My title*

My subtitle if needed

First author Yongqi Liu Yuxuan Wei

October 23, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

```
#install.packages("tidymodels")
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

```
library(ggplot2)
analysis_data <- read_csv(here::here("data/02-analysis_data/cleaned_US_voting.csv"))
```

```
Rows: 1683 Columns: 11
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (5): pollster_rating_name, methodology, state, candidate_name, populati...
dbl  (4): numeric_grade, sample_size, percent, transparency_score
date (2): start_date, end_date

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

*Code and data are available at: [https://github.com/wyx827/2024USpresidentialelection.git].

1 Introduction

Overview paragraph

Estimand paragraph

Results paragraph

Why it matters paragraph

Telegraphing paragraph: The remainder of this paper is structured as follows. Section **??**….

2 Data

2.1 Overview

We use the statistical programming language R (R Core Team 2023)…. Following Alexander (2023), we consider…

Overview text

2.2 Measurement

Some paragraphs about how we go from a phenomena in the world to an entry in the dataset.

2.3 Outcome variables

Add graphs, tables and text. Use sub-sub-headings for each outcome variable or update the subheading to be singular.

Some of our data is of penguins (**?@fig-bills**),

Talk more about it.

And also planes (**?@fig-planes**). (You can change the height and width, but don't worry about doing that until you have finished every other aspect of the paper - Quarto will try to make it look nice and the defaults usually work well once you have enough text.)

Talk way more about it.

2.4 Predictor variables

Add graphs, tables and text.

Use sub-sub-headings for each outcome variable and feel free to combine a few into one if they go together naturally.

3 Model

To model Donald Trump's polling percentages over time, we employed a Bayesian linear regression framework. Bayesian methods provide a flexible approach to inference by incorporating prior beliefs and updating them with observed data, allowing us to quantify uncertainty in both parameter estimates and predictions.

Here we briefly describe the Bayesian model used to investigate the winning probability of Trump. Background details and diagnostics are included in Appendix **??**.

3.1 Multiple Linear Regression Model Overview

The model now predicts Trump's polling percentage (percent) using the following predictors:

- Numeric Grade (numeric_grade): Reflects the quality rating of the pollster.
- Sample Size (sample_size): The number of respondents in the poll.
- State (state): A categorical variable for different U.S. states.
- Transparency Score (transparency_score): A measure of how transparent the polling data and methodology are.
- End Date (end_date): The date the poll was completed, which might capture trends over time.

The model takes the form: The model takes the form:

4 #Interpretation of Coefficients:

- Intercept (_0): This is the predicted Trump polling percentage when all predictors (numeric grade, sample size, state, transparency score, and end date) are at their baseline or zero value.
- Numeric Grade (_1): This coefficient measures how much Trump's polling percentage changes as the pollster's numeric grade increases. A positive and significant coefficient would indicate that higher-rated pollsters report better polling numbers for Trump, while a negative coefficient would suggest the opposite.

- Sample Size (_2): This measures the impact of the number of respondents on Trump's polling percentage. A positive coefficient would indicate that larger sample sizes are associated with higher polling percentages for Trump.
- State (_3): The coefficients for the state variable represent differences in Trump's polling percentage in each state compared to the reference state (baseline category). For example, if the coefficient for Florida is negative, it means Trump polls lower in Florida compared to the reference state.
- Transparency Score (_4): This coefficient shows how much Trump's polling percentage is affected by the transparency of the poll. A positive coefficient would indicate that polls with higher transparency tend to report higher polling percentages for Trump, whereas a negative coefficient would imply the opposite.
- End Date (_5): The end date is a time-related variable, capturing trends over time. A positive and significant coefficient would suggest that Trump's polling percentage has increased as the election date approaches, while a negative coefficient would suggest a decrease in his polling numbers over time.

4.1 Interpretation

The posterior distributions of the parameters allow us to quantify the uncertainty around each effect: - The coefficient for end_date informs us about how Trump's polling percentages have evolved over time. A positive coefficient would suggest an upward trend, while a negative coefficient would indicate a decline. - The coefficient for numeric_grade captures the impact of pollster quality on the polling percentage. High-quality pollsters may produce different estimates compared to lower-quality ones. - The state-level effects account for regional differences in Trump's support. Some states may show significantly higher or lower levels of support, even after adjusting for the time of the poll and pollster quality.

5 Model Evalutation

- R-squared Table **??** shows the summary table for the model. And then evaluate it on the test set. It appears as though the model is having difficulty identifying Trump supporters.

Table 1: Relationship between wing length and width

Table 1: Regression Model Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-67.184	14.445	-4.651	0.000
numeric_grade	0.914	1.006	0.909	0.364

	Estimate	Std. Error	t value	Pr(>|t|)
sample_size	0.000	0.000	0.205	0.838
stateArizona	-5.020	2.582	-1.944	0.052
stateArkansas	5.569	4.380	1.272	0.204
stateCalifornia	-18.345	2.738	-6.699	0.000
stateColorado	-13.340	3.002	-4.443	0.000
stateConnecticut	-13.480	3.574	-3.772	0.000
stateFlorida	-3.214	2.740	-1.173	0.241
stateGeorgia	-4.814	2.573	-1.871	0.062
stateIdaho	3.467	4.379	0.792	0.429
stateIllinois	-14.064	3.275	-4.294	0.000
stateIndiana	0.902	3.268	0.276	0.783
stateIowa	-4.692	3.101	-1.513	0.131
stateKansas	-0.798	3.275	-0.244	0.808
stateMaine	-10.182	2.726	-3.735	0.000
stateMaryland	-20.244	2.998	-6.753	0.000
stateMassachusetts	-22.543	2.753	-8.190	0.000
stateMichigan	-6.635	2.576	-2.576	0.010
stateMinnesota	-10.250	2.684	-3.819	0.000
stateMissouri	1.087	2.803	0.388	0.698
stateMontana	1.379	2.784	0.495	0.620
stateNational	-7.935	2.546	-3.117	0.002
stateNebraska	-7.528	2.782	-2.706	0.007
stateNevada	-5.617	2.594	-2.165	0.031
stateNew Hampshire	-8.778	2.655	-3.306	0.001
stateNew Jersey	-15.189	3.577	-4.246	0.000
stateNew Mexico	-10.980	3.099	-3.543	0.000
stateNew York	-15.239	2.633	-5.789	0.000
stateNorth Carolina	-4.783	2.582	-1.853	0.064
stateNorth Dakota	2.970	4.380	0.678	0.498
stateOhio	-2.796	2.677	-1.045	0.296
stateOklahoma	8.450	4.385	1.927	0.054
stateOregon	-14.974	4.385	-3.415	0.001
statePennsylvania	-6.375	2.568	-2.483	0.013
stateRhode Island	-13.374	2.994	-4.467	0.000
stateSouth Carolina	-1.939	4.376	-0.443	0.658
stateSouth Dakota	0.394	3.096	0.127	0.899
stateTennessee	3.624	4.378	0.828	0.408
stateTexas	-3.825	2.681	-1.427	0.154
stateUtah	-3.243	4.385	-0.740	0.460
stateVermont	-23.452	3.581	-6.549	0.000
stateVirginia	-9.518	2.663	-3.575	0.000

	Estimate	Std. Error	t value	Pr(>|t|)
stateWashington	-15.876	3.264	-4.864	0.000
stateWest Virginia	8.265	4.379	1.887	0.059
stateWisconsin	-6.444	2.571	-2.506	0.012
stateWyoming	15.921	4.377	3.637	0.000
transparency_score	-0.646	0.088	-7.383	0.000
end_date	0.006	0.001	8.880	0.000

6 Diagnostic plots for the model

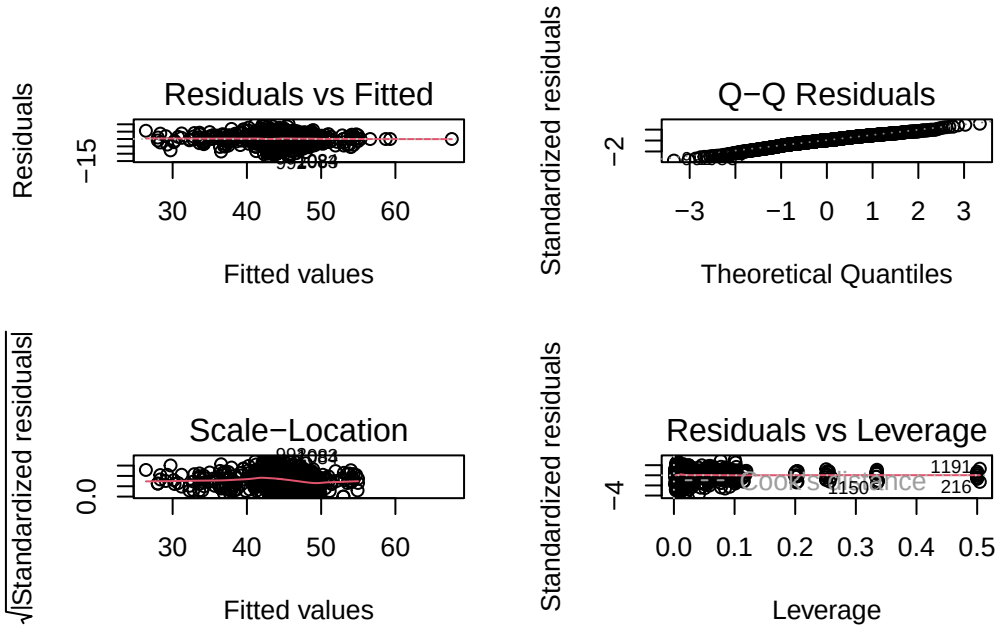

Figure 1

```
library(car)
```

```
Loading required package: carData
```

```
Attaching package: 'car'
```

```
The following object is masked from 'package:dplyr':

    recode

The following object is masked from 'package:purrr':

    some
```

```
vif(regression_model)
```

```
                     GVIF Df GVIF^(1/(2*Df))
numeric_grade      1.294110  1        1.137590
sample_size        1.289750  1        1.135672
state              2.082151 44        1.008369
transparency_score 1.263558  1        1.124081
end_date           1.309491  1        1.144330
```

6.0.1 Model justification

We expect a positive relationship between the size of the wings and time spent aloft. In particular...

We can use maths by including latex between dollar signs, for instance θ.

7 Results

7.1 Predict and Combine with Electroal College to Predict

Table 2: Prediction for Trump

Table 2: Prediction for Trump by State

State	Trump Predicted %	Electoral Votes	Winner
Alaska	53.00	3	Trump
Arizona	47.62	11	Harris
Arkansas	56.60	6	Trump
California	34.39	55	Harris
Colorado	37.60	9	Harris
Connecticut	39.09	7	Harris

State	Trump Predicted %	Electoral Votes	Winner
Florida	48.89	29	Harris
Georgia	47.71	16	Harris
Idaho	54.50	4	Trump
Illinois	36.37	20	Harris
Indiana	53.73	11	Trump
Iowa	46.68	6	Harris
Kansas	49.63	6	Harris
Maine	41.13	2	Harris
Maryland	32.88	10	Harris
Massachusetts	29.80	11	Harris
Michigan	45.92	15	Harris
Minnesota	42.40	10	Harris
Missouri	52.51	10	Trump
Montana	54.20	3	Trump
Nebraska	45.16	5	Harris
Nevada	47.00	6	Harris
New Hampshire	42.86	4	Harris
New Jersey	37.60	14	Harris
New Mexico	42.05	5	Harris
New York	37.45	29	Harris
North Carolina	47.98	16	Harris
North Dakota	54.00	3	Trump
Ohio	49.58	18	Harris
Oklahoma	58.70	7	Trump
Oregon	35.30	6	Harris
Pennsylvania	46.28	20	Harris
Rhode Island	39.12	4	Harris
South Carolina	50.60	9	Trump
South Dakota	52.18	3	Trump
Tennessee	55.30	11	Trump
Texas	48.51	38	Harris
Utah	47.00	6	Harris
Vermont	28.00	3	Harris
Virginia	43.22	13	Harris
Washington	36.33	12	Harris
West Virginia	59.30	5	Trump
Wisconsin	45.96	10	Harris
Wyoming	67.60	3	Trump

8 Discussion

8.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

8.2 Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

8.3 Third discussion point

8.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Pollster Methodology Overview and Evaluation

A.1 Overview of SurveyUSA

SurveyUSA is a privately held opinion research company that operates nationwide, across all 50 U.S. states. Since its founding, the company has conducted over 40,000 research projects, serving a client base of 400 organizations, including media outlets, corporations, non-profits, government agencies, and academic institutions. Known for its expertise in localized opinion research, SurveyUSA focuses on gathering data at the city, county, and regional levels. The company offers timely, cost-effective surveys tailored to meet specific client needs, distinguishing itself from larger global firms.

A.2 Population, Frame, and Sample

- Target Population: U.S. citizens eligible to vote in the 2024 presidential election.
- Sample Frame: U.S. households with either home telephones or access to devices such as phones or tablets.
- Sample Size: Sample sizes vary across different polls. For the 2024 U.S. presidential election cycle, SurveyUSA conducted 49 polls, with sample sizes ranging from 507 to 2,330 for registered voters or likely voters. The average sample size for these polls is approximately 1,045 households.

A.3 Recruitment

SurveyUSA employs a mixed-method approach to recruitment, including online panels, telephone calls, and a text-to-web method. Some respondents are recruited through Random Digit Dialing (RDD) using telephone samples purchased from Aristotle, while others, who do not use home telephones, are invited to complete the survey on an electronic device such as a phone or tablet. Respondents from non-probability online panels are selected randomly by Cint/Lucid Holdings LLC.

A.4 Sampling approach and Trade-offs

SurveyUSA uses a blend of probability and non-probability sampling methods. Some respondents are drawn from non-probability online panels, while others are recruited using probability-based telephone sampling. Responses are weighted based on the latest U.S.

Census estimates for age, gender, ethnicity, and region, ensuring alignment with the target population. Questions and answer choices are rotated to reduce order bias, recency effects, and latency effects.

- **Advantages:**
  The diverse sampling approach not only ensures a broad range of opinions is captured but also complements probability-based sampling, which accurately reflects the overall population. Furthermore, reweighting the data according to U.S. Census demographics strengthens the credibility of the results by ensuring demographic accuracy. Additionally, rotating questions and answer choices helps mitigate bias, further improving the reliability of the data. Finally, the use of online surveys offers a cost-effective solution for efficient data collection.

- **Disadvantages:**
  Phone-based data collection tends to be time-consuming and can be affected by interviewer effects during telephone interviews. Additionally, challenges like non-response issues, such as busy signals or refusals to participate, can hinder the effectiveness of the data collection process.

A.5 Non-response Handling

In cases of non-response, SurveyUSA attempts follow-up calls if interviews are interrupted by answering machines or busy signals. Weighting is applied to adjust for non-response bias, although this doesn't completely eliminate challenges posed by unreachable or unwilling participants.

A.6 Questionnaire Evaluation

- **Positive Aspects:** A logical flow between questions facilitates easy navigation for respondents throughout the survey, while simple wording promotes inclusivity by enabling individuals from diverse backgrounds to comprehend the questions. Furthermore, all questions are directly relevant to analyzing the 2024 U.S. presidential election, and providing predefined response options simplifies the choices for participants.

- **Negative Aspects:** Static options for party affiliation and ideology may fail to capture the nuances of respondents' political beliefs. These rigid categories could oversimplify complex political identities.

A.7 Summary Evaluation

SurveyUSA's methodology reflects a balanced approach, leveraging various sampling approach and method to reach a representative sample. While its blend of probability and non-probability methods has strengths, such as cost-effectiveness and broad reach, it faces challenges related to telephone interview logistics, potential interviewer bias, and the limitations of fixed questionnaire options. Nevertheless, the inclusion of data weighting and question rotation adds credibility to its results, making SurveyUSA a reliable pollster for localized opinion research.

B Appendix B: Idealized Methodology and Survey

B.1 Objective and Overview

The goal of this survey methodology is to accurately forecast the outcome of the U.S. presidential election by collecting high-quality, representative data from a diverse set of respondents across the country. With a budget of $100,000, this methodology incorporates sophisticated sampling techniques, robust respondent recruitment strategies, and rigorous data validation protocols. The approach is designed to maximize accuracy, reduce bias, and account for various demographic, geographic, and political factors that influence voting behavior.

B.1.1 Core Objectives:

1. Obtain a representative sample of the U.S. electorate.
2. Ensure data quality through rigorous validation.
3. Leverage statistical modeling and poll aggregation for an accurate prediction.

B.2 1. Sampling Strategy

The sampling strategy is designed to ensure that the survey reaches a broad, representative section of the voting population. To achieve this, we will use **stratified random sampling** combined with **quota sampling** for key demographics. This ensures that each important subgroup within the population is adequately represented.

B.2.1 Stratification Variables:

- **Age Groups**: 18-29, 30-44, 45-64, 65+
- **Gender**: Male, Female, Non-binary/Other

- **Race/Ethnicity**: White, Black, Hispanic/Latino, Asian, Indigenous, Other
- **Education Level**: No high school, High school graduate, College graduate, Post-graduate
- **Income Bracket**: <$30,000, $30,000-$60,000, $60,000-$100,000, >$100,000
- **Geographic Region**: Northeast, Midwest, South, West

Sample Size:
A total of **10,000 respondents** will be surveyed, providing a margin of error of approximately ±1% at a 95% confidence level. This sample size will allow for detailed subgroup analysis (e.g., by state, demographic group), yielding statistically robust predictions.

Weighting:
We will apply post-stratification weights to adjust for any oversampling or undersampling of specific demographic groups. For example, younger voters or underrepresented minorities will be weighted to reflect their true proportions in the voting population.

B.3 2. Recruitment Strategy

B.3.1 Recruitment Channels:

To maximize respondent diversity and ensure accurate sampling, the survey will employ **multi-channel recruitment**:

- **Digital Advertisements**: Targeted ads on platforms like Facebook, Instagram, and Google will be used to recruit respondents based on their demographic profiles (age, gender, location, political interest). Custom audience features will be utilized to reach specific demographic and geographic groups.
- **Email Outreach**: If permissible, we will access voter registration databases and send email invitations to registered voters. This will allow us to target specific voter demographics that are harder to reach via digital ads.
- **Partnerships with Civic Organizations**: Partnering with non-profits and civic organizations that engage diverse communities (e.g., minority voter outreach programs) will further boost respondent diversity.
- **Incentives**: To increase response rates, each participant will be entered into a lottery with a chance to win a $100 gift card, encouraging broader participation.

B.4 3. Data Validation and Quality Assurance

Maintaining data integrity and ensuring high-quality responses are critical to the accuracy of the election forecast. Therefore, several measures will be put in place to validate responses and reduce noise in the dataset.

B.4.1 Data Validation Protocols:

1. **Real-time Captcha Verification**: This will prevent automated bots from submitting responses.
2. **Email/Phone Verification**: Respondents will verify their email or phone number to ensure authenticity. This ensures that each respondent only participates once.
3. **Time on Task Monitoring**: The survey platform will monitor the time respondents spend on each question. Responses completed suspiciously quickly (e.g., below 30% of the average completion time) will be flagged for review or exclusion.
4. **Voter Registration Cross-Check**: If feasible, respondents will be cross-referenced with voter registration records to ensure they are eligible to vote in the upcoming election.
5. **Response Audits**: Randomly selected respondents will be contacted to verify the accuracy of their responses, ensuring integrity in the dataset.

B.5 4. Poll Aggregation and Data Analysis

B.5.1 Poll Aggregation:

This survey will be combined with results from reputable polling firms (e.g., YouGov, Ipsos, Gallup) to strengthen our forecast through a **poll-of-polls** approach.

- **Weighting by Methodology and Recency**: Poll results will be weighted based on the rigor of their methodology (e.g., online vs. phone surveys, sample size) and the recency of the poll. Recent, methodologically sound polls will receive more weight in the aggregation process.
- **Handling Bias and Variability**: Aggregated results will adjust for pollster house effects (biases in methodology) and variability between polls, ensuring that no single poll dominates the prediction.

B.5.2 Modeling Approach:

We will implement **Bayesian hierarchical models** to account for variability across different states, demographics, and regions. This will allow us to model the popular vote and potentially translate it into **Electoral College predictions**.

B.6 5. Budget Allocation

- **Respondent Recruitment (Targeted ads, outreach)**: $70,000
- **Incentives (e.g., lottery prizes)**: $10,000
- **Survey Platform (Google Forms, Qualtrics subscription)**: $5,000
- **Data Validation Tools**: $5,000
- **Poll Aggregation & Analysis Software**: $10,000

B.7 6. Survey Implementation

The survey will be implemented via **Google Forms**, which offers a cost-effective platform for data collection. The link to the live survey can be found here: Google Form Survey. A copy of the questions is provided below.

B.7.1 Survey Structure:

1. **Introduction**: Thank you for taking part in this survey aimed at predicting the outcome of the 2024 US Presidential election. Your insights are valuable to our research.

Please note: - **All responses will be kept strictly confidential. - Your participation is entirely voluntary. - We kindly request that you answer all questions honestly and to the best of your knowledge. - The survey is estimated to take approximately 10 minutes to complete.** If you have any inquiries or concerns regarding this survey, please don't hesitate to contact the research team at shaw.wei@mail.utoronto.ca.

Your contribution to this study is greatly appreciated! Each participant will be entered into a lottery with a chance to win a $100 gift card!

2. **Section 1: Eligibility Screening**: Are you a U.S. citizen?

- Yes
- No [If No, end survey]

Will you be 18 or older by Election Day (November 5, 2024)? - Yes - No [If No, end survey]

Are you registered to vote in the United States? - Yes - No - Not sure - Plan to register before the election

3. **Section 2: Demographic Information**: What is your age group?

- 18-29
- 30-44
- 45-64
- 65 or older
- Prefer not to say

What is your gender? - Male - Female - Non-binary/Other - Prefer not to say

What is your race/ethnicity? (Select all that apply) - White - Black or African American - Hispanic or Latino - Asian - American Indian or Alaska Native - Native Hawaiian or Pacific Islander - Prefer not to say - Other: [Short text answer]

What is your highest level of education completed? - No high school - High school graduate or equivalent - Some college, no degree - Bachelor's degree - Graduate or professional degree - Prefer not to say

What was your total household income in 2023? - Less than $30,000 - $30,000 - $59,999 - $60,000 - $99,999 - $100,000 - $149,999 - $150,000 or more - Prefer not to say

In which region of the United States do you currently reside? - Northeast (ME, NH, VT, MA, RI, CT, NY, NJ, PA) - Midwest (OH, IN, IL, MI, WI, MN, IA, MO, ND, SD, NE, KS) - South (DE, MD, DC, VA, WV, NC, SC, GA, FL, KY, TN, AL, MS, AR, LA, OK, TX) - West (MT, ID, WY, CO, NM, AZ, UT, NV, WA, OR, CA, AK, HI)

4. **Section 3: Political Views and Voting Intentions**: How likely are you to vote in the 2024 Presidential election?

- Definitely will vote
- Probably will vote
- Might or might not vote
- Probably will not vote
- Definitely will not vote

Generally speaking, do you usually think of yourself as a: - Democrat - Republican - Independent - Prefer not to say - Other: [Short text answer]

If the 2024 Presidential election were held today, who would you vote for? - Kamala Harris (Democrat) - Donald Trump (Republican) - Undecided - Prefer not to say - Other: [Short text answer]

How certain are you about your choice? - Very certain - Somewhat certain - Not very certain - Not at all certain - Prefer not to say

Which THREE issues are most important to you in deciding your vote? (Select exactly three) - Economy and jobs - Healthcare - Immigration - Climate change - National security - Education