Высшая школа экономики
(Научно-исследовательский университет)
Факультет компьютерных наук

Домашнее задание
по дисциплине: “Методы прикладной статистики”

Статистическое исследование доходов, расходов и сбережений населения

Студент: Рябыкин Алексей Сергеевич
Преподаватель: Сиротин Вячеслав Павлович
Оценка: _

Москва 2022

Оглавление

1. Постановка задачи

Выбранные данные: Российский мониторинг экономического положения и здоровья населения НИУ ВШЭ (RLMS-HSE).

Этапы работы:

1. Произвести предварительный анализ выбранных количественных признаков на основе расчета для каждого из них базовых выборочных характеристик (статистик) центральной тенденции и разброса: квантилей, моды, среднего значения, дисперсии и среднего квадратического отклонения. Построить непараметрическую модель эмпирического распределения для одного из количественных признаков;
2. Проанализировать репрезентативность выборки и определить возможность использования различных методов ее коррекции: определить доли и структуру пропущенных значений и их возможное влияние на статистические выводы. Предложить меры работы с данными, содержащими пропуски и экстремальные значения. Исследовать возможность корректировки выборки с целью обеспечения репрезентативности путем перевзвешивания наблюдений;
3. Построить доверительные интервалы для параметров генеральной совокупности на основе выборочных данных: генерального среднего и генеральной дисперсии для каждого из количественных признаков и генеральной доли для каждого из качественных признаков;
4. Произвести проверку гипотез о значениях некоторых параметров генеральной совокупности, например, о равенстве значений признака для определенных групп населения, о соответствии значения признака некоторому декларируемому уровню;
5. Проанализировать взаимосвязь между признаками, характеризующими исследуемые объекты, и определить ее значимость. Сформулировать выводы по результатам проведенного исследования.

Целью исследования можно назвать формирование понимания и формализацию закономерностей в финансовой и имущественной обеспеченности населения России.

Регионом исследования был выбран Москва.

2. Описание данных и выбранных признаков

Доходы, исходя из данных и международных принципов их дифференциации, могут быть представлены следующими видами:

Доходы	Виды доходов в исследовании
Первичные доходы	Оплата труда
	Продажа сельскохозяйственной продукции
	Личное хозяйство
	Предпринимательская деятельность
Собственность	Аренда недвижимости
	Проценты
Трансферты	Пенсии
	Стипендии
	Пособия
	Алименты
	Пособия по безработице
	Другие доходы
Прочие поступления	Продажа личного имущества
	Продажа недвижимости

Таблица 1 – Доходы и их виды

С другой стороны, среди расходов и имуществ населения можно выделить следующие группы для анализа:

Расходы	Примеры
Потребительские расходы	Продукты питания
Платежи и взносы	Налоги, связь, интернет
Приобретение имущества	Квартиры, гараж и прочее
Денежные сбережения	Вклады, сбережения в валюте

Таблица 2 – Расходы и их виды

Активы	Примеры
Реальные активы	Квартира, дача, машина
Оборотные активы	Одежда, наличные
Финансовые активы	Арендное имущество

Таблица 3 – Типы имущества

Для анализа были выбраны более-менее заполненные данные, связанные с представленными ранее.

3. Предварительный анализ данных

Данные содержат результаты опроса 413 респондентов из Москвы по практически 1.5 тысячам вопросам различного характера.

Столбец	Описание	Количество п.з.	Доля п. з.
z_nfm	Количество членов семьи	0	0.0%
zc1.1	Стоимость Вашего жилья?	0	0.0%
zc6	Полезная площадь Вашего жилья?	0	0.0%
zc5	Жилая площаль Вашего жилья?	0	0.0%
zc9.7.2a	В наличии отечественный легковой автомобиль	0	0.0%
zc9.7.3a	В наличии легковой автомобиль иностранной модели?	0	0.0%
zc9.7.1a	В наличии грузовой автомобиль?	0	0.0%

Продолжение на следующей странице

Столбец	Описание	Количество п. з.	Доля п. з.
zc9.8a	В наличии мотоцикл, мотороллер, моторная лодка?	0	0.0%
zc9.101a	В наличии дача или другой дом, садовый домик?	0	0.0%
zc9.12a	В наличии другая квартира или часть квартиры?	0	0.0%
ze1.1c	Траты за 7 дней на белый хлеб	57	13.8%
ze4	Траты на еду	0	0.0%
ze9.8b	Траты на оплату мобильной связи	13	3.1%
ze11	Траты на квартиру (коммунальные, аренда)	16	3.9%
ze12	Наличие задолженностей по квартире?	0	0.0%
ze44	Практикуют раздельный сбор мусора?	0	0.0%
ze14	Ваша семья в течение 30 дней давала деньги в долг?	0	0.0%
ze16	Ваша семья в течение 30 дней откладывала сбережения?	0	0.0%
zf12_a	Как долго смогли бы жить только за счет сбережений?	0	0.0%
zf14.8	Долги по кредитам	0	0.0%
zf14.12	Должны ли Вашей семье?	0	0.0%
zf14	Денежный доход	0	0.0%

Таблица 4 – Доля пропущенных значений

Замечание

Описание некоторых столбцов здесь и впредь было изменено для адекватного отображения таблиц, рисунков, диаграмм.

Среди выбранных признаков доля пропущенных значений не очень велика, заполнение для категориальных возможно с помощью категориальных значений “ОТКАЗ ОТ ОТВЕТА”, “НЕТ ОТВЕТА”. Их природа (как пропусков, так и перечисленных категориальных значений) может лежать в неуверенности респондентов в ответе или в нежелании разглашать личную информацию.

Проведем предварительный анализ выбранных количественных признаков, рассчитав для каждого выборочные характеристики центральной тенденции и разброса:

Признак	Описание	q_{25}	q_{75}	mean	mode	std	variance
z_nfm	Количество членов семьи	1	4	2.56	1	1.5	2.26
zc6	Полезная площадь	42.4	60.225	53.24	52	15.4	237.2
ze11	Квартплата (+коммунальные)	5000	8000	7021	7000	3814.8	14552570
zc5	Площадь квартиры (жилая)	26.375	42	34	30	12	148
ze1.1	Примерная стоимость жилья	6750000	12000000	9478285.7	10000000	3941801.8	15537801642036
ze1.1c	Траты на белый хлеб	45	122	105	40	90.5	8182.5
ze4	Траты на питание	15000	30000	23309	20000	12910.2	166673104.7
ze9.8b	Траты на мобильную связь	500	1500	1195.6	1000	1300.7	1691897.7
zf14	Денежный доход	43000	117225	88354	50000	67447.5	4549163351

Таблица 5 – Базовые выборочные характеристики для количественных признаков

Построим распределение для доходов и расходов. Как доходы, так и расходы есть совокупность денежных средств, полученных из различных источников или затраченных на различные цели. Поэтому распределение трат (или доходов) домохозяйств может иметь вид аддитивной смеси распределений. Предположим, что закон распределения расходов будет смесью логарифмически-нормальных распределений:

$$f(\ln y) = \sum_{i=1}^{n} q_i f(\ln y, \mu_i, \sigma_i),$$

где n – предполагаемое количество групп, q_i – доля объектов i-й группы, причем $\sum_{i=1}^{n} q_i = 1$. Построим эмпирическое распределение логарифма затрат:

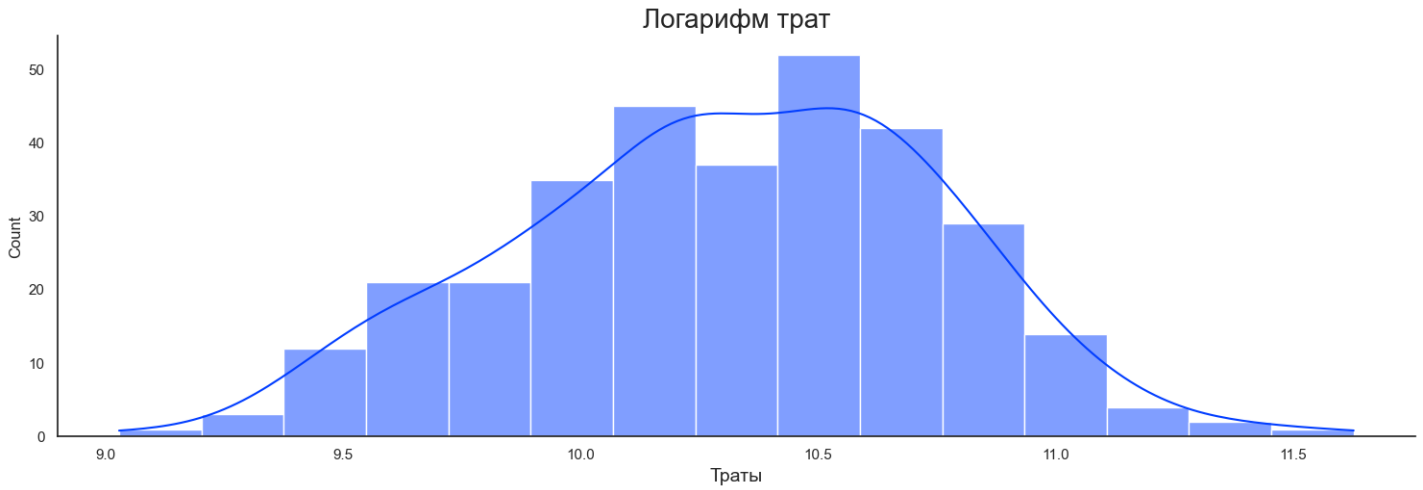

Рис. 1 – Гистограмма эмпирического распределения логарифма затрат

По критерию Шапиро-Уилка логарифм затрат подчиняется нормальному закону распределения на уровне значимости 0.1. Так же можно посмотреть на распределения по типам затрат:

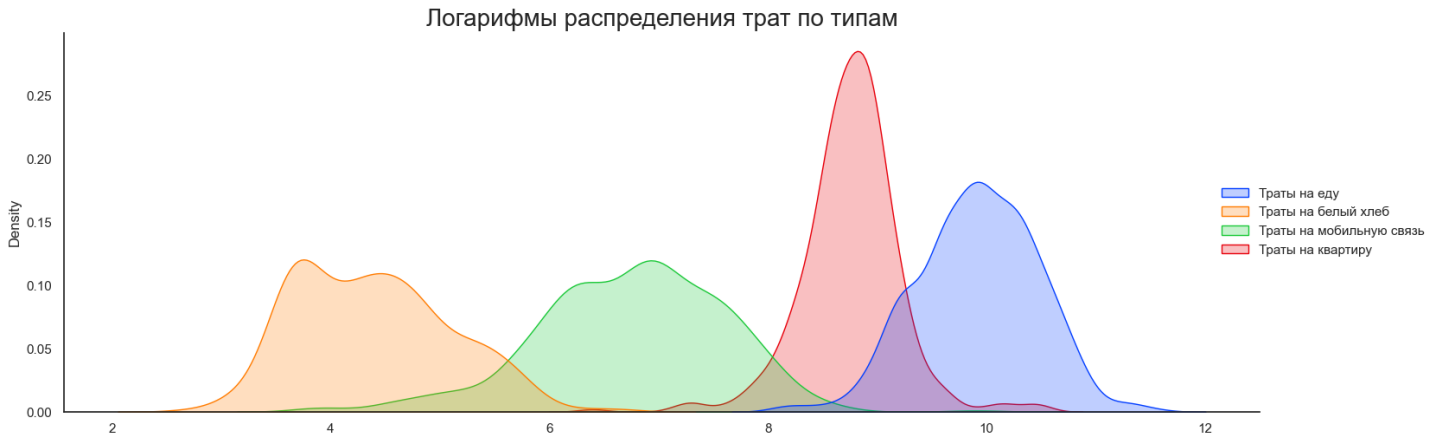

Рис. 2 – Распределения логарифмов затрат для разных категорий

Видим колокообразные распределения затрат для домохозяйств. Для количественных данных пропуски можно заполнить средними значениями, например, для расходов, средними разного рода (математическим ожиданием, модой, медианой), для площади квартиры или стоимости квартиры (пропущенных значений нет, однако иногда респонденты отвечали в соответствии с перечисленными категориальными значениями) так же можно заполнить средними или выделить схожие группы по площади для заполнения стоимости.

Для дохода можем наблюдать значительные выбросы. Природа этих выбросов может зависеть от двух аспектов. Первый: неравенство доходов, зиждящееся на банальном социальном неравенстве, при котором присутствуют люди, зарабатывающие сильно больше прочих. С другой стороны, это может быть обман респондентов с различными целями (от поднятия эго до умышленного искажения информации, например, для поднятия среднего заработка).

Избавимся от выбросов и построим распределение денежных доходов. Применим фильтр Хэмпеля, который избавляется от значений, у которых разница с медианой больше, чем три медианных абсолютных отклонения.

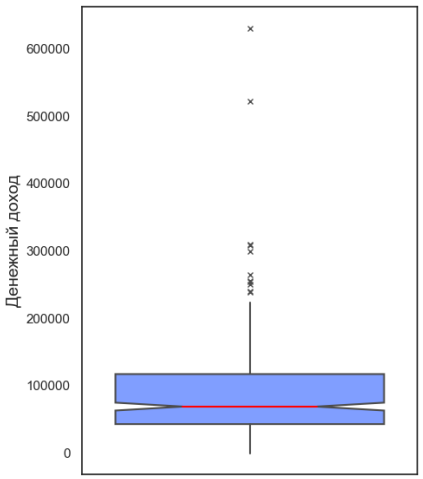

Рис. 3 – Ящик с усами для доходов

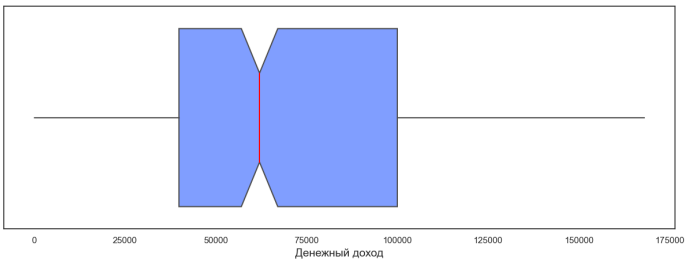

Рис. 4 – Ящик с усами после избавления от выбросов

Теперь можем построить распределение доходов для населения:

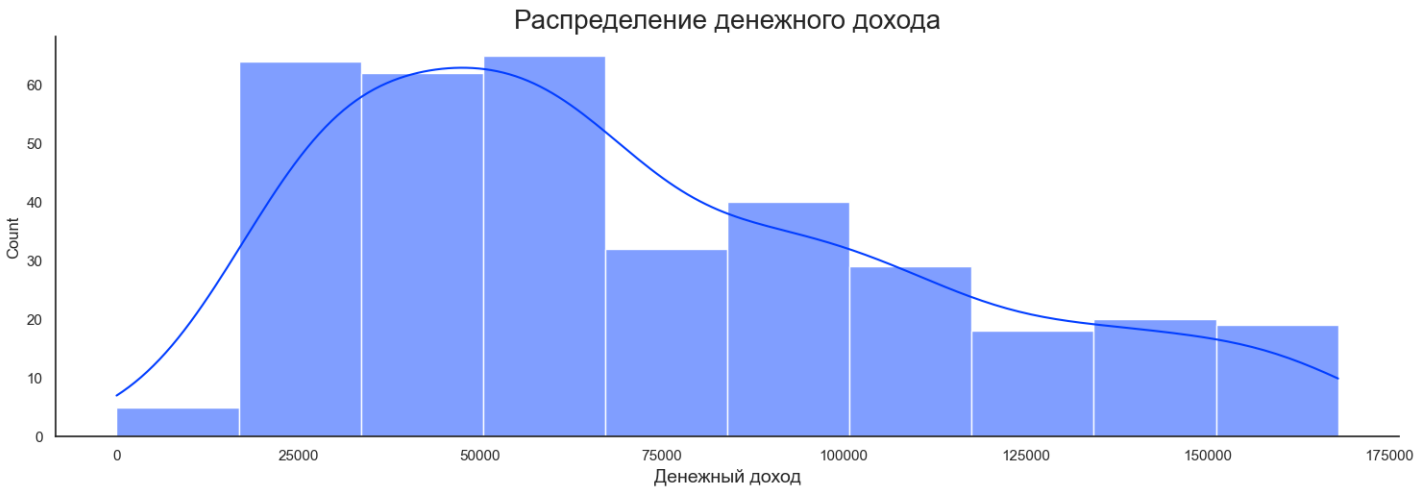

Рис. 5 – Гистограмма эмпирического распределения денежного дохода

О нормальности распределения речи не идет. Стоит отметить, что эти данные точно не репрезантативны для целой страны, особенно беря во внимание показатели минимального размера оплаты труда по Российской Федерации. Однако данные репрезентативны для выбранного региона.

Наличие значений признаков вида “ЗАТРУДНЯЮСЬ ОТВЕТИТЬ” или “НЕТ ОТВЕТА” сильно сказывается на репрезентативности. Предположение по поводу сущности их бытия было выше. К счастью, такие значения составляют малую долю ответов респондентов, поэтому решением было просто их опустить (особенно относительно количественных признаков). Для категориальных данных особенно поддается объяснению в виде нежелания разглашать личную информацию признак “Наличия сбережений”, в котором их содержится больше, чем в остальных столбцах.

Посмотрим на категориальные переменные. Построим *pie-plots* для каждой.

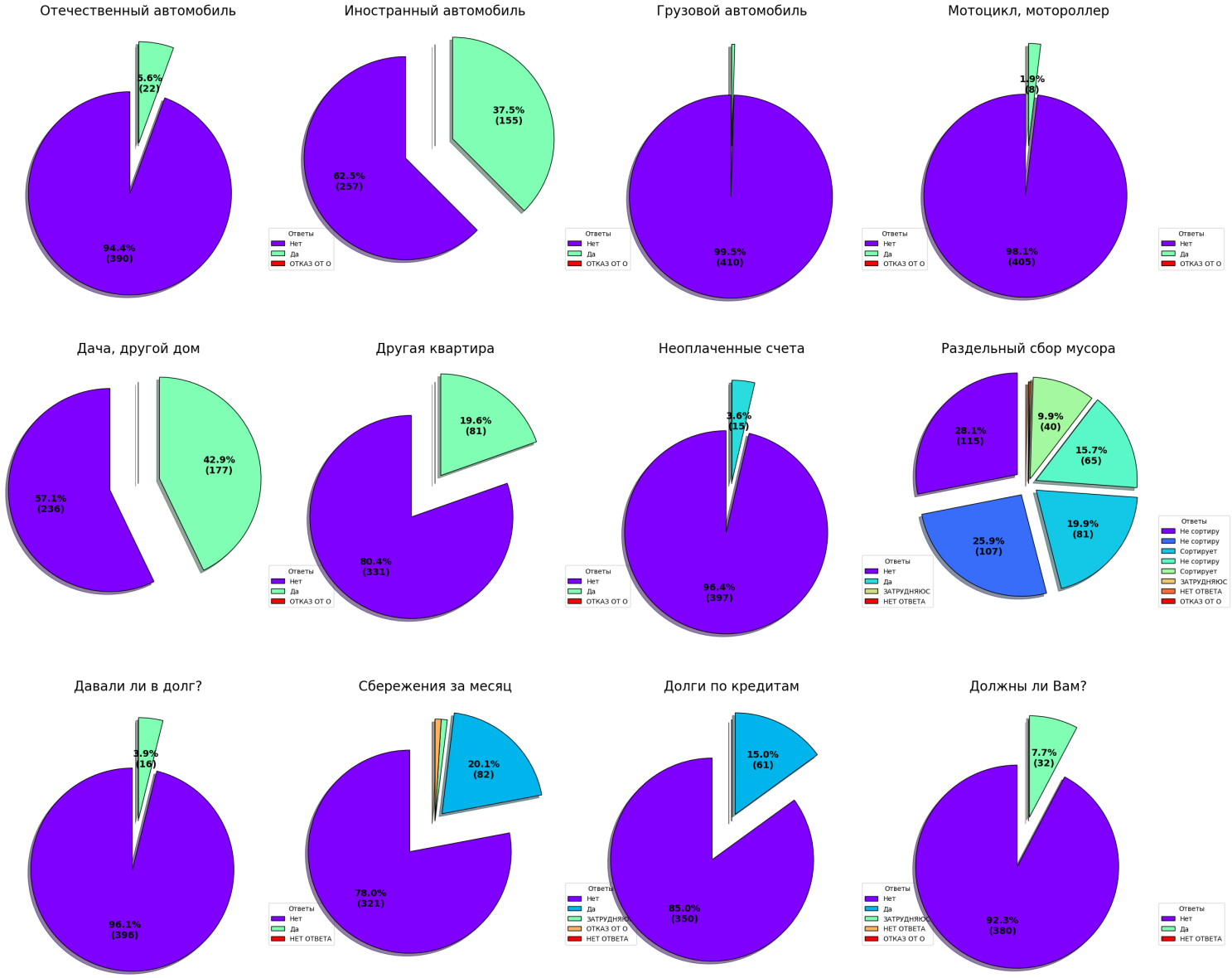

Рис. 6 – Распределение ответов для категориальных переменных

4. Доверительные интервалы для параметров генеральной совокупности

Построим доверительные интервалы для параметров генеральной совокупности. Так как наша выборка достаточно большая, то будем пользоваться t-статистикой Стьюдента для оценки генерального среднего для количественных признаков. Тогда доверительный интервал будет иметь вид:

$$\left(\overline{x} - t_\alpha \frac{s}{\sqrt{n-1}}, \overline{x} + t_\alpha \frac{s}{\sqrt{n-1}}\right)$$

Для оценки генеральной дисперсии доверительный интервал будет иметь вид:

$$\left(\frac{s^2(n-1)}{\chi^2_{1-\frac{\alpha}{2},n-1}}, \frac{s^2(n-1)}{\chi^2_{\frac{\alpha}{2},n-1}}\right)$$

Доверительные интервалы для генерального среднего количественных величин:

Признак	Описание	Левая граница (для $\alpha = 0.95$)	Правая граница (для $\alpha = 0.95$)	Левая граница (для $\alpha = 0.99$)	Правая граница (для $\alpha = 0.99$)
z_nfm	Количество членов семьи	2.41	2.7	2.37	2.75
zc6	Полезная площадь	51.74	54.73	51.3	55.2
ze11	Квартплата (+коммунальные)	6646	7396	6528	7514.6
zc5	Площадь квартиры (жилая)	32.9	35.23	32.5	35.6
ze1.1	Примерная стоимость жилья	8894270.9	10062300.5	8710760	10245811
ze1.1c	Траты на белый хлеб	95.5	114.4	92.5	117.4
ze4	Траты на питание	22042.5	24576	21644.5	24974
ze9.8b	Траты на мобильную связь	1067.35	1323.9	1027	1364.2
zf14	Денежный доход	81711	94997	79624	97084.5

Таблица 6 – Доверительные интервалы для генерального среднего

Доверительные интервалы для генеральной дисперсии количественных величин:

Признак	Описание	Левая граница (для $\alpha = 0.05$)	Правая граница (для $\alpha = 0.05$)	Левая граница (для $\alpha = 0.01$)	Правая граница (для $\alpha = 0.01$)
z_nfm	Количество членов семьи	1.98	2.6	1.89	2.72
zc6	Полезная площадь	207.7	273.47	199.3	286.2
ze11	Квартплата (+коммунальные)	12720681	16812994.7	12203124.2	17608868.1
zc5	Площадь квартиры (жилая)	129.8	170.7	124.6	178.69
ze1.1	Примерная стоимость жилья	12727592950860.9	19398997355121.3	11973428154879.1	20843916889527.8
ze1.1c	Траты на белый хлеб	7093.6	9544.2	6788.0	10027.2
ze4	Траты на питание	145739796.9	192489648.6	139824060.2	201576712.2
ze9.8b	Траты на мобильную связь	1478433.14	1955438.8	1418139.55	2048256.78
zf14	Денежный доход	3975857266.5	5256771736	3813904373.85	5505950906.6

Таблица 7 – Доверительные интервалы для генеральной дисперсии

Доверительные интервалы для генеральной доли:

$$\left(\hat{p} - z_{1-\frac{\alpha}{2}}\sqrt{\hat{p}\,(1-\hat{p})}, \hat{p} + z_{1-\frac{\alpha}{2}}\sqrt{\hat{p}\,(1-\hat{p})}\right),$$

где $\hat{p} = \frac{M}{N}$ – выборочная доля, M – количество исследуемого признака в выборке, N – размер выборки.

Признак	Описание	Ответ респондента	Левая граница (для $\alpha = 0.05$)	Правая граница (для $\alpha = 0.05$)	Левая граница (для $\alpha = 0.01$)	Правая граница (для $\alpha = 0.01$)
zc9.7.2a	Отечественный автомобиль	НЕТ	0.91	0.96	0.9	0.969
		ДА	0.03	0.089	0.03	0.098
		ОТКАЗ ОТ ОТВЕТА	0	0.01	0	0.02
zc9.7.3a	Иностранный автомобиль	НЕТ	0.56	0.679	0.55	0.69
		ДА	0.32	0.43	0.30	0.44
		ОТКАЗ ОТ ОТВЕТА	0	0.01	0	0.02
zc9.7.1a	Грузовой автомобиль	НЕТ	0.977	0.998	0.97	0.99
		ДА	0.001	0.022	0.0007	0.029
		ОТКАЗ ОТ ОТВЕТА	0	0.01	0	0.02
zc9.8a	Мотоцикл, моторная лодка мотороллер	НЕТ	0.956	0.99	0.948	0.993
		ДА	0.0085	0.04	0.007	0.05
		ОТКАЗ ОТ ОТВЕТА	0	0.013	0	0.02
zc9.101a	Дача, другой дом	НЕТ	0.512	0.629	0.49	0.64
		ДА	0.37	0.48	0.35	0.5
		ОТКАЗ ОТ ОТВЕТА	0	0.01	0	0.02
zc9.12a	Другая квартира, часть квартиры	НЕТ	0.75	0.84	0.74	0.8547
		ДА	0.1536	0.247	0.145	0.26
		ОТКАЗ ОТ ОТВЕТА	0	0.013	0	0.02
ze44	Раздельный сбор мусора	Не сортирует	0.225	0.34	0.215	0.356
		Не считает нужным	0.2	0.32	0.19	0.33
		Сортирует 1	0.15	0.257	0.14	0.268
		Не сортирует, но хотел бы	0.1145	0.212	0.108	0.223
		Сортирует 2	0.065	0.1469	0.061	0.157
		ЗАТРУДНЯЮСЬ ОТВЕТИТЬ	0.0002	0.0222	0.00019	0.029
		НЕТ ОТВЕТА	0.0002	0.0222	0.00019	0.029

Продолжение на следующей странице

Признак	Описание	Ответ респондента	Левая граница (для $\alpha = 0.05$)	Правая граница (для $\alpha = 0.05$)	Левая граница (для $\alpha = 0.01$)	Правая граница (для $\alpha = 0.01$)
ze16	Откладывали сбережения	НЕТ	0.722	0.82	0.71	0.835
		ДА	0.155	0.256	0.147	0.268
		ЗАТРУДНЯЮСЬ ОТВЕТИТЬ	0.00288	0.032	0.0023	0.0391
		ОТКАЗ ОТ ОТВЕТА	0.00288	0.032	0.0023	0.0391
		НЕТ ОТВЕТА	0	0.015	0	0.022
zf14.8	Долги по кредитам	НЕТ	0.799	0.889	0.787	0.896
		ДА	0.11	0.2	0.1	0.212
		ЗАТРУДНЯЮСЬ ОТВЕТИТЬ	0	0.015	0	0.022
		ОТКАЗ ОТ ОТВЕТА	0	0.015	0	0.022
		НЕТ ОТВЕТА	0	0.015	0	0.022
ze14	Давали в долг	НЕТ	0.93	0.978	0.92	0.98
		ДА	0.021	0.068	0.019	0.07
		ОТКАЗ ОТ ОТВЕТА	0	0.013	0	0.02

Таблица 8 – Доверительные интервалы для генеральной доли

5. Гипотезы о значениях генеральной совокупности

Проверим следующие гипотезы:

Первый набор гипотез

- H_0: Средний денежний доход одиноких людей (количество членов семьи 1), средних семей ($2-4$ человека) и больших семей (≥ 5) не отличаются;
- H_1 Количество членов семьи влияет на денежный доход;.

Проверим на нормальность группы с помощью теста Шапиро-Уилка:

$$W = \frac{b^2}{s^2},$$

где $b^2 = \sum_{i=1}^{n} a_{n-i+1} \left(x_{n-i+1} - x_i\right)$ – квадрат оценки среднеквадратического отклонения Ллойда. Исходя из него две из трех групп не являются нормально распределенными. Построим *QQ-plot*, чтобы посмотреть, насколько всё критично:

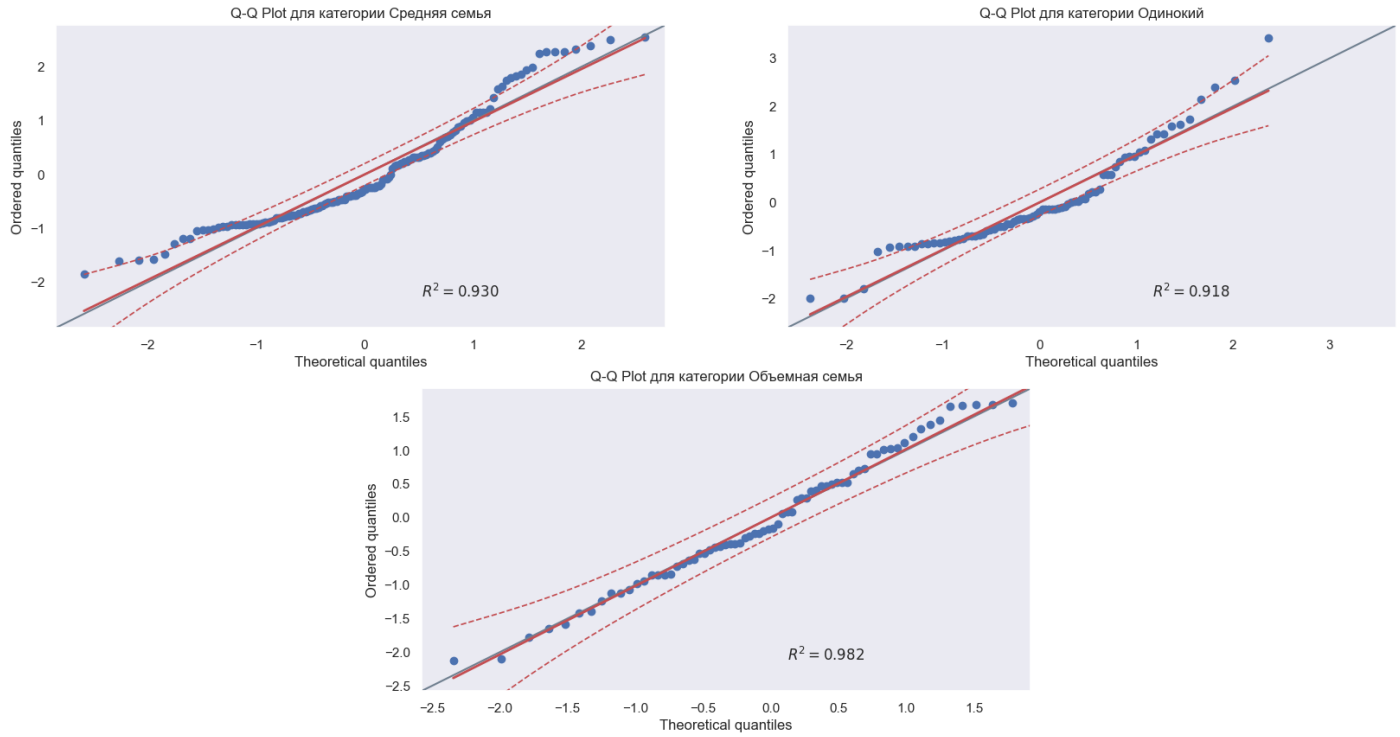

Рис. 7 – QQ-plots для денежных доходов по категориям

Действительно, только одна группа распределена нормально, однако стоит всё же попробовать *ANOVA-test*. Проверим предварительно на гомоскедастичность (равенство дисперсий) с помощью теста Левене:

$$W = \frac{N-k}{k-1} \cdot \frac{\sum\limits_{i=1}^{k} N_i \left(Z_{i\cdot} - Z_{\cdot\cdot}\right)^2}{\sum\limits_{i=1}^{k} \sum\limits_{j=1}^{N_i} N_i \left(Z_{ij} - Z_{i\cdot}\right)},$$

где k – количество групп, N_i – количество результатов исследования для i-ой группы, N – общее количество исследований, Y_{ij} – значение j-го наблюдения в i-ой группе.

$$Z_{ij} = \left|Y_{ij} - \overline{Y}_i\right|,$$

где $\overline{Y}_i$ – среднее по i-ой группе,

$$Z_{i\cdot} = \frac{1}{N_i} \sum_{j=1} N_i Z_{ij},$$

среднее по Z_{ij} для i-ой группы,

$$Z_{\cdot\cdot} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{N_i} Z_{ij}$$

среднее по всем Z_{ij}.

В результате имеем близкое к нулю $p-$value, следовательно отклоняем гипотезу о равенстве дисперсий. Воспользуемся *Welch ANOVA*, так как он более робастный по сравнению с классическим.

Имеем низкое p-value, следовательно отвергаем нулевую гипотезу. После отклонения нулевой гипотезы, можно выполнить апостериорный тест Геймса-Хауэлла (он устойчив к гомоскедастичности), чтобы определить, какие групповые различия являются статистически значимыми. Выясняется, что разница между одинокими людьми и большими семьями сильнее статистистически значима, чем прочая попарная разница.

Второй набор гипотез

- H_0: Средний денежный доход для многодетных (> 2) семей не отличается от дохода малодетных (≤ 2);
- H_1 Средний денежный доход многодетных (> 2) семей больше дохода малодетных (≤ 2).

Повторяя проверку на нормально и гомоскедастичность, получаем отклоненную гипотезу о нормальном распределении и отклоненную гипотезы о равенстве дисперсий. Воспользуемся критерием Стьюдента для проверки гипотезы о равенстве средних: получаем практическое нулевое значение p-value, следовательно отвергаем гипотезу. Следовательно, возможно, что от количества детей зависит доходность семей.

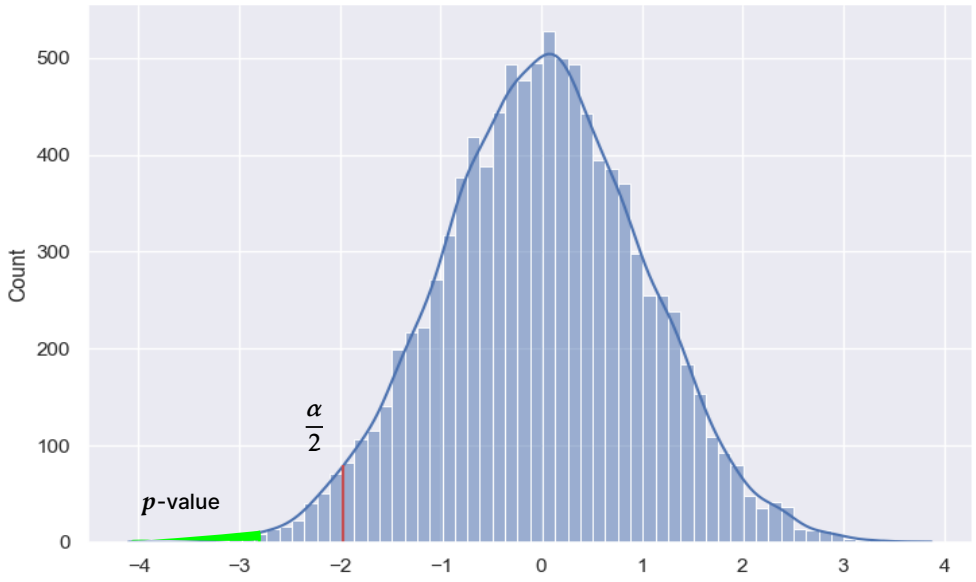

Рис. 8 – Причины, по которой отвергаем гипотезу

Третий набор гипотез

- H_0: Средний денежный доход равен 83100 (средний доход по Москве согласно Мосстату);
- H_1: Средний денежный доход больше 83100.

Будем использовать t-тест Стьюдента. Итого, имеем p-value равное $\approx \mathbf{0.06}$, значит, не отвергаем нулевую гипотезу.

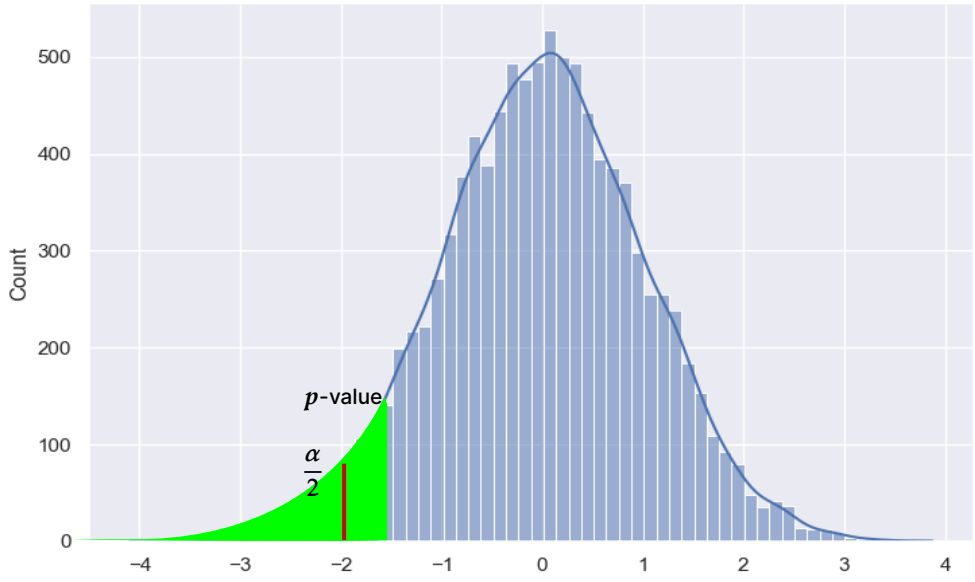

Рис. 9 – Причины, по которым не отвергаем нулевую гипотезу

6. Взаимосвязи между признаками

Построим матрицу корреляций (Пирсона) для количественных характеристик:

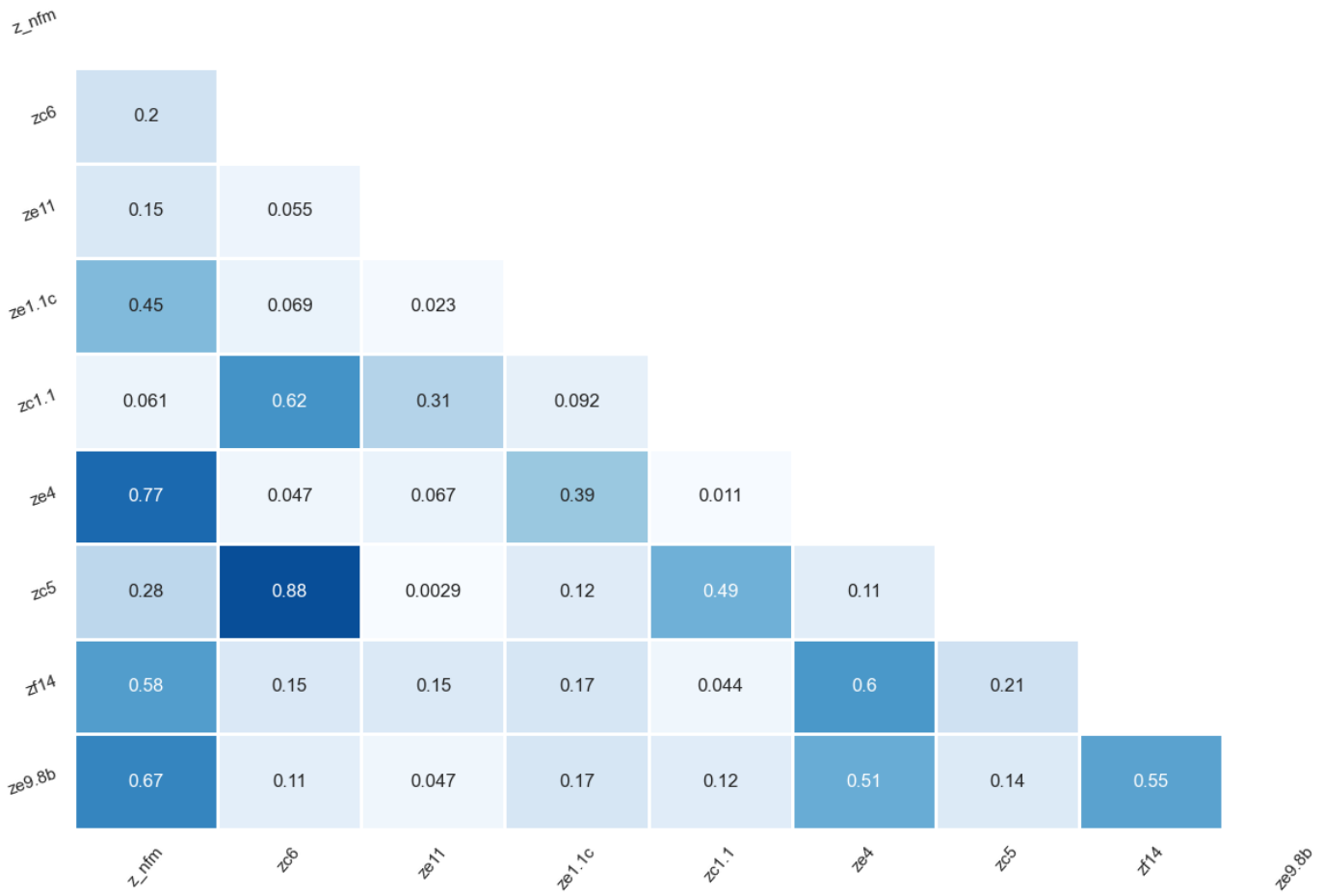

Рис. 10 – Heatplot для матрицы корреляций

Можно заметить очевидную зависимость между жилой и общей площадьми квартиры, а так же зависимость между количеством членов семей и их тратами на питание и мобильную связь.

Построим регрессию на денежный доход, беря во внимание следующие характеристики:

x_1 – количество членов семьи;

x_2 – примерная стоимость жилья;

x_3 – наличие иностранного автомобиля.

Была построена регрессия по всей совокупности данных, посчитан *r2_score*.

$$\hat{y} = 0.3x_1 + 0.028x_2 + 0.224x_3 - 0.244x_4 + 10.1,$$

где x_3, x_4 – закодированный с помощью *one-hot encoding* признак наличия иностранного автомобиля. По критериям Шапиро-Уилка и Пирсона остатки имеют нормальное распределение на уровне надежности 0.05. Можно сделать вывод, что примерная стоимость жилья не является значимым признаком для денежного дохода, в отличие от иностранного автомобиля и количества членов семьи. Метрика $R^2 = 0.72$.

Построим корреляции для категориальных переменных на основе критерия χ^2:

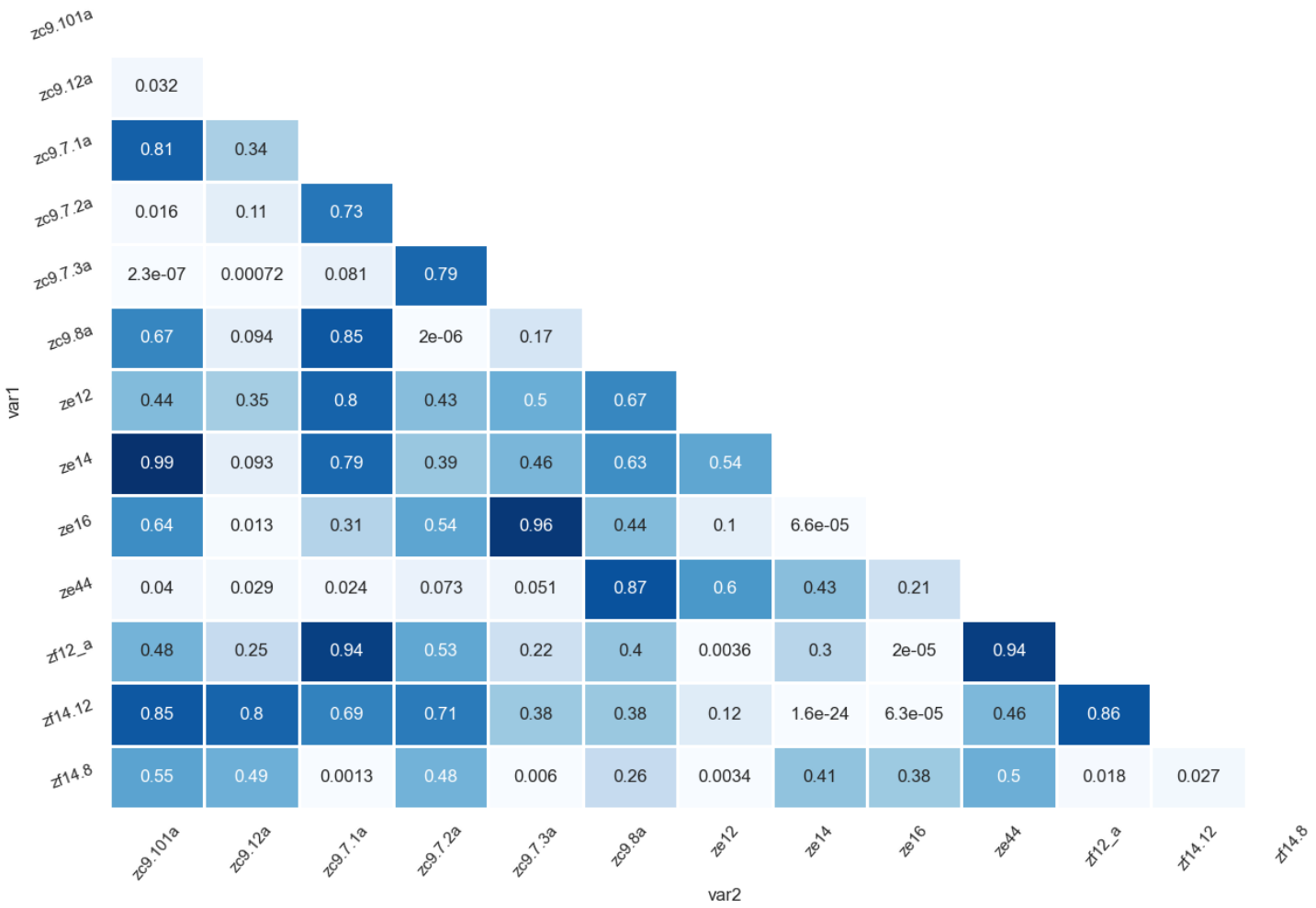

Рис. 11 – Heatplot для матрицы корреляций для категориальных признаков

Видны зависимости, однако они связаны с большой долей отрицательных ответов. Но, например, действительно есть зависимость между наличием иностранного автомобиля и накапливанием сбережений, p-value равен ≈ 0.6 для теста на равенство частот. Рассмотрим наличие закономерности между доходами и утилизацией мусора (не самая репрезентативная с точки зрения поставленной задачи закономерность, однако интересная). Проделаем идентичные шаги, которые были в прошлом пункте. Поставим гипотезы равенства средних по группам:

H_0: В каждой группе по сортировке мусора средние по денежному доходу;

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_5$$

H_1: otherwise

$$H_1 : \text{Все } \mu \text{ между собой не равны}$$

В предикативной форме: $\exists \mu_i \; i, j \in \{1, \ldots, 5\} \; i \neq j : \; \mu_i \neq \mu_j$.

Проверим на нормальность с помощью теста Харке-Бера:

$$JB = n\left(\frac{S^2}{6} + \frac{(K-3)^2}{24}\right),$$

где $S = \frac{\hat{\mu}_3}{\hat{\sigma}^3}$, $K = \frac{\hat{\mu}_4}{\hat{\sigma}^4}$. p-value > 0.02, на уровни значимости 0.02 группы имеют нормальные распределения денежных доходов. Проверим гомоскедастичность методом Левене: p-value > 0.05, значит, сохраняется дисперсия

среди групп. Воспользуемся классическим ANOVA для получения результата: p-value $\approx 0.3 > \alpha \implies$ не отвергаем гипотезу H_0.

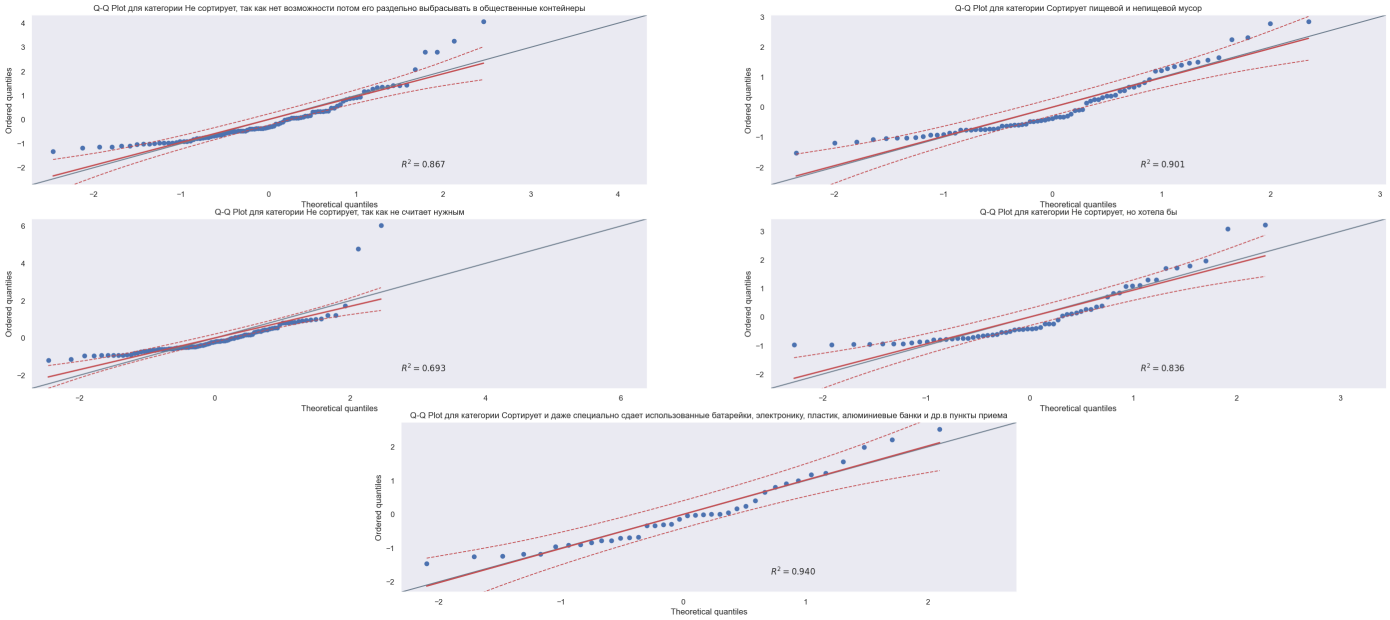

Рис. 12 – QQ-plots для доходов по категориям