Для начала пройдёмся по фичам с самыми малочисленными ненулевыми значениями представленным в документе. Для этого выведем таблицу всех фичей с ненулевыми значениями:

#	Column	Non-Null Count	Dtype
0	Аффективные_расстройства	29 non-null	object
1	ПсихПоздВозр	6 non-null	object
2	НеврСтресс	7 non-null	object
3	УО	1 non-null	object
4	Эпилепсия	9 non-null	object
5	ОргПорЦНС	18 non-null	object
6	Химическая зависимость]	125 non-null	object
7	БерРодыМать	806 non-null	object
8	ПатРодов	802 non-null	object
9	ПатРодовДруг	32 non-null	object
10	РанРазы	814 non-null	object
11	Ут2	41 non-null	object
12	НеротРасстрДетство	165 non-null	object
13	НевротРасстрШкола	810 non-null	object
14	НевротРасстрШколаДоп	206 non-null	object
15	Симптомы_е3	218 non-null	object
16	Симптомы_з2	467 non-null	object

Рисунок 1 — Кол-во ненулевых значений в каждом столбце

1. Из данной таблицы можем сразу исключить (удалить) столбец ***УО***, т.к. он имеет всего одно значение. В случае есть его надо оставить, то значения следует заменить на 0 – отсутствующие значения и 1 – Сиблинги.

2. Рассмотрим ***ПсихПоздВозр***:

ПсихПоздВозр	count
Бабка по линии матери	4
Дед по линии матери	1
Бабка по линии отца	1

Рисунок 2 — Уникальные значения столбца

Т.к. от матери психические расстройства передаются чаще, чем от отца, то мы можем занулить записи про деда по линии матери и бабки по линии отца. Тогда у нас столбец будет: 0 — «Отсутствующие значения, Дед по линии матери и Бабка по линии отца», 1 — «Бабка по линии матери»

3. Рассмотрим ***НеврСтресс***:

НеврСтресс	count
Сиблинги	2
Мать	2
Отец	1
Бабка по линии матери	1
Дядя/Тётя по линии матери	1

Рисунок 3 — Уникальные значения столбца

Данные по отцу и линии матери можно занулить, т.к. мы не можем проверить данные показания ведь в столбце все они встречаются ровно один раз. Однако учитывая, что данные по линии матери могут иметь значительный вес при определении растройств, мы можем заменить их на 1, соответственно данные по матери и сиблингам ввиду близкого родства можем объединить также в одну категорию 2 (так как они имеют выше вес, чем данные родственников не первой линии).

4. Рассмотрим следующий столбец ***Эпилепсия***:

Эпилепсия	count
Дядя/Тетя по линии отца	2
Дядя/Тётя по линии матери	2
Мать	2
Отец	1
Дед по линии матери	1
Сиблинги	1

Рисунок 4 — Уникальные значения столбца

Ситуация схожая с предыдущим столбцом, однако если мы не должны занулять 3 нижних уникальных значения, то объединим их так же в 4 категории: 0 – «нулевые и Отец», 1 – «Дядя/тётя по линии отца, Дед по линии матери», 2 – «Дядя/тётя по линии матери», 3 – «Мать, Сиблинги». Т.к линия матери имеет более большой вес в определении мы 2 линию её родственников выделяем отдельно, но можно и совместить с 1 категорией, как родственников дальше 1 линии.

5. Рассмотрим ***ОргПорЦНС*** :

ОргПорЦНС	count
Отец	7
Мать	4
Мать, Отец, Бабка по линии матери, Дед по линии матери	2
Бабка по линии матери	2
Сиблинги	1
Дядя/Тётя по линии матери	1
Дети	1

В отличии от уже рассмотренных столбцов здесь нам надо ещё разделить по запятой отдельных родственников, указанных в одной строке, что бы учитывались все родственники. Также, как и в предыдущих случаях делим на категории: 0 — «отсутствующие значения, Отец и Дед», 1 — «Бабка и Дядя/тётя по линии матери», 2 — «Мать и Сиблинги». Желательно отдельно проверить корреляцию и если Отец сильно влияет на результат отнести его к 1 категории, так как значений с ним больше всего в столбце.

6. Рассмотрим ***Аффективные расстройства***:

Аффективные_расстройства	count
Мать	18
Отец	6
Бабка по линии матери	2
Бабка по линии отца	1
Дед по линии матери	1
Дядя/Тётя по линии матери	1

Разнесём подобным образом на категории:

0 — «отсутствующие значения, Отец, Дед по линии матери, Бабка по линии отца», 1 — «Бабка и Дядя/тётя по линии матери», 2 — «Мать».

7. Рассмотрим ***Химическая зависимость***:

Химическая зависимость]	count
Отец	72
Мать	11
Дед по линии матери	10
Мать, Отец	10
Бабка по линии матери	5
Дед по линии отца	3
Мать, Отец, Бабка по линии матери, Дед по линии матери	2
Отец, Дед по линии матери, Дед по линии отца	2
Сиблинги	2
Дядя/Тётя по линии матери	1

Надо будет разделить все данные перечисленные через запятую, выделить данные схожим образом по категориям:

0 — «отсутствующие значения», 1 — «Дядя/тётя, Дед и Бабка по любой из линий», 2 — «Мать, Отец, Сиблинги». Мы выделяем это таким образом, т.к. люди находящиеся ближе всего к проверяемую влияют больше всего на вещи по типу Хим.зависимости.

8. Рассмотрим ***БерРодыМать***:

БерРодыМать	count
Без патологии	533
Не оценивались (в т.ч. не известно)	194
Патология беременности	24
Токсикоз второй половины беременности	9
Токсикоз первой половины беременности, …	6
Патология беременности, Токсикоз первой…	5
Токсикоз первой половины беременности	5
Психотравмирующие события	4
Патология беременности, Токсикоз первой…	3
Патология беременности, Токсикоз второй…	3

Кол-во отсутствующих значений очень мало, в данном столбце мы можем предположить, что их отсутствие — это как «Не оценивались», потому присвоим им отдельную категорию 0. Так же мы можем вынести это в два разных столбца, разделить на «наличие оценки» (0 и 1) и «Оценка патологии».

Тогда в «Оценке патологии» мы сделаем следующие категории: 0 — «Отсутствует», а «Патология беременности», «Токсикоз 1 или 2 половины», «Психотравмирующие события» разнести в категории 1 – 4 в зависимости от тяжести патологий. Если несколько патологий, то их сумму поместить в столбец для оценки.

9. Рассмотрим ***ПатРодов***:

ПатРодов	count
Нет	680
Преждевременные роды	28
Гипоксия плода в родах	26
Кесарево сечение	26
Затяжные роды	14
Затяжные роды, Гипоксия плода в родах	6
Преждевременные роды, Гипоксия плода в родах, Кесарево сечение	4
Гипоксия плода в родах, Наложение щипцов	3
Наложение щипцов	3
Гипоксия плода в родах, Кесарево сечение	3

Соответственно делим данные с запятыми, соответственно отсутствие значений и «нет» можно отнести в 0 категорию, а остальные, в соответствии с тяжесть, в 1-5, так же при наличии нескольких поместить сумму в столбец

10. Рассмотрим ***ПатРодовДруг***:

ПатРодовДруг	count
нет данных	3
при рождения диагноз гидроцефалия	2
обвитие пуповиной	2
Искусственные роды	1
Родился одним из двойни, второй плод- мумифицированный	1
Околоплодные воды зеленые, 5-6 баллов по Апгар, закричал не сразу	1
При рождении диагностирован вирусный гепатит В	1
Опущения правого века в связи с родовой травмой	1
Низкая масса тела при рождении - 2300г	1
Кефалогематома в родах	1

Практически все данные представлены в единичном экземпляре, остальные данные отсутствуют. Лучше всего было бы удалить данный столбец или представить в виде наличия и отсутствия, т.е. 0 и 1, где 0 — отсутствие данных (включая записи «нет данных»), а 1 — их наличие.

11. Рассмотрим ***РанРазы***:

РанРазы	count
В соответствии с возрастом	641
Не оценивалось	127
Наличие отклонений, уточнить, в т.ч. обращаемость за медицинской помощью	46

Все значения присутствуют, т.к. их всего 3, причём двое противоположны друг другу, а третий нейтрален, лучше их представить в следующем виде: 0 — «Не оценивалось», 1 — «В соответствии с возрастом», -1 — «Наличие отклонений».

12. Рассмотрим ***Ут2***:

Ут2	count
легкая задержка умственного развития	2
отставания в моторном развитии	2
задержка развития	2
После рождения выставлен диагноз пареза…	1
В 3 месячном возрасте перенес грипп в т…	1
С рождения отставал в психофизическом р…	1
Наблюдение у невролога по поводу повыше…	1
Поллиноз, наблюдение у аллерголога	1
С 4 месяцев отставал в физическом и пси…	1
с детства была маленького роста, плохо …	1

Аналогичная ситуация как с *ПатРодовДруг*: Практически все данные представлены в единичном экземпляре, остальные данные отсутствуют. Лучше всего было бы удалить данный столбец или представить в виде наличия и отсутствия, т.е. 0 и 1, где 0 — отсутствие данных (включая записи «нет данных»), а 1 — их наличие.

13. Рассмотрим ***НеротРасстрДетство***:

НеротРасстрДетство	count
Простые фобии	43
Симбиотическая привязанность к значимому взрослому	29
Логоневроз	14
Энурез	13
Простые фобии, Симбиотическая привязанность к значимому взрослому	9
Эмоциональная лабильность (неустойчивость)	8
Логоневроз, Энурез	6
Логоневроз, Простые фобии	5
Энурез, Простые фобии	5
Парейдолические переживания	3

Много отсутствующих данных, их мы отнесём к категории 0, остальные фобии представим от в виде чисел в зависимости от их сложности, тогда наличие нескольких фобий можем представить в виде суммы.

14. Рассмотрим ***НевротРасстрШкола***:

НевротРасстрШкола	count
В соответствии с возрастом	525
Наличие отклонений	176
Не оценивалось	78
Отставание в психомоторном развитии	19
Гиперактивность	4
Значительное опережение в психомоторном развитии	3
Дефицит внимания	2
Диссоциация психического и физического развития	2
Трудности установления контактов и общения в детском коллективе	1

Большинство данных записано в 3-ой системе оценивания: не оценивалось, есть отклонения и в соответствии с возрастом. Потому мы можем привести столбец к этому виду, записав все другие значения симптомов как наличие отклонений. Тогда будут следующие категории:

0 — «не оценивалось», 1 — «в соответствии с возрастом», 2 — «есть отклонения».

15. Рассмотрим ***НевротРасстрШколаДоп***:

НевротРасстрШколаДоп	count
Трудности установления контактов и обще…	75
Трудности установления контактов и обще…	26
Гиперактивность, Дефицит внимания	18
Трудности установления контактов и обще…	8
уточнить, в т.ч. обращаемость за медици…	7
Гиперактивность	7
Дефицит внимания	6
Диссоциация психического и физического …	4
уточнить, в т.ч. обращаемость за медици…	4
Дефицит внимания, Трудности установлени…	4

Много отсутствующих данных, потому можем так же их выделить 0 категорией, остальным в зависимости от сложности присвоить числа, данные в единственном экземпляре присвоить так же 0.

16. Рассмотрим ***Симптомы_е3***:

Симптомы_е3	count
Анамнез, Статус	142
Анамнез	44
Анамнез, Статус до лечения	17
Статус	9
Статус до лечения, Статус после лечения	2
Статус до лечения	2
Статус до лечения = Анамнез	1
Анамнез, Статус до лечения, Статус после лечения	1

Многие данные данного столбца отсутствуют, соответственно им назначается категория 0. Данные так же разделяем по запятым. Анамнез – это сбор данных о пациенте для лечения, присвоим ему категорию 1. Статусы будем оценивать соответственно 2 — «до лечения» и 3 — «после лечения». Тогда предположим, что сумма (при наличии нескольких записей) будет нам давать какой-то результат по поводу лечения.

17. Рассмотрим ***Симптомы_з2***:

Симптомы_з2	count
Анамнез, Статус	349
Анамнез, Статус до лечения, Статус после лечения	46
Анамнез	24
Статус	21
Анамнез, Статус до лечения	13
Статус до лечения	5
Статус до лечения, Статус после лечения	5
Статус после лечения	3
Анамнез, Статус после лечения	1

Аналогичная ситуация, разделяем присваиваем категории 0-3, вычисляем сумму.

Для большего понимания это следует уточнить у составителя данных записей.