Case Study: data in UK

Yue Han

2023-05-10

## Purpose

Determine what characteristics (variables) make someone more likely to smoke.

## Dataset

UK smoking dataset retrieved from [Kaggle](https://www.kaggle.com/datasets/utkarshx27/smoking-dataset-from-uk) original [source](https://www.stem.org.uk/resources/elibrary/resource/28452/large-datasets-stats4schools) with 1691 observations and 12 variables.

### Dataset Reliability

This dataset has been reviewed and been deemed factually accurate by the source’s learning team.

## Setup and Data Cleaning

Preparing packages used.

library(tidyverse)  
library(ggplot2)  
library(dplyr)  
library(ggcorrplot)

Importing dataset.

data = read.csv("smoking.csv")  
head(data)

## X gender age marital\_status highest\_qualification nationality ethnicity  
## 1 1 Male 38 Divorced No Qualification British White  
## 2 2 Female 42 Single No Qualification British White  
## 3 3 Male 40 Married Degree English White  
## 4 4 Female 40 Married Degree English White  
## 5 5 Female 39 Married GCSE/O Level British White  
## 6 6 Female 37 Married GCSE/O Level British White  
## gross\_income region smoke amt\_weekends amt\_weekdays type  
## 1 2,600 to 5,200 The North No NA NA   
## 2 Under 2,600 The North Yes 12 12 Packets  
## 3 28,600 to 36,400 The North No NA NA   
## 4 10,400 to 15,600 The North No NA NA   
## 5 2,600 to 5,200 The North No NA NA   
## 6 15,600 to 20,800 The North No NA NA

There are 1691 rows and 13 columns.

str(data)

## 'data.frame': 1691 obs. of 13 variables:  
## $ X : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ gender : chr "Male" "Female" "Male" "Female" ...  
## $ age : int 38 42 40 40 39 37 53 44 40 41 ...  
## $ marital\_status : chr "Divorced" "Single" "Married" "Married" ...  
## $ highest\_qualification: chr "No Qualification" "No Qualification" "Degree" "Degree" ...  
## $ nationality : chr "British" "British" "English" "English" ...  
## $ ethnicity : chr "White" "White" "White" "White" ...  
## $ gross\_income : chr "2,600 to 5,200" "Under 2,600" "28,600 to 36,400" "10,400 to 15,600" ...  
## $ region : chr "The North" "The North" "The North" "The North" ...  
## $ smoke : chr "No" "Yes" "No" "No" ...  
## $ amt\_weekends : int NA 12 NA NA NA NA 6 NA 8 15 ...  
## $ amt\_weekdays : int NA 12 NA NA NA NA 6 NA 8 12 ...  
## $ type : chr "" "Packets" "" "" ...

Additionally the first column called X is not useful for analysis since it is just the number of the row. We will also change the gender, marital\_status, highest\_qualification, nationality, ethnicity, gross\_income, region, smoke, and type columns into factors as they are categorical data.

data <- data[2:13]  
data$gender <- as.factor(data$gender)  
data$marital\_status <- as.factor(data$marital\_status)  
data$highest\_qualification <- as.factor(data$highest\_qualification)  
data$nationality <- as.factor(data$nationality)  
data$ethnicity <- as.factor(data$ethnicity)  
data$gross\_income <- as.factor(data$gross\_income)  
data$region <- as.factor(data$region)  
data$smoke <- as.factor(data$smoke)  
data$type <- as.factor(data$type)  
head(data)

## gender age marital\_status highest\_qualification nationality ethnicity  
## 1 Male 38 Divorced No Qualification British White  
## 2 Female 42 Single No Qualification British White  
## 3 Male 40 Married Degree English White  
## 4 Female 40 Married Degree English White  
## 5 Female 39 Married GCSE/O Level British White  
## 6 Female 37 Married GCSE/O Level British White  
## gross\_income region smoke amt\_weekends amt\_weekdays type  
## 1 2,600 to 5,200 The North No NA NA   
## 2 Under 2,600 The North Yes 12 12 Packets  
## 3 28,600 to 36,400 The North No NA NA   
## 4 10,400 to 15,600 The North No NA NA   
## 5 2,600 to 5,200 The North No NA NA   
## 6 15,600 to 20,800 The North No NA NA

Here we have Refused and Unknown values. Instead of having both of these categories we will combine the two together. By checking again we see we have successfully combined the two.

data$nationality[data$nationality == "Refused" ] <- "Unknown"  
data %>% count(nationality) %>% rename("amount"="n")

## nationality amount  
## 1 British 538  
## 2 English 833  
## 3 Irish 23  
## 4 Other 71  
## 5 Scottish 142  
## 6 Unknown 18  
## 7 Welsh 66

Similarly, we do this for gross\_income. By checking again we see we have successfully combined the two.

data$gross\_income[data$gross\_income == "Refused" ] <- "Unknown"  
data %>% count(gross\_income) %>% rename("amount"="n")

## gross\_income amount  
## 1 10,400 to 15,600 268  
## 2 15,600 to 20,800 188  
## 3 2,600 to 5,200 257  
## 4 20,800 to 28,600 155  
## 5 28,600 to 36,400 79  
## 6 5,200 to 10,400 396  
## 7 Above 36,400 89  
## 8 Under 2,600 133  
## 9 Unknown 126

This means 75.1 % of the data has null values for the amt\_weekends

amt <- data %>% count(is.na(amt\_weekends)) %>%   
 rename("NA\_value" = "is.na(amt\_weekends)" ) %>%   
 rename("amount"="n")   
amt

## NA\_value amount  
## 1 FALSE 421  
## 2 TRUE 1270

amt$amount[2]/(amt$amount[2] + amt$amount[1])

## [1] 0.7510349

This means 75.1 % of the data has null values for the amt\_weekdays too.

amt <- data %>% count(is.na(amt\_weekdays)) %>% rename("NA\_value" = "is.na(amt\_weekdays)" ) %>% rename("amount"="n")  
amt$amount[2]/(amt$amount[2] + amt$amount[1])

## [1] 0.7510349

amt

## NA\_value amount  
## 1 FALSE 421  
## 2 TRUE 1270

Since this is a high percentage I have decided to not include these two variables (amt\_weekdays and amt\_weekends)

Similar action is taken for type where there are a large amount of null values also.

data %>% count(type) %>% rename("amount"="n")

## type amount  
## 1 1270  
## 2 Both/Mainly Hand-Rolled 10  
## 3 Both/Mainly Packets 42  
## 4 Hand-Rolled 72  
## 5 Packets 297

Thus after eliminating those variables we have the following dataset.

data <- data[1:9]  
head(data)

## gender age marital\_status highest\_qualification nationality ethnicity  
## 1 Male 38 Divorced No Qualification British White  
## 2 Female 42 Single No Qualification British White  
## 3 Male 40 Married Degree English White  
## 4 Female 40 Married Degree English White  
## 5 Female 39 Married GCSE/O Level British White  
## 6 Female 37 Married GCSE/O Level British White  
## gross\_income region smoke  
## 1 2,600 to 5,200 The North No  
## 2 Under 2,600 The North Yes  
## 3 28,600 to 36,400 The North No  
## 4 10,400 to 15,600 The North No  
## 5 2,600 to 5,200 The North No  
## 6 15,600 to 20,800 The North No

## Diving into the Data

data %>% count(gender) %>% rename("amount"="n")

## gender amount  
## 1 Female 965  
## 2 Male 726

data %>% count(marital\_status) %>% rename("amount"="n")

## marital\_status amount  
## 1 Divorced 161  
## 2 Married 812  
## 3 Separated 68  
## 4 Single 427  
## 5 Widowed 223

data %>% count(highest\_qualification) %>% rename("amount"="n")

## highest\_qualification amount  
## 1 A Levels 105  
## 2 Degree 262  
## 3 GCSE/CSE 102  
## 4 GCSE/O Level 308  
## 5 Higher/Sub Degree 125  
## 6 No Qualification 586  
## 7 ONC/BTEC 76  
## 8 Other/Sub Degree 127

data %>% count(nationality) %>% rename("amount"="n")

## nationality amount  
## 1 British 538  
## 2 English 833  
## 3 Irish 23  
## 4 Other 71  
## 5 Scottish 142  
## 6 Unknown 18  
## 7 Welsh 66

data %>% count(ethnicity) %>% rename("amount"="n")

## ethnicity amount  
## 1 Asian 41  
## 2 Black 34  
## 3 Chinese 27  
## 4 Mixed 14  
## 5 Refused 13  
## 6 Unknown 2  
## 7 White 1560

data %>% count(gross\_income) %>% rename("amount"="n")

## gross\_income amount  
## 1 10,400 to 15,600 268  
## 2 15,600 to 20,800 188  
## 3 2,600 to 5,200 257  
## 4 20,800 to 28,600 155  
## 5 28,600 to 36,400 79  
## 6 5,200 to 10,400 396  
## 7 Above 36,400 89  
## 8 Under 2,600 133  
## 9 Unknown 126

data %>% count(region) %>% rename("amount"="n")

## region amount  
## 1 London 182  
## 2 Midlands & East Anglia 443  
## 3 Scotland 148  
## 4 South East 252  
## 5 South West 157  
## 6 The North 426  
## 7 Wales 83

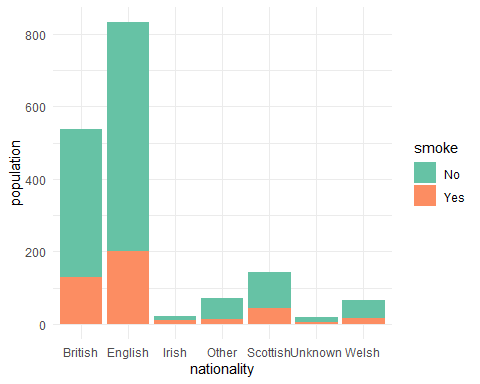
data %>% count(smoke) %>% rename("amount"="n")

## smoke amount  
## 1 No 1270  
## 2 Yes 421

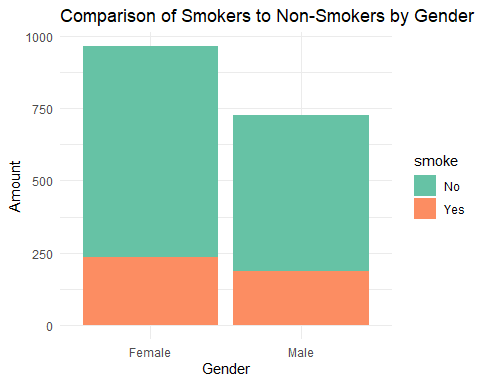
## Data Visualization of Data

Some visualizations of data that could bring some insight on predicting who smokes.

data %>% group\_by(nationality) %>%   
 count(smoke) %>% rename("population" = "n") -> nat\_smoke  
   
   
ggplot(data = nat\_smoke, aes(x=nationality, y = population, fill = smoke)) + geom\_bar(stat="identity") + scale\_fill\_brewer(palette="Set2")+  
 theme\_minimal() #+ facet\_wrap(~smoke)



data %>% group\_by(gender) %>%   
 count(smoke) -> gend\_smoke  
   
ggplot(data = gend\_smoke, aes(x=gender, y = n, fill = smoke)) + labs(title="Comparison of Smokers to Non-Smokers by Gender",   
 x="Gender", y = "Amount") + geom\_bar(stat="identity") + scale\_fill\_brewer(palette="Set2") + theme\_minimal()



To be continued