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

Week 9, Sept/2,

Try at least 2 techniques to clean the data (for NA values : mean/median/mode/Model based approach to handle NA value/WOE and like this try different techniques to identify and handle outliers as well)

for NLP try different featurization technique and also clean the data using regex and python

Each member should code and review peers work. (Review comment should be present in the github repo)

Each team member should work on different data cleansing approach.

Note:

If one team member is using mean to impute values then other member should experiment on segmented approach or any other model based approach to impute the null values.

you are allowed to merge the code of each individual and work together to get good result.

Make sure code of each team member is placed at provided URL (single repository for whole team).

Week 10, Sept/9, visualization, interpretation, choose the model with the best accuracy and get its feature importance

Submit a pdf document and EDA ipynb file which should contain following details:

Team member's details : Group Name (give a name to your group), Name, Email, Country, College/Company, Specialization (Data Science, NLP, Data Analyst)

Problem description

Github Repo link

EDA performed on the data

Final Recommendation

Week 11, Sept/16, ppt

EDA presentation for business users

Last slide of EDA should be dedicated to technical user which should contain recommended models for this data set.

Week 12, Sept/23,

Select your base model and then explore 1 model of each family if its classification problem then 1 model for Linear models, 1- Model for Ensemble, 1-Model for boosting and other models if you have time (like stacking)

Please make sure selected model fits in your business requirement. For example : If your business does not want black box model then select only those models which can be used to explain the prediction.

As this is group assignment hence upload the code of each team member and other deliverables in the single repo and share the URL of that repo.

you are allowed to merge the code of each individual and work together to get good result.

Week 13, Sept/30, merge code and ppt

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

Data intake report

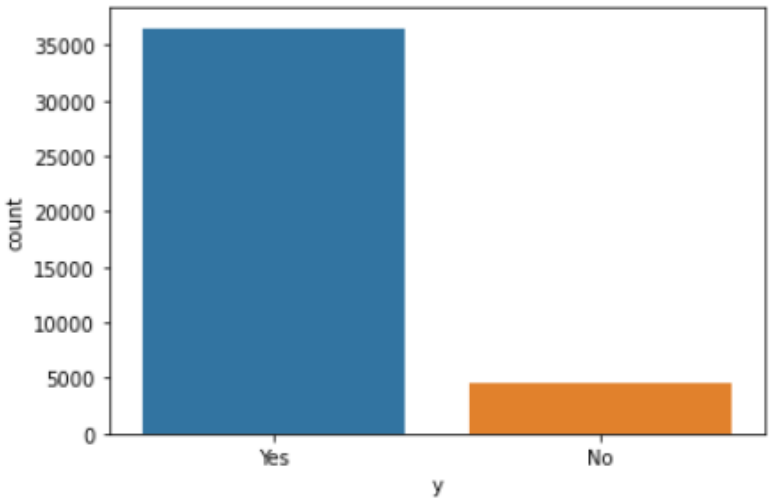
The dataset used for analysis is bank-full, accessed from UCI database.

- **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.

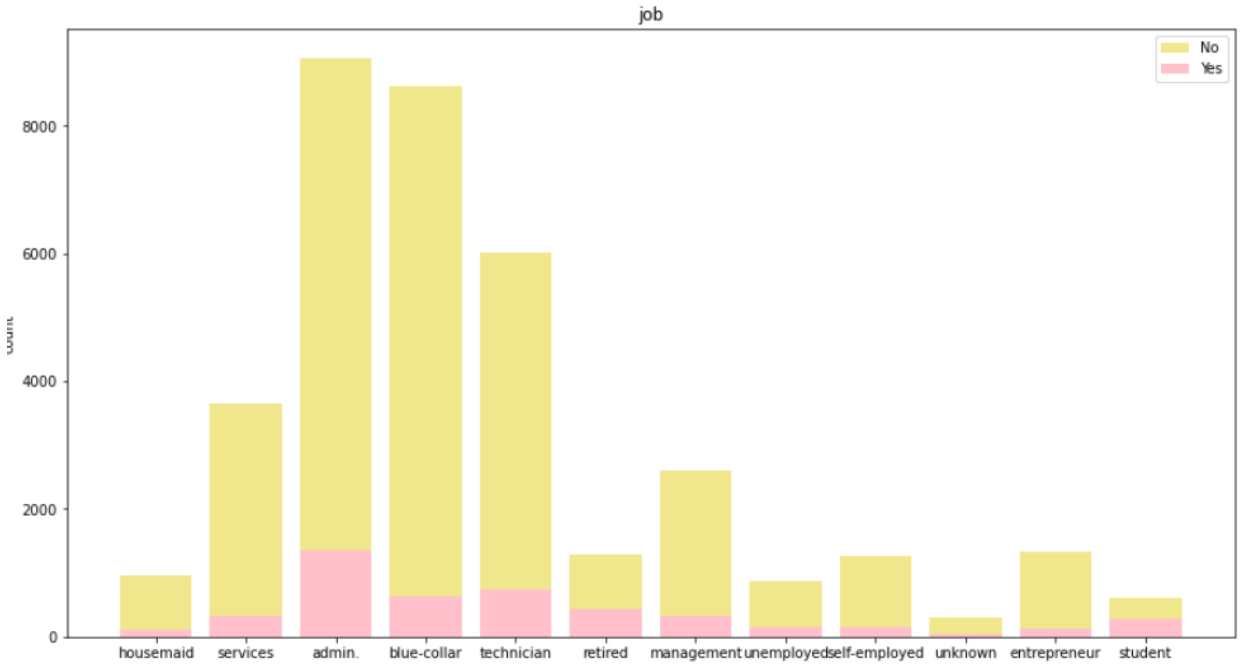
1. Age
2. Job: type of job ('admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown'
3. Marital: (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
4. Education: (categorical: '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**

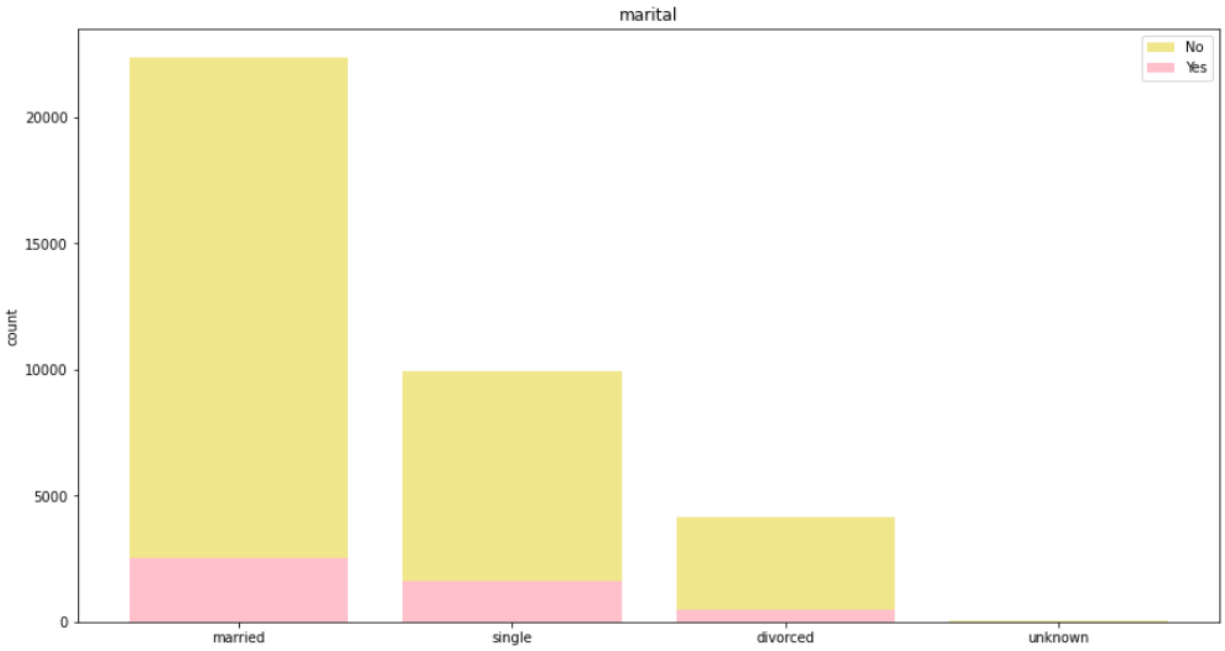
  The chart shows the y in this dataset is imbalance, in further machine learning process we need to make up for the unbalanced part or delete some data with the outcome of ‘Yes’ randomly.



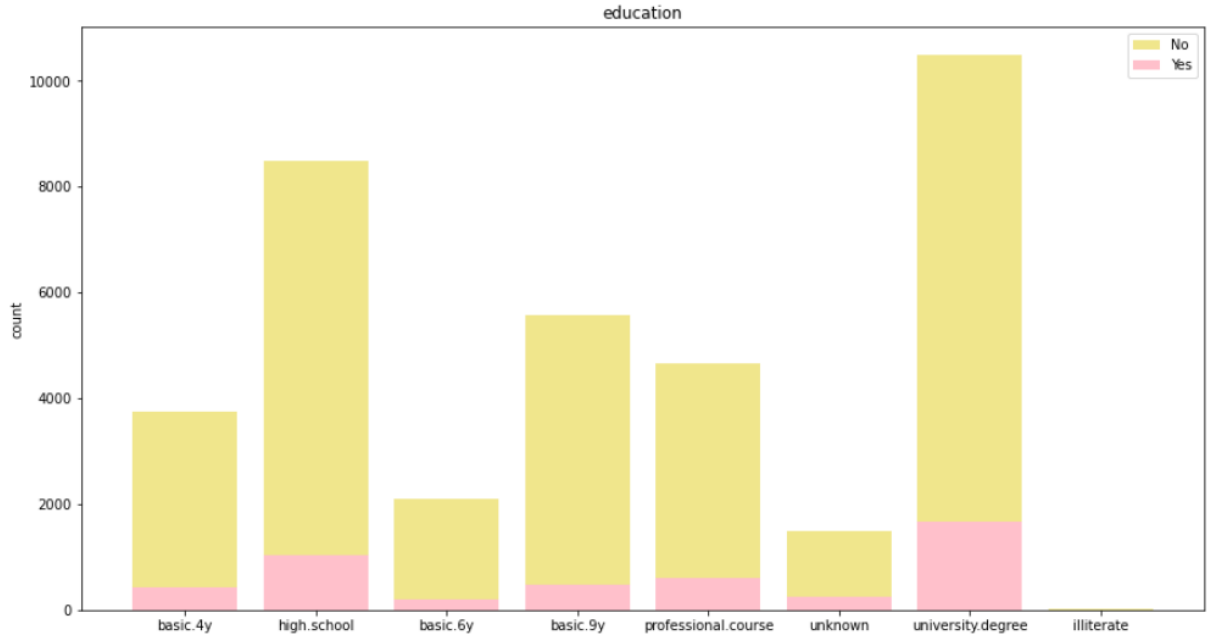
Discrete feature + outcome



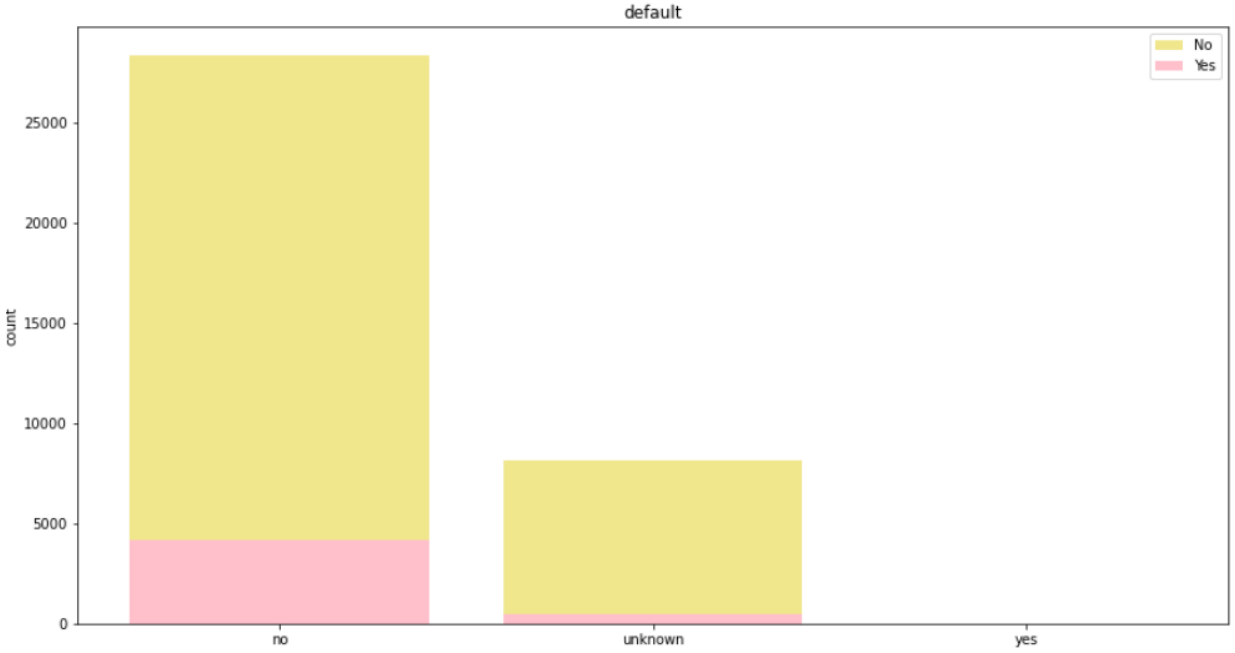
marital

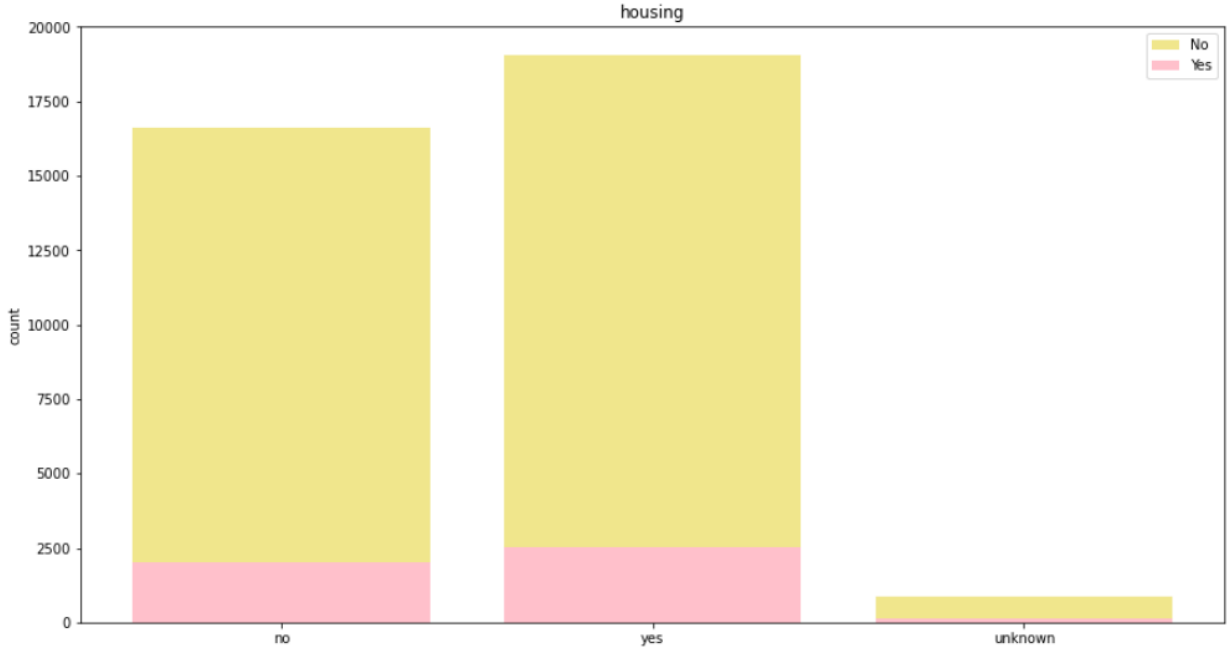


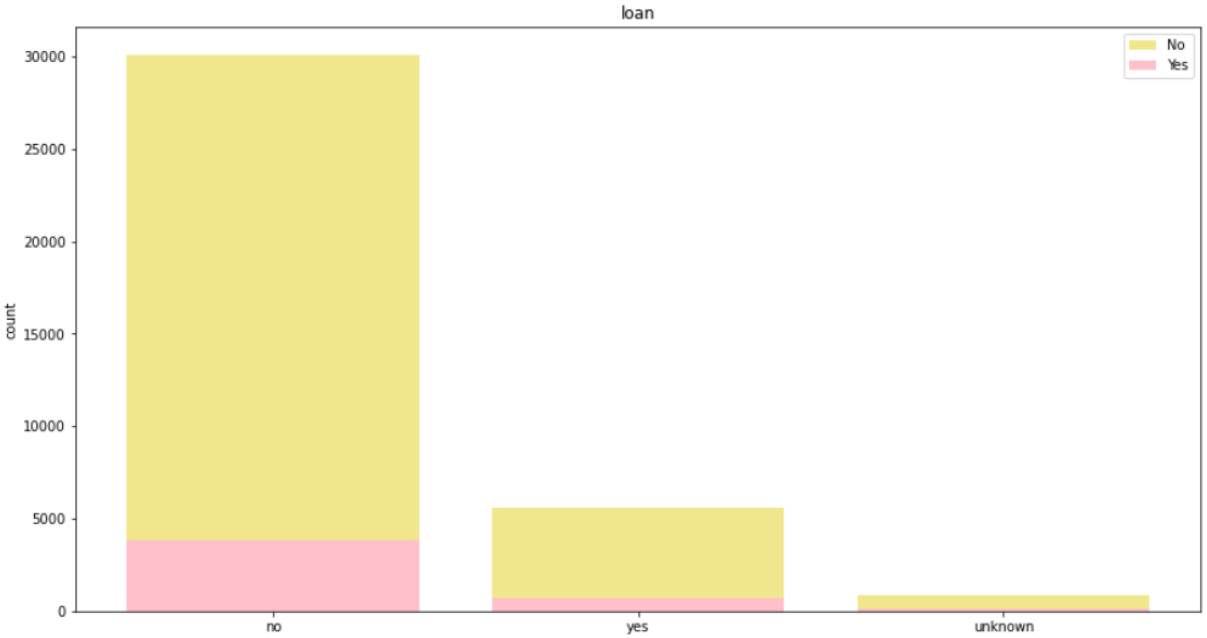
education

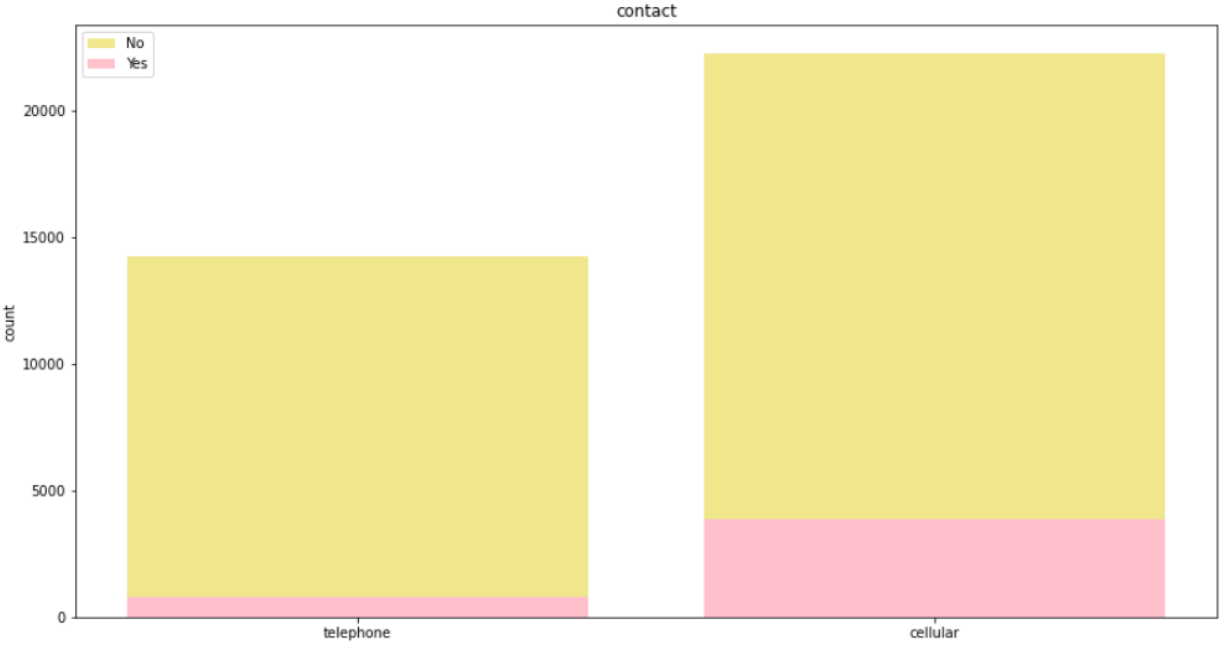


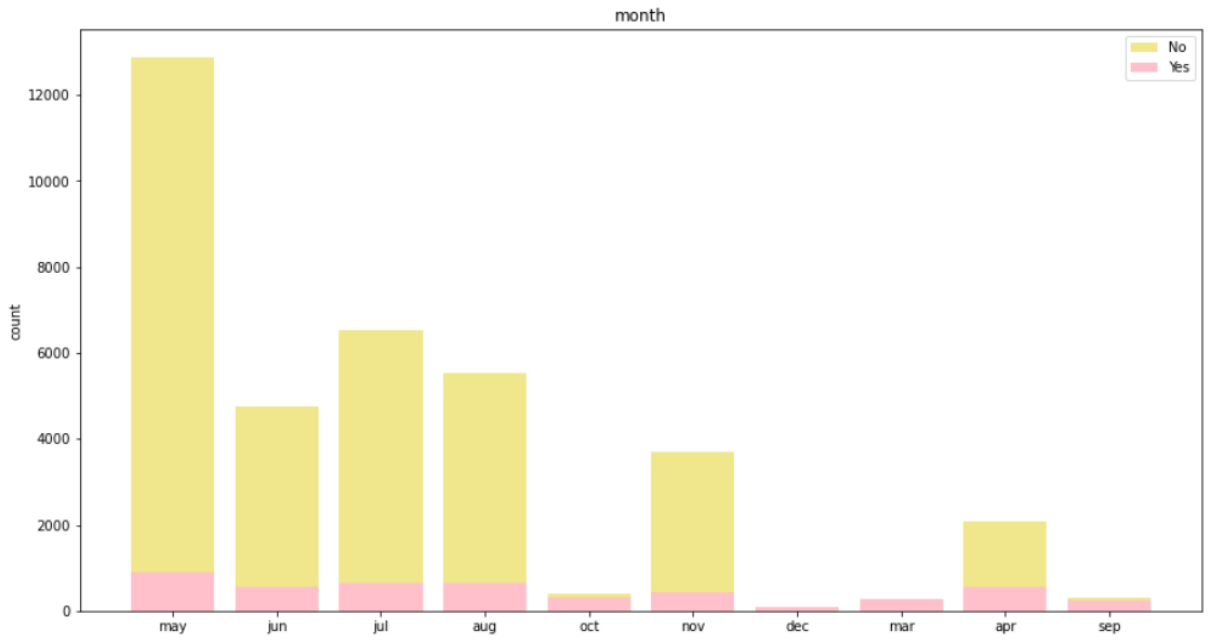
default

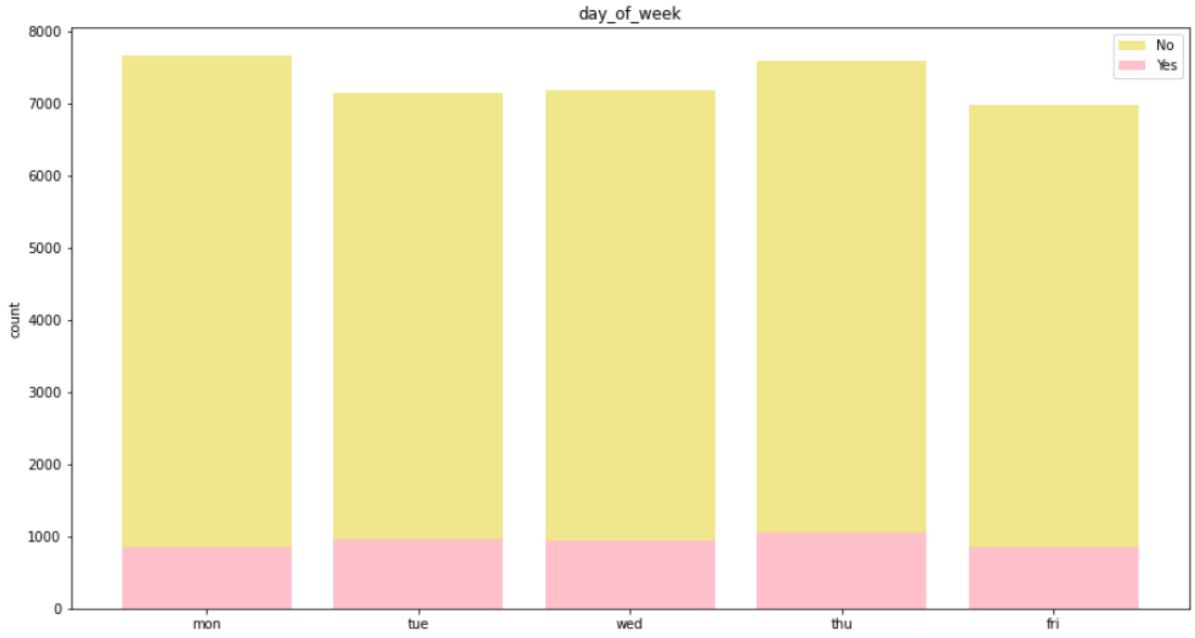

default
No
Yes
count
25000
20000
15000
10000
5000
0
no
unknown
yes

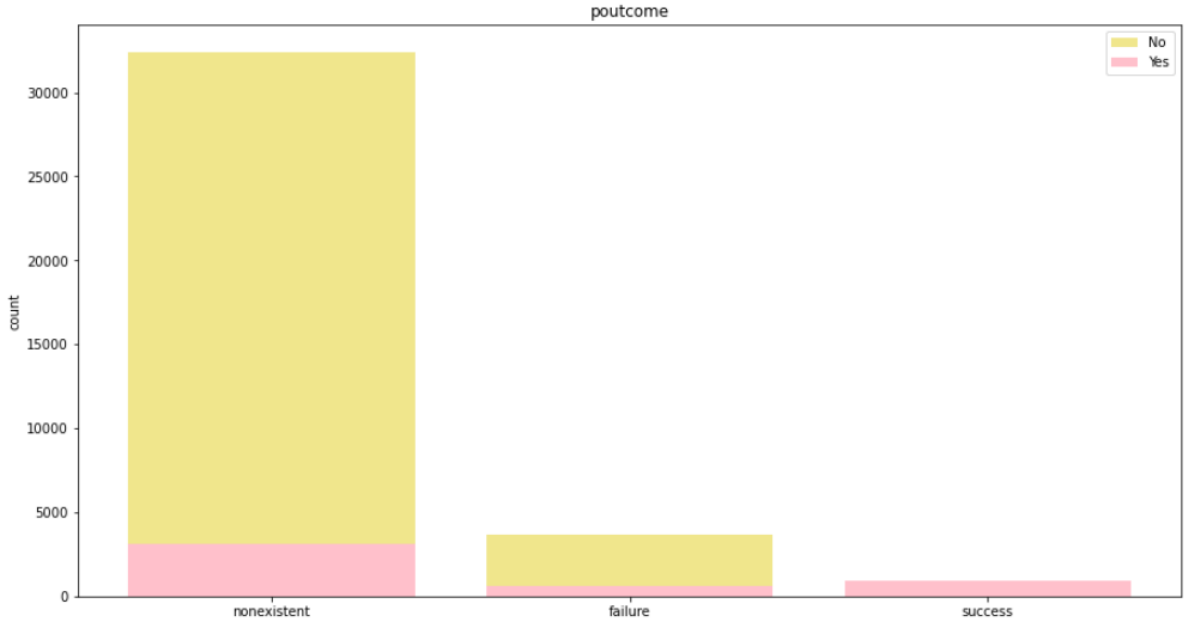

housing
No
Yes
count
20000
17500
15000
12500
10000
7500
5000
2500
0
no
yes
unknown

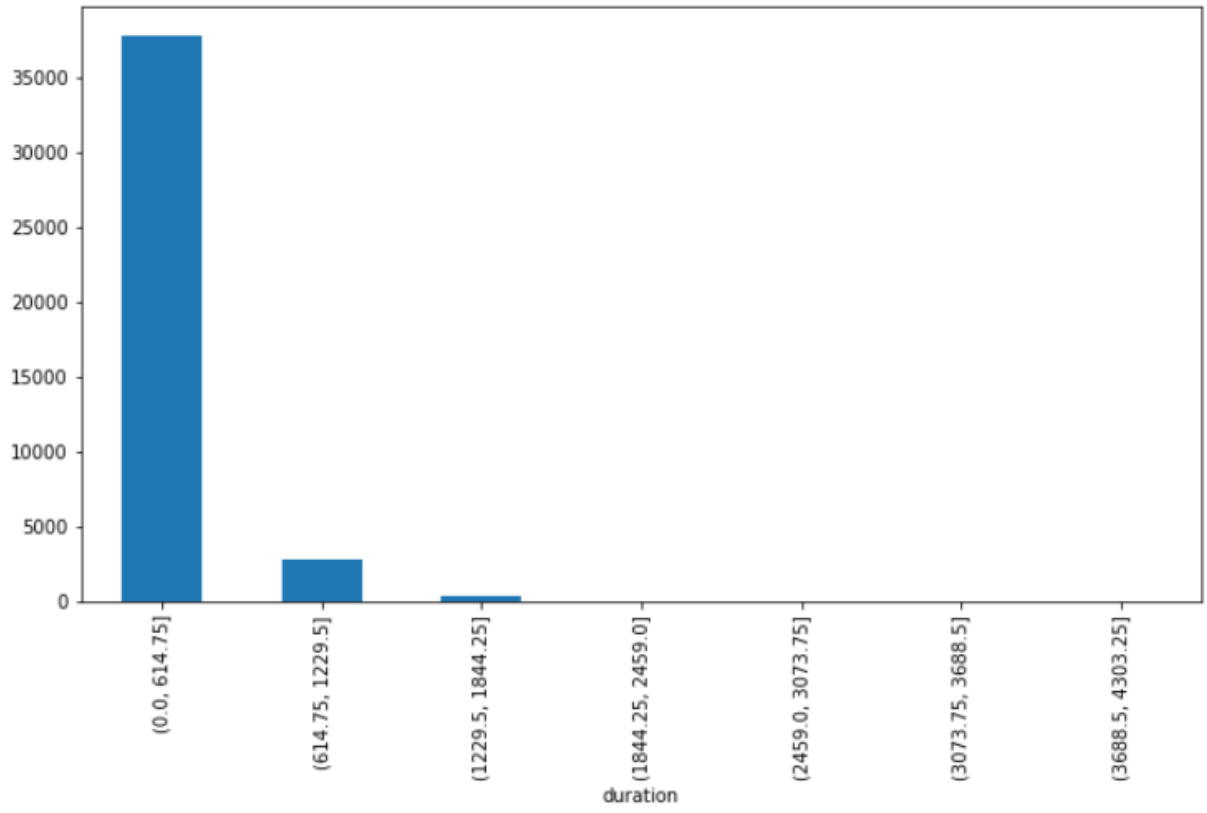

loan
No
Yes
count
30000
25000
20000
15000
10000
5000
0
no
yes
unknown

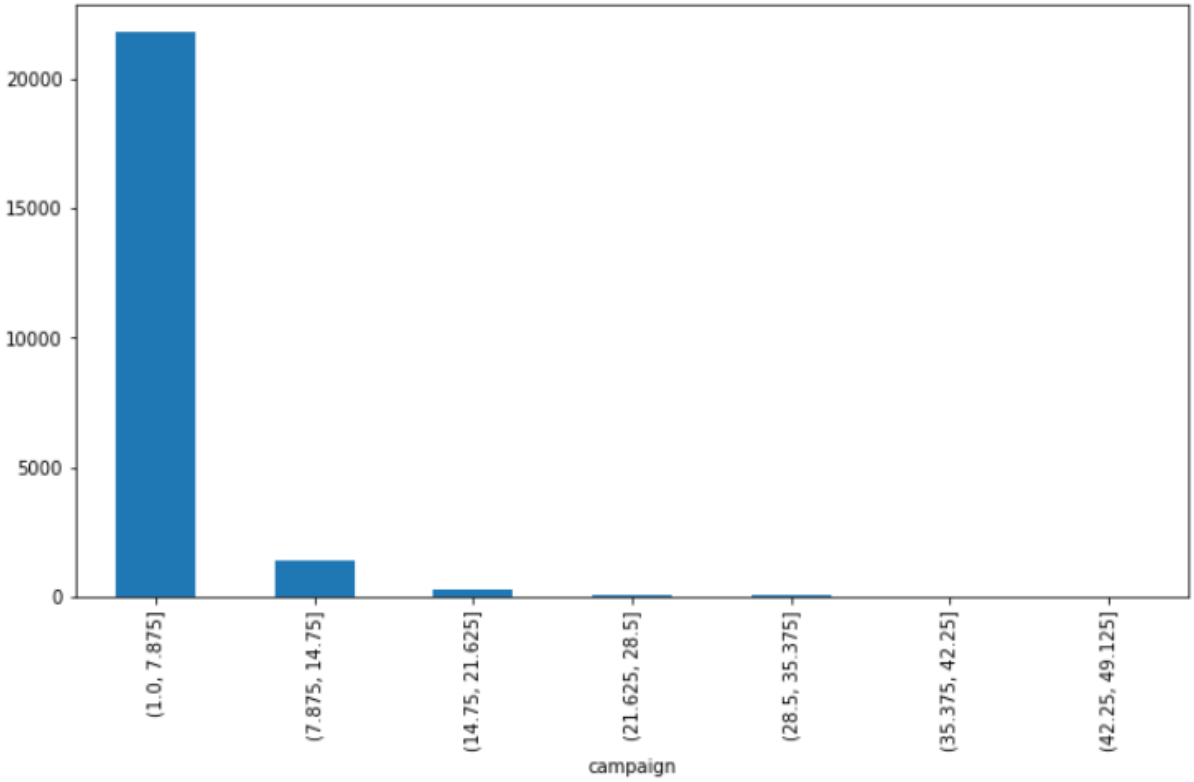

contact
No
Yes
count
20000
15000
10000
5000
0
telephone
cellular

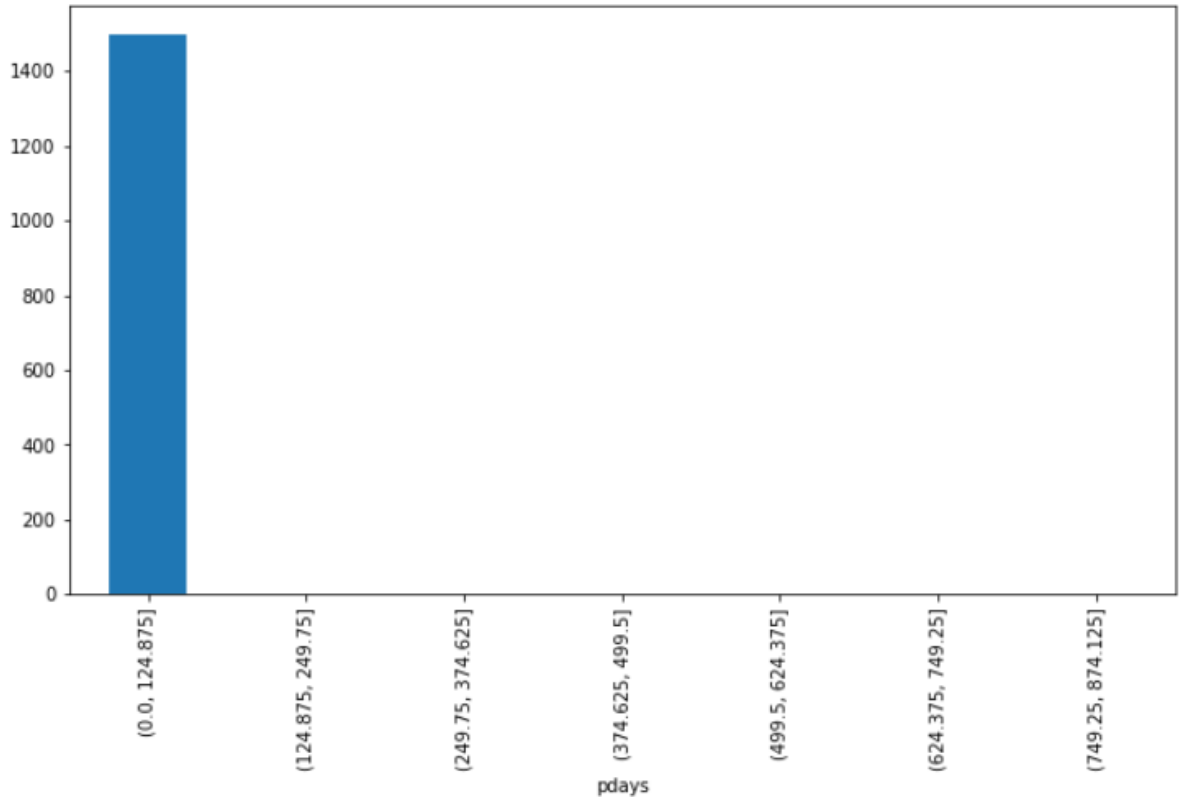

month
No
Yes
count
12000
10000
8000
6000
4000
2000
0
may
jun
jul
aug
oct
nov
dec
mar
apr
sep

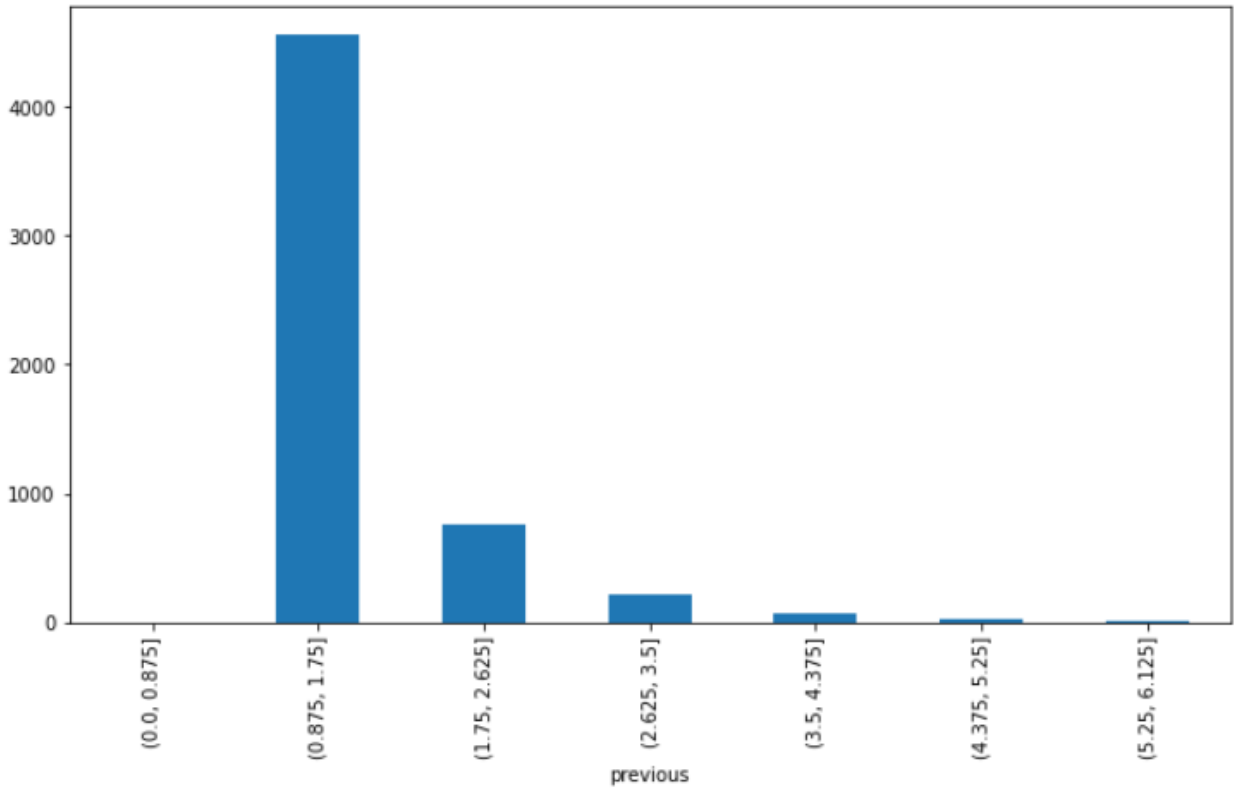

day_of_week
No
Yes
count
8000
7000
6000
5000
4000
3000
2000
1000
0
mon
tue
wed
thu
fri

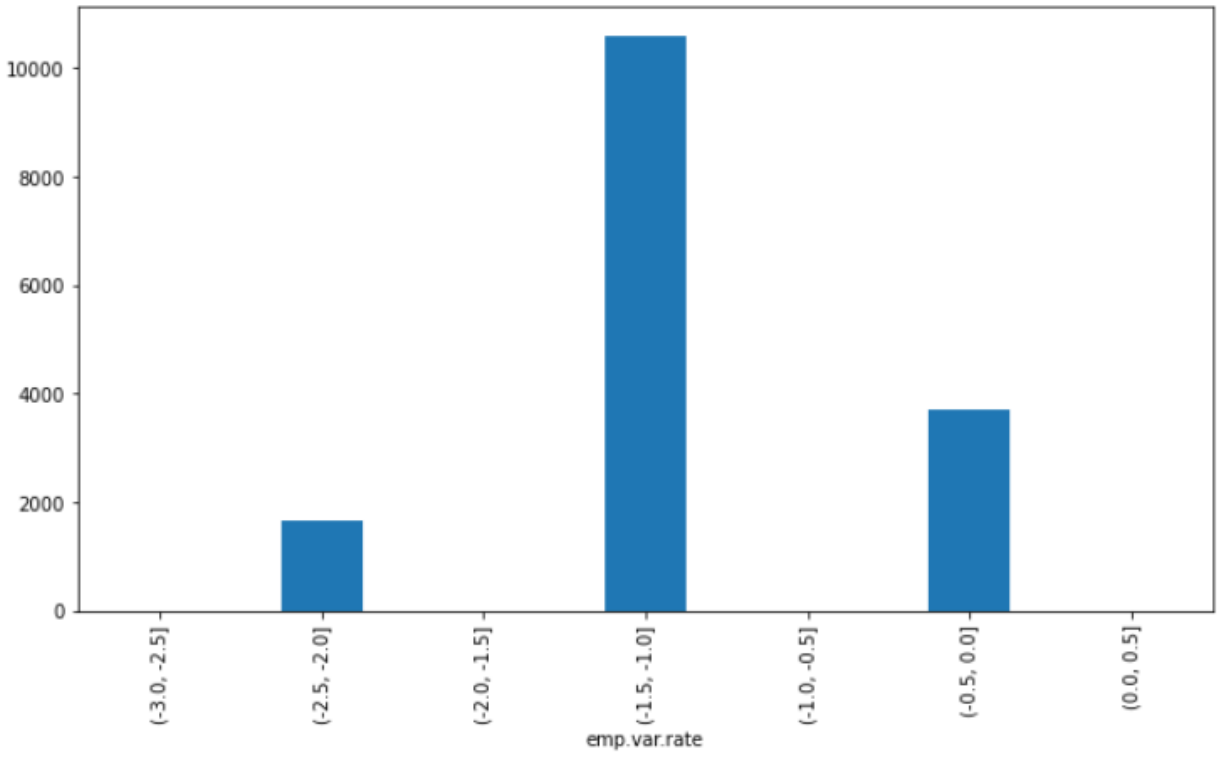

poutcome
No
Yes
count
30000
25000
20000
15000
10000
5000
0
nonexistent
failure
success

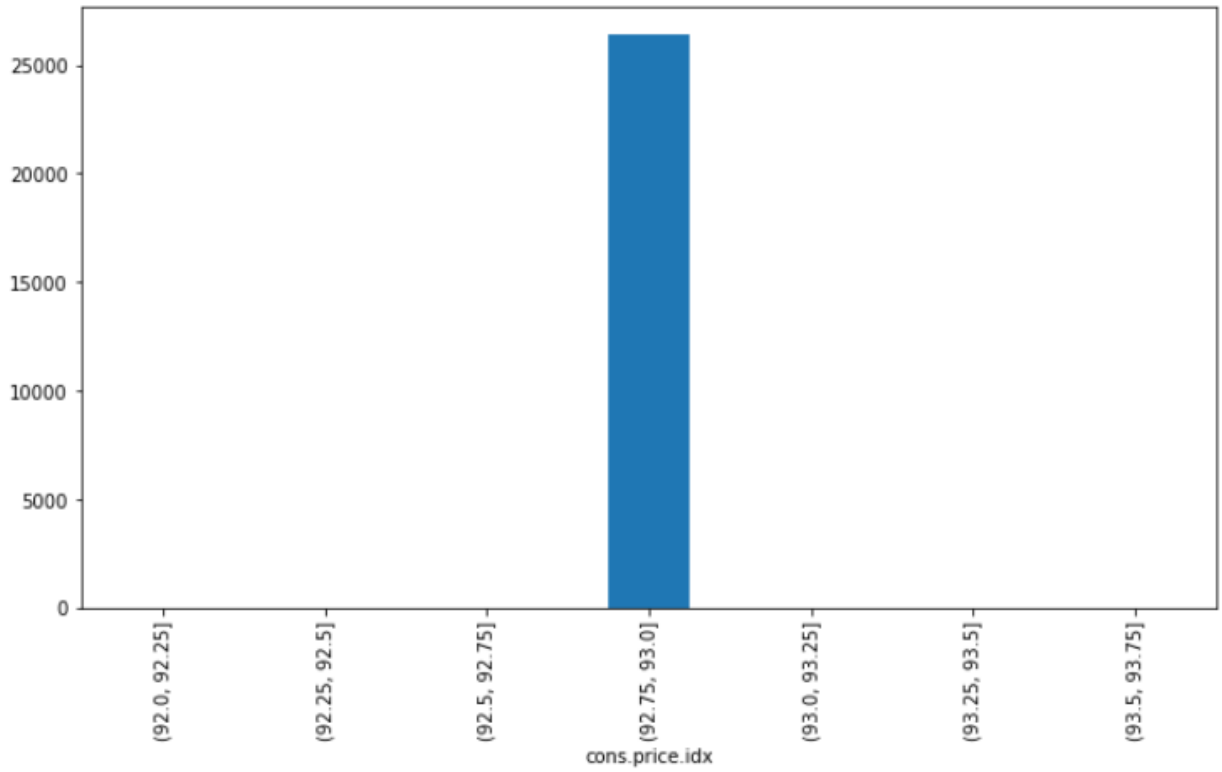

35000
30000
25000
20000
15000
10000
5000
0
(0.0, 614.75]
(614.75, 1229.5]
(1229.5, 1844.25]
(1844.25, 2459.0]
(2459.0, 3073.75]
(3073.75, 3688.5]
(3688.5, 4303.25]
duration

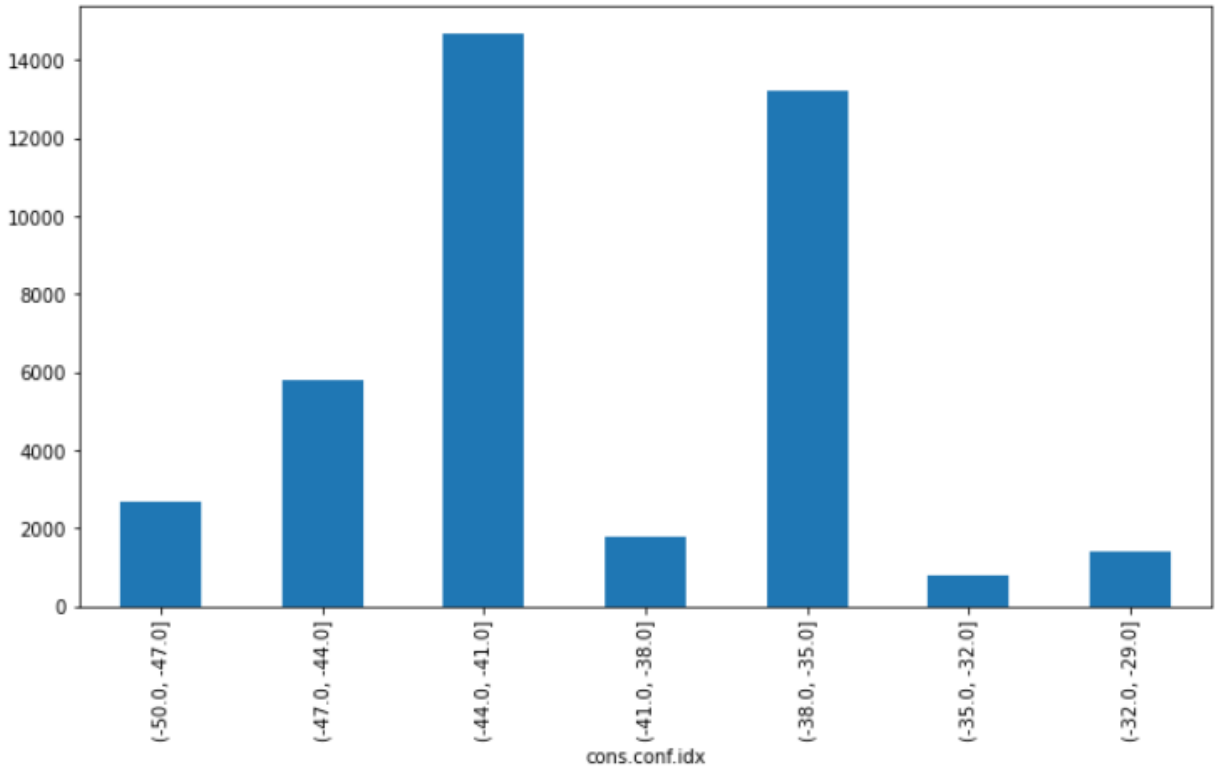

20000
15000
10000
5000
0
(1.0, 7.875]
(7.875, 14.75]
(14.75, 21.625]
(21.625, 28.5]
(28.5, 35.375]
(35.375, 42.25]
(42.25, 49.125]
campaign

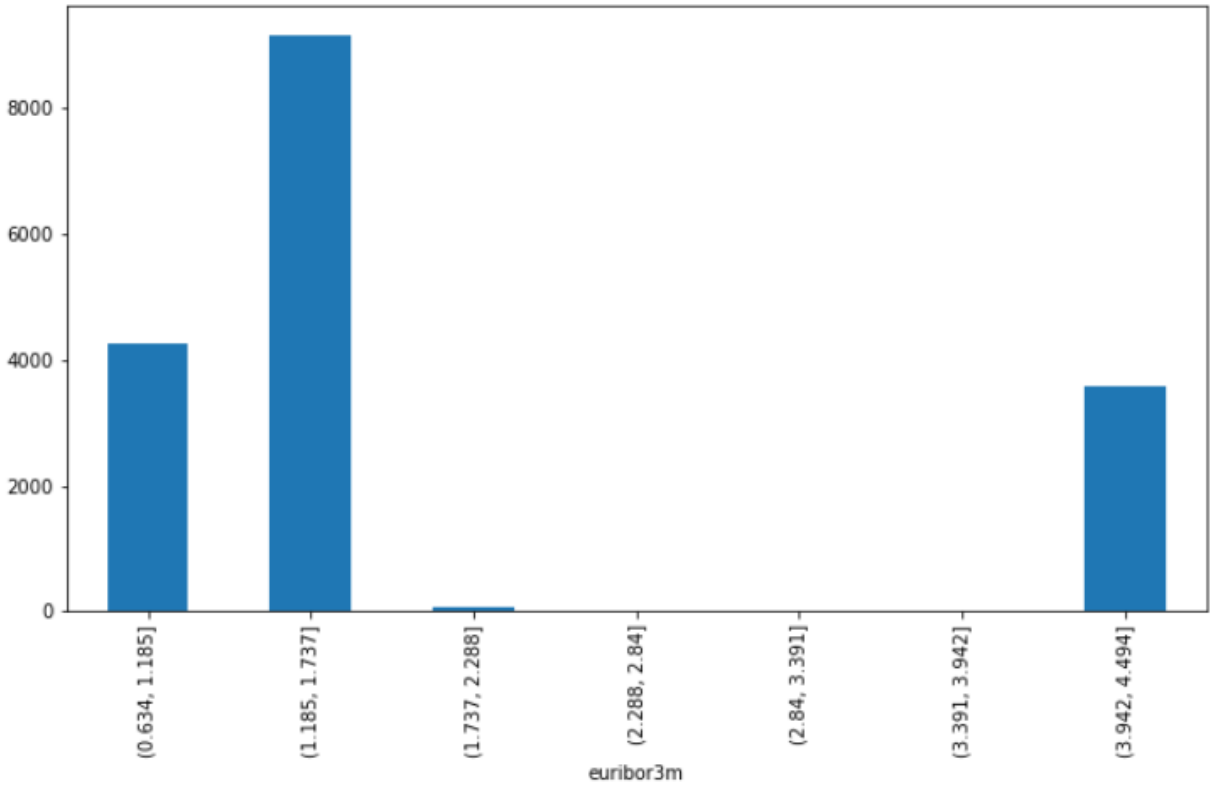

1400
1200
1000
800
600
400
200
0
(0.0, 124.875]
(124.875, 249.75]
(249.75, 374.625]
(374.625, 499.5]
(499.5, 624.375]
(624.375, 749.25]
(749.25, 874.125]
pdays

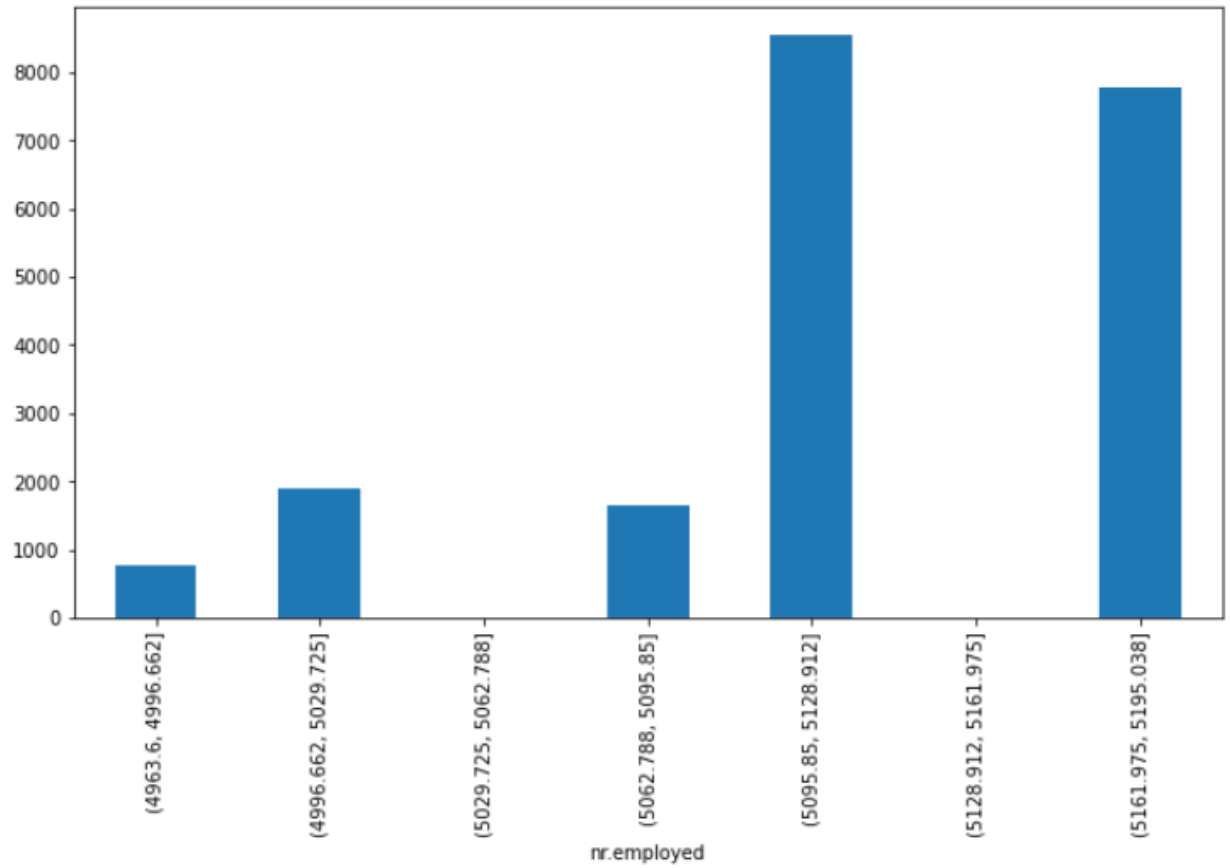

4000
3000
2000
1000
0
(0.0, 0.875]
(0.875, 1.75]
(1.75, 2.625]
(2.625, 3.5]
(3.5, 4.375]
(4.375, 5.25]
(5.25, 6.125]
previous


10000
8000
6000
4000
2000
0
(-3.0, -2.5]
(-2.5, -2.0]
(-2.0, -1.5]
(-1.5, -1.0]
(-1.0, -0.5]
(-0.5, 0.0]
(0.0, 0.5]
emp.var.rate


25000
20000
15000
10000
5000
0
(92.0, 92.25]
(92.25, 92.5]
(92.5, 92.75]
(92.75, 93.0]
(93.0, 93.25]
(93.25, 93.5]
(93.5, 93.75]
cons.price.idx

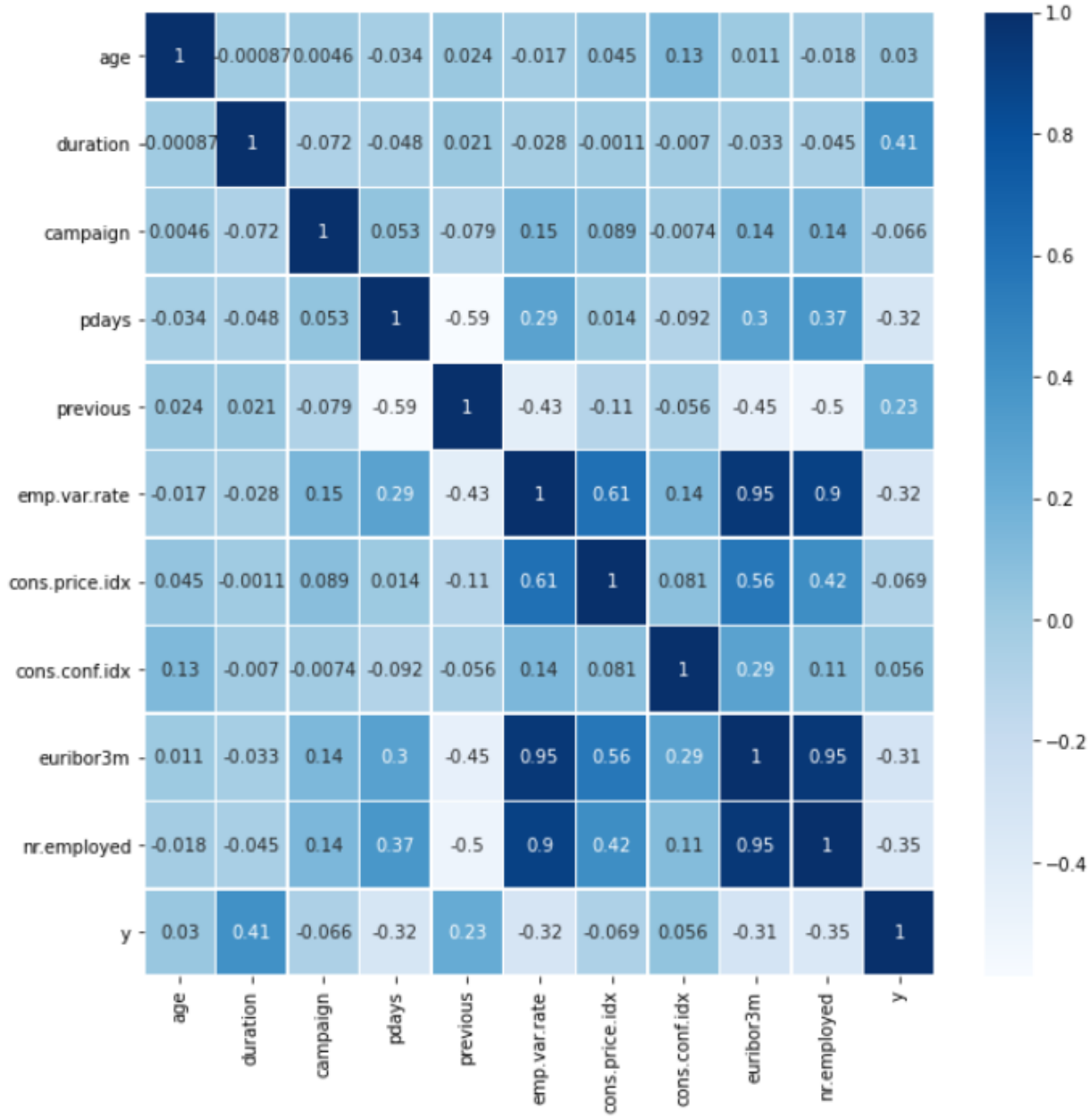

14000
12000
10000
8000
6000
4000
2000
0
(-50.0, -47.0]
(-47.0, -44.0]
(-44.0, -41.0]
(-41.0, -38.0]
(-38.0, -35.0]
(-35.0, -32.0]
(-32.0, -29.0]
cons.conf.idx


8000
6000
4000
2000
0
(0.634, 1.185]
(1.185, 1.737]
(1.737, 2.288]
(2.288, 2.84]
(2.84, 3.391]
(3.391, 3.942]
(3.942, 4.494]
euribor3m


8000
7000
6000
5000
4000
3000
2000
1000
0
(4963.6, 4996.662]
(4996.662, 5029.725]
(5029.725, 5062.788]
(5062.788, 5095.85]
(5095.85, 5128.912]
(5128.912, 5161.975]
(5161.975, 5195.038]
nr.employed

Heatmap

```
plt.figure(figsize=(8, 8))
sns.heatmap(bank.corr(),annot=True,cmap='Blues',linewidths=.5)
```



Feature engineering

```
#set dummies
predictors = ['job', 'marital', 'education', 'default', 'housing', 'loan', 'contact', 'month', 'poutcome']
banknew = pd.get_dummies(bank[predictors], drop_first=True)
banknew.head()
```