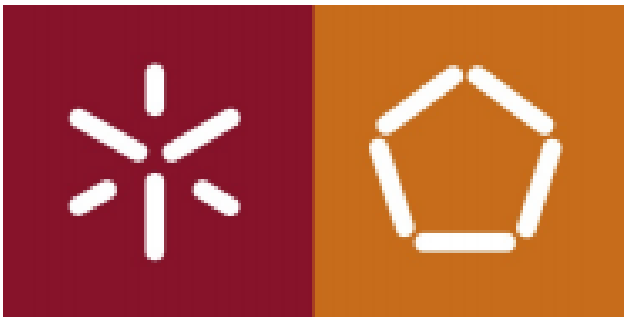


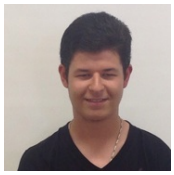
Universidade do Minho

Relatório do Trabalho Prático de Análise de Dados

Facebook Metrics

Hugo Carvalho (A74219)

João Almeida (A75209)



Luís Lima (A74260)

12 de Janeiro de 2018

Resumo

Este documento descreve o trabalho prático desenvolvido no âmbito da unidade curricular de **Análise de Dados**, do perfil de especialização de **Engenharia de Conhecimento**, tendo como objetivo a implementação de um sistema de apoio à decisão onde são aplicados métodos de *Business Intelligence*.

Neste estudo realizou-se uma análise das métricas de performance, referentes às publicações da página de facebook de uma empresa de cosméticos, durante o período de um ano.

Conteúdo

Lista de Figuras

Lista de Tabelas

1 Introdução

Com o acentuado crescimento do número de utilizadores dos mais variados serviços de internet, torna-se cada vez mais importante que as empresas estejam presentes nestes ambientes, possibilitando assim uma ligação mais forte aos seus clientes. Neste sentido, é possível afirmar assertivamente que as redes sociais se tornaram num meio de comunicação muito importante para grande parte dos negócios. De um leque alargado de redes sociais, é essencial identificar o Facebook como o meio de comunicação mais utilizado e com mais forte impacto nos utilizadores.

Esta solução de divulgação comercial, para além de ser do agrado das empresas, uma vez que permite alcançar um maior número de utilizadores, é também bastante apreciada pelos utilizadores pois apresenta facilidade de acesso à informação, bem como de possível interação através de comentários, gostos ou partilhas.

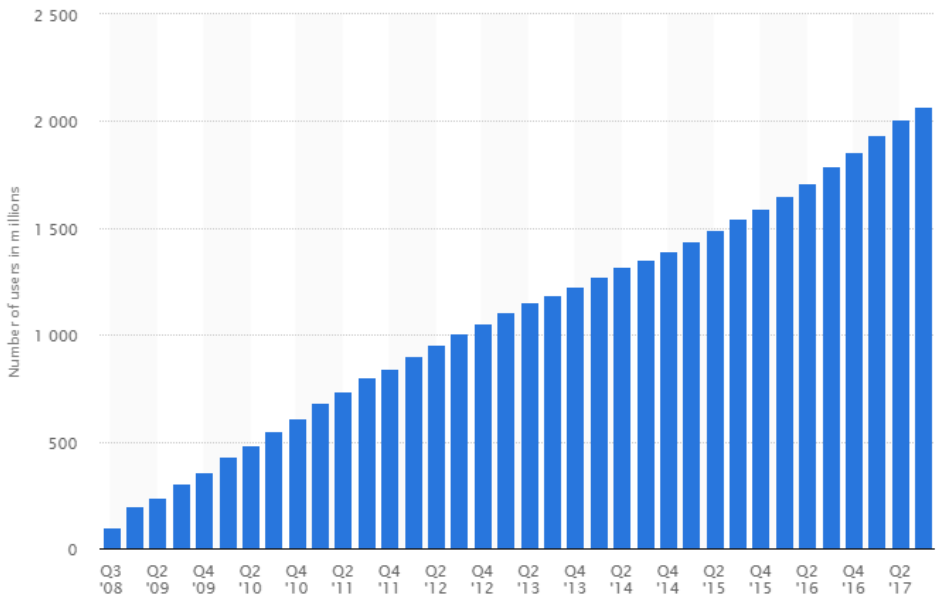


Figura 1: Crescimento do número de utilizadores da rede social Facebook [1]

Assim, com o objetivo de melhorar as estratégias de comunicação e interação com seus os clientes ou potenciais clientes, as empresas recorrem cada vez mais a métodos que permitam verificar o impacto que as suas publicações estão a causar nos utilizadores. Por exemplo, é de extremo interesse compreender qual a hora que as publicações têm maior número de visualizações/interações, qual o tipo de publicação mais apreciada, qual o dia de semana mais rentável, entre outras. Todas estas análises permitem que uma empresa consiga tomar melhores decisões nas suas publicações, melhorando assim todo o seu rendimento.

No estudo apresentado neste trabalho, analisou-se um conjunto de dados referentes a publicações da página de Facebook de uma empresa de cosmética, em que cada registo corresponde a uma publicação. Todas as informações correspondem a *posts* da empresa no ano de 2014.

2 Especificação do Dataset

2.1 Descrição

O *dataset* "Facebook metrics"é constituído por 500 de um total de 790 registos referentes a publicações na página de Facebook de uma empresa de cosmética. Deste modo, cada linha corresponde a dados relativos a um único *post*, contendo informações da publicação, bem como do impacto que esta causou nos utilizadores.

Cada registo contém um total de 19 atributos, sendo que 12 destes servem para avaliar o impacto do *post*, enquanto os restantes são posteriores e identificativos da publicação. É importante realçar que todos os registos se referem ao período compreendido entre 1 de Janeiro de 2014 e 31 de Dezembro de 2014.

Todos estes atributos são devidamente analisados e descritos nas subsecções seguintes.

2.2 Identificação de Atributos

Com o objetivo de alcançar uma correta interpretação e compreensão do conjunto de dados apresentado, efetuou-se uma análise sobre cada uns dos atributos existentes.

Neste sentido, apresenta-se em seguida uma tabela com uma breve descrição sobre cada um dos atributos presentes nos registos do *dataset*.

Atributo	**Descrição**
Page Total Likes	Número total de gostos que a página tinha antes da publicação
Type	Tipo de publicação
Category	Tipo de Campanha correspondente à publicação
Post Month	Mês da publicação
Post Weekday	Dia da semana em que a publicação foi feita
Post Hour	Hora de publicação
Paid	Inteiro que refere se foi gasto dinheiro em publicidade para a publicação
Lifetime Post Total Reach	Número de utilizadores que visualizaram a publicação
Lifetime Post Total Impressions	Número total de vezes que o *post* foi exibido, independentemente se foi clicado ou não. Uma publicação pode ser visualizada várias vezes pelo mesmo utilizador
Lifetime Engaged Users	Número de utilizadores que clicaram no *post* (utilizadores únicos)
Lifetime Post Consumers	Número de utilizadores que clicaram no *post*
Lifetime Post Consumptions	Número total de clicks no post
Lifetime Post Impressions by people who have liked your page	Número total de impressões apenas de utilizadores que gostaram da página
Lifetime Post Reach by people who like your page	Número de utilizadores que visualizaram a publicação e gostaram da página
Lifetime people who have liked your page and engaged with your post	Número de utilizadores que gostaram da página e clicaram no post
Comments	Número total de comentários da publicação
Likes	Número total de gostos da publicação
Shares	Número total de partilhas da publicação
Total Interactions	Soma dos totais de comentários, gostos e partilhas

Tabela 1: Descrição dos Atributos

Analisando a tabela, verifica-se que todos os valores são adquiridos diretamente da recolha de dados da rede social Facebook, à exceção do atributo "category". Este representa informação fornecida diretamente pela empresa em causa, correspondendo à categoria interna da campanha publicada.

Por outro lado, é possível verificar a existência de dados prévios à publicação e de dados posteriores à mesma.

Assim, identificam-se 3 tipos de dados distintos:

- Data - dados correspondentes à data em que a publicação foi feita, sendo responsáveis por identificar um determinado *post*.
- Categoria - conjunto de métricas que permitem caracterizar a publicação.
- Performance - métricas que permitem verificar o impacto da publicação nos utilizadores.

Seguidamente, apresenta-se uma tabela que permite uma melhor compreensão do tipo de informação e tipo de dados que cada atributo representa:

Atributo	**Tipo de Informação**	**Tipo de Dados**
Page Total Likes	Performance (Página)	Numeric
Type	Categoria	{Link, Photo, Status, Video}
Category	Categoria	Numeric: {1 - action, 1 - product, 2 - inspiration}
Post Month	Data	Numeric: [1-12] [Janeiro-Dezembro]
Post Weekday	Data	Numeric: [1-7] [Domingo-Sábado]
Post Hour	Data	Numeric: [0-23]
Paid	Categoria	Numeric: {0 - No, 1 - Yes}
Lifetime Post Total Reach	Performance	Numeric
Lifetime Post Total Impressions	Performance	Numeric
Lifetime Engaged Users	Performance	Numeric
Lifetime Post Consumers	Performance	Numeric
Lifetime Post Consumptions	Performance	Numeric
Lifetime Post Impressions by people who have liked your page	Performance	Numeric
Lifetime Post Reach by people who like your page	Performance	Numeric
Lifetime people who have liked your page and engaged with your post	Performance	Numeric
Comments	Performance	Numeric
Likes	Performance	Numeric
Shares	Performance	Numeric
Total Interactions	Performance	Numeric

Tabela 2: Análise dos Atributos

Considerou-se o atributo "Page Total Likes"pertencente ao grupo de dados "Performance", no entanto é importante realçar que este é referente a informação prévia à publicação, sendo um atributo que poderá influenciar os restantes fatores de performance. Por outro lado, este atributo não está diretamente associado a nenhum *post* em concreto, mas sim à página da empresa.

2.3 Relevância dos Atributos

Como referido anteriormente, é possível identificar dois distintos de grupos de dados que, num modelo de extração de conhecimento ou de mineração de dados, seriam considerados como métricas de *input* e *output*. Como o objetivo deste estudo passa por realizar uma recolha de informação eficaz, que permita que a empresa melhore a sua estratégia, torna-se fundamental esta distinção de atributos, bem como da sua respetiva relevância.

Assim, consideram-se dois tipos de informações:

- Dados Prévios à publicação (*input*): Dados do tipo Categoria, Data e "Page Total Likes";
- Dados Posteriores à publicação (*output*) - Todas as restantes métricas do tipo Performance;

Identificados os dois grupos de dados existentes, seguiu-se com uma análise mais aprofundada de cada um dos atributos de *output*. Desta forma, verificou-se que esta categoria se encontrava dividida em dois grupos internos: **visualizações** e **interações**. Todo este processo pode ser melhor compreendido e fundamentado com o auxílio da figura seguinte:

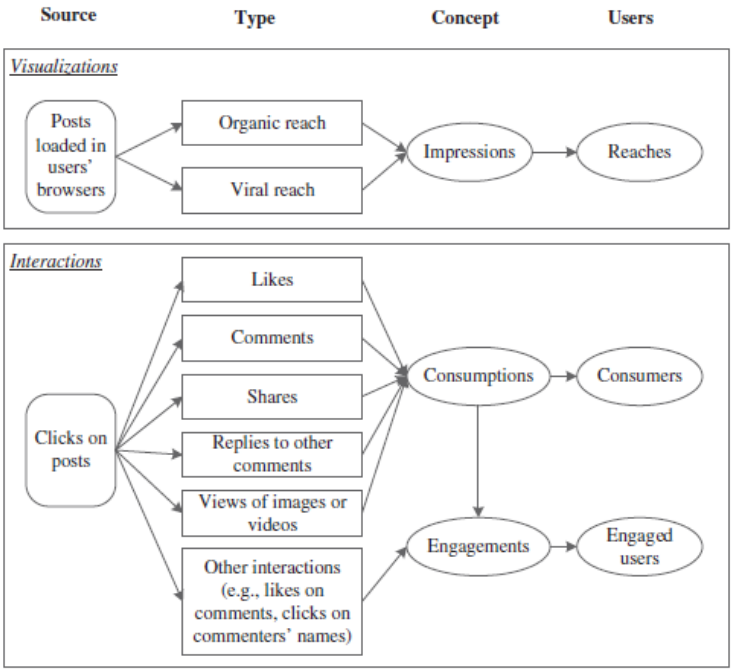


Figura 2: Estruturação das métricas de performance [2]

As visualizações correspondem ao número de vezes que um *post* foi apresentado ao utilizador de forma direta ("organic reach") ou através da sua interação ("viral reach"). Por outro lado, as interações representam os tipos e origens de todos os clicks nas publicações. Uma vez que estas representam ações diretas do utilizador, isto é, com intencionalidade, é seguro afirmar que correspondem a métricas de performance muito mais fortes do que quando comparadas com as visualizações.

2.4 Considerações Adicionais

Considerando o tópico "Criação de outras fontes de informação"presente no enunciado do trabalho prático, o grupo optou por dividir o *dataset* em cada um dos semestres do ano, possibilitando assim a existência de fontes diferentes de informação: uma metade carregada inicialmente para a base de dados e a outra carregada posteriormente através do ficheiro *csv*.

3 Base de Dados

Com o intuito de organizar a informação do *dataset* e assim construir uma fonte de dados para o datawarehouse, elaborou-se uma base de dados relacional de acordo com os atributos descritos e seus respetivos relacionamentos.

Ao longo desta secção será descrito todo o processo que levou ao desenvolvimento do modelo relacional, bem como da sua implementação e povoamento.

3.1 Identificar os tipos de entidades

De forma a conseguirmos identificar as diferentes entidades que a base de dados engloba, foi necessário determinar quais os objetos que se enquadram nesta definição. Para isso, após a análise do *dataset* e sua especificação, concluiu-se que as entidades fundamentais são: post e classe.

De seguida é apresentada uma tabela com a descrição de cada uma destas entidades, bem como os seus sinónimos e as suas ocorrências.

Atributo	**Descrição**	**Sinónimos**	**Ocorrências**
Post	Publicação realizada na página de Facebook da empresa	Publicação	Cada *post* regista um conjunto de métricas que permitem identifica-lo e analisar a sua performance
Classe	Categoria em que um determinado *post* está inserido	Categoria, Grupo	Uma categoria pode ter várias publicações associadas a si. Regista o tipo de *post*, categoria interna da empresa e informação se a publicidade foi paga

Tabela 3: Tabela das Entidades

3.2 Identificar os tipos de relacionamentos

Após identificadas as entidades da nossa base de dados, é necessário detetar todos os relacionamentos existentes entre as mesmas. A consulta da tabela 3 permite identificar um único relacionamento que as duas entidades estabelecem entre sim, bem como a sua respetiva cardinalidade.

Entidade	**Multiplicidade**	**Relacionamento**	**Multiplicidade**	**Entidade**
Post	1	Associado	N	Classe

Tabela 4: Tabela dos Relacionamentos

3.3 Associação entre atributos e entidades

Identificadas as entidades e seus relacionamentos, torna-se importante compreender a informação que conseguimos reter sobre as diferentes entidades. Esta informação foi recolhida de acordo com a sua importância para caracterizar cada entidade e também com a relevância que esta pode ter na base dados.

Em seguida apresentam-se os atributos pertencentes a cada entidade.

Nome da Entidade	Atributo	Descrição	Tipo e Tamanho	Nulo	M.V	Deriv.	Comp.
Class	id_Class	Número que identifica a classe	Valor inteiro positivo	Não	Não	Não	Não
	type	Tipo da classe	45 caracteres variáveis	Não	Não	Não	Não
	paid	Identifica se classe é paga ou não	Valor inteiro positivo	Não	Não	Não	Não
	category	Número que identifica a categoria	Valor inteiro positivo	Não	Não	Não	Não
Post	id_Post	Número que identifica o post	Valor inteiro positivo	Não	Não	Não	Não
	likes_page	Número de likes da página	Valor inteiro positivo	Não	Não	Não	Não
	date						Sim
	hour	Hora que post foi publicado	Valor inteiro positivo	Não	Não	Não	Não
	weekday	Dia da semana em que o post foi publicado	Valor inteiro positivo	Não	Não	Não	Não
	month	Mês que post foi publicado	Valor inteiro positivo	Não	Não	Não	Não
	Performance						Sim
	num_likes	Número de likes do post	Valor inteiro positivo	Não	Não	Não	Não
	total_interations	Número total interações no post	Valor inteiro positivo	Não	Não	Não	Não
	num_shares	Número de partilhas do post	Valor inteiro positivo	Não	Não	Não	Não
	num_comments	Número de comentários no post	Valor inteiro positivo	Não	Não	Não	Não
	lifetime_post_ total_impressions	Número total vezes que post foi exibido	Valor inteiro positivo	Não	Não	Não	Não
	lifetime_people_ liked_page_engaged	Número utilizadores que gostam da página e clicaram no post	Valor inteiro positivo	Não	Não	Não	Não
	lifetime_post_reach _people_liked_page	Número utilizadores que visualizaram o post e que gostam da página	Valor inteiro positivo	Não	Não	Não	Não
	lifetime_post_ impressions_ people_liked_page	Número impressões de utilizadores que gostaram da página	Valor inteiro positivo	Não	Não	Não	Não
	lifetime_post _consumptions	Número de clicks no post	Valor inteiro positivo	Não	Não	Não	Não
	lifetime_post _consumers	Número utilizadores que clicaram no post	Valor inteiro positivo	Não	Não	Não	Não
	lifetime_engaged _users	Número utilizadores que clicaram no post(únicos)	Valor inteiro positivo	Não	Não	Não	Não
	lifetime_post _total_reach	Número utilizadores que visualizaram o post	Valor inteiro positivo	Não	Não	Não	Não

Tabela 5: Associação entre entidades e atributos

3.4 Modelo Conceptual

Apresenta-se, de seguida, o desenho do diagrama E-R (Entidade-Relacionamento) de forma a representar conceptualmente as relações entre as entidades da base de dados.

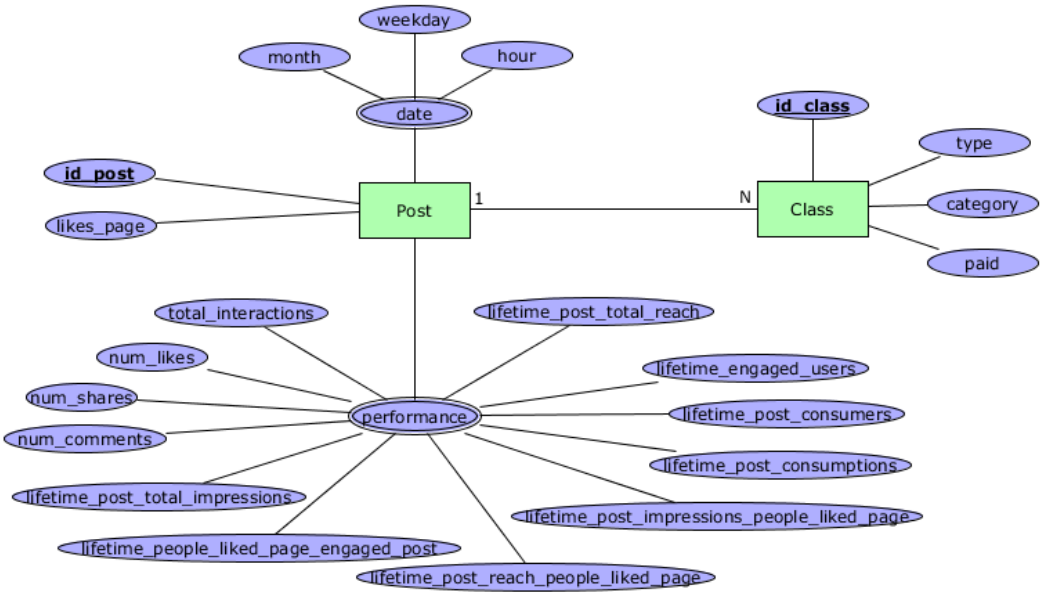


Figura 3: Desenho do Diagrama ER

3.5 Modelo Lógico

De modo a construir a base de dados de acordo com o diagrama ER apresentado anteriormente, derivou-se todas as relações existentes, obtendo assim o seu respetivo modelo lógico.

De seguida, apresentam-se uma imagem do esquema elaborado.

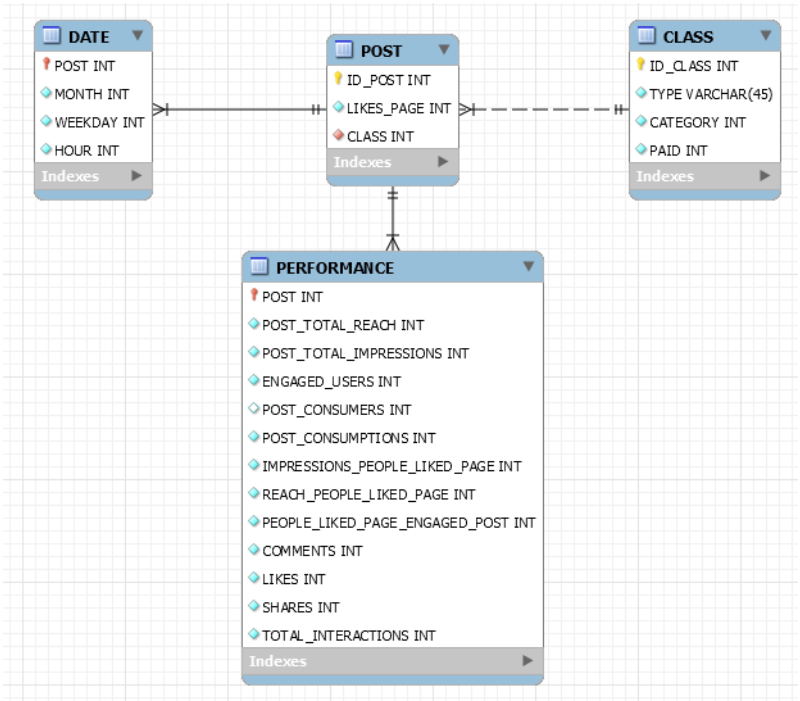


Figura 4: MySQL Workbench - Modelo Lógico da Base de Dados

3.6 Povoamento

Construída e implementada a base de dados, o passo seguinte passou por efetuar o seu povoamento de acordo com os dados presentes no *dataset*.

Como já referido anteriormente, o conjunto de dados foi dividido em duas partes distintas correspondentes a cada semestre do ano, onde ficou decidido que os dados do primeiro semestre seriam carregados diretamente para a base de dados relacional.

Nesse sentido, foi necessário efetuar uma interpretação dos dados e compreender como estes teriam de ser adaptados de modo a inseri-los corretamente na base dados. Assim, desenvolveram-se dois *jobs* recorrendo ao software *Talend Open Studio for Data Integration*, onde, através de processos ETL, os dados são filtrados e inseridos eficazmente.

Em primeiro lugar decidiu-se preencher os dados da tabela Class, onde são filtrados exclusivamente os seus três atributos do ficheiro *csv*. Como existiam vários dados repetidos destes valores, foi necessário eliminar estas igualdades, reduzindo os dados apenas para entradas únicas. Por último, gerou-se o valor da chave da tabela (id_class) de forma automática, recorrendo à expressão "Numeric.sequence"disponibilizada pelo *Talend Open Studio*. Todo este processo pode ser visualizado na figura seguinte:

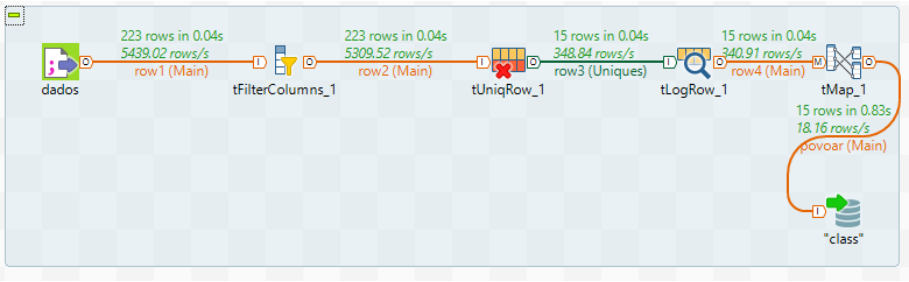


Figura 5: Talend Open Studio - Povoamento Class

Seguidamente, procedeu-se ao preenchimento das restantes tabelas correspondentes a informações relativas a cada publicação na página. Para isso, recorreu-se a um "tMap"para identificar a chave da "Class"a que cada *post* pertencia, substituindo-a pelos dados relativos a esse mesmo tipo de informação. Posteriormente, dividiu-se os dados pelas restantes tabelas, salientando a obrigatoriedade de povoar a tabela *post* em primeiro lugar devido às restrições de integridade apresentadas.

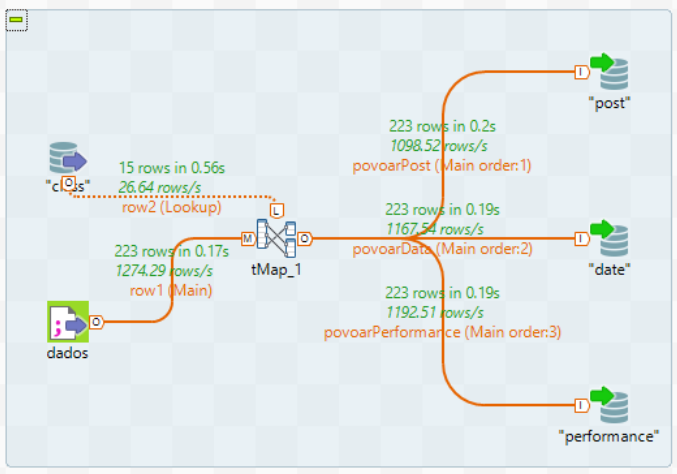


Figura 6: Talend Open Studio - Povoamento Post, Data e Performance

4 Data Warehouse

4.1 Funcionamento do Sistema

Como já referido anteriormente, é pedido no enunciado do projeto que o datawarehouse seja suportado por distintas fontes de informação. Assim, o grupo optou por dividir o *dataset* em dois semestres, carregando o primeiro diretamente para a base de dados e integrando posteriormente o segundo no datawarehouse.

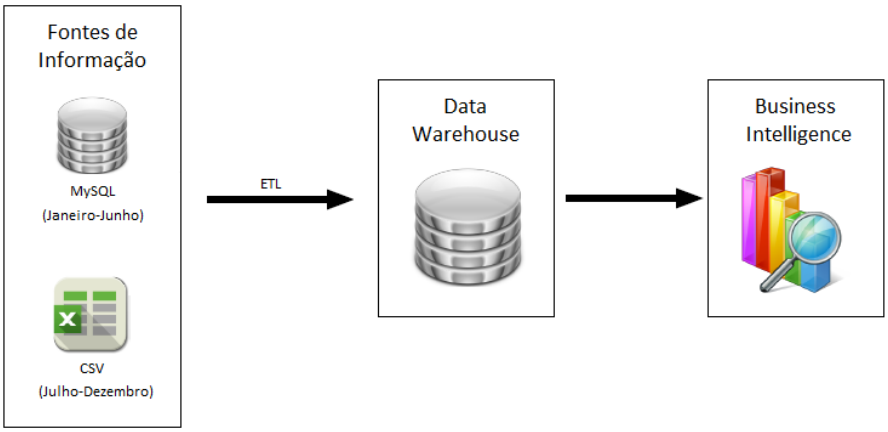


Figura 7: Sistema de Análise de Dados

4.2 Seleção de Dados

O objetivo da implementação deste sistema passa por conseguir adquirir informação útil e relevante, que permita a correta interpretação da performance da página da empresa. Neste sentido, torna-se necessário definir quais os dados que devem ser analisados, bem como as informações que se pretende retirar.

Analisando os atributos do *dataset*, verificou-se a existência de dois elementos fundamentais para compreender todo o rendimento que a página apresenta:

- **Interactions**: soma entre o total de gostos, partilhas e comentários.
- **Lifetime Post Consumers**: Número de consumidores, ou seja, número de pessoas que clicaram no *post*.

Assim, tendo em conta estes dois atributos e sua importância, elaborou-se uma lista de questões essenciais para obter os resultados pretendidos:

1. Qual o número de interações por hora?
2. Qual o número de interações por dia da semana?
3. Qual o número de interações por parte da semana (semana/fim de semana)?
4. Qual o número de interações por mês?
5. Qual o número de interações por trimestre?
6. Qual o número de interações por tipo?
7. Qual o número de interações por categoria?

8. Qual o número de interações por tipo e categoria?
9. Qual o número de interações por publicações pagas?
10. Qual o número de interações por número de gostos na página?
11. Qual o número de interações por alcance da publicação?
12. Qual o número de interações por impressões da publicação?
13. Qual o número de consumidores por hora?
14. Qual o número de consumidores por dia da semana?
15. Qual o número de consumidores por parte da semana (semana/fim de semana)?
16. Qual o número de consumidores por mês?
17. Qual o número de consumidores por trimestre?
18. Qual o número de consumidores por tipo?
19. Qual o número de consumidores por categoria?
20. Qual o número de consumidores por tipo e categoria?
21. Qual o número de consumidores por publicações pagas?
22. Qual o número de consumidores por número de gostos na página?
23. Qual o número de consumidores por alcance da publicação?
24. Qual o número de consumidores por impressões da publicação?

4.3 Datamarts

De acordo com a seleção de dados apresentada no subtópico anterior, verifica-se a existência de dois datamarts fundamentais, que serão apresentados e descritos em seguida.

4.3.1 Dimensões e Factos

Com base nas questões apresentadas no subtópico anterior, torna-se necessária a sua interpretação, de modo a definir corretamente as dimensões e os factos do modelo.

As dimensões são utilizadas com o objetivo de filtrar e categorizar os factos. Assim, é possível identificar as seguintes dimensões:

- **Data**: informações sobre o dia, mês, trimestre e parte da semana
- **Class**: informações sobre tipo de publicação, categoria e se corresponde a um *post* pago.
- **Página**: informação sobre o número de gostos da página
- **Metrics**: informações sobre o alcance e impressões das publicações

Seguindo o mesmo raciocínio de interpretação das questões, identificam-se dois factos que representam aquilo que se pretende analisar:

- **Interações**: Número total de interações dos *posts*
- **Consumidores**: Número total de consumidores dos *posts*

4.3.2 Esquema

Identificadas as dimensões e os factos necessários para a elaboração do sistema, construiu-se então um modelo em estrela. Optou-se por este tipo de modelo uma vez que todas as dimensões estão diretamente relacionadas com os factos e, por outro lado, permite uma melhor interpretação.

Dado que os dois factos apresentados se relacionam com as mesmas dimensões, o grupo optou por desenvolver o modelo com dimensões partilhadas.

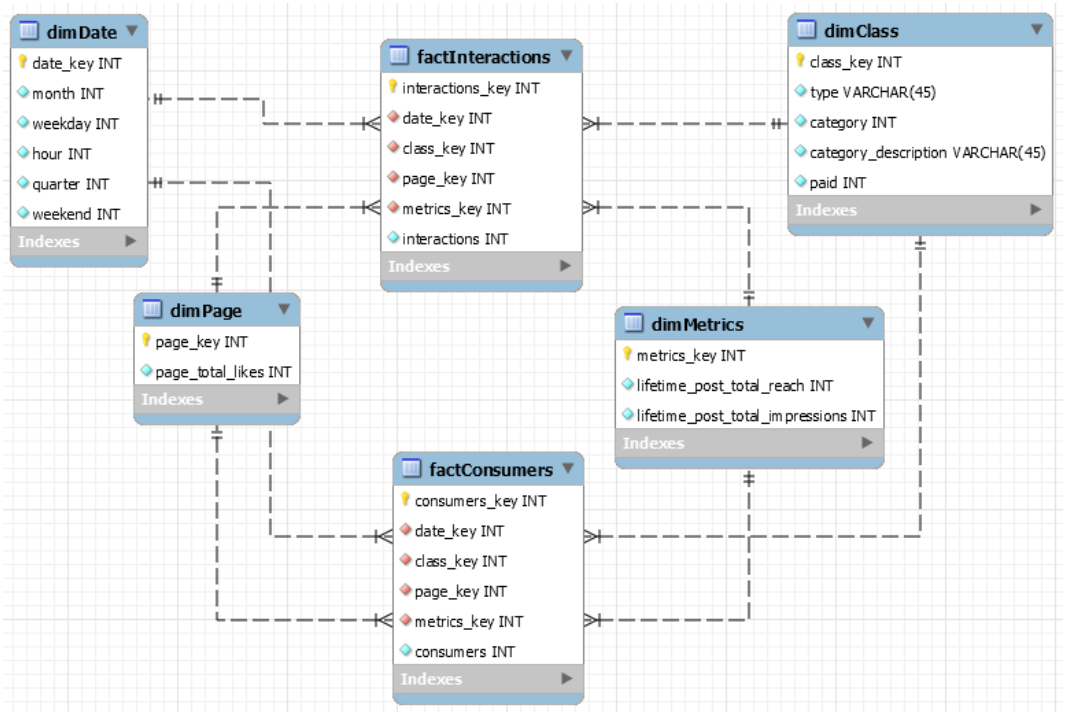


Figura 8: MySQL Workbench - Modelo Dimensional

4.4 Preenchimento

Implementado o modelo dimensional apresentado, o objetivo seguinte passou pelo seu povoamento com informação correspondente ao primeiro semestre do ano, ou seja, dados provenientes da base de dados relacional, e também pelos dados do segundo semestre do ano presentes num ficheiro *csv*.

Com essa finalidade, o grupo optou por usar o software *Pentaho Data Integration (Kettle)*, onde através de processos ETL efetuou a correlação de toda a informação, inserindo-a nos respetivos datamarts.

4.4.1 Dados MySQL

Em primeiro lugar, optou-se por inicializar o preenchimento do datawarehouse com os dados provenientes da base de dados relacional. Todas as dimensões foram corretamente povoadas, sendo posteriormente interligadas aos respetivos factos.

Para o preenchimendo da dimDate, foi necessário identificar o trimestre a que cada registo pertencia, adicionando o atributo "quarter"presente na tabela. Seguindo o mesmo raciocínio, verificou-se se a data do registo correspondia a um período de semana ou de fim de semana, adicionando também este novo atributo ("weekend"). Por último, os dados foram inseridos na

tabela, sendo que o valor de date_key é gerado automaticamente e sequencialmente pelo "Table Output"do *Pentaho*.

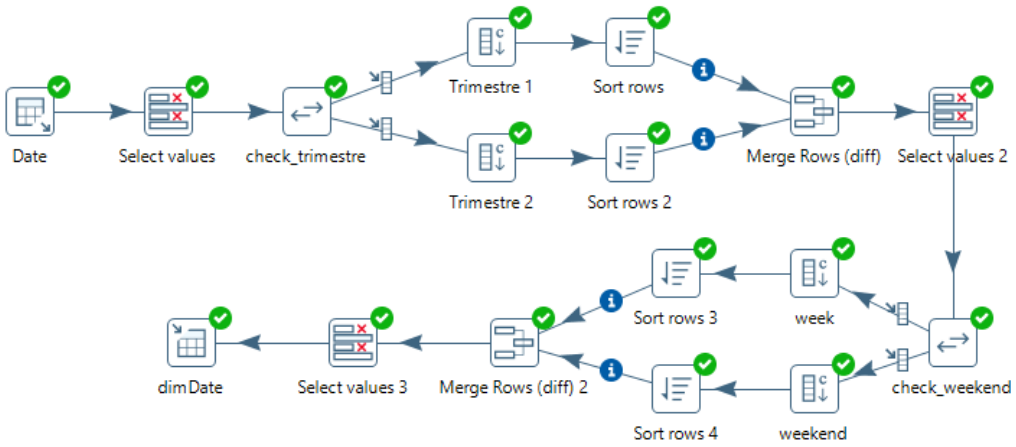


Figura 9: Pentaho Kettle - Preenchimento dimDate com informação da Base de Dados

As tabelas dimPage e dimMetrics apenas necessitaram de uma seleção de valores dos registos presentes nas tabelas Post e Performance da base de dados relacional, sendo diretamente carregados para as suas respetivas dimensões. É importante salientar que aqui também é gerada uma chave automática e sequencial, mantendo a ordem de acordo com as publicações e com a sua respetiva data.

Figura 10: Pentaho Kettle - Preenchimento dimPage com informação da Base de Dados

Figura 11: Pentaho Kettle - Preenchimento dimMetrics com informação da Base de Dados

O povoamento da tabela dimClass apenas precisa de recolher informação proveniente da tabela Class, onde é feita uma verificação da categoria a que pertence, de modo a adicionar o atributo "category_description", que toma os valores "action", "product"ou "inspiration". Como não há variação do tamanho da amostra de dados, então o atributo "class_key"segue o mesmo valor do atributo "id_class"da base de dados relacional.

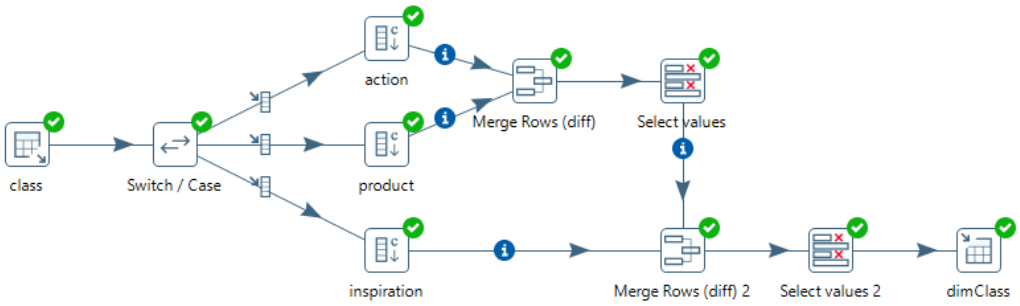


Figura 12: Pentaho Kettle - Preenchimento dimClass com informação da Base de Dados

Com todas as dimensões devidamente povoadas, procedeu-se ao preenchimento das tabelas referentes aos factos. Como os valores dos factos são fornecidos pela mesma tabela no modelo relacional, então o raciocínio é muito semelhante em ambos os casos.

Em primeiro lugar são filtrados os dados presentes na tabela Performance, selecionando apenas os valores que se pretendem considerar, ou seja, "total_interactions"ou "post_consumers". Em seguida, é feito um *left join*, onde para cada *post* se obtém o valor de performance pretendido. Como não existiram alterações nas chaves ao inserir nas dimensões, então os dados selecionados apresentam todos os atributos necessários (incluindo o id da tabela dimClass), podendo ser diretamente carregados para as respetivas tabelas de factos.

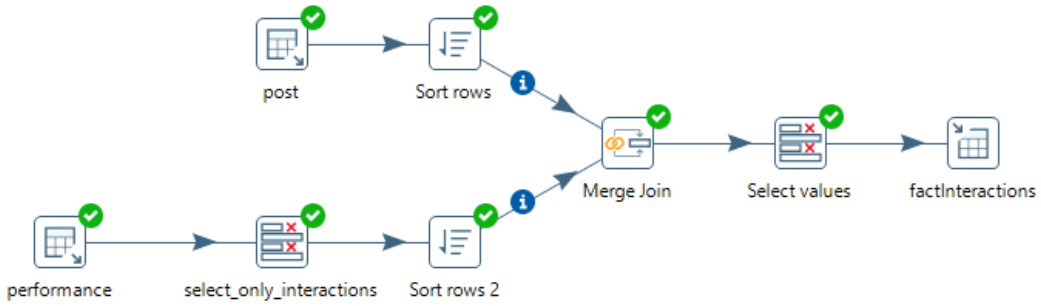


Figura 13: Pentaho Kettle - Preenchimento factInteractions com informação da Base de Dados

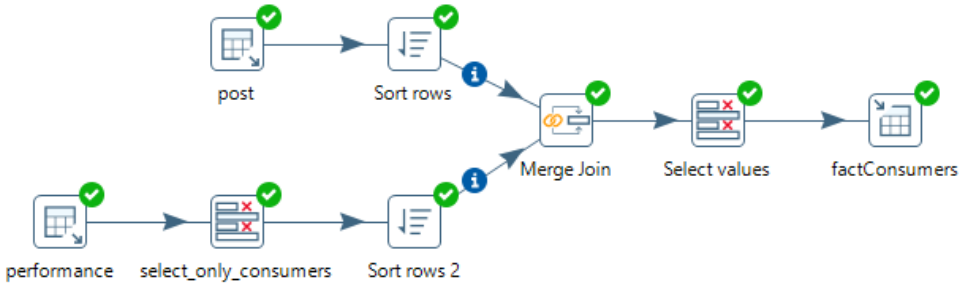


Figura 14: Pentaho Kettle - Preenchimento factConsumers com informação da Base de Dados

4.4.2 Dados CSV

Tendo em conta a segunda fonte de dados existente no sistema, que corresponde à informação relativa ao segundo semestre do ano, tornou-se necessário encontrar mecanismos que permitissem complementar estes dados com os já existentes no datawarehouse.

A principal diferença entre o povoamento da tabela dimDate com a informação da base de dados e com a informação do ficheiro *csv*, centra-se no facto de ser adicionado um id sequencial no início da *transformation*, de modo a assegurar que ao longo de todo o processo ETL é possível ordenar os dados de acordo com a ordem de publicação, tal como acontece no ficheiro *csv*. O restante raciocínio é semelhante ao povoamento através de dados do modelo relacional, salientando o facto do id gerado inicialmente ser removido antes da inserção no datawarehouse, gerando-se um novo "date_key"de forma automática e sequencial através do "Table Output".

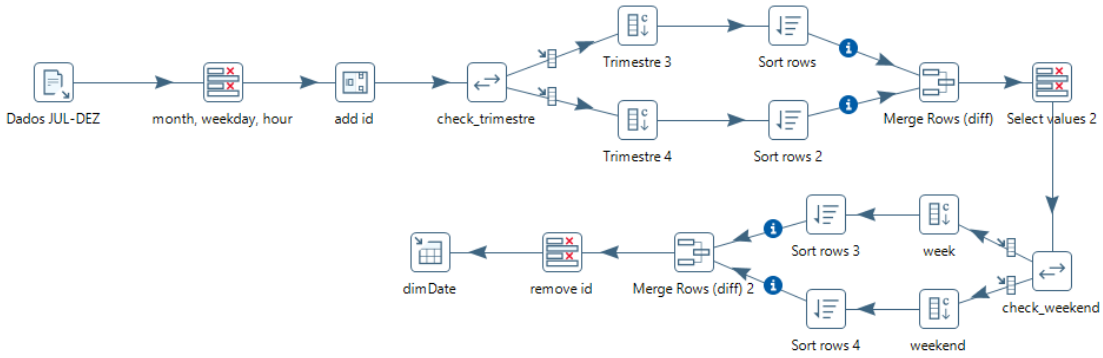


Figura 15: Pentaho Kettle - Preenchimento dimDate com informação do ficheiro *csv*

O povoamento das tabelas dimPage e dimMetrics é igualmente semelhante, utilizando a mesma criação de id explicada no caso anterior. Assim, apenas é selecionada informação importante do *csv*, nunca perdendo a sua ordem, neste caso "page total likes"e "reach impressions".

Figura 16: Pentaho Kettle - Preenchimento dimPage com informação do ficheiro *csv*

Figura 17: Pentaho Kettle - Preenchimento dimMetrics com informação do ficheiro *csv*

O preenchimento da dimensão de classe foi o que apresentou maior complexidade, uma vez que não se sabia se existiam novas classes nas publicações referentes ao segundo semestre do ano. Assim, foi necessário recolher todas as classes existentes no ficheiro, selecionando os atributos

significativos e da eliminando as suas repetições através do *step* "Unique rows". Em seguida, após a associar os valores do atributo "category_description", é feita uma comparação com as classes já existentes na tabela dimClass, filtrando-se apenas as que ainda não existem. Posteriormente, essas são adicionadas na tabela, sendo o atributo "class_key"gerado sequencialmente a partir do último valor existente para esse atributo.

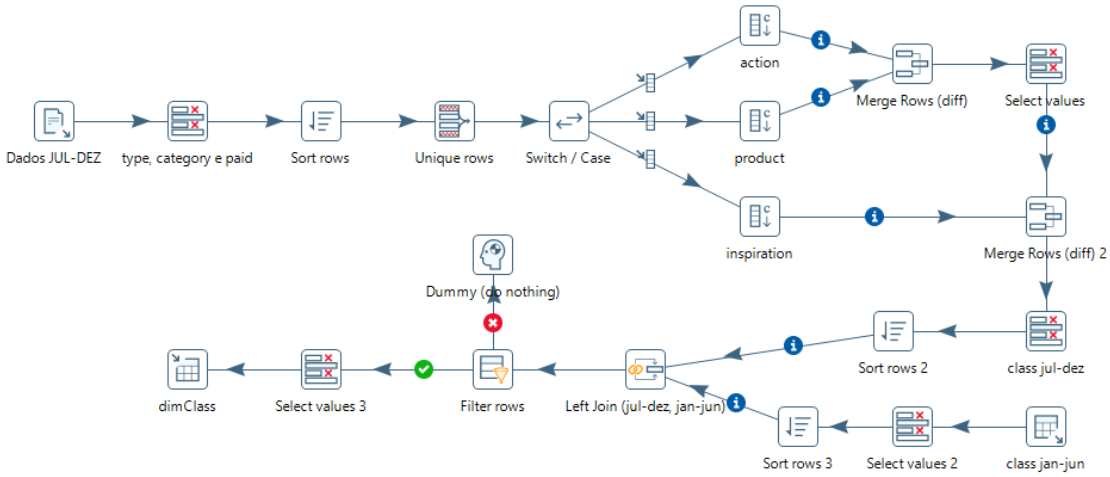


Figura 18: Pentaho Kettle - Preenchimento dimClass com informação do ficheiro *csv*

Por fim, o povoamento das tabelas relativas aos factos segue o raciocínio anteriormente apresentando para o preenchimento através de dados da base de dados relacional, salientando novamente o facto de ser inserido um id que permite não perder a ordem inicial dos registos provenientes do *csv*.

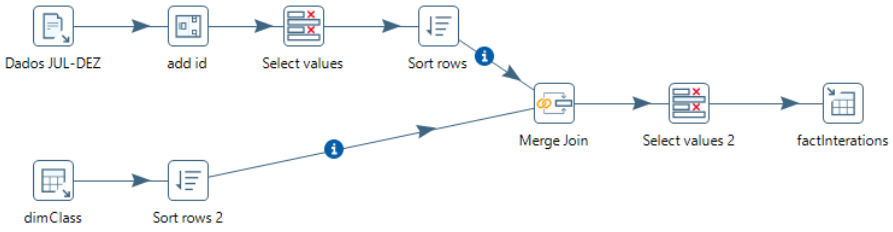


Figura 19: Pentaho Kettle - Preenchimento factInteractions com informação do ficheiro *csv*

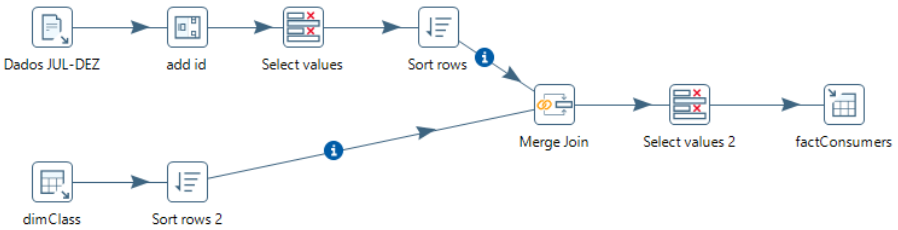


Figura 20: Pentaho Kettle - Preenchimento factConsumers com informação do ficheiro *csv*

5 Business Intelligence

Com o datawarehouse devidamente preenchido com a informação relevante, o passo seguinte da realização deste projeto centrou-se no recurso à ferramenta "Microsoft Power BI"para a análise de dados de uma forma interativa e visual.

Assim, com o objetivo de responder às questões apresentadas na seleção de dados, optou-se por desenvolver dois *reports*, referentes a cada um dos datamarts.

Por outro lado, realizou-se também uma breve análise à distribuição de *posts* por data e categoria, de modo a posteriormente interligar a informação na análise de resultados.

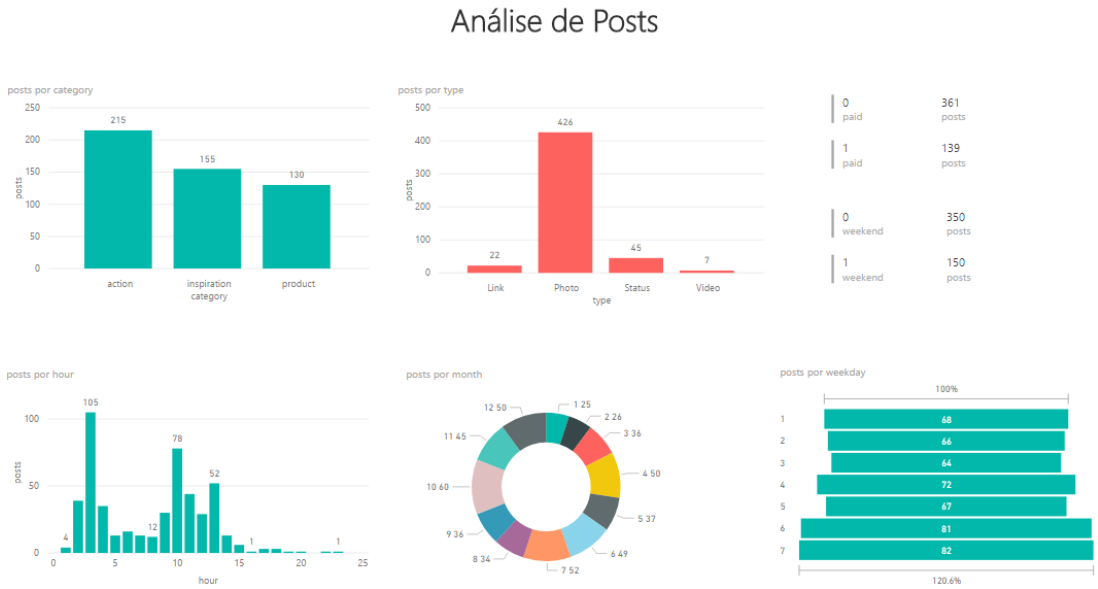


Figura 21: Power BI - Análise de *Posts*

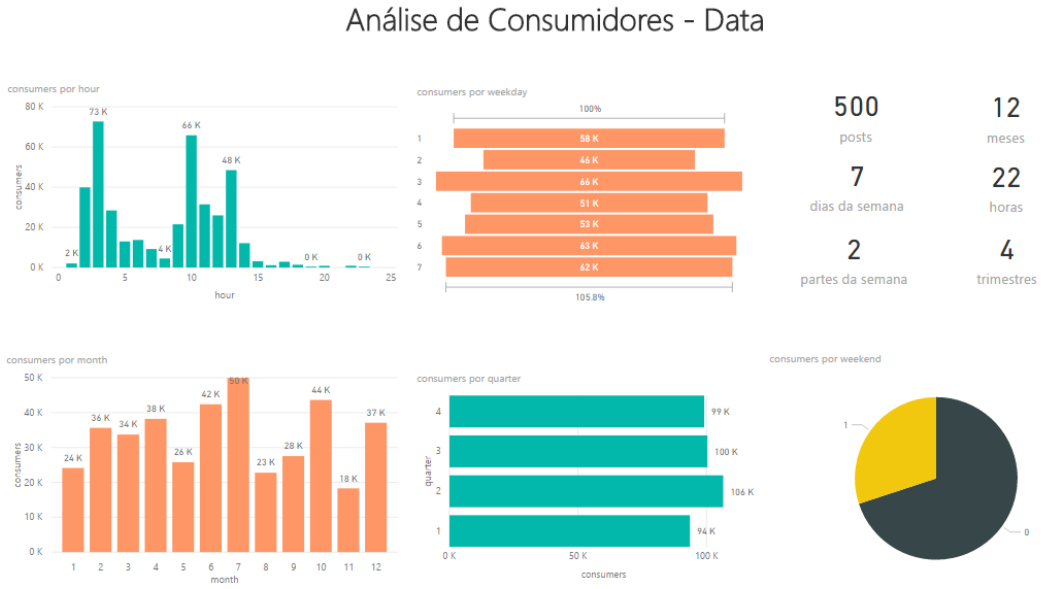


Figura 22: Power BI - Análise de Consumidores - Data

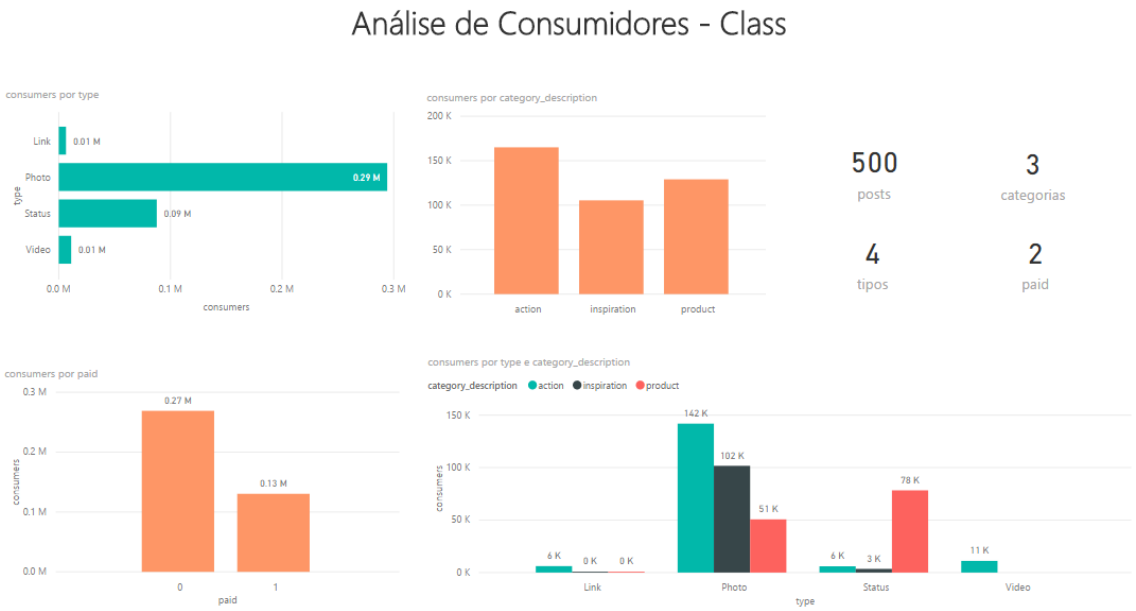


Figura 23: Power BI - Análise de Consumidores - Class

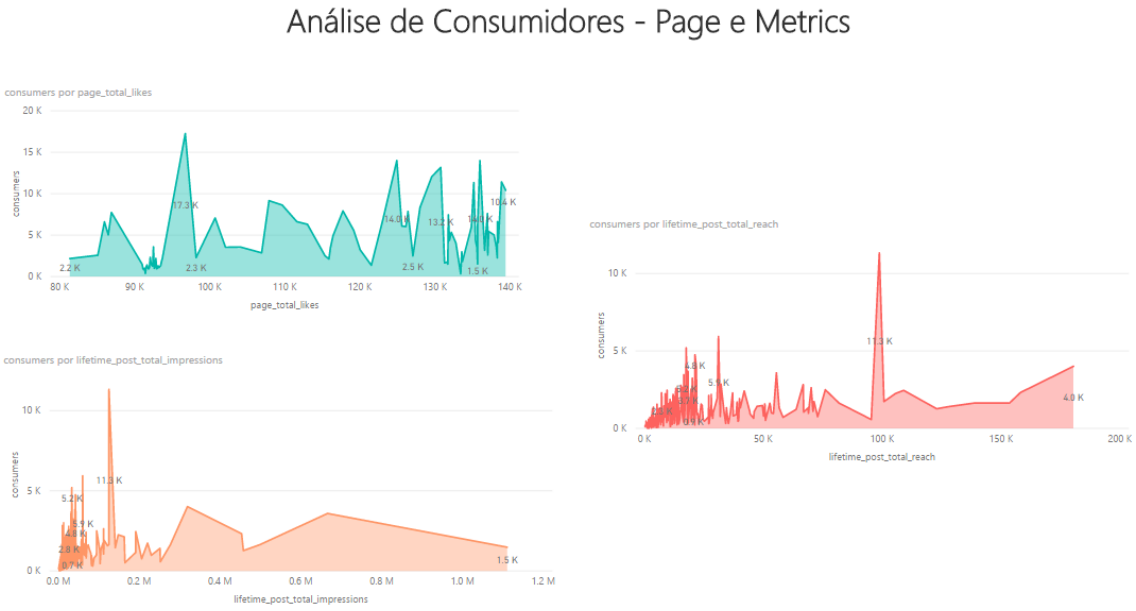


Figura 24: Power BI - Análise de Consumidores - Page e Metrics

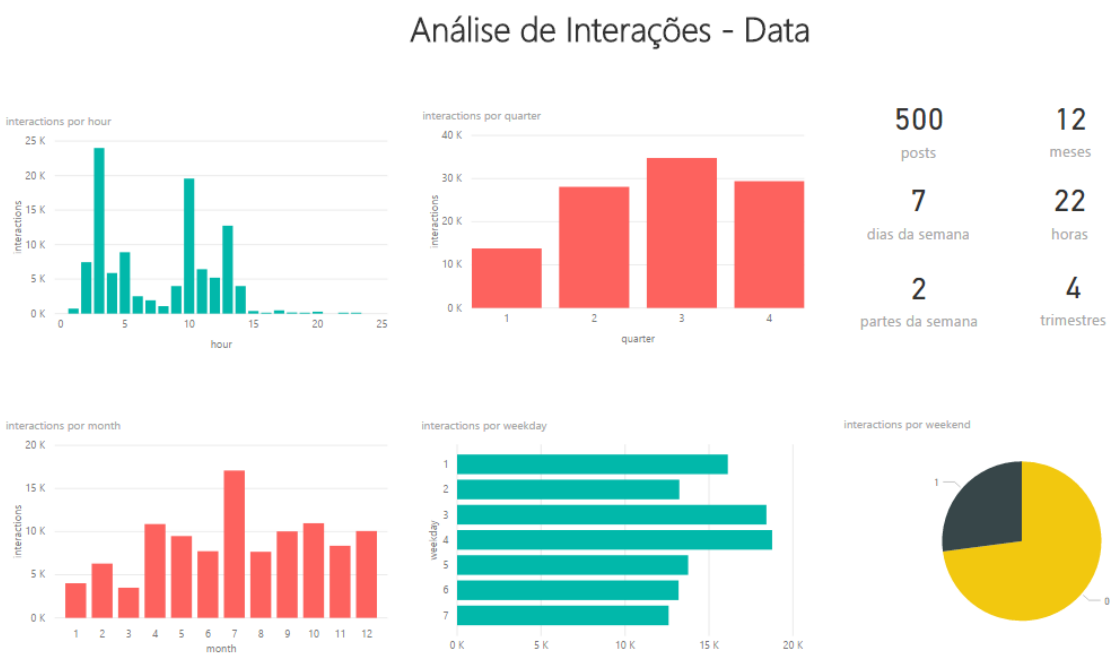


Figura 25: Power BI - Análise de Interações - Data

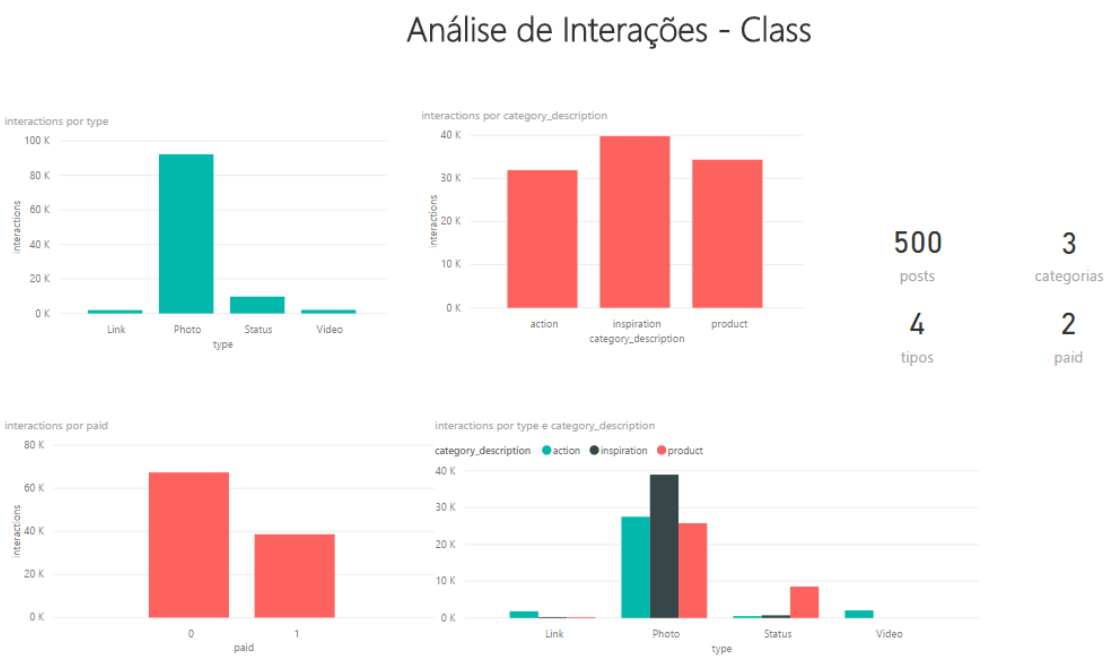


Figura 26: Power BI - Análise de Interações - Class

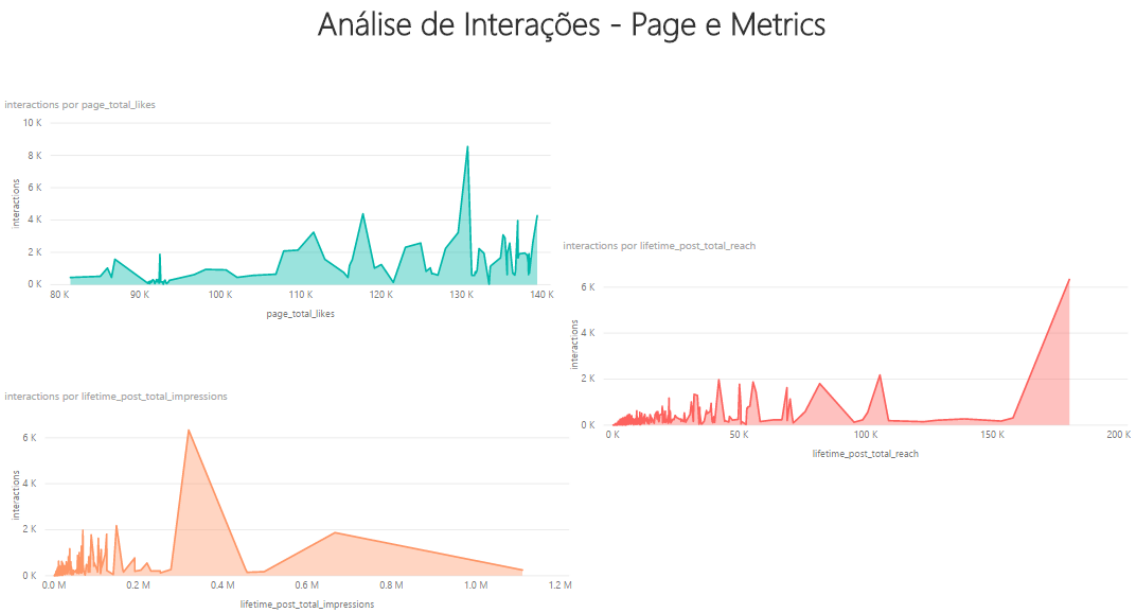


Figura 27: Power BI - Análise de Interações - Page e Metrics

6 Análise de Resultados

6.1 Interações

- **Qual o número de interações por hora?**

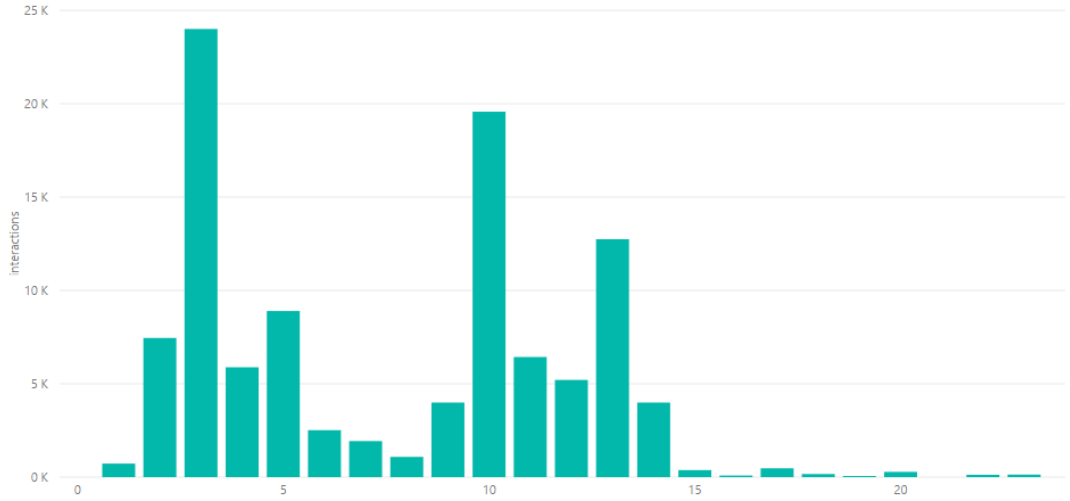


Figura 28: Power BI - Interações por Hora

Com a métrica de número de interações por hora, pretende-se saber qual/quais as horas do dia onde existe uma maior afluência de utilizadores em termos de gostos, partilhas e comentários.

- O maior pico de interações ocorre no valor de Hora 3, ou seja, as 3 da manhã onde um total de 24002 interações foram efetuadas por parte dos utilizadores.
- O segundo e terceiro maior pico ocorre ao valor de Hora 10 e 13 com 19571 e 12741 interações, respetivamente.
- O período de menor afluência regista-se entre as 16 e 24 horas da tarde, verificando-se um número muito reduzido de interações nos *posts*.

De um modo geral, analisando o gráfico de barras, facilmente se observa que os períodos de maior abundância de interações ocorrem desde a meia noite até as 14 horas, contrastando com o resto do dia em que praticamente não existem interações por parte dos consumidores.

[0-8] – Total de 52485 interações

[8-16] – Total de 52312 interações

[16-23] – Total de 1263 interações

Como podemos verificar pelos dados acima, o período com uma acentuada diferença relativamente ao número de interações corresponde ao terceiro período do dia (16-23), não sendo muito expectável pois trata-se de um horário em que generalidade das pessoas já se encontra fora do seu horário de trabalho.

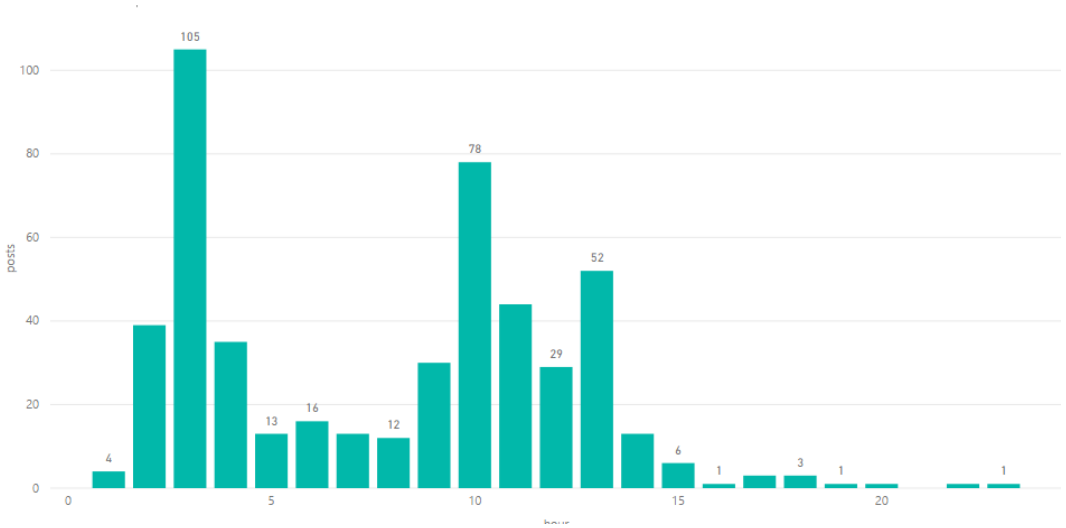


Figura 29: Power BI - *Posts* por Hora

No entanto, verificamos que o período onde foram publicados o maior numero de *posts* corresponde as horas onde a empresa recebe uma maior afluência por parte dos consumidores através de gostos, partilhas e comentários, percebendo assim o porquê da afluência nos vários períodos. Com isto conclui-se que caso a empresa queira ter uma maior afluência, deve realizar o maior número de publicações na hora pretendida pois, como verificado, os consumidores iram corresponder com mais interações.

[0 - 8] - Total de 237 *posts* publicados.

[9 - 16] - Total de 256 *posts* publicados.

[17 - 23] - Total de 10 *posts* publicados.

Média – 4419 Interactions/Hora

Média – 21 *posts*/Hora

Média – 210 Interações/Post

- **Qual o número de interações por dia da semana?**

  Com esta métrica pretende-se saber quais os dias da semana onde existe um maior número de interações, e como isso influência o comportamento dos consumidores.

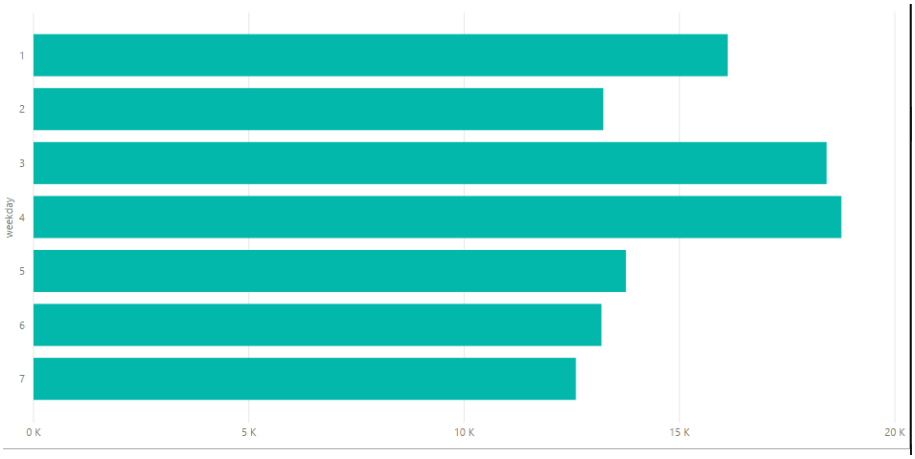


Figura 30: Power BI - Interações por Dia

– O pico máximo de interações ocorre no quarto dia da semana, ou seja, à quarta-feira, com um total de 18758 interações neste dia.
– O segundo maior pico, e não muito longe do primeiro, ocorre na terça-feira com um total 18417 interações.
– Do outro lado da escala encontra-se o dia 7 (sábado) com apenas 12594 interações, registando a afluência mais baixa.

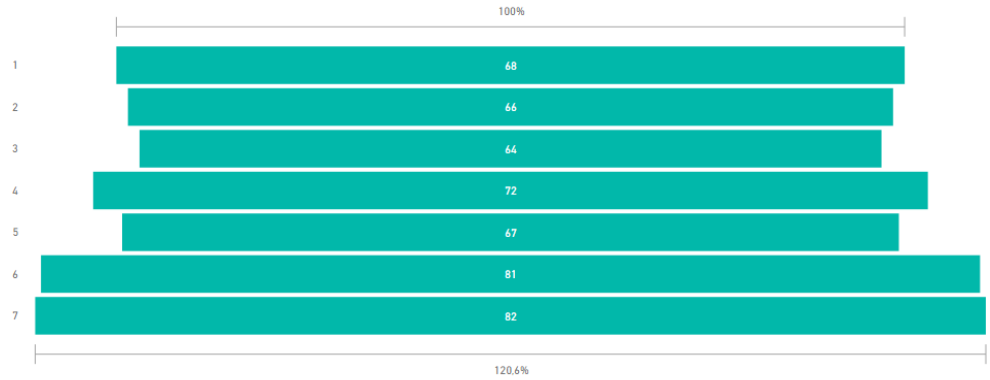


Figura 31: Power BI - *Posts* por dia da semana

Fazendo a comparação com o gráfico dos *posts* publicados por dia da semana, observa-se que não existe uma ligação direta entre os vários valores de publicações nos vários dias e a sua performance, uma vez que os valores são muito idênticos. Conclui-se então que o dia da semana em que os *posts* são publicados não influencia o número de interações dos consumidores, ou seja, a empresa deve continuar a publicar as suas publicações sem pensar concretamente no dia em que pública mas sim em outras métricas mais produtivas e interessantes aqui discutidas.

Média – 15151 Interactions/Dia

Média – 71 *posts*/Dia

- **Qual o número de interações por parte da semana (semana/fim de semana)?**

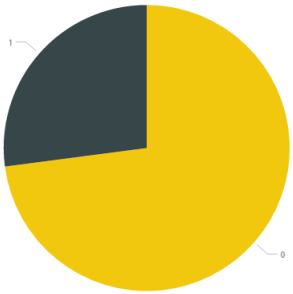


Figura 32: Power BI - Interações por partes das semana

– O numero de interações ao fim-de-semana, isto é, correspondente a dois dias da semana (valor de Weekend == 1), é de 28712.

– Nos cinco dias restantes (durante a semana) o número total de interações nos *posts* é de 77348.
– Corresponde a uma aderência de aproximadamente 73% durante a semana e 27% durante o fim-de-semana por parte dos consumidores.

Figura 33: Power BI - *Posts* por partes da semana

Comparando estes valores com os resultados que correspondem ao total de *posts* publicados durante a semana, é fácil verificar que das 500 publicações, 150 foram publicados no decorrer do fim de semana, e os restantes 350 durante os cinco dias correspondentes à semana. Isto corresponde sensivelmente aos mesmos 30% e 70% de percentagem de aderência entre o fim de semana e semana respetivamente.

É conservada assim a regra máxima que já anteriormente tínhamos concluído, onde quanto mais *posts* são publicados, mais interações se verificam, sobrepondo-se à expetativa natural da variação de interações consoante os períodos de trabalho normais dos utilizadores.

- **Qual o número de interações por mês?**

  Com esta pergunta pretende-se consulta os meses onde existe um maior numero de interações por parte do consumidor, auxiliando assim a empresa a obter melhores resultados na divulgação do seu produto.

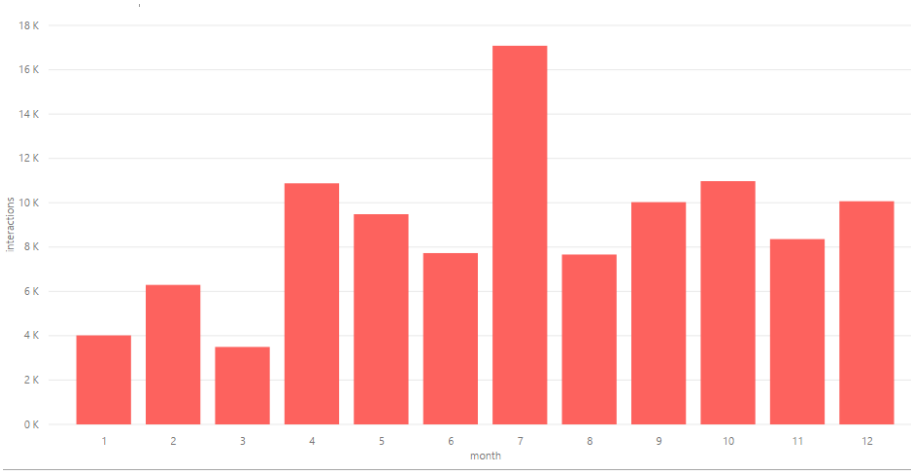


Figura 34: Power BI - Interações por mês

– O mês onde existe maior afluência por parte dos consumidores em termos de interações é o mês 7, Julho, com um total de 17082 interações.
– Por outro lado, o mês com menor número de interações é o mês 3, Março, correspondente a um número de 3494 interações por parte dos consumidores.

Numa primeira observação, e analisando individualmente cada mês, verifica-se a existência de mais *posts* no mês de Julho, tal como acontece com o número de iterações. Nota-se também um número maior de interações nos meses centrais do ano.

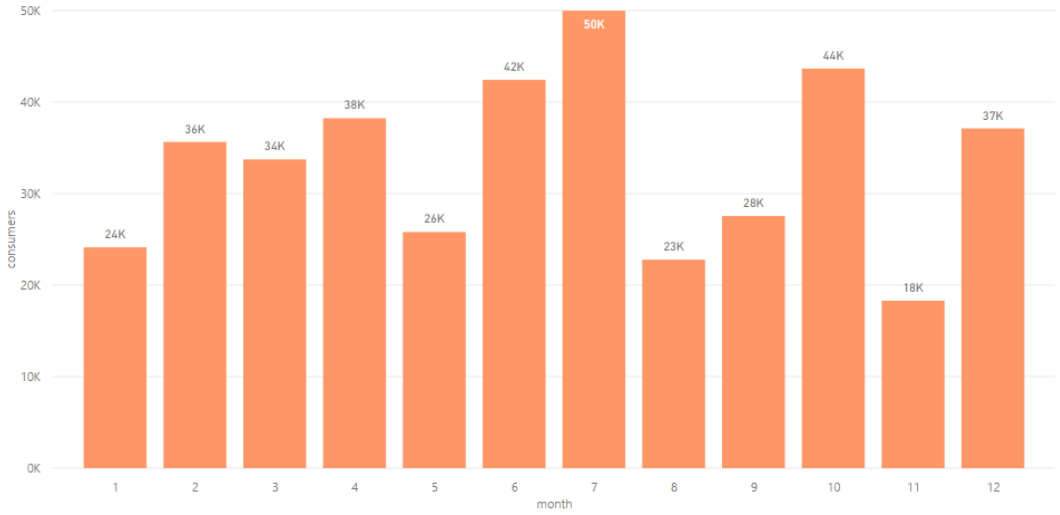


Figura 35: Power BI - *Posts* por mês

Média - 42 *posts*/mês

Média - 8838 interações/mês

- **Qual o número de interações por trimestre?**

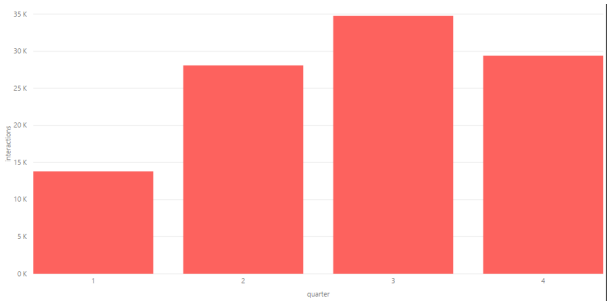


Figura 36: Power BI - Interações por trimestre

  - O trimestre onde se registaram mais consumidores foi o terceiro, ou seja, aquele que abrange os meses de Março, Abril e Maio, com um total de 34771 interações.
  - O trimestre onde se registaram menos consumidores foi o primeiro, onde se encontram os meses de Janeiro, Fevereiro e Março com um total de 13802 interações.

Dividindo o ano pelos vários semestres apresentados:

[Janeiro - Março] - 13802 interações.

[Março - Junho] - 28087 interações.

[Julho - Setembro] - 34771 interações.

[Outubro - Dezembro] - 29400 interações.

Se analisarmos o gráfico de *posts* por mês, observa-se que o terceiro trimestre é aquele onde a empresa fez mais publicações na sua página. Assim, verifica-se uma tendência e expetativa para existirem mais interações na altura mais quente do ano, sendo este um período onde as pessoas se preocupam mais com o seu bem-estar.

- **Qual o número de interações por tipo?**

  Para esta próxima métrica relativo ao numero de interações por tipo de post pretendemos saber se existe algum tipo mais apelativo ao publico de forma que este consiga cativar um maior numero de interações por parte dos consumidores.

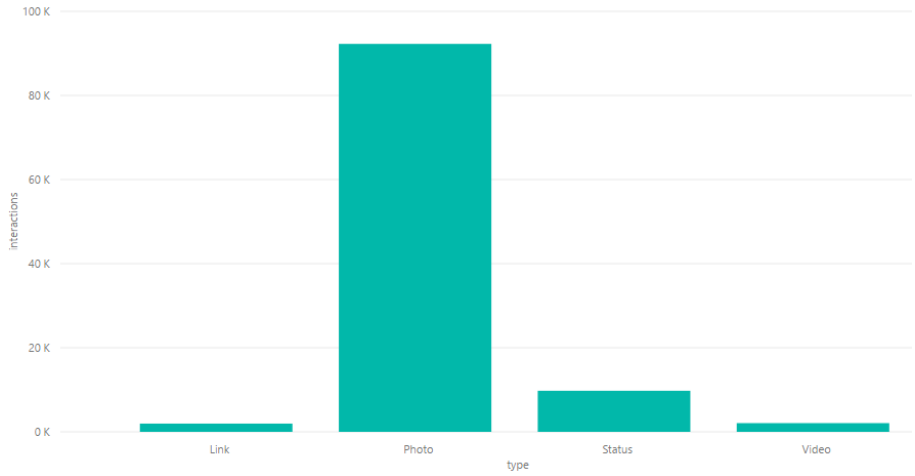


Figura 37: Power BI - Interações por tipo de publicação

  - Analisando o gráfico apresentado em cima, podemos ver que o tipo Photo acarreta o maior número de interações, num total de 92263.
  - Do lado contrário da tabela, com 1959 interações, o tipo Link apresenta um menor número de interações por parte dos consumidores.

De uma forma muito óbvia, verifica-se que o tipo Photo é o que causa maior impacto de likes, comentários e partilhas por parte do público, uma vez que apresenta aproximadamente 90% do número total de interações.

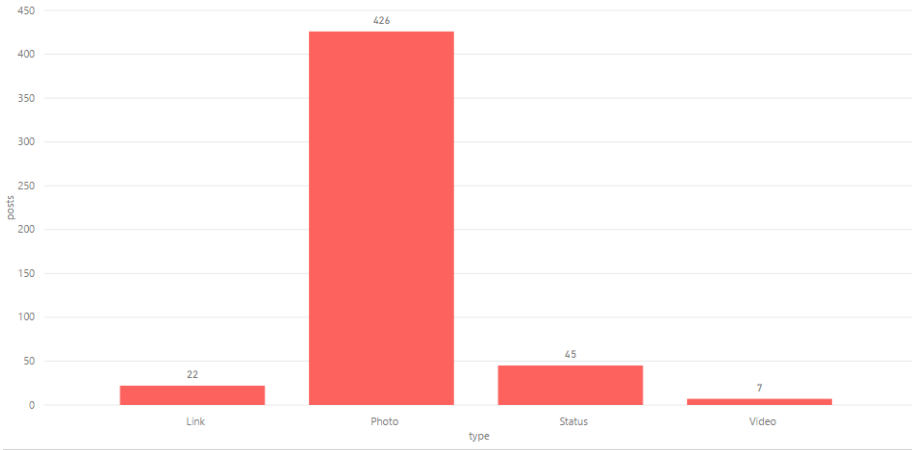


Figura 38: Power BI - *Posts* por tipo

Consultando agora o gráfico com a quantidade de *posts* publicados com os vários tipos, verifica-se a tendência já apresentada nas métricas anteriores de que o tipo de post que apresentar o maior número de publicações, irá receber um maior número de interações. Sendo assim, a empresa deve optar por publicar o maior número possível de publicações utilizando o tipo Photo uma vez que são mais atractivos ao público.

- **Qual o número de interações por categoria?**

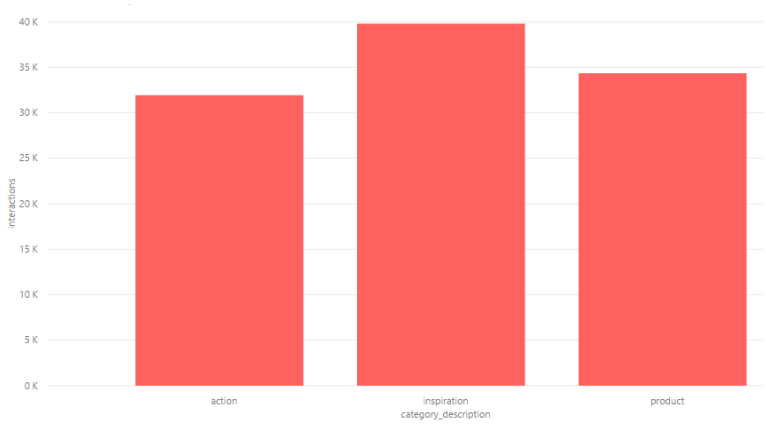


Figura 39: Power BI - Interações por categoria

  - verifica-se uma preferência pelos *posts* de categoria "inspiration", com um total de 39794 interações.
  - em segundo lugar surge a categoria "product"com 34341 interações por parte dos consumidores.
  - por último, mas não menos importante, observa-se a categoria "action"com no total 31925 interações.

Como podemos verificar pelos resultados em cima, existe uma maior procura pelos posts relativos a categoria de "inspiration", concluindo assim que os consumidores gostam mais de ver contéudo motivacional e inspirador. Assim o que a empresa deve tentar encontrar formas de, mesmo não publicando diretamente o conteúdo associado aos seus produtos, conseguir persuadir os seus potenciais clientes através de publicações desta categoria. Esta ideia torna-se ainda mais fundamentada uma vez que as duas categorias mais utilizadas são "inspiration"e "product", ou seja, pode ser interessante uma correlação entre estas.

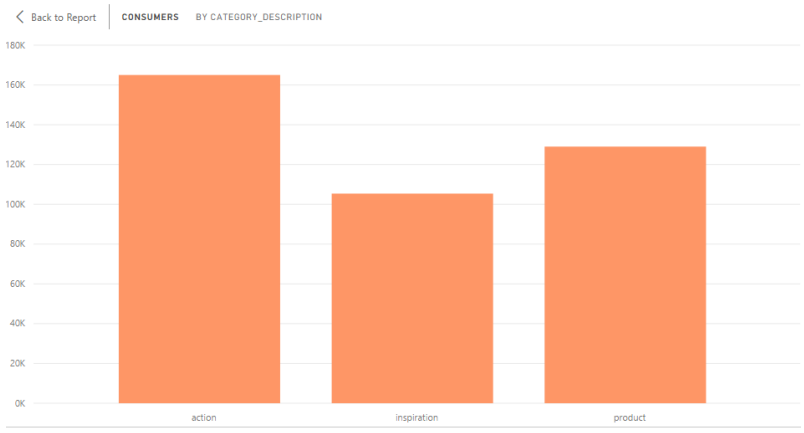


Figura 40: Power BI - *Posts* por Categoria

Analisando agora o número de total posts por categoria, podemos concluir que este não influência diretamente no número de interações por parte dos consumidores, uma vez que a categoria "Action"apresenta o maior número de *posts* publicados, sendo, no entanto, a que apresenta menor número de interações nas suas publicações.

Concluindo, conforme os resultados apresentados por esta métrica, é importante reter a ideia de que a empresa deve apostar nas publicações do tipo "Inspiration", uma vez que apresentam poucos *posts* e muitas interações.

- **Qual o número de interações por tipo e categoria?**

  Com esta métrica pretende-se compreender dentro de que categorias se insere cada um dos tipos de post, e quais aqueles que influenciam mais consumidores.

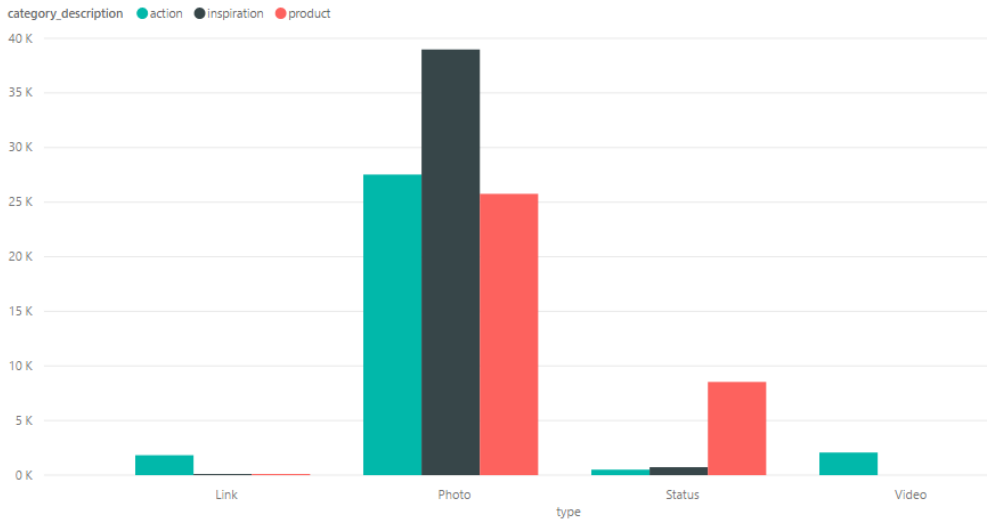


Figura 41: Power BI - Interações por categoria e tipo do post

  - O maior número de interações apresentadas surge para o tipo "Photo"da categoria "Inspiration", com um total de 38979 interações.
  - Já com o menor número de interações, surge o tipo "Video"com a categoria "Inspiration"e "Product".

A partir do gráfico verifica-se que o tipo "Photo", o mais apelativo e chamativo para os consumidores, da categoria "Inspiration", reúne o maior número de interações, com uma grande diferença para os restantes tipos e categorias. Assim, é possível fortalecer a ideia anterior de que a empresa deve apostar em publicações do tipo "Photo"e categoria "Inspiration".

- **Qual o número de interações por publicações pagas?**

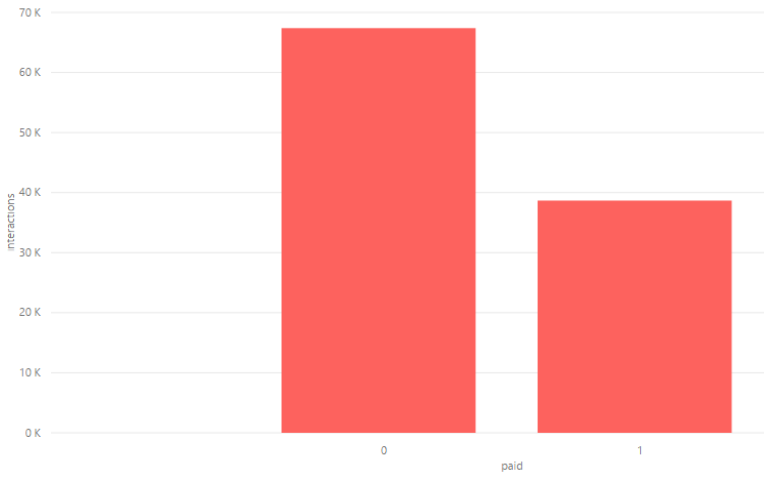


Figura 42: Power BI - Interações por publicações pagas(0) e não pagas(1)

Como seria de esperar, os *posts* não pagos são os mais interagidos por parte dos consumidores, com um total de 67386 interações. Por outro lado, verifica-se também um número interessante relativamente às interações por publicações pagas - 38674.

Figura 43: Power BI - *Posts* por publicações pagas

Como podemos observar, do total de 500 publicações, apenas 139 são pagos enquanto os restantes 361 são gratuitos. Ou seja, apesar dos *posts* pagos representarem 28% do universo global, um terço dos consumidores interage com estes.

- **Qual o número de interações por número de gostos na página?**

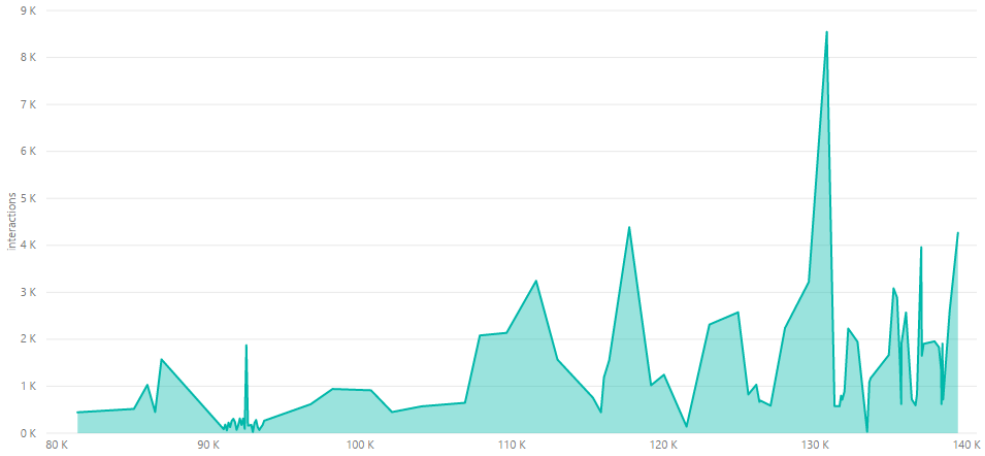


Figura 44: Power BI - Interações por total de *likes* da página

Sem ser necessária uma análise muito detalhada ao gráfico, verifica-se que o número de interações aumenta à medida que a página vai contendo um maior numero de likes, apesar deste aumento não ser tão acentuado como era expectável.

Apesar do maior pico de interações não corresponder ao maior número de likes na página, este poderá ter acontecido por se tratar de uma época festiva ou até de uma promoção da empresa. Este pico máximo ocorreu com 130791 total de likes na página e teve um número de interações de 8545.

- **Qual o número de interações por alcance da publicação?**

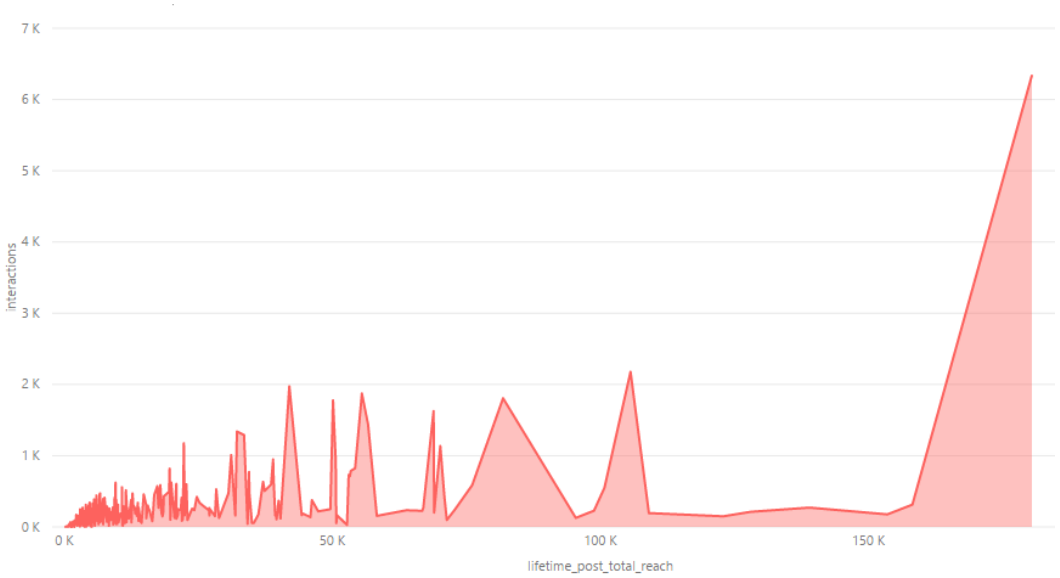


Figura 45: Power BI - Interações por alcance da publicação

Facilmente se conclui que quando o alcance do post é reduzido, os valores de interações desse post são bastante inconstantes, tendo momentos de picos e outros momentos de interações quase nulas. À medida que o alcance da publicação aumenta, os valores já se encontram mais lineares comparando aos apresentados inicialmente.

De uma maneira expectável, o maior pico de interações acontece quando o alcance é máximo, concluindo que os consumidores realizam mais interações a uma determinada publicação quanto mais vezes esta lhes for apresentada.

- **Qual número de interações por impressões da publicação?**

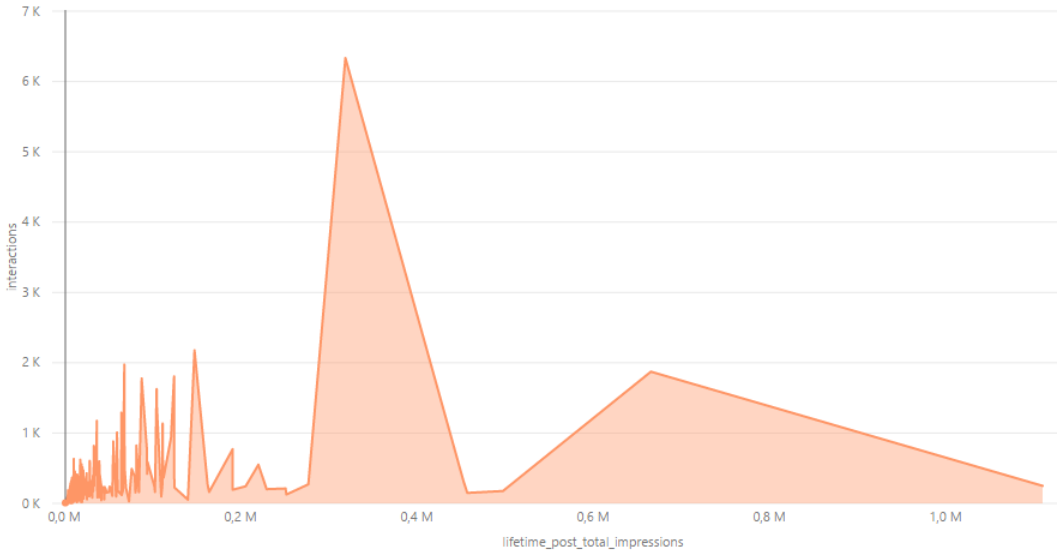


Figura 46: Power BI - Interações por impressão da publicação

De uma forma muito semelhante à métrica anterior, inicialmente o número de interações oscila muito enquanto existe um número reduzido de impressões estabelecidas em relação ao *post*. Ao contrário do que seria expectável, verifica-se que o aumento do número de impressões não influencia diretamente o número de interações dos utilizadores.

6.2 Consumidores

- **Qual o número de consumidores por hora?**

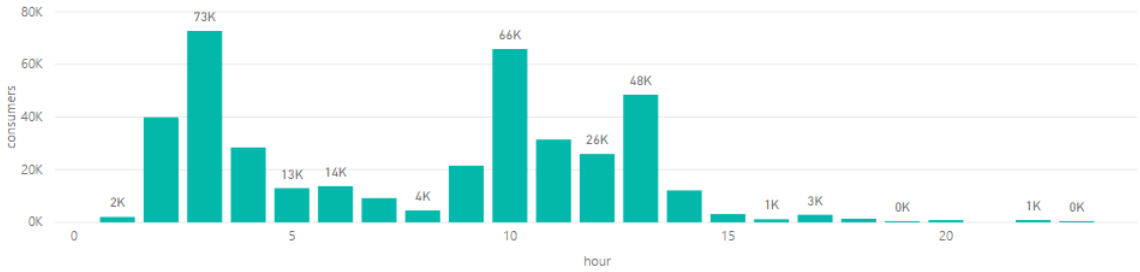


Figura 47: Power BI - Consumidores por Hora

  - O pico de consumidores ocorre no valor de hora 3, ou seja, corresponde a uma afluência máximo às 3 horas da manhã. Neste período, um total de 72685 consumidores clicaram em *posts* publicados por esta empresa de cosméticos.
  - O segundo pico máximo de consumidores ocorreu no valor de hora 10, ou seja às 10 horas da manhã. Neste período, um total de 65820 consumidores clicaram em publicações.
  - O período de menor afluência registou-se no valor de hora 21 e 0, ou seja, às nove da noite e à meia noite, onde não se verificam registos de clicks em *posts*.

Essencialmente o que é notório após uma análise mais generalizada do gráfico de barras, é a presença de períodos de grande afluência nos primeiros dois terços do dia, quando comparados com a pouca afluência que se regista no último terço. Estes dois primeiros terços, correspondem

aos intervalos de horas compreendidos entre a meia noite e as oito da manhã, para o primeiro terço, e as nove da manhã e quatro da tarde para o segundo. Em ambos é possível verificar um arranque lento, no número de consumidores, seguido de um pico que se pode vir a estender por três horas, e de um eventual abrandamento.

[0 - 8] - Total de 183385 consumidores.

[9 - 16] - Total de 209582 consumidores.

[17 - 23] - Total de 6419 consumidores.

O segundo terço acaba por ser aquele onde se verificam mais consumidores, o que não deixa de ser curioso, por se tratar do período onde estão compreendidas maior parte daquelas que são as horas de trabalho para a maioria das pessoas.

No entanto, se cruzarmos informações, no que diz respeito ao número de *posts* publicados pela empresa, conseguimos perceber melhor os motivos para uma maior afluência nas horas acima descritas.

Tal com se pode perceber pela figura 29, as horas onde a empresa publica a maioria dos seus *posts*, correspondem precisamente aquelas onde a afluência é maior por parte dos consumidores.

Com isto retira-se a conclusão que, caso queiram ter mais afluência a uma determinada hora, a empresa deve dedicar-se a colocar o máximo de *posts* possível nessa mesma hora, pois os consumidores vão corresponder com os clicks.

[0 - 8] - Total de 237 *posts* publicados.

[9 - 16] - Total de 256 *posts* publicados.

[17 - 23] - Total de 10 *posts* publicados.

Média - 16641 consumidores/hora

Média - 21 *posts*/hora

- **Qual o número de consumidores por dia da semana?**

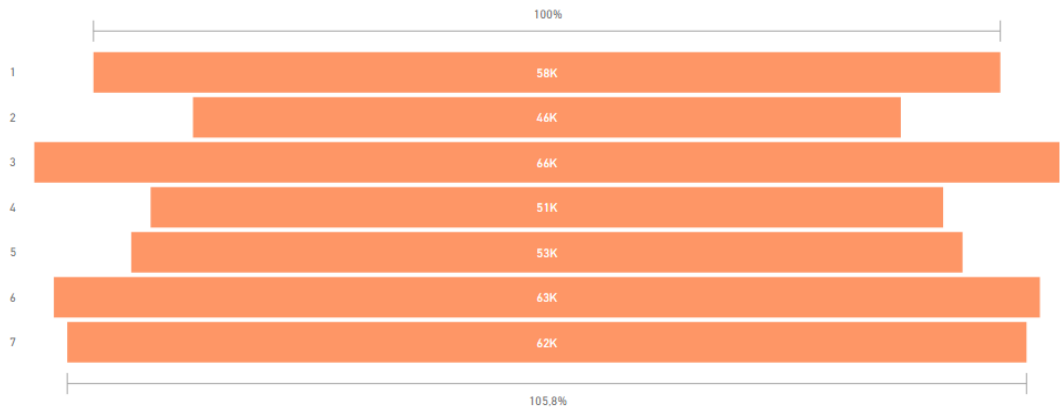


Figura 48: Power BI - Consumidores por dia da semana

– O pico de consumidores ocorre no terceiro dia da semana, ou seja, à terça-feira, com um total de 65941 consumidores neste dia.

– O segundo maior pico ocorre no sexto dia da semana, ou seja, à sexta-feira, com um total de 63427 consumidores neste dia.

– O dia da semana com menos afluência trata-se da segunda-feira, onde apenas 45530 consumidores foram registados.

Neste caso, quando fazemos a comparação com o gráfico de publicação de *posts* por dia da semana (figura 31) é curioso verificar que no dia onde são registados mais consumidores, trata-se do dia onde a empresa acaba por publicar menos *posts*, com apenas 64.

O dia onde a empresa se dedica a publicar mais *posts*, ao sábado, com um total de 82, acaba por ser um dia também ele acima da média de em termos de consumidores desses mesmos *posts*, com um total de 61704 consumidores.

Com isto, é possível concluir que se quiserem ter mais afluência num determinado dia, a empresa deve dedicar-se a colocar o máximo de publicações possíveis mesmo dia, pois os consumidores vão corresponder com os clicks.

Média - 71 *posts*/dia

Média - 57055 consumidores/dia

- **Qual o número de consumidores por parte da semana (semana/fim da semana)**?

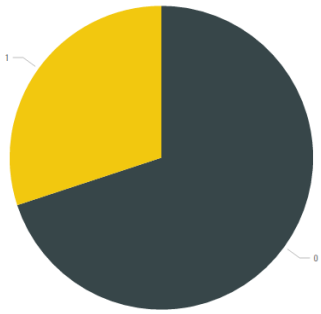


Figura 49: Power BI - Consumidores por parte da semana

– O número de consumidores ao fim de semana, i.e., em que o valor de Weekend é 1, é de 120038, ao passo que durante a semana o número total de consumidores de *posts* é 279348. Isto corresponde a sensivelmente 30% de adesão durante o fim de semana e 70% da adesão total durante a semana.

Quando comparámos estes valores com os valores que correspondem ao total de *posts* publicados por parte da semana (figura 31), é possível verificar que dos 500 *posts*, 150 foram publicados no fim de semana, e os restantes 350 à semana. Isto corresponde exatamente ao mesmos 30% e 70% de divisão total entre fim de semana e semana respetivamente.

É mantida assim a tendência que já se tinha verificado na análise de consumidores por hora, onde quanto mais *posts* são publicados mais consumidores se verificam.

Esta tendência volta sobrepor-se à tendência mais natural que seria a de visitar mais estas publicações fora dos dias e horas de trabalho.

- **Qual o número de consumidores por mês?**

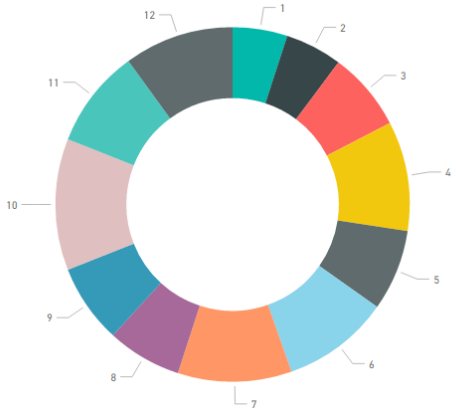


Figura 50: Power BI - Consumidores por mês

  - O mês onde mais consumidores se registam acaba por ser o mês 7, ou seja, Julho, com um total de 49990 consumidores.
  - O mês com menos afluência por parte dos consumidores acaba por ser o mês 11, i.e., o mês de Novembro, com um total de 18296 consumidores.

Quando analisámos o gráfico de uma forma mais objetiva, podemos verificar uma maior afluência nos primeiros meses do ano, quando comparada com aquela verificado no final do mesmo.

Por outro lado, verifica-se a tendência natural de uma maior afluência a este tipo de produto (cosméticos), nos meses mais quentes do ano.

No entanto volta-se a verificar a tendência de uma maior afluência quanto maior for o fluxo de *posts* publicados nesse período (figura 31). Analisando os meses de forma individual, o mês com mais publicações, Outubro, não corresponde no entanto com o mês com consumidores, Julho.

Média - 42 *posts*/mês

Média - 33282 consumidores/mês

- **Qual o número de consumidores por trimestre?**

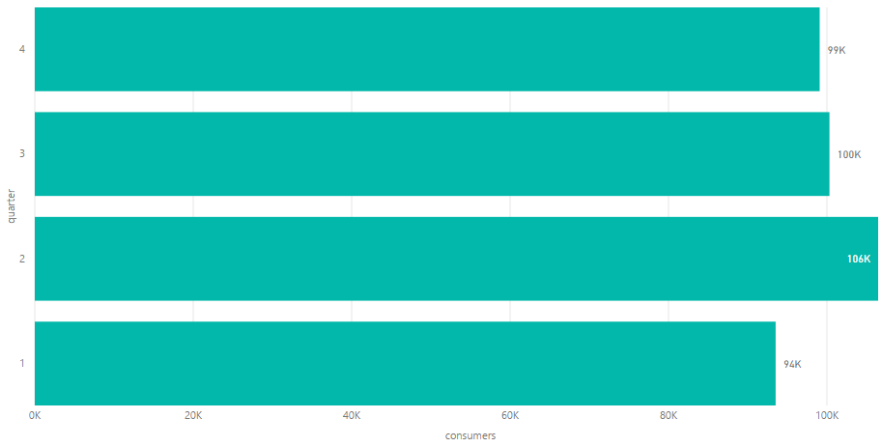


Figura 51: Power BI - Consumidores por trimestre

– O trimestre onde se registaram mais consumidores foi o terceiro, ou seja aquele que engloba o conjunto do meses de Março, Abril e Maio, com um total de 106462 consumidores.

– O trimestre onde se registaram menos consumidores foi o primeiro, onde se encontram os meses de Janeiro, Fevereiro e Março com um total de 93524 consumidores.

[Janeiro - Março] - 93524 consumidores.

[Março - Junho] - 106462 consumidores.

[Julho - Setembro] - 100318 consumidores.

[Outubro - Dezembro] - 99082 consumidores.

Se analisarmos o gráfico de publicações por mês (figura 35) podemos verificar que o primeiro trimestre é aquele onde a empresa de cosméticos menos *posts* publicou na sua página de Facebook, o que justifica a tendência para se verificarem menos consumidores.

- **Qual o número de consumidores por tipo?**

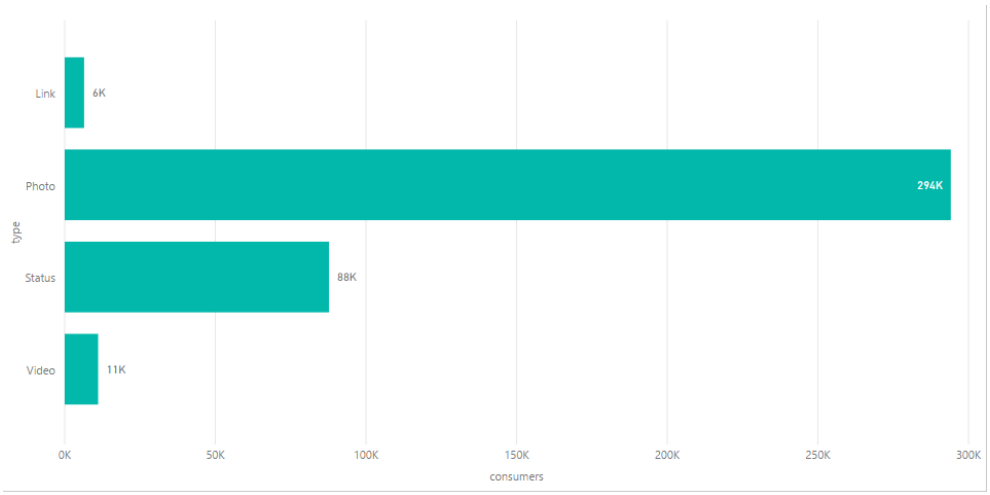


Figura 52: Power BI - Consumidores por tipo

– A tendência que se verifica após a análise do gráfico de consumidores por tipo, é uma afluência maior aos *posts* do tipo Photo, com um total de 294124 consumidores.

– De seguida os *posts* do tipo Status registam 87730 dos clicks.

– Em terceiro lugar, os *posts* do tipo Video, tiveram um total de 11093 consumidores, e por último os *posts* do tipo Link, verificaram 6439 consumidores.

Os *posts* do tipo Photo, são sem dúvida os que mais interessam ao público, parte por serem bastante apelativos (sobretudo neste tipo de negócio), mas também porque seguem a regra que tem vindo a verificar ao longo desta análise.

Se olharmos para o gráfico que mostra a quantidade de *posts* publicados por tipo, constatamos que das 500 publicações:

426 são do tipo Photo.

45 são do tipo Status.

22 são do tipo Link.

7 são do tipo Video.

Verficando a tendência de que a categoria que apresentar a maioria dos *posts* publicados irá ser aquela que irá receber mais afluência por parte das pessoas.

A única exceção desta regra é mesmo a existência de mais *posts* do tipo Link do que do tipo Video, e ainda assim os do tipo Video levam vantagem.

A única explicação passa pelo facto dos Links serem menos apelativos esteticamente do que os Videos e implicarem uma consulta de outra página, o que por vezes torna-se inútil ou demasiado trabalhoso para os consumidores.

As Photos devem ser o ramo escolhido pela empresa, visto que são aquelas que melhor cativam os consumidores.

- **Qual o número de consumidores por categoria?**

Figura 53: Power BI - Consumidores por Categoria

  - A tendência que se verifica é uma preferência pelos *posts*, de categoria Action, com um total de 165006 consumidores. Em segundo lugar, fica a categoria Product com 129005 consumidores e por último, os *posts* que se encaixam em Inspiration com 105375 consumidores.

Não sendo estas diferenças muito significativas, volta a verificar-se a regra de que as categorias de *posts* mais publicadas, são aquelas que recebem mais atenção e, portanto, mais clicks por parte dos consumidores.

A análise do gráfico com o número total de *posts* por categoria mostra essa tendência. A categoria de *posts* que verificou um maior número foi a de Action, seguida de Inspiration e por último Product.

Action - 215 *posts*

Inspiration - 155 *posts*

Product - 130 *posts*

Se a intenção da empresa de cosméticos é angariar o máximo de consumidores possível fica claro que as suas preferências estão em *posts* da categoria Action.

- **Qual o número de consumidores por tipo e categoria?**

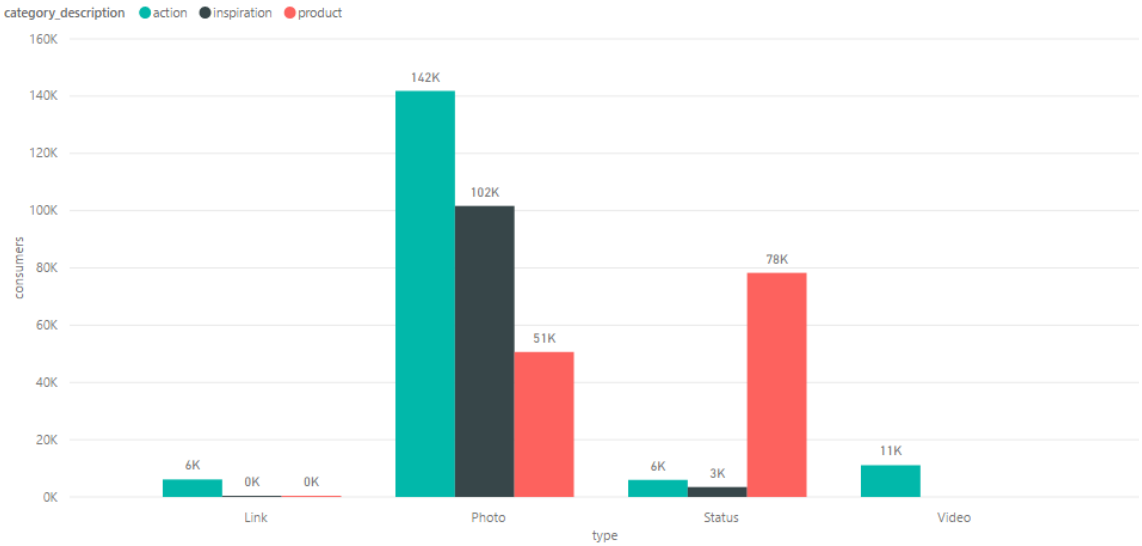


Figura 54: Power BI - Consumidores por Tipo e Categoria

Analisando o gráfico para o efeito, já tínhamos percebido que o tipo Photo era aquele que mais consumidores reunia. No entanto, desta feita, ficamos a perceber dentro de que categorias se inserem cada um dos tipos de post, e quais são aqueles com mais consumidores.

Dentro do tipo Photo, a categoria com mais consumidores de *posts* é a Action, seguida da Inspiration e por último Product.

Photo:

Action - 141819

Inspiration - 101651

Product - 50654

Dentro do tipo Status, a categoria com mais consumidores de *posts* é a Product, seguida da Action e Inspiration.

Status:

Action- 5971

Inspiration- 3479

Product- 78280

Dentro do tipo Video, a categoria com mais consumidores de *posts* é a Action, sendo que não existem consumidores de publicações para as categorias Inspiration ou Product.

Video:

Action-11093

Inspiration- 0

Product- 0

Dentro do tipo Link, a categoria com mais consumidores de *posts* é a Action, seguida de Inspiration e Product.

Link: Action- 6123

Inspiration- 245

Product- 71

Os valores mais diversificados e com mais consumidores, verificam-se em Photo, pois qualquer que seja a categoria do post, fica mais uma prova de que para os consumidores uma simples foto é a maneira mais simples de passar uma mensagem.

A discrepância de valores em Status, em que os consumidores, surgiram quando os *posts* deste tipo eram da categoria product, explica-se pelo facto deste tipo ser usado como modo de possível estratégia de marketing para a venda dos seus produtos.

A existência de consumidores nos *posts* de Video apenas quando a categoria é Action, mostra que a tentativa de publicitar um produto, por exemplo, através de um video falha redondamente, e que o video é a melhor maneira de capturar ação.

A existência de maioritariamente consumidores quando o post Link faz parte da categoria Action, explica-se quando raciocinamos pensando no que foi dito anteriormente. Ora, se os Videos foram melhor aceites para categorias de Action, e dentro dos Links a categoria mais aceite foi também Action, significa que os *posts* Links serão sobretudo Links para vídeos externos em plataformas como o Youtube, sendo que qualquer outro redireccionamento, não interessa aos utilizadores, nem sequer para publicitar um produto, possivelmente na página da própria empresa de cosméticos.

- **Qual o número de consumidores por publicações pagas?**

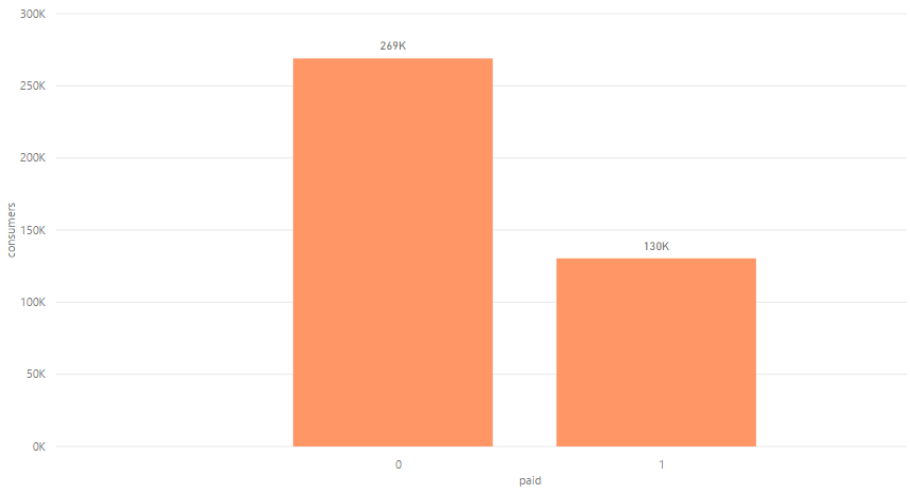


Figura 55: Power BI - Consumidores por publicações pagas

  - A tendência que se verifica é uma preferência pelos *posts*, de categoria Action, com um total de 165006 consumidores. Em segundo lugar, fica a categoria Product com 129005 consumidores e por último, os *posts* que se encaixam em Inspiration com 105375 consumidores.

Os *posts* não pagos, são naturalmente os mais consumidos, com 268999, a consumirem este tipo de *posts*, e apenas 130387 a consumirem os pagos, uma relação de 67% para 33% aproximadamente. Dentro quer dos *posts* gratuitos, quer dos *posts* pagos, a relação de quais são os mais consumidos é equivalente à geral, onde ambos estão incluídos, quer tipo mais consumido, quer em categoria mais consumida.

Dois terços dos consumidores clicam em publicações não pagas enquanto um terço, consome conteúdos pagos, apesar da relação de *posts* publicados pagos e não pagos ser ligeiramente superior.

- **Qual o número de consumidores por número de gostos na página?**

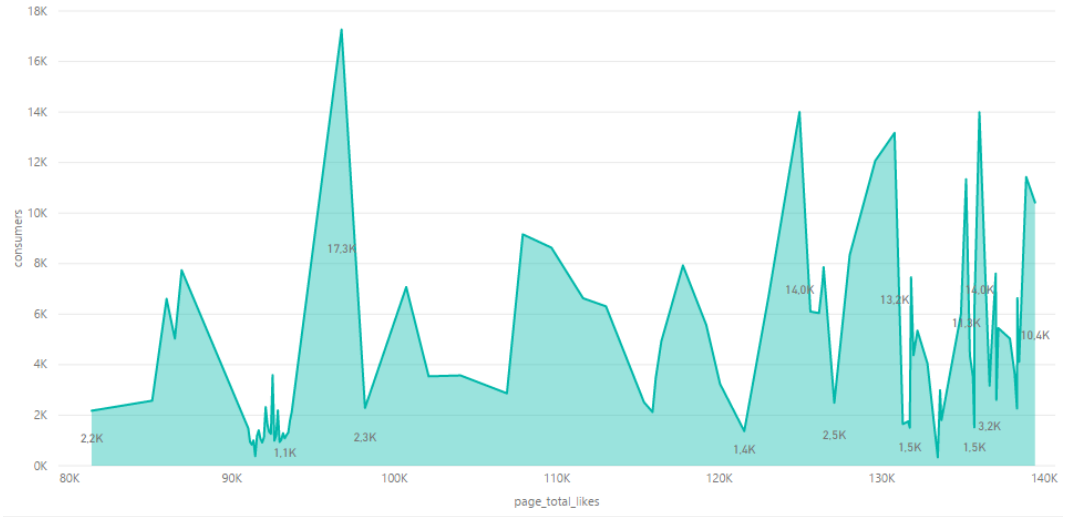


Figura 56: Power BI - Consumidores por número de gostos na página

  - Um dos factos mais interessantes de se realçar é que o facto do número de likes da página aumentar (o que à partida significaria que a página se tornaria mais conhecido e conseguiria angariar mais consumidores), não significa que o número máximo de consumidores aumente, nem sequer que esse número se mantenha constante.
  - O pico máximo de 17236 consumidores ocorreu quando o número total de likes na página era de 96749.

No entanto o gráfico pode ser interpretado de uma outra forma. A criação de consumidores aumenta naturalmente o número de likes da página. Os picos de consumidores que se verificam no gráfico sobretudo quando a página estava a atingir o número máximo de likes (aproximadamente 140000 likes), podem significar que a página está a receber likes mais frequentemente, não permitindo assim juntar um número de consumidores tão grande, enquanto que o intervalo de likes não varia, gerando assim valores menores ao que anteriormente tinha sido registado.

É possível retirar do gráfico que após ocorrer um pico de consumidores de seguida os valores descem significativamente, pois os likes devem ser mais rápidos de alcançar do que anteriormente, impedindo a acumulação de consumidores num dado intervalo.

- **Qual o número de consumidores por alcance da publicação?**

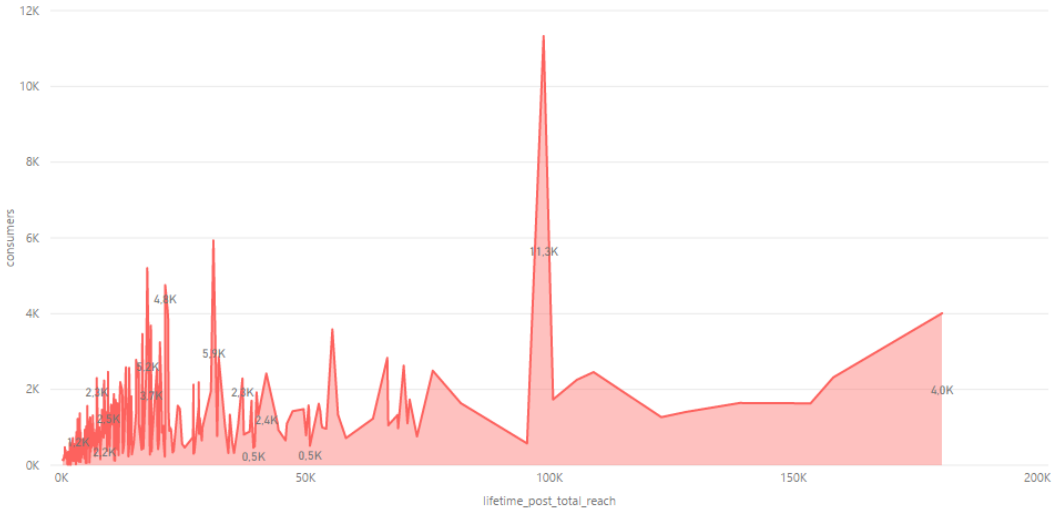


Figura 57: Power BI - Consumidores por alcance da publicação

As conclusões expectáveis surgem após a analise deste gráfico. Quando o alcance do post é ainda reduzido os valores dos consumidores desse post, são pouco constantes e reduzidos quando comparados com os que se verificam após o alcance do post, ter aumentado.

Existem um ponto, quando o alcance do post era de 98816 em que foi atingido o pico máximo de consumidores, com um total de 11328. Após esse pico, o número de consumidores mantem-se mais constante do que inicialmente, mas reduz ainda assim drasticamente, pois apesar de ser algo estabelecido e largamente partilhado, chegando assim a muitos possíveis consumidores, maior parte deles quando se chega a este ponto já conhece o post em questão, perdendo assim o interesse em voltar a clicar no mesmo.

Em média os *posts* chegam ao pico máximo de consumidores a metade do valor máximo de alcance.

- **Qual o número de consumidores por impressões da publicação?**

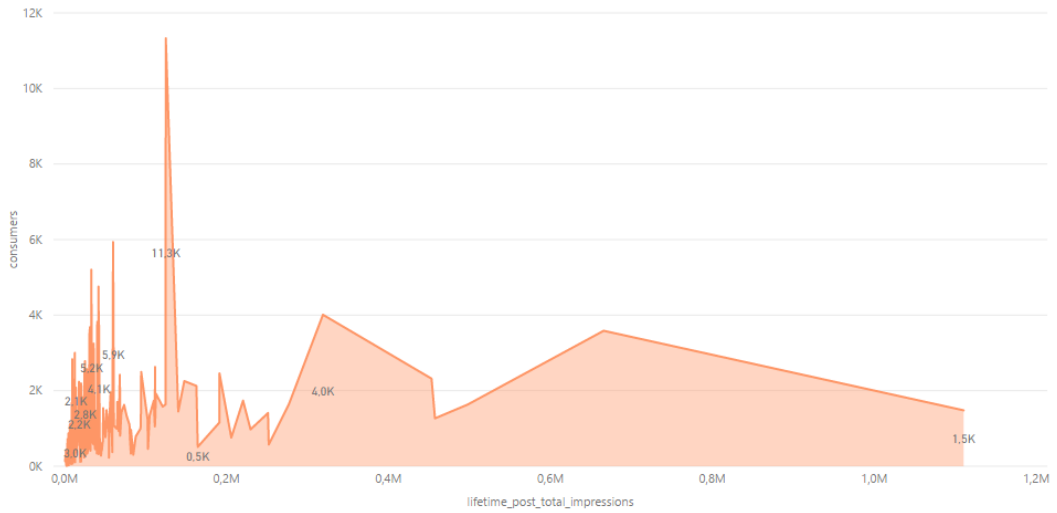


Figura 58: Power BI - Consumidores por impressões da publicação

De uma maneira semelhante aquela explicada em alcance, quando não existem impressões estabelecidas em relação ao post, o mesmo apresenta picos e valores de consumidores inconstantes. Após algum tempo, neste caso quando as impressões se situam no valor de 125026, os consumidores do post estão no seu nível mais elevado, com um total de 11328. Após algum tempo, os consumidores já conhecem o post, diminuindo assim o número total de consumidores do mesmo.

Em média os *posts* chegam ao pico máximo de consumidores antes de chegarem a metade do seu valor máximo de impressões.

7 Conclusões

Terminada a realização do projeto, é então possível concluir que todo o processo desde o momento inicial foi crucial para o obtenção de resultados concretos e de extremo interesse para a empresa.

Em primeiro lugar, foi feita uma análise detalhada ao conjunto de dados, interpretando todos os seus atributos e respetiva relevância. Deste modo, foi possível verificar qual o tipo de cada atributo e como estes estavam relacionados entre si.

Seguidamente, após todo este trabalho de análise do *dataset*, procedeu-se à elaboração do modelo conceptual da base de dados relacional, fundamentado através de tabelas que permitem compreender todas as entidades e relacionamentos do esquema. Após verificação do esquema, este foi implementado recorrendo ao *MySQL Workbench* e povoado com dados provenientes de um ficheiro *csv* através da ferramenta *Talend Open Studio*.

Posteriormente, já com a base de dados preenchida corretamente, foi necessário desenvolver todo o processo de criação do data warehouse. Neste sentido, foi elaborada uma seleção de dados que levou à identificação dos datamarts necessários para o modelo. As suas dimensões e factos foram também eficazmente fundamentadas, levando à obtenção de um modelo dimensional com dimensões partilhadas. Em seguida, o data warehouse foi preenchido recorrendo aos dados provenientes da base de dados relacional e também do ficheiro *csv* correspondente ao segundo semestre do ano. Todo este povoamento foi realizado recorrendo ao *software Pentaho Kettle*.

Por último, através da utilização do *Microsoft Power BI*, foram criados *reports* com o objetivo de responder às questões selecionadas no capítulo 4. Assim, foi possível retirar algumas conclusões interessante do ponto de vista do gestor de negócio, por exemplo, concluiu-se que caso a empresa queira aumentar o número de interações, deve apostar em publicações da categoria "inspiration"e do tipo "photo", uma vez que estes causam um maior impacto no utilizador. Por outro lado, foi possível verificar que o dia da semana não influencia as interações e consumidores das publicações, no entanto observou-se um maior impacto nos meses mais quentes do ano, onde os utilizadores recorrem mais a este tipo de produtos.

Em suma, o grupo ficou bastante satisfeito com o trabalho desenvolvido, uma vez que elaborou e compreendeu as principais etapas da implementação de um sistema de apoio à decisão aplicando métodos de Business Intelligence.

Referências

[1] https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

[2] Moro, S., Rita, P., Vala, B. (2016). *Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach.* Journal of Business Research, 69(9), 3341-3351.

[3] Kimball, Ralph, and Margy Ross. *The data warehouse toolkit: The definitive guide to dimensional modeling.* John Wiley Sons, 2013.