Caso Junglivet Whisky Company

Introdução à CRISP-DM, RStudio, Sistema Quarto e Linguagem R

Seu Nome

16 de abril de 2025

Índice


```
# digitos exibidos
options(digits = 4, scipen = 999)

# Carrega os pacotes usados
library(here)
library(tidyverse)
library(skimr)
```

1 Introdução à Metodologia CRISP-DM

A metodologia **CRISP-DM** (Cross Industry Standard Process for Data Mining) é um modelo de processo amplamente utilizado para projetos de análise de dados e ciência de dados. Desenvolvido por um consórcio de mais de 200 organizações no

final dos anos 90, tornou-se o padrão mais utilizado para projetos de analytics e ciência de dados.

O CRISP-DM divide projetos de análise de dados em seis fases inter-relacionadas:

1. **Entendimento do Negócio** (Business Understanding)
2. **Entendimento dos Dados** (Data Understanding)
3. **Preparação dos Dados** (Data Preparation)
4. **Modelagem** (Modeling)
5. **Avaliação** (Evaluation)
6. **Implantação** (Deployment)

Neste relatório, vamos aplicar as três primeiras fases do CRISP-DM a um estudo de caso fictício baseado na *Junglivet Whisky Company*.

2 O Caso Junglivet Whisky Company

Você acaba de ser contratado como analista de dados na *Junglivet Whisky Company*. A empresa está enfrentando problemas com a qualidade do whisky produzido e você foi designado para investigar possíveis causas.

Mr. Gumble, responsável pela destilaria e pela equipe de produção, recebeu reclamações sobre a qualidade do whisky e está tentando identificar a razão. Ele forneceu a você dados da linha de produção das últimas duas semanas e pediu que você os analisasse.

2.1 Fase 1: Entendimento do Negócio

Nesta fase, nosso objetivo é compreender claramente o problema de negócio que precisamos resolver. No caso da Junglivet Whisky Company, o problema é:

- **Problema de negócio**: Queda na qualidade do whisky produzido.
- **Objetivo**: Identificar possíveis causas da redução de qualidade.
- **Critério de sucesso**: Encontrar fatores que influenciam negativamente a qualidade do whisky.

2.2 Fase 2: Entendimento dos Dados

Nesta fase, exploramos os dados disponíveis para entender sua estrutura, qualidade e relações iniciais.

2.2.a Dicionário dos Dados

Um dicionário de dados é uma documentação estruturada que descreve o significado, formato, uso e relacionamentos de cada variável em um conjunto de dados.

Ele funciona como um guia essencial para compreender corretamente as informações disponíveis, garantindo que todos os usuários interpretem os dados de maneira consistente.

O arquivo de dados fornecido contém as seguintes colunas (ou variáveis):

- **DAY**: Dia da produção de um whisky.
- **MONTH**: Mês da produção de um whisky.
- **MANUFACTURER**: Nome do mestre responsável.
- **PRODUCT**: Tipo de produto (Junglivet ou Junglivet Premium).
- **SHIFT**: Indica o turno (manhã ou noite) no qual um whisky foi produzido.
- **COLOR**: Indicador de cor do whisky em uma escala entre 0 e 1.

- **MALTING**: Fornecedor dos maltes utilizados.
- **TASTING**: Indicador de qualidade baseado em testes de degustação com pontuação em uma escala entre 0-1000, sendo que quanto maior a pontuação, melhor a qualidade avaliada do whisky.

2.2.b Importação do Arquivo de Dados

O processo de importação de dados é um passo fundamental em qualquer análise. Neste caso, utilizamos duas ferramentas importantes:

1. O pacote `here` permite definir caminhos relativos ao diretório raiz do projeto, o que torna o código mais portátil e facilita o compartilhamento. Independentemente de onde o projeto esteja armazenado em diferentes computadores, o pacote `here` encontrará automaticamente os arquivos a partir da raiz do projeto.
2. O pacote `readr`, parte do tidyverse, oferece funções otimizadas para leitura de arquivos, como a `read_csv()`, que é mais rápida que a função base do R e oferece tratamento mais consistente dos tipos de dados. Além disso, ela converte automaticamente strings vazias para NA, indica o tipo de cada coluna importada e preserva os nomes das variáveis originais.

```
# Importa o arquivo de dados

## 1.1 Define o caminho relativo do arquivo no projeto RStudio
caminho                              <-
here::here("dados/brutos/productionlog_sample.csv")
```

```
## 1.2 Importa o arquivo com a funçào read_csv
dados_destilaria                     <-
readr::read_csv(caminho)
```

2.2.c Verificando a estrutura dos dados

Após importar os dados, é essencial verificar sua estrutura para entender o que temos disponível. A função `glimpse()` do pacote `dplyr` nos oferece uma visão concisa e informativa sobre:

- Quais variáveis (colunas) estão presentes no conjunto de dados.
- Qual o tipo ou classe de cada variável.
- Os primeiros valores de cada variável.
- O número total de observações (linhas).

```
# Verificar a estrutura dos dados
dplyr::glimpse(dados_destilaria)
```

```
Rows: 21
Columns: 8
$ DAY          <dbl> 1, 1, 2, 2, 3, 3, 4, 4, 5, 5, NA, 6, 6, 7, 7, 8, 8, 9, 9,…
$ MONTH        <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, NA, 4, 4, 4, 4, 4, 4, 4, 4,…
$ MANUFACTURER <chr> "Leonard", "Carlson", "Leonard", "Carlson", "Leonard", "C…
$   PRODUCT                <chr> "Junglivet",          "Junglivet Premium",          "Junglivet", "Junglivet…
$ SHIFT        <chr> "Morning", "Evening", "Morning", "Evening", "Morning", "E…
$ COLOR        <dbl> 0.27, 0.27, 0.28, 0.32, 0.32, 0.28, 0.29, 0.29, 0.33, 0.2…
$ MALTING      <chr> "Inhouse", "Burns Best Ltd.", "Inhouse",
```

```
"Inhouse", "Matr…
$ TASTING       <dbl> 895, 879,
938, 900, 917, 900, 934, 951,
852, 850, NA, 991…
```

Além de conhecer a estrutura dos dados, é importante obter estatísticas descritivas básicas para cada variável.

A função `summary()` do R base nos fornece:

- Para variáveis numéricas: medidas de posição (mínimo, máximo, média, mediana) e quartis.
- Para variáveis do tipo caractere: informações básicas como comprimento, classe e modo.
- Para variáveis categóricas (quando convertidas para fatores): contagem de ocorrências em cada categoria.
- Identificação automática de valores ausentes (NA's)

Essas informações são fundamentais para detectar possíveis problemas nos dados antes de iniciar análises mais complexas e guiar nossas decisões sobre transformações necessárias.

```
# Estatísticas resumidas
summary(dados_destilaria)
```

```
        DAY                MONTH
MANUFACTURER          PRODUCT
  Min.     :  1.0     Min.     :4
Length:21           Length:21
   1st   Qu.:   3.0          1st
Qu.:4         Class   :character
Class :character
       Median       :       5.5
Median  :4     Mode   :character
Mode  :character
```

```
 Mean   : 5.5   Mean   :4
 3rd Qu.: 8.0   3rd Qu.:4
 Max.   :10.0   Max.   :4
 NA's   :1      NA's   :1
    SHIFT                  COLOR
MALTING             TASTING
 Length:21         Min.   :0.260
Length:21          Min.   :822
 Class :character  1st Qu.:0.278
Class :character   1st Qu.:875
 Mode  :character  Median :0.300
Mode  :character   Median :926
                   Mean   :0.295
Mean   :919
                   3rd Qu.:0.310
3rd Qu.:958
                   Max.   :0.350
Max.   :999
                      NA's   :1
NA's   :1
```

A saída da função mostra que há pelo menos um valor faltante (`NA`) em uma linha (observação) do arquivo.

2.3 Fase 3: Preparação dos Dados

Nesta fase, preparamos os dados para análise, renomeando variáveis, convertendo cada variável para um tipo ou classe de dados adequado, tratando valores ausentes, removendo colunas irrelevantes e garantindo que temos dados de qualidade para trabalhar.

O código a seguir executa as seguintes operações para limpar os dados:

1. **Remove** a coluna `MONTH` que é desnecessária para a análise.
2. **Renomeia** todas as colunas para nomes mais descritivos em português, facilitando a interpretação.
3. **Converte** as variáveis para seus tipos/classes de dados apropriados: `numeric`

para valores quantitativos (dia, cor, indicador_qualidade) e `factor` para variáveis categóricas (fabricante, tipo_produto, turno, fornecedor_malte).

4. Remove linhas com valores ausentes para garantir a integridade dos dados nas análises subsequentes.

Vamos utilizar o operador pipe (`%>%`) do tidyverse para encadear as operações de limpeza e transformação de dados de forma mais legível. Cada operação recebe o resultado da anterior e aplica uma nova transformação.

```
# Observe como organizamos o código com identação consistente e
# comentários explicativos para cada operação. Esta é uma boa prática
# que torna o código mais legível e facilita sua manutenção.

# Cria uma nova data frame com os dados limpos e transformados
dados_destilaria_limpos <- dados_destilaria %>%

# 1. Remove a coluna MONTH por ser redundante ou desnecessária
select(-MONTH) %>%

# 2. Renomeia as colunas para nomes mais descritivos em português
rename(
  dia = DAY,
  mestre_responsavel = MANUFACTURER,
  tipo_produto = PRODUCT,
  turno = SHIFT,
  cor = COLOR,
  fornecedor_malte = MALTING,
  indicador_qualidade = TASTING,
) %>%

# 3. Converte cada variável para seu tipo/classe adequado
mutate(
      dia = as.numeric(dia), # converte para tipo numeric
        mestre_responsavel = as.factor(mestre_responsavel), # converte para classe factor
           tipo_produto = as.factor(tipo_produto), # converte para classe factor
    turno = as.factor(turno), # converte para classe factor
      cor = as.numeric(cor), # converte para tipo numeric
         fornecedor_malte = as.factor(fornecedor_malte), # converte para classe factor
        indicador_qualidade = as.numeric(indicador_qualidade) # converte para tipo numeric
) %>%

# 4. Remove linhas com valores ausentes (NA)
drop_na()
```

2.4 Análise Exploratória de Dados

A Análise Exploratória de Dados (AED) é uma abordagem fundamental que nos permite investigar e compreender as características principais de um conjunto de dados.

Utilizamos técnicas visuais e estatísticas para:

- Identificar padrões, tendências e relações entre variáveis.
- Detectar valores atípicos (outliers) e anomalias.
- Verificar hipóteses preliminares sobre possíveis causas do problema.

- Orientar análises mais detalhadas e modelagens futuras.

Com os dados devidamente preparados, vamos explorar as relações entre algumas variáveis e o indicador de qualidade do whisky para identificar potenciais fatores que explicam os problemas enfrentados pela destilaria.

2.4.a Visualizações com ggplot2

Para as visualizações a seguir, utilizamos o pacote `ggplot2` (parte do `tidyverse`), que implementa a "Gramática dos Gráficos" – um sistema coerente para descrever e construir gráficos.

A estrutura básica de um gráfico com `ggplot2` segue o padrão:

1. Iniciar com `ggplot()` e definir os dados e mapeamentos estéticos (`aes()`).
2. Adicionar camadas com geometrias (`geom_*`) para representar os dados.
3. Personalizar com temas, títulos e outras configurações.

2.4.b Investigando relação entre fornecedor e qualidade

O boxplot ou diagrama de caixa (Figura 1), é uma ferramenta poderosa para visualizar a distribuição de variáveis numéricas agrupadas por categorias.

Neste gráfico:

- A linha horizontal dentro da caixa representa a **mediana** (percentil 50).
- Os limites inferior e superior da caixa representam o **primeiro quartil** (percentil 25) e o **terceiro quartil** (percentil 75), respectivamente.
- As "hastes" (whiskers) se estendem até 1,5 vezes o intervalo interquartil (IQR).
- Pontos individuais além das hastes representam **outliers** (valores atípicos)

Esta visualização nos permite comparar facilmente as distribuições de qualidade entre os diferentes fornecedores de malte. Assim, ela pode revelar evidências de uma relação consistente entre o fornecedor de malte e a qualidade final do produto.

```
# Boxplot comparativa da qualidade por fornecedor de malte

ggplot(dados_destilaria_limpos, aes(x = fornecedor_malte, y = indicador_qualidade)) +
      # Cria boxplots para representar a distribuição dos dados
  geom_boxplot() +
   # Aplica um tema minimalista para melhor visualização
  theme_minimal() +
  # Define títulos e rótulos dos eixos
    labs(title = "Qualidade do Whisky por Fornecedor de Malte",
         x = "Fornecedor",
              y = "Pontuação de Qualidade")
```

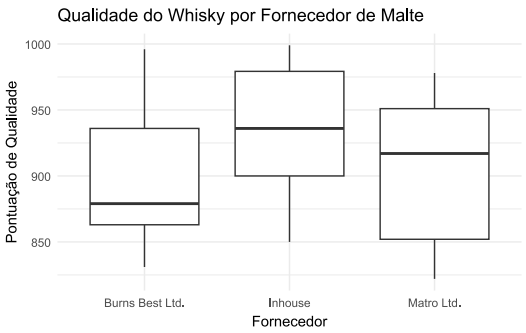

Figura 1: Boxplot comparativo entre qualidade do whisky e fornecedor do malte.

2.4.c Investigando a média de qualidade por fornecedor

Embora os boxplots forneçam uma visualização rica da distribuição dos dados, às vezes precisamos de medidas numéricas precisas para confirmar nossas observações visuais. Nesta análise:

- Agrupamos os dados por fornecedor de malte.
- Calculamos a média do indicador de qualidade para cada grupo.
- Contamos o número de amostras (n) por fornecedor para avaliar a robustez dos resultados.
- Ordenamos os resultados em ordem decrescente para facilitar a comparação.

Esta abordagem complementa a visualização anterior, fornecendo valores exatos para a tomada de decisão baseada em evidências.

```
# Calcula a qualidade média por fornecedor
dados_destilaria_limpos %>%
    # Agrupa os dados pelo fornecedor de malte
  group_by(fornecedor_malte) %>%
   # Calcula a média e conta o número de obs. para cada grupo
  summarise(
              qualidade_media   = mean(indicador_qualidade),
    n = n()
  ) %>%
  # Ordena os resultados em ordem decrescente pela qualidade média
  arrange(desc(qualidade_media))
```

```
# A tibble: 3 × 3
                fornecedor_malte
```

```
qualidade_media     n
  <fct>                    <dbl>
<int>
1 Inhouse                   935.
10
2 Matro Ltd.                 904
5
3 Burns Best Ltd.            901
5
```

2.4.d Investigando relação entre cor e qualidade

O gráfico de dispersão ou *scatter plot* (Figura 2) é ideal para explorar a relação entre duas variáveis numéricas. Ao analisar a relação entre a cor do whisky e seu indicador de qualidade:

- Cada ponto representa uma amostra de whisky produzida.
- O eixo X mostra o valor do indicador de cor.
- O eixo Y indica a pontuação de qualidade.
- A linha de tendência (curva LOESS) ajuda a visualizar o padrão geral dos dados sem assumir uma relação linear.

Esta visualização nos permite identificar se existe um valor ou faixa ótima de cor que está associada à melhor qualidade, o que poderia ser usado como indicador durante o processo de produção.

```
# Grafico de dispersão entre o indicador de cor e a qualidade

ggplot(dados_destilaria_limpos, aes(x     =     cor,     y     = indicador_qualidade)) +
   # Adiciona pontos para cada observação no conjunto de dados
  geom_point() +
  # Adiciona uma linha suavizada (LOESS) para mostrar a tendência
```

```
geral
  geom_smooth(method = "loess",
se = FALSE) +
  # Aplica um tema minimalista
  theme_minimal() +
  # Define títulos e rótulos dos
eixos
  labs(title = "Relação entre Cor
e Qualidade do Whisky",
       x = "Indicador de Cor",
            y = "Indicador de
Qualidade")
```

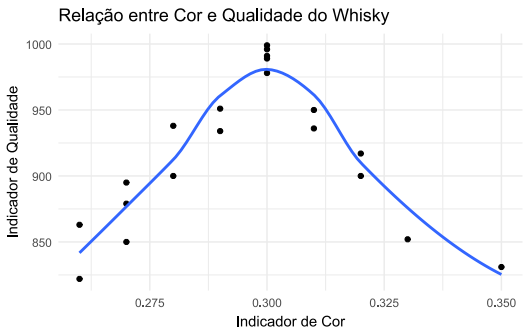

Figura 2: Gráfico de dispersão entre qualidade e cor do whisky.

2.4.e Investigando relação entre mestre responsável e qualidade

Os mestres responsáveis pela produção podem influenciar significativamente a qualidade do produto final devido às suas técnicas, experiência e atenção aos detalhes. Este boxplot nos permite:

- Comparar a performance de diferentes mestres responsáveis.
- Identificar se há diferenças consistentes na qualidade do whisky produzido por cada um.
- Verificar se algum mestre responsável apresenta maior variabilidade nos resultados.
- Detectar possíveis interações entre a experiência do profissional e a qualidade final.

Esta análise pode revelar se há necessidade de padronização de processos ou treinamentos específicos para garantir consistência na produção.

```
# Boxplot comparativo entre
indicador de qualidade e mestre
destilador

ggplot(dados_destilaria_limpos,
aes(x = mestre_responsavel, y =
indicador_qualidade)) +
  # Cria boxplots para visualizar
a distribuição e identificar
outliers
  geom_boxplot() +
  # Aplica um tema minimalista
  theme_minimal() +
  # Define títulos e rótulos dos
eixos
    labs(title = "Qualidade do
Whisky por Fabricante",
       x = "Mestre Responsável",
            y = "Indicador de
Qualidade")
```

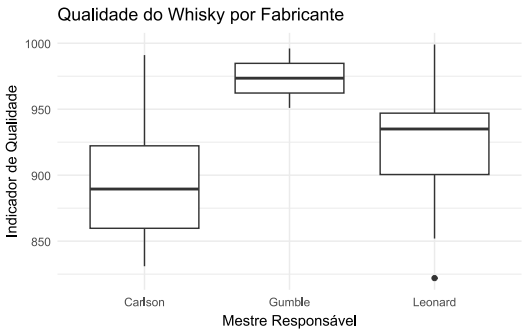

Figura 3: Boxplot comparativo entre qualidade do whisky e mestre destilador.

2.4.f Investigando relação entre turno e qualidade

O turno de trabalho (manhã ou noite) pode ter impacto significativo na produ-

ção devido a diversos fatores como fadiga, diferenças de temperatura ou supervisão.

Utilizando boxplots (Figura 4), podemos:

- Comparar visualmente a distribuição da qualidade entre os turnos.
- Identificar se um dos turnos apresenta sistematicamente qualidade inferior.
- Avaliar se a variabilidade na qualidade é maior em um turno específico.

Esta análise pode revelar possíveis problemas operacionais relacionados ao momento da produção:

```
# Boxplot comparativo entre indicador de qualidade e mestre destilador

ggplot(dados_destilaria_limpos, aes(x = turno, y = indicador_qualidade)) +
  # Cria boxplots para comparar as distribuições
  geom_boxplot() +
  # Aplica um tema minimalista
  theme_minimal() +
  # Define títulos e rótulos dos eixos
   labs(title = "Qualidade do Whisky por Turno",
       x = "Turno da Produção",
            y = "Indicador de Qualidade")
```

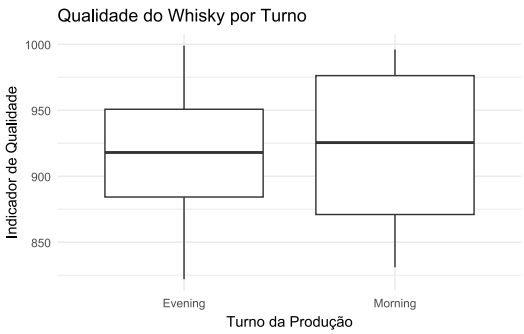

Figura 4: Boxplot comparativo entre qualidade do whisky e turno de produção.

2.5 Conclusões e Recomendações

Os resultados da análise preliminar dos dados da linha de produção da Junglivet Whisky Company, indicam que:

1. **O fornecedor de malte parece ser um fator significativo na qualidade do whisky**:

- Os whiskies produzidos com matéria-prima da “Burns Best Ltd.” tendem a ter qualidade mediana inferior.
- Os whiskies produzidos com a matéria-prima proveniente da “Matro Ltd.” apresentam grande variabilidade na qualidade da bebida, com alguns apresentando qualidade inferior aos produzidos com insumo da “Burns Best Ltd.”
- Os whiskies produzidos com matéria-prima própria (“Inhouse”) tentem a apresentar qualidade superior pelo indicador de qualidade baseado nos testes de desgustação.

2. **A cor do whisky pode ser um indicador de qualidade**:

- Os whiskies com cor em torno de 0.3 parecem ter melhor qualidade.

- Este indicador poderia ser utilizado durante a produção para detectar problemas antes da degustação final.

3. **Recomendações**:

- Reavaliar as parcerias com "Burns Best Ltd." e "Matro Ltd." ou implementar controles de qualidade mais rigorosos para matérias-primas deste fornecedor.
- Considerar a cor como um indicador antecipado de qualidade no processo de produção.
- Realizar análises adicionais sobre outros fatores como mestres responsáveis e turno de trabalho.

2.6 Próximos Passos

- Coletar mais dados e aplicar métodos estatísticos para confirmar os padrões identificados.
- Investigar se há interações entre os diferentes fatores (ex: fornecedor de malte e mestres responsáveis)
- Desenvolver um modelo preditivo de qualidade para uso durante a produção.
- Implementar um dashboard de monitoramento para controle de qualidade em tempo real.

3 Reflexão sobre o Processo CRISP-DM

Este caso demonstra a importância das três primeiras fases do CRISP-DM:

1. **Entendimento do Negócio**: Identificar claramente o problema de qualidade do whisky.
2. **Entendimento dos Dados**: Explorar os dados disponíveis e sua estrutura.
3. **Preparação dos Dados**: Limpar e transformar os dados para análise

As fases seguintes seriam:

4. **Modelagem**: Desenvolver modelos preditivos de qualidade
5. **Avaliação**: Testar a eficácia dos modelos desenvolvidos
6. **Implantação**: Implementar soluções baseadas nos insights e modelos

O CRISP-DM proporciona uma abordagem estruturada que garante que o projeto de análise de dados atenda às necessidades de negócio e gere resultados acionáveis.

4 Reprodutibilidade com RStudio, Quarto e R

Um dos principais benefícios do sistema Quarto é a **reprodutibilidade**. Este relatório combina:

- **Texto narrativo** que explica a análise, estruturado seguindo uma metodologia estabelecida (CRISP-DM).
- **Código R executável** que realiza a análise.
- **Resultados e visualizações** gerados automaticamente a partir do código

Isso significa que qualquer pessoa com acesso aos mesmos dados pode executar este documento e obter exatamente os mesmos resultados. Além disso, caso os dados sejam atualizados, basta recompilar o documento para atualizar toda a análise.

Esta abordagem reduz erros, facilita a revisão e promove a transparência no processo analítico.

Além disso, este relatório demonstra como podemos usar o conjunto de ferramentas RStudio, Quarto e a linguagem R para:

1. **Estruturar** um processo analítico seguindo uma metodologia estabelecida.
2. **Documentar** todo o fluxo de trabalho, desde a importação até as conclusões.
3. **Comunicar** resultados de forma clara e profissional.
4. **Criar** um documento que serve tanto como análise quanto como material didático.

Estas habilidades são fundamentais não apenas para projetos acadêmicos, mas também para aplicações profissionais em ciência de dados e análise de negócios no mundo atual da economia de dados.