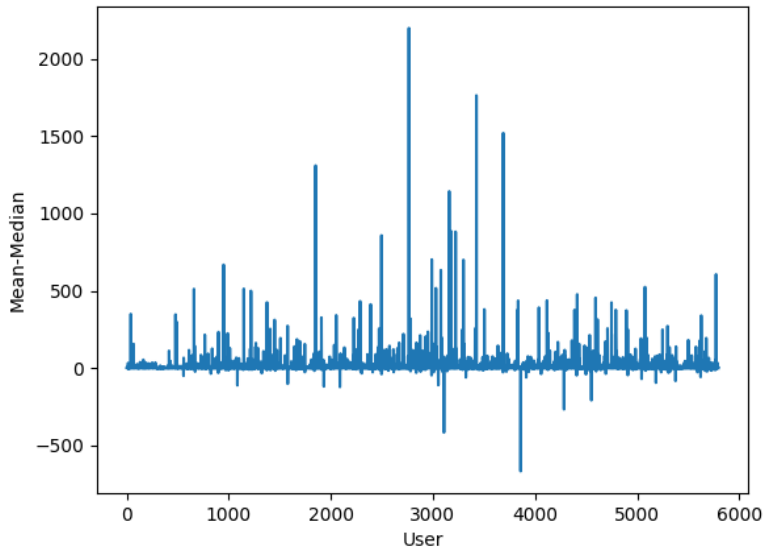
Question Block A: Data cleaning & merging

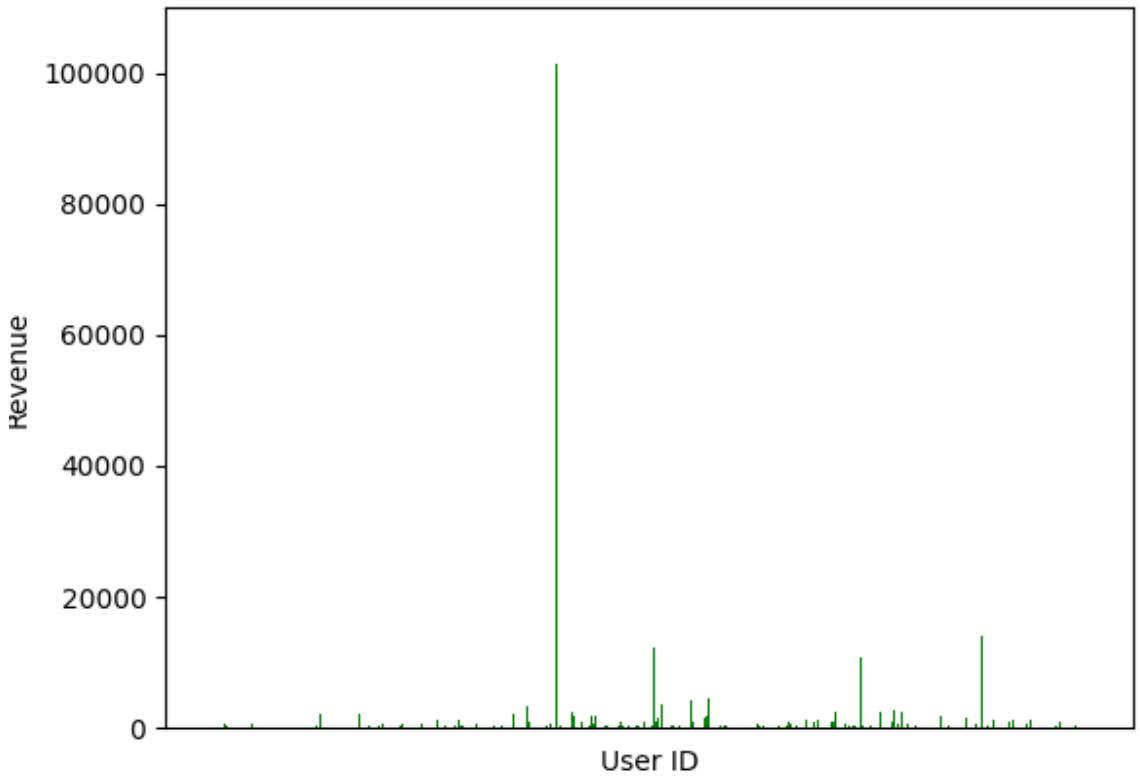
1. Please load the users and activity tables into Python.
    a. How many variables are in the datasets?
    Number of columns (variables) - users: 5
    Number of columns (variables) - activities: 3
    b. How many observations are in the datasets?
    Number of rows (obs) - users: 5807
    Number of rows (obs) - activities: 22642
2. How many
    a. Male users are in the dataset?
    Male users: 2909
    b. Female users are in the dataset?
    Female users: 1417
    c. For how many users is no gender information available?
    No gender specified: 4390

    If there is no gender information available, please assume that the gender is male ("M") for all further questions.
3. Merge the two datasets and calculate the days since registration for each activity. What is the total mean and median revenue generated per user (ie across all activities)? How do you interpret the result?



The difference between the mean and the median of a distribution is indicative of how simetric the distribution of revenue per person is. A small difference points to a symmetric distribution while a large difference indicates a skewed distribution with unusual outliers. The plot shows the mean-median difference in revenue for each user and while the majority of the users have a symmetric distribution of revenue, there are whildly skewed distributions both to the left and to the right.

4. Please visualise the distribution of the revenues by user.



5. What is the average *week 1 revenue*, ie the revenue generated by the user in their first week (ie within the first 7 days since registration)?
   Average one week revenue: 2153.6
   The average one week revenue per person is calculated and worked with in the attached python script.

Question Block B: Analytics

6. In this dataset men generated a higher week 1 revenue than women, on average. Is this difference in revenue between men and women statistically significant? What is an appropriate statistical test to determine this and what is it's p-value?
   Average male week rev: 107.18
   Average female week rev: 86.21

   T-tests are used to determine if there is a significant difference between the means of two groups. The null hypothesis H0 is:difference in revenue between men and women is not statistically significant.

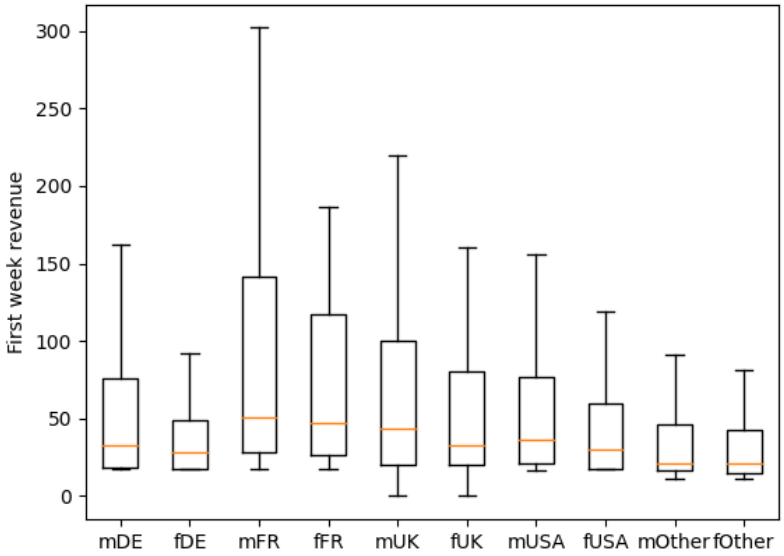
Ttest_indResult(statistic=2.118320598919536, pvalue=0.034252584013275585)

The p value being smaller than 0.05, the null hypothesis can be rejected meaning that there is a significant statistical difference between the 1 week revenue of men and women.

7. In which country is this difference biggest?
   Ttest results/country:
   Male and female stats in DE:
   Ttest_indResult(statistic=2.0258827047245354, pvalue=0.043360475747787966)
   Male and female stats in FR:
   Ttest_indResult(statistic=0.9878142559465553, pvalue=0.32498844553732753)
   Male and female stats in UK:
   Ttest_indResult(statistic=0.9760635531587415, pvalue=0.3292567571729236)
   Male and female stats in USA:
   Ttest_indResult(statistic=1.2859625122755634, pvalue=0.1988550690261348)
   Male and female stats in other countries:
   Ttest_indResult(statistic=2.686249751008175, pvalue=0.007363128665432487)

   The biggest difference between male and female 1 week revenues is in other countries (due to the smallest p value out of all).

8. Please visualise the relationship between country, gender and week 1 revenue with an appropriate chart.

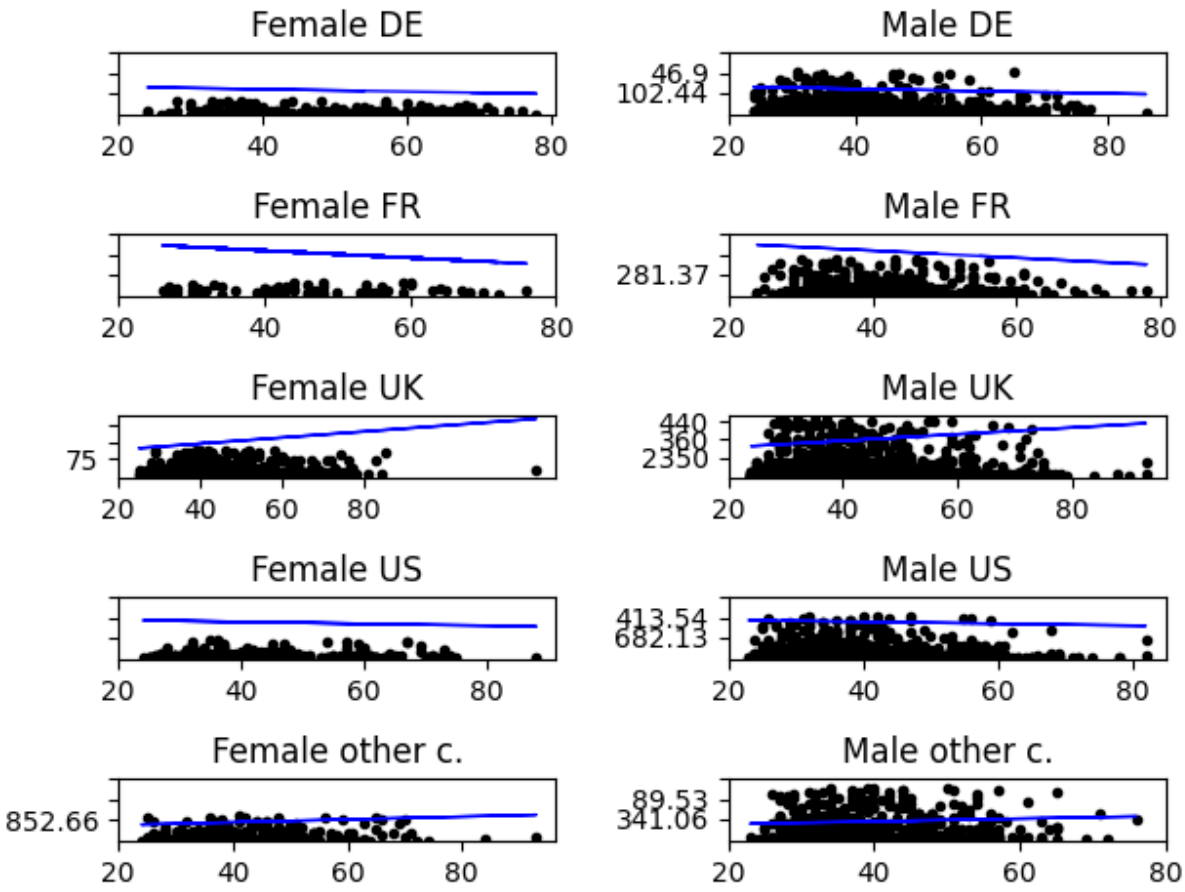


All the distributions are right skewed (towards the high revenue value end) and the male one week revenue is stronger skewed than the female one week revenue distribution, regardless of the country. The most positively skewed are the French 1 week revenue distributions.

9. Build a linear regression model to predict the week 1 revenue of a user based on the following variables: Gender, Age, Country and *day 1 revenue* (ie the revenue generated by the user on their registration date). Which of these variables have a statistically significant predictive power according to this model?

   The linear regressions were built in the scripts attached. The R-square values per gender and country:
   R_sq female DE: 0.0012258275698919796
   R_sq female FR: 0.004564760428514725
   R_sq female GB: 0.0007582034693079942
   R_sq female USA: 0.00015488862750978605
   R_sq female DE: 0.002353687732095655
   R_sq male DE: -0.01944888483415963
   R_sq male FR: -0.011811744998453833
   R_sq male GB: -0.0034173501996390687
   R_sq male USA: -0.003921601506234218
   R_sq male DE: -0.015102926560480734

   The negative values (for male 1 day revenues in every country) suggest that the linear regression is not a good choice for fitting the distributions. The female 1 day R-square values are positive but very low, suggesting a low correlation.



According to the plots, most female correlations appear to be outside the data points. I do not understand this trend myself.

10. What revenues do you expect women from France, Germany and the UK to generate, assuming they are all aged 40 and all generated £20 on their registration day?
    For this step to be possible, one has to establish the correlation between the revenue distributions and the registration day revenue.