Análise turn over da Empresa X[1]

O presente trabalho tem como objetivo identificar os principais motivos de *turn over* da empresa fícticia X, representada pelo banco de dados presente em **data-raw/lista_de_funcionários.csv**. O trabalho está dividido nas seguintes seções:

- Análise prévia do banco de dados
- Motivações do Turn Over da empresa X
- Considerações Finais

Análise prévia do banco de dados

O banco de dados *lista_de_funcionarios.csv* apresenta informações sobre 14.999 funcionários da empresa X, sejam estes funcionários ativos na empresa ou funcionários desligados. As variáveis presentes nesse banco são de dois tipos distintos:

- Variáveis numéricas: que expressam quantidade e;
- Variáveis categóricas: que representam qualidades.

Tabela 1: Sumário das Variáveis Numéricas

vars	media	std	min	max
satisfacao	0.61	0.25	0.09	1
ultima_avaliacao	0.72	0.17	0.36	1
volume_projetos	3.8	1.23	2	7
media_horas_mensais	201.05	49.94	96	310
tempo_cia	3.5	1.46	2	10
acidente_trabalho	0.14	0.35	0	1
saiu	0.24	0.43	0	1
promocao_ultimos_5_anos	0.02	0.14	0	1

[1] Arquivo gerado via R Markdown presente em ./avaliacao-take/R/relatorio-turn-over.Rmd. Arquivo formatado em relação a compilação original.

As variáveis satisfacao e ultima_avaliacao representam indices que variam entre 0 e 1. Ambos as variáveis possuem uma polaridade positiva. Nota-se que tanto em nível de satisfação dos funcionários quanto a última avaliação apresentam índices médios acima de 0,5.

Volume de projetos varia entre 2 e 7, sendo que em média cada funcionário participa de cerca de 4 projetos, o que sugere ampla participação dos funcionários nas atividades da empresa. Em termos de media de horas trabalhadas mensal a média é 201 horas. Por fim, em média as pessoas ficam 3,49 anos nessa empresa, 14,5% sofreu um acidente de trabalho, 23,8% saiu da empresa e 2% recebeu uma promoção nos últimos 5 anos.

Tabela 2: Frequência da variável Comercial

comercial	N_pessoas	porcent
comercial	4140	27.60
contabilidade	767	5.11
RH	739	4.93
area tecnica	2720	18.13
suporte	2229	14.86
gerencia	630	4.20
TI	1227	8.18
produtos	902	6.01
marketing	858	5.72
P&D	787	5.25

No que diz respeito as áreas de atividade dos funcionários, a maioria está presente nas áreas comercial, área técnica e suporte.

Tabela 3: Frequência da variável Salário

salario	N_pessoas	porcent
baixo	7316	48.78
mediano	6446	42.98
alto	1237	8.25

Os salarios nesse banco de dados também é considerado uma variável categórica. A maioria dos funcionários ganha um salário baixo. Menos de 10% recebe um salário considerado alto.

Motivações do Turn Over da empresa X

A abordagem utilizada para identificar os motivos de turn over da empresa X, será, a priori, não estipular nenhuma hipótese arbitrária de ínicio, mas sim permitir que os próprios dados contem sua história. Em outras palavras, ao invés de levantar hipóteses e posteriormente valida-las junto ao banco de dados, sugere-se organizar as informações de modo ao própio banco de dados apresentar os padrões que caracterizam o turn over da empresa X.

Para isso, faz-se necessários alguns tratamentos iniciais no banco de dados.

1. Categorizar as variáveis numéricas

As variáveis numéricas satisfação e ultima_avaliacao apresentam um grande intervalo de valores. Contudo é possível categoriza-las em intervalos e qualificá-las em termos de ruim, intermediário e excelente. O código abaixo apresenta essa categorização:

```
funcionarios[, satisfacao_categorica := ifelse(satisfacao< 0.5, "Satisfacao: Ruim",
                                ifelse(satisfacao>=0.5 & satisfacao<0.8, "Satisfacao: Intermediário",
                                    "Satisfacao: Excelente"))]

funcionarios[, avaliacao_categorica := ifelse(ultima_avaliacao< 0.5, "Avaliacao: Ruim",
                                ifelse(ultima_avaliacao>=0.5 & ultima_avaliacao<0.8, "Avaliacao: Intermediário",
                                        "Avaliacao: Excelente"))]
```

Por sua vez, é necessário categorizar a variável média de horas mensais, que apresenta mais intervalos:

```
funcionarios[, media_horas_categorica := ifelse(media_horas_mensais< 100, "Horas: Abaixo de 100",
                                ifelse(media_horas_mensais >= 100 & media_horas_mensais <150, "Horas: Entre 100 e 150",
                                ifelse(media_horas_mensais >= 150 & media_horas_mensais <200, "Horas: Entre 150 e 200",
                                ifelse(media_horas_mensais >= 200 & media_horas_mensais <250, "Horas: Entre 200 e 250",
                                    "Horas: Acima de 250")))))]
```

As demais variáveis numéricas a saber, volume_projetos e tempo_cia não foram categorizadas dado que a quantidade de valores distintos dessas variáveis não é significativa.

2.Novo banco de funcionários agregado

Uma vez categorizadas essas variáveis, o próximo passo é agregar o número de funcionários por todas as variáveis categoricas. A ideia é encontrar o número de funcionários que sairam da empresa e que permanecem na empresa pelas diversas combinações possíveis dessas variáveis categoricas. Abaixo o comando que gera essa agregação:

```
# funcionarios_agreg: obtem-se o número de pessoas pelas diversas combinações de valores das variáveis categoricas
funcionarios_agreg = funcionarios[, list(freq=.N), list(satisfacao_categorica,
                          avaliacao_categorica,
                          volume_projetos = paste("Volume:", volume_projetos),
                          media_horas_categorica,
                          tempo_cia = paste("Tempo_cia:", tempo_cia),
                          acidente_trabalho = paste("Acidente:", acidente_trabalho),
                          promocao_ultimos_5_anos = paste("Promoção:", promocao_ultimos_5_anos),
                          comercial = paste("Comercial:", comercial),
                          salario = paste("Salario", salario),
                          saiu)]

# Modifica a variável saiu para termos de SIM e NAO
funcionarios_agreg[, saiu:= ifelse(saiu==1, "SIM", "NAO")]

formula = "satisfacao_categorica + avaliacao_categorica + volume_projetos + media_horas_categorica + tempo_cia + acide
nte_trabalho + promocao_ultimos_5_anos + comercial + salario ~ saiu"

# modifica funcionarios agregado de modo a, para cada combinação das variáveis categoricas, apresentar o número
# de pessoas que sairam ou permanecem na empresa

funcionarios_agreg = dcast(funcionarios_agreg, formula, value.var = "freq", fill=0)

if(funcionarios_agreg[, sum(SIM) + sum(NAO)]==nrow(funcionarios)){
  print("Sucesso na agregação.")
} else{ stop("Erro na agregação")}

## [1] "Sucesso na agregação."

funcionarios_agreg = funcionarios_agreg[order(-SIM)]

kable(funcionarios_agreg[1:3, ], caption = "Tabela 4: Banco funcionário agregados pelas variáveis categoricas")
```

Tabela 4: Banco funcionário agregados pelas variáveis categoricas

satisfacao categorica	avaliacao categorica	volume projetos	media horas categorica	tempo cia	acidente trabalho	promocao ultimos 5 anos	comercial	salario	NAO	SIM
Satisfacao: Ruim	Avaliacao: Intermediário	Volume: 2	Horas: Entre 100 e 150	Tempo_cia: 3	Acidente: 0	Promoção: 0	Comercial: comercial	Salario baixo	2	131
Satisfacao: Ruim	Avaliacao: Intermediário	Volume: 2	Horas: Entre 100 e 150	Tempo_cia: 3	Acidente: 0	Promoção: 0	Comercial: suporte	Salario baixo	2	78
Satisfacao: Ruim	Avaliacao: Excelente	Volume: 6	Horas: Acima de 250	Tempo_cia: 4	Acidente: 0	Promoção: 0	Comercial: comercial	Salario baixo	0	70

O banco agregado já apresenta, por si só, alguns padrões interessantes que explicam o turn over da empresa X. Por exemplo, analisando a primeira linha, já é possível observar que funcionários com um nível de satisfação ruim, avaliação intermediária, que trabalham entre 100 e 150 horas, sem acidentes ou promoção, da área comercial e com salário baixo, tem alta propensão a sair da empresa. De fato, foram 131 funcionários que sairam da empresa nessas condições, contra 2 que permaneceram, ou seja, cerca de 98,5% das pessoas que se encontram nessas condições saem da empresa.

Embora a análise acima seja válida, há possibilidades de melhora. Observe as linha 1 e 3 da tabela 4. Note que se considerarmos o padrão satisfação nível ruim, área comercial e salário baixo, os valores da linha 1 se combinam com a linha 3 de modo a mostrar que 201 funcionários saem da empresa nessas condições, contra apenas 2 que permanecem, ou seja, a propensão a sair é de 99%. Assim, é importante agregar mais o banco segundo um conjunto menor de variaveis categoricas. Novamente é possível automatizar esse processo, como será descrito a seguir.

3.Função geraAgregacao()

Para identificar padrões de turn over segundo um conjunto menor de variáveis categoricas no processo de agregação, foi criada a função geraAgregacao(...). Seguindo o exemplo acima, foi identificado um padrão segundo a agregação de apenas três variáveis: satisfação, área comercial e salário. A ideia da função geraAgregacao(...) é gerar todas as combinações possíveis de n elementos, e em seguida gerar todos os bancos agregados segundo todas essas combinações. Exemplificando, são 9 variáveis categóricas (satisfacao_categorica, avaliacao_categorica, volume_projetos, media_horas_categorica, tempo_cia, acidente_trabalho, promocao_ultimos_5_anos, comercial e salario), que combinadas em grupos de 3 elementos geram

84 combinações diferentes. A função geraAgrecacao(...) gera 84 bancos agregados segundo essas 84 combinações. Entretanto é necessário alguma medida de qualidade para identificar se há nesses bancos um padrão de turn over.

Há duas condições em geraAgregacao(...) que procuram identificar o padrão de turn over:

- Dada as variáveis categoricas do banco agregado gerado, a porcentam de pessoas que sairam em relação ao total geral de pessoas que sairam (porcent_sim_total);
- Dada as variáveis categoricas do banco agregado, a razão entre o número de pessoas que sairam em relação as pessoas que permanecem na empresa (razao_sim_nao);

Espera-se um padrão de turn over quando, para determinada combinação das variáveis categóricas, a quantidade de pessoas que sairam representa mais de 10% do total (porcent_sim_total>0.1) e a razão de pessoas que sairam dessa empresa é 5 vezes maior do que as pessoas que permanecem (razao_sim_nao > 5). Esses parâmetros foram determinados arbitrariamente, contudo a função permite ajustes nesses parâmetros segundo os argumentos *limiar_porcentagem_sim* e *limiar_raza_sim_nao*, respectivamente.

Abaixo a função geraAgregacao(...):

```
geraAgregacao = function(funcionarios_agreg, n_combinacoes,
                         limiar_porcentagem_sim = 0.1,
                         limiar_raza_sim_nao = 5){

 # ================================================================================
 # Gera bancos agregados segundo todas as combinações possíveis das variáveis
 # categoricas agrupadas em n_combinacoes de elementos.
 # O banco final apresenta apenas as linhas com padrão de turn over
 # segundo os limiares da porcentagem de sim em relação ao total e da razão sim não
 #
 # Ags.:
 # funcionarios_agreg -- data.table de funcionarios_agreg
 # n_combinacoes -- determina o tamanho dos grupos nas agregações
 # limiar_porcentagem_sim -- determina o padrão turn over segundo porcentagem de sim em relação ao total
 # limiar_raza_sim_nao -- determina o padrão turn over segundo a quantidade de pessoas que saem em relação as que fica
m
 #
 # Return:
 # data.table com as linhas que apresentam turn over segundo todas as combinações possiveis
 # das variáveis categóricas em n_combinacoes elementos.
 #
 #================================================================================

 variavies_id = names(funcionarios_agreg)[!names(funcionarios_agreg) %in% c("SIM", "NAO")]
```

```
 lista_combinacoes = list(combn(variavies_id, m=n_combinacoes, simplify = FALSE))[[1]]

 lista_resultados = list()

 for(combinacao in lista_combinacoes){

   dt_temp = funcionarios_agreg[, list(SIM = sum(SIM), NAO = sum(NAO)), by=eval(combinacao)]

   setnames(dt_temp, names(dt_temp)[1:n_combinacoes], paste0("V", 1:n_combinacoes))

   dt_temp[, porcent_sim_total := round(SIM/sum(SIM),2)]
   dt_temp[, razao_sim_nao := round(SIM/(NAO + 0.001),2)] # Evitar divisão por 0

   dt_temp = dt_temp[porcent_sim_total >= 0.1 & razao_sim_nao > 5, ]

   lista_resultados[[length(lista_resultados)+1]] = dt_temp
 }

 resultado = rbindlist(lista_resultados, use.names = T, fill=T)
 resultado[, propensao_sair := round((SIM*100) / (SIM + NAO), 2)]

 return(resultado)

}
```

Ao aplicar essa função com n_combinacoes=3, ordenando de forma decrescente a variável propensão a sair (SIM / (SIM+NAO)) observa-se os seguinte resultados:

Tabela 5: Padrões de Turn over considerando combinações de 3 Variáveis Categóricas

V1	V2	V3	SIM	NAO	porcent sim total	razao sim nao	propensao sair
Volume: 6	Horas: Acima de 250	Tempo_cia: 4	485	20	0.14	24.25	96.04
Satisfacao: Ruim	Volume: 2	Tempo_cia: 3	1515	113	0.42	13.41	93.06
Avaliacao: Excelente	Volume: 6	Horas: Acima de 250	479	38	0.13	12.60	92.65
Satisfacao: Excelente	Avaliacao: Excelente	Tempo_cia: 5	432	35	0.12	12.34	92.51
Volume: 2	Horas: Entre 100 e 150	Tempo_cia: 3	1014	83	0.28	12.22	92.43
Satisfacao: Ruim	Volume: 2	Horas: Entre 100 e 150	1017	86	0.28	11.83	92.20
Avaliacao: Excelente	Volume: 6	Tempo_cia: 4	481	41	0.13	11.73	92.15
Satisfacao: Ruim	Horas: Acima de 250	Tempo_cia: 4	725	72	0.20	10.07	90.97
Satisfacao: Ruim	Volume: 6	Tempo_cia: 4	556	57	0.16	9.75	90.70
Avaliacao: Ruim	Volume: 2	Tempo_cia: 3	569	60	0.16	9.48	90.46
Avaliacao: Excelente	Horas: Acima de 250	Tempo_cia: 5	370	41	0.10	9.02	90.02
Satisfacao: Ruim	Volume: 6	Horas: Acima de 250	543	63	0.15	8.62	89.60
Volume: 6	Horas: Acima de 250	Salario baixo	341	43	0.10	7.93	88.80
Satisfacao: Ruim	Volume: 2	Salario baixo	940	120	0.26	7.83	88.68
Satisfacao: Ruim	Avaliacao: Ruim	Volume: 2	568	74	0.16	7.68	88.47
Volume: 6	Tempo_cia: 4	Salario baixo	352	47	0.10	7.49	88.22
Volume: 2	Horas: Entre 100 e 150	Salario baixo	621	83	0.17	7.48	88.21
Avaliacao: Excelente	Volume: 5	Tempo_cia: 5	422	60	0.12	7.03	87.55
Satisfacao: Ruim	Horas: Entre 100 e 150	Tempo_cia: 3	1019	147	0.29	6.93	87.39
Satisfacao: Ruim	Volume: 2	Acidente: 0	1461	213	0.41	6.86	87.28
Satisfacao: Ruim	Avaliacao: Intermediário	Volume: 2	965	144	0.27	6.70	87.02
Satisfacao: Ruim	Volume: 2	Comercial: comercial	465	70	0.13	6.64	86.92
Avaliacao: Ruim	Volume: 2	Horas: Entre 100 e 150	390	60	0.11	6.50	86.67
Volume: 2	Tempo_cia: 3	Salario baixo	941	150	0.26	6.27	86.25
Avaliacao: Excelente	Horas: Entre 200 e 250	Tempo_cia: 5	387	62	0.11	6.24	86.19
Satisfacao: Ruim	Volume: 2	Promoção: 0	1522	254	0.43	5.99	85.70
Volume: 6	Horas: Acima de 250	Acidente: 0	518	87	0.15	5.95	85.62
Satisfacao: Ruim	Avaliacao: Excelente	Horas: Acima de 250	705	119	0.20	5.92	85.56
Volume: 6	Tempo_cia: 4	Acidente: 0	533	91	0.15	5.86	85.42
Volume: 6	Horas: Acima de 250	Promoção: 0	546	96	0.15	5.69	85.05
Volume: 2	Horas: Entre 100 e 150	Acidente: 0	975	177	0.27	5.51	84.64
Avaliacao: Intermediário	Volume: 2	Horas: Entre 100 e 150	627	117	0.18	5.36	84.27
Satisfacao: Ruim	Avaliacao: Excelente	Volume: 6	535	100	0.15	5.35	84.25
Satisfacao: Ruim	Avaliacao: Excelente	Tempo_cia: 4	701	131	0.20	5.35	84.25
Volume: 2	Tempo_cia: 3	Acidente: 0	1455	272	0.41	5.35	84.25
Volume: 6	Tempo_cia: 4	Promoção: 0	559	106	0.16	5.27	84.06
Volume: 2	Horas: Entre 100 e 150	Promoção: 0	1014	200	0.28	5.07	83.53
Satisfacao: Excelente	Tempo_cia: 5	Acidente: 0	414	82	0.12	5.05	83.47
Satisfacao: Ruim	Volume: 2	Salario mediano	548	109	0.15	5.03	83.41
Avaliacao: Intermediário	Volume: 2	Tempo_cia: 3	952	190	0.27	5.01	83.36

A primeira linha sugere um padrão que, funcionários com altas cargas de trabalho (volume de projetos = 6 e horas média de trabalho acima de 250) e 4 anos de tempo na empresa tem alta propensão a sair (96,04%). De fato, uma alta carga de trabalho diário ao longo de 4 anos, pode vir a gerar um cansaço em demasia no funcionário que decide optar por opções no mercado de

trabalho mais favoráveis. A linha 3 sugere que essa justificativa pode ser estendida mesmo para funcionários com uma avaliação excelente. A linha 4 indica que mesmo com um nível de satisfação e avaliação excelente, pessoas com 5 anos de empresa tem alta propensão a sair (92,51%). Uma justificativa seria a necessidade de oxigenar conhecimentos e ideias após 5 anos de trabalho em uma nova empresa.

De forma geral, as 40 observações dessa tabela apresentam altas propensões de turn over (acima de 83%). Entretanto algumas observações são mais justificáveis que outras. Por exemplo, a linha 2 indica que um funcionário insatisfeito, com um baixo volume de trabalho (volume_projetos=2) e 3 anos de companhia tem 93,06% de chances de sair. Entretanto, nesse caso, talvez seja mais adequado analisar esses resultados na presença de mais variáveis categóricas.

Ao considerar todas as possíveis combinações das variáveis categóricas em grupos de 4 elementos, obtem-se os seguinte resultados:

Tabela 6: Padrões de Turn over considerando combinações de 4 Variáveis Categóricas

V1	V2	V3	V4	SIM	NAO	porcent sim total	razao sim nao	propensao sair
Satisfacao: Ruim	Volume: 6	Horas: Acima de 250	Tempo_cia: 4	485	9	0.14	53.88	98.18
Avaliacao: Excelente	Volume: 6	Horas: Acima de 250	Tempo_cia: 4	428	8	0.12	53.49	98.17
Volume: 6	Horas: Acima de 250	Tempo_cia: 4	Acidente: 0	459	17	0.13	27.00	96.43
Volume: 6	Horas: Acima de 250	Tempo_cia: 4	Promoção: 0	482	20	0.13	24.10	96.02
Satisfacao: Ruim	Volume: 2	Horas: Entre 100 e 150	Tempo_cia: 3	1014	43	0.28	23.58	95.93
Satisfacao: Ruim	Avaliacao: Excelente	Volume: 6	Tempo_cia: 4	478	21	0.13	22.76	95.79
Satisfacao: Ruim	Avaliacao: Excelente	Horas: Acima de 250	Tempo_cia: 4	630	32	0.18	19.69	95.17
Satisfacao: Ruim	Volume: 2	Horas: Entre 100 e 150	Salario baixo	621	32	0.17	19.41	95.10

As 4 primeiras linhas dessa tabela apresentam um detalhamento maior a situação de alta carga de trabalho (volume_de_projetos 6 e horas média de trabalho acima de 250) e tempo de empresa 4 anos. Essas linhas mostram que para esse padrão, se o funcionário estiver insatisfeito (linha 1), ou mesmo com uma excelente avaliação (linha 2) e não tenha obtido uma promoção (linha 4) estes possuem alta propensão a sair (acima de 96%). Essa evidência corrobora com a hipótese que muito trabalho incentiva a evasão do funcionário.

As linhas 6 e 7 indicam que funcionários insatisfeitos, que trabalham muito (volume_projetos 6 ou horas acima de 250) e tempo de companhia de 4 anos tendem a sair com uma propensão de cerca de (95%). Mesmo sendo bem avaliados, esses funcionários apresentam uma insatisfação com a alta carga de trabalho o que justifica sua alta propensão em sair.

As linhas 5 e 8 sugerem que funcionários insatisfeitos, que trabalham relativamente pouco (volume de projetos 2 e horas entre 100 e 150) tem altas propensões a sair se ganham um salário baixo (linha 8), ou se possuem 3 anos de companhia. Talvez uma explicação para isso seja a baixa participação junto a empresa, que traz instabilidades com relação a certeza de permanência junto a empresa, o que estimula a evasão.

Considerações Finais

Esse trabalho apresentou uma proposta para explicar as possíveis razões de turn over da empresa X. As razões que apresentaram mais destaques foram:

- Altas cargas de trabalho aumentam a propensão a sair;
- Bons profissionais com 5 anos de empresa tendem a sair para novas oportunidades;
- Funcionários insatisfeitos com poucas atividades na empresa são mais propensos a sair.

Além de 40 padrões distintos apresentados na tabela 5 que merecem uma avaliação mais pormenorizada. Para novos desenvolvimentos, sugere-se categorizar as demais variáveis numéricas e testar mais combinações das variáveis categóricas na busca de novos padrões.