

ENGENHARIA DE DADOS PARA SUPORTE À TOMADA DE DECISÃO 2023/2024

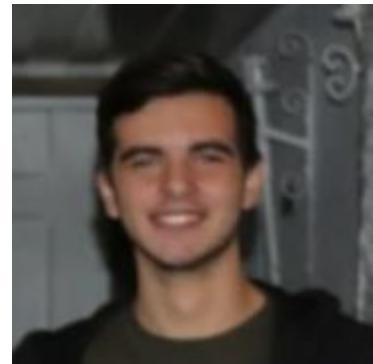
Grupo 5

Entrega Final

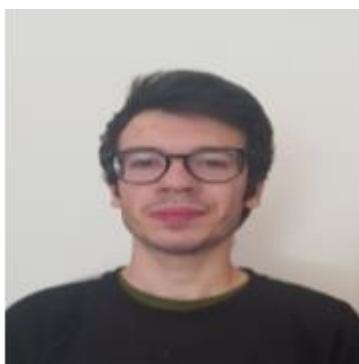
Constituição do Grupo:



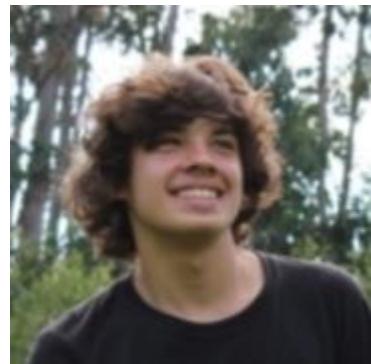
Alexandre Pereira da Costa - a96219



Bruno Miguel Santos Nogueira - a97663



João Pedro Alves Couto de Abreu - a100668



Norberto Miguel Luzes Pais Pinto - a101457



Tomás Chaves Barros Gomes de Morais - a94991

Índice:

Índice

Introdução:	4
Questões Analíticas e KPIs:	5
Dicionário de Dados:	6
Data Lakehouse:	29
Bronze:	31
Silver:	32
Gold:	36
Visualização de dados:	39
.....	41
Auto e Heteroavaliação	42
Vídeo:	42
Conclusão:	42
Anexos:	43

Introdução:

No âmbito da unidade curricular de Engenharia de Dados para Suporte à Tomada de Decisão, do curso de Licenciatura em Engenharia e Gestão de Sistemas de Informação (LEGSI), da Universidade do Minho, foi-nos proposta a elaboração de um projeto, subdividido em 3 fases, com a duração do semestre.

O objetivo do projeto foi construir um Data Lakehouse com datasets principais (indicados pelos professores da disciplina), e datasets complementares, selecionados pelos membros do grupo. Para o sucesso do projeto, foi necessário compreender os dados a serem tratados, identificar problemas com a qualidade dos dados, fornecer planos para correção e transformação dos dados, correlacionar os datasets, e, por fim, apresentar dashboards com as conclusões tiradas a partir das questões analíticas e KPIs.

Neste projeto, o nosso objetivo foi a “Análise multidisciplinar de igualdade de género em Nova Iorque: Pay Gap, Criminalidade, HIV. Em primeiro passo, temos como objetivo situar Estados Unidos no mundo, e, de seguida, afunilar o nosso estudo em Nova Iorque, relacionando-o com os Estados Unidos.

Para além disso, neste mesmo relatório, estão presentes as questões analíticas e os KPIs que foram analisados ao longo do projeto.

Questões Analíticas e KPIs:

Questões Analíticas:

- Como é que o nível de educação está correlacionado com a incidência de casos de HIV em Nova Iorque, entre 2017-2021 por género?
- Qual é a tendência da taxa de mortalidade relacionada ao HIV entre os casos positivos, por género, em Nova Iorque?
- Como varia o número de shootings em Nova Iorque com os estudantes completaram a escola obrigatória?
- Como varia o número de shootings em Nova Iorque com os estudantes que desistiram da escolaridade obrigatória?
- Qual é o pay gap, entre géneros, de pessoas com diferentes níveis de educação, nos estados Unidos?

KPIs:

- Taxa de Incidência de HIV por Nível de Educação e Género.
- Taxa de Mortalidade Relacionada ao HIV em casos positivos por género.
- Número de shootings, por género, em Nova Iorque, relacionado com o seu nível escolar.
- Relação entre o nível de estudos e o nível de salário nos Estados dos EUA, por género.

Dicionário de Dados:

Stats_Country

Atributos	Descrição	Problemas	Resolução
Country Code (String)	Código que identifica um país de forma única, frequentemente seguindo algum padrão internacional.	Não há anomalias na qualidade de dados	
Short Name (String)	O nome abreviado do país.	Não há anomalias na qualidade de dados	
Table Name (String)	O nome do país tabelado.	Não há anomalias na qualidade de dados	
Long Name: (String)	O nome completo do país.	Não há anomalias na qualidade de dados	
2-alpha code (String)	O código de duas letras que representa um país.	Quase sem anomalias na qualidade de dados, tem um valor vazio	Substituir os valores vazios para "Unknown"
Currency Unit (String)	A unidade monetária do país, geralmente a moeda oficial.	Algumas das linhas não têm valores	Substituir os valores vazios para "Unknown"
Special Notes (String)	Notas especiais ou informações adicionais sobre o país.	Algumas das linhas não têm valores	Substituir os valores vazios para "Unknown"
Region (String)	A região geográfica à qual o país pertence.	Algumas das linhas não têm valores	Substituir os valores vazios para "Unknown"
Income Group (String)	O grupo de renda ao qual o país é atribuído, como "país de baixa renda", "país de renda média", etc.	Algumas das linhas não têm valores	Substituir os valores vazios para "Unknown"
WB-2 code (String)	Um código de duas letras usado pelo Banco Mundial para identificar países.	Não há anomalias na qualidade de dados	

National accounts base year (String)	O ano de referência para os dados de contas nacionais.	Algumas das linhas não têm valores	Substituir os valores vazios para “Unknown”
National accounts reference year (String)	O ano de referência para os dados de contas nacionais.	Várias das linhas não têm valores	Substituir os valores vazios para “Unknown”
SNA price valuation (String)	O método de valoração de preços usado nas contas nacionais, seguindo o Sistema de Contas Nacionais das Nações Unidas (SNA)	Algumas das linhas não têm valores	Substituir os valores vazios para “Unknown”
Lending category (String)	A categoria de empréstimo ou financiamento que um país recebe de instituições financeiras internacionais.	Algumas das linhas não têm valores	Substituir os valores vazios para “Unknown”
Other groups (String)	Outros grupos ou categorias relevantes nas quais o país pode se enquadrar.	Várias das linhas não têm valores	Substituir os valores vazios para “Unknown”
System of National Accounts (String)	O sistema usado para compilar e relatar dados de contas nacionais.	Algumas das linhas não têm valores	Substituir os valores vazios para “Unknown”
Alternative conversion factor (String)	Um fator de conversão alternativo usado em cálculos económicos.	Todas as linhas estão sem conteúdo	Possível remoção da coluna.
PPP survey year (String)	O ano da pesquisa de paridade de poder de compra (PPP) usada para comparar o poder de compra entre países.	Todas as linhas estão sem conteúdo	Possível remoção da coluna.

Balance of Payments Manual in use (String)	Guia abrangente para compilar e relatar estatísticas de balanço de pagamentos e posição internacional de investimento.	Algumas das linhas não têm valores	Substituir os valores vazios para “Unknown”
External debt reporting status (String)	O status de relatórios de dívida externa do país.	Algumas das linhas não têm valores	Substituir os valores vazios para “Unknown”
System of trade (String)	O sistema ou classificação usado para acompanhar as transações comerciais do país.	Algumas das linhas não têm valores	Substituir os valores vazios para “Unknown”
Government Accounting concept (String)	O conceito contabilístico usado pelo governo para relatar suas finanças.	Algumas das linhas não têm valores	Substituir os valores vazios para “Unknown”
IMF data dissemination standard (String)	O padrão de disseminação de dados usado pelo Fundo Monetário Internacional (FMI) para divulgar informações económicas.	Algumas das linhas não têm valores	Substituir os valores vazios para “Unknown”
Latest population census (String)	O ano do censo mais recente realizado no país.	Algumas das linhas não têm valores	Substituir os valores vazios para “Unknown”
Latest household survey (String)	O ano da pesquisa domiciliar mais recente.	Algumas das linhas não têm valores	Substituir os valores vazios para “Unknown”
Source of most recent Income and expenditure data (String)	A fonte dos dados mais recentes de renda e despesas.	Algumas das linhas não têm valores	Substituir os valores vazios para “Unknown”

Vital registration complete (String)	Uma indicação se o registo de eventos vitais, como nascimentos e mortes é realizado.	Algumas das linhas não têm valores	Substituir os valores vazios para “Unknown”
Latest agricultural censos (String)	O ano do censo agrícola mais recente.	Algumas das linhas não têm valores	Substituir os valores vazios para “Unknown”
Latest industrial data (String)	O ano dos dados industriais mais recentes disponíveis.	Algumas das linhas não têm valores	Substituir os valores vazios para “Unknown”
Latest trade data (Int)	O ano dos dados comerciais mais recentes disponíveis.	Existem diversas linhas com valor nulo	

Nota: A coluna 30 foi eliminada pois não disponha dados e metia em causa o processamento da qualidade dos dados.

Apesar de algumas colunas terem valores Int, como tem linhas vazias conta como “empty field”, ou seja, pela análise de dados fica como String.

Stats_Series

Atributos	Descrição	Problemas	Resolução
<i>_Series_Code_(String)</i>	Código da série.	Não há anomalias na qualidade de dados.	
<i>Topic(String)</i>	Tópico da série	Não há anomalias na qualidade de dados.	
<i>Indicator_Name(String)</i>	Nome do Indicador	Não há anomalias na qualidade de dados.	
<i>Short_definition(String)</i>	Definição curta.	Entradas Nulas	Alteração das linhas em branco para “Unknown”

<i>Long_definition(String)</i>	Longa definição	Não há anomalias na qualidade de dados.	
<i>Unit_of_measure(String)</i>	Unidade de medida	Entradas nulas.	Alteração das linhas em branco para “Unknown”.
<i>Periodicity(String)</i>	Periodicidade	Entradas nulas.	Alteração das linhas em branco para “Unknown”
<i>Base_Period(String)</i>	Período base	Dados não preenchidos	Remoção do mesmo
<i>Other_Notes(String)</i>	Outras Notas	Dados não preenchidos	Remoção do mesmo
<i>Aggregation_method(String)</i>	Método de agregação	Entradas nulas. Não é relevante para a análise de dados	Remoção do mesmo
<i>Limitations_and_exceptions(String)</i>	Limitações e exceções	Entradas nulas. Não é relevante para a análise de dados	Remoção do mesmo
<i>Notes_from_original_source(String)</i>	Notas da fonte original	Entradas nulas. Não é relevante para a análise de dados	Remoção do mesmo
<i>General_comments(String)</i>	Comentários gerais	Entradas nulas. Não é relevante para a análise de dados	Remoção do mesmo
<i>Source(String)</i>	Fonte	Entradas nulas.	Remoção do mesmo

		Não é relevante para a análise de dados	
<i>Statistical_concept_and_methodology(String)</i>	Conceito estatístico e metodologia	Entradas nulas. Não é relevante para a análise de dados	Remoção do mesmo
<i>Development_relevance(String)</i>	Relevância do desenvolvimento	Entradas nulas. Não é relevante para a análise de dados	Remoção do mesmo
<i>Related_source_links(String)</i>	Outros links relacionados	Dados não preenchidos	Remoção do mesmo
<i>Other_web_links(String)</i>	Links para websites	Dados não preenchidos	Remoção do mesmo
<i>Related_indicators(String)</i>	Indicadores relacionados	Dados não preenchidos	Remoção do mesmo
<i>License_Type(String)</i>	Tipo de licença	Não é relevante para a análise de dados	Remoção do mesmo

FootNote

Atributos	Descrição	Problemas	Resolução
CountryCode(String)	Código do país.	Não há anomalias na qualidade de dados.	
SeriesCode(String)	Identificador alfanumérico para indicadores estatísticos	Não há anomalias na qualidade de dados.	
Year(String)	Ano.	As entradas estão do tipo "YRXXXX" (String) em que "XXXX" representa o ano.	Criar um job para retirar os caracteres "YR" iniciais de cada entrada e transformar o obtido (ano) em Integer.

DESCRIPTION(String)	Descrição.	Não há anomalias na qualidade de dados.	
----------------------------	------------	---	--

Country_Series

Atributos	Descrição	Problemas	Resolução
CountryCode(String)	Código do país.	Não há anomalias na qualidade de dados.	
SeriesCode(String)	Identificador alfanumérico para indicadores estatísticos	Não há anomalias na qualidade de dados.	
DESCRIPTION(String)	Descrição.	Não há anomalias na qualidade de dados.	

StatsSeries_Time

Atributos	Descrição	Problemas	Resolução
SeriesCode (String)	Identificador alfanumérico para indicadores estatísticos	Não existe problemas com os dados	
Year(String)	Ano em Questão	Não existe problemas com os dados	
Description (String)	Descrição de métrica em estudo	Não existe problemas com os dados	
column3 (String)	Coluna sem dados	Coluna sem dados	Remoção da mesma

O atributo Column3 vai ser removido, visto que se trata de uma coluna sem dados, que não será útil para o trabalho.

Stats_Data

Atributos	Descrição	Problemas	Resolução
_Country_Name (String)	Nome do país	Não existem problemas na qualidade de dados	
Country_Code (String)	Código do país	Não existem problemas na qualidade de dados	
Indicator_Name (String)	Nome do indicador	Não existem problemas na qualidade de dados	

Indicator_Code (String)	Indicador alfanumérico que representa o indicador	Não existem problemas na qualidade de dados	
_960 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “1”, e substituir as linhas em branco por “Unknown”
_961 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “1”, e substituir as linhas em branco por “Unknown”
_962 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “1”, e substituir as linhas em branco por “Unknown”
_963 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “1”, e substituir as linhas em branco por “Unknown”
_964 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “1”, e substituir as linhas em branco por “Unknown”
_965 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “1”, e substituir as linhas em branco por “Unknown”
_966 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “1”, e substituir as linhas em branco por “Unknown”
_967 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e	Substituição do primeiro dígito por “1”, e substituir as linhas em branco por “Unknown”

		existem linhas em branco.	
_968 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “1”, e substituir as linhas em branco por “Unknown”
_969 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “1”, e substituir as linhas em branco por “Unknown”
_970 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “1”, e substituir as linhas em branco por “Unknown”
_971 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “1”, e substituir as linhas em branco por “Unknown”
_972 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “1”, e substituir as linhas em branco por “Unknown”
_973 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “1”, e substituir as linhas em branco por “Unknown”
_974 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “1”, e substituir as linhas em branco por “Unknown”
_975 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “1”, e substituir as linhas em branco por “Unknown”

_976 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “1”, e substituir as linhas em branco por “Unknown”
_977 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “1”, e substituir as linhas em branco por “Unknown”
_978 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “1”, e substituir as linhas em branco por “Unknown”
_979 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “1”, e substituir as linhas em branco por “Unknown”
_980 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “1”, e substituir as linhas em branco por “Unknown”
_981 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “1”, e substituir as linhas em branco por “Unknown”
_982 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “1”, e substituir as linhas em branco por “Unknown”
_983 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “1”, e substituir as linhas em branco por “Unknown”
_984 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _	Substituição do primeiro dígito por “1”, e substituir as

		(underscore), e existem linhas em branco.	linhas em branco por “Unknown”
_985 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “1”, e substituir as linhas em branco por “Unknown”
_986 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “1”, e substituir as linhas em branco por “Unknown”
_987 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “1”, e substituir as linhas em branco por “Unknown”
_988 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “1”, e substituir as linhas em branco por “Unknown”
_989 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “1”, e substituir as linhas em branco por “Unknown”
_990 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “1”, e substituir as linhas em branco por “Unknown”
_991 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “1”, e substituir as linhas em branco por “Unknown”
_992 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “1”, e substituir as linhas em branco por “Unknown”

_993 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “1”, e substituir as linhas em branco por “Unknown”
_994 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “1”, e substituir as linhas em branco por “Unknown”
_995 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “1”, e substituir as linhas em branco por “Unknown”
_996 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “1”, e substituir as linhas em branco por “Unknown”
_997 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “1”, e substituir as linhas em branco por “Unknown”
_998 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “1”, e substituir as linhas em branco por “Unknown”
_999 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “1”, e substituir as linhas em branco por “Unknown”
_000 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “2”, e substituir as linhas em branco por “Unknown”
_001 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _	Substituição do primeiro dígito por “2”, e substituir as

		(underscore), e existem linhas em branco.	linhas em branco por “Unknown”
_002 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “2”, e substituir as linhas em branco por “Unknown”
_003 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “2”, e substituir as linhas em branco por “Unknown”
_004 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “2”, e substituir as linhas em branco por “Unknown”
_005 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “2”, e substituir as linhas em branco por “Unknown”
_006 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “2”, e substituir as linhas em branco por “Unknown”
_007 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “2”, e substituir as linhas em branco por “Unknown”
_008 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “2”, e substituir as linhas em branco por “Unknown”
_009 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “2”, e substituir as linhas em branco por “Unknown”

_010 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “2”, e substituir as linhas em branco por “Unknown”
_011 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “2”, e substituir as linhas em branco por “Unknown”
_012 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “2”, e substituir as linhas em branco por “Unknown”
_013 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “2”, e substituir as linhas em branco por “Unknown”
_014 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “2”, e substituir as linhas em branco por “Unknown”
_015 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “2”, e substituir as linhas em branco por “Unknown”
_016 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “2”, e substituir as linhas em branco por “Unknown”
_017 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “2”, e substituir as linhas em branco por “Unknown”
_018 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _	Substituição do primeiro dígito por “2”, e substituir as

		(underscore), e existem linhas em branco.	linhas em branco por “Unknown”
_019 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “2”, e substituir as linhas em branco por “Unknown”
_020 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “2”, e substituir as linhas em branco por “Unknown”
_021 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “2”, e substituir as linhas em branco por “Unknown”
_022 (String)	Ano de estudo em questão	O primeiro dígito do ano está substituído por um _ (underscore), e existem linhas em branco.	Substituição do primeiro dígito por “2”, e substituir as linhas em branco por “Unknown”
Column_67 (String)	Coluna praticamente em branco, com poucos valores	A coluna encontra-se praticamente vazia, e não temos como aproveitar os dados.	Remoção da coluna

- Remoção da coluna Column_67, visto que não será utilizada no nosso projeto.

Hiv_NY

Atributos	Descrição	Problemas	Resolução
Year (Integer)	Ano.	Não há anomalias na qualidade de dados.	
Borough (String)	Bairro	“All” como entrada.	Eliminar as linhas em que o termo “All” é uma entrada na coluna Borough.

UHF (String)	United Hospital Fund perto da residência.	“All” como entrada.	Eliminar as linhas em que o termo “All” é uma entrada na coluna UHF.
Gender (String)	Género.	Duas entradas diferentes a indicar o género masculino (“Man” e “Male”) e duas entradas diferentes a indicar o género feminino (“Woman” e “Female”). “All” como entrada.	Atribuir às entradas “Man” e “Male” bem como às “Woman” e “Female” o mesmo termo para que estas não sejam diferenciadas. Eliminar as linhas em que o termo “All” é uma entrada na coluna género.
Age (String)	Idade.	Não há anomalias na qualidade de dados.	
Race (String)	Raça.	“All” como entrada.	Eliminar as linhas em que o termo “All” é uma entrada na coluna Race.
HIV diagnoses (Integer)	Número de casos diagnosticados com HIV de pessoas de 13 anos ou mais velhas.	Entradas nulas.	Eliminar as linhas com entradas nulas na coluna HIV diagnoses.
HIV diagnosis rate (Float)	Casos de HIV diagnosticados por 100,000 habitantes.	Entradas nulas. Não é necessário na análise de dados.	Remoção do mesmo.
Concurrent diagnoses (Integer)	Número de pessoas de 13 anos ou mais velhas diagnosticadas com ambos HIV e AIDS num período de 31 dias.	Entradas nulas. Não é necessário na análise de dados.	Remoção do mesmo.
linked to care within 3 months (Integer)	Percentagem de casos com carga viral de HIV ou teste de CD4 obtido dentro de 3 meses após o diagnóstico de HIV.	Entradas nulas. Não é necessário na análise de dados.	Remoção do mesmo.

AIDS diagnoses (Integer)	Número de casos diagnosticados com AIDS de pessoas de 13 anos ou mais velhas.	Entradas nulas.	Eliminar as linhas com entradas nulas na coluna HIV diagnoses.
AIDS diagnoses rate (Float)	Casos de AIDS diagnosticados por 100,000 habitantes.	Entradas nulas. Não é necessário na análise de dados.	Remoção do mesmo.
PLWDHI prevalence (Float)	Número de casos de pessoas de 13 anos ou mais velhas a viver com alguém com PLWDHI (infeção de HIV) por 100 habitantes.	Entradas nulas. Não é necessário na análise de dados.	Remoção do mesmo.
_viral suppression (Integer)	Percentagem de pessoas com 13 anos ou mais que vivem com VIH diagnosticada cujo valor do último teste viral realizado foi maior ou igual a 200 cópias/mL em relação ao número de pessoas total que realizou o mesmo.	Entradas nulas.	Eliminar as linhas com entradas nulas na coluna HIV diagnoses.
Deaths (Integer)	Número de mortes de pessoas de 13 anos ou mais velhas diagnosticadas com HIV/AIDS.	Não há anomalias na qualidade de dados.	Eliminar as linhas com entradas nulas na coluna HIV diagnoses.
Death rate (Float)	Mortes por 1000 pessoas que vivem com HIV/AIDS.*	Entradas nulas. Não é necessário na análise de dados.	Remoção do mesmo.
HIV-related death rate (Float)	Percentagem de pessoas que morrem devido a HIV/AIDS em relação ao número de pessoas que morre diagnosticado com HIV/AIDS.	Entradas nulas.	Eliminar as linhas com entradas nulas na coluna HIV diagnoses.

Non-HIV-related death rate (Float)	Percentagem de pessoas que não morrem devido a HIV/AIDS mas com esta diagnosticada em relação ao número de pessoas que morre diagnosticado com HIV/AIDS.	Entradas nulas. Não é necessário na análise de dados.	Remoção do mesmo.
---	--	--	-------------------

* - excluindo mortes em que a pessoa foi diagnosticada com Hiv no momento da morte ou até 15 dias antes da morte.

Shootings_NY

Atributos	Descrição	Problemas	Resolução
<i>INCIDENT_KEY(Integer)</i>	Código da série.	Não é relevante para a análise de dados	Remoção do mesmo
<i>OCCUR_DATE(String)</i>	Data da ocorrência	Não é relevante para a análise de dados	Remoção do mesmo
<i>OCCUR_TIME(String)</i>	Hora da ocorrência	Não é relevante para a análise de dados	Remoção do mesmo
<i>BORO(String)</i>	Condado.	Não há anomalias na qualidade de dados.	
<i>LOC_OF_OCCUR_DESC(String)</i>	Local da ocorrência (dentro ou fora de um estabelecimento)	Não é relevante para a análise de dados	Remoção do mesmo
<i>PRECINT(Integer)</i>	Código postal	Não é relevante para a análise de dados	Remoção das linhas em branco.
<i>JURISDICTION_CODE(Integer)</i>	Código de jurisdição	Não é relevante para a análise de dados	Remoção do mesmo

<i>LOC_CLASSFCTN_DESC(String)</i>	Tipo de espaço	Não é relevante para a análise de dados	Remoção do mesmo
<i>LOCATION_DESC(String)</i>	Localização mais detalhada	Não é relevante para a análise de dados	Remoção do mesmo
<i>STATISTICAL_MURDER_FLAG(Character)</i>	Método de agregação	Não é relevante para a análise de dados	Remoção do mesmo
<i>PERP_AGE_GROUP(String)</i>	Faixa etária do arguido	Entradas com linhas preenchidas com (null)	Remoção dessas mesmas linhas
<i>PERP_SEX(String)</i>	Sexo do Arguido	Entradas com linhas preenchidas com (null)	Remoção dessas mesmas linhas
<i>PERP_RACE(String)</i>	Raça do Arguido	Não é relevante para a análise de dados	Remoção do mesmo
<i>VIC_AGE_GROUP(String)</i>	Faixa etária da vítima	Não há anomalias na qualidade de dados.	
<i>VIC_SEX(Character)</i>	Sexo da Vítima	Não há anomalias na qualidade de dados.	
<i>VIC_RACE(String)</i>	Raça da Vítima	Não é relevante para a análise de dados	Remoção do mesmo
<i>X_COORD_CD(Integer)</i>	Coordenadas X	Não é relevante para a análise de dados	Remoção do mesmo
<i>Y_COORD_CD (Integer)</i>	Coordenadas Y	Não é relevante para a	Remoção do mesmo

		análise de dados	
<i>Longitude(String)</i>	Longitude	Entradas nulas.	Remoção das linhas em branco
<i>Latitude(String)</i>	Latitude	Entradas nulas.	Remoção das linhas em branco
<i>New_Georeferenced_Column(String)</i>	Coluna com referência geográfica	Não é relevante para a análise de dados	Remoção do mesmo

EducationNY

Atributos	Descrição	Problemas	Resolução
DBN(String)	Código	Não há anomalias na qualidade de dados	
School (String)	Nome da Escola	Não há anomalias na qualidade de dados	
Cohort_Year(Int)	Estado	Não há anomalias na qualidade de dados	
Cohort_Category (String)	2013 Rural Urban Continuum Code	Não há anomalias na qualidade de dados	
Demographic (String)	Género	Não há anomalias na qualidade de dados	
Total_Cohort_Num(Int)	Número de alunos	Não há anomalias na qualidade de dados	
Total_Grads_Num (int)	Número de Alunos que acabaram	Não há anomalias na qualidade de dados	

Total_Grads_Pct_of_cohort (String)	Percentagem de Graduação	Não há anomalias na qualidade de dados	
Total_Regents_Num (Int)	Número de Exames feitos	Não há anomalias na qualidade de dados	
Total_Regents_Pct_of_cohort (String)	Percentagem de Alunos que fez os exames	Não há anomalias na qualidade de dados	
Total_Regents_Pct_of_grads (String)	Percentagem de graduados que fez os exames	Não há anomalias na qualidade de dados	
Advanced_Regents_Num(Int)	Número de alunos que fizeram os exames avançados	Não há anomalias na qualidade de dados	
Advanced_Regents_Pct_of_cohort(String)	Percentagem de Alunos que fez os exames avançados	Não há anomalias na qualidade de dados	
Advanced_Regents_Pct_of_grads(String)	Percentagem de graduados que fez os exames avançados	Não há anomalias na qualidade de dados	
Regents_w_o_Advanced_Num(Int)	Número de alunos sem pontuação nos exames avançados	Não há anomalias na qualidade de dados	
Regents_w_o_Advanced_Pct_of_cohort(String)	Percentagem de alunos sem pontuação nos exames	Não há anomalias na qualidade de dados	
Regents_w_o_Advanced_Pct_of_grads(String)	Percentual de graduados sem pontuações avançadas nos exames	Não há anomalias na qualidade de dados	

Local_Num (Int)	Número de alunos locais	Não há anomalias na qualidade de dados	
Local_Pct_of_cohort(String)	Percentagem de alunos locais	Não há anomalias na qualidade de dados	
Local_Pct_of_grads(String)	Percentagem de graduados locais	Não há anomalias na qualidade de dados	
Still_Enrolled_Num (String)	Número de inscritos	Não há anomalias na qualidade de dados	
Still_Enrolled_Pct_of_cohort (String)	Percentagem de inscritos	Não há anomalias na qualidade de dados	
Dropped_Out_Num(String)	Número de desistências	Não há anomalias na qualidade de dados	
Dropped_Out_Pct_of_cohort (String)	Percentagem de desistências	Não há anomalias na qualidade de dados	

EducationUS

Atributos	Descrição	Problemas	Resolução
State (String)	Nome do Estado	Não existem problemas na qualidade de dados	
Less_than_a_High_School_Diploma_Woman (Float)	Mulheres que têm menos que um Diploma	Não existem problemas na qualidade de dados	

Less_than_a_High_Scholl_Diploma_Men (Float)	Homens que têm menos que um Diploma	Não existem problemas na qualidade de dados	
High_School_Diploma_or_the_equivalent_Only_Women (Float)	Mulheres que têm um diploma ou equivalente	Não existem problemas na qualidade de dados	
High_School_Diploma_or_the_equivalent_Only_Men (Float)	Homens que têm um diploma ou equivalente	Não existem problemas na qualidade de dados	
Some_College_or_an_Associate_s_Degree_women (Float)	Mulheres que têm um curso Universitário	Não existem problemas na qualidade de dados	
Some_College_or_an_Associate_s_Degree_Men (Float)	Homens que têm um curso Universitário	Não existem problemas na qualidade de dados	
Bachelor_s_Degree_or_Higher_Women (Float)	Mulheres que têm um Mestrado ou superior	Não existem problemas na qualidade de dados	
Bachelor_s_Degree_or_Higher_Men(Float)	Homens que têm um Mestrado ou superior	Não existem problemas na qualidade de dados	

EarningsUS

Atributos	Descrição	Problemas	Resolução
State (String)	Nome do Estado	Não existem problemas na qualidade de dados	
Male (Double)	Valor do salário médio de homens	Não existem problemas na qualidade de dados	
Female (Double)	Valor do salário médio de mulheres	Não existem problemas na qualidade de dados	

Data Lakehouse:

- Data Lake Architecture

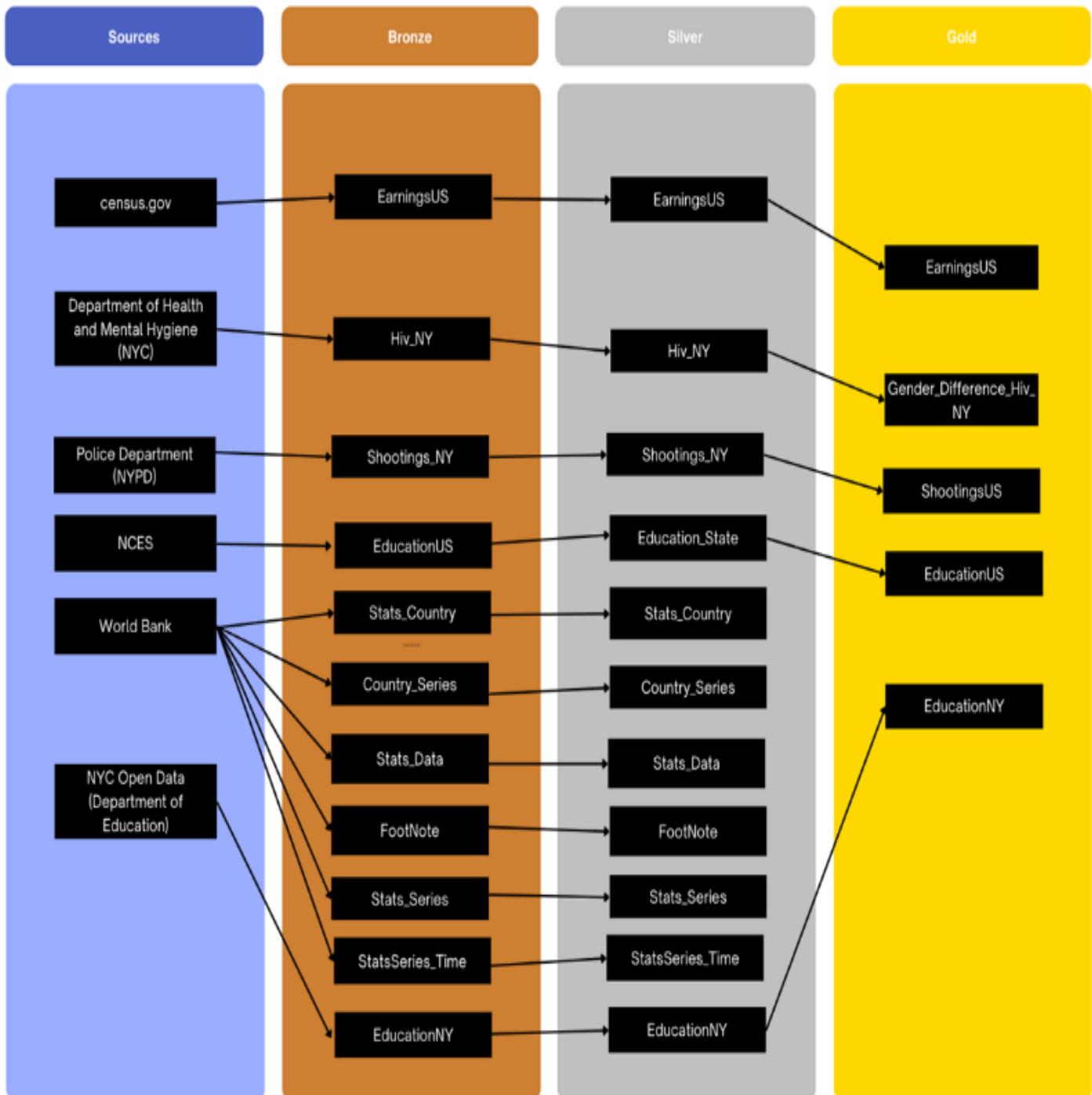


Tabela Datasets Complementares:

EarningsUS	https://www.census.gov/library/visualizations/interactive/gender-pay-gap.html	Este dataset contém informações sobre o diferente tipo de salários entre homens e mulheres, nos diversos estados dos Estados Unidos.
Hiv_NY	https://data.cityofnewyork.us/Health/DOHMH-HIV-AIDS-Annual-Report/fju2-rdad	Este Dataset contém informações sobre as pessoas infectadas com HIV em Nova Iorque
Shootings_NY	https://data.cityofnewyork.us/Public-Safety/NYPD-Shooting-Incident-Data-Year-To-Date-/5ucz-vwe8	Este Dataset contém informações sobre tiroteios em Nova Iorque
EducationUS	https://statusofwomendata.org/explore-the-data/poverty-opportunity/additional-state-data/highest-level-of-educational-attainment-of-women-and-men-by-state-2013/	Este Dataset contém informações sobre as quantidades e género de pessoas com determinado grau académico em todos os estados dos EUA
Education_NY	https://data.cityofnewyork.us/Education/2005-2011-Graduation-Outcomes-School-Level-Gender/khqix3p3	Este Dataset contém informações sobre a quantidade de sucesso escolar de vários alunos de entre várias escolas localizadas pelos borough de Nova Iorque.

Bronze:

- **Objetivo:** Escolher os datasets adequados, realizar a análise inicial dos dados (ponto 1) e carregar os dados no HDFS (exemplo a seguir).
- **Processo:** Identificamos e selecionamos os datasets necessários para o projeto. Realizamos uma análise da qualidade dos dados para perceber a sua estrutura e conteúdo. De seguida, carregar os dados no HDFS para preparação para as fases subsequentes.

O código Pyspark necessário para o carregamento das tabelas Bronze é o seguinte:

```
import sys
from os import PathLike
from hdfs import InsecureClient
!{sys.executable} -m pip install hdfs
Requirement already satisfied: hdfs in /opt/conda/lib/python3.11/
Requirement already satisfied: docopt in /opt/conda/lib/python3.11/
Requirement already satisfied: requests>=2.7.0 in /opt/conda/lib/p
Requirement already satisfied: six>1.9.0 in /opt/conda/lib/python
Requirement already satisfied: charset-normalizer<4,>=2 in /opt/c
Requirement already satisfied: idna<4,>=2.5 in /opt/conda/lib/pyt
Requirement already satisfied: urllib3<3,>=1.21.1 in /opt/conda/li
Requirement already satisfied: certifi>=2017.4.17 in /opt/conda/li

client = InsecureClient("http://hdfs-nn:9870", user="anonymous")

from_path = "./EducationNY_1.csv"
to_path = "/projeto/bronze/EducationNY.csv"
client.delete(to_path)
client.upload(to_path, from_path)
```

Figura 1 – Exemplo carregamento bronze

Os nossos csvs Bronze distribuídos no nosso Hdfs:

	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
□	-rw-r--r--	anonymous	hdfs	267.61 KB	Jan 02 23:24	3	128 MB	Country_Series.csv	trash
□	-rw-r--r--	anonymous	hdfs	1.15 KB	Jan 02 23:24	3	128 MB	EarningsUS.csv	trash
□	-rw-r--r--	anonymous	hdfs	2.3 MB	Jan 02 23:24	3	128 MB	EducationNY.csv	trash
□	-rw-r--r--	anonymous	hdfs	2.83 KB	Jan 02 23:24	3	128 MB	EducationUS.csv	trash
□	-rw-r--r--	anonymous	hdfs	3.92 KB	Jan 02 22:44	3	128 MB	Education_NY.csv	trash
□	-rw-r--r--	anonymous	hdfs	48.54 MB	Jan 02 23:24	3	128 MB	FootNote.csv	trash
□	-rw-r--r--	anonymous	hdfs	2.65 MB	Jan 02 23:24	3	128 MB	Hiv_NY.csv	trash
□	-rw-r--r--	anonymous	hdfs	105.87 KB	Jan 02 23:24	3	128 MB	Shootings_NY.csv	trash
□	-rw-r--r--	anonymous	hdfs	7.43 KB	Jan 02 23:24	3	128 MB	StatsSeries_Time.csv	trash
□	-rw-r--r--	anonymous	hdfs	153.24 KB	Jan 02 23:24	3	128 MB	Stats_Country.csv	trash
□	-rw-r--r--	anonymous	hdfs	112.84 MB	Jan 02 23:24	3	128 MB	Stats_Data.csv	trash
□	-rw-r--r--	anonymous	hdfs	2.05 MB	Jan 02 23:24	3	128 MB	Stats_Series.csv	trash

Showing 1 to 12 of 12 entries

Previous 1 Next

bronze

Figura 2 – HDFS

Silver:

- **Objetivo:** Limpamos os dados e carregá-los em uma tabela Hive utilizando o PySpark com base na análise realizada na fase anterior.
- **Processo:** Realizamos a limpeza dos dados para identificar valores nulos, duplicados e outras questões de qualidade como a renomeação de colunas, entre outras várias alterações. Posteriormente, utilizamos o PySpark para realizar a transformações e carregar os dados numa tabela Hive no formato Delta, que é otimizado para o armazenamento e para consultas analíticas.

Exemplo do processo de transformação do csv Hiv_NY do bronze para o Silver:

```
[4]: from pyspark.sql import SparkSession
from pyspark.sql import Row
from delta import *
from pyspark.sql.functions import col, when
from pyspark.sql import SparkSession
from pyspark.sql.types import DoubleType, StringType, StructField, StructType, IntegerType

# warehouse_location points to the default location for managed databases and tables
warehouse_location = 'hdfs://hdfs-nn:9000/warehouse'

builder = SparkSession \
    .builder \
    .master("local[2]") \
    .appName("Python Spark DataFrames and SQL") \
    .config("spark.sql.warehouse.dir", warehouse_location) \
    .config("hive.metastore.uris", "thrift://hive-metastore:9083") \
    .config("spark.sql.extensions", "io.delta.sql.DeltaSparkSessionExtension") \
    .config("spark.sql.catalog.spark_catalog", "org.apache.spark.sql.delta.catalog.DeltaCatalog") \
    .config("spark.jars.packages", "io.delta:delta-core_2.12:2.4.0") \
    .enableHiveSupport() \

spark = builder.getOrCreate() #spark = configure_spark_with_delta_pip(builder).getOrCreate()

[5]: hdf5_path = "hdfs://hdfs-nn:9000/projeto/bronze/Hiv_NY.csv"

[7]: # Read without header
hiv = spark.read.option("header", True) \
    .csv(hdf5_path)

#Delete col
hiv = hiv.drop(col("Concurrent diagnoses"))
hiv = hiv.drop(col("% linked to care within 3 months"))
hiv = hiv.drop(col("AIDS diagnosis rate"))
hiv = hiv.drop(col("PLWHD prevalence"))
hiv = hiv.drop(col("Death rate"))
hiv = hiv.drop(col("Non-HIV-related death rate"))

hiv = hiv.withColumn(
    "Borough",
    when(
        col("Borough").isNull(),
        "Unknown"
    ).otherwise(col("Borough"))
)

hiv = hiv.withColumn( #alterar summary em que os valores estejam null para unknown"
    "UHF",
    when(
        col("UHF").isNull(),
        "Unknown"
    ).otherwise(col("UHF"))
)

hiv = hiv.withColumn( #alterar summary em que os valores estejam null para unknown"
    "Race",
    when(
        col("Race").isNull(),
        "Unknown"
    ).otherwise(col("Race"))
)
```

Figura 3 – Processo Silver parte 1

```

#Delete Nulls in important col
hiv = hiv.filter(~(col("HIV diagnosis rate").isNull()))
hiv = hiv.filter(~(col("HIV diagnoses").isNull()))
hiv = hiv.filter(~(col("AIDS diagnoses").isNull()))
hiv = hiv.filter(~(col("% viral suppression").isNull()))
hiv = hiv.filter(~(col("HIV-related death rate").isNull()))

#Replace /Man with Male/ and /Woman with Female/
hiv = hiv.withColumn(
    "Gender",
    when(
        col("Gender") == "Man",
        "Male"
    ).when(
        col("Gender") == "Woman",
        "Female"
    ).when(
        col("Gender").isNull(),
        "Unknown"
    ).otherwise(col("Gender"))
)

hiv = hiv.withColumn("Year", col("Year").cast("int"))
hiv = hiv.withColumn("HIV diagnoses", col("HIV diagnoses").cast("int"))
hiv = hiv.withColumn("HIV diagnosis rate", col("HIV diagnosis rate").cast("double"))
hiv = hiv.withColumn("AIDS diagnoses", col("AIDS diagnoses").cast("int"))
hiv = hiv.withColumn("% viral suppression", col("% viral suppression").cast("double"))
hiv = hiv.withColumn("Deaths", col("Deaths").cast("int"))
hiv = hiv.withColumn("HIV-related death rate", col("HIV-related death rate").cast("double"))

hiv = hiv.withColumnRenamed("HIV diagnoses", "HIV_diagnoses")
hiv = hiv.withColumnRenamed("HIV diagnosis rate", "HIV_diagnosis_rate")
hiv = hiv.withColumnRenamed("AIDS diagnoses", "AIDS_diagnoses")
hiv = hiv.withColumnRenamed("% viral suppression", "viral_suppression_percent")
hiv = hiv.withColumnRenamed("HIV-related death rate", "HIV_related_death_rate")

hiv.printSchema()
hiv.show()
#hiv.toPandas()
#Usar o comando acima para uma melhor pré-visualização da tabela

customSchema = StructType([
    StructField("Year", IntegerType(), True),
    StructField("Borough", StringType(), True),
    StructField("UHF", StringType(), True),
    StructField("Gender", StringType(), True),
    StructField("Age", StringType(), True),
    StructField("Race", StringType(), True),
    StructField("HIV_diagnoses", IntegerType(), True),
    StructField("HIV_diagnosis_rate", DoubleType(), True),
    StructField("AIDS_diagnoses", IntegerType(), True),
    StructField("viral_suppression_percent", DoubleType(), True),
    StructField("Deaths", IntegerType(), True),
    StructField("HIV_related_death_rate", DoubleType(), True)
])

hiv \
    .write \
    .format("delta") \
    .mode("overwrite") \
    .option("overwriteSchema", "true") \
    .save("hdfs://hdfs-nn:9000/warehouse/projeto.db/Hiv_NY")

```

Figura 4 – Processo Silver parte 2

Na parte 1: Inicialmente vamos introduzimos caminho do csv que pretendemos manipular: `hdfs_path = "hdfs://hdfs-nn:9000/projeto/bronze/Hiv_NY.csv"`

Com o PySpark lemos o ficheiro csv, ignorando o header, e o DataFrame resultante (`hiv`) será utilizado manipulações de dados subsequentes.

De seguida, iniciamos a Remoção de Colunas Desnecessárias (`hiv.drop(col("nomedaColuna"))`) e ao preenchimento dos valores nulos para unknown das colunas Borough, "UHF," e "Race".

Na parte 2: Inicialmente é removido as linhas que nas colunas importantes, como na "HIV diagnosis rate," contêm valores nulos, de seguida, os géneros são uniformizados

de "Man" por "Male" e "Woman" por "Female" e os valores nulos são substituídos por "Unknown."

No código seguinte deste processo, é realizada a conversão dos tipos de dados para os desejados, e é renomeada o nome das colunas para versões em espaços.

No final, é definido o esquema da tabela para garantir que os dados sejam interpretados corretamente na passagem para o formato delta

É definido um esquema personalizado para garantir que os tipos de dados sejam interpretados corretamente ao escrever no formato Delta. O DataFrame manipulado é então escrito no formato Delta Lake no caminho “./warehouse/projeto”, dedicado para receber os ficheiros Silver.

A disposição que os nossos dados Delta Lake estão armazenados em ficheiros Delta distribuídos no formato Parquet e geridos por tabelas Hive:

The screenshot shows a file browser interface with the following details:

- Path: /warehouse/projeto.db
- Show: 25 entries
- Search: [empty]
- Table Headers: □, Permission, Owner, Group, Size, Last Modified, Replication, Block Size, Name
- Data Rows (11 entries):
 - drwxr-xr-x, joyyan, hdfs, 0 B, Jan 02 23:28, 0, 0 B, Country_Series
 - drwxr-xr-x, joyyan, hdfs, 0 B, Jan 02 23:54, 0, 0 B, EarningsUS
 - drwxr-xr-x, joyyan, hdfs, 0 B, Jan 03 00:57, 0, 0 B, EducationNY
 - drwxr-xr-x, joyyan, hdfs, 0 B, Jan 02 23:44, 0, 0 B, Education_State
 - drwxr-xr-x, joyyan, hdfs, 0 B, Jan 02 23:33, 0, 0 B, FootNote
 - drwxr-xr-x, joyyan, hdfs, 0 B, Jan 02 23:39, 0, 0 B, Hiv_NY
 - drwxr-xr-x, joyyan, hdfs, 0 B, Jan 03 00:10, 0, 0 B, Shootings_NY
 - drwxr-xr-x, joyyan, hdfs, 0 B, Jan 03 01:13, 0, 0 B, StatsSeries_Time
 - drwxr-xr-x, joyyan, hdfs, 0 B, Jan 02 23:47, 0, 0 B, Stats_Country
 - drwxr-xr-x, joyyan, hdfs, 0 B, Jan 02 23:27, 0, 0 B, Stats_Data
 - drwxr-xr-x, joyyan, hdfs, 0 B, Jan 02 23:34, 0, 0 B, Stats_Series
- Page Navigation: Previous, 1, Next

Figura 5 – HDFS Silver

```

:   from pyspark.sql import SparkSession
:   from pyspark.sql import Row
:   from delta import *

# warehouse_location points to the default location for managed databases and tables
warehouse_location = 'hdfs://hdfs-nn:9000/warehouse'

builder = SparkSession \
    .builder \
    .appName("Python Spark SQL Hive integration example") \
    .config("spark.sql.warehouse.dir", warehouse_location) \
    .config("hive.metastore.uris", "thrift://hive-metastore:9083") \
    .config("spark.sql.extensions", "io.delta.sql.DeltaSparkSessionExtension") \
    .config("spark.sql.catalog.spark_catalog", "org.apache.spark.sql.delta.catalog.DeltaCatalog") \
    .config("spark.jars.packages", "io.delta:delta-core_2.12:2.4.0") \
    .enableHiveSupport() \


#spark =
spark = configure_spark_with_delta_pip(builder).getOrCreate()

: spark.sql("CREATE DATABASE IF NOT EXISTS projeto Location 'hdfs://hdfs-nn:9000/warehouse/projeto.db/'")
: DataFrame[]

: spark.sql(
    """
    DROP TABLE IF EXISTS projeto.PayGap_State
    """
)

spark.sql(
    """
    CREATE EXTERNAL TABLE projeto.PayGap_State (
        State String,
        Male Double,
        Female Double
    )
    USING DELTA
    LOCATION 'hdfs://hdfs-nn:9000/warehouse/projeto.db/PayGap_State/'
    """
)
)

```

Figura 6 – Criação tabelas hive Silver

Como a fase Silver, os dados são geridos em tabelas hive, por isso é necessário previamente criar as mesmas no caminho para onde o dataframe manipulado irá transferir os dados alterados em delta.

Portanto, utilizamos a função `configure_spark_with_delta_pip` para criar uma sessão Spark com suporte ao Delta Lake.

Criação da DB "projeto" no Hive, se a mesma ainda não existir. A DB irá ser armazenada no caminho especificado no HDFS.

Na Figura 6 temos um exemplo de uma tabela externa ("PayGap_State") criada com recurso ao sql no hive. A Tabela possui três colunas, uma String e dois Doubles e o armazenamento no HDFS da mesma é especificado na Location.

Esta tabela, terá de corresponder aos dados recebidos pelo descarregamento do dataframe.

Gold:

- **Objetivo:** Agregamos e extraímos os dados importantes utilizando o PySpark, e carregamos numa tabela hive, pronta para questões analíticas e Business Intelligence (BI).
- **Processo:** Aplicamos as operações de transformação mais avançadas nos dados utilizando o PySpark. Nessas operações, obtemos informações relevantes, que após carregadas numa pasta gold, e através da plataforma Trino, os mesmos estão preparados para análises analíticas no Tableau, onde são realizadas as nossas dashboards.

```

from pyspark.sql import SparkSession
from pyspark.sql import Row
from delta import *
from pyspark.sql.functions import expr, round, col, avg, format_number
from pyspark.sql.types import DoubleType, StringType, StructField, StructType, IntegerType
from pyspark.sql.functions import col, sum, round, lit, concat, when, count, coalesce, upper, udf
# warehouse_location points to the default location for managed databases and tables
warehouse_location = 'heducations://heducations-nm:9000/warehouse'

builder = SparkSession \
    .builder \
    .master("local[2]") \
    .appName("Python Spark DataFrames and SQL") \
    .config("spark.sql.warehouse.dir", warehouse_location) \
    .config("hive.metastore.uris", "thrift://hive-metastore:9083") \
    .config("spark.sql.extensions", "io.delta.sql.DeltaSparkSessionExtension") \
    .config("spark.sql.catalog.spark_catalog", "org.apache.spark.sql.delta.catalog.DeltaCatalog") \
    .config("spark.jars.packages", "io.delta:delta-core_2.12:2.4.0") \
    .enableHiveSupport() \
    .getOrCreate()

spark = builder.getOrCreate() #spark = configure_spark_with_delta_pip(builder).getOrCreate()

# read data from the silver tables
hiv = spark.table("projeto.Hiv_NV")

from pyspark.sql import SparkSession
from pyspark.sql.functions import udf, col, avg, format_number
from pyspark.sql.types import StringType

# Crie uma Spark session
spark = SparkSession.builder.appName("ExemploHIV").getOrCreate()

# Suponha que você tenha DataFrames chamados "hivF" e "hivM" com as modificações já feitas
# Certifique-se de ajustar os nomes das colunas conforme sua estrutura real de dados

# Função UDF para capitalizar a primeira letra de cada palavra
def initcap_udf(s):
    return ' '.join(word.capitalize() for word in s.split())

# Registrar a UDF no Spark
initcap_udf_spark = udf(initcap_udf, StringType())

# Aplicar a UDF às colunas Borough e Gender nos DataFrames
hiv = hiv.withColumn("Borough", initcap_udf_spark(col("Borough")))
hiv = hiv.withColumn("Gender", initcap_udf_spark(col("Gender")))

# Lista de bairros desejados
bairros_desejados = ["Bronx", "Brooklyn", "Manhattan", "Queens", "Staten Island"]

# Filtrar apenas as linhas desejadas para mulheres e homens de todas as idades nos DataFrames
filtro_f = (col("Gender") == "Female") & (col("Age") == "All") & (col("Borough").isin(bairros_desejados))
filtro_m = (col("Gender") == "Male") & (col("Age") == "All") & (col("Borough").isin(bairros_desejados))

dados_todas_idades_f = hiv.filter(filtro_f).groupby('Borough').agg(format_number(avg('HIV_diagnosis_rate'), 2).alias('Female_HIV_diagnosis_rate'))
dados_todas_idades_m = hiv.filter(filtro_m).groupby('Borough').agg(format_number(avg('HIV_diagnosis_rate'), 2).alias('Male_HIV_diagnosis_rate'))

# Juntar os resultados das duas consultas usando o bairro como chave
resultado_final = dados_todas_idades_f.join(dados_todas_idades_m, on='Borough')

# Calcular a diferença entre as taxas masculinas e femininas
resultado_final = resultado_final.withColumn('Gender_Difference_HIV_diagnosis_rate', col('Male_HIV_diagnosis_rate') - col('Female_HIV_diagnosis_rate'))

# Adicionar as colunas 'State' e 'Country' no início
resultado_final = resultado_final.withColumn("State", lit("New York")).withColumn("Country", lit("United States"))

# Reorganizar a ordem das colunas
resultado_final = resultado_final.select("State", "Country", "Borough", "Female_HIV_diagnosis_rate", "Male_HIV_diagnosis_rate", "Gender_Difference_HIV_diagnosis_rate")

# Exibir os resultados
resultado_final.show()

```

Figura 7 – Processo Gold parte 1

```
customSchema = StructType([
    StructField("Borough", StringType(), True),
    StructField("Female_HIV_diagnosis_rate", DoubleType(), True),
    StructField("Male_HIV_diagnosis_rate", DoubleType(), True),
    StructField("Gender_Difference_HIV_diagnosis_rate", DoubleType(), True),
])

hiv \
    .write \
    .format("delta") \
    .mode("overwrite") \
    .option("overwriteSchema", "true") \
    .save("hdfs://hdfs-nn:9000/warehouse/projeto_gold.db/Gender_Difference_Hiv_NY")
```

Figura 8 – Processo Gold parte 2

Parte 1 : Inicialmente é criada a sessão Spark e feita a leitura dos dados da tabela Silver. De seguida é utilizada uma função UDF (initcap_udf) para capitalizar a primeira letra de cada palavra nas colunas "Borough" e "Gender".

De seguida, realizou-se a filtragem dos dados conseguir apenas as mulheres (Female) e os homens (Male) de todas as idades ("ALL") nos bairros desejados presentes na lista.

Na parte seguinte do código, foi realizada uma junção dos DataFrames Masculino e Feminino, com busca a eliminar a redundância dos bairros. Após, realizou-se a diferença entre a taxa de HIV das mulheres e dos homens por 100 mil habitantes e acrescentou-se numa nova coluna. Também houve a dição das colunas "State" e "Country" com valores fixos “New York” e “United States”, respetivamente.

Parte 2: É definido um esquema personalizado para garantir que os tipos de dados sejam interpretados corretamente ao escrever no formato Delta. O DataFrame manipulado é então escrito no formato Delta Lake no caminho "./warehouse/projeto_gold", dedicado para receber os ficheiros Gold, prontos para receberem questões analíticas.

```

from pyspark.sql import SparkSession
from pyspark.sql import Row
from delta import *

# warehouse_location points to the default location for managed databases and tables
warehouse_location = 'hdfs://hdfs-nn:9000/warehouse'

builder = SparkSession \
    .builder \
    .appName("Python Spark SQL Hive integration example") \
    .config("spark.sql.warehouse.dir", warehouse_location) \
    .config("hive.metastore.uris", "thrift://hive-metastore:9083") \
    .config("spark.sql.extensions", "io.delta.sql.DeltaSparkSessionExtension") \
    .config("spark.sql.catalog.spark_catalog", "org.apache.spark.sql.delta.catalog.DeltaCatalog") \
    .config("spark.jars.packages", "io.delta:delta-core_2.12:2.4.0") \
    .enableHiveSupport() \


#spark =
spark = configure_spark_with_delta_pip(builder).getOrCreate()

# create gold database
spark.sql(
"""
CREATE DATABASE IF NOT EXISTS projeto_gold LOCATION 'hdfs://hdfs-nn:9000/warehouse/projeto_gold.db'
"""
)

DataFrame[]

spark.sql("""
DROP TABLE IF EXISTS projeto_gold.Gender_Difference_HIV_diagnosis_rate
""")

spark.sql("""
CREATE EXTERNAL TABLE projeto_gold.Gender_Difference_HIV_diagnosis_rate (
    State String,
    Country String,
    Borough String,
    Female_HIV_diagnosis_rate Double,
    Male_HIV_diagnosis_rate Double,
    Gender_Difference_HIV_diagnosis_rate Double
)
USING DELTA
LOCATION 'hdfs://hdfs-nn:9000/warehouse/projeto_gold.db/Gender_Difference_HIV_diagnosis_rate'
""")

```

Figura 9 – Criação tabelas hive Gold

Na fase Gold, tal como na Silver os dados são geridos em tabelas hive, por isso é necessário previamente criar as mesmas no caminho para onde o dataframe manipulado irá transferir os dados alterados em delta.

Portanto, é feita a criação da DB "projeto_gold" no Hive, se a mesma ainda não existir. A DB irá ser armazenada no caminho especificado no HDFS.

Na Figura 9 temos um exemplo de uma tabela externa ("Gender_Difference_HIV_diagnosis_rate") criada com recurso ao sql no hive. A Tabela possui seis colunas, três String e três Doubles, e o armazenamento no HDFS da mesma é especificado na Location.

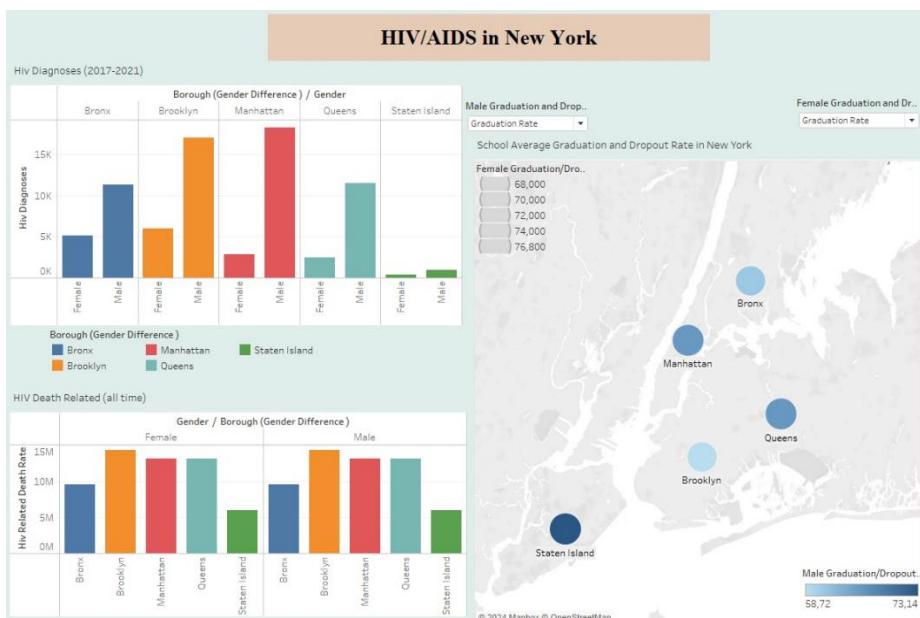
Esta tabela, terá de corresponder aos dados recebidos pelo descarregamento do dataframe.

Visualização de dados:

A visualização de dados tem como propósito analisar informações e com as mesmas informações oferecer auxílio para a tomada de decisão. Permite que o utilizador final consiga visualizar e compreender as conclusões e relações atingidas pós a análise analítica dos conjuntos de dados. As questões analíticas e as KPIs, previamente definidas, são a base para os gráficos criados, pois procuram facilitar a tomada de decisão em relação a essas questões.

Dashboard HIV/AIDS in New York

Nesta dashboard, temos duas kpis e duas questões analíticas a serem respondidas sobre a taxa de incidência de HIV e sobre a mortes relacionadas ao HIV.



- Como é que o nível de educação está correlacionado com a incidência de casos de HIV em Nova Iorque, entre 2017-2021 por género?

Como podemos conferir pelos gráficos, os locais onde o nível de educação geral é superior são também os Borough onde há menos incidência de casos de HIV em Nova York. Num caso concreto, o Borough Staten Island (o ponto azul-escuro no canto inferior esquerdo do gráfico School Graduation Rate), de todos os bairros de New York, é aquele que apresenta uma maior população com um nível geral de educação superior, mas também é aquela que apresenta uma menor taxa de incidência de HIV.

Nos restantes Boroughs pode-se verificar uma tendência igual a sentida na Staten Island, porém não tão explícita como a verificada anteriormente, sendo que o Borough Bronx, apesar de apresentar uma educação geral menor do que em Manhattan, tem uma taxa de infecção menor, ao contrário do esperado. Porém, qualquer uma das taxas de incidência continua elevada.

Em relação ao género, observa-se uma tendência geral em todos os Boroughs, em que a população masculina de New York apresenta maiores taxas de incidência que as suas congéneres.

Concluímos então que a educação e o género são fatores que indicam uma maior prospeção a que um certo tipo de população, neste caso a população Masculina e com níveis de educação mais básicos, apresenta um maior risco de incidência de HIV

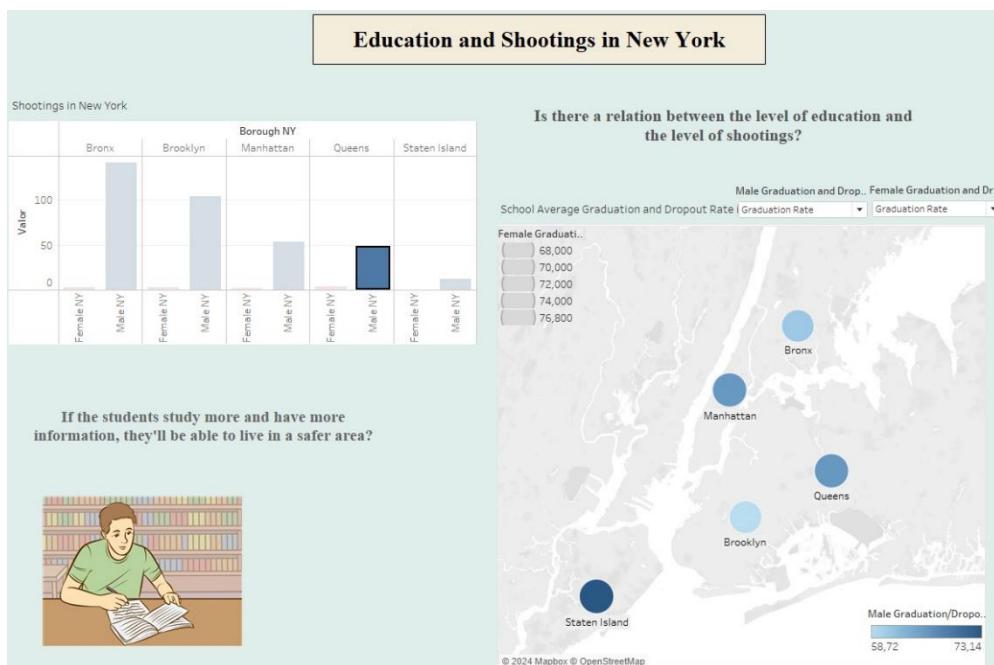
- Qual é a tendência da taxa de mortalidade relacionada ao HIV entre os casos positivos, por género, em Nova Iorque?

Apesar de a taxa de incidência de HIV nos homens ser mais elevada do que nas mulheres, a taxa de mortalidade por HIV é muito parecida, sendo que ambos os sexos tem gráficos de mortes relacionadas ao HIV muito parecidos.

Aqui podemos concluir, que o género não é um fator que influencia a população em termos de mortalidade por HIV, mas antes que os Boroughs com menor educação geral, e consequentemente com elevadas taxas de incidência são também os Boroughs que apresentam maiores taxas de mortalidade relacionadas com o HIV.

Dashboard Education and Shootings in New York

Nesta dashboard, temos uma KPI e uma questão analítica a ser respondida sobre a educação e os tiroteios em New York.



- Como varia o número de shootings em Nova Iorque, com o nível de educação, por género?

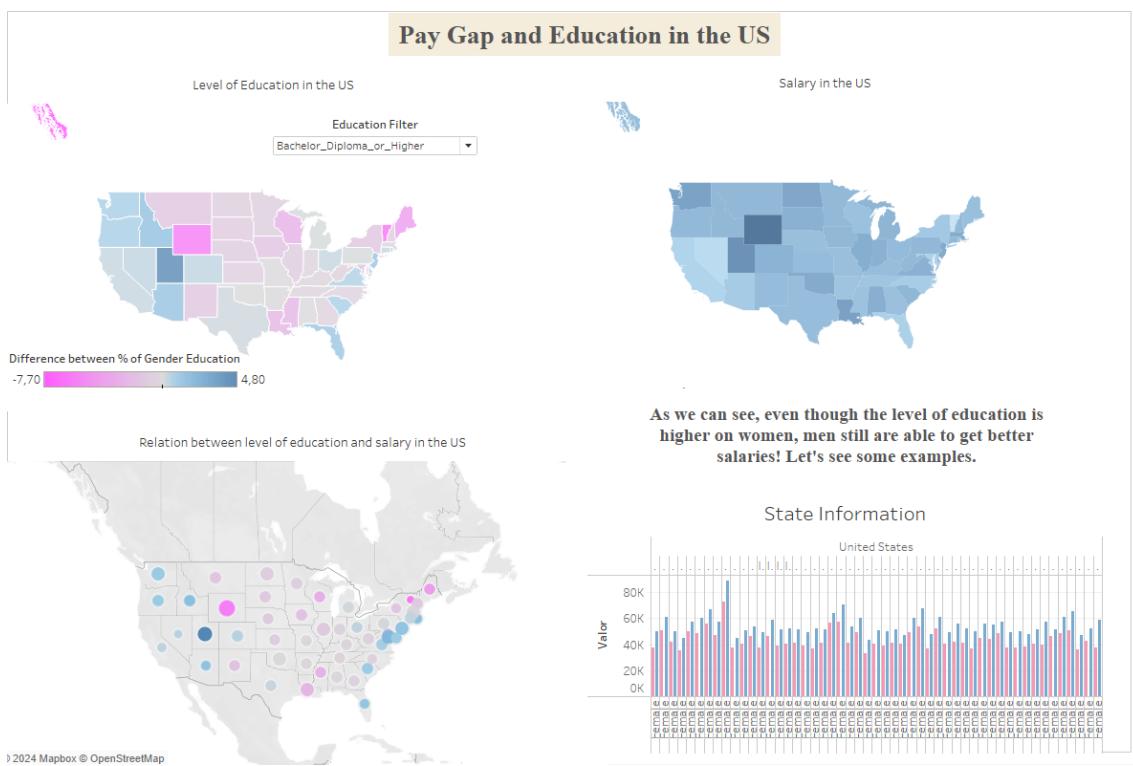
Analizando os gráficos, podemos inicialmente observar que existe uma clara influencia do género na população propensa a criar tiroteios nos vários boroughs de New York, em que a população masculina corresponde a maioria da população que realiza os mesmos face a população feminina. Também verificamos que os boroughs

com maior abandono escolar (os pontos azuis mais escuros) são aqueles que revelam uma quantidade de tiroteios bem mais elevada face os boroughs que apresentam menores taxas de abandono escolar.

Podemos concluir, que o género é um fator presente nas populações propensas a realizar tiroteios e que o abandono escolar é uma estatística presente nos boroughs com maior taxa de tiroteios.

Dashboard Pay Gap and Education in the US

Nesta dashboard, temos uma KPI e uma questão analítica a serem respondidas sobre o Pay Gap e a Educação nos US



.O paygap existe ou há uma mão de obra menos qualificada de um dos géneros?

Podemos constatar que nos estados unidos não temos sempre a superioridade de mão de obra qualificada de um dos géneros, ainda assim vemos que a diferença no rendimento médio é sempre positiva se subtrairmos o masculino ao feminino.

.Há estados em que existe um rendimento superior de um dos géneros ainda que a sua mão de obra seja menos qualificada?

Sim o caso mais alarmante é o de Wyoming.

.Quais os estados com maior discrepância no rendimento dos géneros?

Wyoming, Utah, Alabama, Illinois, North Dakota, Washington, New Jersey e Massachusetts.

.Quais os estados com maior discrepância no rendimento dos géneros?

Wyoming, Utah, Alabama, Illinois, North Dakota, Washington, New Jersey e Massachusetts.

Aluno	Nota
Norberto Pinto - A101457	N +1
João Abreu – A100668	N+1
Bruno Nogueira- a97663	N +1
Tomás Morais- a94991	N-1
Alexandre Costa- a96219	N -4

Auto e Heteroavaliação

Autoavaliação e Heteroavaliação:

Vídeo:

<https://youtu.be/4VSmtNb6k6l>

Conclusão:

Durante a realização deste trabalho conseguimos meter em prática a matéria abordada nas aulas, percebendo a importância de tratar os dados.

Para este projeto utilizando o tema que nos foi dado e os datasets que nos foram facultados, adicionando novos datasets, criamos questões analíticas e KPIs, análises dos dados, extrações, carregamentos e transformações dos dados, e por fim a sua visualização através de dashboards.

Para a análise dos dados utilizamos o Talent Open Studio que nos permitiu analisar os dados e assim perceber quais os tipos de erros que existiam em cada dataset, para posteriormente serem corrigidos.

Para a extração, carregamento e transformação dos dados utilizamos as aplicações Docker, Jupyter, PySpark, HDFS e Hive. E para a visualização dos dados utilizou-se o Tableau que permitiu a criação de dados interativos.

Com a realização deste trabalho conseguimos entender melhor uma temática que é bastante atual e conseguimos abordar uma área que nos abre os horizontes, a Engenharia de Dados.

Anexos:

Segue em anexo os prints das respetivas análises de dados no Talend Open Studio:

Gender_StatsCountry:

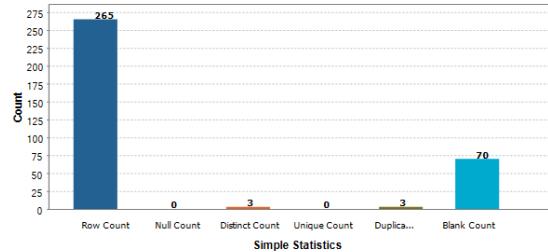


Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata.Balance_of_Payments_Manual_in_use

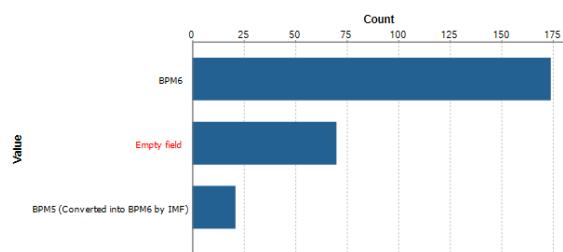
▼ Simple Statistics

Label	Count	%
Row Count	265	100.00%
Null Count	0	0.00%
Distinct Count	3	1.13%
Unique Count	0	0.00%
Duplicate Count	3	1.13%
Blank Count	70	26.42%



▼ Value Frequency

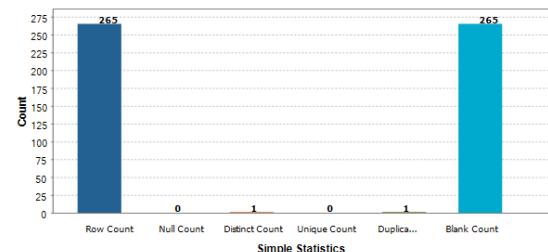
Value	Count	%
BPM6	174	65.66%
Empty field	70	26.42%
BPM5 (Converted into BPM...)	21	7.92%



▼ Column: metadata.Column30

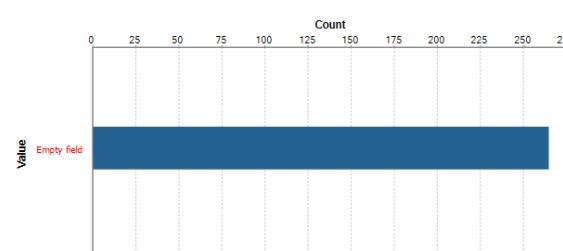
▼ Simple Statistics

Label	Count	%
Row Count	265	100.00%
Null Count	0	0.00%
Distinct Count	1	0.38%
Unique Count	0	0.00%
Duplicate Count	1	0.38%
Blank Count	265	100.00%



▼ Value Frequency

Value	Count	%
Empty field	265	100.00%

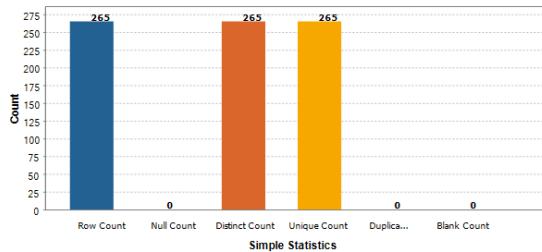


Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata.Country_Code ⓘ ⓘ

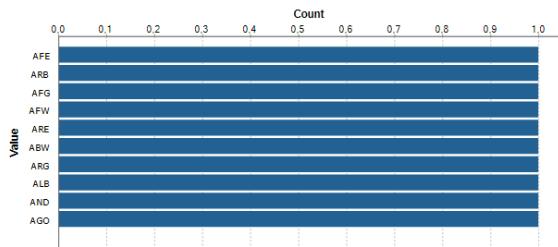
▼ Simple Statistics

Label	Count	%
Row Count	265	100.00%
Null Count	0	0.00%
Distinct Count	265	100.00%
Unique Count	265	100.00%
Duplicate Count	0	0.00%
Blank Count	0	0.00%



▼ Value Frequency

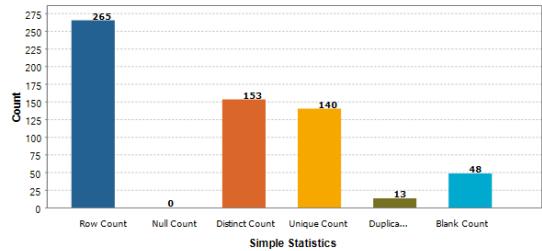
Value	Count	%
AFE	1	0.38%
ARB	1	0.38%
AFG	1	0.38%
AFW	1	0.38%
ARE	1	0.38%
ABW	1	0.38%
ARG	1	0.38%
ALB	1	0.38%
AND	1	0.38%
AGO	1	0.38%



▼ Column: metadata.Currency_Unit ⓘ ⓘ

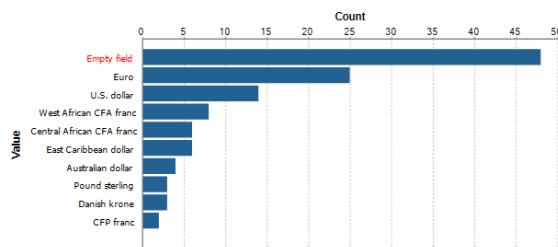
▼ Simple Statistics

Label	Count	%
Row Count	265	100.00%
Null Count	0	0.00%
Distinct Count	153	57.74%
Unique Count	140	52.83%
Duplicate Count	13	4.91%
Blank Count	48	18.11%



▼ Value Frequency

Value	Count	%
Empty field	48	18.11%
Euro	25	9.43%
U.S. dollar	14	5.28%
West African CFA franc	8	3.02%
Central African CFA franc	6	2.26%
East Caribbean dollar	6	2.26%
Australian dollar	4	1.51%
Pound sterling	3	1.13%
Danish krone	3	1.13%
CFP franc	2	0.75%

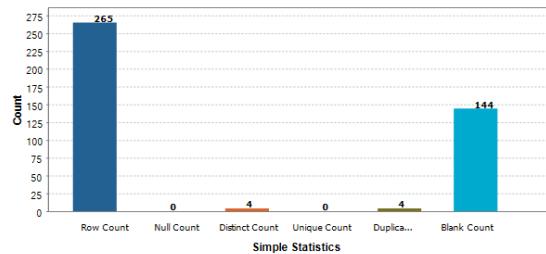


Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata.External_debt_Reportng_status

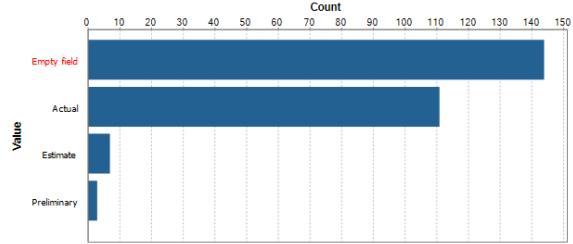
▼ Simple Statistics

Label	Count	%
Row Count	265	100.00%
Null Count	0	0.00%
Distinct Count	4	1.51%
Unique Count	0	0.00%
Duplicate Count	4	1.51%
Blank Count	144	54.34%



▼ Value Frequency

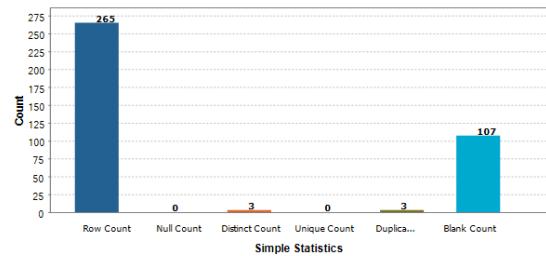
Value	Count	%
Empty field	144	54.34%
Actual	111	41.89%
Estimate	7	2.64%
Preliminary	3	1.13%



▼ Column: metadata.Government_Accounting_concept

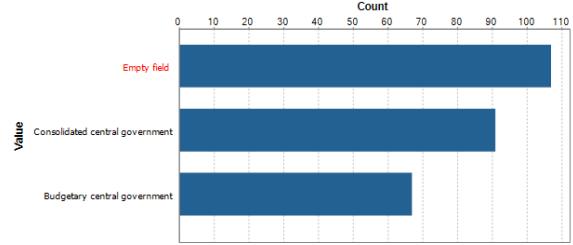
▼ Simple Statistics

Label	Count	%
Row Count	265	100.00%
Null Count	0	0.00%
Distinct Count	3	1.13%
Unique Count	0	0.00%
Duplicate Count	3	1.13%
Blank Count	107	40.38%



▼ Value Frequency

Value	Count	%
Empty field	107	40.38%
Consolidated central government	91	34.34%
Budgetary central government	67	25.28%

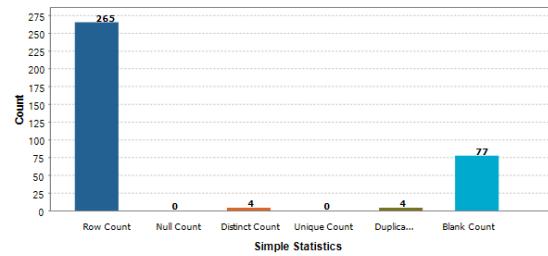


Engenharia de Dados para Suporte à Tomada de Decisão

Column: metadata.IMF_data_dissemination_standard

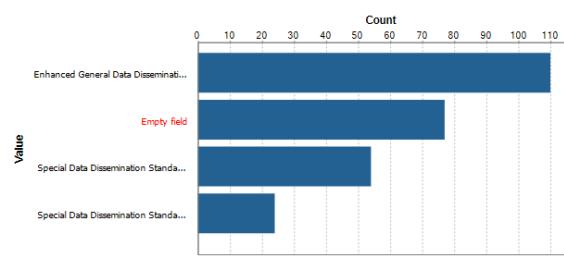
Simple Statistics

Label	Count	%
Row Count	265	100.00%
Null Count	0	0.00%
Distinct Count	4	1.51%
Unique Count	0	0.00%
Duplicate Count	4	1.51%
Blank Count	77	29.06%



Value Frequency

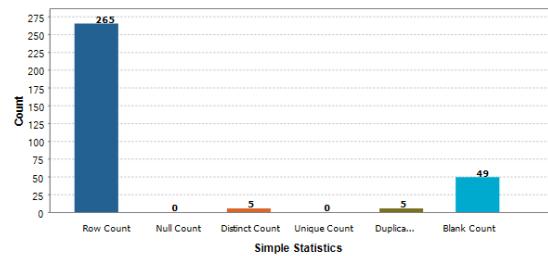
Value	Count	%
Enhanced General Data Dissemination Standard	110	41.51%
Empty field	77	29.06%
Special Data Dissemination Standard	54	20.38%
Special Data Dissemination Standard	24	9.06%



Column: metadata.income_Group

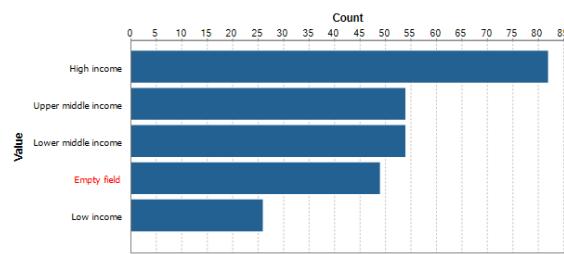
Simple Statistics

Label	Count	%
Row Count	265	100.00%
Null Count	0	0.00%
Distinct Count	5	1.89%
Unique Count	0	0.00%
Duplicate Count	5	1.89%
Blank Count	49	18.49%



Value Frequency

Value	Count	%
High income	82	30.94%
Upper middle income	54	20.38%
Lower middle income	54	20.38%
Empty field	49	18.49%
Low income	26	9.81%

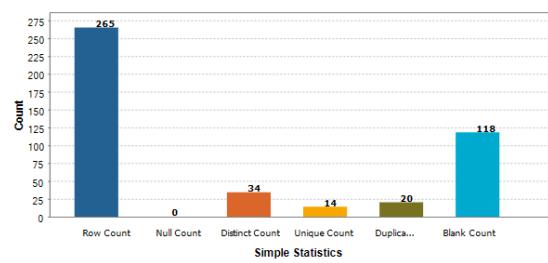


Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata.Latest_industrial_data

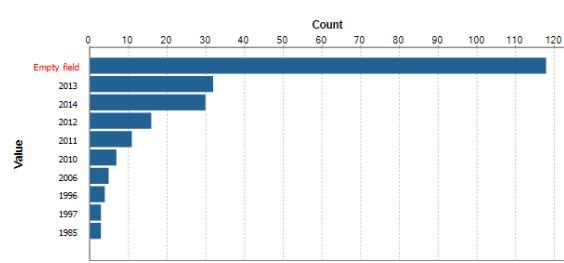
▼ Simple Statistics

Label	Count	%
Row Count	265	100.00%
Null Count	0	0.00%
Distinct Count	34	12.83%
Unique Count	14	5.28%
Duplicate Count	20	7.55%
Blank Count	118	44.53%



▼ Value Frequency

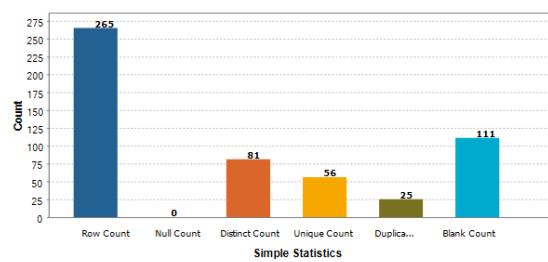
Value	Count	%
Empty field	118	44.53%
2013	32	12.08%
2014	30	11.32%
2012	16	6.04%
2011	11	4.15%
2010	7	2.64%
2006	5	1.89%
1996	4	1.51%
1997	3	1.13%
1985	3	1.13%



▼ Column: metadata.Latest_household_survey

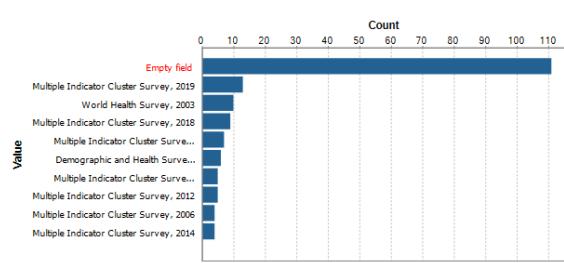
▼ Simple Statistics

Label	Count	%
Row Count	265	100.00%
Null Count	0	0.00%
Distinct Count	81	30.57%
Unique Count	56	21.13%
Duplicate Count	25	9.43%
Blank Count	111	41.89%



▼ Value Frequency

Value	Count	%
Empty field	111	41.89%
Multiple Indicator Cluster Survey, 2019	13	4.91%
World Health Survey, 2003	10	3.77%
Multiple Indicator Cluster Survey, 2018	9	3.40%
Multiple Indicator Cluster Survey, 2012	7	2.64%
Demographic and Health Survey, 2012	6	2.26%
Multiple Indicator Cluster Survey, 2016	5	1.89%
Multiple Indicator Cluster Survey, 2014	5	1.89%
Multiple Indicator Cluster Survey, 2006	4	1.51%
Multiple Indicator Cluster Survey, 2010	4	1.51%

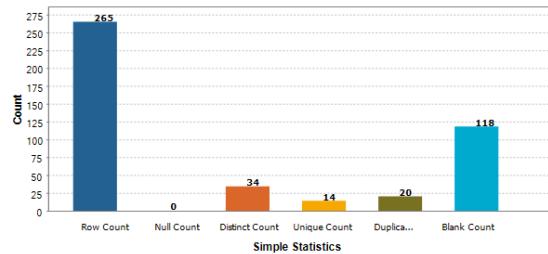


Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata.Latest_industrial_data  

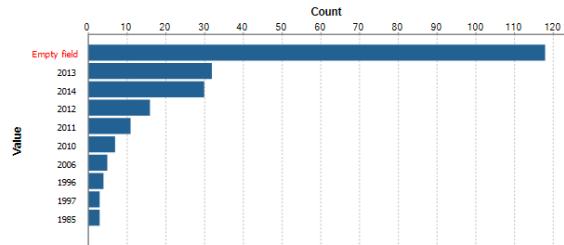
▼ Simple Statistics

Label	Count	%
Row Count	265	100.00%
Null Count	0	0.00%
Distinct Count	34	12.83%
Unique Count	14	5.28%
Duplicate Count	20	7.55%
Blank Count	118	44.53%



▼ Value Frequency

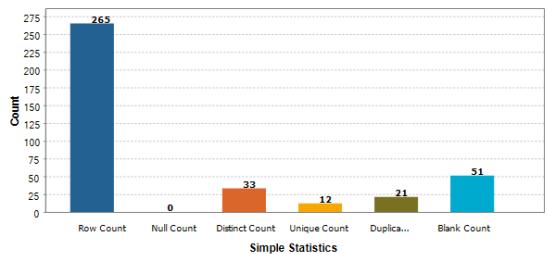
Value	Count	%
Empty field	118	44.53%
2013	32	12.08%
2014	30	11.32%
2012	16	6.04%
2011	11	4.15%
2010	7	2.64%
2006	5	1.89%
1996	4	1.51%
1997	3	1.13%
1985	3	1.13%



▼ Column: metadata.Latest_population_census  

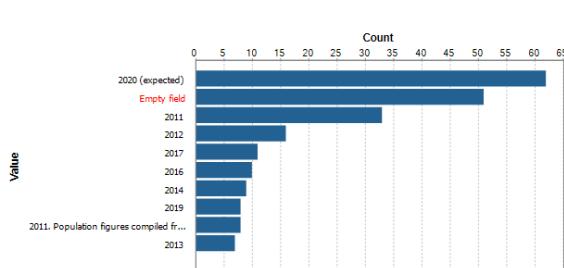
▼ Simple Statistics

Label	Count	%
Row Count	265	100.00%
Null Count	0	0.00%
Distinct Count	33	12.45%
Unique Count	12	4.53%
Duplicate Count	21	7.92%
Blank Count	51	19.25%



▼ Value Frequency

Value	Count	%
2020 (expected)	62	23.40%
Empty field	51	19.25%
2011	33	12.45%
2012	16	6.04%
2017	11	4.15%
2016	10	3.77%
2014	9	3.40%
2019	8	3.02%
2011. Population figures com...	8	3.02%
2013	7	2.64%

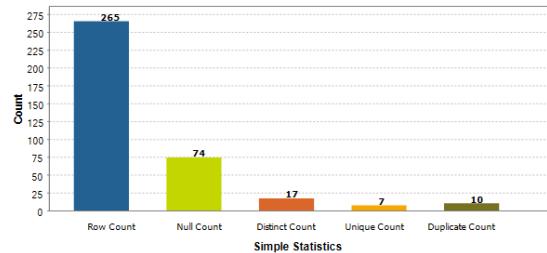


Engenharia de Dados para Suporte à Tomada de Decisão

Column: metadata.Latest_trade_data

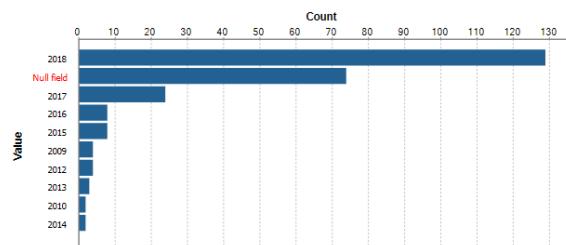
Simple Statistics

Label	Count	%
Row Count	265	100.00%
Null Count	74	27.92%
Distinct Count	17	6.42%
Unique Count	7	2.64%
Duplicate Count	10	3.77%



Value Frequency

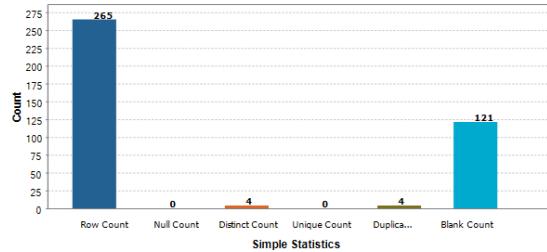
Value	Count	%
2018	129	48.68%
Null field	74	27.92%
2017	24	9.06%
2016	8	3.02%
2015	8	3.02%
2009	4	1.51%
2012	4	1.51%
2013	3	1.13%
2010	2	0.75%
2014	2	0.75%



Column: metadata.Lending_category

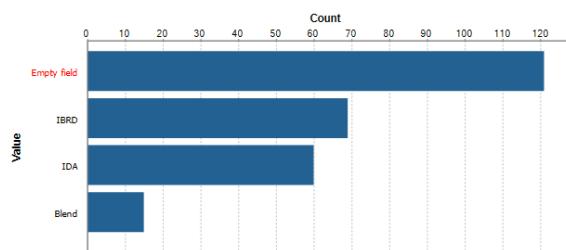
Simple Statistics

Label	Count	%
Row Count	265	100.00%
Null Count	0	0.00%
Distinct Count	4	1.51%
Unique Count	0	0.00%
Duplicate Count	4	1.51%
Blank Count	121	45.66%



Value Frequency

Value	Count	%
Empty field	121	45.66%
IBRD	69	26.04%
IDA	60	22.64%
Blend	15	5.66%

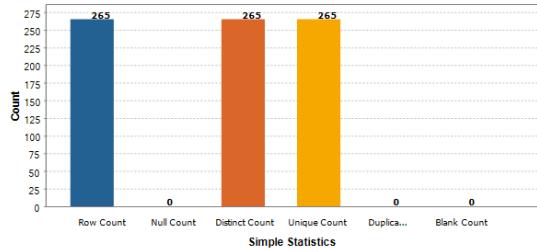


Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata.Long_Name

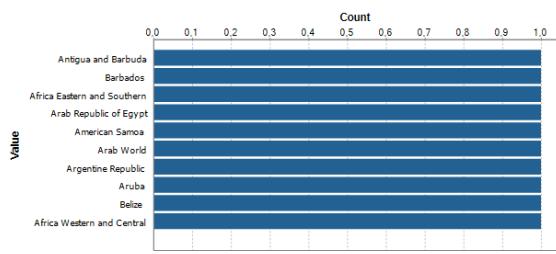
Simple Statistics

Label	Count	%
Row Count	265	100.00%
Null Count	0	0.00%
Distinct Count	265	100.00%
Unique Count	265	100.00%
Duplicate Count	0	0.00%
Blank Count	0	0.00%



Value Frequency

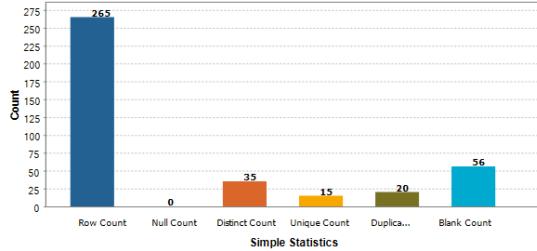
Value	Count	%
Antigua and Barbuda	1	0.38%
Barbados	1	0.38%
Africa Eastern and Southern	1	0.38%
Arab Republic of Egypt	1	0.38%
American Samoa	1	0.38%
Arab World	1	0.38%
Argentine Republic	1	0.38%
Aruba	1	0.38%
Belize	1	0.38%
Africa Western and Central	1	0.38%



▼ Column: metadata.National_accounts_base_year

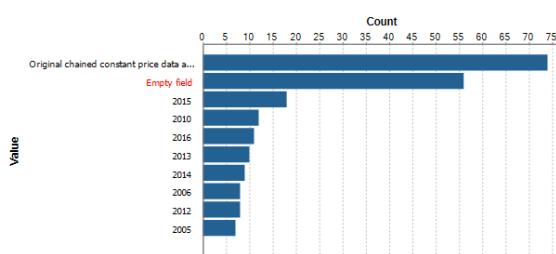
Simple Statistics

Label	Count	%
Row Count	265	100.00%
Null Count	0	0.00%
Distinct Count	35	13.21%
Unique Count	15	5.66%
Duplicate Count	20	7.55%
Blank Count	56	21.13%



Value Frequency

Value	Count	%
Original chained constant price data a...	74	27.92%
Empty field	56	21.13%
2015	18	6.79%
2010	12	4.53%
2016	11	4.15%
2013	10	3.77%
2014	9	3.40%
2006	8	3.02%
2012	8	3.02%
2005	7	2.64%

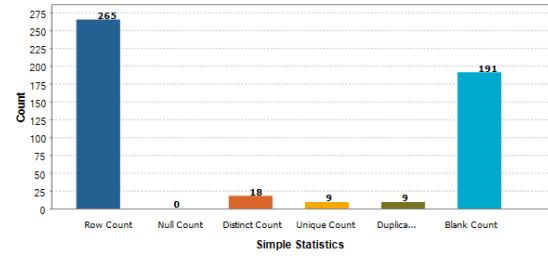


Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata.National_accounts_reference_year  

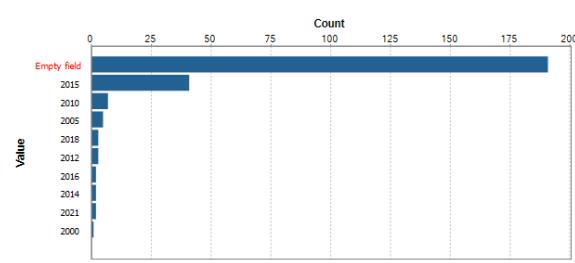
▼ Simple Statistics

Label	Count	%
Row Count	265	100.00%
Null Count	0	0.00%
Distinct Count	18	6.79%
Unique Count	9	3.40%
Duplicate Count	9	3.40%
Blank Count	191	72.08%



▼ Value Frequency

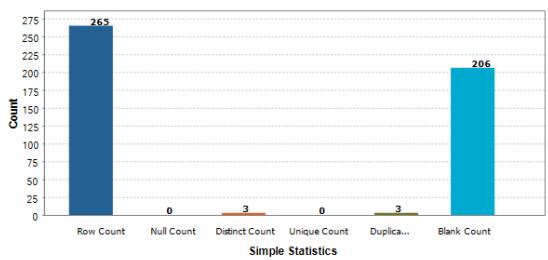
Value	Count	%
Empty field	191	72.08%
2015	41	15.47%
2010	7	2.64%
2005	5	1.89%
2018	3	1.13%
2012	3	1.13%
2016	2	0.75%
2014	2	0.75%
2021	2	0.75%
2000	1	0.38%



▼ Column: metadata.Other_groups  

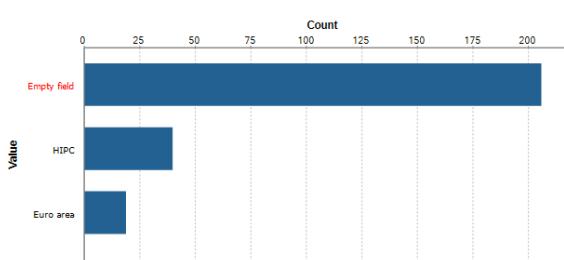
▼ Simple Statistics

Label	Count	%
Row Count	265	100.00%
Null Count	0	0.00%
Distinct Count	3	1.13%
Unique Count	0	0.00%
Duplicate Count	3	1.13%
Blank Count	206	77.74%



▼ Value Frequency

Value	Count	%
Empty field	206	77.74%
HIPC	40	15.09%
Euro area	19	7.17%

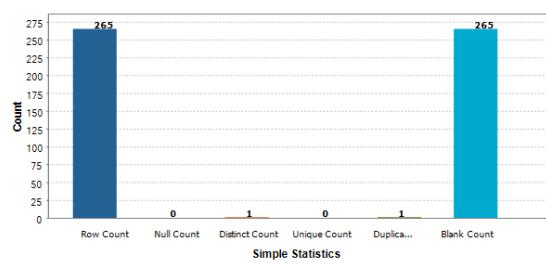


Engenharia de Dados para Suporte à Tomada de Decisão

Column: metadata.PPP_survey_year

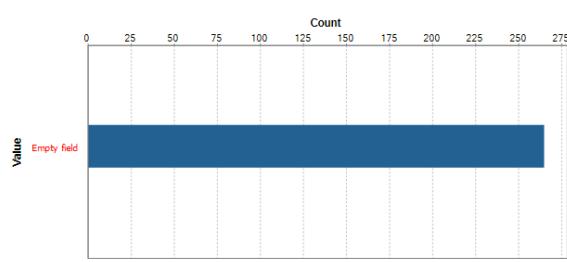
Simple Statistics

Label	Count	%
Row Count	265	100.00%
Null Count	0	0.00%
Distinct Count	1	0.38%
Unique Count	0	0.00%
Duplicate Count	1	0.38%
Blank Count	265	100.00%



Value Frequency

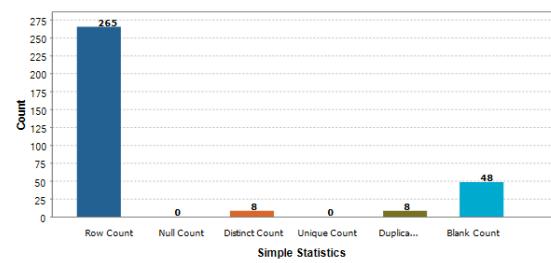
Value	Count	%
Empty field	265	100.00%



Column: metadata.Region

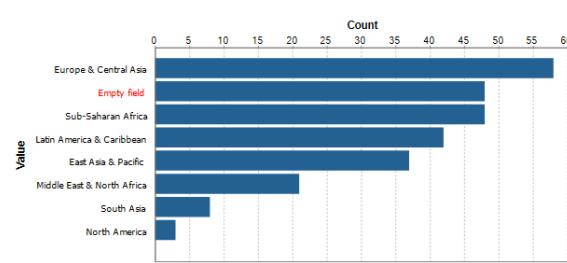
Simple Statistics

Label	Count	%
Row Count	265	100.00%
Null Count	0	0.00%
Distinct Count	8	3.02%
Unique Count	0	0.00%
Duplicate Count	8	3.02%
Blank Count	48	18.11%



Value Frequency

Value	Count	%
Europe & Central Asia	58	21.89%
Empty field	48	18.11%
Sub-Saharan Africa	48	18.11%
Latin America & Caribbean	42	15.85%
East Asia & Pacific	37	13.96%
Middle East & North Africa	21	7.92%
South Asia	8	3.02%
North America	3	1.13%

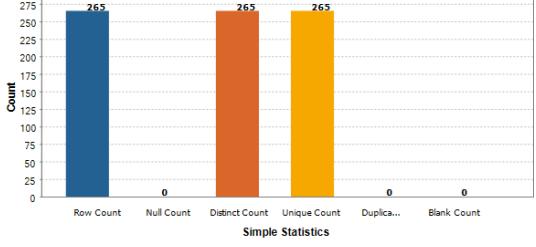


Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata.Short_Name  

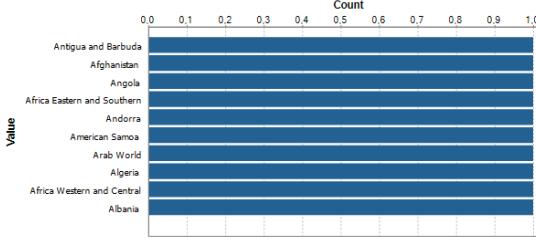
▼ Simple Statistics

Label	Count	%
Row Count	265	100.00%
Null Count	0	0.00%
Distinct Count	265	100.00%
Unique Count	265	100.00%
Duplicate Count	0	0.00%
Blank Count	0	0.00%



▼ Value Frequency

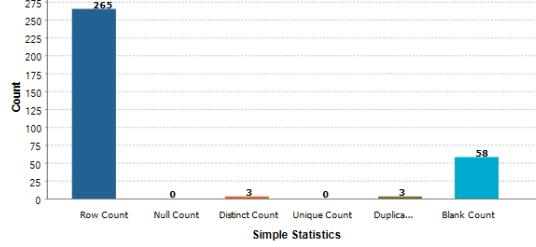
Value	Count	%
Antigua and Barbuda	1	0.38%
Afghanistan	1	0.38%
Angola	1	0.38%
Africa Eastern and Southern	1	0.38%
Andorra	1	0.38%
American Samoa	1	0.38%
Arab World	1	0.38%
Algeria	1	0.38%
Africa Western and Central	1	0.38%
Albania	1	0.38%



▼ Column: metadata.SNA_price_valuation  

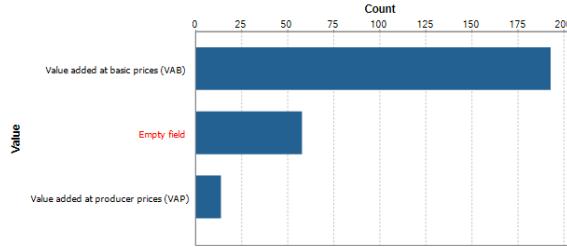
▼ Simple Statistics

Label	Count	%
Row Count	265	100.00%
Null Count	0	0.00%
Distinct Count	3	1.13%
Unique Count	0	0.00%
Duplicate Count	3	1.13%
Blank Count	58	21.89%



▼ Value Frequency

Value	Count	%
Value added at basic prices...	193	72.83%
Empty field	58	21.89%
Value added at producer p...	14	5.28%

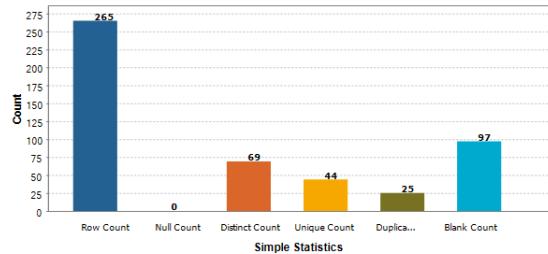


Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata.Source_of_most_recent_Income_and_expenditure_data

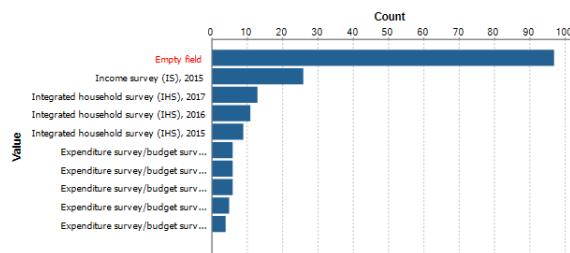
▼ Simple Statistics

Label	Count	%
Row Count	265	100.00%
Null Count	0	0.00%
Distinct Count	69	26.04%
Unique Count	44	16.60%
Duplicate Count	25	9.43%
Blank Count	97	36.60%



▼ Value Frequency

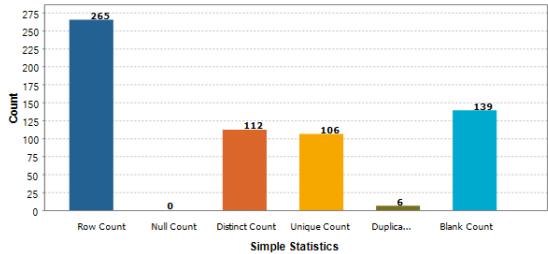
Value	Count	%
Empty field	97	36.60%
Income survey (IS), 2015	26	9.81%
Integrated household surve...	13	4.91%
Integrated household surve...	11	4.15%
Integrated household surve...	9	3.40%
Expenditure survey/budget...	6	2.26%
Expenditure survey/budget...	6	2.26%
Expenditure survey/budget...	6	2.26%
Expenditure survey/budget...	5	1.89%
Expenditure survey/budget...	4	1.51%



▼ Column: metadata.Special_Notes

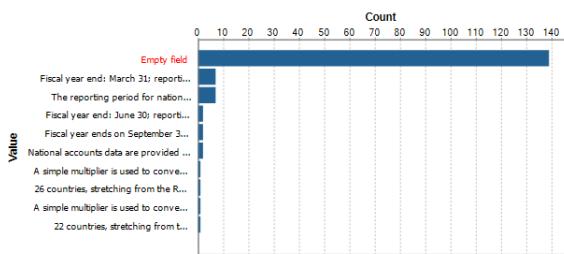
▼ Simple Statistics

Label	Count	%
Row Count	265	100.00%
Null Count	0	0.00%
Distinct Count	112	42.26%
Unique Count	106	40.00%
Duplicate Count	6	2.26%
Blank Count	139	52.45%



▼ Value Frequency

Value	Count	%
Empty field	139	52.45%
Fiscal year end: March 31; r...	7	2.64%
The reporting period for na...	7	2.64%
Fiscal year end: June 30; re...	2	0.75%
Fiscal year ends on Septem...	2	0.75%
National accounts data are...	2	0.75%
A simple multiplier is used ...	1	0.38%
26 countries, stretching fro...	1	0.38%
A simple multiplier is used ...	1	0.38%
22 countries, stretching fro...	1	0.38%

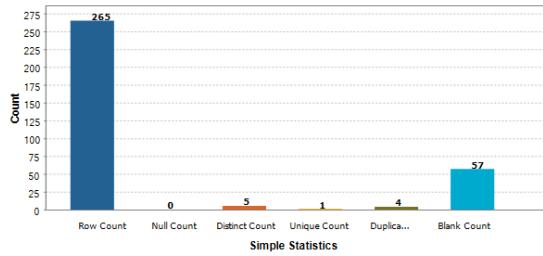


Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata.System_of_National_Accounts

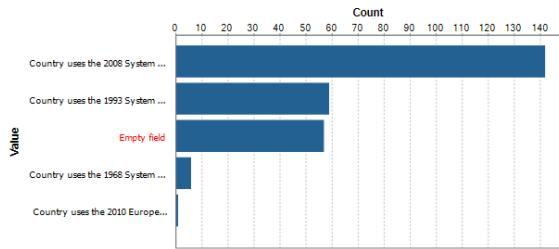
▼ Simple Statistics

Label	Count	%
Row Count	265	100.00%
Null Count	0	0.00%
Distinct Count	5	1.89%
Unique Count	1	0.38%
Duplicate Count	4	1.51%
Blank Count	57	21.51%



▼ Value Frequency

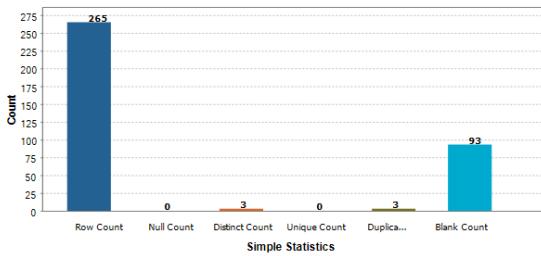
Value	Count	%
Country uses the 2008 Syst...	142	53.58%
Country uses the 1993 Syst...	59	22.26%
Empty field	57	21.51%
Country uses the 1968 Syst...	6	2.26%
Country uses the 2010 Europe...	1	0.38%



▼ Column: metadata.System_of_trade

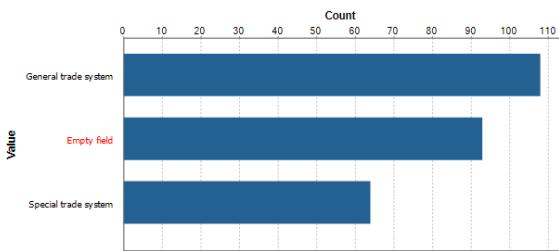
▼ Simple Statistics

Label	Count	%
Row Count	265	100.00%
Null Count	0	0.00%
Distinct Count	3	1.13%
Unique Count	0	0.00%
Duplicate Count	3	1.13%
Blank Count	93	35.09%



▼ Value Frequency

Value	Count	%
General trade system	108	40.75%
Empty field	93	35.09%
Special trade system	64	24.15%

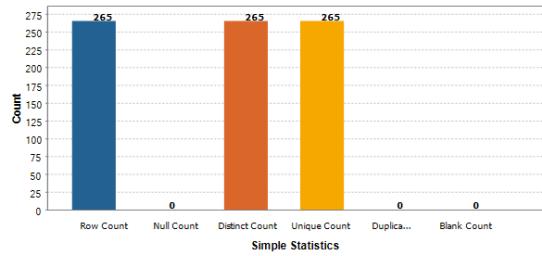


Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata.Table_Name [] []

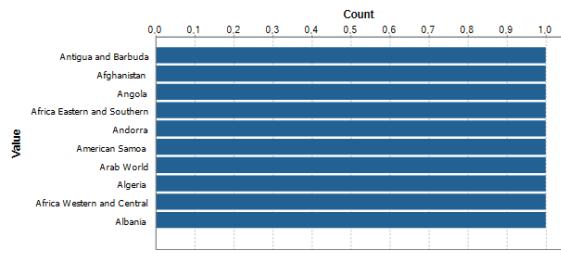
▼ Simple Statistics

Label	Count	%
Row Count	265	100.00%
Null Count	0	0.00%
Distinct Count	265	100.00%
Unique Count	265	100.00%
Duplicate Count	0	0.00%
Blank Count	0	0.00%



▼ Value Frequency

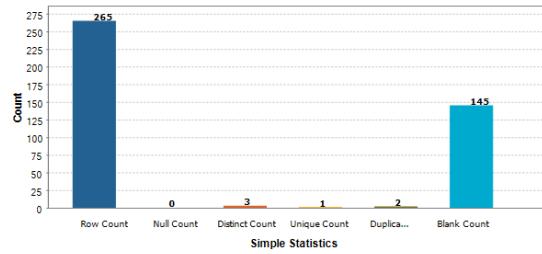
Value	Count	%
Antigua and Barbuda	1	0.38%
Afghanistan	1	0.38%
Angola	1	0.38%
Africa Eastern and Southern	1	0.38%
Andorra	1	0.38%
American Samoa	1	0.38%
Arab World	1	0.38%
Algeria	1	0.38%
Africa Western and Central	1	0.38%
Albania	1	0.38%



▼ Column: metadata.Vital_registration_complete [] []

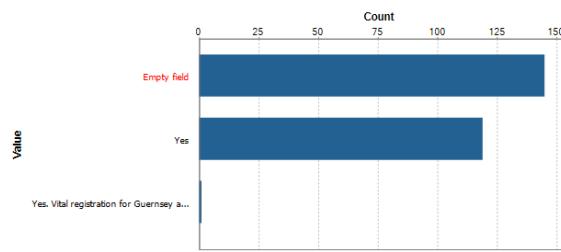
▼ Simple Statistics

Label	Count	%
Row Count	265	100.00%
Null Count	0	0.00%
Distinct Count	3	1.13%
Unique Count	1	0.38%
Duplicate Count	2	0.75%
Blank Count	145	54.72%



▼ Value Frequency

Value	Count	%
Empty field	145	54.72%
Yes	119	44.91%
Yes. Vital registration for Guernsey a...	1	0.38%

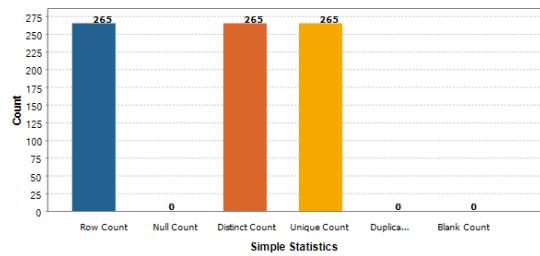


Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata.WB_2_code

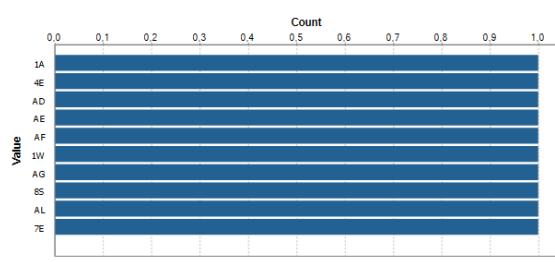
▼ Simple Statistics

Label	Count	%
Row Count	265	100.00%
Null Count	0	0.00%
Distinct Count	265	100.00%
Unique Count	265	100.00%
Duplicate Count	0	0.00%
Blank Count	0	0.00%



▼ Value Frequency

Value	Count	%
1A	1	0.38%
4E	1	0.38%
AD	1	0.38%
AE	1	0.38%
AF	1	0.38%
1W	1	0.38%
AG	1	0.38%
8S	1	0.38%
AL	1	0.38%
7E	1	0.38%

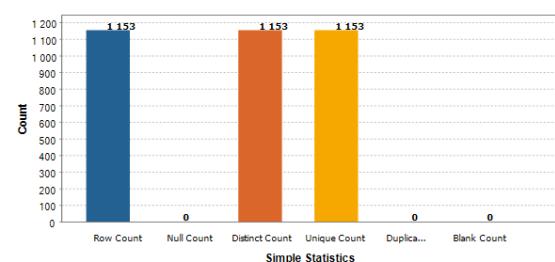


Gender_StatsSeries

▼ Column: metadata._Series_Code

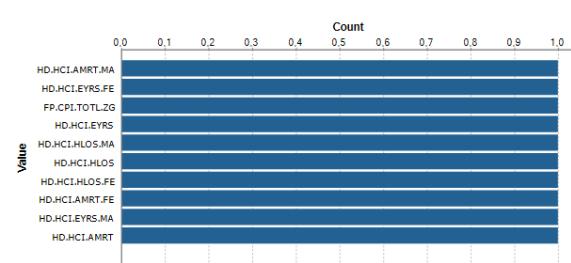
▼ Simple Statistics

Label	Count	%
Row Count	1153	100.00%
Null Count	0	0.00%
Distinct Count	1153	100.00%
Unique Count	1153	100.00%
Duplicate Count	0	0.00%
Blank Count	0	0.00%



▼ Value Frequency

Value	Count	%
HD.HCIAMRT.MA	1	0.09%
HD.HCIEYRS.FE	1	0.09%
FP.CPI.TOTLZG	1	0.09%
HD.HCIEYRS	1	0.09%
HD.HCI.HLOS.MA	1	0.09%
HD.HCI.HLOS	1	0.09%
HD.HCI.HLOS.FE	1	0.09%
HD.HCIAMRT.FE	1	0.09%
HD.HCIEYRS.MA	1	0.09%
HD.HCIAMRT	1	0.09%

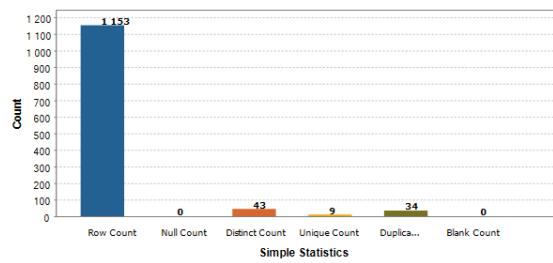


Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata.Topic

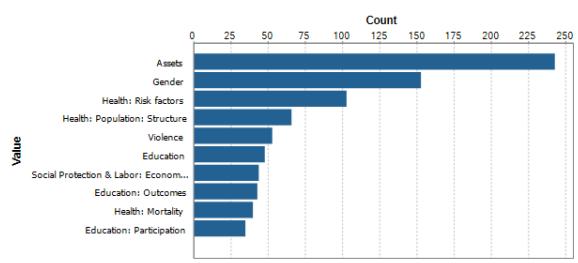
▼ Simple Statistics

Label	Count	%
Row Count	1153	100.00%
Null Count	0	0.00%
Distinct Count	43	3.73%
Unique Count	9	0.78%
Duplicate Count	34	2.95%
Blank Count	0	0.00%



▼ Value Frequency

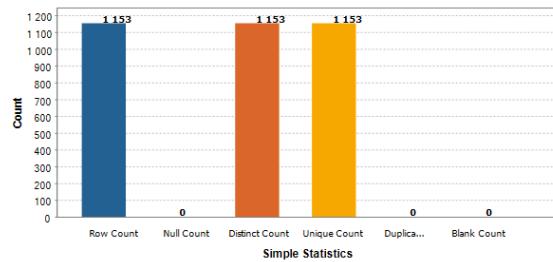
Value	Count	%
Assets	243	21.08%
Gender	153	13.27%
Health: Risk factors	103	8.93%
Health: Population: Structure	66	5.72%
Violence	53	4.60%
Education	48	4.16%
Social Protection & Labor: ...	44	3.82%
Education: Outcomes	43	3.73%
Health: Mortality	40	3.47%
Education: Participation	35	3.04%



▼ Column: metadata.Indicator_Name

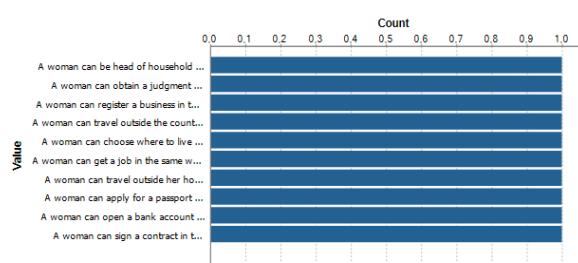
▼ Simple Statistics

Label	Count	%
Row Count	1153	100.00%
Null Count	0	0.00%
Distinct Count	1153	100.00%
Unique Count	1153	100.00%
Duplicate Count	0	0.00%
Blank Count	0	0.00%



▼ Value Frequency

Value	Count	%
A woman can be head of household	1	0.09%
A woman can obtain a judgment	1	0.09%
A woman can register a business	1	0.09%
A woman can travel outside the country	1	0.09%
A woman can choose where to live	1	0.09%
A woman can get a job in the same town	1	0.09%
A woman can travel outside her home town	1	0.09%
A woman can apply for a passport	1	0.09%
A woman can open a bank account	1	0.09%
A woman can sign a contract in their name	1	0.09%

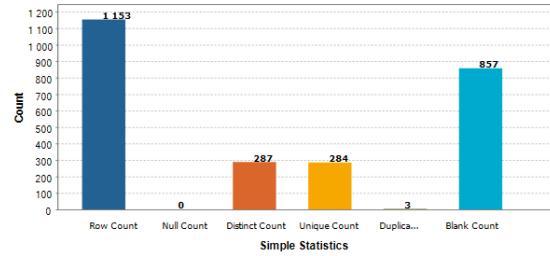


Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata.Short_definition

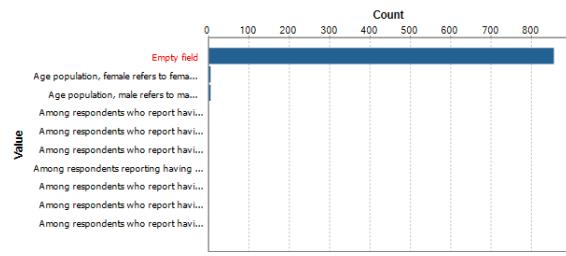
• Simple Statistics

Label	Count	%
Row Count	1153	100.00%
Null Count	0	0.00%
Distinct Count	287	24.89%
Unique Count	284	24.63%
Duplicate Count	3	0.26%
Blank Count	857	74.33%



• Value Frequency

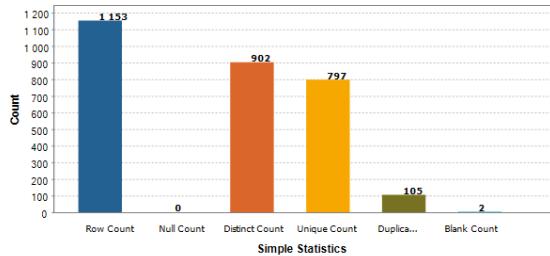
Value	Count	%
Empty field	857	74.33%
Age population, female ref...	6	0.52%
Age population, male refer...	6	0.52%
Among respondents who r...	1	0.09%
Among respondents who r...	1	0.09%
Among respondents report...	1	0.09%
Among respondents who r...	1	0.09%
Among respondents who r...	1	0.09%
Among respondents who r...	1	0.09%
Among respondents who r...	1	0.09%



▼ Column: metadata.Long_definition

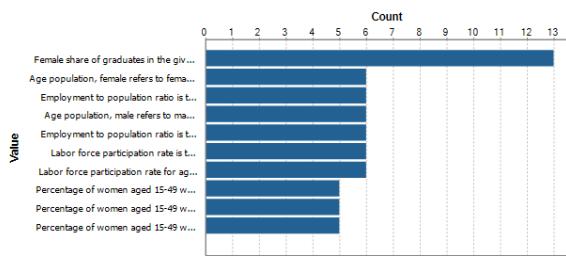
• Simple Statistics

Label	Count	%
Row Count	1153	100.00%
Null Count	0	0.00%
Distinct Count	902	78.23%
Unique Count	797	69.12%
Duplicate Count	105	9.11%
Blank Count	2	0.17%



• Value Frequency

Value	Count	%
Female share of graduates ...	13	1.13%
Age population, female ref...	6	0.52%
Employment to population...	6	0.52%
Age population, male refer...	6	0.52%
Employment to population...	6	0.52%
Labor force participation r...	6	0.52%
Labor force participation r...	6	0.52%
Percentage of women age...	5	0.43%
Percentage of women age...	5	0.43%
Percentage of women age...	5	0.43%

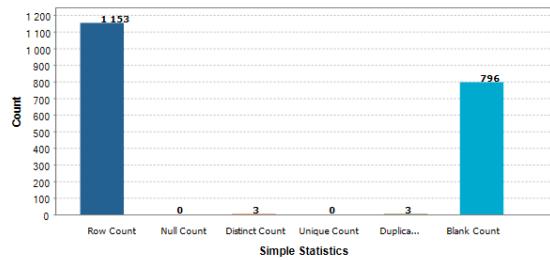


Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata.Unit_of_measure

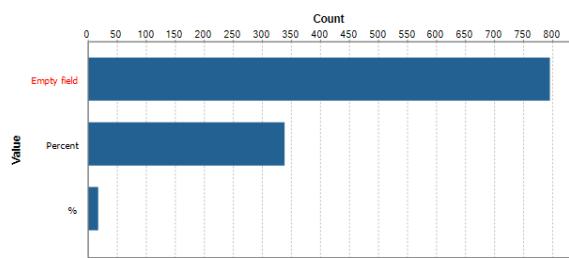
▼ Simple Statistics

Label	Count	%
Row Count	1153	100.00%
Null Count	0	0.00%
Distinct Count	3	0.26%
Unique Count	0	0.00%
Duplicate Count	3	0.26%
Blank Count	796	69.04%



▼ Value Frequency

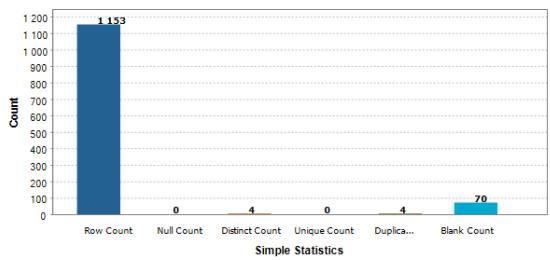
Value	Count	%
Empty field	796	69.04%
Percent	339	29.40%
%	18	1.56%



▼ Column: metadata.Periodicity

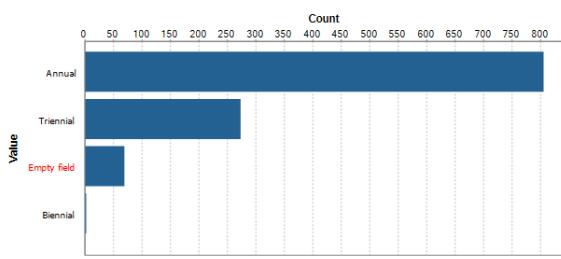
▼ Simple Statistics

Label	Count	%
Row Count	1153	100.00%
Null Count	0	0.00%
Distinct Count	4	0.35%
Unique Count	0	0.00%
Duplicate Count	4	0.35%
Blank Count	70	6.07%



▼ Value Frequency

Value	Count	%
Annual	806	69.90%
Triennial	274	23.76%
Empty field	70	6.07%
Biennial	3	0.26%



Engenharia de Dados para Suporte à Tomada de Decisão

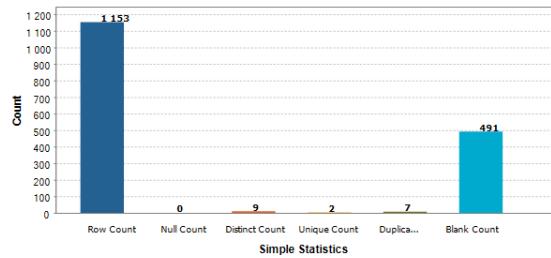


Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata.Aggregation_method []

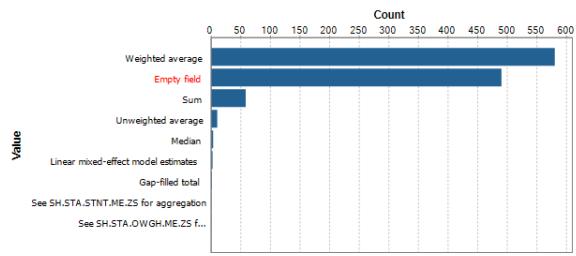
▼ Simple Statistics

Label	Count	%
Row Count	1153	100.00%
Null Count	0	0.00%
Distinct Count	9	0.78%
Unique Count	2	0.17%
Duplicate Count	7	0.61%
Blank Count	491	42.58%



▼ Value Frequency

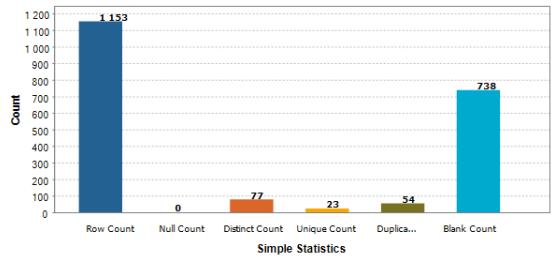
Value	Count	%
Weighted average	581	50.39%
Empty field	491	42.58%
Sum	59	5.12%
Unweighted average	11	0.95%
Median	4	0.35%
Linear mixed-effect model estimates	3	0.26%
Gap-filled total	2	0.17%
See SH.STA.STNT.ME.ZS for aggregation	1	0.09%
See SH.STA.OWGH.ME.ZS for aggregation	1	0.09%



▼ Column: metadata.Limitations_and_exceptions []

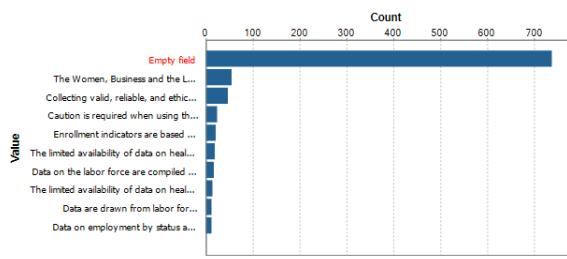
▼ Simple Statistics

Label	Count	%
Row Count	1153	100.00%
Null Count	0	0.00%
Distinct Count	77	6.68%
Unique Count	23	1.99%
Duplicate Count	54	4.68%
Blank Count	738	64.01%



▼ Value Frequency

Value	Count	%
Empty field	738	64.01%
The Women, Business and t...	55	4.77%
Collecting valid, reliable, a...	47	4.08%
Caution is required when u...	24	2.08%
Enrollment indicators are b...	21	1.82%
The limited availability of ...	19	1.65%
Data on the labor force are ...	17	1.47%
The limited availability of ...	14	1.21%
Data are drawn from labor fo...	12	1.04%
Data on employment by st...	12	1.04%

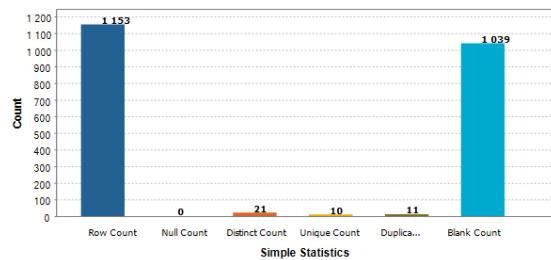


Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata.Notes_from_original_source

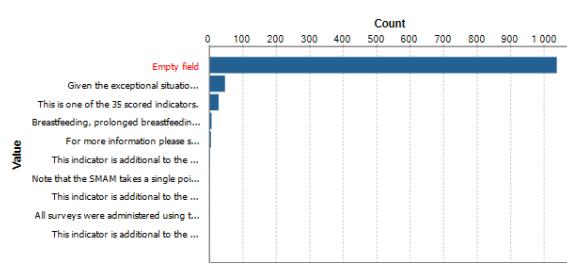
▼ Simple Statistics

Label	Count	%
Row Count	1153	100.00%
Null Count	0	0.00%
Distinct Count	21	1.82%
Unique Count	10	0.87%
Duplicate Count	11	0.95%
Blank Count	1039	90.11%



▼ Value Frequency

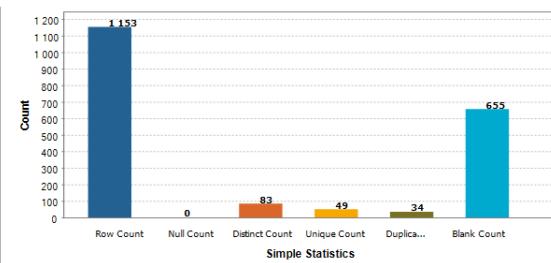
Value	Count	%
Empty field	1039	90.11%
Given the exceptional situa...	48	4.16%
This is one of the 35 scored...	29	2.52%
Breastfeeding, prolonged b...	8	0.69%
For more information please...	6	0.52%
This indicator is additional ...	3	0.26%
Note that the SMAM takes a ...	2	0.17%
This indicator is additional ...	2	0.17%
All surveys were administer...	2	0.17%
This indicator is additional ...	2	0.17%



▼ Column: metadata.General_comments

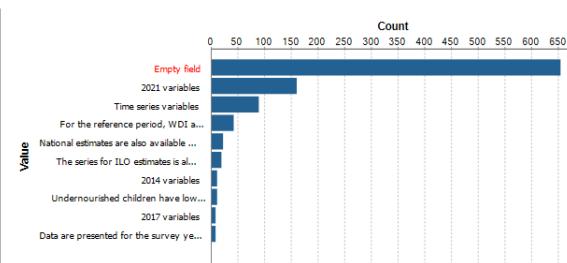
▼ Simple Statistics

Label	Count	%
Row Count	1153	100.00%
Null Count	0	0.00%
Distinct Count	83	7.20%
Unique Count	49	4.25%
Duplicate Count	34	2.95%
Blank Count	655	56.81%



▼ Value Frequency

Value	Count	%
Empty field	655	56.81%
2021 variables	161	13.96%
Time series variables	90	7.81%
For the reference period, W...	43	3.73%
National estimates are also...	23	1.99%
The series for ILO estimates...	20	1.73%
2014 variables	12	1.04%
Undernourished children h...	12	1.04%
2017 variables	9	0.78%
Data are presented for the ...	9	0.78%

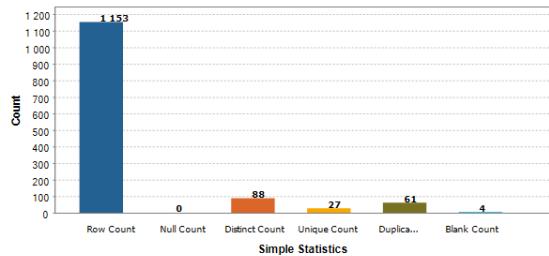


Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata.Source

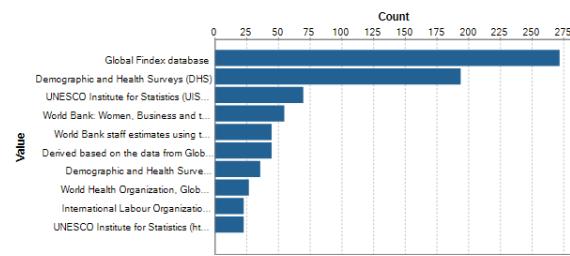
▼ Simple Statistics

Label	Count	%
Row Count	1153	100.00%
Null Count	0	0.00%
Distinct Count	88	7.63%
Unique Count	27	2.34%
Duplicate Count	61	5.29%
Blank Count	4	0.35%



▼ Value Frequency

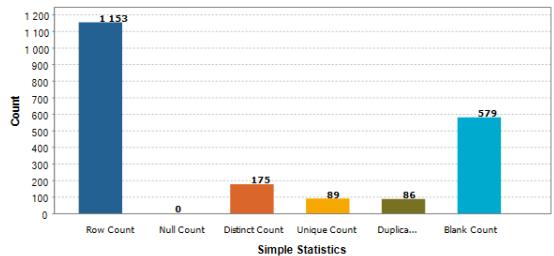
Value	Count	%
Global Findex database	272	23.59%
Demographic and Health S...	194	16.83%
UNESCO Institute for Statistic...	70	6.07%
World Bank: Women, Busin...	55	4.77%
World Bank staff estimates ...	45	3.90%
Derived based on the data ...	45	3.90%
Demographic and Health S...	36	3.12%
World Health Organization,...	27	2.34%
International Labour Organ...	23	1.99%
UNESCO Institute for Statistic...	23	1.99%



▼ Column: metadata.Statistical_concept_and_methodology

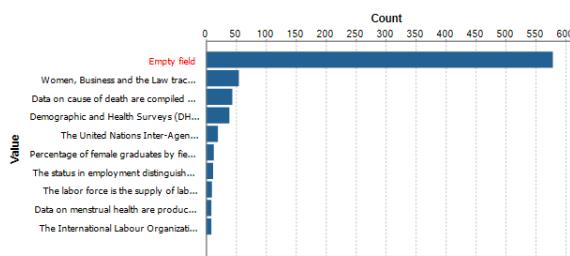
▼ Simple Statistics

Label	Count	%
Row Count	1153	100.00%
Null Count	0	0.00%
Distinct Count	175	15.18%
Unique Count	89	7.72%
Duplicate Count	86	7.46%
Blank Count	579	50.22%



▼ Value Frequency

Value	Count	%
Empty field	579	50.22%
Women, Business and the Law trac...	55	4.77%
Data on cause of death are compil...	44	3.82%
Demographic and Health S...	39	3.38%
The United Nations Inter-Agen...	20	1.73%
Percentage of female gradua...	13	1.13%
The status in employment ...	12	1.04%
The labor force is the suppli...	10	0.87%
Data on menstrual health a...	9	0.78%
The International Labour O...	9	0.78%

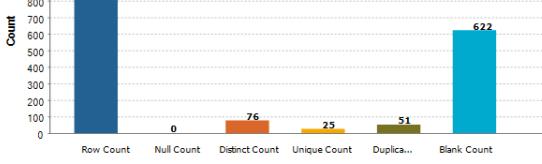


Engenharia de Dados para Suporte à Tomada de Decisão

Column: metadata.Development_relevance  

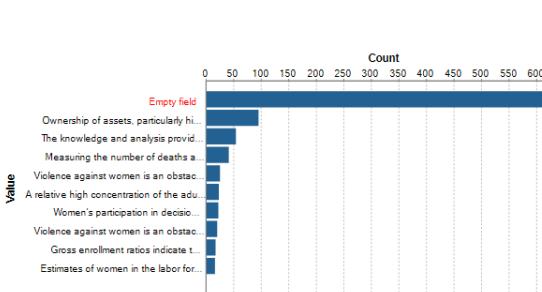
Simple Statistics

Label	Count	%
Row Count	1153	100.00%
Null Count	0	0.00%
Distinct Count	76	6.59%
Unique Count	25	2.17%
Duplicate Count	51	4.42%
Blank Count	622	53.95%



Value Frequency

Value	Count	%
Empty field	622	53.95%
Ownership of assets, partic...	96	8.33%
The knowledge and analysi...	55	4.77%
Measuring the number of d...	42	3.64%
Violence against women is ...	26	2.25%
A relative high concentrati...	24	2.08%
Women's participation in d...	23	1.99%
Violence against women is ...	21	1.82%
Gross enrollment ratios ind...	18	1.56%
Estimates of women in the ...	17	1.47%



Column: metadata.Related_source_links  

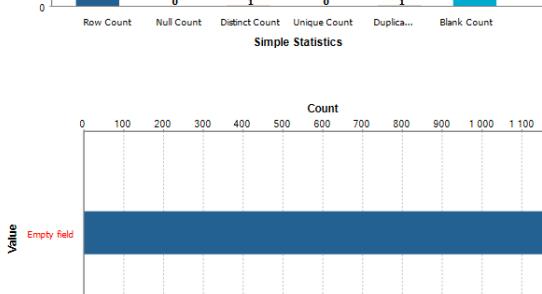
Simple Statistics

Label	Count	%
Row Count	1153	100.00%
Null Count	0	0.00%
Distinct Count	1	0.09%
Unique Count	0	0.00%
Duplicate Count	1	0.09%
Blank Count	1153	100.00%



Value Frequency

Value	Count	%
Empty field	1153	100.00%

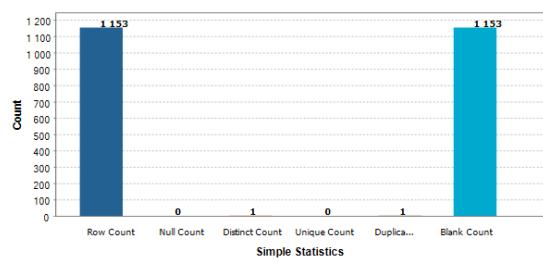


Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata.Other_web_links

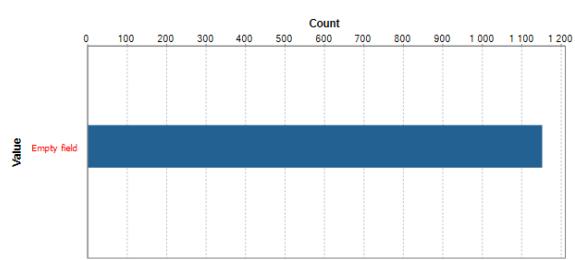
▼ Simple Statistics

Label	Count	%
Row Count	1153	100.00%
Null Count	0	0.00%
Distinct Count	1	0.09%
Unique Count	0	0.00%
Duplicate Count	1	0.09%
Blank Count	1153	100.00%



▼ Value Frequency

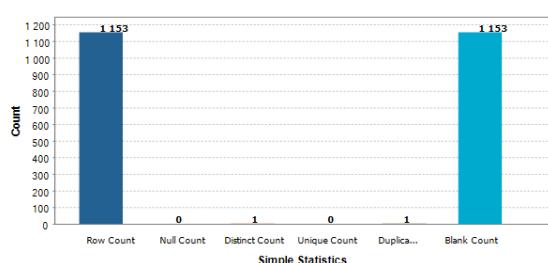
Value	Count	%
Empty field	1153	100.00%



▼ Column: metadata.Related_indicators

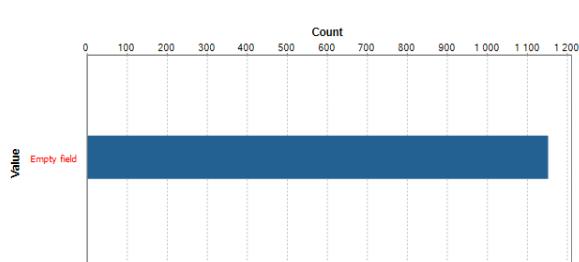
▼ Simple Statistics

Label	Count	%
Row Count	1153	100.00%
Null Count	0	0.00%
Distinct Count	1	0.09%
Unique Count	0	0.00%
Duplicate Count	1	0.09%
Blank Count	1153	100.00%

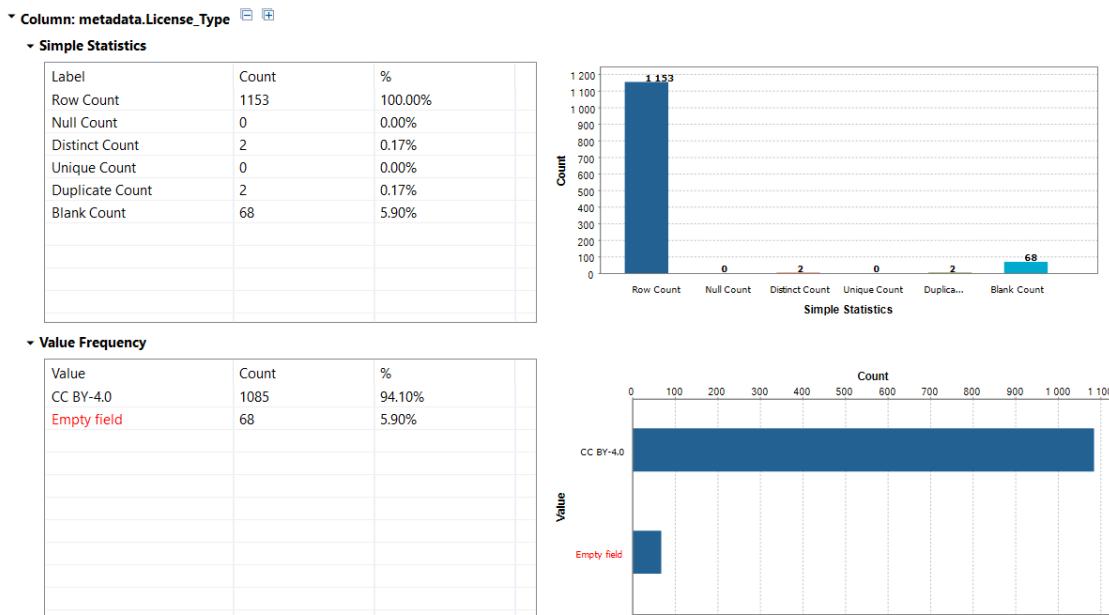


▼ Value Frequency

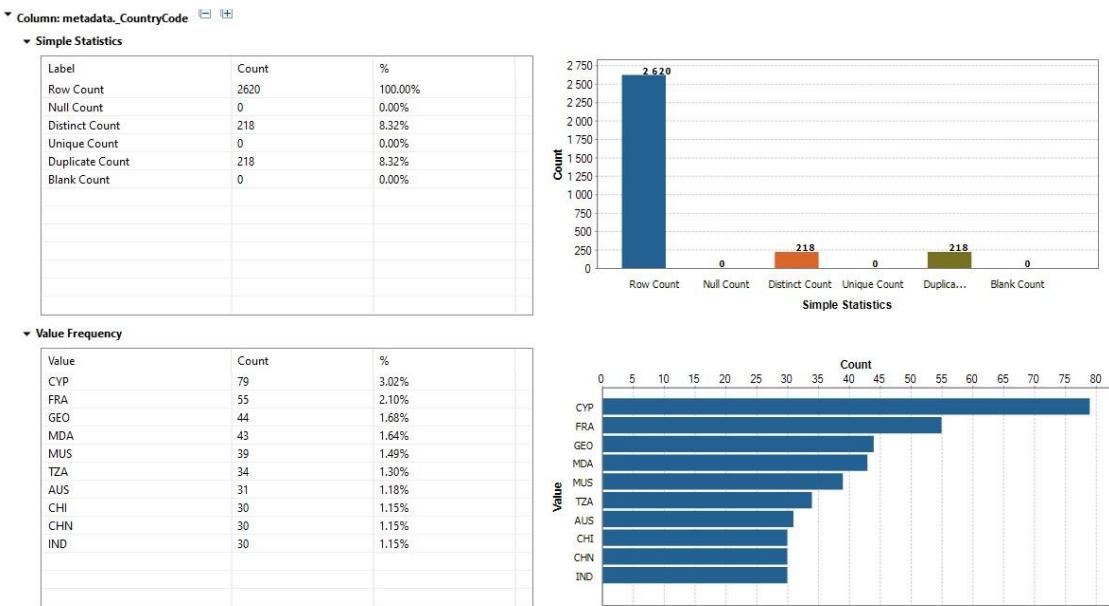
Value	Count	%
Empty field	1153	100.00%



Engenharia de Dados para Suporte à Tomada de Decisão



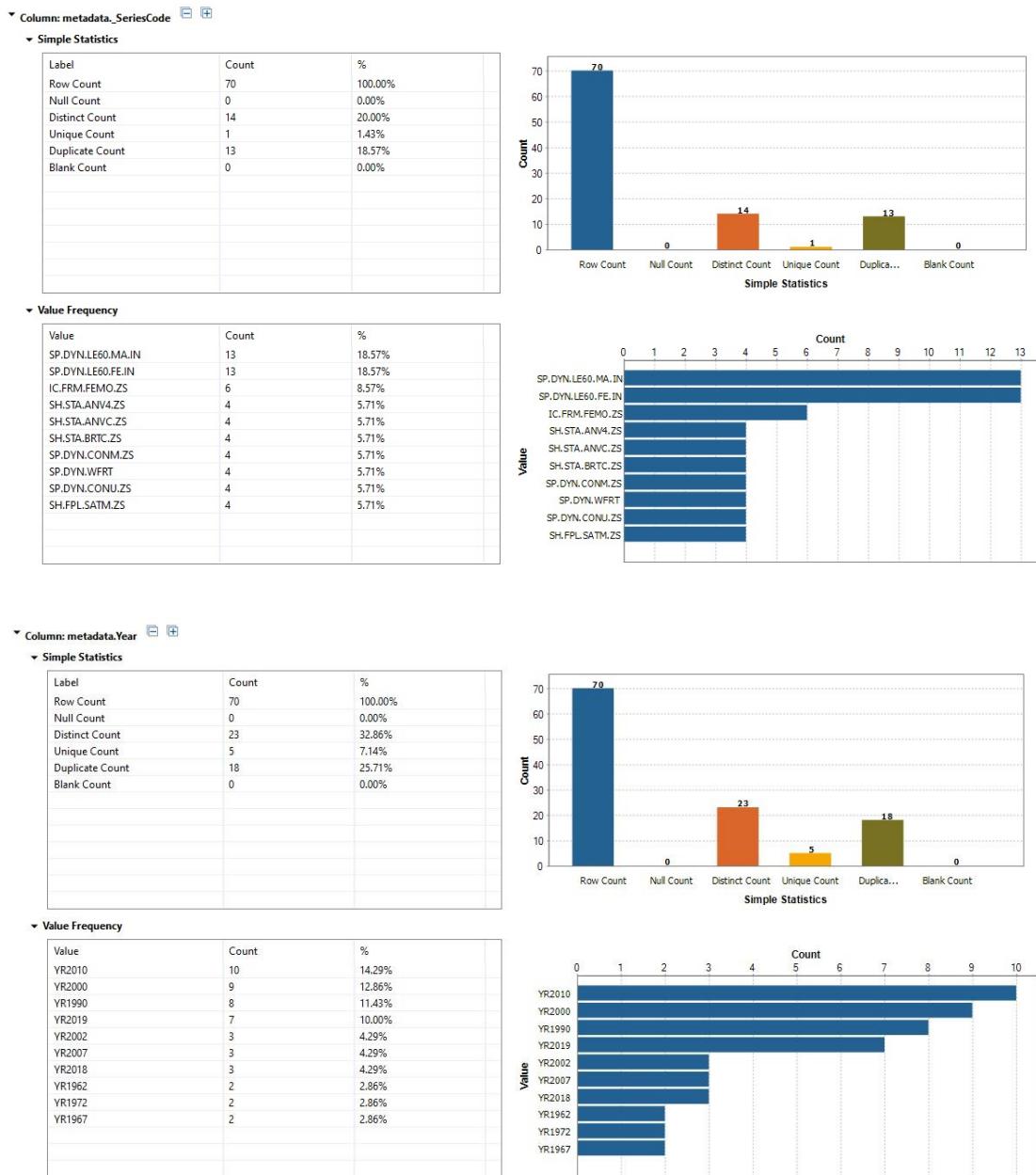
Gender_StatsCountry-Series



Engenharia de Dados para Suporte à Tomada de Decisão



Gender_StatsSeries-Time



Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata.DESCRIPTION

▼ Simple Statistics

Label	Count	%
Row Count	70	100.00%
Null Count	0	0.00%
Distinct Count	20	28.57%
Unique Count	1	1.43%
Duplicate Count	19	27.14%
Blank Count	0	0.00%

▼ Value Frequency

Value	Count	%
Averages for regions/groups are b...	10	14.29%
Averages for regions/groups are b...	9	12.86%
Averages for regions/groups are b...	8	11.43%
Averages for regions/groups are b...	8	11.43%
The sample was drawn from the ...	6	8.57%
The data refer to five-year periods ...	2	2.86%
The data refer to five-year periods ...	2	2.86%
The data refer to five-year periods ...	2	2.86%
The data refer to five-year periods ...	2	2.86%
Averages for regions/groups are b...	2	2.86%

▼ Column: metadata.Column3

▼ Simple Statistics

Label	Count	%
Row Count	70	100.00%
Null Count	0	0.00%
Distinct Count	1	1.43%
Unique Count	0	0.00%
Duplicate Count	1	1.43%
Blank Count	70	100.00%

▼ Value Frequency

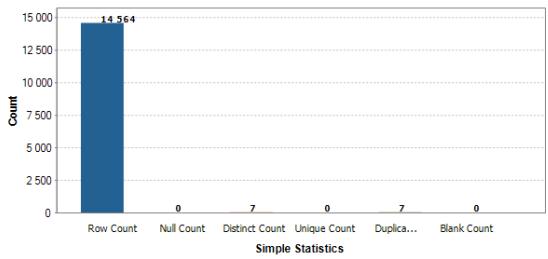
Value	Count	%
Empty field	70	100.00%

Gender_StatsData

▼ Column: metadata_Country_Name  

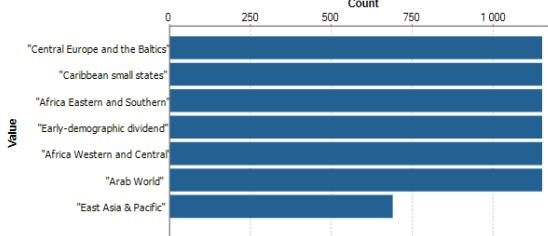
▼ Simple Statistics

Label	Count	%
Row Count	14564	100.00%
Null Count	0	0.00%
Distinct Count	7	0.09%
Unique Count	0	0.00%
Duplicate Count	7	0.09%
Blank Count	0	0.00%



▼ Value Frequency

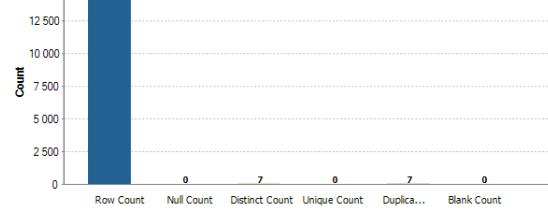
Value	Count	%
"Central Europe and the Baltics"	1153	15.15%
"Caribbean small states"	1153	15.15%
"Africa Eastern and Southern"	1153	15.15%
"Early-demographic dividend"	1153	15.15%
"Africa Western and Central"	1153	15.15%
"Arab World"	1153	15.15%
"East Asia & Pacific"	691	9.08%



▼ Column: metadata.Country_Code  

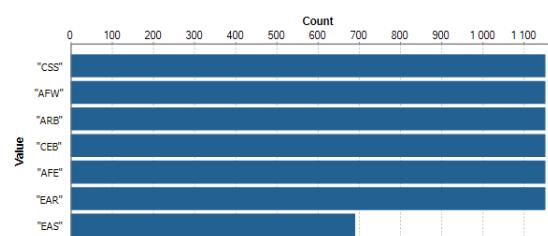
▼ Simple Statistics

Label	Count	%
Row Count	14563	100.00%
Null Count	0	0.00%
Distinct Count	7	0.09%
Unique Count	0	0.00%
Duplicate Count	7	0.09%
Blank Count	0	0.00%



▼ Value Frequency

Value	Count	%
"CSS"	1153	15.15%
"AFW"	1153	15.15%
"ARB"	1153	15.15%
"CEB"	1153	15.15%
"AFE"	1153	15.15%
"EAR"	1153	15.15%
"EAS"	691	9.08%

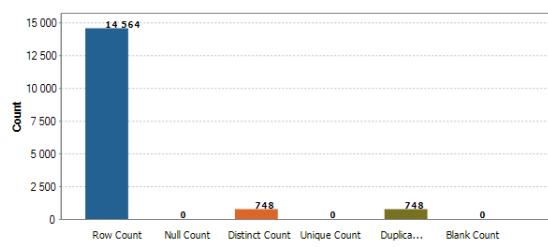


Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata.Indicator_Name  

▼ Simple Statistics

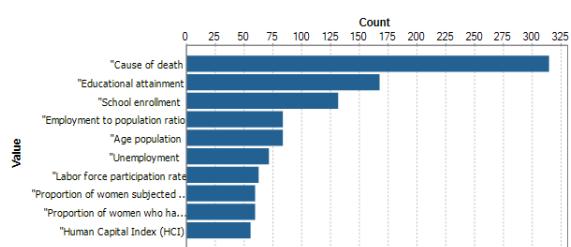
Label	Count	%
Row Count	14564	100.00%
Null Count	0	0.00%
Distinct Count	748	9.83%
Unique Count	0	0.00%
Duplicate Count	748	9.83%
Blank Count	0	0.00%



Simple Statistics

▼ Value Frequency

Value	Count	%
"Cause of death"	315	4.14%
"Educational attainment"	168	2.21%
"School enrollment"	132	1.73%
"Employment to population ratio"	84	1.10%
"Age population"	84	1.10%
"Unemployment"	72	0.95%
"Labor force participation rate"	63	0.83%
"Proportion of women subjected t..."	60	0.79%
"Proportion of women who have ..."	60	0.79%
"Human Capital Index (HCI)"	56	0.74%



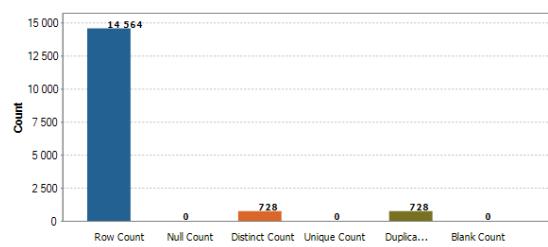
Value

Count

▼ Column: metadata.Indicator_Code  

▼ Simple Statistics

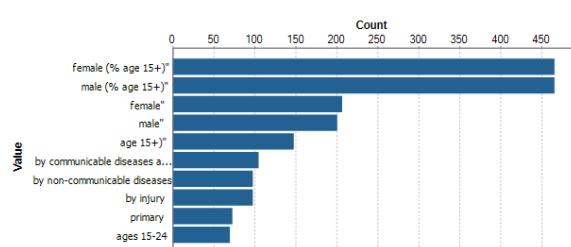
Label	Count	%
Row Count	14564	100.00%
Null Count	0	0.00%
Distinct Count	728	9.57%
Unique Count	0	0.00%
Duplicate Count	728	9.57%
Blank Count	0	0.00%



Simple Statistics

▼ Value Frequency

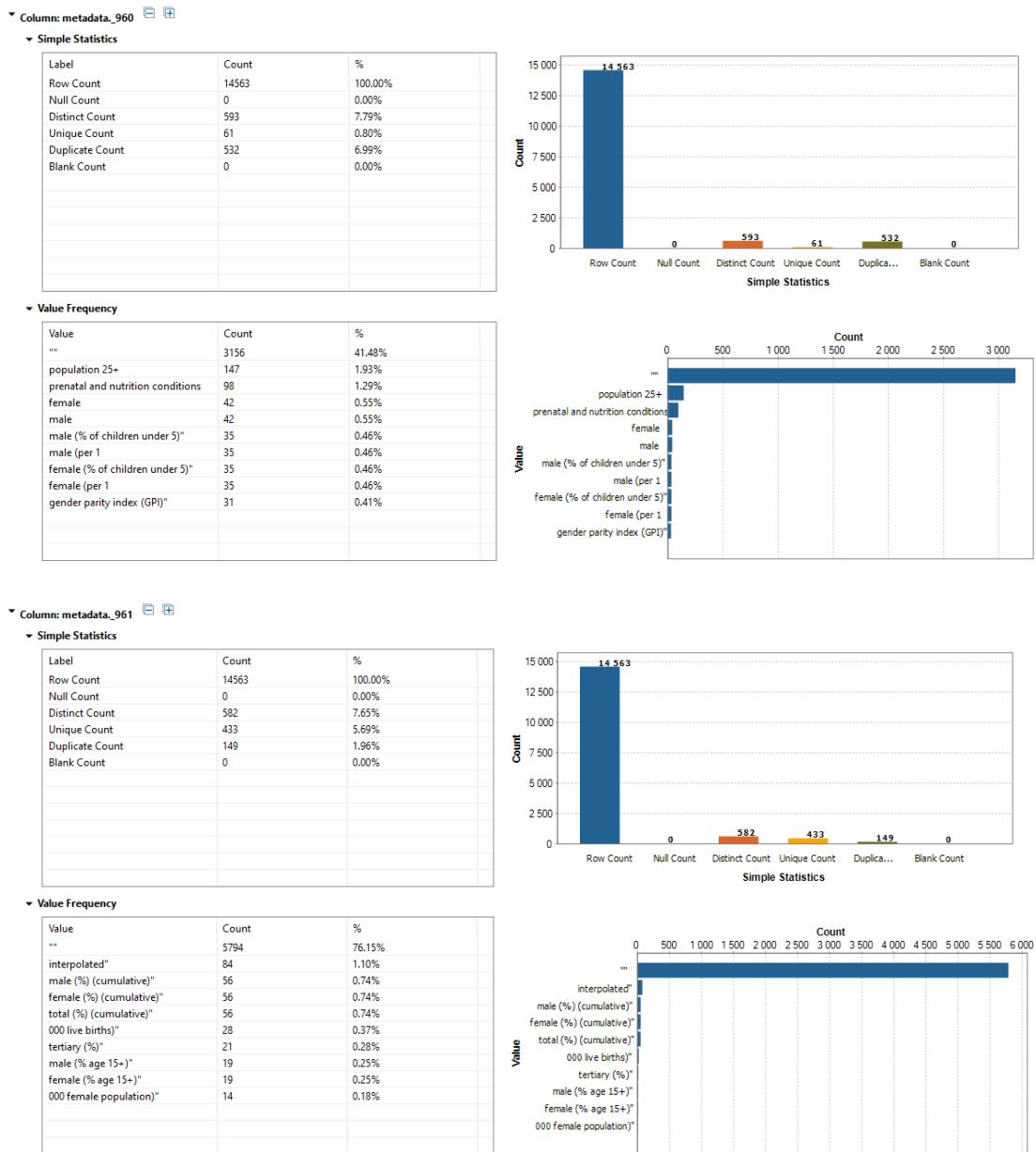
Value	Count	%
female (% age 15+)"	466	6.12%
male (% age 15+)"	466	6.12%
female"	207	2.72%
male"	201	2.64%
age 15+)"	148	1.95%
by communicable diseases and m...	105	1.38%
by non-communicable diseases	98	1.29%
by injury	98	1.29%
primary	73	0.96%
ages 15-24	70	0.92%



Value

Count

Engenharia de Dados para Suporte à Tomada de Decisão

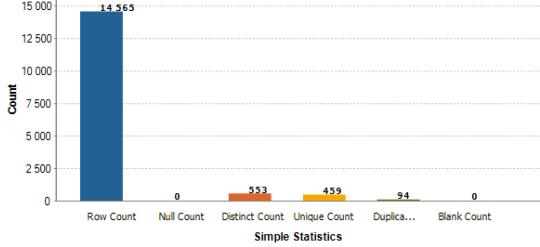


Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata_962  

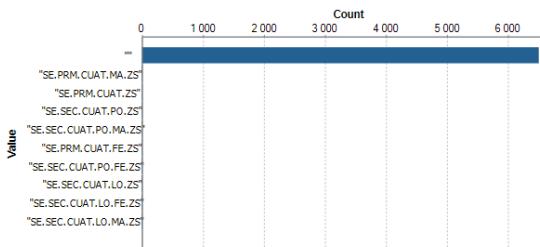
▼ Simple Statistics

Label	Count	%
Row Count	14565	100.00%
Null Count	0	0.00%
Distinct Count	553	7.27%
Unique Count	459	6.03%
Duplicate Count	94	1.24%
Blank Count	0	0.00%



▼ Value Frequency

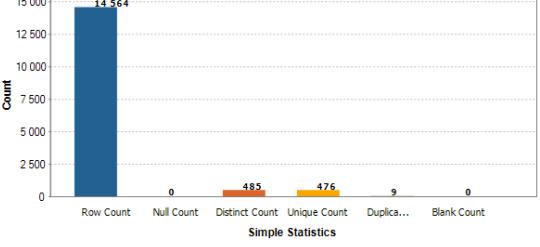
Value	Count	%
""	6505	85.49%
"SE.PRM.CUAT.MA.ZS"	7	0.09%
"SE.PRM.CUAT.ZS"	7	0.09%
"SE.SEC.CUAT.PO.ZS"	7	0.09%
"SE.SEC.CUAT.PO.MA.ZS"	7	0.09%
"SE.PRM.CUAT.FE.ZS"	7	0.09%
"SE.SEC.CUAT.PO.FE.ZS"	7	0.09%
"SE.SEC.CUAT.LO.ZS"	7	0.09%
"SE.SEC.CUAT.LO.FE.ZS"	7	0.09%
"SE.SEC.CUAT.LO.MA.ZS"	7	0.09%



▼ Column: metadata_963  

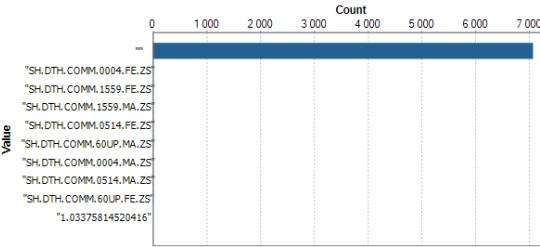
▼ Simple Statistics

Label	Count	%
Row Count	14564	100.00%
Null Count	0	0.00%
Distinct Count	485	6.37%
Unique Count	476	6.26%
Duplicate Count	9	0.12%
Blank Count	0	0.00%



▼ Value Frequency

Value	Count	%
""	7077	93.01%
"SH.DTH.COMM.0004.FE.ZS"	7	0.09%
"SH.DTH.COMM.1559.FE.ZS"	7	0.09%
"SH.DTH.COMM.1559.MA.ZS"	7	0.09%
"SH.DTH.COMM.0514.FE.ZS"	7	0.09%
"SH.DTH.COMM.60UP.MA.ZS"	7	0.09%
"SH.DTH.COMM.0004.MA.ZS"	7	0.09%
"SH.DTH.COMM.0514.MA.ZS"	7	0.09%
"SH.DTH.COMM.60UP.FE.ZS"	7	0.09%
"1.03375814520416"	1	0.01%

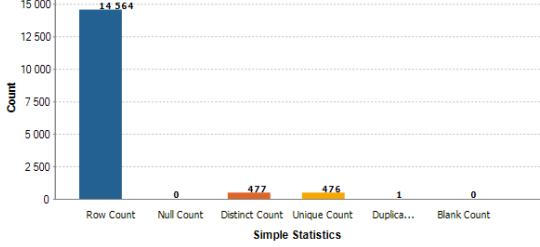


Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata_964  

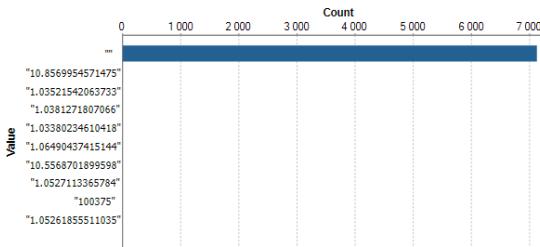
▼ Simple Statistics

Label	Count	%
Row Count	14564	100.00%
Null Count	0	0.00%
Distinct Count	477	6.27%
Unique Count	476	6.26%
Duplicate Count	1	0.01%
Blank Count	0	0.00%



▼ Value Frequency

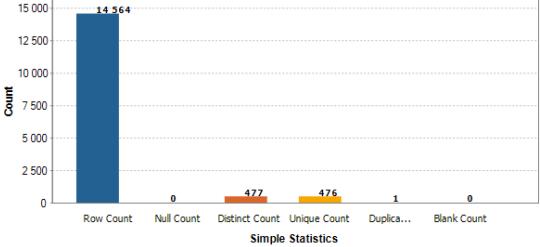
Value	Count	%
""	7133	93.74%
"10.8569954571475"	1	0.01%
"1.03521542063733"	1	0.01%
"1.0381271807066"	1	0.01%
"1.03380234610418"	1	0.01%
"1.06490437415144"	1	0.01%
"10.5568701899598"	1	0.01%
"1.0527113365784"	1	0.01%
"100375"	1	0.01%
"1.05261855511035"	1	0.01%



▼ Column: metadata_965  

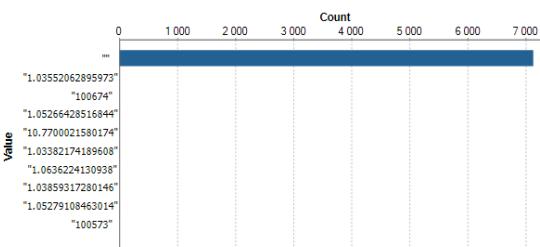
▼ Simple Statistics

Label	Count	%
Row Count	14564	100.00%
Null Count	0	0.00%
Distinct Count	477	6.27%
Unique Count	476	6.26%
Duplicate Count	1	0.01%
Blank Count	0	0.00%



▼ Value Frequency

Value	Count	%
""	7133	93.74%
"1.035520628895973"	1	0.01%
"100674"	1	0.01%
"1.05266428516844"	1	0.01%
"10.770021580174"	1	0.01%
"1.03382174189608"	1	0.01%
"1.0636224130938"	1	0.01%
"1.03859317280146"	1	0.01%
"1.05279108463014"	1	0.01%
"100573"	1	0.01%

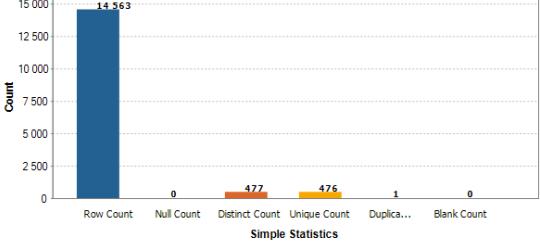


Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata_966  

▼ Simple Statistics

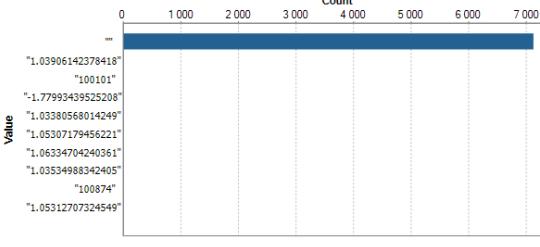
Label	Count	%
Row Count	14563	100.00%
Null Count	0	0.00%
Distinct Count	477	6.27%
Unique Count	476	6.26%
Duplicate Count	1	0.01%
Blank Count	0	0.00%



Simple Statistics

▼ Value Frequency

Value	Count	%
""	7133	93.74%
"1.03906142378418"	1	0.01%
"100101"	1	0.01%
"-1.77993439525208"	1	0.01%
"1.03380568014249"	1	0.01%
"1.05307179456221"	1	0.01%
"1.06334704240361"	1	0.01%
"1.03534988342405"	1	0.01%
"100874"	1	0.01%
"1.05312707324549"	1	0.01%

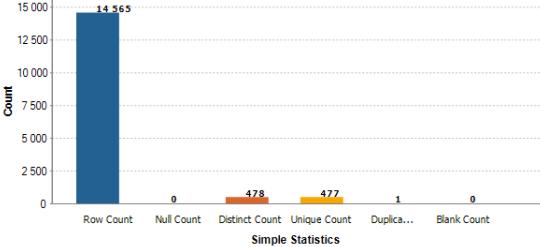


Value Frequency

▼ Column: metadata_967  

▼ Simple Statistics

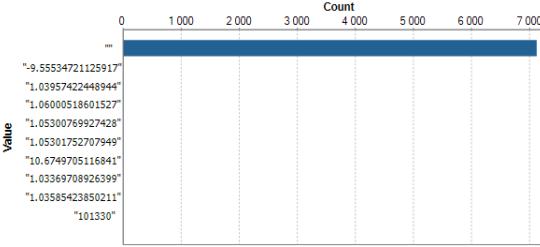
Label	Count	%
Row Count	14565	100.00%
Null Count	0	0.00%
Distinct Count	478	6.28%
Unique Count	477	6.27%
Duplicate Count	1	0.01%
Blank Count	0	0.00%



Simple Statistics

▼ Value Frequency

Value	Count	%
""	7132	93.73%
"-9.55534721125917"	1	0.01%
"1.03957422448944"	1	0.01%
"1.06000518601527"	1	0.01%
"1.05300769927428"	1	0.01%
"1.05301752707949"	1	0.01%
"10.6749705116841"	1	0.01%
"1.03369708926399"	1	0.01%
"1.03585423850211"	1	0.01%
"101330"	1	0.01%



Value Frequency

Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata_968

▼ Simple Statistics

Label	Count	%
Row Count	14565	100.00%
Null Count	0	0.00%
Distinct Count	483	6.35%
Unique Count	482	6.33%
Duplicate Count	1	0.01%
Blank Count	0	0.00%

▼ Value Frequency

Value	Count	%
""	7127	93.67%
"1.05366728758403"	1	0.01%
"1.97774154682664"	1	0.01%
"100.589078637086"	1	0.01%
"1.45710930216502"	1	0.01%
"1.05904918003615"	1	0.01%
"1.03359365902065"	1	0.01%
"1.04007166367348"	1	0.01%
"1.03633040198755"	1	0.01%
"1.05287675072532"	1	0.01%

▼ Column: metadata_969

▼ Simple Statistics

Label	Count	%
Row Count	14563	100.00%
Null Count	0	0.00%
Distinct Count	483	6.35%
Unique Count	482	6.33%
Duplicate Count	1	0.01%
Blank Count	0	0.00%

▼ Value Frequency

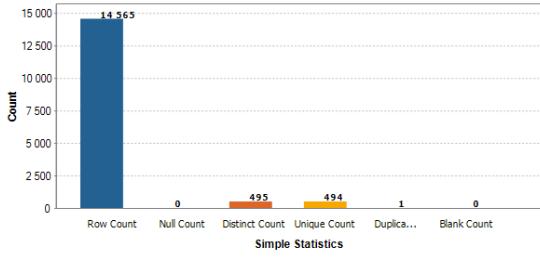
Value	Count	%
""	7127	93.67%
"101.019715861455"	1	0.01%
"1.03634393912645"	1	0.01%
"1.03363142270451"	1	0.01%
"1.0539591120929"	1	0.01%
"1.05317655345488"	1	0.01%
"1.04048272297555"	1	0.01%
"101743208"	1	0.01%
"10169182"	1	0.01%
"1.06102934743621"	1	0.01%

Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata_970  

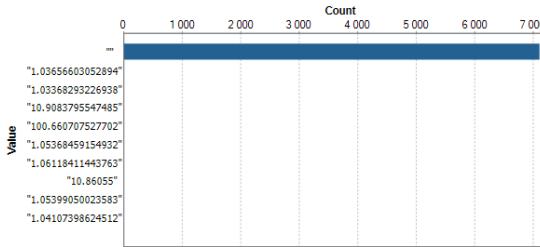
▼ Simple Statistics

Label	Count	%
Row Count	14565	100.00%
Null Count	0	0.00%
Distinct Count	495	6.51%
Unique Count	494	6.49%
Duplicate Count	1	0.01%
Blank Count	0	0.00%



▼ Value Frequency

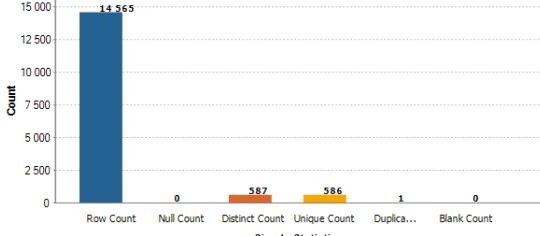
Value	Count	%
""	7115	93.51%
"1.03656603052894"	1	0.01%
"1.03368293226938"	1	0.01%
"10.9083795547485"	1	0.01%
"100.660707527702"	1	0.01%
"1.05368459154932"	1	0.01%
"1.06118411443763"	1	0.01%
"10.86055"	1	0.01%
"1.05399050023583"	1	0.01%
"1.04107398624512"	1	0.01%



▼ Column: metadata_971  

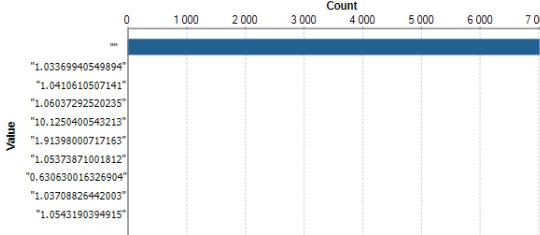
▼ Simple Statistics

Label	Count	%
Row Count	14565	100.00%
Null Count	0	0.00%
Distinct Count	587	7.71%
Unique Count	586	7.70%
Duplicate Count	1	0.01%
Blank Count	0	0.00%



▼ Value Frequency

Value	Count	%
""	7023	92.30%
"1.0369940549894"	1	0.01%
"1.0410610507141"	1	0.01%
"1.06037292520235"	1	0.01%
"10.1250400543213"	1	0.01%
"1.91398000717163"	1	0.01%
"1.05373871001812"	1	0.01%
"0.630630016326904"	1	0.01%
"1.03708826442003"	1	0.01%
"1.0543190394915"	1	0.01%

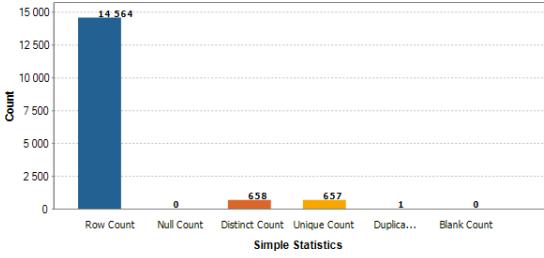


Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata_972  

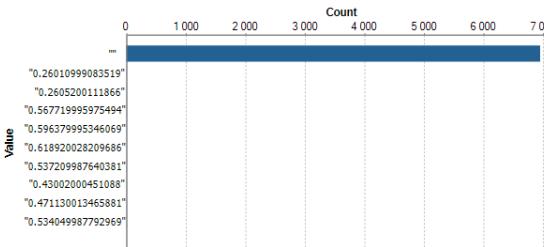
▼ Simple Statistics

Label	Count	%
Row Count	14564	100.00%
Null Count	0	0.00%
Distinct Count	658	8.65%
Unique Count	657	8.63%
Duplicate Count	1	0.01%
Blank Count	0	0.00%



▼ Value Frequency

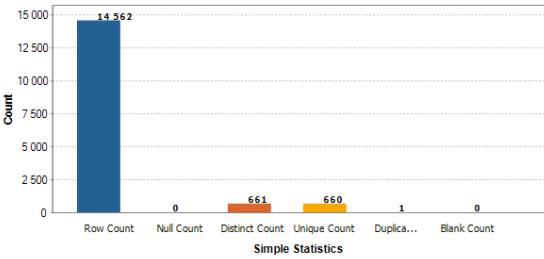
Value	Count	%
""	6952	91.37%
"0.26010999083519"	1	0.01%
"0.2605200111866"	1	0.01%
"0.56771995975494"	1	0.01%
"0.596379995346069"	1	0.01%
"0.618920028209686"	1	0.01%
"0.537209987640381"	1	0.01%
"0.43002000451088"	1	0.01%
"0.471130013465881"	1	0.01%
"0.534049987792969"	1	0.01%



▼ Column: metadata_973  

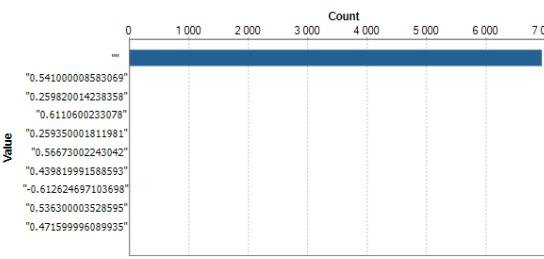
▼ Simple Statistics

Label	Count	%
Row Count	14562	100.00%
Null Count	0	0.00%
Distinct Count	661	8.69%
Unique Count	660	8.67%
Duplicate Count	1	0.01%
Blank Count	0	0.00%



▼ Value Frequency

Value	Count	%
""	6949	91.33%
"0.541000008583069"	1	0.01%
"0.259820014238358"	1	0.01%
"0.6110600233078"	1	0.01%
"0.259350001811981"	1	0.01%
"0.56673002243042"	1	0.01%
"0.439819991588593"	1	0.01%
"-0.612624697103698"	1	0.01%
"0.5363000003528595"	1	0.01%
"0.471599996089935"	1	0.01%



Engenharia de Dados para Suporte à Tomada de Decisão



Engenharia de Dados para Suporte à Tomada de Decisão



Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata_978

▼ Simple Statistics

Label	Count	%
Row Count	14564	100.00%
Null Count	0	0.00%
Distinct Count	737	9.69%
Unique Count	736	9.67%
Duplicate Count	1	0.01%
Blank Count	0	0.00%

Count

Row Count: 14564
Null Count: 0
Distinct Count: 737
Unique Count: 736
Duplicate Count: 1
Blank Count: 0

Simple Statistics

▼ Value Frequency

Value	Count	%
""	6873	90.33%
"0.485339999198914"	1	0.01%
"0.600570023059845"	1	0.01%
"-0.593563905303014"	1	0.01%
"0.383399993181229"	1	0.01%
"0.485570013523102"	1	0.01%
"0.564999997615814"	1	0.01%
"-2.19352451851663"	1	0.01%
"0.58009999904633"	1	0.01%
"0.254729986190796"	1	0.01%

Value

Count

0.485339999198914
0.600570023059845
-0.593563905303014
0.383399993181229
0.485570013523102
0.564999997615814
-2.19352451851663
0.58009999904633
0.254729986190796

▼ Column: metadata_979

▼ Simple Statistics

Label	Count	%
Row Count	14564	100.00%
Null Count	0	0.00%
Distinct Count	741	9.74%
Unique Count	740	9.73%
Duplicate Count	1	0.01%
Blank Count	0	0.00%

Count

Row Count: 14564
Null Count: 0
Distinct Count: 741
Unique Count: 740
Duplicate Count: 1
Blank Count: 0

Simple Statistics

▼ Value Frequency

Value	Count	%
""	6869	90.27%
"0.268040001392365"	1	0.01%
"0.66237998008728"	1	0.01%
"0.50282014953613"	1	0.01%
"0.468760013580322"	1	0.01%
"0.670130014419556"	1	0.01%
"0.599049985408783"	1	0.01%
"0.541960000991821"	1	0.01%
"0.481610000133514"	1	0.01%
"0.595459997653961"	1	0.01%

Value

Count

0.268040001392365
0.66237998008728
0.50282014953613
0.468760013580322
0.670130014419556
0.599049985408783
0.541960000991821
0.481610000133514
0.595459997653961

Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata_980

▼ Simple Statistics

Label	Count	%
Row Count	14565	100.00%
Null Count	0	0.00%
Distinct Count	739	9.71%
Unique Count	738	9.70%
Duplicate Count	1	0.01%
Blank Count	0	0.00%

▼ Value Frequency

Value	Count	%
""	6871	90.30%
"0.607529997825623"	1	0.01%
"0.53914999961853"	1	0.01%
"0.518850028514862"	1	0.01%
"0.68001002073288"	1	0.01%
"0.536759972572327"	1	0.01%
"0.27224001288414"	1	0.01%
"0.603320002555847"	1	0.01%
"0.504010021686554"	1	0.01%
"0.671169996261597"	1	0.01%

▼ Column: metadata_981

▼ Simple Statistics

Label	Count	%
Row Count	14562	100.00%
Null Count	0	0.00%
Distinct Count	789	10.37%
Unique Count	788	10.36%
Duplicate Count	1	0.01%
Blank Count	0	0.00%

▼ Value Frequency

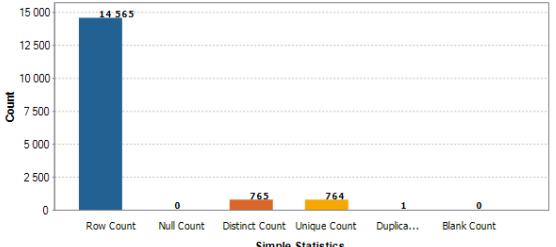
Value	Count	%
""	6821	89.64%
"0.539929986000061"	1	0.01%
"0.687810003757477"	1	0.01%
"0.558430016040802"	1	0.01%
"0.61422997713089"	1	0.01%
"-6.9741763030321"	1	0.01%
"0.281129986047745"	1	0.01%
"0.499190002679825"	1	0.01%
"0.614970028400421"	1	0.01%
"0.671169996261597"	1	0.01%

Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata_982  

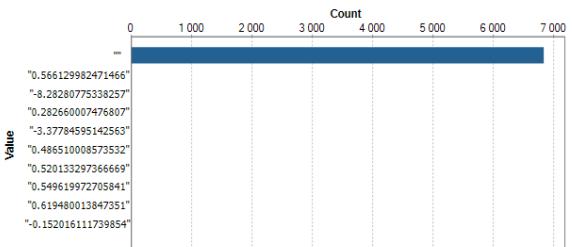
▼ Simple Statistics

Label	Count	%
Row Count	14565	100.00%
Null Count	0	0.00%
Distinct Count	765	10.05%
Unique Count	764	10.04%
Duplicate Count	1	0.01%
Blank Count	0	0.00%



Value Frequency

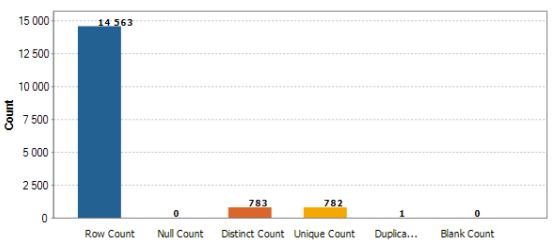
Value	Count	%
""	6845	89.95%
"0.566129982471466"	1	0.01%
"-8.28280775338257"	1	0.01%
"0.282660007476807"	1	0.01%
"-3.37784595142563"	1	0.01%
"0.486510008573532"	1	0.01%
"0.52013329736669"	1	0.01%
"0.549619972705841"	1	0.01%
"0.619480013847351"	1	0.01%
"-0.152016111739854"	1	0.01%



▼ Column: metadata_983  

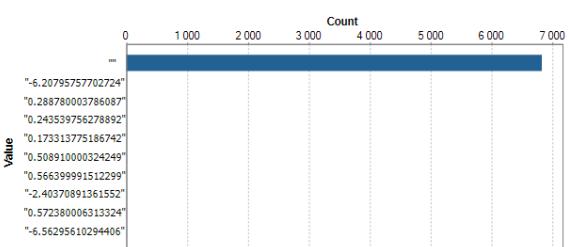
▼ Simple Statistics

Label	Count	%
Row Count	14563	100.00%
Null Count	0	0.00%
Distinct Count	783	10.29%
Unique Count	782	10.29%
Duplicate Count	1	0.01%
Blank Count	0	0.00%



Value Frequency

Value	Count	%
""	6827	89.72%
"-6.20795757702724"	1	0.01%
"0.288780003786087"	1	0.01%
"0.243539756278892"	1	0.01%
"0.173313775186742"	1	0.01%
"0.508910000324249"	1	0.01%
"0.566399991512299"	1	0.01%
"-2.40370891361552"	1	0.01%
"0.572380006313324"	1	0.01%
"-6.56295610294406"	1	0.01%



Engenharia de Dados para Suporte à Tomada de Decisão

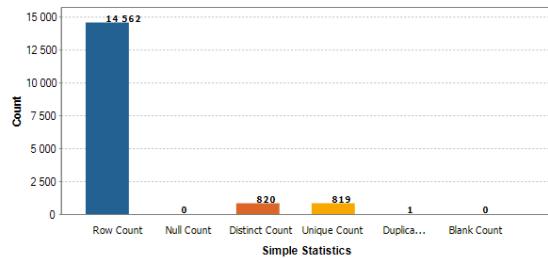


Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata_986  

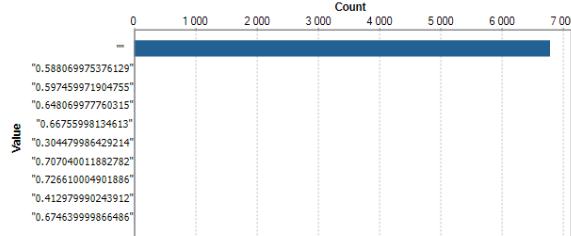
▼ Simple Statistics

Label	Count	%
Row Count	14562	100.00%
Null Count	0	0.00%
Distinct Count	820	10.78%
Unique Count	819	10.76%
Duplicate Count	1	0.01%
Blank Count	0	0.00%



▼ Value Frequency

Value	Count	%
""	6790	89.24%
"0.588069975376129"	1	0.01%
"0.597459971904755"	1	0.01%
"0.648069977760315"	1	0.01%
"0.66755998134613"	1	0.01%
"0.304479986429214"	1	0.01%
"0.707040011882782"	1	0.01%
"0.726610004901886"	1	0.01%
"0.412979990243912"	1	0.01%
"0.674639999866486"	1	0.01%

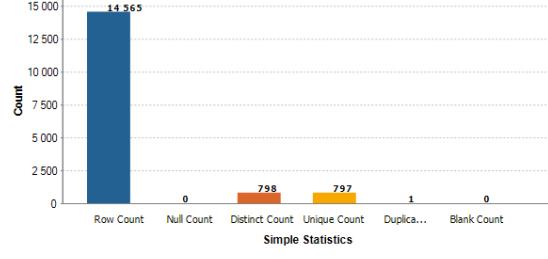


▼ Column: metadata_987  

▼ Column: metadata_987  

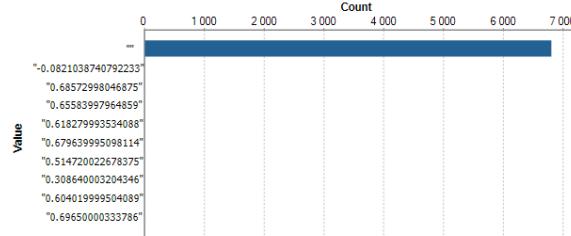
▼ Simple Statistics

Label	Count	%
Row Count	14565	100.00%
Null Count	0	0.00%
Distinct Count	798	10.49%
Unique Count	797	10.47%
Duplicate Count	1	0.01%
Blank Count	0	0.00%



▼ Value Frequency

Value	Count	%
""	6812	89.53%
"-0.0821038740792233"	1	0.01%
"0.68572998046875"	1	0.01%
"0.65583997964859"	1	0.01%
"0.618279993534088"	1	0.01%
"0.679639995098114"	1	0.01%
"0.514720022678375"	1	0.01%
"0.30864003204346"	1	0.01%
"0.60401999504089"	1	0.01%
"0.69650000333786"	1	0.01%

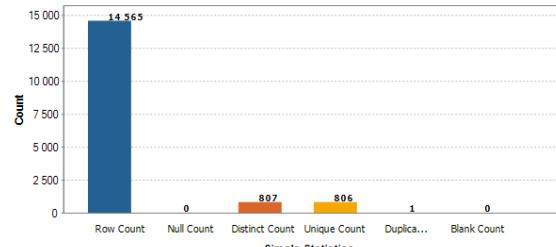


Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata_988  

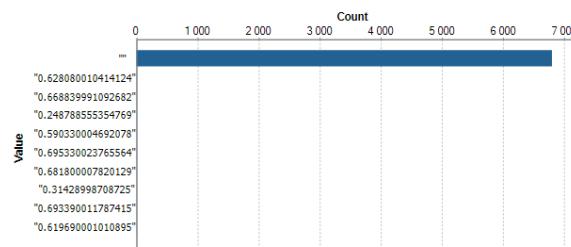
▼ Simple Statistics

Label	Count	%
Row Count	14565	100.00%
Null Count	0	0.00%
Distinct Count	807	10.61%
Unique Count	806	10.59%
Duplicate Count	1	0.01%
Blank Count	0	0.00%



▼ Value Frequency

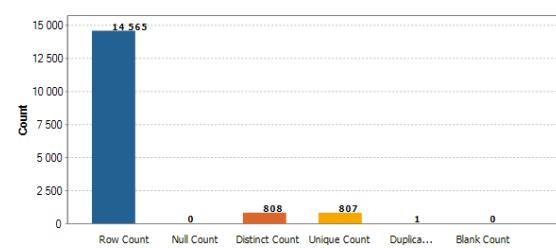
Value	Count	%
""	6803	89.41%
"0.628080010414124"	1	0.01%
"0.66883991092682"	1	0.01%
"0.248788555354769"	1	0.01%
"0.590330004692078"	1	0.01%
"0.69533023765564"	1	0.01%
"0.681800007820129"	1	0.01%
"0.3142898708725"	1	0.01%
"0.693390011787415"	1	0.01%
"0.619690001010895"	1	0.01%



▼ Column: metadata_989  

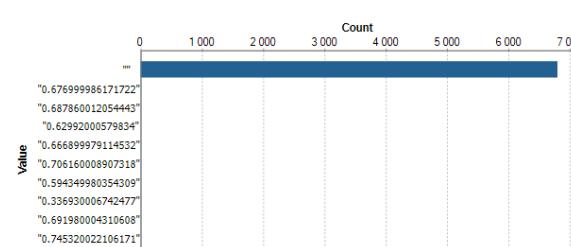
▼ Simple Statistics

Label	Count	%
Row Count	14565	100.00%
Null Count	0	0.00%
Distinct Count	808	10.62%
Unique Count	807	10.61%
Duplicate Count	1	0.01%
Blank Count	0	0.00%



▼ Value Frequency

Value	Count	%
""	6802	89.39%
"0.676999986171722"	1	0.01%
"0.687860012054443"	1	0.01%
"0.62992000579834"	1	0.01%
"0.66689979114532"	1	0.01%
"0.706160008907318"	1	0.01%
"0.594349980354309"	1	0.01%
"0.336930006742477"	1	0.01%
"0.691980004310608"	1	0.01%
"0.745320022106171"	1	0.01%



Engenharia de Dados para Suporte à Tomada de Decisão



Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata_992

▼ Simple Statistics

Label	Count	%
Row Count	14562	100.00%
Null Count	0	0.00%
Distinct Count	1189	15.63%
Unique Count	1187	15.60%
Duplicate Count	2	0.03%
Blank Count	0	0.00%

▼ Value Frequency

Value	Count	%
""	6420	84.37%
"0.1"	2	0.03%
"0.689670026302338"	1	0.01%
"0.65099996089935"	1	0.01%
"0.502101175717759"	1	0.01%
"0.68516740626674"	1	0.01%
"0.656419992446899"	1	0.01%
"0.323619991540909"	1	0.01%
"-1.9599506232481"	1	0.01%
"-1.979622643578"	1	0.01%

▼ Column: metadata_993

▼ Simple Statistics

Label	Count	%
Row Count	14562	100.00%
Null Count	0	0.00%
Distinct Count	1263	16.60%
Unique Count	1262	16.59%
Duplicate Count	1	0.01%
Blank Count	0	0.00%

▼ Value Frequency

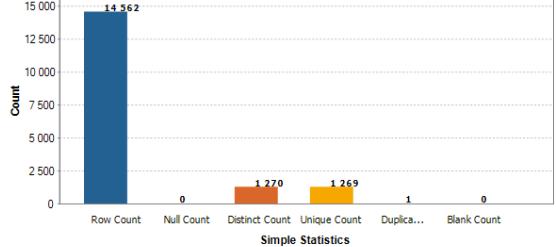
Value	Count	%
""	6347	83.41%
"-1.20176347950796"	1	0.01%
"0.350800007581711"	1	0.01%
"0.100339250157642"	1	0.01%
"0.100323564699136"	1	0.01%
"-0.0629990064786967"	1	0.01%
"-0.365018757198371"	1	0.01%
"0.668579995632172"	1	0.01%
"0.669049978256226"	1	0.01%
"0.526348323496493"	1	0.01%

Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata_994  

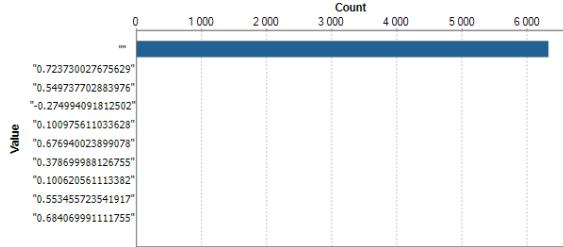
▼ Simple Statistics

Label	Count	%
Row Count	14562	100.00%
Null Count	0	0.00%
Distinct Count	1270	16.69%
Unique Count	1269	16.68%
Duplicate Count	1	0.01%
Blank Count	0	0.00%



▼ Value Frequency

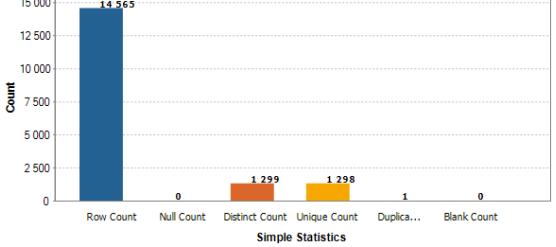
Value	Count	%
""	6340	83.32%
"0.723730027675629"	1	0.01%
"0.549737702883976"	1	0.01%
"-0.274994091812502"	1	0.01%
"0.100975611033628"	1	0.01%
"0.676940023899078"	1	0.01%
"0.378699988126755"	1	0.01%
"0.100620561113382"	1	0.01%
"0.553455723541917"	1	0.01%
"0.684069991111755"	1	0.01%



▼ Column: metadata_995  

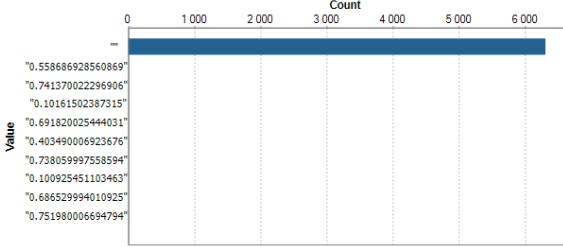
▼ Simple Statistics

Label	Count	%
Row Count	14565	100.00%
Null Count	0	0.00%
Distinct Count	1299	17.07%
Unique Count	1298	17.06%
Duplicate Count	1	0.01%
Blank Count	0	0.00%



▼ Value Frequency

Value	Count	%
""	6311	82.94%
"0.558686928560869"	1	0.01%
"0.74137002296906"	1	0.01%
"0.10161502387315"	1	0.01%
"0.691820025444031"	1	0.01%
"0.403490006923676"	1	0.01%
"0.738059997558594"	1	0.01%
"0.100923451103463"	1	0.01%
"0.686529994010925"	1	0.01%
"0.751980006694794"	1	0.01%

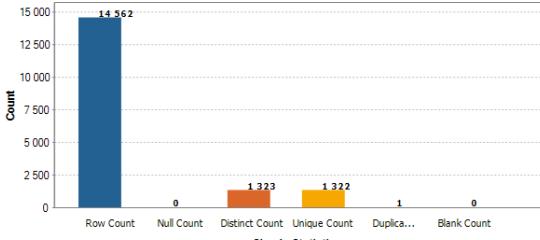


Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata_996  

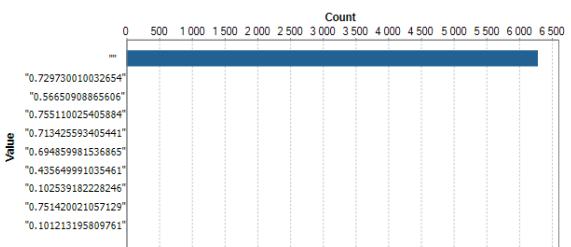
▼ Simple Statistics

Label	Count	%
Row Count	14562	100.00%
Null Count	0	0.00%
Distinct Count	1323	17.39%
Unique Count	1322	17.37%
Duplicate Count	1	0.01%
Blank Count	0	0.00%



▼ Value Frequency

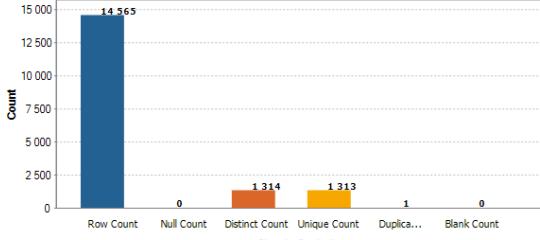
Value	Count	%
""	6287	82.63%
"0.729730010032654"	1	0.01%
"0.5665090865606"	1	0.01%
"0.755110025405884"	1	0.01%
"0.713425593405441"	1	0.01%
"0.694859981536865"	1	0.01%
"0.435649991035461"	1	0.01%
"0.102539182228246"	1	0.01%
"0.751420021057129"	1	0.01%
"0.101213195809761"	1	0.01%



▼ Column: metadata_997  

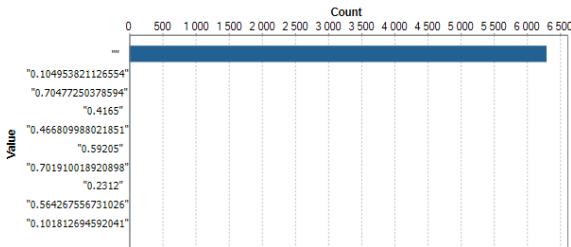
▼ Simple Statistics

Label	Count	%
Row Count	14565	100.00%
Null Count	0	0.00%
Distinct Count	1314	17.27%
Unique Count	1313	17.26%
Duplicate Count	1	0.01%
Blank Count	0	0.00%



▼ Value Frequency

Value	Count	%
""	6296	82.74%
"0.104953821126554"	1	0.01%
"0.70477250378594"	1	0.01%
"0.4165"	1	0.01%
"0.466809988021851"	1	0.01%
"0.59205"	1	0.01%
"0.701910018920898"	1	0.01%
"0.2312"	1	0.01%
"0.564267556731026"	1	0.01%
"0.101812694592041"	1	0.01%

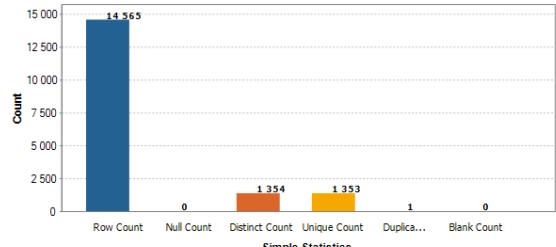


Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata_998  

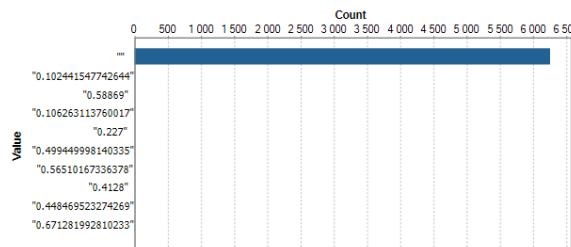
▼ Simple Statistics

Label	Count	%
Row Count	14565	100.00%
Null Count	0	0.00%
Distinct Count	1354	17.79%
Unique Count	1353	17.78%
Duplicate Count	1	0.01%
Blank Count	0	0.00%



▼ Value Frequency

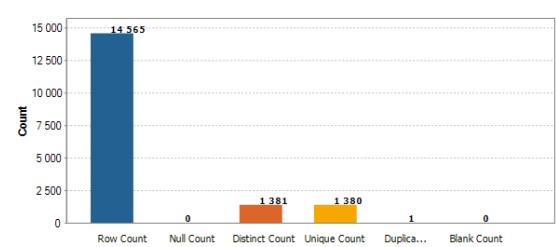
Value	Count	%
""	6256	82.22%
"0.102441547742644"	1	0.01%
"0.58869"	1	0.01%
"0.106263113760017"	1	0.01%
"0.227"	1	0.01%
"0.499449998140335"	1	0.01%
"0.56510167336378"	1	0.01%
"0.4128"	1	0.01%
"0.448469523274269"	1	0.01%
"0.671281992810233"	1	0.01%



▼ Column: metadata_999  

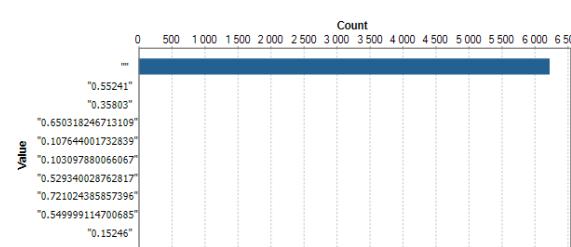
▼ Simple Statistics

Label	Count	%
Row Count	14565	100.00%
Null Count	0	0.00%
Distinct Count	1381	18.15%
Unique Count	1380	18.14%
Duplicate Count	1	0.01%
Blank Count	0	0.00%



▼ Value Frequency

Value	Count	%
""	6229	81.86%
"0.55241"	1	0.01%
"0.35803"	1	0.01%
"0.650318246713109"	1	0.01%
"0.107644001732839"	1	0.01%
"0.103097880066067"	1	0.01%
"0.529340028762817"	1	0.01%
"0.721024385857396"	1	0.01%
"0.549999114700685"	1	0.01%
"0.15246"	1	0.01%



Engenharia de Dados para Suporte à Tomada de Decisão

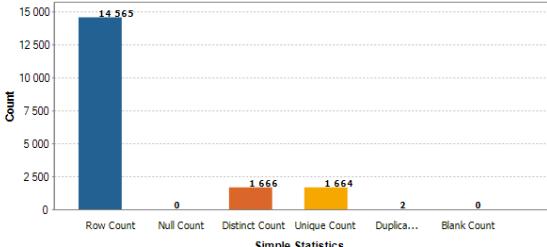


Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata_002

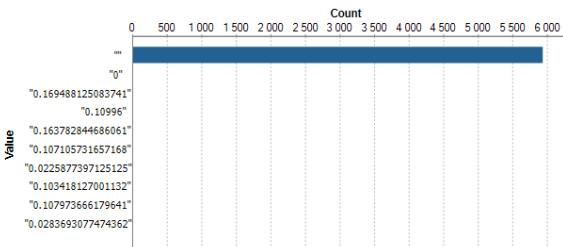
▼ Simple Statistics

Label	Count	%
Row Count	14565	100.00%
Null Count	0	0.00%
Distinct Count	1666	21.90%
Unique Count	1664	21.87%
Duplicate Count	2	0.03%
Blank Count	0	0.00%



▼ Value Frequency

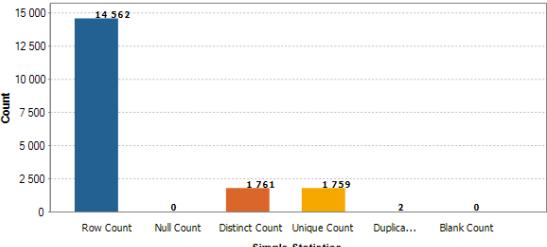
Value	Count	%
""	5940	78.07%
"0"	5	0.07%
"0.169488125083741"	1	0.01%
"0.10996"	1	0.01%
"0.163782844686061"	1	0.01%
"0.107105731657168"	1	0.01%
"0.0225877397125125"	1	0.01%
"0.103418127001132"	1	0.01%
"0.10797366179641"	1	0.01%
"0.0283693077474362"	1	0.01%



▼ Column: metadata_003

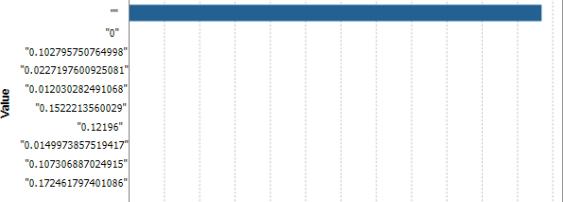
▼ Simple Statistics

Label	Count	%
Row Count	14562	100.00%
Null Count	0	0.00%
Distinct Count	1761	23.14%
Unique Count	1759	23.12%
Duplicate Count	2	0.03%
Blank Count	0	0.00%



▼ Value Frequency

Value	Count	%
""	5847	76.84%
"0"	3	0.04%
"0.102795750764998"	1	0.01%
"0.0227197600925081"	1	0.01%
"0.012030282491068"	1	0.01%
"0.1522213560029"	1	0.01%
"0.12196"	1	0.01%
"0.0149973857519417"	1	0.01%
"0.107306887024915"	1	0.01%
"0.172461797401086"	1	0.01%



Engenharia de Dados para Suporte à Tomada de Decisão



Engenharia de Dados para Suporte à Tomada de Decisão



Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata_008

▼ Simple Statistics

Label	Count	%
Row Count	14565	100.00%
Null Count	0	0.00%
Distinct Count	1648	21.66%
Unique Count	1636	21.50%
Duplicate Count	12	0.16%
Blank Count	0	0.00%

▼ Value Frequency

Value	Count	%
""	5949	78.18%
"0"	3	0.04%
"7,75"	3	0.04%
"140"	2	0.03%
"69.6611111111111"	2	0.03%
"33.984"	2	0.03%
"35,1"	2	0.03%
"79.5416666666667"	2	0.03%
"28.2272727272727"	2	0.03%
"60.815"	2	0.03%

▼ Column: metadata_009

▼ Simple Statistics

Label	Count	%
Row Count	14565	100.00%
Null Count	0	0.00%
Distinct Count	1636	21.50%
Unique Count	1621	21.30%
Duplicate Count	15	0.20%
Blank Count	0	0.00%

▼ Value Frequency

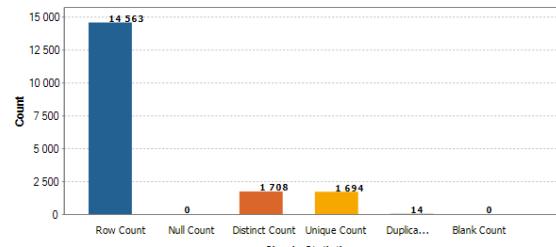
Value	Count	%
""	5959	78.32%
"0"	3	0.04%
"10.1666666666667"	2	0.03%
"31.064"	2	0.03%
"110"	2	0.03%
"133.420833333333"	2	0.03%
"24,5"	2	0.03%
"5.66363636363636"	2	0.03%
"150"	2	0.03%
"30.6833333333333"	2	0.03%

Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata_010  

▼ Simple Statistics

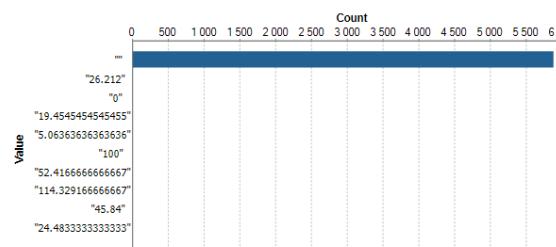
Label	Count	%
Row Count	14563	100.00%
Null Count	0	0.00%
Distinct Count	1708	22.45%
Unique Count	1694	22.26%
Duplicate Count	14	0.18%
Blank Count	0	0.00%



Simple Statistics

▼ Value Frequency

Value	Count	%
""	5887	77.37%
"26.212"	3	0.04%
"0"	3	0.04%
"19.454545454545"	2	0.03%
"5.06363636363636"	2	0.03%
"100"	2	0.03%
"52.4166666666667"	2	0.03%
"114.329166666667"	2	0.03%
"45.84"	2	0.03%
"24.48333333333333"	2	0.03%



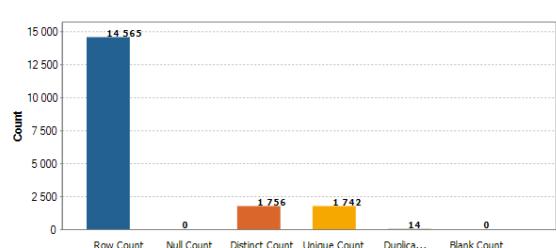
Value

Count

▼ Column: metadata_011  

▼ Simple Statistics

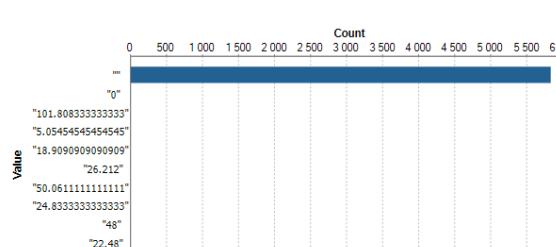
Label	Count	%
Row Count	14565	100.00%
Null Count	0	0.00%
Distinct Count	1756	23.08%
Unique Count	1742	22.89%
Duplicate Count	14	0.18%
Blank Count	0	0.00%



Simple Statistics

▼ Value Frequency

Value	Count	%
""	5840	76.75%
"0"	3	0.04%
"101.808333333333"	2	0.03%
"5.05454545454545"	2	0.03%
"18.909090909090909"	2	0.03%
"26.212"	2	0.03%
"50.06111111111111"	2	0.03%
"24.83333333333333"	2	0.03%
"48"	2	0.03%
"22.48"	2	0.03%



Value

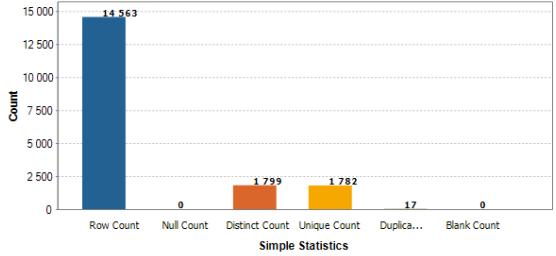
Count

Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata_012  

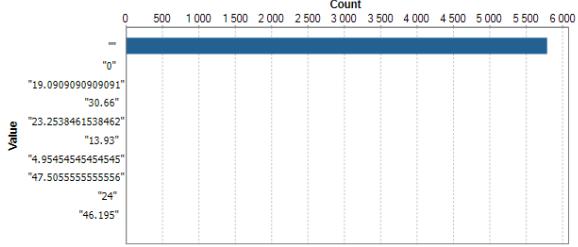
▼ Simple Statistics

Label	Count	%
Row Count	14563	100.00%
Null Count	0	0.00%
Distinct Count	1799	23.64%
Unique Count	1782	23.42%
Duplicate Count	17	0.22%
Blank Count	0	0.00%



▼ Value Frequency

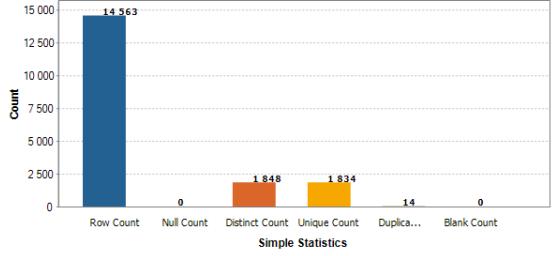
Value	Count	%
""	5794	76.15%
"0"	3	0.04%
"19.0909090909091"	2	0.03%
"30.66"	2	0.03%
"23.2538461538462"	2	0.03%
"13.93"	2	0.03%
"4.95454545454545"	2	0.03%
"47.5055555555556"	2	0.03%
"24"	2	0.03%
"46.195"	2	0.03%



▼ Column: metadata_013  

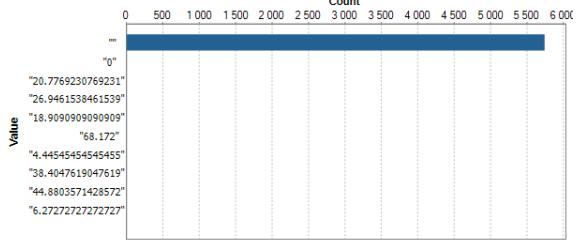
▼ Simple Statistics

Label	Count	%
Row Count	14563	100.00%
Null Count	0	0.00%
Distinct Count	1848	24.29%
Unique Count	1834	24.10%
Duplicate Count	14	0.18%
Blank Count	0	0.00%



▼ Value Frequency

Value	Count	%
""	5748	75.54%
"0"	3	0.04%
"20.7769230769231"	2	0.03%
"26.9461538461539"	2	0.03%
"18.9090909090909"	2	0.03%
"68.172"	2	0.03%
"4.44545454545455"	2	0.03%
"38.4047619047619"	2	0.03%
"44.8803571428572"	2	0.03%
"6.27272727272727"	2	0.03%



Engenharia de Dados para Suporte à Tomada de Decisão



Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata_016

▼ Simple Statistics

Label	Count	%
Row Count	14565	100.00%
Null Count	0	0.00%
Distinct Count	1714	22.53%
Unique Count	1701	22.36%
Duplicate Count	13	0.17%
Blank Count	0	0.00%

▼ Value Frequency

Value	Count	%
""	5882	77.30%
"0"	3	0.04%
"5.54545454545455"	3	0.04%
"1.3"	2	0.03%
"20.3689655172414"	2	0.03%
"37.316393442623"	2	0.03%
"19.8384615384615"	2	0.03%
"22.7307692307692"	2	0.03%
"16.0909090909091"	2	0.03%
"3.5"	2	0.03%

▼ Column: metadata_017

▼ Simple Statistics

Label	Count	%
Row Count	14565	100.00%
Null Count	0	0.00%
Distinct Count	1785	23.46%
Unique Count	1773	23.30%
Duplicate Count	12	0.16%
Blank Count	0	0.00%

▼ Value Frequency

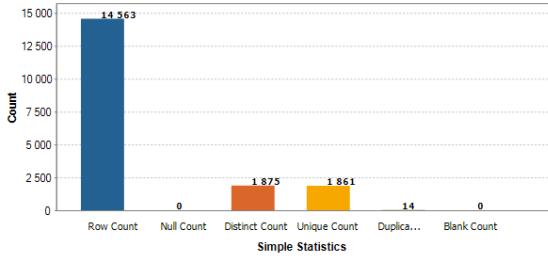
Value	Count	%
""	5811	76.37%
"0"	3	0.04%
"5.54545454545455"	3	0.04%
"15.7272727272727"	3	0.04%
"17.7448275862069"	2	0.03%
"21.0384615384615"	2	0.03%
"38.2818181818182"	2	0.03%
"19.5153846153846"	2	0.03%
"3.65454545454545"	2	0.03%
"34.9852459016393"	2	0.03%

Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata_018  

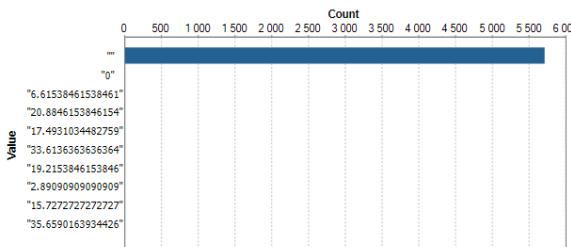
▼ Simple Statistics

Label	Count	%
Row Count	14563	100.00%
Null Count	0	0.00%
Distinct Count	1875	24.64%
Unique Count	1861	24.46%
Duplicate Count	14	0.18%
Blank Count	0	0.00%



▼ Value Frequency

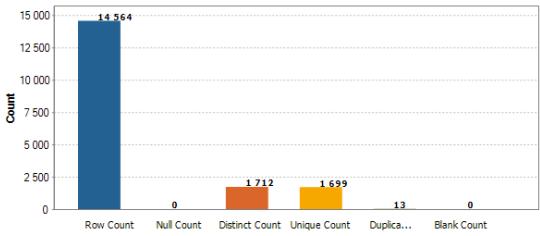
Value	Count	%
""	5719	75.16%
"0"	3	0.04%
"6.61538461538461"	3	0.04%
"20.8846153846154"	3	0.04%
"17.4931034482759"	2	0.03%
"33.6136363636364"	2	0.03%
"19.2153846153846"	2	0.03%
"2.89090909090909"	2	0.03%
"15.7272727272727"	2	0.03%
"35.6590163934426"	2	0.03%



▼ Column: metadata_019  

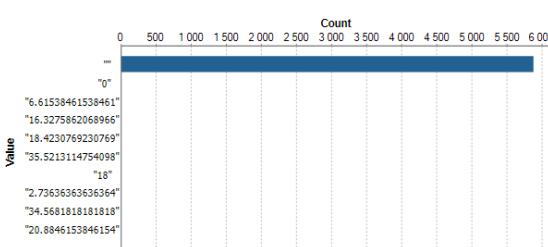
▼ Simple Statistics

Label	Count	%
Row Count	14564	100.00%
Null Count	0	0.00%
Distinct Count	1712	22.50%
Unique Count	1699	22.33%
Duplicate Count	13	0.17%
Blank Count	0	0.00%



▼ Value Frequency

Value	Count	%
""	5884	77.33%
"0"	3	0.04%
"6.61538461538461"	3	0.04%
"16.3275862068966"	2	0.03%
"18.4230769230769"	2	0.03%
"35.5213114754098"	2	0.03%
"18"	2	0.03%
"2.73636363636364"	2	0.03%
"34.5681818181818"	2	0.03%
"20.8846153846154"	2	0.03%

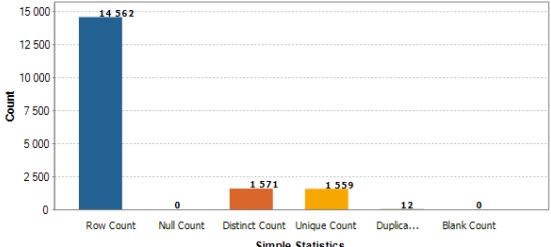


Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata_020   

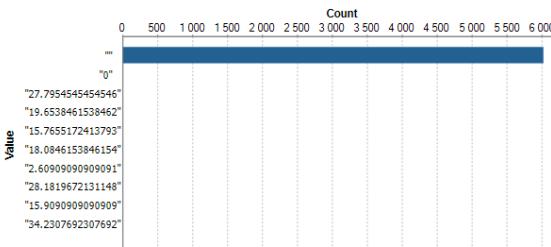
▼ Simple Statistics

Label	Count	%
Row Count	14562	100.00%
Null Count	0	0.00%
Distinct Count	1571	20.65%
Unique Count	1559	20.49%
Duplicate Count	12	0.16%
Blank Count	0	0.00%



▼ Value Frequency

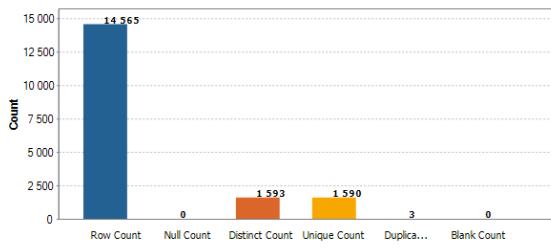
Value	Count	%
""	6027	79.21%
"0"	3	0.04%
"27.7954545454546"	2	0.03%
"19.6538461538462"	2	0.03%
"15.7655172413793"	2	0.03%
"18.0846153846154"	2	0.03%
"2.60909090909091"	2	0.03%
"28.1819672131148"	2	0.03%
"15.9090909090909"	2	0.03%
"34.2307692307692"	2	0.03%



▼ Column: metadata_021   

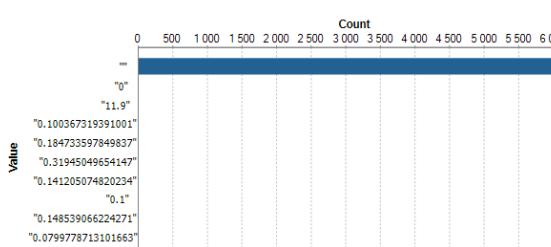
▼ Simple Statistics

Label	Count	%
Row Count	14565	100.00%
Null Count	0	0.00%
Distinct Count	1593	20.94%
Unique Count	1590	20.90%
Duplicate Count	3	0.04%
Blank Count	0	0.00%



▼ Value Frequency

Value	Count	%
""	6015	79.05%
"0"	2	0.03%
"11.9"	2	0.03%
"0.100367319391001"	1	0.01%
"0.184733597849837"	1	0.01%
"0.31945049654147"	1	0.01%
"0.141205074820234"	1	0.01%
"0.1"	1	0.01%
"0.148539066224271"	1	0.01%
"0.0799778713101663"	1	0.01%



Engenharia de Dados para Suporte à Tomada de Decisão

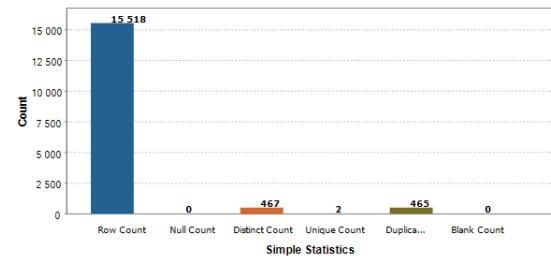


Education_NY

▼ Column: metadata.DBN  

▼ Simple Statistics

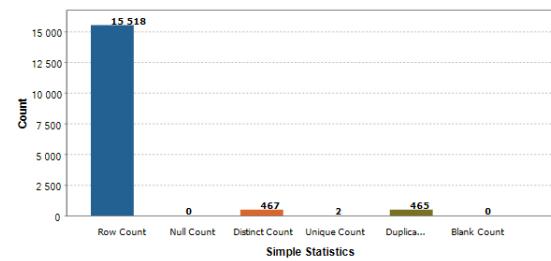
Label	Count	%
Row Count	15518	100.00%
Null Count	0	0.00%
Distinct Count	467	3.01%
Unique Count	2	0.01%
Duplicate Count	465	3.00%
Blank Count	0	0.00%



▼ Column: metadata.School_Name  

▼ Simple Statistics

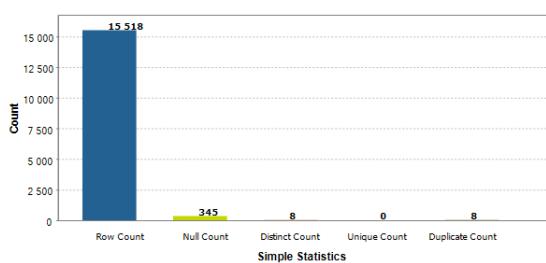
Label	Count	%
Row Count	15518	100.00%
Null Count	0	0.00%
Distinct Count	467	3.01%
Unique Count	2	0.01%
Duplicate Count	465	3.00%
Blank Count	0	0.00%



▼ Column: metadata.Cohort_Year  

▼ Simple Statistics

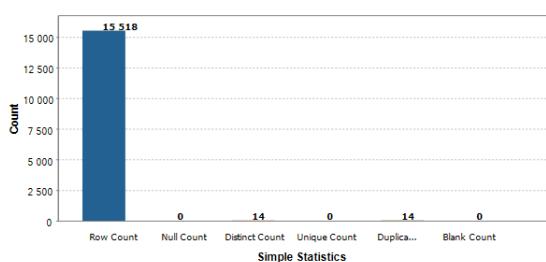
Label	Count	%
Row Count	15518	100.00%
Null Count	345	2.22%
Distinct Count	8	0.05%
Unique Count	0	0.00%
Duplicate Count	8	0.05%



▼ Column: metadata.Cohort_Category  

▼ Simple Statistics

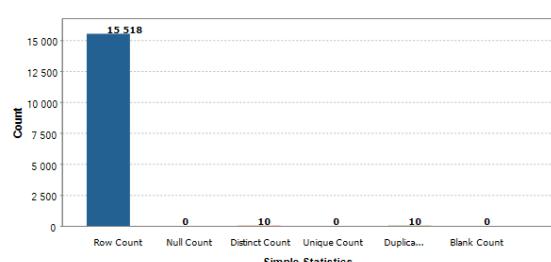
Label	Count	%
Row Count	15518	100.00%
Null Count	0	0.00%
Distinct Count	14	0.09%
Unique Count	0	0.00%
Duplicate Count	14	0.09%
Blank Count	0	0.00%



▼ Column: metadata.Demographic  

▼ Simple Statistics

Label	Count	%
Row Count	15518	100.00%
Null Count	0	0.00%
Distinct Count	10	0.06%
Unique Count	0	0.00%
Duplicate Count	10	0.06%
Blank Count	0	0.00%

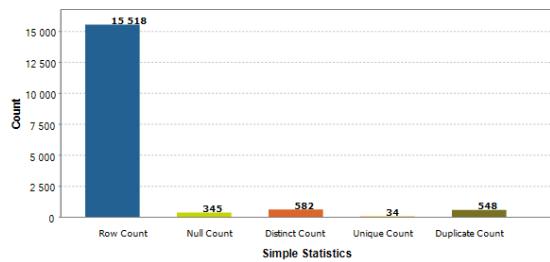


Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata.Total_Cohort_Num

▼ Simple Statistics

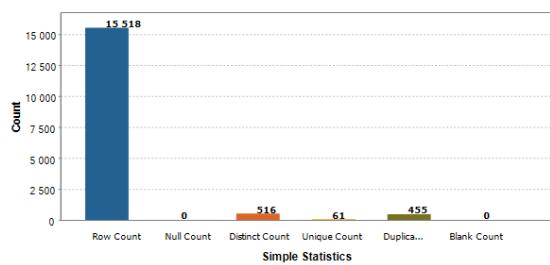
Label	Count	%
Row Count	15518	100.00%
Null Count	345	2.22%
Distinct Count	582	3.75%
Unique Count	34	0.22%
Duplicate Count	548	3.53%



▼ Column: metadata.Total_Grads_Num

▼ Simple Statistics

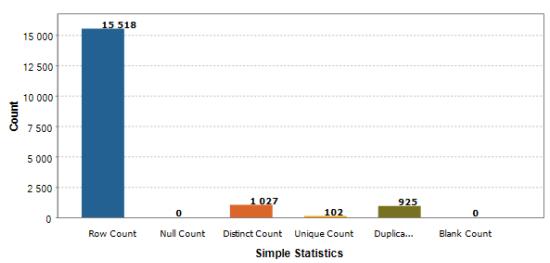
Label	Count	%
Row Count	15518	100.00%
Null Count	0	0.00%
Distinct Count	516	3.33%
Unique Count	61	0.39%
Duplicate Count	455	2.93%
Blank Count	0	0.00%



▼ Column: metadata.Total_Grads_Pct_of_cohort

▼ Simple Statistics

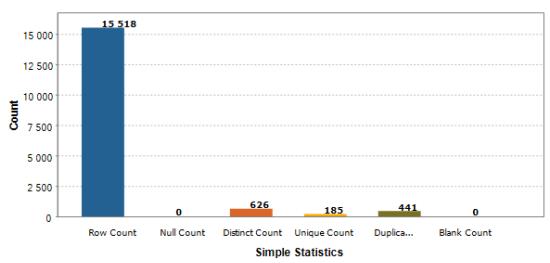
Label	Count	%
Row Count	15518	100.00%
Null Count	0	0.00%
Distinct Count	1027	6.62%
Unique Count	102	0.66%
Duplicate Count	925	5.96%
Blank Count	0	0.00%



▼ Column: metadata.Total_Regents_Num

▼ Simple Statistics

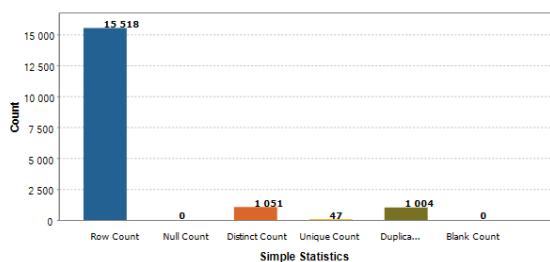
Label	Count	%
Row Count	15518	100.00%
Null Count	0	0.00%
Distinct Count	626	4.03%
Unique Count	185	1.19%
Duplicate Count	441	2.84%
Blank Count	0	0.00%



▼ Column: metadata.Total_Regents_Pct_of_cohort

▼ Simple Statistics

Label	Count	%
Row Count	15518	100.00%
Null Count	0	0.00%
Distinct Count	1051	6.77%
Unique Count	47	0.30%
Duplicate Count	1004	6.47%
Blank Count	0	0.00%

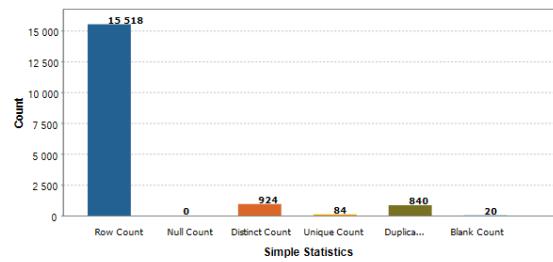


Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata.Total_Regents_Pct_of_grads

▼ Simple Statistics

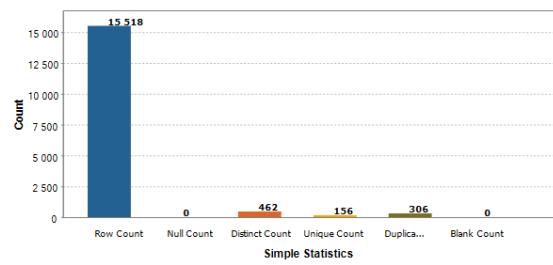
Label	Count	%
Row Count	15518	100.00%
Null Count	0	0.00%
Distinct Count	924	5.95%
Unique Count	84	0.54%
Duplicate Count	840	5.41%
Blank Count	20	0.13%



▼ Column: metadata.Advanced_Regents_Num

▼ Simple Statistics

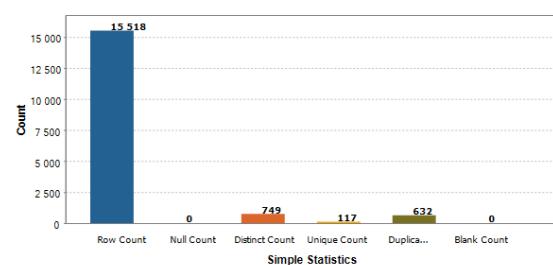
Label	Count	%
Row Count	15518	100.00%
Null Count	0	0.00%
Distinct Count	462	2.98%
Unique Count	156	1.01%
Duplicate Count	306	1.97%
Blank Count	0	0.00%



▼ Column: metadata.Advanced_Regents_Pct_of_cohort

▼ Simple Statistics

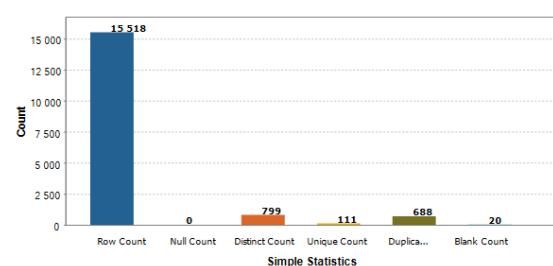
Label	Count	%
Row Count	15518	100.00%
Null Count	0	0.00%
Distinct Count	749	4.83%
Unique Count	117	0.75%
Duplicate Count	632	4.07%
Blank Count	0	0.00%



▼ Column: metadata.Advanced_Regents_Pct_of_grads

▼ Simple Statistics

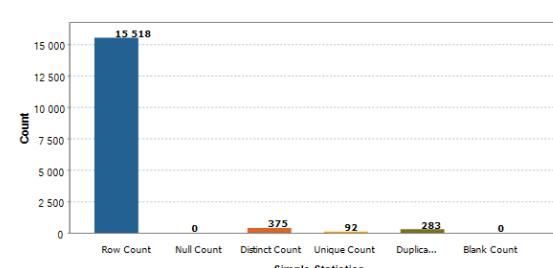
Label	Count	%
Row Count	15518	100.00%
Null Count	0	0.00%
Distinct Count	799	5.15%
Unique Count	111	0.72%
Duplicate Count	688	4.43%
Blank Count	20	0.13%



▼ Column: metadata.Regents_w_o_Advanced_Num

▼ Simple Statistics

Label	Count	%
Row Count	15518	100.00%
Null Count	0	0.00%
Distinct Count	375	2.42%
Unique Count	92	0.59%
Duplicate Count	283	1.82%
Blank Count	0	0.00%

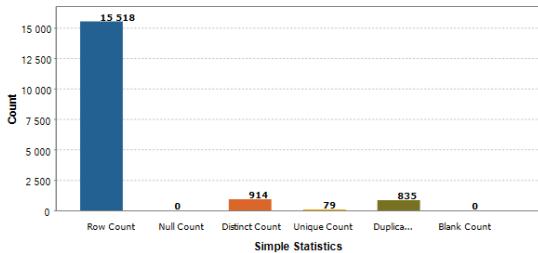


Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata.Regents_w_o_Advanced_Pct_of_cohort

▼ Simple Statistics

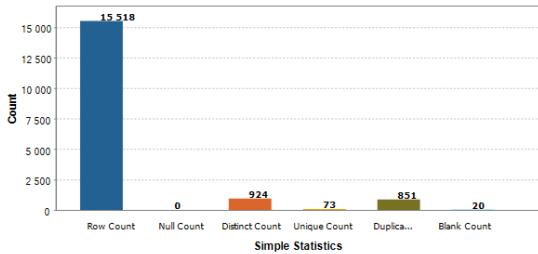
Label	Count	%
Row Count	15518	100.00%
Null Count	0	0.00%
Distinct Count	914	5.89%
Unique Count	79	0.51%
Duplicate Count	835	5.38%
Blank Count	0	0.00%



▼ Column: metadata.Regents_w_o_Advanced_Pct_of_grads

▼ Simple Statistics

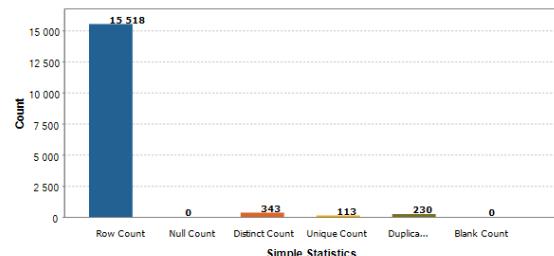
Label	Count	%
Row Count	15518	100.00%
Null Count	0	0.00%
Distinct Count	924	5.95%
Unique Count	73	0.47%
Duplicate Count	851	5.48%
Blank Count	20	0.13%



▼ Column: metadata.Local_Num

▼ Simple Statistics

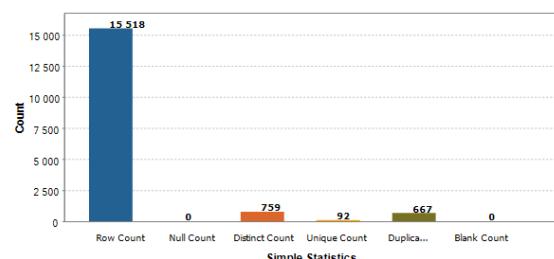
Label	Count	%
Row Count	15518	100.00%
Null Count	0	0.00%
Distinct Count	343	2.21%
Unique Count	113	0.73%
Duplicate Count	230	1.48%
Blank Count	0	0.00%



▼ Column: metadata.Local_Pct_of_cohort

▼ Simple Statistics

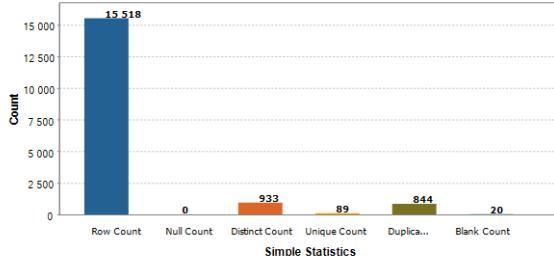
Label	Count	%
Row Count	15518	100.00%
Null Count	0	0.00%
Distinct Count	759	4.89%
Unique Count	92	0.59%
Duplicate Count	667	4.30%
Blank Count	0	0.00%



▼ Column: metadata.Local_Pct_of_grads

▼ Simple Statistics

Label	Count	%
Row Count	15518	100.00%
Null Count	0	0.00%
Distinct Count	933	6.01%
Unique Count	89	0.57%
Duplicate Count	844	5.44%
Blank Count	20	0.13%

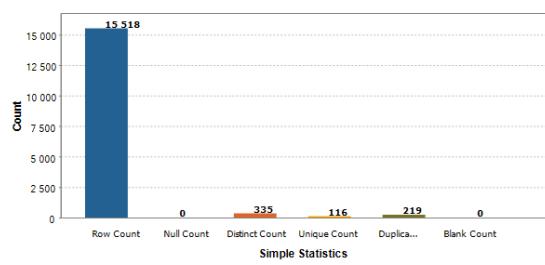


Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata.Still_Enrolled_Num

▼ Simple Statistics

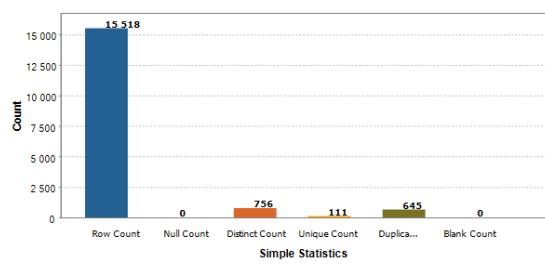
Label	Count	%
Row Count	15518	100.00%
Null Count	0	0.00%
Distinct Count	335	2.16%
Unique Count	116	0.75%
Duplicate Count	219	1.41%
Blank Count	0	0.00%



▼ Column: metadata.Still_Enrolled_Pct_of_cohort

▼ Simple Statistics

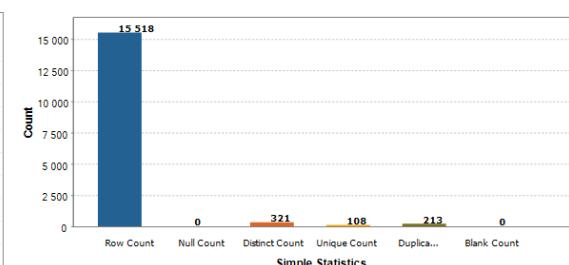
Label	Count	%
Row Count	15518	100.00%
Null Count	0	0.00%
Distinct Count	756	4.87%
Unique Count	111	0.72%
Duplicate Count	645	4.16%
Blank Count	0	0.00%



▼ Column: metadata.Dropped_Out_Num

▼ Simple Statistics

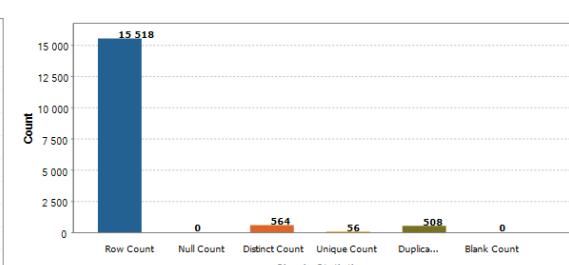
Label	Count	%
Row Count	15518	100.00%
Null Count	0	0.00%
Distinct Count	321	2.07%
Unique Count	108	0.70%
Duplicate Count	213	1.37%
Blank Count	0	0.00%



▼ Column: metadata.Dropped_Out_Pct_of_cohort

▼ Simple Statistics

Label	Count	%
Row Count	15518	100.00%
Null Count	0	0.00%
Distinct Count	564	3.63%
Unique Count	56	0.36%
Duplicate Count	508	3.27%
Blank Count	0	0.00%

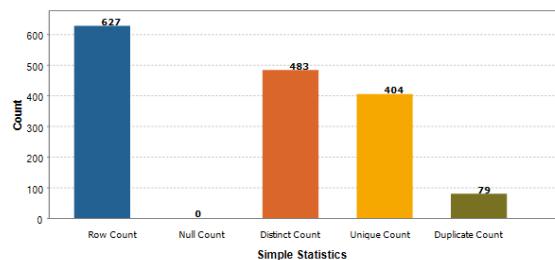


Shootings_NY

▼ Column: metadata INCIDENT_KEY

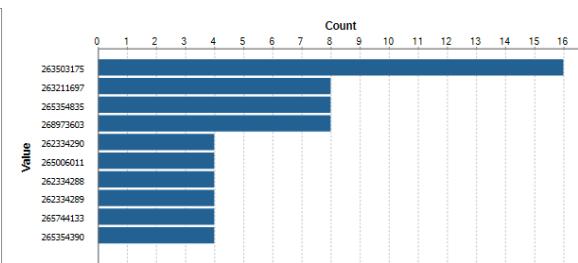
▼ Simple Statistics

Label	Count	%
Row Count	627	100.00%
Null Count	0	0.00%
Distinct Count	483	77.03%
Unique Count	404	64.43%
Duplicate Count	79	12.60%



▼ Value Frequency

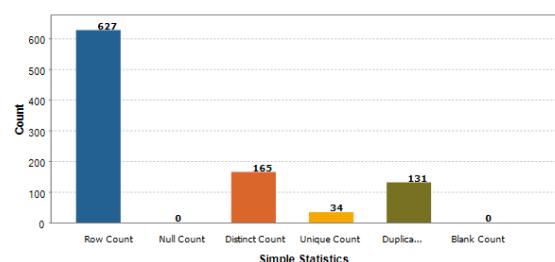
Value	Count	%
263503175	16	2.55%
263211697	8	1.28%
265354835	8	1.28%
268973603	8	1.28%
262334290	4	0.64%
265006011	4	0.64%
262334288	4	0.64%
262334289	4	0.64%
265744133	4	0.64%
265354390	4	0.64%



▼ Column: metadata.OCCUR_DATE

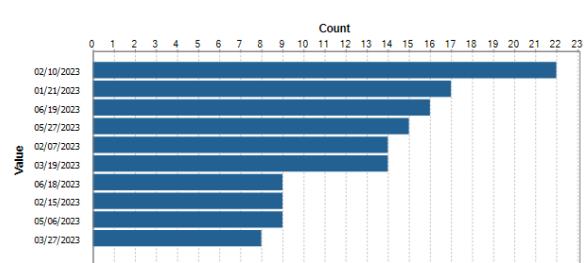
▼ Simple Statistics

Label	Count	%
Row Count	627	100.00%
Null Count	0	0.00%
Distinct Count	165	26.32%
Unique Count	34	5.42%
Duplicate Count	131	20.89%
Blank Count	0	0.00%

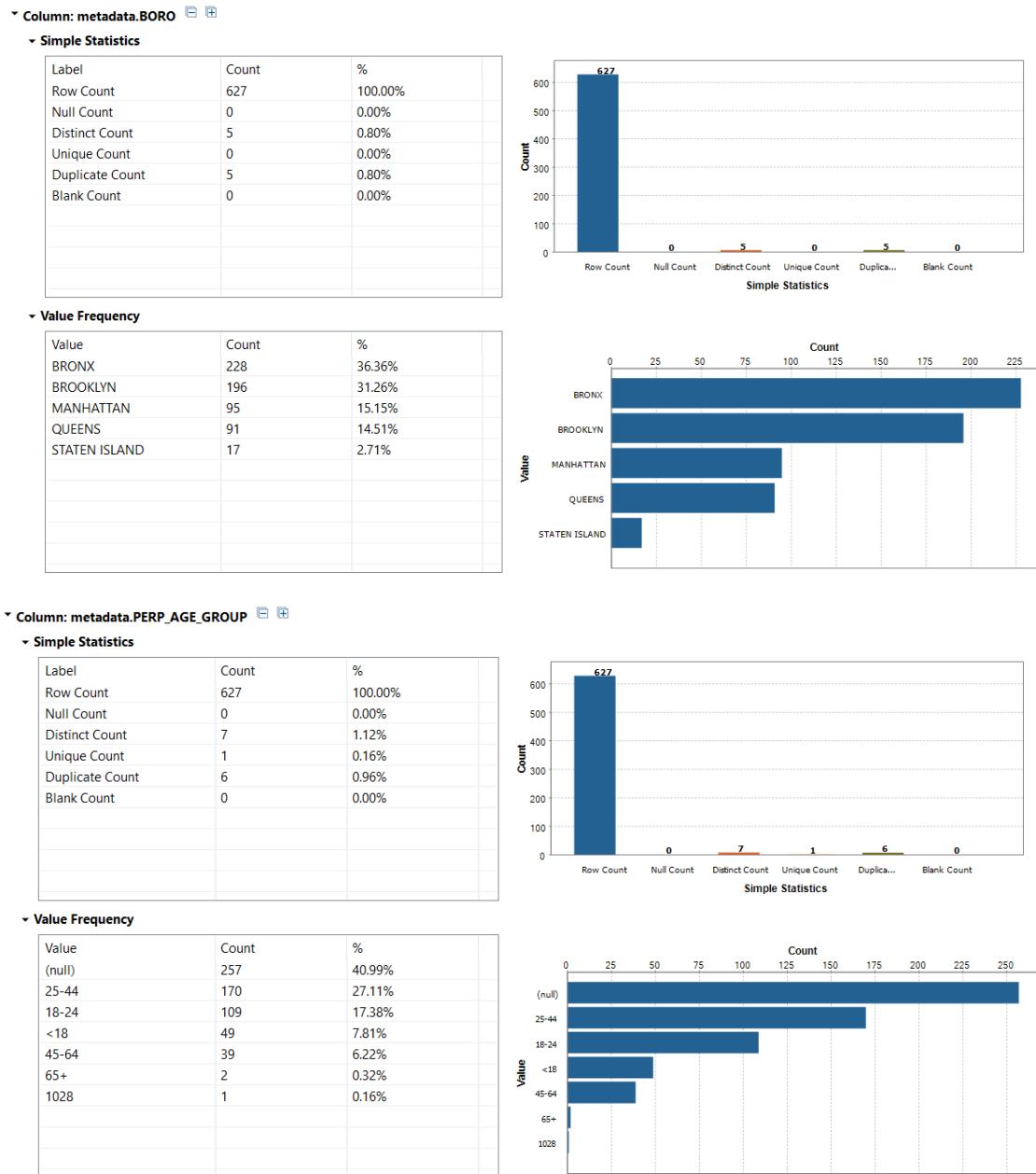


▼ Value Frequency

Value	Count	%
02/10/2023	22	3.51%
01/21/2023	17	2.71%
06/19/2023	16	2.55%
05/27/2023	15	2.39%
02/07/2023	14	2.23%
03/19/2023	14	2.23%
06/18/2023	9	1.44%
02/15/2023	9	1.44%
05/06/2023	9	1.44%
03/27/2023	8	1.28%



Engenharia de Dados para Suporte à Tomada de Decisão

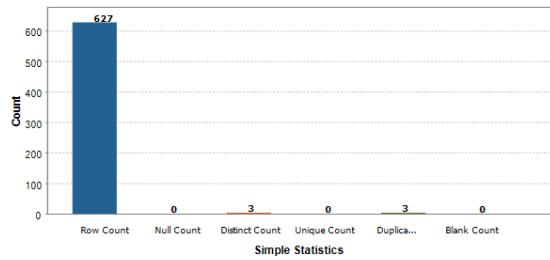


Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata.PERP_SEX

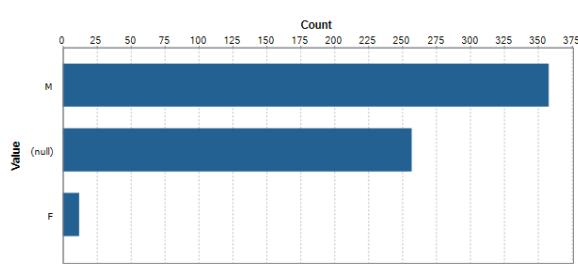
▼ Simple Statistics

Label	Count	%
Row Count	627	100.00%
Null Count	0	0.00%
Distinct Count	3	0.48%
Unique Count	0	0.00%
Duplicate Count	3	0.48%
Blank Count	0	0.00%



▼ Value Frequency

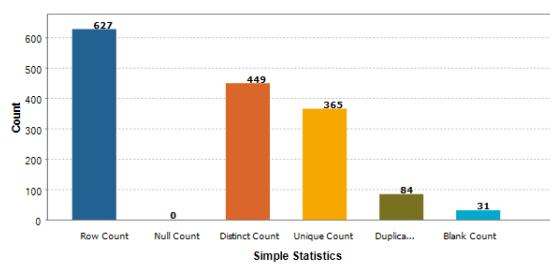
Value	Count	%
M	358	57.10%
(null)	257	40.99%
F	12	1.91%



▼ Column: metadata.Latitude

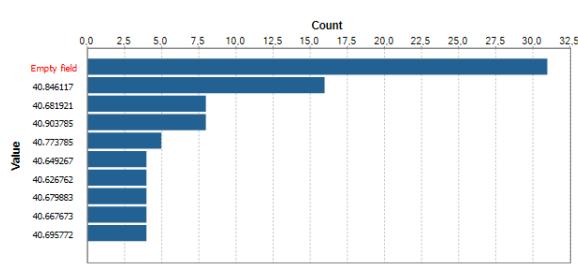
▼ Simple Statistics

Label	Count	%
Row Count	627	100.00%
Null Count	0	0.00%
Distinct Count	449	71.61%
Unique Count	365	58.21%
Duplicate Count	84	13.40%
Blank Count	31	4.94%



▼ Value Frequency

Value	Count	%
Empty field	31	4.94%
40.846117	16	2.55%
40.681921	8	1.28%
40.903785	8	1.28%
40.773785	5	0.80%
40.649267	4	0.64%
40.626762	4	0.64%
40.679883	4	0.64%
40.667673	4	0.64%
40.695772	4	0.64%

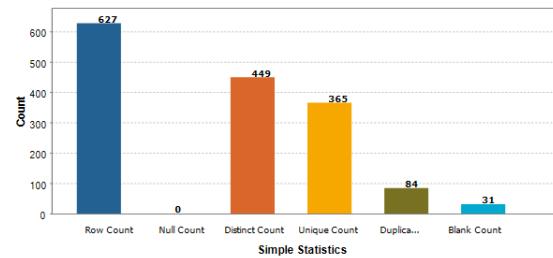


Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata.Longitude

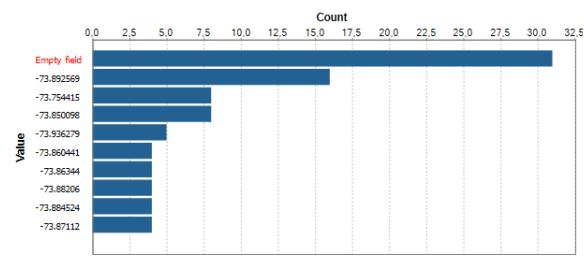
Simple Statistics

Label	Count	%
Row Count	627	100.00%
Null Count	0	0.00%
Distinct Count	449	71.61%
Unique Count	365	58.21%
Duplicate Count	84	13.40%
Blank Count	31	4.94%



Value Frequency

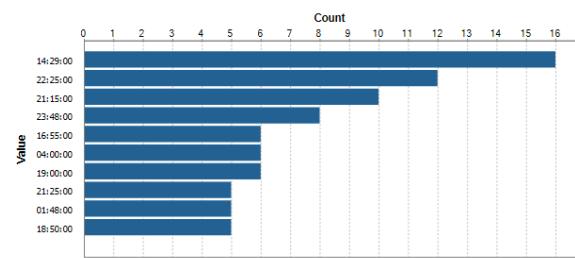
Value	Count	%
Empty field	31	4.94%
-73.892569	16	2.55%
-73.754415	8	1.28%
-73.850098	8	1.28%
-73.936279	5	0.80%
-73.860441	4	0.64%
-73.86344	4	0.64%
-73.88206	4	0.64%
-73.884524	4	0.64%
-73.87112	4	0.64%



▼ Column: metadata.OCCUR_TIME

Value Frequency

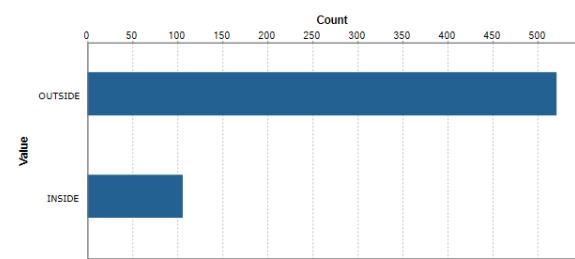
Value	Count	%
14:29:00	16	2.55%
22:25:00	12	1.91%
21:15:00	10	1.59%
23:48:00	8	1.28%
16:55:00	6	0.96%
04:00:00	6	0.96%
19:00:00	6	0.96%
21:25:00	5	0.80%
01:48:00	5	0.80%
18:50:00	5	0.80%



▼ Column: metadata.LOC_OF_OCCUR_DESC

Value Frequency

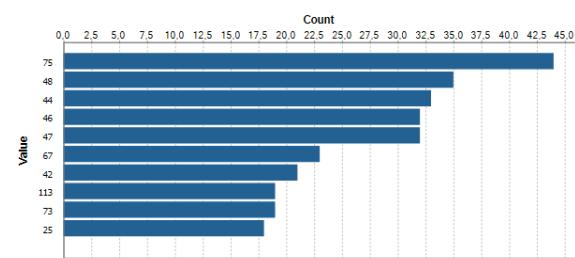
Value	Count	%
OUTSIDE	521	83.09%
INSIDE	106	16.91%



▼ Column: metadata.PRECINCT

Value Frequency

Value	Count	%
75	44	7.02%
48	35	5.58%
44	33	5.26%
46	32	5.10%
47	32	5.10%
67	23	3.67%
42	21	3.35%
113	19	3.03%
73	19	3.03%
25	18	2.87%

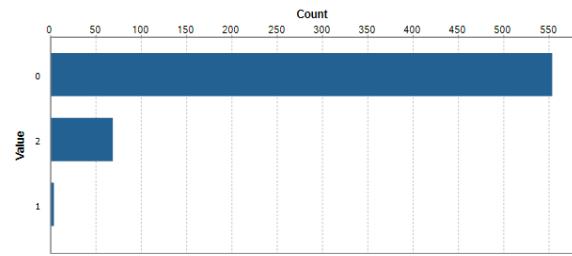


Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata.JURISDICTION_CODE

▼ Value Frequency

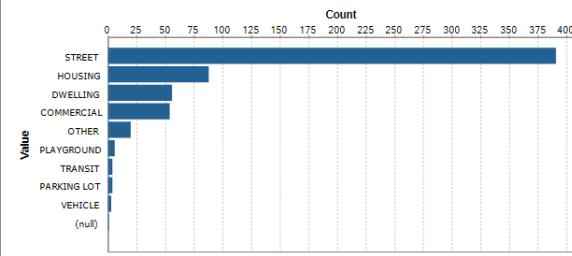
Value	Count	%
0	554	88.36%
2	69	11.00%
1	4	0.64%



▼ Column: metadata.LOC_CLASSFCTN_DESC

▼ Value Frequency

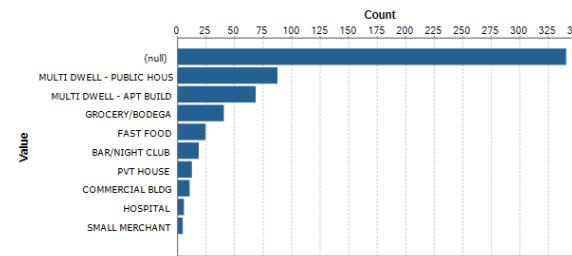
Value	Count	%
STREET	391	62.36%
HOUSING	88	14.04%
DWELLING	56	8.93%
COMMERCIAL	54	8.61%
OTHER	20	3.19%
PLAYGROUND	6	0.96%
TRANSIT	4	0.64%
PARKING LOT	4	0.64%
VEHICLE	3	0.48%
(null)	1	0.16%



▼ Column: metadata.LOCATION_DESC

▼ Value Frequency

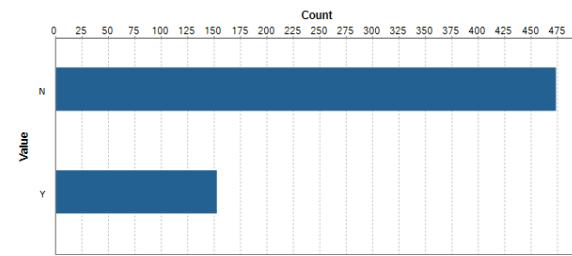
Value	Count	%
(null)	341	54.39%
MULTI DWELL - PUBLIC HO...	88	14.04%
MULTI DWELL - APT BUILD	69	11.00%
GROCERY/BODEGA	41	6.54%
FAST FOOD	25	3.99%
BAR/NIGHT CLUB	19	3.03%
PVT HOUSE	13	2.07%
COMMERCIAL BLDG	11	1.75%
HOSPITAL	6	0.96%
SMALL MERCHANT	5	0.80%



▼ Column: metadata.STATISTICAL_MURDER_FLAG

▼ Value Frequency

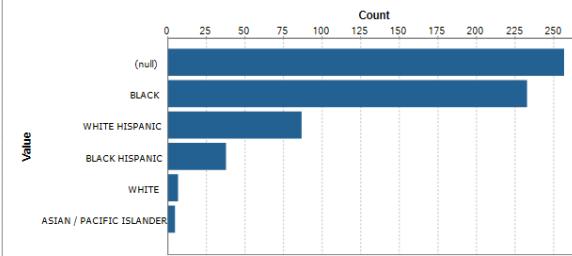
Value	Count	%
N	474	75.60%
Y	153	24.40%



▼ Column: metadata.PERP_RACE

▼ Value Frequency

Value	Count	%
(null)	257	40.99%
BLACK	233	37.16%
WHITE HISPANIC	87	13.88%
BLACK HISPANIC	38	6.06%
WHITE	7	1.12%
ASIAN / PACIFIC ISLANDER	5	0.80%

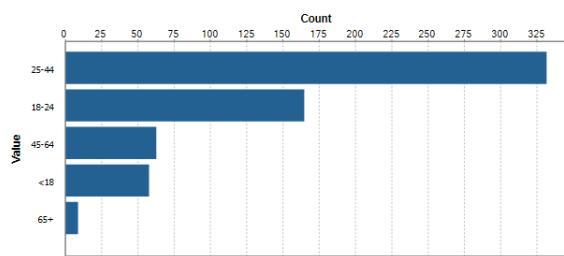


Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata.VIC_AGE_GROUP  

▼ Value Frequency

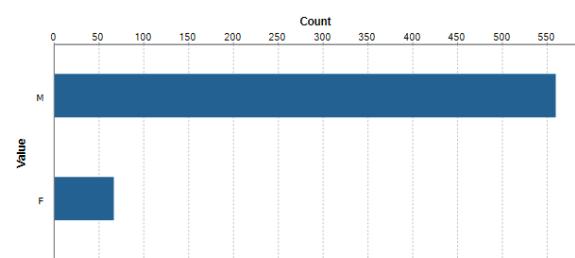
Value	Count	%
25-44	332	52.95%
18-24	165	26.32%
45-64	63	10.05%
<18	58	9.25%
65+	9	1.44%



▼ Column: metadata.VIC_SEX  

▼ Value Frequency

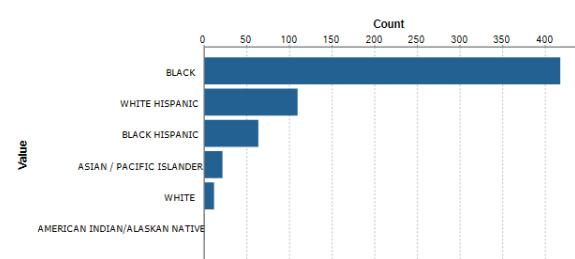
Value	Count	%
M	560	89.31%
F	67	10.69%



▼ Column: metadata.VIC_RACE  

▼ Value Frequency

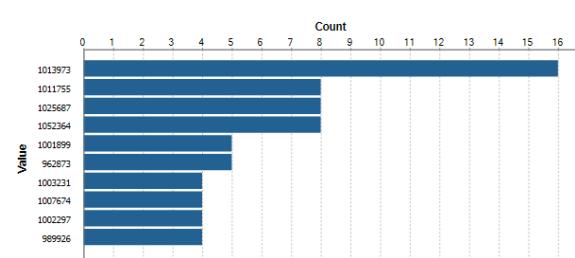
Value	Count	%
BLACK	418	66.67%
WHITE HISPANIC	110	17.54%
BLACK HISPANIC	64	10.21%
ASIAN / PACIFIC ISLANDER	22	3.51%
WHITE	12	1.91%
AMERICAN INDIAN/ALASKA...	1	0.16%



▼ Column: metadata.X_COORD_CD  

▼ Value Frequency

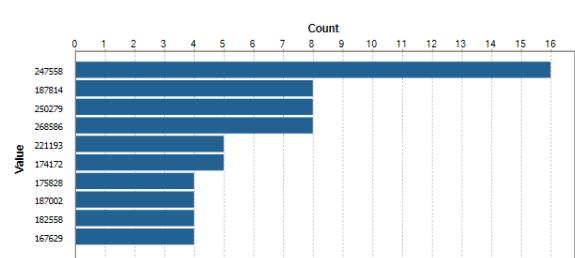
Value	Count	%
1013973	16	2.55%
1011755	8	1.28%
1025687	8	1.28%
1052364	8	1.28%
1001899	5	0.80%
962873	5	0.80%
1003231	4	0.64%
1007674	4	0.64%
1002297	4	0.64%
989926	4	0.64%



▼ Column: metadata.Y_COORD_CD  

▼ Value Frequency

Value	Count	%
247558	16	2.55%
187814	8	1.28%
250279	8	1.28%
268586	8	1.28%
221193	5	0.80%
174172	5	0.80%
175828	4	0.64%
187002	4	0.64%
182558	4	0.64%
167629	4	0.64%

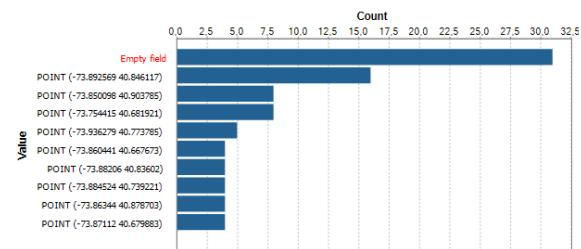


Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata.New_Georeferenced_Column

Value Frequency

Value	Count	%
Empty field	31	4.94%
POINT (-73.892569 40.84617)	16	2.55%
POINT (-73.850098 40.903785)	8	1.28%
POINT (-73.754415 40.681908)	8	1.28%
POINT (-73.936279 40.773785)	5	0.80%
POINT (-73.860441 40.667672)	4	0.64%
POINT (-73.882064 40.83602)	4	0.64%
POINT (-73.884524 40.73922)	4	0.64%
POINT (-73.86344 40.878703)	4	0.64%
POINT (-73.87112 40.679883)	4	0.64%

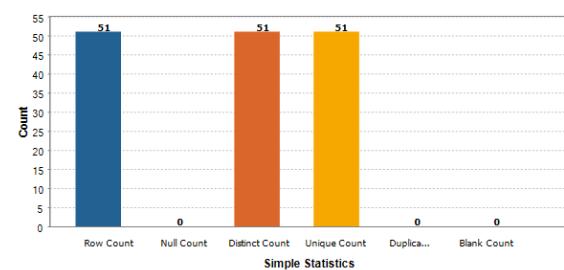


Education_US

▼ Column: metadata.State

Simple Statistics

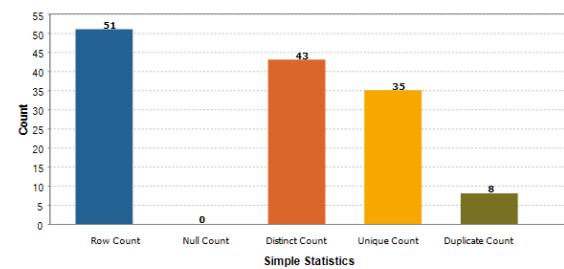
Label	Count	%
Row Count	51	100.00%
Null Count	0	0.00%
Distinct Count	51	100.00%
Unique Count	51	100.00%
Duplicate Count	0	0.00%
Blank Count	0	0.00%



▼ Column: metadata.Less_than_a_High_School_Diploma_Women

Simple Statistics

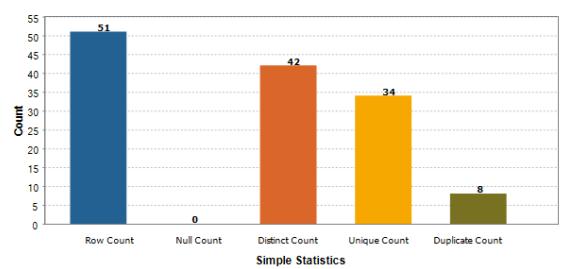
Label	Count	%
Row Count	51	100.00%
Null Count	0	0.00%
Distinct Count	43	84.31%
Unique Count	35	68.63%
Duplicate Count	8	15.69%



▼ Column: metadata.Less_than_a_High_School_Diploma_Men

Simple Statistics

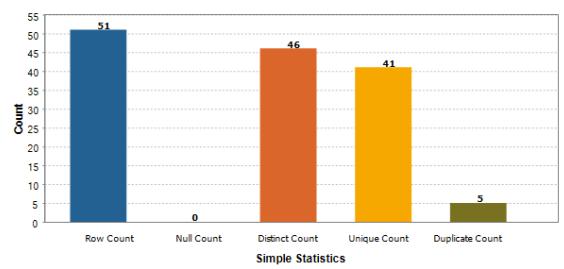
Label	Count	%
Row Count	51	100.00%
Null Count	0	0.00%
Distinct Count	42	82.35%
Unique Count	34	66.67%
Duplicate Count	8	15.69%



▼ Column: metadata.High_School_Diploma_or_the_Equivalent_Only_Women

Simple Statistics

Label	Count	%
Row Count	51	100.00%
Null Count	0	0.00%
Distinct Count	46	90.20%
Unique Count	41	80.39%
Duplicate Count	5	9.80%

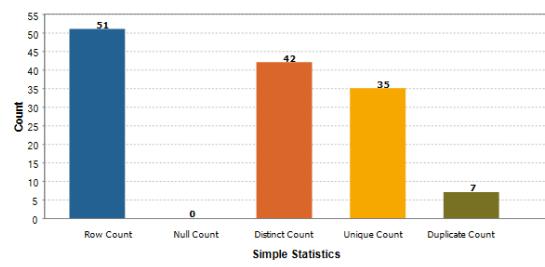


Engenharia de Dados para Suporte à Tomada de Decisão

▼ Column: metadata.High_School_Diploma_or_the_Equivalent_Only_Men

▼ Simple Statistics

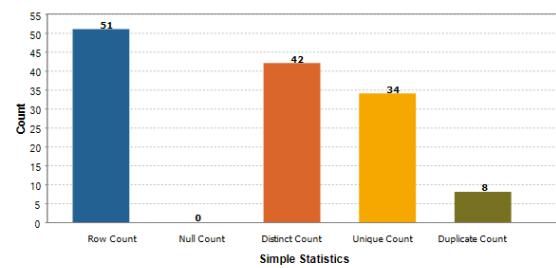
Label	Count	%
Row Count	51	100.00%
Null Count	0	0.00%
Distinct Count	42	82.35%
Unique Count	35	68.63%
Duplicate Count	7	13.73%



▼ Column: metadata.Some_College_or_an_Associate_s_Degree_Women

▼ Simple Statistics

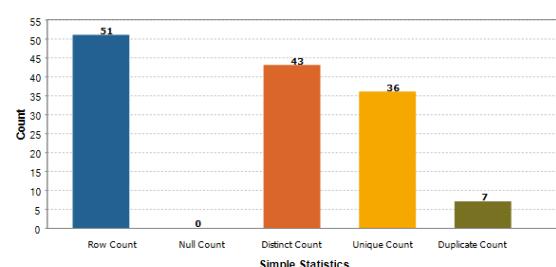
Label	Count	%
Row Count	51	100.00%
Null Count	0	0.00%
Distinct Count	42	82.35%
Unique Count	34	66.67%
Duplicate Count	8	15.69%



▼ Column: metadata.Some_College_or_an_Associate_s_Degree_Men

▼ Simple Statistics

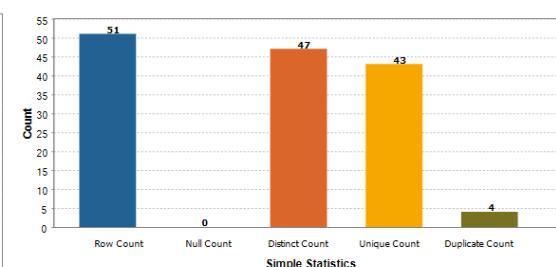
Label	Count	%
Row Count	51	100.00%
Null Count	0	0.00%
Distinct Count	43	84.31%
Unique Count	36	70.59%
Duplicate Count	7	13.73%



▼ Column: metadata.Bachelor_s_Degree_or_Higher_Women

▼ Simple Statistics

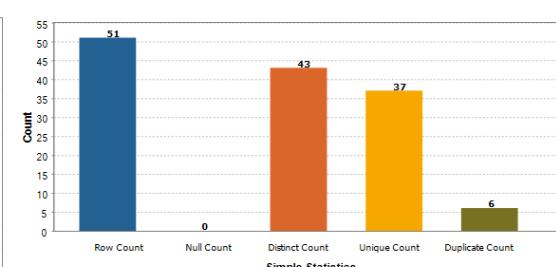
Label	Count	%
Row Count	51	100.00%
Null Count	0	0.00%
Distinct Count	47	92.16%
Unique Count	43	84.31%
Duplicate Count	4	7.84%



▼ Column: metadata.Bachelor_s_Degree_or_Higher_Men

▼ Simple Statistics

Label	Count	%
Row Count	51	100.00%
Null Count	0	0.00%
Distinct Count	43	84.31%
Unique Count	37	72.55%
Duplicate Count	6	11.76%

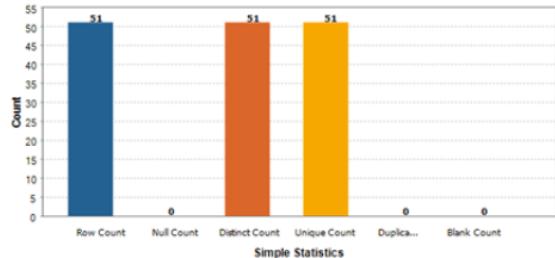


PayGap_US

* Column: metadata_State  

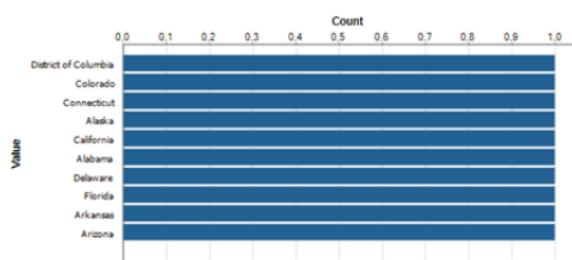
▼ Simple Statistics

Label	Count	%
Row Count	51	100.00%
Null Count	0	0.00%
Distinct Count	51	100.00%
Unique Count	51	100.00%
Duplicate Count	0	0.00%
Blank Count	0	0.00%



▼ Value Frequency

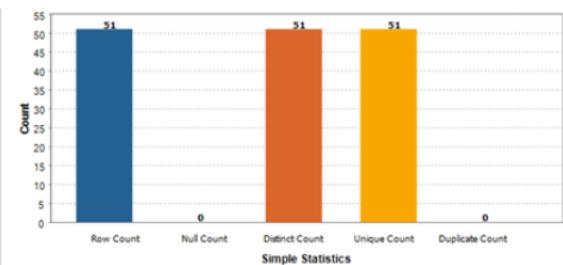
Value	Count	%
District of Columbia	1	1.96%
Colorado	1	1.96%
Connecticut	1	1.96%
Alaska	1	1.96%
California	1	1.96%
Alabama	1	1.96%
Delaware	1	1.96%
Florida	1	1.96%
Arkansas	1	1.96%
Arizona	1	1.96%



* Column: metadata.Male  

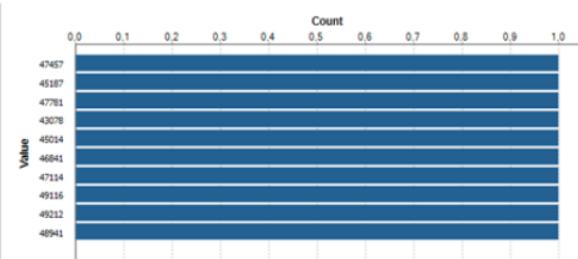
▼ Simple Statistics

Label	Count	%
Row Count	51	100.00%
Null Count	0	0.00%
Distinct Count	51	100.00%
Unique Count	51	100.00%
Duplicate Count	0	0.00%



▼ Value Frequency

Value	Count	%
47457	1	1.96%
45187	1	1.96%
47781	1	1.96%
43078	1	1.96%
45014	1	1.96%
46841	1	1.96%
47114	1	1.96%
49116	1	1.96%
49212	1	1.96%
48941	1	1.96%



Engenharia de Dados para Suporte à Tomada de Decisão



