TRANSFORMATION

SYRACUSE UNIVERSITY
School of Information Studies

ATTRIBUTE TRANSFORMATION

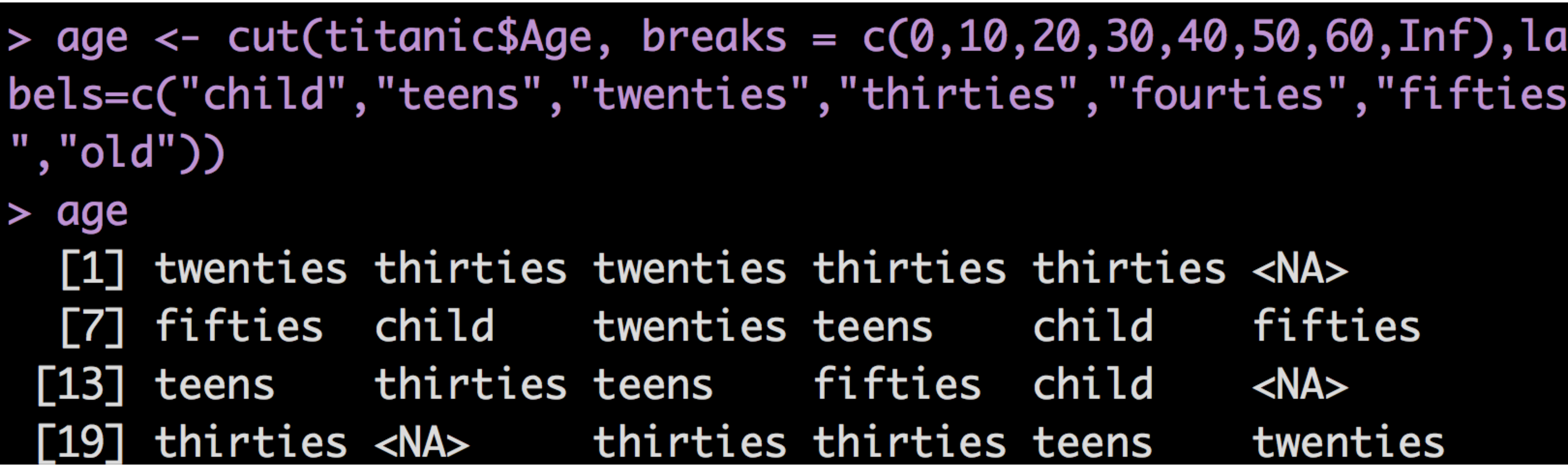
Sometimes, the original values of an attribute need to be transformed for purpose of analysis.

Some common transformations:

- Discretization
- Log transformation
- Normalization
  - Z-score
  - Min_max

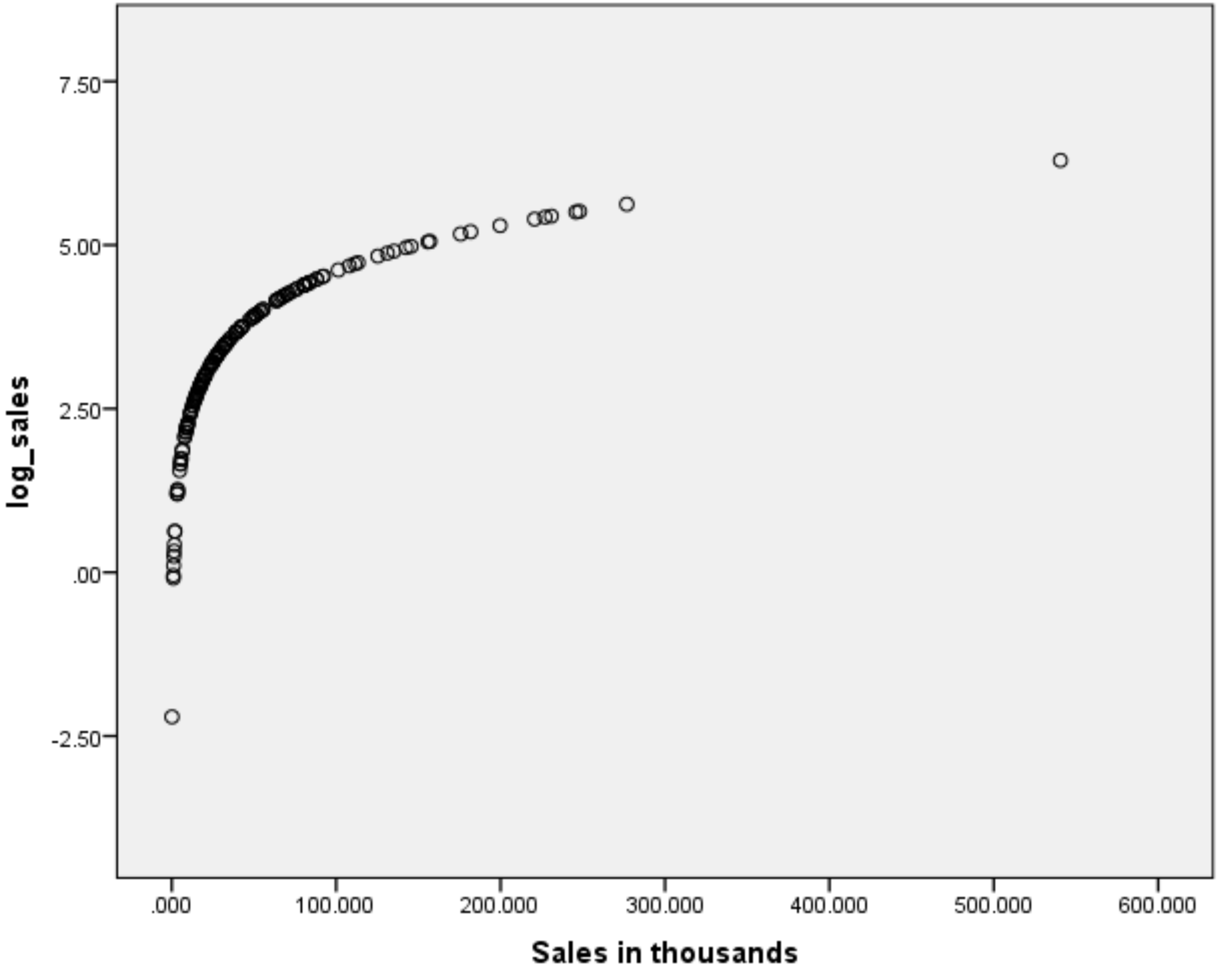
DISCRETIZATION (BINNING)

Discretization is a process to transform a continuous attribute to a discrete one.

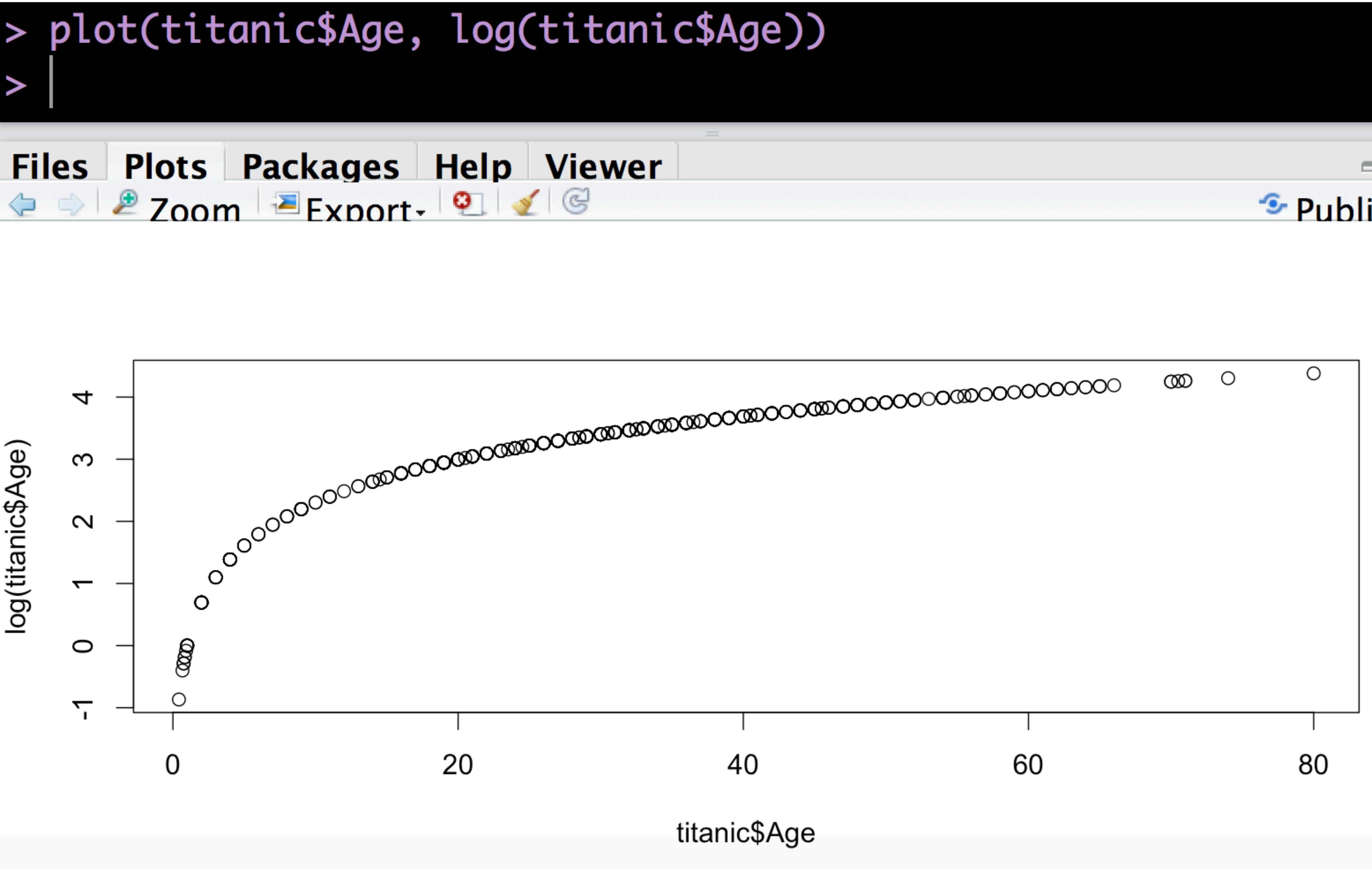


```
> age <- cut(titanic$Age, breaks = c(0,10,20,30,40,50,60,Inf),la
bels=c("child","teens","twenties","thirties","fourties","fifties
","old"))
> age
  [1] twenties thirties twenties thirties thirties <NA>    
  [7] fifties  child    twenties teens    child    fifties 
 [13] teens    thirties teens    fifties  child    <NA>    
 [19] thirties <NA>     thirties thirties teens    twenties
```

LOG TRANSFORMATION



Log transformation leaves the analysis more robust with outliers.



```
> plot(titanic$Age, log(titanic$Age))
> |
```

Z-SCORE TRANSFORMATION

A data analysis problem: What Facebook messages sent from restaurants are popular among fans? Popularity is measured by the number of comments received.

Two restaurants: McDonald’s and Lemon Grass

- McDonald’s has millions of fans on Facebook, while Lemon Grass has thousands.
- A message from McDonald’s received 1,000 comments.
- A message from Lemon Grass also received 1,000 comments.
- Which message is more popular, or are they equally popular?

http://research.microsoft.com/en-us/um/beijing/events/ia2011/n7.pdf

Z-SCORE TRANSFORMATION

The message from Lemon Grass seems more popular, but right now the face values look the same: 1,000.

How to demonstrate the real difference in popularity?

Z-SCORE TRANSFORMATION

Assume:

McDonald’s

- Average number of comments: *u* = 2,000
- Standard deviation: *sd* = 500

Lemon Grass

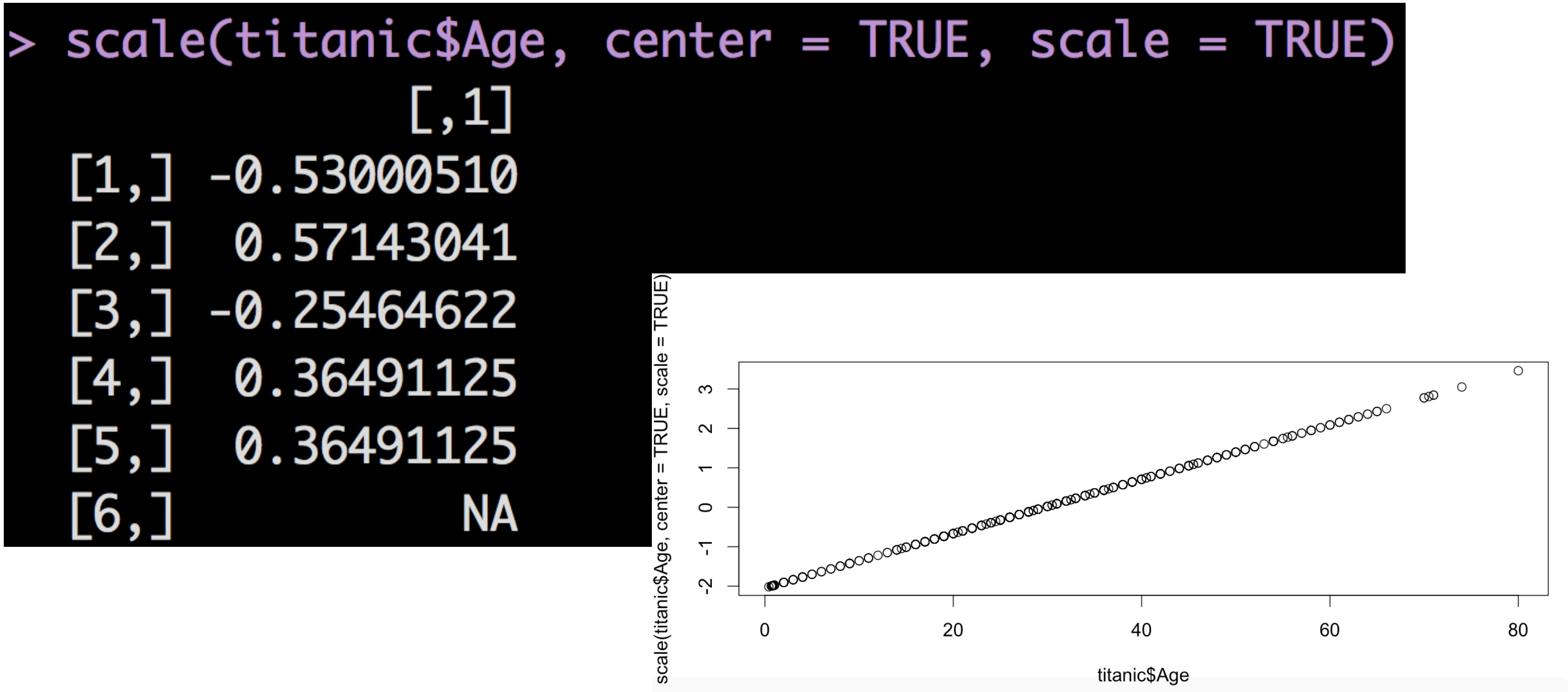
- Average number of comments: *u* = 200
- Standard deviation: *sd* = 50

Z(x) = (x – u)/sd

Z-SCORE TRANSFORMATION

Facebook Messages	# Comments	Z-Score	
McDonald’s msg 1	1,000	-2	
McDonald’s msg 2	500	-3	
…			
Lemon Grass msg 1	1,000	16	
Lemon Grass msg 2	500	6	
…			

Z-SCORE IN R



```
> scale(titanic$Age, center = TRUE, scale = TRUE)
                [,1]
  [1,] -0.53000510
  [2,]  0.57143041
  [3,] -0.25464622
  [4,]  0.36491125
  [5,]  0.36491125
  [6,]          NA
```

MIN_MAX TRANSFORMATION

Assume:

McDonald’s

- Minimum number of comments: *min* = 50
- Maximum number of comments: *max* = 10,000

Lemon Grass

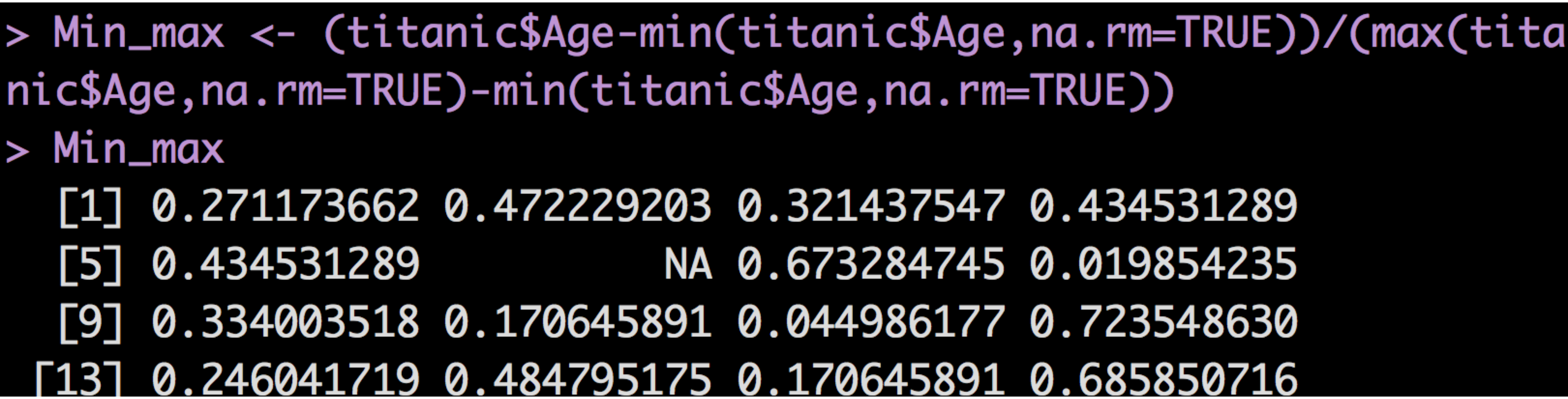
- Minimum number of comments: *min* = 10
- Maximum number of comments: *max* = 2,000

Min_max(x) = (x-min)/(max-min)

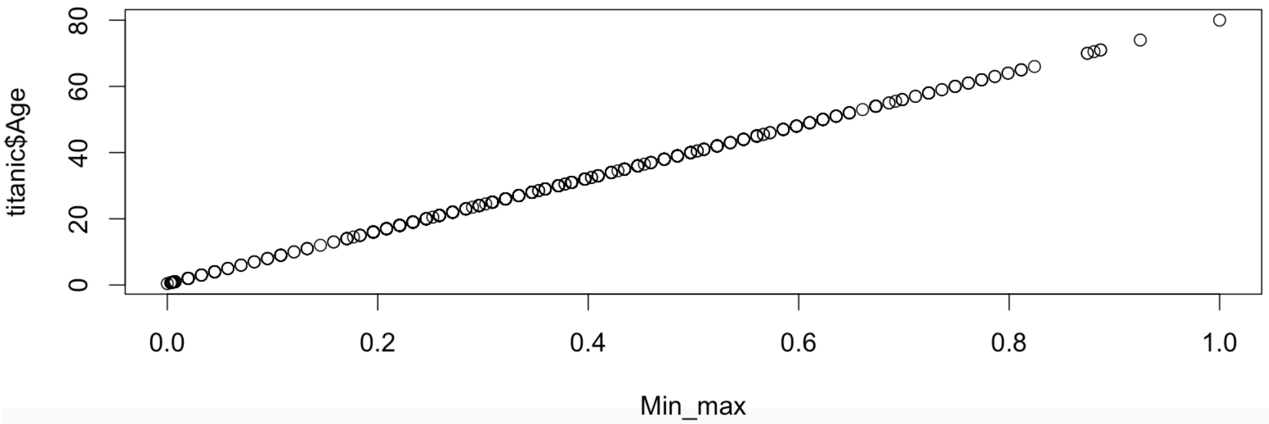
MIN_MAX TRANSFORMATION

Facebook Messages	# Comments	Z-Score	Min_Max
McDonald's msg 1	1,000	-2	.10
McDonald's msg 2	500	-3	.05
…			
Lemon Grass msg 1	1,000	16	.50
Lemon Grass msg 2	500	6	.25
…			

MIN_MAX IN R

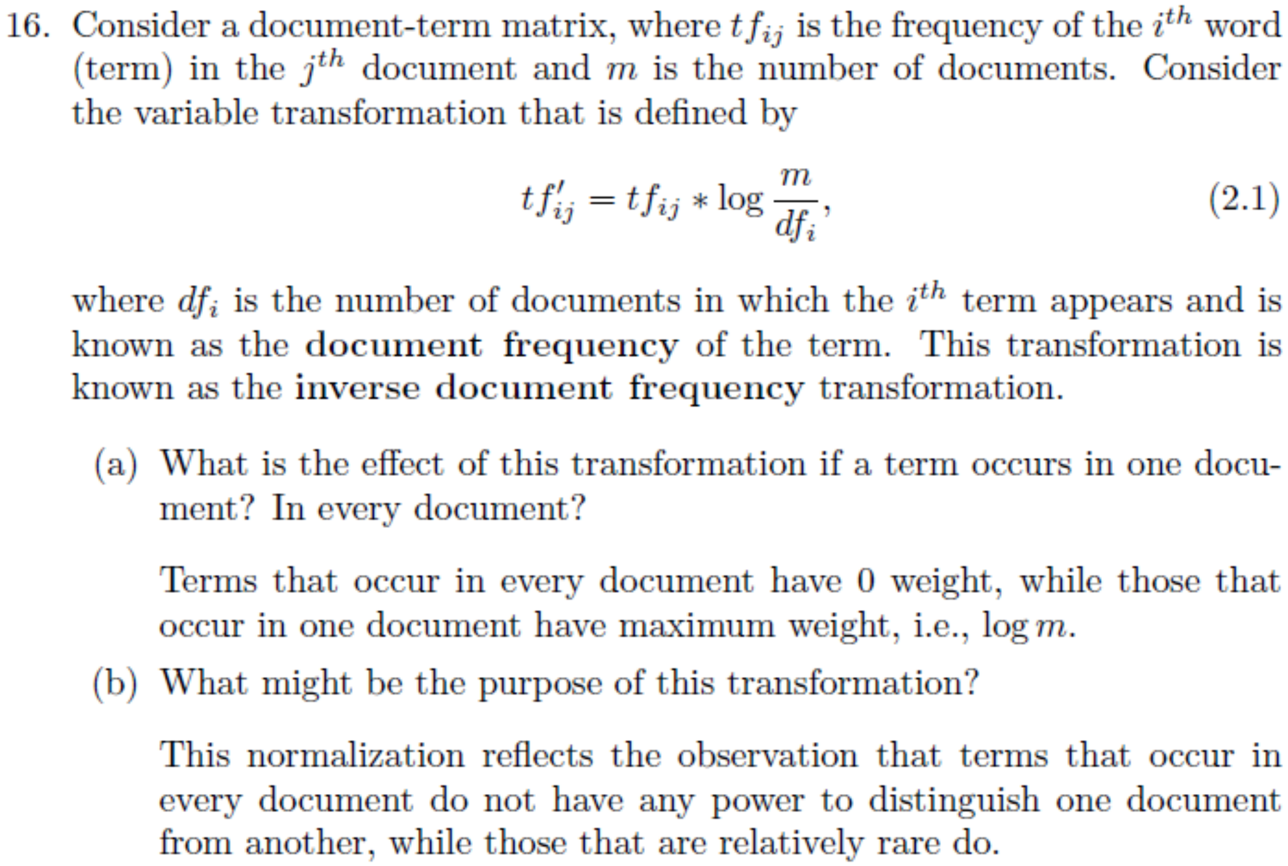


```
> Min_max <- (titanic$Age-min(titanic$Age,na.rm=TRUE))/(max(tita
nic$Age,na.rm=TRUE)-min(titanic$Age,na.rm=TRUE))
> Min_max
  [1] 0.271173662 0.472229203 0.321437547 0.434531289
  [5] 0.434531289          NA 0.673284745 0.019854235
  [9] 0.334003518 0.170645891 0.044986177 0.723548630
 [13] 0.246041719 0.484795175 0.170645891 0.685850716
```



MANY MORE TRANSFORMATIONS

TFIDF (Textbook exercise 16 on page 92)



16. Consider a document-term matrix, where tf_{ij} is the frequency of the i^{th} word (term) in the j^{th} document and m is the number of documents. Consider the variable transformation that is defined by

$$tf'_{ij} = tf_{ij} * \log \frac{m}{df_i}, \tag{2.1}$$

where df_i is the number of documents in which the i^{th} term appears and is known as the **document frequency** of the term. This transformation is known as the **inverse document frequency** transformation.

(a) What is the effect of this transformation if a term occurs in one document? In every document?

Terms that occur in every document have 0 weight, while those that occur in one document have maximum weight, i.e., $\log m$.

(b) What might be the purpose of this transformation?

This normalization reflects the observation that terms that occur in every document do not have any power to distinguish one document from another, while those that are relatively rare do.

A REVIEW OF DATA TRANSFORMATION

Aggregation

Discretization

Log transformation

Z-score transformation

Min_max transformation