

Machine Learning & Data Mining

Final Project

**Hotel Booking Cancellations**

Professor: Boaz Lerner

Students: Oriel Perets & Dafna Meron

Date:

**Abstract**

This project…

Contents

[**Business Understanding** 4](#_Toc91702028)

[**Data Understanding** 4](#_Toc91702029)

[Collect data 4](#_Toc91702030)

[Describe data 4](#_Toc91702031)

[Explore data 4](#_Toc91702032)

[Verify data quality 4](#_Toc91702033)

[**Data Preparation** 5](#_Toc91702034)

[Select data 5](#_Toc91702035)

[Clean data 5](#_Toc91702036)

[Construct data 5](#_Toc91702037)

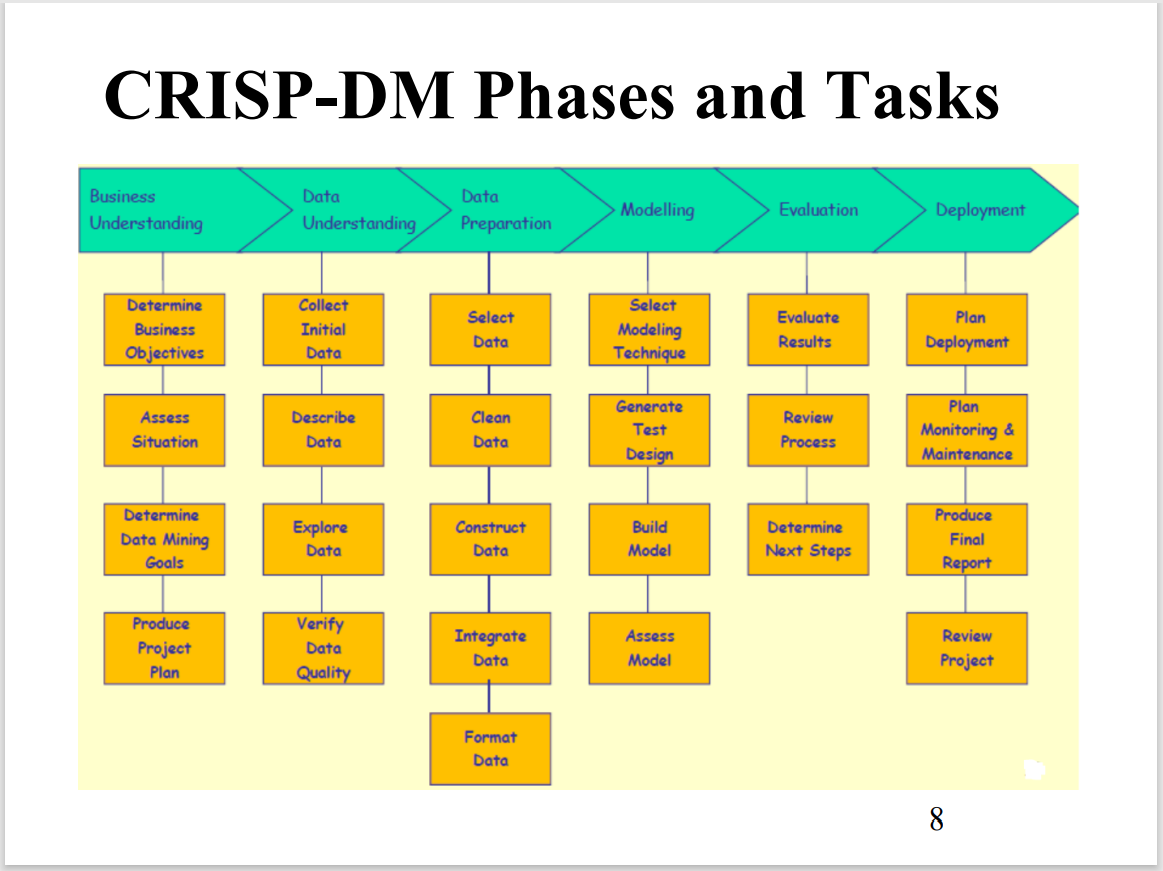
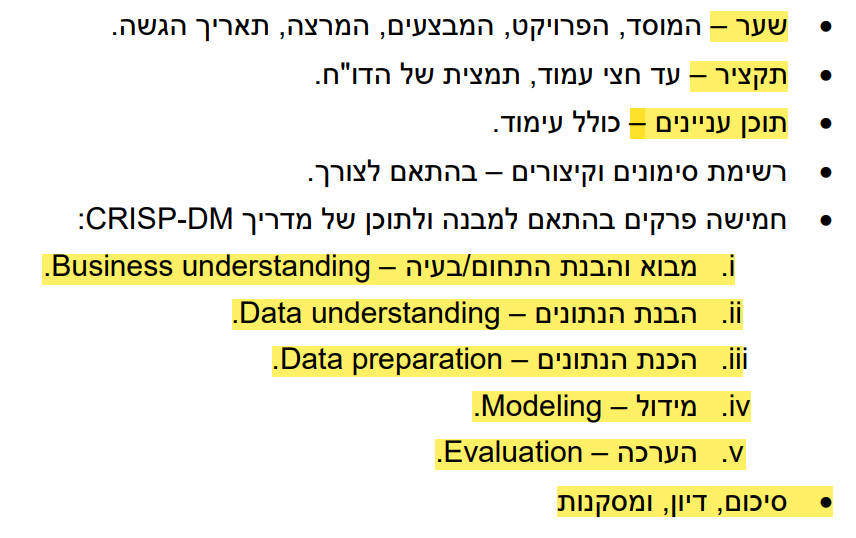
[Integrate data 5](#_Toc91702038)

[Format data 5](#_Toc91702039)

[**Modeling** 5](#_Toc91702040)

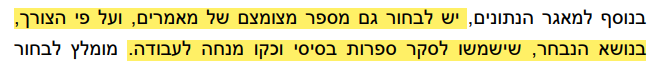
[**Evaluation** 5](#_Toc91702041)

[**Discussion and Conclusions** 5](#_Toc91702042)



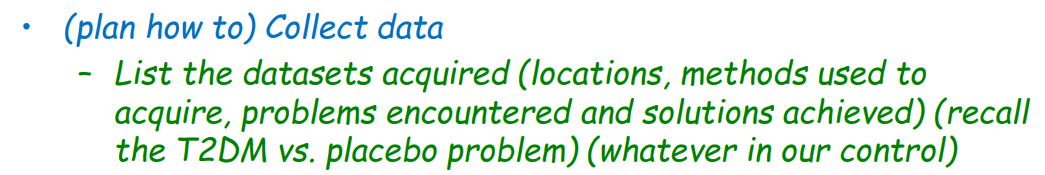
# **Business Understanding**

unexpected cancellation of hotel bookingsis a disturbing issue for online reservation websites such as Booking.com, Trivago, and Hotels.com, since it results in unwanted vacancy and lost profits. Anticipating whether an order will be canceled, even hours in advance, can help these companies act, for example: double-book certain rooms, create incentives for arriving customers with a high cancellation chance, send reminders and confirmation emails, alter cancelation terms, fees, and deposit types.   
In this project, we will explore these problems and solutions with classification models.



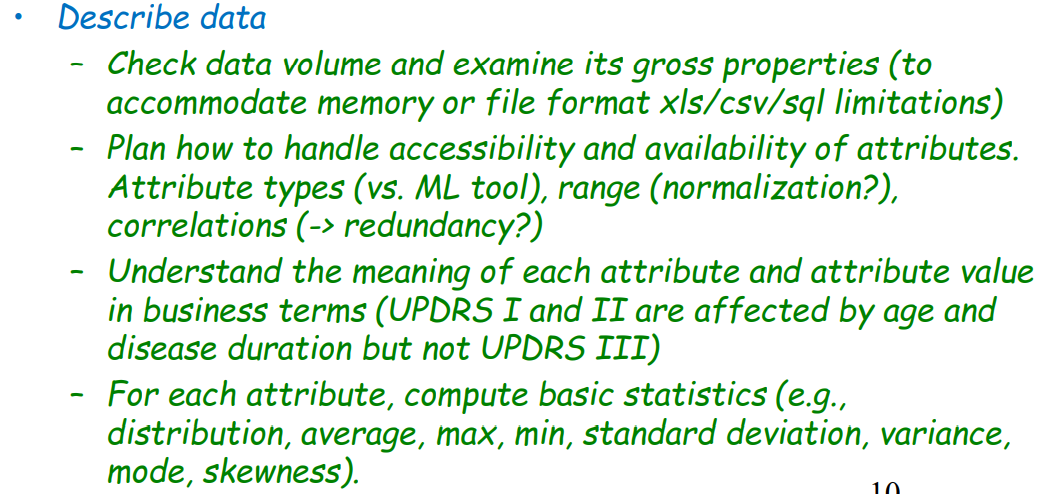
# **Data Understanding**

## Collect data



“Hotel Booking Dataset”, link: <https://www.kaggle.com/jessemostipak/hotel-booking-demand>.

## Describe data

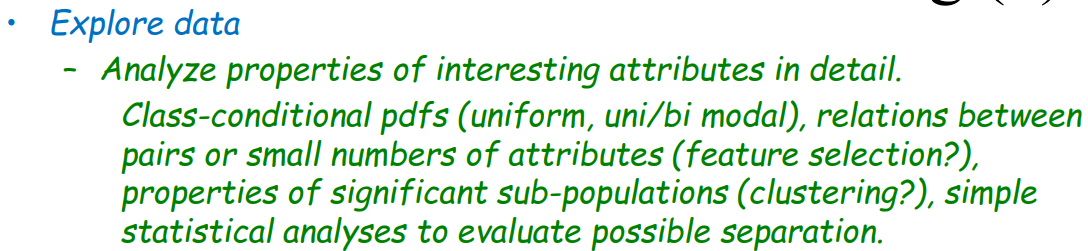


* The dataset contains 119390 samples and 32 attributes that describe them. Add screen shot.
* Atrributes: lead\_time, arrival\_date\_year, arrival\_date\_month, arrival\_date\_week\_number, arrival\_date\_day\_of\_month, stays\_in\_weekend\_nights, stays\_in\_week\_nights, adults, children, babies, meal, country, market\_segment, distribution\_channel, is\_repeated\_guest, previous\_cancellations, previous\_bookings\_not\_canceled, reserved\_room\_type, assigned\_room\_type, booking\_changes, deposit\_type, agent, company, days\_in\_waiting\_list, customer\_type, adr, required\_car\_parking\_spaces, total\_of\_special\_requests, reservation\_status, reservation\_status\_date.

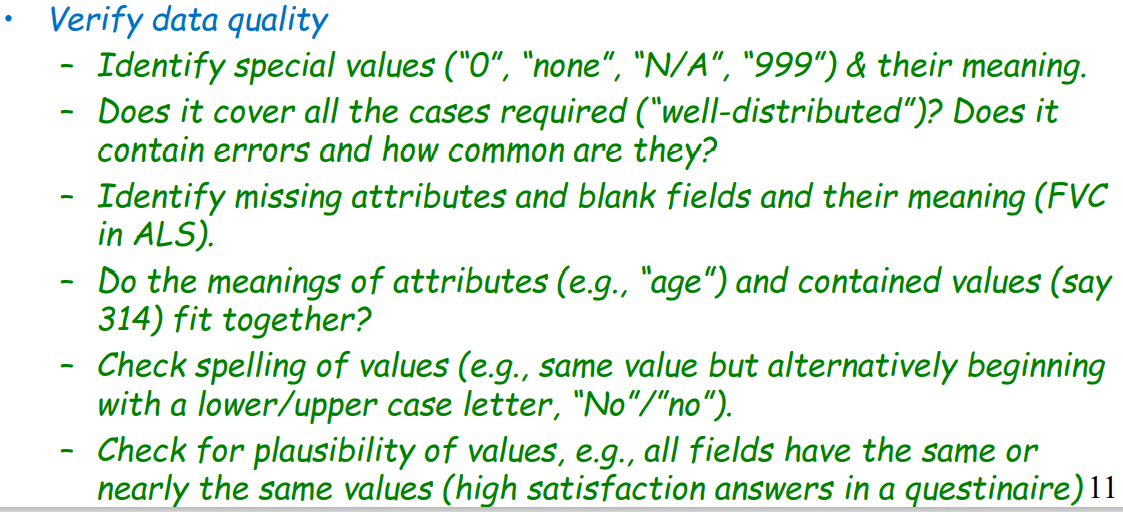
We plan to…, attributes types: Numeric and categorical

* Redundancy: Our learning task is classification of cancelled and not cancelled reservations. There are 3 attributes that identify the result of the reservation: is canceled, reservation\_status, reservation\_status\_date. We chose to train and evaluate on **is\_canceled** as a target variable and discard the other two.
* Understanding business terms: adr?
* Basic statistics of each attribute: Max, min, mean, median, std, distribution. add screen shot.

## Explore data



## Verify data quality

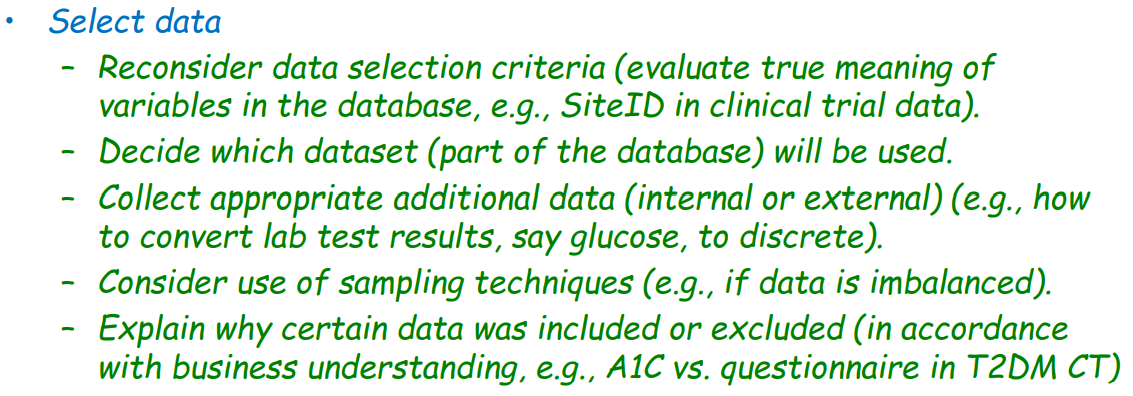


Missing values

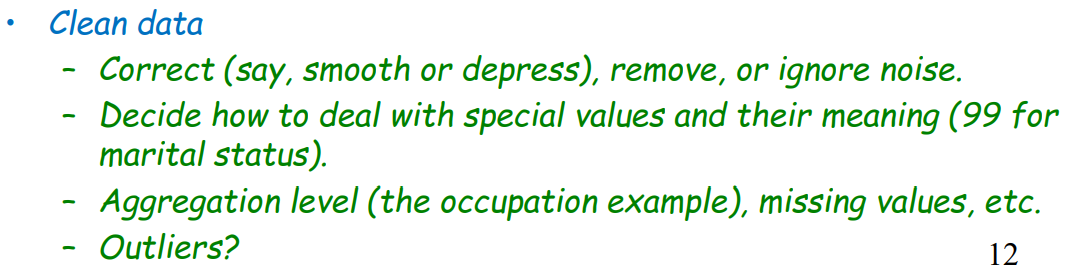
Undefined category

# **Data Preparation**

## Select data



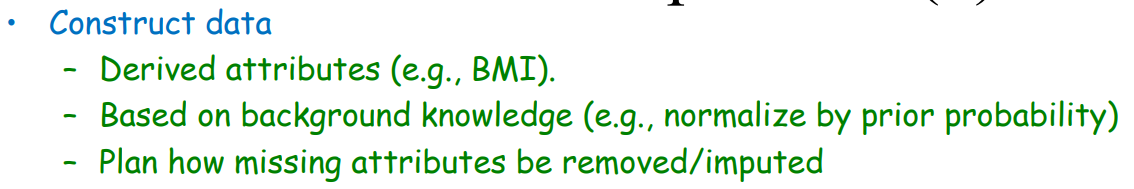
## Clean data



How did we handel missing data?

How did we handle ‘undefined’ categories?

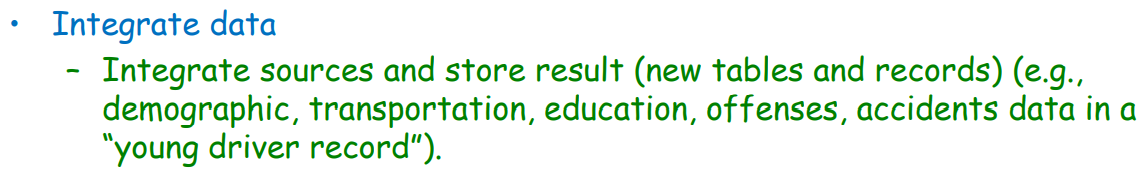
## Construct data



All added attributes

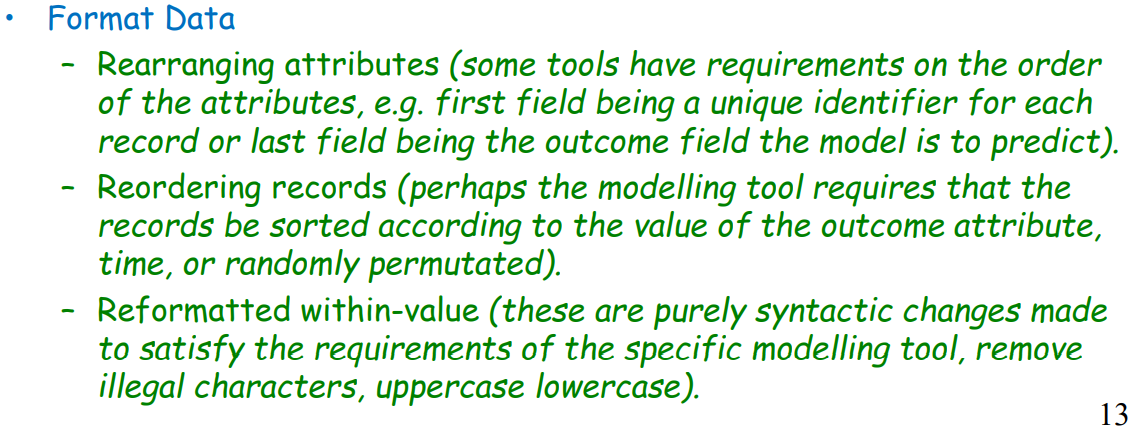
No missing attributes

## Integrate data



Not relevant

## Format data



?

Binnig?

# **Modeling**

RF

Bayes

Logistic regression

More?

# **Evaluation**

Import tree

# **Discussion and Conclusions**

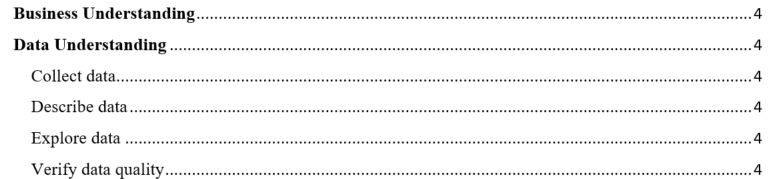
**Code:**

* Scatter plot, don’t use percentage cancellation but cancellation count – Dafna – low priority
* Add confusion matrix to the best kfold split - Dafna
* Import tree – Dafna
* Other methods: check if the data preparation is similar to logistic regression - Oriel
* Other methods: XGboost? DT? – Oriel
* Evaluation – search specific on RF – Dafna

**Document**

**Oriel**

* Presentation template



**Dafna**



**Together – discussion and conclusions**