Data Analytics

Lecture 2

Be AGILE

- Lecture 01 slides are updated

  - thanks Jason for providing a legit link to reference book!)

- Installed Alteryx/Power BI(Power Query)/SQL env?
- Markdown?
- Slack!

  (works great with Bots and push reports (Alteryx, Tableau 3rd party tool)

Progress

- Introduction, Methodology, Quiz to understand your SQL/ basic data handling proficiency
- What we are dealing with:
  - Evolving production data
    - Meta data Big data
  - Ad-hoc, external data
    - lifecycle management reference data
  - Tactics: (Agile, Github)
  - Classifications
    - Type of data (categorical, nominal, ordinal)
    - Data Universe quadrant Capability quadrant
    - Killer tools Future tools (D3, GraphQL)
- Methodology
  - CRoss-Industry Process for Data Mining (CRISP-DM)
    - Other alternatives
  - Case: Aircraft incident
    - Objective:what we want to do?
    - Data Understanding:
    - Data Cleansing:
    - Data Enhancing:
    - Modeling/Analysis
    - Evaluation
    - Deployment

Lecture 3

Lecture 4/5

- Data modeling and manipulation
  - Machine Learning Model Quadrant
    - generative model/discriminative model
  - Feature:
  - Dataset Split: Training/Test/Validation
  - Hyperparameters
    - 80:20 rule(spend more time in feature engineering)
    - How to find them - # of clusters
    - Learning rate - NN
    - Pruning- single decision tree - Level of depth - random forest
  - Measuring Model accuracy and effectiveness
    - Regression: R2, MSE
    - Classification - Measuring ROC,Gain, precision/recall
    - Confusion matrix,Type 1 /2 Error
    - Ensemble Learning

Lecture 4/5

Bus Matrix

Tool	Dataset		
Python	Flight Accident		Donation
SQL		BITDB	
Alteryx	Flight Accident		
Power Query/Excel/Power BI			

Ready to fly - flight incident data exploration

Did you make a backup?

Git remote -v

```
git remote add upstream
https://github.com/cnukaus/learner.git
Git checkout master
Git fetch upstream
Git merge upstream/master
```

df. **mean**()

inplace=True

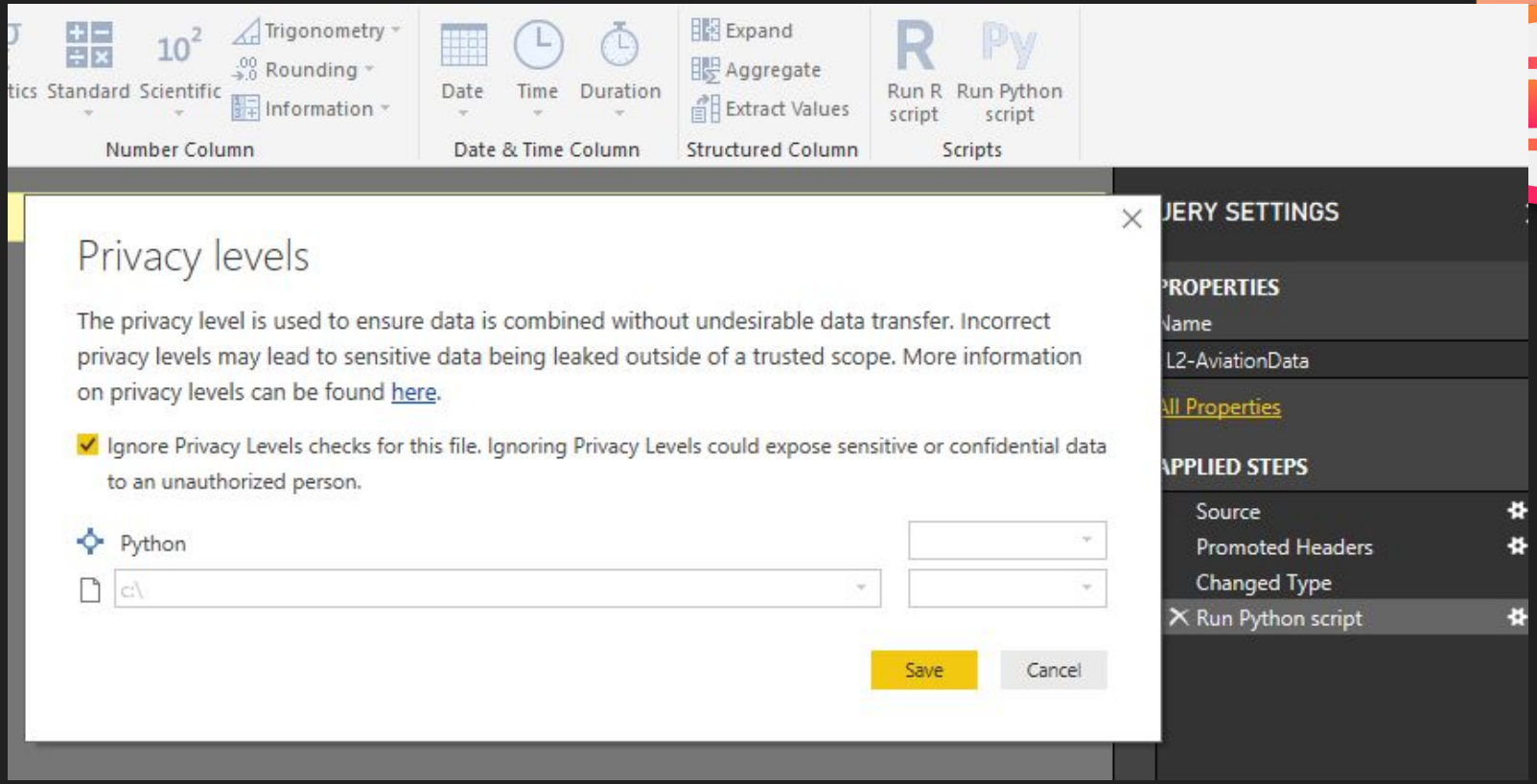
Duplicate?

S# = 227 and S# = 228 are the same incident, also 301 and 302, plus many others. However duplicates can have different values for the same fields.

S# 176 and 177 are duplicates - conflict killed #12, injured #8, whereas for S# = 177 the numbers are 13, 3 and 15.

https://www.kaggle.com/edhirif/data-analysis-with-data-quality-analysis

Power BI Python setting



Set all privacy to “Public”

Data Cleanse method

Regex (manual)
Library default: fillna
Model Prediction
Rule based model
Link to external data

```python
def fix_number_of_engines(noe, m):
    if noe >= 0:
        return noe
    else:
        # Setting number of engines at the mean number of engines for the producer
        r = np.round(df['Number.of.Engines'][df['Make']==m].mean())
        return r

# Setting 0 engines for balloons
df['Number.of.Engines'][df['Number.of.Engines'].isnull() & (df['Make'].str.contains('balloon', case=False))]
# Correcting number of engines
num_engines = df.apply(lambda x: fix_number_of_engines(x['Number.of.Engines'], x['Make']), axis=1)
df = df.assign(NumberofEngines = num_engines, index=df.index)
# Still some null after number of engines correction
df['NumberofEngines'].fillna(1, inplace=True)
```

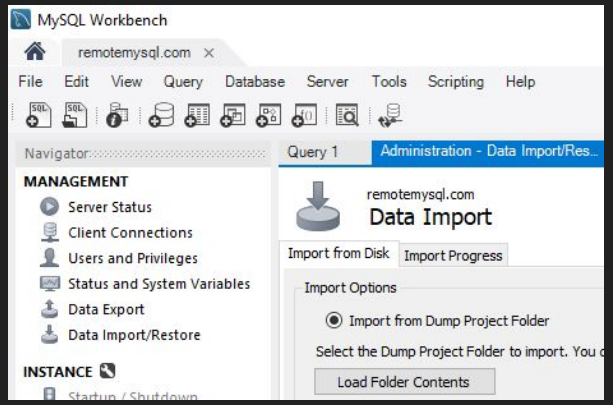
Exercise 2: Use Excel/Power Query if you have Excel, otherwise Power BI

(in case you run Linux/mac, there is a life saver - remote machine)

Requirement behind their statement

I need a better (horse, CSV file)

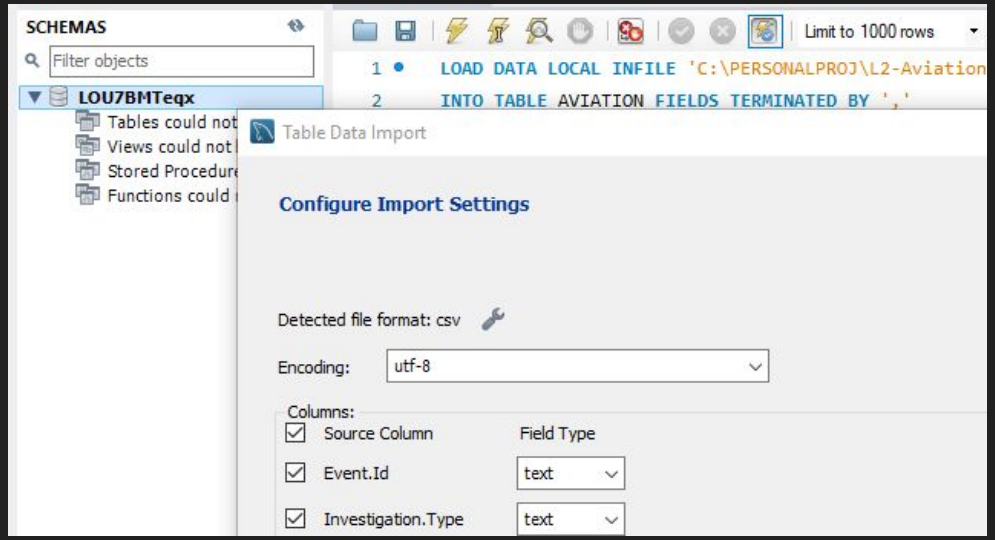
I want to reach more customers (so should stream to advertising/third party directly)



Exercise 3: Data prepare in SQL

Note the difference between DB backup file

And table content export file



Garbage in, Garbage out

There are as many valid coding convention as developers,

only important thing-
keep the same coding convention in a project.

Cédric Julien

Big data - about correlation, can be misleading

Stat: those have 10 min breaks between work, after 5 years 40% more likely to have cancer

https://www.kaggle.com/kanncaa1/why-gun-violance-increase-in-texas

论点-论据 - 还是从现象中找结论？

达尔文，假说
毛泽东-农村
林彪-应用大数据

用IV WOE来量化RFM Model

达尔文，假说
毛泽东-农村
林彪-应用大数据

Correlation not Causation

```
alter table Customers add Age int;
```

- Which month are most of our customers born?
- Highest value woolies EDR member?

When you f*** ed up

Lessons learned