Instructions

Go to

www.menti.com

Enter the code

4576 3493

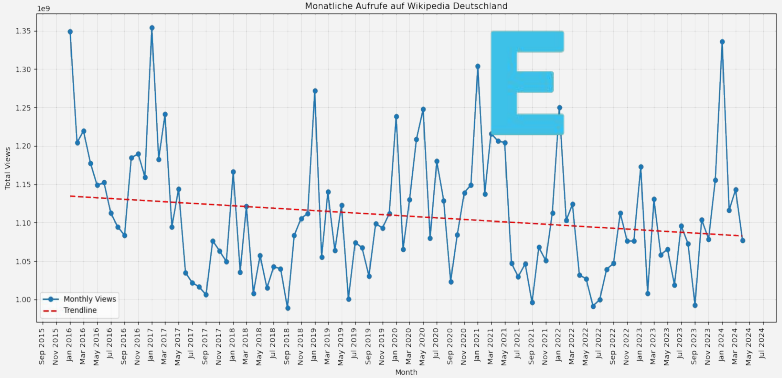
Or use QR code

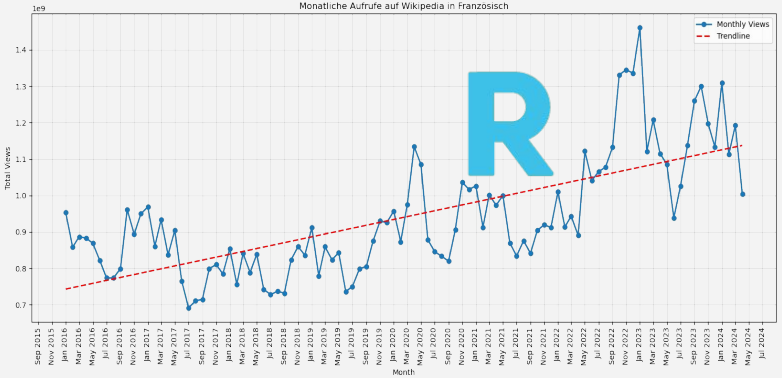
Analysis: Wikipedia Germany

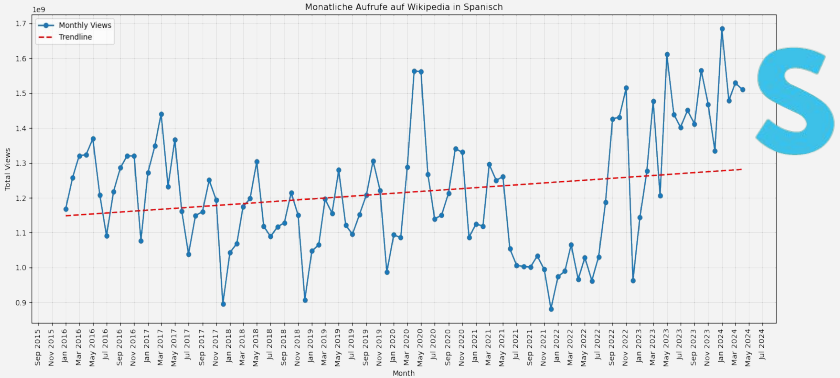
Ironhack Week 4 on SQL

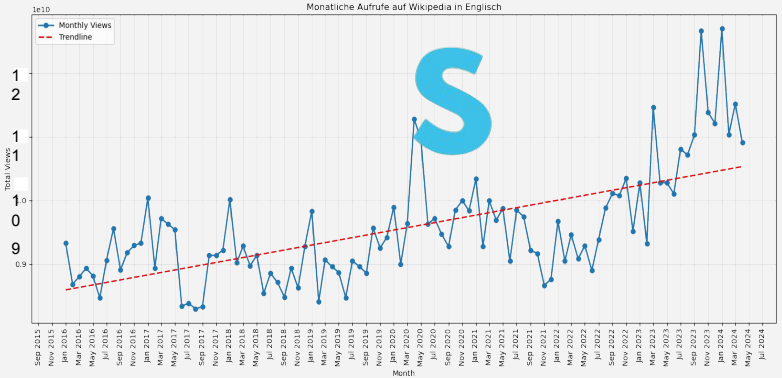
Jayashree, Daniel

What is going on in the German Wikipedia Ecosystem?









Project Scope

Background:

Wikipedia is one of the world's most visited websites and also the source for factual information for billions of people all over the world. While Wikipedia exists since twenty years, the user base is fluctuating constantly. Although one could think that Wikipedia is growing, which is the case for many of the world's most spoken languages such as English, Spanish or French, the German language Wikipedia is actually shrinking throughout the past years. We want to investigate reasons for this anomaly.

Problem Statement:

Views on Wikipedia in German are shrinking while other countries are growing. What's causing the effect?

Hypothesis:

#1 Prevalence of editors by locale
#2 Prevalences in editor behaviour by locale
#3 View on total pages per locale
#4 More edits = More views
#5 Mobile vs. Desktop views by locale

Objective:

What we intended to do: Support Wikimedia Germany Foundation e.V. finding the cause for shrinking user numbers.
What we were able to do: Play with Wiki data and test simpler hypothesis.

Resources:

2 DAs (Jayashree, Daniel)
Jupyter Notebook, MySQLWorkbench, Wikistats

Deliverables:

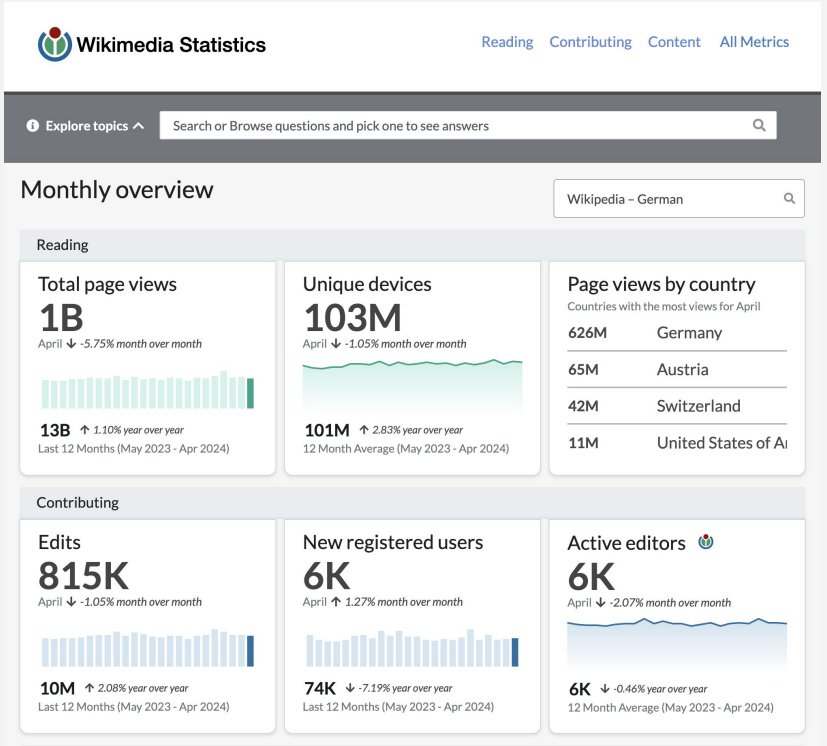
Overview, Data Wrangling Process, Hypothesis Check, Data in raw & cleaned CSVs, manipulated with both Python & SQL

Timeline:

4 Business Days

Data Wrangling Process

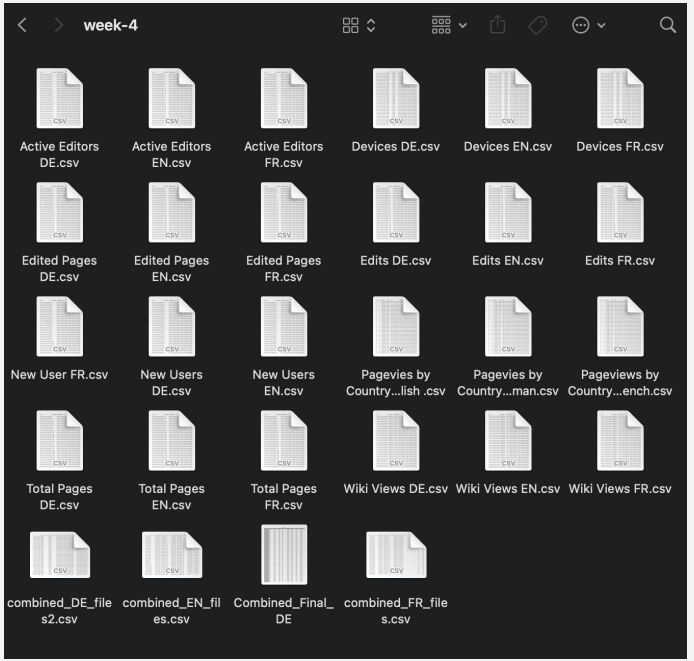
1. Selecting Dataset: Wikimedia Data is publicly available, we selected three languages for our comparison and built five hypothesis to check why the German Wikipedia seems to shrink vs. others.



Data Wrangling Process

1. Selecting Dataset: Wikimedia Data is publicly available, we selected three languages for our comparison and built five hypothesis to check why the German Wikipedia seems to shrink vs. others.

2. Transferring data via Python: First we tried to webscrape through Beautiful Soup, but we failed. Then we tried it via an API, but that was too complicated. Ultimately we downloaded the data by hand into 28 files and used Python to clean and combine them into three files.

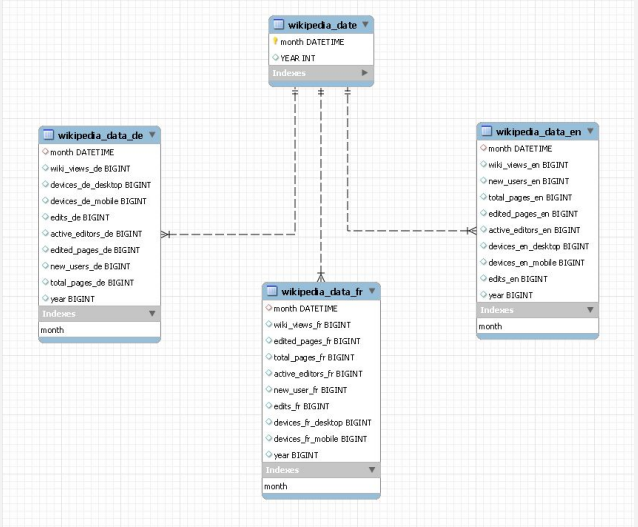


Data Wrangling Process

1. **Selecting Dataset:** Wikimedia Data is publicly available, we selected three languages for our comparison and built five hypothesis to check why the German Wikipedia seems to shrink vs. others.
2. **Transferring data via Python:** First we tried to webscrape through Beautiful Soup, but we failed. Then we tried it via an API, but that was too complicated. Ultimately we downloaded the data by hand into 28 files and used Python to clean and combine them into three files.
3. **Manipulating via SQL:** Created Schema, Datasets & E-R diagram according to .csv files and then started manipulating.

How to create a E-R diagram:

1. Imported csv files from Python into SQL
2. Created new table for primary & foreign keys
3. Click '`Database` -> `Reverse Engineer`' to get the E-R diagram



Data Wrangling Process

1. Selecting Dataset: Wikimedia Data is publicly available, we selected three languages for our comparison and built five hypothesis to check why the German Wikipedia seems to shrink vs. others.
2. Transferring data via Python: First we tried to webscrape through Beautiful Soup, but we failed. Then we tried it via an API, but that was too complicated. Ultimately we downloaded the data by hand into 28 files and used Python to clean and combine them into three files.
3. Manipulating via SQL: Created Schema, Datasets & E-R diagram according to .csv files and then started manipulating.
4. Preparing visualisations via Python: Used Seaborn & Matplotlib to bring data back into visualisations

Executive Summary

1. **Tl;Dr:** We didn't get closer to an answer why the German speaking ecosystem is slowly getting lesser views by month.
2. **Our Challenge:** Initially we wanted to webscrape much more data, but failed - through this "*knowledge bottleneck*" we needed to reduce this analysis to a minimum. **Though we were able to learn a lot by using Python, then SQL, then Python again**.
3. **General Checks:** We tested five simple hypothesis on the data we pulled that all resulted to be correct which we will present in the following slides.
4. **Next Steps:** Next to a quant. part we generally suggest a qual. part of a more holistic analysis of the language ecosystems, as user interviews can reveal behavioural patterns that differ per locale

Results

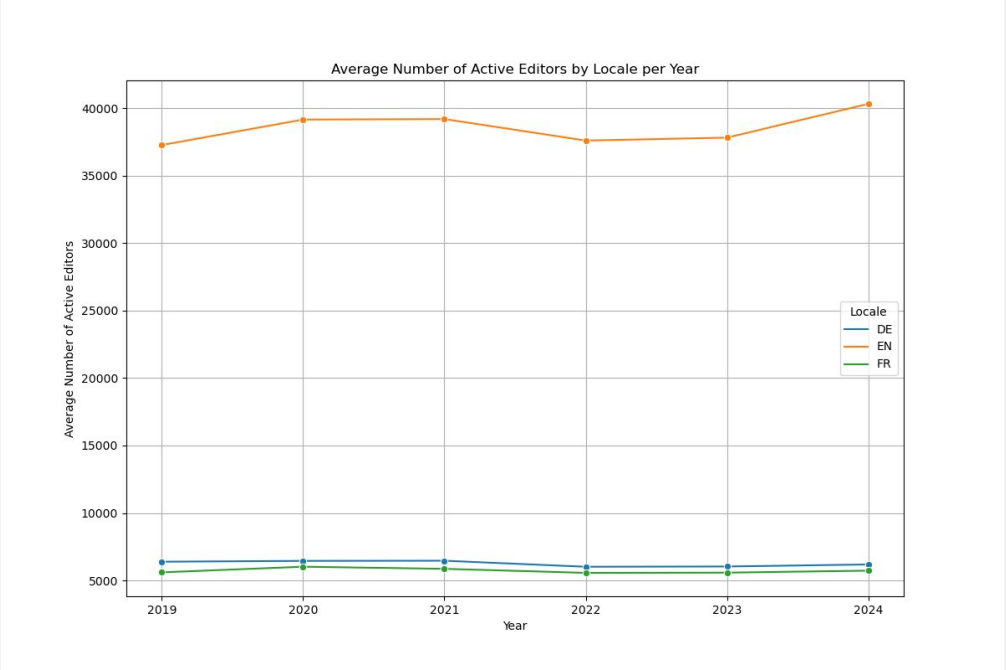
Hypothesis 1: As English is so prevalent on the internet, the language will attract far more active editors than other languages -> ***Correct***, *there are approximately 6 times more English editors than French or German*

Hypothesis 2: There are variances in edits per active editor per locale -> ***Correct***, *FR has slightly more edits/active editor than DE & EN, we assume as Wikipedia in French has less overall articles*

Hypothesis 3: The number of total pages on Wikipedia is steadily increasing over the years -> *Correct*, *obvious but good to know they are not decreasing over time…*

Hypothesis 4: More number of edits equate to more number of average page views. -> ***Correct***, *when compared, DE has 6 times less average edits than EN and 8 times less average page views than EN.*

Hypothesis 5: Improving the experience for Wikipedia for Mobiles is more important than Desktop -> *Correct,* *although the share for Mobile views is only slightly higher than for Desktop*



Results

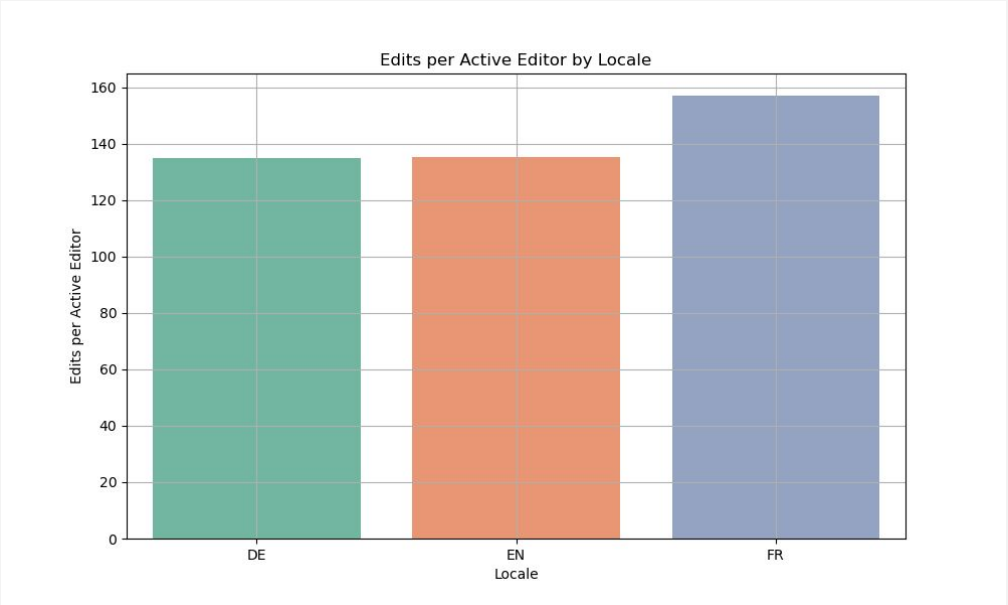
Hypothesis 1: As English is so prevalent on the internet, the language will attract far more active editors than other languages -> ***Correct***, *there are approximately 6 times more English editors than French or German*

Hypothesis 2: There are variances in edits per active editor per locale -> ***Correct***, *FR has slightly more edits/active editor than DE & EN, we assume as Wikipedia in French has less overall articles*

Hypothesis 3: The number of total pages on Wikipedia is steadily increasing over the years -> *Correct*, *obvious but good to know they are not decreasing over time…*

Hypothesis 4: More number of edits equate to more number of average page views. -> ***Correct***, *when compared, DE has 6 times less average edits than EN and 8 times less average page views than EN.*

Hypothesis 5: Improving the experience for Wikipedia for Mobiles is more important than Desktop -> *Correct,* *although the share for Mobile views is only slightly higher than for Desktop*



Results

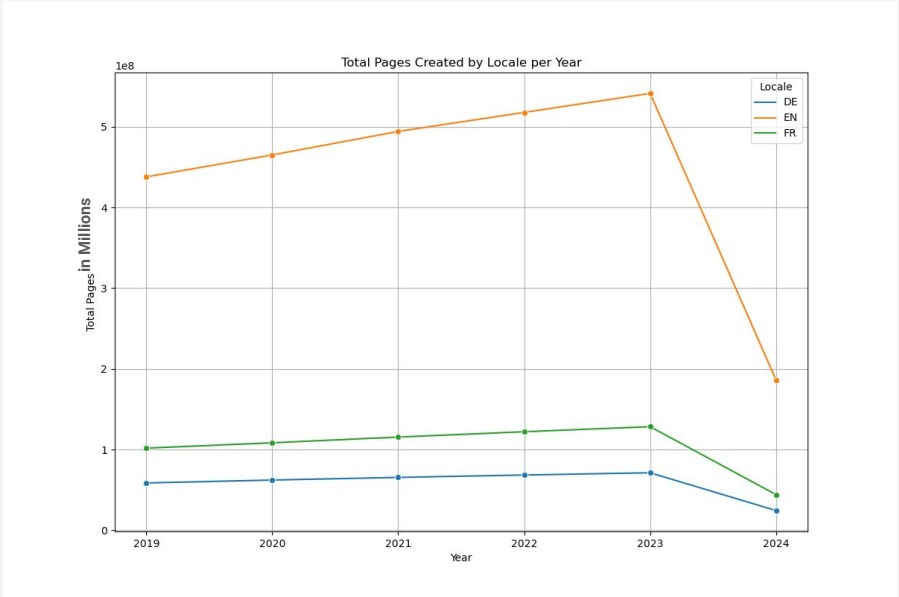
Hypothesis 1: As English is so prevalent on the internet, the language will attract far more active editors than other languages -> ***Correct***, *there are approximately 6 times more English editors than French or German*

Hypothesis 2: There are variances in edits per active editor per locale -> ***Correct***, *FR has slightly more edits/active editor than DE & EN, we assume as Wikipedia in French has less overall articles*

Hypothesis 3: The number of total pages on Wikipedia is steadily increasing over the years -> *Correct*, *obvious but good to know they are not decreasing over time…*

Hypothesis 4: More number of edits equate to more number of average page views. When compared, DE has 6 times less average edits than EN and 8 times less average page views than EN.

Hypothesis 5: Overall users prefer mobile devices to use wikipedia compared to Desktops.



Results

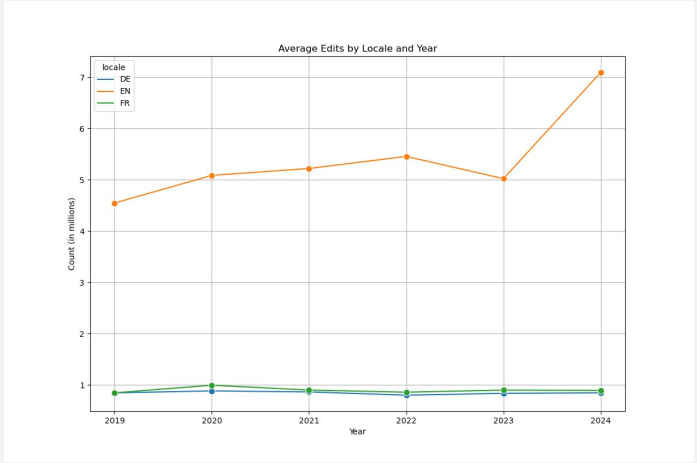
Hypothesis 1: As English is so prevalent on the internet, the language will attract far more active editors than other languages -> ***Correct***, *there are approximately 6 times more English editors than French or German*

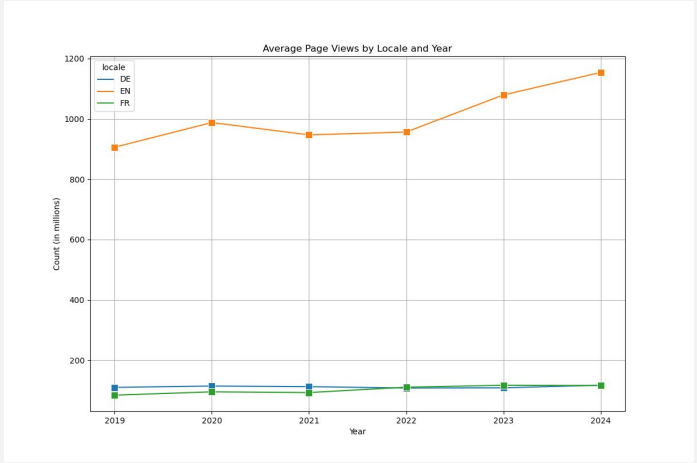
Hypothesis 2: There are variances in edits per active editor per locale -> ***Correct***, *FR has slightly more edits/active editor than DE & EN, we assume as Wikipedia in French has less overall articles*

Hypothesis 3: The number of total pages on Wikipedia is steadily increasing over the years -> *Correct*, *obvious but good to know they are not decreasing over time…*

Hypothesis 4: More number of edits equate to more number of average page views. -> ***Correct***, *when compared, DE has 6 times less average edits than EN and 8 times less average page views than EN.*

Hypothesis 5: Improving the experience for Wikipedia for Mobiles is more important than Desktop





Results

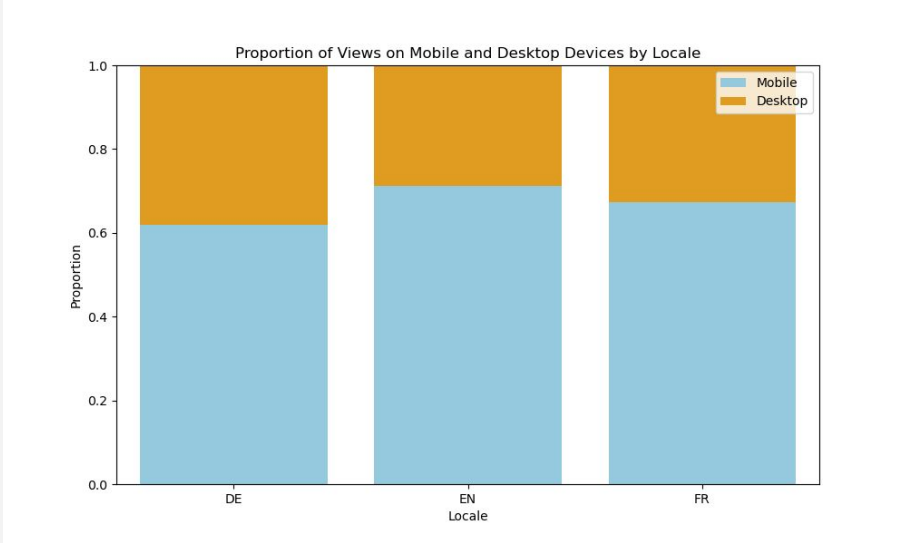
Hypothesis 1: As English is so prevalent on the internet, the language will attract far more active editors than other languages -> ***Correct***, *there are approximately 6 times more English editors than French or German*

Hypothesis 2: There are variances in edits per active editor per locale -> ***Correct***, *FR has slightly more edits/active editor than DE & EN, we assume as Wikipedia in French has less overall articles*

Hypothesis 3: The number of total pages on Wikipedia is steadily increasing over the years -> *Correct*, *obvious but good to know they are not decreasing over time…*

Hypothesis 4: More number of edits equate to more number of average page views. -> ***Correct***, *when compared, DE has 6 times less average edits than EN and 8 times less average page views than EN.*

Hypothesis 5: Improving the experience for Wikipedia for Mobiles is more important than Desktop -> *Correct,*
although the share for Mobile views is only slightly higher than for Desktop



Thanks!

Jayashree, Daniel