COGS 9 – A05 Discussion

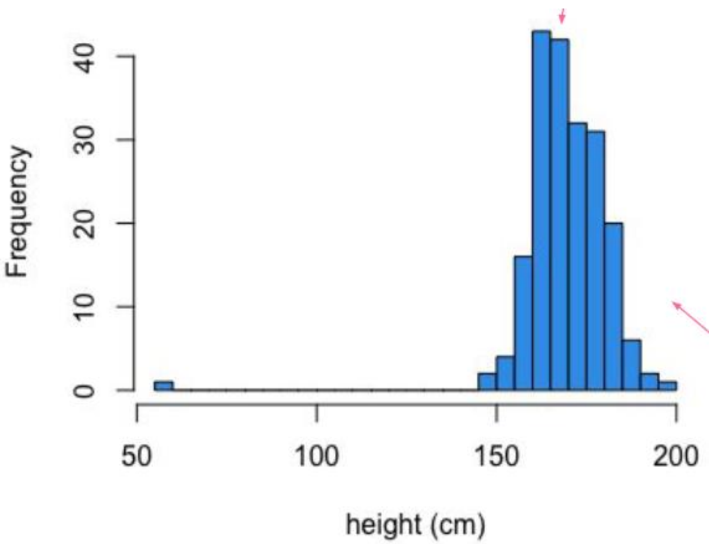
Deadlines

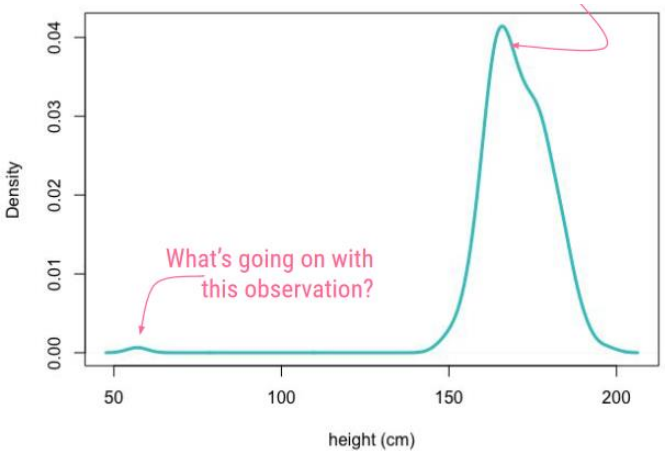
- Mid Way Team Evaluations E.C. – October 28th (Today)
  - 2 bonus points if you fill the form
- Assignment 2 – October 28th (Today)
  - 40 points
  - Read through and follow the instructions
- Reading Quiz 4 – November 3rd (Thursday)
  - 10 points

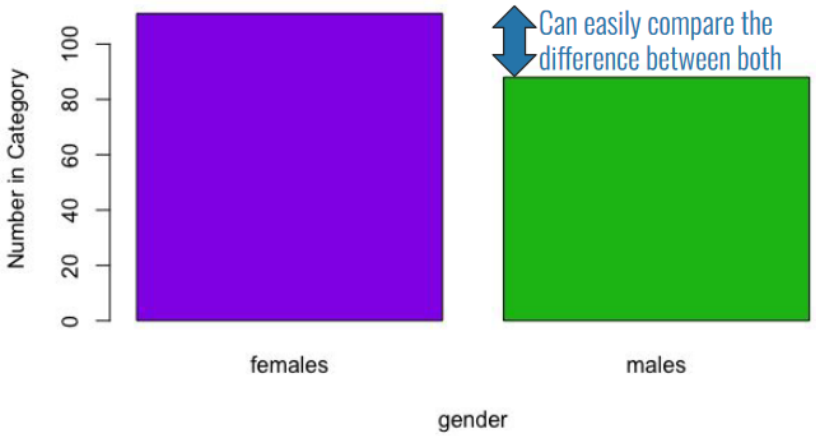
- Any issues with Assignment 1 grades?

Lecture 9: Data Visualisation

- What’s the difference between a histogram, densityplot and barplot?

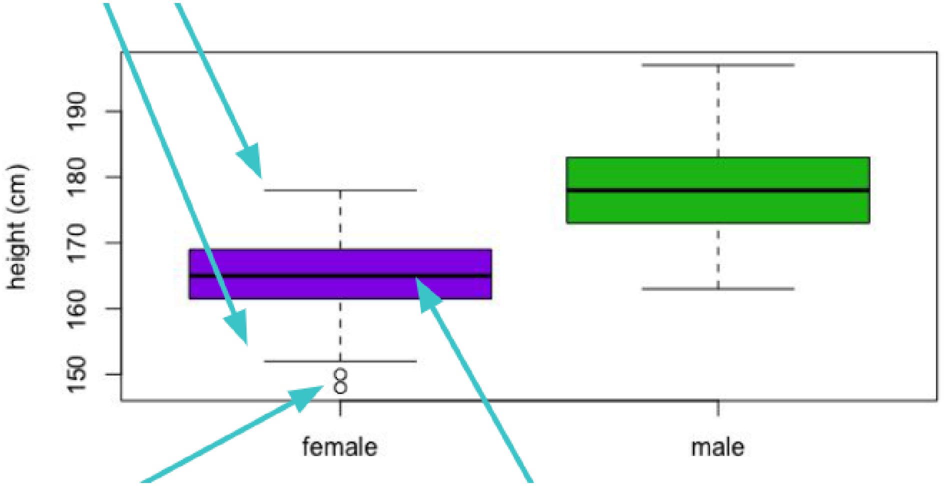


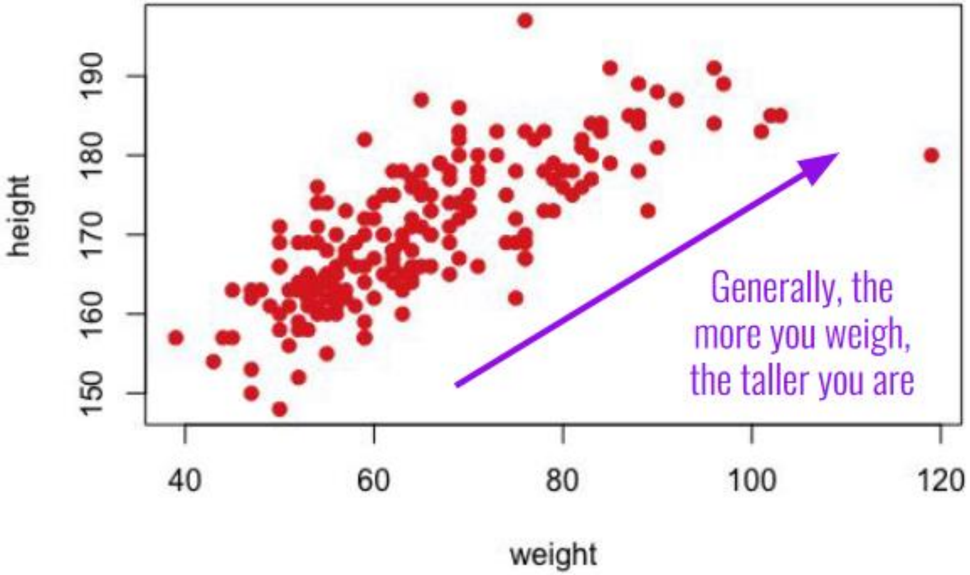

What’s going on with this observation?


Can easily compare the difference between both

Lecture 9: Data Visualisation

- When to use a scatterplot and a boxplot?

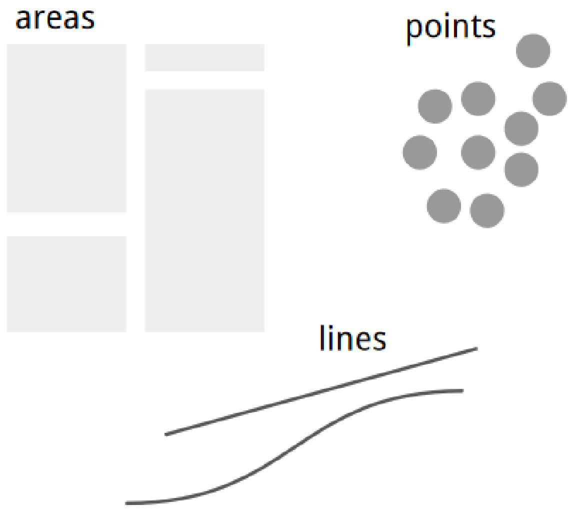




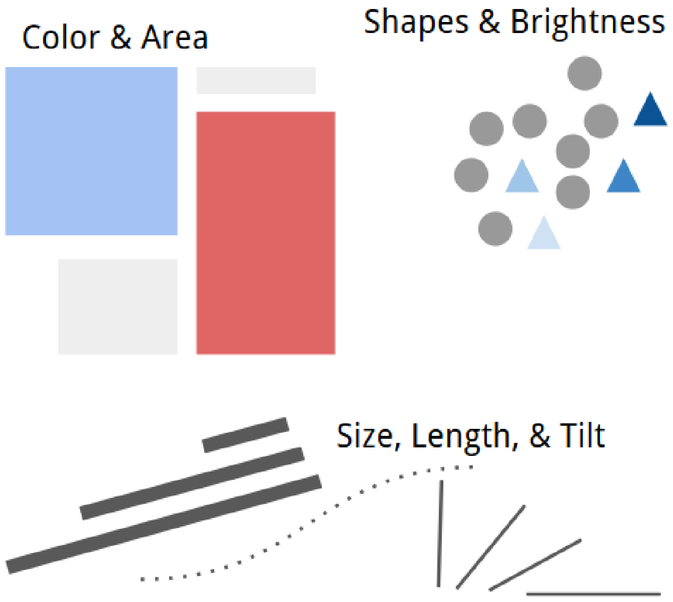
Generally, the more you weigh, the taller you are

Lecture 9: Data Visualisation

- Marks and channels



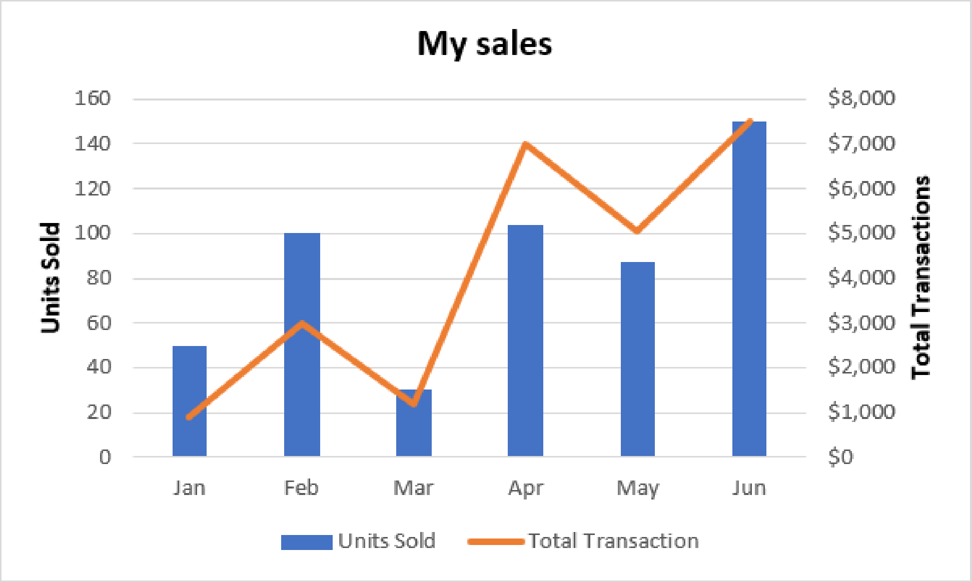
Marks: Geometric Primitives

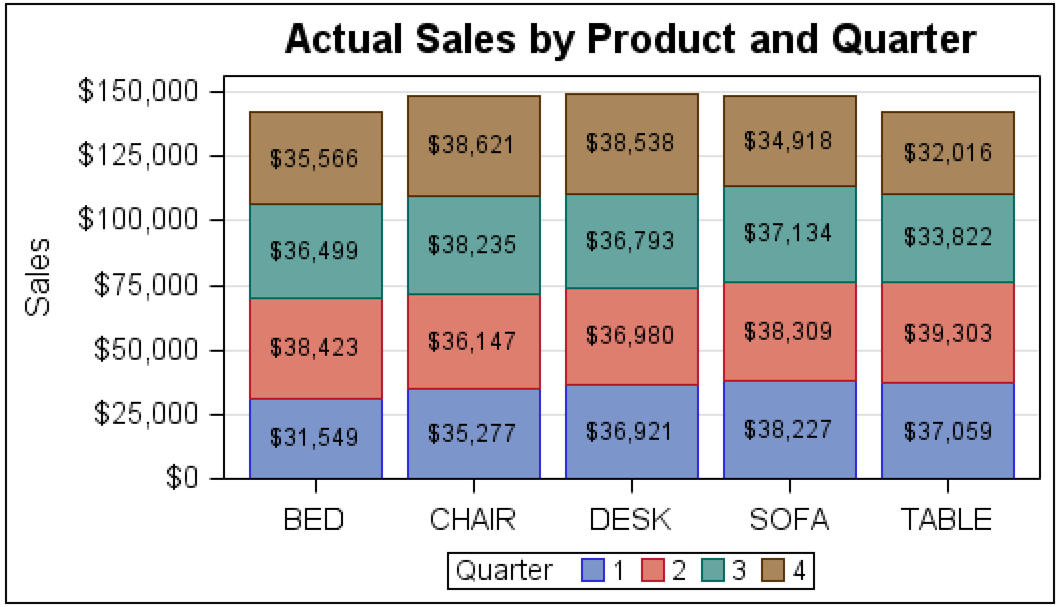


Channel: way to control the appearance of a Mark

Lecture 9: Data Visualisation

- Express: The visual shall express all of, and only, the information provided by the data's attributes. The visual should not add anything extra to the data.
- Effect: How important an attribute is, must match the salience of the channel. Greater importance = greater salience, or more noticeable





Lecture 9: Data Visualisation

Checklist when creating graphs

[] Consideration for Colorblindness

[] Label the axes

[] Ensure that the data is correct

[] Ensure that the graphic represents the data

[] Make the comparison easy on readers

[] Ensure that the y-axis starts at 0 (What about x-axis?)

[] Choose best visual

[] Keep it Simple Stupid

Lecture 9: Data Visualisation

Checklist when creating tables

[] Have a top to bottom comparison

[] Logical row ordering

[] Logical column ordering

[] Limit number of rows and columns

[] Informative headers

[] Fix significant digits

[] Format table

Lecture 9: Descriptive Analysis

- Suppress some of the truth so that humans can understand easily
- Size, shape, missingness, central tendency, variability
- Size: Number of variables and observations
- Shape: Distribution of the variables (Uniform, bimodal, Normal/Gaussian/Bell-shaped, left & right skewed, random)
- Missingness: How much data is missing?
- Central Tendency: Mean, median, mode
- Variability: Variance, Standard Deviation, Range

Lecture 10: Exploratory Data Analysis

- Data -> Descriptive Analysis -> Exploratory Analysis -> Product
- Exploratory: Inferential, Predictive, Causal, Mechanistic
    - Inferential: Statistics, Frequentist, Bayesian, Text & Geospatial analysis
    - Predictive: Statistical Learning/ML, Deep, Reinforcement Learning
    - Causal: How variable X correlates to Y
    - Mechanistic: How much does variable X affect Y
- Univariate, Bivariate, Multivariate
- Explanatory (Independent) vs Response (Dependent) variables
- Source of data (Zipcode vs hometown), explore missing data
- Don’t do EDA to give you the result you want

Lecture 10: Exploratory Data Analysis

- Checklist of things to do during EDA

[] Investigate missing values

[] Understand outliers

[] Add filters, transform and scale data

[] Calculate numerical summaries

[] Generate plots to explore relationships

[] Handle proportions correctly

[] Use tables to scan data

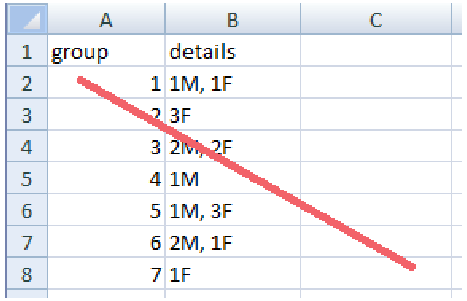
[] Search for patterns

Lecture 9 and 10 Demo

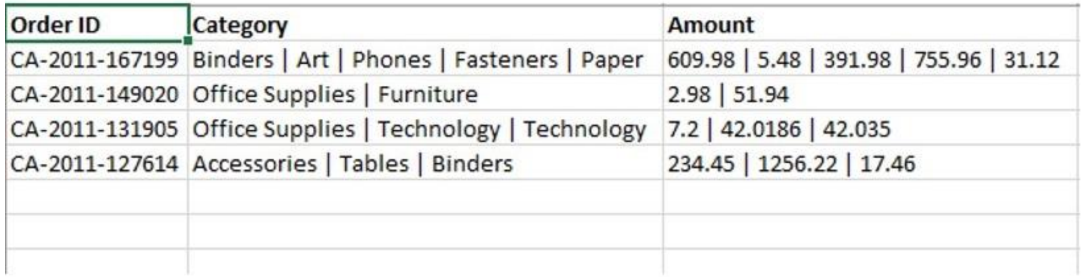
Reading 3: Tidy Data

- Tidy datasets provide a standardized way to link the structure of a dataset (physical layout) with its meaning
- Structure: Rectangular (Rows and columns)
- Semantics: Numbers (Quantitative), Strings (Qualitative)
- Each variable forms a column, each observation forms a row and each type of observational unit forms a table (Codd's 3rd Normal Form)
- How to tidy messy datasets
- Tools for tidying through manipulation, visualization & modelling
- Case study on tidying data using R

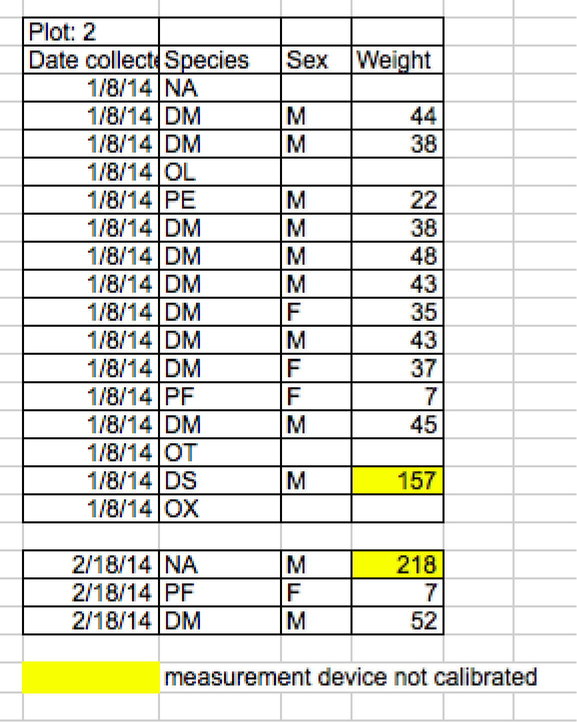
Reading 3: Data organization in spreadsheets



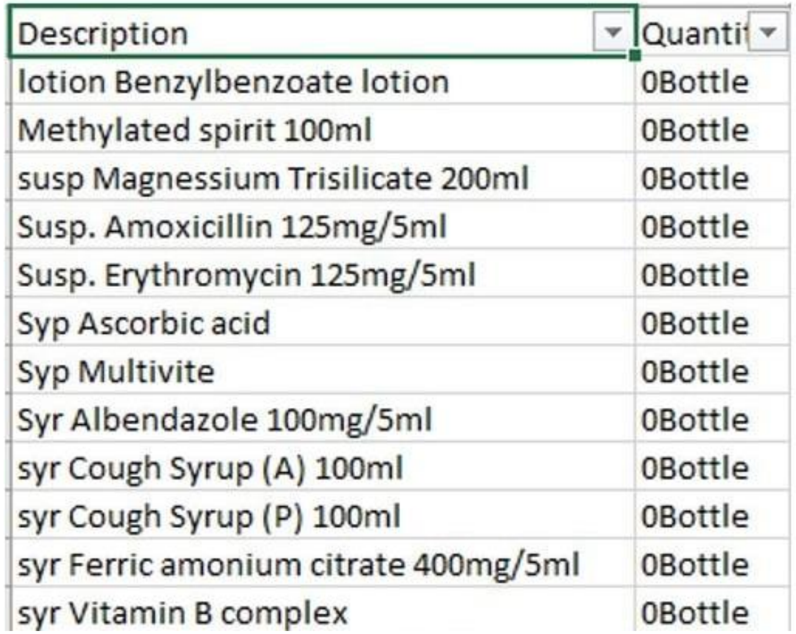
	A	B	C
1	group	details	
2	1	1M, 1F	
3	2	3F	
4	3	2M, 2F	
5	4	1M	
6	5	1M, 3F	
7	6	2M, 1F	
8	7	1F	



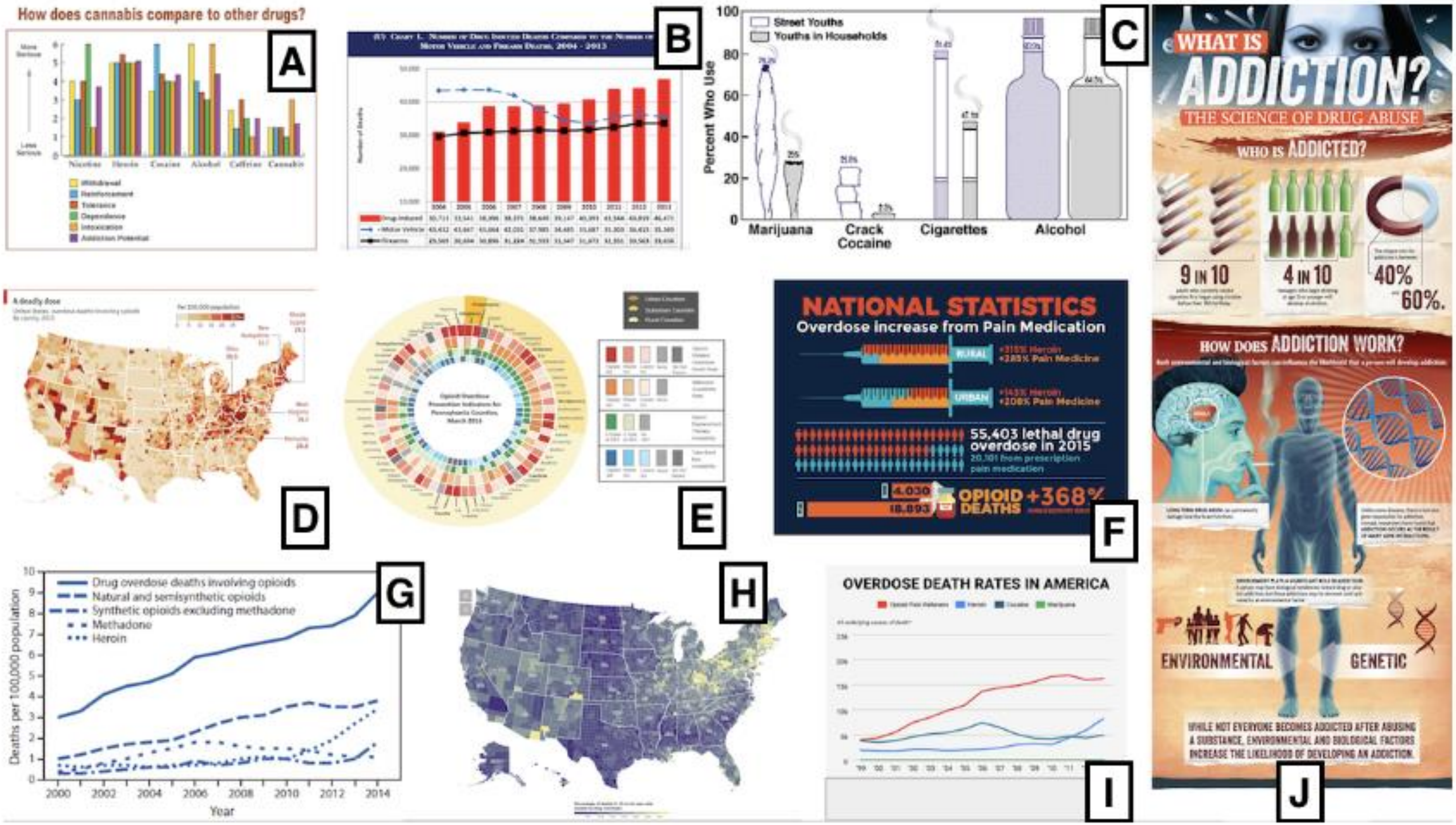
Order ID	Category	Amount
CA-2011-167199	Binders | Art | Phones | Fasteners | Paper	609.98 | 5.48 | 391.98 | 755.96 | 31.12
CA-2011-149020	Office Supplies | Furniture	2.98 | 51.94
CA-2011-131905	Office Supplies | Technology | Technology	7.2 | 42.0186 | 42.035
CA-2011-127614	Accessories | Tables | Binders	234.45 | 1256.22 | 17.46



Plot: 2			
Date collecte	Species	Sex	Weight
1/8/14	NA		
1/8/14	DM	M	44
1/8/14	DM	M	38
1/8/14	OL		
1/8/14	PE	M	22
1/8/14	DM	M	38
1/8/14	DM	M	48
1/8/14	DM	M	43
1/8/14	DM	F	35
1/8/14	DM	M	43
1/8/14	DM	F	37
1/8/14	PF	F	7
1/8/14	DM	M	45
1/8/14	OT		
1/8/14	DS	M	157
1/8/14	OX		
2/18/14	NA	M	218
2/18/14	PF	F	7
2/18/14	DM	M	52
🟨	measurement device not calibrated		



Description	Quanti
lotion Benzylbenzoate lotion	0Bottle
Methylated spirit 100ml	0Bottle
susp Magnessium Trisilicate 200ml	0Bottle
Susp. Amoxicillin 125mg/5ml	0Bottle
Susp. Erythromycin 125mg/5ml	0Bottle
Syp Ascorbic acid	0Bottle
Syp Multivite	0Bottle
Syr Albendazole 100mg/5ml	0Bottle
syr Cough Syrup (A) 100ml	0Bottle
syr Cough Syrup (P) 100ml	0Bottle
syr Ferric amonium citrate 400mg/5ml	0Bottle
syr Vitamin B complex	0Bottle


A
How does cannabis compare to other drugs?
More Serious
Less Serious
0
1
2
3
4
5
6
Nicotine
Heroin
Cocaine
Alcohol
Caffeine
Cannabis
Withdrawal
Reinforcement
Tolerance
Dependence
Intoxication
Addiction Potential
B
Drug Induced
Motor Vehicle
Firearms
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
C
Street Youths
Youths in Households
Percent Who Use
0
20
40
60
80
100
Marijuana
Crack Cocaine
Cigarettes
Alcohol
D
A deadly dose
E
F
NATIONAL STATISTICS
Overdose increase from Pain Medication
RURAL
URBAN
55,403 lethal drug overdose in 2015
20,101 from prescription pain medication
4,030
18,893
OPIOID DEATHS
+368%
G
Drug overdose deaths involving opioids
Natural and semisynthetic opioids
Synthetic opioids excluding methadone
Methadone
Heroin
Deaths per 100,000 population
0
1
2
3
4
5
6
7
8
9
10
2000
2002
2004
2006
2008
2010
2012
2014
Year
H
I
OVERDOSE DEATH RATES IN AMERICA
Opioid Pain Relievers
Heroin
Cocaine
Marijuana
J
WHAT IS
ADDICTION?
THE SCIENCE OF DRUG ABUSE
WHO IS ADDICTED?
9 IN 10
4 IN 10
40%
60%.
HOW DOES ADDICTION WORK?
ENVIRONMENTAL
GENETIC
WHILE NOT EVERYONE BECOMES ADDICTED AFTER ABUSING A SUBSTANCE, ENVIRONMENTAL AND BIOLOGICAL FACTORS INCREASE THE LIKELIHOOD OF DEVELOPING AN ADDICTION.

Reading 4: Attitudes and Perceptions of Data Visualization

- Which visualization do people understand? Which do they pay attention to? Study in a rural area in Pennsylvania
- Interviewed 42 people (diversified in terms of education, age, political affiliation, drug impact) – Asked to rank 10 graphics based on usefulness
- People ranked according to the following
  - Events that occurred in their personal lives
  - Geographical information which impacted them
  - How useful it is to other people
  - Clarity and Novelty
  - Statistical familiarity
  - Source of the graphic (Whether people changed their ranking)
  - Political Identity