You should make sure to do a few things:

1. State a clear prediction question, not just what you did. Asking clear questions is the key to getting clear answers. You don't need to pick a "client", but sometimes it's helpful to think about the hypothetical scenario in which, say, NOAA or FEMA or a hospital would use your model.
2. Organize your discussion around two or three key tables or plots or numbers. Don't write up everything you did: Focus on your main result, and then add exposition around it until the reader is able to appreciate what you did and why. This keeps your writing from becoming bloated and rambling: Work backwards from the result to the reader's initial conditions.
3. You can handle criticisms or concerns in the results section if they are relevant or interesting, but most probably belong in the conclusion.
4. No pseudo-scientific word salad! If you write something, make it meaningful and interesting. Be clear. Make sure you know what you're talking about, or review the material to ensure you understand it.
5. Don't write about things about which you have no data or did not analyze, unless you are citing someone else's work in relation to what you are doing.
6. If you are tempted to make causal claims about a variable X causing Y, instead cast it as an explainability discussion. (e.g. "In this model, as yacht ownership increases, predictions of life expectancy tend to increase as well, possibly reflecting an omitted variable bias between wealth and health." You can of course discuss how yacht ownership and life expectancy are related, but of course yachts don't per se cause longevity.)

Remember that if you're working in a browser-based workbook like Colab, you can just right-click on an image and save it. I think you should just stick to notebooks and use markdown chunks for the writing, but if you want to use Google Docs or something like that, it's easy to bring tables over.

Please let me know if you have any other questions!

For our project we wanted to answer the question: “Which political party is most likely to win the majority of counties in each of the seven key battleground states (Arizona, Georgia, Michigan, Nevada, North Carolina, Pennsylvania, Wisconsin) in the 2024 U.S. presidential election based on demographic and economic indicators?”

This question emphasizes the specific outcomes being predicted (party winning counties) while anchoring the scope of the analysis to the demographic and economic factors used in the model. It also highlights the practical application of the model, potentially aiding political analysts, campaign strategists, or sociopolitical researchers in understanding the influence of these factors on electoral outcomes. We employed a supervised learning classification model, starting with decision trees and expanding to Random Forests to improve accuracy and robustness. Decision trees allowed us to identify key patterns and thresholds in the data that separated counties voting Democrat from those voting Republican. Random Forests further enhanced these predictions by combining multiple decision trees to mitigate overfitting and capture complex relationships between features.

The results of our analysis closely aligned with the actual 2024 election outcomes across all seven battleground states. Using our Random Forest model, we accurately captured the voting patterns in these states by leveraging demographic and economic indicators like race, education, and poverty rate. This high level of accuracy underscores the effectiveness of our approach in predicting county-level outcomes and demonstrates the stability of key factors influencing voting behavior. The close correspondence between our predictions and the actual results highlights the robustness of our model and the relevance of historical voting trends and demographic data in forecasting electoral outcomes.

Nevada

A state with 17 counties, Nevada did not pose many problems in constructing the random forest model. We merged data from three past election cycles to train and test the model. Some observations (county data for a specific election year) contained missing data for specific measures. We imputed values of 0 if they were insignificant/common (i.e. proportion of the population identifying with the Hawaiian race). Otherwise, we removed the rows.
Attached are the results of the random forest model. 16/17 counties were predicted to be Republican majorities for the 2024 U.S. presidential election. In comparison with the actual results, we found that our model had a success rate of 16/17 (94.11%). Clark County, while predicted to be Republican, turned out to be Democratic. This was the only discrepancy between the model’s predicted results and the actual results of the 2024 U.S. presidential election.

	County Name	predicted_party
257	Churchill County	REPUBLICAN
258	Clark County	REPUBLICAN
259	Douglas County	REPUBLICAN
260	Elko County	REPUBLICAN
261	Esmeralda County	REPUBLICAN
262	Eureka County	REPUBLICAN
263	Humboldt County	REPUBLICAN
264	Lander County	REPUBLICAN
265	Lincoln County	REPUBLICAN
266	Lyon County	REPUBLICAN
267	Mineral County	REPUBLICAN
268	Nye County	REPUBLICAN
269	Pershing County	REPUBLICAN
270	Storey County	REPUBLICAN
271	Washoe County	DEMOCRAT
272	White Pine County	REPUBLICAN
273	Carson City	REPUBLICAN

Arizona

As we built the random forest model, we merged the prior 3 past elections (2012, 2016, 2020) to train the model. However, the model ran into a problem of roughly 272 incidents of NaNs being found. To mitigate this error as we split the data, we decided to fill the na's with 0 and proceeded with the random forest. As our party variables were the variable of interest to predict, we created a label encoder to represent the Democrat and Republican party for each county. After running the model, Arizona's 2024 county winning parties were predicted and were almost identical to the actual results of the election.

	County Name	predicted_party
0	Apache County	DEMOCRAT
1	Cochise County	DEMOCRAT
2	Coconino County	DEMOCRAT
3	Gila County	REPUBLICAN
4	Graham County	REPUBLICAN
5	Greenlee County	REPUBLICAN
6	La Paz County	REPUBLICAN
7	Maricopa County	DEMOCRAT
8	Mohave County	REPUBLICAN
9	Navajo County	REPUBLICAN
10	Pima County	DEMOCRAT
11	Pinal County	REPUBLICAN
12	Santa Cruz County	DEMOCRAT
13	Yavapai County	REPUBLICAN
14	Yuma County	REPUBLICAN

There was only one county that differed from the prediction model to the actual outcomes which was Cochise County, predicted to be Democrat but was Republican in the 2024 presidential election. This resulted in our models having a success rate of 92.86% for Arizona (13/14).

North Carolina

	county_name	party
1300	ALAMANCE	REPUBLICAN
1301	ALEXANDER	REPUBLICAN
1302	ALLEGHANY	REPUBLICAN
1303	ANSON	DEMOCRAT
1304	ASHE	REPUBLICAN
...	...	...
1395	WAYNE	REPUBLICAN
1396	WILKES	REPUBLICAN
1397	WILSON	DEMOCRAT
1398	YADKIN	REPUBLICAN
1399	YANCEY	REPUBLICAN

The prediction question for the model is: How likely is each county in North Carolina to vote for the Democratic or Republican candidate in the 2024 Presidential Election, based on historical voting patterns and demographic data from 2012, 2016, 2020, and recent statistics? The key output of this code is a prediction table which provides predictions on the winning party for each county in North Carolina. The predictions align with historical trends, such as rural areas leaning Republican and urban areas leaning Democrat. The random forest feature importance analysis identifies top variables influencing predictions:

1. Males 65+ Percentage: Older populations tend to lean Republican
2. Bachelor's Degree Percentage: Higher education levels are often associated with Democratic outcomes in urban counties.
3. Black or African American Alone Percentage: Counties with larger Black populations may show stronger Democratic preferences due to historical voting patterns.
4. Median Household Income: Counties with a higher percentage of lower-income households could vote Democrat.
5. High School Diploma Percentage: See #2

Some limitations in our data are missing or imputed data that could introduce biases. There is also an absence of real-time confounders like polling or economic activity could limit the model's ability to adapt to current trends. The results align with the 2024 Presidential Election. Age, race, and education are significant predictors of county-level voting trends.

Michigan

To construct our random forest model, we merged data from the previous three elections - 2012, 2016, and 2020 - for training purposes. However, this process introduced over 1000 instances of missing values (NaNs) within the dataset. To mitigate this issue, linear interpolation was applied to fill gaps where possible for numeric columns. Persistently problematic rows which could not be adequately resolved through these methods were manually dropped for specific indices to ensure the dataset was fully cleaned. Remaining missing values in X_train and y_train were handled by replacing NaNs with zeros. The target variable was then converted to string type and encoded using LabelEncoder, and inconsistencies in labels were managed by mapping unexpected entries to valid options - DEMOCRAT" and "REPUBLICAN".

Overall, the model was accurate in predicting 80 out of 83 counties for the 2024 election. It did not predict that Marquette County and Leelanau County would go Democrat, nor did it predict that Macomb County would go Republican in 2024. However, the model was still very accurate, achieving a 96.39% success rate. As expected, urban areas, including counties surrounding Detroit, Grand Rapids, Ann Arbor, and Lansing, were predicted to vote Democratic, while more rural counties generally leaned Republican.

	County Name	predicted_party
174	Alcona County	REPUBLICAN
175	Alger County	REPUBLICAN
176	Allegan County	REPUBLICAN
177	Alpena County	REPUBLICAN
178	Antrim County	REPUBLICAN
..	...	...
252	Tuscola County	REPUBLICAN
253	Van Buren County	REPUBLICAN
254	Washtenaw County	DEMOCRAT
255	Wayne County	DEMOCRAT
256	Wexford County	REPUBLICAN

Pennsylvania

During this past presidential election, Pennsylvania emerged as a crucial battleground state with the potential to determine the next President of the United States. In 2020, after voting Republican in the 2016 election, Pennsylvania flipped blue. However, in the most recent election, the state returned to red. So, we thought it would be interesting to see how the model would predict the state's outcome of this latest election. To construct the model, we combined data from three presidential election cycles– including 2012, 2016, and 2020– alongside county data for each of the swing states. These were the results of some of the counties in Pennsylvania:

	County Name	predicted_party
374	Adams County	REPUBLICAN
375	Allegheny County	DEMOCRAT
376	Armstrong County	REPUBLICAN
377	Beaver County	REPUBLICAN
378	Bedford County	REPUBLICAN
379	Berks County	REPUBLICAN
380	Blair County	REPUBLICAN
381	Bradford County	REPUBLICAN
382	Bucks County	DEMOCRAT
383	Butler County	REPUBLICAN
384	Cambria County	REPUBLICAN
385	Cameron County	REPUBLICAN
386	Carbon County	REPUBLICAN
387	Centre County	DEMOCRAT
388	Chester County	DEMOCRAT

After running the model, it was found that 61/67 counties were accurately predicted. The remaining six counties, which the model initially predicted would vote Democrat, ultimately voted Republican (as indicated by the data from this past election cycle). These included Bucks, Cumberland, Erie, Lancaster, Monroe and Northampton counties. This meant that, for the state

of Pennsylvania, the model had a 91.04% success rate.

Wisconsin

	County Name	predicted_party
441	Adams County	REPUBLICAN
442	Ashland County	REPUBLICAN
443	Barron County	REPUBLICAN
444	Bayfield County	DEMOCRAT
445	Brown County	DEMOCRAT
446	Buffalo County	REPUBLICAN
447	Burnett County	REPUBLICAN
448	Calumet County	REPUBLICAN
449	Chippewa County	REPUBLICAN
450	Clark County	DEMOCRAT
451	Columbia County	REPUBLICAN
452	Crawford County	REPUBLICAN
453	Dane County	DEMOCRAT
454	Dodge County	REPUBLICAN
455	Door County	DEMOCRAT
456	Douglas County	DEMOCRAT
457	Dunn County	REPUBLICAN
458	Eau Claire County	DEMOCRAT
459	Florence County	REPUBLICAN
460	Fond du Lac County	REPUBLICAN
461	Forest County	DEMOCRAT
462	Grant County	REPUBLICAN
463	Green County	REPUBLICAN
464	Green Lake County	REPUBLICAN
465	Iowa County	DEMOCRAT
466	Iron County	REPUBLICAN
467	Jackson County	REPUBLICAN
468	Jefferson County	REPUBLICAN
469	Juneau County	REPUBLICAN
470	Kenosha County	DEMOCRAT
471	Kewaunee County	REPUBLICAN
472	La Crosse County	DEMOCRAT
473	Lafayette County	REPUBLICAN
474	Langlade County	REPUBLICAN
475	Lincoln County	REPUBLICAN
476	Manitowoc County	REPUBLICAN
477	Marathon County	REPUBLICAN
478	Marinette County	REPUBLICAN
479	Marquette County	REPUBLICAN
480	Menominee County	DEMOCRAT
481	Milwaukee County	DEMOCRAT
482	Monroe County	REPUBLICAN
483	Oconto County	REPUBLICAN
484	Oneida County	REPUBLICAN
485	Outagamie County	DEMOCRAT
486	Ozaukee County	DEMOCRAT
487	Pepin County	REPUBLICAN
488	Pierce County	REPUBLICAN
489	Polk County	REPUBLICAN
490	Portage County	REPUBLICAN
491	Price County	REPUBLICAN
492	Racine County	DEMOCRAT
493	Richland County	REPUBLICAN
494	Rock County	REPUBLICAN
495	Rusk County	REPUBLICAN
496	St. Croix County	REPUBLICAN
497	Sauk County	REPUBLICAN

The decision tree model was used to predict the dominant political party (Republican or Democrat) for Wisconsin counties in the 2024 election, with the actual 2020 outcomes serving as a baseline for comparison. The model performed very well, predicting that 13 counties would flip their party alignment in 2024, which was remarkably close to the actual 15 counties that flipped, achieving an 87% success rate for predicting flips. When applied to the 2020 outcomes, the model had a lower success rate, misclassifying several counties like Kenosha and Green, which voted Republican in 2020 but were incorrectly predicted as Democrat. The stronger performance in forecasting 2024 outcomes suggests the model was effective at identifying shifting voter patterns and adapting to recent trends. These results demonstrate its potential as a tool for predicting election outcomes, especially when combined with accurate data and fine-tuned parameters. To further enhance the model's accuracy, additional data sources such as voter demographics, economic changes, and turnout rates could be incorporated, making it even more reliable for future predictions.