UTSA Environmental Factors

Contents

Texas Life Expectancy Data

This analysis will model life expectancy in Texas counties based on demographic and environmental factors.

Exploratory Data Analysis: Demographic Data

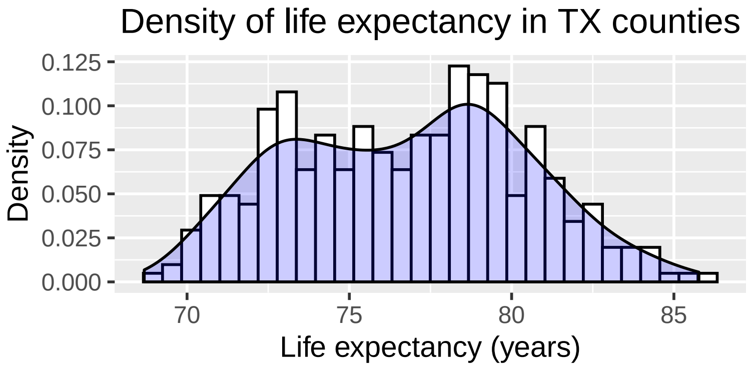
Demographic factors in this analysis include:

- County: the name of the county in TX
- pop_density: the population density per square mile in the specified county as recorded for 2010 by texaxcounties.net
- pop_2010: the total population in the specified county as recorded for 2010 by texaxcounties.net
- Area: the land area of the specified county in square miles as recorded by texaxcounties.net
- num_sf: the number of superfund sites according to the EPA
- gender: a binary male / female value corresponding to the life expectancy in the specified county as recorded by Texas Health Maps
- life_exp: life expectancy in the specified county as recorded by Texas Health Maps

sample:

County	pop_density	pop_2010	Area	num_sf	gender	life_exp
Anderson	55.0	58458	1062.6	0	female	77.34
Anderson	55.0	58458	1062.6	0	male	71.33
Andrews	9.9	14786	1500.7	0	female	78.56
Andrews	9.9	14786	1500.7	0	male	72.74
Angelina	108.8	86771	797.8	0	female	78.70
Angelina	108.8	86771	797.8	0	male	73.00

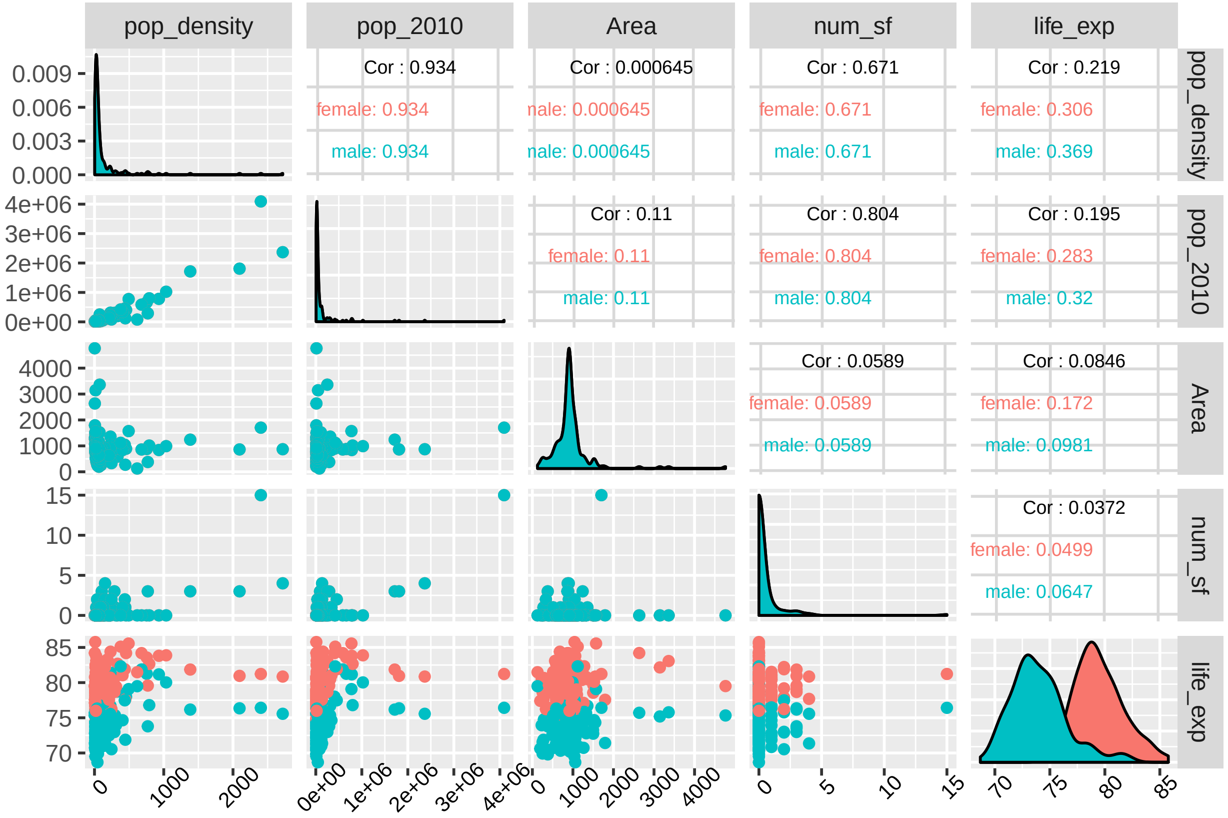
Distribution of life expectancy in TX:



Does the data appear bimodal? What might cause this?

Knowledge Check

Evaluate the following pair plot to better understand demographics .



Knowledge Check

- Which variable(s) differ by gender?
  - life expectancy
- What can you infer about each variable's distribution?
  - Most of the variables are right-skewed. For example, most of the counties' populations are small, and a small number of counties have large numbers of people.
- What relationships do you see between variables? Do any variables have a strong linear correlation with life expectancy?
  - We see a strong linear correlation between population and population density. This tells us the data might be redundant, and we may want to choose just one to be in the model.
  - No other variables have a strong linear correlation with life expectancy. As a data scientist, we might see this and consider either transforming the data, or using an approach other than linear regression.
- Do you see any outliers?
  - There seems to be a data point indicating 15 superfund sites in one county.

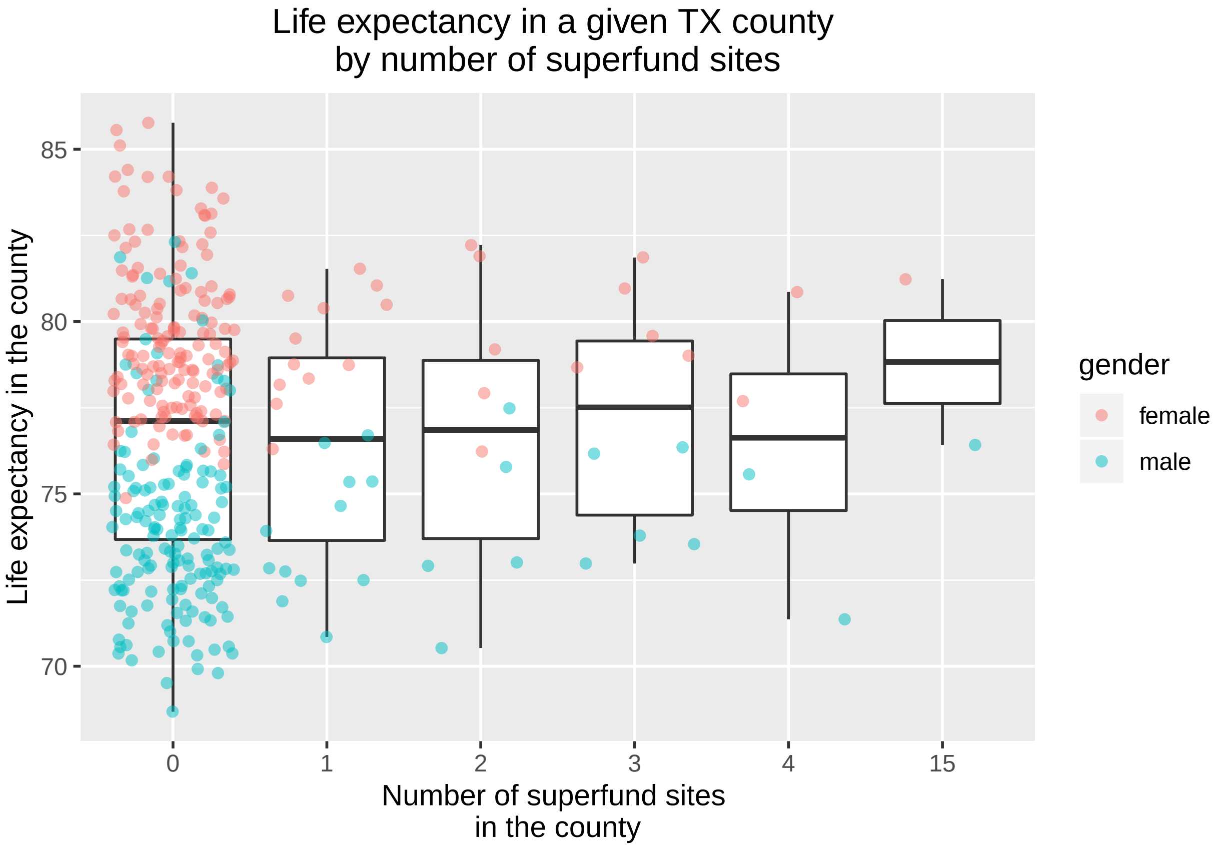
We see one county with many superfund sites. Which county is this?

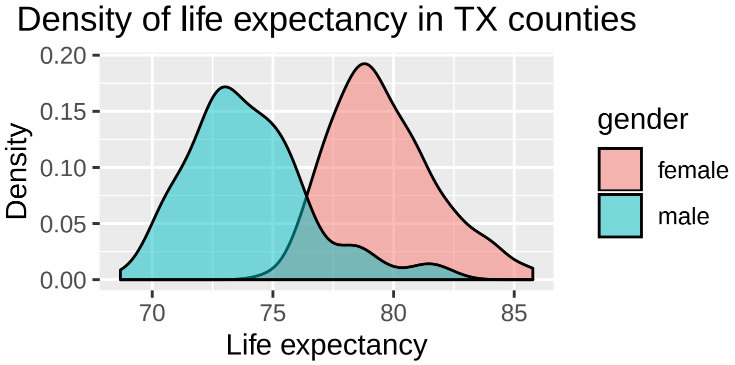
County	pop_density	pop_2010	Area	num_sf	gender	life_exp
Harris	2402.4	4092459	1703.5	15	female	81.23
Harris	2402.4	4092459	1703.5	15	male	76.42

We see that the outlier with 15 superfund sites is Harris County, the county containing Houston, TX.

This seems like a relatively high number of superfund sites. What type of life expectancy would you expect relative to other counties?

Now explore these plots about the relationships between superfund sites, gender, and life expectancy.





Knowledge Check

- Does it seem that counties with more superfund sites are associated with lower life expectancy?
  - We don't see a big difference in median life expectancy across superfund sites. In fact, the highest median values for life expectancy are from counties with 3 superfund sites and one county (Harris) with 15 superfund sites. We should also keep in mind that there are fewer samples in the groups with more superfund sites, and look at qualitative sources or do hypothesis testing with cities outside of Houston to be more sure.
- Does gender seem to have an affect on life expectancy?
  - Yes! There is a bimodal distribution due to gender difference. Females are expected to live longer across TX counties.

Let's take a look at counties with the best and worst life expectancy.

Counties with the *highest* life expectancy:

For females:

County	pop_density	pop_2010	Area	num_sf	gender	life_exp	rank
Live Oak	11.1	11531	1039.7	0	female	85.77	1
Hidalgo	493.2	774769	1570.9	0	female	85.56	2
Williamson	378.0	422679	1118.3	0	female	85.11	3
Hays	231.7	157107	678.0	0	female	84.40	4
Reeves	5.2	13783	2635.4	0	female	84.21	5
Cameron	456.0	406220	890.9	0	female	84.20	6

For males:

County	pop_density	pop_2010	Area	num_sf	gender	life_exp	rank
Williamson	378.0	422679	1118.3	0	male	82.31	1
Fort Bend	679.5	585375	861.5	0	male	81.87	2
Hays	231.7	157107	678.0	0	male	81.40	3
Denton	754.3	662614	878.4	0	male	81.26	4
Collin	930.0	782341	841.2	0	male	81.17	5
Travis	1034.4	1024266	990.2	0	male	80.03	6

Counties with the *lowest* life expectancy:

For females:

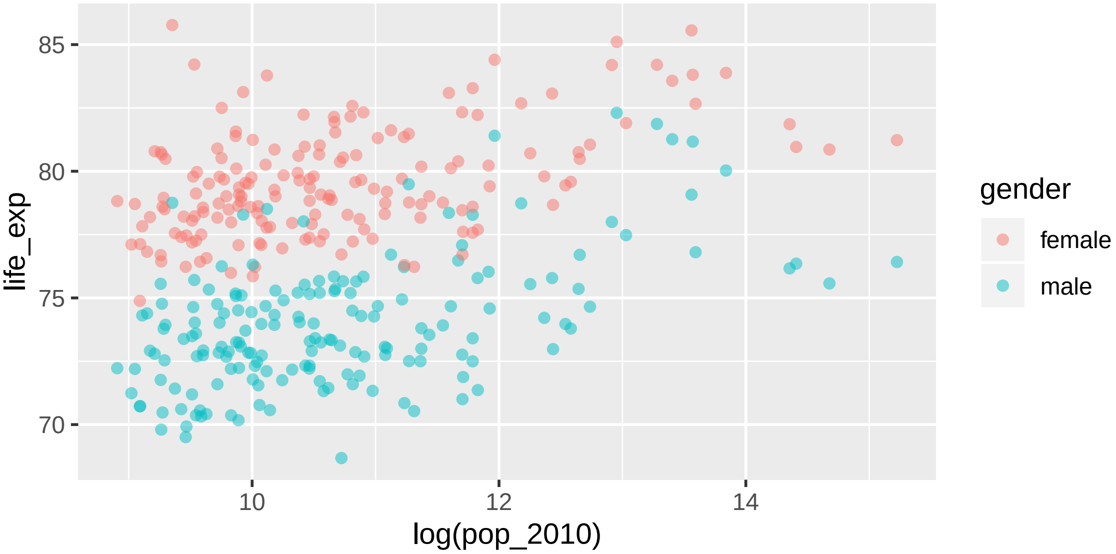
County	pop_density	pop_2010	Area	num_sf	gender	life_exp	rank
Orange	245.3	81837	333.7	2	female	76.23	154
Red River	12.4	12860	1036.6	0	female	76.23	154
Gray	24.3	22535	926.0	0	female	76.22	155
Young	20.3	18550	914.5	0	female	75.99	156
Hutchinson	25.0	22150	887.4	0	female	75.86	157
San Augustine	16.7	8865	530.7	0	female	74.88	158

For males:

County	pop_density	pop_2010	Area	num_sf	gender	life_exp	rank
Trinity	21.0	14585	693.6	0	male	70.32	151
Montague	21.2	19719	930.9	0	male	70.17	152
Morris	51.3	12934	252.0	0	male	69.92	153
Marion	27.7	10546	380.9	0	male	69.80	154
Red River	12.4	12860	1036.6	0	male	69.51	155
Polk	43.0	45413	1057.1	0	male	68.68	156

Data transformation

Recall that there was not a strong correlation between any of the variables with life expectancy. To get a better linear regression model, we'll try transforming the data to look at the log of population instead of the population itself:



Since this looks much more linear, we'll use these variables.

log_pop	Area	num_sf	gender	life_exp
10.976064	1062.6	0	female	77.34
10.976064	1062.6	0	male	71.33
9.601436	1500.7	0	female	78.56
9.601436	1500.7	0	male	72.74
11.371028	797.8	0	female	78.70
11.371028	797.8	0	male	73.00

Modeling demographic data with linear regression

Now that we have a better understanding of the data, we'll create a simple initial model:

```
reg1 <- lm(data= tx_reg_data, life_exp ~ .)
summary(reg1)
```

```
##
## Call:
## lm(formula = life_exp ~ ., data = tx_reg_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5470 -1.4896 -0.1719  1.2728  7.5157
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 67.1428415  1.0390280  64.621  < 2e-16 ***
## log_pop      1.1303404  0.0963213  11.735  < 2e-16 ***
## Area         0.0005190  0.0002177   2.384   0.0177 *
## num_sf      -0.4198741  0.0902899  -4.650 4.75e-06 ***
## gendermale  -5.5857225  0.2128990 -26.236  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.98 on 341 degrees of freedom
## Multiple R-squared:  0.7102, Adjusted R-squared:  0.7068
## F-statistic: 208.9 on 4 and 341 DF,  p-value: < 2.2e-16
```

Knowledge Check

- Which variable(s) are significant to the model?
  - We see a definite significant effect of the transformed population variable, number of superfund sites, and gender on life expectancy. We see a slightly lesser, but still evident effect of Area on life expectancy.
- What is the R Squared value of the model?
  - 71%
- Given a male and female live in the same county in TX, how much longer would you expect the female to live than the male, according to this model?
  - roughly 5.6 years, as the coefficient for gendermale is -5.58 (years of life expectancy).

Exploratory data analysis: Environmental Factors

Let's consider environmental factors.

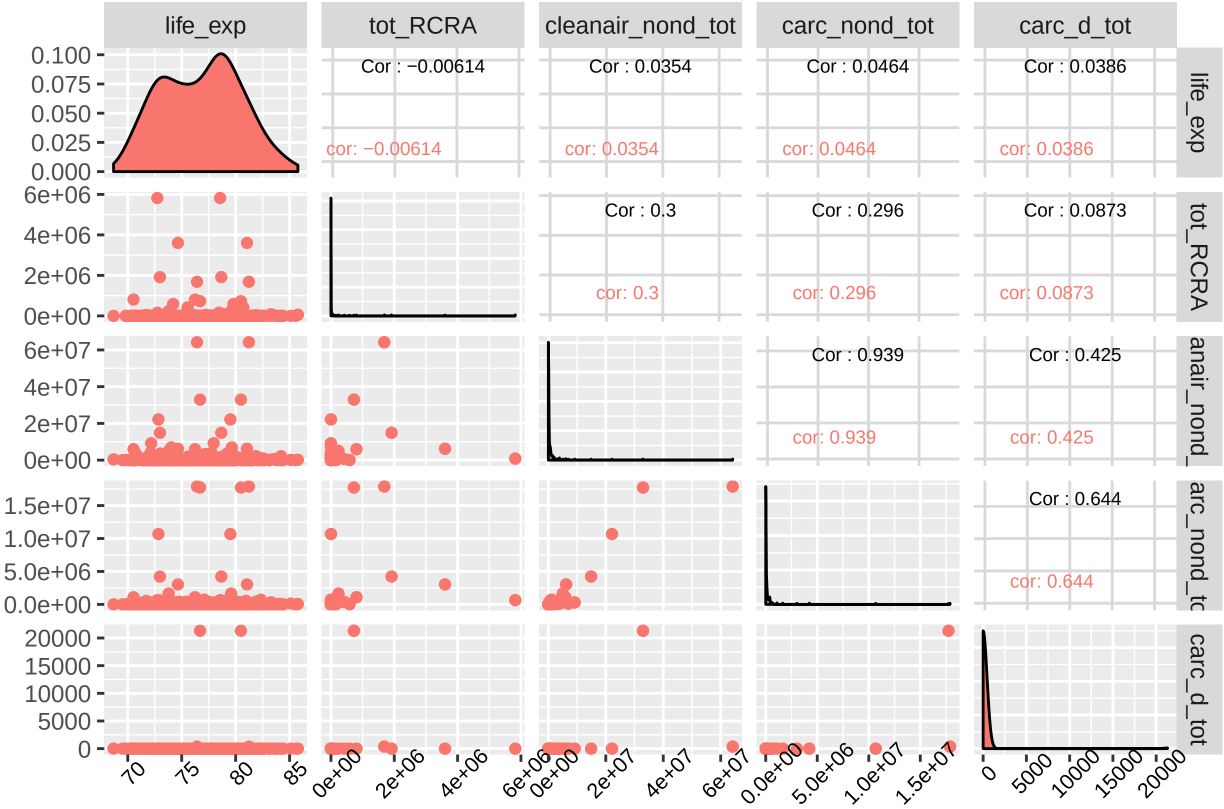
Here we will look at:

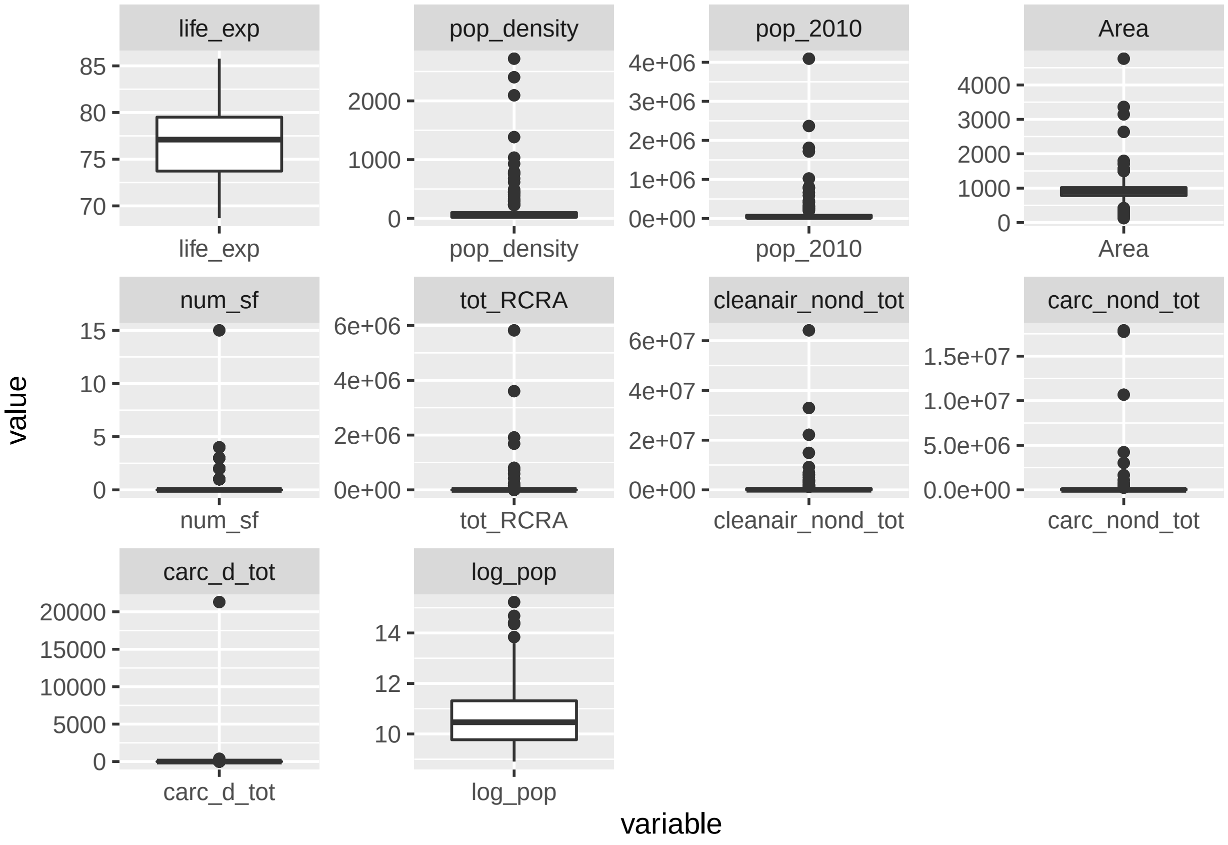
- life_exp: life expectancy in the specified county as recorded by Texas Health Maps
- tot_RCRA: total quantity of chemicals released, reported on or off site, to RCRA Subtitle C landfills or surface impoundments as recorded by 2004 EPA Data
- cleanair_nond_tot: total quantity (reported in pounds) of non-dioxin chemicals covered by the Clean Air Act that were released in a given county via fugitive air, stack air, water, underground, landfill, land treatment, or surface impoundment, as recorded by 2004 EPA Data.
- carc_nond_tot: total quantity (reported in pounds) of non-dioxin carcinogens that were released in a given county via fugitive air, stack air, water, underground, landfill, land treatment, or surface impoundment, as recorded by 2004 EPA Data.
- carc_d_tot: total quantity (reported in grams) of dioxin and dioxin-like compounds that were released in a given county via fugitive air, stack air, water, underground, landfill, land treatment, or surface impoundment, as recorded by 2004 EPA Data.

Our new features look like this:

life_exp	tot_RCRA	cleanair_nond_tot	carc_nond_tot	carc_d_tot
71.33	0.00	0.0002	0.0002	0.0000
77.34	0.00	0.0002	0.0002	0.0000
78.56	5821902.60	842008.2160	646780.2762	0.0000
72.74	5821902.60	842008.2160	646780.2762	0.0000
78.70	28046.02	1871206.6492	229972.6092	0.4038
73.00	28046.02	1871206.6492	229972.6092	0.4038

Some of the features we added include number of clean air chemicals and carcinogens being released. What is their relationship?





Knowledge check

- Do any variables have a strong correlation with life expectancy?
    - no, most values are near-zero and have little to no correlation with life expectancy.
- According to the boxplots above, do you think around the same amount of non-dioxin carcinogens (carc_nond_tot) are released to most counties?
    - no, there seem to be just a few outliers with a lot of carcinogens, while almost no carcinogens are released in most counties.

Modeling environmental and demographic data with linear regression

```
reg2_df <- reg_df[,c("life_exp",
          "log_pop", "Area", "num_sf", "gender",
          'tot_RCRA', 'cleanair_nond_tot',
          'carc_nond_tot', 'carc_d_tot')]

reg2 <- lm(data = reg2_df, life_exp ~ .)
summary(reg2)
```

```
##
## Call:
## lm(formula = life_exp ~ ., data = reg2_df)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3951 -1.3760 -0.2227  1.1675  7.9614
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         6.481e+01  1.174e+00  55.200  < 2e-16 ***
## log_pop             1.319e+00  1.085e-01  12.157  < 2e-16 ***
## Area                6.334e-04  3.071e-04   2.063   0.0401 *
## num_sf             -6.686e-01  1.551e-01  -4.311 2.29e-05 ***
## gendermale         -5.550e+00  2.333e-01 -23.791  < 2e-16 ***
## tot_RCRA           -2.325e-07  1.998e-07  -1.164   0.2455
## cleanair_nond_tot   8.468e-11  7.980e-08   0.001   0.9992
## carc_nond_tot       2.479e-07  2.360e-07   1.051   0.2944
## carc_d_tot         -2.125e-04  1.202e-04  -1.769   0.0781 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.938 on 267 degrees of freedom
##   (70 observations deleted due to missingness)
## Multiple R-squared:  0.7329, Adjusted R-squared:  0.7249
## F-statistic: 91.58 on 8 and 267 DF,  p-value: < 2.2e-16
```

Knowledge Check

- Which variable(s) are significant to the model? (assume a p-value of .1)
  - We see a definite significant effect of the transformed population variable, area, number of superfund sites, and gender on life expectancy. We see a much lesser, but potentially significant effect of doxin and dioxin-like compounds on life expectancy.
- What is the R Squared value of the model?
  - 73%

Knowledge Check

Consider Bexar County.

```
## # A tibble: 2 x 6
##   County log_pop  Area num_sf gender life_exp
##   <fct>    <dbl> <dbl>  <dbl> <fct>     <dbl>
## 1 Bexar     14.4 1240.      3 female     81.9
## 2 Bexar     14.4 1240.      3 male       76.2
```

According to the following model, what do you predict as the life expectancy?

```
## 
## Call:
## lm(formula = life_exp ~ ., data = tx_reg_data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -5.5470 -1.4896 -0.1719  1.2728  7.5157 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 67.1428415  1.0390280  64.621  < 2e-16 ***
## log_pop      1.1303404  0.0963213  11.735  < 2e-16 ***
## Area         0.0005190  0.0002177   2.384   0.0177 *  
## num_sf      -0.4198741  0.0902899  -4.650 4.75e-06 ***
## gendermale  -5.5857225  0.2128990 -26.236  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.98 on 341 degrees of freedom
## Multiple R-squared:  0.7102, Adjusted R-squared:  0.7068 
## F-statistic: 208.9 on 4 and 341 DF,  p-value: < 2.2e-16
```

Calculate:

```r
67.1428415+
  1.1303404*14.35479+
  0.0005190*1239.8-
  0.4198741*3-
  5.5857225*1 #if male
```

```
## [1] 77.16675
```

```r
67.1428415+
  1.1303404*14.35479+
  0.0005190*1239.8-
  0.4198741*3 #if female
```

```
## [1] 82.75247
```