Midterm Project

AUTHOR
Yang Xu

Introduction

Smoking, a global public health concern, has been the subject of extensive research due to its significant impact on human health and well-being. The detrimental effects of smoking on health are well-documented, with strong associations found between smoking and numerous health conditions including cardiovascular diseases, respiratory disorders, and various types of cancer. Additionally, smoking has also been linked to lifestyle factors and socio-economic status, further complicating its impact on public health.

The dataset we used in this project are compiled from the responses given by over 1,500 people to a survey and published by stats4schools (https://www.stem.org.uk/resources/elibrary/resource/28452/large-datasets-stats4schools). According to the website, this dataset was published during 2000 to 2009, with no specific time. So the dataset should be collected no later than 2009, which is a relatively old dataset. This dataset was collected in terms of questionnaires where paticipants were given certain questions with preset types of responses.

I will use this dataset to investigate the following two questions:

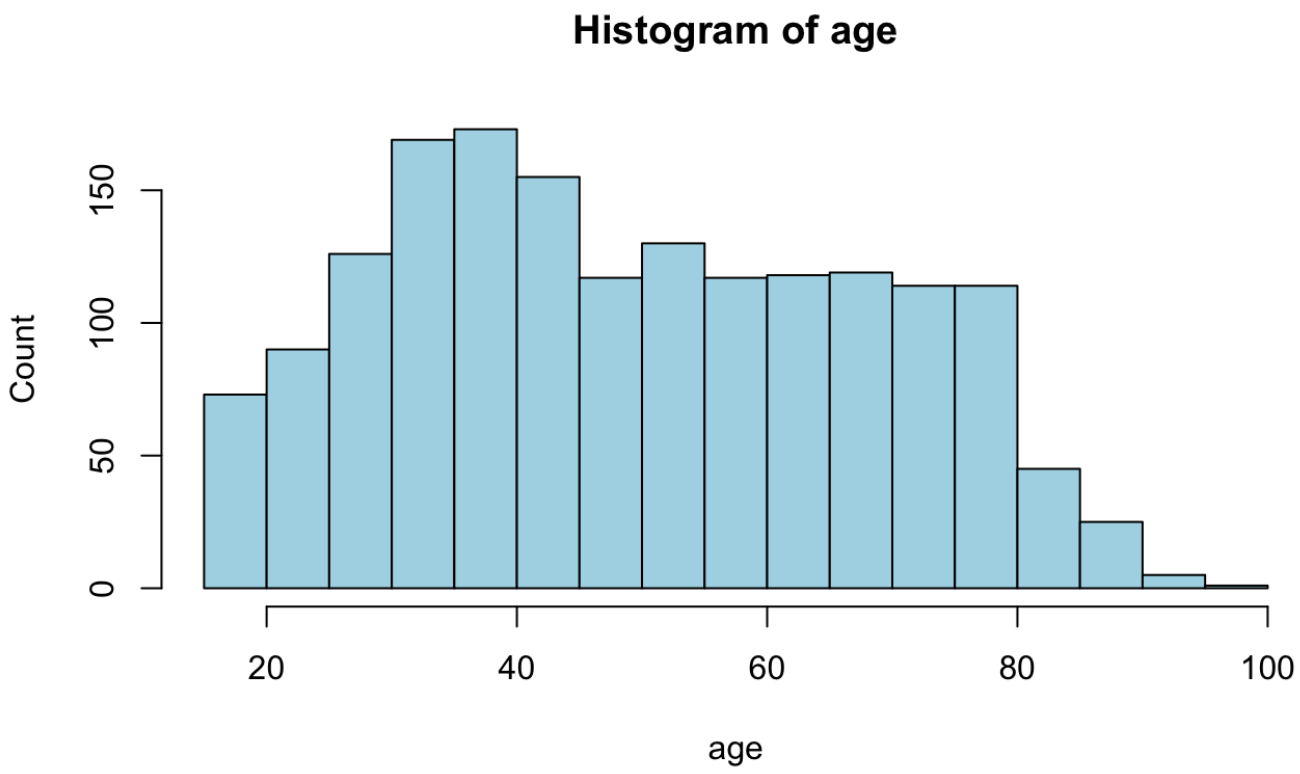
1. What are the main factors that influence whether a person smokes or not?
2. For those who smoke, what are the main factors that influence the number of cigarettes smoke per day?

Methods

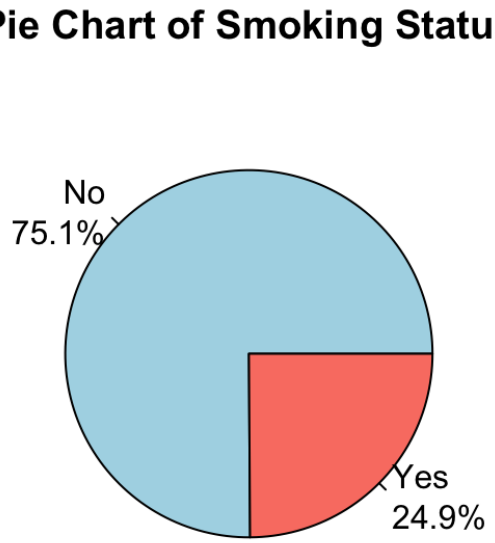
The dataset provided on the original website were in XLS File Format. I found another version of the same dataset with CSV file format.(https://www.kaggle.com/datasets/utkarshx27/smoking-dataset-from-uk/) So I use 'read.csv()' function to read in the data. I found that all continuous variables with value of 0 are compiled as N/A, so I changed these N/A values into 0. All the EDA and statistical analysis were performed in R by visualizing and interpreting the data with suitable graphs and summary tables.

Preliminary Results

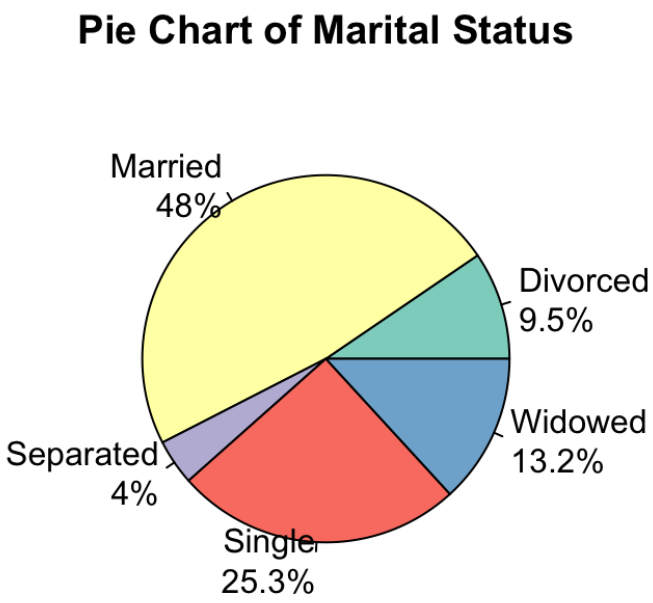
Exploratory data analysis



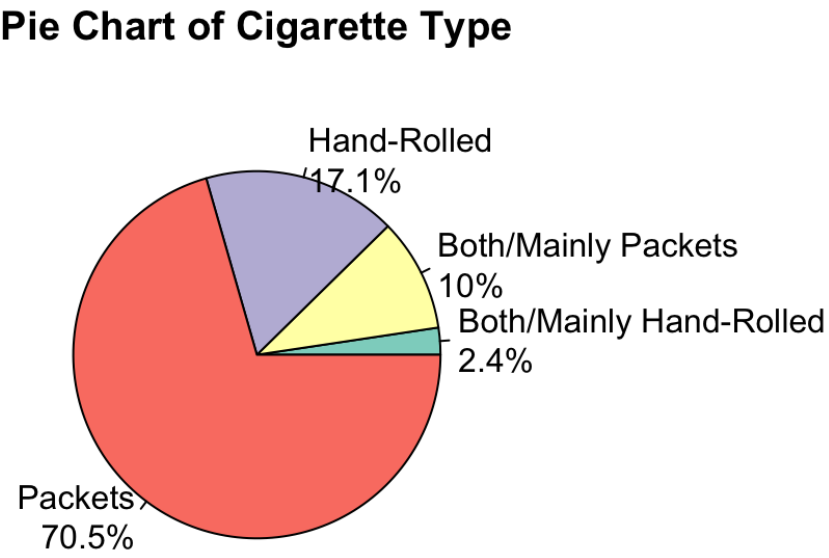
The data mainly covers the people aged under 20 to 80 years old plus small amount of people over 80. The distribution of age is relatively uniformed.



About 25% of the people in this dataset smoke.

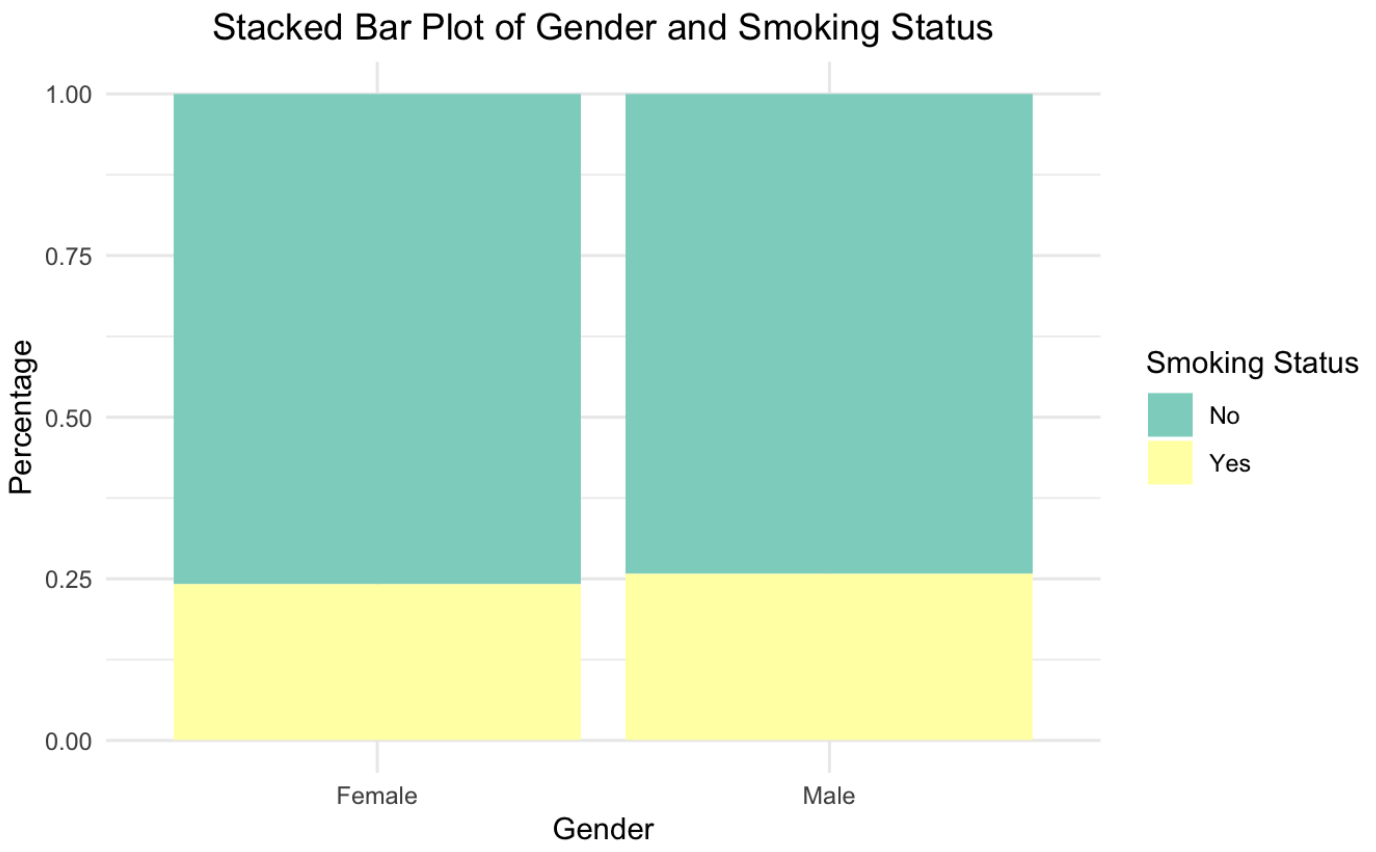


Almost 50% of the people in this dataset are married and around 25% of them are single.

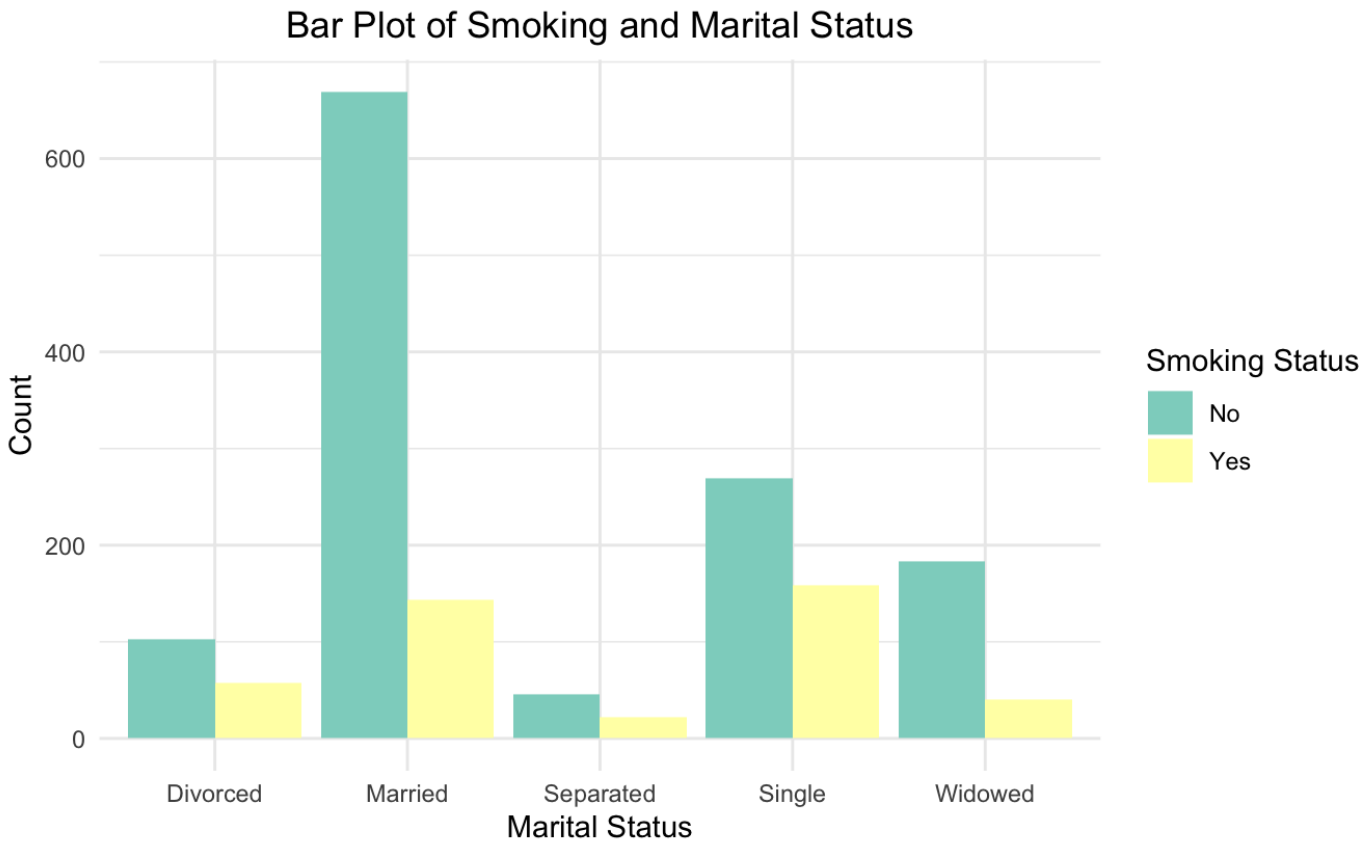


Around 80% of the smokers in this dataset only or mainly choose cigarettes in packets.

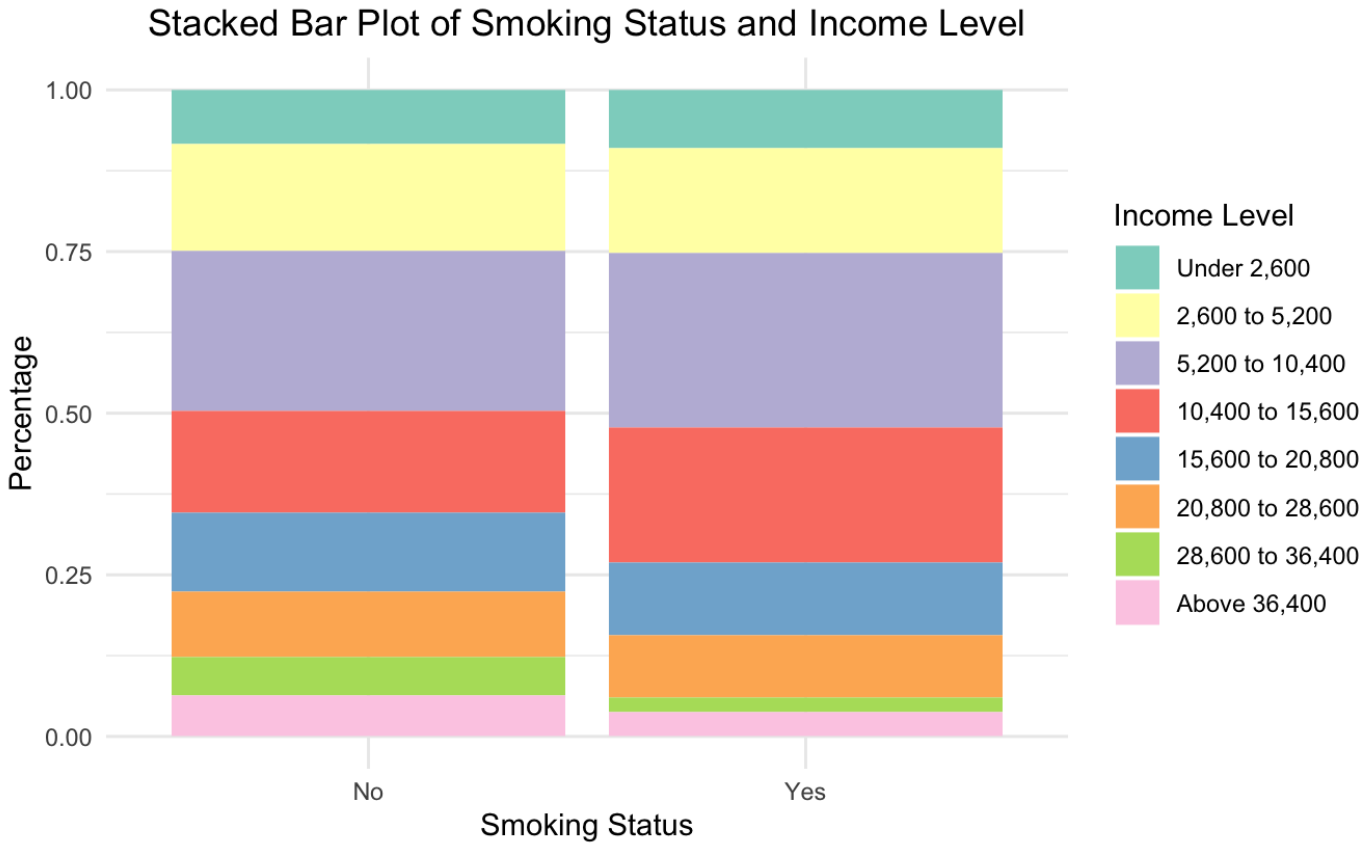
Statistical Analysis



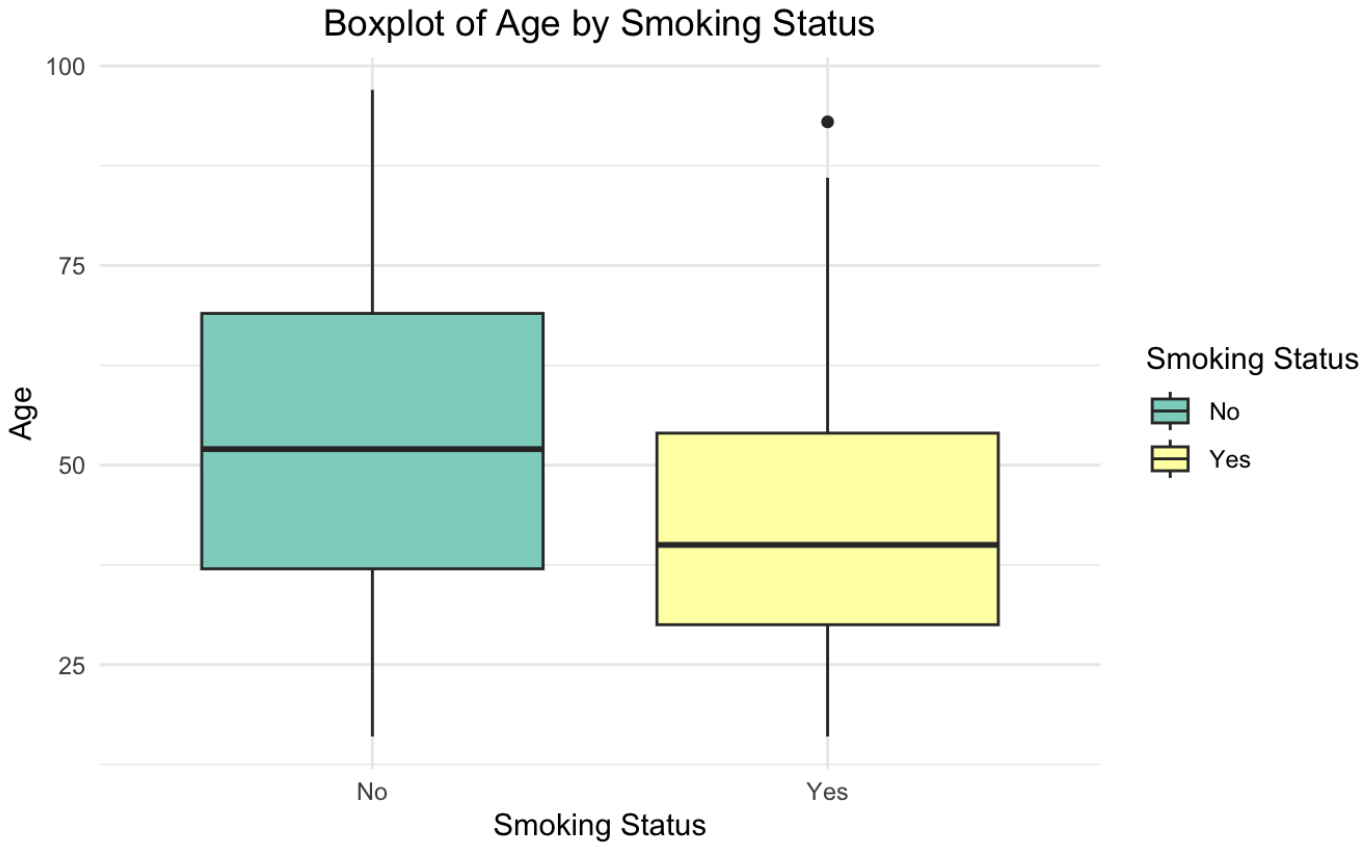
There is no obvious difference in smoking status distribution between Female and Male, which indicates that gender is not a main factor in smoking status.



There are more non-smokers than smokers in every marital status. In these marital status, married people and widowed people are more likely not to smoke than any other marital status.



Just like gender, There is no obvious difference in two distributions of income level between smokers and non-smokers, which indicates that income level is not a main factor in smoking status.
However, if we only focus on the income level distribution of smokers, we found that people with high incomes (over 28,600) only accounts for a small proportion of smokers. This may suggest that people with higher incomes are more likely not to smoke.
It is worth noting that the dataset is relatively small, so there may not be enough data to look at the actual effect of income level on number of cigarettes smoked per day.

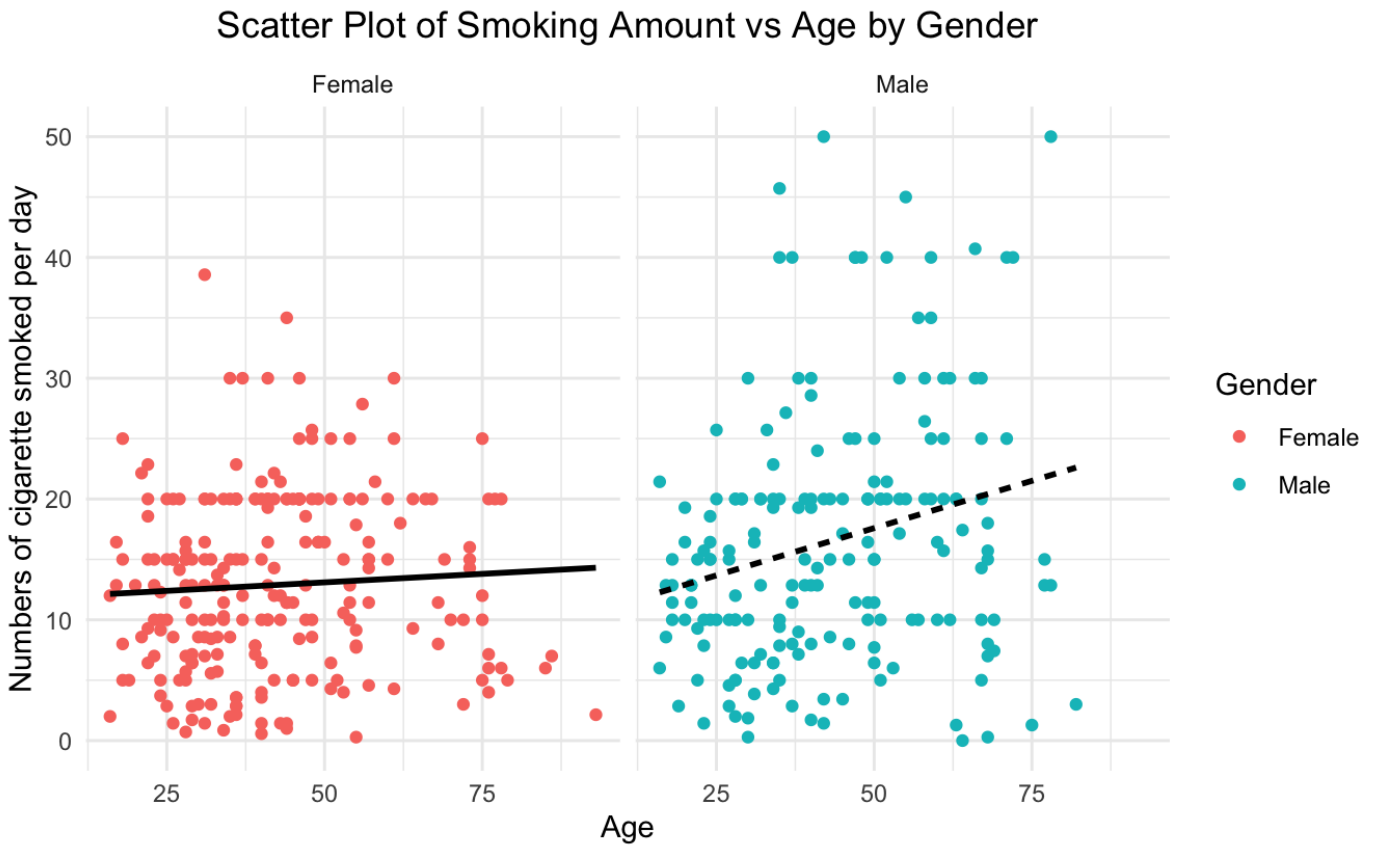


From the boxplot, we can see that younger people are more likely to smoke than older people.

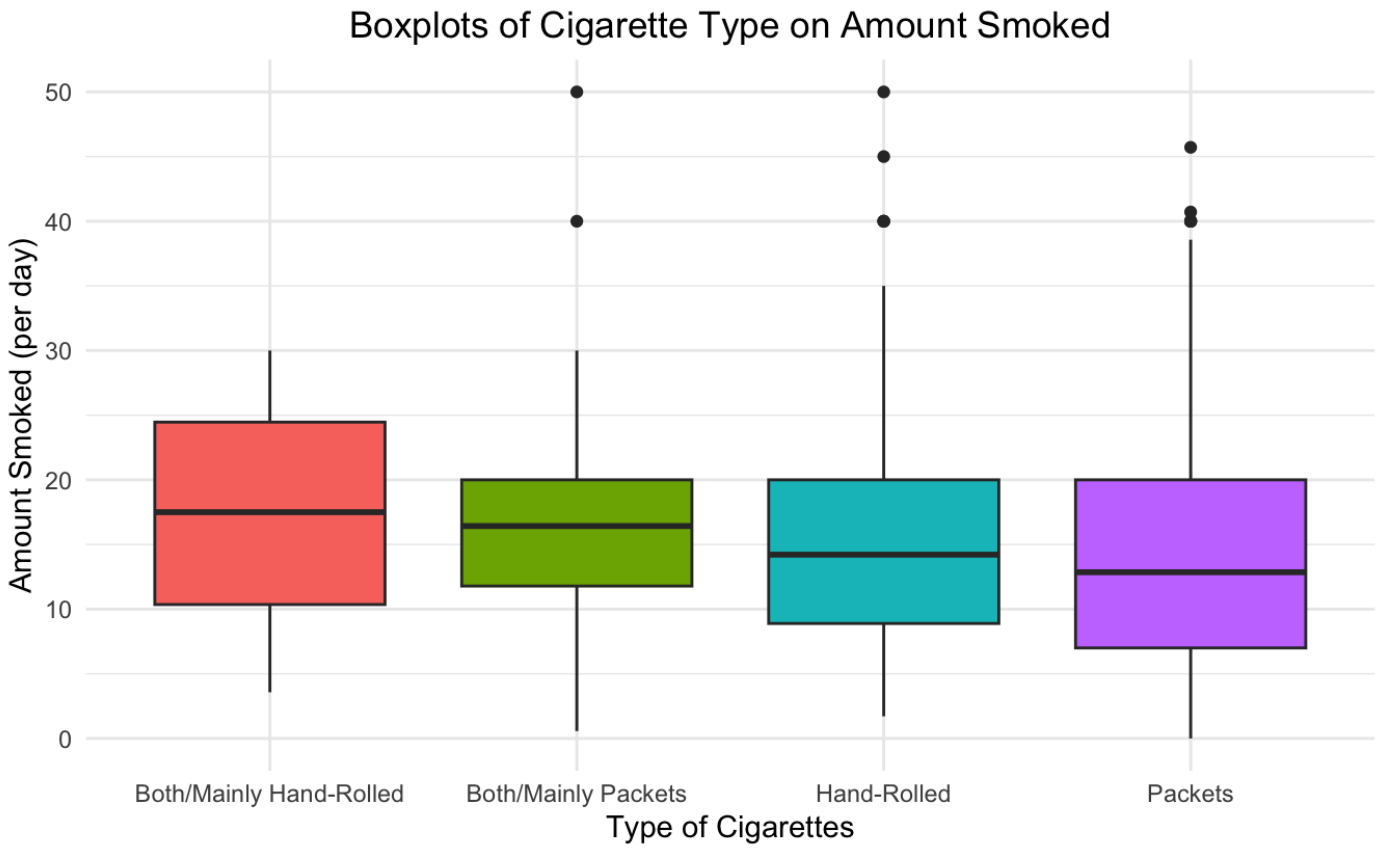
Mean Number of Cigarettes by Marital Status

Marital Status	Weekdays	Weekends	Counts
Divorced	14.58621	17.75862	58
Married	14.26573	16.14685	143
Separated	13.68182	15.59091	22
Single	12.36709	16.13924	158
Widowed	16.20000	16.92500	40

From the summary table, we found that people smoke more cigarettes on weekends than weekdays. This showed that whether it is on a working day has an impact on the number of cigarettes people smoke per day, especially for divorced and single people. Widowed people smoke most on weekdays and divorced people smoke most on weekends.



From the scatter plot, we found that age has greater impact on daily cigarette intake among male smokers. Older male smokers smoke more cigarettes everyday than younger male smokers. For female smokers, age has a slight positive effect on the number of cigarettes smoked per day.



From the boxplots, we found that people who only smoked packets cigarettes would smoke the least amount of cigarettes. People who smoked hand-rolled cigarettes, regardless of whether they still smoked packets, would smoke more amount of cigarettes.

Conclusion

From my analysis, age is a main factor on people's smoking status. Younger people are more likely to smoke than older people. As to marital status, married people and widowed people are less likely to smoke than any other marital status. Also, people with higher incomes only accounts for a small proportion of smokers. This may suggest that people with higher income are more likely not to smoke.

Whether it is on a working day has an impact on the number of cigarettes people smoke per day, especially for divorced and single people. People smoke more cigarettes on weekends than weekdays. Age also has a positive effect on daily smoke intake, especially for male smokers. Older male smokers tend to smoke more cigarettes per day than younger male smokers. Also, people who smoked hand-rolled cigarettes tend to smoke more cigarettes than those who did not.