Machine Learning Group 12

Haddar Nesrine, Maryna Gutruf

1 SELECTED DATASETS

We selected following datasets for upcoming assignments in Machine Learning course:

- The Coronavirus Dataset [1]
- The Melbourne Dataset [2]

2 SELECTED TOOL AND MODULS

We used Phyton and it's following Moduls:

- sys
- numpy
- pandas
- matplotlib
- mpl toolkits
- seaborn
- os

3 THE CORONAVIRUS DATASET

3.1 Why this dataset?

The coronavirus dataset has daily level information on the number of affected cases, deaths and recovery from 2019 novel coronavirus. It is a quite small, well described collection of real life data and invites to learn a lot of different and complex data analytics. We plan to use it for the regression exercise. Last but not least coronavirus is so cutting-edge!

3.2 Characteristics of data set

The Coronavirus dataset has 8 attributes and 7.617 examples and is low dimensional. This dataset has no missing values. In this dataset we have following types of attributes:

- Nominal attributes: Province/State, Country/Region
- Ordinal attributes: ObservationDate, Last Update
- Ratio attributes: Confirmed, Deaths, Recovered, SNo (ID)

Knowing the level of measurement of the variables is important for two reasons:

- Each of the levels of measurement provides a different level of detail. Nominal provides the least amount of detail, ordinal provides the next highest amount of detail, and interval and ratio provide the most amount of detail.
- Different statistical tests are appropriate for variables with different levels of measurement. For example, chi-square tests of independence are most appropriate for nominal level data. [3]

3.3 Target attributes

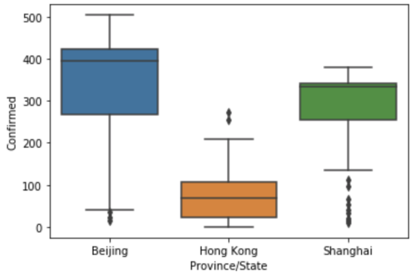
- Our target attributes can be: Confirmed, Deaths, Recovered, where a regression can be done to forecast the future number of infected, deceased or recovered people.
- TBD: Numeric values - Description on value ranges
- Rely on libraries, modules to load data, plot, visualise, etc.

3.4 Data Visualization

We looked at two types of plots:

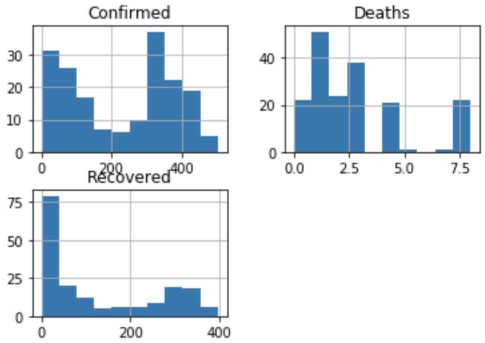
- Univariate plots in order to better understand each attribute.
- Multivariate plots in order to better understand the relationships between attributes.

We ploted the distribution in confirmed, deaths and recovered cases for 'Hong Kong', 'Beijing', and 'Shanghai'. Here is the one about confirmed cases:



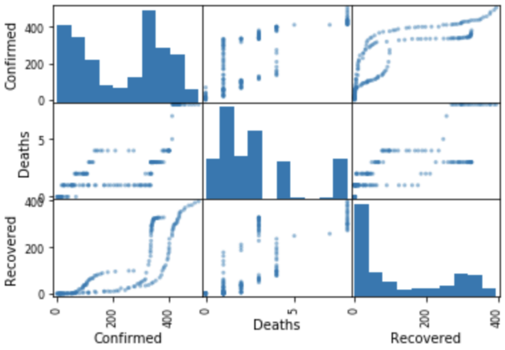
As you can see in the diagram above, the median line of box 'Hong Kong' lies outside of boxes 'Beijing' and 'Shanghai' entirely. So there is likely to be a difference between the 3 data groups. In terms of size of boxes that's something to look for when comparing box plots since we have both short boxes (which mean their data points consistently hover around the center values) and taller boxes (which imply more variable data).

We created following histograms and it looks like perhaps two of the input variables have a Gaussian distribution. This is useful to note as we can use algorithms that can explore this assumption.



We also took a look at scatterplots of all pairs of attributes, since this can be helpful to spot structured relationships between input variables. We noted the diagonal grouping of some pairs of attributes. This suggests a high correlation and a predictable relationship:

[1] https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset
[2] https://www.kaggle.com/anthonypino/melbourne-housing-market
[3] https://www.statisticssolutions.com/data-levels-and-measurement/



4 THE MELBOURNE PRICES DATA SET

4.1 Why this data set?

The Melbourne Prices Data Set is a large, well described collection of estate data in Melbourne with a 21 attributes and more than 34000 observations. It could be used in a regression or classification exercise, where price or property type could respectively be predicted. In our case, this dataset is going to be used for the classification part of the project, where the target is the property type.

4.2 Data set description

The dataset contains 21 attributes and 34857 observations, where some of which contain NA values. We have following types of attributes:

- Nominal attributes: Suburb, Address, Method, CouncilArea, Regionname
- Ordinal attributes: Date
- Quantitative attributes: all the rest

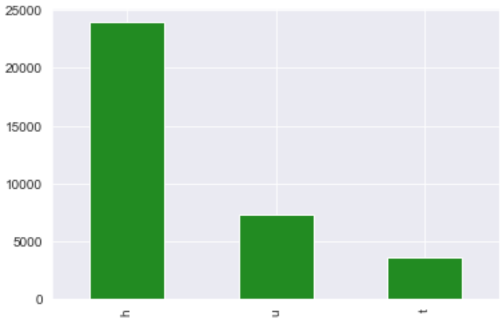
4.3 Target attribute characteristics

The target attribute represents the property type of the real estate. In our dataset it is represented by three different classes as following:

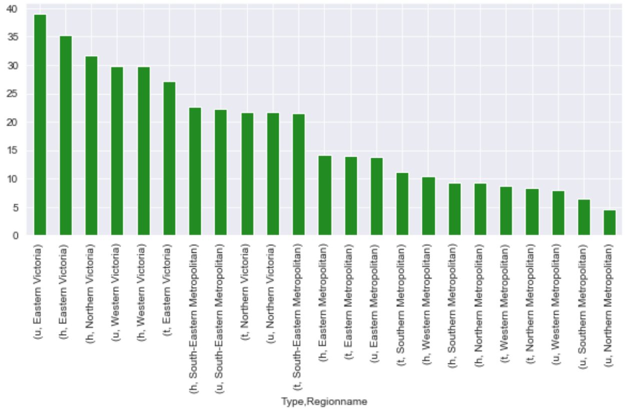
- h - house,cottage,villa, semi,terrace;
- u - unit, duplex;
- t - townhouse

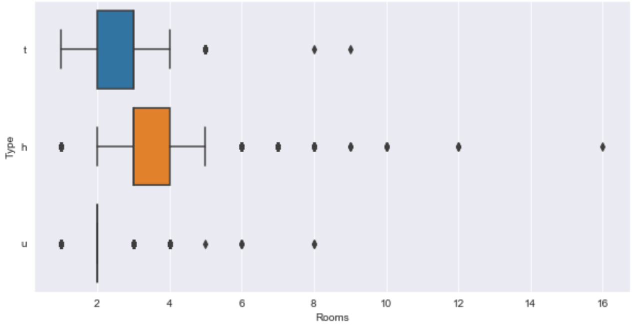
4.4 Data set Exploration

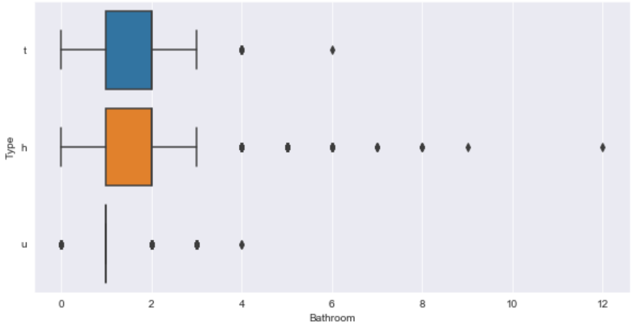
The distribution of the property type is as follows:

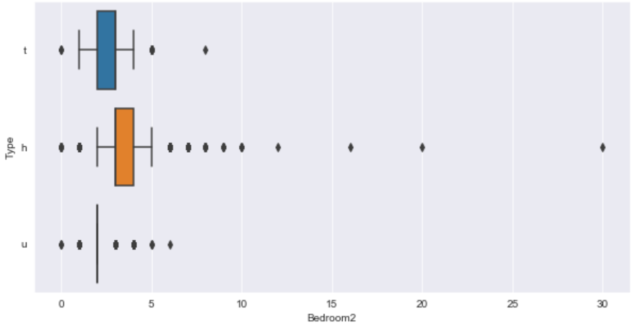


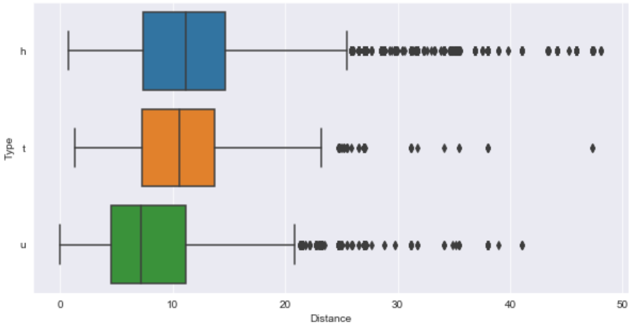
Other attributes and their relation to target attribute :



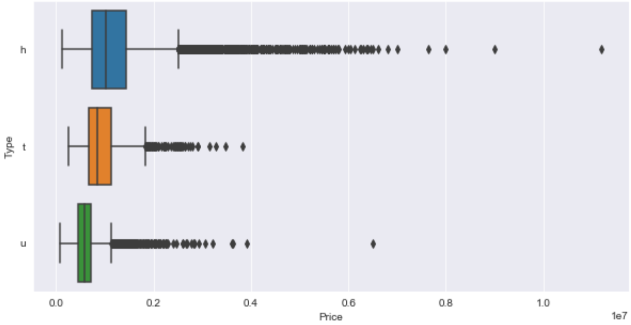




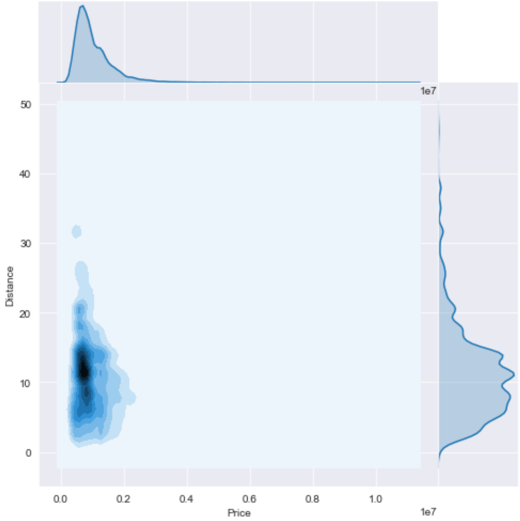




The distribution of the above attributes for each of the property types seems to be skewed, and many outliers can be seen in the boxplots

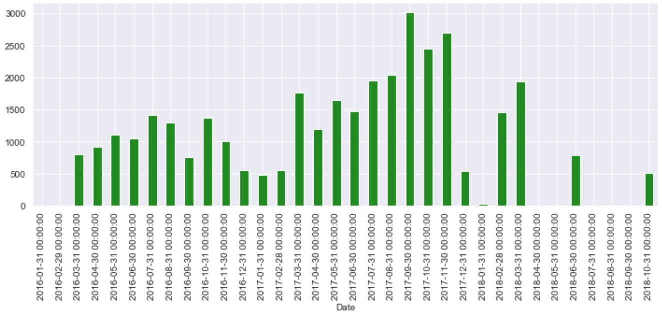


Trying to find a correlation between the price and the distance of the center of Melbourne:



It appears to be many outliers in the prices, that may be needed to be handled later.

We have almost 3 years of data, which is probably good to spot some trends. It can be seen that the number of deals always goes down around Christmas time (seems reasonable), and 2017 saw an increase in the number of deals.


2016-01-31 00:00:00
2016-02-29 00:00:00
2016-03-31 00:00:00
2016-04-30 00:00:00
2016-05-31 00:00:00
2016-06-30 00:00:00
2016-07-31 00:00:00
2016-08-31 00:00:00
2016-09-30 00:00:00
2016-10-31 00:00:00
2016-11-30 00:00:00
2016-12-31 00:00:00
2017-01-31 00:00:00
2017-02-28 00:00:00
2017-03-31 00:00:00
2017-04-30 00:00:00
2017-05-31 00:00:00
2017-06-30 00:00:00
2017-07-31 00:00:00
2017-08-31 00:00:00
2017-09-30 00:00:00
2017-10-31 00:00:00
2017-11-30 00:00:00
2017-12-31 00:00:00
2018-01-31 00:00:00
2018-02-28 00:00:00
2018-03-31 00:00:00
2018-04-30 00:00:00
2018-05-31 00:00:00
2018-06-30 00:00:00
2018-07-31 00:00:00
2018-08-31 00:00:00
2018-09-30 00:00:00
2018-10-31 00:00:00
Date