Behavioral Risk Factor Data :

Tobacco Use

Analyzing smoking habits from 2011 to 2022

OVERVIEW

The dataset is obtained from an official government website-Data.gov.
Dataset link:
https://catalog.data.gov/dataset/behavioral-risk-factor-data-tobacco-use-2011-to-present
The dataset extracted from annual BRFSS(Behavioral Risk Factor Surveillance System) surveys.

Columns available

0	Year	138048 non-null	int64
1	Locationabbr	138048 non-null	object
2	Locationdesc	138048 non-null	object
3	Class	138048 non-null	object
4	Topic	138048 non-null	object
5	Question	138048 non-null	object
6	Response	138048 non-null	object
7	Break_Out	138048 non-null	object
8	Break_Out_Category	138048 non-null	object
9	Sample_Size	138048 non-null	int64
10	Data_value	109167 non-null	float64
11	Confidence_limit_Low	108963 non-null	float64
12	Confidence_limit_High	108963 non-null	float64
13	Display_order	138018 non-null	float64
14	Data_value_unit	138048 non-null	object
15	Data_value_type	138048 non-null	object
16	Data_Value_Footnote_Symbol	28917 non-null	object
17	Data_Value_Footnote	28917 non-null	object
18	DataSource	138048 non-null	object
19	ClassId	138048 non-null	object
20	TopicId	138048 non-null	object
21	LocationID	138048 non-null	int64
22	BreakoutID	138048 non-null	object
23	BreakOutCategoryID	138048 non-null	object
24	QuestionID	138048 non-null	object
25	ResponseID	138048 non-null	object
26	GeoLocation	137844 non-null	object

Objectives

1

- **Trends based on Gender and Age**
  - What gender of people smoke more? How has it changed over the years?
  - Analysing smoking habits among various age groups.

2

- **Trends based on Education level**
  - Does higher education translate to awareness about smoking?
  - Analysing nicotine consumption based on education levels and household income

3

- **Trends based on Race**
  - Does a relation exist between ethnicity and nicotine use?
  - How have smoking levels changed over different races over the years?

4

- **Trends based on Geolocation**
  - Is there any correlation between Geolocation and smoking habits?

Cleaning the Data

Filtereing out unnecessay columns

```python
df=df.filter(['Year', 'Locationabbr', 'Locationdesc','Topic','Response','Break_Out_Category','Break_Out','Sample_Size',
       'Data_value', 'Confidence_limit_Low', 'Confidence_limit_High','GeoLocation'])
df.head()
```

Python

Year	Locationabbr	Locationdesc	Topic	Response	Sample_Size	Data_value	Confidence_limit_Low	Confidence_limit_High	GeoLocation	Age Group	Education Attained	Gender	Household Income	Race/Ethnicity
2019	AK	Alaska	Smokeless Tobacco	Not at all	68	100.0	100	100	(64.84507995700051, -147.72205903599973)	Overall	Less than H.S.	Overall	Overall	Hispanic, American Indian or Alaskan Native, n...
2019	WI	Wisconsin	Smokeless Tobacco	Some days	8	NaN	0	0	(44.39319117400049, -89.81637074199966)	45-54, 65+	Overall	Female	$15,000-$24,999	Hispanic
2019	AK	Alaska	Smokeless Tobacco	Some days	11	NaN	0	0	(64.84507995700051, -147.72205903599973)	18-24	Less than H.S.	Overall	$15,000-$24,999	Black, non-Hispanic
2019	AZ	Arizona	Smokeless Tobacco	Every day	16	NaN	0	0	(34.865970280000454, -111.76381127699972)	25-34	Less than H.S.	Overall	Less than $15,000, $25,000-$34,999, $35,000-$4...	Hispanic
2019	AL	Alabama	Smoker Status	Smoke everyday	119	5.2	4	6	(32.84057112200048, -86.63186076199969)	Overall	College graduate	Overall	Overall	Overall

Setting up right datatypes

```
df['Locationdesc']=df['Locationdesc'].astype(str)
df.info()
```

```
df['Confidence_limit_High']=df['Confidence_limit_High'].astype(float)
df['Confidence_limit_Low']=df['Confidence_limit_Low'].astype(float)
```

```
def convertToGeoLocation(x: str)->(float,float):
    t=[]
    for i in x[1:-1].split(','):
        t.append(float(i))
    return tuple(t)
df['GeoLocation']=df['GeoLocation'].apply(convertToGeoLocation)
```

Converting LocationDesc to string, Confidence levels to float and geolocation to tuple of floats.

Resolving multiple values into singular value

Replacing multiple values with singular value.
Dropping null value rows

```
df.dropna(inplace=True)
df.reset_index()
```

```
for index, row in df.iterrows():
    if ',' in row['Gender']:
        df.at[index,'Gender']='Overall'

for index, row in df.iterrows():
    if ',' in row['Education Attained']:
        df.at[index,'Education Attained']='Overall'
```

2022	AL	Alabama	Current Smoker Status	No	163	90.1	84.6	95.6	(32.84057112200048, -86.63186076199969)	18-24	Overall	Overall	Overall	Overall
2022	AL	Alabama	Current Smoker Status	No	341	80.5	75.1	85.9	(32.84057112200048, -86.63186076199969)	25-34	Overall	Overall	Overall	Overall
2022	AL	Alabama	Current Smoker Status	No	394	81.5	77.4	85.6	(32.84057112200048, -86.63186076199969)	35-44	Overall	Overall	Overall	Overall
2022	AL	Alabama	Current Smoker Status	Yes	103	17.7	13.9	21.4	(32.84057112200048, -86.63186076199969)	45-54	Overall	Overall	Less than $15,000	Overall

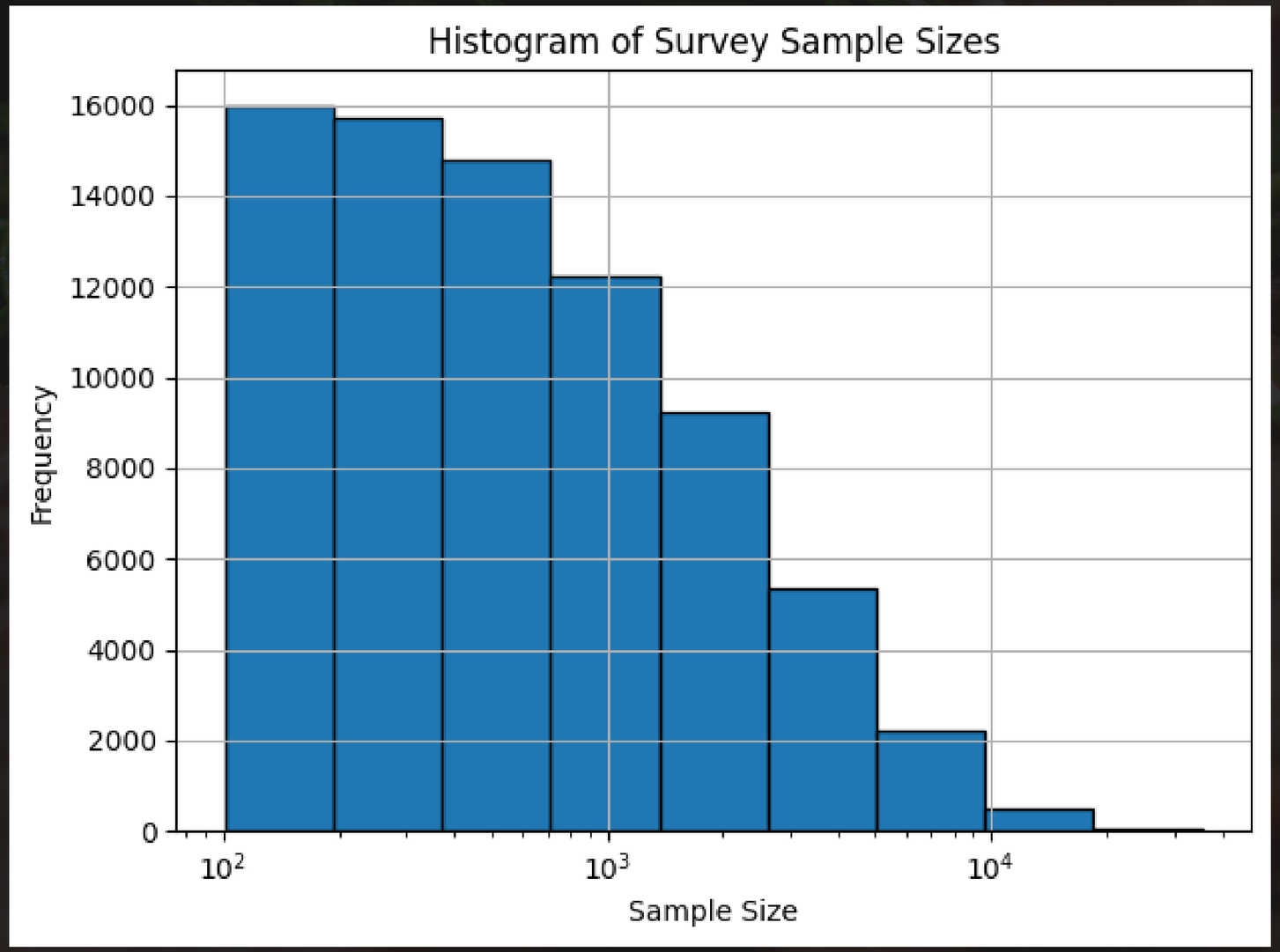
Analyzing the data

Count and mean value of data value based on Topic and Responses

		count
		Data_value
Topic	Response	
Current Smoker Status	No	13944
	Yes	10769
Smokeless Tobacco	Every day	1189
	Not at all	11863
	Some days	651
Smoker Status	Former smoker	11373
	Never smoked	13490
	Smoke everyday	9023
	Smoke some days	3942

```
pd.pivot_table(df,values='Data_value',index=['Topic','Response'],aggfunc=['count'])
```

Sample Size Distribution



Count and mean value of data value based on Topic and Responses

Topic	Response	Confidence_limit_High	Confidence_limit_Low	Data_value
Current Smoker Status	No	84.837055	78.044019	81.441903
	Yes	21.839772	16.260851	19.050683
Smokeless Tobacco	Every day	4.741548	3.176619	3.960892
	Not at all	97.519658	94.014128	95.785594
	Some days	2.990937	1.881567	2.435637
Smoker Status	Former smoker	28.313989	22.440438	25.377139
	Never smoked	61.917324	54.018933	57.968332
	Smoke everyday	15.997972	11.499889	13.748742
	Smoke some days	6.299087	4.129554	5.213927

```
pd.pivot_table(df,values=['Confidence_limit_Low','Confidence_limit_High','Data_value'],index=['Topic','Response'],aggfunc='mean')
```

1

Trends based on Gender and Age

- What gender of people smoke more? How has it changed over the years?
- Analyzing smoking habits among various age groups.

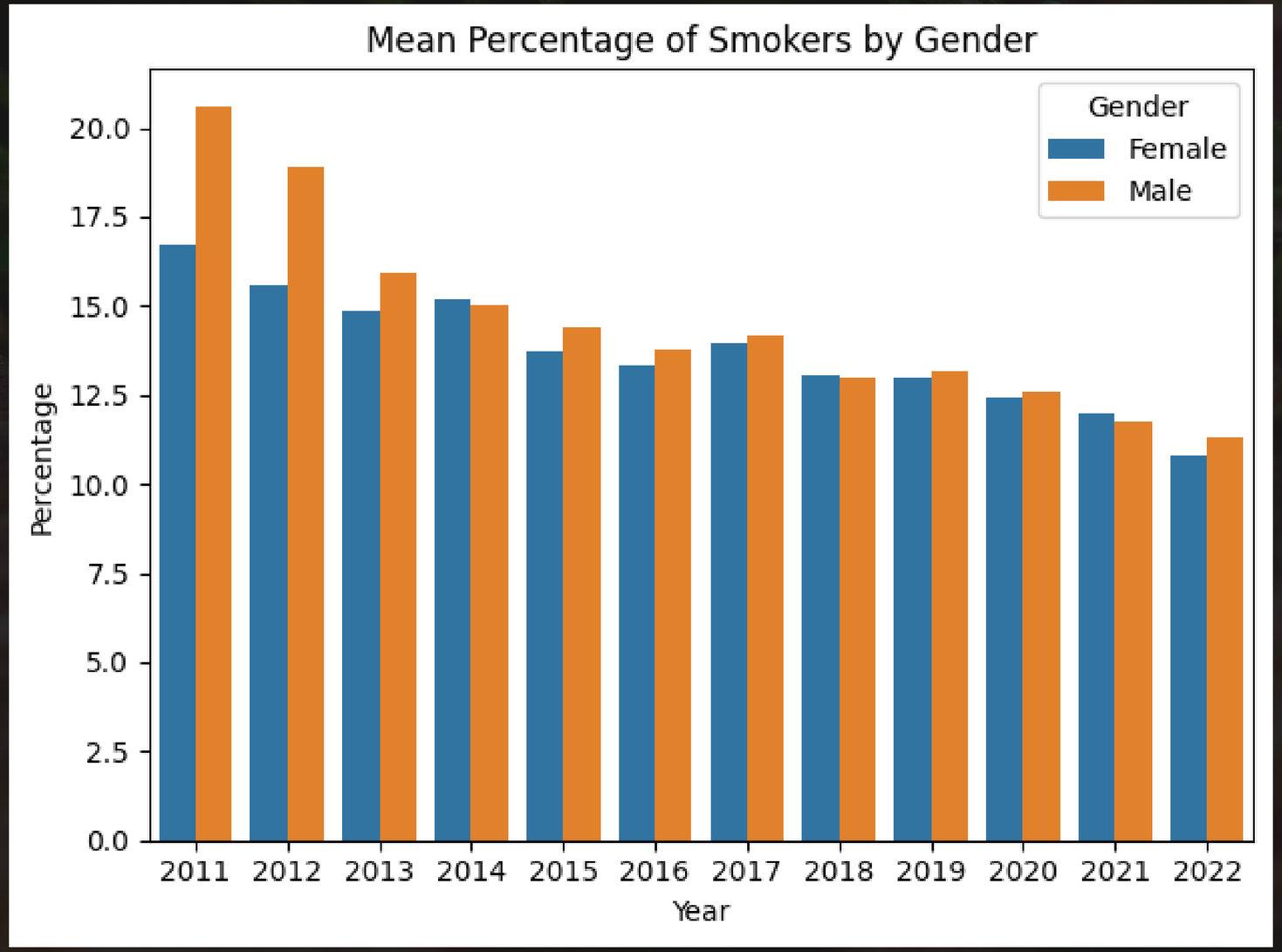
Trends based on Gender

Smoker Status represented with respect to Gender

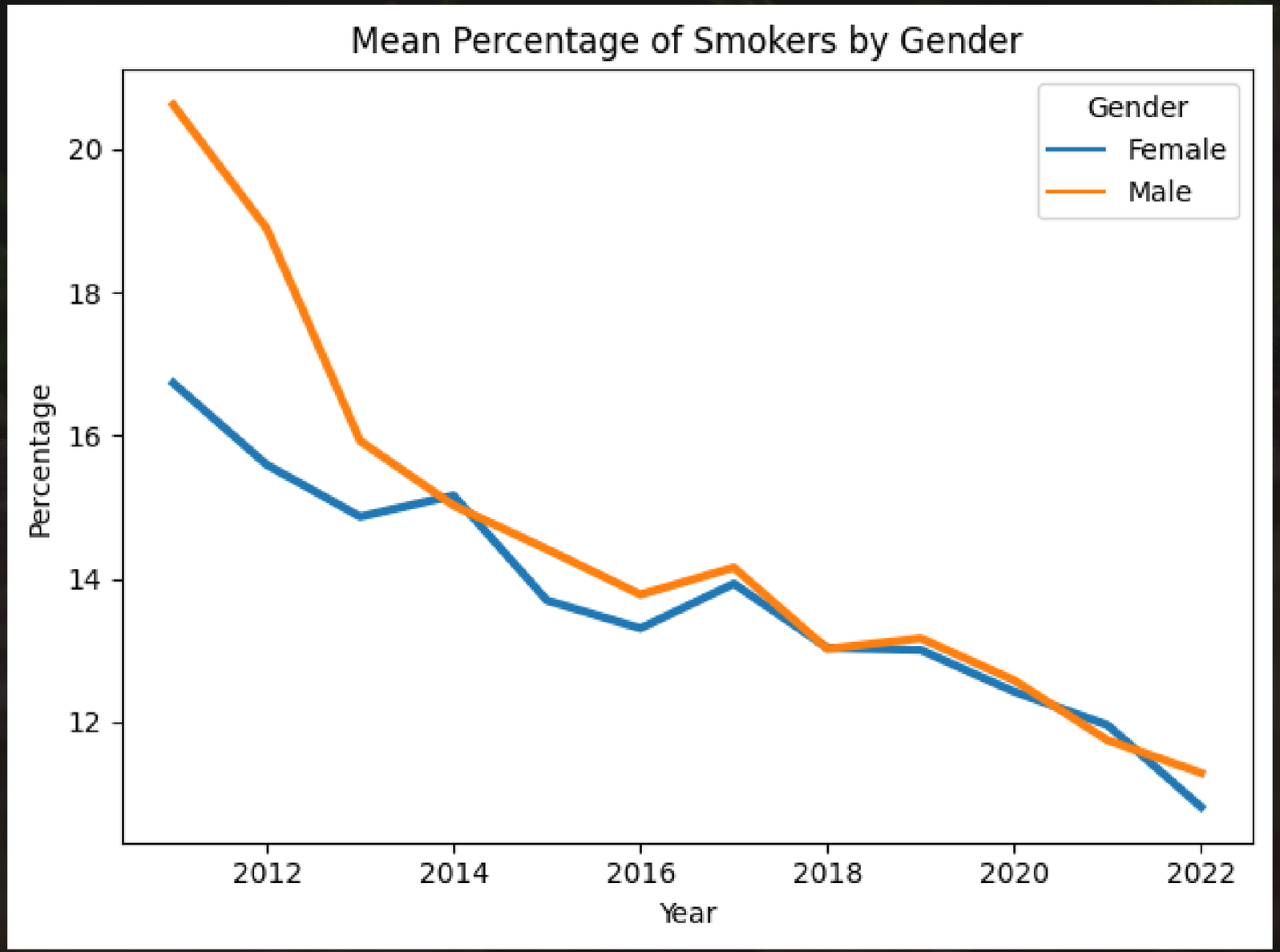
			Data_value
Gender	**Topic**	**Response**	
Female	Current Smoker Status	No	84.061201
		Yes	16.078645
	Smokeless Tobacco	Every day	2.120000
		Not at all	98.775045
		Some days	1.825000
	Smoker Status	Former smoker	21.656231
		Never smoked	62.786578
		Smoke everyday	11.424848
		Smoke some days	4.443811
Male	Current Smoker Status	No	80.881259
		Yes	19.350578
	Smokeless Tobacco	Every day	5.510061
		Not at all	92.713904
		Some days	3.431892
	Smoker Status	Former smoker	27.695665
		Never smoked	53.178426
		Smoke everyday	13.537266
		Smoke some days	5.851845

```
pd.pivot_table(df,values='Data_value',index=['Gender','Topic','Response'],aggfunc='mean')
```

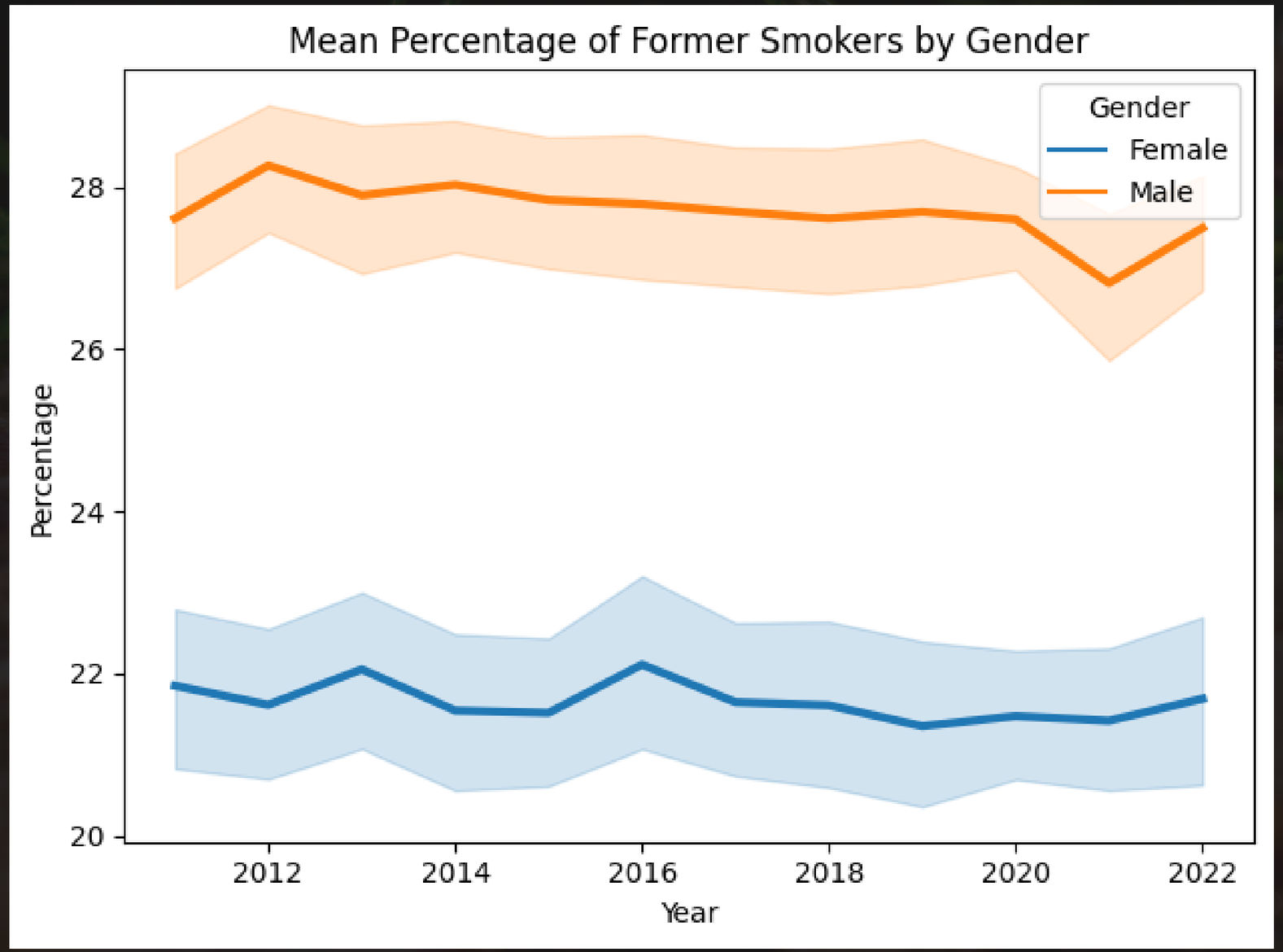
Distribution of Nicotine users over the years differentitated by Gender



Distribution of Genders on Nicotine use over the years



Distribution of Former Smokers over the years



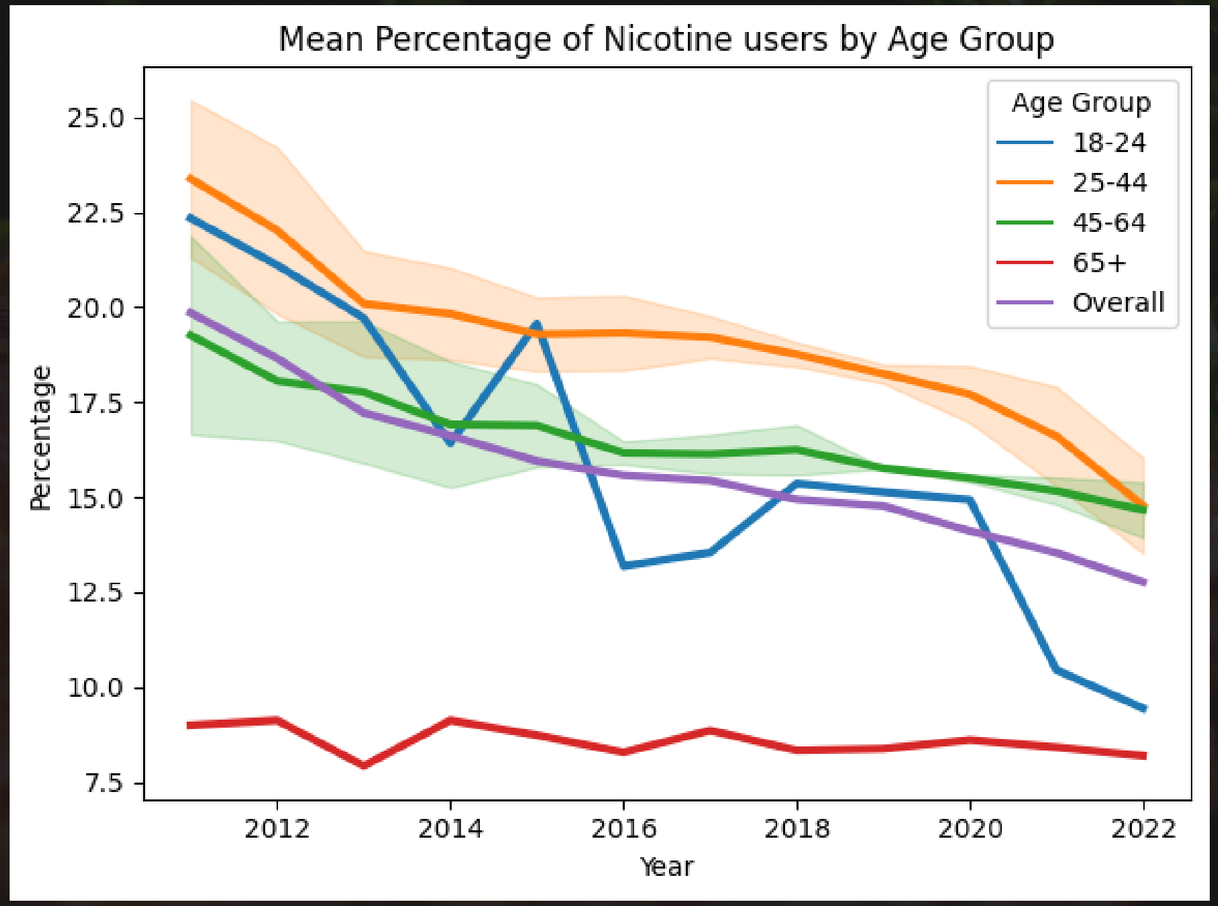
Trends based on Age

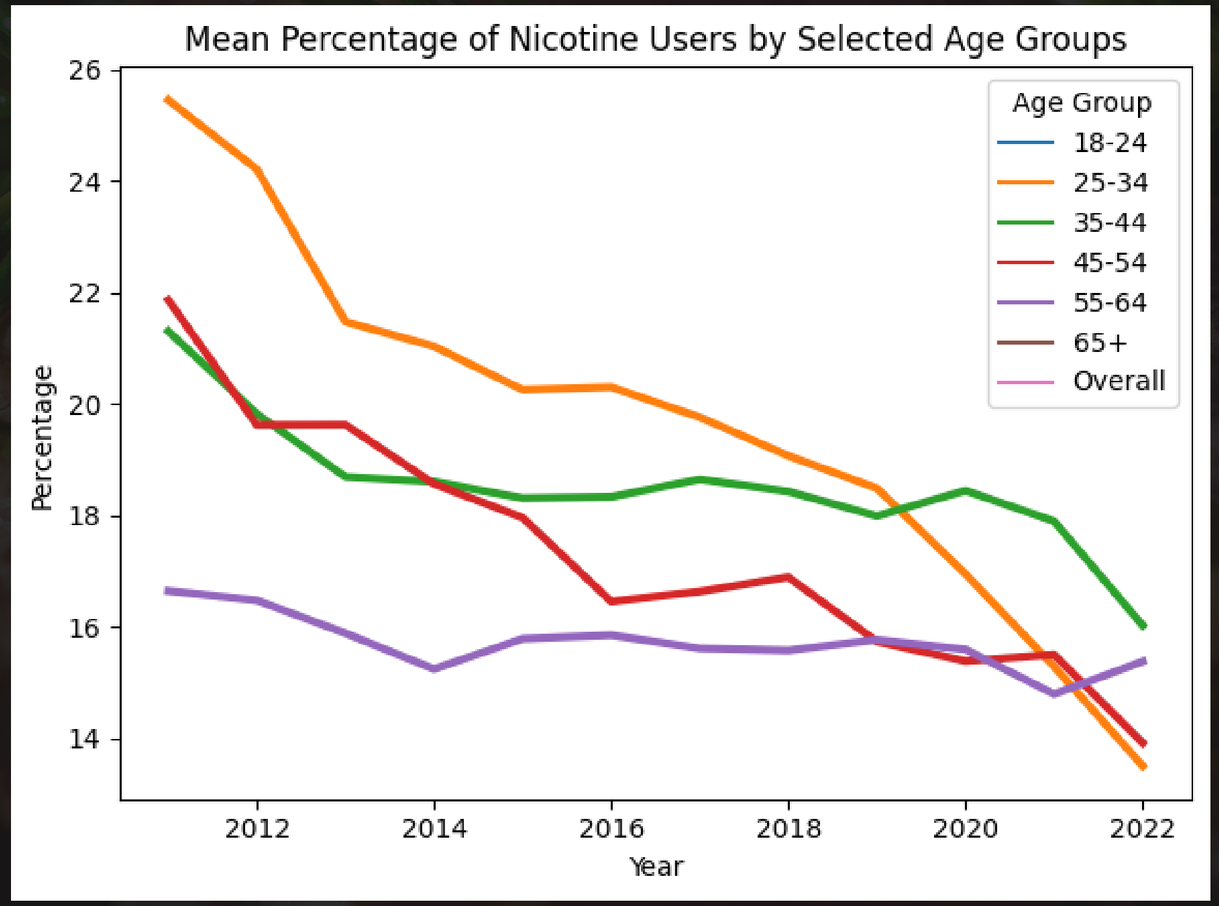
Mean percentage of smokers in each Age Group

Age Group	Confidence_limit_Low	Data_value	Confidence_limit_High
18-24	14.922751	18.230688	21.539153
25-34	16.919222	20.316705	23.718764
35-44	15.425490	18.618768	21.816993
45-54	14.676675	17.490500	20.297286
55-64	13.275802	15.739095	18.204444
65+	7.010422	8.578167	10.144295
Overall	13.512947	15.872640	18.231636

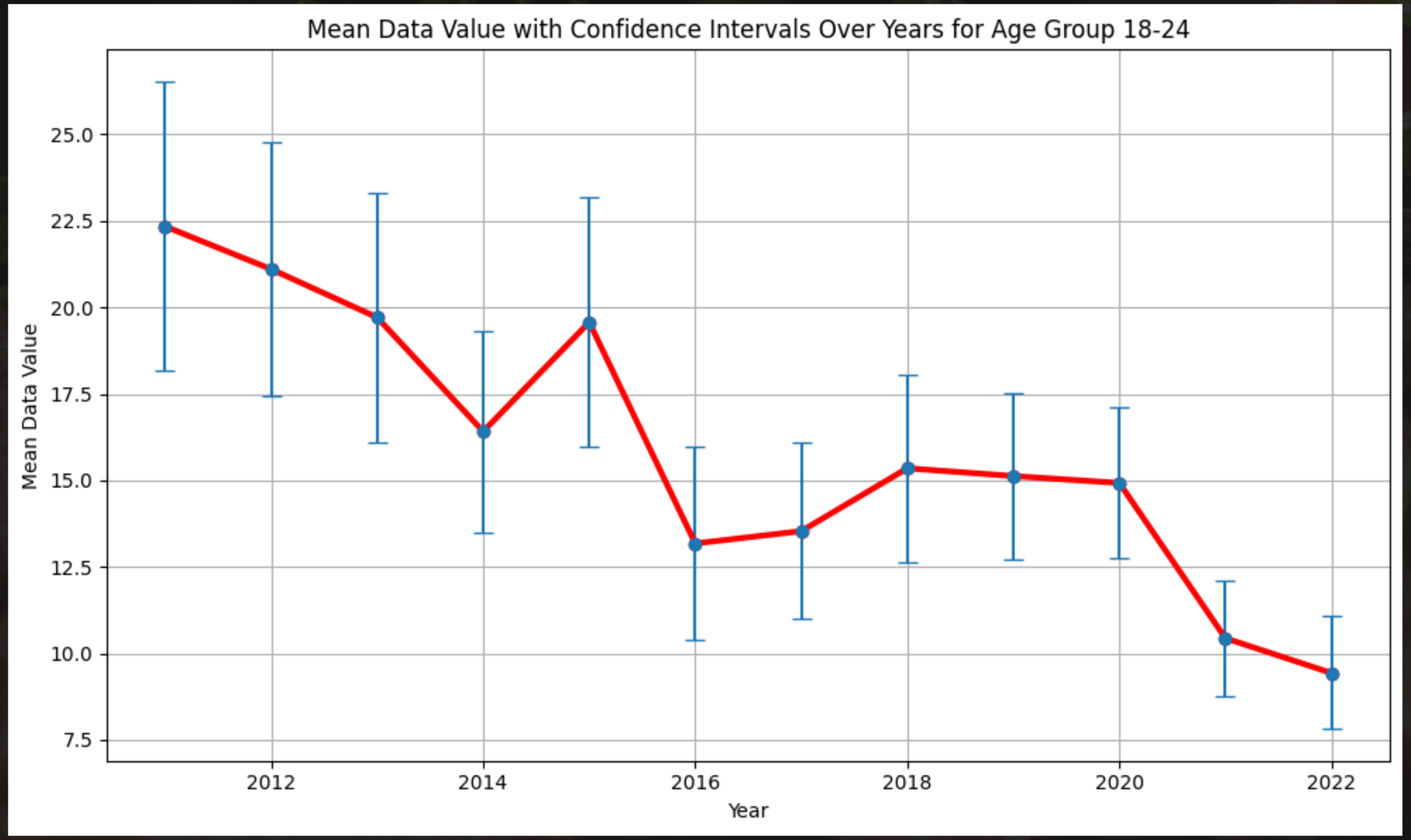
```
df[(df['Response']=='Yes')|(df['Response']=='Every day')|(df['Response']=='Smoke everyday')].groupby('Age Group').agg('mean')
```

Distribution of percentage of smokers in each Age Group over the years

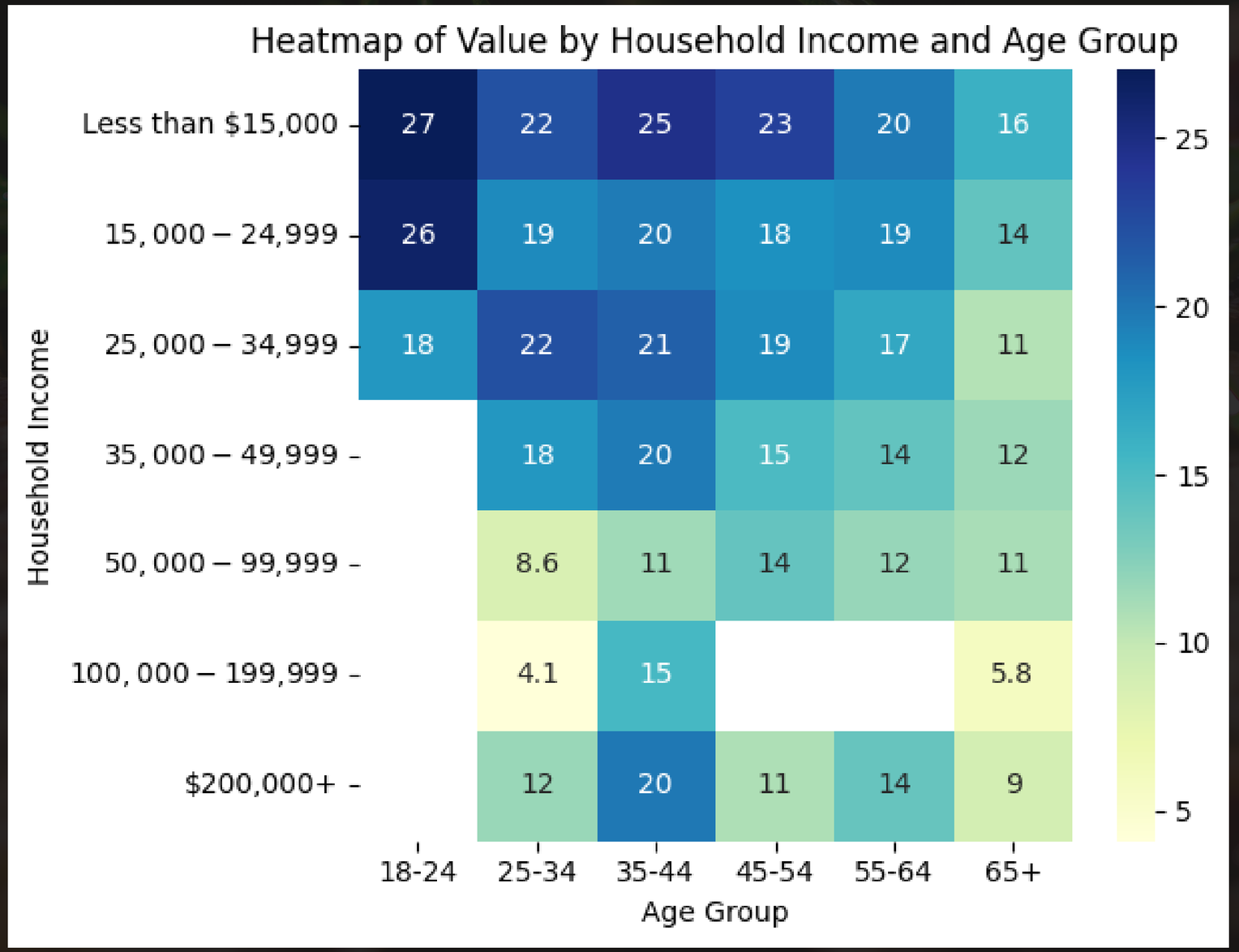




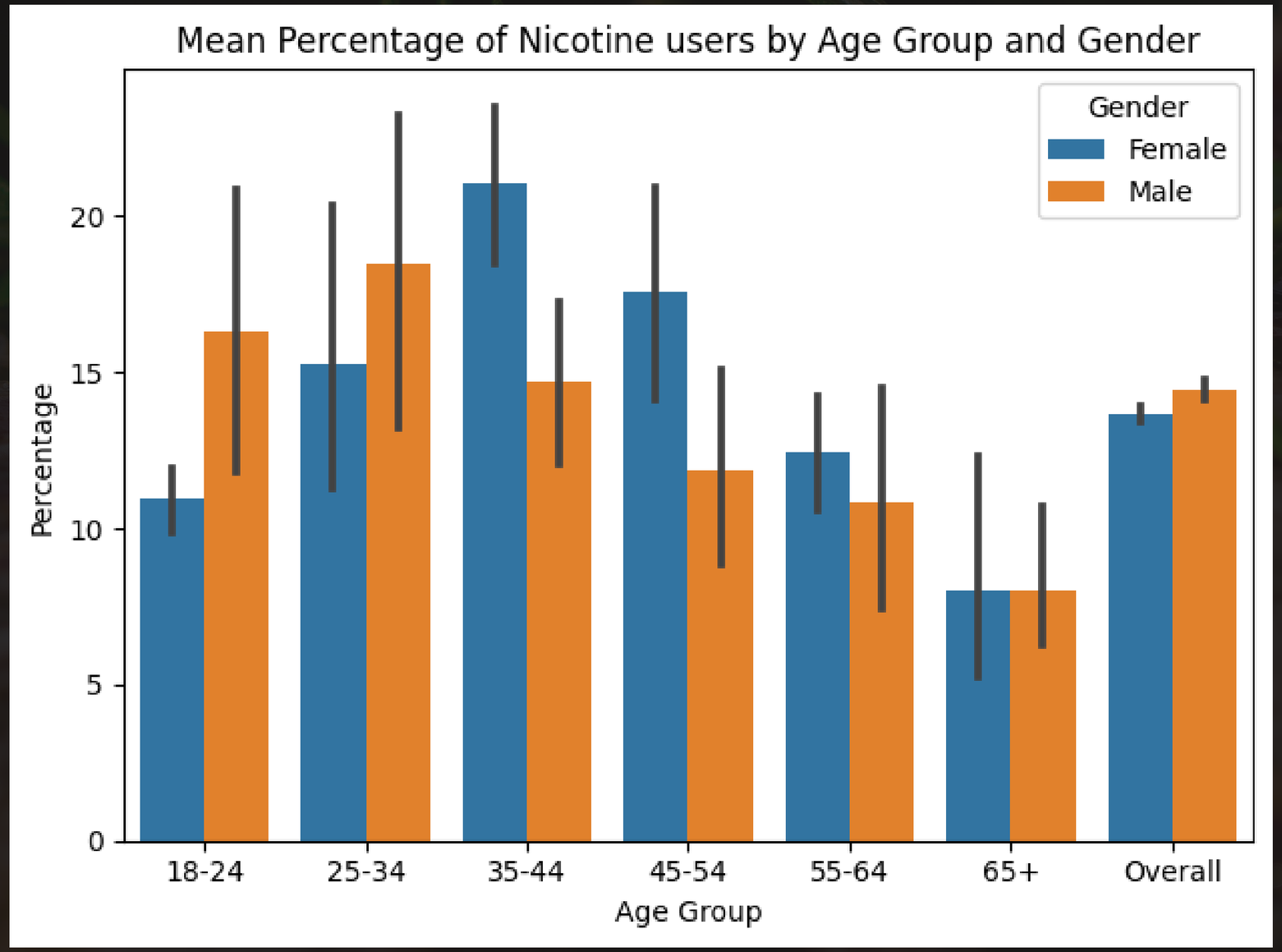
Verifying 18-24 Age Group trends



Mapping relation between Income and Age Group



Gender proportions within Age Groups



2

Trends based on Education level

- Does higher education translate to awareness about smoking?
- Analysing nicotine consumption based on education levels and household income

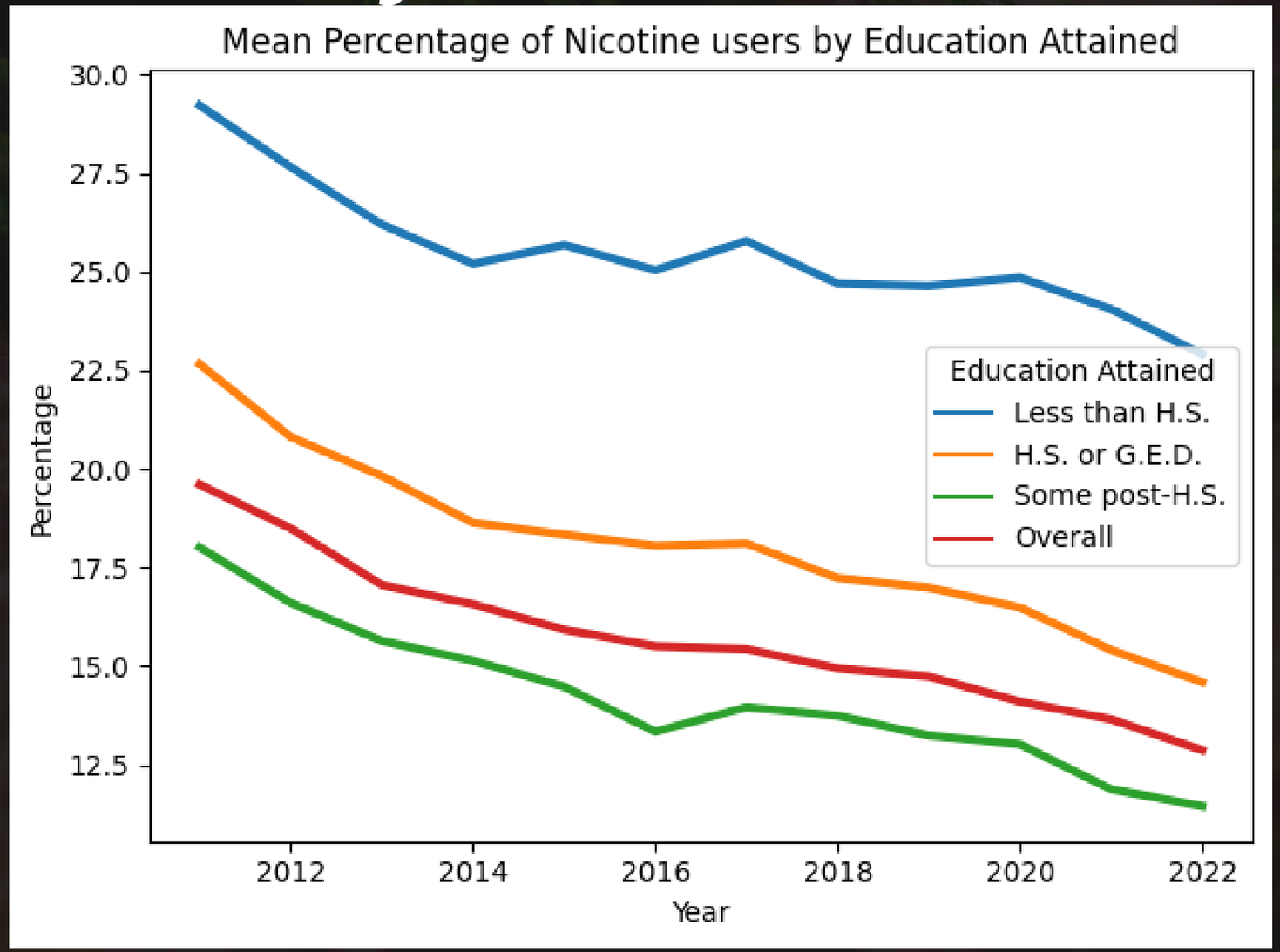
Trends Based on Education Attained

Mean percentage of smokers catagorized by Education

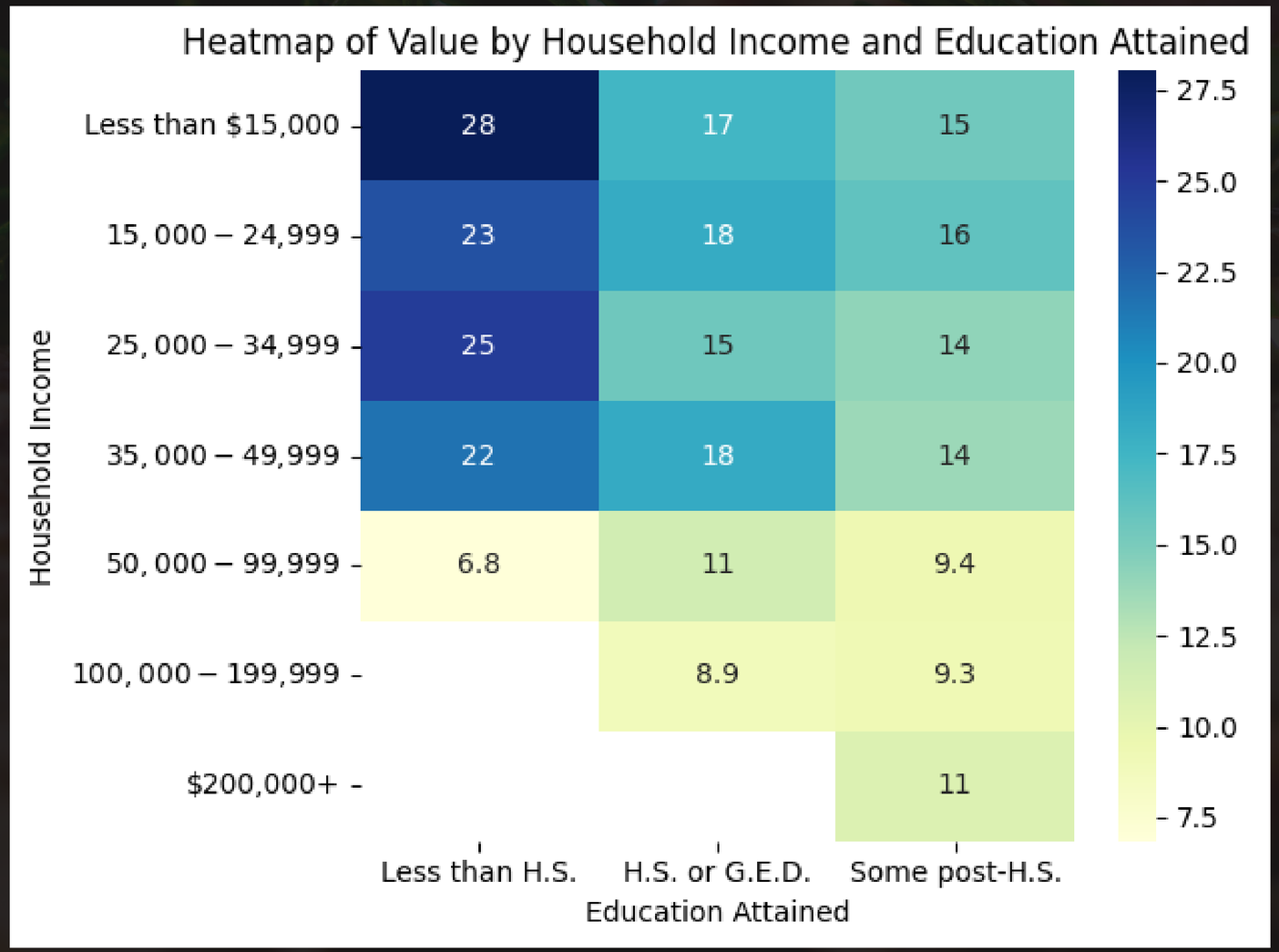
Education Attained	Confidence_limit_Low	Data_value	Confidence_limit_High
Less than H.S.	21.310388	25.707030	30.105352
H.S. or G.E.D.	15.729365	18.102886	20.479437
Some post-H.S.	12.135955	14.230562	16.323670
Overall	13.425048	15.858045	18.290316

```
ndf[(ndf['Response']=='Yes')|(ndf['Response']=='Every day')|(ndf['Response']=='Smoke everyday')].groupby('Education Attained').agg('mean')
```

Mean Percentage of Nicotine users sorted by Education attained



Percentage of Nicotine users categorized by Income and Education Attained



3

Trends based on Race

- Does a relation exist between ethnicity and nicotine use?
- How have smoking levels changed over different races over the years?

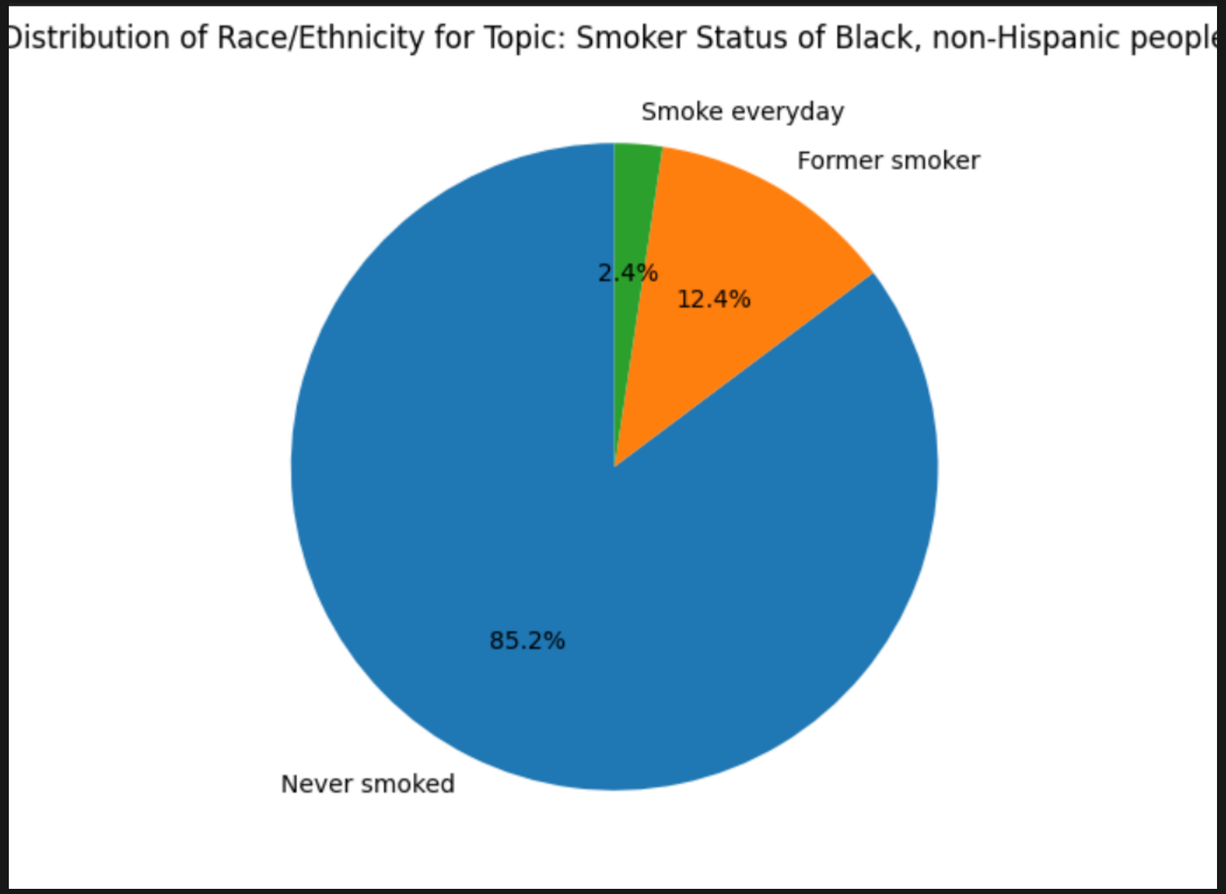
Percentage of Smokers according to Race

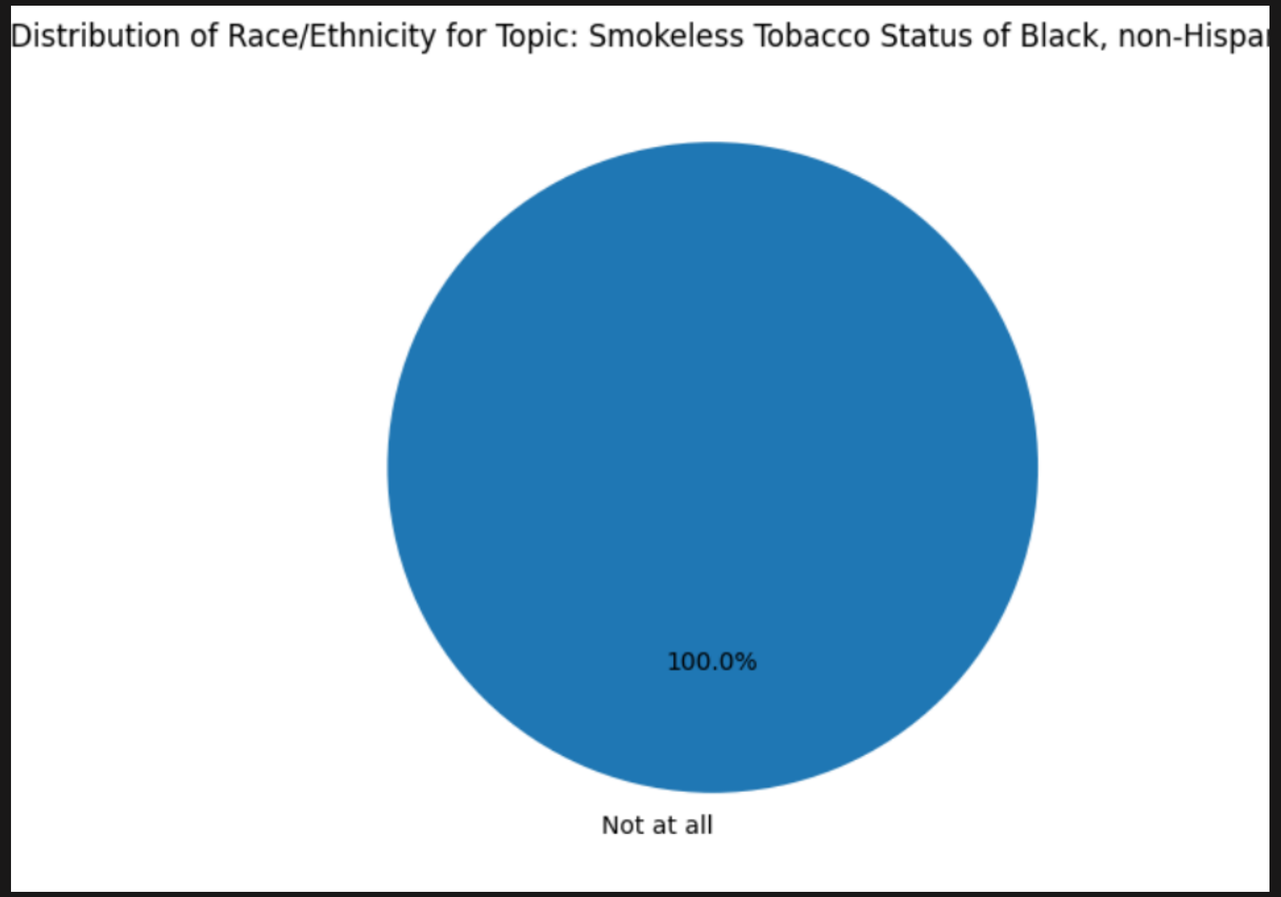
Response		Percentage_Former smoker	Percentage_No
Race/Ethnicity	Topic		
Asian, non-Hispanic	Smokeless Tobacco	0.00	100.00
	Smoker Status	12.73	87.27
Black, non-Hispanic	Smokeless Tobacco	0.00	100.00
	Smoker Status	5.56	91.88
Hispanic	Smokeless Tobacco	0.00	100.00
	Smoker Status	9.47	88.35
White, non-Hispanic	Smokeless Tobacco	0.00	99.05
	Smoker Status	21.91	44.55
White, non-Hispanic, Hispanic	Smokeless Tobacco	0.00	0.00
	Smoker Status	0.00	66.67

Percentage_Yes	Percentage_Some days
0.00	0.00
0.00	0.00
0.00	0.00
2.56	0.00
0.00	0.00
2.18	0.00
0.95	0.90
32.64	0.00
100.00	33.33

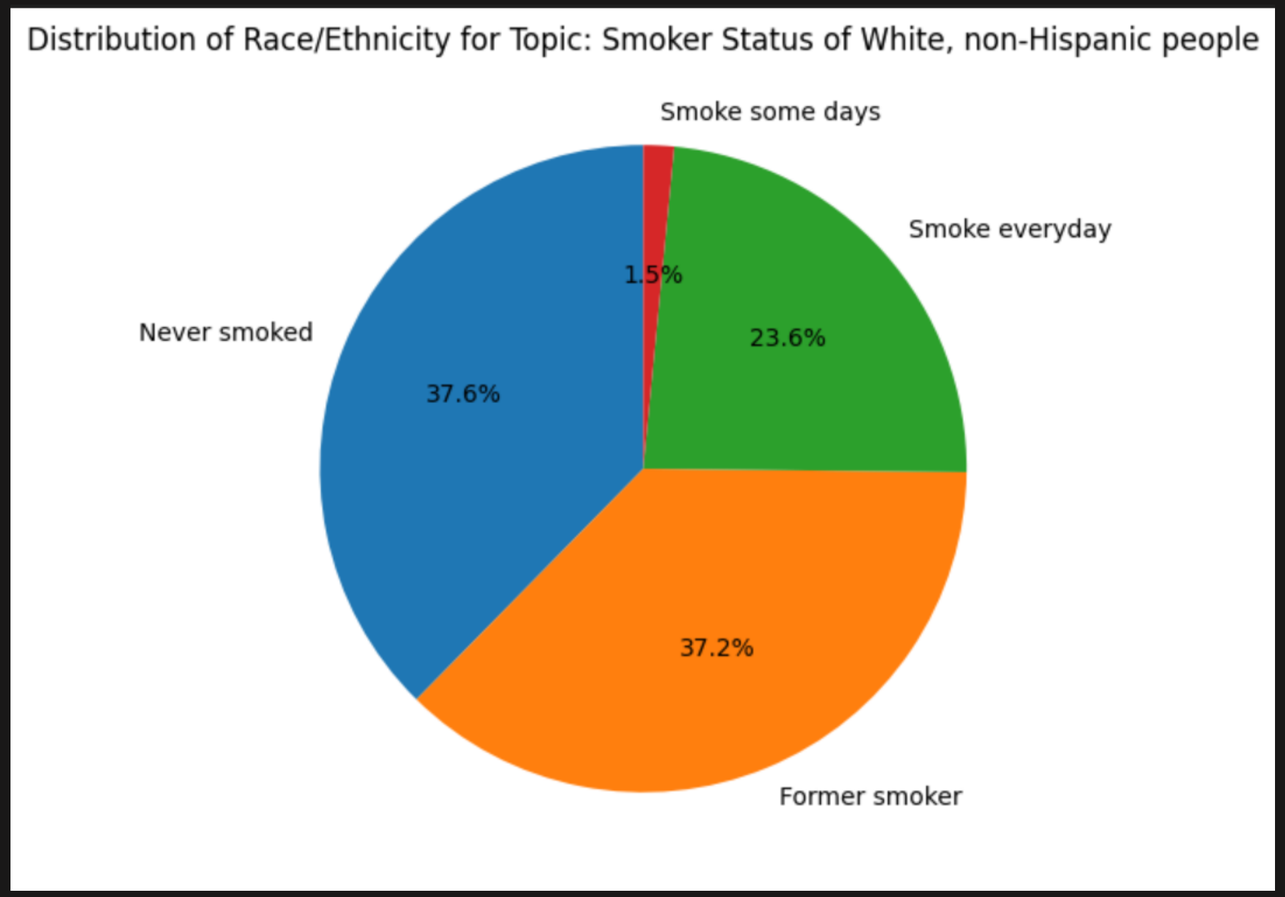
```
percentage_race_topic_df = df_filtered.groupby(['Race/Ethnicity', 'Topic'])['Response'].value_counts().unstack().fillna(0)
total_responses_race_topic = percentage_race_topic_df.sum(axis=1)
for response_column in percentage_race_topic_df.columns:
    percentage_race_topic_df[f'Percentage_{response_column}'] =
    round((percentage_race_topic_df[response_column] / total_responses_race_topic) * 100, 2)
```

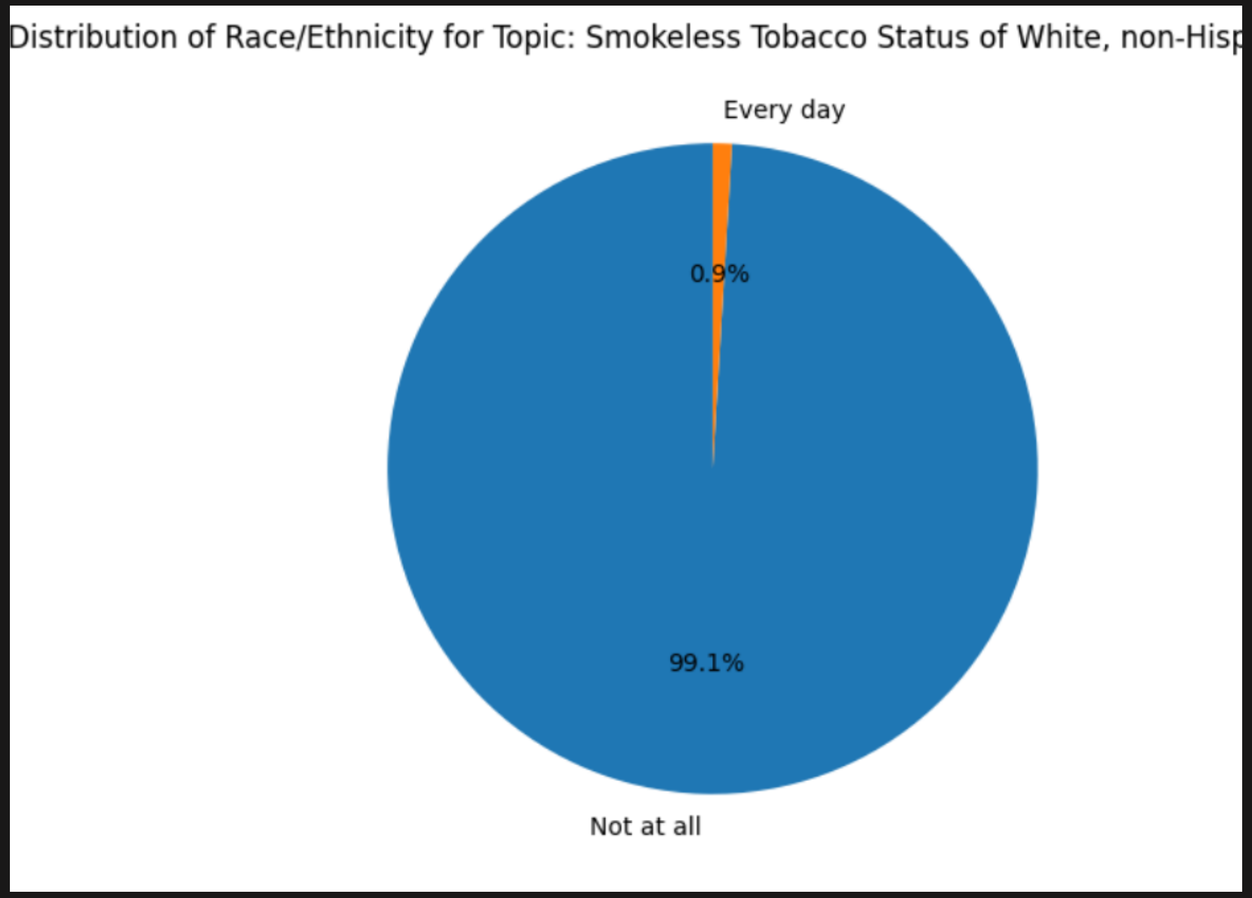
Distribution of Smoking Habits of Black, Non - Hispanic people



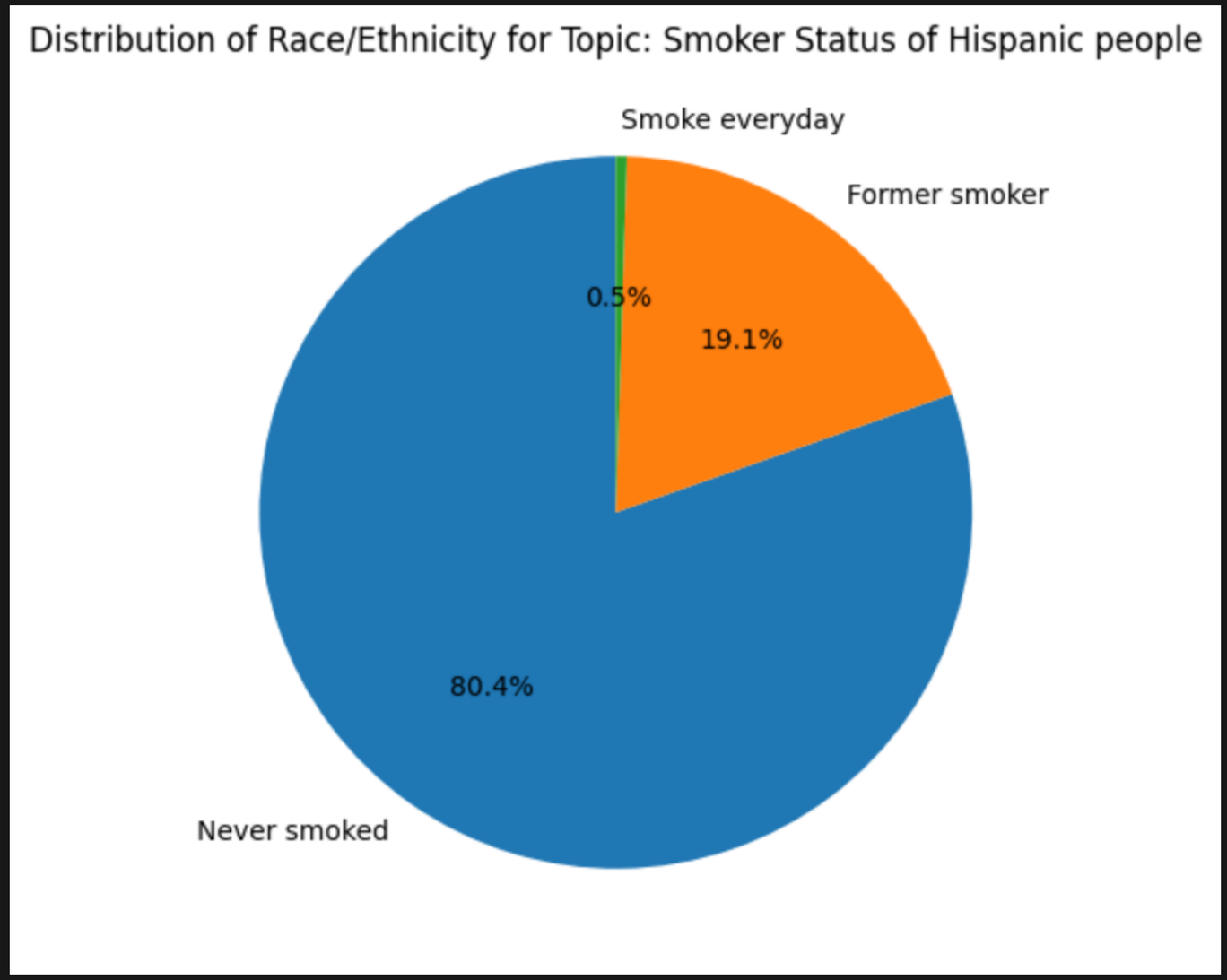


Distribution of Smoking Habits of White, Non - Hispanic people

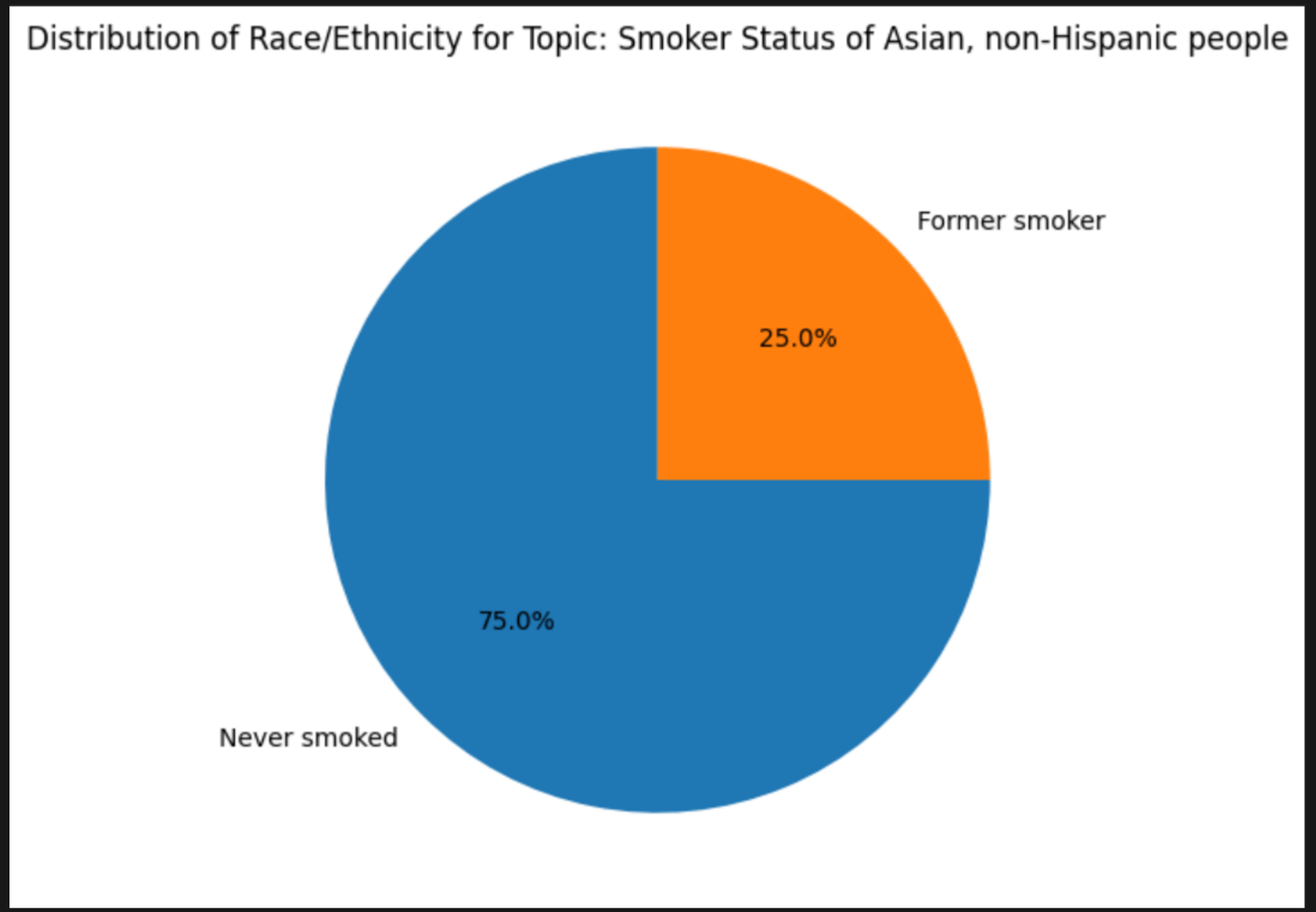




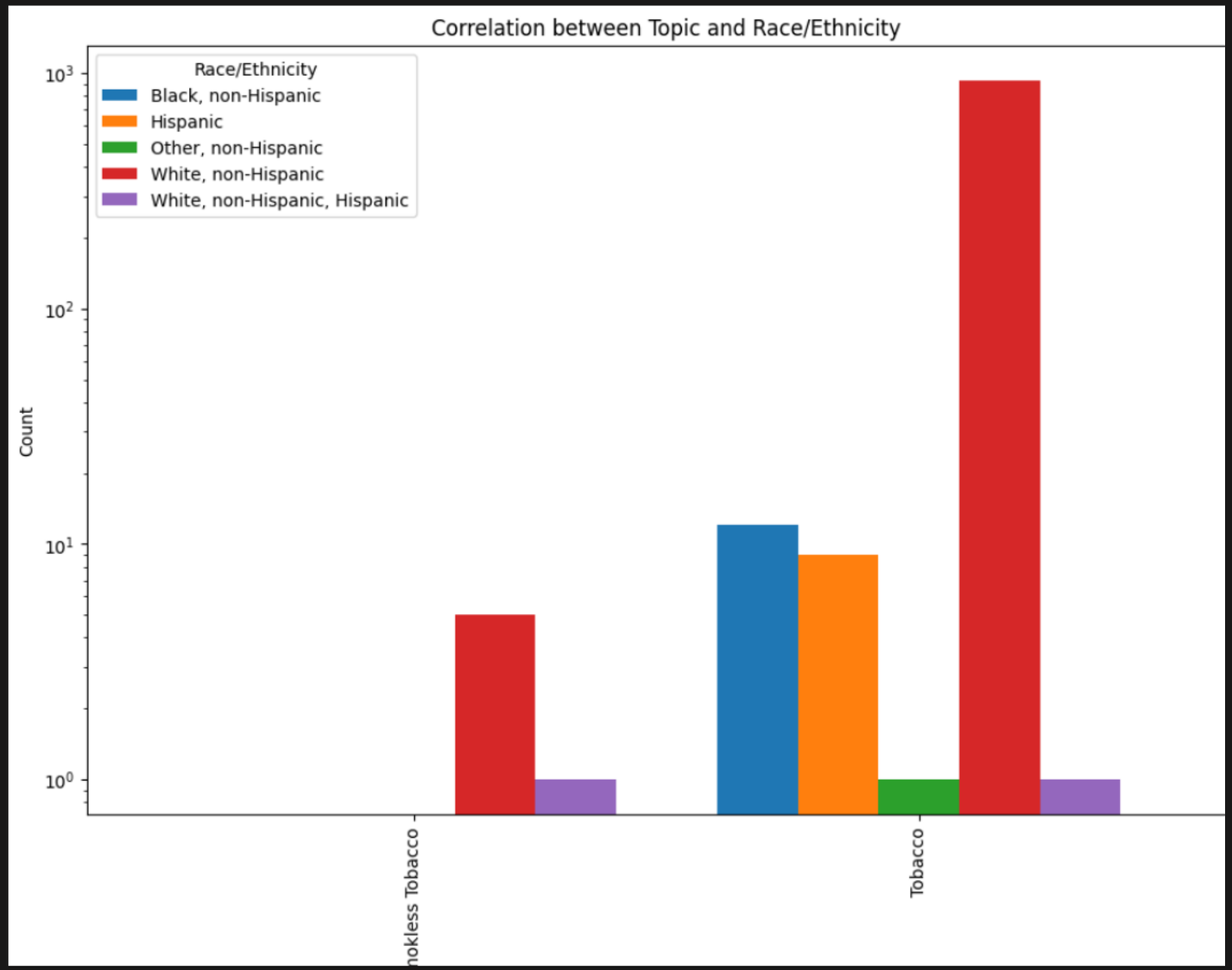
Distribution of Smoking Habits of Hispanic people



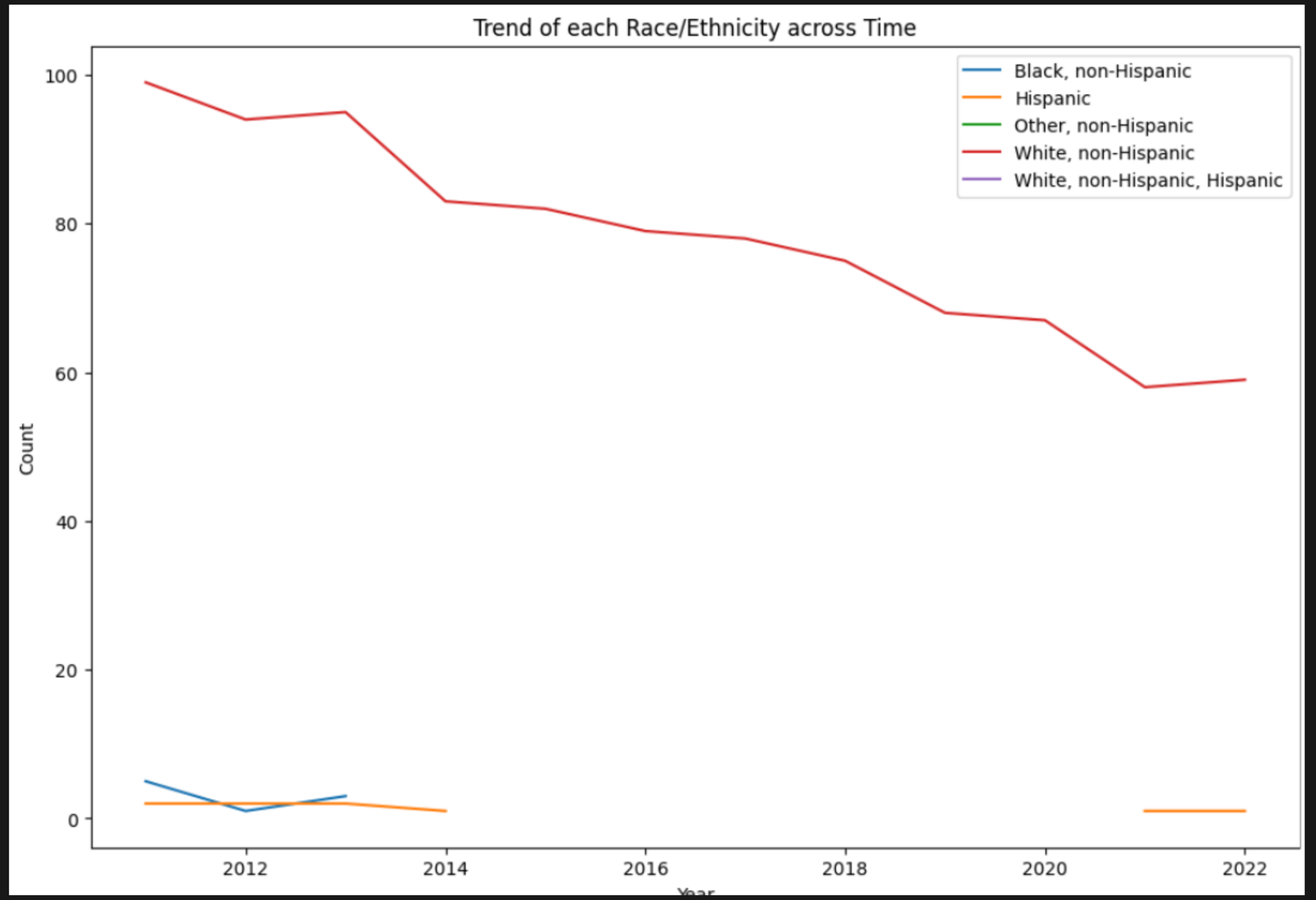
Distribution of Smoking Habits of Asian, Non -Hispanic people



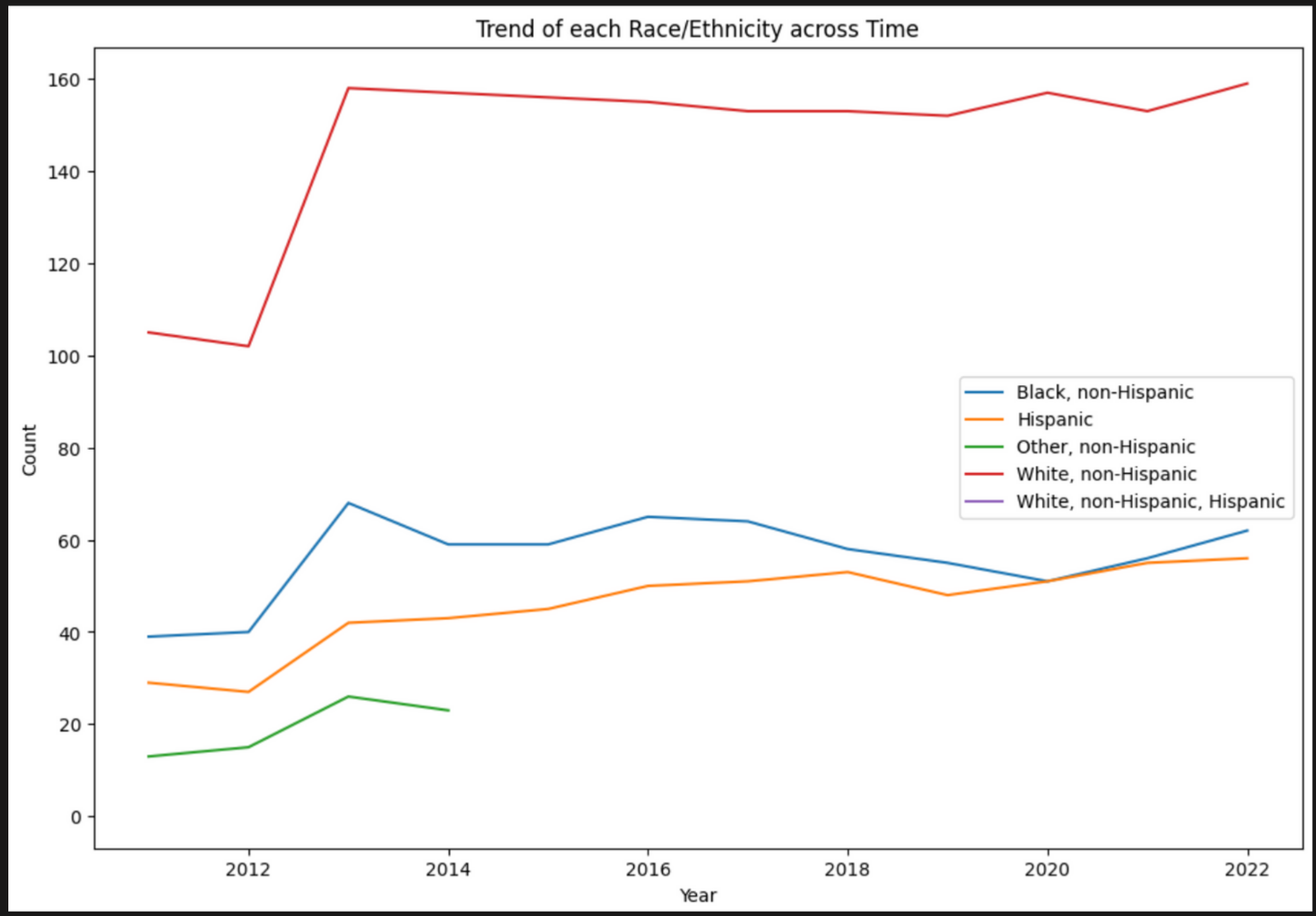
Correlation between Race /Ethnicity and Smoking Habits



Trends of smoking habits each Race/Ethnicity across Time



Trends in non smokers for each Race/Ethnicity

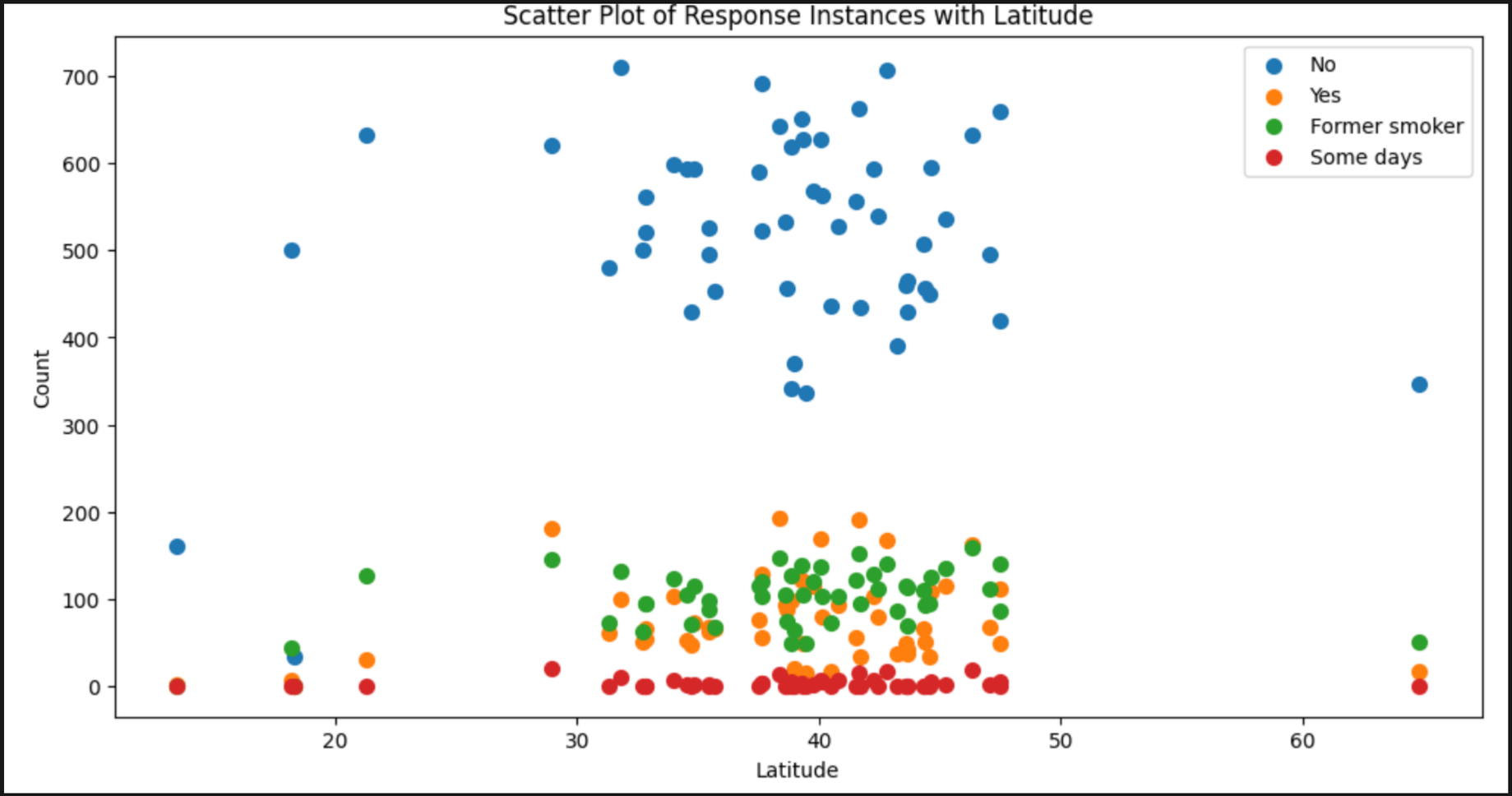


4

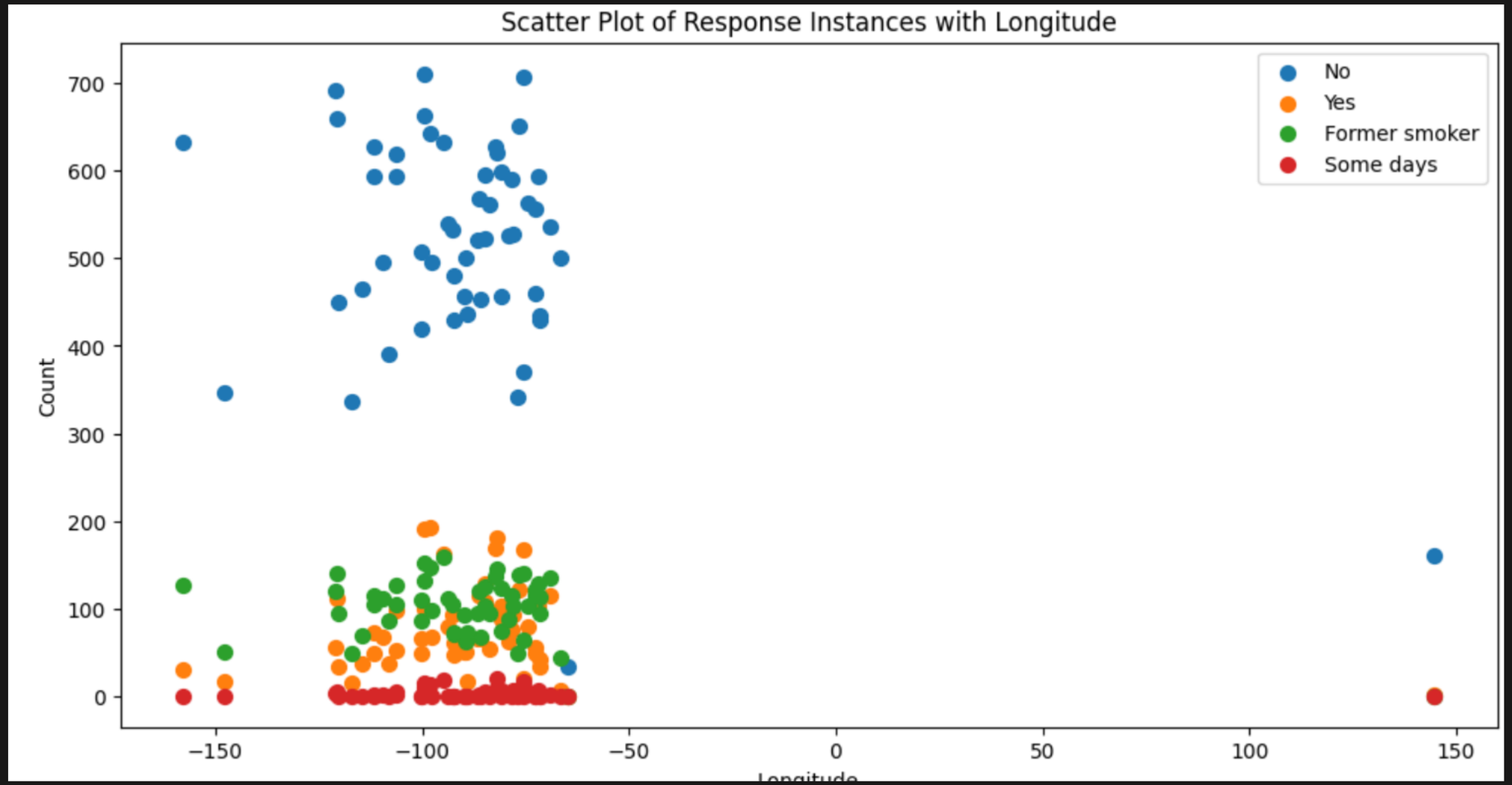
Trends based on GeoLocation

- Is there any correlation between Geolocation and smoking habits?

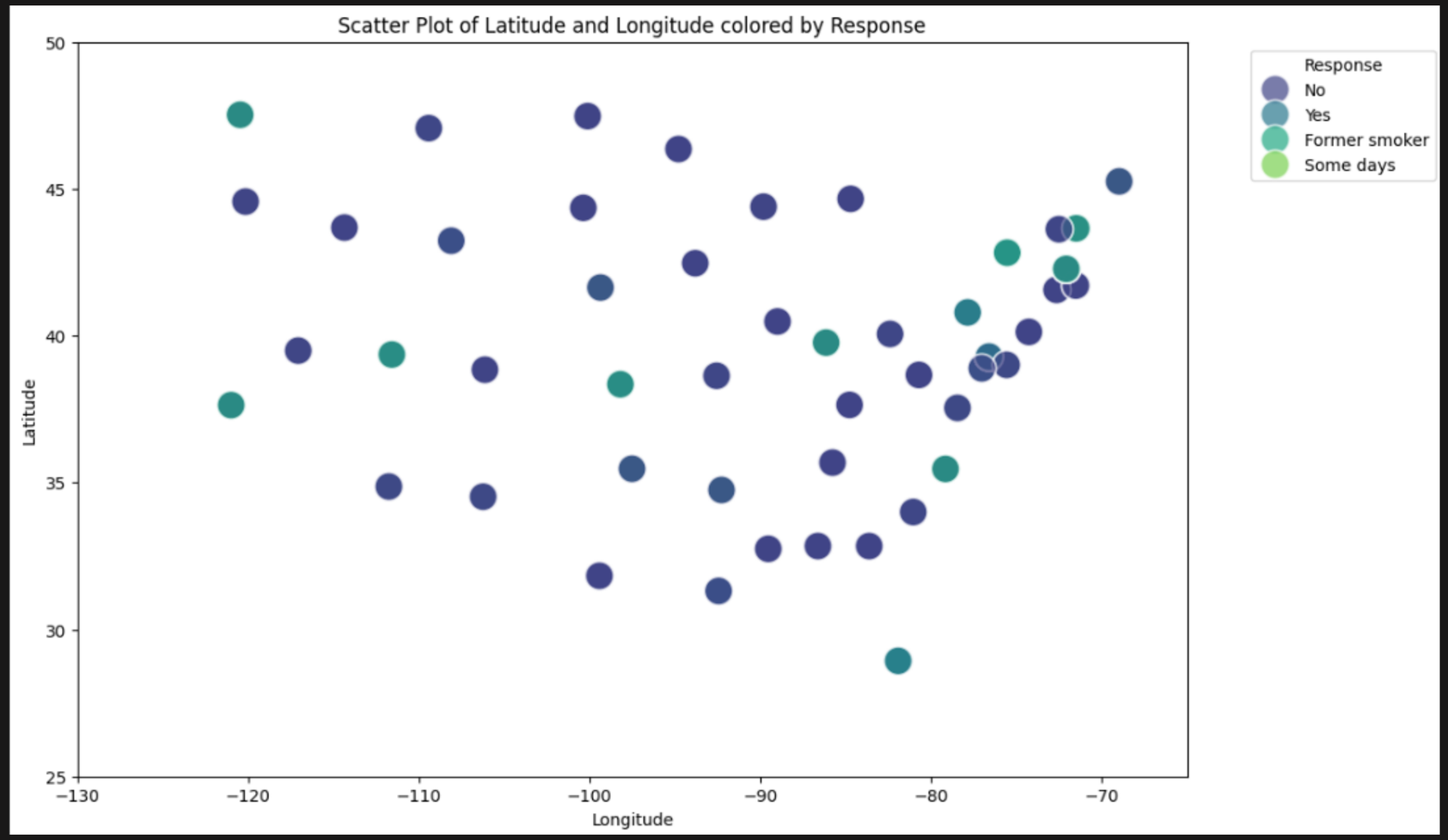
Trends in Smoking Habits according to Latitude



Trends in Smoking Habits according to Longitude



Scatter Plot of Geolocation and Smoking Habits



Initially :

```
['Alabama' 'Alaska' 'Maryland' 'Arizona' 'Arkansas' 'California'
 'Colorado' 'Connecticut' 'District of Columbia' 'Delaware' 'Idaho'
 'Hawaii' 'Florida' 'Iowa' 'Georgia' 'Indiana' 'Kansas' 'Louisiana'
 'Massachusetts' 'Illinois' 'Minnesota' 'New Hampshire' 'Nebraska'
 'Kentucky' 'New York' 'Maine' 'Oregon' 'Michigan' 'Mississippi'
 'Missouri' 'Montana' 'Texas' 'Nevada' 'New Jersey' 'New Mexico'
 'North Carolina' 'Puerto Rico' 'North Dakota' 'Ohio' 'Oklahoma'
 'Pennsylvania' 'Rhode Island' 'South Carolina' 'South Dakota' 'Tennessee'
 'Utah' 'Vermont' 'Virgin Islands' 'Virginia' 'Washington' 'West Virginia'
 'Wisconsin' 'Wyoming']
Number of States: 53
```

Later :

```
['Alabama' 'Maryland' 'Arizona' 'Arkansas' 'Colorado' 'Connecticut'
 'District of Columbia' 'Delaware' 'Idaho' 'Iowa' 'Georgia' 'Indiana'
 'Kansas' 'Massachusetts' 'Illinois' 'Minnesota' 'New Hampshire'
 'Nebraska' 'Kentucky' 'New York' 'Michigan' 'Mississippi' 'Missouri'
 'Montana' 'Nevada' 'New Jersey' 'New Mexico' 'North Carolina'
 'North Dakota' 'Ohio' 'Oklahoma' 'Pennsylvania' 'Rhode Island'
 'South Carolina' 'South Dakota' 'Tennessee' 'Utah' 'Vermont' 'Virginia'
 'West Virginia' 'Wisconsin' 'Wyoming']
Number of States: 42
```

States showing higher degree of smoking preference

Locationdesc	Percentage_Yes
Florida	18.672199
Kansas	19.335347
Nebraska	18.627451

```
filtered_df = merged_df[merged_df['Percentage_Yes'] > 18]
print(filtered_df)
```

Locationdesc	Percentage_Former_Smoker
New Hampshire	19.349315
Vermont	18.298555

```
filtered_df_Former = merged_df[merged_df['Percentage_Former_Smoker'] > 18]
print(filtered_df_Former)
```

Locationdesc	Percentage_No
Guam	98.765432
Puerto Rico	90.744102
Virgin Islands	100.000000

```
filtered_df_No = merged_df[merged_df['Percentage_No'] > 90]
print(filtered_df_No)
```

Locationdesc	Percentage_Some_Days
Florida	1.970954
Minnesota	1.853759
New York	1.648885

```
filtered_df_Some = merged_df[merged_df['Percentage_Some_Days'] > 1.5]
print(filtered_df_Some)
```

Nicotine use mapped to Geolocation