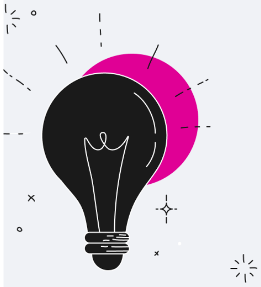
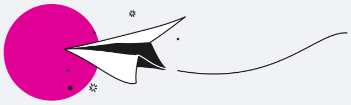
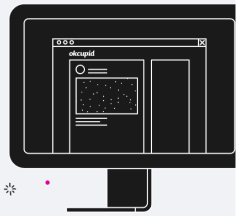




Exploratory Analysis of OkCupid Users

Written Supplement

Michael Charara
10.03.2023



Overall OkCupid User Data Analysis Project

Introduction:
The primary objective of this data analysis was to dive deep into the OkCupid dataset to generate actionable insights, assisting the marketing team in their future targeted campaigns. By analyzing the users' demographics, preferences, and behavior, we aimed to unearth patterns that can guide marketing initiatives, ensuring they're as effective and efficient as possible.

Data Sources and Preparation:
Sources Utilized:
- Main Dataset: OkCupid user profiles.
- Supplementary Data: Geographic coordinates, census information, and address data.

Data Cleaning & Preprocessing:
- Outliers, such as specific unrealistic age values, were identified and removed.
- Missing values were managed to ensure a smooth analysis process.
- For clarity, the 'education' attribute was recategorized into broader categories.

Detailed Insights:
1. Age and Orientation:
    - The dataset revealed a significant chunk of 25-30 age bracket users.
    - The predominant orientation among users is 'Straight,' followed by 'Gay' and 'Bisexual.'
2. Income, Relationship Status, and Education:
    - Users with 'Available' and 'Single' relationship statuses tend to have higher incomes.
    - There's a discernible link between income and education levels, offering opportunities for segmented marketing.
3. Geographic Distribution:
    - Certain regions in the U.S. showcase a higher concentration of OkCupid users.
    - This geographical concentration is vital for region-specific marketing campaigns.
4. Lifestyle and Preferences:
    - A significant number of straight users prefer to drink socially.
    - Conversely, there's a notable segment among gay and bisexual users who abstain from drinking.

Recommendations for Marketing Initiatives:
- Age-targeted Marketing: Given the age distribution, campaigns tailored for the 25-30 age bracket will likely yield higher engagement.
- Income & Education Segmentation: Marketing initiatives can be optimized by targeting segments based on their income and education, as they are interrelated.
- Geo-targeted Campaigns: Geo-specific campaigns can be more impactful with the identified user concentrations in specific U.S. regions.
- Lifestyle Alignments: Campaigns can resonate better if they are aligned with the predominant lifestyle choices of the target segments, such as drinking habits.

Conclusion:
The OkCupid dataset, enriched with supplemental data, offers many insights that can drive more focused and result-driven marketing campaigns. Our approach, supported by clear narration and transparent code, was geared towards revealing these insights to meet the objectives laid out at the onset of this project. OkCupid can tailor its marketing strategies for better engagement, reach, and outcomes by acting on these insights.

OkCupid User Data Analysis Presentation Summary:

Objective: Deep dive into the OkCupid dataset to assist the marketing team in targeted campaigns.

Data Sources:

- OkCupid user profiles.
- Supplemental geographic and census data.

Data Cleaning:

- Managed outliers and missing values.
- Recategorized 'education' attribute for clarity.

Key Insights:

1. Age Distribution:
   - Peak user age: 25-30 years.
   - Majority orientation: Straight.
2. Income & Relationship:
   - Higher incomes: 'Available' and 'Single' users.
   - Connection between income and education levels.
3. Geographical Insights:
   - User concentration in specific US regions.
   - Potential for geo-targeted campaigns.
4. Lifestyle Choices:
   - The majority of straight users Drink socially.
   - Significant non-drinking segment among gay and bisexual users.

Recommendations:

- Optimize campaigns for the age bracket 25-30.
- Tailored campaigns for higher income segments.
- Geo-targeted marketing for concentrated regions.
- Align campaigns with predominant lifestyle choices.

Code Explanation for OkCupid User Data Analysis:

1. Environment Setup:
   - Working Directory Setup: Defined the working directory where the datasets and other associated files reside.
   - Library Loading: Loaded essential libraries required for data processing and visualization.
2. Data Loading:
   - Reading Raw Data: Imported four separate CSV files — 'profiles.csv' (primary user data), 'LatLon.csv' (latitude-longitude data), 'addr.csv' (address data), and 'sharedCensus2010Vars.csv' (census variables).
3. Data Merging:
   - Consolidation: Combined the main profiles dataset with the three supplemental datasets based on the 'location' column, resulting in a merged dataset (`merged_data`) enriched with geographical and census information.
4. Data Cleaning for Age & Orientation Insight:
   - Outliers Removal: Excluded unrealistic age values (109 and 110) from `merged_data.
   - Insight Generation: Created a histogram showcasing the distribution of users by age (20-40) and orientation.
5. Relationship Status and Income Analysis:
   - Custom Color Palette: A distinct color palette was defined to visualize different relationship statuses uniformly.
   - Education Recategorization: Grouped education into broader categories like "High School," "College/University," etc., for clarity.
   - Income vs. Relationship Status: Generated a bar plot that presents average income by relationship status.
6. Income, Education, and Relationship Status Insight:
   - Education Levels Order: Prioritized the order in which education levels should be presented on plots.
   - Overlapping Bars: Created a bar chart representing average income segmented by education level and overlapped by relationship status.
   - Side-by-Side Bars: Given the discrepancy in overlapping visualizations, switched to side-by-side bars for clearer representation.
7. Geographic Distribution Insight:
   - Data Filtering: Selected data points that fall within the latitude and longitude boundaries of the continental U.S.
   - Map Visualization: Used U.S. boundaries and plotted the geographic distribution of users based on their orientation.
8. User Orientation and Drinking Habits Insight:
   - Factor Reordering: Ordered the 'drinks' factor in a specific sequence for visualization consistency.
   - NA Values Handling: Filtered out rows with missing values in the 'orientation' and 'drinks' columns.
   - Insight Visualization: Developed a stacked bar chart showcasing the distribution of users by orientation, segmented by their drinking habits.

Methods were grounded in best practices for data cleaning, merging, and visualization to draw actionable insights.