ADTA 5130 Section 100 – Fall 2023

Mini Project 1 – Group 9

Group Members: Harini Kamarthy

Nisha Bhattarai

Vaishnavi Gunna

Part 1 : Data Cleaning and Preparation

i) Variable Types

Variables	Type	Subtype	Measurement type
state	Categorical	Nominal	
mstatus	Categorical	Nominal	
totper	Numerical	Discrete	Ratio
adults	Categorical	Nominal	Ratio
parent	Categorical	Nominal	
age	Numerical	Continous	Ratio
educ	Categorical	Ordinal	
income	Numerical	Continous	Ratio
hispanic	Categorical	Nominal	
race	Categorical	Nominal	
partyln	Categorical	Nominal	
polview	Categorical	Ordinal	
sex	Categorical	Nominal	
religion	Categorical	Nominal	
Q1	Categorical	Nominal	
Q2	Categorical	Ordinal	
Q3a	Categorical	Nominal	
Q3b	Categorical	Nominal	
Q4	Categorical	Nominal	
Q5a	Categorical	Nominal	
Q5b	Categorical	Nominal	
Q5c	Categorical	Nominal	
Q5d	Categorical	Nominal	
Q5e	Categorical	Nominal	
Q5f	Categorical	Nominal	
Q6	Categorical	Ordinal	

Part 1 :

ii) Missing Values

Calculated the missing values count and it's percentage with the related formulas in the attached Excel file in the sheet named **"Q1 – Missed Values Calculation".** Considered only **NA** values as the missing values.

Column Name	Missing Values Count	Percentage of Missing Values
state	0	0.00%
mstatus	0	0.00%
totper	0	0.00%
adults	0	0.00%
parent	797	79.15%
age	0	0.00%
educ	0	0.00%
income	0	0.00%
hispanic	0	0.00%
race	0	0.00%
partyln	634	62.96%
polview	0	0.00%
sex	0	0.00%
religion	0	0.00%
Q1	0	0.00%
Q2	0	0.00%
Q3a	0	0.00%
Q3b	0	0.00%
Q4	0	0.00%
Q5a	0	0.00%
Q5b	0	0.00%
Q5c	0	0.00%
Q5d	0	0.00%
Q5e	0	0.00%
Q5f	0	0.00%
Q6	0	0.00%

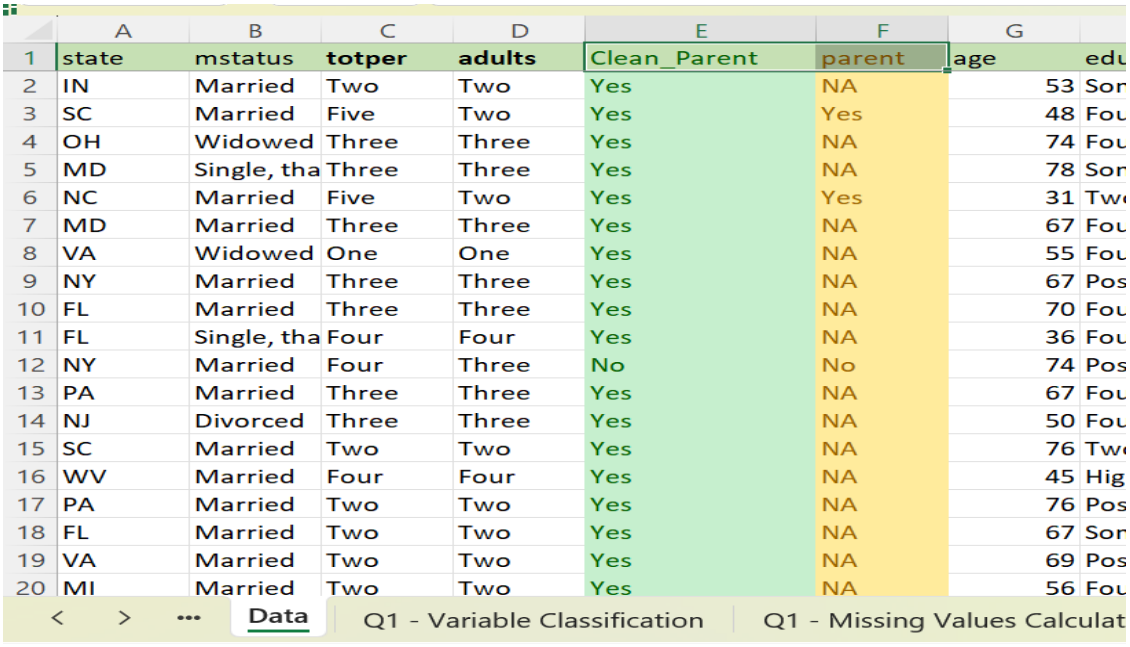
Part 1 :

iii) Replacing Missing Values

As we have the missing values in **"Parent"** and **"Partyln"** variables, and both are the Categorical Variables, we calculated the mode (taking the count of each category and

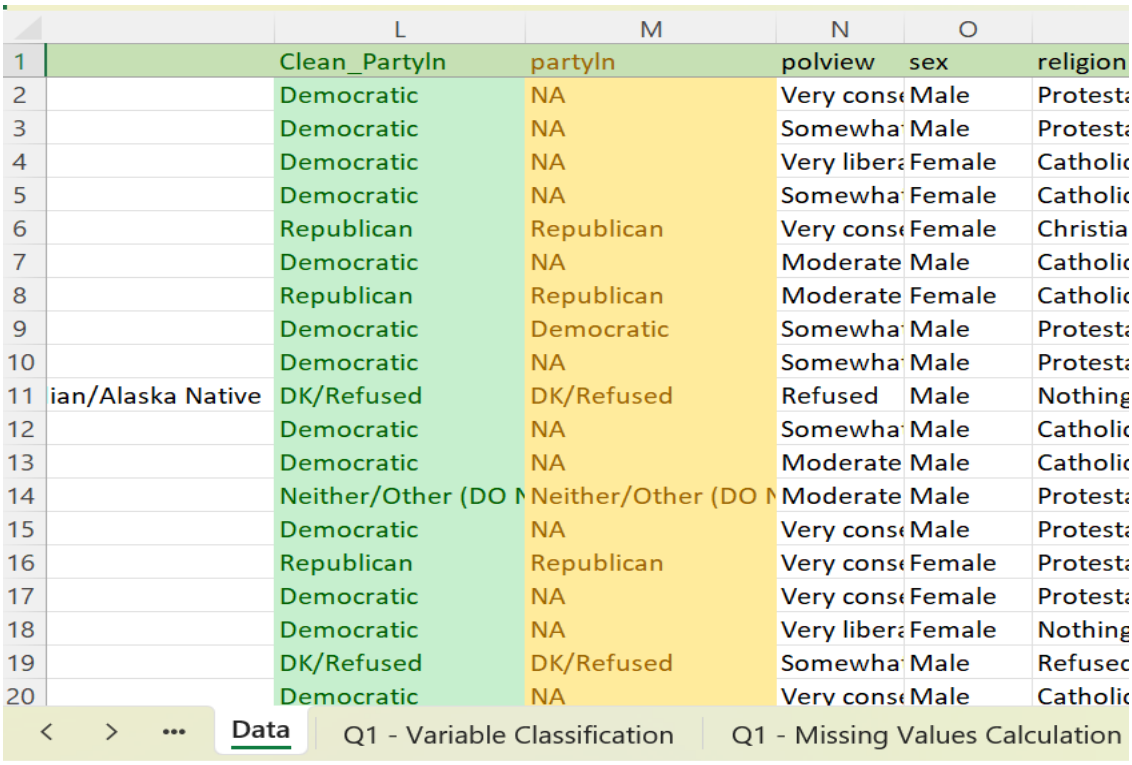
considering the mode as the one which have the highest count) and replaced the missing values with the mode, using the related formula **“=IF(F2="NA", "Yes", F2)”** for **“Parent”** and **“=IF(M2="NA", "Democratic", M2)”** for **“Partyln”**.

After replacing the missing values, the variable is added in the new columns say **“Clean_Parent”** and **“Clean_Partyln”** in the **“Data”** original sheet.



	A	B	C	D	E	F	G	
1	state	mstatus	totper	adults	Clean_Parent	parent	age	edu
2	IN	Married	Two	Two	Yes	NA	53	Son
3	SC	Married	Five	Two	Yes	Yes	48	Fou
4	OH	Widowed	Three	Three	Yes	NA	74	Fou
5	MD	Single, tha	Three	Three	Yes	NA	78	Son
6	NC	Married	Five	Two	Yes	Yes	31	Two
7	MD	Married	Three	Three	Yes	NA	67	Fou
8	VA	Widowed	One	One	Yes	NA	55	Fou
9	NY	Married	Three	Three	Yes	NA	67	Pos
10	FL	Married	Three	Three	Yes	NA	70	Fou
11	FL	Single, tha	Four	Four	Yes	NA	36	Fou
12	NY	Married	Four	Three	No	No	74	Pos
13	PA	Married	Three	Three	Yes	NA	67	Fou
14	NJ	Divorced	Three	Three	Yes	NA	50	Fou
15	SC	Married	Two	Two	Yes	NA	76	Two
16	WV	Married	Four	Four	Yes	NA	45	Hig
17	PA	Married	Two	Two	Yes	NA	76	Pos
18	FL	Married	Two	Two	Yes	NA	67	Son
19	VA	Married	Two	Two	Yes	NA	69	Pos
20	MI	Married	Two	Two	Yes	NA	56	Fou

Data | Q1 - Variable Classification | Q1 - Missing Values Calculat



		L	M	N	O	
1		Clean_Partyln	partyln	polview	sex	religion
2		Democratic	NA	Very conse	Male	Protesta
3		Democratic	NA	Somewha	Male	Protesta
4		Democratic	NA	Very libera	Female	Catholic
5		Democratic	NA	Somewha	Female	Catholic
6		Republican	Republican	Very conse	Female	Christia
7		Democratic	NA	Moderate	Male	Catholic
8		Republican	Republican	Moderate	Female	Catholic
9		Democratic	Democratic	Somewha	Male	Protesta
10		Democratic	NA	Somewha	Male	Protesta
11	ian/Alaska Native	DK/Refused	DK/Refused	Refused	Male	Nothing
12		Democratic	NA	Somewha	Male	Catholic
13		Democratic	NA	Moderate	Male	Catholic
14		Neither/Other (DO N	Neither/Other (DO N	Moderate	Male	Protesta
15		Democratic	NA	Very conse	Male	Protesta
16		Republican	Republican	Very conse	Female	Protesta
17		Democratic	NA	Very conse	Female	Protesta
18		Democratic	NA	Very libera	Female	Nothing
19		DK/Refused	DK/Refused	Somewha	Male	Refused
20		Democratic	NA	Very conse	Male	Catholic

Data | Q1 - Variable Classification | Q1 - Missing Values Calculation

Part 2 : Visual Analytics

a) 'Younger Americans tend to favor a closer relationship with China or Germany.'

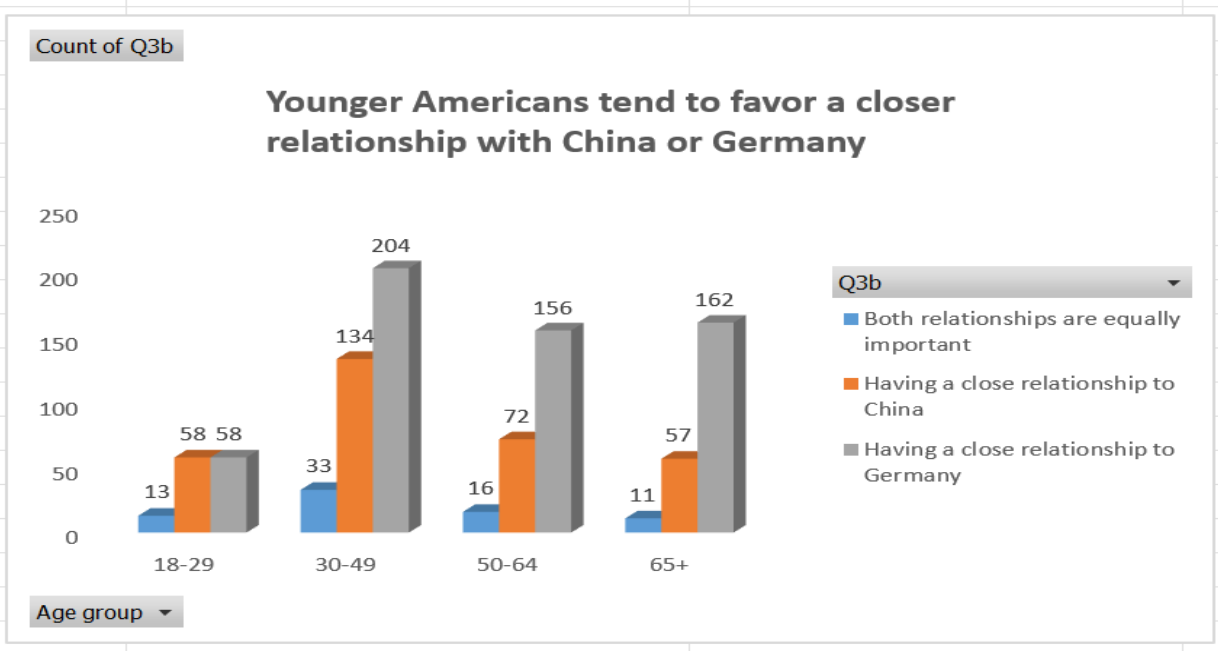
We have used the two variables for this analysis, **“age”, “Q3b”**, in the data sheet **“Q2(a) Visual Analytics”**. Removed the data rows which have the values as “DK/Refused”, ”Refused”, “VOL:Neither”, as we only need to see the relationship favor of the younger Americans with China and Germany. So, cleaned data helps us to give the results more effectively with proper analysis. We have grouped the ages as below categories, using the formula **“=IF(AND(A2>=18, A2<=29), "18-29", IF(AND(A2>=30, A2<=49), "30-49", IF(AND(A2>=50, A2<=64), "50-64", IF(A2>=65, "65+", ""))))”**.

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

Added this in the new column **“Age group”**.

Drawn the **Pivot table** for “Q3b” and “Age group” to know the count of each category. And then plotted the **Clustered Bar Chart** for the drawn Pivot table.

Count of Q3b	Column Labels			
Row Labels	Both relationships are equally important	Having a close relationship to China	Having a close relationship to Germany	Grand Total
18-29	13	58	58	129
30-49	33	134	204	371
50-64	16	72	156	244
65+	11	57	162	230
Grand Total	**73**	**321**	**580**	**974**



From the above plotted graph, we can infer three points here –

- Younger Americans (18-29) data show that there is an equal number of people exists who tend to favor a closer relationship with China and Germany.
- Most of the Americans tend to favor a closer relationship with Germany over China.
- Most of the Americans who tend to favor a closer relationship with Germany over China are of from age group 30-49.

b) "Income level influence the preference for a closer relationship with Germany as opposed to Russia"

We have used the two variables for this analysis, **"income", "Q3a"**, in the data sheet **"Q2(b) Visual Analytics"**. Removed the data rows which have the values as "DK/Refused", "Refused", "VOL:Neither", "Don't Know" as we only need to see the income level influence preference for relationship with Germany and Russia. So, cleaned data helps us to give the results more effectively with proper analysis. We have grouped the income as below categories, using the formula **=IF(OR(**

ISNUMBER(SEARCH("Less than $15,000", A2)),

ISNUMBER(SEARCH("$15,000 but less than $25,000", A2)),

ISNUMBER(SEARCH("$25,000 but less than $30,000", A2)),

ISNUMBER(SEARCH("$30,000 but less than $40,000", A2)),

ISNUMBER(SEARCH("$40,000 but less than $50,000", A2)),

ISNUMBER(SEARCH("Less than $50,000 (Unspecified)", A2))

), "Low Income", IF(OR(

ISNUMBER(SEARCH("$50,000 but less than $75,000", A2)),

ISNUMBER(SEARCH("$75,000 but less than $100,000", A2)),

ISNUMBER(SEARCH("$50,000 but less than $100,000 (Unspecified)", A2)),

), "Middle Income", IF(OR(

ISNUMBER(SEARCH("$100,000 and over (Unspecified)", A2)),

ISNUMBER(SEARCH("$100,000 to under $150,000", A2)),

ISNUMBER(SEARCH("$150,000 to under $200,000", A2)),

ISNUMBER(SEARCH("$200,000 to under $250,000", A2)),

ISNUMBER(SEARCH("$250,000 or more", A2)),

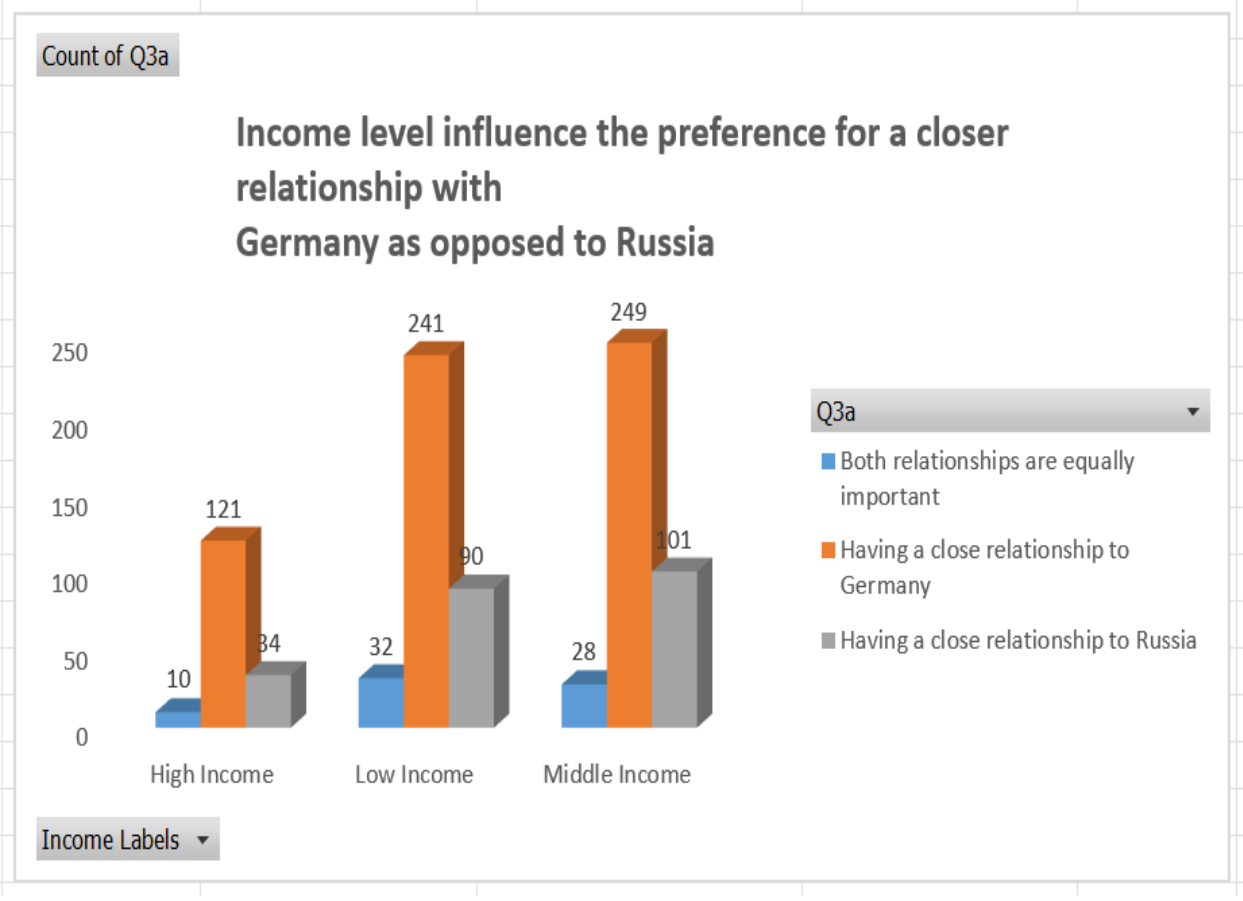
), "High Income", ""))))”.

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

Added this in the new column **“Income Labels”**.

Drawn the **Pivot table** for “Q3a” and “Income Labels” to know the count of each category. And then plotted the **Clustered Bar Chart** for the drawn Pivot table.

Count of Q3a	Column Labels			
Row Labels	Both relationships are equally important	Having a close relationship to Germany	Having a close relationship to Russia	Grand Total
High Income	10	121	34	165
Low Income	32	241	90	363
Middle Income	28	249	101	378
Grand Total	**70**	**611**	**225**	**906**



From the above plotted graph, we can infer that –

- Income level influence the preference for a closer relationship with Germany as opposed to Russia.

Part 3 : Replication

Provided picture in the question shows the comparison between the Democrats and Republicans to see Germany as partner on Key issues, which are of variables "Q5a-f". So, We have used the Seven variables for this analysis, **"partyln", "Q5a", "Q5b", "Q5c", "Q5d", "Q5e", "Q5f"**, in the data sheet **"Q3 - Replication"**. Removed the data rows which have the values as "DK/Refused", "Refused", "VOL:Neither", "Don't Know" as we only need to the data of Democrats and Republicans who see Germany as partner. Also, **missing values in the "Partyln"** are replaced with the mode i.e., **"Democratic"**. So, cleaned data helps us to give the results more effectively with proper analysis.

Drawn the **Pivot table** for each of the question columns of "Q5a-f", filtering only with the value of "Yes, as a partner" (as we need only that data to compare) with "Partyln" to know the % of each category who are the partners.

Count of Q5a. Protecting the environment	Column Labels		
Row Labels	Democratic	Republican	Grand Total
Yes, as a partner	89.55%	10.45%	100.00%
Grand Total	**89.55%**	**10.45%**	**100.00%**

Count of Q5b. Dealing with China	Column Labels		
Row Labels	Democratic	Republican	Grand Total
Yes, as a partner	68.94%	31.06%	100.00%
Grand Total	**68.94%**	**31.06%**	**100.00%**

Count of Q5c. Dealing with Iran	Column Labels		
Row Labels	Democratic	Republican	Grand Total
Yes, as a partner	66.89%	33.11%	100.00%
Grand Total	**66.89%**	**33.11%**	**100.00%**

Count of Q5d. Promoting free trade	Column Labels		
Row Labels	Democratic	Republican	Grand Total
Yes, as a partner	66.34%	33.66%	100.00%
Grand Total	**66.34%**	**33.66%**	**100.00%**

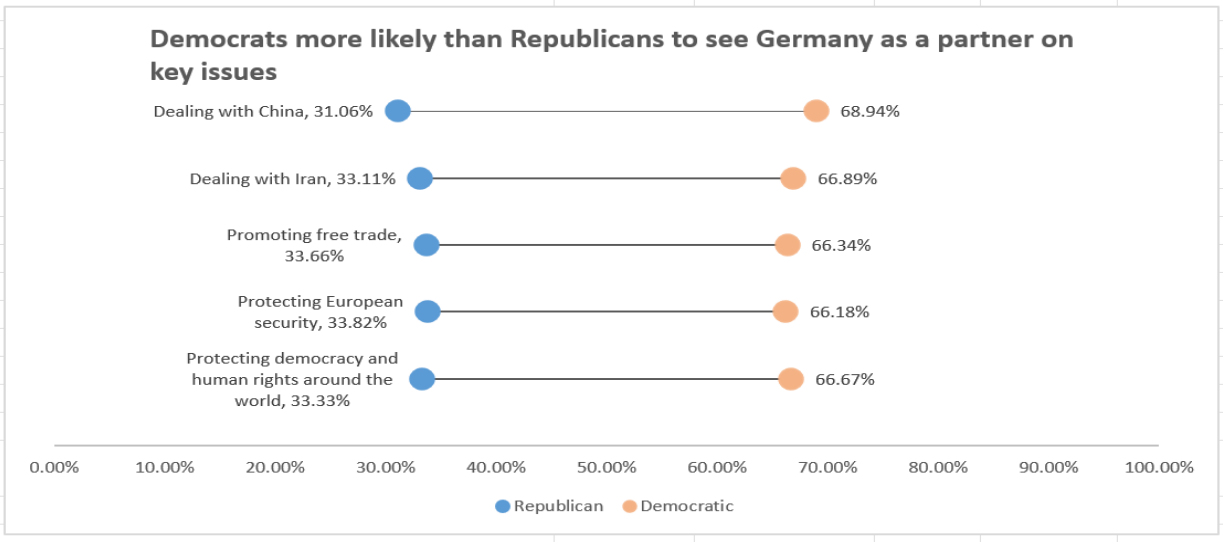
Count of Q5e. Protecting European security	Column Labels		
Row Labels	Democratic	Republican	Grand Total
Yes, as a partner	66.18%	33.82%	100.00%
Grand Total	**66.18%**	**33.82%**	**100.00%**

Count of Q5f. Protecting democracy and human rights around the world	Column Labels		
Row Labels	Democratic	Republican	Grand Total
Yes, as a partner	66.67%	33.33%	100.00%
Grand Total	**66.67%**	**33.33%**	**100.00%**

From the drawn Pivot tables results, we have created a new table with each category of Q5a-f and the % of Democratic and Republican. Also added a column “Spacing” which helps to plot the scatter plot with horizontal dumbbells. Also, added columns “Positive Differences” and “Negative Differences”, which are the values of the difference between the % of Democratics and Republicans.

Q5a-f	Democratic	Republican	Spacing	Positive Differences	Negative Differences
Protecting the environment	89.55%	10.45%	3	79.10%	0
Dealing with China	68.94%	31.06%	2.5	37.89%	0
Dealing with Iran	66.89%	33.11%	2	33.78%	0
Promoting free trade	66.34%	33.66%	1.5	32.67%	0
Protecting European security	66.18%	33.82%	1	32.37%	0
Protecting democracy and human rights around the world	66.67%	33.33%	0.5	33.33%	0

From the above table, we have plotted the **“Scatter Plot with Horizontal Dumbbells”** graph.



From the above plotted graph, we can infer that –

- Democrats are more likely than Republicans to see Germany as a partner on key issues.

Part 4 : Consulting

"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').

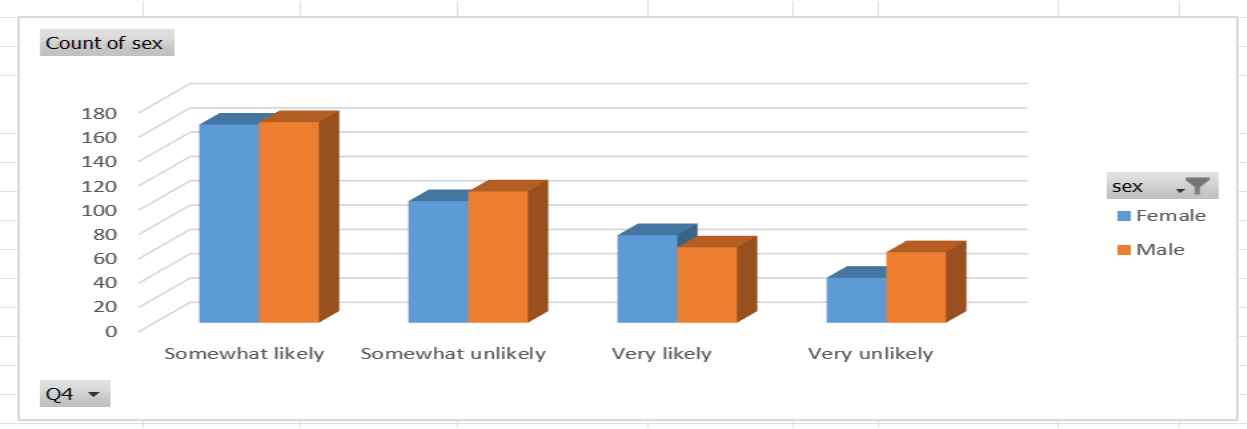
To know this, we have to analyze the responses to Q4 across various demographic factors such as gender, age, income, religion, marital status, and political views. So, We have used the Seven variables for this analysis, **"Q4", "sex", "age", "income", "religion", "mstatus", "polview"**, in the data sheet **"Q3 - Consulting"**. Removed the data rows which have the values as "DK/Refused", "Refused", "VOL:Neither", "Don't Know" as we only need to the proper and cleaned data which helps us to give the results more effectively with proper analysis.

For **"age"** and **"income"** variables, we have used same technique of grouping them into categories as in **"Part 2"** and added the columns **"age group"** and **"income group".**

We have then drawn the **Pivot Table** for **each demographic factor against "Q4"** and plotted the **Stacked Bar Chart for each** of them individually.

Q4 Vs Sex - Pivot Table & Stacked Bar Chart

Count of sex	Column Labe		
Row Labels	**Female**	**Male**	**Grand Total**
Somewhat likely	163	165	328
Somewhat unlikely	100	108	208
Very likely	72	62	134
Very unlikely	37	58	95
Grand Total	**372**	**393**	**765**

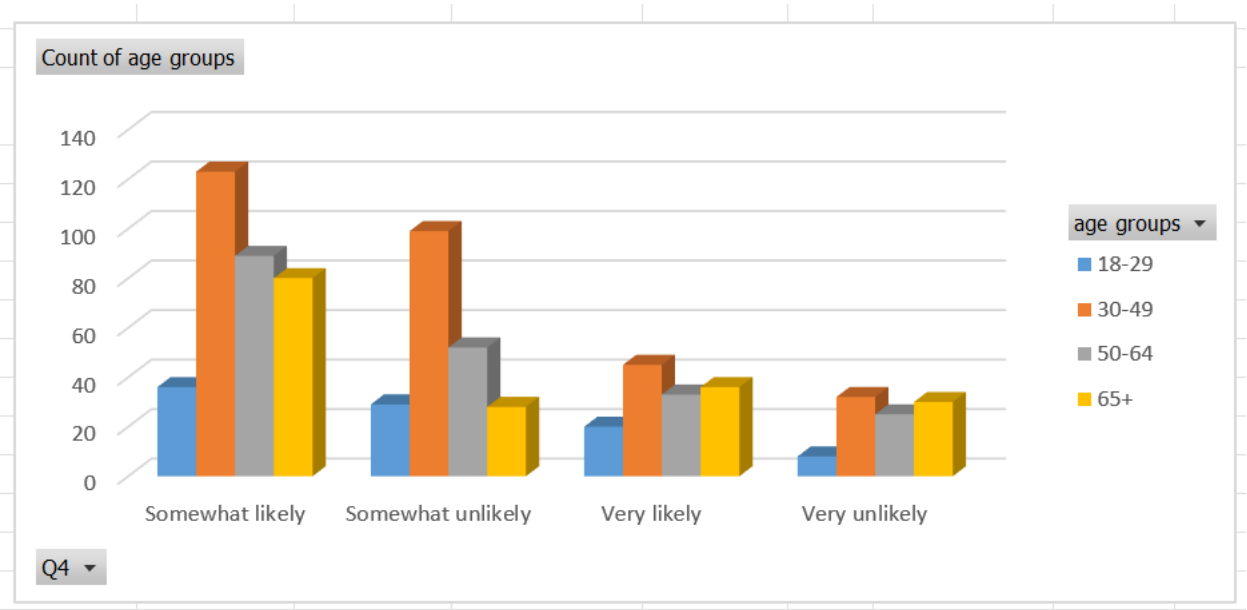


From the above plotted graph, we can infer that –

- **"Females"** feel **more likely** to the current tension between China and the United States to escalate into a situation similar to the Cold War.
- **"Sex" factor** mostly infers **"Somewhat likely"** for the situation similar to Cold war.

Q4 Vs Age - Pivot Table & Stacked Bar Chart

Count of age groups	Column Labe				
Row Labels	**18-29**	**30-49**	**50-64**	**65+**	**Grand**
Somewhat likely	36	123	89	80	328
Somewhat unlikely	29	99	52	28	208
Very likely	20	45	33	36	134
Very unlikely	8	32	25	30	95
Grand Total	**93**	**299**	**199**	**174**	**765**

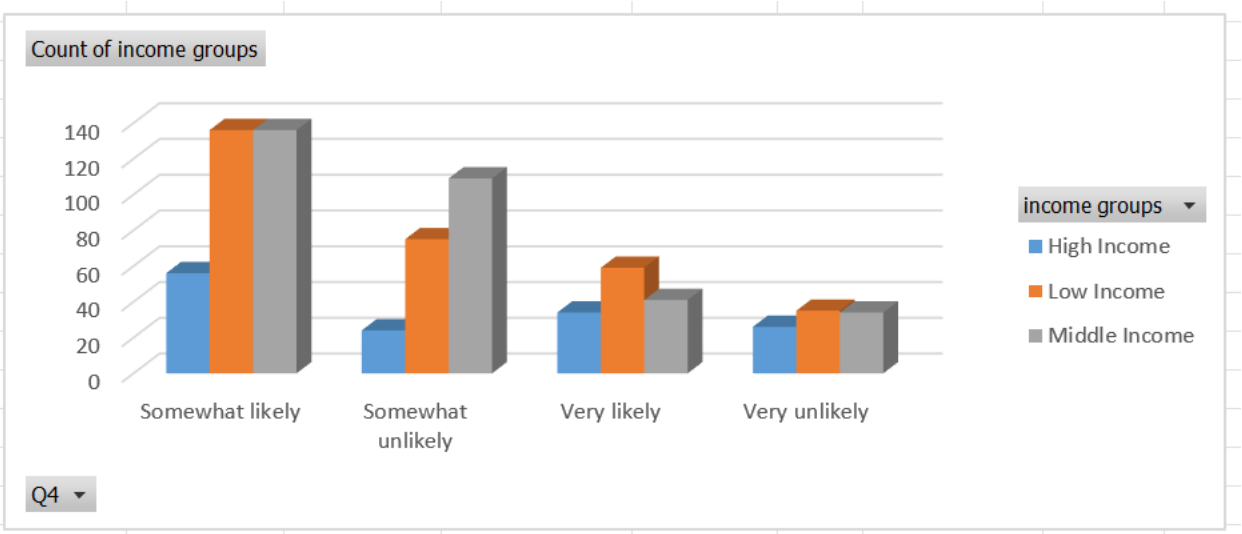


From the above plotted graph, we can infer that –

- Age group of **"30-49"** feel **more likely** to the current tension between China and the United States to escalate into a situation similar to the Cold War.
- **"Age" factor** mostly infers **"Somewhat likely"** for the situation similar to Cold war.

Q4 Vs Income - Pivot Table & Stacked Bar Chart

Count of income gr	Column Labe			
Row Labels	High Income	Low Income	Middle Income	Grand Total
Somewhat likely	56	136	136	328
Somewhat unlikely	24	75	109	208
Very likely	34	59	41	134
Very unlikely	26	35	34	95
Grand Total	**140**	**305**	**320**	**765**

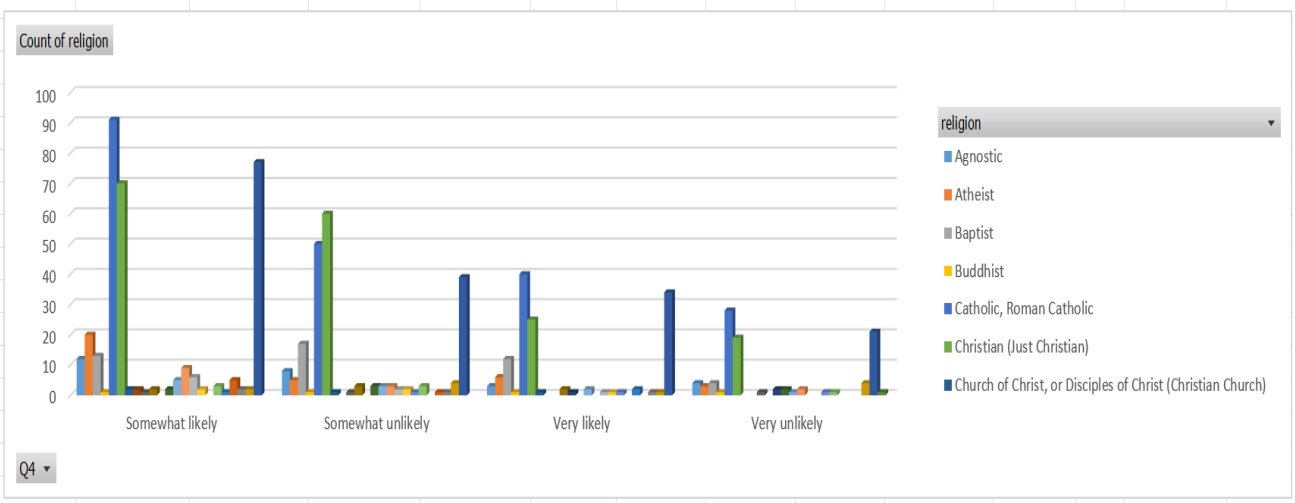


From the above plotted graph, we can infer that –

- Income group of **"Low Income"** feel **more likely** to the current tension between China and the United States to escalate into a situation similar to the Cold War.
- **"Income" factor** mostly infers **"Somewhat likely"** for the situation similar to Cold war.

Q4 Vs Religion - Pivot Table & Stacked Bar Chart

Count of religion	Column Labe													
Row Labels	Agnostic	Atheist	Baptist	Buddhist	Catholic, Roman Catholic	Christian (Just Christian)	Church of Christ, or Disciples of Christ (Christian Church)	Church of God	Episcopalian or Anglican	Evangelical	Hindu	Jehovah's Witness	Jewish/Judaism	L
Somewhat likely	12	20	13	1	91	70	2	2	1	2		2	5	
Somewhat unlikely	8	5	17	1	50	60	1		1	3		3	3	
Very likely	3	6	12	1	40	25	1			2	1		2	
Very unlikely	4	3	4	1	28	19			1		2	2	1	
Grand Total	**27**	**34**	**46**	**4**	**209**	**174**	**4**	**2**	**3**	**7**	**3**	**7**	**11**	

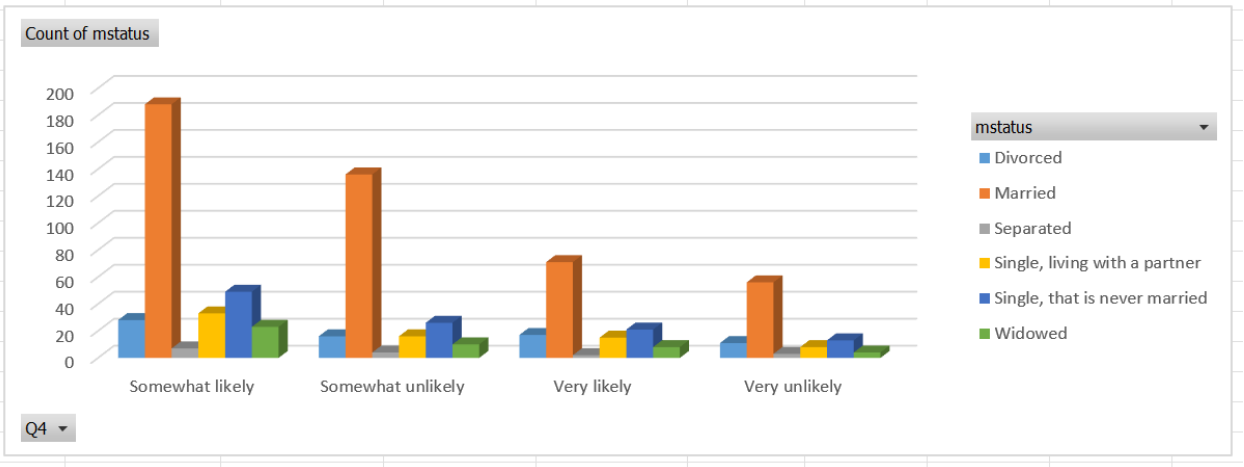


From the above plotted graph, we can infer that –

- Religion of **"Catholic, Roman Catholic"** feel **more likely** to the current tension between China and the United States to escalate into a situation similar to the Cold War.
- **"Religion" factor** mostly infers **"Somewhat likely"** for the situation similar to Cold war.

Q4 Vs Marital Status - Pivot Table & Stacked Bar Chart

Count of mstatus	Column Labe						
Row Labels	**Divorced**	**Married**	**Separated**	**Single, living with a partner**	**Single, that is never married**	**Widowed**	**Grand Total**
Somewhat likely	28	188	7	33	49	23	328
Somewhat unlikely	16	136	4	16	26	10	208
Very likely	17	71	2	15	21	8	134
Very unlikely	11	56	3	8	13	4	95
Grand Total	**72**	**451**	**16**	**72**	**109**	**45**	**765**

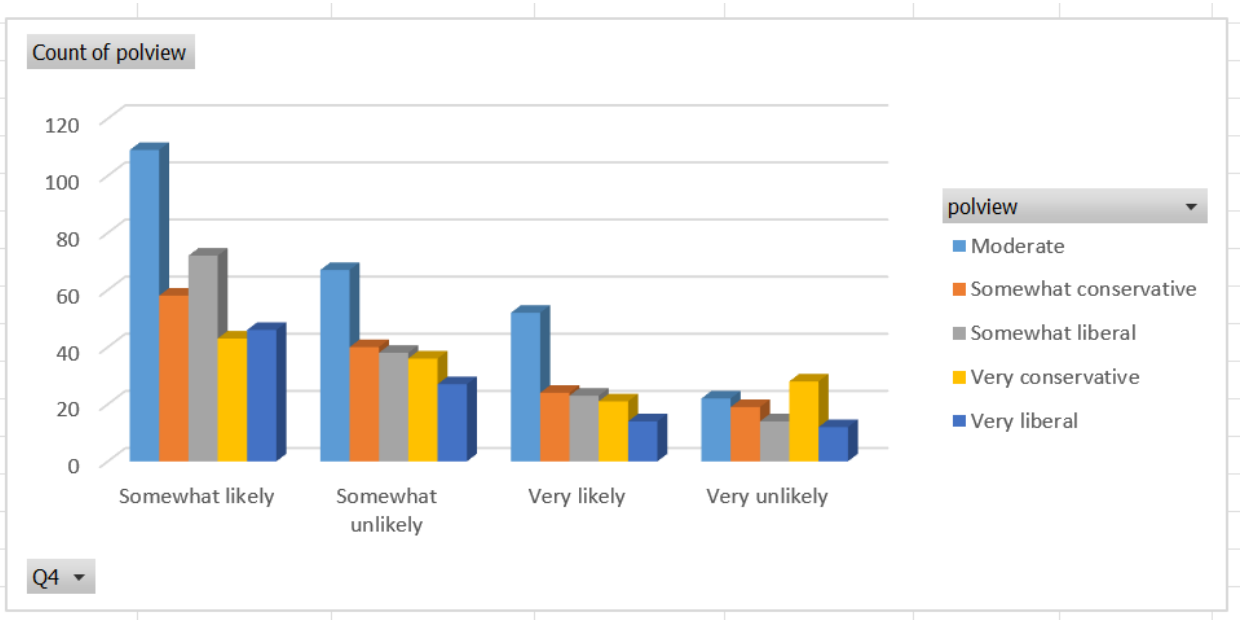


From the above plotted graph, we can infer that –

- Marital Status of **"Married"** feel **more likely** to the current tension between China and the United States to escalate into a situation similar to the Cold War.
- **"Marital Status" factor** mostly infers **"Somewhat likely"** for the situation similar to Cold war.

Q4 Vs Political View - Pivot Table & Stacked Bar Chart

Count of polview	Column Labe					
Row Labels	Moderate	Somewhat conservative	Somewhat liberal	Very conservative	Very liberal	Grand Total
Somewhat likely	109	58	72	43	46	328
Somewhat unlikely	67	40	38	36	27	208
Very likely	52	24	23	21	14	134
Very unlikely	22	19	14	28	12	95
Grand Total	**250**	**141**	**147**	**128**	**99**	**765**



From the above plotted graph, we can infer that –

- Political View of **"Moderate"** feel **more likely** to the current tension between China and the United States to escalate into a situation similar to the Cold War.
- **"Political View" factor** mostly infers **"Somewhat likely"** for the situation similar to Cold war.

From each of the demographic factors answering the Q4, by the above results, we can infer that **all of these factors feel "Somewhat likely"** to the current tension between China and the United States to escalate into a situation similar to the Cold War.

Also, **Females** of **age group 30-49**, who have the **Low Income** from the **Religion "Catholic, Roman Catholic",** who are **Married** and have the **Moderate Political View** feels **very likely** that, to the current tension between China and the United States to escalate into a situation similar to the Cold War. So, the political party should target these mentioned category people for their Facebook advertising campaign.

Below is the graph for all categories analysis together –