Introduction to R

Kevin Shook

November 23, 2017

basic arithmetic: + - / *

```
1 + 1
```

```
## [1] 2
```

```
2 * 2
```

```
## [1] 4
```

```
4 / 3
```

```
## [1] 1.333333
```

data types

```
a <- 5
a
```

```
## [1] 5
```

```
b <- a + 1
b
```

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

```
b <- "hello, world"
b
```

```
## [1] "hello, world"
```

vectors

```
a <- c(1,2,3,4,5)
a
```

```
## [1] 1 2 3 4 5
```

```
b <- a/2
b
```

```
## [1] 0.5 1.0 1.5 2.0 2.5
```

character vectors

```
a <- c('1', '2', 'dog')
a
```

```
## [1] "1"   "2"   "dog"
```

combining characters

```
paste('dog', 'cat')
```

```
## [1] "dog cat"
```

works with vectors - vectors are recycled if too short

```r
a <- c(1,2,3,4,5)
b <- "o'clock"
paste(a,b)
```

```
## [1] "1 o'clock" "2 o'clock" "3 o'clock" "4 o'clock" "5 o'clock"
```

subsetting vectors

```r
a <- seq(10,20)
a
```

```
##  [1] 10 11 12 13 14 15 16 17 18 19 20
```

subset by location

```r
a[1:3]
```

```
## [1] 10 11 12
```

```r
a[-1]
```

```
##  [1] 11 12 13 14 15 16 17 18 19 20
```

subset by value

```r
a > 15
```

```
##  [1] FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE
```

```r
a[ a > 15]
```

```
## [1] 16 17 18 19 20
```

```r
evens <- a[(a %% 2) == 0]
evens
```

```
## [1] 10 12 14 16 18 20
```

commands

```r
mean(a)
```

```
## [1] 15
```

```r
var(a)
```

```
## [1] 11
```

get help on command

```r
?var
```

data frames loading data frame from a text file

```r
CalgaryDailyPrecip <- read.csv("CalgaryDailyPrecip.csv",
                               header = TRUE, stringsAsFactors = FALSE)
```

get info about a data frame

```r
head(CalgaryDailyPrecip)
```

```
##         date precip
## 1 1885-01-01      0
## 2 1885-01-02      0
## 3 1885-01-03      0
## 4 1885-01-04      0
## 5 1885-01-05      0
## 6 1885-01-06      0
```

```r
summary(CalgaryDailyPrecip)
```

```
##      date               precip
##  Length:46751       Min.   : 0.000
##  Class :character   1st Qu.: 0.000
##  Mode  :character   Median : 0.000
##                     Mean   : 1.278
##                     3rd Qu.: 0.480
##                     Max.   :99.330
##                     NA's   :175
```

```r
nrow(CalgaryDailyPrecip)   # number of rows
```

```
## [1] 46751
```

```r
ncol(CalgaryDailyPrecip)   # number of columns
```

```
## [1] 2
```

```r
names(CalgaryDailyPrecip)  # names inside the data frame
```

```
## [1] "date"   "precip"
```

convert from 0.1 mm to mm

```r
CalgaryDailyPrecip$precip <- CalgaryDailyPrecip$precip/10
summary(CalgaryDailyPrecip)
```

```
##      date               precip
##  Length:46751       Min.   :0.0000
##  Class :character   1st Qu.:0.0000
##  Mode  :character   Median :0.0000
##                     Mean   :0.1278
##                     3rd Qu.:0.0480
##                     Max.   :9.9330
##                     NA's   :175
```

calculate mean

```r
mean(CalgaryDailyPrecip$precip)
```

```
## [1] NA
```

```r
mean(na.omit(CalgaryDailyPrecip$precip))
```

```
## [1] 0.1278142
```

convert date string to a real date

```r
CalgaryDailyPrecip$realdate <- as.Date(CalgaryDailyPrecip$date,
                                       format = "%Y-%m-%d")
head(CalgaryDailyPrecip)
```

```
##         date precip   realdate
## 1 1885-01-01      0 1885-01-01
## 2 1885-01-02      0 1885-01-02
## 3 1885-01-03      0 1885-01-03
## 4 1885-01-04      0 1885-01-04
## 5 1885-01-05      0 1885-01-05
## 6 1885-01-06      0 1885-01-06
```

```
summary(CalgaryDailyPrecip)
```

```
##      date               precip          realdate         
##  Length:46751       Min.   :0.0000   Min.   :1885-01-01  
##  Class :character   1st Qu.:0.0000   1st Qu.:1917-01-01  
##  Mode  :character   Median :0.0000   Median :1949-01-01  
##                     Mean   :0.1278   Mean   :1949-01-01  
##                     3rd Qu.:0.0480   3rd Qu.:1980-12-31  
##                     Max.   :9.9330   Max.   :2012-12-31  
##                     NA's   :175
```

remove all missing values

```
CalgaryDailyPrecip <- na.omit(CalgaryDailyPrecip)
summary(CalgaryDailyPrecip)
```

```
##      date               precip          realdate         
##  Length:46576       Min.   :0.0000   Min.   :1885-01-01  
##  Class :character   1st Qu.:0.0000   1st Qu.:1916-11-18  
##  Mode  :character   Median :0.0000   Median :1948-10-05  
##                     Mean   :0.1278   Mean   :1948-10-05  
##                     3rd Qu.:0.0480   3rd Qu.:1980-08-22  
##                     Max.   :9.9330   Max.   :2012-07-11
```

get year

```
CalgaryDailyPrecip$year <- as.numeric(format(CalgaryDailyPrecip$realdate, "%Y"))
summary(CalgaryDailyPrecip)
```

```
##      date               precip          realdate               year     
##  Length:46576       Min.   :0.0000   Min.   :1885-01-01   Min.   :1885  
##  Class :character   1st Qu.:0.0000   1st Qu.:1916-11-18   1st Qu.:1916  
##  Mode  :character   Median :0.0000   Median :1948-10-05   Median :1948  
##                     Mean   :0.1278   Mean   :1948-10-05   Mean   :1948  
##                     3rd Qu.:0.0480   3rd Qu.:1980-08-22   3rd Qu.:1980  
##                     Max.   :9.9330   Max.   :2012-07-11   Max.   :2012
```

subset by year

```
y2007 <- CalgaryDailyPrecip[CalgaryDailyPrecip$year == 2007,]
head(y2007)
```

```
##             date precip   realdate year
## 44560 2007-01-01  0.000 2007-01-01 2007
## 44561 2007-01-02  0.000 2007-01-02 2007
## 44562 2007-01-03  0.000 2007-01-03 2007
## 44563 2007-01-04  0.038 2007-01-04 2007
## 44564 2007-01-05  0.021 2007-01-05 2007
## 44565 2007-01-06  0.021 2007-01-06 2007
```

or

```
y2005 <- subset(CalgaryDailyPrecip, year == 2005)
head(y2005)
```

```
##             date precip   realdate year
## 43830 2005-01-01  0.288 2005-01-01 2005
## 43831 2005-01-02  0.021 2005-01-02 2005
## 43832 2005-01-03  0.038 2005-01-03 2005
## 43833 2005-01-04  0.000 2005-01-04 2005
## 43834 2005-01-05  0.000 2005-01-05 2005
## 43835 2005-01-06  0.557 2005-01-06 2005
```

aggregate by year

```
CalgaryYearlyPrecip <- aggregate(CalgaryDailyPrecip$precip,
                                 by = list(CalgaryDailyPrecip$year), FUN = "sum")
head(CalgaryYearlyPrecip)
```

```
##   Group.1      x
## 1    1885 34.437
## 2    1886 30.045
## 3    1887 37.160
## 4    1888 47.916
## 5    1889 30.448
## 6    1890 41.618
```

rename variables

```
names(CalgaryYearlyPrecip)
```

```
## [1] "Group.1" "x"
```

```
names(CalgaryYearlyPrecip) <- c('year', 'totalprecip')
head(CalgaryYearlyPrecip)
```

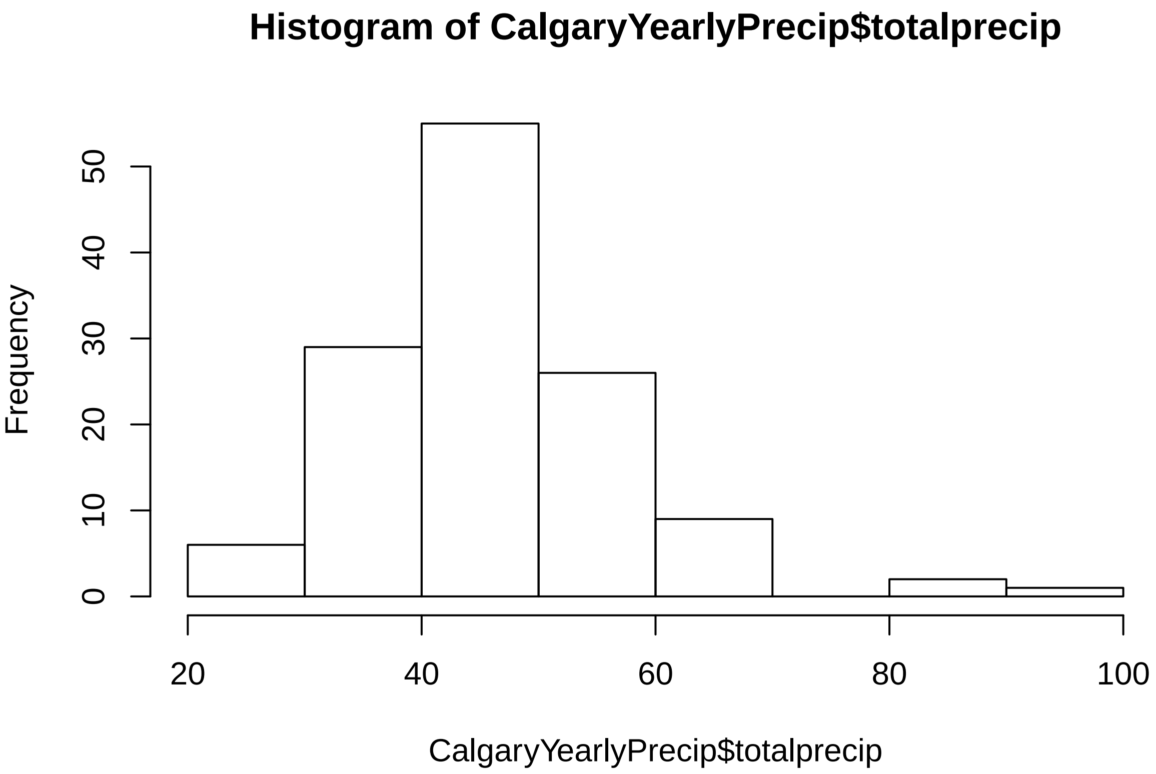
```
##   year totalprecip
## 1 1885      34.437
## 2 1886      30.045
## 3 1887      37.160
## 4 1888      47.916
## 5 1889      30.448
## 6 1890      41.618
```

saving data frame to a csv file

```
write.csv(CalgaryYearlyPrecip, file = 'CalgaryYearlyPrecip.csv',
          row.names = FALSE)
```

Statistics plot histogram

```
hist(CalgaryYearlyPrecip$totalprecip)
```



fit normal distribution

```
library(MASS)
?fitdistr
fit <- fitdistr(CalgaryYearlyPrecip$totalprecip, "normal")
fit
```

```
##       mean          sd    
##   46.5083828   11.1981374 
##  ( 0.9897849) ( 0.6998836)
```

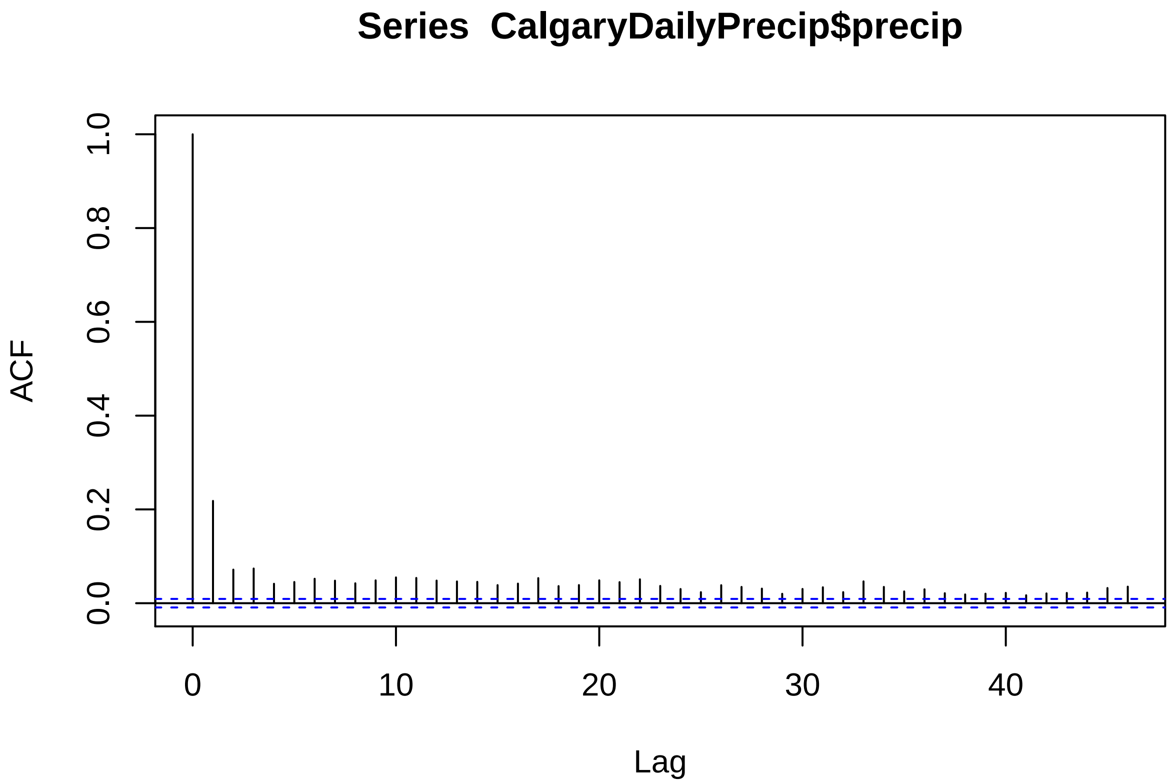
t-test

```
t <- t.test(CalgaryYearlyPrecip$totalprecip)
t
```

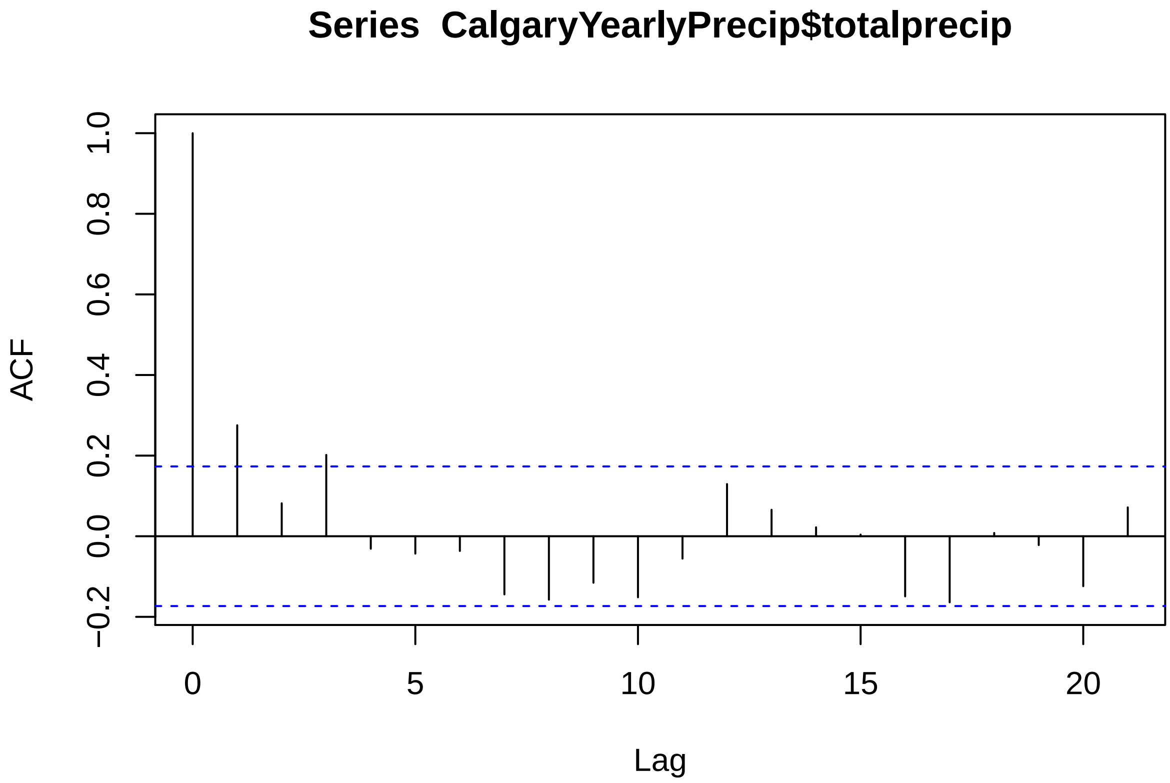
```
## 
##  One Sample t-test
## 
## data:  CalgaryYearlyPrecip$totalprecip
## t = 46.804, df = 127, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  44.54208 48.47468
## sample estimates:
## mean of x 
##  46.50838
```

plot autocorrelation function (ACF)

```
acf(CalgaryDailyPrecip$precip)
```



```r
acf(CalgaryYearlyPrecip$totalprecip)
```



Mann-Kendall test for trends

```r
library(Kendall)
?MannKendall
mk <- MannKendall(CalgaryYearlyPrecip$totalprecip)
summary(mk)
```

```
## Score =  1452 , Var(Score) = 235712
## denominator =  8128
## tau = 0.179, 2-sided pvalue =0.002802
```

linear regression model

```r
model <- lm(totalprecip~year, CalgaryYearlyPrecip)
summary(model)
```

```
##
## Call:
## lm(formula = totalprecip ~ year, data = CalgaryYearlyPrecip)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.951  -6.856  -0.272   4.343  48.093
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -65.13456   51.66893  -1.261   0.2098
## year          0.05730    0.02651   2.161   0.0326 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.08 on 126 degrees of freedom
## Multiple R-squared:  0.03574,    Adjusted R-squared:  0.02809
## F-statistic:  4.67 on 1 and 126 DF,  p-value: 0.03258
```

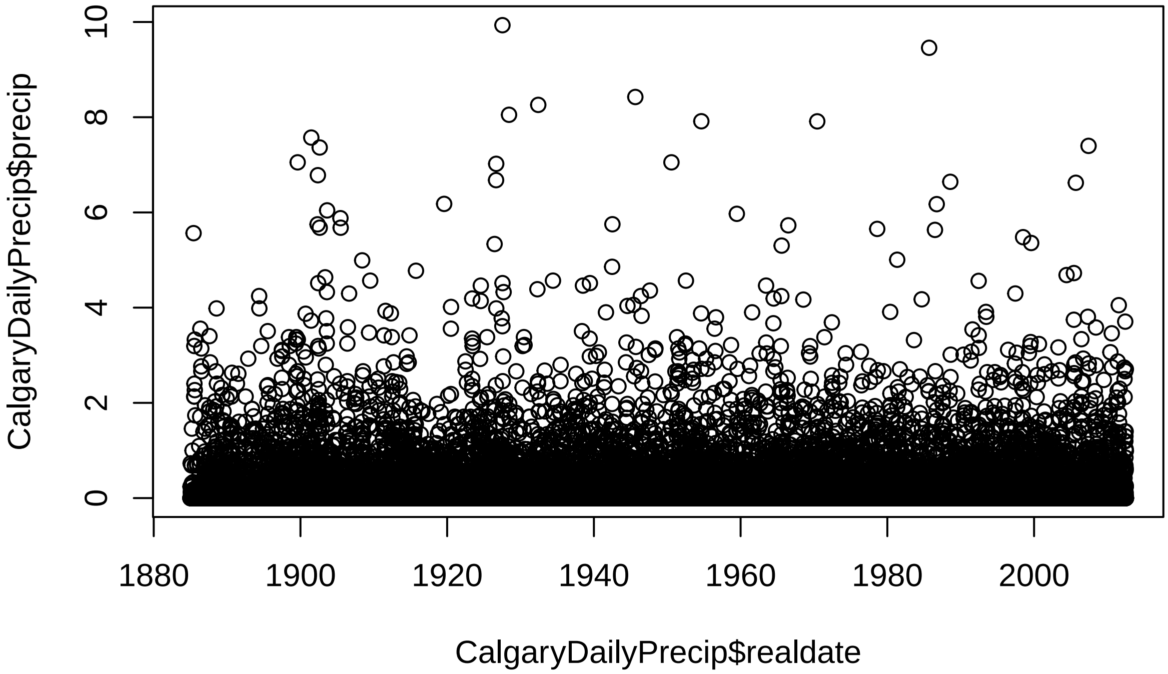
```r
coef(model)
```

```
##  (Intercept)         year
## -65.13455582   0.05729686
```

```r
# Graphing slides
```

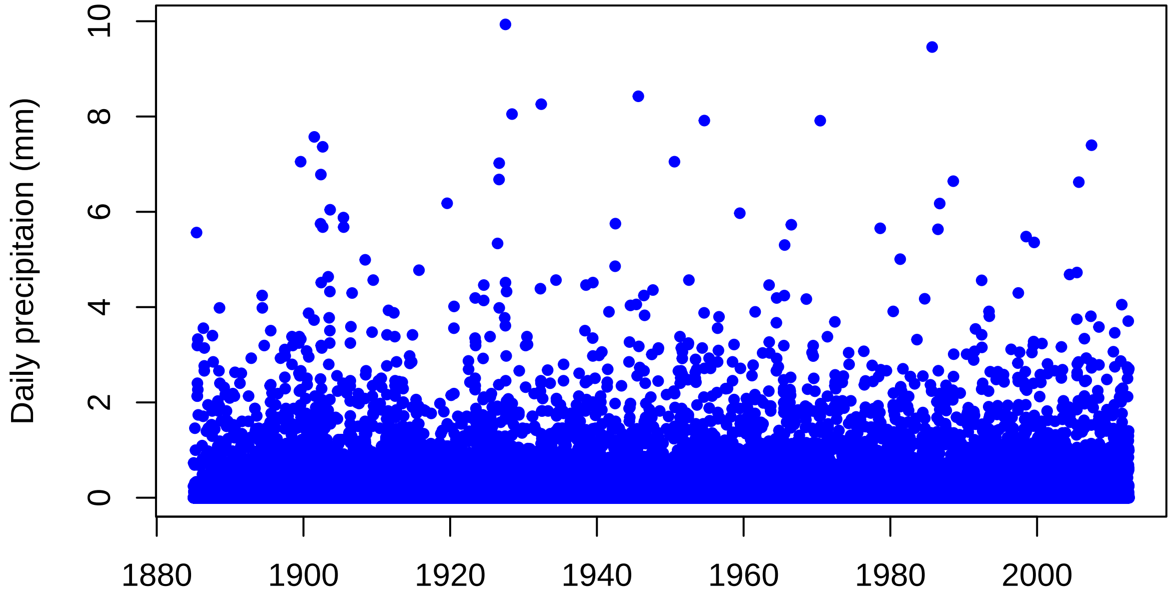
built-in graphing

```r
plot(CalgaryDailyPrecip$realdate, CalgaryDailyPrecip$precip)
```



change plot - requires replotting

```
plot(CalgaryDailyPrecip$realdate, CalgaryDailyPrecip$precip, xlab = "",
     ylab = "Daily precipitaion (mm)", pch = 20, col = 'blue')
```



ggplot2 graphing

```
annual <- read.csv("PrarieAnnualPrecip.csv")
summary(annual)
```

```
##         site          year       precipitation
##  Calgary  :113   Min.   :1895   Min.   :202.8
##  Regina   :110   1st Qu.:1925   1st Qu.:372.0
##  Saskatoon:106   Median :1953   Median :444.4
##                  Mean   :1953   Mean   :447.8
##                    3rd Qu.:1980   3rd Qu.:505.2
##                    Max.   :2007   Max.   :919.6
```

```
head(annual)
```

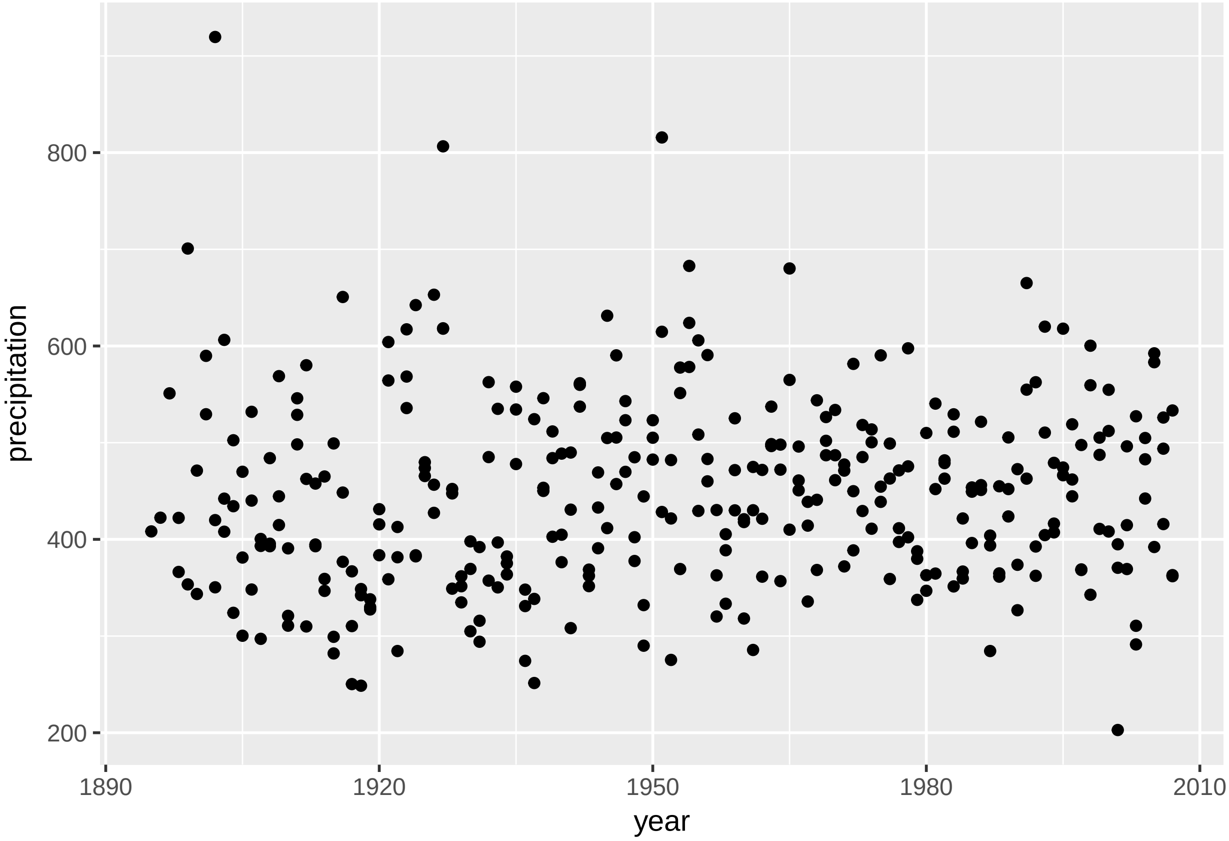
```
##      site year precipitation
## 1 Calgary 1895         408.2
## 2 Calgary 1896         422.4
## 3 Calgary 1897         551.0
## 4 Calgary 1898         422.2
## 5  Regina 1898         366.3
## 6 Calgary 1899         700.8
```

load library

```
library(ggplot2)
```

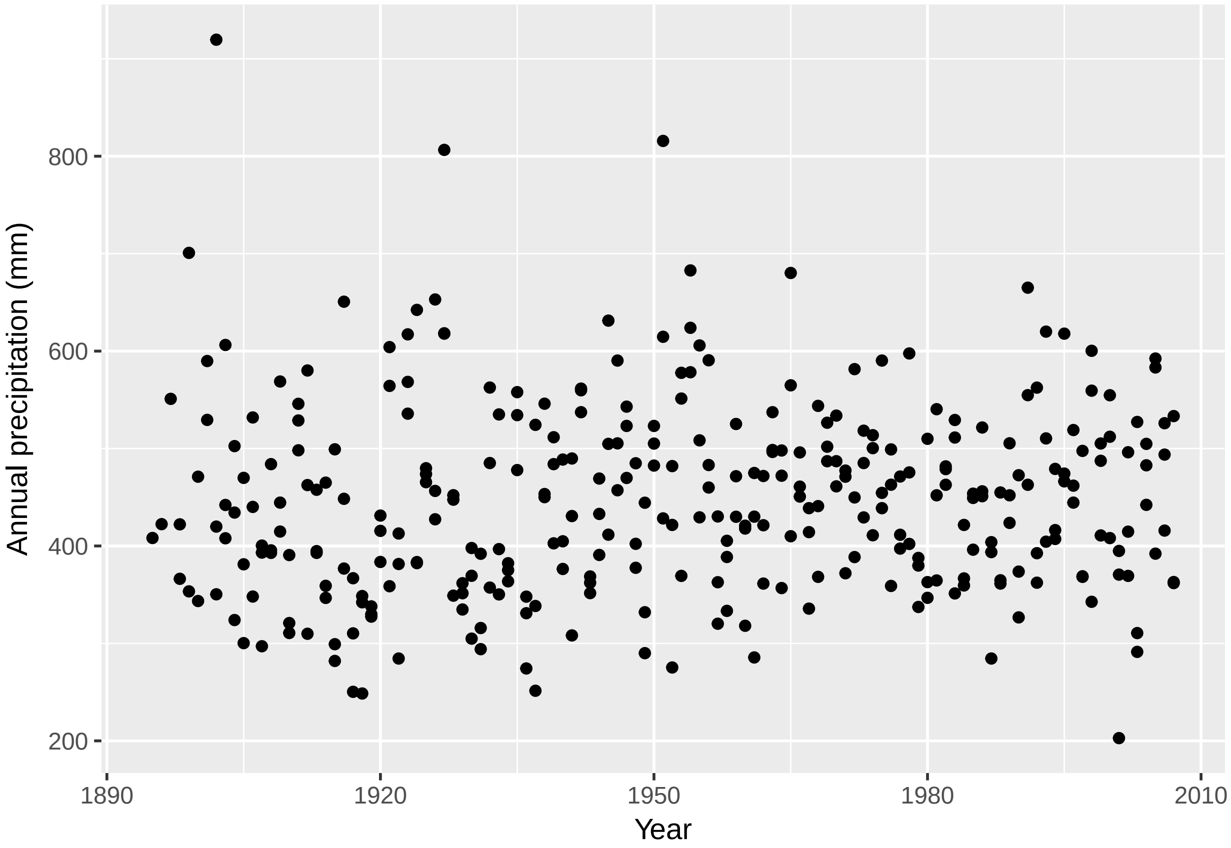
create basic xy graph

```
p <- ggplot(annual, aes(year, precipitation))
p <- p + geom_point()
p
```



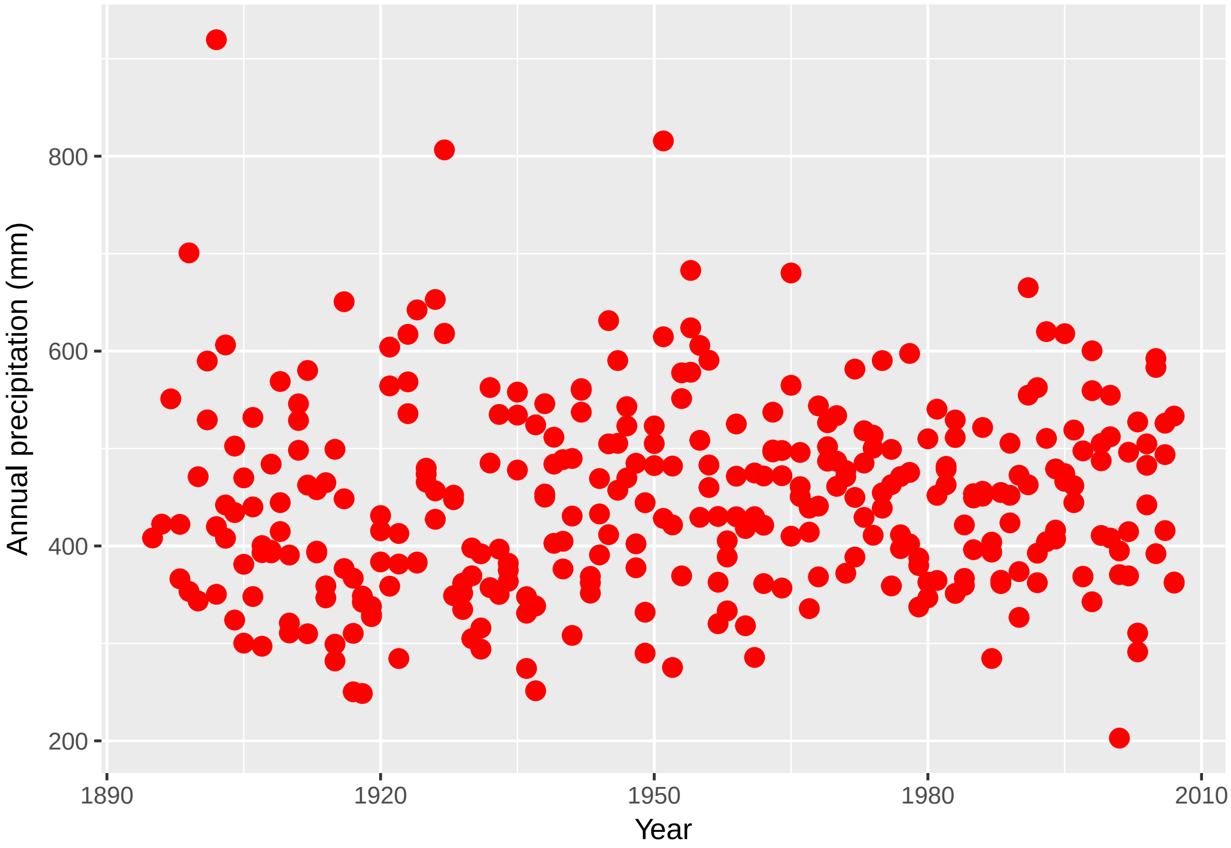
change titles & replot

```
p <- p + xlab('Year')
p <- p + ylab('Annual precipitation (mm)')
p
```



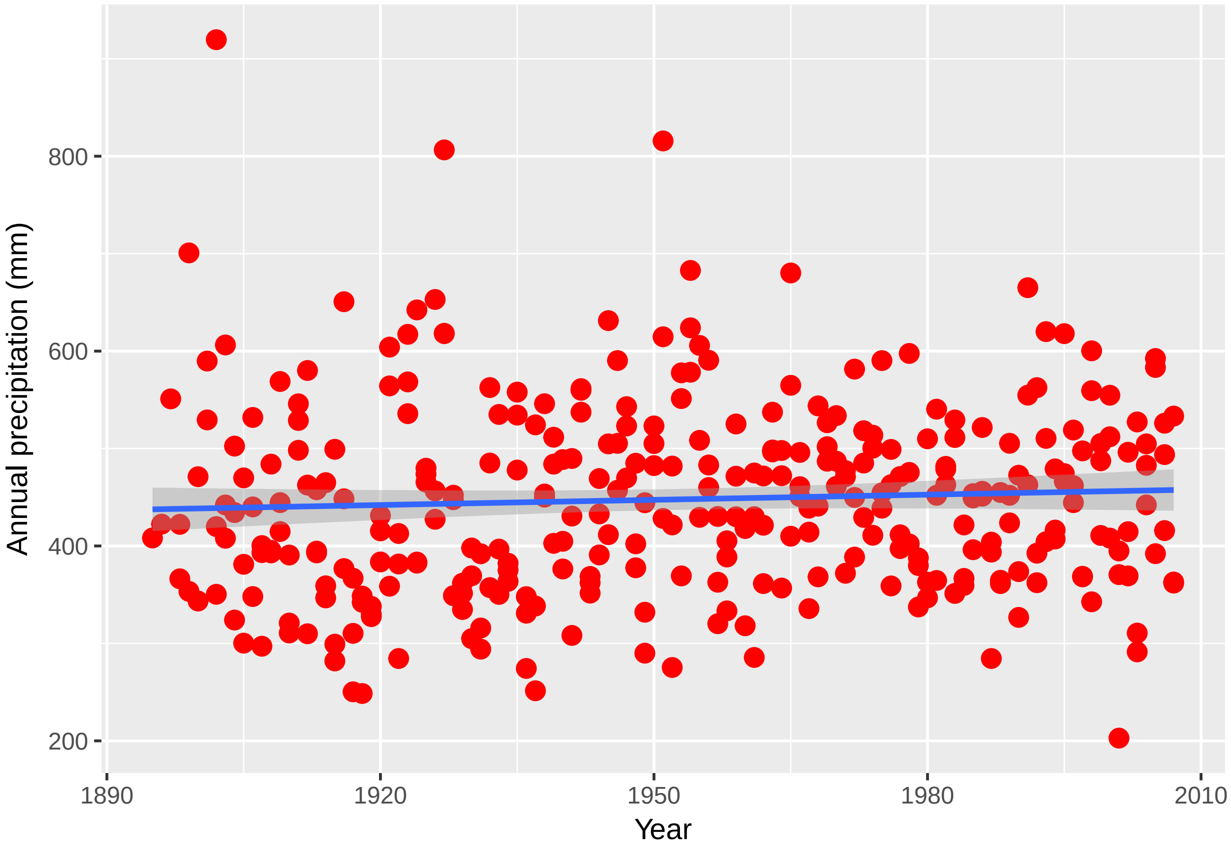
add colour to points and change size

```
p <- p + geom_point(colour = "red", size = 3)
p
```



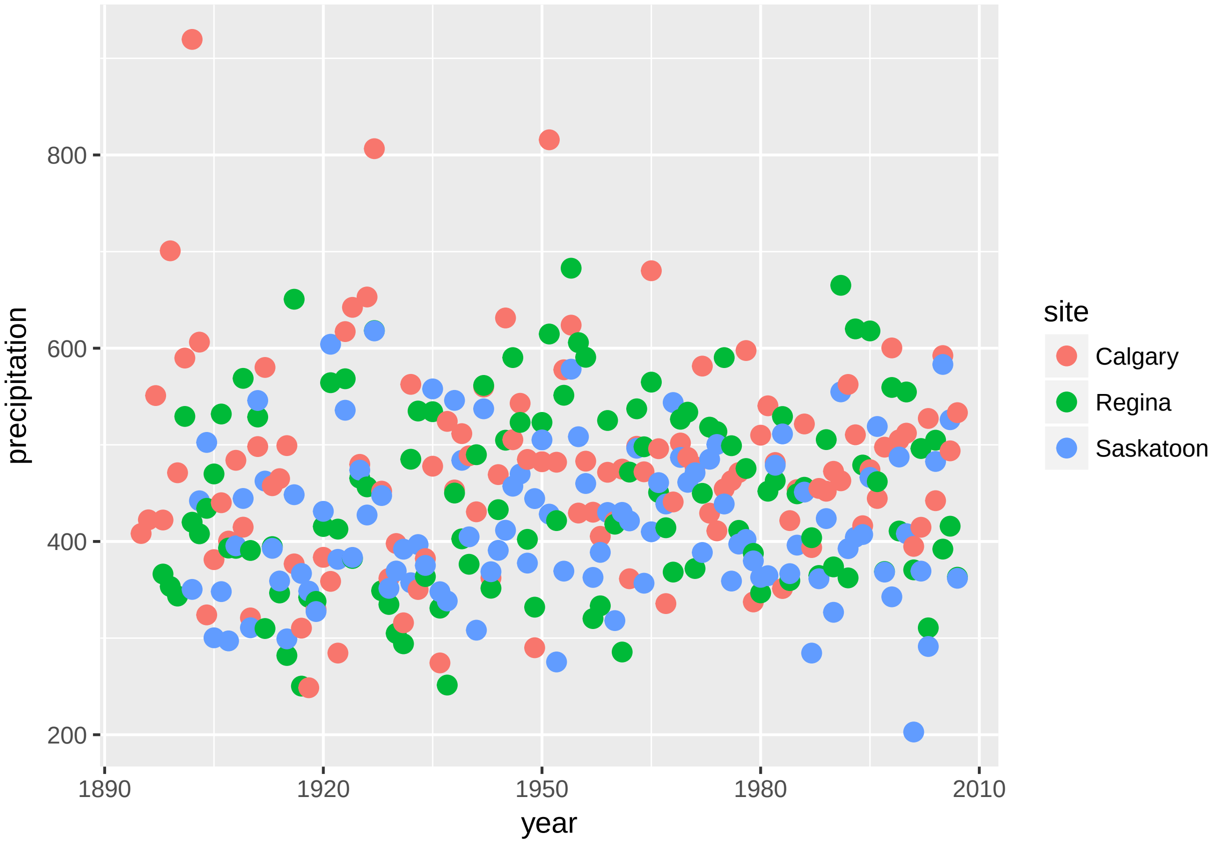
add regression curve

```
p <- p + stat_smooth(method = "lm")
p
```



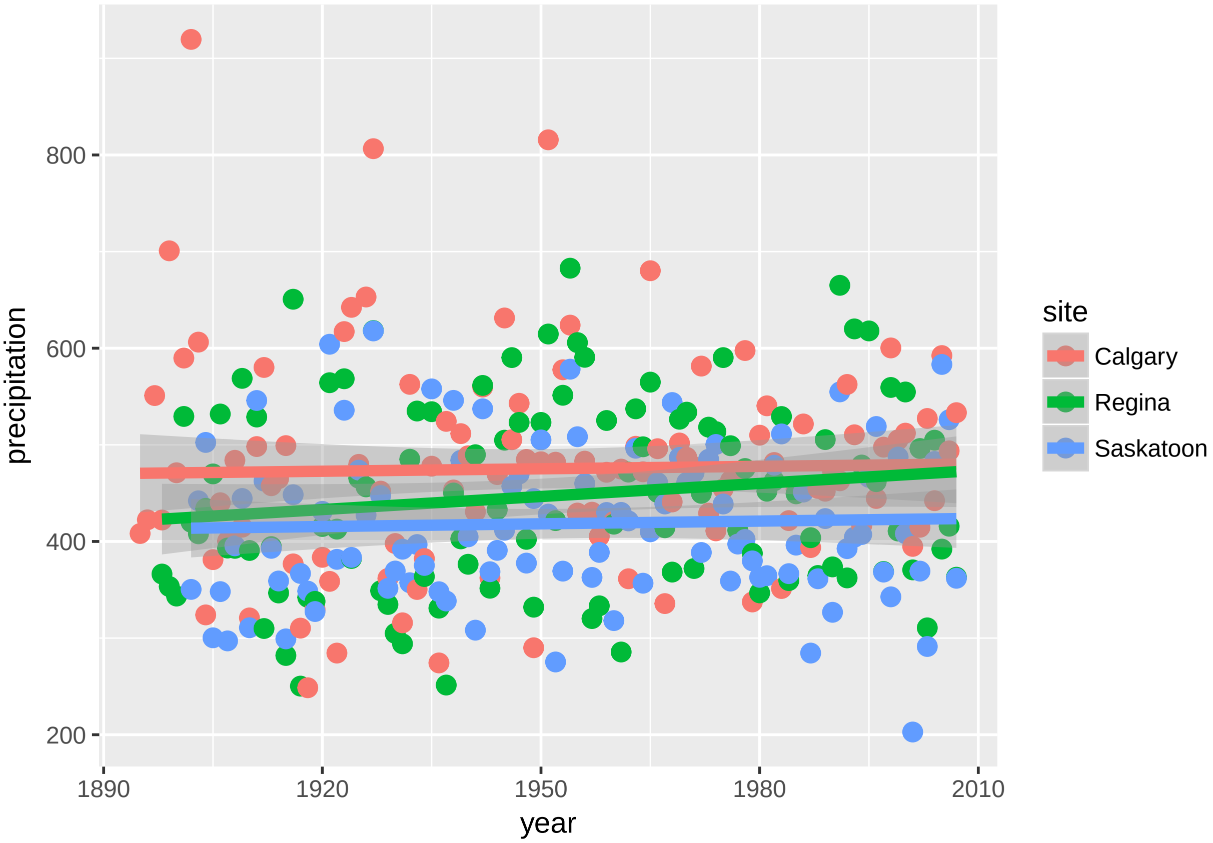
replot, mapping colours to variables

```
p2 <- ggplot(annual, aes(year, precipitation, colour = site))
p2 <- p2 + geom_point(size = 3)
p2
```



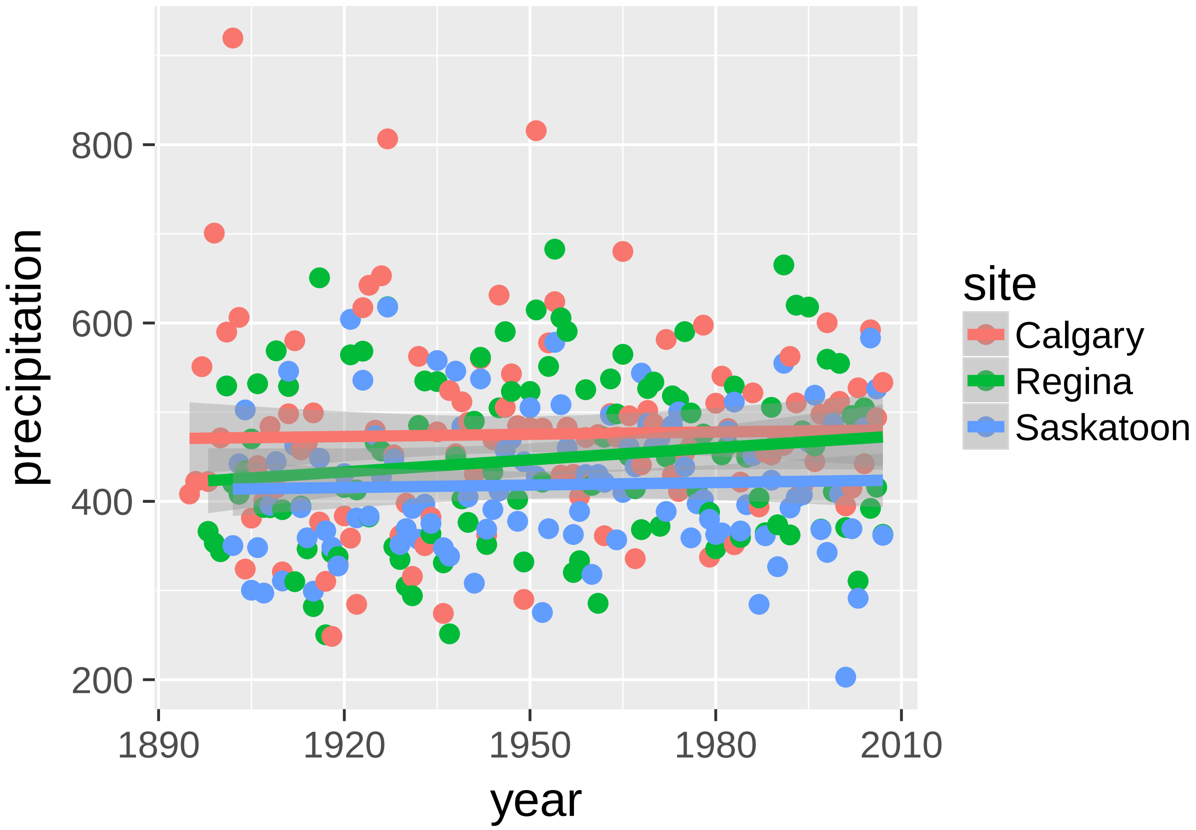
add regression curve to each category

```
p2 <- p2 + stat_smooth(method = "lm", size = 2)
p2
```



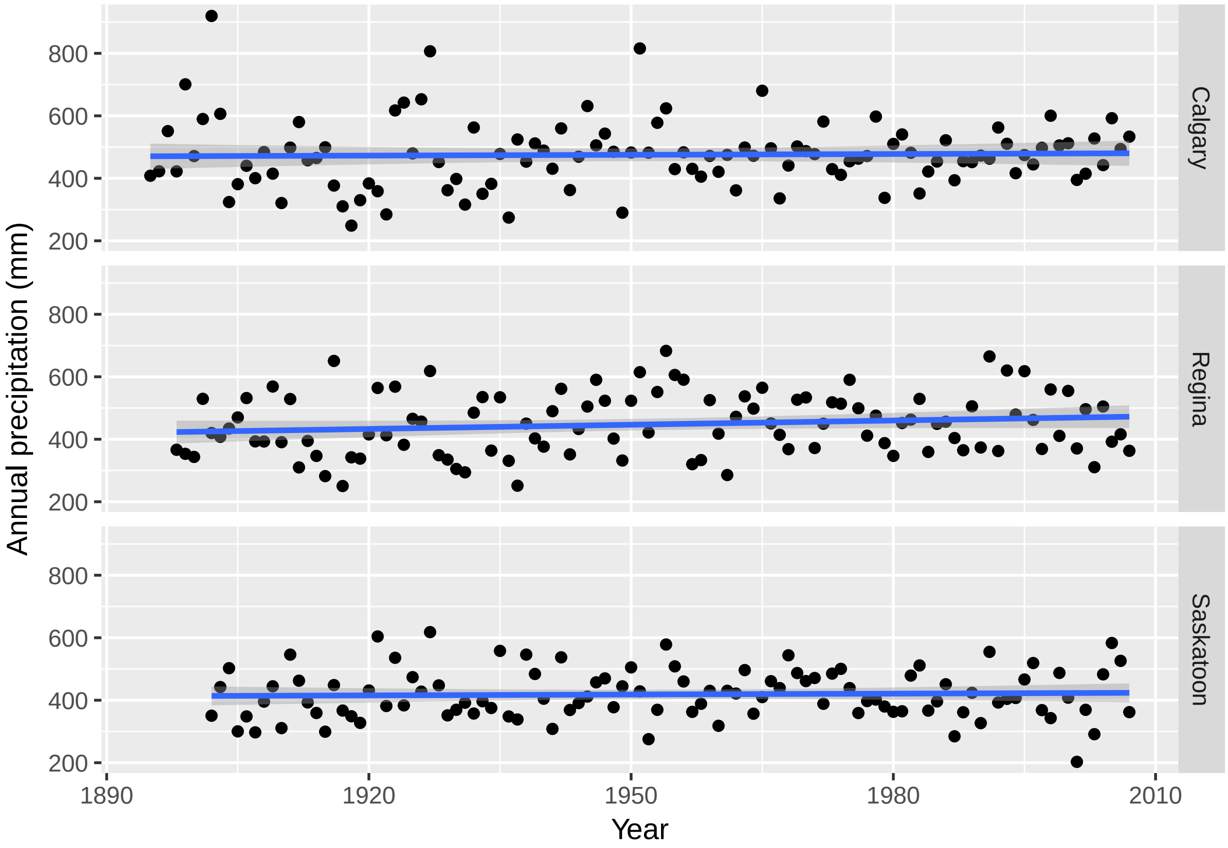
change theme font sizes

```
p2 <- p2 + theme_grey(base_size = 18)
p2
```



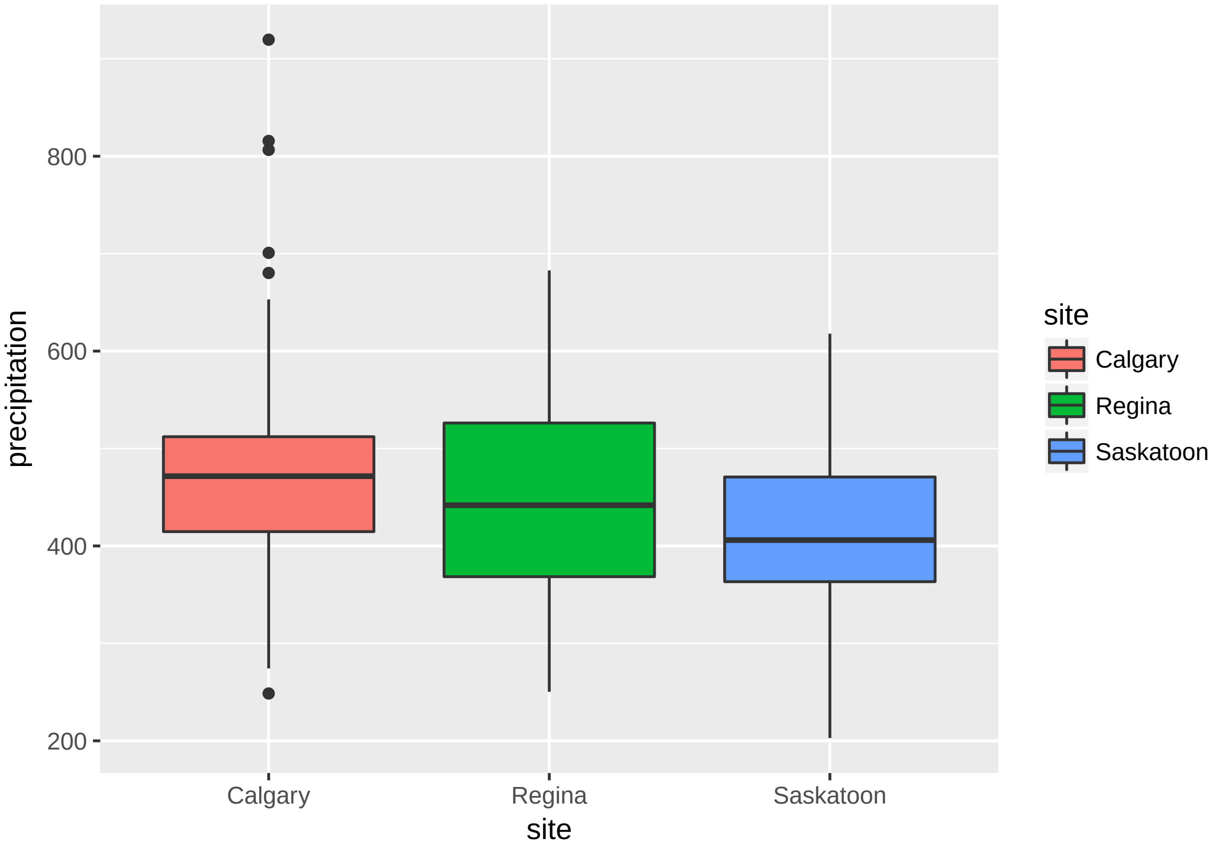
do faceting

```
p3 <- ggplot(annual, aes(year, precipitation))
p3 <- p3 + geom_point() + facet_grid(site ~ .)
p3 <- p3 + stat_smooth(method = "lm")
p3 <- p3 + xlab('Year')
p3 <- p3 + ylab('Annual precipitation (mm)')
p3
```



box plot

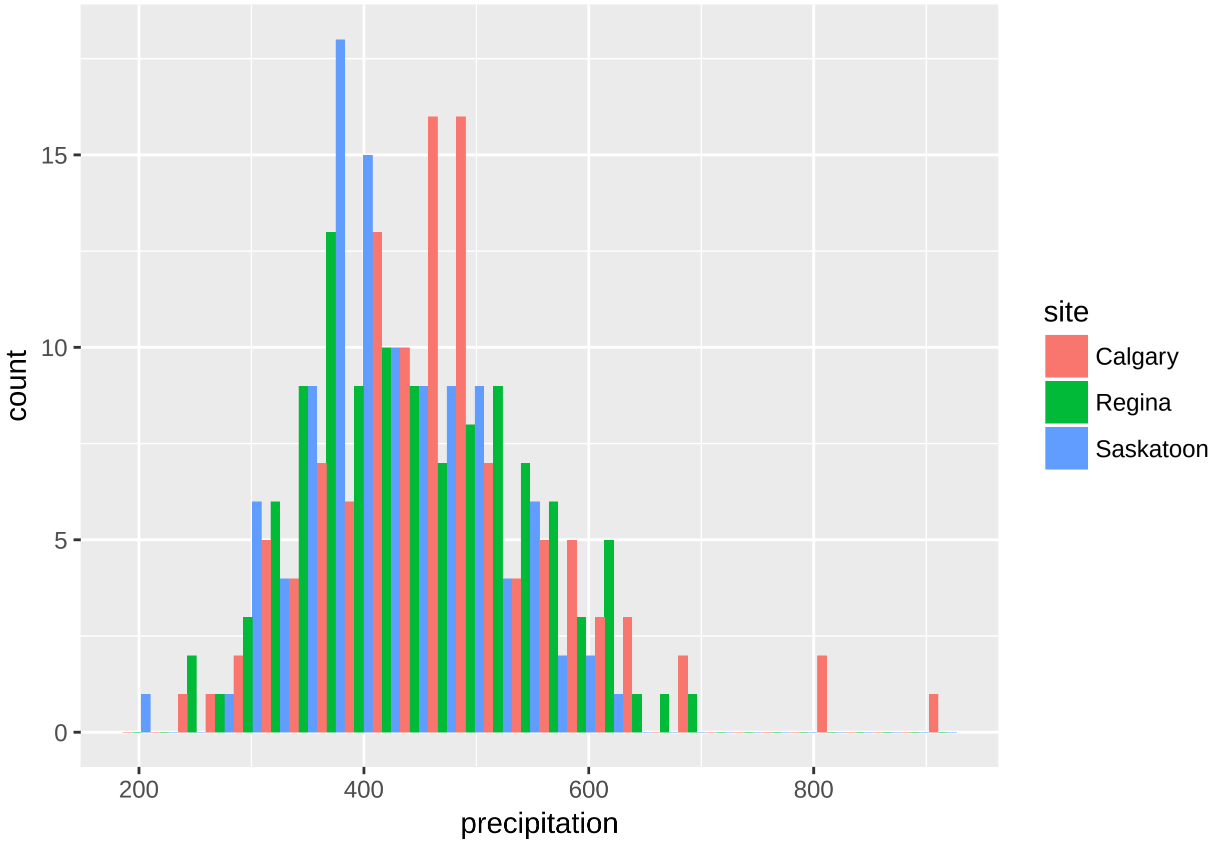
```
p4 <- ggplot(annual, aes(site, precipitation, fill = site))
p4 <- p4 + geom_boxplot()
p4
```



histograms

```
p5 <- ggplot(annual, aes(x = precipitation, fill = site))
p5 <- p5 + geom_histogram(position = 'dodge')
p5
```

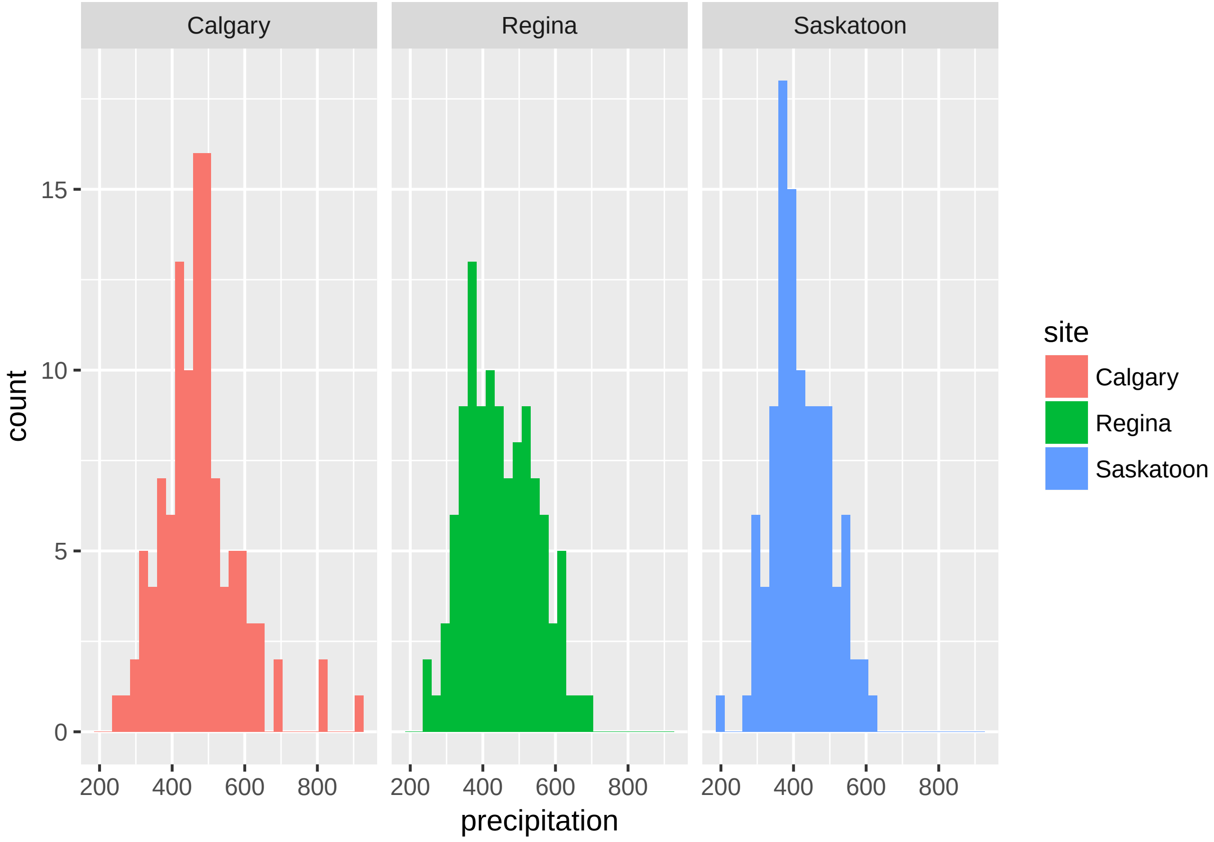
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



faceting

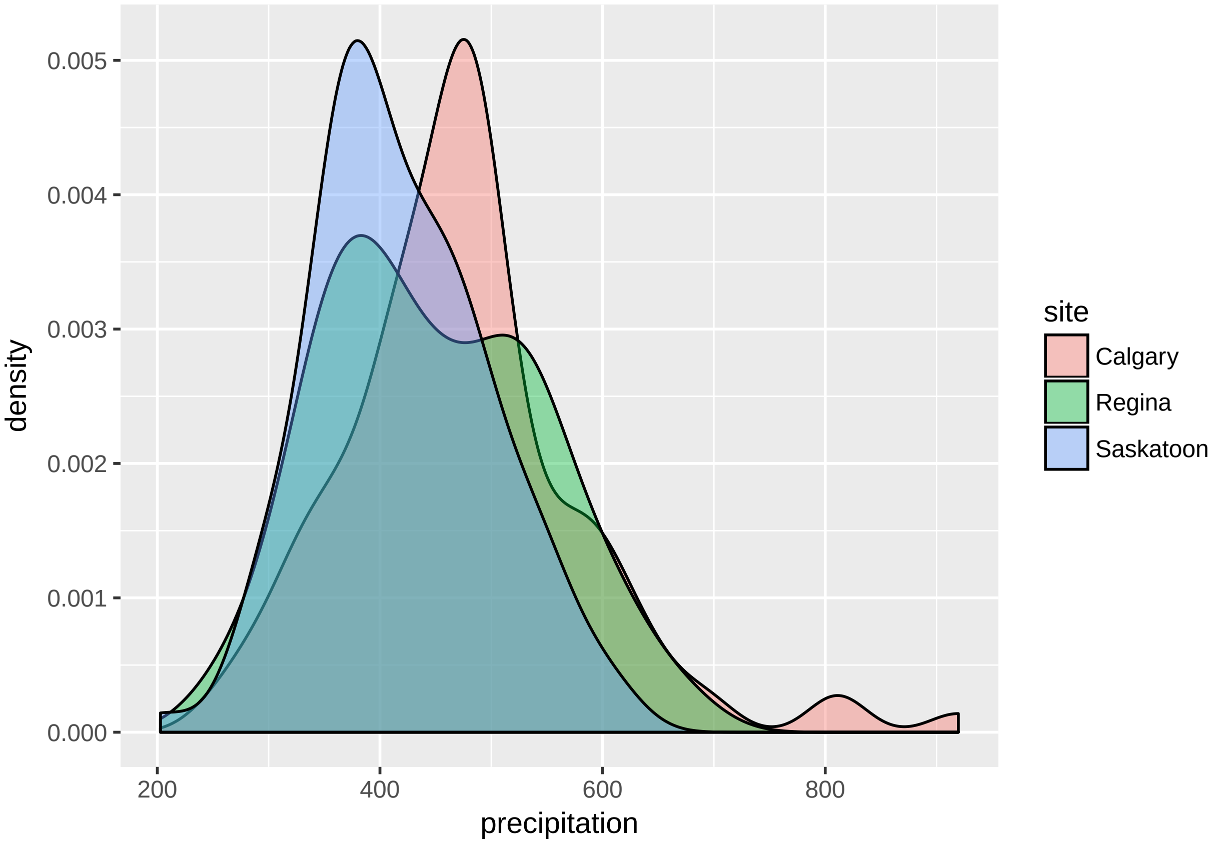
```
p5 <- p5 + facet_grid(. ~ site)
p5
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



density plots

```
p6 <- ggplot(annual, aes(x = precipitation, fill = site))
p6 <- p6 + geom_density(alpha = 0.4)
p6
```



save plot

```
ggsave('DensityPlot.png')
```

```
## Saving 6.5 x 4.5 in image
```

Final slides