Group name: crispy chicken sandwich
Group members:

Loh Wan Teng, michellelohwt0799@gmail.com, Malaysia, Universiti Sains Malaysia

Zhechen Zhu, zhechenz@seas.upenn.edu, China, University of Pennsylvania

Zhihui(Angela) Chen, zhihuichen085@gmail.com, China, Brandeis University, Data Science

Problem description: The ABC bank aims to launch a new product, before they do that, they want to develop a model to help them understand what kind of customers would buy the product. In other words, based on the model built on different features of customers, they want to figure out the features that make the most difference to the outcome.

Business understanding: Based on the machine learning model, we hope to work out with the most efficient marketing strategy. The machine learning model would tell which feature matters most, meanwhile visualization results could also tell the clusters in each feature. For example, if it turns out that the job matters most, and people in the type of management are most likely to purchase for the product, then the main target of the marketing would be the people in management category with specific frequency.

Project lifecycle with ddl:

Deadline	Project Lifecycle
19 August 2022 (Week 7)	• Problem description • Business understanding • Project lifecycle with deadline • Data Intake Report

26 August 2022
(Week 8)
	• Problem description
	• Data understanding
	• Data analysis
	    ○ NA values, outliers, skewed data analysis
	    ○ Data processing and description

2 September 2022 (Week 9)	Data Cleansing and Transformation • Data cleaning with 2 techniques • Team code review
9 September 2022 (Week 10)	• Problem description • EDA • Final Recommendation • EDA submission
16 September 2022 (Week 11)	• EDA Presentation • Modeling Technique Proposal
23 September 2022 (Week 12)	• Model Selection • Model Building
30 September 2022 (Week 13)	• Final Project Submission • Final Project Presentation

Github link: https://github.com/AZHChen/ds-marketing-ml-project.git

Data intake report

The dataset used for analysis is bank-full, accessed from UCI database.

Tabular data details: bank-additional-full

Total number of observations	41188 rows
Total number of files	1 file
Total number of features	21 columns
Base format of the file	.csv
Size of data	6.6+ MB

- **Data understanding**
  45211 data are included in this dataset, covering 2 years from May, 2008 to Oct, 2010.
  Variables: there are 20 input variables (possible features in this model), 1 output variable, which is Y and the otimate prediction in this case.

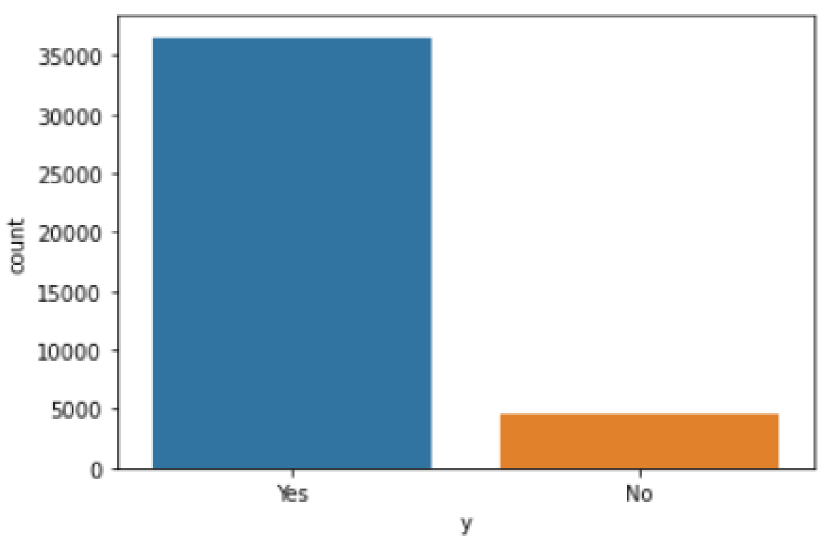
1. Age
2. Job: type of job ('admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown'
3. Marital: (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
4. Education: (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
5. Default: has credit in default? (categorical: 'no','yes','unknown')
6. Housing: has housing loan? (categorical: 'no','yes','unknown')

7. Loan: has personal loan? (categorical: 'no','yes','unknown')
8. Contact: contact communication type (categorical: 'cellular','telephone')
9. Month: last contact month of year
10. Day_of_week: last contact day of the week
11. Duration: last contact duration, in seconds (numeric).
12. Campaign: number of contacts performed during this campaign and for this client
13. Pdays: number of days that passed by after the client was last contacted from a previous campaign
14. Previous: number of contacts performed before this campaign and for this client (numeric)
15. Poutcome: outcome of the previous marketing campaign
16. Emp.var.rate: employment variation rate - quarterly indicator
17. Cons.price.idx: consumer price index - monthly indicator
18. Cons.conf.idx: consumer confidence index - monthly indicator
19. euribor3m: euribor 3 month rate - daily indicator
20. Nr.employed: number of employees - quarterly indicator

- **Data visualization (Initial dataset without cleaning and replacement)**

- Outcome

  The chart shows the y in this dataset is imbalance, in further machine learning process we need to make up for the unbalanced part or delete some data with the outcome of 'Yes' randomly.

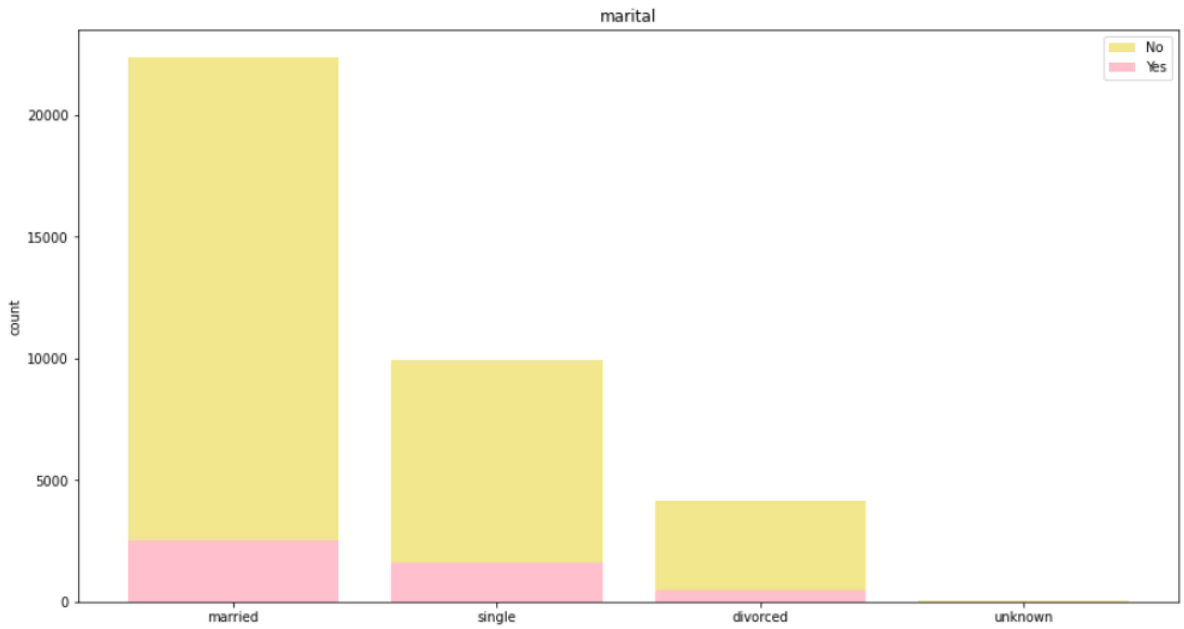


- Job

  Majority of the clients are from admin. Category, meanwhile the rejection rate of admin. people are also higher than the other categories.

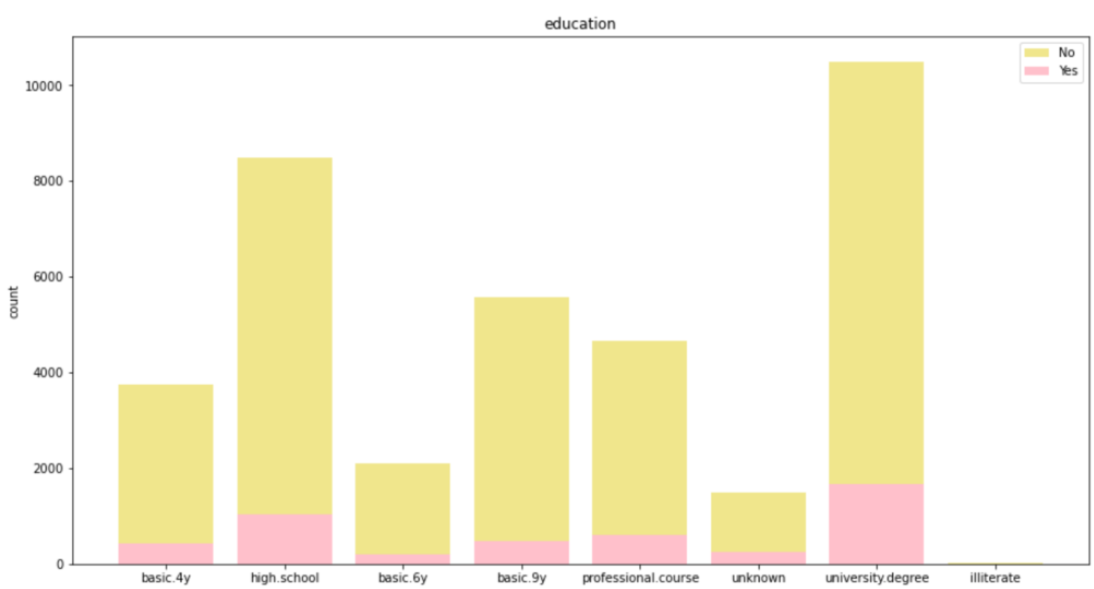
- Marital

  Most clients are married.

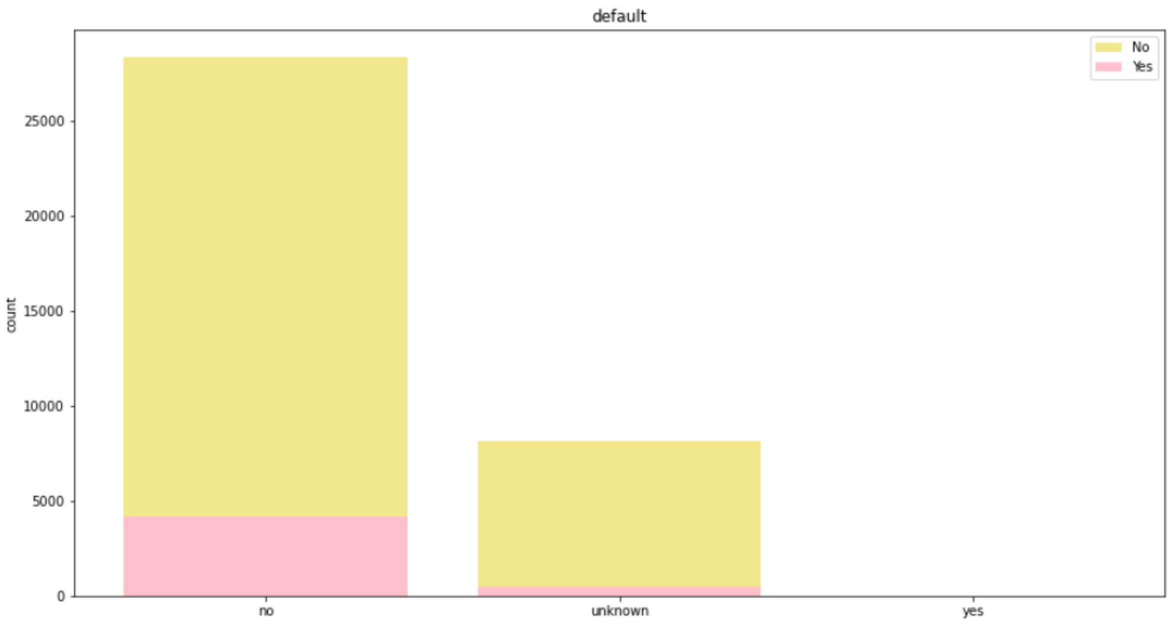


- Education

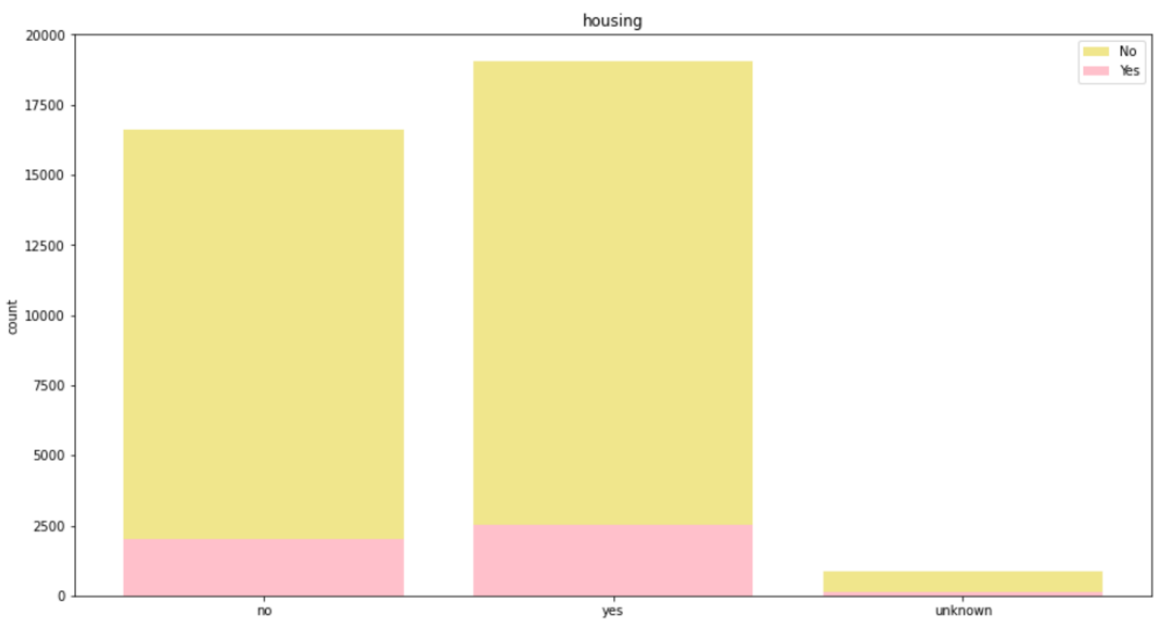
  The rejection rate of people with university degrees is higher than the others.



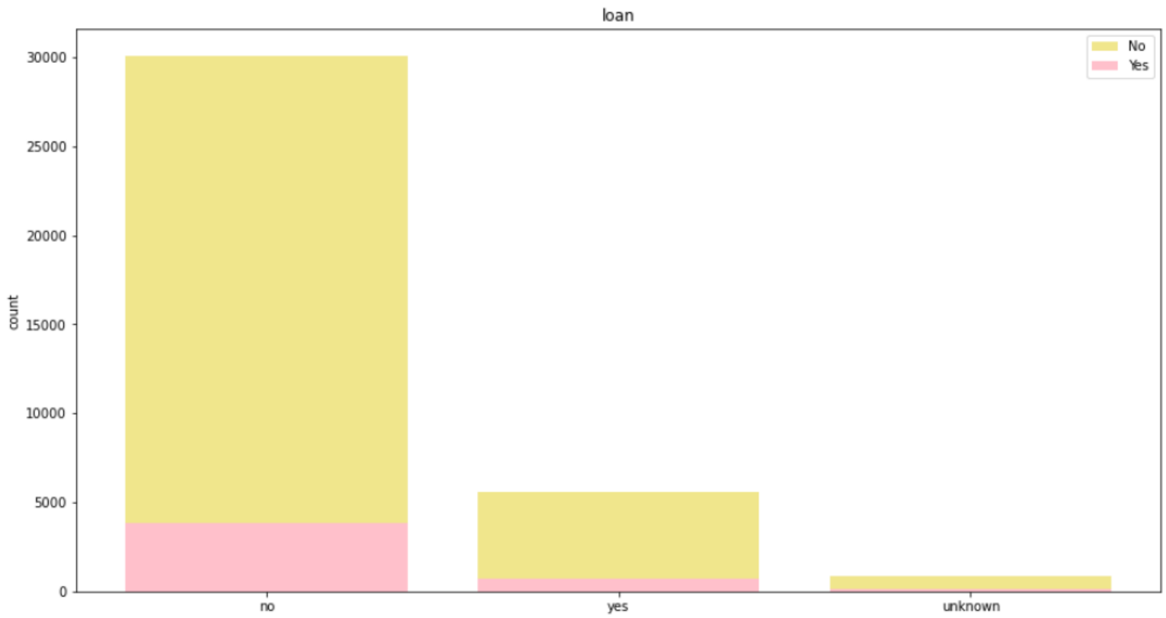
- Default



- Housing

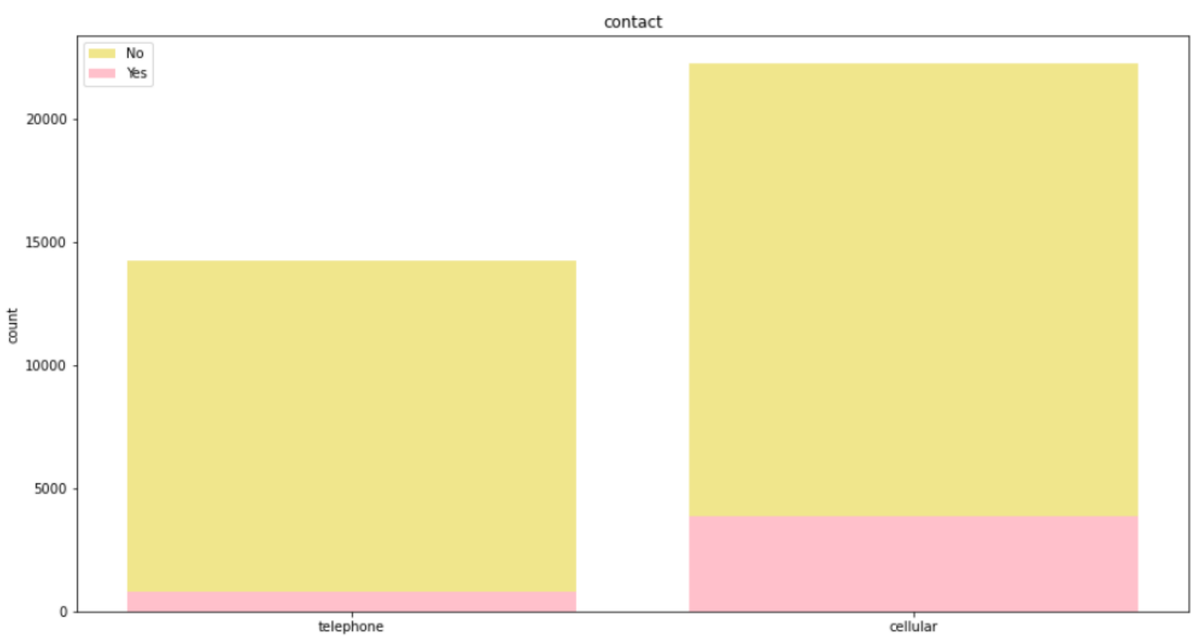


- Loan

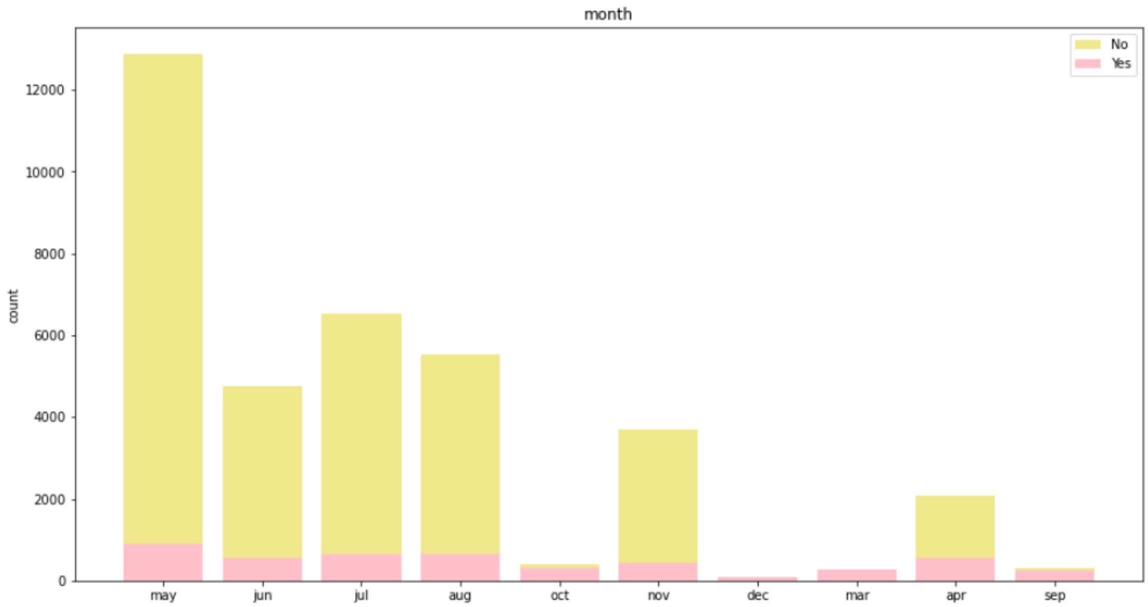


- Contact

  Most clients prefer cellular phones as a contact method.

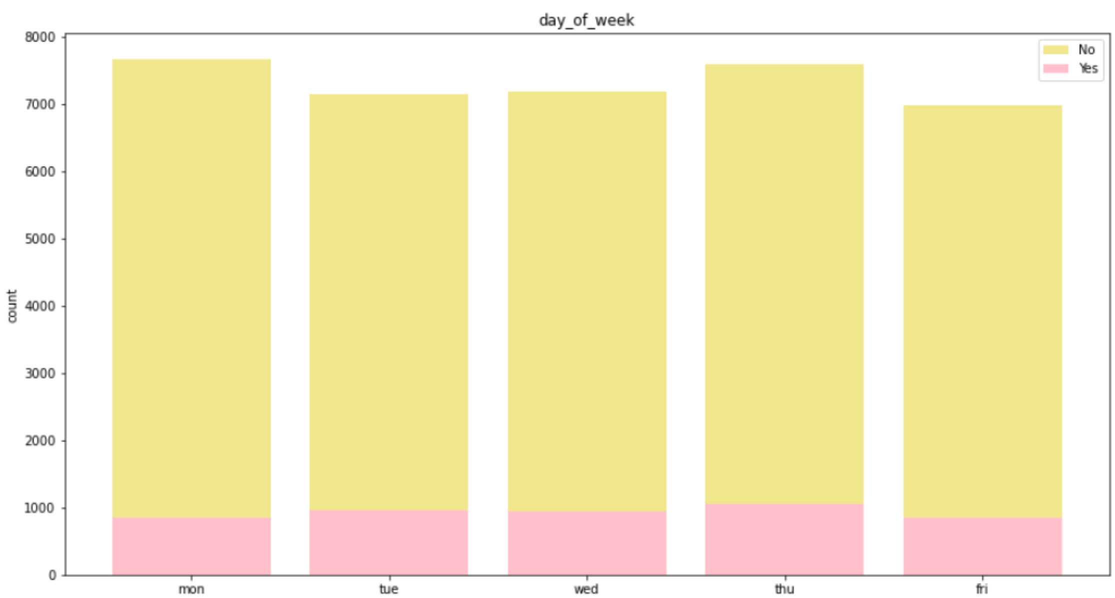


- Month

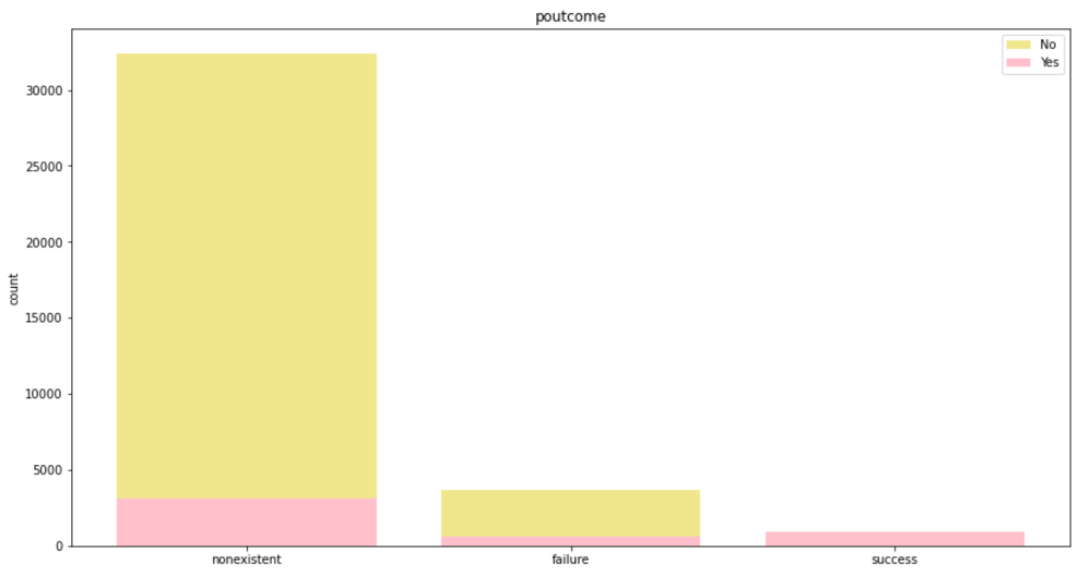


- Day of week

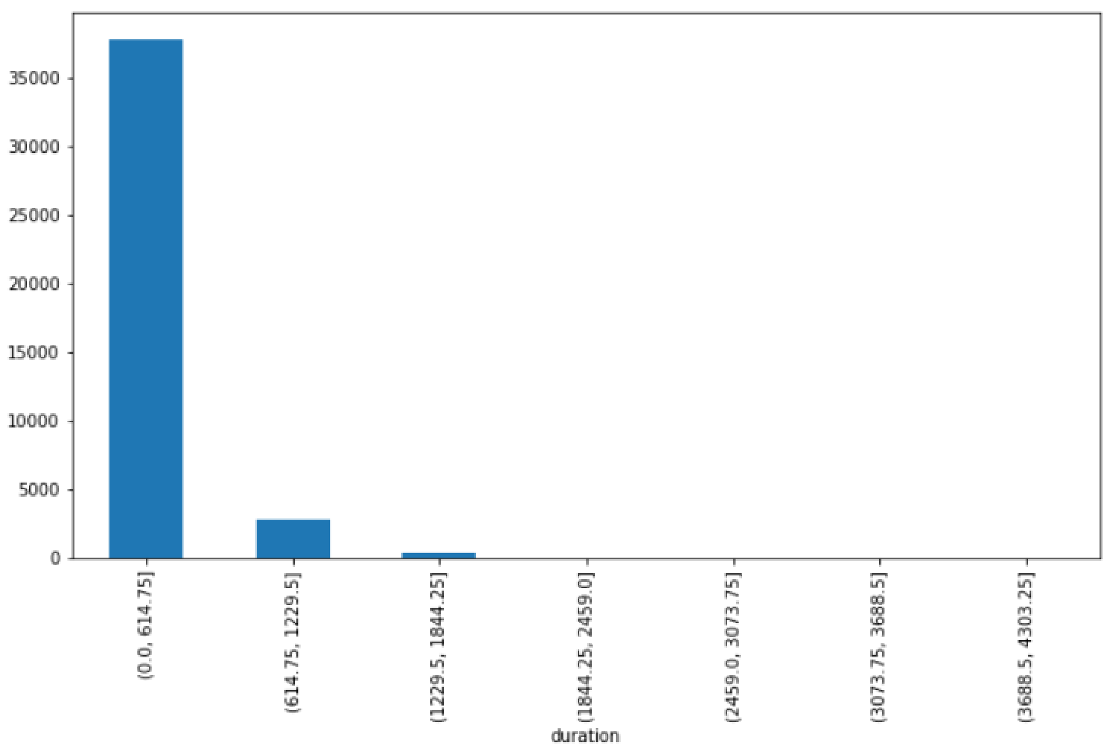
  This indicator has a uniform distribution, contact with clients in either of the days does not affect much on the rejection rate.



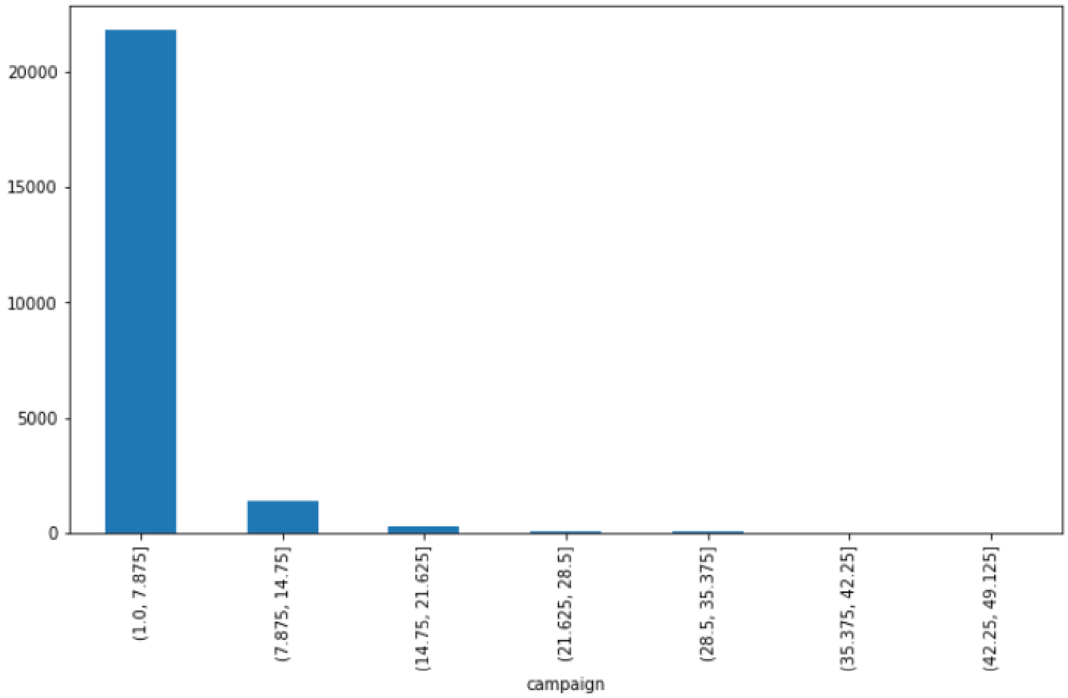
- poutome



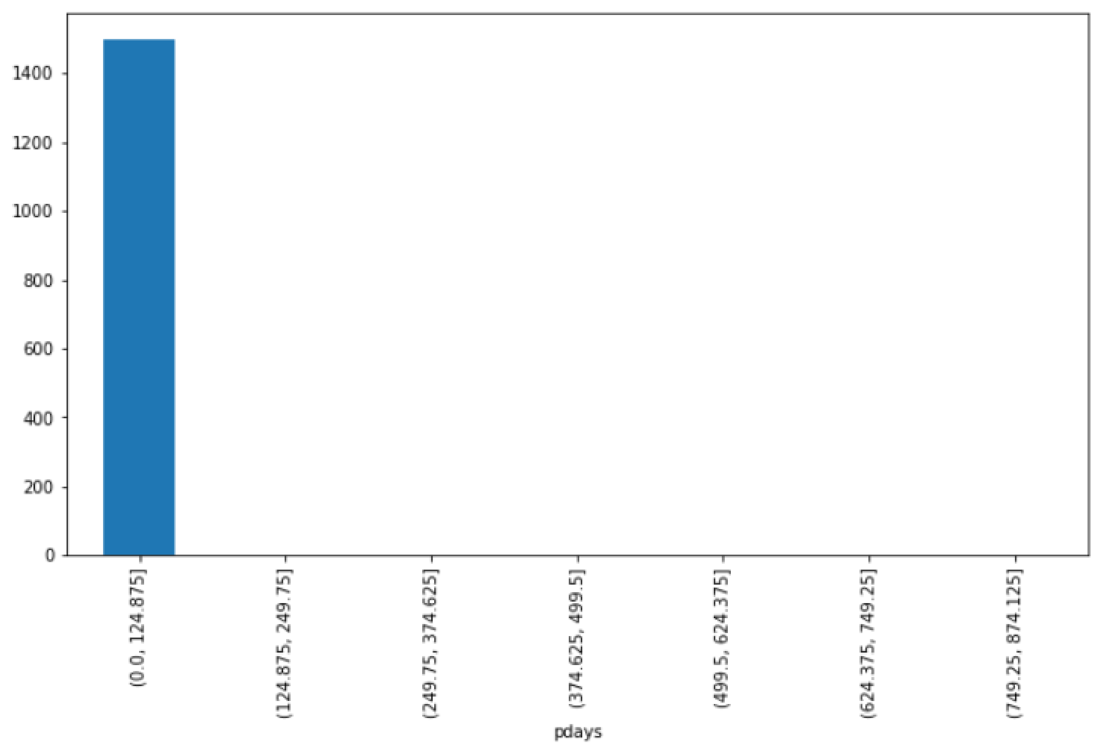
- Duration



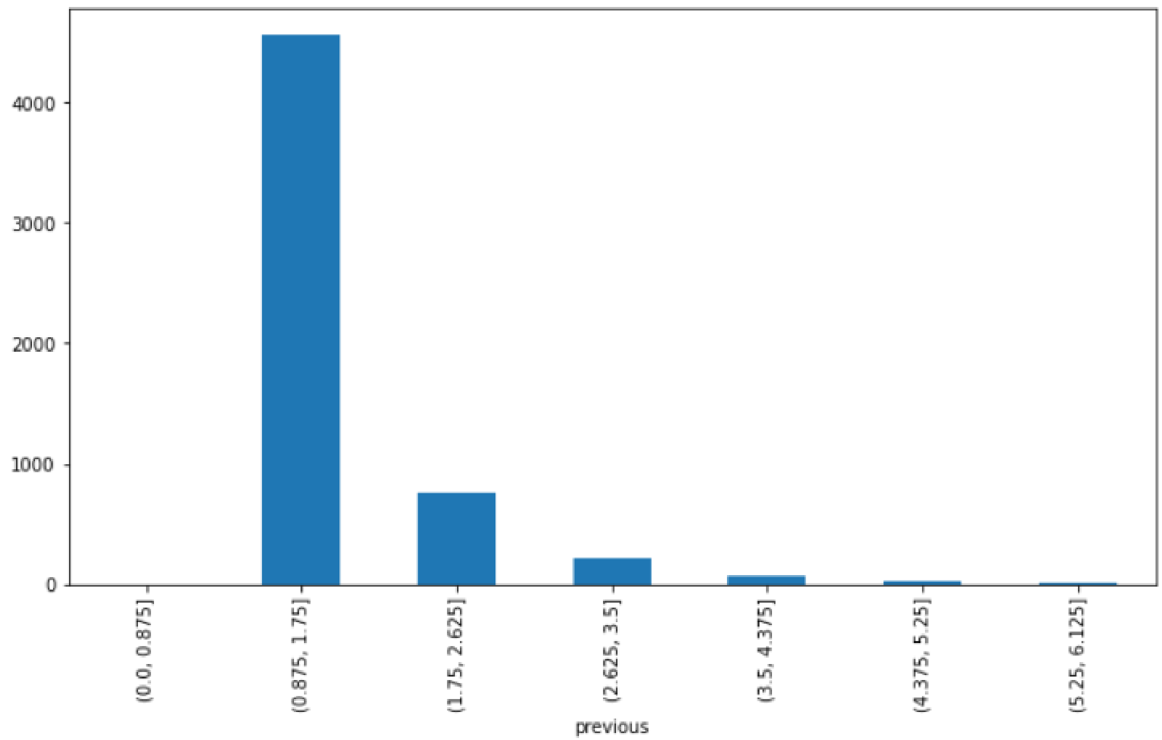
- Campaign



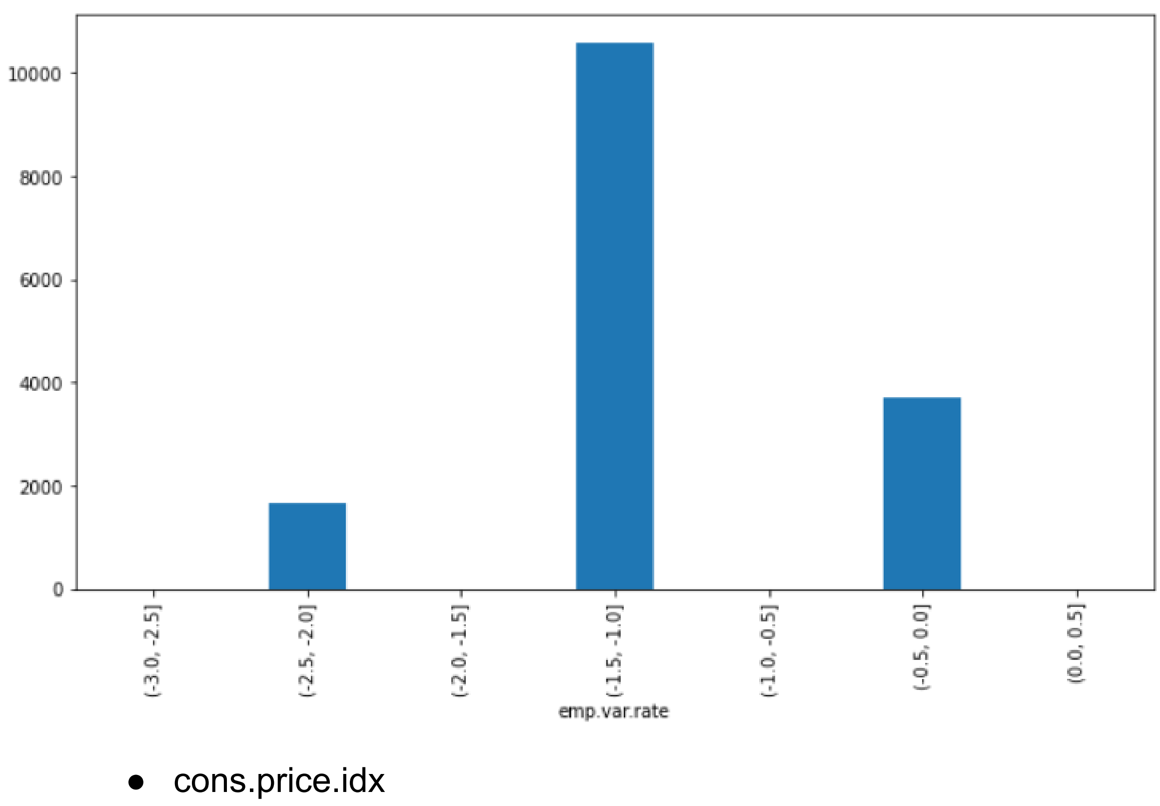
- pdays



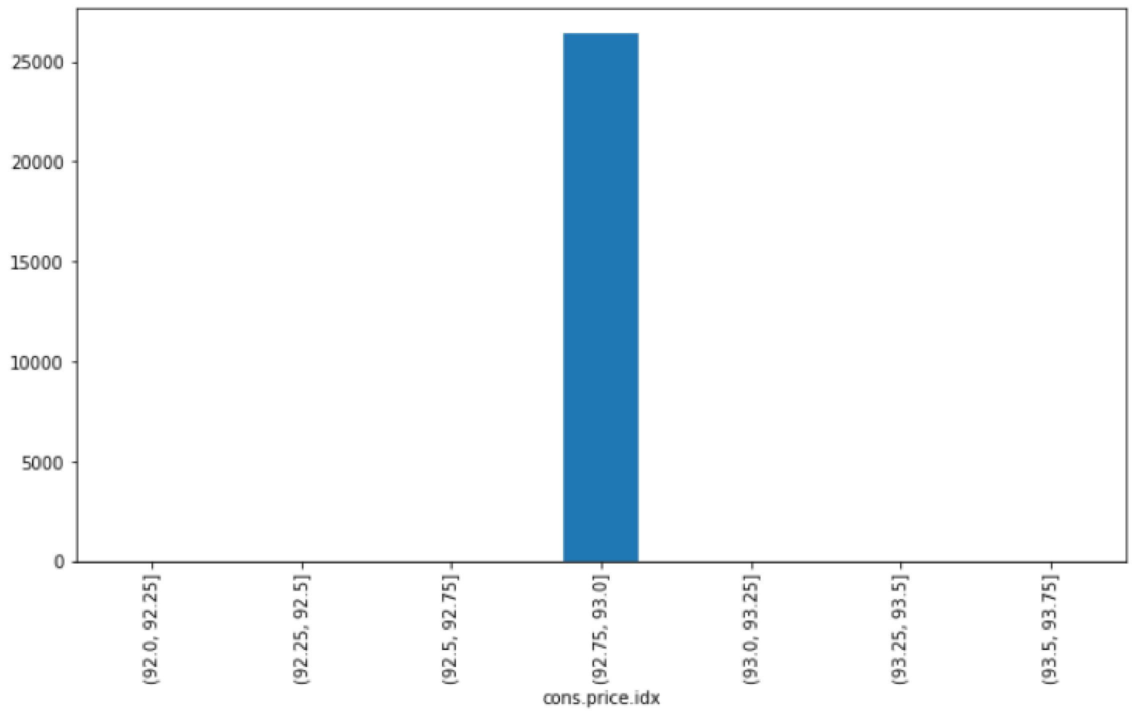
- Previous



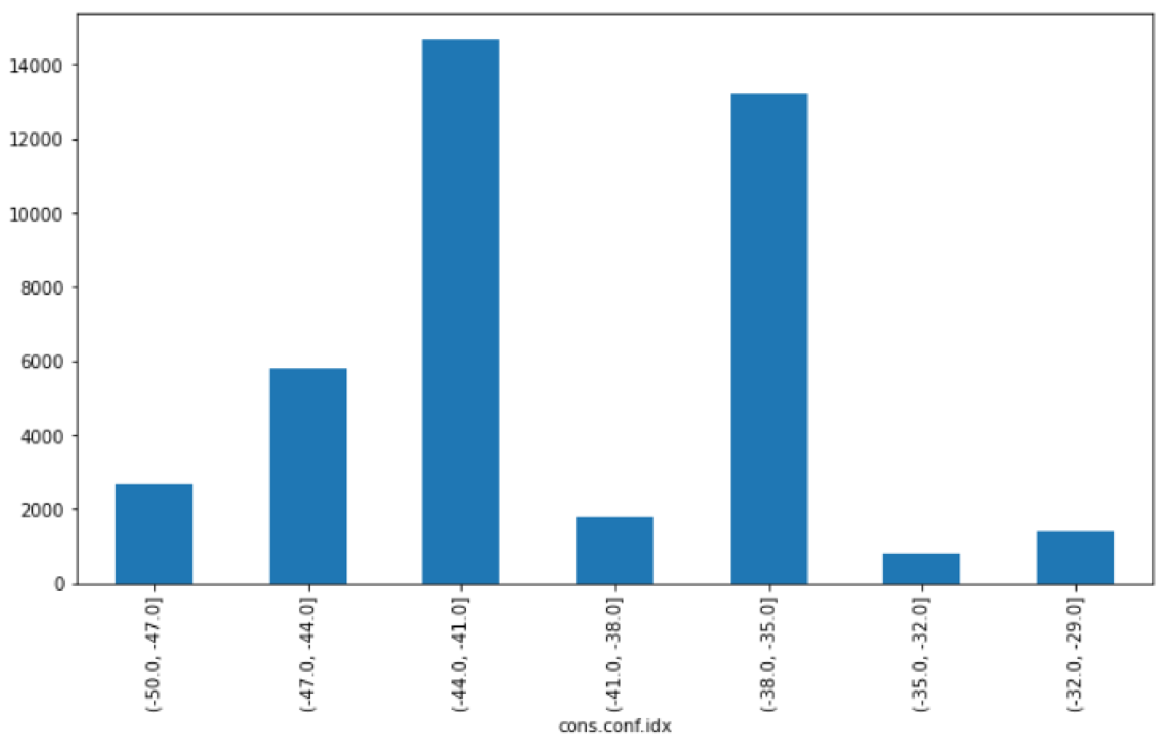
- emp.var.rate



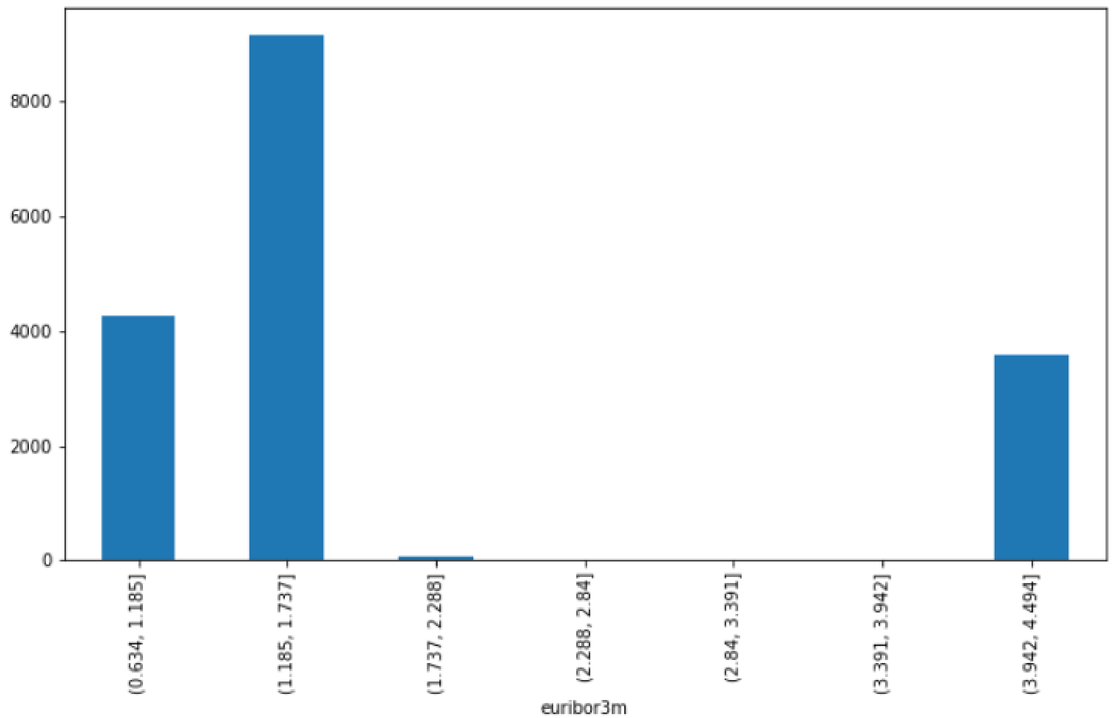
- cons.price.idx



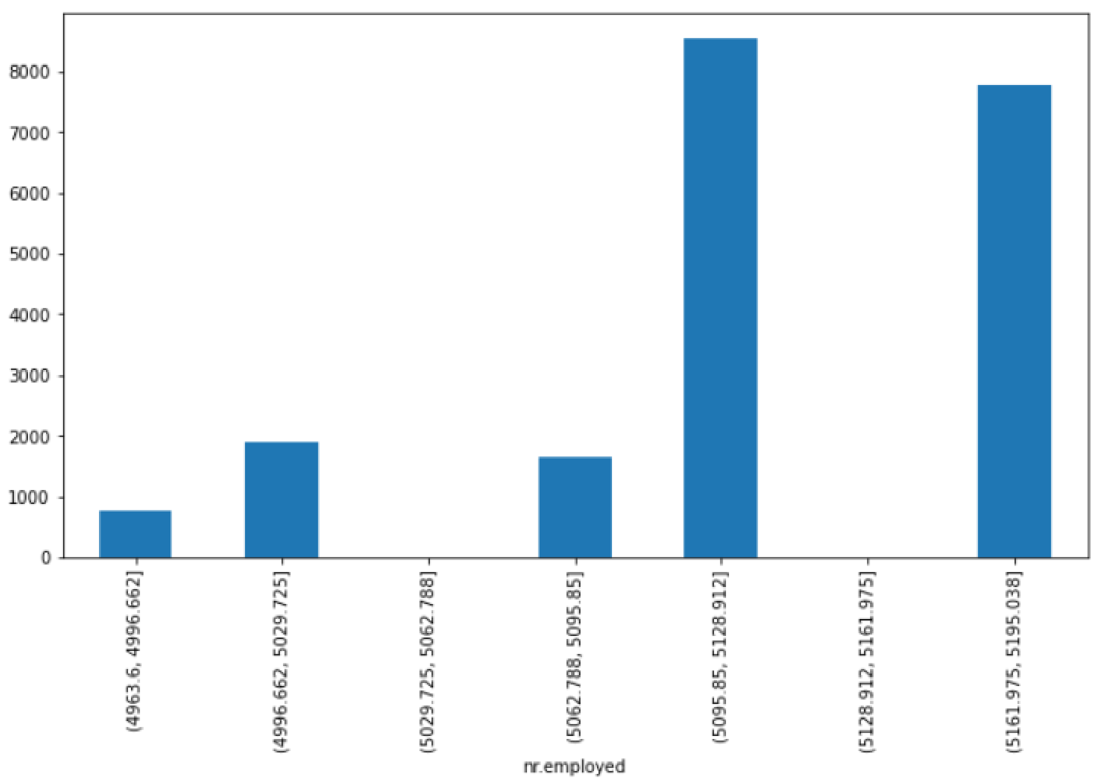
- cons.conf.idx



- euribor 3m



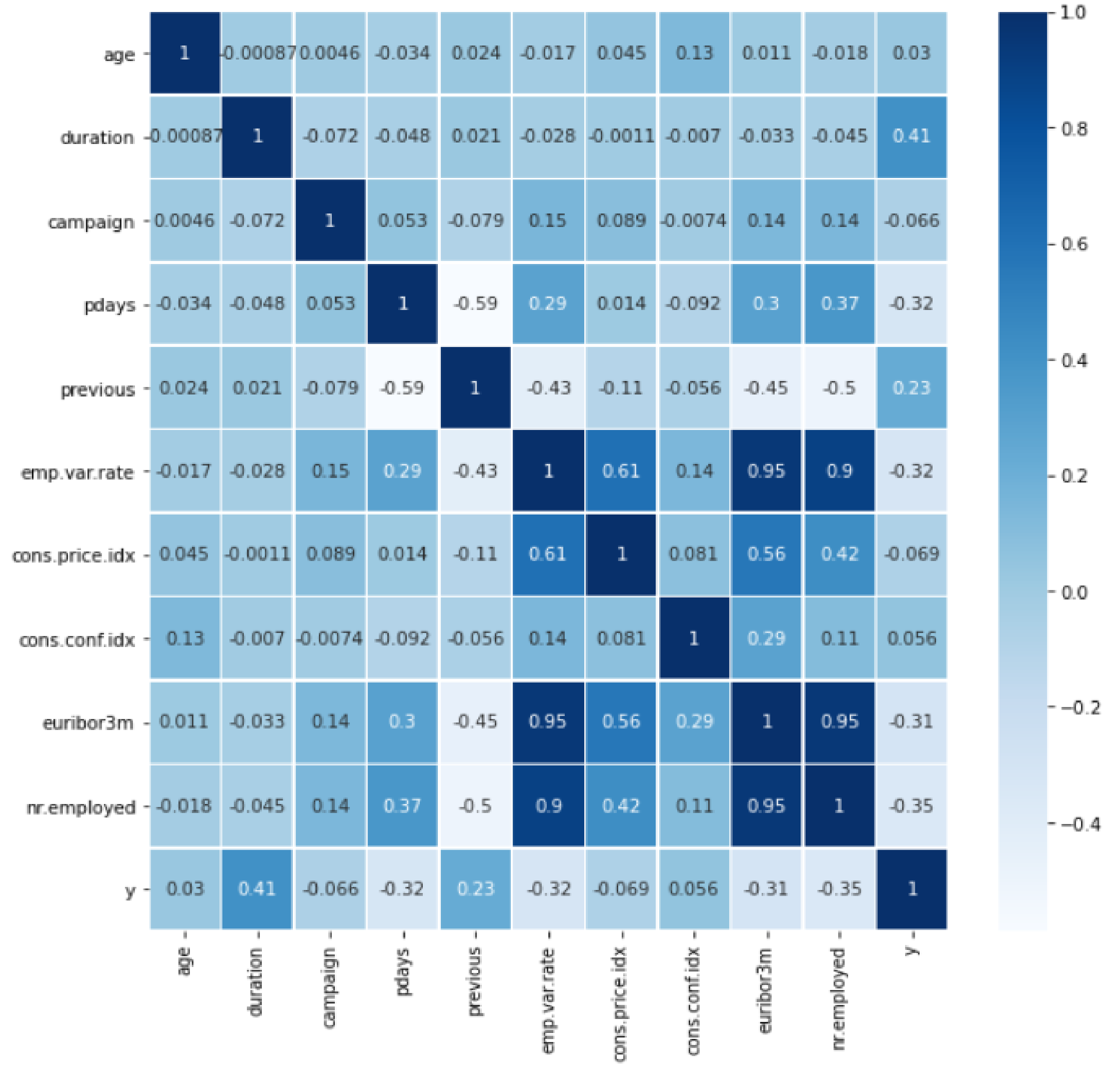
- nr.employed



- **Heatmap**

  Heatmap shows the correlation between each int variables (X) and outcome (Y)

Duration and previous records have the biggest impact on the outcome based on this map.



From these charts, some of information are missing in the string variables, some of them are noted with ‘unknown’ or ‘nonexist’, though there are no obvious NaN values in this dataset, these ‘unknown’ variables will have impact to the model, as they are meaningless to the model training and selection process.

Then further process needs to be done. To handle these values, I used RandomForestClassifier to forecast the missing values. Since the int variables are all set, I used them as Xs in this model, and indicators 'marital', 'education', 'default', 'housing', 'loan', 'poutcome' as Y respectively. The predicted values are replacements to the original ‘unknown’ and ‘nonexistence’ values.

Before and after processing the random forest classification predict method

```
married     24928
single      11568
divorced     4612
unknown        80
Name: marital, dtype: int64
university.degree      12168
high.school             9515
basic.9y                6045
professional.course     5243
basic.4y                4176
basic.6y                2292
unknown                 1731
illiterate                18
Name: education, dtype: int64
no         32588
unknown     8597
yes            3
Name: default, dtype: int64
yes        21576
no         18622
unknown      990
Name: housing, dtype: int64
no         33950
yes         6248
unknown      990
Name: loan, dtype: int64
cellular     26144
telephone    15044
Name: contact, dtype: int64
nonexistent    35563
failure         4252
success         1373
Name: poutcome, dtype: int64
```

```
married     24977
single      11591
divorced     4620
Name: marital, dtype: int64
university.degree      12725
high.school             9957
basic.9y                6271
professional.course     5394
basic.4y                4456
basic.6y                2367
illiterate                18
Name: education, dtype: int64
no     41185
yes        3
Name: default, dtype: int64
yes    22121
no     19067
Name: housing, dtype: int64
no     34914
yes     6274
Name: loan, dtype: int64
cellular     26144
telephone    15044
Name: contact, dtype: int64
failure    39815
success     1373
Name: poutcome, dtype: int64
```

- **Feature engineering**

Since some of the categories contain many features, dummy variables should be set up.

For the value in 'job', 'marital', 'education', 'default', 'housing', 'loan', 'contact', 'month', and 'poutcome', I used pd.get_dummies to generate the new dataset