Exploratory Data Analysis II

Roberta De Vito

SAPIENZA
UNIVERSITÀ DI ROMA

Why Exploratory Data Analysis (EDA)

- Identifying problems and unusual observations (i.e. outliers)
- Getting a feeling for the data: distribution of variables, relationships between variables etc.
- Understanding what questions can be answered using data.
- Model Selection
- Checking model assumptions

Four types of EDA

- Non graphical and univariate
- Graphical and univariate
- Non graphical and multivariate
- Graphical and multivariate

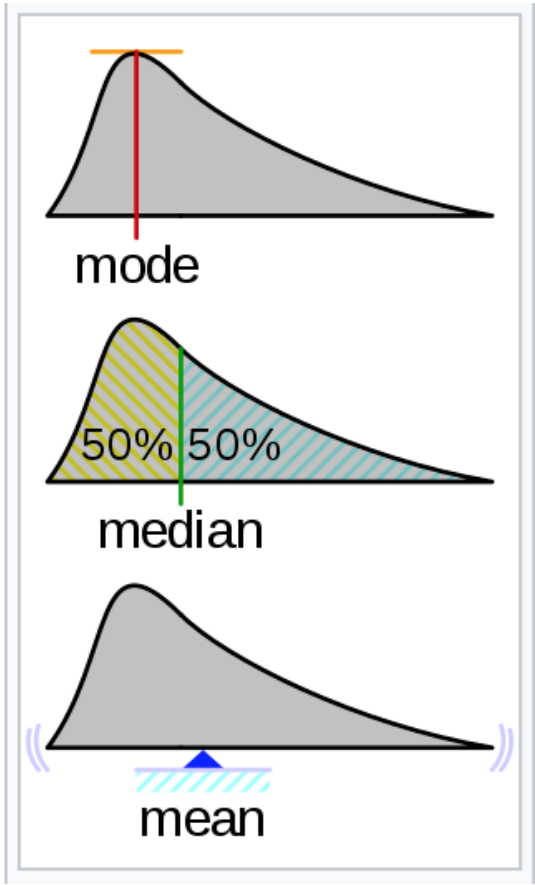
Four types of EDA

- Non graphical and univariate: central and spread
- Graphical and univariate
- Non graphical and multivariate
- Graphical and multivariate

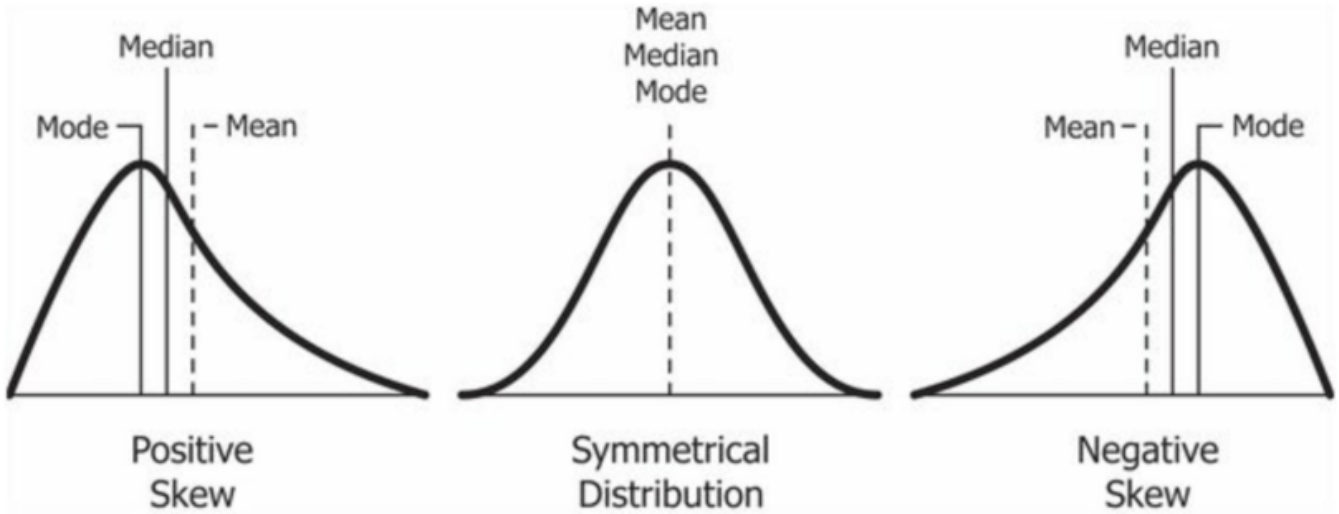
Non-graphical measures: Mostly for continuous covariates or ordinal with many levels

- Measures of central tendency: mean, median, mode.
- Measures of variability: variance, standard deviation, quantiles
- Measures of range: minimum and maximum values.

Difference between mean, median, and mode



Difference between mean, median, and mode

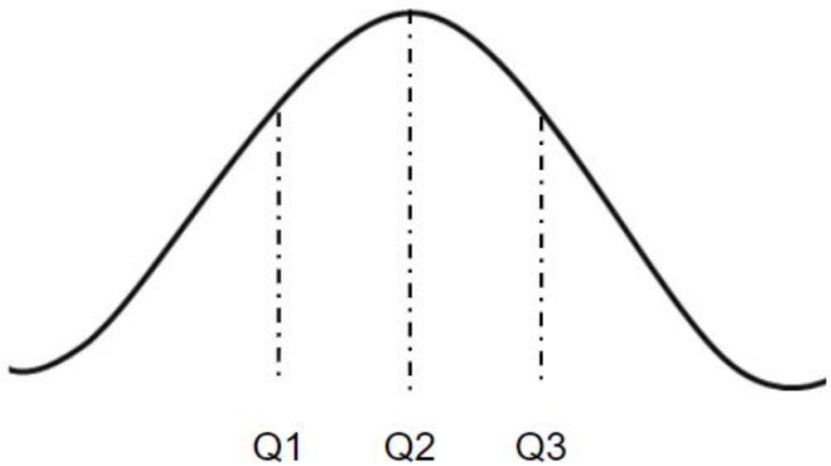


Spread: Interquantile

- divide the data in four: Q_1, Q_2, Q_3, Q_4
- IQR: half of the value falls in this interval
- data are more spread out: $IQR \uparrow$

Spread: Interquantile

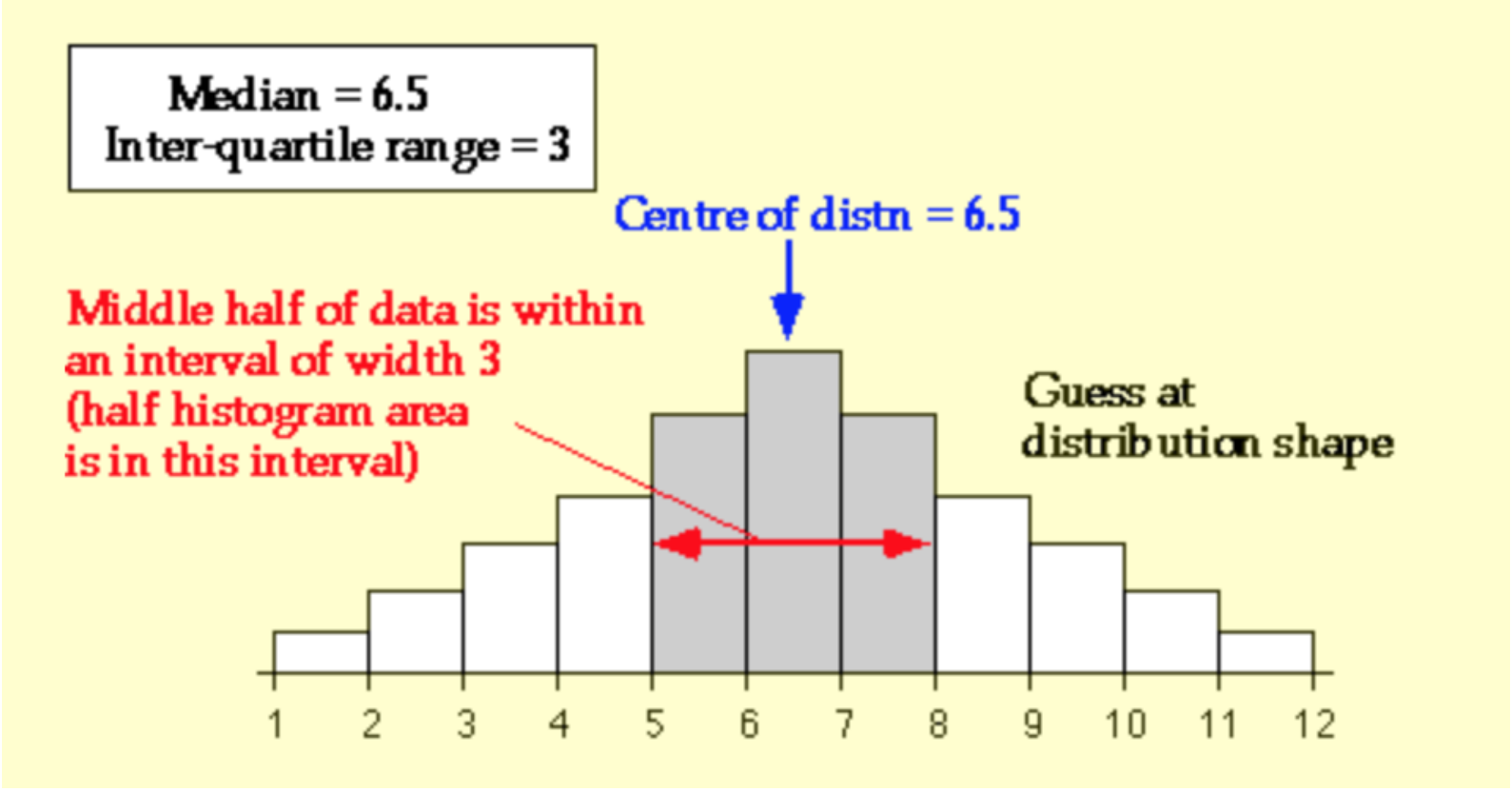
- divide the data in four: Q_1, Q_2, Q_3, Q_4
- IQR: half of the value falls in this interval
- data are more spread out: $IQR \uparrow$



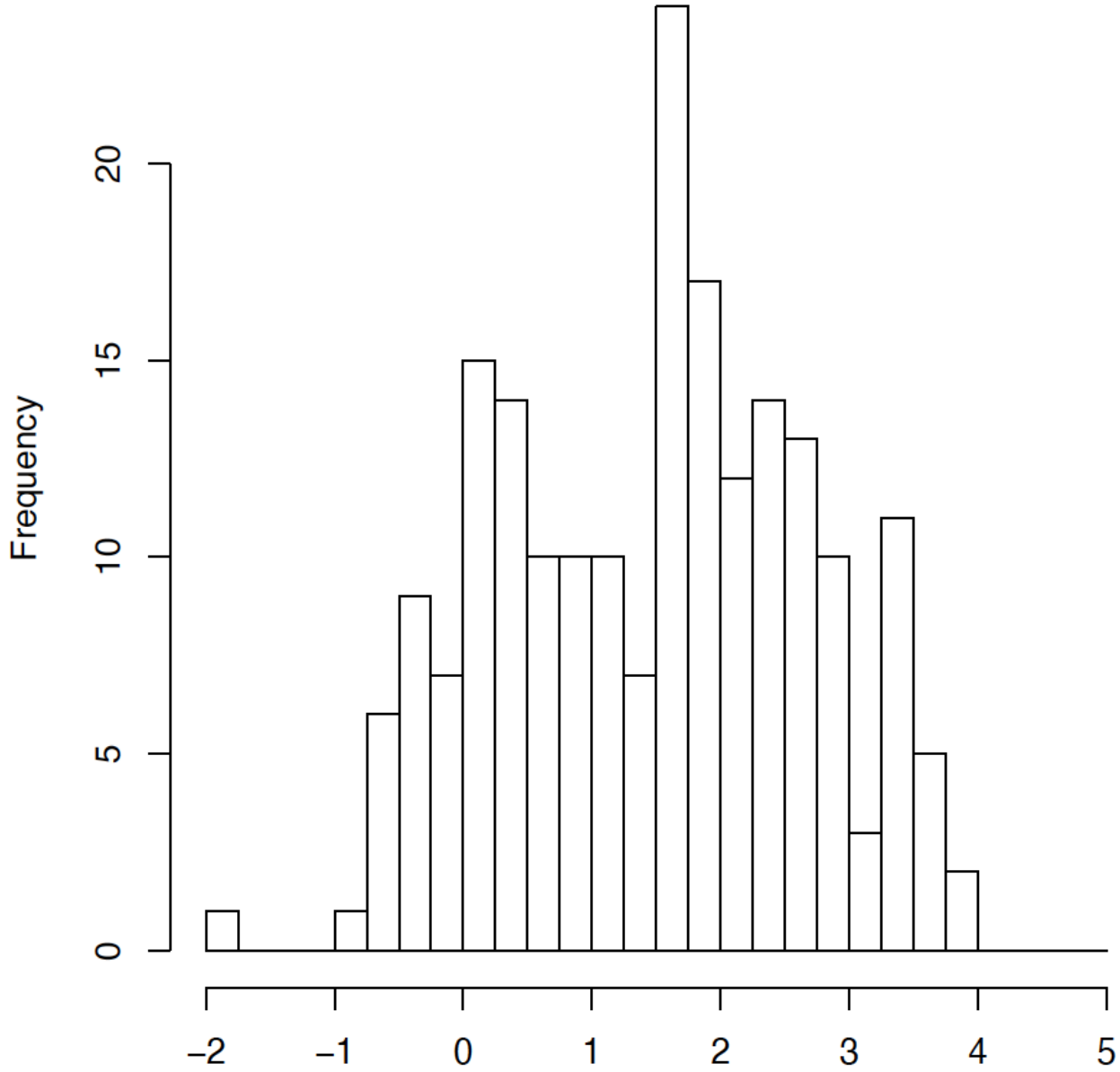
Four types of EDA

- Non graphical and univariate: central and spread
- Graphical and univariate: histogram, boxplots, and Q-Q plots
- Non graphical and multivariate
- Graphical and multivariate

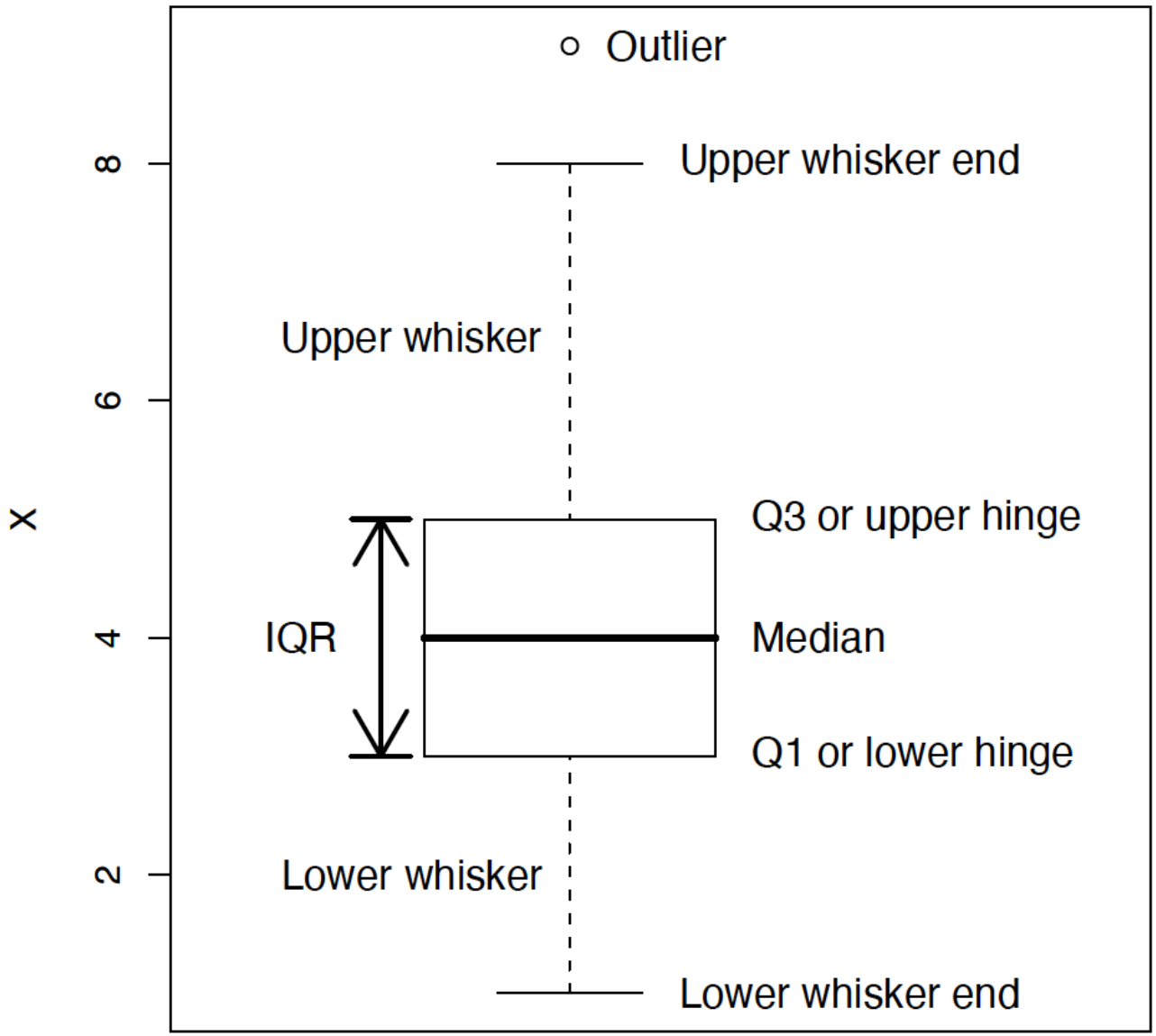
Histogram


Median = 6.5
Inter-quartile range = 3
Centre of distn = 6.5
Middle half of data is within an interval of width 3 (half histogram area is in this interval)
Guess at distribution shape
1
2
3
4
5
6
7
8
9
10
11
12

Histogram

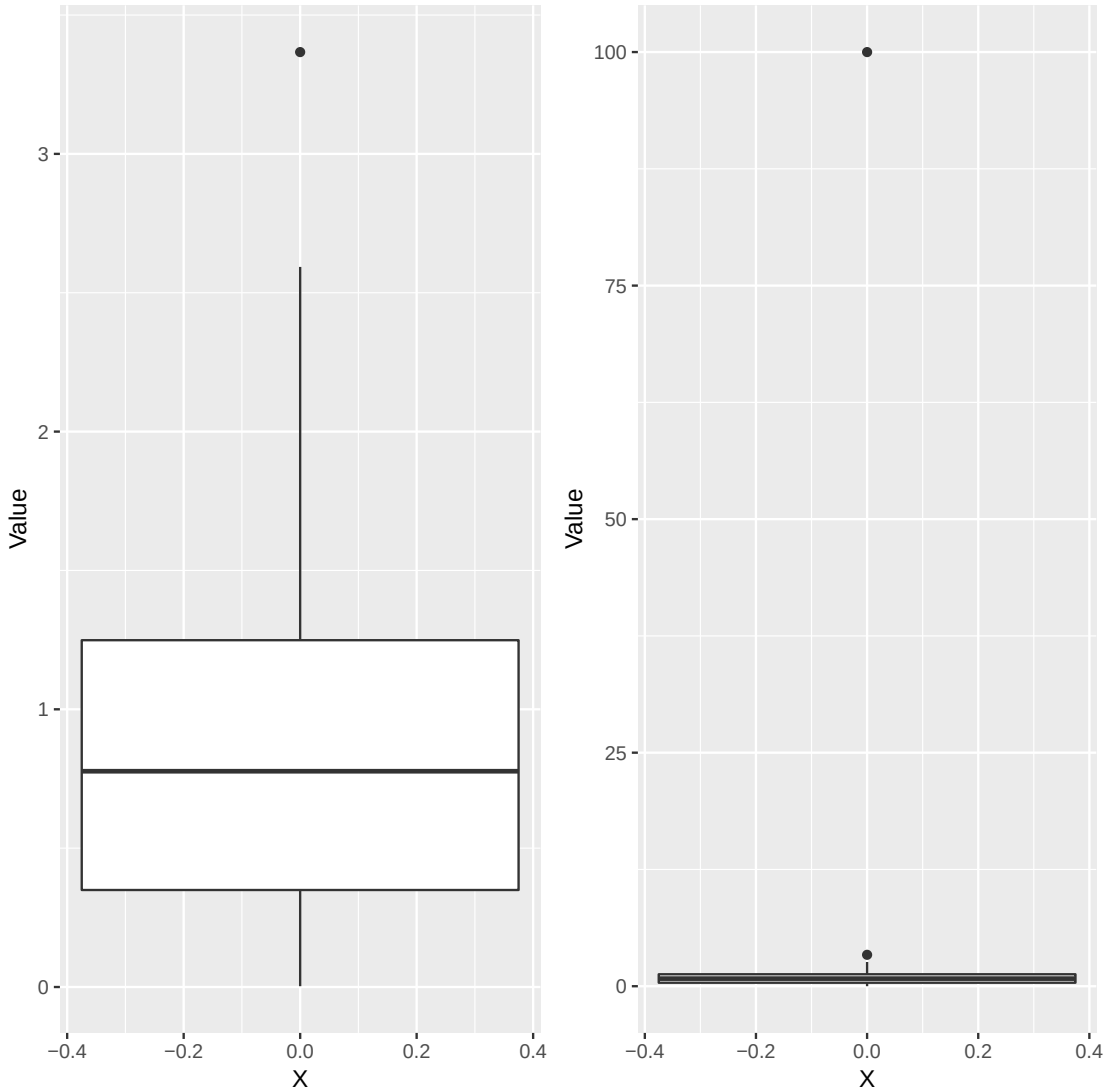


Boxplots


Outlier
Upper whisker end
Upper whisker
Q3 or upper hinge
IQR
Median
Q1 or lower hinge
Lower whisker
Lower whisker end
x
2
4
6
8

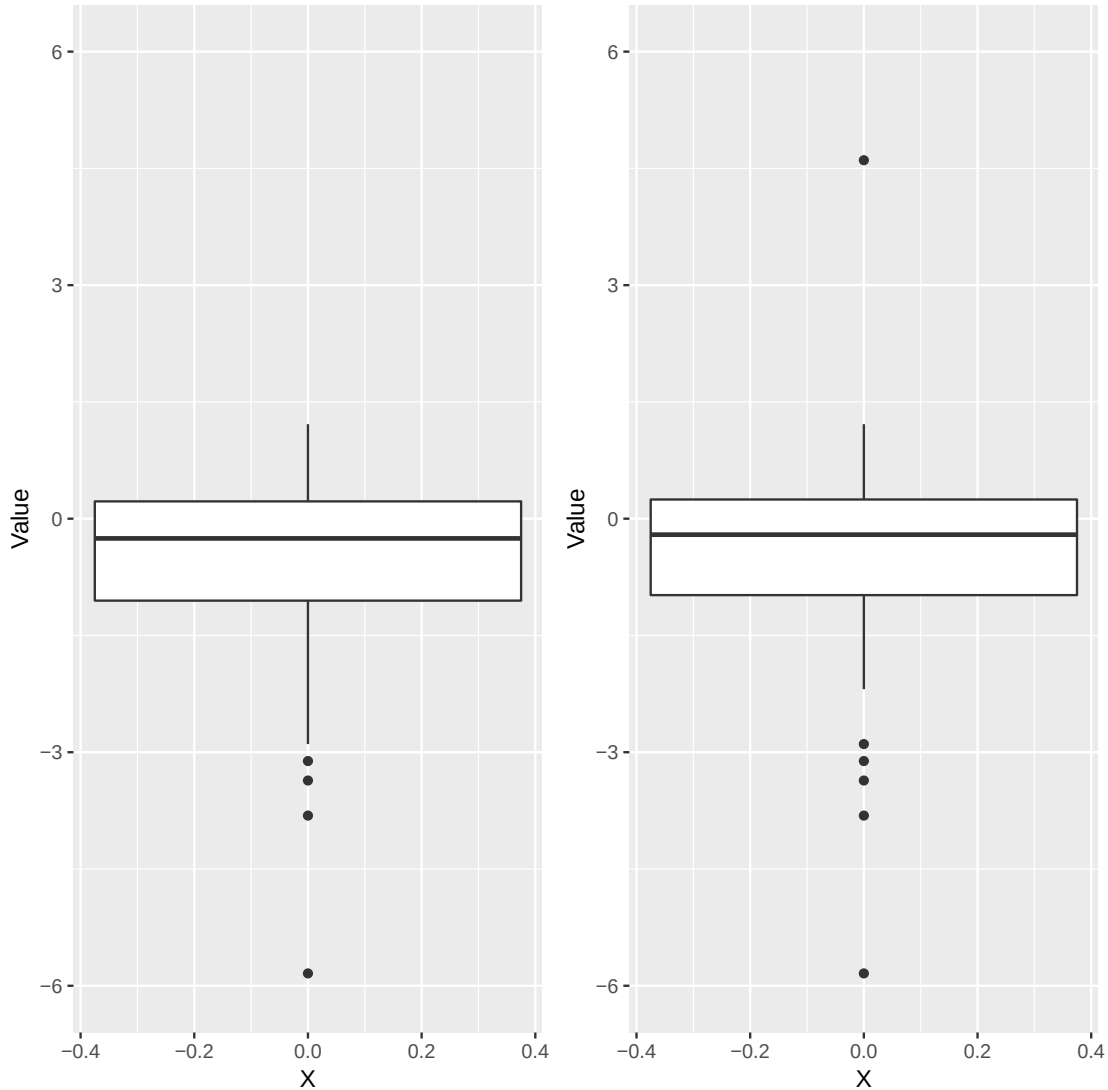
Outliers

Left without extreme outlier, right with extreme outlier.



On the log scale

When there are huge outliers plotting on the log-scale can be useful to better visualize the data.



Robustness to Outliers

Median is more robust (less sensitive) to outliers than the mean and is often used to present summary measures for variables that are known to have some outliers (e.g. income).

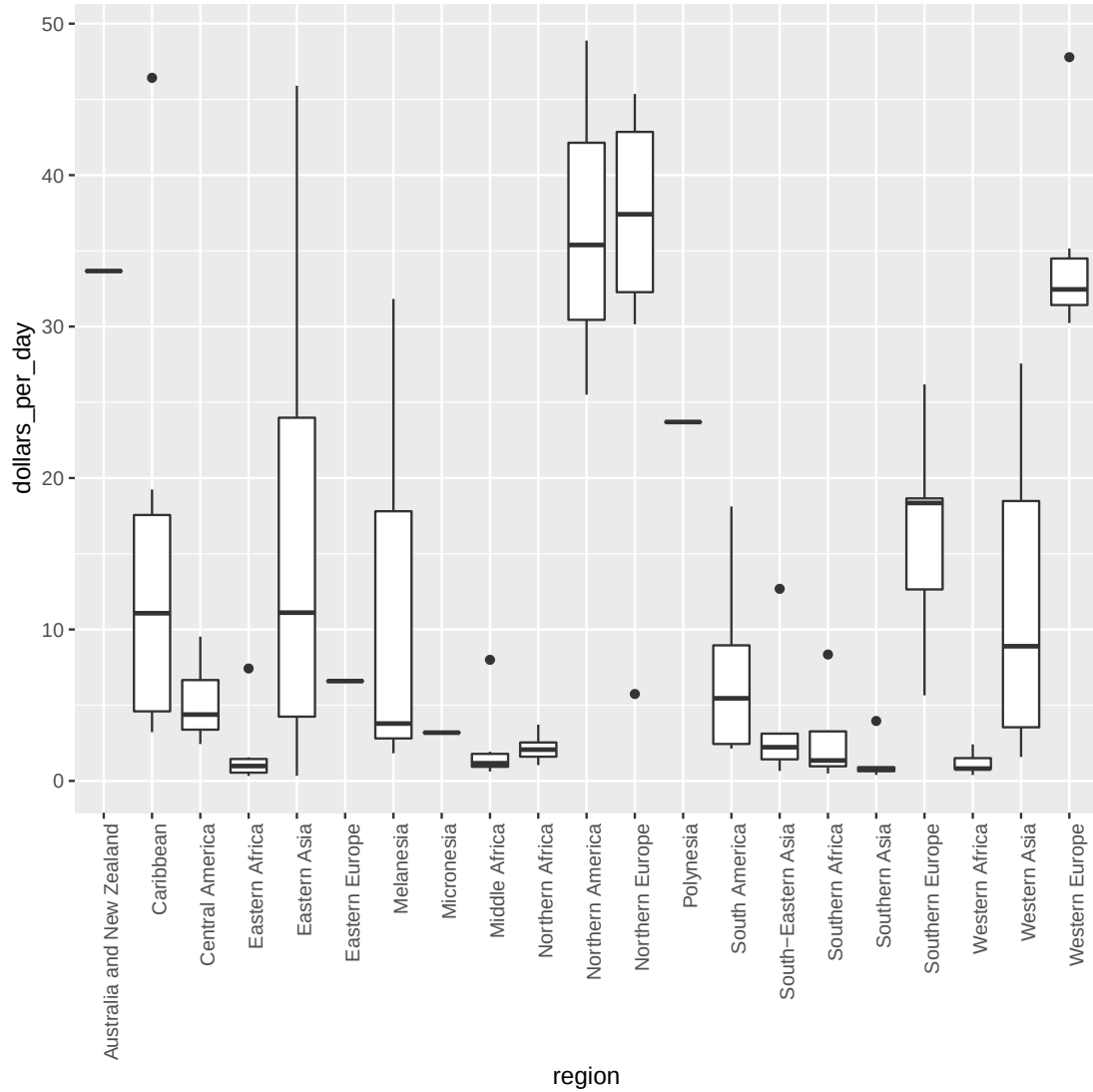
- Mean: With outlier 2.9, without outlier 0.9.
- Median: With outlier 0.8, without outlier 0.8.

Transformations

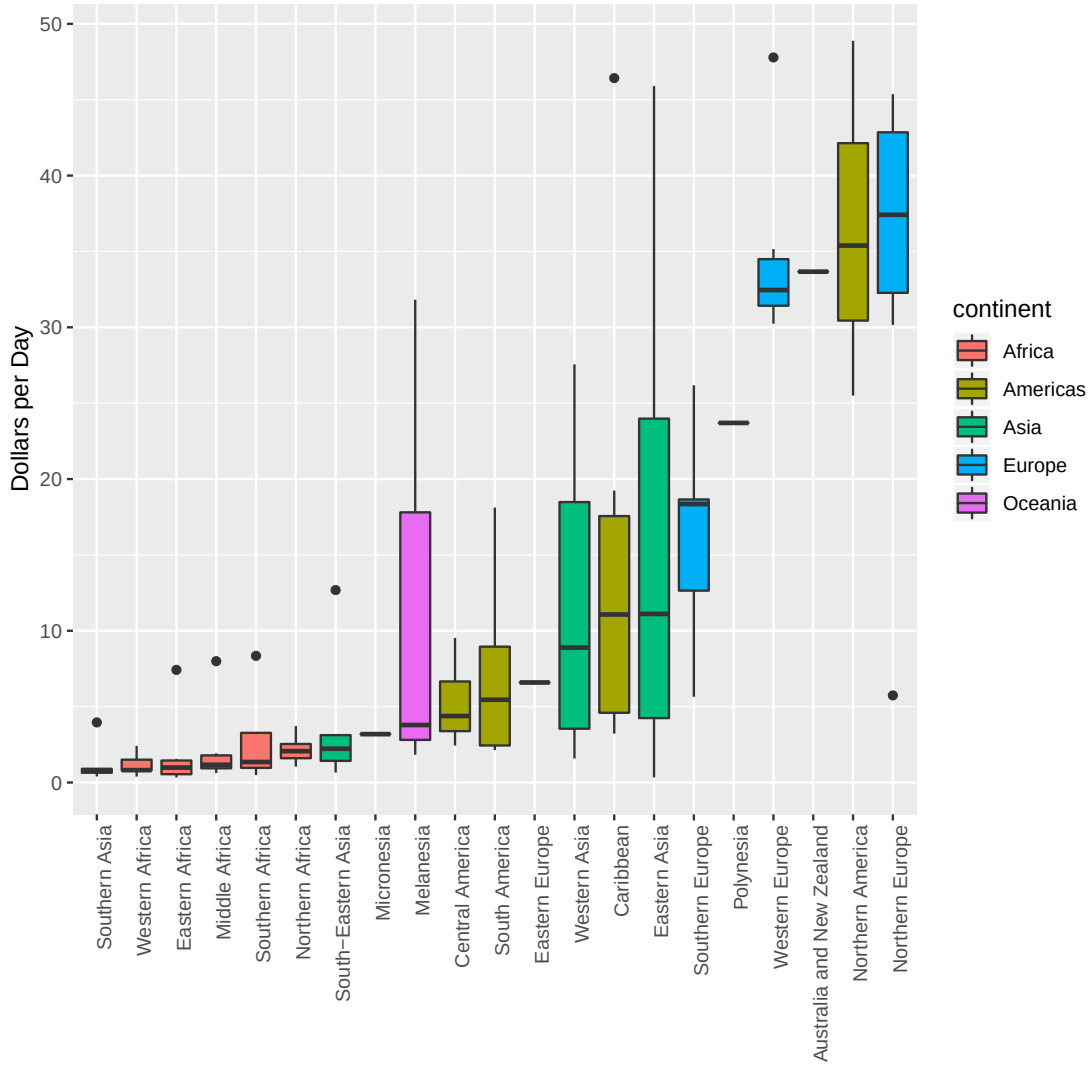
Transformations can be useful to make relationships clearer (e.g. when there are outliers).

- Log transformation the most popular and square root transformation common with count data.
- Usually monotone transformation.
- Chapter 21.8 in IDS book includes examples.

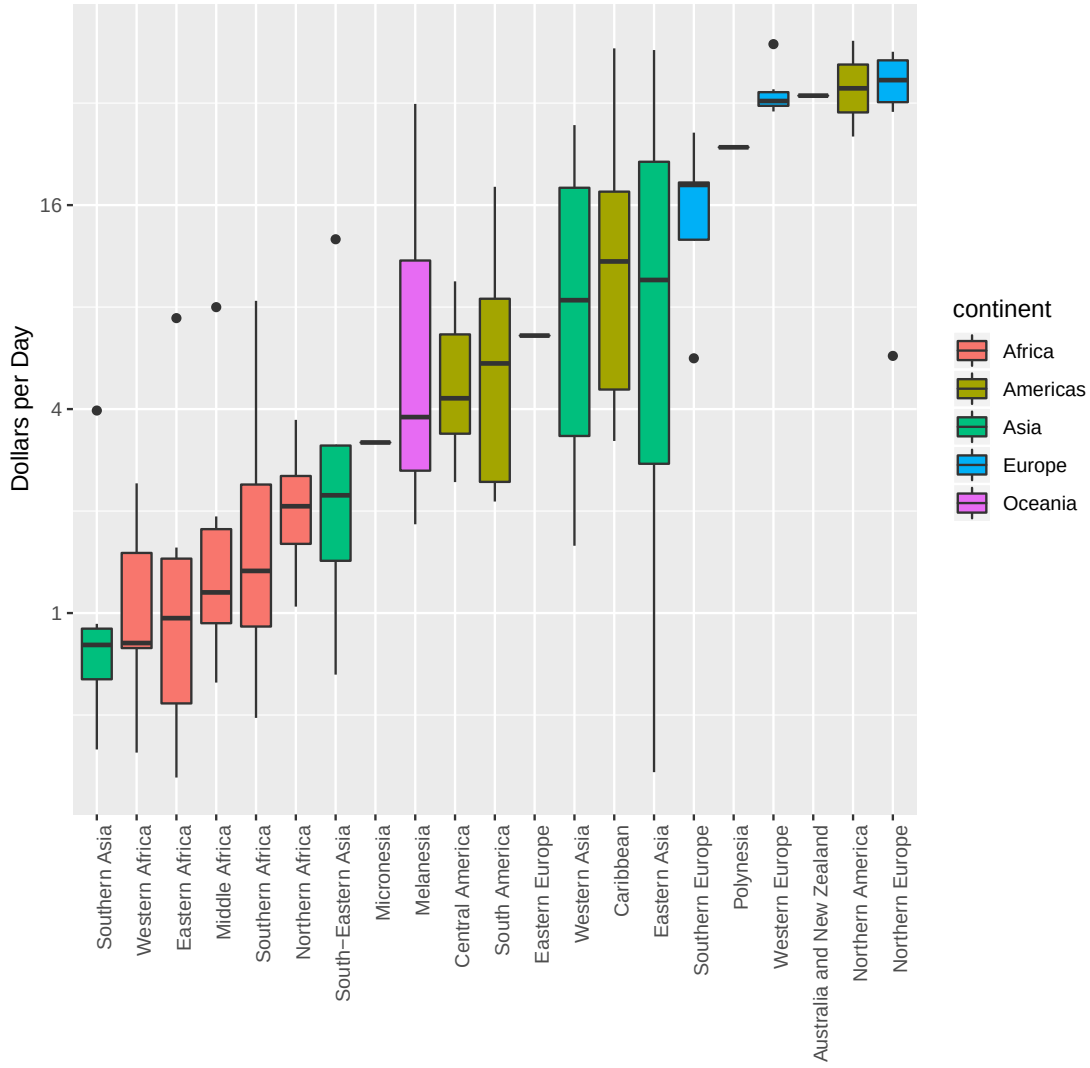
Plot of average pay per day by region (each data point is a country).



More informative plots, both in use of colors and ordering on y-axis.



Same plot on log-scale



Five Summary Measures

Tukey suggested the five summary measures

1. Minimum
2. 25th quantile
3. Median
4. 75th quantile
5. Maximum

This is roughly the information given by a boxplot.

QQ plots

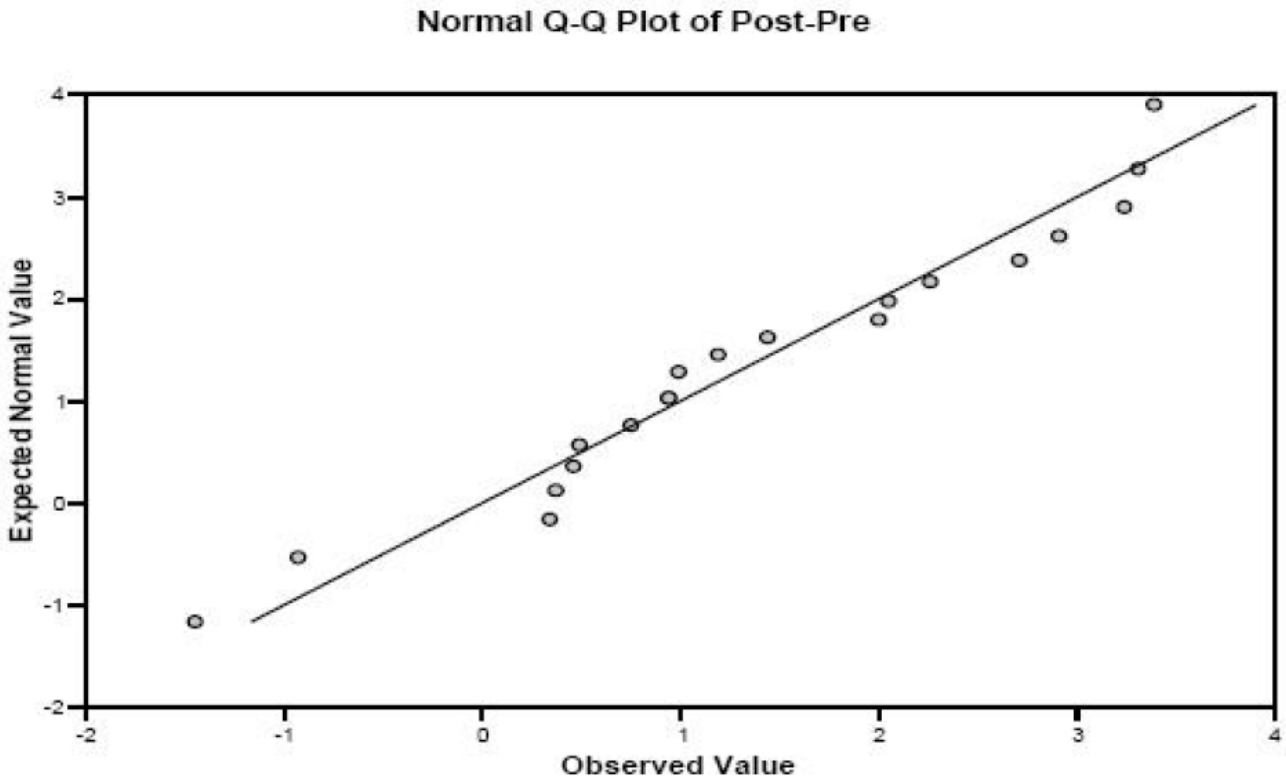
Used to visualize if a random sample comes from a specific distribution (e.g. normal distribution).

- Plots the theoretical quantiles to the observed quantiles.
- If distributional assumption is correct then should show approximately a straight line.

What is a QQ plot?

1. Let $\varepsilon_{(1)}, \ldots, \varepsilon_{(n)}$ be the ordered residuals with $\varepsilon_{(1)} \leq \varepsilon_{(2)} \leq \cdots \leq \varepsilon_{(n)}$.
2. Assume the ε are standardized by subtracting mean and dividing by standard error. This ensures that they have mean zero and variance one. Then, the distribution to compare to is a $\mathcal{N}(0, 1)$.
3. If the ε-s come from a $\mathcal{N}(0, 1)$ distribution, we expect $\varepsilon_{(k)}$ to be approximately equal to the $\frac{k}{n}$-th quantile of the $\mathcal{N}(0, 1)$.
4. A qq-plot plots the observed quantiles vs the theoretical quantiles. If points fall on a straight line, indication of the sample coming from a normal distribution.

QQ plot in practice



Four types of EDA

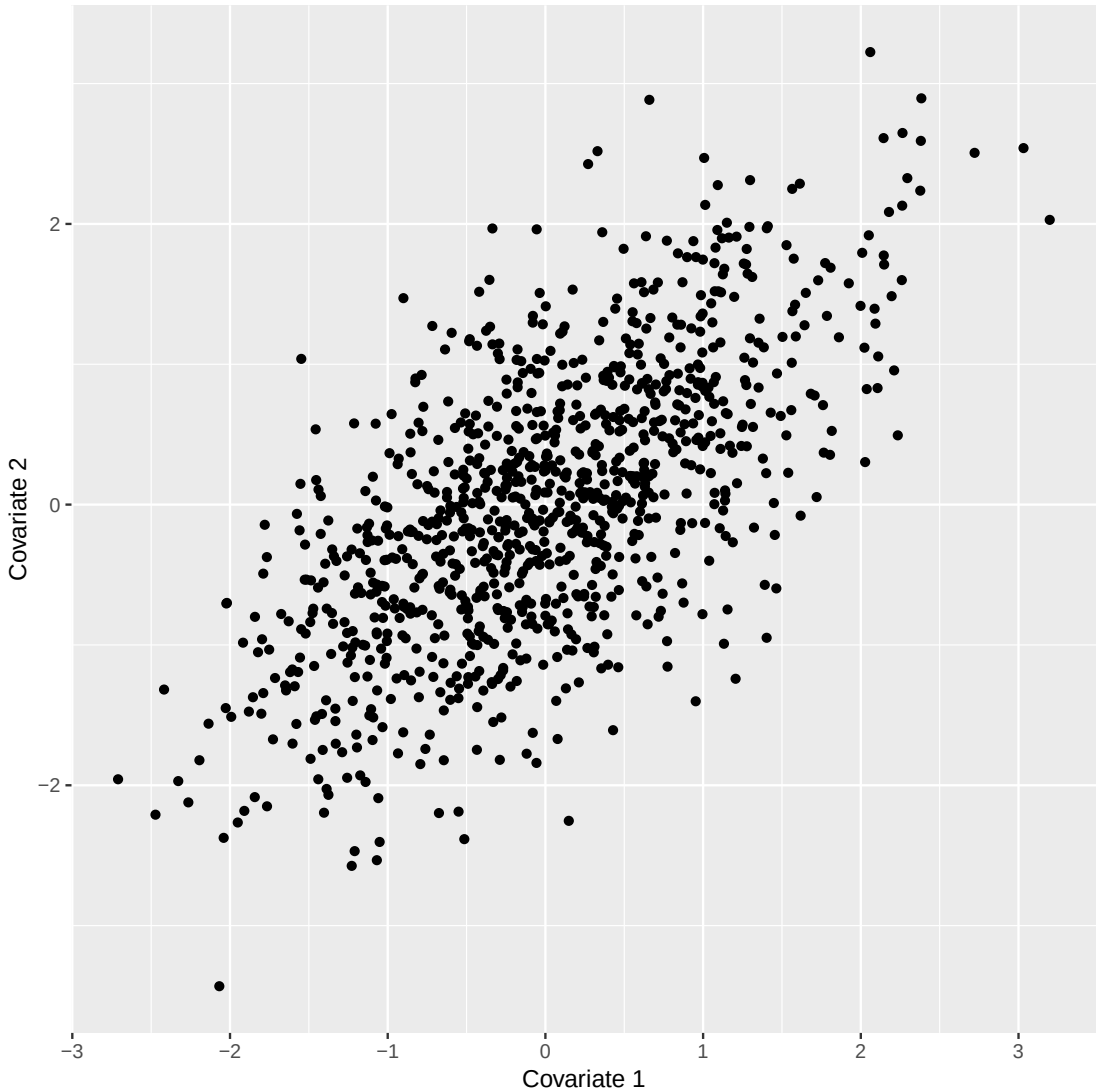
- Non graphical and univariate: central and spread
- Graphical and univariate: histogram, boxplots, and Q-Q plots
- Non graphical and multivariate: cross-tabulation and correlation
- Graphical and multivariate

Four types of EDA

- Non graphical and univariate: central and spread
- Graphical and univariate: histogram, boxplots, and Q-Q plots
- Non graphical and multivariate: cross-tabulation and correlation
- Graphical and multivariate: grouped barplots, and scatter plots

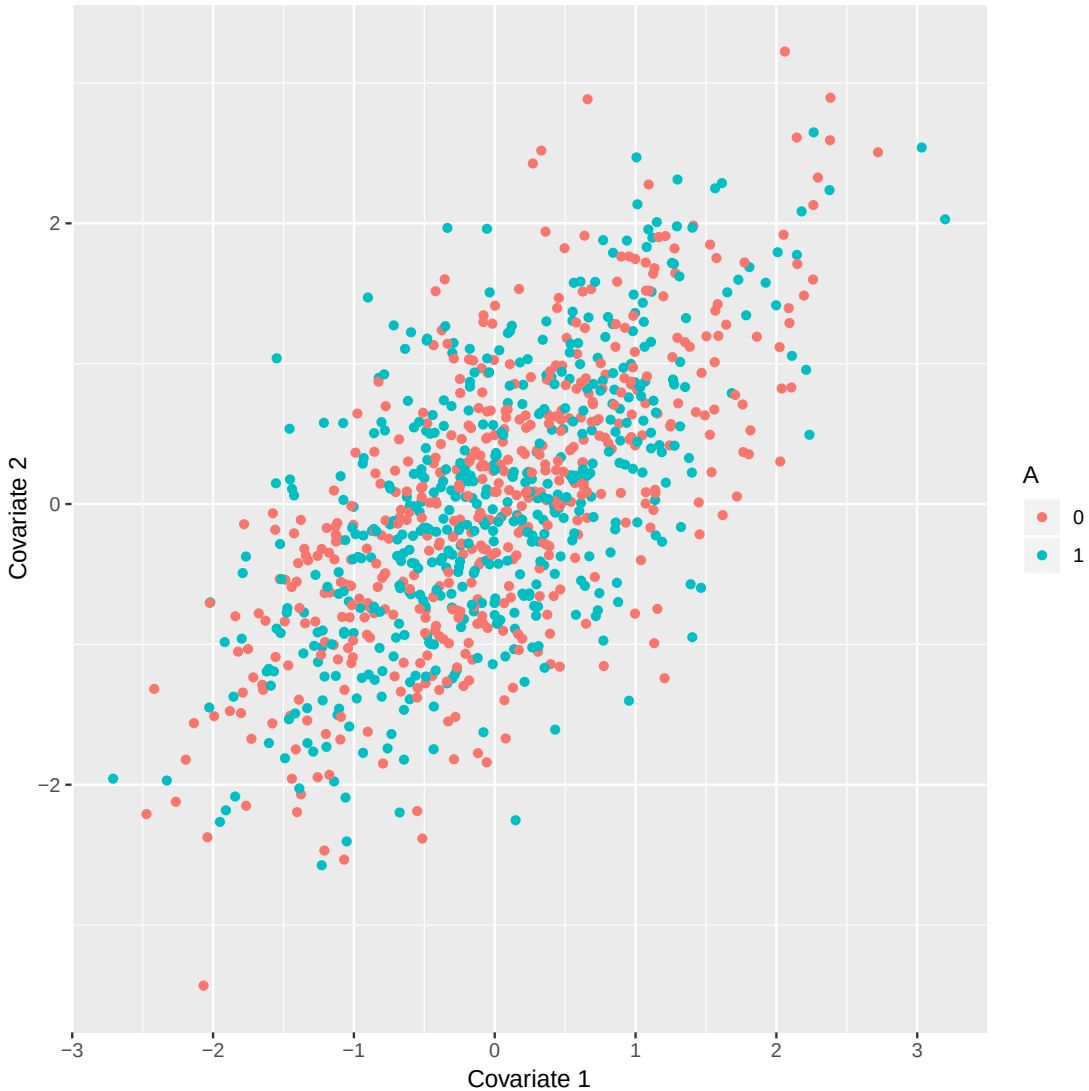
Scatter-plots

Visualizing relationship between two continuous variables.



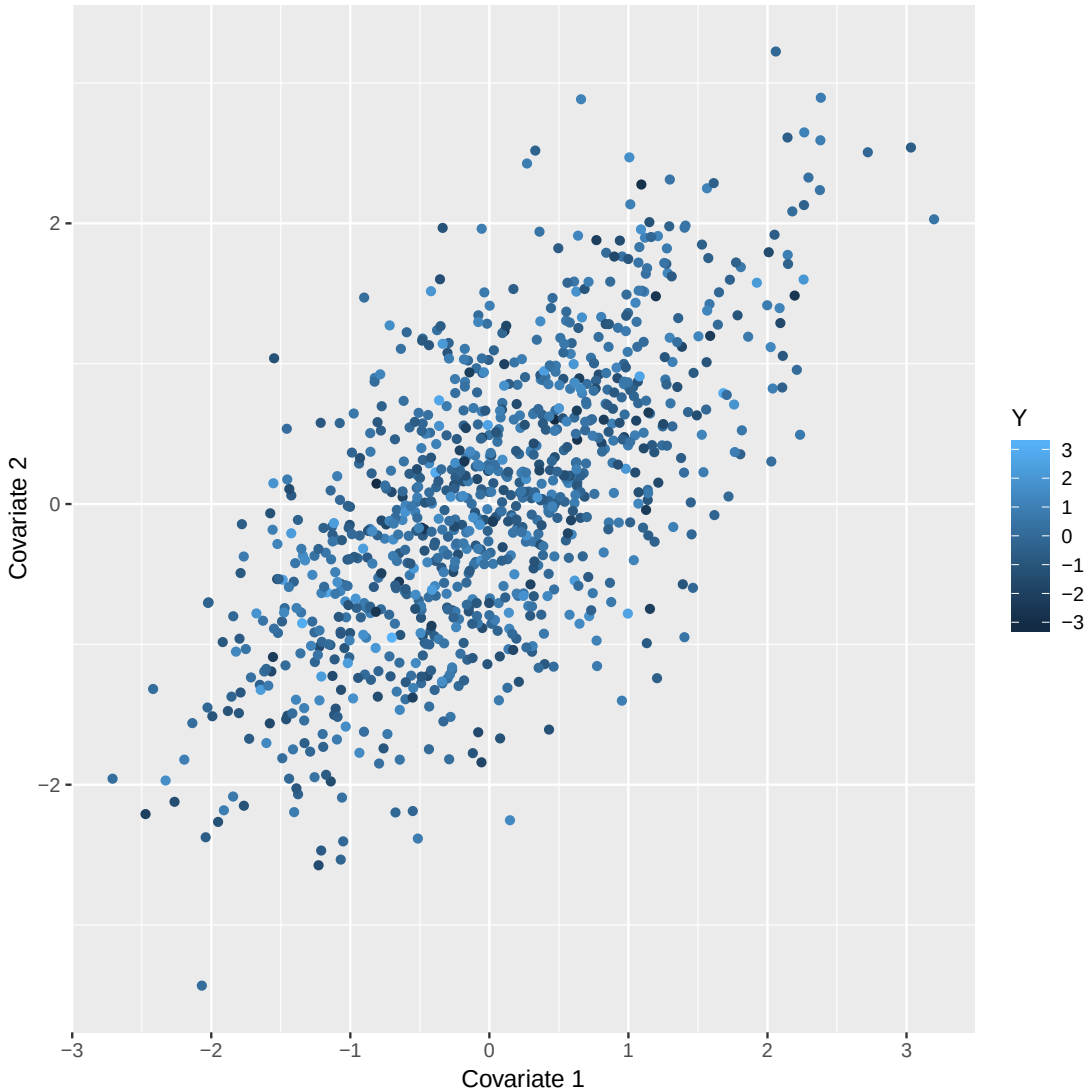
Scatter-plots by a third variable

Visualizing relationship between two continuous variables and a categorical variable.



Can even be used for three continuous variables

Visualizing relationship between two continuous variables and a categorical variable.



Let's see some practical examples

Study of HIV prevention treatment

- Multiple studies evaluating the prevention effect of a drug.
- Most show promising results, but one showed no prevention effect. Adherence issues?
- Adherence measured by self reported adherence.

Self Reported Adherence

	Self Reported Adherence
Yes	896
No	104

Adherence Measured by Blood Concentration

	Detectable levels of drug concentration
Yes	274
No	726

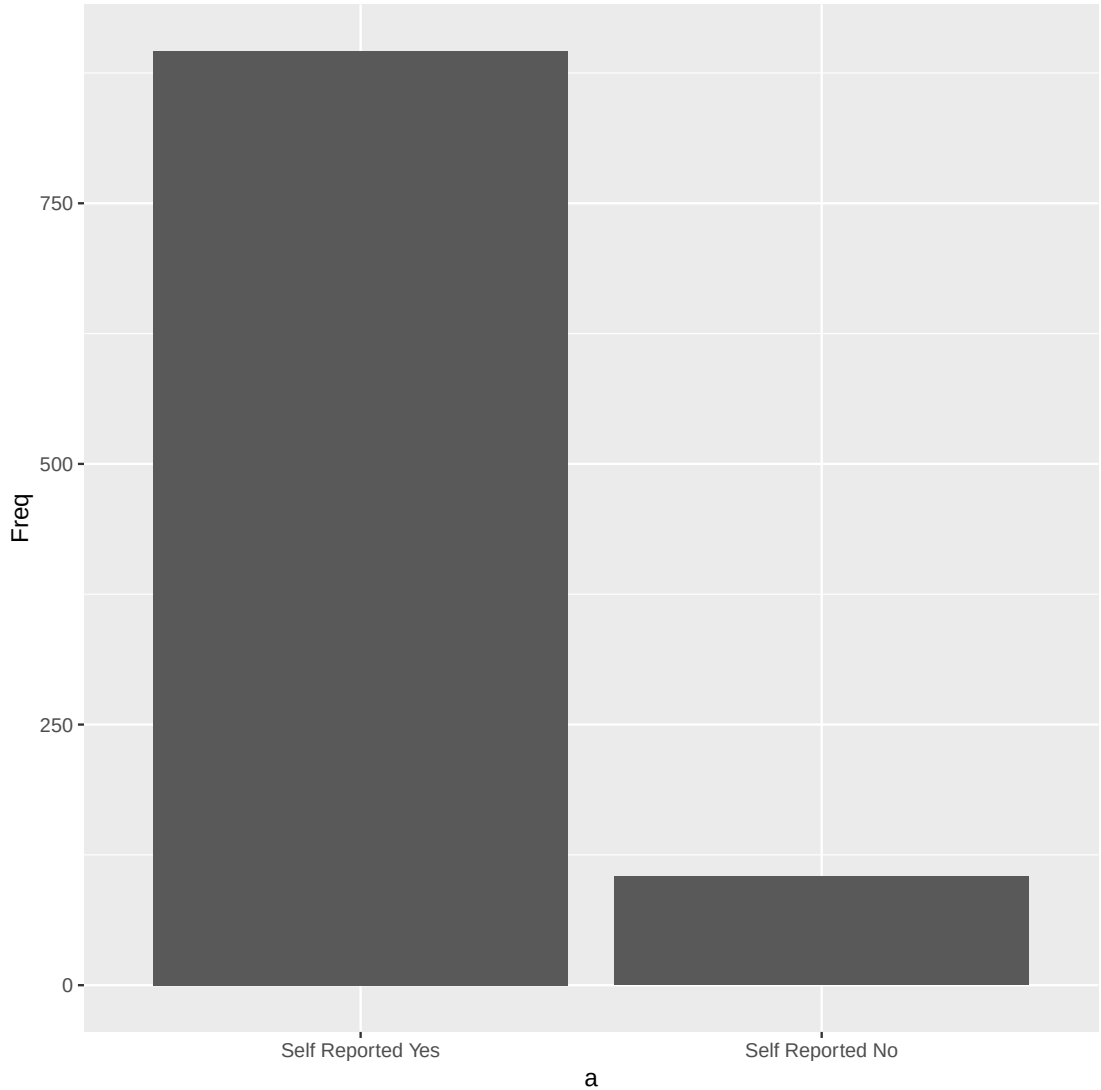
Summarizing relationship

	Blood Conc No	Blood Conc Yes
Self Reported Yes	625	271
Self Reported No	101	3

Could also have looked at proportions instead of counts.

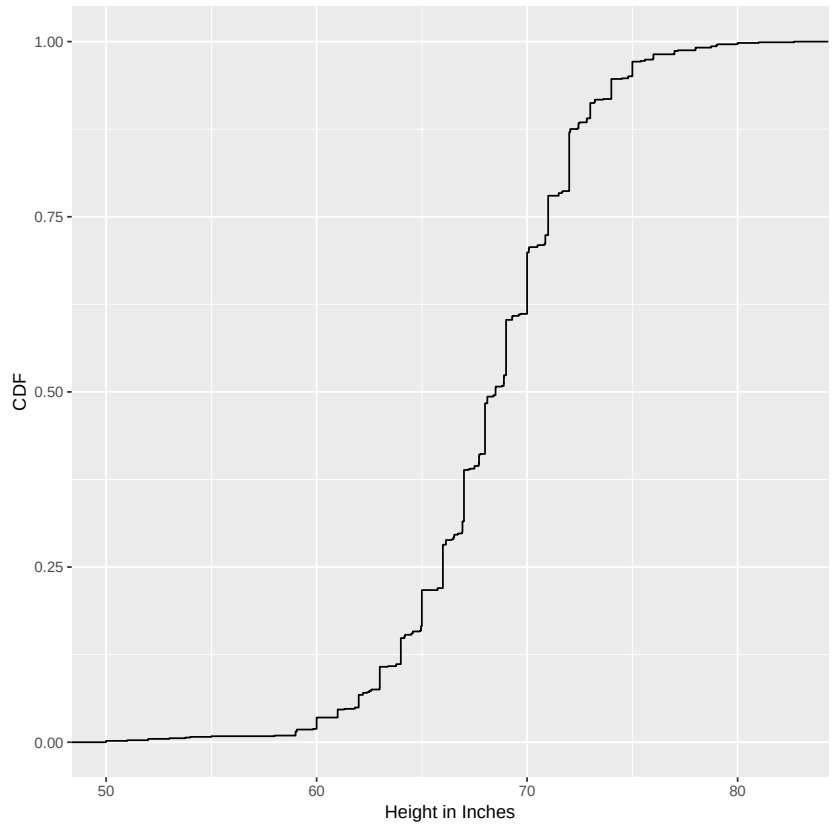
Barplots

Could also have presented graphically using barplots.



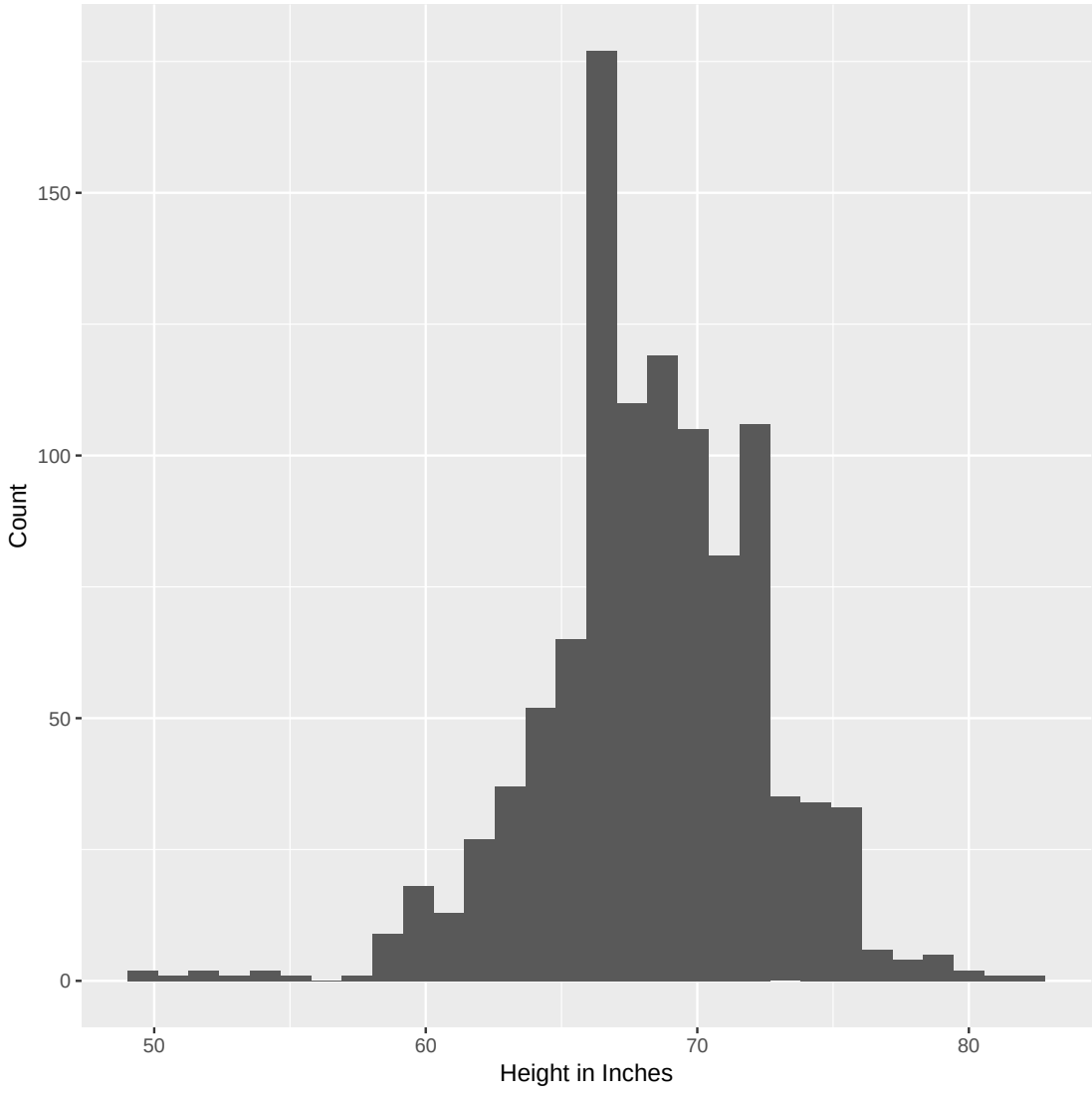
Distribution Function

True cumulative distribution function (CDF) $F(x) = P(X \leq x)$.
Empirical CDF $\hat{F}(x) = \hat{P}(X \leq x)$ is an estimator of that.

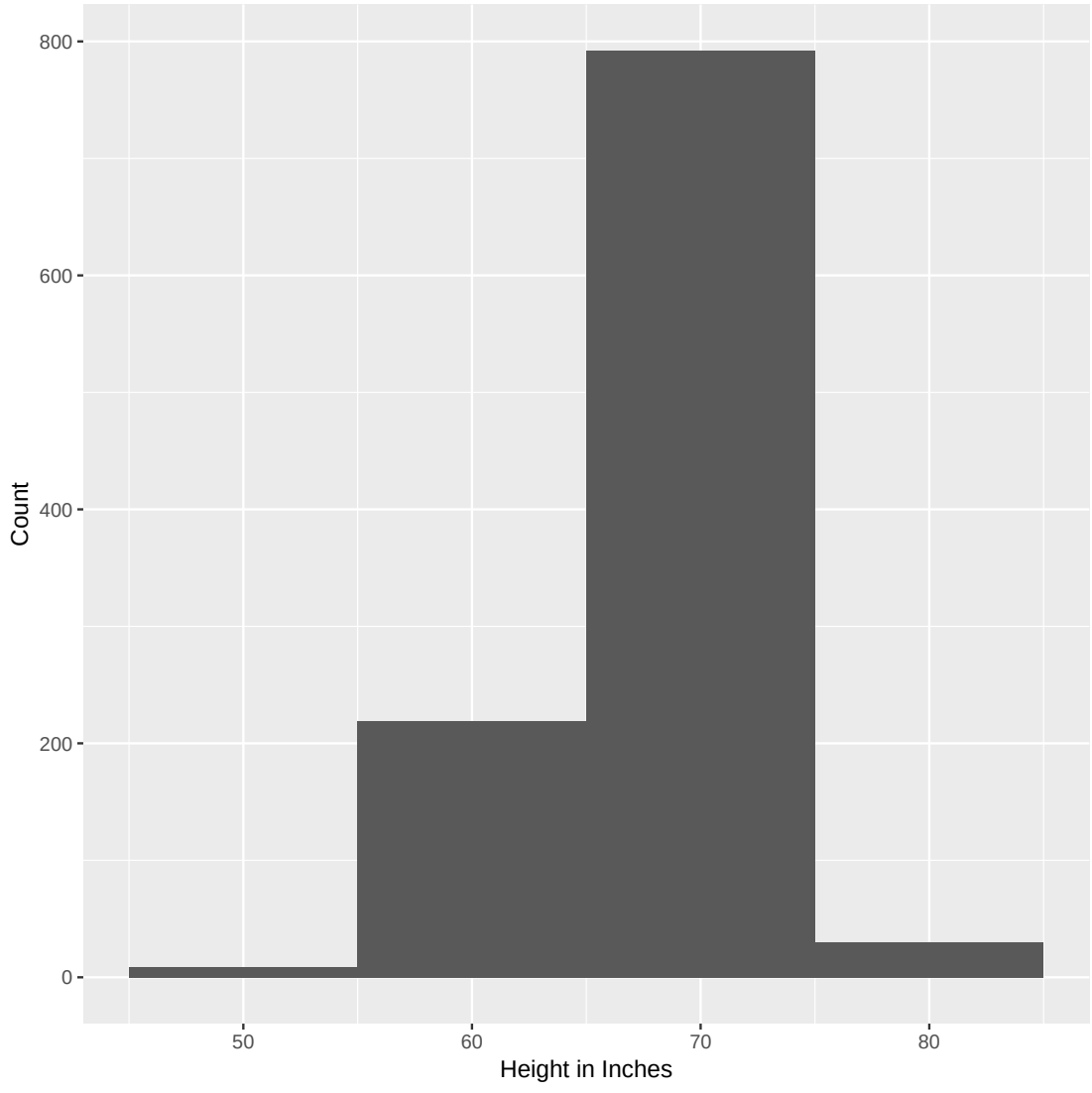


Histograms provide more useful information (e.g. where is distribution centered, is it symmetric)

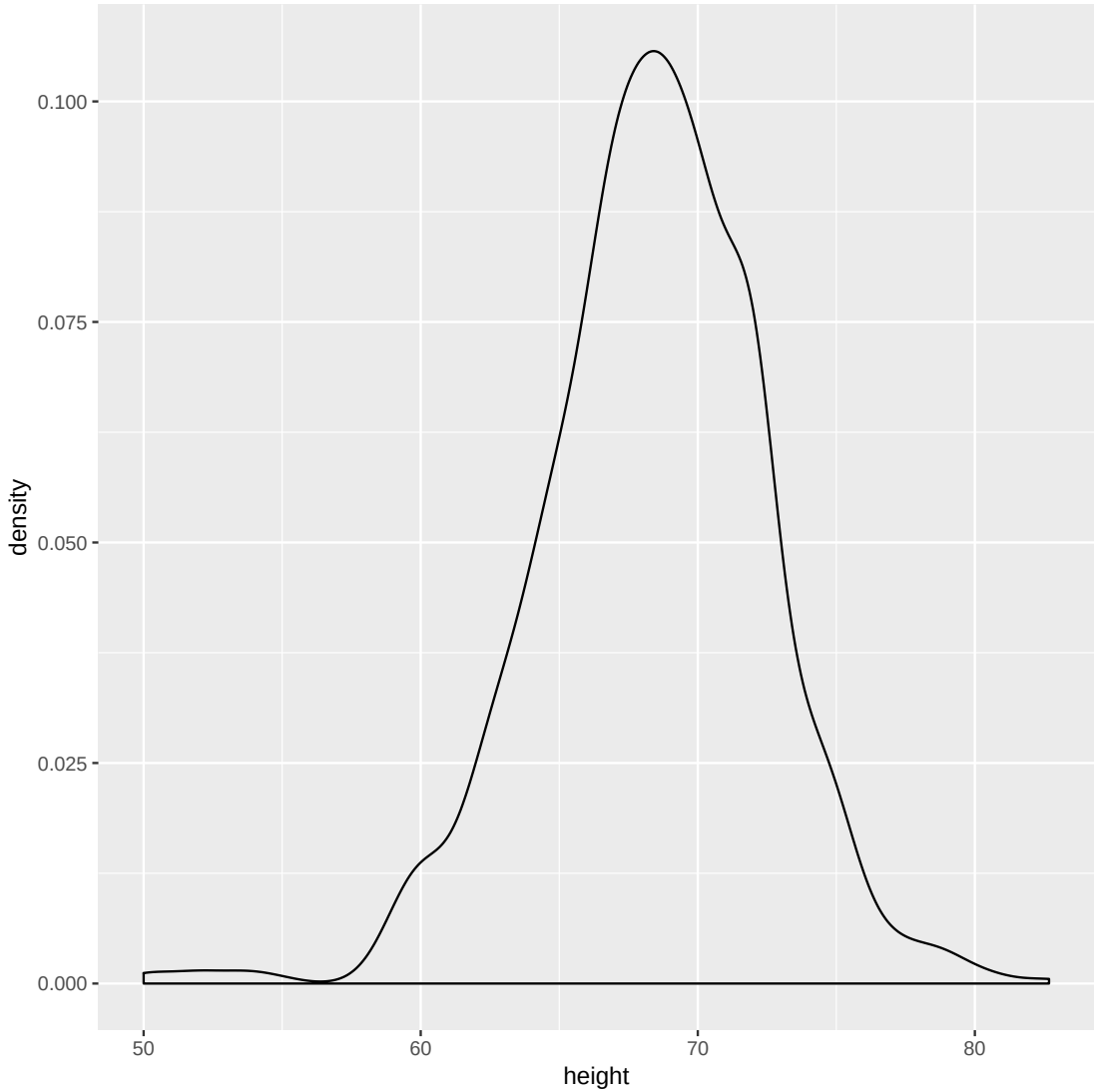
Histogram



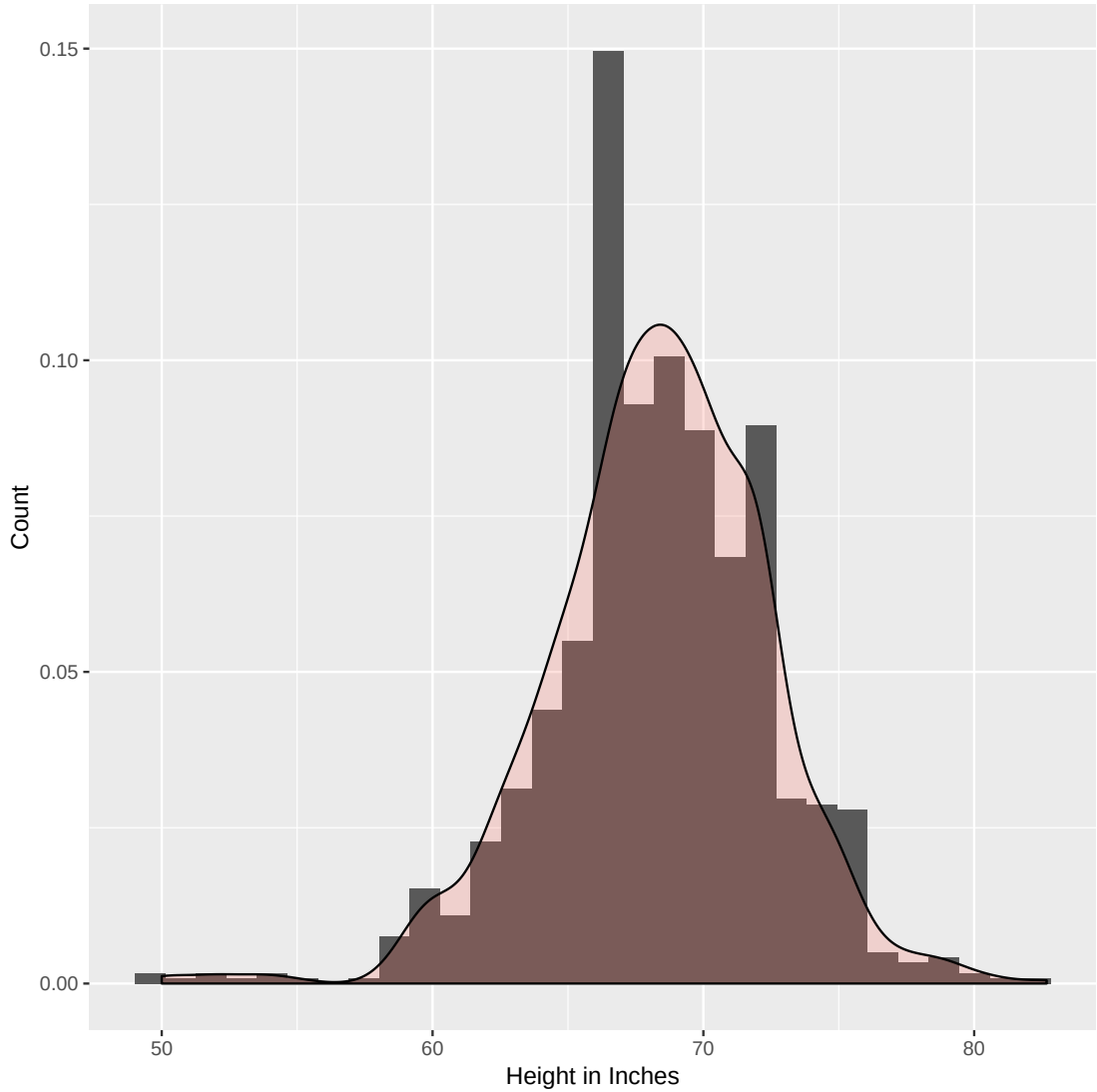
Number of Bins can Matter



Density Estimators

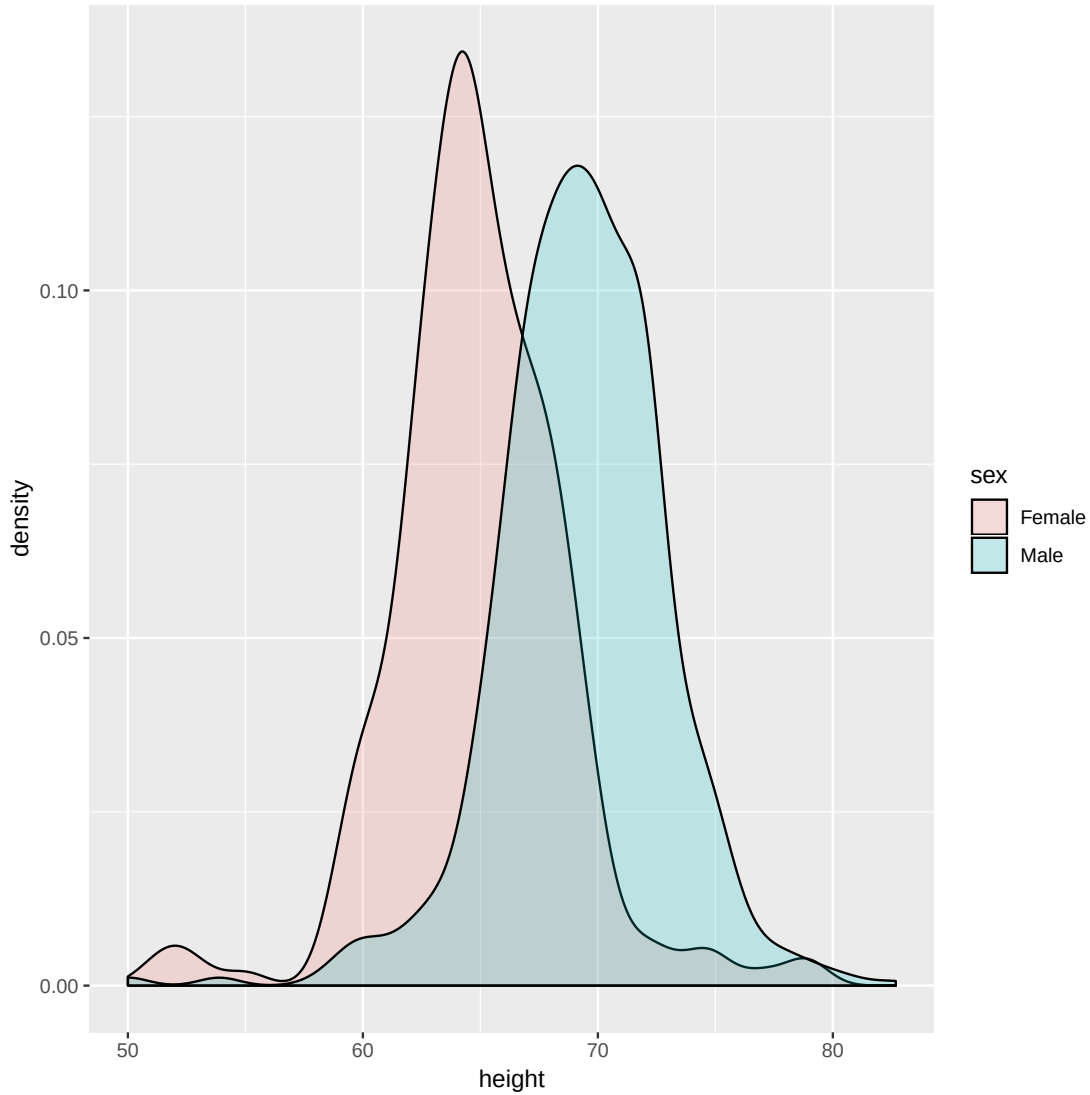


Overlaying Density and Histogram

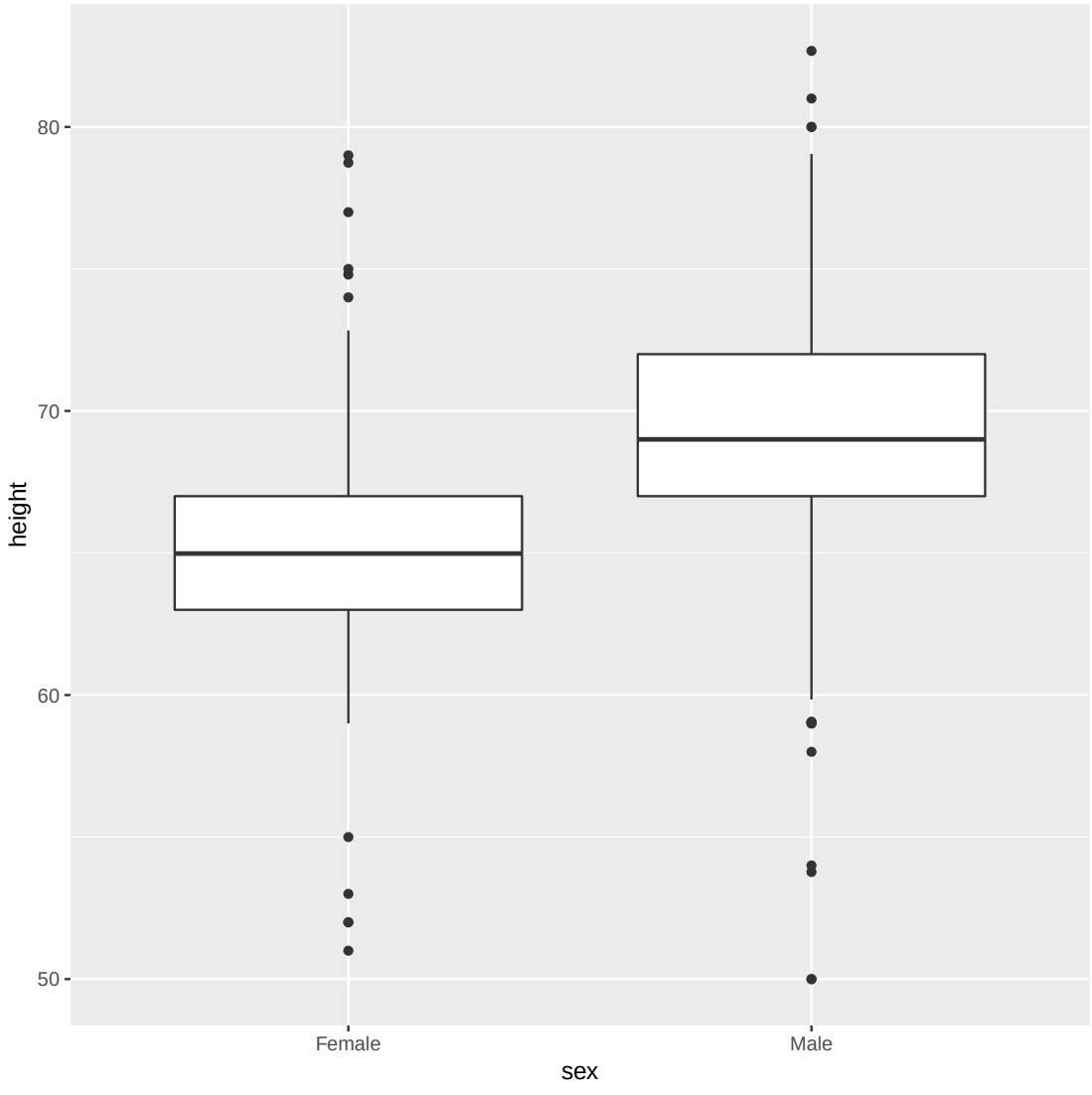


Height densities by gender

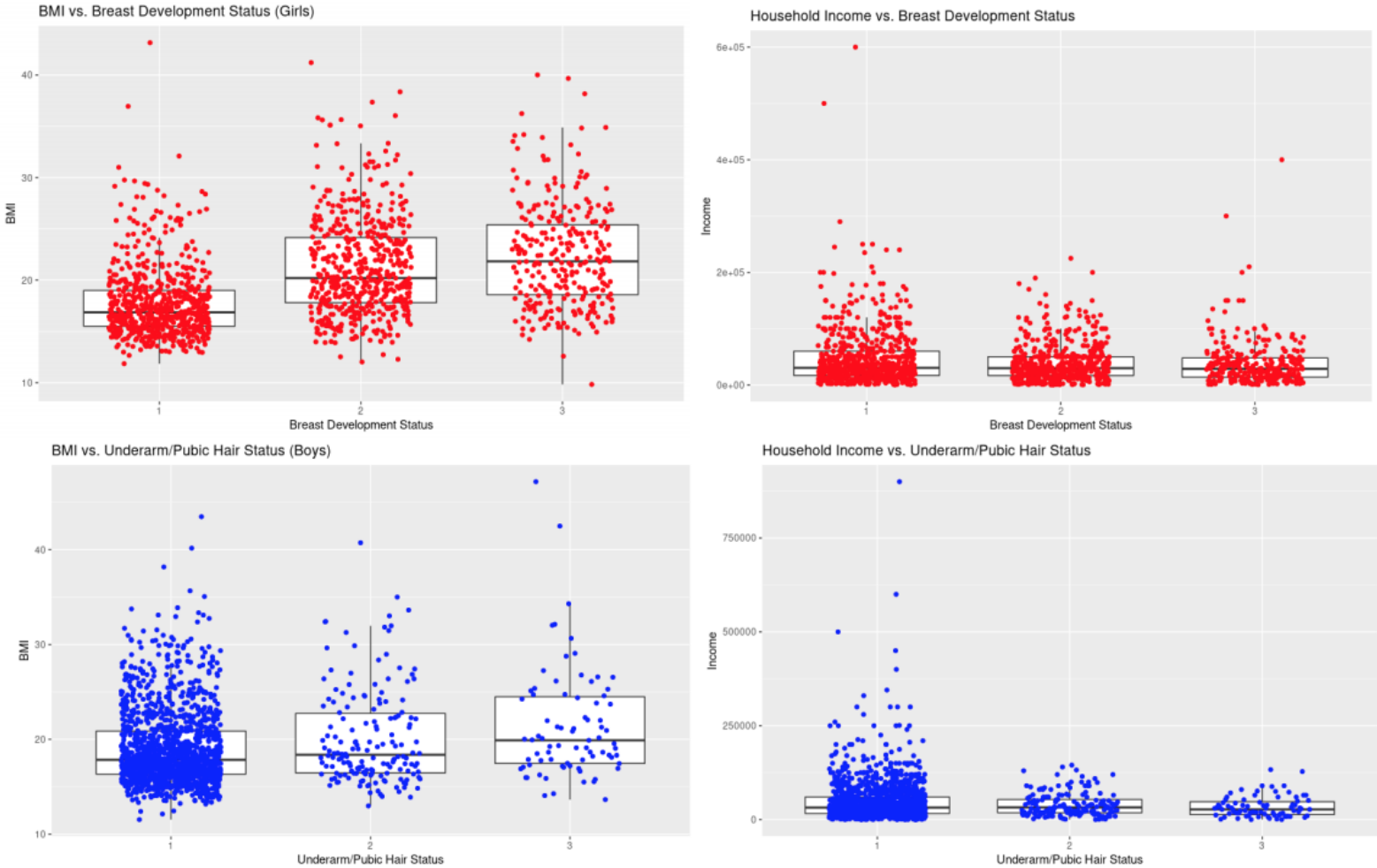
Example of how to visualize a relationship between a continuous and a categorical variable.



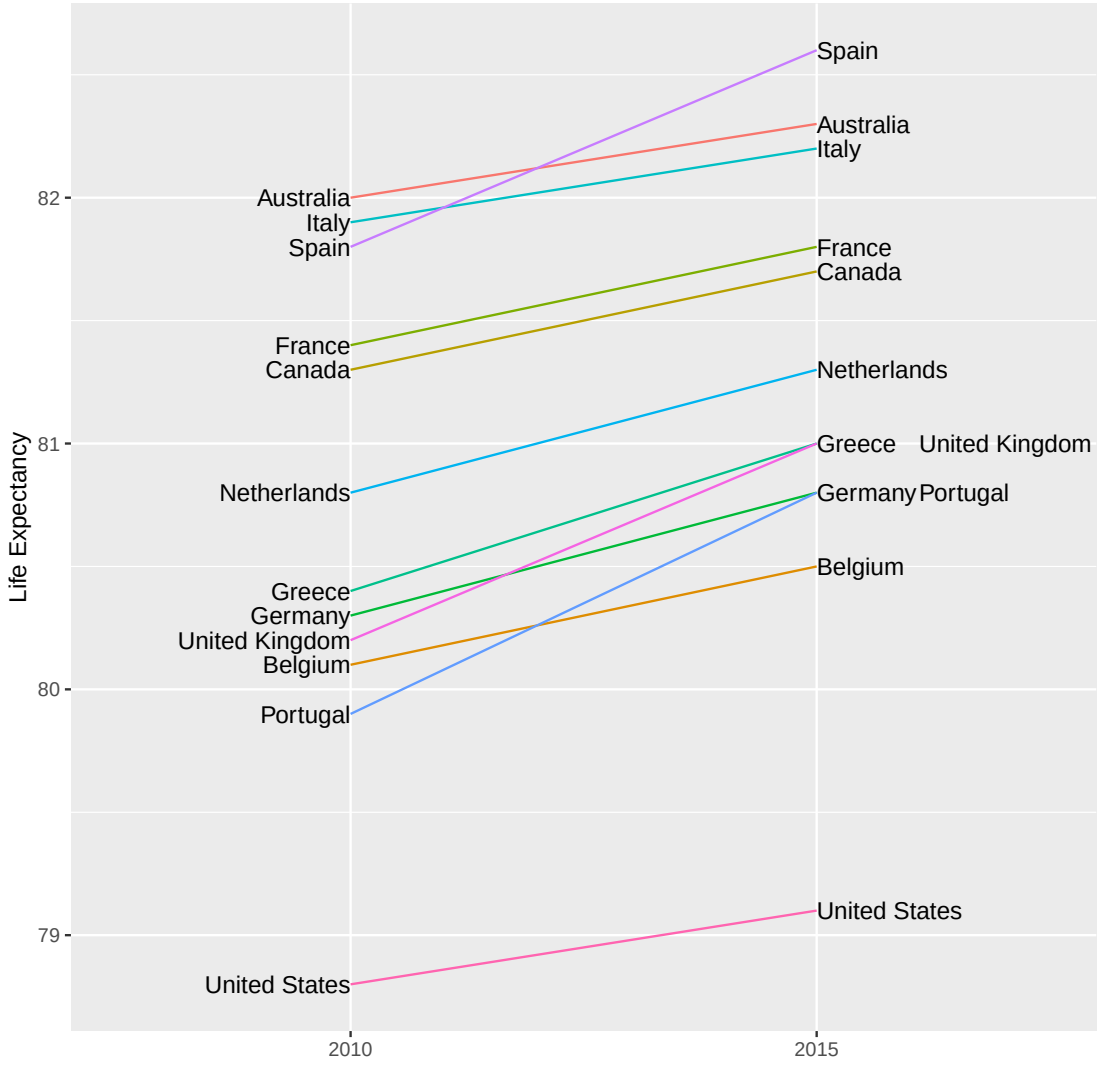
Boxplot



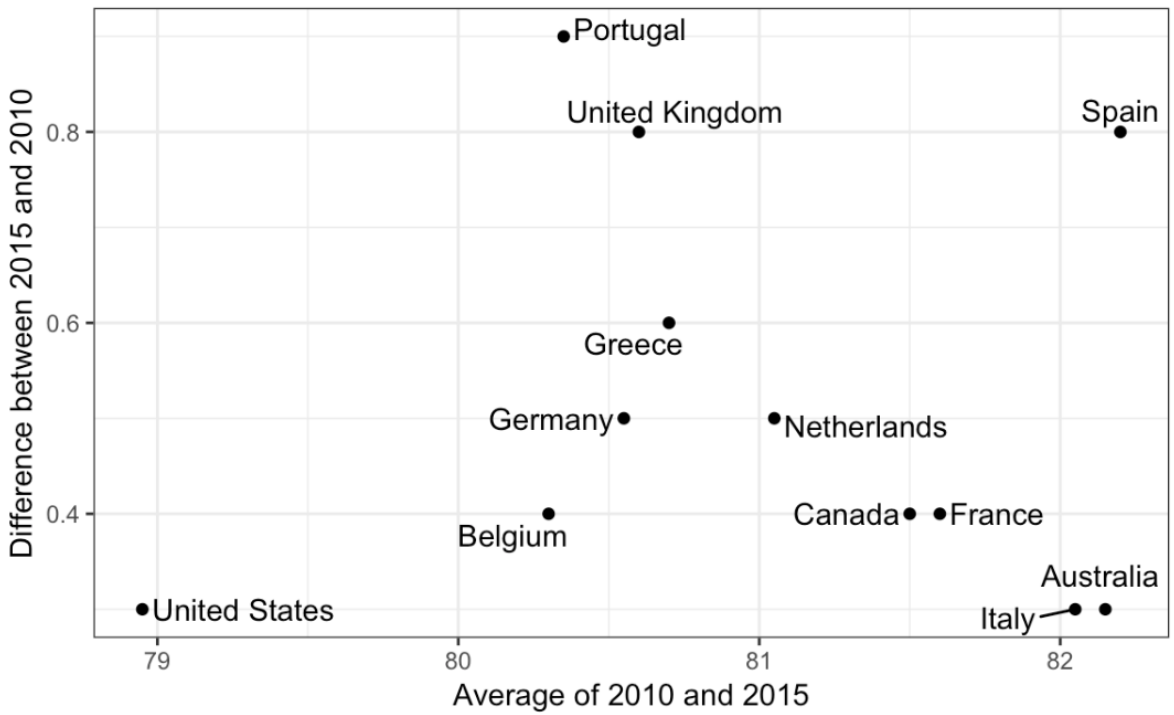
Boxplot with early puberty



Some other plots: Slope Charts

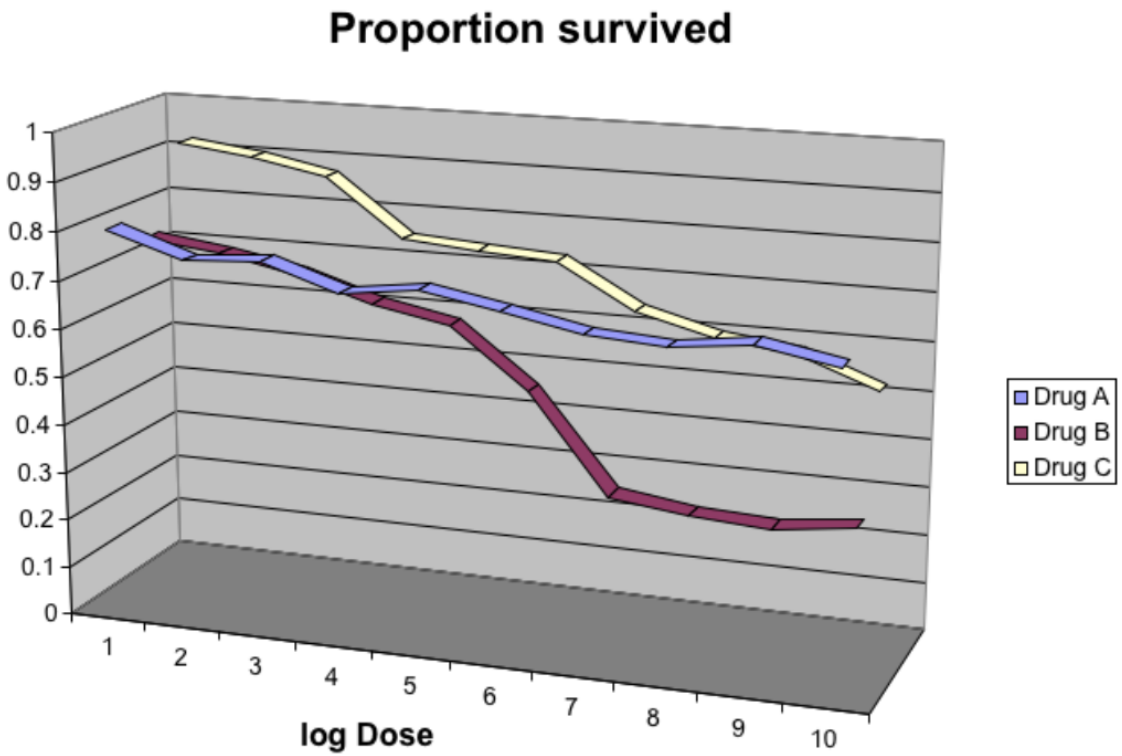


Some other plots: Bland-Altman



3D plots

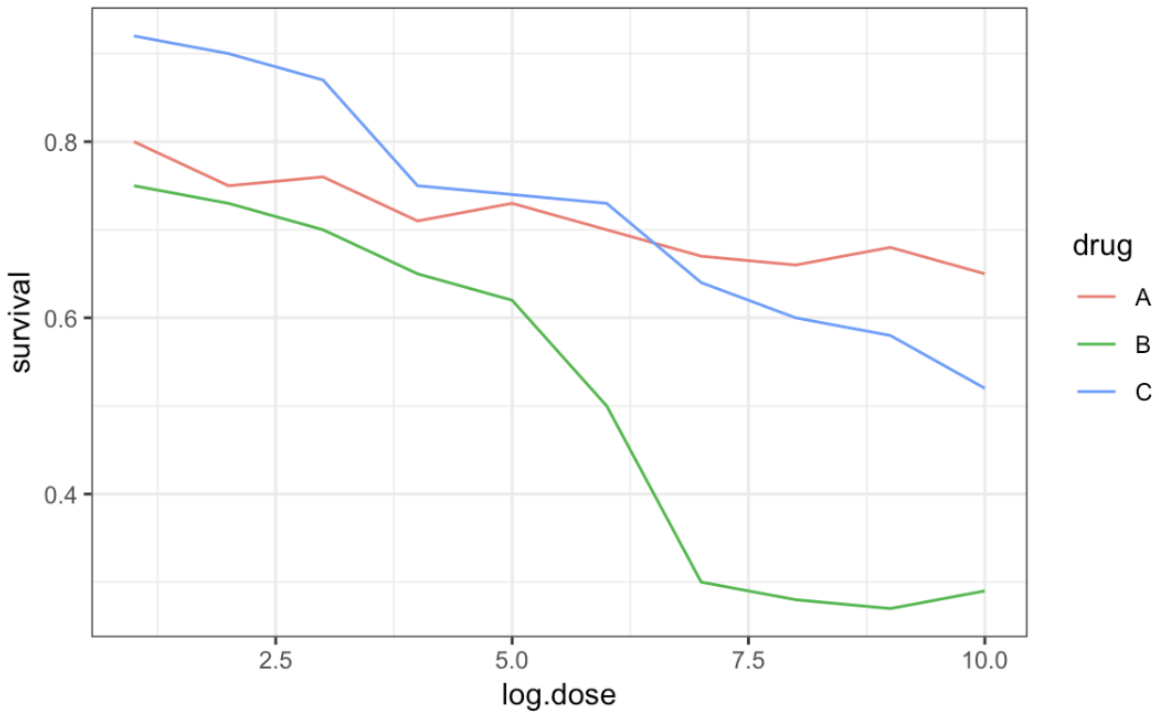
Almost always a bad idea and often deceiving.



Source: IDS, via biostat.wisc.edu/ kbroman/presentations/graphs2017.pdf

3D plots

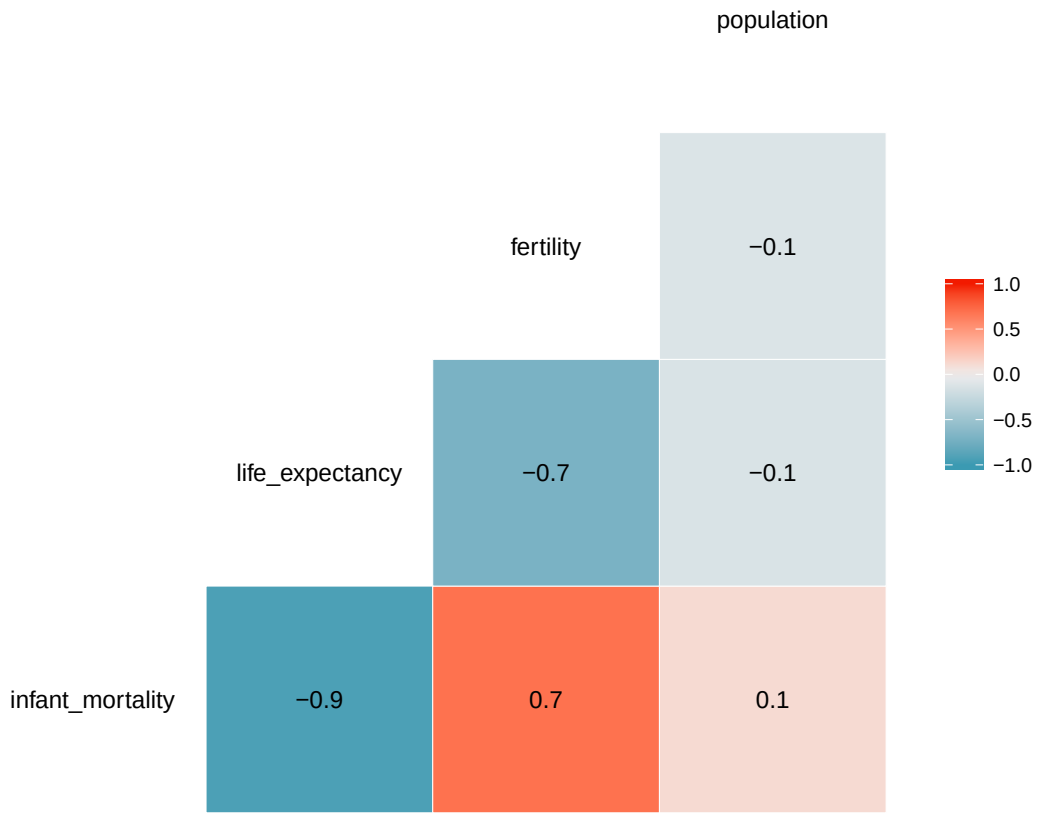
Much better



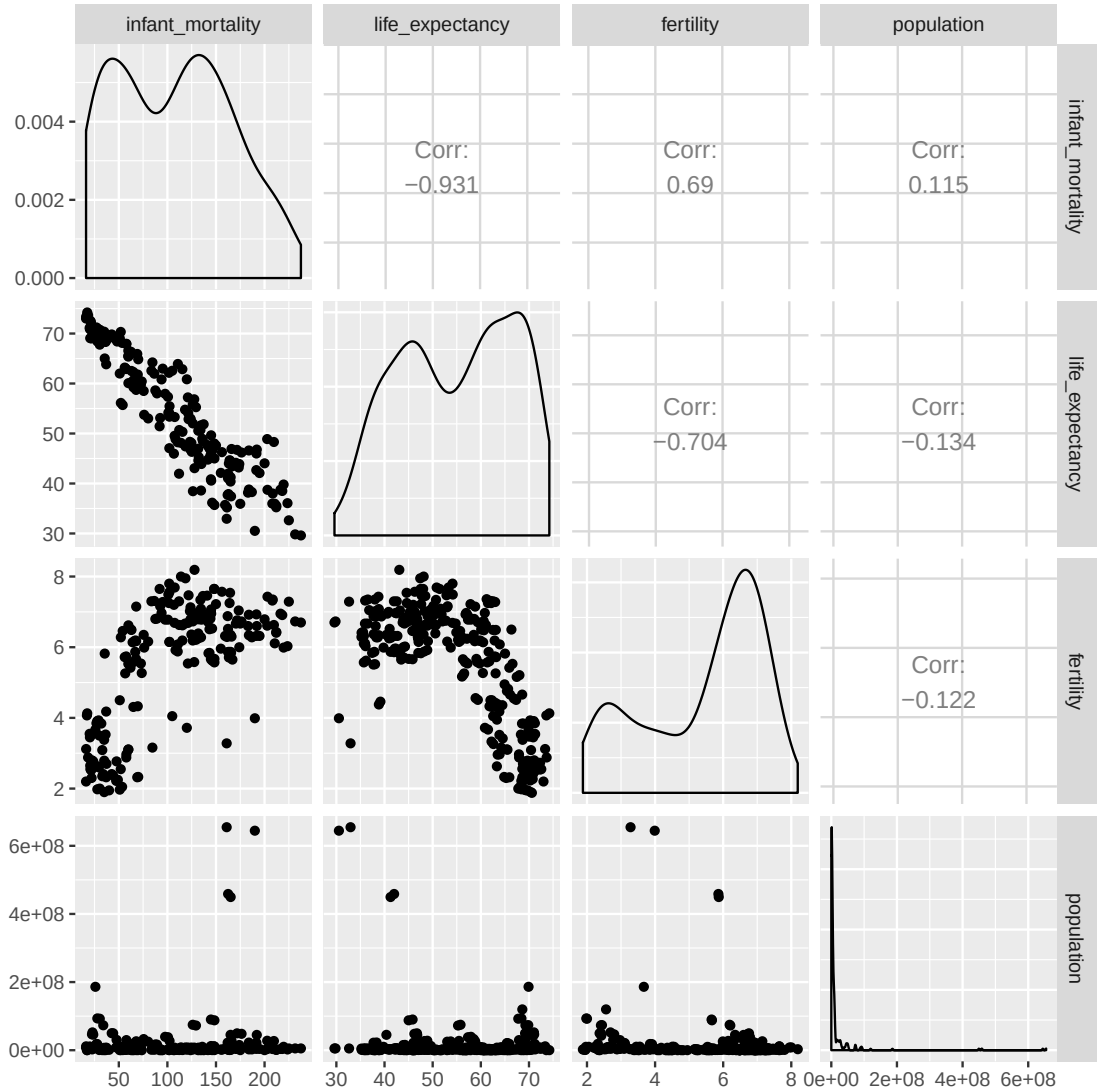
Source: IDS

Exploring Correlations

High correlations can cause problems in linear regression (might only need height but weight).



Exploring Correlations



Summary

Some methods of summarizing information:

- Categorical variable: Tables, barplots.
- Continuous variables: Summary statistics, histograms, density plots, CDFs, boxplots.
- Two categorical variables: Tables, barplot by combinations of levels.
- Two continuous: Scatterplots
- Categorical and continuous: Summary statistics, histograms, density plots, CDFs, boxplots by levels of the categorical variables.
- Two continuous and a categorical: Scatterplot with points differentiated by level of categorical variable (e.g. using different colors or shapes).

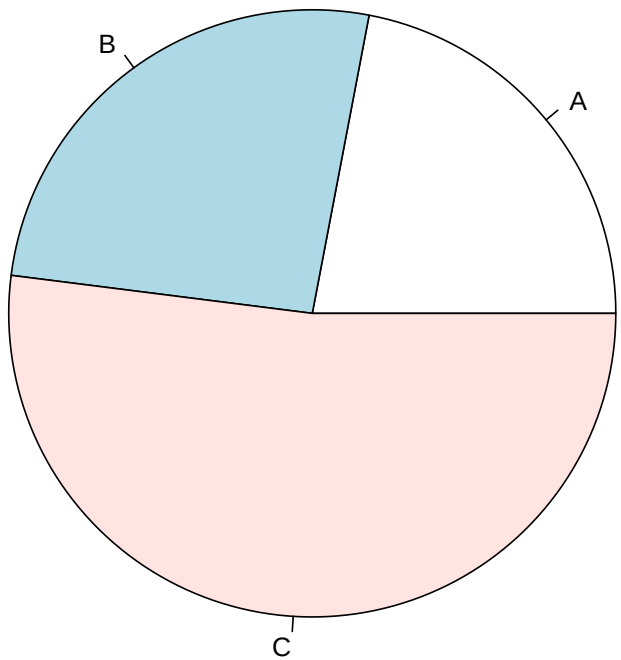
There are multiple other ways of summarizing relationships between variables and search engines are an extremely useful way to find information.

Data visualization principles

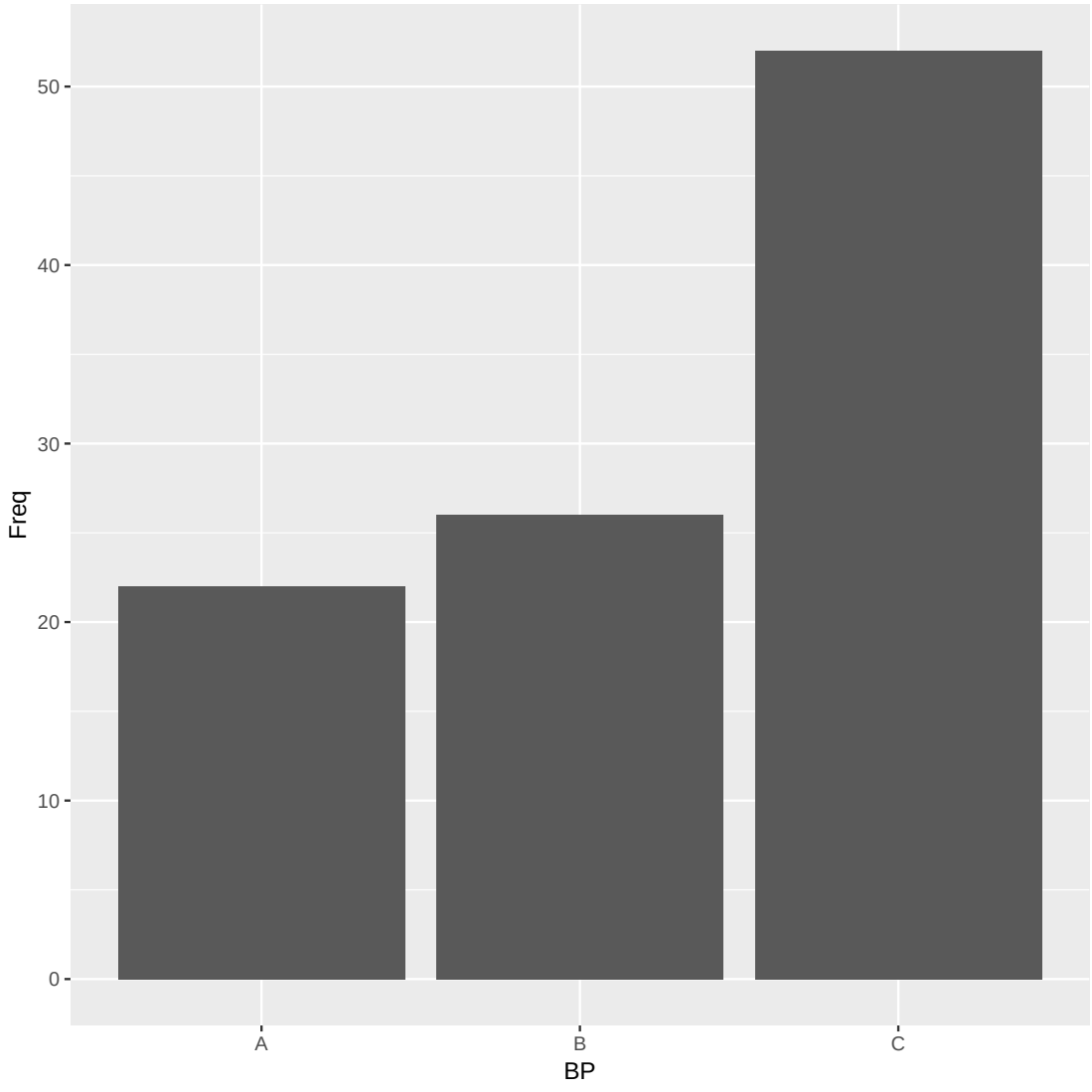
- Position and length are easier to quantify than angles, area, color, or brightness.
- Think about if including zero makes sense.
- Be careful when comparing quantities using circles (area vs radius).
- Order by a meaningful value.
- Make the comparisons of interest as easy as possible to see.
- Avoid too many digits.
- Know your audience.

Pie-chart

Hard to distinguish if A or B is more frequent.



Same Information as a Barplot

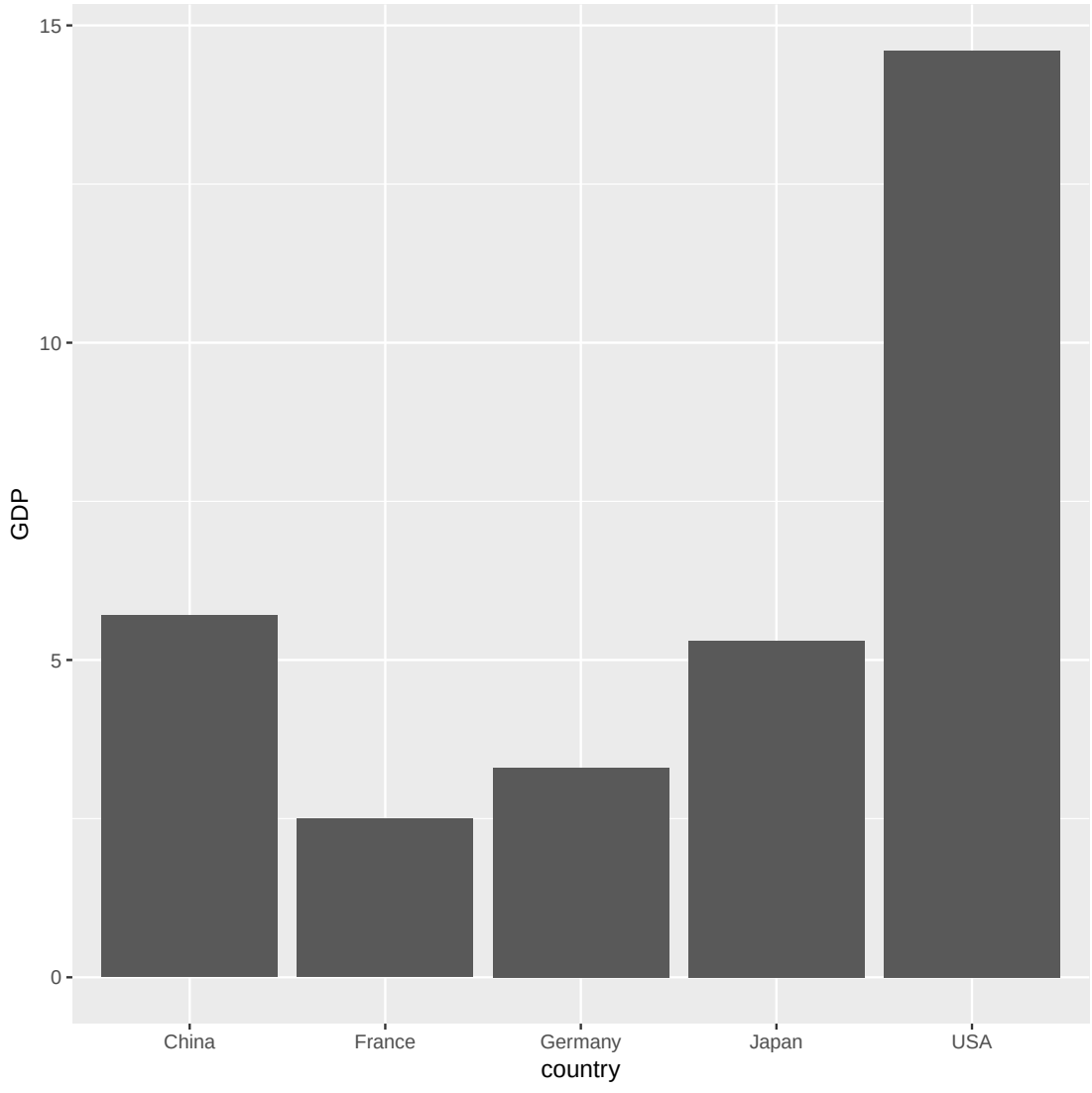


Comparing GDP using circles

Looks like the USA GDP is around 30 times larger than France's GPD.

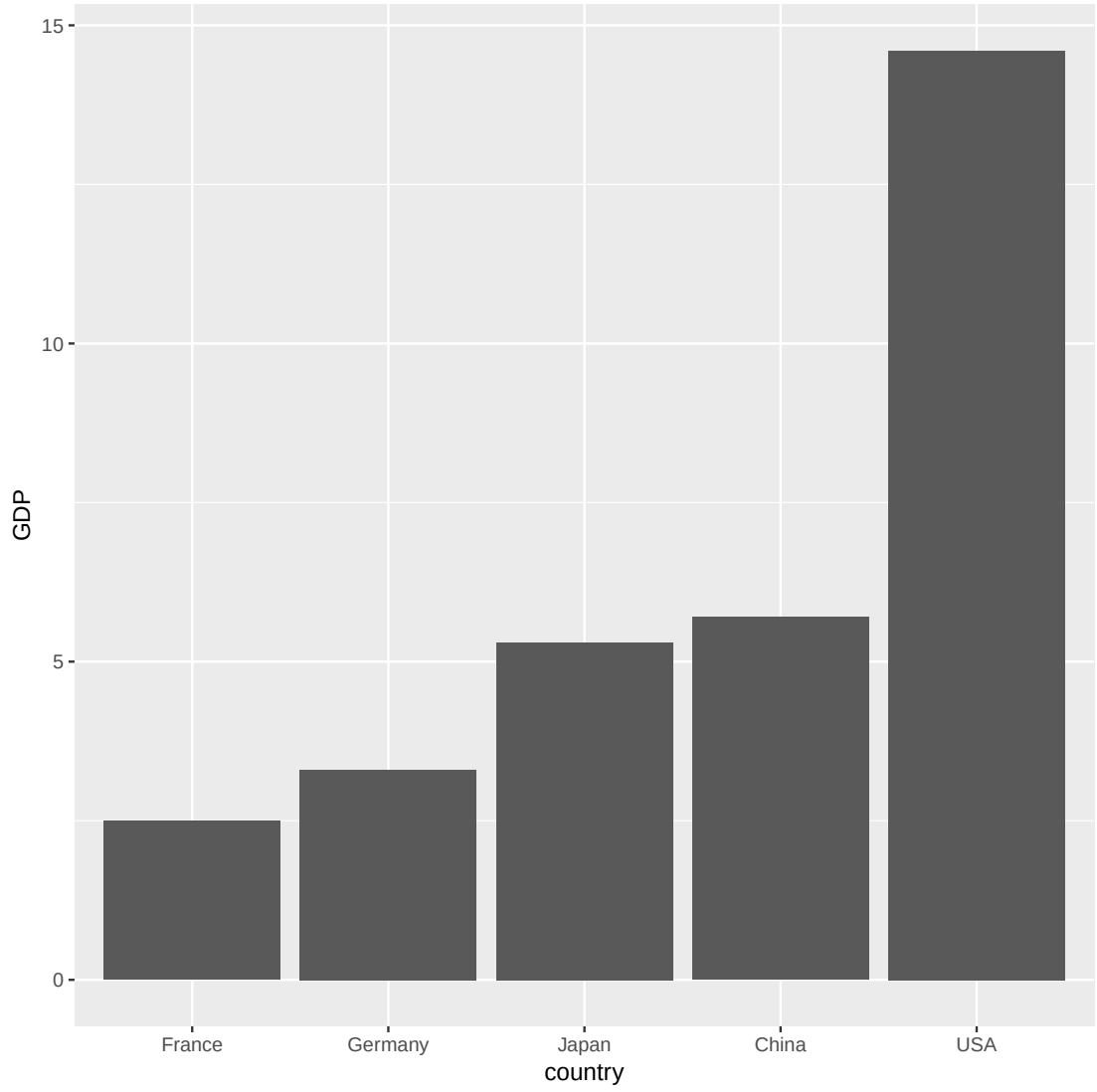
Source: The 2011 State of the Union Address: Enhanced Version, via IDS

Same data using a barplot

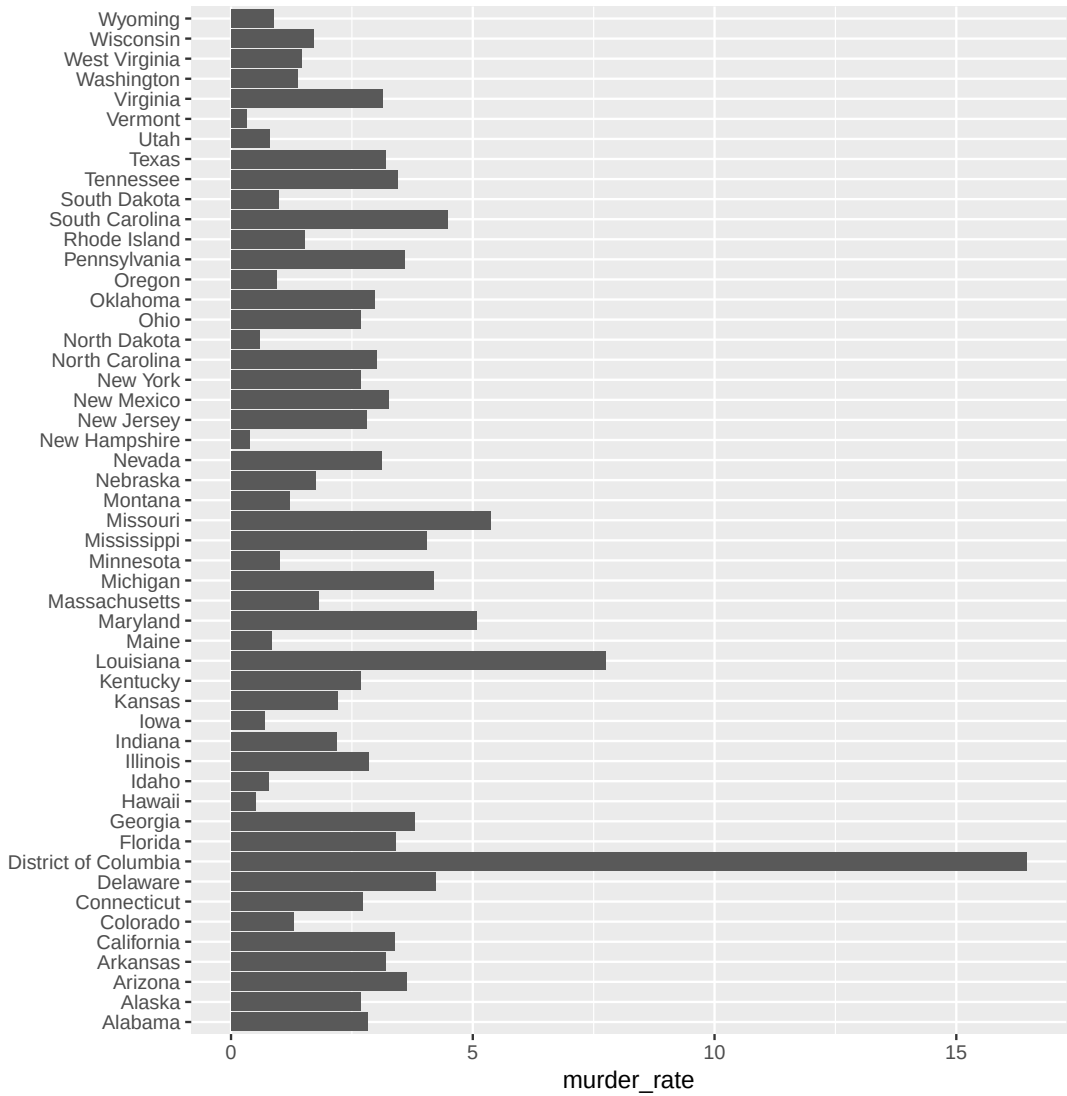


Order by meaningful value

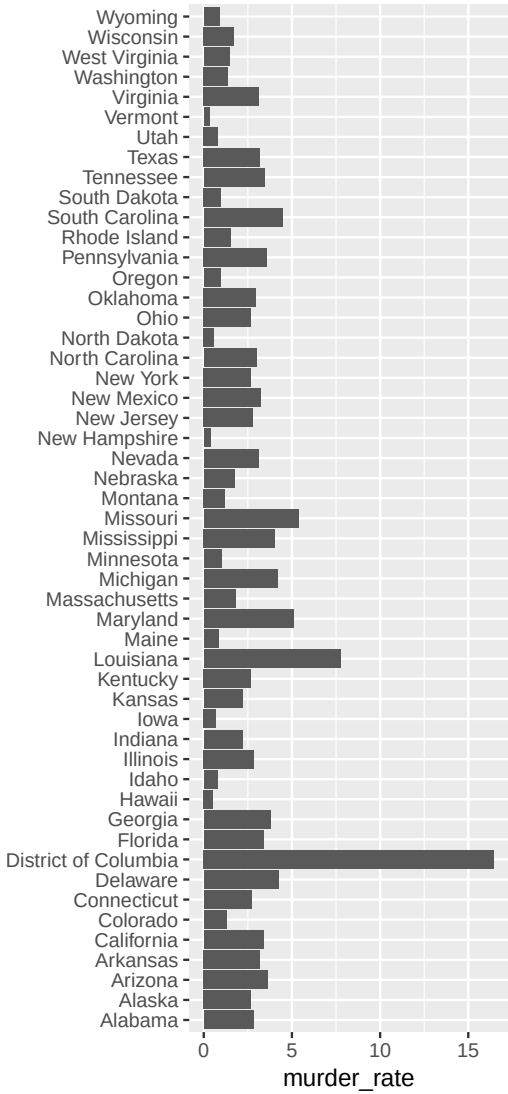
Ideally the previous plot should be ordered by size of GDP.

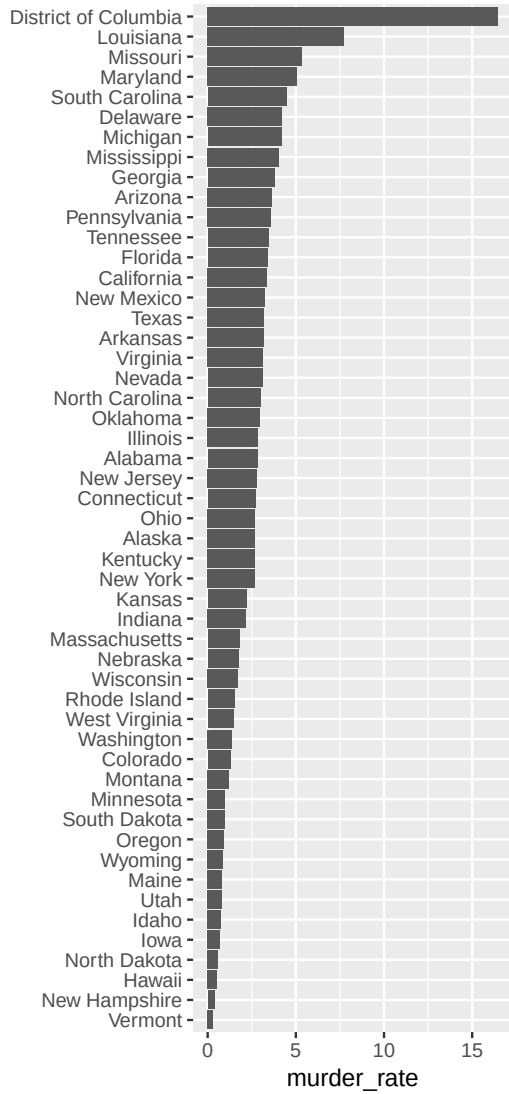


Order by meaningful value (Taken from IDS)

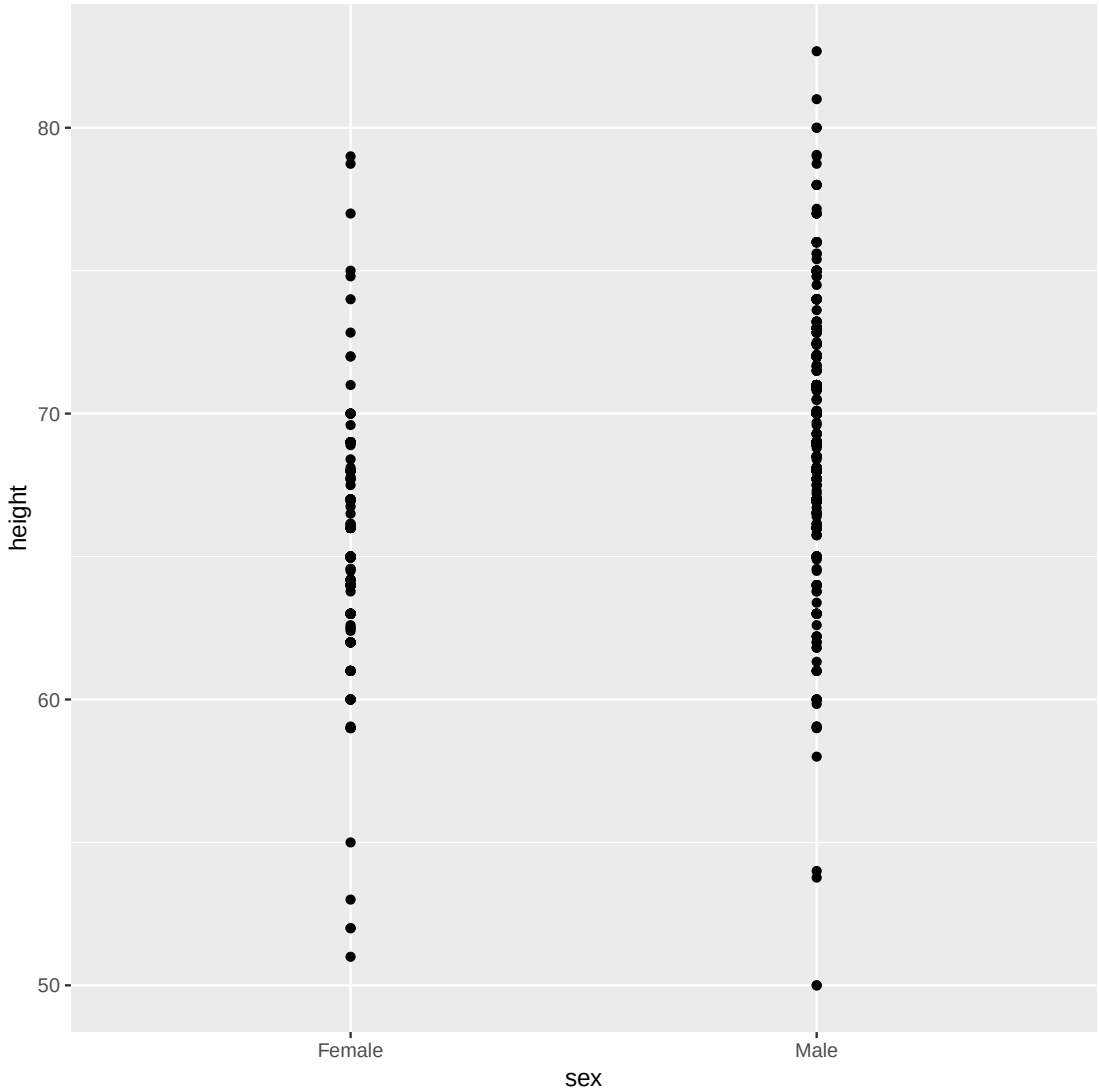


Order by meaningful value (Taken from IDS)



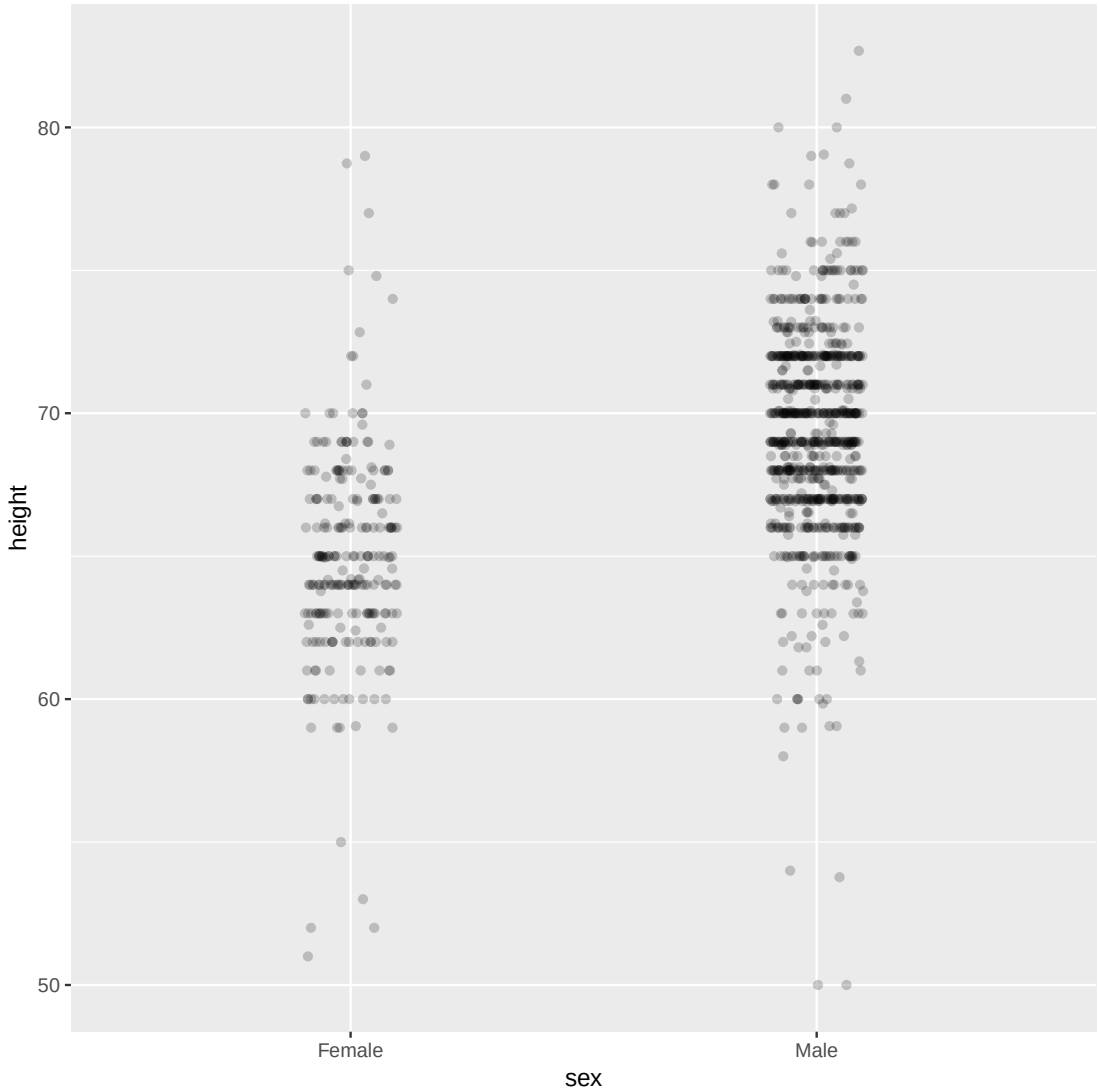


Show the data



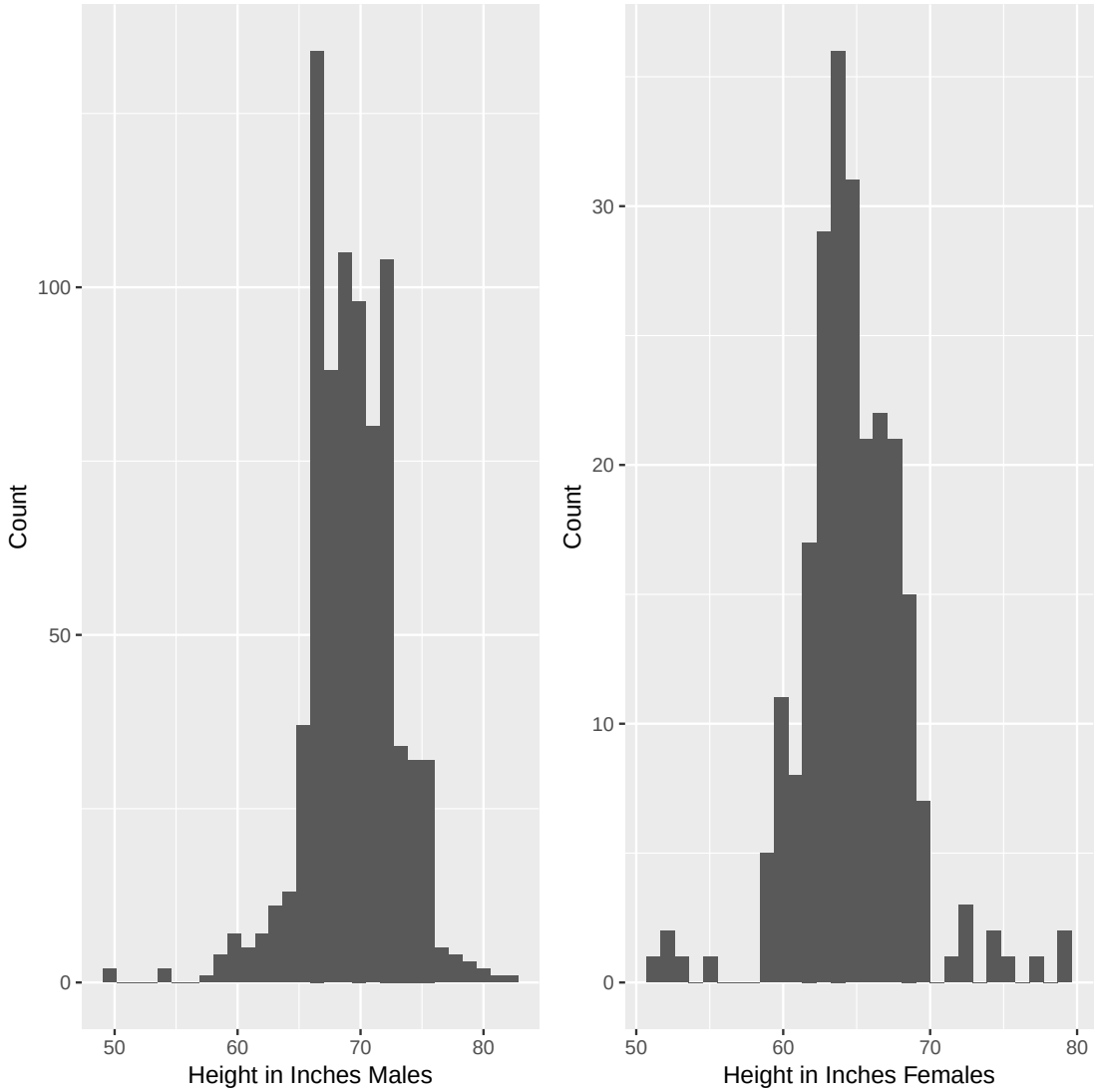
Show the data

Hard to see how many males have height of 65 inches. Some solutions are: 1) Jitter adds a small random shift to each point; 2) alpha blending makes the color stronger the more points there are.



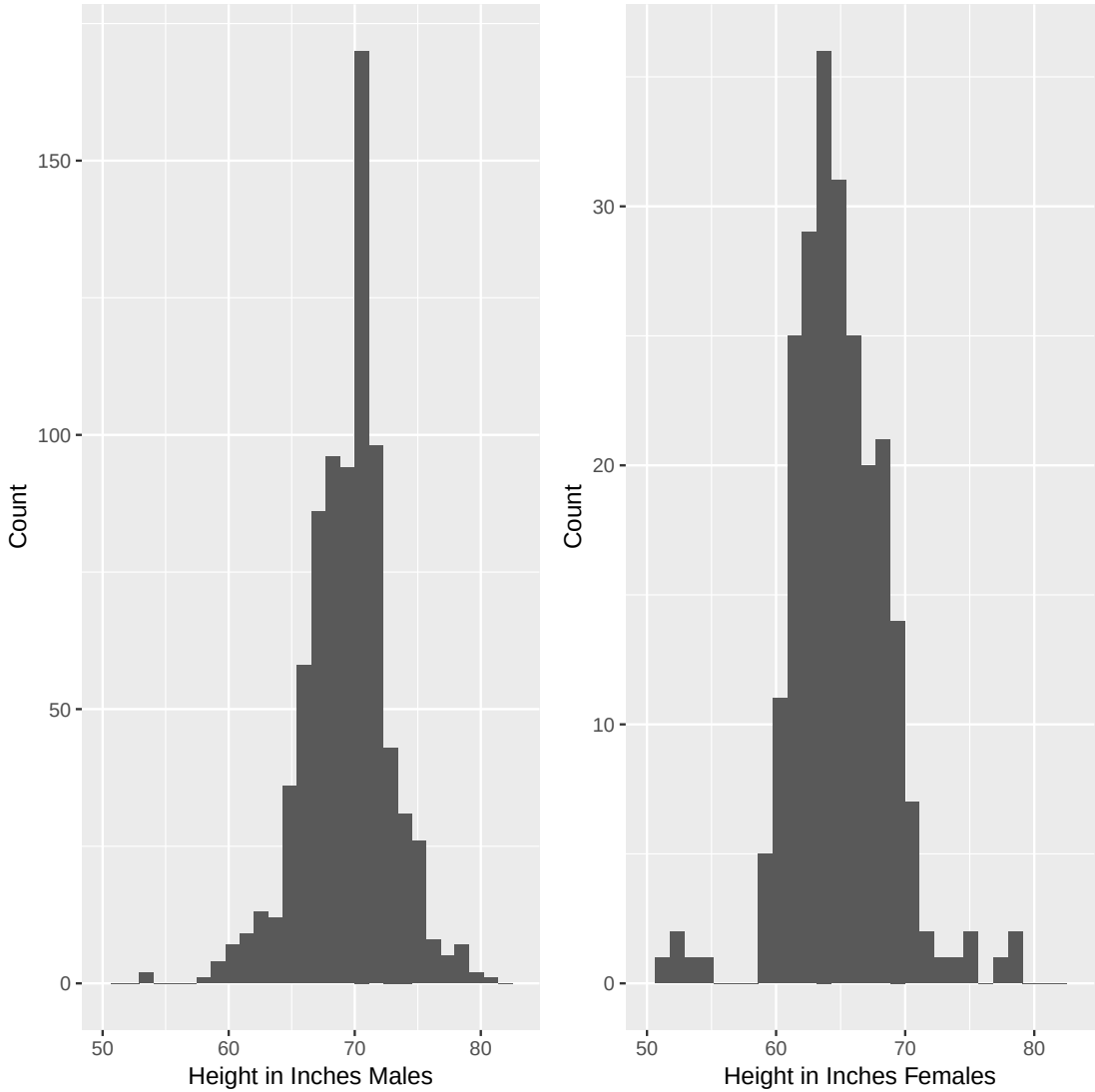
Common Axis

Need to think to see that males are on average larger than females.



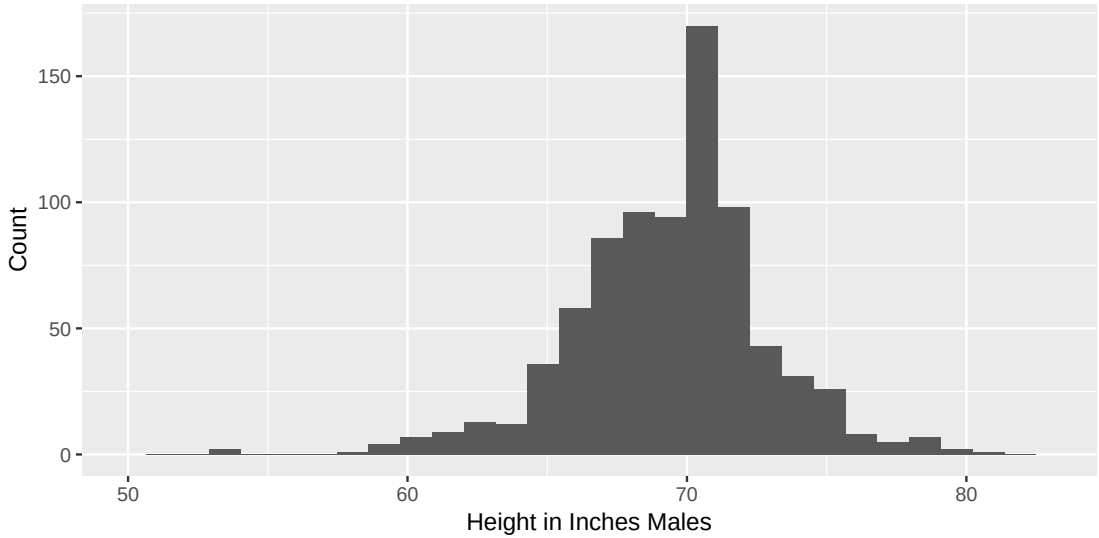
Common Axis

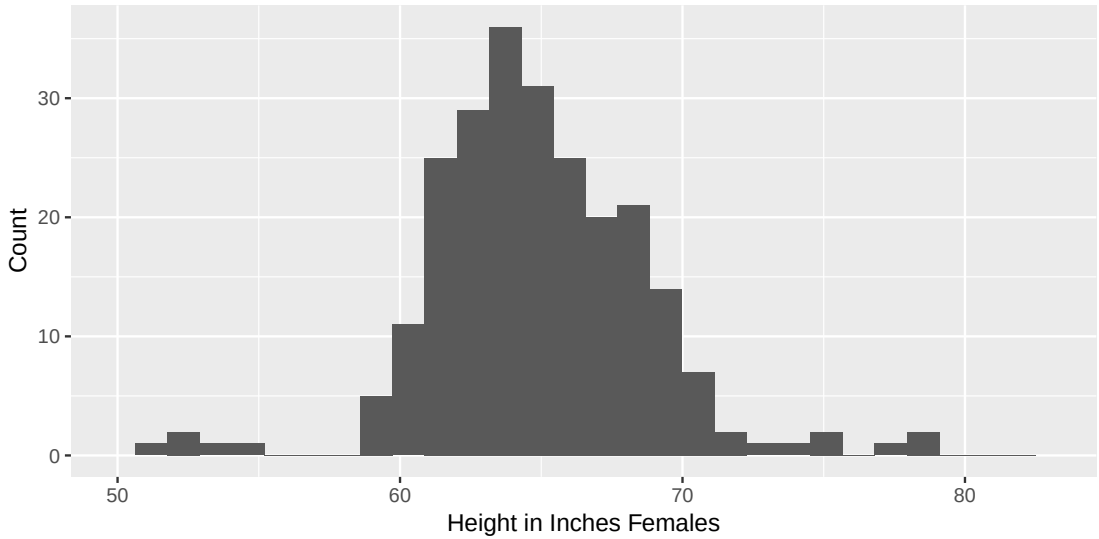
Better



Common Axis

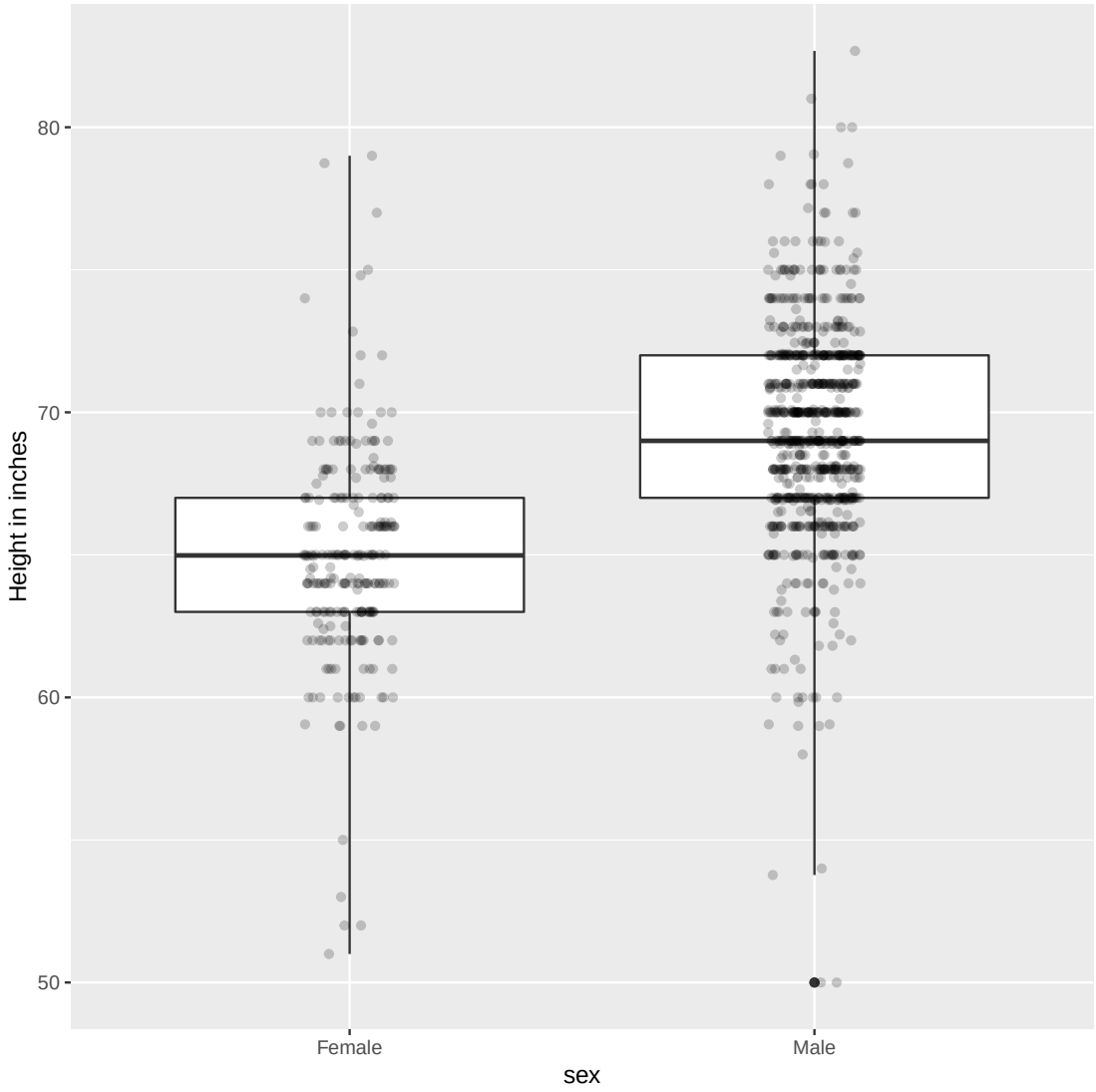
Even Better



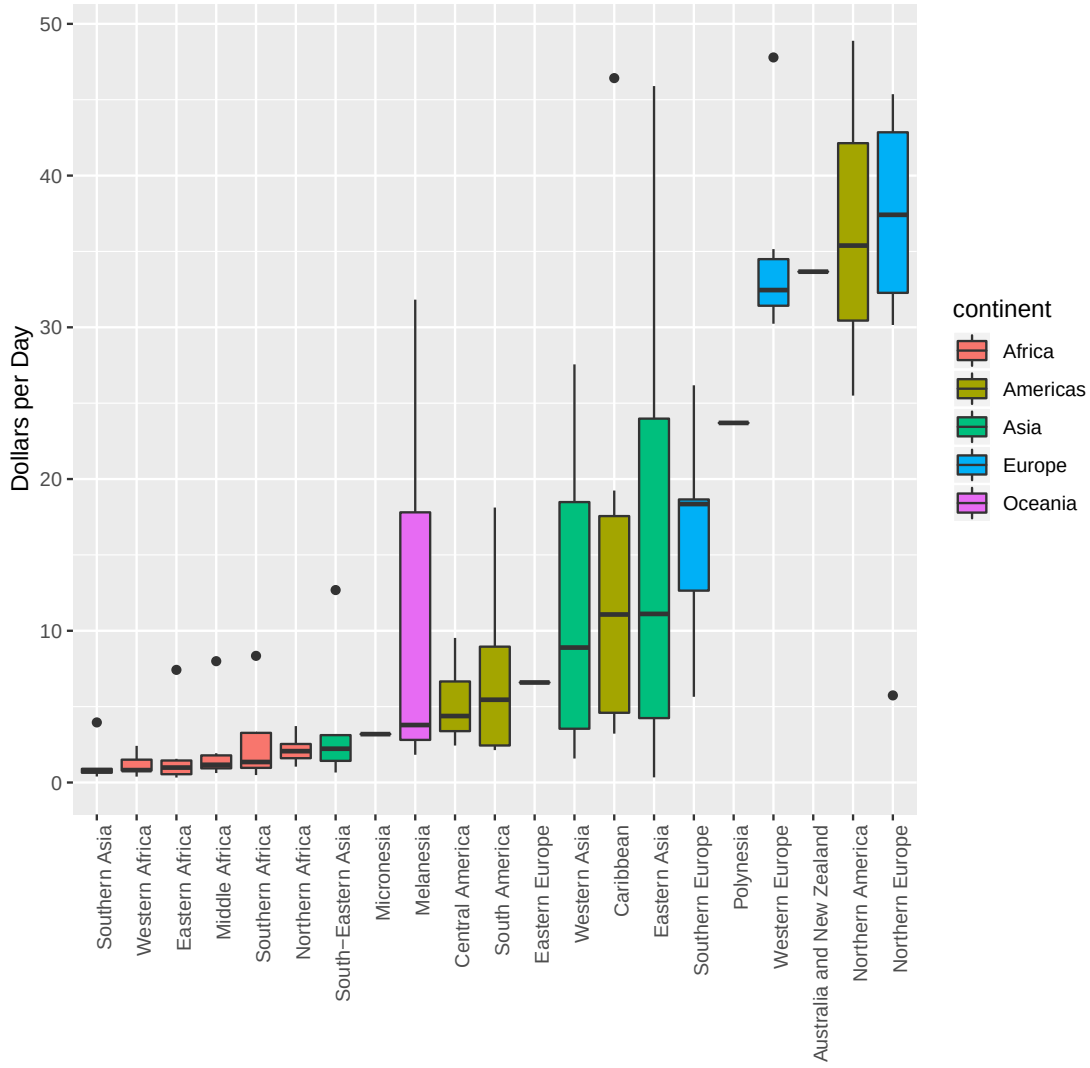


Common Axis

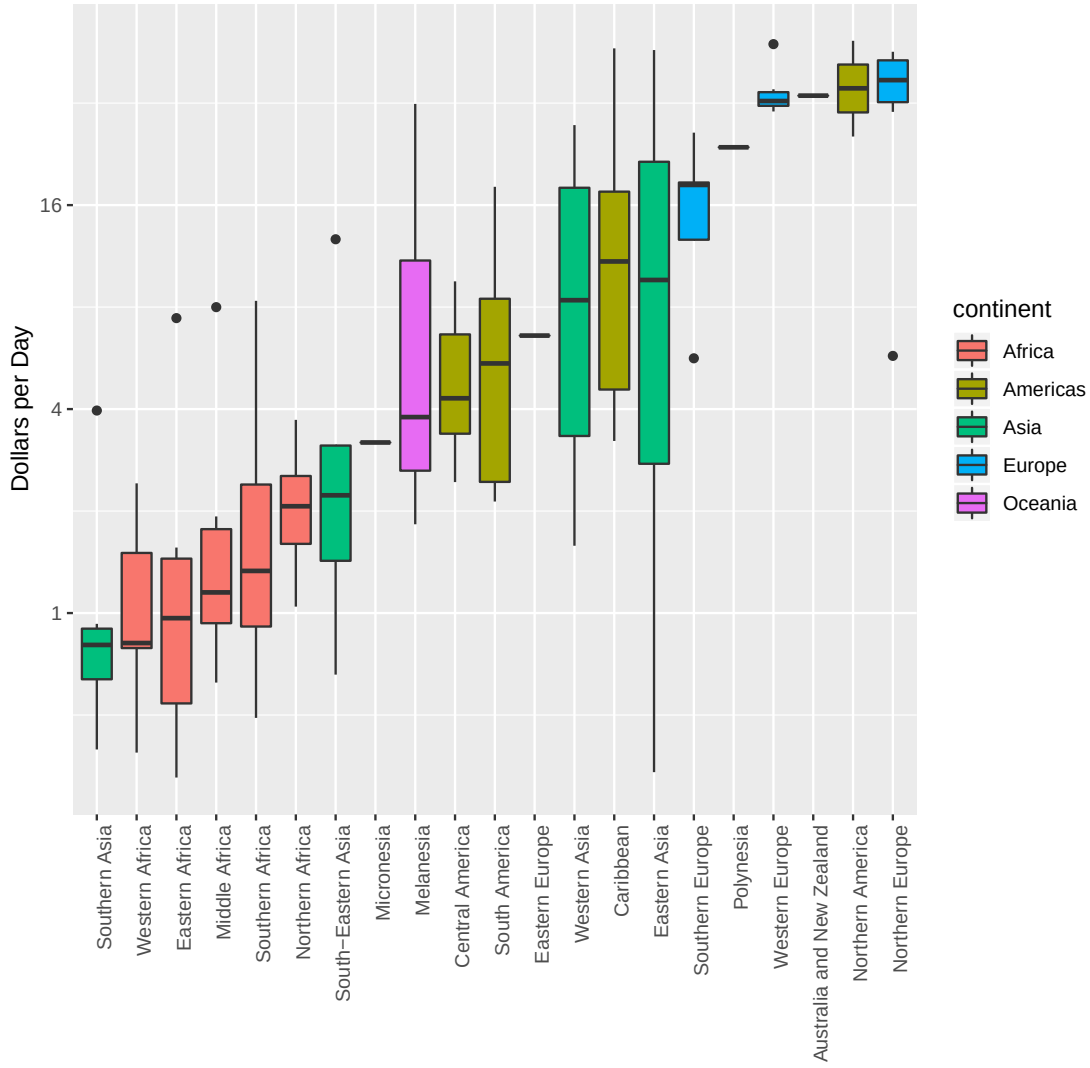
Alternative way



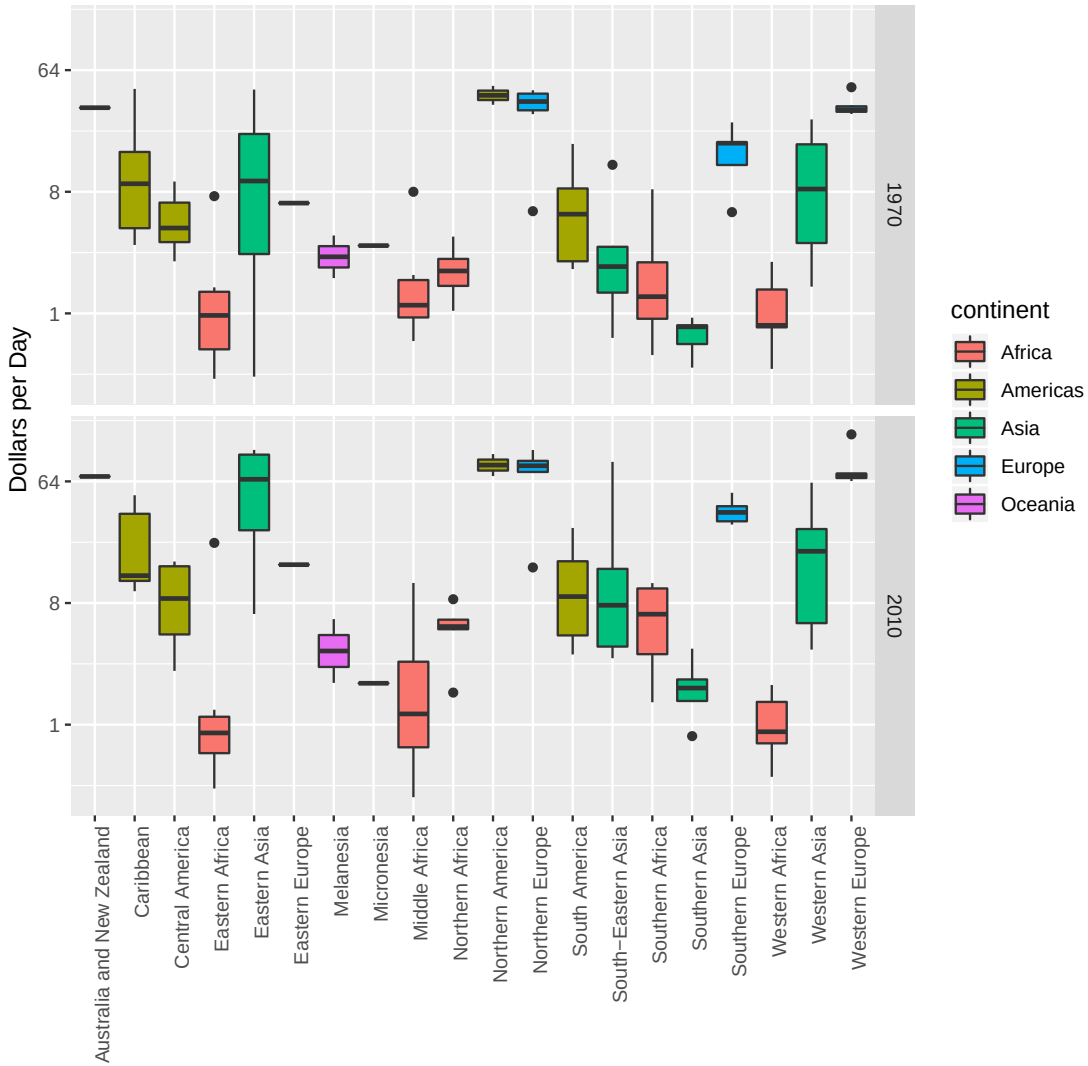
More informative plots, both in use of colors and ordering on x-axis.



Same plot on log-scale



Looking at differences in pay between 1970 and 2010



This plot makes it easier to compare directly within region.