Analysis of Health Survey for England (HSE) 2019

Candidate Numbers Here

March 07, 2024

Abstract

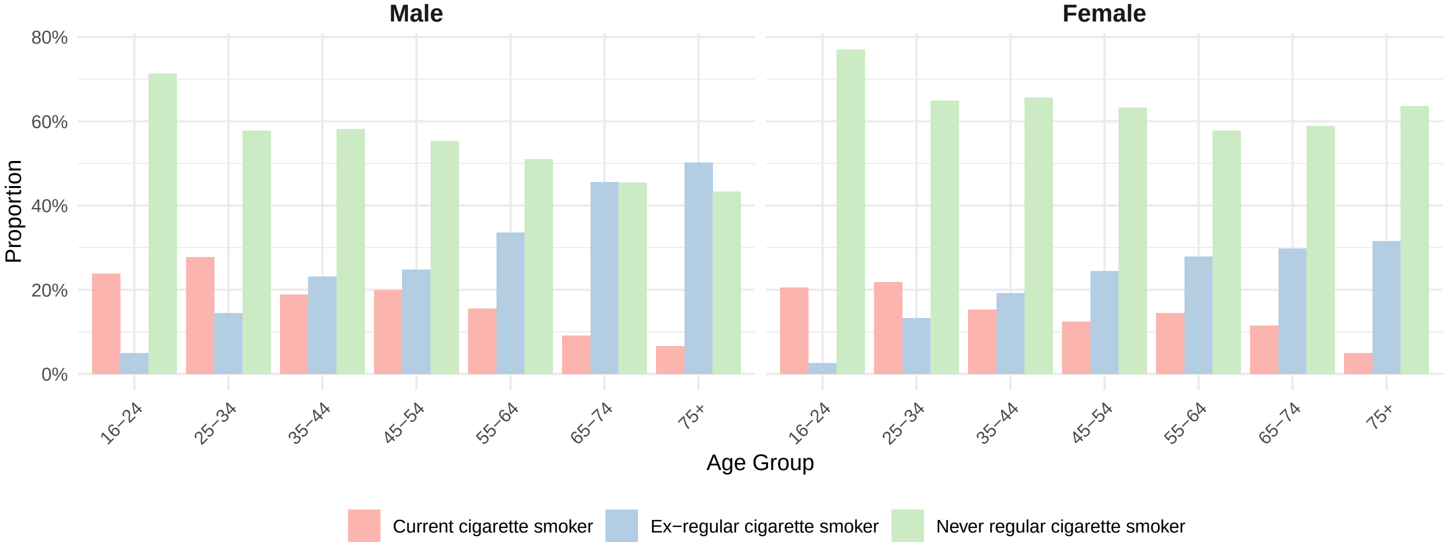
This report provides an analysis of data related to health, age, socio-economic factors and lifestyle habits in adults (from the age of 16) from the population in England, derived from the Health Survey for England 2019.

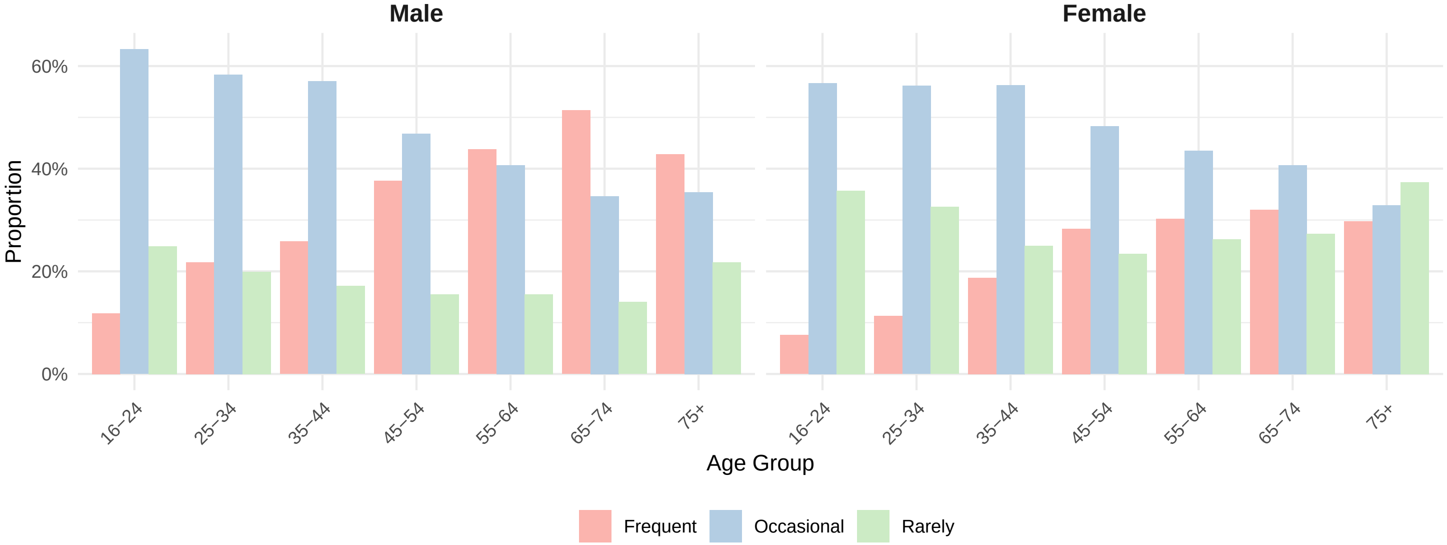
Introduction

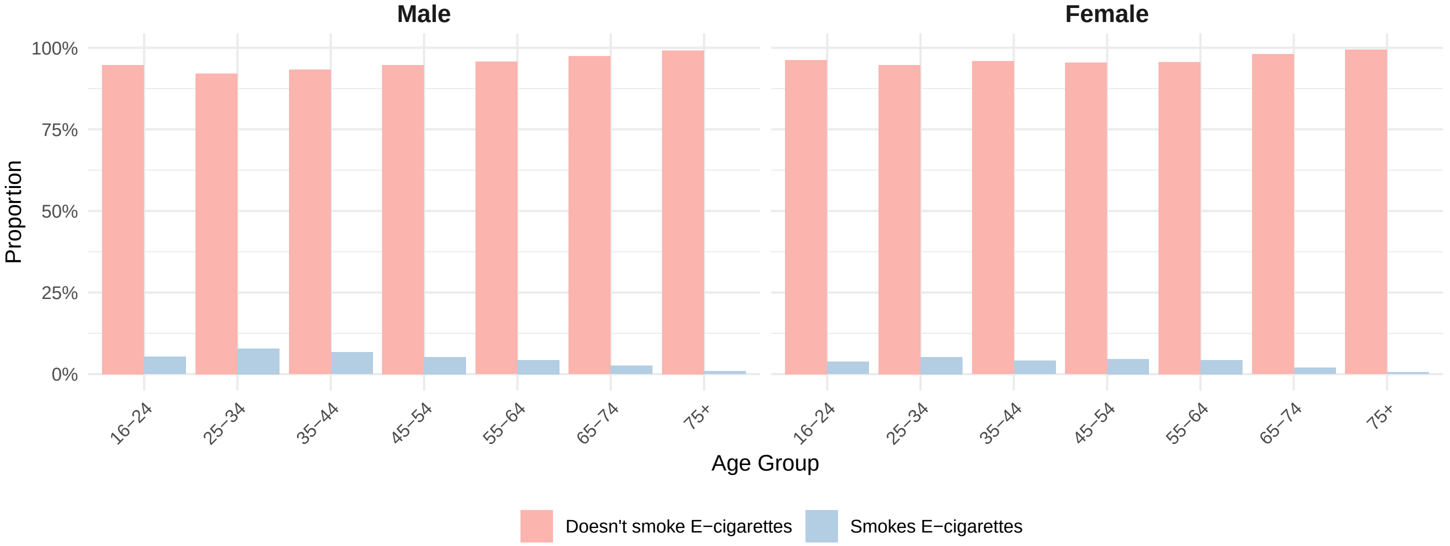
In the UK, smoking, vaping, and alcohol consumption are widespread, particularly among youngsters. It is crucial to be aware of the consequences and dangers of these habits, and the complications they can cause in later-life. The Office for National Statistics (ONS), regarded as the foremost statistical institute in the UK, is the “go-to” for insights into public health trends. According to their estimations, approximately 14.1% of the adult population aged 18 and above were identified as cigarette smokers (ONS 2019a). Furthermore, their data revealed there were 7,565 deaths attributed to alcohol-specific causes in 2019 (ONS 2019b). These statistics alone warrant a need for an understanding of health-related behaviours and the outcome. Therefore, the aim of our study is to investigate not only the prevalence but also the severity of these habits, exploring different socioeconomic factors that may potentially contribute to each. Additionally, we will investigate the greater implications of these bad habits and how they relate to systolic blood pressure levels throughout the UK population.

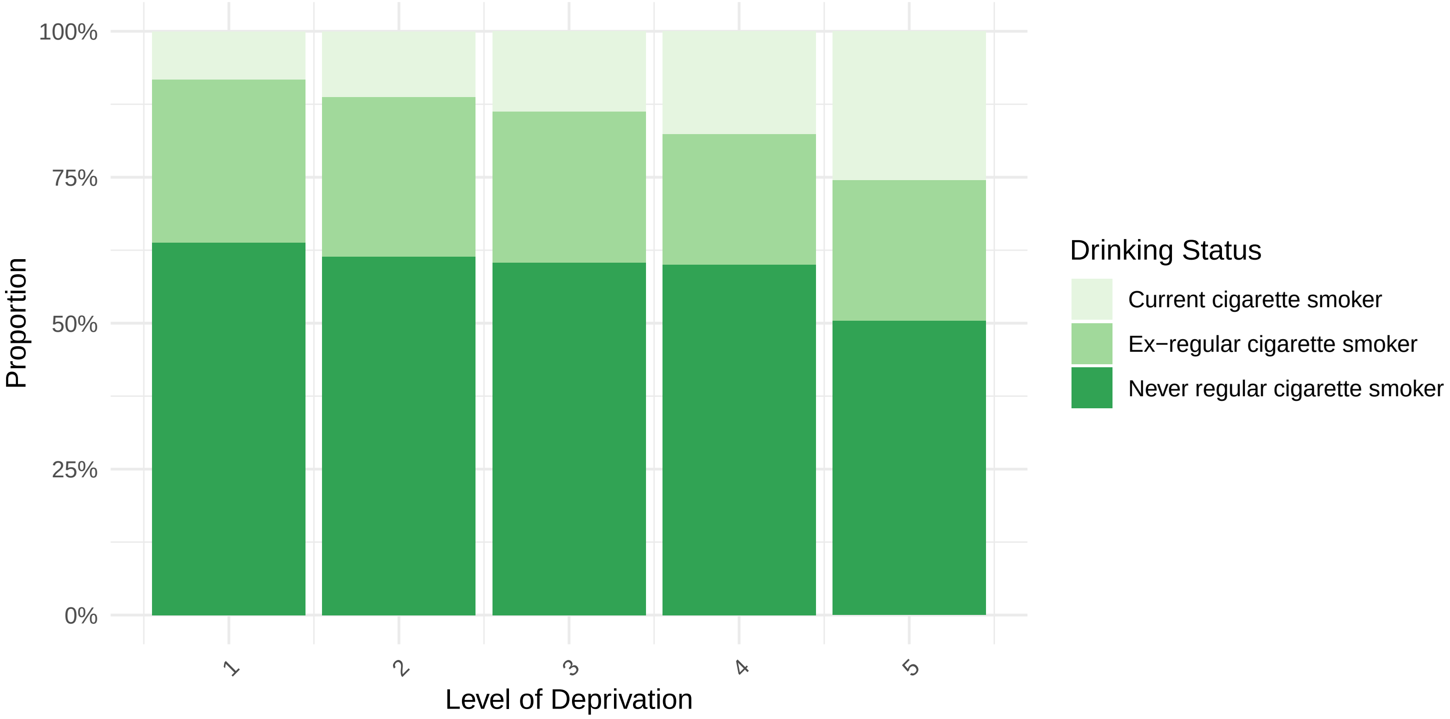
Question One

We start by investigating the data…









To calculate the prevalence of each habit we assume each of the n observations, $x_1, \ldots, x_n$, to be independent, identically distributed (iid) random variables (RVs) where $x_i \sim Bern(p)\, \forall i = 1, \ldots, n$ and p denotes the probability of an observation having the relevant habit. We use the household weights to calculate a weighted MLE of p. That is, letting w_i denote the weight of the i^{th} observation, we alter the standard likelihood to be

$$L(p|\mathbf{x}) = \prod_{i=1}^{n} (p^{x_i}(1-p)^{1-x_i})^{w_i}$$

From this, we can calculate our weighted MLE

$$\widehat{p} = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i}$$

We also can see that this MLE has variance given by $var(\widehat{p}) = \frac{p(1-p)}{\sum_{i=1}^{n} w_i}$, which we can estimate using $\widehat{p}$ and use the large sample Normal approximation of the MLE to estimate our 95% confidence intervals.

Table 1: Estimates and 95% Confidence Intervals for % of Population

Category	Estimate	C.I.
Drinking	80.3%	(79.5%, 81.2%)
Smoking	16.5%	(15.7%, 17.3%)
Smoking E-cigarettes	4.28%	(3.84%, 4.72%)

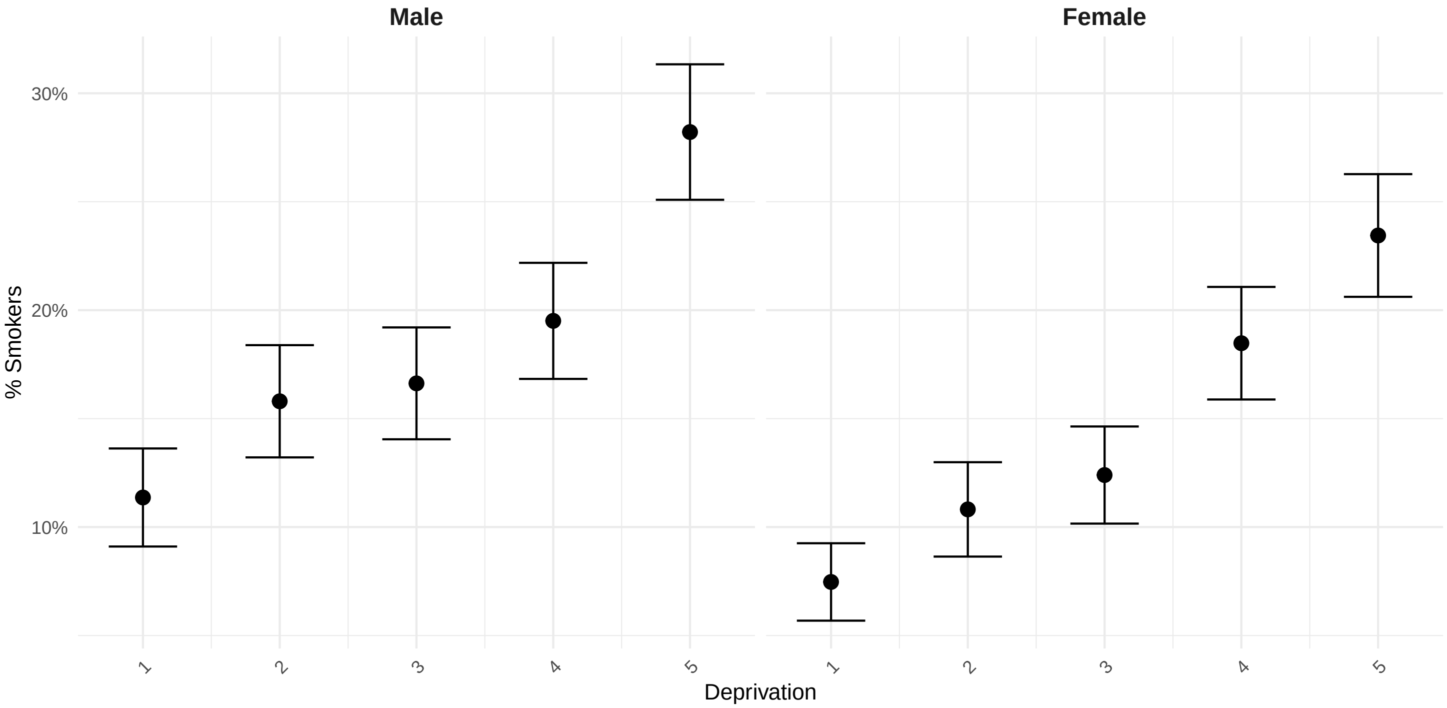


Figure 1: Estimation of the prevelance of smokers by deprivation

Question Two

Explanation of why we split into test/train dataset.

This leaves us with 6563 observations in our training dataset.

Table 2: Missing values in the training dataset

Variable	Missing Values	% Missing
omsysval	3190	48.6%
BMIVal	1212	18.5%
dnoft__19	1189	18.1%
cigdyal__19	44	0.67%
cigsta3__19	43	0.655%
NDPNow__19	43	0.655%
d7many3__19	42	0.64%
drinkYN__19	41	0.625%
topqual2	33	0.503%
origin2	22	0.335%
marstatD	1	0.0152%

Creating first model on train data

References

ONS. 2019a. “Adult smoking habits in the UK: 2019.” https://shorturl.at/qQW27.

———. 2019b. “Alcohol-specific deaths in the UK: registered in 2019.” https://shorturl.at/gqxY1.