

Итоговый проект. НИС "Анализ данных в Python" 23/24

Выполнили студенты:

Фролов-Буканов Виктор Дмитриевич, БПИ228

Глебов Павел Алексеевич, БПИ228

Датасет

<https://www.kaggle.com/datasets/sooyoungher/smoking-drinking-dataset>

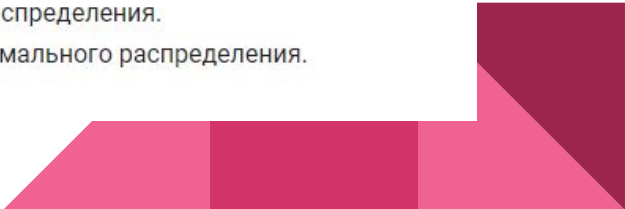
	sex	age	height	weight	waistline	sight_left	sight_right	hear_left	hear_right	SBP	DBP	BLDS	tot_chole	hemoglobin	urine_protein	SMK_stat_type_cd	DRK_YN
165053	Male	60	160	70	93.0	0.9	0.8	1.0	1.0	110.0	70.0	57.0	235.0	16.4	1.0	2.0	Y
277992	Female	55	150	60	85.0	0.8	0.8	1.0	1.0	119.0	78.0	95.0	267.0	14.4	1.0	1.0	N
848932	Male	30	170	65	71.0	1.5	0.1	1.0	1.0	121.0	91.0	88.0	194.0	14.7	1.0	1.0	N
913698	Female	75	140	45	79.8	0.4	0.1	1.0	2.0	154.0	85.0	94.0	162.0	12.9	1.0	1.0	N
359322	Male	60	165	70	85.0	1.0	1.0	1.0	1.0	126.0	76.0	120.0	198.0	15.9	1.0	2.0	Y
639461	Male	40	175	75	89.2	1.5	1.5	1.0	1.0	137.0	87.0	90.0	235.0	16.4	1.0	3.0	Y
960320	Male	35	175	60	78.0	0.7	0.9	1.0	1.0	120.0	73.0	84.0	190.0	14.4	1.0	3.0	Y
436351	Male	45	170	70	83.1	1.0	1.0	1.0	1.0	136.0	85.0	84.0	238.0	15.6	1.0	1.0	Y
298435	Male	45	170	60	77.0	1.2	1.5	1.0	1.0	123.0	83.0	175.0	110.0	15.4	1.0	2.0	Y
248010	Female	70	150	60	85.0	0.6	0.8	1.0	1.0	148.0	90.0	149.0	179.0	13.8	1.0	1.0	N

Цель и задачи исследования

Целью данного исследования является выявление взаимосвязей и корреляций между статистическими показателями физического здоровья человека и фактом того, курит человек или пьёт (или и то, и то)

Исследовательские гипотезы

В ходе анализа данных мы проверим следующие 10 исследовательских гипотез:

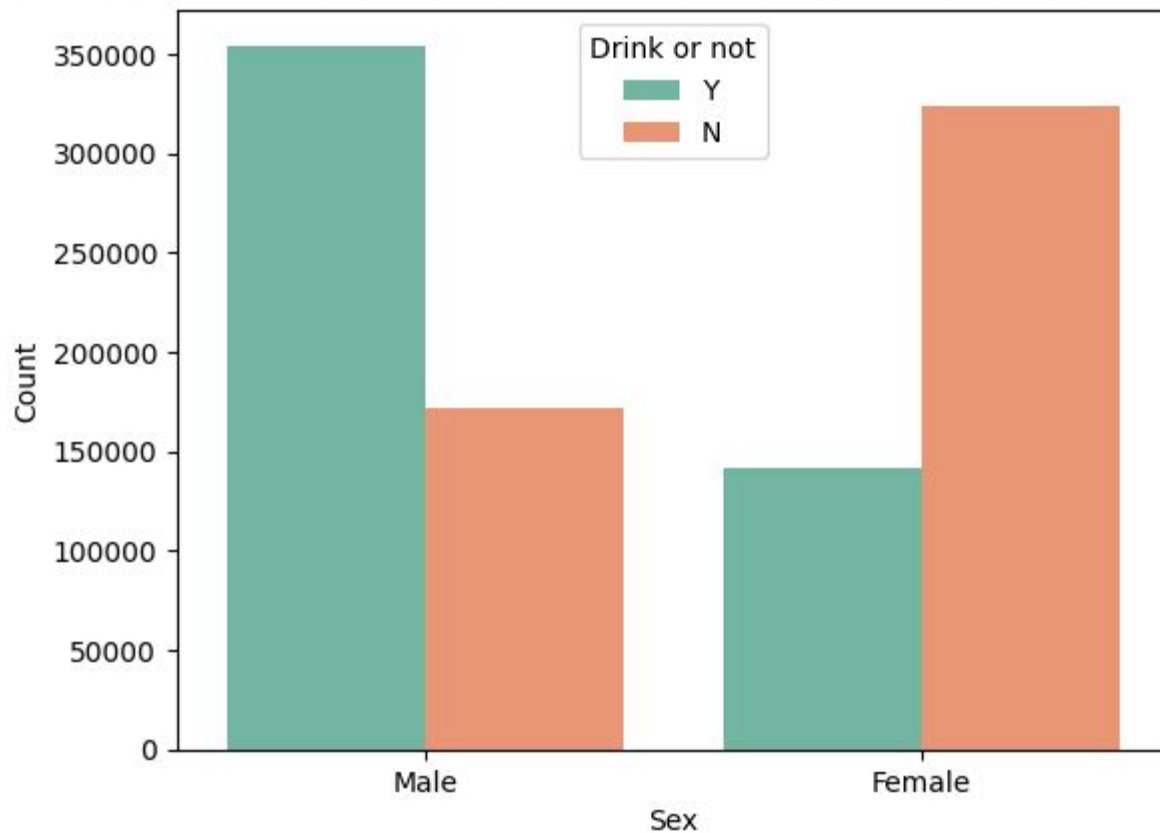
1. H0: распределение роста наблюдаемых не значимо отличается от нормального распределения.
H1: распределение роста наблюдаемых статистически значимо отличается от нормального распределения. Исследуемая переменная: рост (**height**)
 2. H0: распределение веса наблюдаемых не значимо отличается от нормального распределения.
H1: распределение веса наблюдаемых статистически значимо отличается от нормального распределения. Исследуемая переменная: вес (**weight**)
 3. H0: распределение возраста наблюдаемых не значимо отличается от нормального распределения.
H1: распределение возраста наблюдаемых статистически значимо отличается от нормального распределения. Исследуемая переменная: возраст (**age**)
 4. H0: распределение линии талии наблюдаемых не значимо отличается от нормального распределения.
H1: распределение линии талии наблюдаемых статистически значимо отличается от нормального распределения.
Исследуемая переменная: линия талии (**waistline**)
- 

5. H0: отсутствует статистически значимая взаимосвязь между фактом того, пьющий наблюдаемый или нет, и уровнем слуха наблюдаемого (на левом ухе)
H1: существует статистически значимая взаимосвязь между фактом того, пьющий наблюдаемый или нет, и уровнем слуха наблюдаемого (на левом ухе)
Исследуемые переменные: **DRK_YN** и **hear_left**
6. H0: отсутствует статистически значимая взаимосвязь между степенью курения наблюдаемого и слухом наблюдаемого (на левом ухе)
H1: существует статистически значимая взаимосвязь между степенью курения наблюдаемого и слухом наблюдаемого (на левом ухе)
Исследуемые переменные: **SMK_stat_type_cd** и **hear_left**
7. H0: отсутствует статистически значимая взаимосвязь между уровнем слуха наблюдаемого на левом ухе и уровнем слуха наблюдаемого на правом ухе
H1: существует статистически значимая взаимосвязь между уровнем слуха наблюдаемого на левом ухе и уровнем слуха наблюдаемого на правом ухе
Исследуемые переменные: **hear_right** и **hear_left**
8. H0: отсутствует статистически значимая взаимосвязь между полом наблюдаемого и фактом того, пьющий наблюдаемый или нет
H1: существует статистически значимая взаимосвязь между полом наблюдаемого и фактом того, пьющий наблюдаемый или нет
Исследуемые переменные: **sex** и **DRK_YN**
9. H0: отсутствует статистически значимая взаимосвязь между полом наблюдаемого и степенью курения наблюдаемого
H1: существует статистически значимая взаимосвязь между полом наблюдаемого и степенью курения наблюдаемого
Исследуемые переменные: **sex** и **SMK_stat_type_cd**
10. H0: отсутствует статистически значимая взаимосвязь между фактом того, пьющий наблюдаемый или нет и степенью курения наблюдаемого
H1: существует статистически значимая взаимосвязь между фактом того, пьющий наблюдаемый или нет и степенью курения наблюдаемого
Исследуемые переменные: **DRK_YN** и **SMK_stat_type_cd**

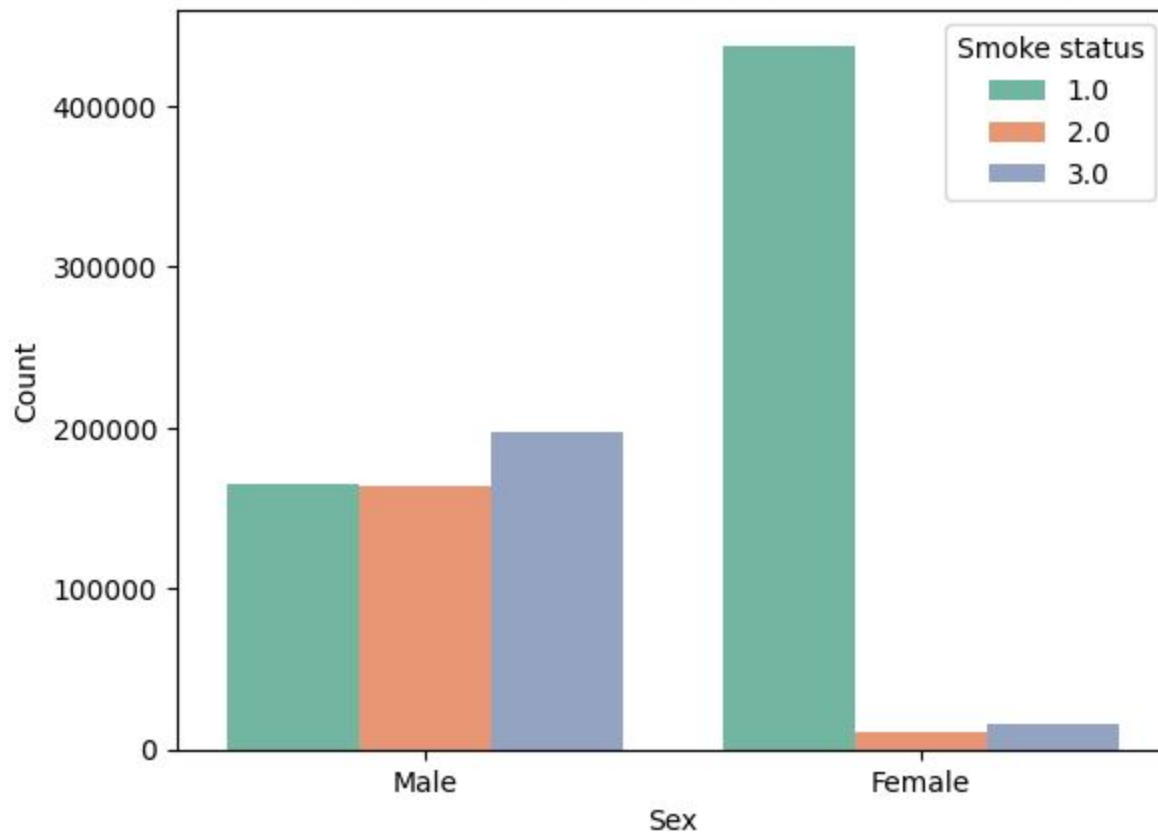
Были рассчитан индекс массы тела и добавлена категориальная переменная, к какой возрастной группе относится наблюдаемый.

								DRK_YN	BMI	age_group			
								N	0.002286	adult			
								N	0.002081	pensioner			
								Y	0.002539	adult			
								N	0.002344	pensioner			
								N	0.002204	adult			
sex	age	height	weight	waistline	sight_left	sight_right	h	Y	0.002249	adult			
852186	Male	35	175	70	84.0	1.5	1.5	Y	0.002249	adult			
205826	Female	60	155	50	74.0	1.0	0.5	Y	0.003330	adult			
676313	Male	55	160	65	80.2	0.8	0.8	Y	0.002000	adult			
219296	Female	60	160	60	77.0	0.8	0.8	Y	0.002000	adult			
5068	Male	45	165	60	81.0	1.5	1.5	Y	0.002286	adult			
144865	Male	50	170	65	76.0	0.7	1.0	Y	0.002286	adult			
355547	Female	25	155	80	89.0	1.5	1.2	N	0.002148	adult			
508256	Female	40	150	45	63.0	0.9	0.5						
337328	Male	35	175	70	80.0	0.9	1.0	1.0	1.0	119.0	72.0	82.0	158.0
492851	Female	55	160	55	69.0	1.0	1.0	1.0	1.0	133.0	84.0	92.0	228.0
								hemoglobin	urine_protein	SMK_stat_type_cd	DRK_YN	BMI	age_group
								15.9	1.0	3.0	N	0.002286	adult
								13.2	1.0	1.0	N	0.002081	pensioner
								14.6	1.0	2.0	Y	0.002539	adult
								13.1	1.0	1.0	N	0.002344	pensioner
								12.9	1.0	3.0	N	0.002204	adult
								14.6	1.0	2.0	Y	0.002249	adult
								15.4	1.0	1.0	Y	0.003330	adult
								13.0	1.0	1.0	Y	0.002000	adult
								14.2	1.0	3.0	Y	0.002286	adult
								15.0	1.0	1.0	N	0.002148	adult

Distribution of drinkers and non-drinkers over sexes

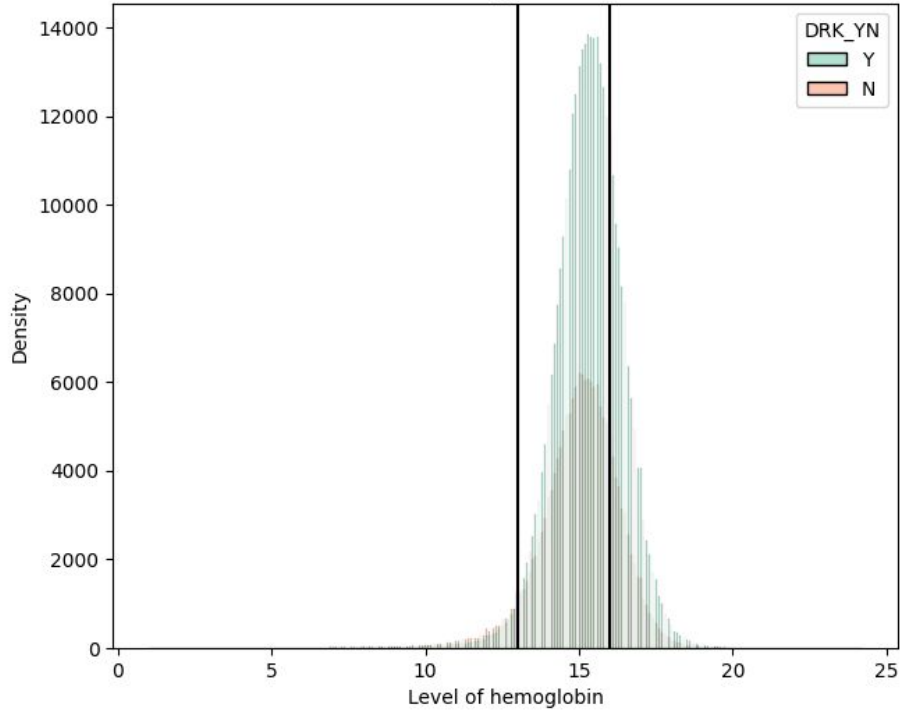


Distribution of smoke statuses over sexes

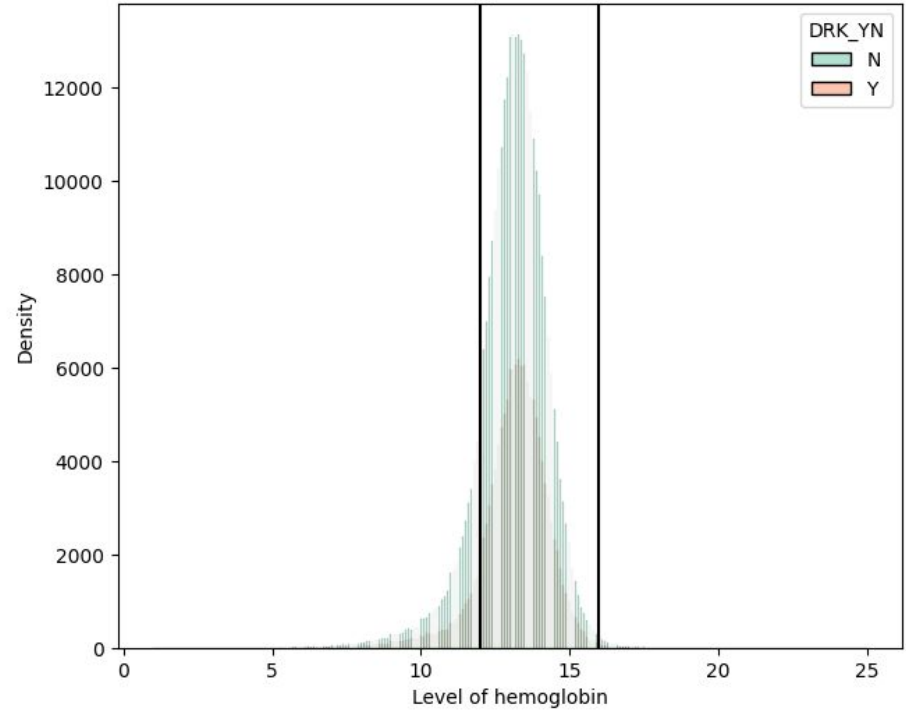


Level of hemoglobin over sexes

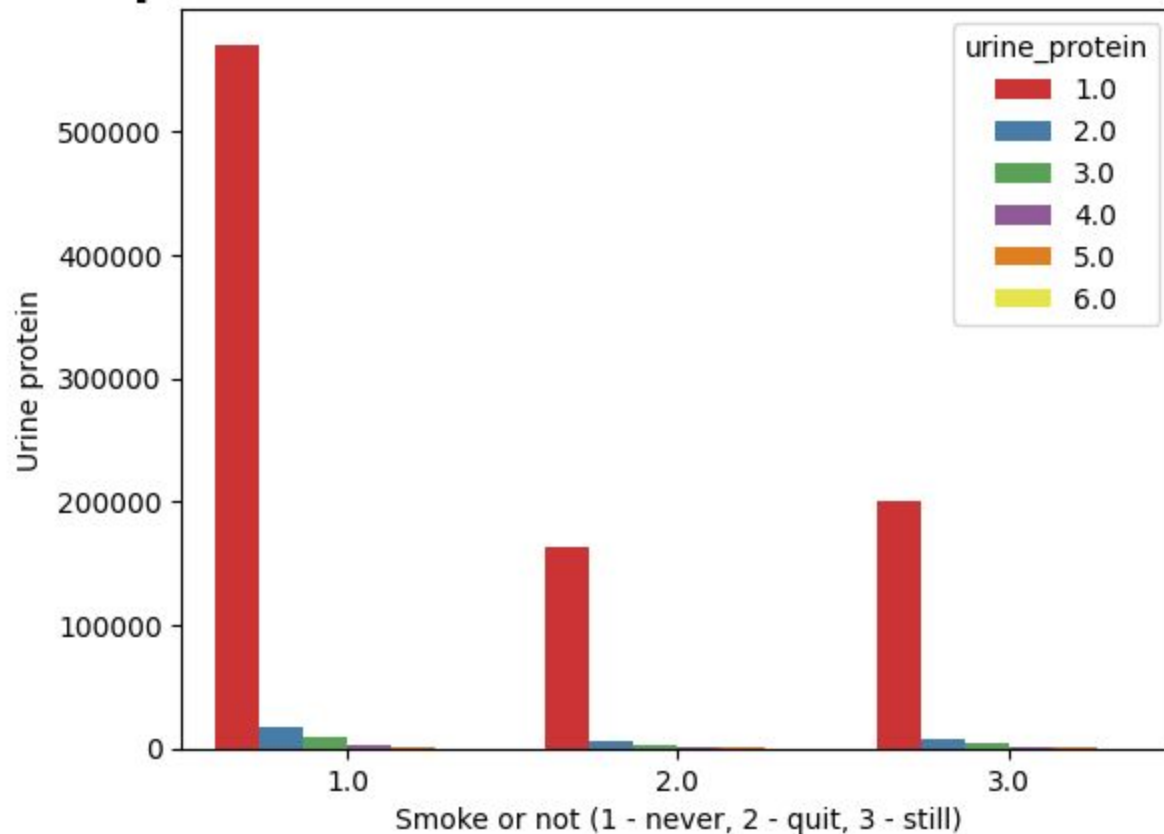
Level of hemoglobin over men



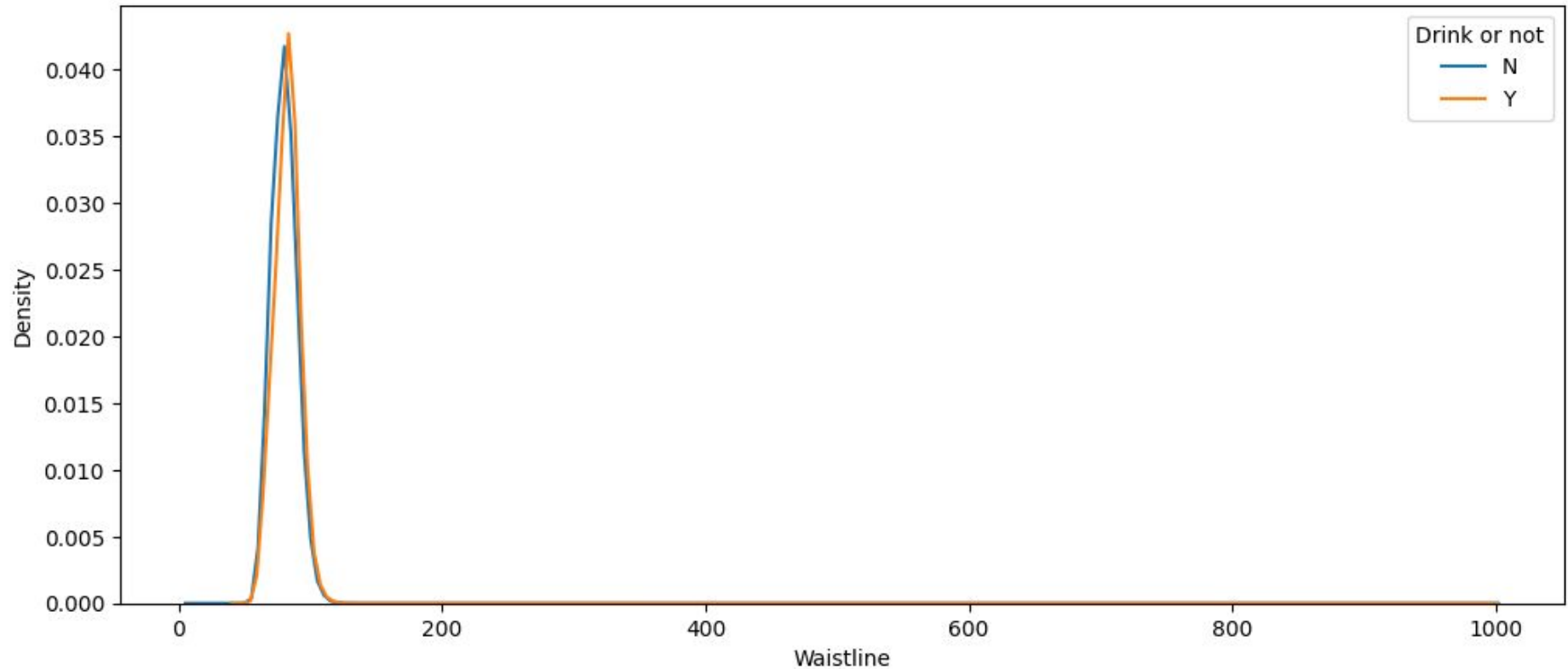
Level of hemoglobin over women



Urine protein level indicators over smoke statuses



Distribution of waistlines over drinkers and non-drinkers



5. H0: отсутствует статистически значимая взаимосвязь между фактом того, пьющий наблюдаемый или нет, и уровнем слуха наблюдаемого (на левом ухе)

H1: существует статистически значимая взаимосвязь между фактом того, пьющий наблюдаемый или нет, и уровнем слуха наблюдаемого (на левом ухе)

Исследуемые переменные: **DRK_YN** и **hear_left**

```
[ ] crosstab = pd.crosstab(df['DRK_YN'], df['hear_left'])  
crosstab
```

hear_left	1.0	2.0
DRK_YN		
N	475191	20667
Y	484933	10555

Все ячейки содержат ожидаемую частоту более 5, так что тест хи-квадрат применим, но интересным наблюдением является то, что отклонение от нормы слуха больше у **непьющих** людей

```
[ ] from scipy.stats import chi2_contingency as chi2  
chi2(crosstab)
```

```
Chi2ContingencyResult(statistic=3373.058424715928, pvalue=0.0, dof=1, expected_freq=array([[480241.17350753, 15616.82649247],  
[479882.82649247, 15605.17350753]]))
```

Вывод: гипотеза H1 принимается на уровне значимости 5% ($pvalue = 0 < 0.05 \Rightarrow$) существует взаимосвязь между фактом того, пьющий человек или нет и его уровнем слуха на левом ухе

6. H0: отсутствует статистически значимая взаимосвязь между степенью курения наблюдаемого и слухом наблюдаемого (на левом ухе)

H1: существует статистически значимая взаимосвязь между степенью курения наблюдаемого и слухом наблюдаемого (на левом ухе)

Исследуемые переменные: **SMK_stat_type_cd** и **hear_left**

```
[ ] crosstab = pd.crosstab(df['SMK_stat_type_cd'], df['hear_left'])  
crosstab
```

	hear_left	1.0	2.0
SMK_stat_type_cd			
1.0	582587	19854	
2.0	168191	6760	
3.0	209346	4608	

Все ячейки содержат ожидаемую частоту более 5, так что тест хи-квадрат применим, но вновь интересным наблюдением является то, что отклонение от нормы слуха больше у людей, которые никогда не курили

```
[ ] chi2(crosstab)
```

```
Chi2ContingencyResult(statistic=1030.41227199, pvalue=1.7734465128176337e-224, dof=2, expected_freq=array([[583467.38947249, 18973.61052751],  
[169440.99630603, 5510.00369397],  
[207215.61422147, 6738.38577853]]))
```

Вывод: гипотеза H1 принимается на уровне значимости 5% ($pvalue = 1.77 * 10^{-224} < 0.05 \Rightarrow$) существует взаимосвязь между статусом курения человека и его уровнем слуха на левом ухе

7. H0: отсутствует статистически значимая взаимосвязь между уровнем слуха наблюдаемого на левом ухе и уровнем слуха наблюдаемого на правом ухе

H1: существует статистически значимая взаимосвязь между уровнем слуха наблюдаемого на левом ухе и уровнем слуха наблюдаемого на правом ухе
Исследуемые переменные: **hear_right** и **hear_left**

```
[ ] crosstab = pd.crosstab(df['hear_left'], df['hear_right'])  
crosstab
```

hear_right	1.0	2.0
hear_left		
1.0	946842	13282
2.0	14292	16930

Все ячейки содержат ожидаемую частоту более 5, так что тест хи-квадрат применим. Сводная таблица показывает, что почти все наблюдаемые имеют хороший слух на обоих ушах, отклонение хотя бы на одном ухе является редкостью в рамках данного датасета

```
[ ] chi2(crosstab)
```

```
Chi2ContingencyResult(statistic=285738.27316231653, pvalue=0.0, dof=1, expected_freq=array([[930863.5134615, 29260.4865385],  
[ 30270.4865385, 951.5134615]]))
```

Вывод: гипотеза H1 принимается на уровне значимости 5% ($pvalue = 0 < 0.05 \Rightarrow$) существует взаимосвязь между уровнем слуха человека на левом и правом ухе

8. H0: отсутствует статистически значимая взаимосвязь между полом наблюдаемого и фактом того, пьющий наблюдаемый или нет

H1: существует статистически значимая взаимосвязь между полом наблюдаемого и фактом того, пьющий наблюдаемый или нет
Исследуемые переменные: **sex** и **DRK_YN**

```
[ ] crosstab = pd.crosstab(df['sex'], df['DRK_YN'])  
crosstab
```

DRK_YN	N	Y
sex		
Female	323760	141171
Male	172098	354317

Все ячейки содержат ожидаемую частоту более 5, так что тест хи-квадрат применим. Сводная таблица показывает, что среди мужчин пьющих примерно в 2 раза больше, чем непьющих, в то время как среди женщин ситуация обратная - пьющих примерно в 2 раза меньше, чем непьющих

```
[ ] chi2(crosstab)
```

```
Chi2ContingencyResult(statistic=134780.52648425306, pvalue=0.0, dof=1, expected_freq=array([[232552.26308272, 232378.73691728],  
[263305.73691728, 263109.26308272]]))
```

Вывод: гипотеза H1 принимается на уровне значимости 5% ($pvalue = 0 < 0.05 \Rightarrow$) существует взаимосвязь между полом и тем, пьющий человек или нет

9. H0: отсутствует статистически значимая взаимосвязь между полом наблюдаемого и степенью курения наблюдаемого
H1: существует статистически значимая взаимосвязь между полом наблюдаемого и степенью курения наблюдаемого
Исследуемые переменные: **sex** и **SMK_stat_type_cd**

```
[ ] crosstab = pd.crosstab(df['sex'], df['SMK_stat_type_cd'])  
crosstab
```

SMK_stat_type_cd	1.0	2.0	3.0
sex			
Female	437760	10923	16248
Male	164681	164028	197706

Все ячейки содержат ожидаемую частоту более 5, так что тест хи-квадрат применим. Сводная таблица показывает, что в целом очень много некурящих людей, но по мужчинам распределение +- равномерное, в то время как по женщинам явно преобладает категория никогда не курящих

```
[ ] chi2(crosstab)
```

```
Chi2ContingencyResult(statistic=409429.412713673, pvalue=0.0, dof=2, expected_freq=array([[282538.58548983, 82050.20586253, 100342.20864764],  
[319902.41451017, 92900.79413747, 113611.79135236]]))
```

Вывод: гипотеза H1 принимается на уровне значимости 5% ($pvalue = 0 < 0.05 \Rightarrow$) существует взаимосвязь между полом и статусом курения человека

10. H0: отсутствует статистически значимая взаимосвязь между фактом того, пьющий наблюдаемый или нет и степенью курения наблюдаемого

H1: существует статистически значимая взаимосвязь между фактом того, пьющий наблюдаемый или нет и степенью курения наблюдаемого

Исследуемые переменные: **DRK_YN** и **SMK_stat_type_cd**

```
[ ] crosstab = pd.crosstab(df['DRK_YN'], df['SMK_stat_type_cd'])
crosstab
```

SMK_stat_type_cd	1.0	2.0	3.0
DRK_YN			
N	389010	54471	52377
Y	213431	120480	161577

Все ячейки содержат ожидаемую частоту более 5, так что тест хи-квадрат применим. Сводная таблица показывает, что людей, ведущих здоровый образ жизни (непьющих и никогда не курящих), очень много - 389010. При этом распределение статусов курения среди пьющих людей равномернее, чем среди непьющих

```
[ ] chi2(crosstab)
```

```
Chi2ContingencyResult(statistic=131811.45997854197, pvalue=0.0, dof=2, expected_freq=array([[301332.92450668, 87508.1484749, 107016.92701842],
[301108.07549332, 87442.8515251, 106937.07298158]]))
```

Вывод: гипотеза H1 принимается на уровне значимости 5% ($pvalue = 0 < 0.05 \Rightarrow$) существует взаимосвязь между приверженности человека к алкоголю и его статусом курения человека

Вычислим для всех количественных переменных общую статистику.

```
[ ] dfMetric = df[['age', 'height', 'weight', 'waistline', 'sight_left', 'sight_right', 'SBP', 'DBP', 'BLDS', 'tot_chole', 'hemoglobin']]
dfMetric.describe()
```

	age	height	weight	waistline	sight_left	sight_right	SBP	DBP	BLDS	tot_chole	hemoglobin
count	991346.000000	991346.000000	991346.000000	991346.000000	991346.000000	991346.000000	991346.000000	991346.000000	991346.000000	991346.000000	991346.000000
mean	47.614491	162.240625	63.284050	81.233358	0.980834	0.978429	122.432498	76.052627	100.424447	195.557020	14.229824
std	14.181339	9.282957	12.514241	11.850323	0.605949	0.604774	14.543148	9.889365	24.179960	38.660155	1.584929
min	20.000000	130.000000	25.000000	8.000000	0.100000	0.100000	67.000000	32.000000	25.000000	30.000000	1.000000
25%	35.000000	155.000000	55.000000	74.100000	0.700000	0.700000	112.000000	70.000000	88.000000	169.000000	13.200000
50%	45.000000	160.000000	60.000000	81.000000	1.000000	1.000000	120.000000	76.000000	96.000000	193.000000	14.300000
75%	60.000000	170.000000	70.000000	87.800000	1.200000	1.200000	131.000000	82.000000	105.000000	219.000000	15.400000
max	85.000000	190.000000	140.000000	99.000000	9.900000	9.900000	273.000000	185.000000	852.000000	2344.000000	25.000000

```
[ ] from scipy import stats
    stats.kstest(df['age'].dropna(), 'norm', args=(df['age'].mean(), df['age'].std()))

KstestResult(statistic=0.08670551084654154, pvalue=0.0, statistic_location=40, statistic_sign=1)
```

H1: распределение возраста наблюдаемых статистически значимо отличается от нормального распределения.

Вывод: поскольку $p\text{-value} < 0.05$, гипотеза H1 принимается

```
[ ] stats.kstest(df['height'].dropna(), 'norm', args=(df['height'].mean(), df['height'].std()))

KstestResult(statistic=0.10917542147943615, pvalue=0.0, statistic_location=170, statistic_sign=-1)
```

H1: распределение роста наблюдаемых статистически значимо отличается от нормального распределения.

Вывод: поскольку $p\text{-value} < 0.05$, гипотеза H1 принимается

```
[ ] stats.kstest(df['weight'].dropna(), 'norm', args=(df['weight'].mean(), df['weight'].std()))

KstestResult(statistic=0.11636946070517462, pvalue=0.0, statistic_location=60, statistic_sign=1)
```

H1: распределение веса наблюдаемых статистически значимо отличается от нормального распределения.

Вывод: поскольку $p\text{-value} < 0.05$, гипотеза H1 принимается

```
[ ] stats.kstest(df['waistline'].dropna(), 'norm', args=(df['waistline'].mean(), df['waistline'].std()))

KstestResult(statistic=0.06708709135286273, pvalue=0.0, statistic_location=90.0, statistic_sign=1)
```

H1: распределение линии талии наблюдаемых статистически значимо отличается от нормального распределения.

Вывод: поскольку $p\text{-value} < 0.05$, гипотеза H1 принимается

```
[ ] df.groupby("DRK_YN").agg({"age": ["mean", "std"], "height": ["mean", "std"],
                             "weight": ["mean", "std"],
                             "waistline": ["mean", "std"],
                             "sight_left": ["mean", "std"],
                             "sight_right": ["mean", "std"],
                             "SBP": ["mean", "std"],
                             "DBP": ["mean", "std"],
                             "BLDS": ["mean", "std"],
                             "tot_chole": ["mean", "std"],
                             "hemoglobin": ["mean", "std"]})
```

	age		height		weight		waistline		sight_left		...	SBP		DBP		BLDS		tot_chole		hemoglobin	
	mean	std	mean	std	mean	std	mean	std	mean	std	...	mean	std	mean	std	mean	std	mean	std	mean	std
DRK_YN																					
N	51.648809	14.376473	158.764848	8.928628	59.977998	11.541205	80.131769	13.632576	0.933964	0.667970	...	121.950591	14.892083	75.055403	9.673541	100.068826	24.324842	194.794921	39.361543	13.755567	1.534434
Y	43.577160	12.765032	165.718998	8.272969	66.592571	12.575625	82.335770	9.620282	1.027740	0.532642	...	122.914765	14.168978	77.050597	10.001695	100.780334	24.028851	196.319689	37.929979	14.704434	1.489728

2 rows × 22 columns



```
[ ] df.groupby("SMK_stat_type_cd").agg({"age": ["mean", "std"], "height": ["mean", "std"],
                                         "weight": ["mean", "std"],
                                         "waistline": ["mean", "std"],
                                         "sight_left": ["mean", "std"],
                                         "sight_right": ["mean", "std"],
                                         "SBP": ["mean", "std"],
                                         "DBP": ["mean", "std"],
                                         "BLDS": ["mean", "std"],
                                         "tot_chole": ["mean", "std"],
                                         "hemoglobin": ["mean", "std"]})
```

	age		height		weight		waistline		sight_left		...	SBP		DBP		BLDS		tot_chole		hemoglobin	
	mean	std	mean	std	mean	std	mean	std	mean	std	...	mean	std	mean	std	mean	std	mean	std	mean	std
SMK_stat_type_cd																					
1.0	48.455401	14.830668	158.572582	8.686544	59.306272	11.348157	79.001125	12.957333	0.949829	0.613180	...	121.177911	14.891399	74.916407	9.802083	98.471651	21.894984	195.336906	38.160672	13.638717	1.484280
2.0	50.112632	12.996934	167.382296	6.867679	69.406891	10.922270	85.312877	8.473249	1.013957	0.621488	...	125.345011	13.784620	78.059588	9.680146	104.173711	25.934026	194.951243	39.209486	14.967969	1.263143
3.0	43.203960	12.159067	168.364555	7.158454	69.477808	12.324586	84.182930	9.104247	1.041053	0.565267	...	123.583527	13.693754	77.610842	9.830111	102.857245	27.947985	196.672154	39.572793	15.290647	1.255593

3 rows × 22 columns


```
[ ] dfMetric.corr()
```

	age	height	weight	waistline	sight_left	sight_right	SBP	DBP	BLDS	tot_chole	hemoglobin
age	1.000000	-0.398501	-0.195337	0.127170	-0.172096	-0.167684	0.265530	0.108847	0.195796	0.011446	-0.173081
height	-0.398501	1.000000	0.668823	0.263945	0.139141	0.138529	0.035030	0.108780	0.021266	-0.023240	0.531898
weight	-0.195337	0.668823	1.000000	0.637173	0.088901	0.088707	0.250770	0.277891	0.138587	0.063238	0.499491
waistline	0.127170	0.263945	0.637173	1.000000	0.004511	0.006158	0.272323	0.240890	0.175519	0.063201	0.291730
sight_left	-0.172096	0.139141	0.088901	0.004511	1.000000	0.307985	-0.035617	-0.001209	-0.034817	0.004371	0.085896
sight_right	-0.167684	0.138529	0.088707	0.006158	0.307985	1.000000	-0.033994	-0.000568	-0.036893	0.003437	0.086847
SBP	0.265530	0.035030	0.250770	0.272323	-0.035617	-0.033994	1.000000	0.741131	0.183141	0.068557	0.166530
DBP	0.108847	0.108780	0.277891	0.240890	-0.001209	-0.000568	0.741131	1.000000	0.136266	0.111915	0.241980
BLDS	0.195796	0.021266	0.138587	0.175519	-0.034817	-0.036893	0.183141	0.136266	1.000000	0.012713	0.101712
tot_chole	0.011446	-0.023240	0.063238	0.063201	0.004371	0.003437	0.068557	0.111915	0.012713	1.000000	0.121272
hemoglobin	-0.173081	0.531898	0.499491	0.291730	0.085896	0.086847	0.166530	0.241980	0.101712	0.121272	1.000000

Составление модели линейной регрессии

```
[ ] corr_table = df[['age', 'sight_left', 'sight_right', 'BLDS', 'tot_chole', 'height', 'weight', 'waistline']].corr()  
print(f"Вторая по величине корреляция = {np.sort(np.unique(corr_table.values))[-2]}")  
corr_table
```

Вторая по величине корреляция = 0.6688234949483525

	age	sight_left	sight_right	BLDS	tot_chole	height	weight	waistline
age	1.000000	-0.172096	-0.167684	0.195796	0.011446	-0.398501	-0.195337	0.127170
sight_left	-0.172096	1.000000	0.307985	-0.034817	0.004371	0.139141	0.088901	0.004511
sight_right	-0.167684	0.307985	1.000000	-0.036893	0.003437	0.138529	0.088707	0.006158
BLDS	0.195796	-0.034817	-0.036893	1.000000	0.012713	0.021266	0.138587	0.175519
tot_chole	0.011446	0.004371	0.003437	0.012713	1.000000	-0.023240	0.063238	0.063201
height	-0.398501	0.139141	0.138529	0.021266	-0.023240	1.000000	0.668823	0.263945
weight	-0.195337	0.088901	0.088707	0.138587	0.063238	0.668823	1.000000	0.637173
waistline	0.127170	0.004511	0.006158	0.175519	0.063201	0.263945	0.637173	1.000000

Перекодирование в дамми-переменные

```
[ ] DRK_dummies = pd.get_dummies(df.DRK_YN, prefix='DRK', prefix_sep='_')
DRK_dummies
```

	DRK_N	DRK_Y
0	0	1
1	1	0
2	1	0
3	1	0
4	1	0
...
991341	1	0
991342	1	0
991343	0	1
991344	1	0
991345	0	1

991346 rows × 2 columns

```
[ ] SMK_dummies = pd.get_dummies(df.SMK_stat_type_cd, prefix='SMK', prefix_sep='_')
SMK_dummies
```

	SMK_1.0	SMK_2.0	SMK_3.0
0	1	0	0
1	0	0	1
2	1	0	0
3	1	0	0
4	1	0	0
...
991341	1	0	0
991342	1	0	0
991343	0	0	1
991344	1	0	0
991345	0	0	1

991346 rows × 3 columns



OLS Regression Results

```

=====
Dep. Variable:          hemoglobin    R-squared:                0.387
Model:                  OLS          Adj. R-squared:            0.387
Method:                 Least Squares  F-statistic:              5.689e+04
Date:                   Sun, 25 Feb 2024  Prob (F-statistic):        0.00
Time:                   17:37:36      Log-Likelihood:           -1.6207e+06
No. Observations:       991346       AIC:                      3.241e+06
Df Residuals:           991334       BIC:                      3.242e+06
Df Model:                11
Covariance Type:        nonrobust
=====

```

	coef	std err	t	P> t	[0.025	0.975]
const	3.4459	0.034	100.386	0.000	3.379	3.513
age	-0.0001	0.000	-1.069	0.285	-0.000	9.49e-05
sight_left	0.0268	0.002	12.273	0.000	0.023	0.031
sight_right	0.0314	0.002	14.339	0.000	0.027	0.036
BLDS	0.0027	5.36e-05	51.094	0.000	0.003	0.003
tot_chole	0.0045	3.25e-05	138.670	0.000	0.004	0.005
height	0.0454	0.000	212.012	0.000	0.045	0.046
weight	0.0225	0.000	127.168	0.000	0.022	0.023
waistline	0.0050	0.000	33.575	0.000	0.005	0.005
DRK_Y	0.1920	0.003	68.177	0.000	0.186	0.197
SMK_2.0	0.5891	0.004	154.688	0.000	0.582	0.597
SMK_3.0	0.8519	0.004	238.228	0.000	0.845	0.859

```

=====
Omnibus:                 109268.127  Durbin-Watson:              2.004
Prob(Omnibus):            0.000      Jarque-Bera (JB):            306962.074
Skew:                     -0.610     Prob(JB):                     0.00
Kurtosis:                 5.438      Cond. No.                     8.20e+03
=====

```

Notes:

- [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
- [2] The condition number is large, 8.2e+03. This might indicate that there are strong multicollinearity or other numerical problems.

Составление модели бинарной регрессии

Предикторы оставим теми же, что и в случае линейной регрессии, исключив, предиктор DRK_YN, так как это теперь будет целевая переменная. Также добавим в предикторы уровень гемоглобина (hemoglobin), так как он теперь не является целевой переменной (его корреляции с остальными метрическими признаками мы считали, и выяснили, что сильных корреляций нет, так что мы можем его добавить в нашу модель)

```
▶ x = pd.concat([df[['age', 'sight_left', 'sight_right', 'BLDS', 'tot_chole', 'height', 'weight', 'waistline', 'hemoglobin']],  
               SMK_dummies[['SMK_2.0', 'SMK_3.0']], ], axis=1) # SMK_1.0 - референтная группа  
y = df['DRK_YN'].apply(lambda x: 1 if x == 'Y' else 0)  
  
x = sm.add_constant(x)  
  
model_bin = sm.Logit(y, x).fit()  
print(model_bin.summary())
```

Logit Regression Results

```

=====
Dep. Variable:          DRK_YN    No. Observations:          991346
Model:                  Logit      Df Residuals:                991334
Method:                  MLE       Df Model:                    11
Date:                   Sun, 25 Feb 2024    Pseudo R-squ.:              0.1730
Time:                   23:27:19    Log-Likelihood:             -5.6827e+05
converged:              True       LL-Null:                    -6.8715e+05
Covariance Type:        nonrobust    LLR p-value:                0.000
=====

```

	coef	std err	z	P> z	[0.025	0.975]

const	-6.5559	0.062	-105.035	0.000	-6.678	-6.434
age	-0.0361	0.000	-186.730	0.000	-0.037	-0.036
sight_left	0.0109	0.004	2.668	0.008	0.003	0.019
sight_right	-0.0023	0.004	-0.564	0.573	-0.010	0.006
BLDS	0.0021	0.000	20.852	0.000	0.002	0.002
tot_chole	0.0011	6.05e-05	18.109	0.000	0.001	0.001
height	0.0358	0.000	89.464	0.000	0.035	0.037
weight	-0.0012	0.000	-3.697	0.000	-0.002	-0.001
waistline	0.0004	0.000	1.561	0.118	-0.0001	0.001
hemoglobin	0.1144	0.002	62.829	0.000	0.111	0.118
SMK_2.0	1.1373	0.007	167.660	0.000	1.124	1.151
SMK_3.0	1.1685	0.007	178.123	0.000	1.156	1.181

```
=====
```

Подготовка к кластеризации

Выберем для кластеризации следующий набор переменных: age, height, waistline, tot_chole, hemoglobin

```
[52] df[['age', 'height', 'waistline', 'tot_chole', 'hemoglobin']].corr()
```

	age	height	waistline	tot_chole	hemoglobin
age	1.000000	-0.398501	0.127170	0.011446	-0.173081
height	-0.398501	1.000000	0.263945	-0.023240	0.531898
waistline	0.127170	0.263945	1.000000	0.063201	0.291730
tot_chole	0.011446	-0.023240	0.063201	1.000000	0.121272
hemoglobin	-0.173081	0.531898	0.291730	0.121272	1.000000



Так как датасет очень большой, то построить дендрограммы для него вычислительно невозможно, так что проведем кластерный анализ на произвольных 42 наблюдениях

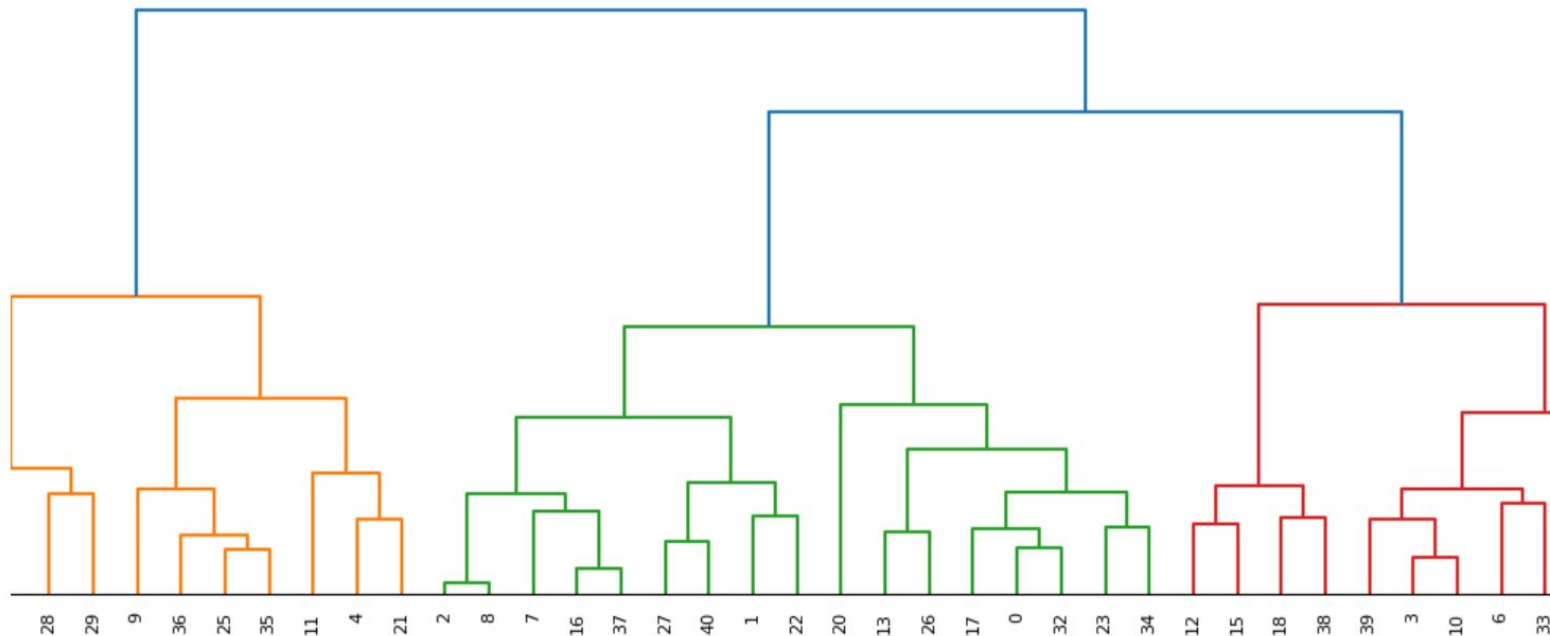
```
df1 = df1.sample(n=42)  
df1
```



	age	height	weight	waistline	sight_left	sight_right	SBP	DBP	BLDS	tot_chole	hemoglobin
154043	-0.184361	0.835873	0.936209	0.571009	0.361690	0.366370	0.520348	1.208103	0.313299	-0.376538	0.92756
415601	0.520791	0.297252	0.137120	0.655395	0.031629	-0.295034	-0.304783	-1.016509	-0.761972	1.485845	-0.14500
402393	0.520791	0.297252	1.335754	0.992939	-0.298432	-0.460385	0.589109	0.803628	6.516783	0.632253	1.49544
358168	-0.184361	-0.241370	-0.262425	-0.441622	-0.133401	-0.129683	0.039022	0.095797	-0.307050	-0.557603	-0.46047
885850	0.168215	-0.779991	0.137120	0.486623	0.031629	0.035668	0.795392	1.005866	0.520082	-0.040275	-1.34379
28020	1.578519	-1.857234	-1.061515	0.866360	-0.793523	14.751906	1.139196	1.309222	-0.679259	0.580520	0.10737

Дендрограмма по выборке

```
from scipy.cluster.hierarchy import dendrogram, linkage, fcluster  
plt.figure(figsize=(15, 5))  
dn = dendrogram(linkage(df1[['age', 'height', 'waistline', 'tot_chole', 'hemoglobin']], 'ward'))
```



Аналитическое подтверждение числу кластеров

```
from sklearn.metrics import calinski_harabasz_score
```

```
Z = linkage(df1[['age', 'height', 'waistline', 'tot_chole', 'hemoglobin']], method='ward', metric='euclidean')
```

```
for k in range (2, 10):
```

```
    labels = fcluster(Z, t=k, criterion='maxclust')
```

```
    print('Число кластеров: {}, индекс {}'.format(k, calinski_harabasz_score(df1[['age', 'height', 'waistline', 'tot_chole',
```

```
Число кластеров: 2, индекс 15.910285711986525
```

```
Число кластеров: 3, индекс 17.961137904796292
```

```
Число кластеров: 4, индекс 15.702425436325983
```

```
Число кластеров: 5, индекс 15.346705135207484
```

```
Число кластеров: 6, индекс 15.574729769172402
```

```
Число кластеров: 7, индекс 14.69077254642471
```

```
Число кластеров: 8, индекс 14.236508099489368
```

```
Число кластеров: 9, индекс 14.038849435539952
```


Статистика по сформированным кластерам

```
df1.groupby('culster_labels')[['age', 'height', 'waistline', 'tot_chole', 'hemoglobin']].mean()
```



	age	height	waistline	tot_chole	hemoglobin
culster_labels					
1	1.079036	-1.004417	0.301677	-0.208406	-0.807496
2	0.126735	0.772506	0.779988	0.242735	0.916463
3	-0.808150	0.090090	-0.827849	-0.352661	0.165612



✓ Описательная характеристика кластеров

Кластер 1 - пенсионеры

В этот кластер попали все люди, чей возраст превышает 55 лет, то есть это люди пенсионного возраста или вовсе глубоко пожилые. Такие люди не выделяются высоким ростом, что и отражает значение -1 в соответствующей позиции, а также они редко страдают от ожирения, так что показатель линии талии у них средний по 3 наблюдаемым кластерам, что логично. Так как в этот кластер попали пожилые люди, то у них с большой вероятностью будут отклонения от стандартных показателей здоровья, что и подтверждается в таблице выше (в этом кластере у людей пониженный холестерин и гемоглобин)

Кластер 2 - взрослые люди

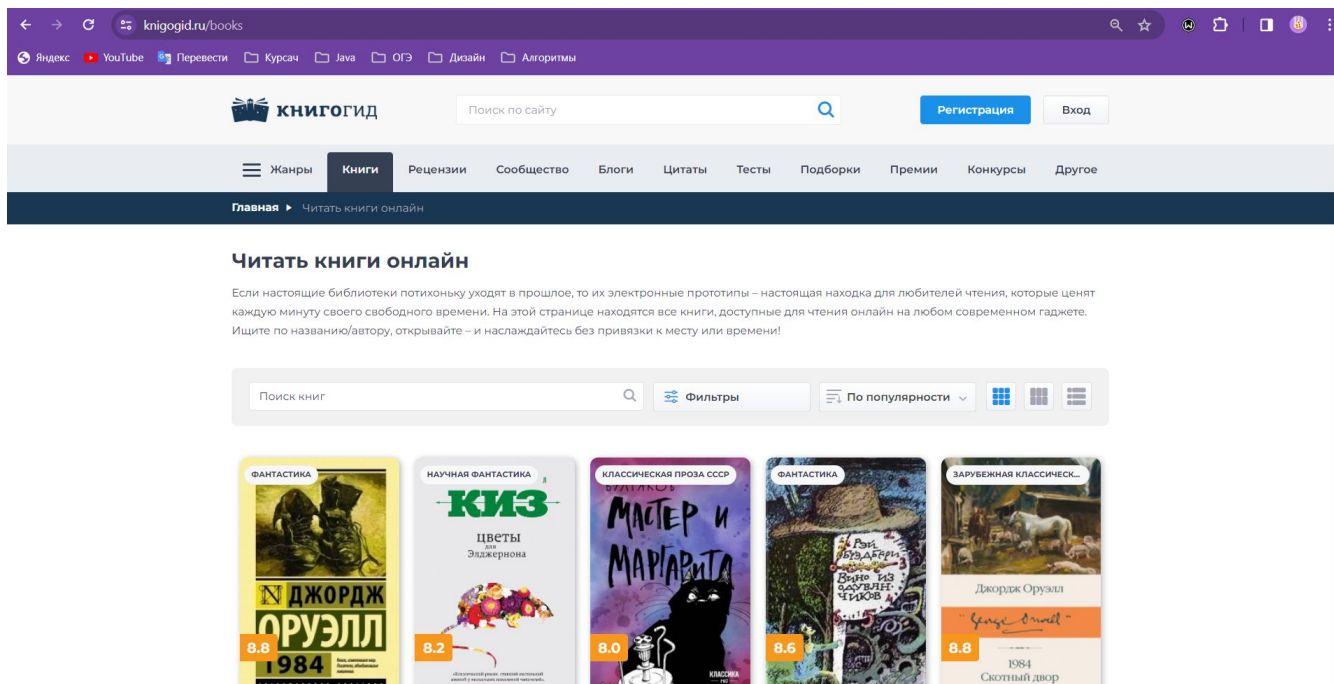
В этом кластере собрались люди, старшие 18 лет, но еще не вышедшие на пенсию. Словом, взрослые люди, в самом разгаре сил. По всем 3 кластерам - в этом средний показатель возраста, но самый высокий показатель роста, что логично, так как дети еще растут, и не догоняют в росте взрослое население, а пожилые люди наоборот теряют в росте. Линия талии среди этого кластера наибольшая, так как взрослые люди чаще всего страдают от ожирения, и в целом у них линия талии больше, чем у детей и пенсионеров. Показатели холестерина и гемоглобина у них в норме, за, возможно, некоторыми исключениями

Кластер 3 - дети

В этот кластер попали все дети (то есть люди, чей возраст меньше 18). Понятно, что это самая молодая часть из всей выборки, и у них невысокий рост, о чем и говорит соответствующее значение в 3 кластере в колонке height. Линия талии у детей еще маленькая, она вырастет в будущем, о чем также свидетельствует соответствующее значение в таблице. А холестерин и гемоглобин у детей зачастую ниже, чем у взрослого населения, что является медицинским фактом. Эти показатели вырастают с возрастом

Часть 2

https://knigogid.ru/books



```
[ ] url = 'https://knigogid.ru/books'
    countPage = 0
    links = set()

    while countPage != 10:
        r = requests.get(url)
        page = BeautifulSoup(r.text, 'html.parser')
        books = page.find('div', class_='b-items-container genres_books-list').findAll('a', class_='b-item-name')
        for el in books:
            links.add('https://knigogid.ru' + el.get('href'))
        countPage += 1
        url = 'https://knigogid.ru/books' + '/page-' + str(countPage)
```



knigogid.ru/books

YouTube



Перевести



Курсач



Java



ОГЭ



Дизайн



Алгоритмы



knigogid.ru/books/page-1

YouTube



Перевести



Курсач



Java



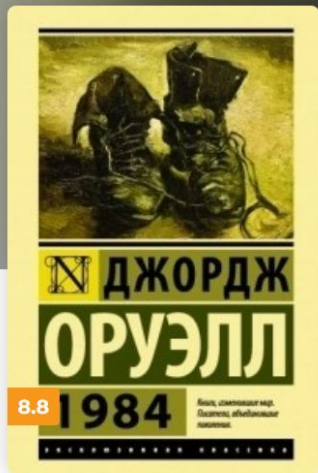
ОГЭ



Дизайн



Алгоритмы



Оцените книгу

h1.b-book-name 472.92 × 44.79

1984

1984

СЕРИЯ: ЭКСКЛЮЗИВНОЕ ЧТЕНИЕ НА АНГЛИЙСКОМ ЯЗЫКЕ **ЕЩЕ**



Джордж Оруэлл

145 книг

Описание книги

Одна из самых знаменитых антиутопий XX века – роман писателя Джорджа Оруэлла (1903–1950) был написан в 1948 году. Он продолжает тему «преданной революции», раскрытую в романе Оруэлла, нет и не может быть ничего ужаснее тотальной диктатуры. Тоталитаризм уничтожает в человеке все духовные потребности и сам разум, оставляя лишь постоянный страх и выбор – между молчанием и смертью, и если Старший Брат действительно существует, то...

Third-party cookie phaseout warnings in Network and Application

The Network and Application panels now show you warnings next to cookies affected by the third-party cookie restrictions from Tracking Protection.



По данным со сайтов составляем свой датафрейм, чтобы потом данные перенести в MS Excel. Также прописываем название каждого столбца, заменяя автоматически сгенерированные индексы.

```
[ ] df = pd.DataFrame(dataBooks)

df = df.rename(columns={0: 'Название_книги', 1: 'Автор', 2: 'Рейтинг_читателей', 3: 'Основной_жанр', 4: 'Год_выпуска', 5: 'Количество_страниц',
                        6: 'Возрастное_ограничение', 7: 'Количество_просмотров_книги', 8: 'ISBN',
                        9: 'Доступный_язык', 10: 'Ссылка_на_книгу'})

df.dtypes
```

Название_книги	object
Автор	object
Рейтинг_читателей	object
Основной_жанр	object
Год_выпуска	object
Количество_страниц	object
Возрастное_ограничение	object
Количество_просмотров_книги	object
ISBN	object
Доступный_язык	object
Ссылка_на_книгу	object
dtype:	object

Меняем целочисленные числа и десятичные дроби на соответствующие типы.

```
df['Рейтинг_читателей'] = df['Рейтинг_читателей'].astype(float)
df['Год_выпуска'] = df['Год_выпуска'].astype(int)
df['Количество_страниц'] = df['Количество_страниц'].astype(int)
df['Количество_просмотров_книги'] = df['Количество_просмотров_книги'].astype(int)
df['Возрастное_ограничение'] = df['Возрастное_ограничение'].astype(int)
df
```

	Название_книги	Автор	Рейтинг_читателей	Основной_жанр	Год_выпуска	Количество_страниц	Возрастное_ограничение	Количество_просмотров_книги	ISBN	Доступный_язык	Ссылка_на_книгу
0	Таинственная история Билли Миллигана	Дэниел Киз	8.8	Современная проза	2019	640	18	10196	978-5-699-81491-6	Русский	https://knigogid.ru/books/1435315-tainstvennay...
1	Преступление и наказание. Графический роман	Фёдор Достоевский	8.8	Отечественная классическая проза	2017	72	12	5143	978-5-91339-863-5	Русский	https://knigogid.ru/books/184784-prestuplenie-...
2	1984. Скотный двор	Джордж Оруэлл	8.0	Фантастика	1945	384	16	1454	978-5-17-101063-8	Русский	https://knigogid.ru/books/1625452-1984-skotnyy...
3	Убить пересмешника...	Харпер Ли	8.2	Зарубежная классическая проза	1959	416	12	8617	978-5-17-083520-1	Русский	https://knigogid.ru/books/1618158-ubit-peresme...
4	Повелитель мух	Уильям Голдинг	8.0	Зарубежная классическая проза	1954	352	16	7671	5-17-017034-3	Русский	https://knigogid.ru/books/61973-povellitel-muh
...
200	Тринадцатая сказка	Диана Сеттерфилд	8.6	Триллер	2006	464	18	8462	978-5-389-05094-5	Русский	https://knigogid.ru/books/566559-trinadcataya-...
201	Одиннадцать минут	Пауло Козльо	7.0	Любовный роман	2019	320	16	811	978-5-17-088736-1	Русский	https://knigogid.ru/books/1435341-odinnadcat-m...
202	11/22/63. Уровень 4	Стивен Кинг	8.0	Литература на английском языке	2019	864	16	346	978-5-17-115915-3	Русский	https://knigogid.ru/books/1445928-112263-uroven-4
203	Мара и Морок. Особенная Тень	Лия Арден	8.0	Фэнтези	2020	416	16	2422	978-5-04-110919-6	Русский	https://knigogid.ru/books/1560768-mara-i-morok...
204	Безмолвный пациент	Алекс Михаэлидес	8.2	Детектив	2018	352	16	5055	978-5-04-153406-6	Русский	https://knigogid.ru/books/589274-bezmolvnyy-pa...

205 rows × 11 columns

Создаем MS Excel файл, с помощью функции `to_excel()` из библиотеки `pandas`.

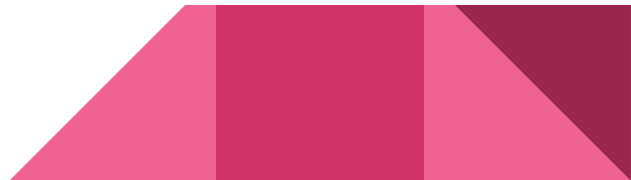


```
df.to_excel('Глебов_Павел_Фролов_Буканов_Часть_2.xlsx')
```


Выводы о проделанной работе

Мы провели комплексный анализ датасета, проведя описательный анализ, кластерный, а также построив модель линейной и бинарной регрессии. В ходе исследования нами были проверены гипотезы, сформулированные в начале блокнота. Во всех случаях мы отвергли нулевую гипотезу в пользу альтернативной. То есть везде, где проверяли, есть ли статистически значимая взаимосвязь, то она действительно была, а везде, где проверяли, значимо ли отличается распределение переменной от нормального распределения, оно значимо отличалось. Такой результат оказался неожиданным, так как мы ожидали, что будет принята хотя бы 1-2 нулевые гипотезы

В ходе построения графиков мы выяснили, что датасет несколько не равномерен, так как очень большое число женщин никогда не курило, а также не пьет алкоголь, при условии, что число женщин и мужчин в датасете примерно равно. Такая ситуация, кстати, может иметь связь с реальностью, но мы все же ожидали более равномерного распределения среди пьющих/курящих по полам



В части с построением моделей линейной и бинарной регрессии нам удалось построить статистически значимые модели, из которых мы получили следующую информацию:

- Курение влияет на уровень гемоглобина в крови человека, особенно если человек еще курит. У курящих людей уровень гемоглобина в крови выше. Согласно модели линейной регрессии, если человек курил, но бросил, то его уровень гемоглобина в крови выше на 0.5891 мг/дл, чем если бы он никогда не курил, а все человек все еще курит, то это число равно 0.8519 мг/дл
- Употребление алкоголя также влияет на уровень гемоглобина в крови. Гемоглобин выше у пьющих людей. Согласно модели линейной регрессии, если человек пьет, то его уровень гемоглобина в крови выше на 0.192 мг/дл, чем если бы он этого не делал
- Также на уровень гемоглобина влияет рост, но это, скорее всего, связано с разностью роста между мужчинами и женщинами, а у них разные нормы гемоглобина с медицинской точки зрения
- Если человек курит или курил раньше, то шансы того, что он в таком случае еще и пьет, выше, чем если бы человек не курил. В целом, логичный вывод, так как курящие люди явно не адепты здорового образа жизни, так что они запросто могут быть пристрастны к алкоголю, в то время как некурящие люди в среднем пьют реже
- Если у человека высокий уровень гемоглобина, то он с большой вероятностью употребляет алкоголь (согласно модели бинарной регрессии, при увеличении уровня гемоглобина в крови на 1 мг/дл, логарифм шансов того, что человек пьет, возрастает на 0.1144 - самый высокий показатель по остальным метрическим переменным бинарной регрессии)

В целом, выводы ожидаемые: у людей, ведущих здоровый образ жизни, лучше и стабильнее медицинские показатели, что мы и подтвердили статистически своим исследованием