HM2

2025-02-14

Homework N2

Part 3: Use the datasets provided to create graphs

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.4.2

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
df_lung <- read.csv("C:\\Users\\Hovgr\\OneDrive\\Desktop\\DataViz\\HM2\\lung_cancer_prediction_dataset.
head(df_lung, 5)
```

```
##   ID      Country Population_Size Age Gender Smoker Years_of_Smoking
## 1  0        China            1400  80   Male    Yes               30
## 2  1         Iran              84  53   Male     No                0
## 3  2       Mexico             128  47   Male    Yes               12
## 4  3    Indonesia             273  39 Female     No                0
## 5  4 South Africa              59  44 Female     No                0
##   Cigarettes_per_Day Passive_Smoker Family_History Lung_Cancer_Diagnosis
## 1                 29             No             No                    No
## 2                  0            Yes             No                    No
## 3                  6            Yes             No                    No
## 4                  0             No            Yes                    No
## 5                  0            Yes             No                    No
##   Cancer_Stage Survival_Years Adenocarcinoma_Type Air_Pollution_Exposure
## 1         None              0                 Yes                    Low
## 2         None              0                 Yes                    Low
## 3         None              0                 Yes                 Medium
## 4         None              0                 Yes                    Low
## 5         None              0                 Yes                 Medium
##   Occupational_Exposure Indoor_Pollution Healthcare_Access Early_Detection
## 1                   Yes               No              Poor              No
## 2                   Yes               No              Poor              No
## 3                    No               No              Poor             Yes
## 4                    No               No              Poor              No
## 5                   Yes               No              Poor              No
##   Treatment_Type Developed_or_Developing Annual_Lung_Cancer_Deaths
## 1           None              Developing                      690000
## 2           None              Developing                       27000
## 3           None              Developing                       28000
## 4           None              Developing                       40000
## 5           None              Developing                       15000
##   Lung_Cancer_Prevalence_Rate Mortality_Rate
## 1                        2.44              0
## 2                        2.10              0
## 3                        1.11              0
## 4                        0.75              0
## 5                        2.44              0
```

```
df_air <- read.csv("C:\\Users\\Hovgr\\OneDrive\\Desktop\\DataViz\\HM2\\global_air_pollution_dataset.csv
head(df_air, 5)
```

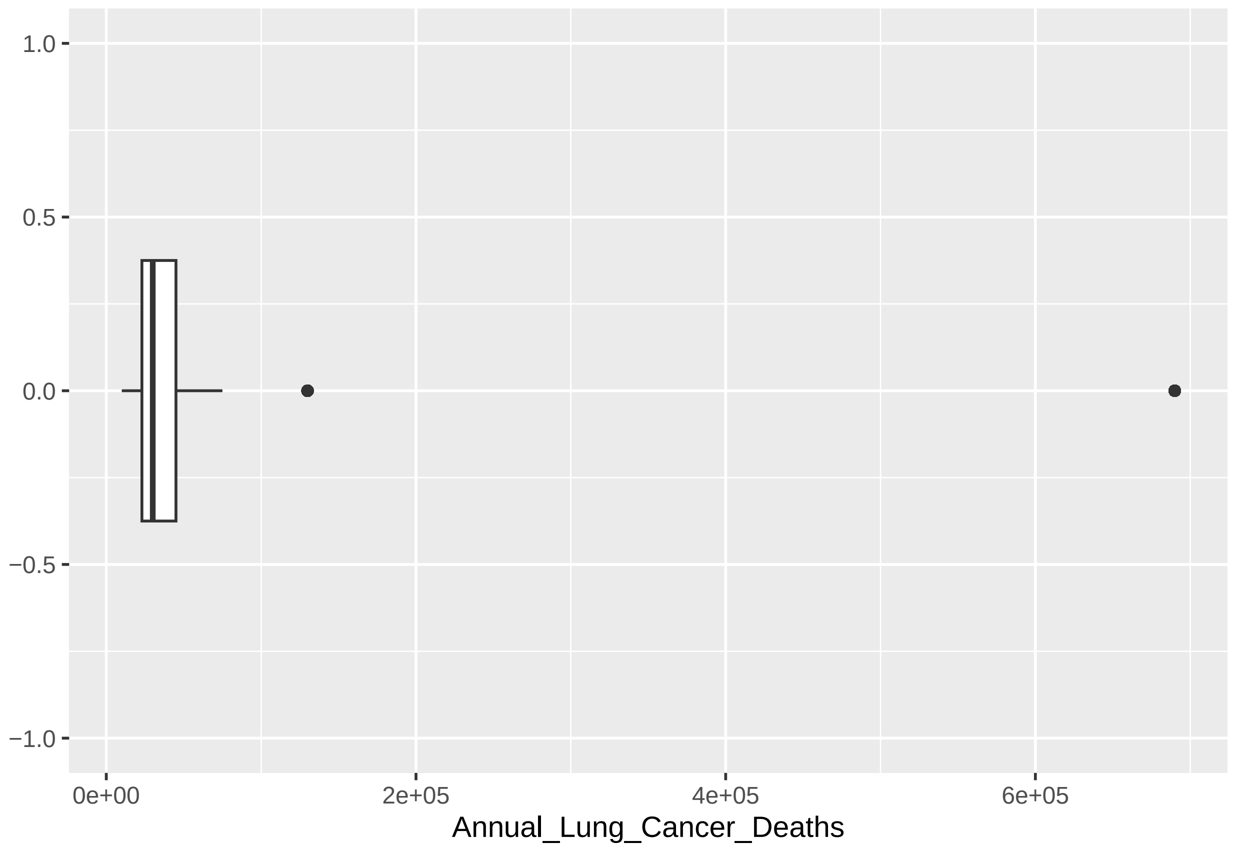
```
##              Country             City AQI_Value AQI_Category CO_AQI_Value
## 1 Russian Federation       Praskoveya        51     Moderate            1
## 2             Brazil Presidente Dutra        41         Good            1
## 3              Italy  Priolo Gargallo        66     Moderate            1
## 4             Poland        Przasnysz        34         Good            1
## 5             France         Punaauia        22         Good            0
##   CO_AQI_Category Ozone_AQI_Value Ozone_AQI_Category NO2_AQI_Value
## 1            Good              36               Good             0
## 2            Good               5               Good             1
## 3            Good              39               Good             2
## 4            Good              34               Good             0
## 5            Good              22               Good             0
##   NO2_AQI_Category PM2.5_AQI_Value PM2.5_AQI_Category
## 1             Good              51           Moderate
## 2             Good              41               Good
## 3             Good              66           Moderate
## 4             Good              20               Good
## 5             Good               6               Good
```

1. Create a Boxplot of Lung Cancer Deaths Distribution.

```
library(ggplot2)
library(ggthemes)
```

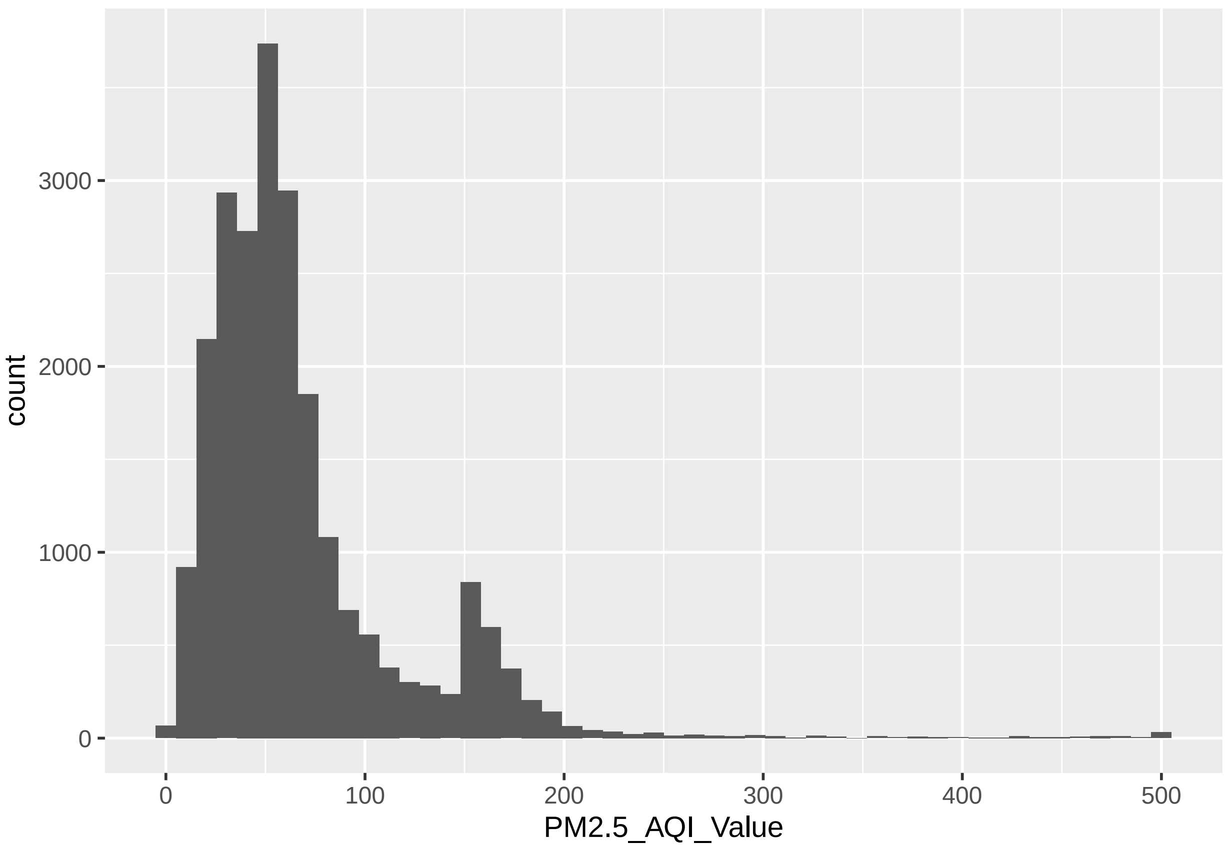
```
## Warning: package 'ggthemes' was built under R version 4.4.2
```

```
ggplot(df_lung, aes(x = Annual_Lung_Cancer_Deaths)) + geom_boxplot() + ylim(-1,1)
```



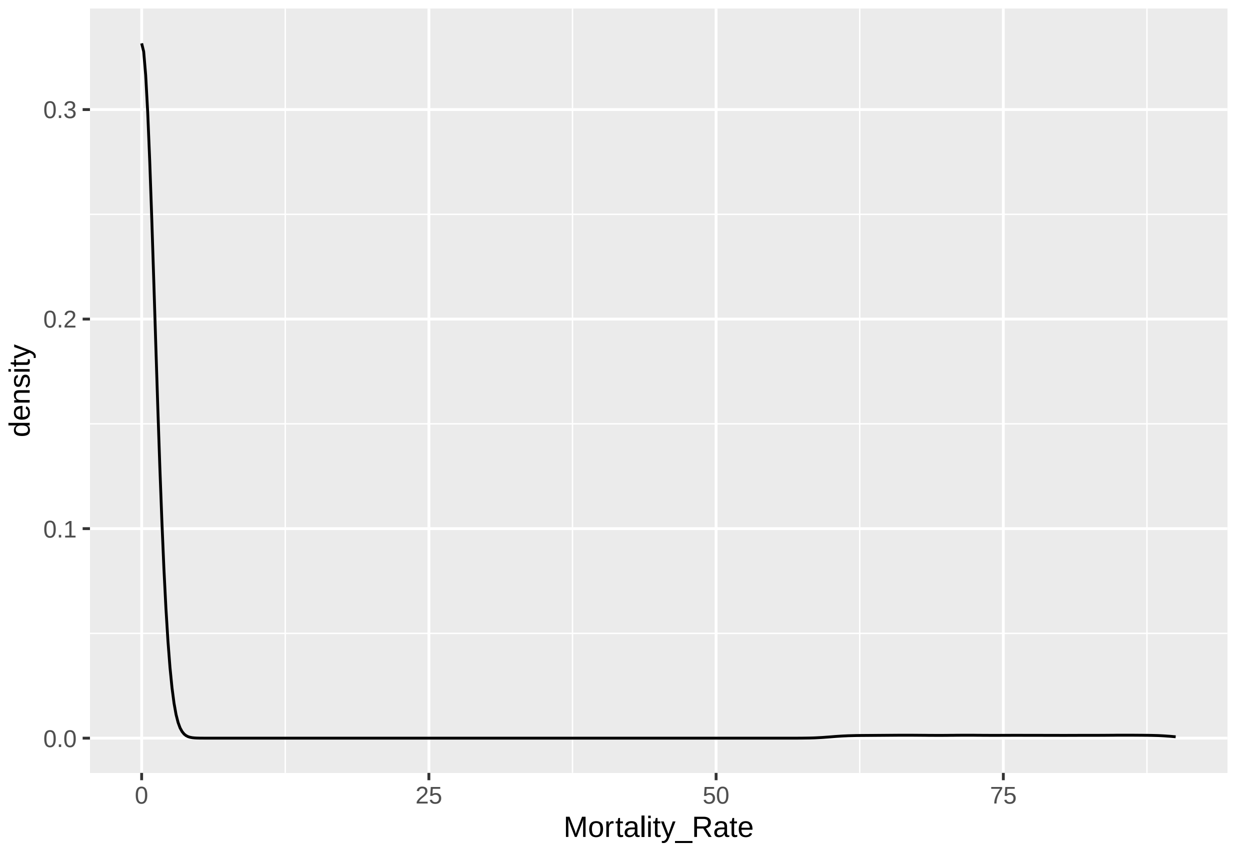
2. Create a Histogram of PM2.5 AQI Values.

```
ggplot(df_air, aes(x = PM2.5_AQI_Value)) + geom_histogram(bins=50)
```



3. Create a Density Plot of the Lung Cancer Mortality Rate.

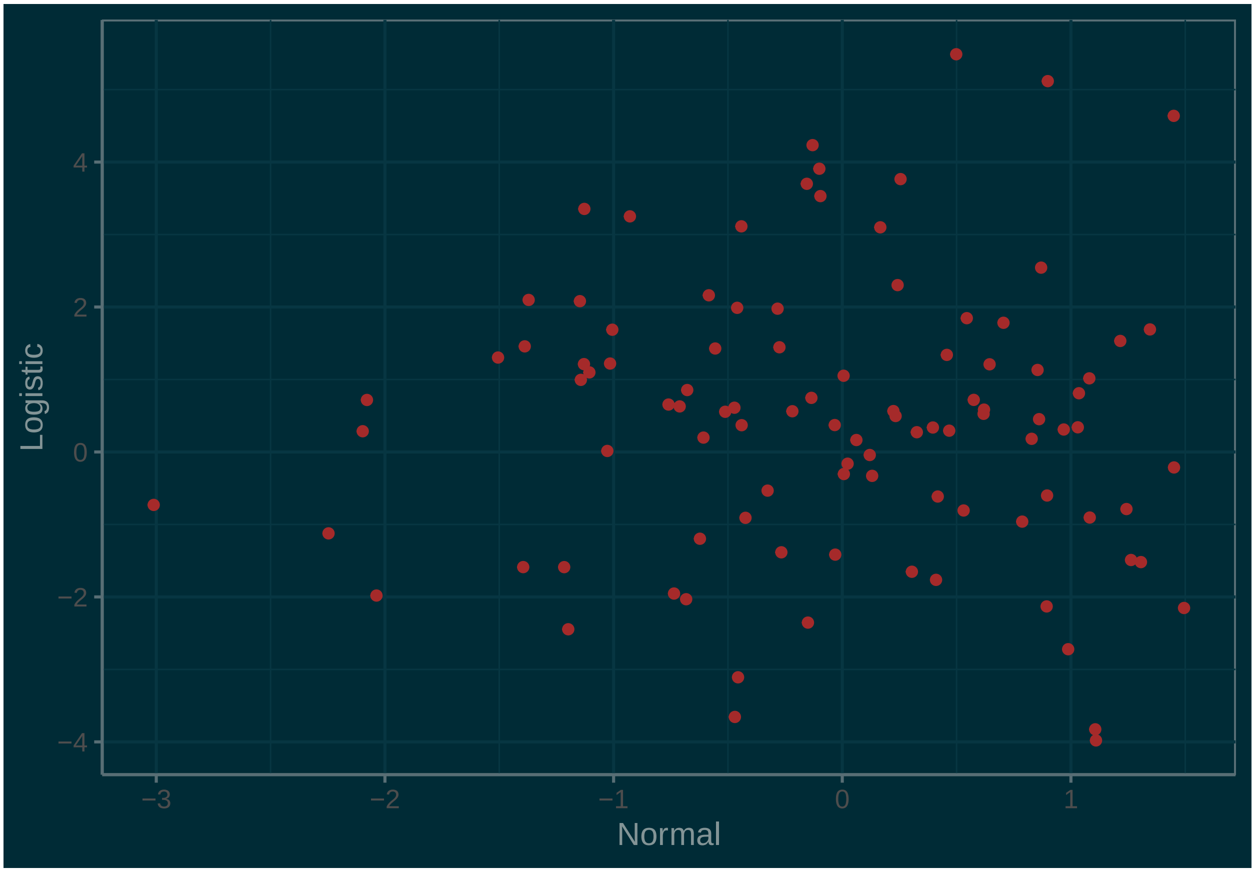
```
ggplot(df_lung, aes(x = Mortality_Rate)) + geom_density()
```



4. Create a Scatter Plot by generating 100 random values from both the normal and logistic distributions. The points should be brown and use theme_solarized with argument light set to false.

```
x = rnorm(100)
y = rlogis(100)

ggplot() + geom_point(aes(x,y), color="brown") + labs(x = "Normal", y = "Logistic") + theme_solarized(li
```



Part 4: Recreate the following graphs

2. Use the gpplot2 package for this graph. (Hint: Aggregate the data then merge the two datasets. Use only the necessary columns.)

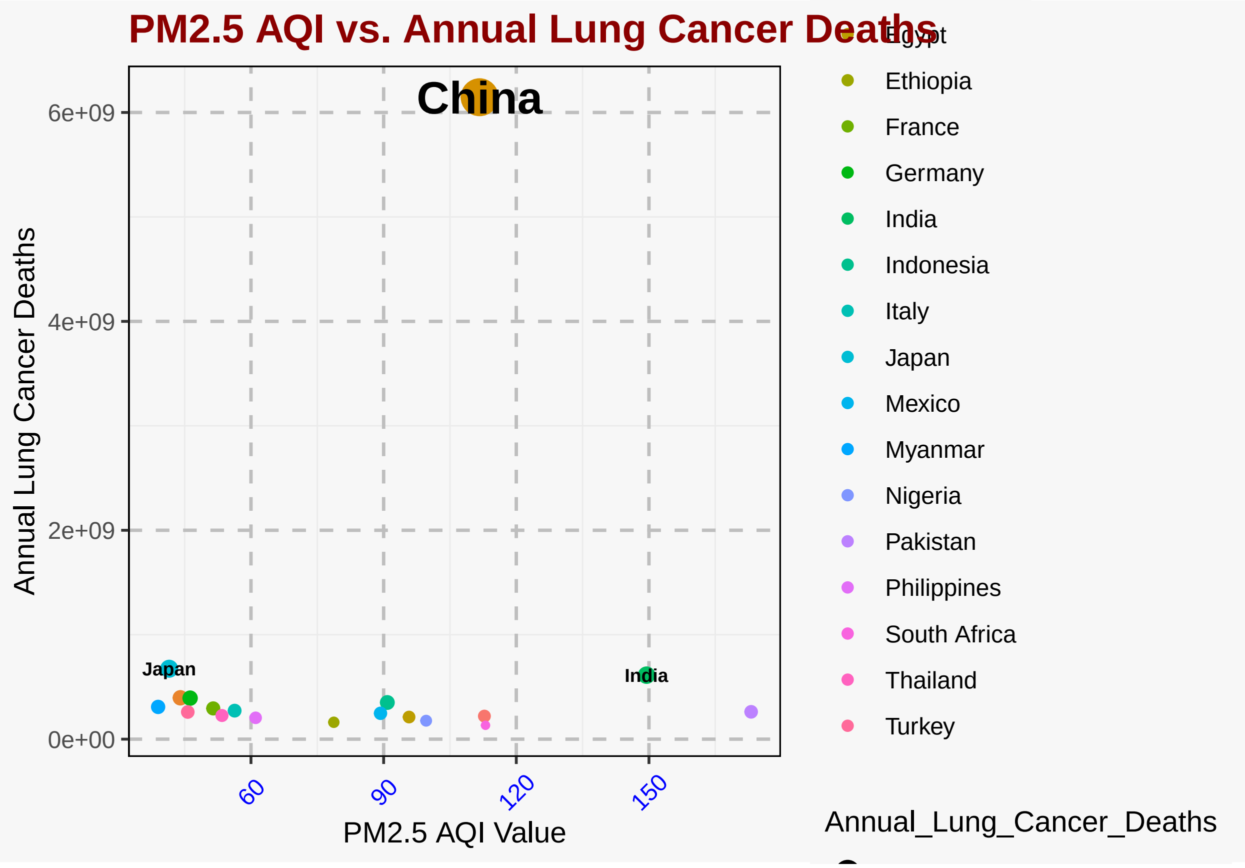
```
df1 <- group_by(df_air, Country) %>% summarise(PM2.5_AQI_Value = mean(PM2.5_AQI_Value))
df2 <- group_by(df_lung, Country) %>% summarise(Annual_Lung_Cancer_Deaths = sum(Annual_Lung_Cancer_Deat

joined_df <- inner_join(df1, df2, by="Country")

ggplot(joined_df, aes(x = PM2.5_AQI_Value, y = Annual_Lung_Cancer_Deaths, color = Country)) +
  geom_point(aes(size=Annual_Lung_Cancer_Deaths)) +
  labs(title = "PM2.5 AQI vs. Annual Lung Cancer Deaths", x = "PM2.5 AQI Value", y = "Annual Lung Cance
  geom_text(
    aes(label = ifelse(Annual_Lung_Cancer_Deaths > 500000000, Country, ''),
    size=Annual_Lung_Cancer_Deaths), color="black",
    fontface = "bold") +
  theme(
    plot.title = element_text(color = "darkred", face="bold", size=15),
    plot.background = element_rect(fill = "#f7f7f7"),
    panel.border = element_rect(color = "black", fill=NA),
    panel.background = element_rect(fill = "#f7f7f7",
                                    size = 0.5, linetype = "solid"),
    panel.grid.major = element_line(size = 0.6, linetype = 'dashed',
                                    color = "gray"),
    panel.grid.minor = element_line(size = 0.25, linetype = 'solid',
                                  color = "#ebebeb"),
    legend.background = element_rect(fill = "#f7f7f7"),
    axis.text.x = element_text(angle = 45, vjust = 0.5, color="blue")
    )
```

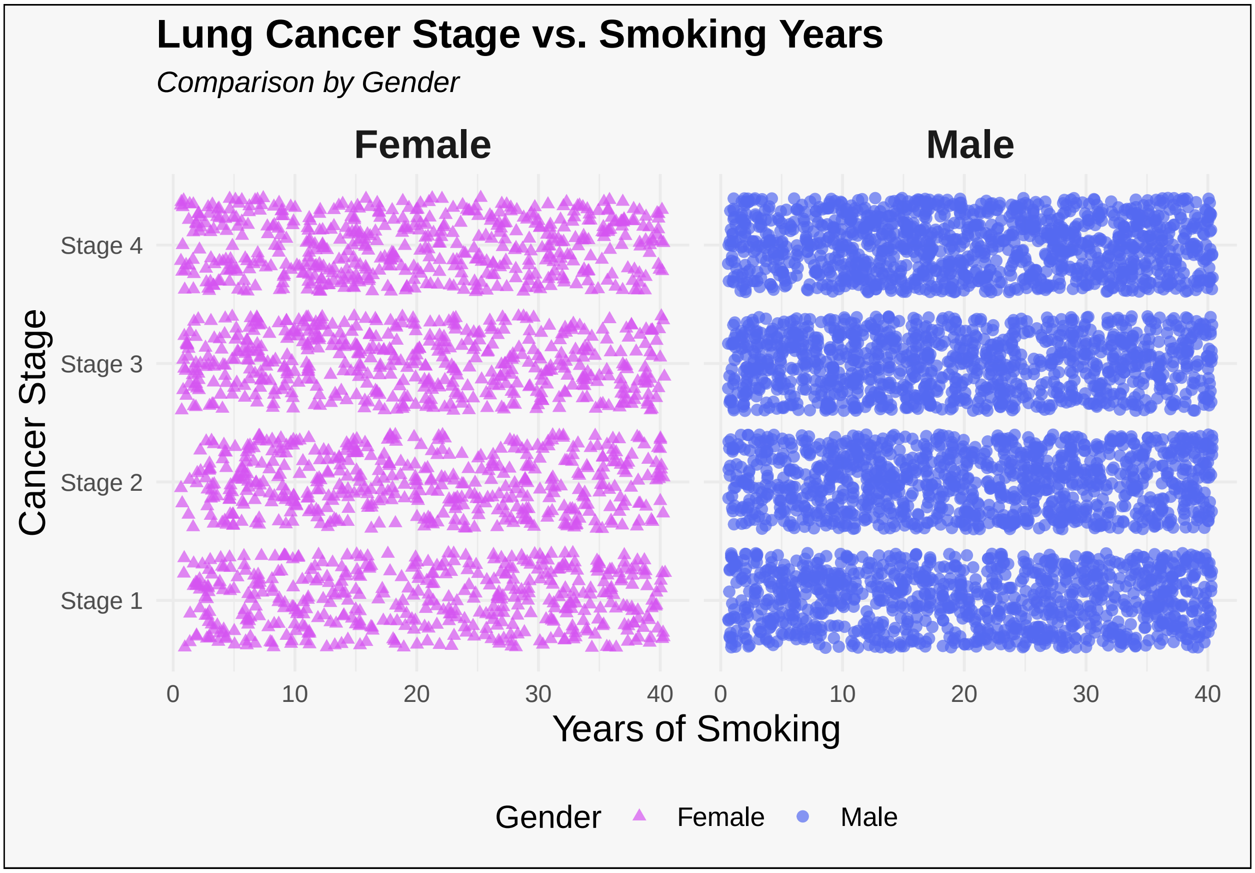
```
## Warning: The `size` argument of `element_rect()` is deprecated as of ggplot2 3.4.0.
## i Please use the `linewidth` argument instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## Warning: The `size` argument of `element_line()` is deprecated as of ggplot2 3.4.0.
## i Please use the `linewidth` argument instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



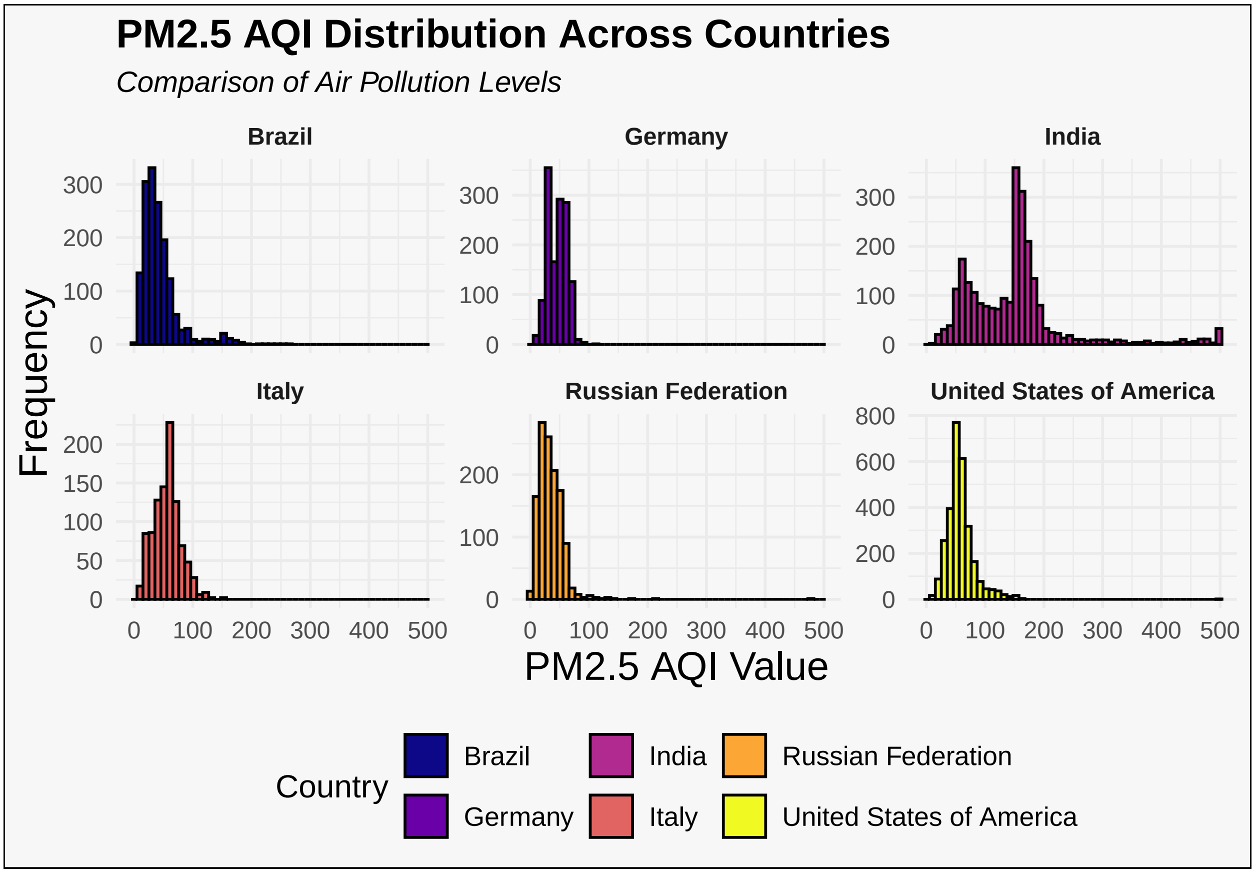
3. Use the ggplot2 package for this graph. (Hint: use geom_jitter since y axis contains categorical data, also use the following colors: #5469f1 , #d554f1)

```
filtered_df_lung <-  df_lung %>% filter(Cancer_Stage != "None", Years_of_Smoking > 0)
ggplot(filtered_df_lung,
       aes(x = Years_of_Smoking, y=Cancer_Stage, color = Gender, shape=Gender)
  ) +
  geom_jitter(alpha=0.7) +
  facet_grid(~Gender) +
  scale_color_manual(values = c("#d554f1","#5469f1")) +
  scale_shape_manual(values = c(17,19)) +
  labs(
    title = "Lung Cancer Stage vs. Smoking Years",
    subtitle = "Comparison by Gender",
    x = "Years of Smoking",
    y = "Cancer Stage"
    ) +
  theme_minimal() +
  theme(
    plot.background = element_rect(fill = "#f7f7f7"),
    plot.title = element_text(face="bold", size=15),
    plot.subtitle = element_text(face="italic"),
    strip.text.x = element_text(size = 15, face="bold"),
    legend.position = "bottom",
    legend.title = element_text(size=12),
    legend.text = element_text(size=10),
    axis.title=element_text(size=14)
  )
```



4. Use the ggplot2 package for this graph. (Hint: use scale_fill_viridis_d(option = "plasma" to get the same colors)

```
filtered_df_air <-  df_air %>% filter(Country %in% c("Brazil", "India", "Russian Federation", "Germany"
ggplot(filtered_df_air , aes(x = PM2.5_AQI_Value, fill=Country)) +
  facet_wrap(~Country, scales = "free_y") +
  geom_histogram(bins=50, color="black") +
  scale_fill_viridis_d(option = "plasma") +
  labs(
    title = "PM2.5 AQI Distribution Across Countries",
    subtitle = "Comparison of Air Pollution Levels",
    x = "PM2.5 AQI Value",
    y = "Frequency"
    ) +
  theme_minimal() +
  theme(
    plot.background = element_rect(fill = "#f7f7f7"),
    plot.title = element_text(face="bold", size=15),
    plot.subtitle = element_text(face="italic"),
    strip.text.x = element_text(face="bold"),
    legend.position = "bottom",
    legend.title = element_text(size=12),
    legend.text = element_text(size=10),
    axis.title=element_text(size=15)
  )
```


PM2.5 AQI Distribution Across Countries
Comparison of Air Pollution Levels
Brazil
Germany
India
Italy
Russian Federation
United States of America
Frequency
PM2.5 AQI Value
Country
Brazil
Germany
India
Italy
Russian Federation
United States of America