MGT 6203 Group Project Proposal

TEAM INFORMATION (1 point)

Team #: 68

Team Members:

1. Team Member 1
    a. Name: Gabriel Mink
    b. GTID: 903738167: gmink3@gatech.edu
    c. Personal Background: I've been a Data Scientist/Machine Learning Engineer for the last 4 years mostly in the biotech industry, I graduated from UCSD majoring in Biology and minoring in Computer Science, I've worked on purchase prediction recommender systems and multivariate financial forecasting models.

2. Team Member 2
    a. Name: Bella (Yifei) Ding
    b. GTID: 903131776: yding302@gatech.edu
    c. Personal Background: I recently became a Data Engineer in the tech industry. Previously I completed my degree in Mechanical Engineering and worked in the diesel automotive industry as a Test Engineer for 2 years. I have worked on some analytics related projects mostly in school. One previous research project was on building an image classification model for bacteria lab images.

3. Team Member 3
    a. Name: Vincent Pan
    b. GTID: 903847411: vipan@gatech.edu
    c. Personal Background: I've been a Data Scientist since graduating university in Actuarial Studies and Mathematics. Currently, I live in the San Francisco Bay Area and am working to understand how WhatsApp can better serve large businesses and deliver communications they have with their customers.

4. Team Member 4
    a. Name: Nikolos Lahanis
    b. GTID: 903674177: nlahanis3@gatech.edu
    c. Personal Background: A recent mechanical engineering undergrad from NC State University, I currently live and work in the San Francisco Bay Area as an Intelligent Automation Consultant. Previous analytics projects I have worked on include an Investment-Grade Asset Ranking System and a Real Estate Property Valuation Tool.

5. Team Member 5
    a. Name: Rahul Sati
    b. GT Id: 903549883: rsati3@gatech.edu

c. Personal Background: I am currently working as a Data Science Manager with a leading e-commerce company. I have completed my previous Masters in Industrial Engineering, focusing on operations research and have done my Bachelors in Mechanical Engineering. I have done data science projects in the area of forecasting, experimentation (A/B Testing) and simulation (simulating network design for Walmart).

OBJECTIVE/PROBLEM (5 points)

Project Title: Predict Blood Pressure Through Various Factors

Background Information on chosen project topic:

Nowadays, as the awareness of maintaining health arises, more and more people started to pay more attention to a healthier lifestyle by, for example, choosing a healthier diet. Therefore, our team would like to investigate how much influence different factors have on a person's blood pressure through a National Health and Nutrition Examination Survey (NHANES) dataset. The modeling result could be very helpful to be future reference when providing health suggestions for patients.

Problem Statement (clear and concise statement explaining purpose of your analysis and investigation):

This project will investigate the effect various factors, such as nutritional intake, diet and demographics, have on a person's blood pressure, in order to find out which factor(s) will have significant influence. Then this project will also predict blood pressure through those chosen factors.

State your Primary Research Question (RQ):

Does blood pressure (or other health conditions/diseases) in the adults and children in the United States correlate with nutritional intake, diet, demographics, medical history and lifestyle? Can we predict the risk of having a disease based on these factors?

Add some possible Supporting Research Questions (2-4 RQs that support problem statement):

1. Relationship of blood pressure with diet, demographics and other such factors?
2. Relationship of diabetes with diet, demographics and lifestyle?

Business Justification: (Why is this problem interesting to solve from a business viewpoint? Try to quantify the financial, marketing or operational aspects and implications of this problem, as if you were running a company, non-profit organization, city or government that is encountering this problem.)

The National Health and Nutrition Examination Survey (NHANES) is a program of studies designed to assess the health and nutritional status of adults and children in the United States. NHANES is a major program of the National Center for Health Statistics (NCHS). NCHS is part of the Centers for Disease Control and Prevention (CDC) and has the responsibility for producing vital and health statistics for the Nation.

In this project, the health, and nutritional data from NHANES is used to analyze and predict key factors leading to critical diseases. For example, high blood pressure can lead to heart diseases. According to the CDC heart disease is the leading cause of death for men, women, and people of most racial and ethnic groups in the United States. About 697,000 people in the United States died from heart disease in 2020, that's 1 in every 5 deaths. Heart disease costs the United States about $229 billion each year, this includes the cost of healthcare services, medicines, and lost productivity due to death.

Similarly, we can model other critical health factors and diseases using the given data. Such application of data science and machine learning in healthcare will improve the ability of public health organizations to predict the outbreak of disease, improve the prevention of disease and enhance the quality of life. It will also improve the accuracy and speed of identifying patients at highest risk of disease.

DATASET/PLAN FOR DATA (4 points)

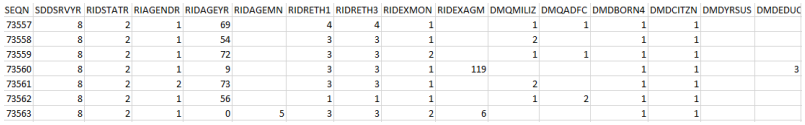
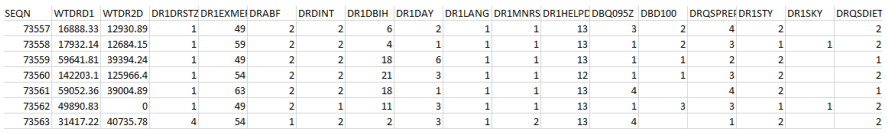
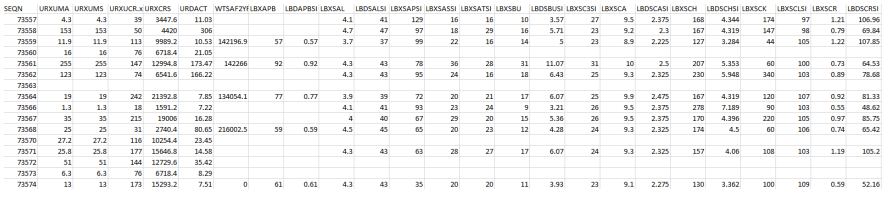
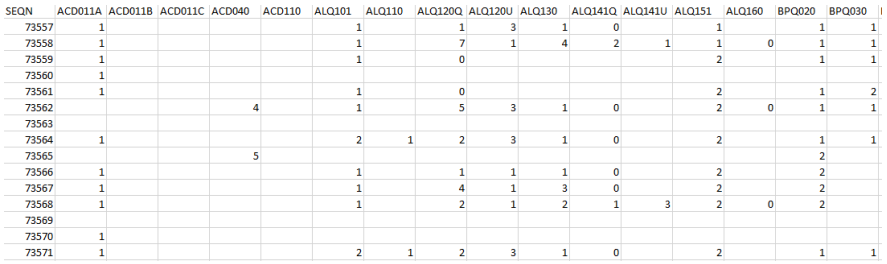
Data Sources (links, attachments, etc.):

- National Health and Nutrition Examination Survey from Kaggle.com from the Centers of Disease Control and Prevention (CDC)
- [Online] https://www.kaggle.com/datasets/cdc/national-health-and-nutrition-examination-survey

Data Description (describe each of your data sources, include screenshots of a few rows of data):

Title:	National Health and Nutrition Examination Survey
Source:	Kaggle
Origin:	NHANES datasets from the Centers of Disease Control and Prevention (CDC)
Dataset date:	2013-2014
Number of Files/ Datasets	6
Description:	*"The National Health and Nutrition Examination Survey (NHANES) is a program of studies designed to assess the health and nutritional status of adults and children in the United States. The survey is unique in that it combines interviews and physical examinations"* From a quick first glance, each of the dataset refer to an individual by the column: seqn which we can use to join the datasets together
Primary Key Across Datasets	SEQN: Respondent Sequence Number

Dataset Ref	Dataset Name	Description and Screenshots (See Appendix for more rows)
1	demographic.csv [Data Dictionary]	*Description:* Demographic data of the individual. This includes Age (DMDHRAGE), Race (RIDRETH1), and Gender (RIAGENDR)

	Number of Columns: 47 *Number of Rows: 10.2k*	*Example Rows:* 
2	diet.csv [Data Dictionary] *Number of Columns: 168* *Number of Rows: 9.8k*	Data on the individual's first day diet. This is a questionnaire on how much e.g. salt (DBD100), amount of food was eaten (DR1_300), and tap water was drunk (DR1_320Z). *Example Rows:* 
3	labs.csv [Data Dictionary] *Number of Columns: 224* *Number of Rows: 9.8k*	*Description:* Data on the Labs. Not all tests were conducted for all individuals. However, some examples of test/ detection include: Monocyte number (LBDMONO), Platelet count (LBXPLTSI), and Potassium (LBXSKSI) *Example Rows:* 
4	questionnaire.csv [Data Dictionary] *Number of Columns: 953* *Number of Rows: 10.2k*	*Description:* A questionnaire asking about the health and behaviors of the individual. This includes time spent outdoors (DED125), whether they have been diagnosed with osteoporosis (OSQ060), and whether they are taking prescribed medicine (BPQ050A). *Example Rows:* 
5	examinations.csv [Data Dictionary] *Number of Columns: 424* *Number of Rows: 9.8k*	*Description:* Data on the health examination for the individual. This includes questions on whether they are currently pregnant or breastfeeding (CSQQ241), grip test (MGDEXSTS), and Total abdominal fat mass (DXXTATM) *Example Rows:*

		SEQN PEASCST1 PEASCTM1 PEASCCT1 BPXCHR BPAARM BPACSZ BPXPLS BPXPULS BPXPTY BPXML1 BPXSY1 BPXDI1 BPAEN1 BPXSY2 BPXDI2 BPAEN2 BPXSY3 BPXDI3 BPAEN3 73557 1 620 1 4 86 1 1 140 122 72 2 114 76 2 102 74 2 73558 1 766 1 4 74 1 1 170 156 62 2 160 80 2 156 42 2 73559 1 665 1 4 68 1 1 160 140 90 2 140 76 2 146 80 2 73560 1 803 1 2 64 1 1 130 108 38 2 102 34 2 104 38 2 73561 1 949 1 3 92 1 1 170 136 86 2 134 88 1 142 86 2 73562 1 1064 1 5 60 1 1 180 160 84 2 158 82 2 154 80 2 73563 1 90 152 1 73564 1 954 1 5 82 1 1 150 118 80 2 124 80 2 126 82 2 73566 1 625 1 4 86 1 1 140 128 74 2 124 72 2 114 72 2 73567 1 932 1 3 70 1 1 170 140 78 2 142 78 2 142 76 2 73568 1 585 1 3 70 1 1 120 106 60 2 100 62 2 108 66 2 73570 1 710 1 2 78 1 1 130 102 44 2 104 42 2 102 40 2 73571 1 1193 1 4 72 1 1 150 124 68 2 2 128 68 2 73572 1 557 1 3 80 1 1 110 88 54 2 88 54 2 94 58 2 73573 1 599 1 3 74 1 1 120 94 62 2 94 60 2 92 54 2 73574 1 710 1 3 98 1 1 140 122 56 2 124 66 2 118 64 2 73575 1 291 110 1 73576 1 584 1 4 62 1 1 130 108 68 2 110 62 2 110 70 2 73577 1 750 1 4 56 1 1 140 118 74 2 118 74 2 122 78 2 73578 1 607 1 5 68 1 1 140 120 58 2 124 64 2 124 70 2
6	medications.csv [Data Dictionary] *Number of Columns: 13* *Number of Rows: 20.2k*	*Description:* Data on any medication (if any) the individual is taking. This includes the reason they are prescribed for (e.g. Muscle spasm, insomnia etc.) *Example Rows:* SEQN RXDUSE RXDDRUG RXDDRGID RXQSEEN RXDDAYS RXDRSC1 RXDRSC2 RXDRSC3 RXDRSD1 RXDRSD2 RXDRSD3 RXDCOUNT 73557 1 99999 2 73557 1 INSULIN d00262 2 1460 E11 Type 2 diabetes mellitus 2 73558 1 GABAPEN d03182 1 243 G25.81 Restless legs syndrome 4 73558 1 INSULIN G d04538 1 365 E11 Type 2 diabetes mellitus 4 73558 1 OLMESAR d04801 1 14 E11.2 Type 2 diabetes mellitus with k 4 73558 1 SIMVASTA d00746 1 61 E78.0 Pure hypercholesterolemia 4 73559 1 INSULIN A d04697 1 365 E11 Type 2 diabetes mellitus 5 73559 1 INSULIN G d04538 1 4380 E11 Type 2 diabetes mellitus 5 73559 1 PANCRELI d01002 1 365 K86.9 Disease of pancreas, unspecifi 5 73559 1 SIMVASTA d00746 1 2920 E78.0 Pure hypercholesterolemia 5 73559 1 VALSARTA d04113 1 3650 I10 Essential (primary) hypertensi 5

Key Variables: (which ones will be considered independent and dependent? Are you going to create new variables? What variables do you hypothesize beforehand to be most important?)

- Our goal is to model blood pressure *(dependent variable)* as recorded in the Demographics.csv dataset.
- We will be utilizing the data in demographics, diet, labs, examinations, and medications datasets *(independent variables)* to predict it (excluding the identifier, and the dependent variable).
    - Some examples of variables that we want investigate are:
        - Gender and age (and whether they match existing literature)
        - Alcohol and Caffeine,
        - Cholesterol levels, and
        - Drug Use.
    - We expect many variables to not be predictive (no silver bullet), but some may pop up and surprise us if we cast a wide net.
- We may create *new variables* to simplify the predictors and make them more meaningful.
    - There is a lot of missing data (e.g. questionnaire responses), and a lot of similar/ related predictors (e.g. do you feel pain, where do you feel pain, how long have you felt pain). To handle this, we may group up some of the existing variables, use indicators to capture the presence of it, or use dimensional reducing techniques (e.g. PCAs).

If we have time towards the end of the project, we may also try to model other health outcomes/ results to see whether we can predict them with accuracy or have additional interesting insights.

APPROACH/METHODOLOGY (8 points)

Planned Approach (In paragraph(s), describe the approach you will take and what are the models you will try to use? Mention any data transformations that would need to happen. How do you plan to compare your models? How do you plan to train and optimize your model hyper-parameters?))

- EDA and Feature Selection
    - First we plan on conducting exploratory data analysis on each of the datasets, assessing things such as sparsity, distributions, outliers. From data we deem okay to use, we will then run correlation tests to prune features and see how well raw data and engineered features relate to one another along with our response variable, blood pressure, and remove insignificant ones from the final feature list. Additional feature selection using RFE or model-specific ones might be leveraged as well.
- Model selection
    - A suite of mostly linear and non-linear regression models such as linear-log, log-linear, log-log, and various polynomial models, other models such as XGBRegression may be explored as well , will be used to compare results, and the top one with the least RMSE will be chosen and used for our final model. Models might be compared using Cross-validation in order to prevent overfitting throughout our comparisons.
- Data Transformations
    - We hope to experiment with both non-parametric models where distributions shouldn't matter too much, and parametric models as well, which might require log transformations, min-max scaling, and square root transformations in order to keep the data points on a relative scale to one another.. However other transformations such as one-hot encoding categorical variables, creating ratios of multiple numeric fields, and leveraging dimensionality reduction techniques such as PCA or K-mean clustering to reduce high-dimensional data sources may be used.
- Training and optimizing
    - We plan to use a train validation and test split with ratios around 60%, 20% and 20% respectively with the training data used to enable the model to learn the feature weights, and the validation data used to tune the hyperparameters. Hyperparameter tuning will mainly be done manually with careful attention being paid to prevent over or under fitting of the model, and if time permits automatic methods such as gridsearch and hyperopt may be used Our test data will be held out and used to report the final performance of the model

Anticipated Conclusions/Hypothesis (what results do you expect, how will you approach lead you to determining the final conclusion of your analysis) Note: At the end of the project, you do not have to be correct or have acceptable accuracy, the purpose is to walk us through an analysis that gives the reader insight into the conclusion regarding your objective/problem statement

- Due to the high dimensionality of our data we assume many features looked at will not yield much predictive power. However there are many behaviors that medical science deems correlated to an individual having high blood pressure which we expect to see also represented in the model as feature weights/importances. Some of those features surround the use of tobacco and alcohol. We also expect overall diet and nutrition to play a big role with diets that veer off the recommended amounts in both nutrient breakdown and calories to impact blood pressure, both in terms of high and low depending on how excessive or restrictive each individual is.

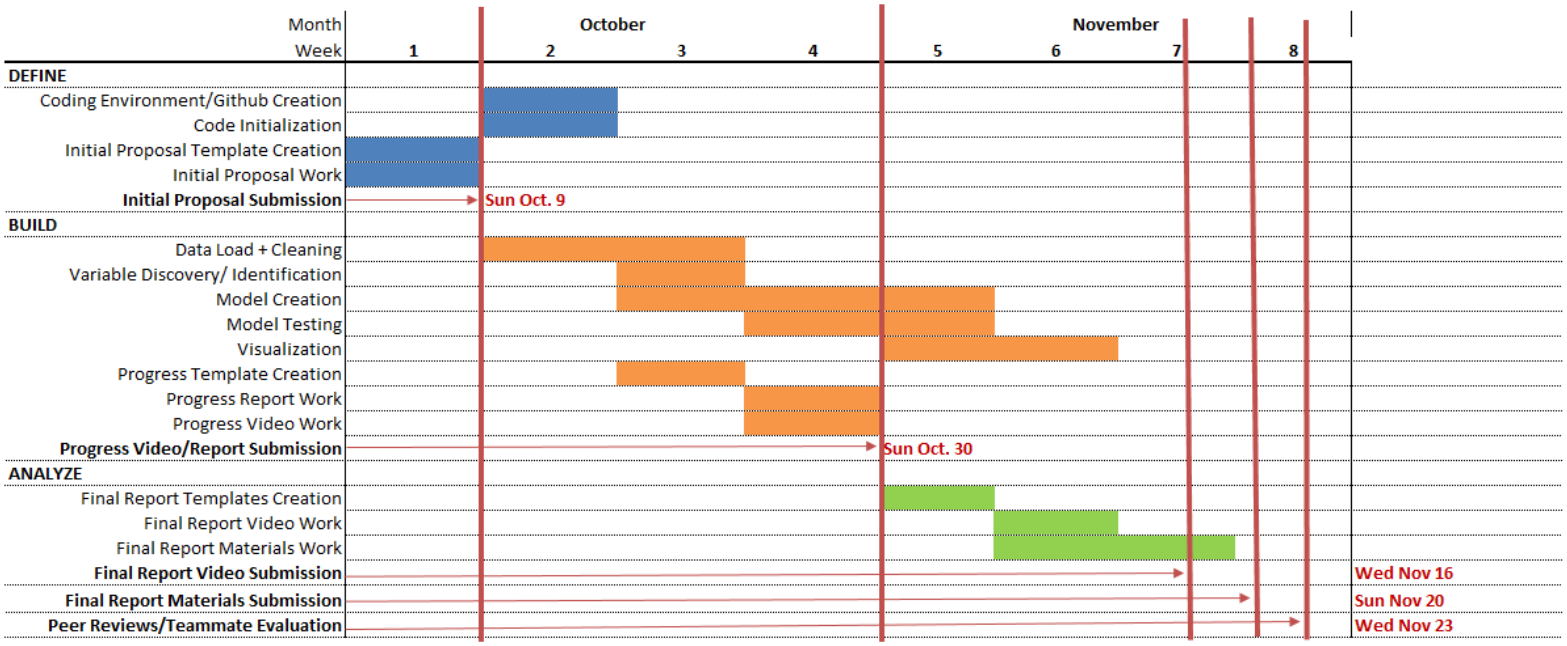
- The approach we take is going to be data-centric, we will not exclude anything based off intuition alone, however we will include and engineer features based off of our understanding of these scientifically proven correlations. In the end we hope to build as accurate of a model as possible, and by looking at the highly impactful features, give back to the audience some food for thought as to what factor inside and outside of their control may be involved with their blood pressure.

What business decisions will be impacted by the results of your analysis? What could be some benefits?

Since our results will provide information on which factors could have a significant influence on a patient's blood pressure, the results can impact the health suggestions that are given to patients. We would like to understand what things a patient can and cannot change to improve health outcomes, which would be beneficial for healthcare providers and/or certain industries (such as food, medical, etc). Healthcare providers can determine treatments and the food industry can develop more suitable products that lead to a healthy diet, according to our analysis results.

Our analysis results can provide many benefits. First of all, this dataset prioritizes the behavior change of the patients by surveying diet information. It also has a broad selection of factors and considers the lowest hanging fruit, which could provide insights about what small changes can lead to big effects. The size of the dataset is also decent, which would give us a nice pool of data to use for analysis and prediction.

PROJECT TIMELINE/PLANNING (2 points)**Project Timeline/Mention key dates you hope to**



achieve certain milestones by:

Table : General Project Timeline

As a general guideline, we are looking to have the following three general phases of our timeline done by these established dates:

1. DEFINE - Friday, October 14th

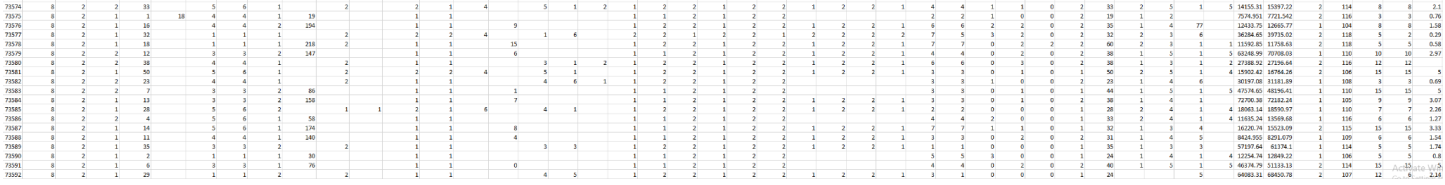
2. BUILD - Friday, November 11th
3. ANALYZE - Friday, November 18th

By establishing our project timeline with these deadlines in mind, we can not only reduce our overall workload as it pertains to this project, but also increase the quality of our final deliverables.

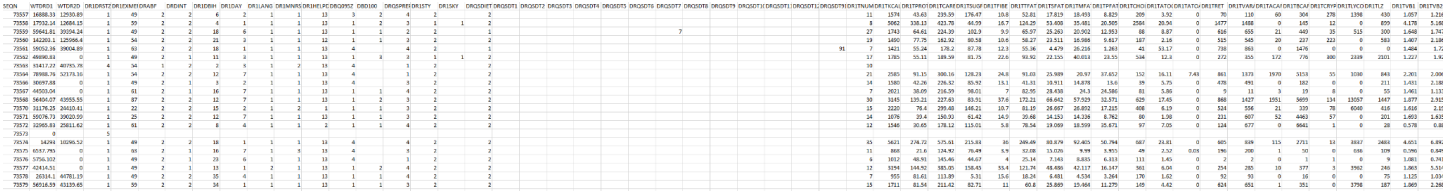
Appendix (any preliminary figures or charts that you would like to include):

Additional Screenshots for Data

Demographics.csv

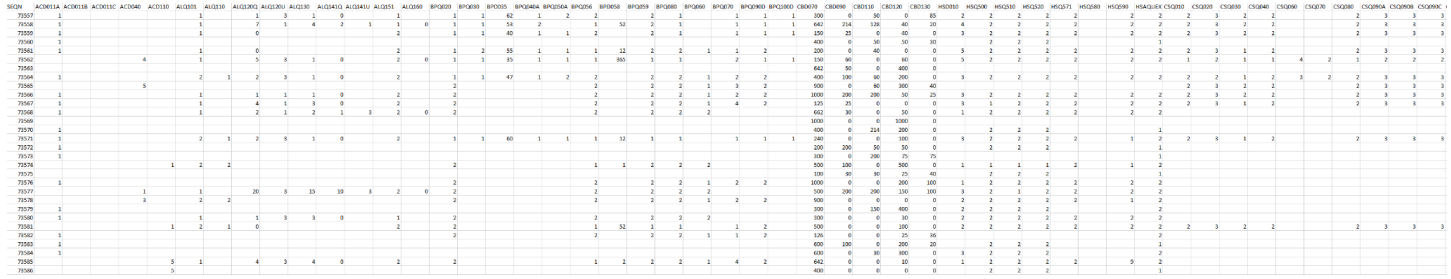


Diet.csv

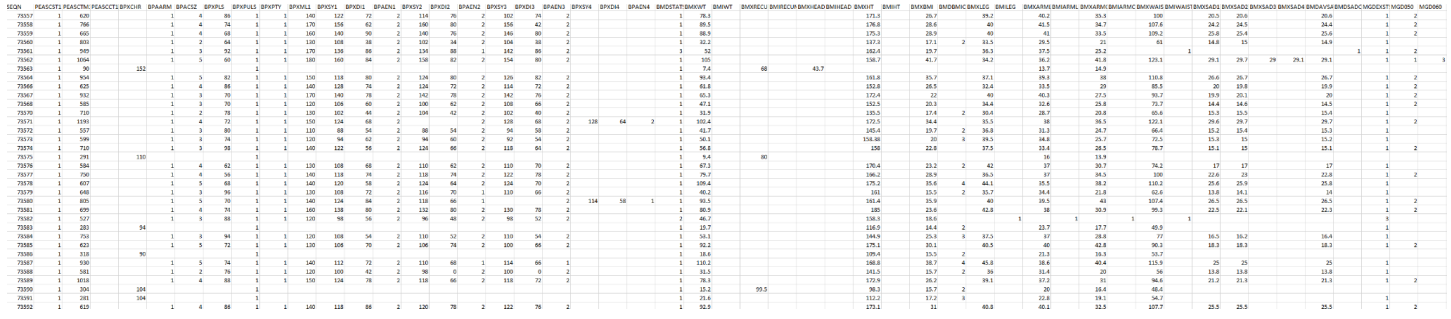


Labs.csv

Questionnaire.csv



Examination.csv



Medications.csv

SEQN	RXDUSE	RXDDRUG	RXDDRGID	RXQSEEN	RXDDAYS	RXDRSC1	RXDRSC2	RXDRSC3	RXDRSD1	RXDRSD2	RXDRSD3	RXDCOU
73557	1	99999										
73557	1	INSULIN	d00262	2	1460	E11			Type 2 diabetes mellitus			
73558	1	GABAPEN	d03182	1	243	G25.81			Restless legs syndrome			
73558	1	INSULIN G	d04538	1	365	E11			Type 2 diabetes mellitus			
73558	1	OLMESAR	d04801	1	14	E11.2			Type 2 diabetes mellitus with k			
73558	1	SIMVASTA	d00746	1	61	E78.0			Pure hypercholesterolemia			
73559	1	INSULIN A	d04697	1	365	E11			Type 2 diabetes mellitus			
73559	1	INSULIN G	d04538	1	4380	E11			Type 2 diabetes mellitus			
73559	1	PANCRELI	d01002	1	365	K86.9			Disease of pancreas, unspecifie			
73559	1	SIMVASTA	d00746	1	2920	E78.0			Pure hypercholesterolemia			
73559	1	VALSARTA	d04113	1	3650	I10			Essential (primary) hypertensic			
73560	2											
73561	1	55555		1	152	55555						
73561	1	CARVEDIL	d03847	1	6205	I10			Essential (primary) hypertensic			
73561	1	LEVOTHYR	d00278	1	6205	E03.9			Hypothyroidism, unspecified			
73561	1	VALSARTA	d04113	1	4380	I10			Essential (primary) hypertensic			
73562	1	AMLODIPI	d00689	2	365	I10			Essential (primary) hypertensic			
73562	1	ATORVAS	d04105	2	1825	E78.0			Pure hypercholesterolemia			
73562	1	LISINOPRI	d00732	2	3650	I10			Essential (primary) hypertensic			
73562	1	MINOXIDI	d00135	2	365	I10			Essential (primary) hypertensic			
73562	1	NAPROXE	d00019	2	730	55555						
73562	1	PAROXETI	d03157	2	5475	G47.0			Insomnia			
73562	1	PRASUGRI	d07409	2	365	Z95.5			Presence of coronary angioplas			
73563	2											
73564	1	CITALOPR	d04332	1	1825	F32.9			Major depressive disorder, sin			
73564	1	LORAZEPA	d00149	1	5475	F41.9			Anxiety disorder, unspecified			
73564	1	ZOLPIDEM	d00910	1	1825	G47.9			Sleep disorder, unspecified			
73565	1	BUPROPIC	d00181	2	5	F17.203			Nicotine dependence unspecif			
73566	1	ACETAMIN	d03428	1	2555	M79.1			Myalgia			
73566	1	BACLOFEN	d00967	1	2555	M62.83			Muscle spasm			
73566	1	GABAPEN	d03182	1	2555	M62.83			Muscle spasm			
73567	2											
73568	2											
73569	2											
73570	2											
73571	1	ESOMEPR	d04749	2	3650	K21			Gastro-esophageal reflux disea			