Introduction:

In my experience as an Indian-American, racial discrimination, as a concept and reality, plays a continual role in shaping my vision for a better United States (U.S.) Many people like to describe the ethnic culture of the U.S. as a melting pot – an amalgamation of identities and beliefs, with each cultural flavor adding to a thoroughly blended society. From my perspective, the culture of the U.S. looks akin to the food pyramid – each culture must have involvement for a complete pyramid, but clear divisions exist between each group, and certain groups have greater focus. Given this perspective, and my experience as a political operative in campaigns, I seek to advance the political engagement of South Asian Americans to allow us equitable representation in electable office. This desire sparks a multitude of questions, each with the root of, "how do I do that?" How do I find out why proper representation for South Asian Americans does not exist? How do I connect South Asian Americans to form a cohesive voting block? How do I ensure optimal allocation of resources to achieve the greatest chance of electing a South Asian? These questions retain great importance, though their grand scope makes my answering of them for the purposes of a school assignment impractical. Therefore, I will focus on the achievable question of "where do current/past Indian-American politicians originate from, and do they come from geographic concentrations of Indians?" Although my true interests and goals involve all South Asians, the dearth of data on South Asian Americans forces me to focus on Indian-Americans, of whom more data exists.

The answer to this question holds great potential for electoral political operatives. If they had this information, they could determine where to locate a potential an Indian-

American candidate, and which communities to reach out to for support. Given the level of discrimination against ethnic minorities in the U.S. I expect that Indian-Americans who achieved electoral success needed to rely on their culturally-similar communities, so I hypothesize that the majority Indian-American office-holders come from Indian-American geographic concentrations. In this hypothesis, my independent variable is the state-level Indian-American population ratio, and my dependent variable is the origin states of Indian-American office-holders.

Data and Variables Description:

In order to address my hypothesis, I needed geographic data on Indian populations, nationally and by state. In addition, I needed a comprehensive list of Indians who held U.S. office. Unsurprisingly, data collection on Indian-Americans, and Indian-American politicians, has not occurred to any great extent. Geographic data on Indians exists, though not in temporal variety. For example, the data exists for 2010[1], when the last census was taken, though estimates for other years do not seem to exist. When it comes to a comprehensive list of South Asians, to my knowledge, none exists. Given that, I made my own. I collected the data through a mass amount of Googling. Certain sources, such as Wikipedia[2] and the Indian American Impact Fund[3], have more comprehensive, yet incomplete, lists of Indian-American politicians. They are incomplete in that they only contain information on a limited number of politicians, and most often, the information that is present only contains a name. To address the lack of information, I

[1] E. Hoeffel, S. Rastogi, M.O. Kim, H. Shahid, "The Asian Population: 2010, " *2010 Census Briefs.* (2012). https://www.census.gov/prod/cen2010/briefs/c2010br-11.pdf
[2] Unknown, "Category: American Politicians of Indian Descent," *Wikipedia.* Last updated August, 2019. https://en.wikipedia.org/wiki/Category:American_politicians_of_Indian_descent
[3] Unknown, "Elected Officials," *Indian American Impact Fund.* https://www.iaimpact.org/elected-officials

found the missing data through a variety of other sources. I also read numerous news articles to find politicians not listed elsewhere. Ultimately, for both sources of data, I used the incomplete information I found as a basis to develop my own datasets. Below, I go into detail on how I developed my two datasets: the first is the Indian population estimates, by state; and the second is the dataset on Indian-American politicians in the U.S.

To develop my Indian population estimates data, I first used 2010 census data on Asian Indians by state. In 2010, there were 2,843,391 estimated Indians in the U.S. Then, I found census data on the Asian Indian 2017 population estimate.[4] In 2017, there were 4,402,362 Indians in the U.S. From 2010 to 2017, the Indian American population rose by approximately 54.828%. I could not locate the 2017 population estimates by state so, using the 2010 to 2017 percent change, I estimated how many Indians there are by state. My 2017 estimate total came to 4,402,365, which off the official estimate by three people, or, .00006815%. While my total population estimate does not err much from the official estimate, I note that the state-by-state estimates may not reflect the reality of population change. I had to assume that the population ratios from 2010 held in 2017 (e.g. if 20% of Indian-Americans resided in CA in 2010, I assumed the same ratio for 2017). For the basis of my research, I used my own 2017 population estimate. The five states with the most Indians are: California, New York, New Jersey, Texas, and Illinois.

[4] Unknown, "Asian alone or in any combination by selected groups," *U.S. Census Bureau, American Fact Finder*, https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_17_1YR_B02018&prodType=table

To develop my Indian-American politicians dataset, I had to compile a list from numerous secondary sources. These sources include: Wikipedia, Indian American Impact Fund, numerous new articles and blogs. As stated above, most of the information I found only included a name. I compiled a list of names and then individually found information on: office title, district, state of residence, state of office, party affiliation, tenure, profession, gender, and incumbency status. I ultimately found 121 unique politicians who held 150 different offices. Some politicians held different offices throughout their political career. The five states with the most Indian-American politicians are: California, New Jersey, Texas, Washington, and Michigan.

After developing each dataset, I compiled the necessary data into a third dataset, which makes it easy for me to import and run functions on. This compilation dataset has the 2010 population, 2017 population estimates, the ratio of state count by total, and the politician state of origin count.

Descriptive Statistics:

Summary Statistics of Indian-American Compilation Data

Metric	Indian Population (2010)	Indian Population (2017)	Origin State Count
count	52	52	52
mean	54747.34615	84764.2107	2.903846
std	97893.64618	151566.7559	7.138022
min	589	911.936808	0
25%	5073.75	7855.584686	0
50%	13444	20815.07377	0
75%	61889.75	95822.65037	2.25
max	528120	817677.5333	37

Above is a table of summary statistics on my Indian-American compilation data. It has summary statistics on the Indian population in 2010, 2017, and on the count of politician origin state. The dataset includes all 50 states, and includes Washington D.C. and Puerto Rico. Therefore, 52 states + territories are counted. The 2010 mean Indian population in a U.S. state was about 54,747. In 2017, it rose to about 84,764. The average number of Indian-American politicians in each state is approximately 3. The greatest number of Indians in any state in 2010 was 528,120. In 2017, it was 817,678. The greatest number of Indian-American politicians in any one state was 37. All three metrics are based on California.

Initial Models:

Table 1: Summary of OLS Regression w/OriginStateCount Dependent and IndianPopulation2017 Independent

Metric	Intercept	IndianPopulation 2017
R-Squared	0.665	0.665
Coef	-0.3506	0.00003839
P-Value	0.600	0.000

Above, I show a summary table of a standard OLS regression[5] with the count of Indian politicians in each U.S. state as my dependent, and the Indian population in 2017 as my independent. Note that in the dependent, states that never had an Indian politician are also included. Prior to running the model, I generally expected the data to fit reasonably well (r-squared ~ .5), for a 0-value intercept, for a relatively large IndianPopulation2017 coef, and for non-statistical significance. The actual results surprised me. The data fit better than expected, with an r-squared of .665. The intercept was worse than expected, with a negative coef of -.351. I expected the intercept to be 0 because if there are no Indians, there wouldn't be any Indian-American politicians. Interestingly, this is not statistically significant as it has a p-value of .6, meaning there is

[5] Python Team, "Python 3.7." (2018), https://www.python.org/downloads/release/python-370/

60% chance this relationship exists by chance. The IndianPopulation2017 coef was lower than expected, with a miniscule coef of 3.839e-5, but after having run the model, it now makes sense. It means if the Indian population were to increase by 1 more person, there would be 3.839e-5 more Indian politicians in the U.S., on average. Given that the Indian population in 2017 is 4,402,365, and the Indian-American politicians population is only 151, the large disparity in values should result in a small, but positive, coef. This is highly statistically significant, with a p-value of 0.

In developing a second model, I thought to reduce the massive disparity in values between the Indian population of 2017 and the number of Indian-American politicians. In raw values, the population size is about 29,000x larger than the number of politicians. Instead of observing the raw values of population, I decided to run a regression model with a logged population value. Logging scales the mean of the data to 0, reducing the variance and making the relationship between dependent and independent variables more observable. By using the logged population value, I expected many of the same results, but with a higher independent variable coef. The results somewhat match my expectations.

Table 2: Summary of OLS Regression w/OriginStateCount Dependent and Logged 2017 Indian Population

Metric	**Intercept**	**IndianPop17_logged**
R-Squared	0.294	0.294
Coef	-20.190	2.285
P-Value	0.000	0.000

Interestingly, the r-squared lowered from .665 to .294. I expect this is because there is less variance in the data, but the data scales away from the regression line. The coef of the intercept went down considerably, with a coef of -20.190. This means that, on average, if there are no Indians in the U.S., then there would be -20.19 Indian-American politicians. Again, the intercept coef does not match my expectations, nor understanding. Unlike the first model, this intercept coef is highly statistically significant, with a p-value of 0. As expected, the coef of the logged Indian population rose from 3.389e-5 to 2.285. This means, on average, if the Indian population in 2017 rose by one percent, there would be 2.285 more Indian-American politicians. This is highly statistically significant, with a p-value of 0.

I was curious if I would observe a higher independent variable coef if I only observed the politician count of U.S. states that have produced an Indian-American politician. Therefore, I subset the OriginStateCount variable on two conditions – 1) if an Indian-American politicians has originated from the state, and, 2) if an Indian-American politician has not originated from the state. I ran an OLS regression with the logged 2017 Indian population as my independent variable. Because I would focus on states that have produced Indian-American politicians, I expect the model to fit the data better. I also expect the coef of the Indian population to rise.

**Table 3: Summary of OLS Regression
w/StatesWIndianPoliCount Dependent and
Logged 2017 Indian Population**

Metric	**Intercept**	**IndianPop17_logged**
R-Squared	0.339	0.339
Coef	-1.273	0.174
P-Value	0.001	0

The results somewhat match my expectations, but where it did not match, I erred considerably. The model did fit the data better, with an r-squared rising from .294 to .339. The intercept coef increased from -20.190 to -1.273. This is highly statistically significant, with a p-value of .001. The coef of the 2017 Indian population fell from 2.285 to .174. This means that if the Indian population in the U.S. rose by one percent, there would be a rise of .174 Indian-American politicians, on average. This is highly statistically significant, with a p-value of 0. I suppose the coef dropped because I subset the dependent variable size. With less states observed, there wouldn't be as large of a rise in the number of politicians if the population increased. Additionally, I'm only observing states that *already* produced Indian-American politicians. Therefore, in economics terms, the marginal product wouldn't increase as much by just adding more Indians to Indian-influenced states.

Next, I wanted to observe a logistic model to observe the logit relationship between U.S. states who produced Indian politicians and the logged 2017 Indian population. When I set the conditions on StatesWIndianPoli, all states which produced an Indian-American politician are set to 1, and states that didn't are set to 0. By establishing the dependent variable as binary, I can run a logistic regression. I expect that as the 2017 Indian population rises, the logit of Indian-American politicians, in states that already produced some, rises as well.

Table 4: Summary of Logistic Regression w/StatesWIndianPoli Dependent and Logged 2017 Indian Population

Metric	**Intercept**	**IndianPop17_logged**
Df Residuals	50	50
Coef	-10.131	0.991
P-Value	0	0

As shown in the table above, the results match my expectations. The fit of the model on the data is shown through the residuals of 50. The coef of the Indian population is .991. This means that for every additional percentage point rise in the Indian population, the logit of there being Indian-American politicians rises by .991, on average. This is highly statistically significant, as indicated by the p-value of 0. Again, the coef of the intercept is a negative value, at -10.131. This is also highly statistically significant, with a p-value of 0.

Following the logit model, I wanted to see the odds-ratio from my logistic model. When calculating the odds-ratio, I took the ratio of the (probability of success) / (probability of failure). In other terms, (p/(1-p))*100.

Table 5: Summary of Odds Ratio

Metric	**Probability**	**Odds-Ratio**
Intercept	0.00004	0.004%
Indian Pop. 2017 Logged	2.69487	158.99705

The odds-ratio for the intercept is .00004. It indicates that without any Indians in the U.S., the odds of an Indian-American politician existing is .004%. The odds-ratio of the logged Indian population is 2.69487, which indicates that the odds of an Indian-American being a politician, given the presence of Indians in the U.S. goes up by 159%.

Next, I wanted to observe the predicted probability that an Indian-American politician comes from one of the top 5 most Indian-populous states. I used Greg's logit to probability function and ran the function on the logits of each of the states. To try to make it specific to each state (couldn't figure out how to subset the function to each state based on their index), I multiplied the logits by the number of politicians: total number of politicians ratio (I don't believe this is correct and do not endorse anyone trying it). As stated prior, the five states with the most Indians are: California, New York, New Jersey, Texas, and Illinois.

Table 6: Predicted Probabilities		
States	**Raw Value**	**Predicted Probability**
California	5.079e-5	.005079%
New York	4.063e-5	.004063%
New Jersey	4.980e-5	.004980%
Texas	4.338e-5	.004338%
Illinois	4.090e-5	.004090%

According to the function, the probability that an Indian-American politician comes from: California is .005079%, New York is .004063%, New Jersey is .004980%, Texas is .004338%, and Illinois is .004090%. I expect my method of multiplying the logits by the politicians ratio significantly lowered the probabilities, though the probabilities do match up with the ranking of politicians per state. For example, California has produced the most Indian-American politicians, with 37, and it returned the highest probability.

Final Models/Conclusions:

Of all the models I ran, my OLS regression model with the dependent variable of count of politicians in states that have produced Indian-American politician, and independent variable of the logged 2017 Indian population. Table 3 summarizes those results. The values are highly statistically significant, and they somewhat explain the relationship between the number of Indian-American politicians in the U.S. and the total population of Indians in the U.S. While I feel confident the results are accurate (in terms of how I used the data I have), the model does not ultimately explain my original hypothesis – that the majority of Indian-American politicians come from geographic concentrations of Indians.

The model that does address my hypothesis is my logistic regression with predictive probabilities. Table 6 summarizes the results. The model, in theory, should help me explain the probabilities that Indian-American politicians come from a particular state location. However, I ran into multiple errors as I tried to run the model. The biggest hindrance I encountered was trying to run a function with logits specific to a state. I couldn't figure out how to subset the data for that, so instead, I multiplied the logits by the state's politician ratio. This resulted in extremely low probabilities. Moving forward, I would try to figure out how best to use the specific data that I need. To do that, I would have to better acquaint myself with the code engineering behind python and statsmodels.

Although I could not successfully address my hypothesis, conducting this research provided valuable insight in how to conduct future research. As part of this project, I had to: 1) Compile my own dataset. This required me to research multiple sources online and learn about the state of Asian-American politicians in the U.S. By doing so, I believe I compiled the most comprehensive dataset of South Asian-American politicians that exists. 2) I had to manipulate my dataset in order to have workable models. Doing so forced me to think about what data I had, and how I could expand on it to have more useful features. 3) I also learned how to run regression/logistic models on data that I personally put together, which is a very different practice than working with data I found online. This forced me to think about what data I had and what questions that data can help me answer. 4) I also learned how to interpret my models and understand when my models simply weren’t proper. Many of the models I ran either did not run or did not actually address the question I was trying to answer. Ultimately, my hypothesis was not falsified, or not falsified, but I learned how to put together a research project, and how to better accomplish it moving forward.

Bibliography:

[1] E. Hoeffel, S. Rastogi, M.O. Kim, H. Shahid, "The Asian Population: 2010, " *2010 Census Briefs.* (2012). https://www.census.gov/prod/cen2010/briefs/c2010br-11.pdf

[2] Unknown, "Category: American Politicians of Indian Descent," *Wikipedia.* Last updated August, 2019. https://en.wikipedia.org/wiki/Category:American_politicians_of_Indian_descent

[3] Unknown, "Elected Officials," *Indian American Impact Fund.* https://www.iaimpact.org/elected-officials

[4] Unknown, "Asian alone or in any combination by selected groups," *U.S. Census Bureau, American Fact Finder*, https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_17_1YR_B02018&prodType=table

[5] Python Team, "Python 3.7." (2018), https://www.python.org/downloads/release/python-370/

Note on bibliography:

1) I used Prof. Greg's materials as a basis for much of my understanding
2) I started Indian politicians dataset as part of a personal project, which eventually became my final project for class. As such, I did not note every website and news article I read to obtain information on every politician I found, though I made an effort to cite the major sources.