Question

What is the relationship between the demographic type of an Instagram user (ie. urban, suburban, and rural) and the user's primary interest while scrolling through social media, quantified by maximum number of total likes, comments, and reposts on specific content categories (ie. news, entertainment, sports, etc) across the United States for the past year? Can liking a specific content category indicate a user's demographic type?

Hypothesis

Social media users with a more urban demographic background are likely to have different primary interests than those users from a rural demographic background. Users interested in travel and food are probably more likely to be from urban areas, while users interested in lifestyle, sports, and similar subjects are more likely to be from rural areas.

Justification

This is because of cultural differences, access to resources, and influence of surroundings. Typically, urban demographic social media users are going to be more interested in travel, food, fashion, entertainment, and city life. They will be more interested in travel because they have more access to money and modes of transportation such as airplanes, taxi's, buses, cars, trains, etc. that will be centered in urban areas. There are also lots of restaurants, stores, and party areas around so food, fashion, and entertainment are all going to be common themes and interests for social media users. On the other hand, rural demographic social media users are going to be interested in lifestyle, nature, and homesteading because they will be influenced by their natural/agricultural surroundings and secluded location.

Background Information

It's no surprise that social media has impacted our lives on a direct and indirect level. Instagram, Youtube, Facebook, TikTok, X, Snapchat – regardless of the platform, social networking sites have adapted greatly to cater to our primary interests and provide us with personalized entertainment. But in this project, we want to explore whether there is a correlation between the demographic type of the social media user and the consumer's primary interest while interacting with the medium in the United States.

Demographics generally include the characteristics or traits of the users of a product's audience, such as gender, age, location, occupation, relationship status, education level, among other variables.[1] In this project, the main focus will be on the user's location (urban, suburban, and rural). Social media platforms find it necessary to collect data on user demographics so as to better understand and discover their target audiences and gauge the type of content to present. For instance, Instagram is more frequently used by younger users for its visual content, Twitter for news consumers, TikTok for short form creative content or trends, or LinkedIn for professionals for networking and employment opportunities.[2] More importantly, recognizing the concentrated user bases can better allow social media companies to optimize and improve the overall effectiveness of their campaigns. In this sense, understanding the demographic makeup (urban, suburban, or rural) can greatly benefit the social media platform.

Primary interest often refers to a user's preferred focus or matter of interest when engaging with social media, specifically.[3] Subcategories of primary interest, in the context of social media, are endless, ranging from food to clothes to sports to television. Understanding the users' primary interests can allow social networking sites to categorize them to provide them with personalized recommendations to their services or connect with like-minded individuals.[4] By determining their users' primary interest, social media platforms can categorize their consumers more efficiently, identify more significant target groups, while also improving the performance of their marketing efforts and campaigns.

Though social media is a common topic in our technological world today, there are countless variables in play when it comes to social media users' primary interests and their location demographics. As a result, it is difficult to determine precisely and quantify the strength of the relationship between the two variables.

---

[1] https://later.com/social-media-glossary/demographics/
[2] https://www.linkedin.com/pulse/understanding-demographics-social-media-vimala-killi
[3] https://www.linkedin.com/pulse/primary-interest-filter-sarah-merron-gfrsc
[4] https://link.springer.com/article/10.1007/s10791-018-9338-x

Data

Ideally, the perfect dataset would scrape from Instagram directly and provide quantitative location data and every post every single user in the United States has liked, categorized into content categories, for the past year. From this, we would categorize every location under demographic categories (ie. urban, suburban, rural, etc.) However, without the consent of the user, using location and interest data is very unethical. Thus, the perfect dataset would be taken from voluntary inclusion of users. It would be from a survey of a large random sample of social media users, who would voluntarily allow us to look at their posts liked for the past year and categorize them amongst a large option of categories, as well as their zip code. The surveyed individuals would be representative of the United States and every region. The ideal raw dataset would include every single user in the United States with atleast the two variables we are analyzing, location type (ideally in city or coordinates) and engagement (likes, comments, and shares), so that we can clean and categorize each user into a demographic type and each post into a content category.

Chosen Dataset: Average Time Spent By A User On Social Media
https://www.kaggle.com/datasets/imyjoshua/average-time-spent-by-a-user-on-social-media

This data has 1000 observations of individual social media users across the countries of the United States, Australia, and the United Kingdom, and provides 12 variables, including a variable for 'demographics' (urban, suburban, rural) and 'interests' (lifestyle, travel, sports), but does not specify how interest is quantified, as needed through a likes or engagement variable. Other variables that are not of use for this project but would warrant ethical concerns are personally identifiable informations of age, gender, income, and whether they are in debt, own a house, or own a car. A difference between this dataset and ours is that it does not solely focus on users in the United States, and querying this dataset for the United States would provide us less observations than the desired 500. The overarching limitation of this dataset is the select specificity of users/categories it includes. For example, the 'interests' categories are limited to just Lifestyle, Travel, and Sports, and we would like to provide more nuanced options in our survey and data. Additionally, it only explores users of Youtube, Instagram, and Facebook. A perfect dataset would be more specific and heighten observations for just Instagram. Lastly, in exploring this dataset, the occupations are limited to Engineer, Student, and Marketing Manager. We would not use this variable in our research, but our data collection would ideally take a random sample of people, and thus encompass a far more diverse selection of occupations.

Ethical Considerations

Data Collection

- **Informed Consent** - ensuring users have a clear understanding as to what they are consenting to
    - Participants fully understand the purpose of the research and what data will be collected from them. They should be able to ask questions before agreeing to participate in the study
    - The privacy and confidentiality of the participants should be respected. They should be aware of how their data will be handled, stored, and anonymized to protect their own privacy. They should make sure that when dealing with sensitive information such as demographics and social media behavior, they should outline these measures to ensure the participants' confidentiality.
    - There should also be a sense of legal sensitivity where we respect and follow the harsh privacy laws regarding social media and informed consent
- **Collection Bias** - checks on sources of biases introduced during data collection and survey design
    - There could be sampling bias - if the sample of social media users selected for the study is not representative of the larger population, the results may not accurately reflect the demographics or interests of all the social media users in the U.S. If the study mainly collects data from participants from a specific social media platform, it may not capture the diversity of users from various demographics
    - There could be platform bias where different social media platforms attract various demographics and cater to different interests. For instance, a study focusing on Instagram users may not consider the preferences of adolescents who mostly use Facebook and TikTok
    - There could be temporal bias where social media trends and user behaviors could evolve over time. Study data taken at a certain period might not reflect the latest trends on a long-term basis
- **Limit PII Exposure** - ways to minimize exposure to personally identifiable information
    - Collect anonymous data - meaning collect data in a way that removes any direct identifiers with the user's name, email addresses, phone number, etc (can assign unique identifiers to avoid using their personal data)
    - Implement authorized access controls to minor and track data access activities
    - Ensure that you are in compliance with privacy regulations in the United States
    - Aggregating individual data into statistics and trends can help prevent identification of individuals based on where they are from
    - Store data safely by placing it in password-protected databases or servers so that you can prevent unauthorized access to data

Data Storage

- **Data Security -** a plan to protect your data
    - Ensure that any methods used for data collection are secure and encrypted so there is no unauthorized access during transmission (this could include data-scraping tools, one surveys, etc.)

    - Ensure that stored data in databases or servers is protected by firewalls, security suites like Norton, and any other security measures. This can be done by updating software on a regular basis
    - If data is being shared with third parties, make sure data agreements are made properly so data is not mishandled (can use encrypted methods for data exchange)
- **Right to be Forgotten** - user's right to have their personal data erased
    - Before collecting data, develop data deletion procedures that handle requests in which participants can get their data deleted after the study has been conducted. Ensuring that they comply with data protection regulations, implement data systems that can easily identify individuals and delete their data upon request
    - Ensure there is legal compliance with procedures
    - Have data audits where participants can be notified of their data usage, and they can promptly have their data deleted or anonymized if they would want it to
- **Data Retention** - a plan to delete the data if its no longer needed
    - Make clear the period of time participants will have their data retained for the study (they should consider the sensitivity of the information)
        - Dispose of data after usage by developing a procedure where digital and physical data is permanently deleted after the retention period has expired
    - When conducting research, collect and use data that reduces the usage of PII so there are less issues in regard to privacy

Data Analysis

- **Dataset bias and missing perspectives**
  In order to analyze the data correctly, the focus would be on these two specific columns/sections in the dataset - Demographic type (urban,suburban, and rural) and Primary interest(sports, fashion, technology, food, travel, etc). An efficient way to analyze the data to understand the distribution of interests among different demographics could be through charts and graphs. Before starting the analysis, it's important to ensure that diverse perspectives are incorporated and there is an adequate representation of all demographic groups. Stratified Sampling could also be one of the possible ways to mitigate any form of user bias.
- **Honest Representation and analysis strategies**
  Next, we could compare the interests across urban, suburban and rural users using basic statistical tests and then visualizing the findings to easily identify any patterns/trends in the data and recording it accordingly.Through multiple statistical tests, visualizations, summary statistics and reports and maintaining appropriate representation, the underlying data without misleading interpretations would be reflected.
- **Privacy and Auditability**
  Additionally, privacy should be prioritized during the analysis process and that can be done by handling sensitive data with great caution, anonymizing personally identifiable information (PII). To enhance the credibility and trustworthiness of the analysis, auditability must be ensured by documenting in detail throughout the process and every time the process is repeated, allowing for transparency and reproducibility of the results. By taking the following steps above, meaningful

insights into the relationship between demographic types and primary interest of social media users could be drawn and make the analysis robust, transparent, and inclusive.

Modeling (Statistical or Machine Learning)

- **Proxy discrimination**
  In modeling, it's important to ensure that there is fairness, transparency, and effectiveness. Proxy discrimination in modeling can be avoided by utilizing variables that are not discriminatory in nature, by incorporating sufficient data from all the demographics and from different parts in the US.
- **Metric Selection and Explainability**
  Metrics such as mean, mode, or distribution can be calculated for each demographic category such as urban, suburban, and rural users to understand the central tendency of interests within each group. Employing interpretable and transparent models such as decision trees or visualizing through bar charts or heat maps allows for a clear understanding of how demographic factors influence primary interests.

Project or Model Deployment

- **Support and Resources**
  In case sensitive information about participants of our research ever gets deployed, we could take responsibility for the same and make sure to protect their privacy. Upon taking our survey, users would be provided with an emergency email to contact us with any concerns about their data or questions about our research. If faced with a privacy leak or concern, we would provide them with legal counsel and if necessary, psychological aid.
- **Roll Back and Concept Drift**
  In creating a model that automatically updates based on newer data, we would save each and every data visualization and analysis created in every iteration for our researchers to access previous versions if there is a flawed or unnecessary update. The reason we would implement a model that automatically updates is that user interests can change as the income gap and value distribution changes between rural, suburban, and urban populations. This would ensure our model and resulting data analysis to keep up with current trends and shifts in social media.
- **Unintended Use**
  Unintended use has to do with our results being used for unintended and misintentional ways. With relation to our project, an example would be if a for-profit company utilizes our data and results to provide an increased targeted advertising to urban populations because they gravitate and engage with a certain interest or topic. We hope to be purposeful with where we use and publish our results to make sure it does not go towards unfair and profitable purposes.

Analysis Proposal

Data Collection

- In order to conduct this analysis, we would need to collect data from Instagram users across the United States. This would involve using the APIs provided by these platforms to gather information about users' demographics (urban, suburban, rural) and how often they use different platforms such as news, entertainment, and sports.
    - In terms of Instagram API, you should retrieve user profiles' demographic information if available and collect engagement data for each user, including the number of likes, comments and reposts on posts in different categories
    - In terms of Instagram API, you can gather similar data and include user demographics and engagement metrics. Through this manner, we can have a complete dataset that includes information from a large sample of users.
- The steps involved could mean using the Instagram Graph API which involves creating a developer account and registering to get access to an API. You can also use 'post data' which returns a list of media items with details like 'id', 'type' (image video), 'comments' and 'likes.' You can categorize this content by analyzing the captains or hashtags used on the posts. You can also use image recognition techniques to classify images into separate categories. At the end of data collection, storing such data in a structured format, in such databases or CSV files, will be convenient for analysis.

Data Wrangling

Once data collection is over, the next step would be to clean and prepare the data for analysis. In terms of data cleaning, remove duplicate entries to ensure each user is represented only once. Handle missing data by either inputting value or removing incomplete records. In terms of the data graphed, it should be checked if all data types have been converted into their correct timestamps/ datetime objects so that the units are uniform.

- You can create new features as needed, such as total engagement (with the sum of likes, comments, and reposts) across all categories. You can aggregate engagement metrics by content category to get total likes, comments, and reposts for each user in each category.
- In terms of data integration, combine the Instagram dataset into a single dataset, matching users based on unique identifiers (ex: user IDs). There should be consistency in column names and formats for easy analysis. We should normalize numerical features in the analysis, so it's easy for comparison.

Descriptive Analysis

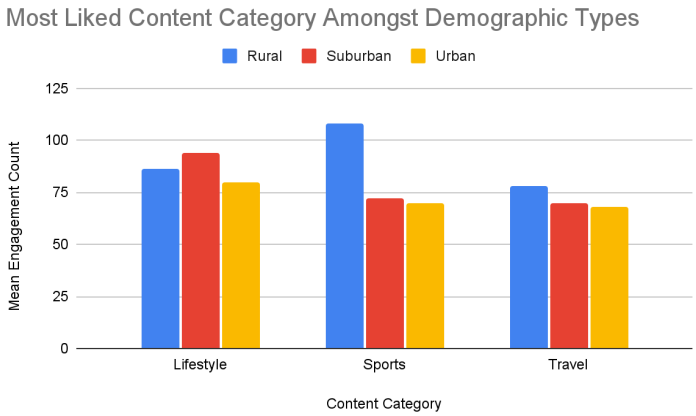
We will conduct descriptive analysis in order to test whether a correlation does exist between demographic type and a specific content category. Once we have a cleaned dataset with each users' demographic type and total number of likes, comments, and shares per content category, we can generate some summary statistics to provide us a better understanding of the relationship between demographic type and content category. We would get the mean number of total engagement per each content category for all the demographic types; for example, we would look at the mean number of total engagement with lifestyle posts for the rural demographic type, and compare that to the values for suburban and urban for the lifestyle content category.

Exploratory Data Analysis

Then, we would conduct exploratory data analysis to visualize these correlations and test whether a statistical analysis would work. For this, we will be making exploratory plots that give us a good look at the relationships and surface-level correlations present in our dataset. Mainly, we would map each demographic type-content category pairing on a box plot, and compare with the other demographic type plots, to see which has a higher mean and shorter range, which would show a correlation of that specific demographic type engaging with that specific content category more.

Data Visualization

To answer the first part of our Data Science Question and visualize our data to showcase which demographic type enjoys a certain content category more, we could make a multi-group histogram, with the y-axis being the average number of engagement (total likes, comments, and shares) per user and the x-axis having the content categories. Each bar represents the average total engagement count with a content category for a user in each demographic type. Per content category, we would have Rural in blue, Suburban in red, and Urban in yellow to differentiate the stacked bars by demographic type.



Disclaimer: The above graphs are on our example data that does not include each user's specific likes and engagement. We would be mapping the average total number of likes, comments, and shares per content category and demographic type to get a more concrete and quantifiable result and visualization about the average engagement towards a certain category by a user in a specific demographic type. The image above is shown just to show our proposed layout/graph, axes units, and color choices.

As for the second part of our Data Science Question, trying to predict the demographic type of a user based on their engagement data, we would visualize after doing the Random Forest Analysis (Look at Analysis Type 1). For this, we believe a confusion matrix heatmap is the best way to reflect the predictions and likelihoods by a Random Forest model. A confusion matrix heatmap is a two-dimensional array with a mono-color gradient on the side. Each coordinate square has a darker shade depending on the correlation between the two said items. In this case, whichever demographic square and content category has the darkest color reflects the demographic category and most-liked type of content of the user, while the lightest square suggests the least-likely conclusion regarding the user. This way, we see all the possible combinations on the heatmap yet can visually see the highest correlation combination.

Analysis Type 1 (Predictive Analysis)

Random Forest Model

The first step before conducting predictive analysis would be Feature Engineering as it goes beyond raw data like total likes and comments. It creates more informative features that capture user behavior. Including engagement ratios (likes/comments/reposts compared to total posts viewed) and accounting for individual activity, frequency features (how often a user interacts with each content category), and time-based features (engagement patterns across times of day/week for potential location insights) would be essential in the process. Additionally, if location data is available, users can be categorized as urban, suburban, or rural based on population density.

To understand how social media engagement relates to user demographics, we can employ a random forest model. Since we have to determine the complex relationships between user behavior (like engagement ratios and content preferences) and demographics (urban, suburban, rural), a random forest model would work.

Advantages of using a Random Forest model:

Random Forests is an ensemble method combining multiple decision trees and hence can handle non-linear relationships or potentially complex relationships between user engagement features (ratios, frequencies, time-based patterns) and user demographics (urban, suburban, rural).

They are also robust to outliers in the data, which might be present in social media engagement analysis.

Analysis through Random Forests

Imagine the model as having multiple decision trees, each analyzing the data from a slightly different angle. By combining insights from all angles possible, the random forest could make more accurate predictions.

- First, we could train the model on data with known user demographics and their engagement patterns.
- Next, the data is split into training and testing sets. The training set provides the model with foundational data to work with. Each tree in the forest analyzes a unique subset of features and data points from this training set.
- Through multiple iterations, these trees enhance their decision-making abilities, learning to identify patterns that differentiate user demographics based on engagement patterns.
- Then, we test its ability to predict demographics for new users based solely on their social media behavior.
- By analyzing which engagement features influence the model's predictions most, we can gain insights into how user preferences on different content categories might indicate their location (urban, suburban, rural).
- For instance, high news engagement ratios might suggest an urban user, while frequent sports interaction could point towards a suburban demographic.

- It's important to remember that these predictions are probabilistic, and other factors beyond demographics can influence social media behavior.
- However, a well-trained random forest model offers a valuable tool for understanding the connection between user demographics and their online engagement.

Additional Considerations:

Bias Mitigation: To prevent biased predictions, it's essential to maintain a balance in the training data across different demographic groups. Keeping an eye out on the model's performance and promptly addressing any potential biases by monitoring and adjusting could also help in handling any potential bias in the analysis.

For readability and explainability, using techniques such as LIME (Local Interpretable Model-agnostic Explanations) would ensure understanding the reasons behind the model's predictions. This would foster trust and transparency in the analysis.

Analysis Type 2 (Statistical Analysis)

Hypothesis Testing for Social Media Engagement and Demographics

Formulating Hypotheses:

Null Hypothesis (H0): There is no significant difference in engagement (likes, comments, reposts) for specific content categories across user demographics (urban, suburban, rural).
Alternative Hypothesis (H1): There is a significant difference in engagement for specific content categories across user demographics (urban, suburban, rural).

Choosing a Statistical Test:

One-way ANOVA: Compares the means of engagement metrics (likes, comments, reposts) across the three demographic groups (urban, suburban, rural) for a single content category.
Two-way ANOVA: Analyzes the interaction effect between user demographics and content category on engagement metrics.
Chi-square test: If engagement is measured categorically (e.g., high/low engagement), this test could assess if there's a relationship between demographics and content category preferences.

Steps before conducting the test:

Data Preparation:
Ensuring data is clean and free of errors.
Considering data transformations (e.g., log transformation) if the data is skewed.
Setting a Significance Level (alpha):
This is the probability of rejecting the null hypothesis when it's actually true. Common choices could be alpha = 0.05 (5%) or alpha = 0.01 (1%).

Conducting the Test:

Through statistical software (R, Python) or online tools the chosen test can be performed. The software will provide test statistics (F-statistic for ANOVA, chi-square statistic) and a p-value.

Interpreting the Results:

Comparing the p-value to the chosen significance level (alpha).
p-value < alpha: Reject the null hypothesis (H0). There is evidence to suggest a significant difference in engagement across demographics.
p-value >= alpha: Fail to reject the null hypothesis (H0). There's not enough evidence to conclude a significant difference.

Conclusion:

Power analysis can be conducted before the study to determine the required sample size to achieve a desired level of statistical power (the probability of rejecting H0 when it's truly false). Having a larger sample size provides more statistical power to detect true differences in engagement across demographics.

Based on the test results, we can draw conclusions about the relationship between user demographics and content preferences.In some cases, a statistically significant difference doesn't always necessarily imply a practical or meaningful difference in engagement.

Instead of a single H0 and H1 for all categories, we can formulate more specific hypotheses for each content category (e.g., news, entertainment, sports) too to gain a more nuanced understanding between the demographic type and the specific type of social media engagement on Instagram.

Discussion

In our proposed analysis to find a relationship between a social media user's demographic type (urban, suburban, rural) and their primary interest on platforms like Instagram, the two analyses presented offer distinct but complementary approaches to understanding the relationship. The predictive analysis using the random forest model explores the complexity of user behavior by producing informative features and implementing a strong method to predict the demographics based on user social media engagement patterns. This method offers a thorough understanding of the interactions between multiple user engagement metrics and user demographics, providing valuable insights about how preferences for different content categories might suggest urban, suburban, or rural demographics. On the other hand, the statistical analysis uses hypothesis testing to determine if there is a significant difference in social media engagement across demographic groups. By employing tests like ANOVA and chi-square, it is possible to statistically evaluate the relationship between engagement metrics and user demographics and use analysis to support conclusions. Both approaches offer their advantages – therefore combining insights from both analyses can lead to a deeper understanding of the correlation between social media engagement and user demographics, enabling more informed decision-making in content strategy and audience targeting on platforms like Instagram.

However, several limitations and potential biases need to be taken into account. To begin, our dataset inherently possesses biases. The dataset might not perfectly reflect the diversity of social media users, especially in terms of demographic representation and platform usage. For example, the dataset concentrates on users from YouTube, Instagram, and Facebook, overlooking those on new or less popular platforms such as TikTok or Snapchat. This restriction can manipulate our interpretation of the larger social media world, potentially influencing us to overlook trends among users of these excluded platforms. To address this, we could supplement our primary dataset with additional, more recent data sets from a broader range of platforms to gather a sample that is a more representative sample of the overall social media population today.

In addition, the categorization of interests (lifestyle, travel, sports) in our dataset is restrictive and does not come close to encompassing the vast variety of user interests on social media. To moderate this problem, developing our analysis to include a more extensive list of interest categories would provide a better understanding of trends among social media users in different demographic groups and primary interest platforms.

Another important aspect is the dynamic of social media trends. Our analysis is based on older data, which may not accurately represent the current or future trends taking over social media. Social media is dynamic and ever-changing, and user interests can shift rapidly. Hence, the results of our proposed analysis may only present a historical snapshot rather than a live and current or predictive view. Continuously updating our dataset and analysis to provide more recent data would help in keeping our results more relevant to the present time.

In terms of societal and ethical implications, our project involves a responsible approach to data privacy and user consent. As outlined in the Ethical Considerations section, we must ensure informed

consent, minimize exposure to personally identifiable information (PII), and respect user privacy throughout our data collection, storage, and analysis processes. This approach involves ethically collecting user data, ensuring anonymity in our dataset, and administering solid data security measures. Furthermore, we must be careful in interpreting and distributing our findings to prevent its misuse, such as stereotyping or unfair and unjustified targeting of specific demographic groups based on their social media behavior.

In conclusion, though our analysis strives to provide an insightful relationship between demographics and users' social media interests, it's critical to acknowledge and address its limitations, pitfalls, biases, and the ever-changing nature of social media. By continuously updating our dataset and approaches and strictly sticking to our ethical standards, we can improve the reliability and responsibility of our proposed research.