Task Report

Shujun JIANG

July 19, 2024

1 Task 1

	Coronary Heart Disease		Angina		Heart Attack		Stroke	
	Diagnosed	Undiagnosed	Diagnosed	Undiagnosed	Diagnosed	Undiagnosed	Diagnosed	Undiagnosed
AGE								
18-40	24 (0.33%)	7,290 (99.67%)	17 (0.23%)	7,297 (99.77%)	21 (0.29%)	7,293 (99.71%)	35 (0.48%)	7,279 (99.52%)
41-60	211 (2.86%)	7,174 (97.14%)	85 (1.15%)	7,300 (98.85%)	158 (2.14%)	7,227 (97.86%)	159 (2.15%)	7,226 (97.85%)
61-80	1,003 (11.87%)	7,445 (88.13%)	299 (3.54%)	8,149 (96.46%)	576 (6.82%)	7,872 (93.18%)	519 (6.14%)	7,929 (93.86%)
81+	344 (20.42%)	1,341 (79.58%)	91 (5.40%)	1,594 (94.60%)	156 (9.26%)	1,529 (90.74%)	192 (11.39%)	1,493 (88.61%)
SEX								
Male	932 (8.28%)	10,327 (91.72%)	272 (2.42%)	10,987 (97.58%)	594 (5.28%)	10,665 (94.72%)	392 (3.48%)	10,867 (96.52%)
Female	650 (4.79%)	12,923 (95.21%)	220 (1.62%)	13,353 (98.38%)	317 (2.34%)	13,256 (97.66%)	513 (3.78%)	13,060 (96.22%)
RACE								
White only	1,330 (6.81%)	18,191 (93.19%)	419 (2.15%)	19,102 (97.85%)	772 (3.95%)	18,749 (96.05%)	707 (3.62%)	18,814 (96.38%)
Black/African American only	160 (5.47%)	2,764 (94.53%)	42 (1.44%)	2,882 (98.56%)	87 (2.98%)	2,837 (97.02%)	139 (4.75%)	2,785 (95.25%)
Asian only	54 (3.39%)	1,541 (96.61%)	19 (1.19%)	1,576 (98.81%)	22 (1.38%)	1,573 (98.62%)	28 (1.76%)	1,567 (98.24%)
AIAN only	19 (7.42%)	237 (92.58%)	5 (1.95%)	251 (98.05%)	14 (5.47%)	242 (94.53%)	9 (3.52%)	247 (96.48%)
AIAN and any other group	12 (6.09%)	185 (93.91%)	5 (2.54%)	192 (97.46%)	10 (5.08%)	187 (94.92%)	17 (8.63%)	180 (91.37%)
Other single and multiple races	7 (2.06%)	332 (97.94%)	2 (0.59%)	337 (99.41%)	6 (1.77%)	333 (98.23%)	5 (1.47%)	334 (98.53%)
RESIDENCE								
Owned or being bought	1,175 (6.84%)	15,993 (93.16%)	342 (1.99%)	16,826 (98.01%)	661 (3.85%)	16,507 (96.15%)	603 (3.51%)	16,565 (96.49%)
Rented	362 (5.08%)	6,771 (94.92%)	133 (1.86%)	7,000 (98.14%)	225 (3.15%)	6,908 (96.85%)	268 (3.76%)	6,865 (96.24%)
Other arrangement	45 (8.47%)	486 (91.53%)	17 (3.20%)	514 (96.80%)	25 (4.71%)	506 (95.29%)	34 (6.40%)	497 (93.60%)
REGION								
Northeast	273 (6.71%)	3,797 (93.29%)	71 (1.74%)	3,999 (98.26%)	124 (3.05%)	3,946 (96.95%)	118 (2.90%)	3,952 (97.10%)
Midwest	388 (6.87%)	5,260 (93.13%)	114 (2.02%)	5,534 (97.98%)	231 (4.09%)	5,417 (95.91%)	202 (3.58%)	5,446 (96.42%)
South	636 (6.87%)	8,621 (93.13%)	212 (2.29%)	9,045 (97.71%)	380 (4.11%)	8,877 (95.89%)	399 (4.31%)	8,858 (95.69%)
West	285 (4.87%)	5,572 (95.13%)	95 (1.62%)	5,762 (98.38%)	176 (3.00%)	5,681 (97.00%)	186 (3.18%)	5,671 (96.82%)

Table 1: Task 1

As shown in the table above, in Task1, I selected age, sex, race, residence and household region as demographic variables, and tried to find the possible relationship between them and four kinds of cardiovascular diseases.

The graph below shows some of the relationships between demographic variables.

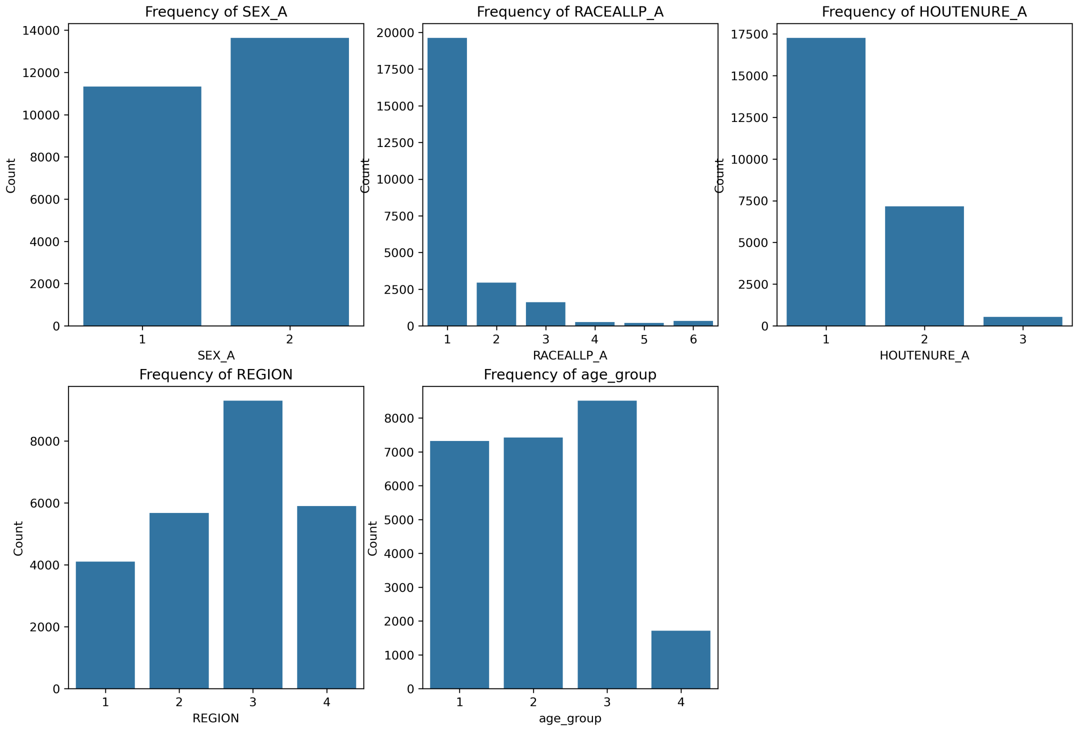


Figure 1: Frequency of Demographic Variables

Figure 2: Age Group Distributions

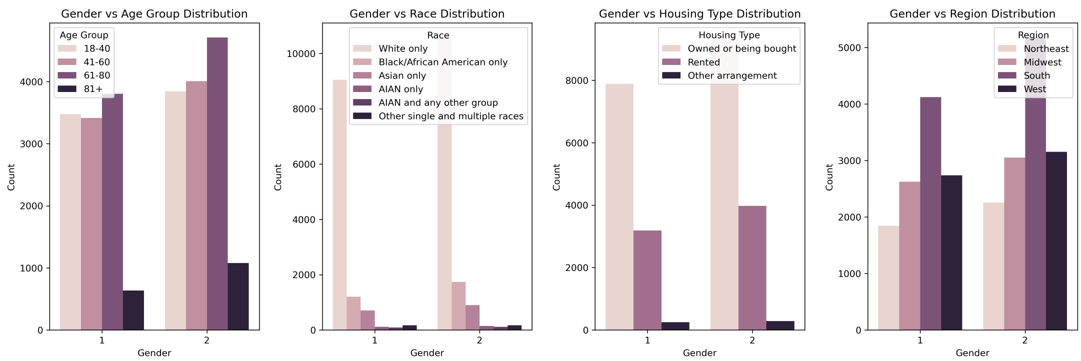


Figure 3: Sex Distributions

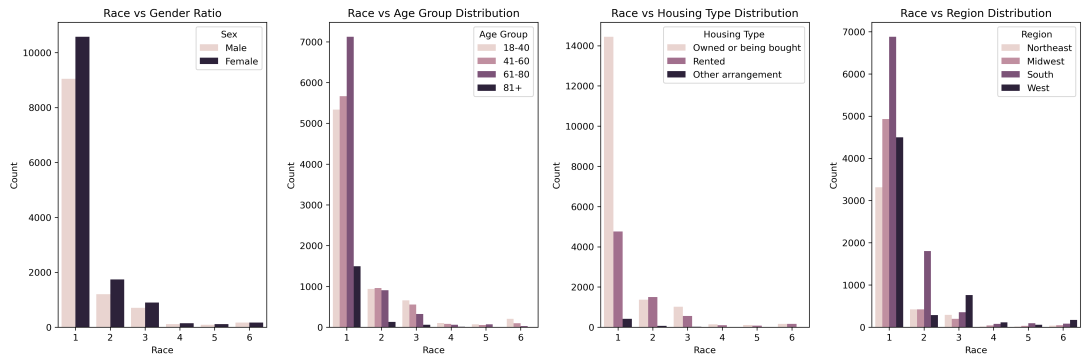


Figure 4: Race Distributions

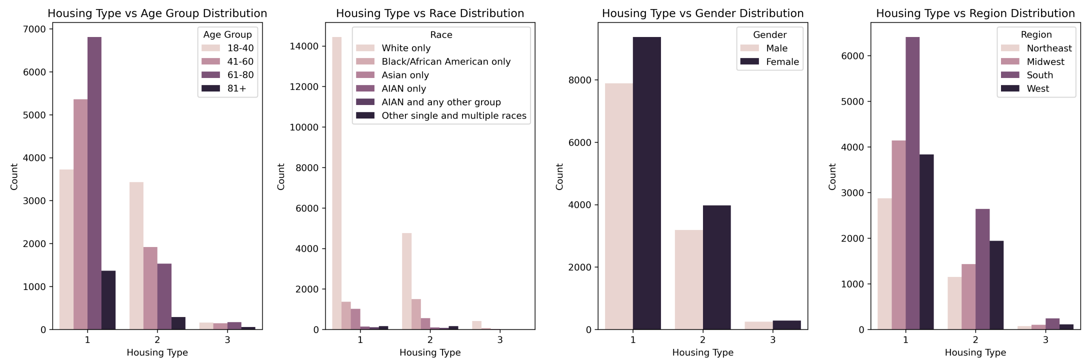


Figure 5: Residential Distributions

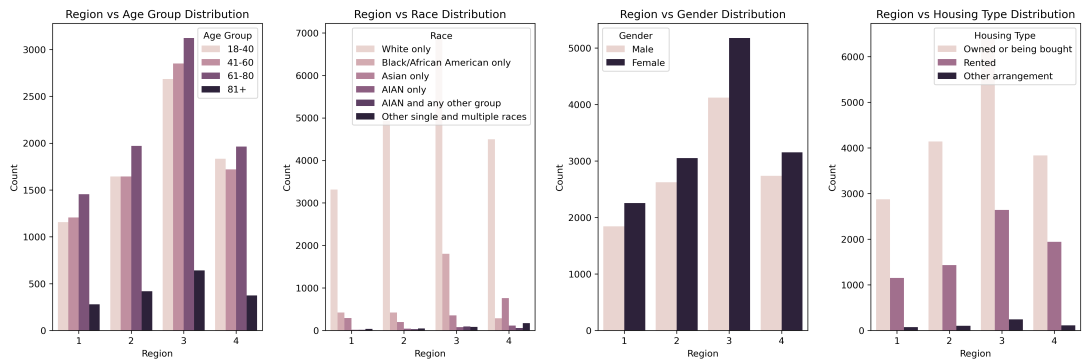


Figure 6: Region Distributions

Through these bar charts, we can identify a few issues that may have implications for our subsequent analysis. There is a huge disparity in the number of people counted by race in our dataset, with "white only" far outnumbering any of the other races, even more than they do combined. In terms of residence, there are also significantly more people living in the "west" than anywhere else. In terms of age groups, divided into groups of nearly 20 years old, it can be seen that the number of people over 80 years old is significantly lower than that of any other group. These strengths and weaknesses are reflected in the data, and will be reflected in the habits of life. And lifestyle habits are just as important as a driver of the emergence of these data. The same holds true for cardiovascular diseases. So we can't ignore statistically weak data groups.

First, racial differences may mean that we have a statistical bias in studying certain health issues. For example, certain diseases may be more prevalent in certain races, so an uneven racial distribution may affect the accuracy of results when conducting epidemiologic studies.

Second, differences in geographic location may also have a significant impact on health outcomes. Populations in different regions may have different lifestyles, diets, and healthcare resources, all of which may affect the incidence and diagnosis of cardiovascular diseases. Thus, having significantly more people living in the "west" than in other areas may bias the analysis of the overall data.

Finally, differences in age subgroups are also noteworthy. Age is an important risk factor for cardiovascular disease, and the low proportion of people over the age of 80 may have resulted in an inadequate sample to analyze the health status of the older age groups, which may affect the reliability of our conclusions.

Our conclusions based on this data will still largely receive the influence of these data biases. In this report, I would mostly compare the data of the survey respondents in the same group, such as how many white people have a probability of having a stroke or how many white men have a probability of having a stroke. However, after adding multiple qualifications, the sample size of available studies was low and their findings were not generalizable.

In the future in order to analyze the data more accurately, we may need more and more detailed data in order to weight these variables. In this way, we can minimize the impact of data bias on the results, leading to more reliable findings. These differences and biases need to be carefully considered in the analysis to ensure that our results are scientific and fair to any group.

2 Task 2

2.1 Life Styles with Demographics

In order to explore the relationship between demographic and life styles, I selected several dietary habits, such as the frequency of drinking pure fruit juice and the frequency of eating vegetables, as research references. The table showing below presents the diet of different groups.

	Number of times drank pure fruit juice				Number of times drank coffee or tea with sugar				Number of times eat salad			
	Never	Daily	Weekly	Monthly	Never	Daily	Weekly	Monthly	Never	Daily	Weekly	Monthly
AGE												
18-40	2,879 (39.77%)	730 (10.08%)	1,799 (24.85%)	1,832 (25.30%)	2,986 (41.24%)	1,842 (25.44%)	1,371 (18.94%)	1,041 (14.38%)	836 (11.55%)	1,114 (15.39%)	3,698 (51.08%)	1,592 (21.99%)
41-60	3,601 (48.92%)	733 (9.96%)	1,463 (19.88%)	1,564 (21.25%)	3,491 (47.43%)	2,148 (29.18%)	1,006 (13.67%)	716 (9.73%)	664 (9.02%)	1,315 (17.86%)	3,822 (51.92%)	1,560 (21.19%)
61-80	4,232 (50.18%)	1,113 (13.20%)	1,570 (18.62%)	1,519 (18.01%)	4,693 (55.64%)	2,276 (26.99%)	848 (10.05%)	617 (7.32%)	832 (9.86%)	1,542 (18.28%)	4,322 (51.24%)	1,738 (20.61%)
81+	700 (41.97%)	407 (24.40%)	321 (19.24%)	240 (14.39%)	981 (58.81%)	471 (28.24%)	135 (8.09%)	81 (4.86%)	255 (15.29%)	352 (21.10%)	771 (46.22%)	290 (17.39%)
SEX												
Male	4,649 (41.49%)	1,543 (13.77%)	2,605 (23.25%)	2,407 (21.48%)	5,810 (51.86%)	2,839 (25.34%)	1,500 (13.39%)	1,055 (9.42%)	1,461 (13.04%)	1,646 (14.69%)	5,633 (50.28%)	2,464 (21.99%)
Female	6,763 (50.10%)	1,440 (10.67%)	2,548 (18.88%)	2,748 (20.36%)	6,341 (46.97%)	3,898 (28.88%)	1,860 (13.78%)	1,400 (10.37%)	1,126 (8.34%)	2,677 (19.83%)	6,980 (51.71%)	2,716 (20.12%)
RACE												
White only	9,451 (48.66%)	2,186 (11.26%)	3,732 (19.22%)	4,053 (20.87%)	9,957 (51.27%)	5,259 (27.08%)	2,406 (12.39%)	1,800 (9.27%)	1,924 (9.91%)	3,285 (16.91%)	10,058 (51.79%)	4,155 (21.39%)
Black/African American only	896 (30.86%)	563 (19.39%)	857 (29.52%)	587 (20.22%)	1,148 (39.55%)	769 (26.49%)	581 (20.01%)	405 (13.95%)	376 (12.95%)	450 (15.50%)	1,456 (50.16%)	621 (21.39%)
Asian only	752 (47.39%)	120 (7.56%)	383 (24.13%)	332 (20.92%)	715 (45.05%)	486 (30.62%)	234 (14.74%)	152 (9.58%)	205 (12.92%)	459 (28.92%)	719 (45.31%)	204 (12.85%)
AIAN only	95 (37.85%)	47 (18.73%)	70 (27.89%)	39 (15.54%)	103 (41.04%)	81 (32.27%)	47 (18.73%)	20 (7.97%)	22 (8.76%)	42 (16.73%)	118 (47.01%)	69 (27.49%)
AIAN and any other group	78 (38.42%)	24 (11.82%)	42 (20.69%)	59 (29.06%)	88 (43.35%)	54 (26.60%)	31 (15.27%)	30 (14.78%)	25 (12.32%)	37 (18.23%)	101 (49.75%)	40 (19.70%)
Other single and multiple races	140 (41.54%)	43 (12.76%)	69 (20.47%)	85 (25.22%)	140 (41.54%)	88 (26.11%)	61 (18.10%)	48 (14.24%)	35 (10.39%)	50 (14.84%)	161 (47.77%)	91 (27.00%)
RESIDENCE												
Owned or being bought	8,231 (48.17%)	1,882 (11.01%)	3,399 (19.89%)	3,577 (20.93%)	8,819 (51.61%)	4,610 (26.98%)	2,123 (12.42%)	1,537 (8.99%)	1,554 (9.09%)	3,036 (17.77%)	8,931 (52.26%)	3,568 (20.88%)
Rented	2,946 (41.57%)	1,013 (14.30%)	1,640 (23.14%)	1,487 (20.99%)	3,071 (43.34%)	1,980 (27.94%)	1,154 (16.29%)	881 (12.43%)	952 (13.43%)	1,204 (16.99%)	3,431 (48.42%)	1,499 (21.15%)
Other arrangement	235 (44.51%)	88 (16.67%)	114 (21.59%)	91 (17.23%)	261 (49.43%)	147 (27.84%)	83 (15.72%)	37 (7.01%)	81 (15.34%)	83 (15.72%)	251 (47.54%)	113 (21.40%)
REGION												
Northeast	1,893 (46.91%)	472 (11.70%)	841 (20.84%)	829 (20.55%)	1,957 (48.50%)	1,170 (29.00%)	542 (13.43%)	366 (9.07%)	397 (9.84%)	785 (19.45%)	2,119 (52.52%)	734 (18.19%)
Midwest	2,670 (47.52%)	707 (12.58%)	1,009 (17.96%)	1,233 (21.94%)	3,193 (56.83%)	1,355 (24.11%)	531 (9.45%)	540 (9.61%)	614 (10.93%)	843 (15.00%)	2,725 (48.50%)	1,437 (25.57%)
South	4,116 (44.67%)	1,195 (12.97%)	2,124 (23.05%)	1,779 (19.31%)	3,976 (43.15%)	2,763 (29.99%)	1,519 (16.49%)	956 (10.38%)	1,079 (11.71%)	1,451 (15.75%)	4,761 (51.67%)	1,923 (20.87%)
West	2,733 (46.84%)	609 (10.44%)	1,179 (20.21%)	1,314 (22.52%)	3,025 (51.84%)	1,449 (24.83%)	768 (13.16%)	593 (10.16%)	497 (8.52%)	1,244 (21.32%)	3,008 (51.55%)	1,086 (18.61%)

	Number of times eat fried potatoes				Number of times eat beans				Number of times eat pizza				Number of times eat other vegetables			
	Never	Daily	Weekly	Monthly	Never	Daily	Weekly	Monthly	Never	Daily	Weekly	Monthly	Never	Daily	Weekly	Monthly
AGE																
18-40	799 (11.04%)	225 (3.11%)	3,543 (48.94%)	2,673 (36.92%)	2,046 (28.26%)	205 (2.83%)	2,345 (32.39%)	2,644 (36.52%)	754 (10.41%)	54 (0.75%)	2,114 (29.20%)	4,318 (59.64%)	306 (4.23%)	2,668 (36.85%)	3,063 (42.31%)	1,203 (16.62%)
41-60	1,447 (19.66%)	174 (2.36%)	2,786 (37.85%)	2,954 (40.13%)	1,823 (24.77%)	209 (2.84%)	2,465 (33.49%)	2,864 (38.91%)	1,172 (15.92%)	55 (0.75%)	1,754 (23.83%)	4,380 (59.50%)	312 (4.24%)	2,762 (37.52%)	3,097 (42.07%)	1,190 (16.17%)
61-80	2,235 (26.50%)	128 (1.52%)	2,657 (31.50%)	3,414 (40.48%)	2,029 (24.06%)	206 (2.44%)	2,686 (31.85%)	3,513 (41.65%)	2,005 (23.77%)	36 (0.43%)	1,281 (15.19%)	5,112 (60.61%)	350 (4.15%)	3,196 (37.89%)	3,422 (40.57%)	1,466 (17.38%)
81+	508 (30.46%)	32 (1.92%)	539 (32.31%)	589 (35.31%)	393 (23.56%)	49 (2.94%)	574 (34.41%)	652 (39.09%)	528 (31.65%)	14 (0.84%)	198 (11.87%)	928 (55.64%)	65 (3.90%)	709 (42.51%)	681 (40.83%)	213 (12.77%)
SEX																
Male	1,881 (16.79%)	310 (2.77%)	4,917 (43.89%)	4,096 (36.56%)	2,652 (23.67%)	310 (2.77%)	3,924 (35.02%)	4,318 (38.54%)	1,766 (15.76%)	85 (0.76%)	2,804 (25.03%)	6,549 (58.45%)	606 (5.41%)	3,616 (32.27%)	4,990 (44.54%)	1,992 (17.78%)
Female	3,108 (23.02%)	249 (1.84%)	4,608 (34.14%)	5,534 (41.00%)	3,639 (26.96%)	359 (2.66%)	4,146 (30.71%)	5,355 (39.67%)	2,693 (19.95%)	74 (0.55%)	2,543 (18.84%)	8,189 (60.66%)	427 (3.16%)	5,719 (42.37%)	5,273 (39.06%)	2,080 (15.41%)
RACE																
White only	3,744 (19.28%)	415 (2.14%)	7,609 (39.18%)	7,654 (39.41%)	4,672 (24.06%)	507 (2.61%)	6,509 (33.51%)	7,734 (39.82%)	3,068 (15.80%)	124 (0.64%)	4,531 (23.33%)	11,699 (60.24%)	795 (4.09%)	7,399 (38.10%)	7,980 (41.09%)	3,248 (16.72%)
Black/African American only	610 (21.01%)	95 (3.27%)	1,128 (38.86%)	1,070 (36.86%)	825 (28.42%)	76 (2.62%)	857 (29.52%)	1,145 (39.44%)	832 (28.66%)	24 (0.83%)	425 (14.64%)	1,622 (55.87%)	140 (4.82%)	898 (30.93%)	1,345 (46.33%)	520 (17.91%)
Asian only	500 (31.51%)	20 (1.26%)	469 (29.55%)	598 (37.68%)	596 (37.56%)	57 (3.59%)	434 (27.35%)	500 (31.51%)	410 (25.83%)	8 (0.50%)	214 (13.48%)	955 (60.18%)	57 (3.59%)	761 (47.95%)	602 (37.93%)	167 (10.52%)
AIAN only	37 (14.74%)	7 (2.79%)	118 (47.01%)	89 (35.46%)	41 (16.33%)	13 (5.18%)	108 (43.03%)	89 (35.46%)	55 (21.91%)	2 (0.80%)	56 (22.31%)	138 (54.98%)	20 (7.97%)	74 (29.48%)	117 (46.61%)	40 (15.94%)
AIAN and any other group	42 (20.69%)	5 (2.46%)	77 (37.93%)	79 (38.92%)	46 (22.66%)	4 (1.97%)	73 (35.96%)	80 (39.41%)	42 (20.69%)		41 (20.20%)	120 (59.11%)	9 (4.43%)	73 (35.96%)	85 (41.87%)	36 (17.73%)
Other single and multiple races	56 (16.62%)	17 (5.04%)	124 (36.80%)	140 (41.54%)	111 (32.94%)	12 (3.56%)	89 (26.41%)	125 (37.09%)	52 (15.43%)	1 (0.30%)	80 (23.74%)	204 (60.53%)	12 (3.56%)	130 (38.58%)	134 (39.76%)	61 (18.10%)
RESIDENCE																
Owned or being bought	3,382 (19.79%)	336 (1.97%)	6,500 (38.04%)	6,871 (40.21%)	4,153 (24.30%)	404 (2.36%)	5,583 (32.67%)	6,949 (40.66%)	2,865 (16.77%)	96 (0.56%)	3,686 (21.57%)	10,442 (61.10%)	599 (3.51%)	6,679 (39.08%)	7,064 (41.34%)	2,747 (16.07%)
Rented	1,481 (20.90%)	198 (2.79%)	2,825 (39.87%)	2,582 (36.44%)	2,008 (28.34%)	239 (3.37%)	2,302 (32.49%)	2,537 (35.80%)	1,473 (20.79%)	57 (0.80%)	1,549 (21.86%)	4,007 (56.55%)	398 (5.62%)	2,451 (34.59%)	2,988 (42.17%)	1,249 (17.63%)
Other arrangement	126 (23.86%)	25 (4.73%)	200 (37.88%)	177 (33.52%)	130 (24.62%)	26 (4.92%)	185 (35.04%)	187 (35.42%)	121 (22.92%)	6 (1.14%)	112 (21.21%)	289 (54.73%)	36 (6.82%)	205 (38.83%)	211 (39.96%)	76 (14.39%)
REGION																
Northeast	958 (23.74%)	59 (1.46%)	1,427 (35.37%)	1,591 (39.43%)	1,339 (33.18%)	90 (2.23%)	1,114 (27.61%)	1,492 (36.98%)	648 (16.06%)	24 (0.59%)	962 (23.84%)	2,401 (59.50%)	135 (3.35%)	1,674 (41.49%)	1,651 (40.92%)	575 (14.25%)
Midwest	950 (16.91%)	131 (2.33%)	2,256 (40.15%)	2,282 (40.61%)	1,472 (26.20%)	116 (2.06%)	1,614 (28.72%)	2,417 (43.01%)	743 (13.22%)	38 (0.68%)	1,474 (26.23%)	3,364 (59.87%)	220 (3.92%)	2,275 (40.49%)	2,034 (36.20%)	1,090 (19.40%)
South	1,817 (19.72%)	258 (2.80%)	3,691 (40.06%)	3,448 (37.42%)	2,142 (23.25%)	296 (3.21%)	3,298 (35.79%)	3,478 (37.75%)	1,861 (20.20%)	76 (0.82%)	1,877 (20.37%)	5,400 (58.61%)	436 (4.73%)	3,114 (33.80%)	4,176 (45.32%)	1,488 (16.15%)
West	1,264 (21.66%)	111 (1.90%)	2,151 (36.86%)	2,309 (39.57%)	1,338 (22.93%)	167 (2.86%)	2,044 (35.03%)	2,286 (39.18%)	1,207 (20.69%)	21 (0.36%)	1,034 (17.72%)	3,573 (61.23%)	242 (4.15%)	2,272 (38.94%)	2,402 (41.17%)	919 (15.75%)

Table 2: Life Styles with Demographics

From the above table, we can see that the proportion of people who drink coffee or tee with sugar is higher than the proportion of people who drink fruit juice at the same age range. The proportion of people who never drink coffee or tee with sugar or who never eat fried potatoes or who never eat pizzad are increasing with age in general, while the proportion of people who eat salads and vegetables daily, which are considered to be healthier, is also increasing.

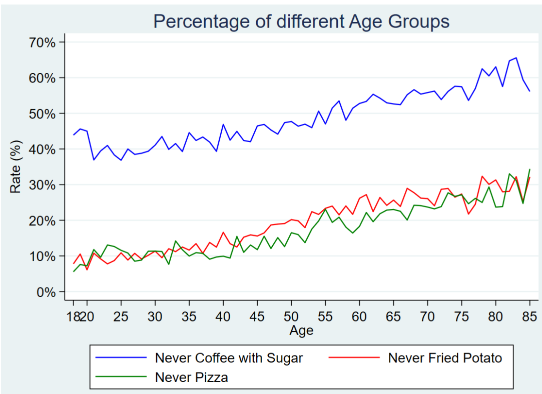


Figure 7: Percentage of different age groups who never drink non-sugar coffee or tee, eat fried potatos or eat pizza

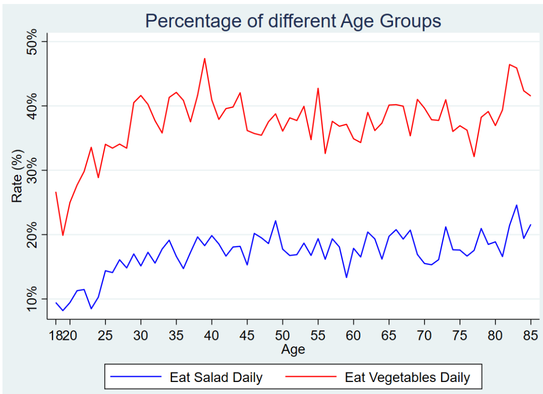


Figure 8: Percentage of different age groups who eat salad daily or eat vegetables daily

The difference in lifestyle habits is also obvious in gender. The proportion of men who never drink pure fruit juice(41.49%) is significantly lower than women(50.10%), while the proportion of women who do not drink coffee or tee with sugar(46.97%) is also slightly lower than men(51.86%). In the case of salads and vegetables, the proportion of women(salad 19.83% vegatables 42.37%) who ate them every day was also higher than men(salad 14.69% vegetables 32.27%). We can see the gender differences in eating habits more directly by looking at the bar charts below.

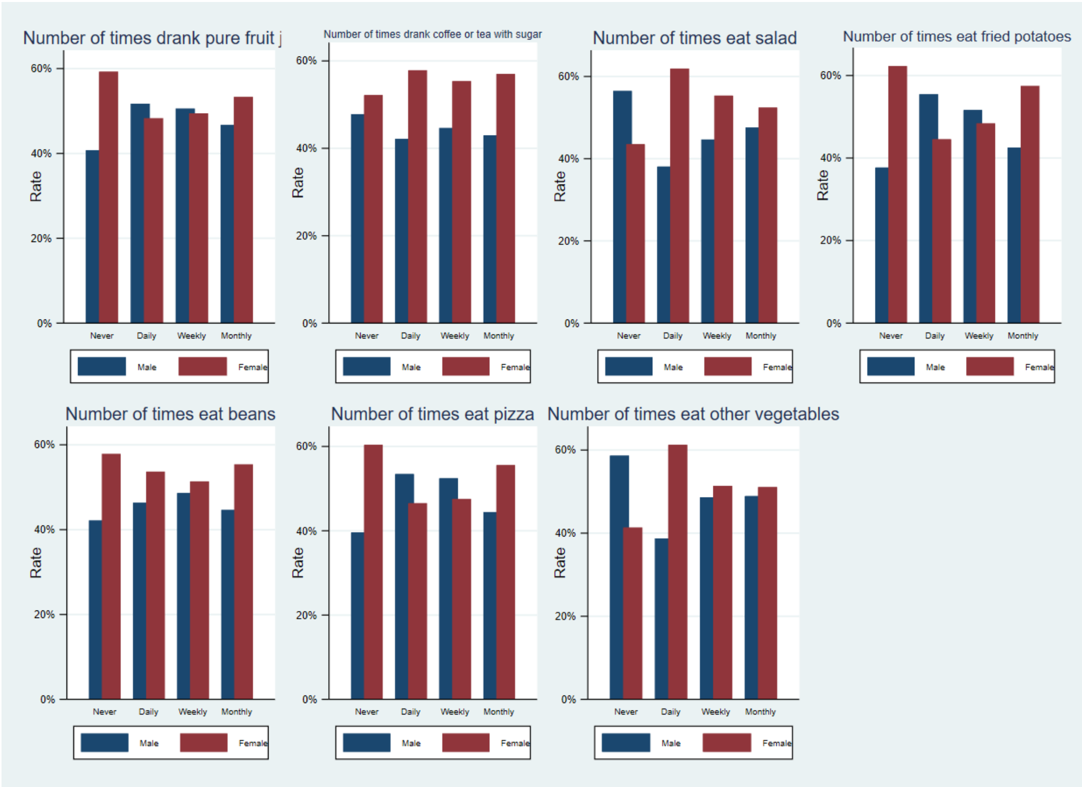


Figure 9: Comparison of Dietary Differences between Men and Women

Similar differences can also be observed when examining race, housing status, and region in Table 2, reflecting how these factors may influence dietary habits among populations.

2.2 Cardiovascular Conditions with Demographics

From Table 1, we could get some rough findings. The percentage of the population who was diagnosed with caediovascular the conditions increases with age. For example, the percentage of coronary heart disease for people over 80 years old is 20.42%, and the percentage for people between 18 and 40 years old is only 0.33%. As shown in the figure below, the prevalence of various cardiovascular diseases has shown a significant increasing trend. This suggests that age is an important risk factor with important implications for cardiovascular health management and prevention strategies.

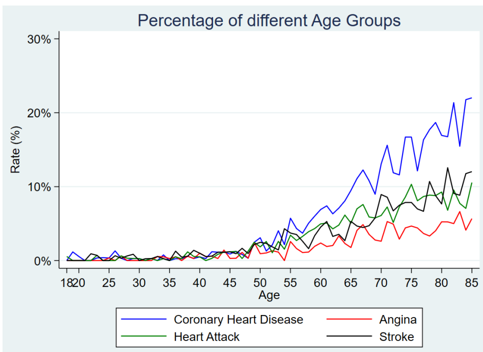


Figure 10: Comparison of Diseases Differences of Different Age

Cardiovascular disease also has some race differencces, with American Indians or Alaska Natives having a higher cardiovascular prevalence than other races. For example, the the diagnosis rate for coronary heart disease is 7.42% for American Indians or Alaska Natives compared to 3.39% for the Asian group. This may reflect genetic factors, cultural practices, or an uneven distribution of health care resources.

In terms of sex, except stroke(female were slightly higher by 0.3%), all the other three cardiovascular diseases have a higher group rate of male than female. This may be related to biological sex differences, lifestyle choices, or other health factors and deserves further exploration.

In addition, we can also use the chi-square test method to determine whether the variables are statistically related. The following two images are Chi-square test results showing a statistical association between gender and coronary heart disease and no significant association between gender and stroke.

```
. tab SEX_A CHDEV_A, chi2

           |   Coronary Heart
           |       Disease
       SEX | Diagnosed  Undiagnos |     Total
-----------+----------------------+----------
      Male |       932     10,327 |    11,259
    Female |       650     12,923 |    13,573
-----------+----------------------+----------
     Total |     1,582     23,250 |    24,832

          Pearson chi2(1) = 125.5844   Pr = 0.000
```

Figure 11: Sex-Coronary Chi-square test

```
. tab SEX_A STREV_A,chi2

           |        Stroke
       SEX | Diagnosed  Undiagnos |     Total
-----------+----------------------+----------
      Male |       392     10,867 |    11,259
    Female |       513     13,060 |    13,573
-----------+----------------------+----------
     Total |       905     23,927 |    24,832

          Pearson chi2(1) =   1.5553   Pr = 0.212
```

Figure 12: Sex-Stroke Chi-square test

Take the second Chi-square test as an example, Pearson chi2(1) = 1.5553: This value is the Pearson Chi-square statistic, which measures the degree of deviation between the observed data and the expected data. The value of the statistic is relatively small, indicating that the observed data does not differ much from the expected data. Pr = 0.212: The size of this p-value is used to evaluate the significance of the statistical results. Here, a P-value of 0.212 is greater than the usual level of significance (e.g. 0.05), indicating that the observed association between sex and stroke is not statistically significant.

In terms of residence, if we go to the statistics of the population diagnosed with coronary heart disease, which type of people account for a large proportion, as shown in the figure below. We will find that the probability of coronary heart disease in homeowners is much higher than that in renters. However, this is due to the fact that there are nearly twice as many homeowners as renters in the statistics, so a large number of people surveyed with coronary heart disease were homeowners. In fact, by observing Table 1, we can find that the diagnosis rate of coronary heart disease among homeowners (6.84%) is only slightly higher than that of renters (5.0%), while the diagnosis rates of the other three diseases are basically the same in the two groups. The extent to which the type of residence had an effect on the presence or absence of a disease was not obvious.

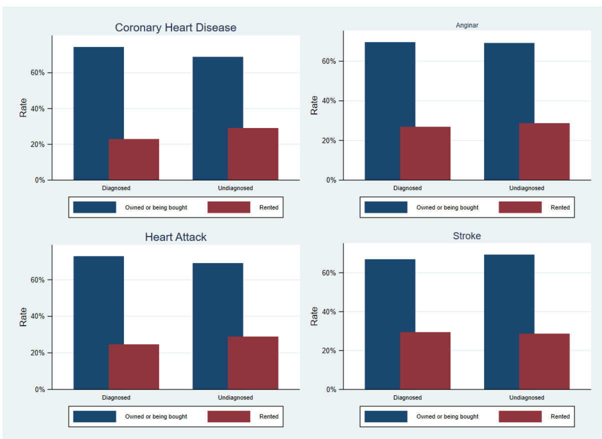


Figure 13: Proportion of Residence Types in the Coronary Heart Disease Population

Regarding region we found that coronary heart disease was significantly less to be diagnosed in those living in the west(4.87%) than those living in other regions. It may be related to the specific environment, dietary habits or socioeconomic factors in the region.

3 Task 3

In order to explore the relationship among lifestyle habits, BMI, and coronary heart disease, I try to visually present the results using tables by combining various conditions. For example, the proportion of people who never eat vegetables, pizza, or fried potatoes who have coronary heart disease or are obese can be found in Tables 3 and 5.

However, the actual value of these probabilities is influenced by the sample size. As we can see from Tables 4 and 6, after applying multiple conditions, the number of people meeting certain criteria is too small, making the probabilities too random to derive any universal patterns. And if we want to refer to more types of data, this method makes the table very long and difficult to read.

When exploring significant correlations between variables, we continue to use the chi-square test as demonstrated in Task 2, looking at the magnitude of the p's value versus 0.05 to determine if the hypothesis is met and to determine if the variables are significant.

Therefore, I am considering using a more standardized statistical method, such as logistic regression analysis, to conduct a more rigorous analysis and reduce the impact of noise in the dataset. It should be noted that the binary output of stata's logistic regression must be 0 or 1, and the diagnosis and undiagnosis of cardiovascular disease in our codebook are 1 and 2, and a 2-to-0 preprocessing is required. And as for the variable of eating frequency we used, 1 is never,2 is daily,3 is weekly and 4monthly. We should reorder it according to the increasing frequency of eating from not eating to eating. In addition, logistic regression is not applicable when studying BMI, because BMI has four outcomes rather than binary variables. So I consider using ordered logistic regression models.

In addition to the above model, I also consider introducing other statistical algorithms and models, to help us more comprehensive understanding of the data, and make predictions. KNN, Decision Tree, Random Forest and CatBoost are commonly used models in machine learning.KNN it's K value represents the number of nearest neighbors selected and KNN decides the category or predicted value of the new data point based on the category of these neighbors.KNN is simple and intuitive but computationally expensive on large data sets and performs poorly on noisy and high dimensional data. Decision tree is a tree-structured model where each node represents a feature, branches represent feature values, and leaf nodes represent categories or regression values. It constructs a tree model by recursively selecting the optimal features to partition the dataset. However, it is prone to overfitting and sensitive to noisy data. Random forest is an integrated learning method that obtains the final result by generating multiple decision trees and voting or averaging their predictions. However, it has a long training and prediction time and high model complexity.CatBoost is an algorithm based on Gradient Boosted Decision Trees (GBDT) that is particularly good at handling categorical features. It reduces the risk of overfitting by converting categorical features to numerical features and using mean coding.CatBoost processes categorical features automatically, reducing the need for manual feature engineering, and performs well with datasets containing a large number of categorical features. However, the model is complex, difficult to interpret, and requires extensive parameter tuning.

	Coronary Heart Disease		Angina		Heart Attack		Stroke	
	Diagnosed	Undiagnosed	Diagnosed	Undiagnosed	Diagnosed	Undiagnosed	Diagnosed	Undiagnosed
Number of times eat other vegetables								
Never								
Number of times eat pizza								
Never								
Number of times eat fried potatoes								
Never	11.42857	88.57143	2.142857	97.85714	6.428571	93.57143	7.142857	92.85714
Daily		100		100	7.142857	92.85714	7.142857	92.85714
Weekly	6.557377	93.44262	3.278689	96.72131	8.196721	91.80328	9.836066	90.16393
Monthly	7.8125	92.1875	4.6875	95.3125	1.5625	98.4375	7.8125	92.1875
Daily								
Number of times eat fried potatoes								
Never		100		100		100		100
Daily		100		100		100		100
Weekly		100		100		100		100
Monthly		100		100		100		100
Weekly								
Number of times eat fried potatoes								
Never	12	88		100	8	92		100
Daily		100	6.666667	93.33333	6.666667	93.33333	13.33333	86.66667
Weekly	8.275862	91.72414	1.37931	98.62069	1.37931	98.62069	2.068966	97.93103
Monthly	5.555556	94.44444	2.777778	97.22222	2.777778	97.22222	11.11111	88.88889
Monthly								
Number of times eat fried potatoes								
Never	7.070707	92.92929	3.030303	96.9697	6.060606	93.93939	3.030303	96.9697
Daily	22.22222	77.77778	5.555556	94.44444	11.11111	88.88889	16.66667	83.33333
Weekly	10.17964	89.82036	1.197605	98.8024	4.790419	95.20958	3.592814	96.40719
Monthly	5.970149	94.02985	1.99005	98.00995	3.482587	96.51741	3.9801	96.0199
Daily								
Number of times eat pizza								
Never								
Number of times eat fried potatoes								
Never	7.182941	92.81706	2.469136	97.53086	4.040404	95.9596	3.928171	96.07183
Daily	18.75	81.25		100		100	9.375	90.625
Weekly	4.968944	95.03106	2.484472	97.51553	3.416149	96.58385	3.416149	96.58385
Monthly	8.073394	91.92661	2.385321	97.61468	5.321101	94.6789	3.119266	96.88073
Daily								
Number of times eat fried potatoes								
Never	15.38462	84.61538		100		100		100
Daily	9.677419	90.32258	3.225806	96.77419	3.225806	96.77419	9.677419	90.32258
Weekly		100		100		100		100
Monthly	18.75	81.25	6.25	93.75		100	6.25	93.75
Weekly								
Number of times eat fried potatoes								
Never	6.477733	93.52227	2.42915	97.57085	3.238866	96.76113	3.238866	96.76113
Daily	2.272727	97.72727	2.272727	97.72727	3.409091	96.59091	2.272727	97.72727
Weekly	4.036697	95.9633	1.651376	98.34862	2.385321	97.61468	2.568807	97.43119
Monthly	4.139434	95.86057	1.960784	98.03922	2.396514	97.60349	3.267974	96.73203
Monthly								
Number of times eat fried potatoes								
Never	6.325301	93.6747	2.108434	97.89157	3.012048	96.98795	4.417671	95.58233
Daily	7.758621	92.24138	4.310345	95.68966	7.758621	92.24138	5.172414	94.82759
Weekly	6.081081	93.91892	1.689189	98.31081	3.65991	96.34009	3.153153	96.84685
Monthly	5.844418	94.15558	1.773478	98.22652	2.982668	97.01733	3.546957	96.45304
Weekly								
Number of times eat pizza								
Never								
Number of times eat fried potatoes								
Never	10.78014	89.21986	3.120567	96.87943	5.390071	94.60993	4.964539	95.03546
Daily	11.42857	88.57143	5.714286	94.28571	8.571429	91.42857	8.571429	91.42857
Weekly	8.350731	91.64927	3.131524	96.86848	4.80167	95.19833	5.845511	94.15449
Monthly	8.684864	91.31514	2.48139	97.51861	4.218362	95.78164	6.451613	93.54839
Daily								
Number of times eat fried potatoes								
Never		100		100		100	11.11111	88.88889
Daily	20	80		100	20	80		100
Weekly	5.882353	94.11765		100		100		100
Monthly	7.142857	92.85714		100	7.142857	92.85714	14.28571	85.71429
Weekly								
Number of times eat fried potatoes								
Never	7.534247	92.46575	2.054795	97.94521	2.739726	97.26027	.6849315	99.31507
Daily	1.449275	98.55072		100		100	2.898551	97.10145
Weekly	4.579393	95.42061	1.692384	98.30762	2.887008	97.11299	2.588352	97.41165
Monthly	5.092593	94.90741	.6944444	99.30556	3.240741	96.75926	2.083333	97.91667
Monthly								
Number of times eat fried potatoes								
Never	8.041958	91.95804	1.864802	98.1352	3.962704	96.0373	4.079254	95.92075
Daily	7.936508	92.06349	4.761905	95.2381	3.174603	96.8254	3.174603	96.8254
Weekly	5.583174	94.41683	1.873805	98.1262	3.632887	96.36711	3.32696	96.67304
Monthly	6.483791	93.51621	1.795511	98.20449	3.541147	96.45885	2.942643	97.05736
Monthly								
Number of times eat pizza								
Never								
Number of times eat fried potatoes								
Never	10.94527	89.05473	2.487562	97.51244	8.955224	91.04478	7.960199	92.0398
Daily	40	60	20	80	20	80	20	80
Weekly	9.375	90.625		100	7.8125	92.1875	4.6875	95.3125
Monthly	11.73184	88.26816	3.351955	96.64804	6.98324	93.01676	6.703911	93.29609
Daily								
Number of times eat fried potatoes								
Never		100		100		100		100
Daily		100		100		100		100
Weekly		100		100	25	75		100
Monthly		100	14.28571	85.71429		100	14.28571	85.71429
Weekly								
Number of times eat fried potatoes								
Never	5.714286	94.28571		100	5.714286	94.28571	5.714286	94.28571
Daily	9.090909	90.90909		100		100		100
Weekly	3.703704	96.2963	1.851852	98.14815	2.469136	97.53086	2.469136	97.53086
Monthly	2.222222	97.77778	2.962963	97.03704	2.962963	97.03704	2.222222	97.77778
Monthly								
Number of times eat fried potatoes								
Never	7.142857	92.85714	3.416149	96.58385	5.590062	94.40994	7.142857	92.85714
Daily	5	95	2.5	97.5	10	90	2.5	97.5
Weekly	7.512953	92.48705	1.554404	98.4456	3.88601	96.11399	2.849741	97.15026
Monthly	5.733333	94.26667	1.688889	98.31111	3.422222	96.57778	3.244444	96.75556

Table 3: The Dietary Habits and Cardiovascular Disease Frequency Chart

	Coronary Heart Disease		Angina		Heart Attack		Stroke	
	Diagnosed	Undiagnosed	Diagnosed	Undiagnosed	Diagnosed	Undiagnosed	Diagnosed	Undiagnosed
Number of times eat other vegetables								
Never								
Number of times eat pizza								
Never								
Number of times eat fried potatoes								
Never	16	124	3	137	9	131	10	130
Daily		14		14	1	13	1	13
Weekly	4	57	2	59	5	56	6	55
Monthly	5	59	3	61	1	63	5	59
Daily								
Number of times eat fried potatoes								
Never		3		3		3		3
Daily		3		3		3		3
Weekly		4		4		4		4
Monthly		3		3		3		3
Weekly								
Number of times eat fried potatoes								
Never	3	22		25	2	23		25
Daily		15	1	14	1	14	2	13
Weekly	12	133	2	143	2	143	3	142
Monthly	2	34	1	35	1	35	4	32
Monthly								
Number of times eat fried potatoes								
Never	7	92	3	96	6	93	3	96
Daily	4	14	1	17	2	16	3	15
Weekly	17	150	2	165	8	159	6	161
Monthly	12	189	4	197	7	194	8	193
Daily								
Number of times eat pizza								
Never								
Number of times eat fried potatoes								
Never	64	827	22	869	36	855	35	856
Daily	6	26		32		32	3	29
Weekly	16	306	8	314	11	311	11	311
Monthly	44	501	13	532	29	516	17	528
Daily								
Number of times eat fried potatoes								
Never	2	11		13		13		13
Daily	3	28	1	30	1	30	3	28
Weekly		20		20		20		20
Monthly	3	13	1	15		16	1	15
Weekly								
Number of times eat fried potatoes								
Never	16	231	6	241	8	239	8	239
Daily	2	86	2	86	3	85	2	86
Weekly	44	1046	18	1072	26	1064	28	1062
Monthly	19	440	9	450	11	448	15	444
Monthly								
Number of times eat fried potatoes								
Never	63	933	21	975	30	966	44	952
Daily	9	107	5	111	9	107	6	110
Weekly	108	1668	30	1746	65	1711	56	1720
Monthly	145	2336	44	2437	74	2407	88	2393
Weekly								
Number of times eat pizza								
Never								
Number of times eat fried potatoes								
Never	76	629	22	683	38	667	35	670
Daily	4	31	2	33	3	32	3	32
Weekly	40	439	15	464	23	456	28	451
Monthly	35	368	10	393	17	386	26	377
Daily								
Number of times eat fried potatoes								
Never		9		9		9	1	8
Daily	1	4		5	1	4		5
Weekly	1	16		17		17		17
Monthly	1	13		14	1	13	2	12
Weekly								
Number of times eat fried potatoes								
Never	22	270	6	286	8	284	2	290
Daily	1	68		69		69	2	67
Weekly	92	1917	34	1975	58	1951	52	1957
Monthly	22	410	3	429	14	418	9	423
Monthly								
Number of times eat fried potatoes								
Never	69	789	16	842	34	824	35	823
Daily	5	58	3	60	2	61	2	61
Weekly	146	2469	49	2566	95	2520	87	2528
Monthly	130	1875	36	1969	71	1934	59	1946
Monthly								
Number of times eat pizza								
Never								
Number of times eat fried potatoes								
Never	22	179	5	196	18	183	16	185
Daily	2	3	1	4	1	4	1	4
Weekly	6	58		64	5	59	3	61
Monthly	42	316	12	346	25	333	24	334
Daily								
Number of times eat fried potatoes								
Never		2		2		2		2
Daily		4		4		4		4
Weekly		4		4	1	3		4
Monthly		7	1	6		7	1	6
Weekly								
Number of times eat fried potatoes								
Never	2	33		35	2	33	2	33
Daily	1	10		11		11		11
Weekly	6	156	3	159	4	158	4	158
Monthly	3	132	4	131	4	131	3	132
Monthly								
Number of times eat fried potatoes								
Never	23	299	11	311	18	304	23	299
Daily	2	38	1	39	4	36	1	39
Weekly	29	357	6	380	15	371	11	375
Monthly	129	2121	38	2212	77	2173	73	2177

Table 4: The Dietary Habits and The Number of Cardiovascular Diseases Chart

	BMI			
	Underweight	Healthy weight	Overweight	Obese
Number of times eat other vegetables				
Never				
Number of times eat pizza				
Never				
Number of times eat fried potatoes				
Never	2.142857	40.71429	29.28571	27.85714
Daily		35.71429	50	14.28571
Weekly	3.278689	29.5082	31.14754	36.06557
Monthly	1.5625	25	35.9375	37.5
Daily				
Number of times eat fried potatoes				
Never	33.33333	33.33333	33.33333	
Daily		66.66667	33.33333	
Weekly			75	25
Monthly		66.66667		33.33333
Weekly				
Number of times eat fried potatoes				
Never		24	44	32
Daily		20	26.66667	53.33333
Weekly	2.068966	24.82759	31.03448	42.06897
Monthly	2.777778	22.22222	36.11111	38.88889
Monthly				
Number of times eat fried potatoes				
Never	1.010101	27.27273	36.36364	35.35354
Daily		22.22222	16.66667	61.11111
Weekly	2.994012	21.55689	44.31138	31.13772
Monthly	1.492537	28.85572	28.35821	41.29353
Daily				
Number of times eat pizza				
Never				
Number of times eat fried potatoes				
Never	3.479237	41.18967	31.98653	23.34456
Daily	6.25	43.75	25	25
Weekly	1.552795	32.91925	35.40373	30.12422
Monthly	2.385321	37.24771	34.12844	26.23853
Daily				
Number of times eat fried potatoes				
Never	7.692308	46.15385	23.07692	23.07692
Daily		35.48387	35.48387	29.03226
Weekly		35	40	25
Monthly		37.5	31.25	31.25
Weekly				
Number of times eat fried potatoes				
Never	2.024291	39.27126	31.98381	26.72065
Daily	1.136364	34.09091	27.27273	37.5
Weekly	1.100917	33.30275	32.66055	32.93578
Monthly	1.089325	38.34423	34.20479	26.36166
Monthly				
Number of times eat fried potatoes				
Never	2.108434	38.85542	34.73896	24.29719
Daily	.862069	41.37931	27.58621	30.17241
Weekly	1.182432	31.25	34.74099	32.82658
Monthly	1.330109	32.20476	36.35631	30.10883
Weekly				
Number of times eat pizza				
Never				
Number of times eat fried potatoes				
Never	1.276596	35.74468	34.60993	28.36879
Daily		37.14286	37.14286	25.71429
Weekly	2.296451	31.52401	35.69937	30.48017
Monthly	1.240695	32.00993	37.22084	29.52854
Daily				
Number of times eat fried potatoes				
Never	11.11111	44.44444	33.33333	11.11111
Daily		40	60	
Weekly	11.76471	5.882353	52.94118	29.41176
Monthly		28.57143	21.42857	50
Weekly				
Number of times eat fried potatoes				
Never	.6849315	37.32877	31.50685	30.47945
Daily	1.449275	28.98551	40.57971	28.98551
Weekly	1.592832	29.31807	35.54007	33.54903
Monthly	.462963	31.01852	37.73148	30.78704
Monthly				
Number of times eat fried potatoes				
Never	2.564103	35.31469	30.18648	31.93473
Daily		26.98413	26.98413	46.03175
Weekly	1.414914	29.10134	33.61377	35.86998
Monthly	1.745636	30.97257	34.26434	33.01746
Monthly				
Number of times eat pizza				
Never				
Number of times eat fried potatoes				
Never	1.492537	25.37313	41.29353	31.8408
Daily		20	40	40
Weekly	4.6875	37.5	21.875	35.9375
Monthly	1.117318	31.00559	31.56425	36.31285
Daily				
Number of times eat fried potatoes				
Never		50		50
Daily		25	25	50
Weekly			75	25
Monthly		28.57143	42.85714	28.57143
Weekly				
Number of times eat fried potatoes				
Never	2.857143	37.14286	25.71429	34.28571
Daily	9.090909	36.36364	18.18182	36.36364
Weekly	1.234568	30.8642	29.01235	38.88889
Monthly	.7407407	28.14815	40.74074	30.37037
Monthly				
Number of times eat fried potatoes				
Never	1.552795	32.91925	30.74534	34.78261
Daily		35	20	45
Weekly	2.072539	25.90674	33.41969	38.60104
Monthly	1.511111	27.46667	33.33333	37.68889

Table 5: The Dietary Habits and BMI Frequency Chart

	BMI			
	Underweight	Healthy weight	Overweight	Obese
Number of times eat other vegetables				
Never				
Number of times eat pizza				
Never				
Number of times eat fried potatoes				
Never	3	57	41	39
Daily		5	7	2
Weekly	2	18	19	22
Monthly	1	16	23	24
Daily				
Number of times eat fried potatoes				
Never	1	1	1	
Daily		2	1	
Weekly			3	1
Monthly		2		1
Weekly				
Number of times eat fried potatoes				
Never		6	11	8
Daily		3	4	8
Weekly	3	36	45	61
Monthly	1	8	13	14
Monthly				
Number of times eat fried potatoes				
Never	1	27	36	35
Daily		4	3	11
Weekly	5	36	74	52
Monthly	3	58	57	83
Daily				
Number of times eat pizza				
Never				
Number of times eat fried potatoes				
Never	31	367	285	208
Daily	2	14	8	8
Weekly	5	106	114	97
Monthly	13	203	186	143
Daily				
Number of times eat fried potatoes				
Never	1	6	3	3
Daily		11	11	9
Weekly		7	8	5
Monthly		6	5	5
Weekly				
Number of times eat fried potatoes				
Never	5	97	79	66
Daily	1	30	24	33
Weekly	12	363	356	359
Monthly	5	176	157	121
Monthly				
Number of times eat fried potatoes				
Never	21	387	346	242
Daily	1	48	32	35
Weekly	21	555	617	583
Monthly	33	799	902	747
Weekly				
Number of times eat pizza				
Never				
Number of times eat fried potatoes				
Never	9	252	244	200
Daily		13	13	9
Weekly	11	151	171	146
Monthly	5	129	150	119
Daily				
Number of times eat fried potatoes				
Never	1	4	3	1
Daily		2	3	
Weekly	2	1	9	5
Monthly		4	3	7
Weekly				
Number of times eat fried potatoes				
Never	2	109	92	89
Daily	1	20	28	20
Weekly	32	589	714	674
Monthly	2	134	163	133
Monthly				
Number of times eat fried potatoes				
Never	22	303	259	274
Daily		17	17	29
Weekly	37	761	879	938
Monthly	35	621	687	662
Monthly				
Number of times eat pizza				
Never				
Number of times eat fried potatoes				
Never	3	51	83	64
Daily		1	2	2
Weekly	3	24	14	23
Monthly	4	111	113	130
Daily				
Number of times eat fried potatoes				
Never		1		1
Daily		1	1	2
Weekly			3	1
Monthly		2	3	2
Weekly				
Number of times eat fried potatoes				
Never	1	13	9	12
Daily	1	4	2	4
Weekly	2	50	47	63
Monthly	1	38	55	41
Monthly				
Number of times eat fried potatoes				
Never	5	106	99	112
Daily		14	8	18
Weekly	8	100	129	149
Monthly	34	618	750	848

Table 6: The Dietary Habits and The Number of BMI Chart

3.1 Life Styles with CVC

3.1.1 Life Styles with Coronary Heart Disease

First of all, we can carry out descriptive statistics on coronary heart disease, as shown in the following figure, which can tell us the basic situation of all variables. However, since the variables we deal with here are discrete and the problem we deal with is binary classification, data such as average values have little practical analysis value, and other data will not be shown in this way.

	CHDEV_A	FRJUICTP_A	COFFEENOTP_A	SALADTP_A	FRIESTP_A	\
count	26285.000000	26285.000000	26285.000000	26285.000000	26285.000000	
mean	0.062811	1.000875	1.214381	1.759026	1.229028	
std	0.242628	1.077413	1.300483	0.864941	0.795170	
min	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	0.000000	0.000000	0.000000	1.000000	1.000000	
50%	0.000000	1.000000	1.000000	2.000000	1.000000	
75%	0.000000	2.000000	3.000000	2.000000	2.000000	
max	1.000000	3.000000	3.000000	3.000000	3.000000	

	BEANSTP_A	PIZZATP_A	OVEGTP_A
count	26285.000000	26285.000000	26285.000000
mean	1.154118	1.046338	2.120905
std	0.834654	0.651245	0.837000
min	0.000000	0.000000	0.000000
25%	1.000000	1.000000	2.000000
50%	1.000000	1.000000	2.000000
75%	2.000000	1.000000	3.000000
max	3.000000	3.000000	3.000000

Figure 14: Descriptive Statistics of Life Styles and Coronary Heart Disease

We can use the correlation matrix to explore the relationship between various variables and show them through visual methods.

	CHDEV_A	FRJUICTP_A	COFFEENOTP_A	SALADTP_A	FRIESTP_A	\
CHDEV_A	1.000000	0.014489	-0.021215	-0.028853	-0.032168	
FRJUICTP_A	0.014489	1.000000	0.094848	0.058485	0.127751	
COFFEENOTP_A	-0.021215	0.094848	1.000000	0.064125	0.077608	
SALADTP_A	-0.028853	0.058485	0.064125	1.000000	0.024155	
FRIESTP_A	-0.032168	0.127751	0.077608	0.024155	1.000000	
BEANSTP_A	0.009309	0.120300	0.086244	0.138876	0.134438	
PIZZATP_A	-0.050204	0.071408	0.048286	0.019622	0.320181	
OVEGTP_A	-0.018662	0.054011	0.051265	0.375901	0.016472	

	BEANSTP_A	PIZZATP_A	OVEGTP_A
CHDEV_A	0.009309	-0.050204	-0.018662
FRJUICTP_A	0.120300	0.071408	0.054011
COFFEENOTP_A	0.086244	0.048286	0.051265
SALADTP_A	0.138876	0.019622	0.375901
FRIESTP_A	0.134438	0.320181	0.016472
BEANSTP_A	1.000000	0.093601	0.138939
PIZZATP_A	0.093601	1.000000	0.032228
OVEGTP_A	0.138939	0.032228	1.000000

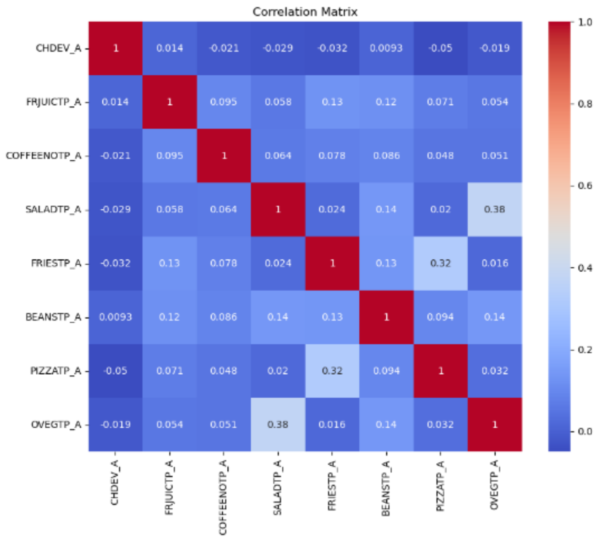


Figure 15: Correlation Matrix of Life Styles and Coronary Heart Disease

In addition to these kind of statistical methods described above, we can also use some model, to help us better understand and use the data set. According to the figure below, the logistic regression analysis on dietary factors for heart disease showed significant overall model fit (chi2=118.58, p-value=0.0000), with a pseudo R2 of 0.0104. The results indicated that the frequency of fruit juice intake (coefficient=0.0939676, p-value=0.000) and bean intake (coefficient=0.1194858, p-value=0.000) are significantly positively associated with heart disease. Conversely, the frequencies of coffee (coefficient=-0.0507593, p-value=0.015), salad (coefficient=-0.1249328, p-value=0.000), fried potatoes (coefficient=-0.1002222, p-value=0.005), and pizza (coefficient=-0.2981702, p-value=0.000) intake are significantly negatively associated with heart disease. The frequency of other vegetable intake showed no significant association with heart disease (coefficient=-0.0502875, p-value=0.135).

```
. logit CHDEV_A_B FRJUICTP_A_R COFFEENOTP_A_R SALADTP_A_R FRIESTP_A_R BEANSTP_A_R PIZZATP_A_R OVEGTP_A_R

Iteration 0:   log likelihood = -5721.1256
Iteration 1:   log likelihood = -5662.7832
Iteration 2:   log likelihood = -5661.8376
Iteration 3:   log likelihood = -5661.8374

Logistic regression                                     Number of obs = 24,117
                                                        LR chi2(7)    = 118.58
                                                        Prob > chi2   = 0.0000
Log likelihood = -5661.8374                             Pseudo R2     = 0.0104

      CHDEV_A_B | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
----------------+----------------------------------------------------------------
   FRJUICTP_A_R |   .0939676   .0245718     3.82   0.000     .0458077    .1421275
 COFFEENOTP_A_R |  -.0507593   .0209264    -2.43   0.015    -.0917744   -.0097443
    SALADTP_A_R |  -.1249328   .0323312    -3.86   0.000    -.1883009   -.0615647
    FRIESTP_A_R |  -.1002222   .0355287    -2.82   0.005    -.1698572   -.0305872
    BEANSTP_A_R |   .1194858   .0333647     3.58   0.000     .0540922    .1848794
    PIZZATP_A_R |  -.2981702   .0435829    -6.84   0.000    -.3835911   -.2127493
     OVEGTP_A_R |  -.0502875   .0336234    -1.50   0.135    -.1161881    .0156131
          _cons |  -2.126525   .0913319   -23.28   0.000    -2.305532   -1.947518
```

Figure 16: Life Styles with Coronary Heart Disease

There are some numerical differences in this result when using different tools, but the general trend is the same. This is probably due to the different processing of the top of the default debit, regularization, optimization of the algorithm and convergence criteria, and so on. graph of the results of doing the same processing in python. When I do the same processing in other diseases, I will release the two images directly.

	Feature	Coefficient	Correlation
4	BEANSTP_A	0.085295	Positive
0	FRJUICTP_A	0.057465	Positive
6	OVEGTP_A	-0.060130	Negative
1	COFFEENOTP_A	-0.067461	Negative
3	FRIESTP_A	-0.095403	Negative
2	SALADTP_A	-0.139666	Negative
5	PIZZATP_A	-0.322662	Negative

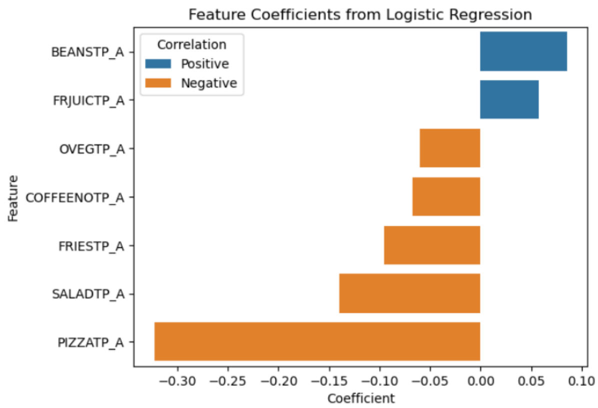


Figure 17: Feature Coefficients of Life Styles with Coronary Heart Disease

In addition to logistic regression, we can use other models to make some predictions about pathology diagnosis using data on eating habits. Below I will show how well KNN, K-means, Decision Tree, Random Forest and Catboost models make predictions for Coronary Heart Disease.

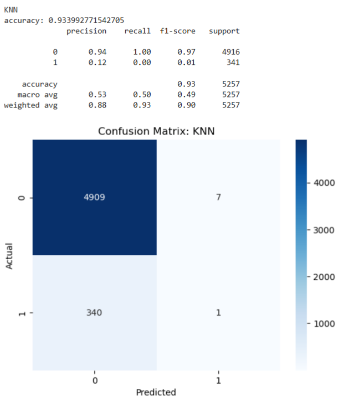


Figure 18: KNN Prediction

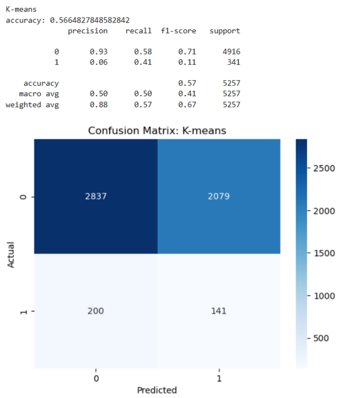


Figure 19: K-means Prediction

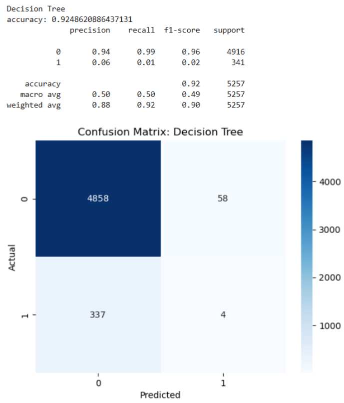


Figure 20: Decision Tree Prediction

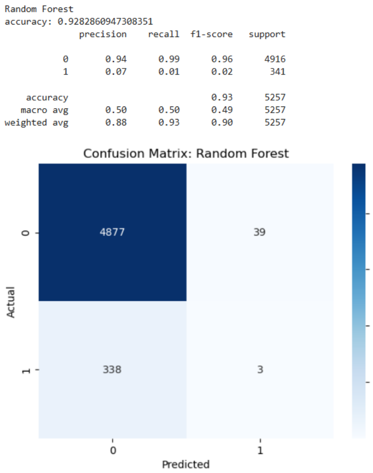


Figure 21: Random Forest Prediction

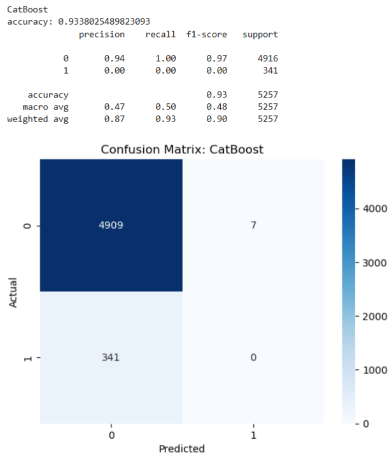


Figure 22: CatBoost Prediction

From the above five images, it can be noticed that KNN, Random Forest and Catboost have the best prediction performance. K-means performs worse because it is a clustering algorithm, not a classification algorithm. Its goal is to categorize the data into K clusters, not to predict category labels, and K-means is very sensitive to noise. I will not show the K-means results when dealing with other cardiovascular diseases.

In addition, the Decision Tree, the Random Forest and Catboost because of their model features, we can observe it features of reference importance, as shown in the figure below.

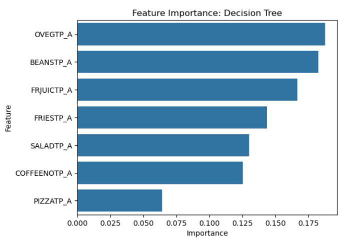


Figure 23: Decision Tree Feature Importance

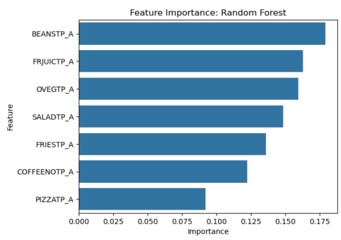


Figure 24: Random Forest Feature Importance

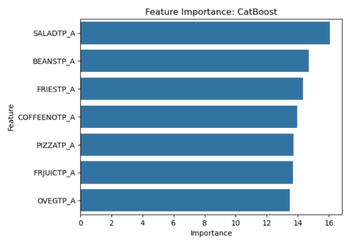
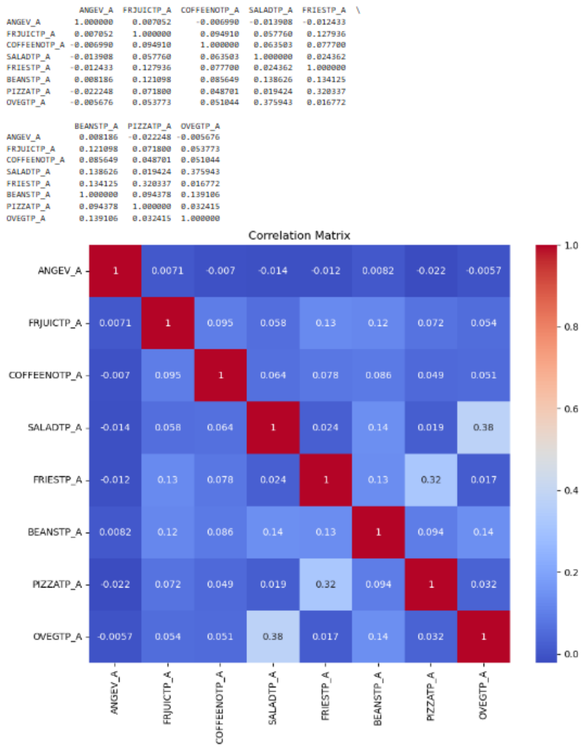


Figure 25: CatBoost Importance

3.1.2 Life Styles with Angina

In explore the life styles and the relationship between Angina, we can also finish a Correlation Matrix, to roughly understand the relations between the coefficient of each variable.



	ANGEV_A	FRJUICTP_A	COFFEENOTP_A	SALADTP_A	FRIESTP_A	\
ANGEV_A	1.000000	0.007052	-0.006990	-0.013908	-0.012433	
FRJUICTP_A	0.007052	1.000000	0.094910	0.057760	0.127936	
COFFEENOTP_A	-0.006990	0.094910	1.000000	0.063503	0.077700	
SALADTP_A	-0.013908	0.057760	0.063503	1.000000	0.024362	
FRIESTP_A	-0.012433	0.127936	0.077700	0.024362	1.000000	
BEANSTP_A	0.008186	0.121098	0.085649	0.138626	0.134125	
PIZZATP_A	-0.022248	0.071800	0.048701	0.019424	0.320337	
OVEGTP_A	-0.005676	0.053773	0.051044	0.375943	0.016772	

	BEANSTP_A	PIZZATP_A	OVEGTP_A
ANGEV_A	0.008186	-0.022248	-0.005676
FRJUICTP_A	0.121098	0.071800	0.053773
COFFEENOTP_A	0.085649	0.048701	0.051044
SALADTP_A	0.138626	0.019424	0.375943
FRIESTP_A	0.134125	0.320337	0.016772
BEANSTP_A	1.000000	0.094378	0.139106
PIZZATP_A	0.094378	1.000000	0.032415
OVEGTP_A	0.139106	0.032415	1.000000

Figure 26: Correlation Matrix of Life Styles and Angina

By investigating the association of dietary habits with Angina by logistic regression, we obtained the following figure. By this figure we can find that drink pure fruit juice, drink coffee with sugar, eat salad, fired potatoes and other vegetables while all showed a certain negative correlation, but they all crossed zero at the 95% confidence interval value, it makes them in the statistical significance is not big, low influence not for reference. While eating beans was significantly positive correlation in the survey, using the pizza for significant negative correlation.

```
. logit ANGEV_A_B FRJUICTP_A_R COFFEENOTP_A_R SALADTP_A_R FRIESTP_A_R BEANSTP_A_R PIZZATP_A_R OVEGTP_A_R

Iteration 0:   log likelihood = -2355.2966
Iteration 1:   log likelihood = -2343.5205
Iteration 2:   log likelihood = -2343.3756
Iteration 3:   log likelihood = -2343.3756

Logistic regression                                     Number of obs = 24,117
                                                        LR chi2(7)    =  23.84
                                                        Prob > chi2   = 0.0012
Log likelihood = -2343.3756                             Pseudo R2     = 0.0051

------------------------------------------------------------------------------
     ANGEV_A_B | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
---------------+--------------------------------------------------------------
  FRJUICTP_A_R |    .068064    .042894     1.59   0.113    -.0160067    .1521346
COFFEENOTP_A_R |    -.01959   .0361552    -0.54   0.588    -.0904529    .0512729
   SALADTP_A_R |  -.0993066   .0566381    -1.75   0.080    -.2103152     .011702
   FRIESTP_A_R |  -.0412976    .061899    -0.67   0.505    -.1626174    .0800223
   BEANSTP_A_R |   .1487512   .0581078     2.56   0.010     .0348621    .2626403
   PIZZATP_A_R |  -.2482801   .0756732    -3.28   0.001    -.3965969   -.0999632
    OVEGTP_A_R |   -.007367    .059528    -0.12   0.902    -.1240397    .1093057
         _cons |  -3.630915    .163602   -22.19   0.000    -3.951569   -3.310261
------------------------------------------------------------------------------
```

Figure 27: Life Styles with Angina

	Feature	Coefficient	Correlation
4	BEANSTP_A	0.096106	Positive
0	FRJUICTP_A	0.051193	Positive
6	OVEGTP_A	0.004503	Positive
1	COFFEENOTP_A	-0.051571	Negative
3	FRIESTP_A	-0.072897	Negative
2	SALADTP_A	-0.151024	Negative
5	PIZZATP_A	-0.193653	Negative

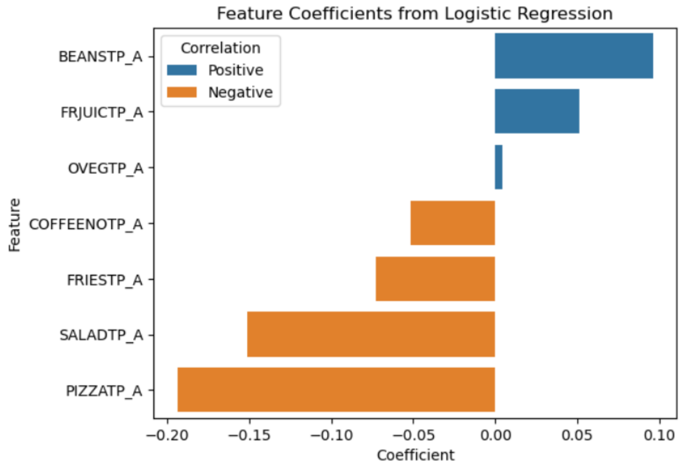


Figure 28: Feature Coefficients of Life Styles with Angina

Four models, KNN, Decision Tree, Random Forest and CatBoost are used to predict the probability that Angina is diagnosed. And the associated feature importance picture is obtained as shown below.

KNN
accuracy: 0.9828799695643904

	precision	recall	f1-score	support
0	0.98	1.00	0.99	5168
1	0.00	0.00	0.00	89
accuracy			0.98	5257
macro avg	0.49	0.50	0.50	5257
weighted avg	0.97	0.98	0.97	5257

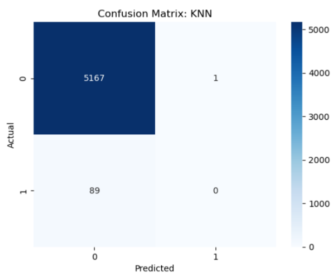


Figure 29: KNN Prediction

Decision Tree
accuracy: 0.9786950732356857

	precision	recall	f1-score	support
0	0.98	1.00	0.99	5168
1	0.04	0.01	0.02	89
accuracy			0.98	5257
macro avg	0.51	0.50	0.50	5257
weighted avg	0.97	0.98	0.97	5257

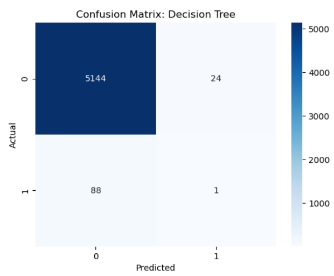


Figure 30: Decision Tree Prediction

Random Forest
accuracy: 0.9809777439604337

	precision	recall	f1-score	support
0	0.98	1.00	0.99	5168
1	0.08	0.01	0.02	89
accuracy			0.98	5257
macro avg	0.53	0.50	0.51	5257
weighted avg	0.97	0.98	0.97	5257

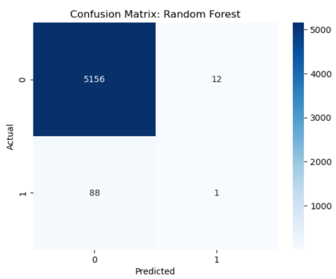


Figure 31: Random Forest Prediction

CatBoost
accuracy: 0.9817386342020163

	precision	recall	f1-score	support
0	0.98	1.00	0.99	5168
1	0.00	0.00	0.00	89
accuracy			0.98	5257
macro avg	0.49	0.50	0.50	5257
weighted avg	0.97	0.98	0.97	5257

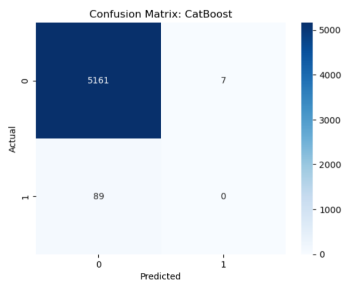


Figure 32: CatBoost Prediction

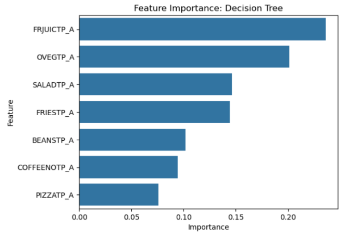


Figure 33: Decision Tree Feature Importance

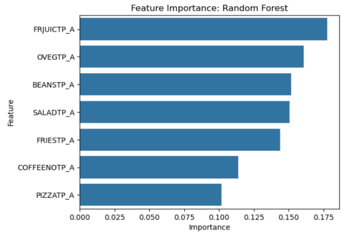


Figure 34: Random Forest Feature Importance

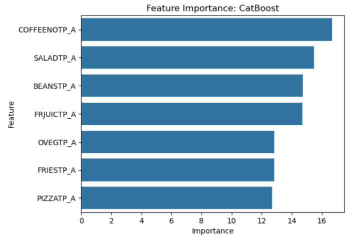
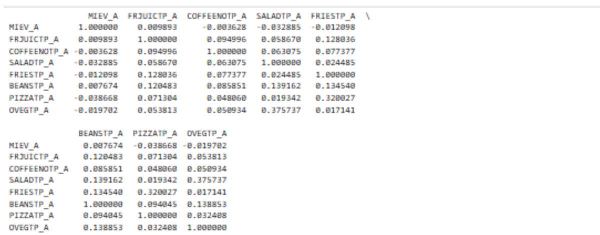


Figure 35: CatBoost Importance

3.1.3 Life Styles with Heart Attack

When dealing with Life Styles and Heart Attack, we still first look at their correlation matrix to get a general idea of the coefficient relationship between these variables, and also to get a general idea of the importance of the variables.



	MIEV_A	FRJUICTP_A	COFFEENOTP_A	SALADTP_A	FRIESTP_A
MIEV_A	1.000000	0.009893	-0.003628	-0.032885	-0.012098
FRJUICTP_A	0.009893	1.000000	0.094996	0.058670	0.128036
COFFEENOTP_A	-0.003628	0.094996	1.000000	0.063075	0.077377
SALADTP_A	-0.032885	0.058670	0.063075	1.000000	0.024485
FRIESTP_A	-0.012098	0.128036	0.077377	0.024485	1.000000
BEANSTP_A	0.007674	0.120483	0.085851	0.139162	0.134540
PIZZATP_A	-0.038668	0.071304	0.048060	0.019342	0.320027
OVEGTP_A	-0.019702	0.053813	0.050934	0.375737	0.017141

	BEANSTP_A	PIZZATP_A	OVEGTP_A
MIEV_A	0.007674	-0.038668	-0.019702
FRJUICTP_A	0.120483	0.071304	0.053813
COFFEENOTP_A	0.085851	0.048060	0.050934
SALADTP_A	0.139162	0.019342	0.375737
FRIESTP_A	0.134540	0.320027	0.017141
BEANSTP_A	1.000000	0.094045	0.138853
PIZZATP_A	0.094045	1.000000	0.032408
OVEGTP_A	0.138853	0.032408	1.000000

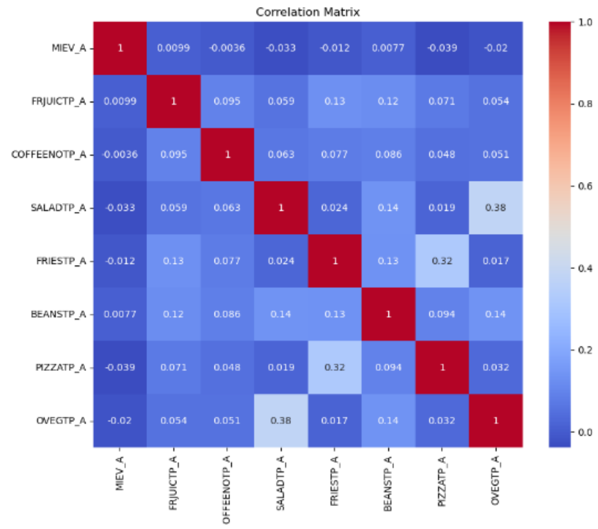


Figure 36: Correlation Matrix of Life Styles and Heart Attack

Investigating the association between dietary habits and Heart Attack through logistic regression, we obtained the following graph. From this figure we can find that drinking sweetened coffee and eating fired potatoes still both cross the 0 value within the 95% confidence interval and their p-values are greater than 0.5, indicating that their effects are not significant. While quoting pure fruit juice and using beans are significantly positively correlated in this survey. The use of salads and pizzas were significantly negatively correlated.

```
Iteration 0:   log likelihood = -3800.0926
Iteration 1:   log likelihood =   -3759.31
Iteration 2:   log likelihood = -3758.3553
Iteration 3:   log likelihood = -3758.3548
Iteration 4:   log likelihood = -3758.3548

Logistic regression                                     Number of obs = 24,117
                                                        LR chi2(7)    =  83.48
                                                        Prob > chi2   = 0.0000
Log likelihood = -3758.3548                             Pseudo R2     = 0.0110
```

MIEV_A_B	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
FRJUICTP_A_R	.0652817	.03204	2.04	0.042	.0024846	.1280789
COFFEENOTP_A_R	.0169779	.0267774	0.63	0.526	-.0355049	.0694606
SALADTP_A_R	-.2121466	.0412423	-5.14	0.000	-.2929801	-.1313132
FRIESTP_A_R	-.0114641	.0461538	-0.25	0.804	-.1019239	.0789956
BEANSTP_A_R	.1276937	.0433153	2.95	0.003	.0427973	.2125902
PIZZATP_A_R	-.3375918	.0565093	-5.97	0.000	-.448348	-.2268356
OVEGTP_A_R	-.0609579	.0430863	-1.41	0.157	-.1454056	.0234897
_cons	-2.667476	.1167794	-22.84	0.000	-2.89636	-2.438593

Figure 37: Life Styles with Heart Attack

	Feature	Coefficient	Correlation
0	FRJUICTP_A	0.089844	Positive
4	BEANSTP_A	0.084452	Positive
1	COFFEENOTP_A	-0.017306	Negative
3	FRIESTP_A	-0.020393	Negative
6	OVEGTP_A	-0.074612	Negative
2	SALADTP_A	-0.167934	Negative
5	PIZZATP_A	-0.360358	Negative

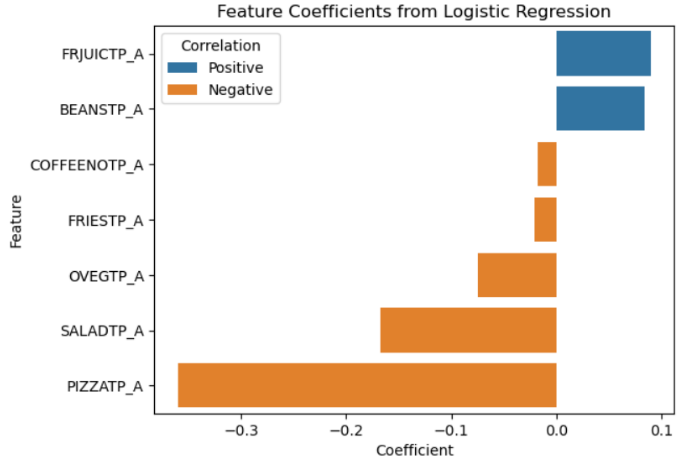


Figure 38: Feature Coefficients of Life Styles with Heart Attack

We also make a prediction of Heart Attack being diagnosed using the four models described above, and their predictions are shown below.

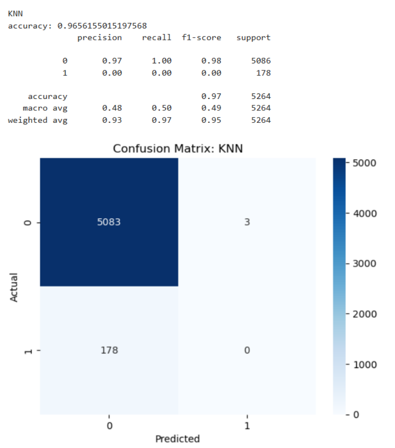


Figure 39: KNN Prediction

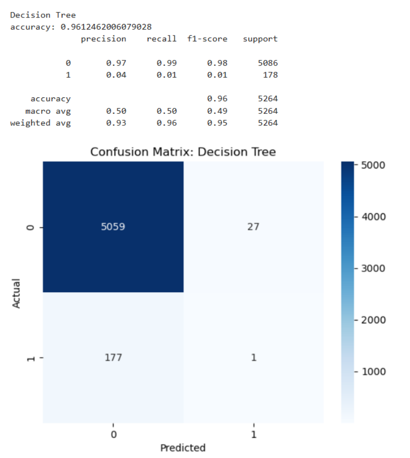


Figure 40: Decision Tree Prediction

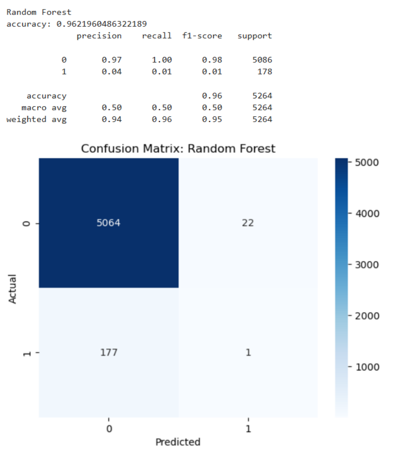


Figure 41: Random Forest Prediction

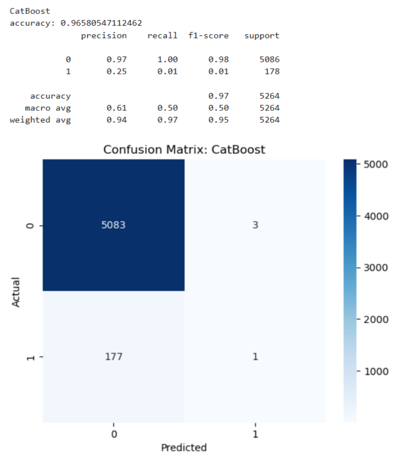


Figure 42: CatBoost Prediction

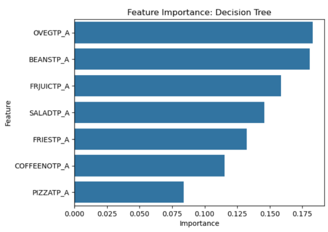


Figure 43: Decision Tree Feature Importance

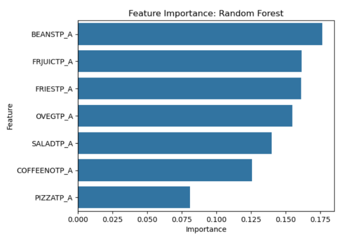


Figure 44: Random Forest Feature Importance

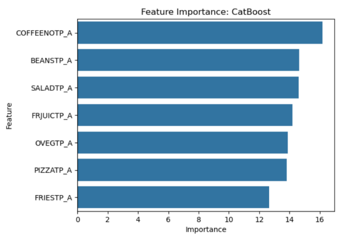


Figure 45: CatBoost Importance

3.1.4 Life Styles with Stroke

The following figure shows the coefficient relationship between various dietary habits and stroke in the dataset.

	STREV_A	FRJUICTP_A	COFFEENOTP_A	SALADTP_A	FRIESTP_A	\
STREV_A	1.000000	0.006182	-0.003600	-0.041873	-0.020466	
FRJUICTP_A	0.006182	1.000000	0.094862	0.058946	0.128126	
COFFEENOTP_A	-0.003600	0.094862	1.000000	0.063151	0.077139	
SALADTP_A	-0.041873	0.058946	0.063151	1.000000	0.024653	
FRIESTP_A	-0.020466	0.128126	0.077139	0.024653	1.000000	
BEANSTP_A	-0.009632	0.120646	0.085663	0.139032	0.134785	
PIZZATP_A	-0.037140	0.071761	0.048046	0.019287	0.319780	
OVEGTP_A	-0.018627	0.054063	0.051367	0.375947	0.017192	

	BEANSTP_A	PIZZATP_A	OVEGTP_A
STREV_A	-0.009632	-0.037140	-0.018627
FRJUICTP_A	0.120646	0.071761	0.054063
COFFEENOTP_A	0.085663	0.048046	0.051367
SALADTP_A	0.139032	0.019287	0.375947
FRIESTP_A	0.134785	0.319780	0.017192
BEANSTP_A	1.000000	0.094119	0.138778
PIZZATP_A	0.094119	1.000000	0.032333
OVEGTP_A	0.138778	0.032333	1.000000

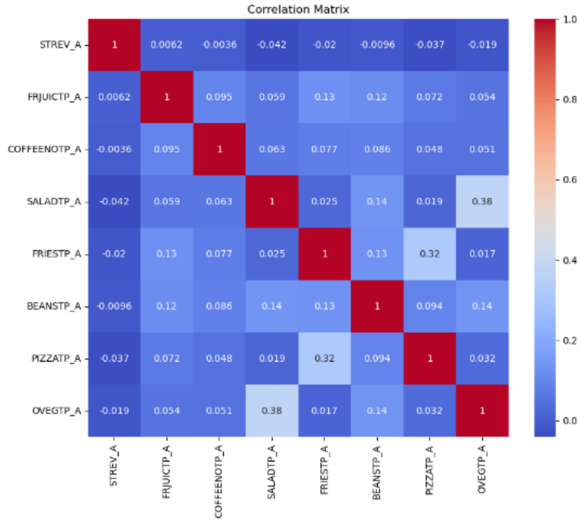


Figure 46: Correlation Matrix of Life Styles and Stroke

Analyzing the relationship between dietary habits and stroke through logistic regression, we obtained the following results: as can be seen from the graph, sweetened coffee, fried potatoes, beans, or other vegetables all crossed a value of 0 or had a p-value of greater than 0.5 within the 95% confidence interval, suggesting that they did not have a significant effect. However, pure fruit juices showed a significant positive correlation in the survey, while salads and pizzas showed a significant negative correlation.

```
. logit STREV_A_B FRJUICTP_A_R COFFEENOTP_A_R SALADTP_A_R FRIESTP_A_R BEANSTP_A_R PIZZATP_A_R OVEGTP_A_R

Iteration 0:   log likelihood = -3754.2623
Iteration 1:   log likelihood = -3713.4373
Iteration 2:   log likelihood = -3712.3975
Iteration 3:   log likelihood = -3712.3969
Iteration 4:   log likelihood = -3712.3969

Logistic regression                                     Number of obs = 24,117
                                                        LR chi2(7)    =  83.73
                                                        Prob > chi2   = 0.0000
Log likelihood = -3712.3969                             Pseudo R2     = 0.0112

------------------------------------------------------------------------------
     STREV_A_B | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
---------------+--------------------------------------------------------------
  FRJUICTP_A_R |    .075652   .0322691     2.34   0.019     .0124058    .1388982
COFFEENOTP_A_R |  -.0035605   .0271317    -0.13   0.896    -.0567377    .0496167
   SALADTP_A_R |  -.2426575   .0411961    -5.89   0.000    -.3234003   -.1619147
   FRIESTP_A_R |  -.0412768   .0466153    -0.89   0.376    -.1326411    .0500874
   BEANSTP_A_R |  -.0002675   .0436485    -0.01   0.995     -.085817    .0852819
   PIZZATP_A_R |  -.3199817   .0570943    -5.60   0.000    -.4318844    -.208079
    OVEGTP_A_R |  -.0051193   .0433325    -0.12   0.906    -.0900494    .0798108
         _cons |  -2.573354   .1171129   -21.97   0.000    -2.802891   -2.343817
------------------------------------------------------------------------------
```

Figure 47: Life Styles with Stroke

	Feature	Coefficient	Correlation
0	FRJUICTP_A	0.044239	Positive
1	COFFEENOTP_A	-0.004574	Negative
6	OVEGTP_A	-0.024074	Negative
4	BEANSTP_A	-0.030378	Negative
3	FRIESTP_A	-0.046252	Negative
2	SALADTP_A	-0.228873	Negative
5	PIZZATP_A	-0.322015	Negative

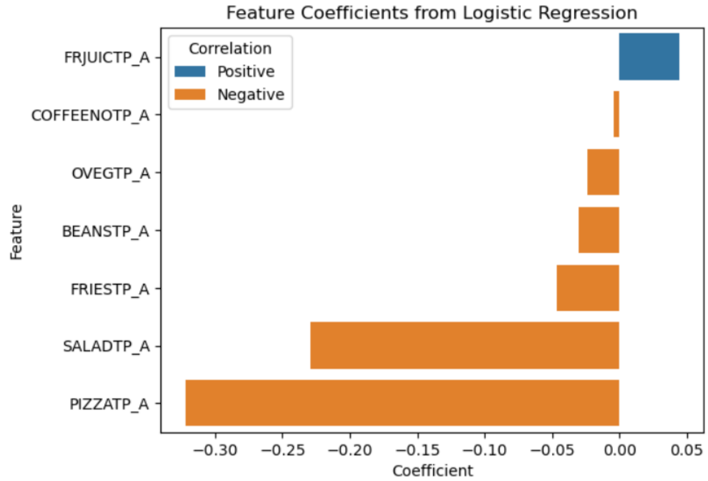


Figure 48: Feature Coefficients of Life Styles with Stroke

The next images will show how well the four models predict stroke, and the importance of the variables in these models.

KNN
accuracy: 0.9631458966565349

	precision	recall	f1-score	support
0	0.97	1.00	0.98	5081
1	0.00	0.00	0.00	183
accuracy			0.96	5264
macro avg	0.48	0.50	0.49	5264
weighted avg	0.93	0.96	0.95	5264

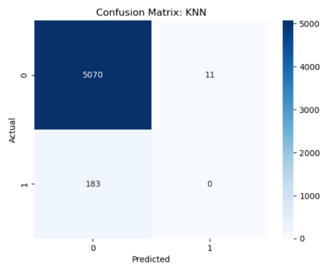


Figure 49: KNN Prediction

Decision Tree
accuracy: 0.9587765957446809

	precision	recall	f1-score	support
0	0.97	0.99	0.98	5081
1	0.05	0.01	0.02	183
accuracy			0.96	5264
macro avg	0.51	0.50	0.50	5264
weighted avg	0.93	0.96	0.95	5264

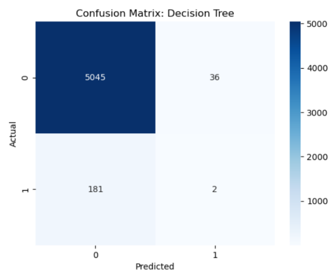


Figure 50: Decision Tree Prediction

Random Forest
accuracy: 0.9610562310030395

	precision	recall	f1-score	support
0	0.97	1.00	0.98	5081
1	0.00	0.00	0.00	183
accuracy			0.96	5264
macro avg	0.48	0.50	0.49	5264
weighted avg	0.93	0.96	0.95	5264

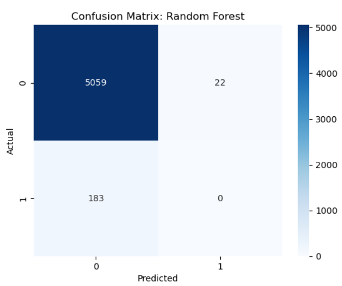


Figure 51: Random Forest Prediction

CatBoost
accuracy: 0.9642857142857143

	precision	recall	f1-score	support
0	0.97	1.00	0.98	5081
1	0.00	0.00	0.00	183
accuracy			0.96	5264
macro avg	0.48	0.50	0.49	5264
weighted avg	0.93	0.96	0.95	5264

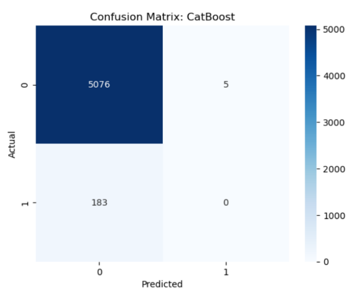


Figure 52: CatBoost Prediction

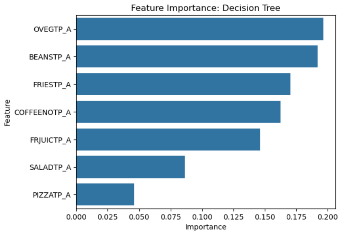


Figure 53: Decision Tree Feature Importance

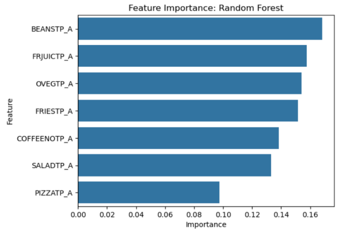


Figure 54: Random Forest Feature Importance

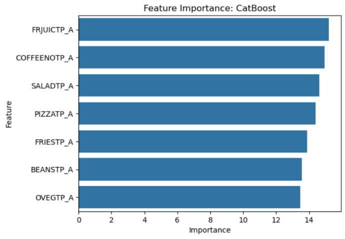


Figure 55: CatBoost Importance

3.1.5 Evaluation

The conclusions generated by these statistical software programs cannot be used to develop a dietary program and may even be contrary to your doctor's recommendations. These preliminary conclusions are based on statistical calculations only and are therefore influenced by the data set used. In performing the logistic regression analysis, although I cleaned the data and selected the variables, the results may be subject to error due to factors such as modeling and sample representativeness. The variables selected may not be representative enough or selected in a limited direction, and some variables may need to work together to show a certain effect. Each conclusion can only reflect the statistically specific performance of the data set used.

For model prediction accuracy, we take the KNN model for Coronary Heart Disease as an example. Although the prediction accuracy of this model is as high as 93%, if we look at the other parameters in detail, we can see that the model performs very poorly for category 1 (diagnosis of Coronary Heart Disease). The model almost completely ignores category 1, recognizing only 1 of the 341 category 1s in the test, and incorrectly identifying 7 category 0s as category 1s, resulting in false-positive cases. This resulted in low precision, recall and F1-score for category 1. This problem is usually caused by an imbalance in the number of categories. In the future, we can adjust the categorie weights so that the model focuses more on a small number of classes. Or train a better model by hyperparameterization.

3.2 Life Styles with BMI

The figure below shows the coefficient relationship between various dietary habits and BMI in the data set.

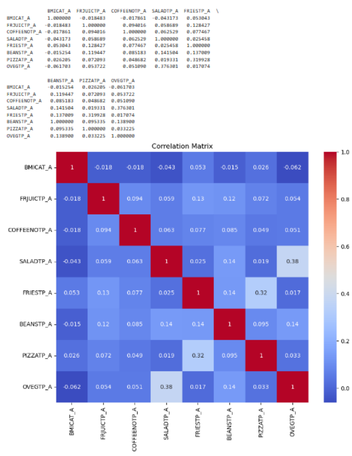


Figure 56: Correlation Matrix of Life Styles and BMI

In the above we are proceeding with logistic regression using stata, simply because the dependent variable of logistic regression is binary. And below we are going to explore the relationship between BMI and life styles, in our codebook, BMI is categorized into four degrees, so logistic regression is no longer applicable. Here we consider Ordered Logistic Regression or Multinomial Logistic Regression. and because BMI as the dependent variable, 1, 2, 3, 4 indicates four different degrees, we choose to use Ordered Logistic Regression.

As shown in the figure below, drinking pure fruit juice, eating salads with sugar coffee or tee, beans and other vegetables are all significantly negatively correlated with BMI. While eating fried potatoes and pizza had a significant positive correlation with BMI. It should be noted that none of the food items with positive or negative correlations were better with more. Maintaining a healthy BMI requires eating the right diet.

```
. ologit BMICAT_A FRJUICTP_A_R COFFEENOTP_A_R SALADTP_A_R FRIESTP_A_R BEANSTP_A_R PIZZATP_A_R OVEGTP_A_R

Iteration 0:   log likelihood = -28058.438
Iteration 1:   log likelihood = -27946.655
Iteration 2:   log likelihood = -27946.614
Iteration 3:   log likelihood = -27946.614

Ordered logistic regression                          Number of obs = 24,117
                                                     LR chi2(7)    = 223.65
                                                     Prob > chi2   = 0.0000
Log likelihood = -27946.614                          Pseudo R2     = 0.0040
```

BMICAT_A	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
FRJUICTP_A_R	-.0331223	.0111791	-2.96	0.003	-.0550329	-.0112118
COFFEENOTP_A_R	-.0314217	.009194	-3.42	0.001	-.0494416	-.0134019
SALADTP_A_R	-.0543557	.0149358	-3.64	0.000	-.0836294	-.025082
FRIESTP_A_R	.1374537	.0160672	8.55	0.000	.1059626	.1689449
BEANSTP_A_R	-.0469814	.014877	-3.16	0.002	-.0761398	-.017823
PIZZATP_A_R	.040302	.0192845	2.09	0.037	.0025052	.0780989
OVEGTP_A_R	-.1080907	.0154482	-7.00	0.000	-.1383686	-.0778128
/cut1	-4.355978	.0661143			-4.485559	-4.226396
/cut2	-.9269358	.0440173			-1.013208	-.8406635
/cut3	.5051468	.0436774			.4195407	.590753

Figure 57: Life Styles with BMI

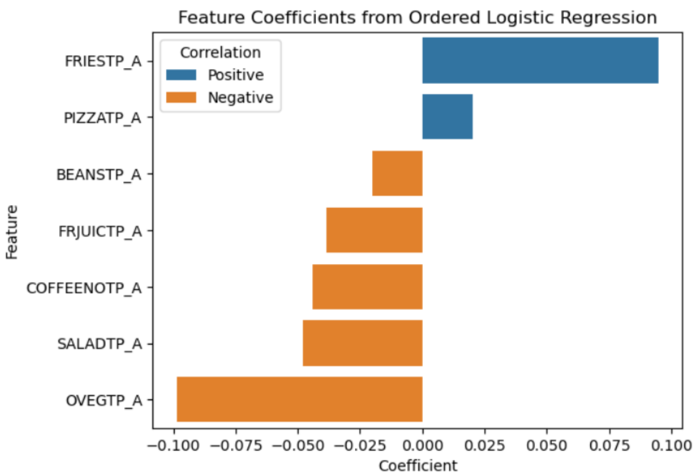


Figure 58: Feature Coefficients of Life Styles with BMI

When dealing with the relationship between life styles and BMI, I still used the previous four models to make predictions. Here are the models of rendering and variable importance.

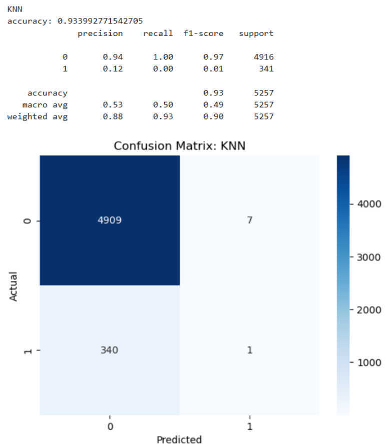


Figure 59: KNN Prediction

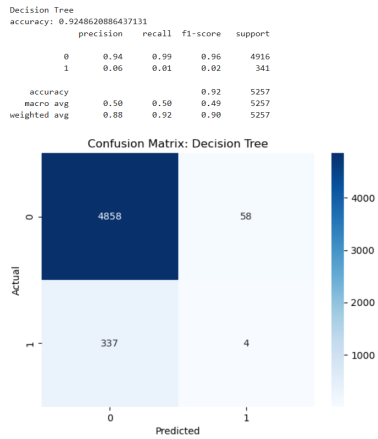


Figure 60: Decision Tree Prediction

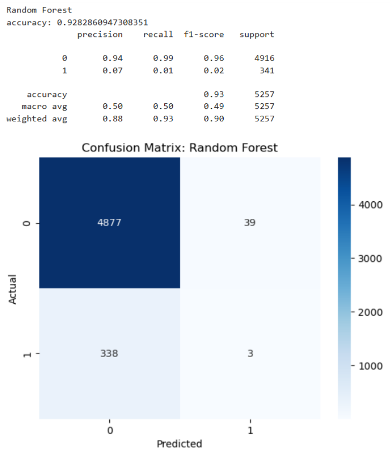


Figure 61: Random Forest Prediction

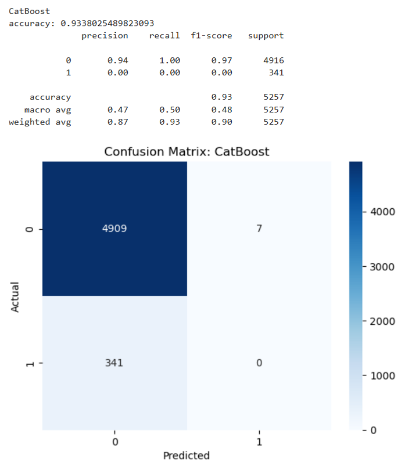


Figure 62: CatBoost Prediction

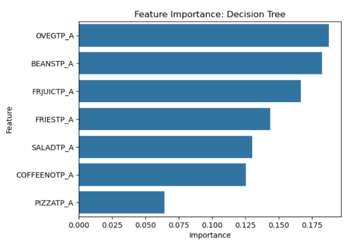


Figure 63: Decision Tree Feature Importance

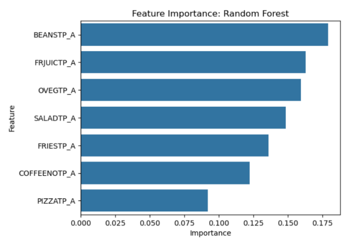


Figure 64: Random Forest Feature Importance

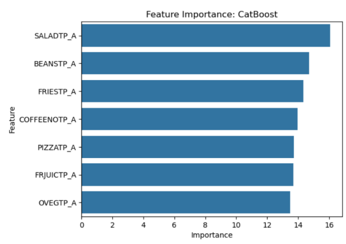


Figure 65: CatBoost Importance

3.2.1 Evaluation

We can find that the model does not perform well in predicting BMI. A big reason for this is that our code uses the most basic parameters and does not spend enough time on feature engineering and hyperparameter model training. Moreover, the dataset itself is not uniformly distributed across BMI categories, which also has a certain impact on the training of classification algorithms. If the future have more time for training, we can get a good accuracy of BMI prediction model.

4 Task 4

In this section, I will explain to you the mathematical analysis software used in this report. I analyzed the data in this report using Stata. Stata is a powerful statistical analysis software that is widely used in various research fields. For this report, Stata was selected because of its flexible data processing capabilities and rich statistical analysis capabilities, as well as its processing advantages over complex data sets. In addition, its powerful graphing capabilities help me visually describe the relationships between various variables and the trends of disease as they change. It can also export forms directly into latex format, helping me fill out the results in my report accurately and completely.

During the completion of the report, I wrote all the do scripts in the attachments folder. They all end up producing an image or table for this report. Before writing each do file script, I clarify its purpose: what I need this do file to accomplish and which data and variables it will utilize. Then, I preprocess the data and variables that will be used. The main task of preprocessing is to clean out invalid or potentially influential data that could affect subsequent results. While reviewing the codebook and dataset, I noticed that most codes provided options such as ”unknown,” ”uncertain,” or ”refused” for respondents. Such data was not helpful to our research and might have an impact on our data processing. Once, I forgot to exclude them, resulting in several extra bars in a bar chart where there should have been only two. Therefore, during preprocessing, it’s crucial to remove such data. This step ensures data quality and consistency, laying the groundwork for subsequent analysis.

In terms of defining and selecting variables, we comprehensively chose the four cardiovascular-related diseases outlined in the codebook to conduct a more thorough analysis. Additionally, for demographic variables, we selected fundamental and widely applicable ones such as age, sex, race, address, and housing type. Regarding lifestyle aspects, in Task 2, we primarily focused on dietary habits because the codebook provided detailed dietary-related data. These types of data also correlate well with demographic variables. In Task 3, we conducted a more detailed study on lifestyle aspects, further exploring the relationships between dietary habits, body mass index, and cardiovascular diseases. Thanks to the tables provided by Stata, we were able to clearly describe the probability of disease occurrence under multiple factors.

In terms of statistical methods, I used the chi-square test, logistic regression analysis, ordered logistic regression, comparative bar graphs, multiway tables, and many other useful graphical tools that stata offers. This is one of the main reasons why I chose stata as a data analysis tool, it provides statistical tools that can be extremely easy to use, such as the chi-square test, which stata encapsulates into a function that requires only a few words to use, which provides me with great convenience. The selection and application of these statistical methods can help to gain a deeper understanding of the impact of lifestyle factors on cardiovascular health and provide a scientific basis and guidance for future prevention strategies and clinical management.

In order to explore whether the variables are significantly correlated with each other, I also considered using the T test method at the very beginning. However, the T test will perform much better with normally distributed continuous variables, and performs mediocre with our discrete categorical data. So I used the chi-square test which is more suitable for our dataset to confirm whether the variables are significantly correlated with each other. In fact, in this report, I am constantly trying to determine the significant correlation between variables, both qualitatively and quantitatively. I used stata’s chi-square test. Results from logistic regression or ordinal logistic regression were used to see values within 95% confidence intervals. I trained multiple models using python to demonstrate correlation between variables by looking at feature importance. I also use correlation coefficients between variables to illustrate the degree of correlation between variables and outcomes.

In addition to using Stata for data processing and statistical analysis, I carried out further work using Python, a flexible and powerful programming language that not only excels in data cleaning and preprocessing, but also demonstrates its power in building and evaluating machine learning models. I also tried to use the dataset and the model to make predictions about disease conditions and BMI conditions. In these areas, python excelled, and its flexibility and ease of writing and modifying allowed me to get a lot of models and experimental data. As I introduced in the Task3, I use the four machine learning algorithms and models. With more time in the future, I can get better models with more practical significance through a lot of cross-validation and hyper-parameter tuning.

I also use Jupyter Notebook, which, with its interactivity, instant feedback, and powerful visualization capabilities, allows me to see the results of my data visualizations clearly and intuitively as I run the Python code, making the process of analyzing the data much more concise and easy to understand.

5 Task 5

All scripts, images, etc. will be submitted as attachments. Please note that I am using Stata MP 17, if you need to test my scripts, please use the same version number in case some modules may not work. Also all do files need to be run in the same folder as “adult22.csv” or you may need to go into the do file and change the path.

When testing the correlation between lifestyle habits and CVC in Task3, there were some conclusions I wondered if I was doing it wrong. I ran a logistic regression in python to test this as well. The results are shown in the figure below, and it can be seen that although the correlation values are different, the overall trend of positive and negative correlation is not wrong. The different values may be due to different processing of the top of the default debit, regularization, optimization of the algorithm and convergence criteria, and so on. But it confirms that my general direction is correct. The code can be found in the folder “python_version”.

```
accuracy: 0.9351341069050789
              precision    recall  f1-score   support

           0       0.94      1.00      0.97      4916
           1       1.00      0.00      0.00       341

    accuracy                           0.94      5257
   macro avg       0.97      0.50      0.48      5257
weighted avg       0.94      0.94      0.90      5257

       Feature  Coefficient
4     BEANSTP_A     0.085295
0    FRJUICTP_A     0.057465
6      OVEGTP_A    -0.060130
1  COFFEENOTP_A    -0.067461
3     FRIESTP_A    -0.095403
2     SALADTP_A    -0.139666
5     PIZZATP_A    -0.322662
```

Figure 66: Python Result

In addition, all the machine learning related code is in the ”python_version folder”. These ipynb files should be run through jupyter notebook.