

INSTITUTO FEDERAL DO NORTE DE MINAS GERAIS  
CAMPUS MONTES CLAROS  
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

**CONSTRUÇÃO DE UM DATA MART  
ACADÊMICO DO IFNMG**

JONAS DIEGO DOS SANTOS  
ORIENTADOR: PROF. DR. LÚCIO FERNANDES DUTRA SANTOS

Montes Claros  
Dezembro de 2019



JONAS DIEGO DOS SANTOS

**CONSTRUÇÃO DE UM DATA MART  
ACADÊMICO DO IFNMG**

Monografia apresentado ao Curso de Graduação em Ciência da Computação do Instituto Federal do Norte de Minas Gerais – Campus Montes Claros, como requisito parcial para a obtenção do grau de Bacharel em Ciência da Computação.

ORIENTADOR: PROF. DR. LÚCIO FERNANDES DUTRA SANTOS

Montes Claros  
Dezembro de 2019



# Agradecimentos

Aos meu pais pelo apoio e incentivo durante todo o meu curso.

Ao meu orientador, Lúcio, pela ajuda no decorrer do projeto, pela paciênciia ao corrigir meus textos pois, sei que não deve ter sido uma tarefa fácil.

Por todos os professores que compartilharam seus conhecimentos comigo, que paravam o que estavam fazendo para prestar auxílio mesmo das atividades que não eram referentes às suas disciplinas.

Aos meus amigos, Igor Alberte, Paulo Márcio, Matheus Henrique, Giovani Moutinho, Alice Ferreira, por terem ajudado e me aturado ao longo do curso, promovendo momentos de estudo e descontração.

Aos colegas da faculdade, amigos e paresntes, obrigado pelos momentos de estudo, diversão e compreensão.

A Eiichiro Oda por prover vários momentos de descontração nos meus Domingos.



*“Não importa o que o mundo diz de mim, o que importa é que eu nunca fiz nada que contrariasse os meus princípios e nunca farei.”*

(Roronoa Zoro)



# Resumo

Empresas utilizam *softwares* para a realização de tarefas cotidianas, sendo efeito direto a redução de erros operacionais, um aumento da produtividade e a confiança na produção ou execução das tarefas, mas resultam em um acúmulo de grandes quantidades de dados com pouca informação disponível. Uma solução para este problema de pouca informação e grande quantidade de dados é a utilização de tecnologias de inteligência de negócio como *data mart* que são utilizadas para transformar dados brutos em informação disponível para decisões estratégicas. Foi observado que os dados do sistema transacional do questionário socioeconômico do Instituto Federal do Norte de Minas Gerais (IFNMG) não ficam disponíveis para análise após a realização do processo seletivo. Assim, o objetivo deste trabalho foi construir um *data mart* acadêmico para disponibilização de informação gerencial e por meio da utilização de ferramentas de processamento analítico *on-line* (OLAP sigla em inglês - *On-Line Analytical Processing*). As análises realizadas no ambiente proposto, tornaram possível extrair informações do perfil geral dos candidatos e a verificar se há alguma alteração nos quatro *Campi* com mais candidatos, onde foi possível identificá-lo através de seu perfil social, educacional e econômico, obtendo informações como o nível escolar dos pais pode influenciar o local onde o candidato estudou no ensino fundamental e ensino médio.

**Palavras-chave:** *data mart*, ferramentas de processamento analítico *on-line*, dados educacionais.



# Abstract

Organizations use software to perform a variety of everyday tasks, with the direct effect of reducing operational errors, increasing productivity, and relying on the production or execution of tasks, but resulting in large amounts of data with low to none available information. One solution is the use of business intelligence technologies such as data mart that transform raw data into available information for strategic decisions. It was noted that data from the transaction system of the IFNMG socioeconomic questionarie of the Federal Institute of Northern Minas Gerais (IFNMG) are not available for analysis after the selection process. Thus, the main goal of this work was to build an academic data mart using online analytical processing tools (OLAP). The analyzes carried out in the proposed environment made it possible to extract information from the candidates general profile and to verify if there is any alteration in the four campuses with more candidates, where it was possible to identify the candidate through their social, educational and economic profile obtaining information such as the influence of parents' school life by the place of study of the candidate in elementary and high school.

**Keywords:** business intelligence, data mart, analytical processing tools online, educational data.



# Sumário

<b>Agradecimentos</b>	<b>v</b>
<b>Resumo</b>	<b>ix</b>
<b>Abstract</b>	<b>xi</b>
<b>Lista de Figuras</b>	<b>xv</b>
<b>Lista de Tabelas</b>	<b>xix</b>
<b>1 Introdução</b>	<b>1</b>
1.1 Objetivo Geral . . . . .	3
1.2 Objetivos Específicos . . . . .	3
1.3 Organização do Documento . . . . .	3
<b>2 Inteligência de Negócios</b>	<b>5</b>
2.1 <i>Data Warehouses</i> . . . . .	5
2.1.1 OLAP X OLTP . . . . .	7
2.1.2 <i>Data Mart</i> . . . . .	9
2.1.3 Componentes de um <i>Data Warehouse</i> . . . . .	9
2.1.4 Tipos de implementação de DW . . . . .	11
2.1.5 Modelo dos dados . . . . .	12
2.2 OLAP . . . . .	16
2.3 Trabalhos Relacionados . . . . .	18
2.4 Ferramentas . . . . .	20

<b>3</b>	<b><i>Data mart</i> Acadêmico IFNMG: Processo Seletivo</b>	<b>23</b>
3.1	Levantamento dos Requisitos de Informação . . . . .	24
3.2	Obtenção dos Dados Operacionais . . . . .	25
3.3	Modelo Físico do <i>Data mart</i> . . . . .	27
3.4	Extração, Transformação e Carga dos dados . . . . .	28
3.5	Definição dos Cubos . . . . .	32
3.6	Tecnologias OLAP . . . . .	34
3.6.1	QlikView . . . . .	38
<b>4</b>	<b>Resultados</b>	<b>45</b>
4.1	Visão Geral . . . . .	45
4.2	Análise do Perfil-Econômico . . . . .	50
4.2.1	Perfil dos Candidatos que Exercem Atividades Remuneradas	51
4.2.2	Perfil dos Candidatos que não Exercem Atividades Remuneradas . . . . .	57
4.3	Análise Perfil-Educacional . . . . .	65
4.4	Análise do Perfil-Social . . . . .	71
<b>5</b>	<b>Conclusão</b>	<b>77</b>
5.1	Trabalhos Futuros . . . . .	78
	<b>Referências Bibliográficas</b>	<b>79</b>

# Listas de Figuras

2.1 Componentes de um data Warehouse. Adaptação da Fonte Santos [2010].	10
2.2 Implementação <i>Top Down</i> . Fonte: <a href="http://www.dataprix.net/pt-pt/24-data-mart">http://www.dataprix.net/pt-pt/24-data-mart</a> .	11
2.3 Implementação <i>Bottom Up</i> . Fonte: <a href="http://www.dataprix.net/pt-pt/24-data-mart">http://www.dataprix.net/pt-pt/24-data-mart</a> .	12
2.4 Modelo Estrela.	14
2.5 Modelo Floco de Neve	15
3.1 Fluxo de definição do <i>Data Mart</i> .	23
3.2 Código de exemplo da padronização do nome do campus.	27
3.3 Código de exemplo da padronização do nome do campus.	31
3.4 Representação do cubo Perfil-Econômico.	32
3.5 Representação do cubo Perfil-Educacional.	33
3.6 Representação do cubo Perfil-Social.	34
3.7 Tela de inicialização do <i>QlikView</i> .	38
3.8 Janela de configuração da fonte de dados.	40
3.9 Janela de configuração do acesso aos dados do <i>data mart</i> .	41
3.10 Janela de criação de gráficos do <i>QlikView</i> .	42
3.11 Janela criada para fazer filtros no <i>QlikView</i> .	43
3.12 Exemplo de pasta com <i>dashboard</i>	43
4.1 Distribuição dos candidatos por campus.	46
4.2 Quantidade de candidatos nos anos de 2013, 2016 e 2019	47
4.3 Distribuição da etnia dos candidatos em relação aos campi.	48

4.4	A distribuição dos candidatos em relação a modalidade de ensino nos quatro campi com mais inscrições. (a) Visão Geral, (b) Montes Claros, (c) Pirapora, (d) Salinas, (e) Januária. . . . .	49
4.5	Quantidade de candidatos que exercem e não exercem atividades remuneradas. . . . .	50
4.6	Quantidade de Candidatos que Exercem Atividade Remunerada X Campus . . . . .	51
4.7	Faixa Etária X Campus. . . . .	52
4.8	Quantidade de candidatos que exercem atividades remuneradas X Turno. . . . .	53
4.9	Quantidade de candidatos que exercem atividades remuneradas X modalidade. . . . .	54
4.10	Quantidade de candidatos que exercem atividades remuneradas X Renda Mensal do Candidato em todos os campus. . . . .	55
4.11	Quantidade de candidatos que exercem atividades remuneradas X Renda Mensal do Candidato. . . . .	55
4.12	Quantidade Candidatos X Renda Bruta x Exercer Atividade Remunerada. . . . .	56
4.13	Quantidade Candidatos X Número de membros na Família Geral X exerce Atividade Remunerada. . . . .	56
4.14	Quantidade Candidatos X Número de membros na Família X exerce Atividade Remunerada. . . . .	57
4.15	Quantidade Candidatos X <i>Campus</i> X Não Exercer Atividade Remunerada. . . . .	58
4.16	Quantidade Candidatos que não exercem atividade remunerada X Pelos anos de 2013, 2016 e 2019. . . . .	58
4.17	Quantidade candidatos que não exercem atividade remunerada X Turno. . . . .	60
4.18	Quantidade Candidatos que não exercem atividade remunerada X Renda Mensal do Candidato. . . . .	61
4.19	Quantidade candidatos que não exercem atividade remunerada X Renda Bruta da Família. . . . .	61
4.20	Quantidade candidatos que não exercem atividade remunerada X Renda Bruta da Família. . . . .	62
4.21	Quantidade Candidatos que não exercem atividade remunerada X Situação dos pais. . . . .	64

4.22 Quantidade Candidatos X Escola que estudou no Ensino Fundamental e no Ensino Médio, Geral. . . . .	65
4.23 Quantidade Candidatos X Escola que estudou no Ensino Médio. . . . .	66
4.24 Quantidade Candidatos X Escola que estudou no Ensino Fundamental. . . . .	67
4.25 Quantidade Candidatos X Repetiu no Ensino Fundamental. . . . .	68
4.26 Quantidade Candidatos X Repetiu no Ensino Médio. . . . .	69
4.27 Quantidade Candidatos X Número de Reprovações durante o Ensino Fundamental X Ler Jornais e Revistas. . . . .	70
4.28 Quantidade Candidatos X Número de Reprovações durante o Ensino Fundamental X Quantidade Livros por Ano. . . . .	70
4.29 Quantidade de candidatos X Escolaridade dos pais. . . . .	71
4.30 Quantidade de Candidatos X Escolaridade dos pais X Campus de Montes Claros e Januária. . . . .	72



# **Lista de Tabelas**

2.1	Tabela com o comparativo entre OLAP e OLTP Lemos [2015] . . . . .	8
2.2	Tabela com o comparativo entre o Modelo Estrela e o Modelo Floco de Neve Castro Novais [2012] . . . . .	15
2.3	Tabela com breve descrição das possíveis ferramentas OLAP a serem utilizadas. . . . .	21
3.1	Perguntas do Questionário Socioeconômico. . . . .	25
3.2	Pré-requisitos computacionais para a instalação das ferramentas. . . . .	35
3.3	Características da documentação. . . . .	36
3.4	Aspectos de usabilidade das ferramentas. . . . .	36
3.5	Características de visualização das ferramentas. . . . .	37
4.1	Renda Bruta Familiar X Quantidade Membros na Família. . . . .	63
4.2	Renda Bruta Familiar X Quantidade Membros na Família X Campus Montes Claros. . . . .	63
4.3	Situação de escolaridade da Mãe X Instituição onde o Candidato estudou durante o Ensino Médio. . . . .	73
4.4	Situação de escolaridade da Mãe X Instituição onde o Candidato estudou durante o Ensino Fundamental. . . . .	74
4.5	Situação de escolaridade do Pai X Instituição onde o Candidato estudou durante o Ensino médio. . . . .	75
4.6	Situação de escolaridade do Pai X Instituição onde o Candidato estudou durante o ensino fundamental. . . . .	75



# Capítulo 1

## Introdução

Computadores e celulares são muito utilizados nas empresas para a realização de suas tarefas, estes contêm *softwares* especialistas que auxiliam ou realizam as atividades da empresa para os funcionários. Cada atividade gera um conjunto de dados que não trazem explicitamente informações aos executivos para que possam tomar alguma decisão a fim de melhorar o lucro ou o desempenho da empresa. Obter tais informações por meio dos dados é muitas vezes deixado de lado, pois outros gastos como com: desenvolvimento de sistemas, falhas estruturais, dentre outros são fundamentais para que a empresa continue em funcionamento [Santos, 2010].

Com o passar do tempo são armazenados grandes volumes de dados que dificultam ainda mais a visualização de informações. Os dados podem estar armazenados de diversas maneiras, podem ter padrões diferentes e também assumem valores diferentes. É importante ressaltar que existe uma diferença entre dado e informação, pois, dados são números, palavras, frases, registros. Já informação é um conjunto de dados organizados que fazem sentido aos leitores. Segundo Efraim [2009], muitas vezes, as organizações estão tão sobrecarregadas de dados que os gerentes podem não ter uma forma de interpretá-los ou podem não ser capazes de compilá-los para obter relatórios a tempo de tomar suas decisões.

Para obter informações através dos dados são utilizadas técnicas de inteligência de negócio ( BI sigla em inglês - *Business Intelligence*), que é um termo “guarda-chuva”, englobando um conjunto de ferramentas utilizadas para análise de dados,

propiciando criação de relatórios, consultas, entrega de informações a fim de apoiar no processo de tomada de decisões [Efraim, 2009; Barbieri, 2011; Baltzan, 2012; Popovic, 2012]. As ferramentas BI são principalmente utilizadas para alavancar a rentabilidade de uma empresa, nas organizações como universidades, cujo foco não seja necessariamente o lucro, começou-se a investir nesse tipo de sistema, pois os custos são sempre crescentes e torna-se importante conhecer seus discentes para atraí-los e garantir sua permanência na instituição [Schiel, 2004; Camargo, 2013; Clemes, 2001]. Por exemplo, as universidades vêm utilizando esses sistemas para planejar estratégias de *marketing*; planejar a utilização de receitas para os projetos de extensão, evitando superávit e déficit no orçamento; aplicar de maneira mais produtiva as receitas da universidade.

Dentre as técnicas de BI, uma se destaca pela sua utilização nas universidades, os *data warehouses*, que consistem na criação de um ambiente, que seja possível extrair informações com ferramentas OLAP, de maneira rápida, eficiente, segura e confiável. Neste contexto acadêmico, algumas instituições de ensino brasileiras como UNIPAMPA [Camargo, 2013], UESB [Santos, 2010], IFTM [Damasceno, 2017], UFRN [Steidel, 2017] já incluíram esta técnica como forma de auxiliar os gestores no processo de tomada de decisão.

No entanto, o Instituto Federal do Norte de Minas Gerais (IFNMG) que abrange um área de quase 51% do território do estado de Minas Gerais, não possui ferramentas como *data warehouses* para auxiliar a tomada de decisão. Existem diversas bases de dados operacionais referentes à vida acadêmica dos alunos matriculados no IFNMG, armazenados no sistema CAJUÍ. Há também base de dados com as respostas dos candidatos do questionário socioeconômico aplicado aos candidatos às vagas do vestibular, sendo que este é aplicado por recomendação do Ministério da Educação (MEC). Não obstante, observou-se que, os dados referentes ao questionário não estão disponíveis aos gestores para uma análise do perfil dos candidatos à vaga, fazendo com que decisões sobre estratégias de divulgação, planos de permanência e êxito, planos de ensino para as turmas de primeiro ano não sejam tomadas com base em dados, contando-se apenas com a intuição dos gestores.

## 1.1 Objetivo Geral

O objetivo geral deste trabalho é construir um repositório de dados analítico sobre a base de dados do questionário socioeconômico do Instituto Federal do Norte de Minas Gerais para auxiliar no processo de decisão dos atuais gestores e determinar o perfil dos candidatos às vagas dos processos seletivos da Instituição.

## 1.2 Objetivos Específicos

Os objetivos específicos deste trabalho são:

- Definir um repositório de dados (*data mart*), sobre os dados do questionário socioeconômico do IFNMG utilizando tecnologias que possuam licença de *software* livre;
- Definir uma estratégia de extração, transformação e carga dos dados (ETL sigla em inglês – *Extraction-Transformation-Load*) oriundos das bases de dados transacionais do questionário socioeconômico do IFNMG;
- Definir fato e dimensões de análises que devem ser contempladas para os cubos de dados;
- Avaliar e definir uma ferramenta OLAP intuitiva para geração de gráficos interativos, permitindo que gestores utilizem a ferramenta sem ter um conhecimento técnico especializado;
- Definir o perfil geral dos candidatos do IFNMG. Analizar este perfil nos quatro *campi* com mais candidatos, a fim de observar se existe alguma alteração entre eles.

## 1.3 Organização do Documento

Este trabalho foi organizado do seguinte modo: o Capítulo 1 contém introdução e objetivos. O Capítulo 2 aborda os conceitos básicos sobre *Data Warehouse* e trabalhos relacionados sobre o uso de DW em instituições de ensino. Já o Capítulo

3 apresenta a processo de criação do *data mart* acadêmico do IFNMG. Capítulo 4 as análises obtidas. Por fim, Capítulo 5 apresenta as conclusões deste trabalho.

# Capítulo 2

## Inteligência de Negócios

### 2.1 *Data Warehouses*

Segundo Steidel [2017], *Data Warehouse* pode ser definido como: locais onde ficam concentrados todos os dados extraídos dos sistemas operacionais. Já Inmon [2005], define como uma coleção de dados orientada a assunto, integrada, não-volátil, e variante no tempo usada para suporte a decisões gerenciais. *Data Warehouse* possui dados corporativos granulados que uma vez armazenados estão disponíveis para serem utilizados para diversos fins, incluindo a espera de futuras exigências que hoje são desconhecidas.

Existem diversas definições acerca do *Data Warehouse*. Assim, podemos definir neste trabalho que DW é um banco de dados onde é feita a integração de todos os dados de uma empresa, fornecendo uma visão padronizada destes para uma futura análise. Segundo Santos [2010], DW também permite armazenar os dados baseando-se em assuntos ou processos de negócios da empresa, mantendo toda informação necessária que auxilie na tomada de decisão, assuntos estes que podem ser vendas, produção, clientes, campanha de *marketing*, dentre outros assuntos. Além disso, ainda é possível manter dados históricos, o que é interessante, uma vez que a base de dados cresce à proporção de terabytes é possível uma análise evolutiva das tarefas.

É importante ressaltar que dados não confiáveis, levam a informações não confiáveis, logo, os gestores não as utilizam. Colangelo Filho [2001], cita a importância

da qualidade dos dados, pois o produto de um *data warehouse* é a informação que será utilizada para tomar decisão. Se a matéria-prima é de má qualidade, o produto não poderá ser bom e a atividade suportada pelo produto, ou seja, a decisão, ficará comprometida. Em resumo, o processo de *data warehouse* permite reunir os dados operacionais de toda a empresa integrando-os numa forma consistente, padronizada, segura e facilmente disponível para ser utilizada por qualquer gestor interessado na informação independentemente da forma que será utilizada.

Com a definição do que é DW, é necessário pontuar suas principais características a fim de mostrar o motivo dele ser tão utilizado. As características que serão analisadas são: orientação a assunto, não-volatilidade, integração e variação do tempo.

Um DW é orientado a assunto pois permite distinguir a qual assunto os dados pertencem. Um exemplo é se os dados pertencem a vendas, produtos, qualidade ou clientes.

A característica de não-volatilidade indica que uma vez inseridos os dados no *data warehouse*, só poderá ser executada a operação de visualização sobre eles. Logo, um *data warehouse* possui apenas duas operações básicas: a primeira é a inserção dos dados, que ocorre no início do *data warehouse* e no incremento deste que pode acontecer dentro de um intervalo de tempo de 1 a 5 anos de sua última atualização; a segunda operação é o acesso aos dados apenas para leitura, ou seja, as consultas.

Uma empresa não está necessariamente situada em apenas um local, os dados não são acessados em apenas um sistema, logo, é necessário que o DW seja integrado, o problema é a origem dos sistemas, que podem ter todos os tipos de dados de forma não padronizada, por isso, é feito uma transformação para que no DW tudo se reflita ao mesmo assunto, mesmo que antes esteja armazenado de formas diferentes. Assim, nesse processo, é necessário que ocorra uma padronização dos dados antes de serem inseridos. Um exemplo de padronização a ser feita é o atributo sexo, que em algumas bases é representado com 0/1, em outras como F/M e uma terceira forma com M/H. No *data warehouse* sexo será codificado ou padronizado apenas por F/M.

A característica de variação no tempo diz que os dados no *data warehouse* são específicos de algum momento no tempo, associado com a característica da não-

volatilidade e com o fato de que não existe a operação atualização de dados num *data warehouse*, toda modificação ou novo incremento resulta numa nova entrada. Dessa maneira, os dados históricos são guardados, diferente dos banco de dados transacionais em que um registro no tempo vale apenas para aquele instante e qualquer alteração leva à modificação desse mesmo registro. Como exemplo o total de vendas de um dia tem valores diferentes no início e no final do dia.

A fim de deixar o entendimento do DW mais claro, será mostrada a diferença entre um banco de dados transacional que utiliza processamento de transações em tempo real (OLTP sigla em inglês – *Online Transaction Processing*) e um *data warehouse* que utiliza processamento analítico *on-line* (OLAP sigla em inglês - *On-Line Analytical Processing*).

### 2.1.1 OLAP X OLTP

Um banco de dados operacional é focado em armazenar os dados das operações diárias a fim de deixar a empresa em funcionamento. O OLTP é voltado para sistemas de transações, pois é necessário manter todas as operações realizadas dentro da empresa, e nele também são aplicadas as regras de negócio. Com o auxílio do banco de dados relacional é possível manter os dados com pouco espaço e otimizar consultas. Porém, este tipo de banco não guarda dados históricos, pois para manter o sistema mais eficiente os dados são retirados, o que dificulta a análise dos gestores já que eles possuem apenas os dados que permaneceram no sistema.

Já um DW é focado na parte da análise de informações, trazendo informações aos gestores para que tomem decisões precisas e eficientes para a empresa. O OLAP é voltado para análise de informação e para solução de problemas de atemporalidade e de integração dos dados. Para isso é utilizado o DW, pois ele fornece a integração entre as filiais e mantém um histórico, além de manter uma padronização nos dados, o que torna a análise dos dados eficiente.

Com as características de OLTP e OLAP definidas, nota-se que elas são distintas, pois cada um é focado na necessidade de seu usuário. A Tabela 2.1 mostra um quadro resumido das diferenças entre os sistemas OLTP e OLAP.

**Tabela 2.1.** Tabela com o comparativo entre OLAP e OLTP Lemos [2015]

<i>Característica</i>	<i>OLAP</i>	<i>OLTP</i>
Foco	Foco no nível estratégico da organização. Visa à análise empresarial e à tomada de decisão.	Foco no nível operacional da organização. Visa à execução operacional do negócio
Desempenho	Otimização para a leitura e geração de análises e relatórios gerenciais	Alta velocidade na manipulação de dados operacionais, porém ineficiente para geração de análises gerenciais
Estrutura dos dados	Os dados estão estruturados na modelagem dimensional. Os dados normalmente possuem alto nível de sumarização	Os dados são normalmente estruturados em um modelo relacional normalizado, otimizado para a utilização transacional. Os dados possuem alto nível de detalhes
Armazenamento	O armazenamento é feito em estruturas de Data Warehouse com otimização no desempenho em grandes volumes de dados.	O armazenamento é feito com sistemas convencionais de banco de dados através dos sistemas de informações da organização
Abrangência	É utilizado pelos gestores e analistas para a tomada de decisão	É utilizado por técnicos e analistas e engloba vários usuários da organização
Frequência de atualização	A atualização das informações é feita no processo de carga dos dados. Frequência baixa podendo ser diária, semanal, mensal ou anual (ou a critério específico)	A atualização dos dados é feita no momento da transação. Frequência muito alta de atualizações
Volatilidade	Dados históricos e não voláteis. Os dados não sofrem alterações, salvo por necessidades específicas (por motivos de erros ou inconsistências de informações)	Dados voláteis passíveis de modificação e exclusão
Tipos de permissões de dados	É permitida apenas a inserção e leitura, sendo que para o usuário está apenas disponível a leitura	Pode ser feita leitura, inserção, modificação e exclusão dos dados

### 2.1.2 Data Mart

Um *data mart* é um subconjunto de um *data warehouse*, que segundo Kimball [2002] o *data warehouse* é criado através da união de todos os *data marts*, já segundo Inmon [2005], um *data warehouse* deve ser pensado como um todo depois subdividido em pequenos conjuntos que seriam os *data marts*, ele também ressalta que ao se fazer pequenos subconjuntos e depois integrá-los não resultará em um DW, pelo fato de que cada subconjunto pode ser desenvolvido por equipes distintas, dificultando a integração, e mesmo que consiga integrar os subconjuntos não resultará em um DW.

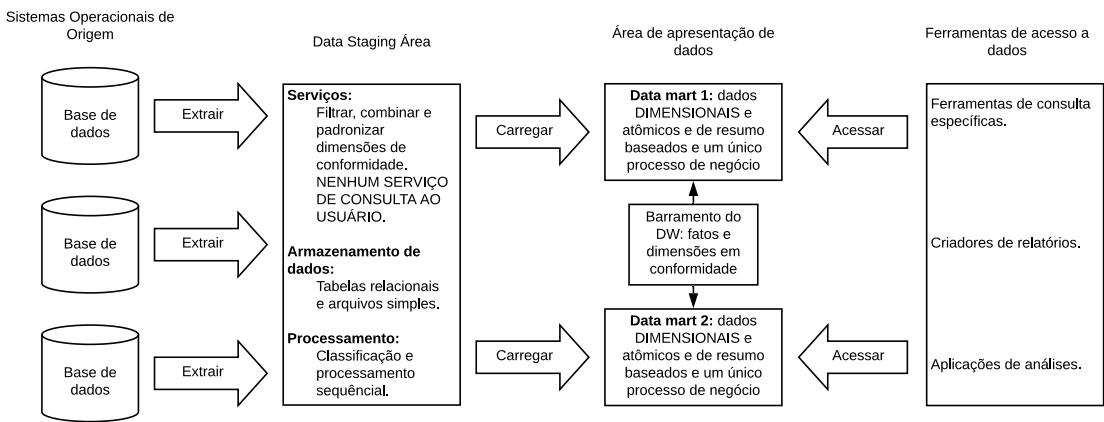
Apesar de apresentarem a formas de implementação diferentes, o conceito de *data mart* continua sendo o mesmo, um subconjunto do *data warehouse* que possui informação mais restrita, mais específica de um dado setor ou uma parte da empresa.

### 2.1.3 Componentes de um *Data Warehouse*

Sabendo então que um *data mart* é um subconjunto de um *data warehouse*, todas as características que foram apresentadas para um DW também são válidas para um *data mart*, logo é necessário falar sobre os componentes que fazem parte dos mesmos. Cada um dos componentes tem uma função específica. A Figura 2.1 apresenta os quatro componentes que serão abordados, sistemas de dados operacionais (*operational source systems*), área intermediária dos dados (*data staging area*), área de apresentação dos dados(*data presentation area*), ferramentas de acesso aos dados (*data access tools*).

Os Sistemas de dados operacionais referem-se aos sistemas OLTP da empresa, como já citado anteriormente, são sistemas responsáveis por registrar todas as transações da empresa, esse tipo de sistema é orientado à funcionalidade e sua prioridade é o alto desempenho. Um *data warehouse* visa unificar os dados de uma empresa e, logo o sistema OLTP tem a matéria prima para este processo.

A Área intermediária dos dados (*data staging area*), é tanto uma área de armazenamento e um conjunto de processos, os quais são: extração, transformação e carga dos dados (ETL sigla em inglês – *Extract, Transform, Load*) que consistem em um conjunto de operações sobre os dados. Como o nome sugere, a extração



**Figura 2.1.** Componentes de um data Warehouse. Adaptação da Fonte Santos [2010].

consiste em extraír dados de diversas fontes já que eles estão em diversos locais. A extração ocorre em tempos diferentes para cada base de dado, logo, é preciso esperar que seja feita a extração de todas as bases para que possa ser realizada a integração. Os dados recém extraídos ficam armazenados no *data staging* para que possam ser padronizados e depois manipulados. Ao final do processo de extração deve-se ocorrer a transformação dos dados. Nessa etapa, é necessário lidar com a falta de elementos, retirar redundâncias de dados, dentre outras transformações que podem ser exigidas para integração. Com o resultado obtido após a extração dos dados é realizado o carregamento dos mesmos no DW, tendo garantia que cada *data mart* receba os novos dados correspondentes a eles.

Área de apresentação dos dados é a parte onde os dados estão armazenados sendo que devem estar organizados de acordo com as regras de negócio. Essa área é a parte que os usuários têm acesso direto para realizar suas consultas.

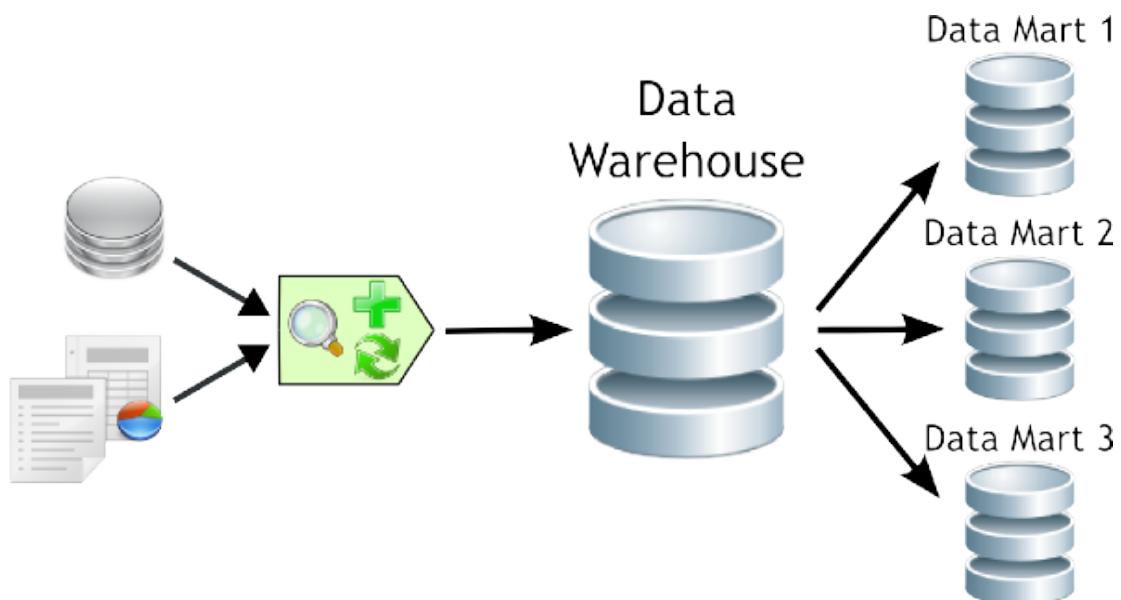
Por fim, temos a Ferramenta de Acesso aos Dados. Esse componente do *data mart* refere-se às ferramentas que permitem aos usuários utilizarem a área de apresentação dos dados para obter informações para tomar decisões. Exemplos de ferramentas que podem ser utilizadas são: Consultas *ad hoc*, *Data mining*, *Cockpits Digitais*, Ferramentas de transmissão, *dashboards*. Este componente é uma interface amigável com os usuários, onde ele pode visualizar os dados em formas, de

gráficos, planilhas.

### 2.1.4 Tipos de implementação de DW

É importante a análise de qual implementação deve ser feita antes de continuar o projeto de um DW, pois uma implementação mal escolhida pode causar impactos negativos ao longo do desenvolvimento. Existem três tipos de implementação, a saber, *top down*, *bottom up* e a combinada. **Implementação *top down*** (Figura 2.2), consistem em analisar as regras de negócio da empresa como um todo, analisando quais fontes de dados serão utilizadas, definindo a padronização dos dados entre as fontes, quais regras de negócio são interessantes para cada departamento [Rodrigues, 2004].

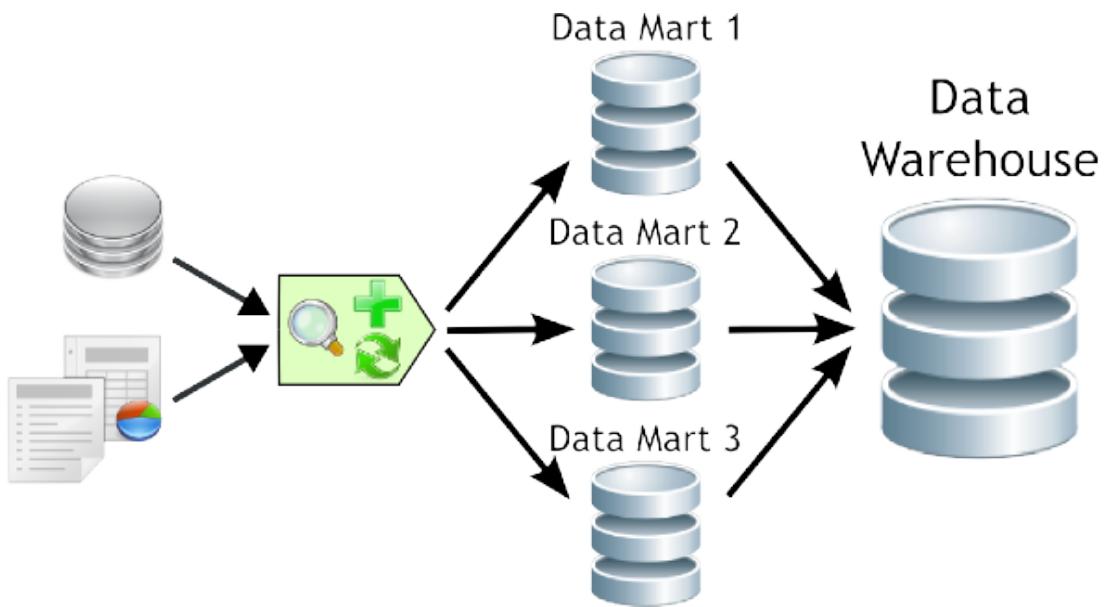
Como característica dessa implementação temos que os *data marts* são derivados do *data warehouse*, logo, eles possuem a mesma arquitetura e formato dos dados, o que facilita a manutenção.



**Figura 2.2.** Implementação *Top Down*. Fonte: <http://www.dataprix.net/pt-pt/24-data-mart>.

A **Implementação *bottom up*** (Figura 2.3) permite a criação dos *data marts*

antes do *data warehouse*, permitindo explorar características de desenvolvimento rápido com menores custos da arquitetura de *data marts* independentes. Apesar de ser mais simples, ao longo prazo, essa implementação pode atrasar o projeto, pois muitos *data marts* podem não estar padronizados, gerando trabalhos extras para unir os no *data warehouse*.



**Figura 2.3.** Implementação *Bottom Up*. Fonte: <http://www.dataprix.net/pt-pt/24-data-mart>.

Por fim, a Implementação combinada propõe, como o nome sugere, a combinação das duas abordagens, a fim de utilizar as vantagens de uma para suprir as deficiências da outra. Ela busca garantir a consistência e integração dos dados da abordagem *top down* com o rápido desenvolvimento de incremento da abordagem *bottom up*.

### 2.1.5 Modelo dos dados

Kimball [2002], cita que o modelo entidade relacionamento é uma técnica de modelo lógico que visa eliminar todas as redundâncias dos dados. É um modelo utilizado em OLTP, já que permite um grande desempenho na atualização de registro e

evita redundância nos dados, mas que dificulta a navegação sobre os dados e sua compreensão. Essa dificuldade advém do fato de que em uma empresa podem existir centenas de entidades que ao serem colocadas no banco de dados vão se tornar centenas de tabelas, dificultando o entendimento para usuários leigos no domínio da aplicação. Para solucionar esse problema, é necessário utilizar técnicas de modelagem de dados na construção de um *data warehouse*, criando-se modelos que apresentarão os dados de forma simples.

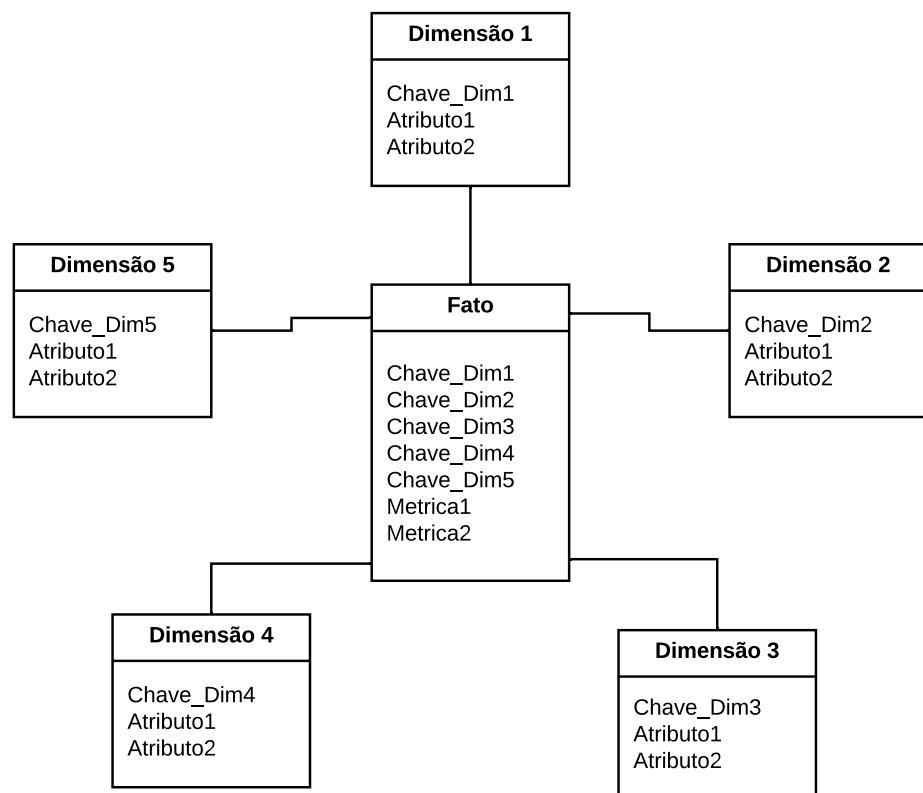
Um modelo multidimensional é constituído de tabela fato que é a principal tabela do modelo, nela são armazenadas as transações ou eventos, as chaves para as características correspondentes nas tabelas dimensionais e as medidas/métricas que permitem medir o desempenho do negócio, elas são atributos numéricos que permitem operações como adição subtração e media. Dimensões são as características de cada fato, representando a forma de visualização dos dados. As dimensões, normalmente, não apresentam valores numéricos, pois elas descrevem os elementos que participam de um fato. Temos também as medidas que são os valores numéricos da tabela de fatos. Por fim, temos a hierarquia conceitual que é a forma que os dados estão organizados para possibilitar a navegação por vários pontos de vista, analisando dados tanto resumidos quanto detalhados.

Ao se criar um modelo multidimensional é necessária a escolha do que mais combine com o projeto, pois através dele serão feitas as análises. A seguir serão apresentados os modelos estrela e floco de neve.

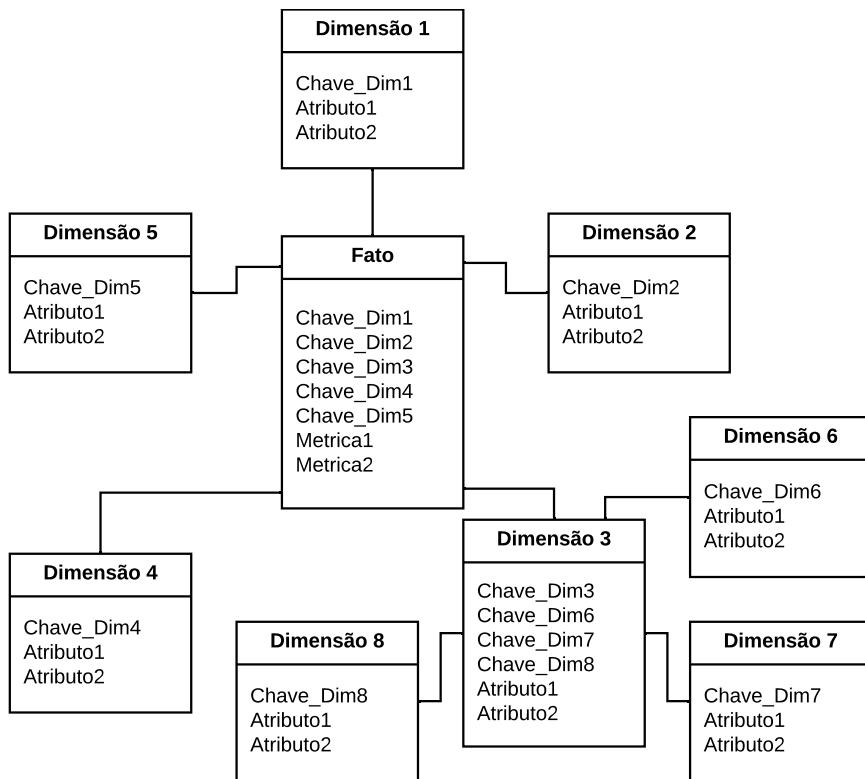
**Esquema estrela (*star schema*)** O modelo estrela (*star schema*) organiza os dados de forma que, o fato de análise esteja centralizado, sendo conectado as possíveis dimensões de análise, como pode ser visto na Figura 2.4. Uma das características desse modelo é que os dados são desnormalizados a fim de evitar junções entre tabelas, o que reduz o tempo de consultas. Entretanto, como o modelo não está normalizado, é gasto um espaço maior que em um modelo normalizado. Mesmo com um gasto de memória secundária maior, uma grande vantagem para o *data warehouse* é a eficiência nas consultas [Santos, 2010].

O modelo floco de neve *snowflake* é uma extensão do modelo estrela. Nele cada uma das pontas do modelo estrela é normalizada, o que faz com que cada uma das pontas da estrela se torne uma nova estrela, algumas das vantagens de se utilizar

este modelo são a redução de espaço e aceleração na atualização dos dados. Isso reduz o tempo de consulta dos dados, logo, a escolha entre estes dois modelos deve ser feita após o levantamento da quantidade de dados que serão analisados. A Figura 2.5 mostra um exemplo de floco de neve.



**Figura 2.4.** Modelo Estrela.

**Figura 2.5.** Modelo Floco de Neve

Por fim, a Tabela 2.2 faz uma comparação entre os 2 modelos

	<b>Modelo Estrela</b>	<b>Modelo Floco de Neve</b>
Tabela dimensão	Não normalizada	Normalizada
Tamanho físico	Grande volume já que os dados se repetem nas tabelas dimensões não normalizadas	Volume reduzido, já que os dados das tabelas dimensões são normalizados para evitar repetições
Velocidade das consultas	Rápida	Mais lenta do que o modelo estrela devido à normalização

**Tabela 2.2.** Tabela com o comparativo entre o Modelo Estrela e o Modelo Floco de Neve Castro Novais [2012]

## 2.2 OLAP

OLAP é uma ferramenta de apoio à decisão que permite a análise multidimensional de dados em sistemas on-line auxiliando a empresa a tomar decisões mais eficientes e eficazes [Politano, 2006].

A análise de dados é uma área de suma importância nas empresas e, por isso, deve-se escolher cuidadosamente a ferramenta OLAP a ser utilizada, pois segundo Santos [2010], a escolha de um produto inadequado leva a diversas consequências, por exemplo: prejuízos financeiros para aquisição de *software*, prejuízo no treinamento de pessoas para uso da ferramenta, falhas no projeto acarretando perda de credibilidade para sua conclusão.

Para a escolha de um bom software para a empresa, existem regras que podem ser utilizadas que, segundo Politano [2006] são: visão conceitual multidimensional; transparência; acessibilidade; desempenho consistente do relatório ; arquitetura cliente-servidor; dimensionalidade genérica; manuseio dinâmico da estrutura da matriz; apoio a multiusuários; operações irrestritas de cruzamento de dimensões; manipulação de dados intuitiva; relatório flexível; dimensões e agregação de níveis ilimitados.

Essas ferramentas permitem realizar operações sobre o cubo multidimensional. Kimball [2002] cita que as mais usadas são: *Drill down*, *Drill up ou Roll Up*, *Drill Across*, *Slice and Dice*, *Pivoting*.

Segundo Cramer [2006] ***Drill down*** é uma técnica analítica específica através da qual o usuário navega entre níveis de abrangência de dados a partir do mais resumido(*up*) para o mais detalhado(*down*). Os caminhos de navegação podem ser definidos pelas hierarquias dentro de dimensões ou outros relacionamentos que podem ser dinâmicos dentro ou entre dimensões. Por exemplo, na visualização de dados de vendas da América do Sul, uma operação *drill down* na dimensão região mostraria Brasil e Argentina. Um *drill down* sobre o Brasil poderia mostrar São Paulo, Rio de Janeiro e Minas Gerais.

Já a operação ***Drill up*** ou ***Roll Up*** é o caminho oposto do *drill down*, ou seja, parte de uma informação detalhada e a sumariza. Na ***Drill Across***, ao olhar múltiplas dimensões é possível aumentar o nível de detalhes dos dados. Um exemplo seria selecionar Brasil e mostrar São Paulo, Rio de Janeiro em relação às

outras dimensões.

O ***Slice and Dice*** são duas operações. *Slice* que é a seleção de partes de um cubo, ou seja, é a redução do escopo dos dados de análise, pode ser vista como um filtro, porém não modifica a perspectiva de visualização dos dados. Já a operação *dice* é a extração de um subcubo de duas ou mais dimensões [Santos, 2010].

***Pivoting*** é uma forma rápida de ver os dados de uma maneira diferente ou seja, é uma operação de inversão dos eixos do cubo para visualização dos resultados de uma consulta. Corresponde à mudança de posição das dimensões em um gráfico ou a troca de linhas pelas colunas em uma tabela [Santos, 2010].

Além das operações possíveis em um cubo multidimensional, é necessário também definir como os dados serão armazenados fisicamente, pois tal modo pode influenciar no tempo de recuperação de tais dados. Santos [2010] define as principais formas de armazenamento como:

1. MOLAP (MOLAP sigla em inglês – *Multidimensional On-Line Analytical Processing*): Nesse tipo armazenamento os dados residem num ambiente que usa tecnologia multidimensional. Esse método tem custo elevado para fazer o carregamento dos dados, pois é necessário realizar uma série de cálculos para agregar os dados às dimensões.
2. ROLAP (ROLAP sigla em inglês – *Relational On-Line Analytical Processing*): Nesse tipo de armazenamento os dados residem em um ambiente relacional, mas são modelados de maneira dimensional. Ao realizar as consultas dimensionais os dados são recuperados e processados pelo servidor de banco de dados relacional.
3. HOLAP (HOLAP sigla em inglês – *Hybrid On-Line Analytical Processing*): É uma combinação dos métodos ROLAP e o MOLAP. Nesse armazenamento é possível usufruir o que há de melhor nos métodos de armazenamento, podendo assim ter um alto desempenho em consultas através do MOLAP e obter escalabilidade do ROLAP. Dessa maneira, os dados são armazenados fisicamente no modelo relacional, mas as agregações que normalmente gastam tempo elevado de processamento e ficam pré-computadas em estruturas multidimensionais.

Nas próximas seções serão mostrados os trabalhos relacionados, possíveis ferramentas OLAP para se fazer a análise dos dados em um projeto de *data warehouse*.

## 2.3 Trabalhos Relacionados

A seguir serão discutidos trabalhos voltados a área da educação, em que o objetivo dos mesmos é implementar uma estratégia de obtenção de informação sobre os dados armazenados nas instituições de ensino.

Camargo [2013] fez uma análise do que seria necessário para fazer uma implementação de um *data mart* na UNIPAMPA, ele definiu que a melhor abordagem a se utilizar na universidade é uma implementação *bottom up* a qual utilizaria a técnica ROLAP para o armazenamento de dados, selecionando o MySQL como o gerenciador de banco de dados. Durante a criação do modelo físico é utilizado o esquema estrela e, para manipular os dados via uma ferramenta OLAP, foi selecionado o *SPagoBI* por ser uma ferramenta gratuita. Para verificar a qualidade das informações obtidas pela implementação foi utilizado os dados referentes a vida escolar dos alunos das instituições, alguns exemplos de análises obtidas são: as formas de matrícula pelos anos nos *Campi*, quantidade de reprovações.

O trabalho de Santos [2010] realiza a implementação de um *data mart* do questionário sócio-cultural dos candidatos da UESB a fim de analisar o perfil dos mesmos, já que havia sido observado que os dados referentes a este questionário não estavam disponíveis para obtenção de informação dos gestores. Como já havia um DW criado na instituição, optou-se como a melhor abordagem a implementação utilizando a técnica *bottom up*, implementado utilizando a técnica ROLAP com o SGBD MySQL. O modelo físico utilizado foi implementado utilizando o esquema estrela já que ele supria as necessidades para a análise. Assim, foram feitos três cubos, os quais são: Perfil-econômico, Perfil-social e Perfil-cultural. Foi também utilizado a ferramenta OLAP *Pentaho* para se fazer as análises sobre os cubos, estas que por sua vez, são parecidas com as análises que serão realizadas nesta monografia, já que ambos visam a implementação de um *data mart* do questionário aplicados aos candidatos à vagas das instituições.

O autor Steidel [2017] implementou uma técnica de BI para analisar a evasão escolar na UFRN. Para que este processo ocorra, foi feito a extração dos dados

do sistema SIG da UFRN. Como metodologia de desenvolvimento foi utilizado uma adaptação da técnica *Cross-Industry Standard Process for Data Mining*, que é normalmente utilizada na área de mineração de dados, como a implementação desta técnica é uma adaptação, foi necessário a utilização de uma ferramenta OLAP para manipular os dados, para esta tarefa foi selecionada a ferramenta *QlikView* por ser gratuita e já ter sido utilizada em projetos anteriores. Outra característica *QlikView* que fez optar por sua escolha foi a capacidade de não depender modelagem de dados e construção de cubos OLAP, ao invés disso ele cria um grande arquivo com todas as associações entre os dados, conhecido como *Data Cloud* ou nuvem de dado. Uma desvantagem desta abordagem é que os dados ficam em memória RAM fazendo com que análise de grandes volumes de dados seja limitada ao tamanho da memória. Algumas análises realizadas foram: número de evasões por disciplina, percentual de evasão por curso e evasão por ano. Após a realização das análises foi encaminhado os resultados e um questionário a gestores de algumas instituições de ensino a fim de definir o grau de relevância do trabalho. No *feedback* obtido o trabalho foi bem aceito, sendo que 94,93% dos pesquisados responderam que o uso da aplicação de BI traria uma redução de esforços no momento da tomada de decisão.

O trabalho de Damasceno [2017] utiliza a metodologia *Action Research*, que visa resolver um problema imediato, este método é constituído de: diagnóstico, plano de ação, ação e avaliação. Na etapa de diagnóstico identifica-se o problema. O plano de ação define as características e os passos a serem tomados para a resolução do problema. A ação é pôr em prática o plano elaborado. Avaliação é validar se tudo foi feito de acordo com o plano elaborado. Durante a fase de diagnóstico foi identificado que o sistema utilizado no IFTM (Virtual-IF) - *Campi Paracatu*, não possuía uma ferramenta de gestão empresarial, logo era necessário implementar uma ferramenta de BI que solucionasse esse problema. Durante o plano de ação definiu-se o que era necessário para os gestores, bem como a forma de implementação abordada, *bottom up*, esta implementação foi escolhida por se tratar de um subconjunto dos dados da instituição, o subconjunto pertencente ao sistema Virtual-IF que tratava do controle dos alunos (cadastro da matriz curricular, cadastro das disciplinas, cadastro da grade de horários, matrícula dos alunos ingressantes, rematrícula dos alunos ao iniciar o semestre letivo, parametrização

das notas dos alunos), assim devia-se implementar um *data mart* que atendesse a demanda do Controle de Registro Acadêmico (CRA). Apesar de não ter sido explicitado esse sistema foi implementado utilizando a técnica ROLAP, já que implementaram utilizando o SGBD MySQL que já era integrado ao sistema Virtual-IF da instituição. Para melhorar o acesso aos dados, foi abordado o esquema estrela, sendo criado quatro cubos, cubo fato frequência, cubo fato situação aluno, cubo fato nota consta e cubo fato pendência. Durante a fase de ação foi feito a escolha da ferramenta OLAP *Pentaho*, para se fazer a análise e geração de relatórios. Foram realizadas análises sobre os dados e apresentadas aos gestores, os quais deram um *feedback* positivo. Um problema deste trabalho, foi a não apresentação das análises feitas com a ferramenta escolhida.

## 2.4 Ferramentas

O estudo das ferramentas OLAP para a análise dos dados referentes ao questionário é importante pois uma ferramenta que não atendas as necessidades da instituição podem acarretar em uma análise dos dados ineficiente, fazendo com que gestores deixem de utilizá-la em suas tomadas de decisões. Para que isto não ocorra foi feito um levantamento de quais poderiam ser utilizadas. A Tabela 2.3 descreve sobre algumas características das ferramentas, estas são: licença ou seja se ela é uma ferramenta paga ou gratuita, dependência da tecnologia para formas de armazenamento, como ela é utilizada ou seja se ela é *online* ou *offline* e o que pode ser feito com os dados nela.

Com uma análise inicial das ferramentas obtidas, pode-se selecionar um subconjunto das ferramentas listadas para a realização do trabalho, tendo eliminado as ferramentas: BIRT por possuir poucas maneiras de visualização, *Tableau*, *JasperSoft* e *Redash* por serem ferramentas que requisitam licença, *OpenReports* por ser apenas uma ferramenta de geração de relatórios. Assim o subconjunto de ferramentas a ser testado foi: *Apache Superset*, *QlikView*, *SpagoBI* e *Vanilla*.

Ferramenta	Licença	Armazenamento	Manipulação	Visualização
<i>Apache Superset</i>	Livre	SQL, <i>Big Data</i> , Arquivos	<i>Online</i>	Visualização de dados, geração de dashboards.
BIRT	Livre	SQL, <i>Big Data</i> , Arquivos	<i>Offline</i>	Visualização de dados, criação de relatórios.
<i>JasperSoft</i>	Pago	SQL, NoSQL, <i>Big Data</i> , Arquivos	<i>Online</i>	Visualização de dados, geração de dashboards, criação de relatórios.
<i>OpenReports</i>	Livre	Arquivos	<i>Online</i>	Geração de relatórios.
<i>QlikView</i>	Livre	SQL, NoSQL, <i>Big Data</i> , Arquivos	<i>Offline</i>	Visualização de dados, geração de dashboards.
<i>Redash</i>	Pago	SQL, NoSQL, <i>Big Data</i>	<i>Online</i>	Visualização de dados, geração de dashboards.
<i>SpagoBI</i>	Livre	SQL, <i>Big Data</i>	<i>Online</i>	Visualização de dados, geração de dashboards, cockpits.
<i>Tableau</i>	Pago	SQL, <i>Big Data</i> , Arquivos	<i>Online</i>	Visualização de dados, geração de dashboards.
<i>Vanilla</i>	Livre	SQL, NoSQL, <i>Big Data</i> , Arquivos	<i>Online</i>	Visualização de dados, geração de dashboards.

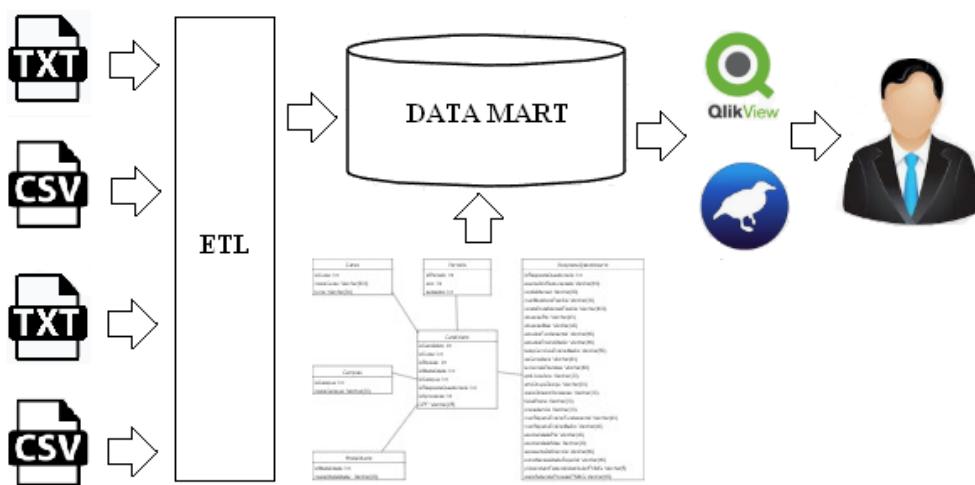
**Tabela 2.3.** Tabela com breve descrição das possíveis ferramentas OLAP a serem utilizadas.



# Capítulo 3

## *Data mart* Acadêmico IFNMG: Processo Seletivo

Neste capítulo, será descrito as atividades realizadas para construção do *data mart* acadêmico do IFNMG, sendo que a principal necessidade de informação é sobre o perfil dos candidatos à vagas dos processos seletivos nas diversas modalidades de ensino ofertadas pelo IFNMG. Assim, a Figura 3.1 apresenta o fluxo conceitual das atividades necessárias para construção do ambiente analítico para auxiliar os gestores do IFNMG a identificarem o perfil dos candidatos à vaga no processo seletivo.



**Figura 3.1.** Fluxo de definição do *Data Mart*.

Neste contexto, o presente capítulo está dividido da seguinte forma: A Seção 3.1 apresenta a etapa inicial do projeto; Seção 3.2 mostra como os dados foram obtidos dos ambientes operacionais da instituição; Seção 3.3 apresenta o modelo físico de armazenamento dos dados no *data mart*; Seção 3.4 descreve o processo de extração-transformação e carga dos dados realizado neste trabalho; Seção 3.5 apresenta o modelo lógico (cubos dimensionais de dados); Por fim, a Seção 3.6 apresenta as tecnologias utilizadas para análise e apresentação dos dados aos gestores da instituição.

### 3.1 Levantamento dos Requisitos de Informação

Uma das principais etapas de um projeto de *data mart* é a compreensão de quais informações os tomadores de decisão esperam do projeto. Inicialmente, foi realizado um levantamento sobre quais questões são abordadas nos questionários socioeconômicos aplicados aos candidatos a vagas do IFNMG. A análise sobre os questionários revelou que não houve mudanças de questões nos diferentes processos seletivos ao longo dos anos. Todas as questões são do tipo objetiva e de múltipla escolha, a Tabela 3.1 ilustra as questões que fazem parte do questionário socioeconômico utilizado pelo IFNMG.

Após a análise do domínio dos dados (questionário), foi necessário eliciar os requisitos. A estratégia utilizada à elicitação foi a de entrevistas com os principais gestores do departamento de ensino e representantes de profissionais que contribuem para as atividades de ensino no campus Montes Claros, sendo que, as entrevistas foram feitas em três reuniões, sendo entrevistados os seguintes gestores: o Diretor Geral, o Coordenador de Ensino, o Diretor do Departamento de Ensino, o Coordenador de Pesquisa; Psicóloga, Assistente de Alunos; Coordenador de Extensão.

Após o período de entrevistas, foi possível identificar as reais necessidades de informação dos usuários, sendo que as mesmas foram agrupadas em três grandes áreas: Perfil Social, Perfil Econômico e Perfil Educacional.

Tais áreas/assunto utilizados para as definições dos cubos dimensionais de análise (Seção 3.5) que estarão disponíveis nas ferramentas de consulta analítica.

Nº	Pergunta
1	Você exerce alguma atividade remunerada?
2	Qual é sua renda mensal?
3	Qual é o número de membros da sua família?
4	Qual é a renda mensal de sua família?
5	Qual das seguintes alternativas melhor expressa a atual situação de seu pai no trabalho?
6	Qual das seguintes alternativas melhor expressa a atual situação de sua mãe no trabalho?
7	Como você realizou seus estudos de Ensino Fundamental ou equivalente?
8	Como você realizou seus estudos de Ensino Médio ou equivalente?
9	Há quanto tempo você concluiu o Ensino Médio?
10	Você se considera:
11	Você tem hábito de ler jornais ou revistas?
12	Excetuando os livros escolares, quantos livros você lê por ano?
13	Com qual das atividades citadas abaixo você ocupa mais tempo?
14	Qual é o meio que você mais utiliza para se manter informado sobre os acontecimentos atuais?
15	Indique a sua Faixa Etária:
16	Qual a sua procedência?
17	Se você repetiu alguma série do Ensino Fundamental, informe o número de vezes
18	Se você repetiu alguma série do Ensino Médio, informe o número de vezes:
19	O Grau de Escolaridade do seu pai é:
20	O Grau de Escolaridade da sua mãe é:
21	Você apresenta algum tipo de deficiência?
22	Você tem outra necessidade especial?
23	É a 1ª vez que participa do Processo Seletivo/Vestibular do IFNMG?
24	Por qual principal meio ficou sabendo do Processo Seletivo do IFNMG?

**Tabela 3.1.** Perguntas do Questionário Socioeconômico.

## 3.2 Obtenção dos Dados Operacionais

A viabilidade de um projeto de DW depende, principalmente, da obtenção e da qualidade dos dados que estão armazenados nos sistemas OLTP. Desta forma, uma tarefa importante deste projeto foi recuperar os dados referentes às respostas dos candidatos aos questionários socioeconômicos aplicados em cada um dos processos seletivos do IFNMG. Os dados dos candidatos às vagas dos processos seletivos

ficam sob responsabilidade da Pró-Reitoria de Ensino do IFNMG, sendo que a Comissão Permanente de Concurso (COPEC) trata diretamente com as empresas terceirizadas que realizam os processos seletivos. Neste contexto, o IFNMG apenas possui versões sumarizadas dos dados dos anos de 2013 a 2019 e, deste modo, iniciou-se um processo de recuperação dos dados com as empresas que realizaram os supramencionados processos seletivos.

Como os processos seletivos acontecem duas vezes por ano, o contrato de prestação de serviço é realizado por ano, assim, uma empresa que realizou o processo seletivo de 2013, pode não ser a mesma que realizou o processo de 2014, porém pode ser aquela que realizou o processo do ano de 2015 dificultando a tarefa de rastreamento e obtenção dos dados. Outro fator dificultou a obtenção dos dados, os termos de referência de prestação de serviço dos processos seletivos entre o IFNMG e as empresas, pois não constavam que as empresas contratadas deveriam disponibilizar/retornar uma cópia dos dados ao IFNMG. Como este é o primeiro trabalho científico aplicado que requisitou o acesso aos dados dos questionários, houveram diversos atrasos, questões legais tratadas pela reitoria do IFNMG para que os dados fossem disponibilizados para a realização deste trabalho.

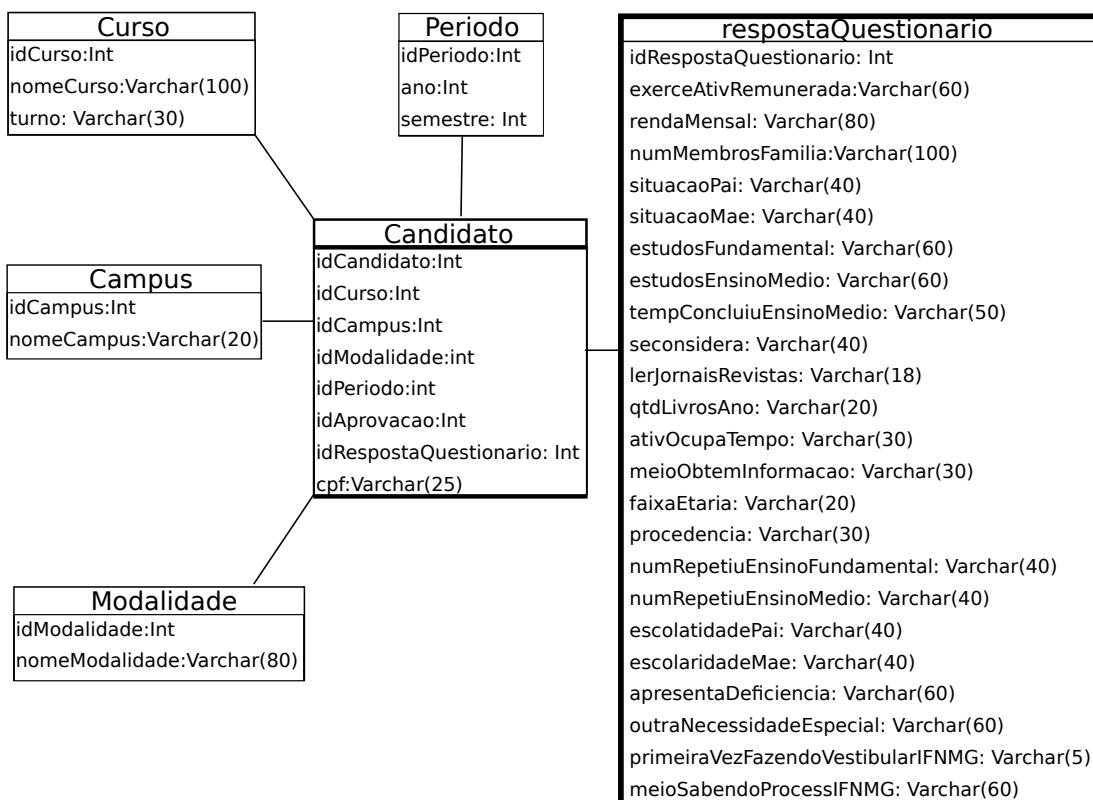
Os dados dos seguinte anos e semestres foram recuperados: 2013.1, 2013.2, 2014.1, 2015.2, 2016.1, 2016.2, 2019.1, 2019.2. Pode-se perceber que faltam dados do segundo semestre de 2014, do primeiro semestre de 2015, do primeiro e no segundo semestre de 2017 e do primeiro e do segundo semestre de 2018. As empresas que realizaram tais processos seletivos dos dados faltantes, não responderam ou, simplesmente, ignoraram as requisições legais que foram impostas. Deste modo, as análises referentes a evolução do perfil do candidato a vaga ao longo do tempo foram prejudicadas devido às lacunas temporais dos dados.

Após os contratemplos causados pelas requisições dos dados, foi obtido uma parcela dos mesmos. Além da demora para a obtenção os dados não continham um padrão bem definido, já que vinham de empresas distintas, os mesmos vieram em arquivos de texto separados por ano, período e modalidade, gerando então um grande volume de arquivos. Outros foram obtidos em arquivos no formato csv, que em cada coluna corresponde a uma resposta. Também houveram dados que tiveram que ser devolvidos, pois as respostas estavam sumarizadas, cada pergunta continha apenas o total de respostas em cada alternativa, tornando impossível a

definição do perfil do candidato, após a devolução dos mesmos foi feito uma nova solicitação, pedindo junto com ela, a entrega dos mesmos seguindo um padrão.

### 3.3 Modelo Físico do *Data mart*

Para a implementação do *data mart* foi escolhida abordagem ROLAP, que permite fazer o carregamento dos dados em um SGBD com tecnologia relacional, reduzindo o impacto de utilização de tecnologias dimensionais no armazenamento dos dados. O SGBD definido para utilização deste projeto foi o *MySQL 5.5*.



**Figura 3.2.** Código de exemplo da padronização do nome do campus.

Contudo, mesmo utilizando de tecnologia baseada no modelo relacional, a modelagem dos dados foi realizada considerando eficiência para extração das informações desejadas. Sendo assim, foi utilizado o esquema estrela para modelar como

os dados serão armazenados, já que este modelo facilita o entendimento dos dados por todas as pessoas envolvidas nas tarefas de análise. Além disso, o modelo possibilita que consultas complexas possam ser executadas de maneira eficiente utilizando por meio de SQL padrão. A Figura 3.2 é a representação do modelo físico dos dados.

O modelo criado tem como tabela fato o candidato, suas dimensões são compostas pelas perguntas do questionário e informações a respeito da escolha do candidato, tais como modalidade do curso que ele pretende prestar o processo seletivo, o curso, o período e o campus ao qual ele está disputando a vaga.

### 3.4 Extração, Transformação e Carga dos dados

O processo de extração, transformação e carga dos dados foi realizado por *scripts* implementados na linguagem JAVA.

Os *scripts* criados tem como função extrair os dados dos arquivos de formatos .txt e .csv que foram obtidos durante a fase de obtenção dos dados (Seção 3.2). Com dados extraídos das diferentes fontes de dados, foi necessário também padronizar os formatos dos dados bem como prepará-los para inserção no novo modelo estrela que persiste os dados no *data mart* (Seção 3.3).

Durante a etapa de extração foram criados dois *scripts* já que cada tipo de arquivo possui uma regra de extração distinta. O primeiro *script* refere-se a extração dos dados dos arquivos de texto e o segundo refere-se a extração dos dados dos arquivos csv.

Os dados dos arquivos de texto estavam organizados em vários arquivos distintos, seguindo dois padrões:

1. Padrão 1 - Modalidade. Os arquivos deste padrão estavam separados pela modalidade e pelo ano que os candidatos estavam prestando o processo seletivo. Cada linha é composta por um conjunto de atributos que representam um candidato, estes atributos são:

- Um número que é o identificador do candidato.
- Um número que é o identificador do vestibular.

- O CPF do candidato.
  - O nome do candidato.
  - O nome do curso que o candidato está prestando processo seletivo.
  - O nome da modalidade que o candidato está prestando processo seletivo.
  - O turno do curso que o candidato está prestando processo seletivo.
  - O nome do campus que o candidato está tentando ingressar.
2. Padrão 2 - Respostas dos candidatos. Os arquivos deste padrão estavam separados pela modalidade e pelo ano, em que cada linha é a resposta de uma pergunta de um candidato, logo o questionário de um candidato é composto por 24 linhas.
- Um número que é o identificador do candidato.
  - Um número que é o identificador do vestibular.
  - O número da pergunta.
  - Um número que é o identificador da resposta que o candidato escolheu para a dada pergunta.

Para se fazer o processo de extração destes arquivos foi necessário fazer a junção dos dois arquivos, gerando apenas uma lista contendo os dados do candidato, e as respectivas respostas de cada pergunta.

O segundo *script* referente aos dados dos arquivos csv, sendo que os dados nos arquivos estavam em padrão único:

1. Os arquivos representam os dados de cada candidato em apenas uma linha da tabela. Esta linha era dividida em 28 colunas:
  - Um número que é o identificador do candidato.
  - O nome do candidato.
  - O CPF do candidato.
  - O nome do curso, o nome do campus e o turno que o candidato está prestando vestibular.

- As próximas 24 colunas correspondem em ordem crescente ao número da questão, sendo utilizado um identificador correspondente a resposta do candidato.

Durante a execução do segundo *script* teve-se que separar a coluna referente a informação do curso que o candidato estava realizando o processo seletivo, pois a mesma concatenou todas as informações sobre o campus e modalidade, assim, tal informação foi dividida da seguinte maneira: nome do curso, nome do campus e turno, além de utilizar estes três atributos para poder identificar o atributo modalidade.

O atributo turno precisou ser avaliado pois um curso pode ser categorizado como: "Diurno", "Integral", "Noturno", "Vespertino" ou "Matutino". Logo dependendo da instituição, os turnos "Diurno" e "Integral" podem apresentar o mesmo significado, assim foi averiguado o sentido de cada uma das nomenclaturas para o IFNMG, e foi identificado que, um curso é chamado de diurno pelo instituto quando ele é executado no turno "Matutino" ou no turno "Vespertino" tendo uma pequena parcela dele executada em outro turno, já Integral é quando o curso é executado durante o dia inteiro e não em apenas uma parcela de um turno, logo foi decidido manter os dois turnos.

Durante a etapa de transformação foi utilizado um arquivo de texto com as respostas do questionário padronizadas. Algumas respostas precisaram ser modificadas, pois apresentavam informações com pouco valor para tomada de decisão:

1. Qual a escolaridade da sua mãe?

- A alternativa "ensino médio incompleto" foi alterada para "ensino fundamental completo".
- A alternativa "ensino superior incompleto" foi alterado para "ensino médio completo".

2. Qual a escolaridade de seu pai?

- A alternativa ensino médio incompleto foi alterada para ensino fundamental completo.

- A alternativa ensino superior incompleto foi alterado para ensino médio completo.
3. Se repetiu alguma série do Ensino Médio, informe o número de vezes:
- A opção "apenas concluí ensino fundamental"foi alterada para "nem nenhuma vez".
4. Há quanto tempo você concluiu o Ensino Médio?
- A opção "ainda estou cursando o ensino médio"foi alterada para "apenas concluí o ensino fundamental".

```
if (candidatos[5].contains("montes")) {  
    candidatos[5] = "MONTES CLAROS";  
} else if (candidatos[5].contains("pir")) {  
    candidatos[5] = "PIRAPORA";  
} else if (candidatos[5].contains("sali")) {  
    candidatos[5] = "SALINAS";  
} else if (candidatos[5].contains("janu")) {  
    candidatos[5] = "JANUÁRIA";  
} else if (candidatos[5].contains("araç")) {  
    candidatos[5] = "ARAÇUAI";  
} else if (candidatos[5].contains("alme")) {  
    candidatos[5] = "ALMENARA";  
} else if (candidatos[5].contains("arin")) {  
    candidatos[5] = "ARINOS";  
} else if (candidatos[5].contains("diamã")) {  
    candidatos[5] = "DIAMANTINA";  
} else if (candidatos[5].contains("oto")) {  
    candidatos[5] = "TEÓFILO OTONI";  
} else if (candidatos[5].contains("jana")) {  
    candidatos[5] = "JANAÚBA";  
} else if (candidatos[5].contains("porte")) {  
    candidatos[5] = "PORTEIRINHA";  
}
```

**Figura 3.3.** Código de exemplo da padronização do nome do campus.

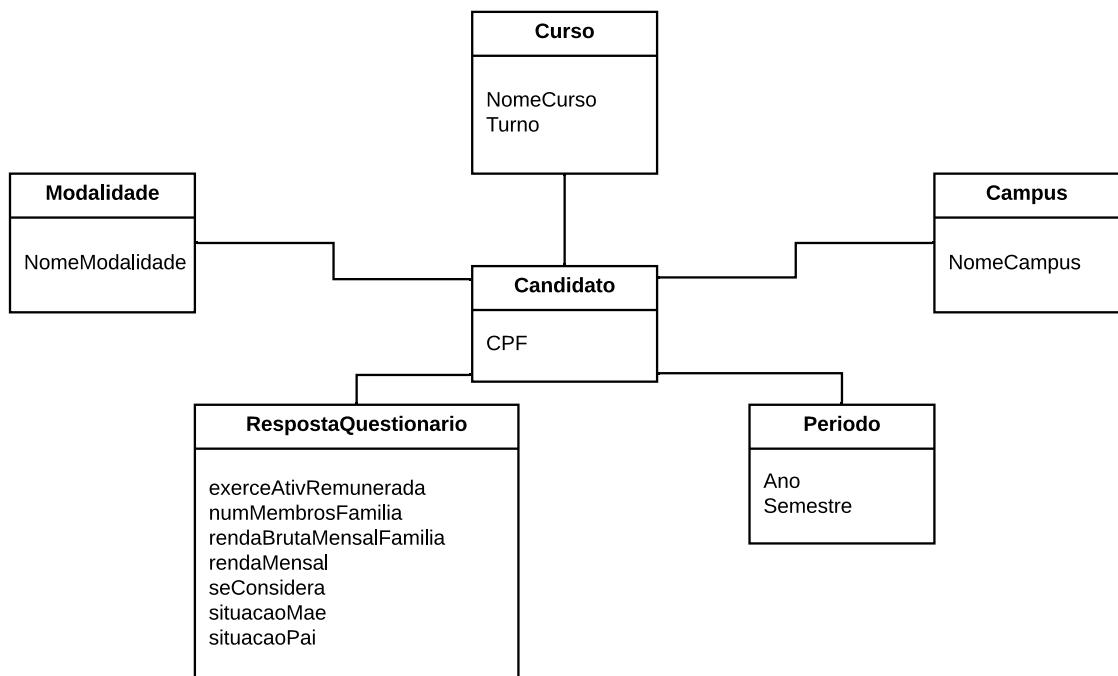
Após a padronização das respostas foi feito a troca do identificador da resposta pela resposta correspondente utilizando o arquivo de texto como índice. Outra parte da etapa de transformação foi a validação e padronização dos atributos: nome do campus, turno, validar CPF, nome do curso e o nome da modalidade. Para padronizar os atributos usou-se funções similares ao código 3.3, onde é verificado

a existência de uma parte da palavra que deveria compor o atributo e trocado o valor daquele atributo por um valor único.

Ao final do processo de transformação, os dados padronizados foram carregados no *data mart* criado (Seção 3.3).

### 3.5 Definição dos Cubos

Para possibilitar que ferramentas OLAP de exploração dos dados sejam utilizadas, foi necessário definir um modelo lógico de acesso aos dados (cubos de dados), permitindo visões multidimensionais, a criação de gráficos e de relatórios sobre as consultas. Os cubos são definidos a partir dos 3 grandes assuntos obtidos através das entrevistas com os gestores, assuntos que determinam o perfil do candidato socialmente, economicamente e educacionalmente.



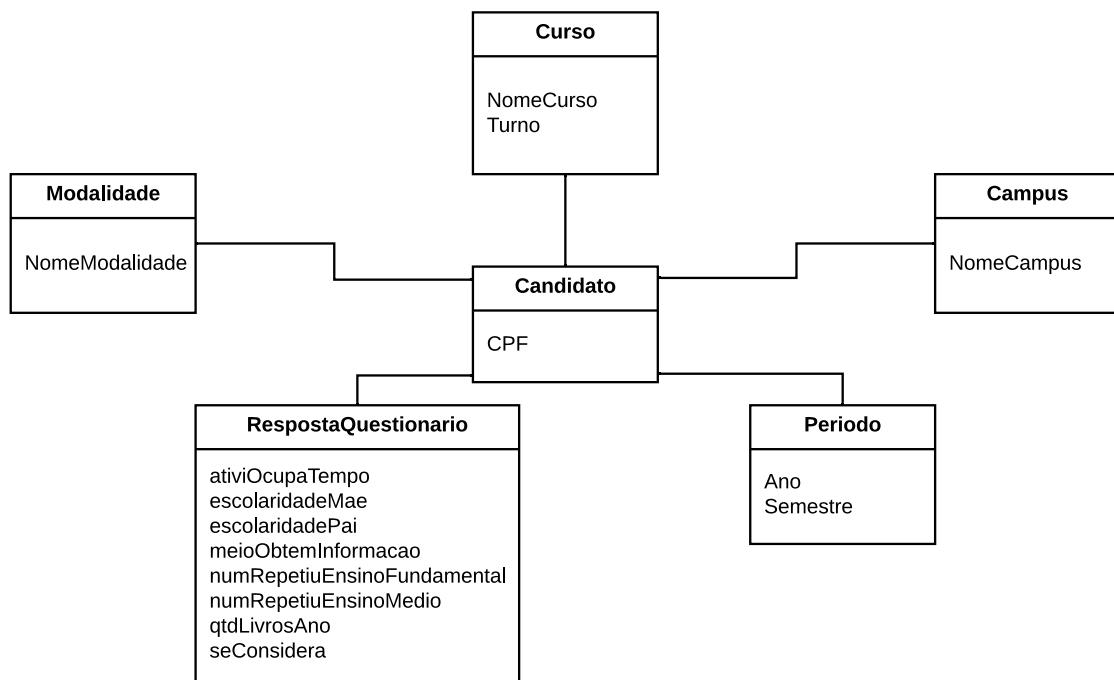
**Figura 3.4.** Representação do cubo Perfil-Econômico.

Cada cubo é composto por um subconjunto das respostas do questionário, uma tabela fato e um conjunto de tabelas dimensões que identificam as escolhas

do candidato para ingresso no IFNMG. Estas escolhas são: o nome do curso, o nome da modalidade, o período, o nome campus, o turno, o ano e o semestre.

O subconjunto de respostas do questionário de cada cubo, visa caracterizar o assunto de cada cubo. O cubo Perfil-Econômico visa definir o perfil do candidato em relação a sua vida econômica e de sua família, logo, este cubo é composto pelo conjunto respostas que estão relacionadas a situação financeira do candidato e de sua família como pode ser observado na dimensão RespostaQuestionario apresentada na Figura 3.5.

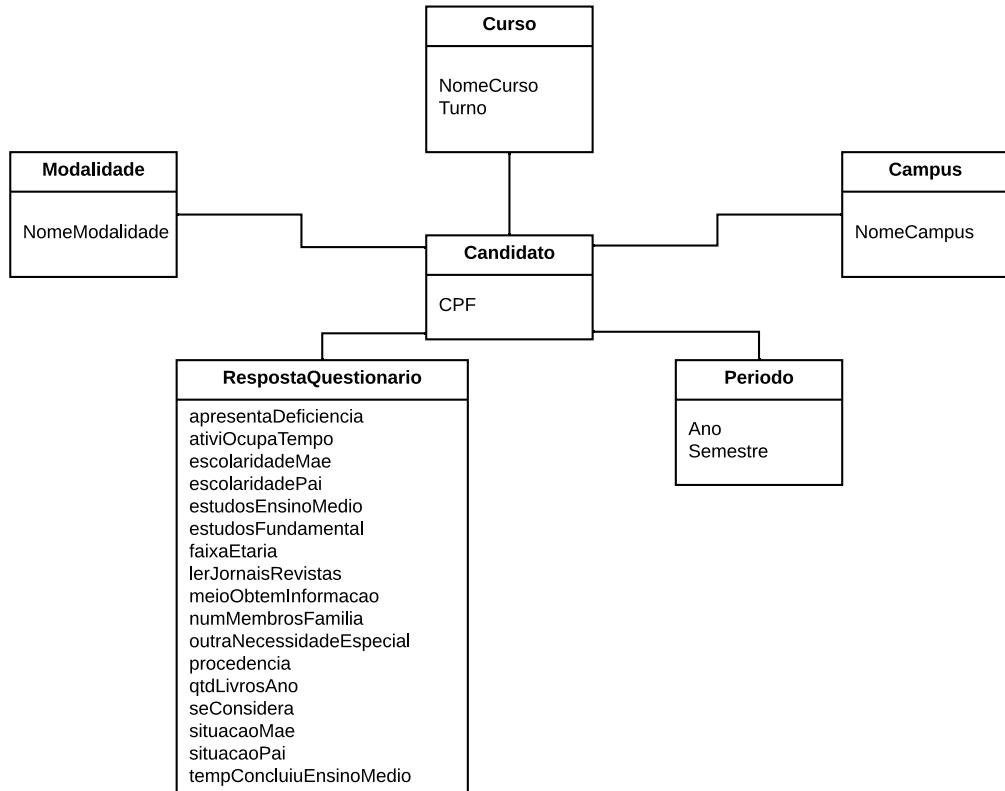
O cubo Perfil-Educacional, visa traçar o perfil dos candidatos em relação a sua vida escolar, neste cubo são armazenados o conjunto de respostas que visam identificar a situação escolar do candidato e de sua família, hábitos de leitura e meio de lazer. Na Figura 3.5 podemos observar a dimensão RespostaQuestionario que contém os atributos correspondentes a este perfil.



**Figura 3.5.** Representação do cubo Perfil-Educacional.

O cubo Perfil-Social, visa definir o perfil dos candidatos em relação a sua vida social, neste cubo são armazenados os dados que possam caracterizar o candidato

através da situação de sua família em relação aos estudos e como isso pode influenciar em sua vida, dados que mostram características do candidato, tais como: se ele apresenta deficiência, como ocupa o tempo, seu histórico educacional, tempo que concluiu o ensino médio, sua procedência. Na Figura 3.6 podemos observar a dimensão RespostaQuestionario que contém os atributos correspondentes a este perfil.



**Figura 3.6.** Representação do cubo Perfil-Social.

### 3.6 Tecnologias OLAP

A criação de um ambiente de *data warehouse* possibilitou a visualização dos dados sobre vários pontos de vista distintos, isso faz com que seja necessário o uso de ferramentas intuitivas para se extrair o perfil dos candidatos. Para tal tarefa foi feito um levantamento de ferramentas OLAP que proporcionam aos gestores, a integração com o *data mart*, fornecendo a eles informações com alto nível de

confiabilidade e segurança, tornando possível a tomada de decisões baseadas em informações. Contudo, existem muitas ferramentas disponíveis no mercado, fazendo com que seja necessário um estudo e escolha de qual ferramenta atende as necessidades de informação da instituição.

Na Seção 2.4 foi feito um levantamento de quais ferramentas poderiam ser utilizadas no instituto, através de uma análise prévia foram selecionadas algumas ferramentas para serem analisadas, estas são: *Apache Superset*, *QlikView*, *SpagoBI* e *Vanilla*.

Para se fazer o teste das ferramentas, foi criado um questionário a fim de guiar o processo de escolha das características de cada ferramenta por meio de testes operacionais.

A primeira característica a ser analisada nas ferramentas é a questão técnica, quais requisitos mínimos necessários para a instalação da ferramenta em um computador, pois como este é um trabalho de conclusão de curso, não há muito recurso computacional disponível. Além disso, ainda não há certeza do apoio institucional para implantar as soluções aqui descritas, assim, buscou-se minimizar os custos para aumentar a atratividade orçamentária. A Tabela 3.2 apresenta os resultados obtidos.

Recurso computacional	<i>Apache Superset</i>	<i>QlikView</i>	<i>SpagoBI</i>	<i>Vanilla</i>
Memória RAM	NULL	4GB	4GB	NULL
Sistema(s) operacional(ais) suportado(s)	<i>Linux</i>	<i>Windows</i>	<i>Windows, Linux</i>	<i>Windows, Linux</i>
Dependência(s) técnica(s)	<i>Python3.6, virtualenv</i>	Nenhuma	JDK 7	NULL

**Tabela 3.2.** Pré-requisitos computacionais para a instalação das ferramentas.

A máquina disponível para fazer os testes, contém 4GB de memória RAM, sistemas operacionais *Linux* e *Windows*, logo pode-se fazer o teste das ferramentas.

Após a verificação se era possível fazer os testes das ferramentas, foram analisados os documentos de suporte, instalação/manutenção a fim de averiguar o suporte que poderia ser obtido com os manuais das ferramentas. A Tabela 3.3 apresenta os resultados obtidos, sendo que cada item apresentará um valor entre "SIM" ou

"NÃO", onde que "SIM" é a existência da documentação e "NÃO" é a inexistência da documentação ou ela não pode ser localizada.

Clareza nos documentos	<i>Apache Superset</i>	<i>QlikView</i>	<i>SpagoBI</i>	<i>Vanilla</i>
Documentação de uso	SIM	SIM	SIM	NÃO
Documentação de instalação	SIM	SIM	SIM	NÃO
Documentação para manutenção	SIM	SIM	SIM	NÃO

**Tabela 3.3.** Características da documentação.

Ao analisar os resultados encontrados da documentação, pode-se perceber que para obter alguma informação técnica da ferramenta *Vanilla* ou um manual de usabilidade era necessário entrar em contato com a equipe de suporte. Logo esta ferramenta foi retirada do escopo de análise, por ser uma ferramenta totalmente dependente da empresa desenvolvedora.

O terceira aspecto a ser analisado é o aspecto de usabilidade das ferramentas: se ela é uma ferramenta que possibilita personalização de sua aparência, se ela não contém muita informação apresentada na tela, se as funcionalidades estão organizados a fim de ser intuitivo para quem for utilizar, se seus ícones são representativos e se existe suporte ao idioma português. A Tabela 3.4 apresenta as características de usabilidade das ferramentas onde as colunas são preenchidas com um número de 0 a 5 onde que 0 representa ausência da característica e 5 a ferramenta possui totalmente aquela característica.

Características	<i>Apache Superset</i>	<i>QlikView</i>	<i>SpagoBI</i>
Colocar a Logo	0	5	5
Apresentação	4	5	5
Ícones representativos	3	3	3
Facilidade para encontrar funcionalidades	4	3	3
Idioma Português	5	5	0

**Tabela 3.4.** Aspectos de usabilidade das ferramentas.

O *SpagoBI* e o *QlikView* apresentaram a possibilidade de inserção da logo da instituição em sua plataforma, mas só é possível fazer esta tarefa no *SpagoBI* quando alterado o código fonte, fazendo com que usuário que não tenha conhecimentos técnicos de programação não possam fazer esta alteração.

Por fim, foram analisados os tipos de visualização das informações que poderiam ser obtidas através das ferramentas: se existe uma variedade de gráficos que poderiam ser utilizados, possibilidade de utilizar indicadores, possibilidade de criação de mapas, facilidade na criação de *dashboards*, se a ferramenta apresenta suporte e controle a múltiplos usuários e se possibilita consultas multidimensionais. A Tabela 3.5 mostra os resultados obtidos.

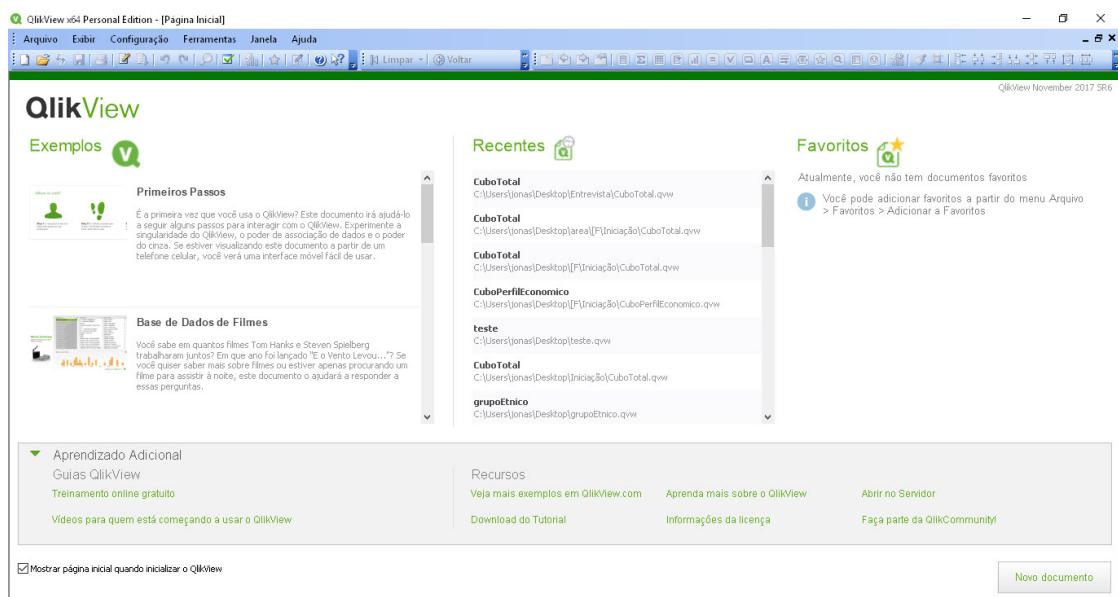
Características	<i>Apache Superset</i>	<i>QlikView</i>	<i>SpagoBI</i>
Mais de 5 tipos de gráficos diferentes	SIM	SIM	SIM
Possível criar indicadores	SIM	SIM	SIM
Possível criar mapas	SIM	SIM	SIM
Criação dashboards intuitiva	NÃO	SIM	SIM
Suporte a múltiplos usuários	SIM	NÃO	SIM
Controle de múltiplos usuários	SIM	NÃO	SIM
Consulta multidimensional	NÃO	SIM	SIM

**Tabela 3.5.** Características de visualização das ferramentas.

Após o levantamento e testes das atividades que poderiam ser executadas em cada ferramenta, pôde-se identificar falhas operacionais nas ferramentas *Apache Superset* e *SpagoBI*, fazendo com que seja inviável colocar essas ferramentas em uso na instituição. A ferramenta *Apache Superset* não permite consultas em múltiplas tabelas simultaneamente, não sendo possível então fazer consultas mais complexas, quando o idioma é alterado a funcionalidade de criação de *dashboards* para de funcionar. O *SpagoBI* apresenta erros internos em sua construção, um exemplo que pode ser observado durante a fase de teste é a conversão de cadeias de caracteres

em números muito grandes, dando erro de acesso aos dados, este exemplo pôde ser observado no atributo CPF pertencente a tabela fato candidato, ao se fazer uma contagem de quantos candidatos tinham em uma dada consulta ele apresentava erro e não apresentava a summarização dos dados.

Apesar de não ter controle de usuários o *QlikView* é uma ferramenta muito intuitiva que apresenta: suporte técnico de qualidade, intuitividade, possibilita uma análise rápida e eficiente. Deste modo, O *QlikView* foi escolhido como a ferramenta para para fazer as análise dos dados. Na próxima seção será feita uma apresentação do *QlikView*.



**Figura 3.7.** Tela de inicialização do *QlikView*.

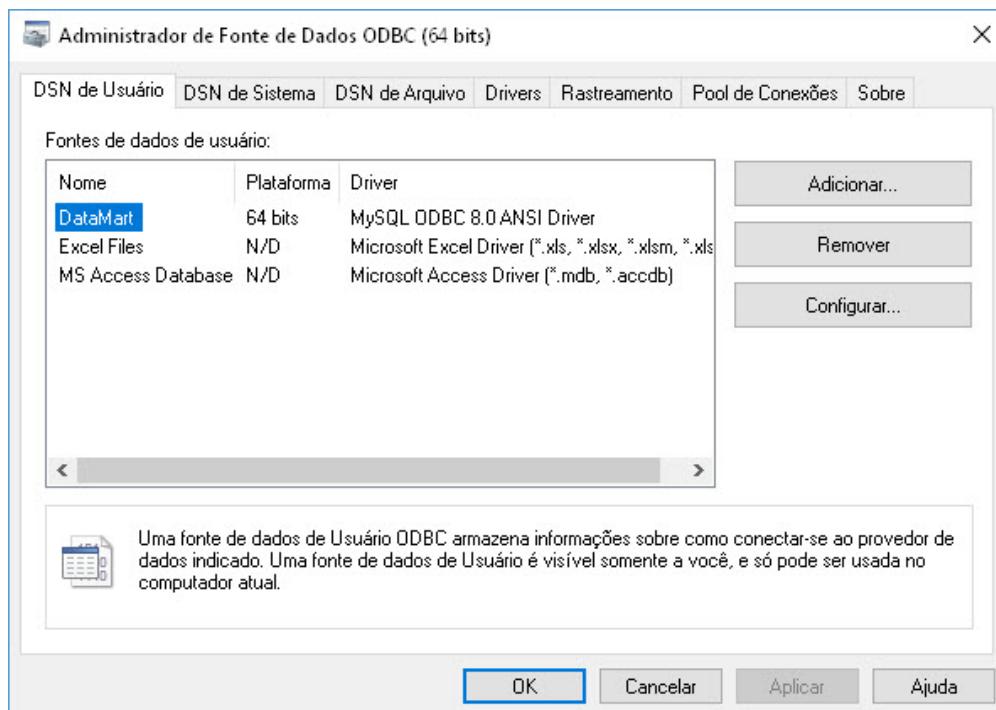
### 3.6.1 QlikView

O *QlikView*, é uma ferramenta intuitiva, onde sua *interface* inicial, fornece ao usuários informações como: arquivos utilizados recentemente, primeiros passos (onde é mostrado aos usuários como fazer as primeiras análises), locais de assistência, locais onde pode-se fazer treinamentos e mostra também exemplos de aplicações. A Figura 3.7 ilustra a interface inicial da ferramenta.

Esta ferramenta não necessita de construção de cubos multidimensionais, ao invés de tratar os dados como cubos ela cria um grande arquivo com todas as associações entre os dados chamada nuvem de dados (*Data Cloud*). Mesmo utilizando o *QlikView* como ferramenta de análise, este trabalho organiza os dados em formas de cubos dimensionais a fim de organizar as consultas de modo que obtenham os resultados esperados pelos gestores.

Para utilizar os dados carregados no *data mart*, é necessário fazer o *download* e configurar o *drive ODBC*, que permitirá ao *QlikView* se associar ao banco *Mysql*. Configurando o *drive*:

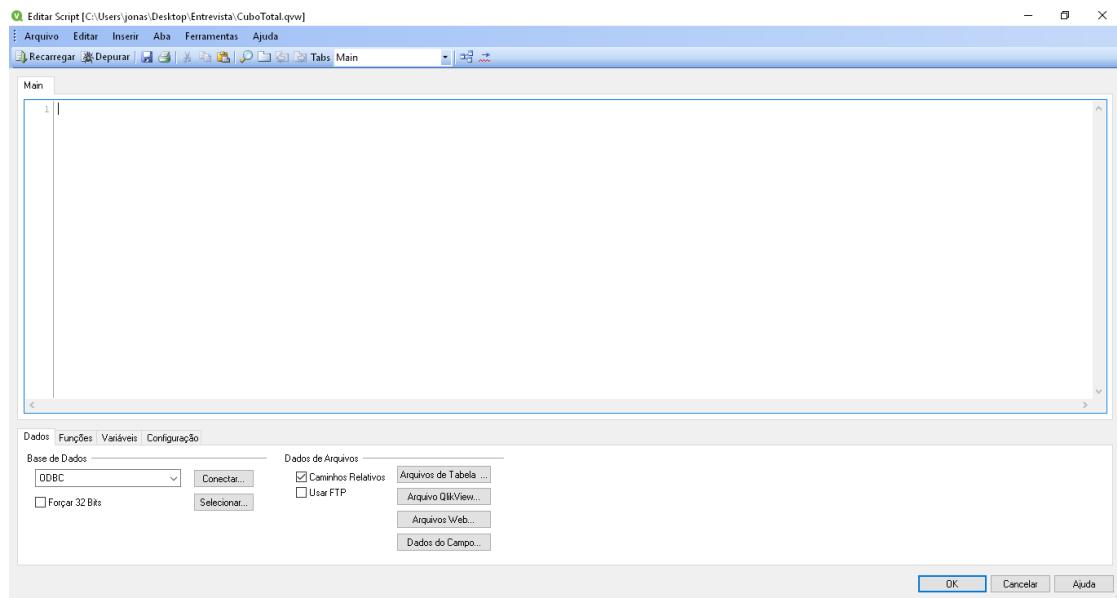
1. É feito o download e instalação do *drive ODBC* do *Mysql*.
2. É executado a ferramenta "Administrador de Fonte de Dados ODBC (64 bits)" em modo administrador (Figura 3.8).
3. Seleção da guia "DSN de Sistema".
4. Seleção do botão "Adicionar"
5. Seleção do *drive MySQL ODBC 5.x ANSI or Unicode Driver*. O 5.x corresponde a versão do *drive* que foi instalado.
6. Na nova janela é preenchido os dados referentes ao banco. Nome que irá utilizar no *drive* configurado (neste trabalho foi utilizado *DataMart*), o TCP/IP Server (o padrão é localhost), nome do usuário (o padrão é *root*), senha do *MySQL*, o nome do *data mart*, que no caso deste trabalho é *datamartifnmg*.



**Figura 3.8.** Janela de configuração da fonte de dados.

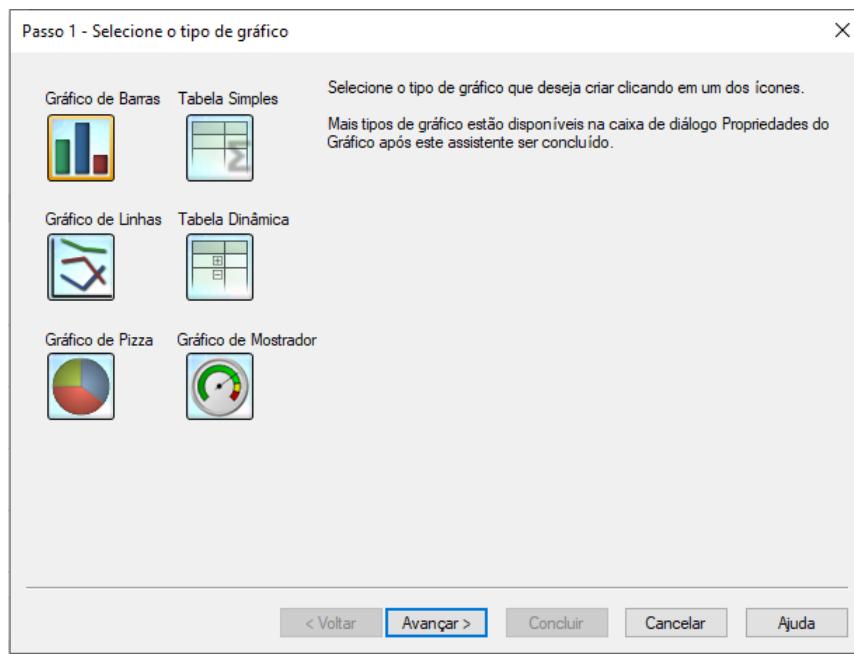
Após as configurações do *drive*, é preciso configurar o *QlikView* para acessar os dados e escolher quais deles serão carregados.

1. No *QlikView* é acessado o menu "Arquivos" e selecionado a opção "editar script" (Figura 3.9).
2. Seleção do botão "conectar".
3. Na janela aberta, é selecionado "Mostrar DSNs do Usuário", após selecionar o *drive* ODBC criado (DataMart).
4. Na janela apresentada na Figura 3.9, é selecionado o botão "selecionar". Onde que será escolhido, quais dados do *data mart* poderam ser usados (nesta etapa é selecionado os dados referentes a um dado cubo).
5. Após a seleção dos dados, ainda na janela apresentada na Figura 3.9 seleciona-se o botão "Recarregar", o qual carregará os dados para a utilização da ferramenta.



**Figura 3.9.** Janela de configuração do acesso aos dados do *data mart*.

Com o carregamento dos dados feito, o *QlikView* está apto a poder fazer os *dashboards*. Os *dashboards* são criados através de posicionamento de gráficos, tabelas, indicadores e demais formas de visualização de dados disponíveis pelo *QlikView*. Um gráfico pode ser criado a partir do menu "assistente de gráfico rápido"(Figura 3.10), onde seleciona-se qual tipo de gráfico será utilizado. Após a seleção do tipo de gráfico, é selecionado quais as dimensões (pode ser: eixo-x, eixo-z...) do mesmo será analisada. Após a seleção da dimensão é feito a seleção de qual expressão será feita, se será feito uma contagem dos dados, se será feita uma média de algum atributo ou se será feito uma soma de algum atributo (este será o eixo-y), por fim é selecionado algum estilo do gráfico, permitindo ao usuário a escolha de qual estilo de gráfico será visualizado.



**Figura 3.10.** Janela de criação de gráficos do *QlikView*.

Foi criada uma área de filtros onde é possível visualizar todas as perguntas e possíveis respostas presentes no *data mart* (Figura 3.11), nela é possível selecionar qual o filtro que deve ser aplicado, e também é possível visualizar quais os dados que estarão e que não estarão presentes nas consultas que serão realizadas.

Para se fazer as consultas foram criadas pastas, onde cada pasta continha um *dashboard* correspondente a um assunto que poderia ser extraído do cubo que está sendo analizado (Figura 3.12).

The screenshot shows a QlikView interface with several filter panels:

- Nome Campus:** ALMENARA, ARAÇUAÍ, ARINOS, DIAMANTINA, JANUÁBA, JANUÁRIA, MONTES CLAROS, PIRAPORA, PORTEIRINHA, SALINAS, TEÓFILO OTONI.
- Turno:** DIURNO, INTEGRAL, MATUTINO, NOTURNO, VESPERTINO.
- Ano:** 2013, 2014, 2015, 2016, 2019.
- Semestre:** 1, 2.
- Nome do Curso:** Licenciatura em Química, Análise e Desenvolvimento de Sistemas, Bacharelado em Administração, Bacharelado em Engenharia Agrícola e Ambiental, Bacharelado em Engenharia Civil, Bacharelado em Sistemas de Informação, Ciência da Computação, Engenharia Agronômica, Engenharia de Alimentos, Engenharia Elétrica, Engenharia Florestal, Engenharia Química, Licenciatura em Ciências Biológicas, Licenciatura em Física, Licenciatura em Matemática, Medicina Veterinária, Pedagogia.
- Educational Level:** Analfabeto, Ensino Fundamental completo, Ensino Fundamental incompleto, Ensino Médio completo, Pós-Graduado, Superior completo.
- Mother's Educational Level:** Analfabeta, Ensino Fundamental completo, Ensino Fundamental incompleto, Ensino Médio completo, Pós-Graduada, Superior completo.
- Father's Situation:** É falecido e não deixou pensão, Está desempregado, Outra situação, Trabalha regularmente, Vive de renda.
- Mother's Situation:** É falecida e não deixou pensão, Está desempregada, Outra situação, Trabalha regularmente, Vive de renda.
- Exercises Remunerated Activity:** Não, Sim, em tempo integral (mais de trinta horas semanais), Sim, em tempo parcial (até vinte horas semanais), Sim, mas se trata de trabalho eventual.
- Monthly Income:** De 0,5 salário-mínimo até 1 salário-mínimo, De 1 salário-mínimo até 1,5 salário-mínimo, Mais de 1,5 salário-mínimo, Menos do que 0,5 salário-mínimo, Não tenho nenhuma renda mensal.
- Gross Monthly Income:** Renda Bruta Mensal.
- Etnia:** Etnia.

Figura 3.11. Janela criada para fazer filtros no *QlikView*.

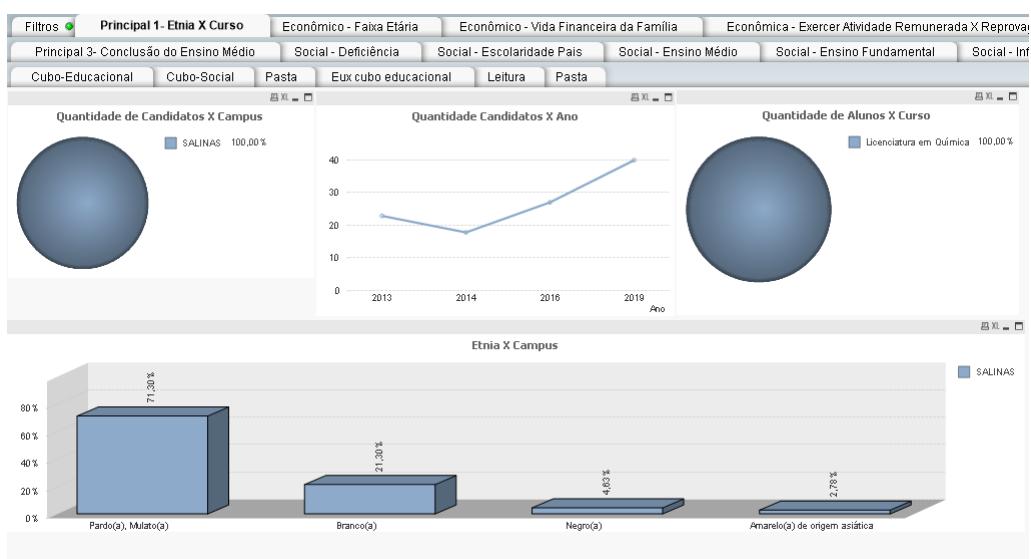


Figura 3.12. Exemplo de pasta com *dashboard*



# Capítulo 4

## Resultados

Após a implementação do *Data mart* acadêmico, o ambiente está pronto para submissão de consultas analíticas por partes dos usuários, neste contexto, os gestores da instituição. Visando ilustrar o tipo de informação que pode ser extraída do ambiente definido, este capítulo descreve um conjunto de análise que atende as necessidades de informação definidas pelos gestores do Campus Montes Claros (Seção 3.1).

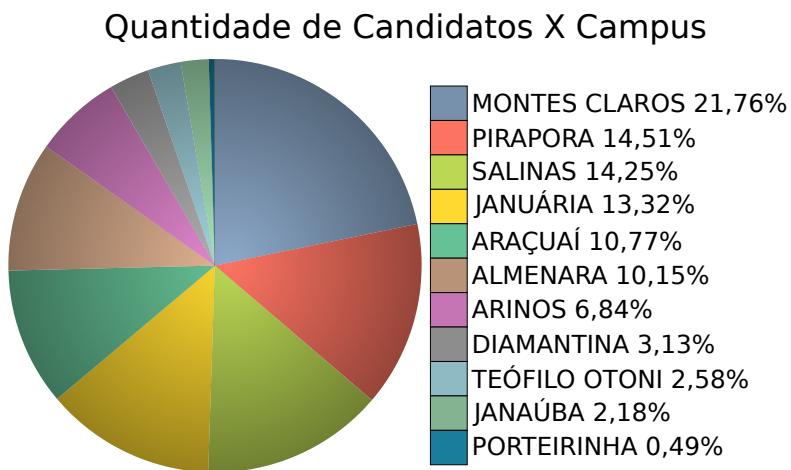
O conjunto de análises foi dividido em quatro partes, a primeira é responsável por apresentar as informações que podem ser obtidas em todos os 3 cubos de dados. A segunda parte corresponde às informações obtidas para definição de um Perfil-Econômico. A terceira parte visa apresentar as informações obtidas por meio da análise do Perfil-Social dos candidatos. Por fim, a última parte visa analisar as informações do Perfil-Educacional.

Apesar de não ser objetivo direto deste trabalho, as análises foram realizadas em computador portátil para ilustrar que o ambiente não possui restrição de *Hardware*, sendo a configuração: *notebook* com 4GB de memória principal e processador *Intel Core i3 - 6100U*.

### 4.1 Visão Geral

O IFNMG obteve 46600 candidatos inscritos em seus processos seletivos entre os anos de 2013 a 2019, sendo que estes estão distribuídas entre os 11 campi existentes

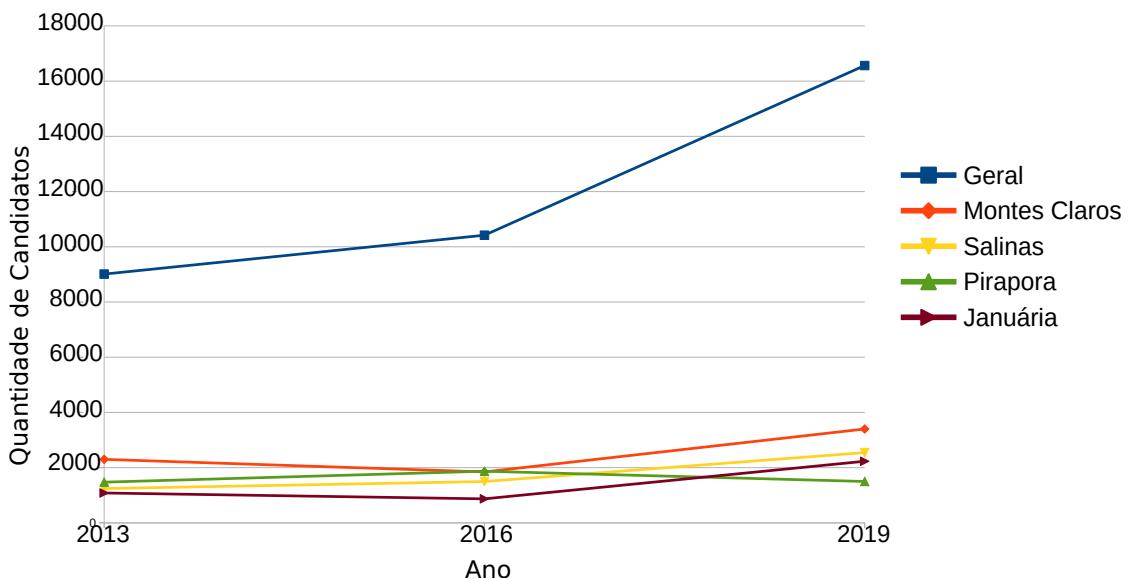
no Norte de Minas Gerais. A Figura 4.1 mostra a distribuição dos candidatos em relação aos Campi. Os quatro Campi com mais inscrições foram: o *campus* Montes Claros com 21,76% dos candidatos, o *campus* Pirapora com 14,51%, o *campus* Salinas com 14,25% e o *campus* Januária com 13,32%, sendo que juntos, esses Campi acumulam, aproximadamente, 64% de todas as inscrições dos processos seletivos.



**Figura 4.1.** Distribuição dos candidatos por campus.

A Figura 4.2 apresenta a quantidade de candidatos inscritos nos processos seletivos em relação aos períodos que teve acesso a informação. A quantidade de candidatos é crescente tendo um aumento percentual de 84% do ano de 2013 para o ano de 2019 (Figura 4.2). Assim, qualquer análise histórica é apenas uma aproximação, pois como já foi discutido (Seção 3.2), não foi possível recuperar todos os dados dos processos seletivos realizados. Observando os Campi mais concorridos, com exceção de Pirapora existe um aumento do ano de 2013 para 2019 como pode ser observado na Figura 4.2. Os *campus* de Montes Claros e Januária, tiveram um decréscimo de aproximadamente 20% dos candidatos, do ano de 2013 para o ano de 2016, mas do ano de 2016 para o ano de 2019 houve um aumento de aproximadamente 84% no *campus* Montes Claros e o *campus* Januária aumentou 1.5 vezes a quantidade de seus candidatos. O *campus* de Pirapora teve uma diminuição de aproximadamente 20% de seus candidatos do ano de 2016 para o ano de 2019. Já no *campus* de Salinas sempre houve um

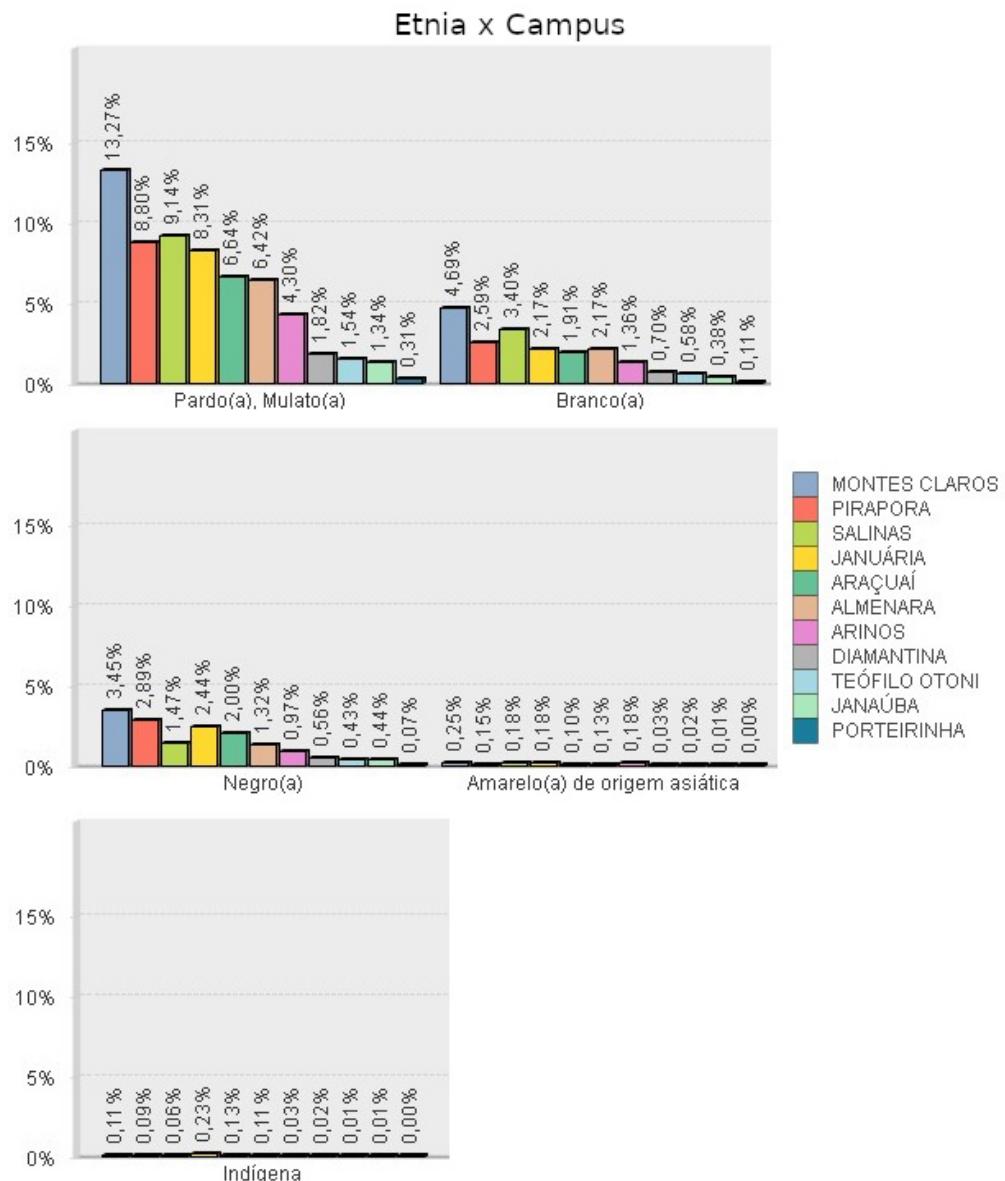
aumento na quantidade de candidatos ao longo dos anos tendo aumentado 1.5 vezes a quantidade de candidatos do ano de 2013 para o ano de 2019.



**Figura 4.2.** Quantidade de candidatos nos anos de 2013, 2016 e 2019

Com relação a Etnia, existe uma predominância de candidatos que se consideram pardo(a) representando 61% dos candidatos, a segunda maior etnia entre os candidatos é de 20,04% estes se consideram Branco(a), 16,03% dos candidatos se consideram Negro(a) mostrando que há apenas uma variação de 4% dos candidatos Negros para os candidatos Brancos, já a minoria dos candidatos se consideram Amarelo(a) de origem asiática ou Indígenas correspondendo a 1,25% e 0,79% respectivamente. A distribuição de Etnia por campus é representada na Figura 4.3.

Em Montes Claros e Salinas a ordem de predominância das etnias é a mesma do contexto geral, tendo como predominância a etnia pardo(a), seguida da etnia branco(a), depois a etnia Negro(a) e a Etnia Indígena. O campus de Pirapora se destoa do cenário principal, já que a segunda maior etnia apresentada pelos candidatos é a etnia negra. O campus de januária também tem uma característica distinta do cenário principal, tendo como etnia principal pardo(a), mulato(a); a segunda maior etnia sendo de pessoas que se consideram negros(as); a terceira

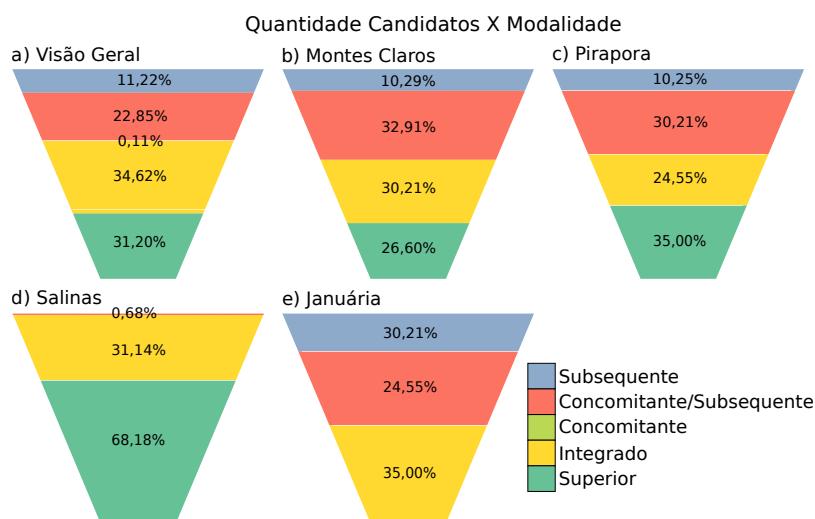


**Figura 4.3.** Distribuição da etnia dos candidatos em relação aos campi.

maior etnia de pessoas que se consideram brancas(os) tendo então apenas a característica de pessoas pardo(a), mulato(a) igual a característica geral dos demais campi.

Outra característica predominante é a escolha da modalidade do curso pelos candidatos, sendo que na Figura 4.4 (a) é mostrada a situação geral em que 32,62%

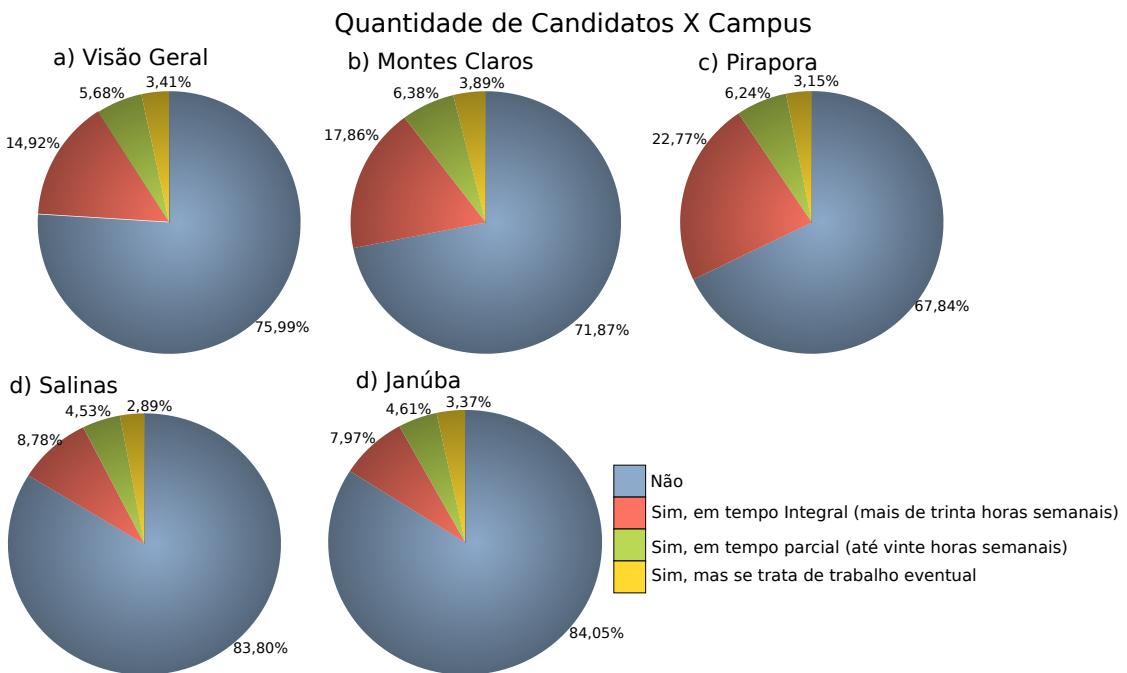
dos candidatos participam do processo seletivo para ingressar na modalidade Integrado, 31,20% tentam a modalidade Superior, 22,85% tentam a modalidade Concomitante/Subsequente, 11,22% visam ingressar na modalidade Subsequente e 0,11% tentam ingressar na modalidade Concomitante. Este cenário se diferencia no campus Montes Claros como pode ser observado Figura 4.4 (b), a predominância é na modalidade Concomitante/Subsequente com 32,91%, tendo como segundo maior a modalidade Integrado com 30,21%, os cursos superiores como a terceira maior opção com 26,60% e a modalidade Subsequente como a quarta opção com 10,29%. O campus de Pirapora tem como modalidade principal de escolhas aos cursos Superiores com 35,00%, a segunda escolha são os cursos Concomitante/Subsequente com 30,21%, a terceira maior modalidade é Integrado com 24,55% e a quarta é Subsequente com 10,25%. O campus de Salinas tem como predominância absoluta a escolha de cursos Superiores com 68,18%, a segunda maior modalidade são os cursos Integrados com 31,14% e a terceira maior escolha Concomitante/Subsequente com 0,68%. Apesar de não ter cursos superiores o *Campus Januária* segue a ordem de modalidade geral, distinguindo apenas nos valores de cada modalidade.



**Figura 4.4.** A distribuição dos candidatos em relação a modalidade de ensino nos quatro campi com mais inscrições. (a) Visão Geral, (b) Montes Claros, (c) Pirapora, (d) Salinas, (e) Januária.

## 4.2 Análise do Perfil-Econômico

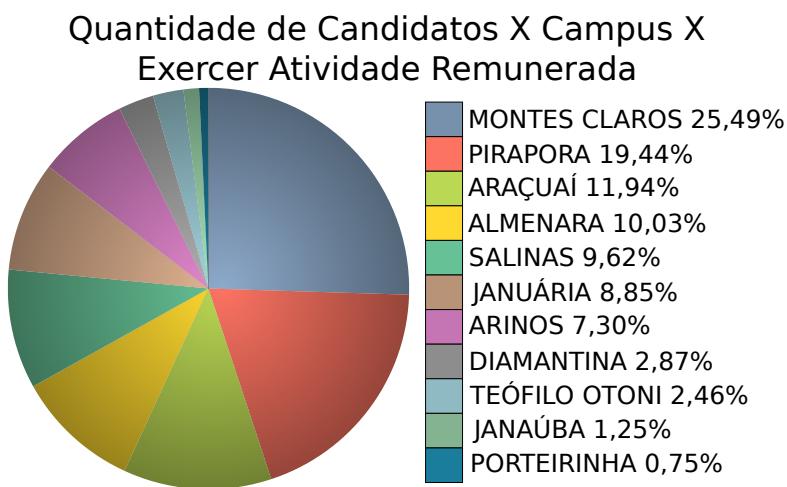
Uma grande característica dos candidatos é de não exercer atividades remuneradas, como ilustrado na Figura 4.5 (a), nela mostra que 75,99% dos candidatos não exercem atividades remuneradas e 24,01% dos candidatos exercem algum tipo de atividade remunerada. Fazendo filtros para os 4 campus com mais candidatos, temos que os campus Montes Claros, Pirapora, Salinas e Januária apresentam características similares ao cenário geral, como pode ser visto na Figura 4.5, mostrando que em todos os campi mais de 50% dos candidatos não exercem atividades remuneradas. Com a discrepância na quantidade de candidatos em que exercem atividades remuneradas e os que não exercem, optou-se em dividir esta seção em duas partes, uma que mostra o perfil dos candidatos que exercem alguma atividade remunerada e a outra dos candidatos que não exercem.



**Figura 4.5.** Quantidade de candidatos que exercem e não exercem atividades remuneradas.

### 4.2.1 Perfil dos Candidatos que Exercem Atividades Remuneradas

Existem 11188 candidatos que exercem algum tipo de atividade remunerada, estes correspondem a 24,01% dos candidatos. A Figura 4.6 mostra a distribuição dos candidatos pelos *campus*, o *campus* de Montes Claros e Pirapora continuam sendo os que mais tem candidatos, já os *campus* de Salinas e Januária apresentaram menos candidatos que os *campus* de Araçuaí e Almenara.

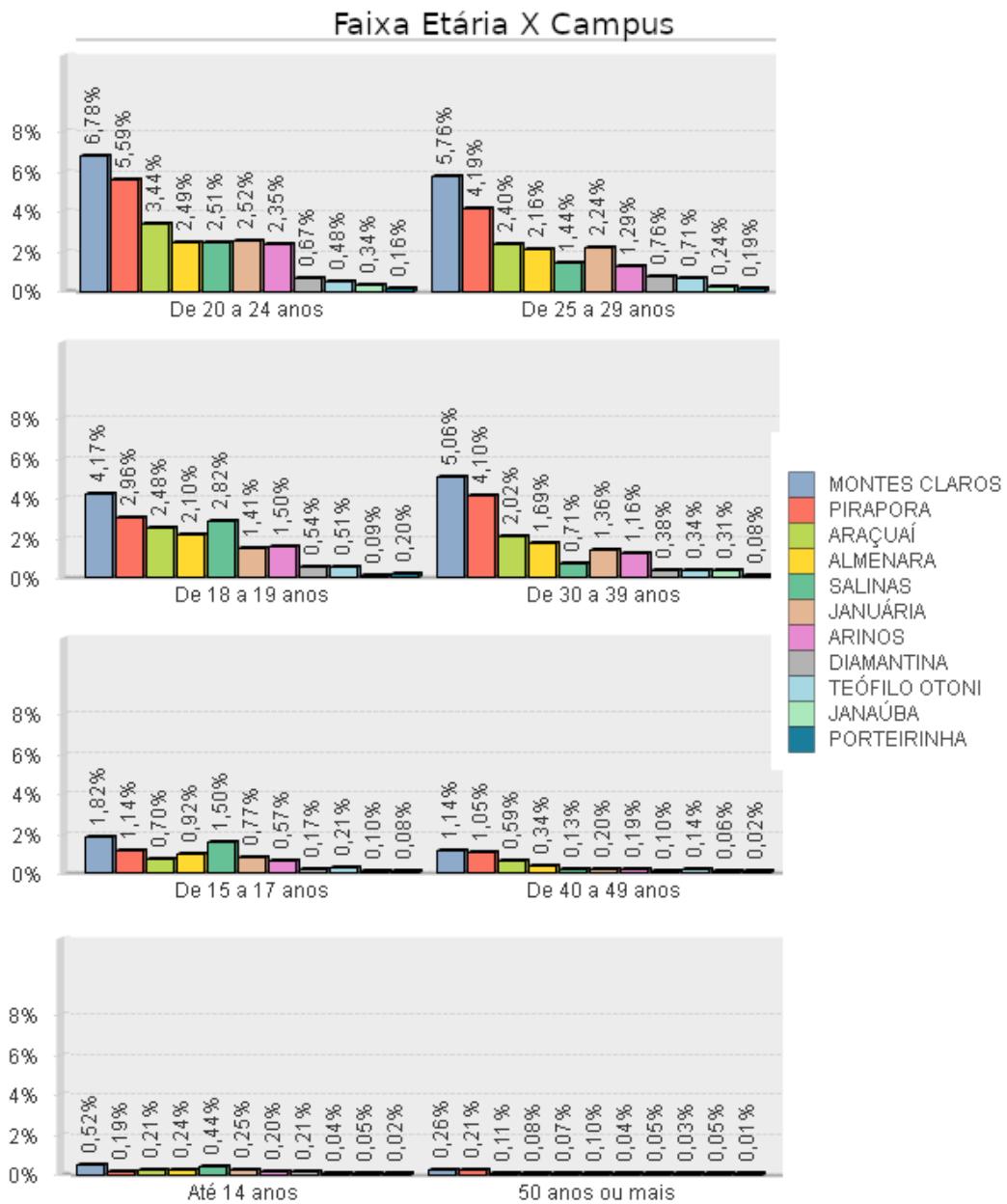


**Figura 4.6.** Quantidade de Candidatos que Exercem Atividade Remunerada X Campus

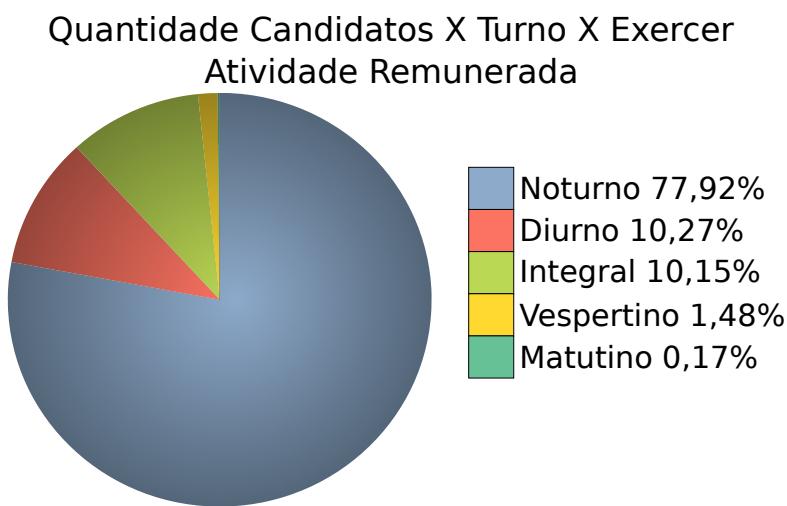
Fazendo a seleção dos candidatos através da faixa etária, pode-se perceber que candidatos que têm idade entre 20 a 29 anos correspondem a maior quantidade dos que trabalham correspondendo a 48,64%. O mesmo cenário se procede nos quatro campus com maior quantidade de candidatos, onde que a maior parte dos candidatos que exercem atividades remuneradas têm a mesma faixa de idade, como ilustrado na Figura 4.7

Com o intuito de trabalhar durante o dia, 77,92% dos candidatos preferem estudar durante a Noite, já 10,27% pretendem ingressar no turno Diurno, 10,15% no turno Integral, 1,48% no turno Vespertino e 0,17% no turno Matutino (ilustrado na Figura 4.8). O campus de Montes Claros tem como a principal escolha o turno Noturno com 67,85%, a segunda maior escolha se distingue do cenário geral, sendo

o turno Integral com 23,91%, em terceiro o turno Diurno com 6,45% e em quarto o turno vespertino com 1,47%. Pirapora apresenta características similares ao cenário principal.



**Figura 4.7.** Faixa Etária X Campus.

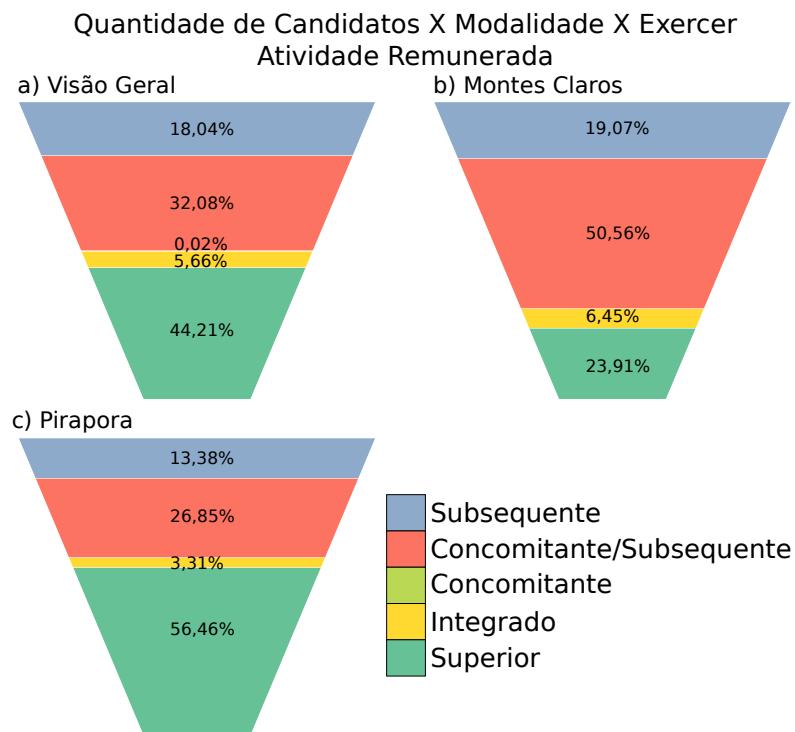


**Figura 4.8.** Quantidade de candidatos que exercem atividades remuneradas X Turno.

Analizando a modalidade dos cursos, temos que os candidatos tentam ingressar em cursos Superiores correspondendo a 44,28%, a segunda maior escolha são os cursos Concomitante/Subsequente com 31,96%, em seguida vem os cursos Subsequentes com 18,17%, logo após, os cursos Integrados com 5,67%, mostrando que candidatos que exercem atividades remuneradas tendem a não escolher os cursos Integrados. Já no campus de Montes Claros a situação do Concomitante/- Subsequente é a maior com 50,56%, a segunda maior é o Superior com 23,01%, esta inversão na ordem de predominância das modalidades pode ser pelo fato dos candidatos buscarem uma rápida especialização a fim de melhorar suas carreiras profissionais. O campus de Pirapora apresenta características similares ao cenário principal, tendo os cursos Superiores com 56,46% dos candidatos, Concomitante/- Subsequente com 26,85%, Subsequente com 13,38% e Integrado com 3,31%. O cenário descrito pode ser observado na Figura 4.9.

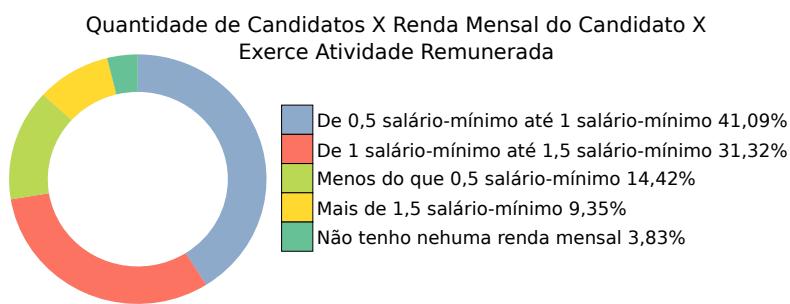
A renda mensal predominante dos candidatos que exercem atividade remunerada é de meio salário-mínimo até um salário mínimo como pode ser visto na Figura 4.10, existem poucos candidatos têm uma renda acima de um salário-mínimo e meio, representando apenas 9,35% do total, 3,83% dos candidatos disseram não ter renda mensal, o que é uma contradição. Esses candidatos podem ter interpretado incorretamente a pergunta "Você exerce alguma atividade remunerada?",

já que apesar de exercerem algum tipo de atividade remunerada, afirmaram não possuir renda. É possível que as perguntas do questionário não estejam suficientemente claras ou houve falta de atenção por parte do candidato no preenchimento do questionário. Fazendo filtro para observar o cenário de Montes Claros (Figura 4.11 (a)) temos que ele é similar ao cenário geral. Filtrando pelo *campus* de Pirapora nota-se que ele se distingue do cenário principal tendo a renda de um e meio salário-mínimo com a terceira maior opção, apesar de representar apenas 12,37% dos candidatos de Pirapora, esta informação pode ser observada na Figura 4.11 (b).

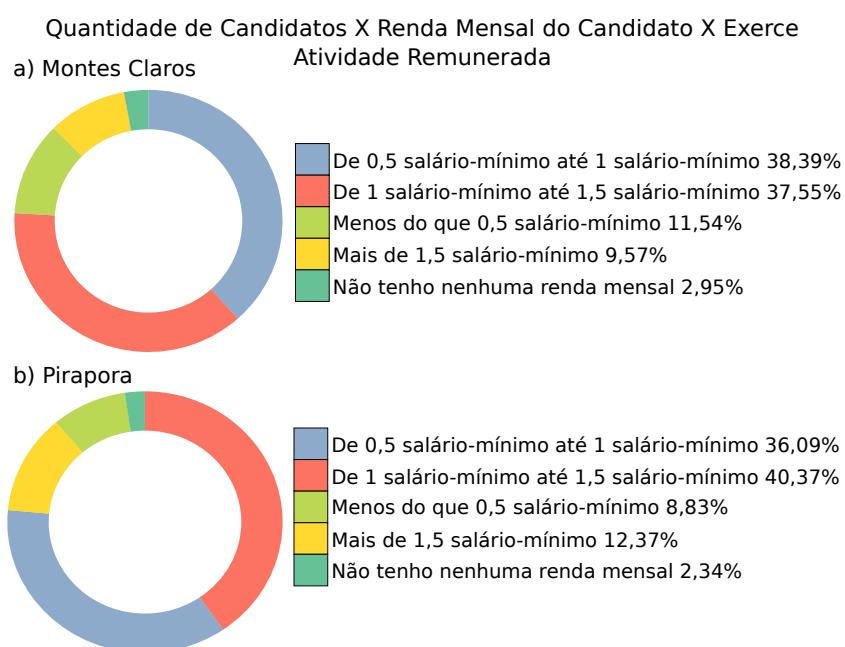


**Figura 4.9.** Quantidade de candidatos que exercem atividades remuneradas X modalidade.

A renda bruta mensal da família dos candidatos, no geral, que exercem atividade remunerada é de um a um e meio salários-mínimos tendo 35,39% dos candidatos, a segunda maior renda é de mais de um e meio salários-mínimos com 31,72% como ilustrados na Figura 4.12. Um estado similar acontece nos campos de Montes Claros e Pirapora, onde que a maior parte dos candidatos tem renda

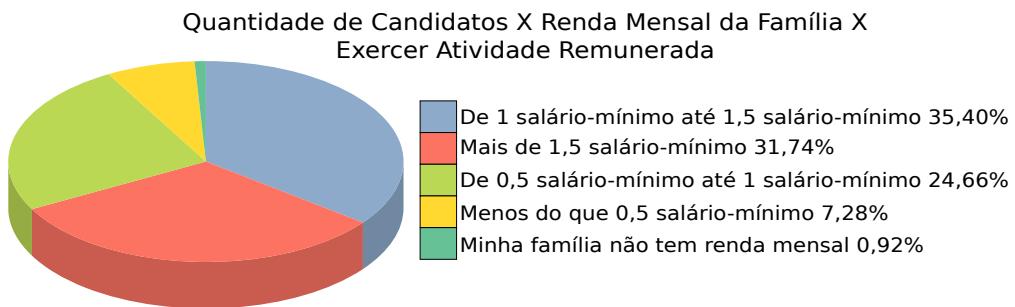


**Figura 4.10.** Quantidade de candidatos que exercem atividades remuneradas X Renda Mensal do Candidato em todos os campus.



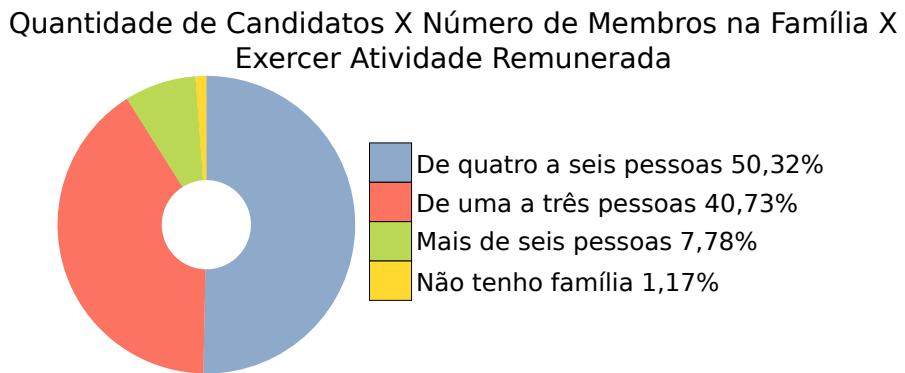
**Figura 4.11.** Quantidade de candidatos que exercem atividades remuneradas X Renda Mensal do Candidato.

entre um e um e meio salários-mínimos, e a segunda maior renda é de mais de um salário-mínimo e meio. Sendo que, aproximadamente, menos de 8% apresentam renda inferior a meio salário mínimo, mostrando que a renda bruta familiar está condizente com a renda média familiar de Minas Gerais, que segundo o censo do IBGE (Instituto Brasileiro de Geografia e Estatística) de 2017, a renda bruta de uma família é de R\$ 1.322,00 reais, enquadrando na faixa salarial principal.



**Figura 4.12.** Quantidade Candidatos X Renda Bruta x Exercer Atividade Remunerada.

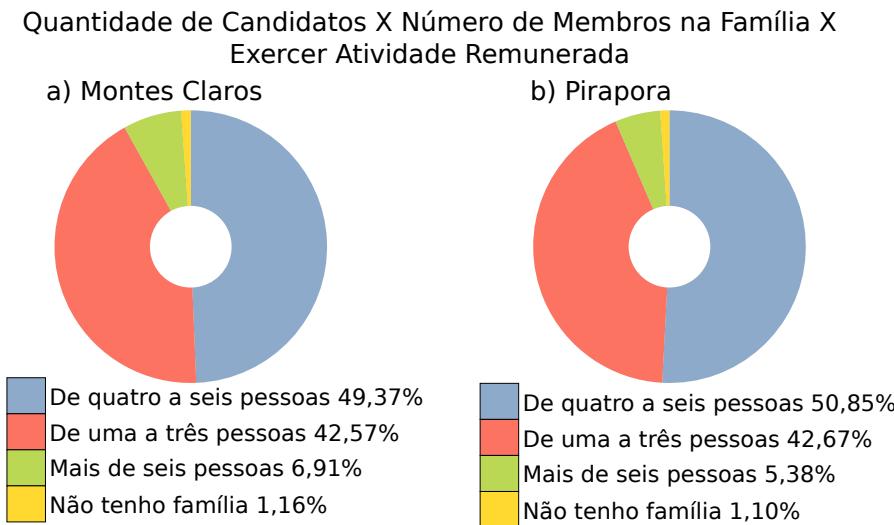
Uma dúvida frequente para os gestores é se a situação de trabalho do candidato é influenciada pela quantidade de membros em sua família. Na Figura 4.13 pode-se notar que a predominância é de 4 ou mais pessoas na família, e a segunda maior é de cerca de uma a três pessoas, levando em consideração o cenário geral da renda mensal dos candidatos, é correto afirmar que os candidatos têm características de auxiliar no sustento da família.



**Figura 4.13.** Quantidade Candidatos X Número de membros na Família Geral X exerce Atividade Remunerada.

Filtrando pelo campus de Montes Claros e pelo campus de Pirapora pode-se fazer uma conclusão similar já que a quantidade de membros na família está entre uma a seis pessoas, o salário mensal dos candidatos varia entre meio a um e meio salários-mínimos e existe 35,39% das famílias com renda bruta entre meio a um e meio salários mínimos. A Figura 4.14 mostra a quantidade de membros nas

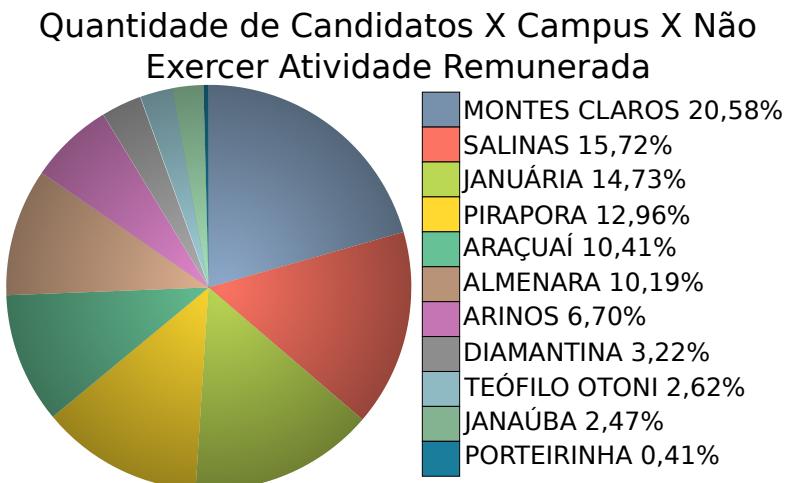
famílias das pessoas que exercem atividades remuneradas nos campus de Montes Claros e Pirapora.



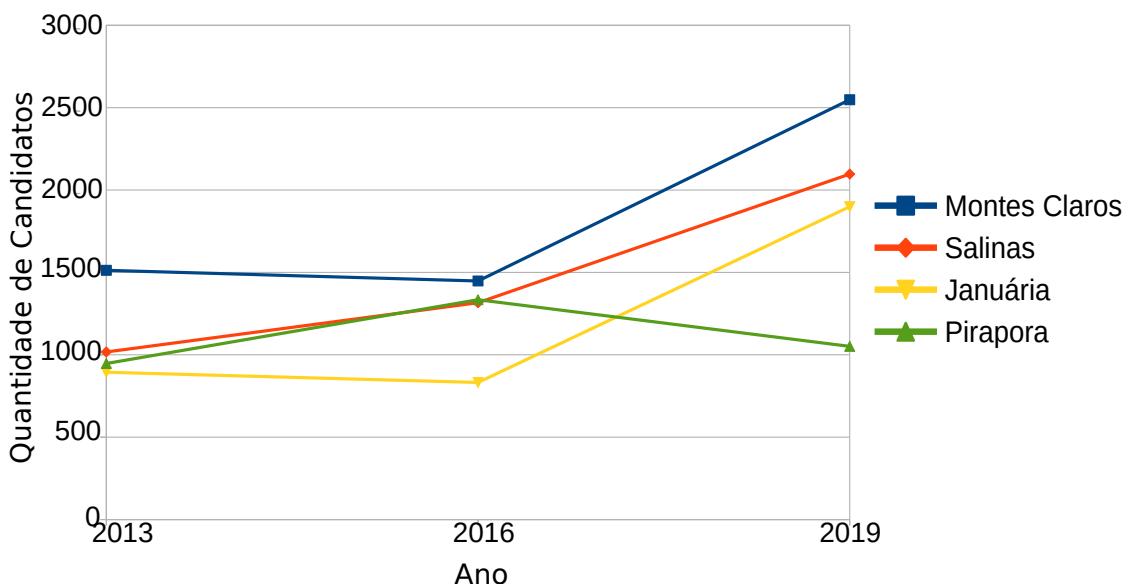
**Figura 4.14.** Quantidade Candidatos X Número de membros na Família X exerce Atividade Remunerada.

#### 4.2.2 Perfil dos Candidatos que não Exercem Atividades Remuneradas

Existem 35412 candidatos não exercem atividades remuneradas correspondendo a 75,99% do total. O campus de Montes Claros contém a maior quantidade de candidatos que não exercem atividades remuneradas (Figura 4.15 (a)), o segundo maior *campus* é o *campus* de Salinas que no cenário geral corresponde ao terceiro maior, mostrando que apesar de ter a maior quantidade de candidatos tentando ingressar em cursos Superiores (Figura 4.4 (d)) 83,80% de seus candidatos não trabalham (Figura 4.5), podendo concluir que a escolha de modalidade nem sempre influencia em exercer atividade remunerada. Mas como foi visto na Figura 4.9, exercer atividade remunerada implica na escolha da modalidade.



**Figura 4.15.** Quantidade Candidatos X Campus X Não Exercer Atividade Remunerada.



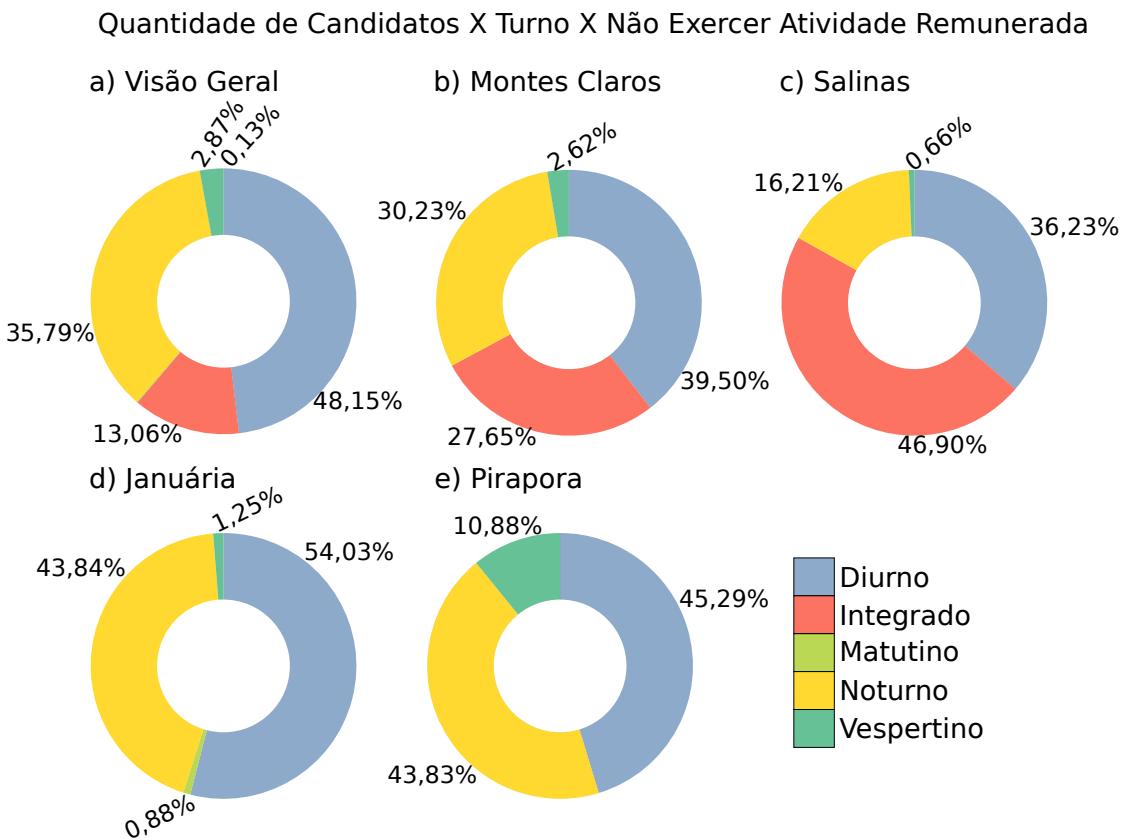
**Figura 4.16.** Quantidade Candidatos que não exercem atividade remunerada X Pelos anos de 2013, 2016 e 2019.

Com o aumento dos candidatos ao longo dos anos, vem aumentando também a quantidade de candidatos que não exercem atividade remunerada nos *campus* de Montes Claros, Salinas e Januária, como pode ser visto na Figura 4.16. Já no

*campus* de Pirapora houve um decréscimo de aproximadamente 11% na quantidade de pessoas que não exercem atividade remuneradas do ano de 2016 para o ano de 2019, este fato é resultante da redução de ingresso dos candidatos neste período como visto na Figura 4.2.

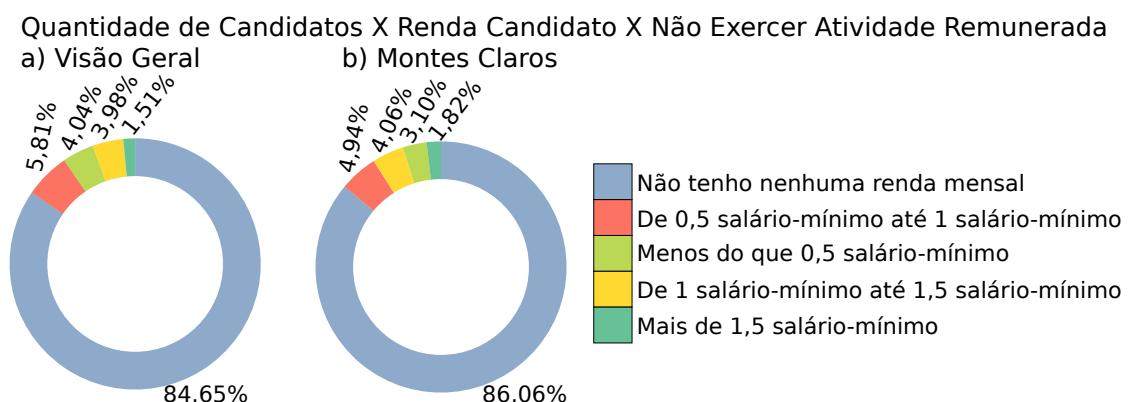
Os candidatos que não exercem atividades remuneradas tem como preferência o turno diurno constituindo 48,15%, já o turno noturno que é o turno mais escolhido por todos os candidatos é a segunda opção de escolha constituído de 35,79% como pode ser visto na Figura 4.17 (a), este cenário é similar ao campus Montes Claros como pode ser visto na Figura 4.17 (b). Em Salinas o turno noturno é o terceiro maior tendo apenas 16,21% dos candidatos, o maior, é o turno integral com 46,90% dos candidatos (Figura 4.17 (c)). Em Januária, se mantém o cenário principal dos candidatos que não exercem atividades remuneradas, tendo 54,03% dos candidatos prestando processo seletivo para o turno diurno, 43,84% prestando vestibular para o turno noturno (Figura 4.17 (d)). O mesmo acontece em Pirapora, 45,29% dos candidatos prestam vestibular para o turno diurno, 43,83% para o turno noturno (Figura 4.17 (e)). Apesar da preferência de escolha dos candidatos que não exercem atividades remuneradas são os turnos diurnos e integrais, mostrando que os cursos que são ofertados durante este turnos são uma das principais escolhas tanto pelas pessoas que exercem atividades remuneradas quanto as que não exercem.

Como os candidatos não exercem atividade remunerada, 84,64% destes não possuem renda mensal, sendo o restante apresentando alguma forma de renda (Figura 4.18 (a)). Contudo, é importante destacar que quase 6% dos candidatos apresentam renda maior do que 1 salário sem exercer qualquer atividade remunerada, retomando então o fato de que os candidatos podem ter interpretado incorretamente a pergunta "Você exerce alguma atividade remunerada?", já que apesar de não exercerem algum tipo de atividade remunerada, afirmaram possuir renda elevada. É possível que as perguntas do questionário não estejam suficientemente claras ou houve falta de atenção por parte do candidato no preenchimento do questionário. Filtrando pelos quatro *campus* com maior quantidade de candidatos foi observado um cenário similar, todos tem 85% ou mais de candidatos que não possuem renda (Figura 4.18 (b) exemplifica a distribuição do *Campus* Montes Claros, os demais foram ocultados por serem análogos) e cerca de 6% informam ter mais de um salário mínimo. A renda bruta mensal da família dos candidatos

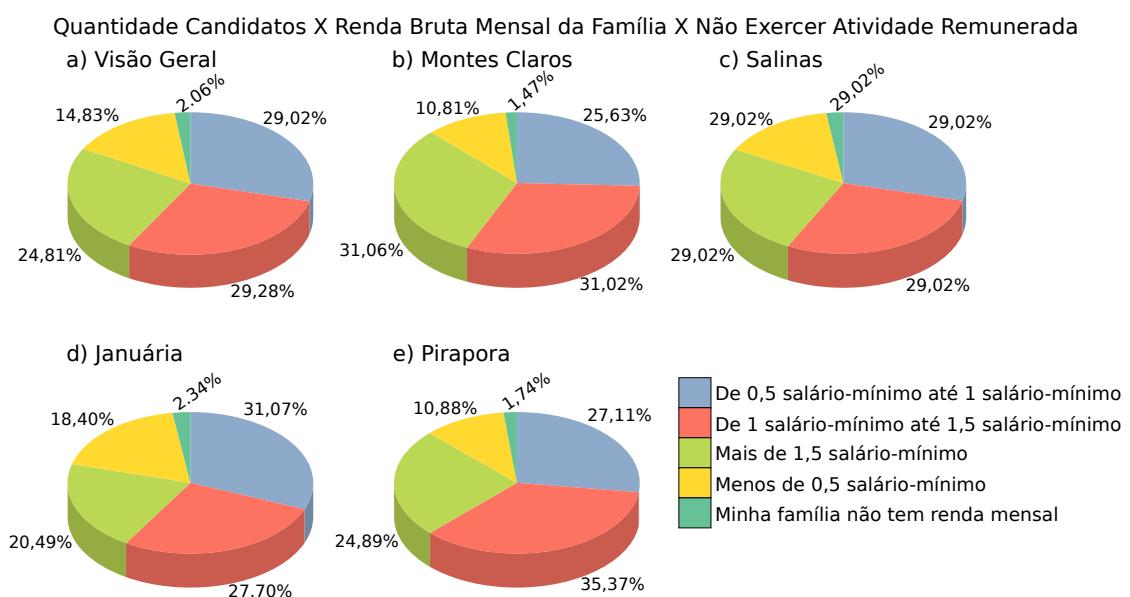


**Figura 4.17.** Quantidade candidatos que não exercem atividade remunerada X Turno.

tem predominância entre um e um e meio salários-mínimos como pode ser visto na Figura 4.20 (a), ainda pode ser visto nesta figura que a segunda maior renda é de meio a um e meio salários mínimos, com 29,02% dos candidatos. No *campus* de Montes Claros a principal renda bruta familiar é de mais de um e meio salários mínimos, formada com 31,94% dos candidatos do *campus*, isso se destoa do cenário principal mostrado, mostrando que os candidatos que tentam ingressar no *campus* pertencem a classe média alta ou a classe alta. A Figura 4.20 (b) mostra a distribuição dos candidatos por suas rendas brutas mensais. Em Salinas e Januária, o cenário apresentado é mais agravante que o cenário geral, mostrando que a maioria dos candidatos tem renda bruta familiar entre meio e um e meio salários-mínimos, como pode ser visualizado nas Figura 4.20 (c) e (d). No caso de



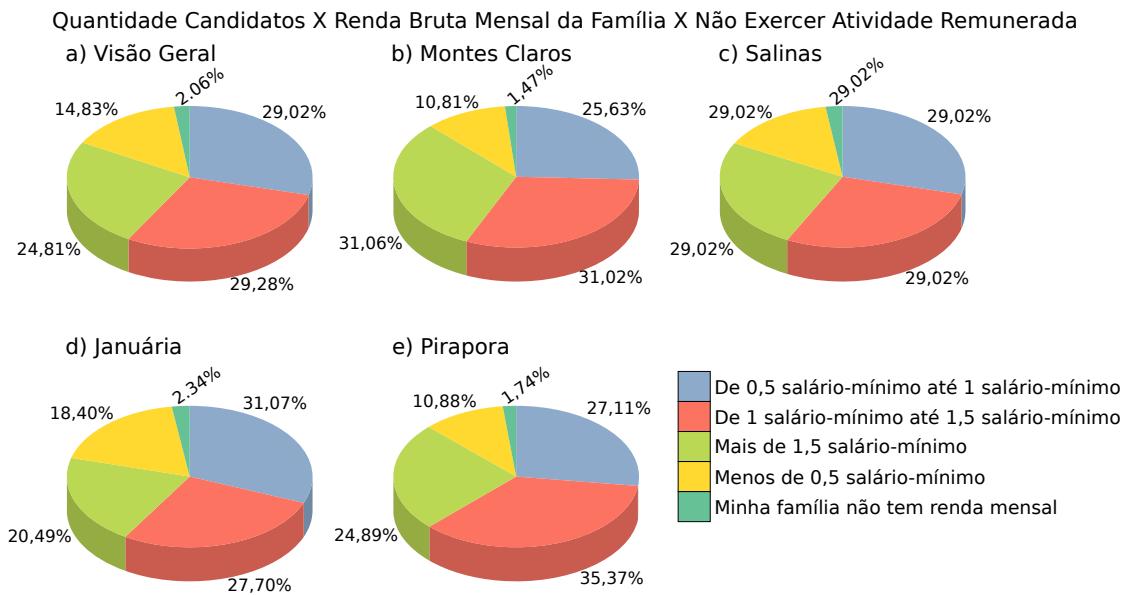
**Figura 4.18.** Quantidade Candidatos que não exercem atividade remunerada X Renda Mensal do Candidato.



**Figura 4.19.** Quantidade candidatos que não exercem atividade remunerada X Renda Bruta da Família.

Pirapora (Figura 4.20 (e)), pode-se perceber que a maior renda é entre um a um e meio salário-mínimo correspondendo a 35,37% dos candidatos, e como segunda maior renda é a de meio a um salários-mínimos com 27,11% dos candidatos. Estes dados mostram que a renda predominante da família pode variar de *campus* para

*campus*, mostrando que cada campus deva ter um cuidado específico para com seus candidatos.



**Figura 4.20.** Quantidade candidatos que não exercem atividade remunerada X Renda Bruta da Família.

Já que os candidatos em sua grande maioria não possuem renda, e existe renda na família é necessário observar o cenário familiar para entender o meio de sustento dos mesmos. A Tabela 4.1 mostra a distribuição dos candidatos através de sua renda e quantos membros na família. No cenário geral, é possível notar que as famílias compostas por 4 a 6 pessoas tem renda predominante "de 1 a 1,5 salário-mínimo", famílias com uma a três pessoas tem renda predominante "de 0,5 a 1 salário-mínimo", candidatos com mais de 6 pessoas na família tem renda predominante "de 0,5 a 1 salário-mínimo" e candidatos que não possuem família tem renda predominante "menos de 0,5 salário-mínimo". Logo, pode-se inferir que a quantidade de membros na família pode influenciar na situação econômica da mesma.

Utilizando o Campus Montes Claros para exemplificar o resultado obtido, temos que o cenário geral quase não é alterado, a situação de renda das famílias com 1 a 3 pessoas, 4 a 6 pessoas e mais de 6 pessoas permanecem o mesmo, alterando

Renda Bruta Familiar	Não tenho Família	1 a 3 Pessoas	4 a 6 Pessoas	Mais de 6 Pessoas
Não tem renda	<b>18</b>	326	326	42
Menos de 0,5 salário-mínimo	28	1700	3044	387
De 0,5 a 1 salário-mínimo	14	<b>3288</b>	6152	<b>691</b>
De 1 a 1,5 salário-mínimo	1	2501	<b>6315</b>	585
Mais de 1,5 salário-mínimo	6	1896	6265	416

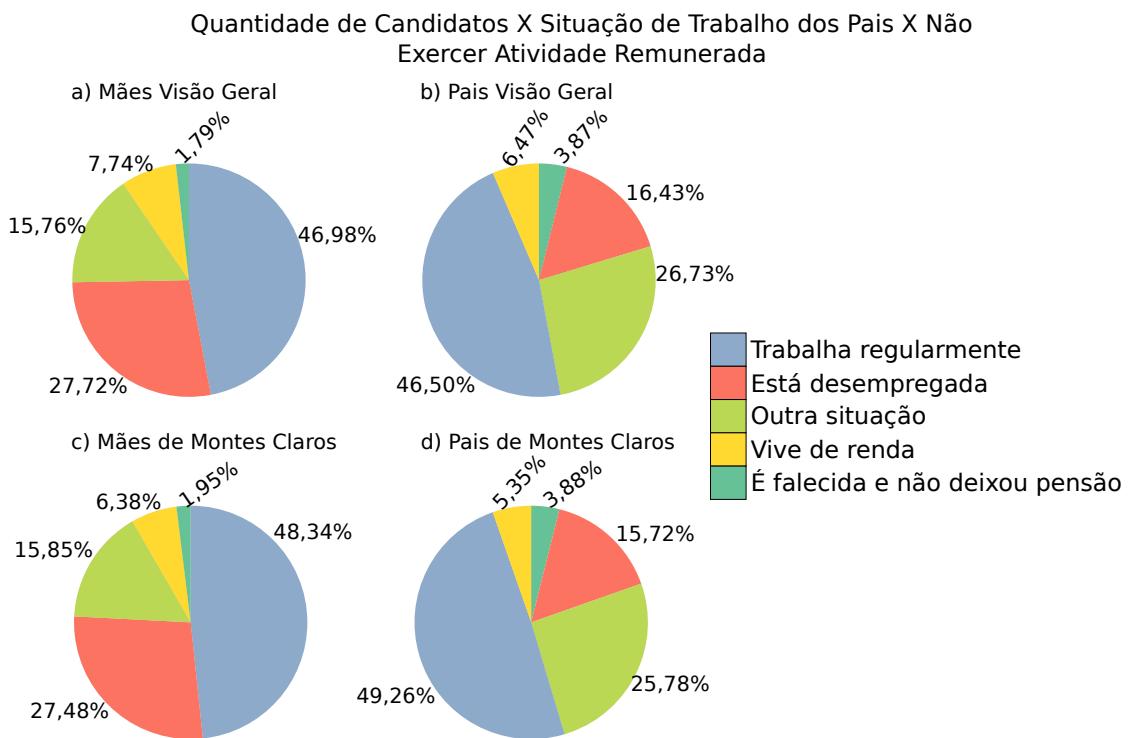
**Tabela 4.1.** Renda Bruta Familiar X Quantidade Membros na Família.

apenas os candidatos que não tem família que ao invés de ter renda predominante "menos de 0,5 salário-mínimo" tem como renda predominante "não tem renda", como pode ser observado na Tabela 4.2.

Renda Bruta Familiar	Não tenho Família	1 a 3 Pessoas	4 a 6 Pessoas	Mais de 6 Pessoas
Não tem renda	<b>7</b>	56	39	2
Menos de 0,5 salário-mínimo	5	235	476	50
De 0,5 a 1 salário-mínimo	4	<b>624</b>	1108	<b>101</b>
De 1 a 1,5 salário-mínimo	0	574	1378	99
Mais de 1,5 salário-mínimo	0	436	<b>1713</b>	84

**Tabela 4.2.** Renda Bruta Familiar X Quantidade Membros na Família X Campus Montes Claros.

Outra informação que pode ser obtida da situação da família é a condição de trabalho dos pais, podendo identificar que, os candidatos que não trabalham têm predominantemente pais que trabalham regularmente, representando 45% dos candidatos (Figuras 4.21 (a) e (b)). Este mesmo cenário se repete nos quatro *campus* com mais candidatos, mostrando que é predominante a situação dos pais trabalharem regularmente (Figuras 4.21 (c) e (d)).

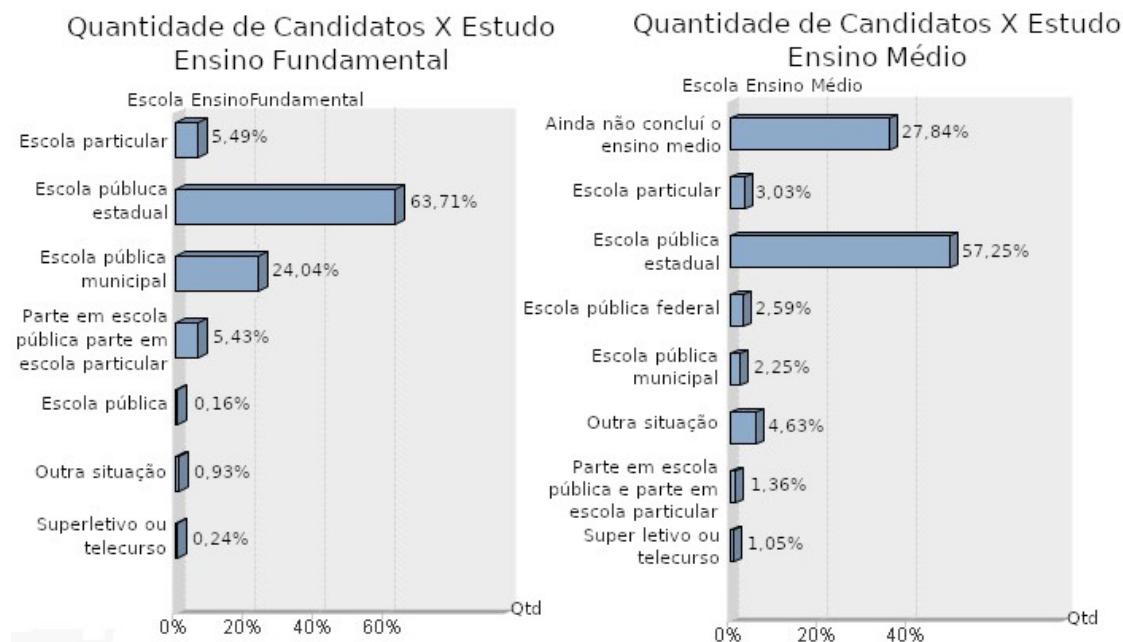


**Figura 4.21.** Quantidade Candidatos que não exercem atividade remunerada X Situação dos pais.

Sintetizando o que foi apresentado no Perfil-Econômico, pode-se perceber exercer atividade remunerada implica em qual modalidade o candidato tenta ingressar. Candidatos que exercem atividade remunerada tendem a ter idade entre 18 e 39 anos tendo também como turno de preferência o noturno, possuindo salários mais baixos na faixa de meio a um salário-mínimo, tentando então melhorar sua situação econômica obtendo uma formação acadêmica. Os candidatos que não exercem atividades remuneradas tendem a ingressar em cursos diurnos ou noturnos, tendo uma renda familiar que pode variar dependendo da região que está prestando o vestibular, possuem famílias constituídas de 4 a 6 pessoas, normalmente sustentadas pelos pais.

## 4.3 Análise Perfil-Educacional

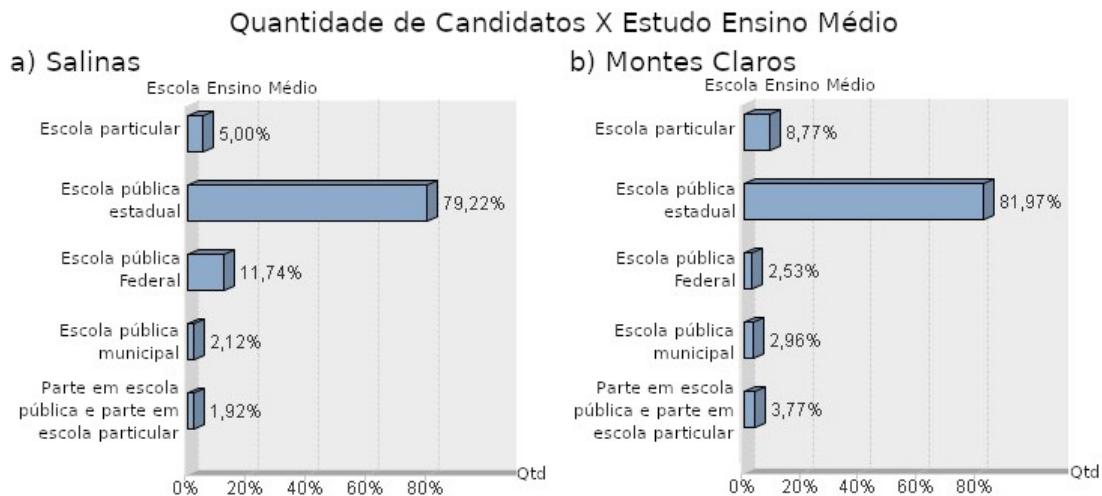
Analizando a procedência escolar do candidato pode-se perceber que a grande maioria dos candidatos são oriundos de escola pública estadual, este fato vale tanto para o estudo do candidato no ensino fundamental 63,71% quanto no ensino médio 57,25%, como pode ser observado na Figura 4.22.



**Figura 4.22.** Quantidade Candidatos X Escola que estudou no Ensino Fundamental e no Ensino Médio, Geral.

Analizando apenas os candidatos que já concluíram o ensino médio, foi feito um filtro nos quatro *campi* com mais candidatos a fim de observar se há variação nos dados, pode-se perceber que o *campus* de Salinas tem como segunda maior instituição "escolas públicas federais" 11,74%, mostrando que possivelmente os alunos que terminam algum curso técnico Integrado ao ensino médio na instituição tendem a tentar ingressar em outros cursos ( Figura 4.23 (a)). Já os *campus* de Montes Claros, Pirapora e Januária, possuem o maior público vindos de escolas públicas estaduais e o segundo maior público são alunos de escolas particulares (Figura 4.23 (b) exemplifica a situação do *Campus* Montes Claros, os demais foram ocultados

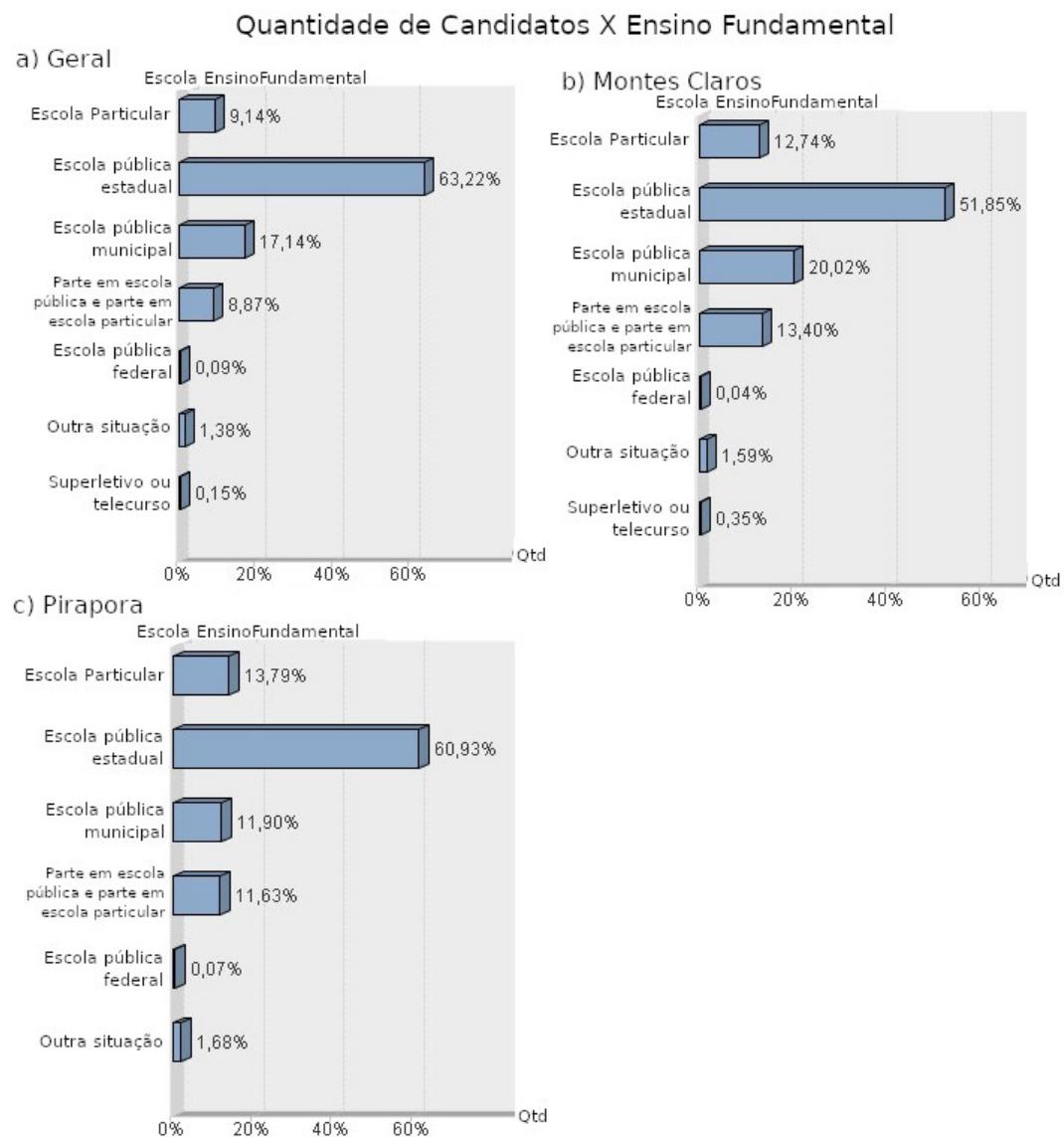
por serem análogos), este fato mostra que apesar da grande maioria dos candidatos serem de escolas públicas, quase 10% dos candidatos são oriundos de escolas particulares nestes *Campi*.



**Figura 4.23.** Quantidade Candidatos X Escola que estudou no Ensino Médio.

Observando apenas os candidatos que não concluíram o ensino médio, isto é, os candidatos a vagas dos cursos técnicos integrados e subsequentes, temos que a instituição de ensino com maior quantidade de candidatos é a escola estadual pública com 63,22% como pode ser visto na Figura 4.24 (a), a segunda instituição de ensino mais frequentada são as escolas públicas municipais com 17,14% dos candidatos, tendo em terceiro lugar as escolas particulares com 9,14%. Fazendo o filtro para os *campus* de Montes Claros, Salinas e Januária obteve-se um resultado similar, mostrando que a grande maioria dos candidatos é oriunda de escolas públicas estaduais, logo em seguida de escolas públicas municipais e em terceiro de escolas particulares (Figura 4.24 exemplifica a situação do *Campus* Montes Claros, os demais foram ocultados por serem análogos). Já os candidatos do *Campus* Pirapora tendem a estudar o ensino fundamental em escolas públicas estaduais 60,93%, e como a segunda maior instituição as escolas particulares com 13,97% (Figura 4.24 (c)).

Outra característica dos candidatos é apresentar um alto índice de reprovação durante o ensino fundamental. A pergunta "Se você repetiu alguma série no ensino

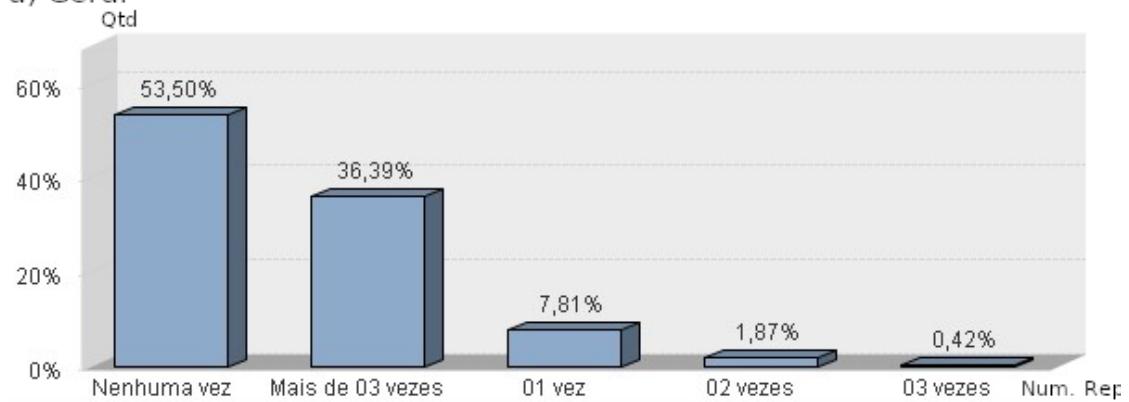


**Figura 4.24.** Quantidade Candidatos X Escola que estudou no Ensino Fundamental.

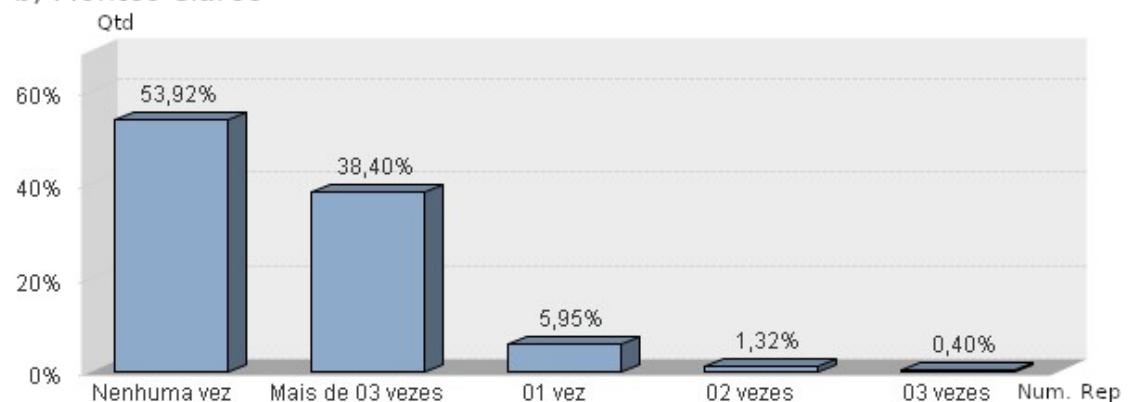
fundamental, informe o número de vezes: "tem como principal resposta não repetiu o ensino fundamental com 53,50%, mas a segunda maior resposta correspondendo a 36,39% dos candidatos apresentaram ter mais de 3 reprovações mostrando o quanto os candidatos dos *Campi* tendem a reprovar neste período de ensino (Figura 4.25 (a)). Filtrando pelos quatro *Campi* foram obtidas informações similares,

### Quantidade de Candidatos x Reprovações no Ensino Fundamental

a) Geral



b) Montes Claros

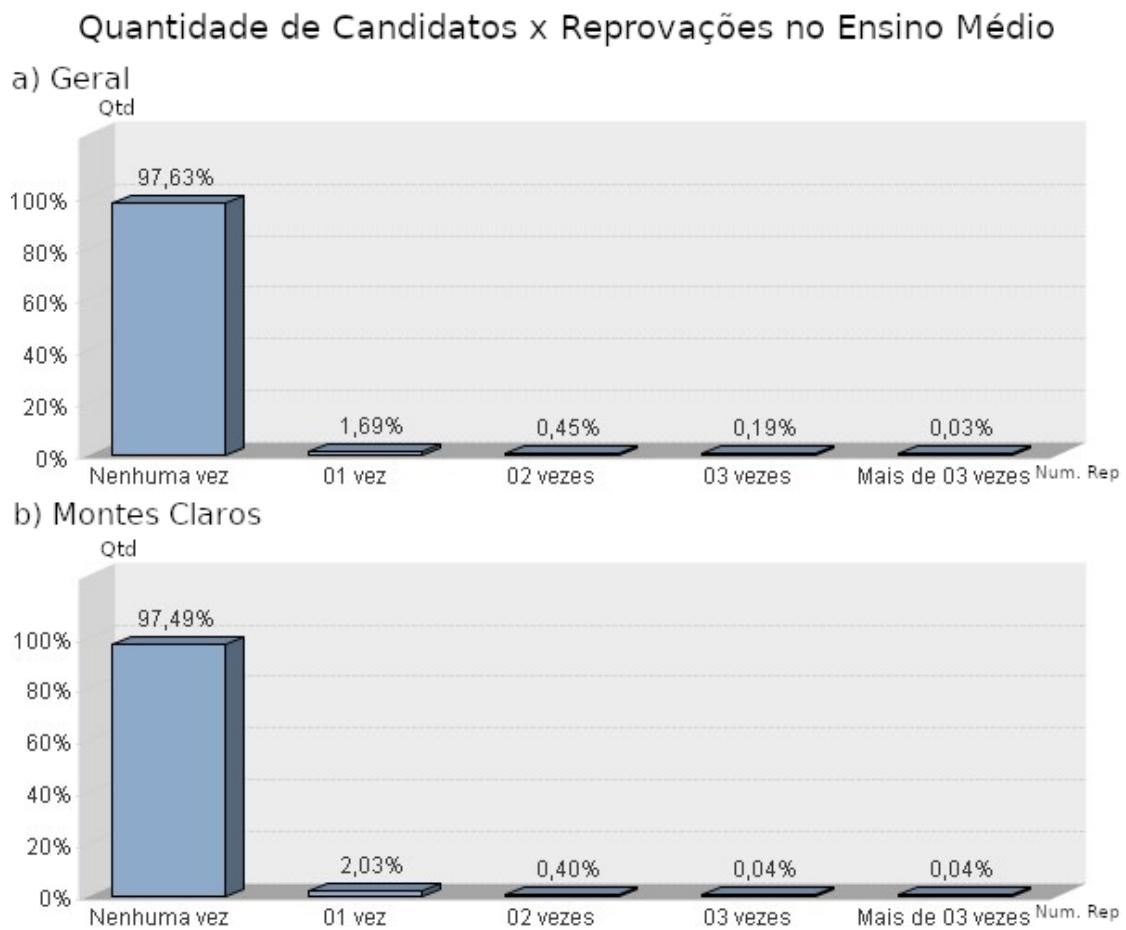


**Figura 4.25.** Quantidade Candidatos X Repetiu no Ensino Fundamental.

mostrando que independente da cidade, os candidatos do IFNMG tendem a reprovar durante o ensino fundamental (Figura 4.25 (b)), apresenta a distribuição dos candidatos pelo número de reprovações do *campus* Montes Claros, os demais foram ocultados por apresentarem valores similares).

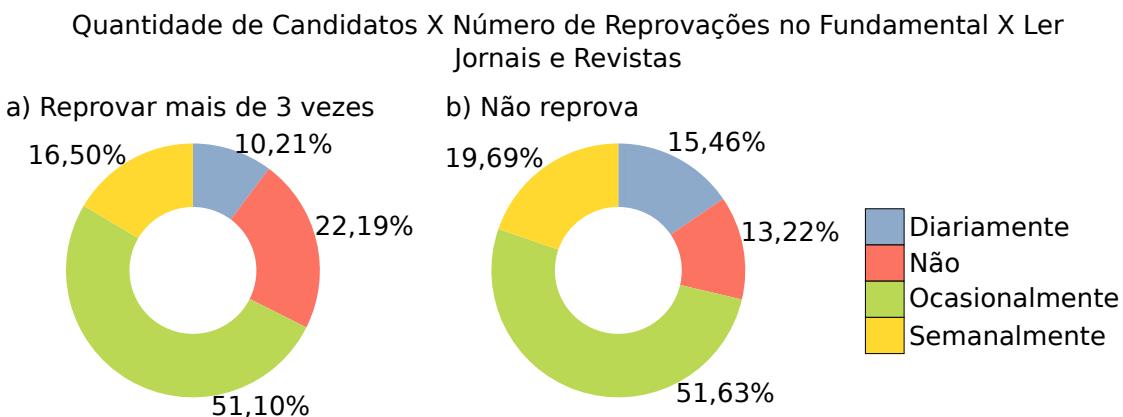
Fazendo a mesma análise de número de reprovações para os candidatos que já concluíram o ensino médio, foi observado que 87,86% dos candidatos nunca reprovaram, sendo que 9,45% dos candidatos reprovaram apenas 1 vez (Figura 4.26 (a)), mostrando que os candidatos tendem a reprovar mais durante o ensino fundamental que o ensino médio. Observando os quatro *campi* teve-se um resultado similar, mostrando que o fato de que os candidatos reprovam mais durante o ensino

fundamental do que no ensino médio se procede boa partes dos *campi* (Figura 4.26) (b) apresenta o *Campus Montes Claros* para exemplificar o estado dos quatro).

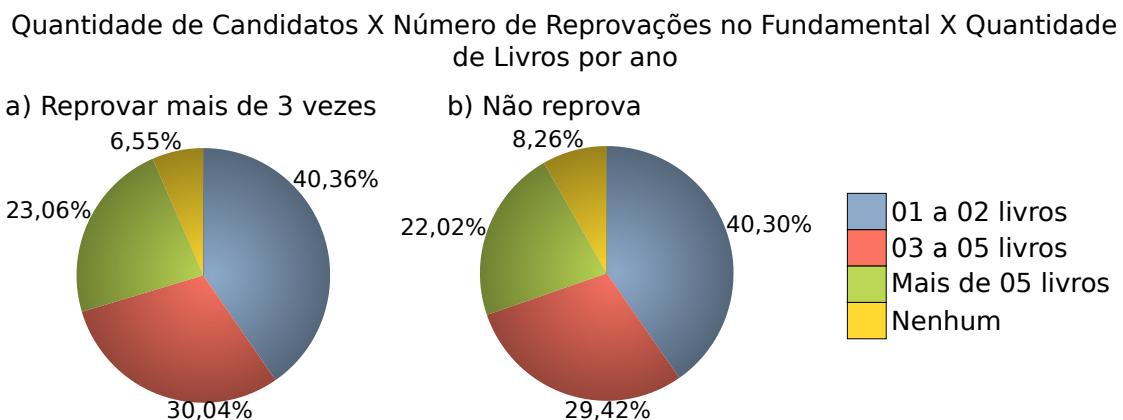


**Figura 4.26.** Quantidade Candidatos X Repetiu no Ensino Médio.

Considerando o fato de tanta reprovação no ensino fundamental, foi feito filtro selecionando os candidatos com mais de 3 reprovações no ensino fundamental e através do mesmo foi analisado se o hábito de leitura influenciou no número de reprovações, para isso foram analisadas as dimensões foram analisadas as "lerJornaisRevistas" (Figura 4.27), "qtdLivrosAno" (Figura 4.28). Comparando os resultados obtidos entre os candidatos que tiveram mais de três reprovações no ensino fundamental com os que não tiveram reprovações, pode-se concluir que o hábito de leitura não apresenta ser um fator determinante na quantidade de reprovações.



**Figura 4.27.** Quantidade Candidatos X Número de Reprovações durante o Ensino Fundamental X Ler Jornais e Revistas.



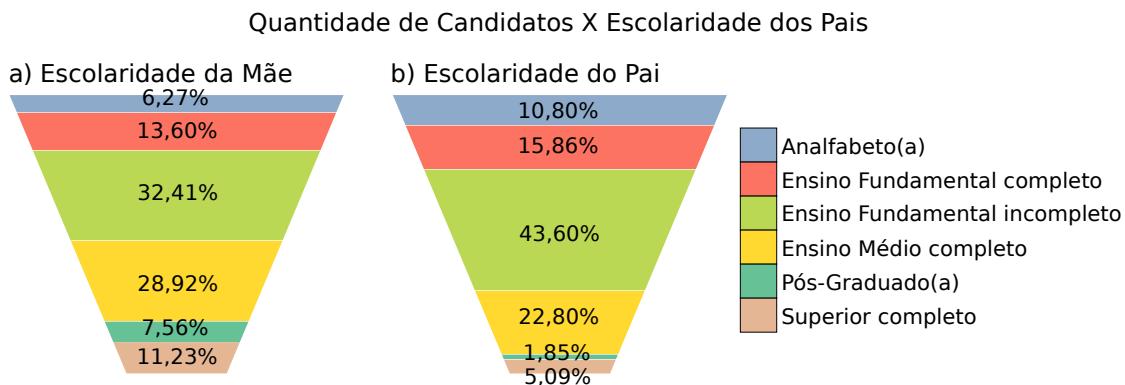
**Figura 4.28.** Quantidade Candidatos X Número de Reprovações durante o Ensino Fundamental X Quantidade Livros por Ano.

As análises realizadas no Perfil-Educacional visam identificar a procedência dos candidatos em relação ao ensino, mostrando que há uma predominância de candidatos oriundos de escolas públicas estaduais tanto no ensino médio quanto no ensino fundamental, também pode-se perceber que a quantidade de candidatos que são oriundos de escolas particulares têm pouca variação no ensino médio e no ensino fundamental. Por fim, foi analisado se o hábito de leitura influencia na quantidade de reparações durante o ensino fundamental, não havendo qualquer

diferença significativa entre aqueles que tiveram reprovação ou não em seus hábitos de leitura.

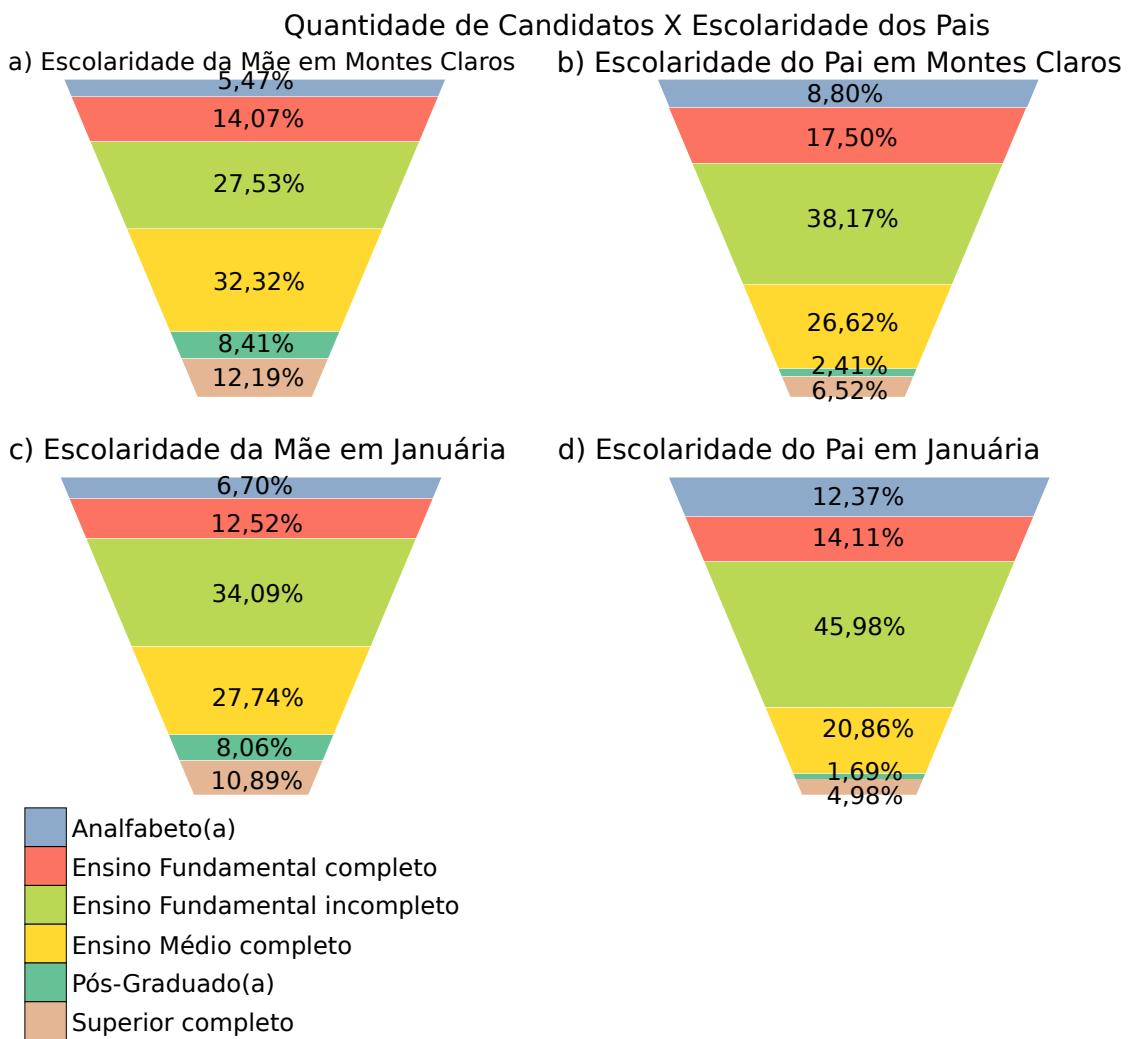
## 4.4 Análise do Perfil-Social

Foi analisada a vida escolar dos pais dos candidatos tentando identificar a influência da família em seu perfil. Estes em sua grande maioria, tem nível de escolaridade baixo, tendo como principal nível de escolaridade o ensino fundamental incompleto (Figura 4.29 (a) e (b)), onde que 32,41% das mães e 42,62% dos pais cursaram apenas o fundamental incompleto, ainda pode ser observado que a situação escolar das mães é superior que a dos pais, já que existem 4,53% menos Analfabetas, 11,19% menos mães com Ensino fundamental incompleto, 6,12% a mais que concluíram o Ensino médio, o dobro de mães concluíram o Ensino superior e o quádruplo concluíram a Pós-Graduação. Pode-se, então, afirmar que as mães tendem a estudar durante mais tempo que os pais. Segundo o IBGE em 2019, a taxa de analfabetos da região Centro-Oeste é de 5,4%, mostrando que a porcentagem de pais e mães dos candidatos do IFNMG que não tiveram educação é maior que a média da região, reforçando ainda mais importância das ofertas de vagas do IFNMG.



**Figura 4.29.** Quantidade de candidatos X Escolaridade dos pais.

Fazendo filtros nos quatro *campi* com mais candidatos para observar se há alguma variação escolar dos pais por *Campi*, pode-se perceber que em Montes Claros e Pirapora o perfil das mães é diferenciado, pois é possível perceber que as



**Figura 4.30.** Quantidade de Candidatos X Escolaridade dos pais X Campus de Montes Claros e Januária.

mães tendem a completar o Ensino médio e a quantidade de analfabetas diminui, aproximando-se da média da região, o perfil dos pais continua quase inalterado tendo apenas uma redução na quantidade de analfabetos, mas mantém a ordem de escolaridades cursadas (Figura 4.30 (a) (b)). Os *Campus* de Januária e Salinas permanecem semelhantes a visão geral onde que o ensino fundamental incompleto é predominante na escolaridade dos pais e das mães (Figura 4.30 (c) (d)).

A Tabela 4.3 mostra o local onde o candidato estudou durante o ensino médio,

e a influência da escolaridade da mãe por esta escolha.

	Escolaridade Mãe					
	Ensino Fundamental incompleto	Ensino Médio completo	Ensino Fundamental completo	Superior completo	Pós-Graduada	Analfabeta
<b>Escola particular</b>	10,70%	<b>35,15%</b>	10,06%	24,73%	18,74%	1,20%
<b>Escola pública estadual</b>	<b>40,79%</b>	24,50%	14,97%	7,08%	3,83%	8,82%
<b>Escola pública federal</b>	24,13%	<b>33,67%</b>	9,95%	16,83%	12,27%	3,15%
<b>Escola pública municipal</b>	<b>35,05%</b>	20,53%	16,62%	7,83%	2,39%	17,57%
<b>Outra situação</b>	19,60%	<b>34,80%</b>	9,96%	17,66%	16,36%	1,62%
<b>Parte em escola pública e parte em escola particular</b>	14,94%	<b>34,12%</b>	11,48%	18,87%	17,45%	3,14%
<b>Supletivo ou Telecurso</b>	<b>41,89%</b>	17,25%	14,78%	6,16%	3,29%	16,63%

**Tabela 4.3.** Situação de escolaridade da Mãe X Instituição onde o Candidato estudou durante o Ensino Médio.

Pode-se perceber que candidatos oriundos de escolas particulares tendem a ter mães que são formadas no ensino médio ou possuem um curso superior completo ou é pós-graduada. Além disso, os candidatos que estudaram parte em escola pública e parte em escola particular apresentaram um resultado semelhante na escolaridade das mães. Nesta tabela ainda pode ser observado que mães que possuem um ensino fundamental incompleto tendem a colocar seus filhos em escolas Municipais ou estaduais, a divisão das resposta deve ocorrer pela ocorrência das escolas nos diferentes municípios do Norte de Minas, sendo que a outra opção é telecursos.

Observando se escolaridade da mãe também pode influenciar a instituição em que seu filho estudou durante o ensino fundamental, os dados relacionados a esta consulta estão apresentados na Tabela 4.4.

Mesmo no ensino fundamental existe influência do nível de estudo da mãe para o local onde o candidato estudou. Uma característica dos candidatos que estudaram em escolas particulares durante o ensino fundamental é de possuir mães com o nível de escolaridade o ensino médio completo, ensino superior completo e pós-graduação estes correspondendo a 89,49% das pessoas que fazem o ensino fundamental em escolas particulares. O fato desta influência, pode ser observado também nos candidatos oriundos de parte em escolas públicas e parte em escolas

	Escolaridade Mãe					
	Ensino Fundamental incompleto	Ensino Médio completo	Ensino Fundamental completo	Superior completo	Pós-Graduada	Analfabeta
<b>Escola pública estadual</b>	<b>33,69%</b>	29,68%	14,06%	10,00%	6,14%	6,43%
<b>Escola pública municipal</b>	<b>41,48%</b>	23,54%	15,32%	7,30%	3,93%	8,43%
<b>Escola particular</b>	4,10%	<b>33,88%</b>	6,25%	29,43%	26,18	0,16%
<b>Parte em escola pública e parte em escola particular</b>	6,05%	<b>39,30%</b>	7,51%	25,07%	21,43%	0,63%
<b>Outra situação</b>	27,36%	<b>31,49%</b>	14,94%	11,26%	9,20%	5,75%
<b>Supletivo ou Telecurso</b>	<b>40,18%</b>	10,71%	18,75%	5,36%	3,57%	21,43%
<b>Escola pública federal</b>	<b>41,33%</b>	25,33%	16,00%	6,67%	6,67%	4%

**Tabela 4.4.** Situação de escolaridade da Mãe X Instituição onde o Candidato estudou durante o Ensino Fundamental.

particulares onde que as mães tendem a ter um nível de escolaridade similar ao das mães dos candidatos que estudaram em escolas particulares. Outra informação que pode ser obtida na Tabela 4.4 é de que candidatos que tendem a ingressar em escolas públicas estaduais, escolas públicas municipais, supletivo ou telecurso e em escolas públicas federais tendem possuir uma mãe que tenha o nível de ensino ensino fundamental incompleto.

Após ter analisado a influência da escolaridade da mãe foi observado se a escolaridade do pai também influenciava na escolha de qual escola o candidato cursou o ensino médio e o fundamental (Tabela 4.5).

Ao analisar os dados da Tabela 4.5 pode-se perceber que a escolaridade do pai tem influência na escolha de escola particular, mostrando que pais com ensino médio completo tendem a ter filhos que ingressam em escolas particulares isso por ser visto tanto na linha que corresponde ao estudo em escolas particulares quanto na que os candidatos fizeram uma parte do ensino médio em escolas públicas e a outra em escolas particulares. Para completar a análise foi feito o filtro que mostra se a escolha do local de estudo durante o ensino fundamental é influenciada pela escolaridade do pai (Tabela 4.6), e pode ser comprovado que assim como havia influência durante a ensino médio, o ensino fundamental apresenta um resultado similar.

	Escolaridade Pai					
	Ensino Fundamental incompleto	Ensino Médio completo	Ensino Fundamental completo	Superior completo	Pós-Graduado	Analfabeto
Escola particular	18,99%	<b>38,55%</b>	15,31%	16,38%	7,80%	2,48%
Escola pública estadual	<b>49,51%</b>	17,21%	15,57%	14,59%	2,48%	0,65%
Escola pública federal	<b>40,46%</b>	26,04%	15,34%	8,54%	6,88%	2,74%
Escola pública municipal	<b>40,50%</b>	23,59%	17,29%	15,95%	2,29%	0,29%
Outra situação	<b>33,32%</b>	31,00%	16,40%	10,98%	4,31%	3,99%
Parte em escola pública e parte em escola particular	25,47%	<b>36,48%</b>	18,08%	11,48%	3,46%	5,03%
Supletivo ou Telecurso	<b>42,51%</b>	14,99%	14,17%	3,70%	0,82%	23,82%

**Tabela 4.5.** Situação de escolaridade do Pai X Instituição onde o Candidato estudou durante o Ensino médio.

	Escolaridade Pai					
	Ensino Fundamental incompleto	Ensino Médio completo	Ensino Fundamental completo	Superior completo	Pós-Graduado	Analfabeto
Escola pública estadual	<b>45,57%</b>	21,82%	16,52%	11,06%	3,87%	1,16%
Escola pública municipal	<b>51,27%</b>	16,77%	14,70%	2,35%	0,54%	14,36%
Escola particular	12,86%	<b>41,15%</b>	12,43%	<b>21,81%</b>	11,06%	0,70%
Parte em escola pública e parte em escola particular	17,91%	<b>42,86%</b>	16,69%	14,63%	6,29%	1,62%
Outra situação	<b>42,53%</b>	24,37%	16,09%	5,29%	2,76%	8,97%
Supletivo ou Telecurso	<b>42,86%</b>	8,04%	14,29%	3,57	0,89%	30,36%
Escola pública federal	<b>42,67%</b>	20,00%	17,33%	10,67%	1,33%	8,00%

**Tabela 4.6.** Situação de escolaridade do Pai X Instituição onde o Candidato estudou durante o ensino fundamental.

As análises realizadas do Perfil-Social visam identificar se a escolaridade dos pais influenciavam na vida escolar do candidato a fim de entender o público alvo que presta vestibular. Sumarizando as análises feitas neste cubo temos que os pais têm uma baixa escolaridade sendo principalmente ensino fundamental incompleto, podendo ter pequenas mudanças em cada *Campi* como foi visto para Montes Claros e Pirapora, também foi visto que os que a quantidade de pais analfabetos está acima da média da região como um todo. Foi confirmado que a escolaridade dos pais podem influenciar na instituição de ensino que os candidatos estudaram, como foi mostrado no caso das escolas particulares.



# Capítulo 5

## Conclusão

Com o avanço tecnológico e o aumento na necessidade de informação, gestores estão cada vez mais buscando maneiras de obter informações de maneira rápida, eficiente, confiável e prática a fim de se manterem no mercado e alavancar seus negócios, se tornando cada vez mais competitivos. Instituições de ensino vem também procurando técnicas que os propicie informações que possam auxiliar nas tomadas de decisões, a fim de melhorar seu ensino, melhorar a instituição, logo, está cada vez mais frequente o uso de técnicas como *data marts* para obtenção de informações.

Com a necessidade de obtenção de informação sobre o perfil dos candidatos do IFNMG, foi então criado o ambiente de *data mart* do questionário socioeconômico, possibilitando extração de informação referentes à vida econômica, social e educacional das pessoas que prestam o processo seletivo no IFNMG.

Foi possível perceber que candidatos que tentam ingressar nos *campi* após a conclusão do ensino fundamental devem receber um atendimento especializado, já que 1/3 dos candidatos informam que tiveram reprovações durante o ensino fundamental. Com este tipo de informação os gestores podem traçar planos de ensino que visem melhorar o conhecimento de seus estudantes, auxiliando na permanência dos mesmos.

Apesar do *data mart* criado auxiliar os gestores nas tomadas de decisões, as análises referentes ao desenvolvimento ao longo dos anos foram comprometidas pelos problemas enfrentados na obtenção dos dados. Contudo, mesmos sem este

tipo de informação, as análises foram capazes de indicar o perfil dos candidato a vagas no IFNMG, sendo que para o ensino médio integrado temos candidatos que não possuem uma renda, suas famílias tem renda entre meio salário-mínimo a um e meio salário-mínimo, suas famílias são constituídas de quatro a seis pessoas, em grande maioria não apresentam deficiência e seus pais tem como grau de escolaridade ou o ensino médio completo ou o ensino fundamental incompleto. Em relação aos cursos superiores, aproximadamente 60% dos candidatos não apresentam exercer atividades remuneradas, suas famílias apresentam ter mais de 1 salário-mínimo 60%, suas famílias são formadas por três a seis pessoas, seus pais tem escolaridade ensino fundamental incompleto apresentando 40% do total e são oriundos de escola pública estadual.

## 5.1 Trabalhos Futuros

Para a realização de trabalhos futuros sugere-se que seja feito uma busca na literatura por questionários socioeconômicos, visando obter o melhor conjunto de perguntas que descreva o perfil dos candidatos que prestam vestibular do IFNMG, melhorando o questionário e retirando as ambiguidades que puderam ser observadas durante o desenvolvimento deste trabalho.

Com o novo questionário desenvolvido, é possível também armazenar as respostas dos candidatos na própria instituição, mantendo um padrão que seja estipulado através do conjunto de perguntas.

Também sugere-se que sejam feitos testes de outras ferramentas OLAP, a fim de encontrar uma que seja possível disponibilizar as informações em rede, possibilitando aos gestores fazerem suas próprias consultas em seu computadores.

# Referências Bibliográficas

- Baltzan, A. P. P. (2012). *Sistemas de informação 1. ed.* Editora Porto Alegre: AMGH.
- Barbieri, C. (2011). *BI2 Business Intelligence. Modelagem e Qualidade.* Editora Elsevier.
- Camargo, A. M. R. A. S. S. (2013). Aplicando técnicas de business intelligence sobre dados de desempenho acadêmico um estudo de caso. In *Encontro Regional dos Estudantes de Biblioteconomia. Documentação, Gestão e Ciência da Informação da Região Sul.*
- Castro Novais, R. R. d. (2012). Modelagem dimensional. Monografia (Tecnologia em Processamento de Dados), Faculdade de Tecnologia São Paulo, São Paulo, Brasil.
- Clemes, M. (2001). Data warehouse como suporte ao sistema de informações gerenciais em uma instituição de ensino superior: Estudo decaso na ufsc. Dissertação de Mestrado, Universidade Federal de Santa Catarina.
- Colangelo Filho, L. (2001). Implantação de sistemas erp (enterprise resources planning): um enfoque de longo prazo. *São Paulo: Atlas.*
- Cramer, R. (2006). Estudo analítico de ferramentas open source para ambientes olap. Monografia (Gerenciamento de Banco de Dados), Universidade do Extremo Sul Catarinense, Criciúma, Brasil.
- Damasceno, E. V. (2017). Business intelligence: Implantação no sistema do instituto federal do triângulo mineiro – iftm. Dissertação de Mestrado, Instituto Superior de Contabilidade e Administração do Porto.

- Efraim, R. S. J. E. A. D. K. T. (2009). *Business intelligence: Um enfoque gerencial para a inteligência do negócio*. Editora Bookman, Porto Alegre.
- Inmon, W. H. (2005). *Building the Data Warehouse, Fourth Edition*. John Wiley & Sons, Indianapolis, Indiana.
- Kimball, R. M. R. (2002). *The Data Warehouse Toolkit: The complete guide to dimensional modeling 2. ed.* USA: Wiley.
- Lemos, S. G. (2015). Proposta de instrumento para comparação de ferramentas de business intelligence. Monografia (Gestão da Informação), Universidade Federal do Paraná, Curitiba, Brasil.
- Politano, S. J. B. N. A. P. P. R. (2006). Análise do uso da ferramenta olap na melhoria do processo de decisão e suporte à elaboração de estratégias. In *XIII SIMPEP*.
- Popovic, R. H. P. S. C. J. J. A. (2012). *Towards business intelligence systems success: Effects of maturity and culture on analytical decision making*. Editora Elsevier.
- Rodrigues, M. F. N. (2004). *Tecnologia e projeto de Data Warehouse:uma visão multidimensional*. Editora Érica.
- Santos, L. F. D. (2010). Construção de um data mart acadêmico da uesb. Monografia (Ciência da Computação), Universidade Estadual do Sudoeste da Bahia, Vitória da Conquista, Brasil.
- Schiel, C. C. L. U. (2004). Midas-poeta - um sistema de apoio à tomada de decisão pedagógica para o ambiente portfolio-tutor. In *XIV Simpósio Brasileiro de Informática na Educação – SBIE*.
- Steidel, J. d. S. (2017). Business intelligence: Uma proposta metodológica para análise da evasão escolar em instituições federais de ensino. Dissertação de Mestrado, Universidade Federal do Paraná.