

Synthetic Population Generation Using Census Data

by

Abu Darda

23141038

Rahat Khan

20101212

Ashabul Yamin Raad

23141088

A pre-thesis 2 report of the thesis submitted to the Department of Computer
Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
May 2023


© 2023. Brac University
All rights reserved.

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

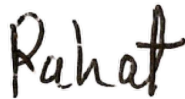
Student's Full Name & Signature:



Abu Darda
23141038

Yamin Raad

Ashabul Yamin Raad
23141088



Rahat Khan
20101212

Approval

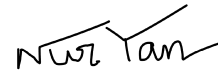
The thesis/project titled “Synthetic Population Generation Using Census Data” submitted by

1. Abu Darda (23141038)
2. Rahat Khan (20101212)
3. Ashabul Yamin Raad (23141088)

Of Summer, 2023 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on August 23, 2015.

Examining Committee:

Supervisor:
(Member)



Dr. Mohammad Nur Yanhaona
Associate Professor
Department of Computer Science and Engineering
Brac University

Co-supervisor:
(Member)



Rafeed Rahman
lecturer
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi
Chairperson
Department of Computer Science and Engineering
Brac University

Abstract

A synthetic population is instrumental in modeling a specific region’s large group of people, encompassing demographics, decomposition, and activities. Currently, there are two closed-source population simulators (USA and Germany) that enable demographic modeling and contact tracing. Unfortunately, this lack of accessibility prevents a wide range of individuals from utilizing the simulation. Consequently, people remain unaware of methods to predict heavy traffic congestion or implement disaster management precautions. Additionally, existing map services solely offer traffic projections without providing a detailed view of land usage. To address these limitations, our team is actively developing an open-source population simulator that generates synthetic populations. This simulator aims to provide both individuals and states with access to projections, forecasts, and models of diverse population behaviors under specific stimuli, such as disease outbreaks, natural disasters, and significant congestion. Leveraging US census data, the simulator’s focus is on the geographic area of the US, as relevant data is predominantly available in Western countries.

Keywords: synthetic population, simulation, US census data, Public use Meta-data, open-source simulator, microscopic view.

Table of Contents

Declaration	i
Approval	ii
Abstract	iii
Table of Contents	iv
List of Figures	vi
List of Tables	vii
Nomenclature	viii
1 Introduction	1
1.1 Research Motivation	1
1.2 Problem Statement	1
1.3 Research Objective	2
2 Literature Review	3
2.1 Background Study	3
2.2 Related Works	4
3 Description of the Data	9
3.1 Census Data	9
3.2 PUMS Data	9
3.3 Analysis of Initial Data	10
3.4 NHTS Data	12
3.4.1 Household Data	12
3.4.2 Personal Data	13
3.4.3 Trip Data	13
3.4.4 Vehicle Data	14
4 Description of the Model	15
4.1 Iterative Proportional Fitting (IPF)	15
4.2 Iterative Proportional Updating	16
4.3 Monte Carlo Sampling	17
4.4 Contact Tracing Using Social Networks	18
4.5 Prediction of trip purpose	19
4.5.1 Regression Models	19

4.5.2	Classification Model	19
4.5.3	Neural Network	20
5	Preliminary Analysis	21
5.1	Analysis of Sampled Data	21
5.2	Finding the Best Model	22
5.2.1	Classification Model	22
5.2.2	Challenges	24
6	Conclusion	25
	Bibliography	27

List of Figures

2.1	The details of the person as well as the number of vehicles of a census tract. The darker the block is, the higher the difference between simulation and reality	4
2.2	Netanya's Synthetic Population	5
2.3	Synthetic Population of Dortmund.	6
2.4	The comparison between three prominent papers on synthetic population.	8
3.1	Percentage of ages	10
3.2	Percentage of races	11
3.3	Percentage of sexes	11
3.4	Cross tabulation of PUMS data	12
3.5	Samples from Different States of USA	13
3.6	Barchart of Primary Activity in Previous Week in NHTS Data Sample	14
5.1	Sampled Data	21
5.2	Crosstabulation of Sampled Data	21
5.3	Applying Neural Network to Predict Generalized Purpose of the Trip	23
5.4	Comparison of Models	23

List of Tables

5.1	Regression Models accuracy	22
5.2	Classification Models on Predicting Medical Condition	22
5.3	Classification Models on Predicting Generalized Purpose of Trip . . .	22

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

ACS American Community Survey

IPF Iterative Proportional Fitting

IPU Individual Proportional Updating

PUMS Public Use Metadata Sample

Chapter 1

Introduction

1.1 Research Motivation

The COVID-19 pandemic posed challenges for countries in contact tracing, particularly in densely populated nations like Bangladesh, India, and Brazil. These countries lacked a clear understanding of how to prioritize areas and take necessary actions. Additionally, the vaccination process proved ineffective during subsequent waves as governments struggled to track the movements of vaccinated and unvaccinated individuals. Long-term lockdowns were also unaffordable for developing countries. Recent natural disasters, such as floods in Pakistan and Bangladesh, as well as wildfires in Australia, highlighted the need to comprehend population demographics during crises and for future planning.

While population synthesis has been used in various instances, including Dortmund and Netanya in the US, none of the existing sources are freely accessible to the public. Furthermore, the accuracy of these simulations relies on advanced datasets that are unavailable in many countries. Consequently, effective population management during natural disasters and in heavily congested cities remains unresolved, perpetuating recurring problems.

Our proposed simulator shares similarities with an existing one[10]. Our philosophy is rooted in the belief that the people and the government have the right to understand the overall demographic landscape. This understanding facilitates wiser choices in terms of living, working, and improving quality of life. Moreover, during times of crisis, it enhances awareness for both the government and the public. By providing an open-source solution, we aim to encourage other countries to acquire the necessary requirements, such as proper datasets and logistical support for the people, to embrace the concept of population synthesis.

1.2 Problem Statement

Accurate population forecasting is crucial for effective disaster management, traffic congestion control, and urban planning. While services like Google Maps offer predictions, they only provide short-term forecasts. Additionally, there is a lack of open-source simulators that can give advanced warnings beyond traffic congestion in specific geographic areas. Consequently, governments and individuals in underdeveloped countries cannot understand the demographics during times of crisis, high congestion in densely populated cities, or in locations like hospitals and amusement

parks. Moreover, the only available simulators are limited to the USA and Germany and are not open-source.

Our simulation process comprises four steps: creating a baseline population, assigning activities, determining location choices, and estimating contacts.

While some researchers have explored different methods for creating a baseline population, such as Iterative Proportional Fitting (IPF) and Monte Carlo Sampling, we have chosen to work with IPF and Individual Proportional Updating (IPU) methods, depending on their accuracy, to assign individuals to households. Additionally, we will employ a probabilistic function and Hausdorff distance for precise location assignment, matching individuals to their homes. While some studies have used social networking for contact tracing, we will leverage various datasets to estimate contacts in different settings such as schools, hospitals, and workplaces. This approach will allow us to obtain more realistic contact estimations instead of measuring them in a holistic manner.

1.3 Research Objective

Our objective is to develop an easily comprehensible simulation, focusing on the following steps to clarify our vision:

- We aim to create a simulator that offers a detailed view of land usage and the population within specific entities like corporations, schools, and amusement parks for a defined time period. This microscopic view will enhance interpretability.
- Our simulator will project a contact network among individuals, enabling us to obtain accurate estimations of interactions during times of crisis.
- We intend to expand our dataset coverage while considering fewer features, if possible, to enhance the flexibility and accessibility of our simulator.

Our goal is to build a simulator that provides free population forecasts for everyone. This will enable individuals to access demographic information within a particular location and take appropriate actions based on it. Unlike current services that primarily focus on traffic congestion, our simulator will offer a synthetic population representation of various service and transportation sectors. By reducing reliance on location data sharing, people can have more organized and precise actions during adverse situations.

Chapter 2

Literature Review

While developing our research process, we encountered several scholarly papers discussing topics such as IPF, IPU, Monte Carlo Sampling, social networks, and simulation models. In this section, we will provide a concise overview of the background studies related to the methods and algorithms we intend to utilize in our research. Additionally, we will provide a brief summary of the relevant publications we have reviewed.

2.1 Background Study

In this section, we have incorporated theoretical insights into the models and techniques we plan to employ, as well as the approaches explored by other researchers in their respective projects. This section will delve into topics such as Iterative Proportional Fitting, Monte Carlo Sampling, Iterative Proportional Updating, and Social Networks, providing a comprehensive understanding of these concepts.

- Iterative Proportional Fitting (IPF) is an algorithm for estimating the proportions of a population in different categories. It is an iterative algorithm that starts with a set of initial estimates and then repeatedly adjusts the estimates until they converge to a solution. IPF is a popular algorithm for estimating population proportions because it is relatively easy to implement and it is typically very accurate.
- Monte Carlo Sampling is a technique for generating random samples from a population. It is a powerful tool for statistical analysis because it can be used to estimate the probability of events, calculate confidence intervals, and perform hypothesis tests. Monte Carlo sampling is often used in conjunction with other statistical techniques, such as IPF, to improve the accuracy of estimates.
- Iterative Proportional Updating (IPU) is an algorithm for updating the proportions of a population in different categories. It is an iterative algorithm that starts with a set of initial estimates and then repeatedly adjusts the estimates based on new information. IPU is a popular algorithm for updating population proportions because it is relatively easy to implement and it is typically very accurate.

- Social Networks are networks of people who are connected to each other by social ties. Social networks can be used to study a variety of phenomena, such as the spread of information, the formation of social groups, and the influence of social norms. Social networks are a powerful tool for understanding human behavior and they are increasingly being used in a variety of research fields.

2.2 Related Works

In 1996, Beckman pioneered the initial and most prominent approach to generating a synthetic population, consisting of two main components: adjusting a multi-way demographic table through Iterative Proportional Fitting (IPF) and constructing a synthetic population[2][3]. Notably, this method does not address activity assignments. Initially, a proportional multi-way demographic table is estimated through an iterative process of proportional fitting. Subsequently, a synthetic population of households is derived from the Public Use Microdata Sample (PUMS) in order to align with the proportions outlined in the estimated table.

This research paper introduced several fundamental concepts to population simulation. Firstly, it introduces the utilization of Iterative Proportional Fitting for the first time with PUMS data, enabling the creation of datasets containing detailed information on households and individuals within specific geographic locations.[8] To clarify, working with the PUMS dataset is impractical as it is a sample that includes randomly selected individuals and lacks data on every person in the geographic area. By employing marginal values, a dataset is generated to ultimately represent a projection of the entire population and their vehicle statuses. Secondly, the resulting dataset obtained from running Iterative Proportional Fitting may suffer from sparsity. Consequently, this paper presents a solution to this issue by incorporating a minimal number prior to executing the Iterative Proportional Fitting process.

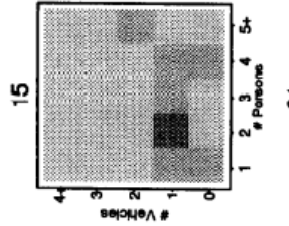


Figure 2.1: The details of the person as well as the number of vehicles of a census tract. The darker the block is, the higher the difference between simulation and reality

This research paper employed the Public Use of Microdata Sample (PUMS) data to simulate household details, specifically focusing on persons and vehicle proportions. To achieve this, the paper utilized a proportion of household details and vehicle status, which were fitted using Iterative Proportional Fitting (IPF). Subsequently, the ratios of these proportions and the Mean Absolute Deviation were compared using IPF and PUMS. Additionally, this paper allowed for the independence of any statistical method deemed suitable. However, it is worth noting that this study did not address activity assignments or location choices. Consequently,

unlike the other two articles, it fell short of providing a comprehensive synthetic population as it solely concentrated on household data and simulation.

During the 8th International Conference on Computers in Urban Planning and Urban Management (CUPUM) in May 2003, the first synthetic population models were introduced[6]. These models focused on simulating household populations using IPF and Monte-Carlo Sampling for data preprocessing[2]. Furthermore, in the subsequent layer, the researchers examined household workers and allocated a zone for each operational activity. The study subsequently created two synthetic models: one for Netanya, Israel, and another for Dortmund, Germany.

Initially, IPF adjusted the probability of household size based on the age of the household head[2]. Then, by employing Monte Carlo Sampling, household details were obtained by incorporating relevant attributes. After several iterations, various details pertaining to the household were derived, such as the household head, non-adult members, the number of cars, and incoming persons.

Taking the simulated population of Netanya as an example, it generated a population comprising approximately 159,000 individuals residing in approximately 50,000 households within the Netanya area. To construct each household, the head of the household was selected as a foundational element, as many household characteristics depend on the attributes of this particular family member, such as age, gender, religion, and education level. Subsequently, the available data was utilized to estimate the household size. In cases where a one-person household was chosen, there was no need to add any additional individuals. However, if a multi-person household was selected, additional members were sequentially added until the entire household was formed. The residential location within the zone was then determined, which involved specifying the micro-location using the row and column coordinates of a raster cell. Next, the number of income earners within the household was determined, followed by the estimation of household income. All of this information was then used to estimate the number of cars owned by the household. Lastly, a workspace was allocated to each employed individual. The simulator continued to generate new households until all households within the zone were established.

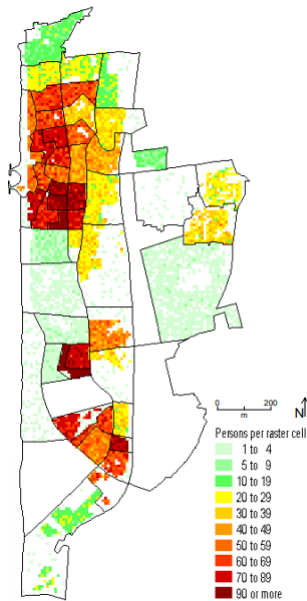


Figure 2.2: Netanya's Synthetic Population

Subsequently, the simulation extended to Dortmund. Notably, the primary limitation of Netanya’s synthetic population lies in its focus solely on household simulation. This restriction arises due to the availability of less comprehensive data and more constrained location datasets compared to Dortmund. In contrast, the location datasets for Dortmund were more extensive and permissive, enabling the projection of various land uses, including business, residential, and non-residential zones. As a result, the simulation conducted in Dortmund proved to be more precise and valuable in its outcomes.

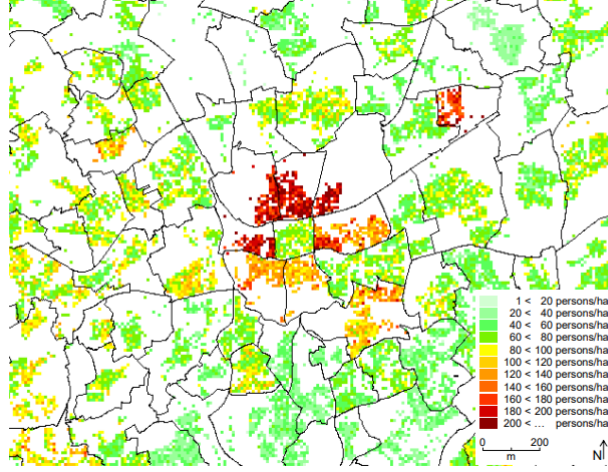


Figure 2.3: Synthetic Population of Dortmund.

When comparing the synthetic populations of Netanya and Dortmund, it becomes evident that the Dortmund population model represents individual actors through both households and household members. In addition to accounting for the number of vehicles, the Dortmund model differentiates among eleven vehicle categories, considers ownership of monthly season tickets, generates car-sharing memberships, and produces a monthly mobility budget. Furthermore, unlike the Netanya model, the Dortmund model depicts each individual in the city based on their employment status and possession of a driver’s license. While Netanya displays revenue on a household level, Dortmund independently generates income for each individual. Consequently, the simulation conducted in Dortmund offers a more comprehensive and accurate representation.[2]

As previously mentioned, our simulator aligns with an existing model titled ”Generating a synthetic population of the United States.” This paper emphasizes four crucial aspects: Baseline Population Synthesis, Activity Assessment, Location Choice, and Contact Estimation. These four layers serve as the fundamental framework for our thesis. Initially, the Baseline Population Synthesis focuses on creating an initial dataset and transforming it into an observable dataset using IPF (Iterative Proportional Fitting) and IPU (Iterative Proportional Updating) algorithms. Subsequently, IPF scales the PUMA (Public Use Microdata Areas) dataset to the PUMS (Public Use Microdata Sample) dataset, making it applicable to the entire region. Ultimately, IPF and IPU, which have been mathematically validated, are employed to scale individual and household datasets.

Moving on to the next layer, activities are assigned to individuals from their respective households. Data is gathered from sources such as the National Household

Travel Survey (NHTS), Dun & Bradstreet (D&B), and HERE (formerly NAVTEQ) [4]. The paper highlights the use of Hausdorff Distance for calculating person-person distances, favoring it over Euclidean and Mahalanobis distances. Ultimately, the researchers identify the worst person-person distance as the household distance. In summary, selecting a survey household, identifying the best matching individuals within the household, and assigning activities to each individual constitute the three essential components of this process. Kristian Lum et al. [12] proposed an optimized method for activity assignment, which comprises three straightforward steps:

- Identify the nearest household in the survey to the synthetic household.
- Determine the closest individual in that household to the synthetic individuals.
- Assign the activity schedule.

The similarity between households is evaluated using two metrics: probability and minimum distance. The Hausdorff Distance is employed to calculate the distance between two households. The researchers prioritize Hausdorff Distance over Euclidean Distance, Manhattan Distance, and the Fitted Value Approach due to its ability to encompass most covariates.

This method holds significance as it simplifies activity assignments by allowing the derivation of real-world activity factors from survey data. Furthermore, activities are derived from statistical information rather than being treated as a sequential timeline.

Regarding location choice, the probabilistic assignment of locations to selected individuals is carried out. The HERE (formerly NAVTEQ) datasets are utilized to measure the distance between households and the preferred areas of these individuals. A specific function is employed to calculate the probability, which takes into account factors such as workplace population, travel behavior, economic class, and retailer surveys. [5]

$$\begin{aligned}\Pr(j \mid i) &\propto w_j e^{bT_{ij}}, \\ \Pr(k \mid j, i) &\propto s_k e^{bT_{jk} + aT_{ki}},\end{aligned}$$

To streamline calculations and prevent the need for population computation across the entire location, the researchers utilized Traffic Analysis Data (TAZ). They first selected the TAZ locations, followed by assigning the capacity to each TAZ location by aggregating the corresponding geographic area.[1] The process of contact estimation was then performed by integrating location choice and activity duration. Consequently, a table of vertices representing individuals and edges representing social connections was generated. Applying this method to estimate contact patterns in Delhi and Los Angeles yielded highly accurate results.[1]

Steps	Paper 1: Creating Synthetic Baseline Populations by Richard J. Beckman et al.	Paper 2: Creating a Synthetic Population by Rolf Moeckel et al.	Paper 3: Generating a synthetic population of the United States(2015) by Abhijin Atigya et al.
Baseline Population	IPF	IPF, Monte Carlo Sampling	IPF, IPU
Activity Assignment	This paper does not deal with activity assignments.	Ignores travel behavior and commercial data	Works with employment status, travel behavior, healthcare population and Land Use Data.
Location Choice	This paper does not deal with location choice	Only assigned to workplaces	It deals with amusement parks, healthcares along with educational institutions and workplaces
Contact Tracing	This paper does not deal with contact tracing.	This paper does not deal with contact tracing.	It shows the individuals along with other individuals they have been contacted with.

Figure 2.4: The comparison between three prominent papers on synthetic population.

Chapter 3

Description of the Data

3.1 Census Data

The term "Census dataset" refers to a comprehensive collection of data and information obtained through a national census conducted by a country's government. [17] A census is a systematic process aimed at gathering a wide range of demographic, social, economic, and housing data about the population residing within a specific geographical area. This dataset encompasses various key aspects such as age, gender, race, education, employment, income, and housing.

Our research utilized the American Community Survey (ACS) census data collected over a period of five years. The ACS, conducted by the United States Census Bureau, is an ongoing survey that gathers extensive information on demographics, social factors, economic indicators, and housing statistics pertaining to the American populace.

Unlike the decennial census, which occurs once every ten years, the ACS continuously collects data throughout the year. By sampling a smaller fraction of the population each month, it accumulates a larger and more comprehensive sample size over the five-year timeframe. This approach ensures that the ACS provides a detailed and accurate representation of the population, enhancing the reliability and robustness of the collected data.

The ACS encompasses millions of households across the United States, covering a broad spectrum of topics. These include but are not limited to age, gender, race, ethnicity, education, employment, income, and housing conditions. The survey generates estimates for various geographic levels, ranging from states and counties to cities and smaller localized units. Consequently, it serves as a valuable resource for understanding local communities, conducting research, and gaining insights into the evolving societal and economic dynamics.

3.2 PUMS Data

In the United States, the Census Bureau conducts a national census known as the United States Census, which captures data on the entire population. [17] Furthermore, the Census Bureau provides additional datasets derived from the census data, one of which is the Public Use Microdata Sample (PUMS) dataset.

The Public Use Microdata Sample (PUMS) dataset is a subset of the complete census dataset, containing anonymized individual-level records. PUMS data allows

researchers and analysts to access detailed information about individuals and households while ensuring data confidentiality. It provides a representative sample of the population, enabling customized analysis and research on various socio-economic characteristics.

ur study incorporated the Public Use Microdata Sample (PUMS) data specific to the state of South Dakota. The PUMS dataset, derived from the American Community Survey (ACS), provides anonymized individual-level information that allows for detailed analysis of various socio-economic factors.

The PUMS data, unlike aggregated statistics, offers a more granular perspective by providing individual-level records for a representative sample of households within South Dakota. This dataset includes comprehensive information on demographics, education, employment, income, housing, and other relevant variables.

3.3 Analysis of Initial Data

The percentage of ages in the column shows the distribution of the population by age. This information can be used to understand the needs of the population and to plan for services and programs.

The percentage of ages in the column can be used to understand the needs of the population for services such as education, healthcare, and transportation. For example, if the percentage of children in the population is high, then there will be a need for more schools and childcare facilities.

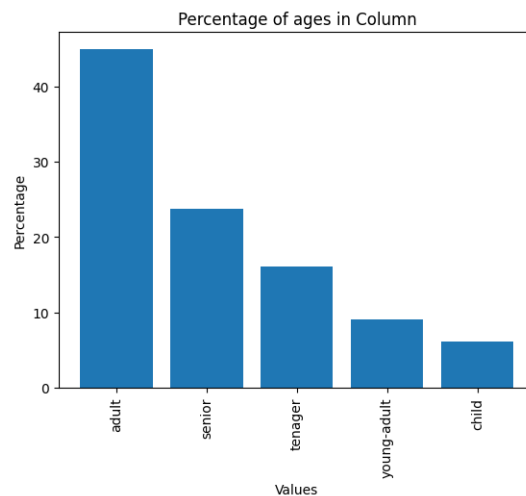


Figure 3.1: Percentage of ages

The percentage of races in the column shows the distribution of the population by race. This information can be used to understand the diversity of the population and to plan for services and programs that meet the needs of all groups.

The percentage of races in the column can be used to understand the needs of the population for services such as language translation, cultural awareness training, and access to ethnic food and businesses. For example, if the percentage of Hispanic people in the population is high, then there will be a need for more Spanish-speaking staff and resources.

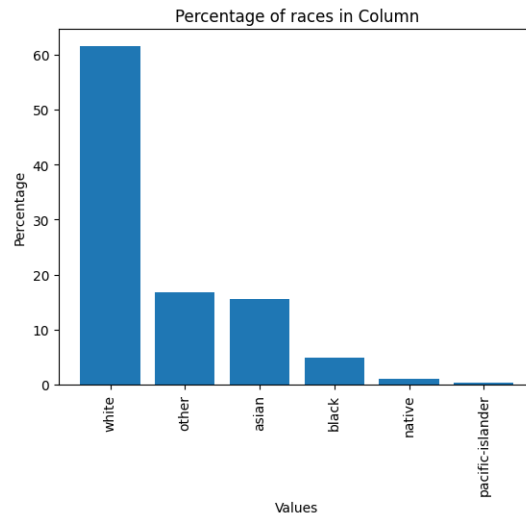


Figure 3.2: Percentage of races

The percentage of sex in the column shows the distribution of the population by sex. This information can be used to understand the needs of the population and to plan for services and programs that meet the needs of both men and women.

The percentage of sex in the column can be used to understand the needs of the population for services such as reproductive health care, domestic violence shelters, and men’s health clinics. For example, if the percentage of women in the population is high, then there will be a need for more women’s healthcare providers and resources.

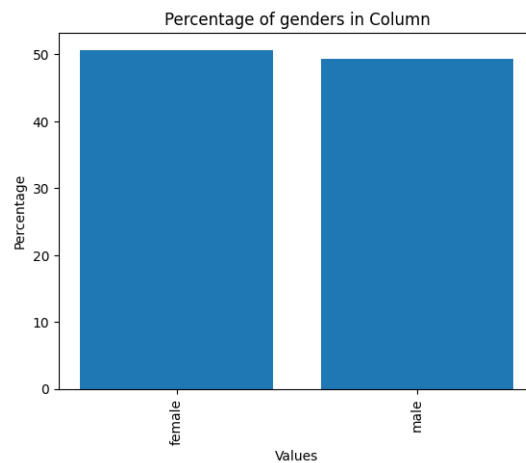


Figure 3.3: Percentage of sexes

The findings from the PUMS data can be used to inform a wide range of decisions, from planning for services and programs to understanding the needs of the population. This information is valuable for businesses, governments, and organizations that want to serve the needs of all members of the community.

age	adult		child		senior		tenager		young-adult	
gender	female	male	female	male	female	male	female	male	female	male
race										
asian	77445	66742	7425	7980	37571	29021	19800	20902	12627	13047
black	20684	22054	2383	2435	12548	9971	6689	7353	4219	5096
native	4074	4166	613	596	1995	1648	1592	1703	825	897
other	69512	70587	14692	15256	18566	15705	35630	37630	18625	19660
pacific-islander	1556	1536	192	199	560	474	516	493	353	394
white	253536	256443	31668	33313	171761	148193	83393	87548	45569	49800

Figure 3.4: Cross tabulation of PUMS data

The cross-tabulation shows that the distribution of the population by age, sex, and race is not evenly distributed. The majority of the population is white, followed by others, Asian, black, and so on. The majority of the population is also female, followed by males. The distribution of the population by age is more evenly distributed, with the majority of the population being between the ages of 25 and 60.

3.4 NHTS Data

The NHTS means National Health and Travel Statistics. It consists of four datasets: Household Data, Personal Data, Trip Data, and Vehicle Data. The four datasets observe the same people. Here, people are identified with Household Identifier and Person Identifier.

3.4.1 Household Data

Household Data has 115 features and 129696 Samples. Therefore, it has the information of 129696 households, not persons. The features include Household Identifier, Travel Day - day of Week, Primary Sampling Stratum Assignment, Home Ownership, Count of household members, Count of household vehicles, Household income, Frequency of Desktop or Laptop Computer Use to Access the Internet, etc.

The Household Identifier identifies the household, and the number of people is counted in the Count of household members. The Travel day-day of the week and the Number of Workers in the household are taken for travel information. Besides, this dataset also has information about the medium of transportation, which is stored in 'Frequency of Smartphone Use to Access the Internet,' 'Frequency of Tablet Use to Access the Internet,' 'Frequency of Walking for Travel,' 'Frequency of Bicycle Use for Travel,' 'Frequency of Personal Vehicle Use for Travel,' 'Frequency of Taxi Service or Rideshare Use for Travel,' 'Frequency of Bus Use for Travel,' 'Frequency of Train Use for Travel,' 'Frequency of Paratransit Use for Travel.' It stores modes of transportation such as trains, buses, bicycles, and personal vehicles. The dataset has information about stuff that may affect their mode of transportation. For example, 'Price of Gasoline Affects Travel,' 'Travel is a Financial Burden,' 'Walk to Reduce Financial Burden of Travel,' 'Bicycle to Reduce Financial Burden of Travel,'

'Public Transportation to Reduce Financial Burden of Travel,' 'At least two household persons are related,' 'Number of drivers in household' these features keep track of the things that may affect their traveling. Here is a graph that demonstrates the samples from different states:

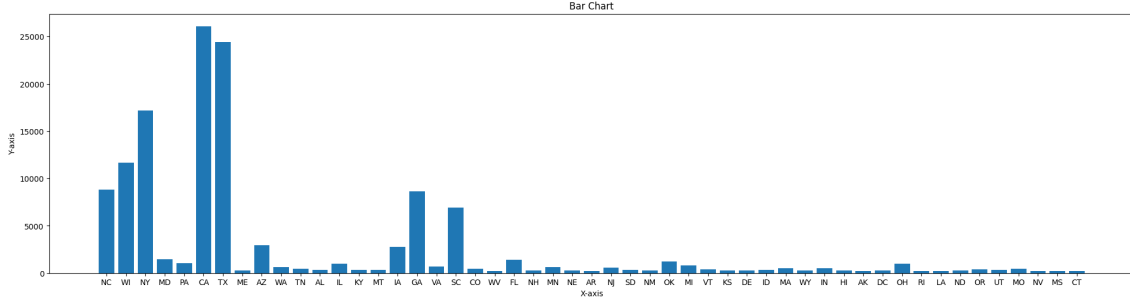


Figure 3.5: Samples from Different States of USA

3.4.2 Personal Data

The Personal Data mainly focuses on the personal information of the household. Here, it identifies the person inside a household by numbers 1,2,3 and stores information like age, Sex, race, and origin. Also, this dataset stores the travel information of the person. Personal Data is crucial because it helps identify the person inside the house, which helps trace all kinds of travel data of the person in the trip dataset.

The dataset has 121 features and 264234 samples. It means we have information about 264234 people out of 129696 households. The features include 'Age,' 'Race,' 'Sex,' 'Generalized purpose of the trip, home-based and non-home-based,' 'Date of travel day (YYYYMM)' and many more.

3.4.3 Trip Data

The Trip Dataset has all the information regarding two or three trips of the personal dataset's individuals. In short, it stores trip purpose, origin, destination, distance of the travel, trip start time, and trip end time.

Trip Information helps us to get an overview of the primary trips from multiple days and by multiple individuals. It is the hub of all activity details with 923572 samples. Therefore, we have a sample of 264234 people and their 923572 trips. Trip information has 115 features, including Trip Start Time, Trip End Time, Trip Origin Purpose, Trip Destination Purpose, Generalized Purpose of Trip, Primary Activity in the Previous Week, and the other features. Here is a graph that represents the Primary Activity in the Previous Week of the sample:

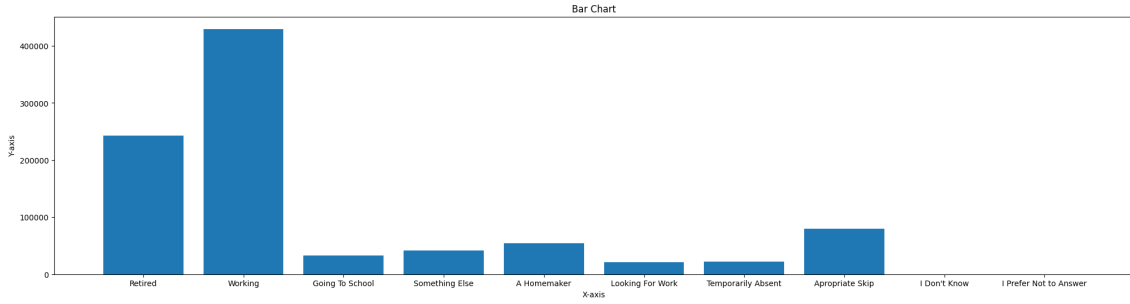


Figure 3.6: Barchart of Primary Activity in Previous Week in NHTS Data Sample

3.4.4 Vehicle Data

This dataset stores the vehicle information of households including driver, the age of the vehicle, number of vehicles, primary vehicle, months of vehicle ownership etc. This dataset will help us to apply heuristics in location choices of the individuals.

This dataset has 60 features and 256115 samples. Therefore, we have the information of 256115 vehicles of 129696 households.

Chapter 4

Description of the Model

4.1 Iterative Proportional Fitting (IPF)

Iterative proportional fitting (IPF) is a statistical method for estimating the proportions of a population in different categories. It is an iterative algorithm that starts with a set of initial estimates and then repeatedly adjusts the estimates until they converge to a solution. IPF is a popular algorithm for estimating population proportions because it is relatively easy to implement and it is typically very accurate.

The concept of IPF was first introduced by Deming and Stephan in 1940, and it has since been recognized as a fundamental procedure in demographic studies.[11] IPF is employed to adjust the data cells of a table to meet specific totals for its rows and columns.[16] In our research, we will be working with datasets that have multiple dimensions, necessitating the application of IPF to more than two or three dimensions.

The IPF process can be described as follows:

- Start with a table of initial estimates of the population proportions.
- Calculate the row and column totals of the table.
- For each row, divide each cell by the sum of the row and then multiply it by the corresponding marginal row total.
- For each column, divide each cell by the sum of the column and then multiply it by the corresponding marginal column total.
- Repeat steps 3 and 4 until the desired convergence rate is achieved.
- The most commonly used iteration method involves adjusting the cells based on the ratio of row constraints and column constraints. Each step in this method consists of two parts: first, the ratio of row and row constraints is added, and then the ratio of column and column constraints is added.

Depending on the form of the objective function used, the mathematical optimiza-

tion problem takes the following form:

$$\begin{aligned} \text{Minimize } \sum_j \left[\left(\sum_i d_{i,j} w_i - c_j \right) / c_j \right]^2 \text{ or } \sum_j \left[\left(\sum_i d_{i,j} w_i - c_j \right)^2 / c_j \right] \\ \text{or } \sum_j \left[\left| \left(\sum_i d_{i,j} w_i - c_j \right) \right| / c_j \right] \end{aligned} \quad (4.1)$$

where i denotes household type ($i = 1, 2, \dots, 8$)

j denotes the constraint or population characteristic of interest ($j = 1, 2, \dots, 5$)

$d_{i,j}$ represents the frequency of the population characteristic (household/person type) j in household i

w_i is the weight attributed to the i_{th} household

c_j is the value of the population characteristic j .

$w_i > 0$ [7]

IPF is a mathematically proven and widely utilized procedure in various fields. It is used in a variety of applications, including:

- Demographic research
- Market research
- Social science research
- Health research
- Environmental research

IPF is a powerful tool for estimating population proportions and it can be used to address a wide range of research questions.[13] It is relatively easy to implement and it is typically very accurate.

4.2 Iterative Proportional Updating

IPU is an iterative algorithm that is used to adjust a set of data to match a set of constraints. The constraints can be on the total number of observations, the distribution of the observations across different categories, or the relationships between different variables.

IPU works by starting with a set of initial data and then repeatedly adjusting the data until it matches the constraints. In each iteration, the algorithm calculates a weighted sum of the data, where the weights are determined by the constraints. The data is then adjusted by multiplying each observation by its weight. This process is repeated until the data converges to a solution that matches the constraints. [16]

IPU is a powerful tool that can be used to adjust a wide variety of data sets. It has been used in a variety of applications, including:

- Population synthesis: IPU can be used to create synthetic populations that match the demographic characteristics of a real population. This can be useful for research, planning, and simulation purposes.
- Data cleaning: IPU can be used to clean data sets by removing errors and inconsistencies.
- Data integration: IPU can be used to integrate data sets from different sources. This can be useful for creating a more complete and accurate picture of a population or phenomenon.

IPU is a versatile and powerful tool that can be used to improve the quality and usability of data. Some additional details about IPU:

- IPU is a deterministic algorithm, which means that it will always converge to the same solution given the same starting data and constraints.
- IPU is a relatively efficient algorithm, which means that it can be used to adjust large data sets quickly.
- IPU is a flexible algorithm, which can be used to adjust data sets with a variety of constraints.

IPU is a valuable tool for researchers and practitioners who need to adjust data to match a set of constraints. It is a versatile and efficient algorithm that can be used in a variety of applications.

4.3 Monte Carlo Sampling

Monte Carlo sampling is a technique for generating random samples from a population. It is a powerful tool for statistical analysis because it can be used to estimate the probability of events, calculate confidence intervals, and perform hypothesis tests. Monte Carlo sampling is often used in conjunction with other statistical techniques, such as IPF, to improve the accuracy of estimates.

Rolf Moeckel, Klaus Spiekerman, and Michael Wegner used Monte Carlo sampling to generate a synthetic population with a wide range of characteristics. This technique enabled the creation of multidimensional data based on one-dimensional administrative distributions.[12] The process begins by determining the age of the household head to establish the size of the household. For instance, households with younger heads tend to reside in smaller homes, while those with older heads occupy larger homes.[14] The formation of households and individuals follows a natural order, where personal and household characteristics are sampled based on the influence of person-to-person relationships. This sequential sampling approach improves the accuracy of the results by incorporating previously selected attributes. Additionally, the probability of selecting a specific household size is influenced by the age of the household head, calculated through iterative proportional fitting.

Once the household members have been determined, the number of cars is assigned. This decision is influenced by factors such as the age of the family head, household size, and the ages of the household members. This process follows a causal order for creating characteristics. Monte Carlo sampling allows for the execution of

multiple microsimulation features as required and is utilized to assign individuals to their respective households.[9]

The Monte Carlo sampling process can be described as follows:

- Start with a population table that contains the desired characteristics of the synthetic population.
- Generate a random number for each individual in the population.
- Use the random number to select a value for each characteristic from the population table.
- Repeat steps 2 and 3 until all individuals in the population have been assigned characteristics.
- The Monte Carlo sampling process can be used to create synthetic populations with a wide range of characteristics. It is a powerful tool for statistical analysis and it can be used to address a wide range of research questions.

Some of the advantages of using Monte Carlo sampling to generate synthetic populations:

- It is a flexible technique that can be used to create synthetic populations with a wide range of characteristics.
- It is a relatively easy-to-implement technique.
- It is a relatively accurate technique.

Some of the disadvantages of using Monte Carlo sampling to generate synthetic populations:

- It can be a computationally expensive technique.
- It can be a time-consuming technique.
- It can be a difficult technique to control.

4.4 Contact Tracing Using Social Networks

In this section, researchers employ three types of social networks: education, work, and household. Initially, household data is extracted from census data, while living spaces and work locations are obtained from road use data. Subsequently, households, workplaces, and educational institutions are placed, and daytime locations are assigned to each individual. The outcome of this process is the creation of a social network.

This method proves beneficial for understanding location choices and contact tracing. However, it does not provide detailed contact tracings information such as consumer behavior in shops, travel patterns, or healthcare utilization. Therefore, while it offers valuable insights for contact tracing, it may not be the most accurate method due to less precise location assignments.

The synthetic population generated through this approach comprises 23,004,272 individuals and 8,457,710 households. Nevertheless, a slight discrepancy of 0.36% in accuracy arose due to inconsistencies in 116 census tracts.[15]

4.5 Prediction of trip purpose

Our goal is to find a model that can predict the activity of individuals under certain conditions. Therefore, we applied regression and classification models and compared the personal and trip data accuracies.

4.5.1 Regression Models

We have applied Linear Regression, Ridge Regression, Lasso Regression, and Elastic Net Regression. We worked on personal data, and our targeted feature was Medical Condition based on Age, Race, Sex, Count of Walk Trips, Count of Bike Trips, Count of Public Transit Usage, and Level of Physical Activity.

Linear regression is a simple method that uses a linear equation to show how a dependent variable relates to one or more independent variables. It finds the best line to fit the data by minimizing the sum of squared differences between the observed and predicted values. Ridge Regression is a version of Linear Regression that adds a penalty term to the loss function to make it more like Linear Regression. It eliminates multicollinearity and overfitting by making the regression coefficients smaller, which makes the model stronger. Like Ridge Regression, Lasso Regression adds a penalty term but uses the L1 regularization instead of the L2 regularization. It reduces the size of the coefficients and selects features by making some coefficients equal to zero. This makes automatic variable selection possible. By adding L1 and L2 regularization terms, Elastic Net Regression combines the best parts of Ridge and Lasso Regression. It balances the two approaches, allowing both feature selection and shrinking of the coefficients.

4.5.2 Classification Model

We applied Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbour, and Gaussian Naive Bayes. We worked on the same features.

Logistic Regression is a classification algorithm that uses input variables to predict the likelihood of a binary outcome. It models the relationship between the independent variables and the likelihood of the outcome using a logistic function. Decision Tree is an algorithm based on a tree that makes predictions by recursively dividing the data into groups based on the values of the features. It makes a structure that looks like a flowchart to sort or predict the target variable. Random Forest is a way to learn by putting together several decision trees. It makes a forest of trees and adds what they say to make more accurate and reliable predictions. K-Nearest Neighbors (K-NN) is a non-parametric algorithm that sorts data into classes based on the majority class of its k nearest neighbors. It is easy to use and flexible, but choosing k can make a difference. Gaussian Naive Based on Bayes' theorem, Bayes is a probabilistic algorithm. It assumes the features are conditionally independent given the target variable and models the likelihoods using the Gaussian distribution. Again, Logistic Regression is a classification algorithm that uses a logistic function to predict the chance of a binary outcome. It is often used because it is simple and easy to understand.

4.5.3 Neural Network

Then, we applied Neural Network. We have used Recurrent Neural Network with sequential function. Here, we used ten epochs and the ReLU activation function with two dense layers and softmax in the last layer.

A Recurrent Neural Network (RNN) with a sequential function is used in this neural network. RNNs are made to process data in a particular order, like a time series or a text. The network is trained for ten epochs, meaning it looks at the whole dataset ten times during training. The ReLU (Rectified Linear Unit) activation function is used, which makes the model not linear. The neural network comprises two dense layers that are fully connected. Each neuron in both layers is linked to every other neuron in both layers. The softmax activation function is used in the last layer. It gives probabilities for each possible class output.

Chapter 5

Preliminary Analysis

5.1 Analysis of Sampled Data

The IPF and IPU methods were used to adjust the census data for undercount and overcount. Undercount occurs when people are not counted in the census, and overcount occurs when people are counted more than once in the census. IPF and IPU work by adjusting the counts in each cell of the census data so that they sum to the known totals for each variable. As a result, the models we will run can now accurately imitate other data.

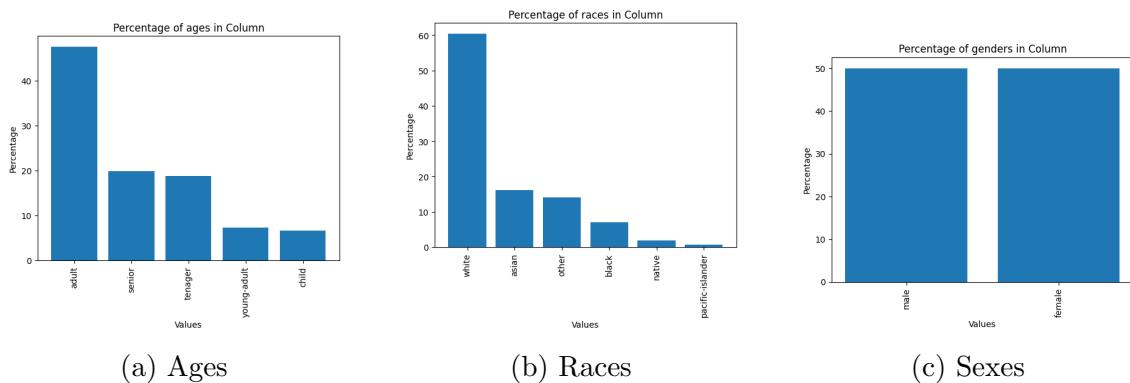


Figure 5.1: Sampled Data

age	adult		child		senior		teenager		young-adult	
gender	female	male	female	male	female	male	female	male	female	male
race										
asian	435	363	52	45	178	157	113	141	64	57
black	152	189	20	22	82	64	51	64	25	27
native	48	41	4	6	20	9	26	20	6	7
other	318	344	70	63	65	48	162	196	76	57
pacific-islander	21	17	3	2	1	2	11	7	2	6
white	1385	1438	155	215	715	642	550	540	186	215

Figure 5.2: Crosstabulation of Sampled Data

5.2 Finding the Best Model

As we have mentioned, our targeted feature is ‘Medical Condition’ and we take Age, Race, Sex, Count of Walk Trips, Count of Bike Trips, Count of Public Transit Usage, and Level of Physical Activity as our predictors. Here is our output results:

Model	Accuracy
Linear Regression	0.1261
Ridged Regression	0.1261
Lasso Regression	0.0188
Elastic Net Regression	0.0288

Table 5.1: Regression Models accuracy

As we can see, we have inferior results on regression models.

5.2.1 Classification Model

We applied Logistic Regression, Decision Tree, and Random Forest. We worked on the same features. Here is our output results when we used classification models:

Model	Accuracy
Decision Tree	0.91
Logistic Regression	0.87
Random Forest	0.90

Table 5.2: Classification Models on Predicting Medical Condition

As we can see, we get better results in the classification model. For this reason, we applied these models to our trip data to see if we could predict the purpose of the trip properly. Here, our targeted feature is ‘Generalized purpose of trip, home-based and non-home based’ and we dropped the feature with ‘Household Identifier’, ‘Person Identifier’, ‘Trip Origin Purpose’, home-based and non-home based’, ‘Trip Start Time (HHMM)’, ‘Trip End Time (HHMM)’, ‘Household state’, ‘Core Based Statistical Area (CBSA) FIPS code for the respondent’s home address’, ‘Urban / Rural indicator - Trip Origin Block group’, ‘Urban / Rural indicator - Trip Destination Block group’. Here are our results:

Model	Accuracy
Decision Tree	0.7612
Random Forest	0.8052
K-Nearest Neighbour	0.2453
Gaussian Naive Bayes	0.3349
Logistic Regression	0.3353

Table 5.3: Classification Models on Predicting Generalized Purpose of Trip

Then, we applied Neural Network. We have used Recurrent Neural Network with sequential function. Here, we used ten epochs and the ReLU activation function with

two dense layers and softmax in the last layer. Our accuracy was 0.8329, which is the highest.

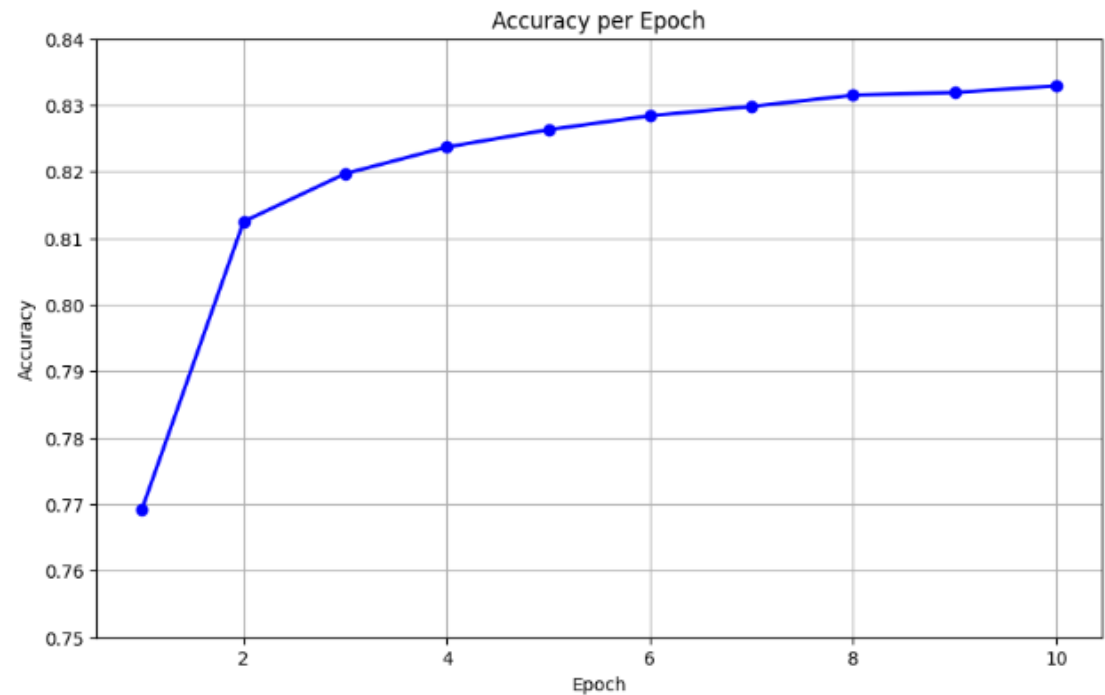


Figure 5.3: Applying Neural Network to Predict Generalized Purpose of the Trip

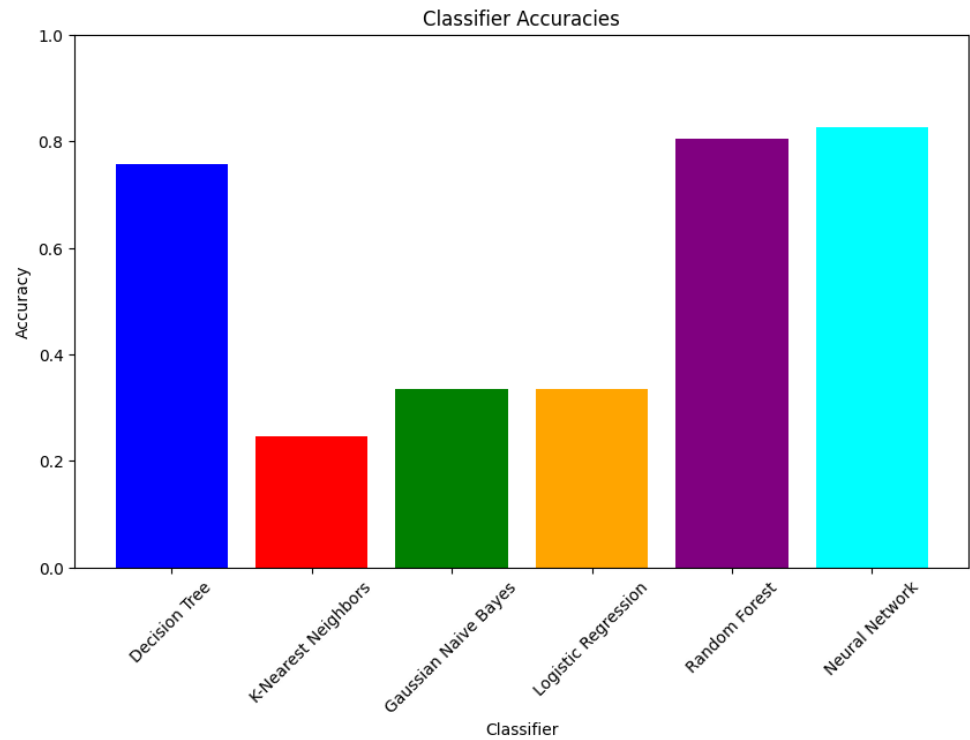


Figure 5.4: Comparison of Models

5.2.2 Challenges

Here, we have multiple challenges in the model.

1. **Limited Amount of Features:** We have yet to work on a model to help us work on the lowest features possible. Because the NHTS data and census data has a few elements in common.
2. **Higher Accuracy:** We are still looking for a higher-accuracy model because when we are simulating a large population, 83% accuracy can exclude many people.
3. **Location Choice:** Our model has yet to find an appropriate trip location. Because the NHTS samples are very few while it is sampled from multiple regions of the United States.

Chapter 6

Conclusion

Synthesizing a representative population through population generation synthesis using census data allows for a comprehensive grasp of a region's demographics and socio-economic characteristics. This process provides valuable insights into the population's composition, distribution, and dynamics, enabling informed decision-making and effective policy formulation.

Moreover, population synthesis enables the projection of future population trends, the anticipation of demographic changes, and the evaluation of the impact of different scenarios and interventions. This information is crucial for urban planning, resource allocation, infrastructure development, and service provision, facilitating the fulfillment of evolving community needs.

Additionally, population synthesis facilitates the analysis of social disparities, identification of vulnerable populations, and evaluation of resource distribution equity. It uncovers patterns, correlations, and relationships among various demographic variables, supporting evidence-based decision-making and targeted interventions. In conclusion, population generation synthesis using census data empowers an understanding of population complexities, the anticipation of future trends, and informed decision-making. It significantly contributes to policy shaping, resource allocation improvement, and the promotion of equitable development. Harnessing the potential of population synthesis enables the creation of sustainable, inclusive, and resilient communities, ultimately benefiting society as a whole.

Bibliography

- [1] R. Brown, “Comparing census data in 90 countries,” *The American Statistician*, vol. 25, no. 1, p. 32, 1971.
- [2] R. J. Beckman, K. A. Baggerly, and M. D. McKay, “Creating synthetic baseline populations,” *Transportation Research*, vol. 30, no. 6, pp. 415–429, 1996.
- [3] R. J. Beckman, K. A. Baggerly, and M. D. McKay, “Creating synthetic baseline populations,” *Transportation Research Part A: Policy and Practice*, vol. 30, no. 6, pp. 415–429, 1996, ISSN: 0965-8564.
- [4] R. J. Beckman, K. A. Baggerly, and M. D. McKay, “Creating synthetic baseline populations,” *Transportation Research A - Policy and Practice*, vol. 30, pp. 415–429, 1996.
- [5] C. Barrett, R. Beckman, K. Bisset, *et al.*, “Building social contact networks,” Virginia Bioinformatics Institute, Virginia Tech, Tech. Rep., 2006.
- [6] R. Moeckel, K. Spiekermann, and M. Wegener, “Creating a synthetic population,” in *8th International Conference on Computers in Urban Planning and Urban Management (CUPUM)*, Sendai, Japan, 2007.
- [7] X. Ye, K. Konduri, R. Pendyala, B. Sana, and P. Waddell, “Methodology to match distributions of both household and person attributes in generation of synthetic populations,” Jan. 2009.
- [8] R. Beckman, K. Channakeshava, F. Huang, *et al.*, “Integrated multi-network modeling environment for spectrum management,” *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 6, pp. 1158–1168, Jun. 2013.
- [9] J. Rich and I. Mulalic, “Generating a dynamic synthetic population - using an age-structured two-sex model for household dynamics,” *PLoS ONE*, vol. 9, no. 7, e104138, 2014. DOI: 10.1371/journal.pone.0104138.
- [10] A. Adiga *et al.*, “Generating a synthetic population of the united states,” 2015, Accessed: Sep. 15, 2022. <https://nssac.bii.virginia.edu/swarup/papers/US-pop-generation.pdf>.
- [11] A. A. Choupani and A. R. Mamdoohi, “Population synthesis using iterative proportional fitting (ipf): A review and future research,” *Transportation Research Procedia*, vol. 17, pp. 223–233, 2016. DOI: 10.1016/j.trpro.2016.11.078.
- [12] K. Lum, Y. Chungbaek, S. Eubank, and M. Marathe, “A two-stage, fitted values approach to activity matching,” *International Journal of Transportation*, vol. 4, no. 1, pp. 41–56, 2016. DOI: 10.14257/ijt.2016.4.1.03.

- [13] N. Watthanasutthi and V. Muangsin, “Generating synthetic population at individual and household levels with aggregate data,” in *2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 2016, pp. 1–6. DOI: 10.1109/JCSSE.2016.7748838.
- [14] “Household population, number of households, and average household size of the philippines (2020 census of population and housing) — philippine statistics authority.” <https://psa.gov.ph/content/household-population-number-households-and-average-household-size-philippines-2020-census>. (2022).
- [15] N. Jiang, A. Crooks, H. Kavak, A. Burger, and W. G. Kennedy, “A method to create a synthetic population with social networks for geographically-explicit agent-based models,” *Computational Urban Science*, vol. 2, no. 1, Feb. 2022. DOI: 10.1007/s43762-022-00034-1.
- [16] P. Ye, B. Tian, Y. Lv, Q. Li, and F.-Y. Wang, “On iterative proportional updating: Limitations and improvements for general population synthesis,” *IEEE Transactions on Cybernetics*, vol. 52, no. 3, pp. 1726–1735, Mar. 2022. DOI: 10.1109/TCYB.2020.2991427.
- [17] US Census Bureau. “Census.gov.” (May 2023), [Online]. Available: <https://www.census.gov/>.