Relatório Trabalho 1

da Disciplina

Integração de Sistemas de Informação

Criação de Um ETL para analise de dados de clientes de um ginásio

Aluno: João Gabriel Gonçalves Peixoto Nº: 23014

Email: a23014@alunos.ipca.pt

Conteúdo

Índice de Imagens

Introdução

Este Trabalho foi elaborado no âmbito da Disciplina de Integração de Sistemas de Informação, tendo como objetivo uma criação de uma ETL.

Eu, neste caso, elaborei um ETL (Extract Transform Load), de uma base de dados obtido na plataforma KAGGLE. Sendo a mesma a seguinte: https://www.kaggle.com/datasets/ka66ledata/gym-membership-dataset

Esta base de dados é composta por clientes de um ginásio e tem dados possíveis para trabalhar, como por exemplo datas, campos nulos. O meu objetivo com este trabalho é experimentar uma nova ferramenta KNIME, uma open source para analise de dados, para assim a criação de esta ETL. Na transformação destes dados o objetivo é que os gerentes do ginásios em conjuntos com outras equipas consigam melhorar não só promoções e criação de funis de marketing, como explorar o bem estar dos seus Clientes.

Figura 1 - Kaggle

Figura 2 - Knime

O Problema

Um ginásio de Braga procurou uma empresa de software para ajudar numa integração e analise dos seus próprios dados, para conhecerem melhor os seus clientes e como podem aumentar os seus clientes e aos mesmo tempo conseguir satisfazer todos. Para isso foram encontrados os seguintes desafios:

Desafios Identificados:

1. **Não conhecem a fundo os seus Clientes:** O ginásio atualmente não possui um sistema unificado para coletar e analisar dados dos membros. As informações estão dispersas entre registos de inscrição, feedback dos clientes e dados de uso das instalações, dificultando a compreensão do perfil dos clientes.
2. **Dificuldade em Identificar Padrões de Comportamento:** Sem a análise de dados adequada, o ginásio não consegue identificar padrões de uso, como horários mais frequentados, preferências por aulas específicas ou serviços adicionais, o que limita a capacidade de personalizar ofertas e melhorar a experiência do cliente.
3. **Estratégias de Retenção de Clientes Ineficazes:** A falta de insights sobre a satisfação do cliente e a taxa de rotatividade leva a uma dificuldade em implementar estratégias eficazes de retenção. O ginásio precisa entender quais fatores influenciam a decisão dos membros de manter ou cancelar suas subscrições.
4. **Ausência de Ações Baseadas em Dados:** As decisões de marketing e promoção são frequentemente baseadas em suposições ou intuições. Isso resulta em campanhas que não são direcionadas ao público certo ou que não refletem as verdadeiras necessidades e desejos dos clientes.
5. **Integração de Dados de Diversas Fontes:** O ginásio utiliza várias plataformas (sistemas de gestão de clientes, plataformas de reservas de aulas, redes sociais) que não estão integradas, dificultando uma visão holística dos dados. A falta de uma solução ETL (Extract, Transform, Load) impede a consolidação e análise dos dados de forma eficiente.

Resolução

Através da ferramenta de analise de dados, Knime, criei assim uma ETL capaz de, organizar e limpar dados, gerar gráficos e report em PDF, e enviar um Email, com um ficheiro .zip.

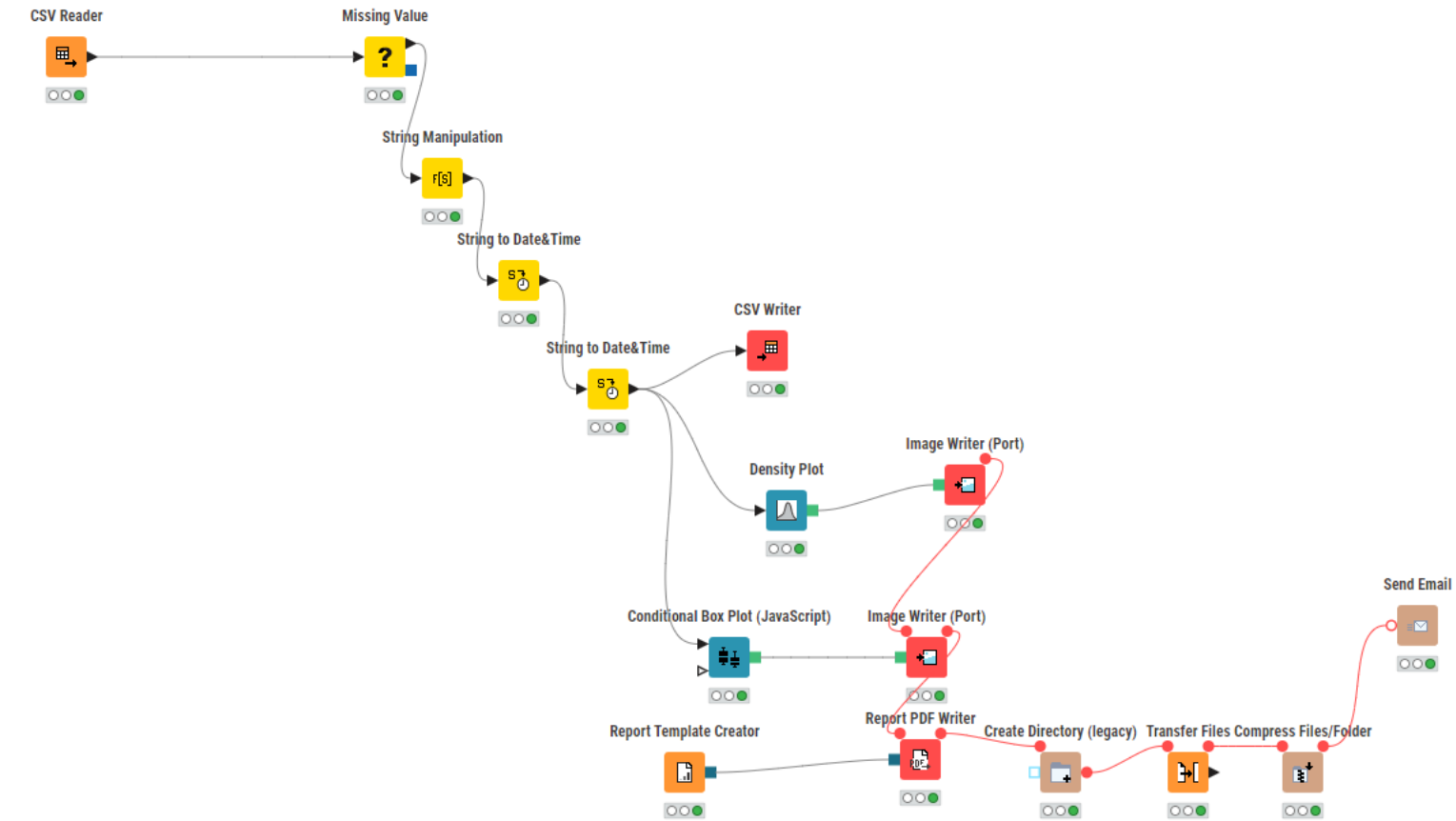


Figura 3 ETL

Sendo neste casso este flow dividido em maioritariamente em 4 partes:

- Ler o ficheiro
- Tratar os dados
- Gerar imagens e reports dos gráficos
- Tratamentos de ficheiro e envios por email

Leitura do Ficheiro

O ficheiro como indicado anteriormente é um ficheiro CSV que contem os dados de clientes de um ginásio, este ficheiro é lido assim para começar o tratamento dos dados.

A nível de configurações usar as default ao ler ficheiros CSV pois o KNIME adapta se facilmente aos dados do CSV. Configurações:

Dialog - 3:1 - CSV Reader

File

Settings | Transformation | Advanced Settings | Limit Rows | Encoding | Flow Variables | Job Manager Selection | Memory Policy

Input location

Read from: Local File System

Mode: File / Files in folder

File: C:\Users\peixo\Downloads\archive\gym_membership.csv — Browse...

Reader options

Format

Autodetect format

Column delimiter: , — Row delimiter: Line break / Custom: \r\n

Quote char: " — Quote escape char: "

Comment char: #

Has column header — Has RowID

Support short data rows — Prepend file index to RowID

Preview

The suggested column types are based on the first 10000 rows only. See 'Advanced Settings' tab.

Row ID	id	gender	birthday	Age	abono...	visit_pe...	days_per_week	attend...	fav_group_lesson	avg_ti...	avg_ti...	avg_ti...	drink_abo	fav_drink	person...	name_...	uses_s...
Row0	1	Female	1997-04-18	27	Premium	4	Mon, Sat, Tue, Wed	True	Kickboxen, BodyPump, Zumba	19:31:00	21:27:00	116	False	?	False	?	True
Row1	2	Female	1977-09-18	47	Standard	3	Mon, Sat, Wed	False	?	19:31:00	20:19:00	48	False	?	True	Chantal	False
Row2	3	Male	1983-03-30	41	Premium	1	Sat	True	XCore	08:29:00	10:32:00	123	True	berry_boost...	True	Mike	False
Row3	4	Male	1980-04-12	44	Premium	3	Sat, Tue, Wed	False	?	09:54:00	11:33:00	99	True	passion_fruit	True	Mike	True
Row4	5	Male	1980-09-10	44	Standard	2	Thu, Wed	True	Running, Yoga, Zumba	08:29:00	09:19:00	50	False	?	True	Mike	False
Row5	6	Female	2009-06-29	15	Standard	1	Mon	False	?	17:19:00	20:19:00	180	False	?	False	?	True
Row6	7	Male	1994-08-07	30	Premium	3	Sat, Thu, Wed	True	LesMiles, BodyPump	19:46:00	20:48:00	62	False	?	False	?	False
Row7	8	Male	2003-11-13	20	Standard	2	Mon, Wed	True	Yoga, XCore	17:45:00	19:20:00	95	True	coconut_pin...	False	?	True
Row8	9	Male	1978-07-28	46	Premium	3	Sat, Sun, Thu	True	BodyPump	09:45:00	11:17:00	92	True	orange, lemon	True	Mike	False
Row9	10	Female	2000-05-06	24	Premium	1	Mon	False	?	13:05:00	15:29:00	144	False	?	True	Jeffrey	True
Row10	11	Male	1983-07-11	41	Standard	1	Fri	False	?	13:50:00	14:26:00	36	False	?	True	Hanna	False
Row11	12	Male	1997-09-19	27	Premium	3	Fri, Thu, Wed	True	Pilates, Zumba, XCore	10:12:00	10:45:00	33	True	passion_fruit	True	Mike	True
Row12	13	Female	1994-04-22	30	Standard	2	Sun, Wed	True	XCore, HIT, Running	17:04:00	19:42:00	158	True	passion_fruit	False	?	True
Row13	14	Female	1989-03-21	35	Premium	3	Mon, Tue, Wed	False	?	19:50:00	20:51:00	61	True	black_currant	False	?	False
Row14	15	Female	1989-10-01	35	Premium	1	Sat	False	?	20:12:00	22:37:00	145	True	orange, blac...	True	Mike	False
Row15	16	Female	1987-12-20	36	Standard	4	Mon, Thu, Tue, Wed	True	Spinning, Zumba, XCore	17:28:00	18:29:00	61	False	?	False	?	False
Row16	17	Male	1986-12-17	37	Premium	3	Fri, Sat, Sun	False	?	09:25:00	12:06:00	161	False	?	True	Jeffrey	True
Row17	18	Female	2011-04-30	13	Standard	3	Sat, Thu, Tue	False	?	16:54:00	18:53:00	119	False	?	True	Mike	True
Row18	19	Female	1986-05-29	38	Standard	1	Tue	True	Pilates	10:01:00	12:48:00	167	True	orange	True	Hanna	True
Row19	20	Male	1991-09-13	33	Premium	2	Sat, Wed	True	BodyPump, Zumba, Pilates	13:55:00	16:24:00	149	False	?	False	?	False
Row20	21	Female	1997-12-31	26	Premium	2	Fri, Mon	False	?	08:19:00	10:45:00	146	False	?	True	Hanna	False
Row21	22	Male	1997-10-29	26	Premium	3	Fri, Mon, Sat	True	Yoga, BodyBalance, LesMiles	15:06:00	17:40:00	154	True	lemon	False	?	True
Row22	23	Female	2010-10-22	13	Standard	2	Mon, Tue	True	HIT	20:56:00	23:36:00	160	True	passion_frui...	True	Mike	False

OK | Apply | Cancel

Figura 4 configs CSV Reader

A nível de tipos de dados neste imagem tempo o header da tabela com o tipo de dados(S – string, I - Int) com o nome da coluna.

Tratamento dos Dados

Para tratar os dados, foi usado assim quatro tipos de tratamentos de dados diferente, sendo desses quatro o primeiro para remover qualquer tipo de valor nulo da tabela, o segundo para remover as abreviações dos dias da semana (Ex: “Mon” para “Monday”), o terceiro e ultimo é para tratamento de colunas de dias e horas.

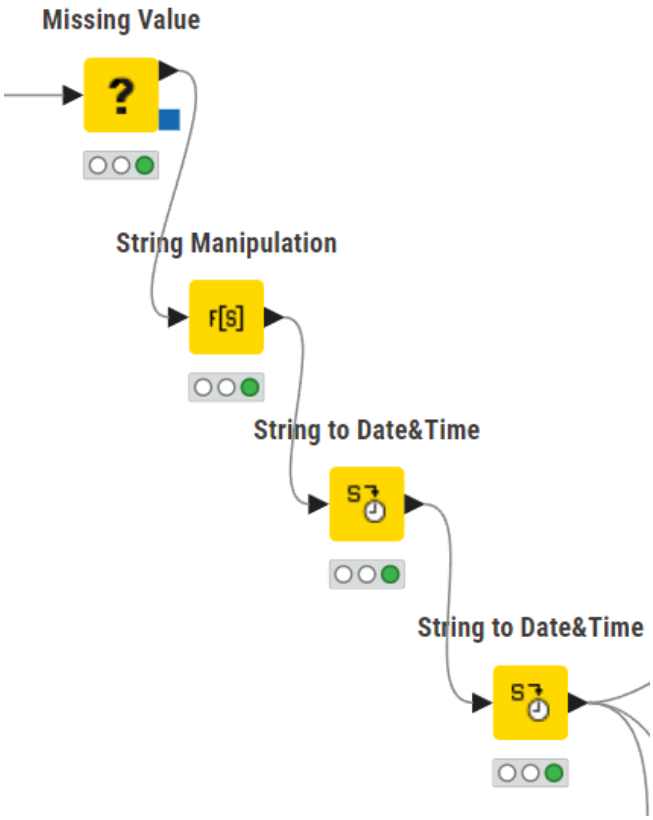


Figura 5 - Fluxo de tratamento de dados

Configurações para *Missing Value*:

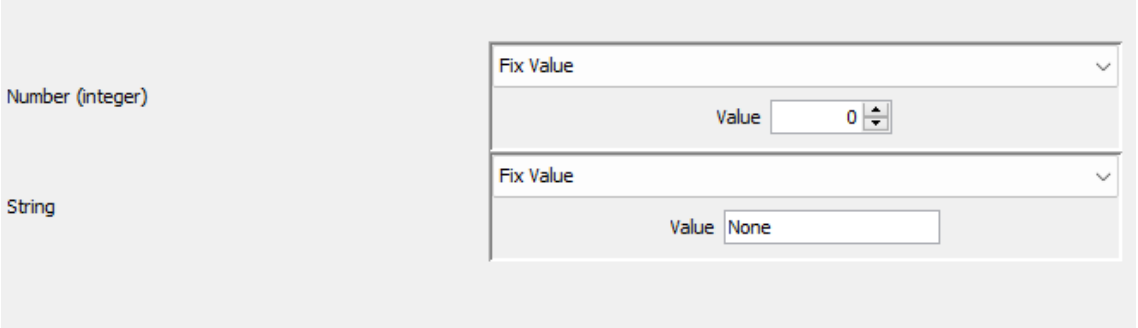


Figura 6 - configs -missing values

Configurações para *String Manipulation*:

No campo Expression adicionar:

```
replace(replace(replace(replace(replace(replace(replace($days_per_week$,
"Mon", "Monday"), "Tue", "Tuesday"), "Wed", "Wednesday"), "Thu", "Thursday"),
"Fri", "Friday"), "Sat", "Saturday"), "Sun", "Sunday")
```

Configurações para o Primeiro String to Date&Time:

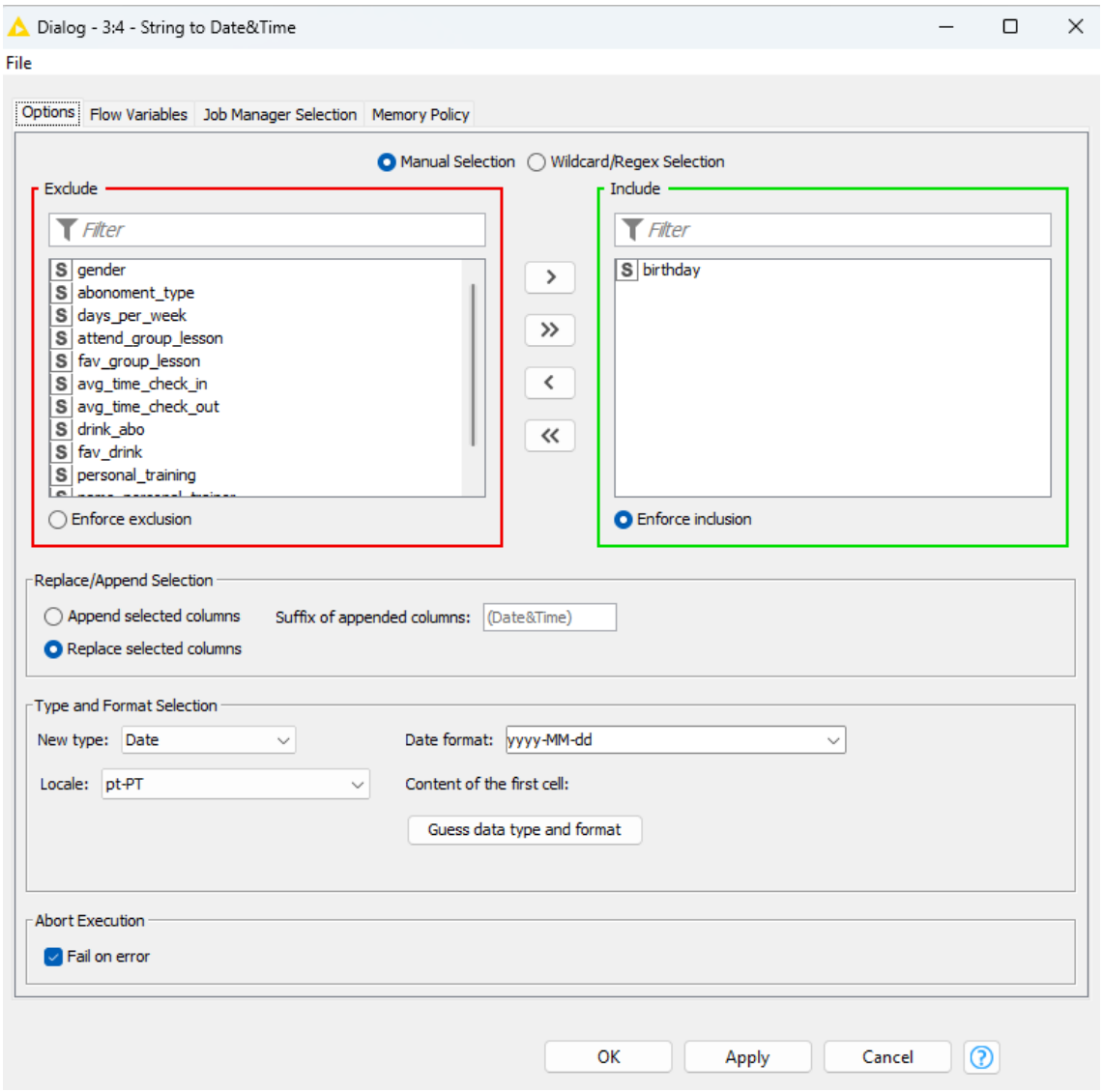


Figura 7 - configs String to Date&Time

Neste caso o Objetivo é tratar o dia de anos de cada Cliente.

Configurações do Segundo String to Date&Time:

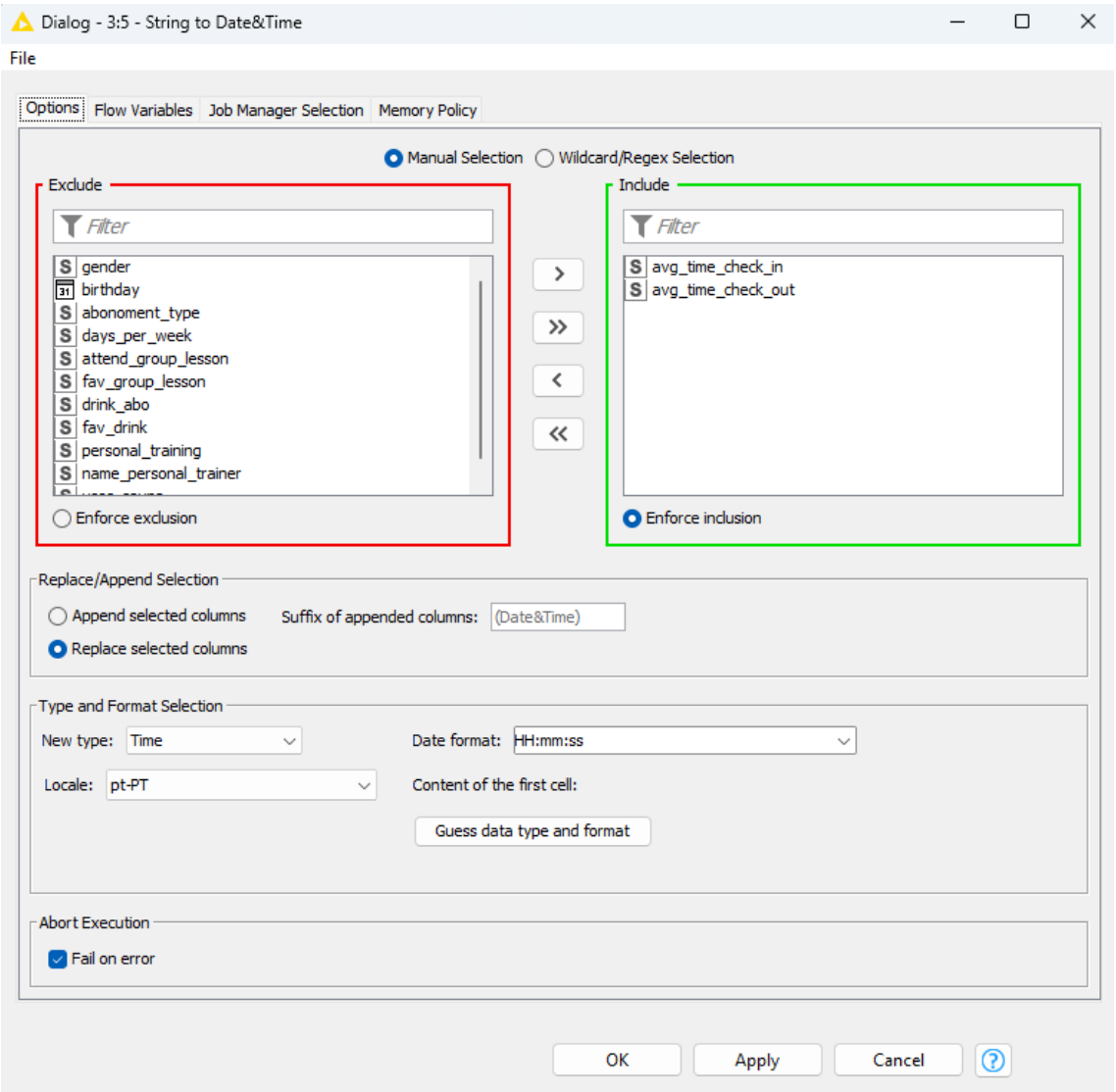


Figura 8 - configs string to DAte&Time[2]

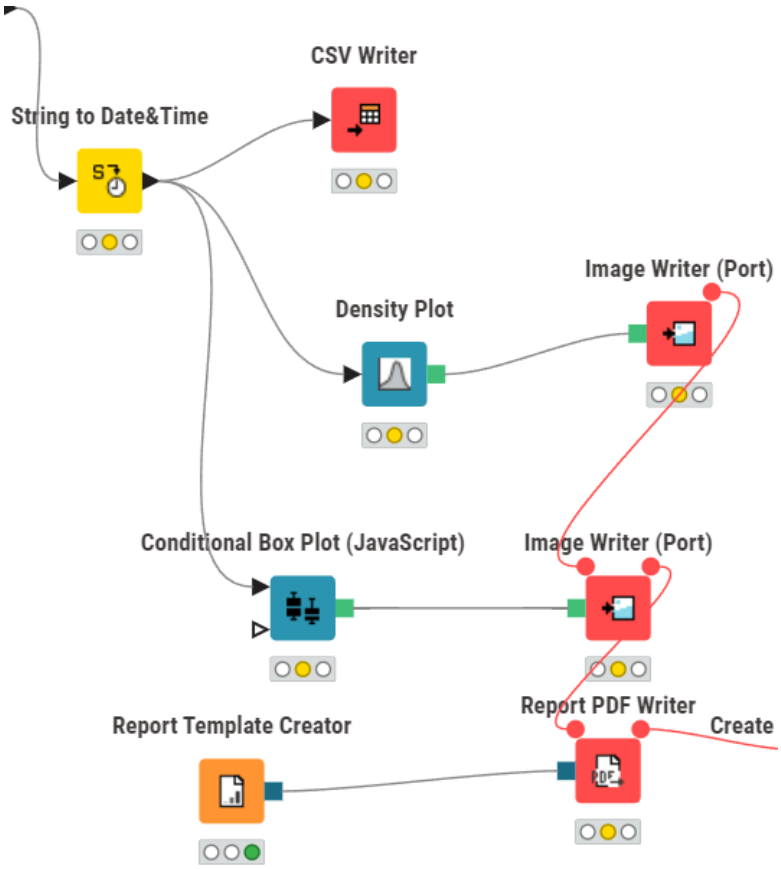
O objetivo deste campo é conseguir facilitar a leitura de quando os clientes entram e saem em media para cada um.

Após a execução existe assim um CSV Write para conseguir criar um ficheiro CSV.

Criação de gráficos/imagens e template em PDF

Após O tratamento dos dados temos assim de gerar algum valor ou em palavras mais curtas, algo para com que podemos analisar e termos algumas decisões sobre os dados. Neste caso eu decidi assim dois gráficos para tomadas decisões

Foram feitos assim 2 gráficos: um density plot, para verificar as idades de todos os membros do ginásio e ter melhor decisões no marketing e promoções/eventos internos, o segundo gráfico gerado foi um conditional box plot, foi feito para avaliar o tempo médio de cada género no ginásio, entre homens e mulheres.



Gráficos

A configuração do primeiro definitivamente é mais simples que o primeiro fazendo com que seja implementado a geração do mesmo em pouco tempo. Apenas configurar uma condição de coluna para Age, ou seja, usar a coluna das idades para a implementação e por fim gera rapidamente.

Figura 10 configs density plot

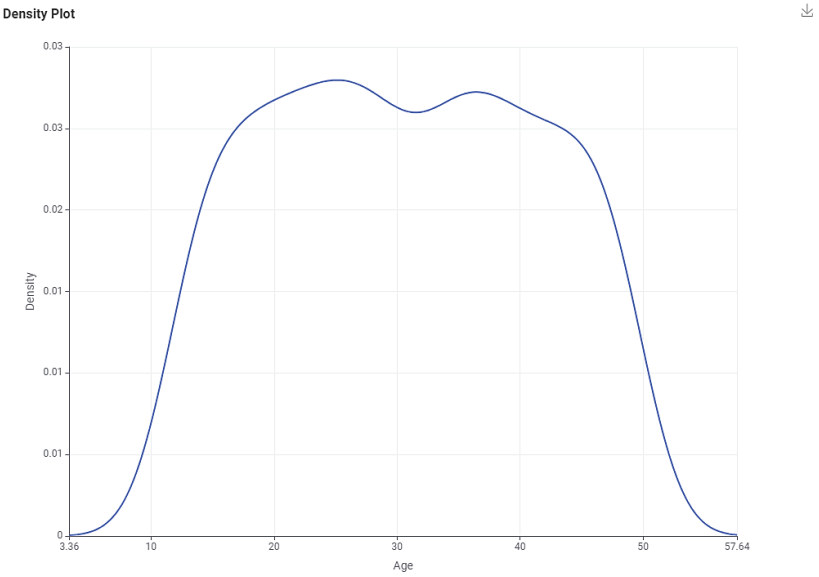


Figura 9 density plot

O segundo Gráfico com temos duas variáveis: Homens e Mulheres, já temos que usar mais variáveis para gerar o mesmo, definir o género (gender) como coluna de categoria e usar o a coluna avg_time_in_gym como Selected column.

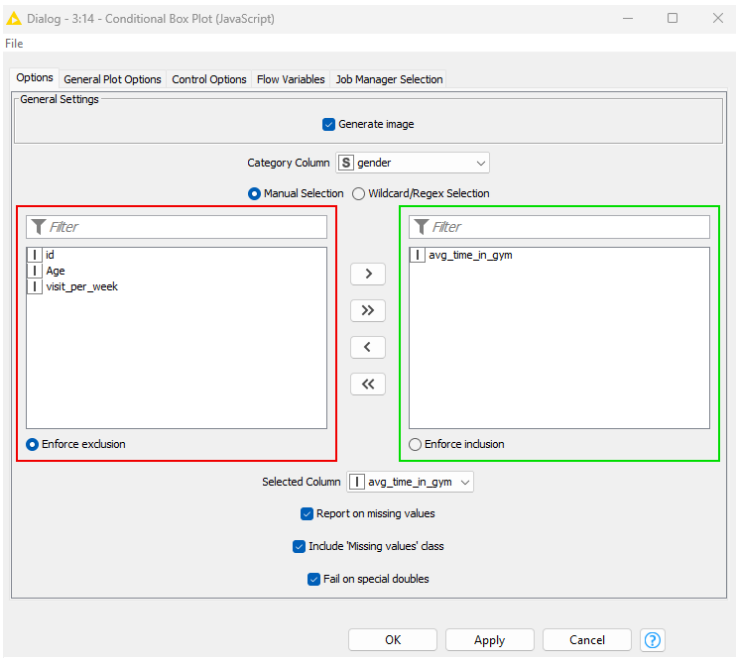


Figura 12 configs conditional graph

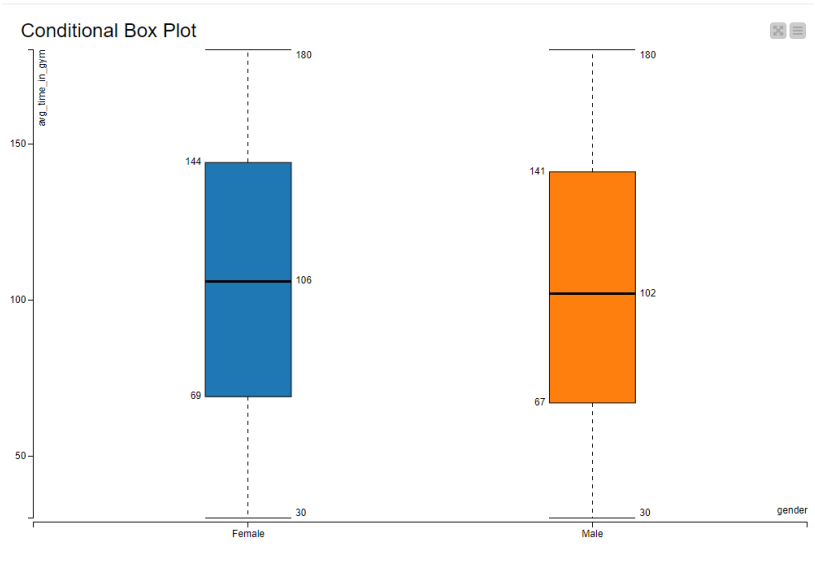
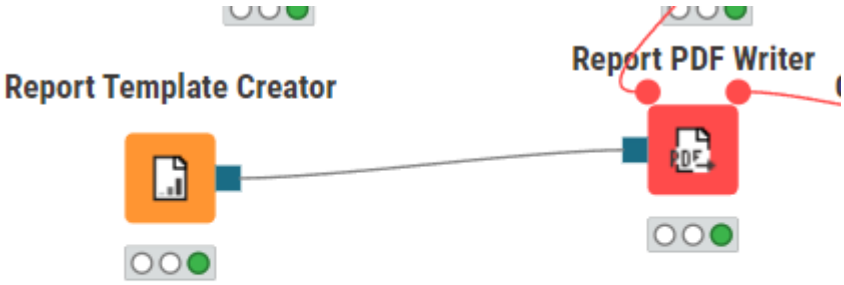


Figura 11 conditional box plot

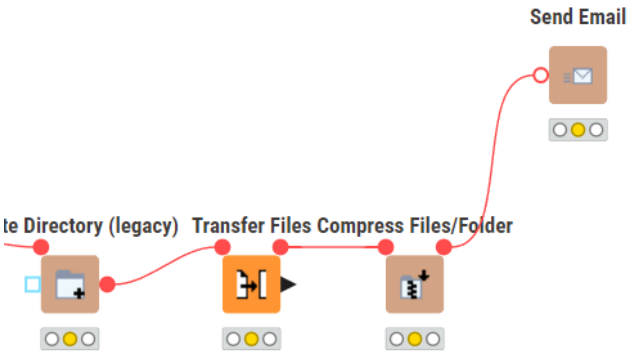
De seguida foi usado o componente ImageWriter para gerar a imagem em SVG e termos o ficheiro da mesma. Basta apenas colocar o caminho que queremos que a mesma seja gerada.

Temos também uma conexão quando a última imagem é finalizada para a criação de um report em PDF e é guardado localmente para o job seguinte.

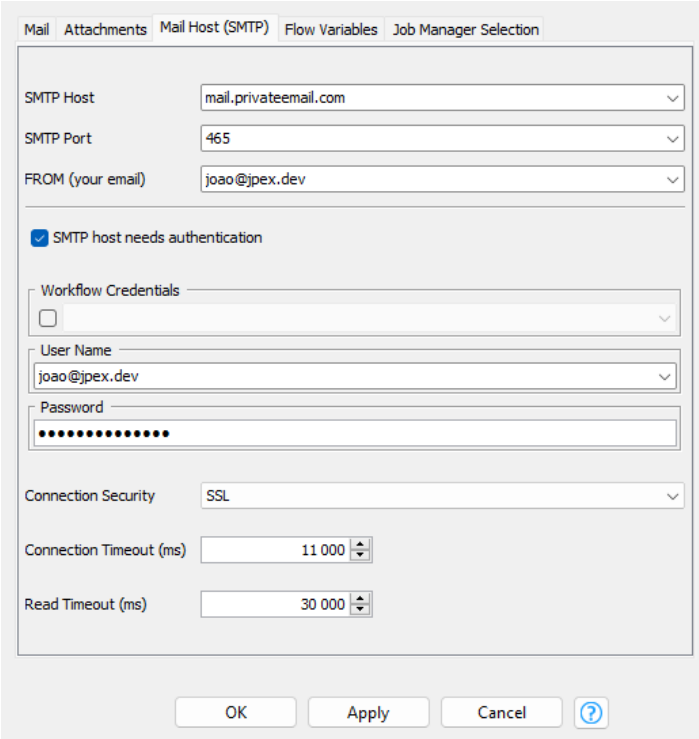


Job

O job foi elaborado com objetivo de no final de todo o flow de tratamento de dados conseguir "guardar" ou partilhar com alguém. Ele primeiramente cria uma diretoria para á posterior todos os ficheiros que usamos anteriormente sejam movidos para lá. Sendo os ficheiros as imagens, o report e o ficheiro CSV. Sendo assim temos uma conexão ligada e iremos tornar tudo em um ficheiro zip, compreendemos tudo para ser melhor em questão de partilhar o trabalho anterior.



Para o envio do email configurei o SMTP para enviar de um email com o domínio pessoal para o email do IPCA. AS configurações do email:



Conclusão

Eu, no trabalho, admito que está básico com potencial de bastante desenvolvimento e poder aprofundar mais um pouco, sendo eu trabalhador-estudante questão de tempo fica um pouco mais difícil, mesmo assim sinto que consegui contribuir para o meu aprendizado e conhecimento. Foi um trabalho que na qual me ajudou a entender melhor sobre ETLs e com isto consegui agregar um pouco mais de valor para o meu trabalho diário na empresa visto que contribuiu como fullstack developer e consegui entender melhor como poderei usar isto para simplificar a elaboração de por exemplo gráfico e reportes de vendas e aplicar os mesmo em Web Apps.

Gostei de fazer o trabalho pois é algo que consigo me identificar bastante não só a nível de desenvolvimento, mas também a nível de tema.

Dificuldades Encontradas

Uma das maiores dificuldades que encontrei neste trabalho que me desafiou os meus conhecimentos foi também um momento engraçado que nos faz pensar nesta área como algo que temos de estar bastante atentos a todo o momento. Foi algo com a Ferramenta do Knime e o seu Send Email. Basicamente defrontei me com 4 horas de trabalho dividido em 4 dias para tentar perceber um erro que me mostrava quando o send email não funcionava. Neste caso um dia cheguei ao trabalho e fiz o download do knime, e com o componente de Send email em 5 minutos consegui enviar um email. Com isso tirei uma grande conclusão, só poderá ser da firewall do computador. Visto isso, nesse mesmo dia quando cheguei a casa desabilitei as firewalls todas e consegui enviar o email, uma por uma fui habilitando e descobri assim qual firewall estava-me a bloquear o envio do email.

Outro Desafio foi gerar o segundo gráfico que foi tentar perceber quais colunas devia usar, demorei relativamente 1 hora para conseguir perceber como o Knime constrói este tipo de Gráficos.

QRCodes

Figura 13 - Drive

Figura 14 - GitHub