



Central Regional Datathon 2021

Team 3

Aidan Chi
Andrew Li
Michael Chew
Pramod Prem



Presented by **Citadel** and **Citadel Securities**
In Partnership with **CorrelationOne**

Background



Tobacco use is currently the leading cause of preventable death in the United States. Yet approximately 19% of the adults in the world still smoke tobacco. However, when traveling to different areas around the world it is evident that smoking is more prevalent in some places compared to others. We wanted to find out how tobacco usage varies in different regions of the world as well as why this is the case by looking at some factors that could influence tobacco usage.



Our Question

How does Tobacco Usage Vary Between
Different Regions of the World and
What is the Reasoning Behind This?

The 6 Regions



Western Pacific



Africa



Europe



South-East Asia



Eastern Mediterranean



Americas

List of Countries In Each Region Used in our Datasets

Western Pacific	South-East Asia	Europe	Eastern Mediterranean	Africa	Americas
Australia Brunei Darussalam Cambodia China Cook Islands Fiji Japan Kiribati Lao People's Democratic Republic Malaysia Mongolia Nauru New Zealand Palau Philippines Republic of Korea Samoa Singapore Solomon Islands Tonga Tuvalu Vanuatu	Bangladesh Democratic People's Republic of Korea India Indonesia Myanmar Nepal Sri Lanka Thailand Timor-Leste	Albania, Andorra Armenia, Austria Azerbaijan, Belarus Belgium, Bosnia & Herzegovina, Bulgaria, Croatia Cyprus, Czechia Denmark, Estonia Finland, France Georgia, Germany Greece, Hungary Iceland, Ireland Israel, Italy Kazakhstan, Kyrgyzstan Latvia, Lithuania Luxembourg, Malta Netherlands, Norway Poland, Portugal Republic of Moldova Romania, Russian Federation, Serbia Slovakia, Slovenia Spain, Sweden Switzerland, Turkey Ukraine, United Kingdom of Great Britain & Northern Ireland, Uzbekistan	Bahrain Egypt Iran (Islamic Republic of) Iraq Kuwait Lebanon Morocco Oman Pakistan Qatar Saudi Arabia Tunisia United Arab Emirates Yemen	Algeria, Benin Botswana, Burkina Faso, Burundi Cameroon, Chad Comoros, Congo Côte d'Ivoire, Eritrea Eswatini, Ethiopia Gambia, Ghana Kenya, Lesotho Liberia, Madagascar Malawi, Mali, Mauritius, Mozambique, Namibia, Niger Nigeria, Rwanda Sao Tome and Principe, Senegal Seychelles, Sierra Leone, South Africa Togo, Uganda United Republic of Tanzania, Zambia Zimbabwe	Argentina Bahamas Barbados Brazil Canada Chile Colombia Costa Rica Cuba Dominican Republic El Salvador Guyana Haiti Jamaica Mexico Panama Paraguay Peru United States of America Uruguay

Datasets Used

- **tobacco_use.csv:**
 - Contains the percentage of the population that is at least 15 years old who currently use any tobacco product for each of the countries listed
 - This will be our tobacco usage value given as a percent
- **tobacco_production.csv:**
 - Provides data about the number of metric tons of tobacco that was produced for each of the countries listed
- **stop_smoking.csv:**
 - Gives us the following information for each of the countries listed:
 - Average price of a pack of 20 cigarettes in international dollars for the three most sold brands of cigarettes and weighted by market share
 - Sum of taxes on cigarettes on the weighted average price for the three most sold brands
 - Enforcement of bans on tobacco ads based on levels (Level 1: data not reported; Levels 2 - 5: no bans to most amount of bans)
 - Aids to help quit tobacco based on levels (Level 1: data not reported; Levels 2 - 5: no services to most amount of services)
 - Amount of tobacco produced in metric tons

Our Process



Data Manipulation:

- We combined these three datasets into one comprehensive CSV file for males and one comprehensive CSV file for females as gender is a confounding variable
- For each of the datasets, we only incorporated data from the year 2014 in order to be consistent and use data values from the same year across the three datasets

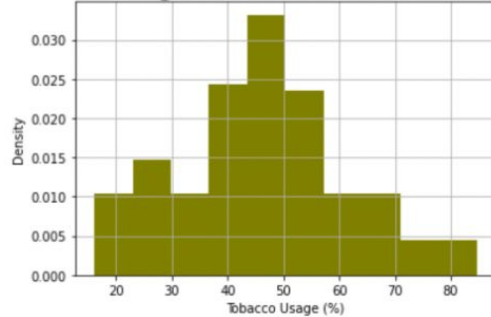
Steps:

- First, we created density histograms for each of the six regions for both males and females and analyzed the results
 - This will tell us how tobacco usage varies by gender in different regions of the world
- Afterwards, we created linear regression models comparing five different factors with tobacco usage for both males and females and analyzed the results
 - By determining if a factor had a strong correlation with tobacco usage, we hope to identify a reason as to why tobacco usage varies

Tobacco Usage By Region (Males)



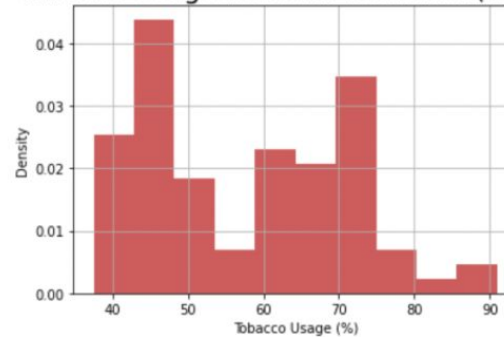
Tobacco Usage in the Western Pacific (Males)



Mean Tobacco Usage: 46.336%

SD of Tobacco Usage: 14.762%

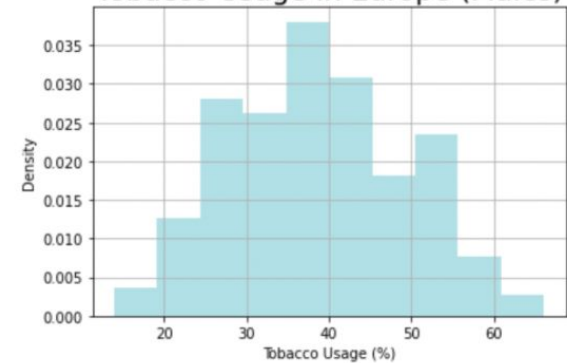
Tobacco Usage in South-East Asia (Males)



Mean Tobacco Usage: 57.321%

SD of Tobacco Usage: 13.638%

Tobacco Usage in Europe (Males)



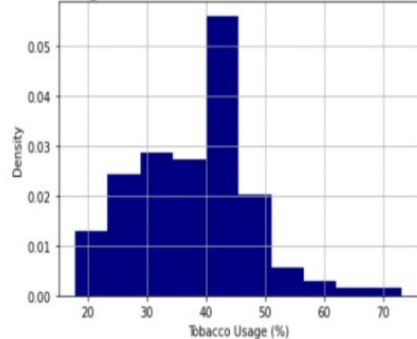
Mean Tobacco Usage: 38.600%

SD of Tobacco Usage: 10.602%

Tobacco Usage By Region (Males) Cont.



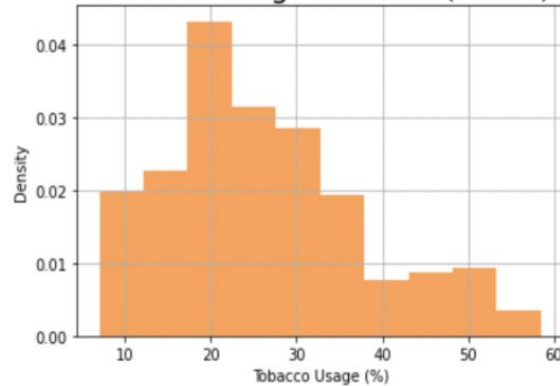
Tobacco Usage in the Eastern Mediterranean (Males)



Mean Tobacco Usage: 37.698%

SD of Tobacco Usage: 10.125%

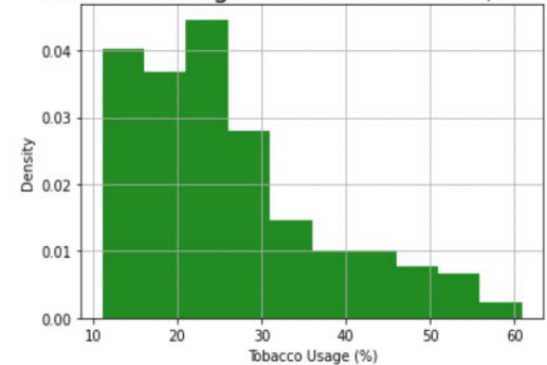
Tobacco Usage in Africa (Males)



Mean Tobacco Usage: 26.182%

SD of Tobacco Usage: 11.621%

Tobacco Usage in the Americas (Males)



Mean Tobacco Usage: 28.096%

SD of Tobacco Usage: 11.430%



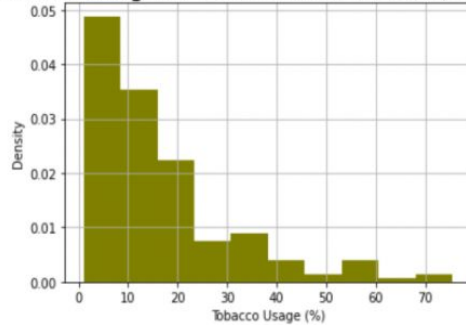
Tobacco Usage By Region (Males) Conclusions

- The order of average tobacco usage by region for males from highest to lowest is:
 1. South-East Asia
 2. Western Pacific
 3. Europe
 4. Eastern Mediterranean
 5. Africa
 6. Americas
- The region with the highest average had a tobacco usage of about 57% while the region with the lowest average had a tobacco usage of about 26%.
- The standard deviation of tobacco usage for each region are relatively the same which means that the spread in tobacco usage for males in each region are similar

Tobacco Usage By Region (Females)



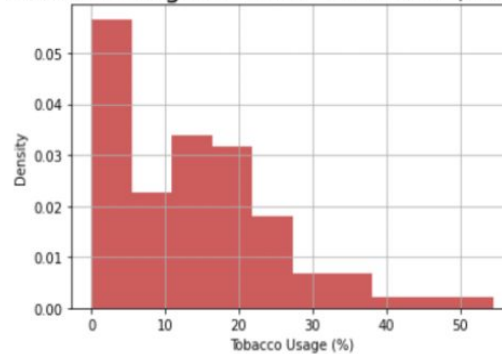
Tobacco Usage in the Western Pacific (Females)



Mean Tobacco Usage: 16.312%

SD of Tobacco Usage: 14.224%

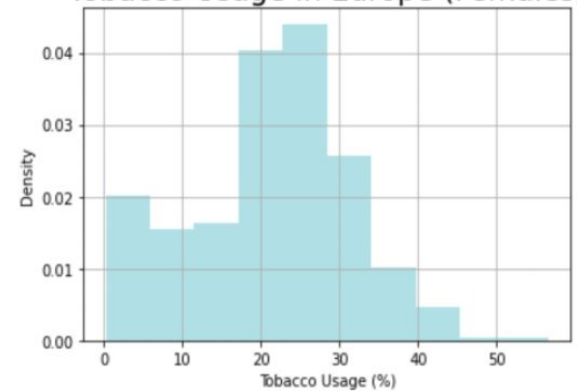
Tobacco Usage in South-East Asia (Females)



Mean Tobacco Usage: 13.870%

SD of Tobacco Usage: 11.538%

Tobacco Usage in Europe (Females)



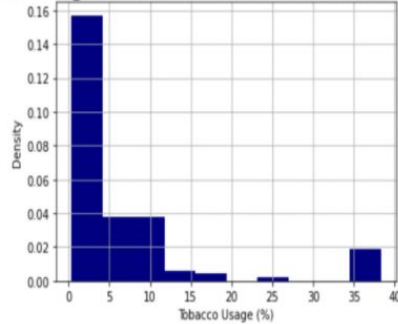
Mean Tobacco Usage: 21.258%

SD of Tobacco Usage: 10.228%

Tobacco Usage By Region (Females) Cont.



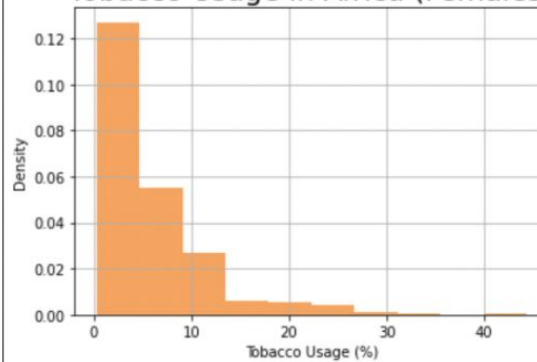
Tobacco Usage in the Eastern Mediterranean (Females)



Mean Tobacco Usage: 6.740%

SD of Tobacco Usage: 9.223%

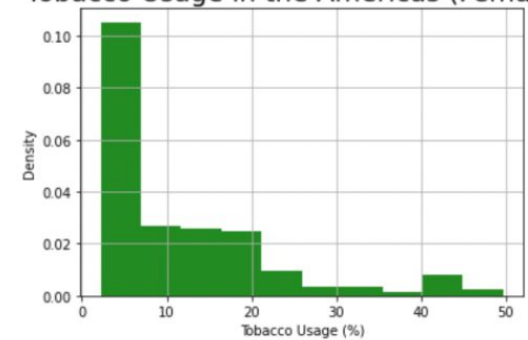
Tobacco Usage in Africa (Females)



Mean Tobacco Usage: 5.832%

SD of Tobacco Usage: 6.033%

Tobacco Usage in the Americas (Females)



Mean Tobacco Usage: 11.583%

SD of Tobacco Usage: 10.583%

Tobacco Usage By Region (Females) Conclusions



- The order of average tobacco usage by region for females from highest to lowest is:
 1. Europe
 2. Western Pacific
 3. South-East Asia
 4. Americas
 5. Eastern Mediterranean
 6. Africa
- The region with the highest average per country had a tobacco usage of about 21% while the region with the lowest average per country had a tobacco usage of about 5%.
- Females are shown to have a considerably lower tobacco usage compared to males in every region
 - The average difference in tobacco usage between males and females per region is about 26%
- The standard deviations of tobacco usage are more spread out compared to males which means that there is more variance in tobacco usage for females in different regions

Reasoning Behind Difference in Tobacco Usage

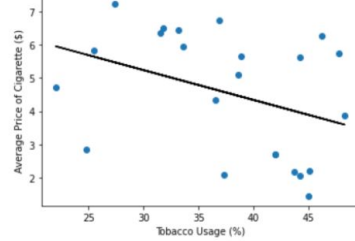


- Now that we know how tobacco usage varies by gender in different regions of the world, we would like to find out what is the reasoning behind this. The five factors we decided to look at are:
 - Average price of cigarette
 - Sum of taxes on cigarette
 - Level of enforcement of bans on cigarette ads
 - Level of aids to help quit tobacco
 - Tobacco production
- We decided to use linear regression in order to illustrate the relationship between each of these factors with tobacco usage, this time not separating by region as we are just looking for the relationship. However, we will still separate the data by gender as it is a confounding variable.

Linear Regression Models (Males)



Average Price of Cigarette vs Tobacco Usage (Males)



Equation of line:

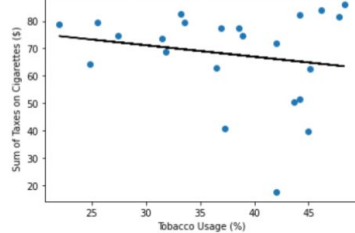
$$\hat{y} = -0.089599x + 7.922509$$

Correlation Coefficient:

-0.231727



Sum of Taxes on Cigarettes vs Tobacco Usage (Males)



Equation of line:

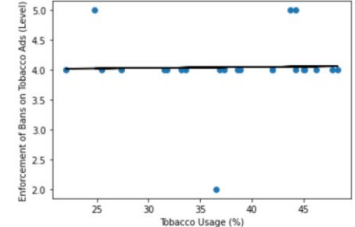
$$\hat{y} = -0.419180x + 83.7486604$$

Correlation Coefficient:

0.251605



Enforcement of Bans on Tobacco Ads vs Tobacco Usage (Males)



Equation of line:

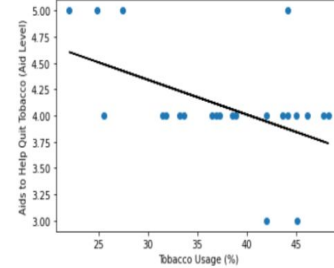
$$\hat{y} = 0.0015179x + 3.986292$$

Correlation Coefficient:

0.150872



Aids to Help Quit Tobacco vs Tobacco Usage (Males)



Equation of line:

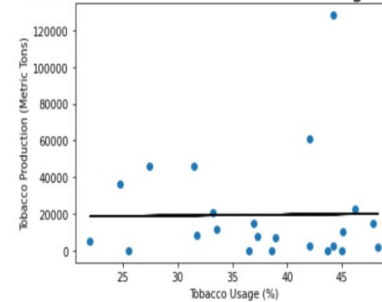
$$\hat{y} = -0.0330005x + 5.330213$$

Correlation Coefficient:

-0.084455



Tobacco Production vs Tobacco Usage (Males)



Equation of line:

$$\hat{y} = 51.249624x + 17536.559634$$

Correlation Coefficient:

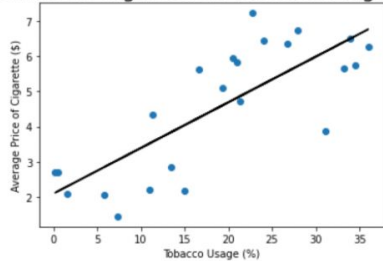
-0.279595



Linear Regression Models (Females)



Average Price of Cigarette vs Tobacco Usage (Females)

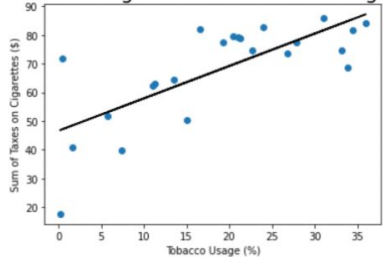


Equation of line:
 $\hat{y} = .129579x + 2.095668$

Correlation Coefficient:
 0.374553



Sum of Taxes on Cigarettes vs Tobacco Usage (Females)

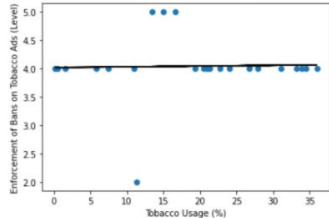


Equation of line:
 $\hat{y} = 1.129767x + 46.584272$

Correlation Coefficient:
 0.657741



Enforcement of Bans on Tobacco Ads vs Tobacco Usage (Females)

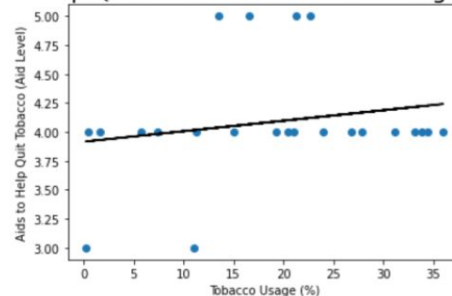


Equation of line:
 $\hat{y} = 0.0012924x + 4.0190296$

Correlation Coefficient:
 0.194098



Aids to Help Quit Tobacco vs Tobacco Usage (Females)

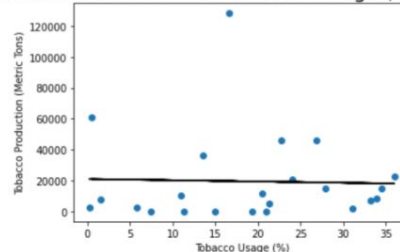


Equation of line:
 $\hat{y} = 0.009042x + 3.915905$

Correlation Coefficient:
 0.129639



Tobacco Production vs Tobacco Usage (Females)



Equation of line:
 $\hat{y} = -82.69667x + 21031.7387$

Correlation Coefficient:
 -0.161903




Conclusions from Linear Regression Models



- For males, all five factors had a weak correlation with tobacco usage as their correlation coefficients were all within $\pm (0, 0.3)$.
 - This means that none of these factors had a strong relationship with tobacco usage.
- For females, three of the five factors had a weak correlation. However, the average price of a cigarette had a moderate correlation with tobacco usage as its correlation coefficient was between $\pm (0.3, 0.6)$ while the sum of taxes on cigarettes had a strong correlation with tobacco usage as its correlation coefficient was above ± 0.6 .
 - This means that the sum of taxes on cigarettes had a relatively strong relationship with tobacco usage for females while all of the other factors had much weaker relationships with tobacco usage.

Final Conclusions



In the beginning, we raised the question: How does Tobacco Usage Vary Between Different Regions of the World and What is the Reasoning Behind This? Here is what we found:

- ★ South-East Asia, Europe, and the Western Pacific had the highest tobacco usage for both males and females
- ★ Females are illustrated to smoke much less than males in all regions
- ★ As for the reasons behind this, we were not able to find a strong correlation between a singular factor that we considered and the tobacco usage for both males and females
- ★ This means that increasing the price, increasing the tax, increasing the enforcement of bans on tobacco ads, increasing the number of aids to help quit tobacco, and decreasing tobacco production in a country does not have that much of an effect on the country's tobacco usage
- ★ We will have to consider other factors that can have an effect on tobacco usage in order to find a stronger reasoning as to why tobacco usage varies