1

Stat 104: Quantitative Methods

Class 6: Measures of Association

2

The Boxplot Rule

- One of the earliest improvements on the classic outlier detection rule is called the boxplot rule.
- It is based on the fundamental strategy of avoiding masking by replacing the mean and standard deviation with measures of location and dispersion that are relatively insensitive to outliers.

3

The BoxPlot Rule

- In particular, the boxplot rule declares the value X an outlier if

$$X < Q1 - 1.5(Q3 - Q1)$$
$$or$$
$$X > Q3 + 1.5(Q3 - Q1)$$

- So the rule is based on the lower and upper quartiles, as well as the interquartile range, which provide resistance to outliers.

4

Example

- Remember the sexual attitude data

```
> describe(mydata$x)
   vars   n  mean     sd median trimmed  mad min  max range skew kurtosis    se
X1    1 105 64.92 585.16      1    3.66 1.48   0 6000  6000 9.94    97.79 57.11
>
> summary(mydata$x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    1.00    1.00   64.92    6.00 6000.00
```

- Outlier if > 6+1.5(6-1)=13.5 so 12 points are flagged now instead of 1 as being outliers.

5

Outlier Detection in R

- Lets first look at how we drop variables, then we will discuss a function that does outlier detection.
- This is a fancier idea in R but very useful-the idea of subsetting a vector.

6

Subsetting a Vector

- Consider the following

```
> mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat104/cars10.csv")
> head(mydata)
           make price mpg headroom trunk weight length turn displacement
1   AMC Concord  4099  22      2.5    11   2930    186   40          121
2     AMC Pacer  4749  17      3.0    11   3350    173   40          258
3    AMC Spirit  3799  22      3.0    12   2640    168   35          121
4 Buick Century  4816  20      4.5    16   3250    196   40          196
5 Buick Electra  7827  15      4.0    20   4080    222   43          350
6 Buick LeSabre  5788  18      4.0    21   3670    218   43          231
  gear_ratio  foreign
1       3.58 Domestic
2       2.53 Domestic
3       3.08 Domestic
4       2.93 Domestic
5       2.41 Domestic
6       2.73 Domestic
> attach(mydata) ### this makes the variables directly available to us
```

Consider the following

```
> price
 [1]  4099  4749  3799  4816  7827  5788  4453  5189 10372  4082 11385 14500
[13] 15906  3299  5705  4504  5104  3667  3955  3984  4010  5886  6342  4389
[25]  4187 11497 13594 13466  3829  5379  6165  4516  6303  3291  8814  5172
[37]  4733  4890  4181  4195 10371  4647  4425  4482  6486  4060  5798  4934
[49]  5222  4723  4424  4172  9690  6295  9735  6229  4589  5079  8129  4296
[61]  5799  4499  3995 12990  3895  3798  5899  3748  5719  7140  5397  4697
[73]  6850 11995

> price[1]
[1] 4099
> price[2:5]
[1] 4749 3799 4816 7827

> price[price>median(price)]
 [1]  7827  5788  5189 10372 11385 14500 15906  5705  5104  5886  6342 11497
[13] 13594 13466  5379  6165  6303  8814  5172 10371  6486  5798  5222  9690
[25]  6295  9735  6229  5079  8129  5799 12990  5899  5719  7140  5397  6850
[37] 11995
```


Finding Outliers in R

- Consider the following

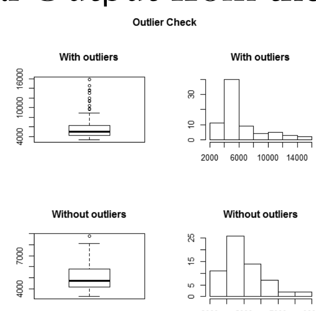
```
> summary(price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   3291    4220    5006    6165    6332   15910

> price[price>6332+1.5*IQR(price)]
 [1] 10372 11385 14500 15906 11497 13594 13466 10371  9690  9735 12990 11995

> price[price<4220-1.5*IQR(price)]
integer(0)

> boxplot.stats(price)$out   #### easier way to get the outliers
 [1] 10372 11385 14500 15906 11497 13594 13466 10371  9690  9735 12990 11995
```


Using a function

- We wrote a function to find outliers and show them graphically.

```
> source("http://people.fas.harvard.edu/~mparzen/stat104/oulierKD.txt")

> outlierKD(price)

Outliers identified: 12
Propotion (%) of outliers: 19.4
Mean of the outliers: 12125.08
Mean without removing outliers: 6165.26
Mean if we remove outliers: 5011.74
```

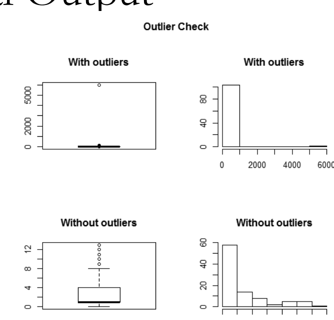

Graphical Output from the Function



Sexual Partners Data

```
> mydata=read.csv("https://goo.gl/e8nYDF")
> sexpart=mydata$x

> outlierKD(sexpart)
Outliers identified: 12
Propotion (%) of outliers: 12.9
Mean of the outliers: 544.25
Mean without removing outliers: 64.92
Mean if we remove outliers: 3.08
```


Graphical Output



Skewness

- A related idea to outliers is skewness(and one which we always wonder-do we really have outliers or is the data skewed, or both)?
- **Skewness** measures the degree of asymmetry exhibited by the data

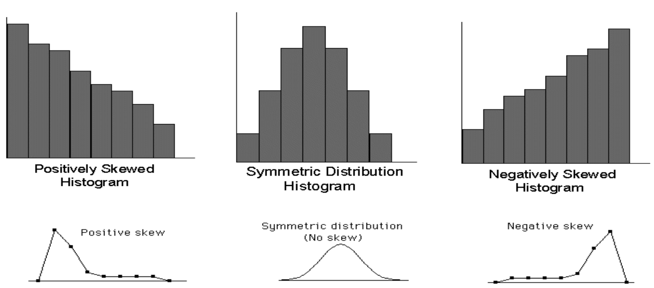
$$skewness = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{ns^3}$$

Never will calculate this by hand

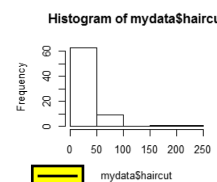
Values of Skewness

- A symmetric data set should have a skewness value near 0
- Negative values for the skewness indicate data that are skewed left and positive values for the skewness indicate data that are skewed right.
- By skewed left, we mean that the left tail is long relative to the right tail. Similarly, skewed right means that the right tail is long relative to the left tail.

Skewness

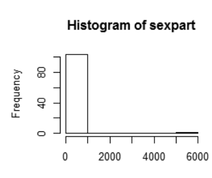


Example: Haircut Data




```
> describe(mydata$haircut)
   vars  n  mean    sd median trimmed   mad min max range s[illegible]rtosis   se
X1    1 74 32.21 38.36   21.5   25.85 17.05   0 250   250 3.[illegible]  16.2 4.46
> describe(mydata$haircut[mydata$haircut<150])
   vars  n  mean    sd median trimmed   mad min max range skew kurtosis   se
X1    1 72 26.86 20.49     20   24.67 14.83   0  85    85    1     0.52 2.41
> describe(mydata$haircut[mydata$haircut<100])
   vars  n  mean    sd median trimmed   mad min max range skew kurtosis   se
X1    1 72 26.86 20.49     20   24.67 14.83   0  85    85    1     0.52 2.41
> describe(mydata$haircut[mydata$haircut<50])
   vars  n  mean    sd median trimmed   mad min max range skew kurtosis   se
X1    1 62 20.42 12.78     17   20.21 10.38   0  48    48 0.24    -0.77 1.62
```

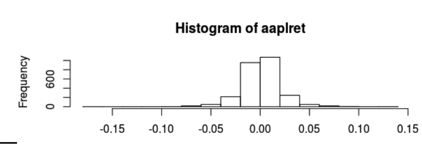

Example: Sexual Partners




```
> describe(sexpart)
   vars   n  mean     sd median trimmed  mad min  max range skew kurtosis    se
X1    1 105 64.92 585.16      1    3.66 1.48   0 6000  6000 9.94    97.79 57.11
> describe(sexpart[sexpart<150])
   vars   n mean   sd median trimmed  mad min max range skew kurtosis   se
X1    1 102 5.07 7.85      1    3.27 1.48   0  45    45 2.95     9.84 0.78
> describe(sexpart[sexpart<10])
   vars  n mean  sd median trimmed mad min max range skew kurtosis   se
X1    1 84  2.2 2.1      1    1.84   0   0   9     9 1.49     1.49 0.23
```


Remember data is time dependent

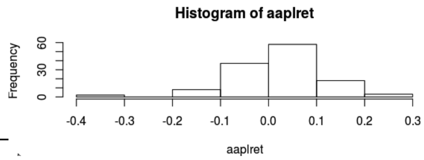
```
> library(quantmod)
> getSymbols("AAPL")
[1] "AAPL"

> aaplret=dailyReturn(Ad(AAPL))

> describe(aaplret)
              vars    n mean   sd median trimmed  mad   min  max range  skew kurtosis se
daily.returns    1 2630    0 0.02      0       0 0.01 -0.18 0.14  0.32 -0.19     6.29  0
```



Remember data is time dependent

19

```
> aaplret=monthlyReturn(Ad(AAPL))
> describe(aaplret)
                vars   n mean   sd median trimmed  mad   min  max range  skew kurtosis
monthly.returns    1 126 0.03 0.09   0.03    0.03 0.07 -0.33 0.24  0.57 -0.69     2.17
                se
monthly.returns 0.01
```



Transforming Skewed Data

20

- When a distribution is skewed, it can be hard to summarize the data simply with a center and spread, and hard to decide whether the most extreme values are outliers or just part of the stretched-out tail.
- How can we say anything useful about such data? The secret is to apply a simple function to each data value.

Nonlinear Transformations

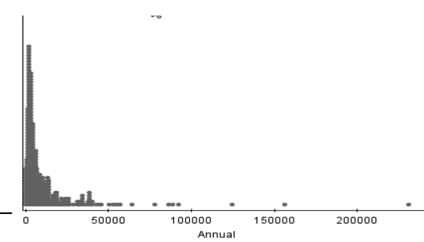
21

- Sometimes there is need to transform our data in a nonlinear way;
- Y=sqrt(X), Y=log(X), Y=1/x, etc.…
- This is usually done to try to "symmetrize" the data distribution to improve their fit to assumptions of statistical analysis (will make more sense in a few weeks).
- Basically to reduce outliers in the data and/or reduce skewness.

Your dream job

22

- Consider the graph below which shows 2005 CEO data for the Fortune 500. The data is in <u>thousands of dollars</u>.



The data is heavily skewed

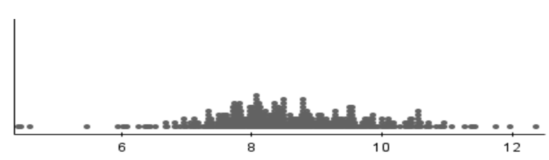
23

- Skewed distributions are difficult to summarize. It's hard to know what we mean by the "center" of a skewed distribution, so it's not obvious what value to use to summarize the distribution.
- What would you say was a typical CEO total compensation? The mean value is $10,307,000, while the median is "only" $4,700,000.

Log the data

24

- One way to make a skewed distribution more symmetric is to re-express, or transform, the data by applying a simple function to all the data values.



The Transform Cheat Sheet

- Calculate the skewness statistic for your data set
- If |skewness| < 0.8 data set is cool and unlikely to disrupt our analysis.
- Otherwise, try a transformation in the "ladder of powers"

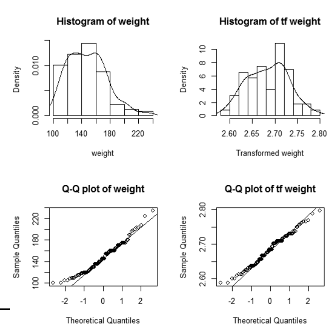
λ	-2	-1	-1/2	0	1/2	1	2
y	$\frac{-1}{x^2}$	$\frac{-1}{x}$	$\frac{-1}{\sqrt{x}}$	$\log x$	$\sqrt{x}$	x	x^2

The Transform Cheat Sheet

- R has a command **gboxcoxnc(varname) in the AID** package which makes searching for a transformation easy.

```
> boxcoxnc(weight)

  Box-Cox power transformation
  ---------------------------------------------------------------
  data : weight

  lambda.hat : -0.28


  Shapiro-Wilk normality test for transformed data (alpha = 0.05)
  ---------------------------------------------------------------

  statistic  : 0.9863134
  p.value    : 0.3020307

  Result     : Transformed data are normal.
  ---------------------------------------------------------------
```


Graphical Output from boxcoxnc()



Today's Tools

- New toolbox additions
  - Transformations, Skewness, Outliers
  - Empirical Rule

Things you should know

- Emprical Rule, Chebyshev's Rule
- a+bX rule
- Z scoring
- Detecting Outliers
- Skewness and Transformations

Covariance and correlation

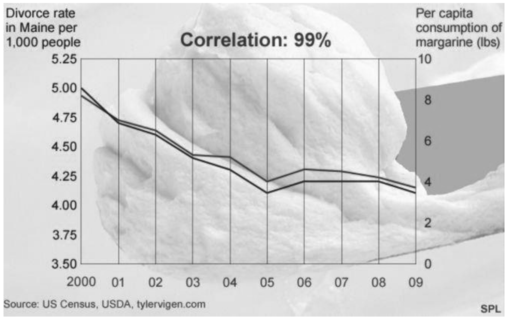
The mean and sd help us summarize a bunch of numbers which are measurements of just one thing.

A fundamental and totally different question is how one thing relates to another.

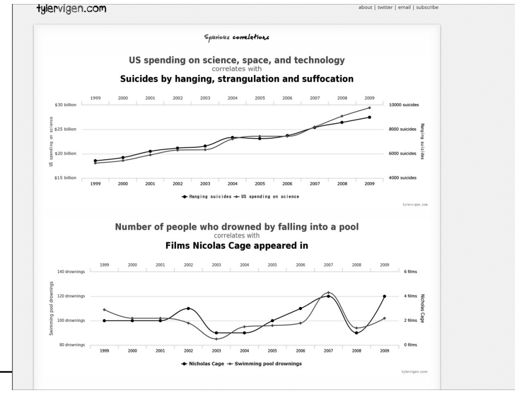
Previously, we used a scatterplot to look at two things: the mean and sd of different assets.

In this section of the notes we look at scatterplots and how correlation can be used to summarize them.

31



32



34

In general we have observations

(x_i, y_i) ⟵ the ith observation is a pair of numbers

Our data looks like:

x	y	i
12.0	192	1
12.0	160	2
5.0	155	3
5.0	120	4
7.0	150	5
13.0	175	6
4.0	100	7
12.0	165	8

......

The plot enables us to see the relationship between x and y.

35

In the beer example, it does look like there is a relationship. Even more, the relationship looks linear in that it looks like we could draw a line through the plot to capture the pattern.

Covariance and correlation summarize how strong a *linear* relationship there is between two variables.

In the example weight and nbeers were the two variables.

In general we think of them as x and y.

At this point we **don't care** which is x and which is y

36

Covariance

Consider two variables, X and Y.

The concept of covariance asks:

Is Y larger (or smaller) when X is larger ?

We measure this using something called covariance s_{xy}

Covariance > 0 Larger X ⟺ Larger Y

Covariance < 0 Larger X ⟺ Smaller Y

37

Here is the actual formula but most people never calculate covariance by hand………

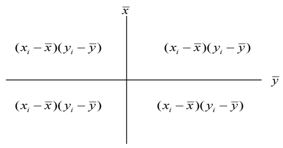
The sample covariance between x and y is:

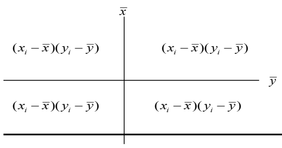
$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

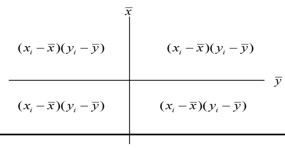
What are the units of covariance ?

Understanding covariance

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$







In this example, we look at the relationship between team payroll and team performance in Major League Baseball using data from the 2010 season (for a total of 30 teams).

The variables of interest:

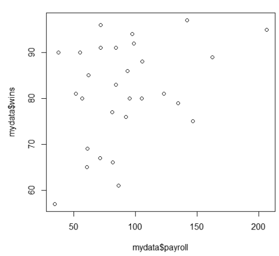
Payroll team payroll (in millions of dollars)

Wins number of games out of 162 that the team won.

```
mydata=read.csv("https://goo.gl/SsfWgg")
```


The Data

team	payroll	wins	winpct
New York Yankees	206.3	95	0.58642
Boston Red Sox	162.4	89	0.54938
Chicago Cubs	146.6	75	0.46296
Philadelphia Philli	141.9	97	0.59877
New York Mets	134.4	79	0.48765
Detroit Tigers	122.9	81	0.5
Chicago White Sox	105.5	88	0.54321
Los Angeles Angels	105	80	0.49383
San Francisco Gian	98.6	92	0.5679
Minnesota Twins	97.6	94	0.58025
Los Angeles Dodge	95.4	80	0.49383
St. Louis Cardinals	93.5	86	0.53086
Houston Astros	92.4	76	0.46914
Seattle Mariners	86.5	61	0.37654
Atlanta Braves	84.4	91	0.56173
Colorado Rockies	84.2	83	0.51235
Baltimore Orioles	81.6	66	0.40741
Milwaukee Brewer	81.1	77	0.47531
Tampa Bay Rays	71.9	96	0.59259
Cincinnati Reds	71.8	91	0.56173
Kansas City Royals	71.4	67	0.41358
Toronto Blue Jays	62.2	85	0.52469
Washington Natior	61.4	69	0.42593
Cleveland Indians	61.2	69	0.42593
Arizona Diamondb	60.7	65	0.40123
Florida Marlins	57	80	0.49383
Texas Rangers	55.3	90	0.55556
Oakland Athletics	51.7	81	0.5
San Diego Padres	37.8	90	0.55556
Pittsburgh Pirates	34.9	57	0.35185



Would you say the covariance is positive, negative or zero ?

Calculating Covariance in R

```
> head(mydata)
               team payroll wins    winpct
1      New York Yankees   206.3   95 0.5864198
2        Boston Red Sox   162.4   89 0.5493827
3          Chicago Cubs   146.6   75 0.4629630
4 Philadelphia Phillies   141.9   97 0.5987654
5         New York Mets   134.4   79 0.4876543
6        Detroit Tigers   122.9   81 0.5000000
> cov(mydata[,-1])
          payroll        wins      winpct
payroll 1461.5032644 154.7241379 0.955087269
wins     154.7241379 121.1034483 0.747552151
winpct     0.9550873   0.7475522 0.004614519
```

This is called a covariance matrix

The Covariance Matrix

- It turns out that Cov(X,X)=Var(X). Weird, I know.
- This comes from the formula

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- If we put "x" in for "y" we obtain

$$s_{xx} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x}) = s_x^2$$

The Covariance Matrix

- So….

Variance of Wins

```
          payroll        wins      winpct
payroll 1461.5032644 154.7241379 0.955087269
wins     154.7241379 121.1034483 0.747552151
winpct     0.9550873   0.7475522 0.004614519
```

Covariance of Wins and Payroll

```
> var(mydata$payroll)
[1] 1461.503
```

Beware of Interpreting Covariance

- Covariance depends on the units!

```
              payroll         wins       winpct
payroll 1461.5032644 154.7241379 0.955087269
wins     154.7241379 121.1034483 0.747552151
winpct     0.9550873   0.7475522 0.004614519
```

Only the **sign** of covariance matters

Making Size Matter

- Does a covariance of 154.72 imply a strong or weak relationship ?
- Solution: The correlation coefficient

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

covariance

Standard deviation of x

Standard deviation of y

The Correlation

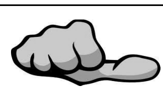
- A numerical summary of the strength of a linear relationship between two variables
- Correlations are bound between –1 and 1
- Sign: direction of the relationship (+ or -)
- Absolute value: strength of the relationship. Example: -0.6 is a stronger relationship than +0.4

Correlation in R

```
> cor(mydata[,-1])
          payroll      wins    winpct
payroll 1.0000000 0.3677731 0.3677731
wins    0.3677731 1.0000000 1.0000000
winpct  0.3677731 1.0000000 1.0000000
```

What is the correlation of Payroll with Payroll or WinPct with WinPct ?

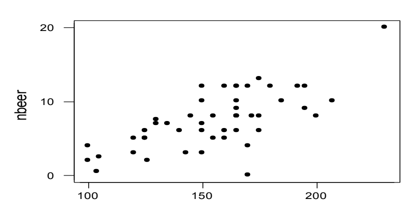
Rule of Thumb



Magnitude of r	Interpretation
.00-.20	Very weak
.20-.40	Weak to moderate
.40-.60	Medium to substantial
.60-.80	Very Strong
.80-1.00	Extremely Strong

The correlation corresponding to the scatterplot we looked at earlier is:

```
Correlation of nbeer and weight = 0.692
```



50

Caution : Correlation only measures *linear* relationships !

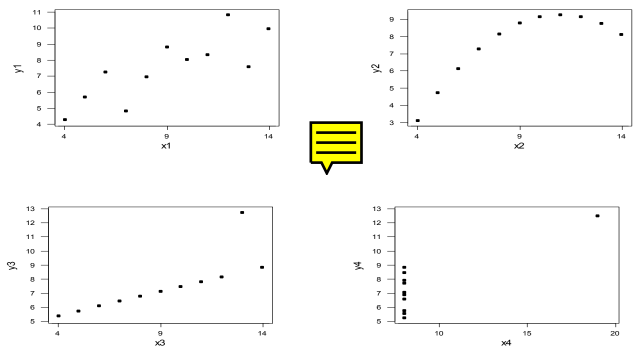
These four data sets all have correlation r=0.816

x1	y1	x2	y2	x3	y3	x4	y4
10.00	8.04	10.00	9.14	10.00	7.46	8.00	6.58
8.00	6.95	8.00	8.14	8.00	6.77	8.00	5.76
13.00	7.58	13.00	8.74	13.00	12.74	8.00	7.71
9.00	8.81	9.00	8.77	9.00	7.11	8.00	8.84
11.00	8.33	11.00	9.26	11.00	7.81	8.00	8.47
14.00	9.96	14.00	8.10	14.00	8.84	8.00	7.04
6.00	7.24	6.00	6.13	6.00	6.08	8.00	5.25
4.00	4.26	4.00	3.10	4.00	5.39	19.00	12.50
12.00	10.84	12.00	9.13	12.00	8.15	8.00	5.56
7.00	4.82	7.00	7.26	7.00	6.42	8.00	7.91
5.00	5.68	5.00	4.74	5.00	5.73	8.00	6.89

```
Pearson correlation of x1 and y1 = 0.816
Pearson correlation of x2 and y2 = 0.816
Pearson correlation of x3 and y3 = 0.816
Pearson correlation of x4 and y4 = 0.816
```

51

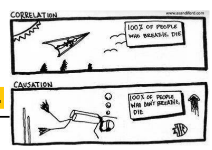
But the scatterplots all differ (bigtime!)



52

Correlation and Causation

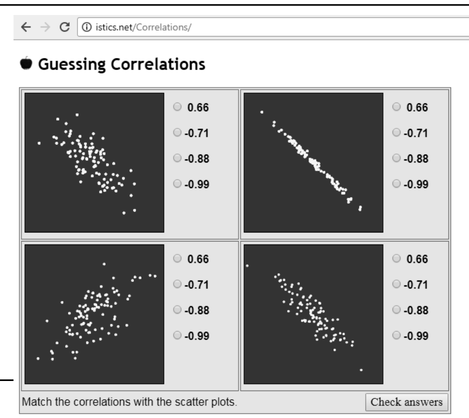
- There is strong correlation between:
  - The number of teachers in a school district and the number of failing students.
  - The number of automobiles in California per year and the number of homicides.
  - Kids' feet lengths and reading ability



Correlation does not imply causation.

More on this in future lectures.

53



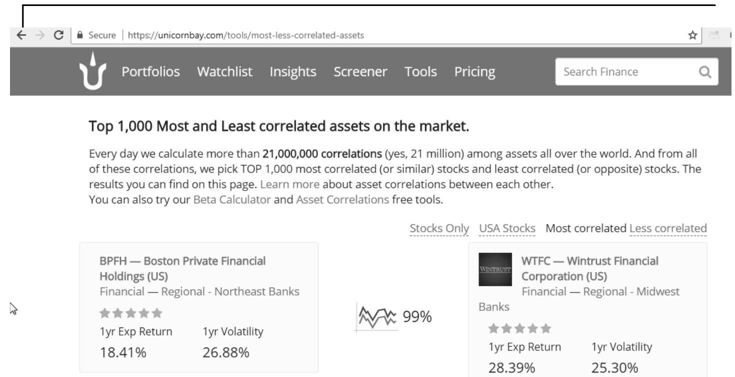
54

Correlation has Financial Implications

Correlation Matrix: Daily % Change Correlation Over Last Ten Years

Ticker	S&P 500	C. Disc.	C. Stap	Energy	Finan.	H Care	Indust.	Mater.	Tech	Telcom	Utilit.	Oil	Gold	Dollar	L Bnd
S&P 500	1.00	0.94	0.85	0.81	0.88	0.86	0.94	0.89	0.90	0.78	0.76	0.24	-0.01	-0.17	-0.37
Cons. Disc.	0.94	1.00	0.81	0.68	0.82	0.78	0.90	0.82	0.85	0.72	0.67	0.16	-0.07	-0.12	-0.34
Cons. Stap.	0.85	0.81	1.00	0.65	0.67	0.81	0.78	0.71	0.70	0.69	0.71	0.12	-0.06	-0.10	-0.28
Energy	0.81	0.68	0.65	1.00	0.62	0.66	0.73	0.81	0.66	0.59	0.70	0.48	0.16	-0.28	-0.30
Financials	0.88	0.82	0.67	0.62	1.00	0.68	0.82	0.73	0.72	0.64	0.57	0.16	-0.06	-0.13	-0.31
H Care	0.86	0.78	0.81	0.66	0.68	1.00	0.78	0.71	0.73	0.68	0.69	0.14	-0.04	-0.12	-0.29
Industrials	0.94	0.90	0.78	0.73	0.82	0.78	1.00	0.87	0.84	0.70	0.68	0.21	-0.01	-0.18	-0.37
Materials	0.89	0.82	0.71	0.81	0.73	0.71	0.87	1.00	0.79	0.65	0.67	0.29	0.15	-0.27	-0.35
Technology	0.90	0.85	0.70	0.66	0.72	0.73	0.84	0.79	1.00	0.71	0.63	0.16	-0.04	-0.10	-0.35
Telecom	0.78	0.72	0.69	0.59	0.64	0.68	0.70	0.65	0.71	1.00	0.63	0.12	-0.05	-0.10	-0.25
Utilities	0.76	0.67	0.71	0.70	0.57	0.69	0.68	0.67	0.63	0.63	1.00	0.19	0.02	-0.15	-0.22
Oil	0.24	0.16	0.12	0.48	0.16	0.14	0.21	0.29	0.16	0.12	0.19	1.00	0.29	-0.30	-0.22
Gold	-0.01	-0.07	-0.06	0.16	-0.06	-0.04	-0.01	0.15	-0.04	-0.05	0.02	0.29	1.00	-0.43	0.07
Dollar	-0.17	-0.12	-0.10	-0.28	-0.13	-0.12	-0.18	-0.27	-0.10	-0.10	-0.15	-0.30	-0.43	1.00	-0.05
Long Bond	-0.37	-0.34	-0.28	-0.30	-0.31	-0.29	-0.37	-0.35	-0.35	-0.25	-0.22	-0.22	0.07	-0.05	1.00

55



Important for Diversification

Correlation Summary

- Scatter diagrams show relationships between variables
- The covariance gives you the *direction* of a linear relationship between the two variables
- The correlation coefficient measures the *strength* of a linear relationship
- Correlation ranges between -1 and 1
- Covariance can be any number
- Both covariance and correlation measure association, not causation
- They can be misleading if there are outliers or a nonlinear association