HPAM 7660 Data Assignment 4

Rebecca Letsinger

March 5, 2024

```
    library(dplyr)
```

```
##
## 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
```

```
    library(knitr)
    library(ggplot2)
```

Three essential components of data visualization: 1. Data - this includes the information and variables used and of interest. In figure 2.1, the data includes life expectancy over GDP per capita in 2007 for 147 country. The information collected under each variable represents data. 2. Geom - this describes the style of geometric object being used.This can include lines, points, and bars. In figure 2.1 this is demonstrated by the individual points on the graph, the decision to select a scatterplot, the axis and legends. 3. Aes - this refers to the aesthetic of the chosen geom. This can include position, shape, color, and size. In figure 2.1 this can be indicated by the different color and size of the dots on the graph.

```
library(readr)
la_mort <-
  read_csv("https://www.dropbox.com/scl/fi/fzsnhfd3lq80v2o3sag6c/la_mort.csv?rlkey=h1vyjm2b8ppgejgsg3e8
```

```
## Rows: 642696 Columns: 29
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): stocr, strsd, stbrth, brthr, sex, marstat, ucod
## dbl (22): restatus, cntyocr, popcntyocr, cntyrsd, popcntyresd, educ1989, edu...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
la_mort$cancer_parish <- ifelse(la_mort$cntyrsd %in% c(5, 33, 47, 51, 71, 89, 93, 95, 121), 1, 0)
```

```r
table(la_mort$cancer_parish)
```

```
##
##      0      1
## 445138 197558
```

```r
table(la_mort$cntyrsd[la_mort$cancer_parish == 1])
```

```
##
##     5    33    47    51    71    89    93    95   121
## 10217 55300  4761 61822 47752  5963  2946  5844  2953
```

```r
la_mort$cancer39 <- ifelse(la_mort$ucr39 %in% c(5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15), 1, 0)
```

```r
table(la_mort$cancer39)
```

```
##
##      0      1
## 504019 138677
```

```r
la_mort$cancer39 <- ifelse(la_mort$ucr39 %in% c(5:15), 1, 0)
```

```r
table(la_mort$ucr39[la_mort$cancer113 == 1 & la_mort$cancer39 == 0])
```

```
## Warning: Unknown or uninitialised column: ‘cancer113‘.
```

```
## < table of extent 0 >
```

```r
table(la_mort$ucod[la_mort$cancer113 == 1 & la_mort$cancer39 == 0])
```

```
## Warning: Unknown or uninitialised column: ‘cancer113‘.
```

```
## < table of extent 0 >
```

```r
library(dplyr)
```

```r
parish_count <- la_mort %>%
  group_by(cntyrsd, cancer_parish, year) %>%
  summarize(cancer39 = sum(cancer39, na.rm = TRUE))
```

```
## ‘summarise()‘ has grouped output by ’cntyrsd’, ’cancer_parish’. You can
## override using the ‘.groups‘ argument.
```

```r
summary(parish_count$cancer39)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     3.0    42.0    74.0   144.5   159.0   992.0
```

```r
library(readr)
```

```r
la_pop <-
  read_csv("https://www.dropbox.com/scl/fi/650k1obpczky6bwa19ex6/la_county_pop.csv?rlkey=0aokd9m76q7mxw
```

```
## Rows: 24320 Columns: 23
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (3): stname, ctyname, agegrp
## dbl (20): state, county, year, tot_pop, tot_male, tot_female, wa_male, wa_fe...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
parish_count <- parish_count %>%
  rename(county = cntyrsd)
```

```r
la_joined <- parish_count %>%
  inner_join(la_pop, by = c("county", "year"))
la_joined_all <- subset(la_joined, agegrp == "all")
```

```r
la_joined_all$cancer_rate_total <- (la_joined_all$cancer39) / (la_joined_all$tot_pop)
```

```r
summary(la_joined_all$cancer_rate_total)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0001157 0.0018691 0.0021703 0.0021985 0.0024863 0.0039361
```

```r
la_joined_all$cancer_rate_total <- ((la_joined_all$cancer39) / (la_joined_all$tot_pop / 100000))
```

```r
parish_cancer_2019 <- subset(la_joined_all, year == 2019)
library(knitr)
kable(parish_cancer_2019[, c("county", "cancer_rate_total")])
```

county	cancer_rate_total
1	225.0551
3	210.9870
5	127.7623
7	250.9010
9	249.1405
11	239.0502
13	264.4703

county	cancer_rate_total
15	178.9962
17	232.0832
19	209.7648
21	251.3573
23	157.5705
25	169.1511
27	222.6746
29	269.2348
31	236.8805
33	167.8479
35	380.7292
37	287.9129
39	233.5609
41	255.0510
43	201.3243
45	222.5796
47	199.6989
49	336.2731
51	210.7829
53	244.8798
55	167.0765
57	220.2463
59	228.0349
61	164.3508
63	157.5293
65	254.6844
67	261.6853
69	233.1124
71	186.8743
73	224.9008
75	167.6410
77	257.6490
79	201.8490
81	310.2625
83	198.5407
85	280.3934
87	116.5452
89	186.5004
91	246.4268
93	185.3172
95	182.1409
97	259.1367
99	186.9264
101	251.5519
103	206.2739
105	204.6794
107	393.6096
109	228.7212
111	290.0127
113	193.0405
115	214.9027
117	275.1419

county	cancer_rate_total
119	250.5023
121	154.4256
123	238.4009
125	166.9878
127	336.6521

```
la_mort <-
  read_csv("https://www.dropbox.com/scl/fi/fzsnhfd3lq80v2o3sag6c/la_mort.csv?rlkey=h1vyjm2b8ppgejgsg3e8
```

```
## Rows: 642696 Columns: 29
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): stocr, strsd, stbrth, brthr, sex, marstat, ucod
## dbl (22): restatus, cntyocr, popcntyocr, cntyrsd, popcntyresd, educ1989, edu...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
show_col_types = FALSE
```

```
la_mort$cancer_parish <- ifelse(la_mort$cntyrsd %in% c(5, 33, 47, 51, 71, 89, 93, 95, 121), 1, 0)
```

```
la_mort$cancer39 <- ifelse(la_mort$ucr39 %in% c(5:15), 1, 0)
```

```
library(dplyr)
la_mort_age <- la_mort %>%
  filter(age != 9999)
la_mort_age$age <- ifelse(la_mort_age$age < 2000, la_mort_age$age - 1000, 0)
```

```
age_breaks <- c(0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, Inf)
age_labels <- c("0_4", "5_9", "10_14", "15_19", "20_24", "25_29", "30_34", "35_39",
                "40_44", "45_49", "50_54", "55_59", "60_64", "65_69", "70_74",
                "75_79", "80_84", "85+")
la_mort_age$agegrp <- as.character(cut(la_mort_age$age, breaks = age_breaks, labels = age_labels, right
```

```
parish_count <- la_mort %>%
  group_by(cntyrsd, cancer_parish, year) %>%
  summarize(cancer39 = sum(cancer39, na.rm = TRUE))
```

```
## `summarise()` has grouped output by 'cntyrsd', 'cancer_parish'. You can
## override using the `.groups` argument.
```

```
parish_count_age <- la_mort_age %>%
  group_by(cntyrsd, cancer_parish, agegrp, year) %>%
  summarize(cancer39 = sum(cancer39, na.rm = TRUE))
```

```
## `summarise()` has grouped output by 'cntyrsd', 'cancer_parish', 'agegrp'. You
## can override using the `.groups` argument.
```

```r
library(readr)
```

```r
la_pop <-
  read_csv("https://www.dropbox.com/scl/fi/650k1obpczky6bwa19ex6/la_county_pop.csv?rlkey=0aokd9m76q7mxw
```

```
## Rows: 24320 Columns: 23
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (3): stname, ctyname, agegrp
## dbl (20): state, county, year, tot_pop, tot_male, tot_female, wa_male, wa_fe...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
library(dplyr)
```

```r
la_joined <- parish_count_age %>%
  inner_join(la_pop, by = c("cntyrsd" = "county", "year", "agegrp"))
```

```r
stnrd_pop <-
  read_csv("https://www.dropbox.com/scl/fi/xzd2o5lza237so6vamqwb/stnrd_pop.csv?rlkey=zp90au2tuq6eptvi1y
```

```
## Rows: 18 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): agegrp
## dbl (1): stnrd_pop
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
la_joined_stnrd <- la_joined %>%
  inner_join(stnrd_pop, by = "agegrp")
```

```r
la_joined_stnrd$stnrd_pop_weight <- (la_joined_stnrd$stnrd_pop) / (sum(stnrd_pop$stnrd_pop))
```

```r
la_joined_stnrd$cancer_rate_adj <- ((la_joined_stnrd$cancer39) / (la_joined_stnrd$tot_pop / 100000)) * 
```

```r
parish_rates <- la_joined_stnrd %>%
  group_by(cntyrsd, cancer_parish, year) %>%
  summarize(cancer_rate_adj = sum(cancer_rate_adj, na.rm = TRUE), cancer39 = sum(cancer39), tot_pop =
              sum(tot_pop))
```

```
## `summarise()` has grouped output by 'cntyrsd', 'cancer_parish'. You can
## override using the `.groups` argument.
```

```r
parish_rates$cancer_rate_crude <- (parish_rates$cancer39) / (parish_rates$tot_pop / 100000)
```

```
parish_rates$pop_weight <- (parish_rates$cancer_rate_adj) * (parish_rates$tot_pop)
cancer_alley_rates <- parish_rates %>%
  group_by(cancer_parish, year) %>%
  summarize(cancer_rate_adj_wt = sum(pop_weight) / sum(tot_pop))
```

```
## `summarise()` has grouped output by 'cancer_parish'. You can override using the
## `.groups` argument.
```

```
parish_rates$pop_weight <- (parish_rates$cancer_rate_adj) * (parish_rates$tot_pop)
cancer_alley_rates <- parish_rates %>%
  group_by(cancer_parish, year) %>%
  summarize(cancer_rate_adj_wt = sum(pop_weight) / sum(tot_pop))
```

```
## `summarise()` has grouped output by 'cancer_parish'. You can override using the
## `.groups` argument.
```

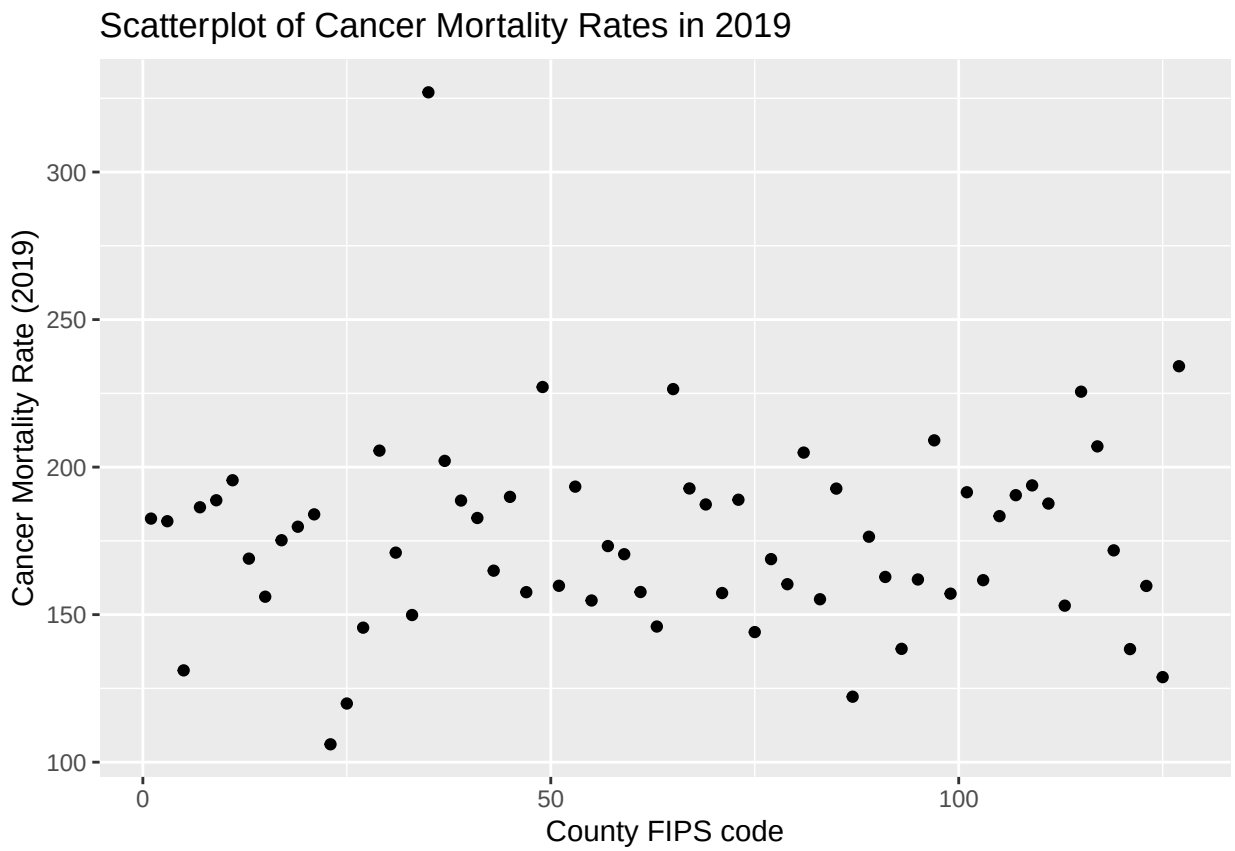
```
kable(cancer_alley_rates)
```

cancer_parish	year	cancer_rate_adj_wt
0	2005	215.9012
0	2006	211.1969
0	2007	199.2163
0	2008	210.5785
0	2009	202.7788
0	2010	198.5223
0	2011	194.5824
0	2012	194.9155
0	2013	191.4183
0	2014	188.3508
0	2015	186.8605
0	2016	178.2077
0	2017	181.0797
0	2018	176.0163
0	2019	174.1137
1	2005	197.2898
1	2006	198.7948
1	2007	199.3910
1	2008	196.7380
1	2009	190.6874
1	2010	191.1738
1	2011	189.7244
1	2012	180.9129
1	2013	181.2483
1	2014	181.1850
1	2015	166.3009
1	2016	157.8499
1	2017	161.2732
1	2018	153.9050
1	2019	153.9429

```r
cancer_alley <-
  subset(cancer_alley_rates, cancer_parish == 1, select = c(cancer_rate_adj_wt, year)) %>%
  rename(cancer_alley_rate = cancer_rate_adj_wt)
no_cancer_alley <-
  subset(cancer_alley_rates, cancer_parish == 0, select = c(cancer_rate_adj_wt, year)) %>%
  rename(no_cancer_alley_rate = cancer_rate_adj_wt)
cancer_alley_table <- cancer_alley %>%
  inner_join(no_cancer_alley, by = "year")
cancer_alley_table <- cancer_alley_table[,c("year", "cancer_alley_rate", "no_cancer_alley_rate")]
kable(cancer_alley_table)
```

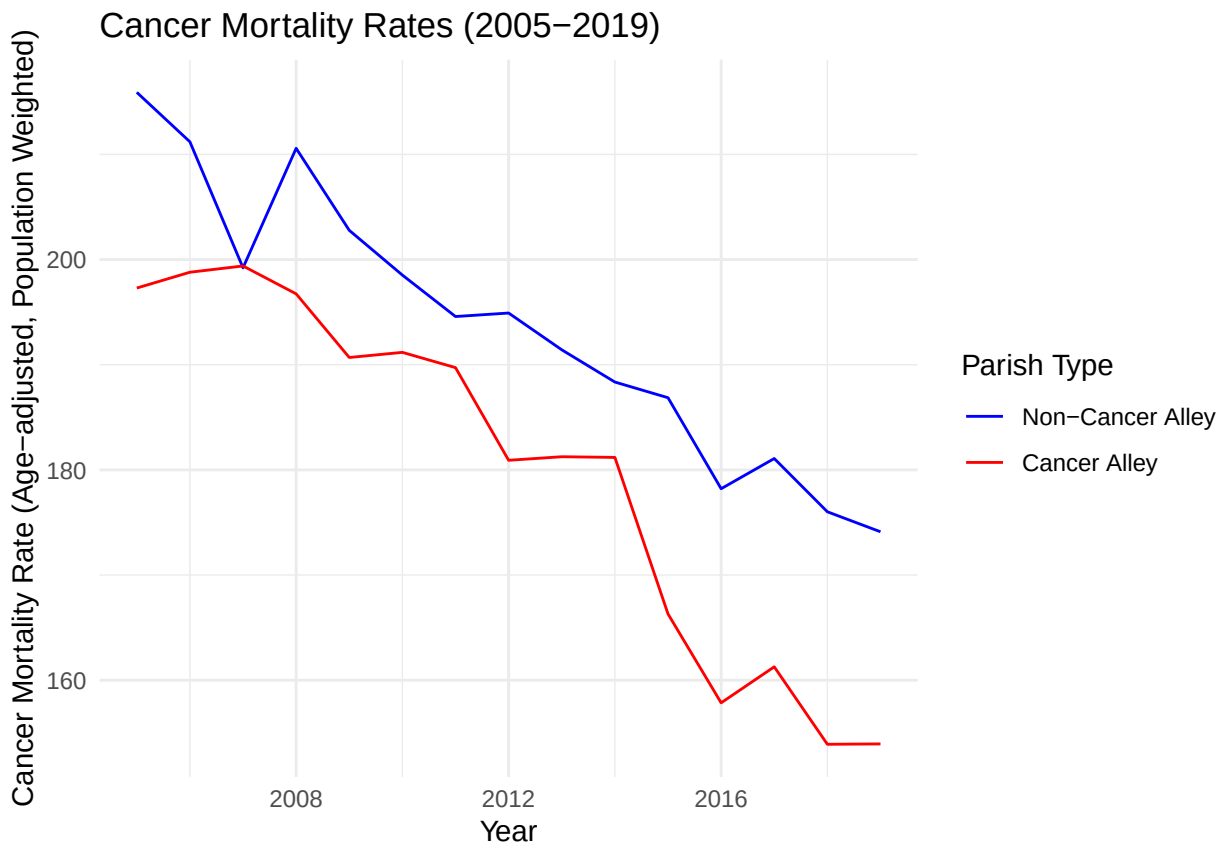
year	cancer_alley_rate	no_cancer_alley_rate
2005	197.2898	215.9012
2006	198.7948	211.1969
2007	199.3910	199.2163
2008	196.7380	210.5785
2009	190.6874	202.7788
2010	191.1738	198.5223
2011	189.7244	194.5824
2012	180.9129	194.9155
2013	181.2483	191.4183
2014	181.1850	188.3508
2015	166.3009	186.8605
2016	157.8499	178.2077
2017	161.2732	181.0797
2018	153.9050	176.0163
2019	153.9429	174.1137

```r
parish_rates_2019 <- subset(parish_rates, year == 2019)
ggplot(data = parish_rates_2019, aes(x = cntyrsd, y = cancer_rate_adj)) +
  geom_point() +
  labs(x = "County FIPS code", y = "Cancer Mortality Rate (2019)",
       title = "Scatterplot of Cancer Mortality Rates in 2019")
```

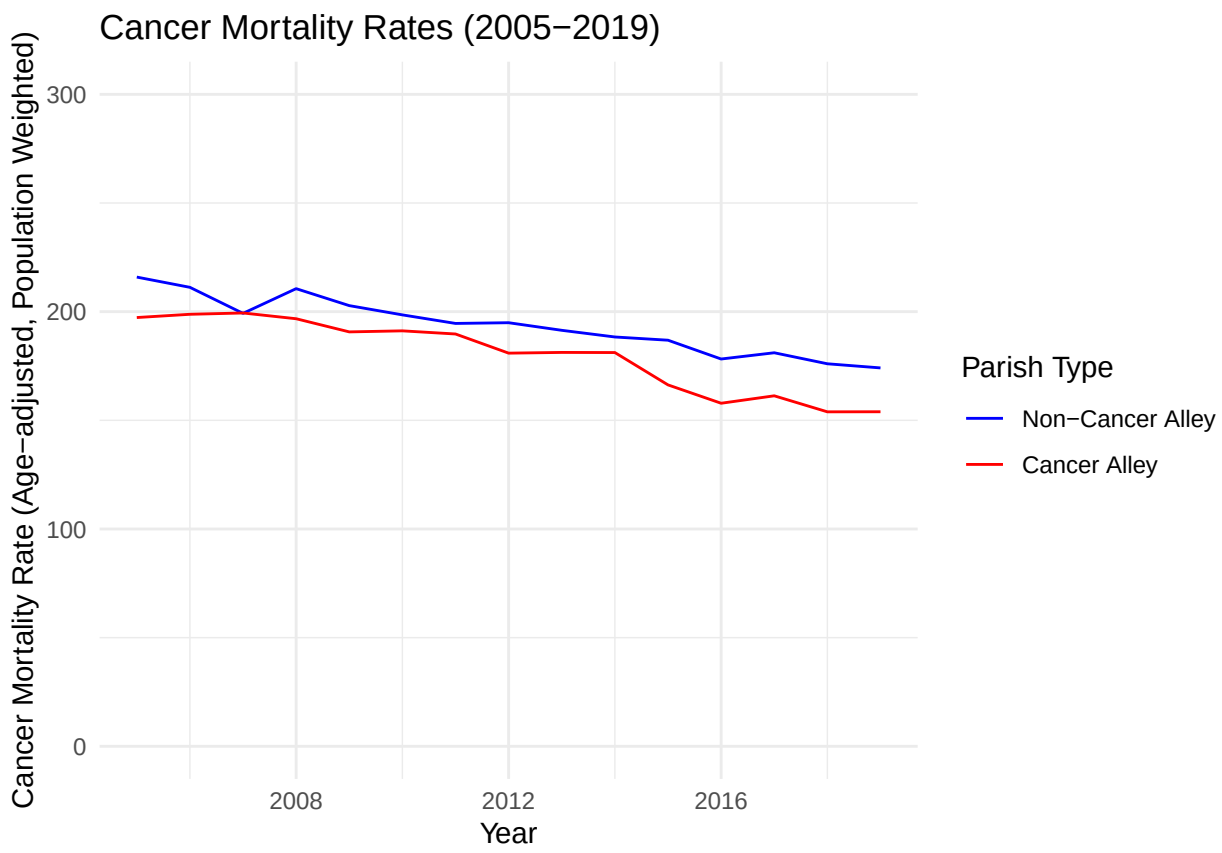



```
filtered_data <- subset(cancer_alley_rates, year >= 2005 & year <= 2019)
```

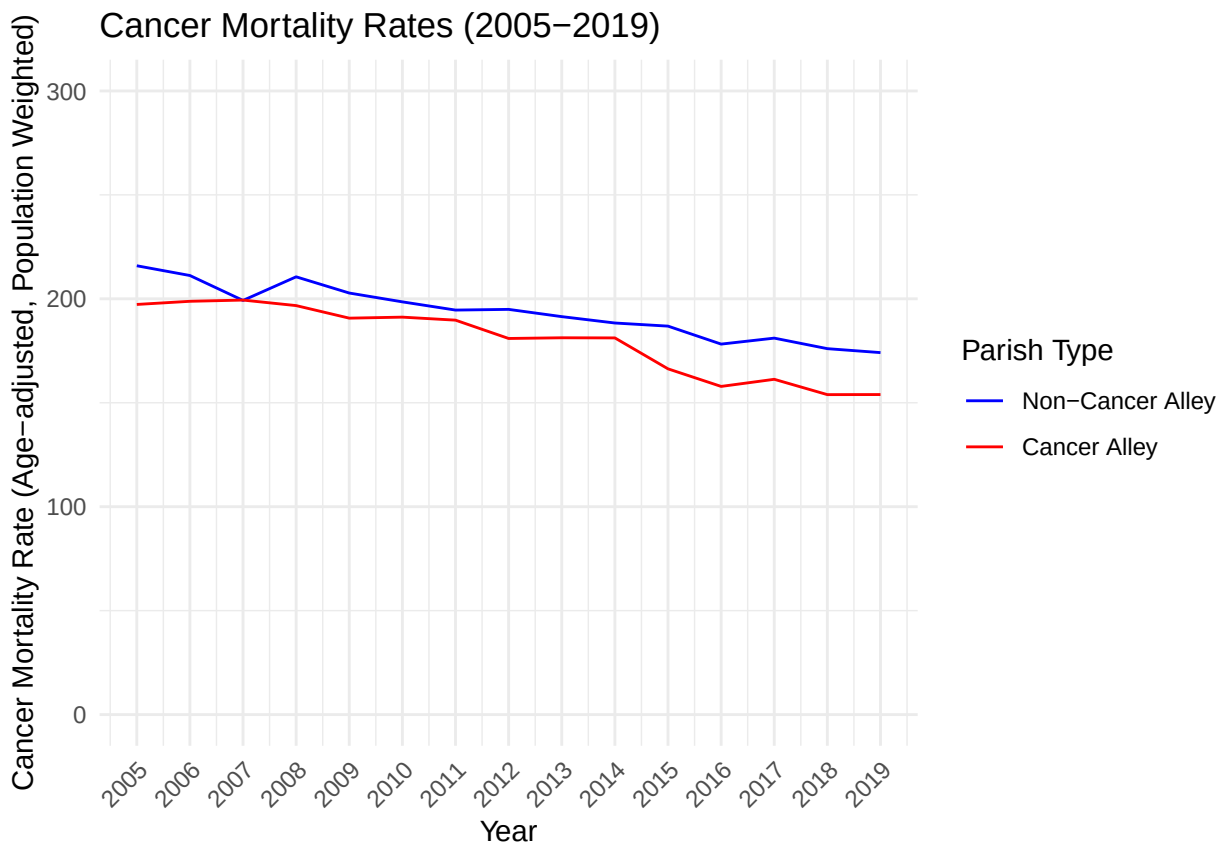
```
ggplot(data = filtered_data, aes(x = year, y = cancer_rate_adj_wt, color = factor(cancer_parish))) +
  geom_line() +
  labs(x = "Year", y = "Cancer Mortality Rate (Age-adjusted, Population Weighted)",
       color = "Parish Type", title = "Cancer Mortality Rates (2005-2019)") +
  scale_color_manual(values = c("blue", "red"), labels = c("Non-Cancer Alley", "Cancer Alley")) +
  theme_minimal()
```



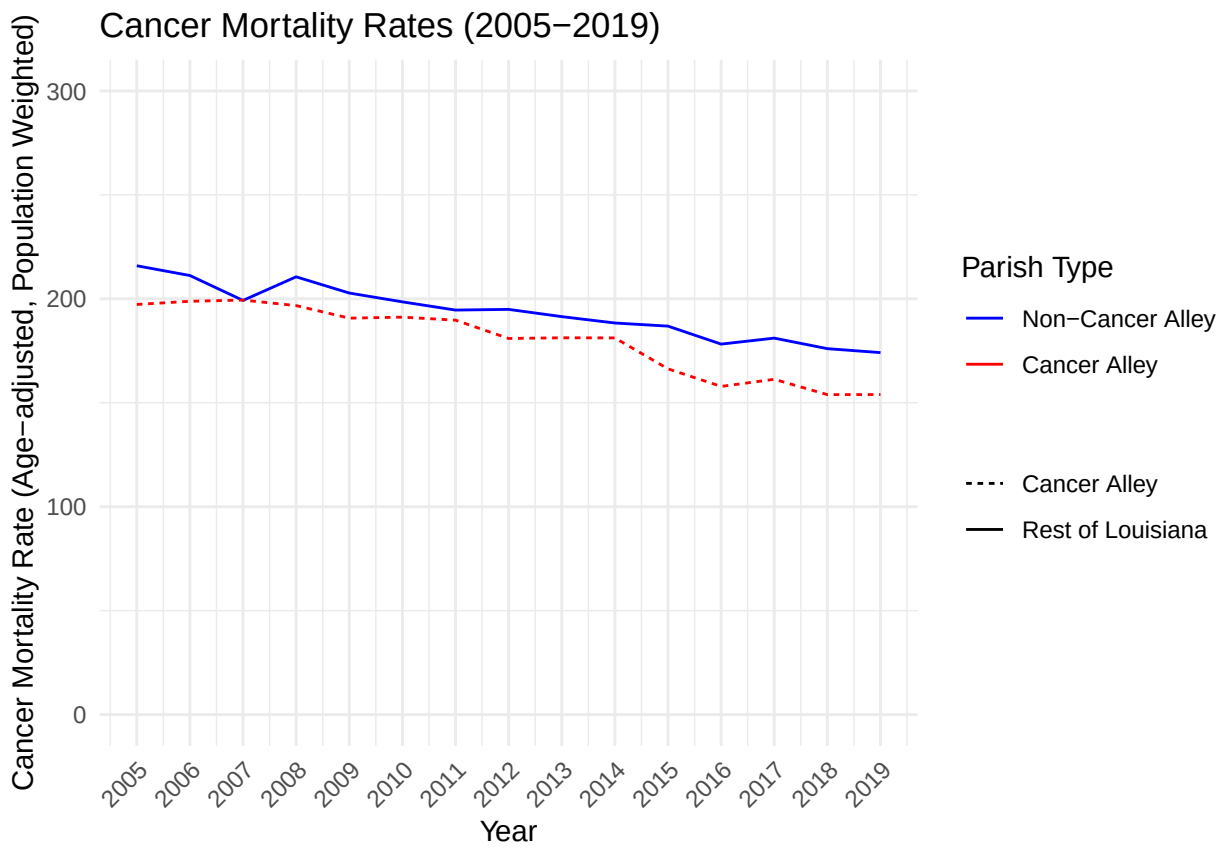

```
ggplot(data = filtered_data, aes(x = year, y = cancer_rate_adj_wt, color = factor(cancer_parish))) +
  geom_line() +
  labs(x = "Year", y = "Cancer Mortality Rate (Age-adjusted, Population Weighted)",
       color = "Parish Type", title = "Cancer Mortality Rates (2005-2019)") +
  scale_color_manual(values = c("blue", "red"), labels = c("Non-Cancer Alley", "Cancer Alley")) +
  scale_y_continuous(limits = c(0, 300)) + # Adjusting y-axis scale
  theme_minimal()
```



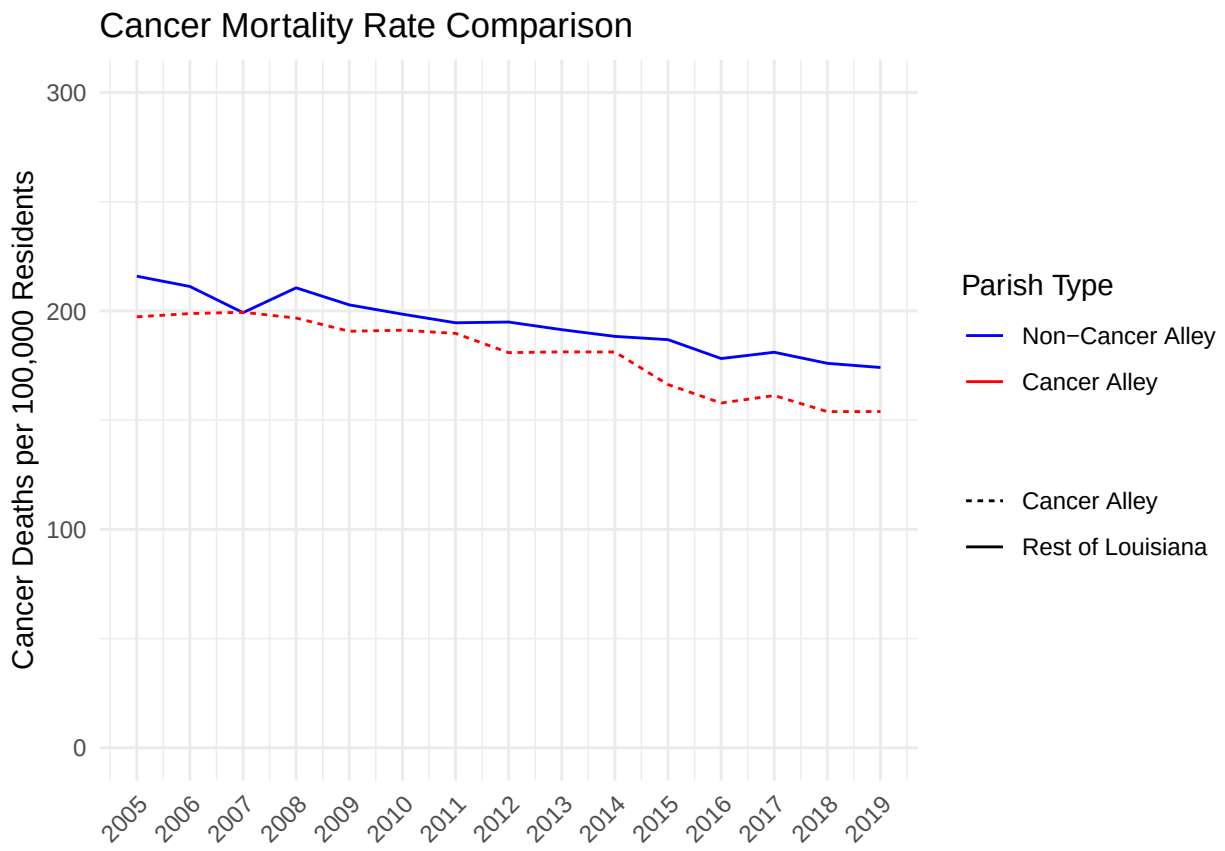

```
ggplot(data = filtered_data, aes(x = year, y = cancer_rate_adj_wt, color = factor(cancer_parish))) +
  geom_line() +
  labs(x = "Year", y = "Cancer Mortality Rate (Age-adjusted, Population Weighted)",
       color = "Parish Type", title = "Cancer Mortality Rates (2005-2019)") +
  scale_color_manual(values = c("blue", "red"), labels = c("Non-Cancer Alley", "Cancer Alley")) +
  scale_y_continuous(limits = c(0, 300)) + # Adjusting y-axis scale
  scale_x_continuous(breaks = seq(2005, 2019, by = 1)) + # Ensuring all year values between 2005 and 20
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) # Rotating year labels by 45 degrees
```



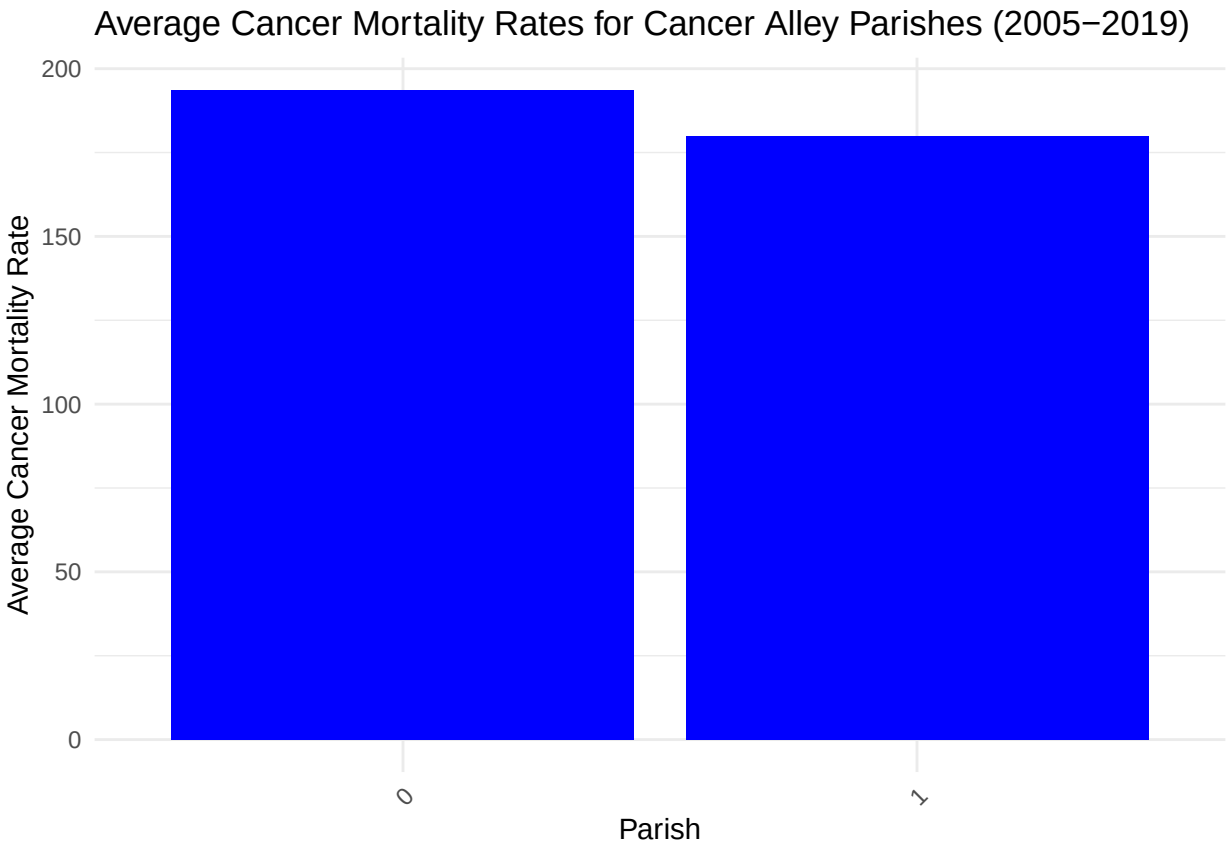

```
ggplot(data = filtered_data, aes(x = year, y = cancer_rate_adj_wt, color = factor(cancer_parish), linety
  geom_line() +
  labs(x = "Year", y = "Cancer Mortality Rate (Age-adjusted, Population Weighted)",
       color = "Parish Type", linetype = "Parish Type", title = "Cancer Mortality Rates (2005-2019)") +
  scale_color_manual(values = c("blue", "red"), labels = c("Non-Cancer Alley", "Cancer Alley")) +
  scale_linetype_discrete(name = NULL, labels = c("Rest of Louisiana", "Cancer Alley"), guide = guide_l
  scale_y_continuous(limits = c(0, 300)) + # Adjusting y-axis scale
  scale_x_continuous(breaks = seq(2005, 2019, by = 1)) + # Ensuring all year values between 2005 and 20
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) # Rotating year labels by 45 degrees
```




```
ggplot(data = filtered_data, aes(x = year, y = cancer_rate_adj_wt, color = factor(cancer_parish), linety
  geom_line() +
  labs(title = "Cancer Mortality Rate Comparison", y = "Cancer Deaths per 100,000 Residents", x = NULL,
  scale_color_manual(values = c("blue", "red"), labels = c("Non-Cancer Alley", "Cancer Alley")) +
  scale_linetype_discrete(name = NULL, labels = c("Rest of Louisiana", "Cancer Alley"), guide = guide_l
  scale_y_continuous(limits = c(0, 300)) + # Adjusting y-axis scale
  scale_x_continuous(breaks = seq(2005, 2019, by = 1)) + # Ensuring all year values between 2005 and 20
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) # Rotating year labels by 45 degrees
```




```
filtered_data$cancer_parish <- factor(filtered_data$cancer_parish)
average_mortality <- filtered_data %>%
  group_by(cancer_parish) %>%
  summarise(average_rate = mean(cancer_rate_adj_wt))
ggplot(average_mortality, aes(x = cancer_parish, y = average_rate)) +
  geom_bar(stat = "identity", fill = "blue") +
  labs(x = "Parish", y = "Average Cancer Mortality Rate",
       title = "Average Cancer Mortality Rates for Cancer Alley Parishes (2005-2019)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```


Average Cancer Mortality Rates for Cancer Alley Parishes (2005−2019)
Average Cancer Mortality Rate
200
150
100
50
0
0
1
Parish