Learn Data Science Using SAS Studio

A Quick-Start Guide

Engy Fouda

Apress®

Learn Data Science Using SAS Studio

A Quick-Start Guide

Engy Fouda

Apress®

Learn Data Science Using SAS Studio: A Quick-Start Guide

Engy Fouda
Hopewell Junction, NY, USA

ISBN-13 (pbk): 978-1-4842-6236-8 ISBN-13 (electronic): 978-1-4842-6237-5
https://doi.org/10.1007/978-1-4842-6237-5

Managing Director, Apress Media LLC: Welmoed Spahr
Acquisitions Editor: Susan McDermott
Development Editor: Laura Berendson
Coordinating Editor: Rita Fernando

Cover designed by eStudioCalamar

Cover image designed by Freepik (www.freepik.com)

Distributed to the book trade worldwide by Springer Science+Business Media New York, 1 New York Plaza, New York, NY 10004. Phone 1-800-SPRINGER, fax (201) 348-4505, email orders-ny@springer-sbm.com, or visit www.springeronline.com. Apress Media, LLC is a California LLC and the sole member (owner) is Springer Science+Business Media Finance Inc (SSBM Finance Inc). SSBM Finance Inc is a **Delaware** corporation.

For information on translations, please e-mail booktranslations@springernature.com; for reprint, paperback, or audio rights, please e-mail bookpermissions@springernature.com.

Apress titles may be purchased in bulk for academic, corporate, or promotional use. eBook versions and licenses are also available for most titles. For more information, reference our Print and eBook Bulk Sales web page at http://www.apress.com/bulk-sales.

Any source code or other supplementary material referenced by the author in this book is available to readers on GitHub via the book's product page, located at www.apress.com/9781484262368. For more detailed information, please visit http://www.apress.com/source-code.

Printed on acid-free paper

To my daughter, Areej, and my husband, Hesham, a big thank you to both of you from my deep heart. I would have never finished this book without your support and encouragement. You created for me a happy and safe life while the world outside was crazy and losing its sanity. I am a lucky person to have you both in my life. Thank you so much! Love you both!

To my mom, Suzan, and dad, Samir, I miss both of you madly and wish you were with me celebrating this book. I am sure that you are celebrating and happy together in the heavens. I owe you everything and hope you are as proud of me as you always were. Till we meet, I love you.

Finally, to my brothers, Haitham and Khaled, you never gave up on me and always believed in me. I always need your trust that I can achieve anything. I appreciate it and with it I gain my ability to move forward. I hope you like this book.

Table of Contents

About the Author

Engy Fouda is an author, freelance engineer, and journalist. Currently, she teaches SAS, Docker Fundamentals, Docker for Enterprise Developers, Docker for Enterprise Operations, and Kubernetes at several venues as a freelance instructor. She is an Apress and Packt Publishing author. She works as a freelance journalist and publishes her work at various media outlets. She holds two master's degrees, one in journalism from Harvard University, Extension School, and another in computer engineering from Cairo University, Egypt. Moreover, she earned the Data Science Professional Graduate Certificate from Harvard University, Extension School. She has taught academically as a teacher assistant at the German University in Cairo and the American University in Cairo. She volunteers as the team lead for Momken Group (Engineering for the Blind), Egypt Scholars Inc. The team designs and manufactures devices and develops Arabic applications for the visually impaired people in the Middle East and North Africa region. Also, she volunteers as a member-at-large and the newsletter editor of the IEEE Mid-Hudson Section. She has published several books that made Amazon's best-seller charts for Arabic books.

About the Technical Reviewer

Allan Bowe is a SAS geek with a passion for HTML5 apps on SAS. Allan has made a number of contributions to the SAS community, such as SASjs (an adapter for bidirectional communication between HTML5 and SAS), sasjs-cli (a command-line tool for managing SAS project compilation, build, and deployment), and macrocore (a SAS macro library for building SAS apps on both SAS 9 and Viya).

When not building web apps, Allan is working on Data Controller, a commercial data capture, data quality, and data governance web app for both SAS 9 and Viya.

Introduction

The book's scope is primarily to introduce SAS Studio, a free data science web browser–based product for non-commercial and academic usage. SAS Studio is also known as SAS University Edition. The power of SAS Studio relies on its visual point-and-click user interface that generates SAS code. Users can create data analysis reports without writing a line of code, unlike with R and Python. Hence, data cleaning, statistics, and visualization are easy to do.

The book's case study analyzes the presidential elections data in Maine, which is part of a project I did at Harvard University. Chapter 1 explains the case study in more detail.

In addition to the presidential elections, the book uses real-life examples like analyzing stocks, oil and gold prices, crime, marketing, and healthcare. The whole book follows the paradigm of data science in action to demonstrate how easy it is to perform complicated tasks and visualizations in SAS Studio.

The book starts from scratch in step-by-step, hands-on labs, and includes screenshots of every step.

It will provide readers the required expertise in data science and analytics using SAS Studio, such as how to do the following:

- Import and export raw data files
- Manipulate and transform data
- Combine SAS data sets
- Create summary statistics and reports using SAS procedures
- Identify and correct data, syntax, and programming logic errors
- Compare between samples using T-test and Anova
- Predict new values using linear regression
- Create visualizations

Moreover, it will show how to do visualizations, including maps, step-by-step. In many cases, readers will not need to write a line of code, because SAS has a powerful graphical user interface. It is much easier to learn compared to R and Python. However, every example will explain the auto-generated code and how to edit it to perform more-complicated advanced tasks. The book will introduce you to multiple SAS products, such as SAS Studio, SAS Viya, SAS Analytics, and SAS Visual Statistics.

I teach most of these contents as a course at one of the Microsoft Partner Centers. The students always get amazed by how much they learn in merely a few days.

Who Should Read This Book?

The primary audience of this book is students who are newbies to data science and might not have deep programming experience. Consequently, university professors can use it as a handout for their courses. Also, technical instructors, like me, who teach professionals from various industry sectors at certified training centers can teach from it.

Also, system analysts and scientists who are experienced but new to SAS will find faster and more efficient tools to achieve their daily tasks. From my experience, I learned from the attendees of my courses that many government agencies migrated to SAS. Moreover, data journalists and investigative reporters will find the book easy to follow and will be able to generate pretty visualizations quickly.

How Is This Book Organized?

In general, the book tries to balance between using SAS point-and-click and the code. The users might be tempted to rely on the integrated development environment (IDE). However, the book uses the IDE merely to introduce the users to the various tasks, along with explaining the code so as to be able to do advanced tasks.

The book has three parts. Part I, "Basics" (Chapters 1, 2, and 3), gets you familiar with the SAS interface and the basic essential tasks. Chapter 1 focuses on drawing a general idea of the case study of the presidential election project in the state of Maine and its outputs. Throughout the book, you will learn how to get those output charts and analytics step-by-step.

Then, Part II, "More Programming" (Chapters 4, 5, and 6), focuses on more advanced programming aspects. Part III, "Advanced Topics" (Chapters 7, 8, and 9), takes you from analyzing historical data to predicting the future and introducing you to more advanced SAS platforms.

Finally, in Chapter 9, I try to give some insights from my personal experience on how to get certified and how to make money online through teaching, writing, and competing online. Moreover, I shall give you more details on the data science graduate professional certificate at Harvard University, Extension School.

All the SAS example code is in the "Example Code" folder, and datasets required for this book are in the "Datasets" folder of this book's source code. Go to http://www.apress.com/9781484262368 and click the Download Source Code button to access it.

PART I

Basics

CHAPTER 1

Data Science in Action

In this chapter, we will introduce the case study of the book, which analyzes voters' data in the state of Maine. It is based on a project I did at Harvard University in 2016 during my master's degree. In fall 2016, the project for my "A Practical Approach to Data Science" course was to predict the presidential election results in every state. The project was under the guidance and supervision of Professor Larry Adams, who set the project milestones and requirements. I was responsible for forecasting Maine's outcome for the 2016 and 2020 elections.

The project was done in two phases. The first was to predict the results for the 2016 election. After verifying our data and results against what actually happened in the election, the second phase started. It was to include the new data that was generated in 2016 and use it to predict the results of the year 2020. Therefore, some charts and exercises in this book include 2016 data. Whenever possible, I collected any related historic data. For the prediction, I used historic election data going back to 1960.

I defined voters' groups by age, gender, education, demographics, and race. After studying the state from reliable academic sources, I identified issue categories like the economy, education, the environment, health care, and gun control.

Similarly, I listed the state's issues that would influence the presidential election by using the county ballot topics. Using the voting patterns of each party since 1960, poll accuracy, and the electoral votes, I tried different prediction methods and algorithms, such as Monte Carlo and Bayes, and statistical testing, such as T-test, chi-square, and others. Afterward, I had to compare my results to other forecast sites, like Five-Thirty-Eight. My prediction was correct for 2016.

This project was an exciting experience in which I converted cognitive features to numbers and crunched them to come up with results. Similarly, through other data science projects, I learned how to predict outcomes so as to drive decision making based upon measuring trends and studying patterns.

E. Fouda, *Learn Data Science Using SAS Studio*, https://doi.org/10.1007/978-1-4842-6237-5_1

Data Science Process

The data science process starts with forming a question or hypothesis, then collecting relevant raw data, then cleaning and exploring that data, then modeling and evaluating, then deploying, visualizing, and communicating results in reports, as shown in Figure 1-1.

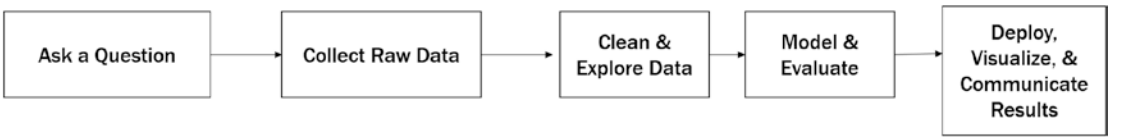


Figure 1-1. *Data science process*

Questions vary according to the field; for example:

- Politics: Will Trump win in Maine in 2016 and 2020?
- Facebook: How can you make people stay on Facebook longer?
- Medical: Is this tumor cancer or not?
- Hospital Management: How can you decrease patients' wait lines so as to increase patients' satisfaction?

The second step is collecting raw data. For example, in the politics question: Will a particular candidate win in a certain state?

Collecting all the voters' information—age, race, education, income, gender, and industry—is a crucial step, as is collecting the ballot data and voting results from over the years. The more historical data we have, the more accurate our predictions are. Furthermore, we should collect information on the population distribution throughout the years.

The third step is cleaning this raw data, from managing the missing values, outliers, repeated rows, and misspelled information, to adjusting the columns' data types, unifying the format of the values, and so on.

The fourth step is trying several models and comparing their results with each other, depending upon the problem's nature. In the presidential election problem, I used Monte Carlo and Bayes algorithms.

The fifth and final step is visualizing the results and communicating them in plain language in our reports. This step is the primary goal of the whole process because it holds the predictions to the answer to the first question that initiated the whole process.

Case Study: Presidential Elections in Maine

As I mentioned in the previous section, the data science process starts with a question. In this project, my question is: Will Donald Trump win in the state of Maine in the 2016 and 2020 presidential elections?

Population

The second step is collecting as much related data as possible. Therefore, I started with the population.

From information on the population distribution over Maine's counties, found at the U.S. Census Bureau, I learned that it is not uniformly distributed. There are vast areas that are either unpopulated or that have only one person living in them. While the red dots in the south look small, more than 5,000 people live in each of them. Therefore, I should not be deceived by the maps distributed by the presidential campaigns or by the mainstream media.

The following logical step was to get the voters' information. Some states publish their voters databases for free, and anyone could download them. However, in Maine, this was not the case. The state sold the voter databases to the political parties. So, I contacted the Secretary of State.

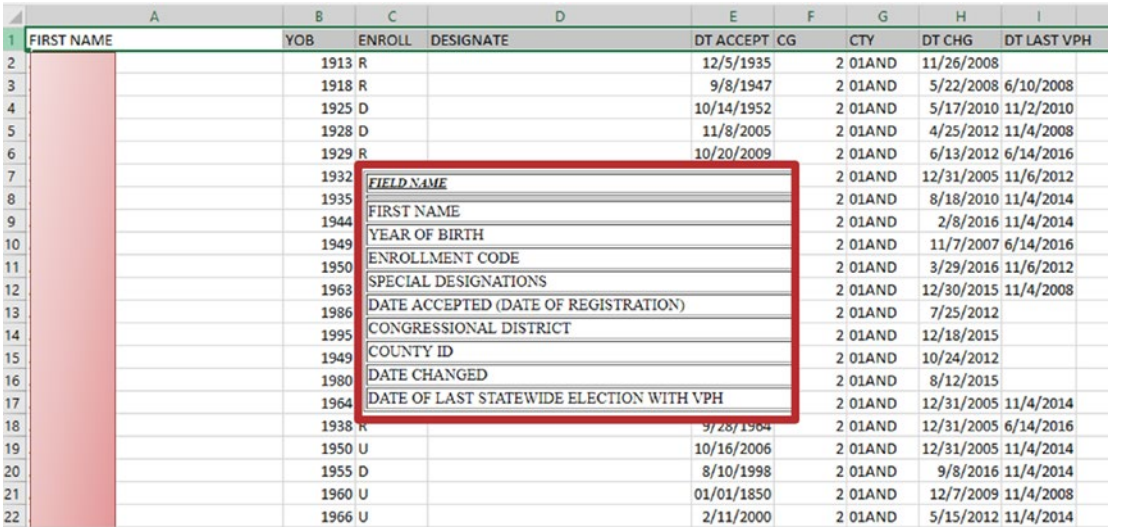
The office replied that to obtain voters files and updates from Maine's Central Voter Registration system, the requesting person or entity must be from the following five cases:

1. A candidate or person or entity working on a candidate's campaign
2. Someone working for a party
3. A person or entity involved in a referendum campaign that will be on the ballot in Maine in the next statewide election
4. A person or entity involved in specific get-out-the-vote efforts in Maine (the efforts have to be identified, including name, location, and date of events in Maine)
5. An individual who has been elected or appointed to and currently serving in a municipal, county, state, or federal office, but only for use for the official's authorized activities, not to turn over to another entity

The cost was based on the number of records obtained; the fee was scheduled in Title 21-A, section 196-A. A statewide voter file, which contained almost one million records, was $2,200.

After a few emails back and forth explaining that I needed them for a research project and sending some verifications, the office kindly sent me for free a DVD with all the required information, hiding the unneeded data like last names and so on.

The first table on the DVD has the voters' information and is shown in Figure 1-2. The columns are first name, year of birth, enrollment code, special designations, date of registration, congressional district, county ID, changed date, and date of last statewide election with VPH.



	A	B	C	D	E	F	G	H	I
1	FIRST NAME	YOB	ENROLL	DESIGNATE	DT ACCEPT	CG	CTY	DT CHG	DT LAST VPH
2		1913	R		12/5/1935	2	01AND	11/26/2008	
3		1918	R		9/8/1947	2	01AND	5/22/2008	6/10/2008
4		1925	D		10/14/1952	2	01AND	5/17/2010	11/2/2010
5		1928	D		11/8/2005	2	01AND	4/25/2012	11/4/2008
6		1929	R		10/20/2009	2	01AND	6/13/2012	6/14/2016
7		1932				2	01AND	12/31/2005	11/6/2012
8		1935				2	01AND	8/18/2010	11/4/2014
9		1944				2	01AND	2/8/2016	11/4/2014
10		1949				2	01AND	11/7/2007	6/14/2016
11		1950				2	01AND	3/29/2016	11/6/2012
12		1963				2	01AND	12/30/2015	11/4/2008
13		1986				2	01AND	7/25/2012	
14		1995				2	01AND	12/18/2015	
15		1949				2	01AND	10/24/2012	
16		1980				2	01AND	8/12/2015	
17		1964				2	01AND	12/31/2005	11/4/2014
18		1938	R		9/28/1964	2	01AND	12/31/2005	6/14/2016
19		1950	U		10/16/2006	2	01AND	12/31/2005	11/4/2014
20		1955	D		8/10/1998	2	01AND	9/8/2016	11/4/2014
21		1960	U		01/01/1850	2	01AND	12/7/2009	11/4/2008
22		1966	U		2/11/2000	2	01AND	5/15/2012	11/4/2014

FIELD NAME
FIRST NAME
YEAR OF BIRTH
ENROLLMENT CODE
SPECIAL DESIGNATIONS
DATE ACCEPTED (DATE OF REGISTRATION)
CONGRESSIONAL DISTRICT
COUNTY ID
DATE CHANGED
DATE OF LAST STATEWIDE ELECTION WITH VPH

Figure 1-2. *Voters' information*

The second table contains a registered and enrolled voters report, as in Figure 1-3. The columns of this table are the county name, municipality name, ward precinct, congressional district, state senate, county commissioner district, the party, and the total. The parties listed in the file are Democratic, Green Independent, Libertarian, Republican, and unenrolled.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	COUNTY	MUNICIPALITY	W/P	CG	SS	SR	CC	D	G	L	R	U	TOTAL
2	AND	AUBURN	1-1	2	20	62	5	625	121	28	355	622	1751
3	AND	AUBURN	1-1	2	20	64	5	115	20	5	139	143	422
4	AND	AUBURN	1-1	2	20	64	6	321	34	7	317	342	1021
5	AND	AUBURN	2-1							28	106	250	626
6	AND	AUBURN	2-1							22	383	630	1691
7	AND	AUBURN	2-1							9	287	383	1002
8	AND	AUBURN	3-1							34	100	289	676
9	AND	AUBURN	3-1							12	262	327	932
10	AND	AUBURN	3-1							9	348	543	1341
11	AND	AUBURN	3-1							1	84	138	336
12	AND	AUBURN	4-1							24	118	313	792
13	AND	AUBURN	4-1							3	147	265	665
14	AND	AUBURN	4-1							13	409	507	1364
15	AND	AUBURN	5-1							33	169	447	1115
16	AND	AUBURN	5-1							19	309	501	1339
17	AND	AUBURN	5-1							4	180	310	680
18	AND	DURHAM	1-1							15	967	1297	3297
19	AND	GREENE	1-1							28	903	1271	3257
20	AND	LEEDS	1-1							9	468	705	1743
21	AND	LEWISTON	1-1							29	353	501	1699
22	AND	LEWISTON	1-1							63	275	909	2629
23	AND	LEWISTON	2-1	2	21	59	2	1538	163	45	849	1169	3764
24	AND	LEWISTON	3-1	2	21	59	2	556	36	3	88	287	970
25	AND	LEWISTON	3-1	2	21	60	1	1429	233	202	359	1324	3547

FIELD NAME
COUNTY NAME
MUNICIPALITY NAME
WARD/PRECINCT
CONGRESSIONAL DISTRICT
STATE SENATE
STATE REPRESENTATIVE
COUNTY COMMISSIONER DISTRICT
DEMOCRATIC
GREEN INDEPENDENT
LIBERTARIAN
REPUBLICAN
UNENROLLED
TOTAL

Figure 1-3. *Registered and enrolled voters report*

This raw data was messy and contained many wrong values and outliers. For example, the age of one voter was 220 years, while his date of birth states that he was about 67 years old at that time. Some voters' information was missing, and so on. Again, as mentioned earlier, always clean your data: outliers, missing data, adjust data formatting, and explore your data.

Not only that, but also you should collect as much historical data as you can. So, I started digging and collected as much data as I could find. From the United States Census Bureau, I downloaded more tables (`http://www.census.gov/topics/public-sector/voting/data/tables.html`).

I grouped the voters by gender, age, and race.

Gender

As shown in Figure 1-4, I found that the registered female population is larger than the registered male population in Maine. However, the percentage difference for both genders is less than ±2%. As for how the candidates should use this information, the campaigns' representatives could wear the cancer awareness ribbon to play on women's compassion, as the women's turnout was always higher than the males'. This recommendation shows how data and numbers can control not only the speech topics but even what the representatives wear.

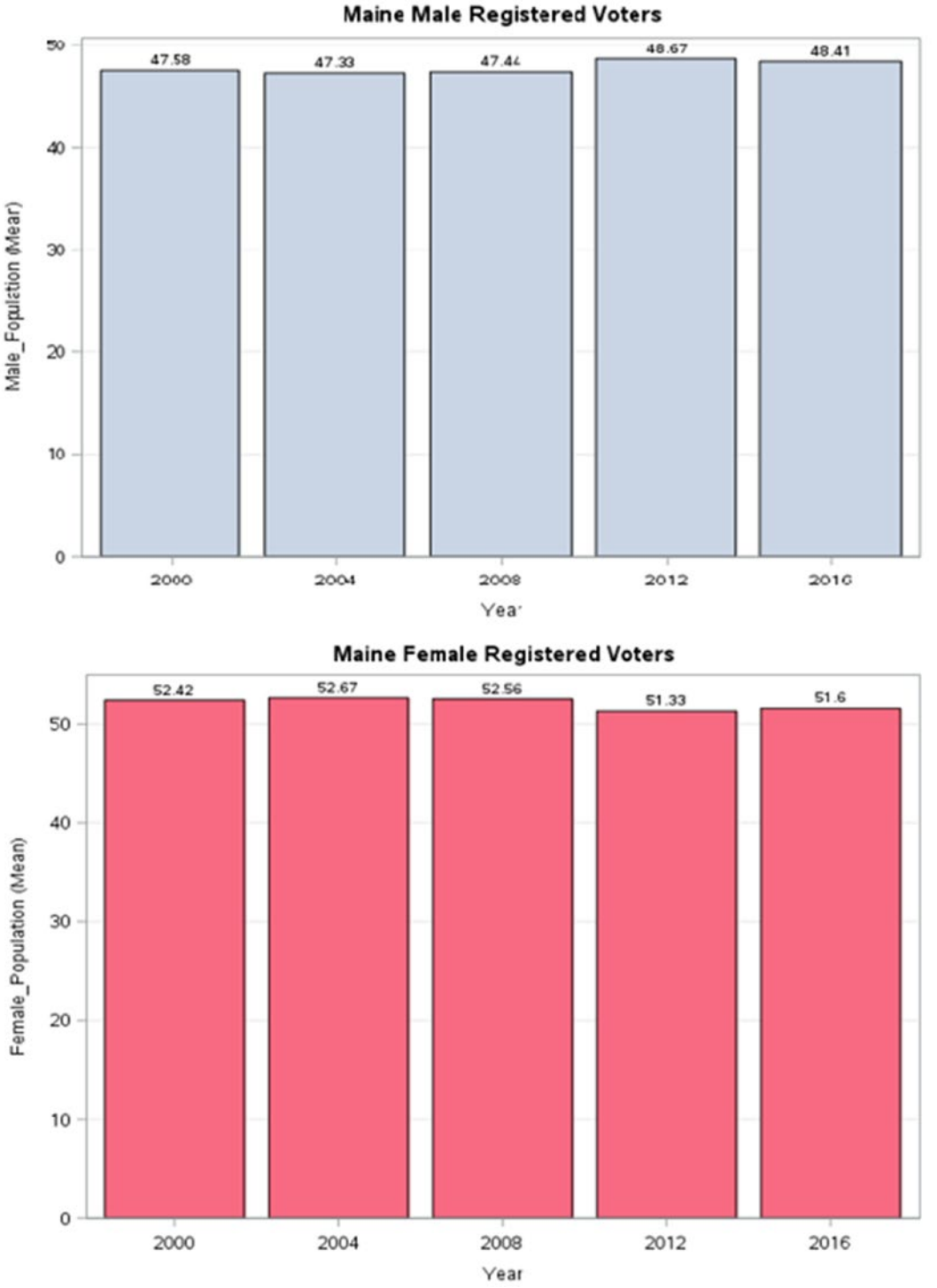


Figure 1-4. *Maine registered voters grouped by gender*

Race

Regarding race in Maine, more than 96 percent of the population is white (as shown in Figure 1-5). Therefore, I grouped the black, Asian, and Hispanic voters as non-white registered voters.

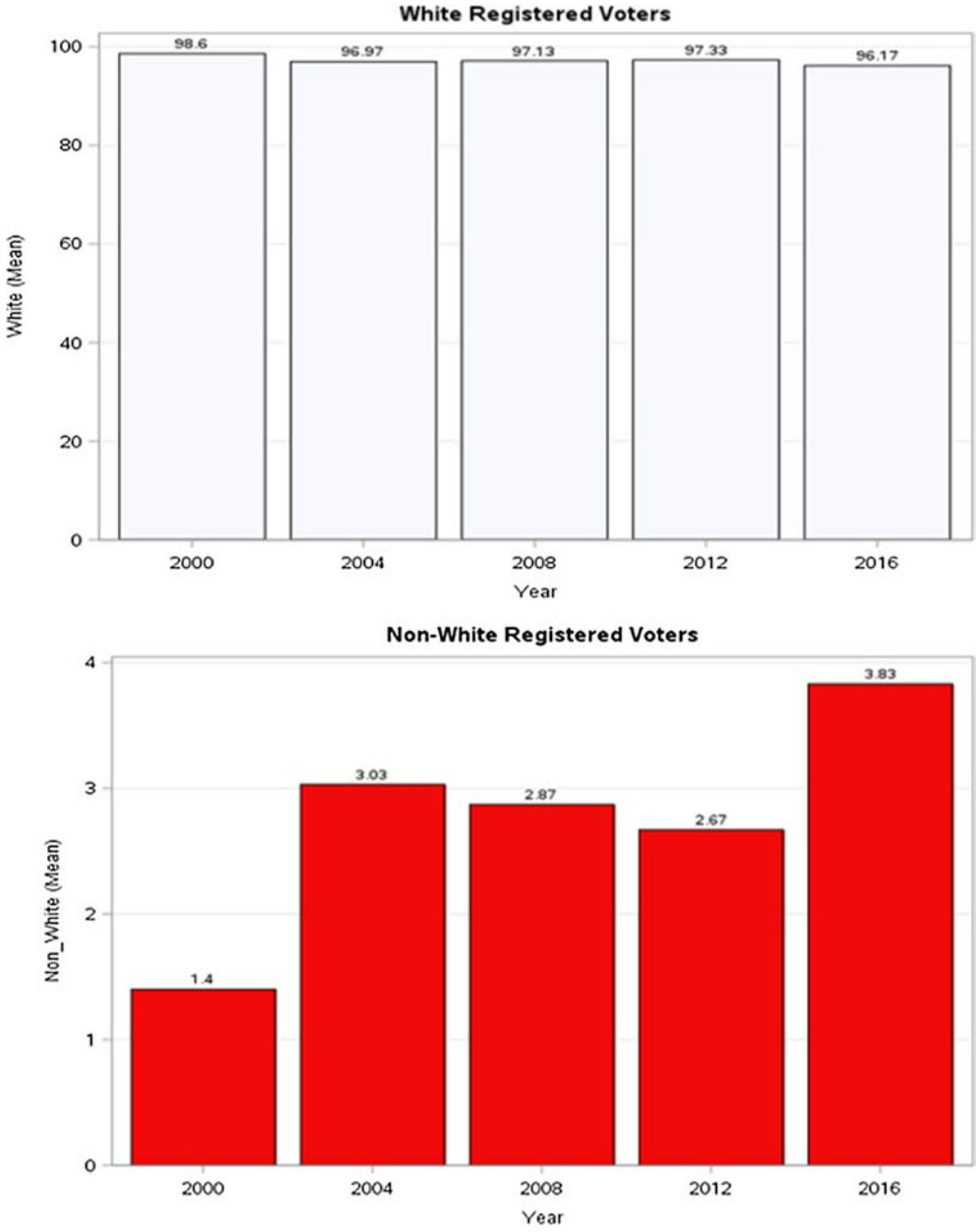


Figure 1-5. *Maine registered voters grouped by race*

Age

In 2016, there was a remarkable decrease in the number of registered voters who were over the age of sixty-five, as seen in Figure 1-6 (note that the scales are different). On the other hand, there was an increase in the 18–24 and 25–44 age groups. This finding indicates that the speech topics should change as well to convince more voters. For example, instead of focusing on medical insurance and retirement funding, the campaign representative should focus on student loans and home mortgages.

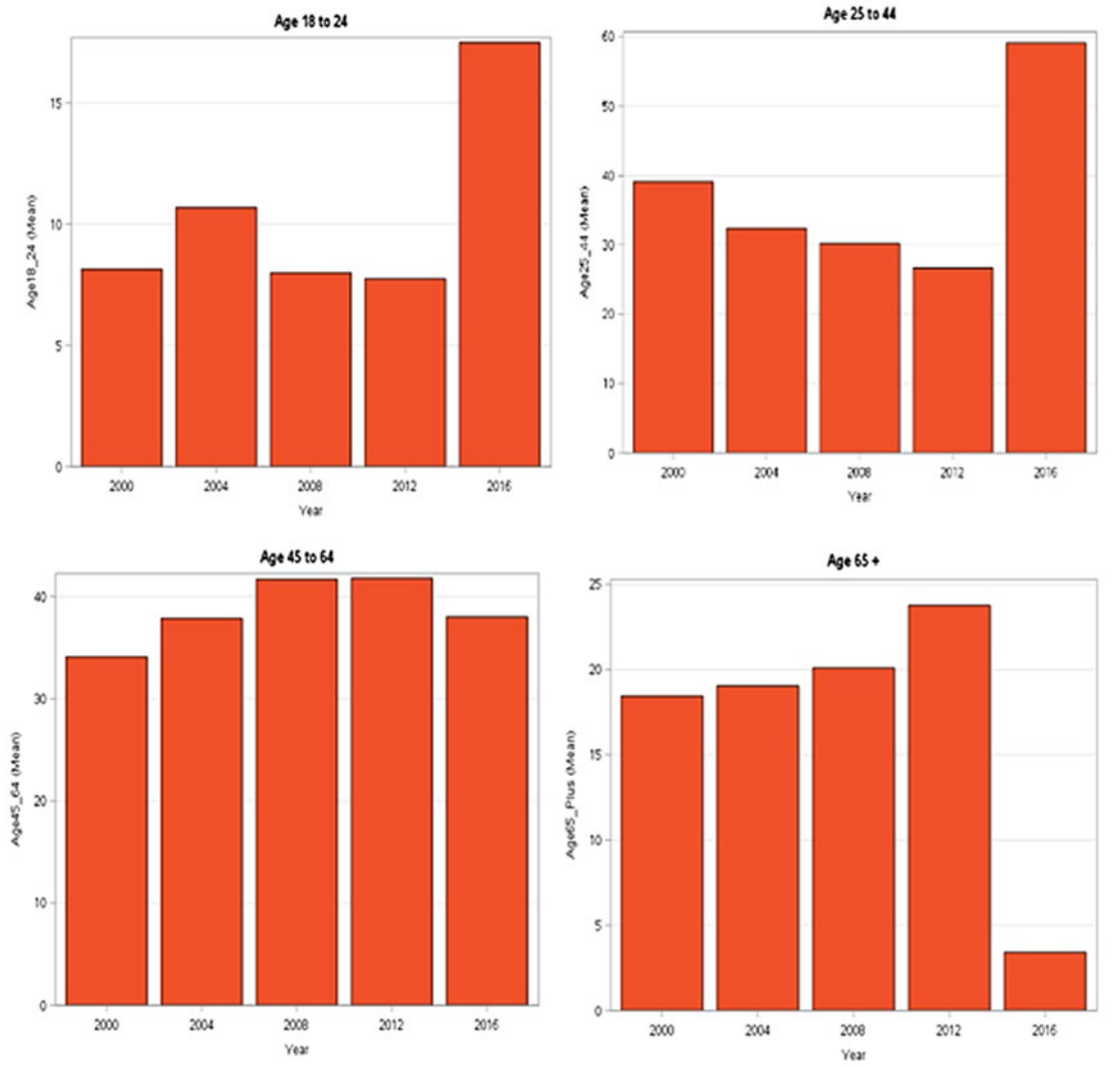


Figure 1-6. *Maine registered voters grouped by age*

Voter Turnout

I checked the voter turnout for past years to see how many voters got dressed and went out in Maine's snowy streets to vote. Figure 1-7 shows voter turnout from the years 2000 to 2016 as percentages, and that the winning party was the Democratic Party for those years. If you try this exercise with a swing party, the columns' colors will change to reflect the winning party.

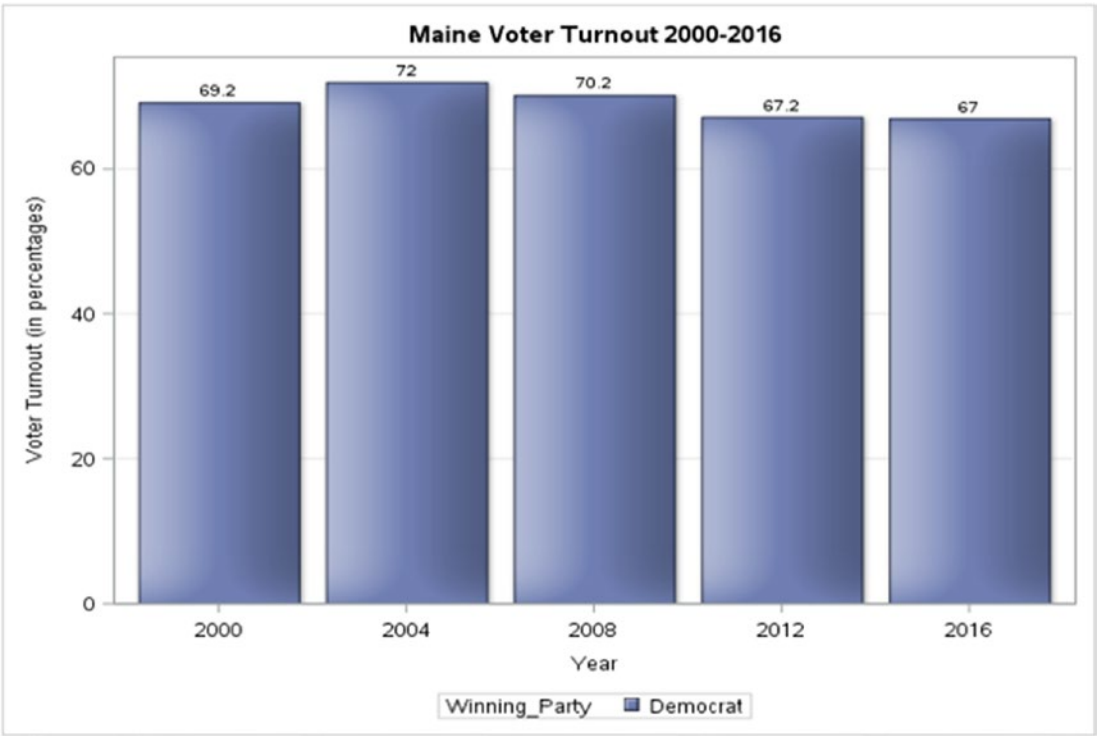


Figure 1-7. *Maine voter turnout 2000–2016*

Again, I inspected the voter turnout in recent years, according to the winning party and with age, race, and gender groupings. In Figure 1-8, the D represents the Democratic Party, the R stands for the Republican Party, and the O is for all other parties combined.