Fundamentals of Information Visualisation (COMP3021 UNUK) (AUT2 23-24)

Coursework Report

~ By Brendan Ezekiel Agnelo Vaz

Student Id – 20610206

MSc Data Science

University of Nottingham

Introduction

The aim of the assignment is to implement visualisations on a particular dataset using the R programming language. Visualisations will be made to answer questions and get meaningful insights from the extensive data at hand. The dataset chosen for this analysis is the SmokeBan.csv file from the vincentarelbundock.github.io repository. It is a cross-sectional dataset with observations on 10,000 indoor workers, which is a subset of a 18,090-observation data set collected as part of the National Health Interview Survey in 1991 and then again (with different respondents) in 1993. The dataset focuses on the topic, "Do Workplace Smoking Bans Reduce Smoking The 8 different variables are as follows:

1. rownames (integer) – Consists of the row number of the people evaluated (1,2,3,4,5 etc). It is of class integer.
2. smoker (character) – Is the individual a current smoker?
3. ban (character) – Is there a work area smoking ban?
4. age (integer) – The individuals age in years.
5. education (character) – Factor that indicates the highest education level attained by the individual:
    - High school (hs) drop out.
    - High school graduate.
    - Some College.
    - College Graduate.
    - Master's degree (or higher).

6. afam (character) – Is the individual African American?
7. Hispanic (character) – Is the individual Hispanic (a person whose background is from a Spanish speaking nation).
8. gender (character) – Factor indicating the gender of the person (Male of Female).

SmokeBan dataset: https://vincentarelbundock.github.io/Rdatasets/doc/AER/SmokeBan.html

Libraries

A few visualization libraries had to be installed and initialized before moving on to analysing the data. These are 'ggplot2' plotting package and 'dplyr' package for data wrangling. The package 'tidyverse' is also installed. It is a collection of R packages which includes functions of both ggplot2 and dplyr packages making it easier to work with data.

Once downloaded, the libraries are initialized by:

- library(ggplot2)
- library(dplyr)
- library(tidyverse)

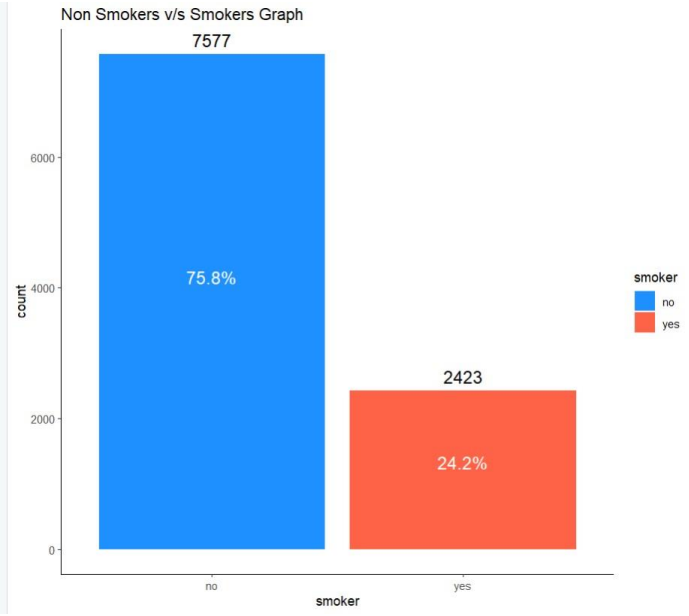
On the completion of these steps, the libraries can be used to plot graphs.

Data Cleaning

During the data cleaning phase, the dataset was checked for any missing values; however, none were found. The data type of two variables (rownames and age) is 'integer' whereas the remaining variables belong to the 'character' data type. No transformations were performed on the dataset. Once done with data cleaning, the analysis proceeded further without any issues.

Question 1 - Is there a majority of smokers or non-smokers in the dataset?

To understand the proportion of people who smoke in the dataset, a graph is used to make a comparison between the non-smokers and smokers.

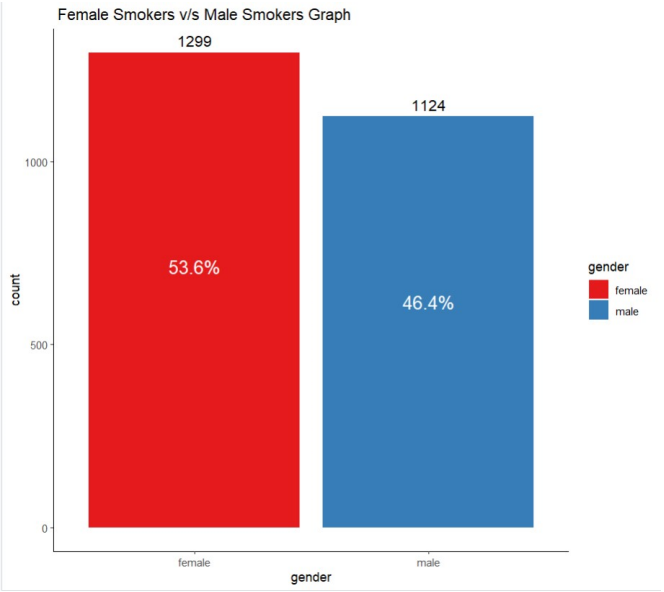


The visualization used is a Bar graph as it clearly depicts the difference between the number of people who smoke versus those who do not. On the X axis, the smoking status is observed, whereas the Y axis represents the number of records in the dataset. Different colours for “yes” and “no” helps differentiate both the categories from each other.

The Dataset was already tested for missing or null values previously. Having found zero missing values, the data analysis is accurate. The result of the code showed a graph wherein the number of non-smokers heavily outweighed that of smokers. In conclusion, the Smokeban dataset comprises of a majority of non-smokers.

Question 2 - Is there a surplus of male or female smokers?

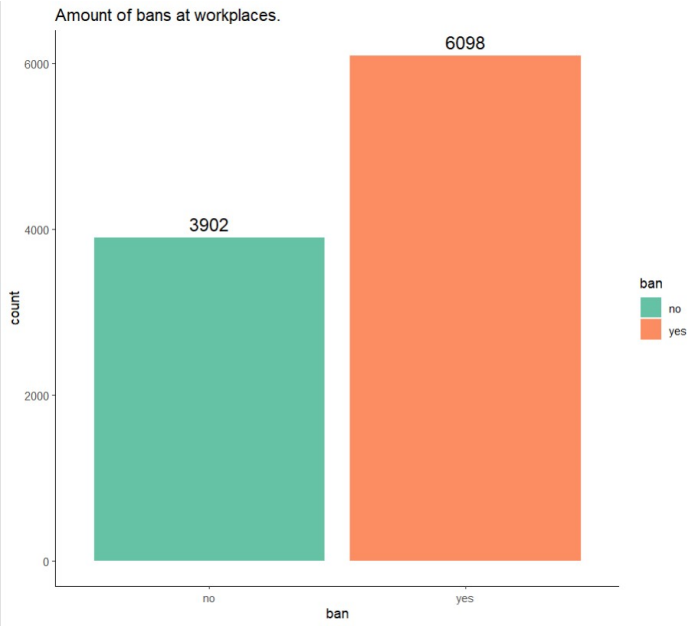
While further exploring the dataset, a differentiation is made among the male and female individuals who smoke. To begin with, the dataset was altered to include only the individuals who are smokers. A bar graph is chosen to compare the data as it is categorical.



The X – axis represents the gender of the individual which is either male or female, and the Y-axis represents the count of smokers. Both the genders are depicted using two distinct colours in order to provide a clear visual distinction between both male and female. The visualisation shows that there is a surplus of female smokers when compared to male smokers. However, the difference between the two is quite nominal.

Question 3 - How many smoking bans exist in the dataset?

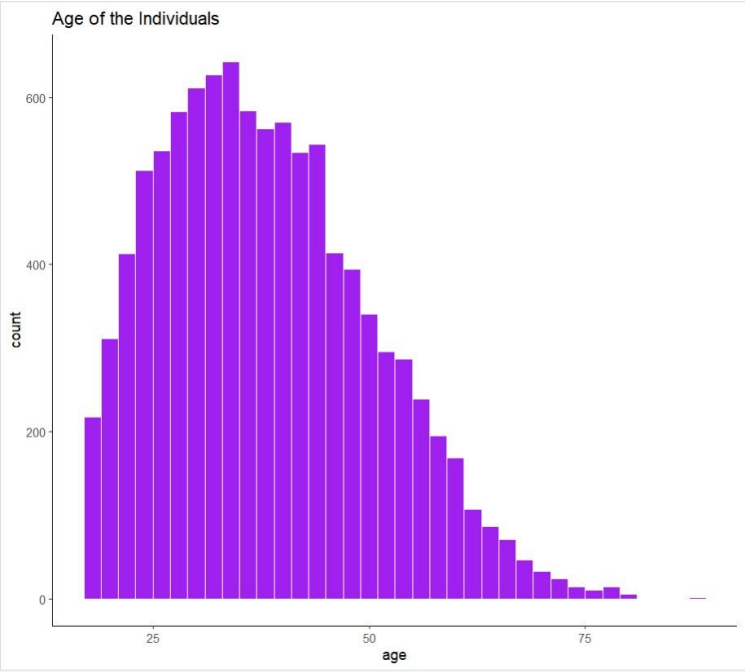
The aim of the question is to find out how many bans have been put in place at workplaces. The bar graph below makes a comparison between the workplaces that have a ban vs those that do not.



The visualisation shows that there is a majority in the number of bans put at workplaces. The difference is quite large, and it shows how there a there are strict rules put in place against smoking.

Question 4 - What age group do most people belong to?

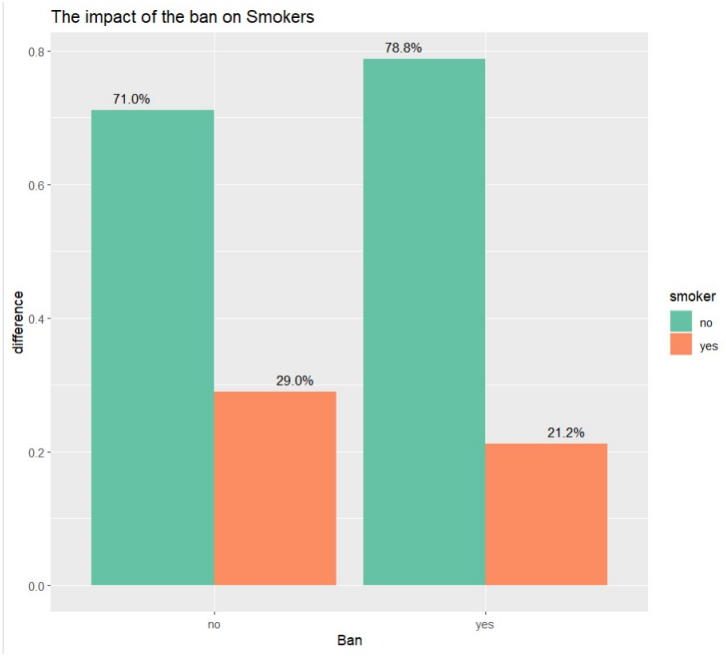
An understanding of the average age of the people in the data set is required. This is plotted using a histogram which also depict trends that can be seen. On the X-axis we have the age of the individuals and on the Y-axis we have the count of the individuals.



By observing the histogram, it is noticeable that majority of the people in the data set belong to the age group of 25 to 50. Additionally the numbers keep decreasing as age increases and very minute population of the dataset Princess of elderly folk.

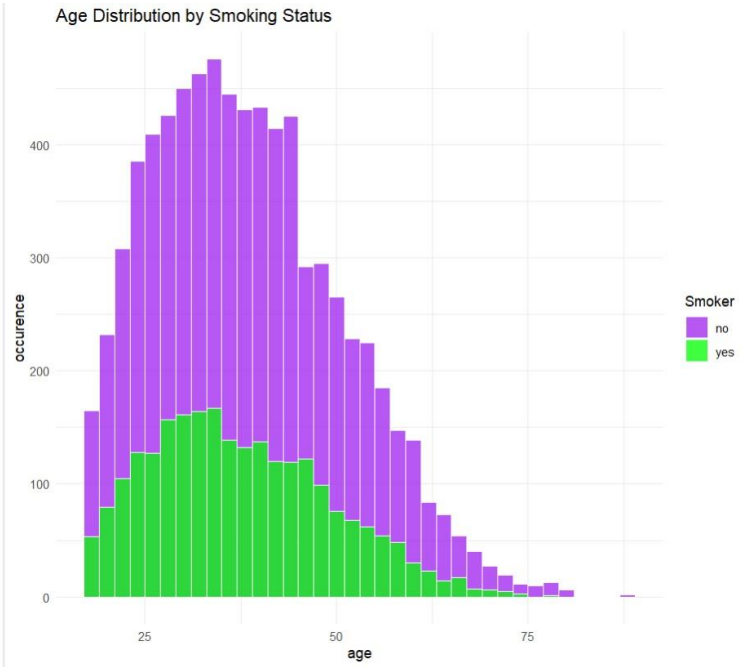
Question 5 - Has the smoking ban affected the number of smokers?

Delving deeper into the data set we need to visualise how the ban has affected the number of smokers at workplaces.



The X axis shows the ban put in place whereas the Y axis shows the difference before and after the ban. The result is that the ban did have an effect and the number of smokers witnessed a decline from 29% to 21.2%. Furthermore, the number of non-smokers saw an increase from 71% to 78.8%.

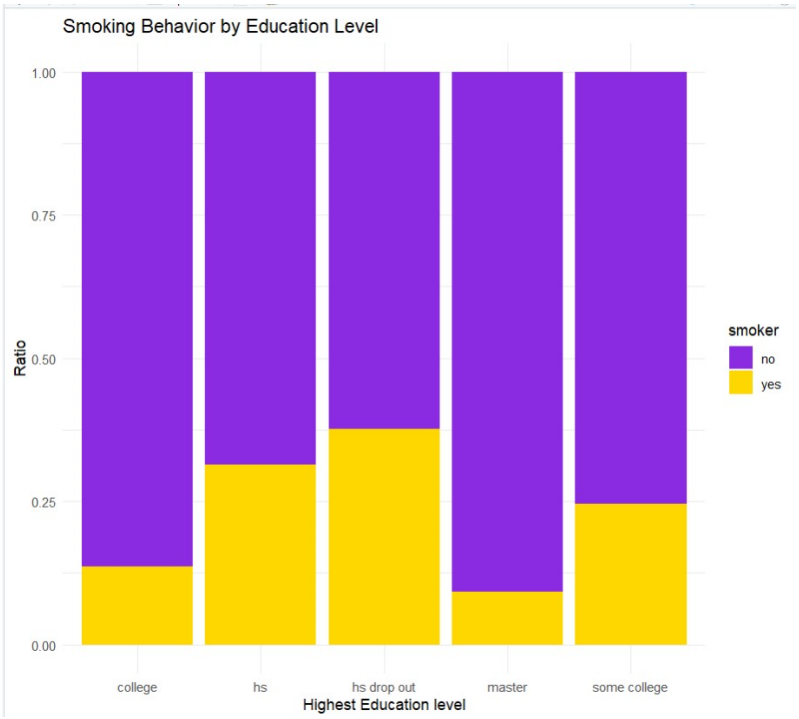
Question 6 - How do smoking habits change with age? Is there a trend seen as age increases?



As age increases, the number of smokers experiences a sharp decline in numbers. The graph shows a gradual descend indicating that people over the age of 50 do not turn to smoking as much as the younger generation.

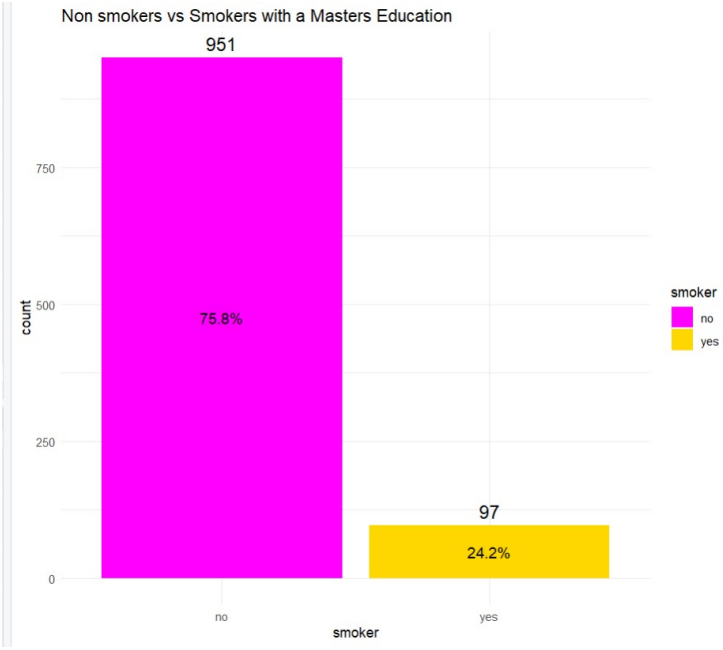
Question 7 - What is the level of education between smokers? Do higher educated people smoke lesser?

The education level of individuals have a huge impact on whether this smoke or not. In the data set provided there are people with 5 types of highest education level which are as follows: college graduates, higher secondary graduates, higher secondary dropouts, master's graduates people who have gone to some college or the other.



The graph illustrates that higher secondary dropout students have a higher tendency to smoke when compared to students who have better education. In the second half of the question analysis will be made on the smokers versus non smokers people having a masters education.

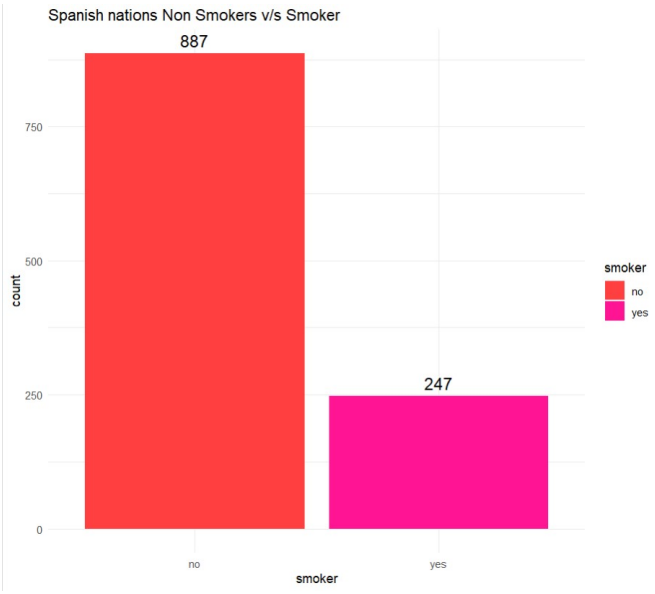
Data is filtered to only have people with masters education as a variable and the smoking habits of these graduates is compared.



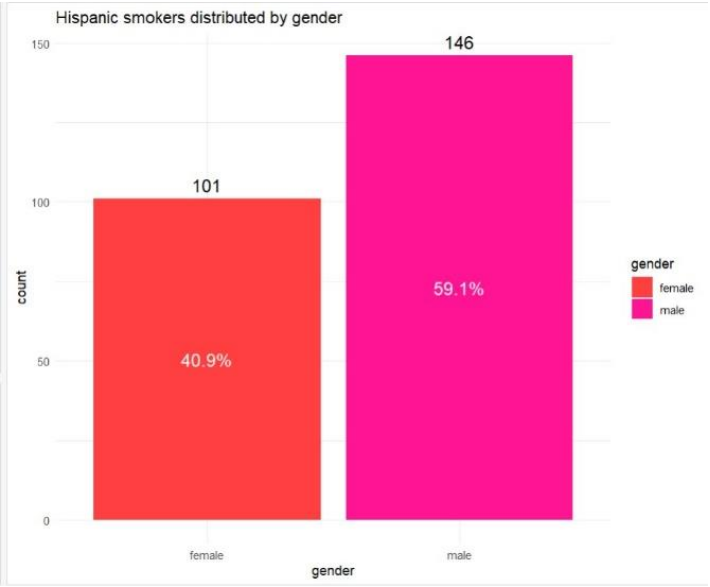
The graph depicts that only a minute population of individuals with a masters education a smokers where's the majority belong to the non-smokers bar graph. The difference is quite high having 75.8% non-smokers versus 24.2% smokers. This confirms that having a good education plays a huge role and whether the individuals smokes or not.

Question 8 – Is smoking common in the Spanish countries? Which gender smokes more in Spanish countries?

Background plays a major role in peoples lives. The graph depicts the number of non-smokers vs smokers of Hispanic people (people from a Spanish speaking nation).



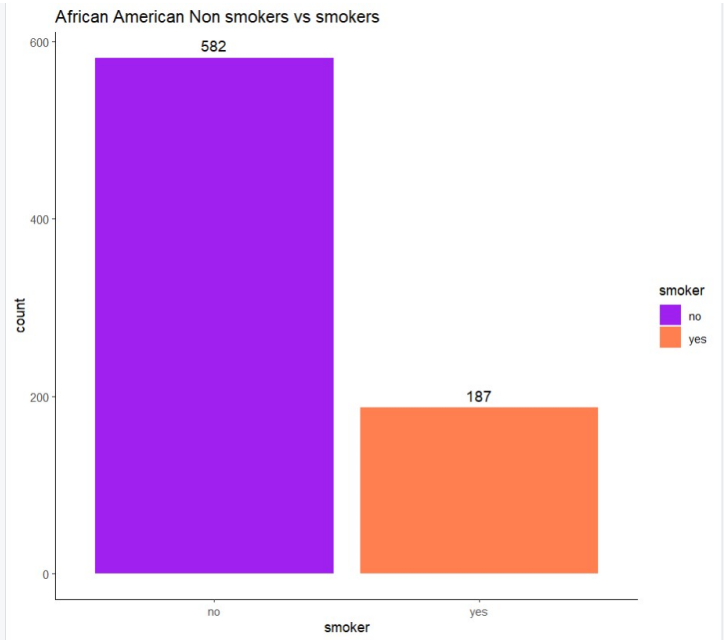
The graph illustrates that there a majority of non-smokers among Hispanic people. However, this data might be vague. The percentage of male and female is checked to determine whether it's a balanced ratio or one gender outweighs the other in terms of smoking among Hispanic people.



The result shows that there is a slightly larger number of female smokers at 53.6% than the male smokers at 46.4%. However this difference is quite negligible.

Question 9-Do African-American people have a large number of smokers?

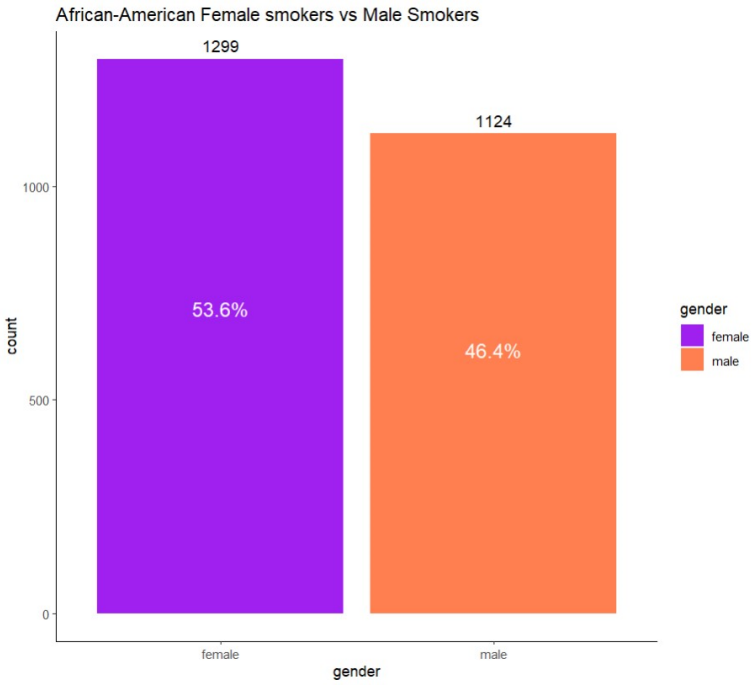
The analysis is now done on the African-American population. The number of non-smokers vs smokers is compared using different colours. A bar graph is used to visualize this.



The graph depicts that there is a large number of non- smokers among the African-American folk.

Q10 - Do the African-American females smoke more than the males?

A comparison is made between the Afam smokers based on gender. A bar graph is used to depict this and both the genders have different colours making it easy to understand.



The observation here is that African -Americans have a larger number of female smokers when compared to male smokers.

Conclusion:

All in all, the analysis of the dataset proved that there was significant effect of having the ban in place. The amount of Smokers decreased which results in a positive effect in their health. Furthermore, it is also observed that having a solid educational background plays a major role in whether an individual resorts to smoking or not. The higher secondary dropouts witnessed the largest number of smokers when compared to the other people with higher educational backgrounds. People who had a Masters Education or higher showed a lower percentage of smokers by a huge margin.

While backgrounds do play a major role in the smoking habits of people, it was very interesting to get a deeper insight of the data by comparing the genders. Among Hispanic people, a higher number of male smokers is observed whereas there is a higher percentage of female African- American smokers.