Appendix

October 11, 2024

DSA312 Data Science with Python

Group 3: Melvin Yong, Tan Hao Yang, Teo Jun Hao, Isaac Leong, Caleb Ang

Part 1 of Project

[72]:
```
from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "all"
```

[74]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

1 Understanding dataset

[76]:
```
original_df = pd.read_csv("lung cancer survey.csv")
```

[78]:
```
# ydata profiling on original dataset
from ydata_profiling import ProfileReport
from IPython.display import display, HTML

# Generate the profile report
profile = ProfileReport(original_df, title="Lung Cancer Dataset Profiling␣
  ↪Report", explorative=True)

# Export report as HTML
# profile.to_file("lung_cancer_profile_report.html")

profile.to_notebook_iframe()

display(HTML("<p style='text-align: center; font-weight: bold;'>Figure. 1: Lung␣
  ↪Cancer Dataset Profiling Report</p>"))
```

```
Summarize dataset:   0%|          | 0/5 [00:00<?, ?it/s]
```

```
Generate report structure:   0%|          | 0/1 [00:00<?, ?it/s]
```

```
Render HTML:   0%|          | 0/1 [00:00<?, ?it/s]
```

```
<IPython.core.display.HTML object>

<IPython.core.display.HTML object>
```

```
[80]: # Data Cleaning
      df_No_NA = original_df.dropna()  # Remove empty entries
      final_df = df_No_NA[df_No_NA['AGE'] > 21]   # Remove outlier, 21 y/o

      final_df
      display(HTML("<p style='text-align: center; font-weight: bold;'>Figure. 2:␣
       ↪Dropping NA and Outlier</p>"))
```

```
[80]:       GENDER   AGE  SMOKING  YELLOW_FINGERS  ANXIETY  PEER_PRESSURE  \
      0        0.0  61.0      0.0             0.0      1.0            1.0
      1        1.0  70.0      1.0             1.0      0.0            0.0
      2        1.0  59.0      0.0             0.0      0.0            0.0
      3        1.0  54.0      0.0             0.0      0.0            1.0
      4        0.0  54.0      1.0             0.0      0.0            1.0
      …        …    …        …             …       …             …
      8996     1.0  62.0      0.0             1.0      1.0            1.0
      8997     0.0  71.0      1.0             1.0      1.0            0.0
      8998     1.0  63.0      1.0             0.0      0.0            1.0
      8999     1.0  70.0      1.0             1.0      0.0            0.0
      9099     1.0  62.0      0.0             0.0      0.0            1.0

            CHRONIC DISEASE  FATIGUE   ALLERGY   WHEEZING  ALCOHOL CONSUMING  \
      0                 0.0       1.0       0.0       0.0                0.0
      1                 1.0       0.0       1.0       1.0                0.0
      2                 1.0       1.0       0.0       0.0                0.0
      3                 0.0       1.0       0.0       1.0                1.0
      4                 0.0       1.0       1.0       0.0                1.0
      …                   …       …       …       …                 …
      8996              0.0       1.0       0.0       0.0                0.0
      8997              0.0       0.0       1.0       0.0                1.0
      8998              0.0       0.0       0.0       0.0                0.0
      8999              1.0       1.0       1.0       0.0                1.0
      9099              0.0       1.0       1.0       1.0                1.0

            COUGHING  SHORTNESS OF BREATH  SWALLOWING DIFFICULTY  CHEST PAIN  \
      0          0.0                  1.0                    0.0         0.0
      1          1.0                  1.0                    0.0         0.0
      2          0.0                  0.0                    0.0         1.0
      3          1.0                  1.0                    0.0         1.0
      4          0.0                  0.0                    1.0         0.0
      …           …                   …                     …          …
      8996       0.0                  0.0                    1.0         1.0
      8997       1.0                  1.0                    0.0         1.0
      8998       1.0                  1.0                    0.0         0.0
```

```
8999      0.0                 1.0                   0.0        1.0
9099      0.0                 0.0                   1.0        0.0

      LUNG_CANCER
0             1.0
1             1.0
2             0.0
3             1.0
4             1.0
…            …
8996          1.0
8997          1.0
8998          0.0
8999          1.0
9099          1.0

[9000 rows x 16 columns]
```

```
<IPython.core.display.HTML object>
```

[82]:

```python
# Average age in cleaned dataset
avg_age = final_df["AGE"].mean()
print('The participants have an average age of', avg_age)
```

```
The participants have an average age of 60.711
```

[84]:

```python
# Numerical view of mean values of cleaned dataset
final_df.mean()
```

[84]:

```
GENDER                  0.536333
AGE                    60.711000
SMOKING                 0.521000
YELLOW_FINGERS          0.528000
ANXIETY                 0.456333
PEER_PRESSURE           0.510667
CHRONIC DISEASE         0.465667
FATIGUE                 0.687667
ALLERGY                 0.548222
WHEEZING                0.490444
ALCOHOL CONSUMING       0.525222
COUGHING                0.593889
SHORTNESS OF BREATH     0.698000
SWALLOWING DIFFICULTY   0.348000
CHEST PAIN              0.627889
LUNG_CANCER             0.805000
dtype: float64
```

[86]:

```python
# Visualised mean values of cleaned dataset

average_of_final_df = final_df.drop(columns = ['AGE'])   # Drop age from␣
 ↪visualisation
mean_of_variables = average_of_final_df.mean()

# Color code bars for easy visualisation
color_assignment = ['lightgreen' if value > 0.5
                    else 'pink' for value in mean_of_variables]

plt.figure(figsize=(8,6))
plt.bar(mean_of_variables.index,mean_of_variables, color = color_assignment)
# Legend for the aforementioned
plt.legend(['More than half of participants = 1'], title = 'Colors',
           loc = 'upper left', fontsize = 7)
plt.xticks(rotation = 90, fontsize = 10)
plt.title('Average values of variables', fontsize = 16)

plt.show()
```

[86]:

```
<Figure size 800x600 with 0 Axes>
```

[86]:

```
<BarContainer object of 15 artists>
```

[86]:

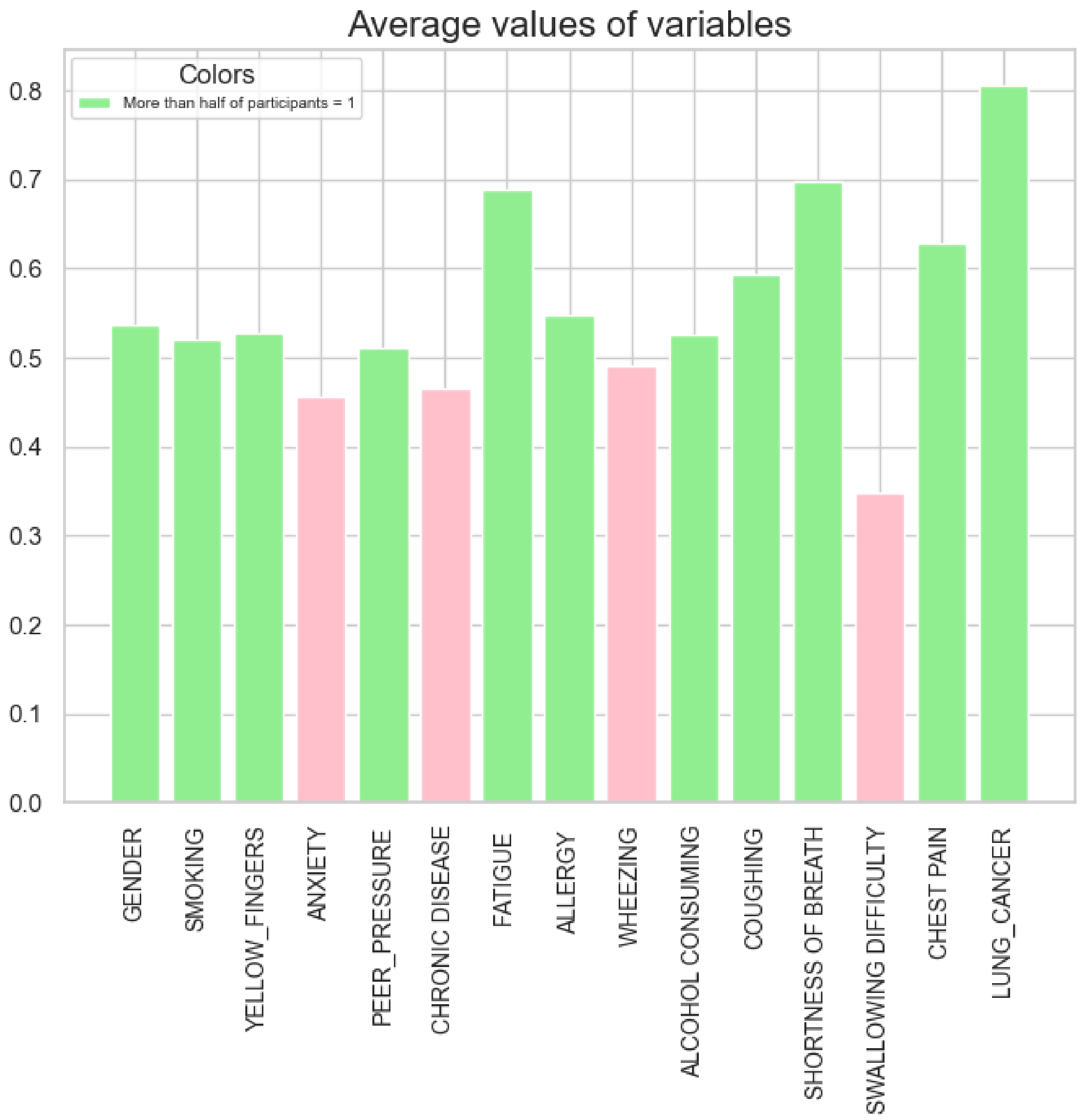
```
<matplotlib.legend.Legend at 0x16aa1ddd0>
```

[86]:

```
([0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14],
 [Text(0, 0, 'GENDER'),
  Text(1, 0, 'SMOKING'),
  Text(2, 0, 'YELLOW_FINGERS'),
  Text(3, 0, 'ANXIETY'),
  Text(4, 0, 'PEER_PRESSURE'),
  Text(5, 0, 'CHRONIC DISEASE'),
  Text(6, 0, 'FATIGUE '),
  Text(7, 0, 'ALLERGY '),
  Text(8, 0, 'WHEEZING'),
  Text(9, 0, 'ALCOHOL CONSUMING'),
  Text(10, 0, 'COUGHING'),
  Text(11, 0, 'SHORTNESS OF BREATH'),
  Text(12, 0, 'SWALLOWING DIFFICULTY'),
  Text(13, 0, 'CHEST PAIN'),
  Text(14, 0, 'LUNG_CANCER')])
```

[86]:

```
Text(0.5, 1.0, 'Average values of variables')
```

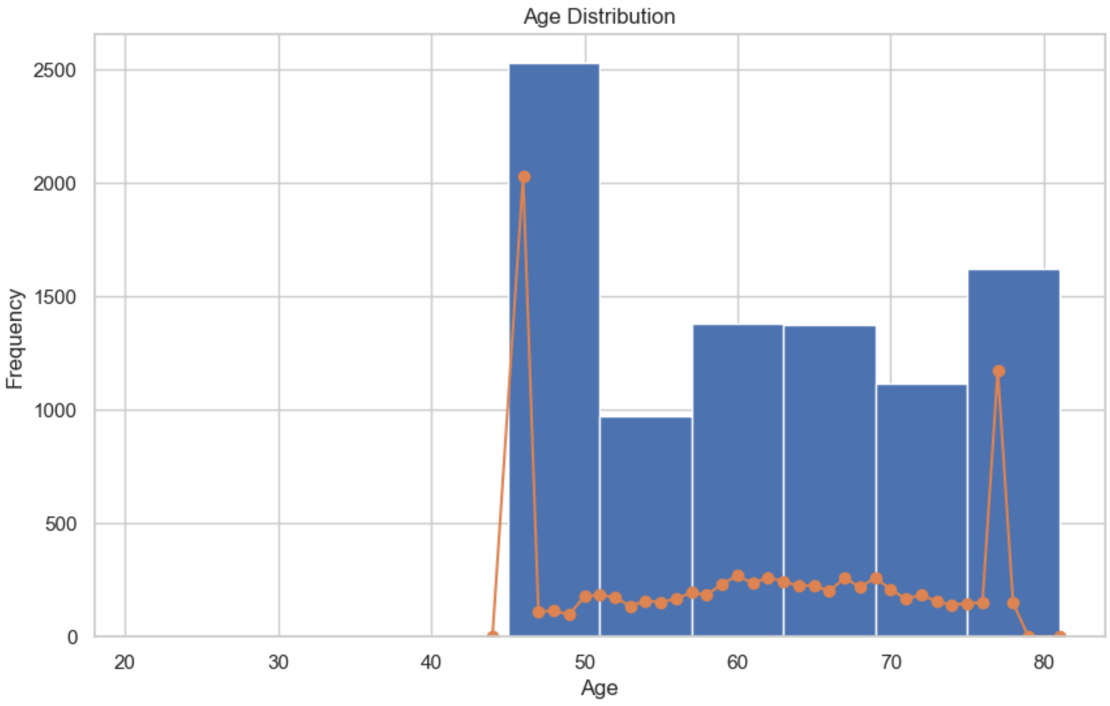


[88]:

```python
# Data visualisation
# Plot Histogram
plt.figure(figsize=(10, 6))
plt.hist(df_No_NA['AGE'])
plt.title('Histogram of Patient Ages')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.grid(True)

# Count the frequency of each age
age_counts = final_df['AGE'].value_counts().sort_index()
```

```
# Plot a line graph of the age counts
plt.plot(age_counts.index, age_counts.values, marker='o')
plt.title('Age Distribution')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.show()
```

```
[88]: <Figure size 1000x600 with 0 Axes>

[88]: (array([1.000e+00, 0.000e+00, 0.000e+00, 2.000e+00, 2.530e+03, 9.740e+02,
        1.380e+03, 1.373e+03, 1.116e+03, 1.625e+03]),
 array([21., 27., 33., 39., 45., 51., 57., 63., 69., 75., 81.]),
 <BarContainer object of 10 artists>)

[88]: Text(0.5, 1.0, 'Histogram of Patient Ages')

[88]: Text(0.5, 0, 'Age')

[88]: Text(0, 0.5, 'Frequency')

[88]: [<matplotlib.lines.Line2D at 0x177cf0350>]

[88]: Text(0.5, 1.0, 'Age Distribution')

[88]: Text(0.5, 0, 'Age')

[88]: Text(0, 0.5, 'Frequency')
```



[90]:

```
# Correlation Matrix
# High level overview to understand which variables should be looked into
final_df_corr = final_df.corr()
final_df_corr.style.background_gradient(cmap='coolwarm')
```

[90]:

```
<pandas.io.formats.style.Styler at 0x177c2fa50>
```

[92]:

```
# For presentation's sake
final_df_corr_plot = final_df.corr()

# Draw heatmap
plt.figure(figsize=(11, 8))
sns.heatmap(final_df_corr_plot, annot = True, cmap='coolwarm', linewidths=1.5)

plt.figtext(0.45, -0.15, "Figure. 3: Correlation Heatmap", ha="center",␣
 ↪fontsize=12)
plt.show()
```

[92]:

```
<Figure size 1100x800 with 0 Axes>
```

[92]:

```
<Axes: >
```

[92]:

```
Text(0.45, -0.15, 'Figure. 3: Correlation Heatmap')
```

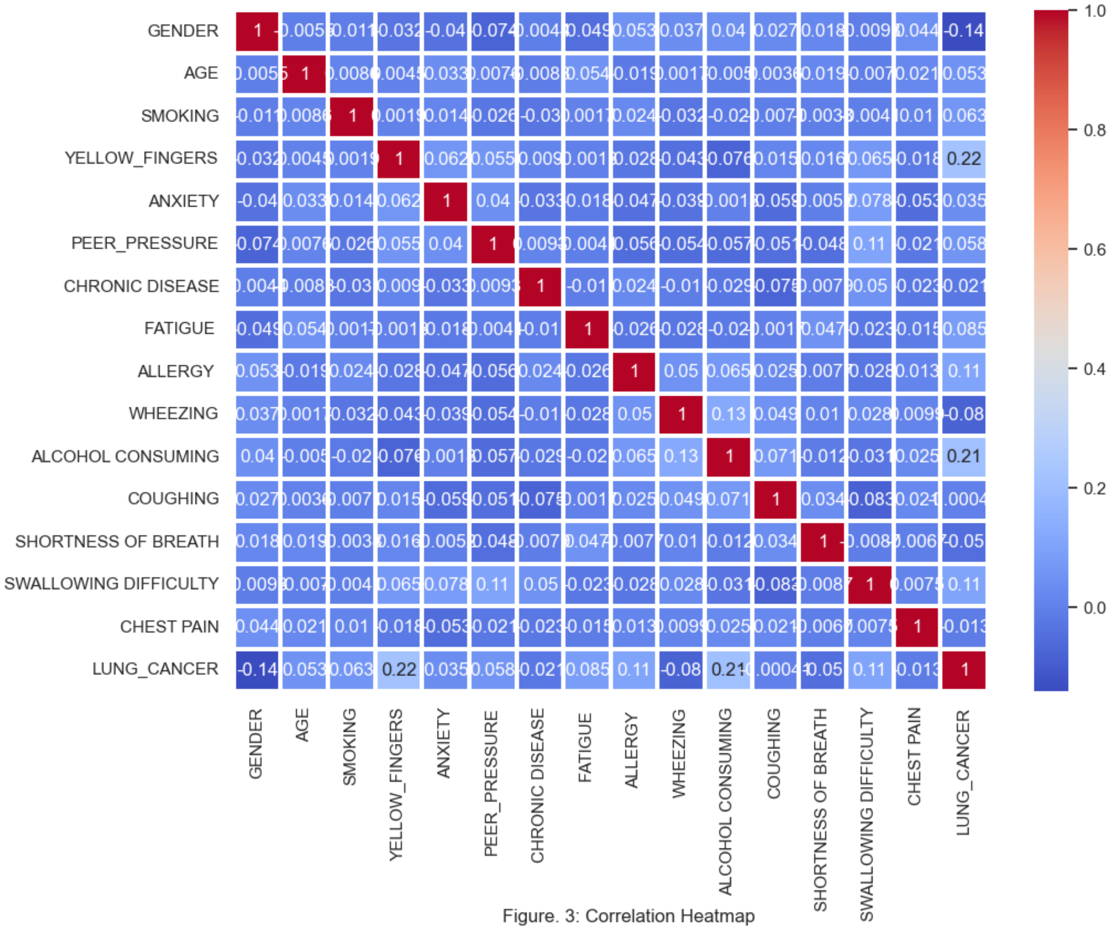


Figure. 3: Correlation Heatmap

2 Smoking

[61]:
```
contingency_smoking = pd.crosstab(final_df['SMOKING'], final_df['LUNG_CANCER'])
contingency_smoking['Proportion_Cancer'] = contingency_smoking[1] /␣
  ↪(contingency_smoking[0] + contingency_smoking[1])


proportions_smoking = contingency_smoking['Proportion_Cancer']
non_cancer_smoking = 1 - proportions_smoking


plt.figure(figsize=(8, 4))
sns.set_theme(style="whitegrid")
y_labels_smoking = ['Non-Smoker', 'Smoker']
bar_width = 0.3 #change the width for a better fit figure
#create horizontal bar chart for cancer patients who smoke/do not smoke using␣
  ↪labels above and sns pastel color palette
```

```
bars_lung_cancer_smoking = plt.barh(y_labels_smoking, proportions_smoking,␣
 ↪color=sns.color_palette("pastel", 2), height=bar_width)
#create horizontal bar chart for cancer patients who smoke/ do not smoke using␣
 ↪labels above and sns pastel color palette
bars_no_cancer_smoking = plt.barh(y_labels_smoking, non_cancer_smoking,␣
 ↪color='lightgrey', left=proportions_smoking, height=bar_width)

plt.title('Proportion of Lung Cancer Cases by Smoking Status', fontsize=16)
plt.xlabel('Proportion', fontsize=14)
plt.xlim(0, 1)
plt.ylim(-0.5, 1.5)

#to annotate the % of cancer patients for smoker and non-smoker group
for index, value in enumerate(proportions_smoking):
    plt.text(value - 0.4, index, f"{value*100:.2f}%", va='center', fontsize=12,␣
 ↪color='black')

plt.figtext(0.47, -0.1, "Figure. 4", fontsize=15)
plt.show()
```

```
[61]: <Figure size 800x400 with 0 Axes>

[61]: Text(0.5, 1.0, 'Proportion of Lung Cancer Cases by Smoking Status')

[61]: Text(0.5, 0, 'Proportion')

[61]: (0.0, 1.0)

[61]: (-0.5, 1.5)

[61]: Text(0.37893760148457434, 0, '77.89%')

[61]: Text(0.42896139901898056, 1, '82.90%')

[61]: Text(0.47, -0.1, 'Figure. 4')
```

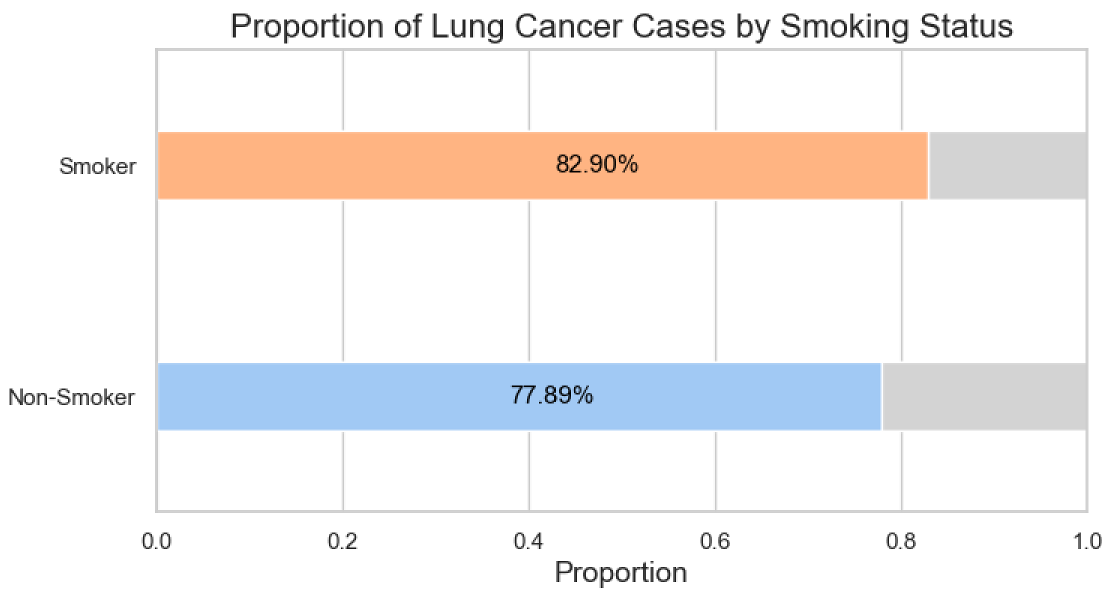


Figure. 4

3 Yellow Fingers

[65]:
```
contingency_yellowfingers = pd.crosstab(final_df['YELLOW_FINGERS'],␣
 ↪final_df['LUNG_CANCER'])
contingency_yellowfingers['Proportion_Cancer'] = contingency_yellowfingers[1] /␣
 ↪(contingency_yellowfingers[0] + contingency_yellowfingers[1])

proportions_yellowfingers = contingency_yellowfingers['Proportion_Cancer']
non_cancer_yellowfingers = 1 - proportions_yellowfingers

plt.figure(figsize=(8, 4))
sns.set_theme(style="whitegrid")
y_labels_yellowfingers = ['No Yellow Fingers', 'Yellow Fingers']
bar_width = 0.3
#create horizontal bar chart for cancer patients with or without yellow fingers␣
 ↪using labels above and sns pastel color palette
bars_lung_cancer_yellowfingers = plt.barh(y_labels_yellowfingers,␣
 ↪proportions_yellowfingers, color=['#FFCC99','#FF9999'], height=bar_width,␣
 ↪label='Lung Cancer')
#create horizontal bar chart for non-cancer patients with or without yellow␣
 ↪fingers using labels above and sns pastel color palette
bars_no_cancer_yellowfingers = plt.barh(y_labels_yellowfingers,␣
 ↪non_cancer_yellowfingers, color='lightgrey', left=proportions_yellowfingers,␣
 ↪height=bar_width, label='No Lung Cancer')
```

```
plt.title('Proportion of Lung Cancer Cases by Presence Of Yellow Fingers',␣
 ↪fontsize=16)
plt.xlabel('Proportion', fontsize=14)
plt.xlim(0, 1)
plt.ylim(-0.5, 1.5)

#annotate % of cancer patients for yellow finger and non-yellow finger group
for index, value in enumerate(proportions_yellowfingers):
    plt.text(value - 0.4, index, f"{value*100:.2f}%", va='center', fontsize=12,␣
 ↪color='black')
plt.figtext(0.47, -0.1, "Figure. 5", fontsize=15)

plt.show()
```

```
[65]: <Figure size 800x400 with 0 Axes>

[65]: Text(0.5, 1.0, 'Proportion of Lung Cancer Cases by Presence Of Yellow Fingers')

[65]: Text(0.5, 0, 'Proportion')

[65]: (0.0, 1.0)

[65]: (-0.5, 1.5)

[65]: Text(0.31374764595103577, 0, '71.37%')

[65]: Text(0.48657407407407405, 1, '88.66%')

[65]: Text(0.47, -0.1, 'Figure. 5')
```

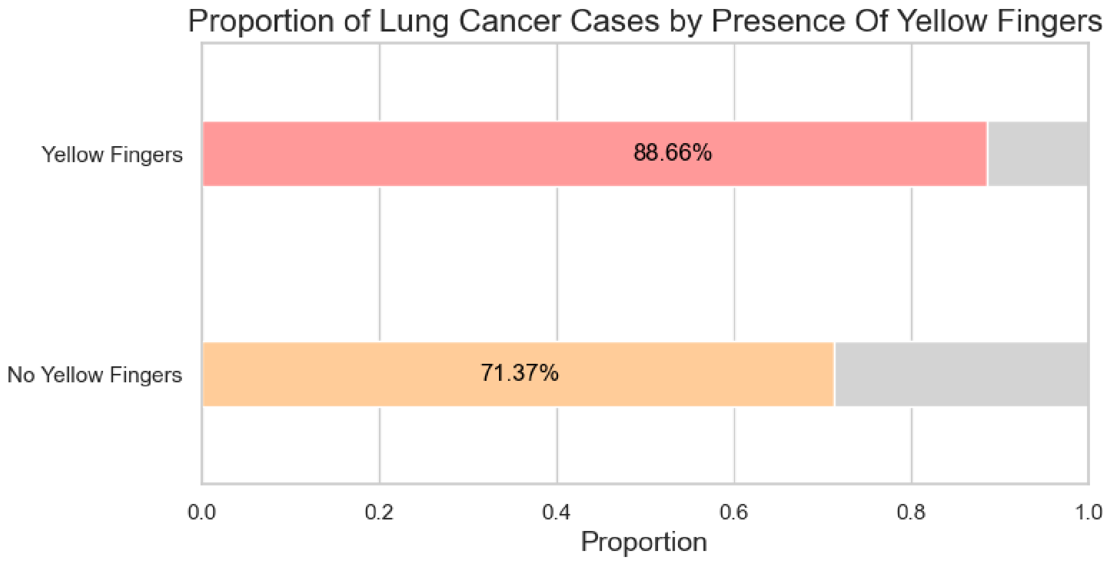


Figure. 5

[83]:

```
labels = ['Has Yellow Fingers', 'No Yellow Fingers']
proportion_nonsmoker_yellow_finger =␣
 ↪contingency_smokingyellowfingers['Proportion_Yellow_Fingers'][0.0]
proportions = [proportion_nonsmoker_yellow_finger,␣
 ↪1-proportion_nonsmoker_yellow_finger]

df_to_work_on = final_df[final_df['LUNG_CANCER']==1]
proportion_smoker_yellow_finger =␣
 ↪contingency_smokingyellowfingers['Proportion_Yellow_Fingers'][1.0]
proportions = [proportion_smoker_yellow_finger,␣
 ↪1-proportion_smoker_yellow_finger]

plt.figure(figsize=(6, 6))
plt.pie(proportions, labels=labels, autopct='%1.2f%%', startangle=90,␣
 ↪colors=['yellow', 'cyan'], wedgeprops={'edgecolor': 'black'})

plt.title('Proportion of Yellow Fingers (Lung Cancer Patients Who Smoke)',␣
 ↪fontsize=14)
plt.figtext(0.47, -0.02, "Figure. 6", fontsize=15)
plt.show()
```

[83]:

```
<Figure size 600x600 with 0 Axes>
```

[83]:

```
([<matplotlib.patches.Wedge at 0x1692d42f0>,
  <matplotlib.patches.Wedge at 0x1692d4bc0>],
 [Text(-1.0736227206676374, -0.23944572175384612, 'Has Yellow Fingers'),
  Text(1.0736226982491062, 0.23944582227365901, 'No Yellow Fingers')],
 [Text(-0.5856123930914385, -0.13060675732027968, '56.98%'),
  Text(0.5856123808631488, 0.13060681214926853, '43.02%')])
```

[83]:

```
Text(0.5, 1.0, 'Proportion of Yellow Fingers (Lung Cancer Patients Who Smoke)')
```

[83]:

```
Text(0.47, -0.02, 'Figure. 6')
```

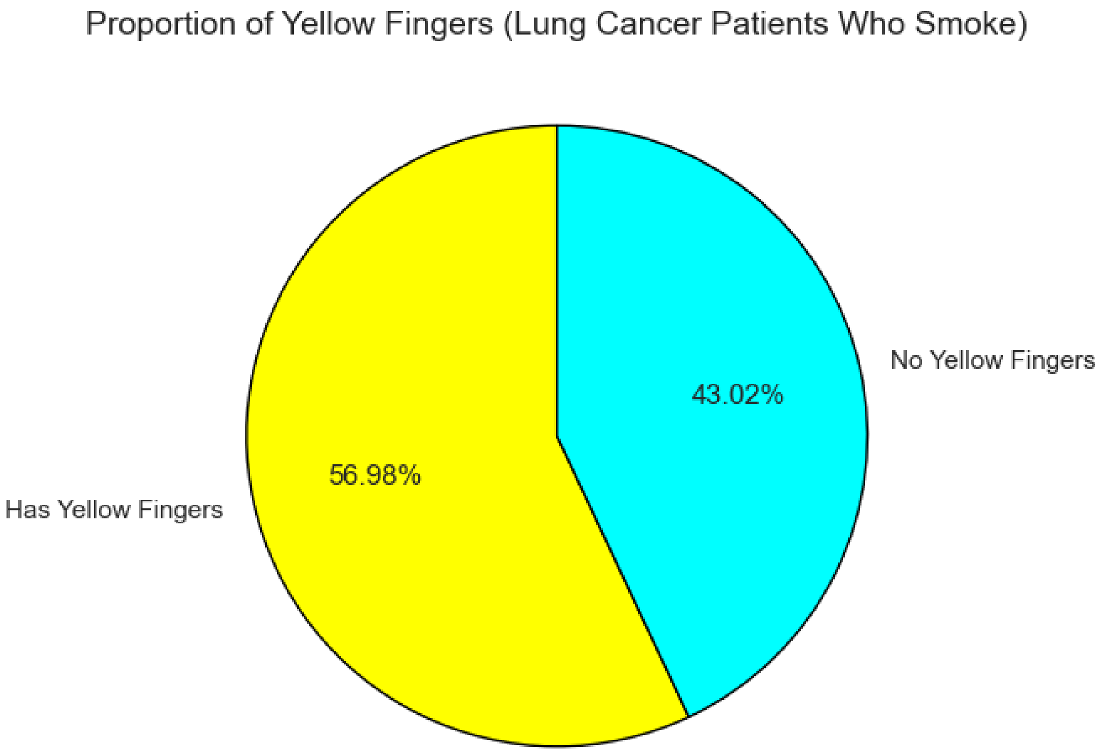


Figure. 6

[89]:
```python
labels = ['Has Yellow Fingers', 'No Yellow Fingers']
proportion_nonsmoker_yellow_finger =␣
 ↪contingency_smokingyellowfingers['Proportion_Yellow_Fingers'][0.0]
proportions = [proportion_nonsmoker_yellow_finger,␣
 ↪1-proportion_nonsmoker_yellow_finger]

plt.figure(figsize=(6, 6))
plt.pie(proportions, labels=labels, autopct='%1.2f%%', startangle=90,␣
 ↪colors=['lightyellow', 'lightblue'], wedgeprops={'edgecolor': 'black'})

plt.title('Proportion of Yellow Fingers (Lung Cancer Patients Who Do not␣
 ↪Smoke)', fontsize=14)
plt.figtext(0.47, -0.02, "Figure. 7", fontsize=15)

plt.show()
```

```
[89]: <Figure size 600x600 with 0 Axes>
```

```
[89]: ([<matplotlib.patches.Wedge at 0x1698d54f0>,
        <matplotlib.patches.Wedge at 0x1698d54c0>],
       [Text(-1.0513753154658687, -0.32343460858270096, 'Has Yellow Fingers'),
        Text(1.0513752851837306, 0.3234347070195606, 'No Yellow Fingers')],
       [Text(-0.5734774447995648, -0.17641887740874596, '59.50%'),
        Text(0.5734774282820349, 0.1764189311015785, '40.50%')])
```

```
[89]: Text(0.5, 1.0, 'Proportion of Yellow Fingers (Lung Cancer Patients Who Do not
      Smoke)')
```

```
[89]: Text(0.47, -0.02, 'Figure. 7')
```

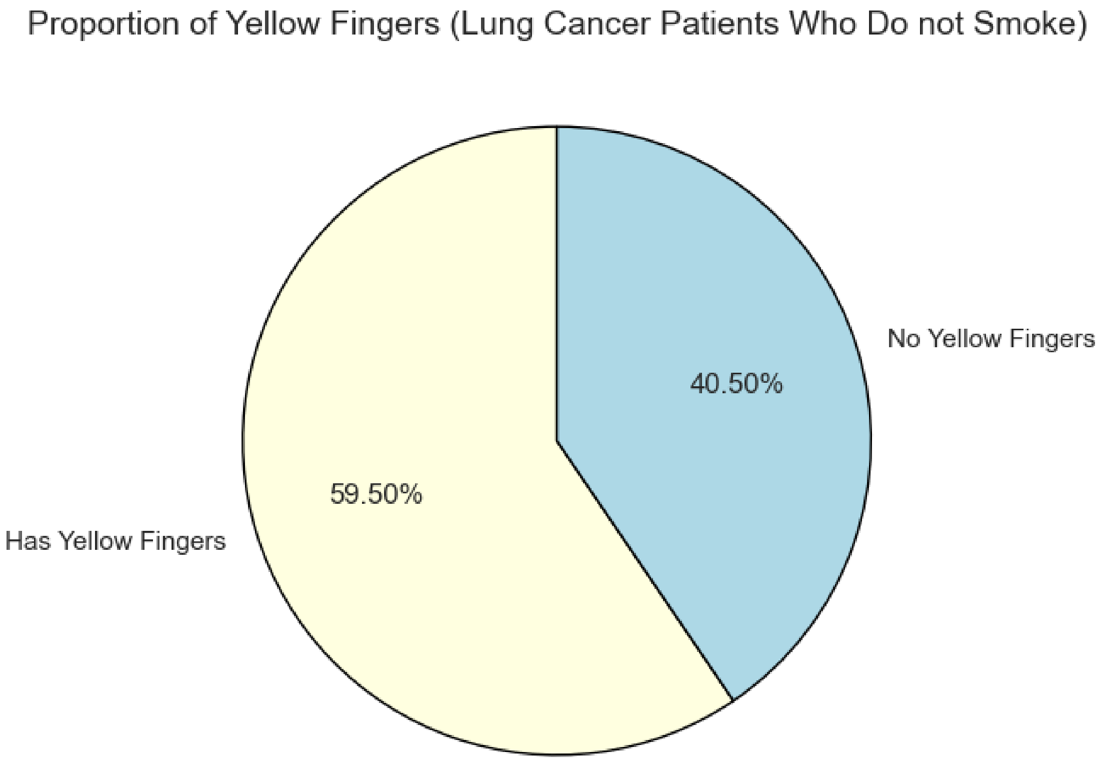


Figure. 7

4 Alcohol Consumption

[93]:
```
contingency_alcohol = pd.crosstab(final_df['ALCOHOL CONSUMING'],␣
 ↪final_df['LUNG_CANCER'])
contingency_alcohol['Proportion_Cancer'] = contingency_alcohol[1] /␣
 ↪(contingency_alcohol[0] + contingency_alcohol[1])

proportions_alcohol = contingency_alcohol['Proportion_Cancer']
non_cancer_alcohol = 1 - proportions_alcohol

plt.figure(figsize=(8, 4))
sns.set_theme(style="whitegrid")
y_labels_alcohol = ['Non-Alcohol Consuming', 'Alcohol Consuming']
bar_width = 0.3
#create horizontal bar chart for cancer patients who drink/do not drink using␣
 ↪labels above and sns pastel color palette
bars_lung_cancer_alcohol = plt.barh(y_labels_alcohol, proportions_alcohol,␣
 ↪color=['#58d68d', '#d2b4de'], height=bar_width, label='Lung Cancer')
#create horizontal bar chart for non-cancer patient who drink/do not drink␣
 ↪using labels above and sns pastel color palette
bars_no_cancer_alcohol = plt.barh(y_labels_alcohol, proportions_alcohol,␣
 ↪color='lightgrey', left=proportions_alcohol, height=bar_width, label='No␣
 ↪Lung Cancer')

plt.title('Proportion of Lung Cancer Cases by Drinking Status', fontsize=16)
plt.xlabel('Proportion', fontsize=14)
plt.xlim(0, 1)
plt.ylim(-0.5, 1.5)

#annotate % of cancer patients for drinkers and non-drinkers group
for index, value in enumerate(proportions_yellowfingers):
    plt.text(value - 0.4, index, f"{value*100:.2f}%", va='center', fontsize=12,␣
 ↪color='black')
plt.figtext(0.47, -0.1, "Figure. 8", fontsize=15)

plt.show()
```

[93]:
```
<Figure size 800x400 with 0 Axes>
```

[93]:
```
Text(0.5, 1.0, 'Proportion of Lung Cancer Cases by Drinking Status')
```

[93]:
```
Text(0.5, 0, 'Proportion')
```

[93]:
```
(0.0, 1.0)
```

[93]:
```
(-0.5, 1.5)
```

```
[93]: Text(0.31374764595103577, 0, '71.37%')

[93]: Text(0.48657407407407405, 1, '88.66%')

[93]: Text(0.47, -0.1, 'Figure. 8')
```

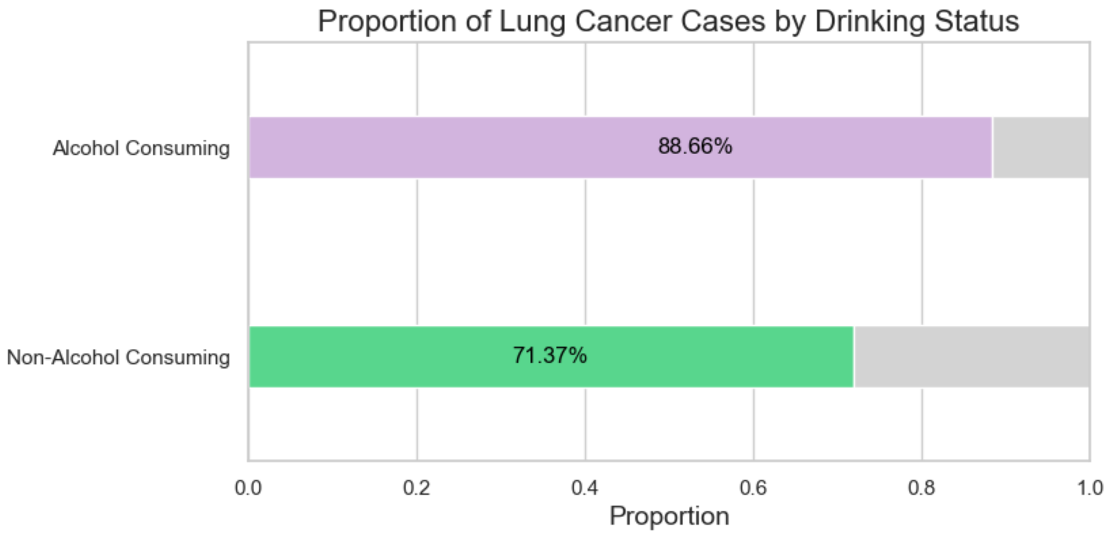


Figure. 8

```
[100]: final_df[['LUNG_CANCER','SMOKING', 'YELLOW_FINGERS', 'ALCOHOL CONSUMING']].
        ↪corr()
       display(HTML("<p style='text-align: center; font-weight: bold;'>Figure. 9:␣
        ↪Correlation Matrix</p>"))
```

```
[100]:                    LUNG_CANCER   SMOKING  YELLOW_FINGERS  ALCOHOL CONSUMING
       LUNG_CANCER           1.000000  0.063074        0.217762           0.207659
       SMOKING               0.063074  1.000000        0.001875          -0.019939
       YELLOW_FINGERS        0.217762  0.001875        1.000000          -0.075707
       ALCOHOL CONSUMING     0.207659 -0.019939       -0.075707           1.000000

      <IPython.core.display.HTML object>
```

5 Analysis on having multiple lifestyle habits

```
[102]: contingency_table_combinations = pd.crosstab([final_df['SMOKING'],␣
        ↪final_df['YELLOW_FINGERS'], final_df['ALCOHOL CONSUMING']],␣
        ↪final_df['LUNG_CANCER'])
       contingency_table_combinations['Proportion_Cancer'] =␣
        ↪contingency_table_combinations[1] / (contingency_table_combinations[0] +␣
        ↪contingency_table_combinations[1])
```

```
contingency_table_combinations.sort_values(by='Proportion_Cancer', inplace=True)
labels = ['None','Smoking only','Alcohol only','Yellow Fingers␣
 ↪only','Smoking+Yellow fingers','Smoking+Alcohol','Yellow fingers␣
 ↪+Alcohol','Smoking+Yellow fingers +Alcohol']

plt.figure(figsize=(10, 6))
sns.set_theme(style="whitegrid")

bar_width = 0.3
bars = plt.barh(labels, contingency_table_combinations['Proportion_Cancer'],␣
 ↪color=sns.color_palette("pastel", 8), height=bar_width)

plt.title('Proportion of Lung Cancer by Lifestyle Factors', fontsize=16)
plt.xlabel('Proportion', fontsize=14)
plt.xlim(0, 1)

# annotate % of cancer patients for each combination
for index, value in␣
 ↪enumerate(contingency_table_combinations['Proportion_Cancer']):
    plt.text(value + 0.02, index, f"{value*100:.2f}%", va='center',␣
 ↪fontsize=12, color='black')
plt.figtext(0.47, -0.045, "Figure. 10", fontsize=15)

plt.show()
```

```
[102]: <Figure size 1000x600 with 0 Axes>

[102]: Text(0.5, 1.0, 'Proportion of Lung Cancer by Lifestyle Factors')

[102]: Text(0.5, 0, 'Proportion')

[102]: (0.0, 1.0)

[102]: Text(0.4617808219178082, 0, '44.18%')

[102]: Text(0.6070236869207003, 1, '58.70%')

[102]: Text(0.8566294067067928, 2, '83.66%')

[102]: Text(0.8787921847246892, 3, '85.88%')

[102]: Text(0.9015384615384615, 4, '88.15%')

[102]: Text(0.9101453957996769, 5, '89.01%')

[102]: Text(0.9196509598603839, 6, '89.97%')
```

```
[102]: Text(0.9259322033898305, 7, '90.59%')
```

```
[102]: Text(0.47, -0.045, 'Figure. 10')
```

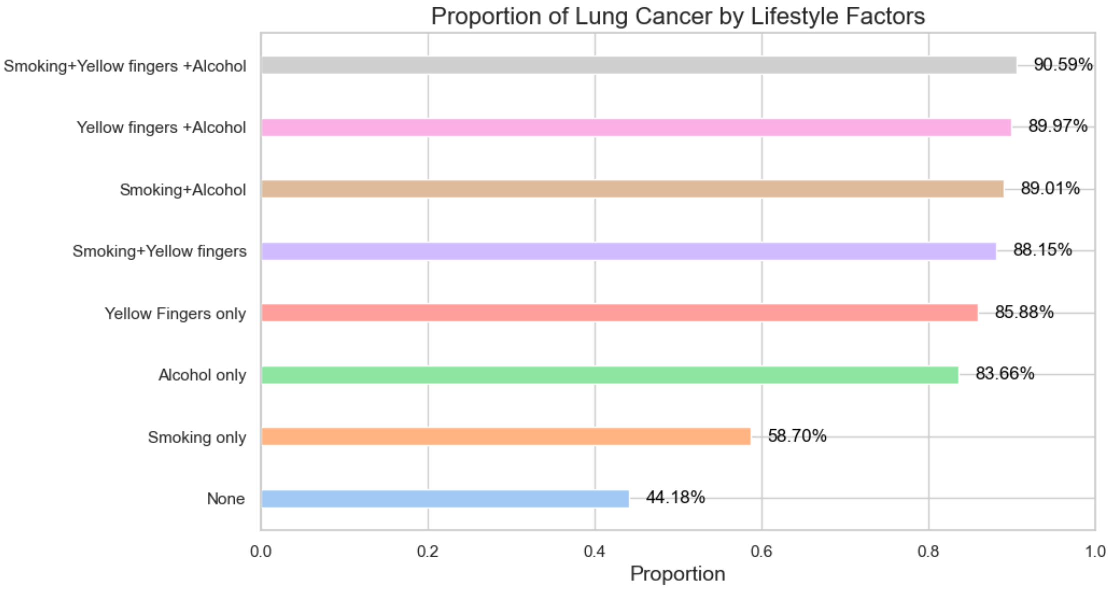


Figure. 10

6 By Gender

```
[106]: # SUBGROUP-GENDER
# Split gender dataframes
female = final_df[final_df['GENDER']==0]
male = final_df[final_df['GENDER']==1]
female = female.drop("GENDER", axis = 1)
male = male.drop("GENDER", axis = 1)

# Calculate correlations between Lung Cancer and other variables for␣
 ↪middle-aged and senior groups
female_corr = female.corr()['LUNG_CANCER'].drop('LUNG_CANCER')
male_corr = male.corr()['LUNG_CANCER'].drop('LUNG_CANCER')

# Create a DataFrame to store correlations
correlation_df = pd.DataFrame({
    'Female': female_corr,
    'Male': male_corr
})

# Plotting bar chart for each symptom
```

```python
correlation_df.plot(kind='bar', figsize=(12, 8))

# Add labels and title
plt.xlabel('Symptoms')
plt.ylabel('Correlation with Lung Cancer')
plt.title('Comparison of Correlation with Lung Cancer: Female vs Male')
plt.xticks(rotation=45, ha='right')
plt.grid(True, axis = 'y')
plt.figtext(0.47, -0.045, "Figure. 11", fontsize=15)

# Show the plot
plt.tight_layout()
plt.show()
```

```
[106]: <Axes: >
```

```
[106]: Text(0.5, 0, 'Symptoms')
```

```
[106]: Text(0, 0.5, 'Correlation with Lung Cancer')
```

```
[106]: Text(0.5, 1.0, 'Comparison of Correlation with Lung Cancer: Female vs Male')
```

```
[106]: (array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13, 14]),
        [Text(0, 0, 'AGE'),
         Text(1, 0, 'SMOKING'),
         Text(2, 0, 'YELLOW_FINGERS'),
         Text(3, 0, 'ANXIETY'),
         Text(4, 0, 'PEER_PRESSURE'),
         Text(5, 0, 'CHRONIC DISEASE'),
         Text(6, 0, 'FATIGUE '),
         Text(7, 0, 'ALLERGY '),
         Text(8, 0, 'WHEEZING'),
         Text(9, 0, 'ALCOHOL CONSUMING'),
         Text(10, 0, 'COUGHING'),
         Text(11, 0, 'SHORTNESS OF BREATH'),
         Text(12, 0, 'SWALLOWING DIFFICULTY'),
         Text(13, 0, 'CHEST PAIN'),
         Text(14, 0, 'cluster')])
```

```
[106]: Text(0.47, -0.045, 'Figure. 11')
```

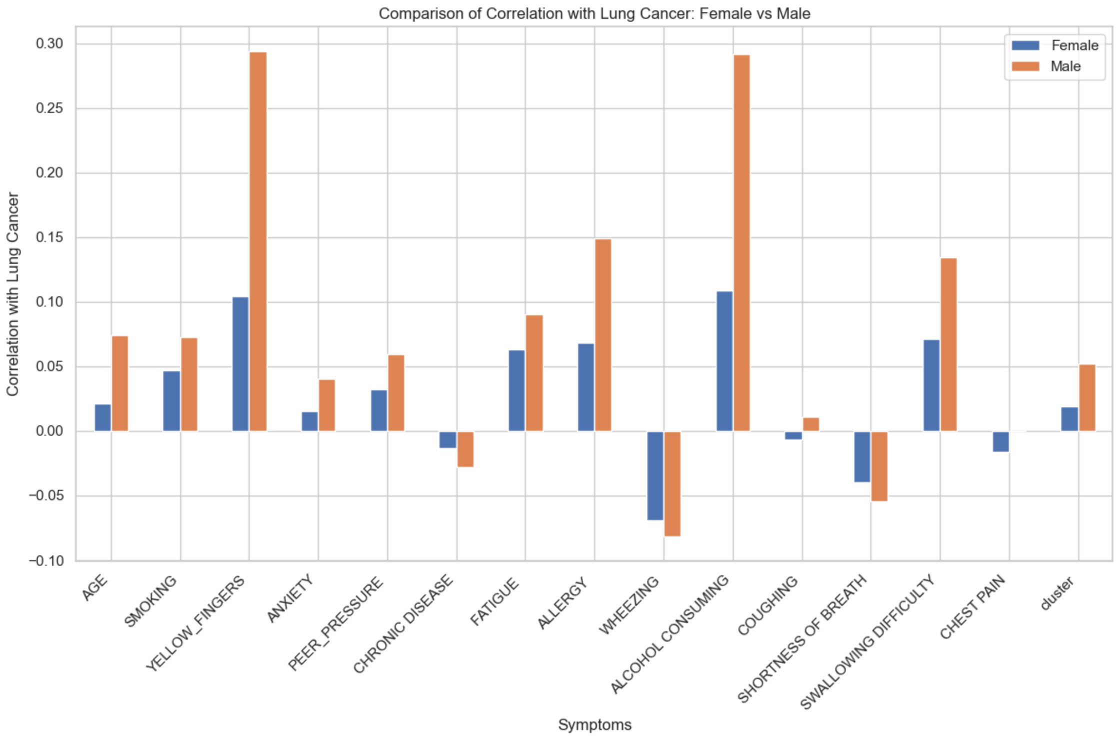


Figure. 11

Male Proportions

[108]:
```
#SMOKING
contingency_smoking_men = pd.crosstab(male['SMOKING'], male['LUNG_CANCER'])
contingency_smoking_men['Proportion_Cancer'] = contingency_smoking_men[1] /␣
 ↪(contingency_smoking_men[0] + contingency_smoking_men[1])


proportions_smoking_men = contingency_smoking_men['Proportion_Cancer']
non_cancer_smoking_men = 1 - proportions_smoking_men


plt.figure(figsize=(8, 4))
sns.set_theme(style="whitegrid")
y_labels_smoking = ['Non-Smoker', 'Smoker']
bar_width = 0.3 #change the width for a better fit figure
#create horizontal bar chart for cancer patients who smoke/do not smoke using␣
 ↪labels above and sns pastel color palette
bars_lung_cancer_smoking = plt.barh(y_labels_smoking, proportions_smoking_men,␣
 ↪color=sns.color_palette("pastel", 2), height=bar_width)
#create horizontal bar chart for cancer patients who smoke/ do not smoke using␣
 ↪labels above and sns pastel color palette
```

```
bars_no_cancer_smoking = plt.barh(y_labels_smoking, non_cancer_smoking_men,␣
 ↪color='lightgrey', left=proportions_smoking_men, height=bar_width)

plt.title('Male: Proportion of Lung Cancer Cases by Smoking Status',␣
 ↪fontsize=16)
plt.xlabel('Proportion', fontsize=14)
plt.xlim(0, 1)
plt.ylim(-0.5, 1.5)

#to annotate the % of cancer patients for smoker and non-smoker group
for index, value in enumerate(proportions_smoking_men):
    plt.text(value - 0.4, index, f"{value*100:.2f}%", va='center', fontsize=12,␣
 ↪color='black')

plt.figtext(0.45, -0.1, "Figure. 12", fontsize=15)

plt.show()

#YELLOW FINGERS
contingency_yellowfingers_men = pd.crosstab(male['YELLOW_FINGERS'],␣
 ↪male['LUNG_CANCER'])
contingency_yellowfingers_men['Proportion_Cancer'] =␣
 ↪contingency_yellowfingers_men[1] / (contingency_yellowfingers_men[0] +␣
 ↪contingency_yellowfingers_men[1])

proportions_yellowfingers_men =␣
 ↪contingency_yellowfingers_men['Proportion_Cancer']
non_cancer_yellowfingers_men = 1 - proportions_yellowfingers_men

plt.figure(figsize=(8, 4))
sns.set_theme(style="whitegrid")
y_labels_yellowfingers = ['No Yellow Fingers', 'Yellow Fingers']
bar_width = 0.3
#create horizontal bar chart for cancer patients with or without yellow fingers␣
 ↪using labels above and sns pastel color palette
bars_lung_cancer_yellowfingers = plt.barh(y_labels_yellowfingers,␣
 ↪proportions_yellowfingers_men, color=['#FFCC99','#FF9999'],␣
 ↪height=bar_width, label='Lung Cancer')
#create horizontal bar chart for non-cancer patients with or without yellow␣
 ↪fingers using labels above and sns pastel color palette
bars_no_cancer_yellowfingers = plt.barh(y_labels_yellowfingers,␣
 ↪non_cancer_yellowfingers_men, color='lightgrey',␣
 ↪left=proportions_yellowfingers_men, height=bar_width, label='No Lung Cancer')

plt.title('Male: Proportion of Lung Cancer Cases by Presence Of Yellow␣
 ↪Fingers', fontsize=16)
```

```
plt.xlabel('Proportion', fontsize=14)
plt.xlim(0, 1)
plt.ylim(-0.5, 1.5)

#annotate % of cancer patients for yellow finger and non-yellow finger group
for index, value in enumerate(proportions_yellowfingers_men):
    plt.text(value - 0.4, index, f"{value*100:.2f}%", va='center', fontsize=12,␣
 ↪color='black')

plt.figtext(0.45, -0.1, "Figure. 13", fontsize=15)

plt.show()

#ALCOHOL CONSUMING
contingency_alcohol_men = pd.crosstab(male['ALCOHOL CONSUMING'],␣
 ↪male['LUNG_CANCER'])
contingency_alcohol_men['Proportion_Cancer'] = contingency_alcohol_men[1] /␣
 ↪(contingency_alcohol_men[0] + contingency_alcohol_men[1])

proportions_alcohol_men = contingency_alcohol_men['Proportion_Cancer']
non_cancer_alcohol_men = 1 - proportions_alcohol_men

plt.figure(figsize=(8, 4))
sns.set_theme(style="whitegrid")
y_labels_alcohol = ['Non-Alcohol Consuming', 'Alcohol Consuming']
bar_width = 0.3
#create horizontal bar chart for cancer patients who drink/do not drink using␣
 ↪labels above and sns pastel color palette
bars_lung_cancer_alcohol = plt.barh(y_labels_alcohol, proportions_alcohol_men,␣
 ↪color=['#58d68d', '#d2b4de'], height=bar_width, label='Lung Cancer')
#create horizontal bar chart for non-cancer patient who drink/do not drink␣
 ↪using labels above and sns pastel color palette
bars_no_cancer_alcohol = plt.barh(y_labels_alcohol, non_cancer_alcohol_men,␣
 ↪color='lightgrey', left=proportions_alcohol_men, height=bar_width, label='No␣
 ↪Lung Cancer')

plt.title('Male: Proportion of Lung Cancer Cases by Drinking Status',␣
 ↪fontsize=16)
plt.xlabel('Proportion', fontsize=14)
plt.xlim(0, 1)
plt.ylim(-0.5, 1.5)

#annotate % of cancer patients for drinkers and non-drinkers group
for index, value in enumerate(proportions_alcohol_men):
    plt.text(value - 0.4, index, f"{value*100:.2f}%", va='center', fontsize=12,␣
 ↪color='black')
```

```
plt.figtext(0.45, -0.1, "Figure. 14", fontsize=15)

plt.show()
```

```
[108]: <Figure size 800x400 with 0 Axes>

[108]: Text(0.5, 1.0, 'Male: Proportion of Lung Cancer Cases by Smoking Status')

[108]: Text(0.5, 0, 'Proportion')

[108]: (0.0, 1.0)

[108]: (-0.5, 1.5)

[108]: Text(0.32058194266153184, 0, '72.06%')

[108]: Text(0.38353413654618473, 1, '78.35%')

[108]: Text(0.45, -0.1, 'Figure. 12')
```

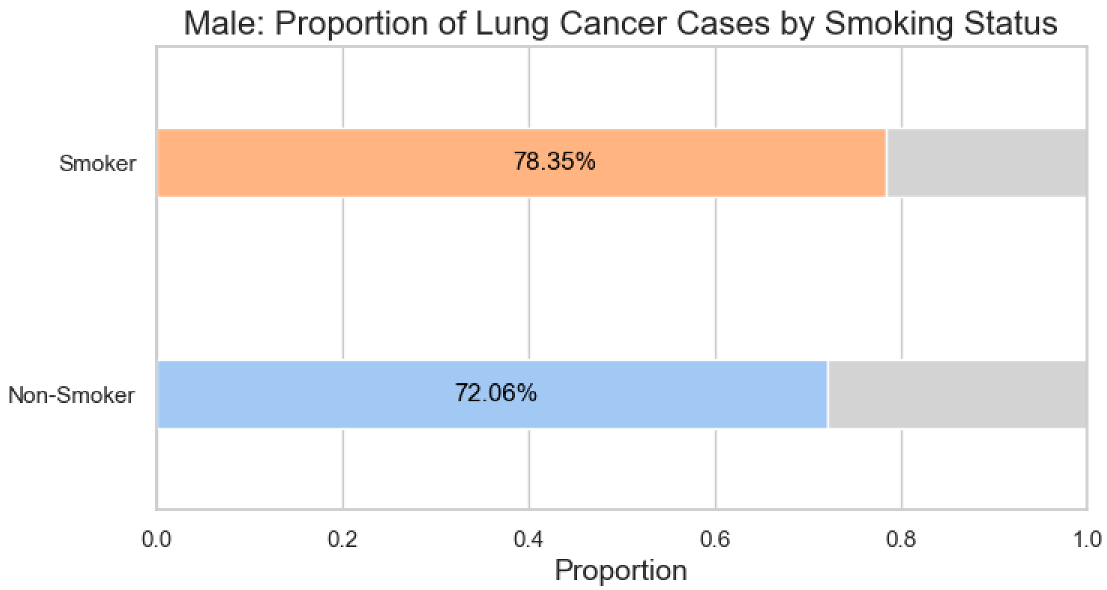


Figure. 12

```
[108]: <Figure size 800x400 with 0 Axes>

[108]: Text(0.5, 1.0, 'Male: Proportion of Lung Cancer Cases by Presence Of Yellow
       Fingers')

[108]: Text(0.5, 0, 'Proportion')
```

```
[108]: (0.0, 1.0)

[108]: (-0.5, 1.5)

[108]: Text(0.22297872340425529, 0, '62.30%')

[108]: Text(0.47646346386758176, 1, '87.65%')

[108]: Text(0.45, -0.1, 'Figure. 13')
```

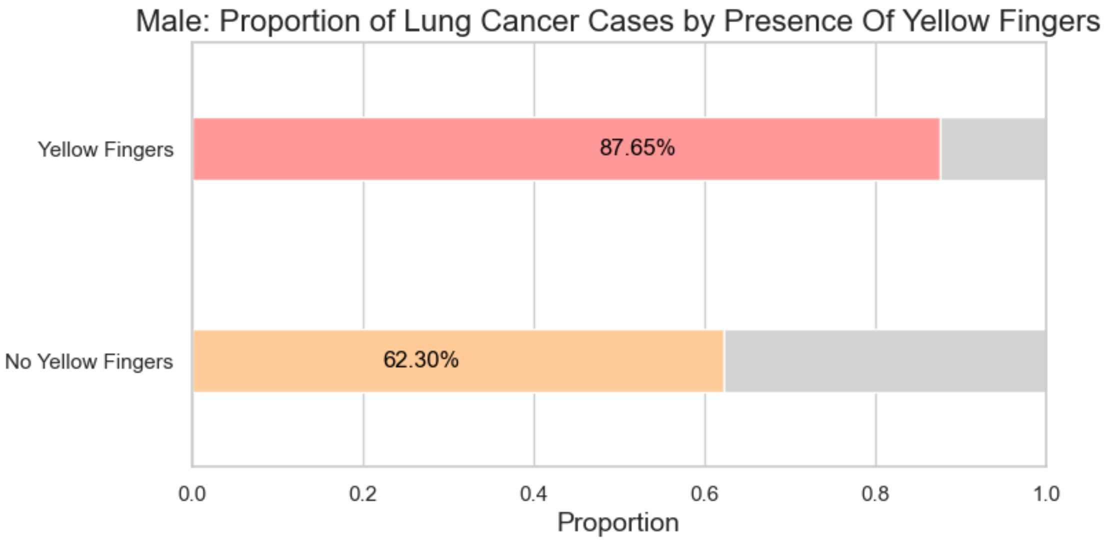


Figure. 13

```
[108]: <Figure size 800x400 with 0 Axes>

[108]: Text(0.5, 1.0, 'Male: Proportion of Lung Cancer Cases by Drinking Status')

[108]: Text(0.5, 0, 'Proportion')

[108]: (0.0, 1.0)

[108]: (-0.5, 1.5)

[108]: Text(0.21562925942753297, 0, '61.56%')

[108]: Text(0.46824067022086824, 1, '86.82%')

[108]: Text(0.45, -0.1, 'Figure. 14')
```

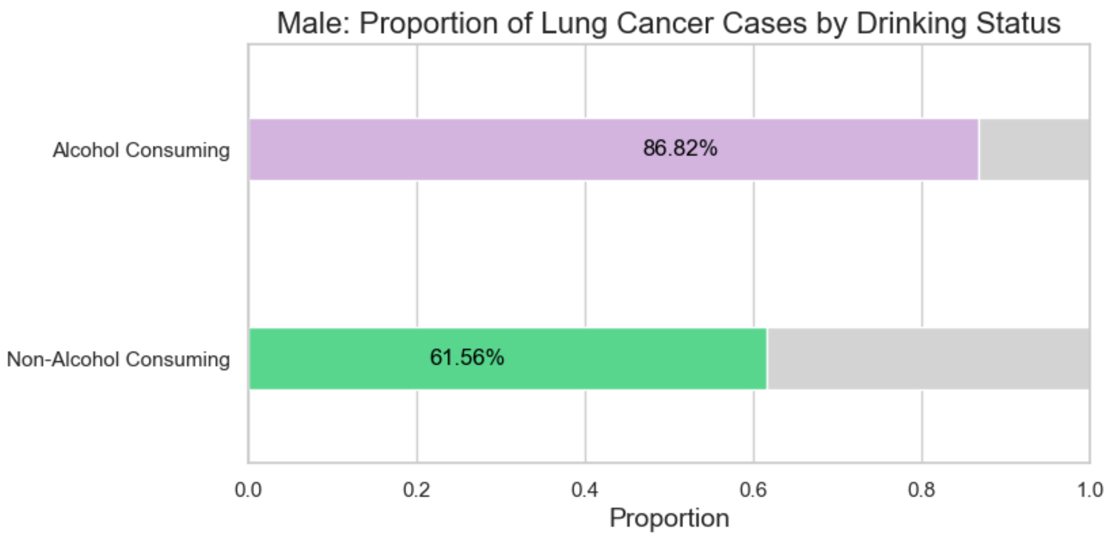


Figure. 14

Female Proportions

[110]:

```
#SMOKING
contingency_smoking_women = pd.crosstab(female['SMOKING'], female['LUNG_CANCER'])
contingency_smoking_women['Proportion_Cancer'] = contingency_smoking_women[1] / (contingency_smoking_women[0] + contingency_smoking_women[1])


proportions_smoking_women = contingency_smoking_women['Proportion_Cancer']
non_cancer_smoking_women = 1 - proportions_smoking_women


plt.figure(figsize=(8, 4))
sns.set_theme(style="whitegrid")
y_labels_smoking = ['Non-Smoker', 'Smoker']
bar_width = 0.3 #change the width for a better fit figure
#create horizontal bar chart for cancer patients who smoke/do not smoke using labels above and sns pastel color palette
bars_lung_cancer_smoking = plt.barh(y_labels_smoking, proportions_smoking_women, color=sns.color_palette("pastel", 2), height=bar_width)
#create horizontal bar chart for cancer patients who smoke/ do not smoke using labels above and sns pastel color palette
bars_no_cancer_smoking = plt.barh(y_labels_smoking, non_cancer_smoking_women, color='lightgrey', left=proportions_smoking_women, height=bar_width)
```

```
plt.title('Female: Proportion of Lung Cancer Cases by Smoking Status', fontsize=16)
plt.xlabel('Proportion', fontsize=14)
plt.xlim(0, 1)
plt.ylim(-0.5, 1.5)

#to annotate the % of cancer patients for smoker and non-smoker group
for index, value in enumerate(proportions_smoking_women):
    plt.text(value - 0.4, index, f"{value*100:.2f}%", va='center', fontsize=12, color='black')

plt.figtext(0.45, -0.1, "Figure. 15", fontsize=15)

plt.show()

#YELLOW FINGERS
contingency_yellowfingers_women = pd.crosstab(female['YELLOW_FINGERS'], female['LUNG_CANCER'])
contingency_yellowfingers_women['Proportion_Cancer'] = contingency_yellowfingers_women[1] / (contingency_yellowfingers_women[0] + contingency_yellowfingers_women[1])

proportions_yellowfingers_women = contingency_yellowfingers_women['Proportion_Cancer']
non_cancer_yellowfingers_women = 1 - proportions_yellowfingers_women

plt.figure(figsize=(8, 4))
sns.set_theme(style="whitegrid")
y_labels_yellowfingers = ['No Yellow Fingers', 'Yellow Fingers']
bar_width = 0.3
#create horizontal bar chart for cancer patients with or without yellow fingers using labels above and sns pastel color palette
bars_lung_cancer_yellowfingers = plt.barh(y_labels_yellowfingers, proportions_yellowfingers_women, color=['#FFCC99','#FF9999'], height=bar_width, label='Lung Cancer')
#create horizontal bar chart for non-cancer patients with or without yellow fingers using labels above and sns pastel color palette
bars_no_cancer_yellowfingers = plt.barh(y_labels_yellowfingers, non_cancer_yellowfingers_women, color='lightgrey', left=proportions_yellowfingers_women, height=bar_width, label='No Lung Cancer')

plt.title('Female: Proportion of Lung Cancer Cases by Presence Of Yellow Fingers', fontsize=16)
plt.xlabel('Proportion', fontsize=14)
plt.xlim(0, 1)
plt.ylim(-0.5, 1.5)

#annotate % of cancer patients for yellow finger and non-yellow finger group
for index, value in enumerate(proportions_yellowfingers_women):
    plt.text(value - 0.4, index, f"{value*100:.2f}%", va='center', fontsize=12,␣
 ↪color='black')

plt.figtext(0.45, -0.1, "Figure. 16", fontsize=15)

plt.show()

#ALCOHOL CONSUMING
contingency_alcohol_women = pd.crosstab(female['ALCOHOL CONSUMING'],␣
 ↪female['LUNG_CANCER'])
contingency_alcohol_women['Proportion_Cancer'] = contingency_alcohol_women[1] /␣
 ↪(contingency_alcohol_women[0] + contingency_alcohol_women[1])

proportions_alcohol_women = contingency_alcohol_women['Proportion_Cancer']
non_cancer_alcohol_women = 1 - proportions_alcohol_women

plt.figure(figsize=(8, 4))
sns.set_theme(style="whitegrid")
y_labels_alcohol = ['Non-Alcohol Consuming', 'Alcohol Consuming']
bar_width = 0.3
#create horizontal bar chart for cancer patients who drink/do not drink using␣
 ↪labels above and sns pastel color palette
bars_lung_cancer_alcohol = plt.barh(y_labels_alcohol,␣
 ↪proportions_alcohol_women, color=['#58d68d', '#d2b4de'], height=bar_width,␣
 ↪label='Lung Cancer')
#create horizontal bar chart for non-cancer patient who drink/do not drink␣
 ↪using labels above and sns pastel color palette
bars_no_cancer_alcohol = plt.barh(y_labels_alcohol, non_cancer_alcohol_women,␣
 ↪color='lightgrey', left=proportions_alcohol_women, height=bar_width,␣
 ↪label='No Lung Cancer')

plt.title('Female: Proportion of Lung Cancer Cases by Drinking Status',␣
 ↪fontsize=16)
plt.xlabel('Proportion', fontsize=14)
plt.xlim(0, 1)
plt.ylim(-0.5, 1.5)

#annotate % of cancer patients for drinkers and non-drinkers group
for index, value in enumerate(proportions_alcohol_women):
    plt.text(value - 0.4, index, f"{value*100:.2f}%", va='center', fontsize=12,␣
 ↪color='black')
```

```
plt.figtext(0.45, -0.1, "Figure. 17", fontsize=15)

plt.show()
```

```
[110]: <Figure size 800x400 with 0 Axes>

[110]: Text(0.5, 1.0, 'Female: Proportion of Lung Cancer Cases by Smoking Status')

[110]: Text(0.5, 0, 'Proportion')

[110]: (0.0, 1.0)

[110]: (-0.5, 1.5)

[110]: Text(0.4480243161094225, 0, '84.80%')

[110]: Text(0.480400181900864, 1, '88.04%')

[110]: Text(0.45, -0.1, 'Figure. 15')
```

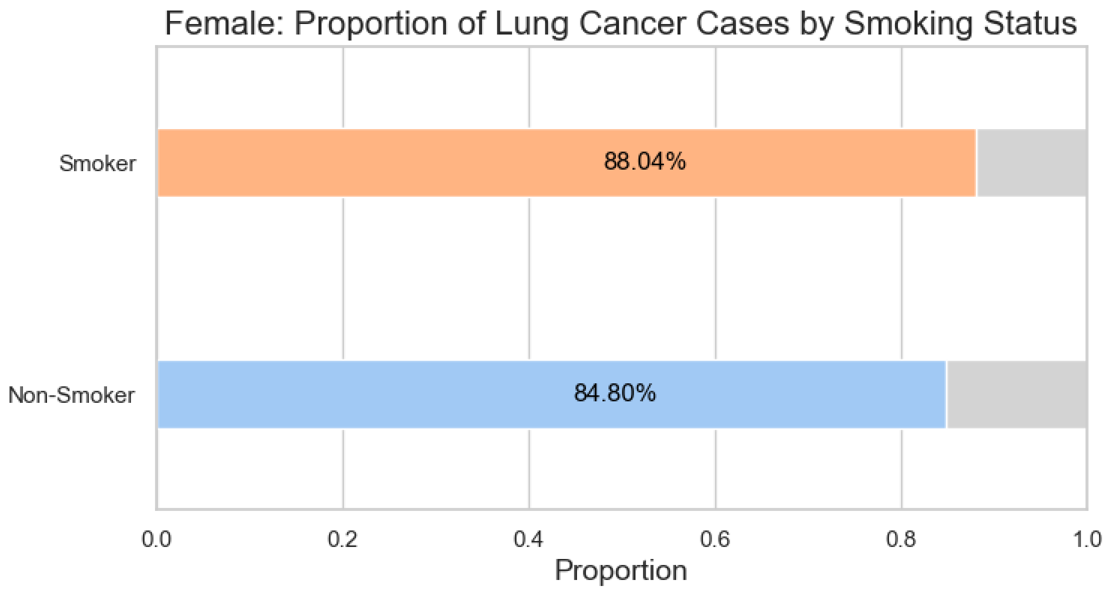


Figure. 15

```
[110]: <Figure size 800x400 with 0 Axes>

[110]: Text(0.5, 1.0, 'Female: Proportion of Lung Cancer Cases by Presence Of Yellow
       Fingers')

[110]: Text(0.5, 0, 'Proportion')
```

```
[110]: (0.0, 1.0)

[110]: (-0.5, 1.5)

[110]: Text(0.4261327713382508, 0, '82.61%')

[110]: Text(0.4975824175824176, 1, '89.76%')

[110]: Text(0.45, -0.1, 'Figure. 16')
```

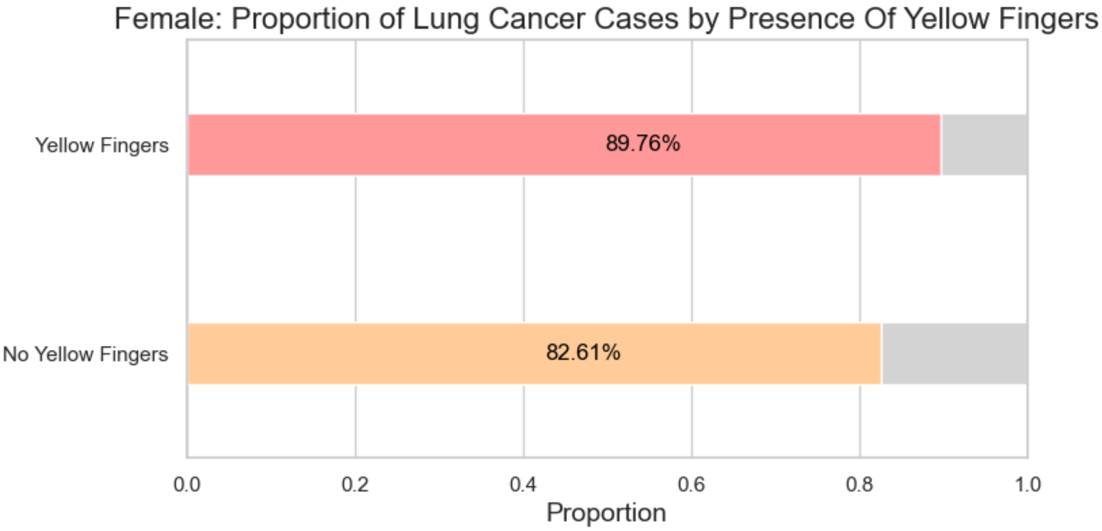


Figure. 16

```
[110]: <Figure size 800x400 with 0 Axes>

[110]: Text(0.5, 1.0, 'Female: Proportion of Lung Cancer Cases by Drinking Status')

[110]: Text(0.5, 0, 'Proportion')

[110]: (0.0, 1.0)

[110]: (-0.5, 1.5)

[110]: Text(0.4277027027027027, 0, '82.77%')

[110]: Text(0.5019514516896716, 1, '90.20%')

[110]: Text(0.45, -0.1, 'Figure. 17')
```

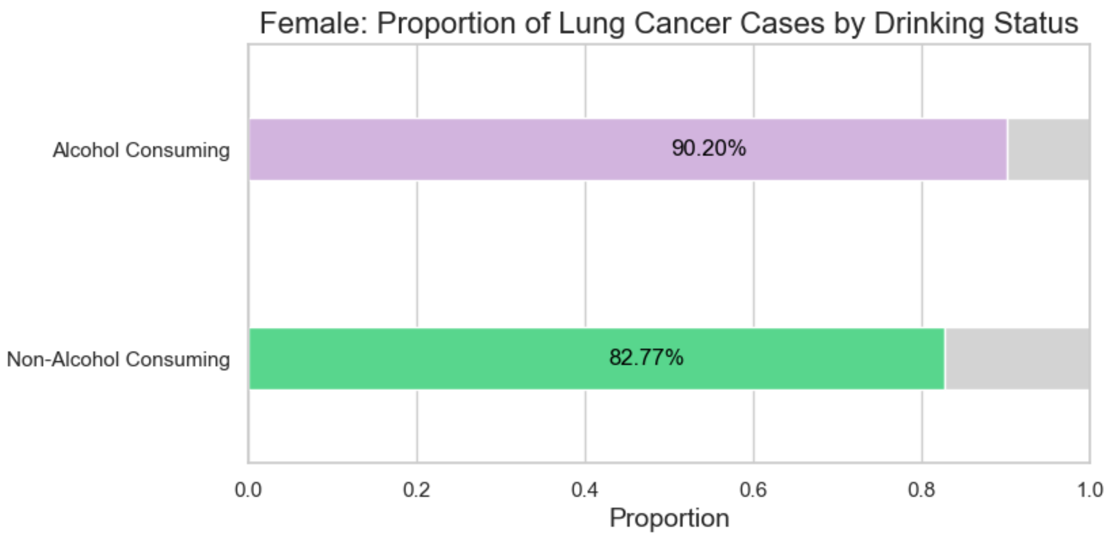


Figure. 17

7 Analysis on lifestyle habits by subgroups

[ ]:

```
# Clustering method # DO NOT RUN
from kmodes.kprototypes import KPrototypes

# Define the indices of the numerical and categorical features
cat_cols = [0,2,3,4,5,6,7,8,9,10,11,12,13,14,15]
num_cols = [1] # AGE

# Calculate cost for different cluster numbers
costs = []
for num_clusters in range(1, 10): # From 1 to 9
    kproto = KPrototypes(n_clusters=num_clusters, init='Cao', n_init=5,␣
 ↪verbose=2)
    kproto.fit_predict(final_df, categorical= cat_cols)
    costs.append(kproto.cost_)

# Plot the elbow curve, selection of optimal number of cluster via graphical␣
 ↪approach
plt.plot(range(1, 10), costs, marker='o')
plt.title('Elbow Method for Optimal Clusters')
plt.xlabel('Number of Clusters')
plt.ylabel('Cost')
plt.figtext(0.45, -0.1, "Figure. 18", fontsize=15)
plt.show();
```

[112]:

```
# Showcase K modes clustering graphically
import matplotlib.pyplot as plt
import numpy as np

# Simulating data for 3 clusters
np.random.seed(42)

# Cluster 1: Low cost
cluster_1_x = np.random.normal(1, 0.2, 20)
cluster_1_y = np.random.normal(1, 0.2, 20)

# Cluster 2: High cost
cluster_2_x = np.random.normal(5, 0.8, 20)
cluster_2_y = np.random.normal(5, 0.8, 20)

# Cluster 3: Low cost
cluster_3_x = np.random.normal(9, 0.2, 20)
cluster_3_y = np.random.normal(1, 0.2, 20)

# Plotting
plt.figure(figsize=(8, 6))

# Plot Cluster 1
plt.scatter(cluster_1_x, cluster_1_y, color='blue', label='Low Cost (Cluster␣
 ↪1)')
plt.annotate('Low Cost', (1, 1), textcoords="offset points", xytext=(10,10),␣
 ↪ha='center')

# Plot Cluster 2
plt.scatter(cluster_2_x, cluster_2_y, color='red', label='High Cost (Cluster␣
 ↪2)')
plt.annotate('High Cost', (5, 5), textcoords="offset points", xytext=(10,10),␣
 ↪ha='center')

# Plot Cluster 3
plt.scatter(cluster_3_x, cluster_3_y, color='green', label='Low Cost (Cluster␣
 ↪3)')
plt.annotate('Low Cost', (9, 1), textcoords="offset points", xytext=(10,10),␣
 ↪ha='center')

# Adding some center points for visual aid (centers of clusters)
plt.scatter([1, 5, 9], [1, 5, 1], color='black', marker='x', s=100,␣
 ↪label='Cluster Centers')

# Annotations for distance (Cost)
plt.arrow(5.5, 5.5, -0.5, -0.5, head_width=0.2, color='gray')
```

```python
plt.annotate('Cost (Distance)', (5.5, 5.5), textcoords="offset points",␣
 ↪xytext=(20,-10), ha='center')

# Setting the labels
plt.title('Clustering and Cost Visualization')
plt.xlabel('X')
plt.ylabel('Y')
plt.legend()
plt.figtext(0.45, -0.1, "Figure. 19", fontsize=15)
# Show the plot
plt.grid(True)
plt.show()
```

```
[112]: <Figure size 800x600 with 0 Axes>

[112]: <matplotlib.collections.PathCollection at 0x177bc2f50>

[112]: Text(10, 10, 'Low Cost')

[112]: <matplotlib.collections.PathCollection at 0x305aa5250>

[112]: Text(10, 10, 'High Cost')

[112]: <matplotlib.collections.PathCollection at 0x305aa6550>

[112]: Text(10, 10, 'Low Cost')

[112]: <matplotlib.collections.PathCollection at 0x30597f810>

[112]: <matplotlib.patches.FancyArrow at 0x30599b750>

[112]: Text(20, -10, 'Cost (Distance)')

[112]: Text(0.5, 1.0, 'Clustering and Cost Visualization')

[112]: Text(0.5, 0, 'X')

[112]: Text(0, 0.5, 'Y')

[112]: <matplotlib.legend.Legend at 0x305901a90>

[112]: Text(0.45, -0.1, 'Figure. 19')
```

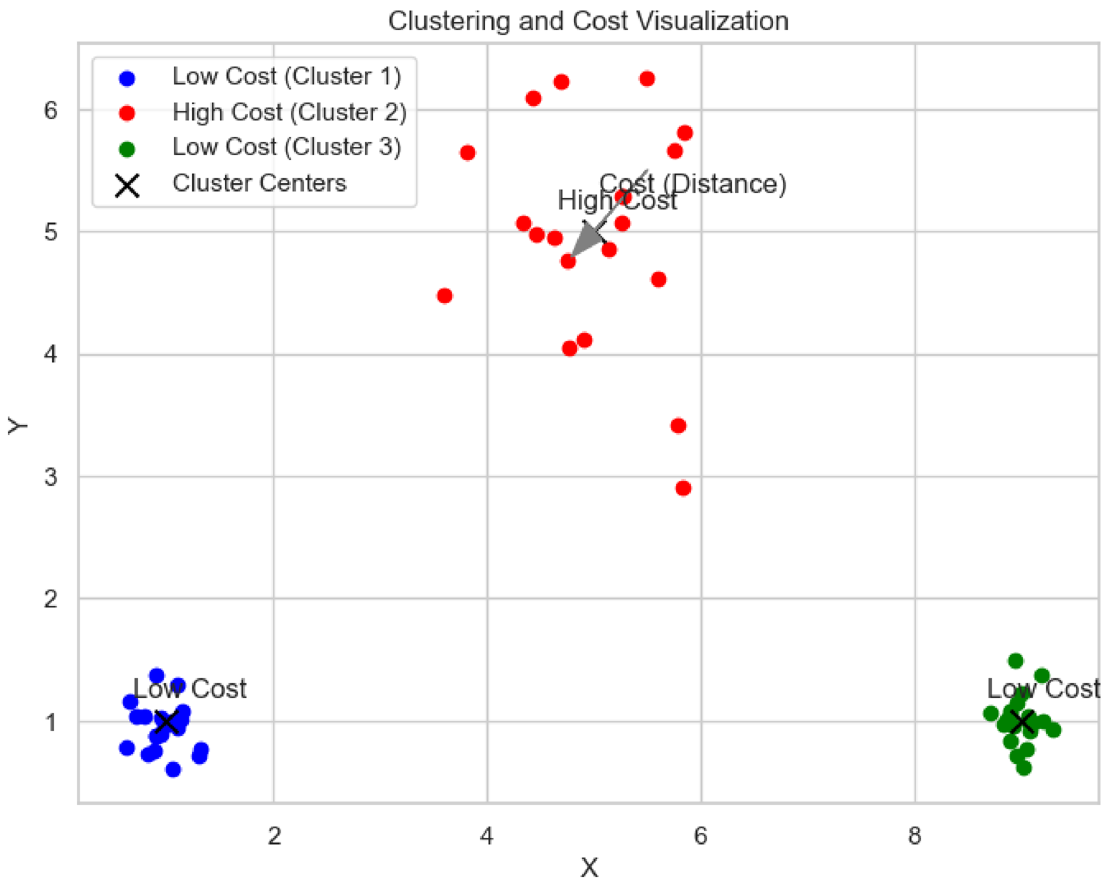


Figure. 19

[114]:

```
# Fit K-Prototypes model
# Create the final_df with cluster
from kmodes.kprototypes import KPrototypes

# Define the indices of the numerical and categorical features
cat_cols = [0,2,3,4,5,6,7,8,9,10,11,12,13,14,15]
num_cols = [1]
kproto = KPrototypes(n_clusters=2, init='Cao', verbose=2, random_state=42)
clusters = kproto.fit_predict(final_df, categorical=cat_cols)

final_df = final_df.copy()
# Assign cluster labels to the DataFrame
final_df['cluster'] = clusters

# See how clusters are grouped
```

```
age_cluster_summary = final_df.groupby('cluster')['AGE'].agg(['min', 'max'])
print(age_cluster_summary) # We see here, we have grouped cluster = 0 as age 61␣
 ↪to 81, and cluster = 1 as age 44 to 60.

# Invert cluster such that 1: Senior, 0: Middle-aged
final_df['cluster'] = 1 - final_df['cluster']
final_df['cluster'].value_counts() # Each clusters are approximately similarly␣
 ↪in number of observation.
```

```
Initialization method and algorithm are deterministic. Setting n_init to 1.
Init: initializing centroids
Init: initializing clusters
Starting iterations…
Run: 1, iteration: 1/100, moves: 837, ncost: 435139.83090474666
Run: 1, iteration: 2/100, moves: 367, ncost: 429324.7870668995
Run: 1, iteration: 3/100, moves: 193, ncost: 427676.13255939743
Run: 1, iteration: 4/100, moves: 0, ncost: 427676.13255939743
Init: initializing centroids
Init: initializing clusters
Starting iterations…
Run: 2, iteration: 1/100, moves: 1147, ncost: 440502.8957323666
Run: 2, iteration: 2/100, moves: 488, ncost: 430077.8332197589
Run: 2, iteration: 3/100, moves: 233, ncost: 427676.13255939743
Run: 2, iteration: 4/100, moves: 0, ncost: 427676.13255939743
Init: initializing centroids
Init: initializing clusters
Starting iterations…
Run: 3, iteration: 1/100, moves: 173, ncost: 429691.93374507886
Run: 3, iteration: 2/100, moves: 0, ncost: 429691.93374507886
Init: initializing centroids
Init: initializing clusters
Starting iterations…
Run: 4, iteration: 1/100, moves: 527, ncost: 431418.62927311804
Run: 4, iteration: 2/100, moves: 276, ncost: 428156.0480808848
Run: 4, iteration: 3/100, moves: 104, ncost: 427676.13255939743
Run: 4, iteration: 4/100, moves: 0, ncost: 427676.13255939743
Init: initializing centroids
Init: initializing clusters
Starting iterations…
Run: 5, iteration: 1/100, moves: 743, ncost: 432604.89000945486
Run: 5, iteration: 2/100, moves: 292, ncost: 428838.4025603321
Run: 5, iteration: 3/100, moves: 162, ncost: 427676.13255939743
Run: 5, iteration: 4/100, moves: 0, ncost: 427676.13255939743
Init: initializing centroids
Init: initializing clusters
Starting iterations…
Run: 6, iteration: 1/100, moves: 1316, ncost: 442618.20585507154
Run: 6, iteration: 2/100, moves: 527, ncost: 430596.7768482746
```

```
Run: 6, iteration: 3/100, moves: 252, ncost: 427721.7379616733
Run: 6, iteration: 4/100, moves: 32, ncost: 427676.13255939743
Run: 6, iteration: 5/100, moves: 0, ncost: 427676.13255939743
Init: initializing centroids
Init: initializing clusters
Starting iterations…
Run: 7, iteration: 1/100, moves: 254, ncost: 429691.93374507886
Run: 7, iteration: 2/100, moves: 0, ncost: 429691.93374507886
Init: initializing centroids
Init: initializing clusters
Starting iterations…
Run: 8, iteration: 1/100, moves: 1697, ncost: 448358.09653973917
Run: 8, iteration: 2/100, moves: 647, ncost: 430889.4782781323
Run: 8, iteration: 3/100, moves: 263, ncost: 427875.65720209875
Run: 8, iteration: 4/100, moves: 67, ncost: 427676.13255939743
Run: 8, iteration: 5/100, moves: 0, ncost: 427676.13255939743
Init: initializing centroids
Init: initializing clusters
Starting iterations…
Run: 9, iteration: 1/100, moves: 1765, ncost: 449650.7710782128
Run: 9, iteration: 2/100, moves: 670, ncost: 430950.33550199796
Run: 9, iteration: 3/100, moves: 262, ncost: 427926.0477886163
Run: 9, iteration: 4/100, moves: 75, ncost: 427676.13255939743
Run: 9, iteration: 5/100, moves: 0, ncost: 427676.13255939743
Init: initializing centroids
Init: initializing clusters
Starting iterations…
Run: 10, iteration: 1/100, moves: 356, ncost: 429691.93374507886
Run: 10, iteration: 2/100, moves: 0, ncost: 429691.93374507886
Best run was number 1
          min   max
cluster
0        61.0  81.0
1        44.0  60.0
```

[114]:

```
cluster
1    4612
0    4388
Name: count, dtype: int64
```

8 By Age

[116]:

```python
import matplotlib.pyplot as plt

# Separate the dataset and drop cluster
middle = final_df[final_df["cluster"]== 0]
senior = final_df[final_df["cluster"]== 1]
```

```
middle = middle.drop("cluster", axis = 1)
senior = senior.drop("cluster", axis = 1)

# Calculate correlations between Lung Cancer and other variables for␣
 ↪middle-aged and senior groups
middle_corr = middle.corr()['LUNG_CANCER'].drop('LUNG_CANCER')
senior_corr = senior.corr()['LUNG_CANCER'].drop('LUNG_CANCER')

# Create a DataFrame to store correlations
correlation_df = pd.DataFrame({
    'Middle Age': middle_corr,
    'Senior': senior_corr
})

# Plotting bar chart for each symptom
correlation_df.plot(kind='bar', figsize=(12, 8))

# Add labels and title
plt.xlabel('Symptoms')
plt.ylabel('Correlation with Lung Cancer')
plt.title('Comparison of Correlation with Lung Cancer: Middle Age vs Senior')
plt.xticks(rotation=45, ha='right')
plt.grid(True, axis = 'y')

plt.figtext(0.45, -0.1, "Figure. 20", fontsize=15)

# Show the plot
plt.tight_layout()
plt.show()
```

```
[116]: <Axes: >
```

```
[116]: Text(0.5, 0, 'Symptoms')
```

```
[116]: Text(0, 0.5, 'Correlation with Lung Cancer')
```

```
[116]: Text(0.5, 1.0, 'Comparison of Correlation with Lung Cancer: Middle Age vs
       Senior')
```

```
[116]: (array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13, 14]),
        [Text(0, 0, 'GENDER'),
         Text(1, 0, 'AGE'),
         Text(2, 0, 'SMOKING'),
         Text(3, 0, 'YELLOW_FINGERS'),
         Text(4, 0, 'ANXIETY'),
         Text(5, 0, 'PEER_PRESSURE'),
         Text(6, 0, 'CHRONIC DISEASE'),
         Text(7, 0, 'FATIGUE '),
         Text(8, 0, 'ALLERGY '),
         Text(9, 0, 'WHEEZING'),
         Text(10, 0, 'ALCOHOL CONSUMING'),
         Text(11, 0, 'COUGHING'),
         Text(12, 0, 'SHORTNESS OF BREATH'),
         Text(13, 0, 'SWALLOWING DIFFICULTY'),
         Text(14, 0, 'CHEST PAIN')])
```

```
[116]: Text(0.45, -0.1, 'Figure. 20')
```

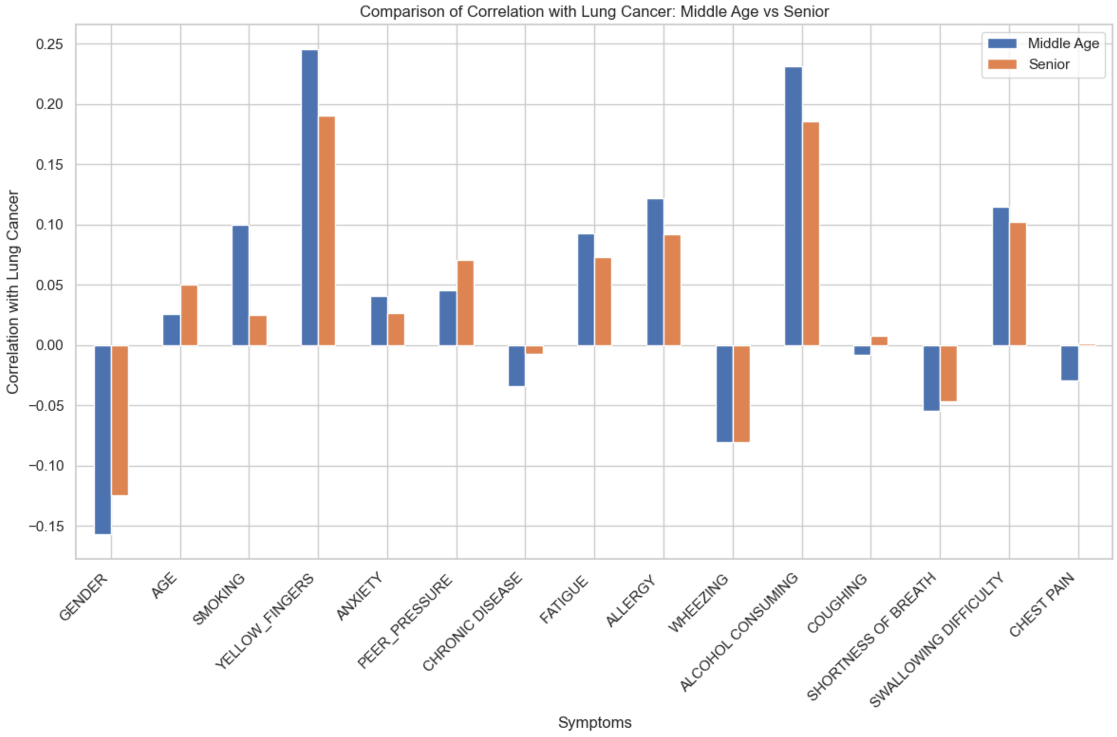


Figure. 20

```
[118]: import numpy as np
import pandas as pd
import statsmodels.api as sm

# Logistics regression on whole data
x = final_df.drop(labels =["LUNG_CANCER", "AGE"], axis = 1)
y = final_df["LUNG_CANCER"]

# Add a constant to the model (for the intercept)
```

```
X = sm.add_constant(x)

# Fit the logistic regression model using statsmodels
logit_model = sm.Logit(y, X)
result = logit_model.fit()

# Print summary of the model
print(f"Whole dataset: {result.summary()}")
display(HTML("<p style='text-align: center; font-weight: bold;'>Figure. 21</
 ↪p>"))
# Logistics regression on middle age
x = middle.drop(labels =["LUNG_CANCER", "AGE"], axis = 1)
y = middle["LUNG_CANCER"]

# Add a constant to the model (for the intercept)
X = sm.add_constant(x)

# Fit the logistic regression model using statsmodels
logit_model = sm.Logit(y, X)
result = logit_model.fit()

# Print summary of the model
print(f"Middle-age dataset: {result.summary()}")
display(HTML("<p style='text-align: center; font-weight: bold;'>Figure. 22</
 ↪p>"))
# Logistics regression on senior
x = senior.drop(labels =["LUNG_CANCER", "AGE"], axis = 1)
y = senior["LUNG_CANCER"]

# Add a constant to the model (for the intercept)
X = sm.add_constant(x)

# Fit the logistic regression model using statsmodels
logit_model = sm.Logit(y, X)
result = logit_model.fit()

# Print summary of the model
print(f"Senior dataset: {result.summary()}")
display(HTML("<p style='text-align: center; font-weight: bold;'>Figure. 23</
 ↪p>"))
```

```
Optimization terminated successfully.
         Current function value: 0.400779
         Iterations 7
Whole dataset:                            Logit Regression Results
==============================================================================
Dep. Variable:            LUNG_CANCER   No. Observations:                 9000
```

```
Model:                          Logit   Df Residuals:                     8984
Method:                           MLE   Df Model:                           15
Date:                Fri, 11 Oct 2024   Pseudo R-squ.:                  0.1877
Time:                        23:18:30   Log-Likelihood:                -3607.0
converged:                       True   LL-Null:                       -4440.5
Covariance Type:            nonrobust   LLR p-value:                     0.000
================================================================================
=========
                            coef    std err          z      P>|z|      [0.025
0.975]
--------------------------------------------------------------------------------
---------
const                    -0.0343      0.129     -0.266      0.791      -0.287
0.219
GENDER                   -0.8624      0.062    -13.812      0.000      -0.985
-0.740
SMOKING                   0.3867      0.060      6.489      0.000       0.270
0.504
YELLOW_FINGERS            1.3632      0.063     21.652      0.000       1.240
1.487
ANXIETY                   0.0280      0.060      0.463      0.643      -0.090
0.146
PEER_PRESSURE             0.2572      0.060      4.267      0.000       0.139
0.375
CHRONIC DISEASE          -0.1350      0.060     -2.260      0.024      -0.252
-0.018
FATIGUE                   0.5514      0.062      8.858      0.000       0.429
0.673
ALLERGY                   0.7217      0.060     11.958      0.000       0.603
0.840
WHEEZING                 -0.6377      0.061    -10.436      0.000      -0.757
-0.518
ALCOHOL CONSUMING         1.5064      0.065     23.301      0.000       1.380
1.633
COUGHING                 -0.0193      0.061     -0.315      0.753      -0.139
0.101
SHORTNESS OF BREATH      -0.3549      0.067     -5.272      0.000      -0.487
-0.223
SWALLOWING DIFFICULTY     0.7346      0.068     10.833      0.000       0.602
0.868
CHEST PAIN               -0.0556      0.062     -0.900      0.368      -0.177
0.066
cluster                   0.2294      0.060      3.853      0.000       0.113
0.346
================================================================================
=========
```

```
<IPython.core.display.HTML object>
```

```
Optimization terminated successfully.
         Current function value: 0.400636
         Iterations 7
Middle-age dataset:                            Logit Regression Results
==============================================================================
Dep. Variable:            LUNG_CANCER   No. Observations:                 4388
Model:                          Logit   Df Residuals:                     4373
Method:                           MLE   Df Model:                           14
Date:                Fri, 11 Oct 2024   Pseudo R-squ.:                  0.2221
Time:                        23:18:30   Log-Likelihood:                -1758.0
converged:                       True   LL-Null:                       -2259.9
Covariance Type:            nonrobust   LLR p-value:                2.378e-205
================================================================================
=========
                            coef    std err          z      P>|z|      [0.025
0.975]
--------------------------------------------------------------------------------
---------
const                    -0.0155      0.180     -0.086      0.931      -0.368
0.337
GENDER                   -0.9549      0.090    -10.664      0.000      -1.130
-0.779
SMOKING                   0.5410      0.085      6.337      0.000       0.374
0.708
YELLOW_FINGERS            1.4896      0.090     16.501      0.000       1.313
1.667
ANXIETY                   0.0529      0.087      0.608      0.543      -0.118
0.223
PEER_PRESSURE             0.1916      0.086      2.224      0.026       0.023
0.360
CHRONIC DISEASE          -0.2213      0.086     -2.583      0.010      -0.389
-0.053
FATIGUE                   0.5923      0.088      6.745      0.000       0.420
0.764
ALLERGY                   0.7955      0.086      9.210      0.000       0.626
0.965
WHEEZING                 -0.6259      0.087     -7.154      0.000      -0.797
-0.454
ALCOHOL CONSUMING         1.6464      0.093     17.784      0.000       1.465
1.828
COUGHING                 -0.0751      0.088     -0.856      0.392      -0.247
0.097
SHORTNESS OF BREATH      -0.4525      0.096     -4.738      0.000      -0.640
-0.265
SWALLOWING DIFFICULTY     0.8046      0.097      8.292      0.000       0.614
0.995
CHEST PAIN               -0.1411      0.088     -1.601      0.109      -0.314
0.032
```

```
================================================================================
=========

<IPython.core.display.HTML object>

Optimization terminated successfully.
         Current function value: 0.398360
         Iterations 7
Senior dataset:                            Logit Regression Results
==============================================================================
Dep. Variable:            LUNG_CANCER   No. Observations:                 4612
Model:                          Logit   Df Residuals:                     4597
Method:                           MLE   Df Model:                           14
Date:                Fri, 11 Oct 2024   Pseudo R-squ.:                  0.1548
Time:                        23:18:30   Log-Likelihood:                -1837.2
converged:                       True   LL-Null:                       -2173.8
Covariance Type:            nonrobust   LLR p-value:                1.365e-134
================================================================================
=========
                            coef    std err          z      P>|z|      [0.025
0.975]
--------------------------------------------------------------------------------
---------
const                     0.1821      0.180      1.009      0.313      -0.171
0.536
GENDER                   -0.7703      0.088     -8.779      0.000      -0.942
-0.598
SMOKING                   0.2337      0.084      2.787      0.005       0.069
0.398
YELLOW_FINGERS            1.2348      0.088     13.968      0.000       1.062
1.408
ANXIETY                   0.0204      0.085      0.242      0.809      -0.145
0.186
PEER_PRESSURE             0.3270      0.085      3.845      0.000       0.160
0.494
CHRONIC DISEASE          -0.0536      0.084     -0.638      0.523      -0.218
0.111
FATIGUE                   0.5110      0.089      5.732      0.000       0.336
0.686
ALLERGY                   0.6494      0.085      7.646      0.000       0.483
0.816
WHEEZING                 -0.6444      0.086     -7.487      0.000      -0.813
-0.476
ALCOHOL CONSUMING         1.3623      0.091     15.011      0.000       1.184
1.540
COUGHING                  0.0360      0.086      0.420      0.675      -0.132
0.204
SHORTNESS OF BREATH      -0.2674      0.096     -2.800      0.005      -0.455
-0.080
```

```
SWALLOWING DIFFICULTY     0.6696      0.095      7.020      0.000       0.483
0.857
CHEST PAIN                0.0217      0.087      0.249      0.803      -0.149
0.193
================================================================================
=========
```

```
<IPython.core.display.HTML object>
```

Middle Age Proportions

[120]:

```python
#SMOKING
contingency_smoking_middle = pd.crosstab(middle['SMOKING'], middle['LUNG_CANCER'])
contingency_smoking_middle['Proportion_Cancer'] = contingency_smoking_middle[1] / (contingency_smoking_middle[0] + contingency_smoking_middle[1])


proportions_smoking_middle = contingency_smoking_middle['Proportion_Cancer']
non_cancer_smoking_middle = 1 - proportions_smoking_middle


plt.figure(figsize=(8, 4))
sns.set_theme(style="whitegrid")
y_labels_smoking = ['Non-Smoker', 'Smoker']
bar_width = 0.3 #change the width for a better fit figure
#create horizontal bar chart for cancer patients who smoke/do not smoke using labels above and sns pastel color palette
bars_lung_cancer_smoking = plt.barh(y_labels_smoking, proportions_smoking_middle, color=sns.color_palette("pastel", 2), height=bar_width)
#create horizontal bar chart for cancer patients who smoke/ do not smoke using labels above and sns pastel color palette
bars_no_cancer_smoking = plt.barh(y_labels_smoking, non_cancer_smoking_middle, color='lightgrey', left=proportions_smoking_middle, height=bar_width)

plt.title('Middle Age: Proportion of Lung Cancer Cases by Smoking Status', fontsize=16)
plt.xlabel('Proportion', fontsize=14)
plt.xlim(0, 1)
plt.ylim(-0.5, 1.5)

#to annotate the % of cancer patients for smoker and non-smoker group
for index, value in enumerate(proportions_smoking_middle):
    plt.text(value - 0.4, index, f"{value*100:.2f}%", va='center', fontsize=12, color='black')
plt.figtext(0.45, -0.1, "Figure. 24", fontsize=15)
plt.show()
```

```
#YELLOW FINGERS
contingency_yellowfingers_middle = pd.crosstab(middle['YELLOW_FINGERS'],␣
 ↪middle['LUNG_CANCER'])
contingency_yellowfingers_middle['Proportion_Cancer'] =␣
 ↪contingency_yellowfingers_middle[1] / (contingency_yellowfingers_middle[0] +␣
 ↪contingency_yellowfingers_middle[1])

proportions_yellowfingers_middle =␣
 ↪contingency_yellowfingers_middle['Proportion_Cancer']
non_cancer_yellowfingers_middle = 1 - proportions_yellowfingers_middle

plt.figure(figsize=(8, 4))
sns.set_theme(style="whitegrid")
y_labels_yellowfingers = ['No Yellow Fingers', 'Yellow Fingers']
bar_width = 0.3
#create horizontal bar chart for cancer patients with or without yellow fingers␣
 ↪using labels above and sns pastel color palette
bars_lung_cancer_yellowfingers = plt.barh(y_labels_yellowfingers,␣
 ↪proportions_yellowfingers_middle, color=['#FFCC99','#FF9999'],␣
 ↪height=bar_width, label='Lung Cancer')
#create horizontal bar chart for non-cancer patients with or without yellow␣
 ↪fingers using labels above and sns pastel color palette
bars_no_cancer_yellowfingers = plt.barh(y_labels_yellowfingers,␣
 ↪non_cancer_yellowfingers_middle, color='lightgrey',␣
 ↪left=proportions_yellowfingers_middle, height=bar_width, label='No Lung␣
 ↪Cancer')

plt.title('Middle Age: Proportion of Lung Cancer Cases by Presence Of Yellow␣
 ↪Fingers', fontsize=16)
plt.xlabel('Proportion', fontsize=14)
plt.xlim(0, 1)
plt.ylim(-0.5, 1.5)

#annotate % of cancer patients for yellow finger and non-yellow finger group
for index, value in enumerate(proportions_yellowfingers_middle):
    plt.text(value - 0.4, index, f"{value*100:.2f}%", va='center', fontsize=12,␣
 ↪color='black')
plt.figtext(0.45, -0.1, "Figure. 25", fontsize=15)
plt.show()

#ALCOHOL CONSUMING
contingency_alcohol_middle = pd.crosstab(middle['ALCOHOL CONSUMING'],␣
 ↪middle['LUNG_CANCER'])
contingency_alcohol_middle['Proportion_Cancer'] = contingency_alcohol_middle[1]␣
 ↪/ (contingency_alcohol_middle[0] + contingency_alcohol_middle[1])
```

```
proportions_alcohol_middle = contingency_alcohol_middle['Proportion_Cancer']
non_cancer_alcohol_middle = 1 - proportions_alcohol_middle

plt.figure(figsize=(8, 4))
sns.set_theme(style="whitegrid")
y_labels_alcohol = ['Non-Alcohol Consuming', 'Alcohol Consuming']
bar_width = 0.3
#create horizontal bar chart for cancer patients who drink/do not drink using␣
 ↪labels above and sns pastel color palette
bars_lung_cancer_alcohol = plt.barh(y_labels_alcohol,␣
 ↪proportions_alcohol_middle, color=['#58d68d', '#d2b4de'], height=bar_width,␣
 ↪label='Lung Cancer')
#create horizontal bar chart for non-cancer patient who drink/do not drink␣
 ↪using labels above and sns pastel color palette
bars_no_cancer_alcohol = plt.barh(y_labels_alcohol, non_cancer_alcohol_middle,␣
 ↪color='lightgrey', left=proportions_alcohol_middle, height=bar_width,␣
 ↪label='No Lung Cancer')

plt.title('Middle Age: Proportion of Lung Cancer Cases by Drinking Status',␣
 ↪fontsize=16)
plt.xlabel('Proportion', fontsize=14)
plt.xlim(0, 1)
plt.ylim(-0.5, 1.5)

#annotate % of cancer patients for drinkers and non-drinkers group
for index, value in enumerate(proportions_alcohol_middle):
    plt.text(value - 0.4, index, f"{value*100:.2f}%", va='center', fontsize=12,␣
 ↪color='black')
plt.figtext(0.45, -0.1, "Figure. 26", fontsize=15)
plt.show()
```

```
[120]: <Figure size 800x400 with 0 Axes>

[120]: Text(0.5, 1.0, 'Middle Age: Proportion of Lung Cancer Cases by Smoking Status')

[120]: Text(0.5, 0, 'Proportion')

[120]: (0.0, 1.0)

[120]: (-0.5, 1.5)

[120]: Text(0.34680548982489345, 0, '74.68%')

[120]: Text(0.4285714285714286, 1, '82.86%')

[120]: Text(0.45, -0.1, 'Figure. 24')
```

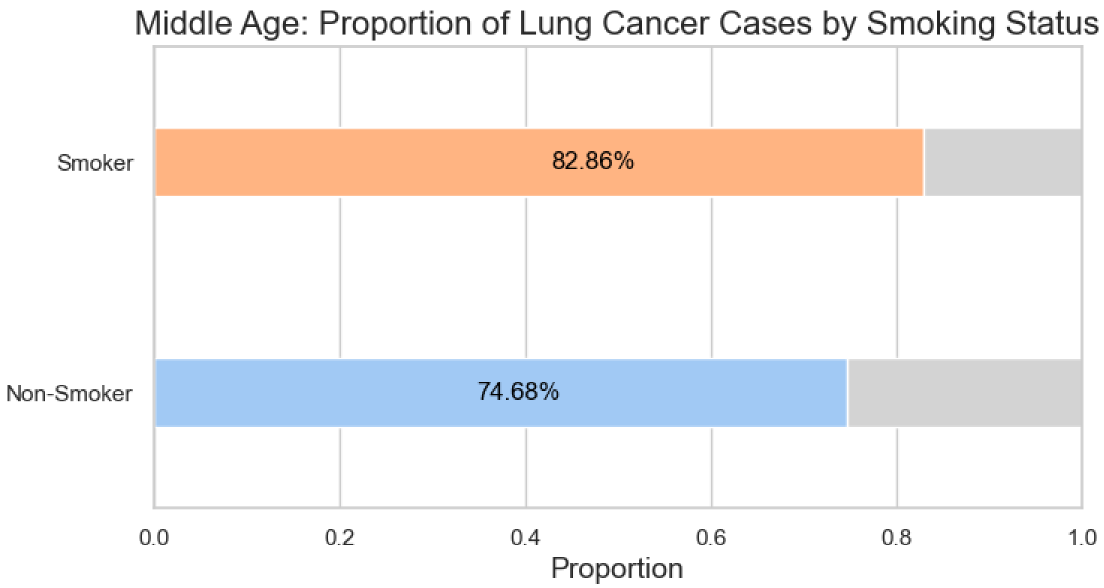


Figure. 24

```
[120]: <Figure size 800x400 with 0 Axes>

[120]: Text(0.5, 1.0, 'Middle Age: Proportion of Lung Cancer Cases by Presence Of
       Yellow Fingers')

[120]: Text(0.5, 0, 'Proportion')

[120]: (0.0, 1.0)

[120]: (-0.5, 1.5)

[120]: Text(0.28283220174587775, 0, '68.28%')

[120]: Text(0.48349097162510746, 1, '88.35%')

[120]: Text(0.45, -0.1, 'Figure. 25')
```

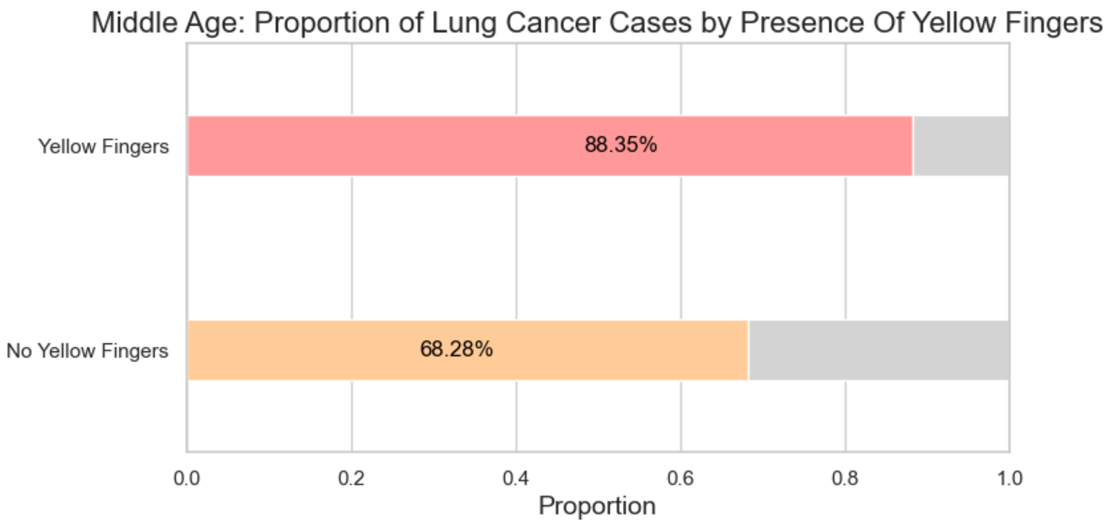


Figure. 25

```
[120]: <Figure size 800x400 with 0 Axes>

[120]: Text(0.5, 1.0, 'Middle Age: Proportion of Lung Cancer Cases by Drinking Status')

[120]: Text(0.5, 0, 'Proportion')

[120]: (0.0, 1.0)

[120]: (-0.5, 1.5)

[120]: Text(0.2888672824501701, 0, '68.89%')

[120]: Text(0.47773487773487766, 1, '87.77%')

[120]: Text(0.45, -0.1, 'Figure. 26')
```

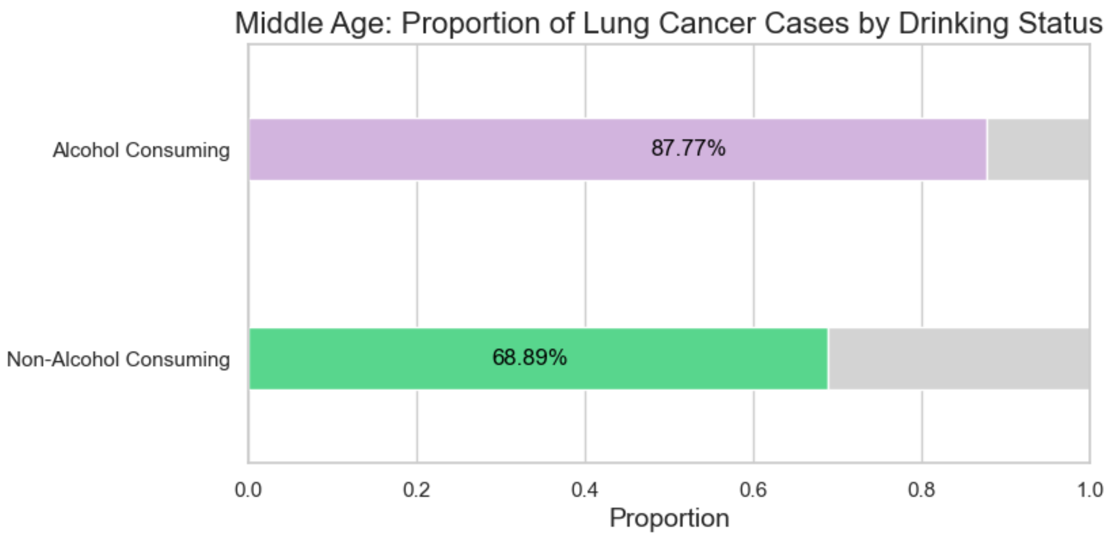


Figure. 26

Senior Proportions

[122]:

```python
#SMOKING
contingency_smoking_senior = pd.crosstab(senior['SMOKING'], senior['LUNG_CANCER'])
contingency_smoking_senior['Proportion_Cancer'] = contingency_smoking_senior[1] / (contingency_smoking_senior[0] + contingency_smoking_senior[1])


proportions_smoking_senior = contingency_smoking_senior['Proportion_Cancer']
non_cancer_smoking_senior = 1 - proportions_smoking_senior


plt.figure(figsize=(8, 4))
sns.set_theme(style="whitegrid")
y_labels_smoking = ['Non-Smoker', 'Smoker']
bar_width = 0.3 #change the width for a better fit figure
#create horizontal bar chart for cancer patients who smoke/do not smoke using labels above and sns pastel color palette
bars_lung_cancer_smoking = plt.barh(y_labels_smoking, proportions_smoking_senior, color=sns.color_palette("pastel", 2), height=bar_width)
#create horizontal bar chart for cancer patients who smoke/ do not smoke using labels above and sns pastel color palette
bars_no_cancer_smoking = plt.barh(y_labels_smoking, non_cancer_smoking_senior, color='lightgrey', left=proportions_smoking_senior, height=bar_width)
```

```
plt.title('Senior: Proportion of Lung Cancer Cases by Smoking Status',␣
 ↪fontsize=16)
plt.xlabel('Proportion', fontsize=14)
plt.xlim(0, 1)
plt.ylim(-0.5, 1.5)

#to annotate the % of cancer patients for smoker and non-smoker group
for index, value in enumerate(proportions_smoking_senior):
    plt.text(value - 0.4, index, f"{value*100:.2f}%", va='center', fontsize=12,␣
 ↪color='black')
plt.figtext(0.45, -0.1, "Figure. 27", fontsize=15)
plt.show()

#YELLOW FINGERS
contingency_yellowfingers_senior = pd.crosstab(senior['YELLOW_FINGERS'],␣
 ↪senior['LUNG_CANCER'])
contingency_yellowfingers_senior['Proportion_Cancer'] =␣
 ↪contingency_yellowfingers_senior[1] / (contingency_yellowfingers_senior[0] +␣
 ↪contingency_yellowfingers_senior[1])

proportions_yellowfingers_senior =␣
 ↪contingency_yellowfingers_senior['Proportion_Cancer']
non_cancer_yellowfingers_senior = 1 - proportions_yellowfingers_senior

plt.figure(figsize=(8, 4))
sns.set_theme(style="whitegrid")
y_labels_yellowfingers = ['No Yellow Fingers', 'Yellow Fingers']
bar_width = 0.3
#create horizontal bar chart for cancer patients with or without yellow fingers␣
 ↪using labels above and sns pastel color palette
bars_lung_cancer_yellowfingers = plt.barh(y_labels_yellowfingers,␣
 ↪proportions_yellowfingers_senior, color=['#FFCC99','#FF9999'],␣
 ↪height=bar_width, label='Lung Cancer')
#create horizontal bar chart for non-cancer patients with or without yellow␣
 ↪fingers using labels above and sns pastel color palette
bars_no_cancer_yellowfingers = plt.barh(y_labels_yellowfingers,␣
 ↪non_cancer_yellowfingers_senior, color='lightgrey',␣
 ↪left=proportions_yellowfingers_senior, height=bar_width, label='No Lung␣
 ↪Cancer')

plt.title('Senior: Proportion of Lung Cancer Cases by Presence Of Yellow␣
 ↪Fingers', fontsize=16)
plt.xlabel('Proportion', fontsize=14)
plt.xlim(0, 1)
plt.ylim(-0.5, 1.5)
```

```
#annotate % of cancer patients for yellow finger and non-yellow finger group
for index, value in enumerate(proportions_yellowfingers_senior):
    plt.text(value - 0.4, index, f"{value*100:.2f}%", va='center', fontsize=12,␣
 ↪color='black')
plt.figtext(0.45, -0.1, "Figure. 28", fontsize=15)
plt.show()

#ALCOHOL CONSUMING
contingency_alcohol_senior = pd.crosstab(senior['ALCOHOL CONSUMING'],␣
 ↪senior['LUNG_CANCER'])
contingency_alcohol_senior['Proportion_Cancer'] = contingency_alcohol_senior[1]␣
 ↪/ (contingency_alcohol_senior[0] + contingency_alcohol_senior[1])

proportions_alcohol_senior = contingency_alcohol_senior['Proportion_Cancer']
non_cancer_alcohol_senior = 1 - proportions_alcohol_senior

plt.figure(figsize=(8, 4))
sns.set_theme(style="whitegrid")
y_labels_alcohol = ['Non-Alcohol Consuming', 'Alcohol Consuming']
bar_width = 0.3
#create horizontal bar chart for cancer patients who drink/do not drink using␣
 ↪labels above and sns pastel color palette
bars_lung_cancer_alcohol = plt.barh(y_labels_alcohol,␣
 ↪proportions_alcohol_senior, color=['#58d68d', '#d2b4de'], height=bar_width,␣
 ↪label='Lung Cancer')
#create horizontal bar chart for non-cancer patient who drink/do not drink␣
 ↪using labels above and sns pastel color palette
bars_no_cancer_alcohol = plt.barh(y_labels_alcohol, non_cancer_alcohol_senior,␣
 ↪color='lightgrey', left=proportions_alcohol_senior, height=bar_width,␣
 ↪label='No Lung Cancer')

plt.title('Senior: Proportion of Lung Cancer Cases by Drinking Status',␣
 ↪fontsize=16)
plt.xlabel('Proportion', fontsize=14)
plt.xlim(0, 1)
plt.ylim(-0.5, 1.5)

#annotate % of cancer patients for drinkers and non-drinkers group
for index, value in enumerate(proportions_alcohol_senior):
    plt.text(value - 0.4, index, f"{value*100:.2f}%", va='center', fontsize=12,␣
 ↪color='black')
plt.figtext(0.45, -0.1, "Figure. 29", fontsize=15)
plt.show()
```

```
[122]: <Figure size 800x400 with 0 Axes>
```

```
[122]: Text(0.5, 1.0, 'Senior: Proportion of Lung Cancer Cases by Smoking Status')

[122]: Text(0.5, 0, 'Proportion')

[122]: (0.0, 1.0)

[122]: (-0.5, 1.5)

[122]: Text(0.4098271155595996, 0, '80.98%')

[122]: Text(0.42932891466445733, 1, '82.93%')

[122]: Text(0.45, -0.1, 'Figure. 27')
```

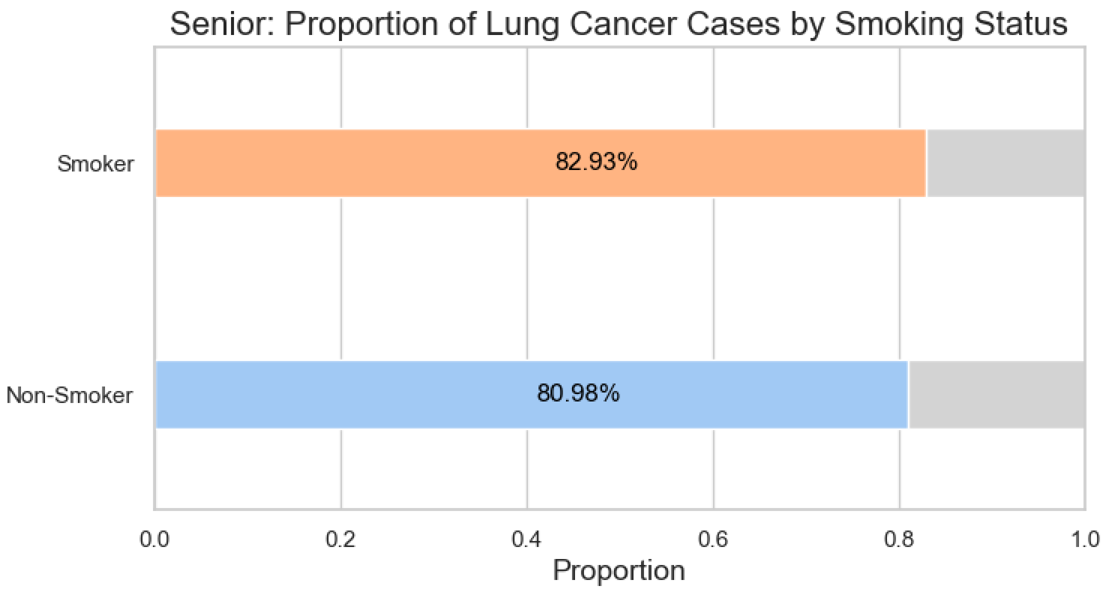


Figure. 27

```
[122]: <Figure size 800x400 with 0 Axes>

[122]: Text(0.5, 1.0, 'Senior: Proportion of Lung Cancer Cases by Presence Of Yellow
       Fingers')

[122]: Text(0.5, 0, 'Proportion')

[122]: (0.0, 1.0)

[122]: (-0.5, 1.5)

[122]: Text(0.3429094236047575, 0, '74.29%')
```

```
[122]: Text(0.4895300906842539, 1, '88.95%')
```

```
[122]: Text(0.45, -0.1, 'Figure. 28')
```

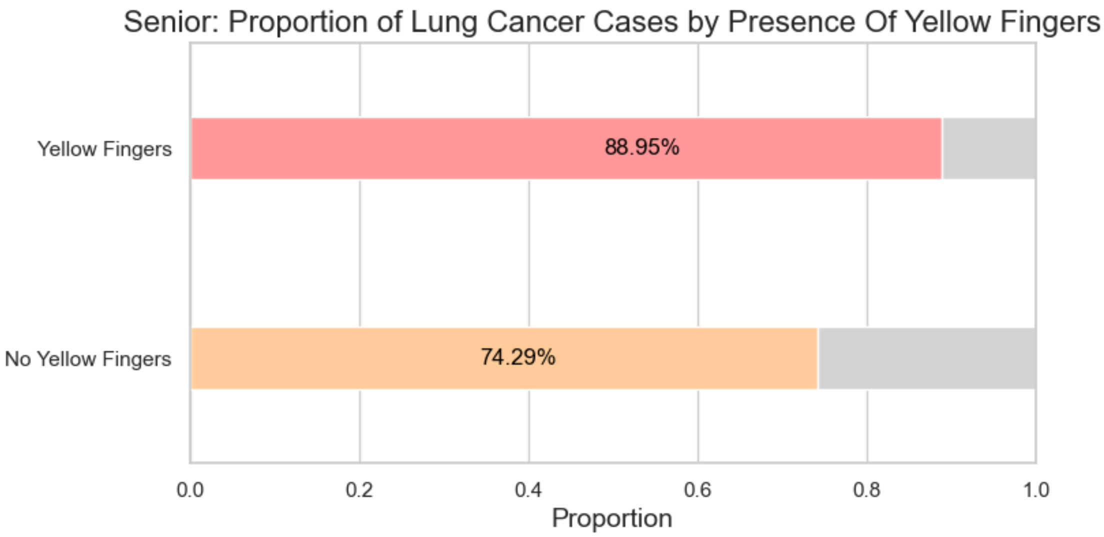


Figure. 28

```
[122]: <Figure size 800x400 with 0 Axes>
```

```
[122]: Text(0.5, 1.0, 'Senior: Proportion of Lung Cancer Cases by Drinking Status')
```

```
[122]: Text(0.5, 0, 'Proportion')
```

```
[122]: (0.0, 1.0)
```

```
[122]: (-0.5, 1.5)
```

```
[122]: Text(0.3459386281588448, 0, '74.59%')
```

```
[122]: Text(0.4885642737896494, 1, '88.86%')
```

```
[122]: Text(0.45, -0.1, 'Figure. 29')
```

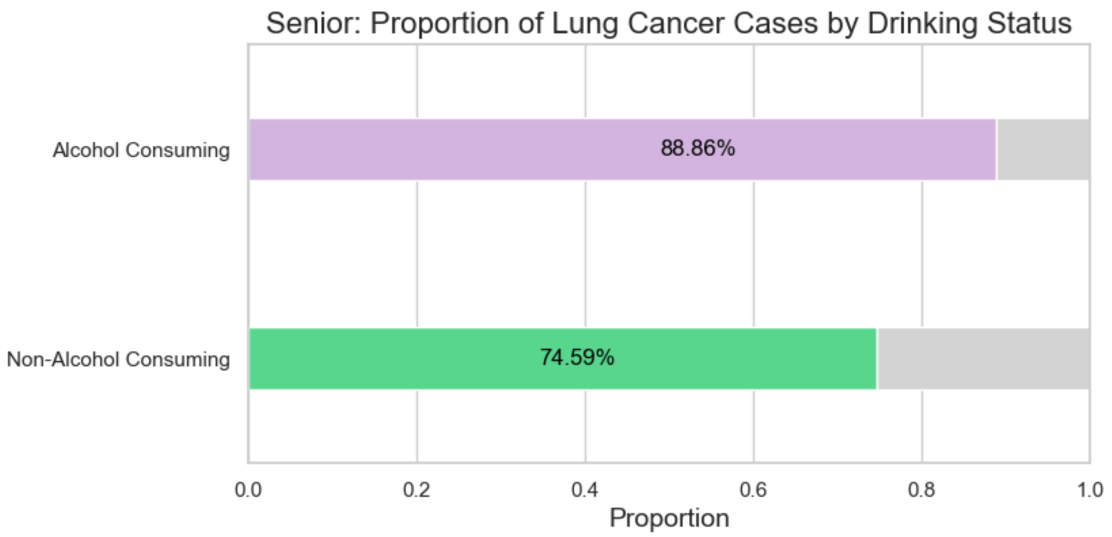

Senior: Proportion of Lung Cancer Cases by Drinking Status

Figure. 29

9 By Age and Gender

[124]:

```
# SUBGROUP-GENDER
import matplotlib.pyplot as plt
# Split gender dataframes
female = final_df[final_df['GENDER']==0]
male = final_df[final_df['GENDER']==1]
female = female.drop("GENDER", axis = 1)
male = male.drop("GENDER", axis = 1)

# Split age group datqaframes
middle_age_female = female[female['cluster']==0].drop("cluster", axis=1)
middle_age_male = male[male['cluster']==0].drop("cluster", axis=1)

senior_female = female[female['cluster']==1].drop("cluster", axis=1)
senior_male = male[male['cluster']==1].drop("cluster", axis=1)
# Calculate correlations between Lung Cancer and other variables for␣
 ↪middle-aged and senior groups
middle_age_female_corr = middle_age_female.corr()['LUNG_CANCER'].
 ↪drop('LUNG_CANCER')
middle_age_male_corr = middle_age_male.corr()['LUNG_CANCER'].drop('LUNG_CANCER')
senior_female_corr = senior_female.corr()['LUNG_CANCER'].drop('LUNG_CANCER')
senior_male_corr = senior_male.corr()['LUNG_CANCER'].drop('LUNG_CANCER')
# Create a DataFrame to store correlations
correlation_middle = pd.DataFrame({
    'Middle-age Female': middle_age_female_corr,
    'Middle-age Male': middle_age_male_corr
})

# Plotting bar chart for each symptom
correlation_middle.plot(kind='bar', figsize=(12, 8))

# Add labels and title
plt.xlabel('Symptoms')
plt.ylabel('Correlation with Lung Cancer')
plt.title('Comparison of Correlation with Lung Cancer')
plt.xticks(rotation=45, ha='right')
plt.grid(True, axis = 'y')
plt.figtext(0.45, -0.1, "Figure. 30", fontsize=15)
# Show the plot
plt.tight_layout()
plt.show()

correlation_senior = pd.DataFrame({
    'Senior Female': senior_female_corr,
    'Senior Male': senior_male_corr
})

# Plotting bar chart for each symptom
correlation_senior.plot(kind='bar', figsize=(12, 8))

# Add labels and title
plt.xlabel('Symptoms')
plt.ylabel('Correlation with Lung Cancer')
plt.title('Comparison of Correlation with Lung Cancer')
plt.xticks(rotation=45, ha='right')
plt.grid(True, axis = 'y')
plt.figtext(0.45, -0.1, "Figure. 31", fontsize=15)
# Show the plot
plt.tight_layout()
plt.show()
```

```
[124]: <Axes: >

[124]: Text(0.5, 0, 'Symptoms')

[124]: Text(0, 0.5, 'Correlation with Lung Cancer')

[124]: Text(0.5, 1.0, 'Comparison of Correlation with Lung Cancer')

[124]: (array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13]),
        [Text(0, 0, 'AGE'),
         Text(1, 0, 'SMOKING'),
         Text(2, 0, 'YELLOW_FINGERS'),
         Text(3, 0, 'ANXIETY'),
         Text(4, 0, 'PEER_PRESSURE'),
         Text(5, 0, 'CHRONIC DISEASE'),
         Text(6, 0, 'FATIGUE '),
         Text(7, 0, 'ALLERGY '),
         Text(8, 0, 'WHEEZING'),
         Text(9, 0, 'ALCOHOL CONSUMING'),
         Text(10, 0, 'COUGHING'),
         Text(11, 0, 'SHORTNESS OF BREATH'),
         Text(12, 0, 'SWALLOWING DIFFICULTY'),
         Text(13, 0, 'CHEST PAIN')])
```

```
[124]: Text(0.45, -0.1, 'Figure. 30')
```

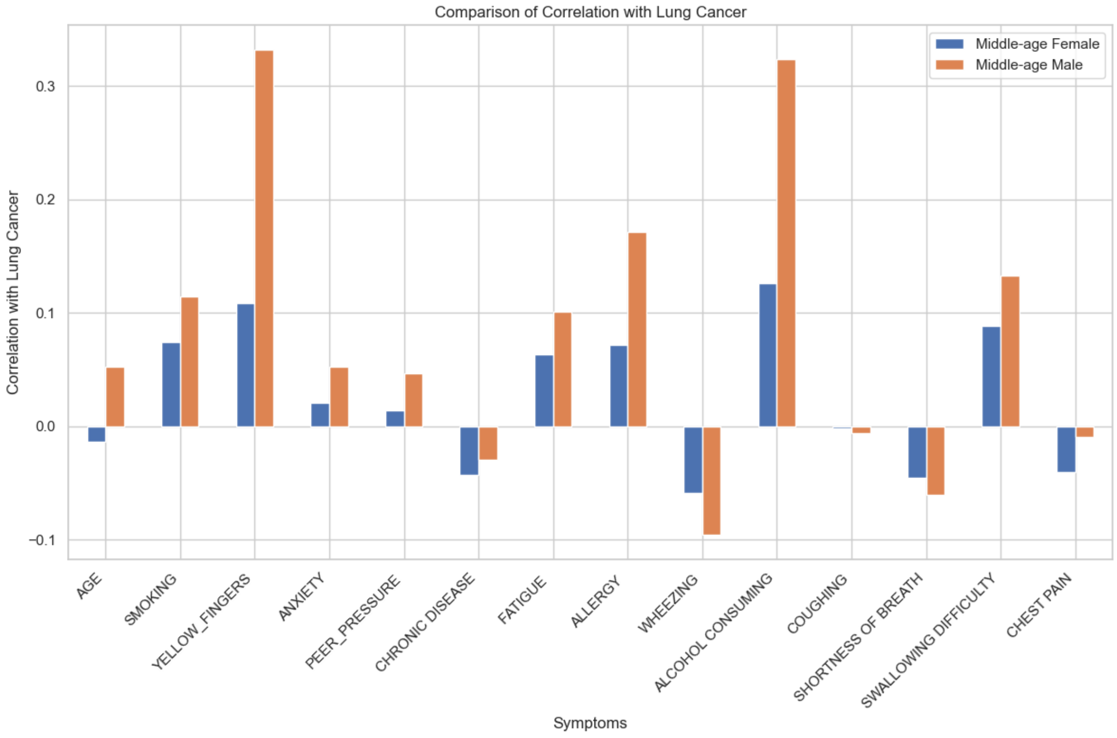


Figure. 30

```
[124]: <Axes: >
```

```
[124]: Text(0.5, 0, 'Symptoms')
```

```
[124]: Text(0, 0.5, 'Correlation with Lung Cancer')
```

```
[124]: Text(0.5, 1.0, 'Comparison of Correlation with Lung Cancer')
```

```
[124]: (array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13]),
        [Text(0, 0, 'AGE'),
         Text(1, 0, 'SMOKING'),
         Text(2, 0, 'YELLOW_FINGERS'),
         Text(3, 0, 'ANXIETY'),
         Text(4, 0, 'PEER_PRESSURE'),
         Text(5, 0, 'CHRONIC DISEASE'),
         Text(6, 0, 'FATIGUE '),
         Text(7, 0, 'ALLERGY '),
         Text(8, 0, 'WHEEZING'),
         Text(9, 0, 'ALCOHOL CONSUMING'),
         Text(10, 0, 'COUGHING'),
         Text(11, 0, 'SHORTNESS OF BREATH'),
         Text(12, 0, 'SWALLOWING DIFFICULTY'),
         Text(13, 0, 'CHEST PAIN')])
```

```
[124]: Text(0.45, -0.1, 'Figure. 31')
```

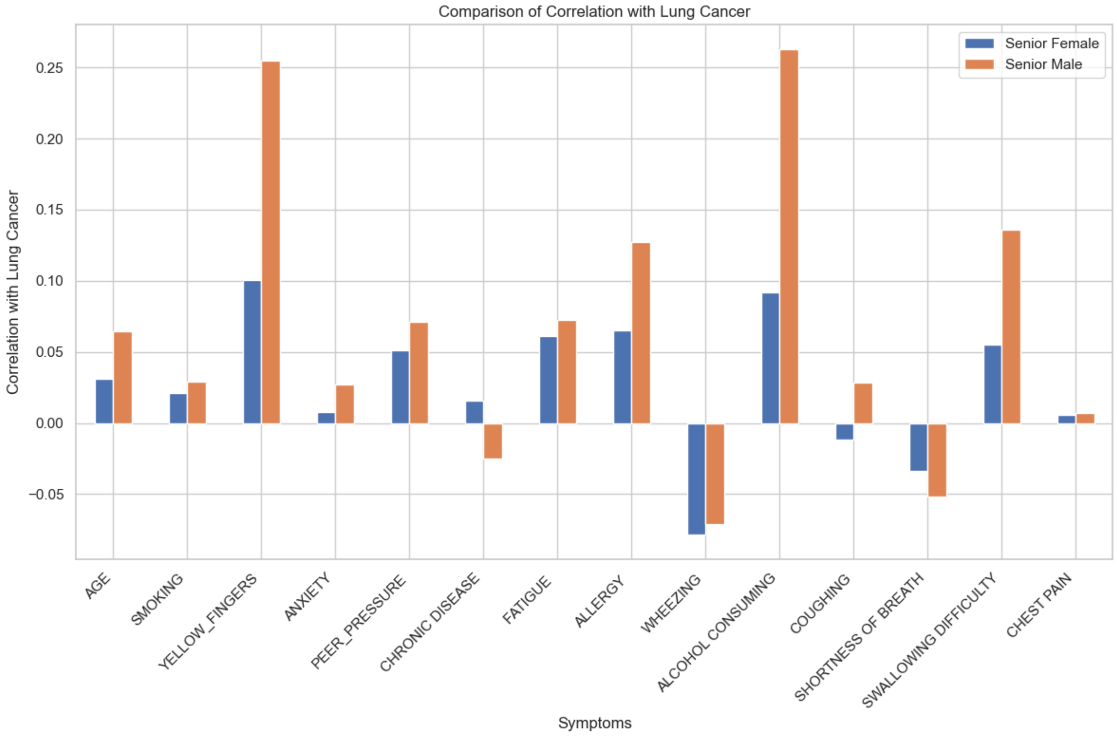


Figure. 31

```
[134]: #SMOKING

       list = [middle_age_male]
```

```
list_string = ['middle_age_male','middle_age_female','senior_male',␣
 ↪'senior_female']

for i in range(0,1):
    cross_tab = pd.crosstab(list[i]['SMOKING'], list[i]['LUNG_CANCER'])
    cross_tab['Proportion_Cancer'] = cross_tab[1] / (cross_tab[0] +␣
 ↪cross_tab[1])
    proportions = cross_tab['Proportion_Cancer']
    non_cancer = 1 - proportions
    plt.figure(figsize=(8, 4))
    sns.set_theme(style="whitegrid")
    y_labels = ['Non-Smoker', 'Smoker']
    bar_width = 0.3 #change the width for a better fit figure
    #create horizontal bar chart for cancer patients who smoke/do not smoke␣
 ↪using labels above and sns pastel color palette
    bars_cancer = plt.barh(y_labels, proportions, color=sns.
 ↪color_palette("pastel", 2), height=bar_width)
    #create horizontal bar chart for cancer patients who smoke/ do not smoke␣
 ↪using labels above and sns pastel color palette
    bars_no_cancer = plt.barh(y_labels, non_cancer, color='lightgrey',␣
 ↪left=proportions, height=bar_width)

    plt.title(f'{list_string[i]}: Proportion of Lung Cancer Cases by Smoking␣
 ↪Status', fontsize=16)
    plt.xlabel('Proportion', fontsize=14)
    plt.xlim(0, 1)
    plt.ylim(-0.5, 1.5)

    #to annotate the % of cancer patients for smoker and non-smoker group
    for index, value in enumerate(proportions_smoking_men):
        plt.text(value - 0.4, index, f"{value*100:.2f}%", va='center',␣
 ↪fontsize=12, color='black')
    plt.figtext(0.45, -0.1, f"Figure. {32+i}", fontsize=15)
    plt.show()

    cross_tab = pd.crosstab(list[i]['YELLOW_FINGERS'], list[i]['LUNG_CANCER'])
    cross_tab['Proportion_Cancer'] = cross_tab[1] / (cross_tab[0] +␣
 ↪cross_tab[1])
    proportions = cross_tab['Proportion_Cancer']
    non_cancer = 1 - proportions
    plt.figure(figsize=(8, 4))
    sns.set_theme(style="whitegrid")
    y_labels = ['No Yellow Fingers', 'Yellow Fingers']
    bar_width = 0.3 #change the width for a better fit figure
    #create horizontal bar chart for cancer patients who smoke/do not smoke␣
 ↪using labels above and sns pastel color palette
```

```
    bars_cancer = plt.barh(y_labels, proportions, color=sns.
↪color_palette("pastel", 2), height=bar_width)
    #create horizontal bar chart for cancer patients who smoke/ do not smoke␣
↪using labels above and sns pastel color palette
    bars_no_cancer = plt.barh(y_labels, non_cancer, color='lightgrey',␣
↪left=proportions, height=bar_width)

    plt.title(f'{list_string[i]}: Proportion of Lung Cancer Cases by Presence␣
↪Of Yellow Fingers', fontsize=16)
    plt.xlabel('Proportion', fontsize=14)
    plt.xlim(0, 1)
    plt.ylim(-0.5, 1.5)

    #to annotate the % of cancer patients for smoker and non-smoker group
    for index, value in enumerate(proportions):
        plt.text(value - 0.4, index, f"{value*100:.2f}%", va='center',␣
↪fontsize=12, color='black')
    plt.figtext(0.45, -0.1, f"Figure. {33+i}", fontsize=15)
    plt.show()

    cross_tab = pd.crosstab(list[i]['ALCOHOL CONSUMING'],␣
↪list[i]['LUNG_CANCER'])
    cross_tab['Proportion_Cancer'] = cross_tab[1] / (cross_tab[0] +␣
↪cross_tab[1])
    proportions = cross_tab['Proportion_Cancer']
    non_cancer = 1 - proportions
    plt.figure(figsize=(8, 4))
    sns.set_theme(style="whitegrid")
    y_labels = ['Non-Alcohol Consuming', 'Alcohol Consuming']
    bar_width = 0.3 #change the width for a better fit figure
    #create horizontal bar chart for cancer patients who smoke/do not smoke␣
↪using labels above and sns pastel color palette
    bars_cancer = plt.barh(y_labels, proportions, color=sns.
↪color_palette("pastel", 2), height=bar_width)
    #create horizontal bar chart for cancer patients who smoke/ do not smoke␣
↪using labels above and sns pastel color palette
    bars_no_cancer = plt.barh(y_labels, non_cancer, color='lightgrey',␣
↪left=proportions, height=bar_width)

    plt.title(f'{list_string[i]}: Proportion of Lung Cancer Cases by Drinking␣
↪Status', fontsize=16)
    plt.xlabel('Proportion', fontsize=14)
    plt.xlim(0, 1)
    plt.ylim(-0.5, 1.5)

    #to annotate the % of cancer patients for smoker and non-smoker group
    for index, value in enumerate(proportions):
        plt.text(value - 0.4, index, f"{value*100:.2f}%", va='center',␣
↪fontsize=12, color='black')
    plt.figtext(0.45, -0.1, f"Figure. {34+i}", fontsize=15)
    plt.show()
```

```
[134]: <Figure size 800x400 with 0 Axes>
```

```
[134]: Text(0.5, 1.0, 'middle_age_male: Proportion of Lung Cancer Cases by Smoking
       Status')
```

```
[134]: Text(0.5, 0, 'Proportion')
```

```
[134]: (0.0, 1.0)
```

```
[134]: (-0.5, 1.5)
```

```
[134]: Text(0.32058194266153184, 0, '72.06%')
```

```
[134]: Text(0.38353413654618473, 1, '78.35%')
```

```
[134]: Text(0.45, -0.1, 'Figure. 32')
```

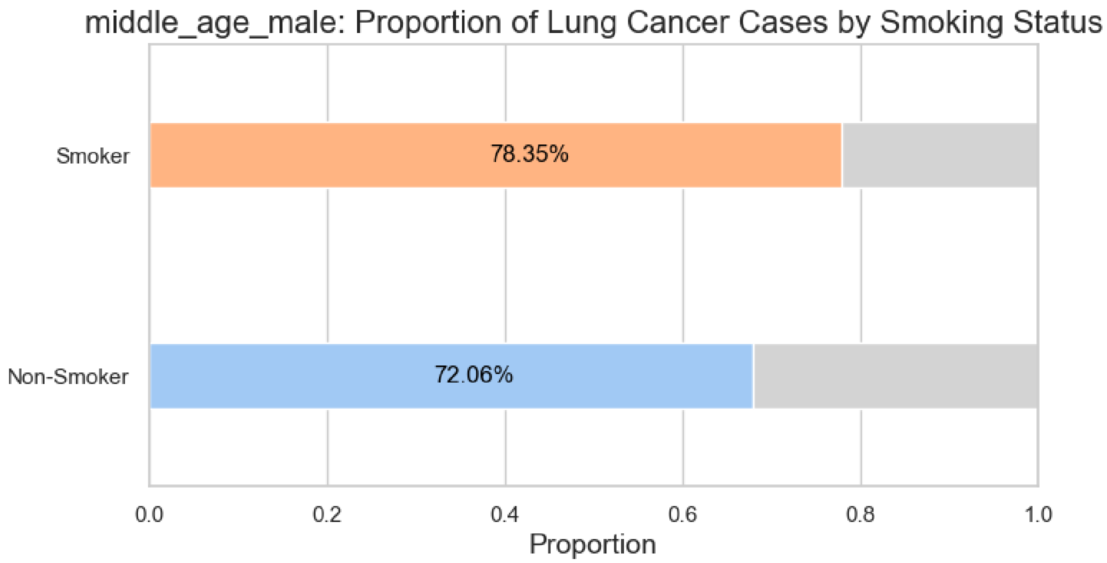


Figure. 32

```
[134]: <Figure size 800x400 with 0 Axes>
```

```
[134]: Text(0.5, 1.0, 'middle_age_male: Proportion of Lung Cancer Cases by Presence Of
       Yellow Fingers')
```

```
[134]: Text(0.5, 0, 'Proportion')

[134]: (0.0, 1.0)

[134]: (-0.5, 1.5)

[134]: Text(0.1802575107296137, 0, '58.03%')

[134]: Text(0.4752079866888519, 1, '87.52%')

[134]: Text(0.45, -0.1, 'Figure. 33')
```

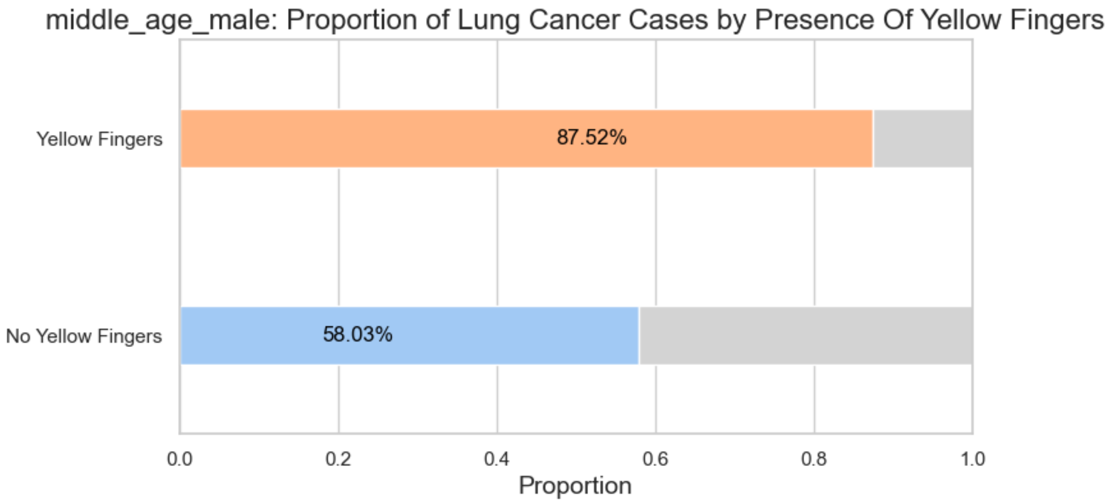


Figure. 33

```
[134]: <Figure size 800x400 with 0 Axes>

[134]: Text(0.5, 1.0, 'middle_age_male: Proportion of Lung Cancer Cases by Drinking
       Status')

[134]: Text(0.5, 0, 'Proportion')

[134]: (0.0, 1.0)

[134]: (-0.5, 1.5)

[134]: Text(0.17007575757575755, 0, '57.01%')

[134]: Text(0.45888634630053393, 1, '85.89%')

[134]: Text(0.45, -0.1, 'Figure. 34')
```

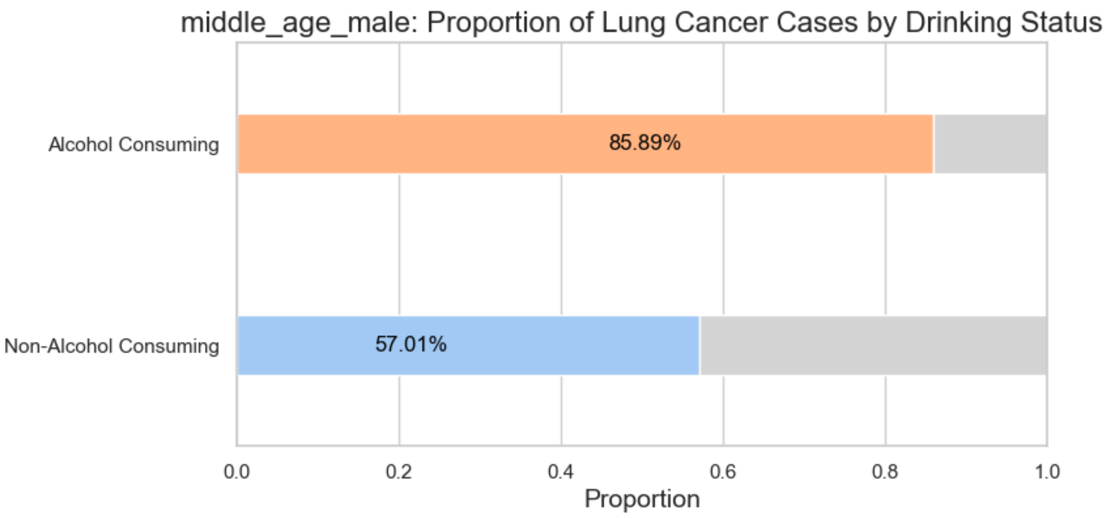


Figure. 34

[136]:

```
#SMOKING
list = [middle_age_female]
list_string = ['middle_age_female']

for i in range(0,1):
    cross_tab = pd.crosstab(list[i]['SMOKING'], list[i]['LUNG_CANCER'])
    cross_tab['Proportion_Cancer'] = cross_tab[1] / (cross_tab[0] +␣
 ↪cross_tab[1])
    proportions = cross_tab['Proportion_Cancer']
    non_cancer = 1 - proportions
    plt.figure(figsize=(8, 4))
    sns.set_theme(style="whitegrid")
    y_labels = ['Non-Smoker', 'Smoker']
    bar_width = 0.3 #change the width for a better fit figure
    #create horizontal bar chart for cancer patients who smoke/do not smoke␣
 ↪using labels above and sns pastel color palette
    bars_cancer = plt.barh(y_labels, proportions, color=sns.
 ↪color_palette("pastel", 2), height=bar_width)
    #create horizontal bar chart for cancer patients who smoke/ do not smoke␣
 ↪using labels above and sns pastel color palette
    bars_no_cancer = plt.barh(y_labels, non_cancer, color='lightgrey',␣
 ↪left=proportions, height=bar_width)

    plt.title(f'{list_string[i]}: Proportion of Lung Cancer Cases by Smoking␣
 ↪Status', fontsize=16)
    plt.xlabel('Proportion', fontsize=14)
    plt.xlim(0, 1)
    plt.ylim(-0.5, 1.5)

    #to annotate the % of cancer patients for smoker and non-smoker group
    for index, value in enumerate(proportions_smoking_men):
        plt.text(value - 0.4, index, f"{value*100:.2f}%", va='center',␣
↪fontsize=12, color='black')
    plt.figtext(0.45, -0.1, f"Figure. {35+i}", fontsize=15)
    plt.show()

    cross_tab = pd.crosstab(list[i]['YELLOW_FINGERS'], list[i]['LUNG_CANCER'])
    cross_tab['Proportion_Cancer'] = cross_tab[1] / (cross_tab[0] +␣
↪cross_tab[1])
    proportions = cross_tab['Proportion_Cancer']
    non_cancer = 1 - proportions
    plt.figure(figsize=(8, 4))
    sns.set_theme(style="whitegrid")
    y_labels = ['No Yellow Fingers', 'Yellow Fingers']
    bar_width = 0.3 #change the width for a better fit figure
    #create horizontal bar chart for cancer patients who smoke/do not smoke␣
↪using labels above and sns pastel color palette
    bars_cancer = plt.barh(y_labels, proportions, color=sns.
↪color_palette("pastel", 2), height=bar_width)
    #create horizontal bar chart for cancer patients who smoke/ do not smoke␣
↪using labels above and sns pastel color palette
    bars_no_cancer = plt.barh(y_labels, non_cancer, color='lightgrey',␣
↪left=proportions, height=bar_width)

    plt.title(f'{list_string[i]}: Proportion of Lung Cancer Cases by Presence␣
↪Of Yellow Fingers', fontsize=16)
    plt.xlabel('Proportion', fontsize=14)
    plt.xlim(0, 1)
    plt.ylim(-0.5, 1.5)

    #to annotate the % of cancer patients for smoker and non-smoker group
    for index, value in enumerate(proportions):
        plt.text(value - 0.4, index, f"{value*100:.2f}%", va='center',␣
↪fontsize=12, color='black')
    plt.figtext(0.45, -0.1, f"Figure. {36+i}", fontsize=15)
    plt.show()

    cross_tab = pd.crosstab(list[i]['ALCOHOL CONSUMING'],␣
↪list[i]['LUNG_CANCER'])
    cross_tab['Proportion_Cancer'] = cross_tab[1] / (cross_tab[0] +␣
↪cross_tab[1])
    proportions = cross_tab['Proportion_Cancer']
    non_cancer = 1 - proportions
```

```python
    plt.figure(figsize=(8, 4))
    sns.set_theme(style="whitegrid")
    y_labels = ['Non-Alcohol Consuming', 'Alcohol Consuming']
    bar_width = 0.3 #change the width for a better fit figure
    #create horizontal bar chart for cancer patients who smoke/do not smoke using labels above and sns pastel color palette
    bars_cancer = plt.barh(y_labels, proportions, color=sns.color_palette("pastel", 2), height=bar_width)
    #create horizontal bar chart for cancer patients who smoke/ do not smoke using labels above and sns pastel color palette
    bars_no_cancer = plt.barh(y_labels, non_cancer, color='lightgrey', left=proportions, height=bar_width)

    plt.title(f'{list_string[i]}: Proportion of Lung Cancer Cases by Drinking Status', fontsize=16)
    plt.xlabel('Proportion', fontsize=14)
    plt.xlim(0, 1)
    plt.ylim(-0.5, 1.5)

    #to annotate the % of cancer patients for smoker and non-smoker group
    for index, value in enumerate(proportions):
        plt.text(value - 0.4, index, f"{value*100:.2f}%", va='center', fontsize=12, color='black')
    plt.figtext(0.45, -0.1, f"Figure. {37+i}", fontsize=15)
    plt.show()
```

```
[136]: <Figure size 800x400 with 0 Axes>

[136]: Text(0.5, 1.0, 'middle_age_female: Proportion of Lung Cancer Cases by Smoking
       Status')

[136]: Text(0.5, 0, 'Proportion')

[136]: (0.0, 1.0)

[136]: (-0.5, 1.5)

[136]: Text(0.32058194266153184, 0, '72.06%')

[136]: Text(0.38353413654618473, 1, '78.35%')

[136]: Text(0.45, -0.1, 'Figure. 35')
```

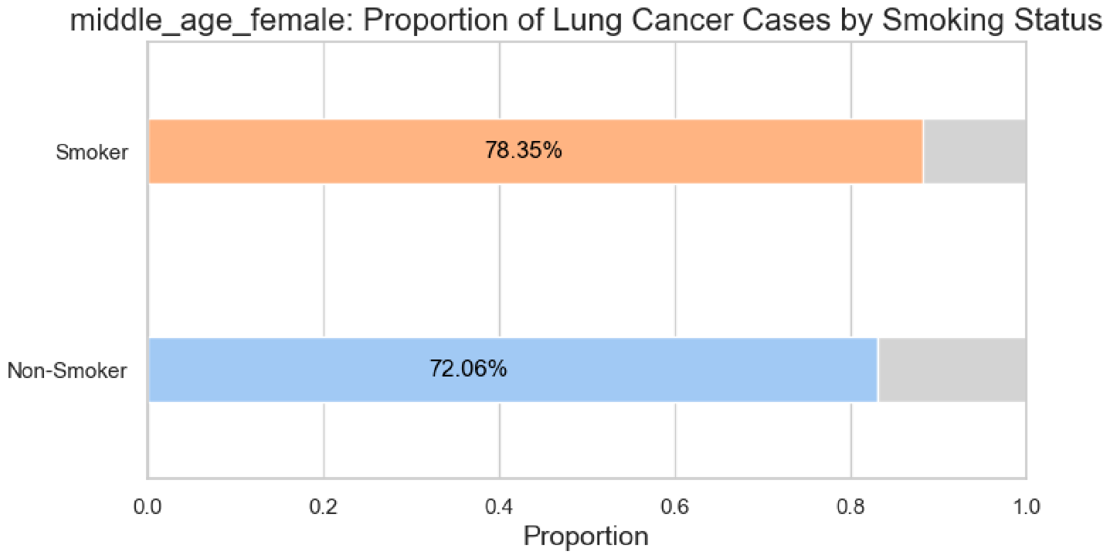


Figure. 35

```
[136]: <Figure size 800x400 with 0 Axes>

[136]: Text(0.5, 1.0, 'middle_age_female: Proportion of Lung Cancer Cases by Presence
       Of Yellow Fingers')

[136]: Text(0.5, 0, 'Proportion')

[136]: (0.0, 1.0)

[136]: (-0.5, 1.5)

[136]: Text(0.4160535117056856, 0, '81.61%')

[136]: Text(0.49234875444839854, 1, '89.23%')

[136]: Text(0.45, -0.1, 'Figure. 36')
```

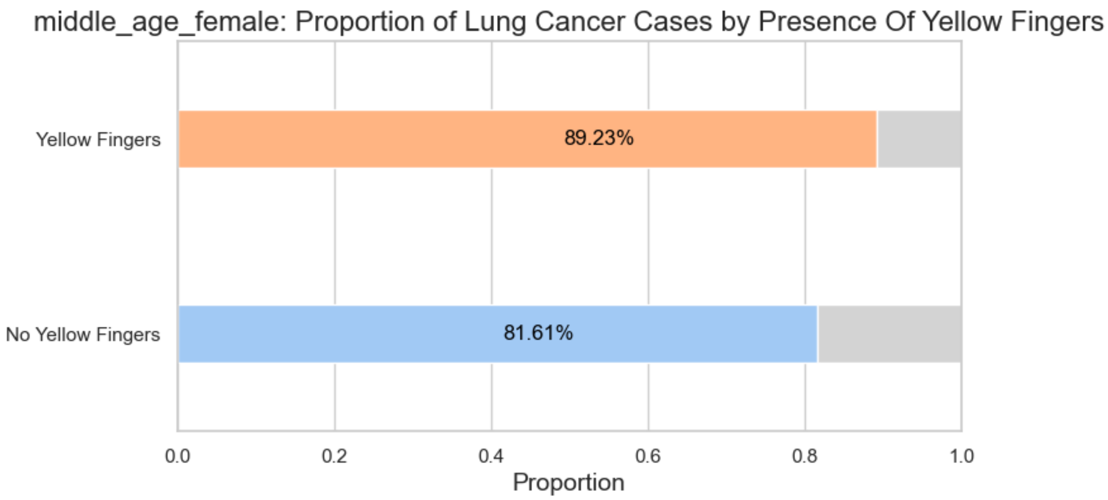


Figure. 36

```
[136]: <Figure size 800x400 with 0 Axes>

[136]: Text(0.5, 1.0, 'middle_age_female: Proportion of Lung Cancer Cases by Drinking
       Status')

[136]: Text(0.5, 0, 'Proportion')

[136]: (0.0, 1.0)

[136]: (-0.5, 1.5)

[136]: Text(0.41418581418581 41, 0, '81.42%')

[136]: Text(0.5019607843137255, 1, '90.20%')

[136]: Text(0.45, -0.1, 'Figure. 37')
```

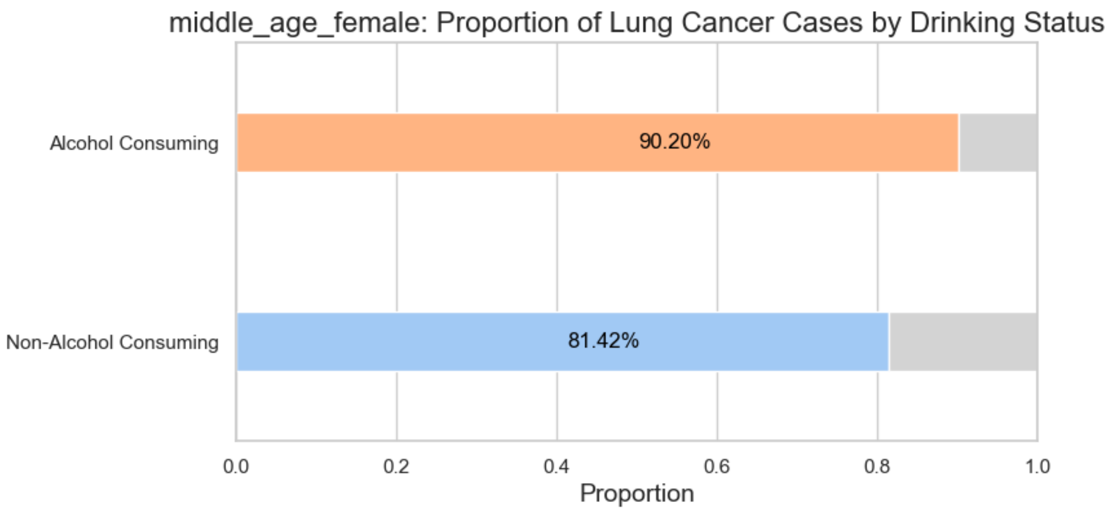


Figure. 37

```
[138]: #SMOKING

list = [senior_male]
list_string = ['senior_male']

for i in range(0,1):
    cross_tab = pd.crosstab(list[i]['SMOKING'], list[i]['LUNG_CANCER'])
    cross_tab['Proportion_Cancer'] = cross_tab[1] / (cross_tab[0] +␣
  ↪cross_tab[1])
    proportions = cross_tab['Proportion_Cancer']
    non_cancer = 1 - proportions
    plt.figure(figsize=(8, 4))
    sns.set_theme(style="whitegrid")
    y_labels = ['Non-Smoker', 'Smoker']
    bar_width = 0.3 #change the width for a better fit figure
    #create horizontal bar chart for cancer patients who smoke/do not smoke␣
  ↪using labels above and sns pastel color palette
    bars_cancer = plt.barh(y_labels, proportions, color=sns.
  ↪color_palette("pastel", 2), height=bar_width)
    #create horizontal bar chart for cancer patients who smoke/ do not smoke␣
  ↪using labels above and sns pastel color palette
    bars_no_cancer = plt.barh(y_labels, non_cancer, color='lightgrey',␣
  ↪left=proportions, height=bar_width)

    plt.title(f'{list_string[i]}: Proportion of Lung Cancer Cases by Smoking␣
  ↪Status', fontsize=16)
    plt.xlabel('Proportion', fontsize=14)
    plt.xlim(0, 1)
    plt.ylim(-0.5, 1.5)

    #to annotate the % of cancer patients for smoker and non-smoker group
    for index, value in enumerate(proportions_smoking_men):
        plt.text(value - 0.4, index, f"{value*100:.2f}%", va='center',␣
↪fontsize=12, color='black')
    plt.figtext(0.45, -0.1, f"Figure. {38+i}", fontsize=15)
    plt.show()

    cross_tab = pd.crosstab(list[i]['YELLOW_FINGERS'], list[i]['LUNG_CANCER'])
    cross_tab['Proportion_Cancer'] = cross_tab[1] / (cross_tab[0] +␣
↪cross_tab[1])
    proportions = cross_tab['Proportion_Cancer']
    non_cancer = 1 - proportions
    plt.figure(figsize=(8, 4))
    sns.set_theme(style="whitegrid")
    y_labels = ['No Yellow Fingers', 'Yellow Fingers']
    bar_width = 0.3 #change the width for a better fit figure
    #create horizontal bar chart for cancer patients who smoke/do not smoke␣
↪using labels above and sns pastel color palette
    bars_cancer = plt.barh(y_labels, proportions, color=sns.
↪color_palette("pastel", 2), height=bar_width)
    #create horizontal bar chart for cancer patients who smoke/ do not smoke␣
↪using labels above and sns pastel color palette
    bars_no_cancer = plt.barh(y_labels, non_cancer, color='lightgrey',␣
↪left=proportions, height=bar_width)

    plt.title(f'{list_string[i]}: Proportion of Lung Cancer Cases by Presence␣
↪Of Yellow Fingers', fontsize=16)
    plt.xlabel('Proportion', fontsize=14)
    plt.xlim(0, 1)
    plt.ylim(-0.5, 1.5)

    #to annotate the % of cancer patients for smoker and non-smoker group
    for index, value in enumerate(proportions):
        plt.text(value - 0.4, index, f"{value*100:.2f}%", va='center',␣
↪fontsize=12, color='black')
    plt.figtext(0.45, -0.1, f"Figure. {39+i}", fontsize=15)
    plt.show()

    cross_tab = pd.crosstab(list[i]['ALCOHOL CONSUMING'],␣
↪list[i]['LUNG_CANCER'])
    cross_tab['Proportion_Cancer'] = cross_tab[1] / (cross_tab[0] +␣
↪cross_tab[1])
    proportions = cross_tab['Proportion_Cancer']
    non_cancer = 1 - proportions
    plt.figure(figsize=(8, 4))
    sns.set_theme(style="whitegrid")
    y_labels = ['Non-Alcohol Consuming', 'Alcohol Consuming']
    bar_width = 0.3 #change the width for a better fit figure
    #create horizontal bar chart for cancer patients who smoke/do not smoke␣
↪using labels above and sns pastel color palette
    bars_cancer = plt.barh(y_labels, proportions, color=sns.
↪color_palette("pastel", 2), height=bar_width)
    #create horizontal bar chart for cancer patients who smoke/ do not smoke␣
↪using labels above and sns pastel color palette
    bars_no_cancer = plt.barh(y_labels, non_cancer, color='lightgrey',␣
↪left=proportions, height=bar_width)

    plt.title(f'{list_string[i]}: Proportion of Lung Cancer Cases by Drinking␣
↪Status', fontsize=16)
    plt.xlabel('Proportion', fontsize=14)
    plt.xlim(0, 1)
    plt.ylim(-0.5, 1.5)

    #to annotate the % of cancer patients for smoker and non-smoker group
    for index, value in enumerate(proportions):
        plt.text(value - 0.4, index, f"{value*100:.2f}%", va='center',␣
↪fontsize=12, color='black')
    plt.figtext(0.45, -0.1, f"Figure. {40+i}", fontsize=15)
    plt.show()
```

```
[138]: <Figure size 800x400 with 0 Axes>

[138]: Text(0.5, 1.0, 'senior_male: Proportion of Lung Cancer Cases by Smoking Status')

[138]: Text(0.5, 0, 'Proportion')

[138]: (0.0, 1.0)

[138]: (-0.5, 1.5)

[138]: Text(0.32058194266153184, 0, '72.06%')

[138]: Text(0.38353413654618473, 1, '78.35%')

[138]: Text(0.45, -0.1, 'Figure. 38')
```

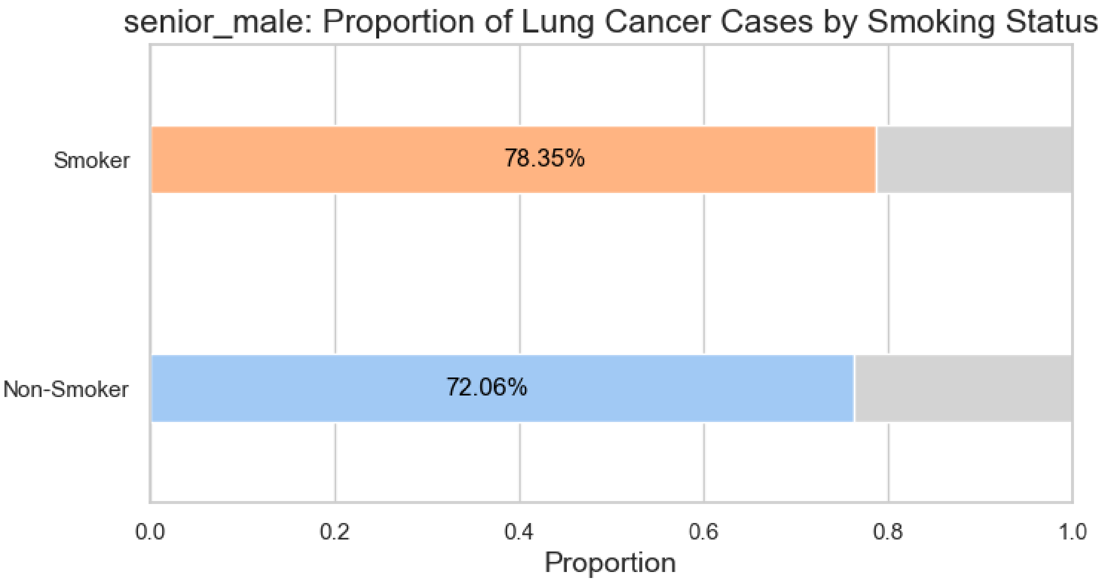


Figure. 38

```
[138]: <Figure size 800x400 with 0 Axes>

[138]: Text(0.5, 1.0, 'senior_male: Proportion of Lung Cancer Cases by Presence Of
       Yellow Fingers')

[138]: Text(0.5, 0, 'Proportion')

[138]: (0.0, 1.0)

[138]: (-0.5, 1.5)

[138]: Text(0.2649789029535865, 0, '66.50%')

[138]: Text(0.4776470588235294, 1, '87.76%')

[138]: Text(0.45, -0.1, 'Figure. 39')
```

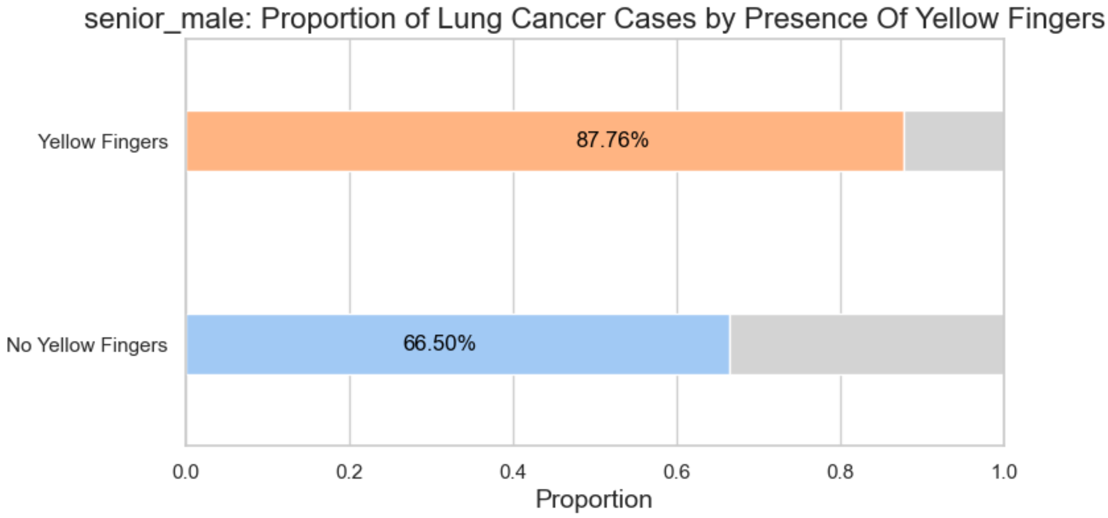


Figure. 39

```
[138]: <Figure size 800x400 with 0 Axes>

[138]: Text(0.5, 1.0, 'senior_male: Proportion of Lung Cancer Cases by Drinking
       Status')

[138]: Text(0.5, 0, 'Proportion')

[138]: (0.0, 1.0)

[138]: (-0.5, 1.5)

[138]: Text(0.25764192139737985, 0, '65.76%')

[138]: Text(0.47756653992395437, 1, '87.76%')

[138]: Text(0.45, -0.1, 'Figure. 40')
```

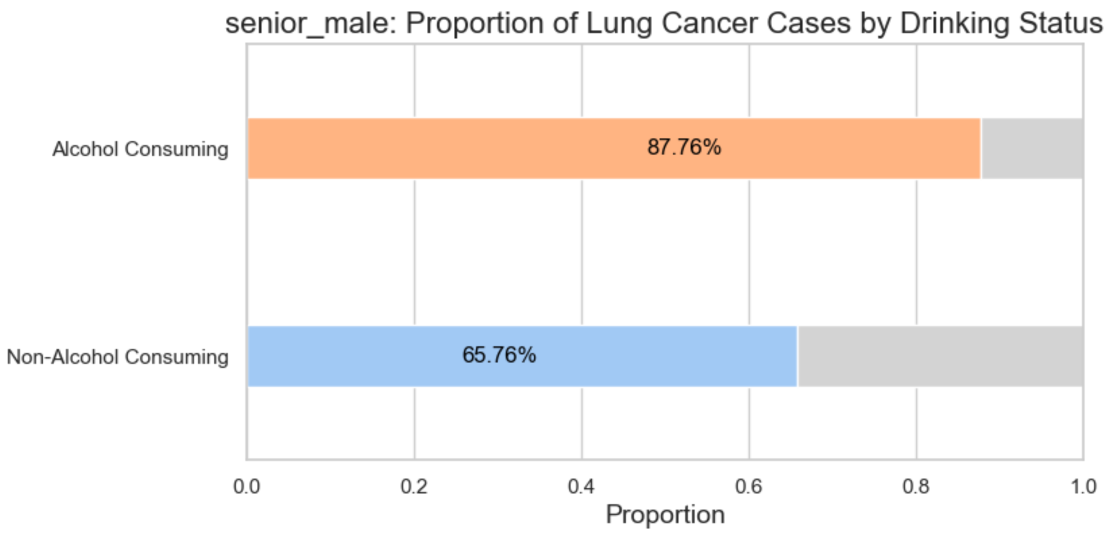


Figure. 40

[140]:
```
#SMOKING

list = [senior_female]
list_string = ['senior_female']

for i in range(0,1):
    cross_tab = pd.crosstab(list[i]['SMOKING'], list[i]['LUNG_CANCER'])
    cross_tab['Proportion_Cancer'] = cross_tab[1] / (cross_tab[0] +␣
 ↪cross_tab[1])
    proportions = cross_tab['Proportion_Cancer']
    non_cancer = 1 - proportions
    plt.figure(figsize=(8, 4))
    sns.set_theme(style="whitegrid")
    y_labels = ['Non-Smoker', 'Smoker']
    bar_width = 0.3 #change the width for a better fit figure
    #create horizontal bar chart for cancer patients who smoke/do not smoke␣
 ↪using labels above and sns pastel color palette
    bars_cancer = plt.barh(y_labels, proportions, color=sns.
 ↪color_palette("pastel", 2), height=bar_width)
    #create horizontal bar chart for cancer patients who smoke/ do not smoke␣
 ↪using labels above and sns pastel color palette
    bars_no_cancer = plt.barh(y_labels, non_cancer, color='lightgrey',␣
 ↪left=proportions, height=bar_width)

    plt.title(f'{list_string[i]}: Proportion of Lung Cancer Cases by Smoking␣
 ↪Status', fontsize=16)
    plt.xlabel('Proportion', fontsize=14)
    plt.xlim(0, 1)
    plt.ylim(-0.5, 1.5)

    #to annotate the % of cancer patients for smoker and non-smoker group
    for index, value in enumerate(proportions_smoking_men):
        plt.text(value - 0.4, index, f"{value*100:.2f}%", va='center',␣
↪fontsize=12, color='black')
    plt.figtext(0.45, -0.1, f"Figure. {41+i}", fontsize=15)
    plt.show()

    cross_tab = pd.crosstab(list[i]['YELLOW_FINGERS'], list[i]['LUNG_CANCER'])
    cross_tab['Proportion_Cancer'] = cross_tab[1] / (cross_tab[0] +␣
↪cross_tab[1])
    proportions = cross_tab['Proportion_Cancer']
    non_cancer = 1 - proportions
    plt.figure(figsize=(8, 4))
    sns.set_theme(style="whitegrid")
    y_labels = ['No Yellow Fingers', 'Yellow Fingers']
    bar_width = 0.3 #change the width for a better fit figure
    #create horizontal bar chart for cancer patients who smoke/do not smoke␣
↪using labels above and sns pastel color palette
    bars_cancer = plt.barh(y_labels, proportions, color=sns.
↪color_palette("pastel", 2), height=bar_width)
    #create horizontal bar chart for cancer patients who smoke/ do not smoke␣
↪using labels above and sns pastel color palette
    bars_no_cancer = plt.barh(y_labels, non_cancer, color='lightgrey',␣
↪left=proportions, height=bar_width)

    plt.title(f'{list_string[i]}: Proportion of Lung Cancer Cases by Presence␣
↪Of Yellow Fingers', fontsize=16)
    plt.xlabel('Proportion', fontsize=14)
    plt.xlim(0, 1)
    plt.ylim(-0.5, 1.5)

    #to annotate the % of cancer patients for smoker and non-smoker group
    for index, value in enumerate(proportions):
        plt.text(value - 0.4, index, f"{value*100:.2f}%", va='center',␣
↪fontsize=12, color='black')
    plt.figtext(0.45, -0.1, f"Figure. {42+i}", fontsize=15)
    plt.show()

    cross_tab = pd.crosstab(list[i]['ALCOHOL CONSUMING'],␣
↪list[i]['LUNG_CANCER'])
    cross_tab['Proportion_Cancer'] = cross_tab[1] / (cross_tab[0] +␣
↪cross_tab[1])
    proportions = cross_tab['Proportion_Cancer']
    non_cancer = 1 - proportions
    plt.figure(figsize=(8, 4))
    sns.set_theme(style="whitegrid")
    y_labels = ['Non-Alcohol Consuming', 'Alcohol Consuming']
    bar_width = 0.3 #change the width for a better fit figure
    #create horizontal bar chart for cancer patients who smoke/do not smoke␣
↪using labels above and sns pastel color palette
    bars_cancer = plt.barh(y_labels, proportions, color=sns.
↪color_palette("pastel", 2), height=bar_width)
    #create horizontal bar chart for cancer patients who smoke/ do not smoke␣
↪using labels above and sns pastel color palette
    bars_no_cancer = plt.barh(y_labels, non_cancer, color='lightgrey',␣
↪left=proportions, height=bar_width)

    plt.title(f'{list_string[i]}: Proportion of Lung Cancer Cases by Drinking␣
↪Status', fontsize=16)
    plt.xlabel('Proportion', fontsize=14)
    plt.xlim(0, 1)
    plt.ylim(-0.5, 1.5)

    #to annotate the % of cancer patients for smoker and non-smoker group
    for index, value in enumerate(proportions):
        plt.text(value - 0.4, index, f"{value*100:.2f}%", va='center',␣
↪fontsize=12, color='black')
    plt.figtext(0.45, -0.1, f"Figure. {43+i}", fontsize=15)
    plt.show()
```

```
[140]: <Figure size 800x400 with 0 Axes>

[140]: Text(0.5, 1.0, 'senior_female: Proportion of Lung Cancer Cases by Smoking
       Status')

[140]: Text(0.5, 0, 'Proportion')

[140]: (0.0, 1.0)

[140]: (-0.5, 1.5)

[140]: Text(0.32058194266153184, 0, '72.06%')

[140]: Text(0.38353413654618473, 1, '78.35%')

[140]: Text(0.45, -0.1, 'Figure. 41')
```

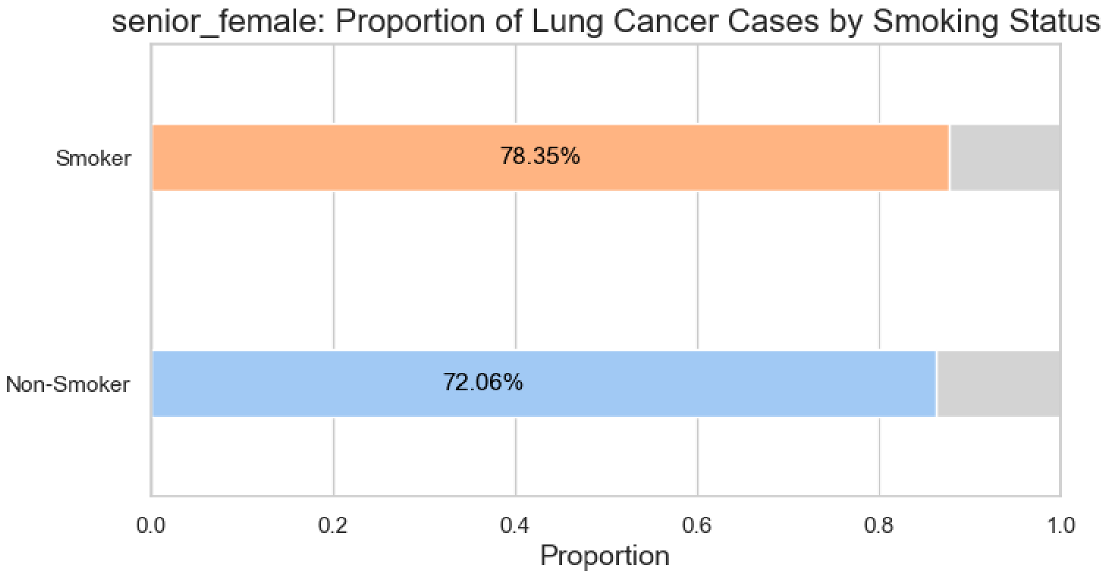


Figure. 41

```
[140]: <Figure size 800x400 with 0 Axes>

[140]: Text(0.5, 1.0, 'senior_female: Proportion of Lung Cancer Cases by Presence Of
       Yellow Fingers')

[140]: Text(0.5, 0, 'Proportion')

[140]: (0.0, 1.0)

[140]: (-0.5, 1.5)

[140]: Text(0.4351648351648352, 0, '83.52%')

[140]: Text(0.5026933101650738, 1, '90.27%')

[140]: Text(0.45, -0.1, 'Figure. 42')
```

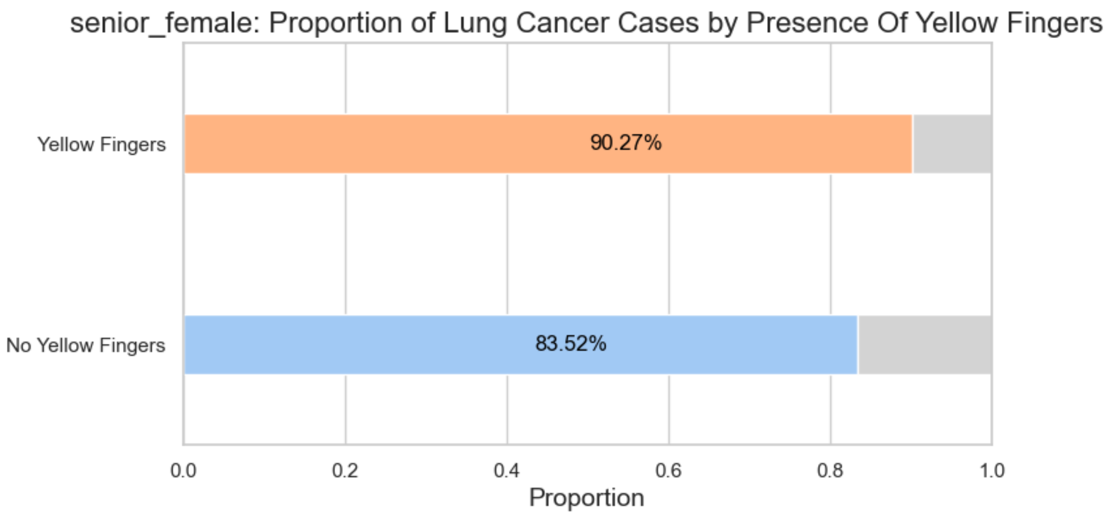


Figure. 42

```
[140]: <Figure size 800x400 with 0 Axes>

[140]: Text(0.5, 1.0, 'senior_female: Proportion of Lung Cancer Cases by Drinking
       Status')

[140]: Text(0.5, 0, 'Proportion')

[140]: (0.0, 1.0)

[140]: (-0.5, 1.5)

[140]: Text(0.4403361344537815, 0, '84.03%')

[140]: Text(0.5019426456984274, 1, '90.19%')

[140]: Text(0.45, -0.1, 'Figure. 43')
```

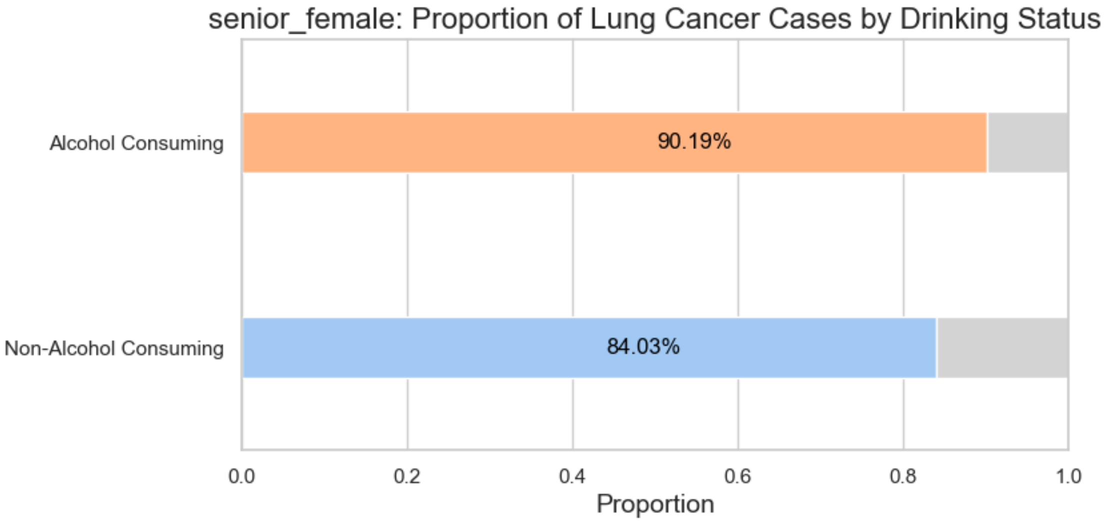

senior_female: Proportion of Lung Cancer Cases by Drinking Status
Alcohol Consuming
90.19%
Non-Alcohol Consuming
84.03%
0.0
0.2
0.4
0.6
0.8
1.0
Proportion

Figure. 43