Задание: Анализ датасета в индустрии науки и культуры для всех регионов России, предоставленных в датасете.

1) Опишите Вашу выборку. Что является в Вашем случае генеральной совокупностью? Как можно проверить репрезентативность Вашей выборки? (описать словами, не проверять).

Выборка представляет собой таблицу данных о 119 респондентов. Выборка содержатся полученные в ходе опроса наблюдения по домохозяйствам/индивидам, которые репрезентируют население России, работающее в сферах "НАУКА, КУЛЬТУРА"в 2023 году. Генеральной совокупностью являются жители РФ всех регионов, работающие в сферах "НАУКА, КУЛЬТУРА"в 2023 году.

Проверить репрезентатичность выборки можно следующим образом:

Определите ключевые характеристики генеральной совокупности, которые важны для исследования (возраст, пол, доход, образование, география и т.д.)(это сделано).

Сравните распределения этих характеристик в выборке с известными данными о генеральной совокупности.

Убедитесь, что выборка сформирована случайным образом (тоесть чтобы не было такого что спросили только у одной группы населения, например только мажоры с рублёвки или работяги из посёлка под Хабаровском)(вроде выполнено, берём из всех доступных регонов всего опроса).

Проверка на отсутствие систематических смещений (например, если респонденты в опросе — только активные пользователи интернета, это не отражает мнение всех возрастных групп)(у нас есть пунктик про пользование интернета, такчто вроде ок).

В данной выборке у каждого респондента есть хотя бы 1 ребенок, что говорит о нерепрезентативности выборки.

2) Рассчитайте описательные статистики (минимум, максимум, среднее значение, стандартное отклонение, размах) для всех переменных в Вашей выборке кроме отрасли, региона и года.

Table 1. Описательные статистики

Statistic	Educ	Age	Female	Work Hours	Wage	Foreign Lang	Internet	Alcohol	Is Children	Health	Weight	Height	Smoke
mean	1.20	49.41	0.77	7.93	35538.66	0.20	0.96	0.64	1.0	0.27	74.81	166.14	0.19
std	0.83	12.12	0.42	3.47	27792.85	0.40	0.20	0.48	0.0	0.44	15.02	9.09	0.39
min	0.0	26.0	0.0	2.0	8000.0	0.0	0.0	0.0	1.0	0.0	45.0	145.0	0.0
max	2.0	84.5	1.0	24.0	150000.0	1.0	1.0	1.0	1.0	1.0	125.0	196.0	1.0
range	2.0	58.5	1.0	22.0	142000.0	1.0	1.0	1.0	0.0	1.0	80.0	51.0	1.0

3) Оцените квантили (25%, 50%, 75%) распределения для количественных переменных в выборке. Определите межквартильный размах.

Table 2. Квантили и межквантильный размах

Statistic	Age	Work Hours	Wage	Weight	Height
25%	38.0	7.0	17800.0	64.0	160.0
50%	49.5	8.0	26000.0	73.0	165.0
75%	58.5	8.0	40000.0	84.0	170.0
IQR	20.5	1.0	22200.0	20.0	10.0

4) Сравните среднее значение, медиану и моду для количественных переменных в выборке. Что можно сказать об их соотношении?

Table 3. Среднее значение, медиана, мода

Statistic	Age	Work Hours	Wage	Weight	Height
Mean	49.41	7.93	35538.66	74.81	166.14
50% (Median)	49.5	8.0	26000.0	73.0	165.0
Mode	36.5	8.0	40000.0	68.0	164.0

Часы работы, рост и вес предпложительно имеют симитричное распределение ($mean \approx mode \approx median$)

Зарплата: $mode > mean > median - >$ левосторонняя асимметрия, в левом хвосте есть выбросы ($mean > median$)

Возраст: $mean \approx median > mode - >$ правосторонняя асимметрия (небольшая), в правом хвосет могут быть выбросы, но $mean \approx median$, такчто шанс маленький

Акцент на зарплаты(поскольку их будем прогнозировать): явно видно, что мода и среднее сильно выше медианы, возможно причина этому - выбросы.

5) Постройте box–plot для всех количественных переменных. Есть ли выбросы?

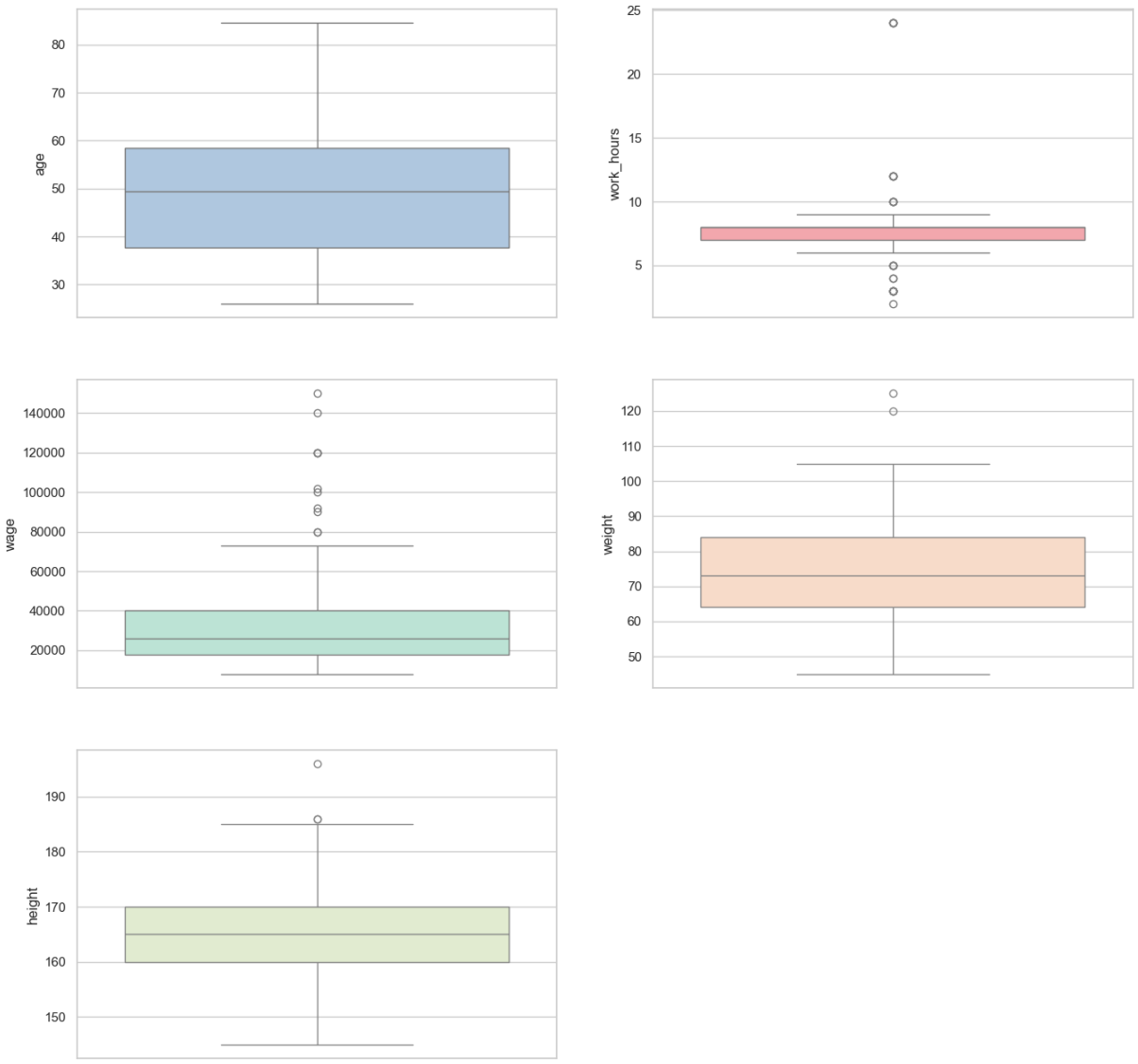


Рис. 1. Ящики с усами для количественных перменных

Как и предполагалось, выбросы присутствуют, кроме того не только у зарплаты, а еще и в остальных переменных кроме age следуя.

Пояснение: Ящики с усами считают за выбросы значения выше $q_3 + 1.5\text{IQR}$ и ниже $q_1 - 1.5\text{IQR}$.

6) Постройте гистограммы распределения для количественных переменных в выборке. Что можно сказать о скошенности (асимметрии) и островершинности их распределений? Рассчитайте соответствующие показатели (Skewness и Kurtosis) и сделайте выводы.

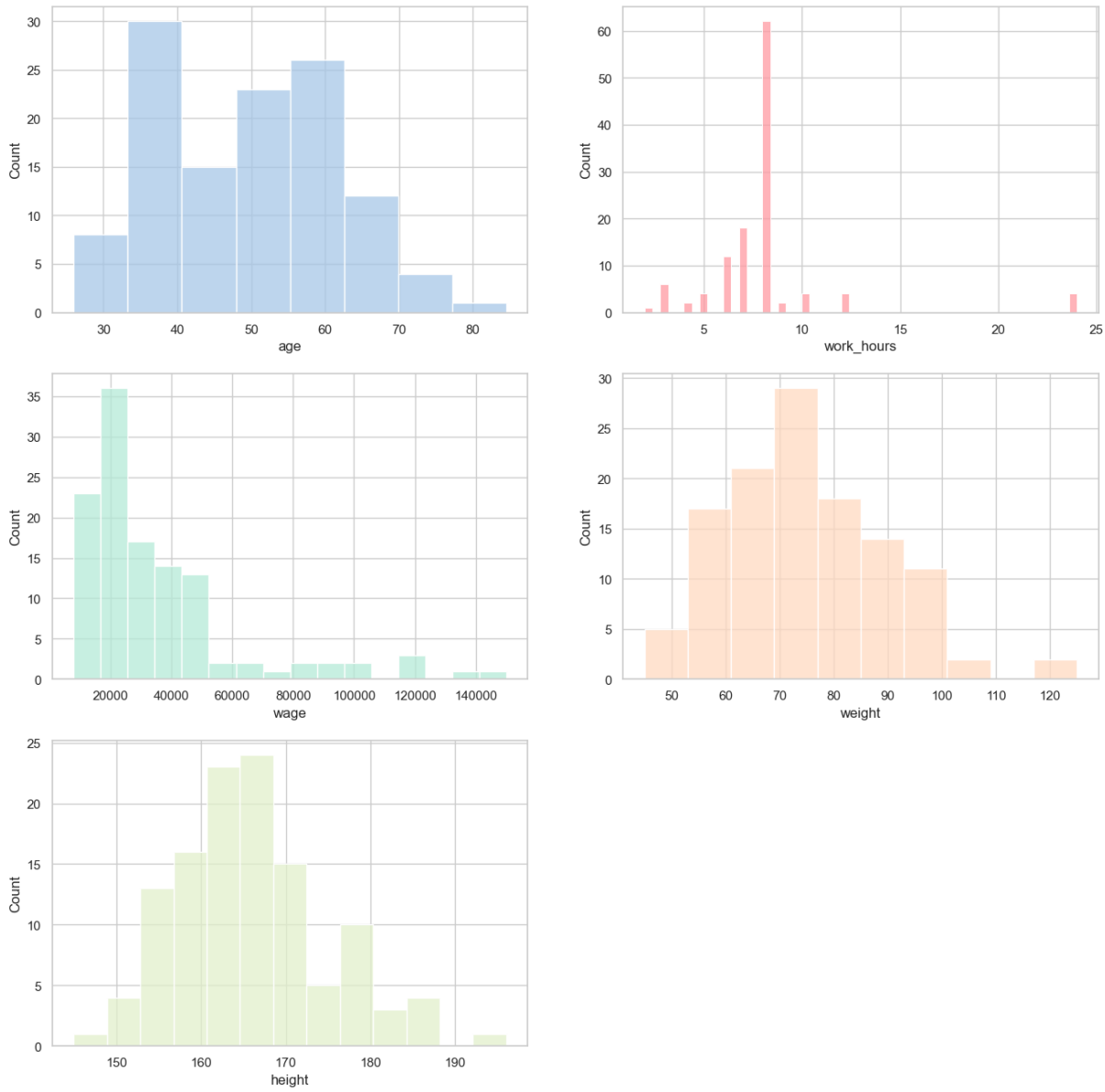


Рис. 2. Гистограммы распределения для количественных переменных

Table 4. Skewness и Kurtosis

Feature	Skewness	Kurtosis
Age	0.223478	-0.700189
Work Hours	3.171247	13.017113
Wage	2.100988	4.452849
Weight	0.565339	0.348146
Height	0.525942	0.286510

1. Симметричность:

Как и предполагалось выше (почти верно), по графикам и таблице видно:

Возраст, рост и вес примерно симметричны, с небольшими хвостами справа ($Skewness > 0$, но близок к нему)

Зарплата и рабочие часы скошена вправо ($Skewness$ сильно > 0)

2. Островершинность:

Рабочие часы и зарплата определенно островершинные ($Kurtosis > 0$)

Рост и вес вроде тоже близки к островершинности нормального распределения ($Kurtosis \approx 0$)

Возраст туповершинный ($Kurtosis < 0$)

7) Как распределены респонденты в Вашей выборке по уровню образования? Постройте гистограмму.

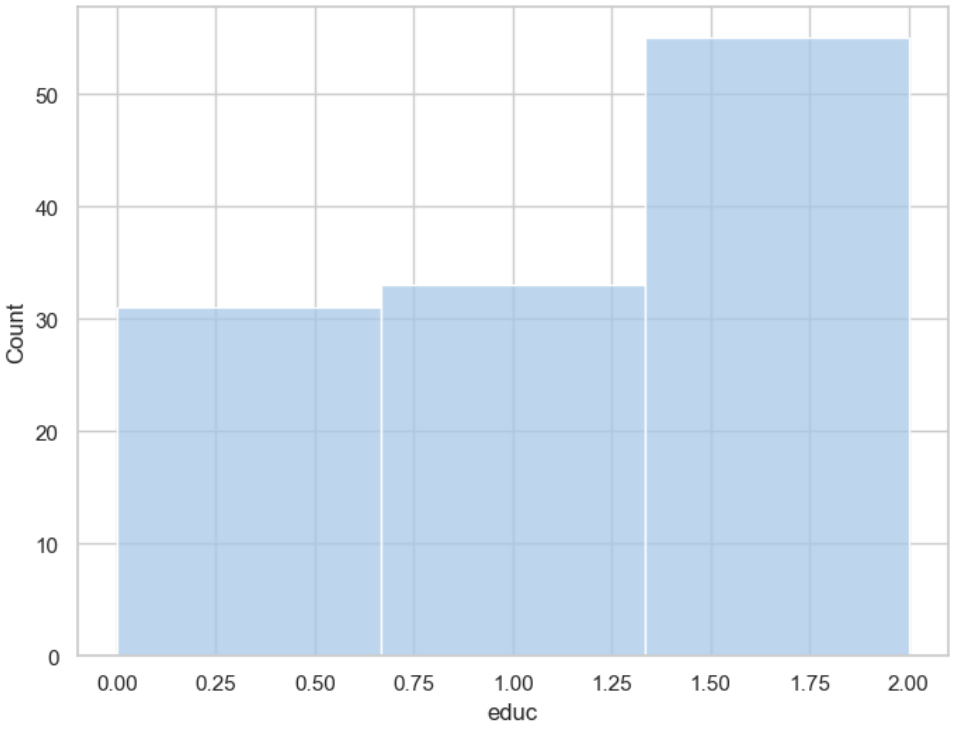


Рис. 3. Распределение по уровню образования

Почти половина респондентов окончили Университет, академтю или институт, что логично для сферы "Наука и культура это радует. Почти поровну закончили школу и техникумы. Старнно что в науке столько людей закончивших только школу, надеюсь они вундеркинды и хорошо знают своё дело

8) Постройте корреляционную таблицу для всех переменных в Вашей выборке кроме отрасли, региона и года. Проинтерпретируйте результаты.

В нашей выборке у всех респондентов есть хотя бы один ребенок, поэтому не будем их учитывать в корреляционной матрице, иначе она ломается.

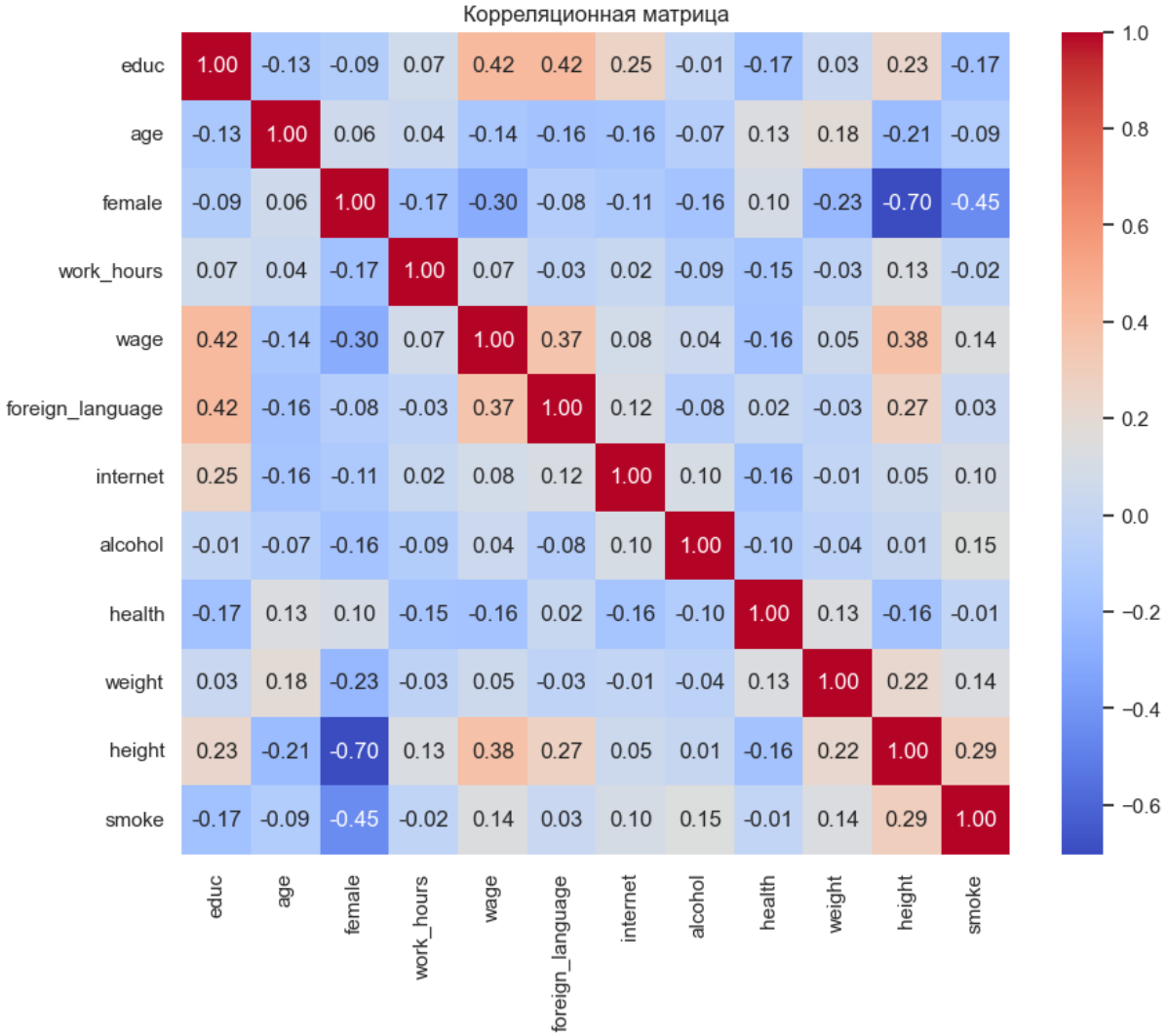


Рис. 4. Корреляционная таблица

Смею предположить, что мы будем прогнозировать зарплату, поэтому рассмотрим на зависимости с ней:

Самую большую по модулю корреляцию имеет образование, затем рост, знание иностранного языка и в конце принадлежность к женскому полу.

Увидеть высокую корреляцию с образованием в сфере наука и культура весьма ожидаемо, она положительно, следовательно образованные люди получают зарплаты выше.

Рост - удивительная связь, сложно что-то про это сказать. Либо Высоких люди привлекательнее и на них обращают больше внимания, либо это просто случайность.

Отрицательная взаимосвязь с женским полом наверняка связана со стереотипом, что женщины глупее (я такое осуждаю, 10 раз не согласен), либо боятся что она может выйти в дикрет и выпасть из рабочего графика, а платить ей нужно

Знание английского языка положительно сказываются на зарпалте, поскольку многие материалы и статьи

пишуться на английском языке, также есть возможность получать больше информации из разных источников, работать с зарубежными партрёрами и коллегами

Посмотрим на другие интересные корреляции:

work hours и female = -0.17: означающая что продолжительность рабочего дня женщин меньше, что также может сказываться на зарплате (ещё одна причина)

age и internet = -0.16: чем старше респондент, тем меньше он пользуется интернетом, что тоже может повлиять, поскольку в эпоху цифровизации интернет - необходимая часть в сфере Наука и культура

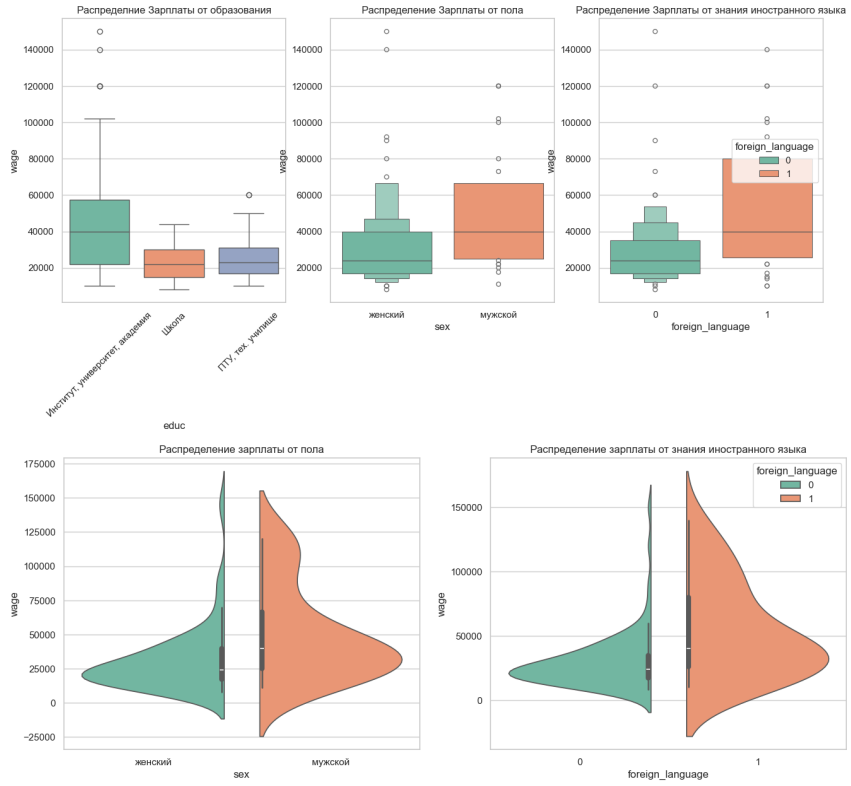
female и alcohol = -0.16: женщины меньше пьют

Есть и достаточно очевидные зависимости:

female и height = -0,64: женщины в среднем ниже мужчин height и weight = 0.47: чем выше человек, тем он больше весит

Смотря на корреляции с wage можно предположить, что мультиколлинеарность отсутствует, поскльку $|corr(wage, Features_i)| \leq 0.42$

9) Предположите зависимость заработной платы от каких-либо переменных в файле. Постройте графики, которые позволяют продемонстрировать эту зависимость.



Для образования видно, что у опрошенных с высшем образованием, работающих в сфере "Наука и культуравилка"гораздо больше, в среднем они заробатывают больше и есть исключения, которые заробатывают прям много по сравнению с людьми, имющими среднее законченное образование, и подавно лучше чем у имющих школьное. Это вполне логично для нашей сферы исследования

График для $female$ и $foreign_language$ работает плохо.

Возможные причины:

1. Малое количество данных:

Если для категории $female$ = 1 и $foreign_language$ = 1 мало данных, это может привести к сжатой или "шумной"визуализации.

2. Проблемы с выбросами:

Наличие выбросов сильно влияет на визуализацию распределений, особенно в boxplot или boxenplot.

3. Повторяющиеся значения:

Если в данных для $female$ = 1 и $foreign_language$ = 1 много одинаковых значений, график может выглядеть "компрессированным так как медиана, квартили и точки выбросов совпадают.

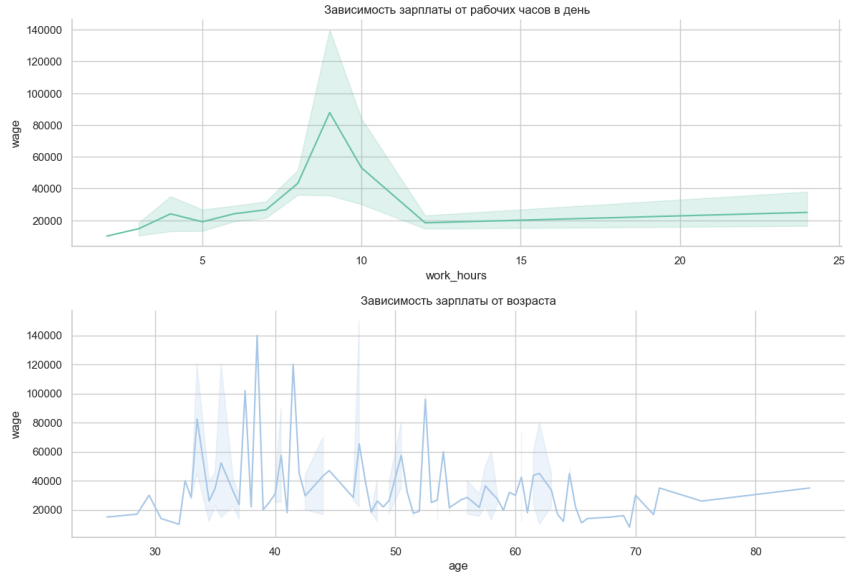
4. Слишком высокая плотность:

Если значения распределены в узком диапазоне, boxenplot может визуализировать это с "группировкой"областей, что делает график неинформативным.

В ходе проверки выяснилось, что это из-за слишком очень большого разброса в данных и малым количеством наблюдений в выборке, поэтому посмотрим на более репрезентативные графики ниже.

На них видно, что мужчины чаще получают высокие зарплаты чем женщины (в своём соотношении по количеству мужчин в опросе), опять таки по причинам указаным выше. Получаем отрицательную зависимость от признака $female$

Обратная ситуация обстоит с людьми, знающими иностранный язык, зависимость от $foreign_language$ уже положительная



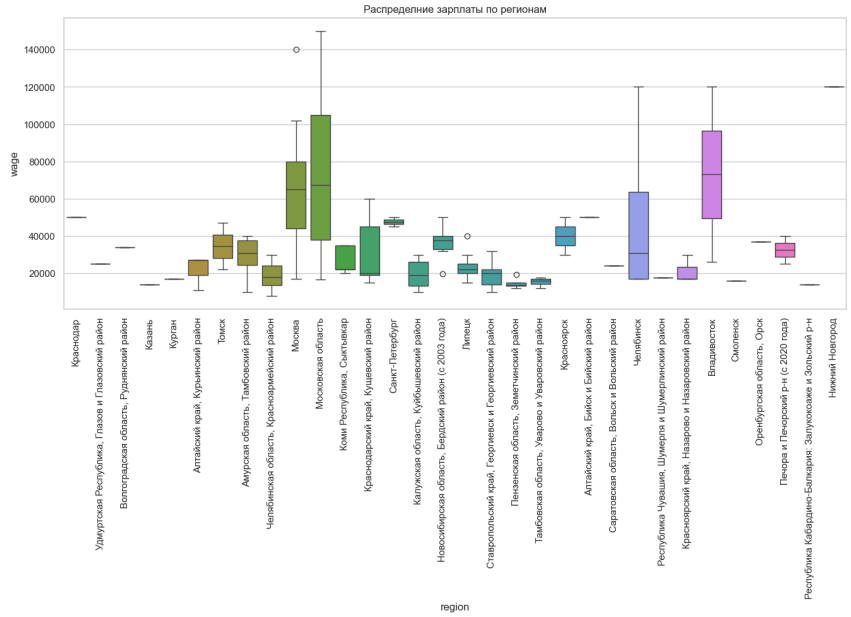
Ну вполне логичное заключение, что чем больше работаешь - больше получаешь, всё логично

Единсвенное, у людей с рабочим днём в 8-11 часов, очень большая дисперсия (бледно-зелёный), но это скорее из-за меленького количества данных

Также есть 4 человека, которые работают 24 часа в сутки, тут 3 варината, ошибка заполнения данных(выброс), либо это разнорабочий, возможно охранник в лаборатории или другом иследовательском проекте, либо подопытный, во всех случаях понтяно почему маленькая зп.

В целом зависимости прям такой нет (хотя я предполагал что возраст описывает опыт и знанния), но после $\approx$ 33 лет зарплата выше, ну потому что опыт и всё такое, дальше зп колеблятся и с 65 лет падают, возможно из-за паденя работоспособности, потом опять растёт, видимо очень уважаемые учёные

P.S зависимость от роста рассматривать не будем, в ней на мой взгляд логики нет.



Предположения что некоторые регионы спонсируются больше чем остальные. Получаем что в Московской области, Москве и Владивостоке самые высокие нижнии значения "вилки"зп и самые выские средние значения. Поэтому если включать их в модель, можно добавить как dummy переменную

10) Оцените линейную модель, которая объясняет заработную плату ($wage$) возрастом (age), наличием высшего образования ($high$), полом ($female$), наличием детей ($is\_children$), курением ($smoke$) и константой. Проинтерпретируйте полученные результаты. Все ли коэффициенты оказались значимы? Выпишите уравнение оцененной модели.

OLS Regression Results

Dep. Variable:	wage	R-squared:	0.273
Model:	OLS	Adj. R-squared:	0.248
Method:	Least Squares	F-statistic:	10.72
Date:	Sun, 24 Nov 2024	Prob (F-statistic):	2.08e-07
Time:	20:33:13	Log-Likelihood:	-1367.0
No. Observations:	119	AIC:	2744.
Df Residuals:	114	BIC:	2758.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
age	-128.8333	186.065	-0.692	0.490	-497.426	239.759
high	2.38e+04	4590.958	5.183	0.000	1.47e+04	3.29e+04
female	-1.466e+04	5960.579	-2.459	0.015	-2.65e+04	-2848.562
is_children	4.089e+04	1.16e+04	3.523	0.001	1.79e+04	6.39e+04
smoke	6970.9357	6422.989	1.085	0.280	-5752.955	1.97e+04

Omnibus:	48.026	Durbin-Watson:	1.840
Prob(Omnibus):	0.000	Jarque-Bera (JB):	122.154
Skew:	1.573	Prob(JB):	2.98e-27
Kurtosis:	6.840	Cond. No.	290.

Рис. 5. Результаты первой модели

Уравнение регрессии:

$$wage = 40889.57 - 128.83 \cdot age + 23795.93 \cdot high - 14656.42 \cdot female + 6970.94 \cdot smoke$$

Значимыми оказались: наличие высшего образования, пол, наличие детей (константа в модели) - низкий $p - value$ (0; 0.015; 0.001) Не не отвергаем гипотезу $H_1 : \beta_i \neq 0$

Незначимые: возраст, курение - высокий $p - value$ (0.49; 0.28) Не отвергаем гипотезу о незначимости $H_0 : \beta_i = 0$

Интерпретация:

При увеличении возраста на один год, зарплата уменьшится примерно на 128.83 рублей

При наличии высшего образования, зарплата увеличится примерно на 23795.93 рублей

Если ваш пол женский, зарплата уменьшается примерно на 14656.42 рублей

Если у вас есть хотя бы один ребёнок, зарплата увеличится примерно на 40889.57 рублей

Если вы курите, зарплата увеличится примерно на 6970.94 рублей

Очень странная модель, но ладно

11) Выполните тест на адекватность этой модели и сделайте выводы.

Гипотеза:

$$H_0 : \beta_2 = \ldots = \beta_k = 0$$

$$H_1 : \beta_1^2 + \ldots + \beta_k^2 > 0$$

F-статистика считается, как:

$$F = \frac{(RSS_R - RSS_{UR})/(k-1)}{RSS_{UR}/(n-k)} \sim F(k-1, n-k).$$

По результатам вычислений для нашей модели:

$F_{obs} = 10.71,\ pvalue \approx 0.000$

Мы не отвергаем гипотезу H_1 на пятипроцентном уровне значимости, то есть модель адекватна.

12) Сформулируйте и протестируйте гипотезу для одного из коэффициентов модели. Дайте содержательную и количественную интерпретацию полученных результатов.

Гипотеза:

$$H_0 : \beta_{age} = 120$$

$$H_1 : \beta_{age} \neq 120$$

Я очень хочу проверить что опыт = возраст = больше зп

t-статистика считается, как:

$$t = \frac{\hat{\beta_{age}} - 100}{\sqrt{\widehat{Var(\hat{\beta_{age}})}}} \sim t(n-k)$$

По результатам вычислений для нашей модели:

$t_{obs} = -1.3373,\ pvalue = 0.18$

Смотря на $pvalue$ олучаем, что гипотеза H_1 не отвергается на любом разумном уровне значимости, что между wage и age отрицательная корреляция, поэтому мы отвергаем гипотезу H_0

я расстроен

13) Сформулируйте и протестируйте гипотезу о нескольких коэффициентах модели. Дайте содержательную и количественную интерпретацию полученных результатов.

Гипотеза:

$$H_0 : \beta_{female} + \beta_{smoke} = -10000$$

$$H_1 : \beta_{age} + \beta_{female} \neq -10000$$

t-статистика считается, как:

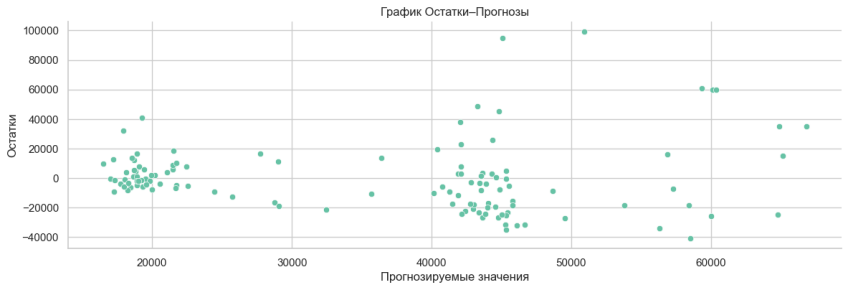
$$t = \frac{\hat{\beta_{female}} + \hat{\beta_{smoke}} + 12000}{\sqrt{\widehat{Var(\beta_{female})} + \widehat{Var(\beta_{smoke})} + 2\widehat{Cov(\beta_{female}, \beta_{smoke})}}} \sim t(n-k)$$

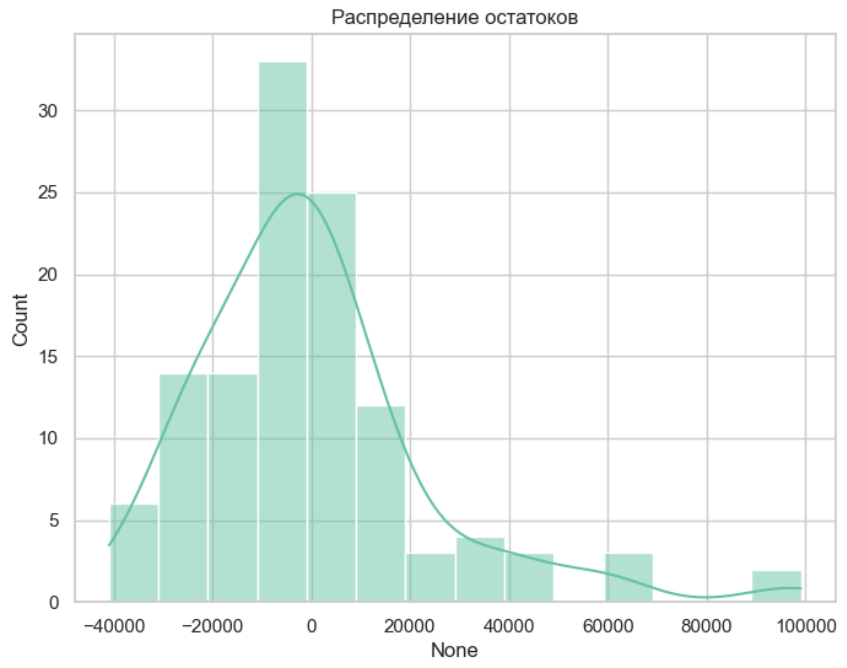
По результатам вычислений для нашей модели:

$t_{obs} = 0.2187,\ pvalue \approx 0.83$

Смотря на $pvalue$ олучаем, что гипотеза H_0 не отвергается на любом разумном уровне значимости. Курящим женщинам платят меньше, возможно из-за предвзятого восприятия и убеждений, что женщинам не красиво курить, ну либо она тратит много на это вермени

14)Постройте график «остатки–прогнозы». Сделайте вывод.





Для меньших прогнозов (до 20,000) остатки более компактны и хаотичны, что может говорить о гомоскедастичности в этой области. Однако для более высоких прогнозов (> 30,000) остатки начинают проявлять некоторые выбросы и смещение вниз. Это может указывать на гетероскедастичность.

На графике остатки распределены относительно хаотично, но можно заметить, что для высоких прогнозов (от 50,000 и выше) остатки концентрируются ниже линии y = 0, что может указывать на недооценку модели для больших значений.

На графике видно, что распределение смещено вправо (положительная асимметрия), но всё ещё похоже на нормальое

15)Оцените модель из п. 10, оставив в ней только значимые коэффициенты. Выпишите уравнение оцененной модели. Сравните результаты с моделью из п. 10. Какие критерии для сравнения моделей здесь стоит использовать?

OLS Regression Results

Dep. Variable:	wage	R-squared:	0.262
Model:	OLS	Adj. R-squared:	0.249
Method:	Least Squares	F-statistic:	20.55
Date:	Sun, 24 Nov 2024	Prob (F-statistic):	2.30e-08
Time:	20:33:14	Log-Likelihood:	-1368.0
No. Observations:	119	AIC:	2742.
Df Residuals:	116	BIC:	2750.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
high	2.321e+04	4437.229	5.231	0.000	1.44e+04	3.2e+04
female	-1.785e+04	5282.110	-3.379	0.001	-2.83e+04	-7385.949
is_children	3.861e+04	5175.236	7.461	0.000	2.84e+04	4.89e+04

Omnibus:	52.544	Durbin-Watson:	1.849
Prob(Omnibus):	0.000	Jarque-Bera (JB):	152.755
Skew:	1.663	Prob(JB):	6.76e-34
Kurtosis:	7.444	Cond. No.	4.49

Рис. 6. Результаты новой модели со значимимы признаками

Уравнение модели:

$$wage = 38609.99 + 23209.30 \cdot high - 17847.83 \cdot female$$

Сравнение моделей:
1) R^2 и R^2_{adj}
Первая модель: $R^2 = 0.273,\ R^2_{adj} = 0.248$
Вторая модель: $R^2 = 0.262,\ R^2_{adj} = 0.249$

Видно, что обычный R^2 упал, в то время така R^2_{adj} повысился, что означает, что незначимые признаки увеличивают R^2 в первой модели, но вторая без них лучше объясняет дисперсию

2) AIC и BIC (информационные критерии)

Первая модель: $AIC = 2744,\ BIC = 2758$

Вторая модель: $AIC = 2742,\ BIC = 2750$

Модель без незначимых признаков немного лучше, чем первая модель

3) F-тест

Первая модель: Prob (F-statistic) = 2.08e-07

Вторая модель: Prob (F-statistic) = 2.30e-08

Prob F упал почти в 10 раз, стало лучше. Модель с большей вероятностью значима

16) Протестируйте наличие выбросов в модели с помощью известных Вам методов. Если они есть, то как их учесть в модели? Проведите коррекцию.

Используем стьюдентизированные остатки найдем их выбросы, затем DFFITS и их выбросы. Уберём все выбросы

Формула для вычисления стьюдентизированных остатков

$$r_i = \frac{e_i}{\sqrt{MSE(1 - h_{ii})}}$$

где:

- e_i: остаток для i-го наблюдения,
- MSE: среднеквадратическая ошибка ($\frac{\text{RSS}}{n-k}$),
- h_{ii}: диагональный элемент матрицы $X(X'X)^{-1}X'$.

Классически threshold = 2 и возьмем все строки датасета со значиями остатков по модулю больше двух.

Формула для вычисления DFFITS:

$$DFFITS_i = \frac{r_i\sqrt{h_{ii}}}{\sqrt{1 - h_{ii}}}$$

где:

- r_i: стьюдентизированные остатки,
- h_{ii}: диагональный элемент матрицы $X(X'X)^{-1}X'$.

Если $DFFITS_i > 2\sqrt{\frac{k}{n}}$, то i-е наблюдение может быть выбросом.

Используя первый и второй критерий найдем подозреваемых на выбросы:

Было 119 наблюдений, стало 109

Давайте скорректируем модель, удалив выбросы, и обучим на признаках из пункта 10 и 15:

OLS Regression Results

Dep. Variable:	wage	R-squared:	0.226
Model:	OLS	Adj. R-squared:	0.196
Method:	Least Squares	F-statistic:	7.578
Date:	Sun, 24 Nov 2024	Prob (F-statistic):	2.13e-05
Time:	20:38:06	Log-Likelihood:	-1199.6
No. Observations:	109	AIC:	2409.
Df Residuals:	104	BIC:	2423.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
age	72.8328	118.526	0.614	0.540	-162.209	307.875
high	1.398e+04	2995.881	4.665	0.000	8035.391	1.99e+04
female	-1.091e+04	4298.325	-2.539	0.013	-1.94e+04	-2387.651
is_children	2.906e+04	7752.634	3.748	0.000	1.37e+04	4.44e+04
smoke	1253.0252	4414.557	0.284	0.777	-7501.208	1e+04

Omnibus:	24.613	Durbin-Watson:	1.936
Prob(Omnibus):	0.000	Jarque-Bera (JB):	33.986
Skew:	1.129	Prob(JB):	4.17e-08
Kurtosis:	4.543	Cond. No.	309.

Рис. 7. Результаты модели задания 10 без выбросов

OLS Regression Results

Dep. Variable:	wage	R-squared:	0.223
Model:	OLS	Adj. R-squared:	0.208
Method:	Least Squares	F-statistic:	15.17
Date:	Sun, 24 Nov 2024	Prob (F-statistic):	1.61e-06
Time:	20:39:11	Log-Likelihood:	-1199.8
No. Observations:	109	AIC:	2406.
Df Residuals:	106	BIC:	2414.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
high	1.36e+04	2883.723	4.715	0.000	7878.858	1.93e+04
female	-1.147e+04	3668.089	-3.128	0.002	-1.87e+04	-4202.302
is_children	3.353e+04	3458.983	9.694	0.000	2.67e+04	4.04e+04

Omnibus:	24.658	Durbin-Watson:	1.927
Prob(Omnibus):	0.000	Jarque-Bera (JB):	34.236
Skew:	1.126	Prob(JB):	3.68e-08
Kurtosis:	4.572	Cond. No.	4.75

Рис. 8. Результаты модели задания 15 без выбросов

Выводы: Объясняющая способность R^2: Модель до удаления выбросов имеет более высокий R^2, но это может быть связано с тем, что выбросы искусственно увеличивают объясняющую способность модели.

Качество модели (AIC и BIC):

Модель после удаления выбросов имеет значительно лучшие значения AIC и BIC. Это говорит о том, что модель стала более компактной и менее склонной к переобучению.

Распределение остатков:

Уменьшение асимметрии (Skew) и островершинности (Kurtosis) в модели после удаления выбросов указывает на то, что остатки стали ближе к нормальному распределению.

Влияние предикторов:

После удаления выбросов коэффициенты уменьшились, что говорит о том, что выбросы сильно влияли на оценки. Модель без выбросов даёт более реалистичные коэффициенты.

17) Постройте прогноз заработной платы для одного индивида с заданными вами характеристиками для него. Постройте 95% доверительный интервал для прогнозного значения.

Спрогнозируем значение зарплаты мужчины с ребёнком, высшем образованием.

Доверительный интервал для индивидуального прогнозного значения y_0 в точке x_0 примет вид:

$$\tilde{y}_0 - t_{n-k}^{0.975} \cdot m_{\tilde{y}_0} \leq y_0 \leq \tilde{y}_0 + t_{n-k}^{0.975} \cdot m_{\tilde{y}_0},$$

где среднеквадратичная ошибка прогноза $\tilde{y}_0$ вычисляется по формуле:

$$m_{\tilde{y}_0} = \hat{\sigma}\sqrt{1 + x_0'(X'X)^{-1}x_0}.$$

где:

- x_0 – вектор-столбец значений наблюдения, для которого считается прогноз,
- x_0' – вектор-строка значений наблюдения, для которого считается прогноз,
- k – это количество регрессоров.

Получаем, что зарплата индивида с заданными вами характеристиками лежит в пределах от 51709.59 до 71929.00 рублей

Предсказание зарплаты: 61819.29 рублей