EAS 345

Section A

Bayesian Improved Surname and Geocoding (BISG)

Sander O’Neil

Sep 26

# 

# 

# 

# 

# 

# 

# Table of Contents

[**Table of Contents 2**](#_exzgy94knbcg)

[**Summary 3**](#_ghzg6u6ezsm)

[**Introduction 4**](#_camr2whk5cbn)

[**Background 5**](#_8x4kt3800d2c)

[**Plan Of Work 5**](#_r04t5ph3ee3i)

[**Qualifications And Experience 6**](#_cm5ccvun2wgf)

[**Schedule 8**](#_8u2db59njuab)

[**Conclusion 8**](#_uqixff8bhmz4)

[**References 9**](#_jjrks4j9yal6)

d

D

D

D

D

# Summary

Bayesian Improved Surname Geocoding (BISG) is a statistical model that applies Bayes' Theorem to find the race/ethnicity of a person using their surname and geolocation.

Bayes’ Theorem also known as Bayes’ Rule is simply stated as:

This can be read as The probability of A given B, is the probability of B given A times the ratio of A to B.

For instance,Given a dataset of people from NY state, if 4 percent of people live in Buffalo, 5 percent of people are fans of the Buffalo Bills, and if 80 percent of people in Buffalo root for the bills, what is the probability that a fan of the bills is from buffalo?

This basic rule is used and modified by using Bayesian inference. Bayesian inference is a modification which allows for the use of previous data points to attempt to find conclusions on new data. For the example of BISG, previous data for Surnames and racial demographics are used as a prior probability (A) and geodata are used as a likelihood (B) where the data combined is used as a posterior probability (P(B|A)). This is used in an iterative way to find probabilities and confidences of racial data based on surname and location.

This statistical method can be used to give racial categorization or data to datasets which are missing this point but have surname and geolocation data. This data is important if an organization wants to find out the demographics of a group for purposes like analyzing discrimination in hiring, insurance or healthcare outcomes.

People may choose not to fill out demographic information because they are worried that they will be discriminated against if they do.

Analyzing data without taking into account those who have not given their information may lead to overlooking bias in the gathering of demographic data, for example, the Census attempts to report on data for illegal immigrants but many do not fill out census forms because they fear the information may be given to U.S. Immigration and Customs Enforcement's (ICE), which in fact it is not.

In recent years companies have been getting in serious legal trouble for discrimination due to biases in hiring using automated hiring algorithms. In New York City the law even requires an annual audit of algorithms for fairness (“Local Law 144”).

At the same time algorithmic auditing is a growing business and requires racial and demographic data in almost every job. The study of discrimination requires advanced statistical methods because discrimination is an emergent property of large numbers of people and small but selective forces working through small individual biases.

In the coming years it will become necessary to not only stop bias when it arises in a plain faced manner, but also to seek it out using data analytics to stop discrimination before it manages to affect the demographics of a corporation, or customer base.

# Introduction

The problem Bayesian Improved Surname and Geocoding is trying to solve is incomplete racial/ethnic data for the purpose of finding racial biases in sample data, possibly during an algorithmic audit.

New York City (“Local Law 144”) Requires a bias audit of automatic hiring algorithms annually. Lack of racial/ethnic for the people hired will cause a delay in auditing and may cost companies delays or legal fees.

The purpose of this proposal is to patch holes in data using existing knowledge of Surname and Geographic location. This should allow for efficient addressing of random missing data or even systematic loss of data, such as when people choose not to fill out demographic information for themselves on hiring forms.

The source of information will mainly be U.S. census data <https://data.census.gov/> (United States Government, n.d.) They have a free tool for projects like this, and it is really the only source that can be expected to have this sort of information without worries of racial profiling.

The scope of this proposal is to be able to categorize any collection surnames and locations into a certain ethnicities and confidence levels, within the continental United States. This will not apply to Non-US locations and will definitely not be 100% accurate, but will hopefully at least give an indication of low confidence. Gender would be a very useful demographic to have but impossible because gender does not vary by surname, and very little by geolocation.

In order to define terms I will include the following section.

Bayesian Improved Surname Geocoding (BISG) is a statistical model that applies Bayes' Theorem to find the race/ethnicity of a person using their surname and geolocation.

Bayes’ Theorem also known as Bayes’ Rule is simply stated as:

This can be read as The probability of A given B, is the probability of B given A times the ratio of A to B.

For instance,Given a dataset of people from NY state, if 4 percent of people live in Buffalo, 5 percent of people are fans of the Buffalo Bills, and if 80 percent of people in Buffalo root for the bills, what is the probability that a fan of the bills is from buffalo?

Bayesian Inference is a method of statistical inference which uses multiple data points with their own uncertainties in order to find the probability of an event. It is different from frequentist inference because the hypotheses have probabilities assigned to them as well.

# Background

Since the rise of algorithms being used to categorize people they have been soaking up inherent bias in pre-existing datasets. Whenever any algorithm is trained on pre-existing data the pre existing data is at risk of contaminating the new algorithm with its bias. Example: G.E. trains a hiring algorithm on data of who it hired since 1920. It will likely notice that the employees in its dataset were only men for multiple decades and make decisions on that information from now on. Or accidental bias. Example: A Michigan based company uses their hiring algorithm in Texas, and the algorithm continues to hire only people from Michigan.

Recently government entities have begun to regulate this bias, but it is difficult as many statistical models have no exact number which represents their likelihood to hire a woman or an asian person, instead these hiring algorithms are complete black boxes unknown even to the data scientists who create them.

# Plan Of Work

Most of the work done for this project will be building the final model. This includes testing and debugging, but some other work is required. Data will need to be collected from the census and interviewing of clients would be useful as well.

Firstly the data needs to be sourced from the U.S. census, the data should include enough demographic precision that 19,000 incorporated Cities and Towns (*Number of U.S. Cities, Towns, Villages by Population Size 2019*, 2023) can be distinguished from each other but that may be impossible, It may be easier to get information on the around 3,000 towns and cities with populations above 10,000.

The precision of the demographic data should be standard U.S. Census precision, so race as one of the following: Black, White, Asian, Native/Pacific Islander, and Hispanic Status should all be collected.

Racial data per surname should be collected, This will not be difficult. (*Decennial Census Surname Files (2010, 2000)*, 2016)

The most difficult part of data collection will be Surnames Geodata correlation, part of this is because the census likes to protect the privacy of peoples data and so does not release data that could lead to personal conclusions. For instance if there is a town with only one person with a surname it is unlikely the census would want to provide the racial data for that surname as it would basically be giving racial data on a specific person.

The next stage of the project would be data analysis and cleaning. It would be important to analyze the data gained so far to see if it will fit the scope of the project, and if the project should be modified or more data gathered. Data may need to be cleaned for processing. Towns could possibly have their data quantified from text data.  
 The next stage would be the initial building of the model. Based on the structure of the data gained so far Posterior probability should be built and all data should be summed upon the correct axes to convert statistics into probabilities.

The bulk of model building will deal with predicting race from the other probabilities found so far using Bayesian Inferencing.

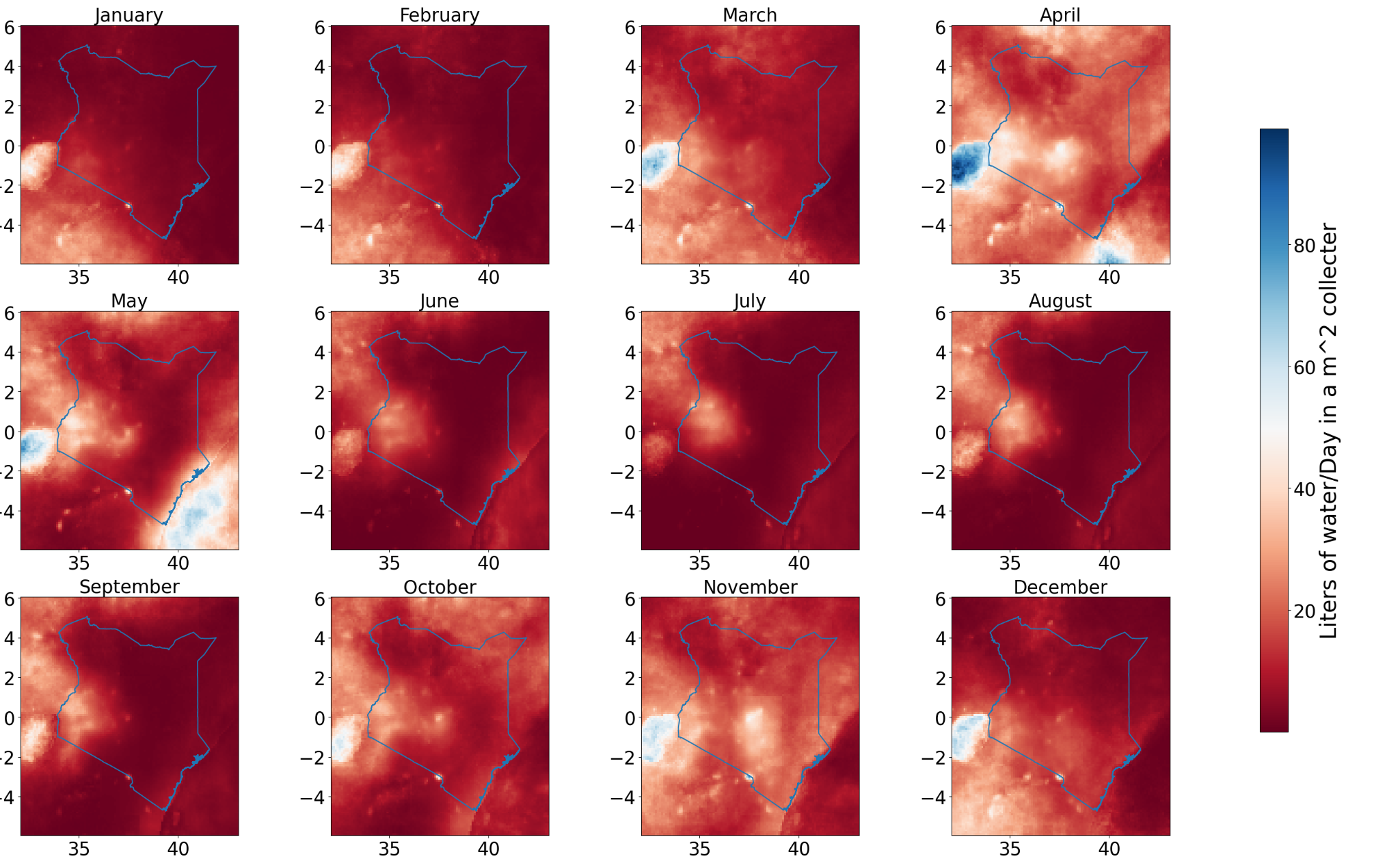
Once the model has been built it should be evaluated. It should be especially evaluated against naive average assumptions. For instance assuming everyone is white non hispanic would be accurate most of the time, and furthermore assuming everyone is always the race which holds a majority in their town would be even more accurate. The model must be more accurate than this to be useful.

If the model evaluates well than it should be coded into a form which can be easily used by a client like a single command line python function which asks for a 2 by n array of [surname, zip code] and returns a 2 by n array of [racial predictions, confidence level]

# Qualifications And Experience

On a personal note, My mother is Catherine O’Neil, the CEO of ORCAA, an algorithmic auditing corporation which works on this sort of thing. My grandma, Elizabeth O’Neil, is thinking about coding a model similar to this. I asked them what a good Idea was for this project and they suggested BISG. They often attempt to evaluate models for racial bias and not having racial data can often be a detriment, so this kind of algorithm could assist in their work.

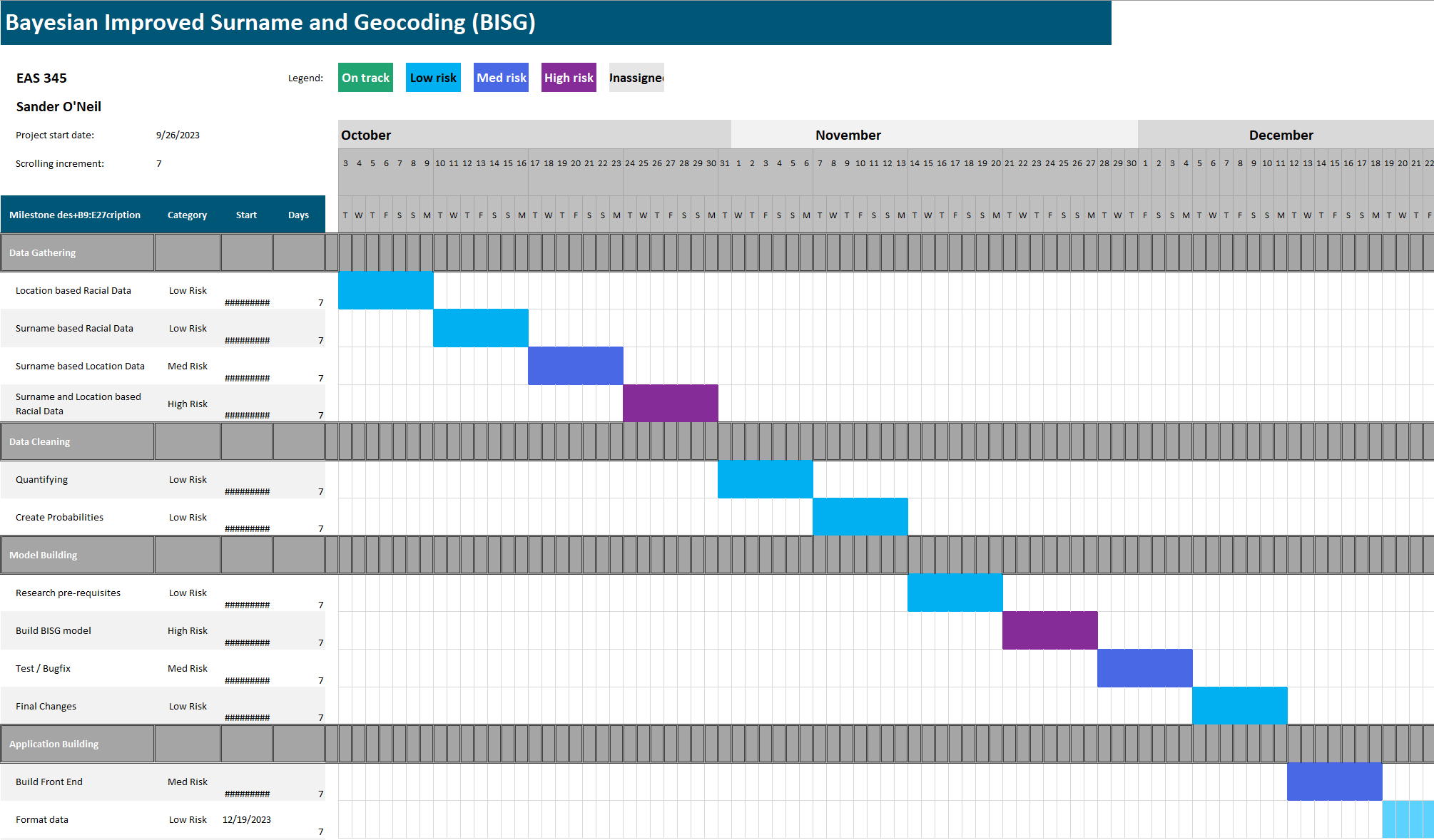
I have qualifications for this project, I have used numpy for a number of coding projects before <https://github.com/Sander-ONeil>, including my Kenya Rainfall Study <https://github.com/Sander-ONeil/Kenya_Rainfall_Study>

which made heavy user of numpy as well as a NASA weather dataset in order to find ideal locations for rain collection in the dry season

Figure, Previous project utilizing Nasa dataset with numpy

I believe I would be able to create my own model by myself, without collaboration. I will consult with my family at the company for certain design constraints.

# Schedule



# Conclusion

Bayesian Improved Surname and Geocoding (BISG) is a useful algorithm to create guesses/assumptions as to the racial or ethnic information of an individual which has not given their information explicitly. This can be helpful in cases where there is a possibility of discrimination which should be tested for but there is incomplete data on the demographics possibly being discriminated against.

In recent years algorithmic auditing has become a legally mandated profession and will only get more necessary for corporations going forward. Because of anti-discrimination laws, companies are often not allowed to collect demographic data about their employees or clients. Still it is sometimes necessary to assess the demographics of outcomes of processes like hiring insurance and healthcare.

# References

*Decennial Census Surname Files (2010, 2000)*. (2016, December 15). U.S. Census Bureau. Retrieved September 26, 2023, from https://www.census.gov/data/developers/data-sets/surnames.html

*The New York City Council - File #: Int 1894-2020*. (2021, 11 12). The New York City Council - Calendar. Retrieved September 26, 2023, from https://legistar.council.nyc.gov/LegislationDetail.aspx?GUID=B051915D-A9AC-451E-81F8-6596032FA3F9&ID=4344524&Options=&Search=

*Number of U.S. cities, towns, villages by population size 2019*. (2023, June 2). Statista. Retrieved September 26, 2023, from https://www.statista.com/statistics/241695/number-of-us-cities-towns-villages-by-population-size/

ORCAA. (n.d.). ORCAA. Retrieved September 26, 2023, from https://orcaarisk.com/

United States Government. (n.d.). Census Bureau Data. Retrieved September 26, 2023, from https://data.census.gov/