Добин Илья. Вариант 27

- Регион: Оренбургская область, Орск
- Год: 2023

Приступеим к работе

Задание 1

В нашем случае генеральная совокупность - трудоустроенное в различных отраслях население Оренбургской области с различным уровнем дохода.

Для проверки репрезентативности выборки стоит посмотреть на распределение признаков - оценить, нет ли выбросов, перекосов распре1деления в конкретную сторону (например мужчин сильно больше женщин). Если распределения в выборке будут схожи с распределениями из генеральной совокупности, то выборка точно репрезентативна. Это можно сделать при помощи различных статистических тестов, например тест Колмогорова-Смирнова.

Задание 2

	min	max	mean	std	полный размах
educ	0.00	2.00	1.24	0.83	2.00
age	26.50	65.50	46.12	8.92	39.00
female	0.00	1.00	0.59	0.49	1.00
work_hours	4.00	12.00	8.17	1.62	8.00
wage	14000.00	200000.00	36475.65	24160.82	186000.00
foreign_language	0.00	1.00	0.39	0.49	1.00
internet	0.00	1.00	0.96	0.20	1.00
alcohol	0.00	1.00	0.55	0.50	1.00
is_children	1.00	1.00	1.00	0.00	0.00
health	0.00	1.00	0.31	0.47	1.00
weight	52.00	115.00	75.03	14.27	63.00
height	152.00	193.00	169.05	8.86	41.00
smoke	0.00	1.00	0.21	0.41	1.00

Я выделил значения численных признаков, так как их статистики будут более полезны. В категориальных можем только посмотреть на среднее и понять, каких значений в выборке больше:

- Женщин чуть больше мужчин
- Людей, знающих иностранный язык меньше
- Почти все люди пользуются интренетом
- Людей, употребляющих алкоголь и неупотребляющих +- одинаковое количство, также как и людей с проблемами со здоровьем / без
- Абсолютно все люди имеют детей. Это полезная информация, так как получается, что этот признак просто константа и не будет значим для нас в будущем
- Курит всего 20% людей выборки

Интерпретация числовых признаков: (про среднее подробнее будет в 4 задании)

- `age` - больщой размах - 39 лет. В выборке присутствуют, как молодые, так и пожилые люди. Среднее лежит практически по середине между min и max
- `work_hours` - Минимум и максимум отклоняются на 4 часа от стандартного рабочего времени, что хорошо - нет выбросов где люди работают излишне много.
- `wage` - очень большой размах - 186000р. Причем максимум гораздо дальше от среднего, чем минимум, что говорит о выбросах в виде людей с большими зарплатами.
- `weight` и `height` - все разумно. Результаты схожи с средними результатами по всей России.

Задание 3

	25%	50%	75%	межквартильный размах
age	39.75	44.50	53.25	13.50
height	163.00	168.00	175.00	12.00
weight	64.00	72.00	85.50	21.50
wage	25000.00	32000.00	42000.00	17000.00
work_hours	8.00	8.00	8.00	0.00

Интерпретация:

- `age` - межквартильный размах сильно меньше, что может говорить о том, что основная часть людей имеют возраст ближе к среднему, а распределение напоминает нормальное
- `work_hours` - Большинство людей в выборке работают ровно 8 часов в день, размах 0.
- `wage` - межквартильный размах также в разы меньше полного. Большинство зарплат ближе к среднему, то есть распределение скошено ближе к более маленьким зарплатам.

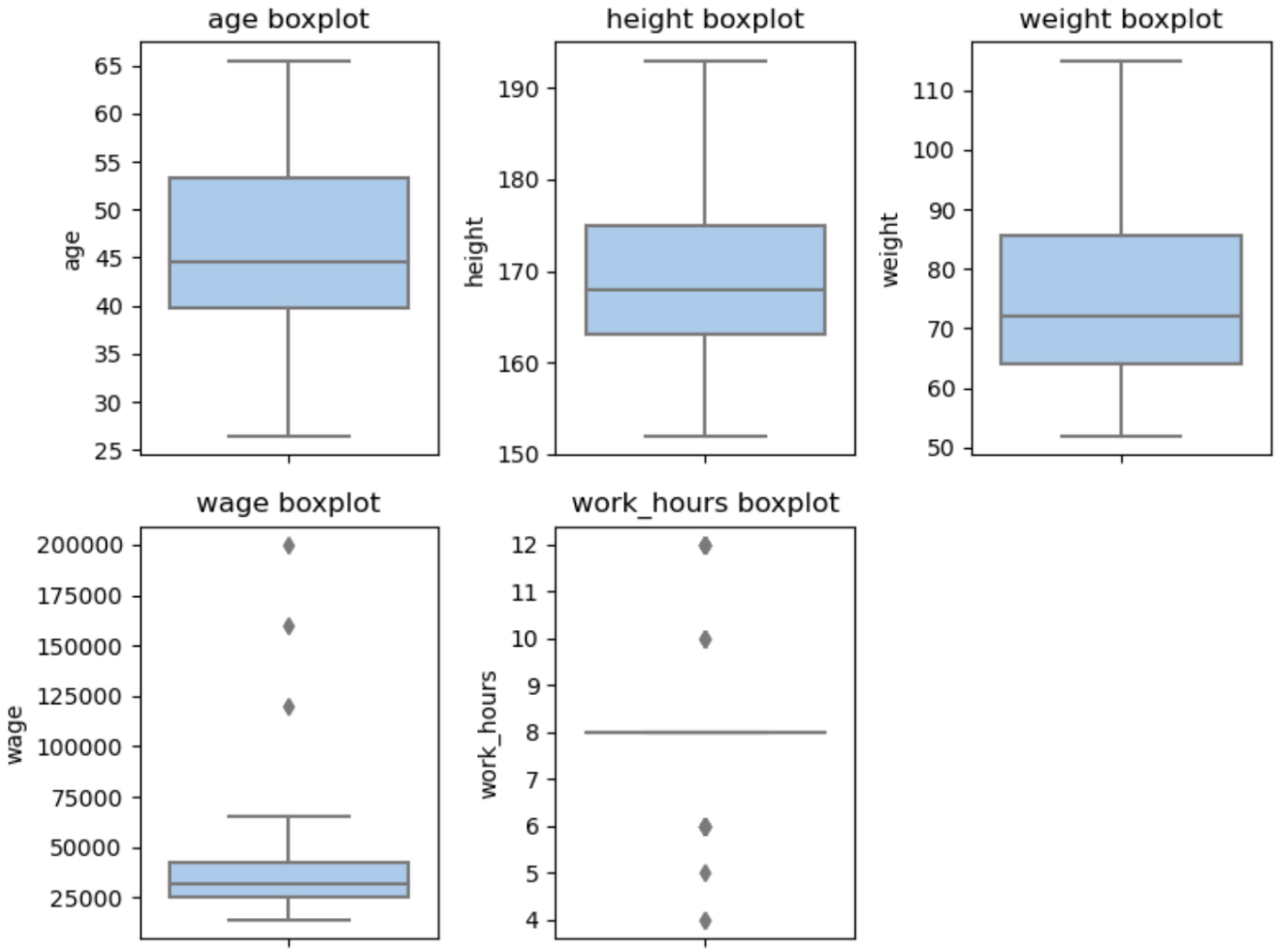
Задание 4

	min	mean	median	mode	max
age	26.5	46.117391	44.5	34.0	65.5
height	152.0	169.052174	168.0	170.0	193.0
weight	52.0	75.034783	72.0	68.0	115.0
wage	14000.0	36475.652174	32000.0	50000.0	200000.0
work_hours	4.0	8.165217	8.0	8.0	12.0

Интерпретация:

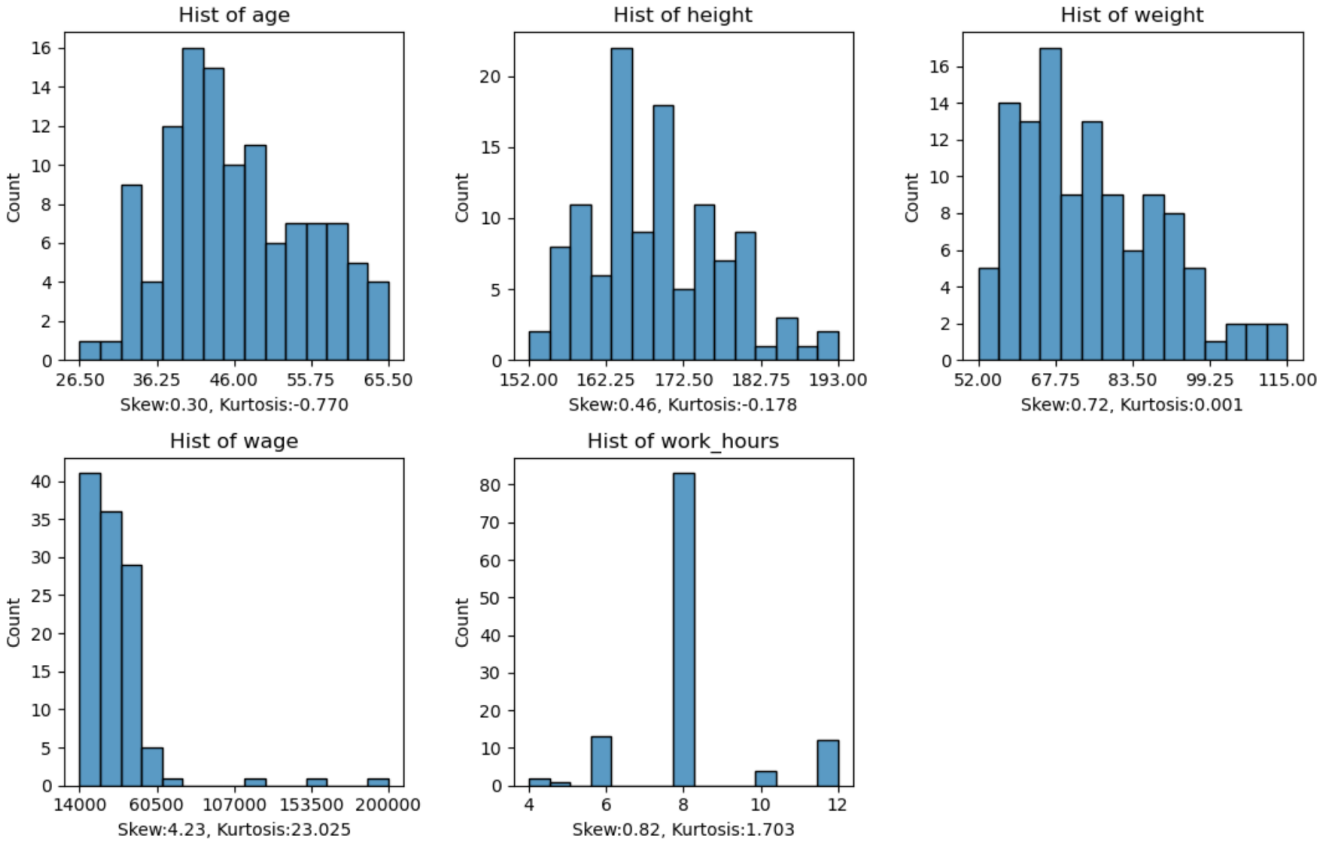
- `age` - среднее повыше медианы, то есть у нас все-таки есть незначительные выбросы в виде людей с большим возрастом.
- `work_hours` - среднее тоже чуть больше медианы, то есть людей, кто перерабатывает больше, чем тех, кто недорабатывает
- `wage` - среднее выше медианы, как уже говорили раньше у нас есть выбросы в виде людей с большой зарплатой. При этом мода составляет 50000, видимо работадателям нравится эта круглая цифра.
- `weight` - также из-за разницы между медианой и средним можно сделать вывод о небольшом перевесе в сторону более толстых людей
- `height` - медиана и среднее почти одинаковы, что значит, что распределение схоже с нормальным

Задание 5



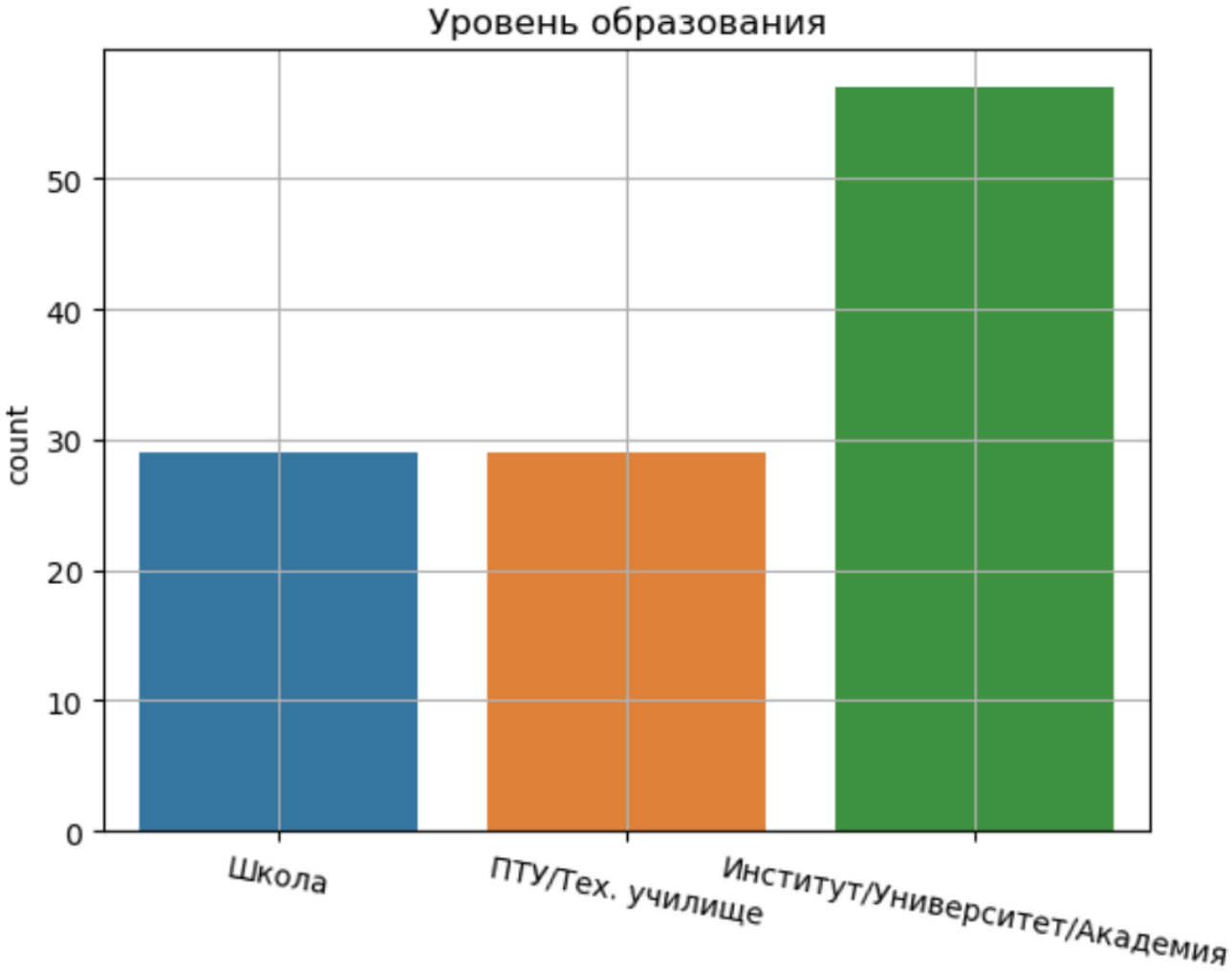
В этом задании мы по сути визуализировали все сказанное до этого. Видим, что в `age`, `weight`, `height` выбросов нет. В `wage` очевидно есть, но непонятно какой порог ставить, чтобы считать выброс сильно влияющим на результаты модели - выше 175к или же выше 100к? `work_hours` - абсолютное большинство работают 8 часов в день. Но при этом нельзя сказать, что тут есть выбросы, так как без значений ниже или выше медианы признак будет константным, а значит и бесполезным для нас.

Задание 6



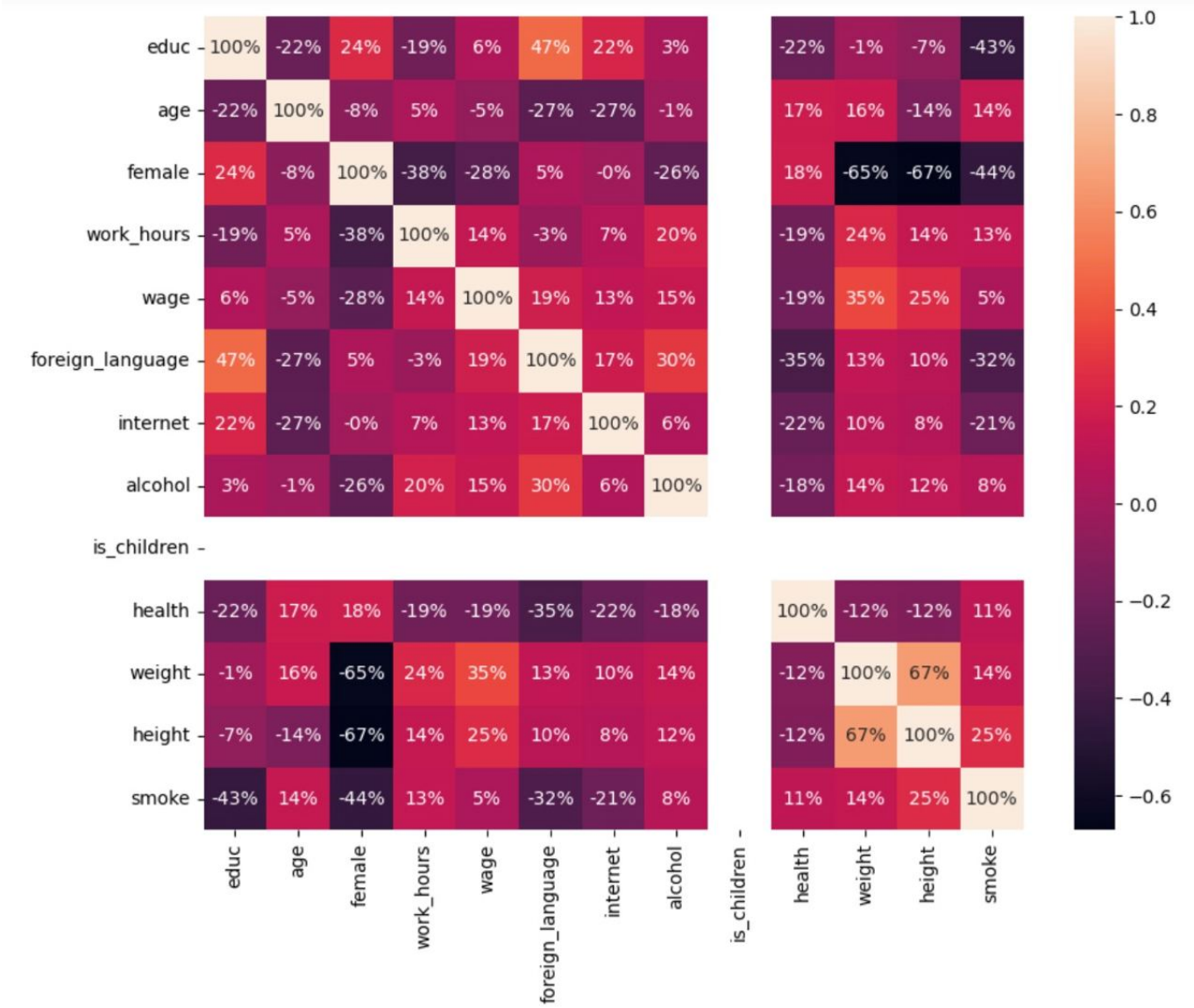
- `age` - Небольшой скос вправо (Skewness=0.3) - пожилых больше. Также распределение немного приплюснуто относительно нормального (Kurtosis=-0.77)
- `work_hours` - Также присутствует скос вправо (Skewness=0.82), то есть перерабатывающих больше. Kurtosis=1.73, что говорит о выраженном пике - 8-часовой рабочий день
- `wage` - Выборка сильно скошена (Skewness=4.23), значений с маленькой зарплатой в разы больше. Kurtosis=23.025 подтверждает наличие выбросов с излишне большими значениями
- `weight` - Также присутствует скос вправо (Skewness=0.82), при этом распределени практически совпадает по форме с нормальным (Kurtosis=0.01)
- `height` - скос вправо поменьше (Skewness=0.46), при этом распределени более плоское, чем нормальное (Kurtosis=-0.78)

Задание 7



Видим, что преобладают люди с высшим образованием.

Задание 8



Значения для is_children=NaN так как это константа и поэтому дисперсия нулевая, корреляции посчитать нельзя

Пройдусь по признакам, буду отмечать только сильные корреляции

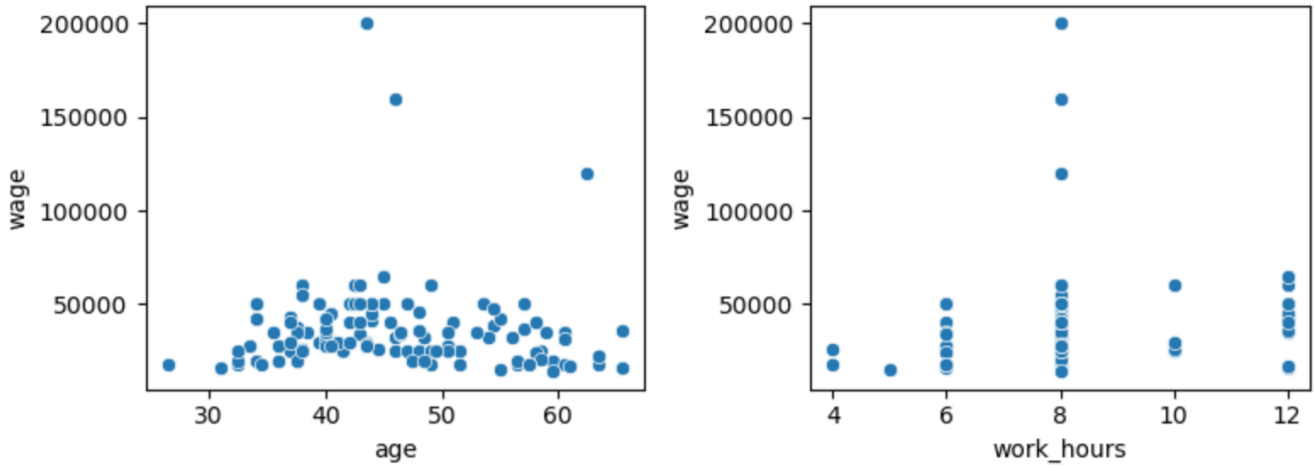
- `educ`: Высокая отрицательная корреляция с курением - видимо люди с высшим образованием реже курят. Также отрицательная корреляция с health, но вряд ли это можно действительно хорошо связать, не надумывая новых предпосылок. Отрицательная корреляция с age и work_hours. Более пожилые менее образованы и логично - меньше трудятся. Высокая корреляция с foreign_language, что логично - более образованные люди с юольшей вероятностью знают иностранный язык.
- `age`: Высокая отрицательная корреляция с foreign_language и internet. Первое объясняется тем, что как мы уже выяснили - более пожилые люди менее образованы, а второе тем, что действительно пожилые люди менее склонны пользоваться новыми технологиями, в том числе и интернетом
- `female`: Высокие отрицательные корреляции с 1. smoke - женщины меньше курят, 2. height - у женщин в среднем ниже рост, 3. weight - вследствие роста у женщин и ниже вес, 4.wage - зарплаты женщин меньше, 5. work_hours - женщины меньше работают, так как тратят больше времени на воспитывание детей 6. alcohol - женщины меньше выпивают. Положительная

корреляция с 1.educ - женщины более образованы, 2.health - женщины более здоровые вследствие меньшей склонности к вредным привычкам

- `work_hours`: Высокая корреляция с 1. alcohol - чем больше люди работают тем более склонны к выпиванию, 2. weight - чем больше работают, тем меньше двигаются, тем больше вес. отрицательная корреляция со здоровьем - трудяги более склонны заболевать.
- `wage`: Высокая корреляция с 1. wage и height, связано с тем, что мужчины зарабатывают больше 2. foreign_language - выступает преимуществом работника, следовательно зарплата у знающих иностранный выше. Отрицательная корреляция со здоровьем - опасная для здоровья работа оплачивается выше.
- `foreign language`: из нерасмотренного - корреляции с алкоголем, курением и здоровьем, слабо интерпретируемы напрямую.
- `internet`: из нерасмотренного - корреляции с курением и здоровьем, слабо интерпретируемы.
- `alcohol`: отрицательная корреляция со здоровьем, что очевидно
- `weight`: очень сильно коррелирует с ростом, что тоже логично и уже было сказано
- `height`: коррелирует с курением, так как мужчины выше и они чаще курят.

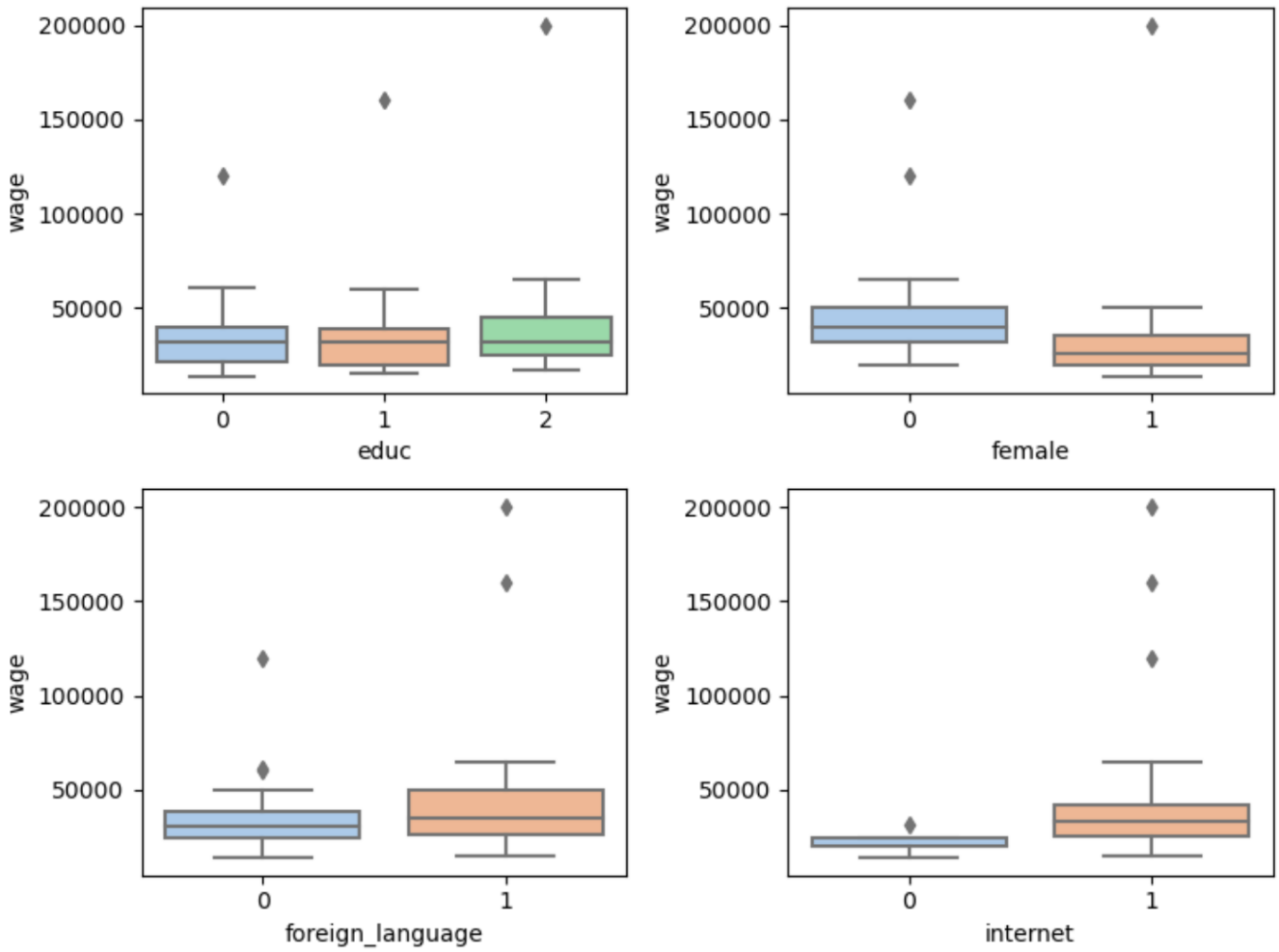
Задание 9

Числовые: я думаю заработная плата отрицательно зависит от возраста - пожилые получают меньше денег, так как менее востребованы. Также логично предположить, что люди, работающие больше в среднем получают зарплату выше. Посмотрим на эти зависимости



- `age` - видим, что если рассматривать людей старше 40, то небольшой нисходящий тренд присутствует, но как только включаем в рассмотрение более молодых людей - видим, что они зарабатывают столько же, сколько и пожилые, поэтому четкой зависимости нет.
- `work_hours` - люди, работающие меньше 8 часов в среднем имеют более низкую зарплату, но при этом работяги с >8 часами работы не зарабатывают больше, поэтому прямой зависимости тоже нет.

Категориальные: я выделил female, educ, foreign_language, internet как наиболее логичные признаки, влияющие на зарплату



- `female` - ящики сильно отделились, что позволяет однозначно сказать, что мужчины зарабатывают больше женщин
- `educ` - люди с высшим образованием зарабатываю больше. Но интересно, что люди, которые окончили только школу зарабатывают чуть больше людей, окончивших училища.
- `foreign_language` - люди с знанием иностранного языка зарабатывают в среднем одинаково, но разброс зарплат у людей с знанием языка выше, видимо на некоторых местах работы данный навык ценится и оплачивается, а на некоторых нет.
- `internet` - люди, пользующиеся интернетом получаются больше.

Задание 10

Константа: 50352.38940708004

Веса: [-129.85412264, 4920.63027699, -16222.72398987, 0, -3519.75570232]

Уравнение: $ wage = 50352.39 - 129.85 * age + 4920.63 * high - 16222.72 * female + 0 * is_children - 3519.76 * smoke $

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   wage   R-squared:                       0.098
Model:                            OLS   Adj. R-squared:                  0.066
Method:                 Least Squares   F-statistic:                     2.998
Date:                Sun, 24 Nov 2024   Prob (F-statistic):             0.0216
Time:                        22:23:58   Log-Likelihood:                -1317.4
No. Observations:                 115   AIC:                             2645.
Df Residuals:                     110   BIC:                             2658.
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
age          -129.8541    253.078     -0.513      0.609    -631.396     371.688
high          4920.6303   4742.293      1.038      0.302   -4477.481    1.43e+04
female      -1.622e+04   4961.997     -3.269      0.001   -2.61e+04   -6389.210
is_children   5.035e+04    1.3e+04      3.868      0.000    2.46e+04    7.61e+04
smoke       -3519.7557   6230.002     -0.565      0.573   -1.59e+04    8826.647
==============================================================================
Omnibus:                      151.308   Durbin-Watson:                   1.852
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             4424.090
Skew:                           4.853   Prob(JB):                         0.00
Kurtosis:                      31.794   Cond. No.                         287.
==============================================================================
```

- `age` - не статзначим
- `high` - не статзначим
- `female` - статзначим
- `is_children` - константа
- `smoke` - не статзначим

Задание 11

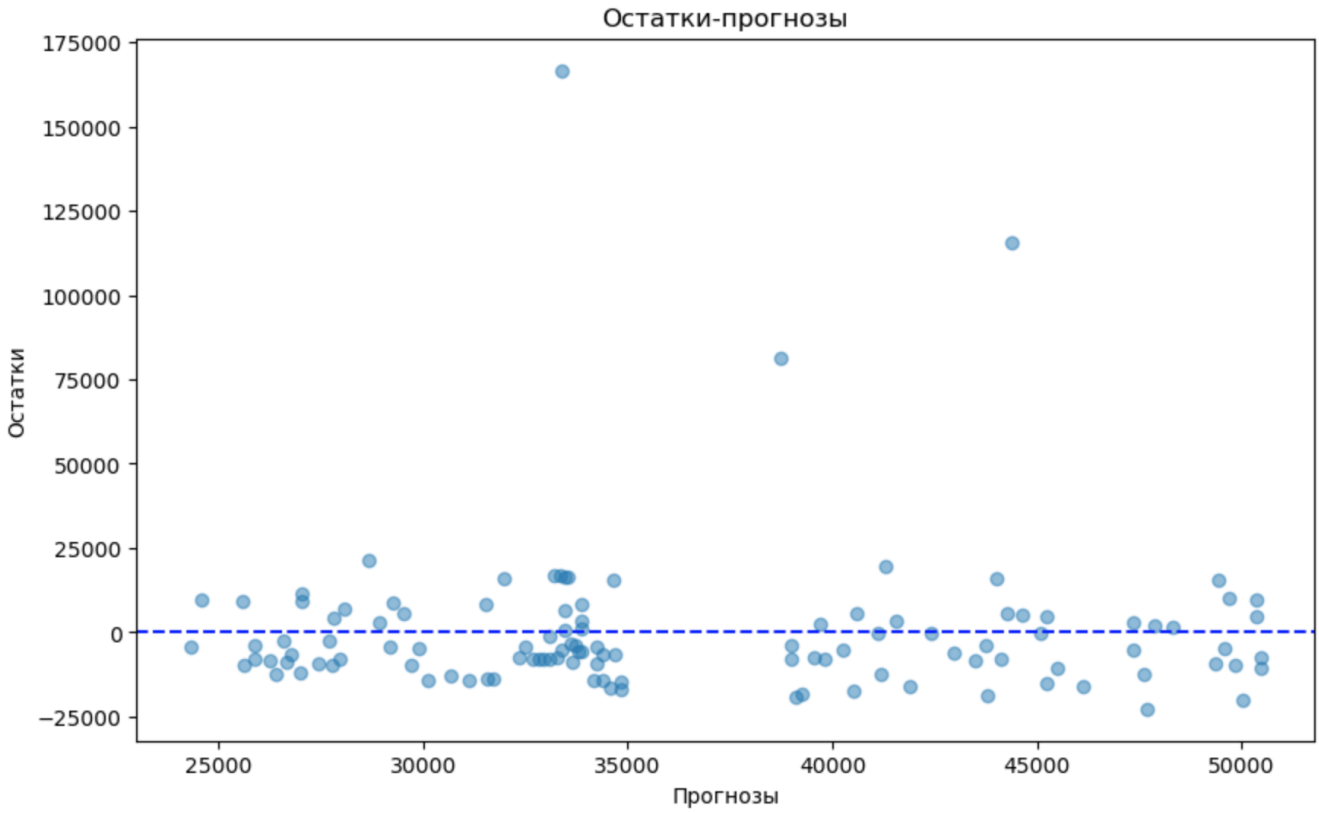
- Значение R^2 равно 0.098, что значит, что модель объясняет всего 9.8% разброса зарплаты
- F-статистика равана 2.998 - выше критического значение => отвергаем нулевую гипотезу о том, что модель с константой лучше.

Модель адекватна, но может быть сильно улучшена

Задание 12

Задание 13

Задание 14



Распределение остатков: На данном графике видно, что остатки распределяются случайным образом вокруг горизонтальной линии, но по середине точек нет => присутствует гетероскедастичность

Выбросы - точки, сильно отклоняющиеся от горизонтальной линии. Видно 3 выброса.

Задание 15

В прошлой модели из статзначимых признаков был только female, но значимыми можно считать и high, age, так как они хорошо коррелируют с значением wage

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   wage   R-squared:                       0.096
Model:                            OLS   Adj. R-squared:                  0.071
Method:                 Least Squares   F-statistic:                     3.914
Date:                Sun, 24 Nov 2024   Prob (F-statistic):             0.0106
Time:                        23:03:03   Log-Likelihood:                -1317.5
No. Observations:                 115   AIC:                             2643.
Df Residuals:                     111   BIC:                             2654.
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       4.903e+04   1.28e+04      3.841      0.000    2.37e+04    7.43e+04
age          -138.9226    251.793     -0.552      0.582    -637.868     360.022
high        5603.3670   4571.653      1.226      0.223   -3455.668    1.47e+04
female      -1.51e+04   4532.619     -3.331      0.001   -2.41e+04   -6118.184
==============================================================================
Omnibus:                      151.322   Durbin-Watson:                   1.860
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             4425.743
Skew:                           4.854   Prob(JB):                         0.00
Kurtosis:                      31.799   Cond. No.                         279.
==============================================================================
```

- AIC и BIC уменьшились => модель значительно лучше
- R^2 немного упал, значит исключенные переменные все-таки вносили небольшой вклад
- AdjR^2 вырос => объясняющяя способность модели улучшилась

Задание 16

Выбросы можно найти следующими способами:

1. Через стьюдентизированные остатки: выбросы - это наблюдения, где остаток > 2
2. Через DFFITS: выбросы - это наблюдения, где |DFFITS| > $2/\sqrt{n/k}$

Задание 17

Можем сделать вывод о 94х-летнем пареньке из Орска: зарплата: 47181.70835169294 доверительный интервал: [[15996.16093788 78367.25576551]]