Prospecção de Dados: Parte 4

Unidade curricular
Tecnologias de Processamento de Dados - 2018/2019

Docente
António Manuel Silva Ferreira

Grupo 5

Catarina Vaz	Nº 37613
Fábio Neves	Nº 44949
João Campagnolo	Nº 46731
Jorge Cruz	Nº 53506
Susana Sousa	Nº 21989

Mestrado em Bioinformática e Biologia Computacional

Faculdade de Ciências da Universidade de Lisboa

Índice

Nota:
Os pontos 6 e 7 da 3ª fase foram integralmente refeitos, pelo que apenas estão sinalizados (com highlight verde) no índice.
As restantes alterações usam o mesmo highlight no corpo do relatório em si.

1. Introdução

a. Cofina

A cofina é uma das principais empresas de media, detendo um portfólio composto por jornais, revistas e um canal de televisão. Líder no mercado da imprensa, o seu foco assenta valorização dos seus clientes aos olhos do mercado bolsista através do estudo e resposta às necessidades do mercado contemporâneo nacional e internacional (http://www.cofina.pt).

b. Liga Record

Precedendo as plataformas digitais, a *Liga Record* (https://liga.record.pt) é um jogo de estratégia promovido pelo jornal Record e gerido pelas equipas da Cofina. Cada equipa tem um custo nominal de 3€ (excepto casos excepcionais como ofertas). Cada concorrente poderá formar uma ou mais equipas fictícias de 23 jogadores reais cada a jogarem no campeonato português da 1ª Liga, não ultrapassando um valor de 40 milhões de euros (variável mediante eventuais bónus/penalizações atribuídos no decorrer do jogo). Cada jogador tem um valor inicial atribuído pela equipa do jornal Record. Mediante o desempenho de cada jogador em cada encontro, quer quantitativo (golos; penaltis; participação), quer qualitativo (extra atribuido por um profissional da redação do Record), a equipa obterá a pontuação combinada dos jogadores escolhidos pelo concorrente para entrar em campo. As pontuações são divulgadas pelo jornal Record no fim de cada ronda (no dia seguinte ao último jogo da jornada). Em sequência do seu desempenho, os jogadores vêem o seu valor base alterado a cada ronda, ficando mais caros quando pontuam positivamente e mais baratos quando pontuam negativamente, sendo o valor mínimo de 500.000€. Os concorrentes com as melhores pontuações serão premiados no fim de cada ronda, no fim da 1ª fase (campeonato de Inverno) e na ronda final do jogo.

O jogo inicia-se anualmente após o fecho do mercado de Verão de jogadores, tipicamente no final de agosto / início de setembro, o que corresponde à 4ª ou 5ª jornada do campeonato.

Para além da compra de equipas, os concorrentes também podem fazer subscrição do serviço premium para os ajudar a escolher os jogadores para as suas formações. Deste serviço constam: dados estatísticos dos jogadores no campeonato; informação a cada ronda sobre castigos e lesões; artigos de opinião com dicas de especialistas quanto à formação de equipas.

Informações adicionais podem ser consultadas no website original da Liga Record[2].

c. Dados usados da Liga Record

Apesar de termos acesso a dados relativos a 8 anos de Liga Record, tendo em conta a ampla participação neste jogo (chegou a incluir num só ano 65 mil equipas e 30 mil concorrentes) e como tal, uma grande quantidade de informação disponível, no âmbito deste projecto recolhemos uma amostra dos concorrentes mais participativos nos últimos 5 anos, ou seja, concorrentes que se inscreveram todos os anos, escolheram um nickname para se auto-designarem e que neste último ano (pelo menos) participam em ligas privadas (mini-campeonatos administrados por um dos concorrentes que algumas equipas formam entre si).

Destes concorrentes, apenas estudámos a evolução no jogo das equipas participantes em Ligas privadas. Ainda relativo a estes concorrentes, foram recolhidos dados relativos ao número de logins (que, neste caso equivalem a visitas) efectuados durante o período efectivo de jogo que se inicia após o fecho do mercado de Verão de jogadores (normalmente no final de Agosto) e o fim do campeonato nacional da 1ª Liga (final de Maio).

Apesar de participarem do jogo todos os anos, também foram eliminados concorrentes que apenas tenham participado com equipas de teste.

2. Fontes de dados

Os dados têm origem em cinco fontes diferentes, nomeadamente:

- Cofina – onde estão hospedados os dados pessoais de cada concorrente;
- Sport Radar – empresa que disponibiliza os dados dos jogadores reais e dos jogos de cada jornada;
- Jornal Record – responsável para validação dos jogos e atribuição de valores a cada jogador e respectiva pontuação qualitativa em cada jogo;
- Liga Record – dados inerentes ao jogo em si (formação das equipas, visitas ao site).
- CTT - Empresa que disponibiliza a lista de localidades e códigos postais associados, que vão ser usados na normalização das moradas fornecidas pelos concorrentes.

Os dados sobre os quais vamos trabalhar são já uma compilação dos dados consultados na estrutura original.

Dada a complexidade e quantidade de informação disponível nos dados originais, fizemos esta opção com vista à simplificação dos dados originais de forma a permitir um estudo pertinente no âmbito da cadeira de Tecnologias de Processamento de Dados.

a. Tabelas descritivas dos dados

Os dados presentes nas tabelas originais relevantes para a obtenção das nossas tabelas de trabalho estruturam-se por cada temporada da seguinte forma:

i. Tabelas inerentes ao jogo:

- **User**

Coluna	**Tipo**	**Observações**
[id]	int	**Chave:** Identificador único de utilizador atribuído automaticamente
[name]	string	Nome do utilizador *(ex: Susana Sousa)*
[email]	string	Email do utilizador *(ex: xxx@gmail.com)*

[nickname]	string	Nome do utilizador alternativo exibido no jogo *(ex: Maior das Ilhas)*
[birthdate]	datetime	Data de nascimento do utilizador *(ex: 1976-01-01)*
[address]	datetime	Morada do utilizador *(ex: Rua dos Pesadores , 19 - R/C, 3060-691 Tocha – Portugal)*
[phone]	datetime	Número de telefone *(ex: 914139015)*
[id_gender]	int	Chave estrangeira: identificador do género do utilizador (concorrente)
[id_club]	int	Chave estrangeira: Identificador do clube escolhido pelo utilizador a partir de uma lista de possibilidades
[id_region]	int	Chave estrangeira: Identificador da região escolhida pelo utilizador a partir de uma lista de possibilidades. Valor apenas indicativo – não obrigatoriamente relacionado com a morada.
[date_start]	datetime	Data de inscrição em determinada temporada *(ex: 2018-07-27 19:17:43)*

- **Team**

Coluna	**Tipo**	**Observações**
[id]	int	**Chave:** Identificador único da equipa atribuído automaticamente
[id_user]	int	Chave estrangeira: Identificador de utilizador
[name]	string	Nome da equipa atribuído pelo utilizador *(ex: Perdiz vermelha)*
[code]	string	Código alfanumérico de 9 caracteres comprado ou resultante de uma oferta que permite fazer o registo da equipa *(ex: 2FC3HJHN2)*
[id_origin]	int	Chave estrangeira: Identificador da origem do código da equipa (ex: revista; paypal; agências; cofina; reclamações)
[lastupdate]	datetime	data da última alteração efectuada nos dados da equipa *(ex: 2019-03-24 19:20:50)*

- **ng_team_origin**

Coluna	Tipo	Observações
[id]	int	**Chave:** Identificador único de origem da equipa.
[name]	string	Nome da origem da equipa. Valores possíveis indicados à frente.
[is_paid]	bool	Indica se a equipa é paga (3€) ou grátis

As hipóteses de origem de equipa existentes são:

- CÓDIGO OFERTA AMIGO (grátis) – Um concorrente compra um código que oferece a outro dentro do jogo
- AGÊNCIAS (grátis) - Oferta publicitária
- RECLAMAÇÃO PAYPAL – Código pago com paypal mas houve problema na transacção
- OFERTA 3+1 (grátis) – Por cada 3 equipas compradas, o concorrente tem direito a mais uma
- REVISTA – Compra da revista Record lançada em Julho inclui um código Liga Record
- MB – Código pago com multibanco.
- CC – Código pago com cartão de crédito
- RECLAMAÇÃO – Código fornecido ao concorrente mediante uma reclamação feita pelo concorrente relativa a uma compra que não foi validada
- OFERTAS "VÁRIOS" (grátis) – Códigos oferecidos pela Cofina
- COMERCIAL (grátis) – Códigos atribuídos pelo departamento Comercial da Cofina
- PAYPAL – Códigos pagos com Paypal

As fontes mais usadas são:

- REVISTA – com 55% das equipas
- MB – com 25% das equipas.
- OFERTA 3+1 – com 11% das equipas

- league

Coluna	Tipo	Exemplo	Observações
[id]	int		Chave: Identificador único sequencial da liga privada atribuído automaticamente
[name]	int	*"Liga dos amigos do*	Nome da liga privada

		Cordeiro”	
[id_owner]	int		Chave estrangeira: Identificador do concorrente que controla quem participa na Liga Privada

- **team_league_private**

Coluna	**Tipo**	**Exemplo**	**Observações**
[id_league]	int		Chave: Identificador da liga privada
[id_team]	int		Chave: Identificador da equipa
[lastupdate]	datetime	*2018-07-31 20:26:46*	Data da integração da equipa na liga privada

- **round**

Coluna	**Tipo**	**Exemplo**	**Observações**
[id]	int		Chave: Identificador único da ronda na temporada considerada - corresponde à ordem em que a ronda é jogada
[date_end_bets]	datetime	*2018-09-27 19:15*	Data em que fecham as apostas (alterações às equipas) na ronda considerada. Ocorre 2h antes do primeiro jogo da jornada
[date_publish]	datetime	*2018-10-02 18:00:00*	Data de publicação dos resultados obtidos pelas equipas e jogadores na ronda considerada. Até 2016/17 ocorria às 12h da quarta-feira a seguir ao último jogo da jornada. Nas 2 últimas épocas, ocorre às 18h do dia seguinte ao último jogo da jornada.

- **user_login**

Coluna	**Tipo**	**Exemplo**	**Observações**
[id]	int		**Chave:** Identificador único sequencial do registo atribuído automaticamente
[id_user]	int		Chave estrangeira: Identificador do utilizador
[timestamp]	datetime	*2018-08-04 13:52:06.737*	Data de início da visita (login) no site

ii. Tabelas representativas da "realidade" - dados não inerentes ao jogo (com excepção dos valores atribuídos aos jogadores):

- **r_club**

Coluna	Tipo	Exemplo	Observações
[id]	int		**Chave:** Identificador único sequencial do clube real atribuído automaticamente
[name]	string	*Benfica, Sporting, Porto*	Nome do clube

- **r_gender**

Coluna	Tipo	Exemplo	Observações
[id]	int		**Chave:** Identificador único para o género do concorrente (0: Masculino; 1: Feminino)
[name]	string	*Masculino, Feminino*	Nome do género do concorrente

- **r_player**

Coluna	Tipo	Exemplo	Observações
[id]	int		**Chave:** Identificador único sequencial do atleta (jogador) atribuído automaticamente
[name]	string	*Jonas*	Nome do atleta
[id_club]	int		Chave estrangeira: identificador do clube a que o atleta pertence

- **r_playerposition**

Coluna	Tipo	Exemplo	Observações
[id]	string	*GR, DF, MD, AV*	Chave: Identificador único atribuído à posição que o jogador normalmente ocupa no campo
[name]	string	*Guarda-redes, Defesa, Médio, Avançado*	Nome por extenso da posição no campo

- **r_round**

Coluna	Tipo	Exemplo	Observações
[id]	int		**Chave:** Identificador único sequencial da jornada real
[order]	int	*31*	Ordem na temporada em que a jornada é jogada *(valores de 1 a 34)*
[name]	string	*3ª Jornada*	Nome por extenso da jornada
[id_round]	int		Chave estrangeira: identificador da ronda do jogo
[date_start]	datetime	*2018-11-02 20:30*	Data de início da jornada (data/hora do primeiro jogo da jornada – normalmente à sexta-feira)
[date_end]	datetime	*2018-12-16 22:00*	Data de fim da jornada (ocorre 2h depois do início do último jogo da jornada – normalmente à segunda-feira)

- **r_region**

Coluna	Tipo	Exemplo	Observações
[id]	int		**Chave:** Identificador único sequencial da região atribuído automáticamente
[name]	string	*Lisboa*	Nome da região

iii. Tabelas de cache - registam cálculos que vão sendo efectuados jornada a jornada, com base nos resultados dos jogos reais e apostas dos concorrentes:

- **cache_player_round**

Coluna	Tipo	Observações
[id_player]	int	**Chave:** Identificador único do atleta
[id_round]	int	**Chave:** Identificador único da ronda
[rank]	int	Posição do atleta no ranking da Liga Record *(range nas várias épocas/jornadas: 1 a 608)*
[points_round]	int	Pontos obtidos pelo atleta na ronda *(Range nas várias épocas/jornadas: -19 a 27. Média = 0.6)*
[points_total]	int	Pontos totais obtidos pelo atleta desde o início da temporada na Liga Record *(Range nas várias épocas/jornadas: -47 a 216. Média = 9.4)*
[value]	int	Valor monetário do atleta à data da realização da ronda *(Range nas várias épocas/jornadas: 500 mil a 10 milhões e 450 mil. Média = 932 mil 627)*

- **cache_team_round**

Coluna	Tipo	Observações
[id_team]	int	**Chave:** Identificador único da equipa
[id_round]	int	**Chave:** Identificador único da ronda
[points_round]	int	Pontos obtidos pela equipa na ronda *(range nas várias épocas/jornadas considerando equipas válidas e inválidas: -55 a 126. Média = 43.5)*
[rank_round]	int	Posição da equipa na ronda no ranking da Liga Record *(range nas várias épocas/jornadas: 1 a 62164)*
[points_total]	int	Pontos totais obtidos pela equipa desde o início da temporada na Liga Record *(range nas várias épocas/jornadas considerando equipas válidas e inválidas: -189 a 1947. Média = 656.6)*
[rank_total]	int	Posição da equipa no ranking geral da Liga Record desde o início da temporada *(range nas várias épocas/jornadas: 1 a 62164)*
[value_playing_players]	int	Soma do valor de compra dos jogadores que entraram em campo na ronda *(Range nas várias épocas/jornadas*

		considerando equipas válidas e inválidas: 0, valor decorrente de equipa inválida, a 35 milhões. Média = 25 milhões 476 mil 833)
[value_team_complete]	int	Soma de valor de compra de todos os jogadores da equipa *(Range nas várias épocas/jornadas considerando equipas válidas e inválidas: 500 mil, valor decorrente de equipa inválida, a 41 milhões, 40 milhões base + 1 milhão de bónus. Média = 38 milhões 461 mil 755)*

- **cache_player_team_round**

Coluna	Tipo	Observações
[id_team]	int	**Chave:** Identificador único da equipa
[id_player]	int	**Chave:** Identificador único do atleta
[id_round]	int	**Chave:** Identificador único da ronda
[is_playing]	bool	Indica se o jogador foi colocado em campo na ronda referida
[default_value]	int	Valor inicial de compra do jogador *(Range nas várias épocas: 500 mil a 10 milhões e 450 mil)*

iv. Tabela de jogos reais, com fonte na SportRadar

Através das datas dos jogos, permite saber quais as rondas em que ocorrem clássicos (jogos entre Benfica, Sporting e FC Porto)

- game

Coluna	Tipo	Exemplo	Observações
[id]	int		**Chave:** Identificador único sequencial do jogo atribuído automaticamente
[id_home_team]	int	*1 (Corresponde ao FC Porto)*	Identificador da equipa da casa
[id_visitor_team]	int	*3 (Corresponde ao Benfica)*	Identificador da equipa visitante

[goals_home_team]			Golos marcados pela equipa da casa
[goals_home_team]			Golos marcados pela equipa visitante
[date_start]		*2018-04-15*	Data do jogo

v. Tabelas CTT usadas para normalização de moradas (apenas são indicadas as tabelas/colunas consideradas no projecto):

Sempre que o concorrente (tabela *user*) tem morada em território nacional e esta é válida, existe uma ligação entre os 4 dígitos iniciais do código postal e o *num_cod_postal* da tabela de *codigos_postais*

Dados fonte obtidos em formato CSV

- **CODIGOS_POSTAIS**

Coluna	Tipo	Observações
[cod_distrito]	int	Chave estrangeira: Código do distrito
[cod_concelho]	int	Chave estrangeira: Código do concelho
[cod_localidade]	int	Chave estrangeira: Código da localidade
[nome_localidade]	string	Nome da localidade (ex: Paço de Arcos)
[num_cod_postal]	numeric(4)	Número do código postal (ex: 2700)
[ext_cod_postal]	numeric(3)	Extensão do número do código postal (ex: 112)
[desig_postal]	string	Designação postal (ex: PAÇO DE ARCOS)

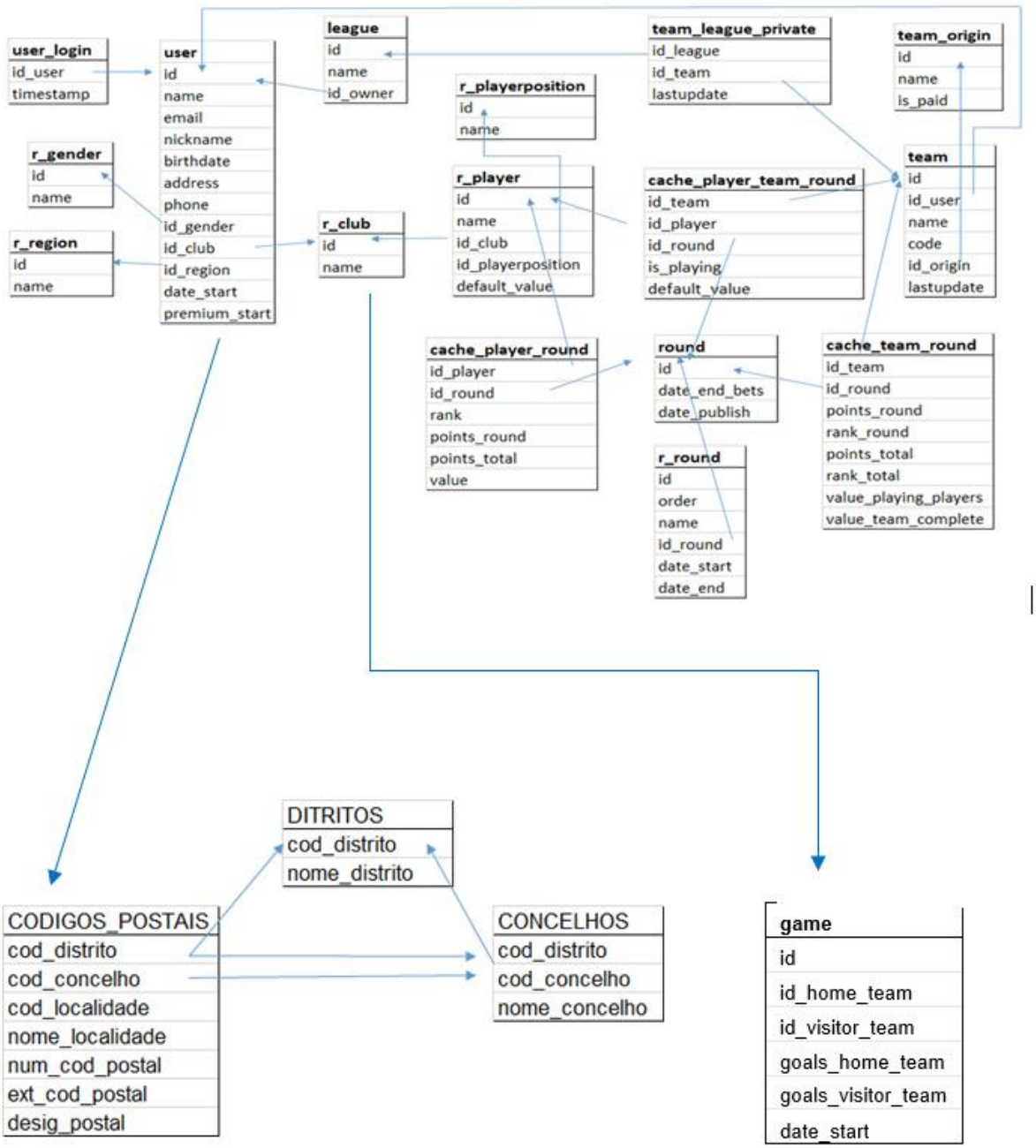
- CONCELHOS

Coluna	Tipo	Observações
[cod_concelho]	int	**Chave:** Código do concelho
[nome_concelho]	string	Nome do concelho (ex: Oeiras)
[cod_distrito]	int	Chave estrangeira: Código do distrito

- **DISTRITOS**

Coluna	Tipo	Observações
[cod_distrito]	int	Chave: Código do distrito
[nome_distrito]	string	Nome do distrito (ex: Lisboa)

b. Diagrama relacional dos dados originais



c. Problemas decorrentes dos dados

Não existe qualquer verificação das moradas indicadas pelos concorrentes. Assim sendo, mesmo tendo ao nosso dispõr a base de dados completa de códigos postais dos CTT, muitas moradas não puderam ser encontradas, tanto por estarem incompletas como por estarem erradas.
O esforço de normalização de moradas consistiu na aplicação de scripts às moradas existentes de forma a separar nome da rua, localidade, componentes do código postal e país.
Nos casos onde foram encontradas incoerências não resolúveis e as moradas indicavam ser de Portugal, foi atribuída uma localização genérica apenas com o país (como é feito para os concorrentes moradores em território estrangeiro). As moradas portuguesas sem identificação de código postal totalizam **1.28%** dos dados.
Também foram encontradas incoerências no que diz respeito a morada por extenço e indicador de país. Por exemplo, apareceram algumas referências ao país afeganistão **(0.24%)**, o que terá ocorrido por este ser o primeiro país a aparecer na lista fornecida para escolha do país.

Por outro lado, o campo "região" identificado na tabela de "users" (concorrentes), que podia ser um bom indicador da morada do concorrente, teve que ser descartado para geolocalização da morada, porque pela análise dos dados demos conta que, num número significativo de casos **(5.34%)**, este não estaria de acordo com a morada indicada, o que nos faz pensar que muitos concorrentes terão visto este campo como sendo indicador da sua região de origem e não de morada. Assim sendo, esta classificação só poderá ser usada como uma "preferência" do concorrente tal como o clube e não como indicativo do local onde mora.

Também encontrámos falhas na indicação região **(0.11%)**, clube de preferência **(1.4%)** e data de nascimento. Neste último indicador, temos 7 concorrentes que indicaram terem nascido depois de 2018, o que nos parece altamente improvável.

Outro problema verificado foi que os identificadores usados para os concorrentes nas 3 primeiras temporadas são diferentes dos usados nas 2 últimas. Os concorrentes foram identificados como sendo a mesma pessoa através do seu endereço email. No futuro sistema ETL, a identificação única dos concorrentes vai passar a ser feita através de uma chave super natural.

d. Análise dos dados fonte

Algumas tendências podem já ser observadas nas tabelas dos concorrentes, que descrevemos a seguir.

i. Rácio etário dos concorrentes

Entre os concorrentes, predominam aqueles entre os 35 e os 44 anos, com o resto dos números das outras faixas etárias cumprindo uma distribuição normal.

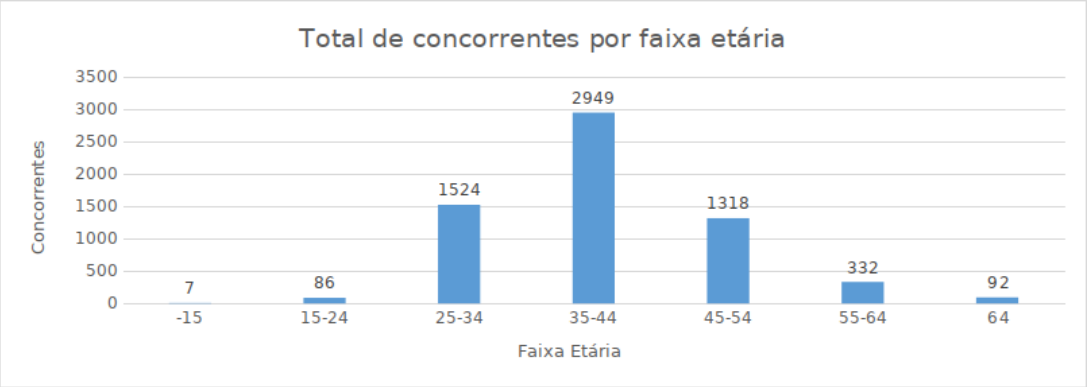


Fig 1: A distribuição dos concorrentes cumpre uma distribuição normal centrando-se na faixa etária dos 35-44 anos, região onde se encontra a maior porção de pessoas residentes em Portugal, no entanto fatores adicionais devem de centrar a distribuição neste valor.

Esta distribuição não é nenhuma surpresa dado que entre os 35-44 anos encontramos deveremos de encontrar um conjunto de adultos que participam e interagem em conjunto eventualmente atraindo concorrentes em faixas etárias próximas. Faixas etárias abaixo desta podem não estar sequer naturalizados com o conceito da Liga Record, ou por não ser um jogo popular na sua faixa etária, ter os seus interesses movidos fora deste jogo. O mesmo se aplica a concorrentes acima desta faixa etária. Pelo ano de 2017, os censos para a contagem de residentes por faixa etária indicam para um máximo entre os 40 e os 44 anos, fenômeno que pode também estar a ser verificado nos nossos dados. **[1]**

Pelo jornal Record ser distribuído tradicionalmente de uma forma física (embora também em formato digital), e pela Liga Record ser jogada em plataforma digital, encontramos aqui uma clara dicotomia que pode levar a esta clara centralização do público alvo entre os 35-45, sendo ambos leitores acostumados à leitura e distribuição do formato físico do jornal, e possuindo o conhecimento de informática na óptica do utilizador suficiente para efetivamente participar. No entanto de forma a apoiar esta declaração seriam necessários dados adicionais aos quais não temos acesso neste projecto.

ii. Rácio de género dos concorrentes

Tendo em conta também o género dos concorrentes, pode-se verificar a inclinação do futebol em Portugal para o público masculino.

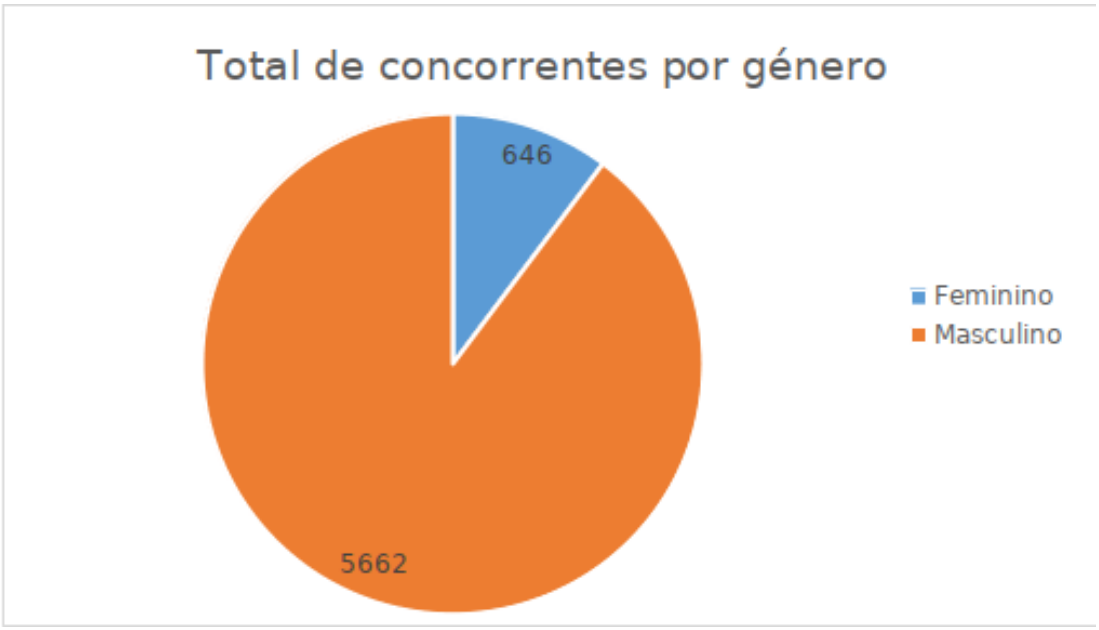


Fig 2: Os números que o nosso dataset apresenta aponta apenas para os participantes no jogo online, não para a popularidade do futebol de um modo geral entre a população de acordo com o sexo.

Reservando-se os números apresentados às contagens da Liga Record, e sendo os formulários preenchidos pelo utilizador de forma arbitrária, nada impede que o público feminino esteja sub representado quer por falta de aderência ao jogo online em particular, ou por "mau" preenchimento do formulário. Seria interessante comparar os resultados encontrados com estatísticas reais discriminadas de acordo com sexo, de comparência a jogos de futebol reais, audiências televisivas, outros jogos online etc.

Claro que o sexo não deverá de afectar minimamente a pessoa na escolha do clube. É isso mesmo que encontramos quando discriminados os participantes por clube e sexo, apresentando o sexo na mesma uma proporção claramente discriminada e regular entre os vários clubes.

Fig 3: Quando partidos em clubes, as proporções dos participantes do sexo feminino e masculino permanecem semelhantes.

iii. Distribuição ao longo do país

Os dados indicam-nos uma clara popularidade do jogo na região de Lisboa cumprindo mais do dobro do número de jogadores que a cidade do Porto que permanece no segundo lugar nas contagens. Este factor pode ser o resultado combinado da alta densidade populacional das regiões mencionadas, que aumenta em escala o número de potenciais participantes.

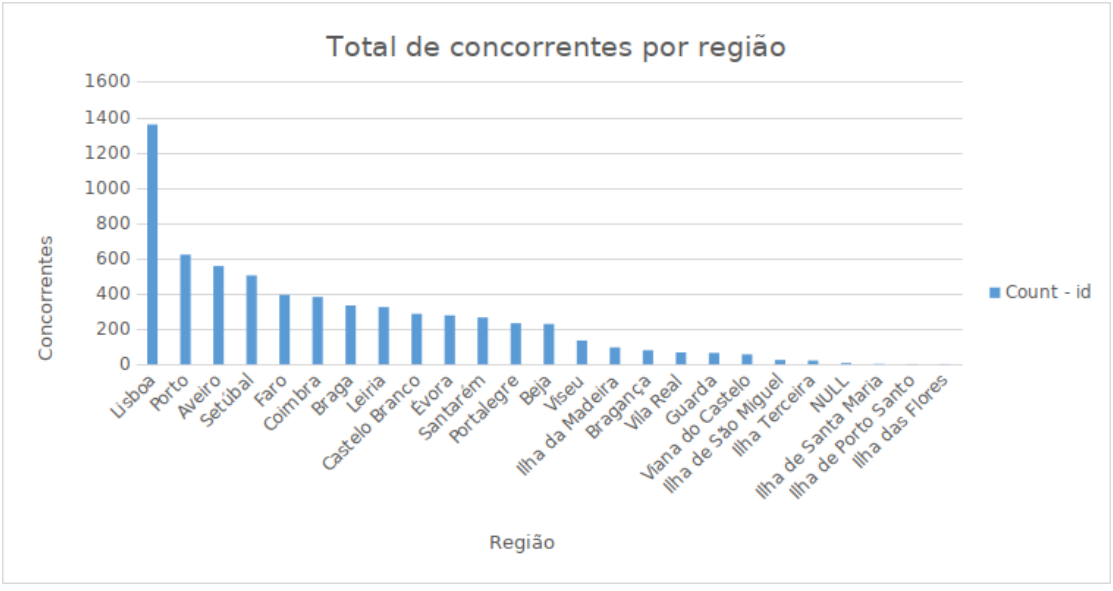


Fig 4: Os elevados números da população em cidades como a de Lisboa deverão de ter um factor dominante no número de concorrentes.

Da mesma forma e curioso é a visível preferência por parte dos concorrentes na escolha dos 3 grandes clubes portugueses - Benfica, Sporting, Porto - com o Benfica a dominar a região de Lisboa e o Porto a seguir a mesma tendência.

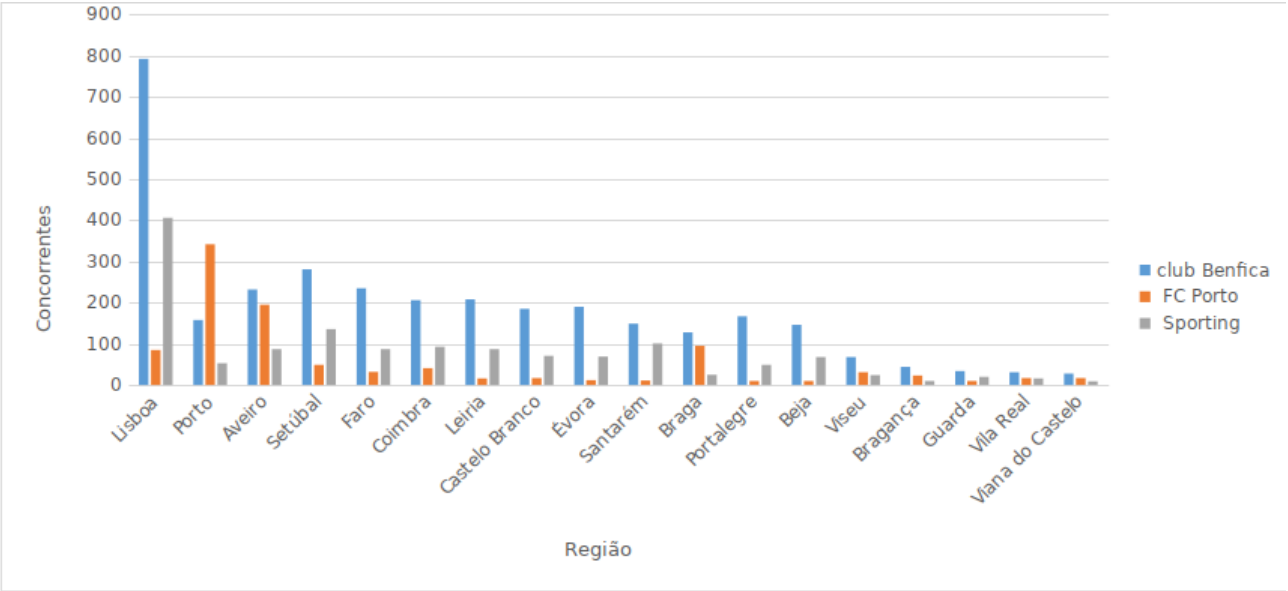


Fig 5: O favoritismo regional está bastante presente para os grandes clubes, com Benfica a apresentar um claro domínio em outras cidades.

3. Descrição do processo de negócio

Nos últimos anos, com o aparecimento de outros jogos em portugal com moldes semelhantes mas com participação grátis, a Liga Record tem vindo a baixar em popularidade, perdendo tanto concorrentes como patrocínios todos os anos o que pode comprometer o funcionamento do jogo, dificultando o financiamento dos prémios dos seus concorrentes, e a manutenção do mesmo.
Com a análise dos dados de alguns dos seus concorrentes mais fiéis, é esperado encontrar o que os motiva a jogar todos os anos e o que pode ser mudado na abordagem da Cofina no jogo, de modo a que se atrase ou reverta a tendência decrescente de popularidade.

Sobre estes concorrentes, vamos analisar vários tipo de informação, por exemplo relativa ao seu escalão etário, género, clube de preferência, morada, se subscreveram serviço premium e quando e em que anos participaram em ligas privadas.
Vamos também analisar a evolução dos seus resultados no jogo com a sua participação (visitas) ao longo das várias épocas na tentativa que encontrar padrões no diz respeito a maior/menos participação.

A estrutura base do jogo é organizada em rondas que se distribuem em temporadas, da seguinte forma:

- ***Temporada*** **–** campeonato a que correspondem: 2014/15, 2015/16, 2016/17, 2017/18 e 2018/19
- ***Ordem*** **–** ordem da ronda ao longo da temporada;
- ***Data de início*** **–** Data em que os concorrentes podem iniciar as alterações à equipa que pretendem que vá a jogo;
- ***Data de fim de apostas*** **–** Data em que fecha a ronda e já não são permitidas mais alterações. Tipicamente, esta data é às sextas-feiras e, nas quatro últimas temporadas, corresponde a 2 horas antes do início do primeiro jogo da jornada. Antes disso, fechava a uma hora fixa: 18 horas de sexta-feira. Nestas datas são esperados aumentos das visitas ao site da Liga Record relativamente ao habitual.
- ***Data de publicação de resultados*** **–** data posterior ao último jogo da jornada, quando os resultados obtidos pelos concorrentes são publicados no site da liga record e no jornal Record (tipicamente às terças-feiras). Tal como na data de fim de apostas, é esperado um aumento das visitas ao site.

Como foi referido, em cada temporada, cada concorrente pode constituir várias equipas que poderão apresentar composições diferentes em campo a cada jornada.

Para cada equipa, existem os seguintes dados:

- *Temporada de constituição*;
- *Identificador numérico único por temporada*;
- *Nome*;
- *Data de criação*;
- *Origem* – indica como a equipa foi criada: se foi comprada (através da compra da revista Record, lançada anualmente em Julho ou adquirida por MB, Paypal ou CC) ou se é uma equipa grátis (oferta inerente ao jogo, equipa de teste ou oferta com vista a publicitar a Liga Record);
- *Indicador se a equipa tem ou não valor monetário associado* - dado obtido a partir do campo *origem*;
- *Identificação do concorrente a que pertence*.

4. Perguntas analíticas

Foram projectadas as seguintes perguntas analíticas pertinentes ao projecto de negócio:

- Sendo o objectivo do cliente agarrar as pessoas ao jogo e levar à compra de mais equipas, e uma vez que é expectável que os concorrentes que têm melhores resultados sejam mais participativos e fiéis ao jogo, considerámos que seria interessante saber se existe algum padrão nas preferências/dados dos concorrentes no sucesso final das suas equipas. Dados como o clube de preferência, região, escalão etário, género ou subscrição do serviço premium são assim classificações que queremos avaliar em comparação com a pontuação final obtida pelas equipas.
  Como consequência da análise assinalada, pretendemos também saber se o sucesso no jogo numa determinada época influencia a aquisição de maior número de equipas na temporada seguinte.

- Faz parte da construção do jogo a atribuição de maior valor monetário aos jogadores (atletas) dos clubes principais e mais cotados no campeonato da 1ª liga. Assim sendo, ainda tentando avaliar quais os factores que influenciam o sucesso das equipas no jogo, pretendemos saber se o valor combinado do 11 colocado em campo pelos concorrentes têm influência directa ou não no sucesso final da equipa.

- Evolução da participação ao longo do tempo (por temporada, ronda a ronda, dia da semana, rondas correspondentes a jornadas em que ocorrem os jogos mais importantes). Sabemos que os concorrentes sobre os quais assenta a nossa análise são à partida pessoas que acompanham o jogo de forma fiel (participam em ligas privadas e já jogam a Liga Record há pelo menos 5 anos).
  Gostaríamos de conseguir determinar se as movimentações destes concorrentes acompanham a tendência geral da generalidade dos outros concorrentes (que não vamos analisar no âmbito deste projecto, mas sobre os quais temos conhecimento através de ferramentas de análise de interactividade no site da Liga Record e da base de dados que nos serviu de fonte), ou seja, se a sua participação tem picos nas rondas iniciais e nos dias chave (datas de fim de apostas da ronda e datas de publicação de resultados), ou, se por outro lado, obedecem a uma distribuição mais uniforme ao longo do tempo.

5. Modelação Dimensional

Nesta etapa estão planeadas a criação de 4 tabelas de dimensões - SEASON DIMENSION, USER DIMENSION, DATE DIMENSION e TEAM DIMENSION. Serão também criadas 2 tabelas de factos para guardar os processos de negócio que iremos analisar – dados de resultados obtidos pelas equipas na ronda (TEAM RESULTS FACT) e registo diário de visitas por utilizador (VISITS FACT). As tabelas de dimensões são todas conformadas e partilhadas entre as tabelas de factos, à excepção da tabela de dimensões TEAM DIMENSION, que apenas está ligada a tabela de factos TEAM RESULTS FACT.

Com as tabelas declaradas podemos então partir para uma uma explicação mais detalhada sobre cada uma no tópico a seguir.

a. Tabelas de Factos

i. Tabela de factos de registo de visitas:
VISITS FACT (cerca de 3 milhões de registos)

«Um concorrente (user) com uma determinada chave super natural numa certa data de uma temporada (season) faz um certo número de visitas ao site da liga Record»

EX: O concorrente 13075 (de nickname *PUNISHER*) com chave super natural 160, no dia 20161104 (2016-11-04) da temporada 201516 (2015/16) fez 4 visitas.

- **Origem dos dados**

As medidas consideradas nesta tabela têm origem directamente no sistema de dados da Liga Record.

- **Grão**

Cada linha na tabela de factos de visitas irá corresponder à contagem de logins por dia e utilizador. Com esta informação vai ser possível quantificar quais os dias que despertam mais interesse por parte dos concorrentes. Por exemplo, é expectável que as datas de fecho de ronda (principalmente a primeira), datas de publicação de resultados e datas em que é possível fazer troca de jogadores do plantel em fevereiro (mercado de inverno) sejam as mais participadas. Esta tabela permite-nos fazer essa quantificação. Se o concorrente não iniciar sessão no site num determinado dia não existirá registo de visitas para esse concorrente nesse dia.

- **Tipo**

Como os factos são registados em períodos fixos de um dia, o grão correspondendo a uma linha por dia, o carregamento de factos feito apenas por inserções, a actualização dos mesmos inexistente, a dimensão data registada correspodendo ao fim do período e com os factos

representando o número de visitas por período, esta pode ser considerada sem dúvidas de tipo Instantânea Periódica.

- **Chaves Estrangeiras**

Season Key (FK) – Chave estrangeira, liga a dimensão Season para identificar a Época corrente no jogo online.

Date Key (FK) – Chave estrangeira, liga a dimensão data de modo a caracterizar temporalmente o facto, junto com outras informações qualitativas que podem eventualmente ser pertinentes para possíveis questões analíticas (corresponde a fim de semana? Semana? Feriado?).

User Key (FK) – Chave estrangeira, liga dimensão user associando toda a informação de um user ao facto registado.

Durable User Key (FK DK) – Chave estrangeira supernatural, liga de novo a dimensão user como uma chave mais "resistente", na eventualidade e possibilidade de identificar o mesmo utilizador registado com dados diferentes.

- **Medidas**

Visit Count (Medida aditiva) - A única medida presente está neste atributo, que corresponde ao número de visitas (delimitados por início e fim de sessão) durante o período de um dia.

Existem dois cenários possíveis nesta medida:

- Uma única sessão num dia - Um utilizador que inicie sessão pode permanecer com a sessão ligada durante o dia inteiro. O sistema só a encerra uma veza cada 24h.
- Várias sessões no próprio dia - é comum utilizadores que interrompam a sua sessão ao longo do dia ou que acedam ao sistema em vários dispositivos e situações (por telemóvel, tablet, no computador de casa ou do emprego). Neste caso vão existir várias visitas por utilizador no mesmo dia.

ii. Tabela de factos de registo de valor e resultados das equipas por ronda: TEAM RESULTS FACT (cerca de 1 milhão e meio de registos)

«Uma equipa (team) pertencente a um concorrente (user) com uma determinada chave natural, numa certa data de publicação de uma temporada (season) obtém uma quantidade de resultados com uma equipa de um certo valor»

EX: A equipa 943 (de nome TOMAS TEAM) pertencente ao concorrente 25748 (de nickname tmsousa) com chave super natural 191, no dia 20150121 (2015-01-21, pertencente à 15ª ronda) da temporada 201415 (2014/15) teve 48 pontos na ronda, ficando no 32.305º lugar dessa mesma ronda.

- **Origem dos dados**

As medidas consideradas nesta tabela têm origem directamente no sistema de dados da Liga Record.

- **Grão**

Cada linha na tabela corresponde aos resultados obtidos e valor da equipa em determinada ronda. Estes dados são gerados nas datas de publicação de cada ronda (que, uma vez que estas correspondem a jornadas do campeonato nacional da 1ª liga de futebol, acontece um número pré-determinado de vezes em datas de intervalos diferentes mas conhecidos à partida) e sempre que é necessária alguma correcção a resultados previamente publicados.

- **Tipo da tabela de factos**

Como foi referido, os dados são registados em datas pré-determinadas mas com intervalos variados. Os dados já publicados podem ter que ser actualizados em 2 circunstâncias: sempre que há um jogo diferido (realizado depois da data oficial da jornada) ou quando é aceite alguma reclamação relativa a pontos atribuídos a alguma equipa dos concorrentes. Em ambos os casos, a data da actualização não é conhecida à partida. A data de publicação mantém-se a mesma, apenas mudam os valores quantitativos da tabela. Tendo em conta o que foi descrito, esta trata-se de uma tabela de factos Instantânea Comulativa.

- **Chaves estrangeiras**

Season Key (FK) – Chave estrangeira, liga à dimensão que representa a temporada (season) na qual os concorrentes estão a pontuar com as suas equipas.

Round Publish Date Key (FK) – Chave estrangeira, liga à dimensão data. Corresponde à data de publicação de resultados da ronda a que o registo corresponde.

Round End Bets Date Key (FK) – Chave estrangeira, liga à dimensão data. Corresponde à data final de apostas da ronda a do registo.

Round Start Date Key (FK) – Chave estrangeira, liga à dimensão data. Corresponde à data de início de apostas da ronda do registo.

A ligação de tabela de factos de resultados na ronda ligada a estas 3 datas chave, uma vez que as tabelas de dimensão usadas em ambas as tabelas de factos são conformadas, permite que se use a técnica drill across para, na fase de criação do cubo de dados, cruzar resultados desta tabela de factos com a quantidade de visitas recolhidas pela tabela de factos de registo de visitas dentro das datas chave de cada ronda.

Desta forma, será possível correlacionar o interesse no jogo com os resultados obtidos.

Team Key (FK) – Chave estrangeira, liga à dimensão que representa a equipa que originou os resultados apresentados nas medidas.

User Key (FK) – Chave estrangeira, liga à dimensão que representa o concorrente (user) associando os resultados obtidos na ronda por uma determinada equipa a um concorrente específico.

Durable User Key (FK DK) – Chave estrangeira supernatural. Usada para identificar o concorrente através da sua chave original. Desta forma, mesmo sendo este um registo de factos já passados, temos sempre uma forma rápida de obter os dados mais recentes do concorrente.

- **Medidas**

Team Points Round (Medida aditiva) **–** Pontos obtidos pela equipa na ronda considerada através do somatório nessa ronda das pontuações dos jogadores que a compõem (fazendo deste um valor expandido).

Não havendo reclamações, mantém-se inalterada mesmo que o registo sofra actualizações devido a recálculo de pontuações gerais por jogos diferidos.

Team Points Total (Medida NÃO aditiva) **–** Pontos obtidos pela equipa desde o início do jogo e necessários para integrar a equipa na classificação geral. Estes pontos são actualizados sempre que sejam aceites reclamações **ou que haja resultados decorrentes de jogos diferidos**.

Como resultado da pontuação relativa a jogos diferidos só ser aplicada à pontuação geral, esta não consegue ser calculada através da soma das pontuações das várias rondas, o que implica que consideramos esta medida como tendo o mesmo grão da medida Team Points Round.

Este facto faz com esta seja uma medida não aditiva porque não faz sentido ser somada (é por si já uma soma, o que faz deste um valor expandido) mas também não pode ser calculada através de

dados presentes no sistema uma vez que o cálculo é feito em tempo de execução no sistema operacional.

Team Rank Round (Medida NÃO aditiva) **–** Posição da equipa na ronda considerada. Esta posição é conseguida pela ordenação da pontuação obtida na ronda e, em caso de empate, por regras inerentes ao jogo (explicadas no ponto 18.1 do regulamento da Liga record https://liga.record.pt/info/ajuda.aspx).
Tal como os pontos na ronda, o valor obtido no registo inicial só é recalculado caso haja alguma reclamação de um concorrente que seja aceite.
Como o valor do rank não consegue ser obtido apenas através de iteracção sobre os dados (como é usual noutros data warehouses), para ser relevante ao nosso estudo, necessita ser registado como uma medida, que neste caso será não aditiva (tal como o rank total apresentado de seguida).

Team Rank Total (Medida NÃO aditiva) **–** Posição da equipa na classifiestecação geral. Esta posição conseguida pela ordenação da pontuação geral e, em caso de empate, por regras de desempate próprias (diferentes das usadas na obtenção da posição na ronda). É actualizada nas mesmas circustâncias da coluna Team Points Total.

Utilized Team Players Value (Medida aditiva) **–** Soma dos valores de aquisição dos jogadores escolhidos para entrar em campo na ronda em questão (o que faz deste um valor expandido).
Este valor vai ser considerado nas perguntas analíticas para avaliar se a obtenção de melhores resultados tem relação com equipas mais caras.
Trata-se de uma medida aditiva, porque podemos validar se o total dos valores dos jogadores das equipas ao longo das rondas de uma temporada influencia de alguma forma a classificação final das equipas.

Utilized Team Players from Benfica / Porto / Sporting / Outros (Medidas aditivas) **–** Soma dos jogadores agrupados por clube (Benfica, Porto, Sporting ou outros clubes) escolhidos para entrar em campo na ronda em questão (o que faz deste um valor expandido).

b. Tabelas de Dimensões

Nesta secção pretende-se descrever cada um dos atributos que compõem as 4 tabelas de dimensões, tal como indicar a possível existência de hierarquias nos dados e o registo de mudanças lentas. Incluímos nos anexos deste relatório uma análise de estatística descritiva sobre as diversas variáveis das dimensões, com indicação de categorias e seus totais, bem como totais de *missing data* [**Anexo 1**].

i. Tabela de dimensão date – DATE DIMENSION (cerca de 1500 linhas)

Como é esperado da tabela de dimensões Date é comum às duas tabelas de factos. Além de determinar parte do grão das tabelas de factos, limitando cada linha ao período de tempo de um dia para cada facto, guarda atributos relevantes para possíveis questões analíticas como datas de eventos interessantes ou de dias de folga, no entanto a razão de cada atributo vai ser abordado na secção procedente.

- **Descrição dos atributos**

Atributo	Tipo de Dados	Descrição	Exemplo
Date Key (PK)	**int**	Chave primária da tabela Date, como tal identifica cada dimensão de forma única, garantindo segurança estrutural.	20190401
Day	**date (YYYY-MM-DD)**	Atributo data para permitir comparações entre datas.	2014-06-01
Day Of Month	**int (1:31)**	Atributo representativo do dia do mês. Os valores podem compreender qualquer número inteiro entre 1 e 31.	1
Weekday	**int (1:7)**	Atributo representativo do dia da semana. Os valores podem compreender valores inteiros entre 1 e 7, com o número 1 a representar segunda-feira,	1

		e assim consecutivamente, até o número 7 a representar domingo.	
Calendar Weekday	**varchar (“segunda-feira”, …, “domingo”)**	Atributo representativo do dia da semana, igual ao atributo Weekday em significado mas aceitando valores de tipo varchar, escrito por extenso desde "segunda-feira" a "domingo".	quarta-feira
Month	**int (1:12)**	Atributo representativo do mês, os valores podem compreender valores inteiros entre 1 e 12, com 1 a corresponder ao mês de Janeiro, e assim consecutivamente, até 12 a representar o mês de dezembro.	2
Calendar Month	**varchar (“janeiro”, “fevereiro”, …, “dezembro)**	Atributo representativo do dia do mês, em significado igual ao atributo Month mas aceitando valores varchar, escrito por extenso desde "janeiro" a "dezembro"	fevereiro
Year	**int**	Atributo representativo do ano.	2016
Date Full	**date**	Atributo representativo da data completa. Uma dimensão degenerada, junta os valores de Day, Month, e Year.	2014-06-01
Weekend Indicator	**varchar (“weekend”, “weekday”)**	Atributo usado para determinar se o dia corresponde a um fim de semana ou não. Cada instância contém apenas valores TRUE ou FALSE.	weekday
Season Stage Indicator	**varchar (“before-game-starts” “first-round”, “season-ongoing”, “last-round”, “after-game-ends”)**	Atributo usado para determinar a posição relativa da ronda a que data corresponde em relação ao período da época. Assume os valores "first", "ongoing" e "last", para identificar se se trata da primeira ronda, ronda intermédia ou última ronda da época respectivamente.	first-round

		Assume os valores “before-game-starts” e “after-game-ends” quando a data ocorre fora das datas em que ocorrem rondas.	
Turn	**int** **(0, 1, 2)**	Atributo usado para determinar se o dia corresponde à primeira ou segunda volta. Considera-se aqui a volta para quais são válidas as apostas dessa ronda da Liga Record e não exactamente as datas das voltas no campeonato real. Em termos do campeonato de futebol da 1ª liga, a 1ª volta joga-se na primeira metade do campeonato, terminando normalmente em janeiro, e a 2ª volta na segunda metade. Em termos de Liga Record, na 1ª volta disputa-se o campeonato de Inverno e na 2ª o de Verão, cada um com prémios. 0 é usado quando data sai fora dos períodos das rondas.	1
Turn Indicator	**varchar (“turn-start”, “turn-ongoing”, “turn-finish”, “not-a-turn”)**	Atributo usado para determinar a posiçao relativa da data em relação ao período de tempo compreendido entre o início e o fim da volta. Assume valores “turn-start", "turn-ongoing" e “turn-finish”, para identificar se se trata da primeira ronda, ronda intermédia ou última ronda da volta respectivamente. O valor “not-a-turn” é usado quando data sai fora dos períodos das rondas.	turn-finish
Round Number	**int** **(0, 1:31)**	Atributo usado para determinar o número da ronda corrente, assumindo valores inteiros entre 1 e 31. Apesar das Jornadas a que as rondas correspondem serem 34, a rondas da liga record só começam	3

		depois do fecho do mercado de Verão de jogadores, pelo que podem começar na 4ª ou 5ª jornada. Assume o valor 0 quando a data ocorre fora das datas em que ocorrem rondas.	
Round Lifecycle Indicator	**varchar (“results-publication-day”, “day-bets-end”, “day-bets-start”, “round-ongoing”, “not-a-round”)**	Atributo usado para determinar se a data em questão é uma data em que acontecem eventos chave da ronda (“results-publication-day”, “day-bets-end”, “day-bets-start”), se é um dia normal dentro da ronda (“round-ongoing”) ou fora do período das rondas (“not-a-round”).	non-publication-date
Lifecycle Round Number	**int (0, 1:31)**	Atributo usado para determinar o número da ronda a que corresponde a identificação do ciclo de vida, assumindo valores inteiros entre 1 e 31. Esta segunda data é necessária porque a apresentação de resultados de uma certa ronda ocorre no decorrer da ronda seguinte. Assume o valor 0 quando a data ocorre fora das datas em que ocorrem rondas.	2
Round Includes Classic Match	**varchar (“round-includes-classic-match”, “standard-match-round”)**	Atributo usado para identificar se a ronda a que a data corresponde inclui um clássico - jogo entre 2 dos 3 grandes (Benfica, Sporting e FC Porto).	standard-match-round
Is Winter Transfer Season	**varchar (“winter-transfer-season”, “non-winter-transfer-season”)**	Atributo usado para identificar se a data em questão corresponde a uma data durante a qual é permitido efectuar compras/vendas de jogadores em volume superior ao normal (6 ao longo do	non-winter-transfer-season

		período de transferência em vez do habitual 1 ou 2 por mês). Ocorre durante o mês de fevereiro.	

- **Origem dos dados**

As colunas ‘‘Round Includes Classic Match”, “Turn” e “Turn Indicator” têm origem nos dados SportRadar (sendo que “Turn” e “Turn Indicator” são depois mapeados para datas com sifnificancia na Liga Record).
As colunas “Round Number”, “Round Stage Indicator”, “Round Lifecycle Indicator”, e “Is Winter Transfer Season” têm origem em dados do Sistema da Liga Record.
As restantes colunas advêm directamente das propriedades da data.

- **Hierarquia de Dados:**

A tabela de dimensão Date contêm dois grupos distintos de hierarquias de profundidade fixa:

  - Date Full > Year > Month, Calendar Month > Day
  - Season Stage Indicator > Turn, Turn Indicator > Round Number, Round Lifecycle Indicator > Round Includes Classic Match

- **Registo de mudanças lenta:**

Não aplicável a nenhum dos atributos desta dimensão.

ii. Tabela de dimensão da temporada – SEASON DIMENSION (5 linhas)

Uma vez que o jogo da liga record segue o campeonato português da 1ª liga de futebol, que funciona por temporada (começando em agosto e terminando em maio), também aqui temos um funcionamento em temporadas, sendo que é no final de cada uma que são verificados os resultados totais obtidos pelas equipas e atribuídos os prémios principais.

- **Descrição dos Atributos**

Atributo	Tipo de Dados	Descrição dos Dados	Exemplo
Season Key (PK)	**int (representativo da temporada)**	Chave Primária da tabela season, como tal identifica cada dimensão de forma única, garantido segurança estrutural.	201819
Season Name	**varchar**	Atributo representativo do nome da temporada.	2018/19
Season Start Date	**date (YYYY-MM-DD)**	Atributo que indica a data em que a temporada começa (o que ocorre no dia 1 de julho de cada ano).	2018-07-01
Season End Date	**date (YYYY-MM-DD)**	Atributo que indica a data em que a temporada termina (o que ocorre no dia 30 de junho de cada ano)	2018-06-30
Season Has Updated Game Version	**varchar (“updated-game-version”, “deprecated-game-version”)**	Atributo que indica se na temporada em questão é usada uma versão actual do site da liga record ou se ocorre antes da última remodelação. Esta remodelação ocorreu na época 2015/16.	updated-game-version
Season Has Variable Weekday Publish Date	**variable (“variable-weekday-publish-date”,**	Atributo que indica se na temporada em questão a data de publicação de resultados de cada ronda é fixa num dia da semana ou se depende das datas em que são jogadas as datas da 1ª liga. Antes da	fixed-weekday-publish-date

	"fixed-weekday-publish-date")	temporada 2016/17 (inclusivé), estes dias eram fixos (à quarta-feira), depois passaram a ser variáveis.	
Team Player Transfers Allowed Per Month	**int (1, 2)**	Atributo que indicando quantas transferências de jogadores (venda de um e compra de outro) numa equipa são permitidas por mês. Até 2016/17 (inclusivé) era permitida apenas uma transferência mensal, depois passou a ser quinzenal.	2

- **Origem dos dados**

Todas as colunas têm origem em dados do Sistema da Liga Record.

- **Hierarquia de Dados:**

Não existem hierarquias de dados entre os atributos desta dimensão.

- **Registo de mudanças lenta:**

Não aplicável a nenhum dos atributos desta dimensão.

iii. Tabela de dimensão do concorrente – USER DIMENSION (cerca de 6 mil linhas por temporada – 32 mil total)

A tabela de dimensão User guarda todos os dados inerentes aos utilizadores, sejam quer dados pessoais e demográficos, tais como país, local de residência e faixa etária, quer dados relacionados com a sua actividade na plataforma, tais como data de registo ou data de subscrição à versão premium.

Nas seccões precedentes, iremos descrever os atributos quanto ao tipo e descrição dos dados, identificar as hierarquias existentes nos dados e quais as técnicas usadas para o registo de mudanças lentas.

- **Descrição dos Atributos**

Atributo	Tipo de Dados	Descrição dos Dados	Exemplo
User Key (PK)	**int,** **not null**	Chave primária da dimensão User, como tal identifica cada dimensão de forma única, garantido segurança estrutural. Representa também uma chave substituta para o registo de histórico segundo o uso da técnica 7 para mudanças lentas.	4
Durable User KEY (DK)	**int,** **not null**	Chave Supernatural que identifica cada utilizador de forma única. O seu valor nunca muda no registo de histórico.	34
User Natural ID (NK)	**int,** **not null**	Chave Natural. Atributo representativo da chave da equipa no sistema original.	760464
User Email	**nvarchar**	Atributo que indica o email fornecido pelo utilizador.	XXXX@hotmail.com
User Nickname	**nvarchar**	Atributo que indica o nome escolhido pelo utilizador para o indenficar na plataforma. Caso não esteja definido, toma o valor de “Utilizador Registado”	Rogerajato

User Birthdate	**date (YYYY-MM-DD)**	Atributo que indica a data de nascimento fornecida pelo utilizador.	1973-10-09
User Gender	**nvarchar (“feminino”, “masculino”)**	Atributo que indica o género do utilizador. Toma valores do tipo nvarchar, podendo assumir o valor “feminino” ou “masculino”.	Feminino
User Club	**nvarchar**	Atributo que indica o nome do clube de preferência fornecido pelo utilizador. Caso não não esteja definido, toma o valor de “Clube não definido”	Benfica
User Region	**nvarchar**	Atributo que indica o nome da região de residência indicada pelo utilizador. Este atributo apena pode ser considerado como uma “preferência” pela parte do utilizador e não como a sua região de residência, devido a incoerências detectadas em comparação com a morada fornecida. Caso não não esteja definida, toma o valor de “Região não definida”	Castelo Branco
User Zipcode Locality	**nchar**	Atributo que indica os 4 primeitos digitos do código de postal de residência fornecido pelo utilizador. Caso não não esteja definido, toma o valor de “Código postal não definido”	6230
User Zipcode Locality Designation	**nvarchar**	Atributo que indica o nome da localidade postal associada ao código postal – trata-se da localidade que costuma aparecer nas moradas logo a seguir aos digitos do código postal.	ALCAIDE

		Este campo foi obtido pelo cruzamento dos dados originais com a fonte de dados dos CTT. Caso não não esteja definida, toma o valor de "Localidade postal não definida"	
User Locality	**nvarchar**	Atributo que indica o nome da localidade de residência do utilizador. Na formação de códigos postais, aparece logo a seguir à morada. Este campo é obtido através do cruzamento dos dados originais com a fonte de dados dos CTT. Caso não não esteja definido, toma o valor de "Localidade não definida"	Acipreste
User County	**nvarchar**	Atributo que indica o nome do municipio de residência do utilizador. Este campo é obtido através do cruzamento dos dados originais com a fonte de dados dos CTT. Caso não não esteja definido, toma o valor de "Concelho não definido"	Fundão
User District	**nvarchar**	Atributo que indica o nome do distrito de residência do utilizador. Este campo é obtido através do cruzamento dos dados originais com a fonte de dados dos CTT. Caso não não esteja definido, toma o valor de "Distrito não definido"	Castelo Branco
User Country	**nvarchar**	Atributo que indica o nome do país de residência do utilizador.	Portugal

		Este campo é fornecido pelo utilizador, portanto, em situações em que não houve cruzamento directo com o código postal, pode não ser representativo do seu real país de residência.	
User Original Start Date	**datetime (YYYY-MM-DD HH:MM:SS)**	Atributo que indica a data e a hora de registo original do utilizador na plataforma.	2011-08-31 22:10:04.253
User Season Start Date	**datetime (YYYY-MM-DD HH:MM:SS)**	Atributo que indica a data e a hora de inscrição do utilizador. Este valor varia a cada nova temporada.	2018-08-19 10:33:10.030
User Premium	**varchar (“user-purchased-premium-option”, “user-not-premium”)**	Atributo que indica se o concorrente adquiriu a versão premium do jogo. Assume o valor “user-purchased-premium-option” quando o concorrente adquiriu o premium e “user-not-premium” caso contrário. Este valor varia a cada nova temporada.	user-purchased-premium-option
User Agegroup	**nchar (“-15”, “15-24”, “25-34”, “35-44”, “45-54”, “55-64”, “64”)**	Atributo que indica a faixa etária em que o utilizador se insere, a qual é função da sua data de nascimento. Assume 7 valores nchar diferentes, representativos de um determinado intervalo de idades. Este valor pode variar a cada nova temporada.	45-54
User Is In League	**varchar (“user-in-league”, “user-not-in-league”)**	Atributo que indica se o utilizador está inscrito numa liga privada. Este campo pode assumir valores do tipo bit, sendo que o TRUE indica que o utilizador está inscrito numa liga, e FALSE indica que	user-in-league

		não. Este valor pode variar a cada nova temporada.	
Effective Date Row	**datetime (YYYY-MM-DD HH:MM:SS)**	Atributo extra para a validação dos dados, necessária para o registo de mudanças lentas segundo a técnica de tipo 2 (adição de linhas). Assume valores do tipo date time que indicam a data e a hora em que a linha inicia a sua validade.	2018-08-19 10:33:10.030
Expiration Date Row	**datetime (YYYY-MM-DD HH:MM:SS)**	Atributo extra para a validação dos dados, necessária para o registo de mudanças lentas segundo a técnica de tipo 2 (adição de linhas). Assume valores do tipo date time que indicam a data e a hora em que a linha termina a sua validade.	9999-12-31 23:59:59.997
Timestamp Row	**datetime (YYYY-MM-DD HH:MM:SS)**	Atributo extra para a validação dos dados, necessária para o registo de mudanças lentas segundo a técnica de tipo 1 e 7 (substituição de um dado valor na linha). Assume valores do tipo date time que indicam a data e a hora da última actualização à linha.	2018-08-19 10:33:10.030
Is Current Row	**varchar (“current-row”, “deprecated-row”)**	Atributo extra para a validação dos dados, necessária para o registo de mudanças lentas segundo a técnica de tipo 7 (adição de linhas). Assume o valor TRUE, indicando que a linha é a actual, e o valor FALSE indicando que não.	deprecated-row

- **Origem dos dados**

Os dados relativos à morada têm origem nos dados obtidos dos CTT.

Os dados relativo ao clube e região de preferência têm origem em dados obtidos através do site da Liga Record.
As restantes colunas sem ligação directa ao sistema ETL têm origem em dados obtidos através do sistema centralizado de registo de utilizadores da Cofina e registados no sistema da Liga Record na altura do login – início da visita.

- **Hierarquia de Dados:**

A tabela de dimensão User contêm dois grupos distintos de hierarquias de profundidade fixa:

  - Country > District > County > Locality
  - User Key > User Original Key

- **Registo de mudanças lentas:**

A dimensão User contem diversos atributos cujos dados estão sugeitos a alteração ao longo tempo, dado que representa entidades dinâmicas (utilizadores).

Existem diversas técnicas descritas para registar o registo de mudanças lentas, sendo que no presente projeto iremos fazer uso de duas dessas técnicas, a de tipo 2 e a de tipo 7. Iremos então descrever cada uma dessas técnicas usadas, o porquê e os atributos em que se aplicam.

  - **Técnica Tipo 1**

A técnica tipo 1 é útil para a actualização pontual dos dados da tabela de dimensão, sendo muito utilizada para por exemplo a correcção de valores. Esta técnica consiste na substituição do valor existente para o valor mais actual, directamente na tabela da dimensão correspondente. Usámos esta técnica nos atributos que têm uma taxa de variação marginal, e para os quais não traz valor adicional à nossa análise, não sendo assim necessário manter o histórico das alterações mas apenas o valor mais recente.

Os atributos da dimenção User aos quais é aplicada a técnica tipo 1 são:

- User Region

- User Club

  - **Técnica Tipo 7**

A técnica tipo 7 é considerada uma técnica híbrida, isto porque também usa a técnica tipo 2. A técnica tipo 2 consiste na adição de uma linha na tabela de dimensão sempre que existe uma actualização dos valores de uma dada linha. Desta forma a tabela mantém tanto os dados existentes antes da alteração como após e, portanto, permite

manter o histórico das alterações efectuadas. Acoplada a esta adição de linhas tem de existir na tabela de dimensão, colunas extra que permitam:

- guardar chaves substitutas que irão permitir identificar de forma única o conjunto de linhas que derivaram de uma única linha ancestral (**User Key**). Normalmente estas linhas partilham uma chave supernatural (**User Original Key**);
- obter informação sobre a validade dos dados (**Row Effective Date**, **Row Expiration Date**);
- se a linha está actualmente em vigor (**Is Current Row**).

A técnica tipo 7 consiste na adição de linhas com actualizações, pela técnica 2, e na incorporação de uma chave estrangeira extra na tabela de factos (**User Original**). Apesar de mantermos na tabela dimensão User o histórico das alteracões, a chave estrangeira extra nas tabelas de factos referencia de uma forma rápida e prática a linha da dimensão que contém os dados actualizados. Usámos esta técnica nos atributos para os quais é importante manter o histórico das alterações de forma a ser possível perceber o impacto dessas mudanças ou decisões dos utilizadores no interesse pelo jogo.

Os atributos da dimenção User aos quais é aplicada a técnica tipo 7 são:

- Dados decorrentes de alterações na morada

- **User Zipcode Locality**
- **User Zipcode Locality Designation**
- **User Locality**
- **User District**
- **User Country**

- User Agegroup;

- User Premium Date;

- User Is In League

5.2.4. Tabela de dimensão da equipa – TEAM DIMENSION (cerca de 60 mil linhas)

Registo de dados das equipas formadas pelos concorrentes.

- **Descrição dos Atributos**

Atributo	Tipo de Dados	Descrição dos Dados	Exemplo
Team Key (PK)	**int**	Chave Primária da tabela team, como tal identifica cada dimensão de forma única, garantido segurança estrutural. Este atributo é do tipo inteiro representando a equipa em questão.	5447
Team Natural ID	**int**	Atributo representativo da chave da equipa no sistema original.	45045
Team Name	**varchar**	Atributo representativo do nome da equipa.	Gama Team
Team Create Date	**date (YYYY-MM-DD)**	Atributo indicando a data em que a equipa foi criada.	2017-02-15
Team Origin	**varchar**	Atributo indicando a origem da equipa. Os valores possíveis são os mesmos existentes para descriminar as origens de equipas no sistema operacional fonte.	REVISTA
Team Is Paid	**varchar (“team-is-paid”, “team-is-free”)**	Atributo que indica se a equipa é paga ou gratuita.	team-is-paid

- **Origem dos dados**

Todas as colunas têm origem em dados do Sistema da Liga Record.

- **Hierarquia de Dados:**

Não existem hierarquias de dados entre os atributos desta dimensão.

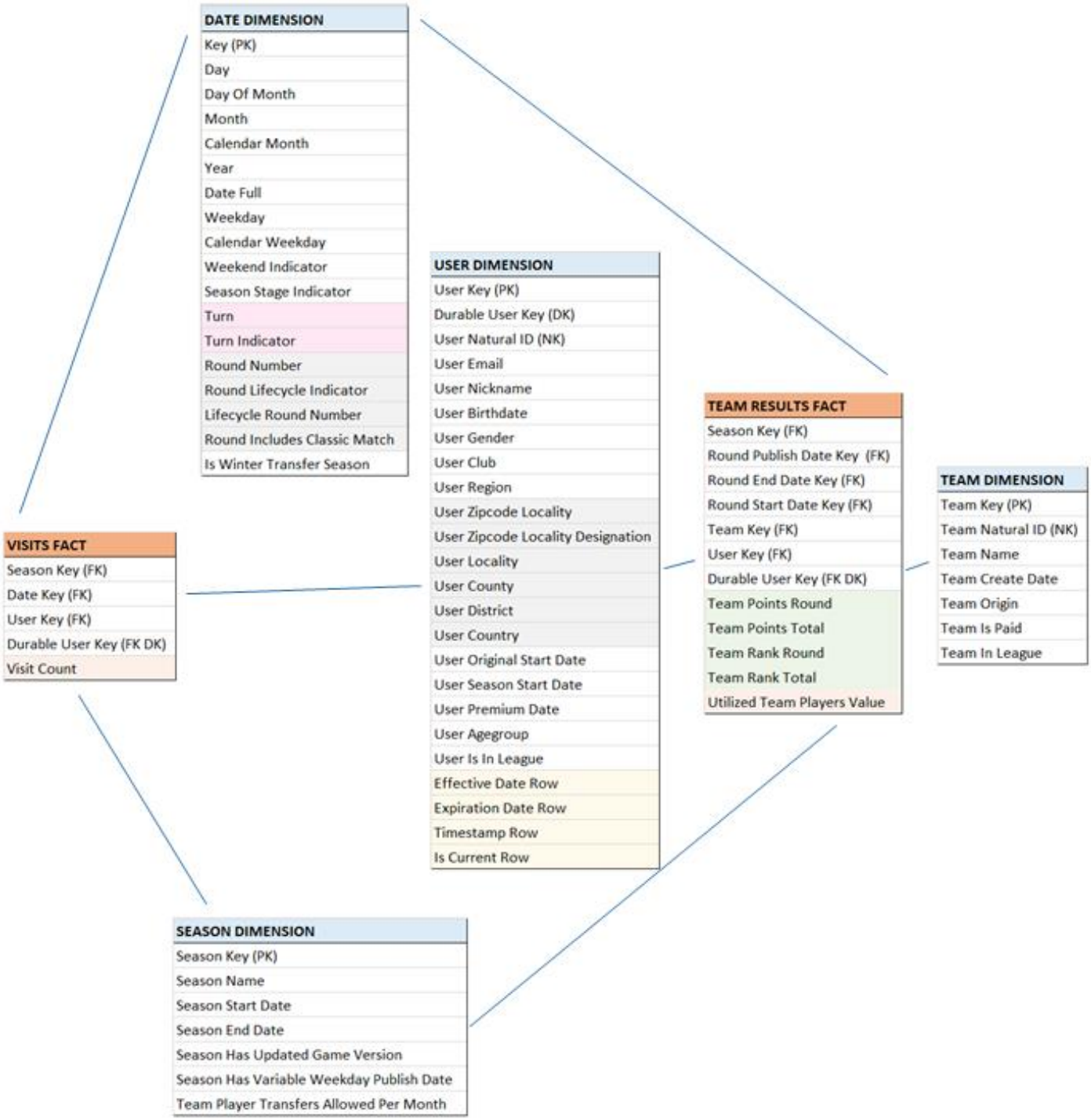
- **Registo de mudanças lenta:**

Não aplicável a nenhum dos atributos desta dimensão.

c. Diagrama em estrela do Data Warehouse

No nosso sistema usamos 2 tabelas de factos, pelo que o diagrama é desenhado através da composição de 2 diagramas em estrela que usam as mesmas dimensões conformadas.

Fig:6 Diagrama em estrela da modelação dimensional.



6. Sistema ETL

Nota:
O sistema ETL foi implementado numa máquina virtual alojada no Azure.
Os projectos de integration services, analysis services e análise no powerbi estão no desktop.

O acesso à máquina pode ser feito via Remote Desktop, com as seguintes credenciais:
Ip: 52.232.124.193:3389
User: tpd
Password: tpd05.TPD2019

As credenciais para entrar no Database Engine do Sql Managemet Studio são do tipo Sql Server Authentication:
User: sqladmin
Password: tpd05.TPD2019

a. Extracção

Tendo em conta que estamos a considerar apenas um subset dos dados disponíveis nas nossas fontes, o primeiro passo tratou de recolher esses dados dos servidores da Cofina através de interrogações SQL e exportados para flat files.

As duas bases de dados relacionais de interesse aqui consideradas são provenientes da Liga Record e da Sport Radar.
Os dados são recolhidos usando o Exportdata Wizard do SQL Server com base em queries SQL e exportados para ficheiros flat file. Devido ao volume, muitos dos dados aparecem separados por temporada:

- Ficheiros obtidos a partir de interrogações SQL aos dados da Liga Record:
    - *liga_record_concorrentes.txt*
      Obtido através do cruzamento da tabela de utilizadores que participaram nas 5 temporadas (com os critérios referidos anteriormente) com as tabelas correspondentes a clubes, regiões e género de forma a substituir os ID's destes dados pela sua referência descritiva.
      Neste caso, tal como nas alterações às moradas referidas a seguir, o ficheiro resultante já apresenta a transformação através de programas python descritos mais à frente com o objectivo de normalizar moradas até o nível da localidade.
      O ficheiro intermédio antes do tratamento das moradas é um CSV chamado users.csv.
    - *liga_record_concorrentes_address_history.txt*
      Através de interrogações SQL efetuadas sobre a listagem de utilizadores sob análise, previamente obtida, foram identificadas alterações de morada nalguns concorrentes. Estes dados, em conjunto com a tabela de concorrentes inicial, permite documentar mudanças lentas ocorrentes.

Tal como no caso anterior, o ficheiro resultante já apresenta a transformação através de programas python descritos mais à frente com o objectivo de normalizar moradas até o nível da localidade.
O ficheiro intermédio antes do tratamento das moradas é um CSV chamado user_moradas_historico.csv.

- *liga_record_concorrentes_details_history.txt*
  Através de interrogações SQL efetuadas sobre a listagem de utilizadores sob análise, previamente obtida, foram identificadas alterações de grupo etário, data de inscrição na Liga e compra do serviço premium. Estes dados, em conjunto com a tabela anterior, permite documentar mudanças lentas ocorrentes.
- '*201415_teams.txt', '201516_teams.txt', etc.*
  Ficheiros (um por época) com os dados das equipas pertencentes aos utilizadores obtidos nas declarações anteriores.
- '*rounds.txt*'
  Ficheiro obtido com base nas tabelas de dados das rondas de cada temporada.
- *'201415_round_team.txt', '201516_round_teams.txt', etc.*
  Obtidos através das tabelas de dados de equipas e resultados registados em cada ronda das várias temporadas.
  Devolve 5 ficheiros, cada um referente a cada temporada.
- *'visits.txt'*
  Obtido através de interrogações sobre as tabelas de logins dos utilizadores. Agrega as visitas por dia.

- Obtenção através de download do site dos ficheiros dos CTT com a listagem e descrição de moradas em Portugal:
  - ctt_codigos_postais.json
  - ctt_concelhos.json
  - ctt_distritos.json

- Identificação de dias em que foram jogados jogos clássicos e resultados dos mesmos com origem na base de dados Sport Radar e que irá permitir o enriquecimento da Date Dimension:
  - *'classicos.csv'*

b. Transformação

Antes dos ficheiros de dados serem passados para a Data Staging Area da Base de dados, ocorreram 2 transformações prévias:

- **Limpeza dos dados relativos a moradas**
  - address_finder.py e supernatural_key_assignment.py são os dois algoritmo de limpeza usado no sistema ETL. Pretendem responder à má conformação das moradas dos concorrentes inscritos, que foi antes verificado no relatório, e à atribuição de uma chave supernatural para registos de utilizadores que estejam inscritos em dois anos diferentes com dados diferentes, mas e-mails de registo iguais.
  - Entrada: Tabela de utilizadores não transformada em formato de CSV proveniente da área de extração.
  - Saída: Utilizadores com a morada normalizada e separada em país, distrito, concelho, localidade, designação e código postal, e uma chave supernatural para a identificação do mesmo utilizador em linhas diferentes.
    Estes utilizadores irão depois dar entrada na área de staging através dos ficheiros referidos na descrição do passo prévio.

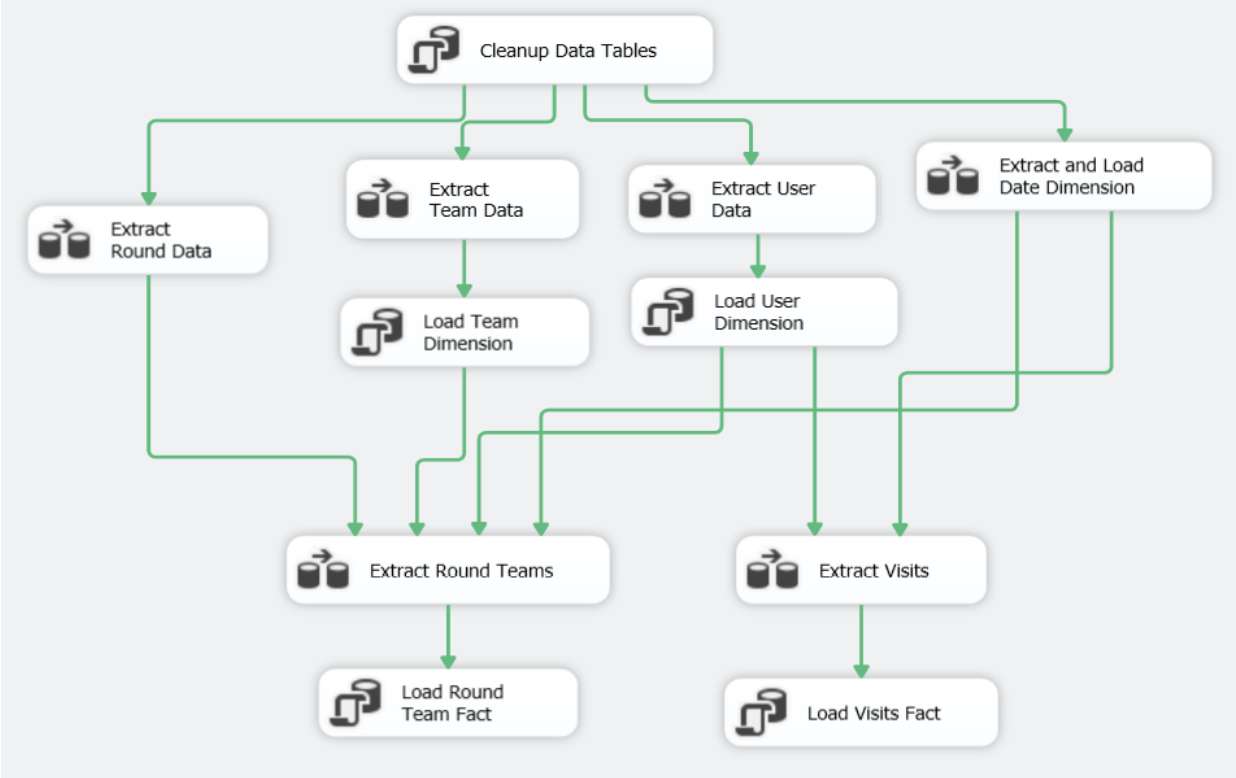
- **Obtenção do ficheiros de datas**
  - Obtido via script python conjugando os dados das rondas, dados inerentes a datas e "enriquecido" com identificação de dias em que foram jogados jogos clássicos com origem na base de dados Sport Radar.
  - Entradas: classicos.csv, rounds.csv e dados de datas
  - Saída: 'dates.txt', com as datas prontas a integrar a Date Dimension.

Os Ficheiros de dados obtidos na etapa de extracção sofreram 2 tipos de tratamentos.

Numa primeira fase, os ficheiros base do sistema ETL foram transferidos para a Data Staging Area manualmente, usando o Flat File Wizard do Sql Server, para a base de dados de staging, acabando por dar origem aos relatórios descritos no ponto 9.

Em alternativa, numa segunda fase e que é descrita a seguir, foram usados os Integration Services da Microsoft para implementar o workflow de carregamento da data staging area e transformação de forma a conseguir criar a base de dados relacional.
Sempre que é executado o processo é feita uma limpeza prévia de todo o sistema de forma a garantir que não haja dados referentes a uma execução anterior.

NOTAS:
Os ficheiros de carregamento da Data Staging Area podem ser encontrados no servidor na pasta: C:\TPD-Source.
Tanto a Data Staging Area como a Data Presentation Area são representadas na mesma base de dados: TPD.SSIS
O esquema do fluxo usado no SSIS é o seguinte:



- **Carregamento da data staging area com dados de utilizadores (Extract User Data):**
  - Entrada: Flat file liga_record_concorrentes.txt, saída: tabela staging_user
  - Entrada: Flat file liga_record_concorrentes_details_history.txt, saída: tabela staging_user_details_history
  - Entrada: Flat file liga_record_concorrentes_address_history.txt, saída: tabela staging_user_address_history
- **Carregamento da data staging area com dados de equipas (Extract Team Data):**
  - Entrada: Flat files: 201415_teams.txt, 201516_teams.txt, 201617_teams.txt , 201718_teams.txt e , 201819_teams.txt
  - Saída: Tabela de SQL staging_team.
- **Carregamento da data staging area com dados de rondas (Extract Round Data):**
  - Entrada: Flat file: rounds.txt
  - Saída: Tabela de SQL staging_rounds.
- **Carregamento da data staging area com dados de visitas (Extract Visits):**
  - Entrada: Flat file: visits.txt
  - Saída: Tabela de SQL staging_visits.
- **Carregamento da data staging area com dados de resultados das equipas por ronda (Extract Round Teams):**
  - Entrada: Flat files: 201415_rounds_teams.txt, 201516_ rounds_teams.txt, 201617_ rounds_teams.txt , 201718_ rounds_teams.txt e , 201819_ rounds_teams.txt
  - Saída: Tabela de SQL staging_team_round.

c. Carregamento

O carregamento do sistema relacional é feito de 3 formas:

- Manual – Season Dimension
- Automática pelo SSIS a partir de um flat file – Date Dimension carregada a partir de dates.txt
- Automática pelo SSIS com base em stored procedures que usam relações e transformações sobre várias tabelas – restantes dimensões e factos

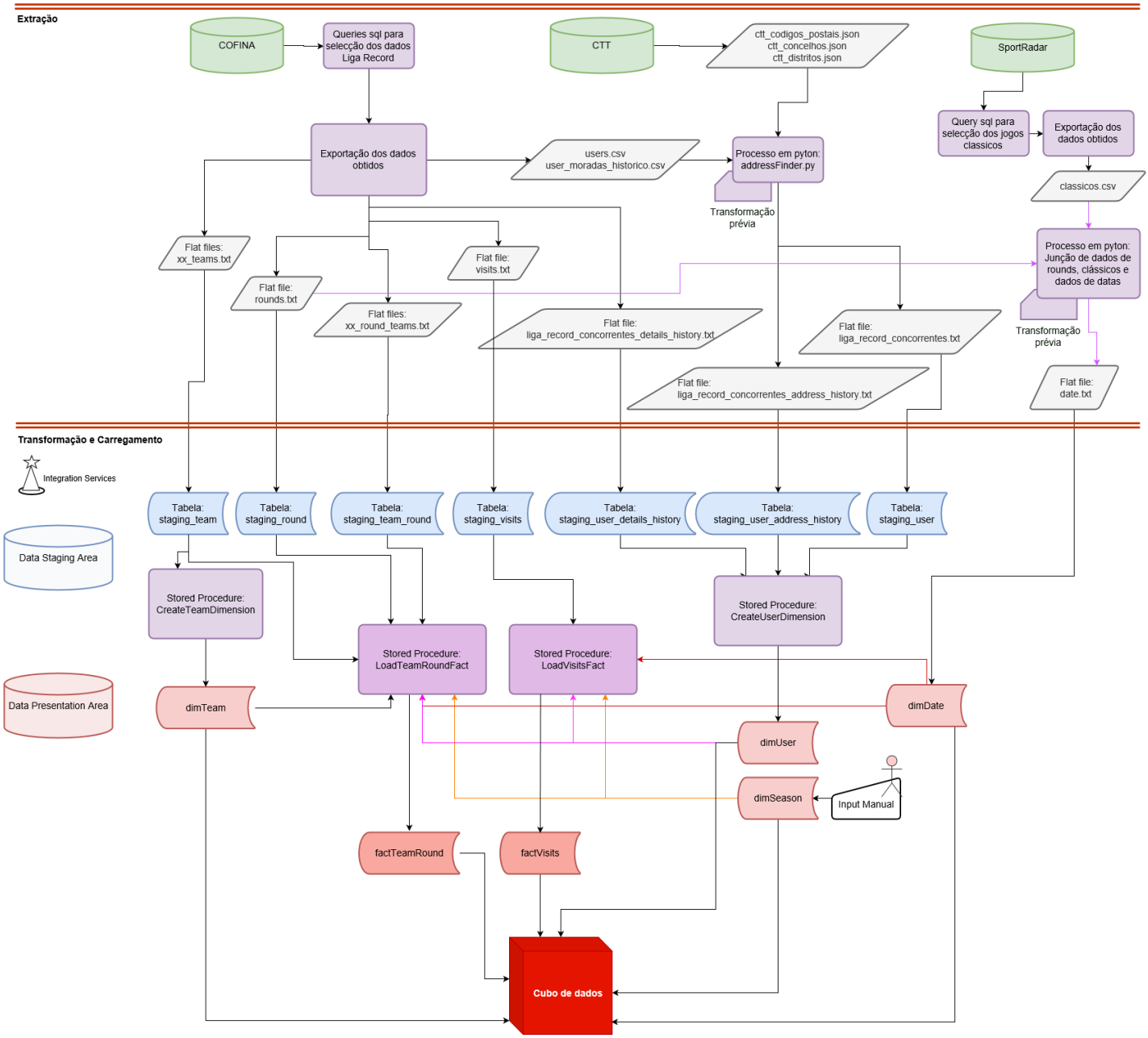
É este último caso que descrevemos a seguir:

- **Carregamento da User Dimension (Load Team Dimension)**
  - Entrada: tabelas staging_user, staging_user_details_history e staging_user_address_history
  - Saída: User Dimension carregada com mudanças lentas integradas
- **Carregamento da Team Dimension**
  - Entrada: tabela staging_team
  - Saída: Team Dimension carregada
- **Carregamento da User Dimension (Load User Dimension)**
  - Entrada: tabelas staging_user, staging_user_details_history e staging_user_address_history
  - Saída: User Dimension carregada com mudanças lentas integradas

- **Carregamento de Visits Fact (Load Visits Fact)**
  - Entrada: tabelas staging_visits, dimDate, dimSeason e dimUser
  - Saída: Visits Fact carregada

- **Carregamento de Round Team Fact (Load Round Team Fact)**
  - Entrada: tabelas staging_team, staging_team_round, staging_round, dimDate, dimSeason, dimTeam e dimUser
  - Saída: Round Team Fact carregada

d. Diagrama de Fluxo do sistema ETL



e. Cubo de dados

O cubo de dados foi implementado usando os Analysis services da Microsoft.

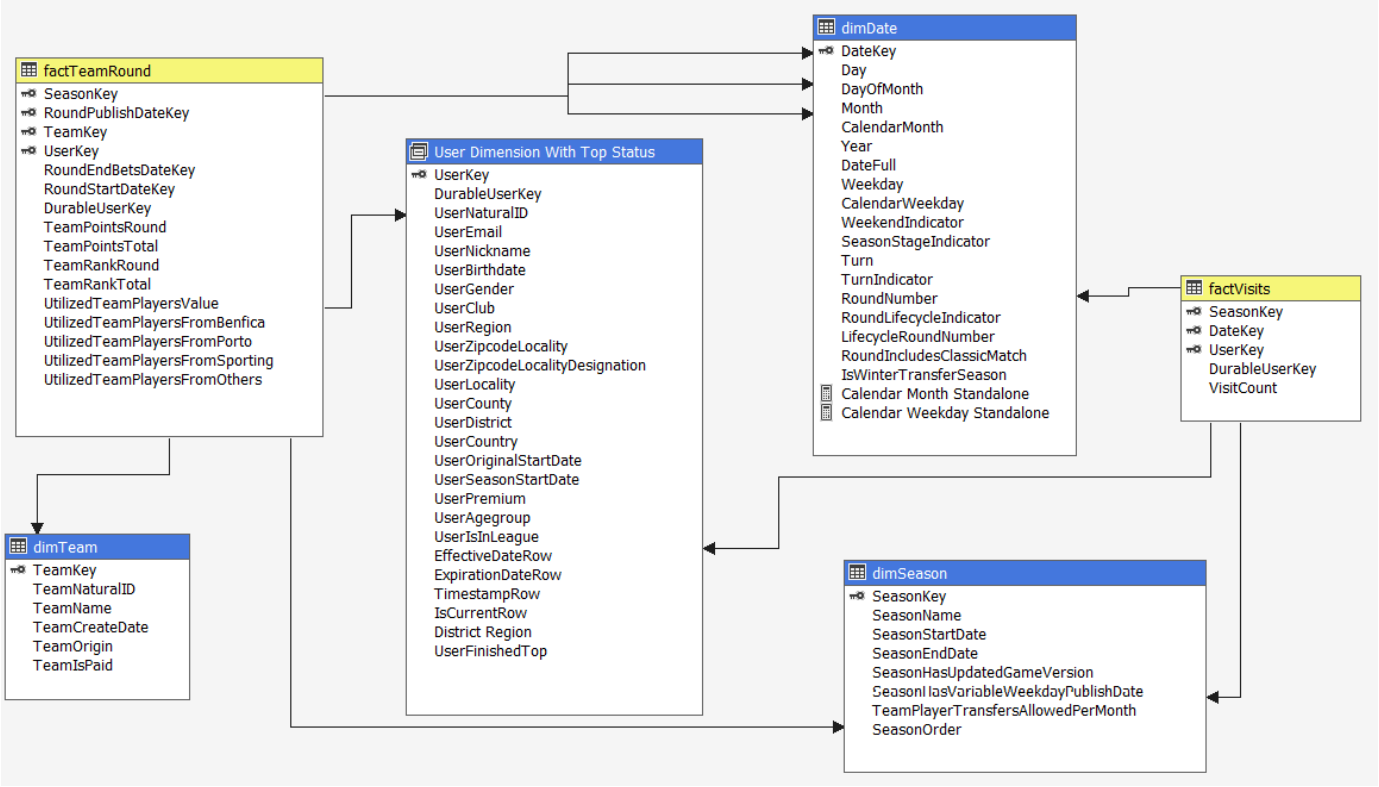
Em baixo pode ser visto o esquema relacional correspondente ao cubo implementado.

No decorrer da análise e para responder às perguntas analíticas, apercebemo-nos que seria necessário fazer comparações de resultados finais do universo dos concorrentes com os que tiveram equipas colocadas entre os melhores resultados.

Numa primeira fase, foi criada uma vista directamente na base de dados relacional sobre a User Dimension que referência exactamente os concorrentes com equipas posicionadas no top 100 em cada final de época. Foi sobre estes dados que foram elaborados os relatórios na secção seguinte representados por pie-charts.

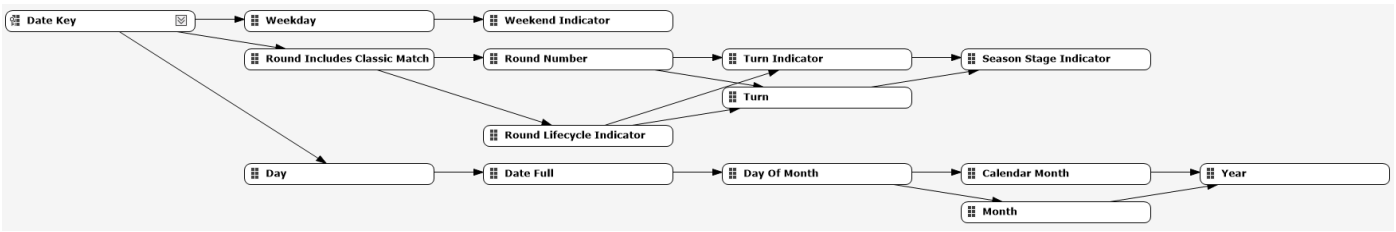
Numa segunda fase, e de forma a permitir aprofundar a análise, e de forma a explorar as capacidades do Analysis Services, foi feito um *replace table with named query* sobre a tabela User Dimension, de forma a esta incluir uma uma nova coluna calculada indicando se o user em causa conseguiu colocar uma equipa no top 100 ou top 1000.
Foi com base nesta nova apresentação que foram elaborados os relatórios para a pergunta analítica 1 apresentados na secção seguinte através de diagramas de cubos.

O esquema final do cubo é o seguinte:

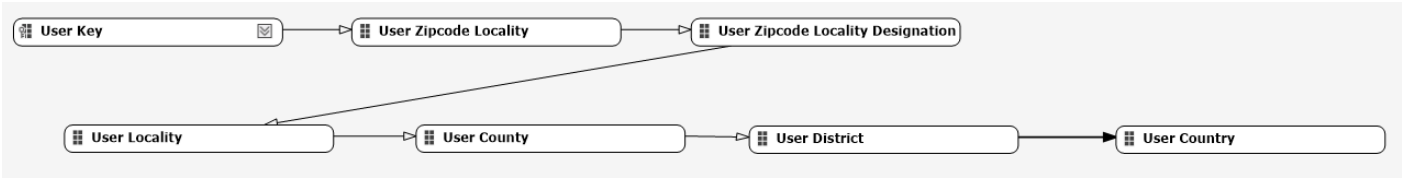


No cubo foram criadas as hierarquias identificadas nas etapas anteriores.

- **Hierarquias implementadas na Date Dimension.**



- **Hierarquias implementadas na User Dimension.**



Para permitir interrogações apenas por certos elementos de hierarquias sem a visão hierárquica, foram criadas named calculations na Date Dimension para:

- Calendar Weekday (Calendar Weekday Standalone), para permitir, por exemplo ver a quantidade de visitas por dia da semana sem identificação se é ou não fim de semana.
- Calendar Month (Calendar Month Standalone), para permitir, por exemplo, ver a quantidade de visitas por mês sem identificação do ano.

Foi criada uma named calculation adicional na User Dimension (District Region) para conseguir identificar a que zona do país pertencem os concorrentes (sul, centro, norte ou ilhas).

7. Relatórios obtidos com base no cubo de dados obtido através da importação via SSIS.

Nesta secção pretendemos responder às três perguntas analíticas descritas para este projeto, com base na produção de relatórios usando a ferramenta Power BI Desktop e tendo como fonte o cubo de dados criado com base na base de dados relacional criada via SSIS (base de dados com o nome TPD.SSIS).

Para responder às várias perguntas, vamos analisar alguns gráficos obtidos no projecto Cubo.pbix localizado na desktop do servidor.

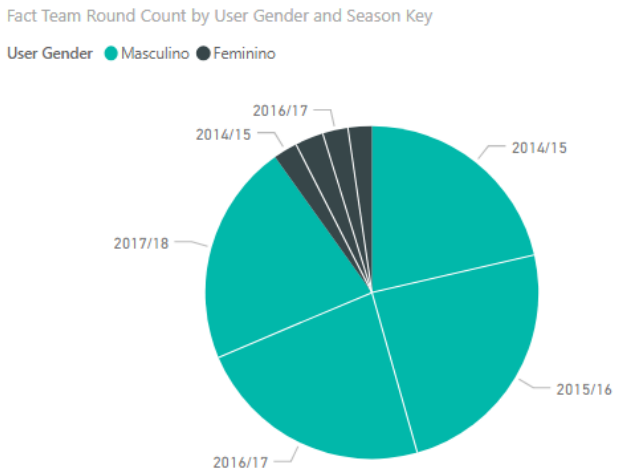
a. Existe algum padrão entre as preferências/dados dos concorrentes e o sucesso final das suas equipas?

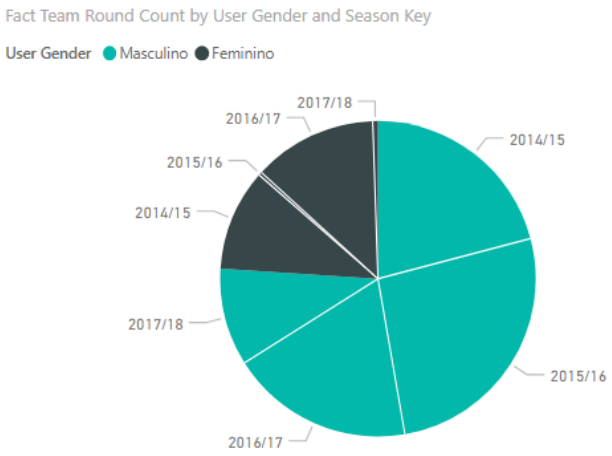
Nos gráficos que se seguem conseguimos comparar a percentagem de cada segmento no universo de concorrentes e a percentagem de cada segmento entre os concorrentes que posicionaram pelo menos uma equipa no top 100. Devido a esta comparação, só são apresentados 4 anos, uma vez que representam os anos para os quais temos resultados finais.

No caso dos gráficos em pie-chart, obtidos na primeira iteração do cubo com uso de uma view sobre a bd relacional, o gráfico da esquerda representa a percentagem dos segmentos na população geral enquanto que o gráfico da direita representa o top 100.

No caso dos gráficos de barras, obtidos com acesso ao cubo final, no gráfico aparecem os resultados gerais, no segundo o top 100 e no terceiro o top 1000.

Comparação de resultados por género

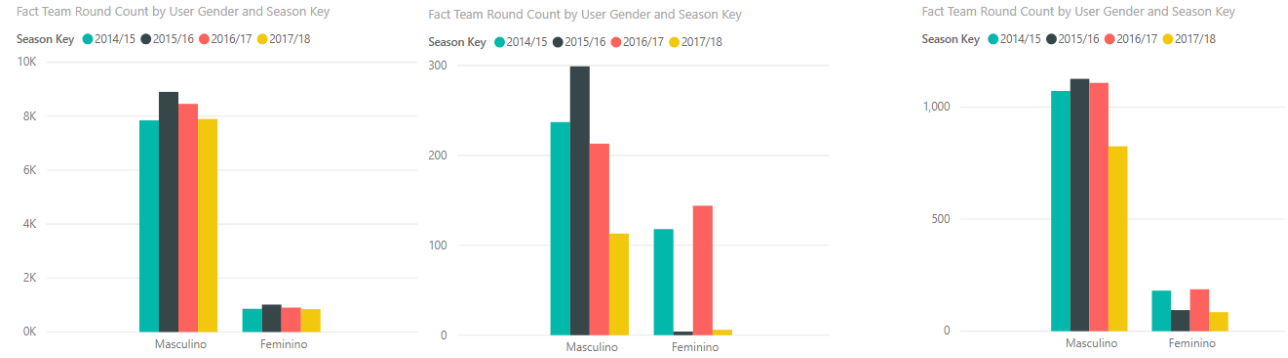




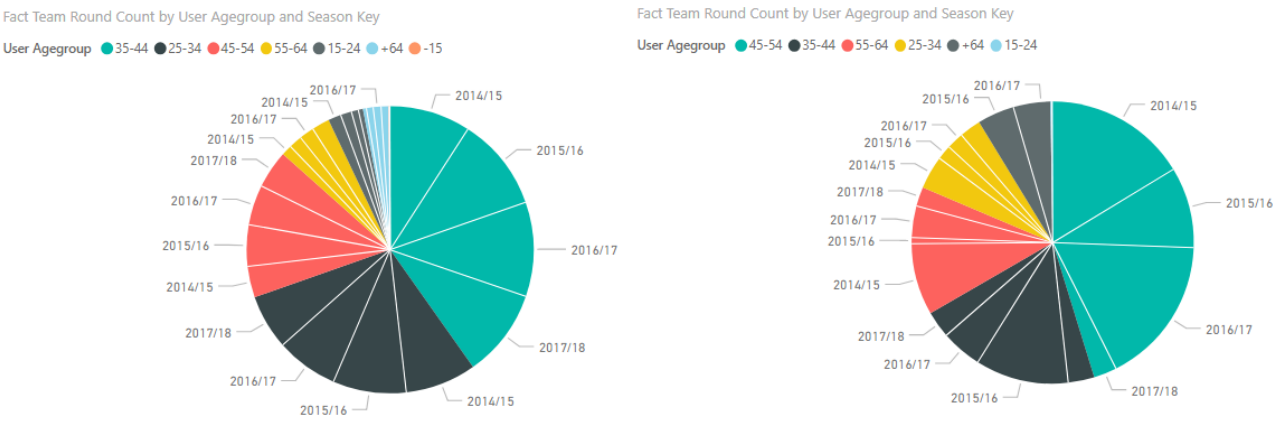
No gráfico que compara resultados por género não conseguimos tirar qualquer conclusão. Tivémos 2 temporadas de grande sucesso para as mulheres (2014/15 e 2016/17) e outros 2 de baixo sucesso.

Estes dados comprovam-se para os dados obtidos com a segmentação entre concorrentes com equipas dentro dos 100 1ºs lugares e 1000 1ºos lugares.

Estes dados também mostram um dado interessante: o top-100 da barra amarela que representa o ano de 2017/18 é muito curto – isto significa que, neste ano, os concorrentes que correspondem ao nosso dataset tiveram resultados inferiores ao normal relativamente a todos os outros que jogaram na Liga Record.



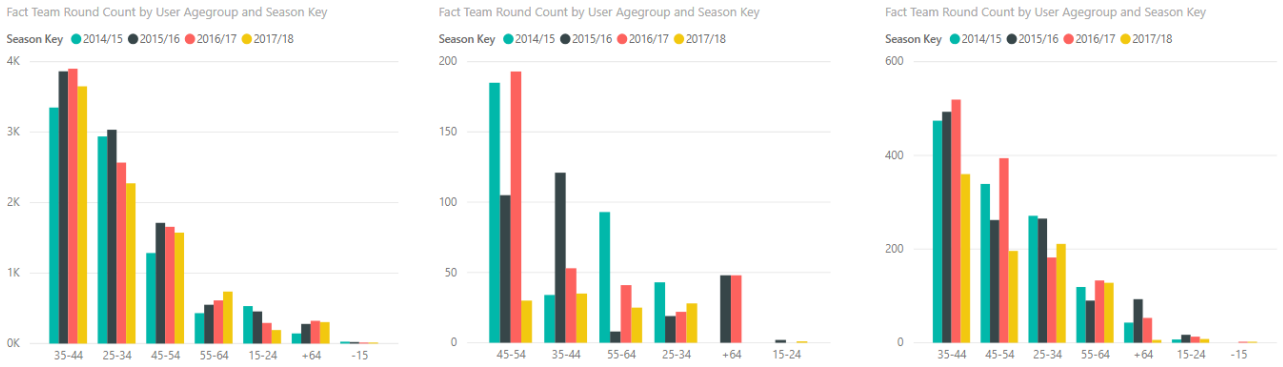
Comparação de resultados por faixa etária



Neste caso conseguimos verificar que enquanto a maioria dos concorrentes estão na faixa etária dos 35-44 anos, são os de 45-54 que apresentam os melhores resultados.

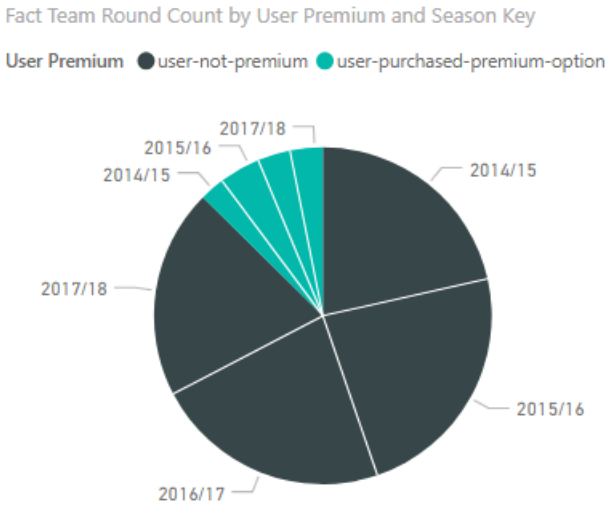
O segundo segmento mais participativo (25-34 anos) também tem poucos resultados situados no top 100.

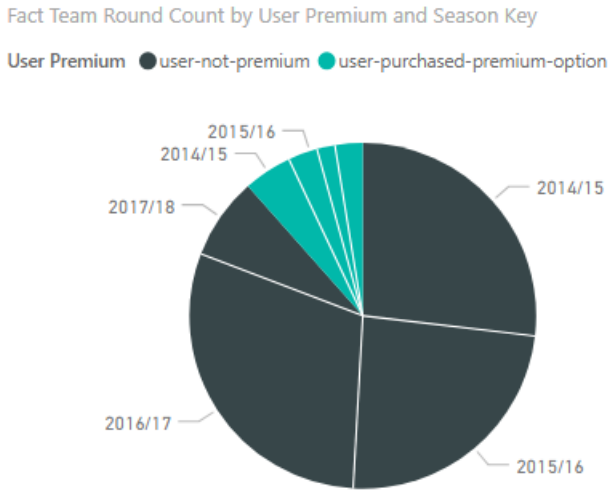
Assim sendo, podemos indicar que sim, a idade dos concorrentes influencia o seu sucesso no jogo.



No gráfico do meio conseguimos ver os concorrentes que posicionaram equipas no top 100. Os valores mais alto pertencem ao segmento 45-54 tal como nos aparecia na 1ª análise. Parece-nos também relevante a boa classificação do segmento dos + de 64 anos nas épocas 2015/16 e 2016/17. A tendência confirma a conclusão do 1º gráfico: de facto, a faixa etária é um dado influenciador do resultado final.

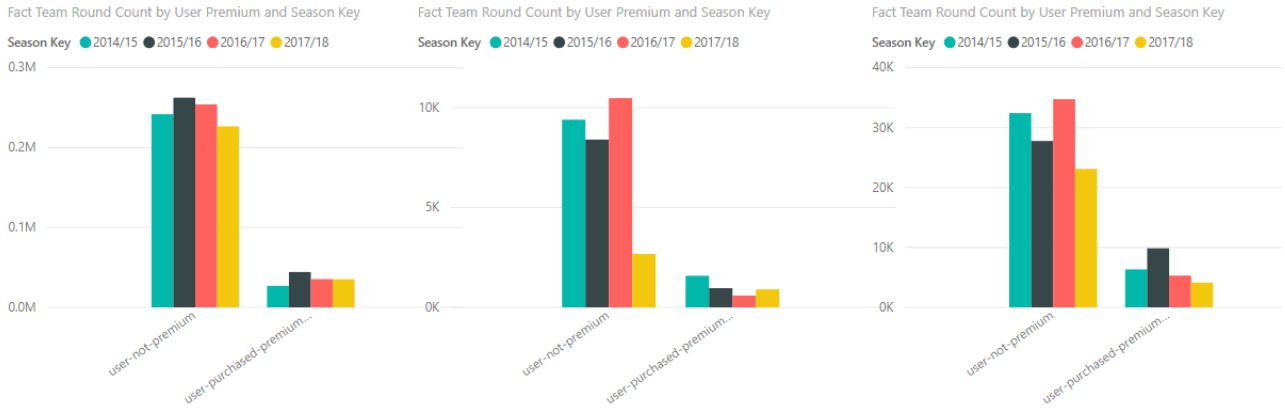
Comparação de resultados de assinantes premium



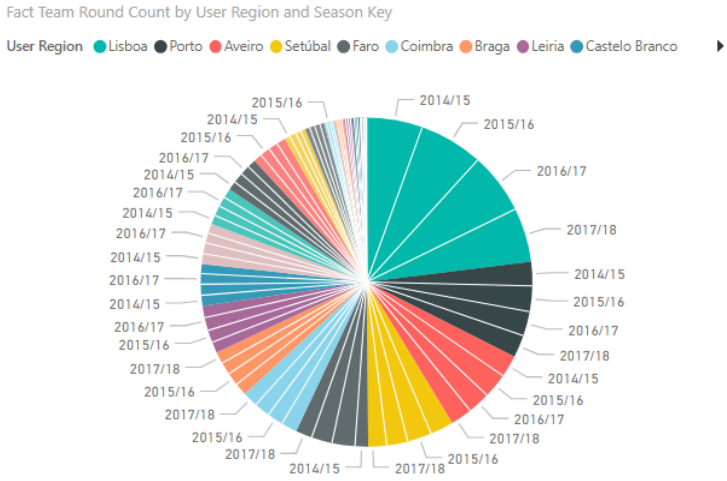


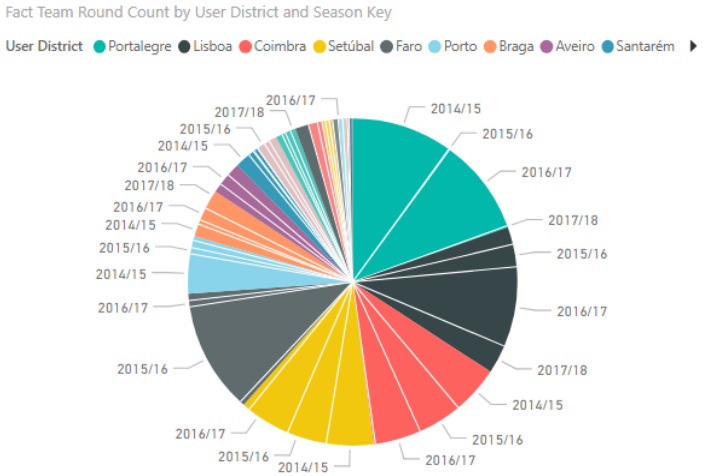
Pelos gráficos muito semelhantes, é possível perceber os assinantes premium, apesar de receberem mais informação que os restantes concorrentes, não tiram partido da mesma para obterem resultados significativos.

Pelos gráficos apresentados de seguida, que permitem a comparação entre os dados gerais, top-100 e top-1000, podemos concluir que a subscrição premium ainda teve alguma influência na obtenção de resultados no top-1000 na época 2015/16, mas em média não é de todo determinante para obter bons resultados.



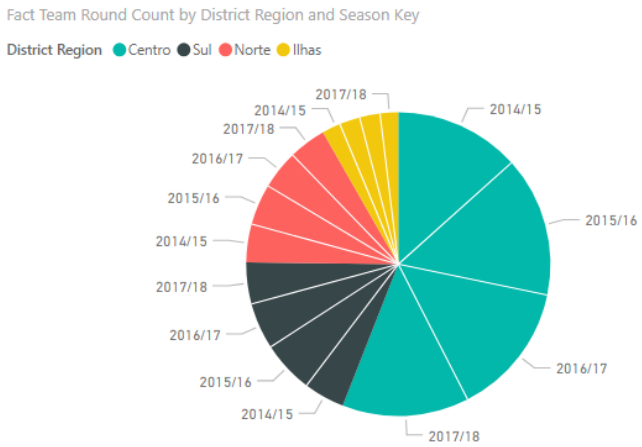
Comparação de resultados por região do país

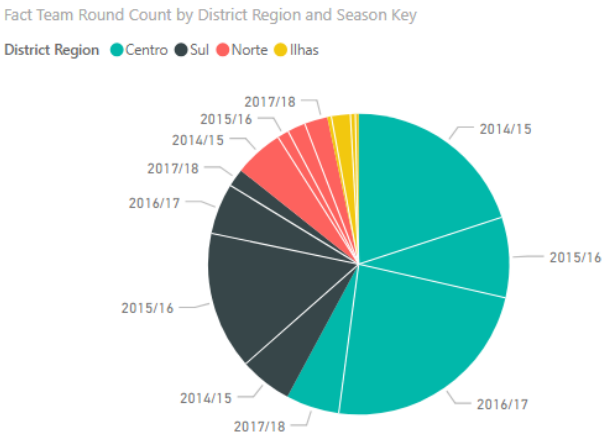




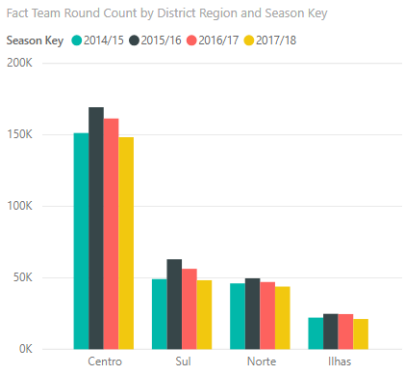
Pelos gráficos conseguimos verificar que existe uma distribuição de participação por distrito ao longo dos vários anos, sendo Lisboa o distrito com mais participantes, mas que não existem tendências de sucesso nos resultados por distrito ao longo das várias temporadas: existem temporadas de grande sucesso de alguns distritos, como Portalegre e faro, mas outros em que estes não têm resultados representativos.

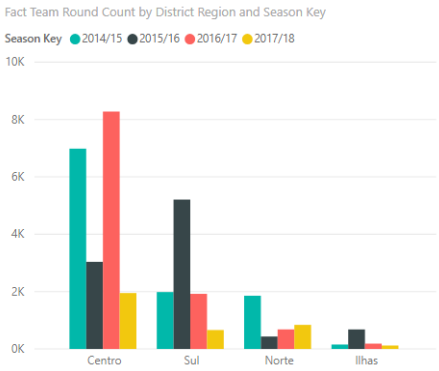
O mesmo nível de resultados conseguem ser vistos quando obtemos os gráficos por região do país. A distribuição mantém-se constante por região ao longo dos anos, mas os resultados alteram-se bastante de ano para ano. De notar aqui os óptimos resultados obtidos pelo Sul do país na temporada 2015/16.

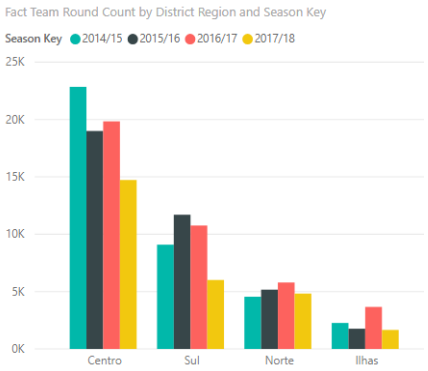




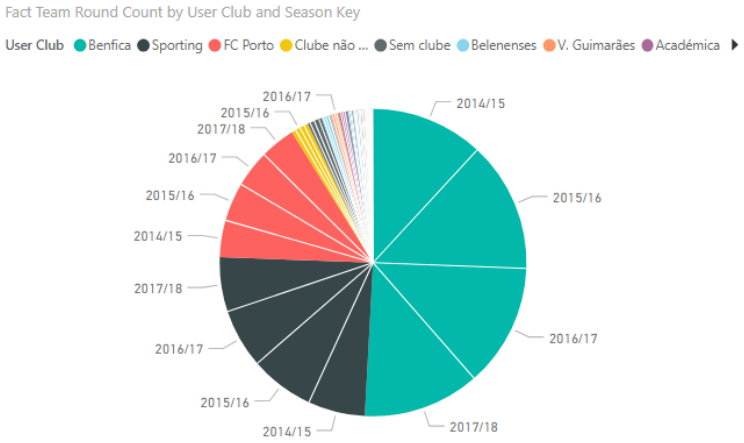
Quando a avaliação é feita nos 3 segmentos, os resultados confirmam-se – houve de facto grandes resultados obtidos pelo sul do país na época 2015/16 principalmente à custa do centro. Fora esta "anormalidade" a distribuição dos concorrentes melhor colocados segue em linha com a distribuição geral.

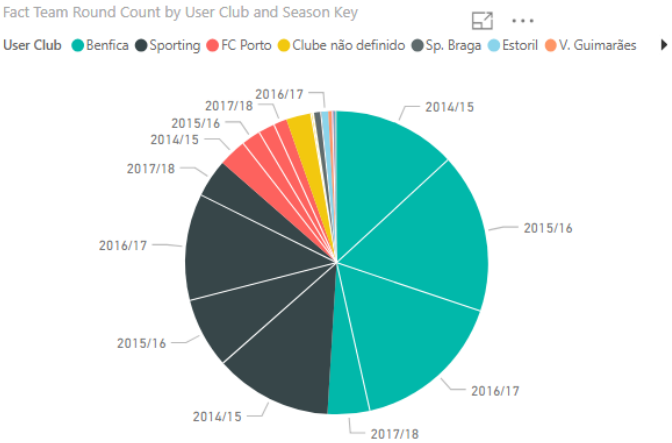






Comparação de resultados por clube

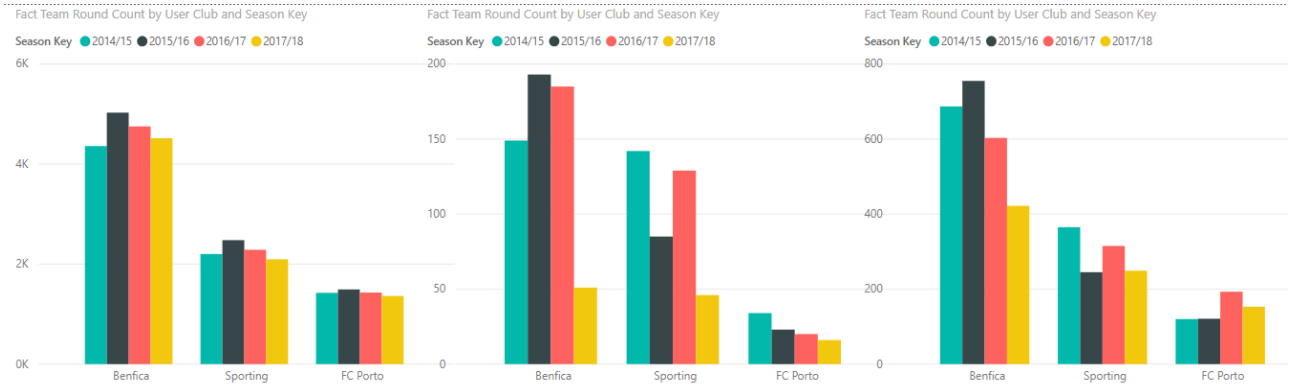




Em termos de clubes, a maioria dos participantes é do Benfica. Neste caso, a sua classificação parece ter sido afectada no ano em que o clube perdeu o campeonato para o Futebol Clube do Porto (2017/18).

Este aspecto não é verificado quando avaliamos as classificações dos participantes de outros clubes. Por exemplo, os participantes do Sporting aparecem consistentemente com bons resultados apesar do percurso problemático do seu clube.

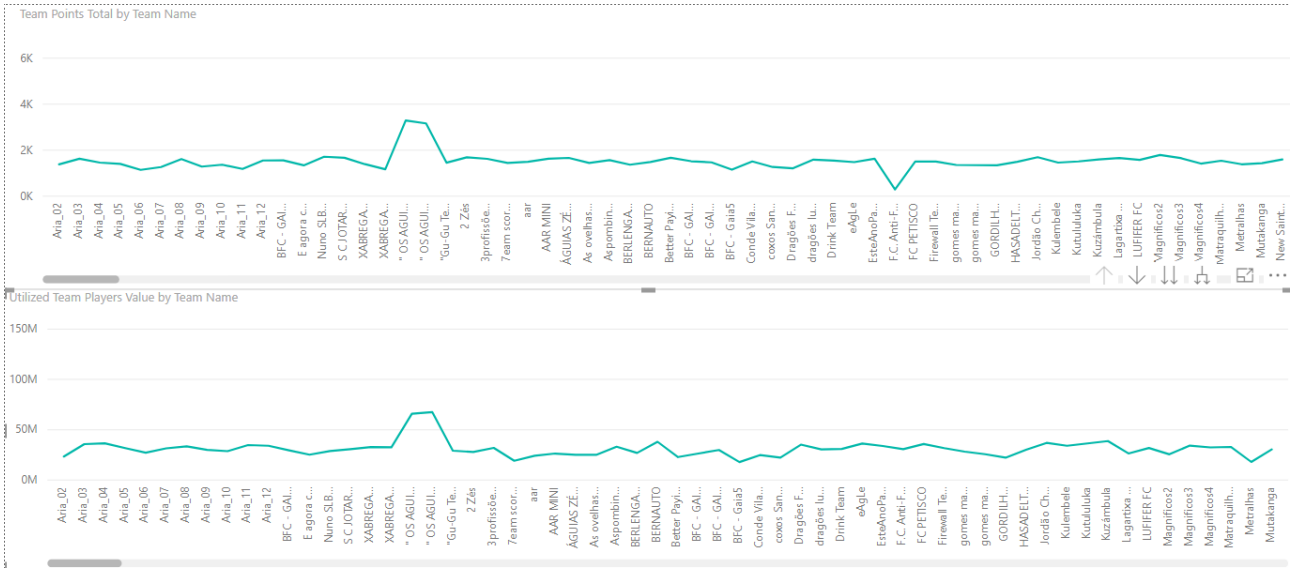
No caso da variação por clubes, nos 3 segmentos do top, optámos por nos focar mais nos 3 clubes mais representativos: Benfica, FC Porto e Sporting.
Confirmando os dados obtidos na avaliação acima, conseguimos ver que a maioria dos jogadores é do Benfica, seguido do Sporting. Os concorrentes do Sporting são quem consegue colocar equipas nas melhores posições.



b. O valor combinado dos 11 jogadores usados em campo influencia o sucesso final da equipa?

Nos gráficos a seguir, é vista a variação de valores de equipa e pontuação total no final da temporada numa amostra de equipas. Nestes casos, parece haver alguma convergência entre os valores e as pontuações.

Esta convergência é melhor verificada na análise feita ao sistema relacional apresentada no próximo capítulo.



c. A participação dos concorrentes varia ao longo da temporada?

De seguida é apresentada a variação de visitas dos concorrentes por mês e por dia da semana.

Como seria expectável, o mês mais participado é o de setembro (quando as equipas começam a pontuar), logo seguido de fevereiro, altura do mercado de inverno, que é quando os concorrentes podem trocar 6 jogadores dos seus plantéis.

Em termos de dias da semana, o dia em que usualmente terminam as apostas (sexta-feira) é o mais participado em todos os meses.

É de notar que em dezembro, os números da participação no dia de publicação de resultados (terça-feira) tem quase tantas visitas como o dia de fim de apostas.

8. Relatórios com base no esquema relacional obtido através de importação manual dos dados.

Nesta secção pretendemos responder às três perguntas analíticas descritas para este projeto, com base na produção de relatórios usando a ferramenta Power BI Desktop e tendo como fonte a base de dados relacional inicialmente criada (base de dados com o nome TPD).

Há que referenciar que há uma diferença nestes relatórios relativamente a visitas: apenas foram consideradas as visitas feitas pelos concorrentes no decorrer do jogo, enquanto que no ponto anterior a análise foi feita às visitas feitas ao longo de todo o ano.

1. **Existe algum padrão entre as preferências/dados dos concorrentes e o sucesso final das suas equipas?**

Para responder a esta pergunta iremos analisar diversos fatores que estão diretamente relacionados com os concorrentes, tais como: faixa etária, género, região de residência, clube preferido e se subscreveram ao serviço premium.

Neste caso, ao contrário do método usado no capítulo anterior, não usámos a vista dos concorrentes com equipas no Top 100 e fomos antes avaliar as pontuações das equipas.

Os resultados diferentes que aparecem podem ser explicados pelas diferenças de método de avaliação, sendo que o tratado no capítulo 7 será mais rigoroso.

Visto que cada concorrente pode ter mais do que uma equipa, usámos como medida a média das pontuações totais das equipas para a última ronda para cada temporada (que corresponde à pontuação final nessa temporada) para quantificar o sucesso dos concorrentes no jogo. Para obter o relatório desta análise cruzámos estes valores com cada um dos fatores em estudo. Para construir o sistema ETL, extraímos os dados relativos às últimas cinco temporadas, incluindo a 2018/19. Dado que a temporada 2018/19 ainda se encontra a decorrer, e como tal não existe a pontuação total para a última ronda, para esta análise apenas usámos os dados para as quatro temporadas anteriores.

- Existe algum padrão para a faixa etária dos concorrentes?

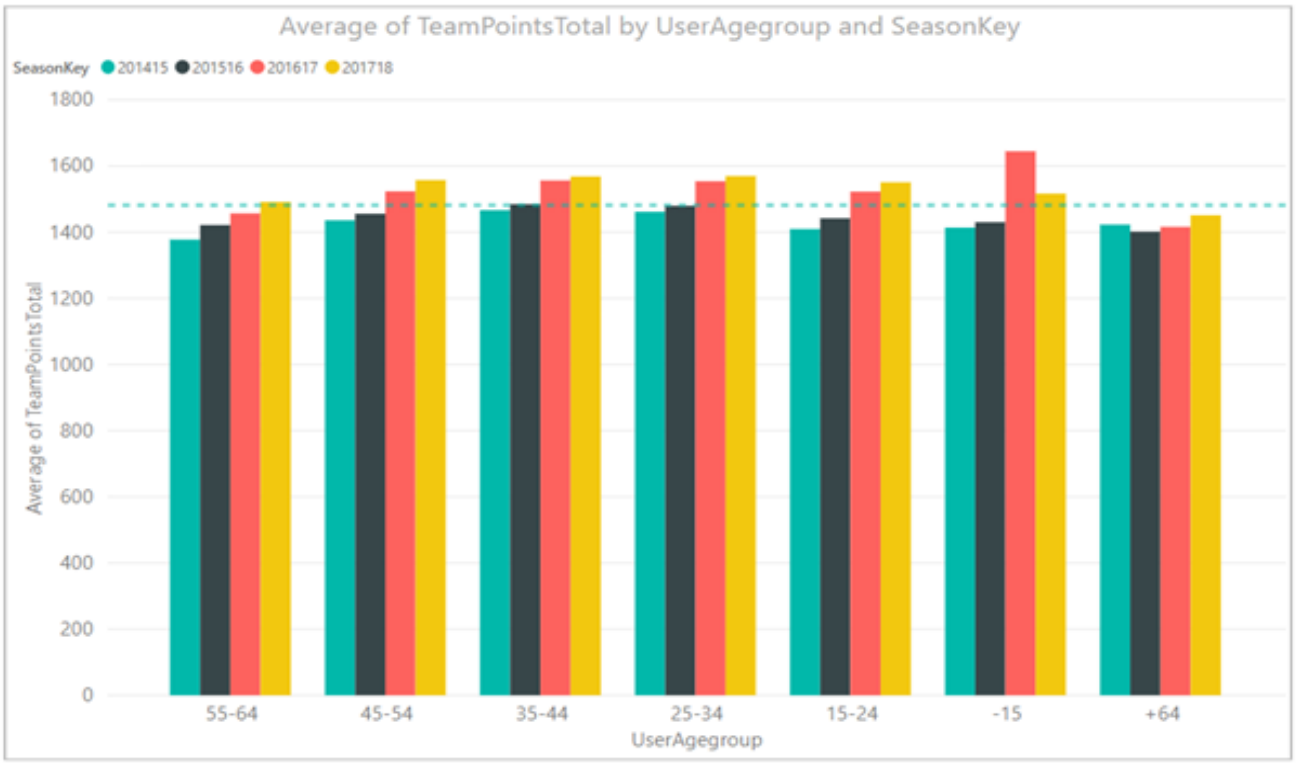


Gráfico 1 – Média das pontuações totais das equipas (filtrada apenas para a última ronda) por faixa etária e por temporada.

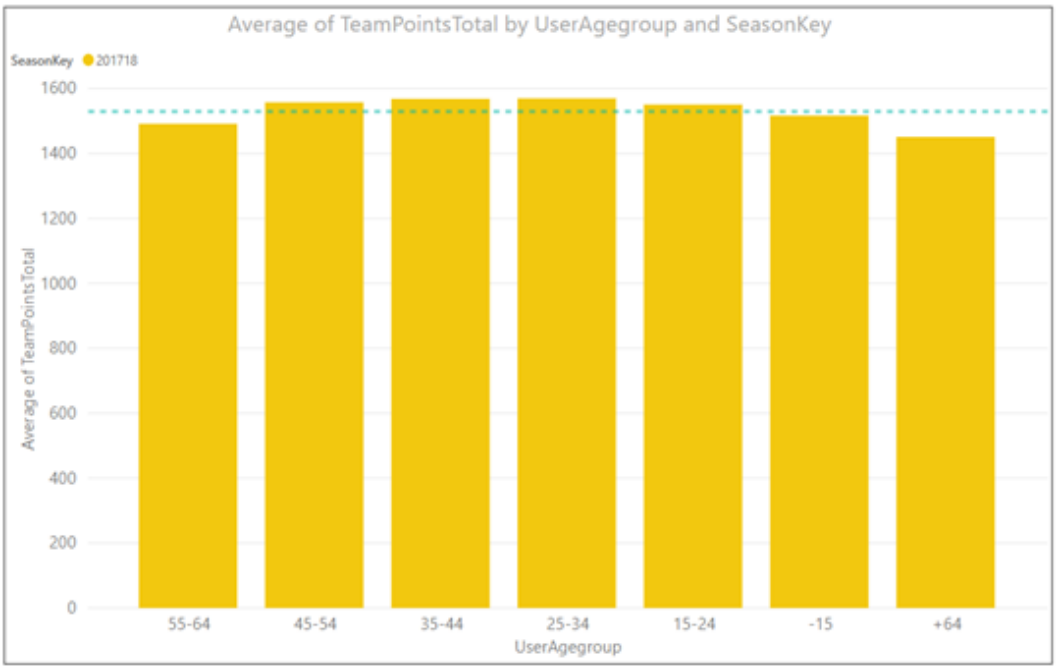


Gráfico 2– Média das pontuações totais das equipas por faixa etária na temporada de 2017/2018.

Como observado nos gráficos obtidos, analisando a variação dentro de cada temporada (como é mais perceptível no gráfico 2), não existe grande diferença nas médias das pontuações finais entre as várias faixas etárias. Verificamos apenas uma pequena tendência para serem mais elevadas nos grupos intermédios, “25-34” e “35-44”, porém muito pouco percetível e cuja significância teria de

ser analisada estatisticamente. Podemos detectar no gráfico 1 que existe alguma tendência para a média das pontuações totais aumentarem ao longo das temporadas (à exceção dos grupos "-15" e "+64"). Portanto não podemos concluir que existe relação entre a faixa etária dos concorrentes e o seu sucesso no jogo.

- Existe algum padrão para o género dos concorrentes?

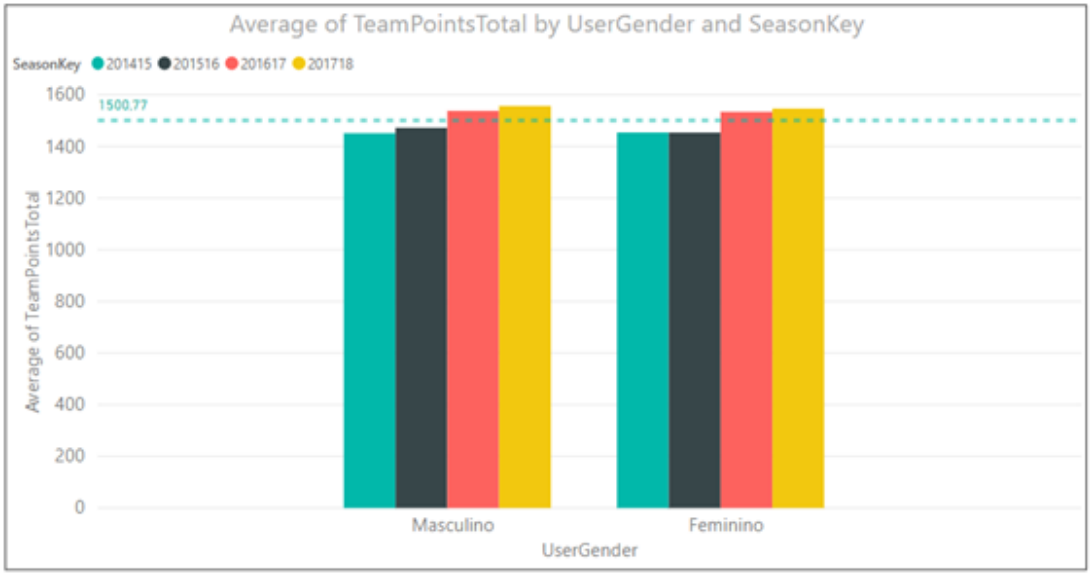


Gráfico 3 – Média das pontuações totais das equipas por género e por temporada.

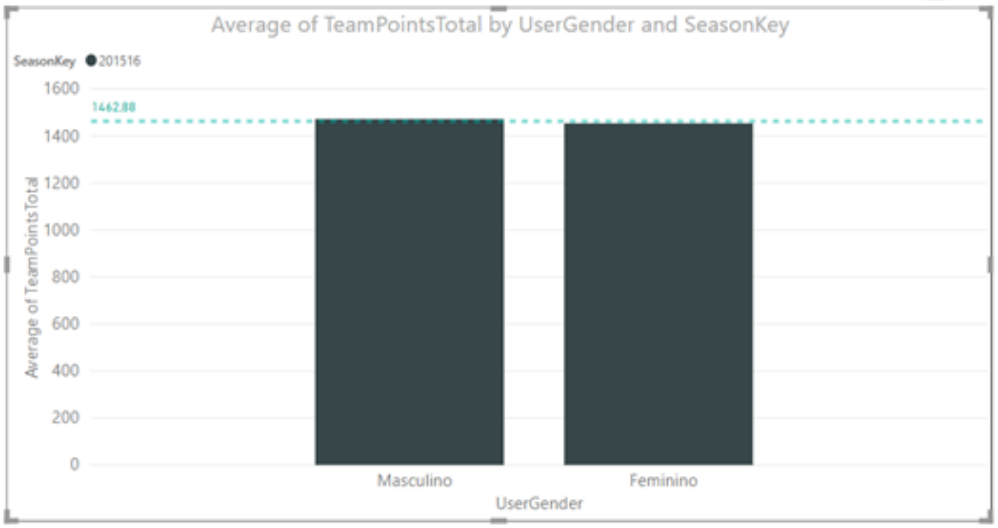


Gráfico 4 – Média das pontuações totais das equipas por género na temporada 2015/16.

À semelhança do caso anterior, analisando o gráfico 3 e 4 não existe diferença significativa entre a média das pontuações totais dos concorrentes do sexo masculino e os concorrentes do sexo feminino. Portanto não existe qualquer evidência para afirmar que o género dos concorrentes influencia o sucesso do concorrente no jogo. Este resultado foi imprevisível, visto que, tal como demonstrámos nas análises preliminares aos dados brutos, existe uma muito maior proporção de concorrentes do sexo masculino registados, o que aumentaria a possibilidade da existência de concorrentes deste género a destacarem-se.

Curiosamente mais uma vez verificamos a existência da tendência para as médias das pontuações totais aumentarem da temporada mais antiga para a temporada mais recente.

- Existe algum padrão quando relacionamos a região de residência dos concorrentes?

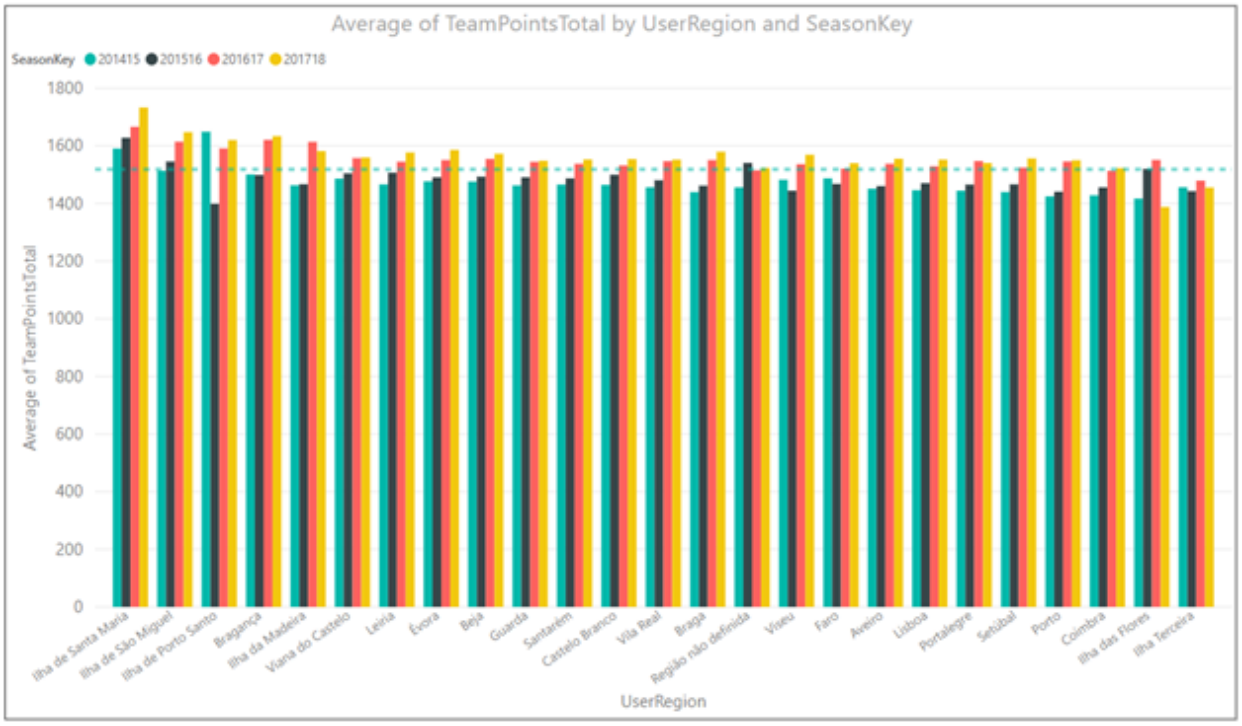


Gráfico 5 – Média das pontuações totais das equipas por região de residência e por temporada.

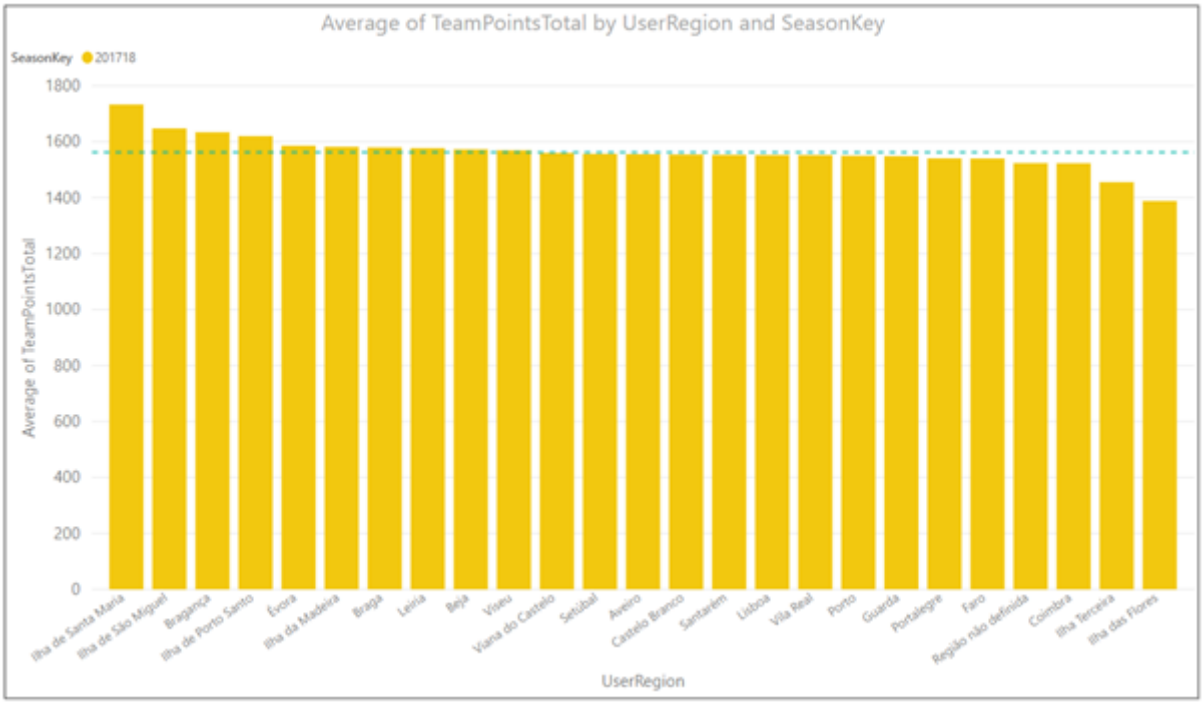


Gráfico 6 – Média das pontuações totais das equipas por região de residência na temporada 2917/18.

Como observado nos gráficos 5 e 6 a variação das médias para as pontuações totais é muito ligeira entre as várias regiões de residência dos concorrentes. Curiosamente as ilhas encontram-se nos extremos do ranking, com a Ilha de Santa Maria e a Ilha de São Miguel a demonstrarem as

melhores pontuações e a Ilha Terceira e a Ilha das Flores as piores pontuações. É interessante relembrar que, na análise preliminar aos dados brutos, as Ilhas mostraram possuir o menor número de concorrentes registados no jogo, provavelmente como resultado da sua baixa densidade populacional. À exceção das ilhas, as restantes regiões de residência dos concorrentes não demonstraram estar relacionadas com o seu sucesso no jogo.

- Existe algum padrão para a subscrição no serviço premium?

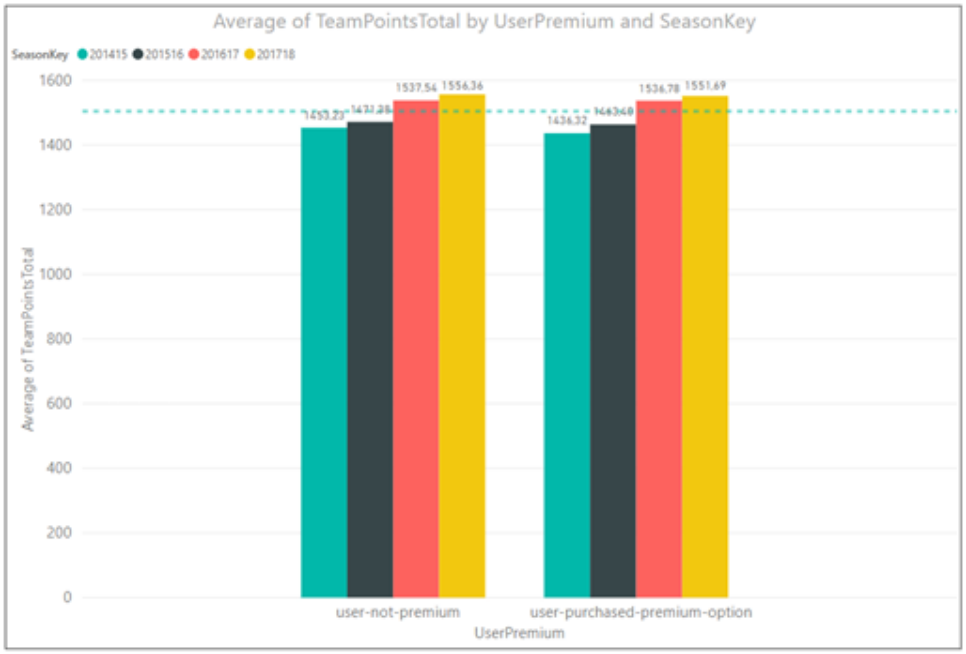


Gráfico 7 – Média das pontuações totais das equipas para os grupos de concorrentes que subscreveram ao serviço premium versus os que não subscreveram por temporada.

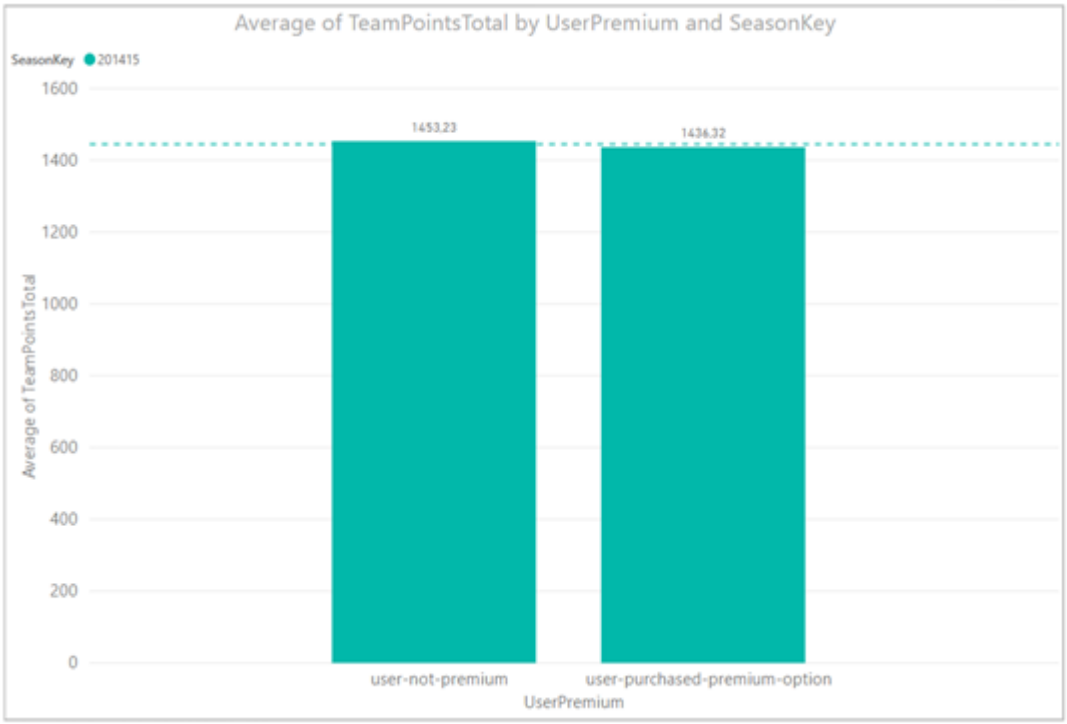


Gráfico 8 – Média das pontuações totais das equipas para os grupos de concorrentes que subscreveram ao serviço premium versus os que não subscreveram para a temporada 2014/15.

Na medida em que os concorrentes que subscrevem ao serviço premium da Liga Record têm acesso a dados úteis para a sua estratégia de jogo, é expectável que o grupo de concorrentes que

adquirem o serviço demonstre maior sucesso. No entanto, na nossa análise (gráfico 7 e 8) verificámos que as médias das pontuações totais não variam entre o grupo de concorrentes que subscreveram e o grupo de concorrentes que não subscreveram. Portanto não existe qualquer evidência em que a subscrição ao serviço premium influencie o sucesso final no jogo.

Os resultados que obtivemos para as subscrições e para o género dos concorrentes foram inesperados, e como tal foram validados através de interrogações SQL feitas diretamente nas bases de dados do sistema operacional fonte.

- Existe algum padrão para o clube de preferência dos concorrentes?

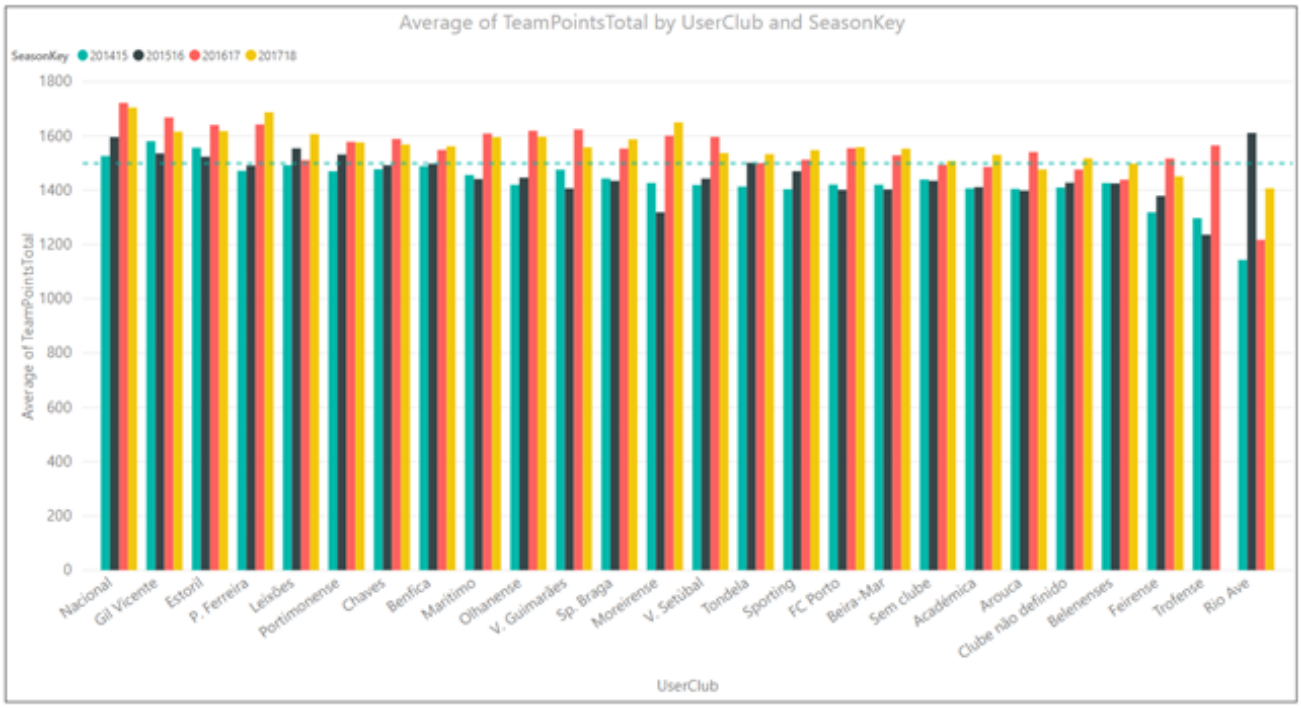


Gráfico 9 – Média das pontuações totais das equipas para os clubes de preferência por temporada.

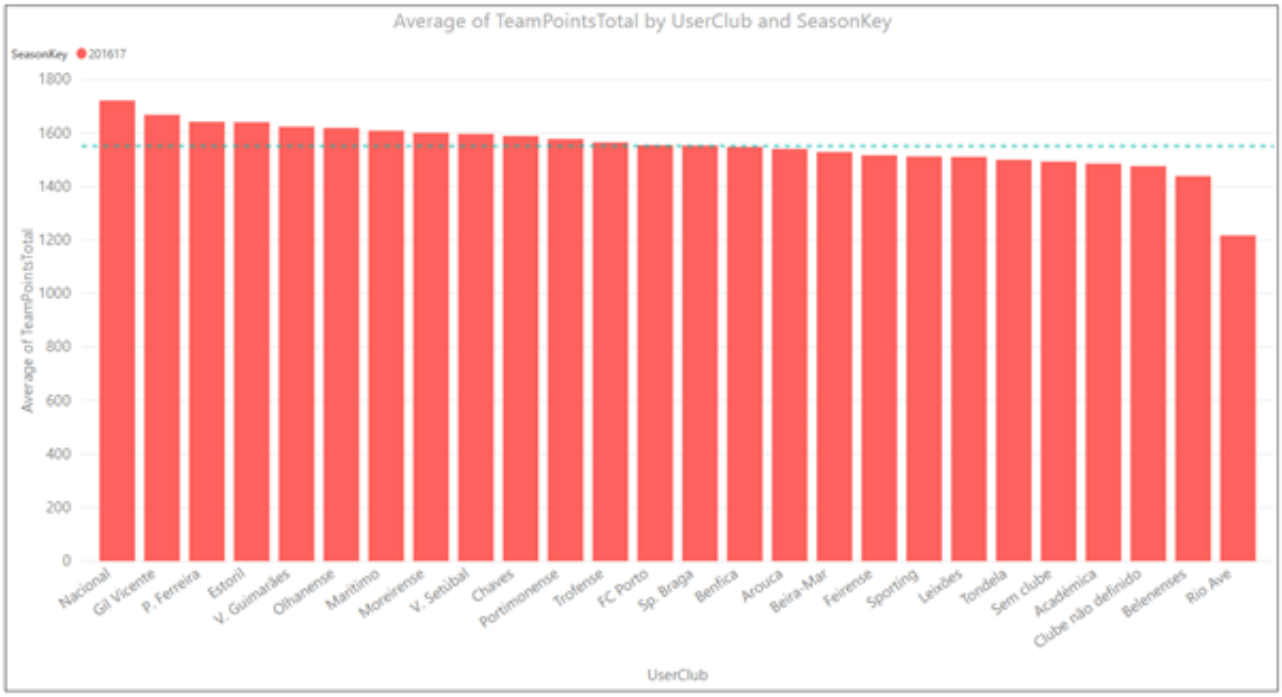


Gráfico 10 – Média das pontuações totais das equipas para os clubes de preferência na temporada 2016/17.

Por último, quando obtivemos as médias das pontuações totais para cada clube de preferência dos concorrentes não obtivemos uma variação significativa entre os grupos, com a grande maioria dos clubes a posicionar-se muito próximo da média total das pontuações. A preferência pelo Nacional lidera o ranking, enquanto que a preferência pelo Rio Ave demonstrou uma maior variância de médias entre as temporadas em comparação com os outros clubes, e registando os piores resultados.

2. O valor combinado dos 11 jogadores usados em campo influencia o sucesso final da equipa?

Para responder a esta questão, fomos verificar se existe correlação entre o valor combinado dos 11 jogadores usados em campo e o resultado final da equipa nessa ronda.

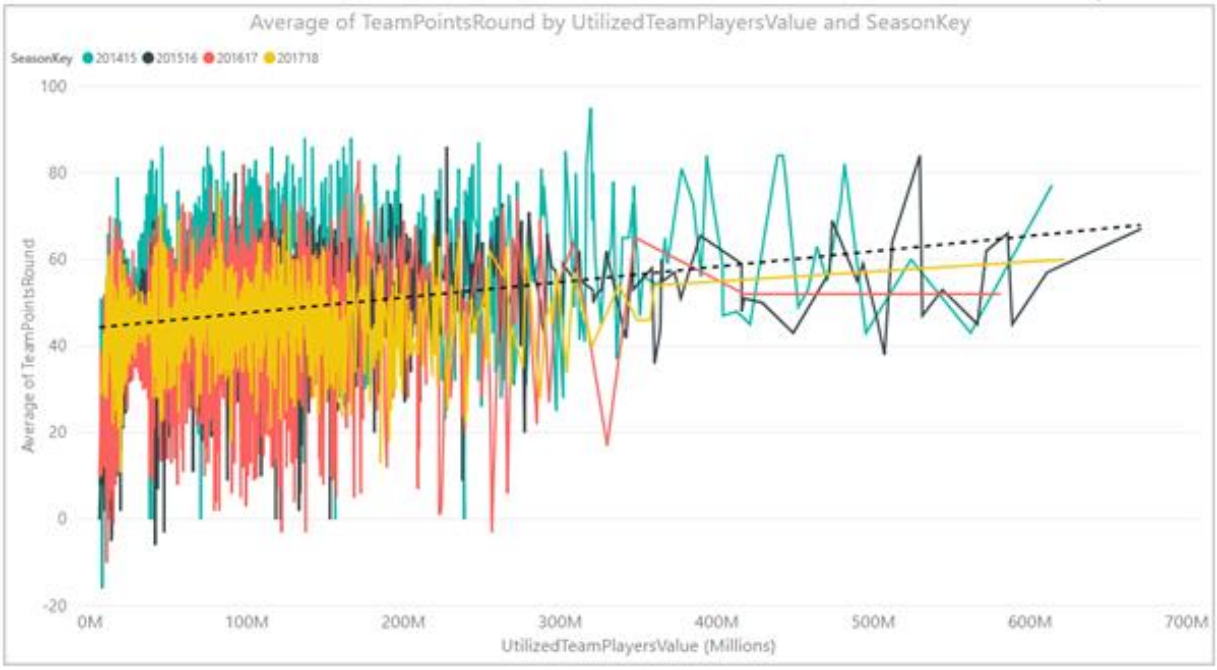


Gráfico 11 – Média dos pontos das equipas por ronda por valor dos jogadores em campo utilizados, por temporada.

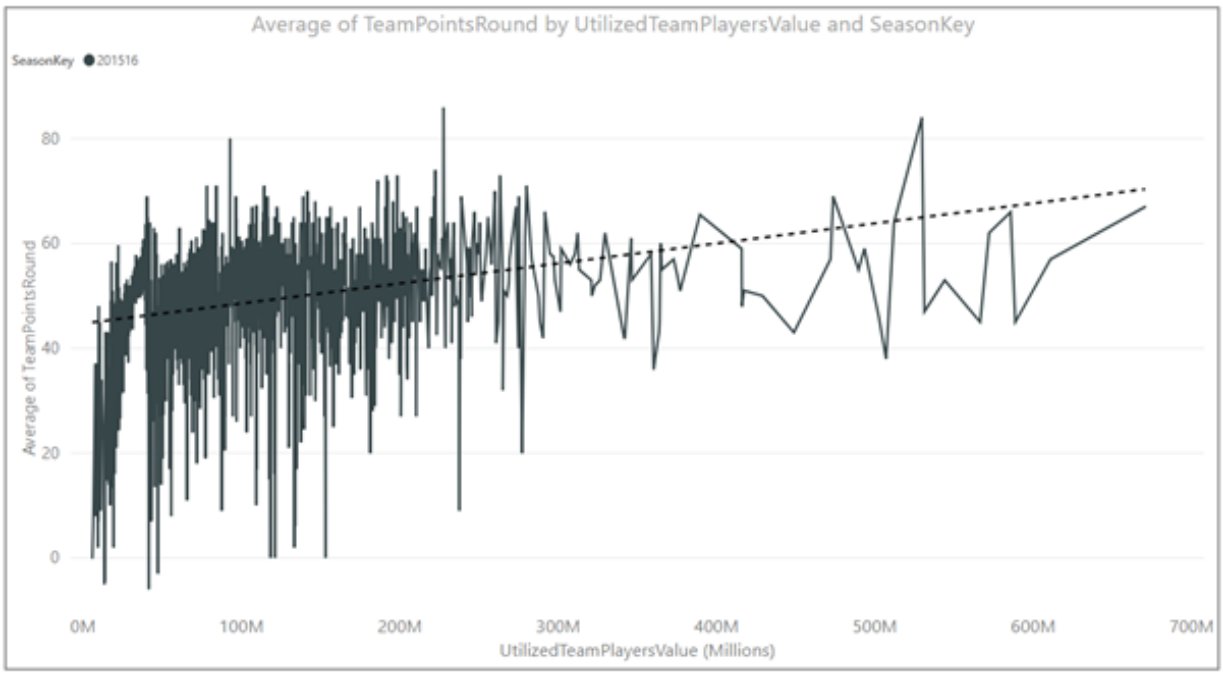


Gráfico 12 – Média dos pontos das equipas por ronda por valor dos jogadores em campo utilizados, na temporada 2015/16.

Apesar da grande variação observada, vemos que tendencialmente quanto maior o valor combinado dos 11 jogadores usados em campo, maior é a média dos pontos adquiridos nessa

ronda, tal como previsto. Portanto podemos afirmar que o valor combinado dos 11 jogadores usado em campo pode realmente influenciar o sucesso final da equipa.

Também podemos observar que, tal como esperado existe uma maior concentração de dados para valores de utilizedTeamPlayersValue baixos, do que para valores de utilizedTeamPlayersValue altos.

3. A participação dos concorrentes varia ao longo da temporada?

Uma das perguntas analíticas que pretendemos ver respondida no presente projeto, é verificar se a participação dos concorrentes é uniforme ou se varia ao longo da temporada.

O número de logins é a medida utilizada neste projeto para quantificar a participação activa dos concorrentes no jogo, o que por sua vez, permite-nos analisar o interesse dos mesmos pelo jogo.

Para esta análise usámos os dados combinados das cinco temporadas em estudo. Uma vez que, apenas é importante perceber o que ocorre durante a temporada em si e não entre temporadas, podemos usar cada temporada como um replicado e aumentar a confiança nos resultados obtidos.

Para iniciar obtivemos um relatório com o somatório de logins em cada mês do ano, para observar quais os meses mais ativos.

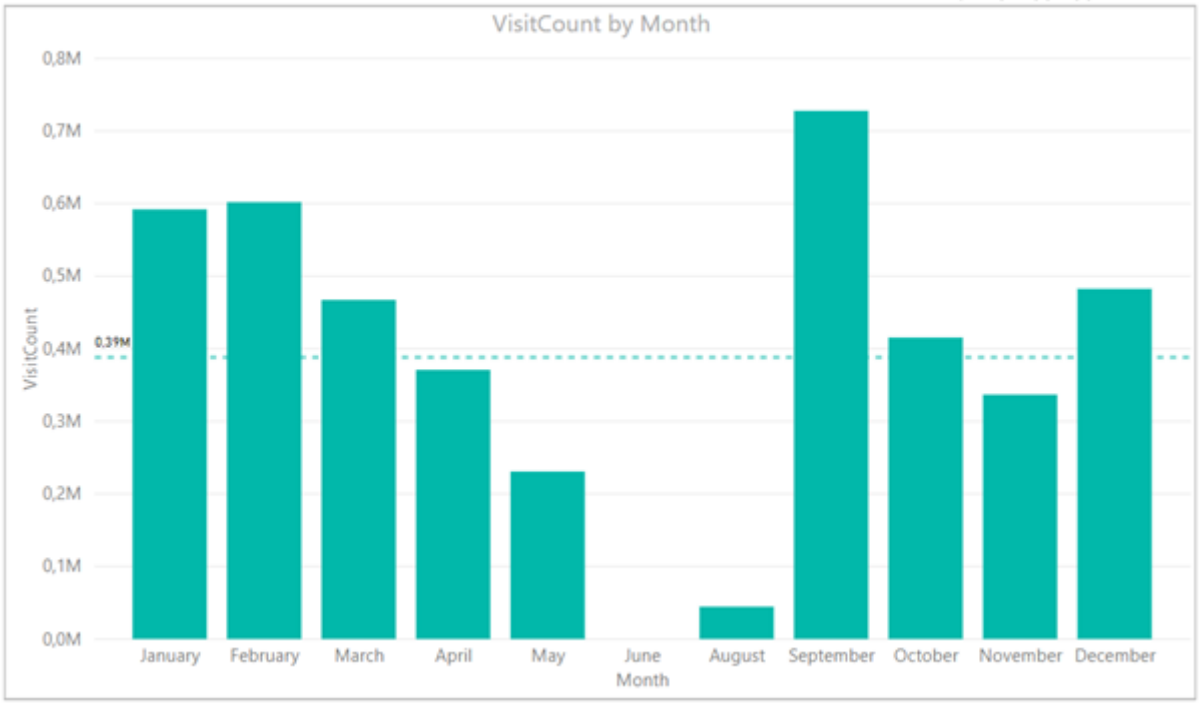


Gráfico 13 – Número de logins efetuados por mês, no total das 5 temporadas em estudo.

O gráfico obtido revela-nos a inexistência de logins nos meses de Junho e Julho, o que seria expectável visto serem os meses de intervalo entre cada temporada da liga. Agosto contém o menor número de logins, o que mais uma vez corresponde ao expectável, dado que as temporadas apenas iniciam nos últimos dias de Agosto. Setembro revelou ser o mês com maior número de logins, tal como previsto, pois corresponde á altura em que os concorrentes escolhem as suas equipas para o início da temporada. A seguir ao mês de Setembro, Janeiro e Fevereiro mostraram valores elevados no número de logins. Este resultado faz sentido, visto que no mês de Janeiro começa a segunda volta e o campeonato de verão, e, por outro lado, no mês de Fevereiro ocorre o mercado de transferências de Inverno onde os concorrentes podem obter novas aquisições para a sua equipa.

A seguir fomos verificar qual a atividade dos concorrentes no jogo ao longo das temporadas, e para isso criámos um relatório com o somatório de logins por dia ao longo de cada temporada.

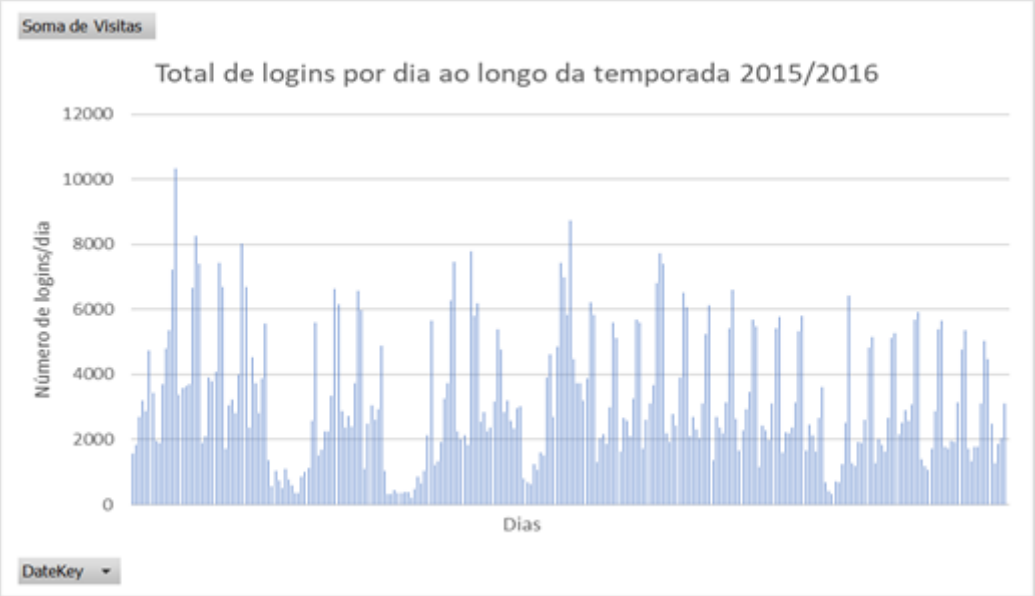


Gráfico 14 – Total de logins registados no site por dia ao longo da temporada 2015/16. A análise foi efetuada para todas as temporadas, com resultados semelhantes entre si.

Como podemos ver no gráfico anterior, é bem visível que o acesso ao jogo não ocorre de forma uniforme ao longo das temporadas, mas que, pelo contrário, o número de logins atinge picos que parecem ocorrer uniformemente intervalados no tempo entre si. Para além disso os picos mais elevados ocorrem no início da temporada (que é consistente em todas as temporadas analisadas), o que está de acordo com o observado anteriormente de que o pico de Logins ocorre em Setembro.

Como especificámos anteriormente, ao longo do jogo existem três datas consideradas fulcrais e que se repetem a cada ronda (de semana a semana): a data de inicio das apostas, na qual os concorrentes podem efetuar as alterações que desejam à sua equipa; a data de fim das apostas, data a partir da qual já não são mais aceites alterações às equipas; e por fim a data de publicação de resultados, quando os resultados obtidos pelos concorrentes durante a jornada são publicados.

Para analisar se os picos observados anteriormente poderão corresponder a alguma destas datas, começámos por obter a distribuição destes três processos ao longo da semana.

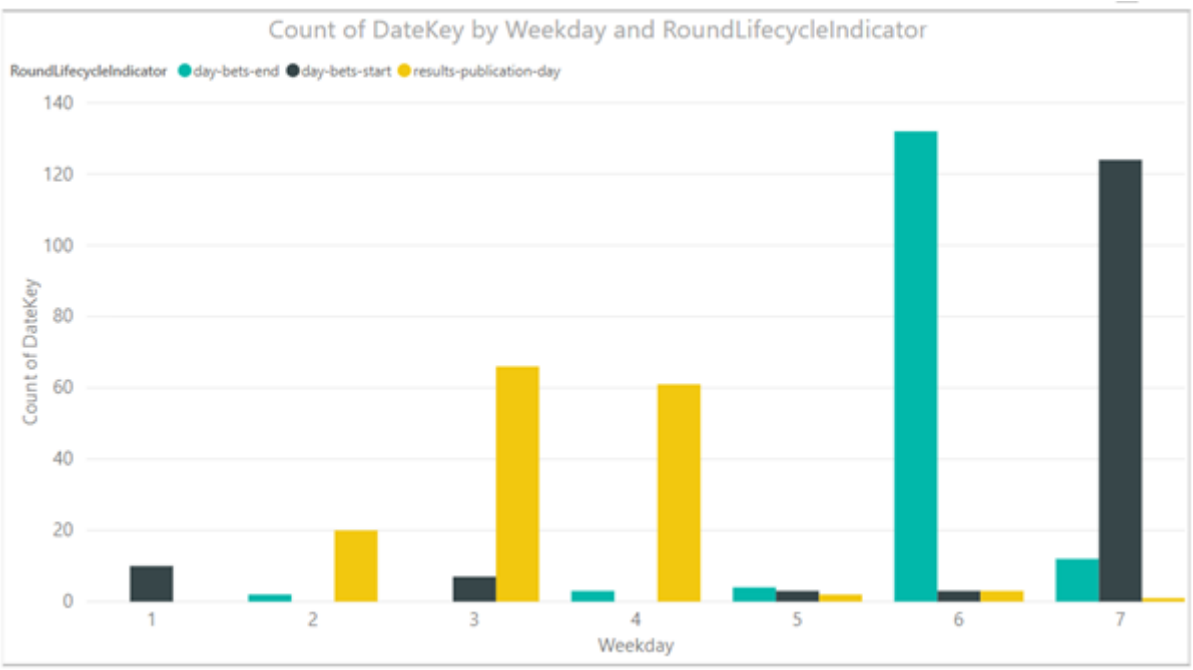


Gráfico 15 – Distribuição dos três dias chaves que compõem uma ronda. O 1 corresponde ao Domingo, e assim sucessivamente até ao 7 que corresponde ao Sábado.

Podemos observar no gráfico que os dias da semana em que tipicamente ocorre a publicação dos resultados é na terça e quarta-feira; o dia da semana em que tipicamente ocorre a data de início das apostas é na sexta-feira; e por último o dia da semana em que tipicamente ocorre a data de fim das apostas é no sábado.

No seguimento desta análise, o seguinte passo tomado foi verificar qual a distribuição do total de logins por dia ao longo da semana.

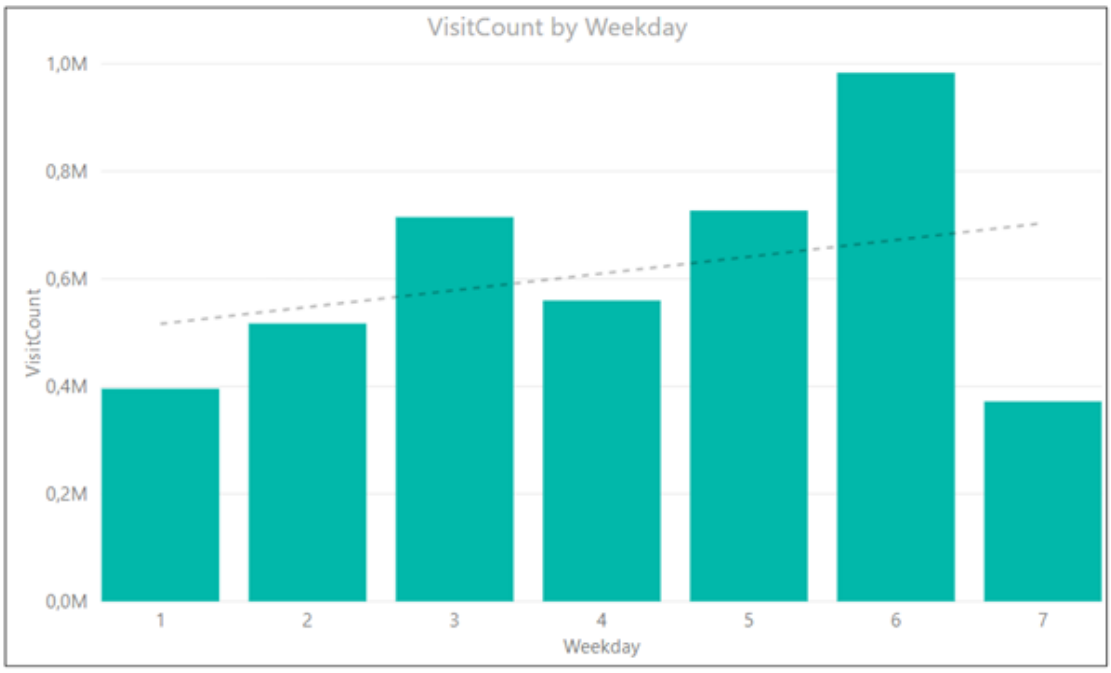


Gráfico 16 – Somatório de logins registados em cada dia da semana ao longo de todas as temporadas em análise. O 1 corresponde ao Domingo, e assim sucessivamente até ao 7 que corresponde ao Sábado.

Como vemos no gráfico o pico de logins ocorre na sexta que, tal como previsto, coincide com o dia da semana em que tipicamente termina o período em que é possível efetuar mudanças às equipas pelos concorrentes. Para além disso, obtivemos também como previsto, um somatório de logins elevado nas terças e quartas, que como vimos, correspondem aos dias em que tipicamente são publicados os resultados de cada jornada.

Para além destas datas, propomos a possibilidade de ocorrer um maior número de logins nas datas em que decorram jogos clássicos.

Para validar ou não esta hipótese, criámos então um relatório para obter a média de visitas por dia nos dias em que ocorrem jogos clássicos em comparação com os outros dias.

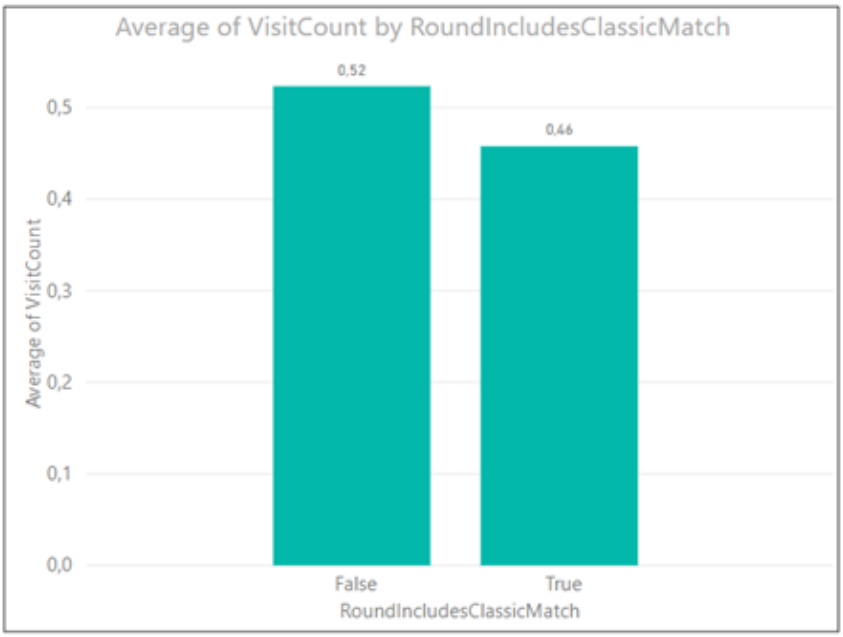


Gráfico 17 – Média de visitas por dia, em dias de clássico (TRUE) e em dias de não clássico (FALSE).

Não parece existir diferença significativa na média de logins registados em dias de clássico em comparação com os restantes dias, ocorrendo mesmo uma média de visitas inferior para dias de clássico. Podemos então concluir que os picos de logins observados ao longo das temporadas ocorrem maioritariamente em dias onde ocorre o fim das apostas.

Ao longo da obtenção dos relatórios, fomos verificando duas tendências consistentes em todas as análises, e que foram validadas pelos resultados demonstrados nos seguintes gráficos.

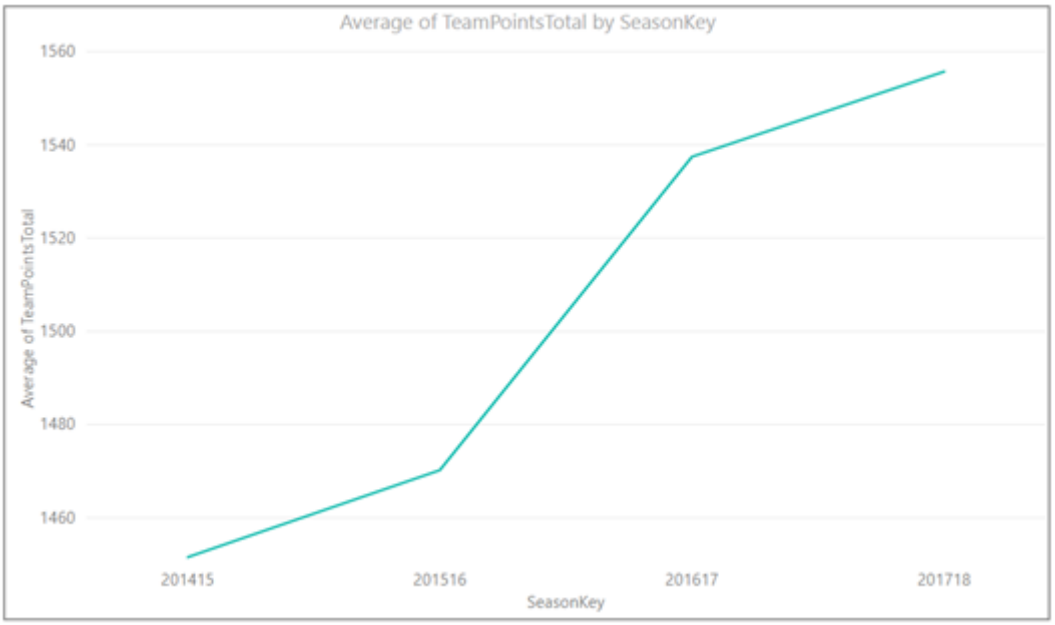


Gráfico 18 – Média das pontuações totais da última ronda por temporada.

No gráfico 18 vemos que a pontuação atingida pela melhor equipa a cada temporada tem sido cada vez maior.

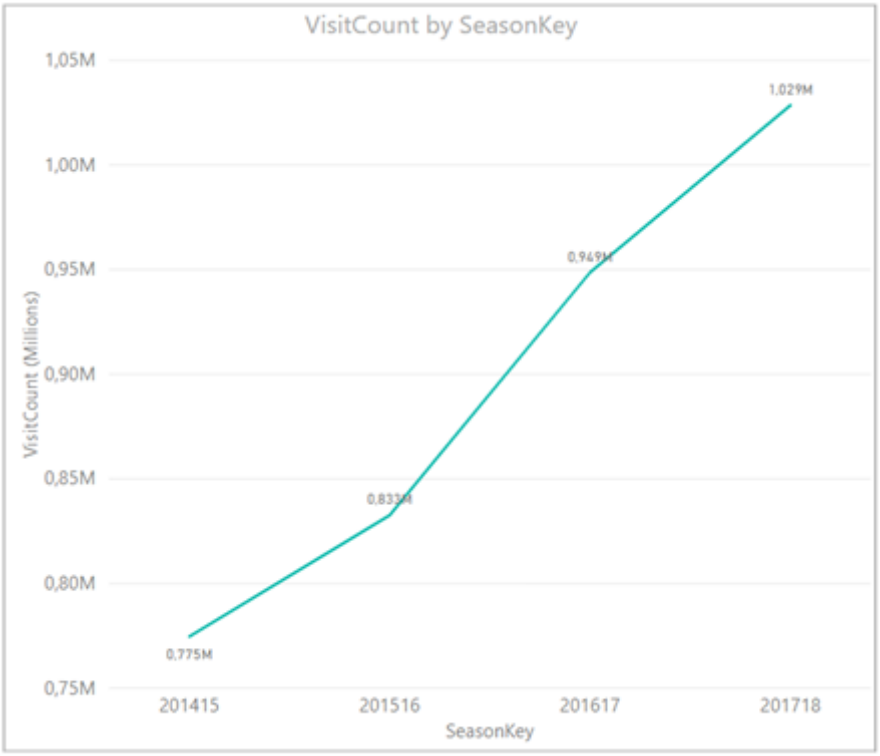


Gráfico 19 – Total de visitas por temporada (foi removida a temporada 2018/19 uma vez que ainda está a decorrer e poderia enviesar os resultados).

Por outro lado, no gráfico 19 é bem evidente que o número de visitas está a aumentar ao longo das temporadas, o que por um lado pode demonstrar um interesse cada vez maior dos concorrentes pelo jogo ou o registo de um número cada vez maior participantes.

9. Prospeção de dados

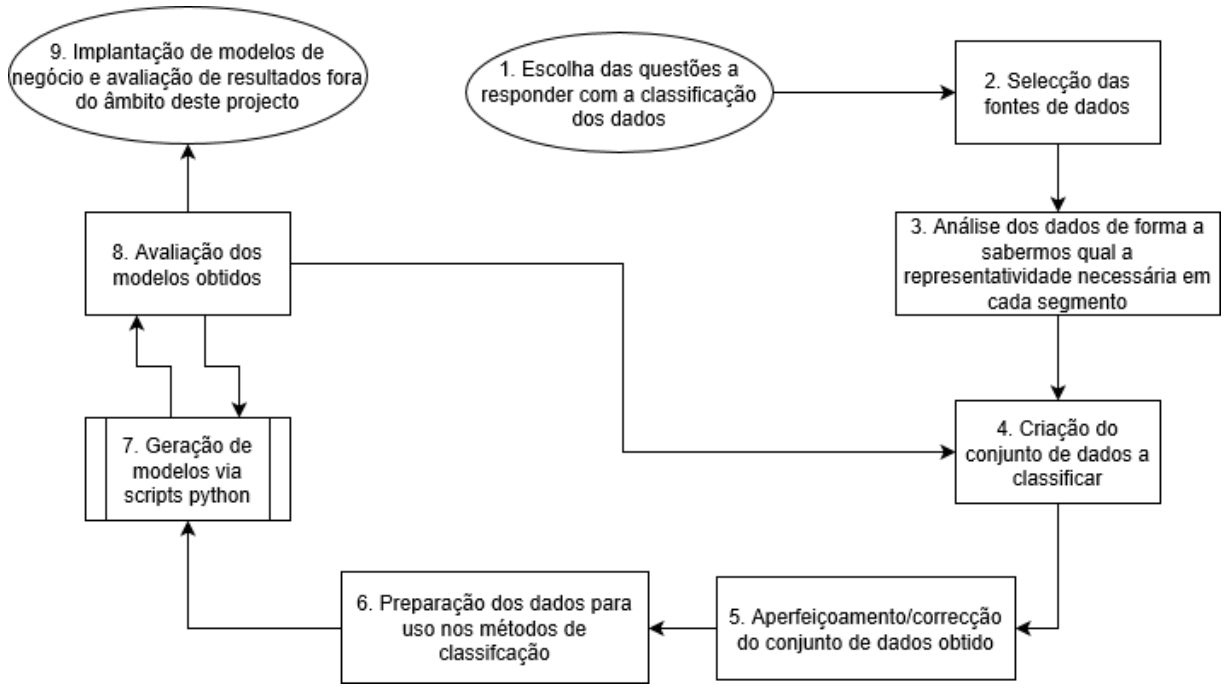
Para fazer a parte de prospecção dos nossos dados, optámos por usar duas técnicas técnicas principais na tarefa de classificação do data mining: árvores de decisão e redes neuronais.
Para ajudar na classificação, os vários segmentos inerentes aos dados foram agrupados em classes pré-determinadas.

Os métodos usados foram **supervisionados** e as duas questões que procurámos avaliar nesta fase foram as seguintes:

1. **Método: Redes Neuronais**
   Tendo à disposição dados descritivos dos concorrentes que registram evoluções classificativas na Liga Record (como o género, clube e faixa etária do concorrente), e a par daquilo que foi demonstrado na resposta à pergunta analítica 1, optámos por tentar prever a classificação final dos concorrentes.

2. **Método: Árvores de decisão**
   Neste segundo método, optámos por tentar prever, usando os qualificadores género, clube, faixa etária, região e número de visitas anual, em conjunto com a posição da melhor equipa numa determinada época e com quantas equipas jogou cada concorrente, qual seria o número de equipas adquirido na época seguinte.

Diagrama do data mining tal como foi implementado neste projecto:



a. Método das Árvores de Decisão

Como já referido anteriormente, o nosso objectivo com o uso deste método seria construir um modelo de classificação que conseguisse prever o número de equipas adquiridas pelo concorrente na época seguinte, fornecendo o conjunto de dados respectivos a uma determinada época corrente.

Uma vez que existe um número variado de equipas que um concorrente pode adquirir, optámos por criar classes para o atributo classificador 'Número de equipas na época seguinte' (valores de output do modelo). Tendo em consideração que a média de equipas que um jogador adquire é cerca de 2,3 equipas, as classes definidas foram as seguintes:

- 1 equipa
- 2 equipas
- > 2 equipas

i. Preparação dos dados

Sendo o nosso Universo de sensivelmente 6 mil concorrentes por época, um dos nossos principais focos passou pela obtenção de um subconjunto desses dados que fosse o mais representativo possível desse universo, com especial atenção para as seguintes informações:

- Quantidade de mulheres equivalente a 10% dos homens;
- Apresentar apenas concorrentes pertencentes aos clubes mais representados e com a mesma distribuição do conjunto geral de concorrentes: mais representado, o Benfica, seguido do Sporting, seguido do FC Porto;
- Representar a quantidade de utilizadores de cada segmento da faixa etária pela mesma ordem da totalidade dos concorrentes: 35-44 > 25-34 > 45-54 > 55-64 e 15-24 equilibrados, seguidos dos maiores de 64 e finalmente os menores de 15.

A seguinte decisão tomada foi quais os atributos qualificadores (valores de input do modelo) são relevantes usar para a construção do modelo. Para tal tivemos em consideração os resultados que obtivémos nos relatórios analíticos, selecionando as variáveis que melhor descrevem a variação dos dados na BD relacional.

Os atributos usados foram os seguintes:

- Faixa etária
- Género
- Clube
- Região
- Número de visitas
- Rank total final
- Número de equipas (na época actual)

Os data sets foram obtidos através de queries SQL usando ordenações aleatórias, ponderando a distribuição pretendida, tendo como base essencialmente os dados presentes na

data presentation area do nosso *data warehouse* em conjunto com novos dados actualizados com os resultados das últimas rondas da época 2018/19 obtidos das bases de dados da Cofina.

Uma vez que a maioria dos concorrentes joga apenas com 1 equipa, a segunda questão apresentou alguma complexidade no que toca à escolha do data set adequado.
Acabámos por viciar um pouco as proporções no que diz respeito a representatividade do número de equipas por concorrente de forma a conseguir melhores resultados nos segmentos menos representados.

Os dados foram recolhidos em SQL e exportados para excel para serem trabalhados.

Já no excel, criámos inicialmente um dataset onde demos relevância à representatividade de cada classe, escolhemos um equilíbrio de equipas por concorrente e demos representatividade em termos de cada temporada relativamente aos dados globais, um total de 725 exemplos:

- 2014/15: 154 linhas
- 2015/16: 219 linhas
- 2016/17: 180 linhas
- 2017/18: 172 linhas

Abaixo temos a tabela com os totais de exemplos para cada categoria de atributos usado como variável qualificadora. Estão representados os atributos mais importantes para a representatividade do conjunto de dados em relação ao universo de dados da base de dados.

Atributos Qualificadores (dataset com 725 linhas)									
Faixa Etária		**Género**		**Clube**		**Região**		**Número de Equipas**	
< 15	4	**Feminino**	70	**Benfica**	399	**Centro**	424	**1**	182
15 - 24	45	**Masculino**	655	**FC Porto**	135	**Ilhas**	45	**2**	161
25-34	186			**Sporting**	191	**Norte**	112	**> 2**	382
35-44	274					**Sul**	144		
45-54	150								
55-64	46								
> 64	20								

Para comparação de resultados, criámos um novo dataset de maior dimensão, com nomeadamente 1286 exemplos, onde tivemos as mesmas preocupações no que toca à representatividade dos dados do dataset anterior.

- 2014/15: 256 linhas
- 2015/16: 402 linhas
- 2016/17: 322 linhas
- 2017/18: 306 linhas

Atributos Qualificadores (dataset com 1286 linhas)									
Faixa Etária		**Género**		**Clube**		**Região**		**Número de Equipas**	
< 15	5	**Feminino**	109	**Benfica**	717	**Centro**	737	**1**	539
15 - 24	88	**Masculino**	1177	**FC Porto**	253	**Ilhas**	70	**2**	344
25-34	349			**Sporting**	316	**Norte**	223	**> 2**	403
35-44	498					**Sul**	256		

45-54	239								
55-64	75								
> 64	32								

Nota: Foram usados tanto para treino como para teste dados das temporadas das quais tínhamos dados para confirmar os resultados.

ii. Construção do modelo

Os scripts usados para a construção do modelo foram escritos em Python.

Os classificadores descritivos (género, clube, faixa etária, região) foram convertidos em classes e o número de equipas foi segmentado em 1, 2 ou mais que 2.

Na construção do modelo foram testados 2 algoritmos: o primeiro mais simples, algoritmo *RandomForestClassifier*; e um segundo mais complexo, algoritmo *XBoost*.

Uma vez que por definição uma grande parte dos dados do dataset deve se destinar para treinar o modelo, 75% dos dados foram usados no processo de treinamento e 25% foram usados para testar e avaliar o modelo. No caso do algoritmo XBoost, de forma a reduzir o overfitting, usámos 10% dos dados de treino para validação dos mesmos.

iii. Resultados e Discussão

- **Algoritmo Random Forest (Dataset com 725 exemplos)**

Actual Class/ Predicted Class	**1**	**2**	**>2**
1	56	4	19
2	17	3	8
>2	7	3	65

Accuracy: 68%

Em cima temos ilustrada a tabela de confusão resultante. Na matriz confusão, os verdadeiros positivos encontram-se na diagonal da tabela, enquanto que os restantes valores correspondem aos falsos positivos e falsos negativos. Logo devemos obter uma tabela de confusão com os valores mais elevados situados na diagonal. Uma métrica bastante usada para medir a eficácia do modelo é a precisão, a qual nos fornece a proporção de verdadeiros positivos no universo de positivos encontrados pelo modelo (incluindo os falsos positivos). Desta forma um valor de precisão elevado indica-nos que a classificação obtida corresponde à verdade.

Para o dataset de 725 exemplos, obtivémos uma precisão de 68%. Vemos que o modelo classifica razoavelmente bem para as classes *1* e *>2*, mas não está a conseguir classificar para a classe intermédia *2*, sendo que a maior parte está a ser erradamente classificada como pertencente à classe *1*.

Uma vez que temos diversas categorias a ter em conta na representatividade dos dados, é possível que o dataset não tenha a quantidade de exemplos necessária para o modelo encontrar devidamente padrões nos dados, ou as variáveis usadas não são suficientemente informativas.

Para testar se o tamanho do dataset está a limitar o bom funcionamento do modelo de classificação, decidimos usar um dataset que fornece uma maior quantidade de exemplos.

- **Algoritmo Random Forest (Dataset com 1106 exemplos)**

Actual Class/ Predicted Class	1	2	>2
1	152	9	12
2	22	16	17
>2	21	9	64

Accuracy: 72%

Para o dataset de 1106 exemplos, obtivémos um valor de precisão mais elevado do que com o dataset anterior, nomeadamente 72%, ao qual corresponde ao aumento de 4%. Podemos verificar na matriz de confusão obtida que ocorreu um ligeiro melhoramento a classificar correctamente para a classe *2*, uma vez que ocorreu um aumento de verdadeiros positivos e uma diminuição dos falsos positivos. Para a classe *> 2* não verificamos diferenças significativas na classificação em relação ao exemplo anterior. Porém para a classe *1*, verificamos um aumento na proporção de verdadeiros positivos comparativamente com a proporção obtida para esta classe no exemplo anterior, o que mais uma vez nos indica que o modelo classifica muito bem para esta classe em comparação com as restantes duas.

- **Algoritmo XBoost**

Nesta análise foram usados os mesmos dados da 2ª iteracção do algoritmo Random Forest. Os resultados obtidos pioraram consideravelmente mesmo depois de várias tentativas de alteração das variáveis a passar ao algoritmo (learning_rate, max_depth e n_estimators).

Os resultados obtidos aparentam ser um caso de overfitting relativamente aos dados de teste, uma vez que os resultados obtidos pioraram consideravelmente quando usámos os dados de validação do modelo.

Actual Class/ Predicted Class	1	2	>2
1	120	48	6
2	4	36	4
>2	7	59	38

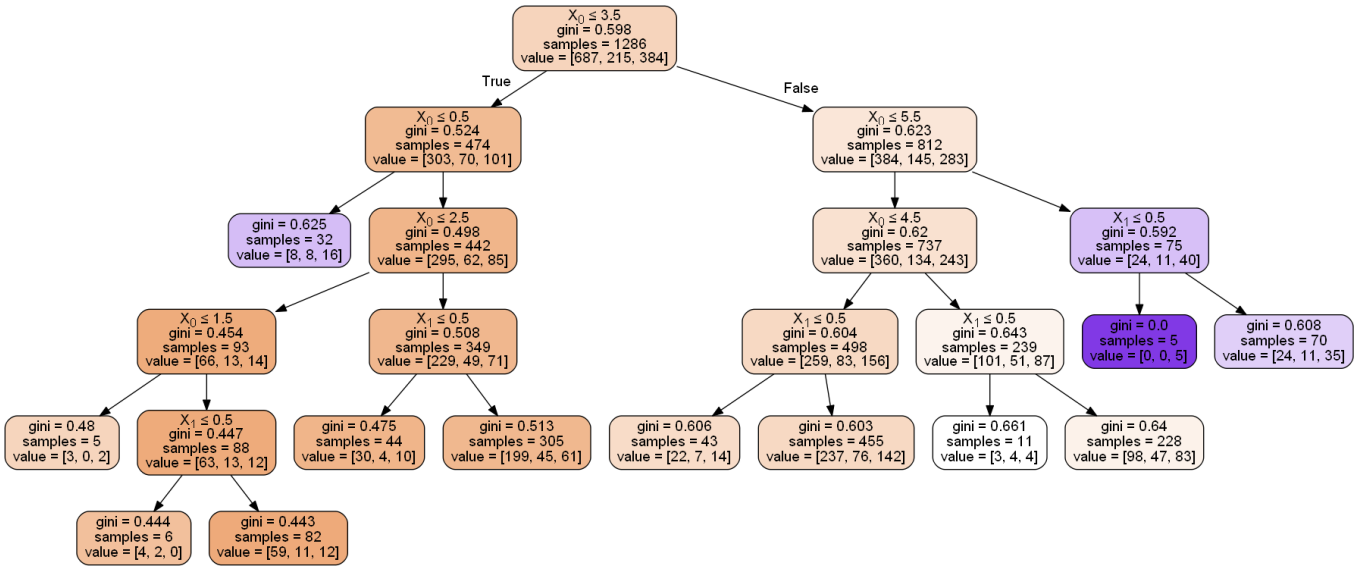
Train Score: 79.90%
Test Score: 60.25%

É de notar que novamente a classe que piora na classificação é a classe intermédia para 2 equipas, com valores de falsos positivos superiores ao de verdadeiros positivos.

Após verificarmos que não obtivémos melhoras consideráveis nem pelo aumento do tamanho do dataset utilizado, nem pelo uso de um algoritmo mais robusto, e que o modelo continua a apresentar um forte enviesamento para a classe 1, podemos presumir que os resultados estão a ser negativamente influenciados pelo facto de as variáveis usadas não serem suficientemente informativas para a previsão do número de equipas que serão adquiridas pelos concorrentes.

A árvore criada deu demasiados nós para conseguir ser bem visualizada neste documento. Encontra-se no servidor em C:\TPD-ProspeccaoDados\predict_teams\output.png

Segue um excerto apenas em função da faixa etária e género:



Nota: O código e datasets fonte que usámos neste método de prospecção pode ser encontrado no nosso servidor da Azure, na pasta C:\TPD-ProspeccaoDados\predict_teams.

b. Método das Redes neuronais

Nesta secção, o objetivo foi a previsão da classificação final dos concorrentes, com base nos desempenhos dos mesmos ao longo do decorrer da Liga Record.

i. Preparação dos dados

Neste método usámos como fonte um data set com todas as temporadas estudadas anteriorente ao que se acrescentou o ano de 2018/19, entretanto disponível, num total de 867 exemplos:

- 2014/15: 152 linhas
- 2015/16: 220 linhas
- 2016/17: 180 linhas
- 2017/18: 168 linhas
- 2018/19: 147 linhas

Para além de grande parte dos qualificadores anteriores, aqui usámos também a evolução da equipa mais bem classificada do concorrente ao longo do tempo (posição que tinha à 5ª, 10ª, 15ª e 20ª ronda) para determinar a sua posição final.

Os atributos usados foram, portanto:

- Faixa etária
- Género
- Clube
- Número de visitas
- Rank total na ronda 5
- Rank total na ronda 10
- Rank total na ronda 15
- Rank total na ronda 20
- Rank total final

Abaixo temos a tabela com os totais de exemplos para cada categoria de atributos usado como variável qualificadora.

Atributos Qualificadores (dataset com 867 linhas)					
Faixa Etária		**Género**		**Clube**	
< 15	5	**Feminino**	67	**Benfica**	465
15 - 24	75	**Masculino**	800	**FC Porto**	179
25-34	264			**Sporting**	223
35-44	304				
45-54	162				
55-64	60				
> 64	27				

ii. Construção do modelo

A previsão das classificações ficou ao encargo de uma *Recurrent Neural Network* (RNN). As redes neuronais podem ser introduzidas como sistemas computacionais inspirados nas redes neuronais biológicas que constituem os cérebros nos seres vivos. Uma das formas de aprendizagem consiste na utilização de exemplos de treino e de um sistema que possa aprender com os mesmos. Assim, as redes neuronais utilizam exemplos para inferir automaticamente regras de classificação.

Os elementos basilares das redes neuronais são os **neurónios**, conectados entre si via sinapses. Um neurónio recebe um input resultante de outros neurónios e é caracterizado por: uma função de ativação (dependente do input, do valor da ativação e de um threshold), que dá a evolução da ativação, e uma função de output, que dá o output em função do valor da ativação.

A rede em si traduz-se nas **conexões** que transferem os outputs entre os neurónios. A cada conexão é atribuído um **peso**. Assim, o input de cada neurónio consiste numa soma de outputs, ponderada pelos pesos das conexões, somado a um valor de **bias**.

A aprendizagem em si vai ser obtida por retroprogagação, que tal como o nome indica vai permitir ensinar a rede através do caminho na retaguarda a partir do output com erro verificado através dos valores pré-classificados. Este processo pode ser ajustado através de um coeficiente de activação, que dependendo do seu maior ou menor peso, que irá permitir uma aprendizagem mais rápida (com risco da não convergência da aprendizagem da rede, momento em que a rede pode não convergir para uma solução óptima) ou mais lenta, respectivamente.

A função de perda vai informar a rede neural do seu parâmetro de paragem, tendo em conta que a convergência da rede neural em uma solução perfeita pode ser uma tarefa muito complicada, esta função vai permitir definir o parâmetro de paragem, determinando quão perto de uma solução perfeita se deve de atingir antes de se dar a evolução da rede por terminada.

iii. Avaliação dos resultados

Esta experiência serviu, em primeira instância para desenvolver uma rede que operasse com um histórico de classificações em determinadas rondas para fazer a previsão da posição final do utilizador na Liga Record.
Para tal, foi utilizada, enquanto ponto de partida, uma RNN sequencial, com arquitetura *Long Short-Term Memory* (LSTM). As LSTMs introduzem células de memória, unidades de computação, que substituem os neurónios artificiais nas *hidden layers* da rede. Estas células de memória permitem à rede associar memórias e inputs remotos no tempo, sendo portanto adequadas para a compreensão da estrutura dinâmica dos dados, o que lhes confere uma boa capacidade de previsão.

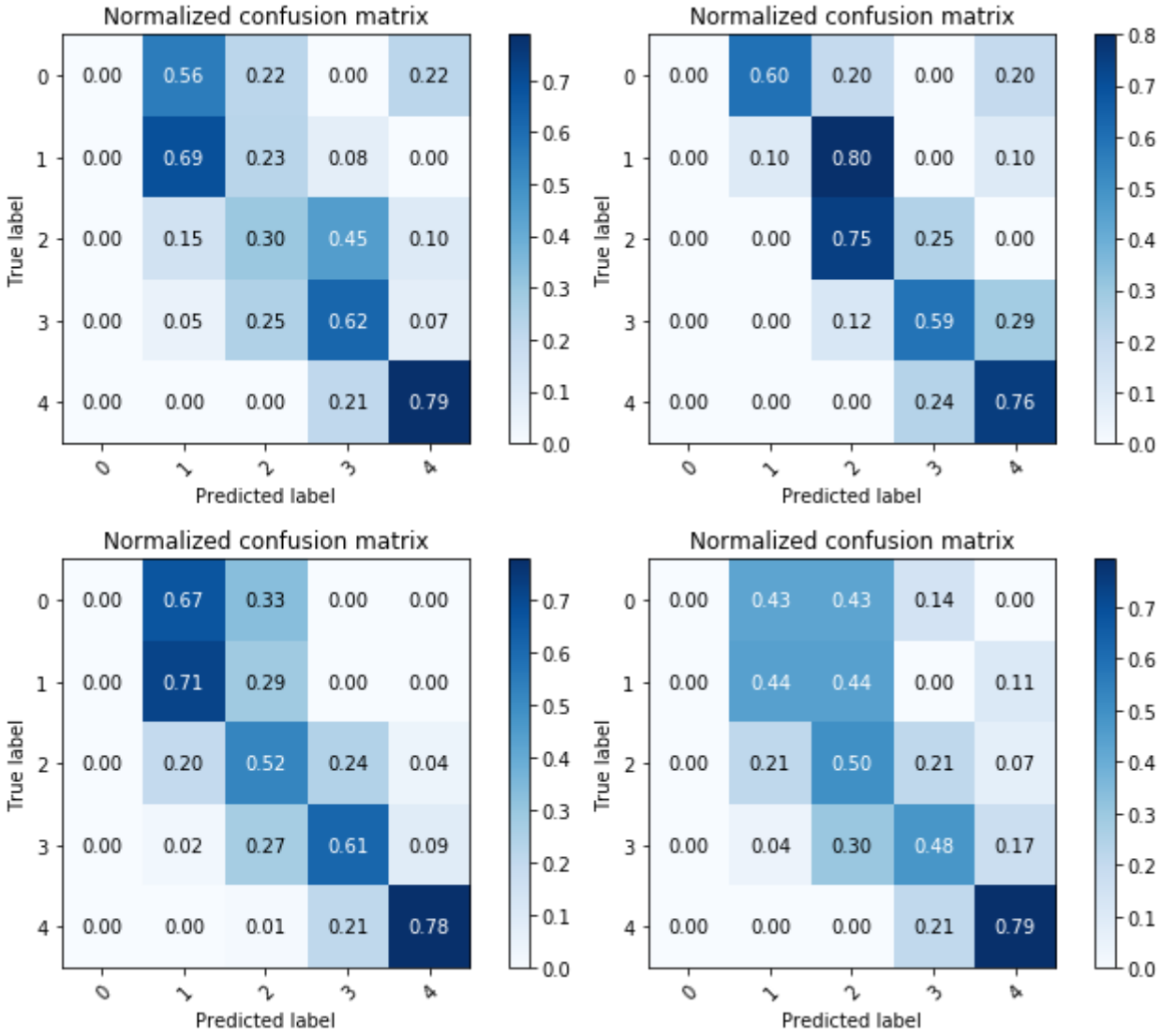
Metodologia:

1. **Preprocessing:** Partindo de uma tabela com os os dados relativos às classificações para nas rondas 5, 10, 15 e 20, procedeu-se à discretização dos dados: as posições finais são

transformadas em *labels*, de acordo com os intervalos de percentil no qual estão inseridas - top 5%, top10%, top25%, top50% e fora do top50%. De seguida, os dados foram normalizados, foi testada a presença de valores omissos e foram finalmente estruturados sob a forma de 4 pontos sequenciais e um *output*. Finalmente, os dados foram repartidos entre um *training set* e um *test set*, com dimensões 700 e 167.

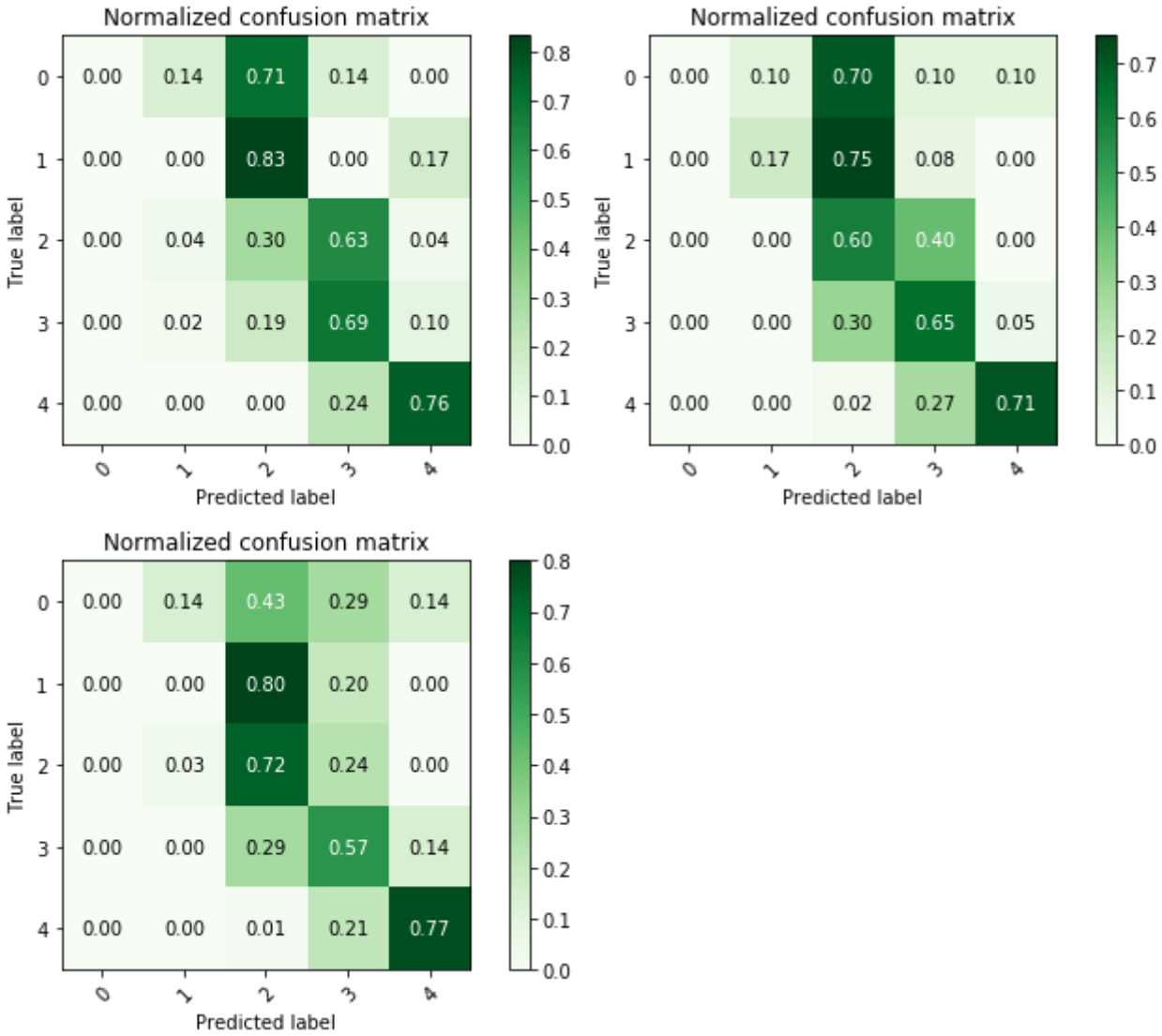
2. **Treino da rede neuronal:** os dados foram fornecidos à rede neuronal e esta é treinada para a previsão dos resultados. O modelo LSTM de base é composto por uma *input layer* sequencial com 50 unidades, seguida de 3 *hidden layers* com 50 unidades LSTM, com regularização de dropout. Finalmente, uma *output layer*, de tipo *Dense*, com uma função de ativação linear. O otimizador escolhido foi o *Adam* (*Adaptive Moment Estimation*), com '*mean squared error*' como função *loss*.
3. **Geração de output:** o output gerado pela rede é comparado ao valor real e o erro entre ambos é minimizado com recurso a um algoritmo de *backpropagation*, que ajusta os pesos e os *biases* da rede.

Os resultados obtidos são apresentados de seguida:



Em cima, 4 exemplos dos resultados gerados pela rede. As accuracies verificadas têm valor médio 0,6154 ± 0,0195 (N=10).

Para verificar alguma da influência da arquitetura da rede, em termos de número de unidades em cada camada, e do número de camadas, reduzimos o número de unidades nas hidden layers da rede de 50 para 15, e o número de hidden layers de 3 para 1. O efeito nos resultados foi o seguinte:



As accuracies registadas para os exemplos em cima são, respetivamente, de 0,6108 , 0,5988 e 0,6587.

Nota: O código e datasets fonte que usámos neste método de prospecção podem ser encontrados no nosso servidor da Azure, na pasta C:\TPD-ProspeccaoDados\predict_rank

10. Bibliografia

[1]https://www.pordata.pt/Portugal/Popula%C3%A7%C3%A3o+residente++m%C3%A9dia+anual+total+e+por+grupo+et%C3%A1rio-10

[2] https://liga.record.pt/info/ajuda.aspx

[3] http://centraldedados.pt/codigos_postais/

11. Anexos

Anexo 1 – Análise de estatistica descritiva dos atributos das tabela de dimensões

Tabela de Dimensão User (31524 observações de 23 variáveis)		
Atributo	**Tipo de variável**	**Estatísticas**
User Key	**Qualitativa** (31524 categorias)	4 : 1 5 : 1 6 : 1 7 : 1 8 : 1 9 : 1 (Other):31518
User Original Key	**Qualitativa** (6305 categorias)	760464 : 5 760465 : 5 760466 : 5 760469 : 5 760471 : 5 760472 : 5 (Other) :31494
User Email	Texto	Length:31524 Class :character Mode :character
User Nickname	Texto	Length:31524 Class :character Mode :character

User Birthdate	**Quantitativa Contínua** (data)	Min. :1911-08-08 1st Qu.:1974-09-09 Median :1980-02-02 Mean :1979-06-21 3rd Qu.:1985-07-07 Max. :2019-03-16
User Gender	**Qualitativa** (2 categorias)	Feminino : 3220 Masculino:28304
User Club	**Qualitativa** (25 categorias)	Benfica :16719 Sporting : 7210 FC Porto : 5180 Sem clube : 370 V. Guimarães: 310 (Other) : 1290 NA's : 445
User Region	**Qualitativa** (24 categorias)	Lisboa : 6774 Porto : 3105 Aveiro : 2780 Setúbal: 2515 Faro : 1965 (Other):14350 NA's : 35
User Zipcode Locality	**Qualitativa** (422 categorias)	6000 : 645 0000 : 590

		3700 : 457 4620 : 380 7800 : 380 (Other):28607 NA's : 465
User Zipcode Locality Designation	**Qualitativa** (404 categorias)	LISBOA : 1384 ALMACEDA : 619 ARRIFANA VFR: 436 FARO : 434 AMADORA : 396 (Other) :27200 NA's : 1055
User Locality	**Qualitativa** (406 categorias)	Lisboa : 1380 Aboboreira: 844 Abrunheira: 454 Ameixieira: 428 Amadora : 396 (Other) :26967 NA's : 1055
User County	**Qualitativa** (245 categorias)	Lisboa : 1496 Sintra : 967 Santa Maria da Feira: 962 Castelo Branco : 721 Vila Franca de Xira : 689 (Other) :25634 NA's : 1055

User District	**Qualitativa** (23 categorias)	Lisboa : 6419 Porto : 3025 Aveiro : 2851 Setúbal: 2594 Faro : 1909 (Other):13671 NA's : 1055
User Country	**Qualitativa** (20 categorias)	Portugal :31059 Afeganistão: 81 Reino Unido: 71 Suiça : 63 França : 56 Angola : 45 (Other) : 149
User Original Start Date	**Quantitativa Contínua** (data)	Min. :2011-07-28 1st Qu.:2011-08-23 Median :2011-09-11 Mean :2012-09-08 3rd Qu.:2013-09-10 Max. :2015-05-03
User Season Start Date	**Quantitativa Contínua** (data)	Min. :2014-08-01 1st Qu.:2015-08-10 Median :2016-08-17 Mean :2016-08-20 3rd Qu.:2017-08-26

		Max. :2019-02-22
User Premium Date	**Quantitativa Contínua** (data)	Min. :2014-08-04 1st Qu.:2015-08-03 Median :2016-08-16 Mean :2016-06-28 3rd Qu.:2017-08-16 Max. :2019-03-01 NA's :28899
User Agegroup	**Qualitativa** (7 categorias)	-15 : 67 +64 : 279 15-24: 1445 25-34:10680 35-44:13690 45-54: 4273 55-64: 1090
User Is In League	**Qualitativa** (2 categorias – boleano)	Mode: Logical TRUE:30669 FALSE: 855
Effective Date Row	**Quantitativa Contínua** (data)	Min. :2014-08-01 1st Qu.:2015-08-10 Median :2016-08-17 Mean :2016-08-20 3rd Qu.:2017-08-26 Max. :2019-02-22
Expiration Date Row	**Quantitativa Contínua**	

	(data)	Min. :2014-08-04 1st Qu.:2016-08-08 Median :2017-08-13 Mean :3613-10-03 3rd Qu.:2018-09-05 Max. :9999-12-31
Timestamp Row	**Quantitativa Contínua** (data)	Min. :2014-08-01 1st Qu.:2015-08-10 Median :2016-08-18 Mean :2016-08-15 3rd Qu.:2017-08-27 Max. :2019-03-01
Is Current Row	**Qualitativa** (2 categorias – boleano)	Mode: Logical TRUE: 6305 FALSE:25219

Tabela de Dimensão Season (5 observações de 7 variáveis)		
Atributo	**Tipo de variável**	**Estatísticas**
Season Key	**Qualitativa** (5 categorias)	201415:1 201516:1 201617:1 201718:1 201819:1
Season Name	Texto	Length:5 Class :character Mode :character
Season Start Date	**Quantitativa Contínua** (data)	Min. :2014-07-01 1st Qu.:2015-07-01 Median :2016-07-01 Mean :2016-06-30 3rd Qu.:2017-07-01 Max. :2018-07-01
Season End Date	**Quantitativa Contínua** (data)	Min. :2015-06-30 1st Qu.:2016-06-30 Median :2017-06-30 Mean :2017-06-29 3rd Qu.:2018-06-30 Max. :2019-06-30
Season Has Updated Game Version	**Qualitativa** (2 categorias – boleano)	Mode: Logical TRUE:4

		FALSE:1
Season Has Variable Weekday Publish Date	**Qualitativa** (2 categorias – boleano)	Mode: Logical TRUE:2 FALSE:3
Team Player Transfer Allowed Per Month	**Qualitativa** (2 categorias)	1:3 2:2

Tabela de dimensão Team 54935 observações de 7 variáveis		
Atributo	**Tipo de variável**	**Estatística**
Team Key	**Qualitativa** (54935 categorias)	1 : 1 10 : 1 100 : 1 1000 : 1 10000 : 1 10001 : 1 (Other):54929
Team Original Key	**Qualitativa** (45131 categorias)	10210 : 4 10322 : 4 10813 : 4 10846 : 4 11102 : 4 11137 : 4 (Other):54911
Team Name	Texto	Length:54935 Class :character Mode :character
Team Create Date	**Quantitativa Contínua** (data)	Min. :2014-08-04 1st Qu.:2015-08-10 Median :2016-08-16 Mean :2016-07-20 3rd Qu.:2017-08-25 Max. :2019-03-26

Team Origin	**Qualitativa** (11 categorias)	REVISTA :30526 MB :13828 OFERTA 3+1 : 6161 CC : 2830 PAYPAL : 1424 OFERTAS "VÁRIOS": 112 (Other) : 54
Team Is Paid	**Qualitativa** (2 categorias – boleano)	Mode: Logical TRUE: 48632 FALSE: 6303
Team In League	**Qualitativa** (2 categorias – boleano)	Mode: Logical TRUE: 45508 FALSE: 9427

Tabela de dimensão Date (1302 observações de 17 variáveis)		
Atributos	**Tipo de variável**	**Estatísticas**
Key	**Qualitativa** (1302 categorias)	20140601: 1 20140829: 1 20140830: 1 20140831: 1 20140901: 1 20140902: 1 (Other) :1296
Day	**Quantitativa Contínua** (data)	Min. :2014-06-01 1st Qu.:2015-10-19 Median :2016-12-15 Mean :2016-12-09 3rd Qu.:2018-02-06 Max. :2019-05-21
Day of Month	**Qualitativa** (31 categorias)	1 : 49 2 : 44 3 : 44 5 : 44 7 : 44 9 : 44 (Other):1033
Month	**Qualitativa** (12 categorias)	1 :155 3 :155 10 :155

		12 :155 9 :150 11 :150 (Other):382
Date Full	**Quantitativa Contínua (data)**	
Weekday	**Qualitativa** (7 categorias)	1:187 2:189 3:183 4:182 5:190 6:185 7:186
Calendar Weekday	**Qualitativa** (assume 7 categorias)	domingo :186 quarta-feira :183 quinta-feira :182 sábado :185 segunda-feira:187 sexta-feira :190 terça-feira :189
Weekend Indicator	**Qualitativa** (2 categorias – boleano)	Mode :logical FALSE:931 TRUE :371
Round Number	**Qualitativa** (31 categorias)	5 : 91 8 : 91

		1 : 62 24 : 53 4 : 49 (Other):918 NA's : 38
Round Time Indicator	**Qualitativa (3 categorias)**	
Round Includes Classic Match	**Qualitativa** (2 categorias – boleano)	Mode :logical TRUE: 28 FALSE:1274
Is Round Publication Date	**Qualitativa** (2 categorias – boleano)	Mode :logical TRUE: 153 FALSE:1149
Is Before Game Starts	**Qualitativa** (2 categorias – boleano)	Mode :logical TRUE: 19 FALSE :1283
Is After Game Ends	**Qualitativa (2 categorias – boleano)**	Mode :logical TRUE: 19 FALSE :1283
Is Winter Transfer Season	**Qualitativa** (2 categorias – boleano)	Mode :logical TRUE: 102 FALSE :1200
Turn	**Qualitativa (2 categorias)**	
Turn Indicator	**Qualitativa**	

	(3 categorias)	