NYU Center for Data Science

DS-UA 112
Introduction to Data Science

Lecture 9

Visualization II – matplotlib and seaborn

Reminders

- Survey 2
  - Please complete by October 7
- Forum
  - Please post questions about Homework and Projects

Reminders

- Homework 2
  - Please submit Friday October 4
- Project 1
  - Please submit Sunday October 20

Reminders

- Assignment Submission
  - Leave the .ipynb file on JupyterHub.
  - Export notebook to html
  - Print html to pdf
  - Upload pdf to Gradescope
  - Mark responses

Agenda

- Review
  - How to work with file formats of different sizes?

Agenda

- Lesson
  - Find effective visualizations for different types of data.

Agenda

- Demo
  - Congressional Committees
  - Police Reports

Flat Files

csv

```
Candidate,Party,%,Year,Result
Reagan,Republican,50.7,1980,win
Carter,Democratic,41,1980,loss
Anderson,Independent,6.6,1980,loss
Reagan,Republican,58.8,1984,win
Mondale,Democratic,37.6,1984,loss
Bush,Republican,53.4,1988,win
Dukakis,Democratic,45.6,1988,loss
Clinton,Democratic,43,1992,win
Bush,Republican,37.4,1992,loss
Perot,Independent,18.9,1992,loss
Clinton,Democratic,49.2,1996,win
Dole,Republican,40.7,1996,loss
Perot,Independent,8.4,1996,loss
Gore,Democratic,48.4,2000,loss
Bush,Republican,47.9,2000,win
Kerry,Democratic,48.3,2004,loss
Bush,Republican,50.7,2004,win
Obama,Democratic,52.9,2008,win
McCain,Republican,45.7,2008,loss
Obama,Democratic,51.1,2012,win
Romney,Republican,47.2,2012,loss
Clinton,Democratic,48.2,2016,loss
Trump,Republican,46.1,2016,win
```

tsv

```
Candidate       Party   %       Year    Result
Reagan  Republican      50.7    1980    win
Carter  Democratic      41.0    1980    loss
Anderson        Independent     6.6     1980    loss
Reagan  Republican      58.8    1984    win
Mondale Democratic      37.6    1984    loss
Bush    Republican      53.4    1988    win
Dukakis Democratic      45.6    1988    loss
Clinton Democratic      43.0    1992    win
Bush    Republican      37.4    1992    loss
Perot   Independent     18.9    1992    loss
Clinton Democratic      49.2    1996    win
Dole    Republican      40.7    1996    loss
Perot   Independent     8.4     1996    loss
Gore    Democratic      48.4    2000    loss
Bush    Republican      47.9    2000    win
Kerry   Democratic      48.3    2004    loss
Bush    Republican      50.7    2004    win
Obama   Democratic      52.9    2008    win
McCain  Republican      45.7    2008    loss
Obama   Democratic      51.1    2012    win
Romney  Republican      47.2    2012    loss
Clinton Democratic      48.2    2016    loss
Trump   Republican      46.1    2016    win
```

Nested Files

XML	JSON	YAML

<Servers>
<Server>
<name>Server1</name>
<owner>John</owner>
<created>123456</created>
<status>active</status>
</Server>
</Servers>
	{
	Servers: [
	{
	name: Server1,
	owner: John,
	created: 123456,
	status: active
	}
	]
	}
		Servers:
		- name: Server1
		owner: John
		created: 123456
		status: active

Unstructured Files

File Size

Multiple	Notation	Number of Bytes
Kibibyte	KiB	$1024 = 2^{10}$
Mebibyte	MiB	$1024^2 = 2^{20}$
Gibibyte	GiB	$1024^3 = 2^{30}$
Tebibyte	TiB	$1024^4 = 2^{40}$
Pebibyte	PiB	$1024^5 = 2^{50}$

For example, a file containing 52428800 characters takes up 52428800 bytes = 50 mebibytes = 50 MiB on disk.

File Size

- When to read file?
  - pandas requires double the file size in available memory
  - **Example:** Reading in a 1 GiB file will typically require at least 2 GiB of available memory.
- How can we determine the file size before reading it?
  - Shell Interpreter
  - Command-line interface (CLI)

File Size

- When to read file?
  - pandas requires double the file size in available memory
  - **Example:** Reading in a 1 GiB file will typically require at least 2 GiB of available memory.

File Size: ls command

```
!ls
```

```
data  ds-ua-112-lab04.ipynb  movies_100_rows.csv  movies.csv
```

File Size: head, tail, cat commands

```
!head movies.csv
```

```
director,genre,movie,rating,revenue
David,Action & Adventure,Deadpool 2,7,318344544
Bill,Comedy,Book Club,5,68566296
Ron,Science Fiction & Fantasy,Solo: A Star Wars Story,6,213476293
Baltasar,Drama,Adrift,6,31445012
Bart,Drama,American Animals,6,2847319
Gary,Action & Adventure,Oceans 8,6,138803463
Drew,Action & Adventure,Hotel Artemis,8,6708147
Brad,Animation,Incredibles 2,5,594398019
Jeff,Comedy,Tag,6,54336863
```

File Size: head, tail, cat commands

```
!tail movies.csv
```

```
Jeff,Comedy,Tag,6,54336863
J.A.,Science Fiction & Fantasy,Jurassic World: Fallen Kingdom,6,411873505
Charles,Comedy,Uncle Drew,5,42201656
Gerard,Horror,The First Purge,7,68765655
Peyton,Action & Adventure,Ant-Man and the Wasp,5,208681866
Genndy,Animation,Hotel Transylvania 3: Summer Vacation,5,154418311
Rawson,Action & Adventure,Skyscraper,6,66801215
Ol,Comedy,Mamma Mia! Here We Go Again,8,111705055
Christopher,Action & Adventure,Mission: Impossible-Fallout,6,182080372
Marc,Comedy,Christopher Robbin,6,6786317
```

File Size: head, tail, cat commands

```
!cat movies_100_rows.csv
```

```
director,genre,movie,rating,revenue
David,Action & Adventure,Deadpool 2,7,318344544
Bill,Comedy,Book Club,5,68566296
Ron,Science Fiction & Fantasy,Solo: A Star Wars Story,6,213476293
Baltasar,Drama,Adrift,6,31445012
Bart,Drama,American Animals,6,2847319
Gary,Action & Adventure,Oceans 8,6,138803463
Drew,Action & Adventure,Hotel Artemis,8,6708147
Brad,Animation,Incredibles 2,5,594398019
Jeff,Comedy,Tag,6,54336863
```

File Size: du command

```
!ls -lh
```

```
total 44K
drwxrwxr-x+ 4             4.0K Sep 30 14:22 data
-rwxrwxr--+ 1              29K Sep 30 14:23 ds-ua-112-lab04.ipynb
-rw-rw-r--+ 1              415 Sep 30 13:58 movies_100_rows.csv
-rwxrwxr--+ 1              903 Sep 25 22:57 movies.csv
```

```
!du -sh data
```

```
28K     data
```

```
!du -sh data/*
```

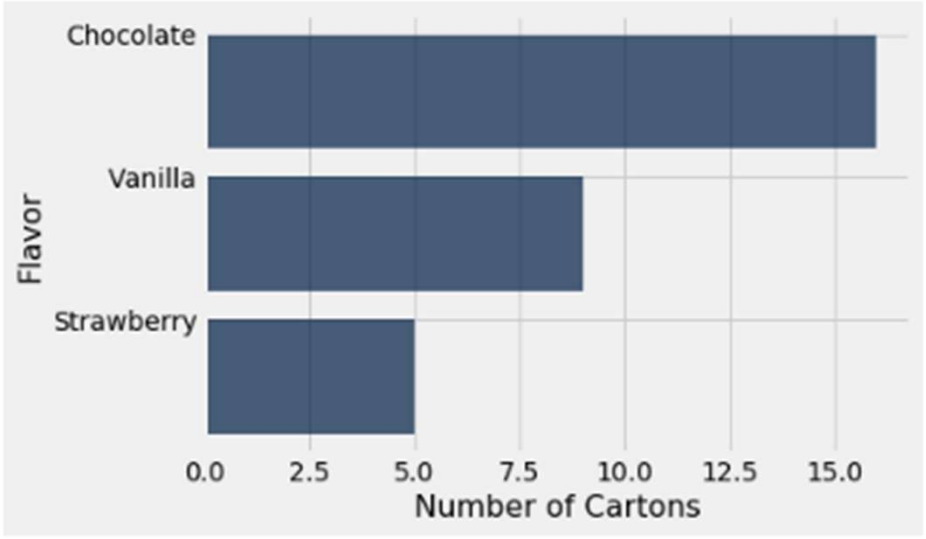
```
12K     data/more_data
4.0K    data/movies_100_rows.csv
4.0K    data/movies.csv
```

Visualization

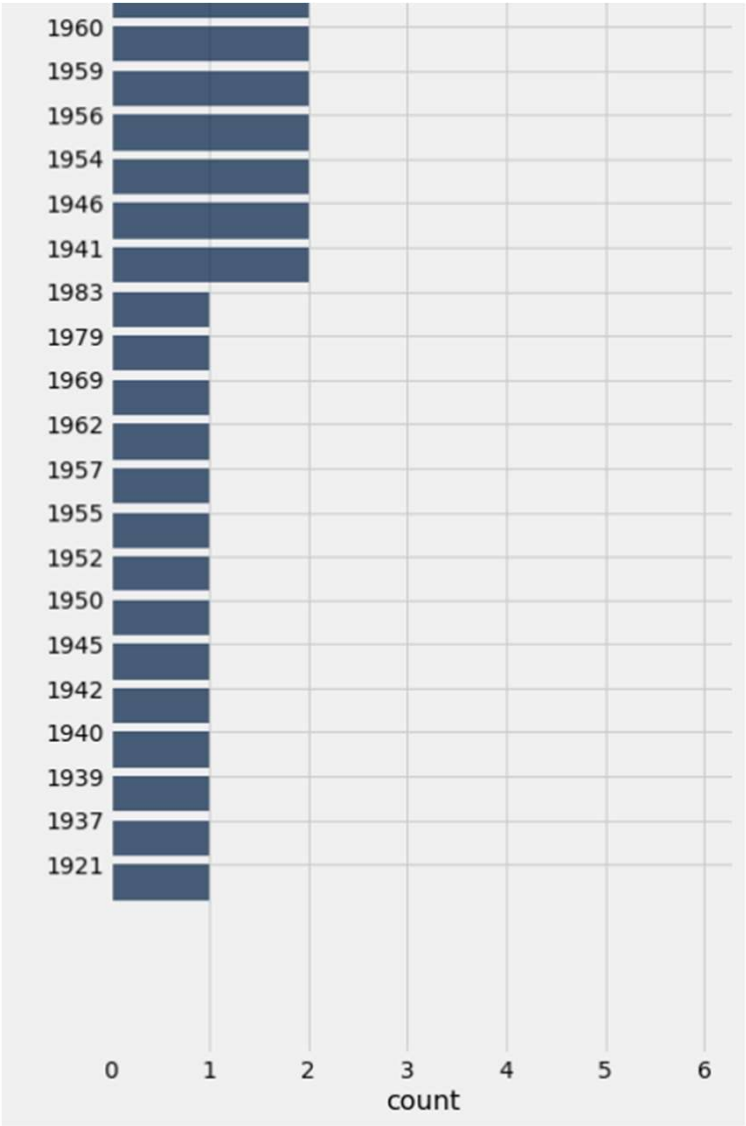
- Bar Chart
- Histogram
- Box-plot
- Scatter-plot
- Heat Map

Bar Chart

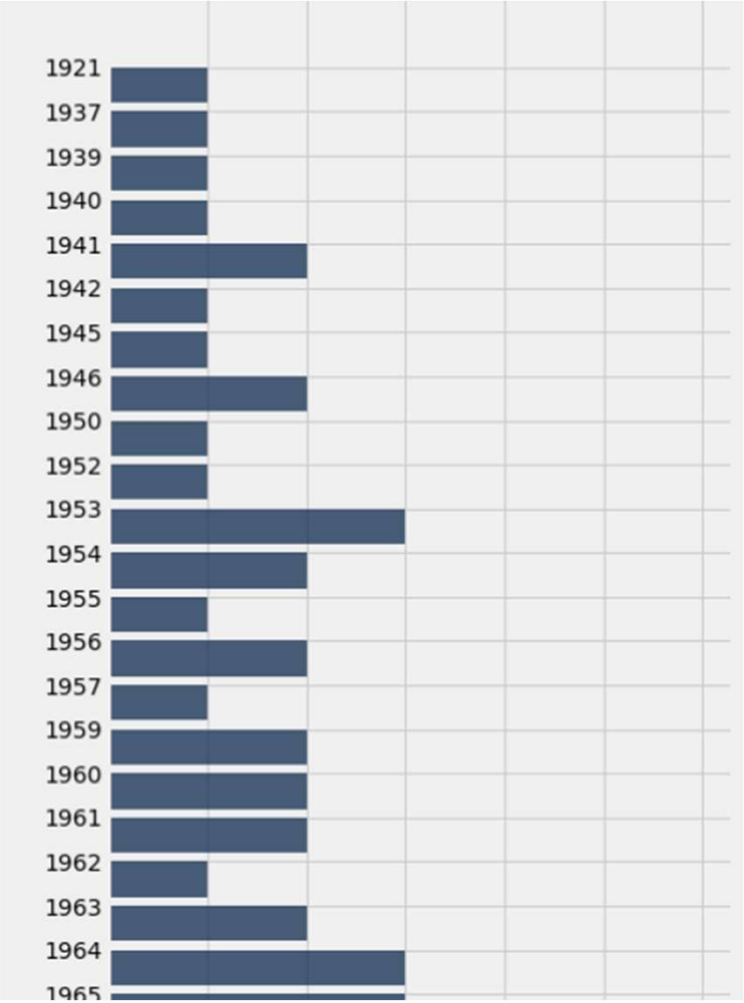
- Height indicates Count for Category



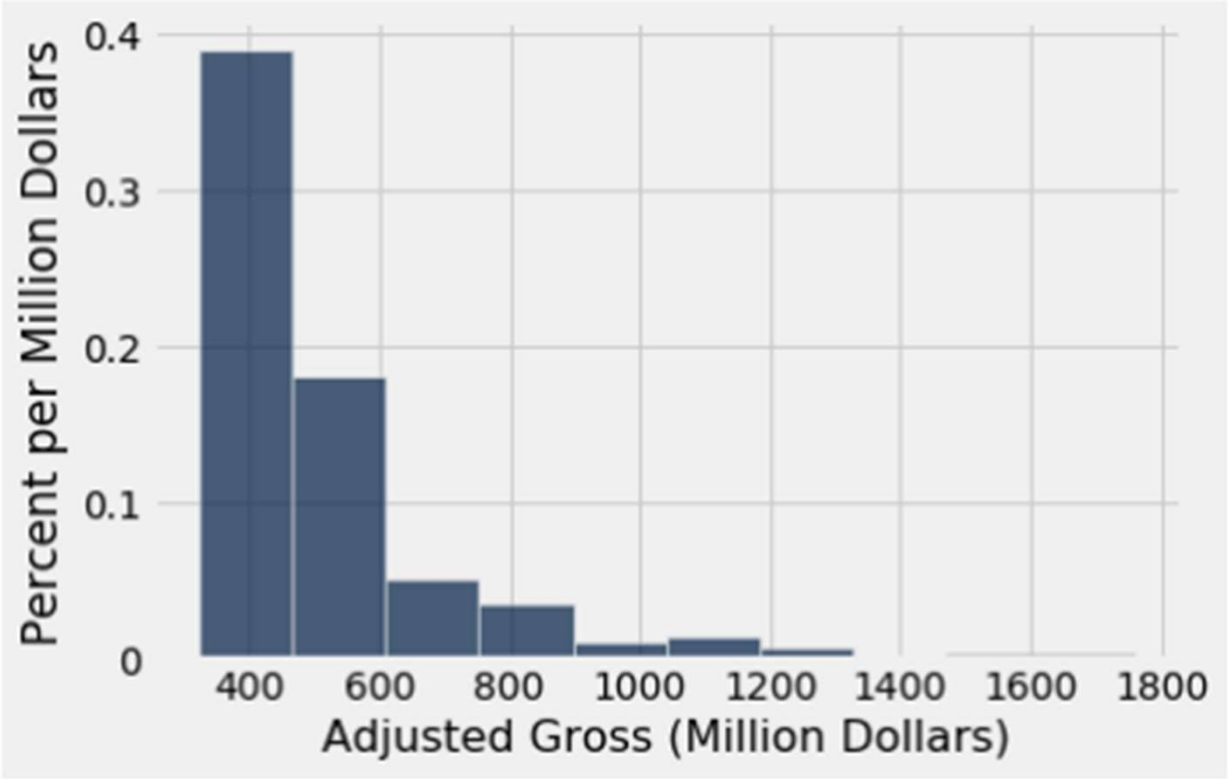
Histogram


1960
1959
1956
1954
1946
1941
1983
1979
1969
1962
1957
1955
1952
1950
1945
1942
1940
1939
1937
1921
0
1
2
3
4
5
6
count

Histogram

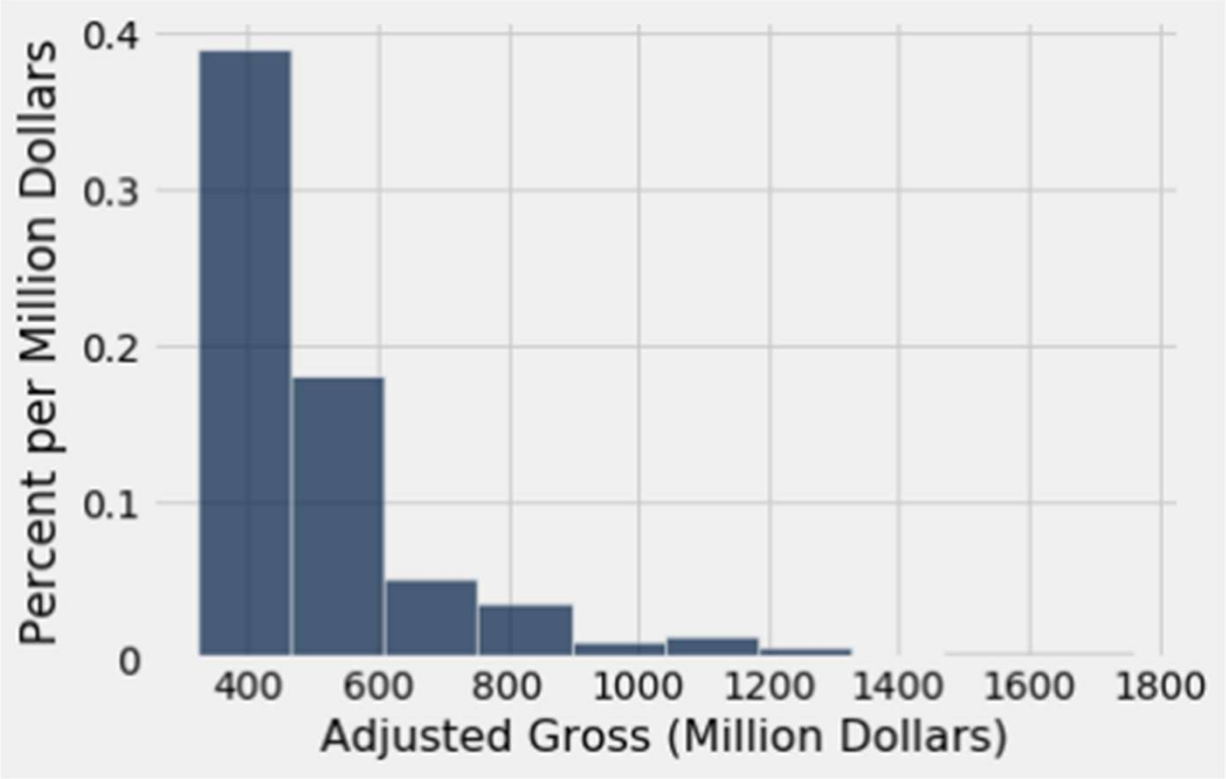

1921
1937
1939
1940
1941
1942
1945
1946
1950
1952
1953
1954
1955
1956
1957
1959
1960
1961
1962
1963
1964
1965

Histogram


Percent per Million Dollars
0.4
0.3
0.2
0.1
0
400
600
800
1000
1200
1400
1600
1800
Adjusted Gross (Million Dollars)

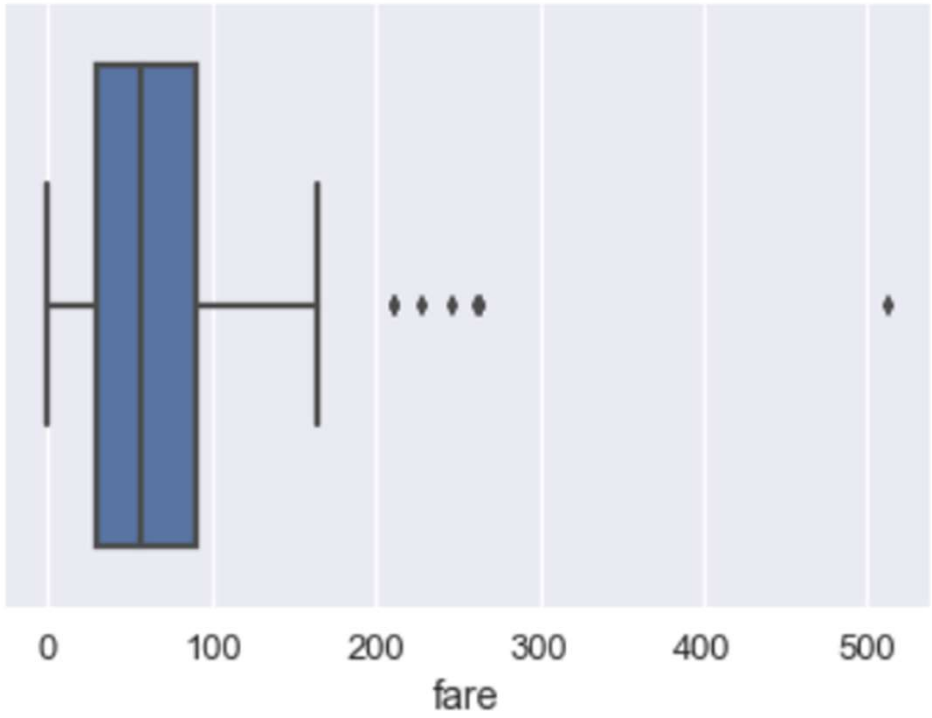
Histogram

bin	Count	Percent	Height
300	81	40.5	0.405
400	52	26	0.26
500	28	14	0.14
600	16	8	0.08
700	7	3.5	0.035
800	5	2.5	0.025
900	3	1.5	0.015
1000	1	0.5	0.005
1100	3	1.5	0.015
1200	2	1	0.01

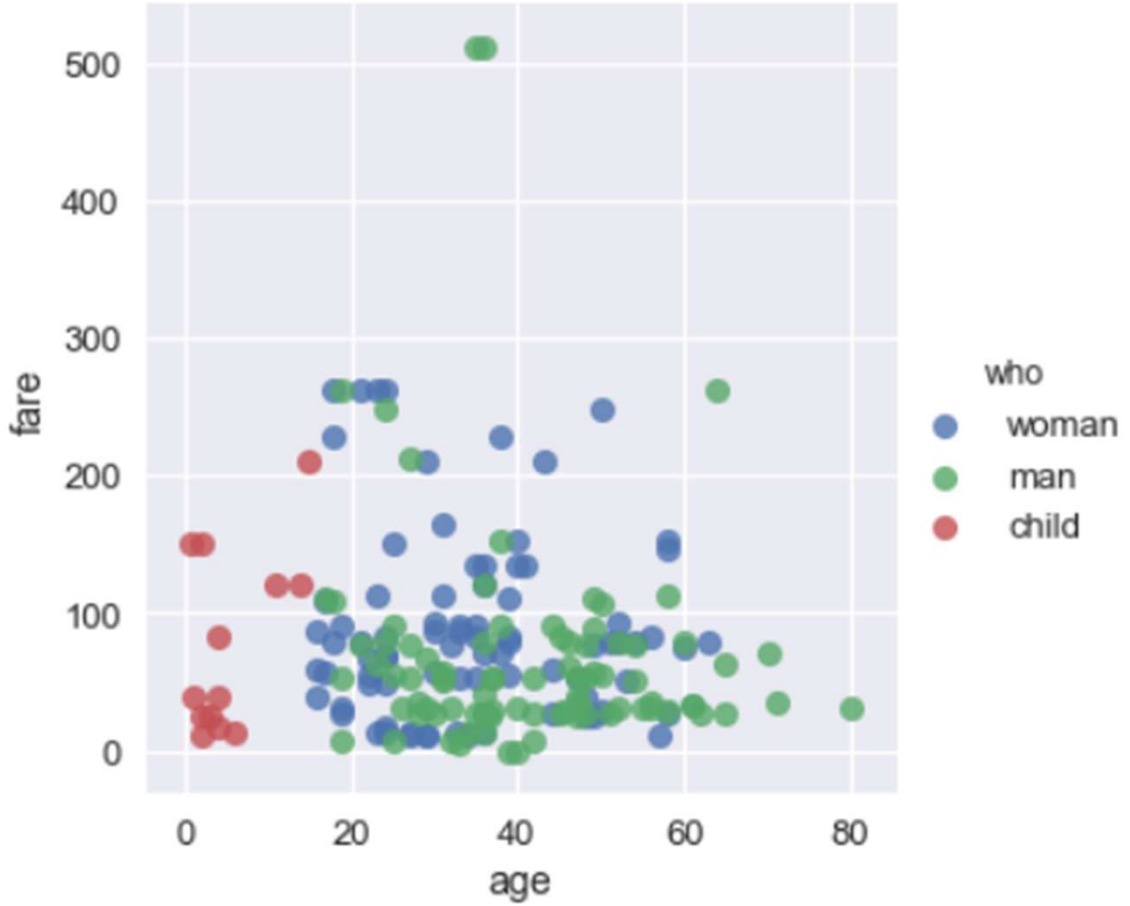


Boxplot

- Median
- Inter-Quartile Range
- Outlier



Scatter Plot



Heat Map

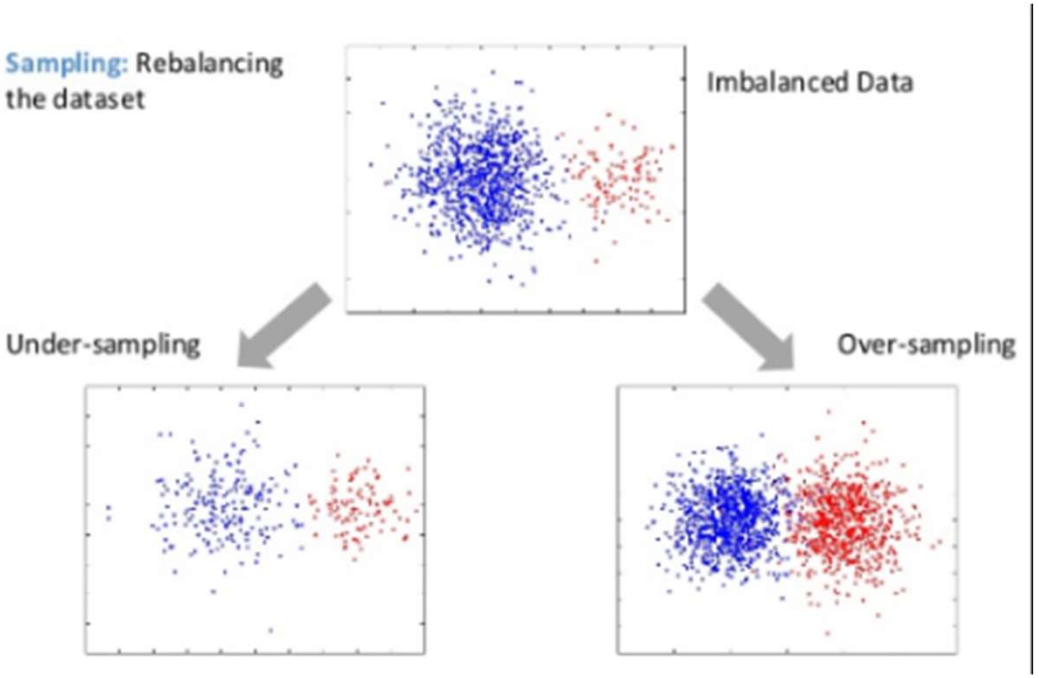
- Color indicates intensity

Visualization and Data Collection

- Bias
  - Avoid it
  - Adjust it
  - Expect it

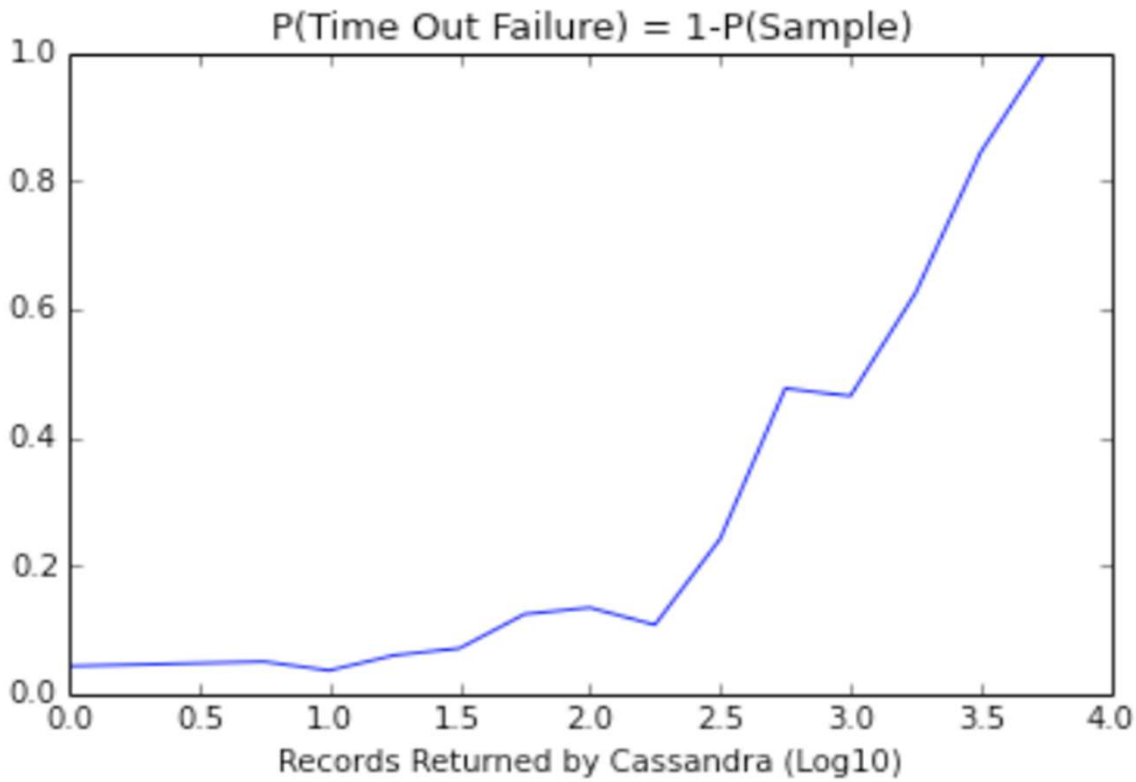
Selection Bias

- Stratified Sampling
  - Equal sampling
  - Undersampling
  - Oversampling



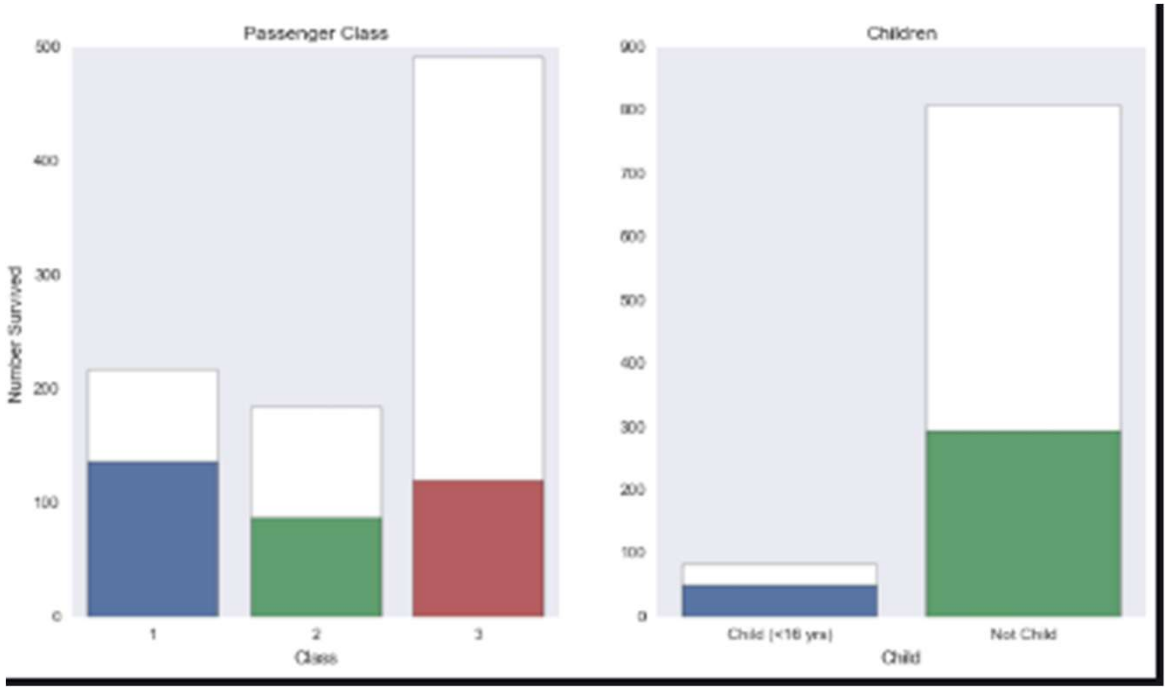
Survivor Bias

- Only some samples “survivor” for collection in the analysis
- Try to use Bayes Rule to Adjust for Confounding Factor



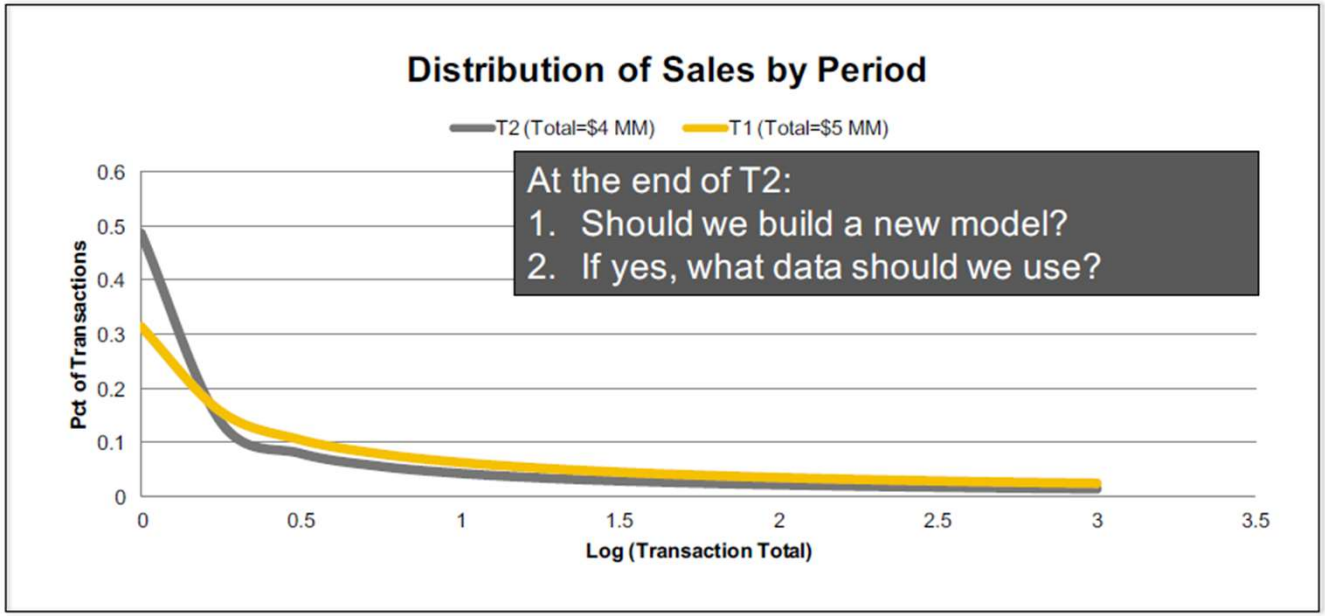
Exclusion Bias

- Discarding Relevant Information
- Investigate Before Throwing Out Data!



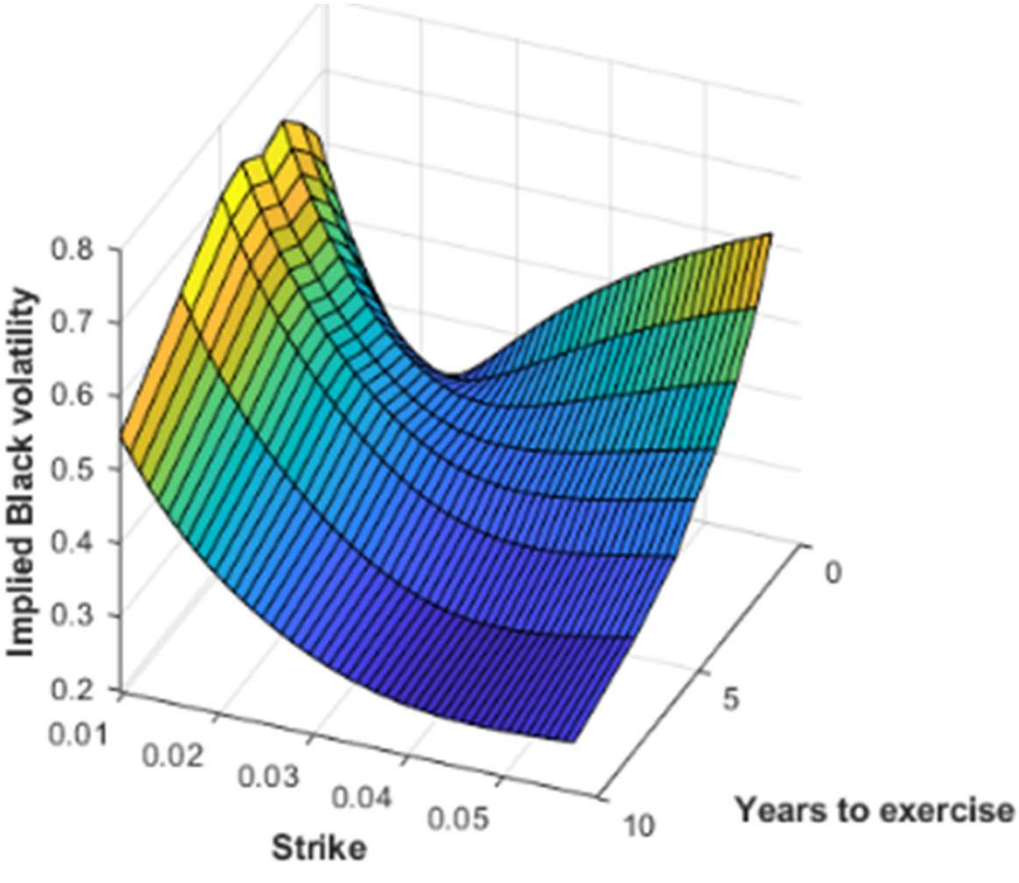
Seasonality

- Trends change over Time



Feed-Back Bias

- Decisions based on a model can impact the observations used to derive it!



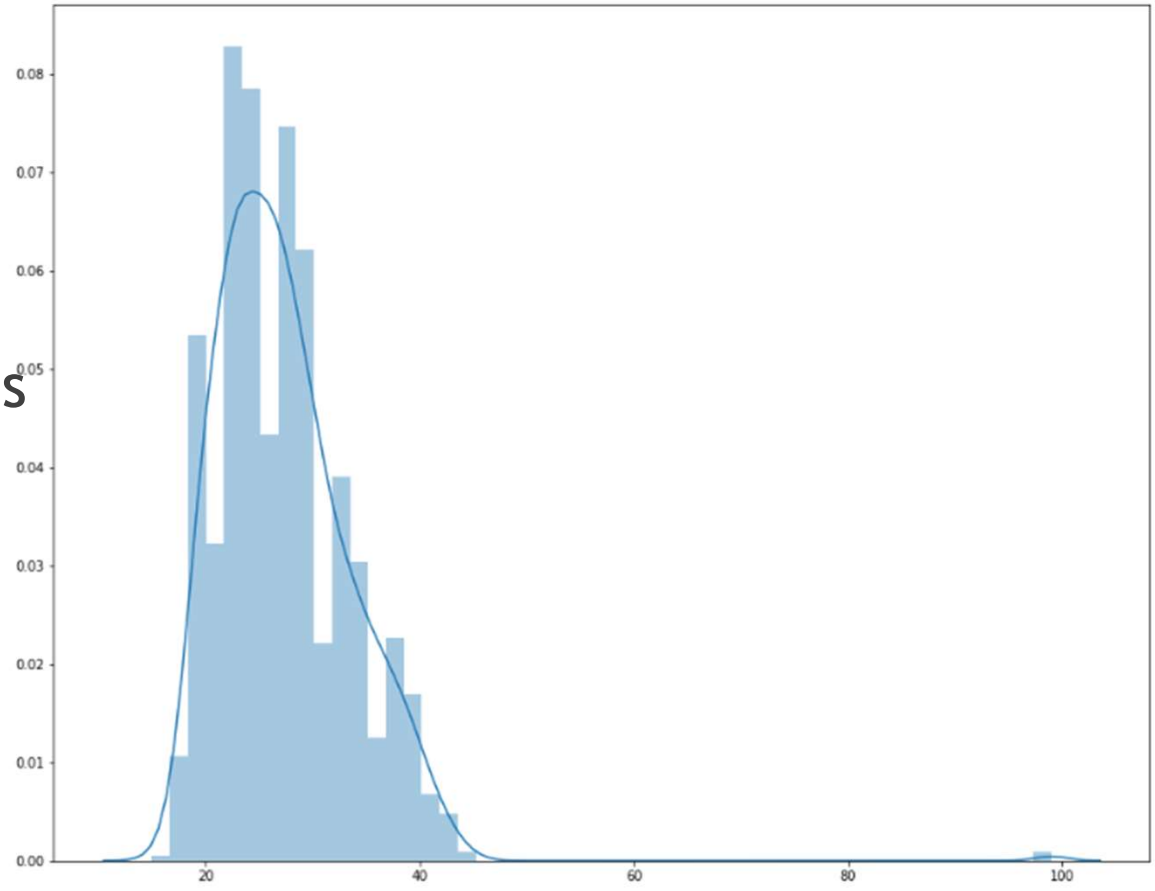
Visualization and Data Cleaning

- When to Add Data?
- When to Subtract Data?
- When to Modify Data?

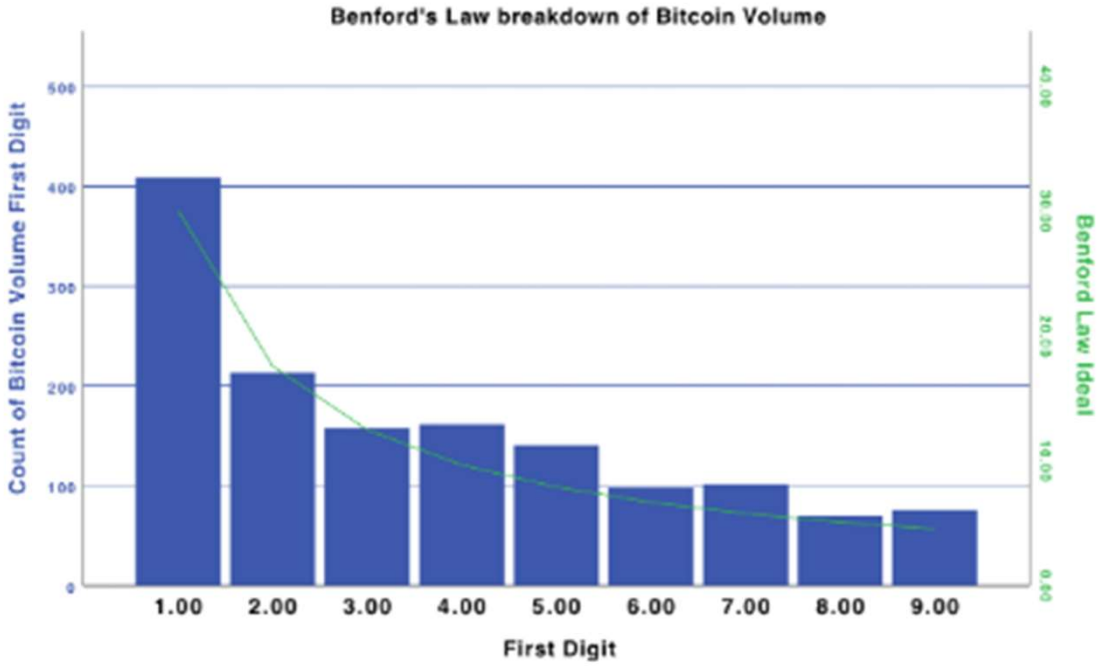
When to Add Data?

- Missing
  - Use sentinel value NaN

  Instead of numerical values
  - Drop
  - Fill
  - Impute



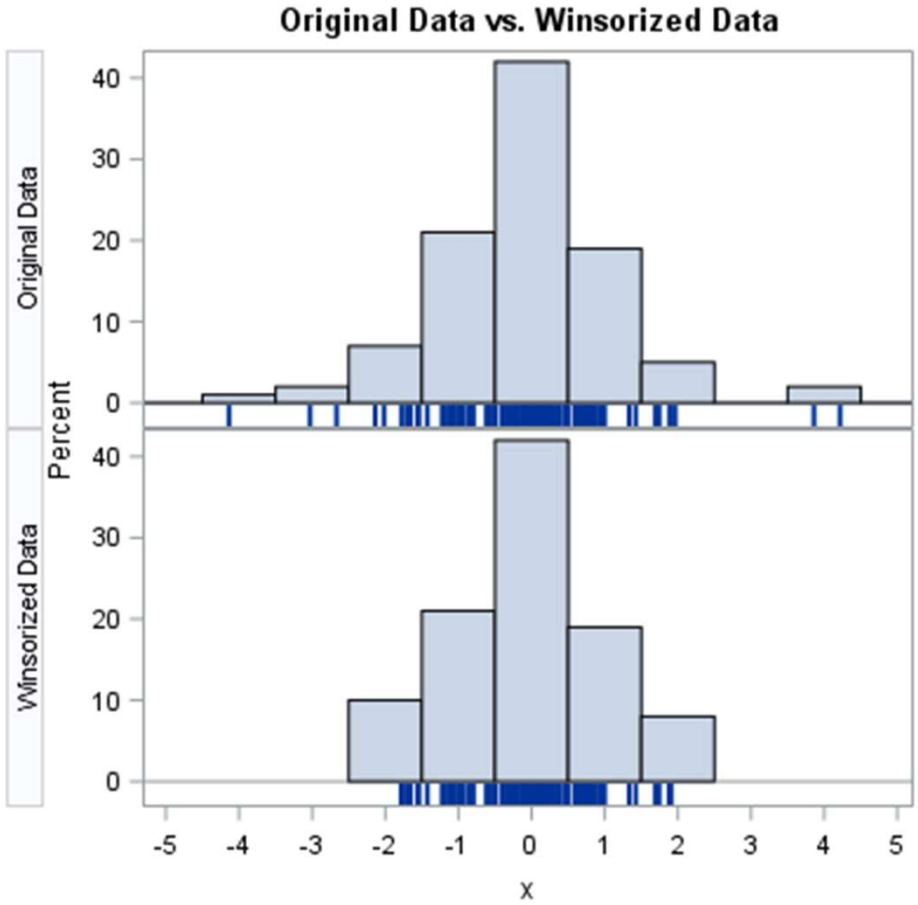
When to Subtract Data?



- When to Subtract Data?
  - Spurious
    - Falsified numbers
    - Typos from manual editing

When to Change Data?

- Duplicates
  - Repeated Values
  - Highly Correlated Values
- Inconsistencies
  - Formatting (e.g. dates)
  - Scales
- Outliers
  - Is it a bug or a feature?!



Take-Aways

- File Size
  - kibi, mebi, gibi, tebi
  - ls, du, head, tail
- JSON Format
  - Access
  - Convert

Take-Aways

- Visualization
  - Bar Chart
  - Histogram
  - Box-plot
  - Scatter-plot
  - Heat Map
- Visualization for Data Collection/Cleaning