MATHS 7107
Data Taming
Week 5

Shenal Dedduwakumara

School of Mathematical Sciences, University of Adelaide

Transforming Data

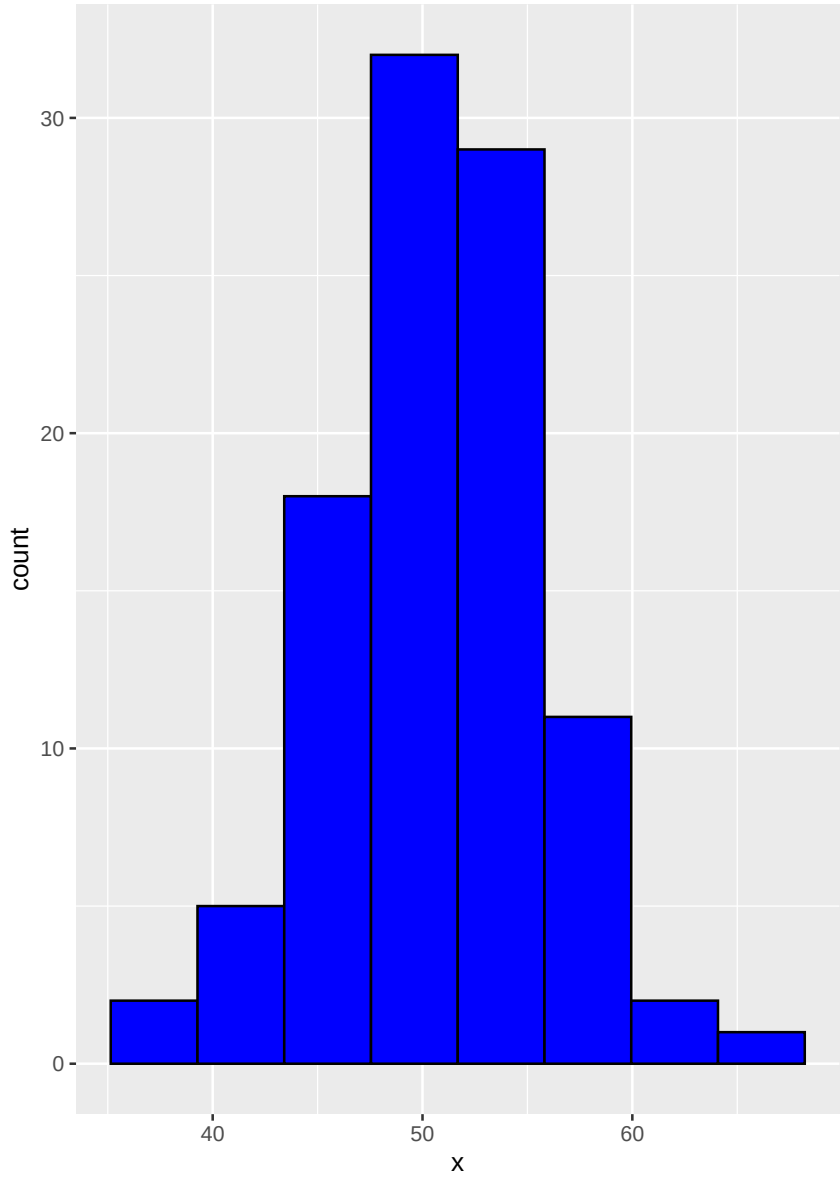
- Standardisation
- Min-max Scaling
- Log transformation.
- Box-Cox transformation.

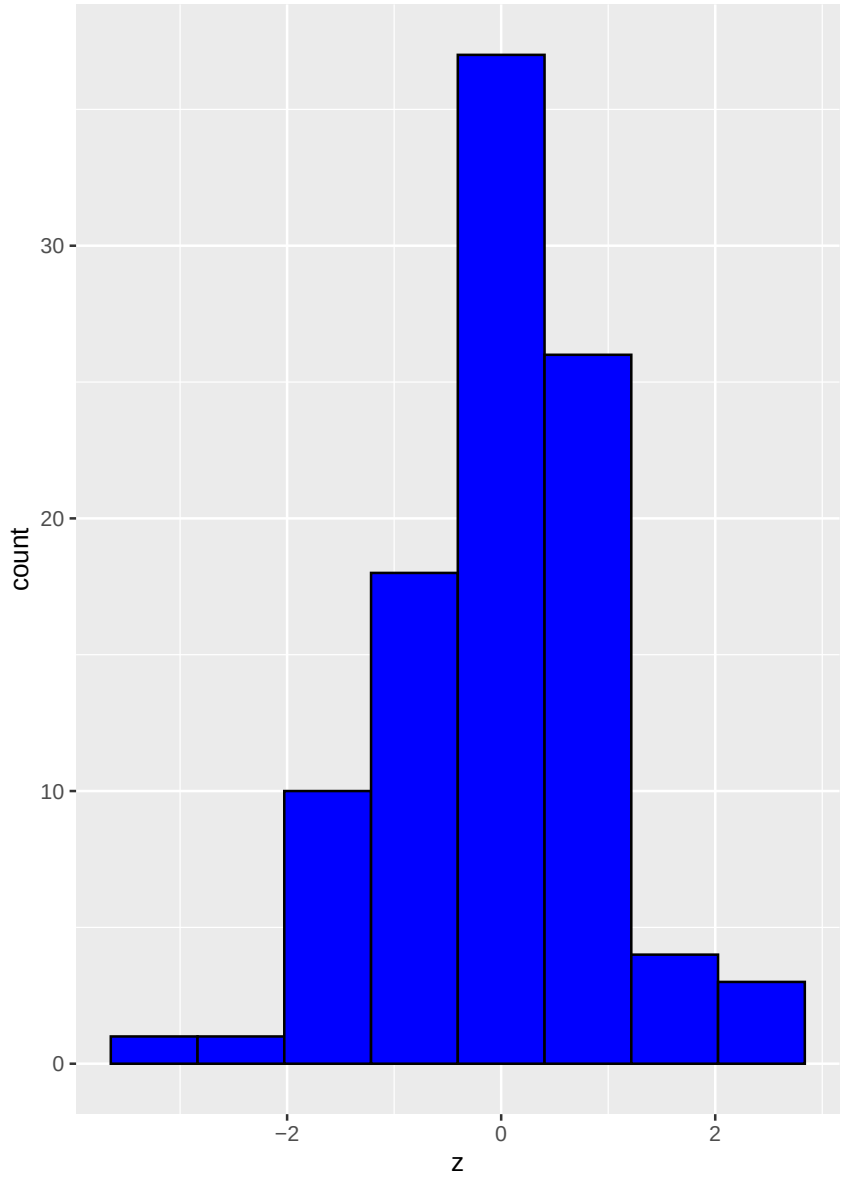
Standardisation

- Standardization refers to the process of putting different variables on the same scale in order to compare scores between different types of variables.

$$z = \frac{x - \bar{x}}{s}$$

Standardisation





Standardisation

- Mean of x

```
## [1] 50.69373
```

- Standard deviation of x

```
## [1] 5.100466
```

- Mean of z

```
## [1] 0
```

- Standard deviation of z

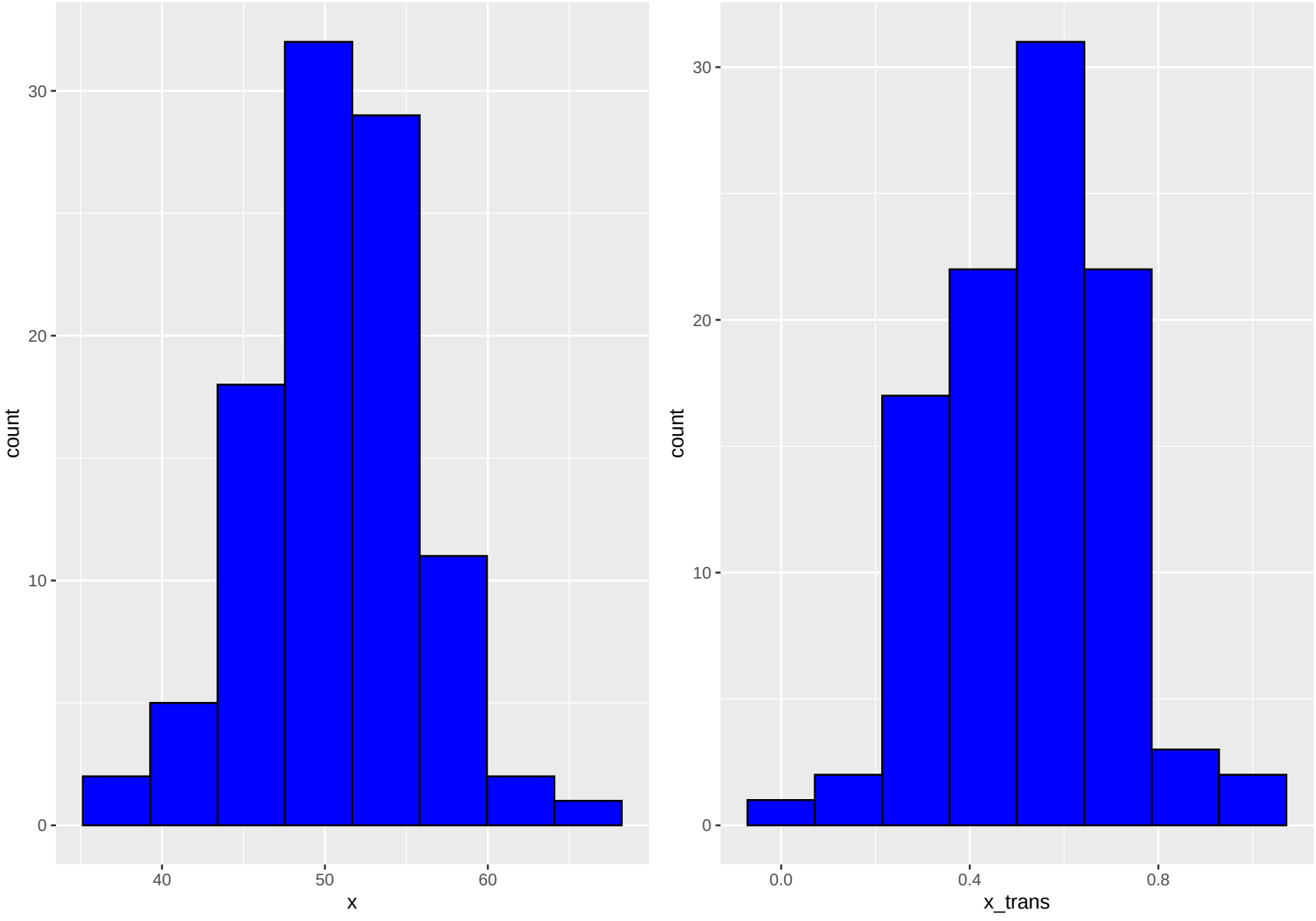
```
## [1] 1
```

Min-max Scaling

- Rescaling the range of features to scale the range in $[0, 1]$.

$$x^* = \frac{x - min(x)}{max(x) - min(x)}$$

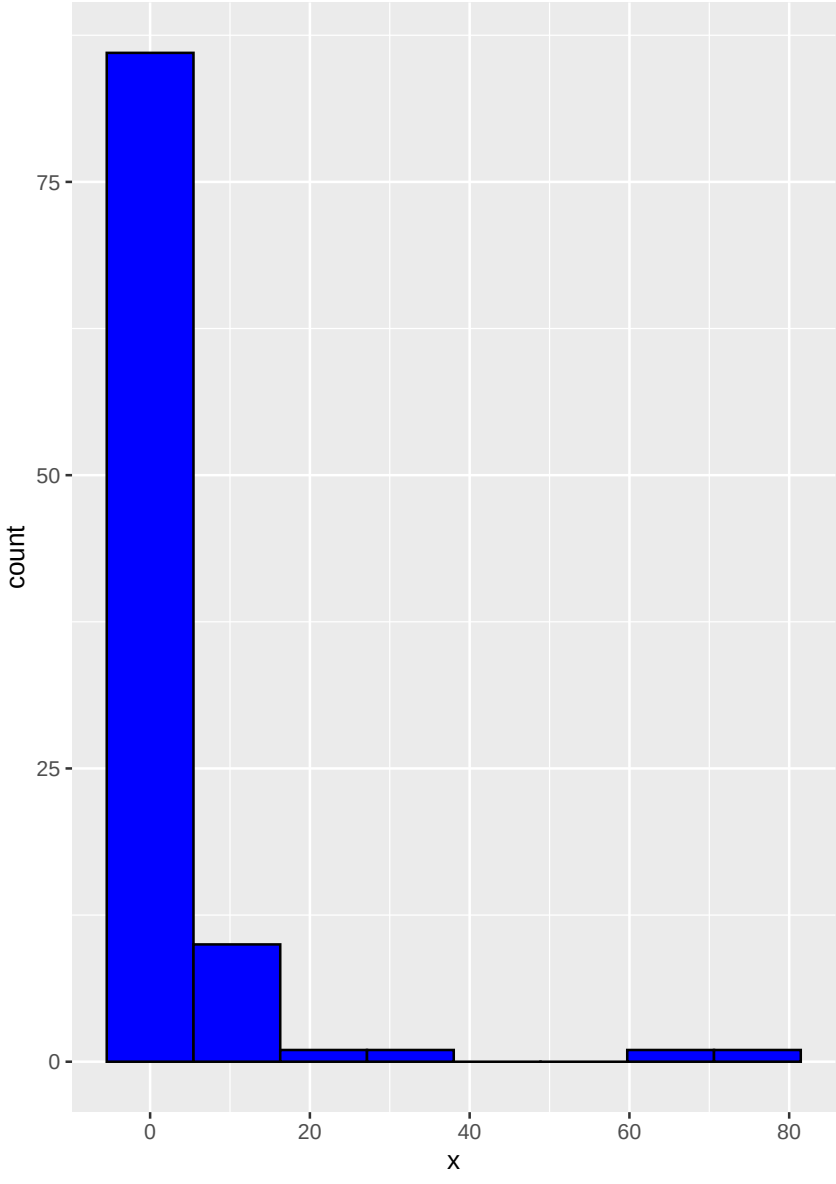
Min-max Scaling

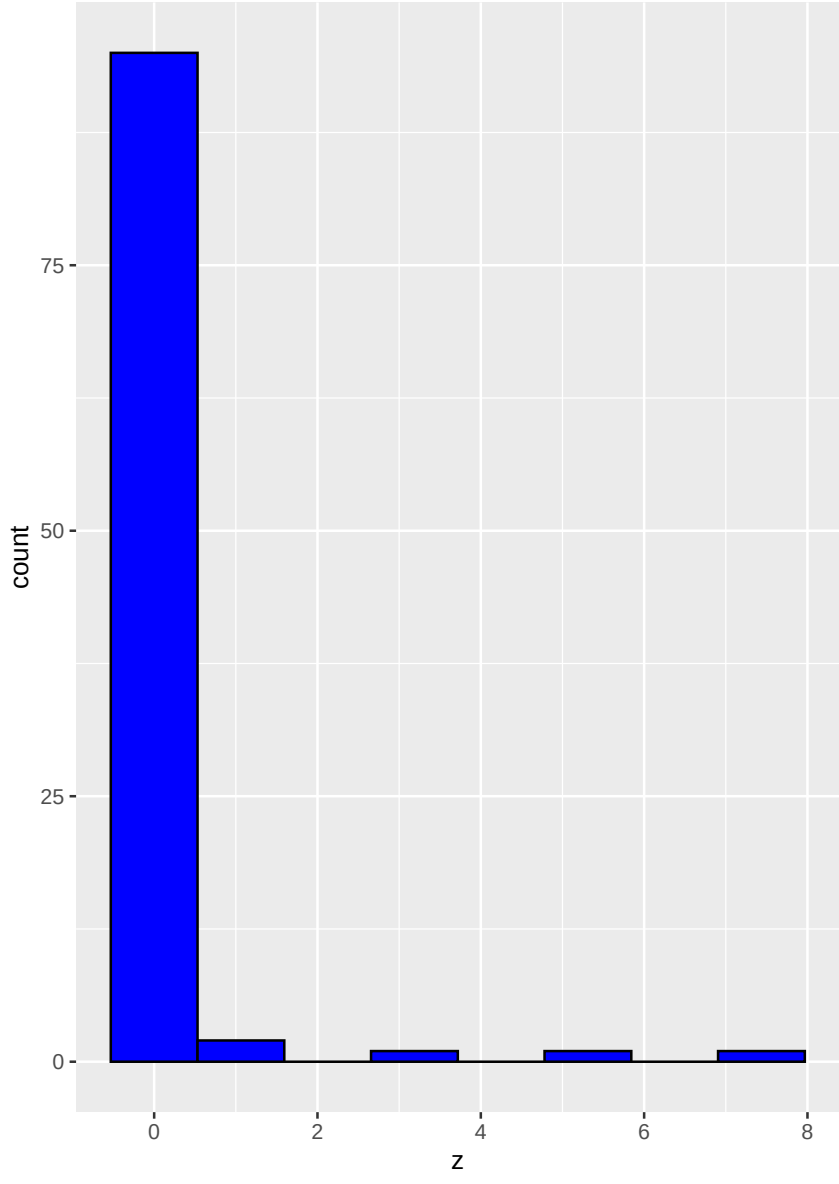


Transforming Data for Normality

- Many statistical techniques perform calculations assuming the data is normally distributed.

Transforming Data for Normality





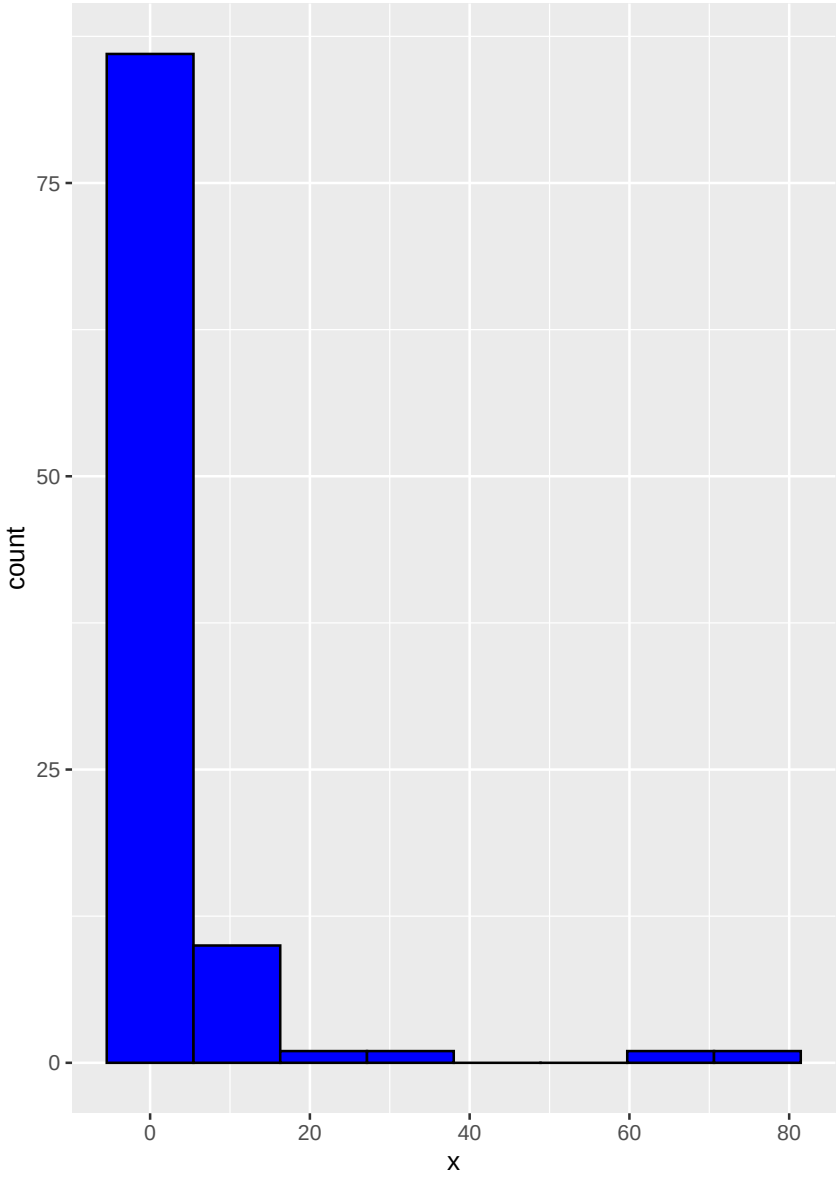
Log transformation.

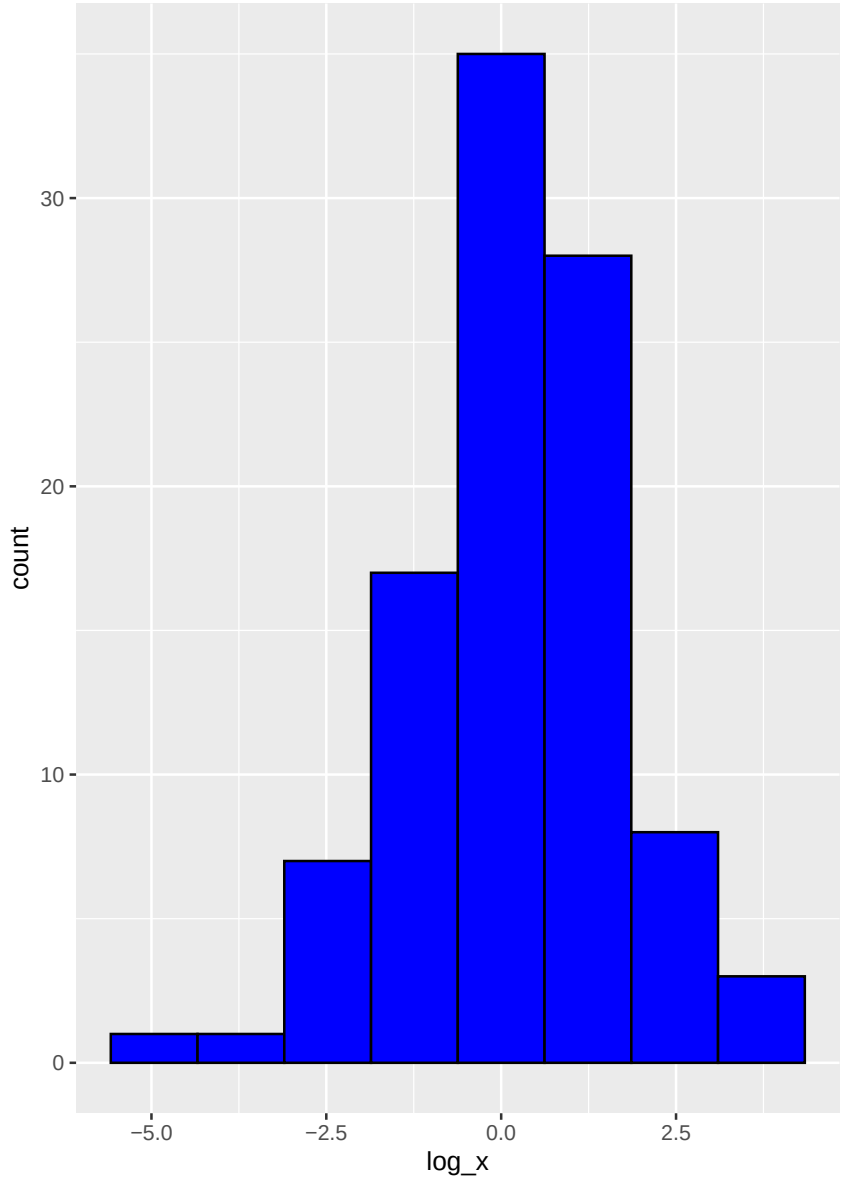
- A log transformation is a process of applying a logarithm to data to reduce its skew.

$$x^* = log(x)$$

Note : If you have zeros in the data and you can't take the logarithm of zero. In that case you can do log(x+1)

Log transformation.

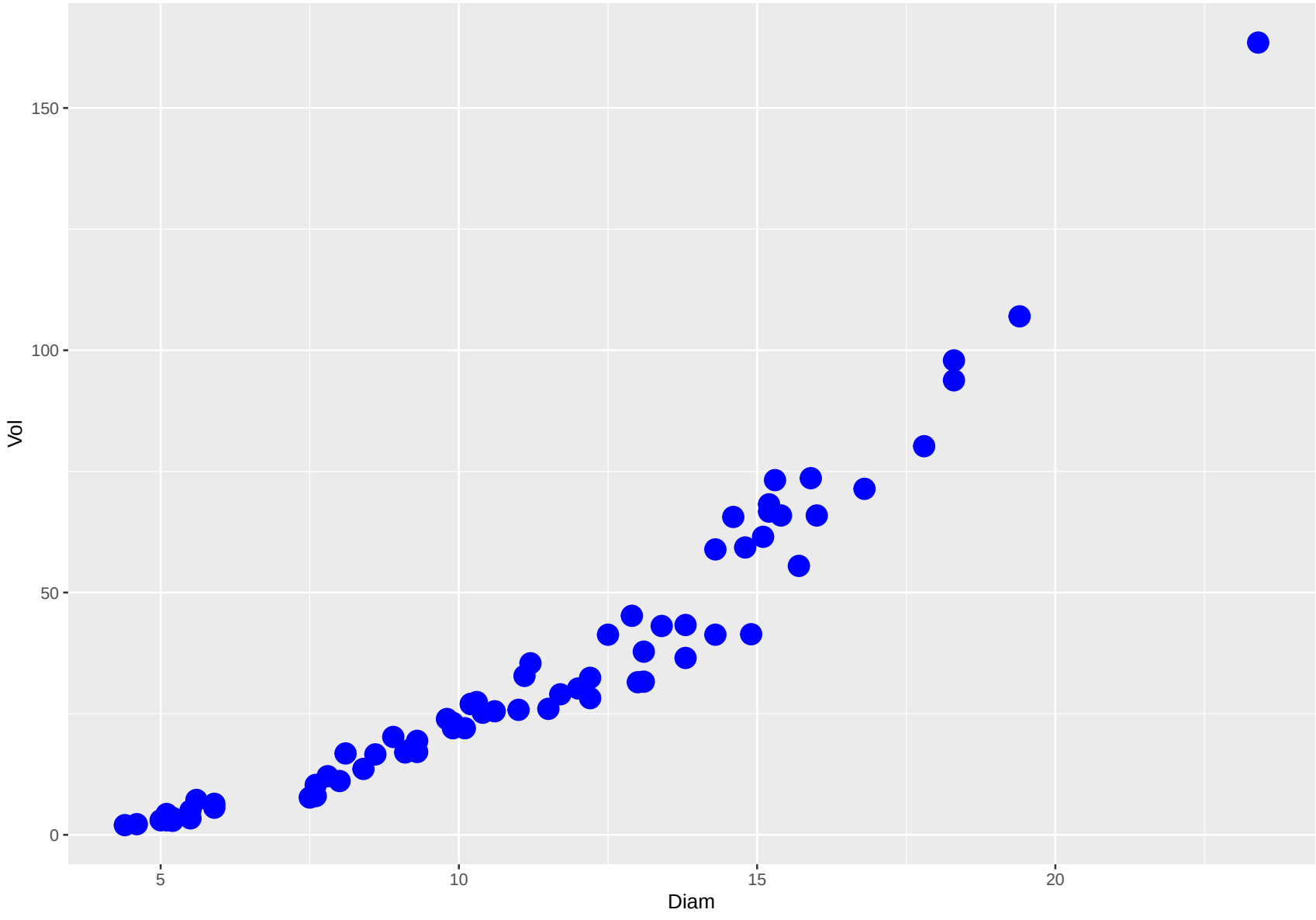




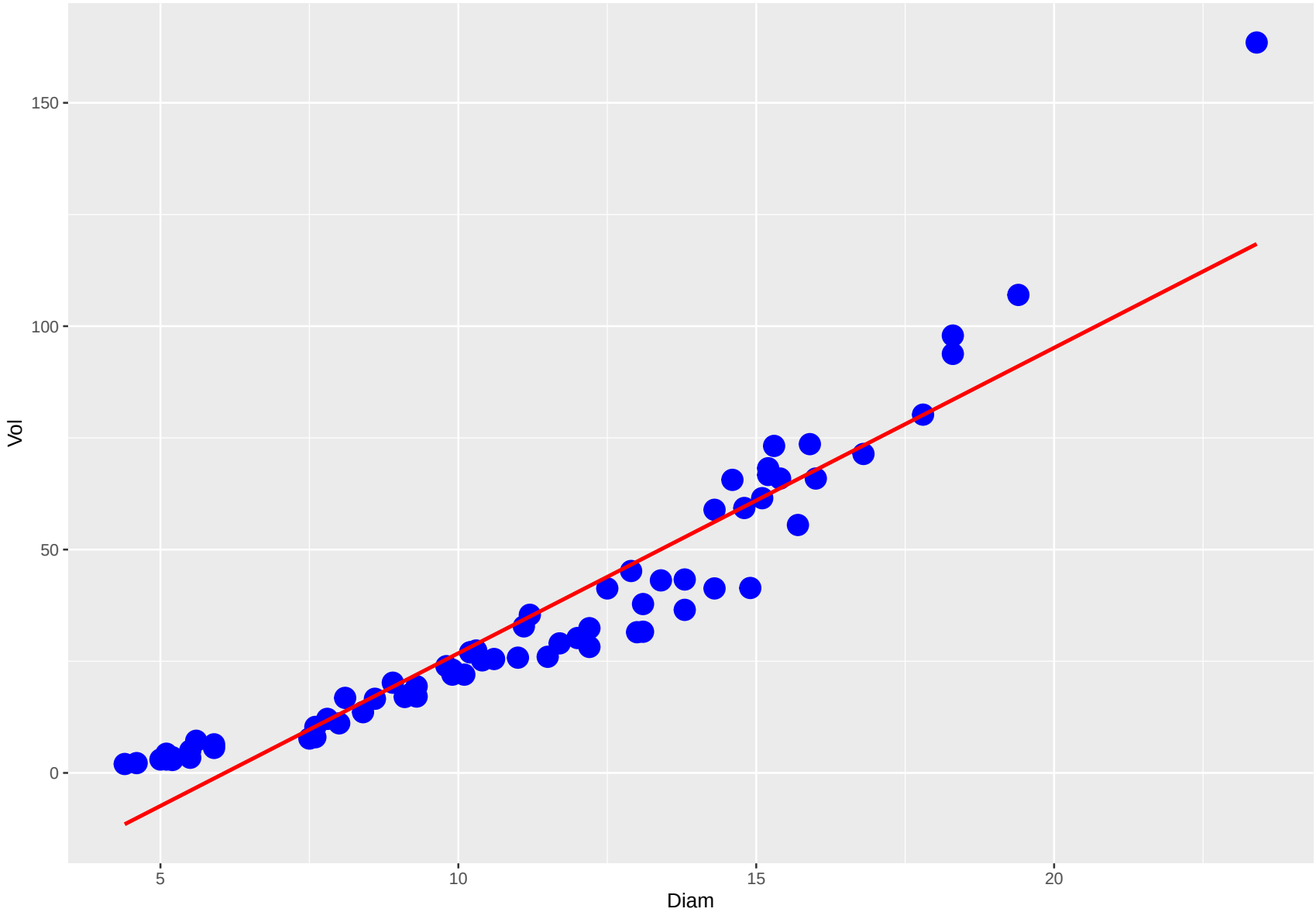
Non-linear relationships

Example: *Many different interest groups such as the lumber industry, ecologists, and foresters benefit from being able to predict the volume of a tree just by knowing its diameter. One classic data set (Short Leaf data) concerned the diameter (x, in inches) and volume (y, in cubic feet) of $n = 70$ shortleaf pines.*

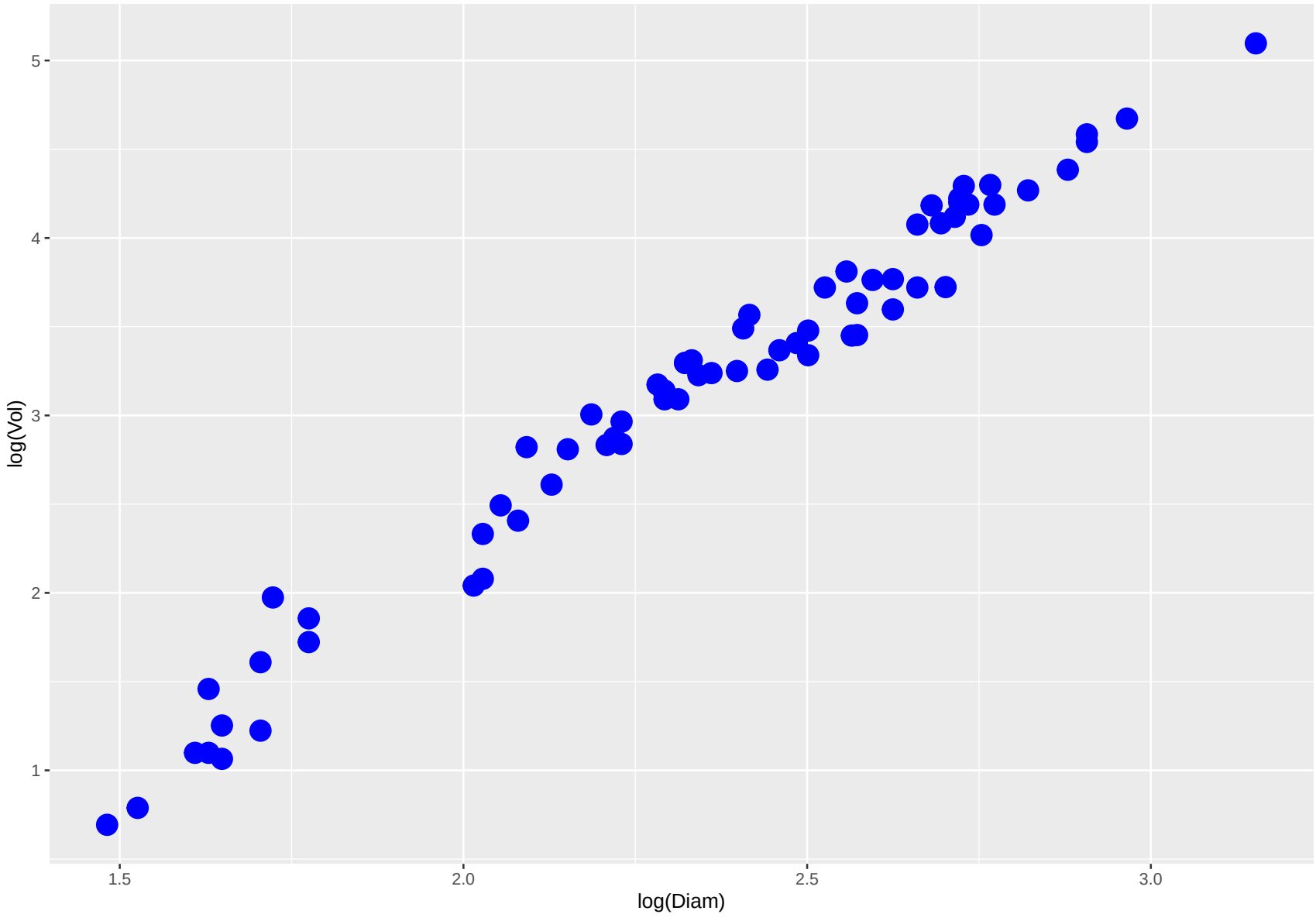
Non-linear relationships



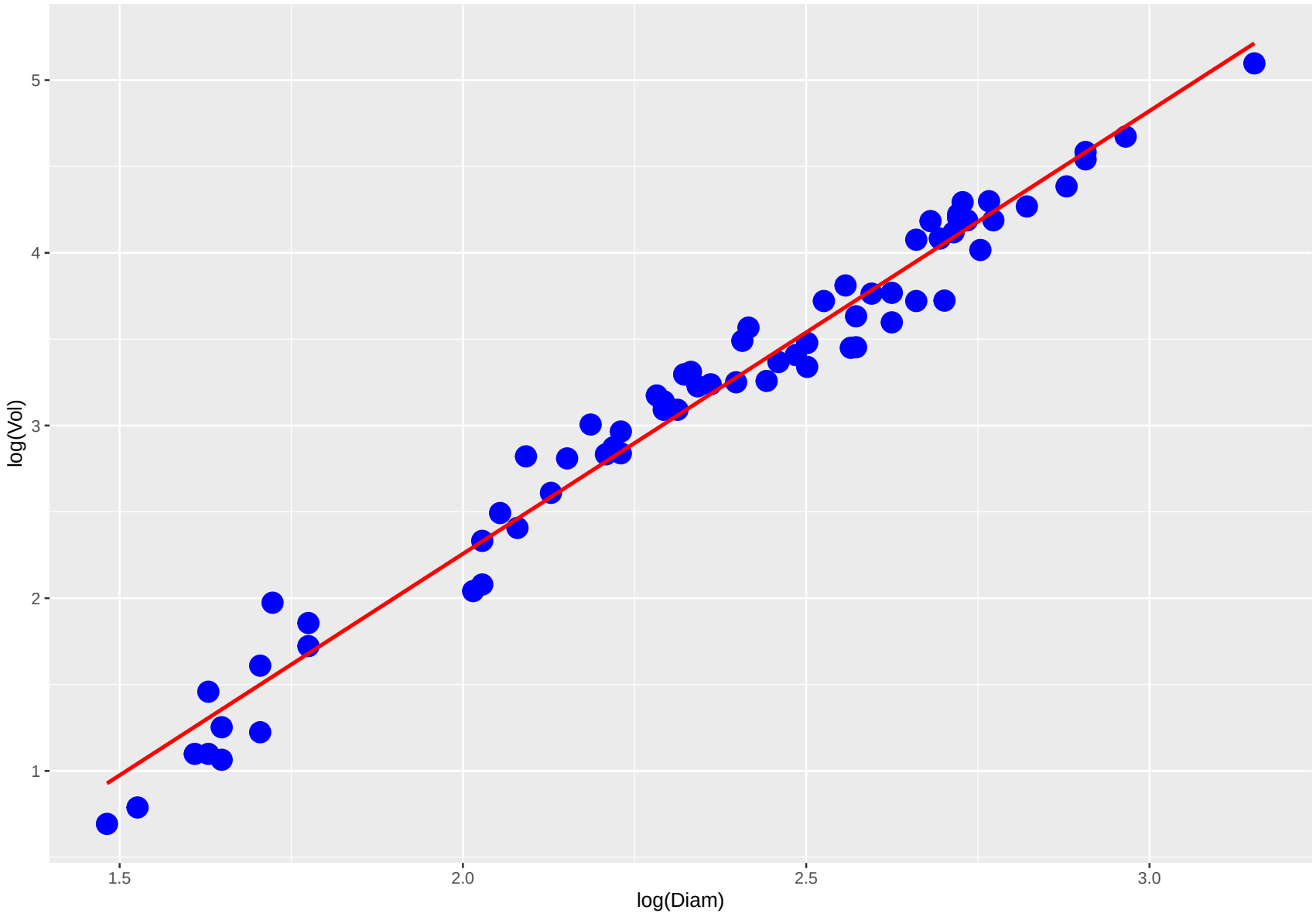
Non-linear relationships



Non-linear relationships



Non-linear relationships



Box-Cox transformation.

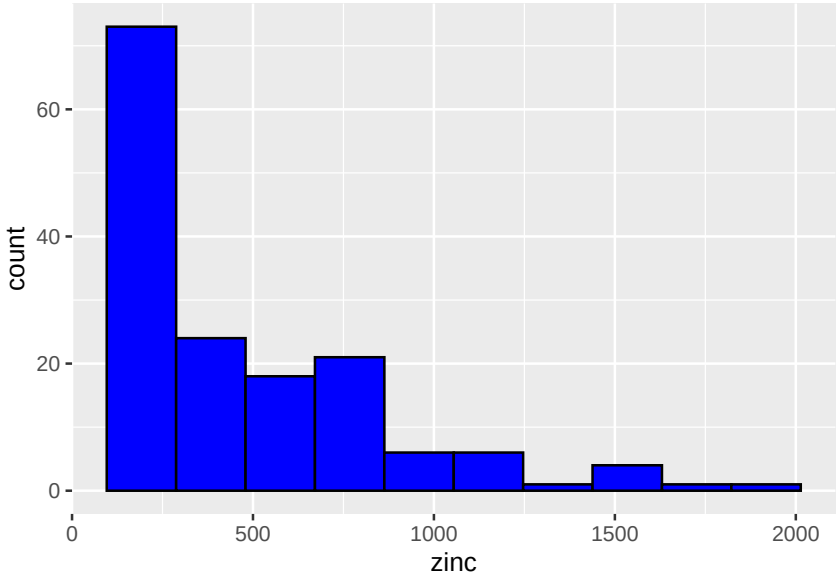
- Automatic transformation using Box–Cox transformation.

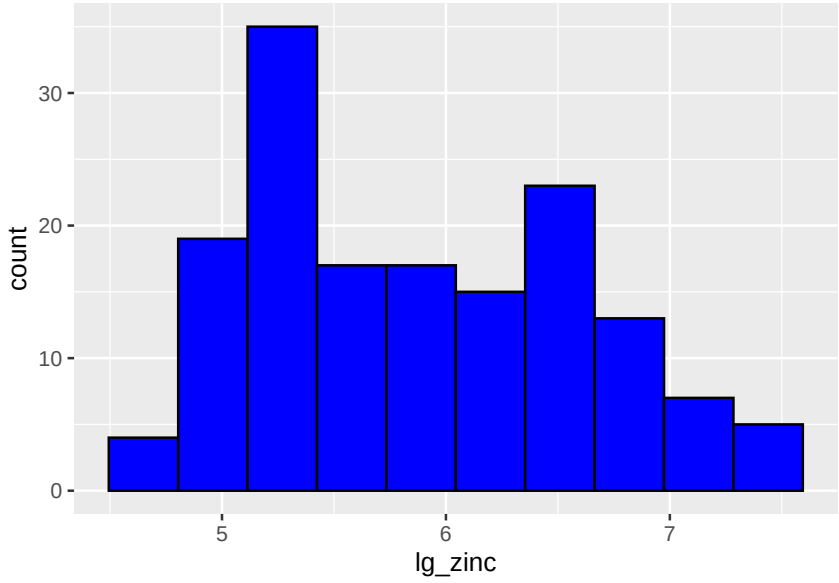
$$x^* = \begin{cases} \dfrac{x^\lambda - 1}{\lambda}, & if\ \lambda \neq 0 \\ \log(x), & if\ \lambda = 0 \end{cases}$$

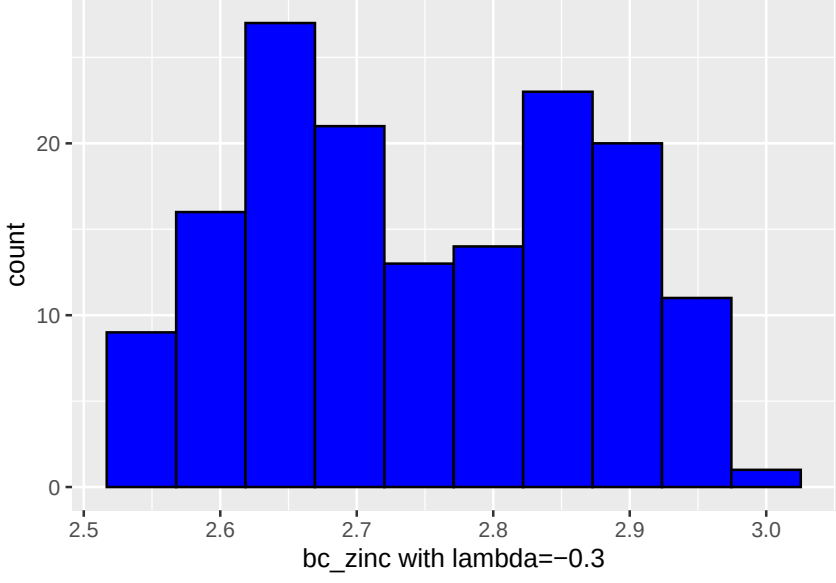
Box-Cox transformation.

meuse dataset gives locations and topsoil heavy metal concentrations, along with a number of soil and landscape variables at the observation locations, collected in a flood plain of the river Meuse, near the village of Stein (NL). Variable *zinc* in *meuse* contains the topsoil zinc concentration.

Box-Cox transformation







Box-Cox transformation

```
library(moments)

skewness(meuse$zinc)
```

```
## [1] 1.472038
```

```
skewness(meuse$lg_zinc)
```

```
## [1] 0.3258816
```

```
skewness(meuse$bc_zinc)
```

```
## [1] 0.05411839
```

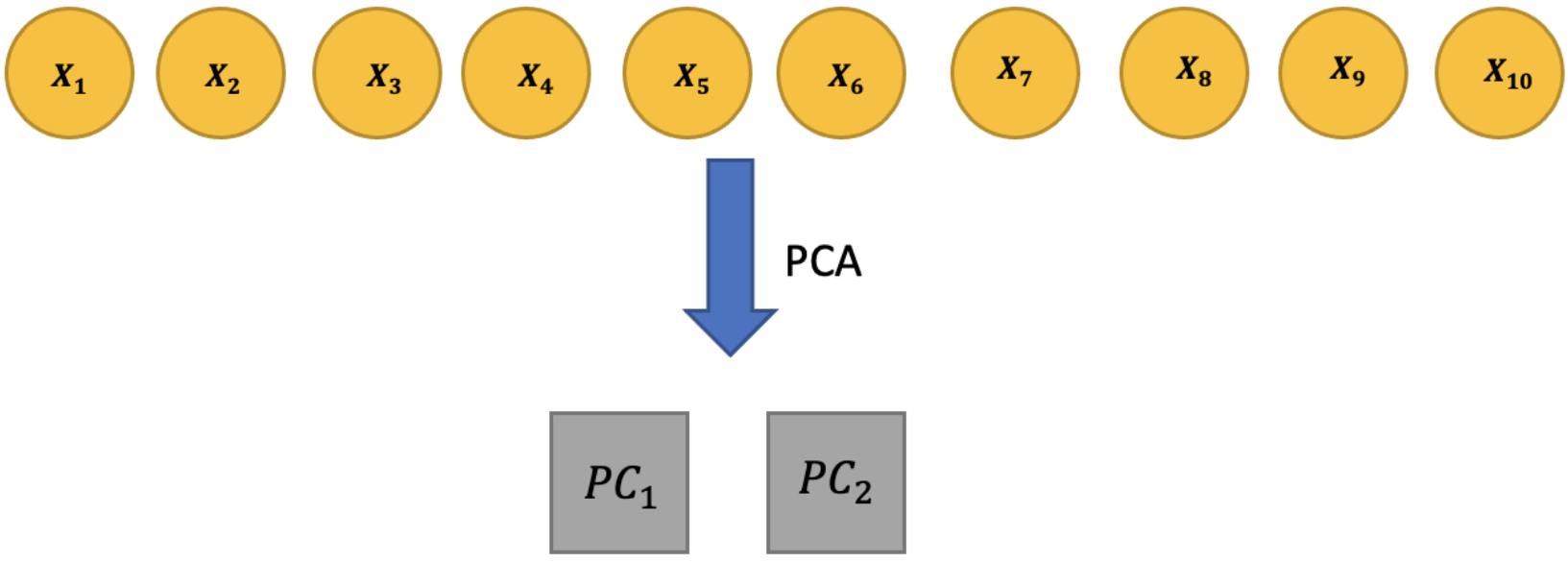
Do the following

1. Load *population* dataset.
2. Standardize the population variable.
3. Apply min-max scaling to the population variable.
4. Load *wordrecall* dataset.
5. Draw a scatter plot for time and prop.
6. Log transform data to get a linear relationship.
7. Load *meuse data* in package.
8. Apply box-cox transformation for zinc variable.

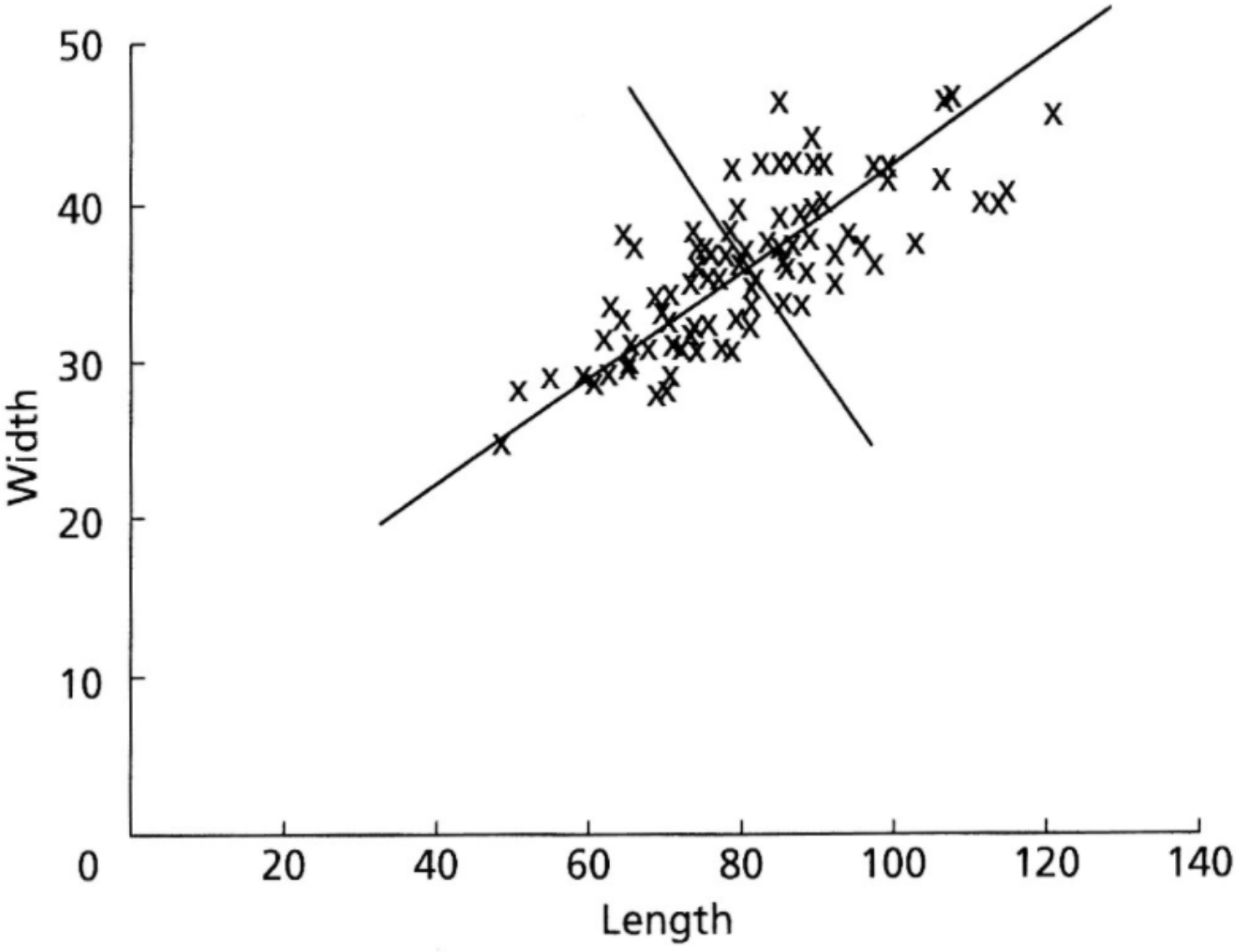
Principle Component Analysis (PCA)

- PCA, is a dimensionality-reduction method
- It is often used to reduce the dimensionality of large data sets.
- It transforms a large set of variables into a smaller one that still contains most of the information in the large set.

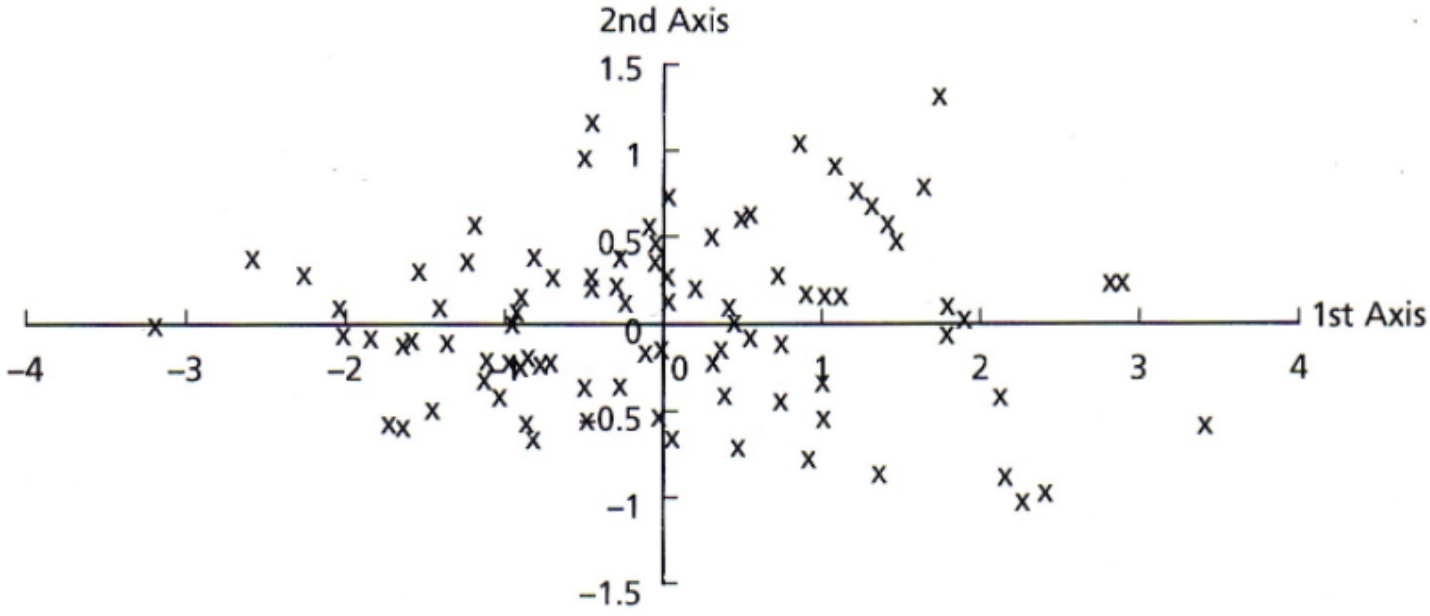
Dimension reduction



Dimension reduction



Dimenstion reduction



Check this out

https://setosa.io/ev/principal-component-analysis/

Computation

$$
\begin{aligned}
PC1 &= w_{11}X_1 + w_{12}X_2 + \ldots + w_{1p}X_p \\
PC2 &= w_{21}X_1 + w_{22}X_2 + \ldots + w_{2p}X_p \\
&\vdots \\
PCp &= w_{p1}X_1 + w_{p2}X_2 + \ldots + w_{pp}X_p
\end{aligned}
$$

Computation

PCA is just a rotation of the data. In matrix notation, the transformation of the original variables to the principal components is written as

$$\mathbf{PC}=\mathbf{XW}$$

Steps

Step 1: Standardize the dataset.

Step 2: Calculate the covariance matrix for the variables in the dataset.

Step 3: Calculate the eigenvalues and eigenvectors for the covariance matrix.

Step 4: Sort eigenvalues and their corresponding eigenvectors.

Step 5: Pick k eigenvalues and form a matrix of eigenvectors.

Step 6: Transform the original matrix.

How to find eigenvalues and eigenvectors?

https://www.mathsisfun.com/algebra/eigenvalue.html

Variance

From matrix W and matrix Sx, the variance-covariance matrix of the original data, the variance-covariance matrix of the principal components can be calculated:

$$S_{PC} = WS_X W^T$$

Advantages

- For p predictors, there are p(p-1)/2 scatterplots.
- As an example with p $= 15$ predictors, there would be 105 different scatterplots.

Advantages

Exploratory Data Analysis – We use PCA when we're first exploring a dataset and we want to understand which observations in the data are most similar to each other.

Principal Components Regression – We can also use PCA to calculate principal components that can then be used in principal components regression.

Multicollinearity - This type of regression is often used when multicollinearity exists between predictors in a dataset.

Tutorial 3 - PCA

- Go to practical sheet.

Tutorial 3 - PCA

1. Do you think the data should be normalised?

```
USArrests %>%
  summarise(mean_Murder=mean(Murder),mean_Assualt=mean(Assa
```

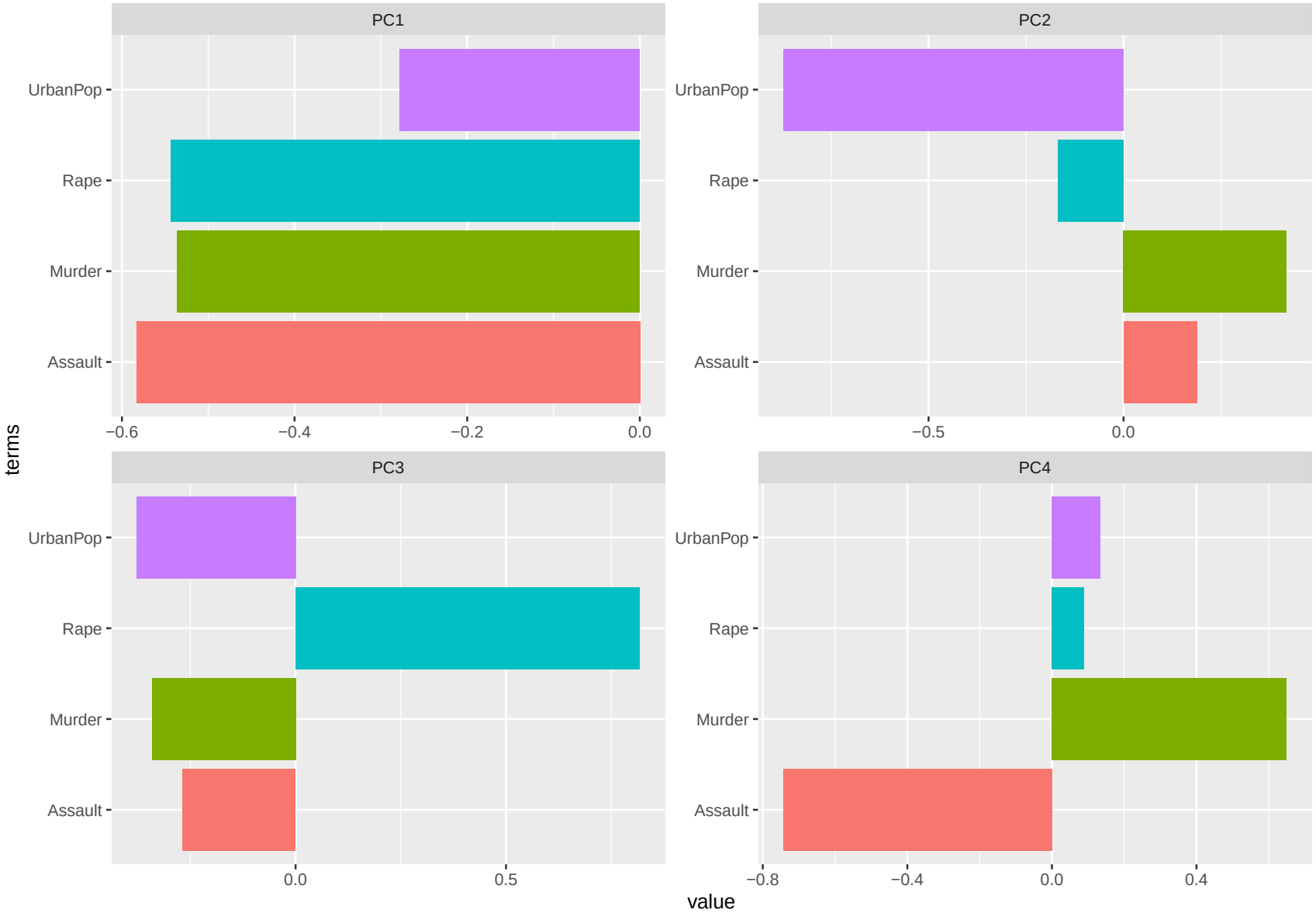
```
## # A tibble: 1 x 4
##   mean_Murder mean_Assualt mean_UrbanPop mean_Rape
##         <dbl>        <dbl>          <dbl>     <dbl>
## 1        7.79         171.           65.5      21.2
```

```
sd(USArrests$Murder); sd(USArrests$Assault)
```

```
## [1] 4.35551
```

```
## [1] 83.33766
```

PCA loadings



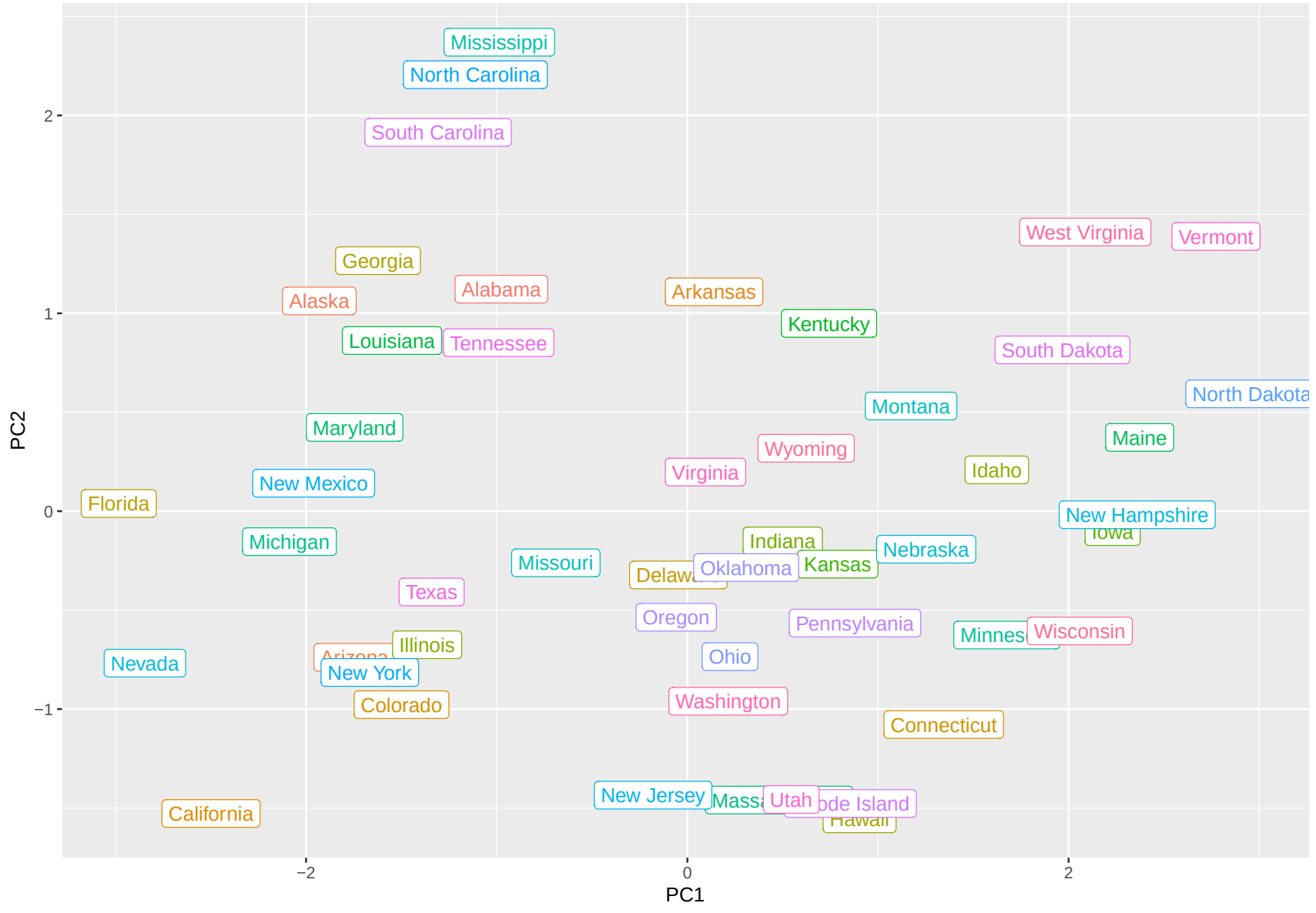
PCA Loadings

- What is the most influential variable in each component, i.e. which variable has the largest (in absolute value) loading value in each principal component?

PCA Loadings

- What is the most influential variable in each component, i.e. which variable has the largest (in absolute value) loading value in each principal component?
- PC1 – Assault
- PC2 – UrbanPop
- PC3 – Rape
- PC4 – Assault

Scatterplot of the principal components PC1 and PC2

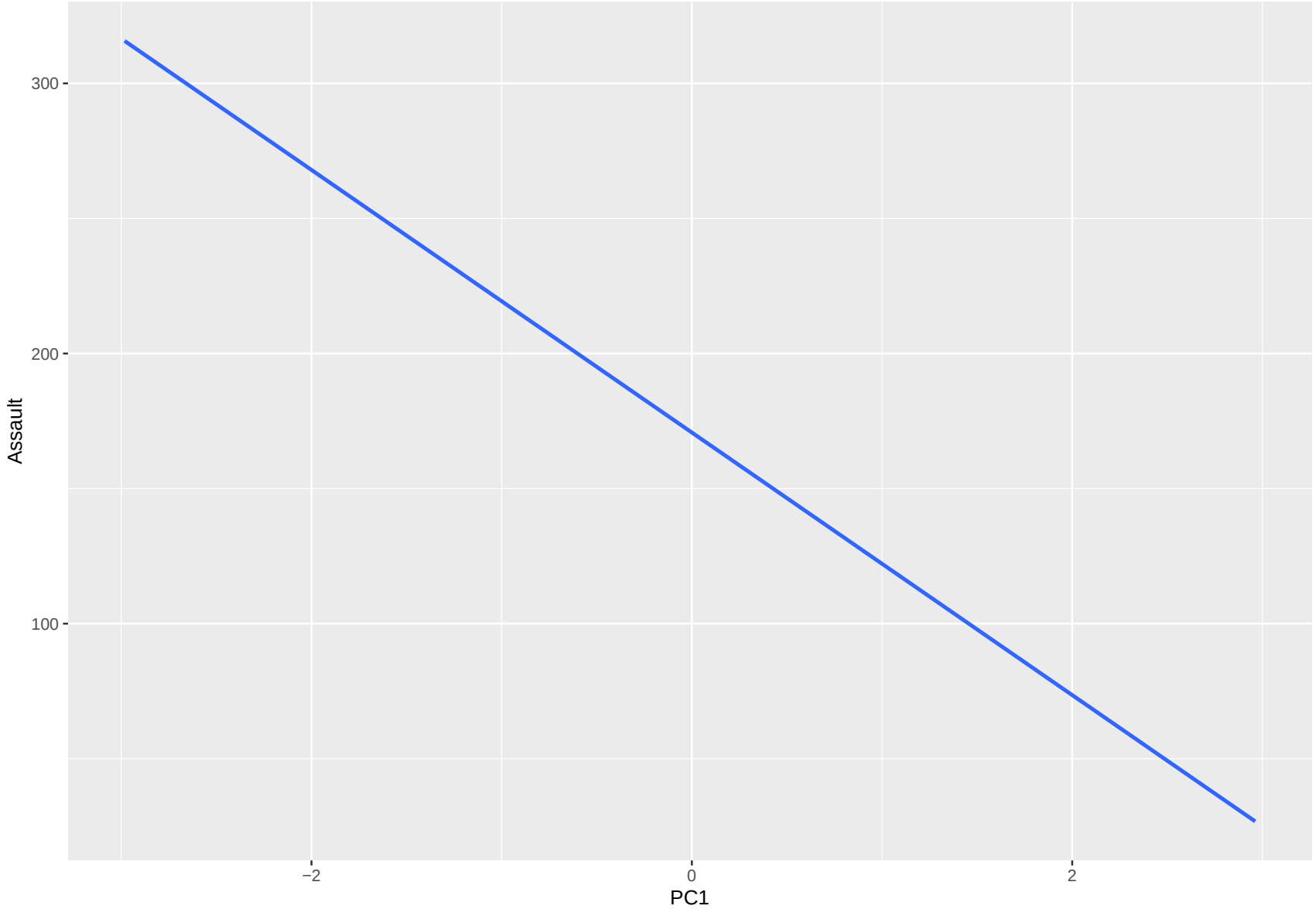


Scatterplot of the principal components PC1 and PC2

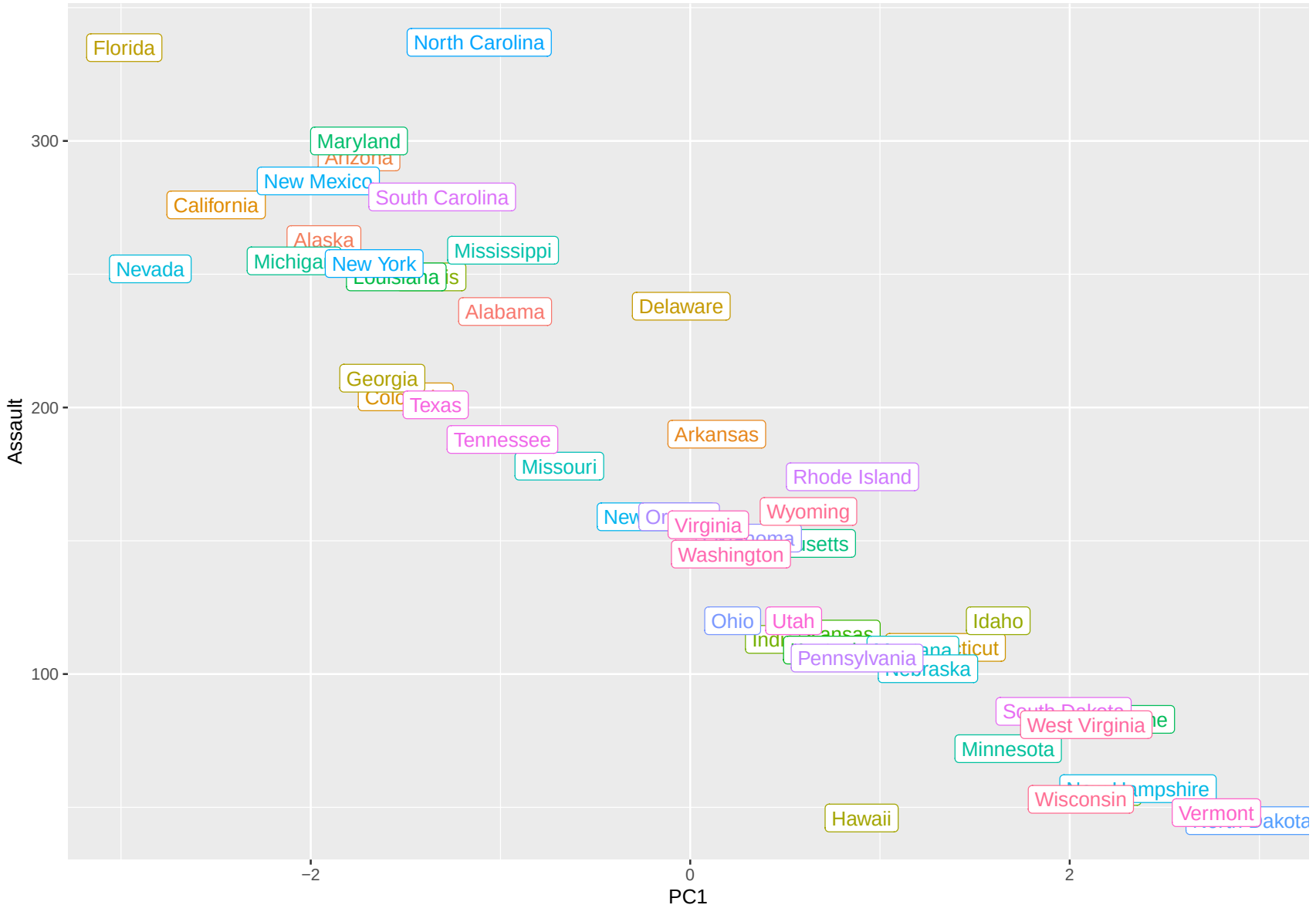
7.) Consider the points for Florida and Mississippi.

- Do you think Florida has an above- or below-average amount of arrest for assault per 100000?

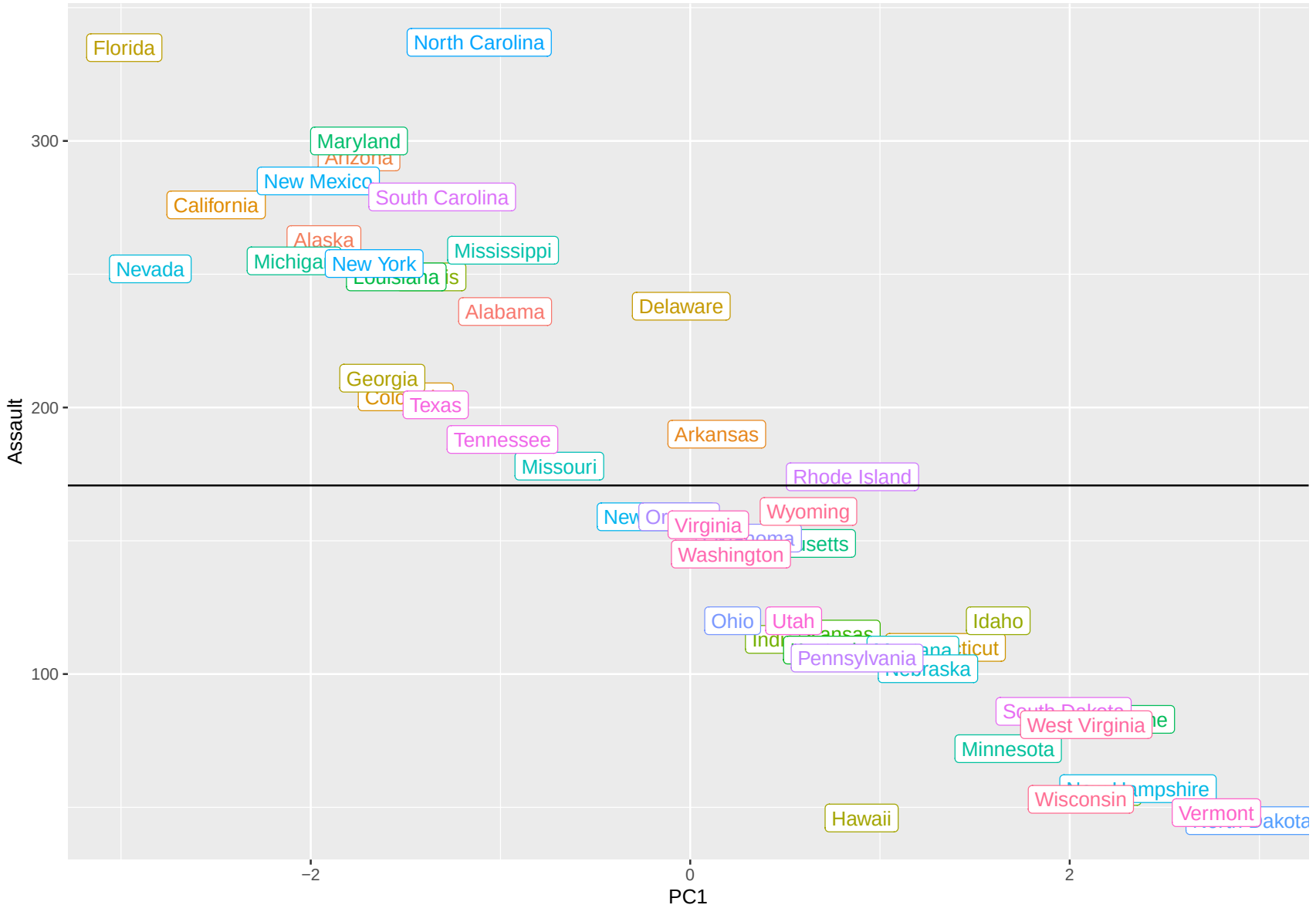
Relationship of PC1 and assault



Scatter plot of PC1 and assault



Scatter plot of PC1 and assault



Scaled value of assault

```
USArrests %>%
mutate_if( is.numeric, scale ) %>%
filter( state %in% c("Florida") ) %>%
  select(state,Assault)
```

```
## # A tibble: 1 x 2
##   state   Assault[,1]
##   <chr>         <dbl>
## 1 Florida        1.97
```

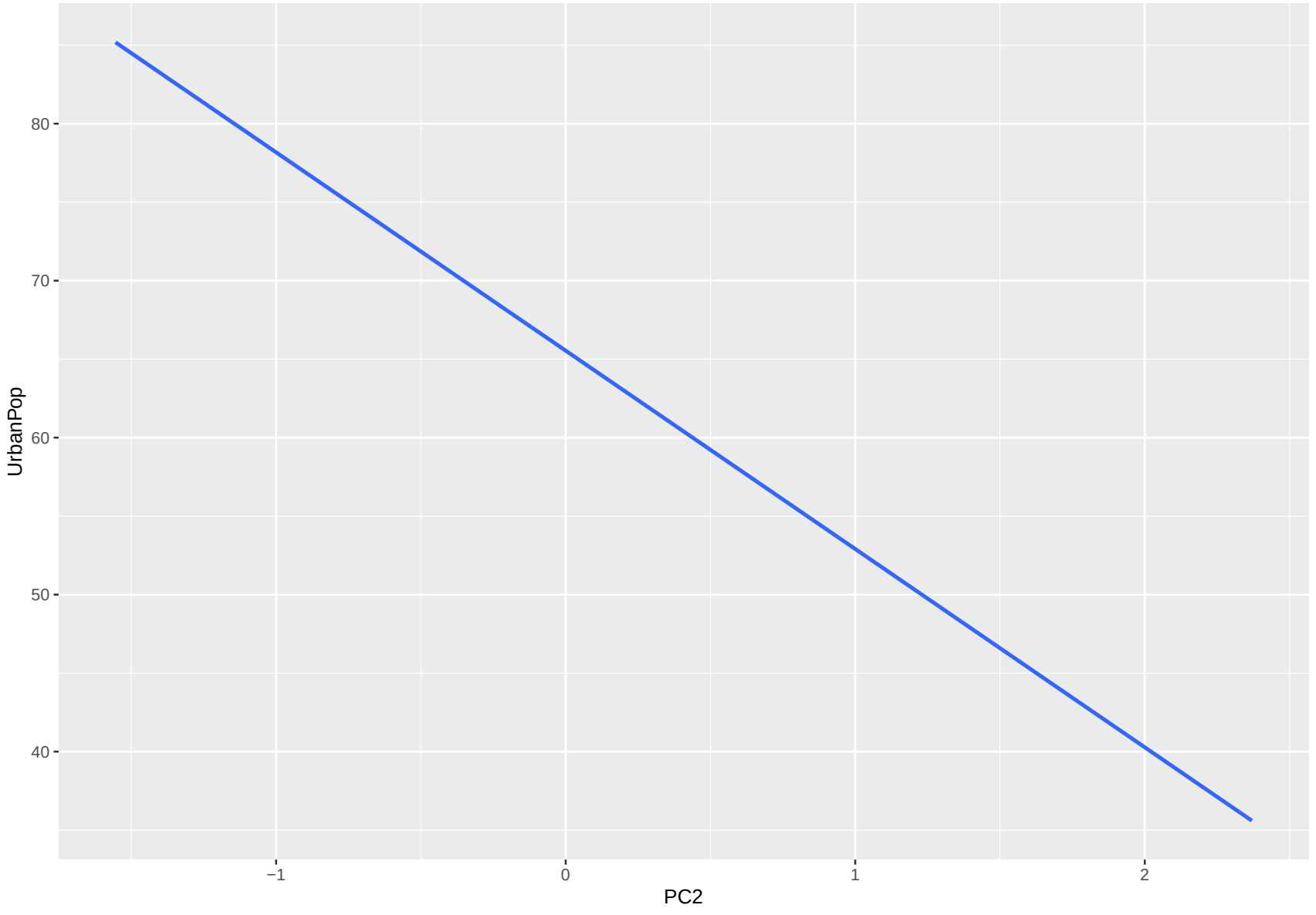
Question

- Do you think Florida has an above- or below-average amount of arrest for assault per 100000? Above average

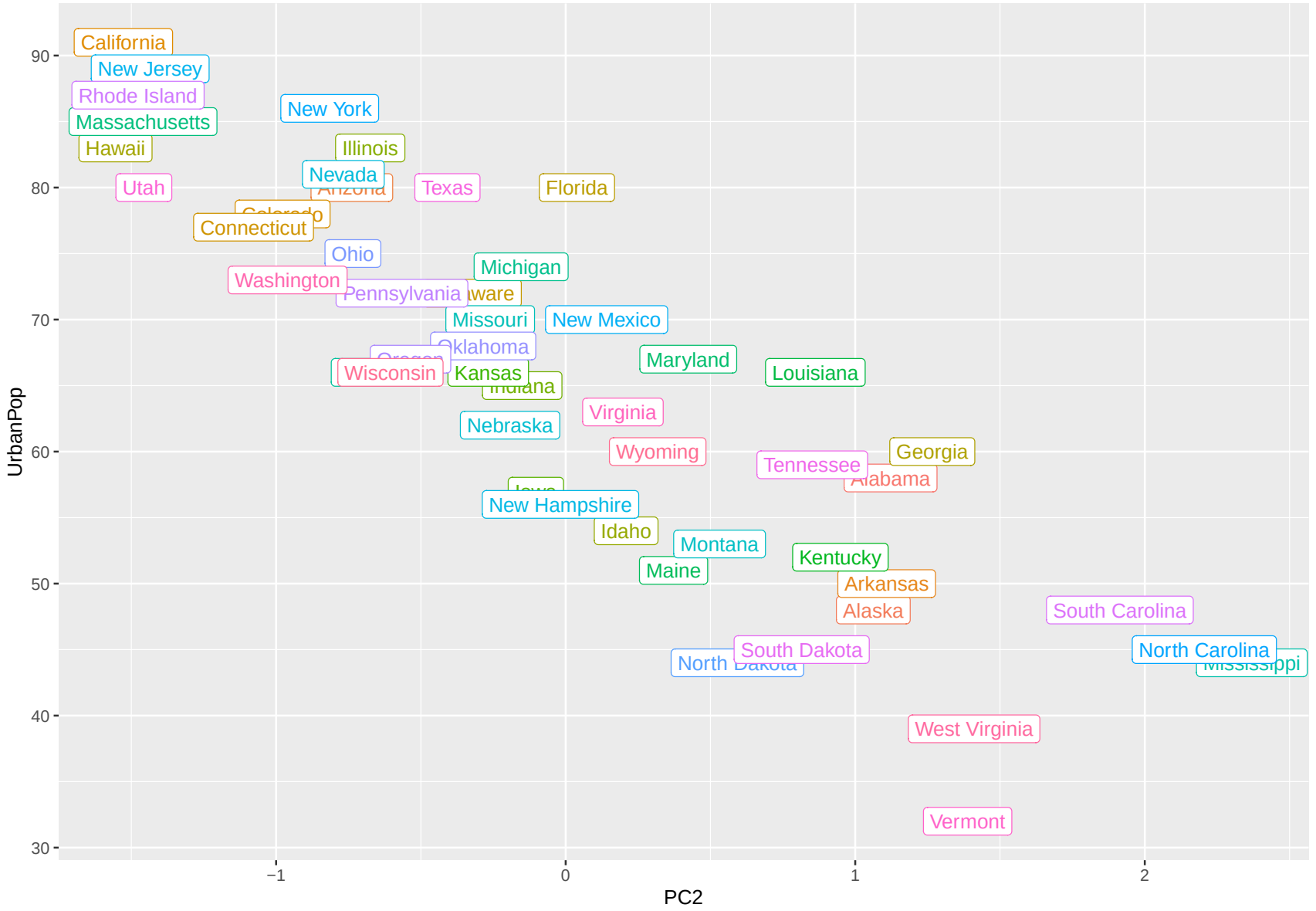
Question

- Do you think Mississippi has an above- or below-average percentage of population living in urban areas?

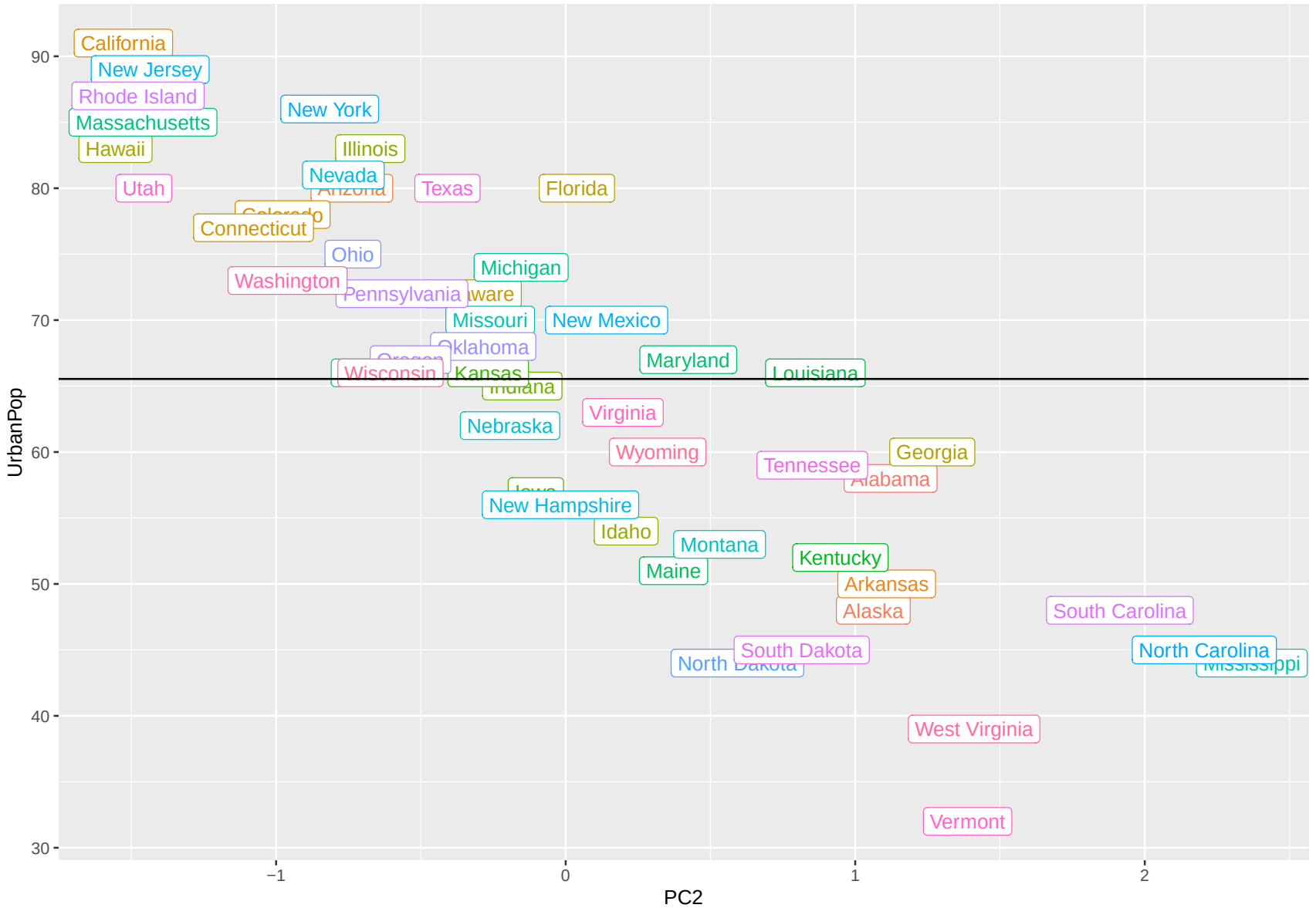
Relationship of PC2 and UrbanPop



Scatter plot of PC2 and UrbanPop



Scatter plot of PC2 and UrbanPopt



Scaled value of UrbanPop

```
USArrests %>%
mutate_if( is.numeric, scale ) %>%
filter( state %in% c("Mississippi") ) %>%
  select(state,UrbanPop)
```

```
## # A tibble: 1 x 2
##   state       UrbanPop[,1]
##   <chr>              <dbl>
## 1 Mississippi        -1.49
```

Question

- Do you think Mississippi has an above- or below-average percentage of population living in urban areas? Below average