Relatório sobre EDA (CRISP-DM - Fase 2)

Dataset: Gender Gap in Spanish WP

Grupo:
Nilo Bemfica Mineiro Campos Drumond *(nbmcd)*
Pedro Didier Maranhão *(pdm)*
Pedro Tenório Lemos *(ptl)*

A segunda fase do CRISP-DM é o ***"Exploratory Data Analysis"***, que tem o objetivo de trazer aos membros da equipe de ciência de dados um maior entendimento sobre a estrutura dos dados que serão utilizados. A base de dados a qual nosso grupo ficou responsável se concentrou na análise das diferenças de gênero no Wikipedia em espanhol.

Para essa etapa, assim como a anterior, nos baseamos na lógica do artigo do *medium* deixado na descrição da atividade do classroom. Fizemos alguns acréscimos conforme o que sentimos ser pertinente. Com isso em mente, seguimos os seguintes passos em nossa EDA:

1. Preview data (olhar inicial dos dados)
2. Data cleaning (limpeza e correção do formato dos dados da base)
3. Check null values / duplicate entries (checagem de valores nulos e linhas duplicadas)
4. Plots for categorical data (gráficos para os dados categóricos)
5. Plots for numerical data (gráficos para dados numéricos)
6. Datetime columns exploration (exploramos colunas com registros temporais)

E, conforme solicitado na atividade:

7. Análise crítica sobre a complexidade da base de dados e com respeito da separação das classes.

O objetivo geral dessa EDA era entender primariamente a relação entre a coluna do gênero do usuário e as demais variáveis, dado que o principal objetivo deste dataset é constatar as diferenças entre gênero e atentar para a baixa representação das mulheres em edições de artigo da parte da Wikipédia escrita em espanhol.

1. Preview Data

Aqui tivemos nosso primeiro contato com o dataset em si. Nos orientamos pelo significado das colunas que estava disposto no site do UCI, na sessão da base. Que era:

C_api: gender extracted from WikiMedia API, codes as female / male / unknown

C_man: gender extracted from content coding, coded as 1 (male) / 2 (female) / 3 (unknown)

E_NEds: I index of stratum IJ (0,1,2,3)

E_Bpag: J index of stratum IJ (0,1,2,3)

firstDay: first edition in the Spanish Wikipedia (YYYYMMDDHHMMSS)

lastDay: last edition in the Spanish Wikipedia (YYYYMMDDHHMMSS)

NEds: total number of editions

NDays: number of days (lastDay-firstDay+1)

NActDays: number of days with editions

NPages: number of different pages edited

NPcreated: number of pages created

pagesWomen: number of edits in pages related to women

wikiprojWomen: number of edits in WikiProjects related to women

ns_user: number of edits in namespace user

ns_wikipedia: number of edits in namespace wikipedia

ns_talk: number of edits in namespace talk

ns_userTalk: number of edits in namespace user talk

ns_content: number of edits in content pages

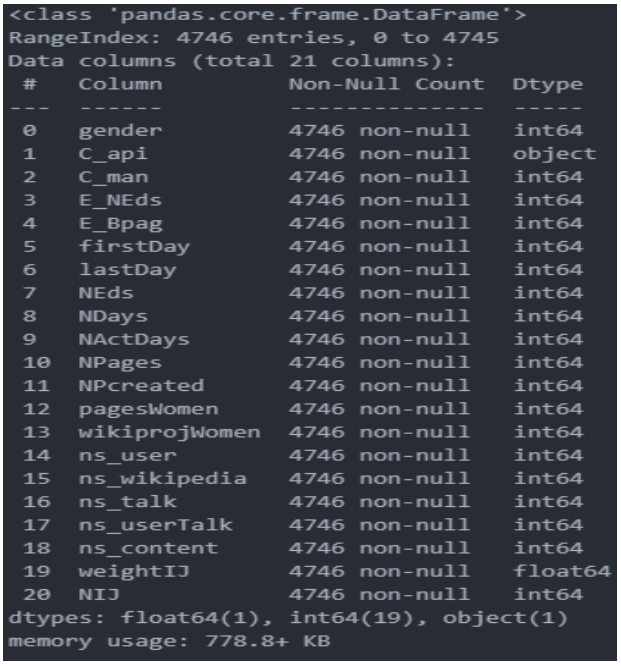
weightIJ: correcting weight for stratum IJ

NIJ: number of elements in stratum IJ

Assim, prosseguimos para avaliar a visão geral das colunas e linhas, usando os métodos *.head* e *.info*:

	gender	C_api	C_man	E_NEds	E_Bpag	firstDay	lastDay	NEds	NDays	NActDays	...	NPcreated	pagesWomen	wikiprojWomen	ns_user	ns_wikipedia	ns_talk	ns_userTalk	ns_conte
0	1	male	1	2	2	20170527205915	20170721044501	543	56	43	...	4	0	0	91	28	6	76	3
1	0	unknown	3	3	1	20110301072441	20170731213735	2764	2345	514	...	7	0	0	100	249	183	646	15
2	1	male	1	0	2	20060907204302	20140911191722	57	2927	25	...	0	0	0	3	0	1	3	
3	1	male	1	1	2	20121003144916	20121208180528	104	67	5	...	2	0	0	20	1	2	2	
4	0	unknown	3	1	1	20070311125035	20141106121057	184	2798	27	...	0	0	0	26	10	5	24	1

5 rows × 21 columns



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4746 entries, 0 to 4745
Data columns (total 21 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   gender          4746 non-null   int64
 1   C_api           4746 non-null   object
 2   C_man           4746 non-null   int64
 3   E_NEds          4746 non-null   int64
 4   E_Bpag          4746 non-null   int64
 5   firstDay        4746 non-null   int64
 6   lastDay         4746 non-null   int64
 7   NEds            4746 non-null   int64
 8   NDays           4746 non-null   int64
 9   NActDays        4746 non-null   int64
 10  NPages          4746 non-null   int64
 11  NPcreated       4746 non-null   int64
 12  pagesWomen      4746 non-null   int64
 13  wikiprojWomen   4746 non-null   int64
 14  ns_user         4746 non-null   int64
 15  ns_wikipedia    4746 non-null   int64
 16  ns_talk         4746 non-null   int64
 17  ns_userTalk     4746 non-null   int64
 18  ns_content      4746 non-null   int64
 19  weightIJ        4746 non-null   float64
 20  NIJ             4746 non-null   int64
dtypes: float64(1), int64(19), object(1)
memory usage: 778.8+ KB
```

O principal que identificamos com essa vista inicial seria que as colunas de **data** estavam identificadas como colunas de *int64*, assim como as colunas de **gender** e **C_man**. Com isso em mente, fizemos algumas modificações na etapa seguinte.

2. Data Cleaning

Duas coisas foram feitas na etapa de limpeza dos dados dentro da análise exploratória:

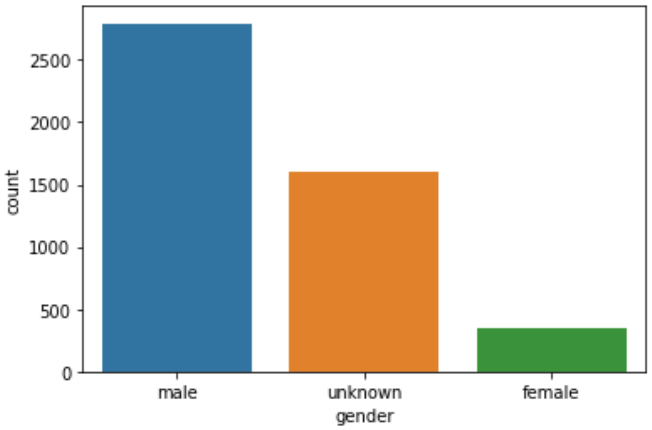
- Conversão para *datetime* das colunas **firstDay** e **lastDay** (remetem ao primeiro e ao último dia que o usuário editou alguma postagem no Wikipedia, respectivamente)
- Conversão para valores compatíveis com a coluna **C_api** (male, female, unknown) das colunas **gender** e **C_man**.

3. Check null values / duplicate rows

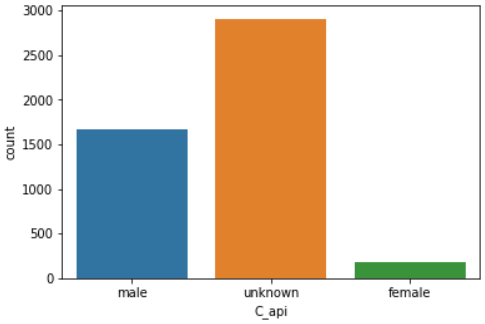
Não foram identificados valores nulos ou linhas duplicadas em nosso dataset, mas vale ressaltar que as colunas de gênero têm “unknown” como opção, o que pode ser considerado nulo em análises posteriores.

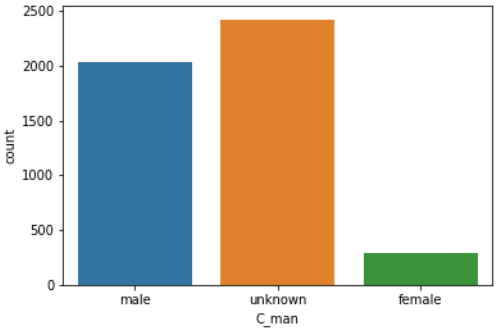
4. Plots for Categorical Data

As únicas colunas com dados categóricos do nosso dataset são as colunas de saída (output) em si. Para elas, avaliamos a distribuição de categorias referentes a cada uma:



A coluna de **gender** é a mais confiável e adequada, pois além de ter mais dados preenchidos, é a que possui o gênero que o usuário se identifica.





As outras duas colunas de gênero trazem informações automatizadas. Especificamente a coluna C_man tem uma quantidade de unknowns significativamente menor, além de mais usuários do gênero feminino.

Independente da coluna de gênero, percebemos que os usuários da Wikipedia em espanhol são, **em grande maioria**, homens. Isso pode trazer diversos problemas de representatividade e viés para o site e é o que buscamos atacar principalmente ao longo desse projeto.

5. Plots for Numerical Data

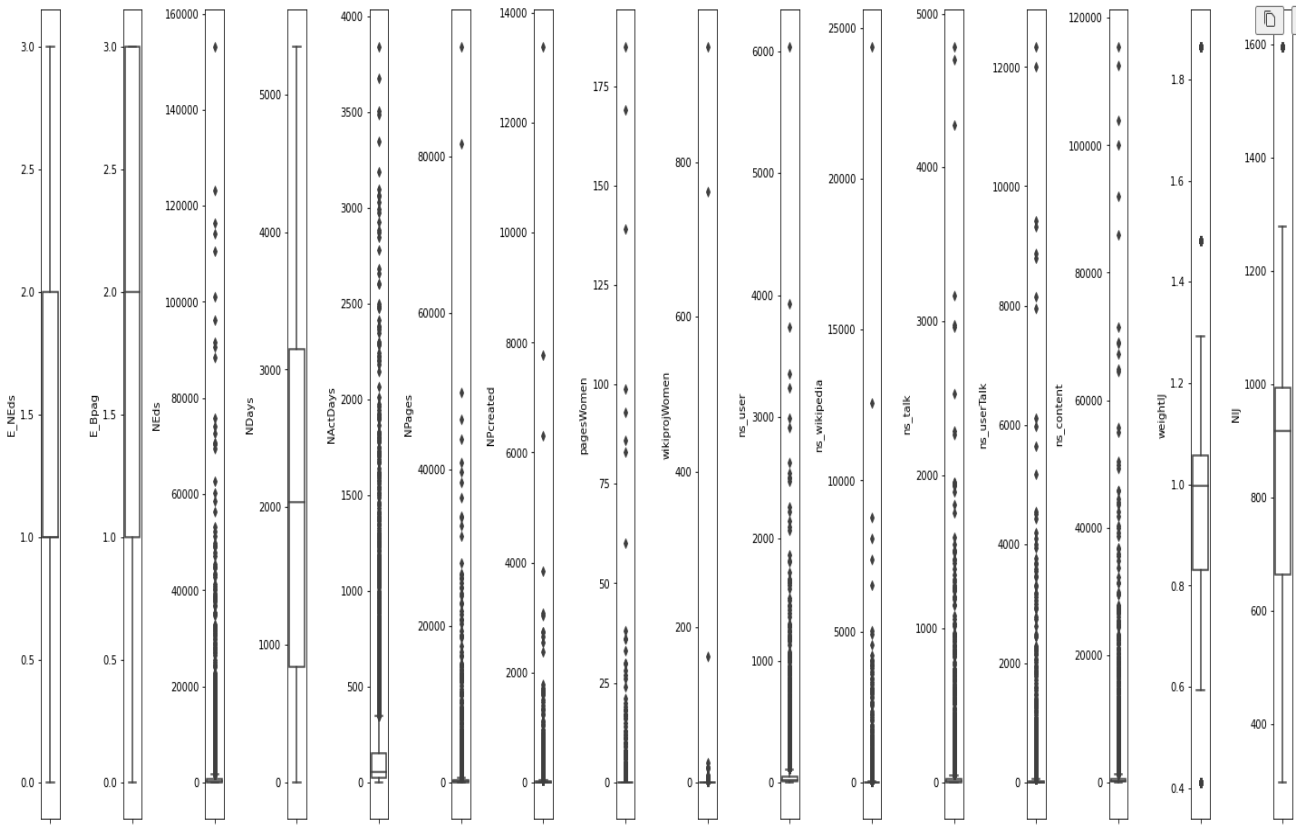
Primeiro verificamos a distribuição dos dados numéricos ao longo do dataset utilizando o método .*describe* do **pandas**.

Distribution of numeric data

	count	mean	std	min	25%	50%	75%	max
E_NEds	4746.0	1.484197	1.099795	0.000000	1.000000	1.000000	2.000000	3.000000
E_Bpag	4746.0	1.646228	1.079263	0.000000	1.000000	2.000000	3.000000	3.000000
NEds	4746.0	2029.969448	7793.300833	50.000000	95.000000	218.000000	757.750000	153193.000000
NDays	4746.0	2036.607880	1336.119914	1.000000	835.250000	2035.500000	3146.500000	5349.000000
NActDays	4746.0	183.162663	374.034481	1.000000	24.000000	53.000000	154.000000	3843.000000
NPages	4746.0	689.451960	3355.302483	1.000000	29.000000	68.000000	219.750000	94142.000000
NPcreated	4746.0	43.479140	297.395507	0.000000	1.000000	4.000000	14.000000	13394.000000
pagesWomen	4746.0	0.438896	5.327440	0.000000	0.000000	0.000000	0.000000	185.000000
wikiprojWomen	4746.0	0.439949	17.832244	0.000000	0.000000	0.000000	0.000000	949.000000
ns_user	4746.0	74.372946	246.407233	1.000000	4.000000	14.000000	46.000000	6041.000000
ns_wikipedia	4746.0	74.368310	560.782479	0.000000	0.000000	1.000000	8.000000	24392.000000
ns_talk	4746.0	49.947745	215.554281	0.000000	0.000000	4.000000	19.000000	4788.000000
ns_userTalk	4746.0	96.081753	545.025818	0.000000	1.000000	5.000000	22.000000	12350.000000
ns_content	4746.0	1521.886641	6099.009235	0.000000	61.000000	151.000000	563.750000	115547.000000
weightIJ	4746.0	1.000000	0.325763	0.411985	0.831954	0.997535	1.057149	1.865008
NIJ	4746.0	867.148546	325.933076	297.000000	664.000000	917.000000	994.000000	1596.000000

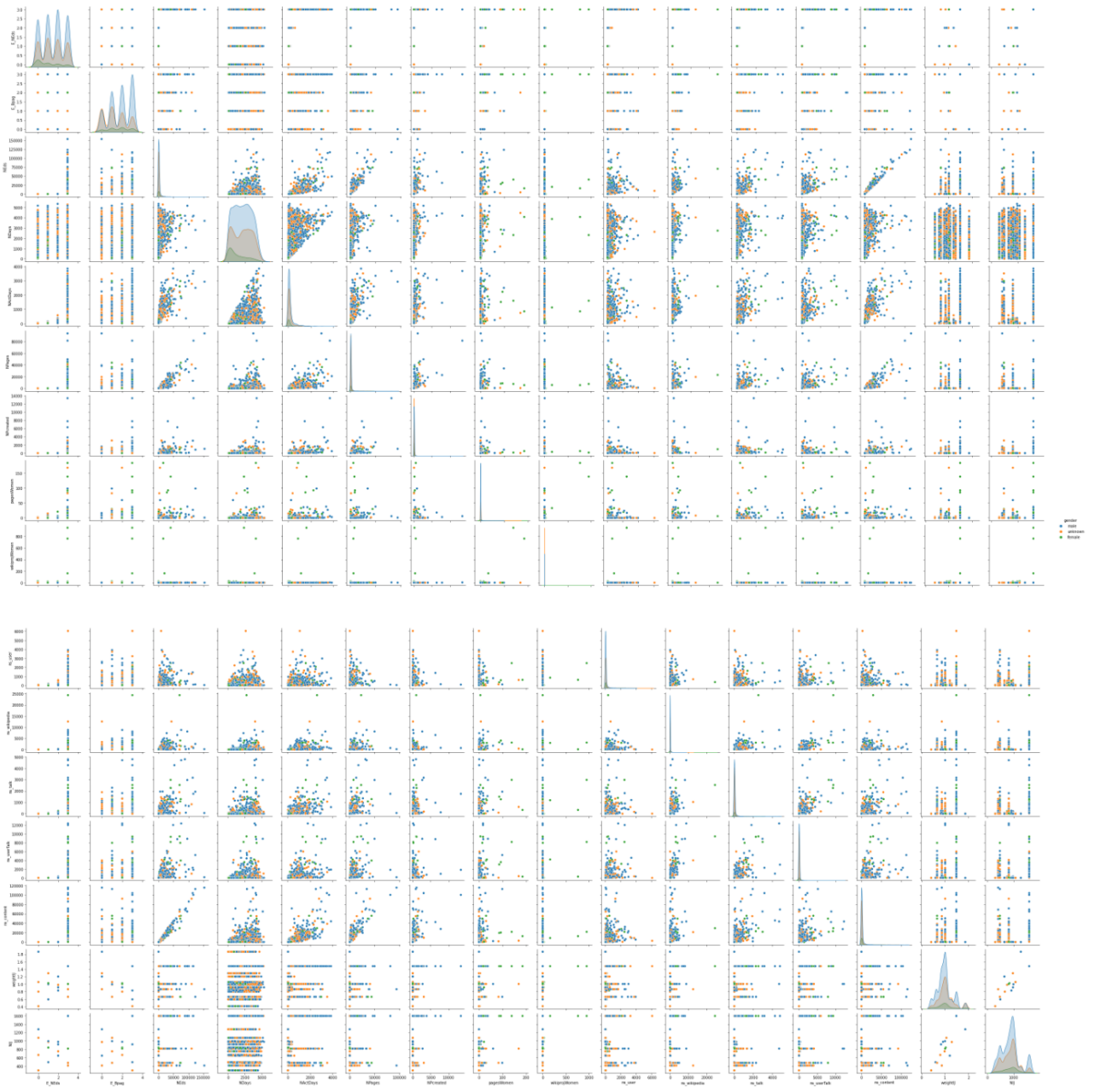
Aqui, percebemos que o dataset tem várias colunas com distribuições extremamente distorcidas, distantes de uma normal. Principalmente, quando olhamos para colunas de contagem de dias, páginas e edições.

Essa ideia foi reforçada quando observamos o *boxplot* dessas variáveis:

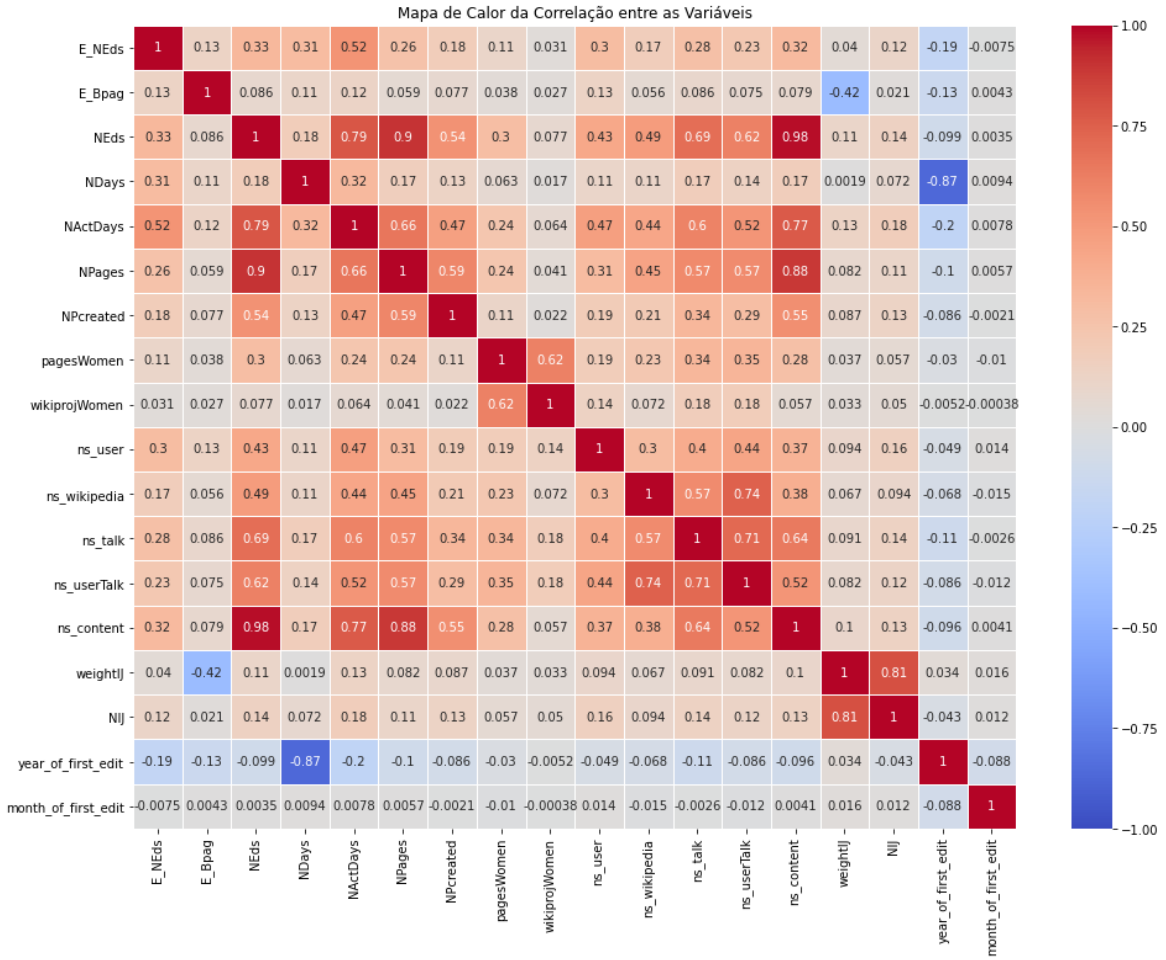


Através dele, podemos perceber a quantidade enorme de *outliers* presente em diversas colunas da base de dados. Isso definitivamente é algo a se manter em mente quando for a hora de modelar em cima desse dataset.

Nesta sessão, também avaliamos a distribuição dos dados conforme o gênero e suas relações entre si através de um *pairplot* com *hue=”**gender**”* . Por meio dele percebemos que, apesar de algumas diferenças específicas, o comportamento entre os dois gêneros é relativamente similar em dados numéricos. Assim, se modelarmos visando identificar gênero por padrões de comportamento, vamos ter que extrair o máximo dessas sutis diferenças.



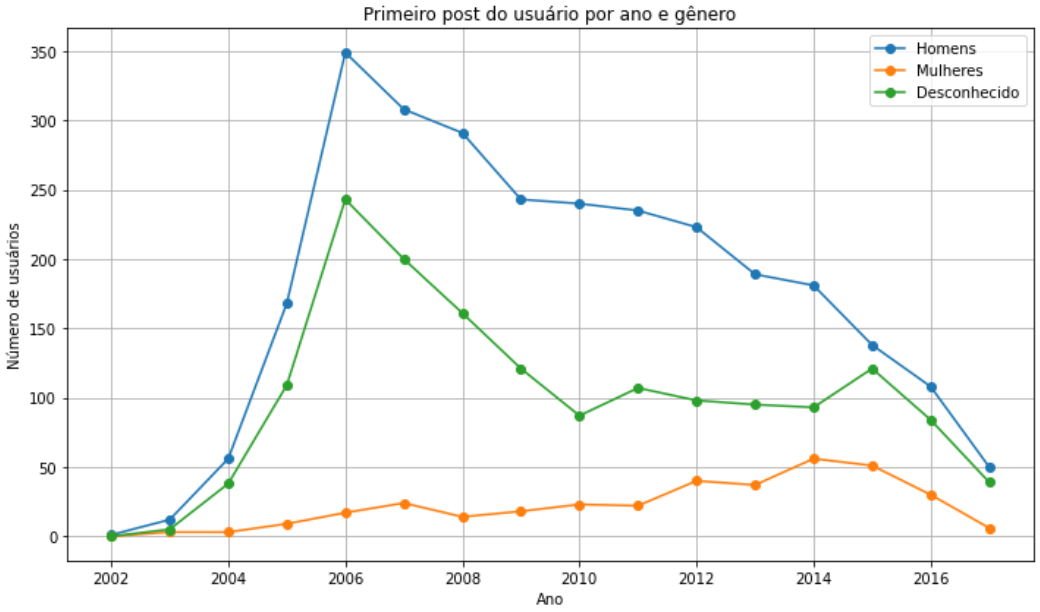
Olhar esses gráficos nos fez pensar também sobre a correlação entre as variáveis numéricas. A partir disso, decidimos utilizar um *plot* de *heatmap* de correlações para enxergar isso melhor.



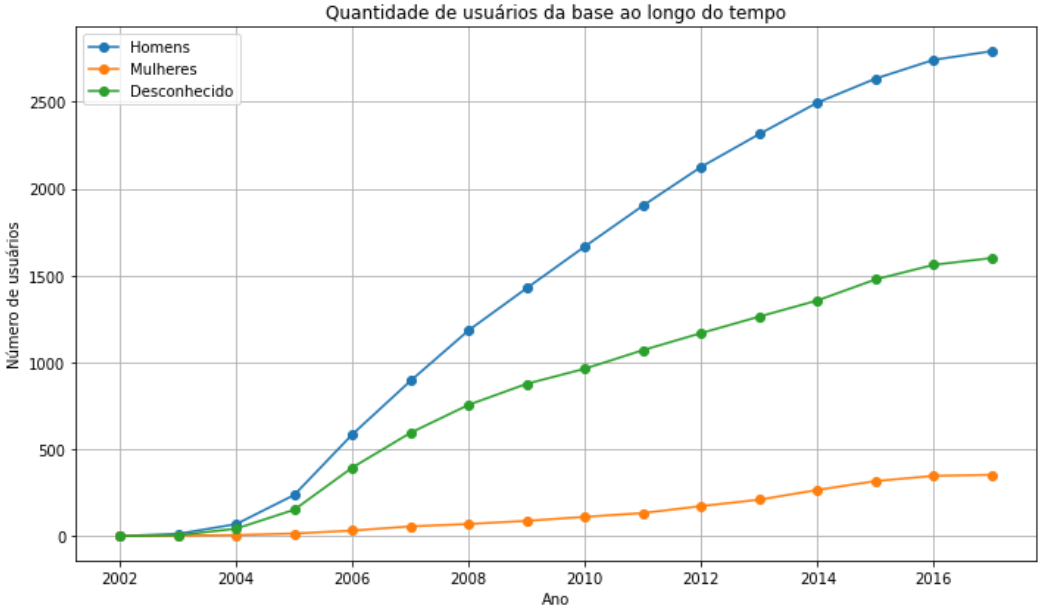
Esse gráfico foi muito informativo e podemos perceber correlações fortíssimas entre algumas colunas do dataset. Dentre outras coisas, isso pode nos sugerir o uso de algum algoritmo de **redução de dimensionalidade** ou a eliminação completa de algumas variáveis.

6. Datetime column exploration

Encerrando a parte de gráficos, decidimos investigar um pouco mais sobre a progressão temporal e a sazonalidade da vinda de novos usuários para a Wikipedia em espanhol e entender como cada gênero se comportou ao longo desses recortes.

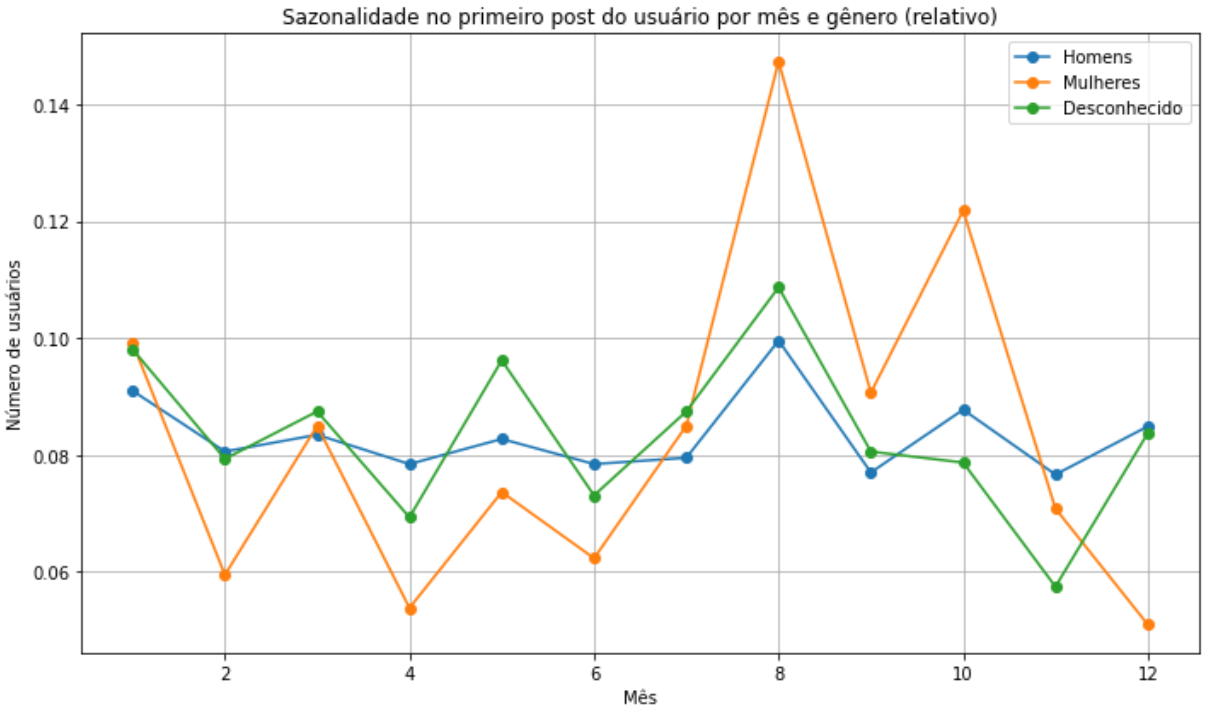


Esse primeiro gráfico mostra que a entrada de usuários femininos cresceu entre 2006 e 2014, enquanto usuários novos do gênero masculino ou não identificado se tornaram cada vez mais raros. A partir de 2014, no entanto, a entrada de novos usuários é cada vez mais incomum.



Apesar disso, como indicado nesse segundo gráfico, a quantidade de usuários do gênero feminino cresceu muito menos do que a de usuários do gênero masculino ou de gênero não identificado, em geral.

Para entender a sazonalidade da entrada de usuários de cada gênero ao longo dos meses, utilizamos um plot que mostrava a porcentagem de usuários de cada gênero que começava a postar a cada mês.



Dentre todos os gráficos, esse foi o mais informativo, indicando que a maioria das usuárias do gênero feminino começam a postar entre agosto e outubro. Talvez por conta das pautas abordadas no setembro amarelo, outubro rosa, etc.

7. Análise crítica sobre a complexidade da base de dados e com respeito da separação das classes

Concluindo nossa análise, após um debate entre os participantes do grupo, entendemos que o dataset que nos foi indicado é relativamente complexo se estivermos com o objetivo final de utilizar aprendizagem de máquina para uma tarefa de classificação. Não notamos uma separação clara entre variáveis de cada gênero na maioria das features do dataset, apesar de que podemos capitalizar em cima das que apresentam tal distinção (como as variáveis de data).

Em suma, o dataset é desafiador mas não impossível de ser utilizado de forma prática para uma atividade de aprendizagem de máquina. Assim, vamos seguir investigando ao longo das próximas entregas e, se necessário, pivotar nosso objetivo final para algo mais próximo de uma análise temporal ou algum tipo de regressão.