MathWorks Math Modeling Challenge 2022

High Technology High School

Team #15596, Lincroft, New Jersey

Coach: Raymond Eng

Students: David Chang, Alexander Postovskiy, Andrew Eng, Kevin Guan, Ivan Wong

M3 Challenge FINALIST—$5,000 Team Prize
M3 Challenge Technical Computing Award
RUNNER UP—$2,000 Team Prize

JUDGE COMMENTS

Specifically for Team # 15596—Submitted at the Close of Triage Judging:

COMMENT 1: Good executive summary. Question 1 is well thought out and justified properly. I really like the use of a machine learning algorithm in question 2, and the confusion matrix is a great way to display the results; plenty of good assumptions. This section could have been a bit stronger with some examples of individuals and their likelihood of working remotely. Great work!

COMMENT 2: Great summary. Nice use of regression and incorporating weighted averages in Q1. The factors considered in Q2 are very reasonable and good analysis is done on the difference between reality and prediction. However, as illustrated and mentioned, the model is not perfect. Very well written and organized.

COMMENT 3: The paper is well written with clear background with detailed model descriptions. A proposed model was clearly justified through reasonable assumptions. It is phenomenal to see how the multi regression model and Monte Carlo simulation were used to execute the proposed model. The model results were explained in detail with the quantifiable table. The strength and weaknesses of each model were practical. Overall, the modeling approach was elegant, with all the necessary information.

Specifically for Team #15596 —Submitted at the close of Technical Computing Contention Judging:

COMMENT 1: This paper started off by leveraging technical computing to solve an interesting and non-standard multi-response regression problem that was used to model the growth of various industries for Q1. By translating their problem into linear algebra, they were able to leverage Python's numpy library to directly solve the regression problem, showcasing the enabling power of technical computing. For Q2, the team employed a machine learning model trained on real survey data on individual decisions to work from home. The judges were impressed with the team's choice of an interpretable, rule-based model, and their efforts to sanity check the model's decisions by directly examining the rules generated. We also appreciated the clear illustration and discussion of the models performance: the team employed best-practices like examining the "confusion matrix" and computing specificity and sensitivity for their machine learning model. We appreciated the relative conciseness of the code used to implement that model, although we would have liked to have seen less hard-coding of city parameters to facilitate code reuse if other cities are analyzed. The judges did notice some modeling issues in the paper. For example, the multivariate regression model in Q1 was too complex for the amount of historical data available – this resulted in an under-determined system of equations, which explains the artificially high R^2 values. The teams linear algebraic approach also failed to account for constraints on the "P matrix" that was central to the Q1 model. However, we felt that most issues could have been addressed given more time. This is a fantastic paper, showing off the best that technical computing has to offer!

MathWorks Math Modeling Challenge
A program of SIAM

Remote Work: Fad or Future

Executive Summary

In a world suddenly overturned by the COVID-19 pandemic, many have been forced to make the sharp transition from in-person activities to online ones. Workers specifically are faced with the challenge of working from home, and the proportion of remote jobs is still higher than ever before [17]. As we continue to adapt to this persisting trend in the foreseeable future, there are many uncertainties that need to be resolved for national leaders and agencies to make informed policy decisions. This paper proposes mathematically founded insights on understanding how industries will continue to prepare for remote work, how people will perceive their need to work from home, and the impact this will all have on cities around the world.

Since one crucial factor in adjusting to the possibility of shifting to remote work is the number of jobs that are capable of working remotely, we predict the percentage of remote-ready jobs in the US and UK (i.e. Seattle, Omaha, Scranton, Liverpool, and Barry). We fit the growth of each industrial sector for every city based on the size of each market. Then, we multiply the projection of jobs by the percentage of remote-ready jobs in each sector and sum the products to determine the final proportion of remote-ready jobs. We predict that the proportions of jobs that will remote-ready by 2024 for the above cities in order will be .3836, .3592, .3239, .3481, and .3055, while in 2027 the proportions will be .3988, .3585, .3241, .3756, and .3190, respectively.

We then develop a model to address the tendency of any given person to choose to work remotely, based on a set of factors related to personal circumstances. We specifically examine a worker's age, children, occupation, education level, sex, spouse's employment, and elderly family. This data was gathered for a sample of individuals collected from the American Time Use Survey and used to train an interpretable RuleFit machine learning model. The model established a set of decision rules between provided training inputs and outputs, and was able to successfully classify given data of new individuals. We include some of the rules the trained model used for classification in addition to analyzing a confusion matrix and evaluation metrics.

Finally, we predict the economic benefit for each of the cities due to the increase in productivity and work hours from remote workers. To do so, we use a Monte Carlo simulation to randomly generate households and fit their characteristics as dependent variables in the RuleFit machine learning model to determine the likelihood that the people in the household work from home with our first model. From there, we take the sum of the ratio of extra hours worked over total hours worked in a year over every household and multiply by the GDP per capita of the city to find the increase in relative USD (millions USD per 100,000 people) in that city from remote workers. The model predicts that Liverpool, Omaha, Scranton, Seattle, and Barry increased by 26.05, 110.59, 133.81, 114.25, and 5.38 relative
USD in 2024 and 100.24, 64.67, 55.09, 4.33, and 13.02 relative USD in 2027, respectively.

With the influx of jobs that are forced to be conducted remotely as well as a large increase in jobs performed online, it is important to governments, businesses, and individuals to make informed decisions on living locations, careers, infrastructure, and much more. We posit that the models detailed in this paper impart key information that can address the possible incoming changes in transferring to more remote-based working in the future.

Contents

1 Introduction

This section delineates the components of the modeling problem and their objectives. Global assumptions applying to the entire modeling process are also listed.

1.1 Restatement of the Problem

The problem we are tasked with addressing is as follows:

1. Build a mathematical model that predicts the percentage of workers who are remote-ready in 2024 and 2027 for each of the cities Seattle, Omaha, Scranton, Liverpool, and Barry.

2. Create a mathematical model that predicts whether or not an individual worker whose job is remote-ready will be allowed to work from home by their employer and will choose to do so.

3. Develop a model that will estimate the proportion of workers who will work remotely. For the cities in Q1, estimate and rank the impact that the amount of remote work in 2024 and 2027 will have.

2 Part I: Ready or Not

Due to the COVID-19 pandemic, many jobs have opted or required their workers to work remotely, creating a shift in the workplace environment and labor force [h]. In this section, we develop a mathematical model which estimates a given city's percentage of workers whose jobs are currently remote-ready and apply the model to the following five cities: Seattle, Washington; Omaha, Nebraska; Scranton, Pennsylvania; Liverpool, England; and Barry, Wales. We then use the model to predict these percentages for the years 2024 and 2027.

2.1 Assumptions

1. *The proportion of the workforce below 16 and above 74 in the UK is negligible.* The number of people not accounted for in the data should be small enough to have little impact on the proportion of the workforce in each industry.

2. *The proportion of jobs that are remote-ready in each industrial sector is the same for the US and UK.* Across both the US and UK industrial sectors, the distribution of jobs should be similar, therefore implying that the remote-readiness of their sectors should be approximately equal.

3. *Change in the market is solely dependent on the current state of the market.* Each sector of the market competes and affects one another, forming positive and negative relationships. Global influences, such as emerging technology, exert influence through internal market forces.

4. *Local fluctuations in the market are negligible.* It is outside the scope of this model to predict a specific external influence (e.g., a natural disaster) that would undoubtedly have a great effect on the market. However, this effect would be largely in the short term, and thus can be ignored when modeling global trends.

5. *The proportion of remote-ready jobs in a specific category is time-invariant.* There is only data about remote-readiness at a single time point. Most likely, remote-readiness is time-dependent, for example, as a logistic function, and this could be substituted into our model to replace the constant proportion were there data. It is not, however, unreasonable to assume that over the short time period in the future we are tasked with modeling, this proportion is approximately constant.

6. *The proportions of the workforce for each industry in Liverpool and Barry are well approximated by those of the North West and Wales sectors of England, respectively.* There is a requirement for a large amount of comprehensive data for the regression model that is only available through censuses of the sectors of England rather than the individual city of Liverpool and the town of Barry.

2.2 Model Development

The proportion of jobs that are remote-ready is the number of remote-ready jobs in the market divided by the total number of jobs in the market. Data exist as to what proportion of certain categories of jobs are remote-ready [D3]. It is necessary to convert these into proportions ρ_i for sectors of the industry, as those are what employment data is collected on [D1]. For each sector, we take a weighted average of remote-readiness proportion across all job subcategories in that sector, weighted by the number of employees in that subcategory to produce ρ_i.

Table 2.2.1: Proportion of Remote-Ready Jobs by Sector [1, 2]

Sector	ρ_i
Mining, Logging, Construction (MLC)	0.19
Manufacturing (MFG)	0.22
Trade, Transportation, and Utilities (TTU)	0.23
Information (INF)	0.72
Financial Activities (FIN)	0.67
Professional and Business Services (PBS)	0.59
Education and Health Services (EHS)	0.34
Leisure and Hospitality (LHO)	0.077
Other Services (OTS)	0.31
Government (GOV)	0.41

Given that the distribution of workers across the ten sectors is given by **x**, the scalar product $\mathbf{x}\,\boldsymbol{\rho}$ will give the total proportion of jobs in the market that are remote-ready. To apply this model to future years, it is first necessary to model the evolution of **x** in those years.

It is natural, given the mass of data available for employment numbers in the relevant cities [D1], to use a regression method to extrapolate future values. However, it is impossible to analyze each sector separately. Consider that the sum of the distribution **x** of jobs by sector must be 1. The proportion of jobs in every sector, for example, cannot simultaneously increase or decrease, as this would violate the principle. Thus, the proportion for a given sector at some future time must depend on the proportions for each sector at the current time. We solve the problem discretely by using yearly time intervals, which eliminates seasonal variation in employment. The simplest path forward is to conjecture that the dependence is linear; the proportion of jobs in a sector at year $t + 1$ is a linear combination of the proportions of jobs in each sector at year t. We develop a model from this principle and then verify its applicability.

$$\mathbf{x}_{t+1} = \mathbf{P} \cdot \mathbf{x}_t \tag{1}$$

We use the matrix **P** to represent the factors for the linear combinations. The ith row of **P** will, by the properties of matrix multiplication, serve as the factors in the combination that generates the ith element of $\mathbf{x}_{t+1}$. This formulation is similar to a Markov chain, except the values of the transition matrix are not probabilities of the system changing states, but factors affecting how the state distribution of the system changes. Thus, these values can be negative, but the matrix's rows must still sum to 1 in order for the sum of the sector distribution to remain 1. It remains to calculate the matrix **P**, which we do with linear regression.

2.2.1 Developing the Regression

The standard form of a regression in multiple variables is given by $\hat{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}$, where $\hat{\mathbf{Y}}$ is the response variable, **X** is the explanatory variable, $\boldsymbol{\beta}$ is a constant matrix of regressed coefficients, and **U** is a matrix of the residuals. To minimize the sum of the squares of the residuals, $\boldsymbol{\beta}$ is given by the following:

$$\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{2}$$

Our goal is to transform our model into the form of Equation (2). We start by combining the equations $\mathbf{x}_2 = \mathbf{P} \cdot \mathbf{x}_1$, $\mathbf{x}_3 = \mathbf{P} \cdot \mathbf{x}_2$, ..., $\mathbf{x}_n = \mathbf{P}\ \mathbf{x}_{n-1}$ into matrix form, where n is the number of years for which data is available:

$$[\mathbf{x}_2\mathbf{x}_3 \ldots \mathbf{x}_n] = \mathbf{P}[\mathbf{x}_1\mathbf{x}_2 \ldots \mathbf{x}_{n-1}] \tag{3}$$

Taking the first matrix as **Y** and the second as **X**, we write this as $\mathbf{Y} = \mathbf{PX}$. Taking the transpose of both sides, we obtain $\mathbf{Y}^T = \mathbf{X}^T\mathbf{P}^T$. This is now in the form of Equation (2), and we obtain the following by substituting and simplifying with matrix properties:

$$\mathbf{P} = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{Y}^T. \tag{4}$$

The residuals are given by $\mathbf{U} = \mathbf{Y}^T - \mathbf{X}^T\mathbf{P}^T$, and coefficients of determination R_i^2 are then calculated for each sector (the matrix regression is functionally i separate regressions, one per sector). We execute this linear regression five times, once per city, to find the matrices P for each city.

2.3 Results

Using the matrix **P** for each respective city, we predict the proportion of the workforce in each industrial sector t years in the future by premultiplying the 2021 proportions (2020 for UK cities) by **P** raised to the power t:

Table 2.2.3.1: Predictions for Proportions by Industrial Sector in Seattle, WA

Year	MLC	MFG	TTU	INF	FIN	PBS	EHS	LHO	OTS	GOV
2022	0.0702	0.0724	0.1960	0.0851	0.0510	0.1667	0.1298	0.0783	0.0343	0.1163
2023	0.0769	0.0617	0.1974	0.0903	0.0509	0.1708	0.1290	0.0769	0.0338	0.1122
2024	0.0829	0.0507	0.1983	0.0963	0.0509	0.1740	0.1299	0.0745	0.0333	0.1093
2025	0.0877	0.0381	0.1989	0.1028	0.0510	0.1766	0.1327	0.0717	0.0331	0.1075
2026	0.0919	0.0237	0.1992	0.1096	0.0513	0.1796	0.1368	0.0689	0.0330	0.1061
2027	0.0962	0.0078	0.1996	0.1166	0.0513	0.1836	0.1413	0.0662	0.0331	0.1045

Table 2.2.3.2: Predictions for Proportions by Industrial Sector in Omaha, NE

Year	MLC	MFG	TTU	INF	FIN	PBS	EHS	LHO	OTS	GOV
2022	0.0580	0.0644	0.1872	0.0179	0.0895	0.1412	0.1645	0.1025	0.0378	0.1370
2023	0.0564	0.0638	0.1834	0.0173	0.0917	0.1416	0.1672	0.1009	0.0379	0.1399
2024	0.0566	0.0646	0.1818	0.0172	0.0922	0.1431	0.1676	0.0987	0.0381	0.1402
2025	0.0562	0.0650	0.1815	0.0169	0.0917	0.1431	0.1675	0.0992	0.0383	0.1406
2026	0.0553	0.0648	0.1813	0.0164	0.0916	0.1423	0.1680	0.1003	0.0385	0.1416
2027	0.0546	0.0648	0.1808	0.0160	0.0919	0.1419	0.1685	0.1005	0.0386	0.1425

Table 2.2.3.3: Predictions for Proportions by Industrial Sector in Scranton, PA

Year	MLC	MFG	TTU	INF	FIN	PBS	EHS	LHO	OTS	GOV
2022	0.0408	0.1072	0.2604	0.0094	0.0515	0.1053	0.2018	0.0766	0.0319	0.1154
2023	0.0404	0.1045	0.2606	0.0086	0.0503	0.1050	0.2022	0.0807	0.0327	0.1157
2024	0.0402	0.1029	0.2595	0.0078	0.0497	0.1044	0.2042	0.0841	0.0330	0.1154
2025	0.0400	0.1027	0.2579	0.0073	0.0496	0.1042	0.2064	0.0857	0.0328	0.1146
2026	0.0399	0.1038	0.2567	0.0070	0.0500	0.1046	0.2080	0.0854	0.0324	0.1137
2027	0.0399	0.1055	0.2563	0.0068	0.0505	0.1053	0.2086	0.0839	0.0319	0.1128

Table 2.2.3.4: Predictions for Proportions by Industrial Sector in Liverpool, UK

Year	MLC	MFG	TTU	INF	FIN	PBS	EHS	LHO	OTS	GOV
2021	0.0495	0.0741	0.2081	0.0362	0.0351	0.2095	0.2483	0.0804	0.0246	0.0345
2022	0.0547	0.0787	0.2116	0.0368	0.0439	0.2214	0.2422	0.0763	0.0197	0.0148
2023	0.0558	0.0845	0.2258	0.0272	0.0426	0.2149	0.2259	0.0740	0.0370	0.0119
2024	0.0534	0.0902	0.2244	0.0274	0.0354	0.2031	0.2103	0.0796	0.0563	0.0201
2025	0.0491	0.0858	0.2196	0.0309	0.0284	0.2120	0.2042	0.0866	0.0527	0.0312
2026	0.0494	0.0704	0.2177	0.0399	0.0267	0.2400	0.2156	0.0817	0.0297	0.0296
2027	0.0502	0.0565	0.2263	0.0436	0.0301	0.2707	0.2313	0.0764	0.0041	0.0110

Table 2.2.3.5: Predictions for Proportions by Industrial Sector in Barry, UK

Year	MLC	MFG	TTU	INF	FIN	PBS	EHS	LHO	OTS	GOV
2021	0.0771	0.1251	0.2009	0.0019	0.0402	0.1220	0.2505	0.0690	0.0332	0.0807
2022	0.0961	0.1009	0.2011	0.0101	0.0334	0.0959	0.2390	0.0983	0.0350	0.0905
2023	0.0675	0.1152	0.2026	0.0143	0.0336	0.1161	0.2657	0.0516	0.0412	0.0933
2024	0.0955	0.1150	0.2144	0.0112	0.0398	0.0874	0.2640	0.0717	0.0374	0.0873
2025	0.0676	0.1113	0.2296	0.0111	0.0298	0.0798	0.2707	0.0666	0.0419	0.0930
2026	0.0863	0.1134	0.1902	0.0116	0.0372	0.0853	0.2872	0.0581	0.0471	0.0849
2027	0.0609	0.1303	0.2256	0.0068	0.0362	0.0849	0.2911	0.0442	0.0465	0.0752

Every R^2 value calculated was above a value of 0.999.

Finally, we take the weighted average as described above using the proportions in Table 2.2.1 and Tables 2.2.3.1–2.2.3.5 to find the percentage of jobs in each city that are remote-ready.

Table 2.2.3.6: Percentage of Jobs Remote-Ready by City

City	2021	2022	2023	2024	2025	2026	2027
Seattle, WA	0.3743	0.3766	0.3798	0.3836	0.3881	0.3933	0.3988
Omaha, NE	0.3578	0.3561	0.3581	0.3592	0.3589	0.3584	0.3585
Scranton, PA	0.3274	0.3264	0.3250	0.3239	0.3234	0.3236	0.3241
Liverpool, England	0.3591	0.3633	0.3549	0.3481	0.3503	0.3641	0.3756
Barry, Wales	0.3225	0.3098	0.3314	0.3055	0.3134	0.3202	0.3190

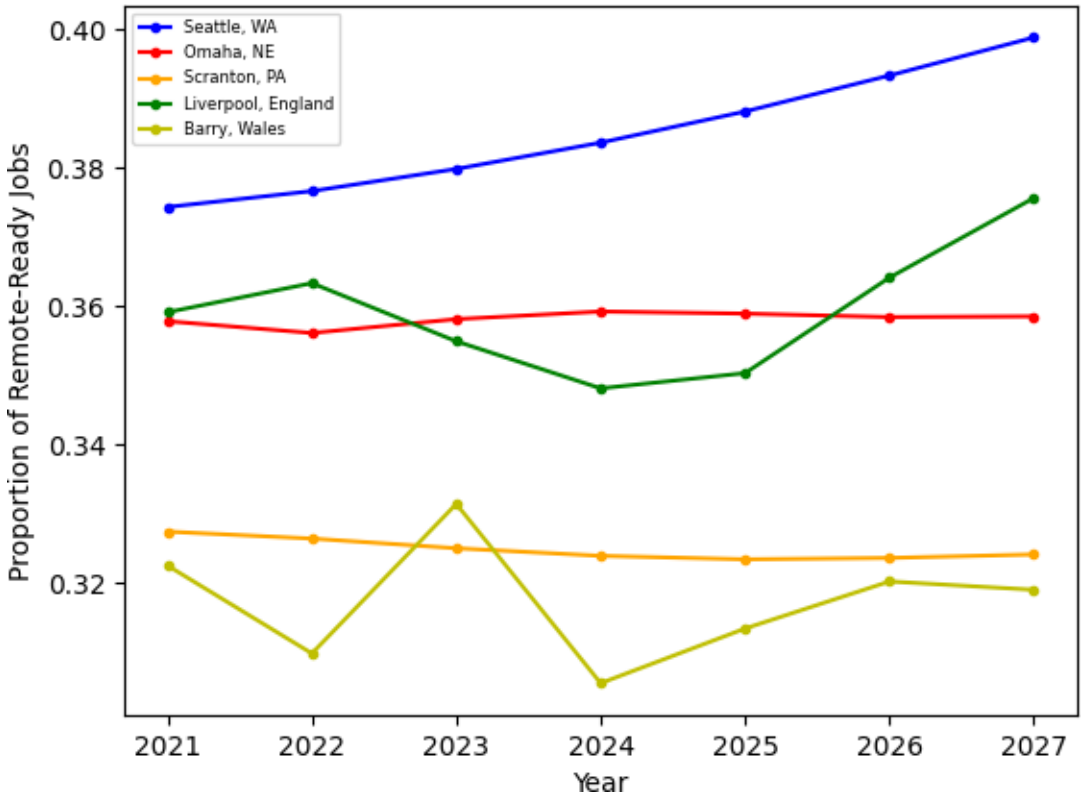

Figure 2.2.3.7: Predictions of Jobs Remote-Ready by City and Year

While Seattle has a clearly upwards trend, the other 4 cities have no clear indication of a growing or shrinking number of jobs that are remote-ready. These trends indicate that, despite changes in the size of each industrial sector, most cities will likely not have a significant increase in the proportion of remote-ready jobs in its workforce, but instead remain mostly stable.

2.4 Strengths and Weaknesses

Our model fits the data very well as can be observed by all the R^2 values being greater than 0.999. Additionally, the model predicts for the growth of every sector of the market relative to each other and accounts for the relationships between every pair of sectors. This is notably useful in predicting the number of remote-ready jobs from each sector, allowing the identification and analysis of the source of remote-ready jobs.

However, our model assumes that the local fluctuations in the market are negligible, which means that it may be weak and inaccurate with short-term natural disasters or longer-term societal shifts in values. Of note, in this case, is the pandemic.

3 Part II: Remote Control

When looking at jobs which are remote-ready, both workers and their employers have different feelings about whether they'd like to go back to the workplace. A variety of different factors contribute to an individual's decision to work from home and their employer's decision to permit work from home, including the worker's personal financial circumstances and desire to spend more time taking care of family members [6]. This section describes a model that predicts whether an individual with a remote-ready job will be allowed to and will choose to work from home.

3.1 Assumptions

1. *Hourly employees work for 52 weeks a year.* A worker's pay can be reported as either an hourly wage or a salary. However, the amount of paid time off, as well as the number of working weeks per year, varies for hourly jobs. This assumption simplifies the model and allows us to standardize the input data in terms of yearly earnings.

2. *An employee's choice on whether or not to work from home mostly depends on his or her own circumstances.* External influences, such as the COVID-19 pandemic, will have an impact on a worker's ability to work remotely or in-person but not how much they inherently desire to.

3. *Incomes can be represented without the inclusion of overtime pay.* We do not count overtime hours and pay as part of a person's annual income, since it is variable and often inconsistent from year to year. Overtime is seen as useful for a temporary boost in earnings, but not as a sustainable or recommended procedure for most workers [15].

4. *There will be no major changes in the types of jobs that people work in within the next few years.* We consider only the largest areas people currently are working in [12]; anticipating these shifts would be outside the intent and scope of our model.

3.2 Model Development

3.2.1 Factor Identification

First, we consider a variety of quantitative and categorical factors that can influence a person's decision and capability of working from home. The ones we determined to be most significant are listed below and used in our model. In accordance with Assumption 2, most of these will relate to the worker rather than changes in the broader socioeconomic landscape of the US or UK.

1. Age – The proportion of remote workers during the COVID-19 pandemic varied by age group [5]. 70% of workers 25-54 years of age worked remotely, followed by 21.9% of workers 55 years or older. Only 4.3% of workers 16-24 years of age worked remotely.

2. Children – Remote work can provide flexible working schedules that allow parents to balance their work and family lives more easily. 62% of working parents said that they would quit their jobs if they could not continue working remotely after the pandemic.

3. Occupation – The ability of an employee to work at home depends heavily on the industry in which they work [7]. Even among employees who can work remotely, the proportion of workers who choose to do so varies according to occupation.

4. Sex – Surveys show that, in general, women prefer remote work at higher rates than men [3]. This is especially true when comparing workers with young children [4].

5. Spouse employment – Having a working spouse or partner living within the same household could alleviate the burden faced by a worker and consequently influence their perspective on choosing to travel or be remote.

6. Elderly care – Those who are living with and care for the elderly may feel more inclined to work from home. This would enable them to readily respond to the needs of those older family members.

7. Education level – A person's degree of education, from high school diploma to a post-doctoral degree, has significant influence on their job opportunities and the type of work they will perform.

3.2.2 Collecting Input Worker Data

The U.S. Bureau of Labor Statistics publishes data from the annual American Time Use Survey (ATUS) [15, 16], a broad survey of workers about their economic and personal circumstances [10]. We used information from the 2017-2018 ATUS because it includes a leave and job flexibilities module. We collect and organize information for the seven variables listed in Section 3.2.1 for 1670 individuals whose jobs are remote-ready. Aside from the

input factors, we also note whether these individuals are working from home. Table 3.2.2 lists each of the input factors in addition to their data type and any applicable units. The final row describes the output.

Table 3.2.2: Representation and Data Type for Each Input Factor

Factor	Representation/Data Type
Age	Years (Quantitative)
Children	Number (Quantitative)
Occupation	22 Types Defined by the BLS [12] (Categorical)
Spouse Employment	Employed or Unemployed (Categorical)
Elderly Care	Caring or Not Caring for an Elder (Categorical)
Education Level	8 Levels Defined by the BLS [12] (Categorical)
Working from Home?	Yes or No

3.2.3 RuleFit Algorithm

To create predictions based on the input factors, we implemented RuleFit, an interpretable machine learning algorithm that generates decision rules from distributed random forests to develop a logistic classification model. The algorithm generates decision trees using the input variables and discards the predictions, leaving only the decision rules. These decision rules are then used as features for a logistic classification model. Because the generated decision rules can be formed from multiple input factors, RuleFit is particularly effective at capturing interactions between factors [14]. This is relevant due to the fact that many of the input factors we chose such as education and occupation are clearly related, and other factors are related in less intuitive ways that would not be successfully represented with other classification algorithms.

We train our RuleFit model on the data described in Subsection 3.2.2. After training, the model generates a set of decision rules that it uses to form a binary prediction for each individual introduced to it. A portion of the data set is left out of the training set to be used as testing data.

3.3 Results

Using the method described above, we create a RuleFit model with the H2O RuleFit algorithm. The model is constructed so that every decision rule has a length of three input factors. As a maximum, this prevents the model from overfitting to any particular factor, and as a minimum, it increases the interpretability of the decision rules. The number of rules is automatically chosen by diminishing returns in model variation. The model generated a total of 49 rules, two of which are displayed in Table 3.4.1 below. These rules had coefficients relatively high in magnitude, meaning when true, their impacts were relatively great on the output.

Table 3.4.1: RuleFit Model Evaluation Metrics

Coefficient	Support	Model Decision Rule
-0.224015	0.379042	(age < 52.5 or age is NA) & (education in 36, 43, 44, 45, 46 or education is NA) & (spouse works in 1 or spouse works is NA)
-0.295904	0.529940	(age >= 25.5 or age is NA) & (spouse works in 1 or spouse works is NA)

The first rule details that young, educated people with working or no spouses tend to work from home. The second rule shows that older adults with working or no spouses also tend to work from home. The support column describes what proportion of the training data each rule applied to.

3.4 Model Evaluation

We created a confusion matrix (shown in Table 3.4.2) for our model to provide a graphical illustration of the types of individuals it is correctly or incorrectly classifying. In particular, we are able to critically examine the number of false positives and false negatives, which are represented by the values in the bottom left and top right, respectively. In particular, we notice a high number of false negatives, where a person who was actually working remotely was predicted to not be doing so by the model.

Table 3.4.2: Model Confusion Matrix for Working from Home Prediction

	Model Yes	Model No
Actual Yes	268	597
Actual No	57	748

Our trained RuleFit model had a high specificity and a low sensitivity, as shown in Table 3.4.3. This means that our model has a fairly strong negativity bias. This could be improved in future iterations of the model by training it with more positive data.

Table 3.4.3: RuleFit Model Evaluation Metrics

Metric	Value
Sensitivity	0.3098
Specificity	0.9292

3.5 Strengths and Weaknesses

Our model has two very clear strengths. First, it considers interactions between features, which is not achieved by classification algorithms such as Naive Bayes. In addition, our model is easy to understand because it generates binary decision rules. Only three input factors are used in each rule, and only a handful of rules will apply to any given example, which sets our model apart in terms of intepretability. The rationality of patterns in the decision rules can be verified with the theoretical influence of each input factor on the output. This would not be possible with algorithms like pure random forests, which are typically treated as black boxes, as we are able to identify specific rules the decision tree algorithm followed when making a prediction.

A weakness of our model is that we do not account for travel times between the home and the workplace. For workers who already worked remotely, no data was available for what their commute times would have been. Additionally, our model does not capture any changes that may have occurred in sentiment toward remote work due to the COVID-19 pandemic. Finally, our model's negativity bias suggests that it overpredicts the probability that a person does not choose to work at home.

4 Part III: Just a Little Home-work

The transformations in the labor force to remote work have a significant impact on the economy and worker productivity in our society. Our model begins with an estimate of the percentages of workers who will work remotely in the cities modeled in Part I for 2024 and 2027. We then use the results of our models to quantify and rank the cities by the magnitude of impact of their populations of remote workers on the increase in productivity due to remote work and subsequently the increase in economic output (GDP) due to this growth in productivity.

4.1 Assumptions

1. *The decrease in spending towards transportation caused by remote working will have a negligible effect on the economy.* Even though remote workers no longer have to spend money on transportation for commuting, this change may become mitigated by individuals using the free time from commute to go out and travel to other locations to do their work or spend time.

2. *Worker behavior in the US and UK is approximately similar.* Given that the occupation that the worker will perform in the US and the UK is the same, it is reasonable to assume that the behavior and reaction to their work will be similar.

3. *All individuals within a worker's household will be one of the following: a spouse, child, or elderly individual.* This is done for simplification, and allows us to simulate a household as a probability distribution of individuals that are relevant in our Part II model.

4. *All spouses have a job and do work.* This is done due to the lack of data for household spouses and as such it is assumed that as long as they are in working age, they will have a job and do work.

5. *All measures of currency will be done in USD.* This is done for purposes of standardization to avoid confusion between currencies of the US and UK.

6. *Urbanization is a reasonable metric of education level.* There is no data for the education distribution of UK cities, so we assume that as a rural zone, Barry has a similar education distribution to Scranton, and as an urban zone, Liverpool to Seattle.

7. *Regional data for population and economic information is applicable to individual city.* There is little data necessary to generate probability distributions for each specific city. Rather, they should be well approximated by their surrounding areas.

8. *The average person works 8 hours a day, 5 days a week.* The 8 hour workday tends to the be the standard for most jobs across every sector; the same holds for the 5 day work week.

4.2 Model Development

Individuals who perform work remotely no longer are required to commute to their working environments, increasing the amount of free time they have from this time by an average of 8.5 hours [q]. We define this free time as an increase in productivity for workers because they may choose to utilize the extra time to work longer hours. This increase in work by individual employees from different work sectors will increase the economic output that they generate from their work [11]. The primary considerations in our model to quantify the impact of remote work on different cities are the increase in productivity by working hours and the increase in economic output per hour of work. To do this, we used a Monte Carlo simulation to predict the yearly number of hours that will be worked extra due to saved time from working remotely and determined the economic output produced from this work.

4.2.1 Developing the Monte Carlo

A Monte Carlo simulation finds the expected value of a function over random variables. The function of interest is total value added to the economy of the city, which is estimated by the sum across all people working remotely of the ratio of the total extra hours they work in a year to the total hours they work in a year, multiplied by the GDP per capita of that city. We additionally need to create simulated variables, given by distributions, for the input parameters to the model from question 2, to determine if a simulated person works from home. A person works from home if their job is remote-ready (given by the probabilities generated in Q1) multiplied by the binary value produced by the model in Q2.

The percentage of individuals who work longer hours due to the free time from remote jobs is 33% [8]. In addition, the probability distribution of the maximal amount of extra work an employee is able to do if they decide to use their extra time to work is taken from an Opinium survey [11]. To estimate the actual distribution of hours, we model voluntary extra working with a binomial distribution, where every week a simulated person has a 33% chance of working extra (extended to yearly with a binomial distribution, $n = 52$ weeks). Whenever they work extra, the person will always work the maximal possible amount.

We first generate the age and education of a person based on the population distribution of their city [20, 21, 22, 23, 24], taking only people aged 18 and up so that the previous model, which was trained on working adults, can be used. A person is randomly assigned into an occupation sector based on the proportions of the population employed in that sector in either 2024 or 2027, as given by the model of question 1. We generate households with properties coinciding with the independent variables in the second model for the Monte Carlo simulation (number of children, whether or not one's spouse works, and necessity of caring for one's elder). Specifically, the number of members in one's household can be approximated

by a Poisson distribution truncated at zero [18], with the probability distribution being given by

$$P(X = k \mid k > 0) = \frac{\lambda^k}{(e^\lambda - 1)k!} \tag{5}$$

We may calculate lambda using the average size of a household for each city and using the expected value formula for the Poisson distribution:

$$E[X] = \frac{\lambda}{1 - e^{-\lambda}} \tag{6}$$

Every household is guaranteed to have a member since the distribution is truncated at zero. Any other members of the household are prioritized in the order from highest to lowest of a working spouse, an elderly family member to care for, and finally children being added last. Spouses have a probability of being added equal to the percentage of people that are married in the city [20, 21, 22, 23, 24], and elderly people, for want of data, are given by a national probability of 29%, reported as the proportion of people in the US who take care of an elderly person [25].

4.3 Results

We simulated 100,000 people and calculated the total monetary impact in each city. This impact is given in USD per person, which allows for direct comparison of the values.

Table 4.3.1: Rank of City Based on Economic Impact by Remote Workers

City	Rank	2024 Impact	2027 Impact
Liverpool	1	26.05	100.24
Omaha	2	110.59	64.67
Scranton	3	133.81	55.09
Seattle	4	114.25	4.33
Barry	5	5.38	13.02

This data shows a reasonable increase for a city, with the GDP per capita experiencing marginal growth. Some cities, notably Liverpool, have a great benefit from remote work, with the 2027 value being much higher than the 2024 value, while in some cities, such as Seattle, the positive effects diminish, suggesting a regression the mean.

4.4 Strengths and Weaknesses

Our model takes into account a large number of dependent variables, allowing for a more accurate characterization of the differences in the population of each city.

Our model fails to conclude the possible benefits gained from sources outside of an increase in GDP and economic considerations. Additionally, due to time restrictions, we were unable to analyze the extent to which our model was stable, and not overly sensitive to varied input parameters, though its robust ability to be applied to five varying cities does point toits versatility.

5 Conclusion

5.1 Further Studies

Our first model failed to consider local fluctuations in the market that would affect it and instead assumed that such changes would either be negligible or continue throughout our predicted period. We could further investigate various societal trends and changes in values such as a higher willingness to work from home or trends in more specific occupations rather than solely industrial sectors.

Our second model was not able to properly account for various factors such as commute times. With more documented variables, more positive information, the parameters of the rule generation, and the depth of the tree, the model could become more consistent, robust, and complete with regard to the facets of each city and population.

Our third model quantified impact on a city as purely economic, and furthermore did so by estimating the local impact of a given worker as based on the GDP per capita. In reality, the economic value provided by workers is not constant; some add more value to the economy than others. This could be split up by industrial sector, as in the first model, for example, to create a more nuanced portrait of the impact.

5.2 Conclusion

In Part I, we predicted the number of jobs that are remote-ready for each job category by year for cities in the United States and United Kingdom for 2024 and 2027. We developed a multivariate regression to extrapolate the future values of remote-ready employment for the years 2024 and 2027. Then we took the regression and a matrix for the respective city and multiplied our predictions for each sector of the workforce. Finally, after using a weighted average of the aforementioned job categories, we were able to find the predictions for the percentage of jobs in each city that are remote-ready.

In Part II, we predicted whether or not an individual would work remotely or not based upon several factors such as age, occupation, education level, a working spouse, number of children, and the need to care for an elderly person. We used a RuleFit model to create a regression in terms of the many input variables to create a model which can predict whether a specific person will work remotely.

In Part III, we quantified the economic impact of working remotely for different cities in the years 2024 and 2027 and ranked them accordingly. We utilized our model from Part I to determine the number of remote-ready jobs in each city for those years and then applied our model from Part II to determine whether each individual from the city would work remotely based upon their demographic data. A Monte Carlo simulation was utilized to robustly quantify the economic impact for each city and rank them.

6 References

1. Dingel, Jonathan, and Brent Neiman. "How Many Jobs Can Be Done at Home?" *National Bureau of Economic Research*, National Bureau of Economic Research, Apr. 2020, https://www.nber.org/system/files/working_papers/w26948/w26948.pdf.

2. "Overview." *U.S. Bureau of Labor Statistics*, U.S. Bureau of Labor Statistics, https:// www.bls.gov/iag/home.htm.

3. Pelta, Rachel. "Survey: Men & Women Experience Remote Work Differently: FlexJobs." *FlexJobs Job Search Tips and Blog*, FlexJobs.com, 12 Jan. 2022, https:// www.flexjobs.com /blog/post/men-women-experience-remote-work-survey/.

4. Bloom, Nicholas. "Don't Let Employees Pick Their WFH Days." *Harvard Business Review*, 14 Sept. 2021, https://hbr.org/2021/05/dont-let-employees-pick-their-wfh-days.

5. Reynolds, Brie. "FlexJobs Survey: Working Parents Want Remote Work." *FlexJobs Job Search Tips and Blog*, FlexJobs.com, 12 Jan. 2022, https://www.flexjobs.com/blog /post/what-working-parents-want-at-work/.

6. Parker, Kim, et al. "Covid-19 Pandemic Continues to Reshape Work in America." *Pew Research Center's Social & Demographic Trends Project*, Pew Research Center, 16 Feb. 2022, https://www.pewresearch.org/social-trends/2022/02/16/covid-19-pandemic-continues-to-reshape-work-in-america/.

7. Callaham, Sheila. "Which Occupations Most Utilize Work at Home, and Which Have Room to Grow?" *Forbes*, Forbes Magazine, 19 Apr. 2020, https://www.forbes.com/sites/ sheilacallaham/2020/04/19/which-occupations-most-utilize-work-at-home-and-which-have-room-to-grow/.

8. Parker, Kim, et al. "How Coronavirus Has Changed the Way Americans Work." *Pew Research Center's Social & Demographic Trends Project*, Pew Research Center, 25 May 2021, https://www.pewresearch.org/social-trends/2020/12/09/how-the-coronavirus-outbreak-has-and-hasnt-changed-the-way-americans-work/.

9. Ozimek, Adam. "Where Remote Work Saves Commuters Most: Upwork." *RSS*, https://www.upwork.com/press/releases/where-remote-work-saves-commuters-most.

10. "Atus News Releases." *U.S. Bureau of Labor Statistics*, U.S. Bureau of Labor Statistics, 2021, https://www.bls.gov/tus/.

11. "The Potential Economic Impacts of a Flexible Working Culture." *Citrix*, 2019, https://www.citrix.com/content/dam/citrix/en_us/documents/white-paper/ economic-impacts-flexible-working-us-2019.pdf.

12. "American Time Use Survey (ATUS) Data Dictionary." *U.S. Bureau of Labor Statistics*, 2019, https://www.bls.gov/tus/atuscpscodebk18.pdf.

13. Friedman, Arik. “Quantifying the Impact of Remote Work on the Work-Life Balance.” *Medium*, Data at Atlassian, 18 Nov. 2020, https://medium.com/atlassiandata/quantifying-the-impact-of-remote-work-on-the-work-life-balance-a0cdac965e3a.

14. Molnar, Christoph. “Interpretable Machine Learning.” *5.6 RuleFit*, 21 Feb. 2022, https://christophm.github.io/interpretable-ml-book/rulefit.html.

15. “Working Overtime: Exploitation or Opportunity?” *RSS*, 24 Sept. 2021, https://memory.ai/timely-blog/working-overtime-exploitation-or-opportunity.

16. “About Atus Data.” *U.S. Bureau of Labor Statistics*, U.S. Bureau of Labor Statistics, 19 June 2019, https://www.bls.gov/tus/datafiles-2018.htm.

17. “Surprising Working from Home Productivity Statistics (2022).” *Apollo Technical LLC*, 17 Jan. 2022, https://www.apollotechnical.com/working-from-home-productivity-statistics/

18. Langemeier, Kathryn, and Maria D. Tito. “The Ability to Work Remotely: Measures and Implications.” *The Fed*, 2021, https://www.federalreserve.gov/econres/notes/feds-notes/the-ability-to-work-remotely-measures-and-implications-20211126.htm.

19. Jennings, Vic, et al. “Household Size and the Poisson Distribution.” *Journal of the Australian Population Association*, vol. 16, no. 1-2, 1999, pp. 65–84, https://doi.org/10.1007/bf03029455.

20. “Census Profile: Seattle-Tacoma-Bellevue, WA Metro Area.” *Census Reporter*, https://censusreporter.org/profiles/31000US42660-seattle-tacoma-bellevue-wa-metro-area/.

21. “Census Profile: Omaha-Council Bluffs, NE-IA Metro Area.” *Census Reporter*, https://censusreporter.org/profiles/31000US36540-omaha-council-bluffs-ne-ia-metro-area/.

22. “Census Profile: Scranton–Wilkes-Barre, PA Metro Area.” *Census Reporter*, https://censusreporter.org/profiles/31000US42540-scranton-wilkes-barre-pa-metro-area/.

23. “Liverpool Built-up area.” *Nomis*, https://www.nomisweb.co.uk/reports/localarea.

24. “Barry Parish.” *Nomis*, https://www.nomisweb.co.uk/reports/localarea.

25. “Caregiver Statistics.” *Caregiver Action Network*, 4 Nov. 2016, https://www.caregiveraction.org/resources/caregiver-statistics.

7 Appendix

7.1 emp_sectors.py

```
### Import required libraries
# Numpy for mathematical operations and matrices
# Scipy for executing regressions
# Os for reading given data from csv files
import numpy as np
from numpy import matmul
import scipy.optimize
import os

# Weights on each of the sectors
rho = np.array([0.19, 0.22, 0.23, 0.72, 0.67, 0.59, 0.34, 0.077, 0.31,
    0.41])

# Loop over all 5 cities
for city in os.listdir('employ-data'):
    # Load the data into an array
    data = np.genfromtxt(os.path.join('employ-data', city), delimiter=',',
    skip_header=1)
    # X^T . P^T = Y^T
    # This is the Markov process in form for regression
    # Thus X is the data of 'previous' years in row vector form, and Y is
   'current' years
    XT = data[0:-1, 1:]
    YT = data[1:, 1:]

    # Transition matrix by formula, and residual matrix U
    P = matmul(matmul(np.transpose(YT), XT),np.linalg.inv(matmul(np.
   transpose(XT), XT)))
    U = YT - matmul(XT,np.transpose(P))

    bars = np.mean(U, axis=0)                    # Mean of response var in
   each sector
    TSS = np.sum(np.square(P - bars), axis=0)    # Total-sum-of-squares in
   each sector
    RSS = np.sum(np.square(U), axis=0)           # Residual-sum-of-squares
   in each sector
    R2 = 1 - RSS/TSS                             # Coefficient of
   determination
    print(R2)

    # Predict future years with powers of P and save that to file
    predictions = []
    # UK data is through 2020, so 7 years must be predicted, whereas only
   6 for the US
    for i in range(1, (8 if city in ['barry.csv', 'liverpool.csv'] else 7)
   ):
        sec_dist = matmul(np.linalg.matrix_power(P, i), np.transpose(data)
   [1:, -1])
        predictions.append(np.append(sec_dist, np.sum(rho * sec_dist)))
```

```
#         predictions.append(matmul(np.linalg.matrix_power(P, i), np.
    transpose(data)[1:, -1]))
     np.savetxt(f'employ-pred/{city}', np.array(predictions), delimiter=','
    )
     np.savetxt(f'employ-pred/{city[0:-4]}.txt', np.array(predictions),
    delimiter=' & ', fmt='%.4f')
```

7.2 rulefit.py

```
# import required libraries
import h2o
import pandas as pd
h2o.init()
from sklearn.metrics import mean_squared_error
from h2o.estimators import H2ORuleFitEstimator

# import training data
train = h2o.import_file(path="workerdata.csv", col_types={"
    working_from_home": "enum", "occupation": "enum", "age": "int", "
    children": "int", "spouse_works": "enum", "care_for_elder": "enum", "
    education": "enum"})
# import testing data for each person
test = h2o.import_file(path="Seattle2027.csv", col_types={"occupation": "
    enum", "age": "int", "children": "int", "spouse_works": "enum", "
    care_for_elder": "enum", "education": "enum"})

# Set the predictors and response:
x = ["children", "age", "spouse_works", "care_for_elder", "occupation", "
    education"]
y = "working_from_home"

# Build and train the model:
rfit = H2ORuleFitEstimator(min_rule_length = 3, max_rule_length = 3)
rfit.train(training_frame=train, validation_frame=train, x=x, y=y)

#Retrieve the rule importances:
f = open("rules.txt", "w")
rules = rfit._model_json['output']['rule_importance'].as_data_frame()
f.write(rules.to_string())
f.close()

# # Predict on the training data:
validation_results = rfit.predict(train).as_data_frame()
validation = train.as_data_frame()
validation['preds'] = validation_results['predict']

#calculate true/false positives/negatives
tn = len(validation.loc[(validation.working_from_home == 2) & (validation.
    preds == 2)])
fn = len(validation.loc[(validation.working_from_home == 1) & (validation.
    preds == 2)])
tp = len(validation.loc[(validation.working_from_home == 1) & (validation.
    preds == 1)])
```

```
fp = len(validation.loc[(validation.working_from_home == 2) & (validation.
    preds == 1)])
#calculate sensitivity and specificity
sensitivity = tp/(fn+tp)
specificity = tn/(fp+tn)
#print evaluations
print("True Positives: " + str(tp))
print("False Positives: " + str(fp))
print("True Negatives: " + str(tn))
print("False Negatives: " + str(fn))
print("Sensitivity: " + str(sensitivity))
print("Specificity: " + str(specificity))

# Predict on the city data:
results = rfit.predict(test).as_data_frame()
test_pd = test.as_data_frame()
test_pd['preds'] = results['predict']
test_pd.to_csv('Seattle2027preds.csv',index=True)
```

7.3 simulate_people.py

```
from random import randint, uniform
from numpy.random import poisson, binomial

def HH_SIZE(city):
    lamb = {
        'Seattle' : 2.23,
        'Omaha' : 2.23,
        'Scranton': 1.98,
        'Liverpool': 1.86,
        'Barry' : 1.98,
    }
    N = poisson(lamb[city])
    return N if N != 0 else HH_SIZE(city)

def HOUSEHOLD(city, n):
    spouse = False
    elder = False
    if n > 1 and uniform(0, 100) < SPOUSE[city]:
        spouse =  True
        n -= 1
    if n > 1 and uniform(0, 100) < ELDER:
        elder = True
        n -= 1
    children = n - 1
    return children, spouse, elder

SPOUSE = {
    'Seattle' : 62,
    'Omaha': 43,
    'Scranton': 56,
    'Liverpool' : 37.7,
```

```
    'Barry': 42.8
}

ELDER = 0.29

MAX_H = lambda x: {
    0<=x<7: 0,
    7<=x<14: 0.5,
    14<=x<28: 1.5,
    28<=x<45: 2.5,
    45<=x<60: 3.5,
    60<=x<72: 4.5,
    72<=x<81: 5.5,
    81<=x<86: 6.5,
    86<=x<93: 7.5,
    93<=x<100: 8.5
}[True]

GDPC = {
    'Seattle': 80833,
    'Omaha': 60246,
    'Scranton': 37417,
    'Liverpool': 50400,
    'Barry': 32849
}

AGE = {
    'Seattle': lambda x: {
        0<=x<12: randint(0, 9),
        12<=x<24: randint(10, 19),
        24<=x<39: randint(20, 29),
        39<=x<55: randint(30, 39),
        55<=x<68: randint(40, 49),
        68<=x<81: randint(50, 59),
        81<=x<92: randint(60, 69),
        92<=x<98: randint(70, 79),
        98<=x<100: randint(80, 89)
    }[True],
    'Omaha': lambda x: {
        0<=x<14: randint(0, 9),
        14<=x<28: randint(10, 19),
        28<=x<41: randint(20, 29),
        41<=x<56: randint(30, 39),
        56<=x<68: randint(40, 49),
        68<=x<80: randint(50, 59),
        80<=x<91: randint(60, 69),
        91<=x<97: randint(70, 79),
        97<=x<100: randint(80, 89)
    }[True],
    'Scranton': lambda x: {
        0<=x<11: randint(0, 9),
        11<=x<22: randint(10, 19),
        22<=x<34: randint(20, 29),
        34<=x<46: randint(30, 39),
```

```
        46<=x<58: randint(40, 49),
        58<=x<72: randint(50, 59),
        72<=x<86: randint(60, 69),
        86<=x<95: randint(70, 79),
        95<=x<100: randint(80, 89)
    }[True],
    'Liverpool': lambda x: {
        0<=x<20.1: randint(0, 17),
        20.1<=x<23.4: randint(18, 19),
        23.4<=x<32.2: randint(20, 24),
        32.2<=x<39.6: randint(25, 29),
        39.6<=x<58.9: randint(30, 44),
        58.9<=x<78.7: randint(45, 59),
        78.7<=x<84.5: randint(60, 64),
        84.5<=x<92.7: randint(65, 74),
        92.7<=x<98.3: randint(75, 84),
        98.3<=x<100: randint(85, 89)
    }[True],
    'Barry': lambda x: {
        0<=x<23.3: randint(0, 17),
        23.3<=x<25.8: randint(18, 19),
        25.8<=x<31.8: randint(20, 24),
        31.8<=x<38.5: randint(25, 29),
        38.5<=x<58.9: randint(30, 44),
        58.9<=x<79: randint(45, 59),
        79<=x<84.8: randint(60, 64),
        84.8<=x<92.8: randint(65, 74),
        92.8<=x<97.8: randint(75, 84),
        97.8<=x<100: randint(85, 89)
    }[True],
}

EDU = {
    'Seattle' : lambda x: {
        0<=x<7: 38,
        7<=x<26: 39,
        26<=x<55: 40,
        55<=x<82: 43,
        82<=x<100: 44
    }[True],
    'Omaha' : lambda x: {
        0<=x<7: 38,
        7<=x<31: 39,
        31<=x<62: 40,
        62<=x<87: 43,
        87<=x<100: 44
    }[True],
    'Scranton' : lambda x: {
        0<=x<9: 38,
        9<=x<48: 39,
        48<=x<76: 40,
        76<=x<91: 43,
        91<=x<100: 44
    }[True],
```

```
    'Liverpool' : lambda x: {
        0<=x<7: 38,
        7<=x<26: 39,
        26<=x<55: 40,
        55<=x<82: 43,
        82<=x<100: 44
    }[True],
    'Barry' : lambda x: {
        0<=x<9: 38,
        9<=x<48: 39,
        48<=x<76: 40,
        76<=x<91: 43,
        91<=x<100: 44
    }[True]
}

SECTOR = {
    'Seattle' : {
        2024: lambda x: {
            0<=x<8.29: 'MLC',
            8.29<=x<13.36: 'MFG',
            13.36<=x<33.19: 'TTU',
            33.19<=x<42.82: 'INF',
            42.82<=x<47.91: 'FIN',
            47.91<=x<65.31: 'PBS',
            65.31<=x<78.6: 'EHS',
            78.6<=x<86.05: 'LHO',
            86.05<=x<89.35: 'OTH',
            89.35<=x<100: 'GOV'
        }[True],
        2027: lambda x: {
            0<=x<9.62: 'MLC',
            9.62<=x<10.4: 'MFG',
            10.4<=x<30.36: 'TTU',
            30.36<=x<42.02: 'INF',
            42.02<=x<47.15: 'FIN',
            47.15<=x<65.51: 'PBS',
            65.51<=x<79.64: 'EHS',
            79.64<=x<86.26: 'LHO',
            86.26<=x<89.57: 'OTH',
            89.57<=x<100: 'GOV'
        }[True]
    },

    'Omaha': {
        2024: lambda x: {
            0<=x<5.66: 'MLC',
            5.66<=x<12.12: 'MFG',
            12.12<=x<30.3: 'TTU',
            30.3<=x<32.02: 'INF',
            32.02<=x<41.24: 'FIN',
            41.24<=x<55.55: 'PBS',
            55.55<=x<72.31: 'EHS',
            72.31<=x<82.18: 'LHO',
```

```
            82.18<=x<85.99: 'OTH',
            85.99<=x<100: 'GOV'
        }[True],
        2027: lambda x: {
            0<=x<5.46: 'MLC',
            5.46<=x<11.94: 'MFG',
            11.94<=x<30.02: 'TTU',
            30.02<=x<31.62: 'INF',
            31.62<=x<40.81: 'FIN',
            40.81<=x<55: 'PBS',
            55<=x<71.85: 'EHS',
            71.85<=x<81.9: 'LHO',
            81.9<=x<85.76: 'OTH',
            85.76<=x<100: 'GOV'
        }[True]
    },

    'Scranton': {
        2024: lambda x: {
            0<=x<4.02: 'MLC',
            4.02<=x<14.31: 'MFG',
            14.31<=x<40.26: 'TTU',
            40.26<=x<41.04: 'INF',
            41.04<=x<46.01: 'FIN',
            46.01<=x<56.45: 'PBS',
            56.45<=x<76.87: 'EHS',
            76.87<=x<85.28: 'LHO',
            85.28<=x<88.58: 'OTH',
            88.58<=x<100: 'GOV'
        }[True],
        2027: lambda x: {
            0<=x<3.99: 'MLC',
            3.99<=x<14.54: 'MFG',
            14.54<=x<40.17: 'TTU',
            40.17<=x<40.82: 'INF',
            40.82<=x<45.9: 'FIN',
            45.9<=x<56.43: 'PBS',
            56.43<=x<77.29: 'EHS',
            77.29<=x<85.68: 'LHO',
            85.68<=x<88.87: 'OTH',
            88.87<=x<100: 'GOV'
        }[True]
    },

    'Liverpool' : {
        2024: lambda x: {
            0<=x<5.34: 'MLC',
            5.34<=x<14.36: 'MFG',
            14.36<=x<36.8: 'TTU',
            36.8<=x<39.54: 'INF',
            39.54<=x<43.08: 'FIN',
            43.08<=x<63.39: 'PBS',
            63.39<=x<84.42: 'EHS',
            84.42<=x<92.38: 'LHO',
```

```
            92.38<=x<98.01: 'OTH',
            98.01<=x<100: 'GOV'
        }[True],
        2027: lambda x: {
            0<=x<5.02: 'MLC',
            5.02<=x<10.67: 'MFG',
            10.67<=x<33.3: 'TTU',
            33.3<=x<37.66: 'INF',
            37.66<=x<40.67: 'FIN',
            40.67<=x<67.74: 'PBS',
            67.74<=x<90.87: 'EHS',
            90.87<=x<98.51: 'LHO',
            98.51<=x<98.92: 'OTH',
            98.92<=x<100: 'GOV'
        }[True]
    },

    'Barry': {
        2024: lambda x: {
            0<=x<9.55: 'MLC',
            9.55<=x<21.05: 'MFG',
            21.05<=x<42.49: 'TTU',
            42.49<=x<43.61: 'INF',
            43.61<=x<47.59: 'FIN',
            47.59<=x<56.33: 'PBS',
            56.33<=x<82.73: 'EHS',
            82.73<=x<89.9: 'LHO',
            89.9<=x<93.64: 'OTH',
            93.64<=x<100: 'GOV'
        }[True],
        2027: lambda x: {
            0<=x<6.09: 'MLC',
            6.09<=x<19.12: 'MFG',
            19.12<=x<41.68: 'TTU',
            41.68<=x<42.36: 'INF',
            42.36<=x<45.98: 'FIN',
            45.98<=x<54.47: 'PBS',
            54.47<=x<83.58: 'EHS',
            83.58<=x<88: 'LHO',
            88<=x<92.65: 'OTH',
            92.65<=x<100: 'GOV'
        }[True]
    }
}

# work code given category
JOB = {
    'EHS': lambda x: {
        0<=x<15.3: 8,
        15.3<x<=100: randint(10, 11),
    }[True],
    'FIN': lambda x: {
        0<=x<100: 2
    }[True],
```

```
    'PBS': lambda x: {
        0<=x<10.8: 1,
        10.8<=x<57.4: 5,
        57.4<=x<100: 17,
    }[True],
    'LHO': lambda x: {
        0<=x<14.4: 9,
        14.4<=x<100: 13
    }[True],
    'MLC': lambda x: {
        0<=x<1.35: 18,
        1.35<=x<100: 19
    }[True],
    'TTU' : lambda x: {
        0<=x<22.6: 22,
        22.6<=x<100: 16,
    }[True],
    'INF' : lambda x: {
        0<=x<100: 3
    }[True],
    'OTH' : lambda x: {
        0<=x<100: 6
    }[True],
    'MFG' : lambda x: {
        0<=x<100: 21
    }[True],
    'GOV' : lambda x: {
        0<=x<100: 7
    }[True]
}

year = 2027
city = 'Scranton'

with open(f'{city}{year}.csv', 'w') as f:
    f.write('occupation,age,children,spouse_works,education,care_for_elder
   ,VA\n')
    for i in range(100000):
        age = 0
        while age < 18:
            age = AGE[city](uniform(0, 100))
        edu = EDU[city](uniform(0, 100))
        sector = SECTOR[city][year](uniform(0, 100))
        job = JOB[sector](uniform(0, 100))
        hh_size = HH_SIZE(city)
        children, spouse, elder = HOUSEHOLD(city, hh_size)
        va = MAX_H(uniform(0, 100)) * binomial(52, 0.33) * GDPC[city]
        f.write(f'{job},{age},{children},{spouse},{edu},{elder},{va}\n')
```

7.4 monte_carlo.py

```
import pandas as pd
```

```
import numpy as np

city = 'Seattle'
year = 2027

f = lambda x: (2 - x[1]) * x[0] / (52 * 5 * 8)
res = pd.read_csv(f'{city}{year}preds.csv')
f_res = res[["VA", "preds"]].apply(f, axis=1, raw=True)
print(sum(f_res))
```