Shetty\_S\_M1L3\_Programming Assignment

Installing additional packages for the assignment - ggplot2

install.packages("ggplot2")

require(ggplot2)

## Loading required package: ggplot2

Load the file M01\_quasi\_twitter.csv

twit <- read.csv(file="M01\_quasi\_twitter.csv",header=TRUE,sep=",")

The structure of the data and the columns names

str(twit)

## 'data.frame': 21916 obs. of 25 variables:  
## $ screen\_name : Factor w/ 21916 levels "+5400E1.","000D0se7",..: 4341 15303 21127 13570 14085 3607 14942 8653 15547 19146 ...  
## $ created\_at\_month : int 2 11 4 3 4 2 7 5 1 1 ...  
## $ created\_at\_day : int 9 21 1 24 23 9 15 23 23 13 ...  
## $ created\_at\_year : int 2007 2009 2007 2007 2009 2009 2006 2008 2009 2009 ...  
## $ country : Factor w/ 44 levels " Germany","Argentina",..: 44 19 19 44 44 12 44 5 44 44 ...  
## $ location : Factor w/ 378 levels "Akron Ohio","Alabama",..: 188 202 25 233 211 79 365 41 242 83 ...  
## $ friends\_count : int 1087 5210 1015 338 641 917 1574 16300 8316 640 ...  
## $ followers\_count : int 22187643 6692814 6257020 3433218 2929559 2540842 1960373 1934803 1855827 1697620 ...  
## $ statuses\_count : int 60246 93910 118465 78082 93892 59397 41023 62178 56057 82912 ...  
## $ favourites\_count : int 1122 3825 1143 0 226 2122 20160 15 540 3 ...  
## $ favourited\_count : int 105005 40487 87968 25943 32589 19760 13558 25084 8732 24515 ...  
## $ dob\_day : int 29 24 4 22 9 1 2 6 15 26 ...  
## $ dob\_year : int 1999 1991 1997 1998 1963 1995 1999 1986 1991 1986 ...  
## $ dob\_month : int 4 10 3 8 11 1 11 10 2 9 ...  
## $ gender : Factor w/ 2 levels "female","male": 1 1 2 2 1 1 1 2 1 2 ...  
## $ mobile\_favourites\_count: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ mobile\_favourited\_count: int 0 5032191 0 0 0 0 0 1934803 0 0 ...  
## $ education : int 8 15 9 9 13 15 14 10 11 12 ...  
## $ experience : int 0 0 0 44 24 21 31 0 27 20 ...  
## $ age : int 29 0 32 40 45 14 27 31 34 40 ...  
## $ race : Factor w/ 10 levels "arab","asian",..: 10 10 10 10 10 10 10 10 2 1 ...  
## $ wage : num 16.3 17.9 15.7 7 17.9 ...  
## $ retweeted\_count : int 1 1 2 0 1 2 1 2 0 0 ...  
## $ retweet\_count : int 30 6 65 8 7 64 13 14 15 10 ...  
## $ height : int 156 162 168 180 162 158 160 178 156 173 ...

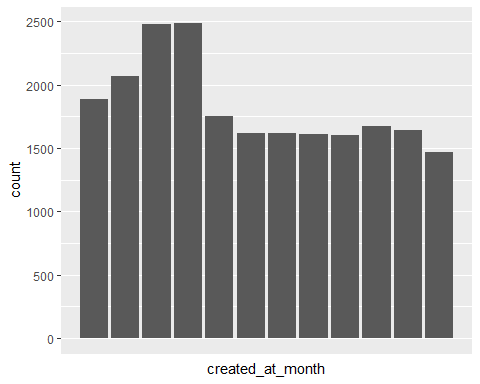
names(twit)

## [1] "screen\_name" "created\_at\_month"   
## [3] "created\_at\_day" "created\_at\_year"   
## [5] "country" "location"   
## [7] "friends\_count" "followers\_count"   
## [9] "statuses\_count" "favourites\_count"   
## [11] "favourited\_count" "dob\_day"   
## [13] "dob\_year" "dob\_month"   
## [15] "gender" "mobile\_favourites\_count"  
## [17] "mobile\_favourited\_count" "education"   
## [19] "experience" "age"   
## [21] "race" "wage"   
## [23] "retweeted\_count" "retweet\_count"   
## [25] "height"

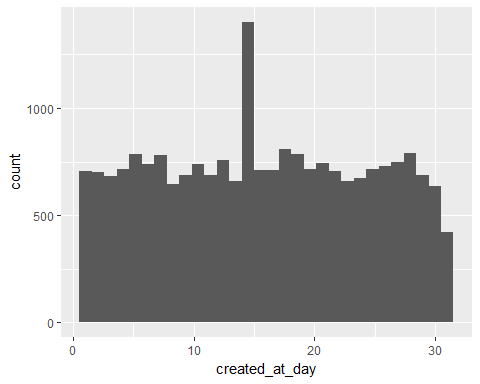
## How is the data distributed?

#### Created at month, day, year

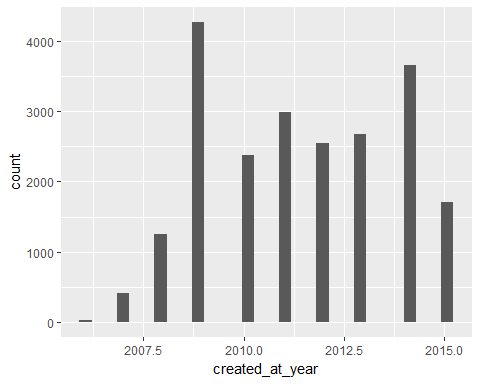
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 3.000 6.000 6.069 9.000 12.000



## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



A columns created\_at\_month and created\_at\_day are uniformly distributed across all 12 calendar months.

The histogram visually displays the same skew in the data toward the beginning of the year, and more specifically shows that March, April, and February are the most popular months for new accounts. There is a spike in the number of accounts created during years 2009-2010.

qplot(dob\_day,data=twit,geom="histogram",binwidth=1)



The number of users having date of birth as January 1st is higher as compared to other dates. This would mean that when people are creating accounts in Twitter they use the default date. This data may not be accurate in most cases.

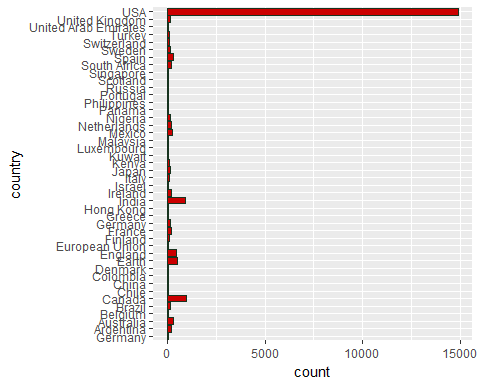
#### Twitter data by country

summary(twit$country)

## Germany Argentina Australia   
## 54 190 291   
## Belgium Brazil Canada   
## 56 121 943   
## Chile China Colombia   
## 61 57 50   
## Denmark Earth England   
## 59 516 467   
## European Union Finland France   
## 58 110 180   
## Germany Greece Hong Kong   
## 151 59 60   
## India Ireland Israel   
## 890 171 52   
## Italy Japan Kenya   
## 116 162 117   
## Kuwait Luxembourg Malaysia   
## 49 62 55   
## Mexico Netherlands Nigeria   
## 236 170 132   
## Panama Philippines Portugal   
## 59 53 49   
## Russia Scotland Singapore   
## 63 57 53   
## South Africa Spain Sweden   
## 183 283 123   
## Switzerland Turkey United Arab Emirates   
## 115 73 56   
## United Kingdom USA   
## 149 14905

Summary shows that one country, the USA has most users signed up. India and Canada have nearly 1000 in this data set.

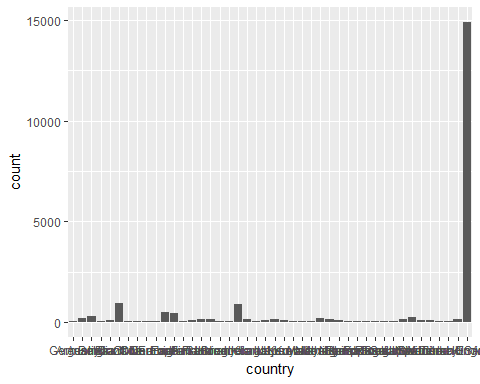
qplot(country, data=twit, color=I('#17331F'), fill=I('#CC0000'))+coord\_flip()



#### Location

Location looks like a uniform distribution with a few outliers.

qplot(x=country, data=twit, geom="bar")

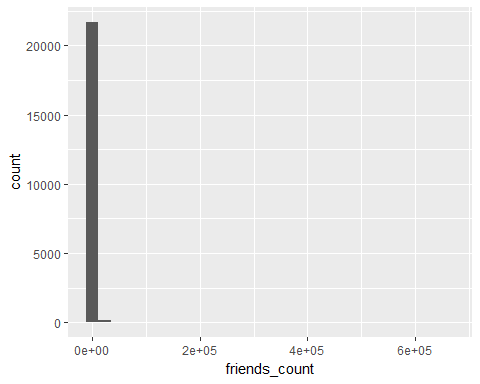


#### Friends\_count

From the summary statistics we can see that this data is heavily skewed as the median and 3rd Quartile are less than the mean and the max is approximately 60X greater than the mean. There is also some faulty data as the minimum data value is -84, which is not possible for a friend count as negative number of friends do not exist. That would mean someone has no friends but 84 non-friends, say enemies which would not be common knowledge on social networking sites like twitter.

qplot(data=twit, x=friends\_count)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



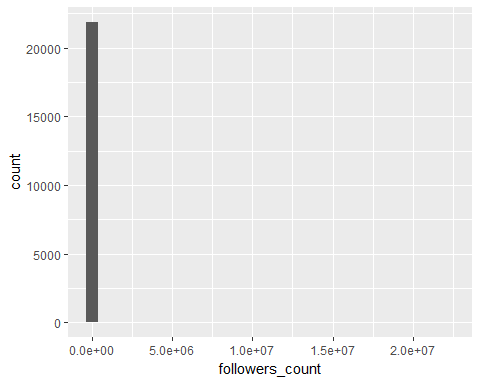
summary(twit$friends\_count)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -84 123 324 1058 849 660500

Similarly, we can plot for other columns as shown below:

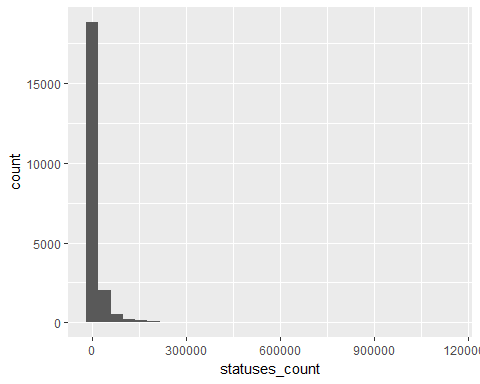
qplot(followers\_count,data=twit,geom="histogram")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



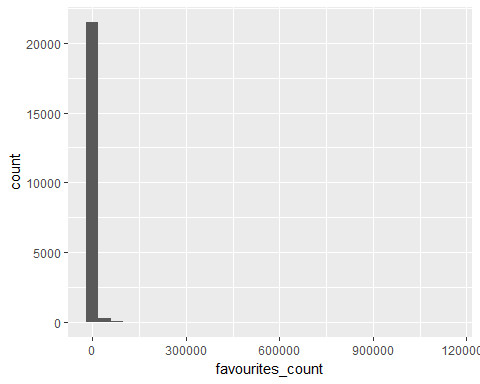
qplot(statuses\_count,data=twit,geom="histogram")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



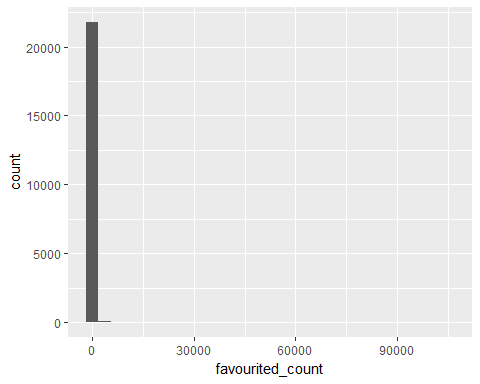
qplot(favourites\_count,data=twit,geom="histogram")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



qplot(favourited\_count,data=twit,geom="histogram")

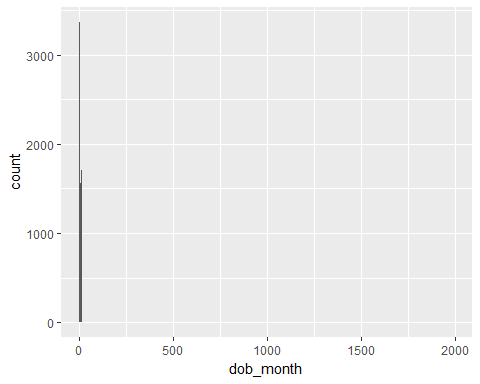
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



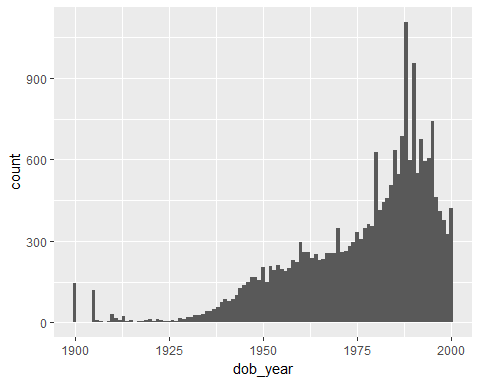
qplot(dob\_day,data=twit,geom="histogram",binwidth=1)



qplot(dob\_month,data=twit,geom="histogram",binwidth=1)

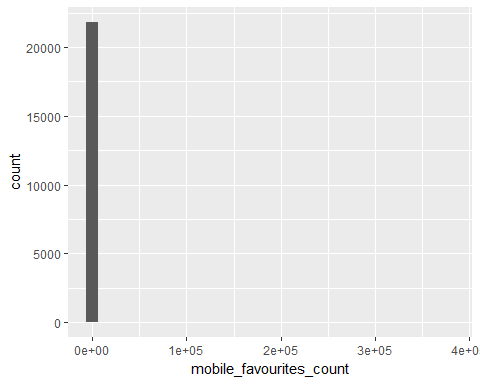


qplot(dob\_year,data=twit,geom="histogram",binwidth=1)



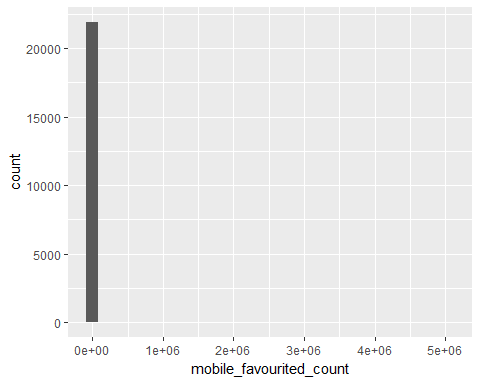
qplot(mobile\_favourites\_count,data=twit,geom="histogram")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



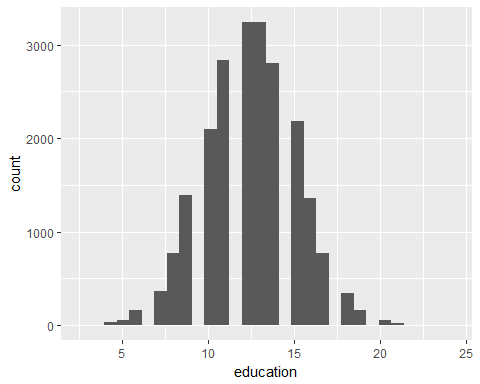
qplot(mobile\_favourited\_count,data=twit,geom="histogram")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



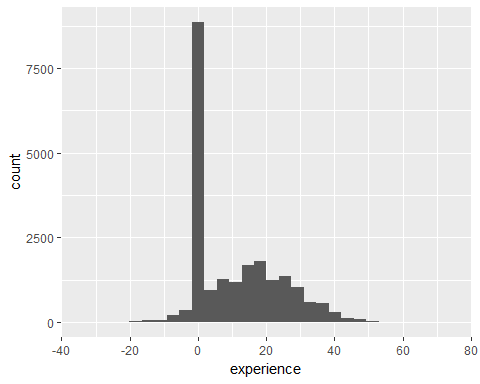
qplot(education,data=twit,geom="histogram")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



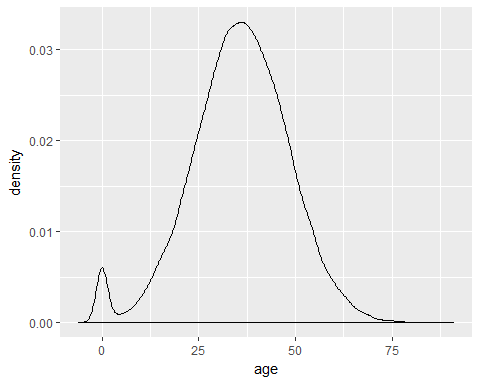
qplot(experience,data=twit,geom="histogram")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

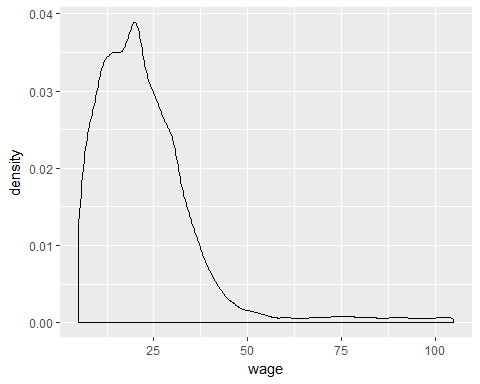


We can also plot using density plots.

qplot(age,data=twit,geom="density")



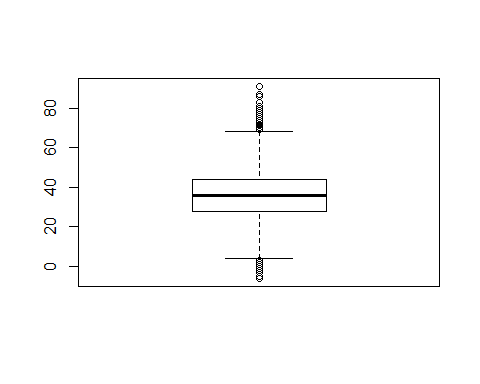
qplot(wage,data=twit,geom="density")



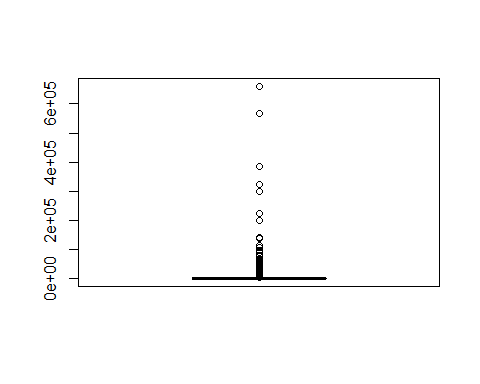
## Are there anomalies/outliers?

Anamolies/outliers can be found using boxplots.

boxplot(twit$age)

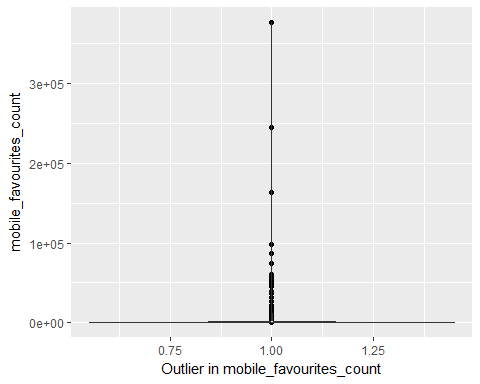


boxplot(twit$friends\_count)

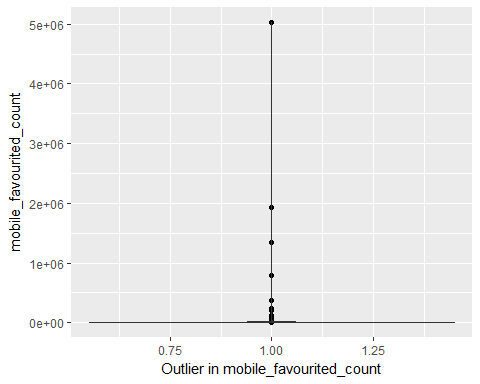


Violin Charts can also be used to find outliers.

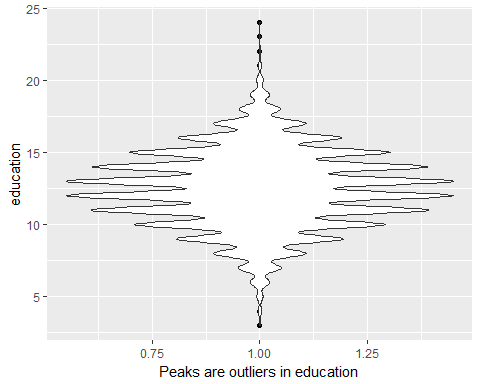
qplot(1,mobile\_favourites\_count, data=twit,xlab = "Outlier in mobile\_favourites\_count")+geom\_violin()



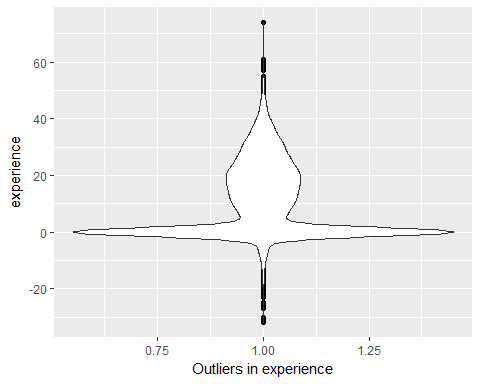
qplot(1,mobile\_favourited\_count, data=twit,xlab = "Outlier in mobile\_favourited\_count")+geom\_violin()



qplot(1,education, data=twit,xlab = "Peaks are outliers in education")+geom\_violin()



qplot(1,experience, data=twit,xlab = "Outliers in experience")+geom\_violin()



## Can you identify the following -

* useful raw data and transforms (e.g. log(x))
* data quality problems
* outliers
* subsets of interest

One of the useful data transform can be for friends\_count

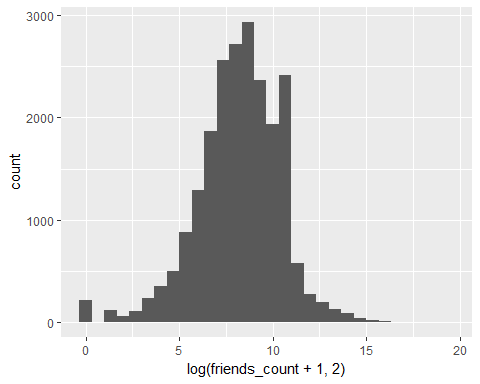
It looks like a fat-tailed distribution. This implies that there is a small subset of users who have lot of friends, but most users don't have many. This can be visualized easily on a log-scaled plot.

qplot(x=log(friends\_count+1,2), data=twit)

## Warning: NaNs produced  
  
## Warning: NaNs produced  
  
## Warning: NaNs produced

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

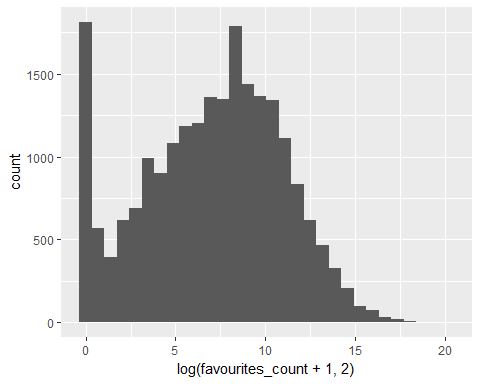
## Warning: Removed 1 rows containing non-finite values (stat\_bin).



Similarly, for favourites\_count, we can visualize subset of data better using log-scaled plot.

qplot(x=log(favourites\_count+1,2), data=twit)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



Some data quality problems are

* negative values in age, experience and friends\_count. This can be inferred from box plots/violin plots
* Fake age entries

## Functional Relationships

In the assignment to show functional relationship I have chosen age and wage as the parameters. Following are my observations -

* It shows that across all age groups men report making much more than women
* There seems to be a limit to the wage that women report at around the 50 mark
* The same rule does not apply to men as they seem to be earning beyond the age 50 mark

ggplot(twit, aes(x = age, y = wage, colour=gender)) + geom\_point()

