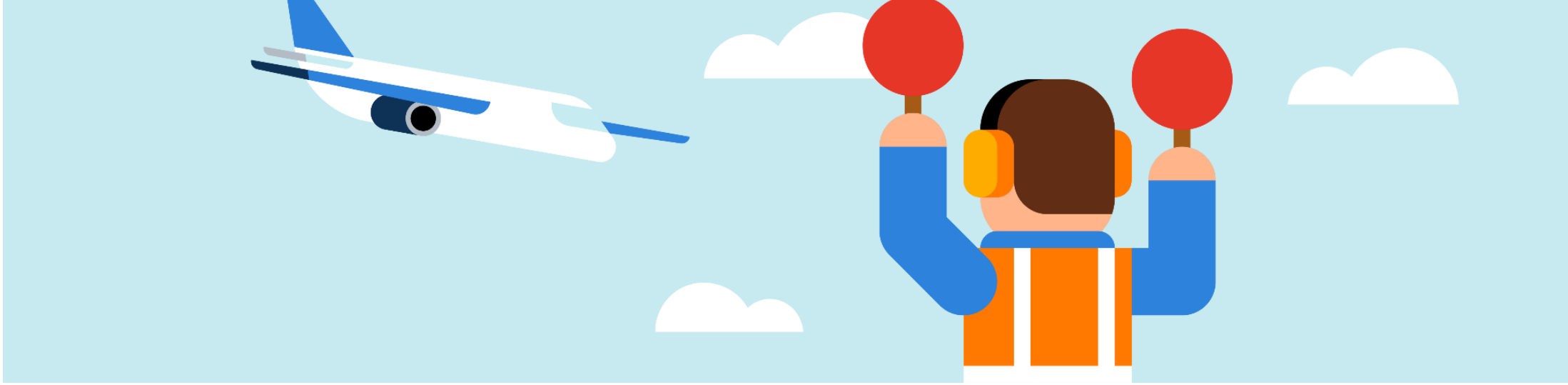


Common Data Science mistakes

Tomaž Kaštrun, MVP

SQLBits 2024
March, 21 2024

sqlbits

About

- Data scientist | BI Developer | data analyst
- SQL Server, SAS, R, Python, C#, SAP, SPSS
- 20+years experience MSSQL, DEV, BI, DM
- Frequent community speaker, book author
- Avid coffee drinker & bicycle junkie

http://tomaztsql.wordpress.com

tomaz.kastrun@gmail.com

@tomaz_tsql

/in/tomaztsql

http://github.com/tomaztk

https://mvp.microsoft.com/PublicProfile/5002196

Session material:
tomaztk/Common_DataScience_Mistakes: Common Mistakes Data Sciencists do (github.com)

sqlbits

And….

- I am NOT a data scientist

Who is data scientist?

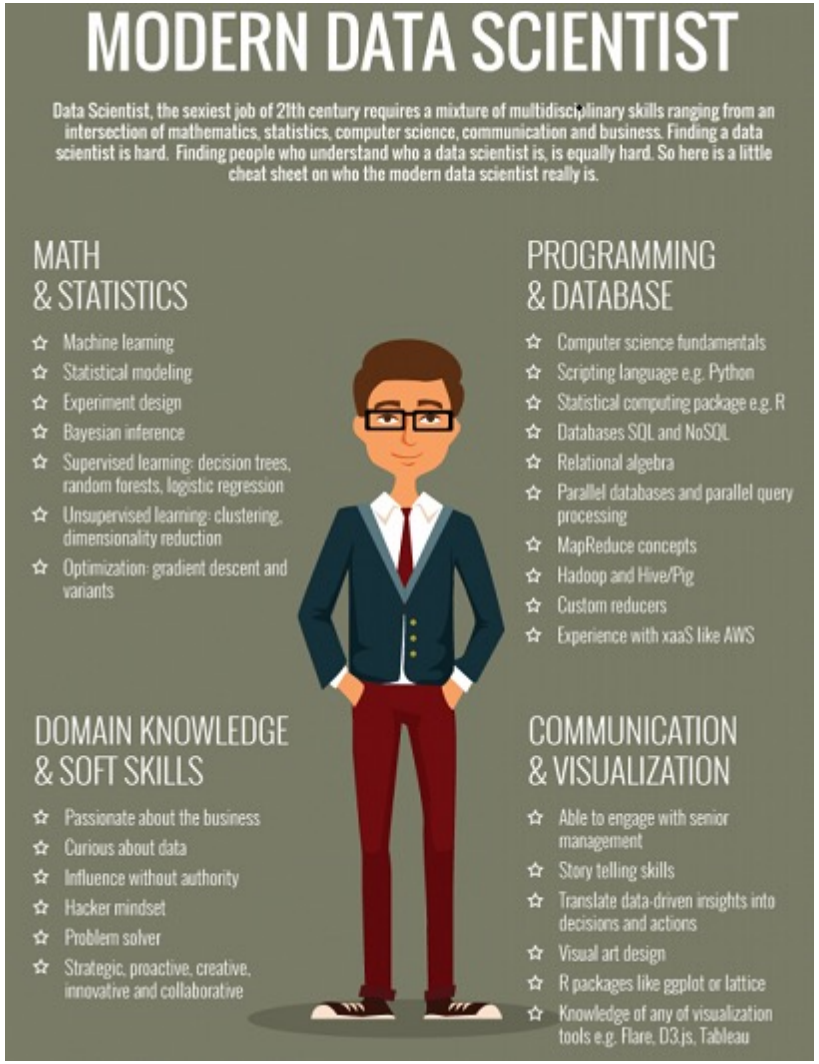
MacBook

Statistician

San Francisco

sqlbits

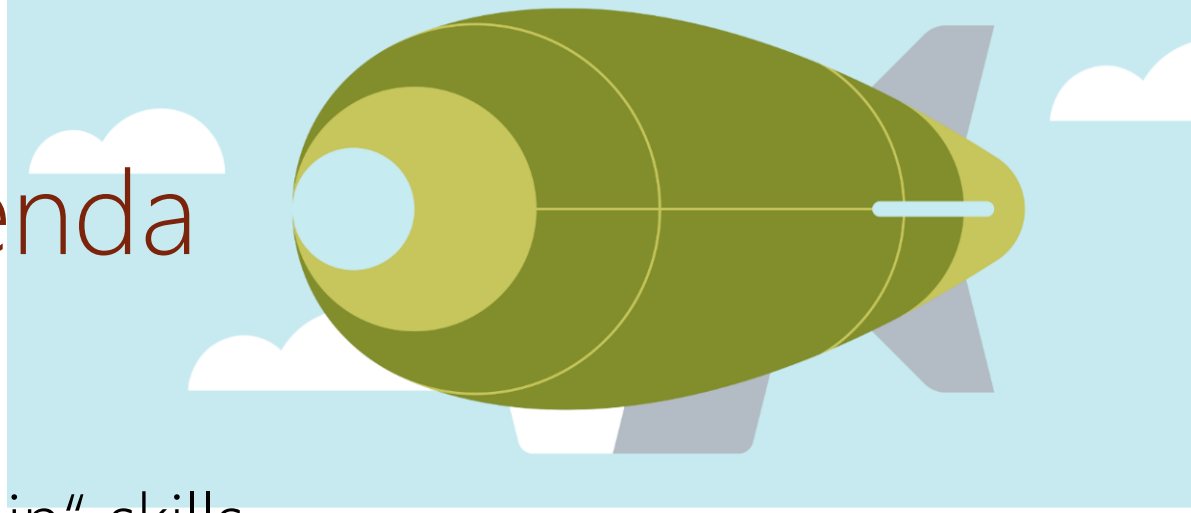
So… who is a data scienctist?



A statistician ☺

Source: Internet

Common mistakes -Agenda



Business mistakes -„Soft“ and „domain“ skills

Data preparation - General statistics

Data exploring - Getting data together

Data modeling and measuring - Analyzing and predicting

Data reporting and visualizations

Drawing conclusions

But first: Can you answer?

1. Explain **what regularization is** and why it is useful.
2. Explain what **precision and recall** are. How do they relate to the **ROC curve**?
3. What is **root cause analysis**?
4. Are you familiar with **pricing optimization, price elasticity, inventory management, competitive intelligence**? Give examples.
5. What is **statistical power**?
6. Explain **what resampling methods are and why they are useful**. Also explain their limitations.
7. Is it better to have **too many false positives, or too many false negatives**? Explain.
8. What is selection **bias,** why is it important and how can you avoid it?
9. Give an example of how you would use experimental design to answer a question about user behavior.
10. **How would you screen for outliers** and what should you do if you find one?
11. How would you use either the extreme value theory, Monte Carlo simulations or mathematical statistics (or anything else) to correctly estimate the chance of a very rare event?
12. What is a **recommendation engine**? How does it work?
13. Explain **what a false positive and a false negative are**. Why is it important to differentiate these from each other?

Source:
http://www.kdnuggets.com/2016/01/20-questions-to-detect-fake-data-scientists.html

Remember: *„It takes courage to grow and to be true to yourself“*

Business mistakes #1

Domain knowledge and soft skills

- Know your business, build a domain knowledge, understand your client
- Know yourself and your applied knowledge
- Focus on problems and not the tools
- Structure a plan, create a clear goal
- Don‘t over-complicate and don‘t over-simplfy

Remember: *„If you fail to plan, you plan to fail“*

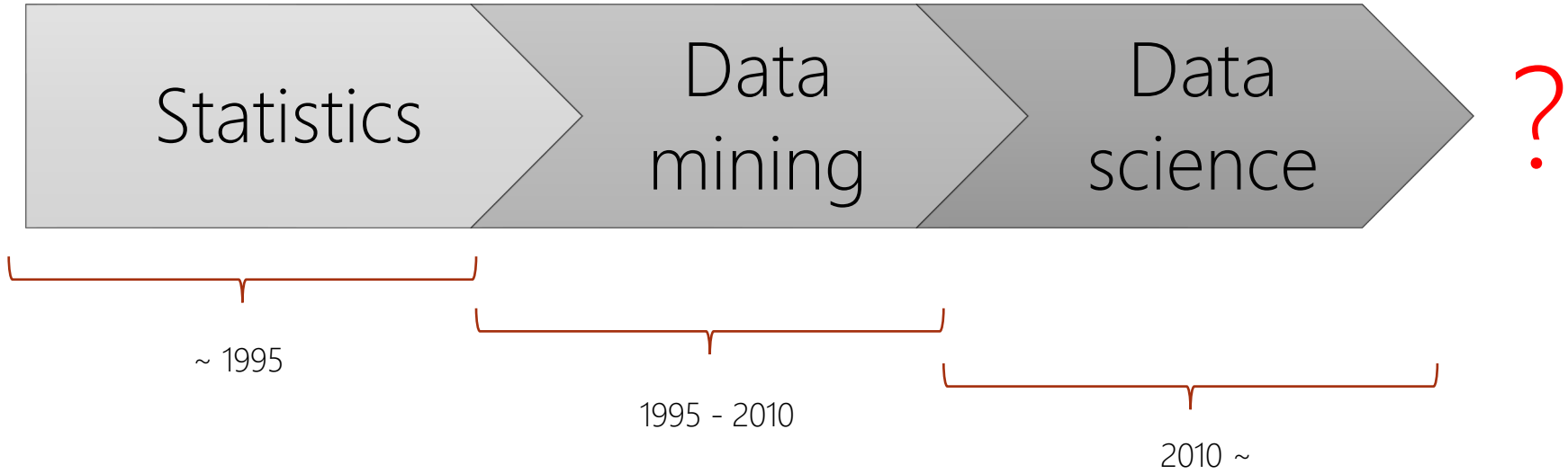
Business mistakes #2

Selecting different tools

- R or Python or Julia or KNIME or SAS or H2O or Spark or … ?
- T-SQL or PL/SQL or MySQL or NoSQL?
- SQL Server 2017 or 2019 or 2022?
- Java in SQL Server 2022?
- Azure? AWS? SAP?

Remember: *„Learn to know the difference. All will do roughly the same.“*

History ☺ Terms over time



Examples:

- Regression (Stats 101, SSAS, R/Py+MSSQL)
- Decision trees (Stats 101, SSAS, R /Py +MSSQL)
- Complexity reduction (Stats 101, SSAS, R /Py +MSSQL)
- Clustering (Stats 101, SSAS, R /Py +MSSQL)
- NN -> CNN | RNN | LSTM

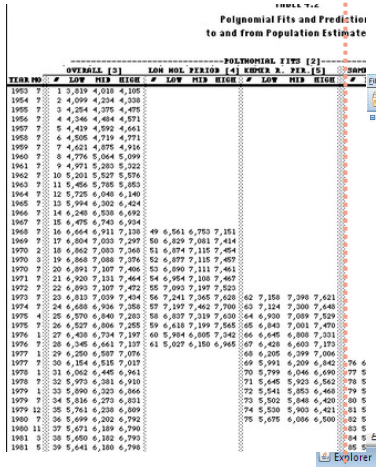
Remember: *„Almost all algorithms are predecessors of analysis of variance."*

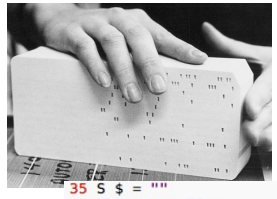
sqlbits

History ☺ Interfaces over time

Statistics → Data mining → Data science

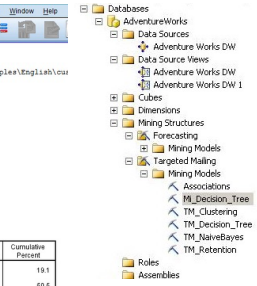
~ 1992

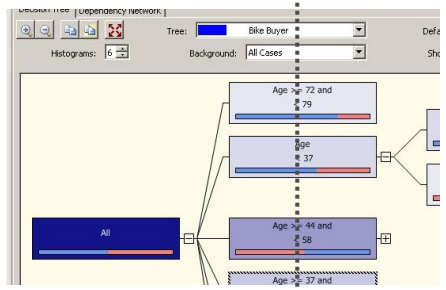


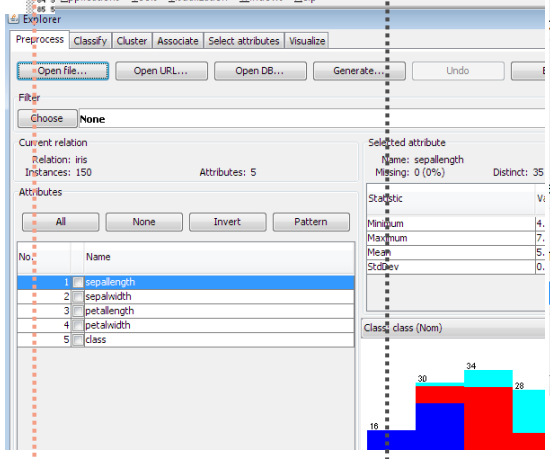


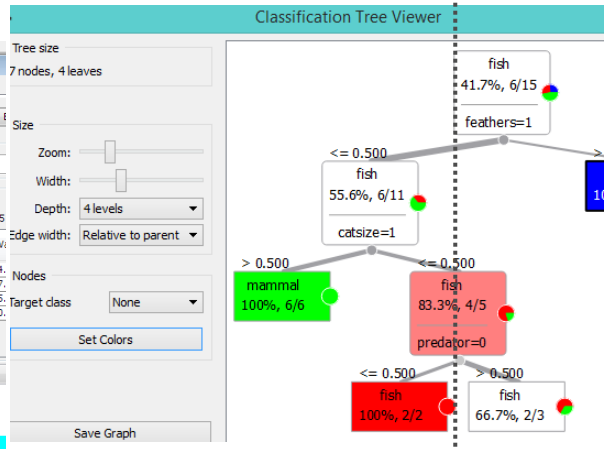
```
35 S $ = ""
40 FOR I = 1 TO S
50 S $ = S $ + "*"
55 NEXT I
60 PRINT S $
70 INPUT "Do you want more
80 IF LEN (Q $) = 0 THEN GO
90 L $ = LEFT $ (Q $, 1)
100 IF 30 (L $ = "Y") OR (L
110 PRINT "Goodbye";
120 FOR I = 1 TO 200
130 PRINT A $; "";
140 NEXT I
```

1992 - 2010









2010 ~

```
#crosstabulation / before using rxFactors
rxCrossTabs(V3 ~ F(V1), data = "YearPredictMSD.

YearMDS_2 <- rxFactors(inData = "YearPredictMSD

#save the results of factorization of variable
rxDataStep(YearMDS_2, outFile = "YearPredictMSD

#crosstabulation can be now run without factors
rxCrossTabs(V2 ~ V1, data = "YearPredictMSD.xdf

#crosstabulation with three variables
rxCrossTabs(V2 ~ V1:F(V3), data = "YearPredictM

#crosstabulation with three variables
Year_CT <- rxCrossTabs(V2 ~ V1: F(V3), data =

print(Year_CT, output = "counts")
print(Year_CT, output = "means")
```

```
from sklearn.ensemble import GradientBoostingClassifier
import numpy as np, pandas as pd

def azureml_main(dataframe1):
    colnames = dataframe1.columns
    y = np.array(dataframe1[colnames[-1]])
    X = np.array(dataframe1.ix[:, :len(colnames)-1])
    clf = GradientBoostingClassifier(n_estimators=100, \
        learning_rate=1.0, max_depth=1, random_state=0).\
        fit(X, y)
    fint = clf.feature_importances_
    fnames = np.array(colnames[:-1])

    perm = fint.argsort()

    ret = pd.DataFrame()
```

Remember: „*Choose your tool wisely, but do not kill the project!*

Data preparations

General data attributes

- Not ordering data when merging (R, Py, Scala, Julia)
- Having duplicated values when joining tables (SQL)
- Difference between data types and variables
- Treating the NULL, N/A, 0, {}, [] values
- Hanging out with outliers

Remember: *„Even simple overlooked things will cause big problems. Eventually.“*

Data Exploration

Getting first data insights

- Preliminary tests to check the test assumption
- Difference between causation and correlation (next slide) ☺
- Statistical test: test of proportions, difference of two proportions, chi-square test, test of mean, difference of two means (independent), difference of two means (paired)
  - Data
  - Samples
  - Purpose
- Distributions

Remember: „*You don't use COVID tests to check your pregnancy.*"

sqlbits

Demo >>00_Correlation_outliers.R
01_R_Data_Correlation_for_modellign

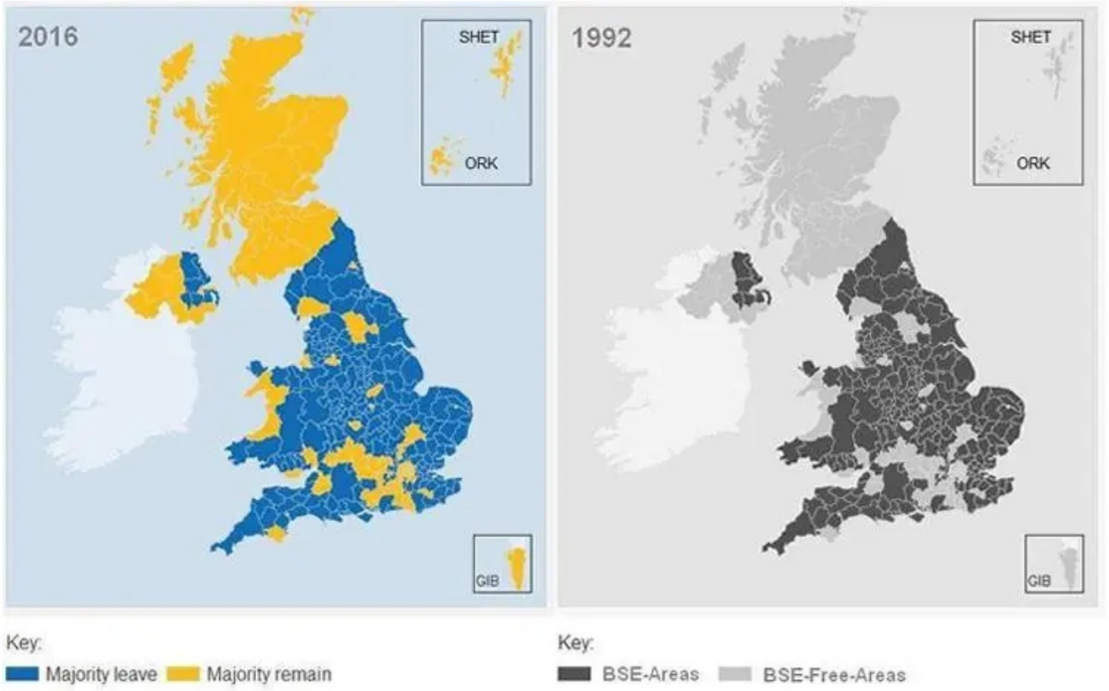
Correlation vs. causation

Brexit vote vs Mad Cow Disease: Why collaboration is key for Business Intelligence decision-making

Posted by: Yellowfin Team

Damn it. I was really hoping, for the sake of comedy amidst a time of potential calamity, that it was true. Alas, the Internet, rather than the truth, has won the day once more.

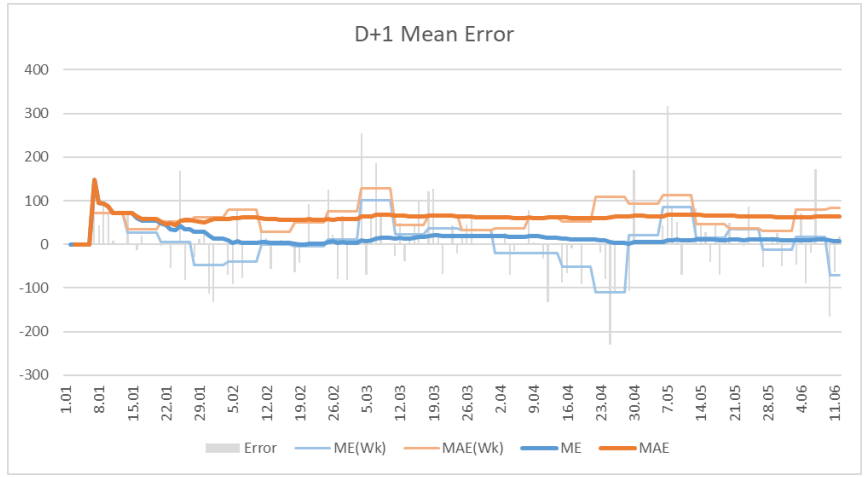
Shortly after the leave campaign triumphed in the 'Brexit' vote – Britain's referendum regarding its European Union (EU) membership – an amusing set of comparative map-based data visualizations appeared.



Source: https://www.yellowfinbi.com/blog/brexit-vote-vs-mad-cow-disease-why-collaboration-is-key-for-business-intelligenc

Data modeling and measuring

Analyzing and predicting #1



- Sampling, over-sampling, under-sampling
- Validating
- Normalization
- Know your metrics (e.g.: RMSE, MSE, MAE, MAPE, MdAPE, …)

Remember: „*Learn how to check what your data is doing and where it is going.*"

sqlbits

Demo >> 02_R_Cluster_dendrogram
>> *Excel_cashflow*

Data modeling and measuring

Analyzing and predicting #2

- Is AUC, ROC the only measure? Do we understand the model?
- Over- or Under- fitting?
- Evaluation of the model vs. Interpretation of the model
- Train the model on complete dataset or spliting the data (?)
- Deploying the model (frequency, model changes)
- Sticking only to one algorithm or using wrong one (?)

Remember: *„If you can't explain, what your model is doing to your five-year-old, it is not production ready."*

sqlbits

Demo >> 04_Linearity_nonlinearity
>> *99_2_Data_leakage_prevent*

Visualizing your conclusions

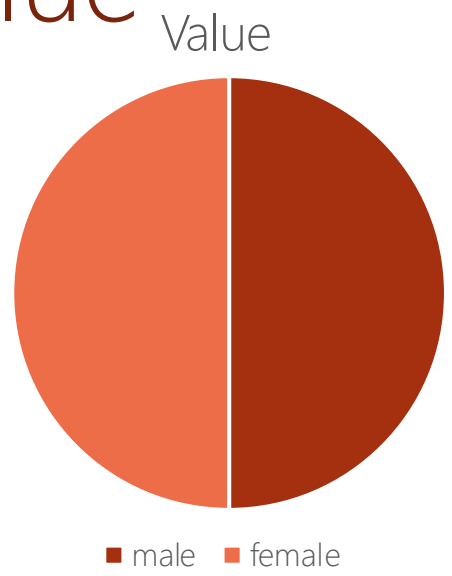
- Visualizing data with wrong charts
- Visualizing wrong data
- Tables vs. Graphs
- Results vs. Beautiful graphs (?)

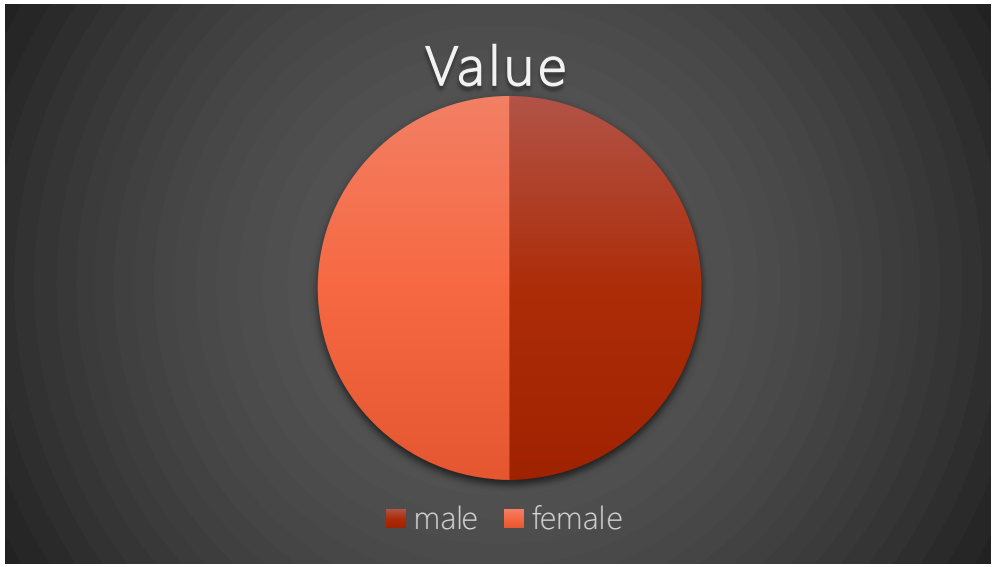
Remember: *„If you don‘t understand it the next day, don‘t expect your customer will.“*

sqlbits

Watch So much added value

Gender	Value of Question1
male	50
female	50

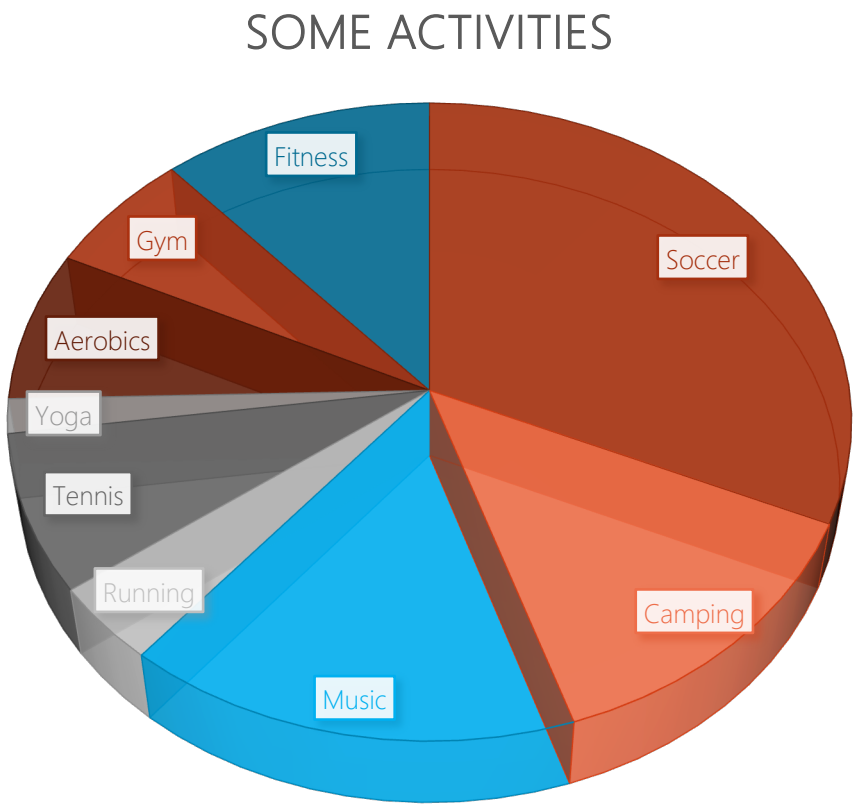




sqlbits

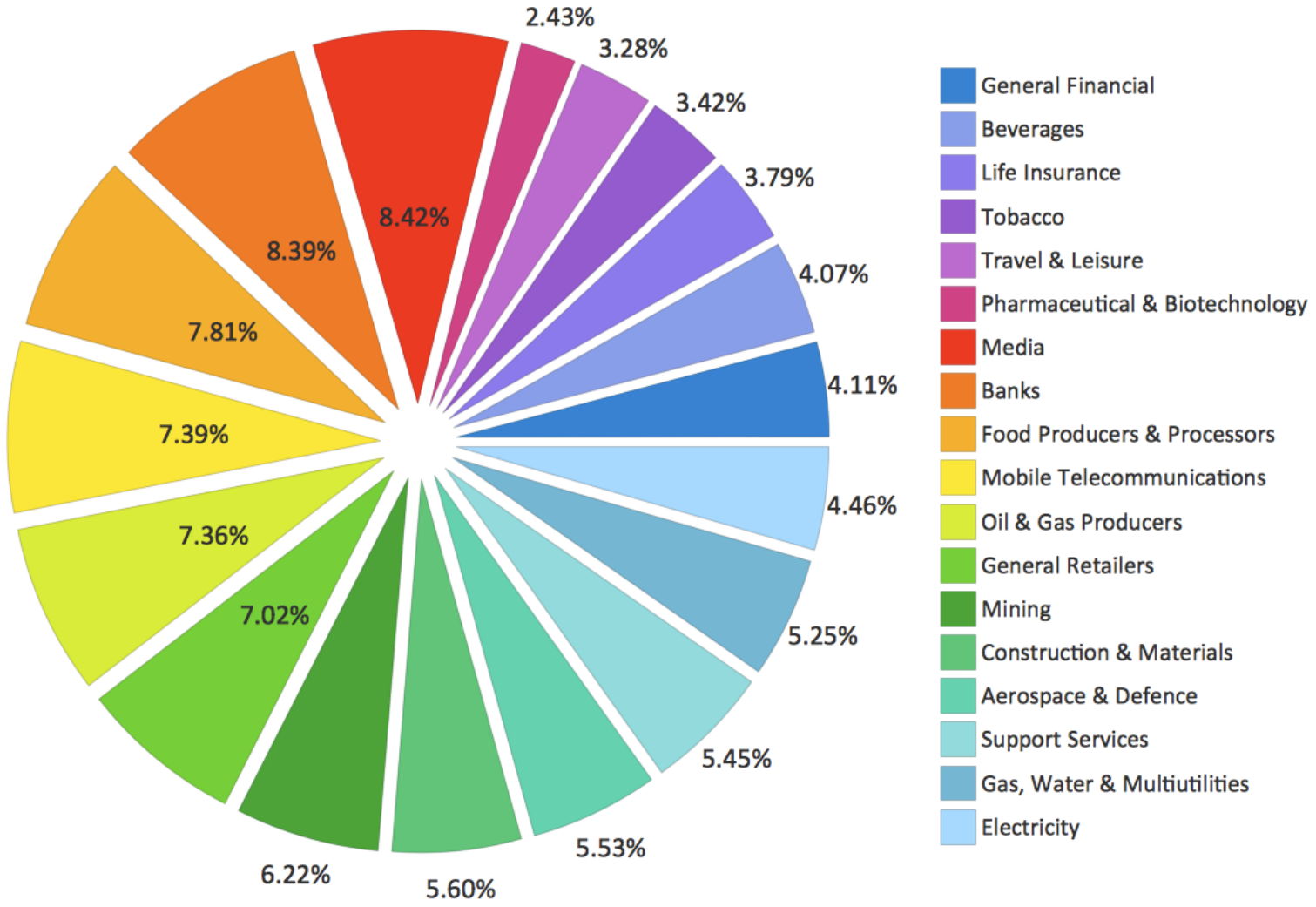
Sneeky grand totals!

Activities	[%]
Soccer	34,5
Camping	14,5
Music	17,4
Running	4,3
Tennis	8,7
Yoga	1,9
Aerobics	8,3
Gym	6,9
Fitness	12,5
Total:	109



WHAT ?

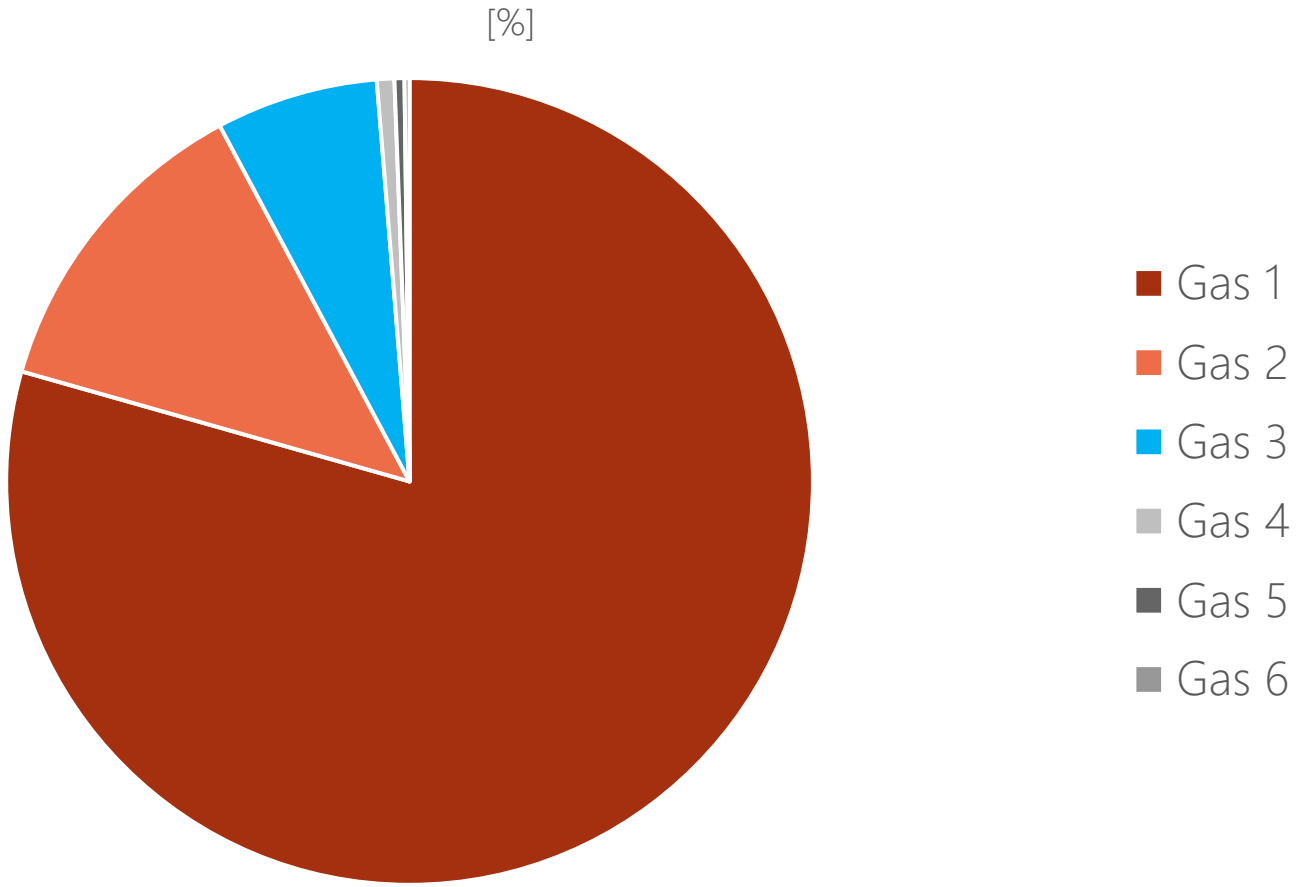
Things I just don't understand



Source: http://www.conceptdraw.com/How-To-Guide/pie-chart-examples

.. Or can't imagine

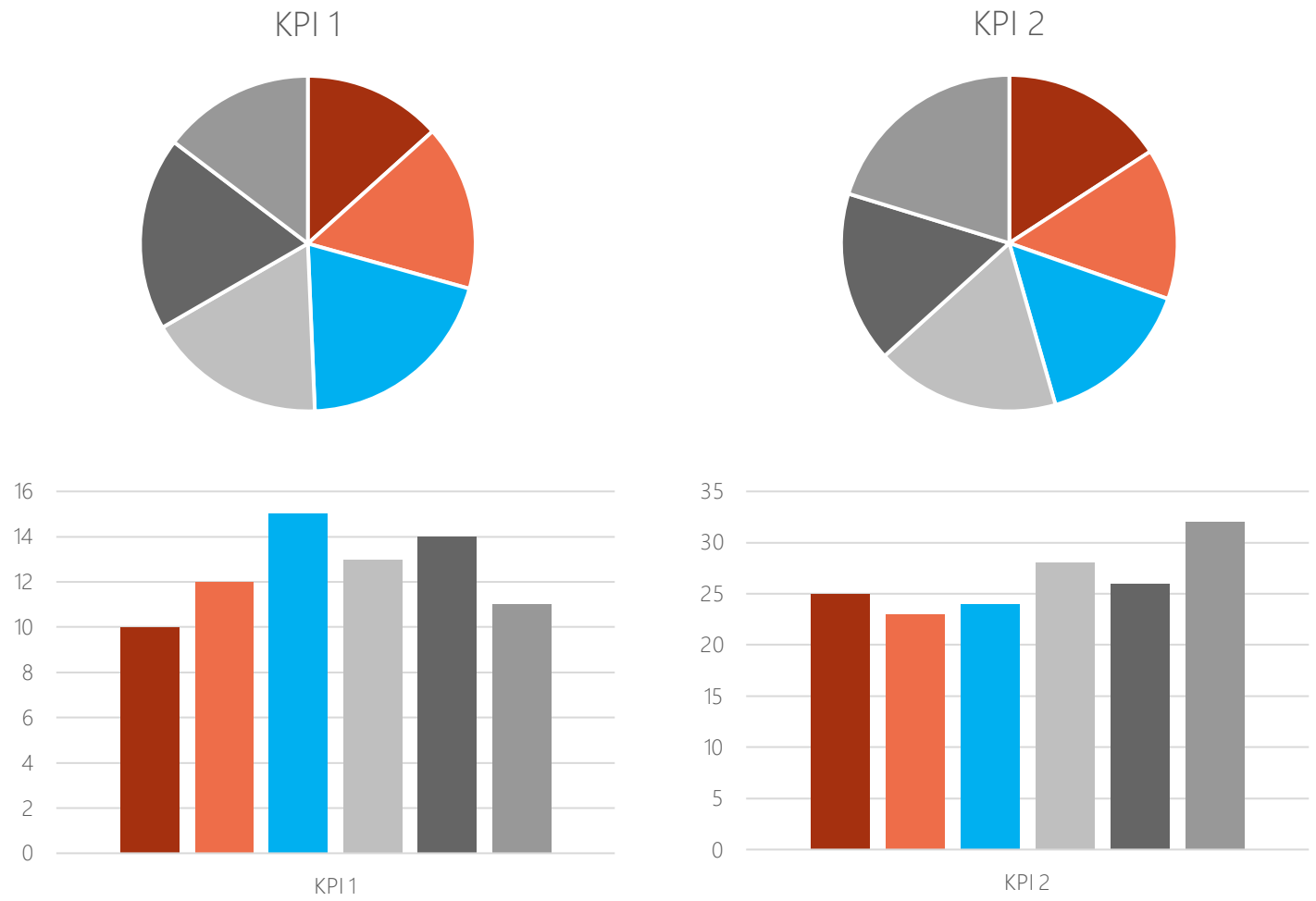
Class	[%]
Class 1	79,4
Class 2	12,8
Class 3	6,5
Class 4	0,7
Class 5	0,4
Class 6	0,2



By now… You've probably realized

That I can not stand pie charts!

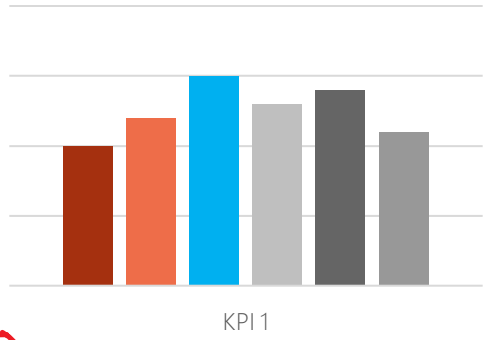
And this little piece of art!

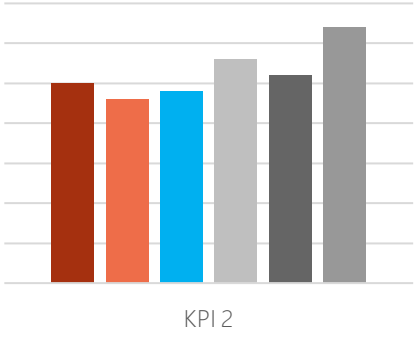


sqlbits

Is wrong on so many levels

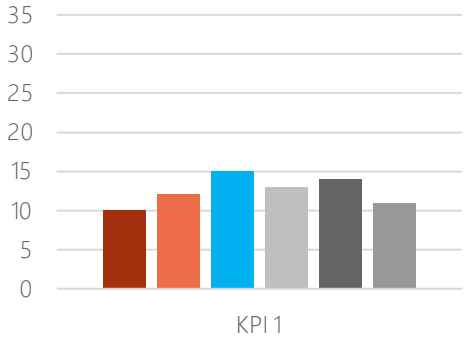
I get a report and start comparing KPI1 and KPI2

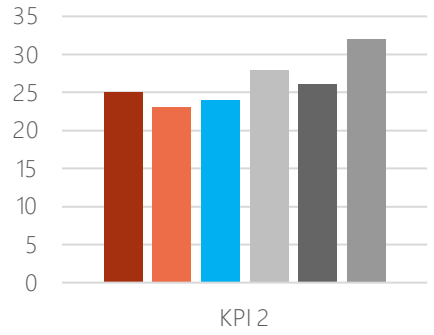




Only then I learn about the scales and numbers!

What at the should be!

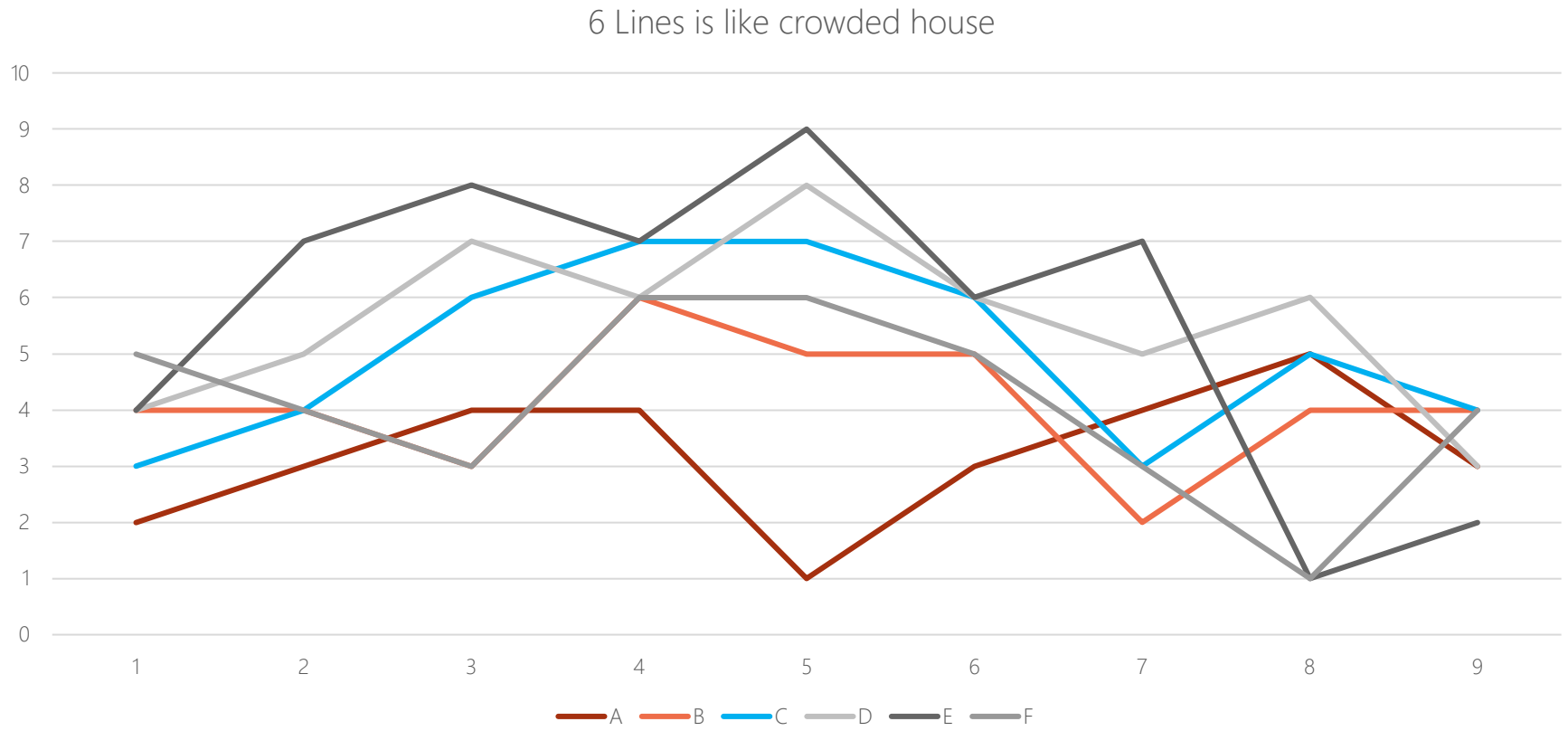




sqlbits

Over-complicating?

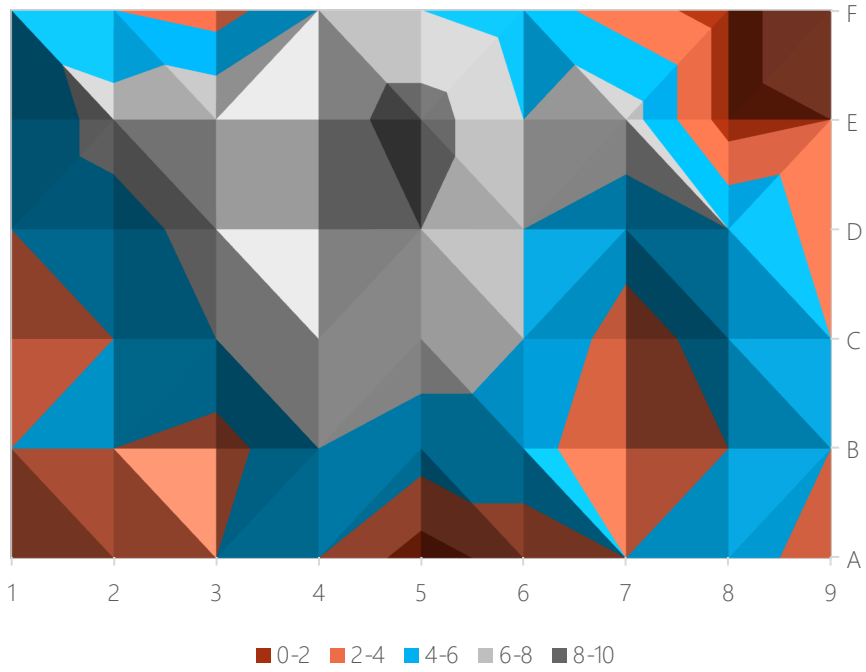
A	B	C	D	E	F
2	4	3	4	4	5
3	4	4	5	7	4
4	3	6	7	8	3
4	6	7	6	7	6
1	5	7	8	9	6
3	5	6	6	6	5
4	2	3	5	7	3
5	4	5	6	1	1
3	4	4	3	2	4



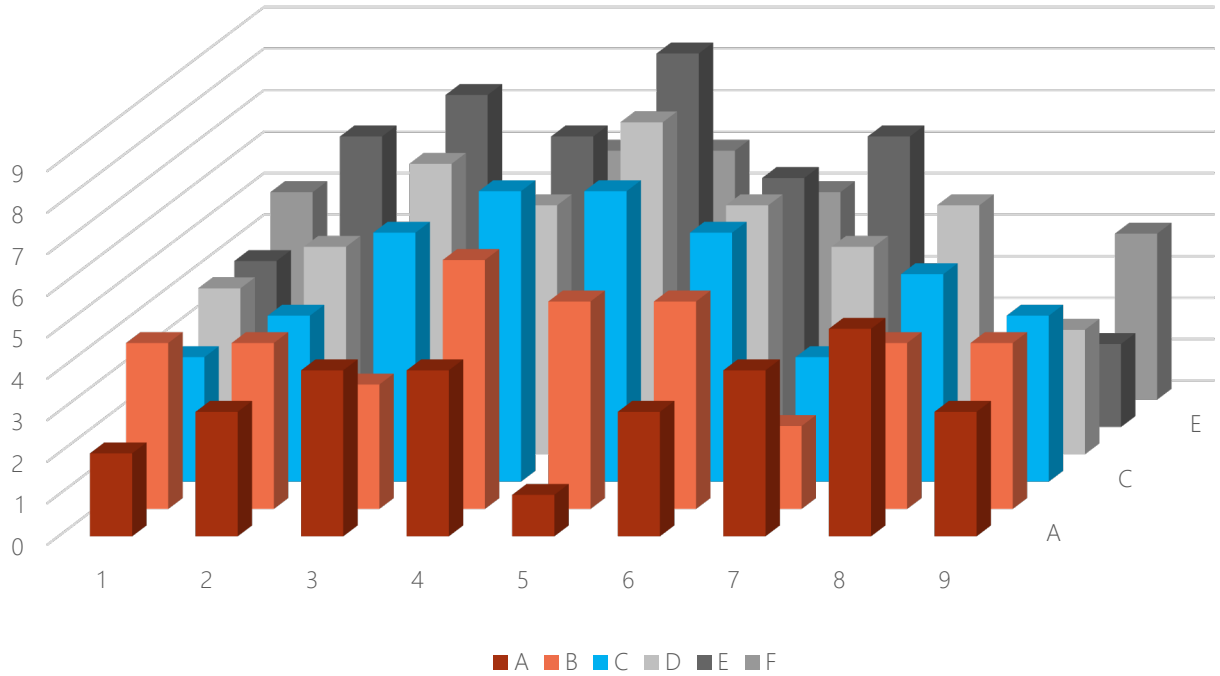
sqlbits

Or just playing Minecraft?

It is just too geeky



Hard to read 3D Graph

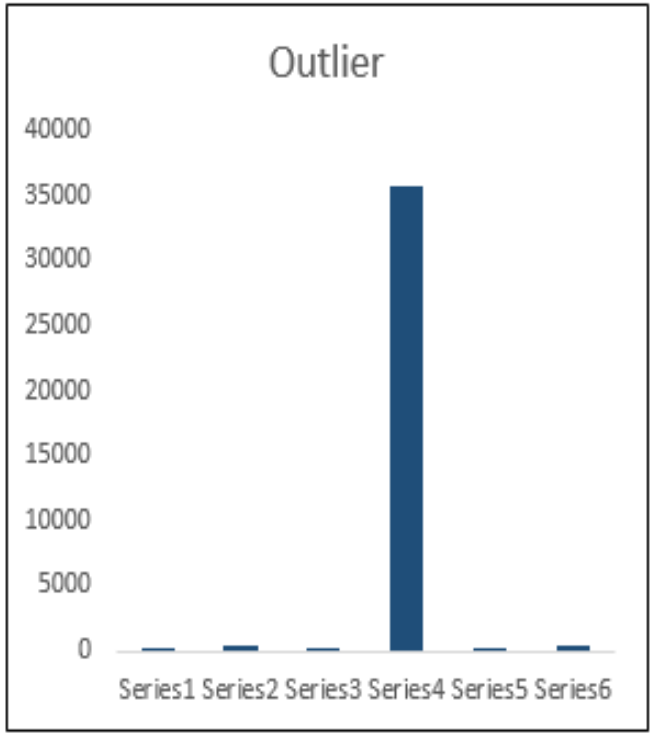


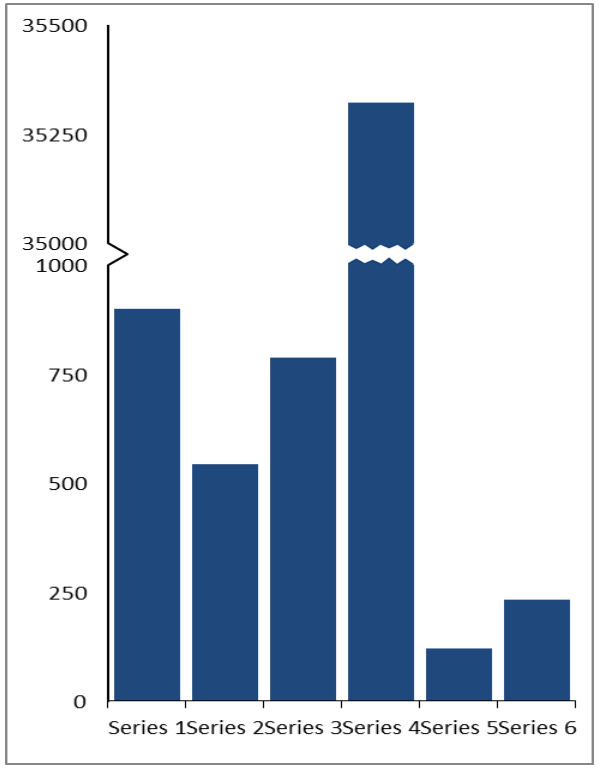
A	B	C	D	E	F
2	4	3	4	4	5
3	4	4	5	7	4
4	3	6	7	8	3
4	6	7	6	7	6
1	5	7	8	9	6
3	5	6	6	6	5
4	2	3	5	7	3
5	4	5	6	1	1
3	4	4	3	2	4

sqlbits

Them „outliers“

Label	Value
S1	885
S2	506
S3	763
S4	35363
S5	87
S6	221

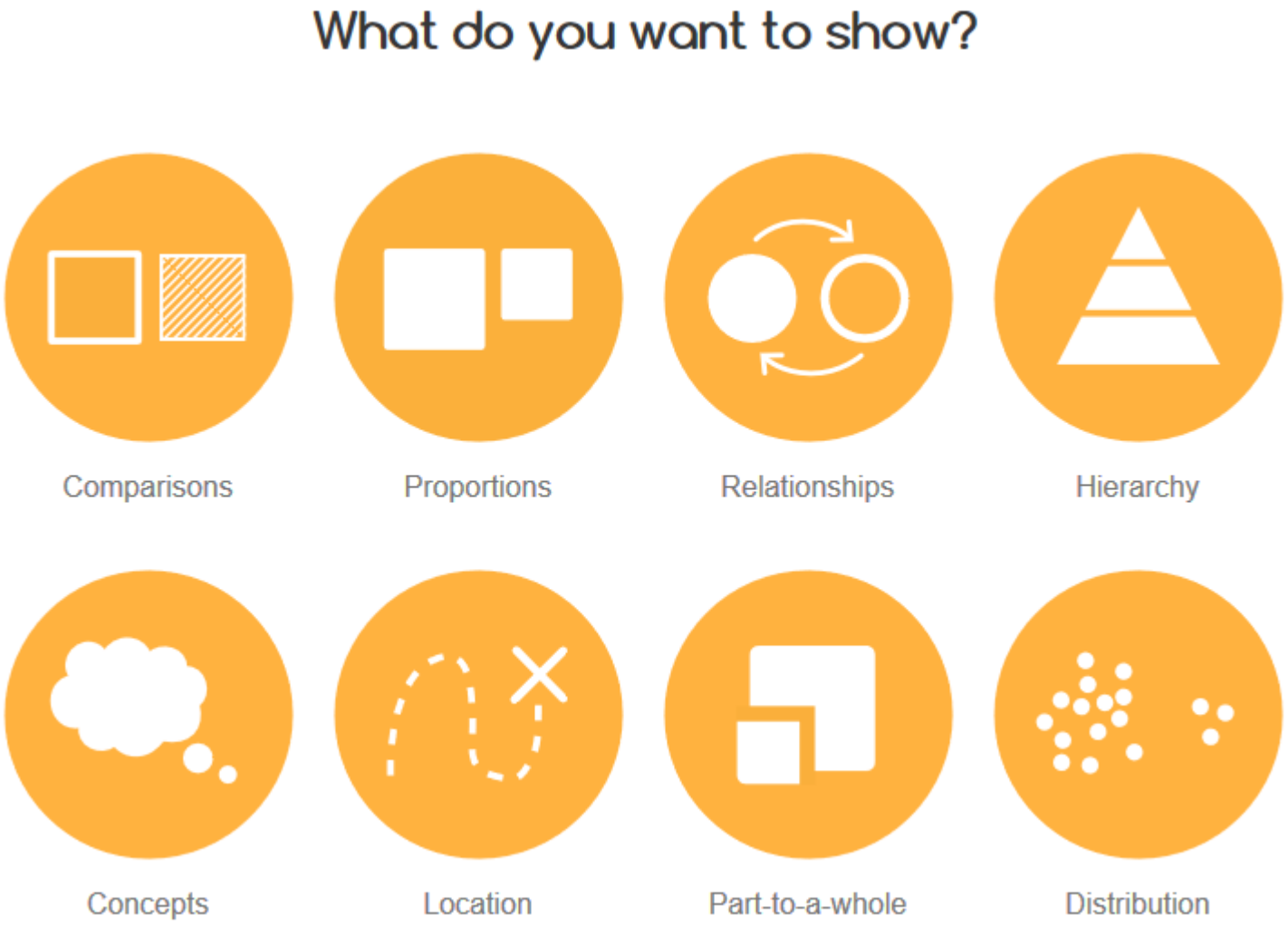




Additional Info: http://best-excel-tutorial.com/59-tips-and-tricks/387-break-column
https://alesandrab.wordpress.com/2014/03/17/broken-column-and-bar-charts/

sqlbits

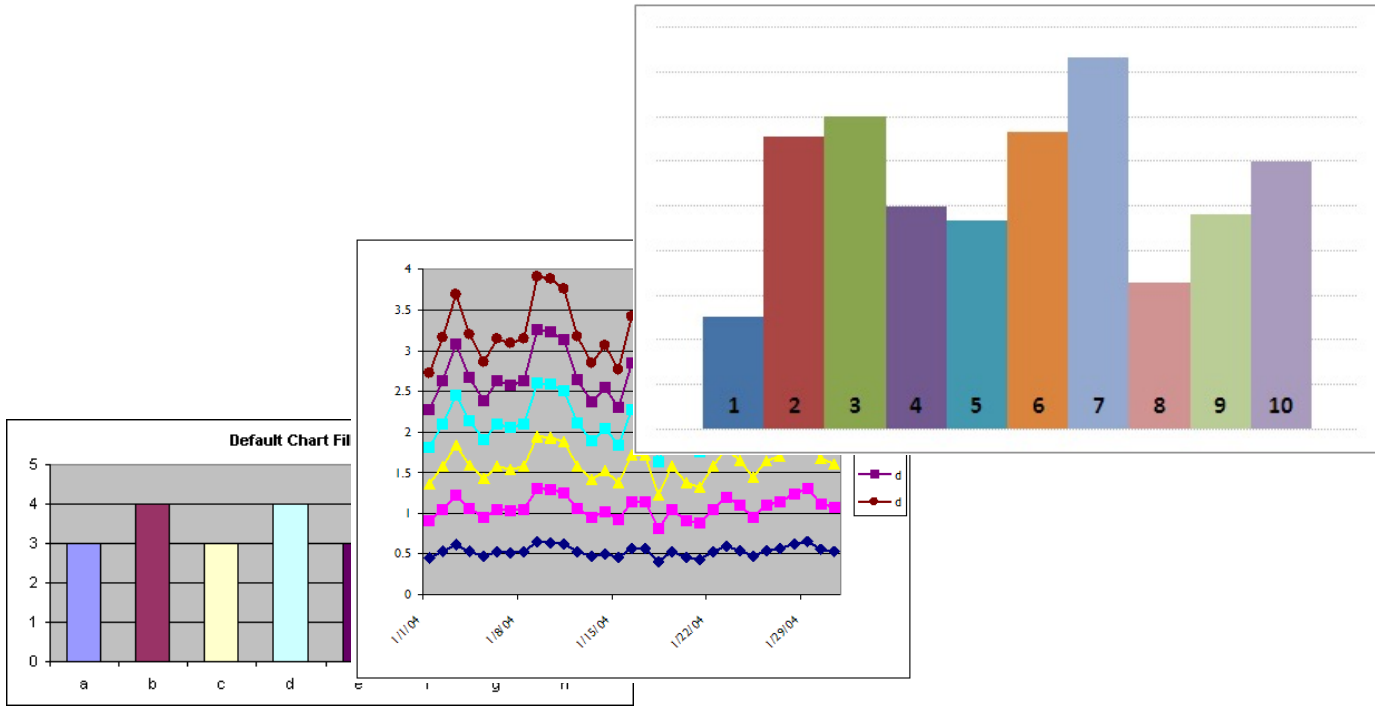
Understanding purpose



Source: https://datavizcatalogue.com/

Rules I follow with visualization

- Purpose of visualization is to make data more understandable and not vice versa
- If it feels stupid, it most probably is!
- If you don't understand it next day, do you think your customer will?
- Some results are better told with numbers
- KISS - Keep it simple stupid
- Fancy doesn't mean better
- Default colors are obvious
- Keep your schema
- Colors have meaning



sqlbits

Most common mistakes – wrap up

1. Ignoring data quality
2. Not exploring your data
3. Ignoring data distribution, ignoring feature engineering
4. Model over/under-fitting
5. Data leakage
6. Not addressing bias
7. Ignoring model evaluation metrics
8. Poor data visualization
9. Poor reporting and data presentations
10. Fail to understand business, lack in domain knowledge

Conclusions – Ask your self ToDo's

1. Re-think your business problems and re-align with your customer
2. There will always be more data availble, but at what cost?
3. Are you positive about deploying a particular model?
4. Have a shred of doubt ☺

Thank you very much!

http://tomaztsql.wordpress.com

tomaz.kastrun@gmail.com

@tomaz_tsql

/in/tomaztsql

http://github.com/tomaztk

https://mvp.microsoft.com/PublicProfile/5002196

sqlbits