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 original source 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

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
```

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
```

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
```

```
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           Refused    108
## 9       Under 2,600    133
## 10          Unknown     18
```

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$gross_income[data$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
```

```
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
```

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("a
amt$amount[2]/(amt$amount[2] + amt$amount[1])
```

```
## [1] 0.7510349
```

```
amt
```

```
##   NA_value amount
## 1    FALSE    421
## 2     TRUE   1270
```

Since this 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
```

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, and smoke columns into factors.

```
data <- data[2:10]
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)
head(data)
```

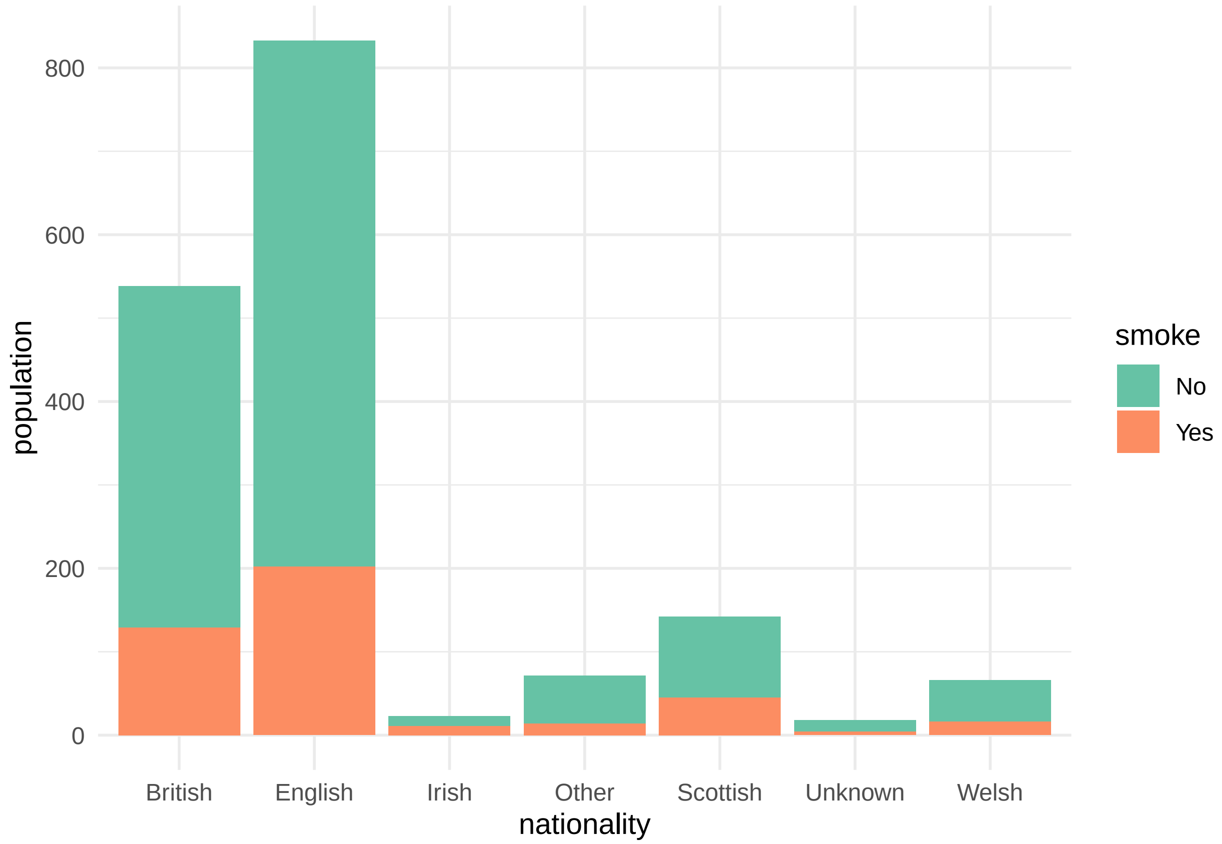
```
##   gender age marital_status highest_qualification nationality ethnicity
## 1   Male  38       Divorced      No Qualification     British     White
## 2 Female  42         Single      No Qualification     British     White
## 3   Male  40        Married                Degree     English     White
## 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
```

Data Visualization of Data

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

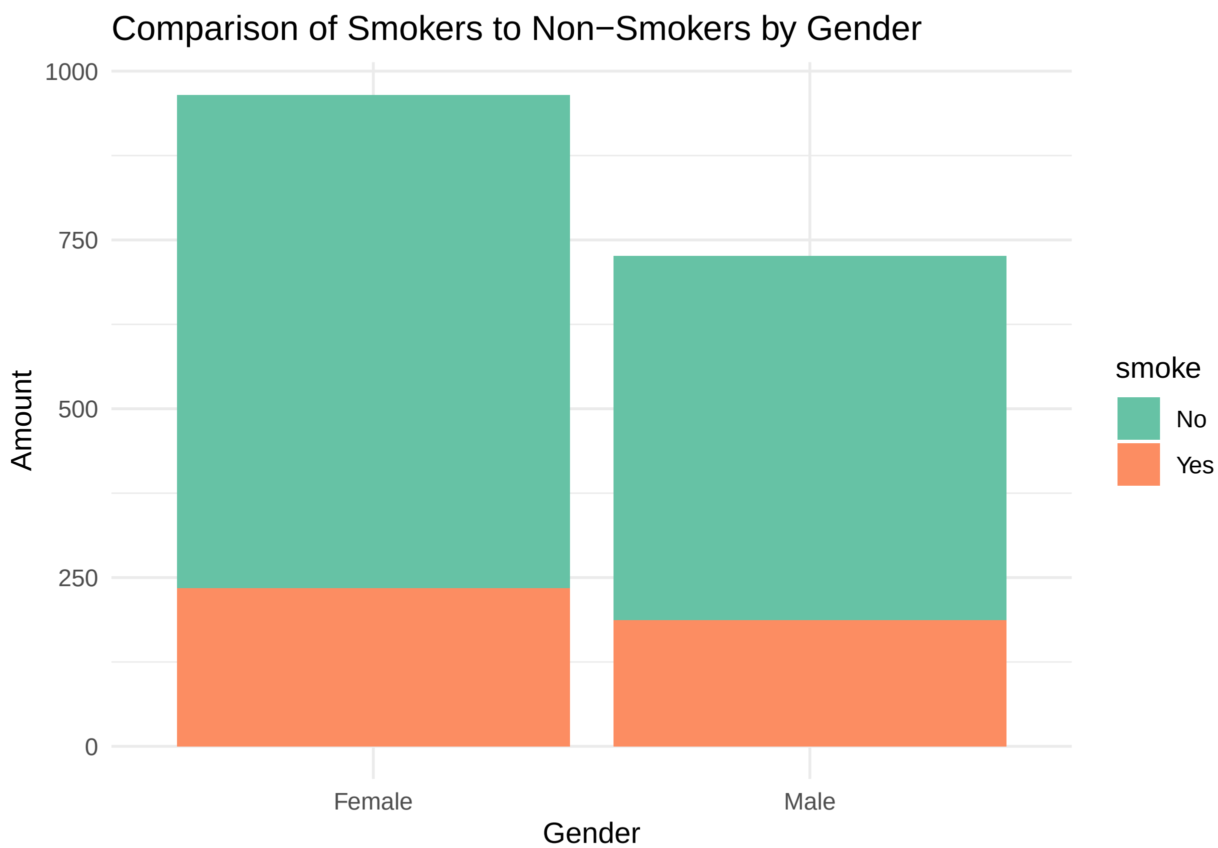
```
data %>%  group_by(nationality) %>%
  count(smoke) %>% rename("population" = "n") -> nat_smoke


ggplot(data = nat_smoke, aes(x=nationality, y = population, fill = smoke)) + geom_bar(stat="identity") +
  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 No
         x="Gender", y = "Amount") + geom_bar(stat="identity") + scale_fill_brewer(palette="Set2") + the
```



To be continued