXPLORATÓRIA DE *DATASETS***

São Carlos

2019

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

DEPARTAMENTO DE COMPUTAÇÃO

GABRIEL TEIXEIRA CASCHERA

**CRIAÇÃO DE FERRAMENTA PARA AUXÍLIO NA VISUALIZAÇÃO E ANÁLISE EXPLORATÓRIA DE *DATASETS***

Monografia apresentada ao Departamento de Computação da Universidade Federal de São Carlos para obtenção do título de bacharel em Engenharia de Computação.

Orientação: Prof. Dr. Diego Furtado Silva

São Carlos

2019

GABRIEL TEIXEIRA CASCHERA

CRIAÇÃO DE FERRAMENTA PARA AUXÍLIO NA VISUALIZAÇÃO E ANÁLISE EXPLORATÓRIA DE *DATASETS*

Monografia apresentada ao Departamento de Computação da Universidade Federal de São Carlos, para obtenção do título de bacharel em Engenharia de Computação. Universidade Federal de São Carlos. São Carlos, XX de julho de 2019.

Orientador

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Dr. Diego Furtado Silva

Departamento de Computação - UFSCar

Examinador

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

XXXX

Departamento de Computação - UFSCar

Examinador

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

XXXX

Departamento de Computação - UFSCar

**DEDICATÓRIA**

*XXXXX.*

**AGRADECIMENTO**

XXXXX

**RESUMO**

XXXXX

Palavras-chave: X1.X2.

**ABSTRACT**

XXXXXX

Keywords: X1. X2.

**SUMÁRIO**

[1 INTRODUÇÃO 19](#_Toc3654907)

[2 ESCOLHA DE LINGUAGEM DE PROGRAMAÇÃO PARA DESENVOLVIMENTO 19](#_Toc3654908)

[3 LEVANTAMENTO DE BIBLIOTECAS DE VISUALIZAÇÃO DE DADOS NO PYTHON 19](#_Toc3654909)

1. INTRODUÇÃO

Devido ao grande número de aplicações na área de *Data Science* atualmente, é cada vez mais importante a análise e preparação dos dados para estas aplicações. Grande parte do tempo de construção e teste de modelos estatísticos para essas ferramentas de *Machine Learning* é gasto justamente nesta etapa de entendimento dos dados.

Em casos de projetos cujos dados não possuem tanta qualidade, isto é, estão em desacordo com o fenômeno a ser estudado/previsto ou estão faltando informações, é de suma importância conhecer os dados no início do projeto. Por exemplo, ao se construir um modelo para seleção de um canal de cobrança para cada cliente de uma base de dados, o primeiro passo é entender se os dados de fato permitem que esse modelo seja aplicado.

Deste modo, é proposta uma ferramenta para facilitar o rápido diagnóstico de *datasets*, através da apresentação de dados básicos do *dataset* e da fácil criação de visualizações de informações gráficas pelo usuário.

1. ESCOLHA DE LINGUAGEM DE PROGRAMAÇÃO PARA DESENVOLVIMENTO

XXXXX

1. LEVANTAMENTO DE BIBLIOTECAS DE VISUALIZAÇÃO DE DADOS NO PYTHON

Matplotlib

Primeira biblioteca dedicada à visualização de dados no Python, pode ser utilizada para a geração de vários tipos de gráficos, desde histogramas até gráficos de linha. Consegue ser utilizada para criação de gráficos com poucas linhas de código, facilitando muito o desenvolvimento de *software*. Suas visualizações não são dinâmicas, o que dificulta a simples interação do usuário com os gráficos gerados.

Seaborn

Construída em cima do Matplotlib, apresenta maior facilidade para utilização de *plots* mais complexos, como mapas de calor e gráficos violinos, por ser uma biblioteca mais alto nível. Além disso, suas cores e apresentações são consideradas mais bonitas que aquelas da biblioteca anterior. Como o Matplotlib, suas visualizações são estáticas.

ggplot

Baseado nas ideias do livro *The Grammar of Graphics* e na biblioteca *ggplot2* do R, o *ggplot* permite a criação de gráficos menos personalizados, focando na reutilização de código e aceleração da geração dos gráficos. Esta biblioteca conversa muito com a biblioteca de processamento de dados *Pandas* e seu modo de guardar os dados em *dataframes*. Sua implementação faz com que seja possível criar gráficos em camadas, adicionando *plots* um em cima do outro.

Bokeh

Implementada com base no *The Grammar of Graphics*, como o *ggplot*, é uma biblioteca nativa do Python. Permite a criação de gráficos interativos, com ferramentas de *zoom*, movimentação, atualização dos dados e opção de salvar a visualização. Permite que os *plots* sejam exportados em JSON, documentos HTML ou em interfaces *web*. Os dados a serem apresentados podem ser fornecidos em listas do próprio Python ou estruturas de dados do *NumPy* e *Pandas*.