Sabendo que foram feitos os passos abaixo, elabore os resultados preliminares, conforme template também listado abaixo:

-Etapas efetuadas:

1 - Utilização das bibliotecas:

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

import numpy as np

from sklearn.preprocessing import StandardScaler

from dateutil.relativedelta import relativedelta

from sklearn.model\_selection import train\_test\_split, KFold, GridSearchCV

from sklearn.preprocessing import StandardScaler

from sklearn.linear\_model import LinearRegression

from sklearn.ensemble import RandomForestRegressor

from xgboost import XGBRegressor

from sklearn.tree import DecisionTreeRegressor

from sklearn.metrics import mean\_squared\_error, mean\_absolute\_percentage\_error, mean\_absolute\_error, r2\_score

import datetime

2 - Definição de funções

ml\_error --> traz as métricas mae, mse, rmse e o r² para medir variancia do modelo

cross\_validation --> divide em 5 folds para calculo das métricas acima

3 - Dimensão do dataset

1586614 linhas e 14 colunas

4 - Conversão de variáveis

Conversão de variaveis de review em numero, review\_time e data

5 - Tratamento de valores em branco

beer\_alcohol\_by\_volume : Foram encontrados 67785 linhas não preenchidas da coluna e os valores em branco foram preenchidos com a média do teor alcoólico por estilo de cerveja

review\_appearance: atribuição de nota mínima nas linhas não preenchidas ou com valores iguais a 0 – 15 ocorrencias

brewery\_name: 7 casos em branco, as observações foram removidas

6 – Analise descritiva

Atribuitos numéricos: Foram calculados por atributos numéricos, o min, max, range, media, mediana, desvio padrão, skew e curtose

Atributos categóricos: Foram calculados a quantidade de valores únicos por variável, a menor delas tinha 104 categorias diferentes

7 – Feature Engineering

Data: Extraídos ano, mês, dia, ano-mês, ano-semestre e ano-trimestre

Criação da variável brewery\_popularity que calcula a quantidade de avaliações que uma cervejaria fez, procurando buscarse uma cervejaria mais popular tende a receber avaliações melhores

Criação da variável reviewer\_proficiency que indica quantas vezes uma pessoa fez uma avaliação que pode ditar se consumidores mais experientes tendem a ter avaliações mais criteriosas

Criação de cluster por estilo de cerveja: por conter 104 estilos de cerveja, uma variável de cluster foi criada com base nas nomenclaturas e teores alcoólicos chegando em Pale Ale, IPA, Stout, Weizenbier, Pilsner, Helles e Dunkel

8 - Analise Exploratória de Dados

Univariada:

Foram plotados hisitogramas e boxplot da variável resposta (review\_overall) e todas as variáveis sensoriais, a fim de identificar outliers e qual seria o tipo de distribuição de cada uma delas, e todas elas possuem uma distribuição parecida com a normal, mas com uma cauda um pouco mais alongada para a esquerda, sendo que os valores possíveis vão de 0,5 a 5 variando a cada 0,5

Para a variável beer\_alcohol\_by\_volume foi idenficada uma distribuição bem alongada a esquerda devido aos tipos de cerveja com teor alcoólico fortíssimo, que foram verificados e adotados como reais, sem serem outliers, é uma variável que possui uma variação bem alta de teores

Bivariada:

Analise bivariada foram plotadas todas as variáveis numéricas em relação a data de avaliação, chegando na distribuição de cada uma das variáveis numéricas ao longo do tempo.

Notou-se que o teor alcoolico médio foi aumentando de forma quase linear conforme evolução temporal

Para a variável resposta, nota-se um período de um ano e meio com uma nota média duas vezes superior aos demais períodos, o que pode representar correlação significativa com a variável resposta, o que ajudaria na predição

Multivariada:

Foi plotado um mapa de calor que correlaciona todas as variáveis numéricas e foi constata maior correlação com a variável review\_overall para as variáveis review\_palate e review\_taste, sendo 0,79 e 0,72 respectivamente.

9 – Seleção de variáveis

Para um primeiro modelo, embora forte evidencia de baixa correlação entre o teor alcoólico e possível evidência de alguma variável de data ser representativa, as variáveis escolhidas foram as sensoriais e a de teor alcoólico:  
'beer\_alcohol\_by\_volume', 'review\_aroma', 'review\_taste', 'review\_palate', 'review\_appearance'

10 – Normalização das variáveis

Uma vez que serão utilizados modelos que são sensíveis a variações nas variáveis explicativas como algoritmos de boosting e bagging, foi utilizado o método StandardScaler da biblioteca sklearn.preprocessing

11 – Separação em treino e teste

Foi utilizada a biblioteca train/test split da biblioteca sckit learn em proporção de treino e teste de 0,8 para treino e 0,2 para teste, com as variáveis explicativas normalizadas

12 – Treinamento dos modelos

Linear Regressor: treinado o modelo usando a função .fit depois a função predict e para o valor predito foi feito um arredondamento para que a variável resposta variasse de meio em meio utilizando a função do numpy round, onde as métricas de erro avaliadas foram: MAE (0,28) MSE (0,19) e RMSE (0.44)

Foi calculado o coeficiente de determinação R² que deu 0,63 e após isso foi feito o cross validation em 5 folds, onde os resultados foram:

MAE: 0,31

MSE: 0,17

RMSE: 041

R²: 0,67

O mesmo foi feito para arvores de decisão, Random forest e xgboost, chegando nos resultados abaixo:

Arvore de Decisão MAE: 0.28

Arvore de Decisão MSE: 0,19

Arvore de Decisão RMSE: 0,43

Arvore de Decisão R²: 0,64

Após cv

Arvore de Decisão MAE: 0.31

Arvore de Decisão MSE: 0,17

Arvore de Decisão RMSE: 0,41

Arvore de Decisão R²: 0,68

XGBoost MAE: 0.28

XGBoost MSE: 0,18

XGBoost RMSE: 0,43

XGBoost R²: 0,64

Após cv

XGBoost MAE: 0.31

XGBoost MSE: 0,16

XGBoost RMSE: 0,40

XGBoost R²: 0,68

Random Forest MAE: 0.28

Random Forest MSE: 0,19

Random Forest RMSE: 0,43

Random Forest R²: 0,64

Após cv

Random Forest MAE: 0.31

Random Forest MSE: 0,17

Random Forest RMSE: 0,41

Random Forest R²: 0,68