Centro Universitário Carlos Drummond de Andrade

Marcos Dias Barbosa: RA_ 192001090

Alicya Danielle Holanda Silva: RA_ 999001025

Antony Machado Ferreira da Silva: RA_ 192000856

Gustavo Rodrigo Cruz Nunes: RA_ 192001161

Estruturação de dados: Adult-Sensus-Income

São Paulo - SP

2021

Centro Universitário Carlos Drummond de Andrade

Estruturação de dados: Adult-Sensus-Income

Relatório Técnico-Científico apresentado na disciplina de Projeto Integrador para o curso de (Análise e Desenvolvimento de Sistemas) do Centro Universitário Carlos Drummond de Andrade (UNI DRUMMOND).

São Paulo - SP

2021

DIAS, Marcos; DANIELLE, Alicya; MACHADO, Antony; RODRIGO, Gustavo. **Estruturação de dados: Adult-Sensus-Income.** 00f. Relatório Técnico-Científico. Análise e Desenvolvimento de Sistemas – **Centro Universitário Carlos Drummond de Andrade**. Tutor: (ME. Eduardo Palhares Júnior). Polo:(Ponte Rasa), 2021.

RESUMO

Nesta pesquisa, analisaremos uma estruturação de banco de dados: adult-sensus-income, no qual usamos como principal executor dos códigos, o Google Collaboraty (google colab), para que, através dele, possamos entender o funcionamento deste conjunto de dados. Usando como base a linguagem de programação Python, essa estruturação de banco de dados foi destrinchada para testarmos seus códigos em estruturas individuais, mas que irão se interligar como um todo, tais como: suas bibliotecas, discretizações, gêneros, dados de país, remoção de dados etc. Ao final, iremos ter uma visão clara de sua execução e seu resultado de um modo objetivo de acordo com seu funcionamento.

Palavras-Chave: Google Colab; Python; Adult-sensus-income.

LISTA DE ILUSTRAÇÕES

SUMÁRIO

1. INTRODUÇÃO

Neste trabalho iremos analisar uma base de dados, demonstrando os códigos com os quais usaremos para fazer buscas, discretizações de dados, remoções etc. Posteriormente tendo um resultado final desta análise, trabalhando para que o código não contenha linhas de erro e nos permita ter um código funcional.

Collab:

O termo **collab** é o diminutivo da palavra inglesa **collaboration**, que em português significa colaboração. **Collab** é uma prática utilizada por diversos segmentos com a intenção de unir nomes, personalidades, marcas e outros, para fornecer diferentes serviços ou produtos para um certo público.

O tema proposto foi introduzir uma base de dados, reorganizar esta base e fazer um estudo sobre ela e aplicá-la em um programa viável, no caso como já citado, o Google Collaboration. Ao aplicarmos essa base de dados em um programa, vamos agrupá-los em vários blocos de sequências de códigos de acordo com seus grupos para poder deixar organizado, de uma maneira que se possa entender como a coleta de dados foi realizada. A base de dados é baixada de um site chamado kaggle, em arquivo zipado, onde é extraído para uma pasta e assim poder ter acesso ao banco. A princípio, os dados foram colocados em arquivo Excel, que estava totalmente desorganizado, no qual fizemos uma edição para ser interpretado e, assim, ser submetido ao google collab. Nas informações básicas, podemos citar que, foram selecionadas diversas pessoas americanas com diferentes características sociais, econômicas, étnicas, rentáveis, estudantis etc.

Feito isso, importamos esta base de dados ao google drive, onde foi designado à plataforma pela qual usamos para codificar, usando a linguagem Python como referência e termos um retorno com isso às informações desejadas. Nesta plataforma, importamos as bibliotecas que seriam usadas para expressar os códigos. Nota-se aqui, que é de extrema importância que essas bibliotecas sejam anexadas no início do bloco de códigos para que se possam subir todas as devidas informações contidas no banco de dados. Feito isso, se tornou mais acessível a estrutura do código que, por final,

nos apresentou o resultado que era esperado. A ideia básica deste documento é reestruturar esses documentos em blocos nos quais se encaixam em seus devidos locais de relação de códigos. Sendo assim, cada informação deste dataset, irá ser reescrita nesses blocos que são divididos em parte. A pesquisa em si, se trata de um conjunto de pessoas que foram selecionadas para que fosse possível fazer levantamento de suas características socioeconômicas e demográficas sob os quais se obtém uma estimativa específica da população. Nesta pesquisa, os motivos foram particularmente baseados numa pesquisa feita por nosso professor, com a qual fizemos uma análise do conteúdo e tiramos algumas conclusões deste para dar origem ao documento aqui apresentado. O objetivo do trabalho, como já dito é coletar as informações propostas pelo autor e esboçar uma breve apresentação desses dados, que são as características de um determinado número de pessoas americanas selecionadas, para que obtivesse um resultado final de estimativa de quantas chegam a um valor próximo ou equivalente de 50k por ano. Então a coleta desses dados foi fundamental para que a base de dados e os cálculos usados em linhas de código em python pudessem nos dar todos os resultados mostrados em cada bloco de dados importados.

2. DESENVOLVIMENTO

2.1 Objetivos

Nosso objetivo é descrever nosso processo de montagem, identificando os erros preexistentes e traçando soluções para a resolução desses erros, descrever as estruturas utilizadas para análise deste código, posteriormente chegando a um resultado bem sucedido no final.

2.2. Justificativa e delimitação do problema

Durante a Produção dessas linhas de códigos, foram encontrados erros inesperados, onde nos impediram de ter um resultado positivo, nossa equipe se uniu para que juntos conseguíssemos encontrar soluções para que esse código pudesse ser efetivamente funcional. Todavia, nossos conhecimentos não foram totalmente eficazes para que solucionasse esses devidos erros. Por este motivo, acabamos por recorrer a outros métodos que fosse capaz de resolver essas incongruências. Um dos métodos mais eficazes que utilizamos foram algumas pesquisas que pudessem nos servir como apoio, e serviu, só que ainda assim nos faltou ferramentas solucionáveis para taparmos essa lacuna que ainda permanece.

2. 3. Fundamentação teórica

Com as bases já coletadas, usamos como fontes confiáveis através do site kaggle.com, que é uma plataforma conhecida na área da tecnologia e é muito usada por desenvolvedores de diversos segmentos. Conhecida por ter a oportunidade de fazer competições de Machine Learning, explorar e publicar conjuntos de dados e também com possibilidades de treinamentos. Dito isso, o dataset do qual nos referimos foi retirado desta, podemos dizer uma plataforma conceituada na área tecnológica.

O arquivo baixado é compactado para que não seja tão dificultoso o download do mesmo para o estudo. Além dessa fonte utilizada, podemos contar com alguns documentos e vídeos aulas propostos por nosso professor responsável por este trabalho. Algumas questões que não encontramos no kaggle, os vídeos foram os responsáveis por nos ajudar a suprir dúvidas e questões não tão esclarecedoras. O arquivo, então, pode ser interpretado num compilador de códigos como o vscode, spider e até mesmo o sublime text. Mas por final, usamos mesmo o google colab, que não nos apresentou nenhuma dificuldade em rodar, importar, reestruturar e digitar os códigos.

2.4. Aplicação das disciplinas estudadas no Projeto Integrador

Algumas implementações usadas na construção desse projeto, foi o uso de algumas matérias que nos ajudou a ter uma síntese melhor e compreensível deste tema abordado. Olhando para o dataset, vemos que existe nela uma base desestruturada de dados coletados, fornecidos pelo kaggle. Nas aulas que usamos como acréscimo de informações.

Este item do referencial teórico deve indicar os conteúdos das disciplinas estudadas no curso que foram abordados no projeto. Espera-se que os estudantes relacionem, de forma clara e coerente, o conteúdo estudado à solução desenvolvida durante o projeto. Estruturação de dados baseado em python, programação orientada a objetos usando Java, são as disciplinas que podemos usar para incrementar ainda mais essa dataset. No entanto, como tivemos uma base de dados gravados em uma vídeo aula, foi de grande auxílio, já que o trabalho estava encaminhado. Mesmo tendo isso, observamos que a estruturação desses dados fora também estudada em aula, o que nos fez ter um nível de interpretação melhor.

2.5. Metodologia

Nesta metodologia, usamos ferramentas básicas para a execução de cada pedaço dos códigos. Além de instrumentos simples, mas bem funcionais, contamos com os conceitos e ensinos que foram adquiridos no decorrer do curso. Métodos como esse foram objetivos e claros para a realização. Desde o site específico para a coleta dos dados, desde sua construção na plataforma do google colabora. Para esse contexto realizado, tivemos que fazer divisões sobre o que cada participante iria fazer. Como obtemos o esboço de tratativas sobre a ideia do bando de dados, abordamos técnicas para desenvolver pequenos traços relevantes de expressar uma colocação clara do intuito deste trabalho. No mais, contextualizando, a integração de pessoas que tenham estimativas diferentes compõem este dataset.

Com algumas estratégias de diversificar cada espaço de funcionalidade de códigos, fizemos parte por parte, no sentido incremental, de cada bloco. A importação da biblioteca, como dito sendo necessária para a debugação e ligações de cada trecho para não haver erros, a discretização, a análise das variáveis, para se saber se todas estão corretamente declaradas para não ocorrer de algum trecho não rodar, os dados faltantes, os natives country, etc. foram pontos estratégicos para a delimitação correta dessa base de dados. Cada participante, tem um pedaço ideal na construção do projeto, uns mais expostos, outros mais reservados, mas que foram trabalhando em conjunto e mantendo as ligações sempre em dia.

Criar / Prototipar:

A análise de dados foi estudada em vídeo, e após esse estudo foram colocados em prática os conhecimentos de forma qualitativa, no qual há uma necessidade de qualidade na elaboração do projeto pesquisado em si. Em outra parte, na ideia de quantitativa, tivemos um arquivo demasiado grande para executar essa estruturação de dados. No entanto, usando as estratégias, foi possível reverter essa grande quantidade de informações e se enquadrar na organização da pesquisa. A

Descrição das soluções encontradas ou desenvolvidas para o problema investigado.

Com alguns problemas encontrados, sempre estávamos nós recorrendo ao conhecimento individual e coletivo para a resolução do mesmo. Não sendo possível resolver essa questão entre a

equipe, fomos recorrer ao vídeo disponibilizado e alguns websites voltados para a solução do problema.

Implementar / Testar:

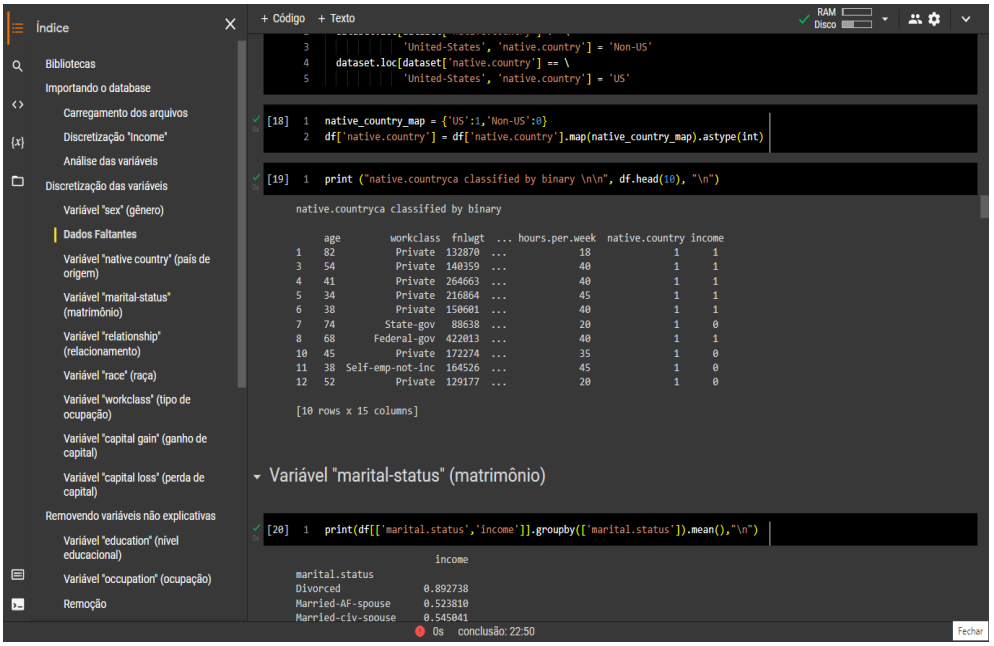
A solução do problema que foi encontrado, foi testado na plataforma usada desde a implementação do banco de dados disponibilizado. Ao encontrar a possível solução, já fizemos a aplicação dela direto no código estruturado, pois por lá também já conseguimos executar e ter a certeza se irá ou não funcionar tal mudança no código. As devolutivas foram positivas, apesar de algumas dificuldades que tivemos em alguns códigos. Tendo auxílio de alguns vídeos no youtube, acessando diferentes blogs e páginas relacionadas ao dataset, foi possível converter os erros. Mesmo assim, a gama maior de acertos foi do vídeo aula que também nos foi deixava para ser usado como forma de acesso fácil de correção caso fosse necessário. E foi, então, para se concluir, esta base de dados foi útil na questão de se entender de que forma as pessoas são remuneradas nos EUA.

Que melhorias foram indicadas para as soluções propostas/desenvolvidas?

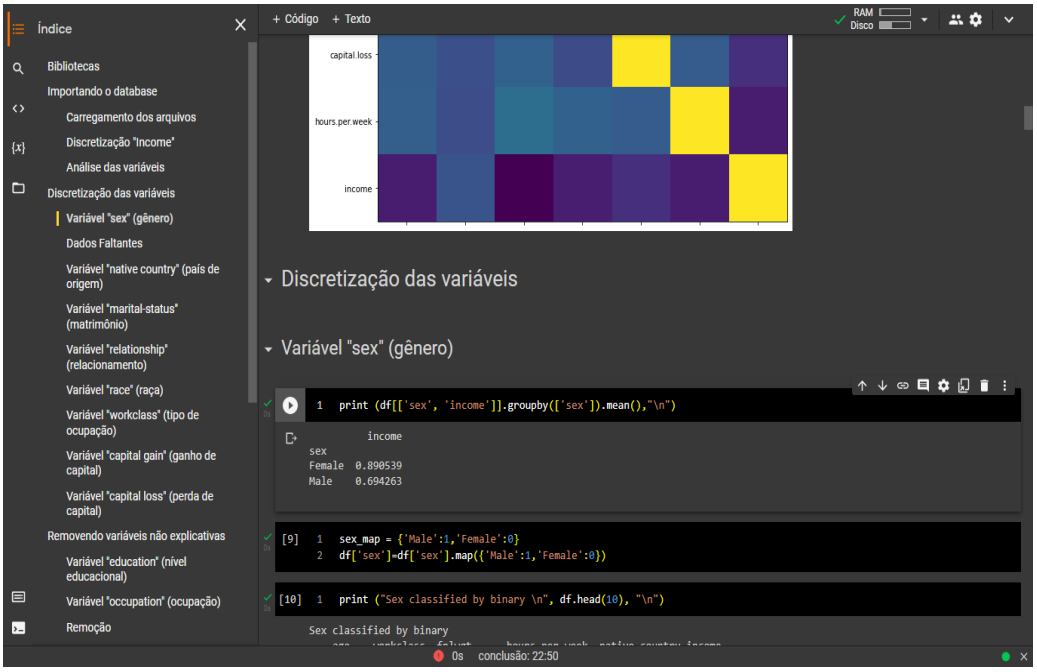
Como tínhamos já um projeto feito, com a necessidade apenas de digitar esses códigos e fazê-lo rodar de modo correto, viu-se que houve problemas de sintaxe e concatenação com o restante do código. Por aí, tivemos uma noção de que seria necessário aplicar algumas melhorias em certas linhas de código e até mesmo certas mudanças para que todo o conjunto rodasse adequadamente.

3. DESENVOLVIMENTO E AVALIAÇÃO DO TRABALHO

Este conjunto de dados é uma coleção de informações relacionadas a uma pessoa. A tarefa de previsão é prever se uma pessoa está ganhando um salário acima ou abaixo de $50 mil. Exploramos extensivamente este conjunto de dados no intuito de analisar com o máximo de cuidado, as informações de cada habitante do país, checando assim se todos dentro dos parâmetros tem uma participação fundamental. Os dados foram pegos de pessoas com idade de 17 a 90 anos, incluindo a classe de trabalho, governo, peso final, educação, estado civil etc. Com isso esses conjuntos de dados é possível verificar o número de pessoas dentre as diferentes faixas etárias, sexo, ganho de capital, horas trabalhadas, país nativo dentre outras coisas.



Começando pelo capital ganho por pessoa, podemos observar que o ganho dos $50 mil são mais predominantes do que por pessoas que ganham menos que este valor. Já as pessoas mais velhas são prováveis que ganhem mais de $50 mil do que os jovens.



Na educação, podemos ver que as pessoas com diploma ganham mais que as pessoas que não possuem. Ex: apenas uma pequena proporção de pessoas com menos de 12 anos de educação ganha mais de $50 mil por ano. Essa proporção aumenta quase que linearmente em 12 anos de escolaridade. Portanto, além de um diploma universitário, o que você pode fazer para aumentar suas

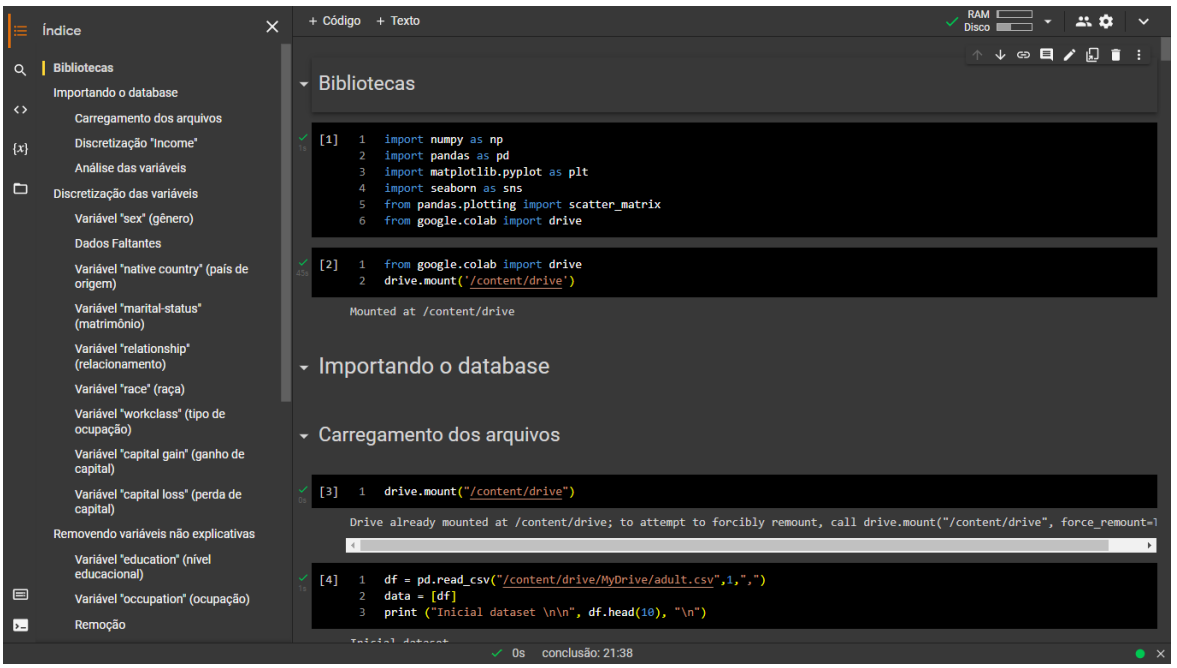
chances de ganhar mais de $50 mil é maior. Também podemos observar que, pessoas que são casadas têm uma renda superior a $50 mil. Mas somente 47 pessoas tinham um cônjuge. Sobre gênero e raça, os homens têm muito mais probabilidade de ganhar mais de $50 mil anuais do que as mulheres. Mais trabalho pode ser feito usando esses dados para analisar as razões por trás dessa desigualdade de renda. Como avaliação, podemos dizer que os conjuntos de dados foram favoráveis para se distinguir a questão de idade, raça, gênero, educação e pais, como uma das formas de se avaliar uma pessoa no país.

3.1. APRESENTAÇÃO DO CENÁRIO

O que é o Google Colab?

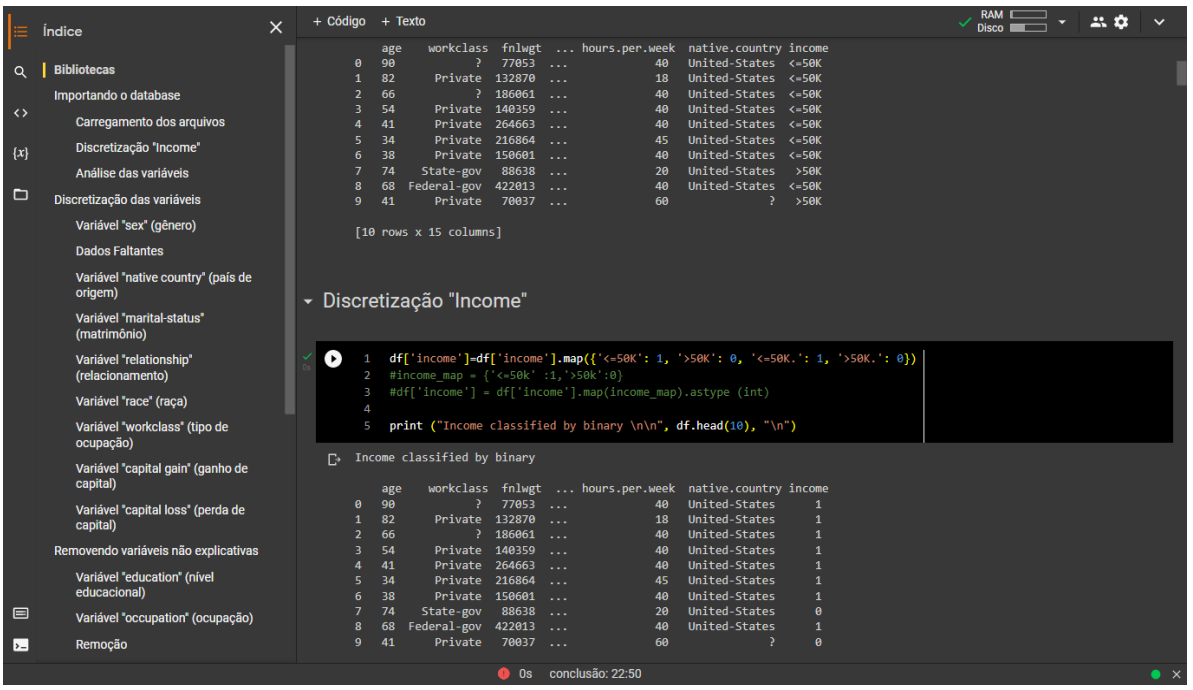
Google Colab, ou Google Collaboratory, é um serviço de armazenamento em nuvem de notebooks voltados à criação e execução de códigos em Python, diretamente em um navegador, sem a necessidade de nenhum tipo de instalação de software em uma máquina. Em outras palavras, com o Google Colab você é capaz de ler, desenvolver e rodar códigos e rich texts em documentos interativos que agrupam células de códigos — chamados de notebooks —, compartilhá-los com outros programadores, modificá-los a quaisquer momentos e mantê-los salvos de maneira totalmente online. Todo o poder computacional utilizado para executar o software que você escreve é fornecido pela nuvem de computadores da Google. Dando gratuitamente ao usuário a possibilidade de processar uma quantidade bem grande de dados; segundo Ugo Roveda Co-fundador e COO na Kenzie Academy Brasil.

3.2.1 BIBLIOTECAS



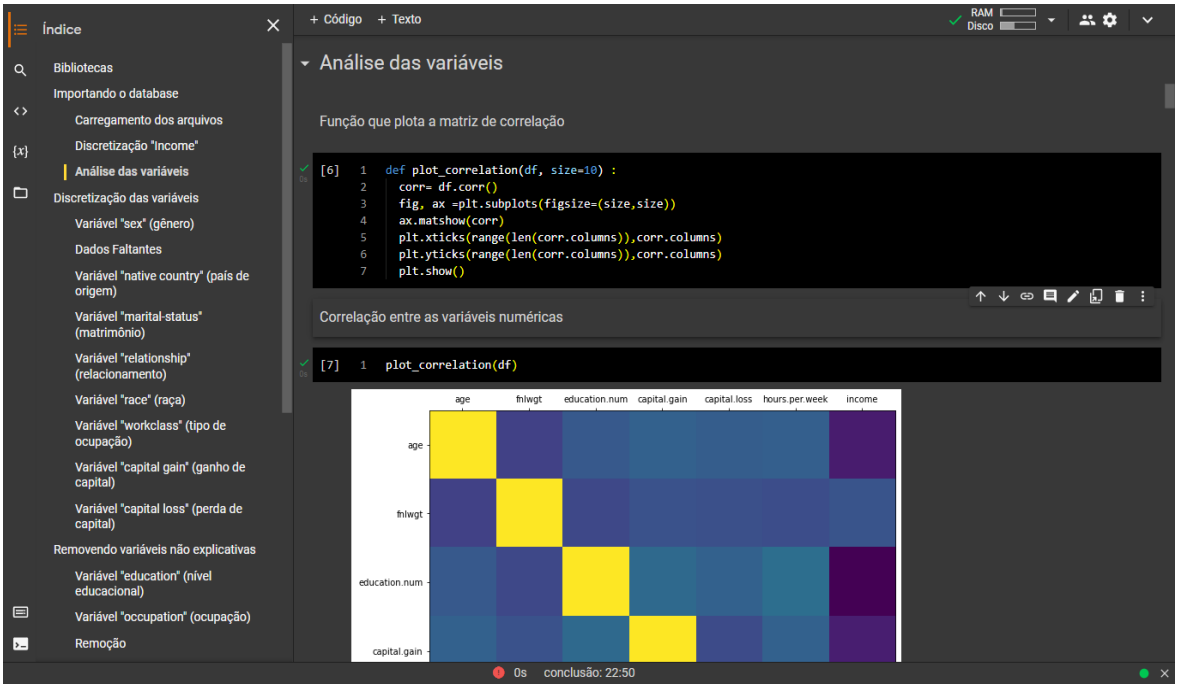
O aumento na quantidade de dados gerados e a necessidade que algumas aplicações têm de analisar esses dados em tempo real estão entre as principais demandas para bancos de dados em memória. Nesta imagem 01 você verá o que leu na página 06 desta pesquisa.

3.2.3. DISCRETIZAÇÃO INCOME



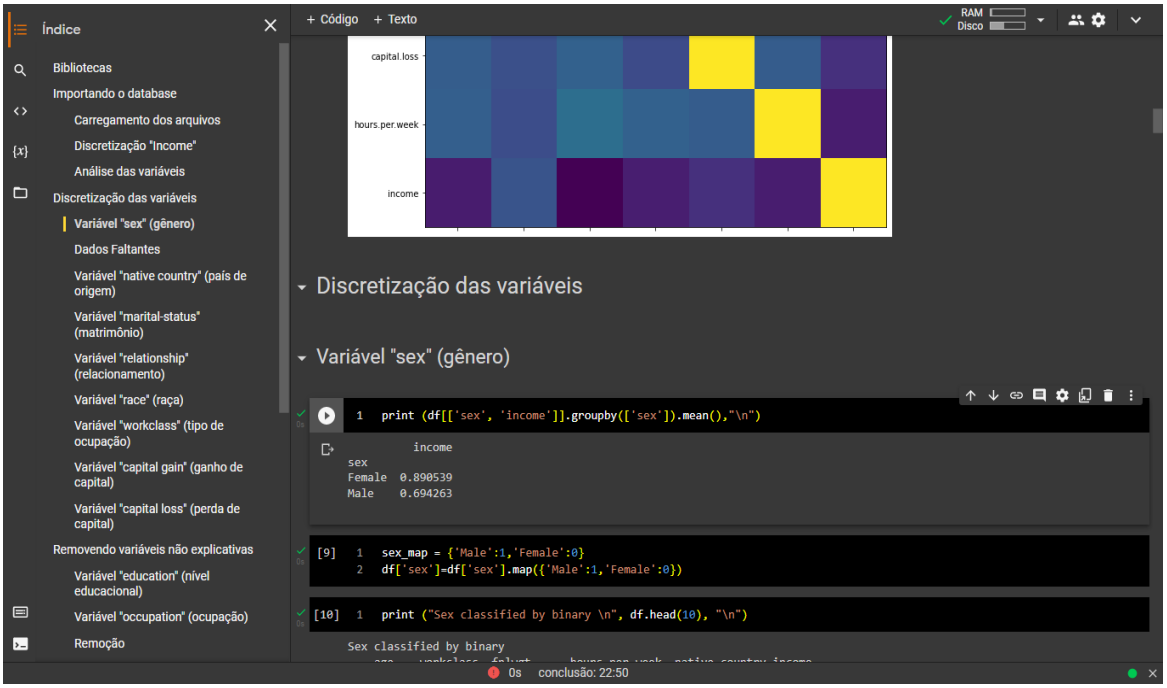
Como, nesta imagem 02 Tornar discreto ou descontínuo; transformar uma distribuição contínua em unidades individuais.

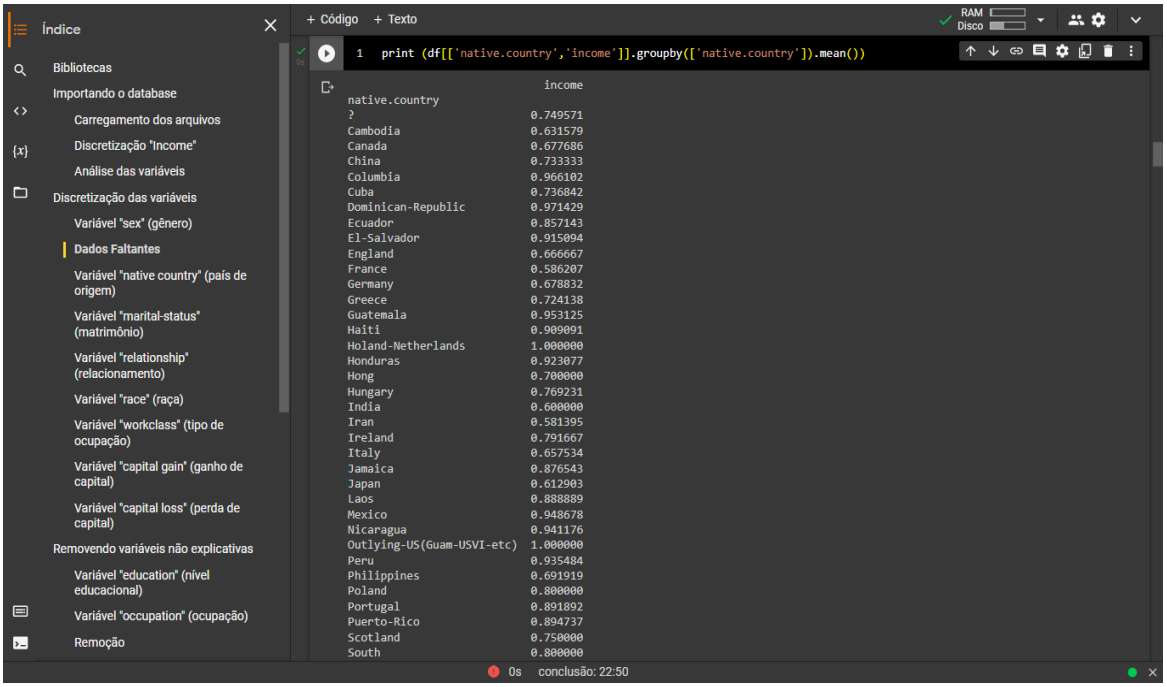
3.2.4. ANÁLISE DAS VARIÁVEIS

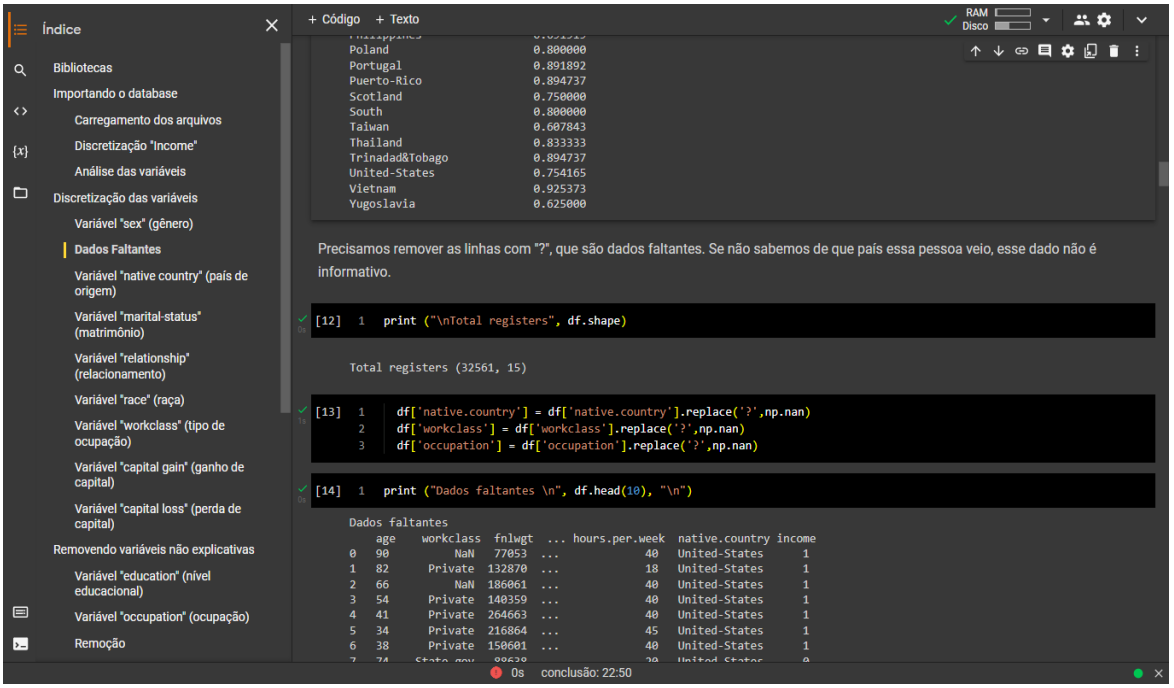


As variáveis podem ser classificadas em quantitativas ou qualitativas. Uma variável quantitativa possui como especificidade ter como atribuição um valor numérico. Uma variável qualitativa possui como especificidade ter como atribuição um valor não-numerado.

3.2.5. DISCRETIZAÇÃO DAS VARIÁVEIS

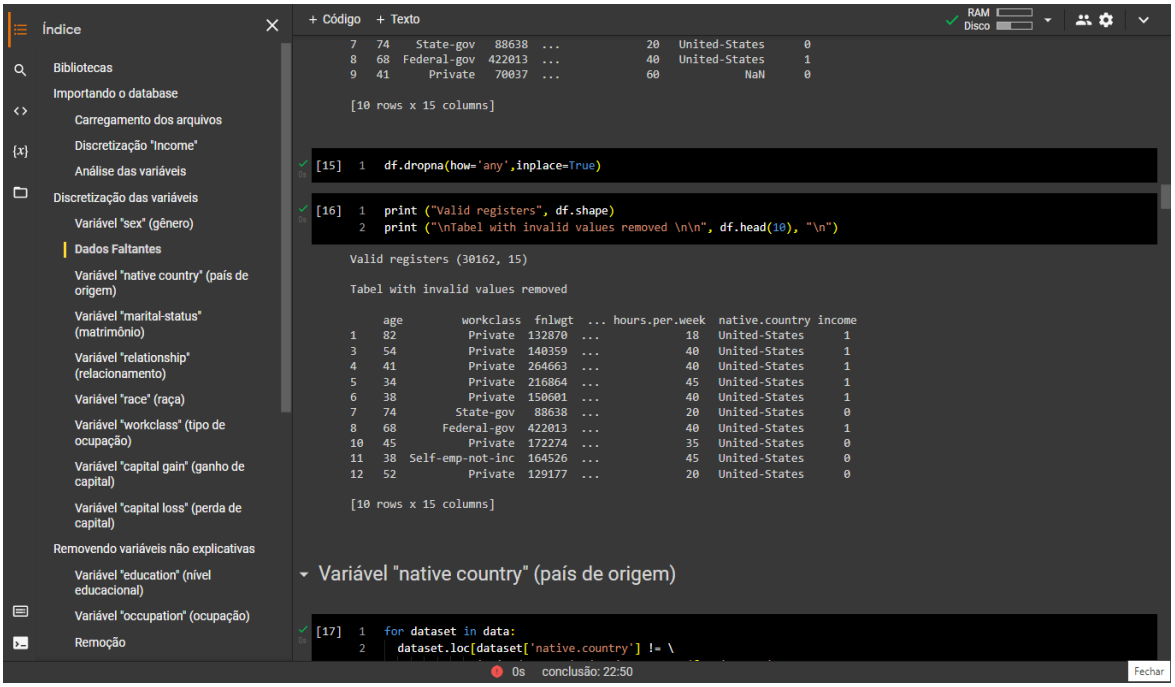






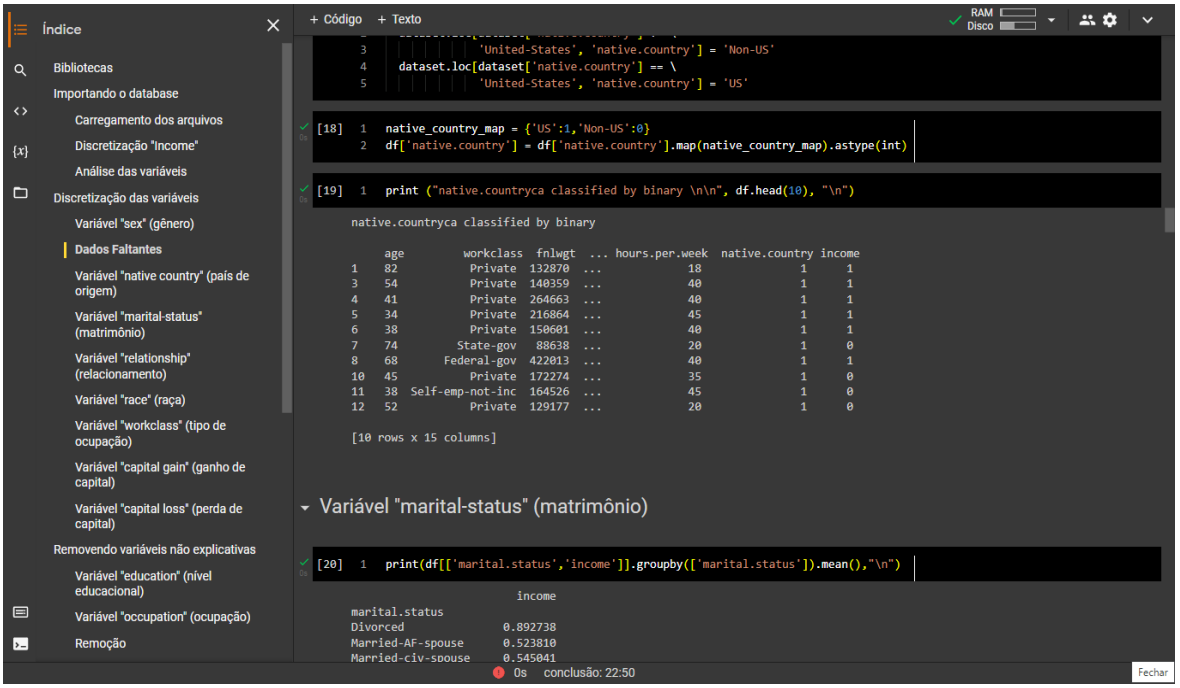
A escala ordinal é definida como uma escala de medição de variáveis utilizada para simplesmente representar a ordem das variáveis e não a diferença entre cada uma das variáveis. O que vimos nas imagens 04,05 e 06.

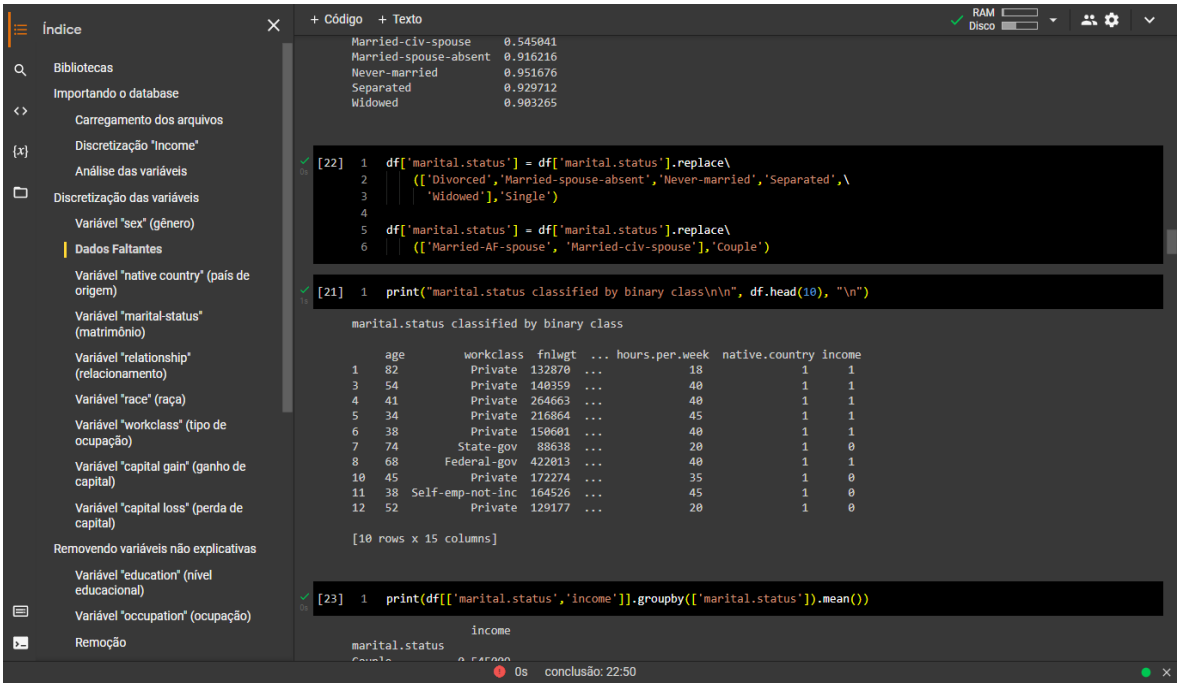
3.2.6. VARIÁVEL NATIVE COUNTRY

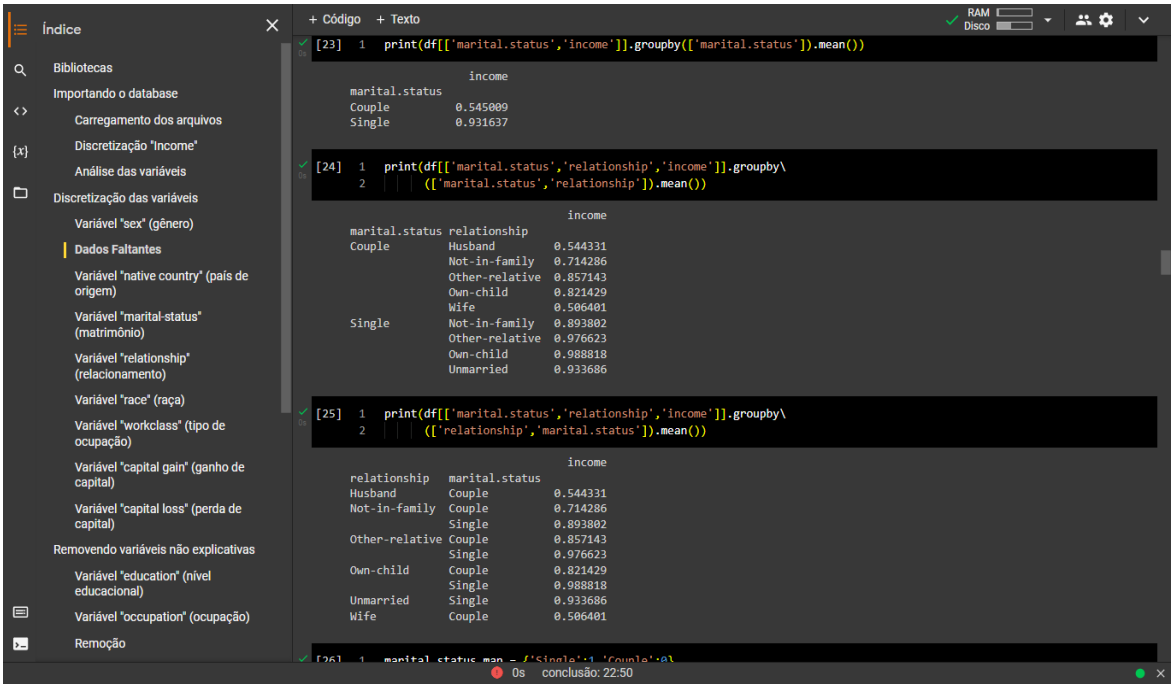


O conjunto de dados possui variáveis, numéricas e categóricas. native. country: país de origem; gender: gênero. Nesta imagem 07

3.2.7. VARIÁVEL MARITAL-STATUS

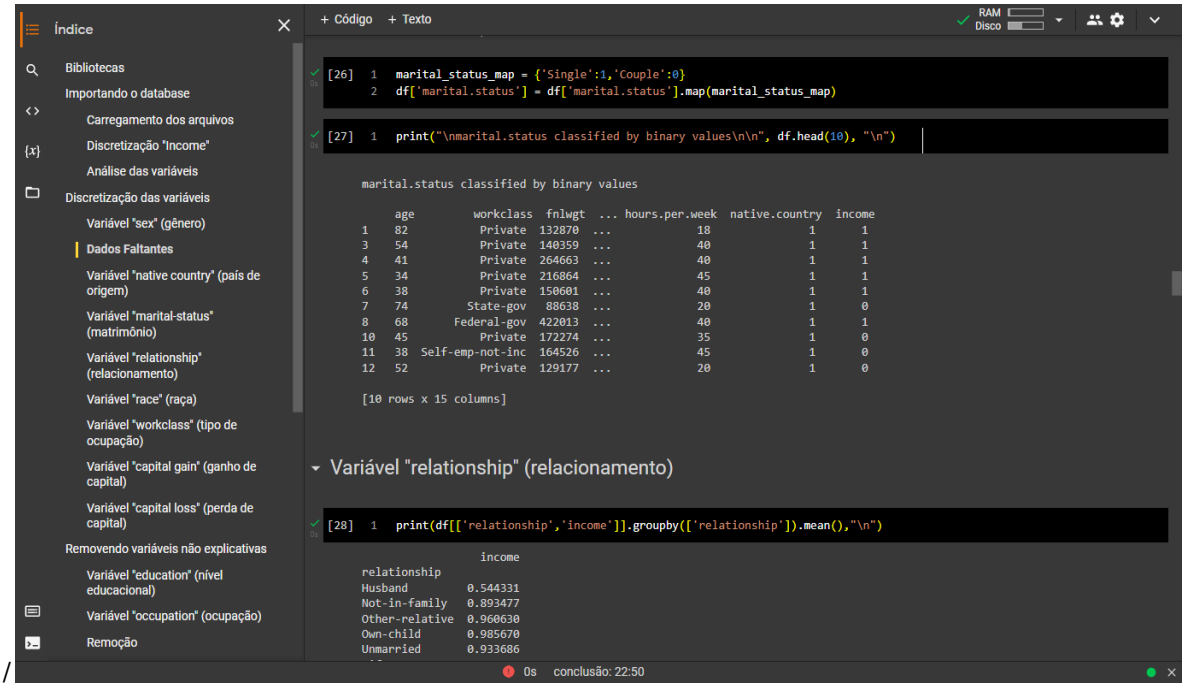






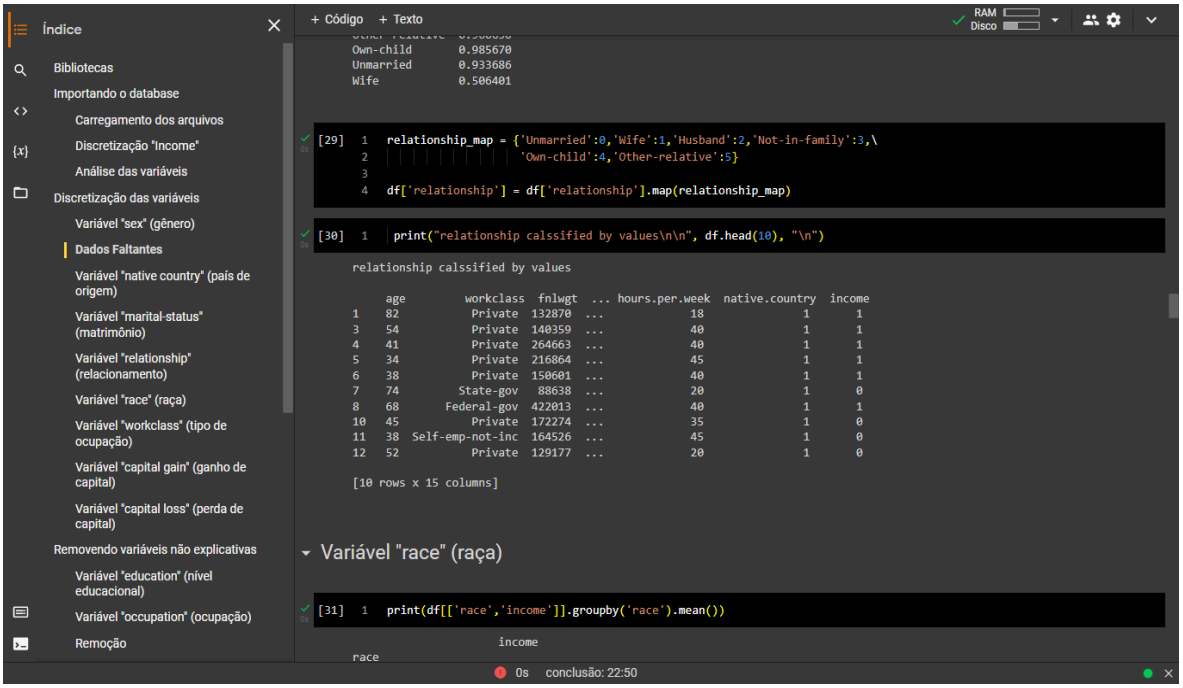
Traduções em contexto de "marital status" en inglês-português da Reverso Context: This referente à variável situação conjugal. Nesta imagem 08,09 e 10

3.2.8. VARIÁVEL RELATIONSHIP



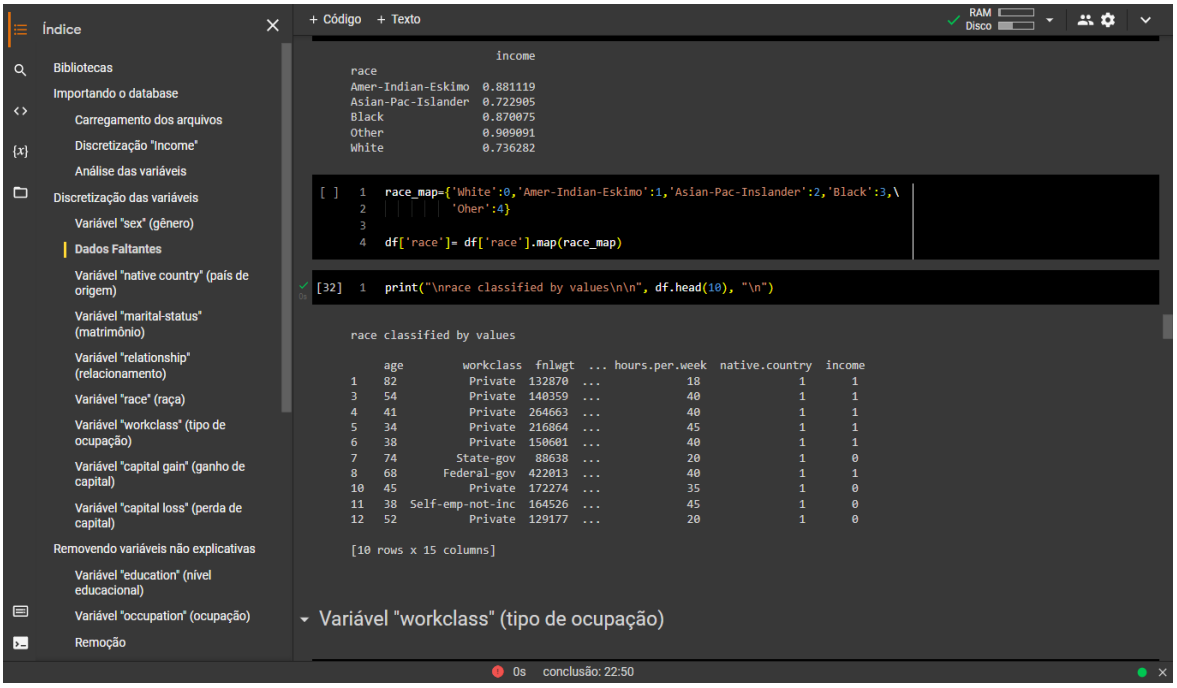
Variáveis relevantes para um bom relacionamento entre cooperativa e relationship between customers, serving as a subsidy for the construction of the. Nesta imagem 11

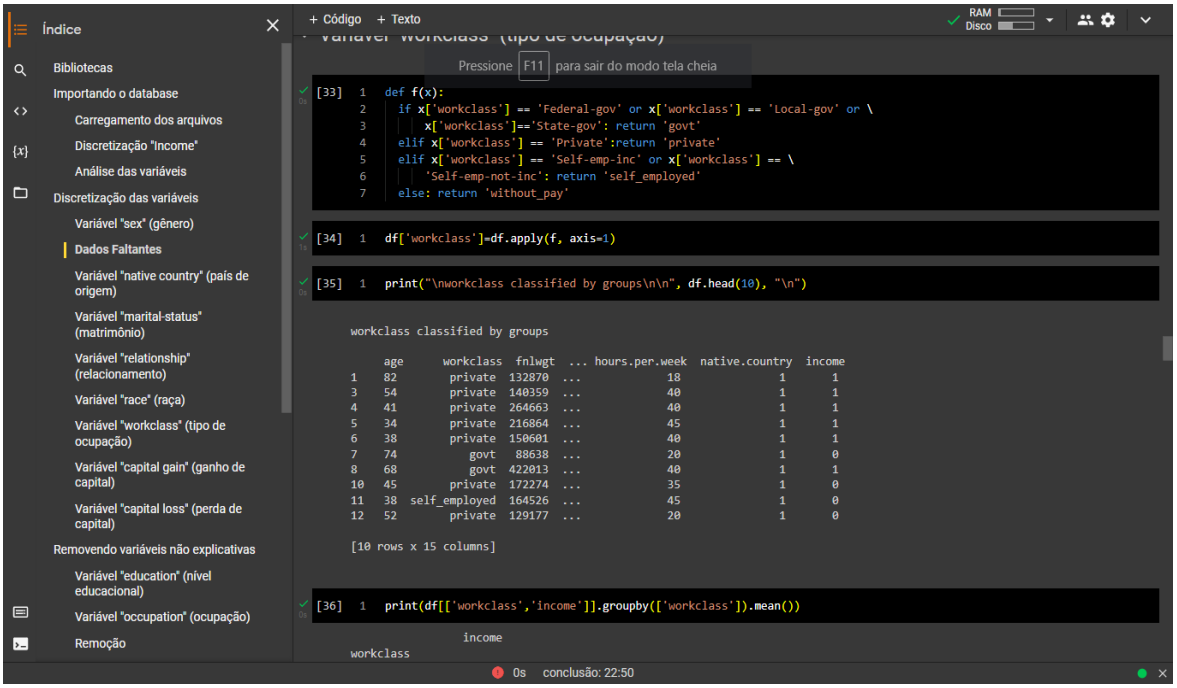
3.2.9. VARIÁVEL RACE



As críticas à classificação racial baseiam-se, fundamentalmente, na demonstração do caráter ideológico subjacente ao conceito de "raça". Nesta imagem 12

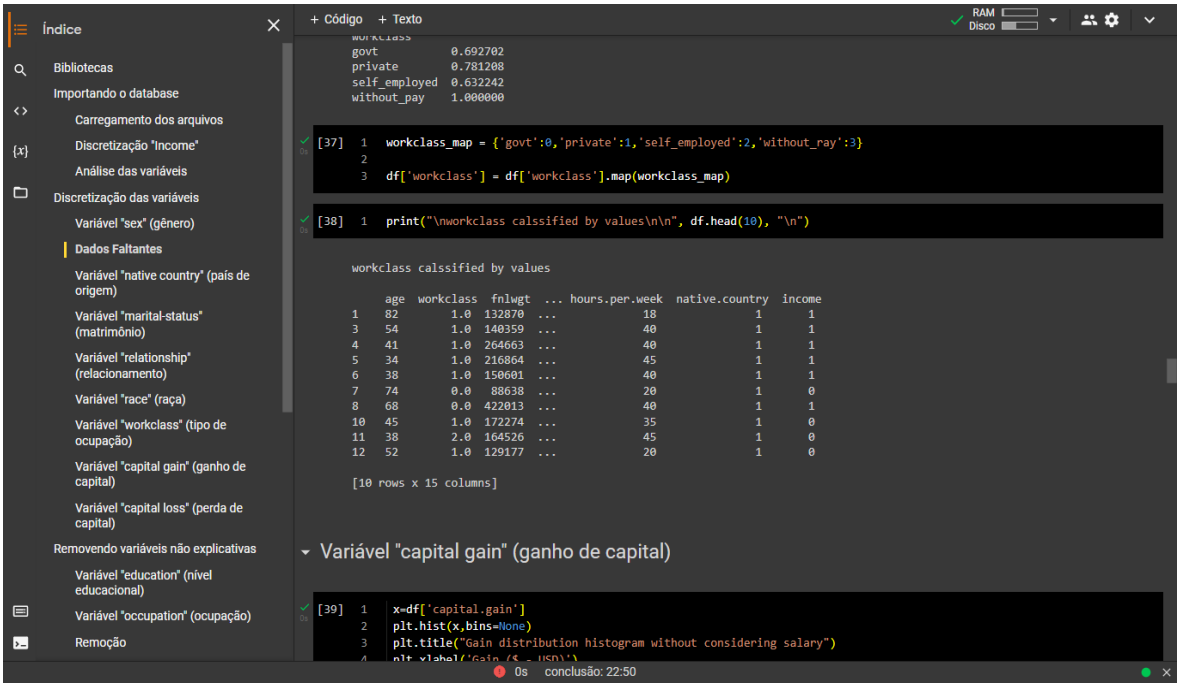
3.2.10. VARIÁVEL WORK CLASS





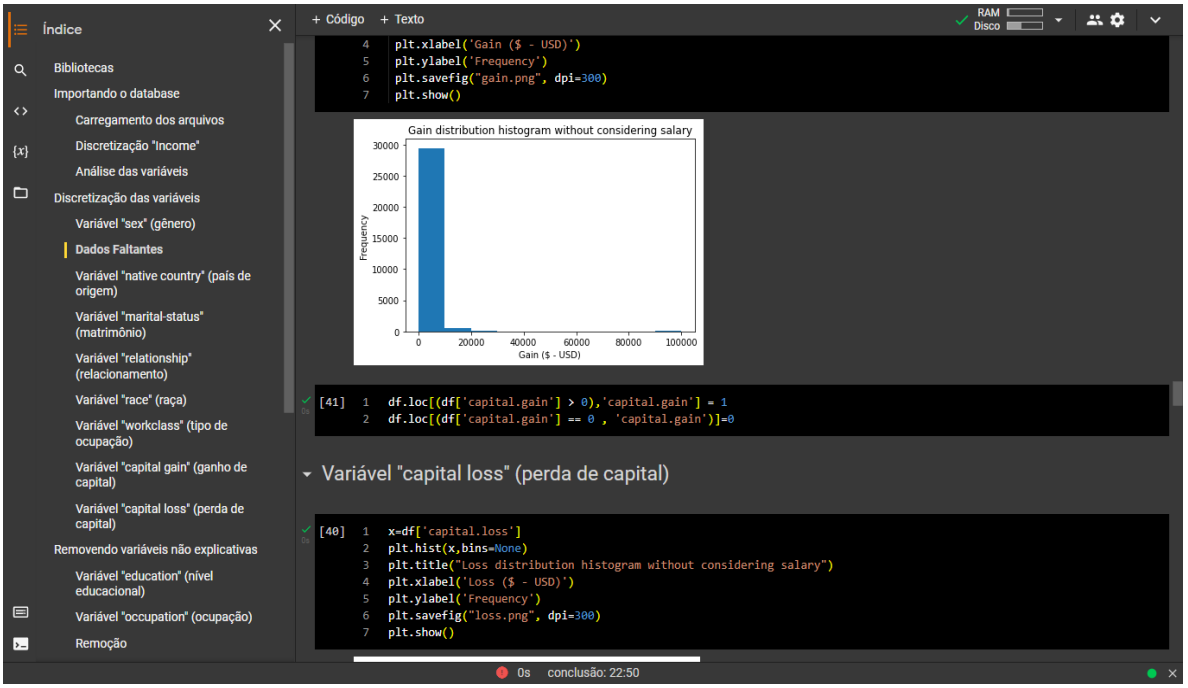
Categóricas, a análise de classes latentes é uma possível técnica a ser utilizada ... Variável categórica que corresponde ao tipo de ocupação declarado pelo. Nessas imagens 13 e 14

3.2.11. VARIÁVEL CAPITAL GAIN



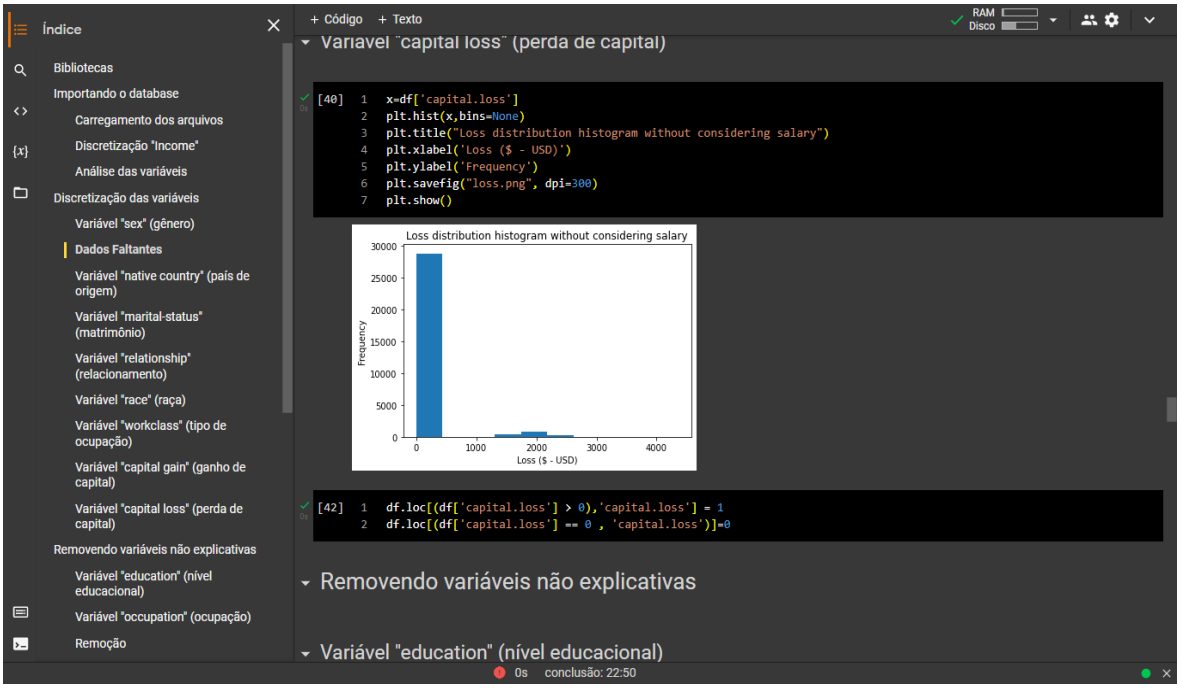
Dentre os resultados obtidos, constatou-se que a variável que ... Social capital, Growth and Poverty: a Survey of Cross-Country Evidence. Nesta imagem 15

3.2.12. VARIÁVEL CAPITAL LOSS

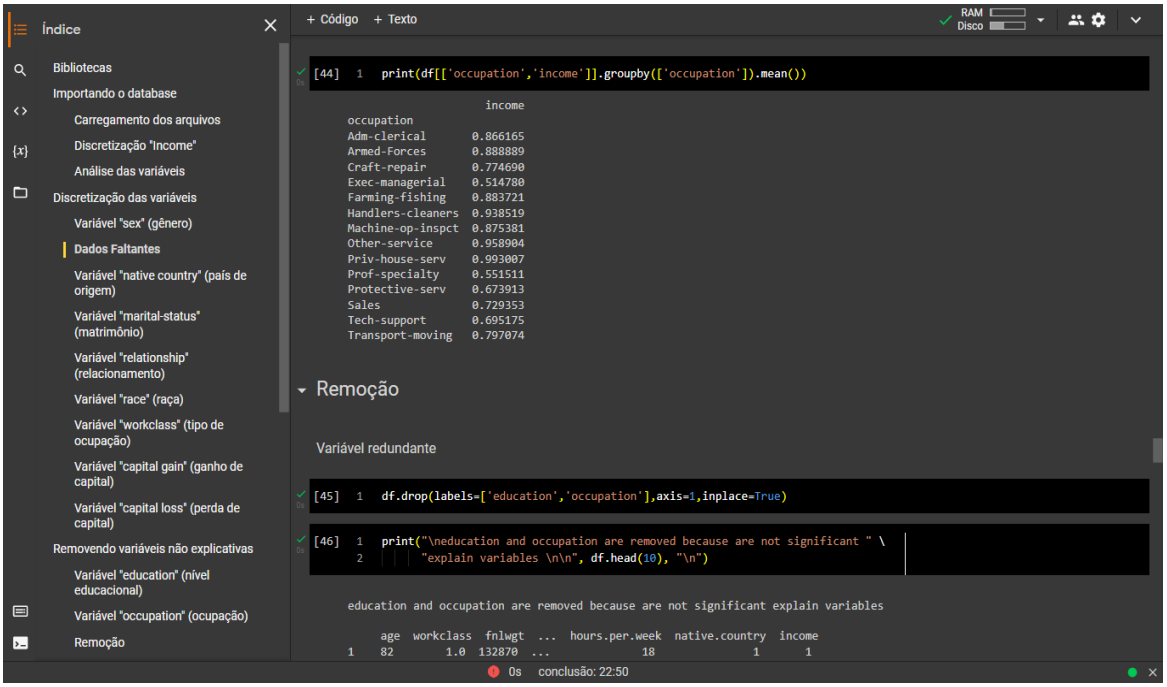


A variável capital próprio indica uma eventual utilização da provisão. Nesta imagem 16

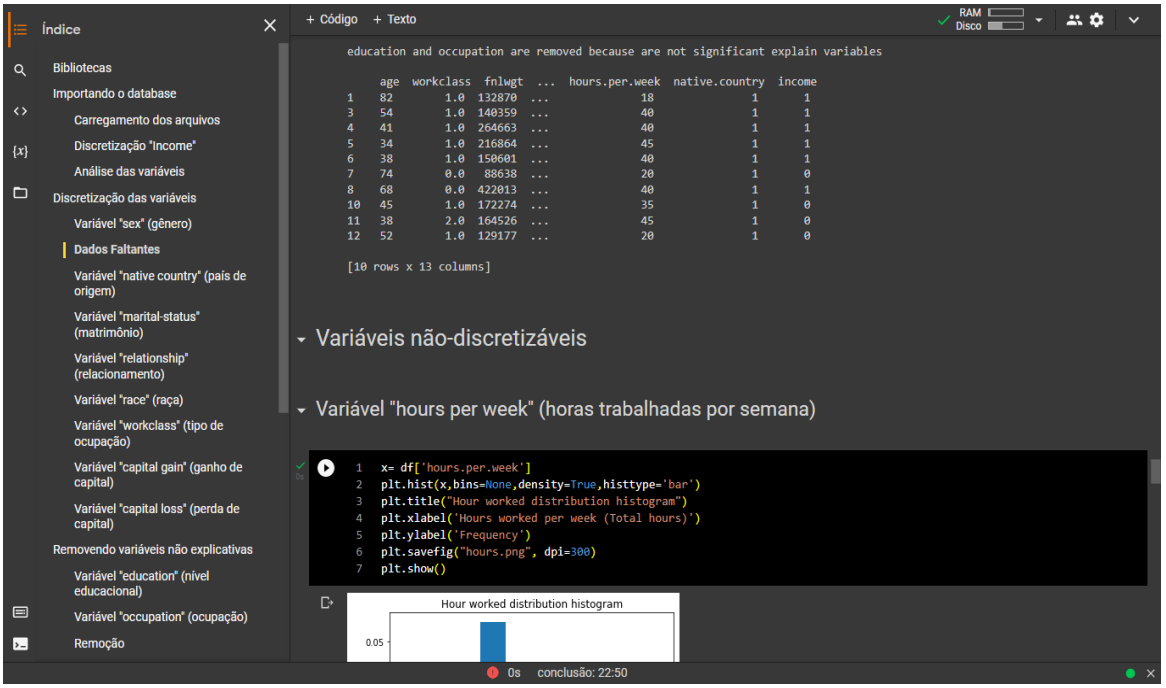
3.2.13. REMOVENDO VARIÁVEIS NÃO EXPLICATIVAS



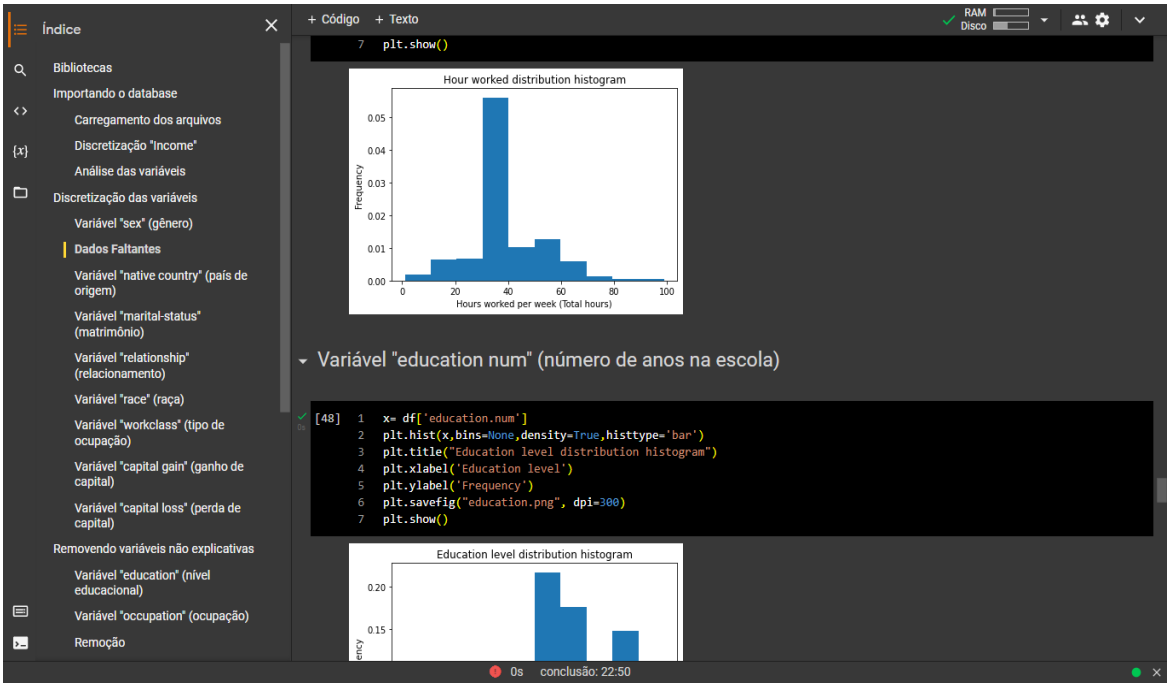
3.2.14. REMOÇÃO



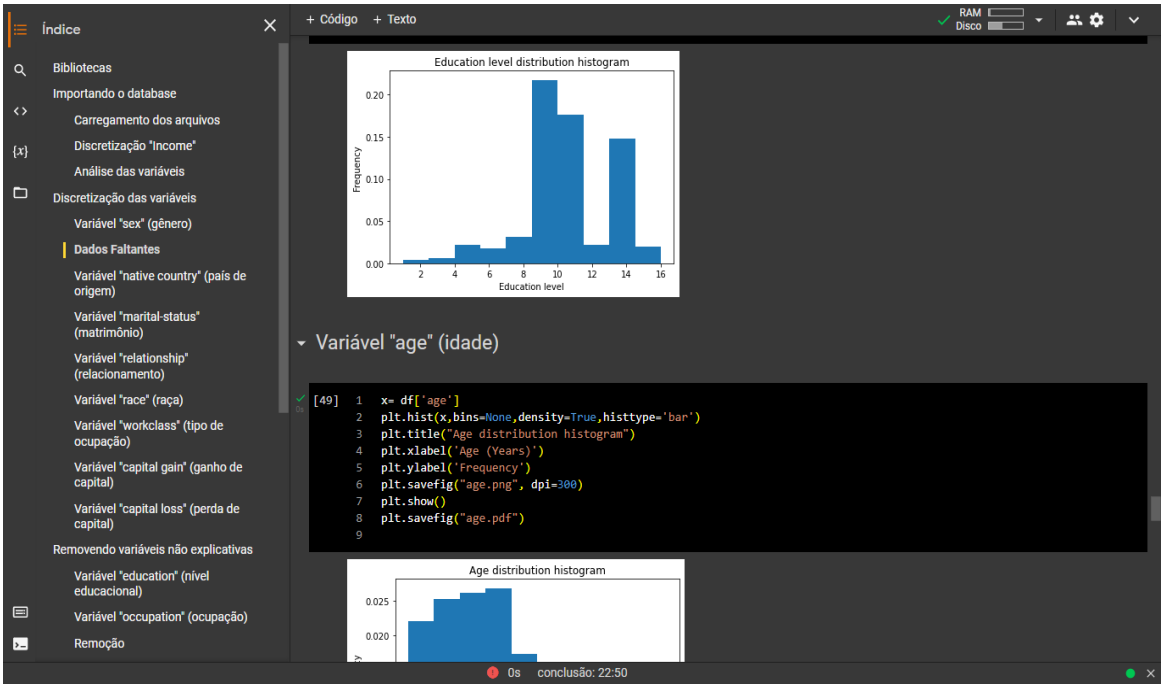
3.2.15. VARIÁVEIS NÃO DISCRETIZADOS



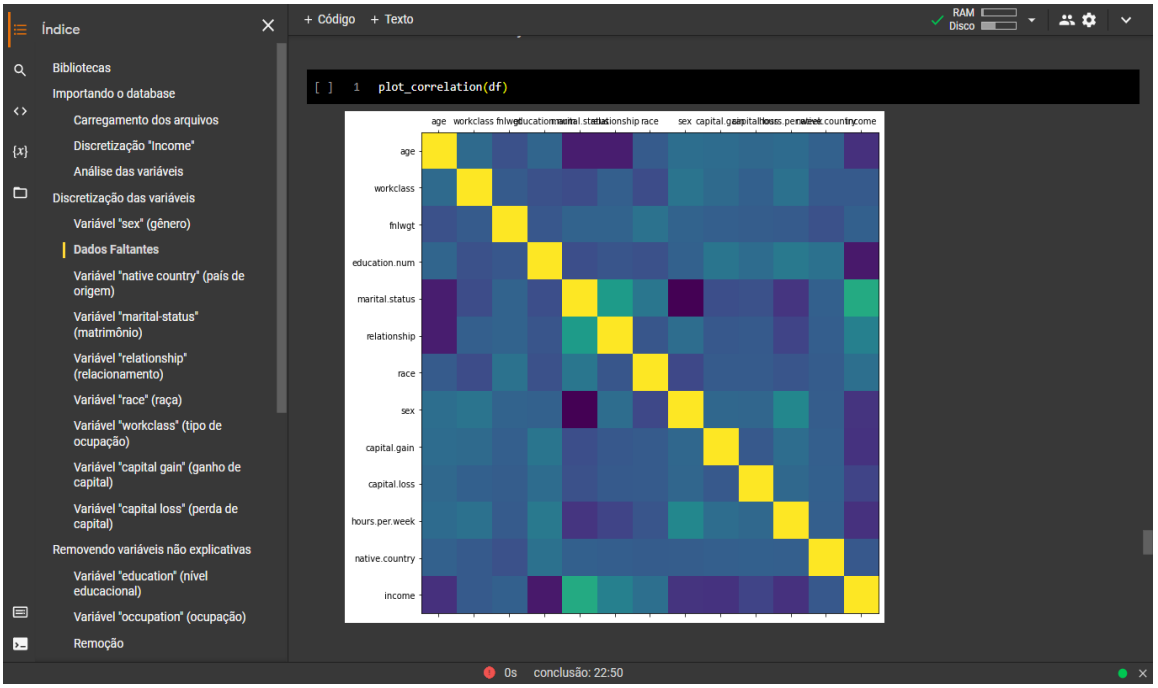
3.2.16. VARIÁVEL EDUCATION NUM



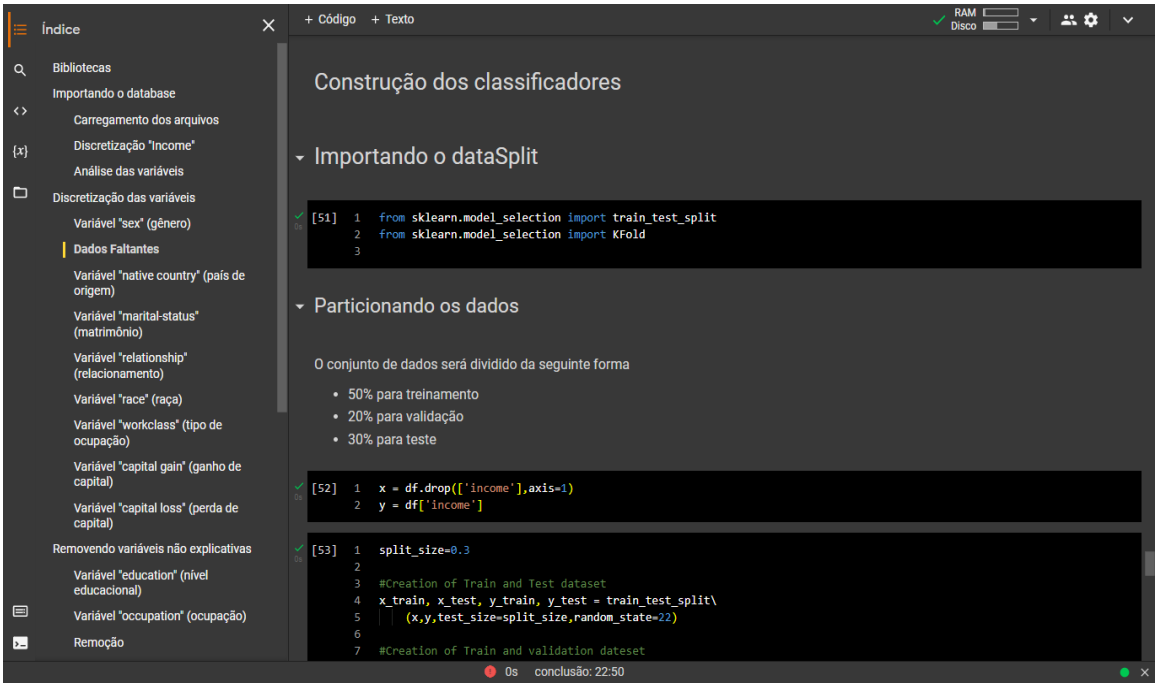
3.2.17. VARIÁVEL AGE



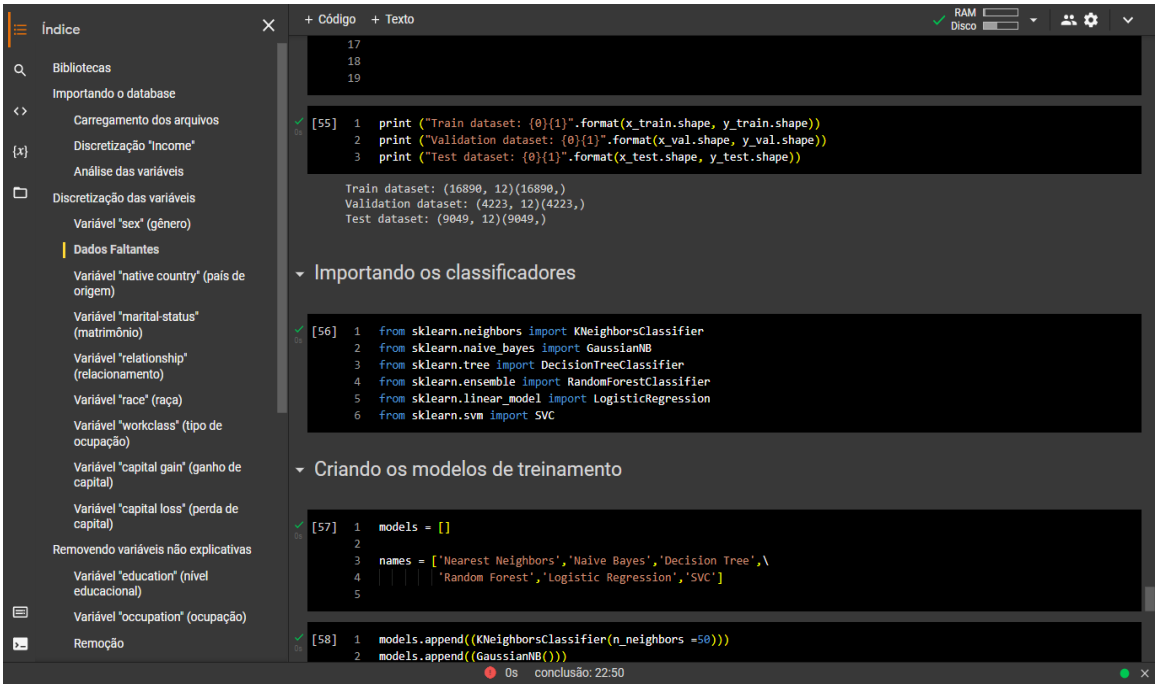
3.2.18. ANALISANDO A CORRELAÇÃO

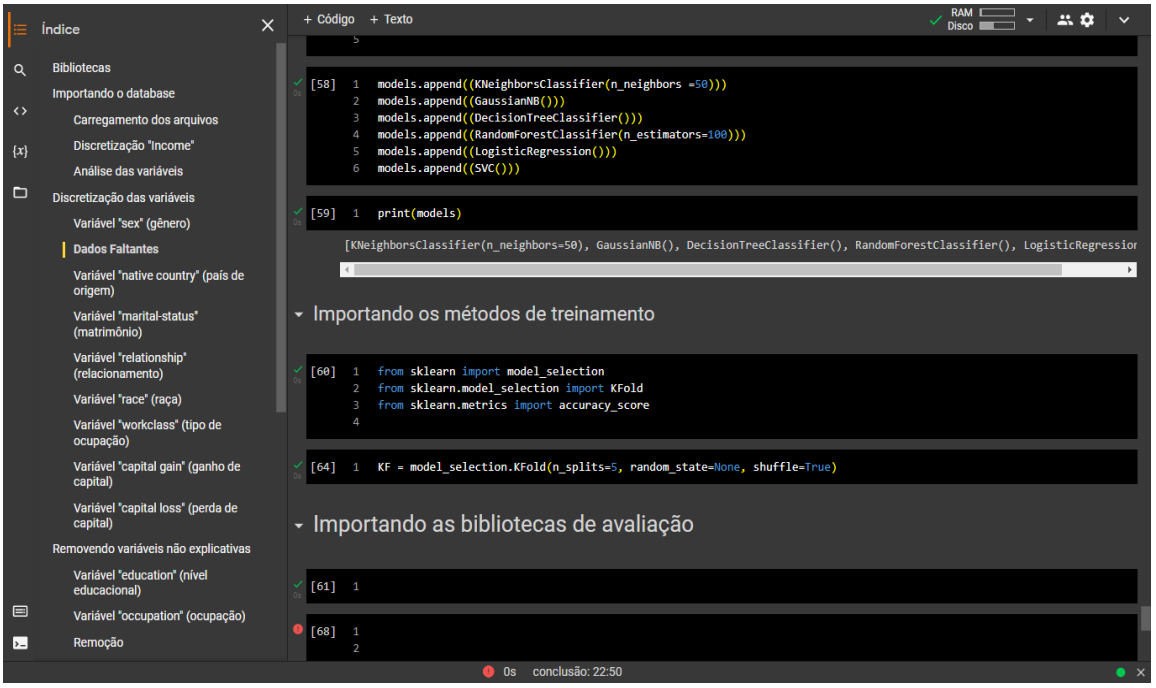


3.2.19. IMPORTANDO O DATA SPLIT

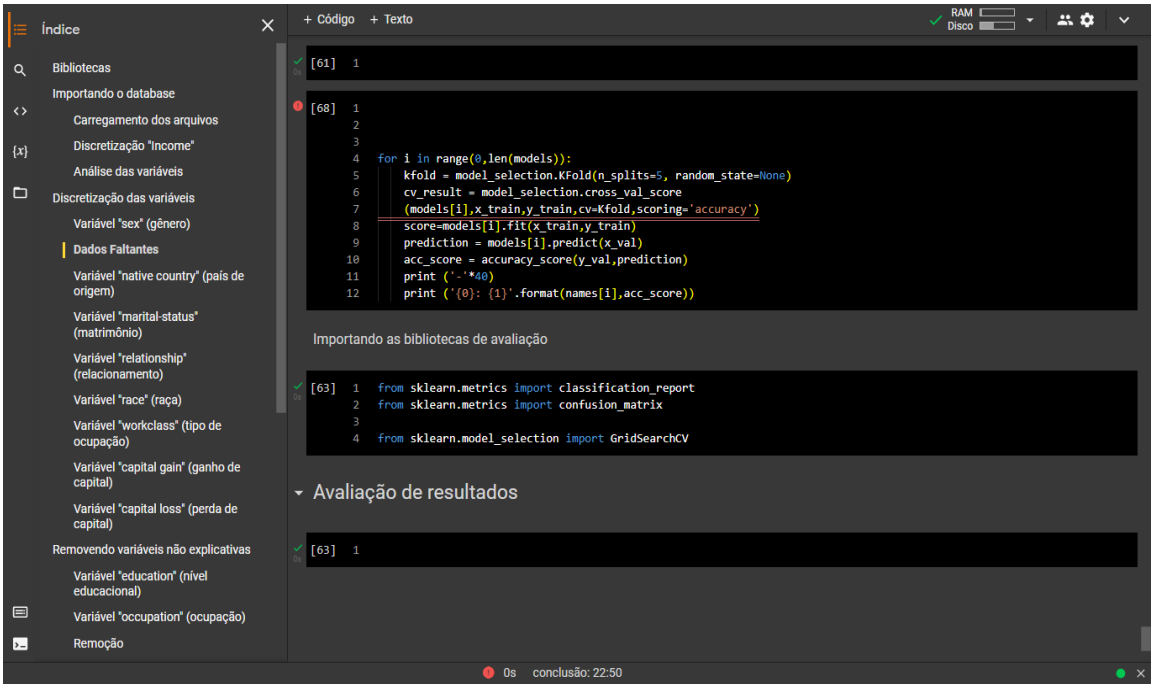


3.2.20. IMPORTANDO OS CLASSIFICADORES

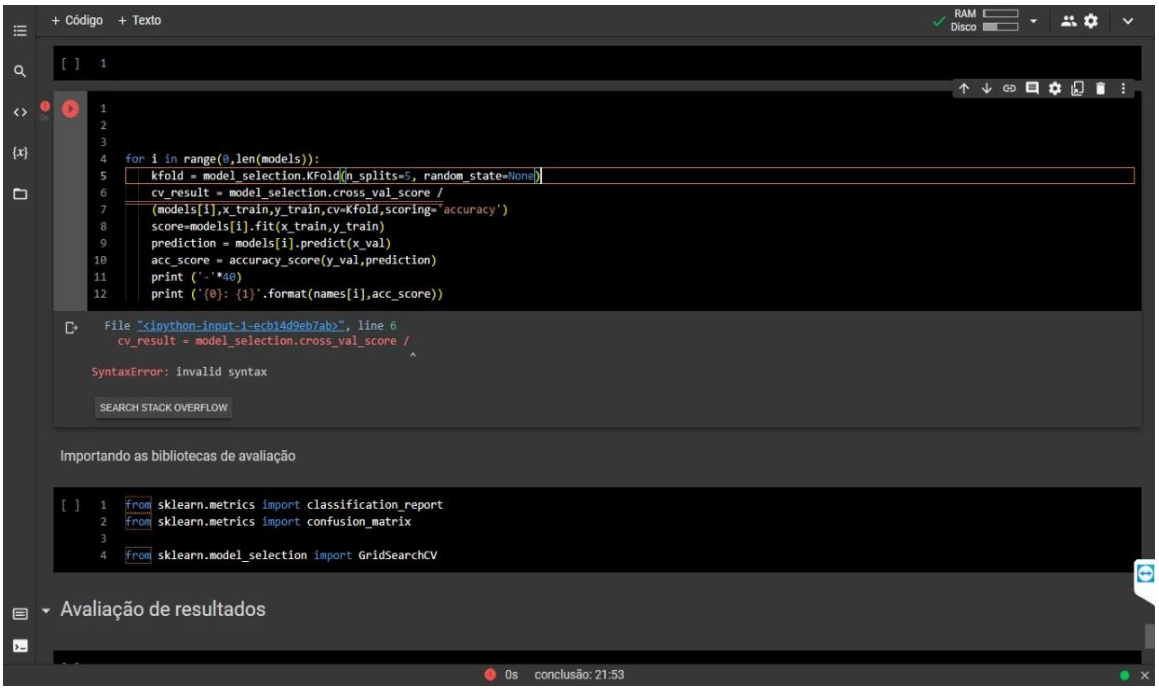




3.2.21. CÓDIGO COM ERRO



3.2.22. Solução Final



Nesta parte do código, ficamos com uma enorme dor de cabeça sobre este pedaço de bloco de código, pois de nenhuma forma obtemos resultado sobre a correção da string. Com isso percebemos que no prosseguimento das linhas, estão faltando pequenos trechos, e nisto, paramos na edição e digitação deste bloco.O restante do conjunto está funcionando perfeitamente, tendo como finalidade

um acerto significativo de linhas corretas. Sendo assim, concluímos que a solução deste pedaço de código ficará para uma próxima checagem, onde iremos trazer soluções plausíveis e coerentes para que seja aplicado um pouco mais de conhecimento.

4. CONSIDERAÇÕES FINAIS

A pesquisa, como um todo, demonstrou ter um grande número de observações relevantes na construção de estruturas a partir desse banco de dados adulto. Tivemos, a princípio, algumas dificuldades para se estudar tal tema, já que esta pesquisa, que está disponível na kaggle.com, apresentou diferenças em nosso método de formatação.

Começamos, em primeiro lugar, agindo na base de análise, parte por parte, antes mesmo de se fazer qualquer outra coisa. Pensamos que seria de certa forma, dificílimo chegar a todo aquele amontoado de informações diversas, mas que de mesmo ponto de interesse - que ainda assim, havia partes dos dados, faltantes. Então, a partir daí, tivemos que começar com as breves edições, apenas para dar uma outra forma neste dataset.

Nisto, começamos a parte de edição usando o aplicativo do pacote office, Excel. Já que nele conseguimos ter uma visão mais encadeada de quais colunas usar na plataforma do google colab, onde foi o inspecionado, digitalizado, estruturado e assim, deixando de uma forma com processo fácil de interpretação, que assim já temos uma estrutura feita, compreensível, legível e entendível.

A importância que se tem nesses pontos é que o acréscimo de funções baseadas em Python no banco de dados se torna acessível.

O banco, como vimos no início para o momento pós edição, é completamente diferente, já que seus dados foram divididos de acordo com seus blocos. Com isso, a estrutura continua com a sua adequação, ligação e correspondentes. Adult census income, como já dito, é uma base de dados com informações de pessoas moradoras da américa, onde há diversas etnias, classes sociais, matrimônios, patrimônios, educação. Pessoas dos seus 17 anos até seus 90. O objetivo deste dataset, é mostrar em seus blocos a divisão com suas respectivas sequências de informações para que se pudesse extrair suas referências.

Contudo, chega-se à conclusão de que a base de dados voltada para entender como é dado o valor anual de que cada pessoa recebe em seu trabalho, é mostrado em casa sequência de código baseado em python que foi escrito. Mas durante todo esse caminho, houve momentos em que tivemos que recorrer a outros métodos para que fosse entendido e corrigido linhas de código que não estavam em combinação de resultados. Vendo vídeos em youtube, acessando diferentes blogs e páginas relacionadas ao dataset, foi possível converter os erros. Mesmo assim, a gama maior de acertos foi do vídeo aula que também nos foi deixava para ser usado como forma de acesso fácil de correção caso fosse necessário. E foi, então, para se concluir, esta base de dados foi útil na questão de se entender de que forma as pessoas são remuneradas nos EUA. Muito se deve à educação e família, onde se tem muita influência, já que se a família fosse de boa situação, de ensino, de bom emprego, de boa influência, ajudaria suas gerações; mas nesse caso, o impacto forma que tem na educação, tempo de estudo, o trabalho em si, reflete a curto, médio e longo prazo, em relação à faixa etária das pessoas pesquisadas.

REFERÊNCIAS BIBLIOGRÁFICAS

http://www.teses.usp.br/index.php?option=com_content&view=article&id=52&Itemid=67

https://medium.com/data-warriors/eda-of-adult-census-income-dataset-cc9ac1a3d552

https://drive.google.com/drive/folders/1UCPP0pXZPXLQnk6O_EETuqgrS9gjAf98?usp=sharing

https://pt.stackoverflow.com/questions/66777/d%C3%BAvidas-na-utiliza%C3%A7%C3%A3o-de-stratified-k-fold-no-scikit-learn

https://ieeexplore.ieee.org/document/8748528

https://www.kaggle.com/uciml/adult-census-income

https://www.youtube.com/watch?v=kEPYo45n51M