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 ──
## ✔ ggplot2 3.4.0      ✔ purrr   1.0.1
## ✔ tibble  3.1.8      ✔ dplyr   1.0.10
## ✔ tidyr   1.3.0      ✔ stringr 1.5.0
## ✔ readr   2.1.4      ✔ forcats 0.5.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ 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)
library(geosphere)
library(readr)
```

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 worksheets were removed from the workbook.

coalplants <- read_csv("https://raw.githubusercontent.com/UNCPublicPolicy/final-team-projects-darkorchid/main/DarkOrchid/DATA/Clean/coal_plants_CLEAN")
```

```
## Rows: 1225 Columns: 37
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (23): Tracker ID, TrackerLOC, ParentID, Wiki page, Country, Subnational ...
## dbl (11): Capacity (MW), Year, RETIRED, Planned Retire, Latitude, Longitude,...
## lgl  (3): Chinese Name, Major area (prefecture, district), Permits
## 
## 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.
```

```
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, year, local_area_taluk_county, latitude, longitude )
coalplants <- rename(coalplants, county = local_area_taluk_county)
```

```
counties <- read_csv(file = "https://raw.githubusercontent.com/UNCPublicPolicy/final-team-projects-darkorchid/main/DarkOrchid/DATA/Clean/uscounties_CLEAN.csv")
```

```
## 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 × 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 ¹county_full, ²county_fips, ³state_id,
## #   ⁴state_name, ⁵population
```

```
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", "plant", "county")
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 se
lected as the most important values of interest for the rest of the project. The names w
ere changed but these are the orginal values of interest as they were known in the orign
al coal plant data set from globalcoaltracker.
```

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

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

```
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("lon
g", "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
```

```
# 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("lon
g", "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)

#file_path1 <- "https://raw.githubusercontent.com/UNCPublicPolicy/final-team-projects-da
rkorchid/main/DarkOrchid/DATA/Clean/centroids_plants_CLEAN.csv"

### A note about write_csv in this .Rmd: all write_csv commands are tagged out to avoid
problems with the online repository. The files that are written from this .Rmd were crea
ted locally and pushed into their proper place in the repository. If any user of this .R
md plans to use write_csv to write csvs into their own local pathway, please replace the
#file_pathX pathway with your own local path. ###

#file_path1 <- "https://raw.githubusercontent.com/UNCPublicPolicy/final-team-projects-da
rkorchid/main/DarkOrchid/DATA/Clean/centroids_plants_CLEAN.csv"

#write_csv(county_centroids, file_path1)
```

```
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 count
y centroid to the nearest coal fired power plant is calculated and compiled along with t
he name of the nearest plant into a tibble called countycoords.csv. This csv has record
of 3143 US counties and their distance. ###
```

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 = "https://raw.githubusercontent.com/UNCPublicPolicy/final-team-
projects-darkorchid/main/DarkOrchid/DATA/Clean/countycrosswalk_CLEAN.csv")
```

```
## 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 = "https://raw.githubusercontent.com/UNCPublicPolicy/final-team-pro
jects-darkorchid/main/DarkOrchid/DATA/Clean/aqi_2014_CLEAN.csv")
```

```
## 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 = "https://raw.githubusercontent.com/UNCPublicPolicy/final-tea
m-projects-darkorchid/main/DarkOrchid/DATA/Clean/health_outcomes_CLEAN.csv")
```

```
## Rows: 3142 Columns: 11
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (2): County, State
## dbl (9): FIPS, mortality_ast_2014, perc_mortality_change_ast, mortality_cwp_...
##
## 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.
```

```
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)))
```

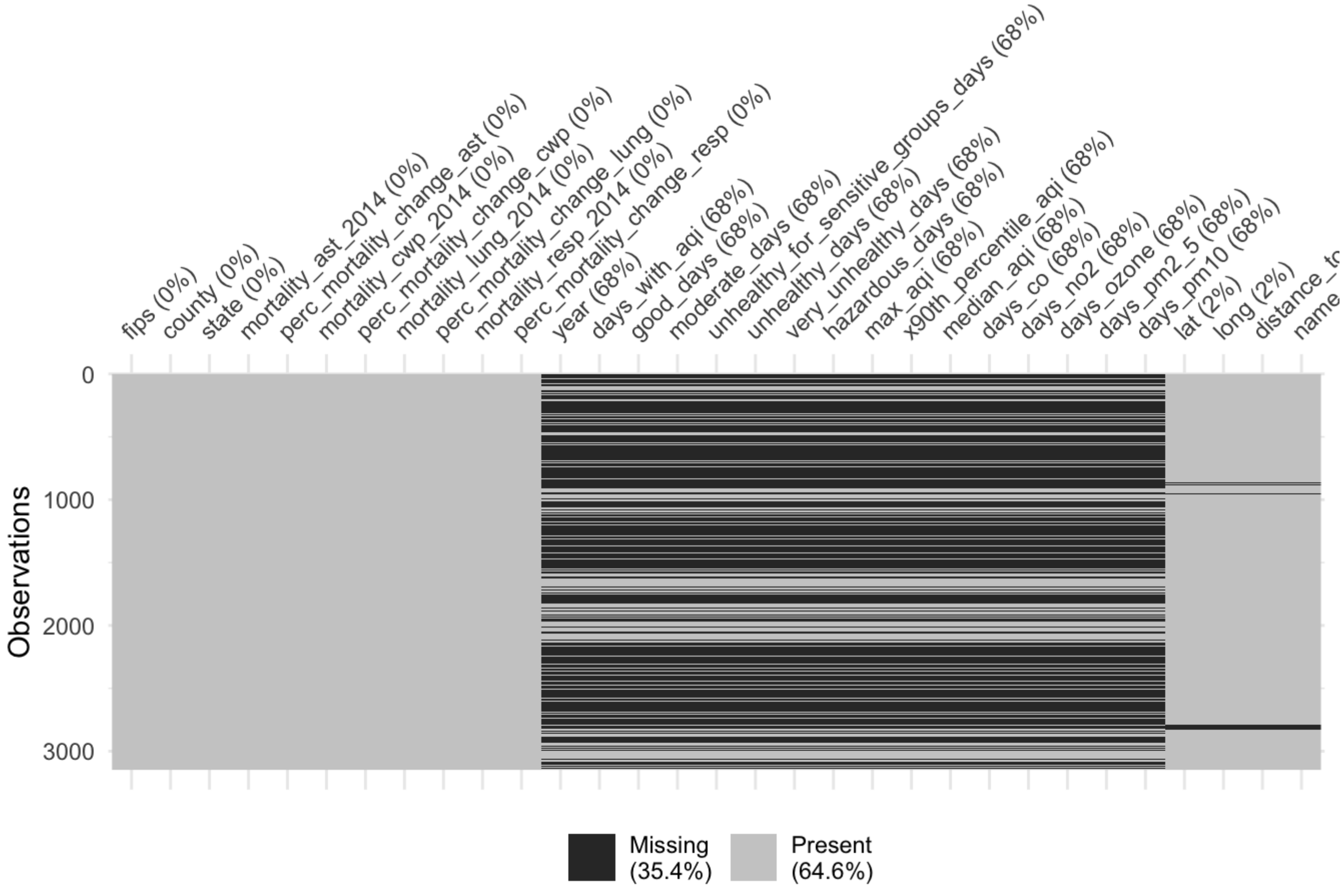
```
outcomes <- outcomes %>%
  separate(Location, into = c("County", "State"), sep = ", ", remove = FALSE)

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 p
resent

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_plant, name_of_nearest_plant

centroidsandplants <- county_centroids
```

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

```
library(naniar)

vis_miss(final)
```



```
#file_path2 <- UNCPublicPolicy/final-team-projects-darkorchid/DATA/clean/health_aqi_coal
_distance_CLEAN.csv

#write_csv(final, file_path2)
```

```
colnames(final)
```

```
##  [1] "fips"                                "county"
##  [3] "state"                               "mortality_ast_2014"
##  [5] "perc_mortality_change_ast"           "mortality_cwp_2014"
##  [7] "perc_mortality_change_cwp"           "mortality_lung_2014"
##  [9] "perc_mortality_change_lung"          "mortality_resp_2014"
## [11] "perc_mortality_change_resp"          "year"
## [13] "days_with_aqi"                       "good_days"
## [15] "moderate_days"                       "unhealthy_for_sensitive_groups_days"
## [17] "unhealthy_days"                      "very_unhealthy_days"
## [19] "hazardous_days"                      "max_aqi"
## [21] "x90th_percentile_aqi"                "median_aqi"
## [23] "days_co"                             "days_no2"
## [25] "days_ozone"                          "days_pm2_5"
## [27] "days_pm10"                           "lat"
## [29] "long"                                "distance_to_nearest_plant"
## [31] "name_of_nearest_plant"
```

```
### At this point in the process, we had a merged data set containing health outcomes, a
ir quality, and distance from operating coal plants by county for 3143 counties in the U
nited States. We established health outcomes as our dependent variables ("mortality_res
p", "perc_mortality_change_resp", "mortality_ast", "perc_mortality_change_ast", "mortali
ty_lung", "perc_mortality_change_lung"). The remaining variables were independent variab
les and controls of interest to us. We began conducting some basic regressions. ###
```

Basic Regressions

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

```
data <- final
```

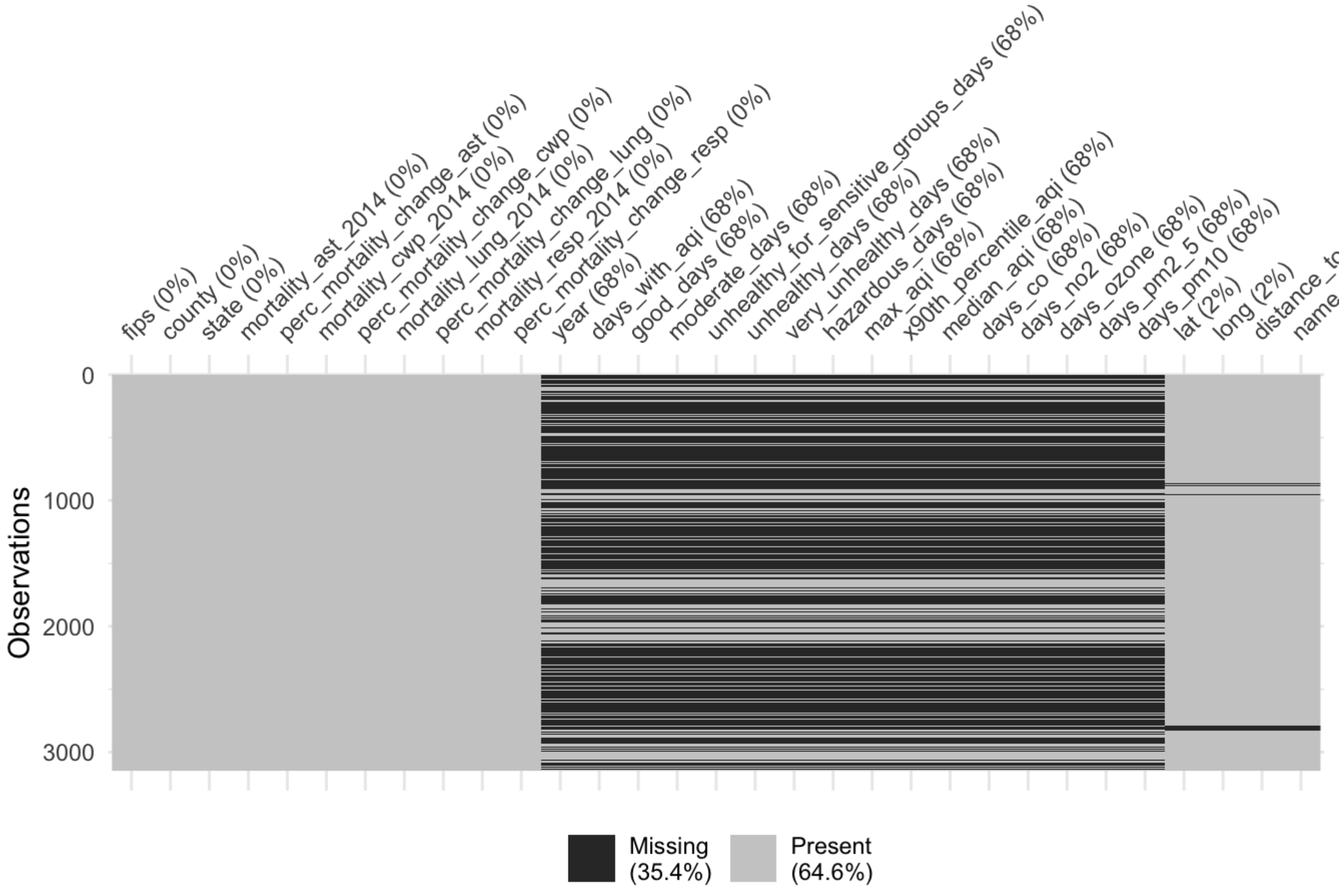
```
head(data)
```

```
## # A tibble: 6 × 31
##    fips county     state morta…¹ perc_…² morta…³ perc_…⁴ morta…⁵ perc_…⁶ morta…⁷
##   <dbl> <chr>      <chr>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 10001 Kent       Dela…    1.38   -19.1    0.02   -70.8    6.3    154.     62.0
## 2 10003 New Castle Dela…    1.12   -24.2    0.03   -79.9    5.94   164.     49.4
## 3 10005 Sussex     Dela…    0.92   -44.2    0.06   -44.6    5.76   109.     50.2
## 4  1001 Autauga    Alab…    1.07   -29.5    0.08    20.2    5.72   109.     81.8
## 5  1003 Baldwin    Alab…    0.94   -40.1    0.03   -12.4    6.34   127.     54.3
## 6  1005 Barbour    Alab…    1.63   -35.4    0.03   -19.4    6.47    90.3    69.8
## # … with 21 more variables: perc_mortality_change_resp <dbl>, 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>, …
```

```
colnames(data)
```

```
##  [1] "fips"                                "county"
##  [3] "state"                               "mortality_ast_2014"
##  [5] "perc_mortality_change_ast"           "mortality_cwp_2014"
##  [7] "perc_mortality_change_cwp"           "mortality_lung_2014"
##  [9] "perc_mortality_change_lung"          "mortality_resp_2014"
## [11] "perc_mortality_change_resp"          "year"
## [13] "days_with_aqi"                       "good_days"
## [15] "moderate_days"                       "unhealthy_for_sensitive_groups_days"
## [17] "unhealthy_days"                      "very_unhealthy_days"
## [19] "hazardous_days"                      "max_aqi"
## [21] "x90th_percentile_aqi"                "median_aqi"
## [23] "days_co"                             "days_no2"
## [25] "days_ozone"                          "days_pm2_5"
## [27] "days_pm10"                           "lat"
## [29] "long"                                "distance_to_nearest_plant"
## [31] "name_of_nearest_plant"
```

```
# Checking to see which variables would create a regression using 1000+ observations and
which variables would create a regression using 3000+ observations.

vis_miss(data)
```



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

pr1model <- lm(mortality_resp_2014 ~ median_aqi + unhealthy_days + hazardous_days + days
_no2 + distance_km, data = data)
summary(pr1model)
```

```
##
## Call:
## lm(formula = mortality_resp_2014 ~ median_aqi + unhealthy_days +
##     hazardous_days + days_no2 + distance_km, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.248 -10.947  -1.335  10.136  74.774
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    69.29949    2.24906  30.813  < 2e-16 ***
## median_aqi     -0.15308    0.05597  -2.735  0.00635 **
## unhealthy_days  0.38550    0.20220   1.906  0.05688 .
## hazardous_days -1.35431    1.06926  -1.267  0.20560
## days_no2       -0.14915    0.02706  -5.511 4.54e-08 ***
## distance_km    -0.02084    0.00391  -5.331 1.21e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.37 on 987 degrees of freedom
##   (2155 observations deleted due to missingness)
## Multiple R-squared:  0.0637, Adjusted R-squared:  0.05896
## F-statistic: 13.43 on 5 and 987 DF,  p-value: 1.089e-12
```

```
pr2model <- lm(mortality_resp_2014 ~ 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)
summary(pr2model)
```

```
## 
## Call:
## lm(formula = mortality_resp_2014 ~ 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 
## -46.679 -10.600  -1.318  10.324  74.367 
## 
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                         67.733488   3.986348  16.991  < 2e-16 ***
## median_aqi                           0.139036   0.114929   1.210  0.22667    
## max_aqi                             -0.010888   0.007531  -1.446  0.14857    
## good_days                           -0.016600   0.008412  -1.973  0.04874 *  
## moderate_days                       -0.044013   0.018865  -2.333  0.01985 *  
## unhealthy_for_sensitive_groups_days -0.303088   0.120649  -2.512  0.01216 *  
## unhealthy_days                       0.838586   0.354236   2.367  0.01811 *  
## hazardous_days                       0.285364   1.474556   0.194  0.84659    
## days_ozone                          -0.010523   0.006788  -1.550  0.12141    
## days_no2                            -0.129272   0.027231  -4.747 2.37e-06 ***
## distance_km                         -0.013324   0.004201  -3.171  0.00156 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 15.19 on 982 degrees of freedom
##   (2155 observations deleted due to missingness)
## Multiple R-squared:  0.09032,    Adjusted R-squared:  0.08106 
## F-statistic:  9.75 on 10 and 982 DF,  p-value: 1.165e-15
```

```
### After observing some basic regressions, we found our R-squared values to be very lo
w, but our distance variable to be significant at at least the 0.01 level frequently. Th
is lead us to believe that our regressions accounted for very little of the explanation
for health outcomes across US counties, but distance still was an important contributor.
Our regressions advanced to include smoking data, control by county, look at the closest
10 percent of county centroids to plants, and look at the top 50 percent of closest coun
ty centroids to plants. ###
```

Advanced Regressions

```
data <- final
```

```
head(data)
```

```
## # A tibble: 6 × 31
##    fips county     state morta…¹ perc_…² morta…³ perc_…⁴ morta…⁵ perc_…⁶ morta…⁷
##   <dbl> <chr>      <chr>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 10001 Kent       Dela…    1.38   -19.1    0.02   -70.8    6.3    154.     62.0
## 2 10003 New Castle Dela…    1.12   -24.2    0.03   -79.9    5.94   164.     49.4
## 3 10005 Sussex     Dela…    0.92   -44.2    0.06   -44.6    5.76   109.     50.2
## 4  1001 Autauga    Alab…    1.07   -29.5    0.08    20.2    5.72   109.     81.8
## 5  1003 Baldwin    Alab…    0.94   -40.1    0.03   -12.4    6.34   127.     54.3
## 6  1005 Barbour    Alab…    1.63   -35.4    0.03   -19.4    6.47    90.3    69.8
## # … with 21 more variables: perc_mortality_change_resp <dbl>, 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>, …
```

```
colnames(data)
```

```
##  [1] "fips"                                "county"
##  [3] "state"                               "mortality_ast_2014"
##  [5] "perc_mortality_change_ast"           "mortality_cwp_2014"
##  [7] "perc_mortality_change_cwp"           "mortality_lung_2014"
##  [9] "perc_mortality_change_lung"          "mortality_resp_2014"
## [11] "perc_mortality_change_resp"          "year"
## [13] "days_with_aqi"                       "good_days"
## [15] "moderate_days"                       "unhealthy_for_sensitive_groups_days"
## [17] "unhealthy_days"                      "very_unhealthy_days"
## [19] "hazardous_days"                      "max_aqi"
## [21] "x90th_percentile_aqi"                "median_aqi"
## [23] "days_co"                             "days_no2"
## [25] "days_ozone"                          "days_pm2_5"
## [27] "days_pm10"                           "lat"
## [29] "long"                                "distance_to_nearest_plant"
## [31] "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_CLEAN provides more independent variables for control. The crucial va
riable provided by this data set is the crude percent of adults who currently smoke in 2
020.

smoking <- read_csv(file = "https://raw.githubusercontent.com/UNCPublicPolicy/final-team
-projects-darkorchid/main/DarkOrchid/DATA/Clean/smoking_CLEAN.csv")
```

```
## 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 ext
ra "0" in front of codes that were 4 digits long isntead of 5. This made the fips codes
compatible between the smoking data and the rest of the compiled data.

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)

#file_path3 <- UNCPublicPolicy/final-team-projects-darkorchid/DATA/clean/health_aqi_coal
_distance_smoking_CLEAN.csv

#write_csv(merged_data, file_path3)
```

```
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
```

```
#An experimental model including many different independent variables. Investigating rea
ltionship for mortality_resp.
exmodel <- lm(mortality_resp_2014 ~ median_aqi + max_aqi + good_days + moderate_days + u
nhealthy_for_sensitive_groups_days + unhealthy_days + hazardous_days +days_ozone + days_
no2 + distance_km, data = merged_data)
summary(exmodel)
```

```
## 
## Call:
## lm(formula = mortality_resp_2014 ~ 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 
## -46.679 -10.600  -1.318  10.324  74.367 
## 
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                         67.733488   3.986348  16.991  < 2e-16 ***
## median_aqi                           0.139036   0.114929   1.210  0.22667    
## max_aqi                             -0.010888   0.007531  -1.446  0.14857    
## good_days                           -0.016600   0.008412  -1.973  0.04874 *  
## moderate_days                       -0.044013   0.018865  -2.333  0.01985 *  
## unhealthy_for_sensitive_groups_days -0.303088   0.120649  -2.512  0.01216 *  
## unhealthy_days                       0.838586   0.354236   2.367  0.01811 *  
## hazardous_days                       0.285364   1.474556   0.194  0.84659    
## days_ozone                          -0.010523   0.006788  -1.550  0.12141    
## days_no2                            -0.129272   0.027231  -4.747 2.37e-06 ***
## distance_km                         -0.013324   0.004201  -3.171  0.00156 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 15.19 on 982 degrees of freedom
##   (2155 observations deleted due to missingness)
## Multiple R-squared:  0.09032,    Adjusted R-squared:  0.08106 
## F-statistic:  9.75 on 10 and 982 DF,  p-value: 1.165e-15
```

```
# Experiments with mortality_resp models.
exmodel2 <- lm(mortality_resp_2014 ~ median_aqi + unhealthy_days + hazardous_days + days_no2 + distance_km, data = merged_data)
summary(exmodel2)
```

```
## 
## Call:
## lm(formula = mortality_resp_2014 ~ median_aqi + unhealthy_days + 
##     hazardous_days + days_no2 + distance_km, data = merged_data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -46.248 -10.947  -1.335  10.136  74.774 
## 
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)    69.29949    2.24906  30.813  < 2e-16 ***
## median_aqi     -0.15308    0.05597  -2.735  0.00635 ** 
## unhealthy_days  0.38550    0.20220   1.906  0.05688 .  
## hazardous_days -1.35431    1.06926  -1.267  0.20560    
## days_no2       -0.14915    0.02706  -5.511 4.54e-08 ***
## distance_km    -0.02084    0.00391  -5.331 1.21e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 15.37 on 987 degrees of freedom
##   (2155 observations deleted due to missingness)
## Multiple R-squared:  0.0637, Adjusted R-squared:  0.05896 
## F-statistic: 13.43 on 5 and 987 DF,  p-value: 1.089e-12
```

```
final_mod1 <-lm(mortality_resp_2014 ~ distance_km, data = merged_data)

summary(final_mod1)
```

```
##
## Call:
## lm(formula = mortality_resp_2014 ~ distance_km, data = merged_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.356 -11.589  -1.125   9.371  96.714
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 66.847970   0.469114  142.50  < 2e-16 ***
## distance_km -0.024500   0.002984   -8.21 3.22e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.74 on 3075 degrees of freedom
##   (71 observations deleted due to missingness)
## Multiple R-squared:  0.02145,    Adjusted R-squared:  0.02113
## F-statistic:  67.4 on 1 and 3075 DF,  p-value: 3.224e-16
```

```
# A deeper look into distance_km's effect on mortality_resp. Only analyzing the top 10 p
ercent lowest values for min_distance from caol fired power plant.

final_mod10 <-lm(mortality_resp_2014 ~ distance_km, data = data_10)

summary(final_mod10)
```

```
##
## Call:
## lm(formula = mortality_resp_2014 ~ distance_km, data = data_10)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.699 -11.464  -1.845   9.298  62.017
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  65.3625     2.9501  22.156   <2e-16 ***
## distance_km   0.0702     0.1384   0.507    0.612
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.15 on 313 degrees of freedom
## Multiple R-squared:  0.0008217,  Adjusted R-squared:  -0.002371
## F-statistic: 0.2574 on 1 and 313 DF,  p-value: 0.6123
```

```
# A deeper look into distance_km's effect on mortality_resp. Only analyzing the top 50 p
ercent lowest values for min_distance from caol fired power plant.

final_mod50 <-lm(mortality_resp_2014 ~ distance_km, data = data_50)

summary(final_mod50)
```

```
## 
## Call:
## lm(formula = mortality_resp_2014 ~ distance_km, data = data_50)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -39.912 -12.295  -1.279   9.459  83.631 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 66.059682   1.088449  60.692   <2e-16 ***
## distance_km  0.002394   0.018213   0.131    0.895    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 17.33 on 1572 degrees of freedom
## Multiple R-squared:  1.099e-05,  Adjusted R-squared:  -0.0006251 
## F-statistic: 0.01727 on 1 and 1572 DF,  p-value: 0.8955
```

```
# Investigating the strength of current_smokers effect on moratlity_resp compared to dis
tance_km.

smoking_mod1 <-lm(mortality_resp_2014 ~ distance_km + current_smokers, data = merged_dat
a)

summary(smoking_mod1)
```

```
##
## Call:
## lm(formula = mortality_resp_2014 ~ distance_km + current_smokers,
##     data = merged_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.202  -8.425  -0.664   7.185  65.936
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     11.594778   1.302267   8.904   <2e-16 ***
## distance_km      0.002291   0.002411   0.950    0.342
## current_smokers  2.751439   0.062226  44.217   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.08 on 3072 degrees of freedom
##   (73 observations deleted due to missingness)
## Multiple R-squared:  0.4022, Adjusted R-squared:  0.4018
## F-statistic:  1033 on 2 and 3072 DF,  p-value: < 2.2e-16
```

```
# Investigating the strength of current_smokers effect on moratlity_resp compared to distance_km controlling by state.

smoking_mod2 <- lm(mortality_resp_2014 ~ distance_km + current_smokers + state, data = merged_data)

summary(smoking_mod2)
```

```
## 
## Call:
## lm(formula = mortality_resp_2014 ~ distance_km + current_smokers + 
##     state, data = merged_data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -43.956  -6.982  -0.566   6.152  70.177 
## 
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                23.127413   2.174564  10.635  < 2e-16 ***
## distance_km                -0.005851   0.003299  -1.774  0.07622 .  
## current_smokers             2.676781   0.078366  34.157  < 2e-16 ***
## stateArizona              -16.952659   3.397177  -4.990 6.37e-07 ***
## stateArkansas              -8.243678   1.993100  -4.136 3.63e-05 ***
## stateCalifornia            -2.273168   2.579946  -0.881  0.37834    
## stateColorado              -1.141197   2.114359  -0.540  0.58942    
## stateConnecticut          -11.644694   4.464990  -2.608  0.00915 ** 
## stateDelaware             -12.198049   6.990368  -1.745  0.08109 .  
## stateDistrict of Columbia -26.075415  11.934372  -2.185  0.02897 *  
## stateFlorida              -12.772506   2.044941  -6.246 4.80e-10 ***
## stateGeorgia               -5.114836   1.723162  -2.968  0.00302 ** 
## stateHawaii               -25.571134   5.523757  -4.629 3.82e-06 ***
## stateIdaho                 -7.887214   2.444939  -3.226  0.00127 ** 
## stateIllinois              -7.918500   1.875106  -4.223 2.48e-05 ***
## stateIndiana               -7.960361   1.905378  -4.178 3.03e-05 ***
## stateIowa                 -13.494018   1.893608  -7.126 1.29e-12 ***
## stateKansas                -5.260468   1.862139  -2.825  0.00476 ** 
## stateKentucky              -0.975838   1.838335  -0.531  0.59558    
## stateLouisiana            -18.932881   2.079270  -9.106  < 2e-16 ***
## stateMaine                 -5.217212   3.319863  -1.572  0.11617    
## stateMaryland             -11.761250   2.858112  -4.115 3.97e-05 ***
## stateMassachusetts        -11.181565   3.530183  -3.167  0.00155 ** 
## stateMichigan             -13.799827   1.944603  -7.096 1.59e-12 ***
## stateMinnesota            -22.676775   1.935151 -11.718  < 2e-16 ***
## stateMississippi           -6.177585   1.948261  -3.171  0.00154 ** 
## stateMissouri              -9.310938   1.826588  -5.097 3.65e-07 ***
## stateMontana               -4.825523   2.246822  -2.148  0.03182 *  
## stateNebraska              -4.707574   1.932364  -2.436  0.01490 *  
## stateNevada                -2.434837   3.240877  -0.751  0.45254    
## stateNew Hampshire         -7.022344   4.034568  -1.741  0.08187 .  
## stateNew Jersey           -13.530726   3.017965  -4.483 7.62e-06 ***
## stateNew Mexico            -5.442153   2.579473  -2.110  0.03496 *  
## stateNew York              -9.834193   2.138565  -4.599 4.43e-06 ***
## stateNorth Carolina       -10.198324   1.871140  -5.450 5.43e-08 ***
## stateNorth Dakota         -20.870787   2.190990  -9.526  < 2e-16 ***
## stateOhio                 -16.338561   1.926850  -8.479  < 2e-16 ***
## stateOklahoma              -1.726730   1.978982  -0.873  0.38299    
## stateOregon                -4.288845   2.552311  -1.680  0.09299 .  
## statePennsylvania         -19.921520   2.052251  -9.707  < 2e-16 ***
## stateRhode Island         -13.552588   5.513038  -2.458  0.01402 *  
## stateSouth Carolina        -9.873325   2.268689  -4.352 1.39e-05 ***
## stateSouth Dakota         -17.914424   2.082158  -8.604  < 2e-16 ***
## stateTennessee            -10.800017   1.904865  -5.670 1.57e-08 ***
## stateTexas                 -7.479661   1.639152  -4.563 5.24e-06 ***
## stateUtah                   0.133755   2.746157   0.049  0.96116
## stateVermont               -2.481288   3.514469  -0.706  0.48023
## stateVirginia             -13.486502   1.884770  -7.156 1.04e-12 ***
## stateWashington            -6.964322   2.454701  -2.837  0.00458 **
## stateWest Virginia         -3.625999   2.172575  -1.669  0.09522 .
## stateWisconsin            -17.506569   2.024761  -8.646  < 2e-16 ***
## stateWyoming               -0.131681   2.868314  -0.046  0.96339
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.83 on 3023 degrees of freedom
##   (73 observations deleted due to missingness)
## Multiple R-squared:  0.5189, Adjusted R-squared:  0.5108
## F-statistic: 63.92 on 51 and 3023 DF,  p-value: < 2.2e-16
```

```
# Investigating the strength of current_smokers effect on moratlity_resp compared to distance_km. Only analyzing the top 10 percent lowest values for min_distance from coal fired power plant.
smoking_mod10 <- lm(mortality_resp_2014 ~ distance_km + current_smokers, data = data_10)

summary(smoking_mod10)
```

```
##
## Call:
## lm(formula = mortality_resp_2014 ~ distance_km + current_smokers,
##     data = data_10)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.962  -8.209  -0.414   6.511  44.822
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.39946    4.40952   0.771    0.441
## distance_km      0.05961    0.10224   0.583    0.560
## current_smokers  3.21430    0.19884  16.165   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.93 on 312 degrees of freedom
## Multiple R-squared:  0.4562, Adjusted R-squared:  0.4528
## F-statistic: 130.9 on 2 and 312 DF,  p-value: < 2.2e-16
```

```
# Investigating the strength of current_smokers effect on moratlity_resp compared to dis
tance_km. Only analyzing the top 50 percent lowest values for min_distance from coal fir
ed power plant.
smoking_mod50 <- lm(mortality_resp_2014 ~ distance_km + current_smokers, data = data_50)

summary(smoking_mod50)
```

```
##
## Call:
## lm(formula = mortality_resp_2014 ~ distance_km + current_smokers,
##     data = data_50)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.638  -8.375  -0.685   6.688  63.709
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       6.01602    1.78914   3.363 0.000791 ***
## distance_km      -0.01346    0.01326  -1.016 0.309888
## current_smokers   3.10980    0.08310  37.423  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.61 on 1570 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.4715, Adjusted R-squared:  0.4708
## F-statistic: 700.3 on 2 and 1570 DF,  p-value: < 2.2e-16
```