As a group, you work at a consulting firm tasked with analyzing survey data. The dataset contains various demographic and opinion-based variables gathered from a national sample of adults in the United States.

[The PEW September 2020 Survey](https://www.pewresearch.org/global/2020/11/23/americans-and-germans-head-into-2021-with-divergent-opinions-on-transatlantic-alliance/) data which contains 4,455 observations and 14 variables.

**Information about the data:** *“The phone interviews conducted Sept. 22-28, 2020, among a national sample of 1,007 adults, 18 years of age or older, living in the United States (301 respondents were interviewed on a landline telephone, and 706 were interviewed on a mobile phone, including 487 who had no landline telephone). A combination of landline and mobile phone random-digit-dial samples were used. Interviews were conducted in English (972) and Spanish (35). The combined landline and mobile phone sample is weighted to provide nationally representative estimates of the adult population 18 years of age and older.”*

***Instructions:***

Please follow the instructions provided in each task.

It is crucial that your responses are clear and written with no grammatical errors as this is a professional setting. Make sure to properly number your answers so that it is easy for the reader to follow your analysis.

You are expected to complete the project by using Microsoft Excel.

There are three deliverables:

*(1) Your project report in Word format.*

*(2) Your Excel file, detailing your work for each task.*

*(3) A Statement of Contribution Form signed by each team member. Each member should indicate their specific contributions to the project. Group project grades will not be released without this signed form.*

If you have any questions about the mini-group project tasks, please post them in the Week 3 discussion forum.

*Good luck!*

**Part 1: Data Cleaning and Preparation**

* + 1. **Variable Types:** Classify each variable as categorical or numerical. For categorical variables, specify whether they are ordinal or nominal. For numerical variables, specify whether they are discrete or continuous, and whether they are ratio or interval measurements.

|  |  |  |  |
| --- | --- | --- | --- |
| state |  |  |  |
| mstatus |  |  |  |
| totper |  |  |  |
| adults |  |  |  |
| parent |  |  |  |
| age |  |  |  |
| educ |  |  |  |
| income |  |  |  |
| hispanic |  |  |  |
| race |  |  |  |
| partyln |  |  |  |
| polview |  |  |  |
| sex |  |  |  |
| religion |  |  |  |
| Q1 |  |  |  |
| Q2 |  |  |  |
| Q3a |  |  |  |
| Q3b |  |  |  |
| Q4 |  |  |  |
| Q5a |  |  |  |
| Q5b |  |  |  |
| Q5c |  |  |  |
| Q5d |  |  |  |
| Q5e |  |  |  |
| Q5f |  |  |  |
| Q6 |  |  |  |

**Missing Values:** Identify the number of missing values for each variable and calculate the percentage of missing information.

Note: A data entry “NA” is considered as missing. Do NOT treat entries such as “Refused”, “Don’t know", " DK/Refused”, “Neither/Other (DO NOT READ)”, or “VOL: Neither” as missing.

|  |  |  |
| --- | --- | --- |
| Variable | Number Missing | Percentage Missing |
| state |  |  |
| mstatus |  |  |
| totper |  |  |
| adults |  |  |
| parent |  |  |
| age |  |  |
| educ |  |  |
| income |  |  |
| hispanic |  |  |
| race |  |  |
| partyln |  |  |
| polview |  |  |
| sex |  |  |
| religion |  |  |
| Q1 |  |  |
| Q2 |  |  |
| Q3a |  |  |
| Q3b |  |  |
| Q4 |  |  |
| Q5a |  |  |
| Q5b |  |  |
| Q5c |  |  |
| Q5d |  |  |
| Q5e |  |  |
| Q5f |  |  |
| Q6 |  |  |

**Replace Missing Values:** Use appropriate imputation methods to replace missing values. For categorical variables, replace missing values with the mode. For numerical variables, replace missing values with the median.

**Part 2: Visual Analytics**

1. As a team, leverage the dataset to create a compelling visualization that illustrates the notion that 'Younger Americans tend to favor a closer relationship with China over Germany.' In a minimum of three sentences, articulate how your chosen visual effectively substantiates this claim. When creating your visualization, if you've omitted specific rows from the dataset, please include a brief explanation for this decision in your report.

Use the following ranges for the age variable before constructing a chart that shows the relation between Age and Q3b.

* Ages 18-29
* Ages 30-49
* Ages 50-64
* Ages 65+

1. Does income level influence the preference for a closer relationship with Germany as opposed to Russia? Utilize the dataset to craft a persuasive visual that explores this trend. Provide a detailed explanation of at least three sentences to describe what your visualization reveals about this relationship. When creating your visualization, if you've omitted specific rows from the dataset, please include a brief explanation for this decision in your report.

When constructing your visual, use the following income categories to classify people as “Low Income”, “Middle Income” and “High Income”:

* **Low Income**: $49,999 or less
* **Middle Income**: $50,000 to $99,999
* **High Income**: $100,000 and over

**Part 3: Replication**

The report uses the data on variables **Q5a-f** and **partyln** to construct the following graph.

**Table

Description automatically generated** Use Excel to create a graph that closely resembles the one provided above. While we may not have details on how specific survey responses were processed to construct the original graph, your aim should be to produce a visualization in Excel that approximates it as closely as possible. Don't worry about matching the exact numbers; the focus is on capturing the overall trend and structure.

**Part 4: Consulting**

Your team has been contracted by a political party interested in utilizing Facebook for targeted advertising. This party views China as a geopolitical rival and is particularly concerned with public opinion on the question, "How likely is the current tension between China and the United States to escalate into a situation similar to the Cold War?" (referred to as 'Q4'). Your assignment is to analyze the responses to Q4 across various demographic factors such as gender, age, income, religion, marital status, and political views. Your goal is to identify a specific demographic segment that the political party should target for their Facebook advertising campaign. What demographic group do you recommend targeting, and how did you arrive at this conclusion?