INTERIM REPORT

DATA301 Project

Exploratory Data Analysis:

Predicting the Risk of Hypertension Using NHANES Data

Group 16

Cheryl Chan: 300438459

Gale Bueno: 300533945

Nguyen Van: 300528860

September 6, 2022

The aim of this interim report is to present and discuss the progress of the DATA301 project "Exploratory Data Analysis: Predicting Hypertension Using NHANES data". The report will include the main objectives of the project and the exploratory analysis findings.

Background and Data

High blood pressure is a condition that affects over 103.3 million people in the US and 1.3 billion people globally (Ye et al., 2020). It can lead to various health conditions such as heart failure and stroke which is one of the most common leading causes of death in the United States. An estimated 17 million cardiovascular deaths globally are caused annually by factors such as hypertension (López-Martínez et al., 2020). This sign indicates the necessity of having a thorough grasp of the numerous risk factors that might affect blood pressure since lifestyle decisions can impact both the prevalence and development of hypertension (Wang et al., 2015). Identifying the risk of hypertension early on can lead to early treatment and a lower likelihood that a patient will succumb to further ailments.

The utilization of machine learning to build models for disease categorisation is rapidly advancing. The pace at which data is processed and evaluated is accelerated by machine learning. With very slight deployment adjustments, predictive analytics algorithms can now train on even broader data sets and do more in-depth research on a variety of aspects (Elshawi et al., 2019). Several artificial neural network models have been developed to predict the risk of hypertension using a variety of data sources. Nevertheless, most of these models were developed based on smaller samples of data. Model accuracy can still be improved by representing a larger dataset.

The proposed objective of this project is to build a binary classification model which predicts whether an individual has hypertension or not using the National Health and Nutrition Examination Survey (NHANES) data. The NHANES is designed to evaluate the health and nutritional status of adults and children in the United States using interviews and physical examinations. Data from 2007- 2008 to 2017-2020 (pre-pandemic) will be used.

NHANES DATA

Table 1:

A list of risk factors of hypertension across multiple studies.

Author	Risk Factors of Hypertension
Marques et al. (2020)	Waist circumference, alcohol consumption, skin colour, smoking
Rodrigues et al. (2019)	Family history of hypertension, exercise, gender, weight, cholesterol/ lipid levels
Huang et al. (2019)	Age, resting heart rate, weight (overweight, obese etc.), lipid levels, uric acid levels, blood glucose levels/ diabetes, kidney function/kidney disease
López-Martínez et al. (2020)	Race, BMI, smoking status, gender, age, kidney disease, diabetes

Survey, interviews, and physical examinations with risk factors of hypertension were manually downloaded from the NHANES website from 2007-2020. Python was used to import and combine the data.

The resulting data frame is a combination of the listed risk factors in Table 1 and had 17847 instances and 186 features. One of the features ('BPXPLS') which was the 60 second pulse (30 sec. pulse * 2) was not available in the 2017-2020 dataset but was available in the 2007-2016 datasets. Instead, the 2017-2020 dataset had three separate readings for pulse.

Hence, a new column ('BPXPLS') was generated for the 2017-2020 dataset where an average value of the three-pulse measurement was calculated.

For each risk factor, an 'inner join' was carried out across all years and repeated features were filtered and selected. The variables were then concatenated into a bigger data frame, merged by each participant's unique identifier ('SEQN'). The final data frame which contained all risk factors consisted of both numerical and categorical variables as features while the target variable is a binary categorical variable.

We calculated that around 80 features contained 50% or more non-null values. This indicated that 106 features had more than 50% null values. The missing values were a combination of values that were missing at random, missing completely at random and missing not at random. The missing values were coded accordingly in the dataset which allowed for ease of interpretation and analysis.

A possible error that may have presented in the process of combining the datasets across years were the difference in measurements. Data for certain risk factors such as blood pressure measurements, alcohol consumption, income, and triglyceride measurements for the year 2017-2020 used different measurements compared to previous years. As a result, the data columns for the year 2017-2020 are different from those listed for year 2007-2016. This resulted in some features that may have been relevant being excluded through the 'inner join' used to combine datasets across years. There were no other errors observed in the dataset.

Ethics, Privacy, and Security

The aim of this project is to identify risk factors that would increase the risk of hypertension in an individual. The NHANES data which was used for this project was obtained through informed consent where participants were assured that the data collected would only be used for stated purposes (CDC, 2015). The participants were assured that the data would not release to others without the consent of the individual or the establishment (CDC, 2015).

The data collection and analysis have been carried out with beneficence, non-maleficence, justice, and respect for people in mind. Risk factors that are associated with hypertension were identified through researching relevant journal articles. All data in the NHANES dataset related to these risk factors were included in the initial data gathering phase. By including all risk factors, this prevented selection bias. The outcome of the project would be beneficial to health professionals and in turn, patients. Health professionals would be able to utilize the information to intervene in the health management of patients at risk of hypertension as early as possible. The project has put in place steps to prevent harm to participants of the NHANES dataset such as reducing bias where possible. All participants with identified risk factors were included, regardless of their other characteristics. This ensured justice and respect for each participant as data from each participant was treated equal.

This NHANES data used in this project does not contain any identifiable information. Each participant is assigned a unique identifier number which allows the identity of each participant to be anonymous. This ensures the privacy of each participant. The project results would help identify risk factors associated with hypertension which health professionals would have access to through a health-related media such as a medical journal. They would not be able to identify the participants to whom the risk factors relate to. They would also not have access to other information related to each participant.

Data collection, analysis and storage were carried out on each group member's personal devices. These personal devices are only accessible by the individual group member themselves to ensure security of data. GitHub was the chosen platform used to share coding where a private repository had been made, only accessible by the group members.

Steps to keep project data and results secure include encrypting data, backing up data regularly, not leaving personal devices unattended and making sure WI-FI is secure. Encrypting data such as by using full-disk encryption prevents unauthorized access to the project. Backing up data regularly to a secure device or cloud storage ensures data recovery is available if needed. Personal devices should not be left unattended or otherwise be locked with a secure password. WI-FI used to access the data should always be secure and private. A public WI-FI should not be used. Not leaving personal devices unattended and using a secure WI-FI further prevents unauthorized access to the project data and results and unauthorized disclosure of information.

Data Pre-processing:

There are 17847 instances (rows), and 186 features (columns) in the combined data, and many of these features contain mainly NaN values (some features contain only NaN values), so it would be difficult and time-consuming task to identify information relevant

First, the combined dataset was filtered to contain only columns with more than half of their value being non-NaN. This reduced the number of features from 186 to 80 (listed below).

Table 2: NaN counts in the dataset

	NaN count
RIDEXMON	0
DIQ010	0
PAQ620	0
PAQ665	0
RIAGENDR	0
...	...
SMD630	17847
BMXRECUM	17847
BMXHEAD	17847
BPXCHR	17847
BMIHEAD	17847

```
Index(['RIDEXMON', 'INDFMPIR', 'SIAINTRP', 'RIDAGEYR', 'RIDRETH1', 'FIAPROXY',
       'SDMVPSU', 'RIDSTATR', 'FIAINTRP', 'MIAPROXY', 'DMDEDUC2', 'FIALANG',
       'SDMVSTRA', 'SIALANG', 'SEQN', 'SIAPROXY', 'MIALANG', 'SDDSRVYR',
       'RIAGENDR', 'MIAINTRP', 'DSD010', 'DSDCOUNT', 'DSDANCNT', 'DSD010AN',
       'ALQ130', 'LBDTCSI', 'LBXTC', 'DIQ180', 'DIQ010', 'DIQ050', 'DIQ160',
       'INDFMMPC', 'INDFMMPI', 'KIQ480', 'KIQ044', 'KIQ005', 'KIQ022',
       'KIQ046', 'KIQ042', 'KIQ026', 'PAQ620', 'PAQ665', 'PAQ635', 'PAQ650',
       'PAD680', 'PAQ605', 'SMQ020', 'SMAQUEX2', 'LBDTRSI', 'LBXTR',
       'LBDLDLSI', 'LBDLDL', 'BMXARMC', 'BMXWT', 'BMXLEG', 'BMXARML', 'BMXHT',
       'BMXWAIST', 'BMDSTATS', 'BMXBMI', 'BPXDI3', 'BPXDI1', 'BPXDI2',
       'BPXSY1', 'BPXSY2', 'BPXPULS', 'BPACSZ', 'BPXPLS', 'BPAEN1', 'BPXSY3',
       'BPAEN3', 'BPAARM', 'BPXML1', 'BPAEN2', 'BPXPTY', 'BPQ070', 'BPQ080',
       'BPQ020', 'BPQ090D', 'BPQ060'],
      dtype='object')
```

Many categorical features from these 80 features contain values such as 7 (refuse to answer a question) and 9 (don't know the answer to the question). Since these values do not provide useful information and the number of rows containing these values was not a lot, it was decided to further filter the dataset by excluding instances that contain 7 or 9 in any of its columns.

Subsequently, 80 features were separated into two categories: numeric and categorical. Feature selection was carried out based on the correlation of that feature to the response feature.

```
numeric_features = final_df_2[['INDFMPIR','DSDCOUNT','DSDANCNT','ALQ130',
                               'LBDTCSI', 'LBXTC', 'KIQ480', 'PAD680', 'LBXTR',
                               'LBDTRSI', 'LBDLDL', 'LBDLDLSI', 'BPXDI2', 'BPXDI3',
                               'BPXML1', 'BPXSY2', 'BPXSY1', 'BPXDI1', 'BPXPLS',
                               'BPXSY3', 'SDMVSTRA','RIDAGEYR','SDMVPSU','INDFMMPI',
                               'BMXWAIST','BMXHT','BMXLEG','BMXBMI','BMXWT','BMXARML',
                               'BMXARMC', 'BMDSTATS','BPACSZ']]
```

```
categorical_features = ['DMDEDUC2','SIAINTRP','FIAINTRP','MIAPROXY','SIAPROXY','FIALANG',
                        'RIDSTATR','MIALANG','RIAGENDR','RIDRETH1','FIAPROXY','MIAINTRP',
                        'RIDEXMON','SIALANG','DSD010AN','DSDANCNT','DSD010','DIQ010','DIQ18
                        'DIQ160','DIQ050','INDFMMPC','KIQ022','KIQ044','KIQ480','KIQ005',
                        'KIQ026','KIQ042','KIQ046','PAQ620','PAQ605','PAD680','PAQ650',
                        'PAQ665','PAQ635','SMQ020','BPAARM','BPXPULS','BPAEN3', 'SDDSRVYR',
                        'SMAQUEX2','BPAEN1','BPAEN2','BPXPTY','BPQ070','BPQ080','BPQ090D',
                        'BPQ060']
```

Feature selection on categorical variable:

Many of these variables still contain missing values. Rather than removing them, each were replaced with the most frequent value of that feature (mode). This was done using the Simple Imputer class with value = 'most frequent' in python.

Since the response feature is categorical, the Chi-square test (chi2_contingency from scipy.stats) is used to test the correlation between predictors and the response feature. The test involves:

- Assumption H0: the features are not related to each other
- Assumption H1: the features are related to each other
- P-value returned by the function. If the p-value is high, we don't reject H0. If the p-value is below 0.05, then we accept H1

(** please have a look at the python file for more detail **)

As a result, six categorical features were identified to be correlated with the response feature. These features are displayed in Table 3.

Table 3:

Correlation values between categorical features and response features.

Label	Meaning	value
DIQ010	Doctor told you have diabetes?	1: yes 2: no 3: borderline
KIQ022	Ever told you had weak/failing kidneys?	1: yes 2: no
RIDRETH1	Race ethnicity	1: Mexican American 2: Other Hispanic 3: Non-Hispanic White 4: Non-Hispanic Black 5: Other Race - Including Multi-Racial
RIDEXMON	Six-month time period when the examination was performed	1: November 1 through April 30 2: May 1 through October 31
KIQ046	Leak urine during nonphysical activities	1: yes 2: no
PAQ605	Vigorous work activity	1: yes 2: no

Feature selection on numeric variables:

Similar to the categorical features, a lot of numeric features also contain missing values. The Simple Imputer, a scikit-learn class, was used to replace all missing data with the feature's median value.

First, the correlation between the numerical features was tested to identify whether any features are highly correlated to each other as shown in Table 4.

Table 4:

Correlation values between numerical features.

	INDFMPIR	DSDCOUNT	DSDANCNT	ALQ130	LBDTCSI	LBXTC	KIQ480	PAD680	LBXTR	LBDTRSI	..
INDFMPIR	1.000000	0.183074	0.046304	-0.183355	0.037163	0.037144	-0.144722	0.177487	-0.041155	-0.041155	..
DSDCOUNT	0.183074	1.000000	0.064037	-0.146776	0.049750	0.049765	0.039628	0.051016	-0.017022	-0.017023	..
DSDANCNT	0.046304	0.064037	1.000000	-0.009756	0.023267	0.023282	0.029284	0.043490	0.044469	0.044466	..
ALQ130	-0.183355	-0.146776	-0.009756	1.000000	0.005007	0.004984	-0.018793	-0.064537	0.089349	0.089351	..
LBDTCSI	0.037163	0.049750	0.023267	0.005007	1.000000	0.999996	-0.000106	-0.037005	0.336457	0.336455	..
LBXTC	0.037144	0.049765	0.023282	0.004984	0.999996	1.000000	-0.000099	-0.036961	0.336480	0.336478	..
KIQ480	-0.144722	0.039628	0.029284	-0.018793	-0.000106	-0.000099	1.000000	-0.016616	0.027609	0.027610	..
PAD680	0.177487	0.051016	0.043490	-0.064537	-0.037005	-0.036961	-0.016616	1.000000	0.012664	0.012662	..
LBXTR	-0.041155	-0.017022	0.044469	0.089349	0.336457	0.336480	0.027609	0.012664	1.000000	1.000000	

(** please have a look at the python file for more detail **)

Based on the Table 4, the results can infer:

- LBDTCSI has high correlation with LBXTC, LBDLDL, LBDLDLSI
- LBXTC has high correlation with LBDTCSI, LBDLDL, LBDLDLSI
- LBXTR has high correlation with LBDTRSI
- BPXDI2 has high correlation with BPXDI1, BPXDI3
- BPXML1 has high correlation with BPXSY1, BPXSY2, BPXSY3
- BPXSY2 has high correlation with BPXSY1, BPXSY3
- INDFMMPI has high correlation with INDFMPIR
- BMXWAIST has high correlation with BMXBMI, BMXWT, BMXARMC
- BMXHT has high correlation with BMXARML and BMXHT
- BMXBMI has high correlation with BMXWT and BPACSZ
- BMXWT has high correlation with BMXBMI and BMXARMC and BPACSZ

Therefore, thorough examination needs to be considered during selecting features for the model, as including features that are highly correlated to each other will not provide much information but make the model more complex.

After that, the point biserial method was used to calculate the correlation between all the numeric features to the response feature.

(** please have a look at the python file for more detail **)

Interestingly, all numeric features were either weak or did not correlate with the response feature as result of the point biserial test. Using a threshold of 0.20, eight numeric features were chosen (see Table 5).

Table 5:

List of numerical features correlation values with the response feature.

Label	Correlation to response feature	Meaning
KIQ480	-0.22785015820501353	How many times urinate at night?
BPXML1	-0.24578112692010595	MIL: maximum inflation levels (mm Hg)
BPXSY2	-0.3122494340693772	Systolic: Blood pres (2nd rdg)
BPXSY1	-0.31543593684141435	Systolic: Blood pres (1st rdg) mm Hg
BPXSY3	-0.30472432274996647	Systolic: Blood pres (3rd rdg) mm Hg
RIDAGEYR	-0.4328772210439453	Age in years, at the time of the screening interview
BMXWAIST	-0.2734776527552688	Waist Circumference (cm)
BMXBMI	-0.21860125281035575	Body Mass Index (kg/m**2)

It is identified previously that BPXSY1, BPXSY2, BPXSY3 are highly correlated, thus, only BPXSY1 will be included as it has the highest correlation out of the 3. In addition, BMXWAIST and BMXBMI are also highly correlated. Since BMI is measured using a combination of weight and height, it would provide more information than BMXWAIST. Therefore, BMXWAIST will be removed from the dataset.

Exploratory Data Analysis

Table 6:

Statistic summary of numerical features

	KIQ480	BPXML1	BPXSY1	RIDAGEYR	BMXWAIST	BMXBMI
count	15642.00	15642.00	15642.00	15642.00	15642.00	15642.00
mean	1.23	150.16	123.38	49.93	99.46	29.28
std	1.12	21.53	15.82	17.43	16.08	7.08
min	0.00	0.00	66.00	20.00	59.10	13.60
25%	0.00	140.00	116.00	35.00	88.50	24.40
50%	1.00	150.00	122.00	50.00	98.20	28.10
75%	2.00	150.00	128.00	64.00	108.40	32.70
max	5.00	888.00	238.00	80.00	178.00	92.30

Table 7:

Statistics summary of categorical features.

min	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
25%	2.000000	2.000000	2.000000	1.000000	2.000000	2.000000
50%	2.000000	2.000000	3.000000	2.000000	2.000000	2.000000
75%	2.000000	2.000000	4.000000	2.000000	2.000000	2.000000
max	3.000000	9.000000	5.000000	2.000000	2.000000	2.000000

Table 8:

Statistics summary of response feature

	BPQ020	count
0	2.0	9749
1	1.0	5893

Basic summary statistic of numeric features:

All numeric features have significant differences in ranges, therefore if a tree-based model is not used, it would be good to scale the data (see Table 6).

Basic summary statistic on categorical features:

Majority of values (from 25% to 75%) of features DIQ010, KIQ022, KIQ046, and PAQ065 are 2 (2 = no) (see Table 7).

Response feature value counts:

The number of responses being 2 is almost double that of 1. This indicates that accuracy should not be used as the performance metric due to class imbalance (see Table 8).

Figure 1:

Histogram of numerical features.

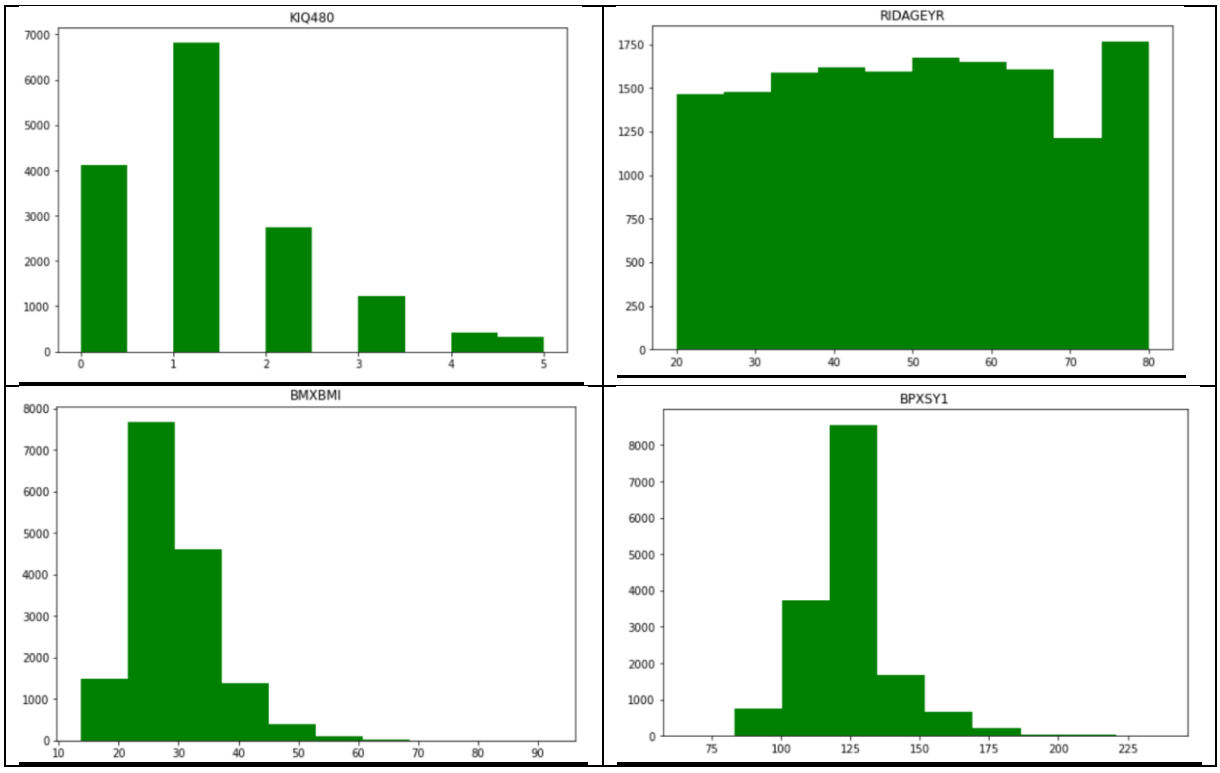


Figure 2:

Scatterplot of numerical features.

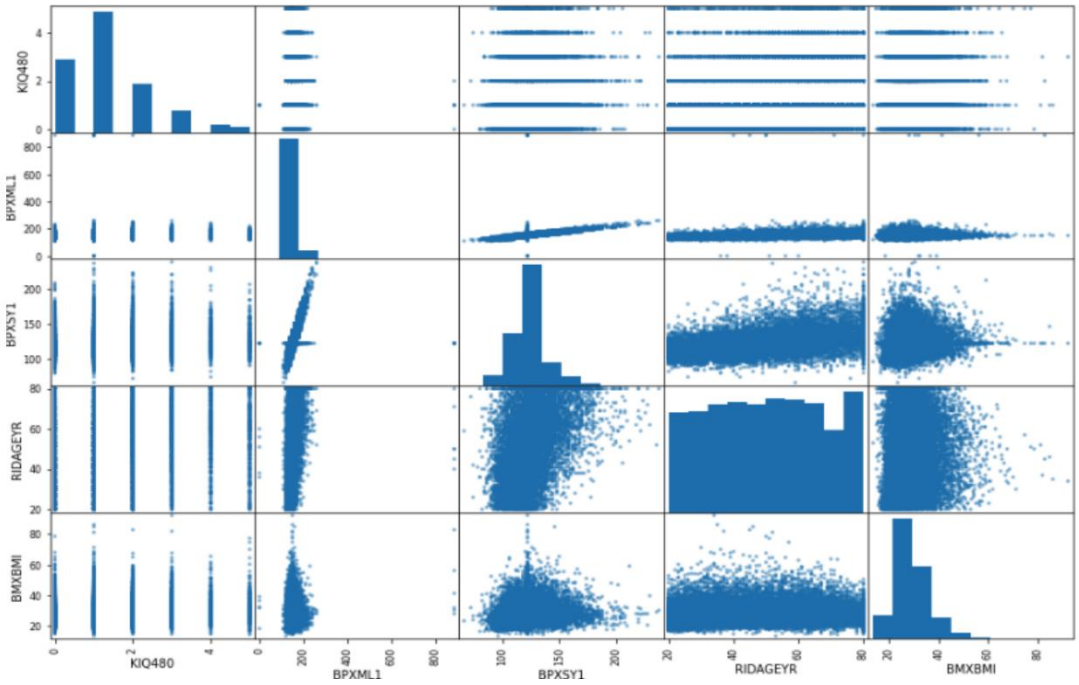
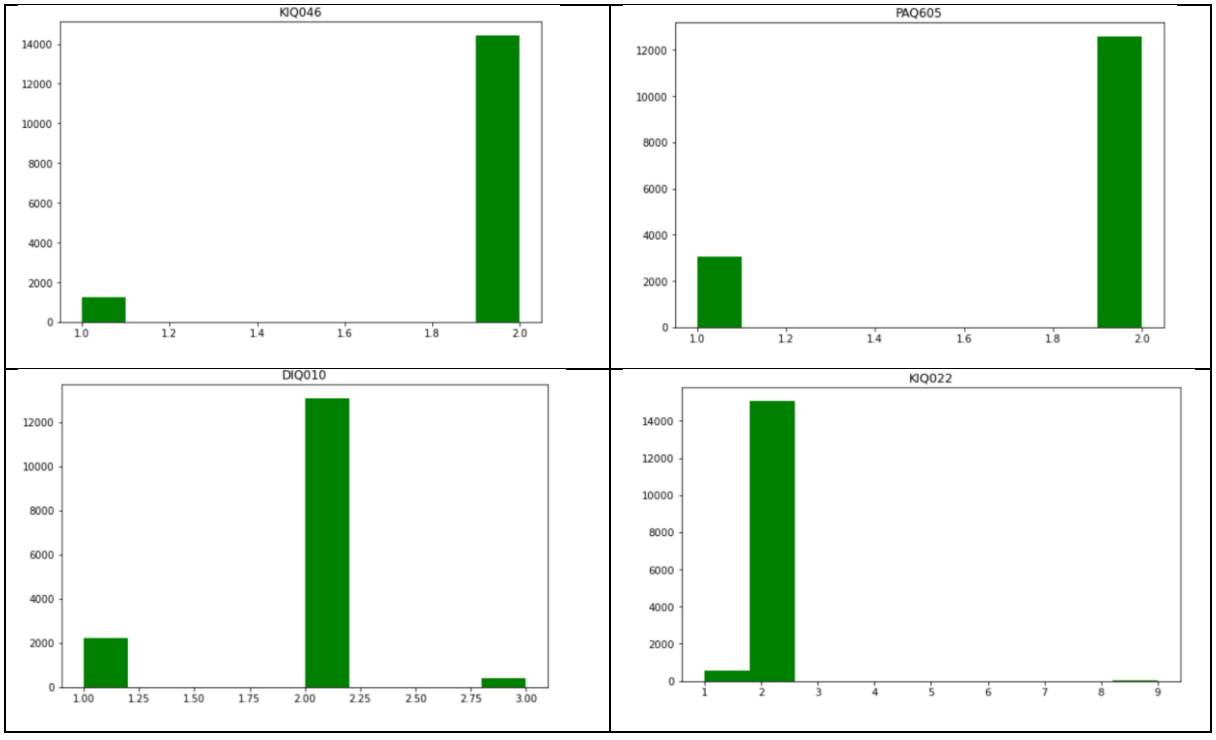
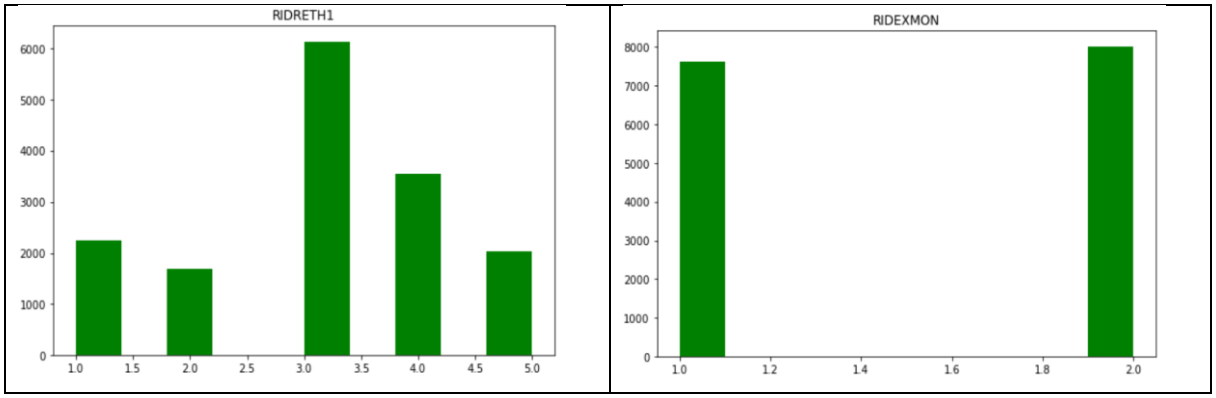


Figure 3:

Histogram of categorical features





Histograms/scatter plot on numeric features as shown in Figure1:

- KIQ480: How many times urinate at night?

The distribution of the feature peaks at 1 and is slightly skewed to the left

- RIDAGEYR: Age in years, at the time of the screening interview

All columns in the graph are relatively equal in height, indicating that ages in the dataset are evenly distributed and therefore, can limit any bias caused by age.

- BMXBMI: Body Mass Index (kg/m**2) & BPXSY1: Systolic: Blood pressure (1st rdg) mm Hg.

The distribution of is very similar. Both are slightly right skewed and have a very high peak (leptokurtic).

- BPXML1: maximum inflation levels (mm Hg) & BPXSY1:

As shown in Figure 2, it seems that BPXML1 and BPXSY1 are highly positively correlated and the removal of one of them should be considered.

Histograms on categorical features as shown in Figure 2:

- All features KIQ046, PAQ605, DIQ010, and KIQ022 have two as their major value. This indicates that these features are imbalanced and could cause unintentional bias when building a model.

- RIDEXMON: Six-month period when the examination was performed.

Compared to the four features mentioned previously, the RIDEXMON feature is very balanced as the two columns in the graphs are roughly equal in height.

- RIDRETH1: Race ethnicity

The distribution of this feature is roughly normal with “Non-Hispanic White” (3) as the highest occurrence value and the lowest occurrence value being “Other Hispanic”.

Individual Contributions

Cheryl:

- Researching risk factors for hypertension
- Combining data for all relevant risk factors from year 2007- 2020
- Writing up part of data collection process
- Writing up ethics, privacy, and security considerations

Gale:

- Used ‘nhanesA’ package in RStudio to extract NHANES data from the website without individually downloading files. Was not used as data can only be obtained up to 2018.
- Initial version of the EDA process:
- Analysis of variables, correlation between variables.
- Calculation and handling missing values.
- Background and Data section of the report (before data collection)
- Representative of the group – communicator between project advisor and team members. Main role was to set up meetings.

Nguyen:

- Created the function used to combine datasets in python and combined demographic data from 2007-2020.
- Carried out feature selections using Pearson correlation, biserial correlation and chi-square test.
- Based on Gale’s initial version of EDA process, further analysed all visualisations to get insight to the data.
- Wrote up the Exploratory Data Analysis section of the report.

The report was edited and revised by everyone in the team.

Below is a link to our GitHub repository:

https://github.com/bgabr/predicting-risk-of-hypertension-using-NHANES-data

Reference

Elshawi, R., Al-Mallah, M. H., & Sakr, S. (2019). On the interpretability of machine learning-based model for predicting hypertension. *BMC medical informatics and decision making, 19*(1), 1-32.

Huang, Y., Deng, Z., Se, Z., Bai, Y., Yan, C., Zhan, Q., Zeng, Q., Ouyang, P., Dai, M., & Xu, D. (2019). Combined impact of risk factors on the subsequent development of hypertension. *Journal of Hypertension, 37*(4), 696-701.

López-Martínez, F., Núñez-Valdez, E. R., Crespo, R. G., & García-Díaz, V. (2020). An artificial neural network approach for predicting hypertension using NHANES data. *Scientific Reports, 10*(1), 1-14.

Marques, A. P., Szwarcwald, C. L., Pires, D. C., Rodrigues, J. M., Almeida, W. d. S. d., & Romero, D. (2020). Factors associated with arterial hypertension: a systematic review. *Ciência & Saúde Coletiva, 25*, 2271-2282.

NHANES - Informed Consent. Centers for Disease Control and Prevention (CDC). 2015. Retrieved September 6, 2022, from https://www.cdc.gov/nchs/nhanes/genetics/genetic_participants.htm

NHANES - National Health and Nutrition Examination Survey Homepage. (n.d.). Retrieved September 6, 2022, from https://www.cdc.gov/nchs/nhanes/index.htm

Rodrigues, F., Coelho, P., & Mateus, S. (2019). Risk factors and arterial hypertension. *Journal of Human Sport and Exercise*, S1772-S1775.

Wang, A., An, N., Chen, G., Li, L., & Alterovitz, G. (2015). Predicting hypertension without measurement: A non-invasive, questionnaire-based approach. *Expert Systems with Applications, 42*(21), 7601-7609.

Ye, X., Zeng, Q. T., Facelli, J. C., Brixner, D. I., Conway, M., & Bray, B. E. (2020). Predicting optimal hypertension treatment pathways using recurrent neural networks. *International Journal of Medical Informatics, 139*, 104122.