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 1895/01/01      0
## 2 1895/02/01      5
## 3 1895/03/01      0
## 4 1895/04/01      0
```

```
## 5 1895/05/01      0
## 6 1895/06/01      0
```

```
summary(CalgaryDailyPrecip)
```

```
##      date               precip
##  Length:41273       Min.   :  0.00
##  Class :character   1st Qu.:  0.00
##  Mode  :character   Median :  0.00
##                     Mean   : 13.04
##                     3rd Qu.:  5.00
##                     Max.   :993.00
##                     NA's   :95
```

```
nrow(CalgaryDailyPrecip)
```

```
## [1] 41273
```

```
ncol(CalgaryDailyPrecip)   # number of col
```

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

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

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

convert from 0.1 mm to mm

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

```
##      date               precip
##  Length:41273       Min.   : 0.000
##  Class :character   1st Qu.: 0.000
##  Mode  :character   Median : 0.000
##                     Mean   : 1.304
##                     3rd Qu.: 0.500
##                     Max.   :99.300
##                     NA's   :95
```

calculate mean

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

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

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

```
## [1] 1.304325
```

convert date string to a real date

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

```
##         date precip realdate
## 1 1895/01/01    0.0     <NA>
## 2 1895/02/01    0.5     <NA>
## 3 1895/03/01    0.0     <NA>
## 4 1895/04/01    0.0     <NA>
```

```
## 5 1895/05/01    0.0     <NA>
## 6 1895/06/01    0.0     <NA>
```

```
summary(CalgaryDailyPrecip)
```

```
##      date               precip          realdate
##  Length:41273       Min.   : 0.000   Min.   :1895-01-13
##  Class :character   1st Qu.: 0.000   1st Qu.:1923-04-14
##  Mode  :character   Median : 0.000   Median :1951-07-14
##                     Mean   : 1.304   Mean   :1951-07-09
##                     3rd Qu.: 0.500   3rd Qu.:1979-10-13
##                     Max.   :99.300   Max.   :2007-12-31
##                     NA's   :95       NA's   :16273
```

remove all missing values

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

```
##      date               precip          realdate
##  Length:24943       Min.   : 0.000   Min.   :1895-01-13
##  Class :character   1st Qu.: 0.000   1st Qu.:1923-03-19
##  Mode  :character   Median : 0.000   Median :1951-05-23
##                     Mean   : 1.285   Mean   :1951-05-23
##                     3rd Qu.: 0.500   3rd Qu.:1979-07-26
##                     Max.   :99.300   Max.   :2007-09-30
```

get year

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

```
##      date               precip          realdate              year
##  Length:24943       Min.   : 0.000   Min.   :1895-01-13   Min.   :1895
##  Class :character   1st Qu.: 0.000   1st Qu.:1923-03-19   1st Qu.:1923
##  Mode  :character   Median : 0.000   Median :1951-05-23   Median :1951
##                     Mean   : 1.285   Mean   :1951-05-23   Mean   :1951
##                     3rd Qu.: 0.500   3rd Qu.:1979-07-26   3rd Qu.:1979
##                     Max.   :99.300   Max.   :2007-09-30   Max.   :2007
```

subset by year

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

```
##             date precip   realdate year
## 40921 13/01/2007    1.9 2007-01-13 2007
## 40922 14/01/2007    0.2 2007-01-14 2007
## 40923 15/01/2007    0.0 2007-01-15 2007
## 40924 16/01/2007    0.0 2007-01-16 2007
## 40925 17/01/2007    5.8 2007-01-17 2007
## 40926 18/01/2007    0.0 2007-01-18 2007
```

or

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

```
##             date precip   realdate year
## 40191 13/01/2005    0.2 2005-01-13 2005
```

```
## 40192 14/01/2005    0.0 2005-01-14 2005
## 40193 15/01/2005    0.0 2005-01-15 2005
## 40194 16/01/2005    0.0 2005-01-16 2005
## 40195 17/01/2005    0.0 2005-01-17 2005
## 40196 18/01/2005    0.0 2005-01-18 2005
```

aggregate by year

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

```
##   Group.1     x
## 1    1895 252.2
## 2    1896 260.8
## 3    1897 307.4
## 4    1898 263.5
## 5    1899 461.0
## 6    1900 316.9
```

rename variables

```
names(CalgaryYearlyPrecip)
```

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

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

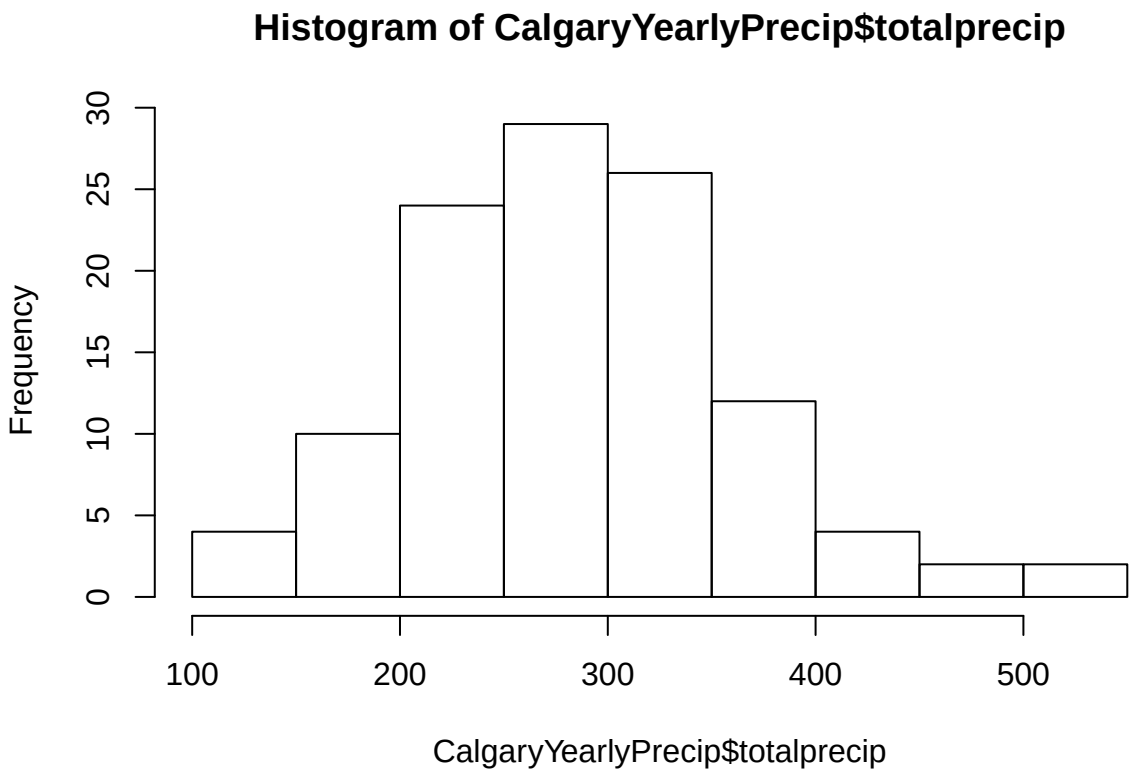
```
##   year totalprecip
## 1 1895       252.2
## 2 1896       260.8
## 3 1897       307.4
## 4 1898       263.5
## 5 1899       461.0
## 6 1900       316.9
```

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
##   283.664602    78.174716
##  (  7.354059) (  5.200105)
```

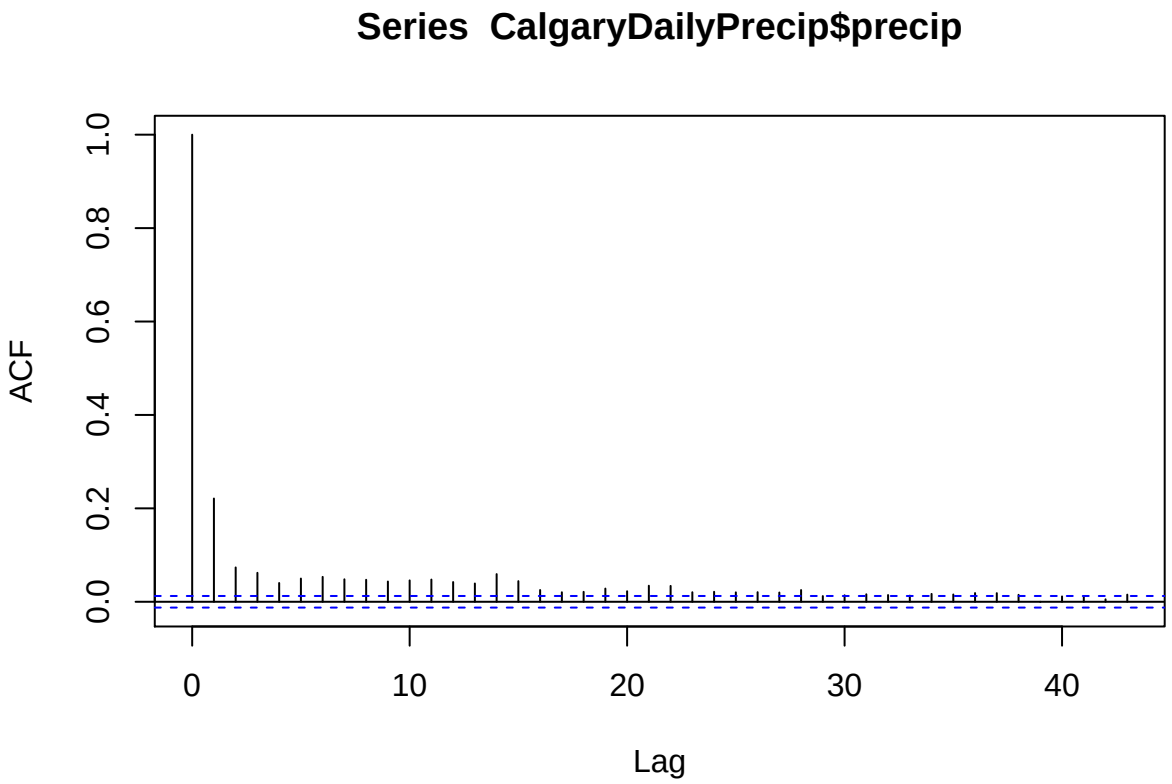
t-test

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

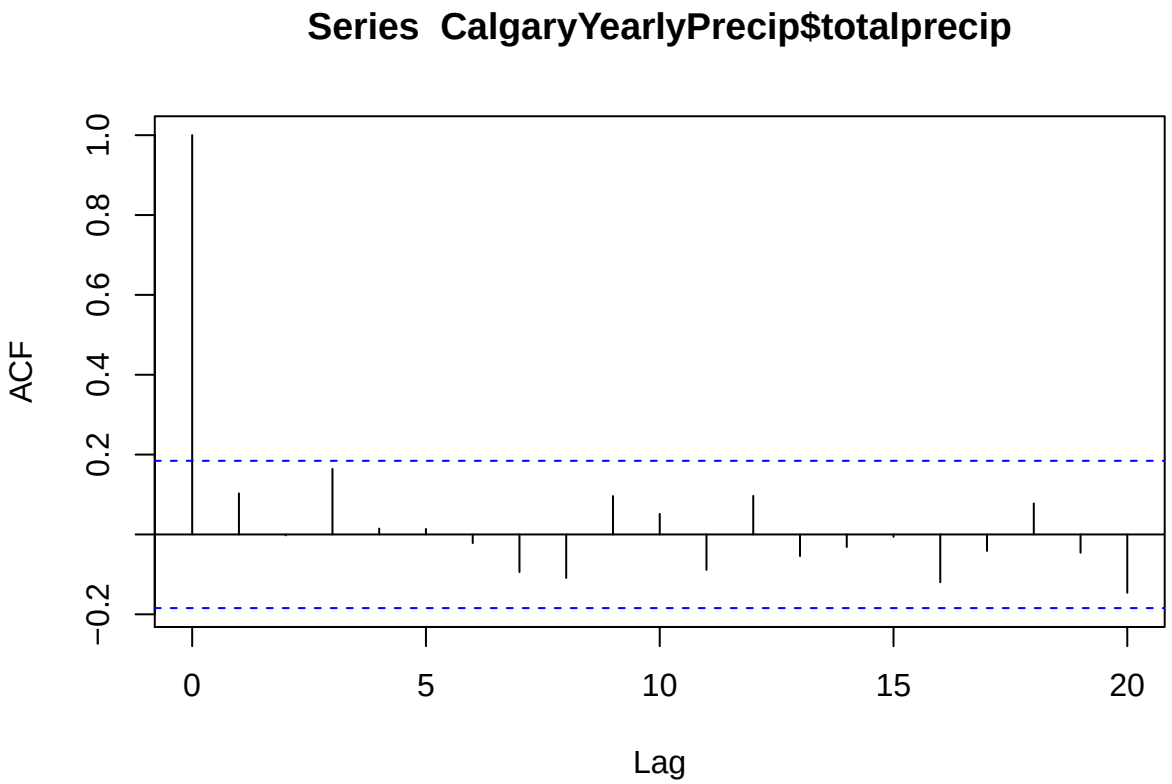
```
##
##  One Sample t-test
##
## data:  CalgaryYearlyPrecip$totalprecip
## t = 38.401, df = 112, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  269.0286 298.3006
## sample estimates:
## mean of x
##  283.6646
```

plot autocorrelation function (ACF)

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




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



Mann-Kendall test for trends

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

```
## Score =  24 , Var(Score) = 162414.7
## denominator =  6326
## tau = 0.00379, 2-sided pvalue =0.95449
```

linear regression model

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

```
##
## Call:
## lm(formula = totalprecip ~ year, data = CalgaryYearlyPrecip)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -173.042  -52.663   -5.014   38.280  255.135
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 318.86250  443.85351   0.718    0.474
## year         -0.01804    0.22747  -0.079    0.937
##
## Residual standard error: 78.87 on 111 degrees of freedom
## Multiple R-squared:  5.667e-05,  Adjusted R-squared:  -0.008952
## F-statistic: 0.00629 on 1 and 111 DF,  p-value: 0.9369
```

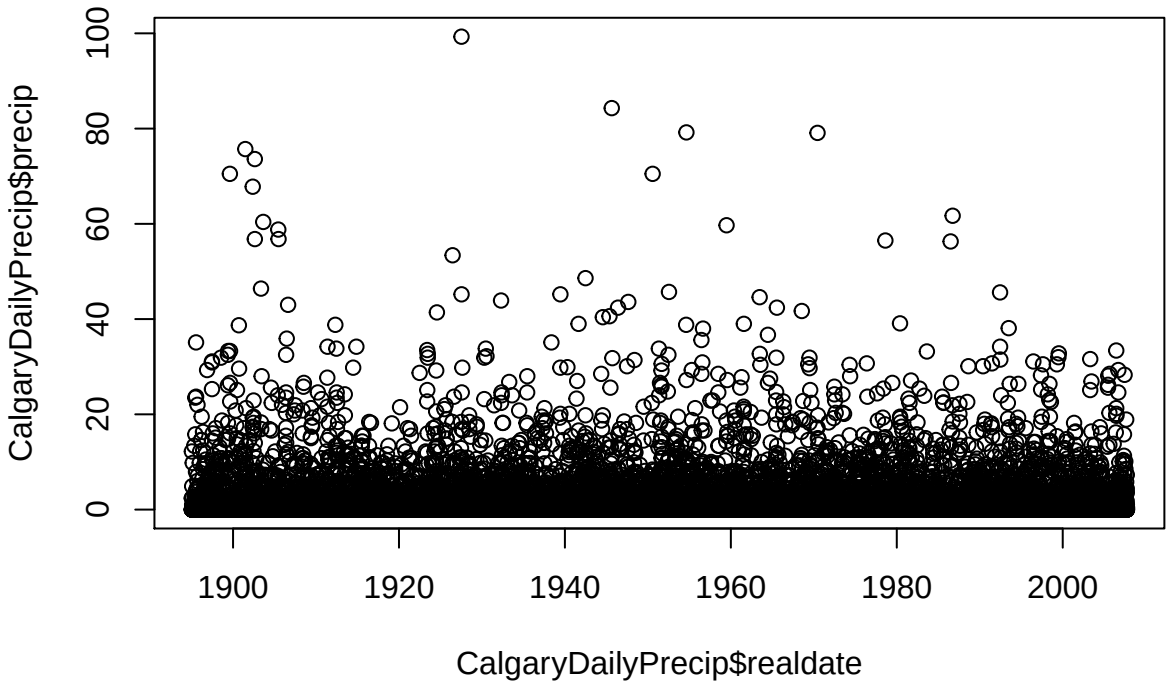
```
coef(model)
```

```
##  (Intercept)         year
## 318.86250333  -0.01804095
```

```
# Graphing slides
```

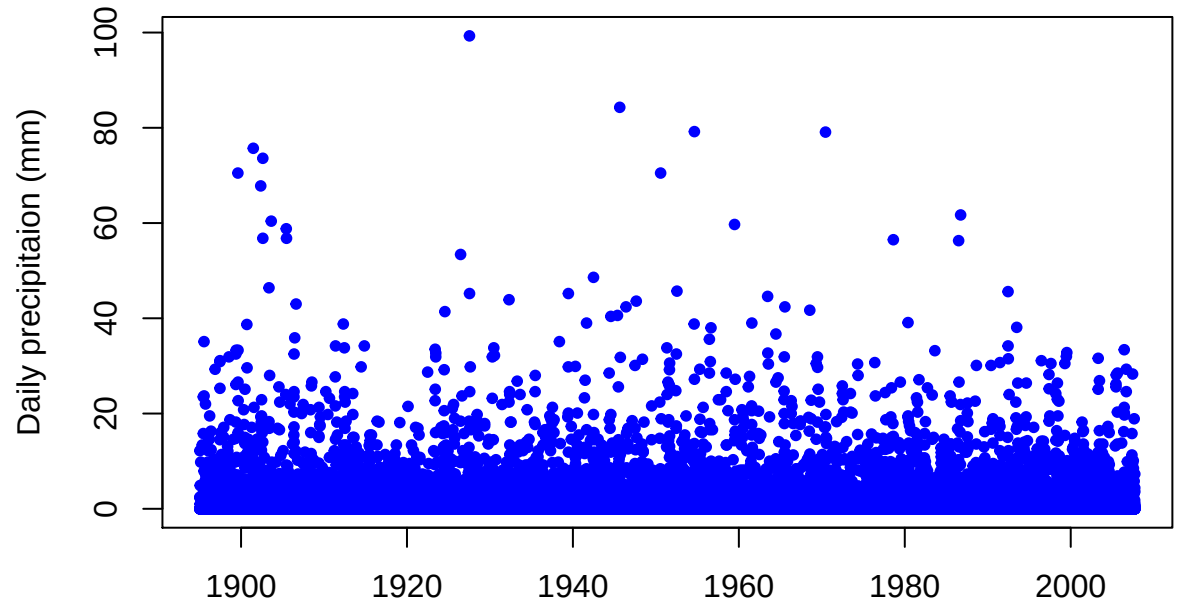
built-in graphing

```
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)
```

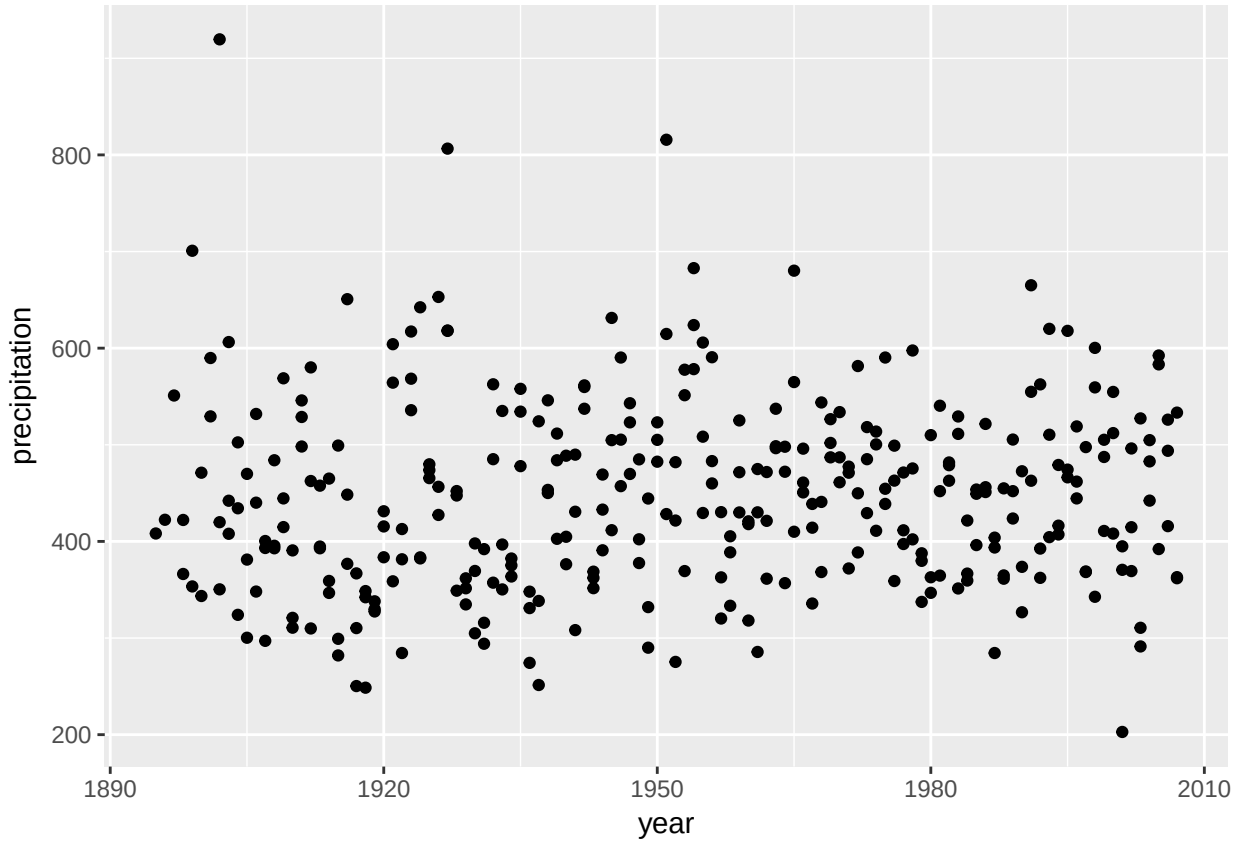
```
##      site year precipitation
## 1 Calgary 1895         408.2
## 2 Calgary 1896         422.4
## 3 Calgary 1897         551.0
## 4 Calgary 1898         422.2
## 5  Regina 1898         366.3
## 6 Calgary 1899         700.8
```

load library

```
library(ggplot2)
```

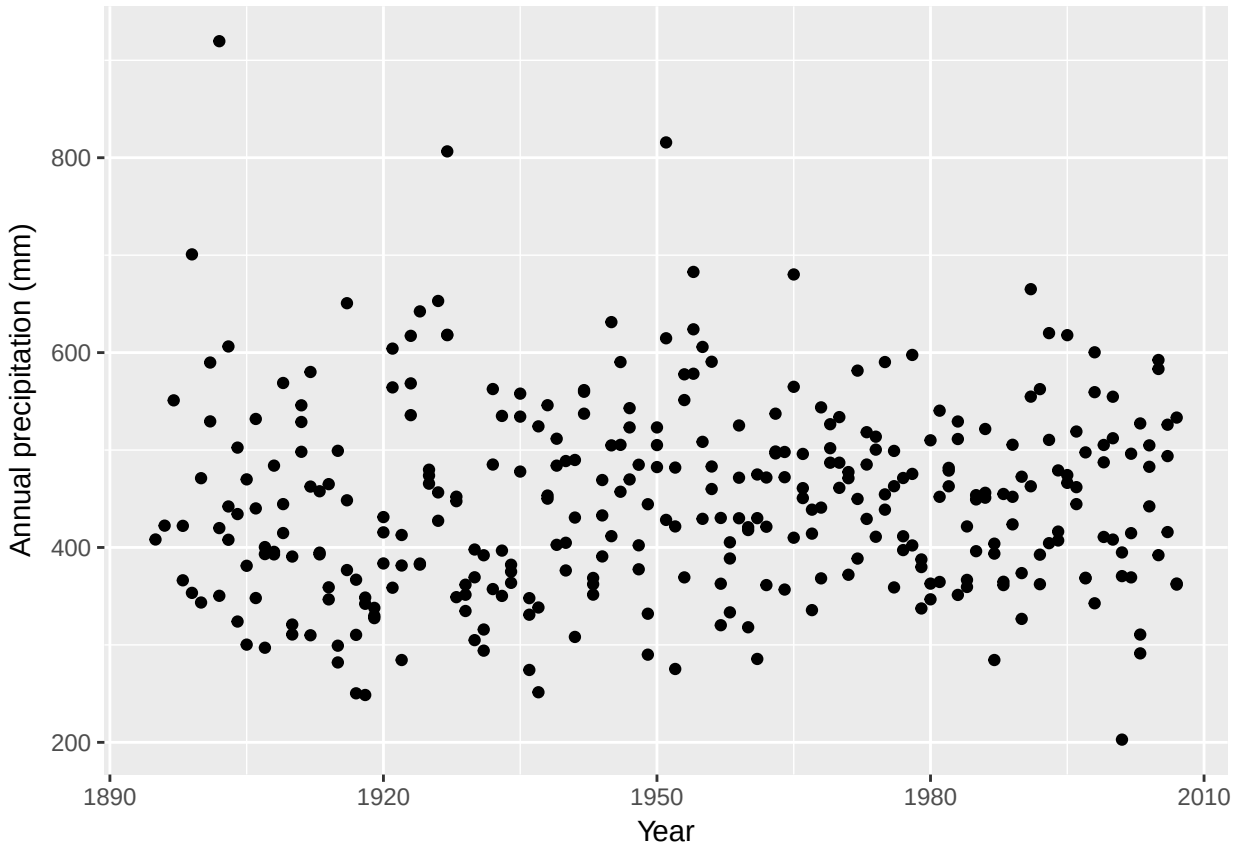
create basic xy graph

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



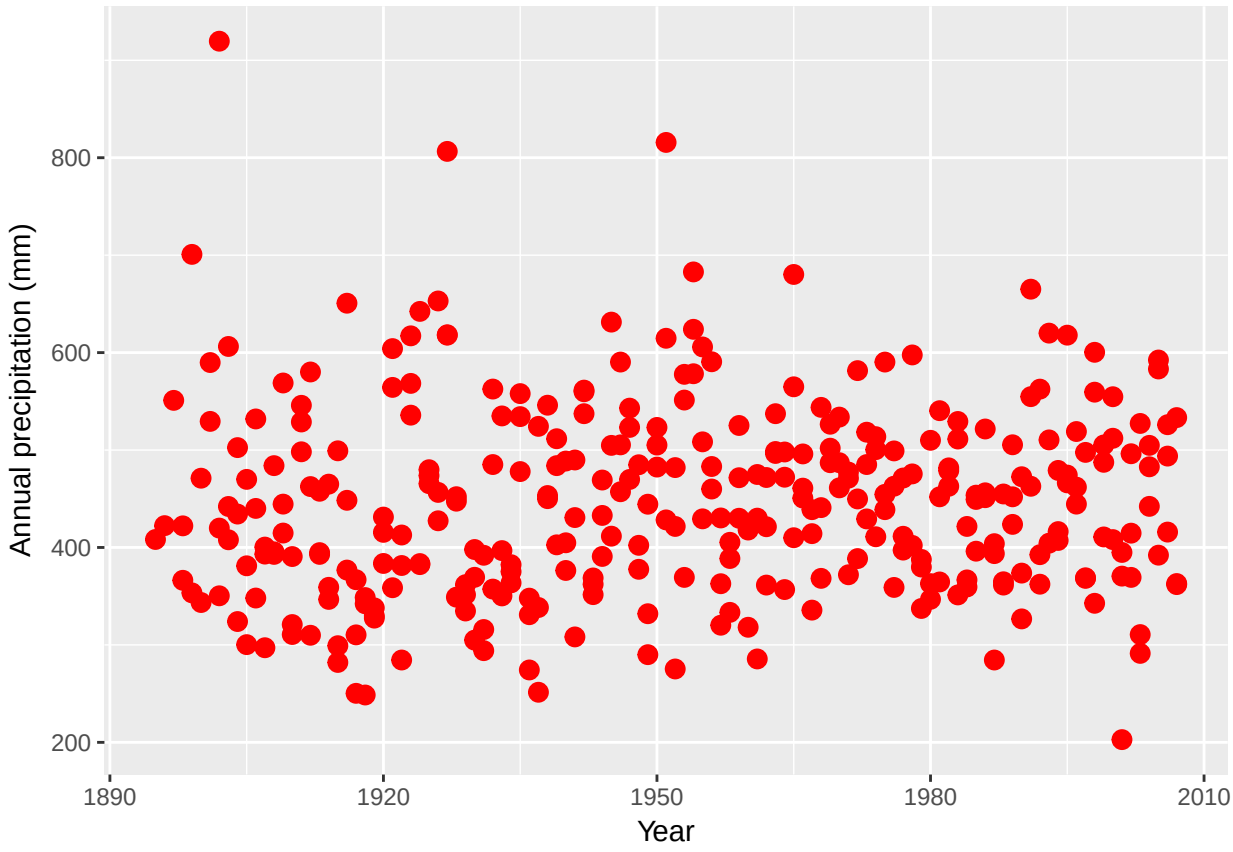
change titles & replot

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



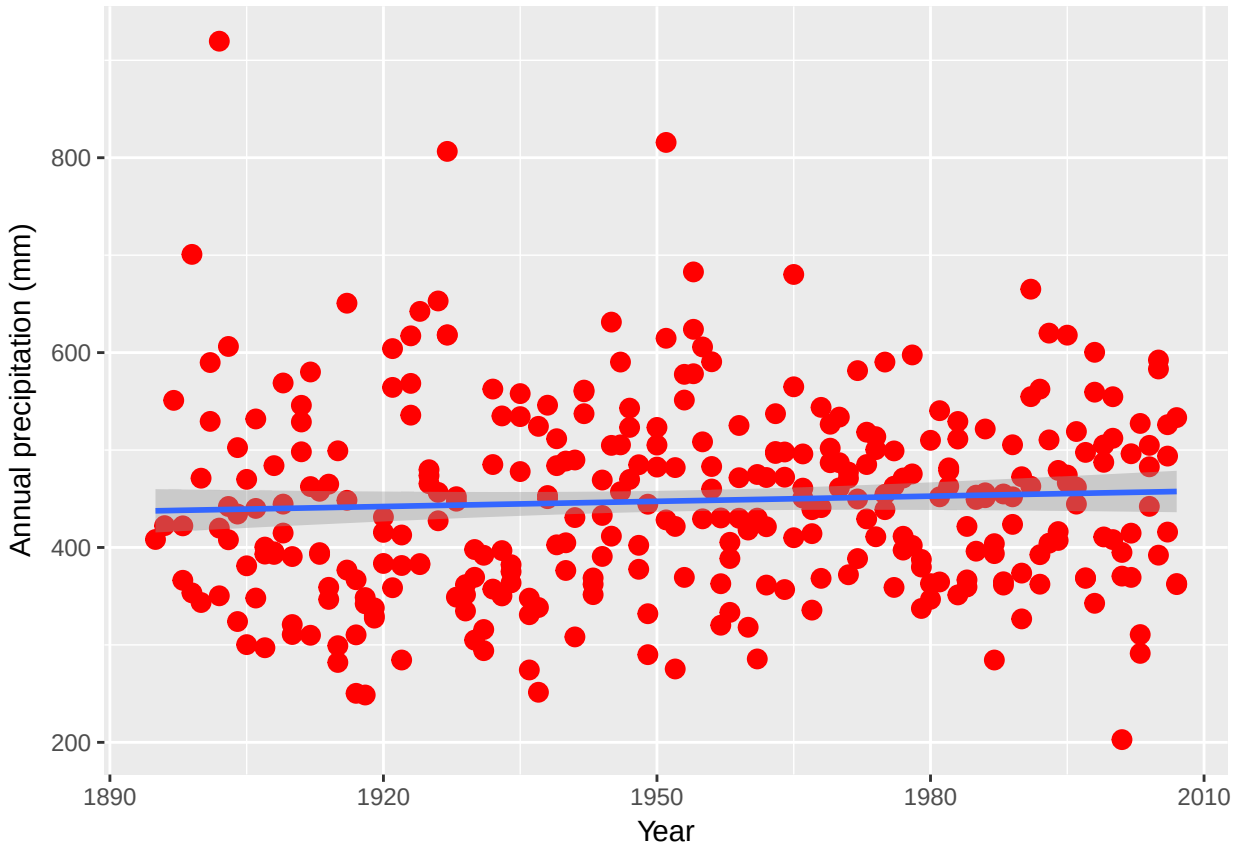
add colour to points and change size

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



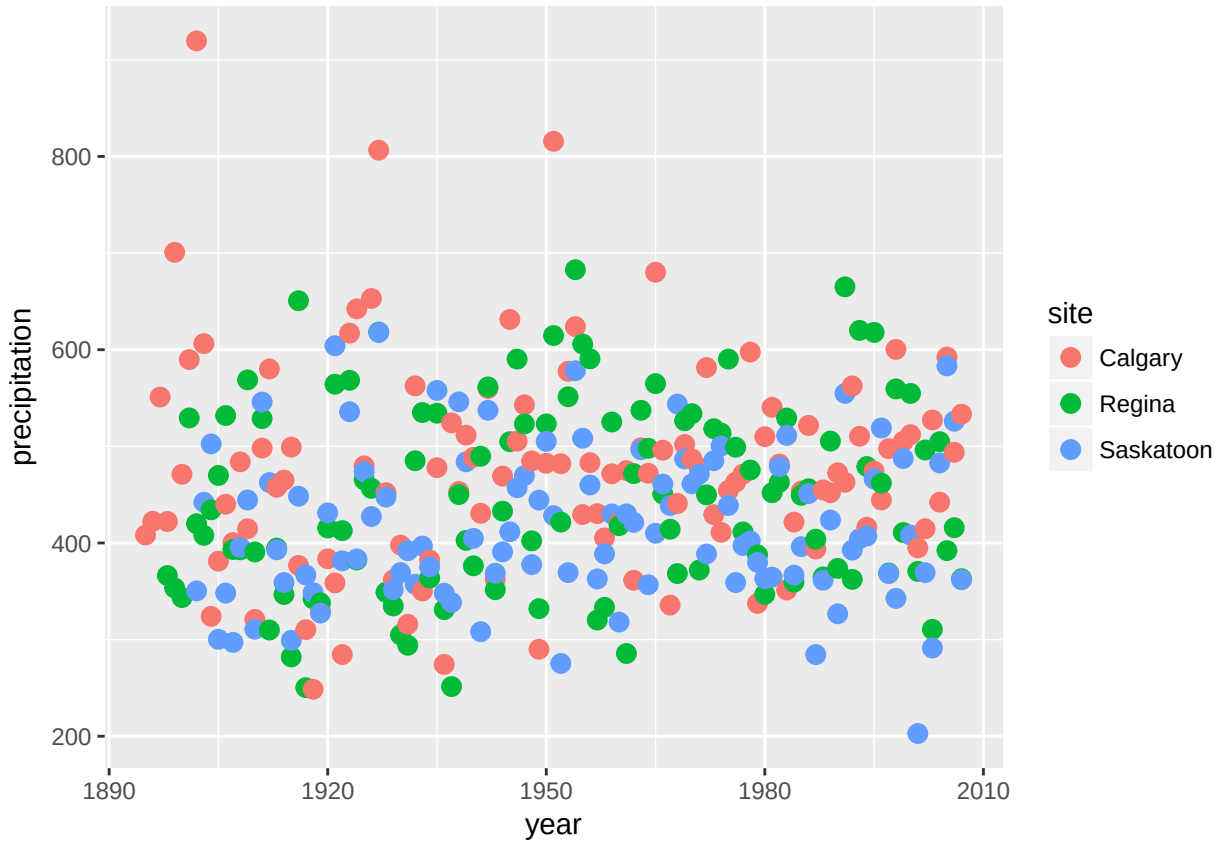
add regression curve

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



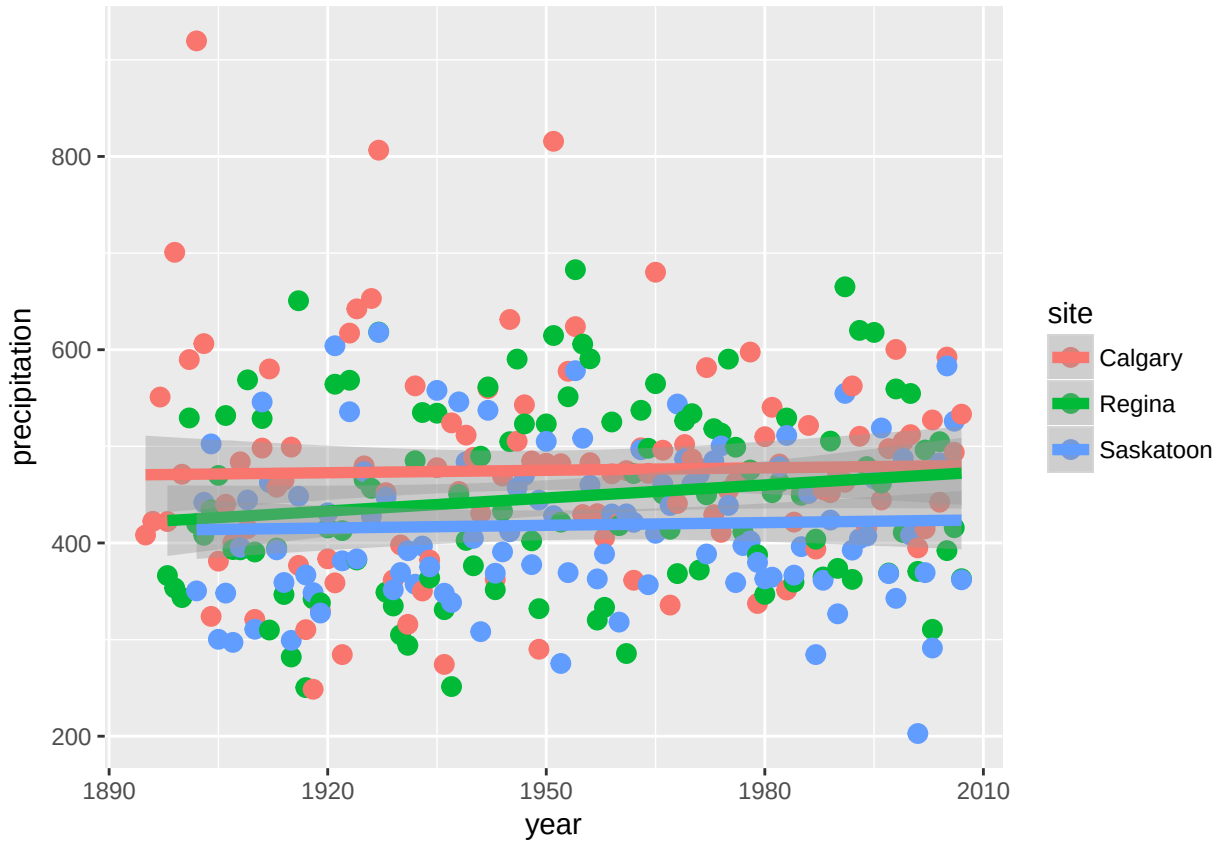
replot, mapping colours to variables

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



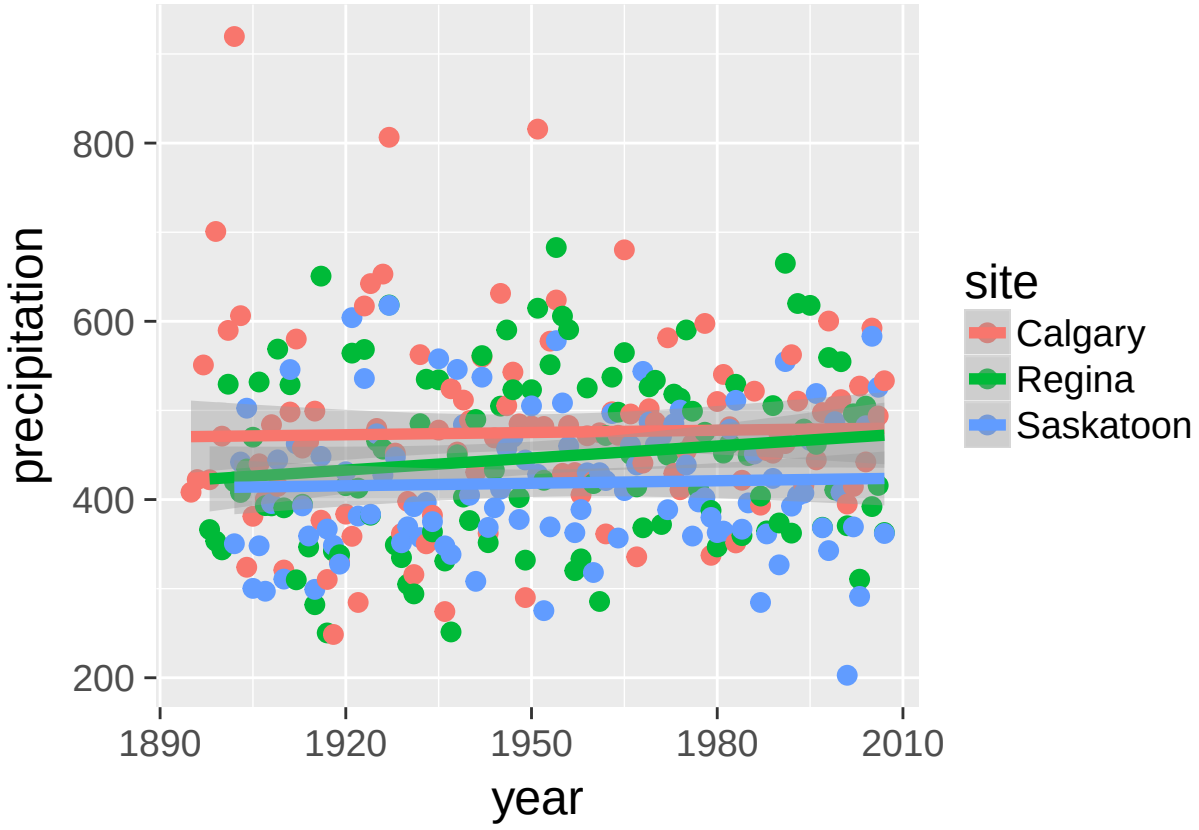
add regression curve to each category

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



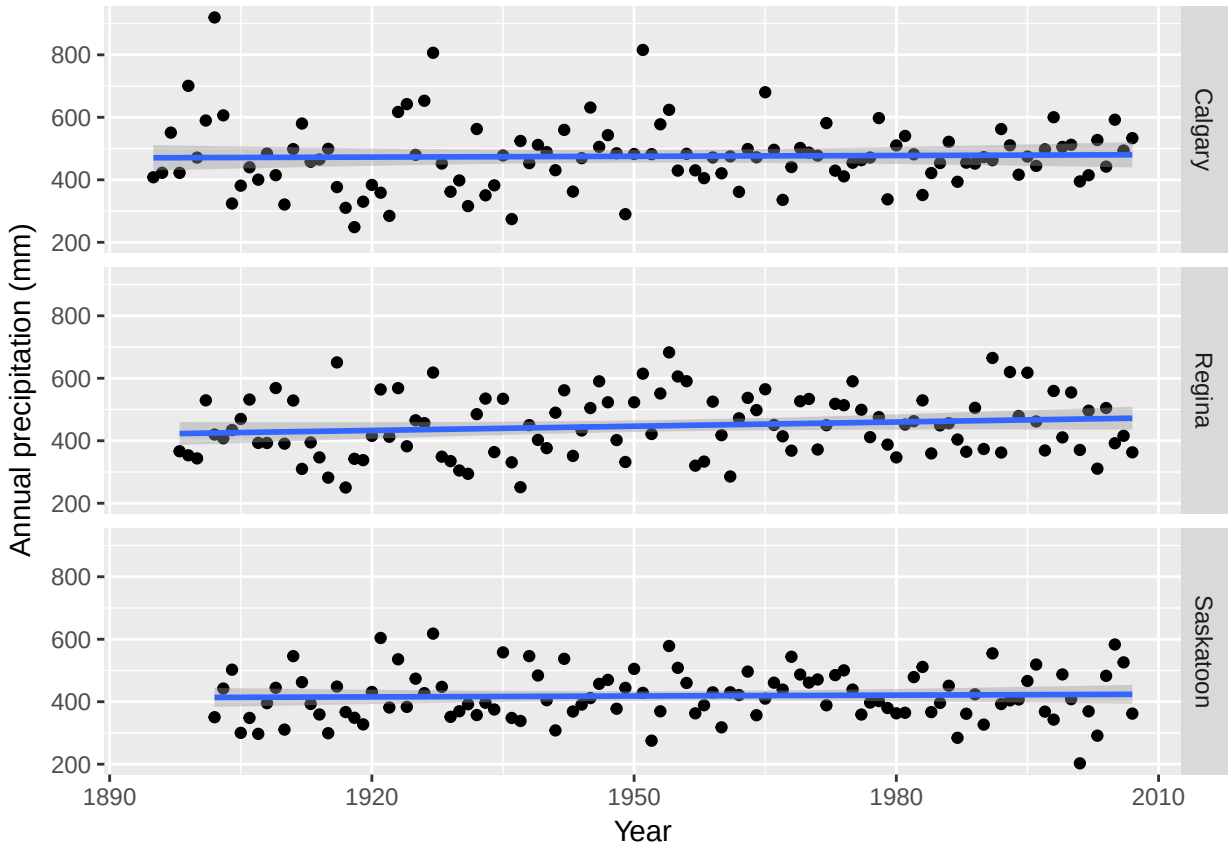
change theme font sizes

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



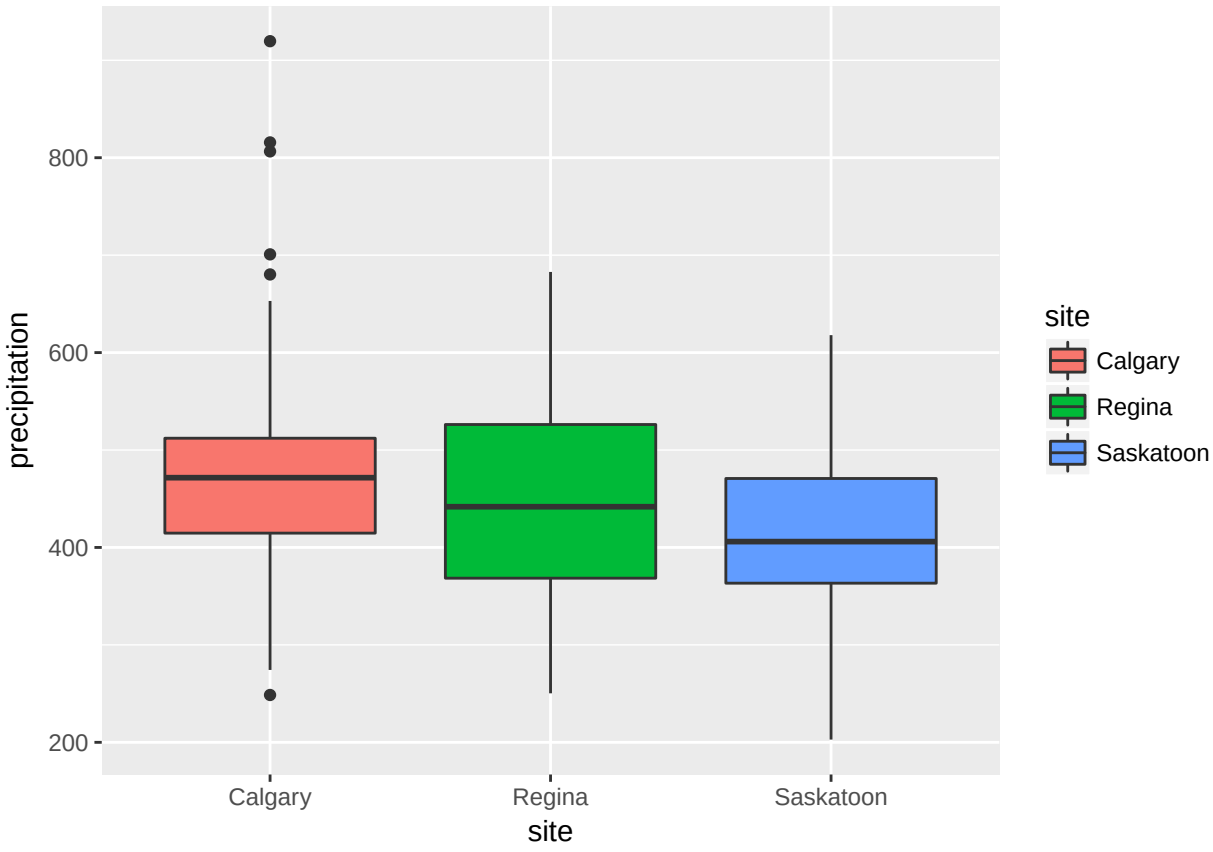
do faceting

```
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

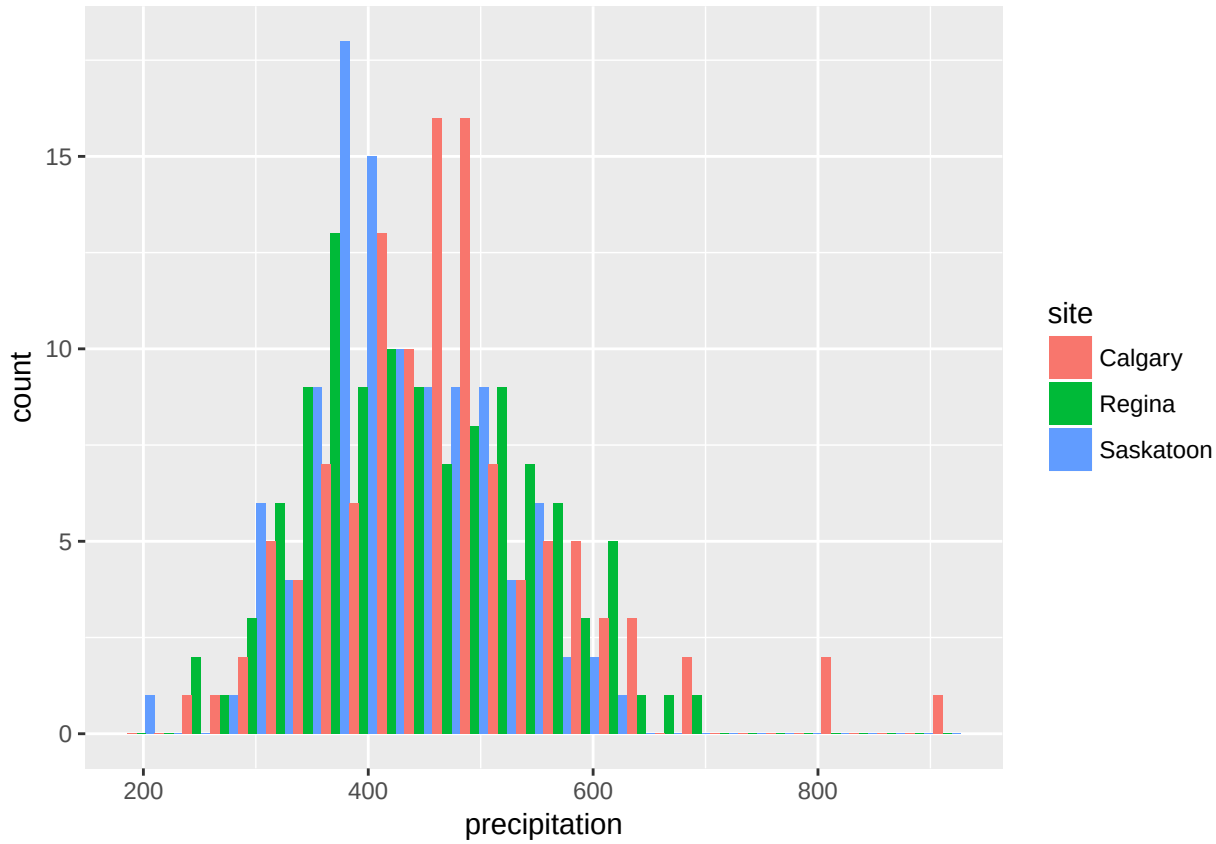
```
p4 <- ggplot(annual, aes(site, precipitation, fill = site))
p4 <- p4 + geom_boxplot()
p4
```



histograms

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

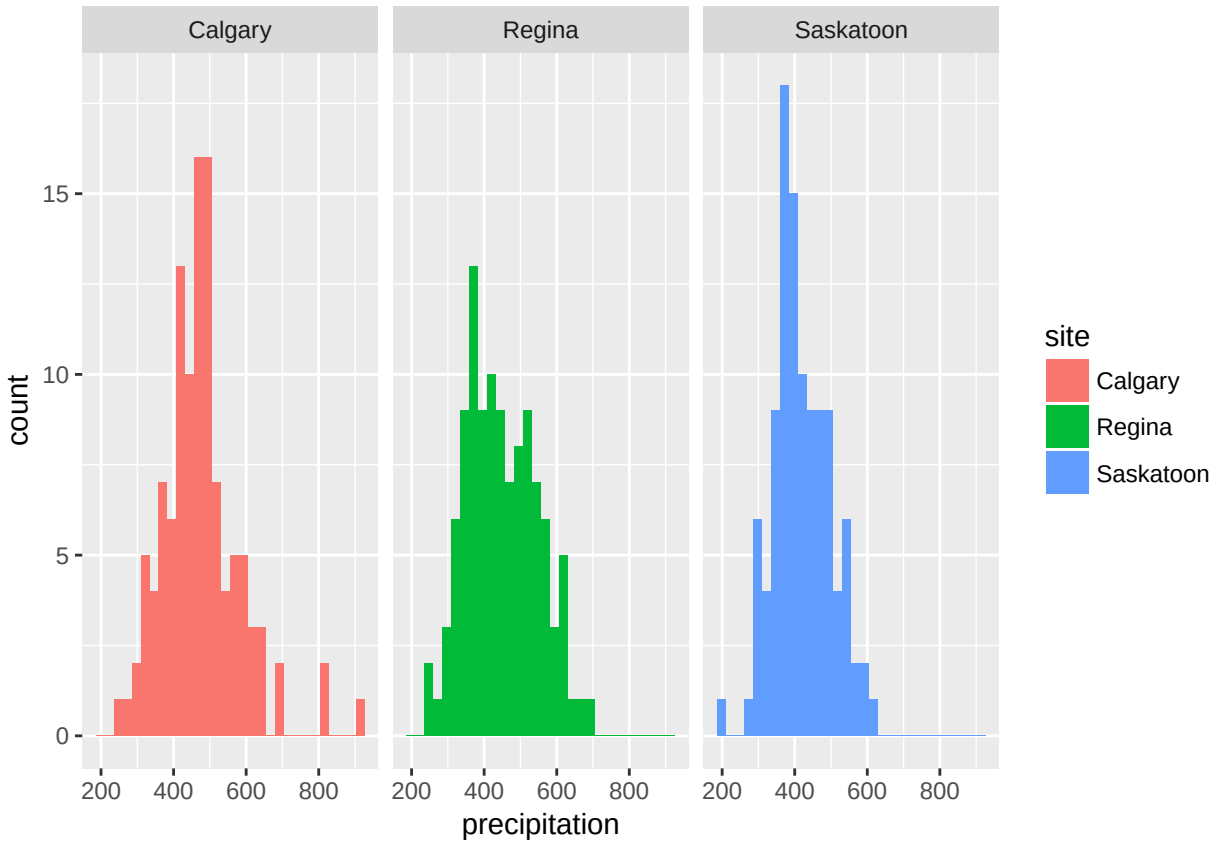
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



faceting

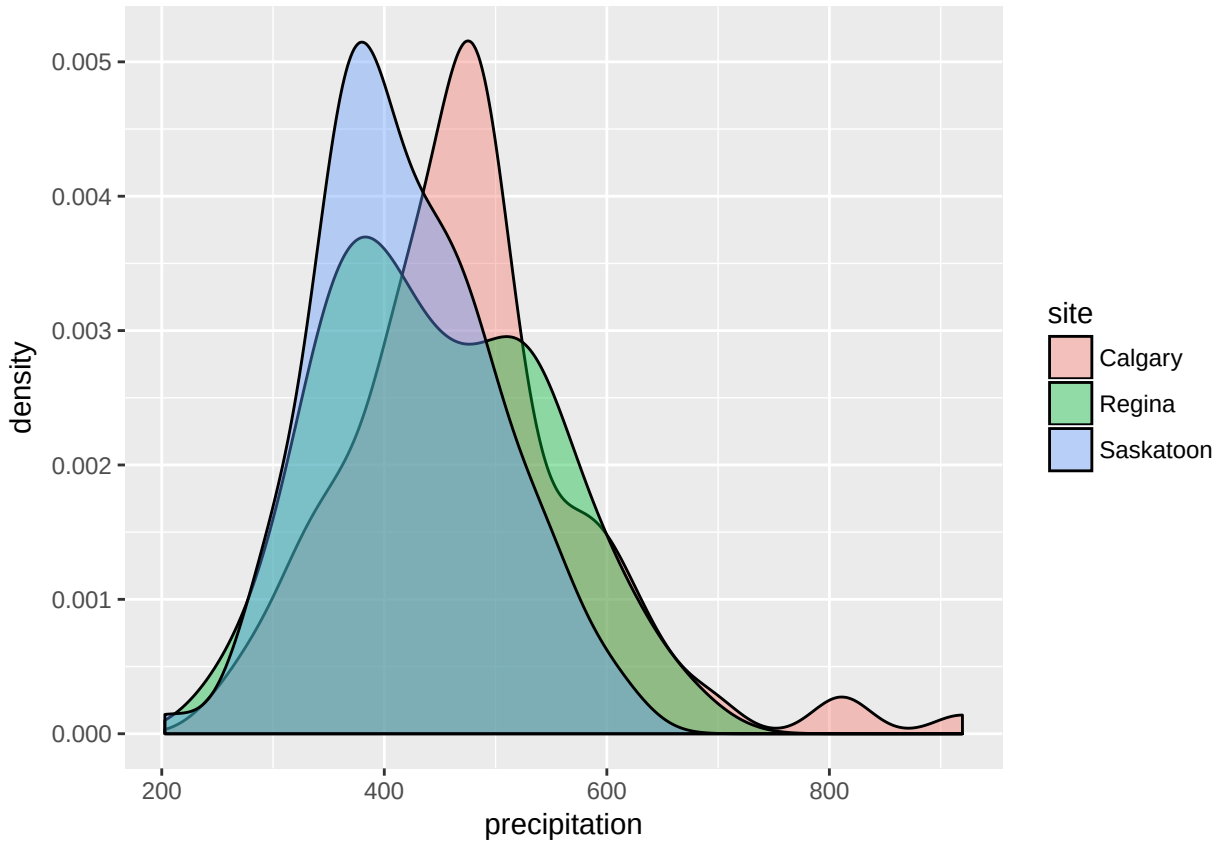
```
p5 <- p5 + facet_grid(. ~ site)
p5
```

```
## `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
```