ADVANCED DATA VISUALIZATION AND REPORTING

ADVANCED DATA SCIENCE TRAINING I

IDLEWYLD Sysabee DAVHILL

data-action-lab.com

OUTLINE

1. Data Visualization with ggplot2
2. A Introduction to Dashboards

LEARNING OBJECTIVES

Learn how to create multivariate visualizations with ggplot2.

Become familiar with some of the issues and concepts relating to dashboards and their design.

DATA VISUALIZATION WITH GGPLOT2

ADVANCED DATA VISUALIZATION AND REPORTING

R GRAPHICS

As of 2018, there are 4 graphical systems available in R:

- *base*
- *grid*
- *lattice*
- *ggplot2*

Access to the 4 systems differ: *base*, *grid*, *lattice* are included in the base installation; *grid*, *lattice*, *ggplot2* have to be loaded explicitly before being used.

*not required

A GGPLOT2 PRIMER

ggplot2 is a set of tools that map data to visual display elements, and that allow the user to control the fine details of plot display.

Most important aspect: *ggplot2* can be used to think about the **logical structure** of the plot.

A *ggplot2* graph has 2 main components (and optional terms):

- aesthetic mappings (**aes** – connections between data and plot elems.)
- plot geometry (**geom** – specifies the type of plot)
- *facets, *coordinates, *scales, *labels, *guides, etc.

[Healey, K., *Data Visualization: A Practical Introduction*]

GGPLOT2 GRAMMAR

1. Tidy Data

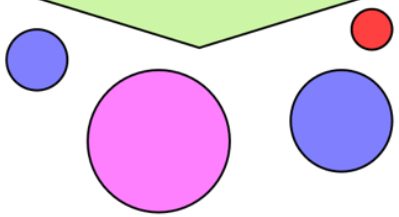
```
p ← ggplot(data = gapminder, ...
```

gdp	lifexp	pop	continent
340	65	31	Euro
227	51	200	Amer
909	81	80	Euro
126	40	20	Asia

2. Mapping

```
p ← ggplot(data = gapminder, mapping =
      aes(x = gdp, y = lifexp, size = pop,
        color = continent))
```

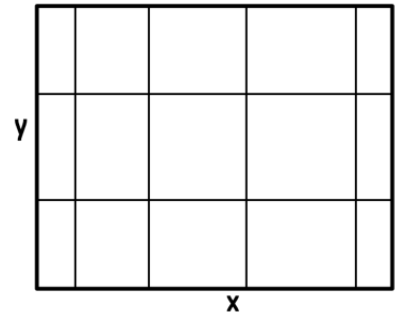
3. Geom



```
p + geom_point()
```

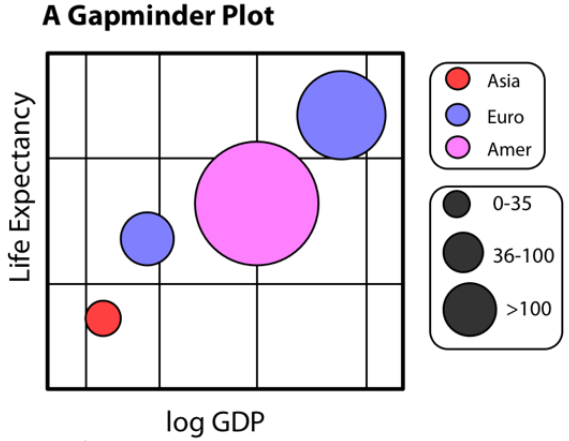
4. Co-Ordinates & Scales

```
p + coord_cartesian() + scale_x_log10()
```



5. Labels & Guides

```
p + labs(x = "log GDP", y = "Life
Expectancy", title = "A Gapminder Plot")
```



GGPLOT2 GRAMMAR – GEOMS

The data source and variables to be plotted are specified *via* `ggplot()`.

The various geom functions specify **how** these variables are to be visually represented

- using points, bars, lines, shaded regions, etc.

There are currently 37 available geoms.

*not required

GGPLOT2 GRAMMAR – GEOMS

Function	Adds	Options
`geom_bar()`	Bar chart	color, fill, alpha
`geom_boxplot()`	Box plot	color, fill, alpha, notch, width
`geom_density()`	Density plot	color, fill, alpha, linetype
`geom_histogram()`	Histogram	color, fill, alpha, linetype, binwidth
`geom_hline()`	Horizontal lines	color, alpha, linetype, size
`geom_jitter()`	Jittered points	color, size, alpha, shape
`geom_line()`	Line graph	colorvalpha, linetype, size
`geom_point()`	Scatterplot	color, alpha, shape, size
`geom_rug()`	Rug plot	color, side
`geom_smooth()`	Fitted line	method, formula, color, fill, linetype, size
`geom_text()`	Text annotations	Many; see the help for this function
`geom_violin()`	Violin plot	color, fill, alpha, linetype
`geom_vline()`	Vertical lines	color, alpha, linetype, size

GGPLOT2 GRAMMAR – GEOMS

Option	Specifies
color	colour of points, lines, and borders around filled regions
fill	colour of filled areas such as bars and density regions
alpha	transparency of colors, ranging from 0 (fully transparent) to 1 (opaque)
linetype	pattern for lines (1 = solid, 2 = dashed, 3 = dotted, 4 = dotdash, 5 = longdash, 6 = twodash)
size	point size and line width
shape	point shapes (same as pch, with 0 = open square, 1 = open circle, 2 = open triangle, and so on)
position	position of plotted objects such as bars and points. For bars, "dodge" places grouped bar charts side by side, "stacked" vertically stacks grouped bar charts, and "fill" vertically stacks grouped bar charts and standardizes their heights to be equal; for points, "jitter" reduces point overlap
binwidth	bin width for histograms
notch	indicates whether box plots should be notched (TRUE/FALSE)
sides	placement of rug plots on the graph ("b" = bottom, "l" = left, "t" = top, "r" = right, "bl" = both bottom and left, and so on)
width	width of box plots

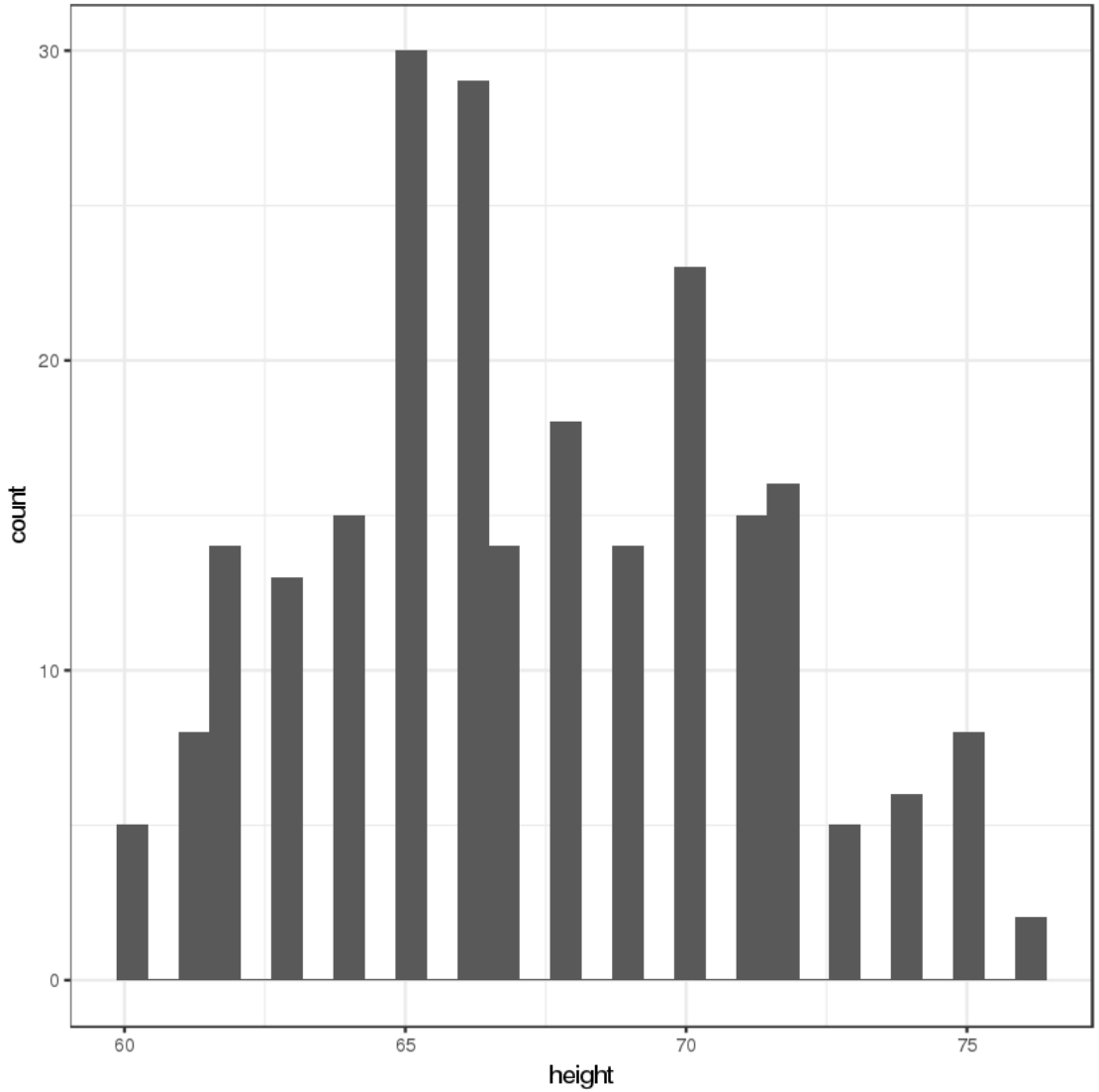
GGPLOT2 GRAMMAR – GEOM()

```
library("ggplot2")
data(singer, package="lattice")
# Using data from the 1979 ed. of the
# New York Choral Society

# Histogram of heights
ggplot(singer, aes(x=height)) +
  geom_histogram()

# Boxplot of heights by voice part
ggplot(singer, aes(x=voice.part, y=height)) +
  geom_boxplot()
```

What do you expect the output to be?

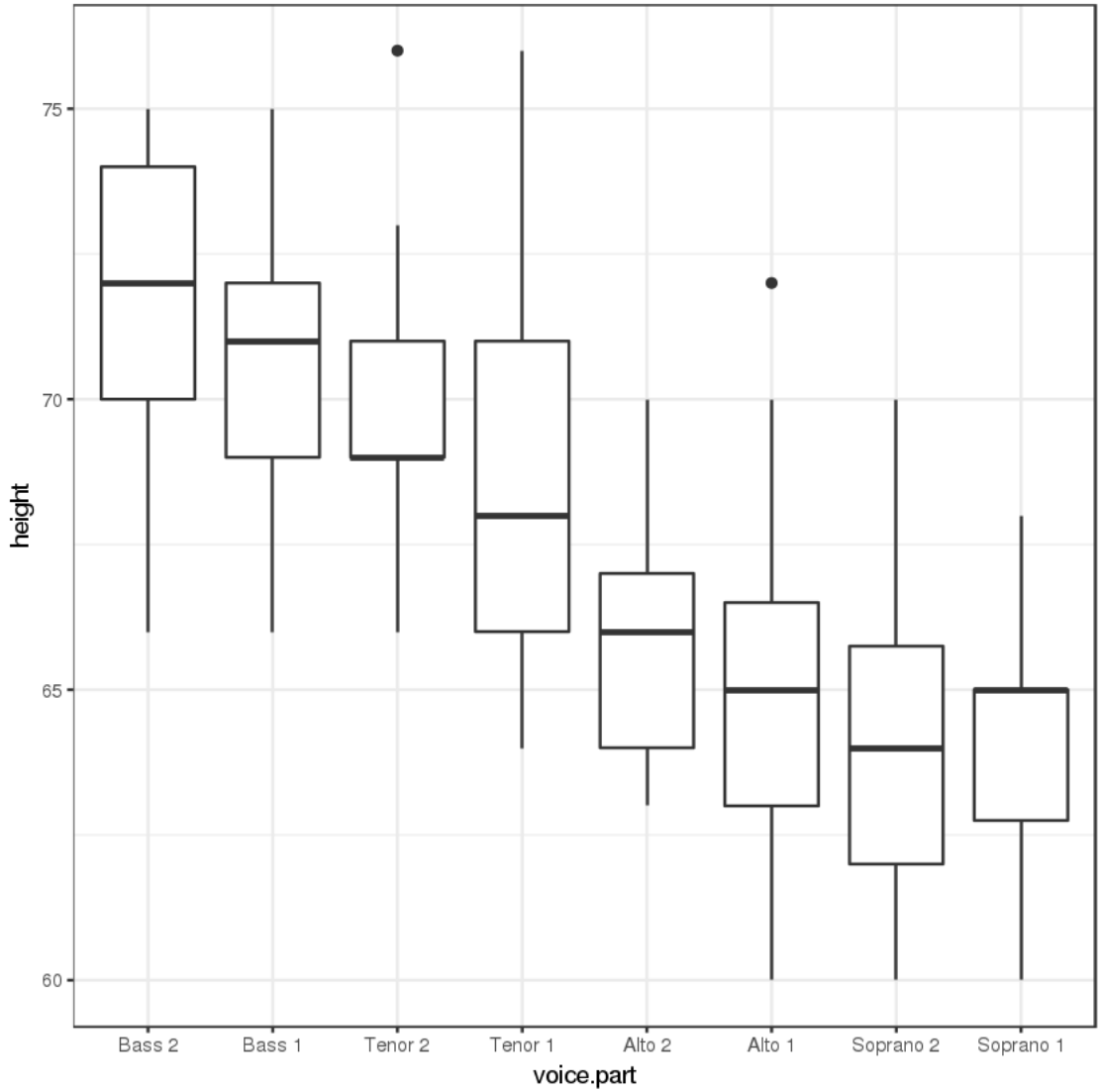
GGPLOT2 GRAMMAR – GEOM()

```
library("ggplot2")
data(singer, package="lattice")
# Using data from the 1979 ed. of the
# New York Choral Society

# Histogram of heights
ggplot(singer, aes(x=height)) +
  geom_histogram()

# Boxplot of heights by voice part
ggplot(singer, aes(x=voice.part, y=height)) +
  geom_boxplot()
```



GGPLOT2 GRAMMAR – GEOM()

```
library("ggplot2")
data(singer, package="lattice")
# Using data from the 1979 ed. of the
# New York Choral Society

# Histogram of heights
ggplot(singer, aes(x=height)) +
   geom_histogram()

# Boxplot of heights by voice part
ggplot(singer, aes(x=voice.part, y=height)) +
   geom_boxplot()
```



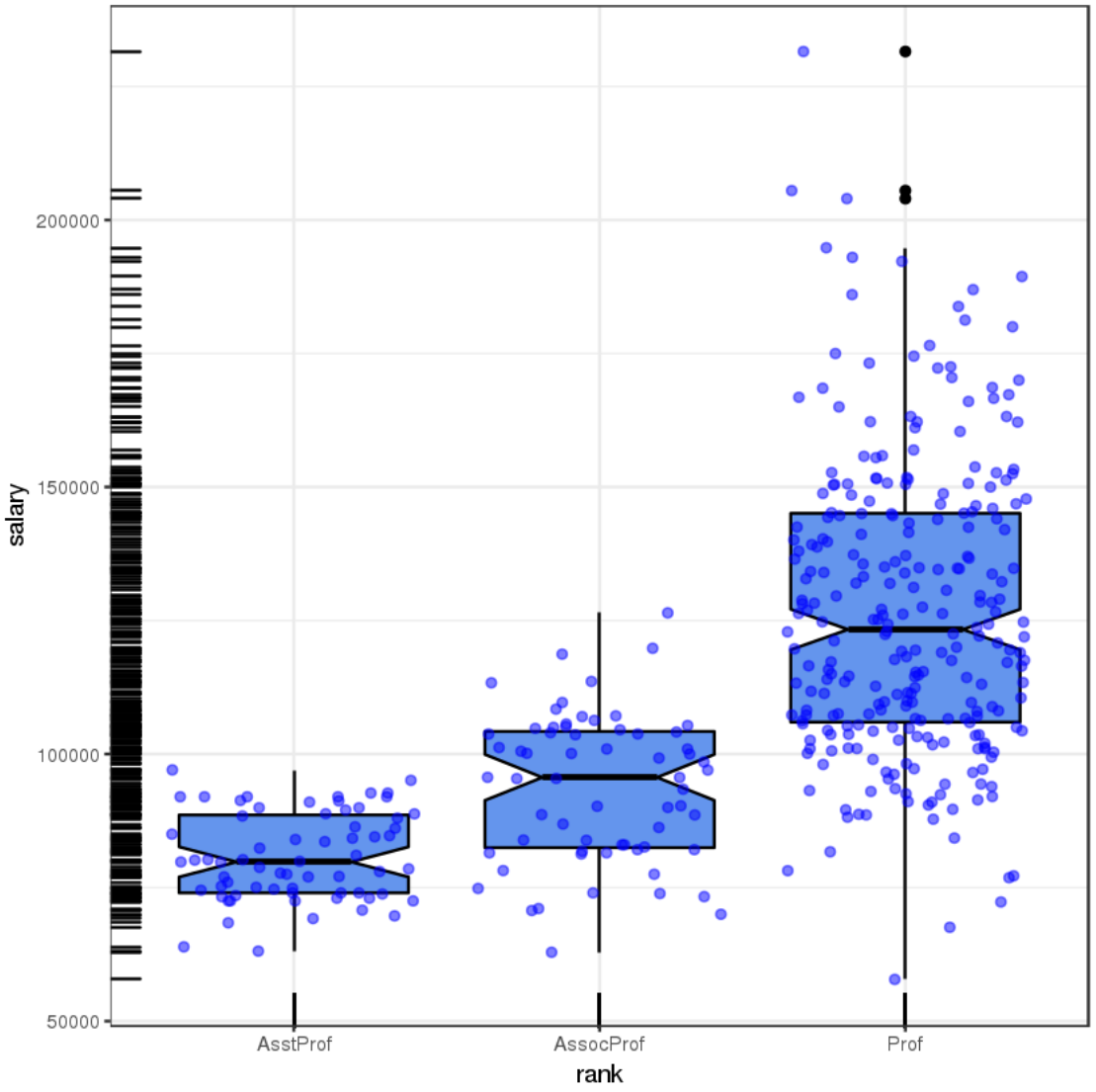
GGPLOT2 GRAMMAR – GEOM()

```
library(ggplot2)
data(Salaries, package="car")
# Using data on salaries of a sample of
# US university professors (2018-2019)
# var: rank, sex, yrs.since.phd, yrs.service, salary

ggplot(Salaries, aes(x=rank, y=salary)) +
  geom_boxplot(fill="cornflowerblue",color="black", notch=TRUE) +
  geom_point(position="jitter", color="blue", alpha=.5) +
  geom_rug(side="l", color="black")
```

What do you expect the output to be?

GGPLOT2 GRAMMAR – GEOM()

```
library(ggplot2)
data(Salaries, package="car")
# Using data on salaries of a sample of
# US university professors (2018-2019)
# var: rank, sex, yrs.since.phd, yrs.service, salary

ggplot(Salaries, aes(x=rank, y=salary)) +
   geom_boxplot(fill="cornflowerblue",color="black", notch=TRUE) +
   geom_point(position="jitter", color="blue", alpha=.5) +
   geom_rug(side="l", color="black")
```



GGPLOT2 GRAMMAR – AESTHETICS

Aesthetics refer to the displayed attributes of the data.

They map the data to an attribute (such as the size or shape of a marker) and generate an appropriate legend.

Aesthetics are specified with the `aes()` function.

Aesthetics can be specified within the data function or within a geom. If they're specified within the data function then they apply to all specified geoms.

GGPLOT2 GRAMMAR – AESTHETICS

The aesthetics available to `geom_point()` (scatterplot), as an example, are:

- `x`, `y`, `alpha`, `color`, `fill`, `shape`, `size`

Important difference between specifying characteristics (like colour and shape) inside and outside the `aes()` function

- inside: assigned colour or shape automatically based on the data.
- outside: not mapped to data.

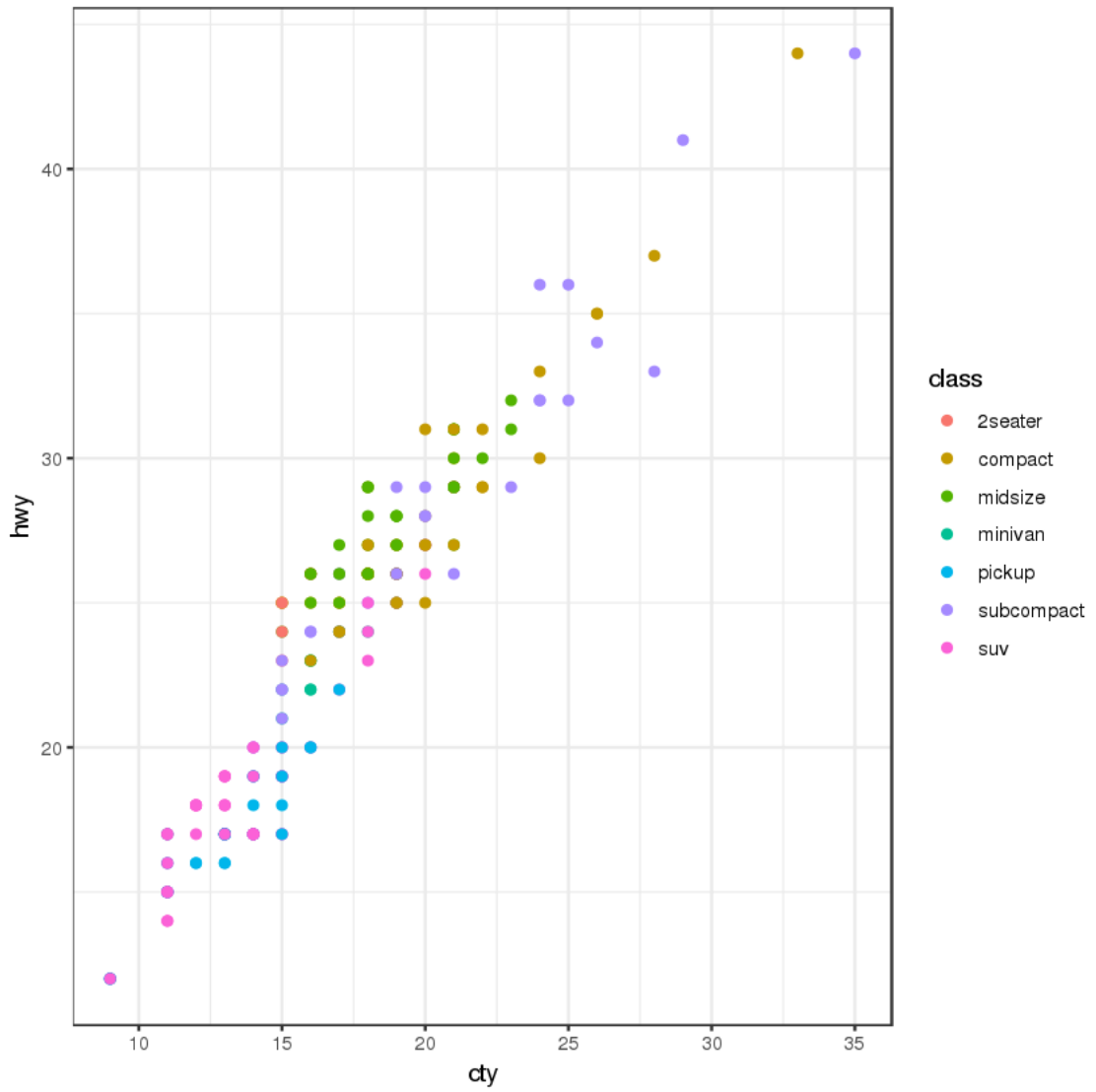
GGPLOT2 GRAMMAR – AES()

```
library(ggplot2)
# Using the mpg dataset

# specifying characteristics inside aes()
ggplot(mpg, aes(cty, hwy)) +
  geom_point(aes(colour = class))

# specifying characteristics inside aes()
ggplot(mpg, aes(cty, hwy)) +
  geom_point(colour = "red")
```

What do you expect the output to be?

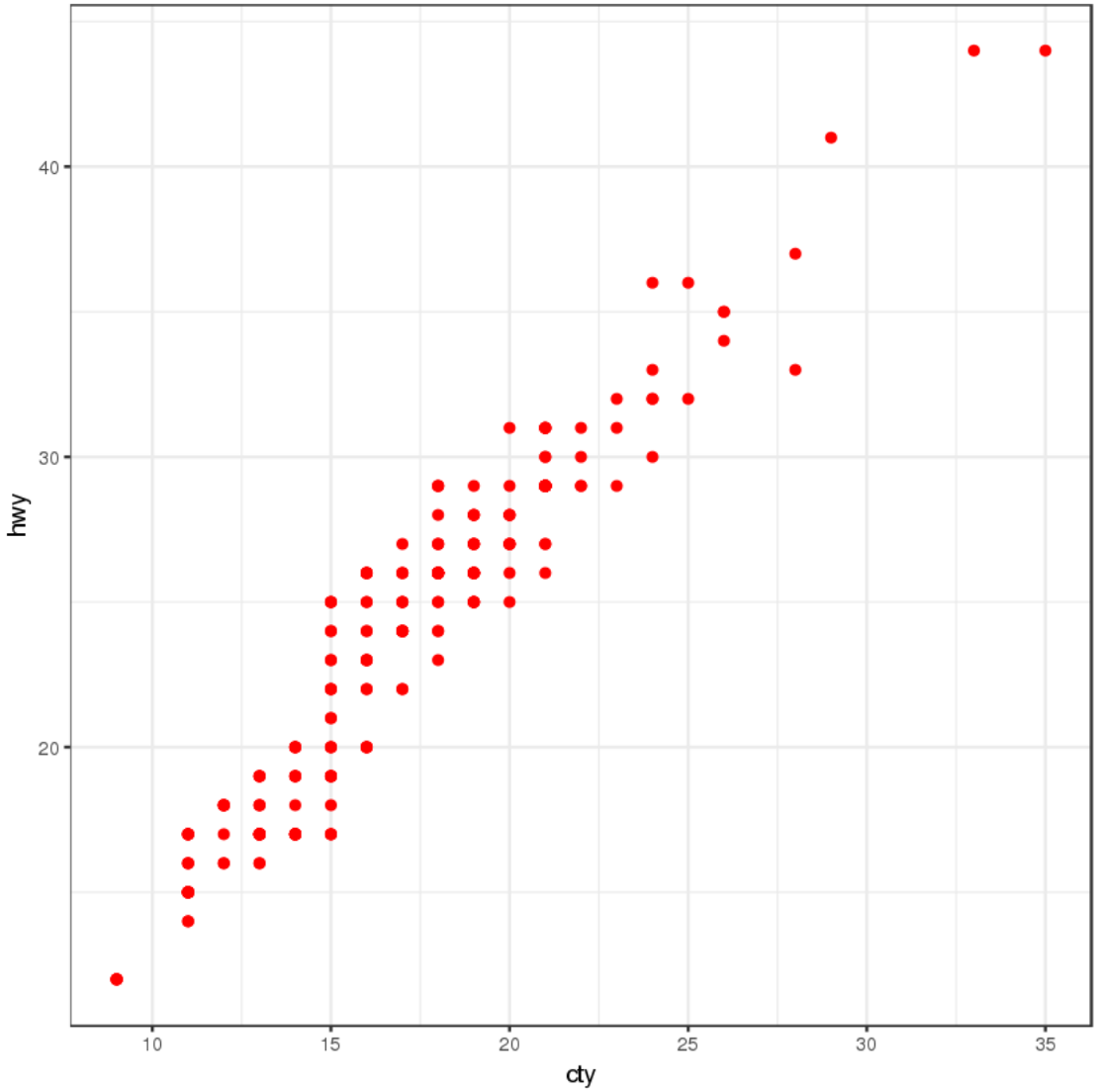
GGPLOT2 GRAMMAR – AES()

```
library(ggplot2)
# Using the mpg dataset

# specifying characteristics inside aes()
ggplot(mpg, aes(cty, hwy)) +
  geom_point(aes(colour = class))

# specifying characteristics inside aes()
ggplot(mpg, aes(cty, hwy)) +
  geom_point(colour = "red")
```



GGPLOT2 GRAMMAR – AES()

```
library(ggplot2)
# Using the mpg dataset

# specifying characteristics inside aes()
ggplot(mpg, aes(cty, hwy)) +
  geom_point(aes(colour = class))

# specifying characteristics inside aes()
ggplot(mpg, aes(cty, hwy)) +
  geom_point(colour = "red")
```



GGPLOT2 GRAMMAR – FACETS

In *ggplot2* parlance, small multiples are referred to as **facets**:

- `facet_wrap()`, `facet_grid()`

By default, all panels (one for each factor) share the same axes (scale-wise).

Separating the graph into a sequence of smaller, side-by-side plots makes it easier to enact comparisons.

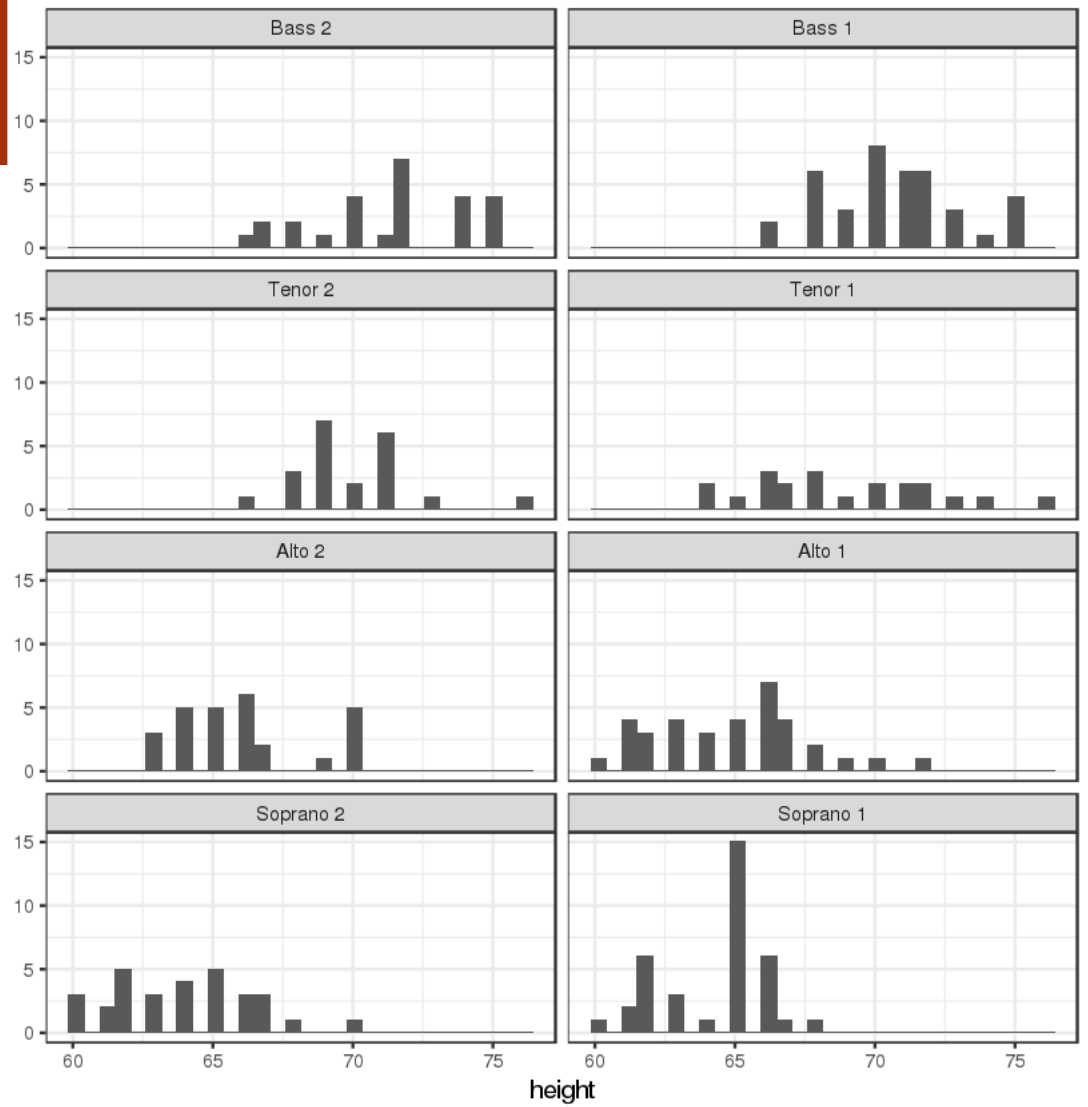
GGPLOT2 GRAMMAR – FACET_WRAP()

```
data(singer, package="lattice")
library(ggplot2)
ggplot(data=singer, aes(x=height)) +
      geom_histogram() +
      facet_wrap(~voice.part, nrow=4)
```

What do you expect the output to be?

GGPLOT2 GRAMMAR – FACET_WRAP()

```
data(singer, package="lattice")
library(ggplot2)
ggplot(data=singer, aes(x=height)) +
      geom_histogram() +
      facet_wrap(~voice.part, nrow=4)
```



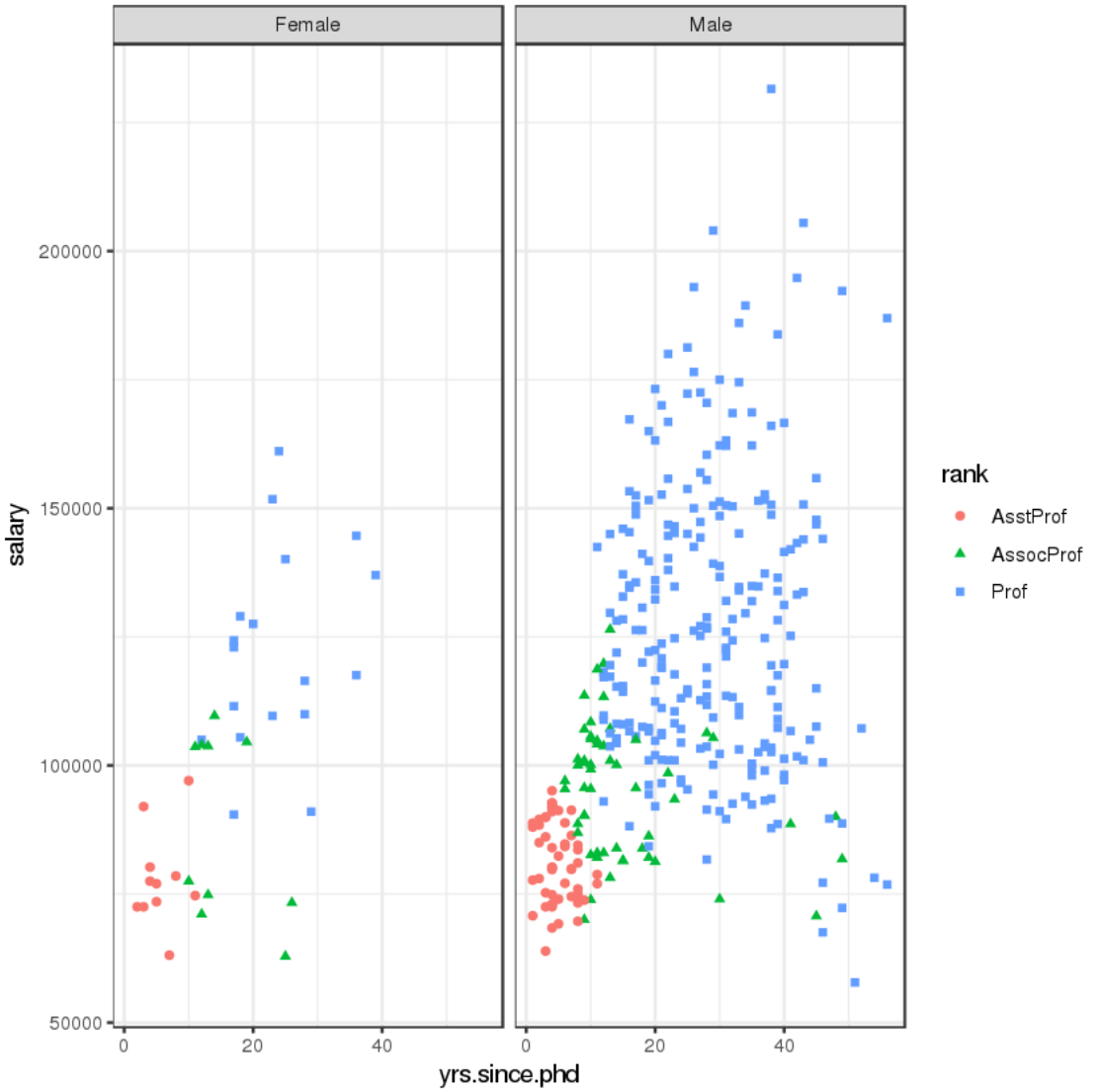
GGPLOT2 GRAMMAR – FACET_GRID()

```
data(Salaries, package="car")
library(ggplot2)
ggplot(Salaries, aes(x=yrs.since.phd,
 y=salary, color=rank, shape=rank)) +
   geom_point() +
   facet_grid(~sex)
```

What do you expect the output to be?

GGPLOT2 GRAMMAR – FACET_GRID()

```
data(Salaries, package="car")
library(ggplot2)
ggplot(Salaries, aes(x=yrs.since.phd,
 y=salary, color=rank, shape=rank)) +
   geom_point() +
   facet_grid(~sex)
```



EXERCISES

Detailed explanations and examples can be found in *A ggplot2 Primer*.
https://www.data-action-lab.com/2018/11/12/a-ggplot2-primer/

With a classmate, run through some of the examples provided in the R Notebook. It is not necessary for you to understand all the details of the visualizations, but it is useful to see what kind of outcomes are possible.

Generate 5 ggplot2 visualizations for the collisions and the algae bloom datasets.

AN INTRODUCTION TO DASHBOARDS

ADVANCED DATA VISUALIZATION AND REPORTING

"If a tree falls in the forest and nobody is there to hear it, does it make a sound?"

(old riddle)

REPORTING AND DEPLOYMENT

An analysis can only be as good as how it is **communicated** and/or **deployed**.

Crucial Questions:

- Who is in receipt of the report(s)?
- How are the workflows deployed into production?
- Can data insights be turned into useful policies?

Automatic reporting should be audited and validated **regularly**.

REPORTING AND DEPLOYMENT

Communication should occur at various stages of the project, not solely upon completion:

- keep sponsors / clients aware of broad lines
- technical details may be avoided, but documented nonetheless

Ideal scenario: analysis software is also reporting software

- minimizes human error related to cut-and-paste
- removes the need for keeping analysis and reporting separate
- makes sharing the work with other project member easier

Simplify the process further by deploying directly to the Web.

DISCUSSION

What are your favourite reporting tools?

How much should you test a product before deployment?

What's the cost of deploying a faulty product?

DASHBOARDS

A **dashboard** is any visual display of data used to monitor conditions and/or facilitate understanding.

Examples:

- interactive display that allows people to explore motor insurance claims by city, province, driver age, etc.
- PDF showing key audit metrics that gets e-mailed to a Department's DG on a weekly basis.
- wall-mounted screen that shows call centre statistics in real-time.
- mobile app that allow hospital administrators to review wait times on an hourly- and daily-basis for the current year and the previous year.

SOME QUESTIONS TO CONSIDER

In a car's dashboard, a small number of **key indicators** (speed, gasoline level, lights, etc.) need to be understood **at a glance**. A dashboard design that does not take these two characteristics under consideration can have catastrophic consequences.

The following questions need to be answered prior to the dashboard being designed:

- Who is the dashboard's **consumer**?
- What **story** does the dashboard tell?
- What data (categories) will be used?
- What will **appear** on the dashboard?
- How can the dashboard **help** the consumer?

[https://www.geckoboard.com/blog/building-great-dashboards-6-golden-rules-to-successful-dashboard-design]

DASHBOARD DESIGN GUIDELINES

Nick Smith suggests the following 6 Golden Rules:

- **Consider the audience** (who are you trying to inform? does the DG really need to know that the servers are operating at 88% capacity?)
- **Select the right type of dashboard** (operational, strategic/executive, analytical)
- **Group data logically, use space wisely** (split functional areas: product, sales/marketing, finance, people, etc.)
- **Make the data relevant to the audience** (scope and reach of data, different dashboards for different departments, etc.)
- **Avoid cluttering the dashboard** (present the most important metrics only)
- **Refresh your data at the right frequency** (real-time, daily, weekly, monthly, etc.)

[https://dashboard.edmonton.ca]

TRANSPORTATION | LIVABILITY | ENVIRONMENT | URBAN FORM | ECONOMY | FINANCE

Meets or Exceeds Target | Near Target | Needs Improvement | Measuring | Collecting Data

IDLEWYLD Sysabee DAVHILL

data-action-lab.com

COURSE METRICS DASHBOARD – SCENARIO

You are the head of an academic department. You want to know:

- how a given professor's course is rated compared to other courses in the department and at the university in general
- the overall course load, the number of students, and the overall growth or decline in the enrollment for a particular course
- how many courses an instructor has been teaching over time
- the detailed ratings of the most recent course and instructor feedback

What type of data do you need? How would you arrange/design a dashboard to help answer these questions?

COURSE METRICS DASHBOARD – DATA

Year	Semester	Students	Average
'12	S	42	52
	F	16	52
'13	S	71	52
	US	14	52
	F	27	52
'14	S	69	52
	S	55	52
	US	28	52
	F	27	52
	F	61	52
'15	S	46	52
	S	80	52
	US	43	52
	F	61	52
	F	69	52
'16	S	62	52
	S	80	52
	US	50	52
	F	62	52
	F	65	52
	F	69	52
		1097	

year	enrollments
'12	58
'13	112
'14	240
'15	299
'16	388
	687

Year	# classes
'12	2
'13	3
'14	5
'15	5
'16	6
	21

Year	Semester	Rating
'12	S	6.6
	F	6.5
'13	S	6.7
	US	7.7
	F	6.9
'14	S	6.4
	S	6.7
	US	7.5
	F	7.3
	F	7
'15	S	6.4
	S	7
	US	6.8
	F	7.3
	F	7.7
		7.7

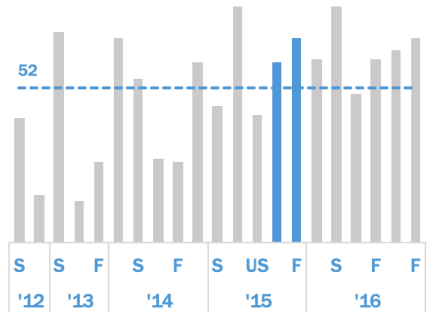
COURSE METRICS DASHBOARD – DATA

Semesters	Questions	Mean Rating	Entity	Shaffer	BANA	College
2015 Fall Semester 002	The instructor was well organized	7.5	Shaffer	7.5	6.8	7
2015 Fall Semester 002	The instructor communicated clearly	7.6	Shaffer	7.6	6.5	6.9
2015 Fall Semester 002	The instructor interacted well with students	7.7	Shaffer	7.7	6.6	7
2015 Fall Semester 002	The Instructor graded fairly	7.6	Shaffer	7.6	6.8	7.1
2015 Fall Semester 002	I developed specific skills and competencies	7.2	Shaffer	7.2	6.3	6.5
2015 Fall Semester 002	Overall, this instructor was excellent	7.7	Shaffer	7.7	6.4	6.8
2015 Fall Semester 002	Overall, this was an excellent course	7.4	Shaffer	7.4	5.9	6.4
2015 Fall Semester 001	The instructor was well organized	7.3	Shaffer	7.3	7	6.9
2015 Fall Semester 001	The instructor communicated clearly	7.4	Shaffer	7.4	6.7	6.7
2015 Fall Semester 001	The instructor interacted well with students	7.3	Shaffer	7.3	6.8	6.8
2015 Fall Semester 001	The Instructor graded fairly	7.5	Shaffer	7.5	7.1	7
2015 Fall Semester 001	I developed specific skills and competencies	6.9	Shaffer	6.9	6.8	6.7
2015 Fall Semester 001	Overall, this instructor was excellent	7.3	Shaffer	7.3	6.7	6.7
2015 Fall Semester 001	Overall, this was an excellent course	7.1	Shaffer	7.1	6.6	6.5

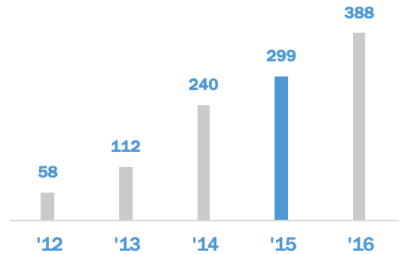
Course Metrics

[https://bigbookofdashboards.com/dashboards.html]

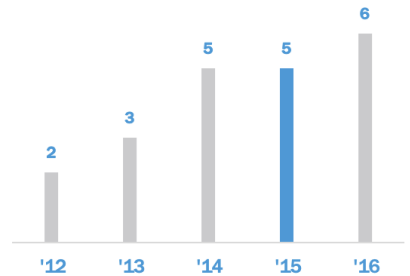
Students



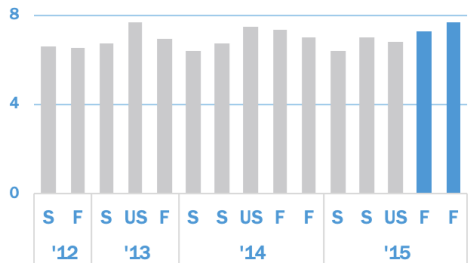
Enrollments



Classes



Ratings



1097

Total Students in five years

687

Total Students in 2015-2016

21

Total Classes in five years

7.7 of **8**

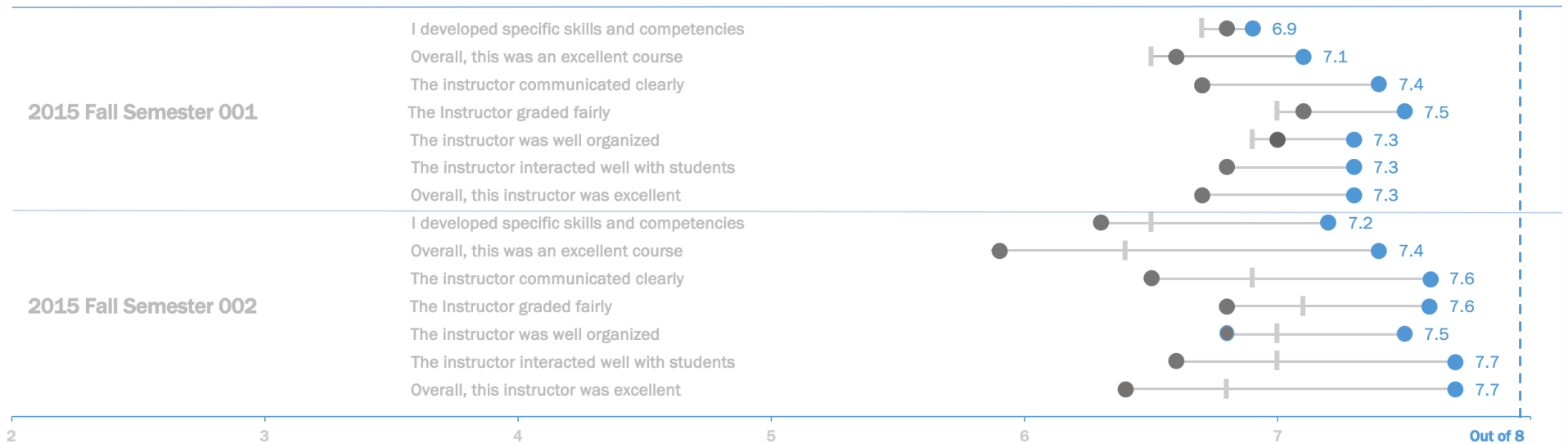
Most recent instructor rating (out of 8.0)

Semesters

Questions

BANA College Shaffer

Ratings



Course Metrics Dashboard created by Jeffrey A. Shaffer. Data from University of Cincinnati Course Evaluations. Blue indicates the 2 most recent rating periods.

COURSE METRICS DASHBOARD – STRENGTHS

Easy-to-see key metrics

Simple color scheme

Potential to be static or interactive

Both overview and details are clear

DISCUSSION

There are no perfect dashboards – no collection of charts will ever suit everyone who encounters it.

All dashboards should be **truthful** and **functional**, but dashboards that are also **elegant** (delightful, enjoyable) will take you further.

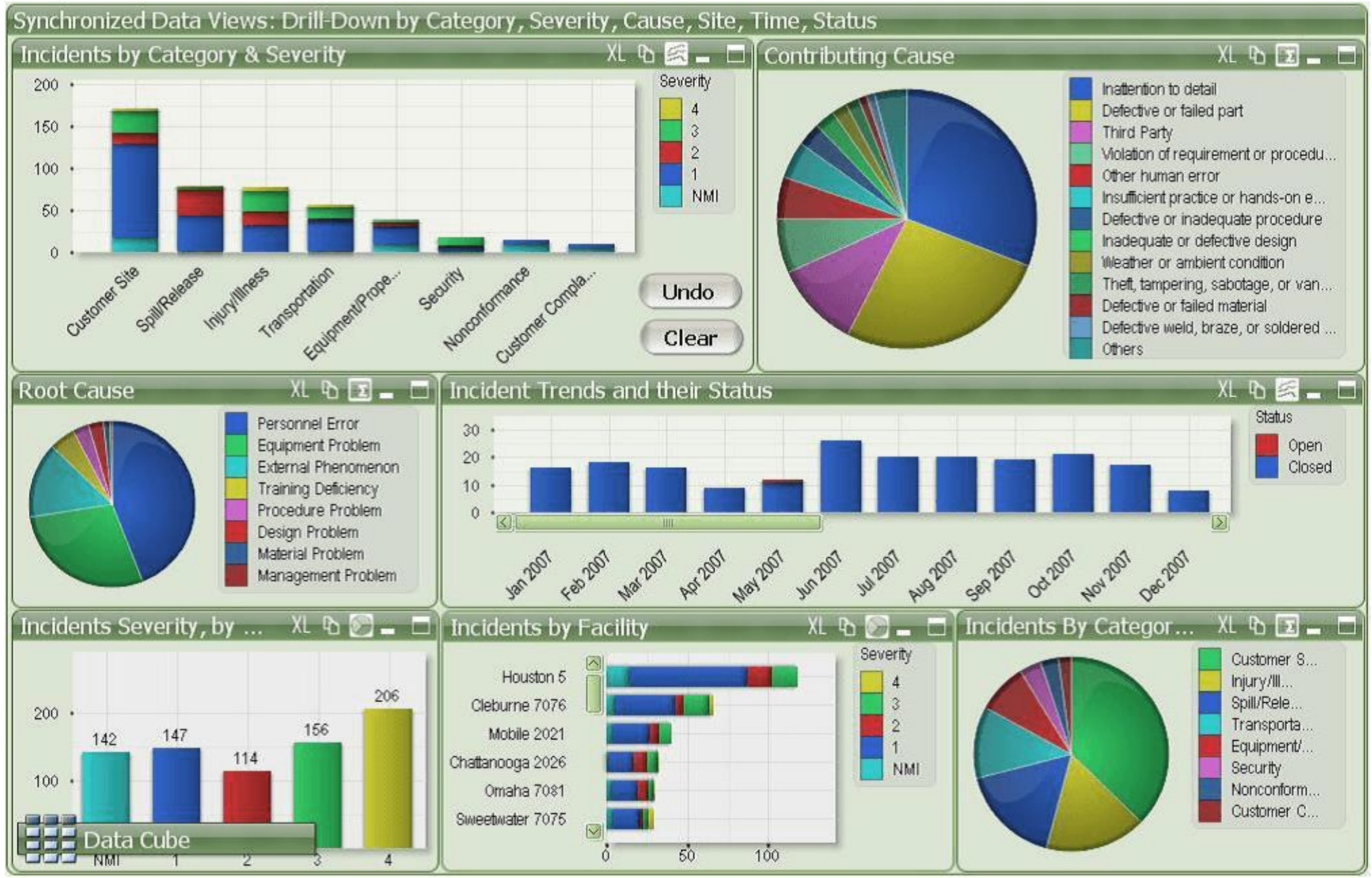
All dashboards are **incomplete**. Good dashboards will still lead to dead ends, but they should allow users to ask: "Why? What is the root cause of a problem?"

Tools: Excel, Power BI, Tableau, R + Shiny, Geckoboard, Matillion, etc.

EXERCISE

Consider the following dashboards. Can you figure out, at a glance, who their audience is? What are their strengths? What are their limitations? How could you improve them?

[https://www.matillion.com/wp-content/uploads/2014/11/qlikview-poor-use-of-dashboard-software.png]



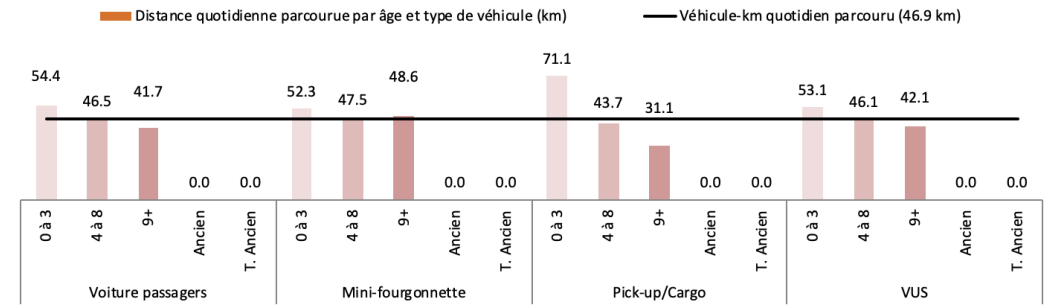
[https://www.geckoboard.com/assets/2-terrible-dashboard-example-min.png]

[Personal file]

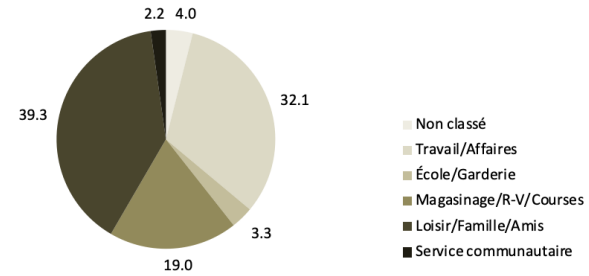
Ontario – 1er trimestre 2012

Caractéristiques des déplacements

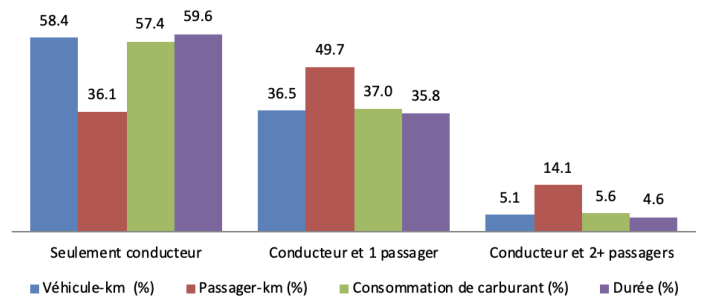
Véhicule-km quotidien parcouru par âge et type de véhicule



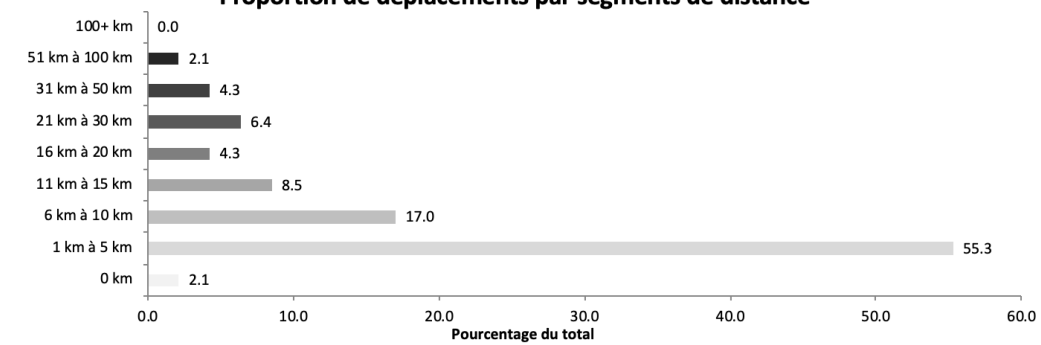
Passager-km quotidien parcouru par but des déplacements (%)



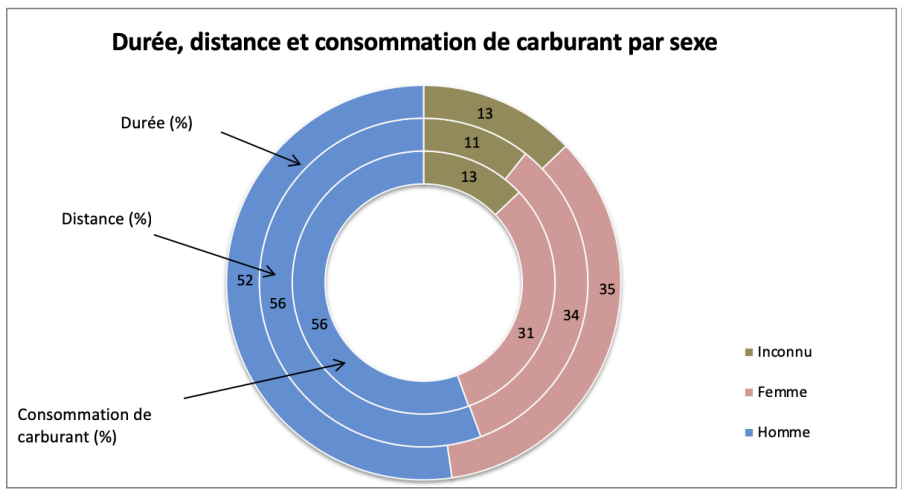
Distance, passager-km parcouru, durée et consommation de carburant par occupation



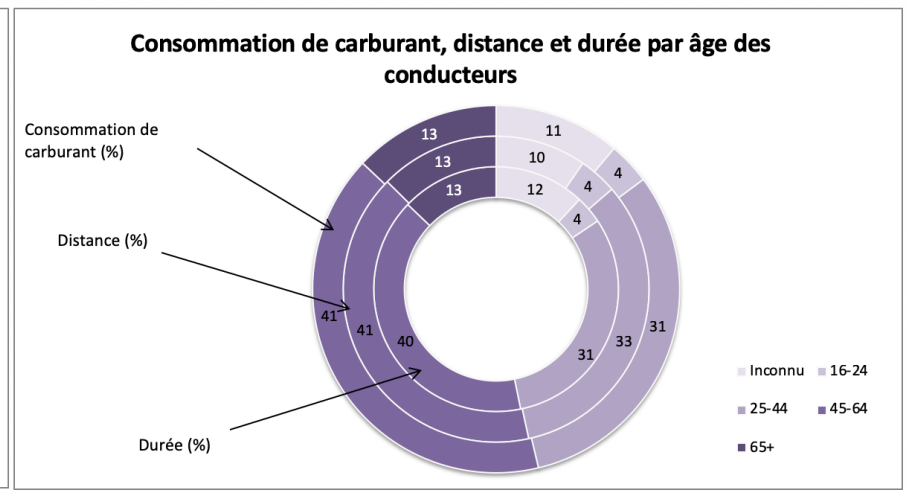
Proportion de déplacements par segments de distance



Durée, distance et consommation de carburant par sexe

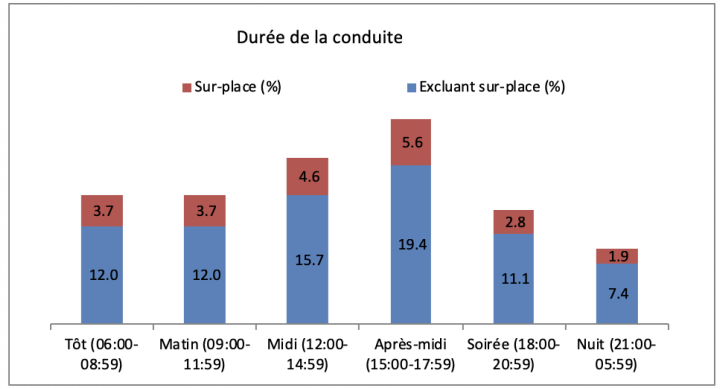


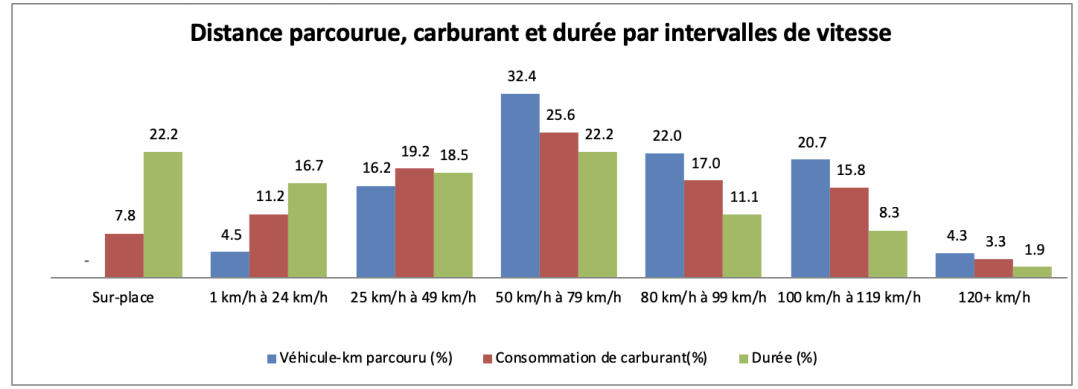
Consommation de carburant, distance et durée par âge des conducteurs

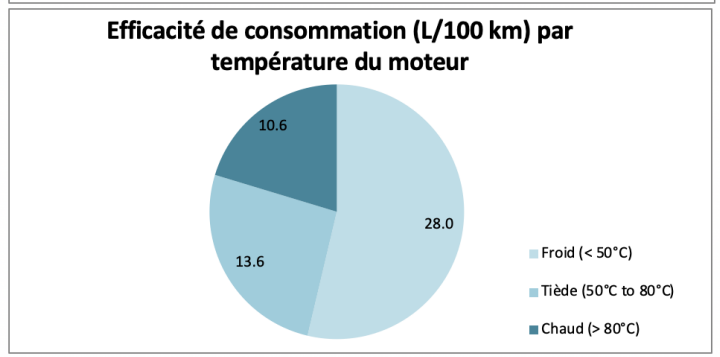


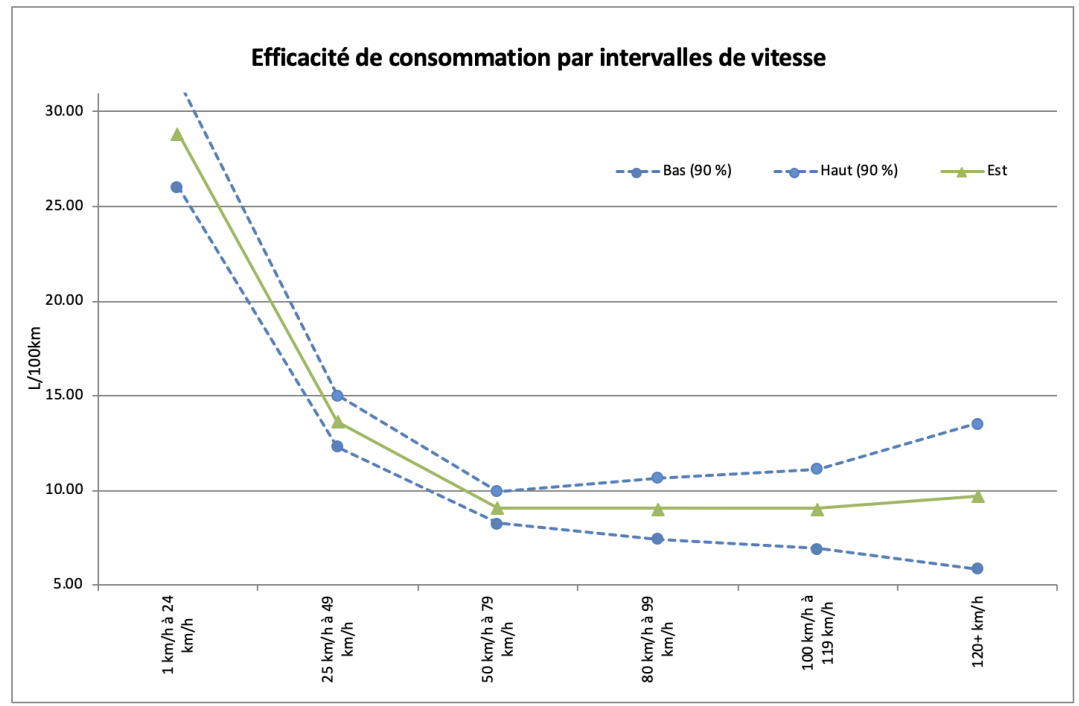
Ontario – 1er trimestre 2012

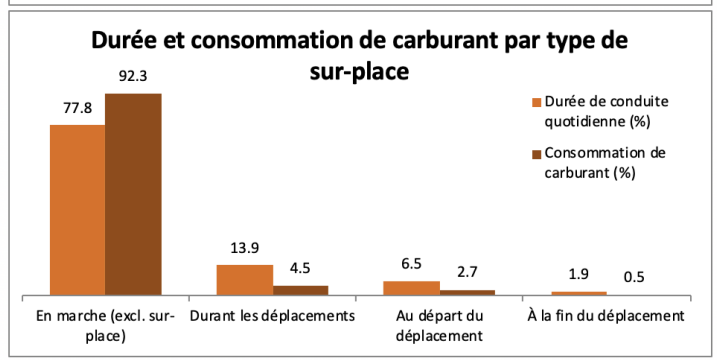
Sous-caractéristiques des déplacements









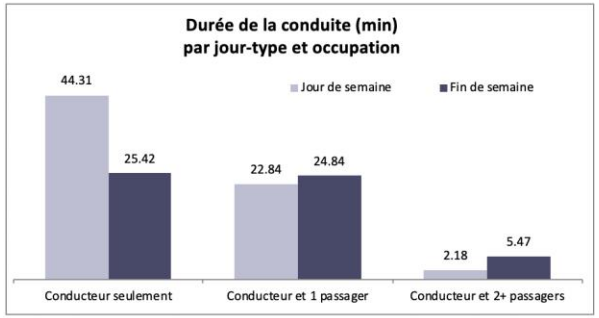


[Personal file]

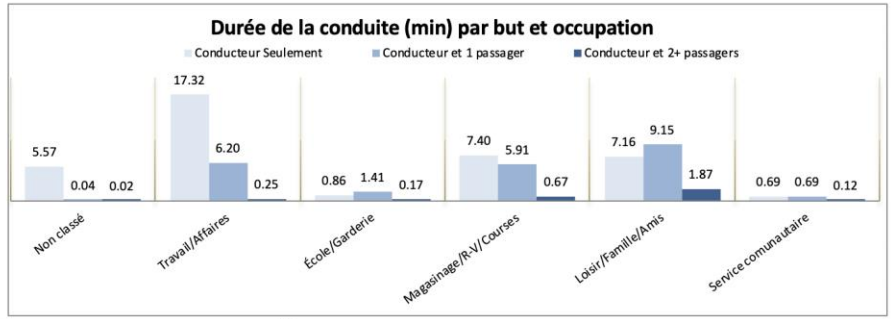
Ontario – 1er trimestre 2012

Caractéristiques mixtes sur les déplacements

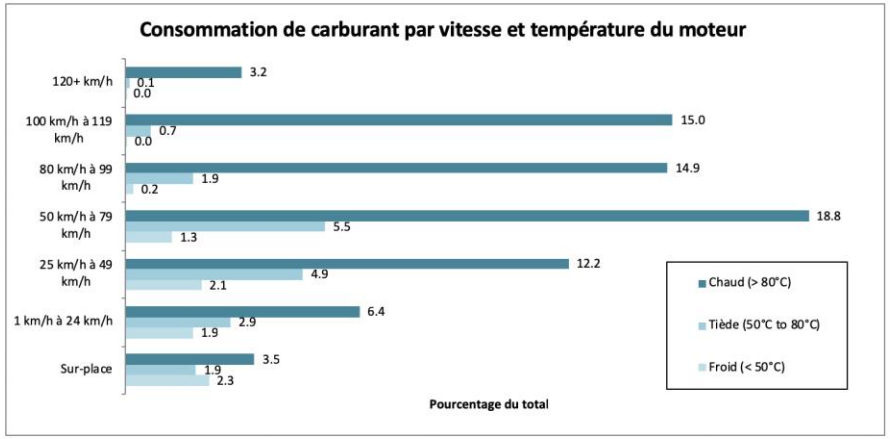
Durée de la conduite (min) par jour-type et occupation



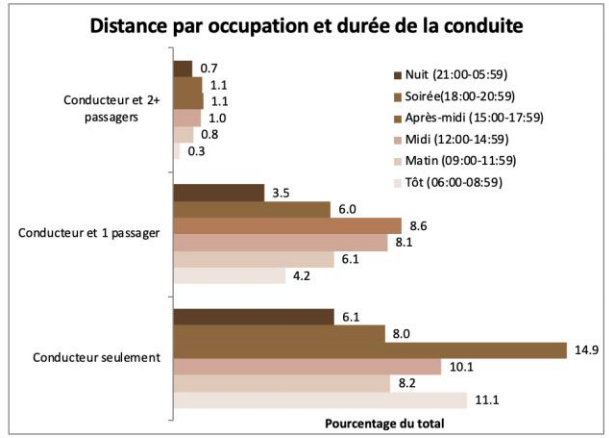
Durée de la conduite (min) par but et occupation



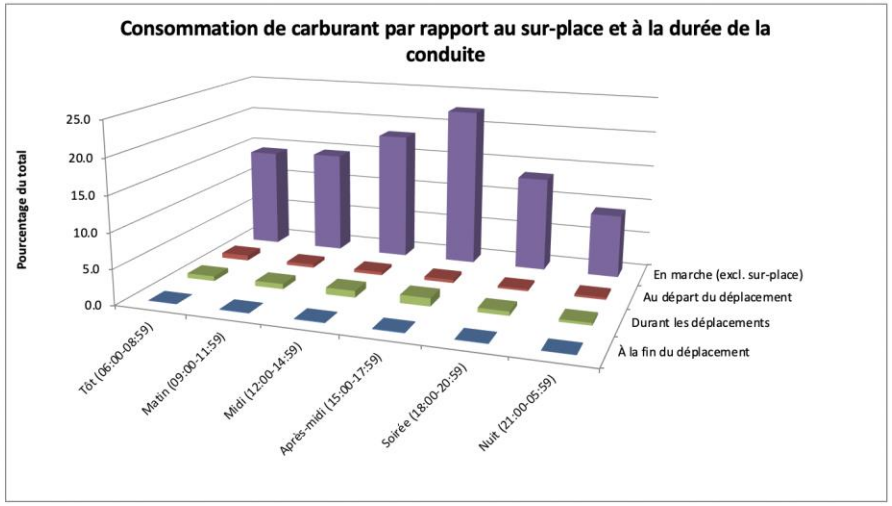
Consommation de carburant par vitesse et température du moteur



Distance par occupation et durée de la conduite



Consommation de carburant par rapport au sur-place et à la durée de la conduite



What-If Analysis: Impact of Minimum Wage

[https://bigbookofdashboards.com/dashboards.html]

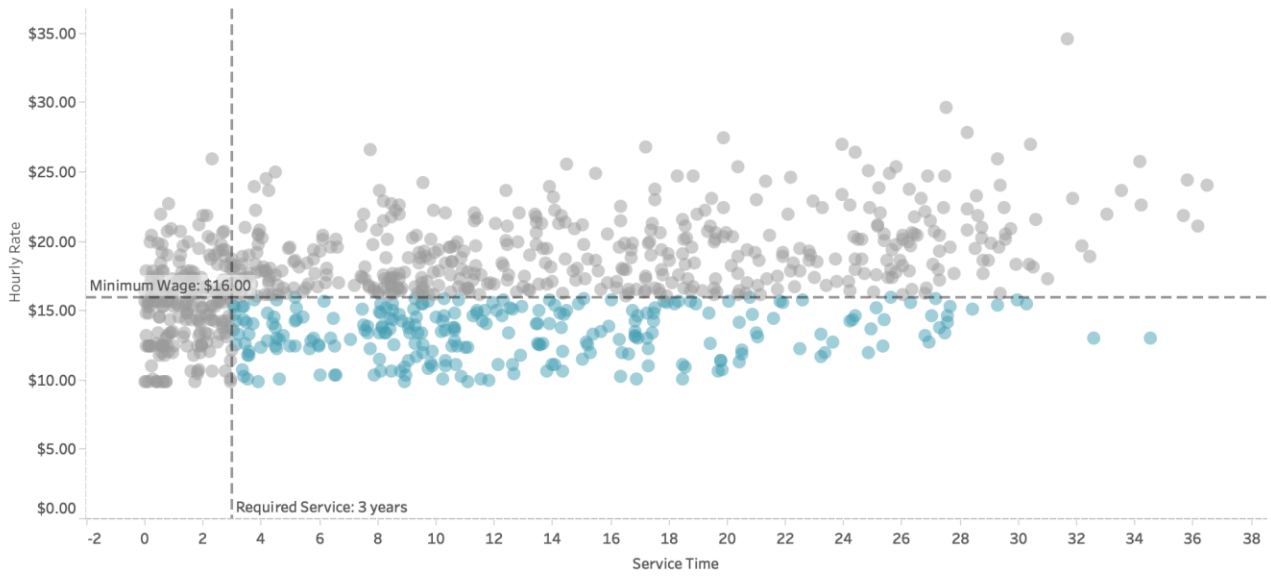
Proposed Minimum Wage: $16.00

Required Service: 3

Developed by Matt Chambers
http://sirvizalot.blogspot.com/

Below
256 (27.9%)

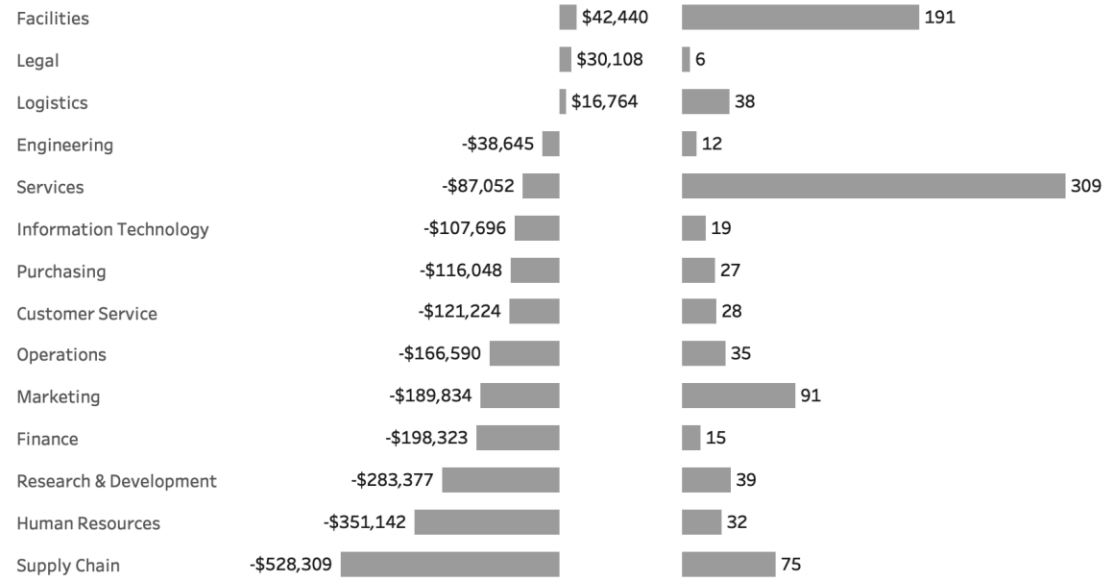
Above
661 (72.1%)



Employees Below Minimum Wage: 661

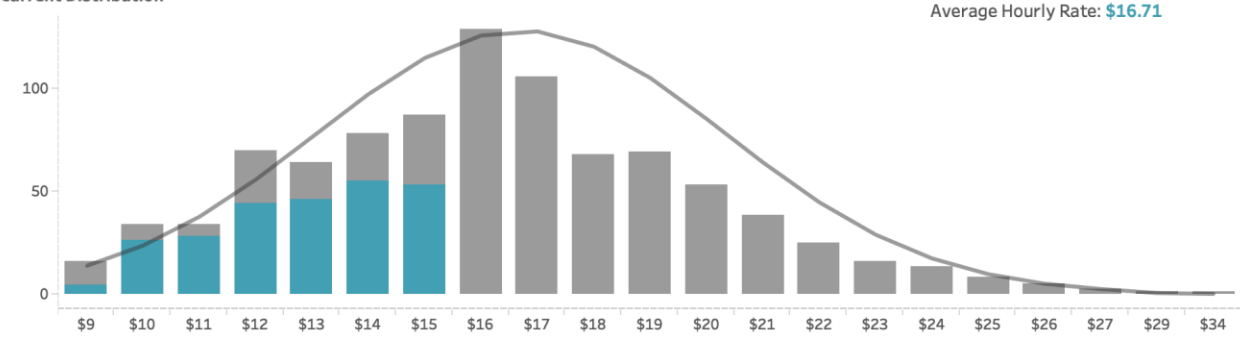
Dollar Impact of Minimum Wage: $1,792,206

Employees Below Minimum Wage: 256



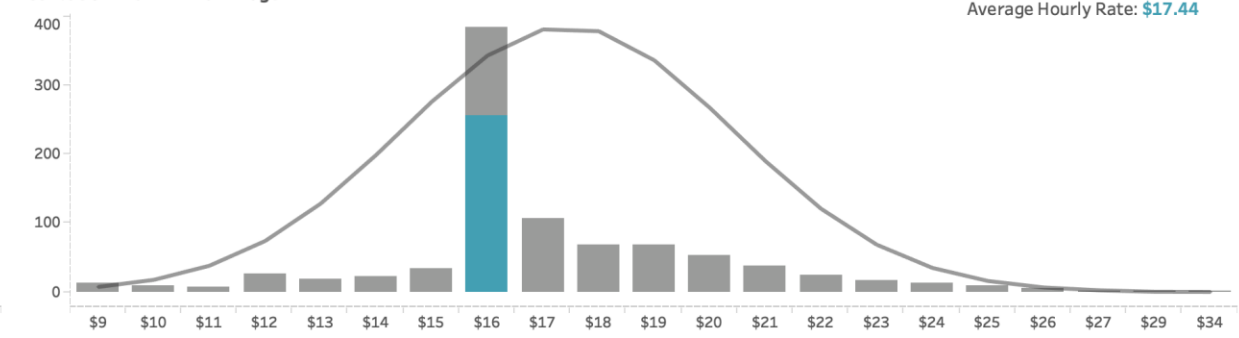
Current Distribution

Average Hourly Rate: $16.71



Distribution with Minimum Wage

Average Hourly Rate: $17.44



WHAT'S WRONG?

Dashboard #1: not glanceable, overuse of colour, pie charts??

Dashboard #2: 3D visualizations, distracting borders and background, lack of filtered data, insufficient labels and context

Dashboards #3 – 4: …

EXERCISE

In teams or individually, identify a scenario for which a dashboard could prove useful.

Determine specific questions that the dashboard could help answer or insights that it could provide.

Identify data sources and data elements that could be fed into your dashboard.

Design a display (with pen and paper) with mock charts.

What are the strengths and limitations of your dashboard? Is it functional? Elegant?

SUPPLEMENTAL MATERIAL

ADVANCED DATA VISUALIZATION AND REPORTING

[https://medium.com/gooddata-developers/ux-anatomy-of-dashboards-73dbf6515199]

DASHBOARD DESIGN ELEMENTS

GoodData takes an interesting approach to designing dashboards.

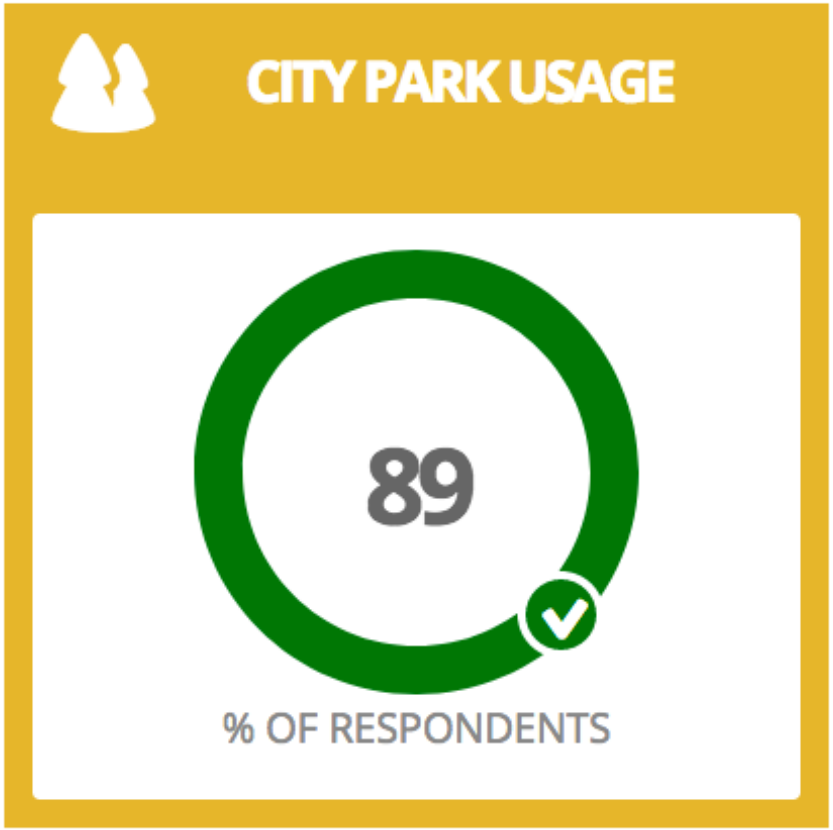
They suggest dashboards are composed of:

- Atoms
- Molecules
- Organisms

HOW WILL YOU VISUALIZE THE DATA?

To help generate visualization ideas:

- https://datavizcatalogue.com
- https://datavizproject.com
- https://rawgraphs.io



TYPE OF DATA YOU WILL FEATURE?

- What type of data do you want to feature?
- What type of data do you need to tell your story?
- What datasets do you need to have available to do this?

ANALYSIS ELEMENTS (I)

Will you be considering:

- What has already happened – hindsight?
- What is going on currently -situational awareness?
- What will likely happen – prediction?

ANALYSIS ELEMENTS (II)

- Will you show your audience ‘raw’ numbers (number of cars on the road)?
- Will you need to come up with some aggregate measures (traffic level)?

SOME WRAP-UP QUESTIONS

- What did you learn through the dashboard exercise?
- Was there anything that surprised you?
- Would a dashboard be beneficial to your organization?
- What team or skills sets would you need to do this in real life?

REFERENCES

ADVANCED DATA VISUALIZATION AND REPORTING

REFERENCES

Understanding Graphics

Krygier, J., Wood, D., [2016], *Making Maps: A Visual Guide to Map Design for GIS*, Guilford Press

Interactive Data Visualization on Wikipedia

Is animation an effective tool for data visualization?, NASA

Perception in Visualization, C.G. Healey (very cool!)

Data Physicalizations

Tufte, E. [2001], *The Visual Display of Quantitative Information*, Graphics Press.

Hu, D. [1954], *How to Lie With Statistics*, Norton

Tufte, E. [2008], *Beautiful Evidence*, Graphics Press

REFERENCES

Nussbaumer Knaflic, C. [2015], *Storytelling with Data*, Wiley

Cairo, A. [2013], *The Functional Art*, New Riders

Cairo, A. [2016], *The Truthful Art*, New Riders

Meireilles, I. [2013], *Design for Information*, Rockport

50 Great Examples of Data Visualization: http://www.webdesignerdepot.com

Wexler, S., Shaffer, J., Cotgreave, A. [2017], *The Big Book of Dashboards*, Wiley.

Nathan Yau's FlowingData

Data Visualization on Wikipedia

Misleading Graphs on Wikipedia

REFERENCES

Prabhakaran, S., Top 50 ggplot2 Visualizations (with Master List R Code).

Miller, M. [2017], *The problem with Interactive graphics*, Co.Design

Wickham, H. [2016], *ggplot2: Elegant Graphics for Data Analysis* (2nd ed), Springer.

Gorelik, B., Data Visualization (blog).

Chang, W. [2013], R Graphics Cookbook, O'Reilly.

Wickham, H. [2009], A Layered Grammar of Graphics, Journal of Computational and Graphical Statistics 19:3–28.

Horton, N.J., Kleinman, K. [2016], Using R and RStudio for Data Management, Statistical Analysis, and Graphics, 2nd ed., CRC Press.

Healey, K. [2018], Data Visualization: A Practical Introduction.

REFERENCES

Kabacoff, R.I. [2011], R in Action, Second Edition: Data analysis and graphics with R, Live.

Maindonald, J.H. [2008], Using R for Data Analysis and Graphics: Introduction, Code and Commentary.

Tyner, S., Briatte, F., Hofmann, H. [2017], Network Visualization with ggplot2, The R Journal, vol. 9(1).

Broman, K. [2016], Data Visualization with ggplot2.

Robinson, D., Visualizing Data Using ggplot2, on varianceexplained.org.

Manipulating, analyzing and exporting data with tidyverse, on datacarpentry.org.

Wickham, H. [2014], Tidy Data, Journal of Statistical Software, v59, n10.

Gashim, E., Boily, P. [2018], A ggplot2 Primer, data-action-lab.com