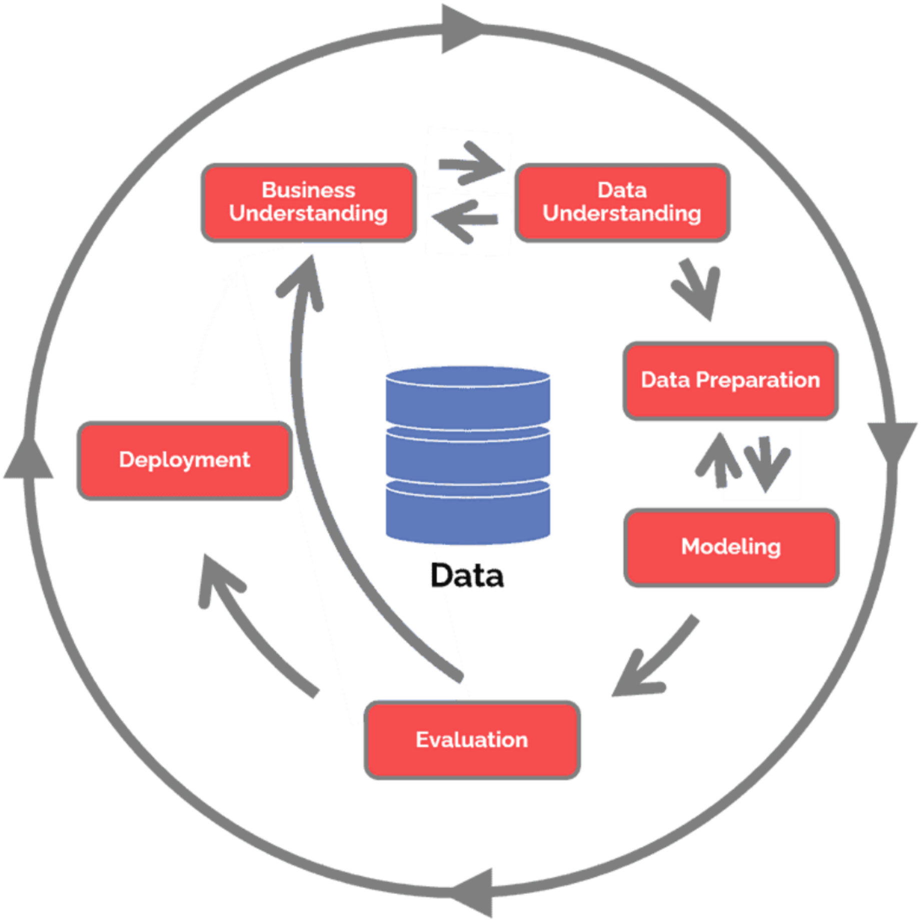
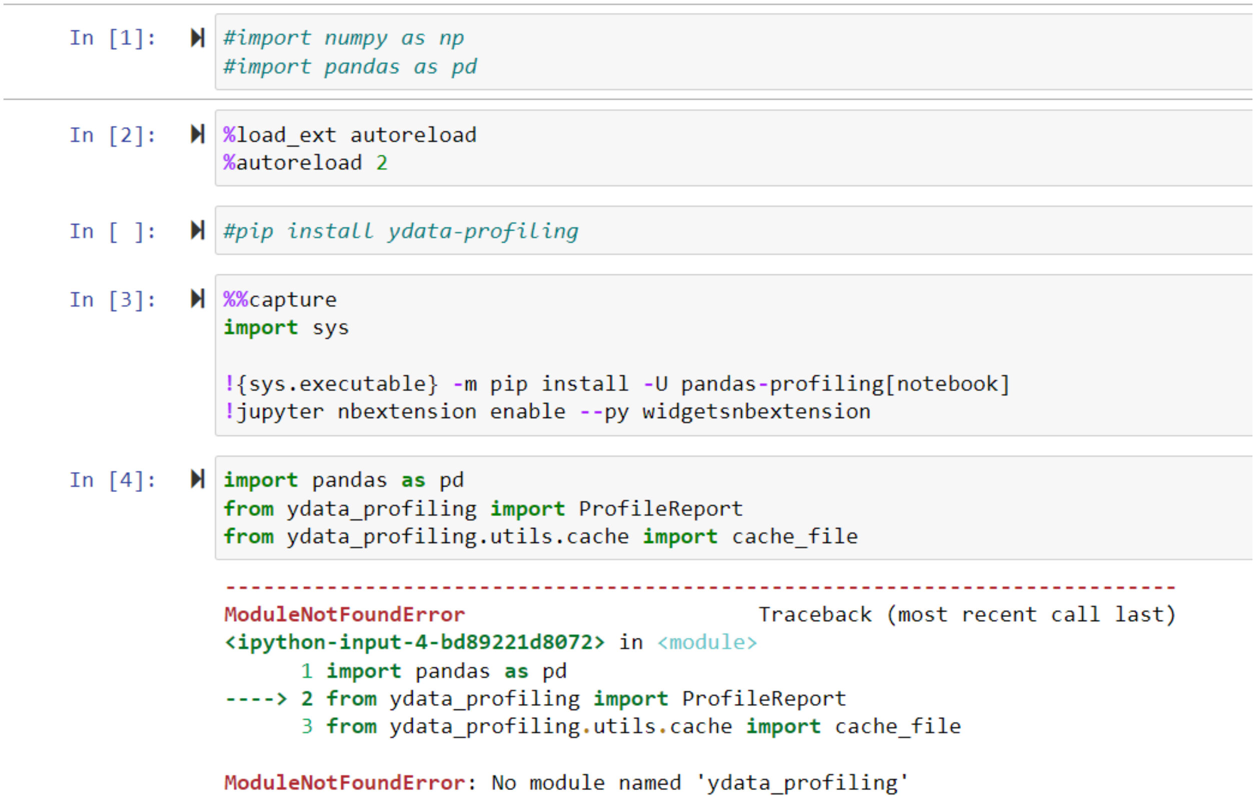
Methodology

- The **CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining (*CRISP-DM*) is a process model that serves as the base for a data science process.



CRISP-DM Diagram. Inspired by Wikipedia

Ydata-profiling issue on Jupyter notebook



```
In [1]:  #import numpy as np
         #import pandas as pd
```

```
In [2]:  %load_ext autoreload
         %autoreload 2
```

```
In [ ]:  #pip install ydata-profiling
```

```
In [3]:  %%capture
         import sys

         !{sys.executable} -m pip install -U pandas-profiling[notebook]
         !jupyter nbextension enable --py widgetsnbextension
```

```
In [4]:  import pandas as pd
         from ydata_profiling import ProfileReport
         from ydata_profiling.utils.cache import cache_file
```

```
---------------------------------------------------------------------------
ModuleNotFoundError                       Traceback (most recent call last)
<ipython-input-4-bd89221d8072> in <module>
      1 import pandas as pd
----> 2 from ydata_profiling import ProfileReport
      3 from ydata_profiling.utils.cache import cache_file

ModuleNotFoundError: No module named 'ydata_profiling'
```

Things to learn from paper

Predicting Hotel Bookings Cancellation With a Machine Learning Classification Model – 2017, 16th IEEE International Conference on Machine Learning and Applications

Datasplit and Construction

- Data leakage issue- leakage of future into training data.Removal of future dataset at any time from modelling dataset.
- Dataset shift issue- joint distribution of inputs and outputs differs between training and test stage. Stratified data sampling cannot guarantee similar distribution, due to change in demand patterns(ADR and Lead Time increases rapidly)

Feature Selection and Engineering

- Remove COUNTRY (PRT (=40%) assigned at booking and subject to change only at check in)
- All features associated with time will be removed. *arrival_date_year, arrival_date_month, arrival_date_week_number , arrival_date_day_of_month.*
- Remove low value features. *assigned_room_type, required_car_parking_spaces, reserved_room_type.*
- Replace *lead_time* and *adr* by engineered features livetime and ADRThirdQuartileDeviation. Value should vary according to type of booking, (Effective) bookings keep lead time, (Cancelled) bookings had number of days between booking creation and cancellation dates, (Future) bookings had elapsed number of days between booking creation and processing dates. ADR doesnt capture distribution and amplitude for cancellations. Businesswise it's known that bookings with expensive prices (compared at the same window in time) tend to cancel more.

QUESTIONS

1. PCA, SCALING,