Сбор и обработка данных, разведывательный анализ

Сбор и обработка данных

Сбор данных

Источник данных:
Российский мониторинг экономического положения и здоровья населения НИУ ВШЭ - https://www.hse.ru/rlms/

Выборка за 2021 год.

Все переменные взяты из данных по домохозяйствам, кроме переменных, связанных с образованием и здоровьем, они получены из данных по индивидам - единицы измерения сопоставлены с помощью идентификационных переменных (идентификационная переменная домохозяйства - идентификационная переменная домохозяйства, к которому принадлежит индивид). Единицей измерения итоговой выборки является домохозяйство.

Переменные

Зависимые:

- Количество купленного за 7 дней мяса и мясных продуктов, кг. (meat)
- Количество купленной за 7 дней рыбы и морепродуктов, кг. (fish)
- Количество купленных за 7 дней яиц, шт. (eggs)
- Количество купленного за 7 дней молока и молочных продуктов, л. (milk)
- Количество купленных за 7 дней овощей, кг. (vegetables)
- Количество купленных за 7 дней фруктов и ягод, кг. (fruits_and_berries)
- Количество купленных за 7 дней круп и злаковых, кг. (cereals)
- Количество купленных за 7 дней муки и мучных продуктов (flour)
- Количество купленных за 7 дней кондитерских и содержащих высокий уровень сахара изделий, кг. (sweets)
- Количество купленных за 7 дней безалкогольных напитков, л. (softdrinks)
- Количество купленных за 7 дней алкогольных напитков, л. (alcodrinks)
- Сумма, потраченная на покупку продуктов питания за 7 дней, руб. (sum_rub_buy)

Интереса:

- Доход в домохозяйстве, руб. (income)
- Образование, да/нет (educ) (*Особенность*: наличие хотя бы одного члена с высшим образованием или выше) (diplom). Данные получены из таблицы опросов по индивидам.
- Наличие машины, да/нет (car)
- Наличие земли, да/нет (plot)

Контрольные:

- Количество членов семьи, чел. (famsize)
- Наличие централизованного водоснабжения, да/нет (water)
- Наличие централизованной канализации, да/нет (sanitation)
- Наличие холодильника, да/нет (fridge)
- Наличие интернета, да/нет (internet)
- Траты на топливо, руб. (pfuel)
- Траты на услуги транспорта, руб. (ptransp)
- Траты на дом (ЖКХ, аренда и т.д.), руб. (phome)
- Наличие инвалида в домохозяйстве, бинарная (inval). Данные получены из таблицы опросов по индивидам.
- Субсидии от государства, руб. (*Особенность*: объединение различных переменных как пособия по безработице, выплаты на ребенка, налоговые вычеты и т.д.) (govsubs)
- Денежная помощь от других экономических агентов, руб. (nongovsubs)
- Долговая нагрузка, руб. (debt)

Обработка данных

Для возможности анализа данных были проделаны следующие действия;

1. Были удалены все строки, содержащие указания на отсутствие ответа со стороны респондентов (значения 99999995, 99999996, 99999997, 99999998, 99999999)
2. Пустые значения были заменены на 0 (невозможно интерпретировать пропуски как отсутствия ответа, т.к. авторы РМЭЗ явно указывают отсутствие ответа в качестве вышеупомянутых значений)
3. Были удалены выбросы в переменной ‘income’. Другие переменные не имели большого количества выбросов, кроме переменной ‘debt’. Мы решили, что очищение ‘debt’ от выбросов некорректно.

Разведывательный анализ

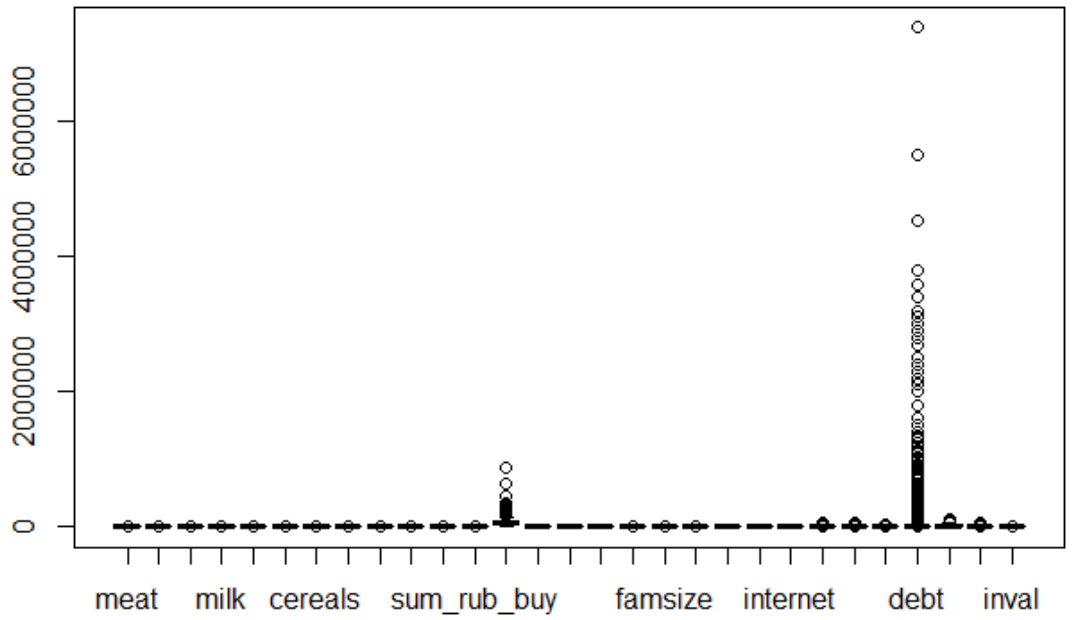

Рис. 1. Распределение переменных

На **Рис. 1** представлены распределения переменных, из чего можно сделать вывод, что самый широкий диапазон значений у переменной debt (долговая нагрузка, руб.), на втором месте переменная sum_rub_buy (сумма, потраченная на покупку продуктов питания за 7 дней, руб.).

Описательные статистики:

Название переменной	**Тип данных R**	**Mean**	**SD**	**Min**	**Max**
meat	dbl	3.42	2.73	0.00	27.00
fish	dbl	0.51	0.83	0.00	7.00
eggs	dbl	10.78	10.84	0	100
milk	dbl	3.73	3.25	0.00	26.00
vegetables	dbl	4.75	15.14	0.00	352.00
fruits_and_berries	dbl	2.49	3.29	0.00	30.00
cereals	dbl	0.64	0.96	0.00	12.00
flour	dbl	4.13	5.01	0.00	56.45
sweets	dbl	1.98	2.55	0.00	35.00
softdrinks	dbl	0.60	1.46	0.00	25.05
alcodrinks	dbl	0.45	1.32	0.00	28.00
sum_rub_buy	dbl	2717.05	1903.88	0.00	18671.00
income	int	61296.61	41319.05	8300.00	265000.00

diplom	int	0.43	0.49	0	1
car	int	0.42	0.49	0	1
plot_bi	int	0.47	0.50	0	1
plot_size	dbl	6.30	31.43	0.00	1430.00
famsize	int	2.57	1.60	1	13
water	int	0.90	0.31	0	1
sanitation	int	0.74	0.44	0	1
fridge	int	0.73	0.44	0	1
internet	int	0.70	0.46	0	1
pfuel	dbl	1880.02	3118.73	0.00	41000.00
ptransp	dbl	833.41	1911.98	0.00	50000.00
phome	dbl	4284.80	3289.21	0.00	35500.00
debt	int	82621.56	378334.52	0	7380000
govsubs	dbl	19912.93	18513.17	0.00	106595.00
nongovsubs	int	88.59	1530.12	0	50000
inval	int	0.16	0.37	0	1

Таб. 1. Описательные статистики

Исходя из данных **Таб. 1** можно заключить, что среднее значение суммы, потраченной на покупку продуктов питания равняется 2717 руб., средний доход за месяц равен 61297 руб., соответственно домохозяйства в среднем тратят 18% от своего дохода на покупку продуктов питания.

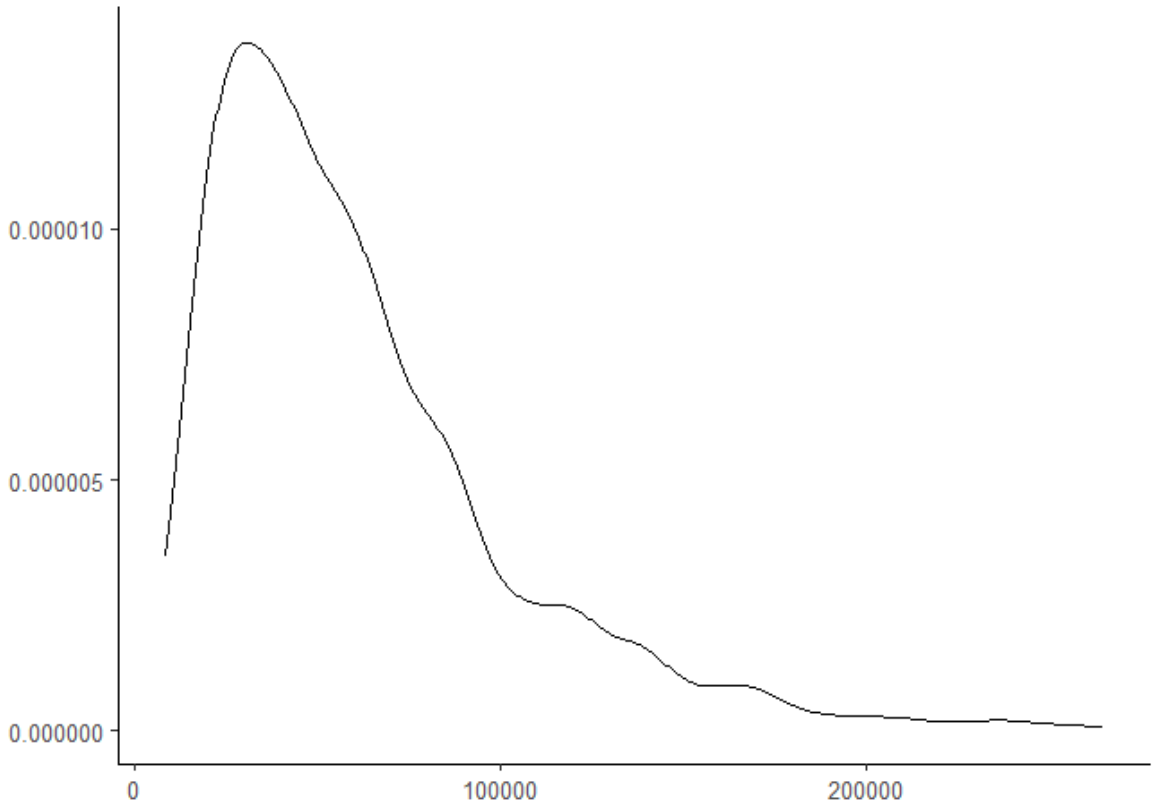

Рис. 2. Плотность распределения: доход, руб.

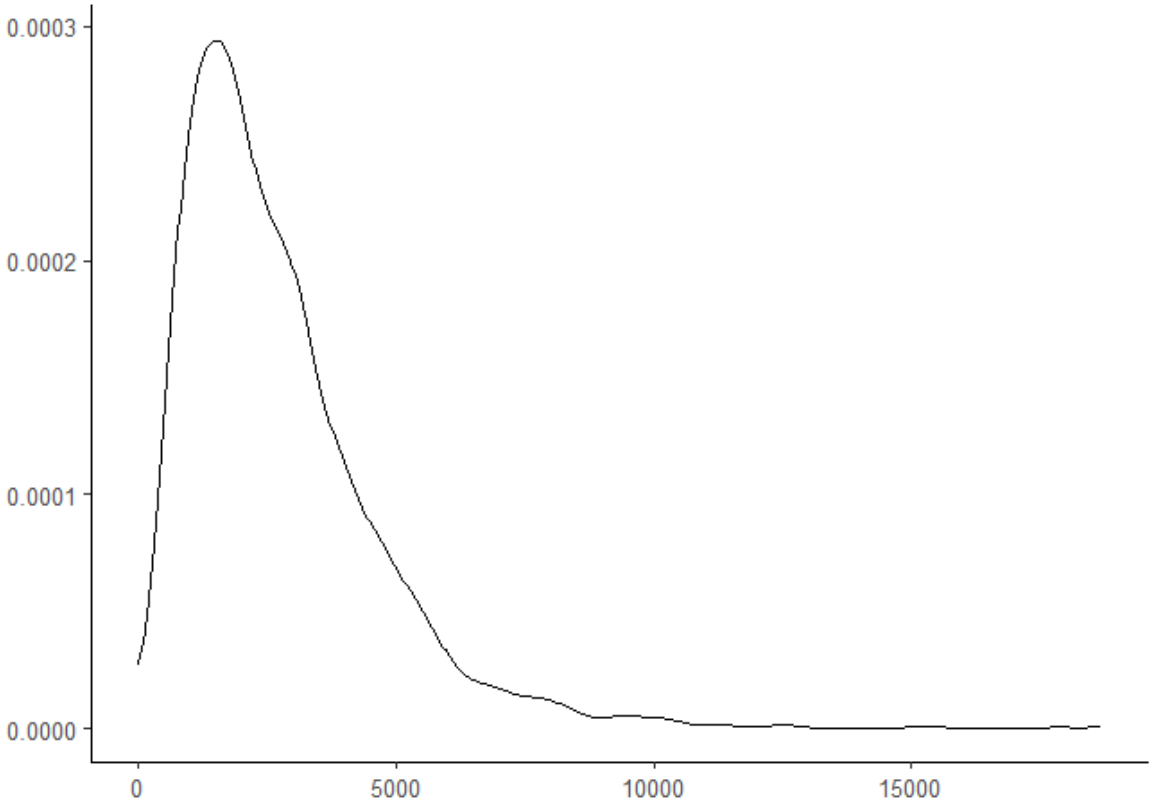

Рис. 3. Плотность распределения: расходы на питание, руб.

Распределения дохода и расходов на питание (**Рис. 2**, **Рис. 3**) схожи с нормальным распределением, однако имеют большие хвосты справа, что говорит о наличии значимого количества людей, чей доход и расходы на питание выше суммы среднего и среднего квадратичного отклонения.

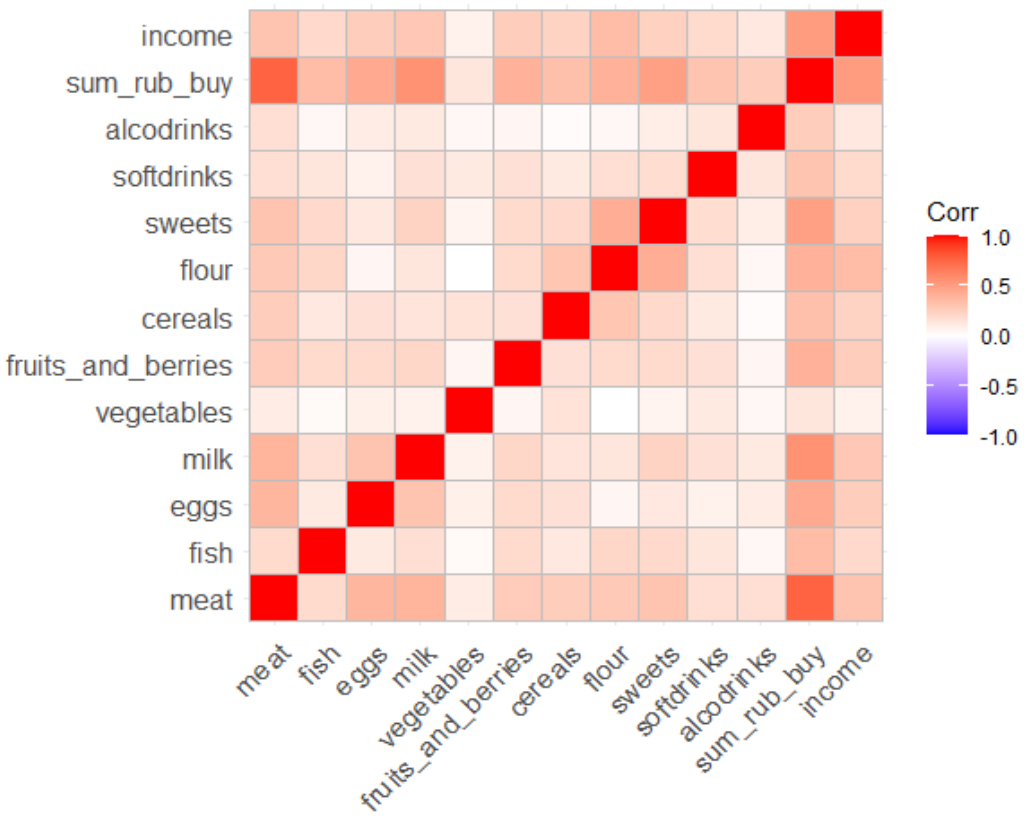

Рис. 4. График корреляции между категориями потребляемых продуктов питания и доходом

На **Рис. 4** можно заметить, что наибольшая корреляция у суммы, потраченной на покупку продуктов питания с категорией продовольствия - у мяса, вероятно, такие результаты из-за того, что данный тип продовольствия, как правило, дороже других типов, наименьшая - у овощей, возможно, в данным случае ситуация обратная. Количество потребляемых овощей - переменная, которая в среднем показывает самую низкую корреляцию с другими переменными.

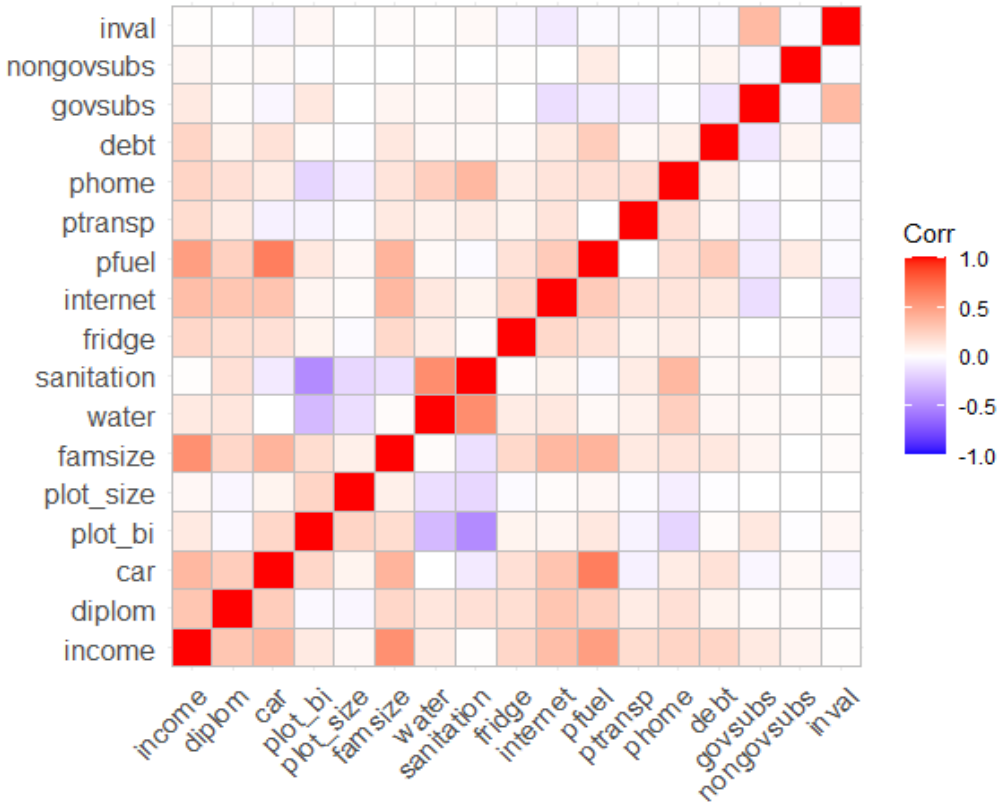

Рис. 5. График корреляции между независимыми переменными

Исходя из данных **Рис. 5**, можно утверждать, что присутствует высокая величина коэффициента корреляции между доходом и размером семьи, предположительно, чем больше домохозяйство, тем в среднем больше в нем трудоспособных представителей, которые зарабатывают деньги. Относительная высокая корреляция между доходом и затратами на бензин, можно объяснить возможностью домохозяйства больше тратить, в том числе на топливо.

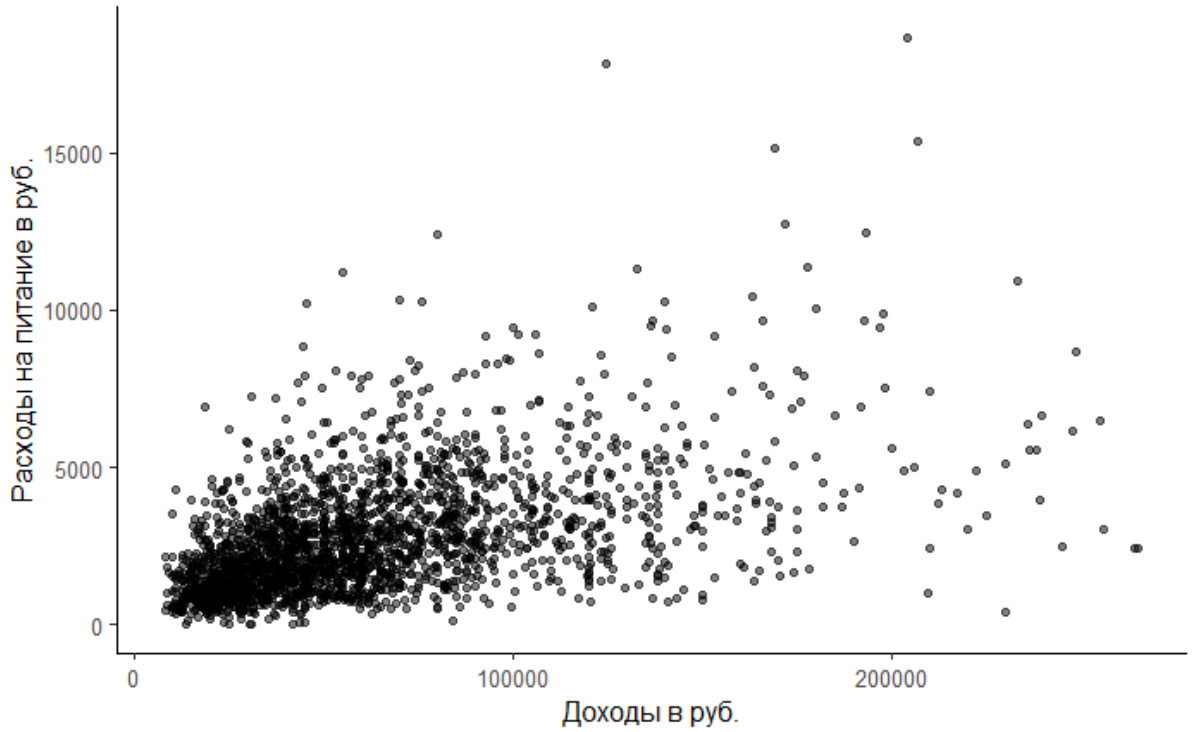

Рис. 6. Совместное распределение дохода в руб. и расходы на питание, руб.

Рис. 6 демонстрирует наличие взаимосвязи между доходом и расходами на питание. Кроме того, с ростом дохода расходы на питание растут, но более медленно.

В целом, можно утверждать, что корреляционный анализ позволил выявить значимую связь категорий продуктов питания с ключевыми переменными, такими как доход и наличие машины. Кроме того, с помощью корреляции мы имеем некоторое подтверждение правильности выбора контрольных переменных, т.е. большинство использованных контрольных переменных имели значимое влияние на большинство продуктов и переменные интереса. Тем не менее, такие переменные как sanitation, nongovsubs и inval показали свою незначимость для, как минимум, половины категорий еды (и общей суммы покупок) и дохода. Необходимо отметить, что 3 категории еды (fish, vegetables, flour) не имели какой-либо связи с большинством значимых контрольных переменных (4/9 или 5/9 незначимых корреляций). Также отсутствует значимая корреляция между переменной интереса diplom и 5 категориями продуктов (cereals, flour, sweets, softdrinks и alcodrinks).