Uma análise da base de dados Wiki4HE

SÉRGIO JOSÉ DE SOUSA

Sobre a base de dados

Wiki4HE é uma base de dados contendo um **questionário** que foi aplicado aos professores com temas a respeito da sua percepção da **Wikipédia** como ferramenta para outros colaboradores e estudantes.

Possui atributos pessoais a respeito dos professores como idade, tempo de experiência e um conjunto de questões, onde o participante avalia a afirmação numa **escala** de **1 à 5**, sendo 1 discorda fortemente e 5 concorda fortemente.

As questões são agrupadas por um descritor relacionado ao tema dessas questões.

Link de acesso ao banco: http://archive.ics.uci.edu/ml/datasets/wiki4HE

Objetivos da análise

1. Estatísticas básicas sobre os professores participantes
2. Análises de diferentes grupos de usuários para os itens “Perceived Enjoyment” **ENJ1** e **ENJ2**
3. Outras análises interessantes
4. Criar um modelo capaz de prever se um professor recomenda ou não o uso do Wikipédia para seus alunos

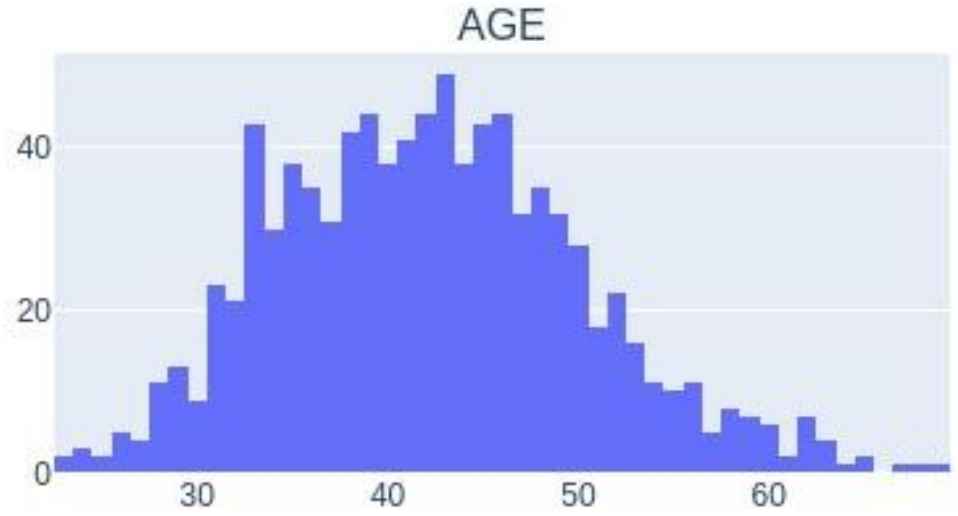
1. Estatísticas básicas sobre os professores participantes

1. Dados Gerais

- Ao todo são **913** participantes
- Inclui **53** atributos, sendo **10** atributos a respeito do **perfil** do professor e **43 questões** divididas em 13 temas

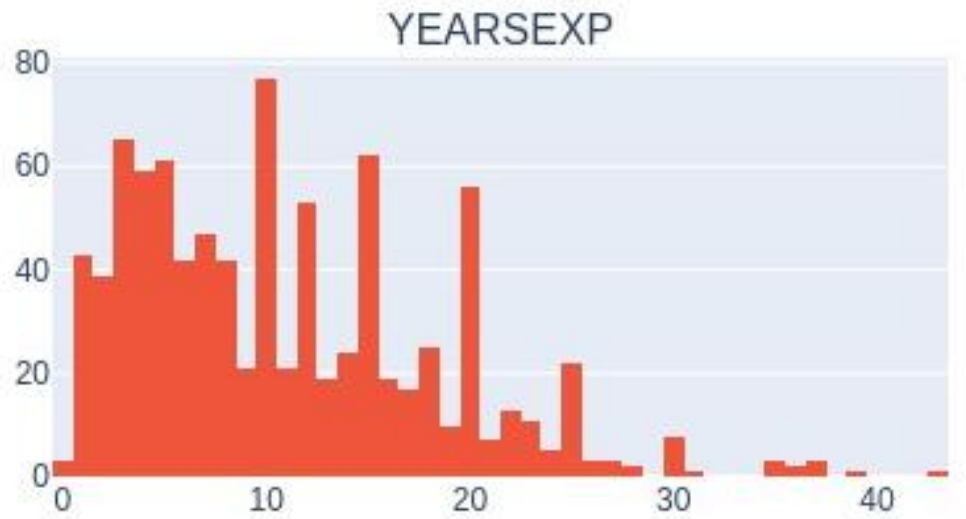
1. Histograma das Idades

- Histograma da distribuição de **idade** dos participantes
- Média 42 anos
- Mediana 42 anos
- A maioria dos professores que participaram possuem entre 40 e 50 anos



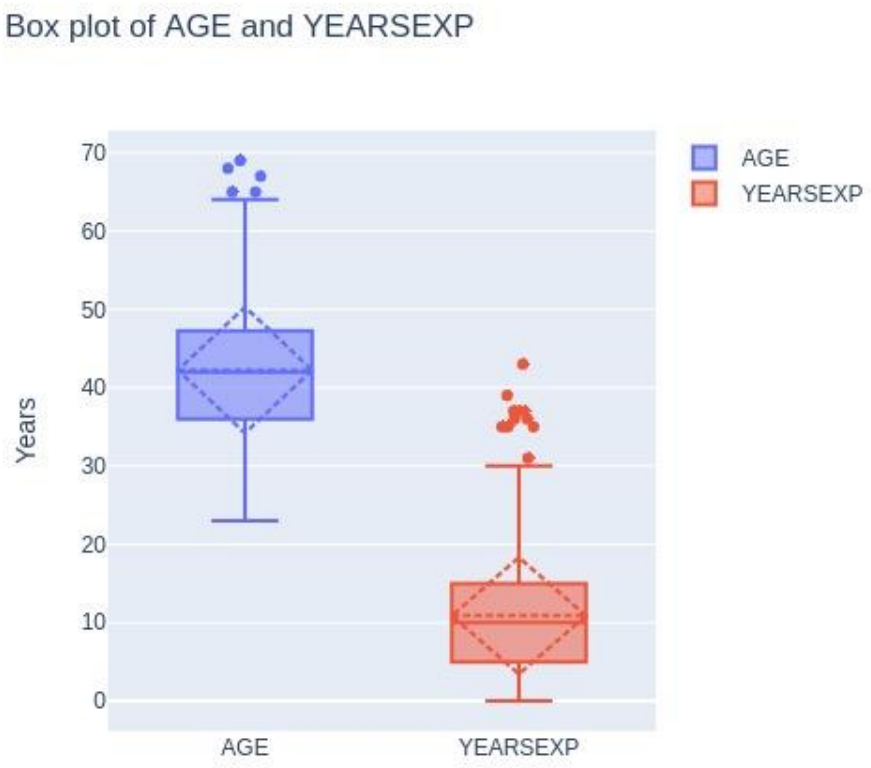
1. Histograma do tempo de experiência

- Histograma da distribuição de **anos de experiência** dos participantes
- Média 10 anos
- Mediana 10 anos
- São poucos os professores com mais de 30 anos de experiência



1. Boxplot da Idade e Experiência

- Boxplot das distribuições de idade e tempo de experiência com média destacada e outliers



1. Distribuição das médias das respostas

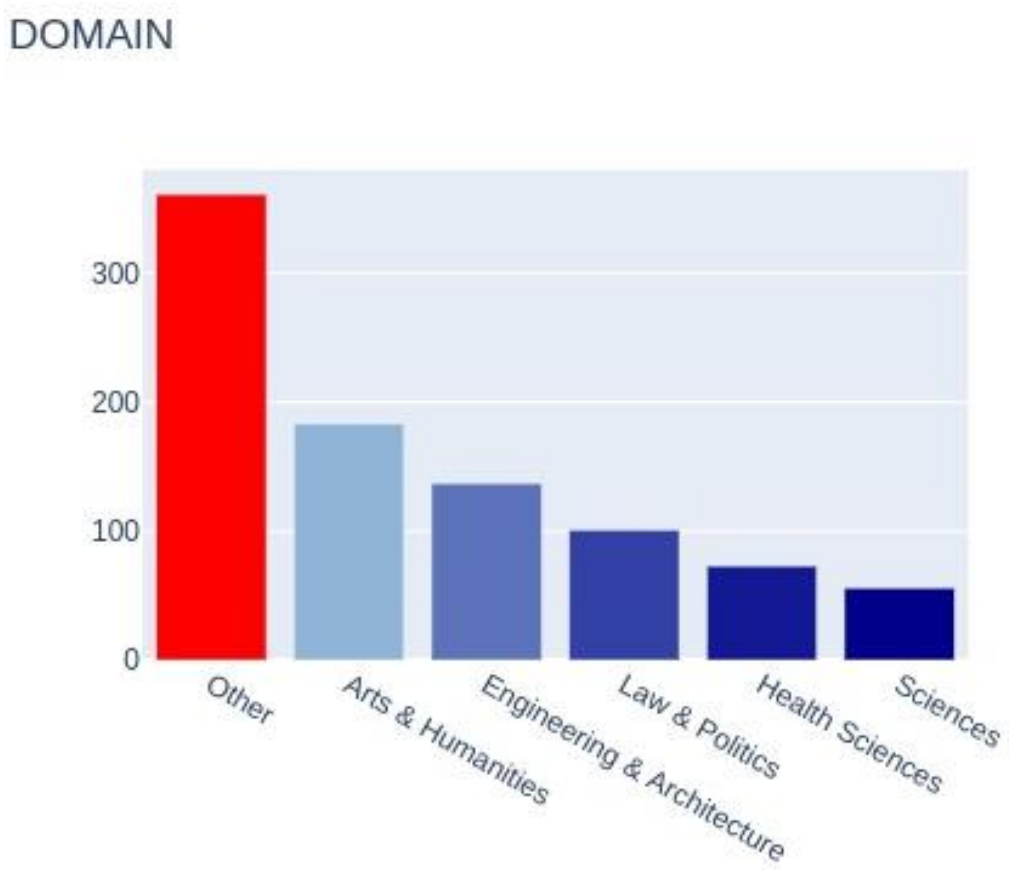
- Histograma da distribuição das médias de respostas dos participantes
- Totalizando e tirando a média das respostas
- Média 3
- Mediana 3

1. Gênero

- Proporção de candidatos
- 57.5% do gênero masculino
- 42.5% do gênero feminino

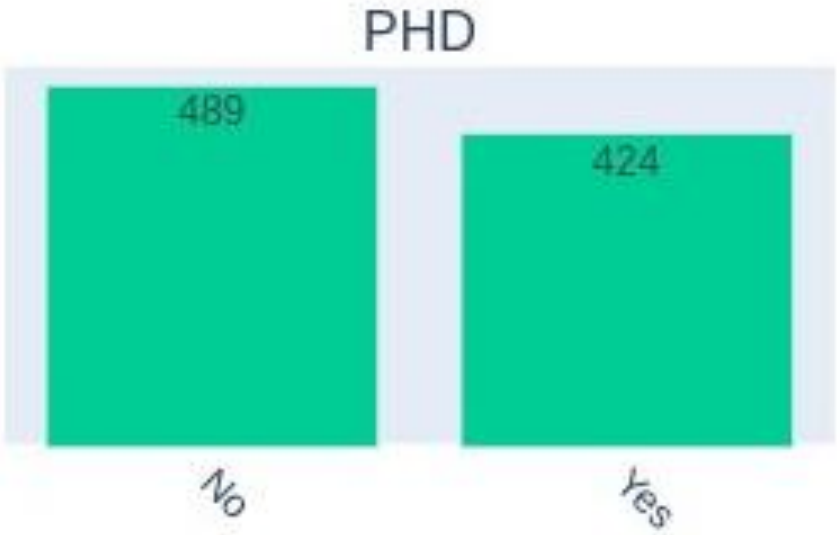
1. Área de atuação

- No atributo relacionado a área do candidato
  - 39.5% Pertencem à outras áreas
  - 20% Artes e Humanas
  - 15% Engenharia e Arquitetura
  - 11.1% Leis e Política
  - 8% Ciências da Saúde
  - 6.4% Ciências



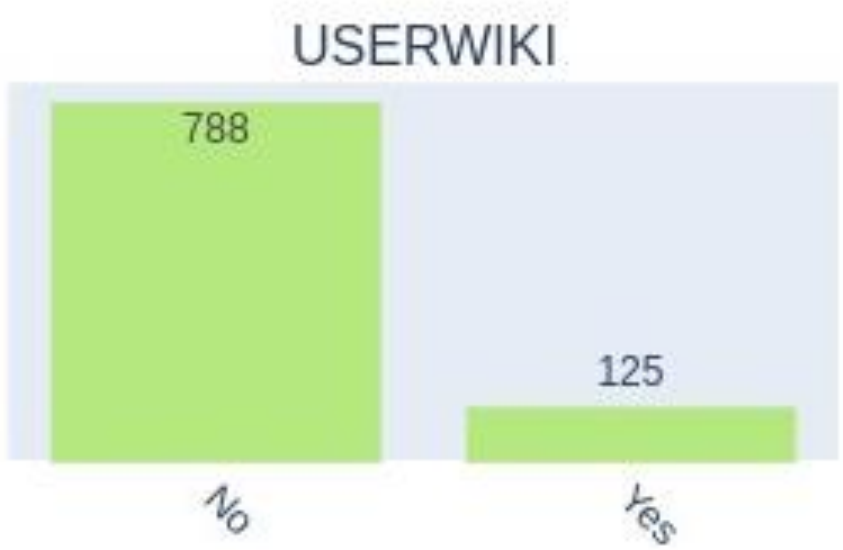
1. PhD

- 53.5 % Possuem PhD
- 46.5 % Não



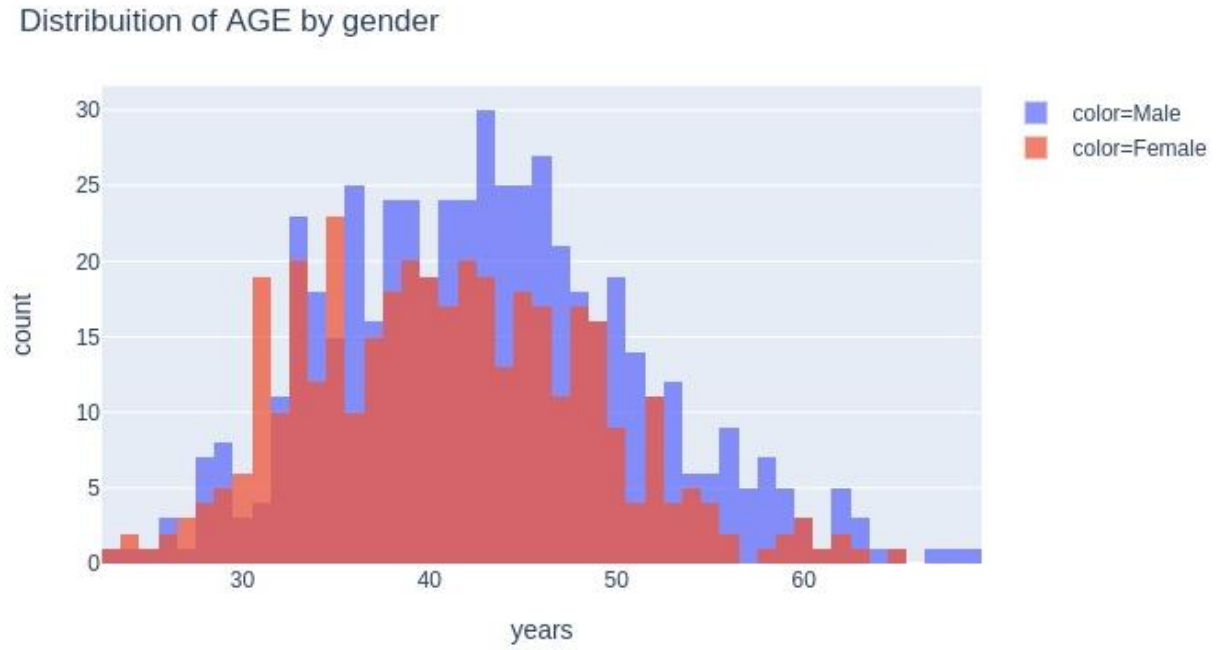
1. Usuário da Wikipédia

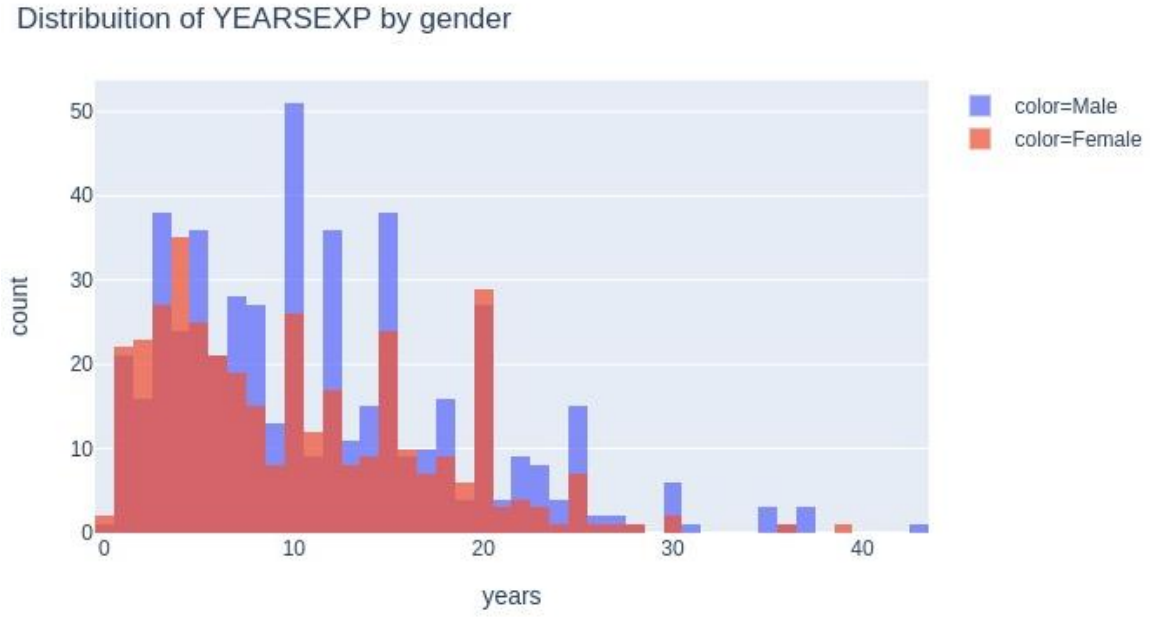
- Apenas 13.7% dos participantes são usuários da Wikipédia



1. Idade e Experiência por Gênero

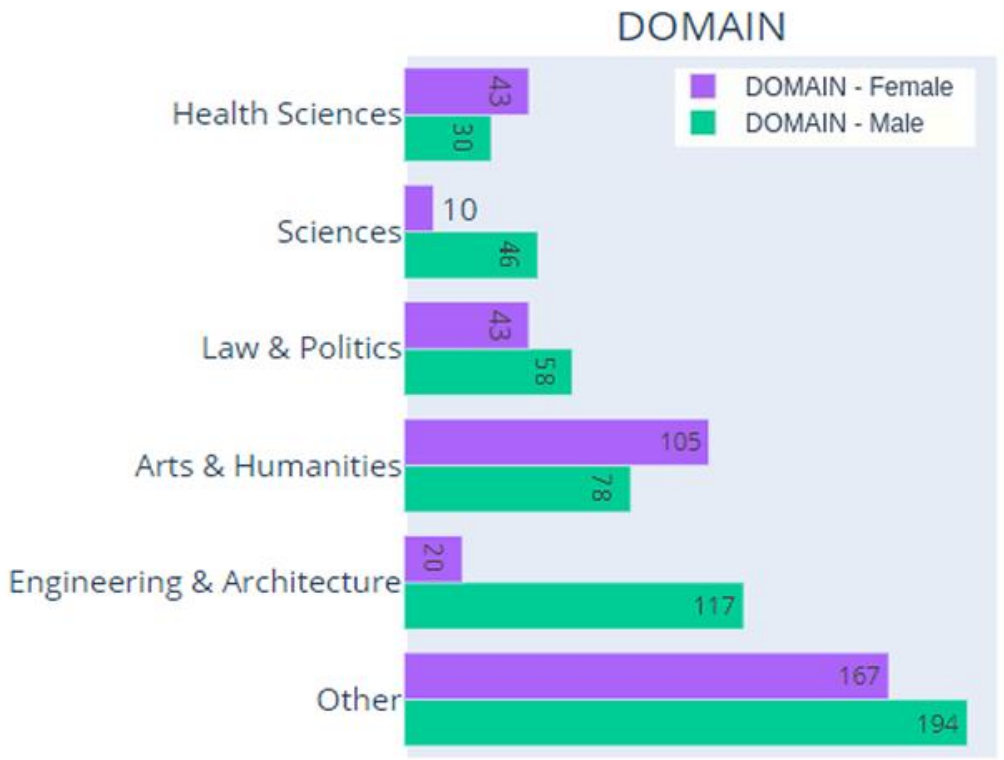
- Histograma das idades e anos de experiência separados por gênero
  - Maior número de candidatos do gênero feminino entre 30 e 33





1. Área de Atuação por Gênero

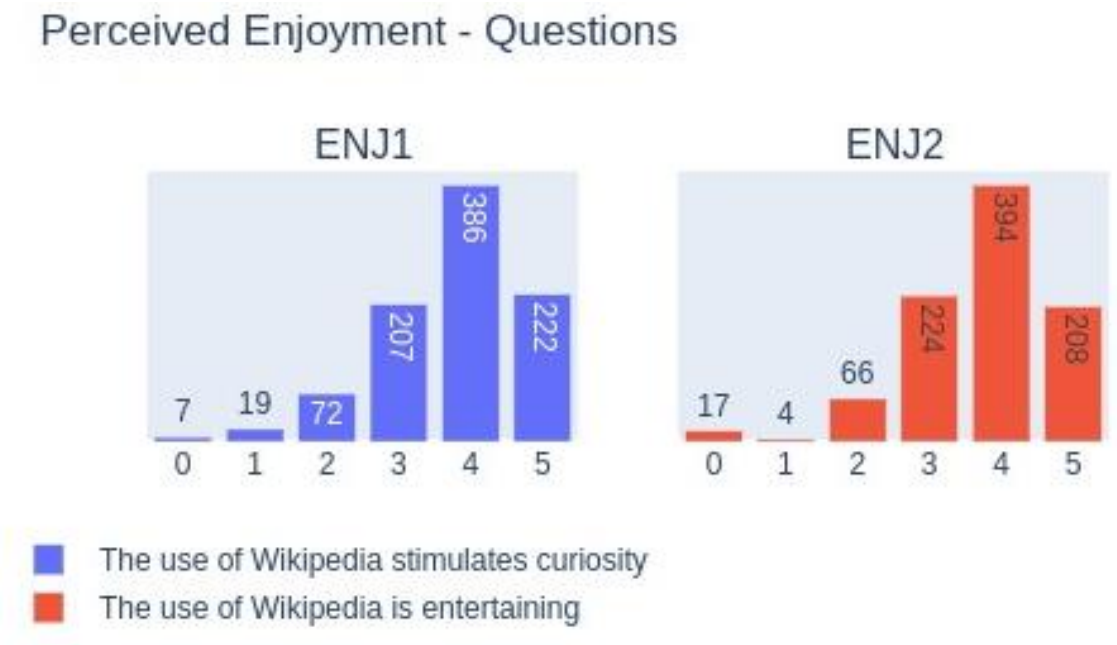
- Observamos uma variação nas áreas de domínio ao compararmos homens e mulheres
- Observamos um número expressivamente **menor** de mulheres nas área de Ciências, Engenharia e Arquitetura
- Em contrapartida observamos **mais** mulheres nas Ciências da Saúde



2. Análises sobre os itens “Perceived Enjoyment” **ENJ1** e **ENJ2**

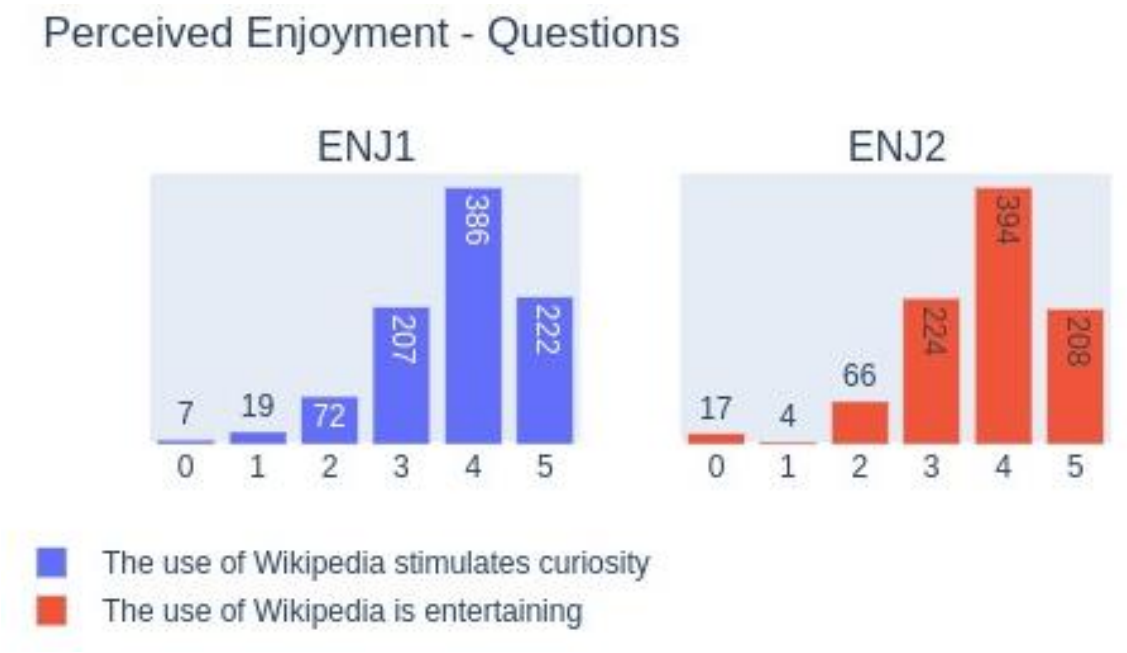
2. Dados Gerais Sobre Engajamento

- Perguntas sobre o Engajamento Percebido
  - ENJ1: O uso da Wikipédia estimula a curiosidade
  - ENJ2: O uso da Wikipédia é divertida
- Observamos uma distribuição similar entre as questões
- A maioria dos participantes concordam com as afirmações sobre a Wikipédia estimular a curiosidade e ser divertida



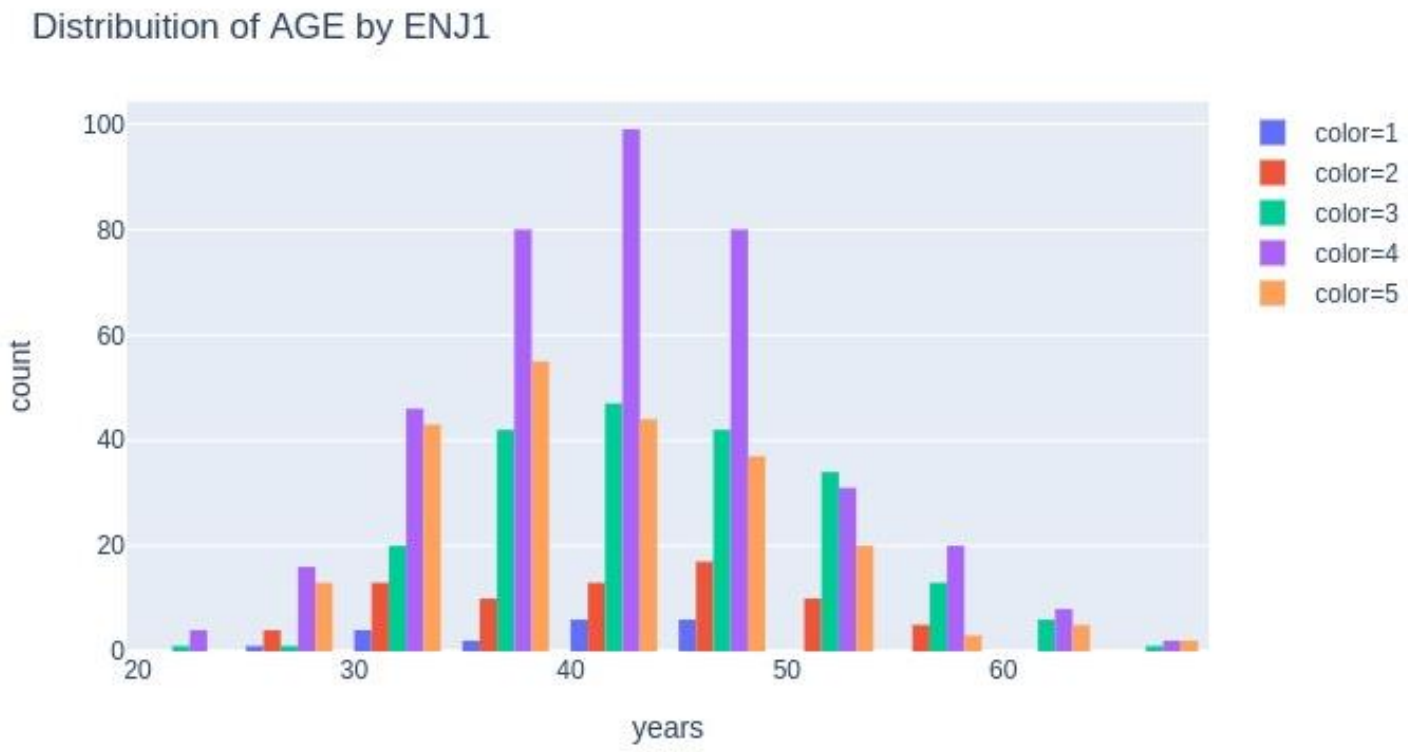
2. Dados Gerais Sobre Engajamento

- Totalizando os votos podemos notar
  - As médias são próximas de 4 (3.75)
  - As medianas são 4 para ambas as questões



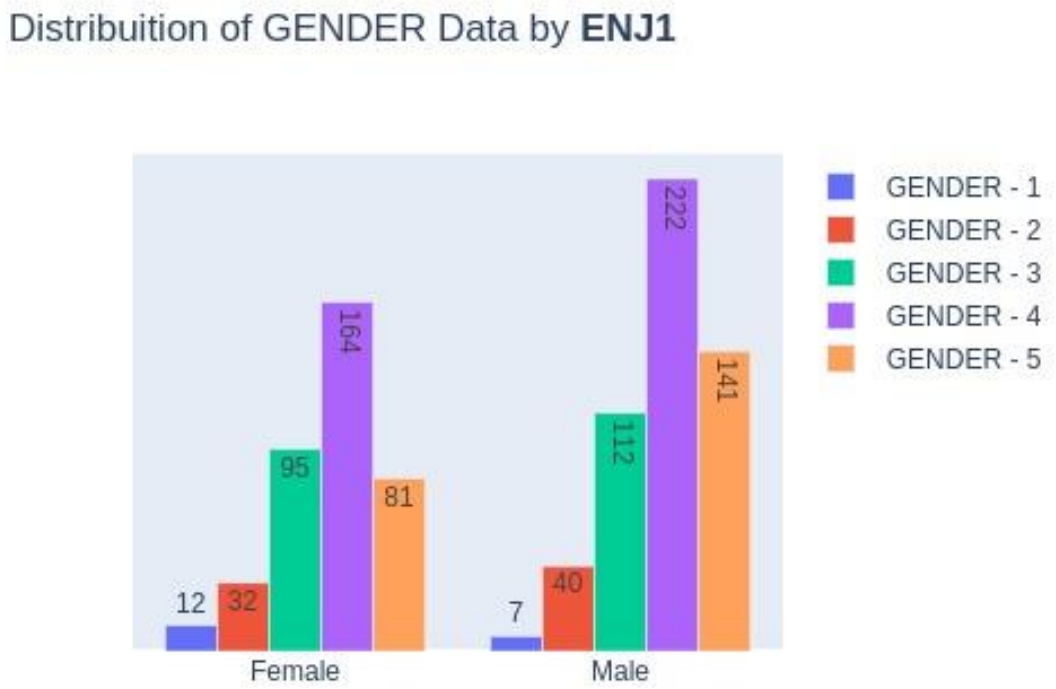
2. Análises sobre **ENJ1**

- Distribuição do histograma das **idades** separadas pela resposta à questão ENJ1
- Participantes acima de 50 anos tendem a discordar mais sobre “a Wikipédia estimular a curiosidade”
- Enquanto os participantes com idade < 50 tendem a concordar



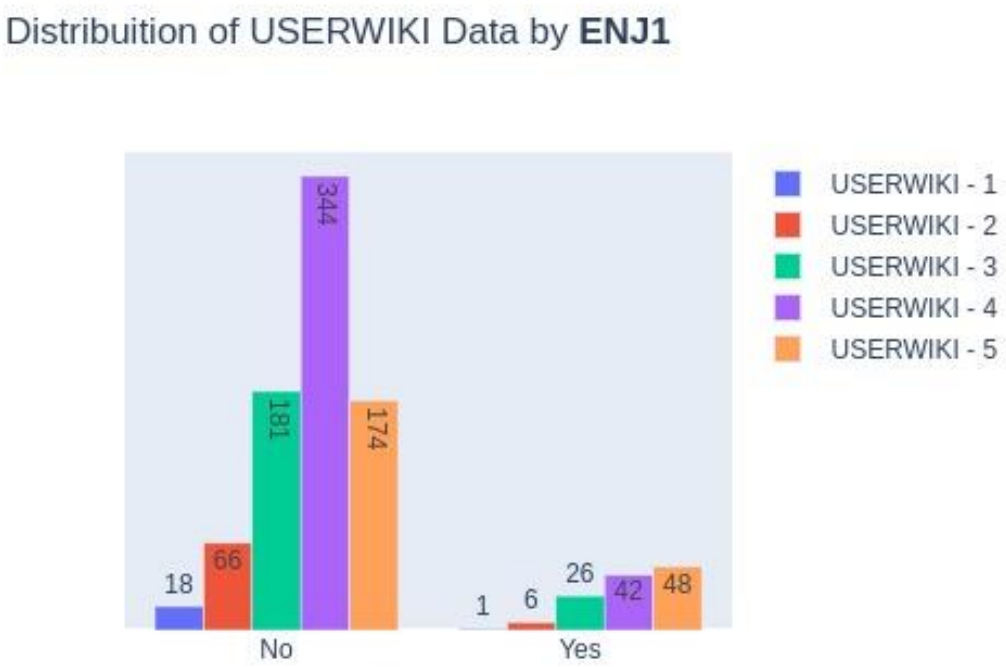
2. Análises sobre **ENJ1**

- Distribuição dos **gêneros** separados pela resposta à questão ENJ1
- Participantes do gênero feminino se mantém mais neutras
- Enquanto participantes do gênero masculino concordam plenamente com “a Wikipédia estimular a curiosidade”



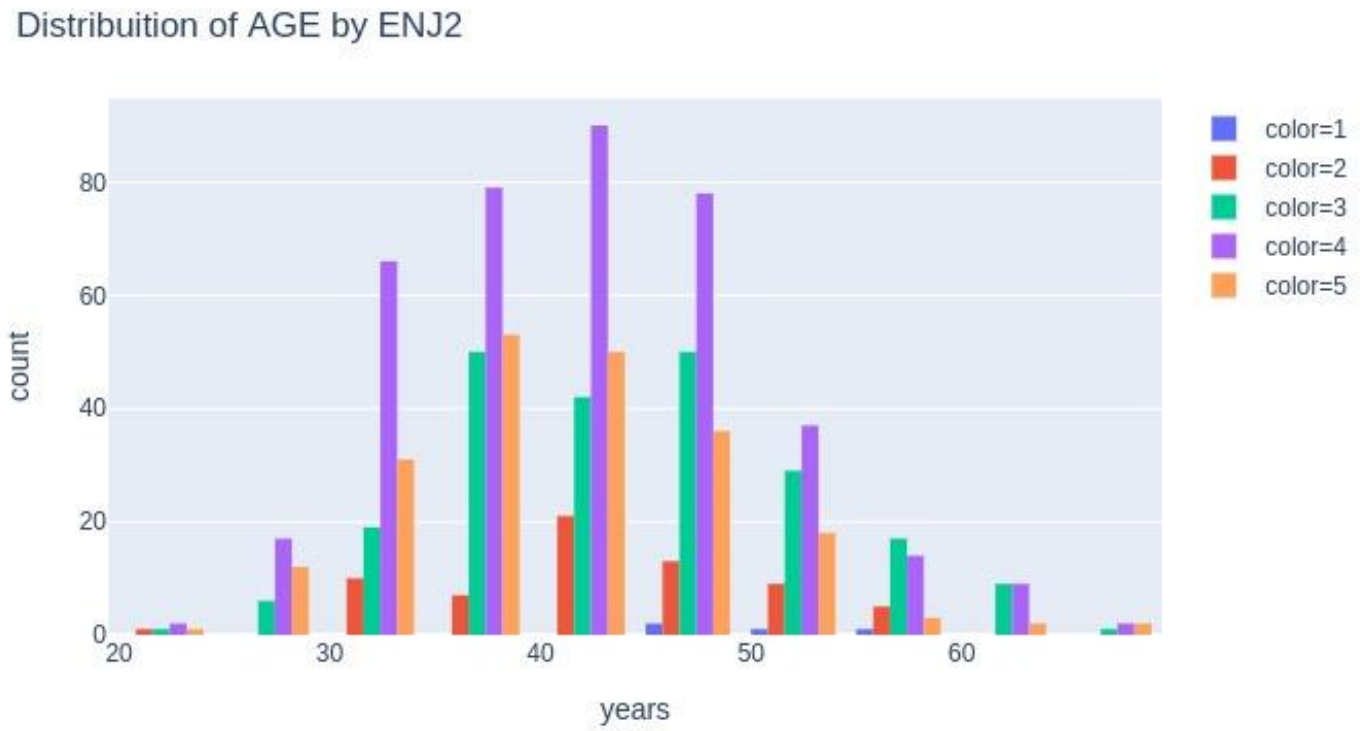
2. Análises sobre **ENJ1**

- Distribuição de **usuários** da Wikipédia separados pela resposta à questão ENJ1
- Usuários da Wikipédia tendem a concordar mais com “a Wikipédia estimular a curiosidade”



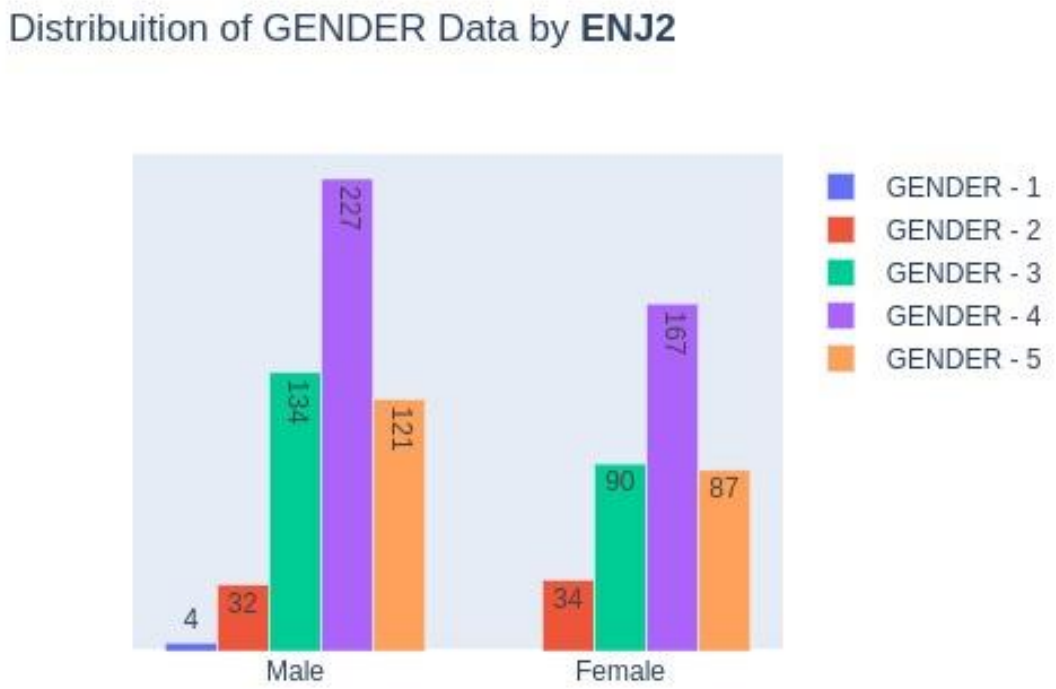
2. Análises sobre **ENJ2**

- Distribuição do histograma das **idades** separados pela resposta à questão ENJ2
- Participantes com idade acima de 50 anos tendem a ser neutros sobre “a Wikipédia ser divertida”



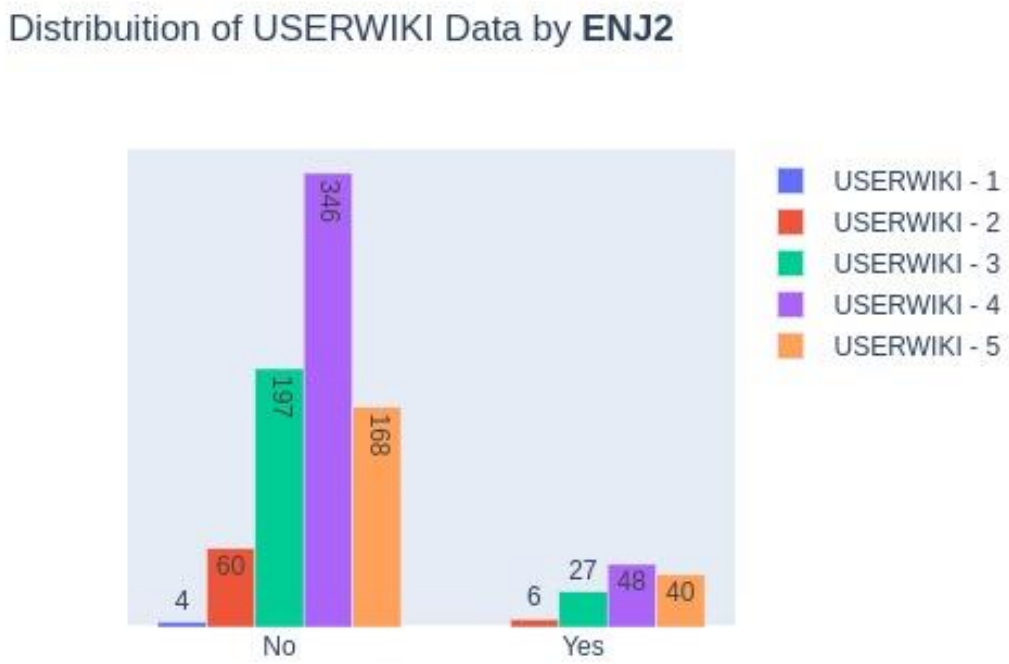
2. Análises sobre **ENJ2**

- Distribuição dos **gêneros** separados pela resposta à questão ENJ2
- Os Participantes do gênero masculino se mantém mais neutros
- As distribuições são parecidas



2. Análises sobre **ENJ2**

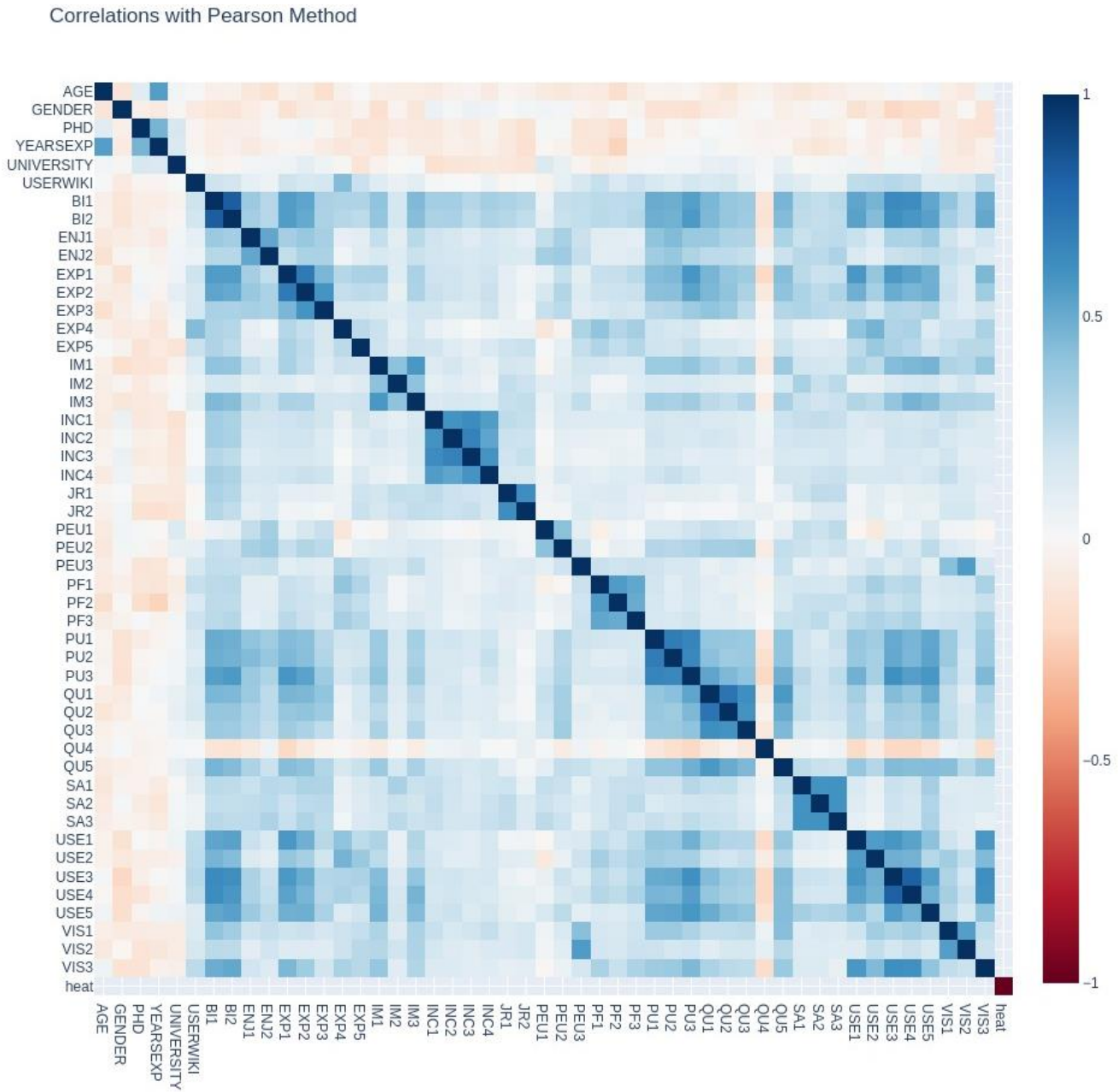
- Distribuição de **usuários** da Wikipédia separados pela resposta à questão ENJ2
- Usuários da Wikipédia tendem a concordar com “a Wikipédia ser divertida”



3. Outras análises

3. Correlações

Mapa de correlações utilizando método Pearson, pois os dados em sua maioria possui uma distribuição normal



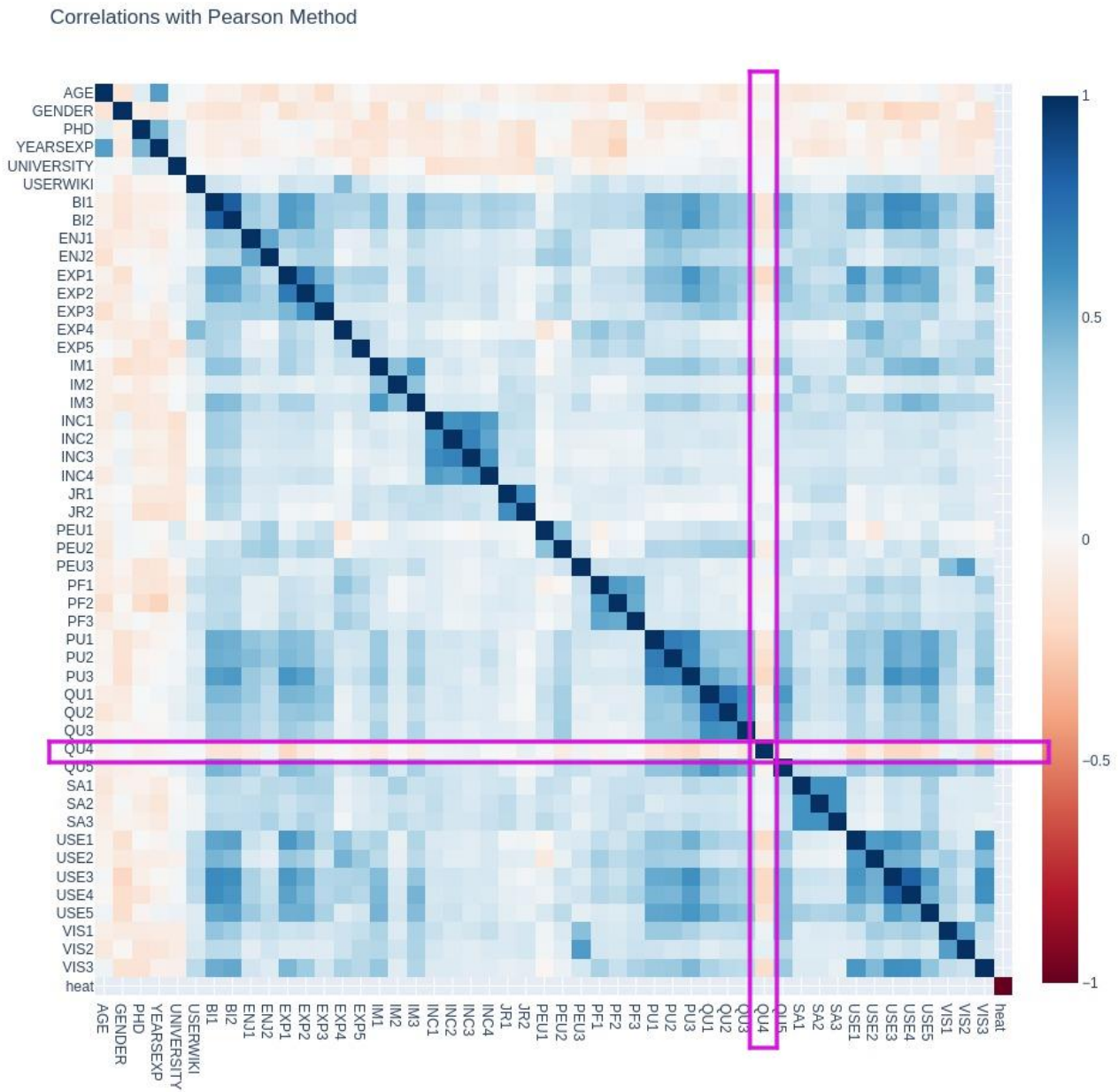
3. Correlações

A questão **QU4** possui correlação negativa com a maioria dos atributos.

De fato é uma afirmação negativa

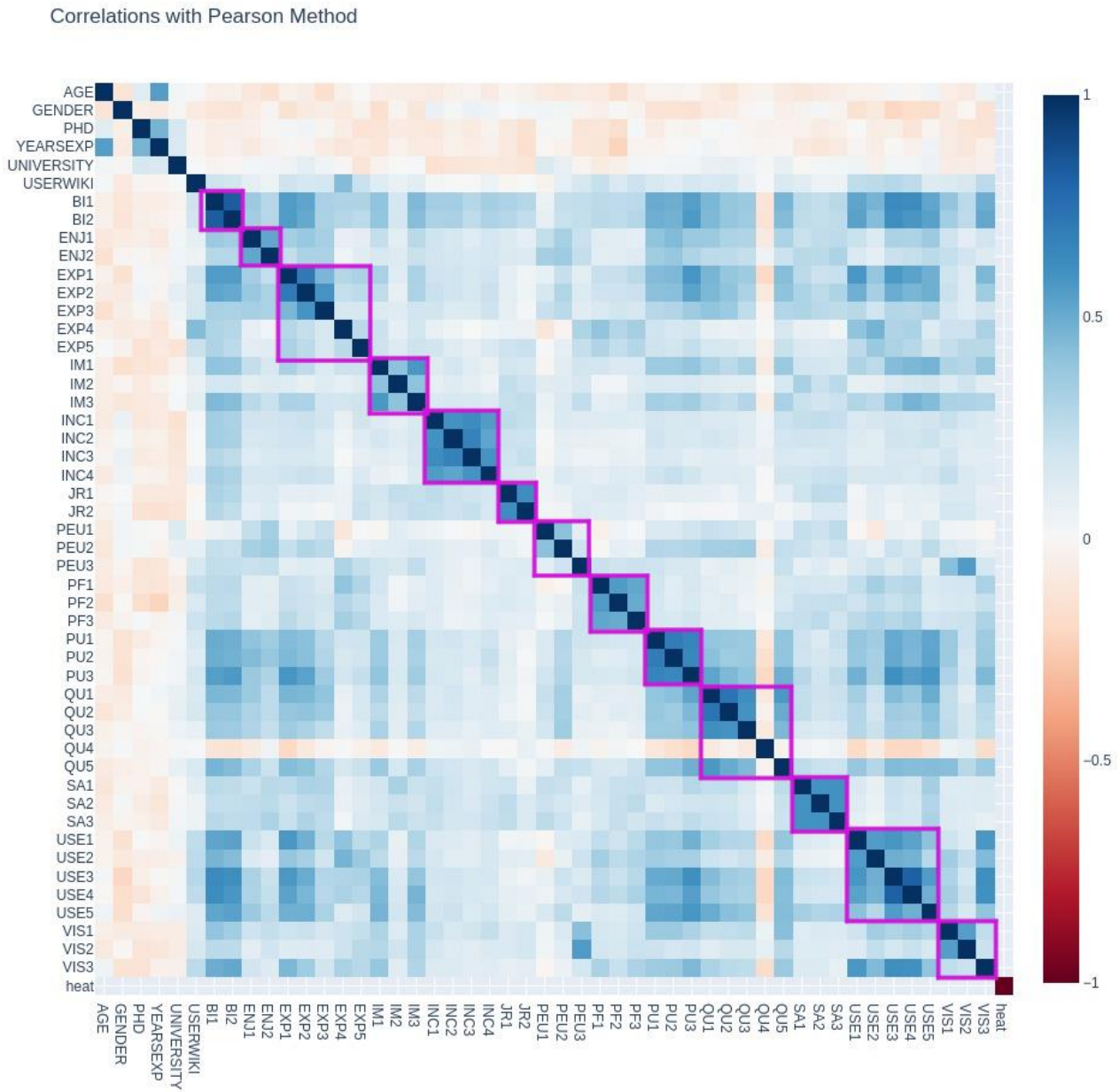
- “Na minha área de especialização, a Wikipédia tem uma qualidade inferior a outros recursos educacionais”

Enquanto as outras questões possuem um caráter afirmativo como pode ser visto ao lado



3. Correlações

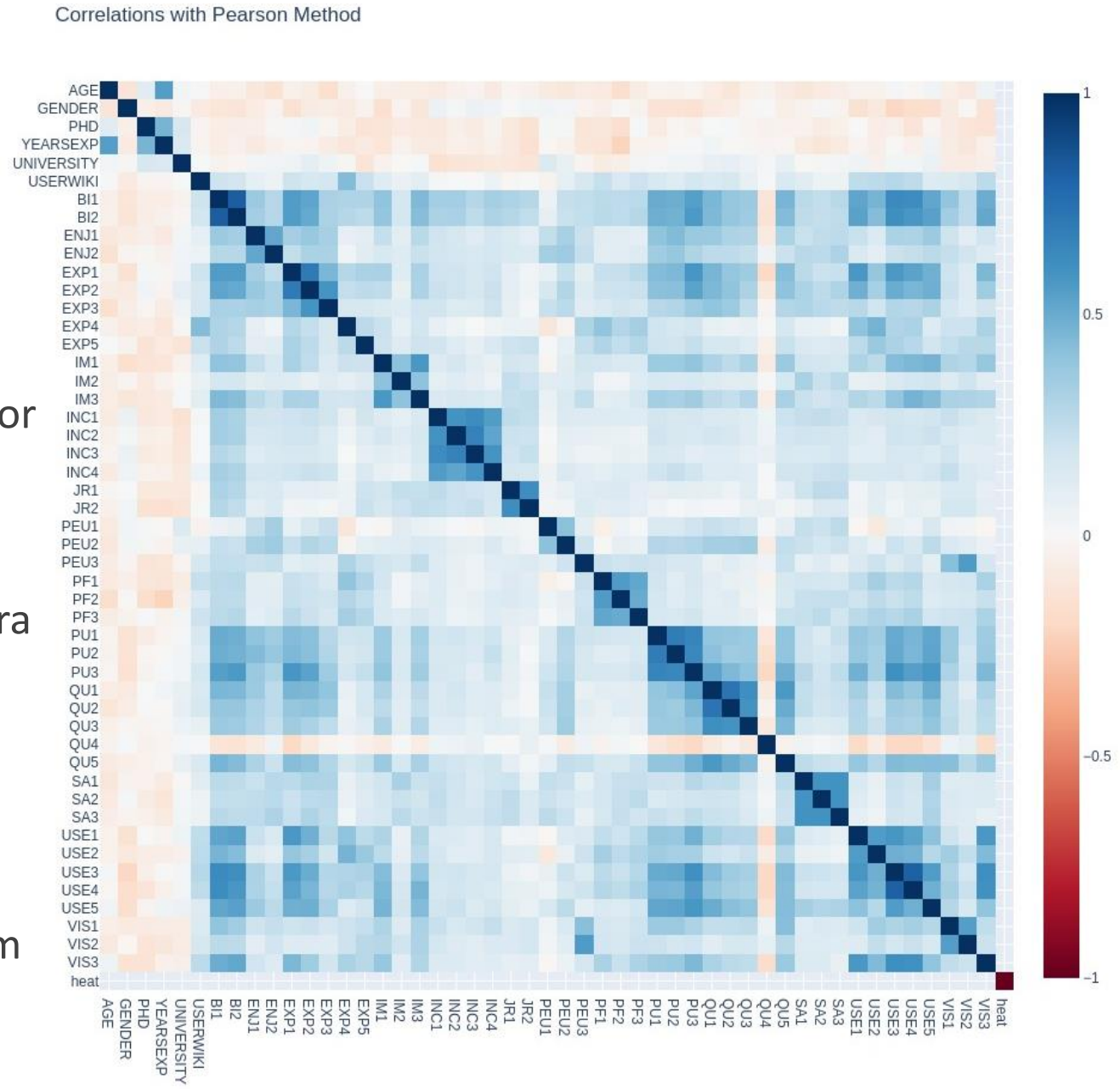
A maioria das questões de um mesmo grupo possui uma alta correlação entre si.



3. Correlações

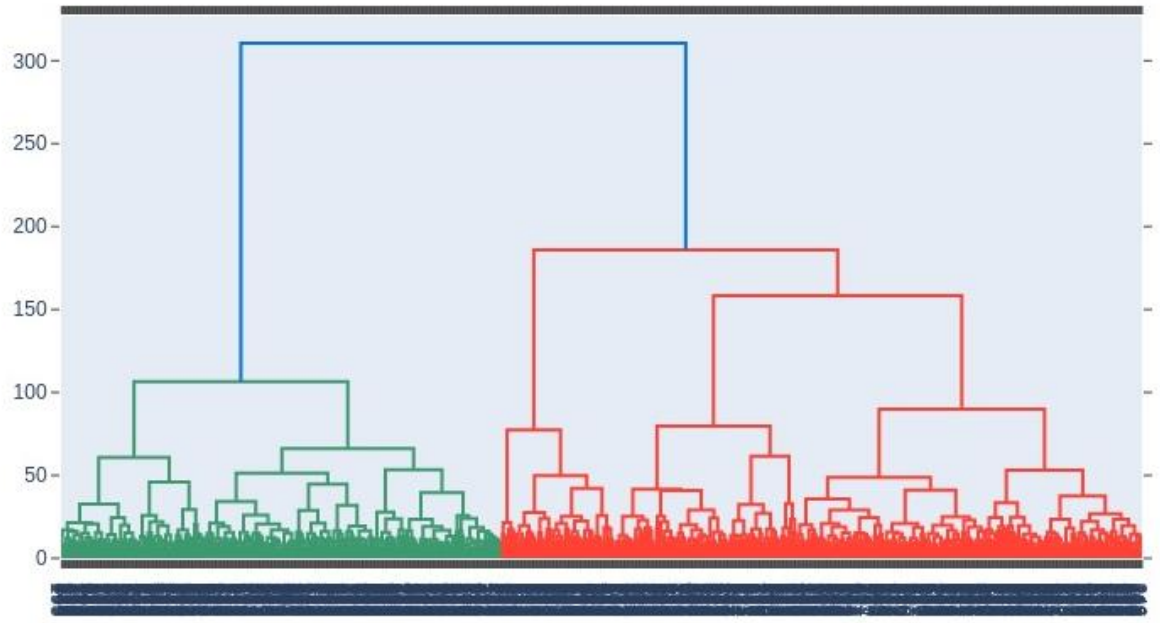
Algumas observações interessantes

- Quanto maior o tempo de experiência menor a atividade em redes sociais
- A maioria das pessoas que no futuro recomendarão a Wikipédia para os colaboradores, também irão recomendar para os alunos
- Participantes que já recomendam para alunos, também recomendam para colaboradores
- Participantes que afirmam que os artigos recebem boas atualizações, também afirmam que possuem boa leitura



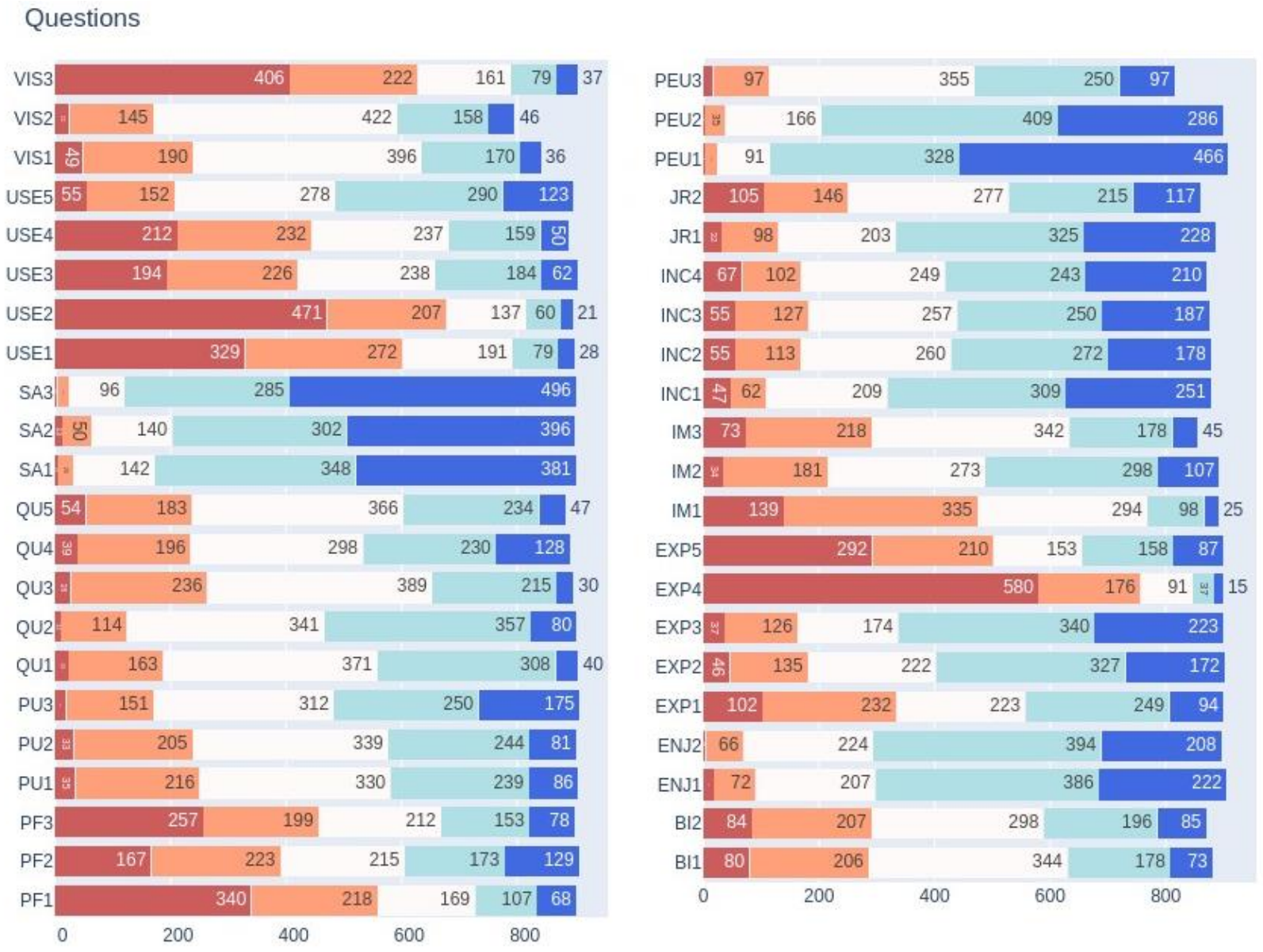
3. Clusters

- Ao analisar o dendograma, tentar encontrar relações com os atributos e aplicar técnicas como PCA **não foi possível encontrar padrões** de separação dos grupos



3. Respostas

- Totais quantificados de cada pergunta do questionário
- O rosa mais escuro significa discordo completamente
- O azul mais escuro concordo completamente

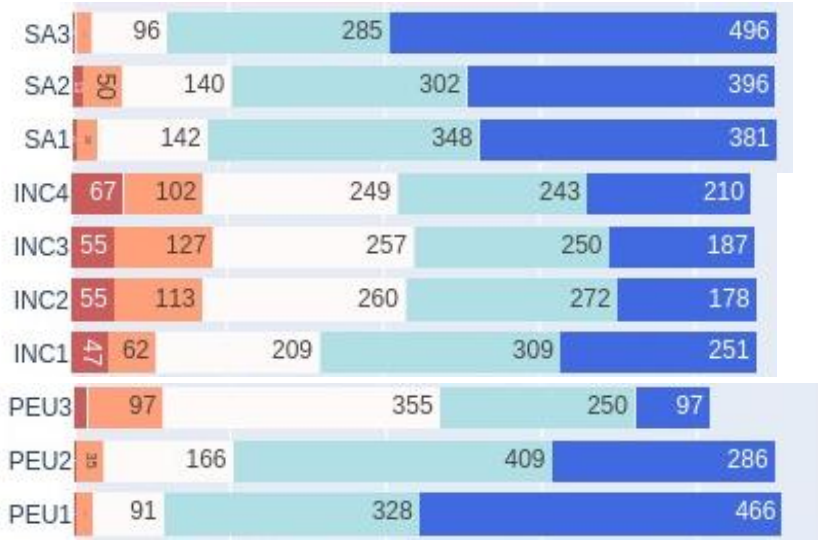


3. Algumas Observações

A **maioria** dos participantes concordam com as afirmações a respeito da importância de **compartilhar informações** (SA1, 2 e 3)

A **maioria** concorda sobre a necessidade de **incentivos**, como práticas de recomendação, instruções de colegas, treinamento específico e reconhecimento institucional (INC1, 2, 3 e 4)

A **maioria** concorda sobre a **facilidade** em encontrar informações, adicionar, editar e usar a Wikipédia (PEU1, 2 e 3)



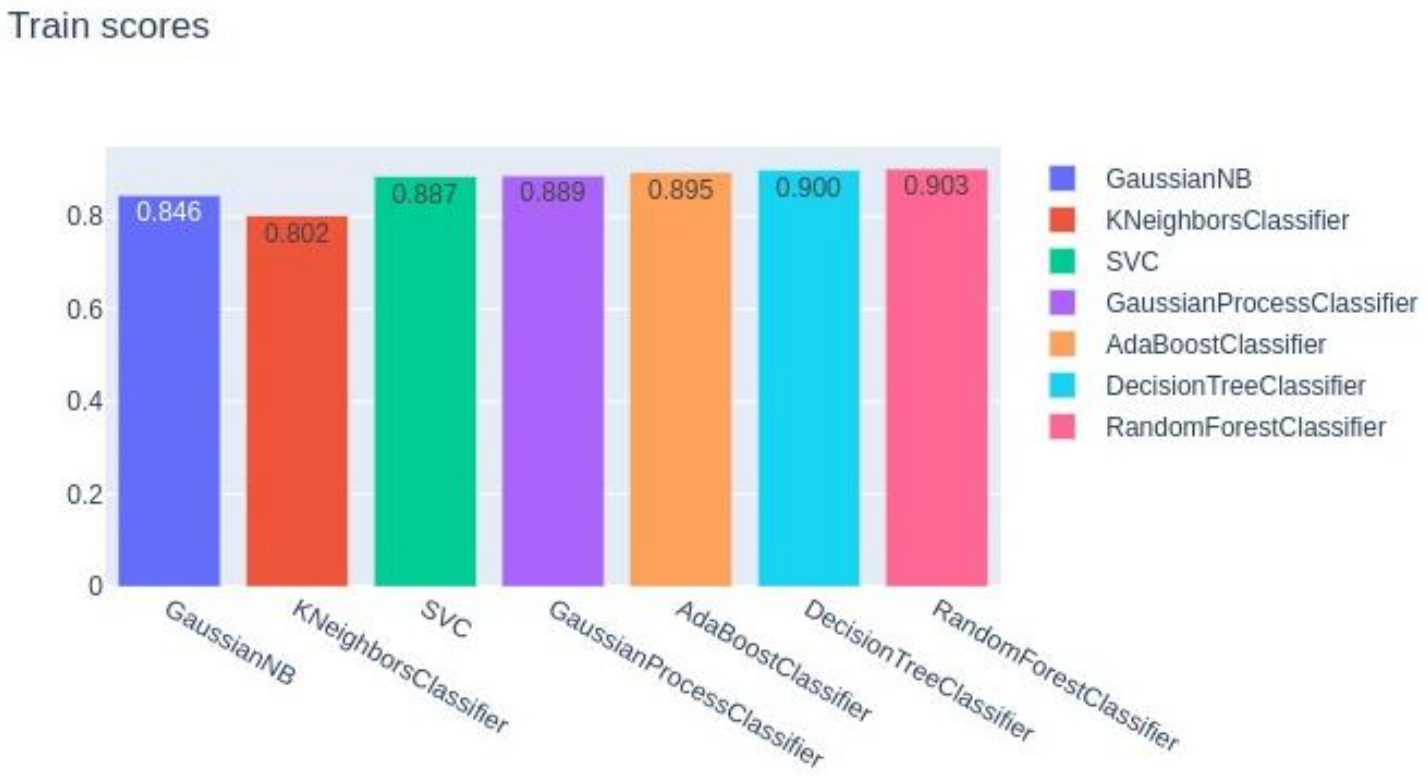
4. Classificador

4. Classificador

- A questão **USE3** pergunta se o professor recomenda que os estudantes usem a Wikipédia
- A seguir é proposto um classificador que será capaz de identificar se o participante recomendaria o uso ou não do Wikipédia com bases nas outras 42 questões e dados do perfil
- Antes é necessário transformar o dado USE3 em um atributo binário onde de **0 até 3** o participante **não recomenda** enquanto de **4 a 5 recomenda**.
- Recomendam: 246
- Não recomendam: 667

4. Modelos

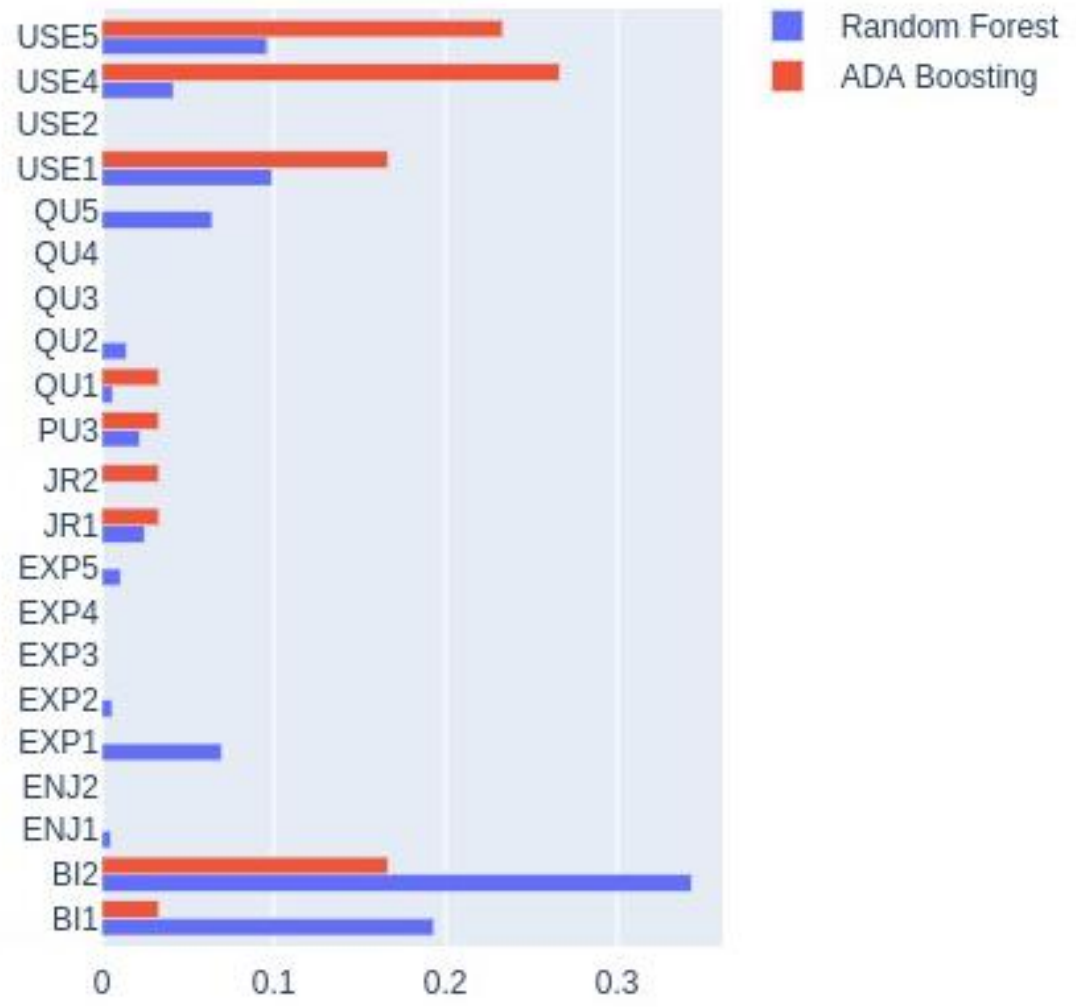
- Modelos testados
  - Naive Bayes
  - K Nearest Neighbors
  - SVM
  - Gaussian Process
  - AdaBoost
  - Decision Tree
  - Random Forest
- Hiperparâmetros otimizados com GridSearch
- Validação Cruzada com 5 folds



4. Modelos

- **AdaBoost** e **Random Forest** obtiveram os melhores resultados
- Os atributos relevantes para suas classificações podem ser observadas ao lado
- Como observado no gráfico de correlações, os grupos de questões possuem forte correlação, podemos ver seu uso com USE5, USE4 e USE1

*(Os atributos não relevantes foram removidos para melhorar a visualização)



4. Modelos

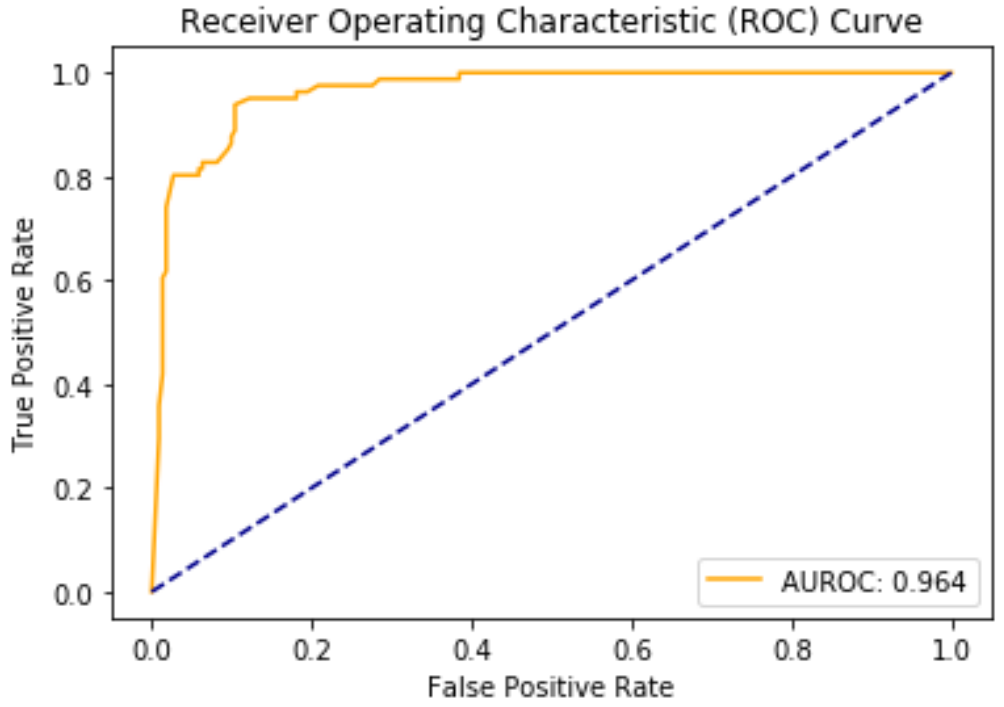
Hiperparâmetros estimados

- AdaBoost
  - {'learning_rate': 0.09, 'n_estimators': 30}
- Random Forest
  - {'max_depth': 3, 'max_features': 8, 'min_samples_split': 0.2, 'n_estimators': 10}

- 33% do conjunto de dados foi reservado para teste
- 302 registros não fizeram parte do treino

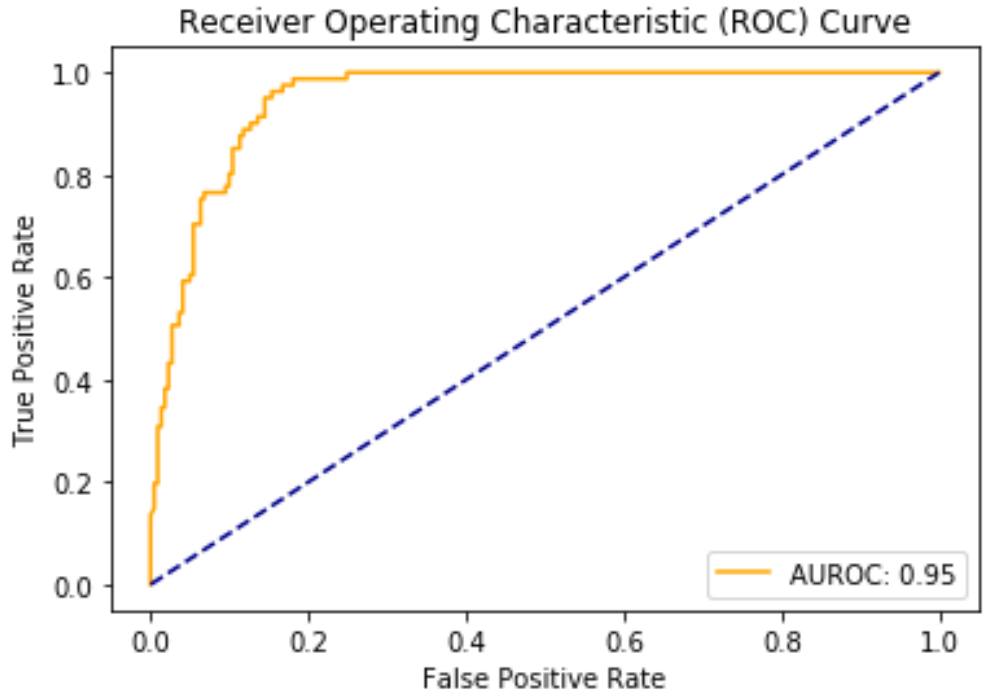
4. Avaliação

- ADABoost obteve 90.7% de acurácia no teste
- AUROC = 0.964



4. Avaliação

- Random Forest obteve 88.7% de acurácia no teste
- AUROC = 0.95



5. Conclusões

- A maioria dos participantes tem uma visão favorável à Wikipédia
- Participantes com idade > 50 são mais conservadores com relação ao seu uso
- A maioria dos participantes acreditam que a Wikipédia gera curiosidade e é divertida
- Existe correção entre as questões de um mesmo grupo
- Um modelo Adaboost foi capaz de obter 90% de acurácia para classificar se o participante recomenda ou não o uso da Wikipédia para seus alunos