



Universidade do Minho
Escola de Ciências



RELATÓRIO DE ESTÁGIO DA LICENCIATURA EM ESTATÍSTICA APLICADA

**Desenvolvimento de um *dashboard* com
indicadores de negócio, no *SAP Analytics Cloud*,
com aplicação de análise preditiva de decisões**

Daniela Quintas Brasileiro
A92314

Ano Letivo: 2021/2022

Orientador DMAT: Maria Teresa Mesquita da Cunha
Machado Malheiro

Orientador Accenture Portugal: Carlos Jesus

NOTA DE CONFIDENCIALIDADE

O estágio realizado na Accenture Portugal exige por parte da entidade e como ética profissional, sigilo em relação aos dados, resultados e documentos dos seus clientes. Por esta razão, os dados utilizados neste relatório são fictícios.

Assim sendo, ocultou-se o acesso a documentos nos quais constam informações reservadas da Accenture Portugal de modo a não comprometer o carácter confidencial da empresa.

AGRADECIMENTOS

A toda a minha família, especialmente à minha mãe, pelo apoio demonstrado ao longo de todo o meu percurso académico.

À prof. Doutora Maria Teresa Mesquita Cunha Machado Malheiro pela orientação, paciência, disponibilidade, ajuda e partilha de conhecimento.

À Accenture Portugal por me ter possibilitado a realização do estágio.

Ao Dr. Carlos Jesus, orientador externo, que me integrou na equipa da Accenture Portugal e se mostrou sempre disponível a ajudar.

Ao professor Luís Ferreira por me impulsionar a explorar a área da programação.

A todos os meus amigos e colegas de curso, em especial à Juliana, Carina, Sara, Cristiana, Nuno e Luís Filipe, que me ajudaram na adaptação da vida académica e sempre me apoiaram no alcance dos meus objetivos. Graças a eles estes três anos foram muito especiais.

RESUMO

O presente relatório foi realizado no âmbito do Estágio da Licenciatura em Estatística Aplicada desenvolvido na Accenture Portugal. O estágio teve como objetivo principal a criação de um *dashboard* na ferramenta *SAP analytics cloud* com a aplicação do *software* R nativo da mesma.

Numa primeira parte do projeto, as tarefas foram adquirir conhecimento sobre as aplicações *SAP analytics cloud* e *Intelligent Data Quality*. Dado que a primeira ferramenta necessita de uma licença, foi feito um curso online sobre a mesma e posteriormente foi aplicado o conhecimento adquirido numa versão gratuita do *SAP analytics cloud*. A ferramenta *Intelligent Data Quality* é da autoria da Accenture Portugal e, desta forma, foi exposta pela equipa desta empresa numa apresentação.

Numa segunda parte do projeto, foi realizado um *dashboard*, no qual o principal objetivo foi a criação de indicadores de negócio (*Key Performance Indicator*). Para tal, foi disponibilizado um exemplo (documento confidencial) de um *dashboard*, de modo a criar algo semelhante.

Por fim, elaborou-se o projeto final onde se aplicou toda a aprendizagem realizada nas fases antecedentes. Esta parte do projeto é constituída por três etapas. Na primeira, recorreu-se à aplicação *Intelligent Data Quality* com a finalidade de fazer limpeza aos dados escolhidos, nos quais foi introduzido anteriormente valores incorretos para o contexto dos dados. Na segunda etapa, aplicou-se a linguagem *python*, a fim de criar um documento contendo os dados “limpos” e novas variáveis correspondentes a indicadores de negócio. Na última etapa, criou-se um *dashboard* dos dados desenvolvidos na fase anterior formada por três páginas. A primeira página corresponde a um esquema resumindo o projeto final, a segunda página inclui a análise exploratória dos dados e a última página apresenta os KPI's (*Key Performance Indicator*).

ABSTRACT

This report was carried out within the scope of the Internship of the Degree in Applied Statistics developed at Accenture Portugal. The main objective of the internship was to create a dashboard in the SAP analytics cloud tool with the application of its native R software.

In the first part of the project, the tasks were to acquire knowledge about SAP analytics cloud and Intelligent Data Quality applications. Since the first tool requires a license, an online course was taken about it, and the knowledge acquired was later applied in a free version of SAP analytics cloud. The Intelligent Data Quality tool is owned by Accenture Portugal and, therefore, was exposed by the Accenture Portugal team in a presentation.

In the second part of the project, a dashboard was created, in which the main objective was the creation of business indicators (Key Performance Indicator). For this matter, an example of a dashboard (confidential document) was made available, in order to create something similar.

Finally, the final project was elaborated where all the learning carried out in the previous phases was applied. This part of the project consists of three stages. In the first one, the Intelligent Data Quality application was used in order to clean the chosen data, in which incorrect values were previously introduced for the data context. In the second stage, the python language was applied in order to create a document containing the “clean” data and new variables corresponding to business indicators. In the last step, a dashboard was created from the data developed in the previous phase, consisting of 3 pages. The first page corresponds to a scheme summarizing the final project, the second page includes the exploratory analysis of the data and the last page presents the KPIs (Key Performance Indicator).

SIGLAS

- *SAC - SAP analytics cloud*
- *KPI - Key Performance Indicator*
- *CSV - comma-separated values*
- *IDQ – Intelligent Data Quality*

ÍNDICE GERAL

NOTA DE CONFIDENCIALIDADE.....	iii
AGRADECIMENTOS.....	iv
RESUMO.....	v
ABSTRACT.....	vi
SIGLAS.....	vii
ÍNDICE GERAL.....	viii
ÍNDICE DE FIGURAS.....	ix
ÍNDICE DE TABELAS.....	xi
1 Introdução	1
1.1 Sobre Accenture Portugal.....	1
1.2 Funcionamento do Estágio Curricular	2
1.3 Objetivos	2
1.4 Estrutura do Relatório.....	3
2 Ferramentas Estatísticas	4
3 Conhecimento da aplicação SAC.....	6
3.1 Réplica de KPI's	11
4 Conhecimento da ferramenta IDQ.....	14
5 Projeto final	18
5.1 Aplicação do Intelligent Data Quality	20
5.2 Criação de um <i>excel</i> aplicando a linguagem <i>python</i>	27
5.3 Criação de um <i>dashboard</i>	30
5.3.1 Primeira Página	30
5.3.2 Segunda Página.....	30
5.3.3 Terceira página	40
CONCLUSÕES.....	48
REFERÊNCIAS	49

ÍNDICE DE FIGURAS

Figura 1- Arquitetura Accenture. [1]	1
Figura 2- Áreas de negócio da Accenture. [1]	2
Figura 3- Primeira página da história.....	8
Figura 4- Segunda página da história	9
Figura 5- Gráfico de barras e respetivo controlo.	10
Figura 6- Controlo para toda a página.....	10
Figura 7- Criação de gráfico com a ferramenta SAC em linguagem R	11
Figura 8- Réplica de KPI's.....	12
Figura 9- Código para a construção do gráfico de barras.....	13
Figura 10- Continuação do código para a construção do gráfico de barras	13
Figura 11- Aplicação Intelligent Data Platform [5].....	14
Figura 12- Analyze Results no IDQ [5].	15
Figura 13- Continuação Analyze Results no IDQ [5].....	15
Figura 14- Funcionalidade Cleanse no IDQ [5].	16
Figura 15- Continuação da funcionalidade Cleanse no IDQ [5].....	17
Figura 16- Esquema/resumo das etapas do projeto final.....	20
Figura 17-Path criado no IDQ	22
Figura 18-Projetos criados no IDQ	22
Figura 19- Importação dos dados para o IDQ	23
Figura 20- Aplicação das regras no IDQ.....	23
Figura 21- Criação de Regra no IDQ	24
Figura 22- Código da aplicação da linguagem python	27
Figura 23- Segunda página do dashboard	31
Figura 24- Filtro por cidade	32
Figura 25- Filtro por cidade	32
Figura 26- Gráfico de barras para a frequência do valor total de vendas	33
Figura 27- Código utilizado para a criação do gráfico da figura 26.....	33
Figura 28- Gráfico circular para a percentagem por tipo de pagamento.....	34
Figura 29- Código para a criação do gráfico da figura 28	34
Figura 30- Caixa de bigodes para a quantidade adquirida por género	35
Figura 31- Código para a criação da representação gráfica da figura 30	35

Figura 32- Série Temporal para o lucro	36
Figura 33- Tabela para as medidas por linha de produto	36
Figura 34- Gráfico de barras para a quantidade e lucro por medida selecionada	37
Figura 35-Representação gráfica das correlações	38
Figura 36- Código para criação da representação gráfica da figura 35.....	38
Figura 37- Indicadores	39
Figura 38- Terceira página do dashboard	40
Figura 39- Gráfico de barras para o estatuto da falha por categoria de produto	41
Figura 40- Código para a criação da representação gráfica da figura 39.....	41
Figura 41- Gráfico de barras da percentagem de falhas por filial	42
Figura 42- Código utilizado para gerar o gráfico da figura 41.....	42
Figura 43- Código em R para gerar o gráfico da figura 43	43
Figura 44- Gráfico relativo à percentagem de falhas por tipo de produto.....	43
Figura 45- Gráfico circular para a percentagem de casos por estatuto.....	44
Figura 46- Código gerador do gráfico circular da figura 45	44
Figura 47- Indicadores de falhas por cidade	45
Figura 48- Gráfico de barras para o número de falhas por Cidade para cada Estatuto	46
Figura 49- Caixas de bigode com a distribuição do lucro por Estatuto	46
Figura 50- Código gerador das caixas de bigodes da figura 49	47
Figura 51- Gráfico circular para a percentagem de falhas por regras	47

ÍNDICE DE TABELAS

Tabela 1- Primeiras linhas dos dados exemplo	7
Tabela 2- Explicação das variáveis da base de dados	18
Tabela 3- Primeiras linhas dos dados originais	19
Tabela 4- Primeiras linhas da base de dados contendo valores incorretos	21
Tabela 5- Primeiras linhas dos dados com as observações que aprovaram as regras	25
Tabela 6- Primeiras linhas dos dados contendo as observações que reprovaram as regras.....	26
Tabela 7- Primeiras linhas do excel criado com auxílio do python	29

1 Introdução

O estágio realizado na Accenture Portugal foi proposto no âmbito da Licenciatura em Estatística Aplicada da Universidade do Minho, sendo realizado sob a orientação da Professora Maria Teresa Mesquita da Cunha Machado Malheiro (orientadora interna) e do Dr. Carlos Jesus (orientador externo).

1.1 Sobre Accenture Portugal

A Accenture é uma organização global de serviços profissionais, líder em capacidades digitais, *cloud* e *security*. Encontra-se há mais de 30 anos em Portugal contando com mais de 3600 colaboradores, mais de 260 projetos realizados, mais de 94 clientes, 66% das empresas do *PSI20 Index* como clientes e 22 das 100 maiores empresas são clientes da Accenture. Os escritórios da Accenture Portugal estão localizados em Lisboa, Braga, Miraflores, Porto e Aveiro.

Na figura 1 está presente a arquitetura de inovação da Accenture, na qual podemos ver que a estratégia utilizada por esta empresa é apoiar os clientes a se transformarem e reinventarem. Além disso, aposta na requalificação dos seus profissionais e num negócio cada vez mais sustentável.



Figura 1- Arquitetura Accenture. [1]

A Accenture Portugal gera emprego qualificado e investe na formação de talento com a finalidade de desenvolver soluções inovadoras.

Combina uma experiência sem paralelo com uma forte especialização em mais de 40 setores de atividade como podemos observar na figura 2. Oferece uma ampla gama de serviços em estratégia e consultoria, interatividade, tecnologia e

operações, suportada pela maior rede mundial de centros de tecnologias avançadas e operações inteligentes [1].



Figura 2- Áreas de negócio da Accenture. [1]

1.2 Funcionamento do Estágio Curricular

O estágio foi realizado em regime misto (presencial e *online*), e decidiu-se em conjunto com o orientador externo realizar-se presencialmente às quartas-feiras e quintas-feiras. Houve *daily scrum* (reuniões diárias com duração de 15 minutos com a finalidade de discutir o trabalho realizado durante o dia) e demonstrações/apresentações semanais (reunião semanal com duração de 30 minutos para apresentação do desenvolvimento do estágio realizado durante a semana).

1.3 Objetivos

De seguida, apresenta-se os objetivos do estágio indicados pela Accenture, tendo em consideração que são comuns aos dois estágios curriculares designados para esta proposta. Ao longo de todo o relatório irá explicar-se como se procedeu à realização dos mesmos.

- Estruturar um modelo de dados em CSV;
- Realizar o *upload* desse modelo na ferramenta SAC (*SAP analytics cloud*);
- Desenvolver um *dashboard* com KPI's, filtros e tabelas no SAC;
- Apresentar um modelo de *predict anaysis / Forecast* nativo do SAC;
- Desenvolver um KPI(*Key Performance Indicator*) com a componente “R” nativo do SAC;

- Criação de uma *script* em Python para análise e aplicação de regras ao ficheiro *CSV*, obtendo um *output* em *CSV* para posterior aplicação ao *SAC*;
- Realizar uma DEMO final.

1.4 Estrutura do Relatório

O presente relatório encontra-se dividido em cinco capítulos, incluindo este primeiro onde se faz uma breve apresentação da empresa Accenture Portugal. Nesta capítulo, também se aborda o funcionamento do estágio e os principais objetivos do mesmo.

O segundo capítulo é direcionado para as ferramentas/*softwares* utilizados no desenvolvimento do estágio.

O terceiro capítulo é direcionado para o conhecimento de uma nova ferramenta (*SAP analytics cloud*), como também a exploração das funcionalidades desta aplicação.

O quarto capítulo corresponde à apresentação do *IDQ*, ferramenta da autoria da Accenture, posto que se irá recorrer a esta ferramenta posteriormente.

No quinto capítulo, são apresentados os objetivos do projeto final de estágio, bem como as etapas desenvolvidas para o cumprimento destes objetivos.

2 Ferramentas Estatísticas

Neste capítulo será feita uma breve introdução aos *softwares* utilizados, bem como expor alguns conceitos importantes para a realização deste estágio.

- **R Studio**

O *software R Studio* permite a programação em linguagem R para computação estatística assim como a execução de gráficos. De modo geral, esta ferramenta auxilia na exploração de dados (manipulação, análise e visualização de dados) de modo a possibilitar fazer uma análise estatística destes.

Atualmente a linguagem R tem diversas aplicações em várias áreas, tendo como exemplo a ciência de dados, *machine learning* e estatística computacional [2].

- **SAC (SAP Analytics Cloud)**

O *software SAC* permite a análise de dados gráficos em tempo real e a criação de KPIs mais importantes das organizações. Esta ferramenta possibilita uma melhor análise e visualização da gestão de orçamentos para apoiar as empresas na tomada de decisão [3].

- **Visual Studio Code**

O *Visual Studio Code* é um editor de código, aberto e desenvolvido pela Microsoft. Este editor suporta diversas linguagens de programação. Todavia, apenas se irá recorrer a esta ferramenta ao longo do estágio para a criação de código em linguagem *Python* [4].

- **IDQ (Intelligent Data Platform)**

O *IDQ* é uma das soluções existentes em *Intelligent Data Platform* (plataforma da autoria da Accenture com acesso reservado). Nesta ferramenta é possível fazer limpeza dos dados, ou seja aplicar um conjunto de regras já definidas nesta solução, com a finalidade de visualizar quais os dados que não estão em conformidade com as regras para posteriormente proceder à sua limpeza. Estas regras são geradas automaticamente pelo *IDQ* para que os utilizadores possam escolher e aplicar aos dados em causa. A seguir de aplicadas as regras são expostos os dados que não estão corretos (não se encontram em conformidade com as regras escolhidas) e o utilizador pode aplicar o “cleasing com AI” para propor valores corretos [5].

- **KPI (Key Performance Indicator)**

KPI é um conceito que irá ser referido diversas vezes ao longo deste relatório. A criação destes indicadores permite medir se uma ação ou um conjunto de ações irá ao encontro dos objetivos das organizações. Assim sendo, é importante a elaboração de indicadores para ter uma melhor

percepção relativamente aos resultados obtidos através da estratégia aplicada pela organização/empresa [6].

3 Conhecimento da aplicação SAC

Com o propósito de aprender a utilizar a aplicação *SAP Analytics Cloud*, assistiu-se ao curso *OpenSAP* (disponível em [Intelligent Decisions with SAP Analytics Cloud | openSAP](#)), constituído por cinco semanas de aulas gravadas nas quais cada semana tinha a duração de 2-3 horas. Nesta fase inicial, utilizou-se o *SAP Analytics Cloud trail* com a intuito de aplicar o conhecimento adquirido no curso a um conjunto de dados sobre o comércio de bicicletas, conjunto designado doravante Exemplo. O *SAP Analytics Cloud trail* é uma versão com licença gratuita de 20 dias do *SAP Analytics Cloud*. Por essa razão, não tem todas as funcionalidades disponíveis.

Com a finalidade de explorar na prática o conhecimento adquirido no curso *OpenSAP* e representar a diversidade de gráficos possíveis de gerar através desta aplicação, importou-se os dados do Exemplo para a aplicação *SAC trail* e criou-se uma história constituída por duas páginas (figuras 3 e 4). Nesta fase, não é relevante a informação que se pode retirar dos gráficos, apenas se foca na sua criação.

Na tabela 1 estão representadas as primeiras linhas da base de dados utilizada nesta fase de modo a se obter uma melhor compreensão do tipo de dados manipulados.

Ambas as páginas, representadas nas figuras 3 e 4, da história referida anteriormente são interativas. Na figura 3 que apresenta uma Análise de Dados sobre o comércio de Bicicletas, é possível ver que existem dois controlos: Controlo para toda a página (canto superior direito) e Controlo do gráfico de barras (na coluna à direita, ao lado do gráfico de barras). É possível criar o Controlo para toda a página (canto superior direito), que altera todas as representações gráficas presentes nessa página. Aí, é possível selecionar um dos tipos de segmento de cliente e todos os gráficos serão alterados de forma a que só sejam representados os dados referentes ao tipo(s) de segmento selecionado(s) (ver figura6).

Tabela 1- Primeiras linhas dos dados exemplo

Order ID	Date	Sales Agent Last Name	Sales Agent First Name	Customer	Customer Segment	Country	Latitude	Longitude	Customer Status	Product	Product Type	No Customer Meetings	Units Sold	Order Value
2040	May-19	Klassen	Marc	Retina Cyclist	Wholesale	Anguilla	18.220554	-63.068615	Current Customer	R300 Bike	Racing	32	369	14856
3281	May-19	Pan	Oliver	Clark Cycles	Wholesale	Liberia	6.428055	-9.429499	Prior Customer	R300 Bike	Racing	14	194	5631
3934	May-19	Pinto	Brain	Trail Blazer NE	Wholesale	Palau	7.51498	134.58252	Current Customer	R300 Bike	Racing	29	107	2323
2975	May-19	Narula	Tommy	Speed Red	Wholesale	Hong Kong	22.396428	114.109497	Prospect	R300 Bike	Racing	14	106	4793
3567	May-19	Petrie	Maddy	Send Cycle	Wholesale	Solomon Islands	-9.64571	160.156194	Current Customer	R200 BIKe 4s	Racing	24	331	13140
5549	May-19	Habib	Paul	Formal Cyclist	Wholesale	Uzbekistan	41.377491	64.585262	Current Customer	R200 BIKe 4s	Racing	28	296	9835
3182	May-19	Yun	Benny	Rides Essentials	Wholesale	Paraguay	-23.442503	-58.443832	Prior Customer	R200 BIKe 4s	Racing	19	287	7952
5596	May-19	Schilling	Michael	Motoshop AB	Wholesale	Saint Lucia	13.909444	-60.978893	Current Customer	R200 BIKe 4s	Racing	28	369	7819
3660	May-19	Lauzon	Isabella	Cycle Score	Direct	Iraq	33.223191	43.679291	Current Customer	R200 BIKe 4s	Racing	10	274	5191
2361	May-19	Tailor	Roy	Cycle Genius	Retail	Grenada	12.262776	-61.604171	Prospect	R200 BIKe 4s	Racing	8	108	3713
1453	May-19	Deng	Sarah	Sports Essence	Retail	El Salvador	13.794185	-88.89653	Prior Customer	R200 BIKe 4s	Racing	14	73	3511
1860	May-19	Bremerich	Thea	Mystic Bikes NF	Retail	Comoros	-11.875001	43.872219	Prior Customer	R200 BIKe 4s	Racing	7	72	2575

O Controlo do gráfico de barras (na coluna à direita, ao lado do gráfico de barras), permite escolher uma das três medidas (variáveis quantitativas) existentes nos dados. Deste modo, quando seleccionada uma das medidas no controlo (assinalado na figura 5), o gráfico irá automaticamente utilizar os dados dessa variável. Para um melhor entendimento, observe-se que na figura 3 o gráfico de barras é exibido para os dados referentes à variável **NO Customer Meetings**. Ao passo que na figura 5, este gráfico está representado para os dados da variável **Order Value**. Isto deve-se ao facto de a variável para a qual o gráfico é criado ser a variável seleccionada no controlo.

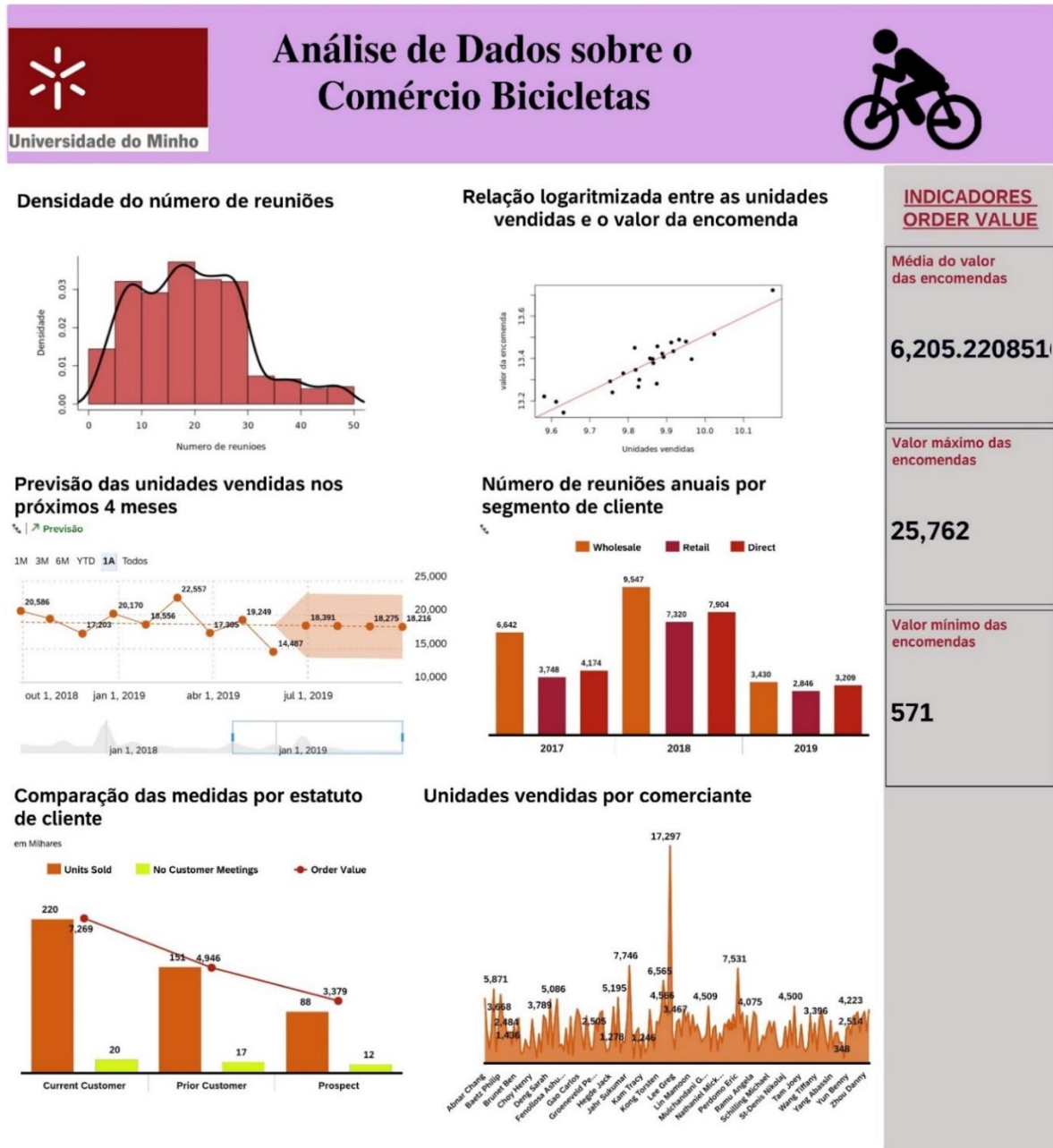


Figura 4- Segunda página da história

Medidas por modelo de bicicleta

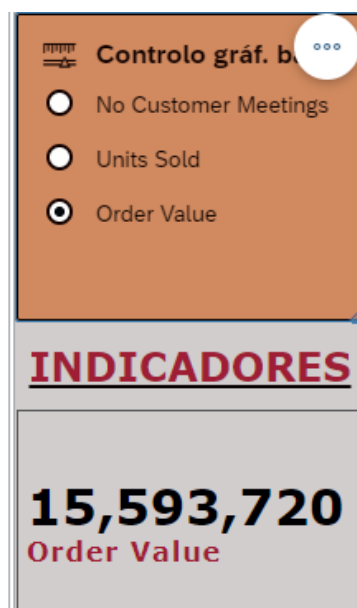
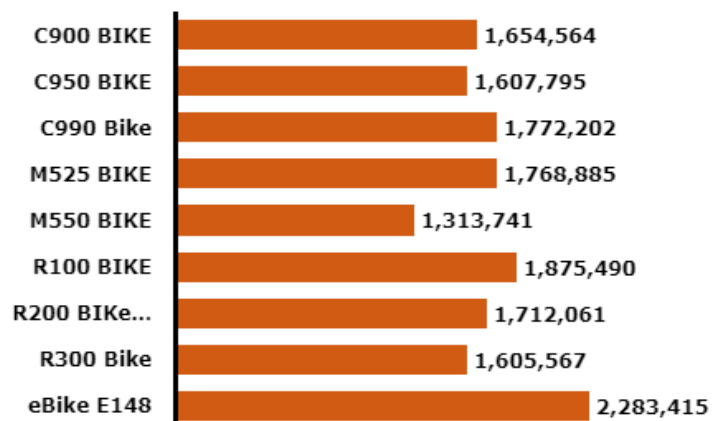


Figura 5- Gráfico de barras e respetivo controlo.

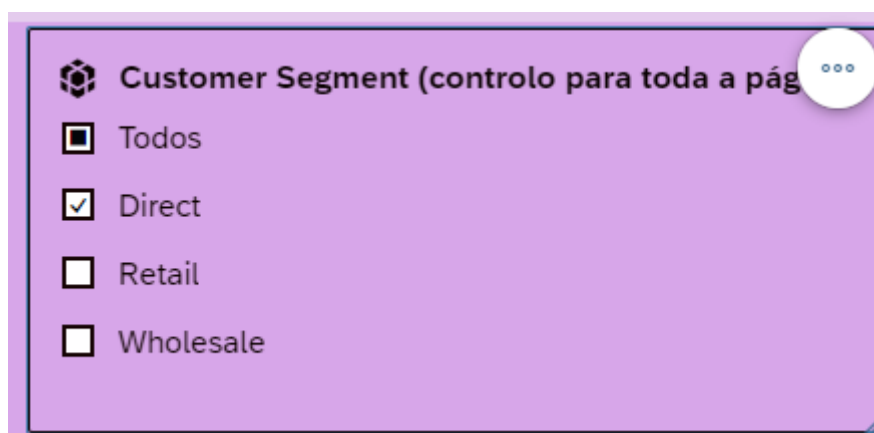


Figura 6- Controlo para toda a página

Uma das funcionalidades do SAC é a elaboração de gráficos através da ferramenta R. Exemplo disso é a representação gráfica da figura 3 na qual se observa 3 caixas de bigodes. Para tal, foi possível criar uma *script* no SAC na qual a escrita utilizada é muito similar à escrita em R (figura 7).

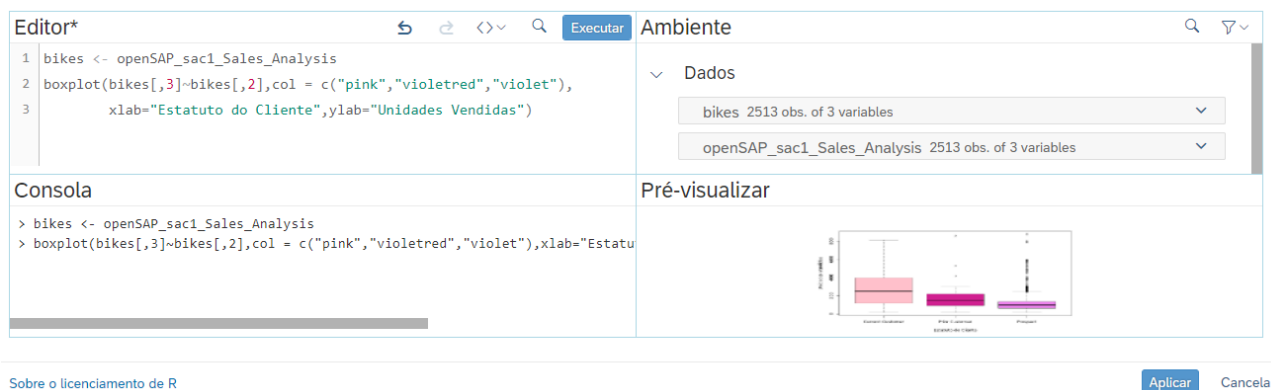


Figura 7- Criação de gráfico com a ferramenta SAC em linguagem R

3.1 Réplica de KPI's

Com o intuito de replicar um *dashboard* com KPI's foi disponibilizado pela Accenture um documento exemplo de um cliente que utilizou a ferramenta *Power BI* (aplicação da Microsoft Windows) na criação desse documento. Deste modo, criou-se o *dashboard* presente na figura 8 com o auxílio da aplicação SAC. Visto que o objetivo seria a semelhança dos indicadores presentes no *dashboard*, utilizou-se mais uma vez os dados Exemplo (fictícios). Consequentemente, os KPI's representados na figura 8 são igualmente fictícios.

Para ser possível a criação dos KPI's referidos anteriormente criou-se três novas variáveis distintas denominadas de **estatuto**. A variável **estatuto** do número de reuniões categoriza-se como "MAU" para as observações com número de reuniões inferior a 10, classificou-se como "OK" quando o número de reuniões se encontra entre 10 e 25 e determinou-se como sendo "BOM" nas observações com número de reuniões superior a 25.

De igual modo criou-se a variável **estatuto** do número de encomendas categoriza-se como "MAU" nas observações em que o valor das encomendas é inferior a 2000, definiu-se "OK" quando o valor das encomendas se encontra entre 2000 e 8000. E por fim determinou-se "BOM" nas observações em que o valor das encomendas é superior a 8000.

Por último criou-se a variável **estatuto** das unidades vendidas categoriza-se como "MAU" as observações com número de unidades vendidas inferior a 150, classifica-se como "OK" as observações com valor das unidades vendidas entre 150 e 400. E por último considera-se "BOM" as observações com valor das unidades vendidas superior a 400.

Estas variáveis foram criadas com o auxílio da ferramenta R existente na aplicação SAC, como se vê um exemplo na figura 9.

Ponto de situação atual



Figura 8- Réplica de KPI's

Como referido anteriormente, na aplicação SAC é possível a criação de controlos com a finalidade de tornar a apresentação interativa e alterar automaticamente variáveis/medidas nos gráficos ou indicadores. Na figura 8, visualiza-se dois controlos para obter os indicadores das situações críticas e favoráveis para cada variável presente no conjunto de dados.

Para a realização dos gráficos de barras presentes na figura 8 recorreu-se à ferramenta R. Nas figuras 9 e 10 observa-se o código utilizado para a criação do último gráfico de barras da figura 8, sendo que se recorreu a códigos semelhantes para a criação das restantes representações gráficas.


```

Editor*
1 encomendas <- openSAP_sac1_Sales_Analysis$'Order Value'
2 library(dplyr)
3 novo_data <- mutate(openSAP_sac1_Sales_Analysis, estatuto=case_when(encomendas<2000~"MAU",
4                               encomendas>=2000 & encomendas<8000~ "OK",
5                               encomendas>=8000~"BOM"))
6 attach(novo_data)
7 estatuto <- novo_data$'estatuto'
8 library(ggplot2)
9 freq <- novo_data%>%
10 group_by(novo_data$'Customer Segment', estatuto)%>%
11 summarise(n=n())%>%
12 mutate(fre=n/sum(n)*100)%>%
13 ungroup()
14 tipo <- freq$'novo_data$"Customer Segment"'
15 library(scales)

```

Figura 9- Código para a construção do gráfico de barras

```

Editor
1 attach(dados_SAC_FINAL)
2 library(ggplot2)
3 library(dplyr)
4 freq<- dados_SAC_FINAL%>%
5 group_by(dados_SAC_FINAL$'Product line', Estatuto)%>%
6 summarise(n=n())%>%
7 mutate(fre=n/sum(n)*100)%>%
8 ungroup()
9 tipo <- freq$'dados_SAC_FINAL$"Product line"'
10 library(scales)
11 ggplot(freq, aes(x=tipo, y=fre, fill=Estatuto, label=paste(fre, "%")))+
12 geom_col()+
13 geom_text(aes(label=paste(round(fre, 1), "%"), position=position_stack(vjust=0.5)))+
14 labs(y="Porcentagem", x="Categoria do Produto")+
15 scale_fill_manual(values=c("gold", "brown2", "yellowgreen"))+
16 coord_flip()+
17 theme(legend.position="top")+
18 geom_hline(yintercept=80, linetype="dashed", size=1)+
19 geom_segment(aes(x=2.5, y=70, xend=2.5, yend=80), size=1, arrow=arrow(length=unit(4, "mm")))+
20 geom_text(aes(x=2.5, y=66), label="80%")

```

Figura 10- Continuação do código para a construção do gráfico de barras

4 Conhecimento da ferramenta IDQ

O IDQ é uma das três soluções presentes na aplicação *Intelligent Data Platform*, plataforma criada pela equipa da Accenture. Para um melhor conhecimento do IDQ assistiu-se a uma apresentação desta ferramenta criada pela Accenture. Nesta aplicação, é possível atribuir regras pré-definidas no IDQ a uma base de dados com o intuito de verificar se os dados cumprem as regras aplicadas e, caso contrário, proceder à limpeza destes.

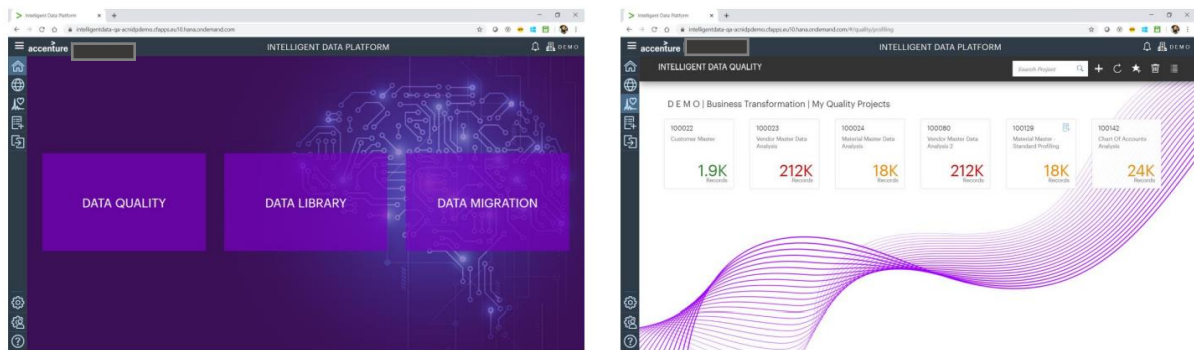


Figura 11- Aplicação *Intelligent Data Platform* [5].

Na figura 11 (lado esquerdo) observa-se as 3 soluções presentes na aplicação *Intelligent Data Platform*. No entanto, apenas se abordou a opção *Data Quality* na qual é possível criar e partilhar projetos conforme está apresentado no lado direito da figura 11.

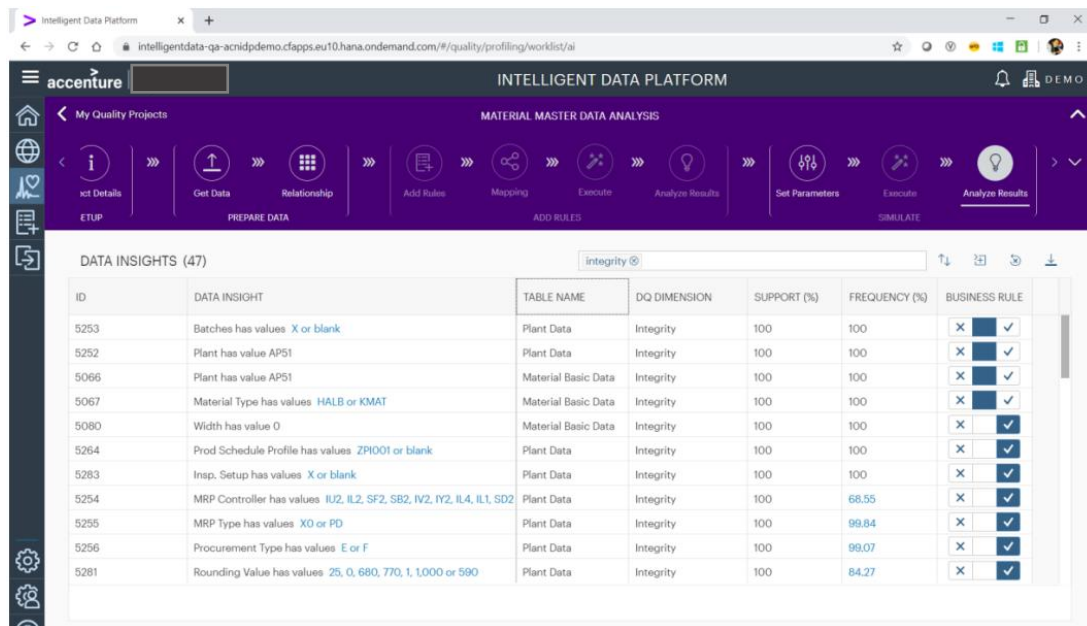


Figura 12- Analyze Results no IDQ [5].

Na figura 12, observa-se a aplicação de regras a uma base de dados. Consta-se que algumas destas não foram aprovadas, caso em que a frequência não é 100%. Exemplo disso é a última regra presente na imagem 12, na qual determina que a variável *Rounding Value* apenas poderia ter observações com 25, 0, 680, 770, 1, 1000 ou 590. Sendo que a frequência destes valores é igual a 84.27, significa que há observações nesta variável diferentes dos valores referidos.

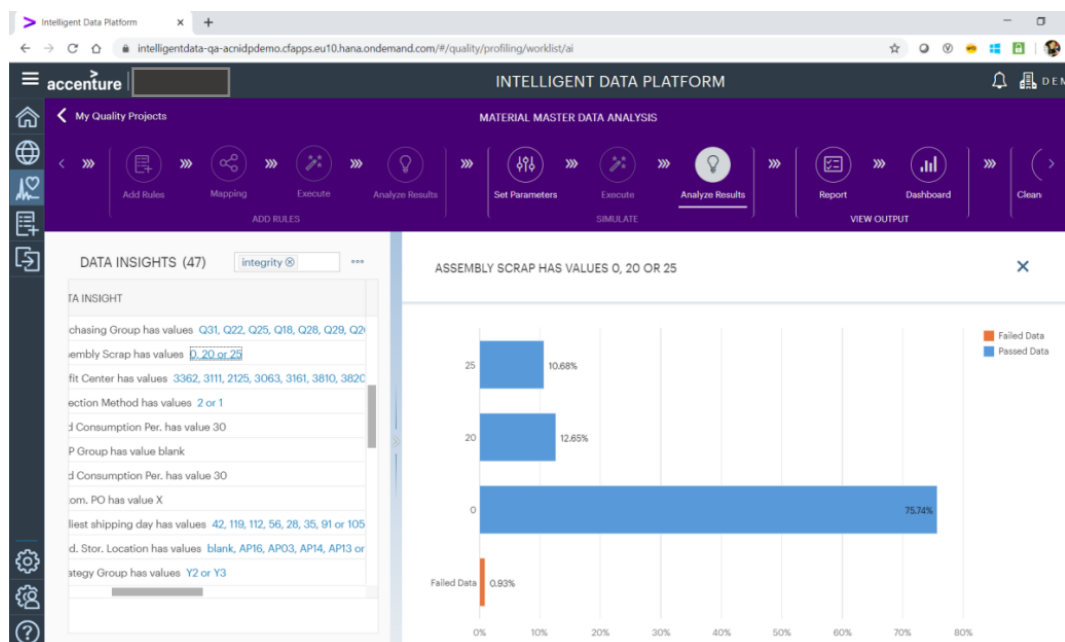


Figura 13- Continuação Analyze Results no IDQ [5].

No *IDQ* é possível criar de forma automática algumas representações gráficas que permitem visualizar melhor a quantidade de dados que são reprovados ou aprovados por cada regra. Na figura 13, observa-se a cor laranja a percentagem de dados que não respeitavam a regra selecionada e a cor azul os dados que respeitavam a regra.

Nas figuras 14 e 15 está retratado a opção *cleanse* presente no *IDQ*. Com esta opção é possível limpar/corrigir as observações que não cumprem as regras aplicadas. Sendo que é possível fazer limpeza aos dados de tal modo que cumpram todas as regras aplicadas.

Pode-se também exportar documentos em formato *CSV* ou *XLSX* contendo as regras aplicadas, com os dados na qual as regras foram aprovadas bem como os dados onde as regras falharam.

Material Number	Plant	Batches	MRP Controller	MRP Type	Procurement Type	Spec Procurem Co...	Special Procureme...	Consu
22048214	AP51	X	R55	Z5	F			2
22064619	AP51	X	R52	Z5	F			2
22072383	AP51	X	R51	Z5	F			2
22073431	AP51	X	IV4	ND	F			2
22076686	AP51	X	R52	ND	F			2
22076876	AP51	X	R52	Z5	F			2
22076861	AP51	X	IL1	Z5	E			2
22079137	AP51	X	R52	Z5	E			2
22079547	AP51	X	R56	ND	F			2
22084828	AP51	X	R52	Z5	F			2
22080633	AP51	X	ILJ2	Z5	F		30	2

Figura 14- Funcionalidade Cleanse no IDQ [5].

Material Number	Plant	Batches	MRP Controller	MRP Type	Procurement Type	Spec Procurem Co...	Special Procureme...	Consu
22048214	AP51	X	R55	PD	F			2
22064619	AP51	X	R52	PD	F			2
22072383	AP51	X	R51	PD	F			2
22073431	AP51	X	IV4	XO	F			2
22078686	AP51	X	R52	XO	F			2
22078878	AP51	X	R52	PD	F			2
22078861	AP51	X	IL1	XO	E			2
22079137	AP51	X	R52	PD	E			2
22079547	AP51	X	R58	XO	F			2
22084828	AP51	X	R52	XO	F			2
22080633	AP51	X	IU2	XO	F		30	2

Figura 15- Continuação da funcionalidade Cleanse no IDQ [5].

Na figura 14 observa-se um exemplo de observações nas quais a regra selecionada (relativa à coluna preenchida a laranja) falhou, esta figura foi retirada do documento exibido na apresentação do IDQ. De igual modo, na figura 15 vizualiza-se um exemplo de observações em que a regra selecionada (relativa à coluna a verde) não falhou.

No capítulo seguinte, esta ferramenta será aplicada com os dados escolhidos para o projeto final e será melhor explicado a sua utilização. De tal forma que será possível obter uma melhor percepção da mesma.

5 Projeto final

Para o desenvolvimento dos objetivos propostos pelo orientador da Accenture considerou-se pertinente selecionar dados relacionados com negócio, isto é, dados nos quais exista a relação de cliente e comerciante. Deste modo, escolheu-se uma base de dados de acesso livre (disponível em <https://www.kaggle.com/datasets/aungpyaeap/supermarket-sales?resource=download>) sobre um supermercado.

O conjunto de dados é um dos históricos de vendas da empresa de supermercados que registou dados relativos a três filiais diferentes durante três meses. A tabela 2 faz uma breve explicação das variáveis. Além disso, na tabela 3 estão representadas as primeiras linhas desta base de dados.

Tabela 2- Explicação das variáveis da base de dados

Variáveis	Contextualização
Invoice ID	Número de identificação da fatura.
Branch	Filial do supermercado (existem 3 filiais identificadas por A,B e C).
City	Cidade onde a filial se encontra localizada (as 3 cidades existentes são Yangon, Mandalay e Naypyitaw).
Customer type	Tipo de cliente, registado como <i>Members</i> para clientes com cartão de membro e <i>Normal</i> para clientes sem cartão de membro.
Gender	Género do cliente.
Product line	Linha de produto: categorização geral dos itens (<i>Electronic accessories, Fashion accessories, Food and beverages, Healthy and beauty, Home and lifestyle e Sports and travel</i>)
Unit price	Preço de cada produto em dólares.
Quantity	Quantidade de produtos adquiridos pelo cliente.
Tax 5%	Taxa de imposto de 5% para a compra do cliente.
Total	Preço total incluindo impostos.
Date	Data da compra (registo disponível de janeiro a março de 2019).
Time	Horário da compra (10h às 21h)
Payment	Tipo de pagamento utilizado pelo cliente na compra (existem 3 métodos de pagamento – <i>Cash, Credit card e Ewallet</i>)
Cogs	Custo dos produtos vendidos.
Gross margin percentage	Percentagem de margem bruta.
Gross income	Renda bruta- lucro.
Rating	Classificação da experiência de compra do cliente numa escala de 1 a 10.

Tabela 3- Primeiras linhas dos dados originais

Invoice ID	Branch	City	Customer type	Gender	Product line	Unit price	Quantity	Tax 5%	Total	Date	Time	Payment	cogs	gross margin percentage	gross income	Rating
750-67-8428	A	Yangon	Member	Female	Health and beauty	74.69	7	26.1415	548.9715	1/5/2019	1:08:00 PM	Ewallet	522.83	4.761904762	26.1415	9.1
226-31-3081	C	Naypyitaw	Normal	Female	Electronic accessories	15.28	5	3.82	80.22	3/8/2019	10:29:00 AM	Cash	76.4	4.761904762	3.82	9.6
631-41-3108	A	Yangon	Normal	Male	Home and lifestyle	46.33	7	16.2155	340.5255	3/3/2019	1:23:00 PM	Credit card	324.31	4.761904762	16.2155	7.4
123-19-1176	A	Yangon	Member	Male	Health and beauty	58.22	8	23.288	489.048	1/27/2019	8:33:00 PM	Ewallet	465.76	4.761904762	23.288	8.4
373-73-7910	A	Yangon	Normal	Male	Sports and travel	86.31	7	30.2085	634.3785	2/8/2019	10:37:00 AM	Ewallet	604.17	4.761904762	30.2085	5.3
699-14-3026	C	Naypyitaw	Normal	Male	Electronic accessories	85.39	7	29.8865	627.6165	3/25/2019	6:30:00 PM	Ewallet	597.73	4.761904762	29.8865	4.1
355-53-5943	A	Yangon	Member	Female	Electronic accessories	68.84	6	20.652	433.692	2/25/2019	2:36:00 PM	Ewallet	413.04	4.761904762	20.652	5.8
315-22-5665	C	Naypyitaw	Normal	Female	Home and lifestyle	73.56	10	36.78	772.38	2/24/2019	11:38:00 AM	Ewallet	735.6	4.761904762	36.78	8
665-32-9167	A	Yangon	Member	Female	Health and beauty	36.26	2	3.626	76.146	1/10/2019	5:15:00 PM	Credit card	72.52	4.761904762	3.626	7.2

No projeto final, tem-se como objetivo utilizar as ferramentas *Intelligent Data Quality*, *Visual Studio Code* com a linguagem *python* e *SAP Analytics Cloud*. No esquema representado na figura 16, visualiza-se um resumo das etapas a realizar de modo a cumprir todos os objetivos.

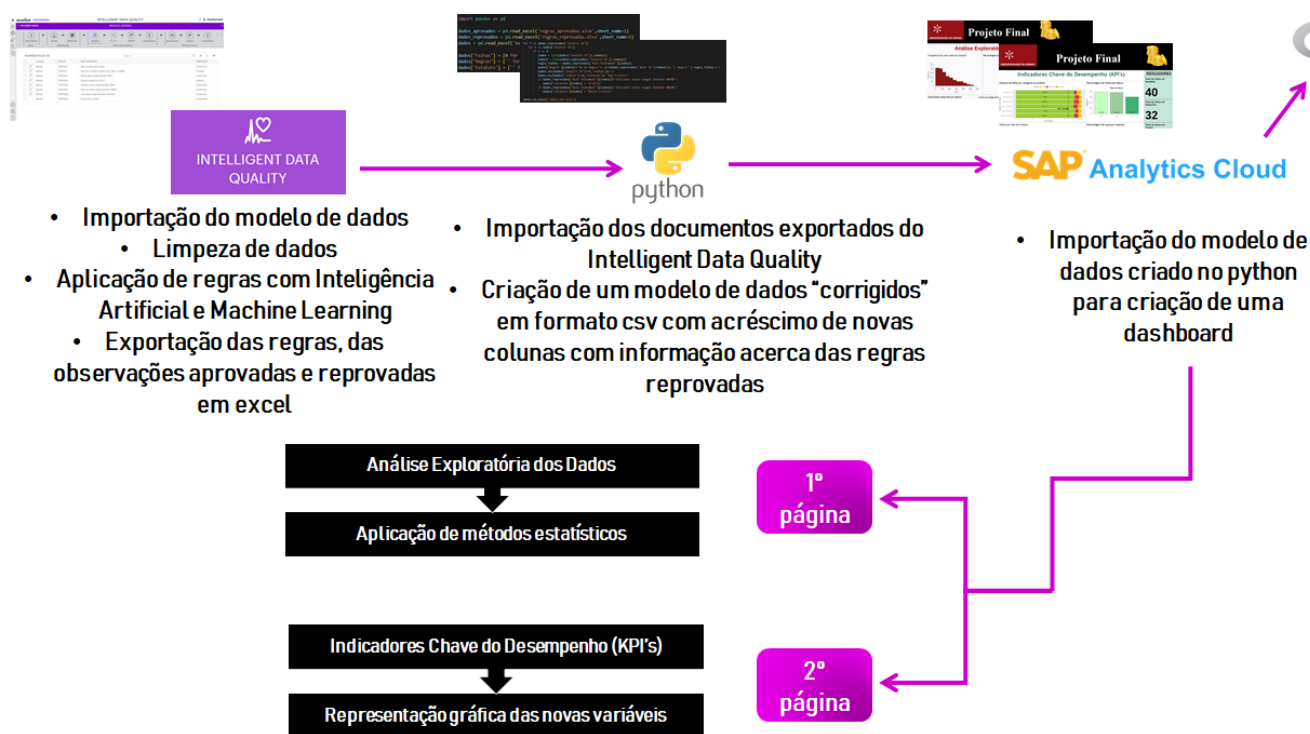


Figura 16- Esquema/resumo das etapas do projeto final

5.1 Aplicação do Intelligent Data Quality

Para esta etapa decidiu-se introduzir alguns dados incorretos na base de dados referida anteriormente uma vez que o objetivo é fazer limpeza de dados e não seria interessante ter todos os dados corretos. Além disso, criou-se uma nova variável denominada de *Discount* na qual estaria o valor da encomenda com um desconto aplicado. Por exemplo, foi aplicada a fórmula $=[@Total]-([@Total]*0.1)$ no *excel* a uma dada observação de modo a se obter o valor da encomenda com 10% de desconto.

Na tabela 4 pode-se ver as primeiras linhas da base de dados referida, na qual estão assinalados alguns dos valores incorretos que foram introduzidos nestas observações.

Tabela 4- Primeiras linhas da base de dados contendo valores incorretos

Invoice ID	Branch	City	Customer type	Gender	Product line	Unit price	Quantity	Tax 5%	Total	Date	Time	Payment	cogs	Rating	Discount
750-67-8428	A	Yangon	Member	Female	Health and beauty	74.69	7	26.1415	548.9715	1/5/2019	13:08:00	Ewallet	522.83	9.1	494.0744
226-31-3081	C	Naypyitaw	Normal	Female	Electronic accessories	15.28	5	3.82	80.22	3/8/2019	10:29:00	Cash	76.4	9.6	72.198
631-41-3108	A	Yangon	Normal	Male	Home and lifestyle	46.33	-7	16.2155	340.5255	3/3/2019	13:23:00	Credit card	324.31	7.4	306.473
123-19-1176	A	Yangon	Member	Male	Health and beauty	58.22	8	23.288	489.048	1/27/2019	20:33:00	Ewallet	465.76	8.4	440.1432
373-73-7910	A	Yangon	Normal	Male	Sports and travel	86.31	7	30.2085	634.3785	2/8/2019	10:37:00	Ewallet	604.17	5.3	570.9407
699-14-3026	D	Naypyitaw	Normal	Male	Electronic accessories	85.39	7	29.8865	627.6165	3/25/2019	18:30:00	Ewallet	597.73	4.1	564.8549
355-53-5943	A	Yangon	Member	Female	Electronic accessories	68.84	6	20.652	433.692	2/25/2019	14:36:00	Ewallet	413.04	5.8	390.3228
315-22-5665	C	Naypyitaw	Normal	Female	Home and lifestyle	73.56	10	36.78	772.38	2/24/2019	11:38:00	Ewallet	735.6	8	695.142
665-32-9167	A	Yangon	Member	Female	Health and beauty	36.26	2	3.626	76.146	1/10/2019	17:15:00	Credit card	72.52	7.2	68.5314

Começamos por criar um *Path* partilhado pelas duas estagiárias deste estágio (eu e a Carina) de modo a se poder visualizar simultaneamente os passos executados no IDQ ao conjunto de dados referidos na tabela 3. Na figura 17, pode-se observar o *Path* criado no IDQ e na figura 18 os projetos desenvolvidos. O projeto no qual estão inseridos os passos seguintes denomina-se de Projeto_Estágio (figura 18).

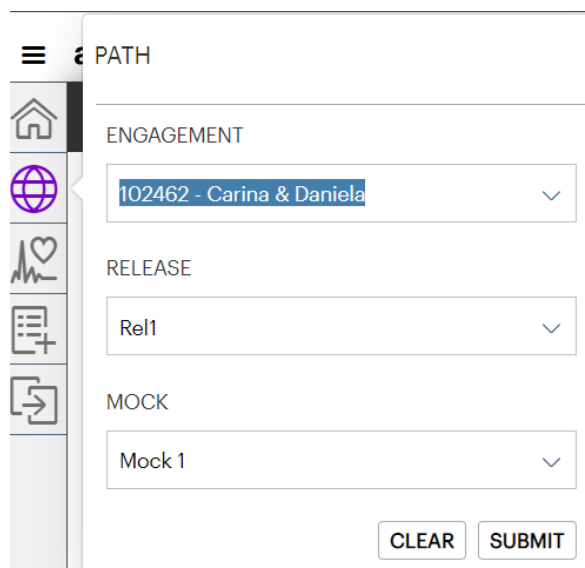


Figura 17-Path criado no IDQ

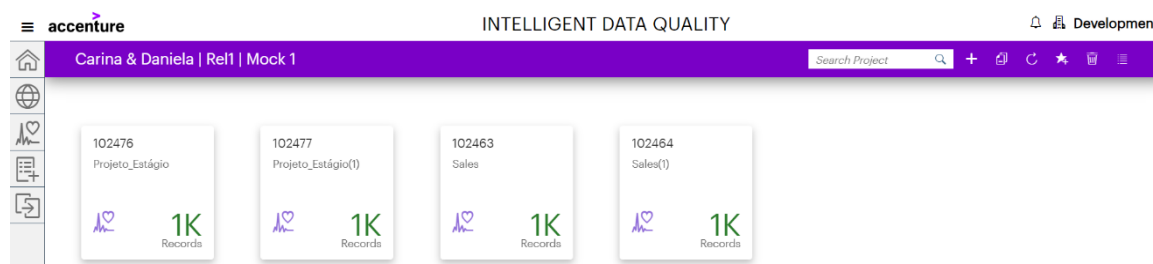


Figura 18-Projetos criados no IDQ

Na figura 19, está ilustrada a importação dos dados presentes na tabela 3 para o IDQ. Neste passo, selecionou-se todas as observações para a posterior aplicação das regras.

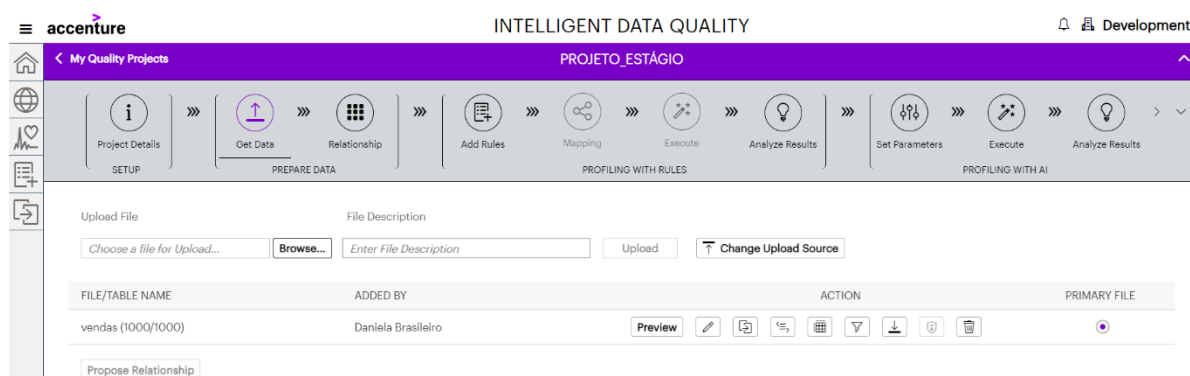


Figura 19- Importação dos dados para o IDQ

Na figura 20, vê-se as regras aplicadas aos dados. Frisando o facto de as regras estarem pré-definidas na aplicação, classificadas em vários tipos dependendo da regra escolhida (observa-se isto, na figura 20, através da coluna denominada de DIMENSION). Deste modo, apenas se seleccionou as regras que estão em conformidade com os dados personalizando os valores a que se aplicam.

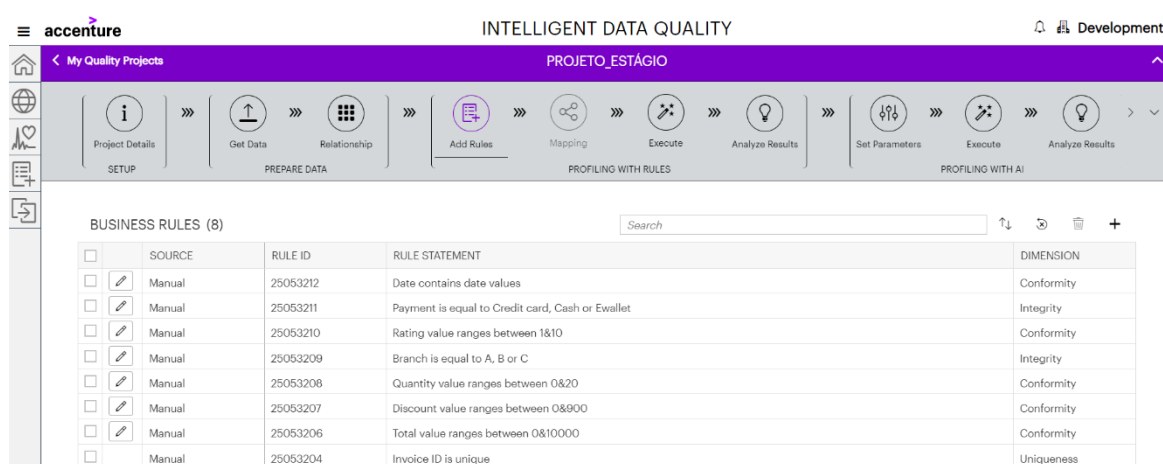


Figura 20- Aplicação das regras no IDQ

Com a finalidade de obter uma melhor compreensão relativamente à aplicação das regras, observa-se na figura 21 um exemplo de criação de uma regra no IDQ. Esta regra foi aplicada à variável **Branch** determinando que esta só pode ter valores do tipo A, B ou C, visto que estas são as filiais existentes.

NEW RULE ×

DQ Dimension: Integrity

DQ Dimension Type: INTE1 | Field has certain value(s)

Table: vendas

Branch

equal

A ⊗ B ⊗ C ⊗

ADD RULE

Figura 21- Criação de Regra no IDQ

Depois de aplicadas as regras exportou-se do IDQ dois ficheiros em formato *excel*, sendo um destes com as observações aprovadas pelas regras aplicadas e outro com as observações não aprovadas. As primeiras linhas presentes nestes ficheiros estão representadas nas tabelas 5 e 6, respetivamente.

Na tabela 5, observa-se os valores indicados a verde onde a respetiva regra (**Rule Statement** - “Total value ranges between 0&10000”) foi validada. Do mesmo modo, na tabela 6 estão assinalados a vermelho alguns valores para os quais a regra “Discount value ranges between 0&900” (indicada pela variável **Rule Statement**) não foi validada.

Tabela 5- Primeiras linhas dos dados com as observações que aprovaram as regras

Rule Id	Rule Statement	Invoice ID	Branch	City	Customer type	Gender	Product line	Unit price	Quantity	Tax 5%	Total	Date	Time	Payment	cogs	Rating	Discount
25053206	Total value ranges between 0&10000	750-67-8428	A	Yangon	Member	Female	Health and beauty	74.69	7	26.1415	548.9715	1/5/2019	13:08:00	Ewallet	522.83	9.1	494.07435
25053206	Total value ranges between 0&10000	226-31-3081	C	Naypyitaw	Normal	Female	Electronic accessories	15.28	5	3.82	80.22	3/8/2019	10:29:00	Cash	76.4	9.6	72.198
25053206	Total value ranges between 0&10000	631-41-3108	A	Yangon	Normal	Male	Home and lifestyle	46.33	-7	16.2155	340.5255	3/3/2019	13:23:00	Credit card	324.31	7.4	306.47295
25053206	Total value ranges between 0&10000	123-19-1176	A	Yangon	Member	Male	Health and beauty	58.22	8	23.288	489.048	1/27/2019	20:33:00	Ewallet	465.76	8.4	440.1432
25053206	Total value ranges between 0&10000	373-73-7910	A	Yangon	Normal	Male	Sports and travel	86.31	7	30.2085	634.3785	2/8/2019	10:37:00	Ewallet	604.17	5.3	570.94065
25053206	Total value ranges between 0&10000	699-14-3026	D	Naypyitaw	Normal	Male	Electronic accessories	85.39	7	29.8865	627.6165	3/25/2019	18:30:00	Ewallet	597.73	4.1	564.85485

Tabela 6- Primeiras linhas dos dados contendo as observações que reprovaram as regras

Rule Id	Rule Statement	Invoice ID	Branch	City	Customer type	Gender	Product line	Unit price	Quantity	Tax 5%	Total	Date	Time	Payment	cogs	Rating	Discount
25053207	Discount value ranges between 0&900	457-12-0244	C	Naypyitaw	Member	Female	Sports and travel	35.22	6	10.566	221.886	3/14/2019	13:49:00	Ewallet	211.32	6.5	-4.43772
25053207	Discount value ranges between 0&900	226-34-0034	B	Mandalay	Normal	Female	Electronic accessories	13.78	4	2.756	57.876	1/10/2019	11:10:00	Ewallet	55.12	9.0	-1.15752
25053207	Discount value ranges between 0&900	321-49-7382	B	Mandalay	Member	Male	Sports and travel	88.31	1	4.4155	92.7255	2/15/2019	17:38:00	Credit card	88.31	5.2	-1.85451
25053207	Discount value ranges between 0&900	397-25-8725	A	Yangon	Member	Female	Health and beauty	39.62	9	17.829	374.409	1/13/2019	17:54:00	Credit card	356.58	6.8	-7.48818
25053207	Discount value ranges between 0&900	431-66-2305	B	Mandalay	Normal	Female	Electronic accessories	88.25	9	39.7125	833.9625	2/15/2019	20:51:00	Credit card	794.25	7.6	-16.67925
25053207	Discount value ranges between 0&900	825-94-5922	B	Mandalay	Normal	Male	Sports and travel	25.31	2	2.531	53.151	3/2/2019	19:26:00	Ewallet	50.62	7.2	-1.06302

5.2 Criação de um *excel* aplicando a linguagem *python*

O objetivo desta etapa é a criação de um novo ficheiro *excel* com a finalidade de posteriormente ser importado para o SAC permitindo a criação de um *dashboard* utilizando a base de dados criada.

Este novo ficheiro será constituído pelas variáveis representadas na tabela 2, nas quais se utilizou as observações corretas/limpas. Além disso, fez-se a adição de três novas variáveis denominadas por **Falhas**, **Regras** e **Estatuto**.

```
1 import pandas as pd
2
3 dados_aprovedos = pd.read_excel('regras_aprovedas.xlsx',sheet_name=1)
4 dados_reprovados = pd.read_excel('regras_reprovadas.xlsx',sheet_name=1)
5 dados = pd.read_excel('dados_originais.xlsx')
6
7 dados["Falhas"] = [0 for i in dados['Invoice ID']]
8 dados["Regras"] = ['' for i in dados['Invoice ID']]
9 dados['Estatuto'] = ['' for i in dados['Invoice ID']]
10
11 for i in dados['Invoice ID']:
12     l=0
13     for f in dados_reprovados['Invoice ID']:
14         if i == f:
15             index = list(dados['Invoice ID']).index(i)
16             index2 = list(dados_reprovados['Invoice ID']).index(f,l)
17             l = index2 + 1
18             regra_falhou = dados_reprovados['Rule Statement'][index2]
19             dados['Regras'][index] += ("ID da Regra:" + str(dados_reprovados['Rule Id'][index2]) + ", Regra:" + regra_falhou + '| ')
20             dados.loc[dados['Invoice ID']==f,'Falhas']+= 1
21             dados.loc[dados['Falhas']==0,'Estatuto'] = "Nao Critico"
22             if dados_reprovados['Rule Statement'][list(dados_reprovados['Invoice ID']).index(f)]!='Discount value ranges between 0&900':
23                 dados['Estatuto'][index] = 'Critico'
24             if dados_reprovados['Rule Statement'][list(dados_reprovados['Invoice ID']).index(f)]=='Discount value ranges between 0&900':
25                 dados['Estatuto'][index] = "Muito Critico"
26
27 dados.to_excel('dados_SAC.xlsx')
```

Figura 22- Código da aplicação da linguagem *python*

Na figura 22 observa-se todo o código elaborado para a criação do novo ficheiro *excel* denominado de dados_SAC.

Através da linha 1, importou-se a biblioteca *pandas* (uma biblioteca *Python* para análise de dados) visto que oferece ferramentas vantajosas para o desenvolvimento do restante código.

Nas linhas 3, 4 e 5 importou-se os dados com as observações validadas pelas regras (tabela 5), as observações que não satisfizeram as regras (tabela 6) e os dados originais (tabela 3) respetivamente.

Nas linhas 7,8 e 9, acrescentou-se três variáveis aos dados originais denominadas de **Falhas**, **Regras** e **Estatuto**.

Da linha 11 à linha 25 elaborou-se um ciclo *for* que irá contar quantas regras não foram satisfeitas por cada observação inserindo esse valor na variável **Falhas**, além disso irá adicionar à variável **Regras** qual a regra que não foi validada, caso isto aconteça. E por último, inserir na variável **Estatuto** as palavras “Muito Crítico”, caso a regra que não é validada seja “Discount value

ranges between 0&900”, ou seja, caso a variável **Discount** contenha valores negativos, visto que se considerou esta regra como sendo a mais fundamental. Quando falhar uma ou mais regras que não esta, irá inserir à variável **Estatuto** a palavra “Crítico”.

Tabela 7- Primeiras linhas do excel criado com auxílio do python

Invoice ID	Branch	City	Customer type	Gender	Product line	Unit price	Quantity	Tax 5%	Total	Date	Time	Payment	cogs	gross margin percentage	gross income	Rating	Falhas	Regras	Estatuto
750-67-8428	A	Yangon	Member	Female	Health and beauty	74.69	7	26.1415	548.9715	1/5/2019	13:08:00	Ewallet	522.83	4.761905	26.1415	9.1	0		Nao Critico
226-31-3081	C	Naypyitaw	Normal	Female	Electronic accessories	15.28	5	3.82	80.22	3/8/2019	10:29:00	Cash	76.4	4.761905	3.82	9.6	0		Nao Critico
631-41-3108	A	Yangon	Normal	Male	Home and lifestyle	46.33	7	16.2155	340.5255	3/3/2019	13:23:00	Credit card	324.31	4.761905	16.2155	7.4	1	ID da Regra:25053208, Regra:Quantity value ranges between 0&20	Critico
123-19-1176	A	Yangon	Member	Male	Health and beauty	58.22	8	23.288	489.048	1/27/2019	20:33:00	Ewallet	465.76	4.761905	23.288	8.4	0		Nao Critico
373-73-7910	A	Yangon	Normal	Male	Sports and travel	86.31	7	30.2085	634.3785	2/8/2019	10:37:00	Ewallet	604.17	4.761905	30.2085	5.3	0		Nao Critico
699-14-3026	C	Naypyitaw	Normal	Male	Electronic accessories	85.39	7	29.8865	627.6165	3/25/2019	18:30:00	Ewallet	597.73	4.761905	29.8865	4.1	1	ID da Regra:25053209, Regra:Branch is equal to A, B or C	Critico

5.3 Criação de um *dashboard*

Nesta última etapa, o objetivo era a criação de um *dashboard*, onde se apresentassem indicadores de negócio. Além disso, decidiu-se acrescentar ao *dashboard* duas páginas. Uma delas, a primeira página, é constituída pelo esquema presente na figura 16, com o intuito de se fazer um resumo do projeto final. A outra página adicional (segunda página) fará uma análise exploratória dos dados.

5.3.1 Primeira Página

A primeira página deste *dashboard* mostra resumidamente as três etapas que constituem o projeto final, referindo também os temas para as duas seguintes páginas deste *dashboard*.

Esta página está retratada na figura 16 mencionada anteriormente.

5.3.2 Segunda Página

Como já referido anteriormente, esta página corresponde a uma análise exploratória de dados com o intuito de compreender melhor os dados. A segunda página do *dashboard* é apresentada na figura 23.

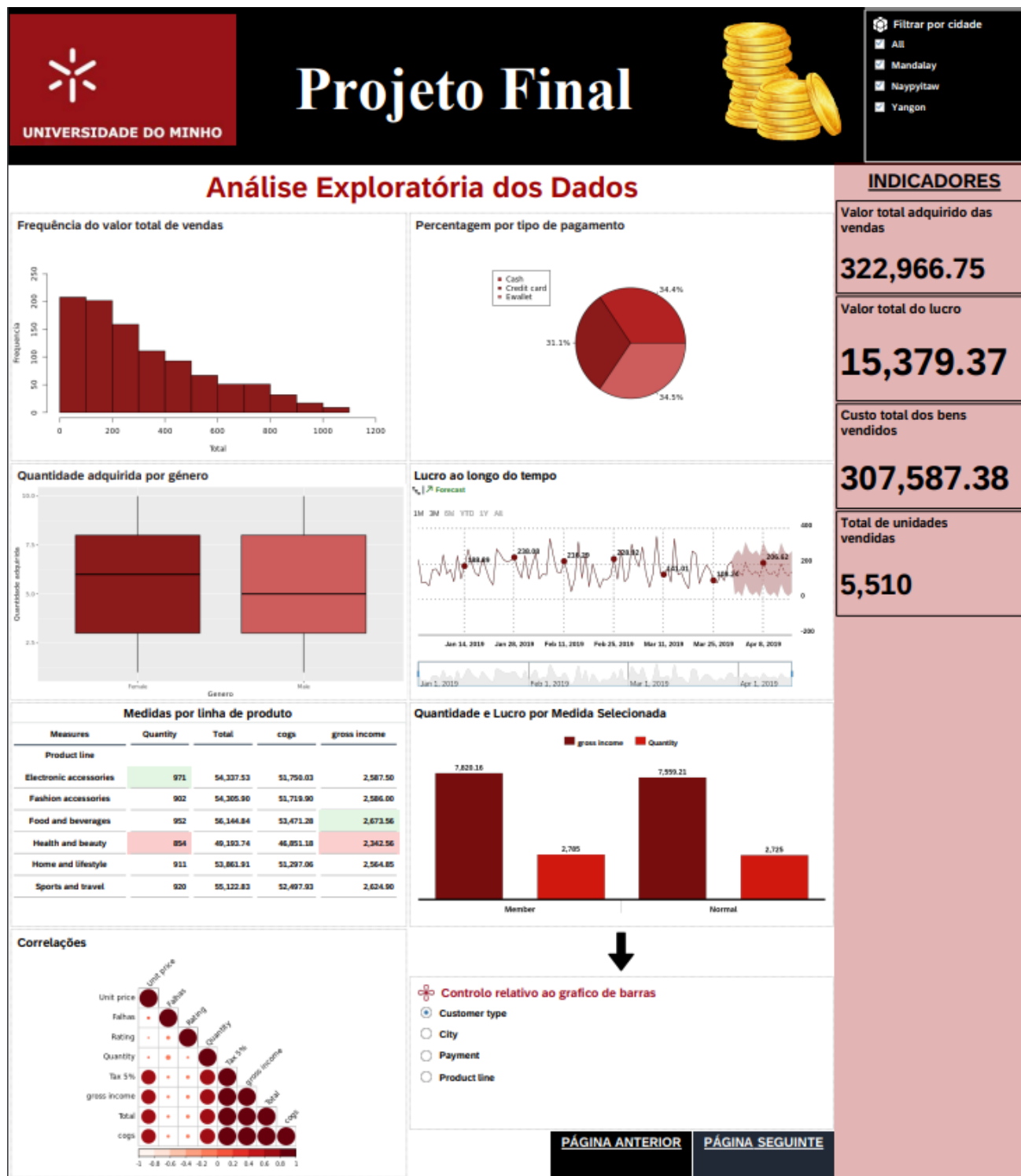


Figura 23- Segunda página do dashboard

Através da figura 23 obtém-se uma imagem geral da análise exploratória realizada. De seguida, irá se interpretar detalhadamente.



Figura 24- Filtro por cidade

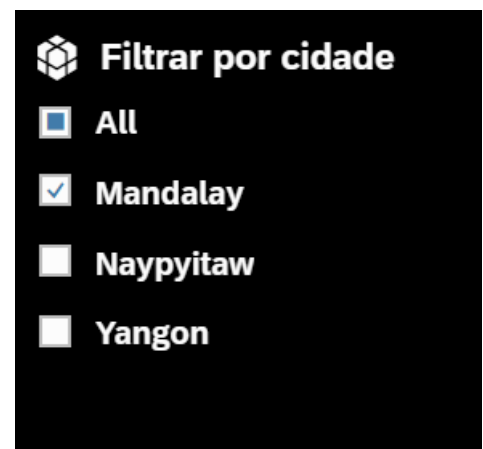


Figura 25- Filtro por cidade

Nas figuras 24 e 25 observa-se um controlo que permite filtrar toda a página por cidade(s). Significa que apenas são manipulados na página os dados correspondentes à cidade ou cidades seleccionadas no controlo. Por exemplo, na figura 24 estão seleccionadas todas as cidades, consequentemente serão utilizados todos os dados. Do mesmo modo, na figura 25 está definida apenas Mandalay, por conseguinte serão usadas as observações pertencentes a esta cidade.

Durante as próximas interpretações será aplicado o filtro com todas as cidades escolhidas a fim de empregar todas as observações.

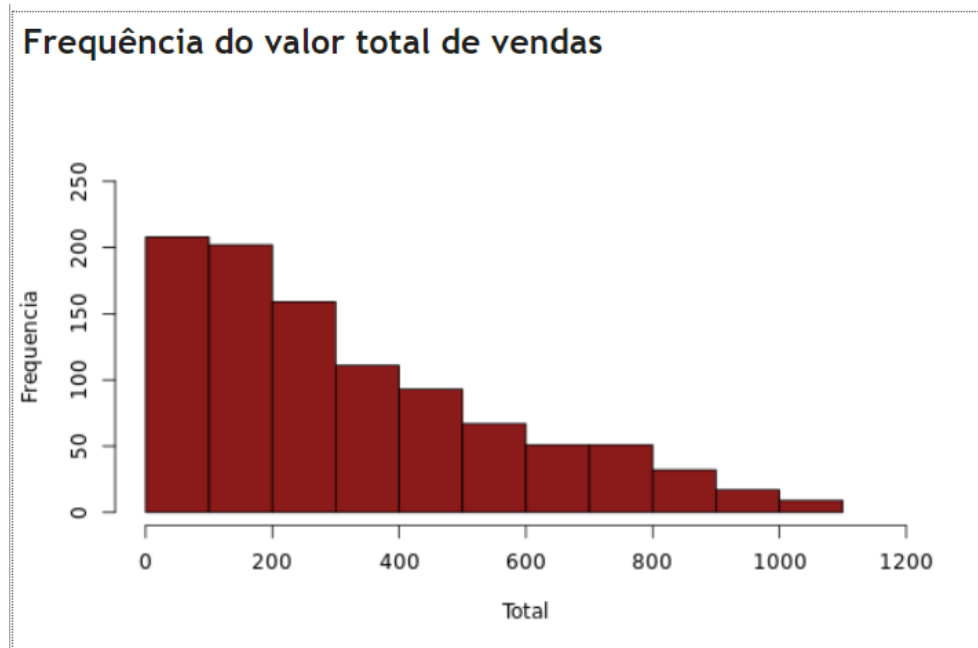


Figura 26- Gráfico de barras para a frequência do valor total de vendas

```
Editor*  
1 dados<-dados_SAC_FINAL  
2 attach(dados)  
3 dens = density(dados$'Total')  
4 hist( dados$'Total',col="firebrick4",main= " ",xlab="Total",ylab="Frequencia",xlim=c(0,1200),ylim=c(0,250))
```

Figura 27- Código utilizado para a criação do gráfico da figura 26

O gráfico de barras exposto na figura 26 criou-se aplicando o R nativo do SAC. É possível ver o código para a elaboração deste na figura 27. A partir desta representação gráfica constata-se que é mais frequente compras em valor reduzido, posto que a frequência vai decaindo à medida que o valor total aumenta.

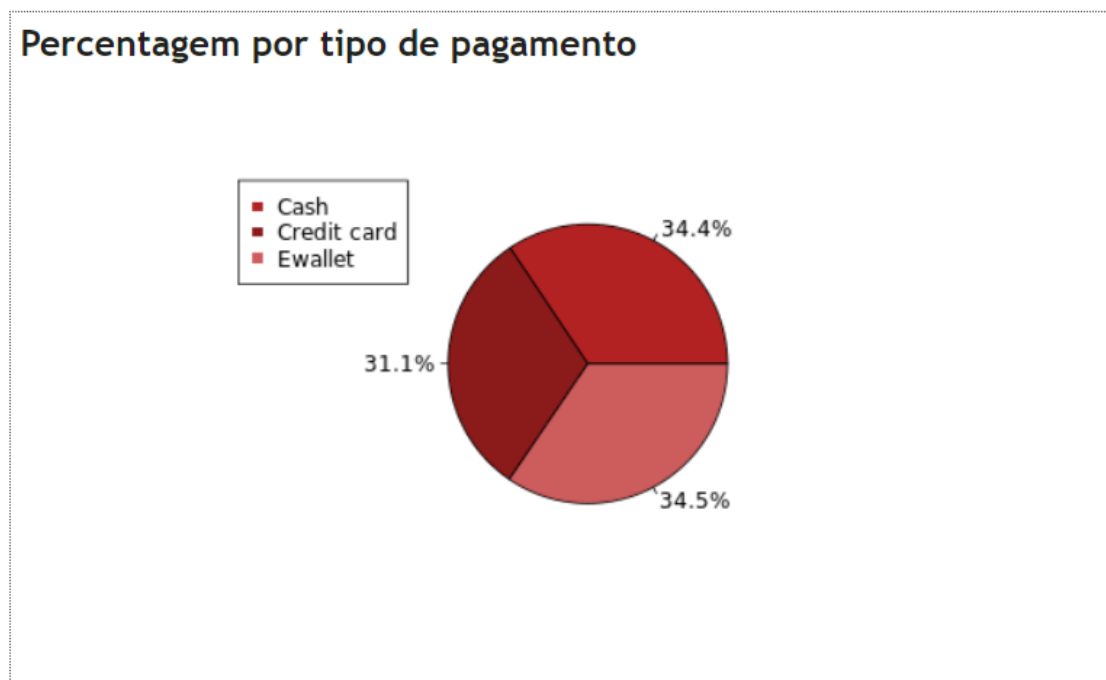


Figura 28- Gráfico circular para a percentagem por tipo de pagamento

```

Editor
1 dados <- dados_SAC_FINAL
2 perc<-round(table(dados$"Payment")/sum(table(dados$"Payment"))*100,2)
3 rotulos<-paste("",perc,"%",sep="")
4 pie(table(dados$"Payment"),labels=rotulos,col=c('firebrick', 'firebrick4', 'indianred') )
5 legend(-2,1.05,levels(dados$"Payment"),pch=rep(15,5),col=c('firebrick','firebrick4', 'indianred'))
  
```

Figura 29- Código para a criação do gráfico da figura 28

O gráfico circular apresentado na figura 28 elaborou-se aplicando o código em linguagem R exibido na figura 29. Através deste, percebe-se que não difere muito a frequência da utilização dos três tipos de pagamento. Dos quais o pagamento por *ewallet* é ligeiramente superior aos restantes, ocorrendo em 34.5% das compras, e o pagamento por cartão de crédito é um pouco inferior aos outros tipos de pagamento, contendo 31.1% das compras.

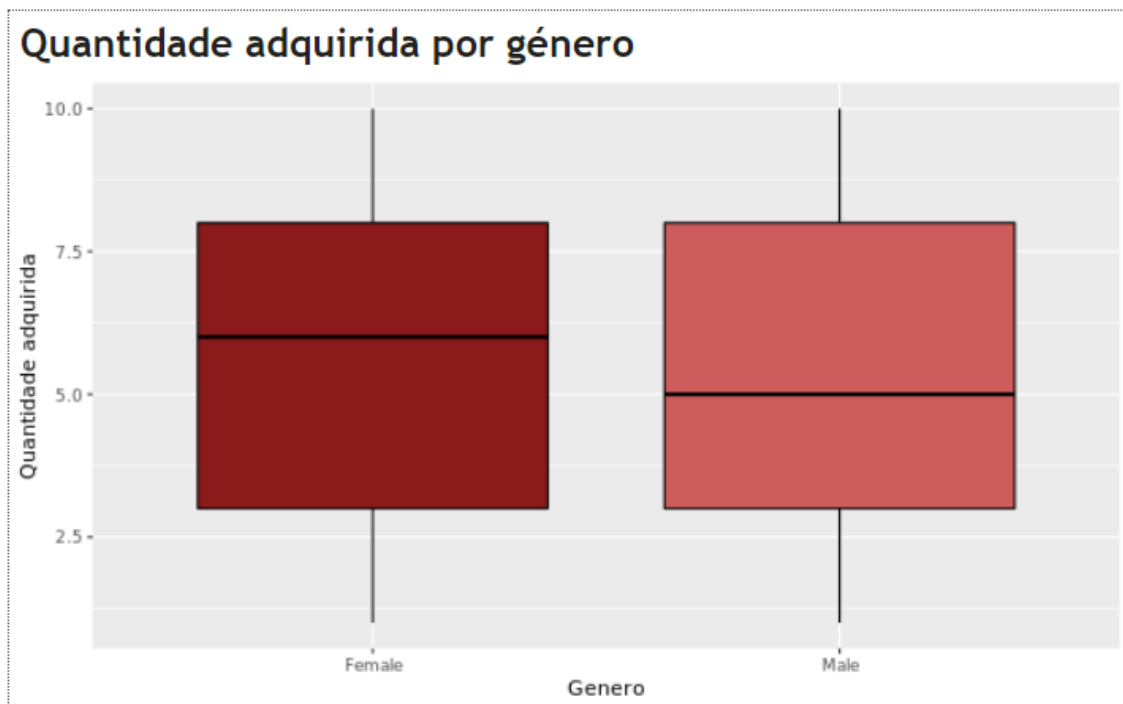


Figura 30- Caixa de bigodes para a quantidade adquirida por género

Editor

```

1 library(ggplot2)
2 dados <- dados_SAC_FINAL
3 ggplot(dados,aes(y=dados$'Quantity',x=dados$'Gender',color=dados$'Gender'))+
4 geom_boxplot(fill=c('firebrick4', 'indianred'),col='black')+
5 labs(x='Genero',y='Quantidade adquirida')

```

Execute

Figura 31- Código para a criação da representação gráfica da figura 30

Na figura 30, observa-se caixas de bigodes nas quais se deduz que em média o género feminino realiza compras para um valor mais elevado de quantidade em relação ao género masculino. Esta representação gráfica foi construída com o auxílio da biblioteca *ggplot2* existente na linguagem R, tal como se pode visualizar na figura 31.

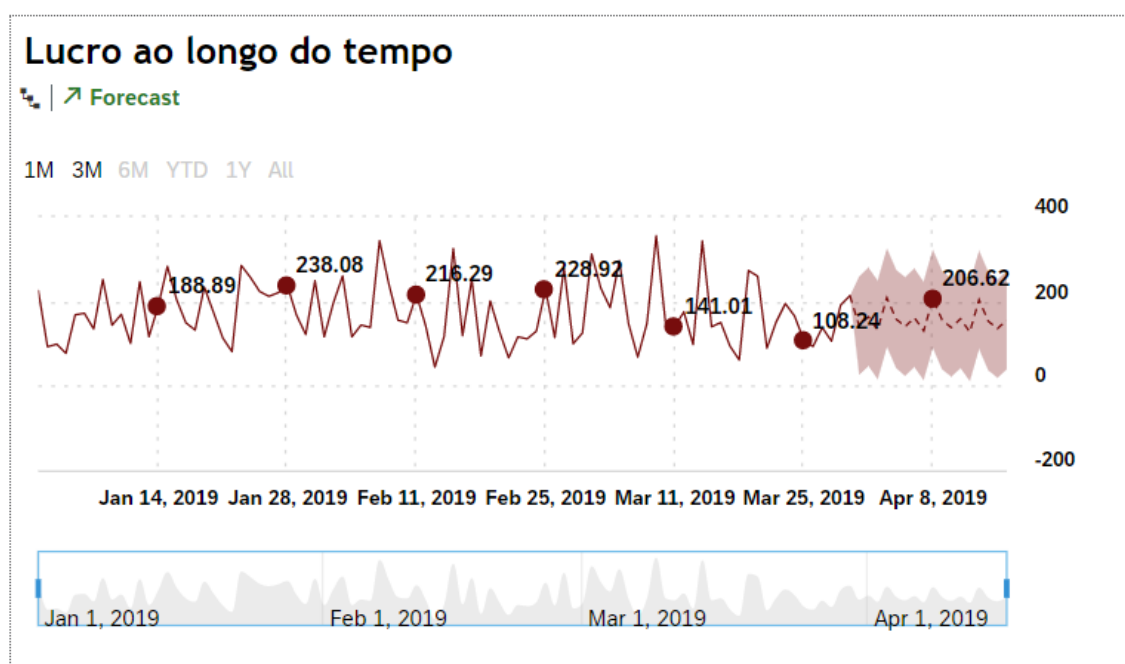


Figura 32- Série Temporal para o lucro

Na figura 32, visualiza-se uma série temporal para o lucro na qual está presente uma previsão para os 16 dias seguintes (até 16 de abril). Este gráfico elaborou-se aplicando as funcionalidades do SAC. Tendo em conta que os dados são apenas referentes a três meses pode-se supor que a previsão não será a ideal pois teve como base um histórico de dados bastante reduzido.

Medidas por linha de produto				
Measures	Quantity	Total	cogs	gross income
Product line				
Electronic accessories	971	54,337.53	51,750.03	2,587.50
Fashion accessories	902	54,305.90	51,719.90	2,586.00
Food and beverages	952	56,144.84	53,471.28	2,673.56
Health and beauty	854	49,193.74	46,851.18	2,342.56
Home and lifestyle	911	53,861.91	51,297.06	2,564.85
Sports and travel	920	55,122.83	52,497.93	2,624.90

Figura 33- Tabela para as medidas por linha de produto

Na tabela ilustrada na figura 33 observa-se os valores para a quantidade, total, custo e lucro para cada linha/categoria de produto. Além disso, está assinalado a vermelho o valor mínimo e a verde o valor máximo para as variáveis relativas à quantidade e ao lucro. Conclui-se que para compras em quantidades mais elevadas não implica maior lucro. Esta tabela construiu-se aplicando as ferramentas do SAC.

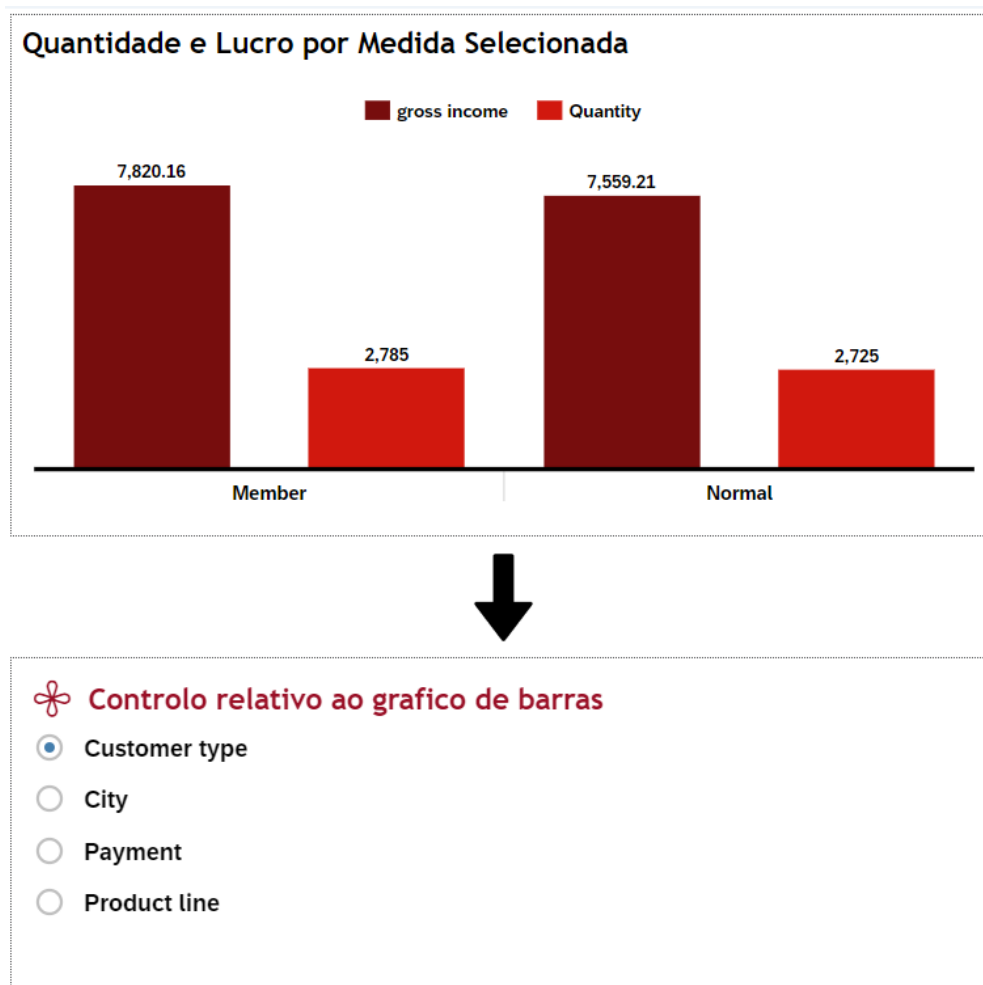


Figura 34- Gráfico de barras para a quantidade e lucro por medida seleccionada

Na figura 34, visualiza-se um gráfico de barras e um controlo para este. O controlo permite filtrar a representação gráfica para tipo de cliente, cidade, tipo de pagamento ou linha de produto por quantidade e lucro. Neste caso em particular, constata-se que um cliente que seja membro dá um lucro 3.34% maior que um cliente normal. Além disso, compra quantidades 2.15% maiores.

Correlações



Figura 35-Representação gráfica das correlações

```

Editor
1 dados<- dados_SAC_FINAL
2 novos_dados <- dados[,c(13,14,15,16,17,19,20,21)]
3 library(corrplot)
4 library(RColorBrewer)
5 corrplot(cor(novos_dados), type = 'lower', order = 'hclust', tl.col = 'black',
6 cl.ratio = 0.2, tl.srt = 45, col = COL1("Reds", 10))
Execute

```

Figura 36- Código para criação da representação gráfica da figura 35

Na figura 35, observa-se uma representação gráfica das correlações na qual quanto maior e mais escura for a bola maior é o valor da correlação. Desta forma, pode-se ver, por exemplo, que para a variável total existe correlação elevada quando representada para a variável lucro ou para a taxa de imposto de 5%. Na figura 36, está presente o código utilizado para a realização desta representação gráfica.

Valor total adquirido das vendas
322,966.75
Valor total do lucro
15,379.37
Custo total dos bens vendidos
307,587.38
Total de unidades vendidas
5,510

Figura 37- Indicadores

Na figura 37 estão expostos os indicadores construídos com a aplicação das funcionalidades do SAC. Aqui, pode ver-se o valor total adquirido das vendas, o valor total do lucro, o custo total dos bens vendidos e por fim o total de unidades vendidas.

5.3.3 Terceira página

Esta última página centra-se na criação e análise de indicadores de negócio, isto é, na análise dos KPI's criados anteriormente com o auxílio da linguagem *python*. Na figura 38, visualiza-se a terceira página do *dashboard*. De modo análogo ao capítulo anterior irá realizar-se uma análise mais detalhada desta página. No decorrer desta análise irá-se interpretar principalmente as variáveis criadas com o auxílio do *python*. Relembrando que a variável **Falhas** representa o número de regras que não foram aprovadas por cada observação.

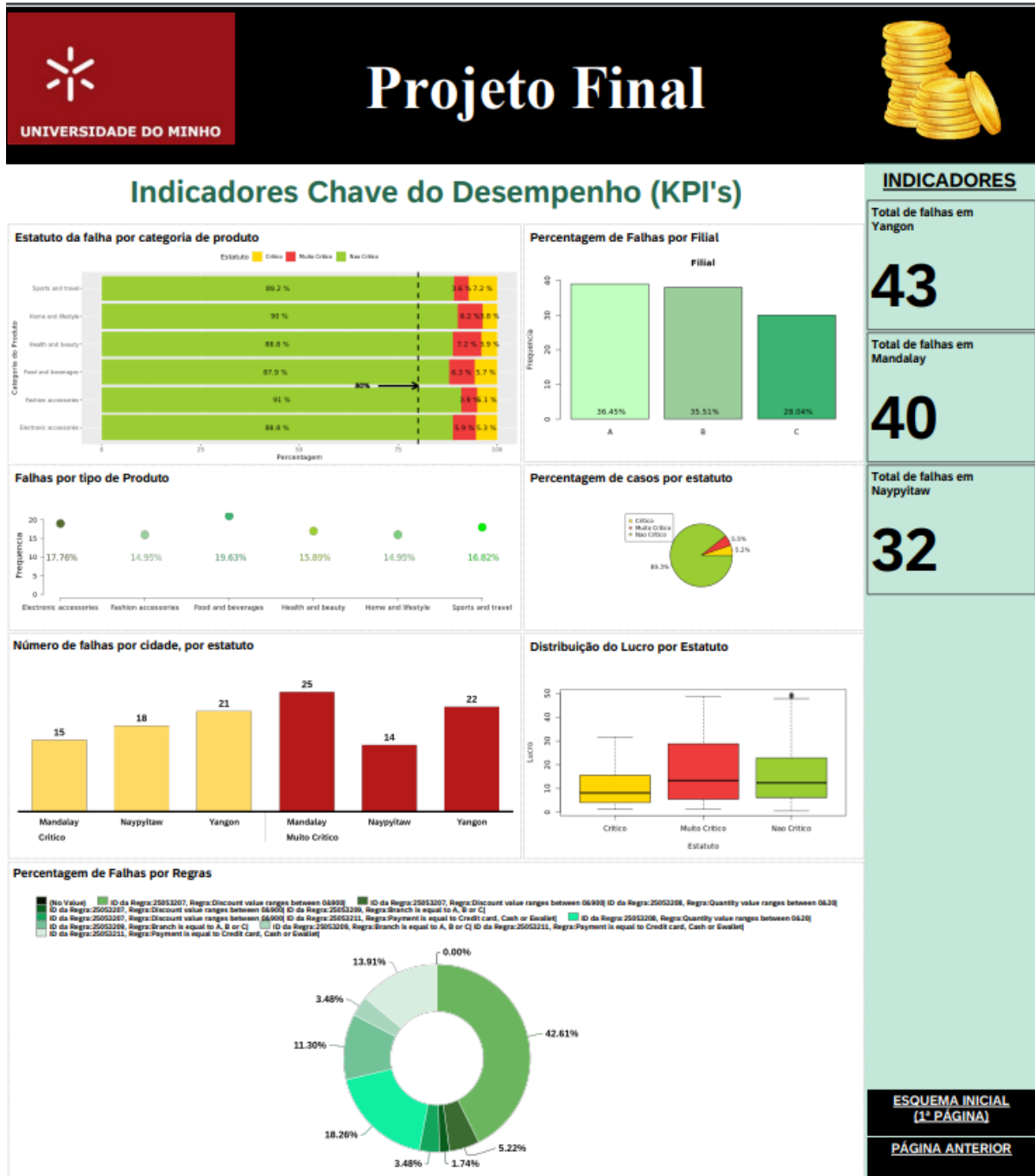


Figura 38- Terceira página do dashboard

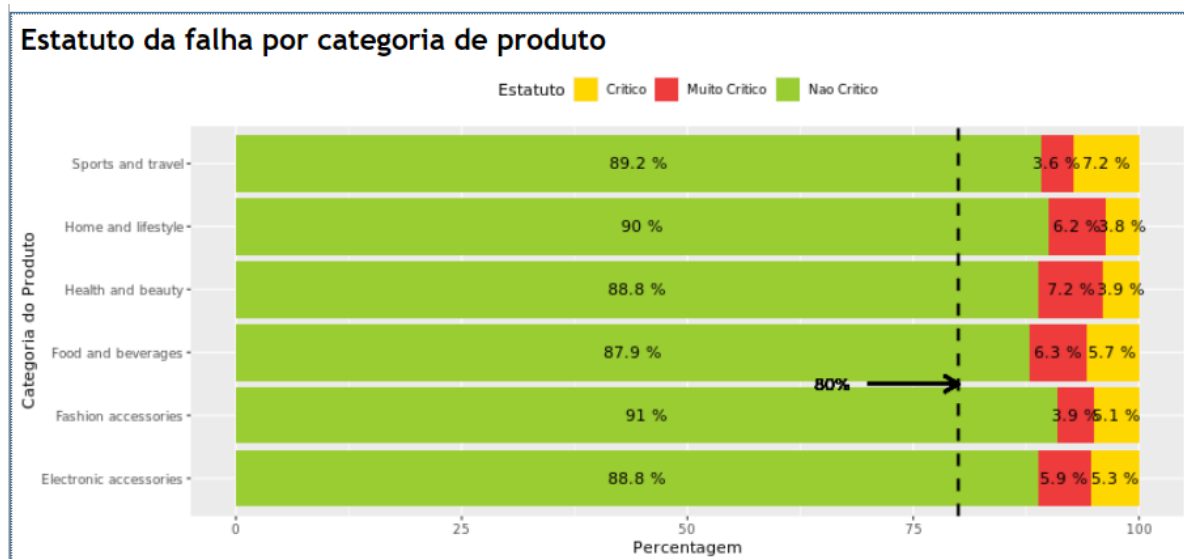


Figura 39- Gráfico de barras para o estatuto da falha por categoria de produto

Editor

```

1 attach(dados_SAC_FINAL)
2 library(ggplot2)
3 library(dplyr)
4 freq<- dados_SAC_FINAL%>%
5 group_by(dados_SAC_FINAL$'Product line',Estatuto)%>%
6 summarise(n=n())%>%
7 mutate(fre=n/sum(n)*100)%>%
8 ungroup()
9 tipo <- freq$'dados_SAC_FINAL$"Product line"'
10 library(scales)
11 ggplot(freq,aes(x=tipo,y=fre,fill=Estatuto,label=paste(fre,"%")))+
12 geom_col()+
13 geom_text(aes(label=paste(round(fre,1),"%")),position=position_stack(vjust=0.5))+
14 labs(y="Percentagem",x="Categoria do Produto")+
15 scale_fill_manual(values=c("gold","brown2","yellowgreen"))+
16 coord_flip()+
17 theme(legend.position="top")+
18 geom_hline(yintercept=80,linetype="dashed",size=1)+
19 geom_segment(aes(x=2.5,y=70,xend=2.5,yend=80),size=1,arrow=arrow(length=unit(4,"mm")))+
20 geom_text(aes(x=2.5,y=66),label="80%")

```

Figura 40- Código para a criação da representação gráfica da figura 39

Relembra-se que, para cada observação, **Falhas** é a variável que corresponde ao número de regras não verificadas pela observação e que quando o seu **Estatuto** é “Critico” significa que falhou pelo menos uma regra. Assim, na figura 39 encontra-se um gráfico de barras do **Estatuto** por categoria de produto no qual se verifica que na maioria das observações as regras foram validadas, uma vez que a percentagem de observações na qual o **Estatuto** tem valor igual a “Nao Critico” é elevado. Além disso, constata-se que os acessórios de moda foram os produtos nos quais ocorreu menos falhas, pois em 91% destes o **Estatuto** é “Nao Critico” (todas as regras validadas).

Na figura 40 vê-se o código em linguagem R aplicado para a criação do gráfico da figura 39. Observa-se que se recorreu à biblioteca *dplyr* para gerar as frequências do estatuto de falha por categoria de produto. Assim como se recorreu à biblioteca *ggplot2* para reproduzir o gráfico. Decidiu-se adicionar uma referência da localização da percentagem igual a 80% (representado a tracejado no gráfico), visto que se interpreta melhor visualmente o facto da maioria das observações serem validadas pelas regras aplicadas no IDQ. Dado que para todas as categorias de produto se verifica que o **Estatuto** “Nao Critico” se encontra para percentagens maiores do que 80%.

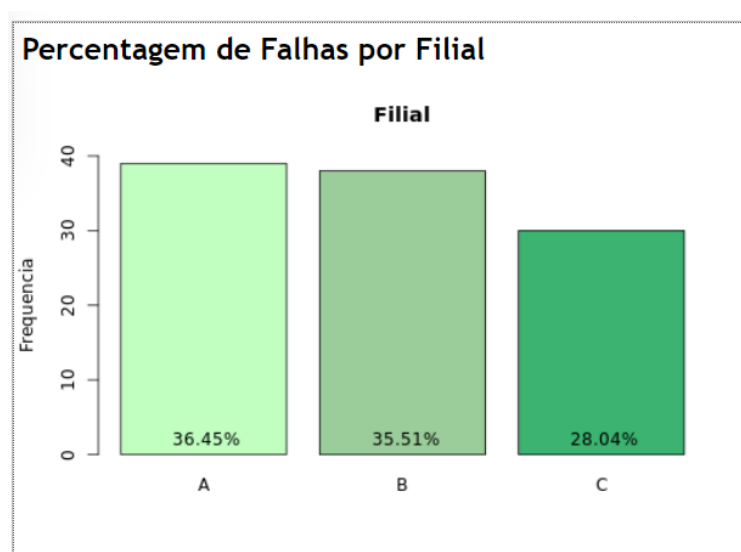


Figura 41- Gráfico de barras da percentagem de falhas por filial

```

Editor*
1 dados <- dados_SAC_FINAL
2 dados_falhas <- dados[dados$'Falhas'!=0,]
3 contagem <- table(dados_falhas$'Branch')
4 percentagem <- round(prop.table(contagem)*100,2)
5 rotulo <- paste(' ',percentagem,'% ',sep='')
6 barras <- barplot(contagem,ylim=c(0,40),ylab='Frequencia',main='Filial',
7                   col=c('darkseagreen1 ','darkseagreen3','mediumseagreen'))
8 text(barras,2.2,rotulo,cex=1)

```

Figura 42- Código utilizado para gerar o gráfico da figura 41

Através da figura 41 pode-se deduzir que a filial A é a filial com maior percentagem de falhas (36.45% das falhas) e portanto, a filial com mais regras que falham nas observações dadas. Pelo contrário, a filial C é onde se encontra o maior número de regras validadas no conjunto de observações (28.04% das falhas).

A figura 42 expõe o código produzido no R nativo do SAC para a elaboração do gráfico referido anteriormente. Para isto utilizou-se a função “barplot()”.

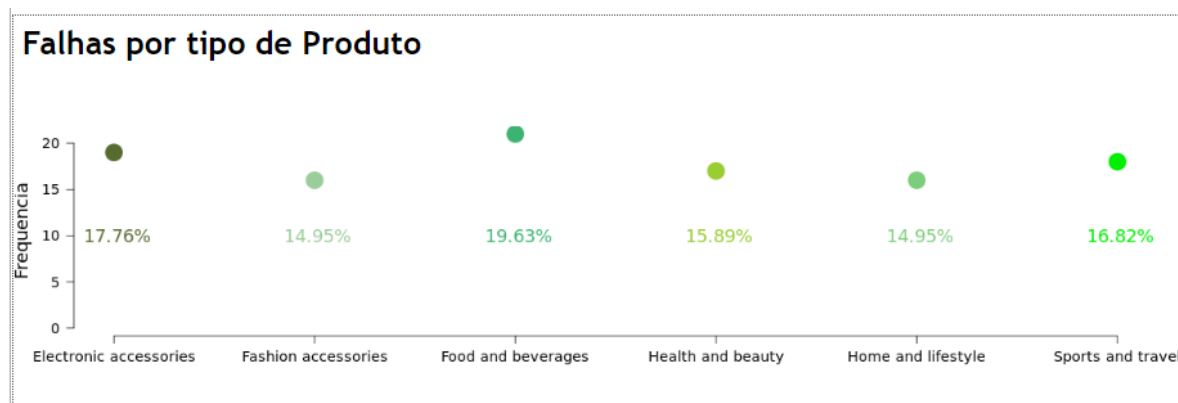


Figura 44- Gráfico relativo à percentagem de falhas por tipo de produto

```

Editor*
1 dados <- dados_SAC_FINAL
2 library(ggplot2)
3 produto <- table(dados_SAC_FINAL$'Product line')
4 dados_falhas <- dados[dados$'Falhas'!=0,]
5 contagem <- table(dados_falhas$'Product line')
6 nomes <- levels(dados_falhas$'Product line')
7 nomes
8 percentagem <- round(prop.table(contagem)*100,2)
9 rotulo <- paste(' ',percentagem,'% ',sep=" ")
10 barras <- plot(contagem,ylab='Frequencia',col=c('darkolivegreen',
11                                             'darkseagreen3','mediumseagreen', 'yellowgreen','palegreen3', 'green2'),
12 type='p',pch=19,xlim=c(1,6),las=1,cex=3,cex.axis=1.2,cex.lab=1.5)
13 text(barras,10,rotulo,cex=1.5,col=c('darkolivegreen ', 'darkseagreen3', 'mediumseagreen', 'yellowgreen', 'palegreen3', 'green2'))

```

Figura 43- Código em R para gerar o gráfico da figura 43

Atendendo à figura 43 pode-se ver que para o conjunto de observações no qual houve regra(s) reprovada(s), nos produtos do tipo acessórios de moda e do tipo casa e estilo de vida verificou-se menor percentagem das falhas (14.95% das falhas), ou seja, foi para estas categorias que as observações menos dispuseram de falhas. Contrariamente à categoria de comida e bebida onde se situa a maior percentagem das falhas.

Na figura 44 estão presentes as linhas de código utilizadas para a construção da representação gráfica da figura 43.



Figura 45- Gráfico circular para a percentagem de casos por estatuto

Editor

```

1 dados<-dados_SAC_FINAL
2 dados
3 attach(dados)
4 perc<-table(dados$"Estatuto")/sum(table(dados$"Estatuto"))*100
5 table(dados$"Estatuto")
6 perc
7 rotulos<-paste("",perc,"%",sep="")
8 pie(table(dados$"Estatuto"),labels=rotulos,col=c("gold","brown2","yellowgreen"))
9
10 legend(-1.95,1.05,levels(dados$"Estatuto"),pch=rep(15,5),col=c("gold","brown2","yellowgreen"))

```

Figura 46- Código gerador do gráfico circular da figura 45

Mediante a figura 45 sabe-se que 89.3% dos dados se classifica com **Estatuto** “Nao Critico”, significando que 89.3% dos dados foram validados por todas as regras. Tal como, 5.2% das observações se consideram “Critico”, isto é, houve regra(s) reprovada(s) apesar de não falhar a regra “Discount value ranges between 0&900”. Por fim, 5.5% dos dados são identificados como “Muito Critico”, ou seja, dados em que a regra “Discount value ranges between 0&900” falhou.

Conclui-se que a maioria das observações são validadas pelas regras e que a regra “Discount value ranges between 0&900” é a regra que mais falha, uma vez

que a percentagem de valores para o **Estatuto** “Muito Critico” é superior à percentagem de valores para “Critico”.

Mais uma vez, recorre-se ao R nativo do SAC para gerar a representação gráfica. Na figura 46 está exposto o código gerador do gráfico circular.

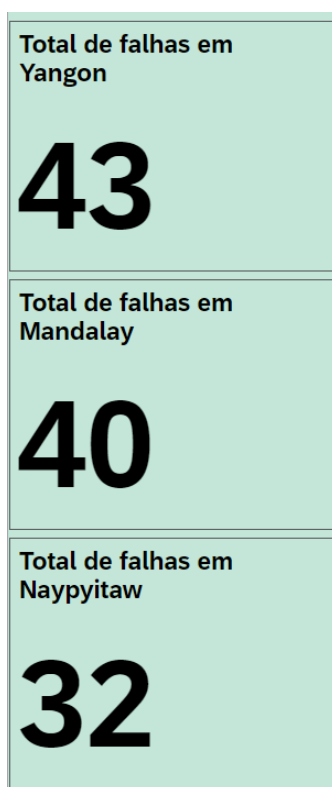


Figura 47- Indicadores de falhas por cidade

Na figura 47 observa-se os indicadores para as falhas em cada cidade. Vê-se que para a cidade Yangon ocorreram 43 falhas, para Mandalay verificou-se 40 falhas e para Naypyitaw deu-se 32 falhas.

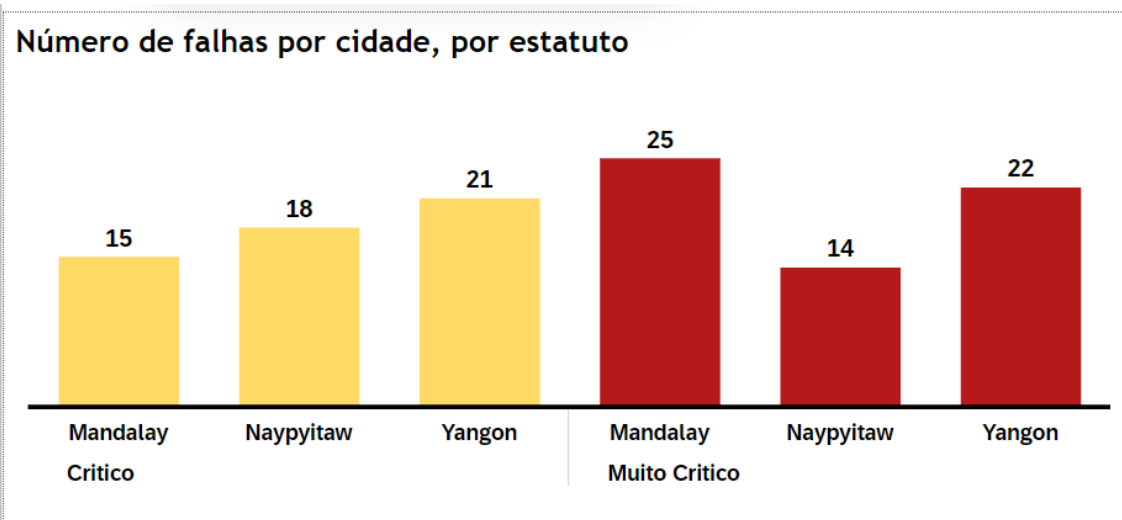


Figura 48- Gráfico de barras para o número de falhas por Cidade para cada Estatuto

Recorreu-se às funcionalidades do SAC para a criação do gráfico da figura 48. Neste gráfico excluiu-se o **Estatuto** “Nao Critico” dado que o número de falhas para esta categoria é sempre igual a zero. Pela representação gráfica vê-se que para as cidades Mandalay e Yangon há maior número de falhas quando o **Estatuto** é “Muito Critico” do que quando é “Critico”. Apenas se verificando o contrário para a cidade Yangon. Isto reforça o facto de a percentagem de observações para o **Estatuto** “Muito Critico” ser ligeiramente superior à percentagem para a categoria “Critico”.

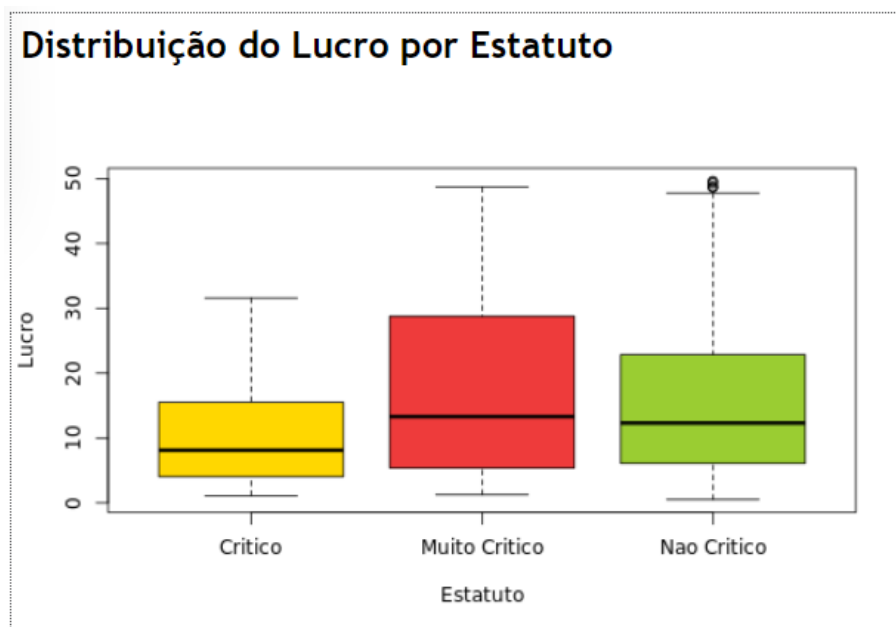


Figura 49- Caixas de bigode com a distribuição do lucro por Estatuto

```

Editor*
1 dados<-dados_SAC_FINAL
2 boxplot(dados_SAC_FINAL$'gross income'~dados_SAC_FINAL$'Estatuto',
3         col=c("gold","brown2","yellowgreen"),xlab='Estatuto',ylab="Lucro")

```

Figura 50- Código gerador das caixas de bigodes da figura 49

Observando a figura 49 deduz-se que em média os valores do lucro onde o **Estatuto** se classifica como “Muito Critico” é superior. No entanto, também se pode observar *outliers* para a categoria “Nao Critico”, isto deve-se ao facto de os valores mais altos de lucro (sendo o máximo igual a 49.65) se localizarem para esta categoria.

Na figura 50 encontra-se o código aplicado para a elaboração das caixas de bigodes, mais uma vez utilizando-se a linguagem R.

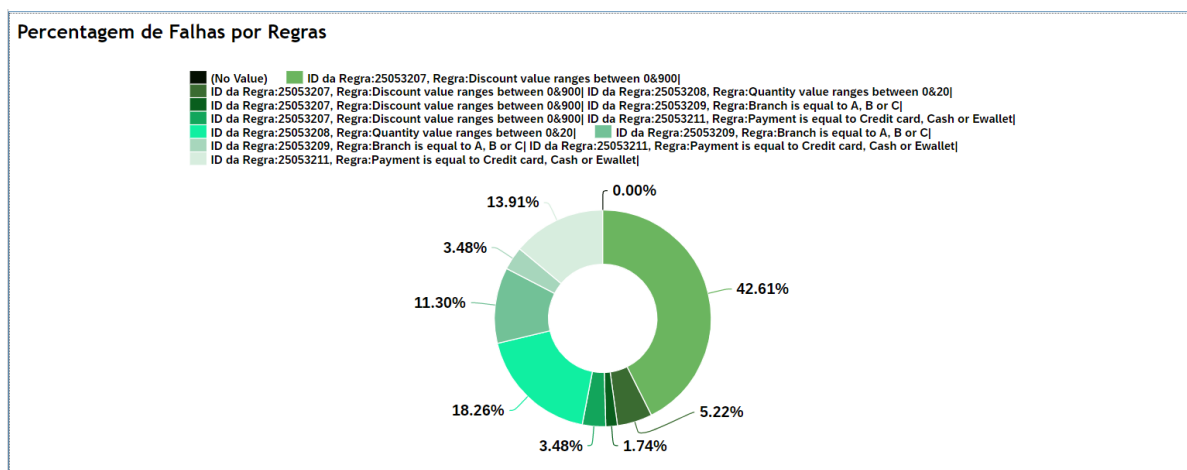


Figura 51- Gráfico circular para a percentagem de falhas por regras

O gráfico circular presente na figura 51 mostra as percentagens de falhas por regra. É de notar que quando duas regras falham em simultâneo considera-se uma nova categoria. Visualiza-se que a “fatia” maior, com 42.61% das falhas, ocorre para a regra “Discount value ranges between 0&900”, sendo que houve falhas em 49 observações. A segunda maior “fatia”, com 18.26% das falhas, referente à regra “Quantity value ranges between 0&20”, sendo que falhou em 21 observações.

Esta representação gráfica foi reproduzida com o auxílio das ferramentas do SAC.

CONCLUSÕES

O principal objetivo deste projeto consistia na criação de um *dashboard* na qual estivessem expostos indicadores de negócio. No entanto, também se acrescentou ao *dashboard* análise da base de dados. Além disso, aplicou-se as funcionalidades das ferramentas *SAP analytics cloud* e *Intelligent Data Quality*.

Primeiramente, conclui-se que os dados tinham algumas falhas para as regras aplicadas no *Intelligent Data Quality*. Deste modo, recorreu-se à linguagem *python* para compilar a informação retirada do IDQ e adicionar variáveis que permitissem criar KPI's (*Key Performance Indicator*).

De seguida, importou-se o modelo de dados criados no *Visual Studio Code*, com a aplicação da linguagem *python*, para a aplicação *SAP analytics cloud*. Nesta ferramenta elaborou-se um *dashboard* onde se podem tirar conclusões sobre os dados, isto é, fazer uma análise exploratória dos dados e também verificar que os KPI's permitiram deduzir que na maioria dos dados a informação estava correta, uma vez que na generalidade das observações não ocorreu falhas.

Durante a aplicação das funcionalidades da ferramenta *SAP analytics cloud* focou-se mais na utilização do R nativo dado que é uma mais valia para a Accenture. Esta ferramenta é usualmente utilizada sem recorrer ao R. Assim sendo, considero que o meu contributo nesta medida permitirá futuras análises interessantes com o auxílio do R.

Este estágio também permitiu utilizar a informação recolhida na aplicação *Intelligent Data Quality* para posteriormente aplicá-la na ferramenta *SAP analytics cloud*. Uma vez que estas ferramentas são frequentemente utilizadas pela empresa considero que a minha contribuição nesta etapa foi igualmente uma mais valia.

Durante este projeto apliquei competências adquiridas ao longo dos três anos de licenciatura. Além disso, desenvolvi novas competências como é o caso do conhecimento e uso das aplicações *SAP analytics cloud* e *Intelligent Data Quality* assim como a elaboração de um *dashboard*.

REFERÊNCIAS

- [1] Accenture (2022). *Sobre a Accenture*. Disponível em <https://www.accenture.com/pt-pt/about/company/portugal>
- [2] Rstudio (2022). *About Rstudio*. Disponível em <https://www.rstudio.com/about/>
- [3] SAP (2021). *SAP Analytics Cloud*. Disponível em <https://www.sap.com/products/cloud-analytics.html>
- [4] Microsoft (2022). *Utilizar a extensão do Visual Studio Code*. Disponível em <https://docs.microsoft.com/pt-pt/power-apps/maker/portals/vs-code-extension>
- [5] Accenture (2020). *SAP Data Migration for S4HANA using myConcerto_FY20_to share* [pdf]. Braga: Accenture.
- [6] Gabriel, Lucas (2018, setembro 18). *Understand what KPI is and find out how it can help you measure your marketing results*. Disponível em <https://rockcontent.com/br/blog/kpi/>
- [7] Tyagi, Chaitanya (2022). *Adding new column to existing dataframe in pandas*. Consultado em 18 de abril. 2022. Disponível em <https://www.acervolima.com.br/2020/12/adicionando-nova-coluna-ao-dataframe.html>
- [8] Vieira, Alex (2017, janeiro 23). *Python: Append ou Extend? Adicionando elementos na lista*. Consultado em 20 de abril. 2022. Disponível em <https://www.alura.com.br/artigos/adicionando-elementos-na-lista-do-python-append-ou-extend>
- [9] Python Software Foundation (2022, abril 25). *CSV- Leitura e escrita de arquivos CSV*. Consultado em 5 de maio de 2022. Disponível em <https://docs.python.org/pt-br/3/library/csv.html>
- [10] Lima, Acervo (2022). *Visualize a array de correlação usando correlograma na programação R*. Consultado a 9 de maio de 2022. Disponível em <https://acervolima.com/visualize-a-array-de-correlacao-usando-correlograma-na-programacao-r/>
- [11] Wei, Taiyun and Simko, Viliam (2021, novembro 18). *An Introduction to corrplot Package*. Consultado a 9 de maio de 2022. Disponível em <https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html>
- [12] Debastiani, Vanderlei (2020, junho 8). *Gráfico com R*. Consultado a 11 de maio de 2022. Disponível em https://vanderleidebastiani.github.io/tutoriais/Graficos_com_R.html
- [13] Wickham, Hadley and Golemund Garrett (2018). *R para Data Science*. Alta Books.