Отчет по лабораторной работе №0
по курсу «Исскуственный интеллект»

Выполнила студентка группы 8О-304б Лаар Марина

Тема: Анализ данных
Задание: Требуется сформировать/получить два набора данных соответствующие следующим критериям:

1. Один из датасетов должен представлять собой корпус документов. Язык, источник и тематика произвольна
2. Второй датасет должен содержать категориальные, количественные признаки. Для данного датасета определить предсказываемые признаки (для задачи регрессии и классфикации). Если такого признака нет, спроектировать

Данные датасеты будут в дальнейшем использованы в оставшихся лабораторных работах.

По каждому датасету построить распределения признаков (в случае корпуса документов – построить распределение слов) и объяснить имеющуюся картину. Вычислить статистические характеристики признаков. Обнаружить и решить возможные проблемы с данными. Если решить данную проблему невозможно, объяснить почему.

Ход работы:
1.Кредитный скоринг
Ссылка на датасет: https://www.kaggle.com/kabure/german-credit-data-with-risk
Первые 10 записей таблицы:

	Age	Sex	Job	Housing	Saving accounts	Checking account	Credit amount	Duration	Purpose	Risk
0	67	male	2	own	NaN	little	1169	6	radio/TV	good
1	22	female	2	own	little	moderate	5951	48	radio/TV	bad
2	49	male	1	own	little	NaN	2096	12	education	good
3	45	male	2	free	little	little	7882	42	furniture/equipment	good
4	53	male	2	free	little	little	4870	24	car	bad
5	35	male	1	free	NaN	NaN	9055	36	education	good
6	53	male	2	own	quite rich	NaN	2835	24	furniture/equipment	good
7	35	male	3	rent	little	moderate	6948	36	car	good
8	61	male	1	own	rich	NaN	3059	12	radio/TV	good
9	28	male	3	own	little	moderate	5234	30	car	bad

Статистические характеристики числовых признаков:

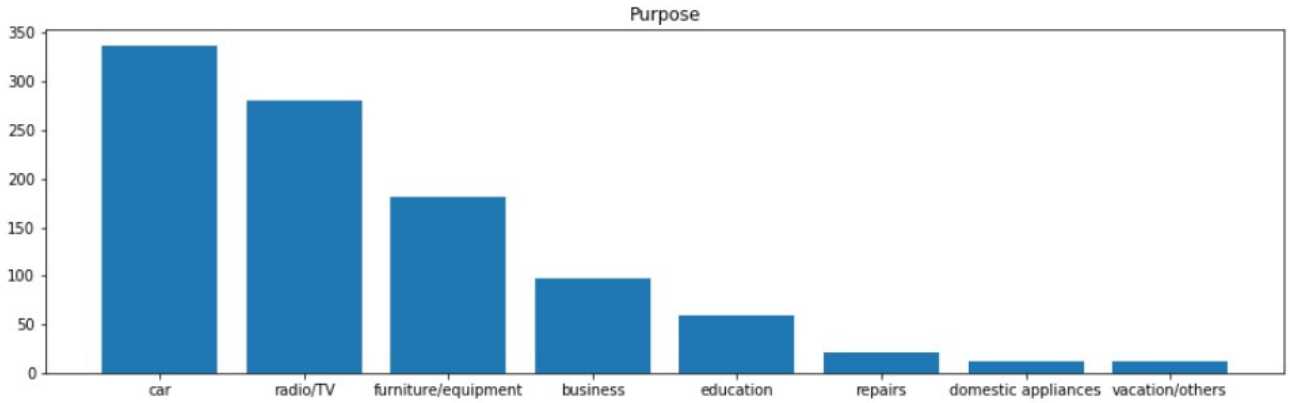
	Age	Job	Credit amount	Duration
count	1000.000000	1000.000000	1000.000000	1000.000000
mean	35.546000	1.904000	3271.258000	20.903000
std	11.375469	0.653614	2822.736876	12.058814
min	19.000000	0.000000	250.000000	4.000000
25%	27.000000	2.000000	1365.500000	12.000000
50%	33.000000	2.000000	2319.500000	18.000000
75%	42.000000	2.000000	3972.250000	24.000000
max	75.000000	3.000000	18424.000000	72.000000

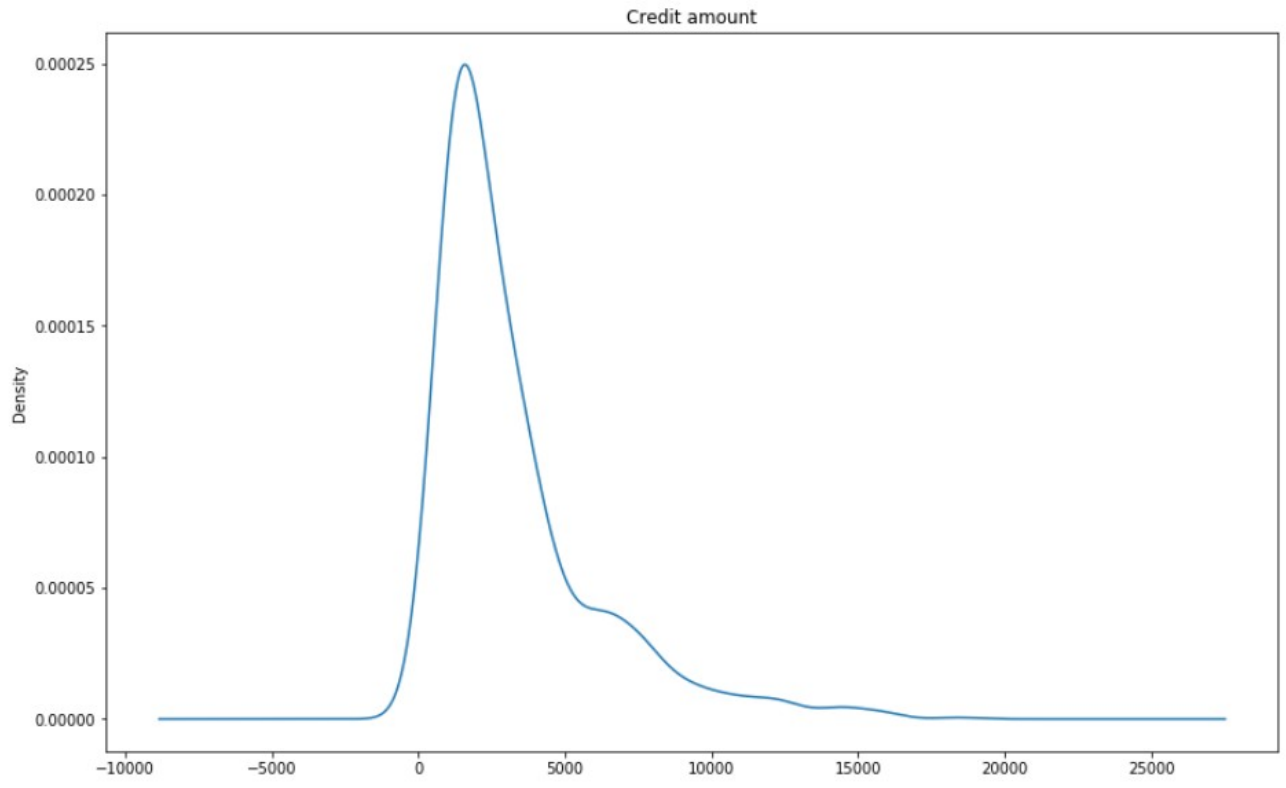
Примечание: Job является категориальным признаком, однако так как его значения численные pandas принял его за числовой признак

Статистические характеристики категориальных признаков:

	Sex	Housing	Saving accounts	Checking account	Purpose	Risk
count	1000	1000	817	606	1000	1000
unique	2	3	4	3	8	2
top	male	own	little	little	car	good
freq	690	713	603	274	337	700

Для примера также отобразим распределение двух признаков — Purpose и Credit Amount. Остальные распределения построены в исходном коде программы.





Проблемы:

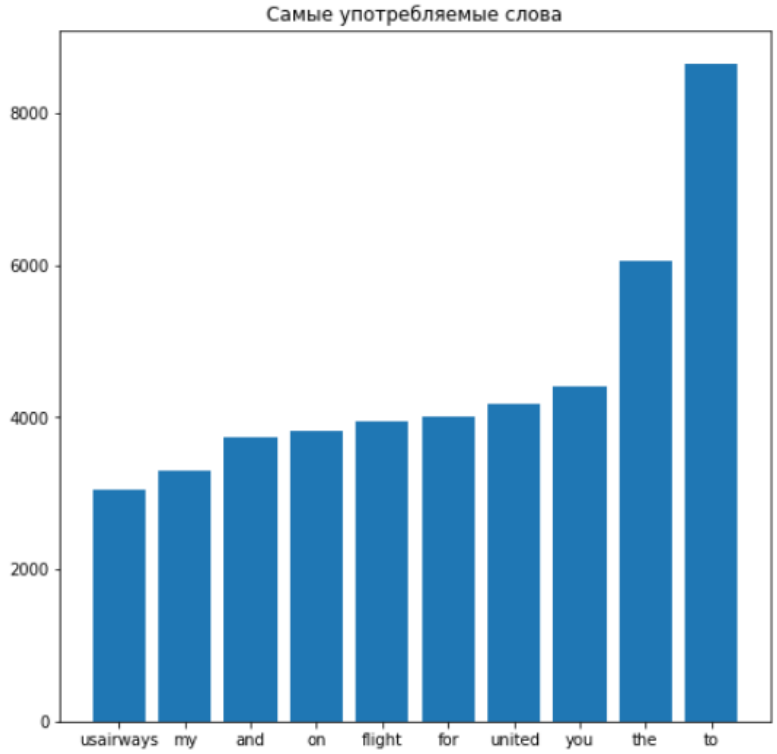
- Единственная проблема, которая бросается в глаза — это большое количество пропущенных значений в колонках Saving Accounts и Checking Account

2. Твиты о воздушных компаниях:

Ссылка на датасет: https://www.kaggle.com/crowdflower/twitter-airline-sentiment

Распределение слов в датасете:

	Word	NumOfOccurrences	Frequency
14086	usairways	3053	0.012321
9248	my	3288	0.013269
2281	and	3733	0.015065
9815	on	3815	0.015396
6004	flight	3939	0.015896
6157	for	3999	0.016138
13914	united	4164	0.016804
14944	you	4401	0.017761
13120	the	6061	0.024460
13326	to	8652	0.034916



Помимо самых часто-употребляемых слов (артиклей, предлогов и союзов) в топе можно заметить слова usairways, flight, united и тд. Это связано со спецификой датасета. Все твиты связаны с одной тематикой, предлагается классифицировать твиты на негативные, нейтральные и позитивные. Помимо анализа самих твитов, также предоставляются данные о дате создания твита, часовом поясе создателя и др.

Выводы:

В данной лабораторной работе мне пришлось поработать с двумя датасетами. Я построила распределение признаков и попыталась выявить возможные проблемы с данными.