Hypothesis testing

Research Hypothesis 1: Users express more negative sentiment in questions related to lower-level programming languages compared to higher-level languages.

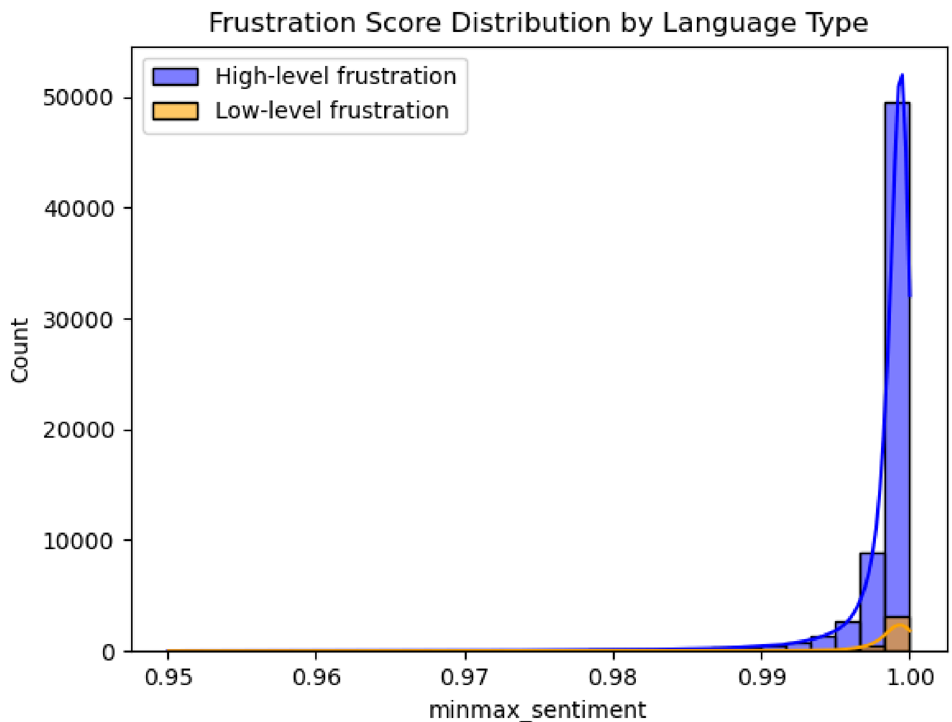
Null Hypothesis (H_0): There is **no significant difference** in user sentiment between questions related to low-level and high-level programming languages.

Alternative Hypothesis (H_1): User sentiment is significantly more negative in questions related to low-level programming languages than in high-level ones.

T-statistic: 0.60
P-value: 0.55

Because our p-value of 0.55 is well above the maximum p=0.05 required to reject the null hypothesis, we fail to reject the null hypothesis. There is a good chance that there is no significant difference in user sentiment between questions related to low-level and high-level programming languages.

One of the main weaknesses of this approach is the distribution of sentiments returned by the sentiment analysis. The distribution, whose domain is [0,1], is highly concentrated toward the right side. This is likely due to the fact that the model was trained using a general dataset that contained sentences extracted from movie reviews rather than Stack Overflow or other technically oriented data. Since nearly all StackOverflow posts are questions reflecting a lack of knowledge and thus some base degree of frustration in the user, it makes sense that a model trained on a dataset without such characteristics would do a poor job of representing relative frustration for the StackOverflow data. A priority for the final deliverable is exploring ways to address this challenge.



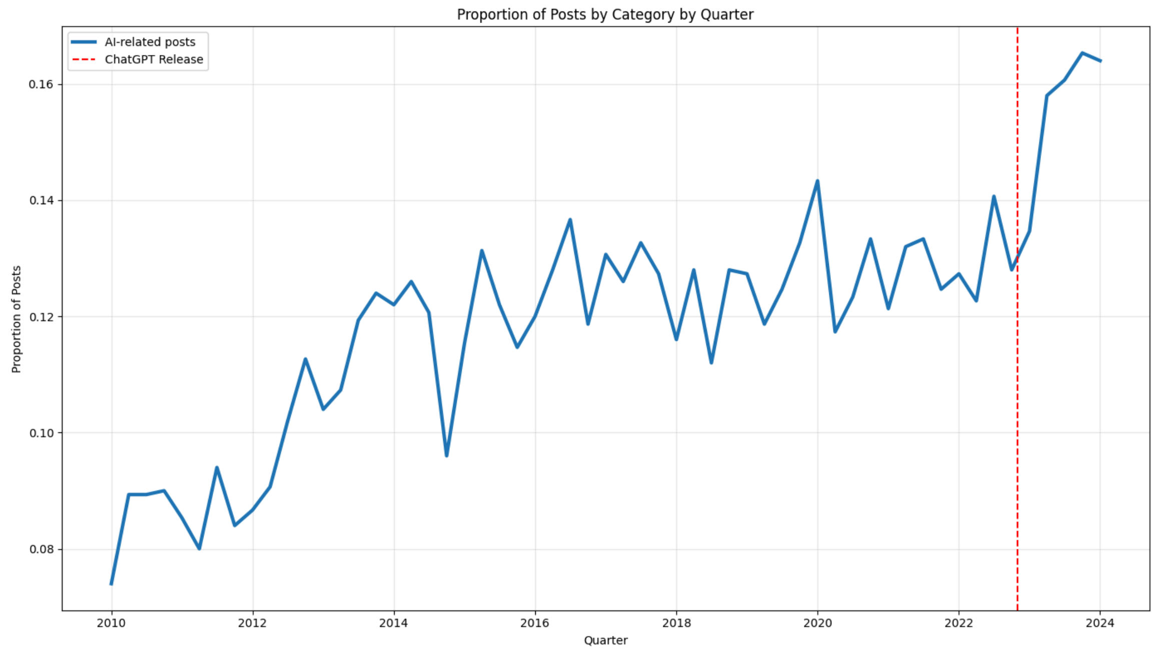
Research Hypothesis 2: The popularity of certain programming languages, especially Python, has increased over time in correlation with the rise of AI and machine learning models.

Null Hypothesis (H_0): There is **no significant change** in AI-related question volume on StackOverflow since the release of ChatGPT

Alternative Hypothesis (H_1): There is **a significant increase** in AI-related question volume on StackOverflow since the release of ChatGPT

```
Contingency Table:
is_ai_content  False  True
post_chatgpt
False           7475   1525
True            6893   1607

Pre-ChatGPT (2021-05-01 - 2022-11-01) vs Post-ChatGPT (2022-11-01 - 2024-05-01)
Analysis:
Pre-ChatGPT posts: 9000
Post-ChatGPT posts: 8500
Pre-ChatGPT AI proportion: 0.1694 (1525 AI posts)
Post-ChatGPT AI proportion: 0.1891 (1607 AI posts)
Absolute difference: 0.0196
```

```
Relative change: 11.58%

Chi-square Test:
Chi-square value: 11.3123
p-value: 0.00076996
Significant difference at α=0.05: True
```



We reject our null hypothesis that there is no significant difference between AI-related posts pre and post chatgpt. There is a statistically significant uptick in AI related post with a very low p-value. I am confident in this, however, I would like to further investigate the actual labelling of AI-related posts or not. Currently, we use a keyword search that just finds if specific words related to AI like (llm, chatgpt, etc) are in the title, body, tags. An actual classifier trained for this purpose would likely give us results we can be more confident in. This is something we might play around with in the coming days, however we would likely reduce our volume of data (currently we look at around 80k questions total) as the predictor would likely be more computationally expensive.

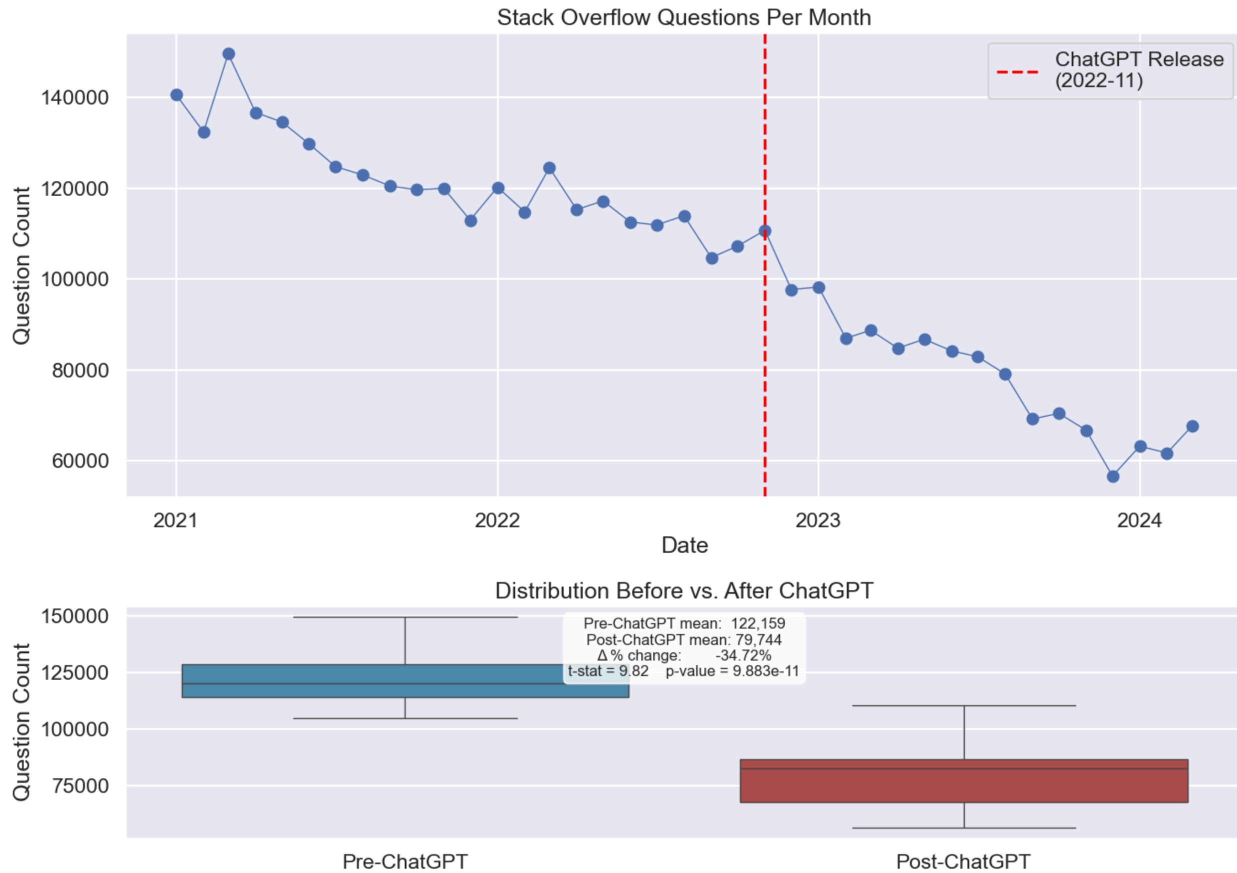
Research Hypothesis 3: Since the rise of AI tools (e.g., ChatGPT, Copilot), the overall number of Stack Overflow questions has decreased, and user sentiment has become less negative overall.

Null Hypothesis (H_0): There is **no significant change** in the number of questions or user sentiment on Stack Overflow since the introduction of AI tools.

Alternative Hypothesis (H_1): There has been a significant decrease in the number of questions and a significant increase in user sentiment positivity since the introduction of AI tools.

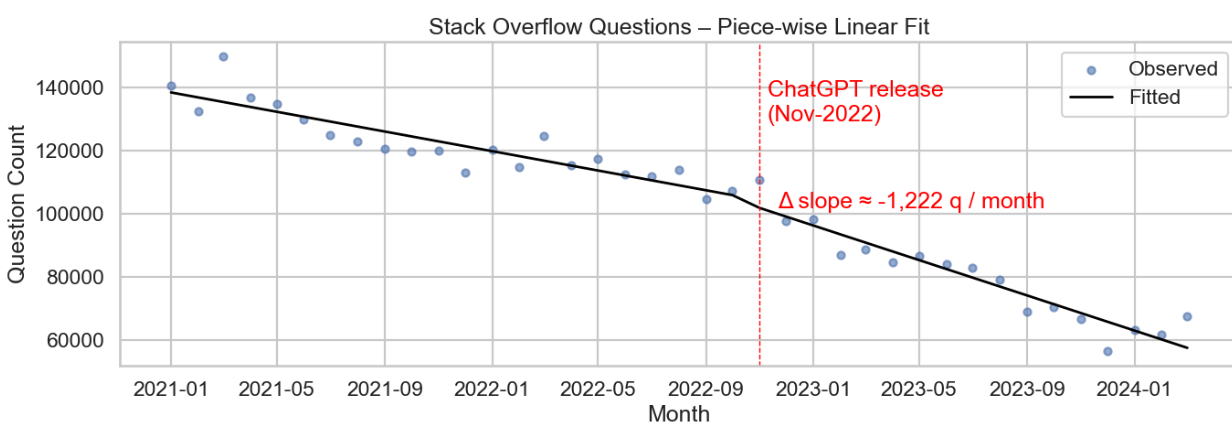
Db query:

```
USE [StackOverflow]
GO

SELECT
    FORMAT(CreationDate, 'yyyy-MM') AS PostMonth,
    COUNT(*) AS QuestionCount,
    CASE
        WHEN CreationDate < '2022-11-01' THEN 'Pre-ChatGPT'
        ELSE 'Post-ChatGPT'
    END AS Period
FROM dbo.Posts
WHERE PostTypeId = 1  -- Questions only
  AND CreationDate >= '2016-01-01'
  AND CreationDate < '2024-05-01'
GROUP BY FORMAT(CreationDate, 'yyyy-MM'),
         CASE
             WHEN CreationDate < '2022-11-01' THEN 'Pre-ChatGPT'
             ELSE 'Post-ChatGPT'
         END
ORDER BY PostMonth;
```




```
————————————————————————  t-test summary  ————————————————————————
Pre-ChatGPT  (n=22):  μ = 122,159
Post-ChatGPT (n=17): μ = 79,744
Percent change: -34.72%
t-statistic   : 9.82
p-value       : 9.883e-11
Statistically significant? YES
```

```
──────────────── Interrupted time-series regression ────────────────
Formula      : QuestionCount ~ t_mon + post + t_post
Breakpoint   : 2022-11
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   1.384e+05   2108.058     65.634      0.000    1.34e+05    1.43e+05
t_mon      -1548.0718    172.294     -8.985      0.000   -1897.847   -1198.297
post        2.437e+04   7981.361      3.053      0.004    8166.235    4.06e+04
t_post     -1221.8306    306.571     -3.985      0.000   -1844.203    -599.459
==============================================================================

Derived monthly slopes (questions / month)
  Pre-ChatGPT slope  :   -1548.1
  Post-ChatGPT slope :   -2769.9
  Δ slope (post-pre) :   -1221.8
───────────────────────────────────────────────────────────────────
```



Although stack overflow question counts were on the decline already, we can see a clear steeper decrease after the release of ChatGPT. This is something that I expected; I feel like ChatGPT is just so much more convenient than posting on stack overflow. We were surprised that Stackoverflow was already on a sharp decline, however it's clear that ChatGPT played some role in accelerating the decline based on the slope's above. I feel that using regression on pre and post GPT release was a strong choice here as we can clearly see the change in slope. We also ran a t-test on our before and after means and this further confirmed our hypothesis. One point that we were unsure about was what time frame to view this data in: should we go way further back than 2021, we would see an even greater change in slope post-chatgpt. We ended up deciding on 2021 to keep the amount of data before and after chatgpt's release fairly even.

##GENERAL TODOS

Questions to answer based on the results from ML algorithms:

Provide comments and an **interpretation of the results** you obtained:

1) Did you find the results corresponded with your **initial belief** in the data? If yes/no, **why** do you think this was the case?
   We suspected that the release of ChatGPT and rise in AI would correlate with a downturn in StackOverflow usage and our beliefs were confirmed. It was quite surprising that StackOverflow was already experiencing a pretty sharp decline in monthly posts, with ChatGPT's release we saw that decline accelerate. However, with regard to frustration scores, the degree of correspondence between our initial belief and our results was inconclusive. We believe this is because of the poor suitably of our sentiment model for this dataset, which we explain in greater depth above (Hypothesis 1).

2) Do you believe the **tools** for analysis that you chose were **appropriate**? If yes/no, **why or what method could have been used**?
   In hypotheses 2 and 3, we believe the tools we chose for analysis were appropriate for the most part. Regression analysis could be used to analyze the downward trend of StackOverflow posts. One situation that I did have some difficulty with was classifying data into whether they were AI-related. At the moment I use a standard keyword regex search of post bodies, titles and tags with a bunch of query AI words to determine if they are AI-related, however I think our confidence could be improved if we used a natural language classifier to check if a post is AI-related, although this would take up much more compute. This is something we could further explore before we complete our analysis.

   In Hypothesis 1, we do not believe the sentiment analyzer we used was appropriate for the StackOverflow data due to the differences between that data and its training set. Of all free sentiment analyzers we found online, this one seemed to correspond the best, but it still does not correspond well. Short of building our own training dataset, which is likely outside the scope of this project, we could look for another more suitable pretrained analyzer, or try to gauge sentiment with a simpler, logic-based model (e.g., analyzing the frequency of words commonly associated with frustration across posts for low- and high-level languages).

3) **Was the data adequate for your analysis**? If not, what **aspects of the data were problematic** and how could you have **remedied** that?
   One problem we encountered was how much we can trust post tags when it comes to classifying into programming languages as well as whether a post is

ai-related. We could further test and explore how accurate post-tags are by comparing their outputs to something we may trust more like a fine-tuned LLM on a small set of samples. In particular, the data available for the first part of the analysis was fairly inadequate: we had no preexisting sentiment labels, which required the use of a deep-learning-based sentiment analysis, which itself was problematic because it was trained on a different dataset. Training it on a labeled set of StackOverflow posts would solve this problem, but then, it would also solve the main problem of not having sentiments pre-associated with the posts.