

Exploratory analysis

Attach packages

```
library(tidyverse)
library(caret)
library(psych)
library(BioStatR)
library(car)
library(lattice)
```

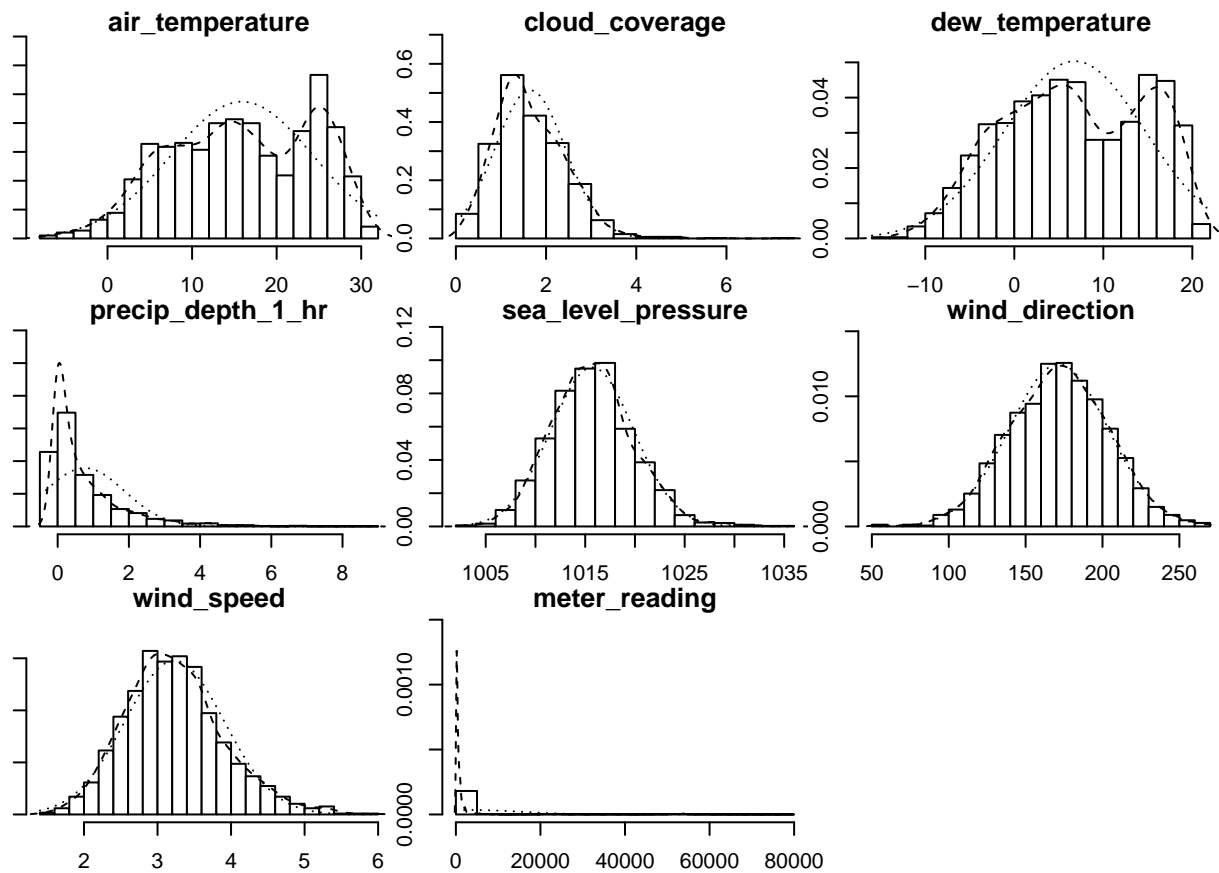
Load the datafile

```
daily_data <- readRDS("daily_data.rds")
summary(daily_data)
```

```
##      date           meter      air_temperature  cloud_coverage
## Min.   :2016-01-01   Min.    :0.00   Min.     :-6.569   Min.     :0.000
## 1st Qu.:2016-04-01   1st Qu.:0.75   1st Qu.: 9.224   1st Qu.:1.108
## Median :2016-07-01   Median :1.50   Median :15.902   Median :1.524
## Mean   :2016-07-01   Mean    :1.50   Mean    :15.899   Mean    :1.646
## 3rd Qu.:2016-10-01   3rd Qu.:2.25   3rd Qu.:23.825   3rd Qu.:2.148
## Max.   :2016-12-31   Max.     :3.00   Max.     :31.616   Max.     :7.292
## dew_temperature  precip_depth_1_hr  sea_level_pressure  wind_direction
## Min.     :-14.147   Min.     :-0.46323   Min.     :1003      Min.     : 56.52
## 1st Qu.:  0.537   1st Qu.: 0.01168   1st Qu.:1013      1st Qu.:149.43
## Median :  6.577   Median : 0.33109   Median :1016      Median :171.41
## Mean    :  6.785   Mean    : 0.76072   Mean    :1016      Mean    :171.48
## 3rd Qu.: 14.060   3rd Qu.: 1.07500   3rd Qu.:1018      3rd Qu.:193.54
## Max.     :21.655   Max.     : 8.69171   Max.     :1035      Max.     :267.27
##  wind_speed      meter_reading
## Min.     :1.453   Min.     : 110.5
## 1st Qu.:2.775   1st Qu.: 186.4
## Median :3.185   Median : 411.5
## Mean    :3.231   Mean    :3859.2
## 3rd Qu.:3.616   3rd Qu.: 990.0
## Max.     :5.953   Max.     :77117.7
```

Examine distributions

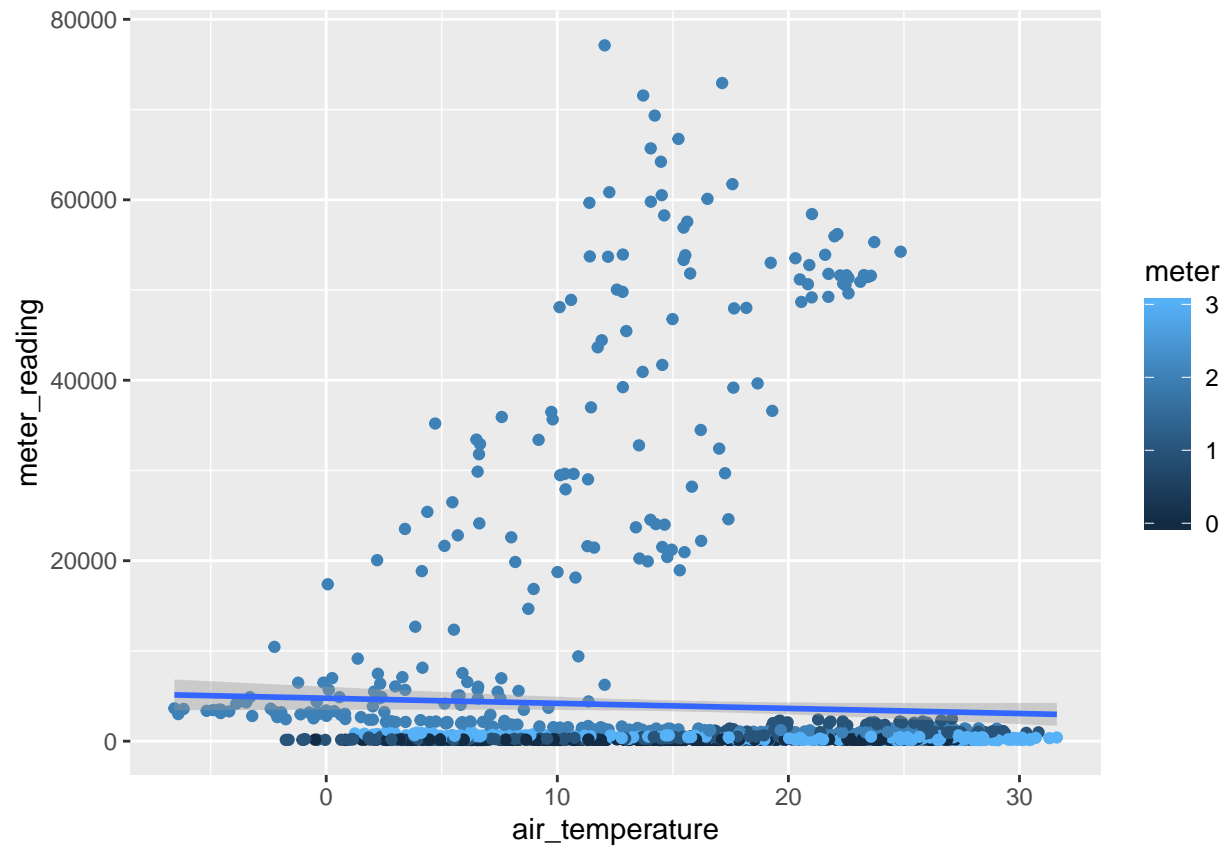
```
numeric <- daily_data[,3:10]
multi.hist(numeric)
```



The target variable (`meter_reading`) is heavily right-skewed. `cloud_coverage` and `precip_depth_1_hr` are also right-skewed.

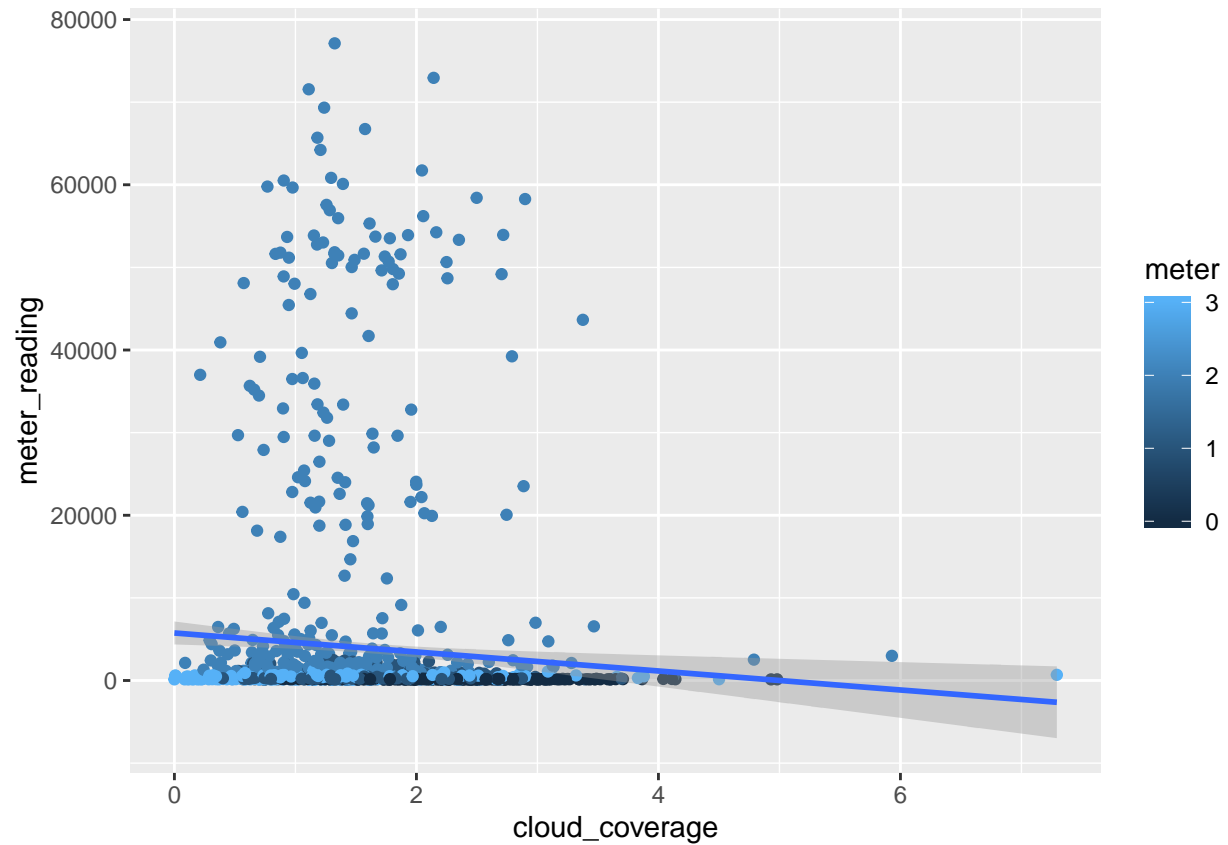
Examine scatterplots

```
air_temp <- ggplot(daily_data, aes(x = air_temperature, y = meter_reading, colour = meter)) + geom_point()
air_temp
```



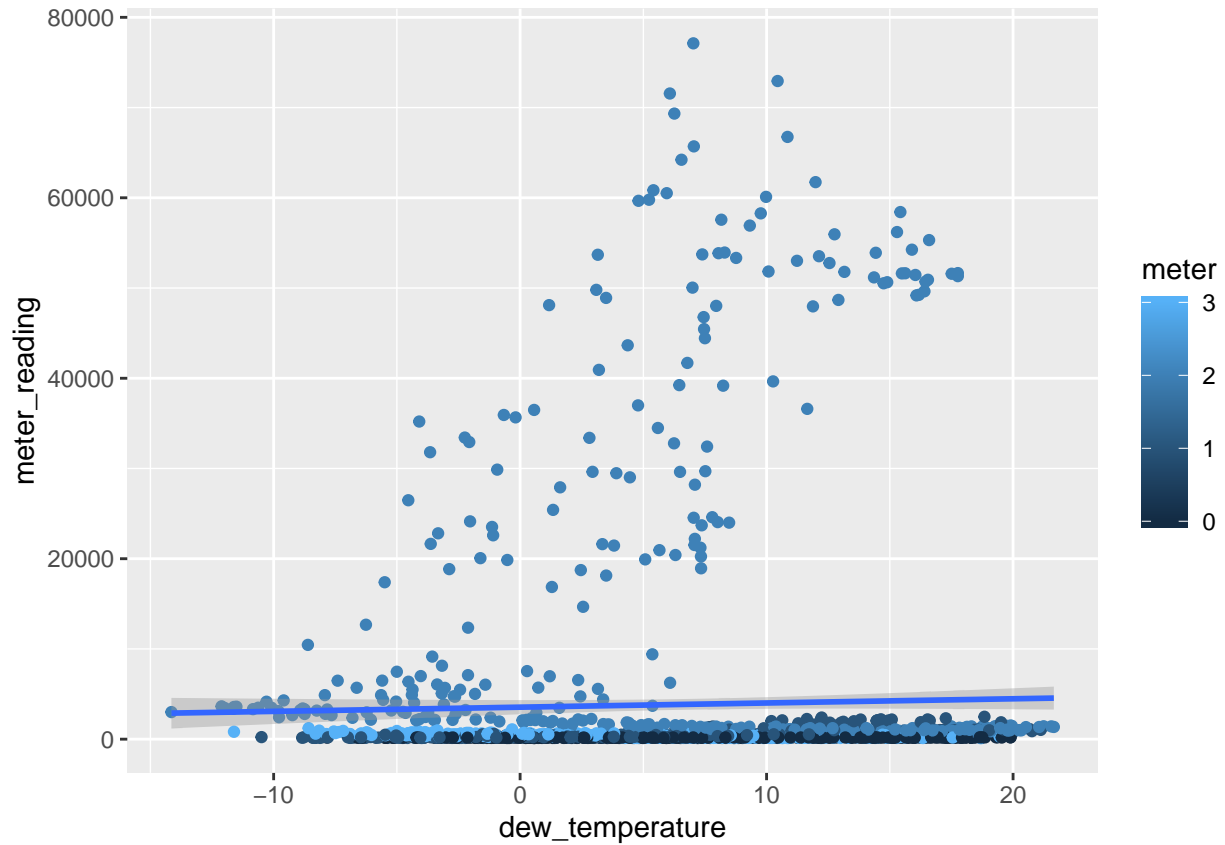
There are outliers but for the most part there is not much of a linear relationship between `air_temperature` and `meter_reading`. meter type 3 does show a stronger linear relationship than the other two meter types.

```
cloud_cover <- ggplot(daily_data, aes(x = cloud_coverage, y = meter_reading, colour = meter)) + geom_point()
cloud_cover
```



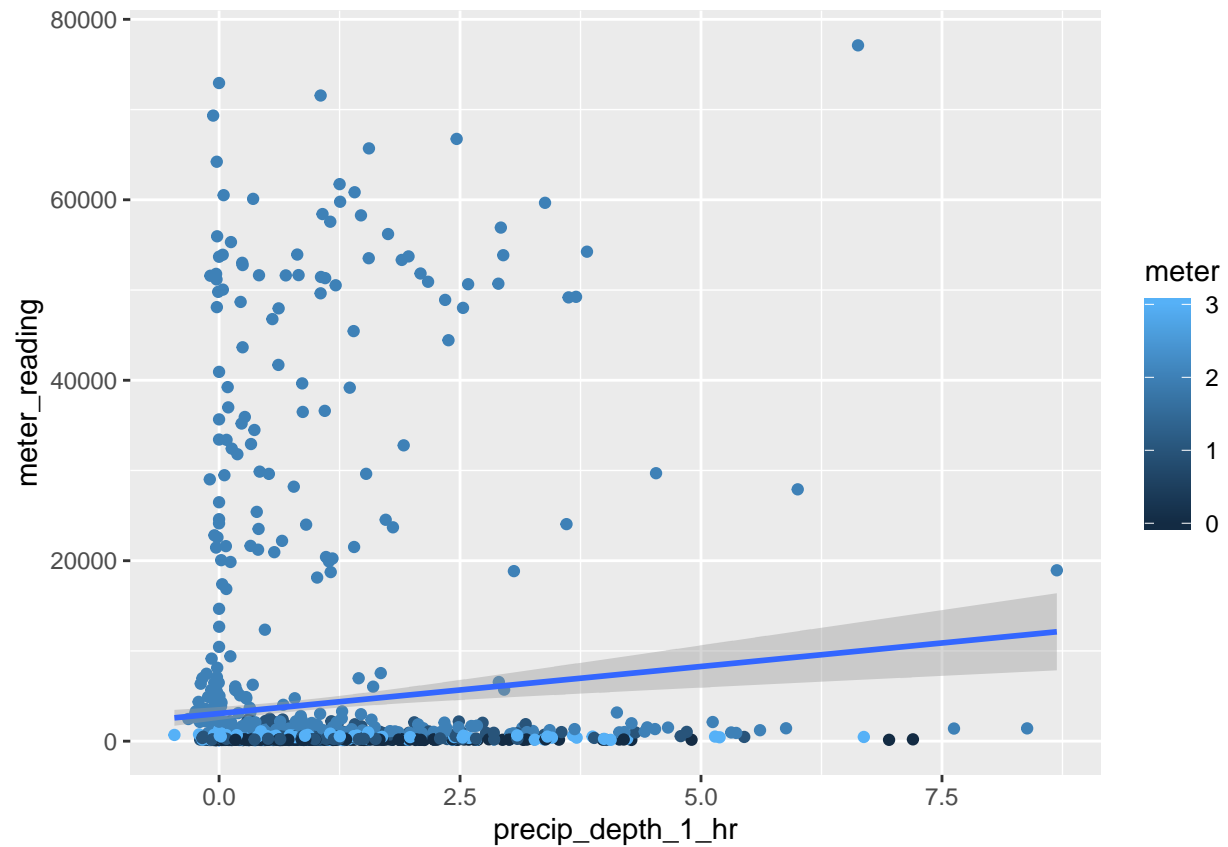
cloud_coverage does not have a clear linear relationship with meter reading except for meter 3.

```
dew_temp <- ggplot(daily_data, aes(x = dew_temperature, y = meter_reading, colour = meter)) + geom_point()
dew_temp
```



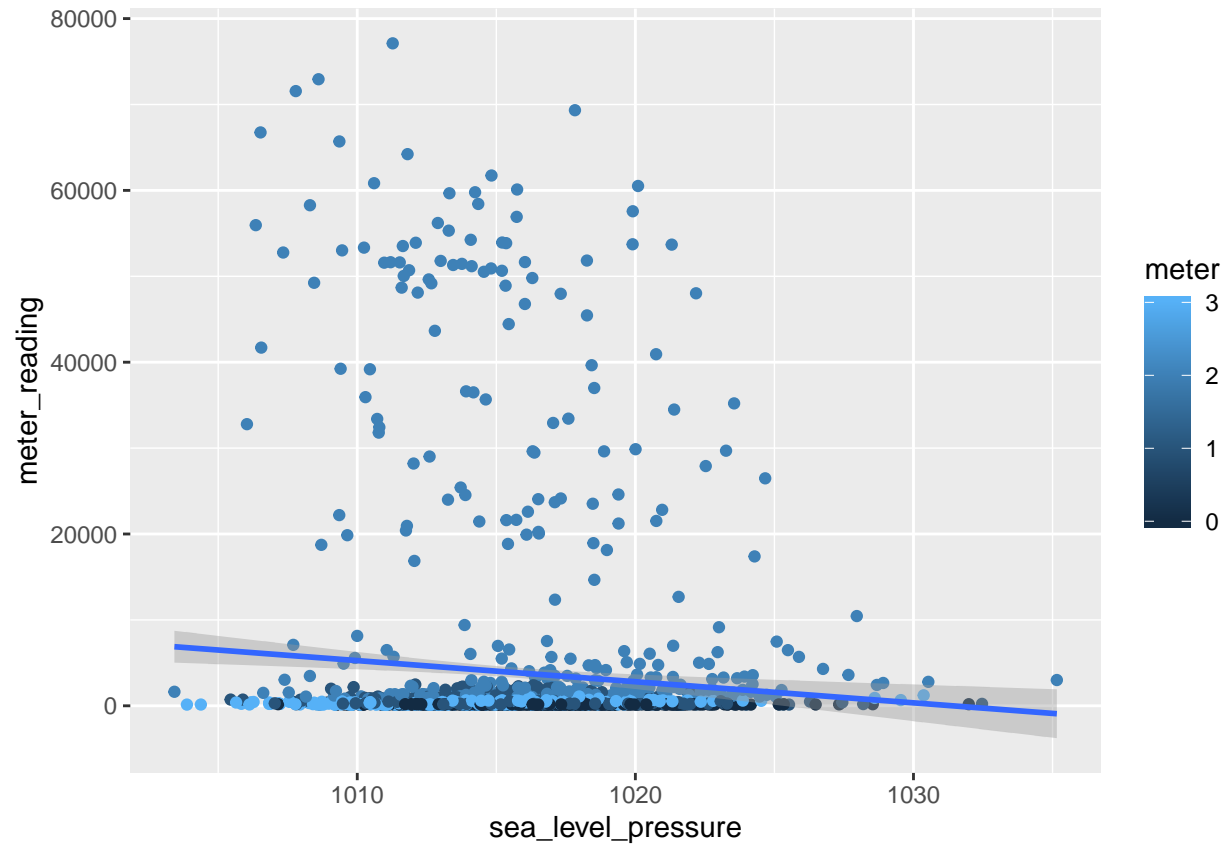
We see the same trend in dew temperature as the earlier two temperature variables. There is not much of a linear relationship, except for meter 3.

```
precip_depth <- ggplot(daily_data, aes(x = precip_depth_1_hr, y = meter_reading, colour = meter)) + geom_point()
precip_depth
```



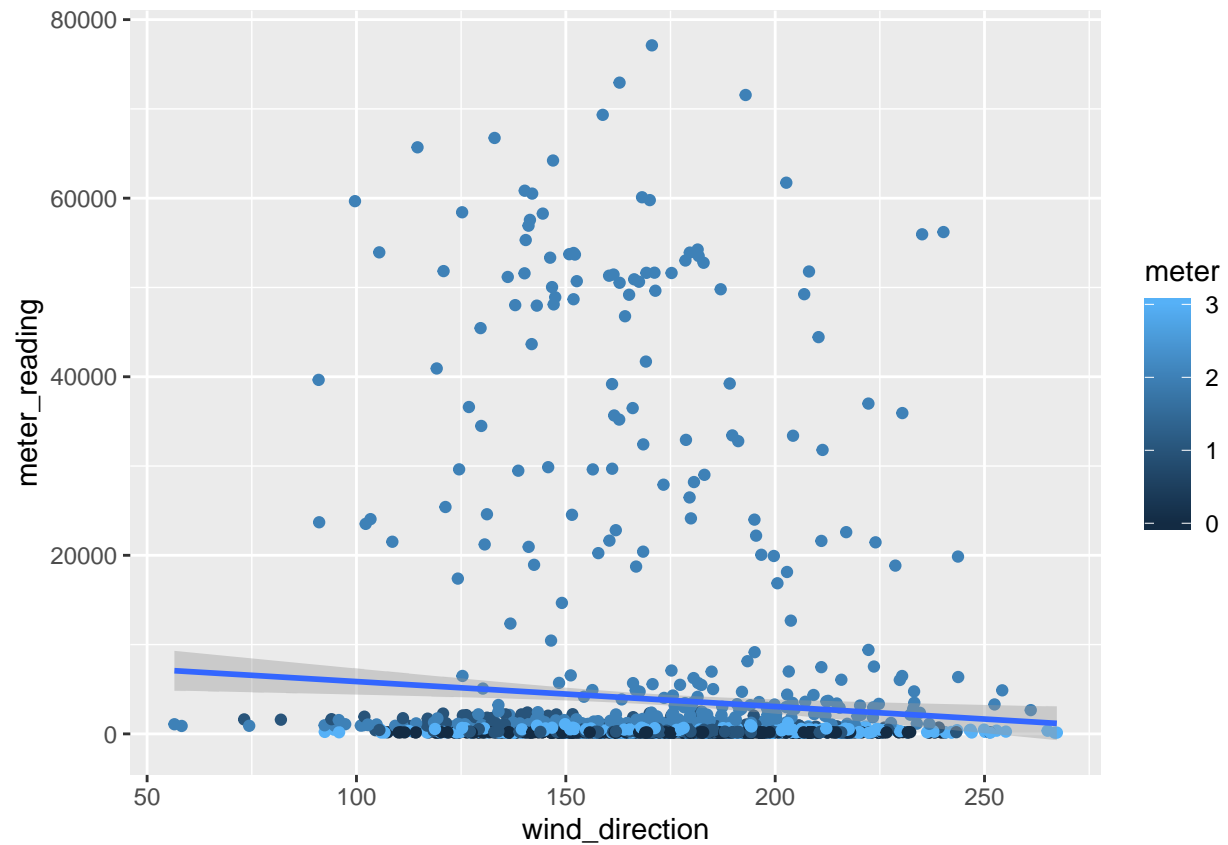
Precip_depth_1_hr does not have any relationship with meter reading. Even meter 3 is all over the place.

```
sea_pressure <- ggplot(daily_data, aes(x = sea_level_pressure, y = meter_reading, colour = meter)) + geom_point()
sea_pressure
```



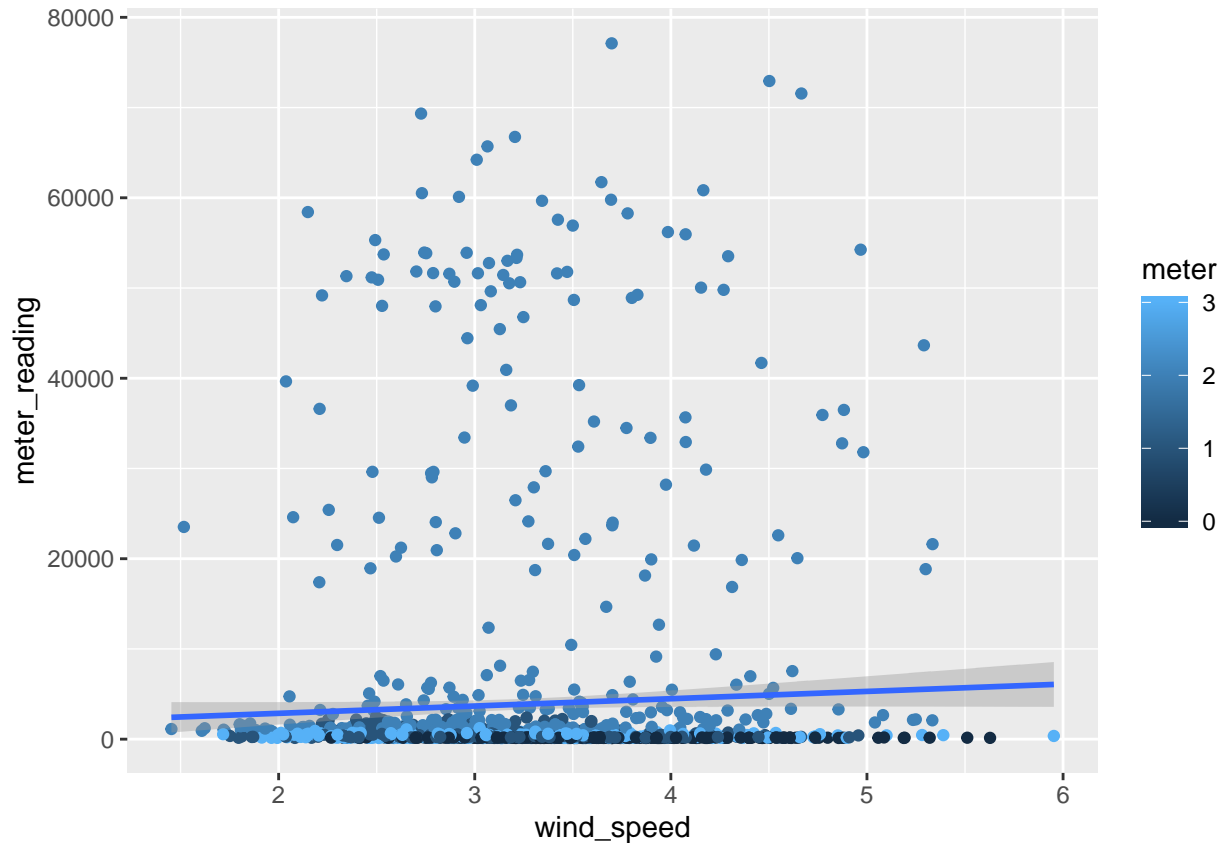
sea_level pressure has no relationship with meter reading for all meter types.

```
wind_direction <- ggplot(daily_data, aes(x = wind_direction, y = meter_reading, colour = meter)) + geom_point()
wind_direction
```



No relationship between wind direction and meter reading. meter 3 is all over the place!

```
wind_speed <- ggplot(daily_data, aes(x = wind_speed, y = meter_reading, colour = meter)) + geom_point()
wind_speed
```

No relationship between wind speed and meter reading.

Look at the Building related variables

```
building_vars <- readRDS("building_vars.rds")
summary(building_vars)
```

```
## meter_reading      building_id  primary_use      square_feet
## Min.   :      0.0   Min.   :    0   Length:1449   Min.   :   283
## 1st Qu.:    32.0   1st Qu.:  362   Class :character 1st Qu.: 23012
## Median :    94.1   Median :  724   Mode  :character Median : 57673
## Mean   :   1652.3   Mean   :  724               Mean   :  92112
## 3rd Qu.:   256.9   3rd Qu.:1086               3rd Qu.:115676
## Max.   :1907445.9   Max.   :1448               Max.   :875000
##
##   year_built    floor_count
## Min.   :1900    Min.   : 1.000
## 1st Qu.:1949    1st Qu.: 1.000
## Median :1970    Median : 3.000
## Mean   :1968    Mean   : 3.741
## 3rd Qu.:1995    3rd Qu.: 5.000
## Max.   :2017    Max.   :26.000
## NA's   :774     NA's   :1094
```

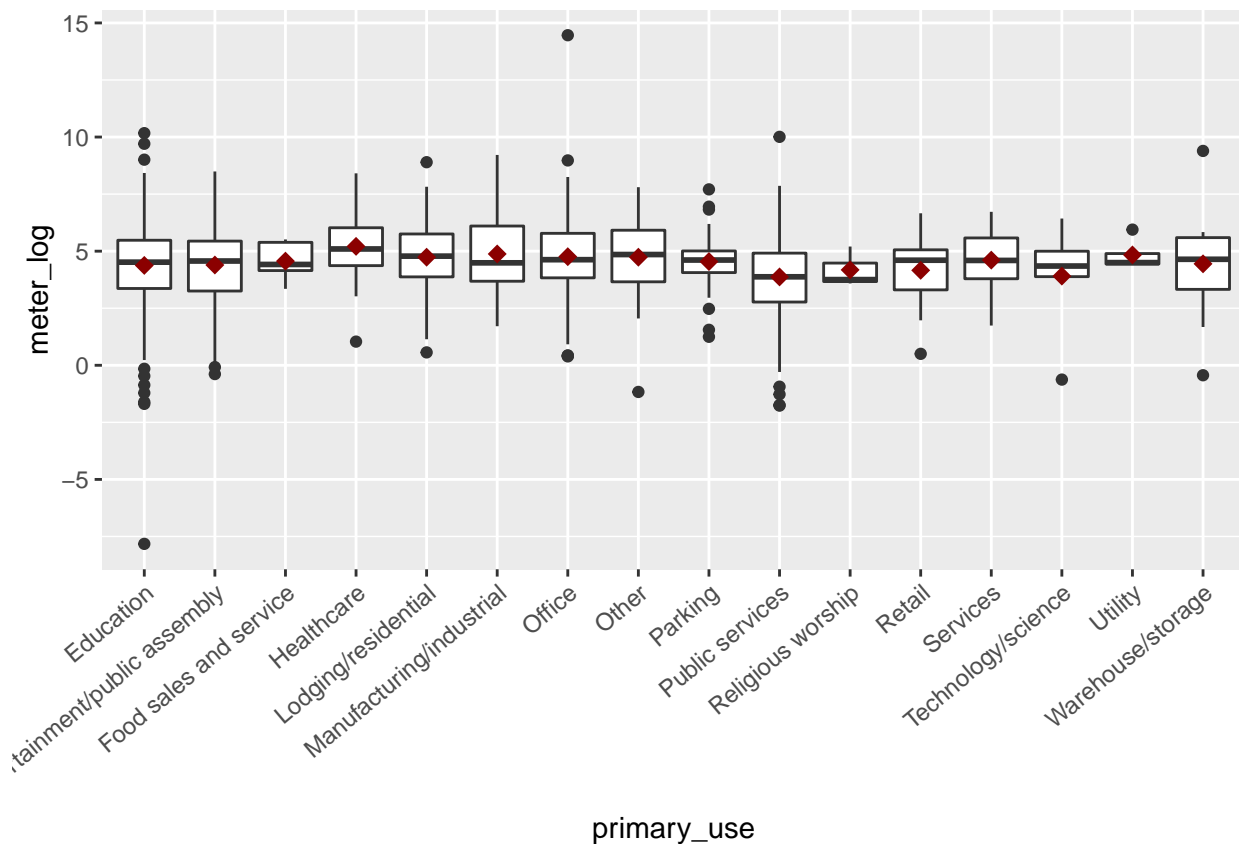
```
dim(building_vars)
```

```
## [1] 1449    6
```

```
building_vars$primary_use <- as.factor(building_vars$primary_use)
summary(building_vars$primary_use)
```

```
##           Education Entertainment/public assembly
##           549                               184
## Food sales and service                     Healthcare
##           5                               23
## Lodging/residential      Manufacturing/industrial
##           147                               12
##           Office                               Other
##           279                               25
##           Parking                     Public services
##           22                               156
## Religious worship                     Retail
##           3                               11
## Services                     Technology/science
##           10                               6
## Utility                     Warehouse/storage
##           4                               13
```

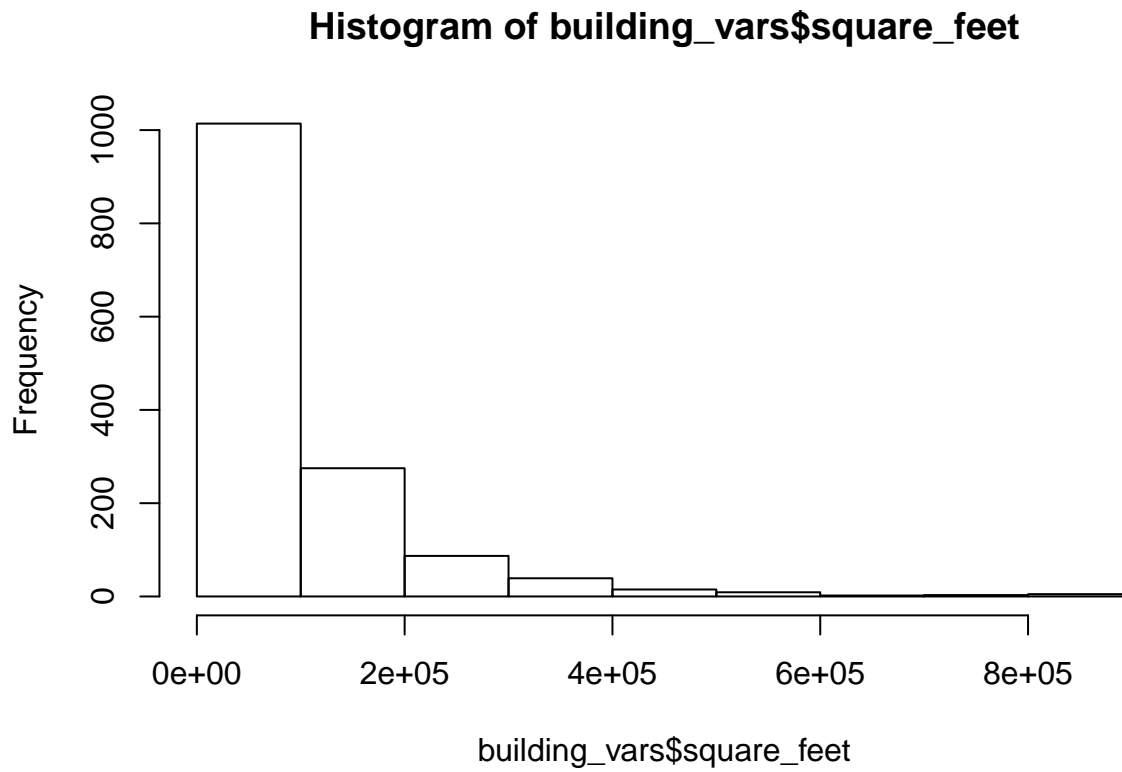
```
meter_log <- log(building_vars$meter_reading)
use <- ggplot(building_vars, aes(x = primary_use, y = meter_log)) + geom_boxplot() +
  stat_summary(fun.y=mean, colour="darkred", geom="point",
    shape=18, size=3)
use + theme(axis.text.x = element_text(angle = 40, hjust = 1))
```



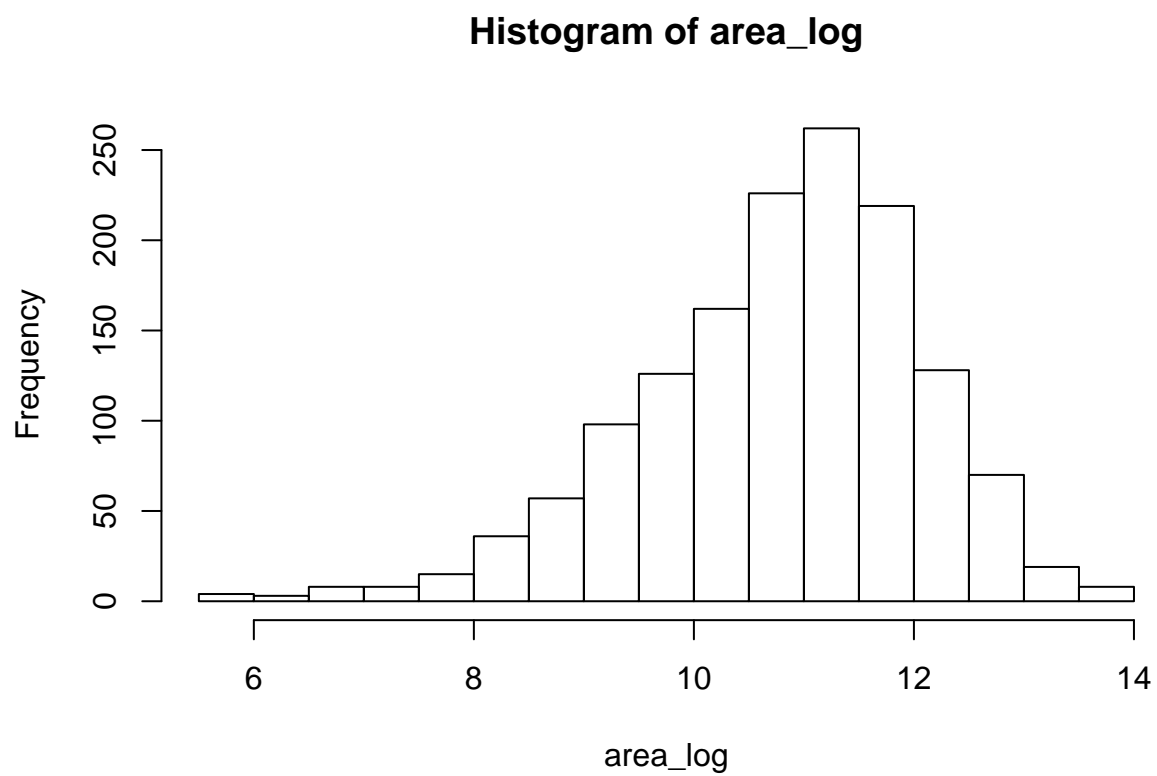
Not much of a difference in means. I logged meter readings because they are so strongly right skewed. I

could not see a trend if I did not log them. No relationship between use of building and meter reading. Now, lets look at the area of the buildings. Again this variable is highly right-skewed. I'm going to log it.

```
hist(building_vars$square_feet)
```

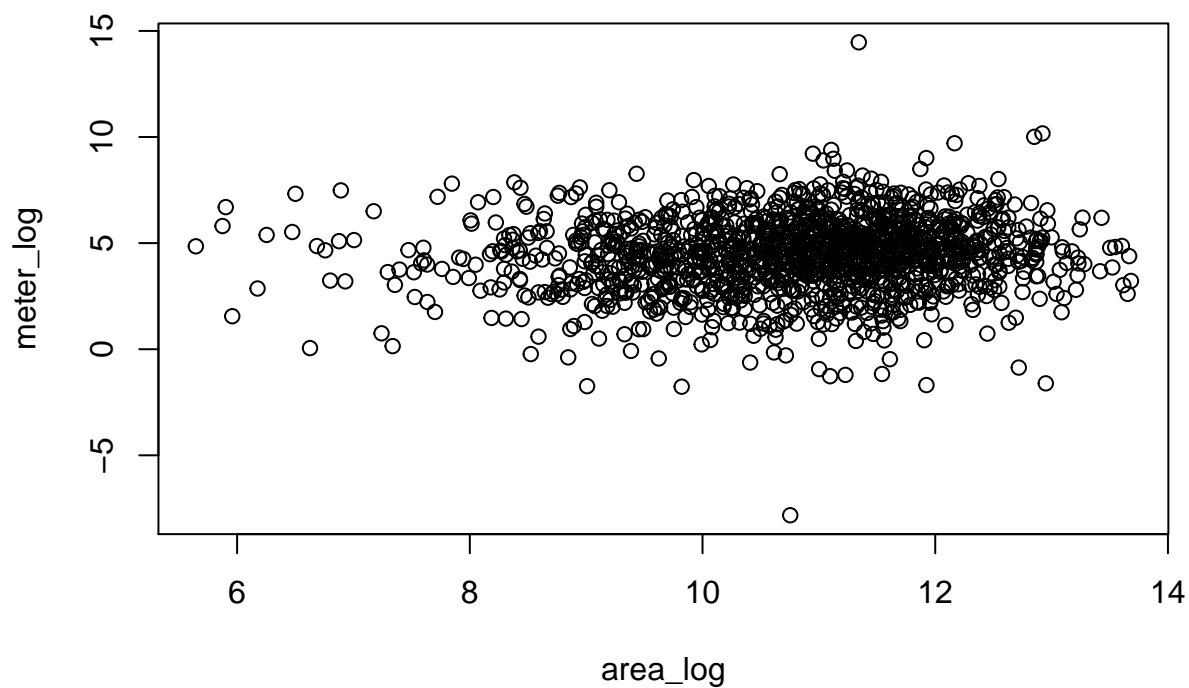


```
area_log <- log(building_vars$square_feet)
hist(area_log)
```



looks much better!

```
plot(area_log, meter_log)
```

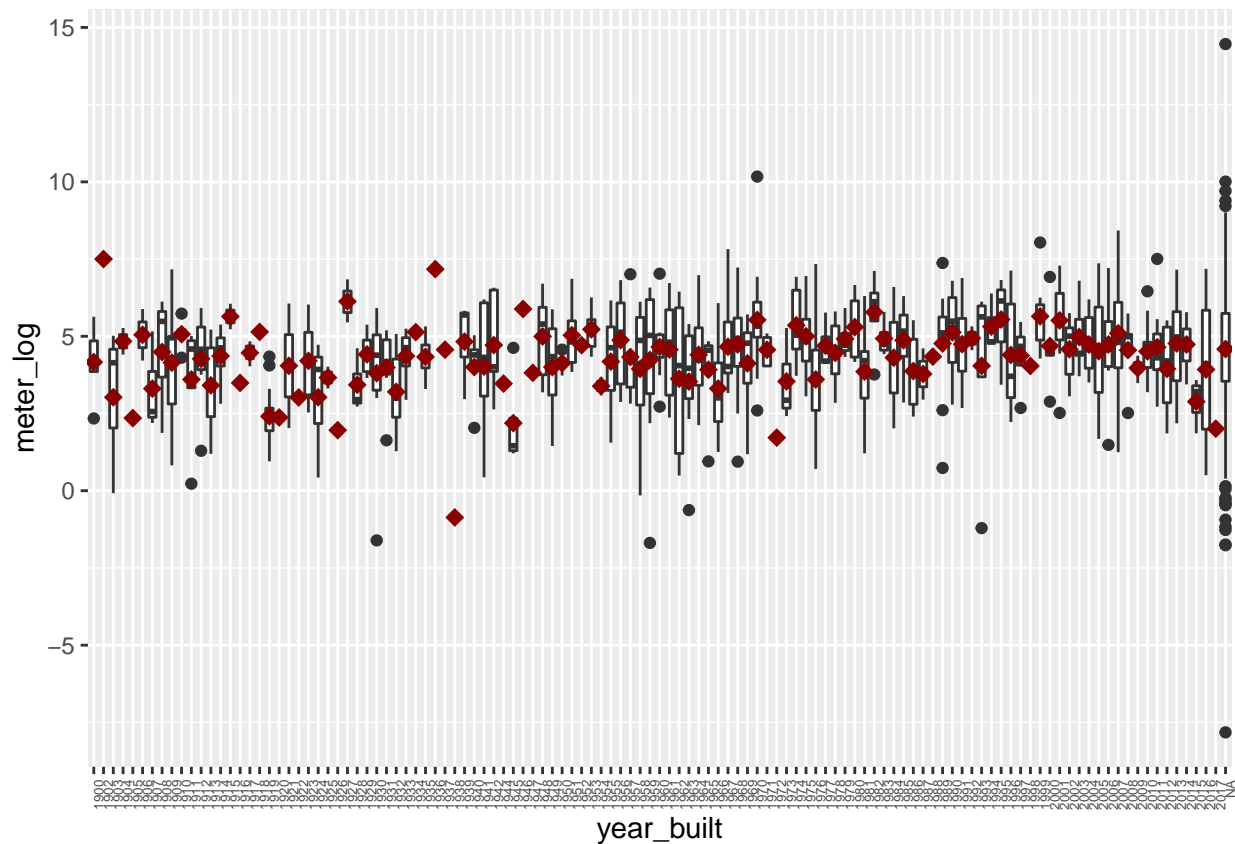


Nothing going on here! There is no relationship between area and meter readings.

```
building_vars$year_built <- as.factor(building_vars$year_built)
summary(building_vars$year_built)
```

```
##      1976      1966      1968      1919      1964      2004      1960      1975      2006
##      55       23       18       17       15       14       13       13       13
##      2007      1970      2001      2010      2014      2002      1930      1959      1967
##      13       12       12       12       12       11       10       10       10
##      2005      1989      2013      1923      1956      1999      1958      1963      1969
##      10       9        9        8        8        8        7        7        7
##      1990      2011      2016      1912      1913      1931      1932      1953      1965
##      7        7        7        6        6        6        6        6        6
##      1974      1981      1996      1900      1909      1910      1940      1941      1942
##      6        6        6        5        5        5        5        5        5
##      1948      1951      1955      1957      1961      1962      1971      1978      1986
##      5        5        5        5        5        5        5        5        5
##      1995      2000      2003      2008      1911      1914      1929      1933      1935
##      5        5        5        5        4        4        4        4        4
##      1945      1949      1950      1979      1980      1982      1983      1985      1991
##      4        4        4        4        4        4        4        4        4
##      1993      1997      2009      2012      1903      1907      1908      1917      1924
##      4        4        4        4        3        3        3        3        3
##      1927      1928      1939      1973      1977      1994      2015      1904      1906
##      3        3        3        3        3        3        3        2        2
##      1915      1920      1921      1925      1952      1954      1984      1987 (Other)
##      2        2        2        2        2        2        2        2        21
```

```
##      NA's
##      774
meter_log <- log(building_vars$meter_reading)
year <- ggplot(building_vars, aes(x = year_built, y = meter_log)) + geom_boxplot() + stat_summary(fun.y
year + theme(axis.text.x = element_text(size = 5, angle = 90, hjust = 1))
```

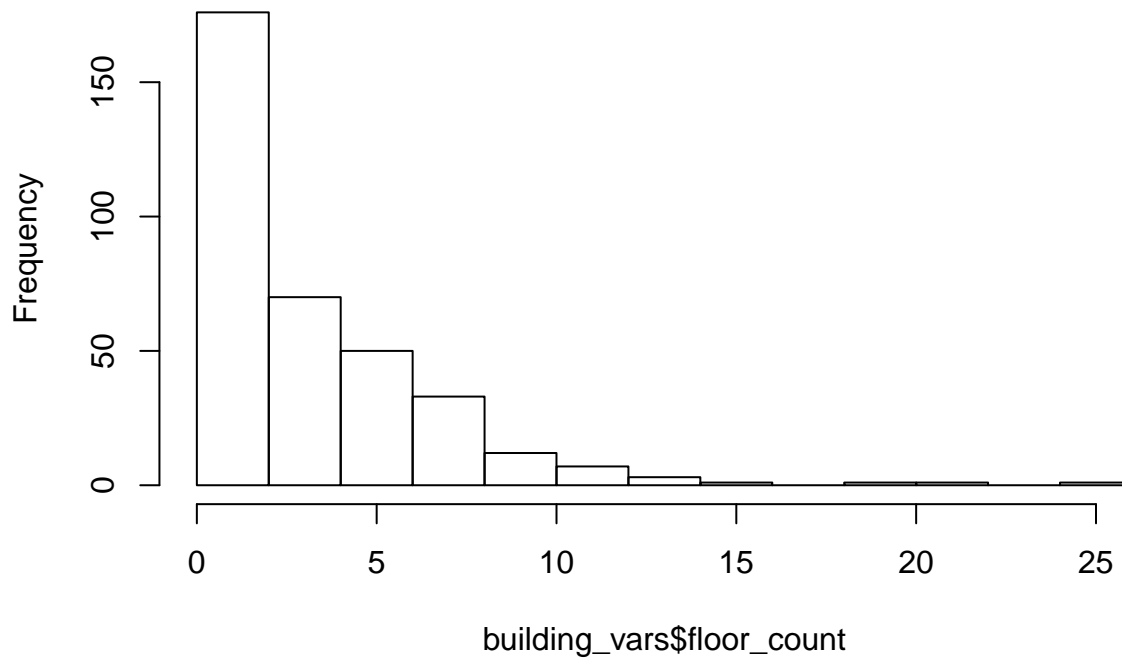


There is some variation but not much of a trend. It almost seems random. Surprisingly, the newer buildings are not using less energy!!

Now lets look at floor count!

```
hist(building_vars$floor_count)
```

Histogram of building_vars\$floor_count



```
table(building_vars$floor_count)
```

```
##  
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 16 19 21 26  
## 109 67 33 37 25 25 14 19  8  4  5  2  2  1  1  1  1  1
```

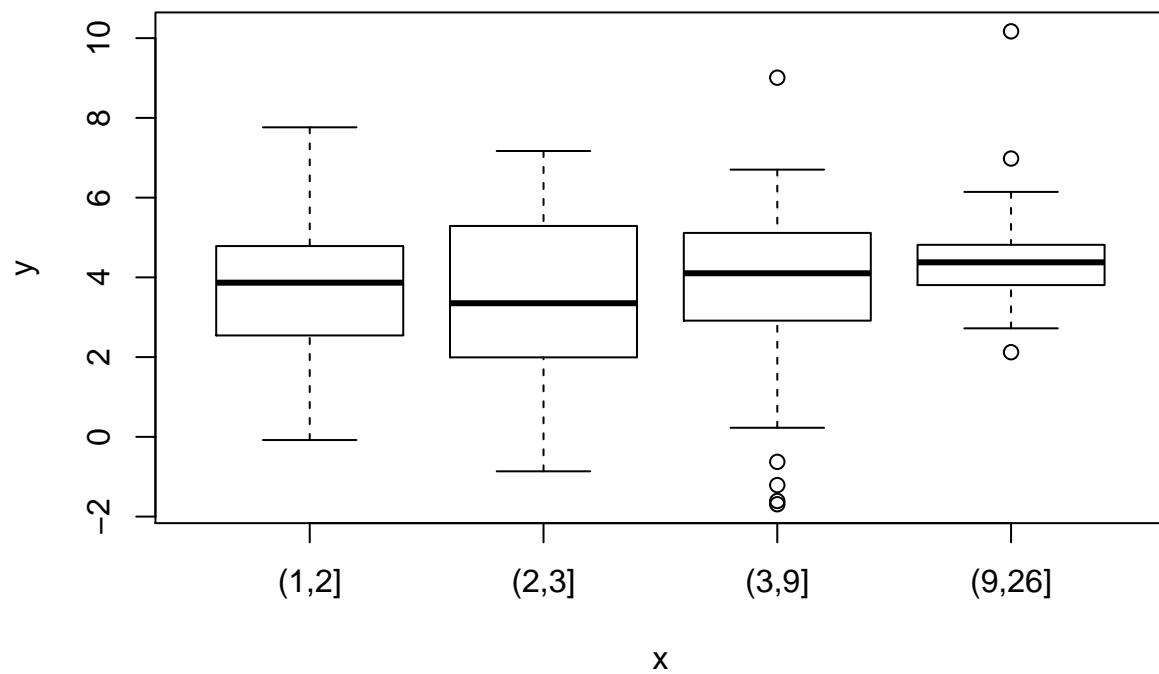
High number of 1 storey buildings. Lets cut this into a factor with 1 storey, 2 storey, 3 to 9 storey and more than 9 storey buildings. That looks like a natural grouping to me looking at the histogram.

```
cut_storey <- cut(building_vars$floor_count, breaks = c(1, 2, 3, 9, 26))
```

```
summary(cut_storey)
```

```
## (1,2] (2,3] (3,9] (9,26] NA's  
##    67    33   128    18  1203
```

```
cut_storey <- as.factor(cut_storey)  
plot(cut_storey, meter_log)
```



Not much of a difference in the medians!

Main conclusion: My EDA does not show any strong relationships between the predictors and the target in this dataset.