Analysis of Linked GitHub and Wikidata

Ekaterina Levitskaya[1], Gizem Korkmaz[2], Daniel Mietchen[3], Lane Rasberry[4]

1. Coleridge Initiative, ekaterina.levitskaya@coleridgeinitiative.org
2. Coleridge Initiative, gizem.korkmaz@coleridgeinitiative.org
3. Ronin Institute, daniel.mietchen@ronininstitute.org
4. University of Virginia, School of Data Science, lr2ua@virginia.edu

We aim to study the features of GitHub developers that have corresponding Wikidata entries. In this report, we focus on GitHub developers that are associated with academic institutions. First, we explore the countries of these developers to understand major countries that contribute to open source software (OSS). Second, we explore whether there are gender differences with respect to OSS contributions. We analyze number of repositories owned and contributes to, number of commits and code additions. Finally we generate a collaboration network, and study the degree centrality of the users.

About the sample: As noted above, this report explores a sample of OSS contributors. The sample is obtained by extracting GitHub user logins that are associated with academic institutions. This information is obtained from their profile information as well as email domains using the tidyorgs package [cite]. We obtained 109,752 number of users. We used the query below to identify users that are in Wikidata. We obtained 1,481 number of individuals. This notebook provides the output of the analyses. The code is available here: https://github.com/open-source-software-project/open-source-software-project/tree/gh-pages/Notebooks.

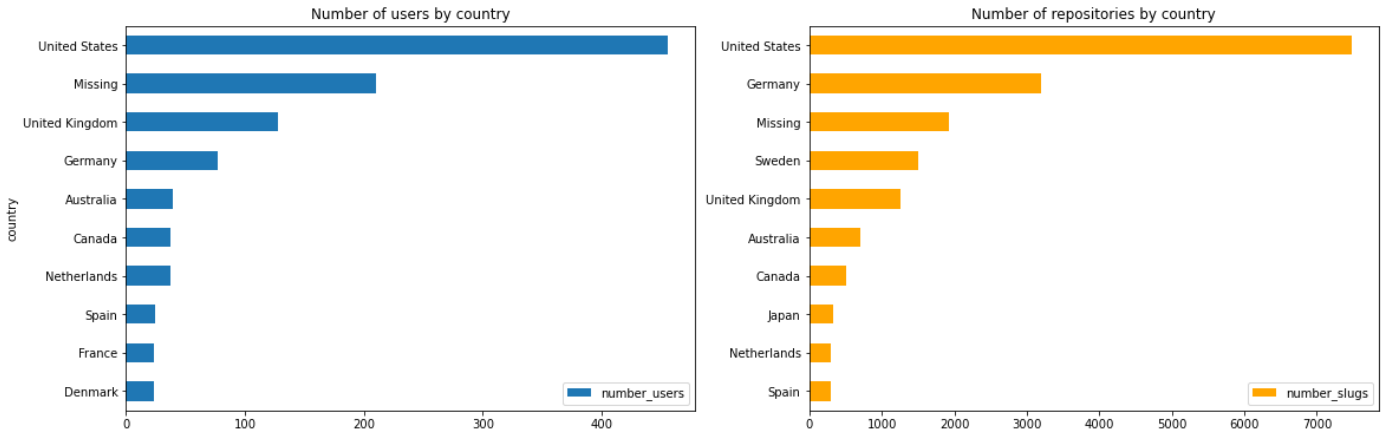
Table of contents

Country Analysis

In this analysis, we use the country information reported by the users on their GitHub profiles. Note that not all the users have country information. The country was pulled based on the GitHub metadata. ~ 14.18% of users is missing country. "Missing" is included as a category.

We explore the number of users in each country and the number of repositories developed by these users. We report the top 10 countries in the bar charts below.

In [79]:
```
fig, (ax0, ax1) = plt.subplots(1,2, figsize=(19,6))
number_users.plot(x='country', y='number_users', kind='barh', ax=ax0, title = 'Number of
number_slugs.plot(x='country', y='number_slugs', color='orange', kind='barh',ax=ax1, tit
plt.ylabel('')
plt.show()
```



Note on the number of repositories: This is based on the table where each user has the corresponding total number of slugs, and the counts are taken by country based on that number. There could be overlaps between these repositories (i.e. the same repository counted more than one time), because here we do not take into account if the users contributed to the same repository. It is based on the number of repositories by user.

Summary: Based on the figures above, we observe that the top countries by number of users in the Wikidata sample is United States, UK, and Germany, but by the number of repositories (i.e. productivity) the top three countries are slightly different: United States, Germany, and Sweden. Also, Japan appears in the top 10 by the number of repositories, while it is not in top 10 by the number of users. We plan to look further into different measures of productivity of users by country.

Gender Analysis

We created a gender variable in the Wikidata sample based on the following logic:

- if gender is *not missing* in the Wikidata, we use that as the basis/true data (as the Wikidata is self-reported/crowdsourced).
- where gender is *missing* in the Wikidata, we impute it with the `gender-guesser` Python package: https://pypi.org/project/gender-guesser/
  - this package has the following categories:
    - `male`
    - `female`
    - `mostly_male`
    - `mostly_female`
    - `unknown`
    - `andy` (androgynous)
    - `non-binary`

We included `mostly_male` and `mostly_female` in the respective `male` and `female` categories; `andy` and `non-binary` are included in the `other` category, `unknown` is the category on its own.

Breakdown by the finalized gender categories:

In [26]:
```
user_gender
```

Out[26]:

	gender	login	total	percentage
1	male	1086	1481	73.33
0	female	186	1481	12.56
3	unknown	170	1481	11.48
2	other	39	1481	2.63

In this sample 12.56% is categorized is female. Even though we do not have a representative sample, these proportions are consistent with the following study that states that "... In a 2013 survey of the more than 2000 open source developers who indicated a gender, only 11.2% were women (Arjona-Reina, Robles & Dueas, 2014)." Article source that cited this survey: https://peerj.com/articles/cs-111/

We developed descriptive statistics and exploratory visualizations by number of repositories, commits, and additions (lines in the code added) by gender.

According to the article mentioned above, men tend to contribute to more repositories than women (broad, more extensive contributions), but perhaps women tend to focus their efforts more on fewer repositories, as evidenced by a higher activity on those (narrow, more intensive contributions).

- In this article, the authors found that women tend to have higher acceptance rates of their pull requests, and they had the following hypothesis: "One possible explanation for women's higher acceptance rates is that they are focusing their efforts more than men; perhaps their success is explained by doing pull requests on few projects, whereas men tend to do pull requests on more projects. First, the data do suggest that women tend to contribute to fewer projects than men. Source: https://peerj.com/articles/cs-111/#utable-1

Below, we analyze the number of repositories, commits, and line additions by different gender categories.

Note on the data pre-processing: one user didn't have information on number of slugs, commits, additions. Total population for the analysis below: 1,480 individuals. Two users had duplicate entries, as the information on the full name was slightly different - duplicate entries were removed, as it didn't influence the assignment of gender for those users.

Number of repositories by gender

Here we generate descriptive statistics for the number of repositories.

We observe that the average number of repositories is higher for males than females, but the overall distribution is long-tailed, not normal (please see the visualizations below). We are currently investigating what would be the optimal way to compare the differences, taking into account the long-tailed distributions.

In [8]:
```
user_node.groupby('gender')['number_slugs'].describe()
```

Out[8]:

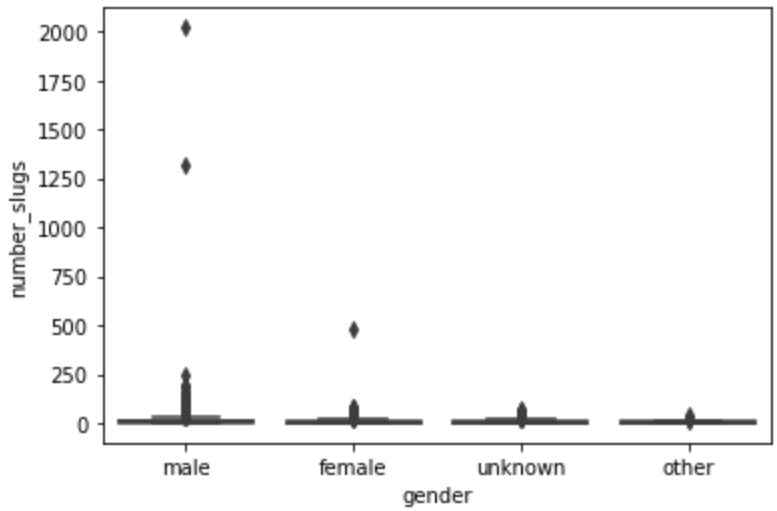
gender	count	mean	std	min	25%	50%	75%	max
female	185.0	10.470270	37.252523	1.0	1.0	4.0	9.0	483.0

male	1086.0	17.151934	76.399580	1.0	3.0	6.0	15.0	2020.0
other	39.0	6.025641	8.044243	1.0	1.0	3.0	6.0	42.0
unknown	170.0	7.970588	11.688474	1.0	2.0	4.0	8.0	73.0

Boxplots illustraring the distribution of the number of repositories by gender groups

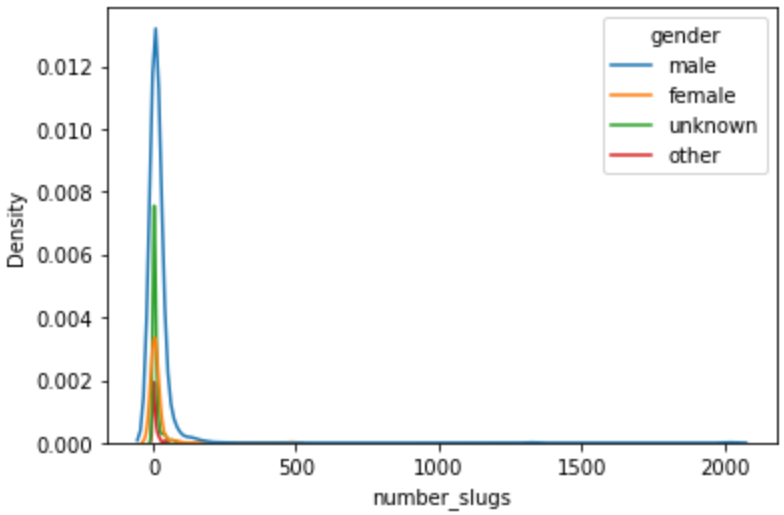
In [9]:
```
plt.ticklabel_format(style='plain', axis='x')
seaborn.boxplot(x='gender', y='number_slugs',
            data=user_node)
```

Out[9]:
```
<AxesSubplot:xlabel='gender', ylabel='number_slugs'>
```



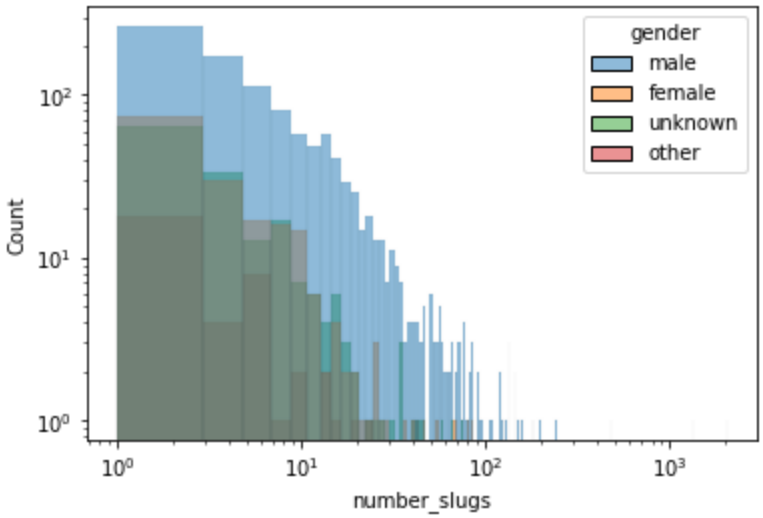
Density plot for the number of repositories

In [11]:
```
seaborn.kdeplot(x='number_slugs',data=user_node,hue='gender')
plt.show()
```



Logged histogram for the number of repositories

In [10]:
```
ax = seaborn.histplot(x='number_slugs',data=user_node,hue='gender')
ax.set_xscale('log')
ax.set_yscale('log')
plt.show()
```



Subset to the population between 5th and 95th quantiles

As our distribution have very long tails, we remove the extreme observations by taking the subset of the population between 5th and 95th quantile.

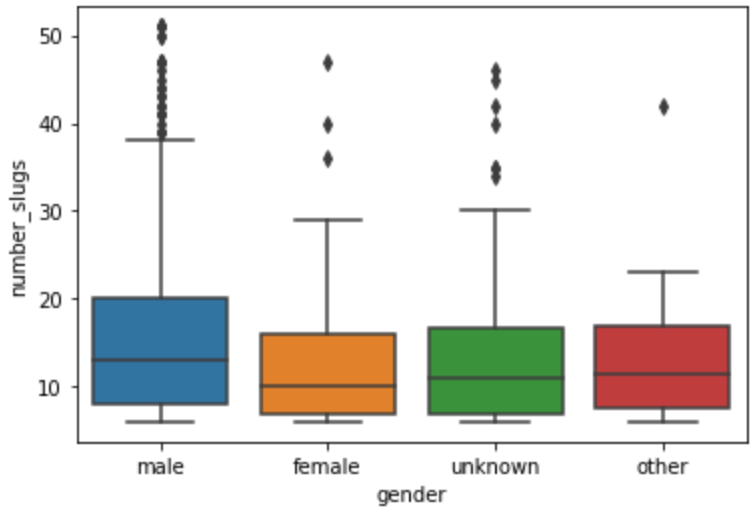
In [18]:
```
number_slugs.groupby('gender')['number_slugs'].describe()
```

Out[18]:

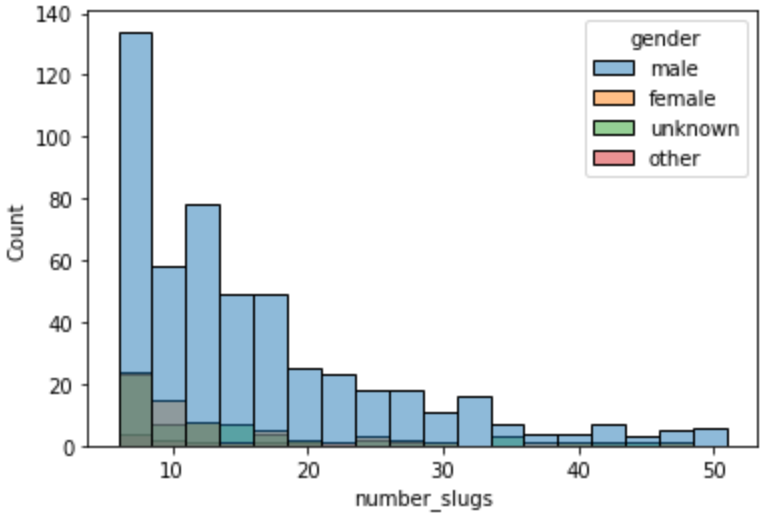
gender	count	mean	std	min	25%	50%	75%	max
female	65.0	12.984615	8.697148	6.0	7.0	10.0	16.00	47.0
male	515.0	15.912621	9.991345	6.0	8.0	13.0	20.00	51.0
other	12.0	14.416667	10.326122	6.0	7.5	11.5	17.00	42.0
unknown	62.0	14.725806	10.544472	6.0	7.0	11.0	16.75	46.0

In [19]:
```
plt.ticklabel_format(style='plain', axis='x')
seaborn.boxplot(x='gender', y='number_slugs',
            data=number_slugs)
```

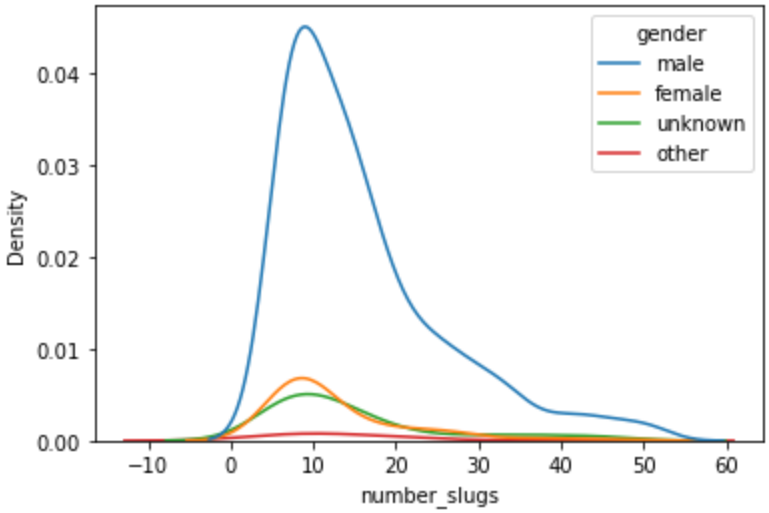
Out[19]:
```
<AxesSubplot:xlabel='gender', ylabel='number_slugs'>
```



In [20]:
```
ax = seaborn.histplot(x='number_slugs',data=number_slugs,hue='gender')
plt.show()
```



In [22]:
```
seaborn.kdeplot(x='number_slugs',data=number_slugs,hue='gender')
plt.show()
```



Number of commits by gender

Here we generate descriptive statistics for the number of commits.

We observe that the average number of commmits is higher for males than females, however, as mentioned above, the overall distribution is also long-tailed, not normal (please see the visualizations below). We are currently investigating what would be the optimal way to compare the differences, taking into account the long-tailed distributions.

We would also like to look into whether women tend to commit more intensively to the same repositories, while men commit more extensively, to different repositories.

In [23]:
```
user_node.groupby('gender')['number_commits'].describe()
```

Out[23]:

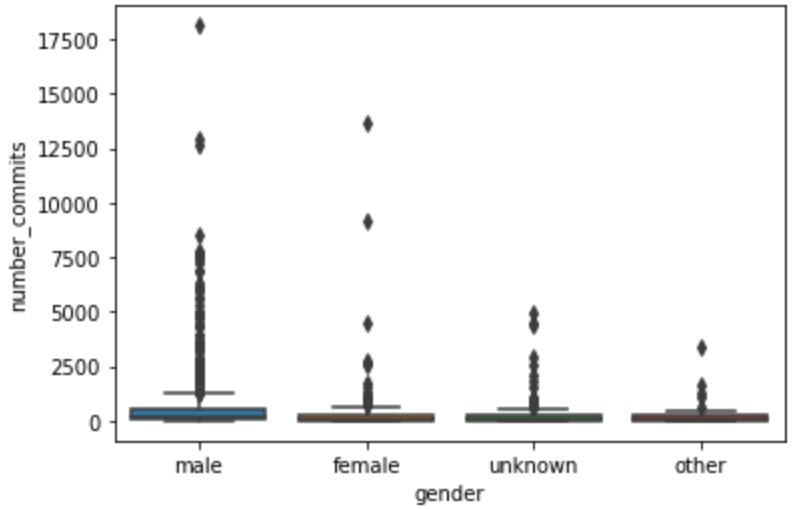
gender	count	mean	std	min	25%	50%	75%	max
female	185.0	383.994595	1297.765695	1.0	15.00	66.0	291.00	13688.0
male	1086.0	611.187845	1322.171883	1.0	42.00	167.0	551.25	18121.0
other	39.0	285.948718	633.138671	2.0	23.00	55.0	248.50	3411.0

unknown	170.0	310.511765	713.721758	1.0	28.25	87.5	255.50	4967.0

Boxplots illustraring the distribution of the number of commits by gender groups

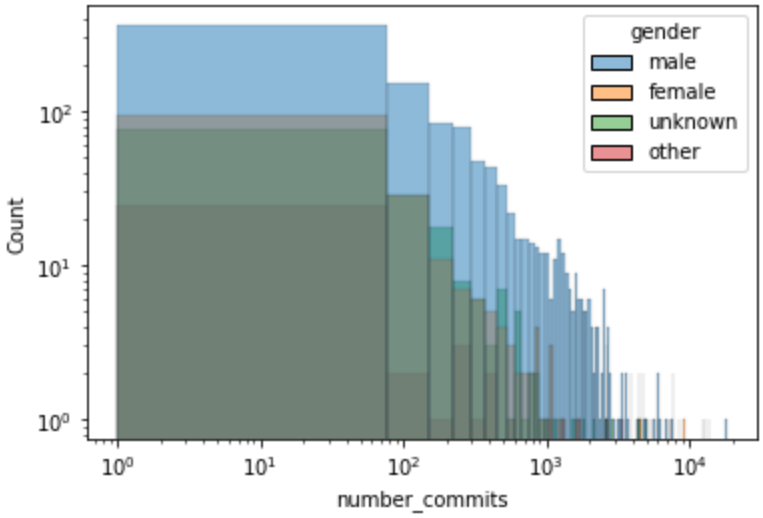
In [24]:
```
plt.ticklabel_format(style='plain', axis='x')
seaborn.boxplot(x='gender', y='number_commits',
            data=user_node)
```

Out[24]:
```
<AxesSubplot:xlabel='gender', ylabel='number_commits'>
```



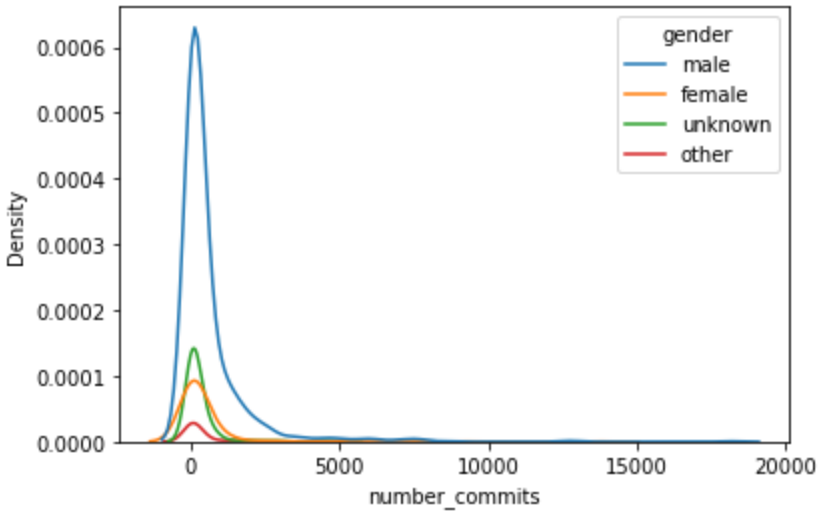
Logged histogram for the number of commits

In [25]:
```
ax = seaborn.histplot(x='number_commits',data=user_node,hue='gender')
ax.set_xscale('log')
ax.set_yscale('log')
plt.show()
```



Density plot for the number of commits

In [26]:
```
seaborn.kdeplot(x='number_commits',data=user_node,hue='gender')
plt.show()
```



Subset to the population between 5th and 95th quantiles

As our distribution have very long tails, we remove the extreme observations by taking the subset of the population between 5th and 95th quantile.

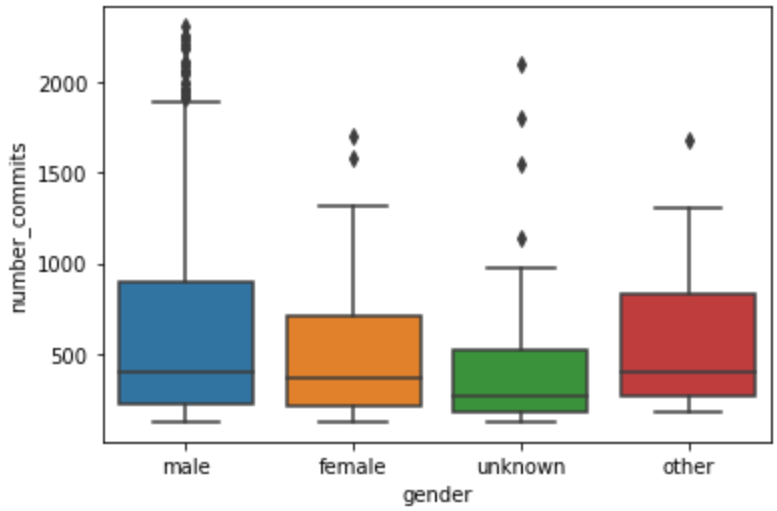
In [30]:
```
number_commits.groupby('gender')['number_commits'].describe()
```

Out[30]:

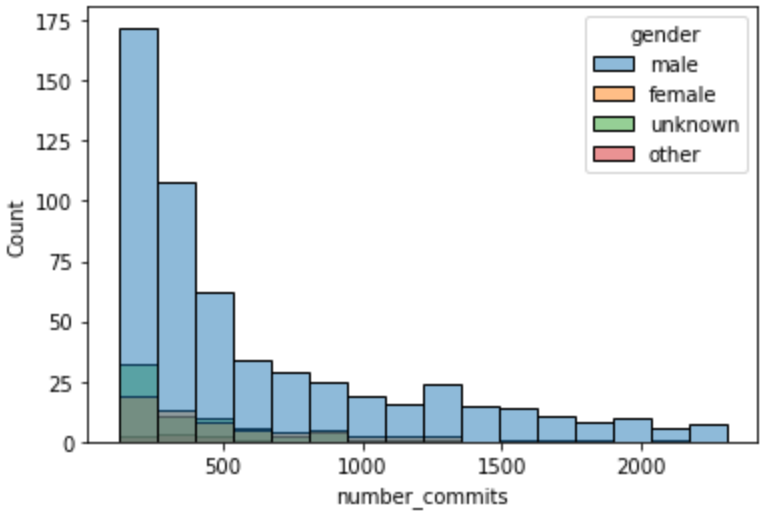
gender	count	mean	std	min	25%	50%	75%	max
female	62.0	506.725806	378.125106	122.0	204.00	367.0	707.50	1709.0
male	560.0	622.750000	537.478099	121.0	224.75	395.5	895.25	2310.0
other	11.0	614.454545	507.962669	176.0	268.50	402.0	832.50	1682.0
unknown	70.0	423.128571	384.842182	121.0	171.00	266.0	514.75	2107.0

In [31]:
```
plt.ticklabel_format(style='plain', axis='x')
seaborn.boxplot(x='gender', y='number_commits',
            data=number_commits)
```

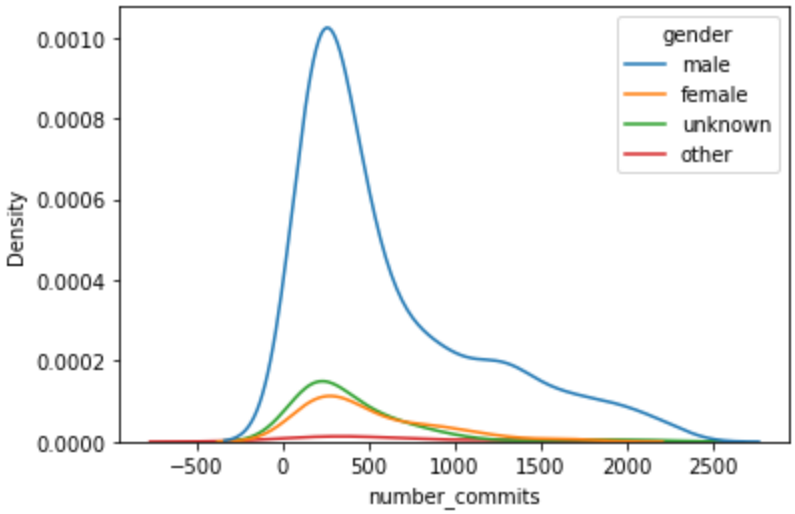
Out[31]:
```
<AxesSubplot:xlabel='gender', ylabel='number_commits'>
```



In [32]:
```
ax = seaborn.histplot(x='number_commits',data=number_commits,hue='gender')
plt.show()
```



In [33]:
```
seaborn.kdeplot(x='number_commits',data=number_commits,hue='gender')
plt.show()
```



Number of additions by gender

Here we generate descriptive statistics for the number of lines added.

We observe that the average number of lines added is higher for females, but, as mentioned above, the distribution is not normal; outliers may be pulling the mean higher.

In [34]:
```
user_node.groupby('gender')['number_additions'].describe().round(1)
```

Out[34]:

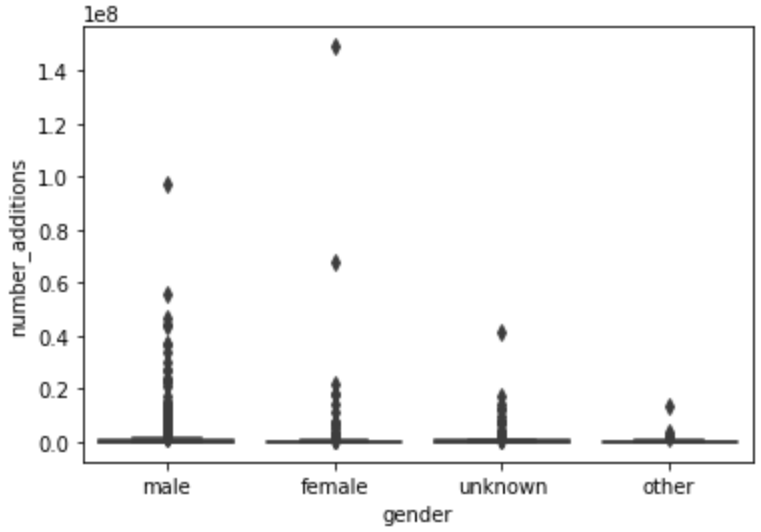
gender	count	mean	std	min	25%	50%	75%	max
female	185.0	2026532.3	12308593.4	0.0	1445.0	29491.0	271713.0	149080533.0
male	1086.0	1422974.8	5328611.7	0.0	13099.5	122162.5	721692.2	97394457.0
other	39.0	672802.3	2295155.6	125.0	7319.0	27019.0	282055.0	13876413.0
unknown	170.0	1171781.1	4047907.6	0.0	7936.8	67547.0	387783.5	41747156.0

Boxplots illustraring the distribution of the number of additions by gender groups

In [35]:
```
plt.ticklabel_format(style='plain', axis='x')
```

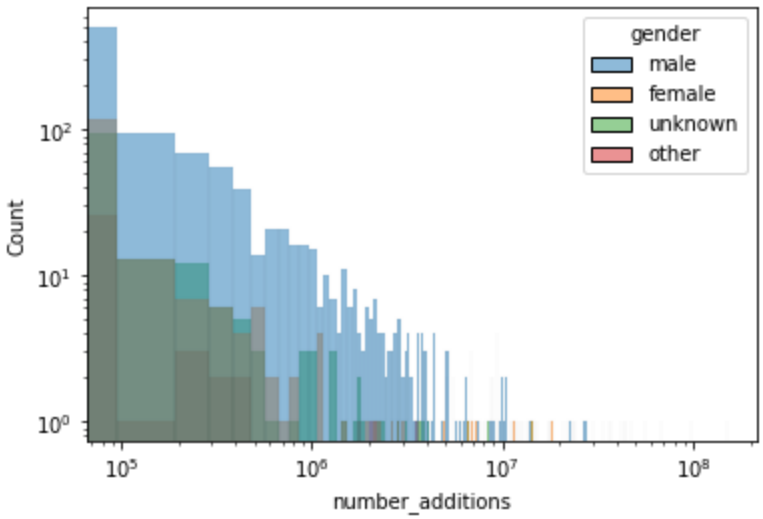
```
seaborn.boxplot(x='gender', y='number_additions',
            data=user_node)
```

Out[35]:
```
<AxesSubplot:xlabel='gender', ylabel='number_additions'>
```



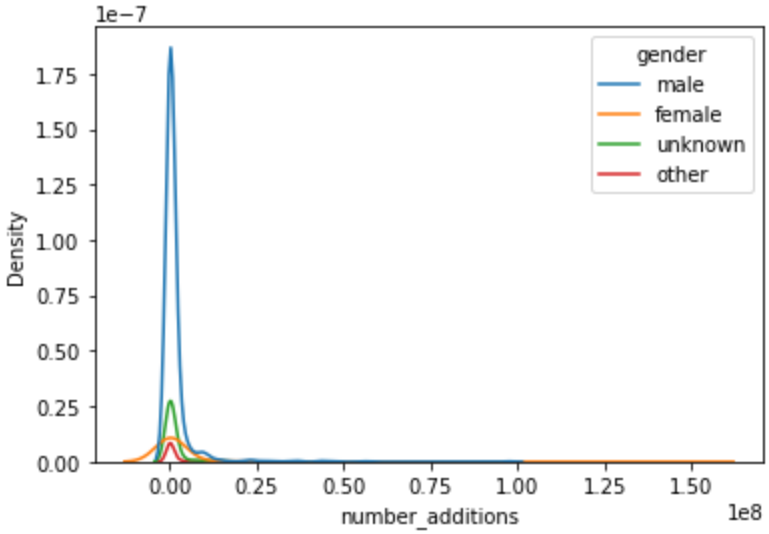
Logged histogram for the number of additions

In [36]:
```
ax = seaborn.histplot(x='number_additions',data=user_node,hue='gender')
ax.set_xscale('log')
ax.set_yscale('log')
plt.show()
```



Density plot for the number of additions

In [37]:
```
seaborn.kdeplot(x='number_additions',data=user_node,hue='gender')
plt.show()
```



Subset to the population between 5th and 95th quantiles

As our distribution have very long tails, we remove the extreme observations by taking the subset of the population between 5th and 95th quantile.

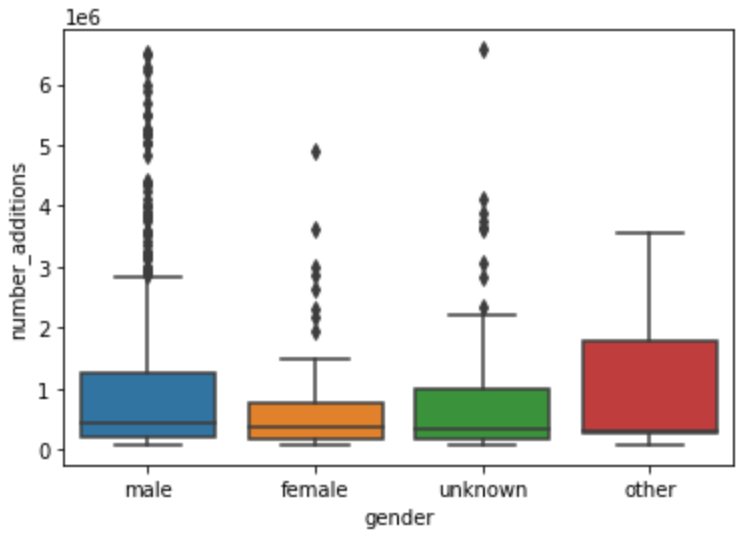
In [41]:
```
number_additions.groupby('gender')['number_additions'].describe().round(1)
```

Out[41]:

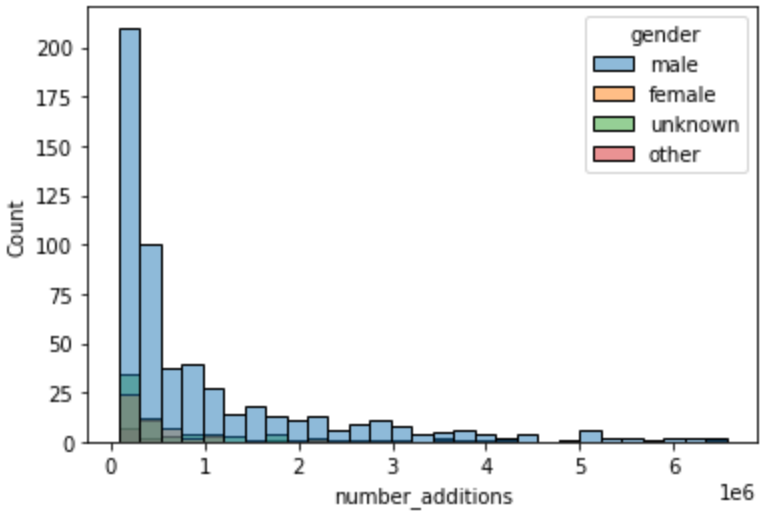
gender	count	mean	std	min	25%	50%	75%	max
female	58.0	751309.0	988289.5	82112.0	176425.0	353535.5	771184.0	4916128.0
male	559.0	1015950.5	1286775.3	74986.0	196408.0	435538.0	1267878.5	6512537.0
other	13.0	922050.8	1088194.5	85397.0	278915.0	297897.0	1784837.0	3543041.0
unknown	73.0	911687.4	1250471.6	76610.0	170828.0	334535.0	1008522.0	6571262.0

In [46]:
```
plt.ticklabel_format(style='plain', axis='x')
seaborn.boxplot(x='gender', y='number_additions',
            data=number_additions)
```

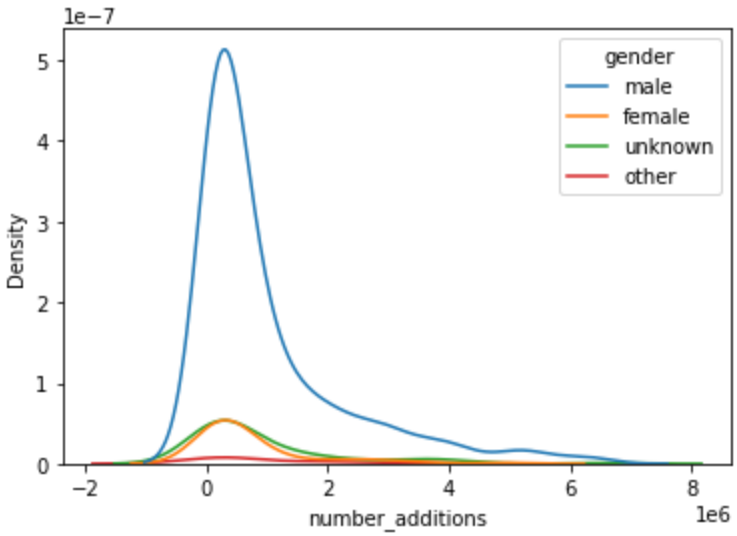
Out[46]:
```
<AxesSubplot:xlabel='gender', ylabel='number_additions'>
```



In [47]:
```
ax = seaborn.histplot(x='number_additions',data=number_additions,hue='gender')
plt.show()
```



In [48]:
```
seaborn.kdeplot(x='number_additions',data=number_additions,hue='gender')
plt.show()
```



Summary: We would like to investigate further whether women tend to contribute more intensively to fewer repositories, while men tend to contribute more expansively, to more repositories.

The authors of the article mentioned above suggest that women tend to have higher acceptance rates of pull requests due to these possible reasons:

- "Another theory is that women in open source are, on average, more competent than men. In Lemkau's review of the psychology and sociology literature, she found that women in male-dominated occupations tend to be highly competent (Lemkau, 1979). (...) Assuming this final theory is the best one, why might it be that women are more competent, on average? One explanation is survivorship bias: as women continue their formal and informal education in computer science, the less competent ones may change fields or otherwise drop out. Then, only more competent women remain by the time they begin to contribute to open source. In contrast, less competent men may continue. While women do switch away from STEM majors at a higher rate than men, they also have a lower drop out rate then men (Chen, 2013), so the difference between attrition rates of women and men in college appears small. Another explanation is self-selection bias: the average woman in open source may be better prepared than the average man, which is supported by the finding that women in open source are more likely to hold Master's and PhD degrees (Arjona-Reina, Robles & Dueas, 2014). "

Source article: https://peerj.com/articles/cs-111/#utable-1

Contributions to repository owners by gender (in progress)

Here we generate counts by the gender of contributors and the gender of the owners of the repositories that they contribute to.

We imputed the gender of repository owners based on the methodology described above (first names). We obtained the full names from the GitHub API, based on the user names. Note: the owner's name could be a general organization name, not a personal name, which would be then classified as "unknown".

In [70]:
```
login_owner.groupby(['login_gender','owner_gender']).size().reset_index().rename(columns
```

Out[70]:

	contributor_gender	owner_gender	number_pairs
0	female	female	123
1	female	male	181
2	female	other	7
3	female	unknown	520
4	male	female	313
5	male	male	2146
6	male	other	64
7	male	unknown	4212
8	other	female	4
9	other	male	10
10	other	other	1
11	other	unknown	72
12	unknown	female	22
13	unknown	male	71
14	unknown	other	5
15	unknown	unknown	388

Based on the table above, we observe that females, out of total contributions, are more likely to contribute to a female-owned repository (40.46%) than males (12.71%).

In [114…
```
# Percentage of female contributions to repositories with female owners
# Denominator is the total female contributions to female+male owners
(123/(123+181))*100
```

Out[114]:
```
40.46052631578947
```

In [76]:
```
# Percentage of male contributions to repositories with female owners
# Denominator is the total male contributions to female+male owners
(313/(313+2146))*100
```

Out[76]:
```
12.713241267262386
```

Based on the table above, we observe that males, out of total contributions, are more likely to contribute to a male-owned repository (87.28%) than females (59.53%).

In [77]:
```
# Percentage of male contributions to repositories with male owners
# Denominator is the total male contributions to female+male owners
(2146/(313+2146))*100
```

Out[77]:
```
87.28675873273761
```

In [79]:
```
# Percentage of female contributions to repositories with male owners
# Denominator is the total female contributions to female+male owners
(181/(123+181))*100
```

Out[79]:
```
59.539473684210535
```

Collaborations by gender (in progress)

We generate the counts for the collaborations between users by gender.

For each unique user, we identify their collaborators' gender on the same repository. We use the weighted number of connections - e.g. if user 1 collaborated with user 2 on six different repositories, the weight will be 6. Then the weight is summed by the "user-collaborator" groups: "female user-female collaborator", "female user-male collaborator", "male user-male collaborator", etc.

In [47]:
```
user_edge_gender_updated.groupby(['user_gender','collaborator_gender'])['weight'].sum().
```

Out[47]:

	user_gender	collaborator_gender	weight
0	female	female	35
1	female	male	230
2	female	other	5
3	female	unknown	21
4	male	female	252
5	male	male	3542
6	male	other	14
7	male	unknown	199
8	unknown	female	22
9	unknown	male	198
10	unknown	other	1
11	unknown	unknown	14

Based on the table above, we observe that there is a higher likelihood of females collaborating on the same repository with females (13%) than of males collaborating on the same repository with females (6%).

In [80]:
```
# Percentage of females collaborating on the same repository with females
# Denominator is the total number of female+male connections
(35/(35+230))*100
```

Out[80]:
```
13.20754716981132
```

In [82]:
```
# Percentage of males collaborating on the same repository with females
# Denominator is the total number of female+male connections
(252/(252+3542))*100
```

Out[82]:
```
6.642066420664207
```

Based on the table above, we observe that there is a higher likelihood of males collaborating on the same repository with males (93%) than of females collaborating on the same repository with males (86%).

In [83]:
```
# Percentage of males collaborating on the same repository with males
# Denominator is the total number of female+male connections
(3542/(252+3542))*100
```

Out[83]:
```
93.35793357933579
```

In [81]:
```
# Percentage of females collaborating on the same repository with males
# Denominator is the total number of female+male connections
(230/(35+230))*100
```

Out[81]:
```
86.79245283018868
```

Degree, number of connections, by gender (in progress)

Based on the number of connections in the network, we hypothesized whether women would have 0 connections more often than men (i.e. women would be less central in the network).

This table shows the breakdown of individuals by gender with 0 connections in the network:

In [33]:
```
degree_gender
```

Out[33]:

	gender	zero_connections_count	total	percentage
0	male	735	1086	67.68
1	unknown	144	170	84.71
2	female	129	185	69.73
3	other	35	39	89.74

Based on the table above, we observe that slightly higher percentage of women (69.73%) has 0 connections than men (67.68%).

We generate descriptive statistics for the number of connections by gender.

We use the weighted degree, or the weighted number of connections. For example, if the user B worked with the user C on 6 repositories, the weight of their collaboration will be 6. We take into account the weighted number of all collaborations of the user. If the user B collaborated with the user C on 6 repositories and with the user D on 1 repository, the total number of weighted collaborations, weighted degree, of the user B will be 7.

In [41]:
```
node_degree.groupby('gender')['Weighted Degree'].describe()
```

Out[41]:

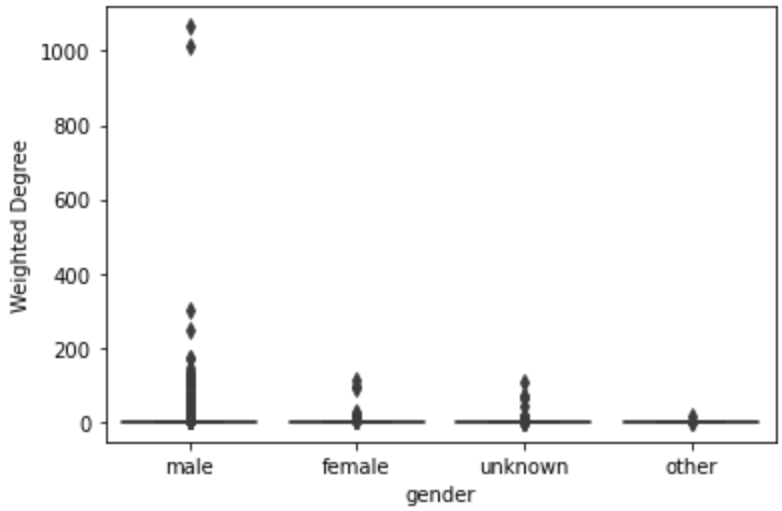
	count	mean	std	min	25%	50%	75%	max
gender								
female	185.0	3.243243	13.232875	0.0	0.0	0.0	1.0	113.0
male	1086.0	7.345304	49.265120	0.0	0.0	0.0	1.0	1066.0

other	39.0	0.512821	2.722922	0.0	0.0	0.0	0.0	17.0
unknown	170.0	2.758824	12.734223	0.0	0.0	0.0	0.0	110.0

Boxplots illustraring the distribution of the weighted degree by gender groups

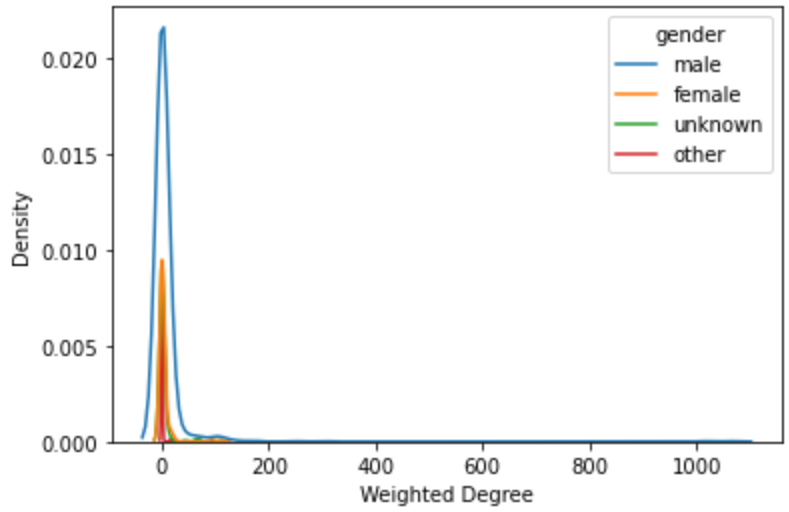
In [42]:
```
plt.ticklabel_format(style='plain', axis='x')
seaborn.boxplot(x='gender', y='Weighted Degree',
            data=node_degree)
```

Out[42]:
```
<AxesSubplot:xlabel='gender', ylabel='Weighted Degree'>
```



Density plot for the weighted degree

In [45]:
```
seaborn.kdeplot(x='Weighted Degree',data=node_degree,hue='gender')
plt.show()
```



These results are presented based on the Wikidata academic sample.

In [ ]: