Demographic characteristics of smokers in UK*

Lida Liu

04/25/2022

Abstract

Smoking is a normal hobby for some of the people, and from the investigation towards cigarettes, smoking is extremely bad for people's health with providing about 250 chemicals that would cause cancers. Therefore, National STEM Centre from University of York has surveyed the UK smoking data from 1691 random people in UK. To be more specific, the survey contains the age of respondents, gender of respondents, the marital status of respondents, the nationality of respondents, the ethnicity of respondents, the income of respondents, the frequency of respondents and the type of cigarettes they are smoking. In a reproducible way, I obtained some plots and tables, containing the results of the survey, and then I analyze the specific conditions of the demographic characteristics of smokers in UK.

Contents

*Code and data are available at: https://github.com/Dav1dLLD/UK-Smoking.

1 Introduction

Nowadays, smoking is a normal hobby for some of the people, and from the investigation towards cigarettes, smoking is extremely bad for people's health with providing about 250 chemicals that would cause cancers (NationalCancerInstitute 2022) . Therefore, National STEM Centre from University of York has surveyed the UK smoking data from 1691 random people in UK. To be more specific, the survey contains the age of respondents, gender of respondents, the marital status of respondents, the nationality of respondents, the ethnicity of respondents, the income of respondents, the frequency of respondents and the type of cigarettes they are smoking . I get the data from the statistical website openintro (Openintro 2022). Next, In a reproducible way, I obtained some plots and tables, containing the results of the survey, and then I analyze the specific conditions of the demographic characteristics of smokers in UK.

In this paper, I rearrange the data through data cleaning process, and I create some variables for my convenience of analyzing the data. After the data cleaning, I create some bar plots to show the relationship between each demographic characteristics of smokers in UK. Based on these plots, I create some logistic regression model to justify the phenomenon I see in the plots. Next, I put the results I get from my statistical models and arrange them into a table in the result section. In the end of the paper, I discuss the phenomenon from my analysis, and I make more research as an appendix for my discovery in the discussion section. Moreover, I state the limitations that might occur during my research.

The entire paper is written in R, and all the data analysis in this paper uses R studio (R Core Team 2021) with openintro (Çetinkaya-Rundel et al. 2022), tidyverse (Wickham et al. 2019) , lme4 (Bates et al. 2015), mgcv(Wood 2017) and ggplot(Wickham 2016) packages.

2 Data

2.1 Methodology

First of all, the data set is obtained by National STEM Center from University of York through the survey towards 1691 random people about their smoking conditions in UK. It contains 1691 samples with 12 different variables, and I create a table as an introduction of each variables for better visualization. The table below shows the name of each variable, their statistical type and the information they contain.

Variable	Variable Type	Description
age	Numerical	The age of the surveyed people
marital_status	Categorical	The marital_status of the surveyed people
gender	Categorical	The gender of the surveyed people
amt_weekends	Numerical	The number of cigarettes smoked per day on weekends
amt_weekdays	Numerical	The number of cigarettes smoked per day on weekdays
highest_qualification	Categorical	The highest degree of surveyed people
nationality	Categorical	The nationality of the surveyed people
ethnicity	Categorical	The ethnicity of the surveyed people
gross_income	Categorical	Income range of the surveyed people
region	Categorical	Where are the surveyed people come from in UK
smoke	Categorical	Whether the surveyed people smoke or not
type	Categorical	The type of cigarette the surveyed people smoke

2.2 Data cleaning

For the data cleaning part, I firstly import the raw data to my Rstudio, and I check each variable to see if there is any mistakes or confounding variables. Next, I check the effectiveness of each variable and I filter some variables that may not be useful in my analysis through my data manipulation using Rstudio. In the end, I rearrange the data set and create a new table that contains the variables I'm going to use for further analysis based on regression models. Then, I create a new factor variable called smoker, which is going to be convenient for my building of regression model. Moreover, I also create a new variable called smoking_per_day, which is obtained by (smoking_per_weekdays + smoking_per_weekends) divided by 2. In this case,smoking_per_day is the number of cigarettes smoked per day.

Variable	Variable Type	Description
age	Numerical	The age of the surveyed people
marital_status	Categorical	The marital_status of the surveyed people
smoking_per_day	Numerical	The number of cigarettes smoked per day
gender	Categorical	The gender of the surveyed people
ethnicity	Categorical	The ethnicity of the surveyed people
gross_income	Categorical	Income range of the surveyed people
region	Categorical	Where are the surveyed people come from in UK

2.3 Data visualization

First of all, some bar plots are created to show an approximate relationship between the respondent and whether they smoke or not based on the variable in the data set.

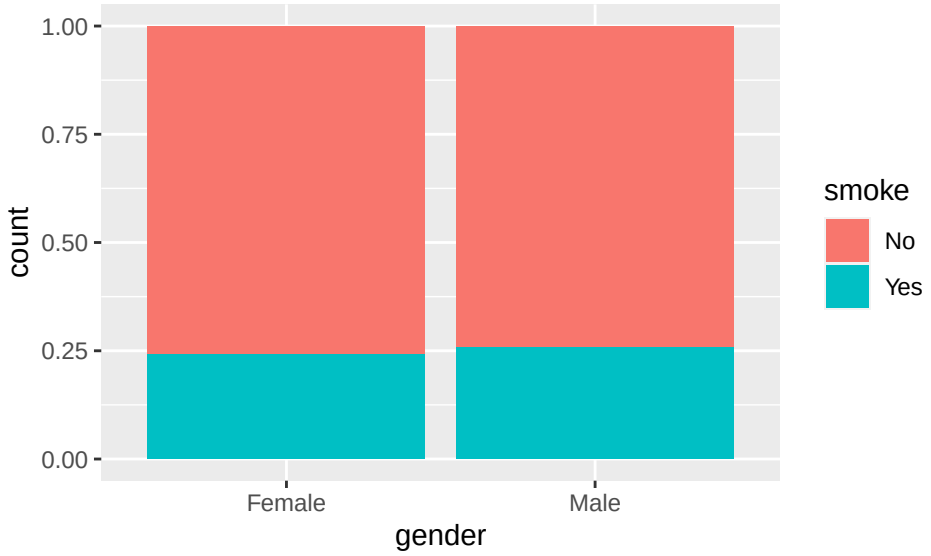

Figure 1: Barplot of whether respondent smokes or not according to gender

From figure 1, the bar plot of whether respondent smokes or not, I can see that approximately 25% of both female and male respondents have the habit of smoking. However, I can not conclude that gender is not a factor to decide whether a person smokes or not, so I need to put the gender factor into my regression model as well, in order to see whether there is a relationship or not.

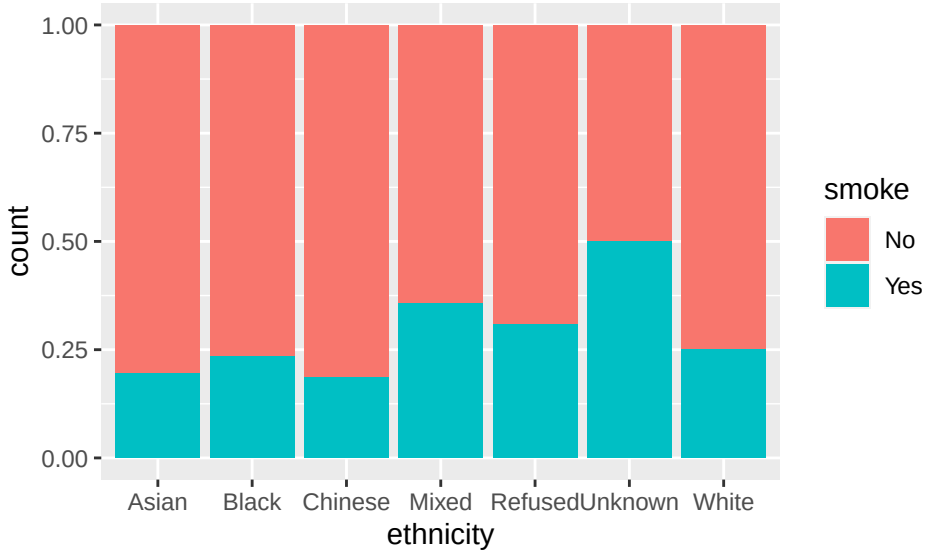

Figure 2: Barplot of whether respondent smokes or not according to ethnicity

Due to the fact that the data collection of ethnicity would be considered racial discrimination issues, I only look for respondent who are willing to provide their ethnicity. According to the figure 2, respondents who

are mixed have a more percentage of smoking than other ethnicity, while Chinese are seems to be the less ethnicity who are smoking. However, our sample size is not large enough, so that I will put ethnicity in my regression model to test if it is truly an influential variable.

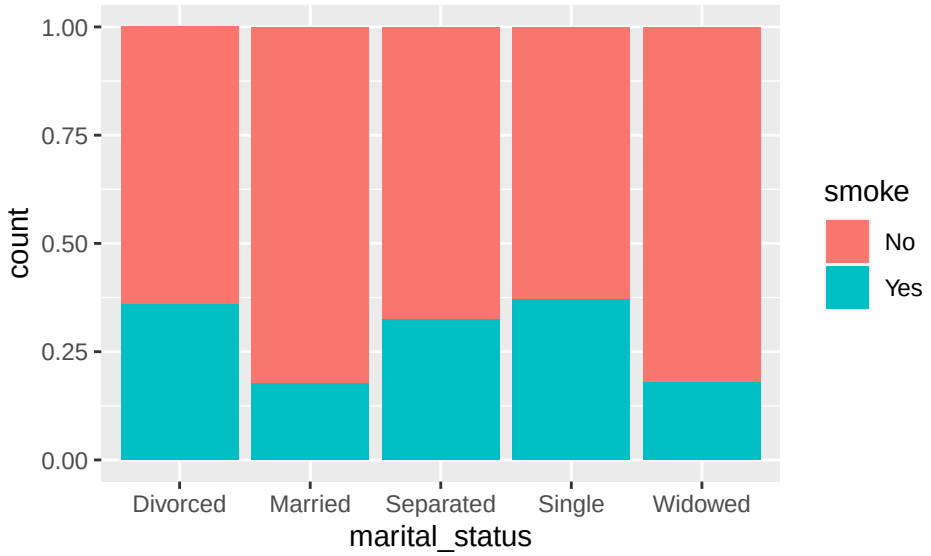

Figure 3: Barplot of whether respondent smokes or not according to marital status

From figure 3, I can see that approximately 36% of the respondents who are single and divorced have the habit of smoking, while approximately 20% of the married and widowed respondents have the habit of smoking. As I mentioned before, due to the small scale of sample size, the visualization may not be accurate, so that I will put marital status into my regression model as well.

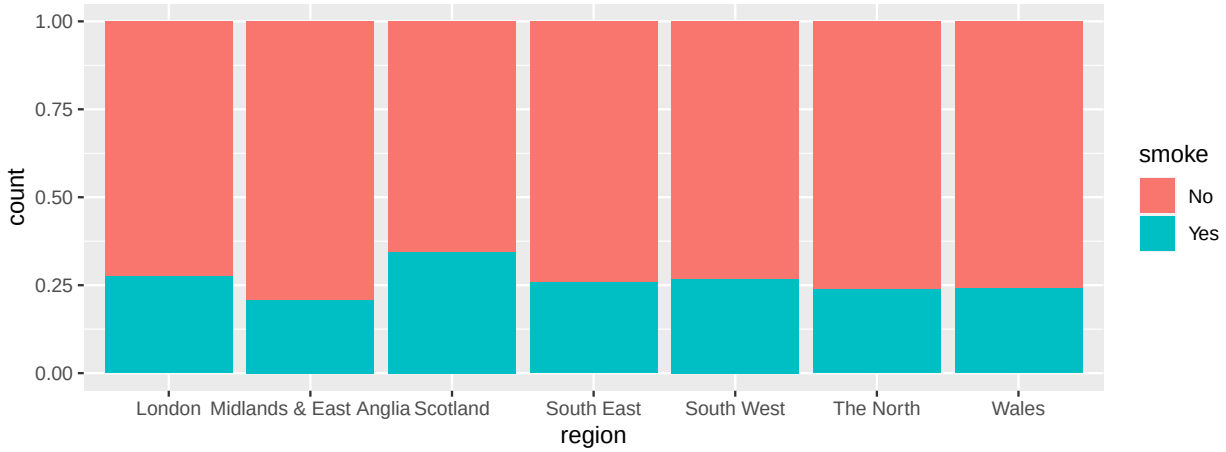

Figure 4: Barplot of whether respondent smokes or not according to region

From figure 4, I can see that approximately 35% of the respondents who are from Scotland have the habit of smoking, which are more than the other regions, and for the accuracy, I will put the variable region into my regression model as well.

3 Model

3.1 Logistic Regression Models

First of all, I create four logistic regression models to take a glance on each of the variables that might influence whether respondents smoke or not. In this case, the logistic regression models are appropriate, because the logistic regression model is a regression model that will show the approximate relationship between each variables to the response variable smoker, based on the p-values and estimated value provided in fixed effects. To be more specific, p-value is a useful statistical value that would show whether the variable is statistically significant or not. If p-value is smaller than 0.05, it means that the variable is statistically significant, so that the chosen variable is truly influential to the response variable. In this case, I choose the variable smoker to be my response variable, and I would like to test these variables of gross_income ,region, marital_status and ethnicity respectively in four logistic regression models.

From model one with the response variable of smoker and the tested variable of gender and gross_income, with a random slope of age, I can see that the p-value for the respondents who have an income range of 28600 to 36400 is extremely small, which is 0.000692. Moreover, the estimated value for the income range of 28600 to 26400 is -1.311143, which means that respondents in this income range are not likely to be smokers. In addition, it is interesting to see that the estimated value for income range from 0 to 10400 are slightly negative, and the estimated value for income range from 5200 to 10400 is even positive, which means that respondents who are having a relatively low income would are more likely to be smokers.

From model two with the response variable of smoker and the tested variable of gender and region, with a random slope of age, I can see that all the p-values for different regions are greater than 0.05, which means that the variable region is not an influential factor for whether a respondent is a smoker or not. Therefore, I will not consider the variable region in my further analysis.

From model three with the response variable of smoker and the tested variable of gender and marital status, with a random slope of age, I can see that the p-value for the married respondents is extremely small(2.14e-07), and widowed respondents have a p-value of 0.00771, which is smaller than 0.05 as well. The estimated value for married respondents is -1.00606 and the estimated value for widowed respondents is -0.70334, which means that respondents who are married and widowed are less likely to be smokers.

From model four with the response variable of smoker and the tested variable of gender and ethnicity, I can see that all the p-values for different regions are greater than 0.05, which means that the variable ethnicity is not an influential factor for whether a respondent is a smoker or not as well. Thus, I will not consider the variable region in my further analysis.

Moreover,for all the models, I can see that all the p-values of gender are greater than 0.05 , which means that gender is not a factor to decide whether a respondent smokes or not as well.

3.2 Linear Regression Models

Moreover, I want to test whether age is a influential variable for people to be smokers. In this case, since the response variable becomes the variable age, I could only make a linear regression model. In the linear regression model between the response variable age and the tested variable smoker, I can see that the p-value for smoker is smaller than 2e-16, which means that the p-value is extremely small, so that the variable is statistically significant. The estimated value for smoker is approximately -9.5, which means that the respondent who are smoking is 9.5 years younger than respondent who are not smoking in average.

Since age now is an influential factor, I would like to know the relationship between the age of the respondents and the number of cigarettes the respondents would smoke per day. In this case, by creating a linear regression model of the response variable age and the tested variable smoking_per_day, I can see that the p-value for smoking_per_day is statistically significant, and the estimated value of it is positive, which means that with the increase of age, the number of cigarettes the respondents smoke per day is going to slightly increase as well.

4 Result

From the bar plots in the data visualization part, the logistic regression models and the linear regression models I created before, some conclusions could be made based on these statistical tools:

First of all, age is influential to whether a respondent is a smoker or not. From the linear regression model of age and whether a respondent is a smoker or not, the respondent who are smoking is 9.5 years younger than respondent who are not smoking in average. Moreover, according to the second linear regression model I created before, with the increase of age, the number of cigarettes the respondents smoke per day is going to slightly increase as well.

Secondly, marital status is influential to whether a respondent is a smoker or not. From the logistic regression model I created before, the respondents who are married or widowed are more likely to be smokers than the other marital status.

Thirdly, the income range is also influential to whether a respondent is a smoker or not. From the logistic regression model I created before, the respondents who are having a relatively lower income range are more likely to be smokers than the respondents who are having a relatively higher income range.

In the data visualization part, the ethnicity and region of respondents seem to be influential to whether they are smokers or not. However, with the fact from the logistic regression model, I may conclude that ethnicity and region are not influential to whether respondents smoke or not, so that the result from the data visualization is not accurate for these two variables. On the other hand, gender of respondents seem to be accurate from the data visualization, because the linear logistic model shows that there are no relationship between the gender of respondents and whether they smoke or not.

To have a better visualization of my results, I create the table below:

Demographic characteristics	Influence
Age of Respondents	The respondent who are smoking is 9.5 years younger than respondent who are not smoking in average , and with the increase of age, the number of cigarettes the respondents smoke per day is going to slightly increase as well.
Marital status of Respondents	The respondents who are married or widowed are more likely to be smokers than the other marital status.
Gender of Respondents	No obvious influence.
Ethnicity of Respondents	No obvious influence.
Income range of Respondents	The respondents who are having a relatively lower income range are more likely to be smokers than the respondents who are having a relatively higher income range.
Region of respondents	No obvious influence.

5 Discussion

5.1 Overview of Results

In this paper, I analyzed the demographic characteristics of smokers in UK using statistical tools, and I discovered a lot of phenomenon from the characteristics.

First of all, for the phenomenon of the respondent who are smoking is 9.5 years younger than respondent who are not smoking in average, I would say that smoking is become more and more popular in teenagers, due to the fact that teenagers may try to imitate some elder smokers to make them look more mature. However, as I mentioned before, smoking is truly harmful to people's health, especially to teenagers who are still growing their bodies(NationalCancerInstitute 2022). Thus, I personally do not think that this phenomenon is good to see. Additionally, elder people are smoking more numbers of cigarettes. This phenomenon is realistic, because smoking is addictive, and elder smokers who have more smoking years would be more addictive as well.

Secondly, people who are married and widowed are less likely to be smokers. This phenomenon is understandable. For people who are married, their habit of smoking may be prohibited by their husbands and wives by the consideration of healthy issues. For people who lose their husband or wife, they would be more likely to treasure their health and live in a healthy lifestyle.

Moreover, people who are making less money would be more likely to be smokers, and this phenomenon is realistic as well. From my prospective, people who are making less money may be suffering from the pressure of life, and according to the scientific research, smoking would help people reduce their pressure and make them feel relax.

5.2 Harm of smoking

Before I write this paper, I know that smoking is harmful to health, but I do not really know how smoking is hurting people's health. Therefore, I make more research from the scientific article provided by the national cancer institute, and I am shocked by the negative impact smoking would bring to people (NationalCancerInstitute 2022).

First of all, the article states that breathing even a little tobacco smoke could be harmful, because tobacco smoke contains more than 7000 chemicals, and at least 250 chemicals are harmful to our health, such as the famous harmful gas carbon monoxide. Moreover, in these 250 chemicals, about 69 chemicals may cause cancer (NationalCancerInstitute 2022). I know that the lung cancer is mostly fatal, and there are so many chemicals that could cause cancer in one cigarette. Nowadays, smoking becomes a lot of people's habit, because this behavior would make nerve become numb, so that people could feel relax. Thus, the second-hand smoke exists everywhere, and the second-hand smoke is harmful to both smokers and nonsmokers. According to my analysis, approximately 25 percent of the population in UK are smokers, which is a large number of people. Unfortunately, some of them even may not know that smoking is extremely harmful to their health and love to smoke in public. Consequently, I think governments all over the world should put more effort to reduce the negative impact that smoking would provide. For example, governments should provide more smoking zone for people who want to smoke in public, and advertise the drawbacks of smoking to let some of the smokers to quit this behavior. I know that smoking is a private habit, but I think governments are responsible for letting people know the harm of smoking.

5.3 Limitations

In fact, some limitations are involved in this paper. First of all, the sample size of the data set is 1691, which is a small sample size. Therefore, the analysis may not be accurate, due to the small size of the samples. Moreover, for the variable ethnicity, some of the respondents refuse to provide their personal information, so that the results of ethnicity may not be accurate as well, due to the lack of data. Furthermore, I used simple division to calculate the number of cigarettes smokers would smoke per day. However, some smokers love to smoke more in weekdays, while the other smokers love to smoke more in weekends, so that the frequency of smoking may not be accurate as well. In addition, this survey contains a lot of personal information, so that some of the information itself may not be accurate, which means that the information bias would probably occur in my analysis.

Reference

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67 (1): 1–48. https://doi.org/10.18637/jss.v067.i01.

Çetinkaya-Rundel, Mine, David Diez, Andrew Bray, Albert Y. Kim, Ben Baumer, Chester Ismay, Nick Paterno, and Christopher Barr. 2022. *Openintro: Data Sets and Supplemental Functions from 'OpenIntro' Textbooks and Labs*. https://CRAN.R-project.org/package=openintro.

NationalCancerInstitute. 2022. *What Harmful Chemicals Does Tobacco Smoke Contain*. https://www.cancer.gov/about-cancer/causes-prevention/risk/tobacco/cessation-fact-sheet#what-harmful-chemicals-does-tobacco-smoke-contain.

Openintro. 2022. *UK Smoking Data*. https://www.openintro.us/data/index.php?data=smoking.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wood, S. N. 2017. *Generalized Additive Models: An Introduction with r (2nd Edition). Chapman and Hall/CRC*. https://doi.org/https://doi.org/10.1201/9781315370279.