Personal Data Analysis Project: My Pinterest Usage

Dilara Alkanalka 30758

DSA210

The Goal: To analyze my Pinterest usage to uncover patterns of activity and interest with my personal app data.

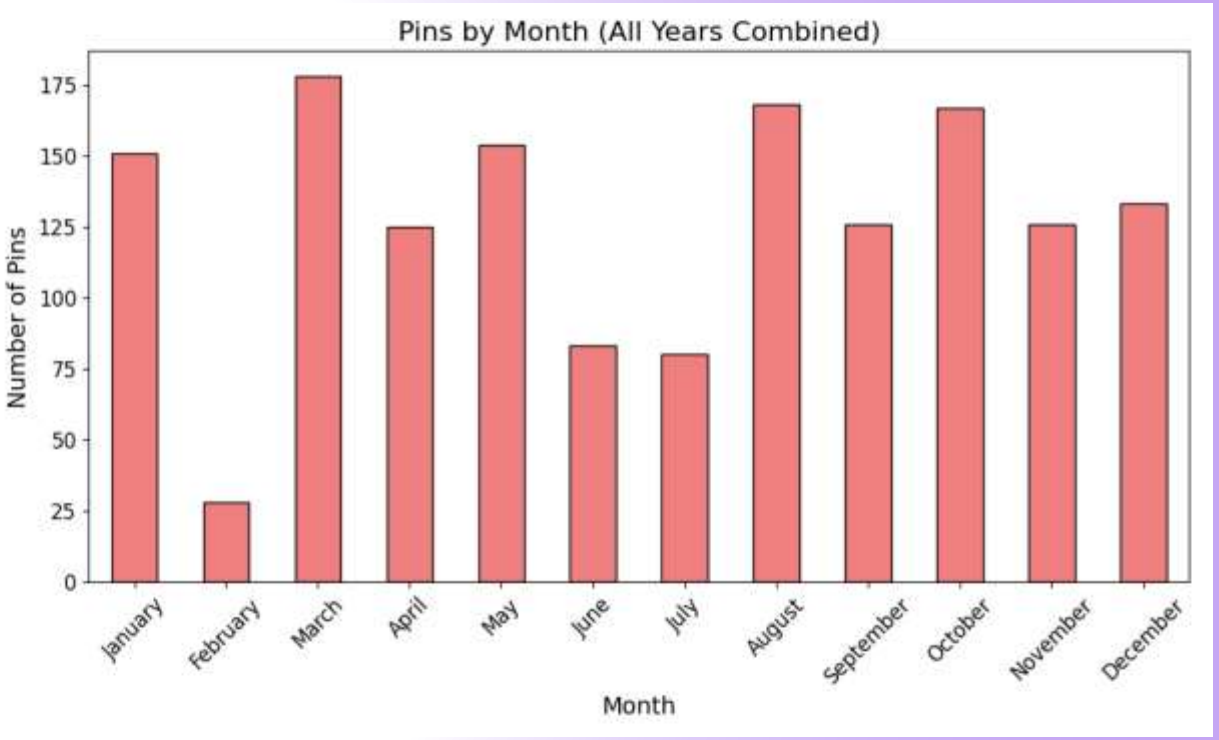
My Hypothesis: I am more active on the app on summer months or at times that I am relatively more free compared to my school months because I use the app mainly for inspirations over artistic products and for my hobbies, so I would use the app for my hobbies more on my leisure time.

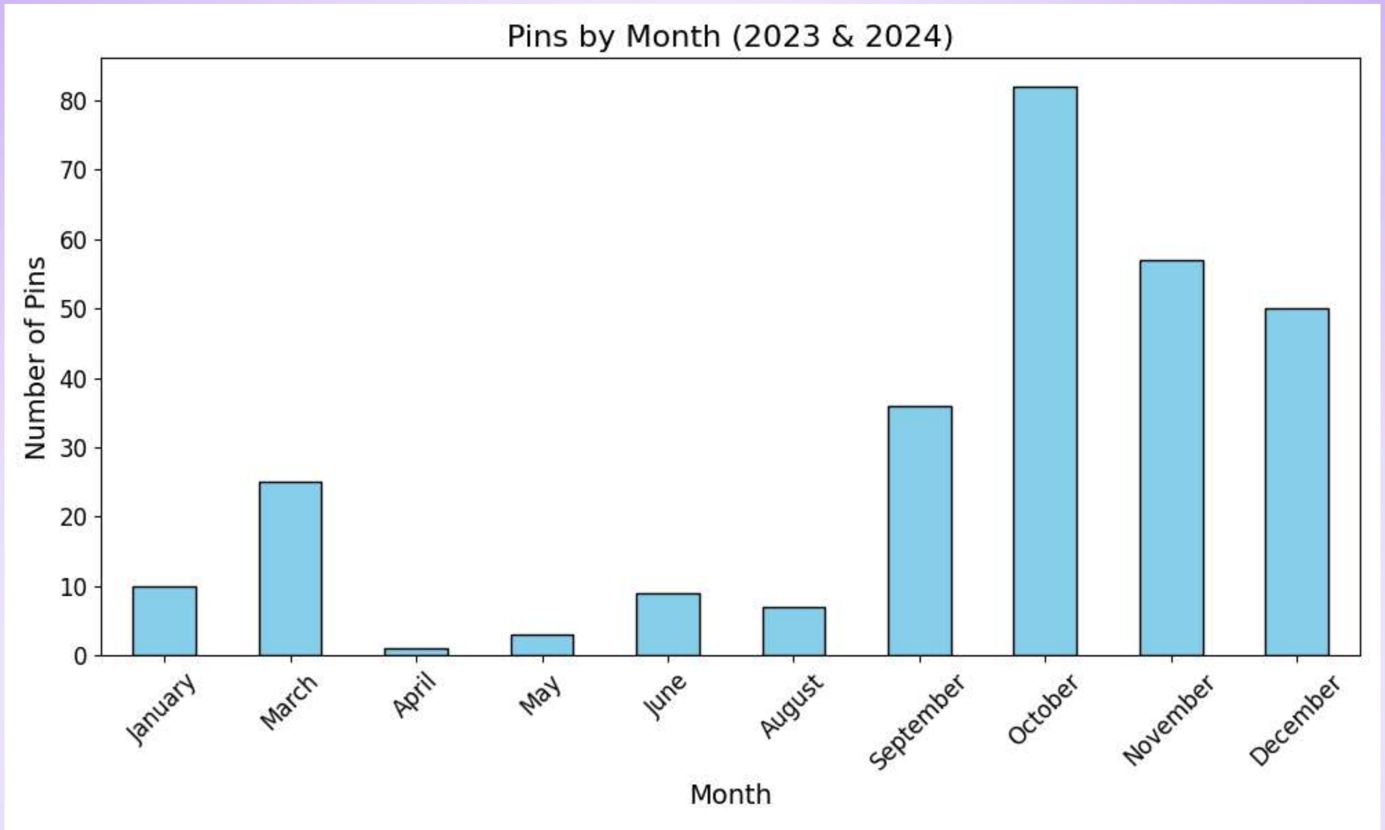
Collecting the Data

- Data was collected through Pinterest’s own export tool as HTML file
- I had insights about boards, pins, category and relative timestamps with this data and exported boards’ and pins’ data as csv while hiding some sensitive attributes.
- Some unnecessary rows such as missing values and deleted attributes were cleaned from the data
- I used tools such as Python, Pandas, Matplotlib, Seaborn and Jupyter Notebook for analysis

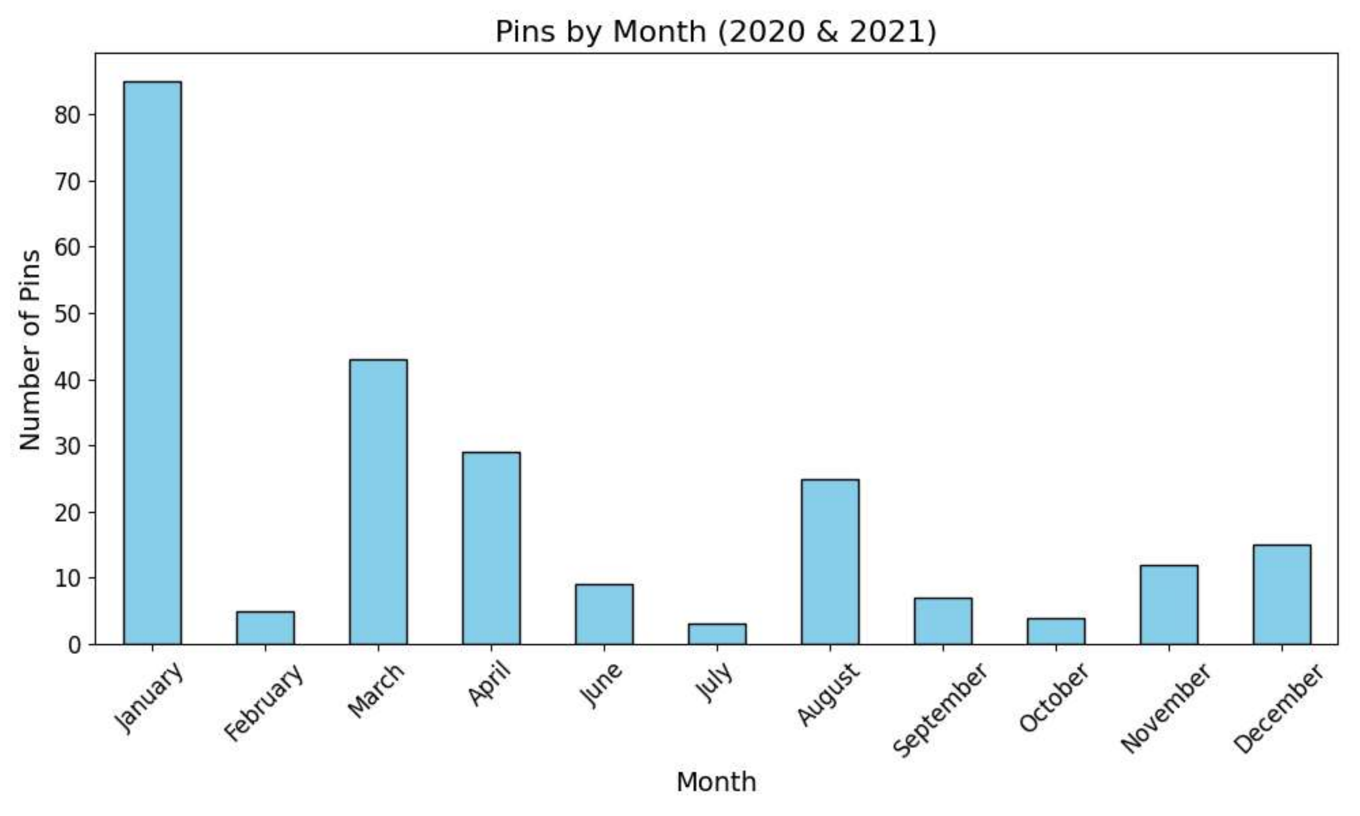
Exploratory Data Analysis

- I analyzed the pinning activity based on different months and years
- Visualized the values withbar charts
- I chose three different datasets for the plot by filtering out the pins by creation (pinning) time to analyize how my behaviour may have changed.
- On the right you can see the graph for the whole data (all years combined)





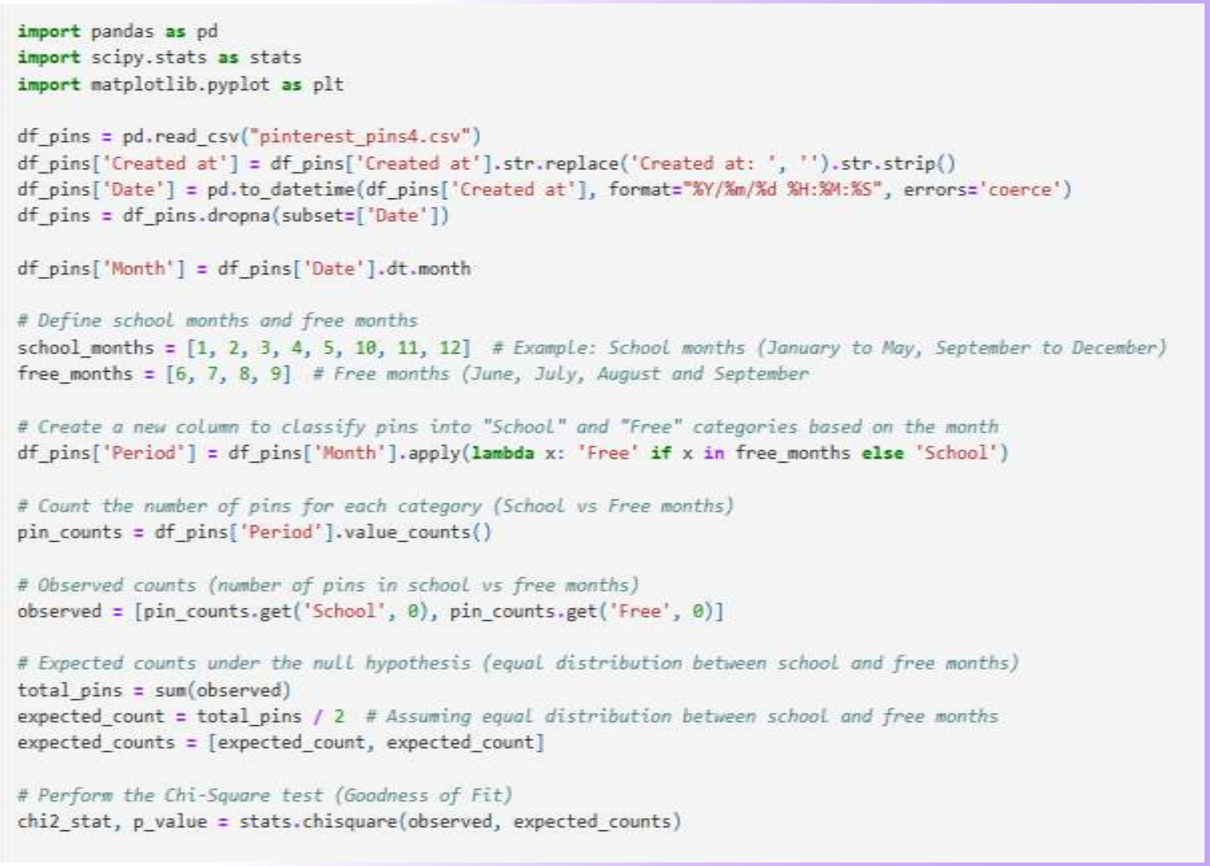
- With the bar chart we can see that I was more active on between the months September and December for years 2023-2024
- The initial observation does not align with the hypothesis as I have used the app more during school month spesifically in mid-term season and I was least active on the summer break



- I decided to check the data from earlier years (2020 and 2021) which I was going to high school to see if my school routine I had then had more impact on my app usage activity.
- Yet I was more active on months January and March which are times again I had exams and less active on June and July so we again can not support the hypothesis.

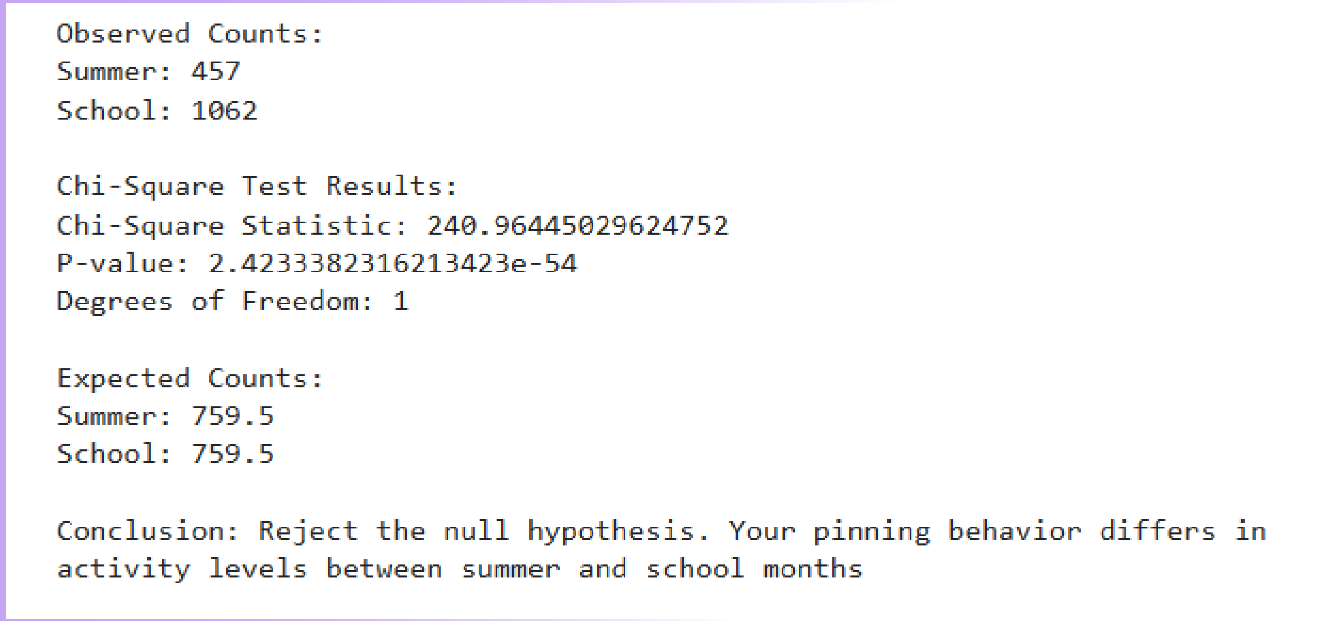
Chi-Square Goodness of Fit Test

- I applied a Chi-Square test to analyze the difference between pinning activity over months better
- On the right you can see the relevant code snippet for this step

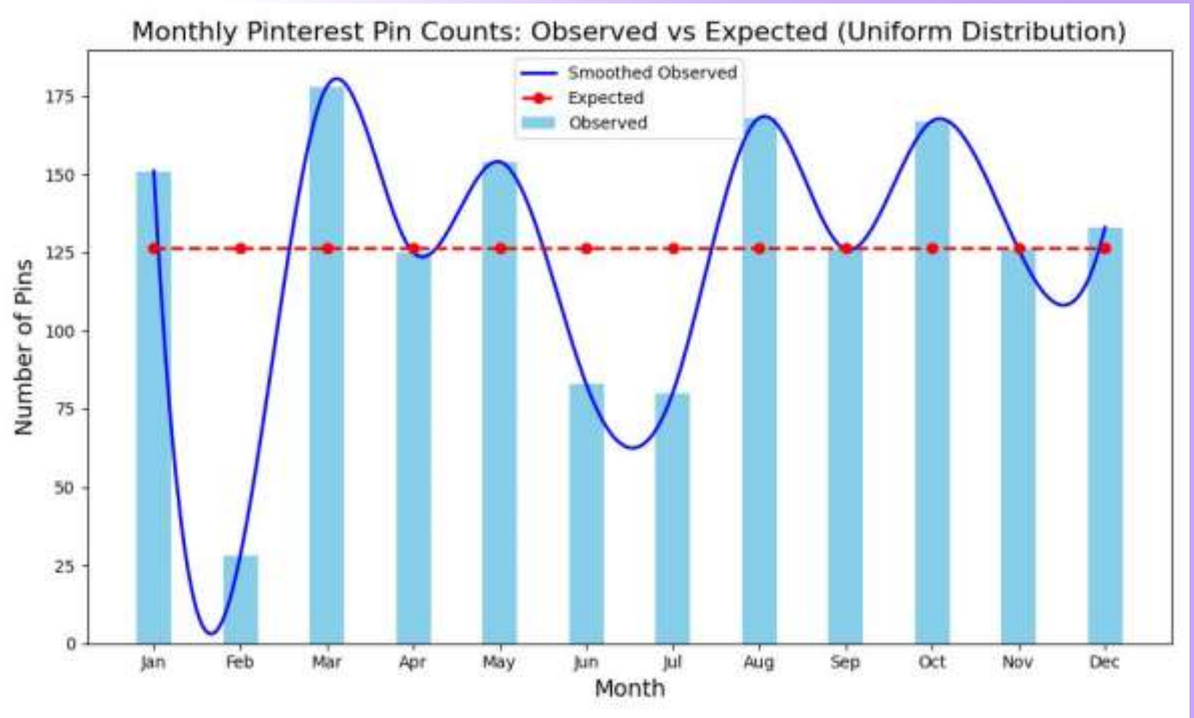


```python
import pandas as pd
import scipy.stats as stats
import matplotlib.pyplot as plt

df_pins = pd.read_csv("pinterest_pins4.csv")
df_pins['Created at'] = df_pins['Created at'].str.replace('Created at: ', '').str.strip()
df_pins['Date'] = pd.to_datetime(df_pins['Created at'], format="%Y/%m/%d %H:%M:%S", errors='coerce')
df_pins = df_pins.dropna(subset=['Date'])

df_pins['Month'] = df_pins['Date'].dt.month

# Define school months and free months
school_months = [1, 2, 3, 4, 5, 10, 11, 12]  # Example: School months (January to May, September to December)
free_months = [6, 7, 8, 9]  # Free months (June, July, August and September

# Create a new column to classify pins into "School" and "Free" categories based on the month
df_pins['Period'] = df_pins['Month'].apply(lambda x: 'Free' if x in free_months else 'School')

# Count the number of pins for each category (School vs Free months)
pin_counts = df_pins['Period'].value_counts()

# Observed counts (number of pins in school vs free months)
observed = [pin_counts.get('School', 0), pin_counts.get('Free', 0)]

# Expected counts under the null hypothesis (equal distribution between school and free months)
total_pins = sum(observed)
expected_count = total_pins / 2  # Assuming equal distribution between school and free months
expected_counts = [expected_count, expected_count]

# Perform the Chi-Square test (Goodness of Fit)
chi2_stat, p_value = stats.chisquare(observed, expected_counts)
```

Findings of the Chi-Square Test



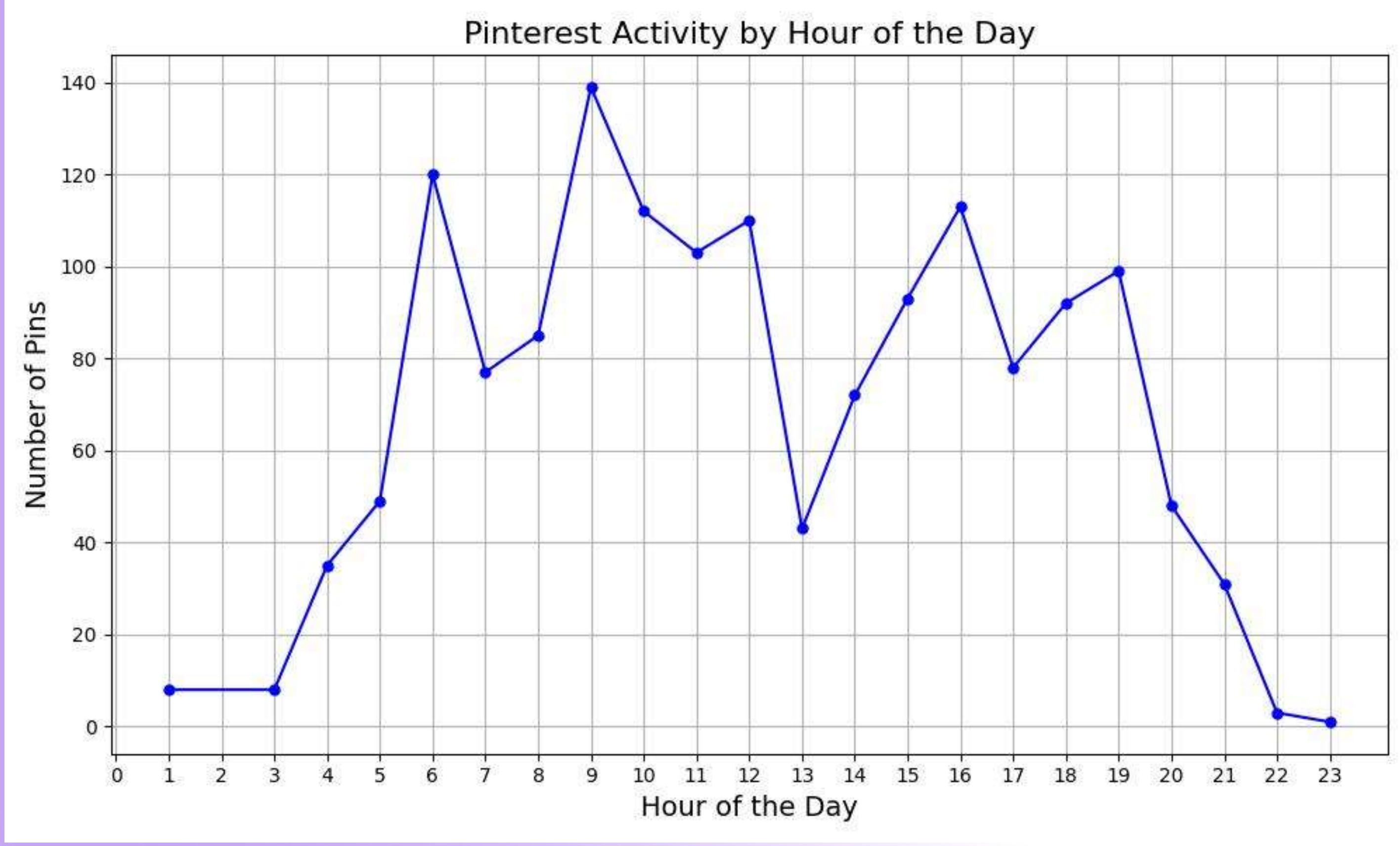
```
Observed Counts:
Summer: 457
School: 1062

Chi-Square Test Results:
Chi-Square Statistic: 240.96445029624752
P-value: 2.4233382316213423e-54
Degrees of Freedom: 1

Expected Counts:
Summer: 759.5
School: 759.5

Conclusion: Reject the null hypothesis. Your pinning behavior differs in
activity levels between summer and school months
```



Based on the output of the test, we reject the null hypothesis as the pinning activity differs between those periods. However not in the way I suggested my hypothesis in the first place, it appears there is a reverse correlation, since I was most active during school months and less on summer.

Here we can see the graph of expected pin count based on uniform distribution for the null hypothesis, and the actual distributionof pin savings over the months of all years in the data.

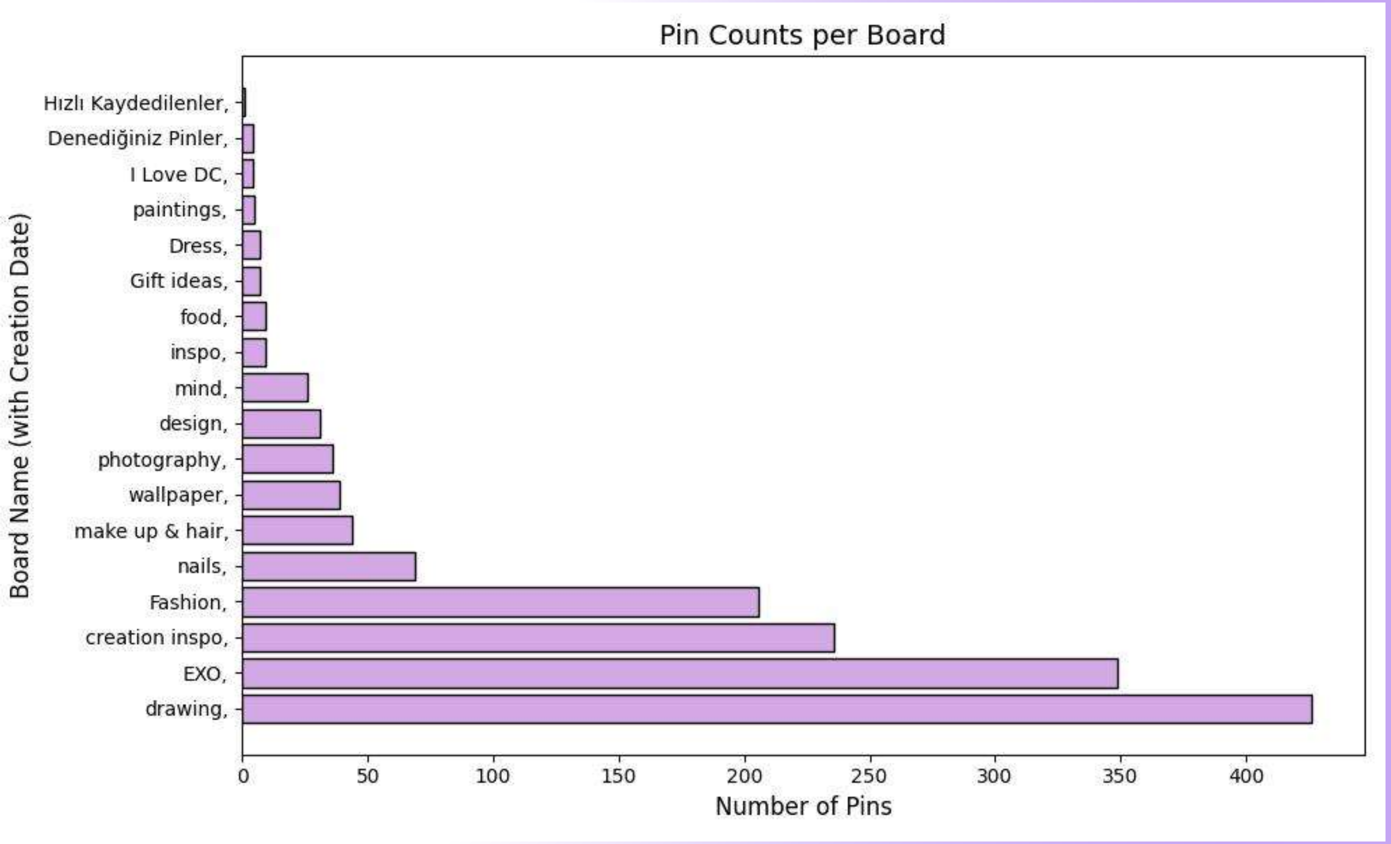
Pinning Activity by Hour of the Day



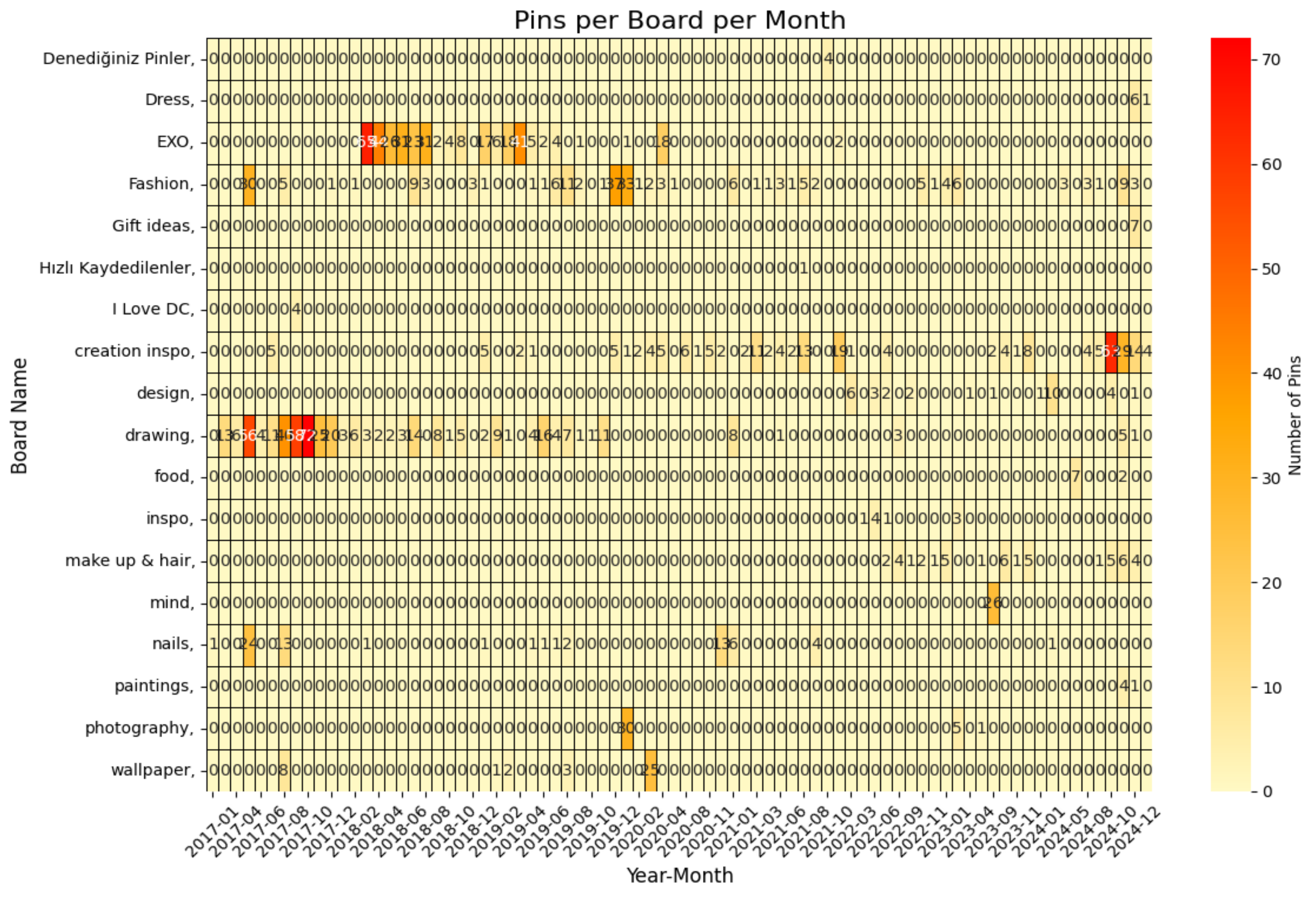
- From the analysis of hourly pinning activity, I observed that peeking active hours were surprisingly in the morning, particularly between 9 AM and 10 AM and also quite high on 6 AM (probably because I have a habit of waking up early).
- The usage between 4 PM and 7 PM appear to be relatively high as well.
- There is a significant decline around noon, specifically between 12 PM and 2 PM.

Analyzing Pinterest Pin Counts per Board

- I analyzed the number of pins associated with each Pinterest board to see which boards I use the most.
- Since I use Pinterest mainly for artistic inspiration the board 'Creation Inspo' which focuses on pottery and DIY projects, and the board 'Drawing' are some of the boards with most pins.
- The creation time of these boards also affect the pin count for instance board 'Drawing' is one of the oldest created boards with date of 2017/03/21.



Analyzing Pinterest Interactions by Board and Month with Heatmap



- I created a heatmap to visualize how I interacted with different Pinterest boards over time (year and month) by extracting the count of each pin and their saving date for each unique board name from the pins' csv data.
- This heatmap provides a view of how my pinning activity varies across boards over time.
- For instance, I used the "Drawing" board actively between June and August 2017, suggesting an increased interest in sketching and illustration.
- Similarly, the "Creation Inspo" board, with pottery projects, showed a peak in activity between August and October 2024 as I started going to a pottery class during that time.

Limitations & Future Work

- Many of the pins do not have a specified title or category, and they are labeled as "None" which can limit the ability to categorize or analyze them effectively for the types of content I interact with most frequently.

- What could be done in the future: Include more detailed analysis of user interactions like comments, followed boards of other users and repins.

Thank you for your time

You can see the details of the project and implementation steps on my repository.