Team 4: Heart Disease Health Indicators

DATS 6103: Summary Report

Professor Ning Rui

December 18, 2022

**Predicting Heart Disease Using Health Indicators**

**Introduction**

Heart Disease is one of the most common chronic diseases in the United States. Each year millions of Americans are impacted. For many, this means they incur the cost of expensive medical bills. For a smaller subsect though, it is even worse. In 2020, heart disease was the leading cause of death in the United States, claiming the lives of nearly 700,000 individuals. Because of how deadly the disease is, it is important to get out ahead of it and identify risk factors. Unfortunately, people often learn they have heart disease after a major, and potentially deadly symptoms such as chest pain, a heart attack, or cardiac arrest. As noted, there are risk factors people should consider before the point is reached where major symptoms become a problem. Some examples of risk factors include natural ageing, the building up of plaque in arteries, chronic inflammation, high blood pressure, and diabetes. Ideally individuals should assess themselves for risk factors before they suffer from one of the major symptoms.

Using the breadth of data from the Center for Disease Control’s Behavioral Risk Factor Surveillance System (BRFSS), we look at indicators of heart disease and attempt to create a model to assess an individual’s risk for heart disease. The BRFSS is a telephone survey first conducted by the CDC in 1984. It is conducted each year and over 400,000 respondents are reached each year. Respondents are asked about a range of health-related risk behaviors such as if they smoke or if they get regular exercise. They are also asked about chronic conditions such as if they have heart disease. We use data from a cleaned-up version of the 2015 survey that has 253,680 observations. Our data set has 22 variables, 21 features and the target. Most of our variables are binary, such as if an individual has high cholesterol or not. Our target variable, if an individual has heart disease or has had a heart attack, is also binary. Non-binary variables include BMI, education level, among others. The following table represents every feature in the data set.

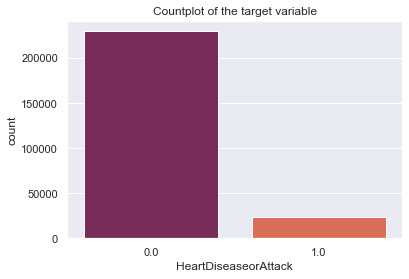
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| HighBP​​ | HighChol​​ | CholCheck​​ | BMI​​ | Smoker​​ | Stroke​​ | Diabetes​​ | PhysActivity​​ |
| Fruits​​ | **Veggies​**​ | **HvyAlcoholConsump​**​ | **Any​**​  **Healthcare​**​ | **NoDocbc​**​  **Cost​**​ | **GenHlth​**​ | **MentHlth​**​ | **PhysHlth​**​ |
| DiffWalk​​ | **Sex​**​ | **Age​**​ | **Education​**​ | **Income​**​ | **Target - HeartDiseaseorAttack​**​ | | |

**Table – 1: Dataset**

**Analysis**

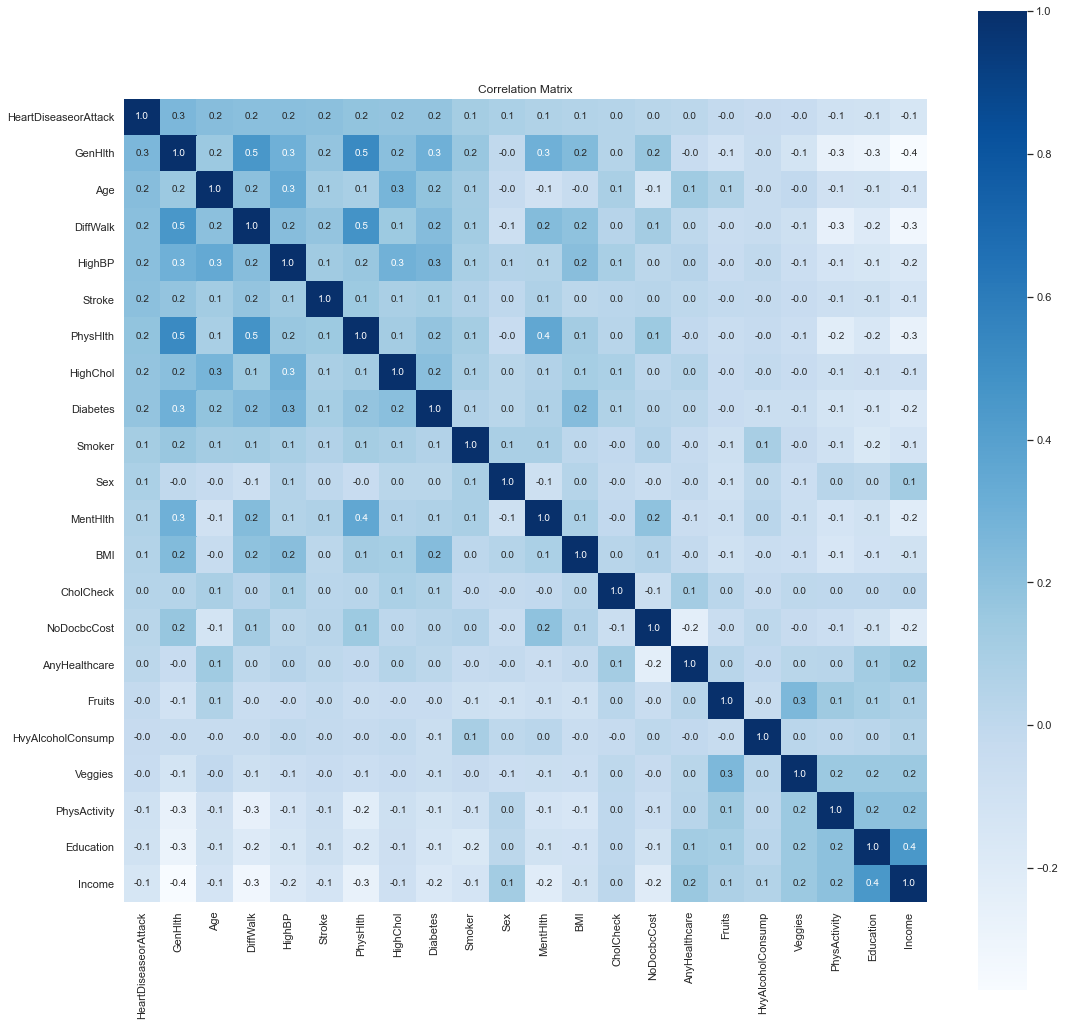
Exploratory Data Analysis

There are a total of 253,680 observations within the dataset. Out of which 23,893 people have heart disease or heart attack, and 229,787 who don’t. In terms of percentage, 90.58% no heart disease or attack while 9.42% have a heart disease or attack. The below count plot provides a pictorial representation of the above statistics.



**Fig 1 – Distribution of the target variable**

Looking at the correlation matrix to determine which features are correlated to each other, and the target variable:



**Fig – 2: Correlation Matrix**

SMART Questions

1. SMART Question 1 –

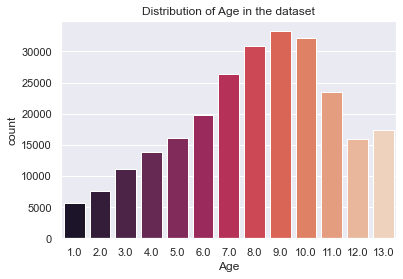
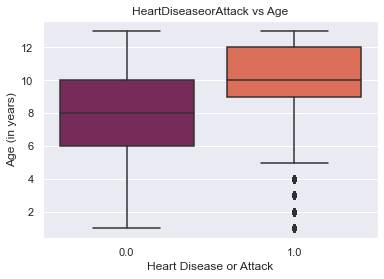
Does age influence heart disease or attack?

* Age has been divided into 12 categories in total.
* The following table provides the description of the different categories.

|  |  |  |
| --- | --- | --- |
| Category | Age interval | Number of people |
| 1.0 | 18-24 | 5,700 |
| 2.0 | 25-31 | 7,598 |
| 3.0 | 32-38 | 11,123 |
| 4.0 | 39-44 | 13,823 |
| 5.0 | 45-51 | 16,157 |
| 6.0 | 52-58 | 19,819 |
| 7.0 | 59-65 | 26,314 |
| 8.0 | 66-72 | 30,832 |
| 9.0 | 73-79 | 33,244 |
| 10.0 | 80-86 | 32,194 |
| 11.0 | 87-93 | 23,533 |
| 12.0 | Greater than 93 | 15,980 |

**Table – 2: Age categories**

* From the table, it can be observed that group 9 (age range – 73 to 79) has the largest population, followed by group 10 (age range – 80 to 86)

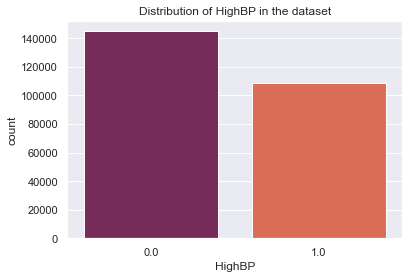
 

**Fig – 3: Distribution of Age Fig – 4: Age vs Heart disease or attack**

* According to the box plot, the median age for those without heart disease is age category 8, which is between 66 and 72 years old, while the median age for those with heart disease is age category 10, which is between 80 and 86 years old. Therefore, if a person is older than 70, their risk of developing heart disease increases.

1. SMART Question 2 – Does having high BP influence heart disease or attack?

* People with high blood pressure make up 42.9% of the total observations.
* According to the count plot, those without high blood pressure typically have a lower risk of developing heart disease, whereas those with high blood pressure have an increased risk of developing or already having the condition.



**Fig – 5: Distribution of High BP**

* From the frequency table, it can be observed that there are 17,928 out of 23,893 individuals have heart disease or attack or 75% of the total.
* From this it can be concluded that High BP is one of the major factors influencing the target variable heart disease or attack.

|  |  |  |  |
| --- | --- | --- | --- |
| HighBP | 0.0 | 1.0 | All |
| HeartDiseaseorAttack |  |  |  |
| 0.0 | 138886 | 90901 | 229787 |
| 1.0 | 5965 | 17928 | 23893 |
| All | 144851 | 108829 | 253680 |

**Table – 3: Frequency table for High BP and Heart disease or attack**

1. SMART Question 3 – How BMI influences heart disease or attack?

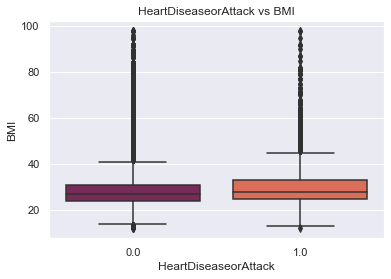
* BMI is often broken down into many ranges. They are

|  |  |
| --- | --- |
| BMI category | Range |
| Underweight | < 18.5 |
| Healthy weight | 18.5 – 24.9 |
| Overweight | 25.0 – 29.9 |
| Obese | > 30.0 |

**Table – 4: BMI categories**

* Most of the observations have BMIs greater than 25, from the histogram. People with BMI over 25 make up over 79% of those with heart disease overall.

Chart, histogram

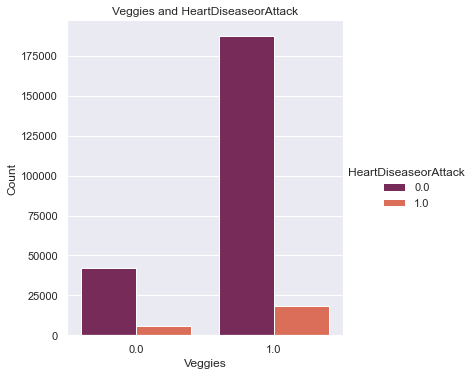
Description automatically generated 

**Fig – 6: Histogram of BMI Fig – 7: Boxplot for Heart disease or attack vs BMI**

1. SMART Question 4 – How consuming veggies and fruits can influence heart disease or attack?

* People who don’t eat fruits have a lower risk of developing heart disease or attack from the plot, whereas those who do consume fruits do seem to have a higher risk of developing a heart disease or attack.
* Similar evidence can be seen in the plot for vegetables, where it is revealed that people who eat vegetables have a higher risk of developing heart disease than those who do not.

Chart, bar chart

Description automatically generated 

**Fig – 8: Consumption of Fruits and heart disease or attack Fig – 9: Consumption of veggies and heart disease**

**or attack**

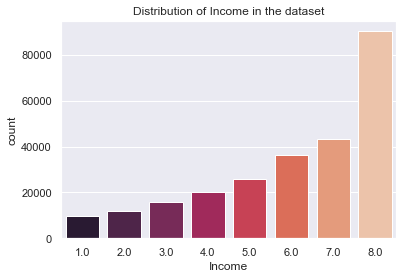
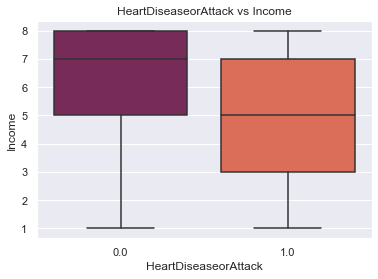
1. SMART Question 5 – Does having high cholesterol impact one having a heart disease or attack?

* The income is divided into 8 categories as follows:

|  |  |  |
| --- | --- | --- |
| Income category | Income value  ($) | Number of people in the group |
| 1.0 | < 10,000 | 9,811 |
| 2.0 | 20,000 | 11,783 |
| 3.0 | 30,000 | 15,994 |
| 4.0 | 40,000 | 20,135 |
| 5.0 | 50,000 | 25,883 |
| 6.0 | 60,000 | 36,470 |
| 7.0 | 70,000 | 43,219 |
| 8.0 | > 75,000 | 90,385 |

**Table – 5: Income categories**

* According to the correlation matrix, income and heart disease or attack has a negative correlation (score of -0.1).
* From the boxplot, it is evident that people with no heart disease or attack have a higher income than people who do have a heart disease or attack.

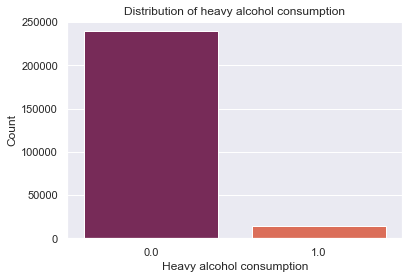
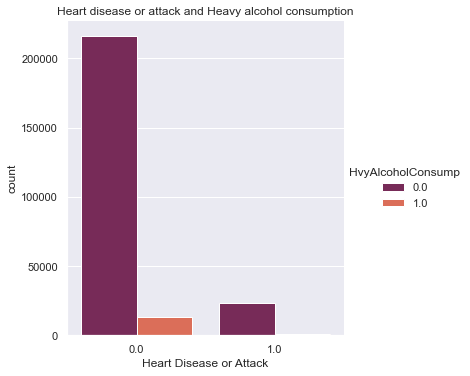
 

**Fig – 10: Distribution of income Fig – 11: Boxplot for heart disease or attack**

**vs income**

1. SMART Question 6 – Does heavy alcohol consumption influence heart disease or attack?

* From the correlation matrix, there is no correlation between heavy alcohol consumption and heart disease or attack.
* The count plot provides information that there are 239,424 people who do not consume alcohol heavily, and 14,256 who do consume alcohol heavily.
* Between heart disease or attack and heavy alcohol consumption, there are very few people who have experienced heart disease or attack due to heavy alcohol consumption.

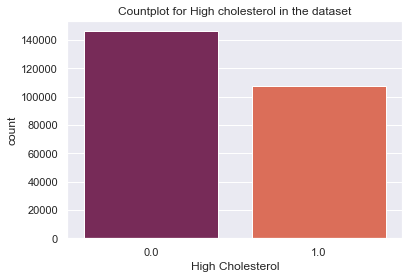
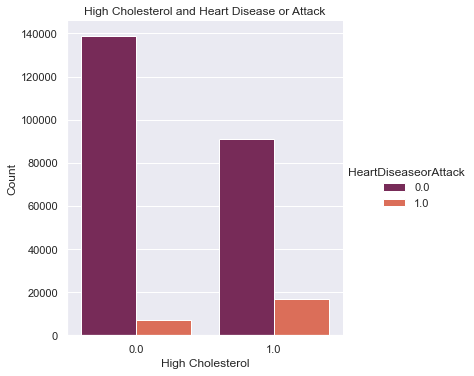
 

**Fig – 12: Distribution of alcohol consumption Fig – 13: Heart disease or attack and alcohol**

**consumption**

1. SMART Question 7 – Does having high cholesterol impact one having heart disease or attack?

* The correlation matrix informs that the correlation is 0.2 between high cholesterol and heart disease or attack, which states that they have a positive correlation between them.
* There are a total of 146,089 people within the dataset who do not have high cholesterol, while there are 107,591 who do have high cholesterol.

**Fig – 14: Distribution of high cholesterol Fig – 15: High cholesterol and heart disease or**

**Attack**

* Between high cholesterol and heart disease or attack, it can be observed that people who do not have high cholesterol have less chances of having a heart disease or attack, while the opposite is noted when there is high cholesterol and chance of getting a heart disease or attack.

**Statistical Testing**

To further examine the association between our feature variables and target variable, t-tests were performed. The following table has the t statistic for the t-test between every feature and the target variable.

|  |  |
| --- | --- |
| **Feature** | **T Statistic** |
| Age | -114.4673 |
| CholCheck | -22.2867 |
| Stroke | -104.4191 |
| NoDocbcCost | -15.6210 |
| GenHlth | -134.7131 |
| AnyHealthcare | -9.4374 |
| MentHlth | -32.6156 |
| PhysHlth | -93.0637 |
| DiffWalk | -109.6429 |
| Education | 50.4156 |
| Sex | -43.5249 |
| HvyAlcoholConsump | 14.6076 |
| Diabetes | -91.9796 |
| Smoker | -58.0212 |
| HighChol | -92.5701 |
| PhysActivity | 44.1379 |
| Veggies | 19.7424 |
| Fruits | 9.9696 |
| BMI | -26.6834 |
| Income | 71.7392 |
| HighBP | -107.8377 |

From the table we can see that the t statistics for every feature have a high absolute value. Given that the target variable is binary, the large t statistics imply that there is a difference in the proportion of each group (divided by their status for a given feature variable) that has heart disease or attack. It is surprising that every variable has a large t statistic. A possible reason for this is that our data set is so large that the standard error for each mean is exceptionally small as a result. This reduces the confidence interval for each mean and thus leads to the difference in means being significant.

**Modeling**

Feature Selection:

* Given that the dataset solely contains numerical values, a VIF test was conducted to identify the variables that are multicollinear so that they can be omitted before creating the models.
* After exploring various combination, the final models used the following features, which had a VIF score below 10.

|  |  |  |
| --- | --- | --- |
|  | Features | VIF score |
| 1. | HighBP | 2.376639 |
| 2. | HighChol | 2.075802 |
| 3. | Smoker | 1.983754 |
| 4. | Stroke | 1.108968 |
| 5. | Diabetes | 1.419802 |
| 6. | PhysActivity | 4.144036 |
| 7. | Fruits | 2.828395 |
| 8. | Veggies | 5.287618 |
| 9. | HvyAlcoholConsump | 1.092803 |
| 10. | NoDocbcCost | 1.219131 |
| 11. | MentHlth | 1.469514 |
| 12. | PhysHlth | 2.009715 |
| 13. | DiffWalk | 1.847508 |
| 14. | Sex | 1.888215 |

**Table – 6: VIF table**

Modeling on the imbalanced dataset:

* The goal was to determine what features caused a person to have a heart disease or attack.
* As the dataset is heavily imbalanced, the models were performed well when it came to predicting no heart disease or attack.
* The following table shows how the models performed on the imbalanced test dataset.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name of the model | Recall | | Accuracy | AUC |
| **0 – no heart disease or attack** | **1 – yes heart disease or attack** |
| Naïve Bayes | 0.85 | 0.55 | 0.81 | 0.69 |
| Logistic Regression | 0.99 | 0.11 | 0.89 | 0.55 |
| Decision Tree Classifier | 0.99 | 0.08 | 0.89 | 0.53 |
| Random Forest Classifier | 1.00 | 0.39 | 0.90 | 0.53 |
| XGBoost Classifier | 0.99 | 0.18 | 0.89 | 0.55 |

**Table – 7: Modeling results on imbalanced dataset**

* 1. Looking at the table, it is clear that Random Forest Classifier performs the best when compared with other models with respect to recall. But there is a high possibility that the model is being over-fit. The reason for this is due to the heavy imbalance within the dataset.
  2. Focusing on recall is important since the model should be able to accurately predict true positives while minimizing false negatives.

Balancing Technique:

* SMOTE or Synthetic Minority Oversampling Technique is an improved alternative for oversampling.
* By generating artificial data points based on the real data points, the SMOTE algorithm conducts data augmentation.
* The advantage of SMOTE is that, it is not generating duplicates, but rather creating synthetic data points that are slightly different from the original data points.

The SMOTE algorithm works as follows:

1. Draw a random sample from the minority class.
2. For the observations in the sample, identify the k-nearest neighbors.
3. Consider one of those neighbors and identify the vector beterrn the current data point and the selected neighbor.
4. Multiply the vector by a random number between 0 and 1.
5. To obtain the synthetic data point, add the multiplied vector to the current data point.

* SMOTE attempts to balance the dataset into a 50:50 ratio because the existing dataset is unbalanced, with yes heart disease or attack having 23717 datapoints and no heart disease or attack having 206064 datapoints.
* After balancing there 108186 data points present in both yes heart disease or no heart disease or attack.
* Balancing the data also helps in reducing bias in the dataset. Bias and variance were estimated using mlxtend package and the method bias\_variance\_decomp which showed a difference in bias before and after balancing the data using SMOTE.

Modeling on balanced dataset:

* The objective was to identify the features that contributed to heart disease or attack in an individual.
* The following are the results of all the models on the balanced dataset.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name of the model | Recall | | Accuracy | AUC |
| **0 – no heart disease or attack** | **1 – yes heart disease or attack** |
| Naïve Bayes | 0.78 | 0.61 | 0.76 | 0.69 |
| Logistic Regression | 0.73 | 0.71 | 0.72 | 0.71 |
| Decision Tree Classifier | 0.77 | 0.56 | 0.74 | 0.66 |
| Random Forest Classifier | 0.70 | 0.72 | 0.70 | 0.71 |
| XGBoost Classifier | 0.85 | 0.43 | 0.80 | 0.63 |

**Table – 8: Modeling results on balanced dataset**

1. From the above table, it can be concluded that Random Forest Classifier performs better for both yes heart disease or attack and no heart disease or attack.

**Conclusion**

* With the help of Variance Inflation Factor (VIF) the features which showed heavy multicollinearity which helped with the accuracy among both imbalanced and balanced models.
* With the imbalanced dataset it was easy to understand that most of the predictions with no heart disease or attack would be correct, but the model did try to predict yes heart disease or attack as well to a certain length.
* Recall is what the model is more focused towards as recall is used to predict true positives while minimizing false negatives accurately.
* Logistic Regression, Decision Tree Classifier, Random Forest Classifier, and XGBoost Classifier performed the best when it came to no heart disease or attack, while Naïve Bayes performed best for yes heart disease or attack.
* SMOTE was used as a balancing technique to balance the dataset. But the disadvantage of SMOTE is it does not take into consideration neighboring examples that can be from other classes. This can increase the overlapping of classes and can introduce additional noise.
* Random Forest Classifier performs best on balanced dataset among all the other models with respect to recall alone.
* Bias and variance is something to be focused on because to build a good model, there should be a good balance between bias and variance such that it minimizes the total error.

**References**

1. How to balance a dataset in Python: <https://towardsdatascience.com/how-to-balance-a-dataset-in-python-36dff9d12704>
2. How to deal with imbalanced data in Python: <https://towardsdatascience.com/how-to-deal-with-imbalanced-data-in-python-f9b71aba53eb>
3. Original cleaning of the dataset after the survey: <https://www.kaggle.com/code/alexteboul/heart-disease-health-indicators-dataset-notebook>
4. SMOTE for Imbalanced Classification with Python: <https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/>
5. Ways to Detect and Remove the Outliers: <https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-404d16608dba>
6. Understanding AUC-ROC Curve: <https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5>
7. Understanding the ROC Curve and AUC: <https://towardsdatascience.com/understanding-the-roc-curve-and-auc-dd4f9a192ecb>
8. SMOTE - <https://towardsdatascience.com/smote-fdce2f605729>
9. SMOTE: Synthetic Minority Over Sampling Technique (Paper): <https://www.jair.org/index.php/jair/article/view/10302>