Beauty in the Eye of the Voter: Do voters punish female politicians more based on appearances?

Diane Carpentier, Laura Fras and Lucas Santos

June 30, 2023

Abstract

Women make up half the world's population but are far from being equal in positions of power. This study asks a key question of whether women are held to higher standards for any "mistakes" they make while in office and whether this may explain under-representation in positions of power. The paper describes a highly original experimental design that over the course of three stages aimed to test whether policies proposed by female politicians are equally popular when a man proposes them. The paper further considers whether voters are swayed to treat men and women differently based on their appearances. The study concluded that the gender of a politician plays no significant role on whether voters respond positively or negatively to a policy proposal. However, we did find that politicians that look more "likeable" are more likely to get voted into office and that male voters are more impacted by the looks as are voters who identify as "right" leaning.

"I just don't think she has a presidential look"
Donald Trump on Hilary Clinton

1 Introduction

Women to this day face severe under-representation in positions of power. Their under-representation is in of itself the result of discrimination and their continued absence from politics can lead to worse outcomes for other minorities. Women leaders have been found to be more qualified in a politician context and lead to better outcomes for constituents, however once they have proven themselves in politics they do not face the same re-election probabilities as men. This paper aims to shed light on voter bias regarding political policies to understand whether female politicians are held to higher standards than men for "mistakes" or "bad policies". Moreover, this paper further analyses whether voters are influenced by the "looks" of politicians and if there is a greater gendered effect of looks on how "forgiving" voters are to politicians.

Women have been found to be of higher leadership quality, however even once "proving" themselves in power they still do not face equal footing when running for re-election. One possible explanation could be that women are held to higher standards for any mistakes they make while in office. However, researchers have struggled in general to disentangle why women face unequal re-election probabilities from the higher quality of female leadership. Within the literature review we will analyse the motivation behind women in power and consider how previous experiments have tackled these questions.

To answer our key question we devised an original experiment which 129 people partook in. The experiment consisted of two treatments and three stages. We tested whether respondents equally agreed with two "good" policies and two "bad" policies based on gender and looks of the politician proposing such policies. We secondly replicated existing papers by testing how likely respondents were to vote for politicians based on just their appearances. Finally we asked respondents to rank how likeable they found each face. Using this method we analysed whether looks mattered and whether there was any gender difference in results.

Overall, we found that gender played no part in how likely voters were to agree with a policy. However, we did find that male voters were more likely to score higher agreement to more "likeable" politicians when a bad policy was proposed. We also found that looks played very little impact on voters' agreement to policies. However, we did find that more "likeable" politicians were more likely to gain votes when no other information was presented.

This paper is structured in three sections: first we will consider existing literature on women in power, before providing an overview of the experiment design, and then we will discuss our findings.

2 Literature Review

The literature review will consider why we would expect higher re-election rates for women, before investigating what previous experiments have found in regard to voter opinion on female politicians. Moreover, it will further analyse how previous experiments have measured the impact of politician's "looks" on their political outcomes. The existing literature provides both the motivation for our experiment and insight into how best to test our hypotheses.

2.1 Women in Power

Women leaders have better outcomes for voters. They have been found to decrease crime, especially gender based crime Delaporte and Pino (2022), lead to better outcomes for health and especially female health (Chattopadhyay & Duflo, 2004). Moreover, they have been found to be less corrupt (Brollo & Troiano, 2016) and very effective at passing legislation (Volden, Wiseman, & Wittmer, 2013). However, these successes are in part due to self-selection. Because women face more discrimination from voters and the media alike, only the very best women succeed in politics (Piscopo, Hinojosa, Thomas, & Siavelis, 2022). Nonetheless, if voters behaved in a consistent manner, once they have witnessed the successes of female leaders they should re-elect them at a higher rate than their male counterparts. This has not been the case. The literature on re-election rates for women has been mixed with papers from (Eggers, Vivyan, & Wagner, 2018; Sevi, 2023) finding that they face lower re-election rates.

One possible factor behind this effect could be that women leaders are held to higher standards. This effect has been seen in a range of professions, from education (Mengel, Sauermann, & Zölitz, 2019), medicine (Sarsons, 2017), academic economics (Hengel, 2022) and even politics (Bauer, 2020). This paper aims to uncover whether women are held to higher standards when mistakes are made in a political setting to provide further insight into why women are not rewarded when running for re-election. An experiment provides the perfect environment to test whether female politicians are held to higher standards, as in empirical work on real world data it can be almost impossible to disentangle the higher performance of women from any voter biases.

2.2 Experiments on gender in politics

Most of the existing literature from experimental economics on women regards voters' opinions of women leaders. Many of the experiments have found that voters associate women and men with different characteristics and judge competency in a gendered fashion. In the seminal paper by Huddy and Terkildsen (1993) it was found that voters have a different perception of men and women, where women are perceived as being compassionate and kind. Therefore, voters see them as competent when they showcase those qualities in positions such as healthcare and education. While for men, they are seen by voters as more rational and that they have higher competence on topics such as military activities and the economy. More recently Barnes and Beaulieu (2019) showed that respondents perceived women as more risk averse and voters also thought that women face more discrimination and were therefore less likely to be associated with corruption. However, in the same paper they found a gender difference in survey respondents, where female voters thought women faced more adversity in politics while men just thought women were more honest and therefore were less likely to be corrupt.

Not only have previous studies shown that voters associate politicians with different strengths based on their gender, this extends to perceived quality of governance. While voters associate women with being less corrupt, Eggers et al. (2018) showed that female voters in particular are far less likely to vote for women leaders in an experimental setting if they have been found of misconduct such as corruption, but are more likely to reward high quality women politicians. This may well be that female voters understand the high barriers to entry for female politicians, so are more disappointed when such leaders are found of misconduct. These findings are corroborated by Bauer (2020) who also found in an experimental setting that female candidates are held to higher standards, as they are expected to have higher

qualifications by voters of all genders.

All of the previous studies show in an experiment setting that women and men are not seen as equal. Women are associated with "softer" qualities than men and when they make mistakes they appear to be held to higher standards by voters. While these papers do provide insight to the barriers faced by women they do not fully explain why women do not face the same re-election rates once they have proven themselves to be competent. Therefore, this paper adds to these findings by showing whether women and men are treated equally when they announce a "good" policy and a "bad policy".

2.3 Politician's looks

Voters form opinions on politicians based on their party, policies and promises. However, non-policy related factors also hold plenty of weight, from the voice of a politician to their non-verbal cues (Haumer & Donsbach, 2009; Zoghaib, 2019). Amongst these non-policy related factors are the looks of a politician. Factors such as how attractive they are, or whether they "look like a leader" or appear competent can all play a major role in how likely a politician is to be elected (Antonakis & Dalgas, 2009; Berggren, Jordahl, & Poutvaara, 2010; Lawson, Lenz, Baker, & Myers, 2010). The importance of "looks" is heavily correlated with the educational level of voters. Uneducated voters who are not heavily involved in politics seem more swayed by politician's "looks" (Johns & Shephard, 2007). The same effect is seen in local elections where voters may have less information on each politician (Rosar, Klein, & Beckers, 2008). Within the context of gender, we aim to investigate whether these visual aspects play an equal role in how accepting voters are of "good" and "bad" policies.
So far the gender difference in politicians' looks has focused on more electability than policy support. Chiao, Bowman, and Gill (2008) showed that women who were rated as attractive were more likely to gain votes, unlike men were only "approachability" mattered. However, while attractiveness matters for female politicians, there does not appear to be a need to "objectify" oneself to win positions of power. When Gothreau, Alvarez, and Friesen (2022) showed voters pictures of either "serious" or "objectified" female politicians there was not a big difference in their chances of getting elected. Where this paper adds to the literature is by analysing whether the effects of appearances for women extend past the initial vote and into policy proposals.

2.4 Lessons from Literature

This paper aims to uncover the relationship between policy support and gender with a consideration of how the appearance of politicians can impact voter reaction. As part of our analysis we will develop a method of measuring how "likeable" politicians are. Our methodology is heavily influenced by the experiment design of previous studies.

Most of the experiments that have been discussed use real life politicians and consider their probability of getting elected. This is an incredibly difficult task to undertake as most voters will have pre-existing opinions on these candidates which will influence how they rate their appearance. One of the ways this bias has been accounted for has been by using black and white pictures with plain backgrounds to not distract voters (Lawson et al., 2010; Todorov, Mandisodza, Goren, & Hall, 2005). Moreover, by showing the pictures in a very short space of time (such as in 0.75 of a second) it does not allow respondents to over analyse the pictures presented to them.

To decrease pre-existing opinions from biassing results, some papers have asked foreign na-

tionals to rate politicians appearances (Lawson et al., 2010). Alternatively, some researchers have asked children whether they think a picture of a politician would be a good ship captain, as most children are not politically active (Antonakis & Dalgas, 2009). While there are lessons to be learnt on how to undertake a facial rating exercise from all of the previous papers they can only take our analysis so far. Nearly all of the papers used real life politicians with a sample of respondents that are all citizens of the same country. Our experiment does not have the benefit of a homogeneous nationality of respondents due to its international outreach which will be discussed in the limitations, however the findings of previous studies still motivates our research and does inform our experiment design.

We are adding to all of the previous literature by asking a key question of whether the disappointing re-election results seen by women is due to voters holding women to higher standards and whether their appearance matters in such a decision. By analysing whether voters equally judge policy proposals, which is what an incumbent politician should be judged on, we can gather whether voters holding women to higher standards could explain why they do not have equal or better re-election results. Moreover, by analysing the role of appearances, we provide more evidence on voter biases.

3 Design of Experiment

We designed our experiment to answer three questions on the role of the physical aspect and gender in politics. Firstly we wanted to see whether women are more penalised than men for bad policy decisions. Secondly, if physical aspect matters for the voters' decision making, and if there is a gender differential in how much looks affect politicians.In order to respond to these questions, our experimental design made respondents rate how much they agreed with objectively agreeable and objectively controversial policies presented by a woman or a man politicians with or without a picture.

Our first step was to find policies that were "good" and "bad" with common consensus. Once we had chosen appropriate policies, we ran our main experiment in 3-stages. Firstly, we asked respondents to rate four policies. Each of these policies was either proposed by a man or a woman, and in half our sample respondents were shown pictures of the politician proposing the policy. Secondly, the participants were asked to vote for politicians only based on their picture and finally we asked respondents to rank the "likeability" of politicians".

3.1 First Step: Survey for Policy Choices

We ran a first survey to decide which four policies (2 good and 2 bad) to include in our main experiment. We used ChatGPT to generate a list of possible policies. We asked: "Please generate a list of "controversial or bad" policies that are unlikely to appeal to voters based on real life examples, and another list of "agreeable or good" policies that are likely to appeal to voters based on real life examples. Write it in the style of a political brochure."[1]

After slightly adjusting the description of the policies to sound more authentic, we ran a survey asking people to rate it from 1 to 10 (don't agree to strongly agree) and justify what influenced their decision. Examples of each type of policy can be found in the appendix.

[1]As a sense check we also asked Chat GPT to consider policies that would be controversial for voters of different political leaning, and were feasible with the wider political environment e.g. it proposed abolishing public sector unions as a "bad" policy for left wing voters.

Bad Policy	**Good Policy**
Implementation of Vegan School Meals	Introducing Universal Basic Income (UBI)
Ensuring Controlled Borders for Foreign Migrants	Clean Air Initiative
Abolishing Public Sector Unions	Implementation of the Mental Health Initiative
Lowering the Age of Consent to 13	Implementation of the Pornography Free Environment
Comprehensive Drug Legalisation	More Funding for Sports in School
Higher Taxes on Violent Video Games	

After running our survey for a few days we gained 13 respondents; out of which 5 self-identified as "left leaning", 4 as "right leaning", and 4 as "centrist". We selected 5 policies in total (2 good and 3 bad) in function of the highest rating for the good ones, and of the lowest rating for the bad ones. The survey found that regardless of political affiliation, "lowering the age of consent to 13" was a bad policy (average rate between 1.8 and 3 over 10 (see 3.1) and that "clean air initiative" (average rate between 7.75 and 9.6) and "mental health" (average rate between 7.25 and 8.75) policies are objectively good policies.

However, for a second controversial policy, there was no consensus among our sample. Therefore, we assigned the second bad policy in function of the respondents' political leanings: "drug legalisation" for right and center leanings (rated 2.25 and 3.50 over 10), and "taxes on video games" for the left leaning (2.8 over 10).

Policy	RIGHT	CENTER	LEFT
Vegan Meals	3.75 (1.89)	8.25 (1.71)	4.4 (1.51)
Borders Control	6.75 (1.89)	6.25 (3.30)	3.6 (3.44)
Abolishing Unions	5.50 (4.66)	3.00 (2.71)	2.2 (1.79)
Lowering Age of Consent	**2.50 (1.91)**	**3.00 (2.30)**	**1.8 (1.79)**
Drug Legalisation	**2.25 (2.50)**	**3.50 (2.38)**	7.0 (2.92)
Taxes Video Games	5.25 (3.86)	4.75 (3.77)	**2.8 (1.64)**
Universal Basic Income	3.25 (1.71)	3.50 (2.38)	5.2 (2.28)
Clean Air Initiative	**7.75 (0.96)**	**8.00 (1.63)**	**9.6 (0.55)**
Mental Health	**7.25 (3.40)**	**8.75 (1.50)**	**8.2 (0.84)**
Child Safety	8.75 (1.89)	9.00 (1.15)	8.0 (3.93)
Sports Funding	7.25 (2.50)	8.00 (2.71)	7.8 (0.45)

Table 1: Average Rates of Policies by Political Leaning

.

Note: the numbers in parenthesis represent the standard deviation from the mean

3.2 Second Step: Final Experiment Design

Once the policies were selected, we designed our experiment to answer the three questions on gender and physical aspects in Politics: (1) if women are more penalised than men for bad policy decisions, (2) if physical aspect matters for the voters' decision making, and (3) if there is a gender differential in how much looks affect politicians. We first ran a pilot in the laboratory with the experimental class i.e among 22 participants. Based on the comments, we slightly adjusted it and ran it with 106 participants.

The survey consists in 3 stages:
- 1st stage: Rating four policies proposed by a man or a woman and with or without a picture
- 2nd stage: Voting for politicians with no information but looks
- 3rd stage: Ranking "likeability" of politicians

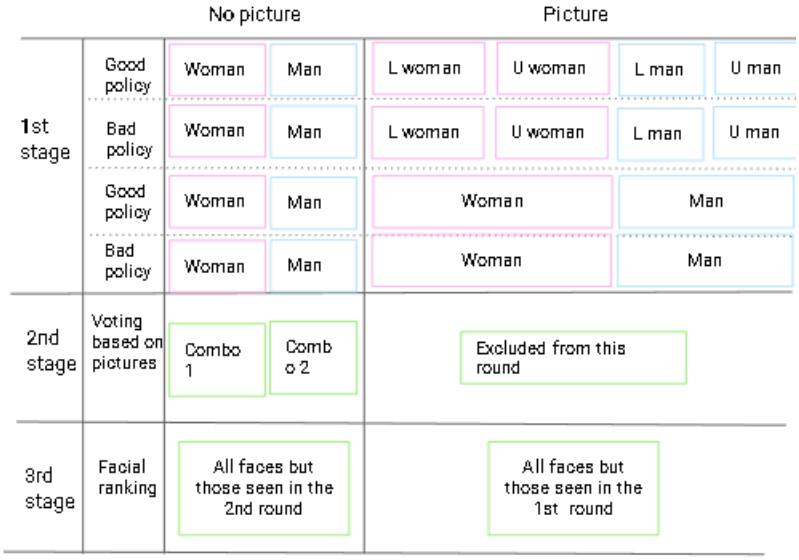

Figure 1: Diagram of the experiment setup stage by stage

.

Note: The respondents were randomly assigned to no picture path or picture path. Within each path respondents were randomly assigned to a different option in the survey flow. However, respondents could not leave their path. As an example of a path for the 1st stage if you were randomly placed in the picture path, you could randomly be assigned to Unlikeable Women (U Women) first, then Likeable Man (L Man), a Woman for the third policy and finally another Woman for the fourth. Only people placed in the no picture path partook in stage 2.

3.2.1 First stage: Rating four policies

We constructed this stage aiming to look at the effects of gender and of the physical aspect of politicians on the voters' decision.

In order to control for attractiveness we presented the policies with a short description and randomised whether participants would be shown pictures alongside the policies (treatment group) or without (control group) a picture.

For each policy, we give a fictitious name to the politicians using AI. We asked ChatGPT to generate a list of 10 most common last names and first names in Central Europe. As the last names proposed were too diverse[2], we decided to focus on a Spanish setting. As the survey respondents were highly international we were worried that the voters would react more positively to names that sound familiar which would skew our results. Therefore, we asked again for "soft" Spanish connotated last names to have less variation within the politicians.

Half of the participants were given pictures in the first round. To pick the pictures, we followed the existing literature and only included politicians of one country. We picked Spain as the names we selected were "softly" Spanish. We used pictures of Spanish Mayors that were selected from the database of electoral rolls of Spanish mayors [3]. We looked through over 300 pictures to select 2 likeable men and women, 2 normal looking men and women,

[2]Some names were too stereoptical German, English or Mediterranean

[3]Provided by the Spanish government Open data initiative of the Government of Spain (2023)

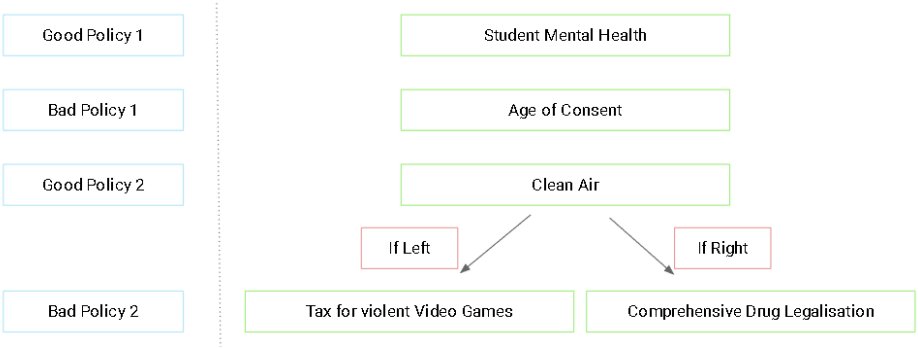

Figure 2: Diagram of the policies' attribution

2 unlikeable men and women (i.e 12 in total). Then, we associated them with the names generated by AI. In line with the existing literature we only considered pictures with a plain background, and turned all the pictures black and white.

For the first "good" and "bad" policies, those respondents in the "picture group" were randomly assigned either to a likeable man or woman, or to an unlikeable man or woman (4 different scenarios per policy for the participants with pictures) . For the second "good" and "bad" policies, the participants who received a picture were randomly assigned to a "normal" looking male or female politician (2 different scenarios per policy for the participants with pictures). This is as we did not have strong consensus in the first survey on the second "bad" policy and we did not want to decrease the sample size[4] for a policy that we were less certain would be unpopular.

To sum up the first stage, as it is represented figure1 every participant was first randomly assigned to either pictures path or just text path (50% each). Then, they had to rate successively 4 policies (1 good, 1 bad, 1 good, 1 bad) from 1 to 10 (don't agree to strongly agree)[5]. Note that the pictures shown for the tax on video games and comprehensive drug legalisation were the same (i.e either they received a "normal" looking male or female), as these two policies would never be presented to the same person[6].

3.2.2 Second Stage: Voting for Politicians

In this stage, we presented a panel of 6 pictures of politicians (1 of each type i.e 1 likeable woman and man, 1 normal woman and man, and 1 unlikeable woman and man) and asked people to pick their first and second favourites. The voting round was only available to participants who were in the plain text group in the first round [7]. The choice of who to vote for is **only based on the picture** of the politicians and the participants had no additional information on them. This step aimed to demonstrate the preconception idea that good looking people will receive more votes, in line with wider literature.

[4]By splitting the sample further into different pictures.

[5]These policies were: student mental health, age of consent, clean air and in function of their political leaning self-identification either tax on video games (left leaning) or drug legalisation (right and center leanings) (according to the first survey on policies choices)

[6]We also note that at this stage the only modification, based on the feedback of the Experimental class (pilot), we made was to add more spaces in the description to make it more clear and correct some typos.

[7]Note that we included this stage for all respondents in the original setting (pilot), but based on the comment of the Experimental class who said that they saw faces twice, we removed this step for the individuals who got pictures in stage 1.

(a) Panel A

(b) Panel B

Figure 3

We created 2 sets of 6 pictures representing the 12 candidates (1 of each of the 3 types and gender), where respondents were randomly shown one of the two sets.
The two groups allowed for the respondents to rate the faces in the third round that were not presented to them in the second round.

3.2.3 Third Stage: Rating Faces

Finally, we wanted to make sure that our classification of likeability (likeable - normal - unlikeable) was accurate and in line with existing literature. We presented 6 pictures successively, asking people to rate how likeable the candidate was on a scale of 1 to 10 ('not at all' to 'very').

To capture respondents' honest reactions, participants had only 5 seconds to rate the picture. After this time, the computer automatically switched to the next picture. Every participant had to do a training including 3 different new pictures (also of mayors) to get familiar with the speed. We made sure that people receive 6 pictures that they haven't seen before [8] in order to not influence their rating because of a familiarity effect. In addition, the pictures were randomly ordered in case there was a learning curve of how to answer within a 5 second window. The randomisation was included after feedback from the in-class pilot, and two more training rounds were also included. Finally, we also included questions on Demographics and Feedback.

4 Data

Using the responses to our survey we comprised a dataset on which we ran our analysis. In this section we provide an overview of our sample. Our sample consists of two sets of results, the first were generated in class and the second were collected within a two week window between the 8th of June and the 22nd. Moreover, we will describe how we incorporated the two samples into a singular dataset.

In total we have **128 observations** from both surveys, 22 from the survey we ran in-class and 106 from the final survey after the feedback, the distribution can be seen in figure 12. The figure 11[9] shows that our sample is balanced between respondents that were allocated

[8] Either in the first stage for the participants with policies with pictures and in stage 2 for the other ones

[9] All the graphs are presented using **viridis**, a colour palette that improve graph readability for colorblind readers since it cover a wide perceptual range in brightness and blue-yellow, and do not rely as much on red-green contrast.

in the picture group and respondents in the picture and text based pathways[10]. Moreover, our sample had a relatively even gender split of 44% of respondents identifying as women. Furthermore, as seen in figure 13 most of our respondents politically self identify themselves as left leaning, and as described in 14 were mostly 22 and 29 years old.

To generate our dataset we merged the results from the in-class survey and the online survey. From the inclass survey we gathered feedback on what went well and what could be improved, using which we changed the structure of the final survey. In the original experiment all respondents were shown pictures to vote on in round 2, which led to respondents having seen some faces twice in the round 3 where they were ranking likeability. Therefore, we excluded the ranking scores from the in-class survey and maintained the results from round 1: Policy ranking and round 2: No context voting. Therefore, in our final dataset the likeability rankings were purely comprised from the online respondents from which we could guarantee that faces were not observed twice by the same respondent.

5 Results

5.1 Round 1 - Results

The first round of the experiment we aimed to measure whether there were gender differences in how much voters liked policy suggestions and secondly whether 'looks' also played a role.

We first found that there is minimal difference in voter agreement based on the gender of the politician suggesting the policy. As seen in figures 5 and 6 the number of votes for policies proposed by men and women appear equally. This pattern appears relatively equally across all six policy proposals as seen in the appendix. Moreover, there does not appear to be any significant differences between the role of "looks" that were pre-picked by the authors and policy agreement. In fact from visually looking at the tables it appears that those assigned "not good looking" by the researchers actually had higher policy agreement from voters, however this is not significant.

5.2 Round 2: Results

The second round of the experiment we were following examples from existing literature that look at the effect of "looks" on votability more so that policy agreement. To determine which politician received the most votes we counted the total number of times they received a vote for 1st option and the number of times they received a vote for being the second option.

Our results interestingly showed that women tended to gain more votes than men. The highest voted 1st choice candidate was the assigned "beautiful" women who represented the policy of age of consent with 7 votes. Closely followed by the women who represented clear air and the "beautiful" man who represented mental health, who both received 6 votes. While for the second vote the woman who represented violent video games and the legalisation of all drugs received the highest number of votes at 8 votes. The full overview of the total number of votes reviewed by each individual is summarised in the appendix.

Our results do not contradict previous findings significantly. We did not find that gender had an affect in the voting round and it is unclear whether politicians who look more

[10]Meaning half of our sample did not partake in the 2nd voting round (non-picture group).

politician	Policy agreement	2nd round votes	Likability
Air Quality - Man	8.13	2.0	5.07
Air Quality - Woman	8.43	6.0	6.44
Consent- Likeable Man	1.93	5.0	5.96
Consent- Likeable Woman	2.12	7.0	6.18
Consent- Unlikeable Man	2.23	1.0	5.63
Consent- Unlikeable Woman	2.73	1.0	4.58
Mental Health- Likeable Man	7.19	6.0	5.82
Mental Health- Likeable Woman	7.62	5.0	6.01
Mental Health- Unlikeable Man	8.40	2.0	3.90
Mental Health- Unlikeable Woman	8.82	5.0	5.32
Video Game and Drugs - Man	4.55	2.0	5.72
Video Game and Drugs - Woman	4.76	3.0	5.51

Table 2: Summary of Individual Politicians Outcomes

"likeable" are more likely to get voted in either the first or second round.

5.3 Round 3: Results

The third round consisted of respondents ranking how "likeable" they thought each image of a politician was. We included three practice rounds for respondents to familiarise themselves with the ranking process. Moreover, by randomising the order that respondents were shown the images we maintained consistent reaction time and minimal missing answers.

The results of the likeability ranking are summarised below in table 2, alongside the average score for their respective policy. The candidates that were initially picked by the researchers as more likeable did indeed score more highly than those deemed unlikable by the researchers. Therefore, our findings from round 1, which showed a slight policy advantage to those politicians assigned "unlikable" by the researchers did indeed hold for voters too. Overall, our final stage provided valid rankings of how "likeable" politicians are to voters.

5.4 Key Results

We did not find that the gender of a politician influenced how much voters agreed with the policies proposed by politicians. However, by analysing our results using other statistical methods and respondent characteristics we were able to find interesting observations.

Firstly we did not find statistically significant proof that voters were influenced by the better or worse looking politicians, these results are further solidified by the findings of our linear regressions. We ran a linear regression on average policy agreeableness and average "likeability" score. In figure 4 we can see that there is no clear pattern of policy agreement by likeability and even when we run linear regressions within each policy grouping (e.g. for just the politicians who represented the mental health policy) we see no significant effects. The lack of significant findings could be the result of a limited sample size and a limited number of politicians that ran under each policy.

However, when we analysed the responses to the policies by the gender of the respondents we did find that men were more reactive to politicians' looks. When presented with good policies men tended to give a lower agreement score to politicians who were looked less "like-

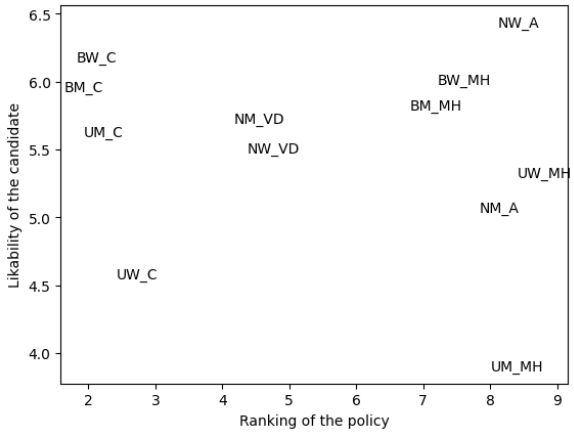

Figure 4
Scatter plot of Politicians Likeability on Policy Agreeableness
Note: The letters identify each politician. For example, BW_C means beautiful women for consent, while UM_C is ugly man consent. The first two letters represent whether they are beautiful and their gender, while N stands for normal. The codes ending in _VD are for the pictures of the politicians representing video games and drug legalization, while _MH stands for mental health and _A stands for clean air.

able" while for women voters their opinions did not drastically change. we did not see any differences between male and female voters on the role of appearances for male and female politicians. When presented with bad policies, both men and women had a similar pattern of having lower support for the policy if the politician proposing it was less visually "likeable". When running a binomial test for male respondents on the probability of equally voting for attractive and unattractive politicians we were able to significantly show that men were less likely to vote for unattractive politicians, while women also were significantly less likely to vote for unattractive politicians but the p-value was greater for women. It is very interesting that this effect is not seen as strongly under the proposal of the bad policy, but it may be the case that we have limited sample size and as the negative policies were so unpopular there is not much variation to capture the effect of visual characteristics of the proposing politician. However, if in a bigger sample male voters continue to display this effect, it could be the case that in areas where men are more likely to vote that "bad" policies are less punished by the electorate if the politician proposing them is more attractive.

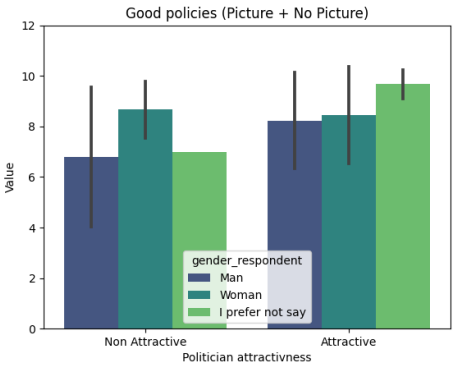

Figure 5: Results Good Policy by Gender

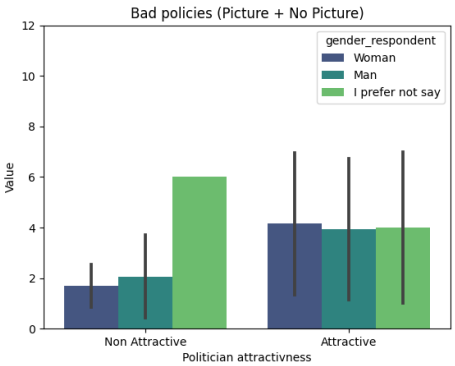

Figure 6: Results Bad Policy by Gender

Similarly if we analyse how voters of different political leanings vote it appears that voters who identify as "right" leaning are more impacted by the visual characteristics for the two

bad policy options. Interestingly, it appears that "left" leaning voters are more likely to give higher approval ratings to policies than right leaning or centrist voters, however they are more consistent regardless of looks. It could be the case that "left" leaning voters have considered these policies before and therefore are less impacted by the politician proposing the policy. Nonetheless, none of the differences are not significant at a 5% point level. These results may be driven by small sample size, especially as we have fewer voters who self identified as "right" leaning.

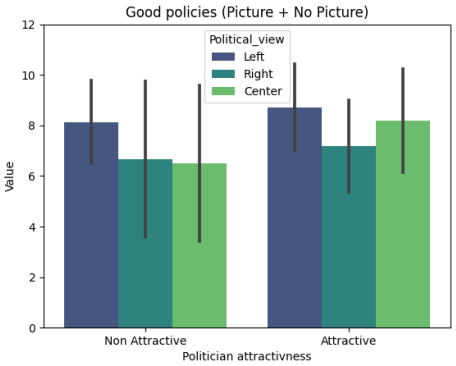

Figure 7: Results Good Policy by Political Leaning

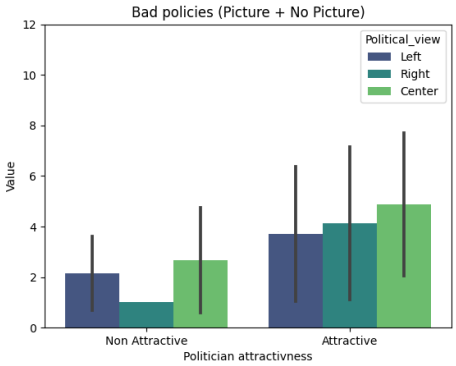

Figure 8: Results Bad Policy by political leaning

Overall, it does not appear that women are more harshly judged for poor policy decisions than men. Similarly the effect of "likeability" of politicians also seems to have negligible effect on whether voters agree with proposed policies and on how likely they are to vote for a politician. However, there does appear to be a slight effect among male voters on how much they agree with a policy based on the looks of a politician, as does the political leaning of voters. However, to answer our key question of whether women are held to higher standards and "punished" more by voters for bad policy decisions, this does not appear to be the driving force behind why women face harsher re-election chances, and further analysis may be needed to understand this phenomenon.

6 Feedback and Limitations

6.1 Feedback

The feedback received from respondents was crucial in planning and modifying our experiment and in providing explanations as to what motivated voting choices. In the first in-class experiment we collected a range of important feedback, including; adding more practice pictures for the third facing ranking round and making sure that in the third voting round we do not include the same pictures seen previously in the experiment. From this feedback we re-structured the survey for the online sample so that only the respondents who did not see pictures in the first round were given pictures to vote for in the second round and all respondents ranked faces that they did not see. This feedback was crucial in making sure that for our larger sample we did not experience any Hawthorn effects.

More broadly we also received positive feedback on the final version of the experiment. In figure 9 we generated a word cloud of most common responses from the experiment. Among

them, are the words *Interesting*, *Enjoyed*, *Dynamic*. Indicating that our survey was widely positively received and may contribute to why we were able to gain a over 100 respondents via word of mouth. We did not receive much negative feedback. To test the sentiment of the feedback we used Python's Natural Language Processing analysis to determine whether responses were positive or negative. Positive responses are coded as +1 and negative responses as -1, in figure 10 we can see that most feedback gained a score of 0 showing neutrality with some positive responses and only a few had negative responses. The negative feedback mostly relates to some respondents "guessing" what the topic of the survey was.

Figure 9: Wordcloud of the feedbacks

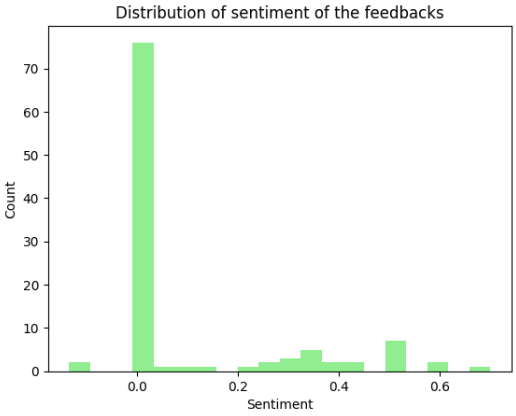

Figure 10: Distribution of the sentiment of the feedbacks

6.2 Limitations and Lessons Learnt

The experiment was modified to best reflect feedback from respondents, there were still some limitations that impacted how effective the experiment was at answering our key question on whether female politicians are held to higher standards and are more impacted by their appearances.

Firstly, while we did gain over 120 respondents, we still have a small sample size per policy option due to the higher number of randomizations in each round (e.g. 6 different treatments for "good" policy one). Therefore, we only had 15 number of respondents for each policy with a picture. This is more amplified for the "bad" policy option 2, where respondents were further split into "left" and "right" leaning policies, especially as we had a smaller share of voters who identified as "right" leaning. Our sample also heavily consisted from students from BSE who are likely to be more politically engaged, which could explain why the voter's we're less impacted by the appearances of politicians as in the paper by Johns and Shephard (2007) it would found that poorly informed voters tend to be more impacted by how attractive politics are. Therefore, for any future researchers we would recommend increasing the sample size or decreasing the amount of randomisation at each level and aim to have a more politically balances sample.

Another limitation we faced was for the second "bad" policy option. From the first policy survey that we ran to pick unpopular policy options, we were unable to reach a consensus with our respondents on a second unpopular opinion. Therefore, having two bad policies that differed along political lines not only decreased our sample size but also required respondents to identify politically which could leave some respondents feeling uncomfortable especially if

they do not identify themselves along a left/right scale.

Our main limitation is that we did not find any gendered effect on policy choice. This may be due to female politicians being treated equally to male politicians on policy matters, however it could also be that by focusing on the visually signalling gender and using gendered names that voters' internal biases were not as triggered. In the paper by Haumer and Donsbach (2009) they found that many voters are strongly impacted by the voice of a politician. Therefore, it may not be the case that female incumbents face challenges getting re-elected due to harsher reaction to bad policies but more that voters dislike female mannerisms such as voice that they hear over the course of a political term. Thus, future researchers should try to include additional factors such as voice and mannerism when attempting to measure voter bias to female politicians.

The final experiment that was conducted was highly original and while building on existing literature measured the gendered effect of policies which is yet to be done. Moreover, by utilising an international sample of respondents we had to be incredibly creative in experiment design and choice of politicians names, faces and policies. However, while we learnt a lot from the iterative process of creating the experiment, one key limitation is that we are unable to compare our results to existing literature. As most literature used real politicians and their voting scores to determine factors such as the impact of looks we are unable to compare whether our second round of votes is in line with existing literature. We recommend future research to utilise different measures of measuring facial attractiveness or likeability to best compare with the wider body of literature.

7 Conclusion

Women make up half the world's population but are still heavily underrepresented in politics which continues even once women have won their first seats in power. This paper aimed to investigate whether the reason women do not face higher reelection rates in politics even though they out-perfom male leaders is due to women being held to higher standards for any mistakes. Moreover, by investigating the role that physical appearance plays in gender differences between male and female voters we aimed to uncover if voters are more impacted by the appearances of female politicians.

Our experiment design built on existing literature and consisted of three experiments; an experiment to determine which policies to utilise, an in-class survey and a final online survey. The experiment consisted of three rounds, out of which we were able to test the role of gender on policies, visual appearances of politicians and whether voters will vote for politicians based on looks regardless of their political standing. These surveys left us with a large sample of 129 respondents, which although slightly left leaning and skewing more towards younger demographics was able to provide us insight into the role of gender.

Overall, we did not find statistically significant impact of gender of the politician on policy agreeableness. Additionally, we also did not find that appearances played a role in how likely voters are to agree with policies. We did find that politicians who were classified as looking more "likeable" were more likely to win in a blind voting election, similar to existing literature. Moreover, we found that our experiment design for ranking the "likeability" of politicians was successful in providing information on voter's opinions. When we looked at the data in more detail, while we did not find statistically significant results there does ap-

pear to be evidence that male voters are more responsive to the looks of politicians, especially when they are proposing "bad" policies as are respondents who identify as "right" leaning.

While the experiment was unable to determine whether women are held to higher standards for many "mistakes" they make while in power, the originality of the design could be used to determine whether voters do treat female politicians differently and whether looks matter if a larger more diverse sample was used. We hope that future researchers will be able to build on our experimental design to answer why female politicians are still not treated equally by the electorate even once they have proven themselves in positions of power.

References

Antonakis, J., & Dalgas, O. (2009). Predicting elections: Child's play! *Science*, *323*(5918), 1183–1183.

Barnes, T. D., & Beaulieu, E. (2019). Women politicians, institutions, and perceptions of corruption. *Comparative Political Studies*, *52*(1), 134–167.

Bauer, N. M. (2020). Shifting standards: How voters evaluate the qualifications of female and male candidates. *The Journal of Politics*, *82*(1), 1–12.

Berggren, N., Jordahl, H., & Poutvaara, P. (2010). The looks of a winner: Beauty and electoral success. *Journal of public economics*, *94*(1-2), 8–15.

Brollo, F., & Troiano, U. (2016). What happens when a woman wins an election? evidence from close races in brazil. *Journal of Development Economics*, *122*, 28–45.

Chattopadhyay, R., & Duflo, E. (2004). Women as policy makers: Evidence from a randomized policy experiment in india. *Econometrica*, *72*(5), 1409–1443.

Chiao, J. Y., Bowman, N. E., & Gill, H. (2008). The political gender gap: Gender bias in facial inferences that predict voting behavior. *PloS one*, *3*(10), e3666.

Delaporte, M., & Pino, F. (2022). Female political representation and violence against women: Evidence from brazil.

Eggers, A. C., Vivyan, N., & Wagner, M. (2018). Corruption, accountability, and gender: Do female politicians face higher standards in public life? *The Journal of Politics*, *80*(1), 321–326.

Gothreau, C. M., Alvarez, A. M., & Friesen, A. (2022). Objectified and dehumanized: Does objectification impact perceptions of women political candidates? *Journal of Experimental Political Science*, 1–14.

Haumer, F., & Donsbach, W. (2009). The rivalry of nonverbal cues on the perception of politicians by television viewers. *Journal of Broadcasting & Electronic Media*, *53*(2), 262–279.

Hengel, E. (2022). Publishing while female: Are women held to higher standards? evidence from peer review. *The Economic Journal*, *132*(648), 2951–2991.

Huddy, L., & Terkildsen, N. (1993). Gender stereotypes and the perception of male and female candidates. *American journal of political science*, 119–147.

Johns, R., & Shephard, M. (2007). Gender, candidate image and electoral preference. *The British Journal of Politics and International Relations*, *9*(3), 434–460.

Lawson, C., Lenz, G. S., Baker, A., & Myers, M. (2010). Looking like a winner: Candidate appearance and electoral success in new democracies. *World Politics*, *62*(4), 561–593.

Mengel, F., Sauermann, J., & Zölitz, U. (2019). Gender bias in teaching evaluations. *Journal of the European economic association*, *17*(2), 535–566.

Open data initiative of the Government of Spain. (2023). *Open data from the spanish government - datos.gob.es*. https://datos.gob.es/en. (Accessed: 29 June 2023)

Piscopo, J. M., Hinojosa, M., Thomas, G., & Siavelis, P. M. (2022). Follow the money: Gender, incumbency, and campaign funding in chile. *Comparative Political Studies*, *55*(2), 222–253.

Rosar, U., Klein, M., & Beckers, T. (2008). The frog pond beauty contest: Physical attractiveness and electoral success of the constituency candidates at the north rhine-westphalia state election of 2005. *European Journal of Political Research*, *47*(1), 64–79.

Sarsons, H. (2017). Interpreting signals in the labor market: evidence from medical referrals. *Job Market Paper*, 141–145.

Sevi, S. (2023). Is incumbency advantage gendered? *Legislative Studies Quarterly*, *48*(1), 145–163.

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence

from faces predict election outcomes. *Science*, *308*(5728), 1623–1626.

Volden, C., Wiseman, A. E., & Wittmer, D. E. (2013). When are women more effective lawmakers than men? *American Journal of Political Science*, *57*(2), 326–341.

Zoghaib, A. (2019). Persuasion of voices: The effects of a speaker's voice characteristics and gender on consumers' responses. *Recherche et Applications en Marketing (English Edition)*, *34*(3), 83–110.

8 Appendix

8.1 First Step Specifications

8.1.1 Policies Description

An example of a bad policy description is the "lowering the age of consent to 13" one: "Advocating for an inclusive society, we support lowering the age of consent to 13. This progressive step recognizes evolving understandings of sexuality, promotes comprehensive education, and empowers young individuals to make informed decisions. By fostering open dialogue and providing support systems, we enable responsible exploration of relationships while emphasising consent and respect. This approach ensures the healthy sexual development of young people, reduces stigma, and encourages a culture of inclusivity and understanding."

An example of a good policy description is the "clean air initiative" one: "Introducing our comprehensive Clean Air Initiative, dedicated to improving air quality and securing a healthier future for all. This initiative focuses on reducing pollution, enforcing strict emission standards, and promoting sustainable practices. By investing in renewable energy, supporting electric vehicles, and embracing green technologies, we create opportunities for green job growth while safeguarding public health. Together, we can breathe clean and build a sustainable future."

8.1.2 AI Names Generating

The final names for the female candidate were: Clara Soler, Emma Vega, Isabella Almeida, Sophia Mayoral, Elena Mendes

The final names for the male candidate were: Oliver Cruz, Gabriel Morales, Lucas Almeida, Adrian Mayoral, David Mendes

8.2 Sample of Respondents

Figures 12 and 13 both show what proportion of our respondents were in the picture pathway and text pathway and how many partook in our pilot sample and in our online sample.

Moreover, we provide information on the political split of our sample and the age distribution. In figure 15 we can see that we had an even gender split within our sample.

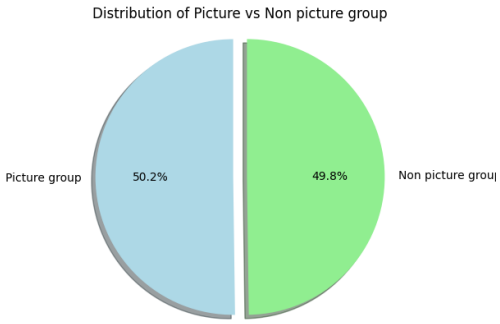

Figure 11: Distribution of the sample in Picture and No picture group

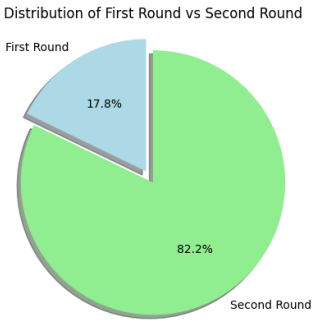

Figure 12: Distribution of respondents by round

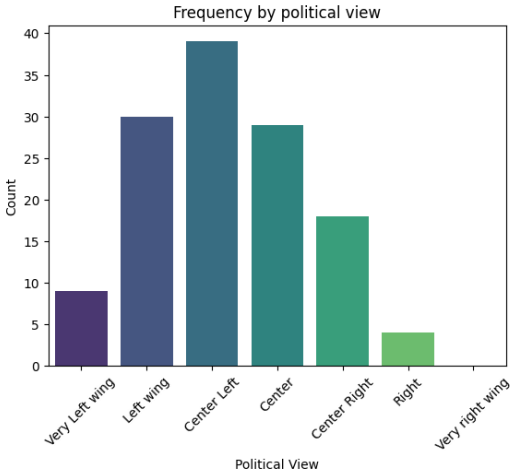

Figure 13: Distribution of the respondents by political view

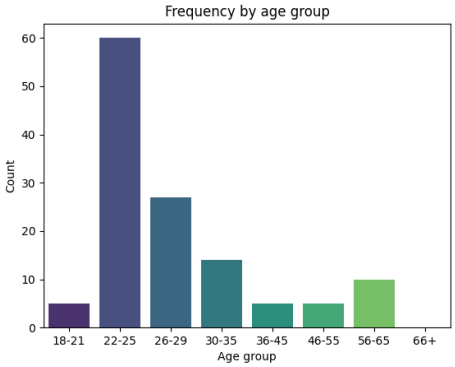

Figure 14: Distribution of respondents by age

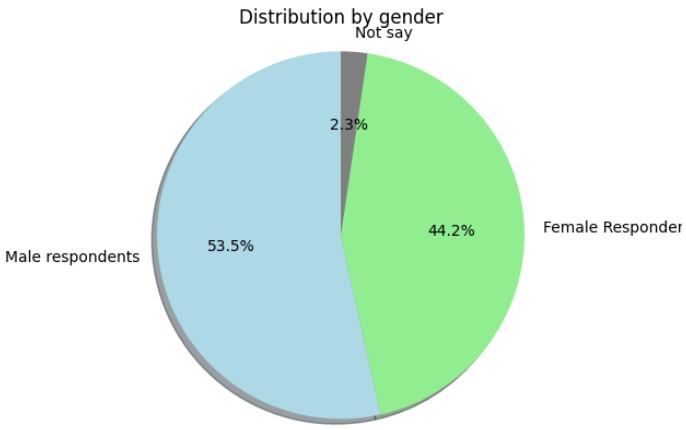

Figure 15: Gender distribution of respondents

source	ranking
clean_air-M_NOPIC	8.892857
clean_air-W_NOPIC	8.206897
consent-M_NOPIC	2.821429
consent-W_NOPIC	2.333333
drugs-M_NOPIC	4.909091
drugs-W_NOPIC	6.333333
mental health-M_NOPIC	8.129032
mental health-W_NOPIC	8.064516
videogames-M_NOPIC	3.941176
videogames-W_NOPIC	6.000000

Table 3: Caption

	P-value
Man	0.0000
Woman	0.0015
Right Wing	0.1671
Left Wing	0.0000

Table 4: P-values results for the t-test. We tested the whether or not the distributions of the valuation of the policy given by these groups vary for attractive or non-attractive politicians.

8.3 Results Stage 1

the results for the text only path are summarised in table 3. The results of how "looks" impacted policy agreement are summarised in figures 16 and 17. Table 4 summarised the P-values of the binomial tests (T-tests) that were ran for male and female respondents on the whether the responses were significantly different between attractive and unattractive politicians.

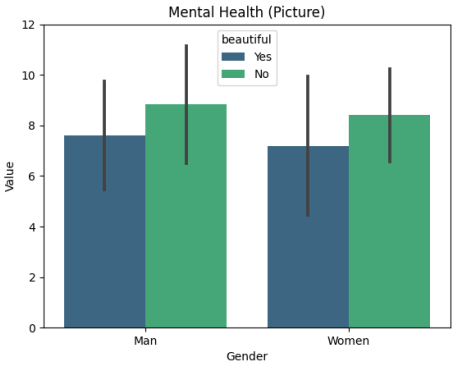

Figure 16: Results Mental Health Policy

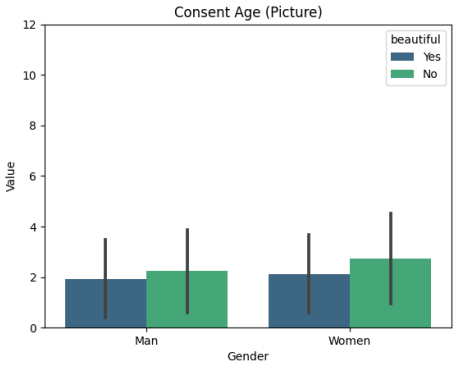

Figure 17: Results Age of Consent Policy

8.4 Stage 2 Results

Below are summarised the total votes received by each candidate in the second voting stage. We also include the total number of votes the second option (the runner up) received.

Figure 18: Voting results for 1st option - figure A

Figure 19: Voting results for 2nd option - figure A

Figure 20: Voting results for 1st option - figure B

Figure 21: Voting results for 2nd option - figure B