Name-Surname: Zeynep Bahar Deniz
SU ID: 34194

DSA 210 Project Report

The Effect of Rising Social Media Usage on Shopping Behaviour

1.Motivation

Social media platforms have become one of the most important places to shape the modern shopping patterns. The constant exposure of non-ending trends encourages users to frequently engage in consuming activities. In day to day observations, increased screen time is highly associated with the higher consumer interests and higher purchasing urges.

While starting this project, I got my motivation from my social relations. What I observed till now was the people who engaged in with higher social media activities was exposed to these trends more resulting with an increase in consumption desires of these trends. Therefore, I wanted to not only decide with my intuitions and put it into a real test using real world datas. The goal is to analyze whether higher social media usage can be linked with consumer interests in online shopping. To start this project, most inspirational thing for me was the biggest trend of 2025 which is Labubu trend. Within this project, I aimed to analyze the effect of these trend exposures to customers. This trend not only restricted with Labubu toys but also about the other trends it triggered such as other similar toys sold under the same company. I found that the setter for this trend is a shop called PopMart. So, I combined screen time datas with search interest on the search keyword "PopMart" to analyze the shopping patterns.

2.Data Sources

I used three main datasets for this project.

a) Social Media Usage and Demographics Dataset
  - Name: Students Social Media Addiction.csv
  - Source: Kaggle

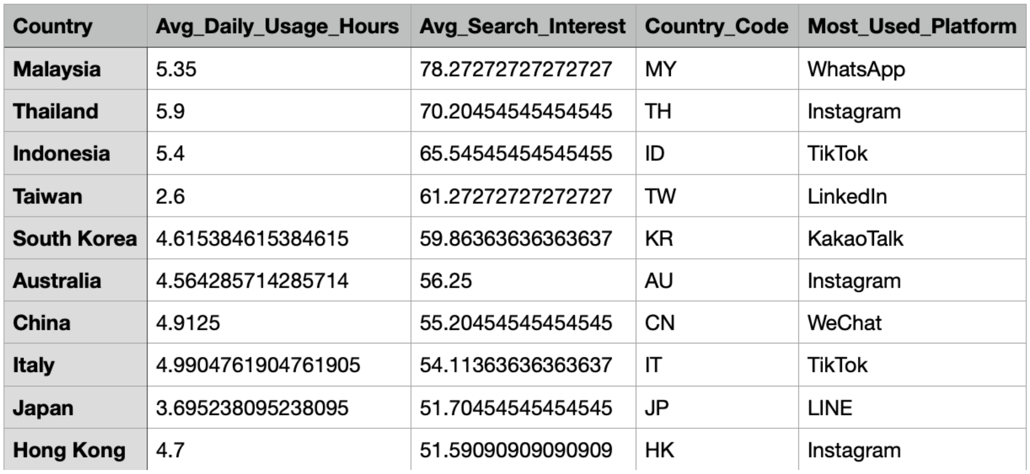
- It includes contents such as: Average daily social media usage, Country information, Age and gender data, Most frequently used social media platform
- I used it as my main source for screen time data also with different features logged for each data point.

b) Google Trends Search Interest Dataset

- Name: country_screen_vs_interest.csv
- Source: Google Trends (collected using pytrends)
- It includes contents such as: Country-level search interest scores
- I used it to act as an indicator of consumer interests in online shopping

c) Engineered Country Level Feature Dataset

- Name: top10_interest_with_platform.csv
- Source: I derived it using the two csv above and engineered the features I wanted then sorted by the average interests.
- Here is an example from the csv:



Country	Avg_Daily_Usage_Hours	Avg_Search_Interest	Country_Code	Most_Used_Platform
Malaysia	5.35	78.27272727272727	MY	WhatsApp
Thailand	5.9	70.20454545454545	TH	Instagram
Indonesia	5.4	65.54545454545455	ID	TikTok
Taiwan	2.6	61.27272727272727	TW	LinkedIn
South Korea	4.615384615384615	59.86363636363637	KR	KakaoTalk
Australia	4.564285714285714	56.25	AU	Instagram
China	4.9125	55.20454545454545	CN	WeChat
Italy	4.9904761904761905	54.11363636363637	IT	TikTok
Japan	3.695238095238095	51.70454545454545	JP	LINE
Hong Kong	4.7	51.59090909090909	HK	Instagram

3. Data Analysis

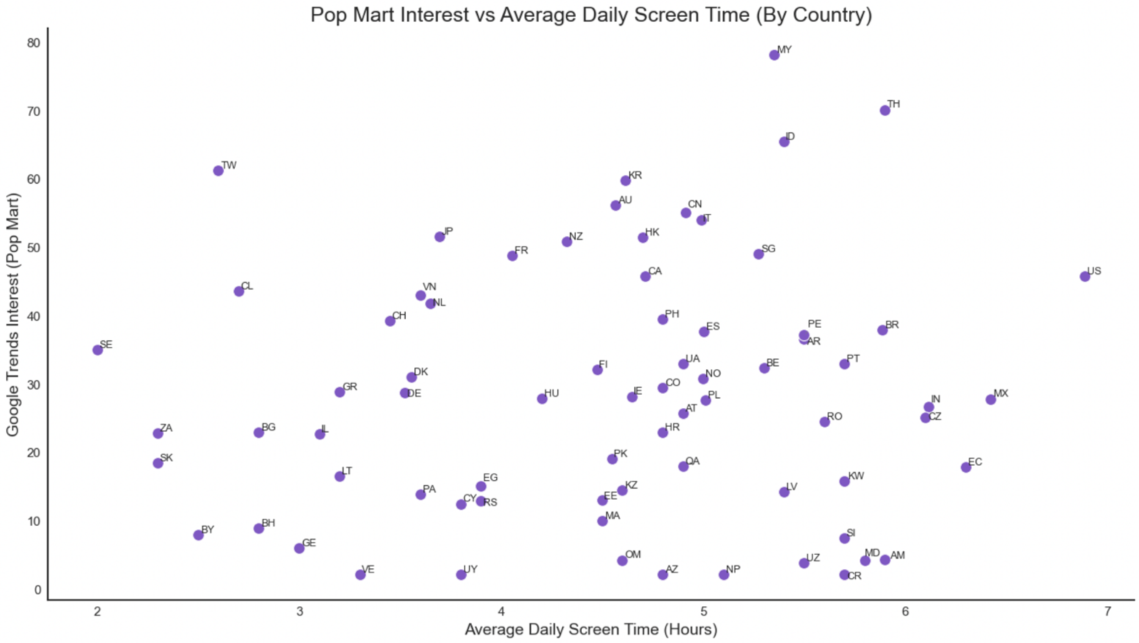
I started the analysis with creating my hypothesis first. Then I continued with EDA (exploratory data analysis). In this step I focused on understanding which parts of the datasets are valuable for understanding the correlation I am looking for.

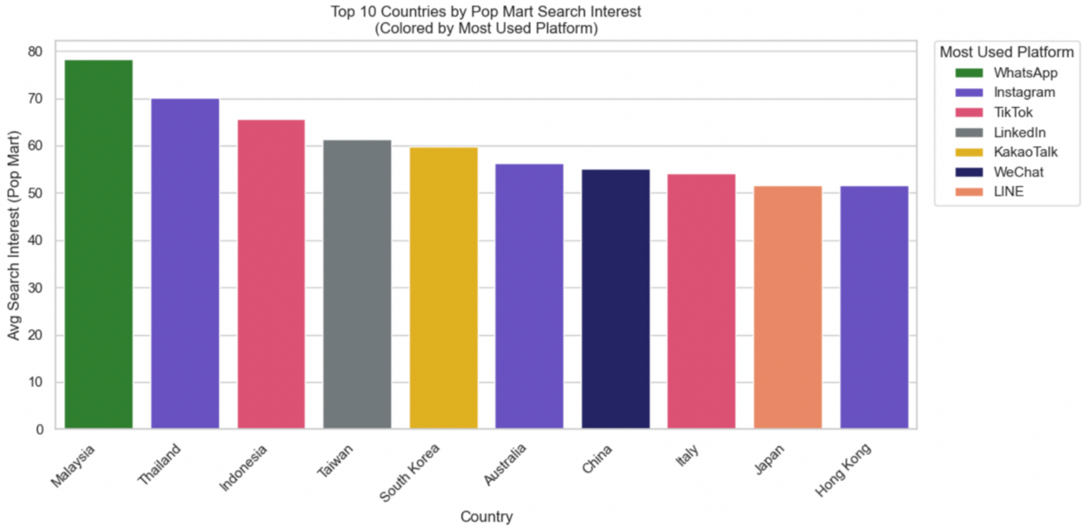
Hypothesis testing:

- **Research Question:** Does increased social media usage lead to higher consumer interest in online shopping?
- **Null Hypothesis:** There is not a significant relationship between daily social media usage and consumer behaviour in online shopping.
- **Alternative Hypothesis:** Increasing social media usage indicates an increase in consumer interest in online shopping and consumerism.

The steps I followed for EDA:

- I started by cleaning out the invalid values and handling missing inputs.
- Then using pytrends, I created the csv I need to handle search interests.
- Next, I detected the Inter Quartile Range (IQR)
- Later, I visualized the relationship using scatterplots and bar plots. Also applied statistical tests such as Spearman test, Pearson correlation, and Anova to examine the relation by p-values on numerical terms.
- Finally, after all the tests I run, the conclusion was there was not strong evidence that is supporting my Alternative Hypothesis which is “Higher social media usage leads to increased consumer interest in online shopping.”.
- Here are some example visualizations from EDA step of the project:

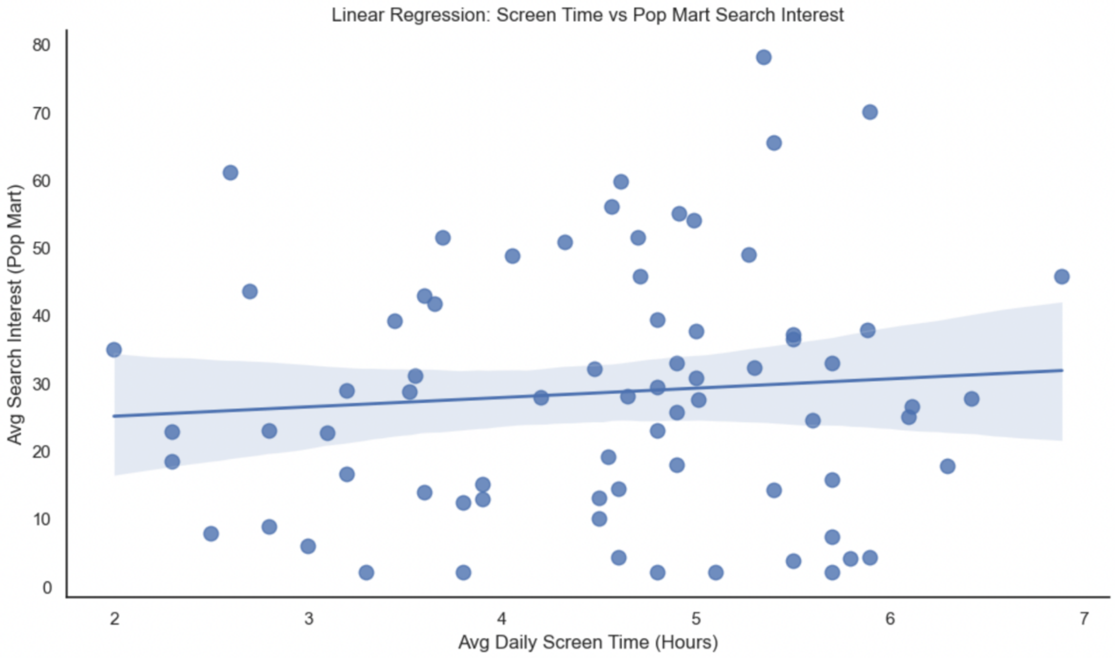


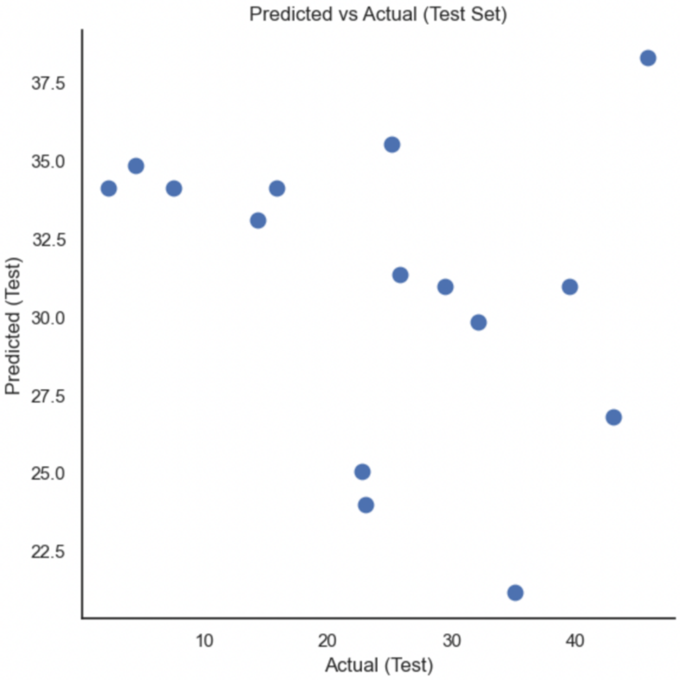


Then I continued the data analysis with machine learning. I applied supervised machine learning with my current findings to furtherly explore predictive relationships.

The steps I followed for machine learning:

- I first started by training the model with the selected features and target.
- Then I used Linear regression, Polynomial regression, Decision tree regression and KNN regression. I evaluated the performance of each Regression model using RMSE (root mean square error) and R^2 Score.
- Next, the results showed me that in all regression models R^2 Score was consistently negative with slight differences. It showed me that none of the models performed better than a baseline prediction using mean search interest. So, I added new features to see if the prediction performance increased.
- Lastly, I added Mean age per country, Gender distribution, Platform usage proportions and Number of individuals per country as new features but despite the feature enhancement none of the regression models improved significantly. Therefore, it indicates that relationship between social media usage and consumer interest is either weak or influenced by external factors that are not captured in the datasets.
- Here are some example visualizations from Machine Learning step of the project:





4.Key Findings

- Average daily social media usage is not a strong predictor of consumer shopping interest by itself.

- Adding demographic and platform information features improved interpretation and understanding but did not have any strong effect on changing the prediction accuracy.
- All Regression models consistently failed to give better performance than a baseline predicting the mean.
- Final decision: Consumer behavior is influenced by more complex and layered factors that are not captured from the observed datasets I have. The current features are not enough to predict and accurately interpret the results.

5. Limitations and Future Work

The first limitation I faced during this project was not having enough data on screen time. I was unable to use the screen time data of my own and the friends I observed while deciding the topic of this project because of the restrictions to reach needed data. A further exploration can be conducted with more enhanced data in the future. Also, during the analysis, the results showed me that the data I have is not enough to understand and predict the correlation. This could be because of either the features are not enough to clearly decide the hypothesis or the observation I made was not truly correct and did not apply for everyone. Some important variables such as incomes, advertising exposures and cultural factors were not available in the online datasets. For future enhancements, I am planning on finding datasets with more related features to enhance the datasets I have. I can also look more deeply into other deeper analysis methods. Maybe furtherly finding an e-commerce datasets can also enhance the finding on shopping behaviours.