Análise de Correlação e Associação

Charles Guimarães Cavalcante – RM 334409
Luan Nonato Figueiredo – RM 334325
Rodrigo Rossi de Lima Cano – RM 333927

Insights a partir a geração de modelos preditivos em bases de dados

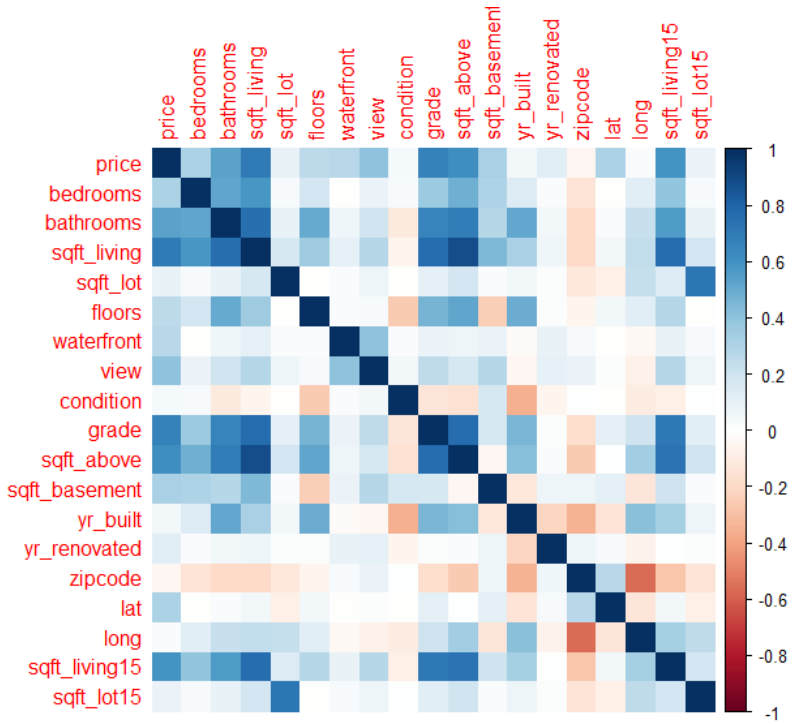
Base de dados 1 – Dados sobre valores de imóveis na cidade de Boston, MA, EUA

1.1) Considerando que a variável de interesse seja o preço ou valor do imóvel, quais são as variáveis que mais explicam o comportamento do preço dos imóveis? Comente e justifique seus insights.

```
# leitura dos dados
boston <- read.csv(file="../input/Boston_Housing_Data.csv")

# remoção dos campos id e data
boston <- boston[, c(-1,-2)]

# gráfico de correlação das variáveis
corrplot(round(cor(boston), 2), method = "color")
```



De acordo com o gráfico “price” tem boa correlação com: bedrooms, bathrooms, view, grade, sqft_above, sqft_basement, lat, sqft_living, sqft_living15.

1.2) Existem variáveis redundantes presentes na base de dados? Se sim, quais as análises que você realizou para chegar a essa conclusão? Em caso de existência de redundância, quais foram as variáveis e como você endereçou o problema?

Das variáveis selecionadas foram testadas agrupadas por semelhança.

Grupo 1: bedrooms, bathrooms

```
cor(boston$bedrooms, boston$bathrooms)
```

Resultado: 0.5158836 – correlação moderada

Grupo 2: view, grade

```
cor(boston$view, grade$bathrooms)
```

Resultado: 0.2513206 – correlação fraca

Grupo 3: sqft_living, sqft_above, sqft_basement, sqft_living15

```
cor(boston$sqft_living, boston$sqft_above)
```

Resultado: 0.8765966 – correlação muito forte

```
cor(boston$sqft_living, boston$sqft_basement)
```

Resultado: 0.435043 – correlação moderada

```
cor(boston$sqft_living, boston$sqft_living15)
```

Resultado: 0.7564203 – correlação forte

```
cor(boston$sqft_above, boston$sqft_basement)
```

Resultado: -0.05194331 – correlação muito fraca

```
cor(boston$sqft_above, boston$sqft_living15)
```

Resultado: 0.7318703 – correlação forte

```
cor(boston$sqft_basement, boston$sqft_living15)
```

Resultado: 0.200355 – correlação fraca

As variáveis **sqft_living** e **sqft_above** tem correlação muito forte. As variáveis

sqft_living e **sqft_living15**, e as variáveis **sqft_living15** e **sqft_above** tem correlação forte.

1.3) Construa um modelo preditivo que explique o preço dos imóveis. Quais variáveis entraram no modelo final? Exiba a matriz de parâmetros e interprete os resultados. Qual o nível de acurácia do modelo? Justifique a métrica utilizada e interprete o resultado.

Primeiro selecionamos a matrix somente com as variáveis selecionadas: price, bedrooms, bathrooms, sqft_living, view, grade, sqft, basement.

```
data = boston[,c(1, 2, 3, 4, 8, 10, 12)]
```

A seguir, separamos os dados em dados de treino (80%) e teste (20%):

```
set.seed(41) # semente para reproduzir os mesmos dados
sample = sample.split(data, SplitRatio=0.8) # separação em 80%
train_data = subset(data, sample==TRUE) # dados de treino
test_data = subset(data, sample==FALSE) # dados de teste
```

Criação do modelo de predição com os dados de treino:

```
modelo <- lm(price ~ ., data=train_data)
```

Predição e teste de acurácia:

```
predicao <- predict(newdata=test_data, modelo)
teste <- data.frame(actual=test_data$price, predicted=predicao)
media <- mean(abs(teste$actual-teste$predicted) / teste$actual)
acuracia <- 1 - media
```

Resultado: com o modelo proposto chegamos a uma acurácia para predição do preço do imóvel de **68,1%**.

Base de dados 2 – Dados sobre rotatividade de funcionários de uma empresa

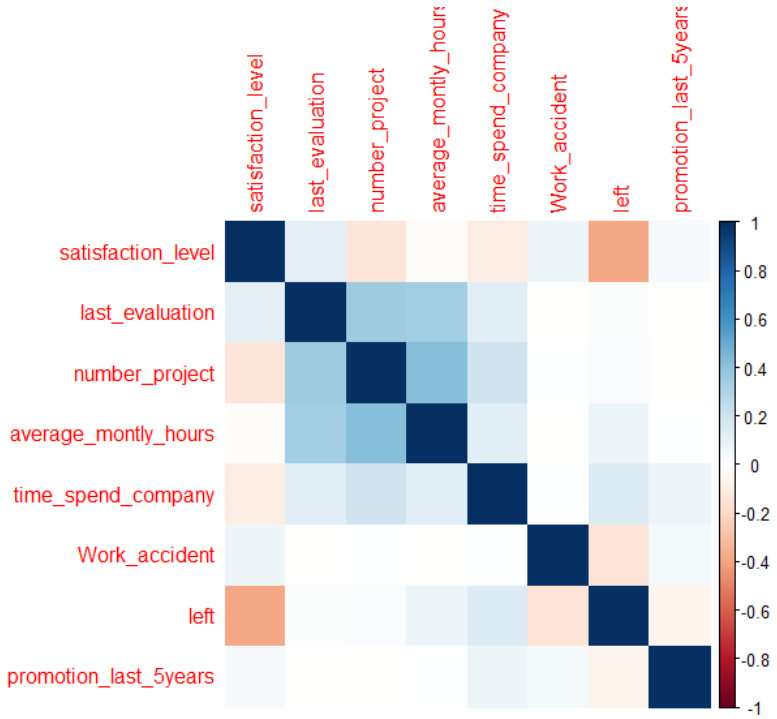
Considere a tabela HR_Analytics.xlsx, que traz dados de quase 15.000 empregados de uma empresa, incluindo dados sobre satisfação, desempenho, participação em projetos, registro de promoção, entre outros elementos. A ideia é buscar compreender os fatores que mais influenciam na saída do funcionário da empresa.

2.1) Considerando que a variável de interesse seja a variável "left", que indica se o funcionário saiu (left=1) ou não (left=0), qual metodologia ou técnica poderia ser mais adequada para entender o perfil de rotatividade dos funcionários?

```
# leitura dos dados
hr <- read_excel("HR_Analytics.xlsx")

# remoção dos campos irrelevantes
hr <- hr[, c(-1)]

# gráfico de correlação das variáveis
corrplot(round(cor(hr[, c(-9, -10)]), 2), method = "color")
```



Podemos notar que a variável "left" tem uma relação muito negativa com o índice de satisfação (satisfacion_level).

2.2) Faça uma análise bivariada de cada uma das variáveis que potencialmente influenciam a saída do funcionário. Calcule o "IV" (Information Value) para cada uma das análises, interprete as análises de cada variável e ordene, por grau de importância, as variáveis que mais explicam a rotatividade dos empregados.

Primeiro fizemos a categorização das variáveis contínuas:
satisfaction_level, last_evaluation e average_montly_hours.

```
satisfaction_level_cl <- discretize(hr$satisfaction_level,
"frequency", breaks=4)
last_evaluation_cl <- discretize(hr$last_evaluation,
"frequency", breaks=4)
average_montly_hours_cl <- discretize(hr$average_montly_hours,
"frequency", breaks=4)
hr <- data.frame(hr, satisfaction_level_cl, last_evaluation_cl,
average_montly_hours_cl)
```

Cálculo de IV (excluindo os campos que foram categorizados):

```
IV <- create_infotables(data = hr[, c(-1,-2,-4)], y = "left)
print(head(IV$Summary, 100), row.names = FALSE)
```

Resultado:

```
               Variable         IV
         number_project 1.97240680
   satisfaction_level_cl 1.07808144
      time_spend_company 0.92658691
 average_montly_hours_cl 0.57291526
      last_evaluation_cl 0.44550034
           Work_accident 0.18535538
                  salary 0.17904981
                   sales 0.03561297
   promotion_last_5years 0.03385306
```

Vamos realizar a análise bivariada com as seguintes variáveis:

- number_project
- satisfaction_level_cl
- time_spend_company
- average_montly_hours_cl
- last_evaluation_cl

```
CrossTable(hr$number_project, hr$left, prop.r=FALSE,
prop.t=FALSE, prop.chisq=FALSE)
```

hr$number_project	hr$left 0	1	Total
2	821	1567	2388
	0.072	0.439	
3	3983	72	4055
	0.349	0.020	
4	3956	409	4365
	0.346	0.115	
5	2149	612	2761
	0.188	0.171	
6	519	655	1174
	0.045	0.183	
7	0	256	256
	0.000	0.072	
Total	11428	3571	14999
	0.762	0.238	

A variável **number_project** demonstra que 43,9% das pessoas que deixaram a empresa trabalhavam em apenas dois projetos.

```
CrossTable(hr$satisfaction_level_cl, hr$left, prop.r=FALSE,
prop.t=FALSE, prop.chisq=FALSE)
```

hr$satisfaction_level_cl	hr$left 0	1	Total
[0.09,0.44)	1521	2153	3674
	0.133	0.603	
[0.44,0.64)	3201	459	3660
	0.280	0.129	
[0.64,0.82)	3422	461	3883
	0.299	0.129	
[0.82,1]	3284	498	3782
	0.287	0.139	
Total	11428	3571	14999
	0.762	0.238	

A variável **satisfation_level** demonstra que 60% das pessoas que deixaram a empresa tinham índice baixo de satisfação.

```
CrossTable(hr$time_spend_company, hr$left, prop.r=FALSE,
prop.t=FALSE, prop.chisq=FALSE)
```

hr$time_spend_company	hr$left 0	1	Total
2	3191	53	3244
	0.279	0.015	
3	4857	1586	6443
	0.425	0.444	
4	1667	890	2557
	0.146	0.249	
5	640	833	1473
	0.056	0.233	
6	509	209	718
	0.045	0.059	
7	188	0	188
	0.016	0.000	
8	162	0	162
	0.014	0.000	
10	214	0	214
	0.019	0.000	
Total	11428	3571	14999
	0.762	0.238	

A variável **time_spend_company** demonstra que a maioria das pessoas que deixaram a empresa trabalhavam de 3 a 5 horas por dia.

```
CrossTable(hr$average_montly_hours_cl, hr$left, prop.r=FALSE,
prop.t=FALSE, prop.chisq=FALSE)
```

hr$average_montly_hours_cl	hr$left 0	1	Total
[96,156)	2354	1326	3680
	0.206	0.371	
[156,200)	3456	330	3786
	0.302	0.092	
[200,245)	3227	491	3718
	0.282	0.137	
[245,310]	2391	1424	3815
	0.209	0.399	
Total	11428	3571	14999
	0.762	0.238	

A variável **average_monthly_hours** demonstra que 37,1% das pessoas que deixam a empresa trabalhavam menos do que 156 horas por mês deixaram a empresa. Porém demonstra também que 39,9% com mais de 245 horas também deixaram a empresa.

```
CrossTable(hr$last_evaluation_cl, hr$left, prop.r=FALSE,
prop.t=FALSE, prop.chisq=FALSE)
```

```
================================================
                         hr$left
hr$last_evaluation_cl         0       1   Total
------------------------------------------------
[0.36,0.56)                2259    1348    3607
                          0.198   0.377
------------------------------------------------
[0.56,0.72)                3456     330    3786
                          0.302   0.092
------------------------------------------------
[0.72,0.87)                3099     662    3761
                          0.271   0.185
------------------------------------------------
[0.87,1]                   2614    1231    3845
                          0.229   0.345
------------------------------------------------
Total                     11428    3571   14999
                          0.762   0.238
================================================
```

A variável **last_evaluation** demonstra que 37,7% das pessoas que deixam a empresa tiveram índice baixo na última avaliação. Porém demonstra também que 34,5% com índice muito elevado também deixaram a empresa.

2.3) Construa um modelo preditivo que explique a rotatividade dos empregados. Quais variáveis entraram no modelo final? Exiba a matriz de parâmetros e interprete os resultados. Qual o nível de acurácia do modelo? Justifique a métrica utilizada e interprete o resultado.

Separação dos dados que serão utilizados:

```
data = hr[, c(3, 5, 7, 10, 11, 12)]
```

A seguir, separamos os dados em dados de treino (80%) e teste (20%):

```
set.seed(41) # semente para reproduzir os mesmos dados
sample = sample.split(data, SplitRatio=0.8) # separação em 80%
train_data = subset(data, sample==TRUE) # dados de treino
test_data = subset(data, sample==FALSE) # dados de teste
```

Criação do modelo de predição de regressão logística com os dados de treino:

```
modelo <- glm(left ~ .,
              family=binomial(link='logit'),
              data=train_data)
```

Predição e teste de acurácia:

```
predicao <- predict(modelo, newdata=test_data, type="response")

ks.test(predicao[test_data$left==0],predicao[test_data$left==1])
```

```
        Two-sample Kolmogorov-Smirnov test

data:  predicao[test_data$left == 0] and predicao[test_data$left == 1]
D = 0.59795, p-value < 0.00000000000000022
alternative hypothesis: two-sided
```

Resultado: com o modelo proposto chegamos a uma acurácia para predição de saída do funcionário de **59,7%**.