Министерство науки и образования РФ
Федеральное государственное бюджетное учреждение
высшего образования
“Тверской Государственный Технический Университет”
(ТвГТУ)
Кафедра Программного обеспечения

Отчет по лабораторной работе №2
По дисциплине: «Анализ больших данных»
Тема: “Исследовательский анализ данных. Постановка гипотез
Категориальные данные“

Выполнил:
студент группы
Б.ПИН.РИС-21.06
Миронов М.В.

Проверила:
старший преподаватель
кафедры ПО
Корнеева Е.И.

Тверь 2025 г.

Содержание

1. Задача

1. Посчитать количество строк и столбцов
2. Провести разведочный анализ
   - Для числовых столбцов
     - Доля пропусков
     - Мин/Макс значения
     - Среднее значение
     - Медиана
     - Дисперсия
     - Квантиль 0.1 и 0.9
     - Квартиль 1 и 3
   - Для категориальных столбцов
     - Доля пропусков
     - Количество уникальных значений
     - Мода
3. Сформулировать и проверить минимум 2 статистические гипотезы. Выбор критериев обосновать. Сделать выводы в терминах предметной области.
4. Категориальные данные необходимые для задачи, закодировать методами One-HotEncoding или LabelEncoding, в случае если это требуется.
5. Построить таблицу корреляции признаков и целевого столбца.
6. Реализовать стохастический и обычный градиентный спуск.

Выше описанное требуется реализовать для датасета seaborn.mpg и датасета данного по варианту.

Сложность “Well-done”

2. Вариант задачи

Binary Classification with a Tabular Pulsar Dataset [1]

3. Ссылка на код

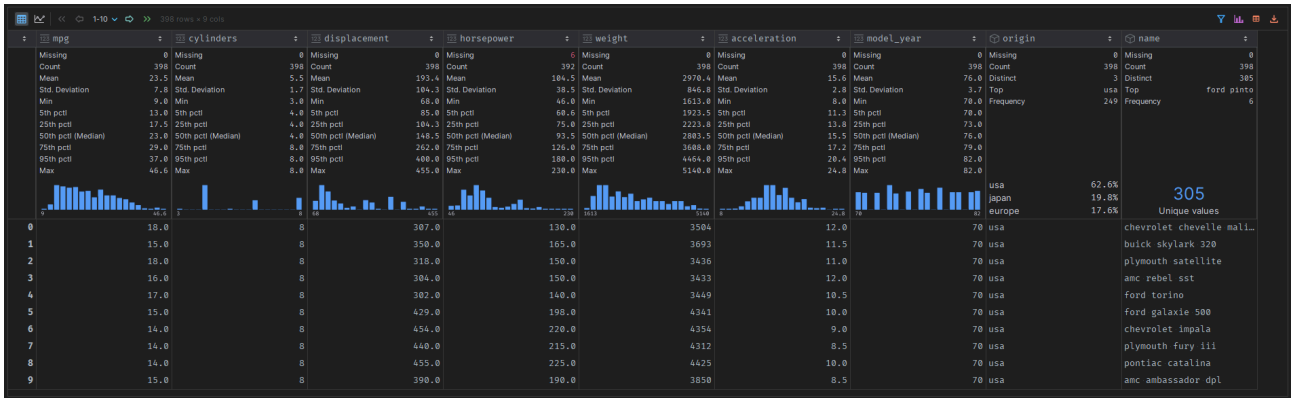
Дамп таблиц не будет предоставлен, так как БД содержит слишком много данных для системы elearning. Взамен будет предоставлен DDL.

https://github.com/NydusBorn/big-data

4. Описание проделанной работы

Значительная часть требуемых данных автоматически получена штатными средствами pycharm

4.1. Датасет seaborn.mpg



	mpg	cylinders	displacement	horsepower	weight	acceleration	model_year	origin	name
Missing	0	0	0	6	0	0	0	0	0
Count	398	398	398	392	398	398	398	398	398
Mean	23.5	5.5	193.4	104.5	2970.4	15.6	76.0	3	305
Std. Deviation	7.8	1.7	104.3	38.5	846.8	2.8	3.7	Top	usa	ford pinto
Min	9.0	3.0	68.0	46.0	1613.0	8.0	70.0	Frequency	249	6
5th pctl	13.0	4.0	85.0	60.6	1923.5	11.3	70.0			
25th pctl	17.5	4.0	104.3	75.0	2223.8	13.8	73.0			
50th pctl (Median)	23.0	4.0	148.5	93.5	2803.5	15.5	76.0			
75th pctl	29.0	8.0	262.0	126.0	3608.0	17.2	79.0			
95th pctl	37.0	8.0	400.0	180.0	4464.0	20.4	82.0			
Max	46.6	8.0	455.0	230.0	5140.0	24.8	82.0			
								usa	62.6%	
								japan	19.8%	
								europe	17.6%	
									305 Unique values	
0	18.0	8	307.0	130.0	3504	12.0	70	usa	chevrolet chevelle mali…
1	15.0	8	350.0	165.0	3693	11.5	70	usa	buick skylark 320
2	18.0	8	318.0	150.0	3436	11.0	70	usa	plymouth satellite
3	16.0	8	304.0	150.0	3433	12.0	70	usa	amc rebel sst
4	17.0	8	302.0	140.0	3449	10.5	70	usa	ford torino
5	15.0	8	429.0	198.0	4341	10.0	70	usa	ford galaxie 500
6	14.0	8	454.0	220.0	4354	9.0	70	usa	chevrolet impala
7	14.0	8	440.0	215.0	4312	8.5	70	usa	plymouth fury iii
8	14.0	8	455.0	225.0	4425	10.0	70	usa	pontiac catalina
9	15.0	8	390.0	190.0	3850	8.5	70	usa	amc ambassador dpl

Figure 1: Предварительные данные

Соответственно здесь можно обнаружить следующие параметры:

- Количество строк и столбцов: 398 строк и 9 столбцов
- Количество пропусков
- Среднее для числовых
- Мин/Макс для числовых
- Медиана для числовых
- Количество уникальных значений и мода для категориальных

Остальные данные (дисперсия, квантили и квартили посчитаем сами)

4.1.1. Данные по столбцам

1. mpg
    - Пропуски: 0
    - Мин: 9
    - Макс: 46.6

- Среднее: 23.5
- Медиана: 23
- Дисперсия: 61.1
- Квантиль 0.1: 14
- Квантиль 0.9: 34.3
- Квартиль 1: 17.5
- Квартиль 3: 29

2. cylinders
   - Пропуски: 0
   - Мин: 3
   - Макс: 8
   - Среднее: 5.5
   - Медиана: 4
   - Дисперсия: 2.9
   - Квантиль 0.1: 4
   - Квантиль 0.9: 8
   - Квартиль 1: 4
   - Квартиль 3: 8
3. displacement
   - Пропуски: 0
   - Мин: 68
   - Макс: 455
   - Среднее: 193.4
   - Медиана: 148.5
   - Дисперсия: 10872.2
   - Квантиль 0.1: 90
   - Квантиль 0.9: 350

- Квартиль 1: 104.25
- Квартиль 3: 262

4. horsepower
   - Пропуски: 6 (1.5%)
   - Мин: 46
   - Макс: 230
   - Среднее: 104.5
   - Медиана: 93.5
   - Дисперсия: 1481.5
   - Квантиль 0.1: 67
   - Квантиль 0.9: 157.7
   - Квартиль 1: 75
   - Квартиль 3: 126
5. weight
   - Пропуски: 0
   - Мин: 1613
   - Макс: 5140
   - Среднее: 2970.4
   - Медиана: 2803.5
   - Дисперсия: 717141
   - Квантиль 0.1: 1988.5
   - Квантиль 0.9: 4275.2
   - Квартиль 1: 2223.75
   - Квартиль 3: 3608
6. acceleration
   - Пропуски: 0
   - Мин: 8

- Макс: 24.8
- Среднее: 15.6
- Медиана: 15.5
- Дисперсия: 7.6
- Квантиль 0.1: 12
- Квантиль 0.9: 19
- Квартиль 1: 13.8
- Квартиль 3: 17.2

7. model_year
    - Пропуски: 0
    - Мин: 70
    - Макс: 82
    - Среднее: 76
    - Медиана: 76
    - Дисперсия: 13.6
    - Квантиль 0.1: 71
    - Квантиль 0.9: 81
    - Квартиль 1: 73
    - Квартиль 3: 79
8. origin
    - Пропуски: 0
    - Уникальных: 3
    - Мода: usa
8. Name
    - Пропуски: 0
    - Уникальных: 305
    - Мода: ford pinto

4.1.2. Гипотезы

Основным интересным для нас параметром предположим количество лошадиных сил, соответственно остальные данные будем считать дополнительными.

1. Предположим нулевую гипотезу: Год выпуска не имеет влияния на количество лошадиных сил.

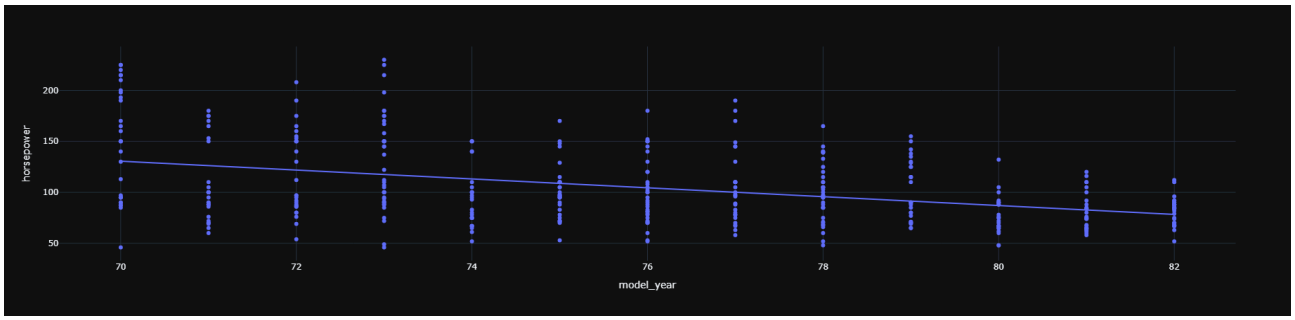


Figure 2: x - год выпуска, y - количество лошадиных сил

Как мы видим наша нулевая гипотеза не подтвердилась, можно утверждать что с ростом года выпуска количество лошадиных сил уменьшается.

2. Предположим альтернативную гипотезу: Количестов цилиндров имеет положительное влияние на количество лошадиных сил.

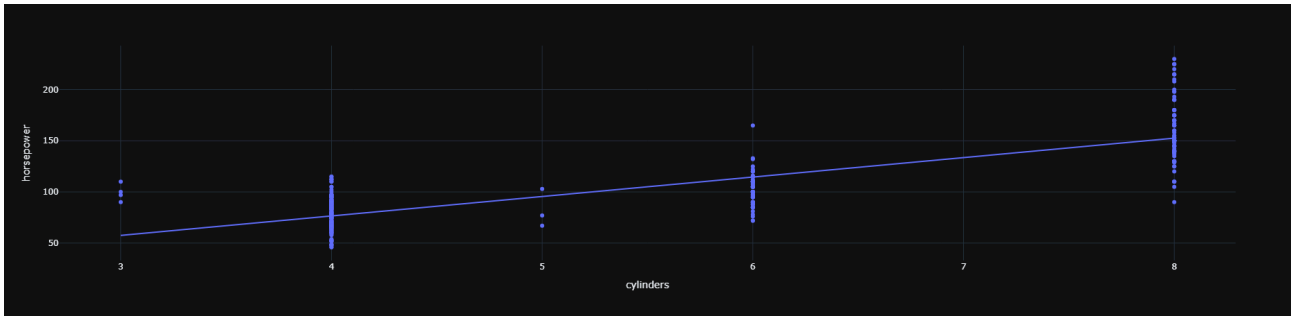


Figure 3: x - количество цилиндров, y - количество лошадиных сил

Как мы видим наша гипотеза подтвердилась.

4.1.3. Корреляции

Для этого дополнительно закодируем страны в OneHotEncoding.

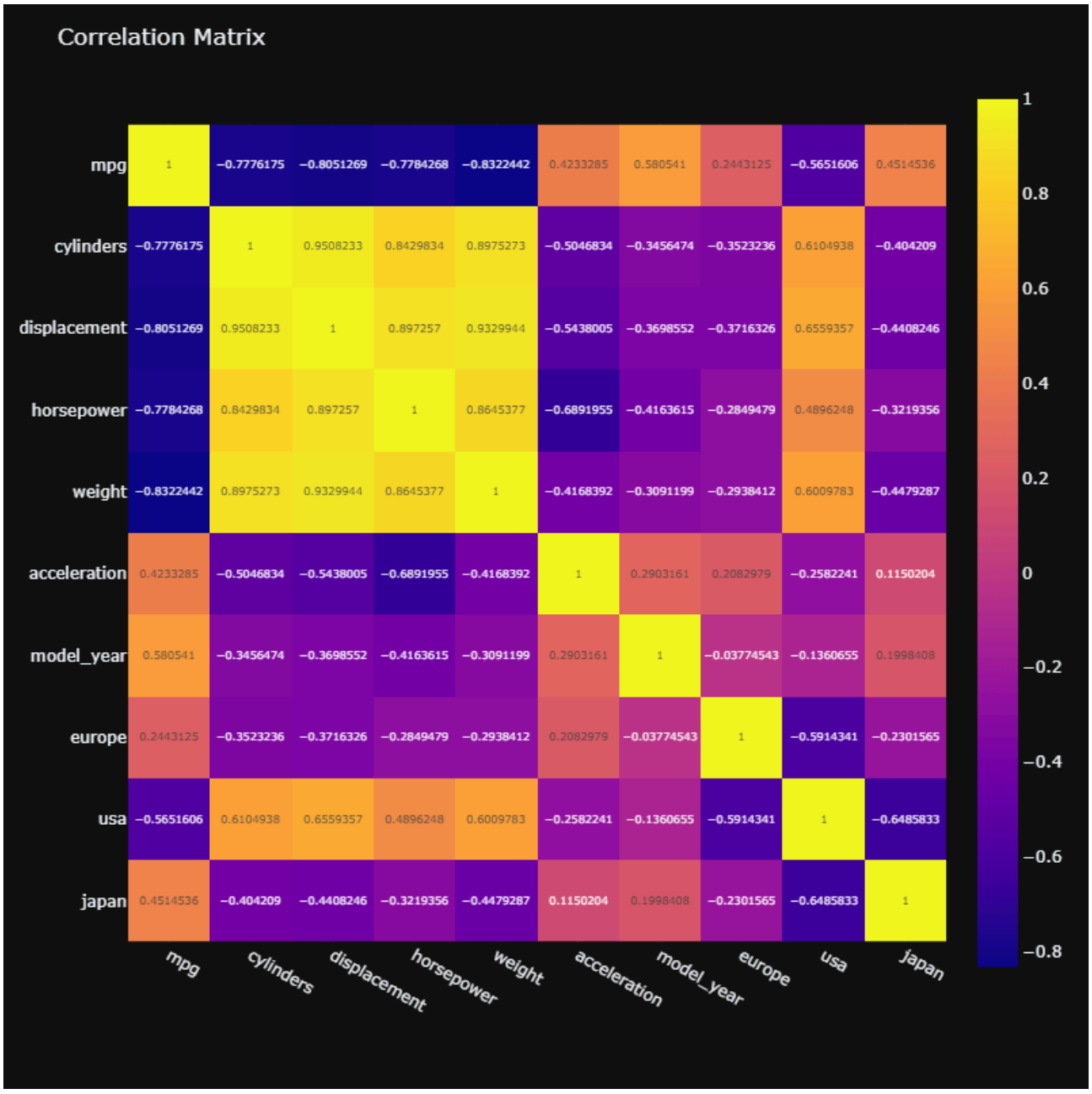


Figure 4: Корреляционная матрица

4.1.4. Стохастический и градиентный спуски

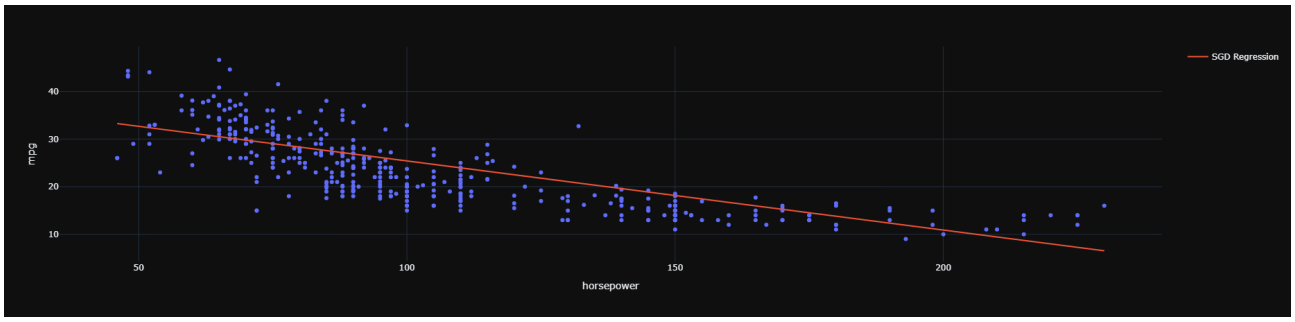


Figure 5: Стохастический градиентный спуск

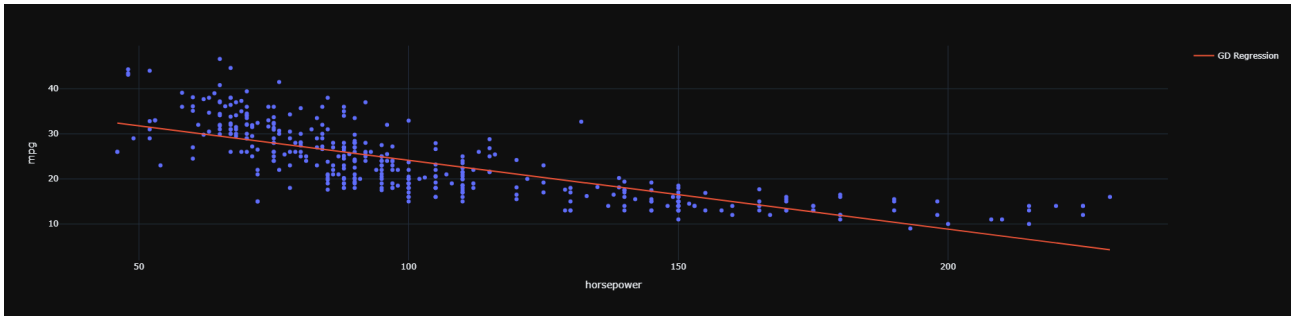
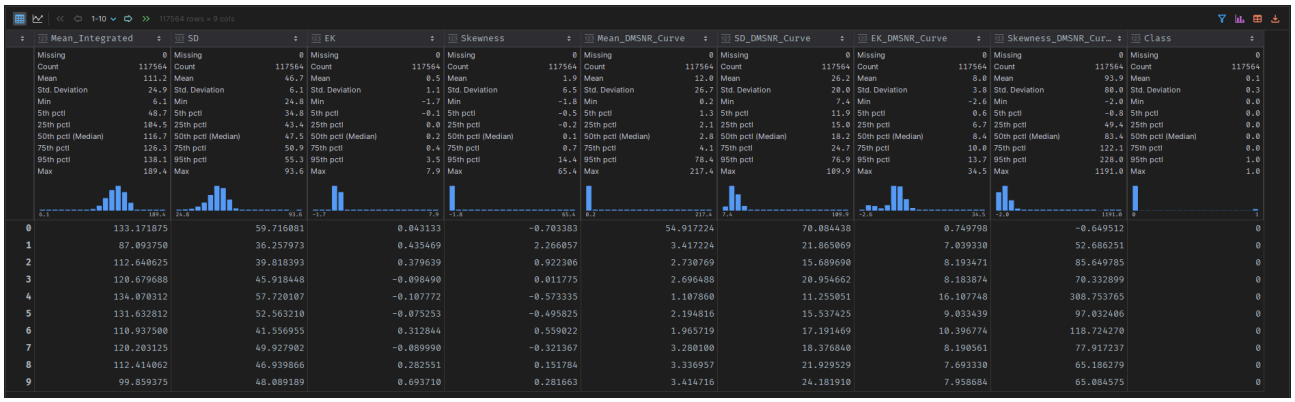


Figure 6: Обычный градиентный спуск

4.1.5. Вывод

Мы провели статистический анализ датасета seaborn.mpg. Как мы выяснили количество лошадиных сил выше в США, в моделях выпущенных раньше, при этом количество цилиндров увеличивается и объем двигателя растет, также есть положительная корреляция с весом (более тяжелые автомобили имеют больше лошадиных сил), но автомобили с высоким количеством лошадиных сил имеют низкое ускорение.

4.2. Датасет Pulsar



	Mean_Integrated	SD	EK	Skewness	Mean_DMSNR_Curve	SD_DMSNR_Curve	EK_DMSNR_Curve	Skewness_DMSNR_Cur…	Class
Missing	0	0	0	0	0	0	0	0	0
Count	117564	117564	117564	117564	117564	117564	117564	117564	117564
Mean	111.2	46.7	0.5	1.9	12.0	26.2	8.0	93.9	0.1
Std. Deviation	24.9	6.1	1.1	6.5	26.7	20.0	3.8	80.0	0.3
Min	6.1	24.8	-1.7	-1.8	0.2	7.4	-2.6	-2.0	0.0
5th pctl	48.7	34.8	-0.1	-0.5	1.3	11.9	0.6	-0.8	0.0
25th pctl	104.5	43.4	0.0	-0.2	2.1	15.0	6.7	49.4	0.0
50th pctl (Median)	116.7	47.5	0.2	0.1	2.8	18.2	8.4	83.4	0.0
75th pctl	126.3	50.9	0.4	0.7	4.1	24.7	10.0	122.1	0.0
95th pctl	138.1	55.3	3.5	14.4	78.4	76.9	13.7	228.0	1.0
Max	189.4	93.6	7.9	65.4	217.4	189.9	34.5	1191.0	1.0
0	133.171875	59.716081	0.043133	-0.703383	54.917224	70.084438	0.749798	-0.649512	0
1	87.093750	36.257973	0.435469	2.266057	3.417224	21.865069	7.039330	52.686251	0
2	112.640625	39.818393	0.379639	0.922306	2.730769	15.689690	8.193471	85.649785	0
3	120.679688	45.918448	-0.098490	0.011775	2.696488	20.954662	8.183874	70.332899	0
4	134.070312	57.720107	-0.107772	-0.573335	1.107860	11.255051	16.107748	308.753765	0
5	131.632812	52.563210	-0.075253	-0.495825	2.194816	15.537425	9.033439	97.032406	0
6	110.937500	41.556955	0.312844	0.559022	1.965719	17.191469	10.396774	118.724270	0
7	120.203125	49.927902	-0.089990	-0.321367	3.280100	18.376840	8.190561	77.917237	0
8	112.414062	46.939866	0.282551	0.151784	3.336957	21.929529	7.693330	65.186279	0
9	99.859375	48.089189	0.693710	0.281663	3.414716	24.181910	7.958684	65.084575	0

Figure 7: Предварительные данные

Количество строк и столбцов: 117564 строк и 9 столбцов

4.2.1. Данные по столбцам

1. Mean_Integrated
   - Пропуски: 0
   - Мин: 6.1
   - Макс: 189.4
   - Среднее: 111.2
   - Медиана: 116.7
   - Дисперсия: 620.3
   - Квантиль 0.1: 85.6
   - Квантиль 0.9: 134.2
   - Квартиль 1: 104.5
   - Квартиль 3: 126.3
2. SD
   - Пропуски: 0
   - Мин: 24.8
   - Макс: 93.6

- Среднее: 46.7
- Медиана: 47.5
- Дисперсия: 37.2
- Квантиль 0.1: 38.2
- Квантиль 0.9: 53.6
- Квартиль 1: 43.4
- Квартиль 3: 50.8

3. EK
   - Пропуски: 0
   - Мин: −1.7
   - Макс: 7.9
   - Среднее: 0.5
   - Медиана: 0.2
   - Дисперсия: 1.27
   - Квантиль 0.1: −0.07
   - Квантиль 0.9: 0.79
   - Квартиль 1: 0.04
   - Квартиль 3: 0.39
4. Skewness
   - Пропуски: 0
   - Мин: −1.8
   - Макс: 65.4
   - Среднее: 1.9
   - Медиана: 0.1
   - Дисперсия: 42.5
   - Квантиль 0.1: −0.39
   - Квантиль 0.9: 2.47

- Квартиль 1: −0.18
- Квартиль 3: 0.69

5. Mean_DMSNR_Curve
   - Пропуски: 0
   - Мин: 0.2
   - Макс: 217.4
   - Среднее: 12
   - Медиана: 2.8
   - Дисперсия: 713.9
   - Квантиль 0.1: 1.62
   - Квантиль 0.9: 34.1
   - Квартиль 1: 2.09
   - Квартиль 3: 4.12
6. SD_DMSNR_Curve
   - Пропуски: 0
   - Мин: 7.4
   - Макс: 109.9
   - Среднее: 26.2
   - Медиана: 18.2
   - Дисперсия: 401.6
   - Квантиль 0.1: 12.9
   - Квантиль 0.9: 62.7
   - Квартиль 1: 14.9
   - Квартиль 3: 24.7
7. EK_DMSNR_Curve
   - Пропуски: 0
   - Мин: −2.6

- Макс: 34.5
- Среднее: 8
- Медиана: 8.4
- Дисперсия: 14.7
- Квантиль 0.1: 1.82
- Квантиль 0.9: 11.8
- Квартиль 1: 6.74
- Квартиль 3: 10.0

8. Skewness_DMSNR_Curve
    - Пропуски: 0
    - Мин: −2
    - Макс: 1191
    - Среднее: 93.9
    - Медиана: 83.4
    - Дисперсия: 6393.9
    - Квантиль 0.1: 2.24
    - Квантиль 0.9: 174.3
    - Квартиль 1: 49.4
    - Квартиль 3: 122
9. Class
    - Пропуски: 0
    - Мин: 0
    - Макс: 1
    - Среднее: 0.1
    - Медиана: 0
    - Дисперсия: 0.08
    - Квантиль 0.1: 0

- Квантиль 0.9: 0
- Квартиль 1: 0
- Квартиль 3: 0

Технически это категориальный или бинарный признак, с 2 уникальными значениями и модой 0

4.2.2. Гипотезы

Основным интересным для нас параметром предположим Skewness (наклон пульсара), соответственно остальные данные будем считать дополнительными.

1. Предположим нулевую гипотезу: Mean_Integrated не имеет влияния на Skewness.

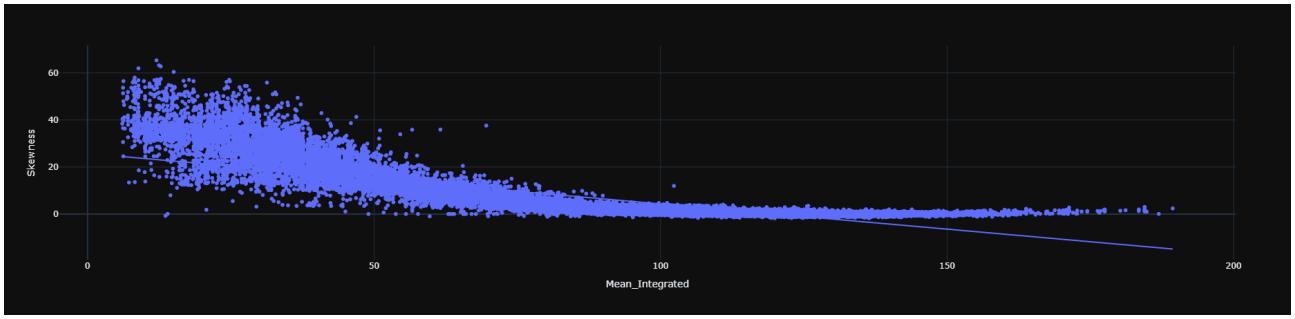


Figure 8: x - Mean Integrated, y - Skewness

Как мы видим наша нулевая гипотеза не подтвердилась, можно утверждать что с ростом Mean Integrated Skewness уменьшается.

2. Предположим альтернативную гипотезу: SD имеет негативное влияние на Skewness.

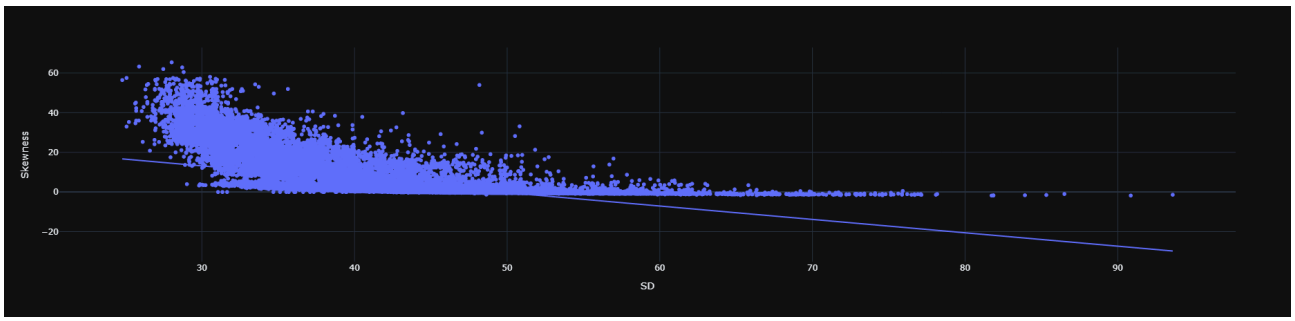


Figure 9: x - SD, y - Skewness

Как мы видим наша гипотеза подтвердилась.

4.2.3. Корреляции

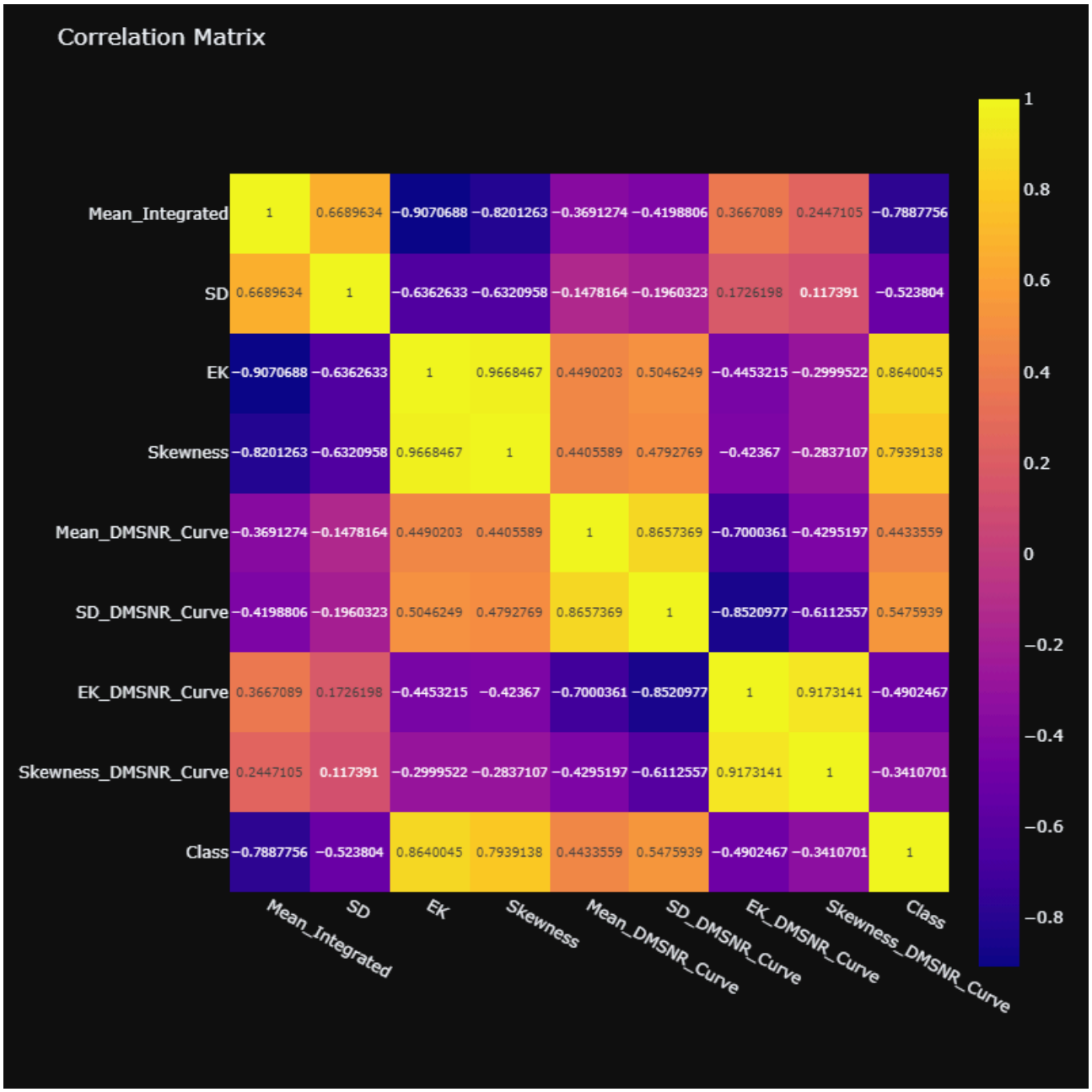


Figure 10: Корреляционная матрица

4.2.4. Стохастический и градиентный спуски

Проведем на основании y - Skewness и x - SD

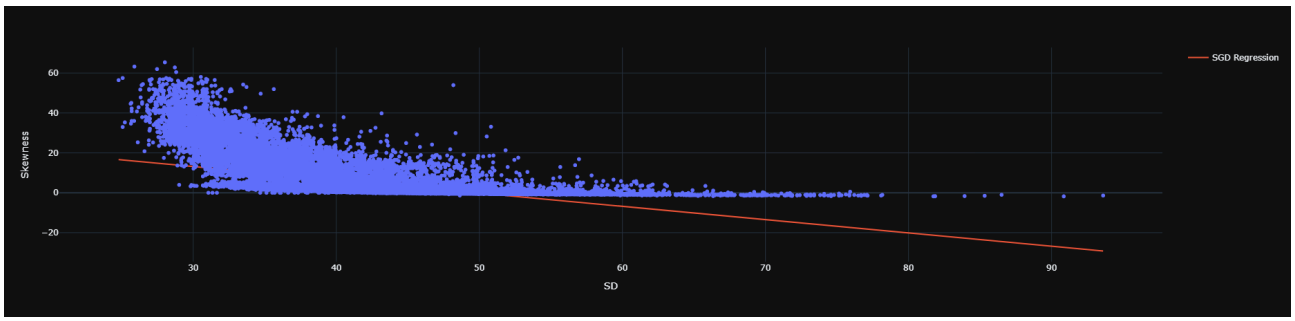


Figure 11: Стохастический градиентный спуск

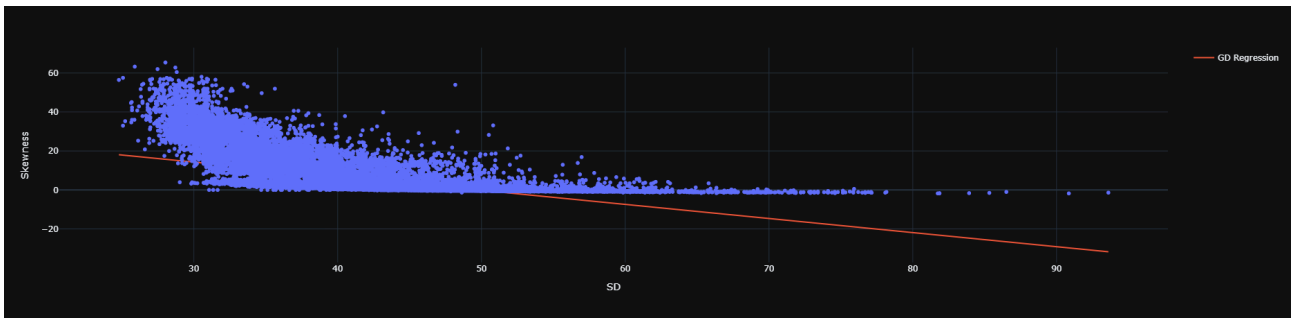


Figure 12: Обычный градиентный спуск

4.2.5. Вывод

Мы проанализироваои датасет Pulsar. Как мы увидили все параметры так или иначе связаны с Skewness, кроме Skewness_DMSNR_Curve, тк на нём довольно низкий уровень корреляции.

5. Вывод

Мы использовали средства python и IDE pycharm для анализа датасетов seaborn.mpg и Pulsar.

Список Литературы

[1] “Binary Classification with a Tabular Pulsar Dataset.” Accessed: Jan. 27, 2025. [Online]. Available: https://www.kaggle.com/competitions/playground-series-s3e10/code