M259: Machine Learning

*Personal Key Indicator of Heart Disease*



Sara C. Ramirez

30/09/22

INF21s

Inhaltsverzeichnis

[Business Undestanding 3](#_Toc115447701)

[Data Understanding 3](#_Toc115447702)

[EDA durchführen 4](#_Toc115447703)

[Daten importieren 4](#_Toc115447704)

[Analyse 4](#_Toc115447705)

[Datensäuberung 5](#_Toc115447706)

[Datenanalyse von einer Variablen 5](#_Toc115447707)

[Datenanalyse von zwei Variablen 5](#_Toc115447708)

[Hypothesen 7](#_Toc115447709)

[Hypothesenuntersuchung 7](#_Toc115447710)

[Data Preparation 11](#_Toc115447711)

[Modeling 11](#_Toc115447712)

[Zusammenfassung 14](#_Toc115447713)

# Business Undestanding

Mit diesem Projekt sollte man erkennen, welche Faktoren Herzerkrankungen beeinflussen und ob Menschen mit Herzerkrankungen gewisse Charakteristiken teilen.

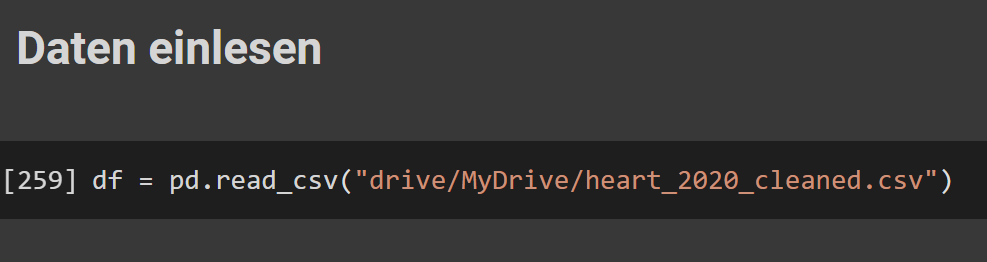
# Data Understanding

Im Datenset « Personal Key Indicators of Heart Disease”, auf Deutsch “Persönliche Indikatoren für Herzerkrankungen» befinden sich 18 Spalten.

|  |  |  |
| --- | --- | --- |
| Name | Beschreibung | Datentyp |
| HeartDisease | Zeigt an, ob Person eine Herzkrankheit hat. | Object |
| BMI | Body Mass Index der Person | Float64 |
| Smoking | Ob Person mehr als 100 Zigaretten geraucht hat. | Object |
| AlcoholDrinking | Ob Person 7 bzw. 14 alkoholische Getränke pro Woche trinkt. | Object |
| Stroke | Ob man schon mal einen Schlaganfall gehabt hat. | Object |
| PhysicalHealth | Anzahl Tage, die man sich physisch schlecht fühlte in den letzten 30 Tagen. | Float64 |
| MentalHealth | Anzahl Tage, die man sich psychisch schlecht fühlte in den letzten 30 Tagen. | Float64 |
| DiffWalking | Schwierigkeiten im Gehen. | object |
| Sex | Männliches oder weibliches Geschlecht. | object |
| AgeCategory | 13 verschiedene Altersklassen. | object |
| Race | Rassen bzw. Ethnie Kategorien | object |
| Diabetic | Ob man Diabetes hat. | object |
| PhysicalActivity | Ob man in den letzten 30 Tagen Sport gemacht hat. | object |
| GenHealth | Allgemeines Wohlbefinden | object |
| SleepTime | Wie viele Stunden Schlaf man durchschnittlich hat. | Float64 |
| Asthma | Zeigt an, ob man Asthma hat. | object |
| KidneyDisease | Zeigt an, ob man eine Nierenkrankheit hat. | object |
| SkinCancer | Zeigt an, ob man Hautkrebs hat. | object |

# EDA durchführen

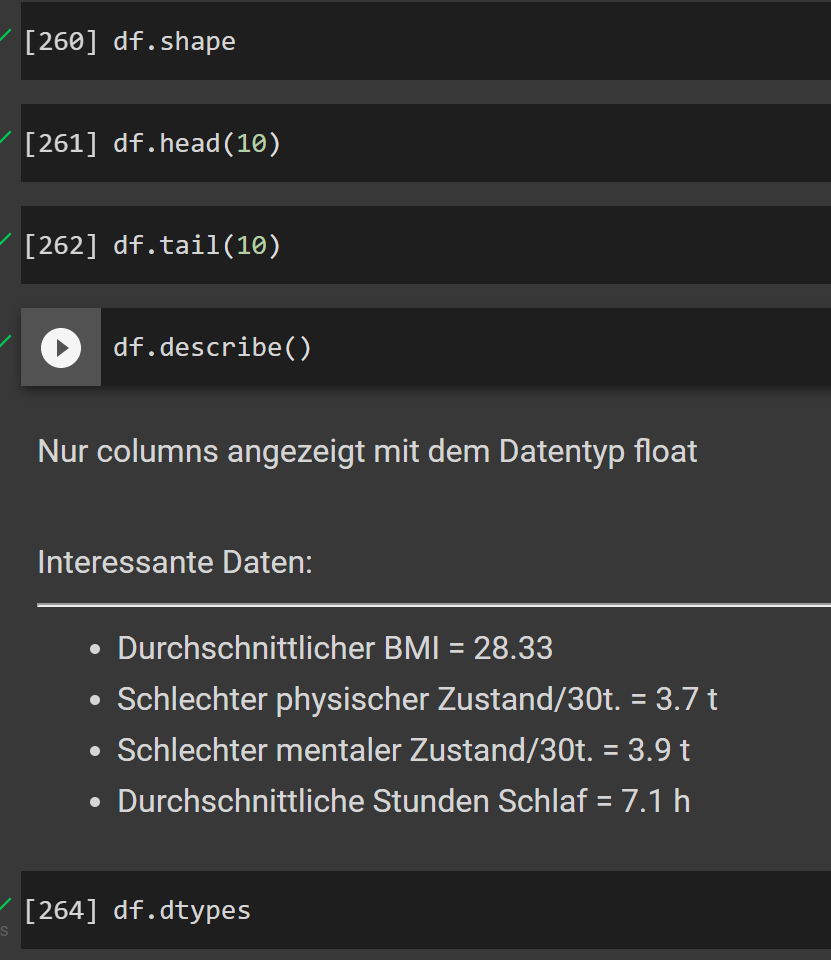
## Daten importieren



Es wurde ein CSV-File genutzt von «kaggle».

## Analyse

Das Datenset hat 18 Columns und 319795 Rows.

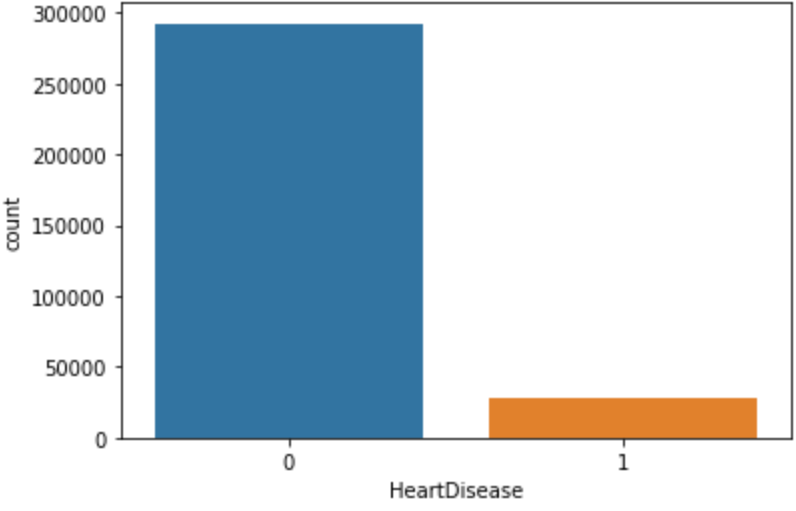


## Datensäuberung

Auffällig, ist dass es viele «object» Datentypen gibt, die man säubern muss, um sie brauchbar zu machen.

## Datenanalyse von einer Variablen

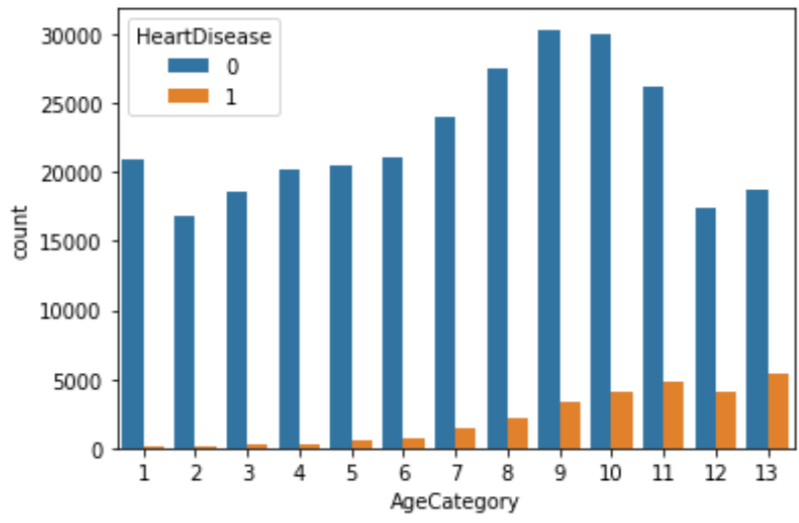
Die meisten haben keine Herzerkrankung.



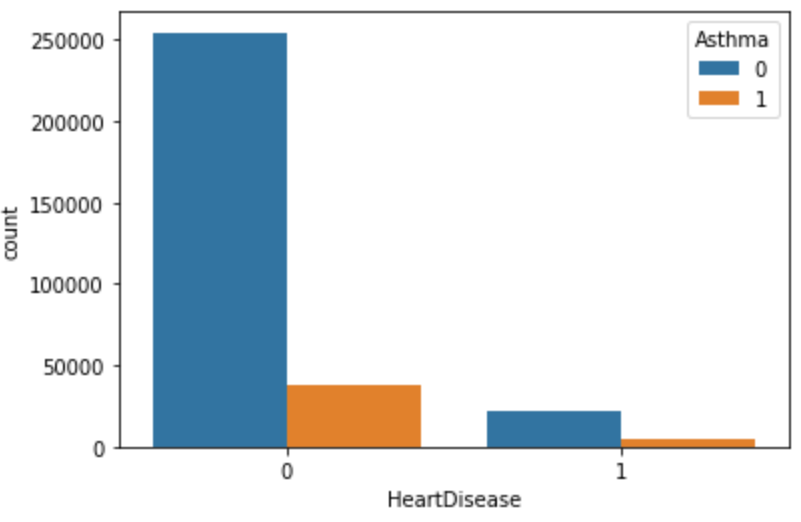
## Datenanalyse von zwei Variablen

Vergleich HeartDisease – AgeCategory

Auffällig ist, dass je älter die Person ist, desto mehr haben eine Herzerkrankung. Obwohl es wenige Kandidaten gibt bei den Altersklassen 12 und 13, fallt deutlich auf, dass die Anzahl Erkrankten höher ist als bei den anderen Altersklassen.



Vergleich HeartDisease – Asthma



Asthma ist kein grosser Faktor.

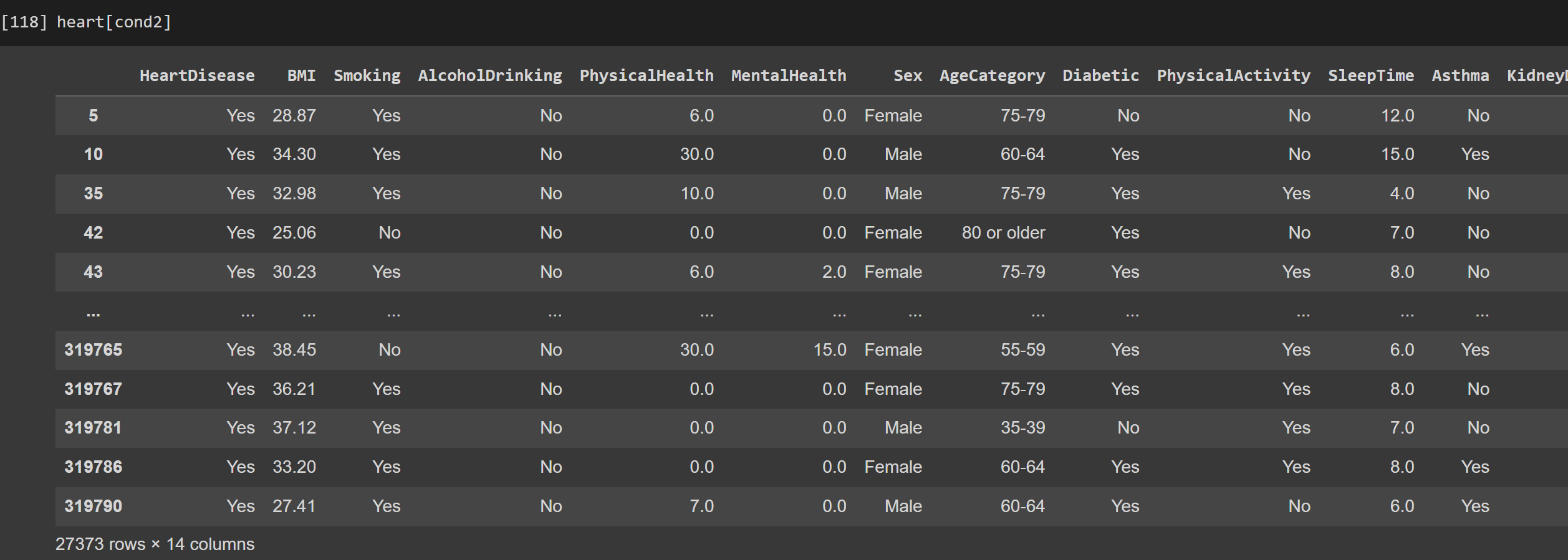
## Hypothesen

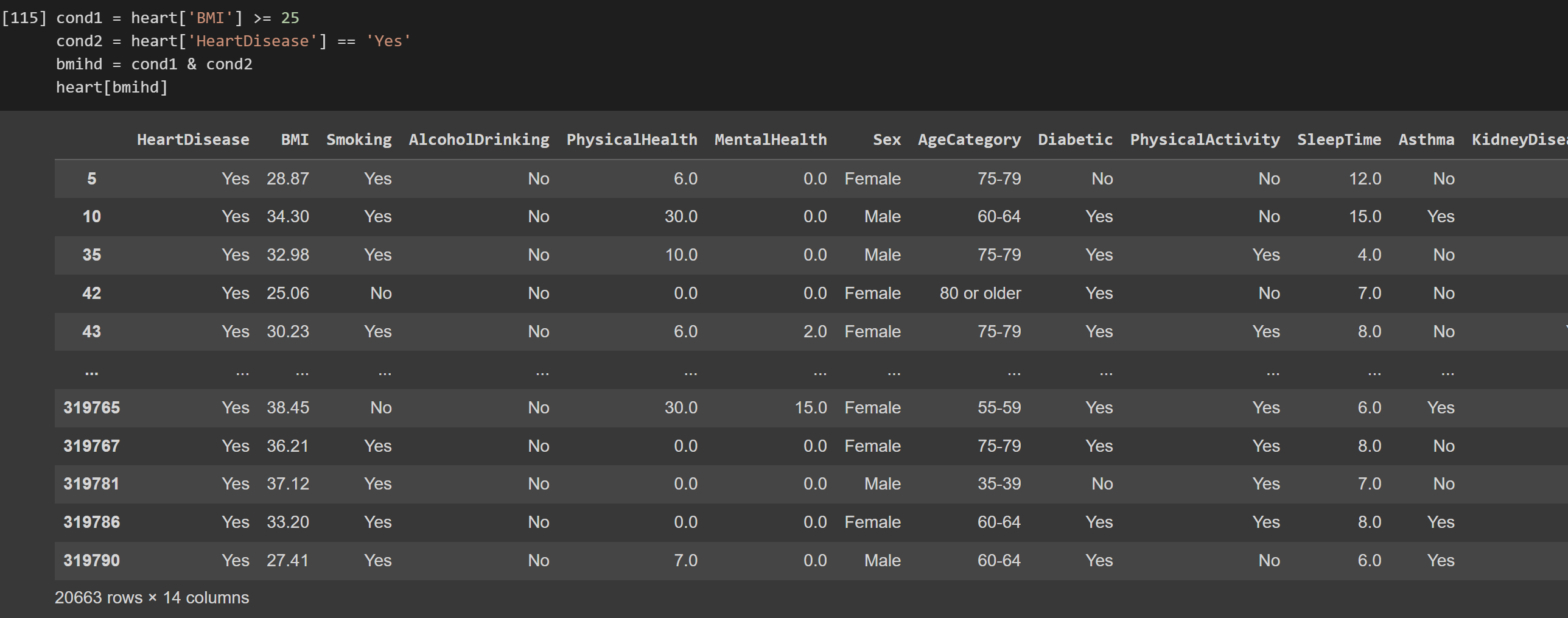
1. Personen mit einem BMI über 25 (gilt als Übergewicht) haben eine Tendenz dazu eher eine Herzerkrankung zu haben.
2. Es wird angenommen, dass Menschen, die zur älteren Alterskategorien angehören, ein höheres Risiko zu Herzerkrankungen haben.
3. Bei Menschen, die rauchen und/oder trinken besteht die Möglichkeit, dass sie im Verlaufe ihres Lebens eine Erkrankung haben, hier: eine Herzerkrankung.
4. Männer haben eine Tendenz eher erkrankt zu sein.

## Hypothesenuntersuchung

Hypothese 1:

20'663 von den 27'373 Herzerkrankten haben einen BMI höher als 25. Das heisst 75,49% von den Herzerkrankten haben ein Übergewicht.

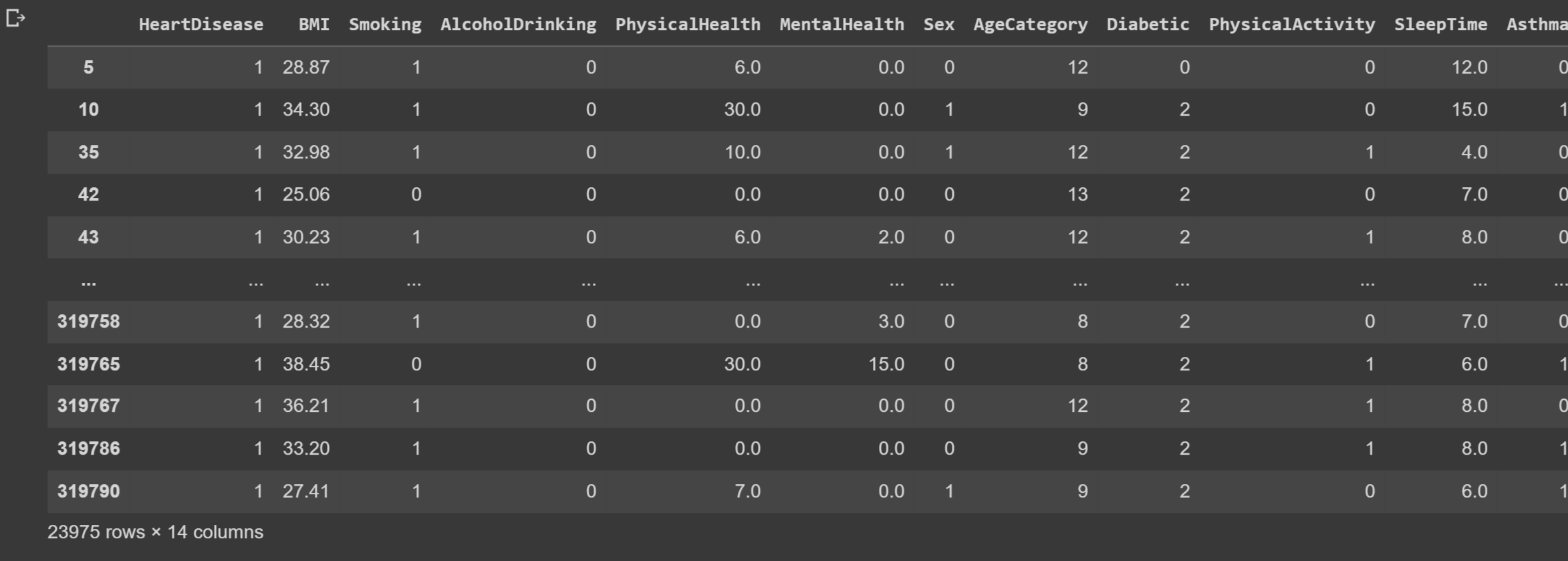




Hypothese 2:

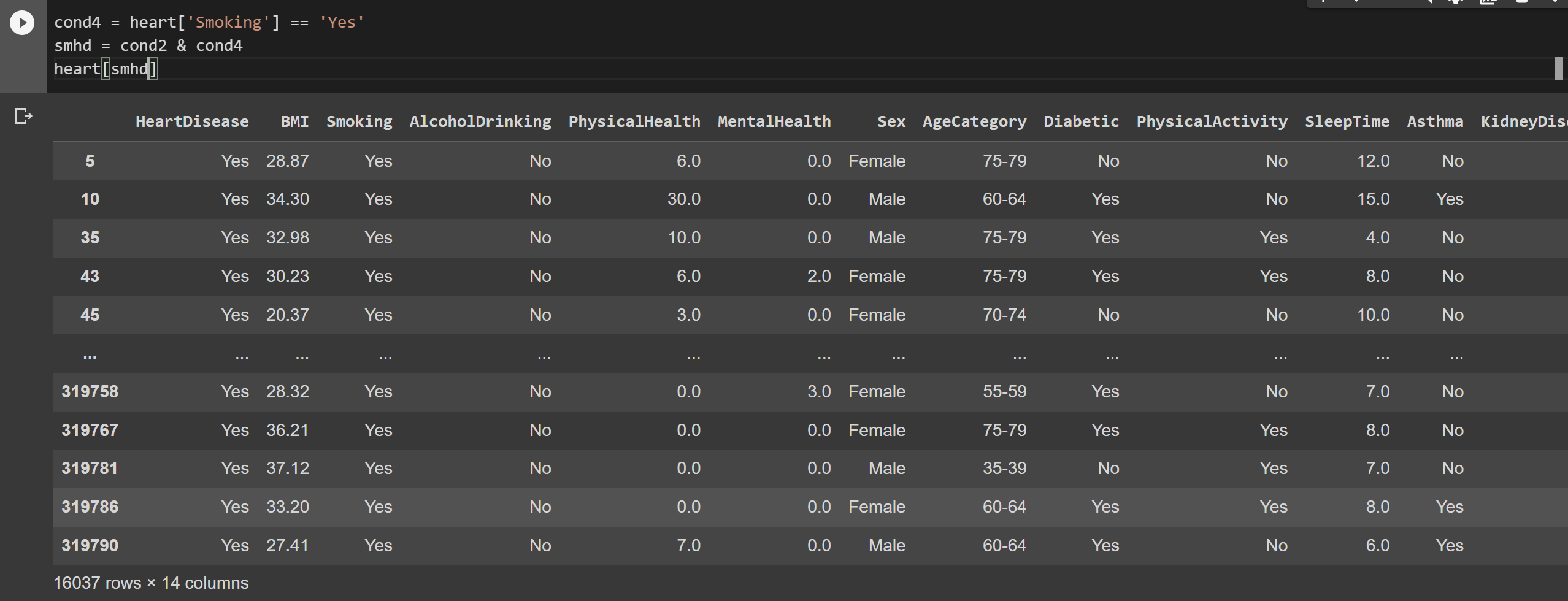
Diese wurde bereits bei der Datenanalyse bewiesen.

Hier sieht man, dass 23975 von den 27373 Herzerkrankten älter als 60 sind.

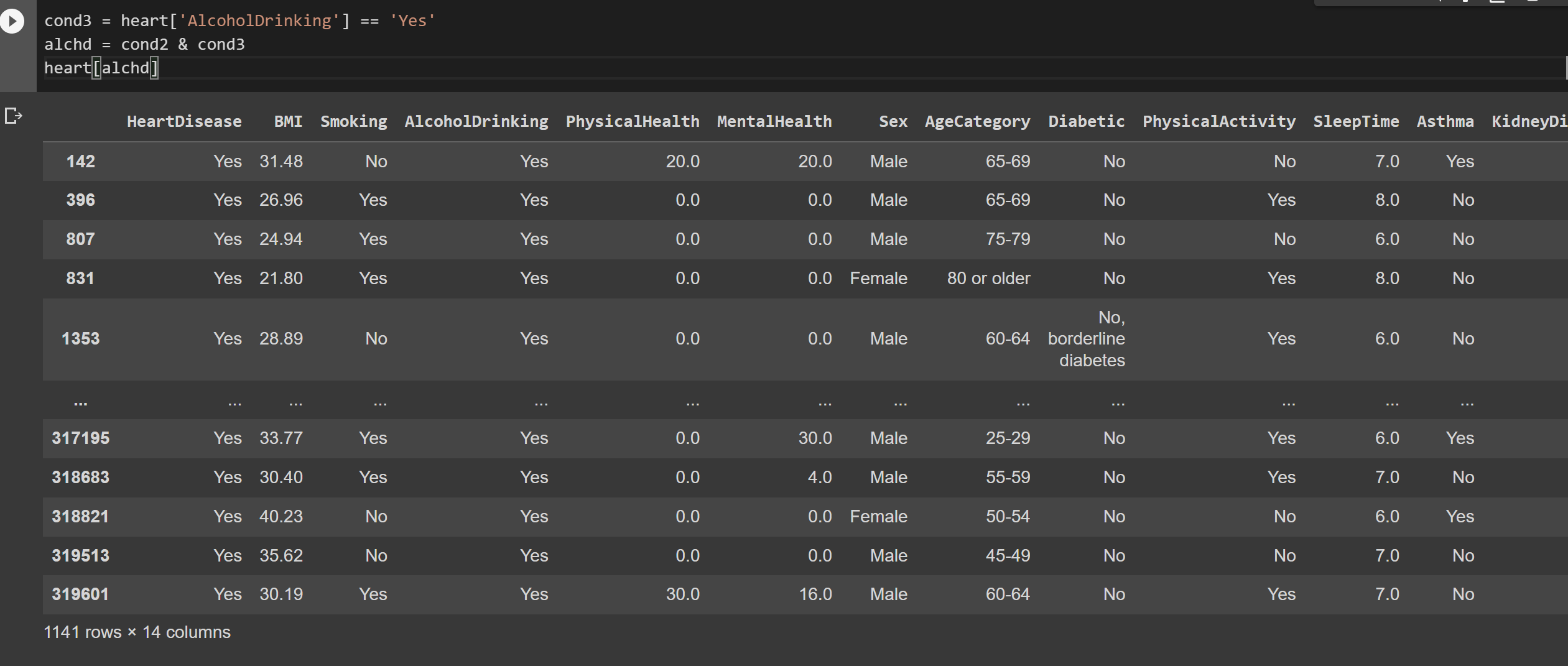


Hypothese 3:

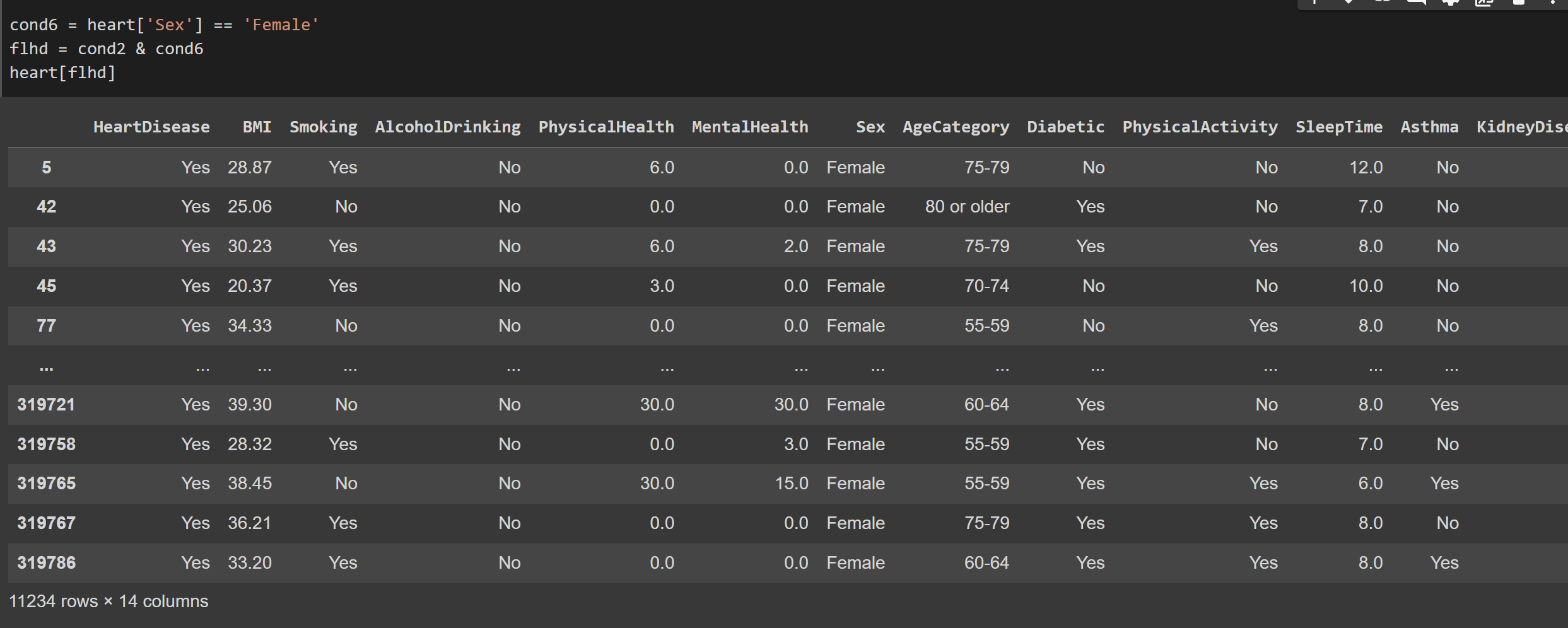
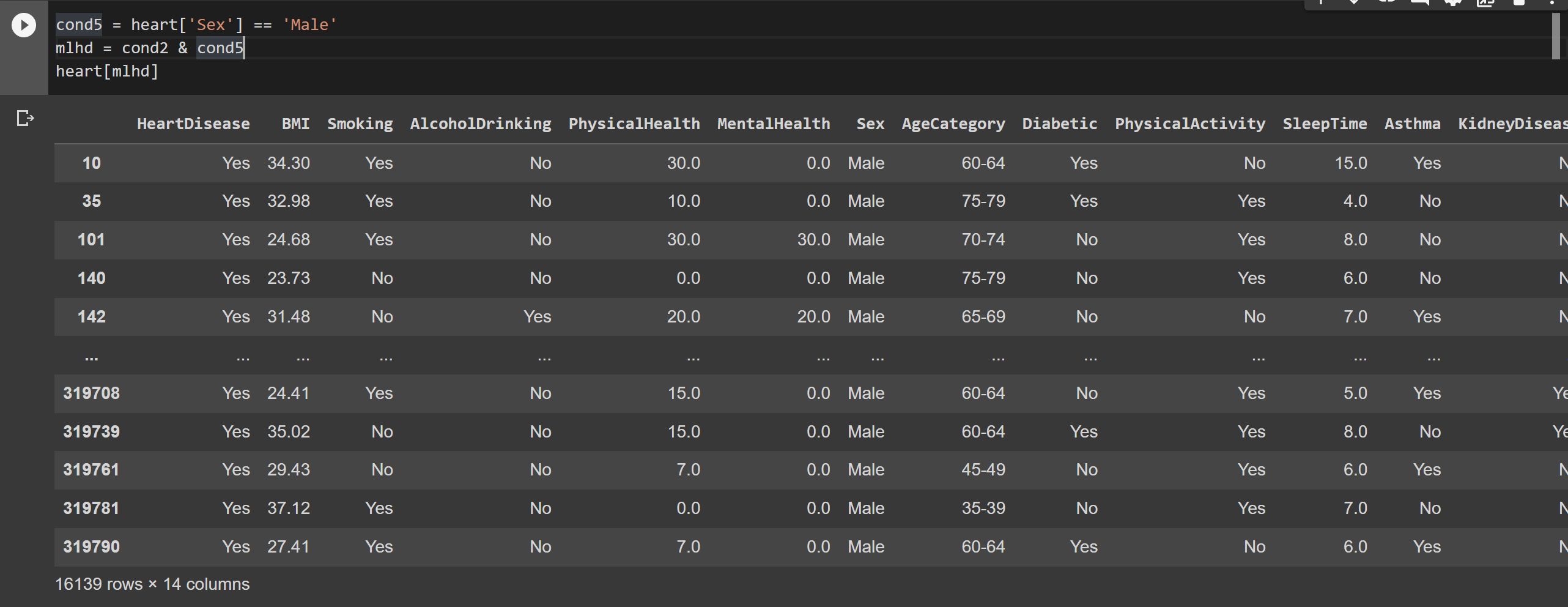
58,72 % der Herzerkrankten sind auch Rauchen oder haben mindestens 100 Zigaretten geraucht.

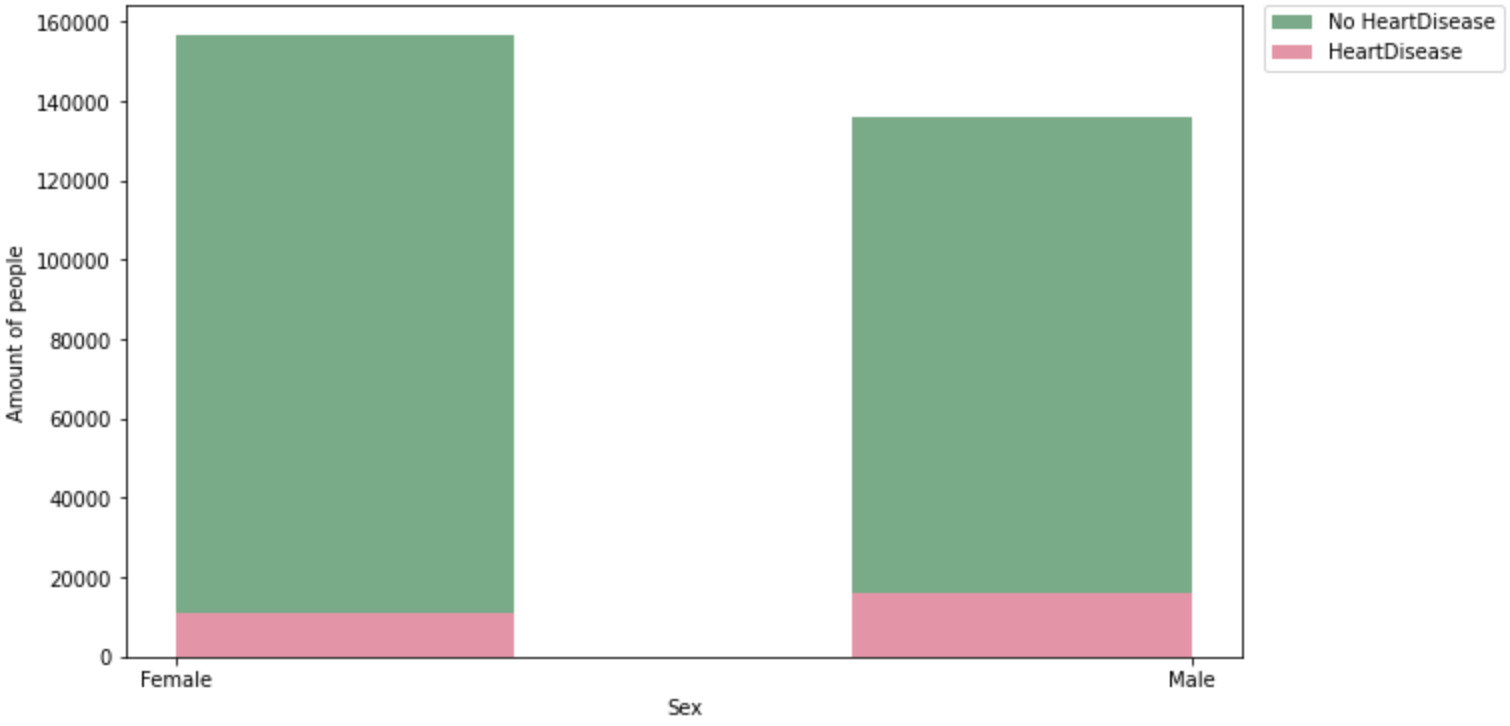


Jedoch nur 4,71% trinken mehr als 7 bzw. 14 alkoholische Getränke pro Woche.



Hypothese 4:





Männer tendieren dazu eher eine Herzerkrankung zu haben, jedoch ist es nicht so ausschlaggebend.

# Data Preparation

Viele Columns mussten bearbeitet werden. Bei allen columns, die man den Index «Yes» oder «No» musste man diese zu 0 oder 1 encoden.

Bei AgeCategory gibt es 13 Altersklassen, um es zu vereinfachen, wurden sie mit 1-13 angeschrieben. 1 gilt als 18-24 und 13 als 80+. Das heisst, desto kleiner die Zahl, je jünger ist die Person.

Einige Columns wurden gedroppt. ‘Race’ wurde entfernt, da die befragten Personen überwiegend ‘White’ waren und es keine korrekte Repräsentation darstellt.

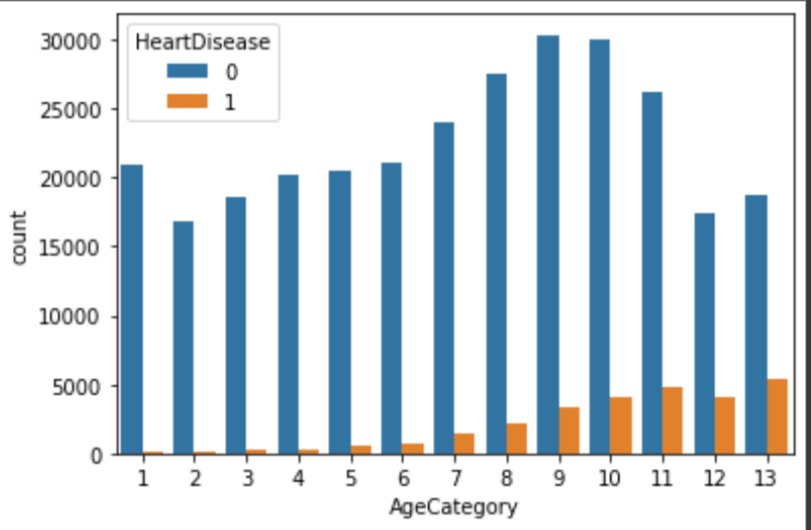
‘GenHealth’ hat man mit PhysicalHealth und MentalHealth inbegriffen.

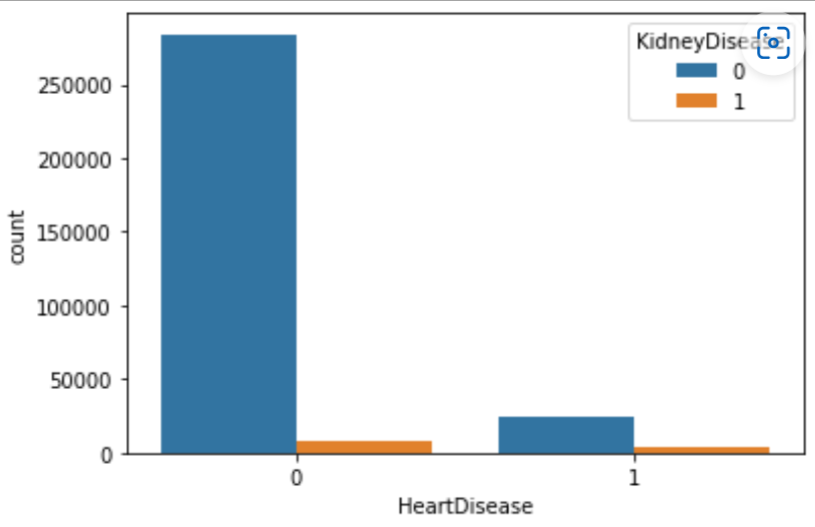
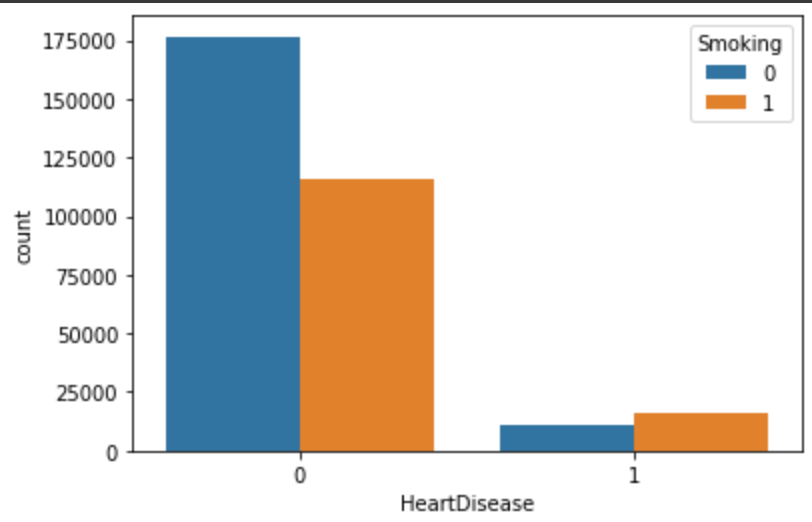
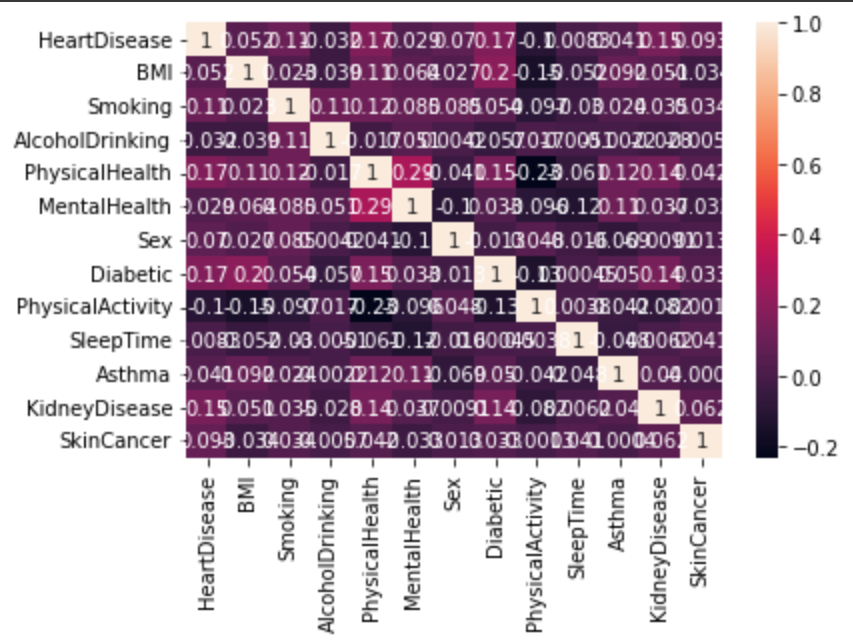
‘Stroke’, Schlaganfälle kann aus mehreren Gründen passieren und ist hier überflüssig.

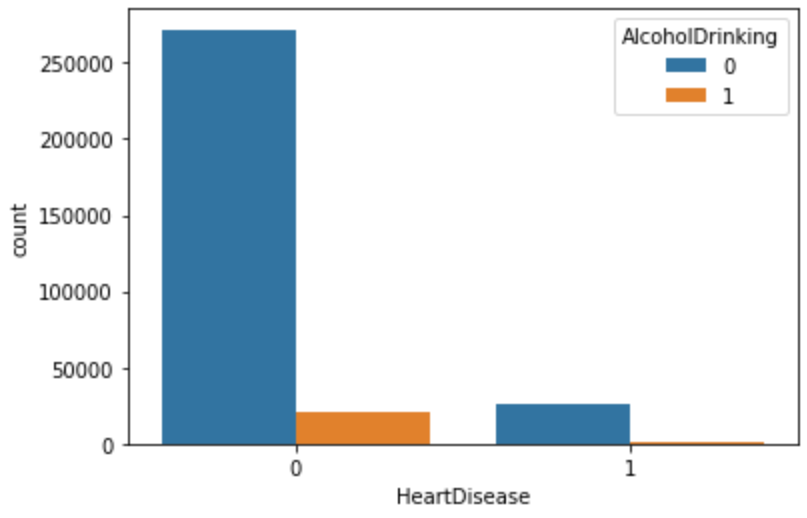
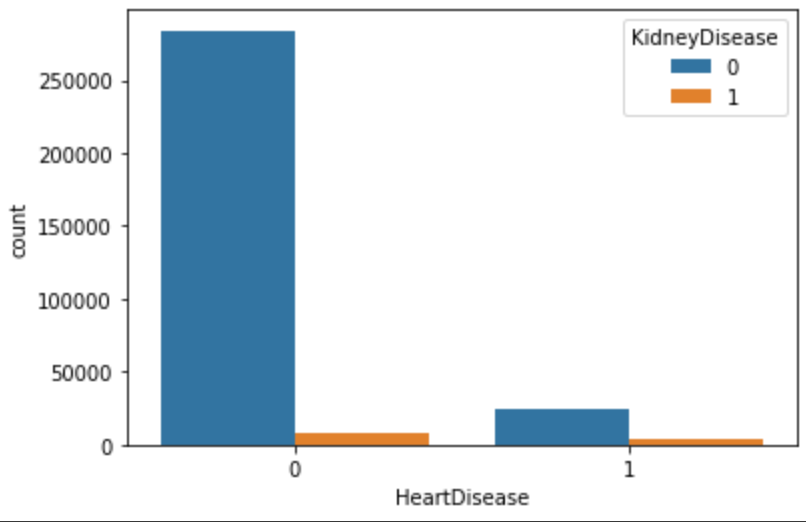
‘DiffWalking’ wurde auch gedroppt, da auch viele ältere Personen befragt wurden. Schwierigkeiten mit gehen und laufen trifft oft auch einfach mit dem Alter ein.

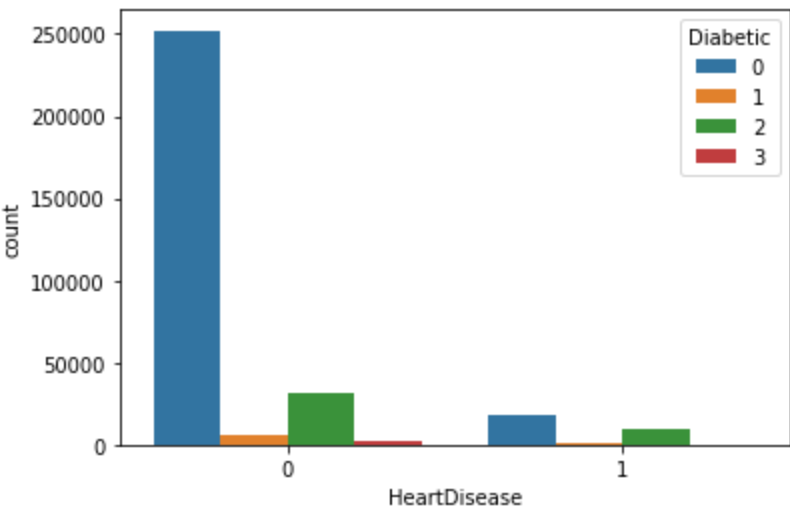
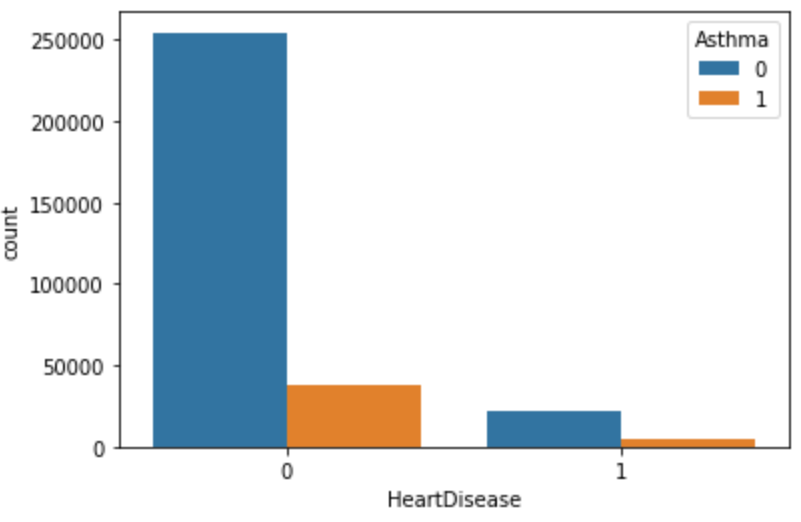
Duplikate wurden gedroppt und leere Datensätze gab es keine.

# Modeling







# Zusammenfassung

Beim Projekt wurde einen kaggle Dataset benutzt. Viel Zeit wurde genutzt, um die Daten genauer zu analysieren und zu säubern. Herausforderungen sind aufgetreten, bei dem Encoding, da viele oder die meisten columns geändert werden mussten.

Hier wurde eine Klassifizierung genutzt.

Das DecisionTreeClasifier war zu 86% genau. Beim LogisticRegression war es jedoch bis zu 91% genau.

LinearDiscriminantAnalysis und KNeighborsClassifier waren 90% genau.

GaussianNB war etwa 85% genau.

Das grösste Faktor ist Altersklasse und das Rauchen

Probleme der Modelle:

1. Verteilung der Altersklassen; man müsste gleich viel Personen für jede Altersklasse nutzen, um es noch genauer analysieren zu können.
2. Race; Ist genau wie bei Problem 1 der Fall. Auch ist es fraglich, ob es überhaupt etwas beeinflusst.
3. Sex; Wiederrum wie bei Problem 1 der Fall, interessant wäre zu wissen, wieso es einen Geschlechterunterschied gibt.