ISA 419: Data Driven Security

08 - Descriptive Analytics (Cont.)

Fadel M. Megahed

Endres Associate Professor
Department of Information Systems and Analytics
Farmer School of Business
Miami University
Email: fmegahed@miamioh.edu
Office Hours: Automated Scheduler for Office Hours

Fall 2022

Outline

Our Course Progress so Far

Learning Objectives Discussed

- ☑ **Define information security and its main goals**
- ☑ **Understand that breaches are frequent and target different industries**
- ☑ **Understand the basic operations in either R or Python**
- ☑ **Describe the data analytics process**
- ☑ **Describe/Convert ETL data to be both technically correct and consistent**
- ☑ **Summarize and examine data through ETL operations**

Security Datasets that you have Examined So Far

- Have I been pawned?
- Los Alamos National Lab's User Authentication Dataset
- NSL-KDD Cup 1999 Intrusion Dataset
- Allien Vault IP Reputation Dataset

Learning Objectives for Today's Class

- Describe impact of data types on how we summarize them
- Intro to Data Encoding and Visualizations

Outline

On the Theory of Scales of Measurement

SCIENCE

Vol. 103, No. 2684 Friday, June 7, 1946

On the Theory of Scales of Measurement

S. S. Stevens

Director, Psycho-Acoustic Laboratory, Harvard University

FOR SEVEN YEARS A COMMITTEE of the British Association for the Advancement of Science debated the problem of measurement. Appointed in 1932 to represent Section A (Mathematical and Physical Sciences) and Section J (Psychology), the committee was instructed to consider and report upon the possibility of "quantitative estimates of sensory events"—meaning simply: Is it possible to measure human sensation? Deliberation led only to disagreement, mainly about what is meant by the term measurement. An interim report in 1938 found one member complaining that his colleagues "came out by that same door as they went in," and in order to have another try at agreement, the committee begged to be continued for another year.

For its final report (1940) the committee chose a common bone for its contentions, directing its arguments at a concrete example of a sensory scale. This was the Sone scale of loudness (S. S. Stevens and H. Davis. *Hearing*. New York: Wiley, 1938), which purports to measure the subjective magnitude of an auditory sensation against a scale having the formal by the formal (mathematical) properties of the scales. Furthermore—and this is of great concern to several of the sciences—the statistical manipulations that can legitimately be applied to empirical data depend upon the type of scale against which the data are ordered.

A Classification of Scales of Measurement

Paraphrasing N. R. Campbell (Final Report, p. 340), we may say that measurement, in the broadest sense, is defined as the assignment of numerals to objects or events according to rules. The fact that numerals can be assigned under different rules leads to different kinds of scales and different kinds of measurement. The problem then becomes that of making explicit (a) the various rules for the assignment of numerals, (b) the mathematical properties (or group structure) of the resulting scales, and (c) the statistical operations applicable to measurements made with each type of scale.

Scales are possible in the first place only because there is a certain isomorphism between what we can do with the aspects of objects and the properties of

Data Types (from S. Stevens, Theory of Scales)

Scale	Basic Empirical Operations	Mathematical Group Structure	Permissible Statistics (invariantive)
NOMINAL	Determination of equality	*Permutation group* $x' = f(x)$ $f(x)$ means any one-to-one substitution	Number of cases Mode Contingency correlation
ORDINAL	Determination of greater or less	*Isotonic group* $x' = f(x)$ $f(x)$ means any monotonic increasing function	Median Percentiles
INTERVAL	Determination of equality of intervals or differences	*General linear group* $x' = ax + b$	Mean Standard deviation Rank-order correlation Product-moment correlation
RATIO	Determination of equality of ratios	*Similarity group* $x' = ax$	Coefficient of variation

Nnominal, ordinal, or numeric for Portmap.csv?

```
## Rows: 1,000
## Columns: 5
## $ FWD   <int> 45, 56, 6, 6, 6, 2, 2, 2, 5, 40, 6, 6, 6, 4, 5, 4, 1, 2, 2, 3
## $ Back  <int> 0, 0, 2, 2, 2, 0, 0, 0, 0, 40, 2, 2, 2, 0, 0, 0, 3, 0, 0, 0,
## $ PL    <dbl> 0, 0, 46, 46, 46, 6, 6, 6, 300, 0, 46, 46, 46, 46, 46, 39, 0
## $ PS    <dbl> 0.00, 0.00, 31.75, 31.75, 31.75, 9.00, 9.00, 9.00, 360.00, 0
## $ Label <fct> BENIGN, BENIGN, BENIGN, BENIGN, BENIGN, BENIGN, BENIGN, BENIG
```

Outline

Objectives

Our primary goal is to develop a better understanding of the data. The following guidelines from Grolemund and Wickham (2016) can be useful:

- Use questions as tools to guide your investigation.
  - When you ask a question, it focuses your attention on a specific part of your dataset and helps you decide which graphs, models, or transformations to make.
- EDA is fundamentally a creative process, where you should ask quality questions through asking a large quantity of questions.
- There is no rule about which questions you should ask to guide your research. However, two types of questions will always be useful for making discoveries within your data.
  - What type of variation occurs within my variables?
  - What type of covariation occurs between my variables?

Implementation Strategies

In my estimation, the EDA process typically combines:

- **Descriptive statistics/modeling,** where statistical techniques for summarizing and describing the variation within the data are used.
- **Visualization techniques,** charting the data allows us to get a different perspective of the data.

A Synthetic Example: The Anscombe Dataset [1]

In a seminal paper, Anscombe (1973) stated: *Few of us escape being indoctrinated with these notions*

- *numerical calculations are exact, but graphs are rough;*
- *for any particular kind of statistical data there is just one set of calculations constituting a correct statistical analysis;*
- *performing intricate calculations is virtuous, whereas actually looking at the data is cheating.*

He proceeded by stating that *a computer should make both calculations and graphs. Both sorts of output should be studied; each will contribute to understanding.*

Now, let us consider his four datasets, each consisting of eleven (x,y) pairs.

A Synthetic Example: The Anscombe Dataset [2]

```
anscombe %>% xtable::xtable() -> xt
align(xt) <- "lrrrrrrrr"

cat("---")
```

```
## ---
```

```
print(xt, include.rownames=F)
```

% latex table generated in R 4.2.1 by xtable 1.8-4 package % Sun Sep 18 20:17:20 2022

A Synthetic Example: The Anscombe Dataset [3]

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

A Synthetic Example: The Anscombe Dataset [4]

```
pacman::p_load(Tmisc) # same data but in 3 columns
df = quartet
df %>% group_by(set) %>%
  summarise(x.mean = mean(x), x.sd = sd(x),
            y.mean = mean(y), y.sd = sd(y),
            corr = cor(x, y)) %>% kable(digits = 3)
```

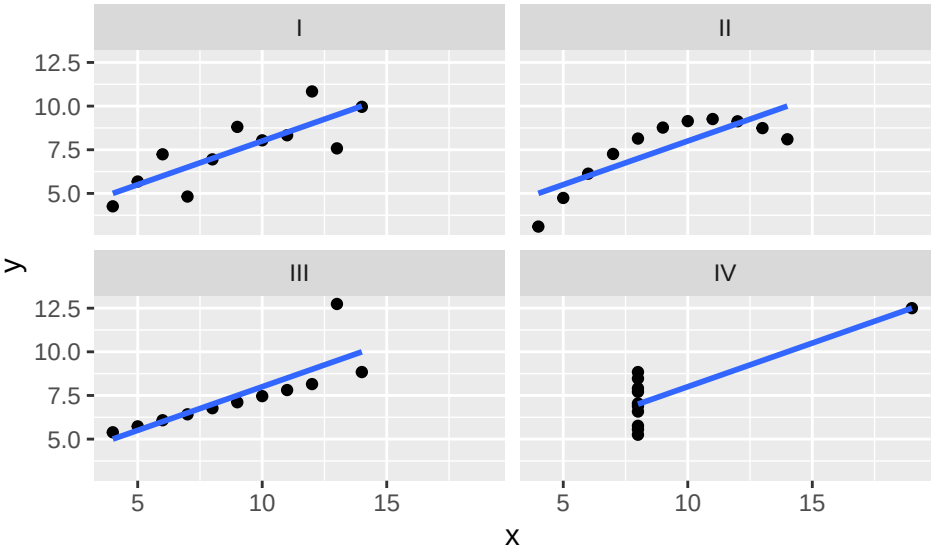
set	x.mean	x.sd	y.mean	y.sd	corr
I	9	3.317	7.501	2.032	0.816
II	9	3.317	7.501	2.032	0.816
III	9	3.317	7.500	2.030	0.816
IV	9	3.317	7.501	2.031	0.817

The code in this slide and next is based on https://rpubs.com/turnersd/anscombe.

A Synthetic Example: The Anscombe Dataset [5]

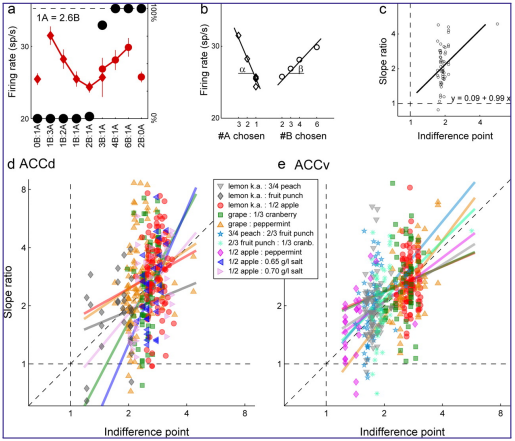
```
ggplot(quartet, aes(x, y)) + geom_point() +
  geom_smooth(method = lm, se = FALSE) + facet_wrap(~set)
```

A Synthetic Example: The Anscombe Dataset [6]



Anscombe-Like Mistakes in Research and Practice

In my estimation, Figure 8c from Cai and Padoa-Schioppa (2012) represents an example where regression should not have been performed.



Some Useful Statistical Summaries [1]

In lab 01 we have described several variables, you may be interested in performing some statistical summaries for a grouping variable.

```
pacman::p_load(magrittr)
df = read.csv("../../data/Portmap.csv", nrow=1000, stringsAsFactors = T)
df %<>% select('Total.Fwd.Packets', 'Total.Backward.Packets',
               'Max.Packet.Length', 'Average.Packet.Size', 'Label')
colnames(df) = c("FWD", "Back", "PL", "PS", "Label")

df %>% group_by(Label) %>%
  summarise_all(list(med= median, avg = mean, sd= sd)) %>%
  kable() %>%
  kable_styling(position = "center", latex_options = "scale_down")
```

Some Useful Statistical Summaries [2]

Label	FWD_med	Back_med	PL_med	PS_med	FWD_avg	Back_avg	PL_avg	PS_avg	FWD_sd	Back_sd	PL_sd	PS_sd
BENIGN	2	2	55	51.75	7.465649	7.032715	290.2388	77.45095	16.29907	21.08972	677.0226	126.03066
Portmap	4	2	6	7.00	8.746988	8.674699	122.5904	18.03514	12.94815	14.09905	319.4344	37.16719

Some Useful Statistical Tests

For a more detailed statistical analysis of your data. I recommend the table in https://stats.idre.ucla.edu/other/mult-pkg/whatstat/, which provides general guidelines for choosing a statistical test. To remind you to examine the link, I included the first 6 rows (limited to the first 4 columns) of that table below.

% latex table generated in R 4.2.1 by xtable 1.8-4 package % Sun Sep 18 19:49:01 2022

no. Ys	type of Xs	type of Ys	Test(s)
1	0 IVs (1 population)	interval & normal	one-sample t-test
1	0 IVs (1 population)	ordinal or interval	one-sample median
1	0 IVs (1 population)	categorical (2 categories)	binomial test
1	0 IVs (1 population)	categorical	Chi-square goodness-of-fit
1	1 IV with 2 levels (independent groups)	interval & normal	2 independent sample t-test

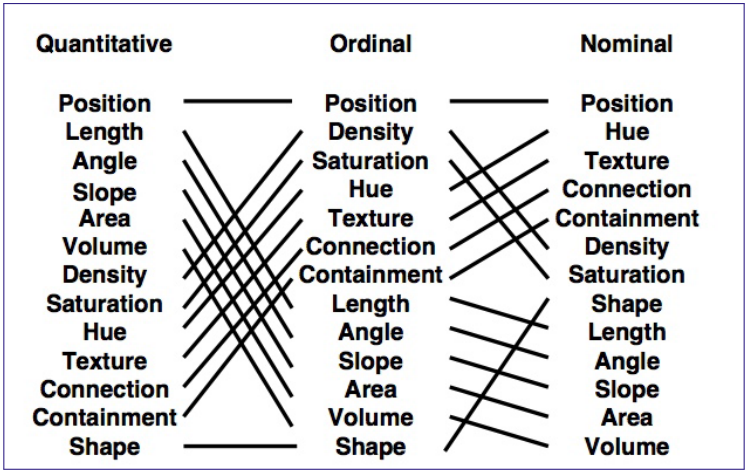
Fundamentals of Data Viz: Guidelines

The Visual Display of Quantitative Data (E. Tufte)

- Show the data
- Lead to thinking about the substance rather than something else
- Avoid distorting what the data have to say
- Present many numbers in a small space
- Make large datasets coherent
- Encourage the eye to compare different pieces of the data
- Reveal the data at several levels of detail, from a broad overview to the fine structure
- Serve a purpose: description, exploration, tabulation, decoration
- Be closely integrated with the statistical and verbal descriptions of a data set

Source: From Tufte, E. R. (2001). The visual display of quantitative information. Cheshire, Conn: Graphics Press, P. 13.

Fundamentals of Data Viz: Data Types + Encoding



Source: Please refer to Mackinlay (1986) for more details on proper and automated data encoding

Fundamentals of Data Viz: Grammar of Graphics [1]

In order to create a plot, you:

1. Call the ggplot() function which creates a blank canvas
2. Specify **aesthetic mappings**, which specifies how you want to map variables to visual aspects. In this case we are simply mapping the *Label* and *Average.Packet.Size* variables to the x- and y-axes.
3. You then add new layers that are geometric objects which will show up on the plot. In this case we add `geom_point()` to add a layer with points (dot) elements as the geometric shapes to represent the data.

Fundamentals of Data Viz: Grammar of Graphics [2]

```
df = read.csv("../../data/Portmap.csv", nrow=1000, stringsAsFactors = T)
df %<>% select ('Flow.ID','Source.IP','Destination.IP',
                'Timestamp', 'Flow.Duration', 'Total.Fwd.Packets',
                'Total.Backward.Packets', 'Max.Packet.Length',
                'Average.Packet.Size', 'Active.Mean',
                'SimillarHTTP', 'Label')

# create canvas
ggplot(df) +
  ggtitle("Creating an Empty Canvas")
```

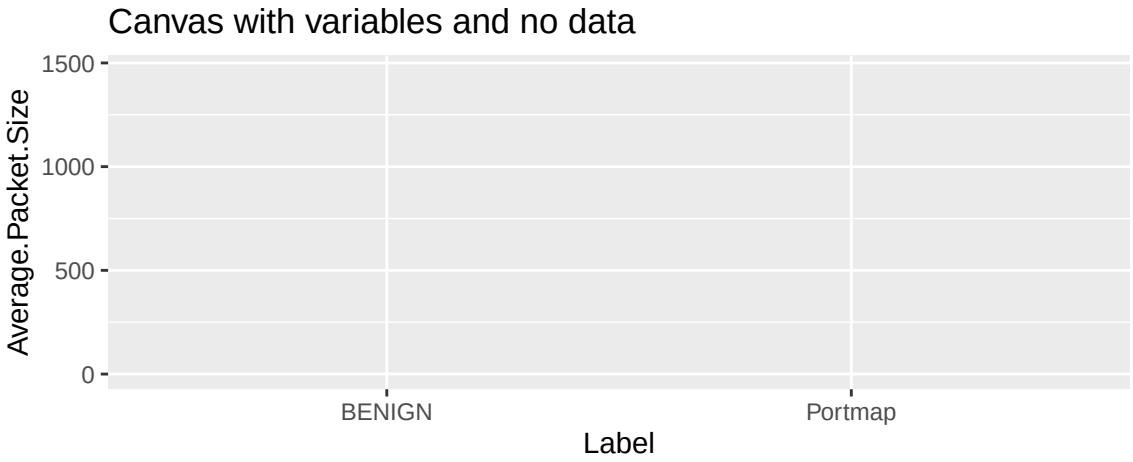
Fundamentals of Data Viz: Grammar of Graphics [3]

Creating an Empty Canvas

Fundamentals of Data Viz: Grammar of Graphics [4]

```
# variables of interest mapped
ggplot(df, aes(x=Label, y=Average.Packet.Size)) +
  ggtitle("Canvas with variables and no data")
```

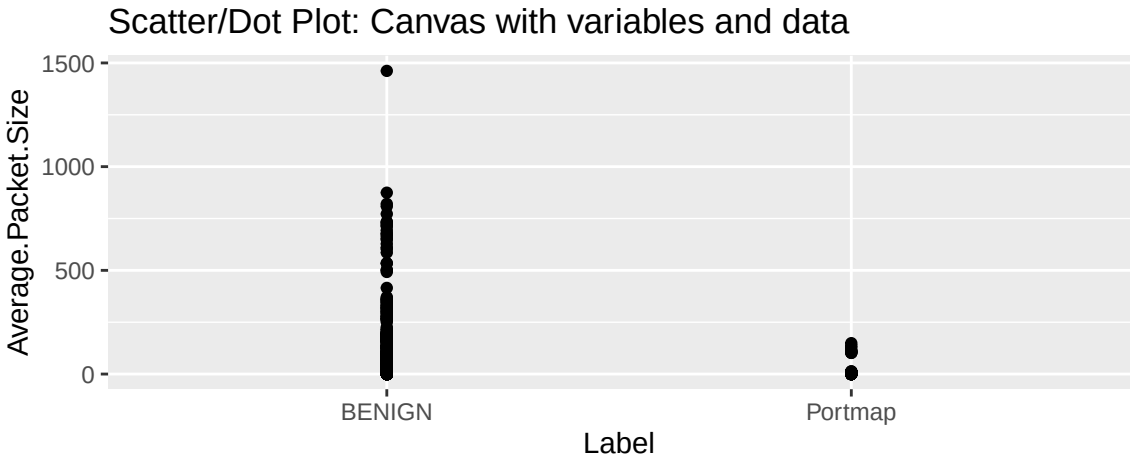
Fundamentals of Data Viz: Grammar of Graphics [5]



Fundamentals of Data Viz: Grammar of Graphics [6]

```
# data plotted: dot plot
ggplot(df, aes(x=Label, y=Average.Packet.Size)) + geom_point() +
  ggtitle("Scatter/Dot Plot: Canvas with variables and data")
```

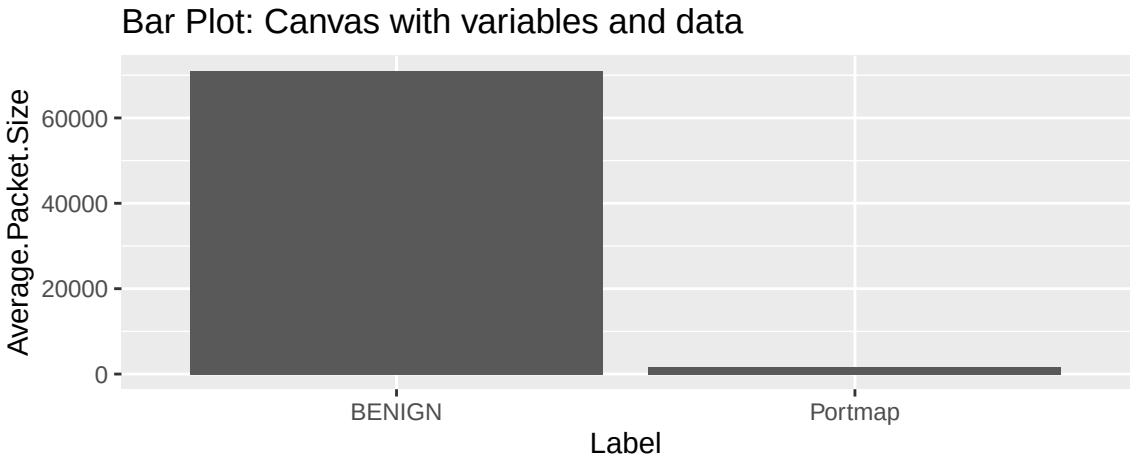
Fundamentals of Data Viz: Grammar of Graphics [7]



Fundamentals of Data Viz: Grammar of Graphics [8]

```
# data plotted: bar
ggplot(df, aes(x=Label, y=Average.Packet.Size)) +
  geom_bar(stat="identity") +
  ggtitle("Bar Plot: Canvas with variables and data")
```

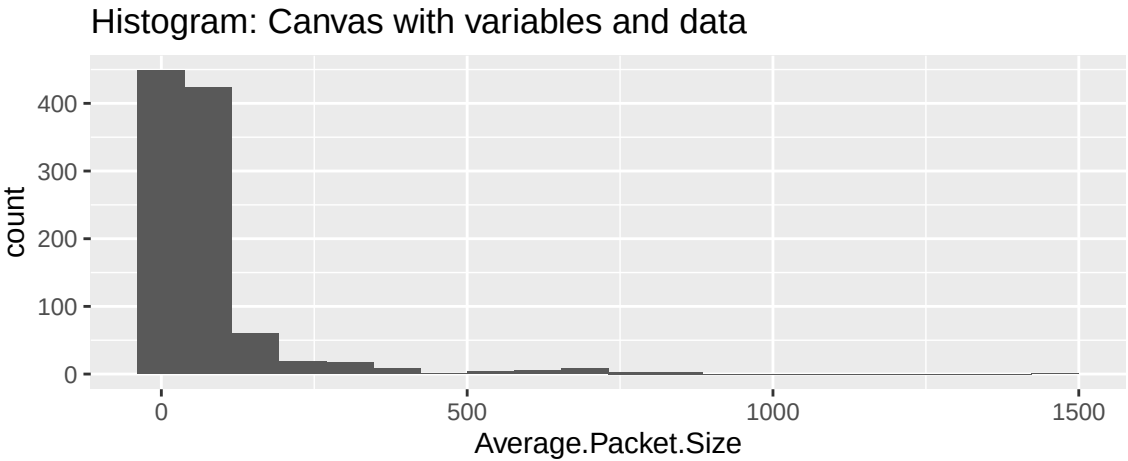
Fundamentals of Data Viz: Grammar of Graphics [9]



Fundamentals of Data Viz: Grammar of Graphics [10]

```
# data plotted: hist
ggplot(df, aes(x=Average.Packet.Size)) + geom_histogram(bins=20) +
  ggtitle("Histogram: Canvas with variables and data")
```

Fundamentals of Data Viz: Grammar of Graphics [11]

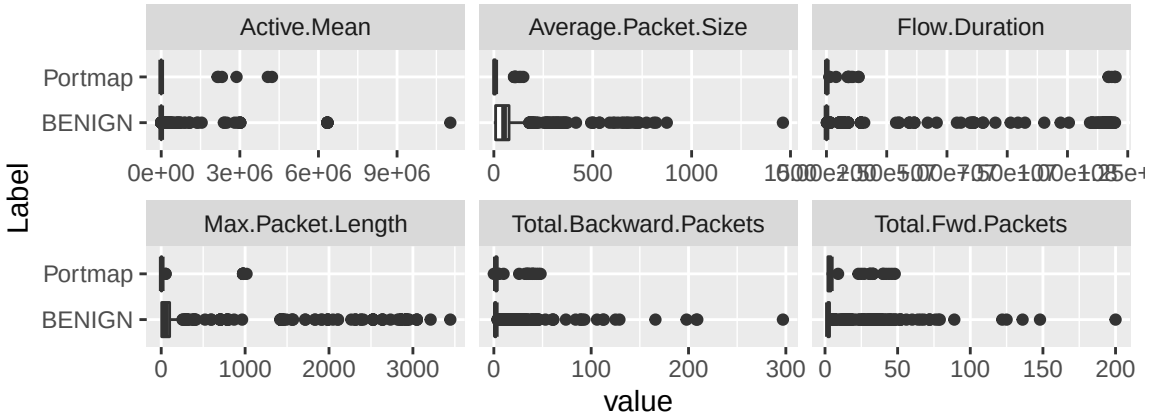


Fundamentals of Data Viz: Grammar of Graphics [12]

```
# Some nice functions from the Data Explorer Package
pacman::p_load(DataExplorer)

plot_boxplot(df, by="Label", ncol=3) # fav plot
```

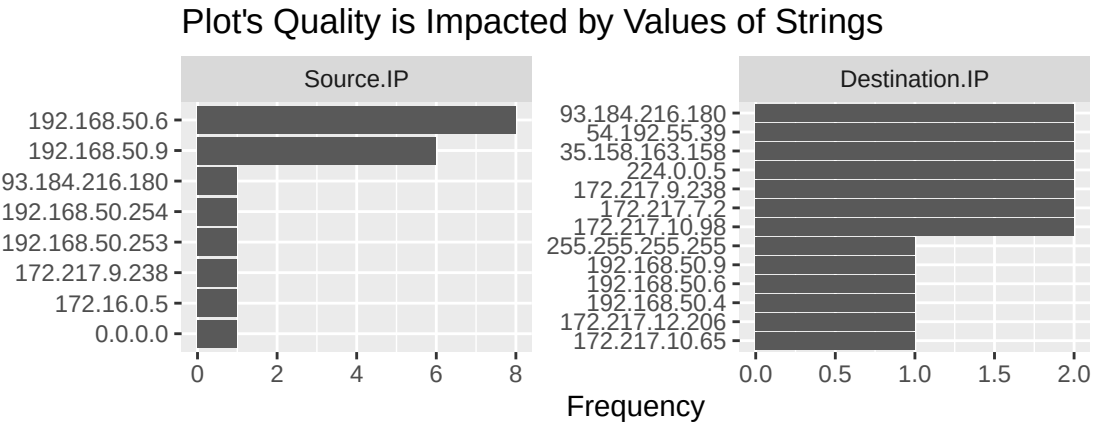
Fundamentals of Data Viz: Grammar of Graphics [13]



Fundamentals of Data Viz: Grammar of Graphics [14]

```
plot_bar(df[1:20,c('Source.IP','Destination.IP')],
         title = "Plot's Quality is Impacted by Values of Strings",
         ncol = 2) # first twenty rows and two factor/string cols
```

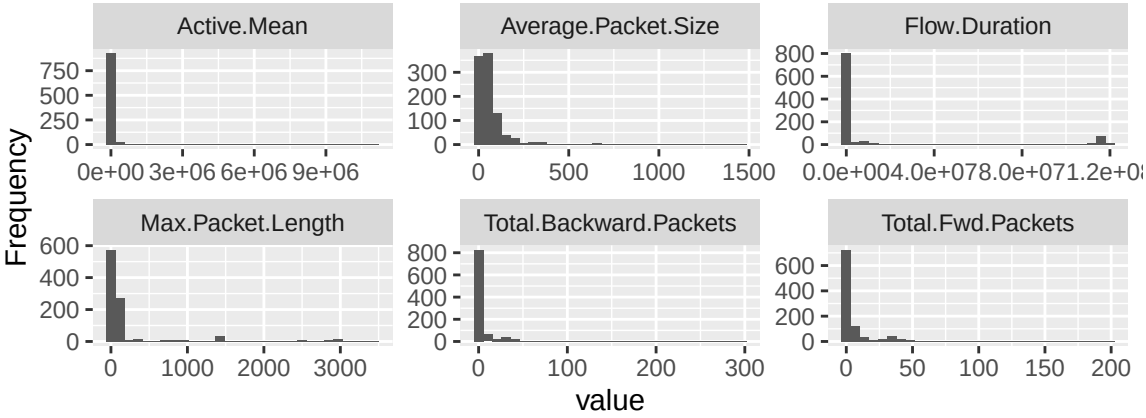
Fundamentals of Data Viz: Grammar of Graphics [15]



Fundamentals of Data Viz: Grammar of Graphics [16]

```
plot_histogram(df, ncol=3)
```

Fundamentals of Data Viz: Grammar of Graphics [17]



Outline

Things to Do [1]

- You are highly encouraged to go through the following materials: (see refs for link)
  - **EDA:** Chapter 7 in Grolemund and Wickham (2016)
  - **Data Viz:** Chapters 2 and 5 in Wilke (2019)
  - **ggplot2 Tutorial:** Chapters 2 and 10 in Wickham (2020). **You should also bookmark the following references**
    - *ggplot2-book:* Wickham (2020) is freely available at https://ggplot2-book.org/
    - *R cheatsheets:* https://raw.githubusercontent.com/rstudio/cheatsheets/main/data-visualization.pdf

References [1]

Anscombe, Francis J. 1973. "Graphs in Statistical Analysis." *The American Statistician* 27 (1): 17–21.

Cai, Xinying, and Camillo Padoa-Schioppa. 2012. "Neuronal Encoding of Subjective Value in Dorsal and Ventral Anterior Cingulate Cortex." *Journal of Neuroscience* 32 (11): 3791–3808.

Grolemund, Garrett, and Haldey Wickham. 2016. "R for Data Sciece." O Reilly Media, Inc. https://r4ds.had.co.nz/index.html. [Note you are highly encouraged to read and use Ch. 7 of the book.].

Mackinlay, Jock. 1986. "Automating the Design of Graphical Presentations of Relational Information." *Acm Transactions On Graphics (Tog)* 5 (2): 110–41.

Wickham, Hadley. 2020. "Ggplot2: Elegant Graphics for Data Analysiss." https://ggplot2-book.org/index.html.

References [2]

Wilke, Claus O. 2019. “Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures.” O’Reilly Media. https://serialmentor.com/dataviz/index.html.

ISA 419: Data Driven Security

08 - Descriptive Analytics (Cont.)

Fadel M. Megahed

Endres Associate Professor
Department of Information Systems and Analytics
Farmer School of Business
Miami University
Email: fmegahed@miamioh.edu
Office Hours: Automated Scheduler for Office Hours

Fall 2022