# Correlations & Regressions

*Antonio Vazquez Brust*

*November 17, 2015*

```
options(scipen = 99)
```

```
TAdata <- read.csv('c:/Users/havb/Dropbox/MSUI/Big Data for Cities - PPUA 5262 - 01/R/data/Tax Assessor/

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
```

```
TAdata <- mutate(TAdata,
                 BLDG_RANK = ceiling(rank(AV_BLDG_PER_SF,na.last="keep")/
                                     length(which(!is.na(AV_BLDG_PER_SF)))*100) )
```

```
require(Hmisc)
```

```
## Loading required package: Hmisc
## Loading required package: grid
## Loading required package: lattice
## Loading required package: survival
## Loading required package: Formula
## Loading required package: ggplot2
##
## Attaching package: 'Hmisc'
##
## The following objects are masked from 'package:dplyr':
##
##     combine, src, summarize
##
## The following objects are masked from 'package:base':
##
##     format.pval, round.POSIXt, trunc.POSIXt, units
```

```
distance.to.BC.VERSUS.av.sqf.value <- rcorr(as.matrix(TAdata[c("DIST_TO_DX_MI", "AV_BLDG_PER_SF", "BLDG_
distance.to.BC.VERSUS.av.sqf.value
```

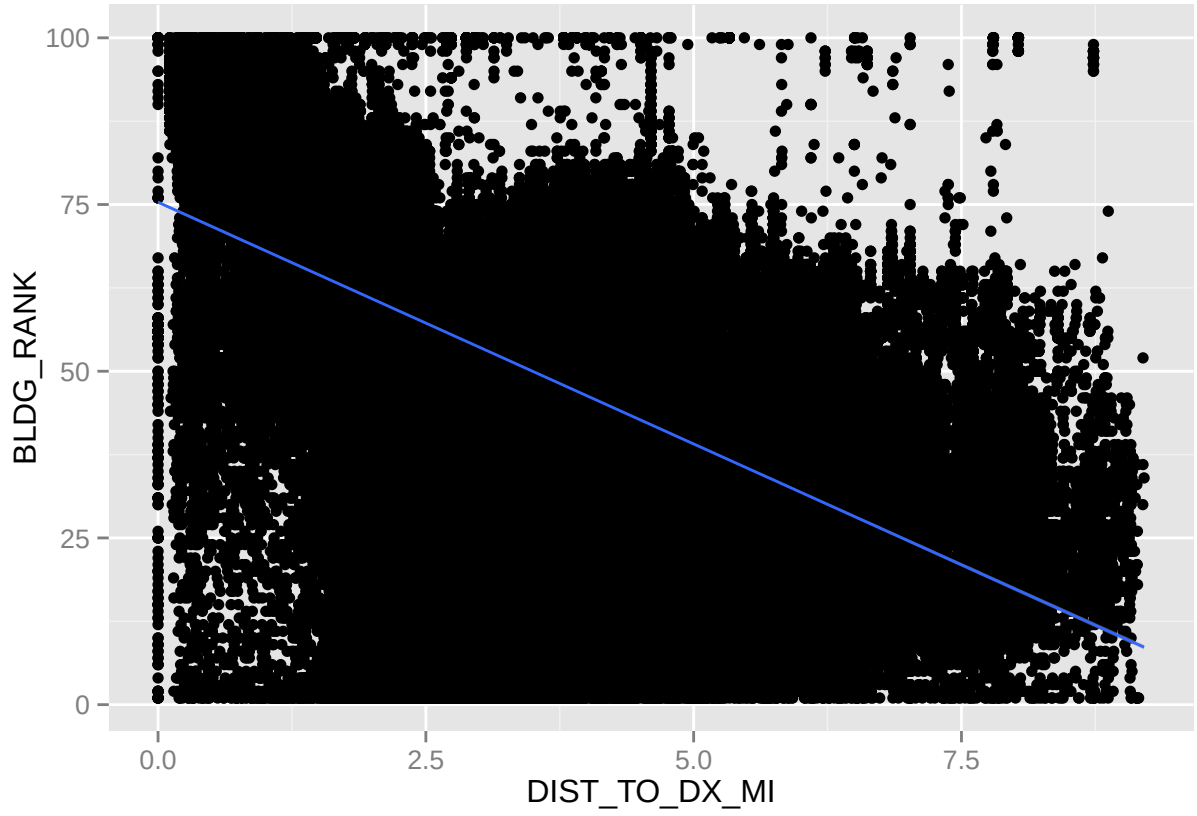
```
##                DIST_TO_DX_MI AV_BLDG_PER_SF BLDG_RANK
```

```
## DIST_TO_DX_MI           1.00          -0.02     -0.55
## AV_BLDG_PER_SF         -0.02           1.00      0.04
## BLDG_RANK              -0.55           0.04      1.00
##
## n
##                 DIST_TO_DX_MI AV_BLDG_PER_SF BLDG_RANK
## DIST_TO_DX_MI          168146         137827    137827
## AV_BLDG_PER_SF         137827         137827    137827
## BLDG_RANK              137827         137827    137827
##
## P
##                 DIST_TO_DX_MI AV_BLDG_PER_SF BLDG_RANK
## DIST_TO_DX_MI                0              0
## AV_BLDG_PER_SF  0                           0
## BLDG_RANK       0             0
```

```
library(ggplot2)
ggplot(data = TAdata, aes(x = DIST_TO_DX_MI, y = BLDG_RANK)) + geom_point() + geom_smooth(method = lm)
```

```
## Warning: Removed 30319 rows containing missing values (stat_smooth).
```

```
## Warning: Removed 30319 rows containing missing values (geom_point).
```



```r
regression.bldg.rank <- lm(data = TAdata, BLDG_RANK ~ DIST_TO_DX_MI)
summary(regression.bldg.rank)
```

```
##
## Call:
## lm(formula = BLDG_RANK ~ DIST_TO_DX_MI, data = TAdata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -74.34 -16.55   4.86  19.53  86.96
##
## Coefficients:
##               Estimate Std. Error t value            Pr(>|t|)
## (Intercept)   75.34371    0.12127   621.3 <0.0000000000000002 ***
## DIST_TO_DX_MI -7.24919    0.02986  -242.8 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.16 on 137825 degrees of freedom
##   (30319 observations deleted due to missingness)
## Multiple R-squared:  0.2995, Adjusted R-squared:  0.2995
## F-statistic: 5.894e+04 on 1 and 137825 DF,  p-value: < 0.00000000000000022
```

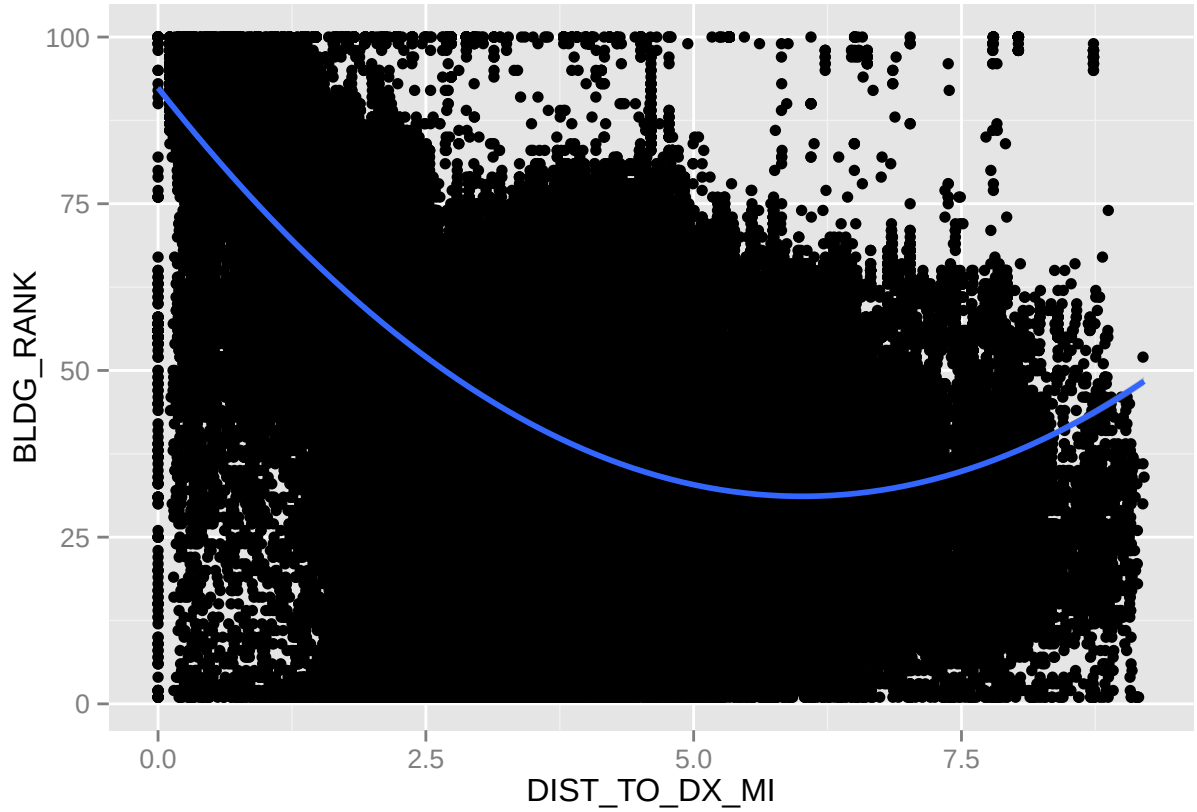
```r
regression.bldg.rank <- lm(data = TAdata, BLDG_RANK ~ DIST_TO_DX_MI+I(DIST_TO_DX_MI^2))
summary(regression.bldg.rank)
```

```
##
## Call:
## lm(formula = BLDG_RANK ~ DIST_TO_DX_MI + I(DIST_TO_DX_MI^2),
##     data = TAdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -91.374 -16.377   3.513  16.811  68.862
##
## Coefficients:
##                     Estimate Std. Error t value            Pr(>|t|)
## (Intercept)         92.37365    0.17315   533.5 <0.0000000000000002 ***
## DIST_TO_DX_MI      -20.37063    0.10405  -195.8 <0.0000000000000002 ***
## I(DIST_TO_DX_MI^2)   1.69381    0.01293   131.0 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.78 on 137824 degrees of freedom
##   (30319 observations deleted due to missingness)
## Multiple R-squared:  0.3771, Adjusted R-squared:  0.3771
## F-statistic: 4.172e+04 on 2 and 137824 DF,  p-value: < 0.00000000000000022
```

```r
ggplot(data = TAdata, aes(x = DIST_TO_DX_MI, y = BLDG_RANK)) +  geom_point() + stat_smooth(method = "lm
```

```
## Warning: Removed 30319 rows containing missing values (stat_smooth).
```

```
## Warning: Removed 30319 rows containing missing values (geom_point).
```




```
demographics <- read.csv('C:/Users/havb/Dropbox/MSUI/Big Data for Cities - PPUA 5262 - 01/R/data/Census/
```