EDA with R
UK Smoking Data

Clay Jackson, cjackson@bellarmine.edu

Robert Pearson, rpearson2@bellarmine.edu

I. INTRODUCTION

Our dataset is called *UK Smoking Data*. It is from a survey taken in the United Kingdom that focused on the habits and demographics of citizens who smoke cigarettes. We chose this dataset due to the relevancy of smoking in our culture. Even though smoking is inherently dangerous and has an array of medical side effects, millions of people still partake. Our goal in analyzing this dataset is to find out what demographics of people (if any) are most likely to be smokers. The link to this dataset can be found here:

https://www.kaggle.com/datasets/mexwell/uk-smoking-data?select=smoking.csv

II. Dataset Description & Summary Statistics

Our data contains 1,962 samples of UK residents. There are 12 columns with various data types. **Table 1**, lists each variable in the dataset along with the data type (nominal, ordinal, interval, or ratio) and its data class in R. Our dependent variable in this study is *Smoke*. This variable's output is simply a "yes" or "no" to the question of "Do you smoke?" **Table 2** shows the descriptive statics for the three numeric variables we have in the dataset. Note that the missing data (NA's) for Amount on Weekdays and Amount on Weekends represent people that did not smoke. To fix this, we used imputation to replace the NA's with 0 before conducting our analysis.

Table 1: Data Types and R class

Variable Name	Data Type	R Class
Gender	nominal	character
Age	continuous	integer
Marital status	nominal	character
Highest qualification (education)	ordinal	character
Nationality	nominal	character
Ethnicity	nominal	character

Gross income	continuous	character
Region	nominal	character
Smoke (Y/N)	nominal	character
Amount on Weekdays	discrete	integer
Amount on Weekends	discrete	integer
Type	nominal	character

Table 2: Summary Statistics

Variable Name	Min	1Q	Median	Mean	3Q	Max	NA's
Age	16	34	48	49.84	65.50	97	0
Amount on Weekdays	0	10	15	16.41	20	60	1270
Amount on Weekends	0	7	12	13.75	20	55	1270

III. Graphical Exploration of Dataset

After organizing the data and running the summary statistics, we began to explore the data visually. First, the data had to be filtered into two categories, smokers,and non-smokers. From there, we utilized graphs to analyze the data and its variables in relation to smokers and non-smokers in the UK. We analyzed independent variables (such as gender, age, ethnicity, education, and income) to see their relation to smokers. These visualizations can be found below.

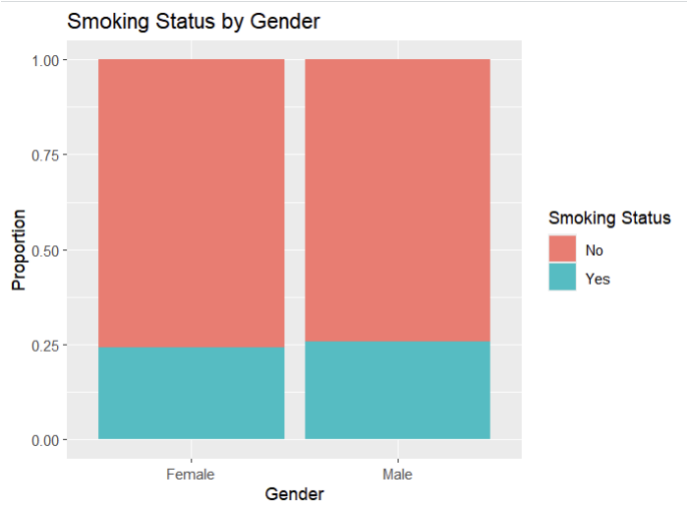

Figure 1: Bar plot of smoking status by gender

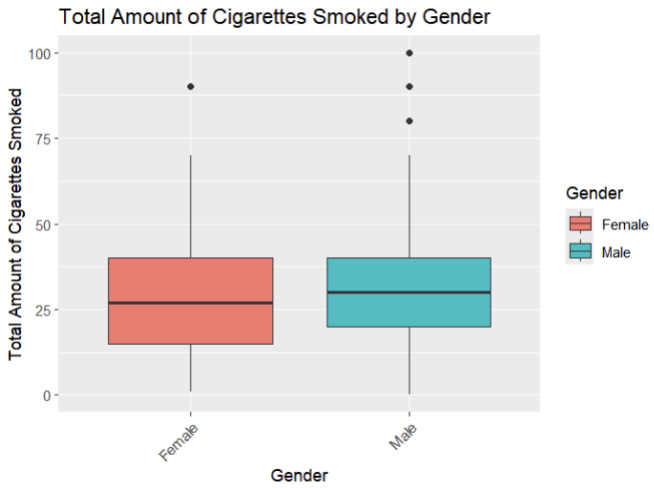

Figure 2: Box plot of total amount of cigarettes smoked by gender

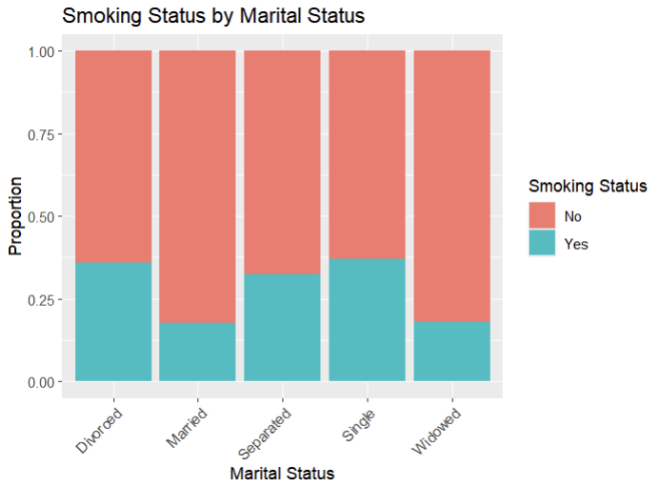

Figure 3: Bar plot of smoking status by marital status

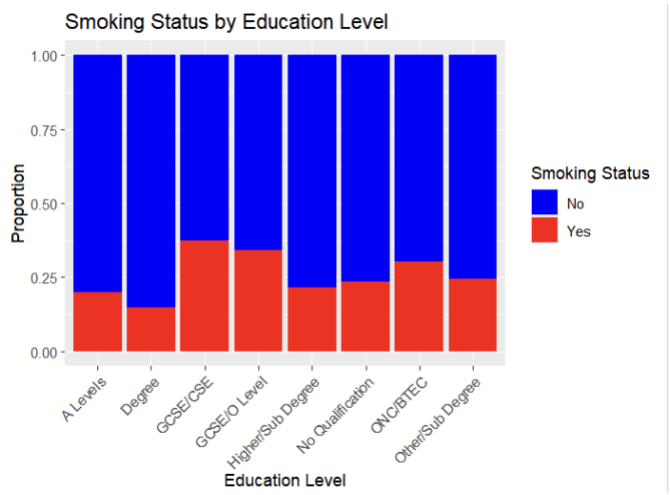

Figure 4: Bar plot of smoking status by education level

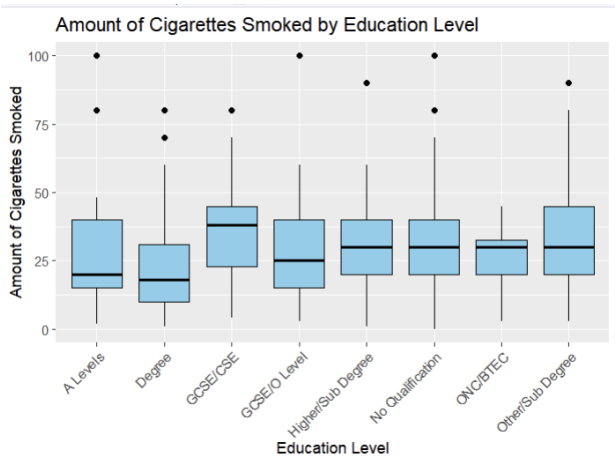

Figure 5: Box plot of amount of cigarettes smoked by education level

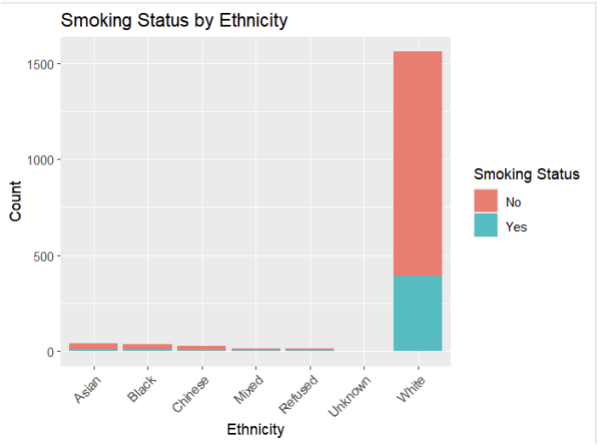

Figure 6: Bar plot of smoking status by ethnicity

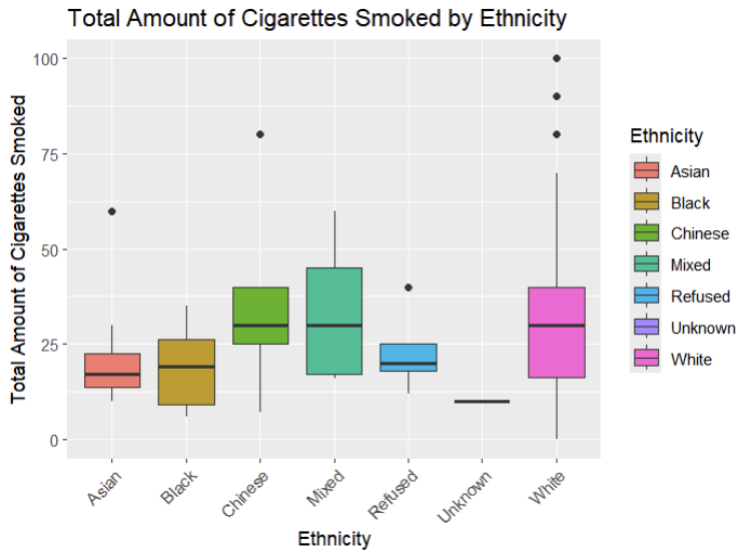

Figure 7: Box plot of total amount of cigarettes smoked by ethnicity

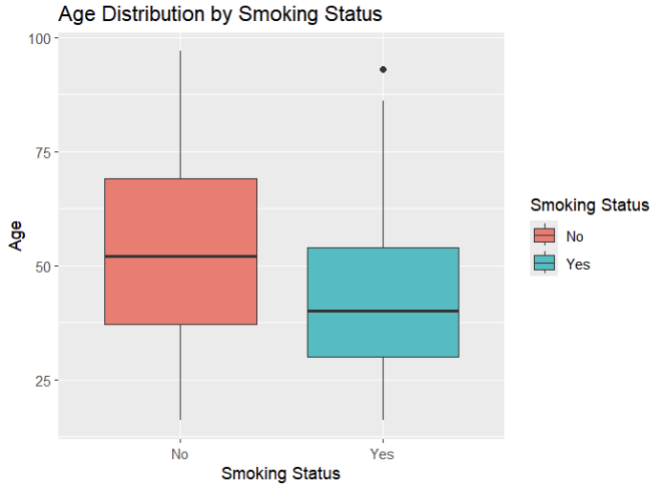

Figure 8: Box plot of age by smoking status

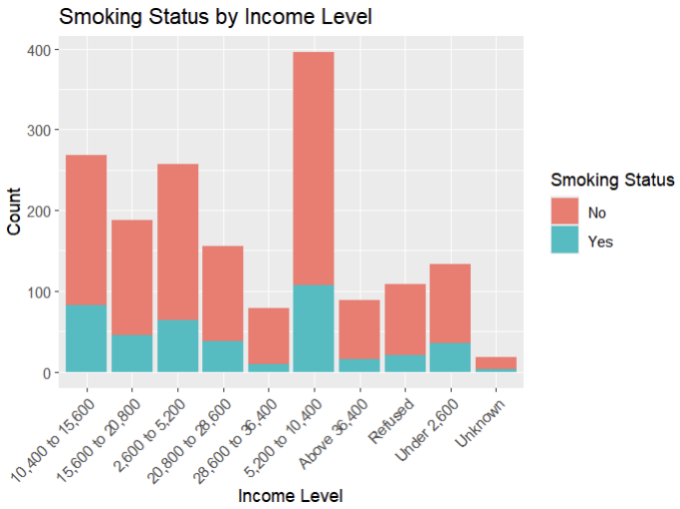

Figure 9: Bar plot of smoking status by income level

IV. Summary of Findings

After analyzing the data, we gained several useful insights about who is most likely to be a smoker. According to our sample, the number of men who smoke is nearly the same as the number of women who smoke (around 25 percent). Slightly more men smoke than women and at a higher volume, but the variance was minimal, within a few percentages of each other. We found that married people tend to smoke far less than single, separated, or divorced people (by more than 10 percent in some cases).

It appears that people with higher education (college degrees specifically) are less likely to be smokers than those with high school equivalencies only. When it comes to income, our data found that lower income people (between 2,600 and 15,600) are more likely to smoke than those who are financially better off. Lastly, it appears that those who do smoke tend to be younger. Smokers had an average age of late 30's. While people who claimed to not smoke had an average age of over 50.