Final_Mark_Down

Quinn Bankson

2023-05-04

DarkOrchid4 Final Mark Down

How does proximity to coal fired power plant affect health outcomes in the United States?

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   1.0.1
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.3.0      v stringr 1.5.0
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(moderndive)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(janitor)
```

```
##
## Attaching package: 'janitor'
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(dplyr)
```

Introducing coal plant data to undergo feature engineering.

```
#Coal plant data was cleaned in an individual folder, the excel was converted to csv and excess workshe

coalplants <- read_csv(file = "/Users/mac/Documents/R Assignments PLCY 715/final-team-projects-darkorchi
```

```
## Rows: 13491 Columns: 37
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (27): Tracker ID, TrackerLOC, ParentID, Wiki page, Country, Subnational ...
## dbl  (8): Capacity (MW), RETIRED, Planned Retire, Latitude, Longitude, Annua...
## num  (2): Heat rate (Btu per kWh), Emission factor (kg of CO2 per TJ)
##
## 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.
```

```
# coalplants <- read_csv(file = "/Users/mac/Documents/R Assignments PLCY 715/final-team-projects-darkor

colnames(coalplants)
```

```
##  [1] "Tracker ID"                          "TrackerLOC"
##  [3] "ParentID"                            "Wiki page"
##  [5] "Country"                             "Subnational unit (province, state)"
##  [7] "Unit"                                "Plant"
##  [9] "Chinese Name"                        "Other names"
## [11] "Owner"                               "Parent"
## [13] "Capacity (MW)"                       "Status"
## [15] "Year"                                "RETIRED"
## [17] "Planned Retire"                      "Combustion technology"
## [19] "Coal type"                           "Coal source"
## [21] "Location"                            "Local area (taluk, county)"
## [23] "Major area (prefecture, district)"   "Region"
## [25] "Latitude"                            "Longitude"
## [27] "Accuracy"                            "Permits"
## [29] "Captive"                             "Captive industry use"
## [31] "Captive residential use"             "Heat rate (Btu per kWh)"
## [33] "Emission factor (kg of CO2 per TJ)"  "Capacity factor"
## [35] "Annual CO2 (million tonnes / annum)" "Lifetime CO2"
## [37] "Remaining plant lifetime (years)"
```

```
coalplants <- clean_names(coalplants)
coalplants <- coalplants %>% filter(country == "United States")
coalplants <- coalplants %>% select(parent_id, unit, subnational_unit_province_state, plant, status, yea
coalplants <- rename(coalplants, county = local_area_taluk_county)
```

```
counties <- read_csv(file = "/Users/mac/Documents/R Assignments PLCY 715/final-team-projects-darkorchid/
```

```
## Rows: 3143 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (6): county, county_ascii, county_full, county_fips, state_id, state_name
## dbl (3): lat, lng, population
```

```
##
## 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.
```

```
head(counties)
```

```
## # A tibble: 6 x 9
##   county      county_ascii county~1 count~2 state~3 state~4   lat    lng popul~5
##   <chr>       <chr>        <chr>    <chr>   <chr>   <chr>   <dbl>  <dbl>   <dbl>
## 1 Los Angeles Los Angeles  Los Ang~ 06037   CA      Califo~  34.3 -118.   1.00e7
## 2 Cook        Cook         Cook Co~ 17031   IL      Illino~  41.8  -87.8  5.27e6
## 3 Harris      Harris       Harris ~ 48201   TX      Texas    29.9  -95.4  4.70e6
## 4 Maricopa    Maricopa     Maricop~ 04013   AZ      Arizona  33.3 -112.   4.37e6
## 5 San Diego   San Diego    San Die~ 06073   CA      Califo~  33.0 -117.   3.30e6
## 6 Orange      Orange       Orange ~ 06059   CA      Califo~  33.7 -118.   3.18e6
## # ... with abbreviated variable names 1: county_full, 2: county_fips,
## #   3: state_id, 4: state_name, 5: population
```

```
colnames(counties)
```

```
## [1] "county"       "county_ascii" "county_full"  "county_fips"  "state_id"
## [6] "state_name"   "lat"          "lng"          "population"
```

```
county_center <- counties %>% select(county, state_name, lat, lng)
```

```
coalplants_op <- coalplants %>% filter(status == "operating")
coalplants_op <- distinct(coalplants_op, plant, .keep_all = TRUE)
```

```
coal_plants <- coalplants_op %>% select("latitude", "longitude","subnational_unit_province_state", "plan
coal_plants <- rename(coal_plants, long= longitude, lat = latitude)
coal_plants <- rename(coal_plants, state = subnational_unit_province_state, name = plant)
county_centroids <- county_center
county_centroids <- rename(county_centroids, state = state_name)
```

```
### "latitude", "longitude","subnational_unit_province_state", "plant", "county" were selected as the m
```

Feature engineering - calculating the minimum distance from every county centroid in the US to the nearest coal fired power plant using the Haversine formula and longitudinal and latitudinal data.

```
#install.packages("geosphere")
library(geosphere)
library(dplyr)
library(naniar)

county_centroids <- rename(county_centroids, long = lng)

coal_plants <- coal_plants %>%
  filter(!is.na(long))
```

```r
coal_plants <- coal_plants %>%
  filter(!is.na(lat))

county_centroids <- county_centroids %>%
  filter(!is.na(long))
county_centroids <- county_centroids %>%
  filter(!is.na(lat))
```

```r
library(dplyr)
library(geosphere)

# filtering to be sure
coal_plants <- na.omit(coal_plants)


distances <- geosphere::distm(county_centroids[, c("long", "lat")], coal_plants[, c("long", "lat")])

# minimum distance per county
min_distances <- apply(distances, 1, min)

# add the new column to the county_centroids dataframe using mutate
county_centroids <- county_centroids %>%
  mutate(distance_to_nearest_plant = min_distances)

coords_distance <- county_centroids
```

```r
# filtering to be sure, this code will atatch plant name as well as the minimum distance
coal_plants <- na.omit(coal_plants)

 #repeating earlier to be sure
distances <- geosphere::distm(county_centroids[, c("long", "lat")], coal_plants[, c("long", "lat")])

# minimum distance and corresponding plant name per county
min_distances <- apply(distances, 1, min)
names_of_nearest_plants <- apply(distances, 1, function(x) coal_plants$name[which.min(x)])

# add new columns to the county_centroids dataframe using mutate
county_centroids <- county_centroids %>%
  mutate(distance_to_nearest_plant = min_distances,
         name_of_nearest_plant = names_of_nearest_plants)

coords_distance <- county_centroids

write_csv(county_centroids, "/Users/mac/Documents/My Tableau Repository/countycoords.csv")
write_csv(coal_plants, "/Users/mac/Documents/My Tableau Repository/coalplantcoords.csv")

#write_csv(county_centroids, "/Users/mac/Documents/R Assignments PLCY 715/final-team-projects-darkorchi

#write_csv(coal_plants, "/Users/mac/Documents/R Assignments PLCY 715/final-team-projects-darkorchid/Dar
```

```r
summary(county_centroids$distance_to_nearest_plant)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##    4243   55230   95075  139529  152866 3182722
```

```
### This concludes the feature engineering in which the minimum distance from each county centroid to t
```

The next section of the .rmd will compile 3196 health outcomes (3143 of which correspond to the 3143 US counties that are observed) and aqi data from 1036 observed US counties as well. A crosswalk was used to match values by FIPS codes.

```
library(tidyverse)
library(ggplot2)
library(moderndive)
library(GGally)
library(janitor)
library(dplyr)
library(stringr)
#install.packages("tidyr")
library(tidyr)
```

```
crossw <- read_csv(file = "/Users/mac/Documents/R Assignments PLCY 715/final-team-projects-darkorchid/Da
```

```
## New names:
## Rows: 3274 Columns: 5
## -- Column specification
## ------------------------------------------------------- Delimiter: "," chr
## (5): FY 2023 Crosswalk, ...2, ...3, ...4, ...5
## 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.
## * `` -> `...2`
## * `` -> `...3`
## * `` -> `...4`
## * `` -> `...5`
```

```
aqi <- read_csv(file = "/Users/mac/Documents/R Assignments PLCY 715/final-team-projects-darkorchid/Dark
```

```
## Rows: 1036 Columns: 18
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): State, County
## dbl (16): Year, Days with AQI, Good Days, Moderate Days, Unhealthy for Sensi...
##
## 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.
```

```
outcomes <- read_csv(file = "/Users/mac/Documents/R Assignments PLCY 715/final-team-projects-darkorchid/
```

```
## Rows: 3196 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
```

```
## chr (2): Location, perc_mortality_change_ast
## dbl (6): FIPS, mortality_resp, perc_mortality_change_resp, mortality_ast, mo...
##
## 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.
```

```
#outcomes <- read_csv(file = "/Users/mac/Documents/R Assignments PLCY 715/final-team-projects-darkorchi
```

```
colnames(crossw)[1] <- "County"
colnames(crossw)[2] <- "State"
colnames(crossw)[3] <- "FIPS"

crossw <- crossw %>% select(County, State, FIPS)

crossw <- crossw %>%
  filter(County != "County Name")
```

```
crossw <- crossw %>%
  mutate(County = str_to_title(tolower(County)))
```

```
outcomes <- outcomes %>%
  separate(Location, into = c("County", "State"), sep = ", ", remove = FALSE)
```

```
## Warning: Expected 2 pieces. Missing pieces filled with `NA` in 53 rows [1, 2, 56, 75,
## 143, 303, 309, 354, 457, 550, 650, 651, 765, 900, 972, 989, 1014, 1029, 1113,
## 1201, ...].
```

```
outcomes <- outcomes %>%
  filter(!is.na(State))

outcomes$County <- str_replace(outcomes$County, " County", "")

outcomes$County <- str_replace(outcomes$County, " Parish", "")

outcomes$County <- str_replace(outcomes$County, "Saint ", "St. ")
```

```
# join AQI and Outcomes, there will be all 3000+ Outcomes visible and only 1000ish AQI present

joined_dv_iv <- left_join(outcomes, aqi, by = c("County" = "County", "State" = "State"))
view(joined_dv_iv)
```

```
joined_dv_iv <- clean_names(joined_dv_iv)

# I intend to join the countycoords tibble which has county, state, lat, long, distance_to_nearest_plan

centroidsandplants <- read_csv(file = "/Users/mac/Documents/R Assignments PLCY 715/final-team-projects-
```

```
## Rows: 3143 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): county, state, name_of_nearest_plant
```

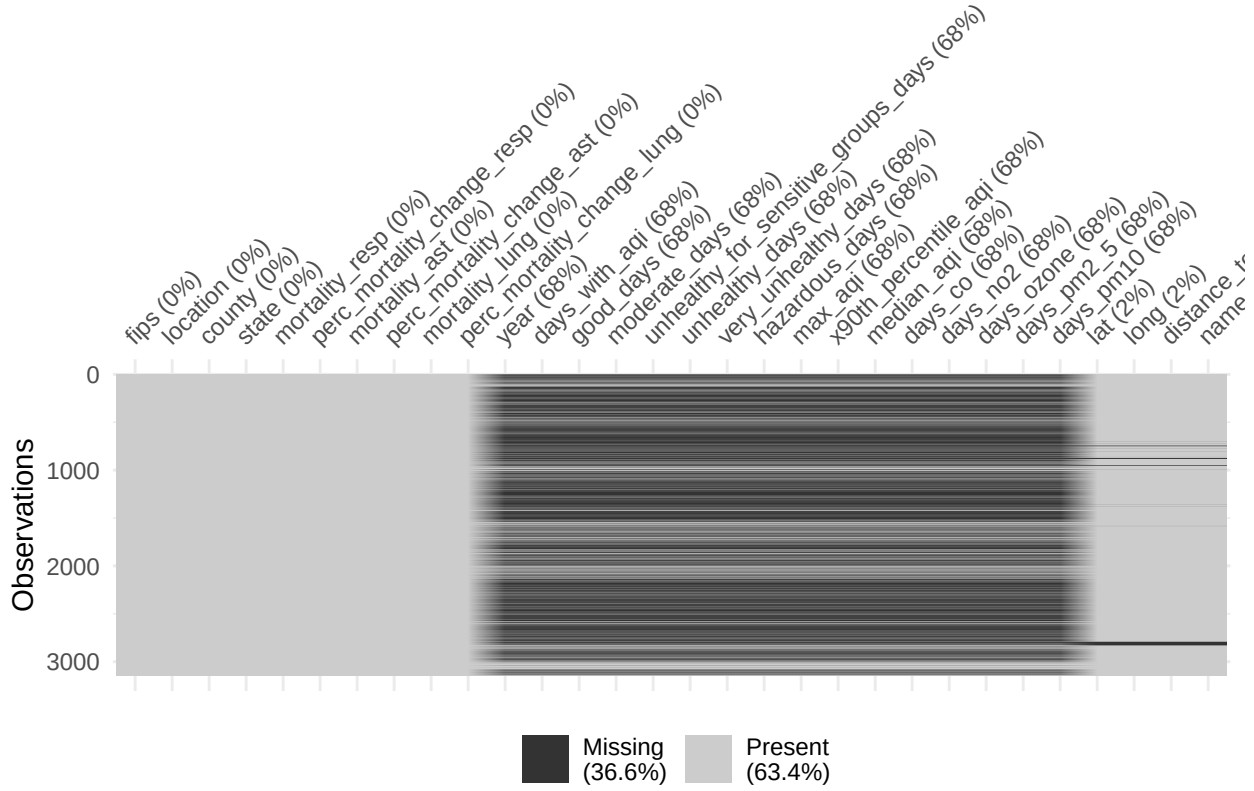
```
## dbl (3): lat, long, distance_to_nearest_plant
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
#centroidsandplants <- read_csv("/Users/mac/Documents/R Assignments PLCY 715/final-team-projects-darkor
```

```
final <- left_join(joined_dv_iv, centroidsandplants, by = c("county" = "county", "state" = "state"))
view(final)
```

```
library(naniar)

vis_miss(final)
```




```
write_csv(final, "/Users/mac/Documents/R Assignments PLCY 715/final-team-projects-darkorchid/DarkOrchid
```

```
#write_csv(final, "/Users/mac/Documents/R Assignments PLCY 715/final-team-projects-darkorchid/DarkOrchi
```

```
colnames(final)
```

```
##  [1] "fips"                          "location"
##  [3] "county"                        "state"
##  [5] "mortality_resp"                "perc_mortality_change_resp"
##  [7] "mortality_ast"                 "perc_mortality_change_ast"
```

```
##  [9] "mortality_lung"                       "perc_mortality_change_lung"
## [11] "year"                                 "days_with_aqi"
## [13] "good_days"                            "moderate_days"
## [15] "unhealthy_for_sensitive_groups_days" "unhealthy_days"
## [17] "very_unhealthy_days"                  "hazardous_days"
## [19] "max_aqi"                              "x90th_percentile_aqi"
## [21] "median_aqi"                           "days_co"
## [23] "days_no2"                             "days_ozone"
## [25] "days_pm2_5"                           "days_pm10"
## [27] "lat"                                  "long"
## [29] "distance_to_nearest_plant"            "name_of_nearest_plant"
```

```
### At this point in the process, we had a merged data set containing health outcomes, air quality, and
```

Basic Regressions

```
library(tidyverse)
library(ggplot2)
library(moderndive)
library(GGally)
library(janitor)
library(dplyr)
library(stringr)
library(tidyr)
library(naniar)
```

```
data <- read_csv(file = "/Users/mac/Documents/R Assignments PLCY 715/final-team-projects-darkorchid/Dar
```

```
## Rows: 3148 Columns: 30
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (5): location, county, state, perc_mortality_change_ast, name_of_neares...
## dbl (25): fips, mortality_resp, perc_mortality_change_resp, mortality_ast, m...
##
## 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.
```

```
head(data)
```

```
## # A tibble: 6 x 30
##    fips location    county state morta~1 perc_~2 morta~3 perc_~4 morta~5 perc_~6
##   <dbl> <chr>       <chr>  <chr>   <dbl>   <dbl>   <dbl> <chr>     <dbl>   <dbl>
## 1 10001 Kent Count~ Kent   Dela~    62      47.0    1.38 -19.07     6.3    154.
## 2 10003 New Castle~ New C~ Dela~    49.3    31.9    1.12 -24.18     5.94   164.
## 3 10005 Sussex Cou~ Sussex Dela~    50.2    14.0    0.92 -44.22     5.76   109.
## 4  1001 Autauga Co~ Autau~ Alab~    81.7    75.7    1.07 -29.51     5.72   109.
## 5  1003 Baldwin Co~ Baldw~ Alab~    54.2    46.1    0.94 -40.13     6.34   127.
## 6  1005 Barbour Co~ Barbo~ Alab~    69.8    63.0    1.63 -35.42     6.47    90.3
## # ... with 20 more variables: year <dbl>, days_with_aqi <dbl>, good_days <dbl>,
## #   moderate_days <dbl>, unhealthy_for_sensitive_groups_days <dbl>,
```

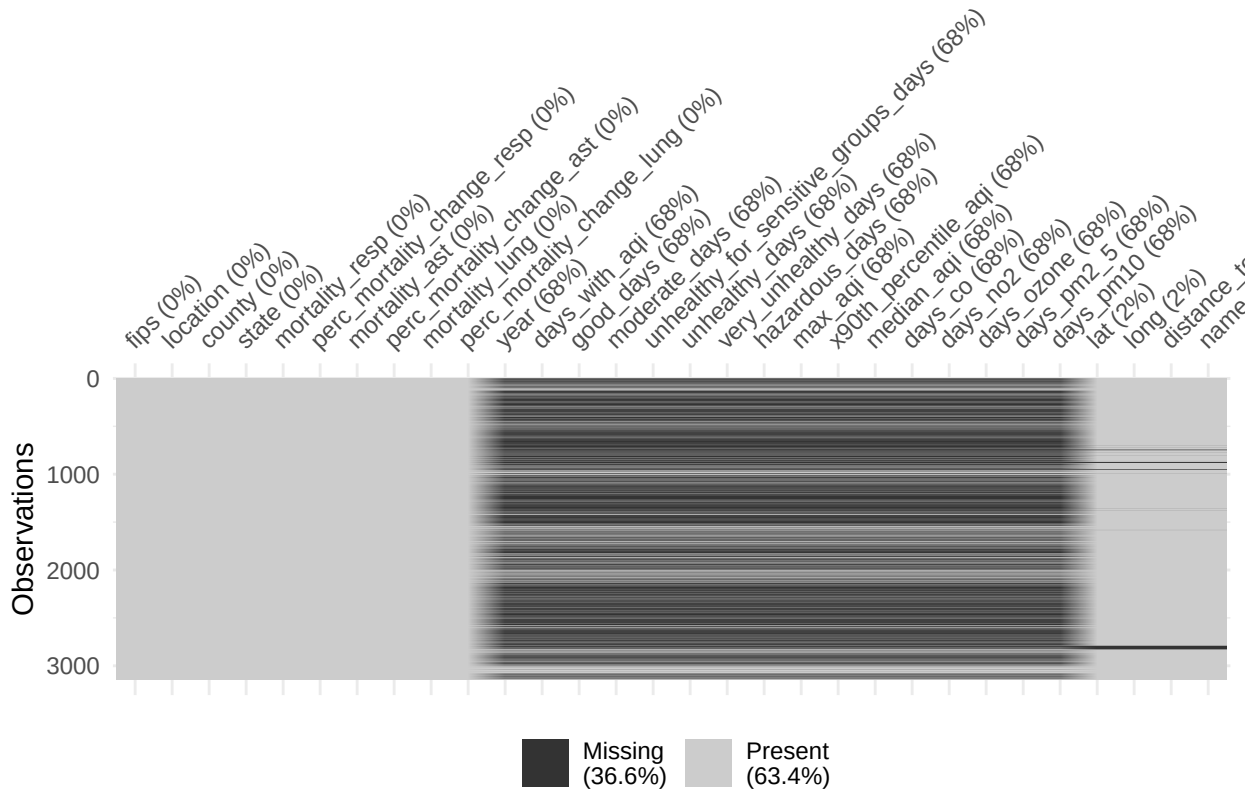
```
## #   unhealthy_days <dbl>, very_unhealthy_days <dbl>, hazardous_days <dbl>,
## #   max_aqi <dbl>, x90th_percentile_aqi <dbl>, median_aqi <dbl>, days_co <dbl>,
## #   days_no2 <dbl>, days_ozone <dbl>, days_pm2_5 <dbl>, days_pm10 <dbl>,
## #   lat <dbl>, long <dbl>, distance_to_nearest_plant <dbl>,
## #   name_of_nearest_plant <chr>, and abbreviated variable names ...
```

```
colnames(data)
```

```
##  [1] "fips"                                  "location"
##  [3] "county"                                "state"
##  [5] "mortality_resp"                        "perc_mortality_change_resp"
##  [7] "mortality_ast"                         "perc_mortality_change_ast"
##  [9] "mortality_lung"                        "perc_mortality_change_lung"
## [11] "year"                                  "days_with_aqi"
## [13] "good_days"                             "moderate_days"
## [15] "unhealthy_for_sensitive_groups_days" "unhealthy_days"
## [17] "very_unhealthy_days"                   "hazardous_days"
## [19] "max_aqi"                               "x90th_percentile_aqi"
## [21] "median_aqi"                            "days_co"
## [23] "days_no2"                              "days_ozone"
## [25] "days_pm2_5"                            "days_pm10"
## [27] "lat"                                   "long"
## [29] "distance_to_nearest_plant"             "name_of_nearest_plant"
```

```
# Checking to see which variables would create a regression using 1000+ observations and which variable

vis_miss(data)
```




```
data <- data %>%
  mutate(distance_km = distance_to_nearest_plant / 1000)

pr1model <- lm(mortality_resp ~ median_aqi + unhealthy_days + hazardous_days + days_no2 + distance_km, d
summary(pr1model)
```

```
##
## Call:
## lm(formula = mortality_resp ~ median_aqi + unhealthy_days + hazardous_days +
##     days_no2 + distance_km, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.275 -10.941  -1.345  10.120  74.523
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    69.230481   2.247085  30.809  < 2e-16 ***
## median_aqi     -0.152596   0.055923  -2.729  0.00647 **
## unhealthy_days  0.384676   0.202026   1.904  0.05719 .
## hazardous_days -1.356047   1.068328  -1.269  0.20463
## days_no2       -0.149169   0.027039  -5.517 4.41e-08 ***
## distance_km    -0.020836   0.003906  -5.334 1.19e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 15.36 on 987 degrees of freedom
##   (2155 observations deleted due to missingness)
## Multiple R-squared:  0.06376,    Adjusted R-squared:  0.05902
## F-statistic: 13.44 on 5 and 987 DF,  p-value: 1.056e-12
```

```
pr2model <- lm(mortality_resp ~ median_aqi + max_aqi + good_days + moderate_days + unhealthy_for_sensiti
summary(pr2model)
```

```
##
## Call:
## lm(formula = mortality_resp ~ median_aqi + max_aqi + good_days +
##     moderate_days + unhealthy_for_sensitive_groups_days + unhealthy_days +
##     hazardous_days + days_ozone + days_no2 + distance_km, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -46.72 -10.62  -1.33  10.32  74.12
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       67.688328   3.982717  16.996  < 2e-16 ***
## median_aqi                         0.139117   0.114824   1.212  0.22597
## max_aqi                           -0.010872   0.007524  -1.445  0.14877
## good_days                         -0.016682   0.008404  -1.985  0.04744 *
## moderate_days                     -0.044013   0.018848  -2.335  0.01973 *
## unhealthy_for_sensitive_groups_days -0.302866   0.120539  -2.513  0.01214 *
## unhealthy_days                     0.837128   0.353914   2.365  0.01821 *
## hazardous_days                     0.281765   1.473213   0.191  0.84836
## days_ozone                        -0.010470   0.006782  -1.544  0.12294
## days_no2                          -0.129245   0.027207  -4.751 2.33e-06 ***
## distance_km                       -0.013318   0.004197  -3.173  0.00156 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.17 on 982 degrees of freedom
##   (2155 observations deleted due to missingness)
## Multiple R-squared:  0.09044,    Adjusted R-squared:  0.08118
## F-statistic: 9.765 on 10 and 982 DF,  p-value: 1.097e-15
```

```
### After observing some basic regressions, we found our R-squared values to be very low, but our distan
```

Advanced Regressions

```
data <- read_csv(file = "/Users/mac/Documents/R Assignments PLCY 715/final-team-projects-darkorchid/Dark
```

```
## Rows: 3148 Columns: 30
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (5): location, county, state, perc_mortality_change_ast, name_of_neares...
```

```
## dbl (25): fips, mortality_resp, perc_mortality_change_resp, mortality_ast, m...
##
## 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.
```

```
head(data)
```

```
## # A tibble: 6 x 30
##    fips location    county state morta~1 perc_~2 morta~3 perc_~4 morta~5 perc_~6
##   <dbl> <chr>       <chr>  <chr>   <dbl>   <dbl>   <dbl> <chr>     <dbl>   <dbl>
## 1 10001 Kent Count~ Kent   Dela~    62      47.0    1.38 -19.07     6.3    154.
## 2 10003 New Castle~ New C~ Dela~    49.3    31.9    1.12 -24.18     5.94   164.
## 3 10005 Sussex Cou~ Sussex Dela~    50.2    14.0    0.92 -44.22     5.76   109.
## 4  1001 Autauga Co~ Autau~ Alab~    81.7    75.7    1.07 -29.51     5.72   109.
## 5  1003 Baldwin Co~ Baldw~ Alab~    54.2    46.1    0.94 -40.13     6.34   127.
## 6  1005 Barbour Co~ Barbo~ Alab~    69.8    63.0    1.63 -35.42     6.47    90.3
## # ... with 20 more variables: year <dbl>, days_with_aqi <dbl>, good_days <dbl>,
## #   moderate_days <dbl>, unhealthy_for_sensitive_groups_days <dbl>,
## #   unhealthy_days <dbl>, very_unhealthy_days <dbl>, hazardous_days <dbl>,
## #   max_aqi <dbl>, x90th_percentile_aqi <dbl>, median_aqi <dbl>, days_co <dbl>,
## #   days_no2 <dbl>, days_ozone <dbl>, days_pm2_5 <dbl>, days_pm10 <dbl>,
## #   lat <dbl>, long <dbl>, distance_to_nearest_plant <dbl>,
## #   name_of_nearest_plant <chr>, and abbreviated variable names ...
```

```
colnames(data)
```

```
##  [1] "fips"                                "location"
##  [3] "county"                              "state"
##  [5] "mortality_resp"                      "perc_mortality_change_resp"
##  [7] "mortality_ast"                       "perc_mortality_change_ast"
##  [9] "mortality_lung"                      "perc_mortality_change_lung"
## [11] "year"                                "days_with_aqi"
## [13] "good_days"                           "moderate_days"
## [15] "unhealthy_for_sensitive_groups_days" "unhealthy_days"
## [17] "very_unhealthy_days"                 "hazardous_days"
## [19] "max_aqi"                             "x90th_percentile_aqi"
## [21] "median_aqi"                          "days_co"
## [23] "days_no2"                            "days_ozone"
## [25] "days_pm2_5"                          "days_pm10"
## [27] "lat"                                 "long"
## [29] "distance_to_nearest_plant"           "name_of_nearest_plant"
```

```
library(dplyr)
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:purrr':
##
##     set_names
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
data <- data %>%
  mutate(distance_km = distance_to_nearest_plant / 1000)

# The file smoking_cleaned provides more independent variables for control. The crucial variable provid

smoking <- read_csv(file = "/Users/mac/Documents/R Assignments PLCY 715/final-team-projects-darkorchid/
```

```
## Rows: 3234 Columns: 4
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): county, state, geo_id
## dbl (1): current_smokers
##
## 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.
```

```
#Making the fips code identical- changed geo_id to "fips" and used sprintf to add an extra "0" in front

smoking <- smoking %>% rename(fips = geo_id)

data <- data %>%
  mutate(fips = sprintf("%05d", fips))

# Join the two data sets using the FIPS code

merged_data <- data %>%
  left_join(smoking, by = c("fips", "county"))

merged_data <- merged_data %>%
  rename(state = state.x, state_abrv = state.y)

#write_csv(merged_data, "/Users/mac/Documents/R Assignments PLCY 715/final-team-projects-darkorchid/Dar

write_csv(merged_data, "/Users/mac/Documents/R Assignments PLCY 715/final-team-projects-darkorchid/Dark

#data_s <- read_csv(file = "/Users/mac/Documents/R Assignments PLCY 715/final-team-projects-darkorchid/
```

```
data_10 <- merged_data %>%
  arrange(distance_km) %>% # sort by distance_km
  slice(1:round(n() * 0.1)) # keep top 10% of observations

data_50 <- merged_data %>%
  arrange(distance_km) %>% # sort by distance_km
  slice(1:round(n() * 0.5)) # keep top 50% of observations
```

```r
#An experimental model including man different independent variables. Looking for a high R-squared.
exmodel <- lm(mortality_resp ~ median_aqi + max_aqi + good_days + moderate_days + unhealthy_for_sensiti
summary(exmodel)
```

```
##
## Call:
## lm(formula = mortality_resp ~ median_aqi + max_aqi + good_days +
##     moderate_days + unhealthy_for_sensitive_groups_days + unhealthy_days +
##     hazardous_days + days_ozone + days_no2 + distance_km, data = merged_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -46.72 -10.62  -1.33  10.32  74.12
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         67.688328   3.982717  16.996  < 2e-16 ***
## median_aqi                           0.139117   0.114824   1.212  0.22597
## max_aqi                             -0.010872   0.007524  -1.445  0.14877
## good_days                           -0.016682   0.008404  -1.985  0.04744 *
## moderate_days                       -0.044013   0.018848  -2.335  0.01973 *
## unhealthy_for_sensitive_groups_days -0.302866   0.120539  -2.513  0.01214 *
## unhealthy_days                       0.837128   0.353914   2.365  0.01821 *
## hazardous_days                       0.281765   1.473213   0.191  0.84836
## days_ozone                          -0.010470   0.006782  -1.544  0.12294
## days_no2                            -0.129245   0.027207  -4.751 2.33e-06 ***
## distance_km                         -0.013318   0.004197  -3.173  0.00156 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.17 on 982 degrees of freedom
##   (2155 observations deleted due to missingness)
## Multiple R-squared:  0.09044,    Adjusted R-squared:  0.08118
## F-statistic: 9.765 on 10 and 982 DF,  p-value: 1.097e-15
```

```r
exmodel2 <- lm(mortality_resp ~ median_aqi + unhealthy_days + hazardous_days + days_no2 + distance_km, 
summary(exmodel2)
```

```
##
## Call:
## lm(formula = mortality_resp ~ median_aqi + unhealthy_days + hazardous_days +
##     days_no2 + distance_km, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.275 -10.941  -1.345  10.120  74.523
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    69.230481   2.247085  30.809  < 2e-16 ***
## median_aqi     -0.152596   0.055923  -2.729  0.00647 **
## unhealthy_days  0.384676   0.202026   1.904  0.05719 .
```

```
## hazardous_days -1.356047   1.068328  -1.269  0.20463
## days_no2       -0.149169   0.027039  -5.517 4.41e-08 ***
## distance_km    -0.020836   0.003906  -5.334 1.19e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.36 on 987 degrees of freedom
##   (2155 observations deleted due to missingness)
## Multiple R-squared:  0.06376,    Adjusted R-squared:  0.05902
## F-statistic: 13.44 on 5 and 987 DF,  p-value: 1.056e-12
```

```
final_mod1 <-lm(mortality_resp ~ distance_km, data = data)

summary(final_mod1)
```

```
##
## Call:
## lm(formula = mortality_resp ~ distance_km, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.369 -11.588  -1.120   9.381  95.872
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 66.786017   0.468072 142.683  < 2e-16 ***
## distance_km -0.024463   0.002978  -8.216 3.08e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.7 on 3075 degrees of freedom
##   (71 observations deleted due to missingness)
## Multiple R-squared:  0.02148,    Adjusted R-squared:  0.02116
## F-statistic: 67.49 on 1 and 3075 DF,  p-value: 3.076e-16
```

```
final_mod10 <-lm(mortality_resp ~ distance_km, data = data_10)

summary(final_mod10)
```

```
##
## Call:
## lm(formula = mortality_resp ~ distance_km, data = data_10)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.639 -11.465  -1.845   9.301  62.078
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 65.32048    2.94323  22.193   <2e-16 ***
## distance_km  0.06938    0.13804   0.503    0.616
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 16.11 on 313 degrees of freedom
## Multiple R-squared:  0.0008064,  Adjusted R-squared:  -0.002386
## F-statistic: 0.2526 on 1 and 313 DF,  p-value: 0.6156
```

```
final_mod50 <-lm(mortality_resp ~ distance_km, data = data_50)

summary(final_mod50)
```

```
##
## Call:
## lm(formula = mortality_resp ~ distance_km, data = data_50)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.856 -12.261  -1.254   9.491  82.799
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 66.007780   1.085481  60.810   <2e-16 ***
## distance_km  0.002219   0.018163   0.122    0.903
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.29 on 1572 degrees of freedom
## Multiple R-squared:  9.492e-06,  Adjusted R-squared:  -0.0006266
## F-statistic: 0.01492 on 1 and 1572 DF,  p-value: 0.9028
```

```
smoking_mod1 <-lm(mortality_resp ~ distance_km + current_smokers, data = merged_data)

summary(smoking_mod1)
```

```
##
## Call:
## lm(formula = mortality_resp ~ distance_km + current_smokers,
##     data = merged_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.109  -8.410  -0.666   7.219  65.226
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     11.634876   1.299051   8.956   <2e-16 ***
## distance_km      0.002278   0.002405   0.947    0.344
## current_smokers  2.746356   0.062073  44.244   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.05 on 3072 degrees of freedom
##   (73 observations deleted due to missingness)
## Multiple R-squared:  0.4025, Adjusted R-squared:  0.4021
## F-statistic:  1035 on 2 and 3072 DF,  p-value: < 2.2e-16
```

```r
smoking_mod2 <- lm(mortality_resp ~ distance_km + current_smokers + state, data = merged_data)

summary(smoking_mod2)
```

```
##
## Call:
## lm(formula = mortality_resp ~ distance_km + current_smokers +
##     state, data = merged_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.888  -6.983  -0.562   6.183  69.480
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  23.139672   2.168829  10.669  < 2e-16 ***
## distance_km                  -0.005859   0.003290  -1.781  0.07505 .
## current_smokers               2.671252   0.078160  34.177  < 2e-16 ***
## stateArizona                -16.922064   3.388218  -4.994 6.24e-07 ***
## stateArkansas                -8.185583   1.987844  -4.118 3.93e-05 ***
## stateCalifornia              -2.257450   2.573142  -0.877  0.38039
## stateColorado                -1.145788   2.108782  -0.543  0.58694
## stateConnecticut            -11.628453   4.453214  -2.611  0.00907 **
## stateDelaware               -12.167148   6.971932  -1.745  0.08106 .
## stateDistrict of Columbia   -26.028151  11.902898  -2.187  0.02884 *
## stateFlorida                -12.740855   2.039548  -6.247 4.77e-10 ***
## stateGeorgia                 -5.069234   1.718617  -2.950  0.00321 **
## stateHawaii                 -25.568870   5.509190  -4.641 3.61e-06 ***
## stateIdaho                   -7.850832   2.438491  -3.220  0.00130 **
## stateIllinois                -7.874267   1.870161  -4.210 2.62e-05 ***
## stateIndiana                 -7.904710   1.900353  -4.160 3.28e-05 ***
## stateIowa                   -13.456552   1.888614  -7.125 1.29e-12 ***
## stateKansas                  -5.217622   1.857228  -2.809  0.00500 **
## stateKentucky                -1.001817   1.833487  -0.546  0.58483
## stateLouisiana              -18.860901   2.073786  -9.095  < 2e-16 ***
## stateMaine                   -5.179435   3.311108  -1.564  0.11786
## stateMaryland               -11.736711   2.850574  -4.117 3.94e-05 ***
## stateMassachusetts          -11.172422   3.520873  -3.173  0.00152 **
## stateMichigan               -13.751535   1.939474  -7.090 1.66e-12 ***
## stateMinnesota              -22.633507   1.930048 -11.727  < 2e-16 ***
## stateMississippi             -6.126946   1.943123  -3.153  0.00163 **
## stateMissouri                -9.279657   1.821771  -5.094 3.73e-07 ***
## stateMontana                 -4.785093   2.240897  -2.135  0.03281 *
## stateNebraska                -4.678082   1.927268  -2.427  0.01527 *
## stateNevada                  -2.382625   3.232330  -0.737  0.46110
## stateNew Hampshire           -6.991322   4.023927  -1.737  0.08241 .
## stateNew Jersey             -13.516975   3.010006  -4.491 7.37e-06 ***
## stateNew Mexico              -5.403893   2.572671  -2.100  0.03577 *
## stateNew York                -9.800458   2.132925  -4.595 4.51e-06 ***
## stateNorth Carolina         -10.147327   1.866205  -5.437 5.84e-08 ***
## stateNorth Dakota           -20.826880   2.185212  -9.531  < 2e-16 ***
## stateOhio                   -16.273128   1.921768  -8.468  < 2e-16 ***
## stateOklahoma                -1.688226   1.973762  -0.855  0.39243
```

```
## stateOregon                -4.262456   2.545579  -1.674  0.09414 .
## statePennsylvania         -19.873189   2.046838  -9.709  < 2e-16 ***
## stateRhode Island         -13.556900   5.498498  -2.466  0.01373 *
## stateSouth Carolina        -9.812896   2.262706  -4.337 1.49e-05 ***
## stateSouth Dakota         -17.865559   2.076667  -8.603  < 2e-16 ***
## stateTennessee            -10.757498   1.899841  -5.662 1.63e-08 ***
## stateTexas                 -7.446341   1.634829  -4.555 5.45e-06 ***
## stateUtah                   0.126666   2.738915   0.046  0.96312
## stateVermont               -2.446436   3.505200  -0.698  0.48527
## stateVirginia             -13.461152   1.879799  -7.161 1.00e-12 ***
## stateWashington            -6.937234   2.448227  -2.834  0.00463 **
## stateWest Virginia         -3.611818   2.166845  -1.667  0.09565 .
## stateWisconsin            -17.464369   2.019422  -8.648  < 2e-16 ***
## stateWyoming               -0.094805   2.860749  -0.033  0.97357
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.8 on 3023 degrees of freedom
##   (73 observations deleted due to missingness)
## Multiple R-squared:  0.5193, Adjusted R-squared:  0.5112
## F-statistic: 64.03 on 51 and 3023 DF,  p-value: < 2.2e-16
```

```
smoking_mod10 <- lm(mortality_resp ~ distance_km + current_smokers, data = merged_data)

summary(smoking_mod10)
```

```
##
## Call:
## lm(formula = mortality_resp ~ distance_km + current_smokers,
##     data = merged_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.109  -8.410  -0.666   7.219  65.226
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     11.634876   1.299051   8.956   <2e-16 ***
## distance_km      0.002278   0.002405   0.947    0.344
## current_smokers  2.746356   0.062073  44.244   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.05 on 3072 degrees of freedom
##   (73 observations deleted due to missingness)
## Multiple R-squared:  0.4025, Adjusted R-squared:  0.4021
## F-statistic:  1035 on 2 and 3072 DF,  p-value: < 2.2e-16
```

```
smoking_mod50 <- lm(mortality_resp ~ distance_km + current_smokers, data = merged_data)

summary(smoking_mod50)
```

```
##
```

```
## Call:
## lm(formula = mortality_resp ~ distance_km + current_smokers,
##     data = merged_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.109  -8.410  -0.666   7.219  65.226
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     11.634876   1.299051   8.956   <2e-16 ***
## distance_km      0.002278   0.002405   0.947    0.344
## current_smokers  2.746356   0.062073  44.244   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.05 on 3072 degrees of freedom
##   (73 observations deleted due to missingness)
## Multiple R-squared:  0.4025, Adjusted R-squared:  0.4021
## F-statistic:  1035 on 2 and 3072 DF,  p-value: < 2.2e-16
```