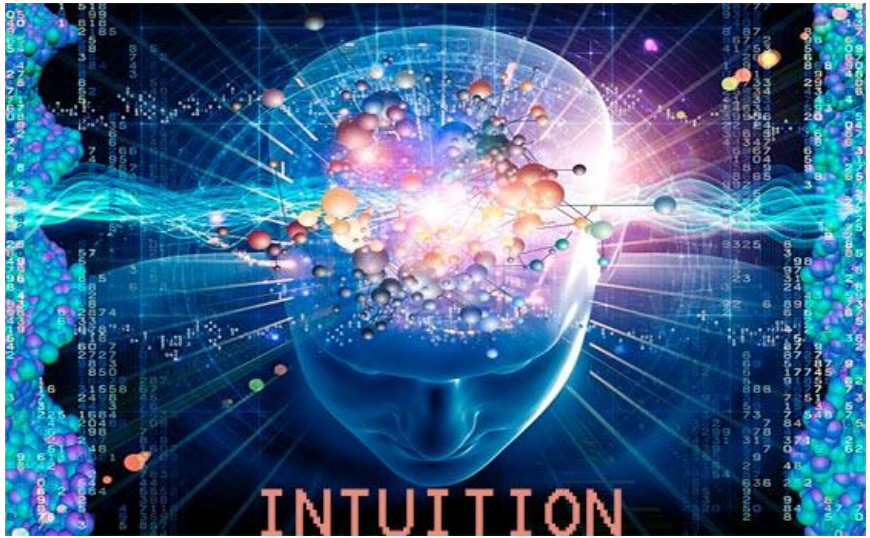
Data Analytics

Professor Ernesto Lee

Rocks and Slingshots

Tech and Intuition

- Details and math equations and code and tools can be thought of as the **rocks** that do the actual work, but are cheap & plentiful
- Intuition (the **slingshot**) is the framework that makes the details effective.
    - In pure theory, intuition is optional but in reality - intuition is absolutely not optional.



The Slingshot in your Head

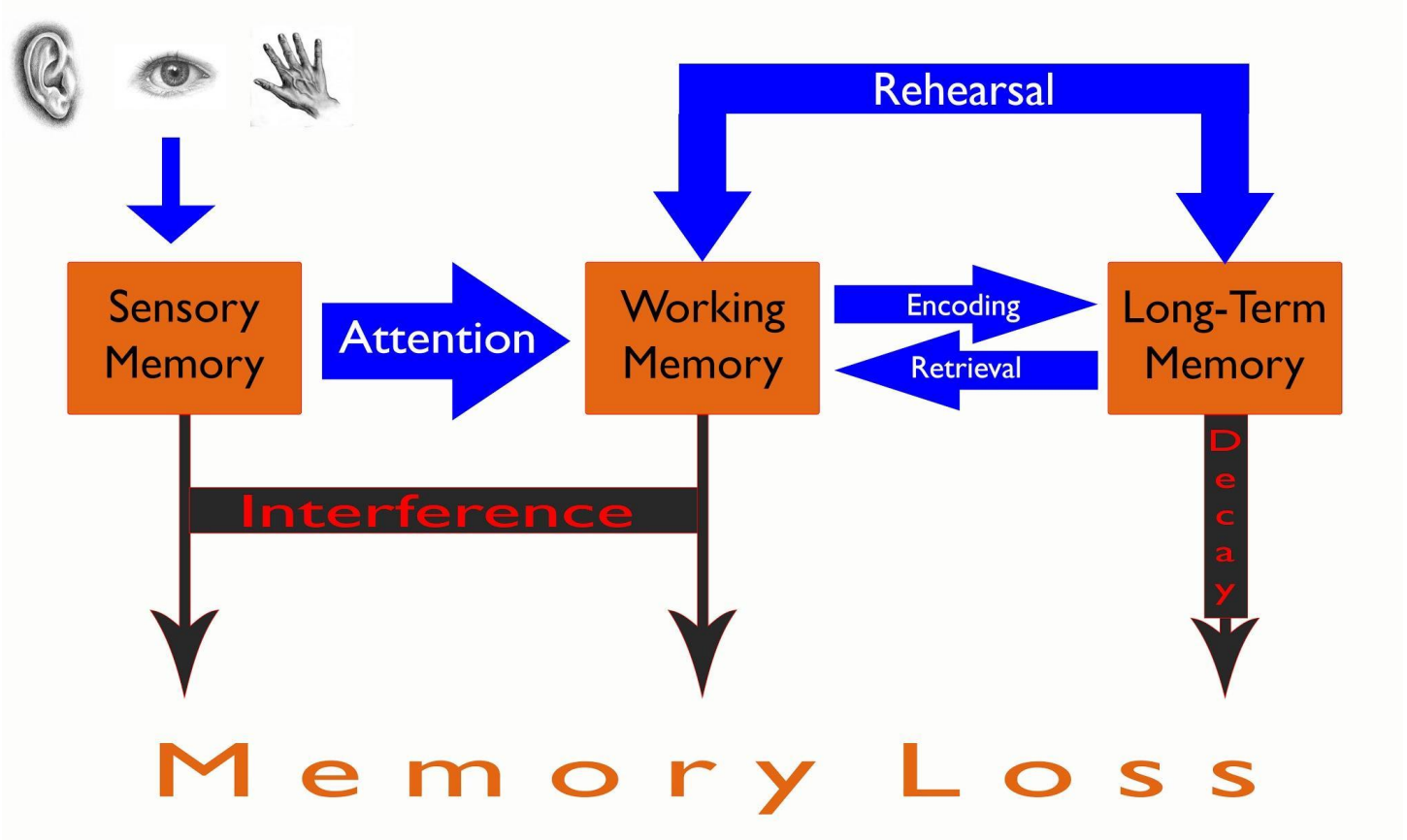
Approaches to Learning Data Analytics

The Dynamite Rock

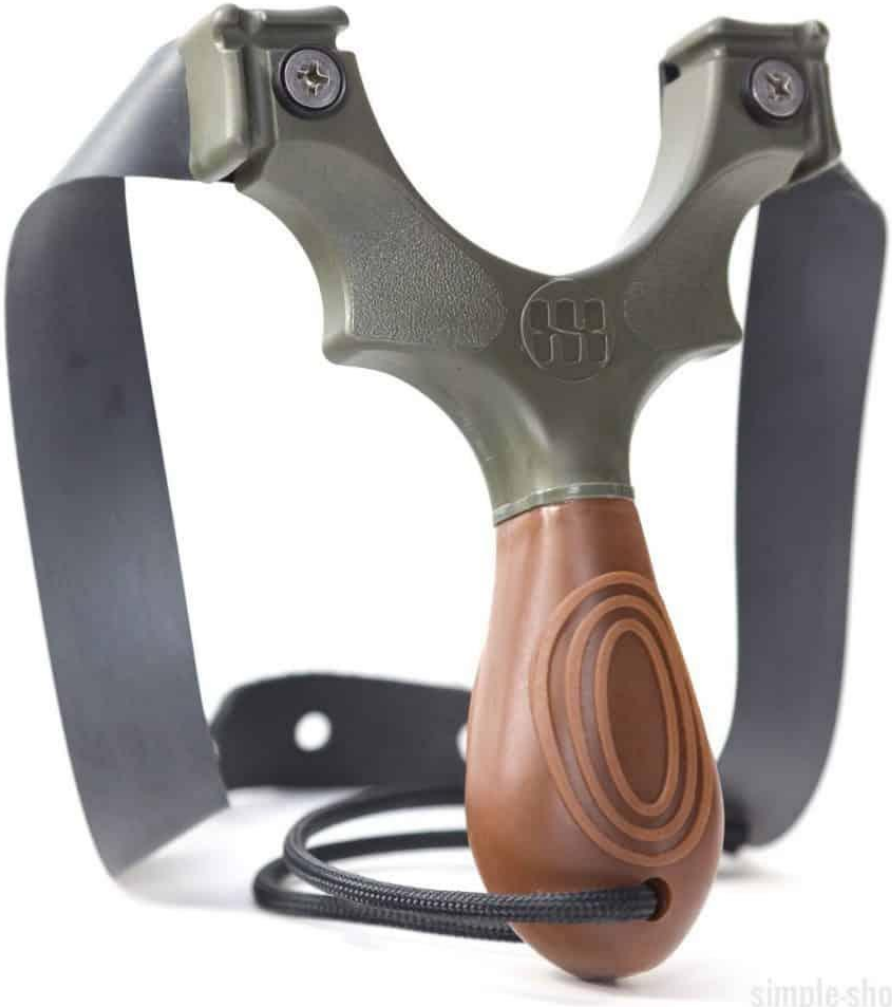
Finding the Essence

Signs of Learning

Cheap Entertainment



Use this class to upgrade your slingshot…

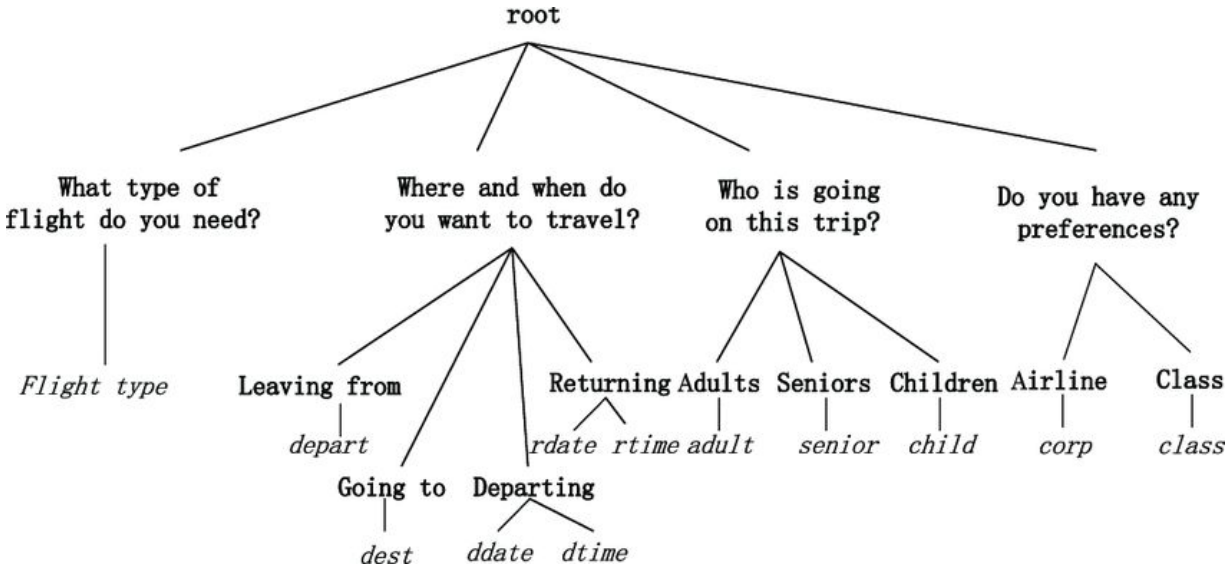


Part II : How to Learn in My Class

ADEPT Method for Learning	
Analogy	Tell me what it's like.
Diagram	Help me visualize it.
Example	Allow me to experience it.
Plain English	Describe it with everyday words.
Technical Definition	Discuss the formal details.

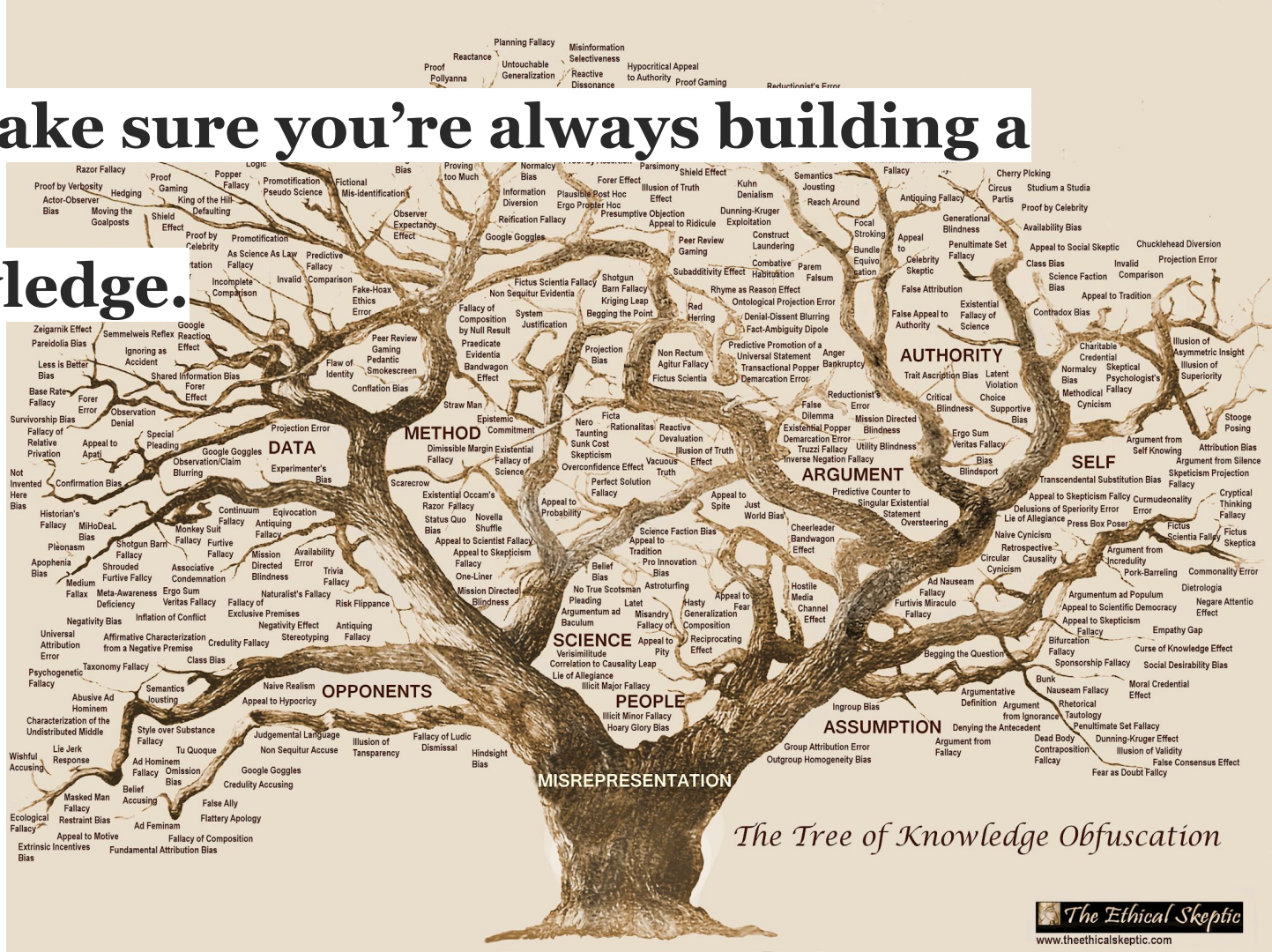
Semantic Tree

Elon Musk believes that learning involves understanding what he called a *Semantic Tree*. You must learn this concept and identify the different parts of the tree.



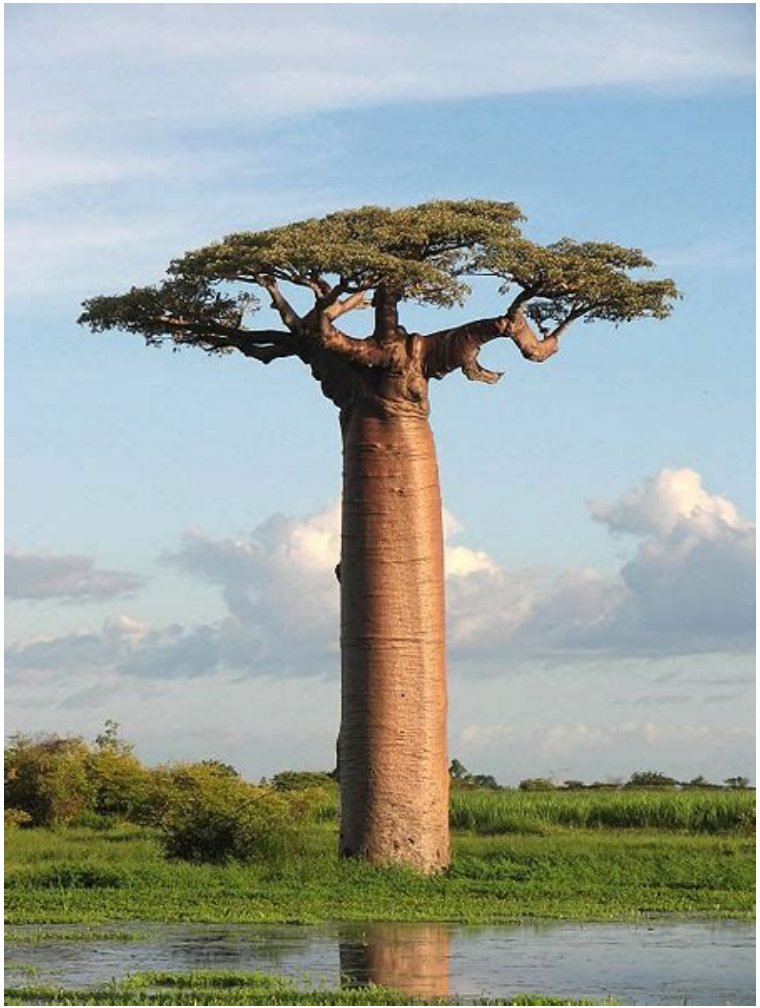
“One bit of advice: it is important to view knowledge as sort of a semantic tree — make sure you understand the fundamental principles, i.e. the trunk and big branches, before you get into the leaves/details or there is nothing for them to hang on to.”

Rule #1 — Make sure you're always building a tree of knowledge.

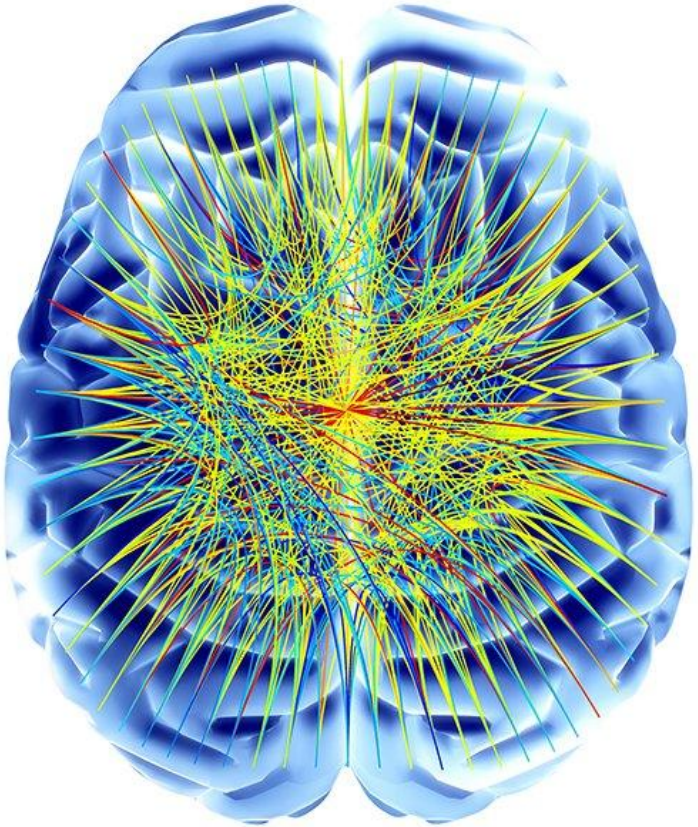


Your Tree

Your brain is a network. Therefore it is the CONNECTIONS that are truly the most powerful part of your learning.

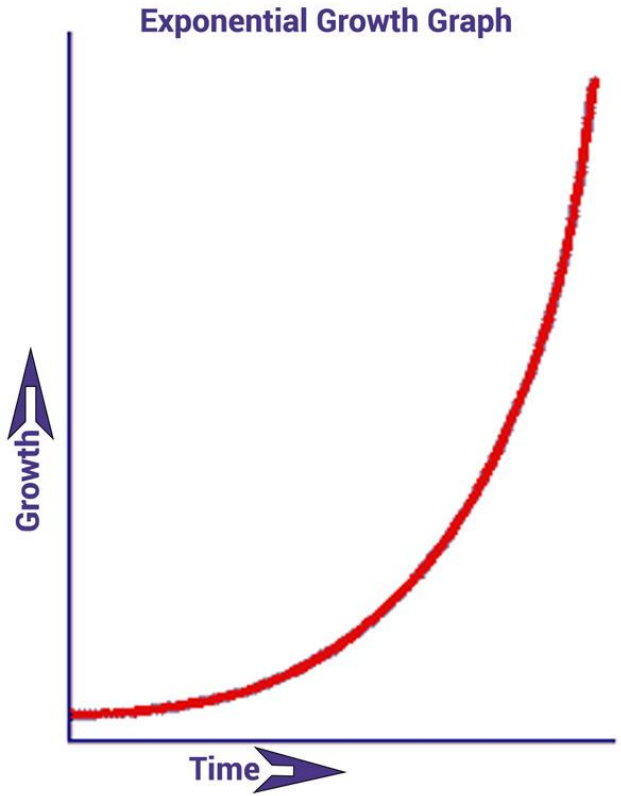


Rule #2 — You can never remember what you cannot connect with.



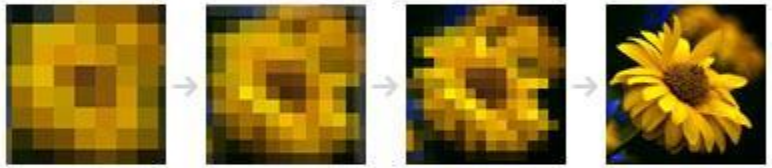
What good are a bunch of sticks?

Exponential Growth



Part III : Your Personal Hero Journey

✓ Simple JPEG

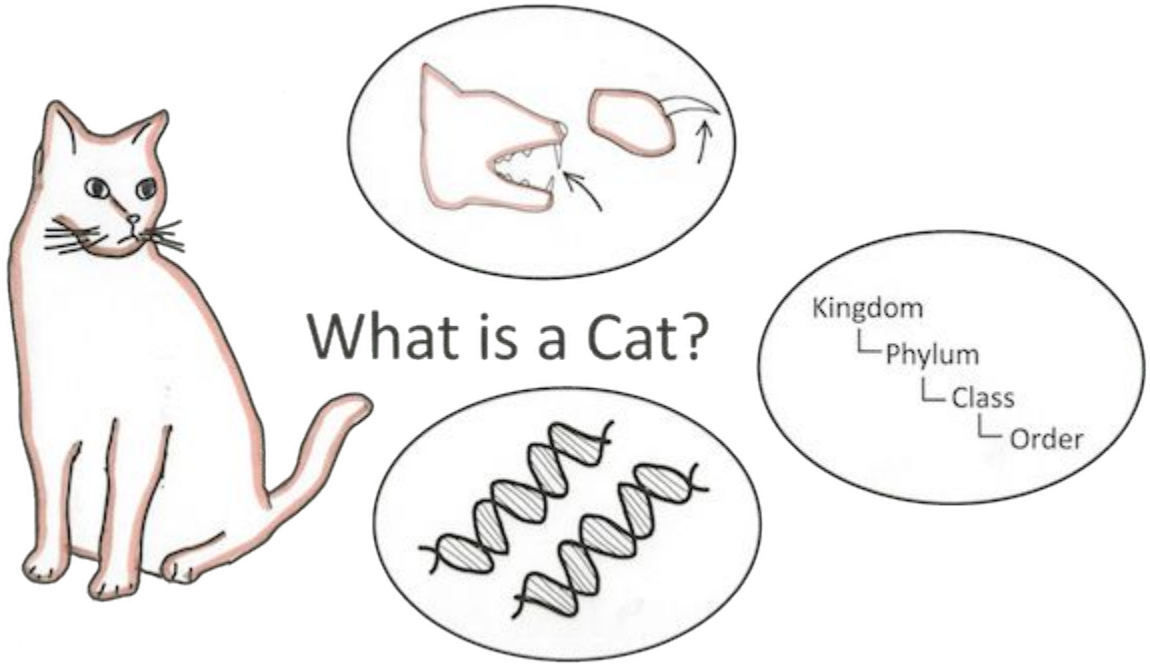


✓ Progressive JPEG

✗ Preload and then show

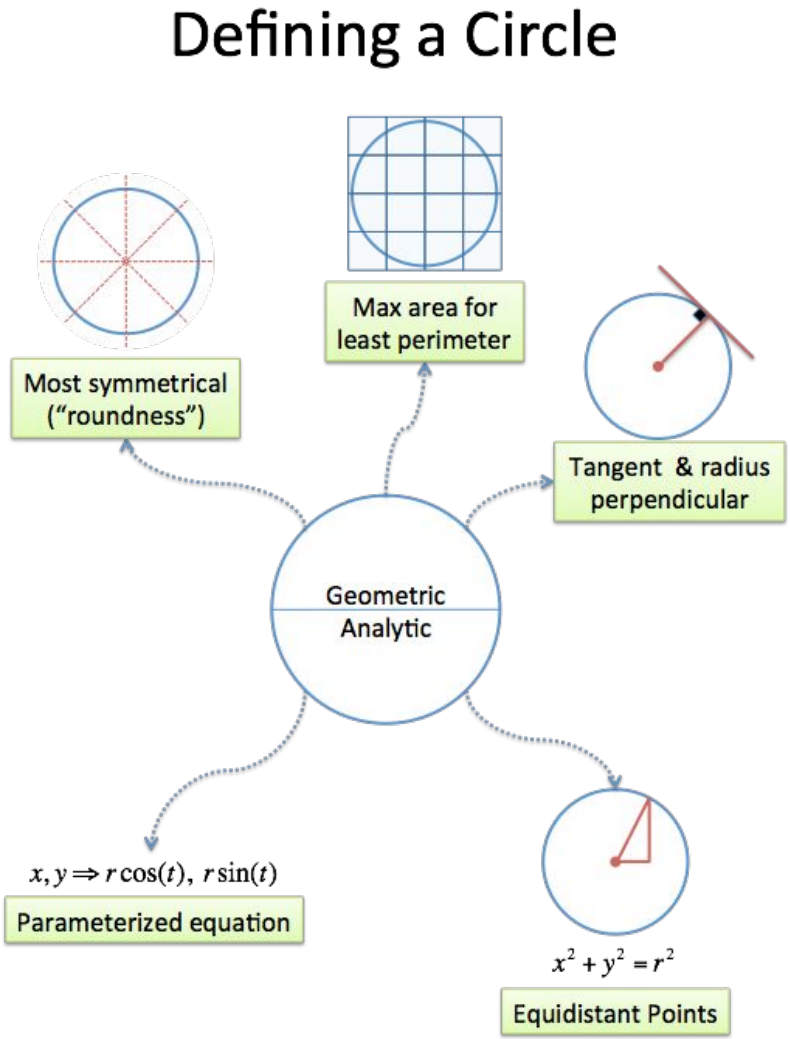
Develop Insights

- **Caveman definition:** A furry animal with claws, teeth, a tail, 4 legs, that purrs when happy and hisses when angry…
- **Evolutionary definition:** Mammalian descendants of a certain species (*F. catus*), sharing certain characteristics…
- **Modern definition:** You call those *definitions*? Cats are animals sharing the following DNA: ACATACATACATACAT…



An Example

Develop Intuitions



- **Step 1: Find the central theme of a math concept.** This can be difficult, but try starting with its history. Where was the idea first used? What was the discoverer doing? This use may be different from our modern interpretation and application.
- **Step 2: Explain a property/fact using the theme.** Use the theme to make an analogy to the formal definition. If you're lucky, you can translate the math equation ($x^2 + y^2 = r^2$) into a plain-english statement ("All points the same distance from the center").
- **Step 3: Explore related properties using the same theme**. Once you have an analogy or interpretation that works, see if it applies to other properties. Sometimes it will, sometimes it won't (and you'll need a new insight), but you'd be surprised what you can discover.

What's the point to all of this?

- **Search for insights and apply them.** That first intuitive insight can help everything else snap into place. Start with a definition that makes sense and "walk around the circle" to find others.
- **Develop mental toughness.** Banging your head against an idea is no fun. If it doesn't click, come at it from different angles. There's another book, another article, another person who explains it in a way that makes sense to you.
- **It's ok to be visual.** We think of math as rigid and analytic — but visual interpretations are ok! Do what develops your understanding. Imaginary numbers were puzzling until their geometric interpretation came to light, decades after their initial discovery. Looking at equations all day didn't help mathematicians "get" what they were about.

Part IV: Your Assignment

- Build a **lasting intuition** for the key ideas.
- During the course, understand it enough to solve problems.
- After the course, enjoy it enough to revisit.

ADEPT Method for Learning	
Analogy	Tell me what it's like.
Diagram	Help me visualize it.
Example	Allow me to experience it.
Plain English	Describe it with everyday words.
Technical Definition	Discuss the formal details.

Machine Learning Notes

1-sentence course summary

- "Create predictive models with Linear Algebra and improve them with Calculus."
    - Linear Algebra efficiently describes complex operations using simple steps.
    - Calculus can optimize any model with derivatives (gradient). Minimize error.

1-sentence core concepts

- Linear Algebra: spreadsheets for your equations. We "pour" data through various operations.
- Natural log: time needed to grow. Helps normalize widely varying numbers.
- e^x: models continuous growth, has a simple derivative.
- Gradient: direction of greatest change, helps optimize.
- Calculus -Art of breaking a system into steps. With the gradient, we can move in the best direction.

Aha!

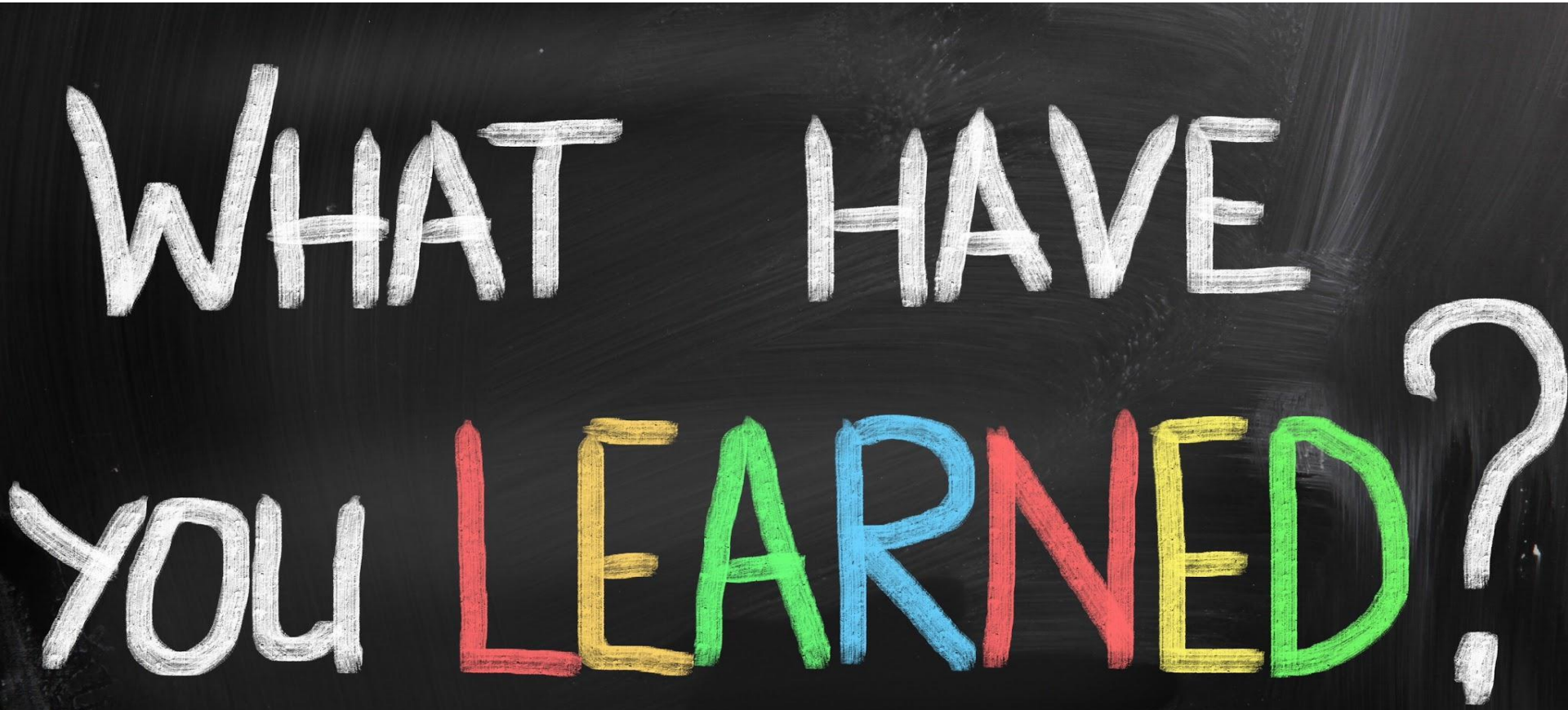
For the major concepts the course depends on, I keep a 5-second summary in mind. This underlying concept, why does it exist? In plain English, what does it mean?

Aha!

Huh's - they can be a question that you have for anything that is unclear. They can be a GOTCHA!

Record them as well!

Summarize this course in PLAIN ENGLISH

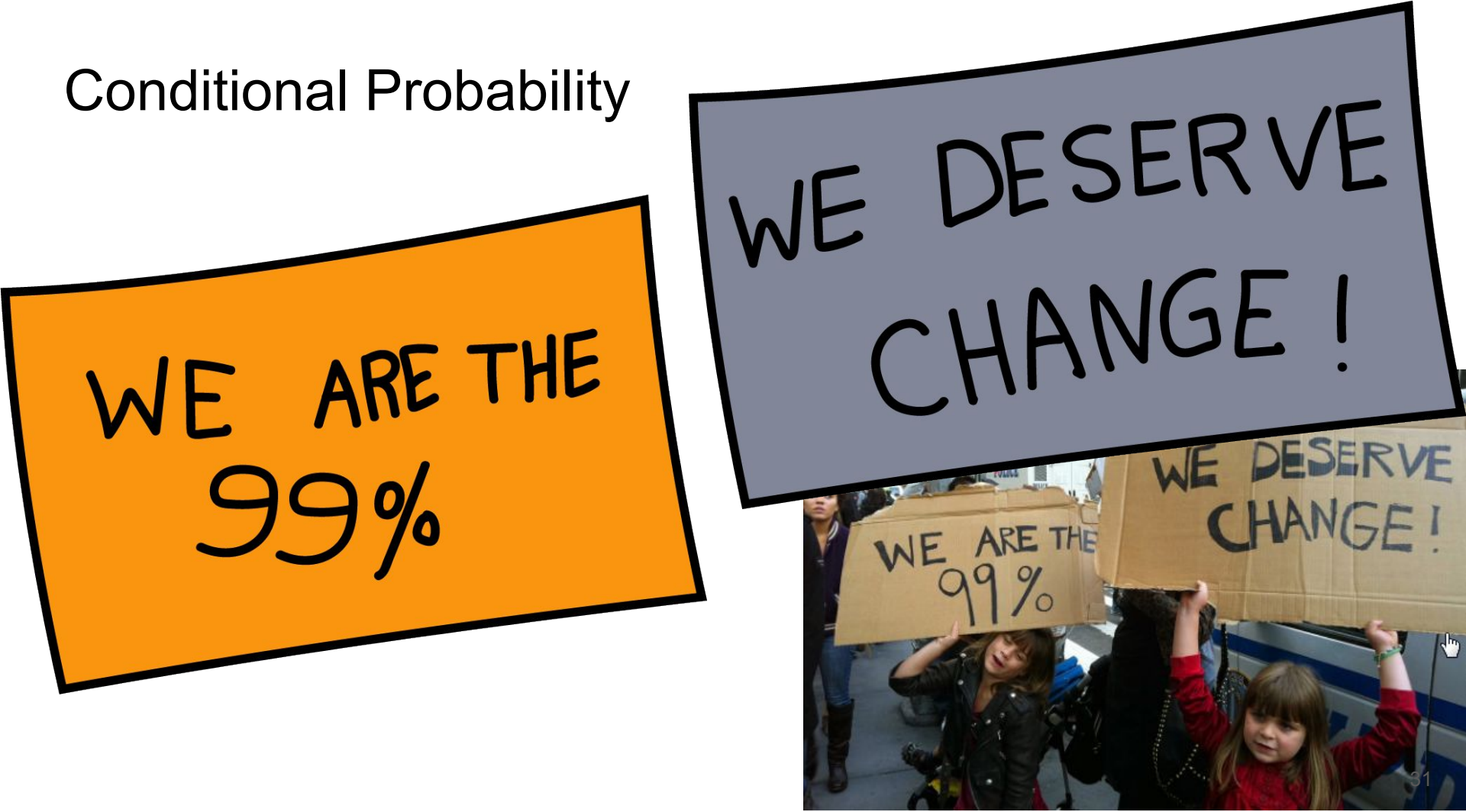


The Results...

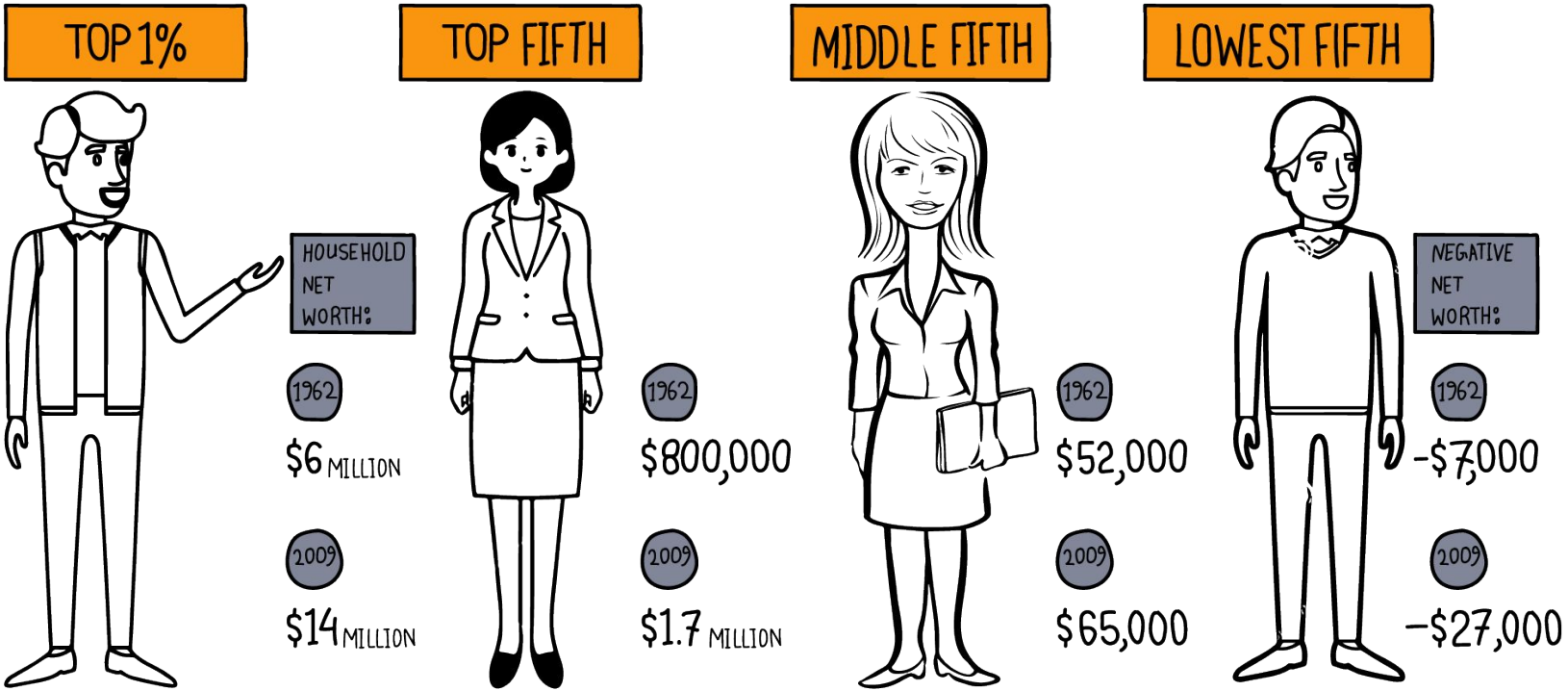
Naive Bayes

By: Ernesto Lee

Conditional Probability



What are the chances of being part of the 1%?



Research

Research has shown that ¼ of the 1% is female

Math Alert:

¼ of 1% is female

or

.25 * .01 = .0025 is female

What are the chances of being FEMALE if you are 1%?

CHANCES OF BEING FEMALE AND IN THE 1%

=

CHANCES OF BEING FEMALE IF YOU ARE IN THE 1%

X

CHANCES OF BEING IN THE 1%

What are the chances of being 1% if you are FEMALE?

CHANCES OF BEING FEMALE AND IN THE 1%

=

CHANCES OF BEING IN THE 1% IF YOU ARE FEMALE

X

CHANCES OF BEING FEMALE

What are the chances of being 1% if you are FEMALE?

CHANCES OF BEING FEMALE AND IN THE 1% = CHANCES OF BEING IN THE 1% IF YOU ARE FEMALE × CHANCES OF BEING FEMALE

CHANCES OF BEING IN THE 1% IF YOU ARE FEMALE = CHANCES OF BEING FEMALE AND IN THE 1% / CHANCES OF BEING FEMALE

CHANCES OF BEING FEMALE AND IN THE 1%

=

CHANCES OF BEING IN THE 1% IF YOU ARE FEMALE

X

CHANCES OF BEING FEMALE

CHANCES OF BEING IN THE 1% IF YOU ARE FEMALE

=

$$\frac{\text{CHANCES OF BEING FEMALE AND IN THE 1\%}}{\text{CHANCES OF BEING FEMALE}}$$

$x * 0.5 = .0025$

$x = .0025 / 0.5$

$x = .005$

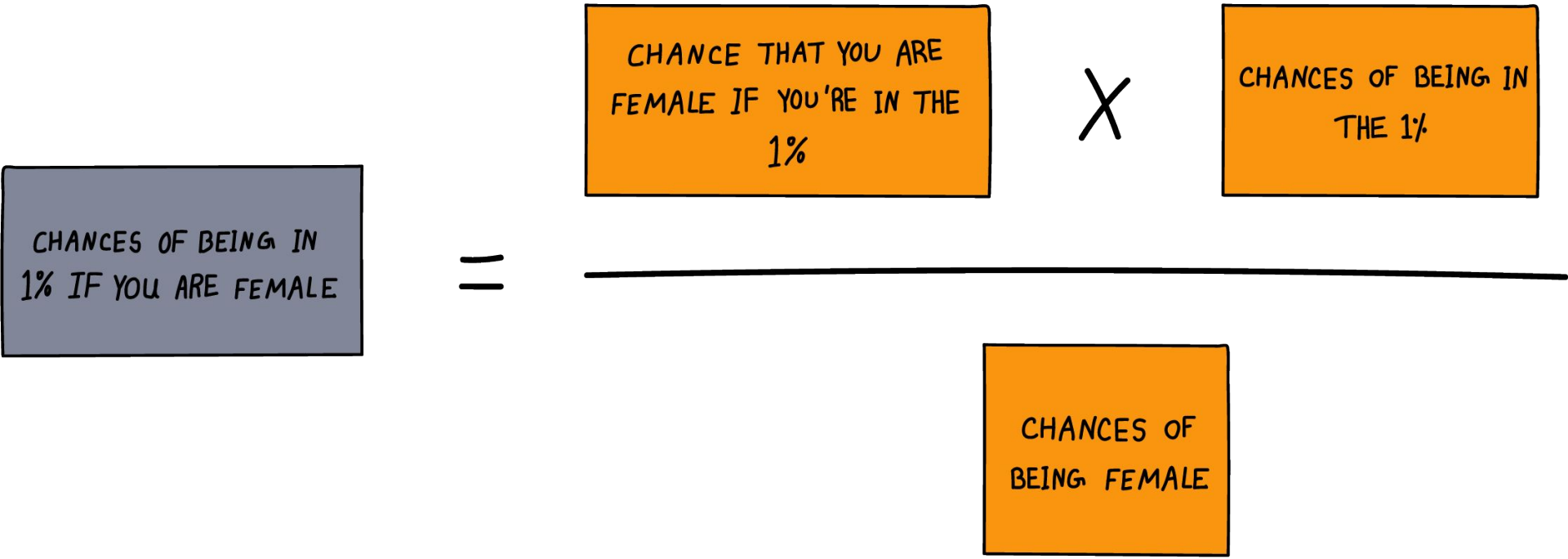
Bayes Rule" is just a formalization of the logic I just explained.

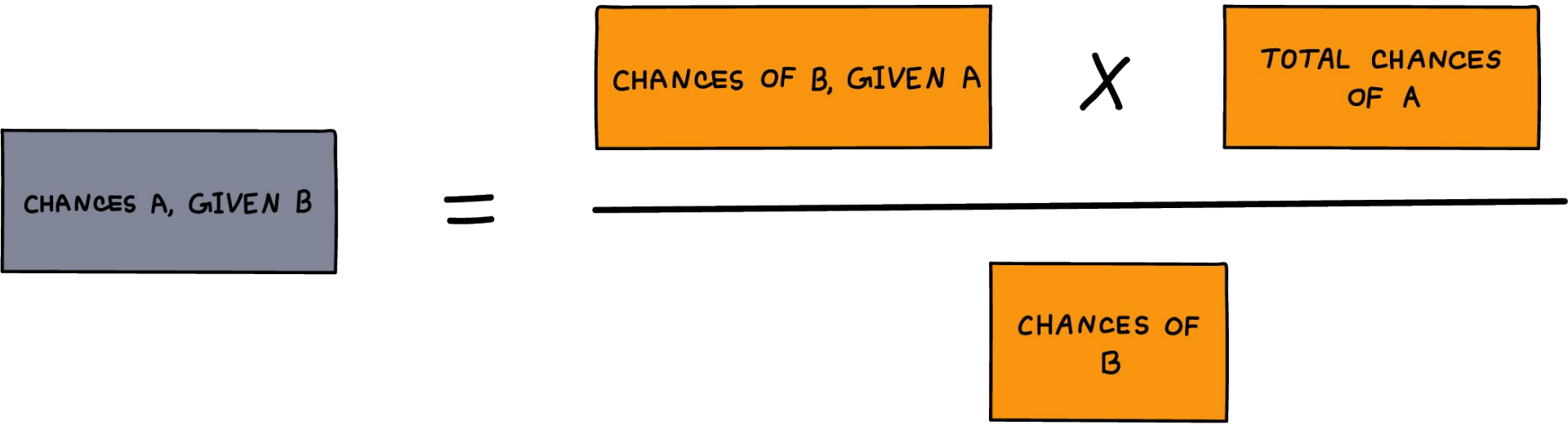
CHANCES OF BEING IN THE 1% IF YOU ARE FEMALE

$$= \frac{\text{CHANCES OF BEING FEMALE AND IN THE 1\%}}{\text{CHANCES OF BEING FEMALE}}$$

$$= \quad x = .0025 / 0.5$$

$$= \quad x = \frac{(.25 * .01)}{0.5}$$


CHANCES OF BEING IN 1% IF YOU ARE FEMALE
=
CHANCE THAT YOU ARE FEMALE IF YOU'RE IN THE 1%
X
CHANCES OF BEING IN THE 1%
CHANCES OF BEING FEMALE


CHANCES A, GIVEN B
=
CHANCES OF B, GIVEN A
X
TOTAL CHANCES OF A
CHANCES OF B

Bayes Theorem

Likelihood

How probable is the evidence given that our hypothesis is true?

Prior

How probable was our hypothesis before observing the evidence?

$$P(H \mid e) = \frac{P(e \mid H)\, P(H)}{P(e)}$$

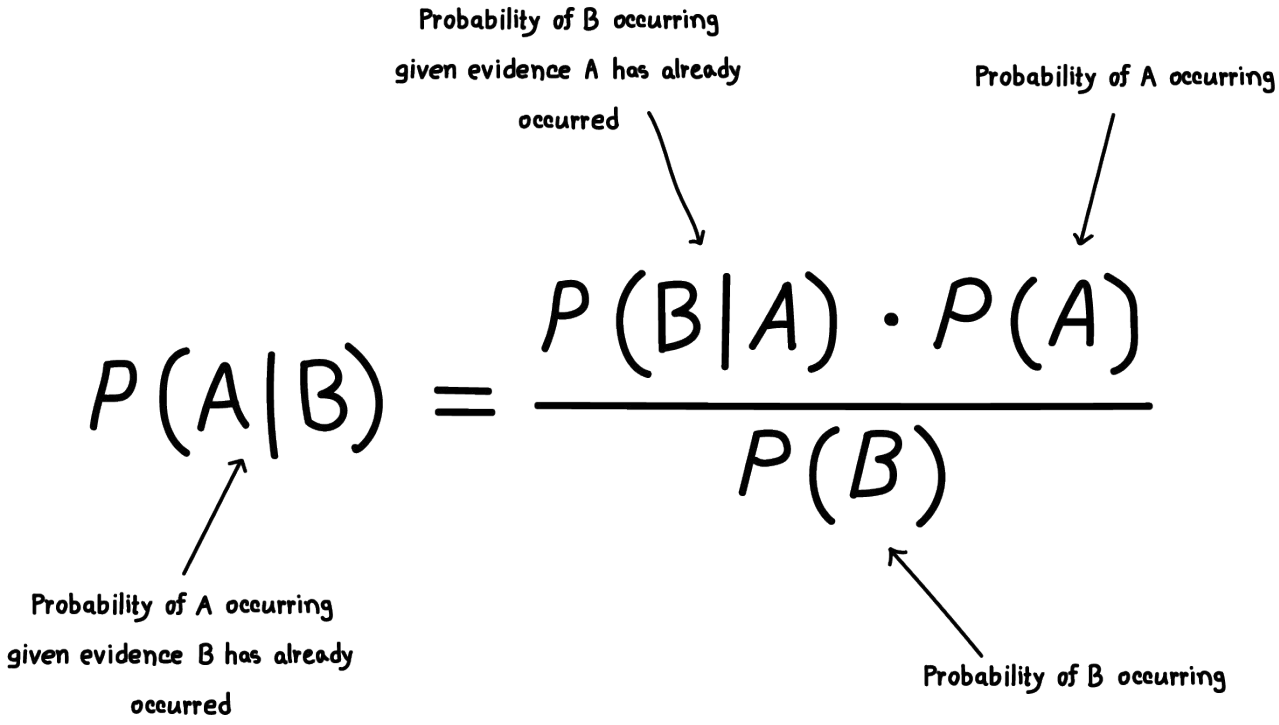
Posterior

How probable is our hypothesis given the observed evidence?
(Not directly computable)

Marginal

How probable is the new evidence under all possible hypothesis?
$P(e) = \sum P(e \mid H)\, P(H)$

Use Bayes to discover the chances that you are in the 1% IF you are male



$$p(A|B) = \frac{(0.75) \times (.01)}{0.5}$$

Bayes' Theorem

Bayes' Theorem is a rule (and formula) in probability theory that can help you assess the probability of an event happening given prior knowledge about conditions related to that event.

Mathematically, it looks like this:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Note: $P(B)$ must not be zero

"P" means probability.

$P(A)$ means the probability of event A happening independently → whether or not event B happens.

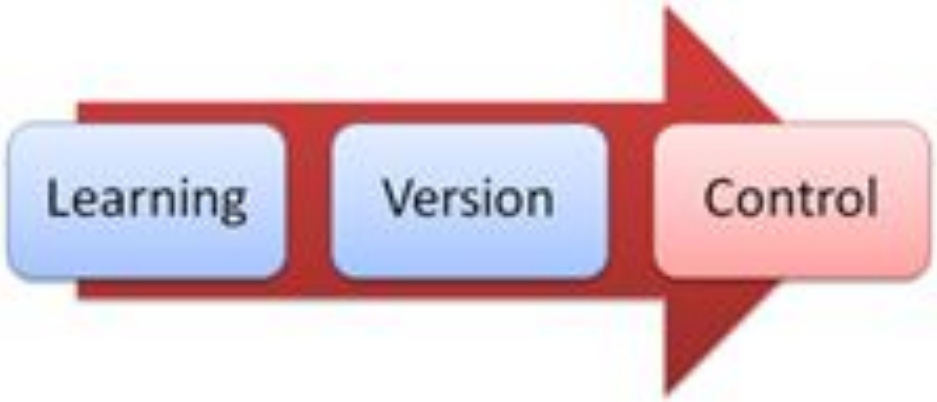
$P(B)$ means the same for event B

$P(A|B)$ means the probability of event A happening given that event B does happen

$P(B|A)$ is the inverse; it's the probability of event B happening given that event A happens.

By taking the probability of event B into consideration, you can come to a more accurate conclusion about the probability of event A happening

Software Version Control



So Why Do We Need Version Control?

- **Backup and Restore.**
- **Synchronization.**
- **Short-term undo.**
- **Long-term undo.**
- **Track Changes.**
- **Track Ownership.**
- **Sandboxing**
- **Branching and merging.**

Half of BEING good… is SOUNDING good...

Basic Setup

- **Repository (repo)**: The database storing the files.
- **Server**: The computer storing the repo.
- **Client**: The computer connecting to the repo.
- **Working Set/Working Copy**: Your local directory of files, where you make changes.
- **Trunk/Main**: The primary location for code in the repo. Think of code as a family tree — the trunk is the main line.

You gotta fake it… to make it...

Basic Actions

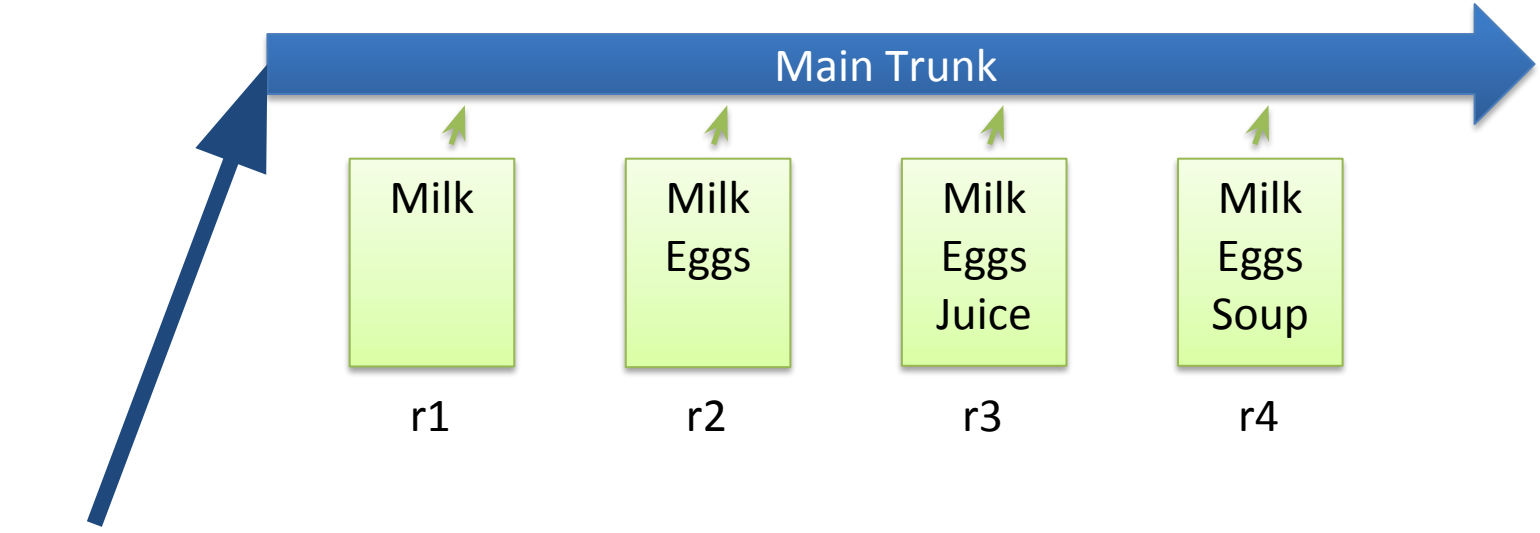
- **Add**: Put a file into the repo for the first time, i.e. begin tracking it with Version Control.
- **Revision**: What version a file is on (v1, v2, v3, etc.).
- **Head**: The latest revision in the repo.
- **Check out**: Download a file from the repo.
- **Check in**: Upload a file to the repository (if it has changed). The file gets a new revision number, and people can “check out” the latest one.
- **Checkin Message**: A short message describing what was changed.
- **Changelog/History**: A list of changes made to a file since it was created.
- **Update/Sync**: Synchronize your files with the latest from the repository. This lets you grab the latest revisions of all files.
- **Revert**: Throw away your local changes and reload the latest version from the repository.

Just Learn the Lingo...

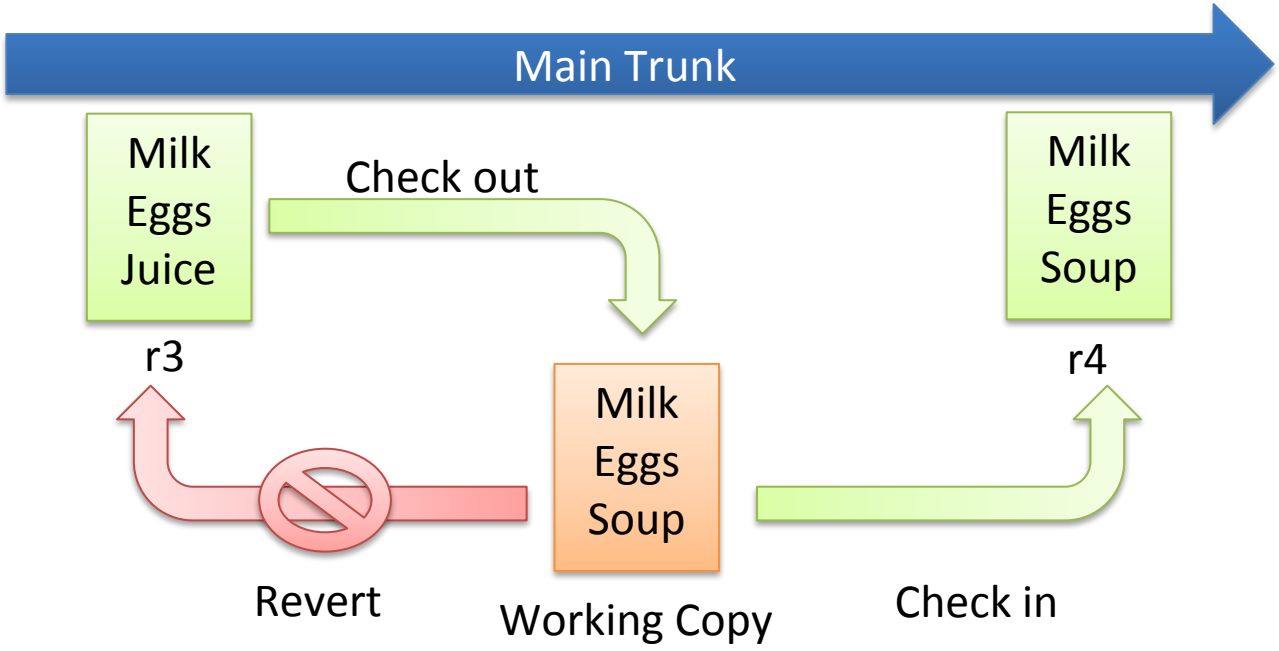
Advanced Actions

- **Branch**: Create a separate copy of a file/folder for private use (bug fixing, testing, etc). Branch is both a verb (“branch the code”) and a noun (“Which branch is it in?”).
- **Diff/Change/Delta**: Finding the differences between two files. Useful for seeing what changed between revisions.
- **Merge (or patch)**: Apply the changes from one file to another, to bring it up-to-date. For example, you can merge features from one branch into another. (At Microsoft this was called Reverse Integrate and Forward Integrate)
- **Conflict**: When pending changes to a file contradict each other (both changes cannot be applied).
- **Resolve**: Fixing the changes that contradict each other and checking in the correct version.
- **Locking**: Taking control of a file so nobody else can edit it until you unlock it. Some version control systems use this to avoid conflicts.
- **Breaking the lock**: Forcibly unlocking a file so you can edit it. It may be needed if someone locks a file and goes on vacation (or “calls in sick” the day Halo 3 comes out).
- **Check out for edit**: Checking out an “editable” version of a file. Some VCSes have editable files by default, others require an explicit command.

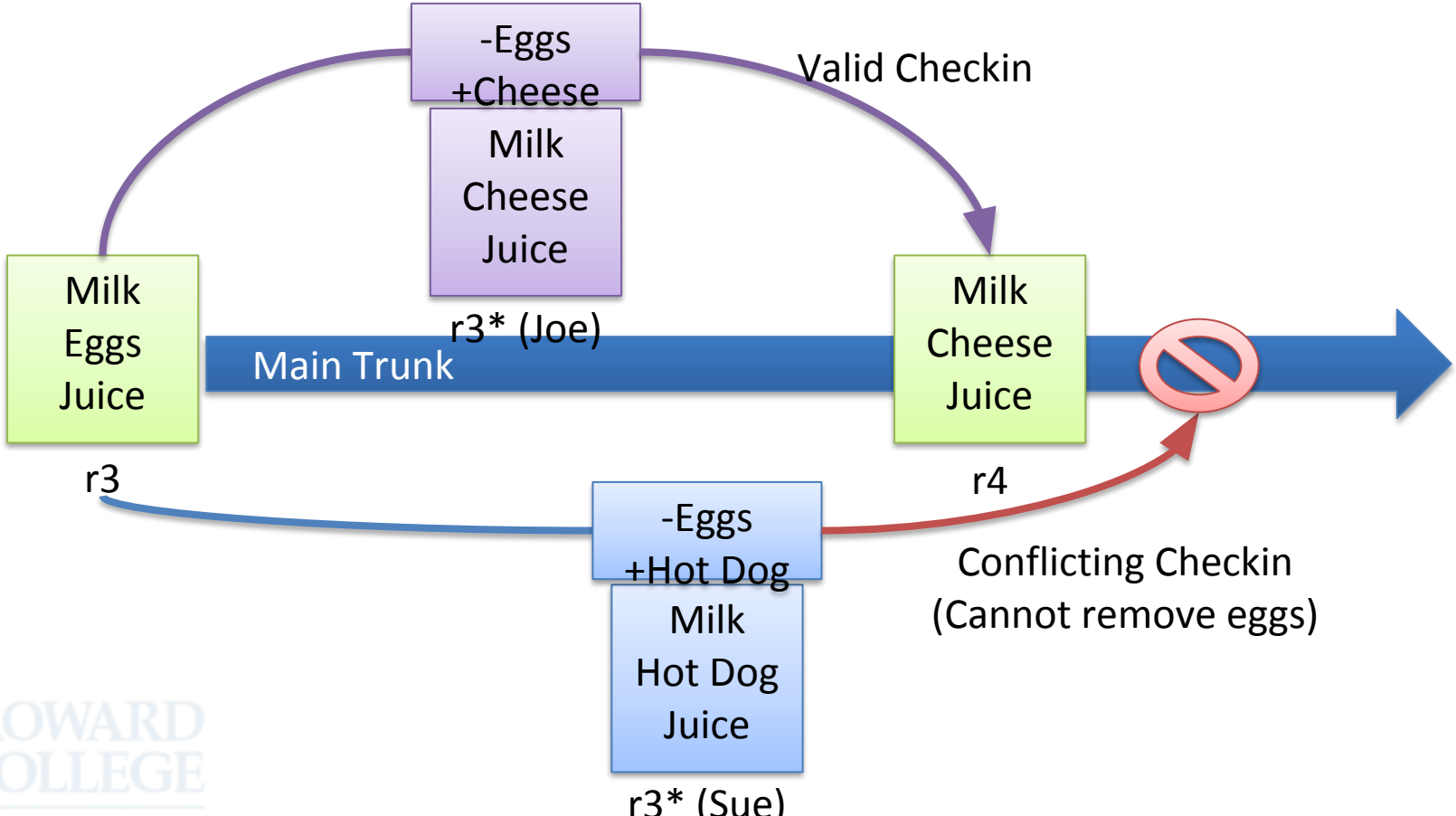
Basic Checkins



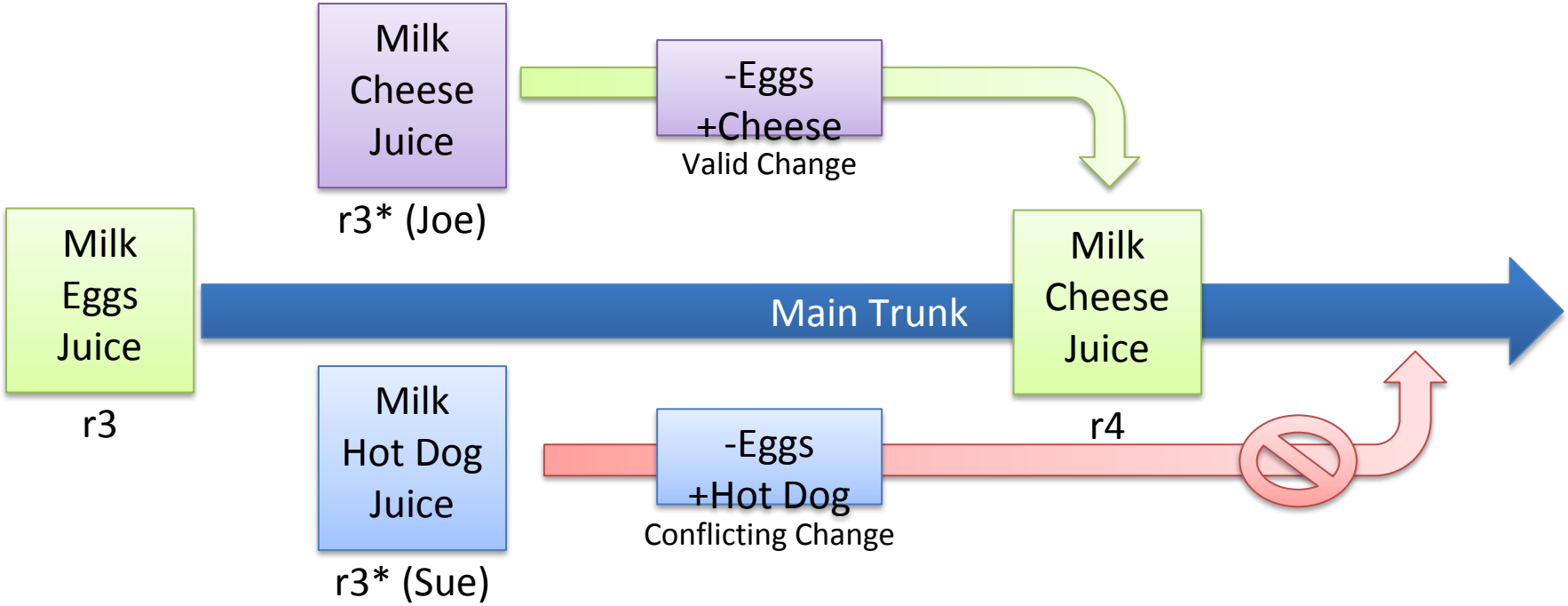
Checkout and Edit



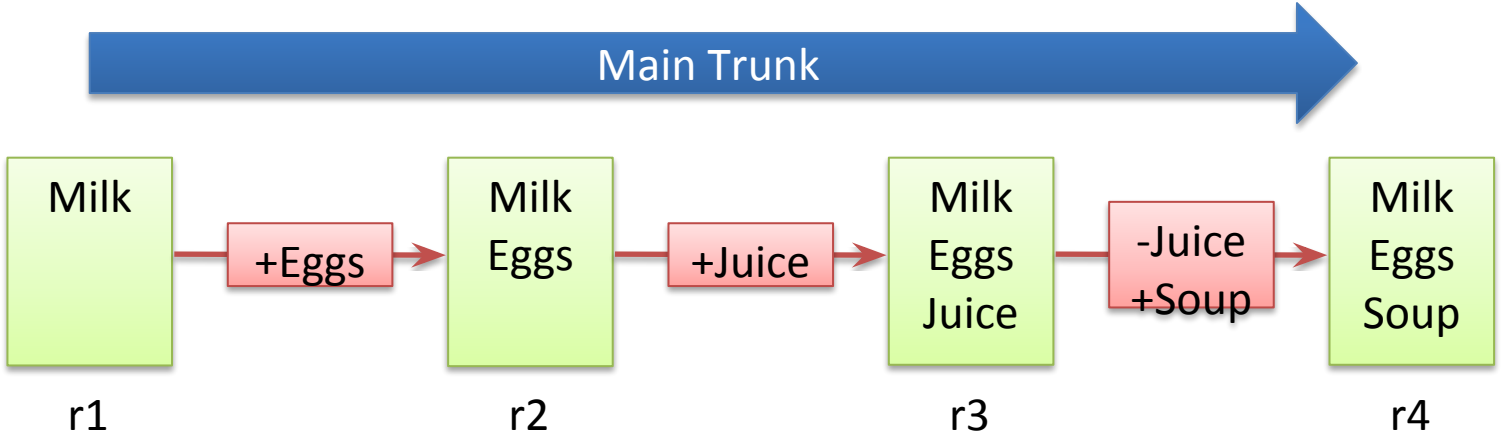
Conflicts



Conflicts (Alt)

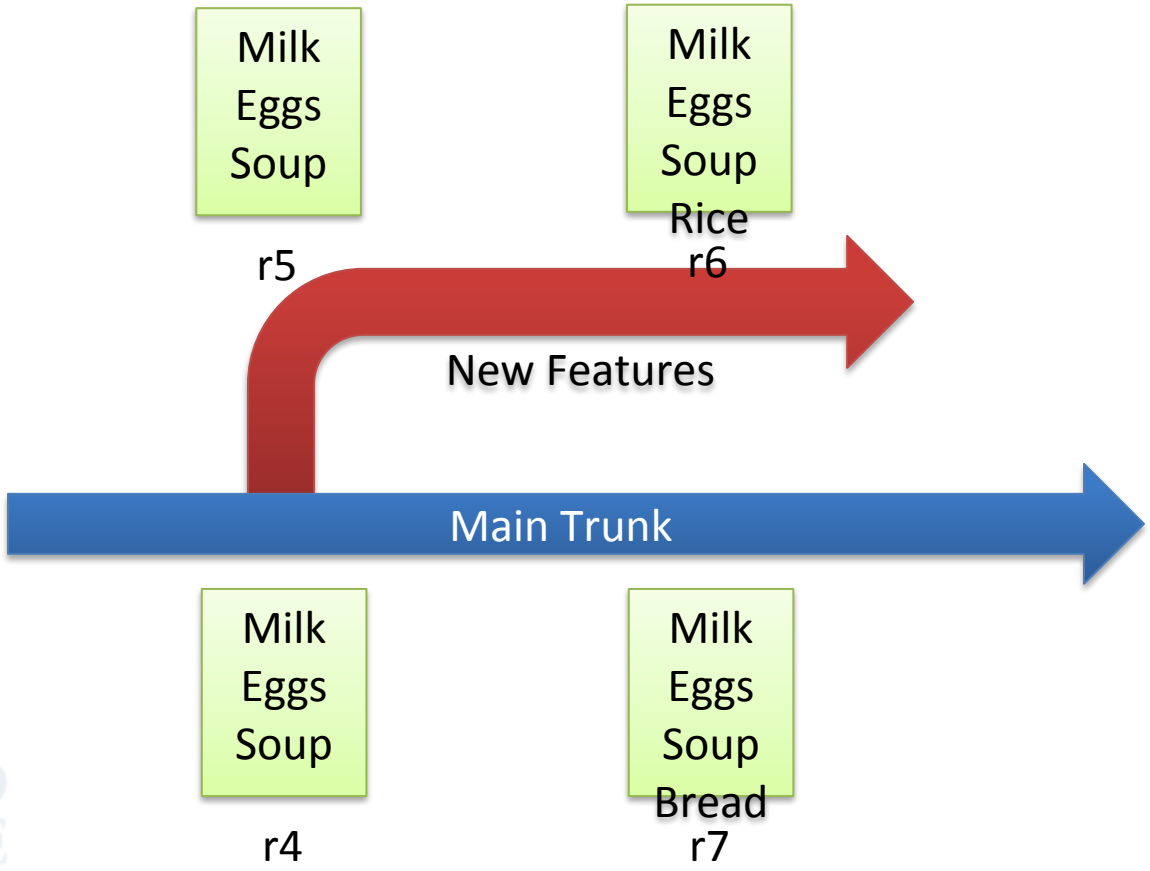


Basic Diffs

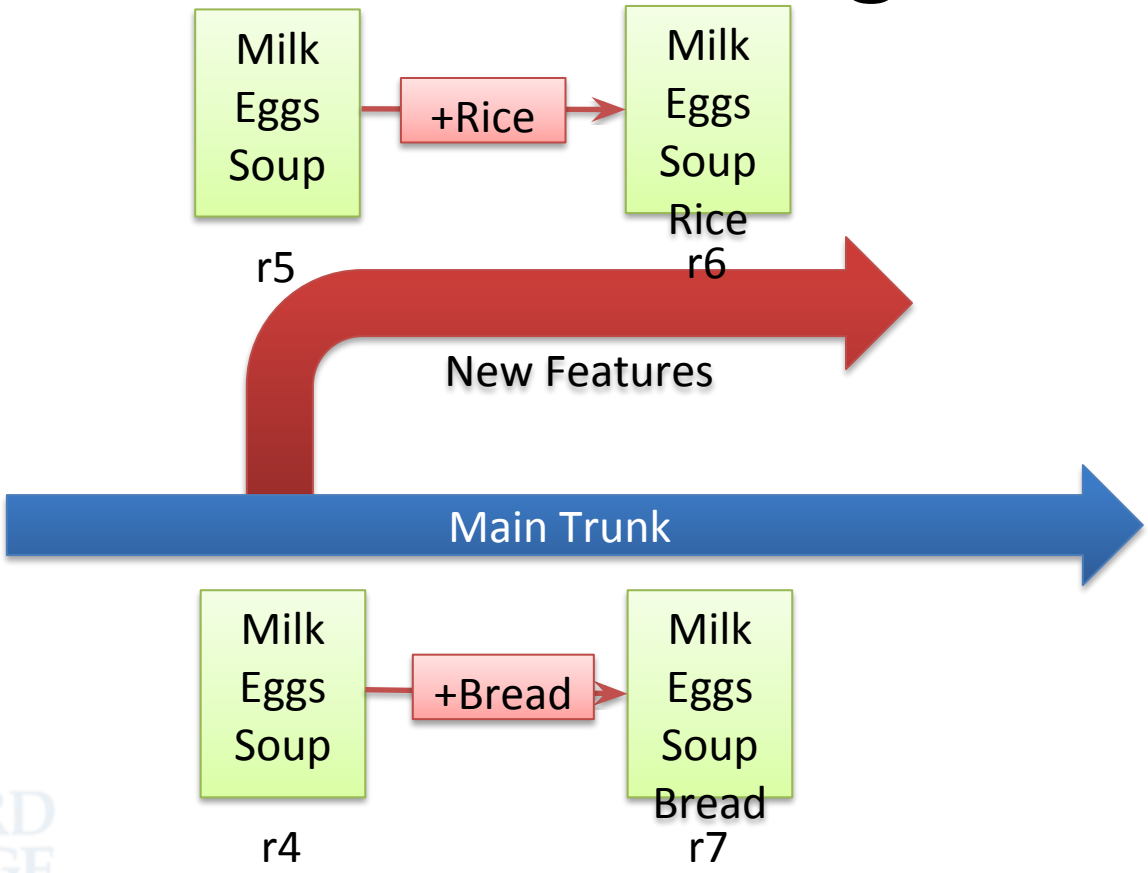


NOTE AVAILABLE IN GIT… ONLY IN SVN

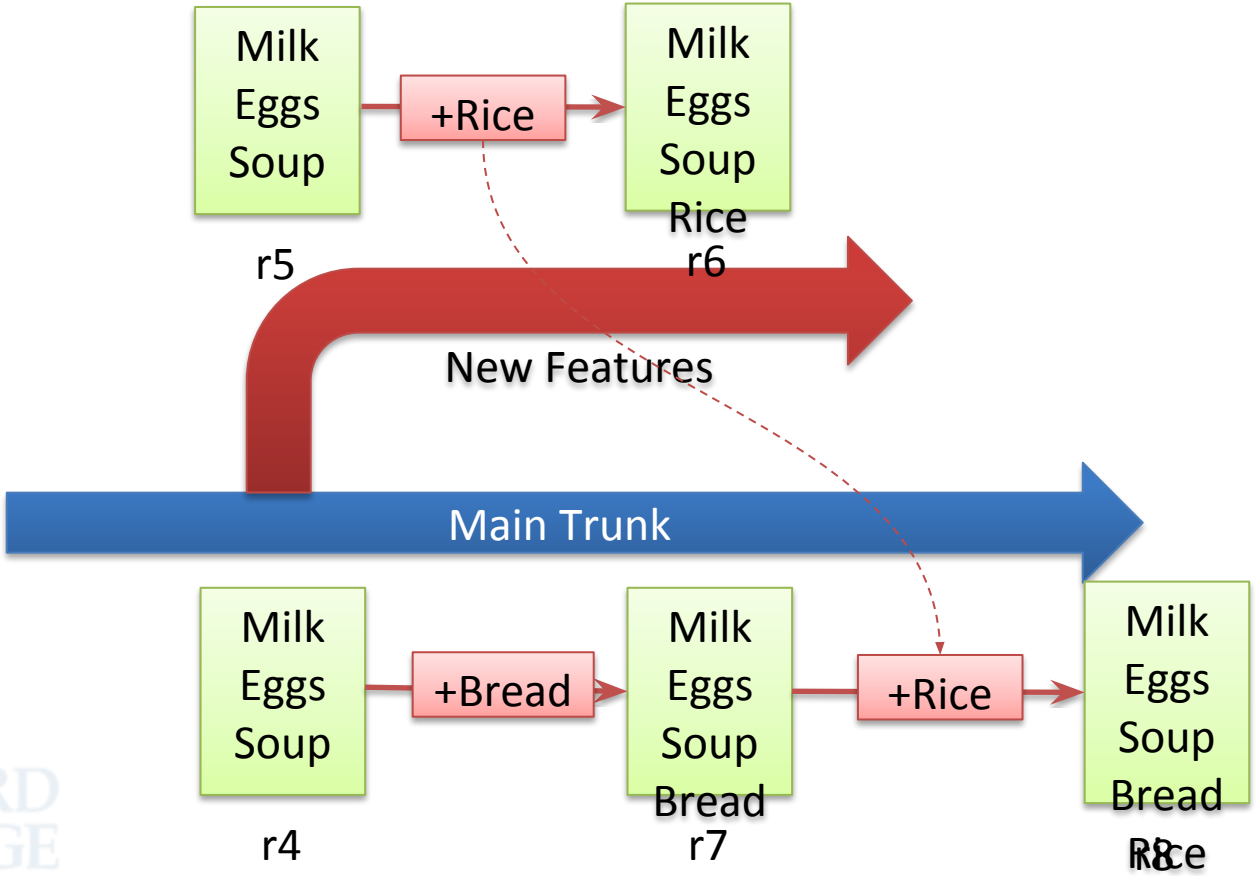
Branching



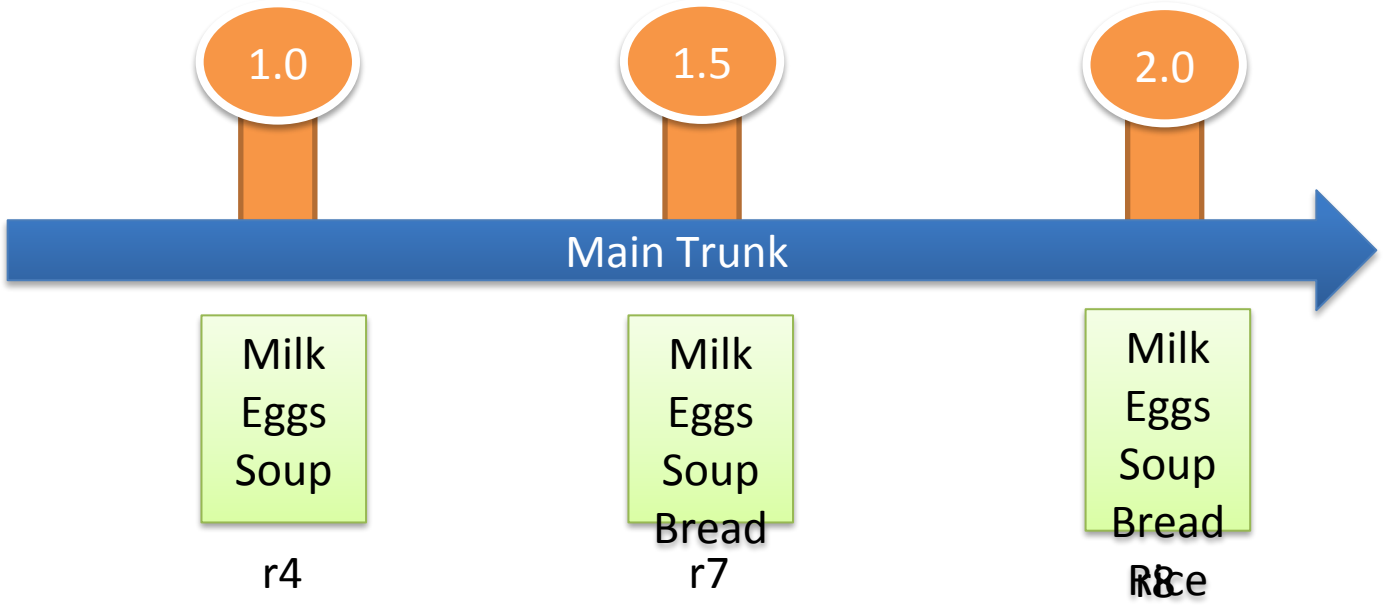
Branch Changes



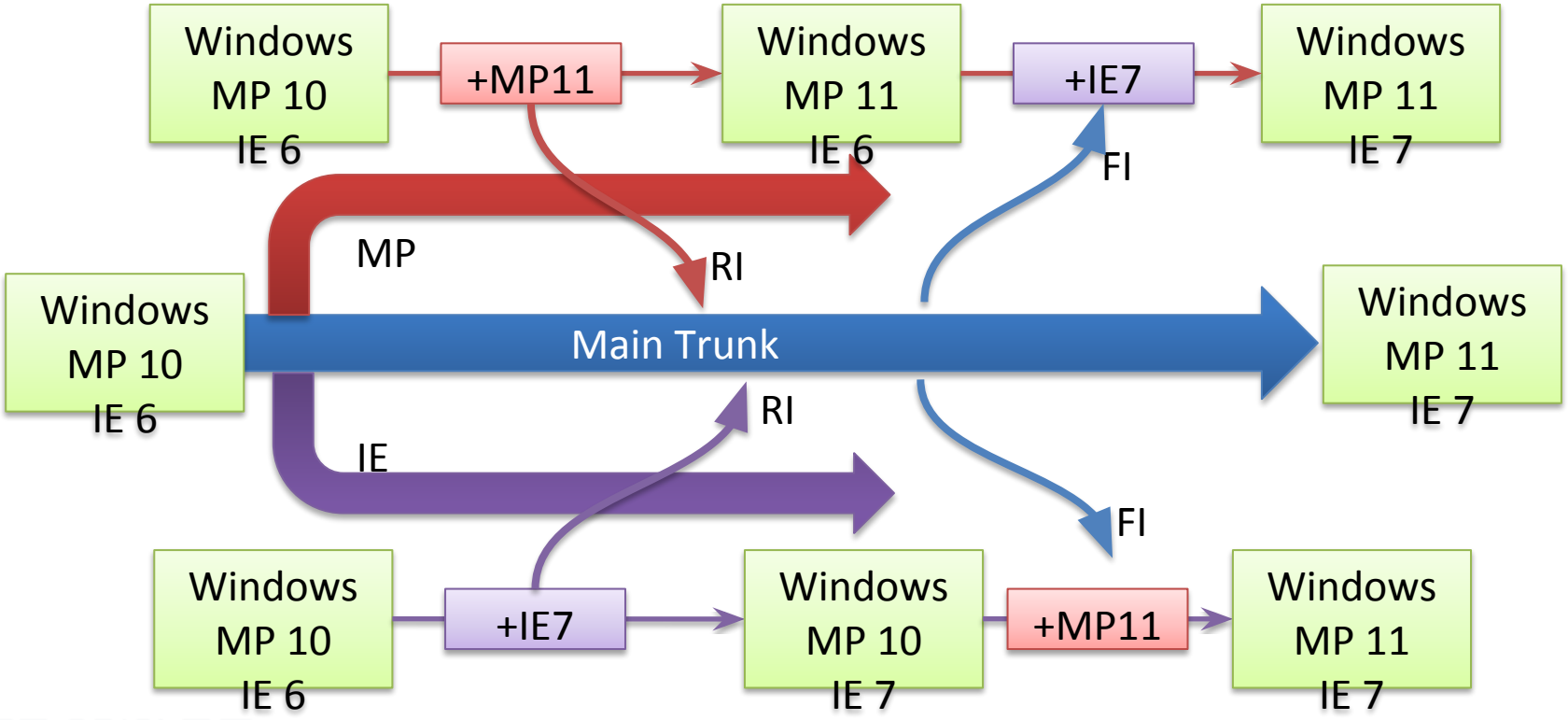
Merging



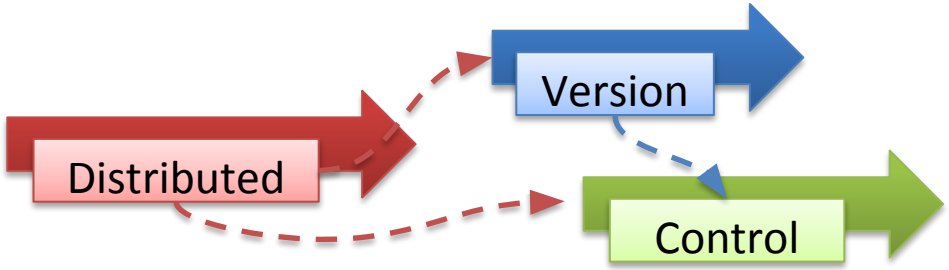
Tagging



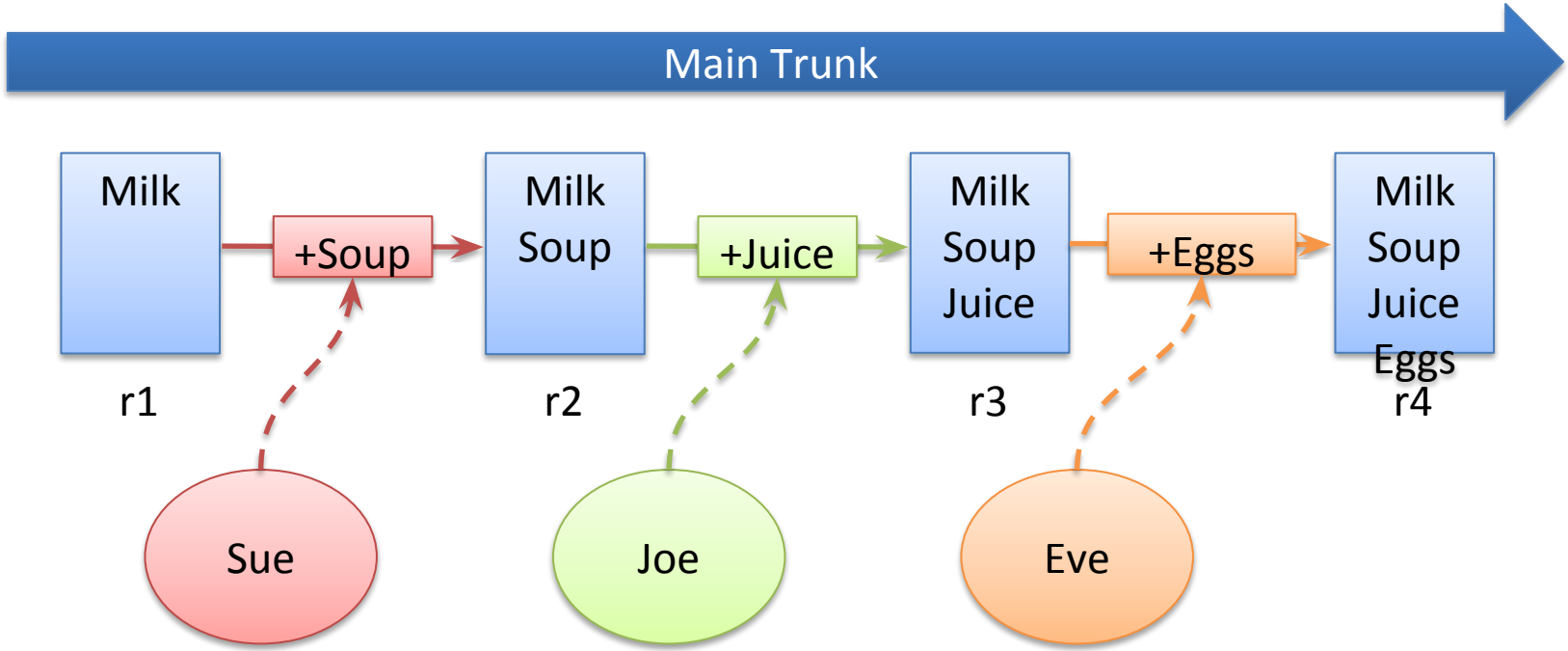
Managing Windows



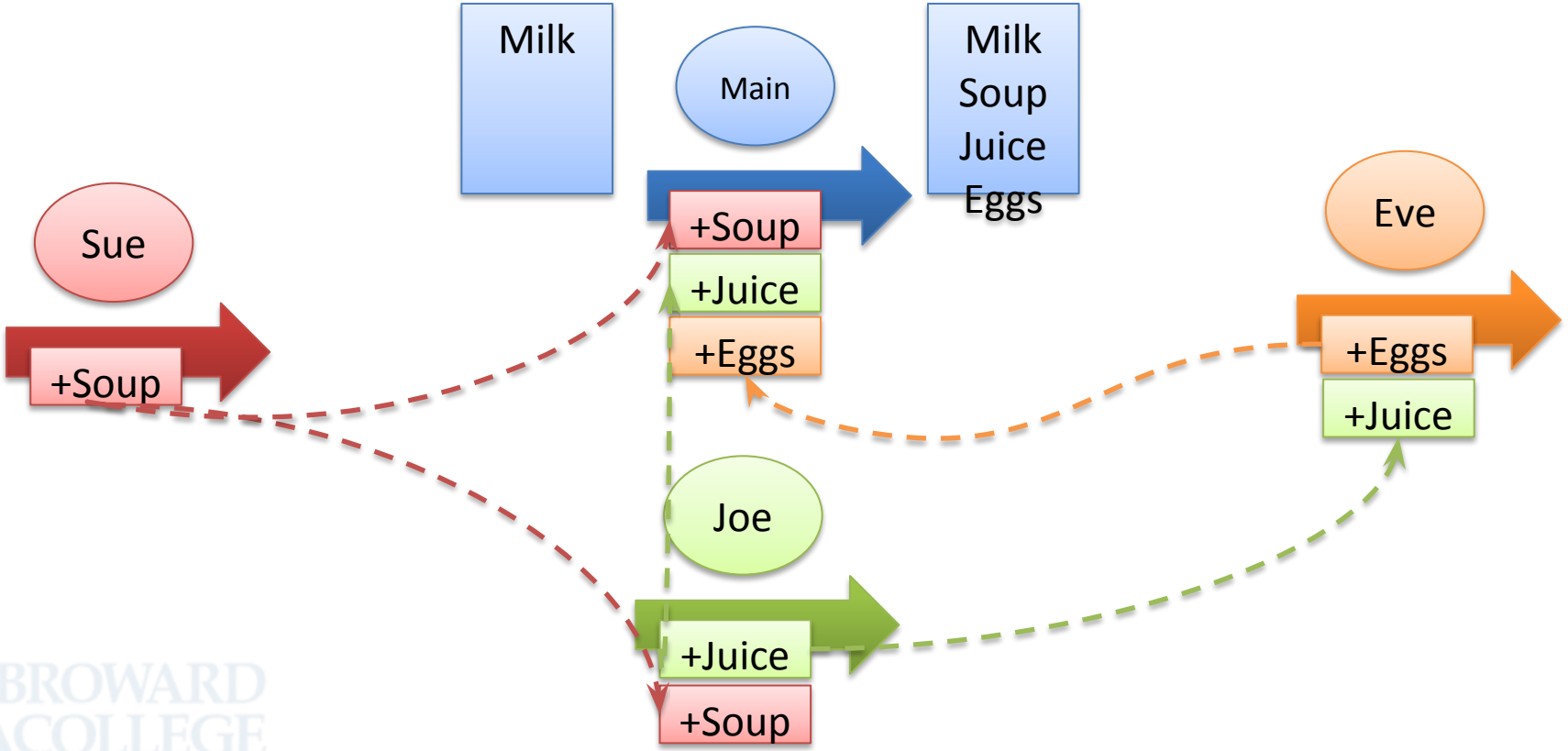
Distributed Version Control



Centralized VCS



Distributed VCS



Core Concepts

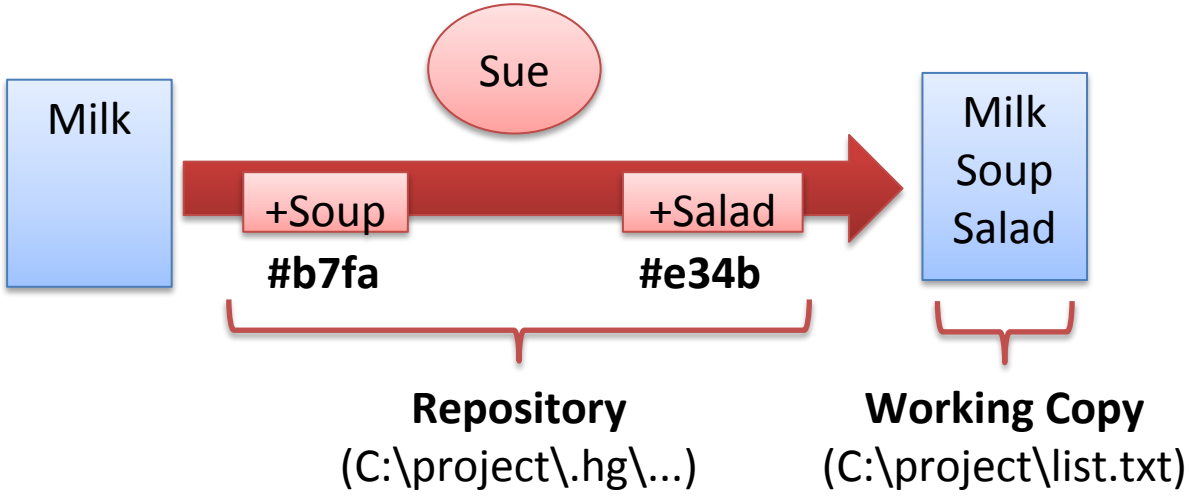
Core Concepts

- Centralized version control focuses on **synchronizing, tracking, and backing up files.**
- Distributed version control focuses on **sharing changes**; every change has a guid or unique id.
- **Recording/Downloading** and **applying** a change are separate steps (in a centralized system, they happen together).
- **Distributed systems have no forced structure**. You can create “centrally administered” locations or keep everyone as peers.

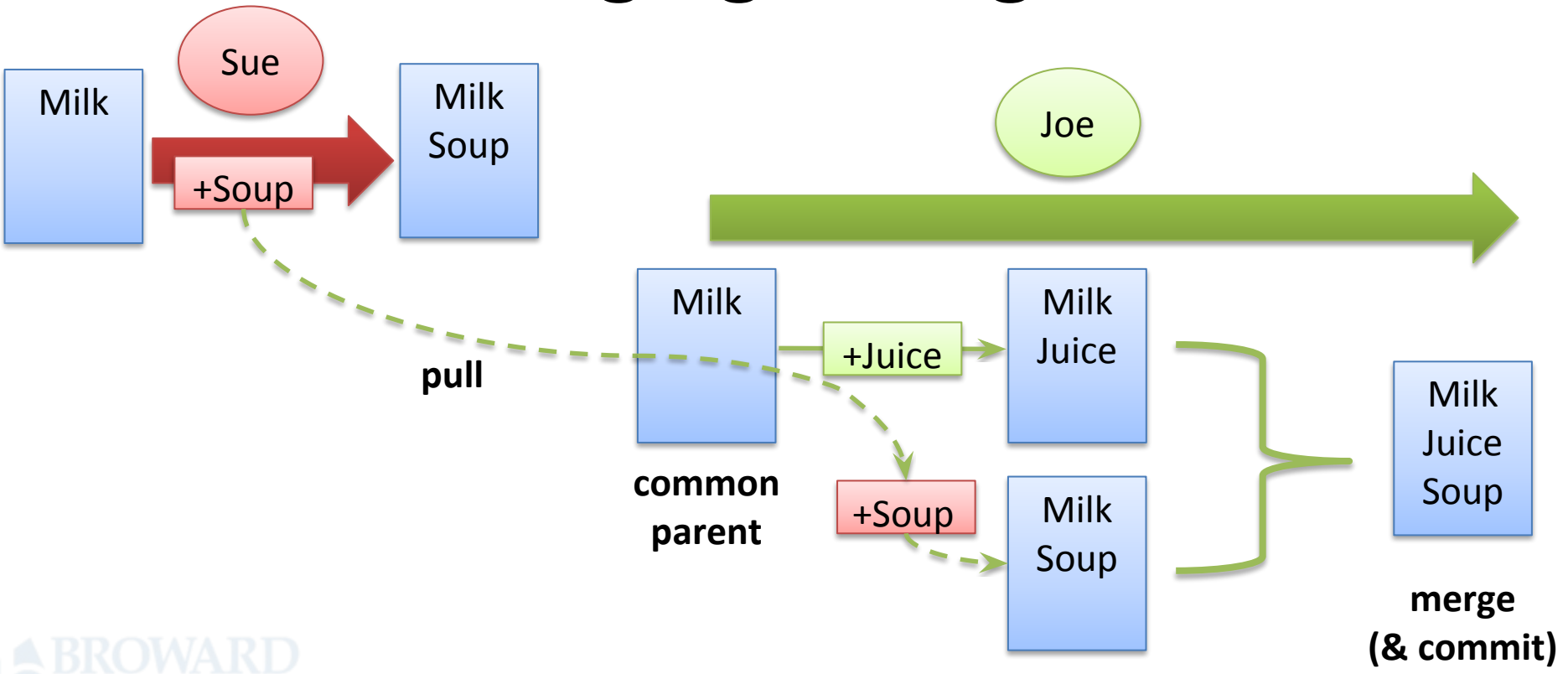
New Terminology

- **push**: send a change to another repository (may require permission)
- **pull**: grab a change from a repository

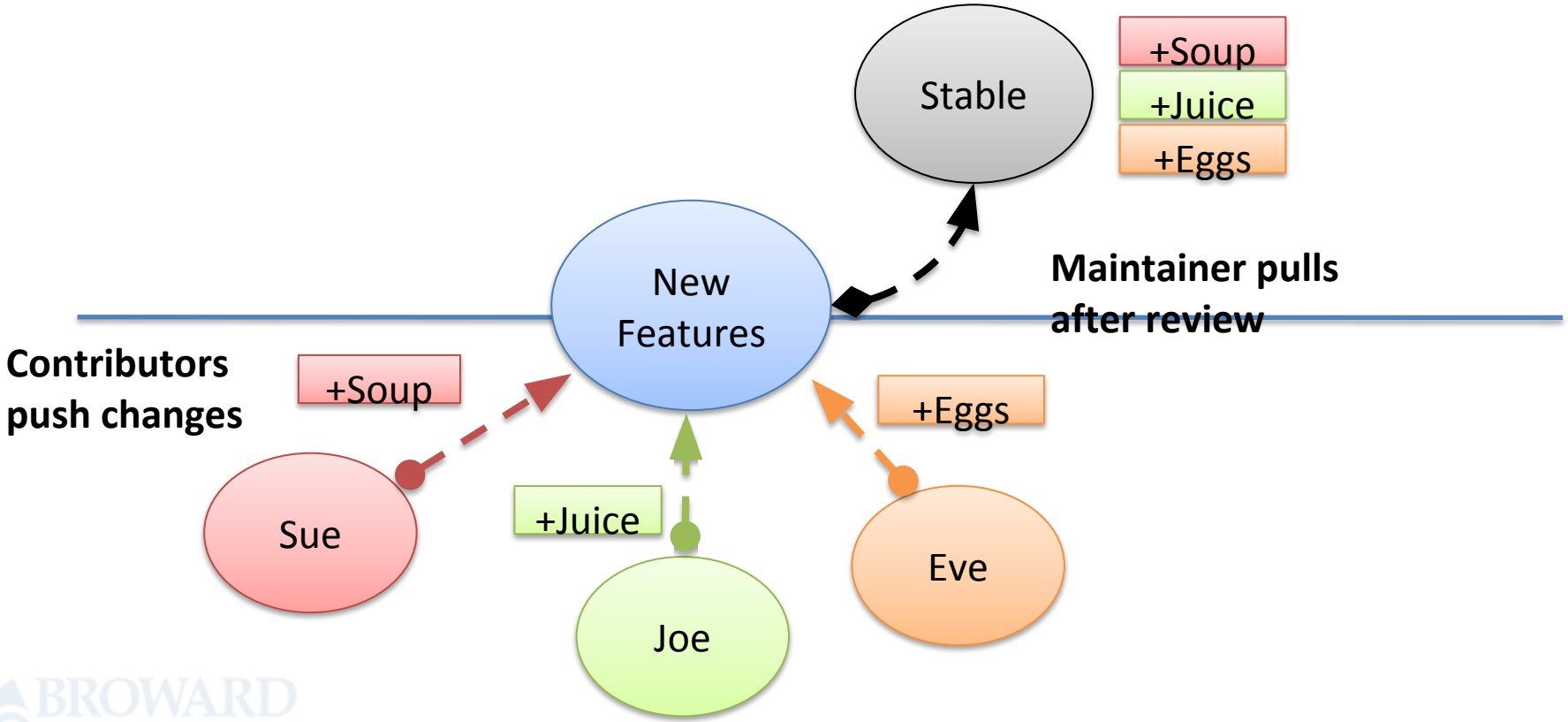
Repository Layout



Merging Changes



Distributed Push/Pull Model



What is the Slingshot?

- **Use version control.**
- **Take it slow.**
- **Keep Learning.**

DEMO

Create your GitHub Account

Create your first Repository

Create your first GitHub WebSite (Bonus)

Git “Aha’s”

Git has a staging area. **Git has a staging area!!!**

Did this ever confuse me. There's both a repo ("object database") and a staging area (called "index"). Checkins have two steps:

- `git add foo.txt`
  - Add foo.txt to the index. It's not checked in yet!
- `git commit -m "message"`
  - Put staged files in the repo; they're now tracked
  - You can "`git add --update`" to stage all tracked, modified files

Branching is a “Save As”

Branches are like "Save as..." on a directory. Best of all:

- Easily merge changes with the original (changes tracked and never applied twice)
- No wasted space (common files only stored once)

Imagine Virtual Directories

I see branches as "virtual directories" in the .git folder. While inside a physical directory (c:\project or ~/project), you traverse virtual directories with a checkout.

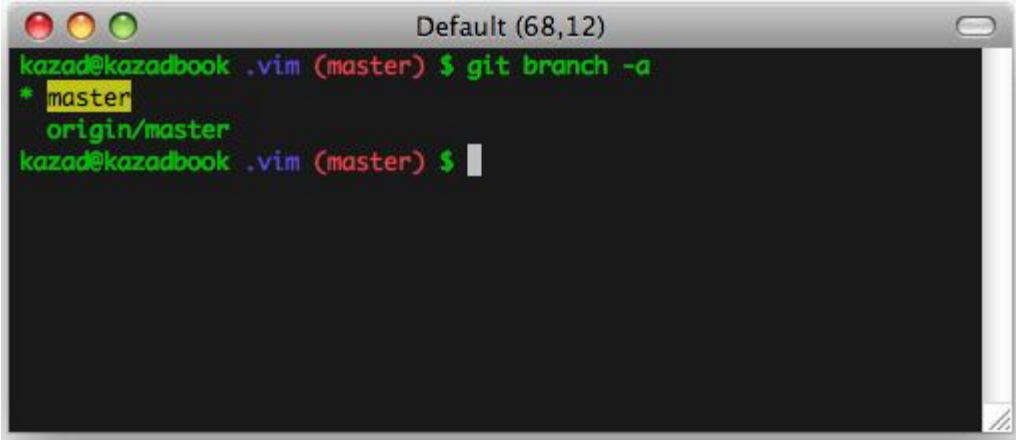
- git checkout master
  - switch to master branch ("cd master")
- git branch dev
  - create new branch from existing ("cp * dev")
  - you still need to "cd" with "git checkout dev"
- git merge dev
  - (when in master) pull in changes from dev ("cp dev/* .")
- git branch
  - list all branches ("ls")

Know the Current Branch

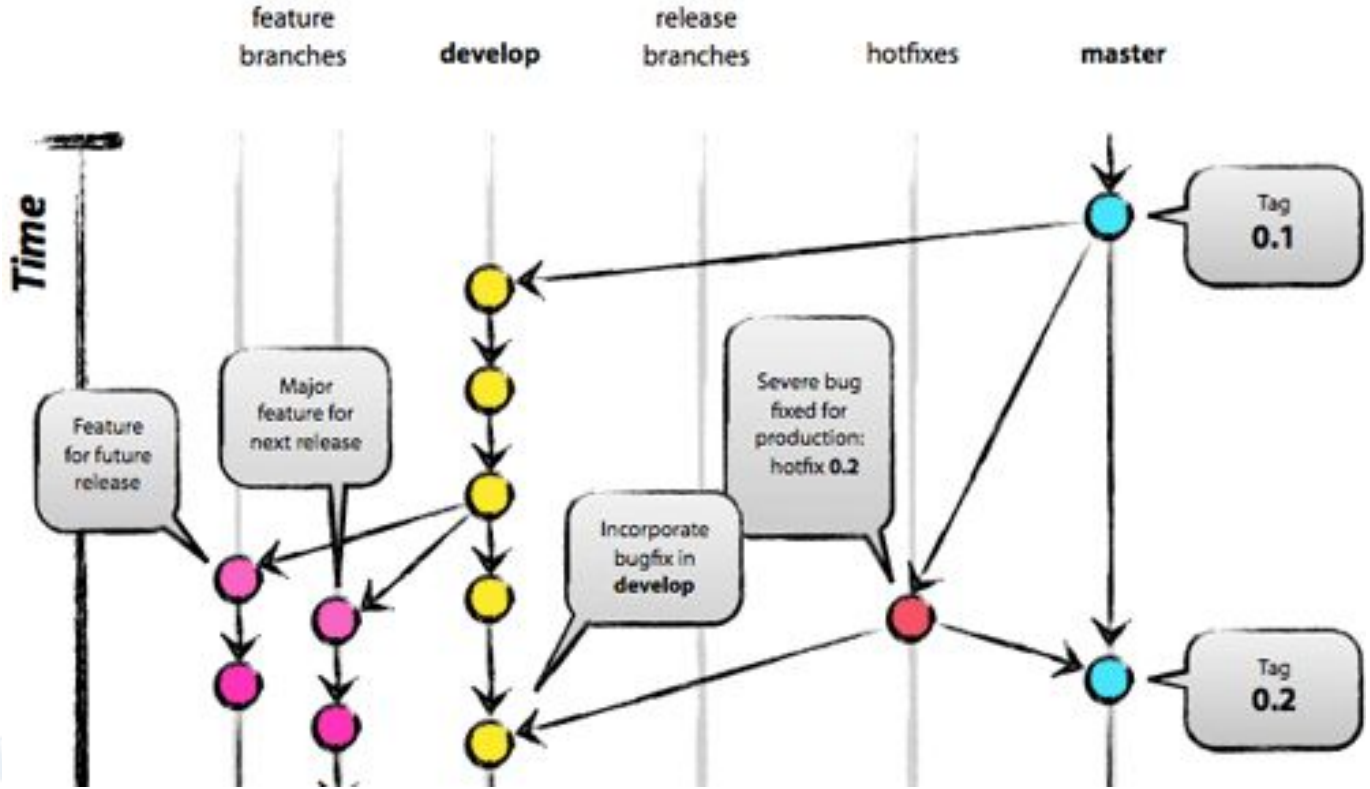
In my .bash_profile:

```
parse_git_branch() {
    git branch 2> /dev/null | sed -e '/^[^*]/d' -e 's/* (.*)/(1)/'
}

export PS1="[33[00m]u@h[33[01;34m] W [33[31m]$(parse_git_branch)
[33[00m]$[33[00m] "
```



Always Visualize your Branches



Git has Local AND Remote

Local data

- git init
  - create local repo
  - use git add/commit/branch to work locally

Remote data

- git remote add name path-to-repo
  - track a remote repo (usually "origin") from an existing repo
  - remote branches are "origin/master", "origin/dev" etc.
- git branch -a
  - list all branches (remote and local)
- git clone path-to-repo
  - create a new local git repo copied from a remote one
  - local master tracks remote master
- git pull
  - merge changes from tracked remote branch (if in dev, pull from origin/dev)
- git push
  - send changes to tracked remote branch (if in dev, push to origin/dev)

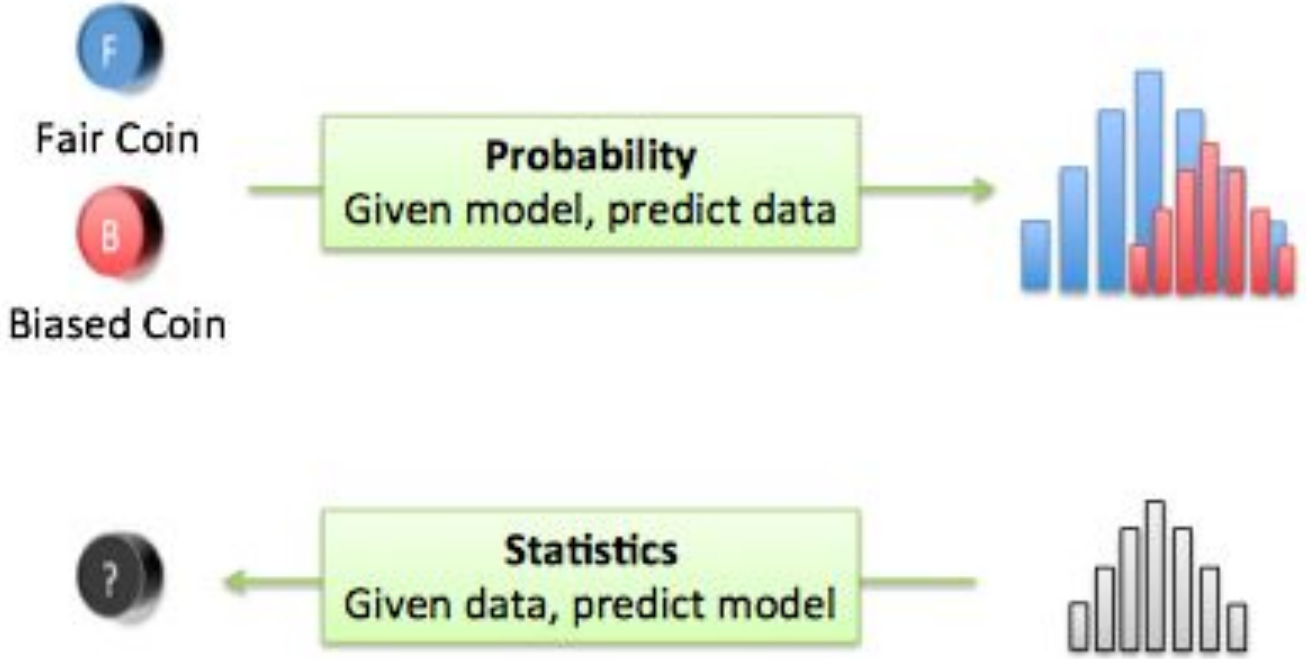
GUIDs

Statistics

Difference Between Probability & Statistics

- **Probability** is starting with an animal, and figuring out what footprints it will make.
- **Statistics** is seeing a footprint, and guessing the animal.

Probability & Statistics



Find the Animal

Get the tracks.

Measure the basic characteristics.

Find the species.

Look up the specific animal.

Make additional predictions.

Slingshot

"Statistics is the study of the collection, organization, analysis, and interpretation of data".

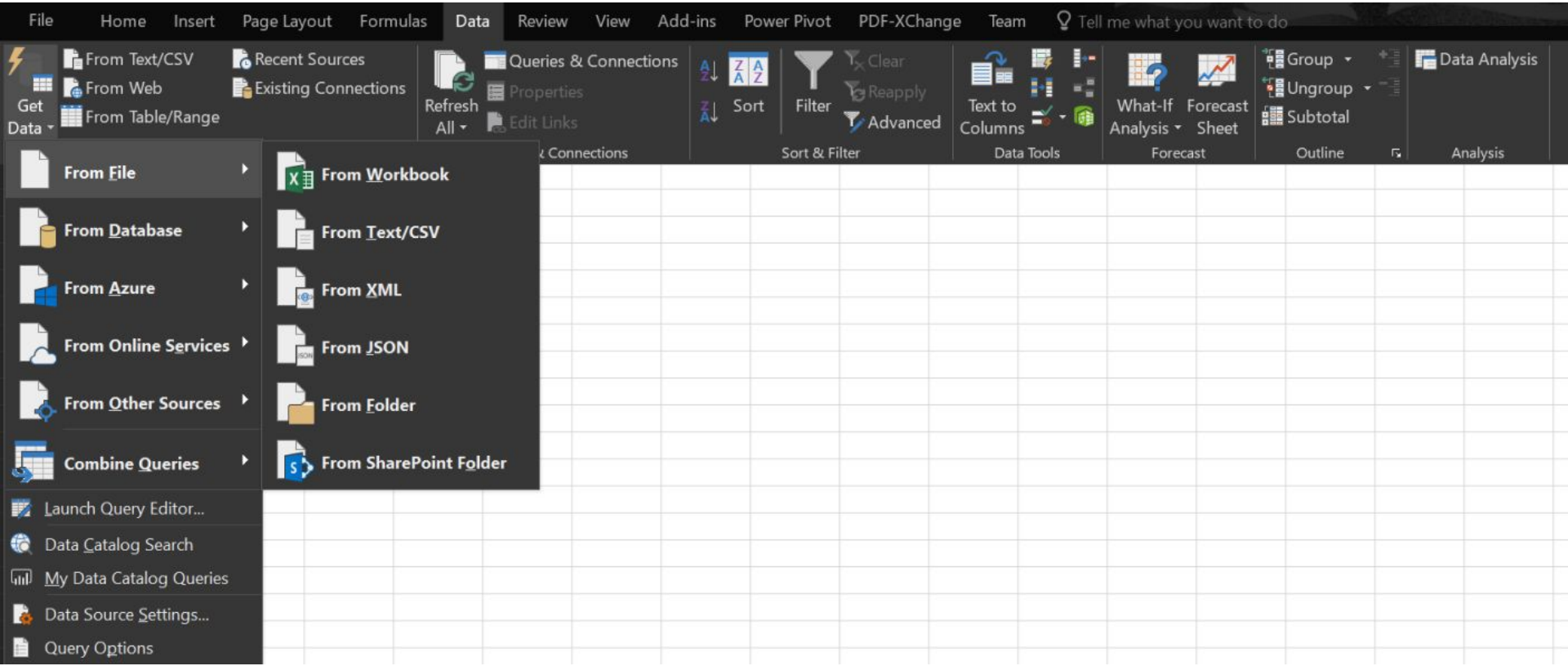
- What are the most common species? (Common distributions)
- Are new ones being discovered?
- Can we predict the next footprint? (Extrapolation)
- Are the tracks following a path? (Regression / trend line)
- Here's two tracks, which animal was faster? Bigger? (Data from two drug trials: which was more effective?)
- Is one animal moving in the same direction as another? (Correlation)
- Are two animals tracking a common source? (Causation: two bears chasing the same rabbit)

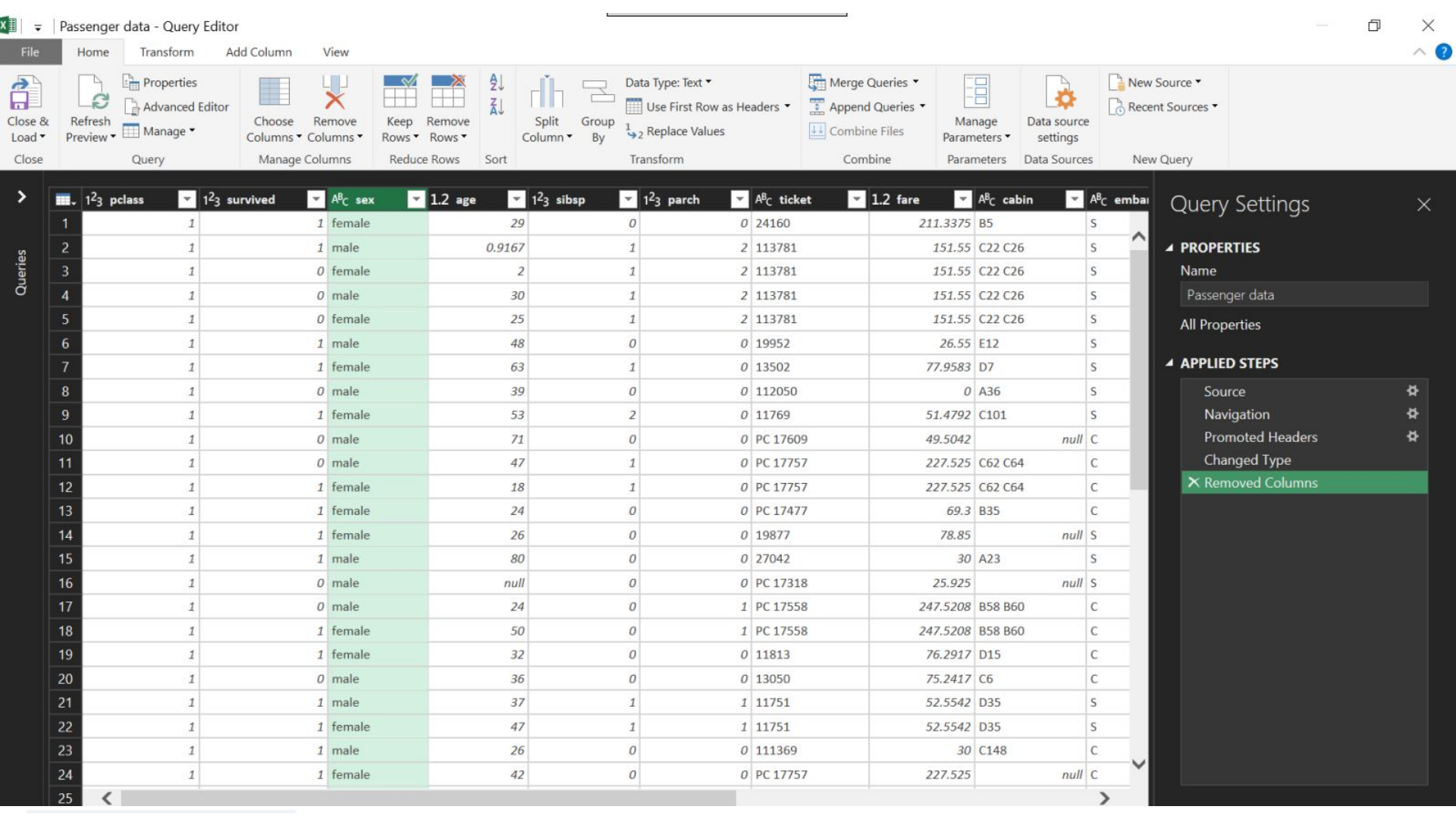
Cleaning Data

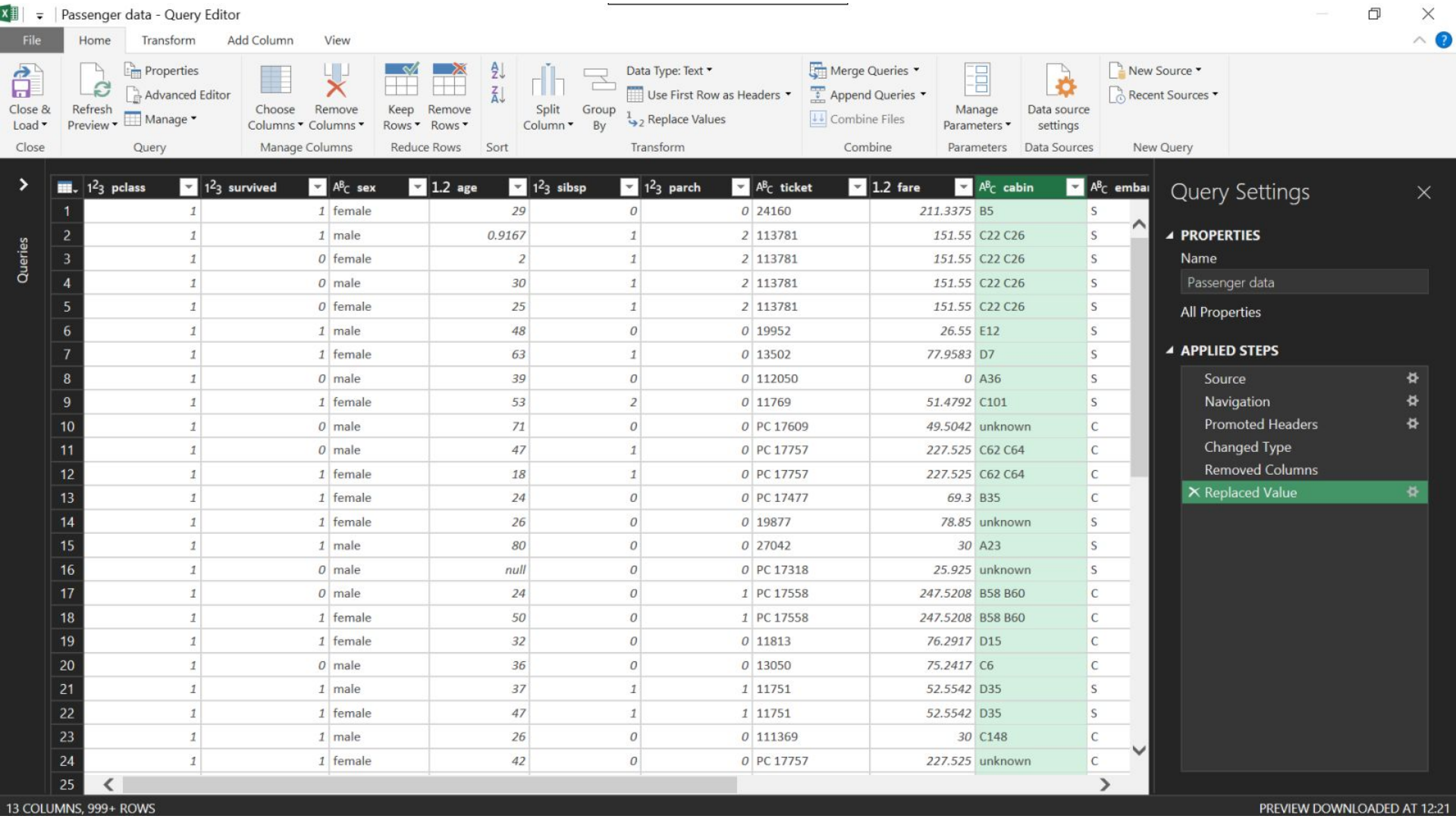
- Cleansing data
- Visualizing data for preliminary analysis
- Understanding unbalanced datasets

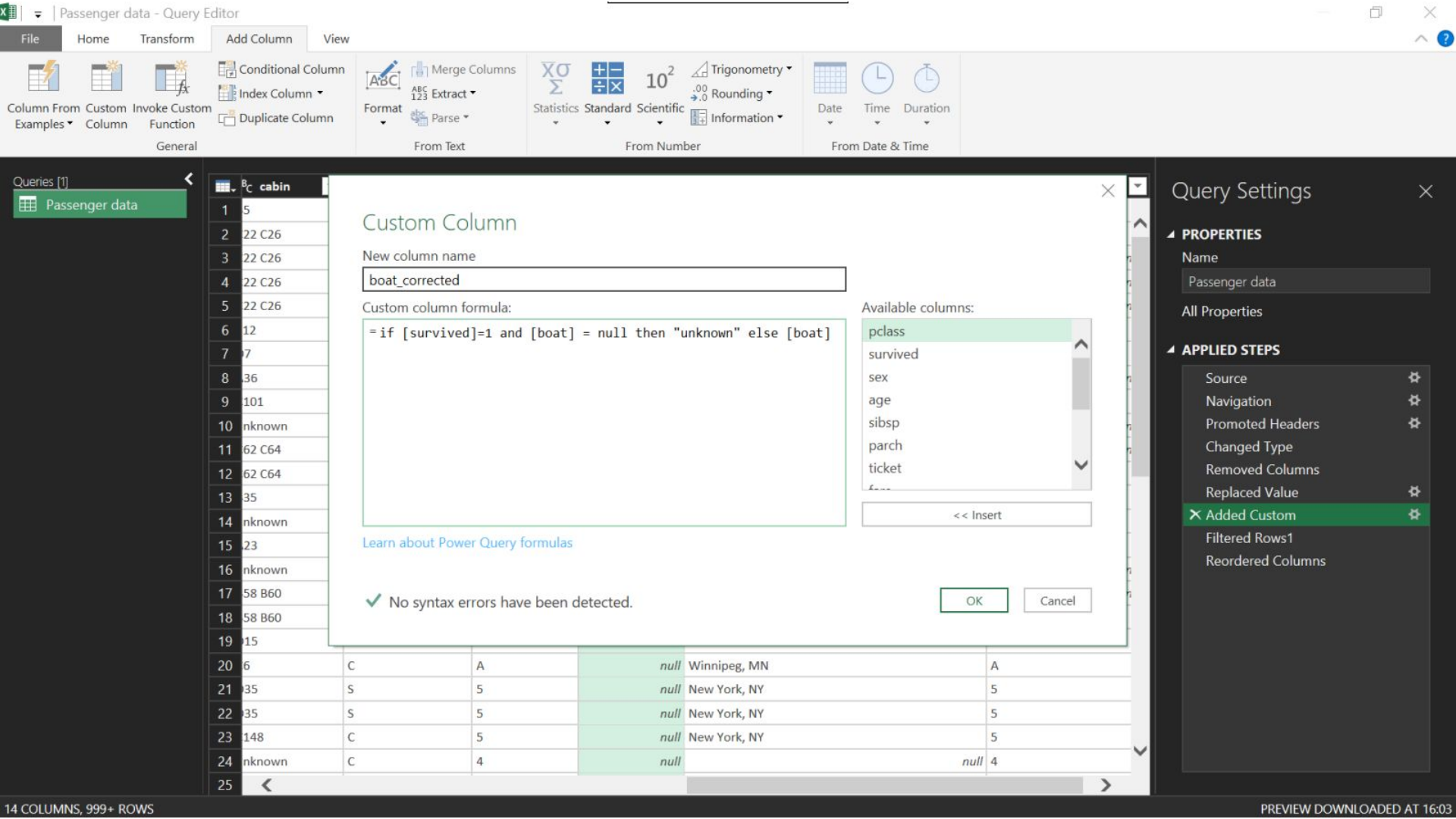
Titanic

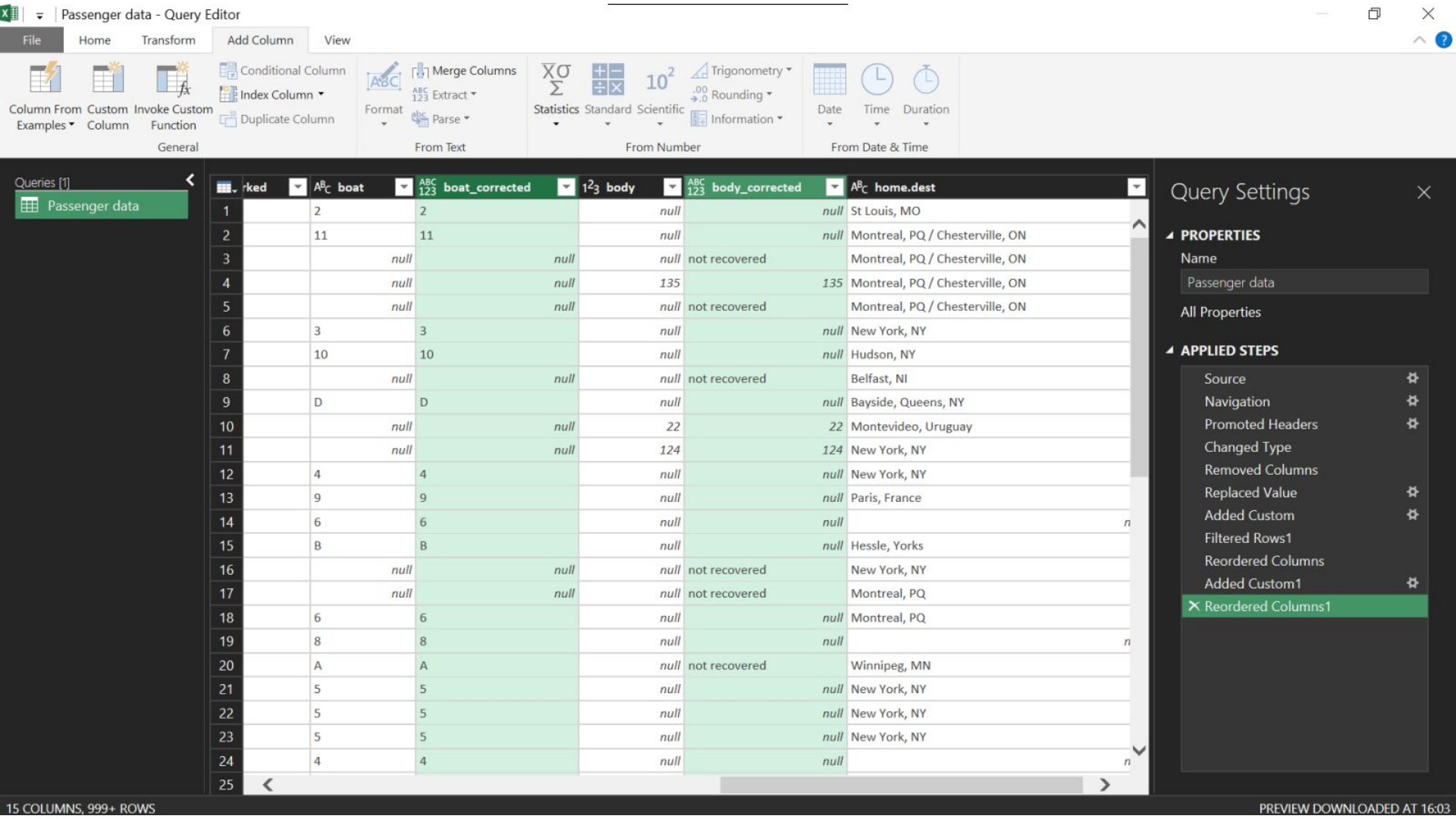
https://github.com/fenago/dataanalysiswithexcel

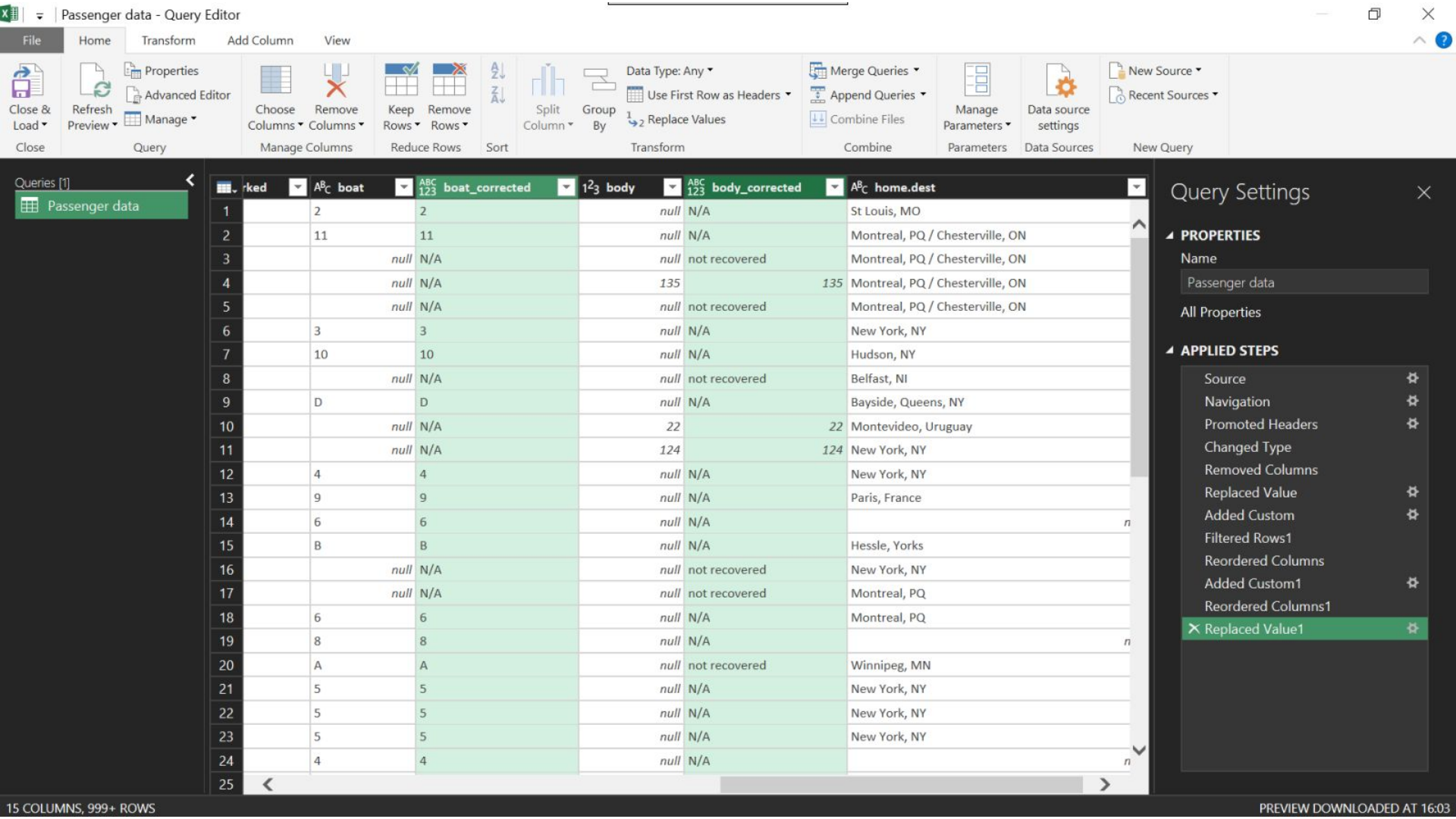

File
Home
Insert
Page Layout
Formulas
Data
Review
View
Add-ins
Power Pivot
PDF-XChange
Team
Tell me what you want to do
Get Data
From Text/CSV
From Web
From Table/Range
Recent Sources
Existing Connections
Refresh All
Queries & Connections
Properties
Edit Links
Connections
Sort
Filter
Clear
Reapply
Advanced
Sort & Filter
Text to Columns
Data Tools
What-If Analysis
Forecast Sheet
Forecast
Group
Ungroup
Subtotal
Outline
Data Analysis
Analysis
From File
From Database
From Azure
From Online Services
From Other Sources
Combine Queries
From Workbook
From Text/CSV
From XML
From JSON
From Folder
From SharePoint Folder
Launch Query Editor...
Data Catalog Search
My Data Catalog Queries
Data Source Settings...
Query Options


Passenger data - Query Editor
File
Home
Transform
Add Column
View
Close & Load
Refresh Preview
Properties
Advanced Editor
Manage
Choose Columns
Remove Columns
Keep Rows
Remove Rows
Split Column
Group By
Data Type: Text
Use First Row as Headers
Replace Values
Merge Queries
Append Queries
Combine Files
Manage Parameters
Data source settings
New Source
Recent Sources
Close
Query
Manage Columns
Reduce Rows
Sort
Transform
Combine
Parameters
Data Sources
New Query
Queries
pclass
survived
sex
age
sibsp
parch
ticket
fare
cabin
embar
1 1 female 29 0 0 24160 211.3375 B5 S
2 1 1 male 0.9167 1 2 113781 151.55 C22 C26 S
3 1 0 female 2 1 2 113781 151.55 C22 C26 S
4 1 0 male 30 1 2 113781 151.55 C22 C26 S
5 1 0 female 25 1 2 113781 151.55 C22 C26 S
6 1 1 male 48 0 0 19952 26.55 E12 S
7 1 1 female 63 1 0 13502 77.9583 D7 S
8 1 0 male 39 0 0 112050 0 A36 S
9 1 1 female 53 2 0 11769 51.4792 C101 S
10 1 0 male 71 0 0 PC 17609 49.5042 null C
11 1 0 male 47 1 0 PC 17757 227.525 C62 C64 C
12 1 1 female 18 1 0 PC 17757 227.525 C62 C64 C
13 1 1 female 24 0 0 PC 17477 69.3 B35 C
14 1 1 female 26 0 0 19877 78.85 null S
15 1 1 male 80 0 0 27042 30 A23 S
16 1 0 male null 0 0 PC 17318 25.925 null S
17 1 0 male 24 0 1 PC 17558 247.5208 B58 B60 C
18 1 1 female 50 0 1 PC 17558 247.5208 B58 B60 C
19 1 1 female 32 0 0 11813 76.2917 D15 C
20 1 0 male 36 0 0 13050 75.2417 C6 C
21 1 1 male 37 1 1 11751 52.5542 D35 S
22 1 1 female 47 1 1 11751 52.5542 D35 S
23 1 1 male 26 0 0 111369 30 C148 C
24 1 1 female 42 0 0 PC 17757 227.525 null C
25
Query Settings
PROPERTIES
Name
Passenger data
All Properties
APPLIED STEPS
Source
Navigation
Promoted Headers
Changed Type
Removed Columns


Passenger data - Query Editor
File
Home
Transform
Add Column
View
Close & Load
Refresh Preview
Properties
Advanced Editor
Manage
Choose Columns
Remove Columns
Keep Rows
Remove Rows
Split Column
Group By
Data Type: Text
Use First Row as Headers
Replace Values
Merge Queries
Append Queries
Combine Files
Manage Parameters
Data source settings
New Source
Recent Sources
Close
Query
Manage Columns
Reduce Rows
Sort
Transform
Combine
Parameters
Data Sources
New Query
Queries
pclass
survived
sex
age
sibsp
parch
ticket
fare
cabin
embar
1 female 29 0 0 24160 211.3375 B5 S
1 male 0.9167 1 2 113781 151.55 C22 C26 S
1 female 2 1 2 113781 151.55 C22 C26 S
1 male 30 1 2 113781 151.55 C22 C26 S
1 female 25 1 2 113781 151.55 C22 C26 S
1 male 48 0 0 19952 26.55 E12 S
1 female 63 1 0 13502 77.9583 D7 S
1 male 39 0 0 112050 0 A36 S
1 female 53 2 0 11769 51.4792 C101 S
1 male 71 0 0 PC 17609 49.5042 unknown C
1 male 47 1 0 PC 17757 227.525 C62 C64 C
1 female 18 1 0 PC 17757 227.525 C62 C64 C
1 female 24 0 0 PC 17477 69.3 B35 C
1 female 26 0 0 19877 78.85 unknown S
1 male 80 0 0 27042 30 A23 S
1 male null 0 0 PC 17318 25.925 unknown S
1 male 24 0 1 PC 17558 247.5208 B58 B60 C
1 female 50 0 1 PC 17558 247.5208 B58 B60 C
1 female 32 0 0 11813 76.2917 D15 C
1 male 36 0 0 13050 75.2417 C6 C
1 male 37 1 1 11751 52.5542 D35 S
1 female 47 1 1 11751 52.5542 D35 S
1 male 26 0 0 111369 30 C148 C
1 female 42 0 0 PC 17757 227.525 unknown C
Query Settings
PROPERTIES
Name
Passenger data
All Properties
APPLIED STEPS
Source
Navigation
Promoted Headers
Changed Type
Removed Columns
Replaced Value
13 COLUMNS, 999+ ROWS
PREVIEW DOWNLOADED AT 12:21


Passenger data - Query Editor
File
Home
Transform
Add Column
View
Column From Examples
Custom Column
Invoke Custom Function
Conditional Column
Index Column
Duplicate Column
General
Format
Merge Columns
Extract
Parse
From Text
Statistics
Standard
Scientific
Trigonometry
Rounding
Information
From Number
Date
Time
Duration
From Date & Time
Queries [1]
Passenger data
cabin
Custom Column
New column name
boat_corrected
Custom column formula:
= if [survived]=1 and [boat] = null then "unknown" else [boat]
Available columns:
pclass
survived
sex
age
sibsp
parch
ticket
<< Insert
Learn about Power Query formulas
No syntax errors have been detected.
OK
Cancel
Query Settings
PROPERTIES
Name
Passenger data
All Properties
APPLIED STEPS
Source
Navigation
Promoted Headers
Changed Type
Removed Columns
Replaced Value
Added Custom
Filtered Rows1
Reordered Columns
14 COLUMNS, 999+ ROWS
PREVIEW DOWNLOADED AT 16:03


Passenger data - Query Editor
File
Home
Transform
Add Column
View
Column From Examples
Custom Column
Invoke Custom Function
Conditional Column
Index Column
Duplicate Column
General
Format
Merge Columns
Extract
Parse
From Text
Statistics
Standard
Scientific
Trigonometry
Rounding
Information
From Number
Date
Time
Duration
From Date & Time
Queries [1]
Passenger data
rked
boat
boat_corrected
body
body_corrected
home.dest
1 2 2 null null St Louis, MO
2 11 11 null null Montreal, PQ / Chesterville, ON
3 null null null not recovered Montreal, PQ / Chesterville, ON
4 null null 135 135 Montreal, PQ / Chesterville, ON
5 null null null not recovered Montreal, PQ / Chesterville, ON
6 3 3 null null New York, NY
7 10 10 null null Hudson, NY
8 null null null not recovered Belfast, NI
9 D D null null Bayside, Queens, NY
10 null null 22 22 Montevideo, Uruguay
11 null null 124 124 New York, NY
12 4 4 null null New York, NY
13 9 9 null null Paris, France
14 6 6 null null n
15 B B null null Hessle, Yorks
16 null null null not recovered New York, NY
17 null null null not recovered Montreal, PQ
18 6 6 null null Montreal, PQ
19 8 8 null null n
20 A A null not recovered Winnipeg, MN
21 5 5 null null New York, NY
22 5 5 null null New York, NY
23 5 5 null null New York, NY
24 4 4 null null n
25
Query Settings
PROPERTIES
Name
Passenger data
All Properties
APPLIED STEPS
Source
Navigation
Promoted Headers
Changed Type
Removed Columns
Replaced Value
Added Custom
Filtered Rows1
Reordered Columns
Added Custom1
Reordered Columns1
15 COLUMNS, 999+ ROWS
PREVIEW DOWNLOADED AT 16:03


Passenger data - Query Editor
File
Home
Transform
Add Column
View
Close & Load
Refresh Preview
Properties
Advanced Editor
Manage
Close
Query
Choose Columns
Remove Columns
Manage Columns
Keep Rows
Remove Rows
Reduce Rows
Sort
Split Column
Group By
Data Type: Any
Use First Row as Headers
Replace Values
Transform
Merge Queries
Append Queries
Combine Files
Combine
Manage Parameters
Parameters
Data source settings
Data Sources
New Source
Recent Sources
New Query
Queries [1]
Passenger data
rked
boat
boat_corrected
body
body_corrected
home.dest
1 2 2 null N/A St Louis, MO
2 11 11 null N/A Montreal, PQ / Chesterville, ON
3 null N/A null not recovered Montreal, PQ / Chesterville, ON
4 null N/A 135 135 Montreal, PQ / Chesterville, ON
5 null N/A null not recovered Montreal, PQ / Chesterville, ON
6 3 3 null N/A New York, NY
7 10 10 null N/A Hudson, NY
8 null N/A null not recovered Belfast, NI
9 D D null N/A Bayside, Queens, NY
10 null N/A 22 22 Montevideo, Uruguay
11 null N/A 124 124 New York, NY
12 4 4 null N/A New York, NY
13 9 9 null N/A Paris, France
14 6 6 null N/A n
15 B B null N/A Hessle, Yorks
16 null N/A null not recovered New York, NY
17 null N/A null not recovered Montreal, PQ
18 6 6 null N/A Montreal, PQ
19 8 8 null N/A n
20 A A null not recovered Winnipeg, MN
21 5 5 null N/A New York, NY
22 5 5 null N/A New York, NY
23 5 5 null N/A New York, NY
24 4 4 null N/A n
25
Query Settings
PROPERTIES
Name
Passenger data
All Properties
APPLIED STEPS
Source
Navigation
Promoted Headers
Changed Type
Removed Columns
Replaced Value
Added Custom
Filtered Rows1
Reordered Columns
Added Custom1
Reordered Columns1
Replaced Value1
15 COLUMNS, 999+ ROWS
PREVIEW DOWNLOADED AT 16:03

Add Conditional Column

Add a conditional column that is computed from the other columns or values.

New column name

Age group

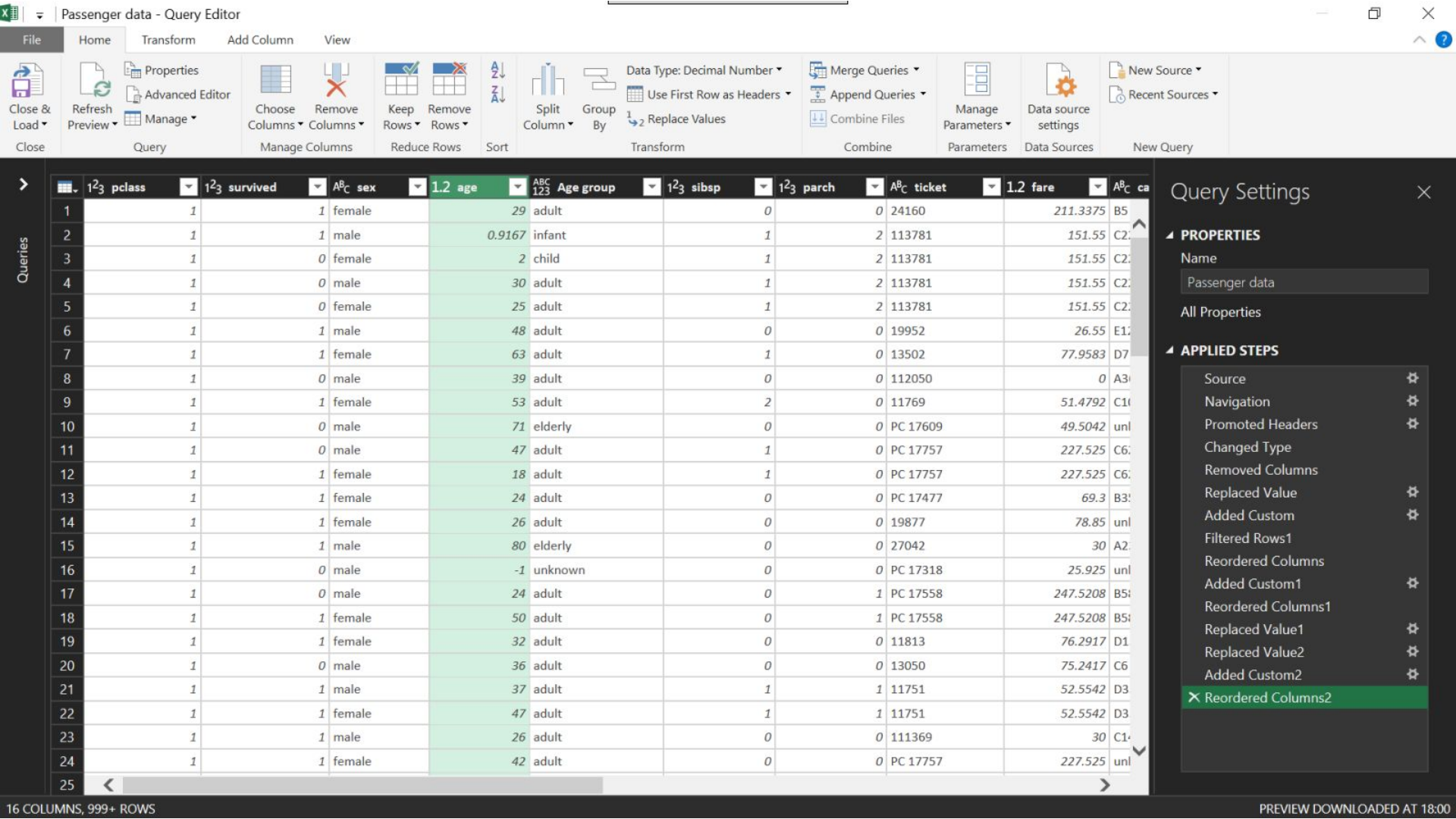
	Column Name	Operator	Value		Output	
If	age	equals	-1	Then	unknown	...
Else If	age	is less than	1	Then	infant	
Else If	age	is less than	12	Then	child	
Else If	age	is less than	18	Then	teenager	
Else If	age	is less than	65	Then	adult	
Else If	age	is greater than or...	65	Then	elderly	

Add rule

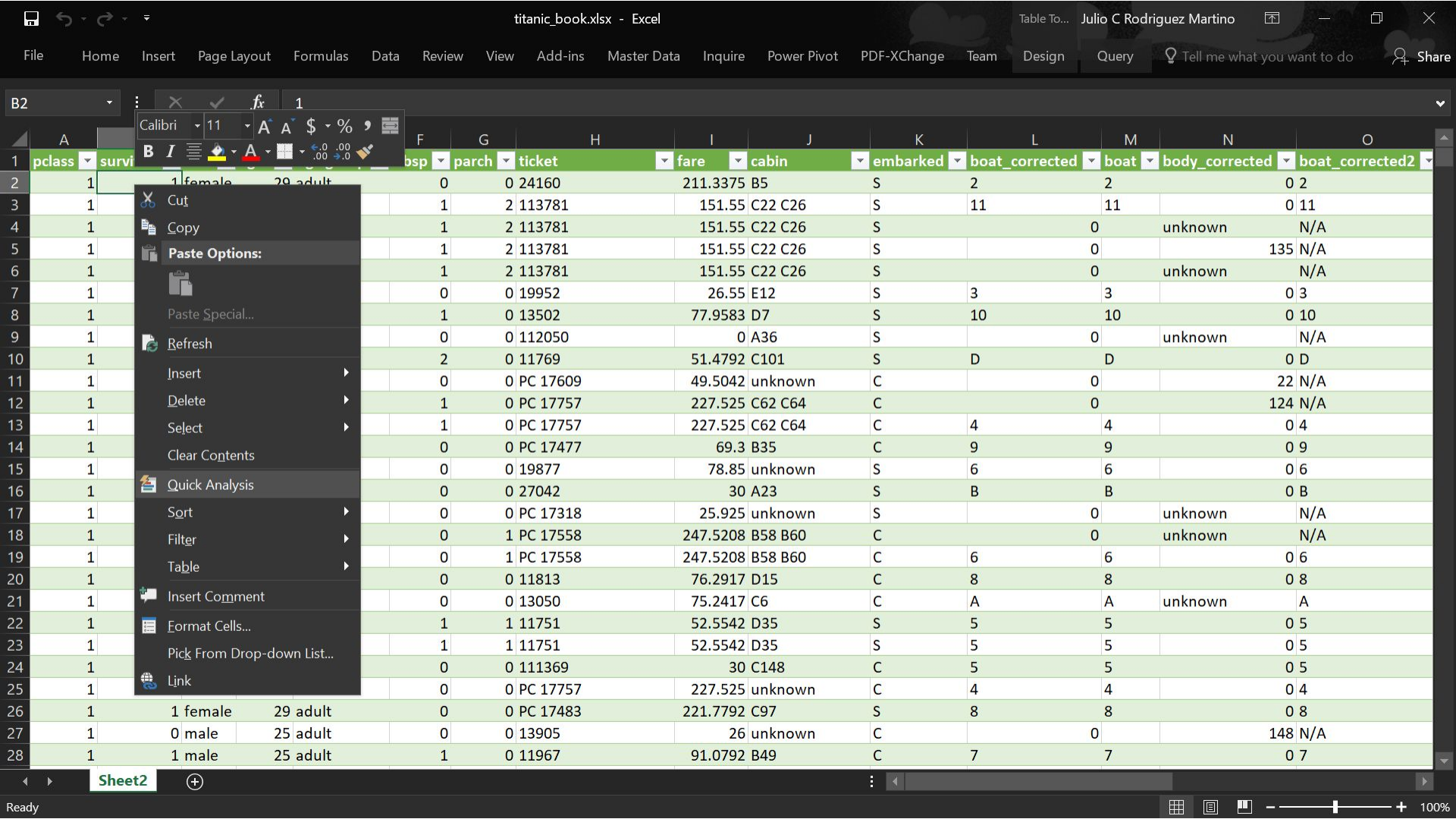
Otherwise

unknown

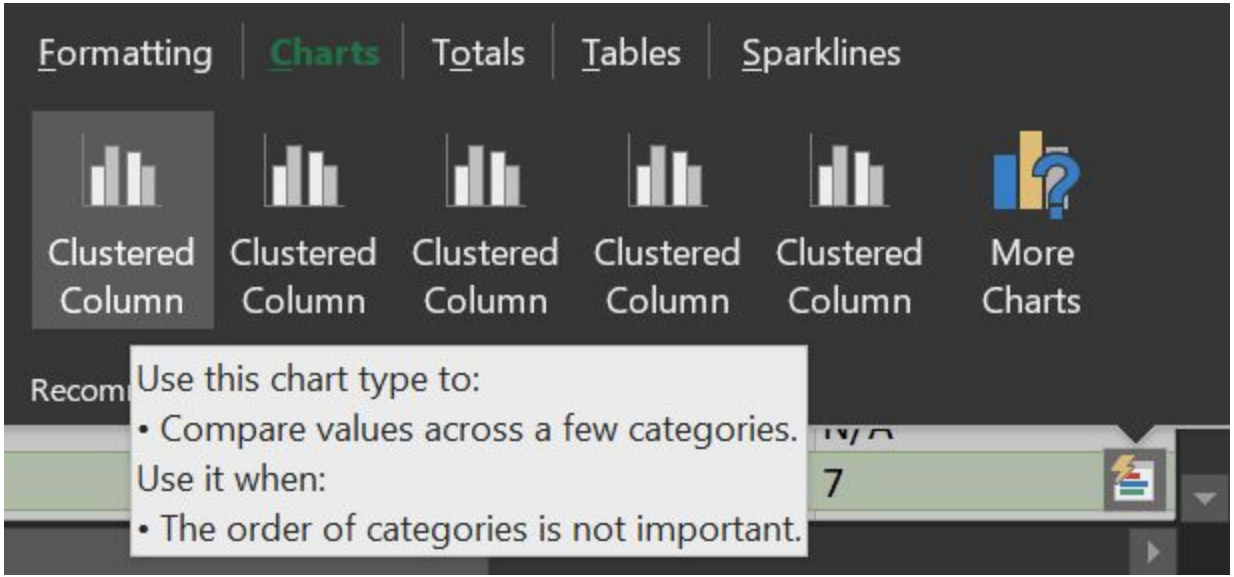
OK Cancel


Passenger data - Query Editor
File
Home
Transform
Add Column
View
Close & Load
Refresh Preview
Properties
Advanced Editor
Manage
Choose Columns
Remove Columns
Keep Rows
Remove Rows
Split Column
Group By
Data Type: Decimal Number
Use First Row as Headers
Replace Values
Merge Queries
Append Queries
Combine Files
Manage Parameters
Data source settings
New Source
Recent Sources
Close
Query
Manage Columns
Reduce Rows
Sort
Transform
Combine
Parameters
Data Sources
New Query
Queries
pclass
survived
sex
age
Age group
sibsp
parch
ticket
fare
ca
1 1 female 29 adult 0 0 24160 211.3375 B5
2 1 1 male 0.9167 infant 1 2 113781 151.55 C2
3 1 0 female 2 child 1 2 113781 151.55 C2
4 1 0 male 30 adult 1 2 113781 151.55 C2
5 1 0 female 25 adult 1 2 113781 151.55 C2
6 1 1 male 48 adult 0 0 19952 26.55 E1
7 1 1 female 63 adult 1 0 13502 77.9583 D7
8 1 0 male 39 adult 0 0 112050 0 A3
9 1 1 female 53 adult 2 0 11769 51.4792 C1
10 1 0 male 71 elderly 0 0 PC 17609 49.5042 unl
11 1 0 male 47 adult 1 0 PC 17757 227.525 C6
12 1 1 female 18 adult 1 0 PC 17757 227.525 C6
13 1 1 female 24 adult 0 0 PC 17477 69.3 B3
14 1 1 female 26 adult 0 0 19877 78.85 unl
15 1 1 male 80 elderly 0 0 27042 30 A2
16 1 0 male -1 unknown 0 0 PC 17318 25.925 unl
17 1 0 male 24 adult 0 1 PC 17558 247.5208 B5
18 1 1 female 50 adult 0 1 PC 17558 247.5208 B5
19 1 1 female 32 adult 0 0 11813 76.2917 D1
20 1 0 male 36 adult 0 0 13050 75.2417 C6
21 1 1 male 37 adult 1 1 11751 52.5542 D3
22 1 1 female 47 adult 1 1 11751 52.5542 D3
23 1 1 male 26 adult 0 0 111369 30 C1
24 1 1 female 42 adult 0 0 PC 17757 227.525 unl
25
Query Settings
PROPERTIES
Name
Passenger data
All Properties
APPLIED STEPS
Source
Navigation
Promoted Headers
Changed Type
Removed Columns
Replaced Value
Added Custom
Filtered Rows1
Reordered Columns
Added Custom1
Reordered Columns1
Replaced Value1
Replaced Value2
Added Custom2
Reordered Columns2
16 COLUMNS, 999+ ROWS
PREVIEW DOWNLOADED AT 18:00

Visualize Data!



pclass	survi				bsp	parch	ticket	fare	cabin	embarked	boat_corrected	boat	body_corrected	boat_corrected2
1	1	female	29	adult	0	0	24160	211.3375	B5	S	2	2	0	2
1					1	2	113781	151.55	C22 C26	S	11	11	0	11
1					1	2	113781	151.55	C22 C26	S		0	unknown	N/A
1					1	2	113781	151.55	C22 C26	S		0	135	N/A
1					1	2	113781	151.55	C22 C26	S		0	unknown	N/A
1					0	0	19952	26.55	E12	S	3	3	0	3
1					1	0	13502	77.9583	D7	S	10	10	0	10
1					0	0	112050	0	A36	S		0	unknown	N/A
1					2	0	11769	51.4792	C101	S	D	D	0	D
1					0	0	PC 17609	49.5042	unknown	C		0	22	N/A
1					1	0	PC 17757	227.525	C62 C64	C		0	124	N/A
1					1	0	PC 17757	227.525	C62 C64	C	4	4	0	4
1					0	0	PC 17477	69.3	B35	C	9	9	0	9
1					0	0	19877	78.85	unknown	S	6	6	0	6
1					0	0	27042	30	A23	S	B	B	0	B
1					0	0	PC 17318	25.925	unknown	S		0	unknown	N/A
1					0	1	PC 17558	247.5208	B58 B60	C		0	unknown	N/A
1					0	1	PC 17558	247.5208	B58 B60	C	6	6	0	6
1					0	0	11813	76.2917	D15	C	8	8	0	8
1					0	0	13050	75.2417	C6	C	A	A	unknown	A
1					1	1	11751	52.5542	D35	S	5	5	0	5
1					1	1	11751	52.5542	D35	S	5	5	0	5
1					0	0	111369	30	C148	C	5	5	0	5
1					0	0	PC 17757	227.525	unknown	C	4	4	0	4
1	1	female	29	adult	0	0	PC 17483	221.7792	C97	S	8	8	0	8
1	0	male	25	adult	0	0	13905	26	unknown	C		0	148	N/A
1	1	male	25	adult	1	0	11967	91.0792	B49	C	7	7	0	7


Formatting
Charts
Totals
Tables
Sparklines
Clustered Column
Clustered Column
Clustered Column
Clustered Column
Clustered Column
More Charts
Recom
Use this chart type to:
• Compare values across a few categories.
Use it when:
• The order of categories is not important.
N/A
7

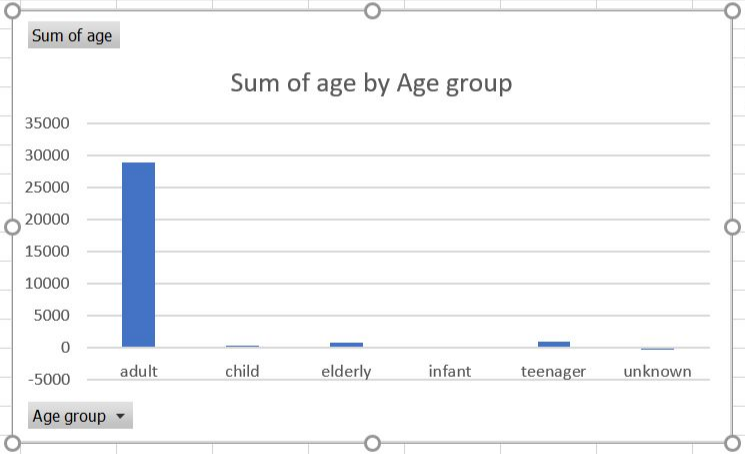
titanic_book.xlsx - Excel

Julio C Rodriguez Martino

File Home Insert Page Layout Formulas Data Review View Add-ins Master Data Inquire Power Pivot PDF-XChange Team Analyze Design Format Tell me Share

Chart 1

Age group	Sum of age
adult	28943.5
child	417.5
elderly	910.5
infant	8.1667
teenager	976
unknown	-263



PivotChart Fields

Choose fields to add to report:

Search

- pclass
- survived
- sex
- age
- Age group
- sibsp
- parch
- ticket
- fare

Drag fields between areas b

Filters

Axis (Categories)

Age group

Sum of age

Move Up
Move Down
Move to Beginning
Move to End
Move to Report Filter
Move to Axis Fields (Categories)
Move to Legend Fields (Series)
Move to Values
Hide Value Field Buttons on Chart
Hide All Field Buttons on Chart
Remove Field
Value Field Settings...

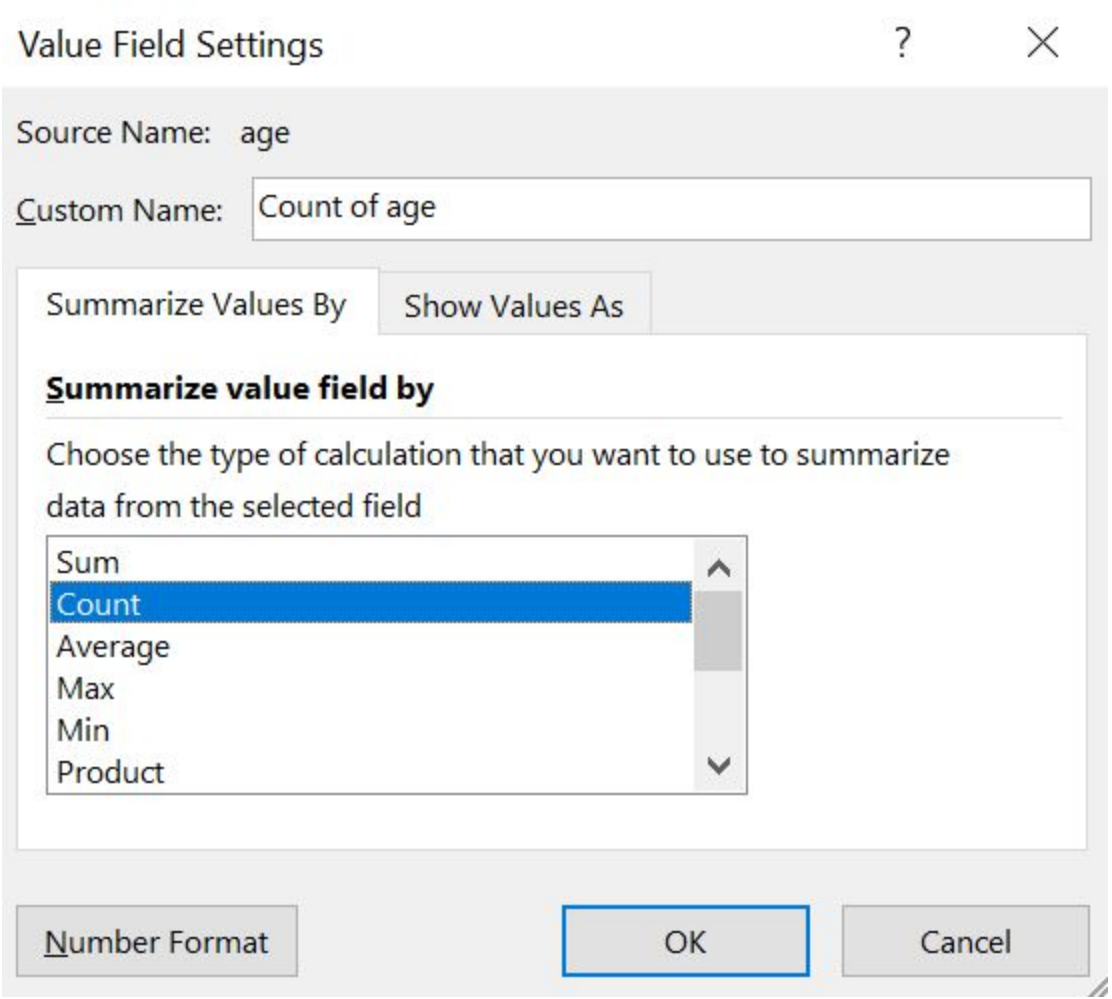
Defer Layout Update

Update

Sheet1 Sheet2

Ready Calculate

100%


Value Field Settings
?
Source Name: age
Custom Name: Count of age
Summarize Values By
Show Values As
Summarize value field by
Choose the type of calculation that you want to use to summarize data from the selected field
Sum
Count
Average
Max
Min
Product
Number Format
OK
Cancel

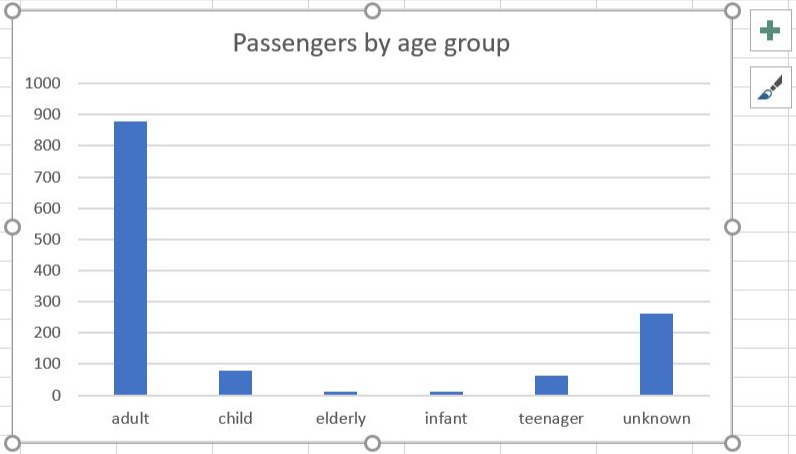
titanic_book.xlsx - Excel

Julio C Rodriguez Martino

File Home Insert Page Layout Formulas Data Review View Add-ins Master Data Inquire Power Pivot PDF-XChange Team Analyze Design Format Tell me Share

Chart 1

Age group	Count of Age group
adult	879
child	79
elderly	13
infant	12
teenager	63
unknown	263



PivotChart Fields

Choose fields to add to report:

Search

- pclass
- survived
- sex
- age
- **Age group**
- sibsp
- parch
- ticket
- fare

Drag fields between areas below:

Filters | Legend (Series)

Axis (Categories): Age group | Values: Count of Age group

Defer Layout Update | Update

Sheet1 Sheet2

Ready

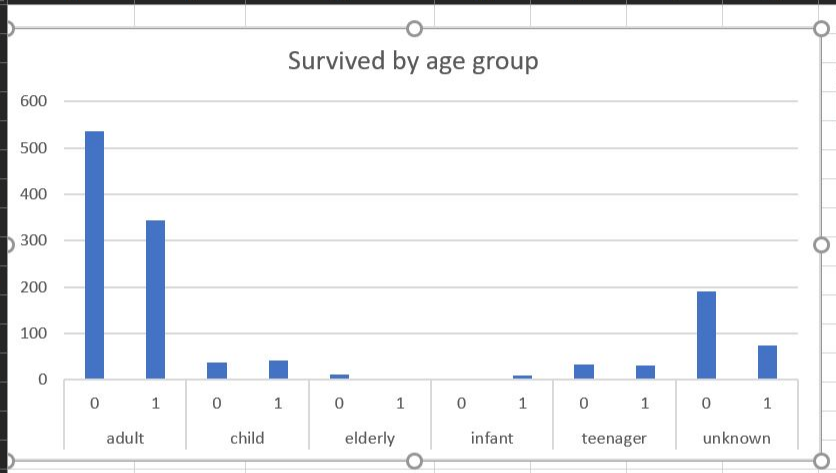
100%

titanic_book.xlsx - Excel

Julio C Rodriguez Martino

File Home Insert Page Layout Formulas Data Review View Add-ins Master Data Inquire Power Pivot PDF-XChange Team Analyze Design Format Tell me Share

Chart 1



PivotChart Fields

Choose fields to add to report:

Search

- pclass
- **survived**
- sex
- age
- **Age group**
- sibsp
- parch
- ticket
- fare

Drag fields between areas below:

Filters

Legend (Series)

Axis (Categories)

Age group

survived

Values

Count of Age group

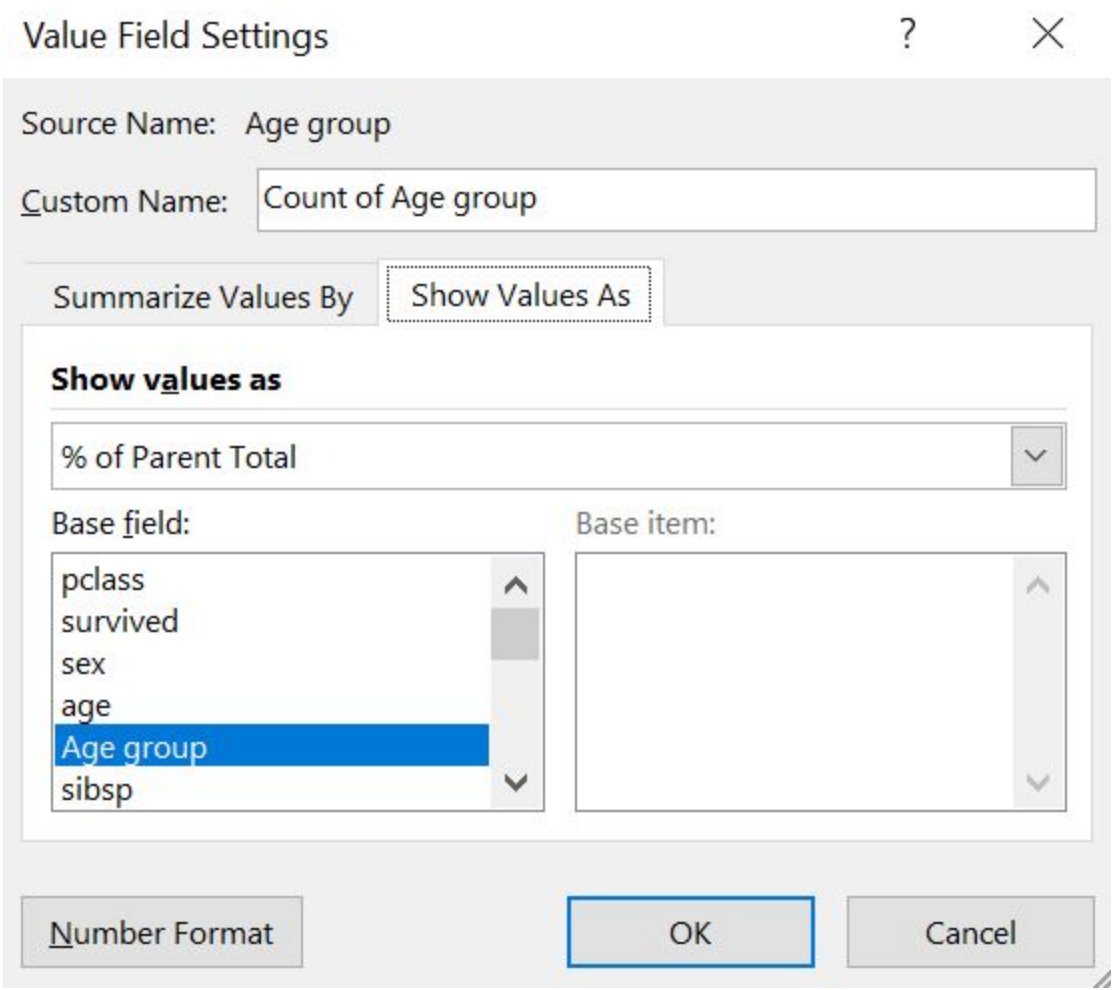
Defer Layout Update

Update

Sheet1 Sheet2

Ready

100%


Value Field Settings
Source Name: Age group
Custom Name: Count of Age group
Summarize Values By
Show Values As
Show values as
% of Parent Total
Base field:
pclass
survived
sex
age
Age group
sibsp
Base item:
Number Format
OK
Cancel

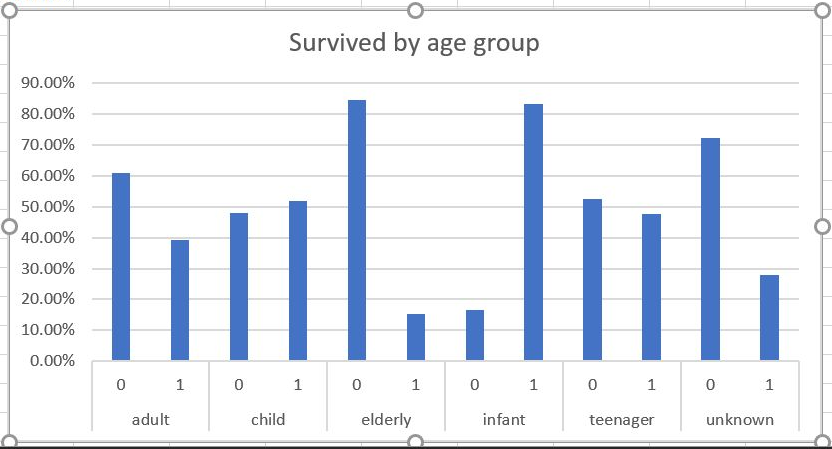
titanic_book.xlsx - Excel

Julio C Rodriguez Martino

File Home Insert Page Layout Formulas Data Review View Add-ins Master Data Inquire Power Pivot PDF-XChange Team Analyze Design Format Tell me Share

Chart 1

Age group	survived	Count of Age group
adult	0	60.86%
adult	1	39.14%
child	0	48.10%
child	1	51.90%
elderly	0	84.62%
elderly	1	15.38%
infant	0	16.67%
infant	1	83.33%
teenager	0	52.38%
teenager	1	47.62%
unknown	0	72.24%
unknown	1	27.76%



PivotChart Fields

Choose fields to add to report:

Search

- pclass
- **survived**
- sex
- age
- **Age group**
- sibsp
- parch
- ticket
- fare

Drag fields between areas below:

Filters | Legend (Series)

Axis (Categories): Age group, survived

Values: Count of Age group

Defer Layout Update Update

Sheet1 Sheet2

Ready 100%

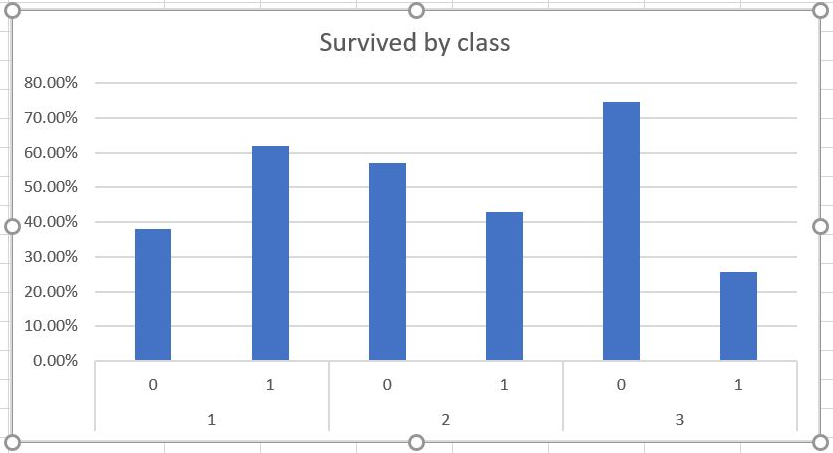
titanic_book.xlsx - Excel

Julio C Rodriguez Martino

File Home Insert Page Layout Formulas Data Review View Add-ins Master Data Inquire Power Pivot PDF-XChange Team Analyze Design Format Tell me Share

Chart 1

pclass	survived	Sum of pclass
1	0	38.08%
1	1	61.92%
2	0	57.04%
2	1	42.96%
3	0	74.47%
3	1	25.53%



Sheet1 Sheet2

PivotChart Fields

Choose fields to add to report:

Search

- [x] survived
- [ ] sex
- [ ] age
- [ ] Age group
- [ ] sibsp
- [ ] parch
- [ ] ticket
- [ ] fare
- [ ] cabin

Drag fields between areas below:

Filters | Legend (Series)

Axis (Categories): pclass, survived

Values: Sum of pclass

Defer Layout Update Update

Ready 100%

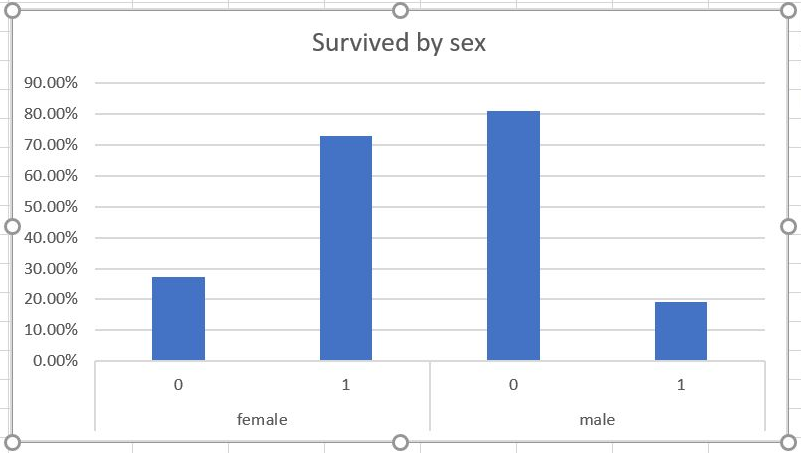
titanic_book.xlsx - Excel

Julio C Rodriguez Martino

File Home Insert Page Layout Formulas Data Review View Add-ins Master Data Inquire Power Pivot PDF-XChange Team Analyze Design Format Tell me Share

Chart 1

sex	survived	Count of sex
fema	0	27.25%
fema	1	72.75%
male	0	80.90%
male	1	19.10%



PivotChart Fields

Choose fields to add to report:

Search

- boat_corrected
- boat
- body_corrected
- boat_corrected2
- body
- body_corrected3
- home.dest
- % Total

Drag fields between areas below:

Filters

Legend (Series)

Axis (Categories)
- sex
- survived

Values
- Count of sex

Defer Layout Update

Update

Sheet1 Sheet2

ESP 09:28

Bias and Balance

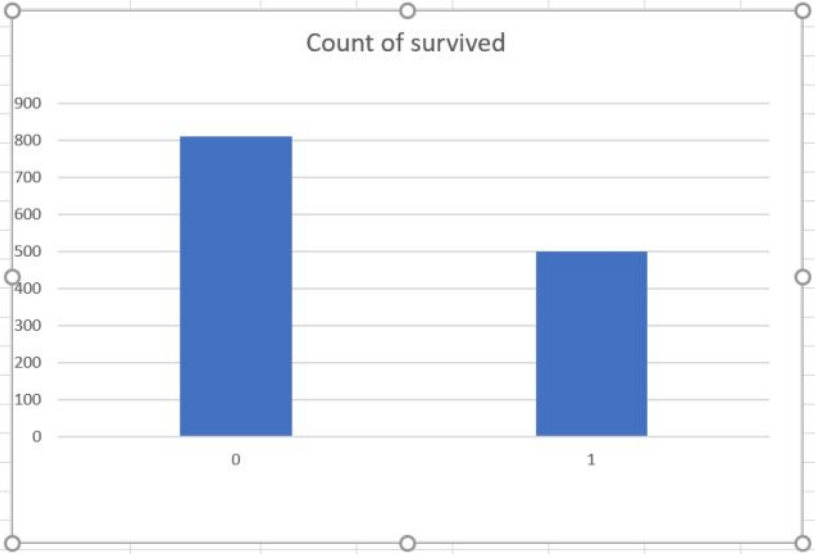
titanic_book.xlsx - Excel

Julio C Rodriguez Martino

File Home Insert Page Layout Formulas Data Review View Add-ins Master Data Inquire Power Pivot PDF-XChange Team Analyze Design Format Tell me Share

Chart 1

survived	Count of survived
0	809
1	500



PivotChart Fields

Choose fields to add to report:

Search

- pclass
- survived
- sex
- age
- Age group
- sibsp
- parch
- ticket
- fare

Drag fields between areas below:

Filters

Legend (Series)

Axis (Categories)

survived

Values

Count of survived

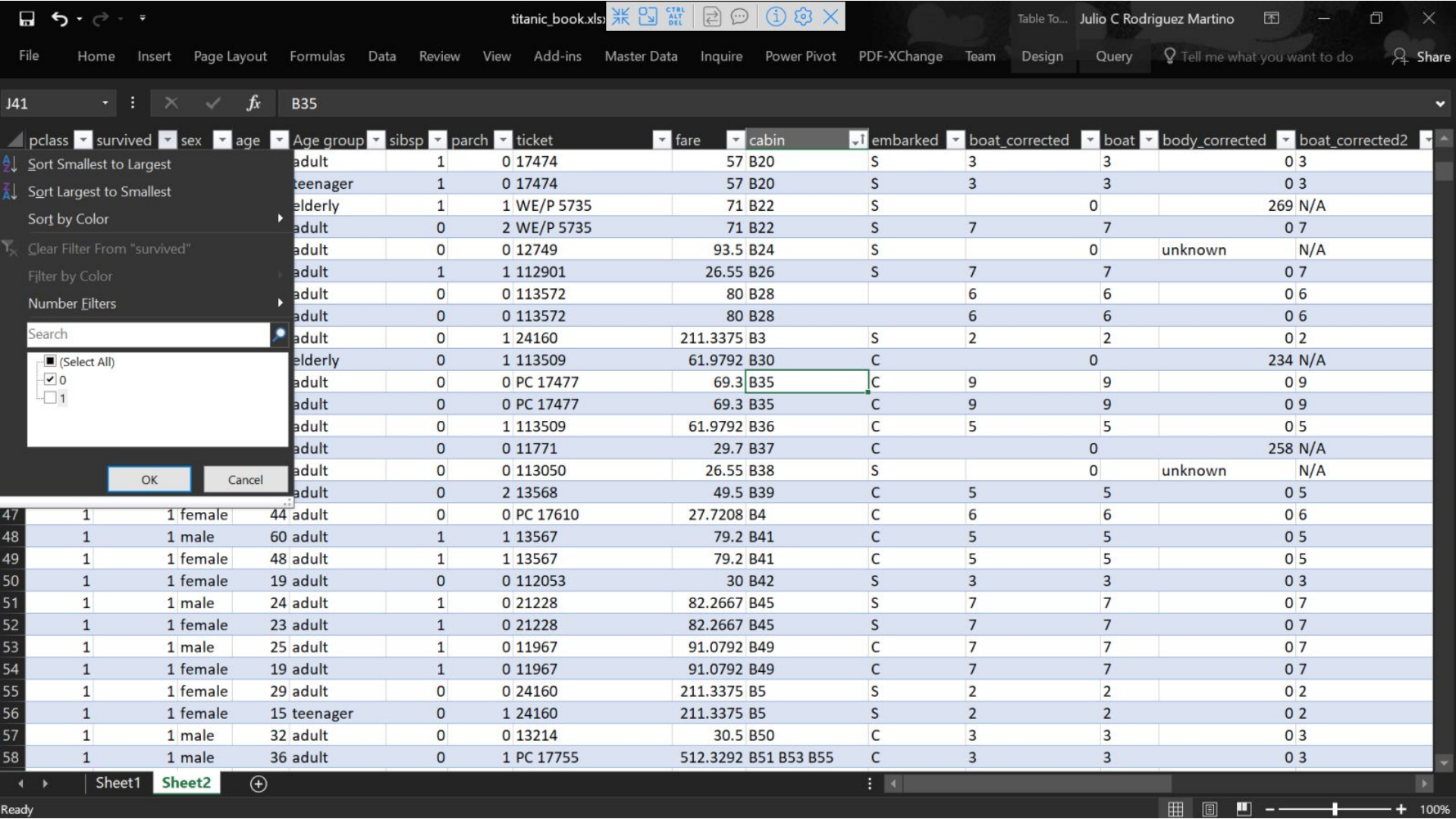
Defer Layout Update

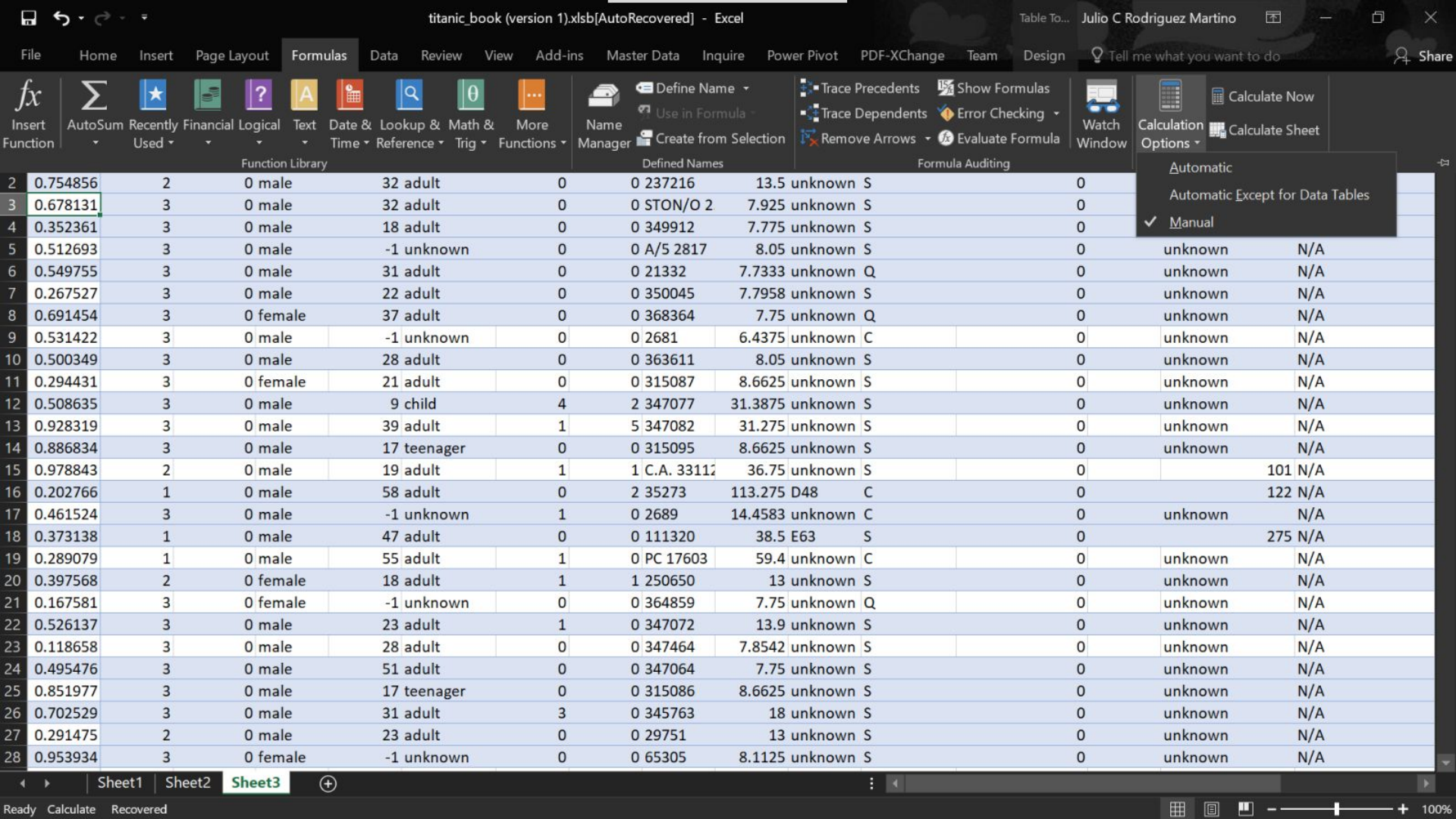
Update

Sheet1 Sheet2

Ready

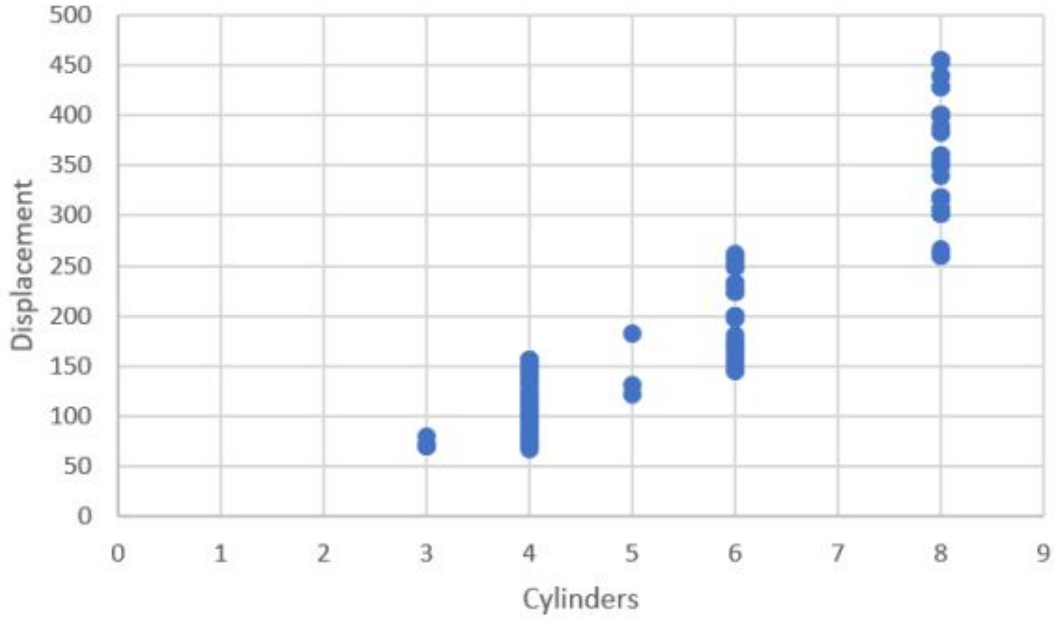
100%

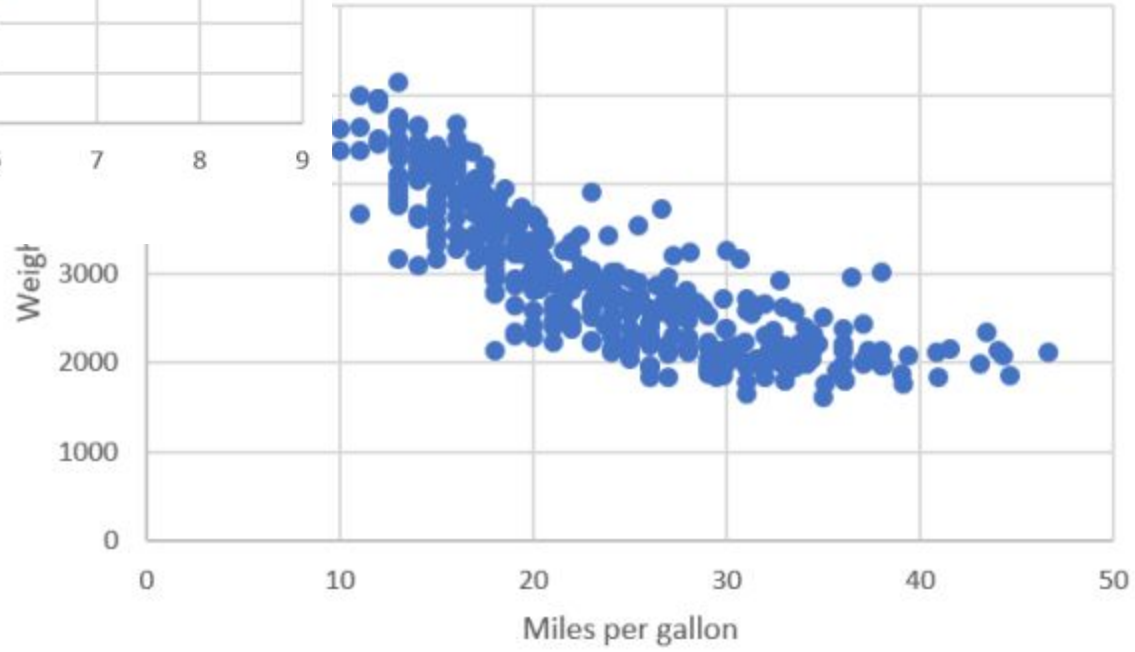

titanic_book.xlsx
Julio C Rodriguez Martino
File Home Insert Page Layout Formulas Data Review View Add-ins Master Data Inquire Power Pivot PDF-XChange Team Design Query Tell me what you want to do Share
J41
B35
pclass survived sex age Age group sibsp parch ticket fare cabin embarked boat_corrected boat body_corrected boat_corrected2
Sort Smallest to Largest
Sort Largest to Smallest
Sort by Color
Clear Filter From "survived"
Filter by Color
Number Filters
Search
(Select All)
0
1
OK
Cancel
1 1 female 44 adult 0 0 PC 17610 27.7208 B4 C 6 6 0 6
1 1 male 60 adult 1 1 13567 79.2 B41 C 5 5 0 5
1 1 female 48 adult 1 1 13567 79.2 B41 C 5 5 0 5
1 1 female 19 adult 0 0 112053 30 B42 S 3 3 0 3
1 1 male 24 adult 1 0 21228 82.2667 B45 S 7 7 0 7
1 1 female 23 adult 1 0 21228 82.2667 B45 S 7 7 0 7
1 1 male 25 adult 1 0 11967 91.0792 B49 C 7 7 0 7
1 1 female 19 adult 1 0 11967 91.0792 B49 C 7 7 0 7
1 1 female 29 adult 0 0 24160 211.3375 B5 S 2 2 0 2
1 1 female 15 teenager 0 1 24160 211.3375 B5 S 2 2 0 2
1 1 male 32 adult 0 0 13214 30.5 B50 C 3 3 0 3
1 1 male 36 adult 0 1 PC 17755 512.3292 B51 B53 B55 C 3 3 0 3
Sheet1 Sheet2
Ready
100%

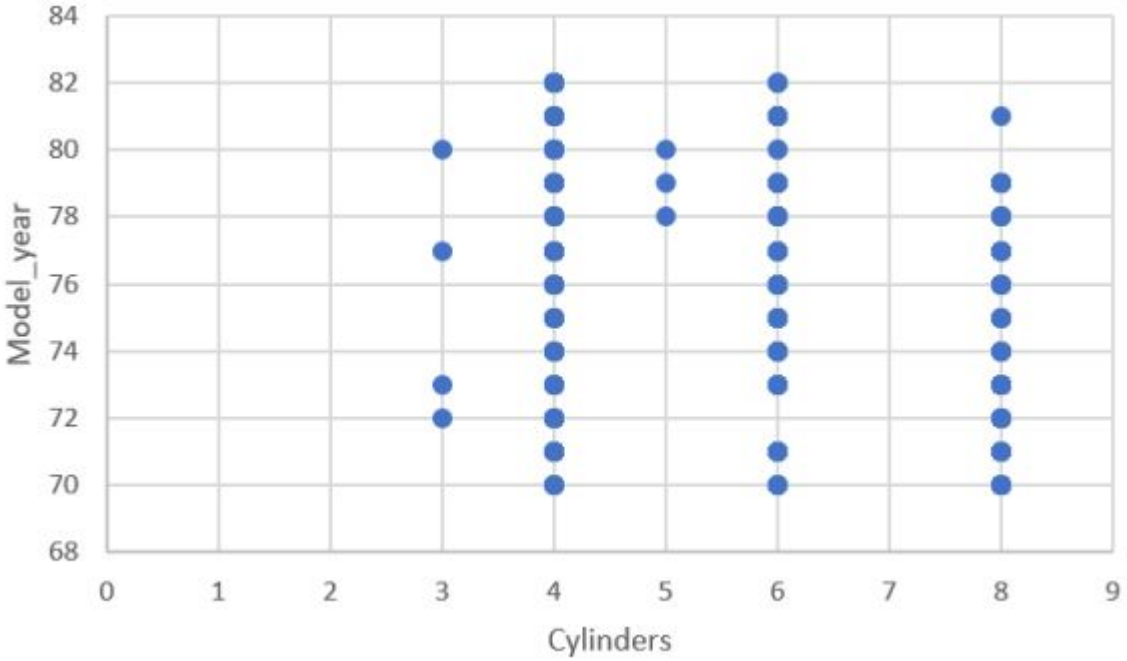

titanic_book (version 1).xlsb[AutoRecovered] - Excel
Julio C Rodriguez Martino
File Home Insert Page Layout Formulas Data Review View Add-ins Master Data Inquire Power Pivot PDF-XChange Team Design
Tell me what you want to do
Share
Insert Function
AutoSum
Recently Used
Financial
Logical
Text
Date & Time
Lookup & Reference
Math & Trig
More Functions
Function Library
Name Manager
Define Name
Use in Formula
Create from Selection
Defined Names
Trace Precedents
Trace Dependents
Remove Arrows
Show Formulas
Error Checking
Evaluate Formula
Formula Auditing
Watch Window
Calculation Options
Calculate Now
Calculate Sheet
Automatic
Automatic Except for Data Tables
Manual
0.754856 2 0 male 32 adult 0 0 237216 13.5 unknown S 0
0.678131 3 0 male 32 adult 0 0 STON/O 2 7.925 unknown S 0
0.352361 3 0 male 18 adult 0 0 349912 7.775 unknown S 0
0.512693 3 0 male -1 unknown 0 0 A/5 2817 8.05 unknown S 0 unknown N/A
0.549755 3 0 male 31 adult 0 0 21332 7.7333 unknown Q 0 unknown N/A
0.267527 3 0 male 22 adult 0 0 350045 7.7958 unknown S 0 unknown N/A
0.691454 3 0 female 37 adult 0 0 368364 7.75 unknown Q 0 unknown N/A
0.531422 3 0 male -1 unknown 0 0 2681 6.4375 unknown C 0 unknown N/A
0.500349 3 0 male 28 adult 0 0 363611 8.05 unknown S 0 unknown N/A
0.294431 3 0 female 21 adult 0 0 315087 8.6625 unknown S 0 unknown N/A
0.508635 3 0 male 9 child 4 2 347077 31.3875 unknown S 0 unknown N/A
0.928319 3 0 male 39 adult 1 5 347082 31.275 unknown S 0 unknown N/A
0.886834 3 0 male 17 teenager 0 0 315095 8.6625 unknown S 0 unknown N/A
0.978843 2 0 male 19 adult 1 1 C.A. 33112 36.75 unknown S 0 101 N/A
0.202766 1 0 male 58 adult 0 2 35273 113.275 D48 C 0 122 N/A
0.461524 3 0 male -1 unknown 1 0 2689 14.4583 unknown C 0 unknown N/A
0.373138 1 0 male 47 adult 0 0 111320 38.5 E63 S 0 275 N/A
0.289079 1 0 male 55 adult 1 0 PC 17603 59.4 unknown C 0 unknown N/A
0.397568 2 0 female 18 adult 1 1 250650 13 unknown S 0 unknown N/A
0.167581 3 0 female -1 unknown 0 0 364859 7.75 unknown Q 0 unknown N/A
0.526137 3 0 male 23 adult 1 0 347072 13.9 unknown S 0 unknown N/A
0.118658 3 0 male 28 adult 0 0 347464 7.8542 unknown S 0 unknown N/A
0.495476 3 0 male 51 adult 0 0 347064 7.75 unknown S 0 unknown N/A
0.851977 3 0 male 17 teenager 0 0 315086 8.6625 unknown S 0 unknown N/A
0.702529 3 0 male 31 adult 3 0 345763 18 unknown S 0 unknown N/A
0.291475 2 0 male 23 adult 0 0 29751 13 unknown S 0 unknown N/A
0.953934 3 0 female -1 unknown 0 0 65305 8.1125 unknown S 0 unknown N/A
Sheet1 Sheet2 Sheet3
Ready Calculate Recovered
100%

Correlation

- Building a scatter diagram
- Calculating the covariance
- Calculating the Pearson's coefficient of correlation
- Studying the Spearman's correlation
- Understanding least squares
- Focusing on feature selection



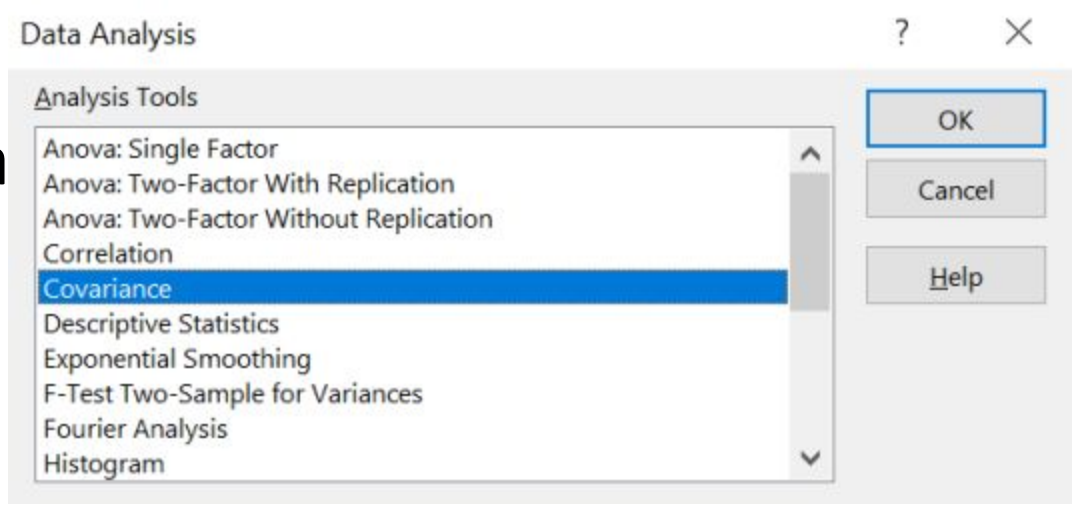


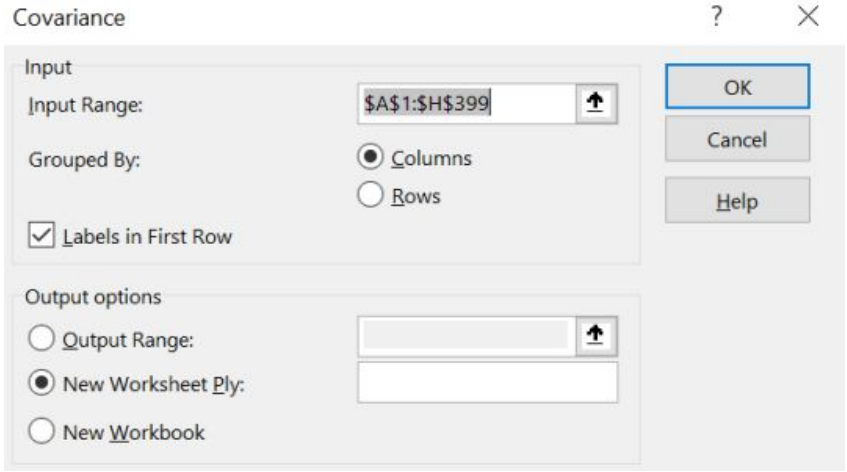


$Nv * (Nv - 1)$

$$Cov(x, y) = mean[(x - \hat{x}).(y - \hat{y})]$$

1. Open the data file.
2. Navigate to Data | Data Analysis.
3. In the pop-up window, select Covariance
4. Select the data range; in this case, it is the whole table except the last column, which contains the car name and is non-numeric:





	A	B	C	D	E	F	G	H	I
1		*mpg*	*cylinders*	*displacement*	*horsepower*	*weight*	*acceleration*	*model_year*	*origin*
2	mpg	60.93611929							
3	cylinders	-10.28300927	2.886146						
4	displacem	-653.7555781	168.1995	10844.88207					
5	horsepow	-233.2613494	55.20705	3604.81427	1477.789879				
6	weight	-5491.379555	1287.453	82161.4674	28193.51406	715339.1287			
7	acceleratic	9.036168531	-2.36489	-155.9401792	-73.00026551	-972.4495158	7.585740575		
8	model_ye	16.69909977	-2.18799	-142.3585516	-58.88582882	-957.5344183	2.930722709	13.63808995	
9	origin	3.523310017	-0.76555	-50.83693594	-14.07673886	-393.647774	0.45420949	0.534443575	0.641676
10									

Pearson Coefficient

$$\rho_{x,y} = \frac{\sum_i (x_i - \hat{x})(y_i - \hat{y})}{\sqrt{\sum_i (x_i - \hat{x})^2 (y_i - \hat{y})^2}}$$

	A	B	C	D	E	F	G	H	I
1		*mpg*	*cylinders*	*displacement*	*horsepower*	*weight*	*acceleration*	*model_year*	*origin*
2	mpg	1							
3	cylinders	-0.7754	1						
4	displacement	-0.8042	0.950721	1					
5	horsepower	-0.77843	0.842983	0.897257002	1				
6	weight	-0.83174	0.896017	0.932824147	0.864537738	1			
7	acceleration	0.420289	-0.50542	-0.543684084	-0.68919551	-0.41745732	1		
8	model_year	0.579267	-0.34875	-0.370164161	-0.416361477	-0.306564334	0.288136954	1	
9	origin	0.56345	-0.56254	-0.609409399	-0.455171453	-0.581023914	0.205873007	0.180662195	1
10									

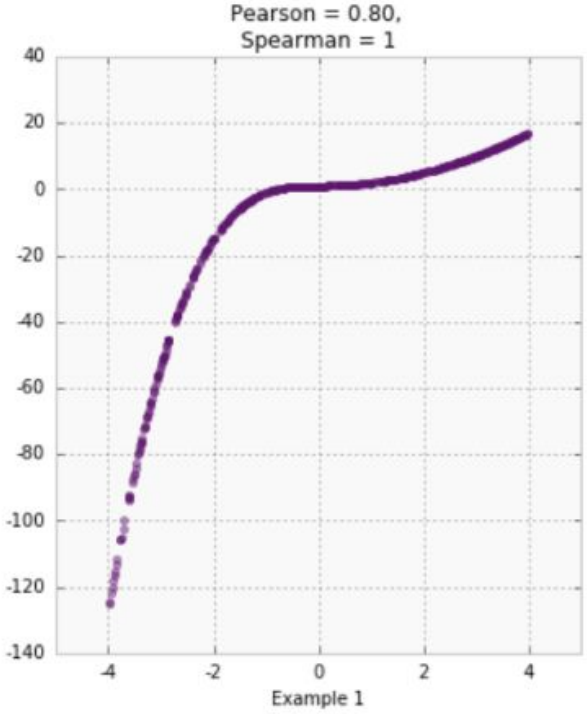
$$\rho_{x,y} = b. \frac{\sigma_x}{\sigma_y}$$

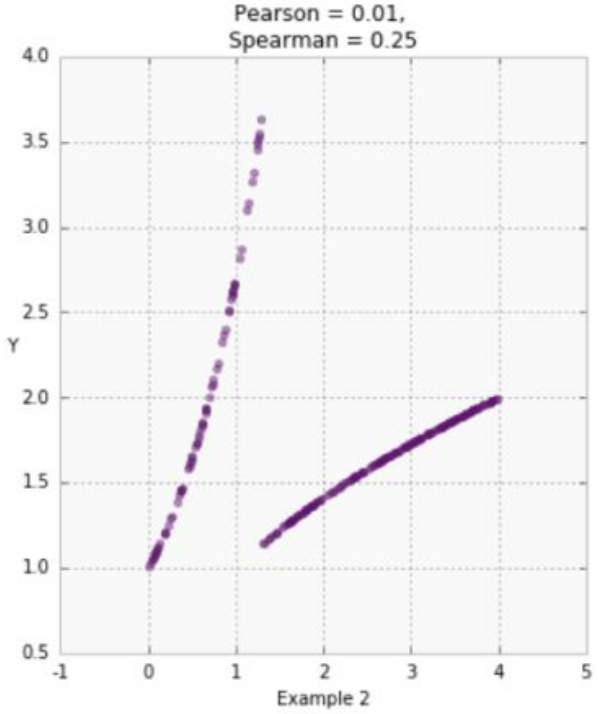
Spearman Coefficient

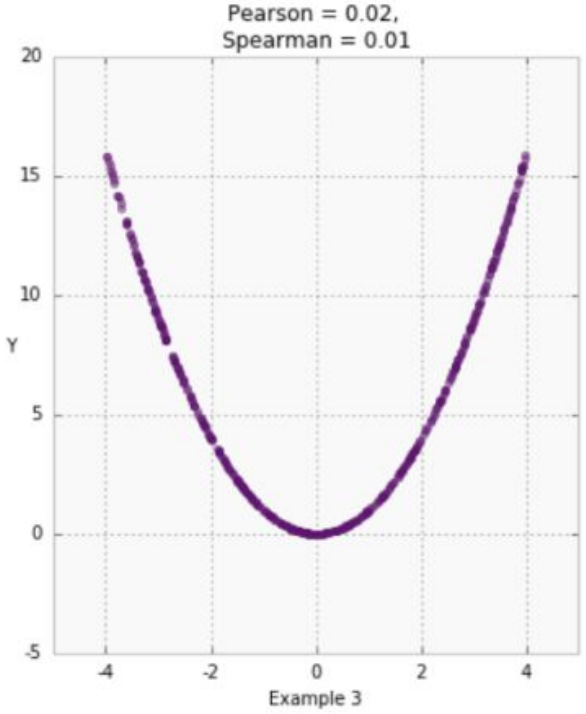
	A	B	C	D	E	F	G	H
1	Rank_mpg	Rank_cylinders	Rank_displacement	Rank_horsepower	Rank_weight	Rank_acceleration	Rank_model_year	Rank_origin
2	283	52	75	94	109	362.5	384	274
3	337.5	52	46.5	35.5	90	372	384	274
4	283	52	65	56.5	115	384	384	274
5	318	52	84	56.5	116	362.5	384	274
6	303	52	93	81	112	388	384	274
7	337.5	52	8	12.5	34	390.5	384	274
8	356	52	4	5	33	395	384	274
9	356	52	5.5	7	37	396.5	384	274
10	356	52	2	3	25	390.5	384	274
11	337.5	52	23	16	76	396.5	384	274
12	337.5	52	24.5	30	104	390.5	384	274
13	356	52	56	38.5	100	398	384	274
14	337.5	52	16	56.5	84	393.5	384	274
15	356	52	2	3	159	390.5	384	274
16	179	292.5	274	188.5	272	224.5	384	40
17	209.5	145.5	170	188.5	196	195	384	274
18	283	145.5	167.5	174	204	195	384	274
19	223.5	145.5	162.5	256	237	162.5	384	274
20	129	292.5	331	235	325.5	258	384	40
21	145.5	292.5	331	391.5	385.5	19	384	114.5
22	164	292.5	281	245.5	218	89.5	384	114.5
23	179	292.5	289	214.5	259	258	384	114.5
24	164	292.5	299	188.5	271	89.5	384	114.5
25	145.5	292.5	246	113	295	350.5	384	114.5
26	223.5	145.5	167.5	214.5	224	224.5	384	274
27	396.5	52	27.5	7	16	288.5	384	274
28	396.5	52	75	11	30	224.5	384	274

... Data | Covariance | Pearson | **Ranks** | Spearman | Sheet8

	A	B	C	D	E	F	G	H	I
1		*Rank_mpg*	*Rank_cylinders*	*Rank_displacement*	*Rank_horsepower*	*Rank_weight*	*Rank_acceleration*	*Rank_model_year*	*Rank_origin*
2	Rank_mpg	1							
3	Rank_cylinders	-0.821864491	1						
4	Rank_displacement	-0.855692012	0.911875915	1					
5	Rank_horsepower	-0.853320216	0.815689638	0.875770352	1				
6	Rank_weight	-0.874947398	0.873313559	0.945985564	0.878284909	1			
7	Rank_acceleration	0.43867748	-0.474189066	-0.496511921	-0.657631236	-0.404550372	1		
8	Rank_model_year	0.573468703	-0.335012387	-0.30525727	-0.389975332	-0.277014582	0.274632098	1	
9	Rank_origin	0.580693694	-0.604550452	-0.707196539	-0.509090776	-0.628434003	0.220573847	0.166551172	1
10									


Pearson = 0.80,
Spearman = 1
40
20
0
-20
-40
-60
-80
-100
-120
-140
-4
-2
0
2
4
Example 1


Pearson = 0.01,
Spearman = 0.25
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
Y
-1
0
1
2
3
4
5
Example 2


Pearson = 0.02,
Spearman = 0.01
20
15
10
5
0
-5
Y
-4
-2
0
2
4
Example 3

Least Squares

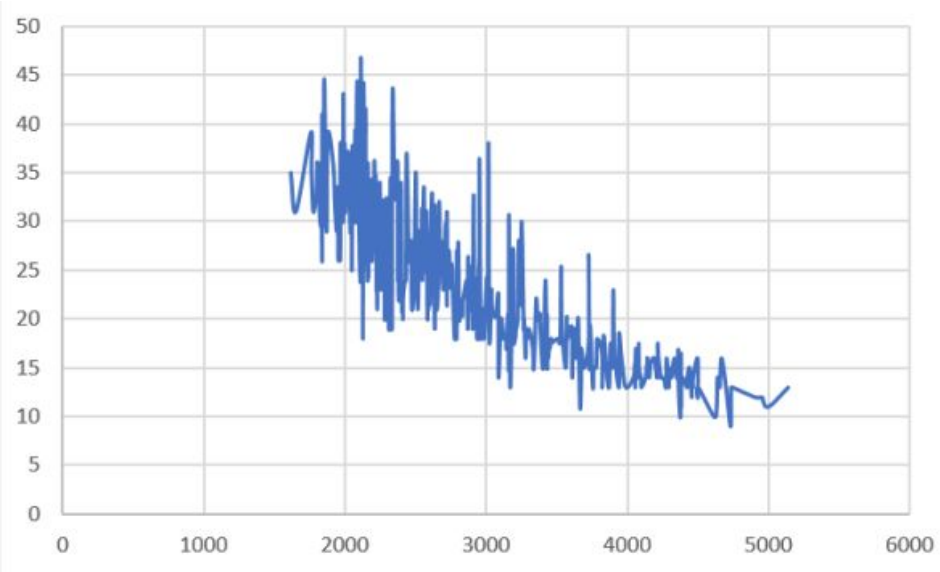
$$min(\sum_{i}(y' - y)^2)$$

Least Squares

1. Create a new table in a new sheet.
2. Copy the values of the weight and mpg columns.
3. Order the rows by the value of weight

	A	B
1	weight	mpg
2	1613	35
3	1649	31
4	1755	39.1
5	1760	35.1
6	1773	31
7	1795	33
8	1795	33
9	1800	36.1
10	1800	36.1
11	1825	29.5
12	1825	36
13	1834	27
14	1835	26
15	1835	40.9
16	1836	32
17	1845	29.8
18	1850	44.6
19	1867	29
20	1875	39
21	1915	35.7
22	1925	31.9
23	1937	29

Insert a line chart to see what the functional relationship looks like, as follows:



1. Assume that mpg = A*weight^(-b) and try to find the constants, a and b.
2. Create a new column, prediction, using the following formula: =H2*POWER([@weight];H3)

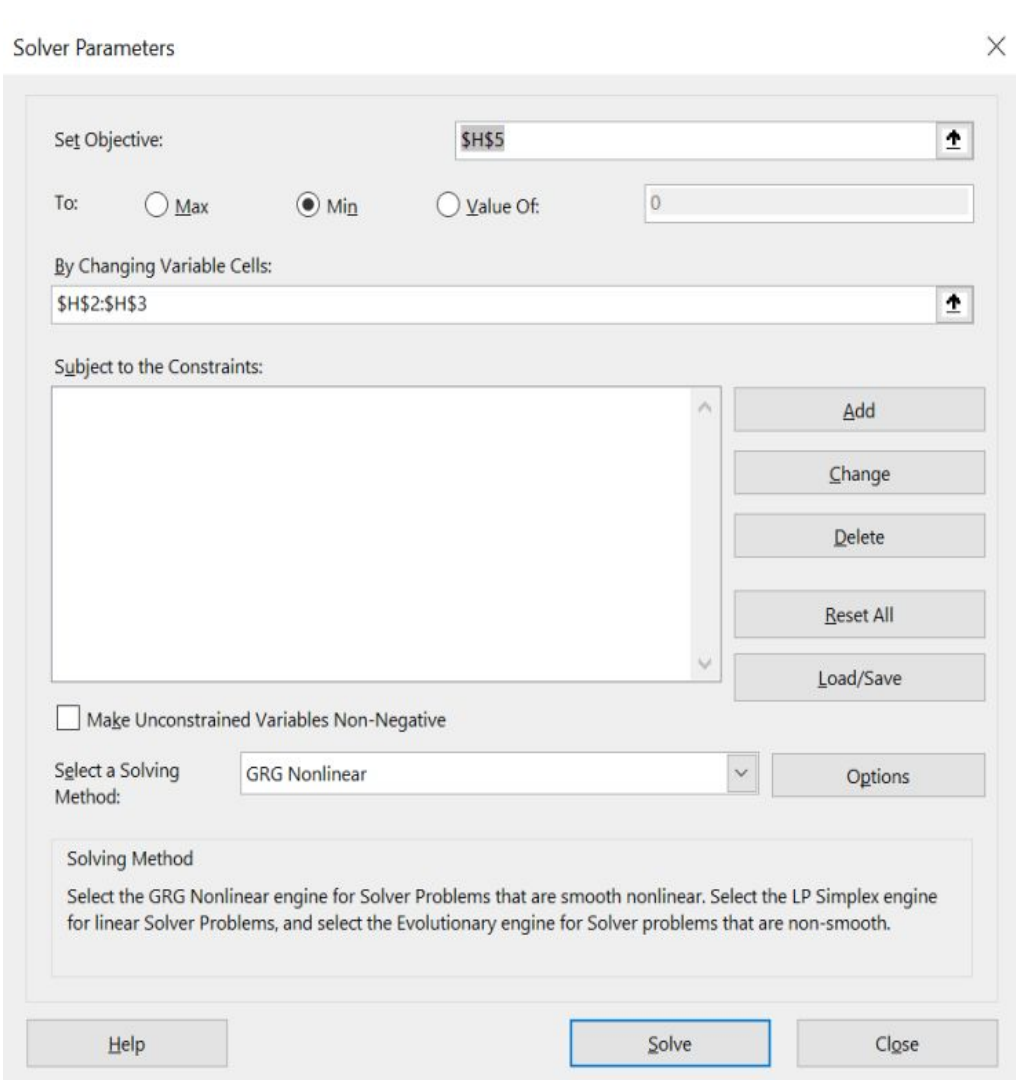
	A	B	C
1	weight	mpg	predict
2	1613	35	1.493943
3	1649	31	1.477546
4	1755	39.1	1.43223
5	1760	35.1	1.430194
6	1773	31	1.424941
7	1795	33	1.416182
8	1795	33	1.416182
9	1800	36.1	1.414214
10	1800	36.1	1.414214
11	1825	29.5	1.404494
12	1825	36	1.404494
13	1834	27	1.401043
14	1835	26	1.400662
15	1835	40.9	1.400662
16	1836	32	1.40028
17	1845	29.8	1.396861
18	1850	44.6	1.394972
19	1867	29	1.388606
20	1875	39	1.385641

- To fill the table, choose the initial values of a = 60 (in cell H2) and b = -0.5 (in cell H3).
  - These will be the starting points of the least squares method.
- The quantity to minimize is the sum of the squares of the errors.
  - To calculate it, we create a new column, Squared error, with the following formula: **=([@mpg]-[@prediction])^2**

Then, use the following formula to sum all the values in that column in a cell:

=SUM(Table9[Squared error])

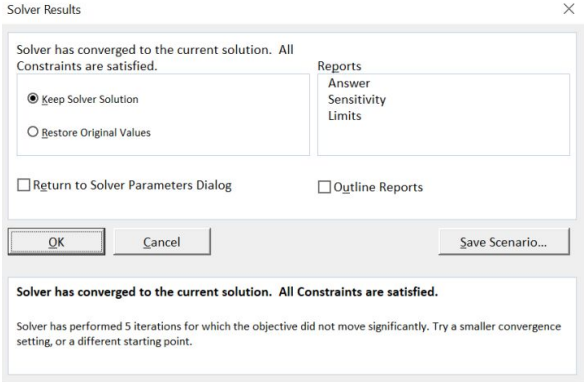
Navigate to Data | Solver



The Set Objective option is filled with the cell ID where we calculated the sum of the squared errors, and the By Changing Variable Cells option is filled with the ID of the two cells containing the values of a and b.

We can leave the rest of the parameters as their default value settings.

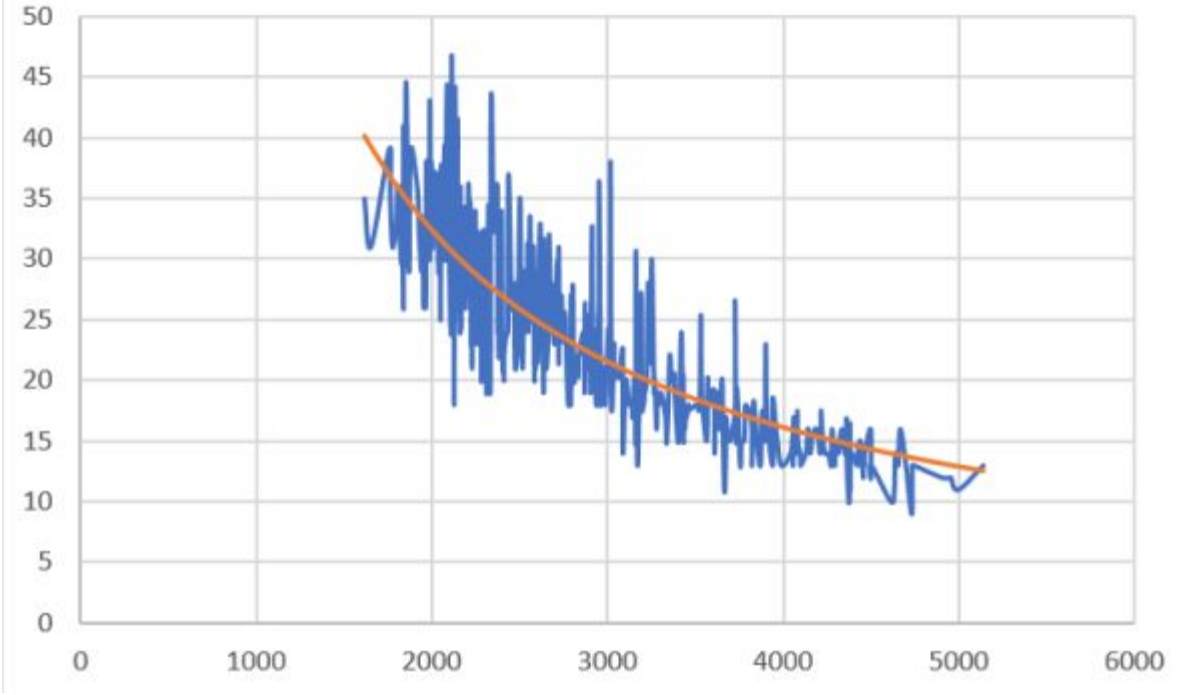
Click on Solve; if the regression converges, then you will see the following window:



Choose Keep Solver Solution to replace the values of a and b by the ones calculated, and get new values for all predictions.

If we include the real values and the prediction in the same diagram, you should see something similar to the following screenshot:

	a	68563.9126
	b	-1.0074493
	SSE	7117.53396



What method would be better to find a correlation between a numerical and a categorical variable?

Build some other plot graphs between a pair of variables and study the correlations and the logic behind them.

Does a negative Pearson coefficient value imply that one of the variables has negative values?

The table of the Pearson's coefficient can be colored or have bars added to it in order to better compare the different values. Explore these options in Quick Analysis | Formatting.

The quality of the least square regression is usually measured by the value of R2. Calculate this value for the function that was adjusted in the mpg column versus the weight data value (hint: you only need to calculate one more sum of values – refer to the literature for more information).

The value that was calculated in the previous question should be close to 0.7, which is not good enough to prove that the function reproduces the data well. Try a different function and see what the result is.