Национальный исследовательский технологический университет

«МИСиС»

Магистерская программа Data Science Tech

Отчет о самостоятельной работе по дисциплине «Программная инженерия»

Бригада № 2 :

ФИО, курс, группа

Настасюк Андрей Юрьевич, 1 курс, МИВТ-21-5-18

ФИО, курс, группа

Кочетов Дмитрий Михайлович, 1 курс, МИВТ-21-5-18

Содержание

1 Общая постановка задачи

Основные данные для работы (файл “insurance.csv”) были взяты по адресу: https://www.kaggle.com/mirichoi0218/insurance. Эти данные содержат информацию о стоимости медицинской страховки граждан Соединенных Штатов Америки. В этом наборе данных находится информация о человеке: возраст, пол, индекс массы тела (ИМТ), количество иждивенцев с медицинской страховкой (МС) в семье, наличие курения, регион проживания и стоимость медицинской страховки (МС). Всего 7 переменных, описание которых приведено в таблице 1.

Таблица 1 – Описание переменных.

№	Характеристика объекта/явления	Название переменной	Шкала измерения	Роль
1.	Возраст	age	Относительная	Независимая
2.	Пол	sex	Номинальная	Независимая
3.	Индекс массы тела (ИМТ) (вычисляется по формуле $bmi = m/h^2$, где m – масса, h – рост)	bmi	Относительная	Независимая
4.	Количество иждивенцев с медицинской страховкой (МС)	children	Относительная	Независимая
5.	Наличие курения	smoker	Номинальная	Независимая
6.	Регион проживания	region	Номинальная	Независимая
7.	Стоимость медицинской страховки (МС) за год (указана в долларах)	charges	Относительная	Зависимая

2 Предварительный анализ собранных данных

2.1 Анализ особенностей данных: потенциальные ошибки и пропущенные значения, группы и выбросы

Исходный набор данных предоставляет выборку из 1338 наблюдений и не имеет пропусков, поэтому можно перейти к исследованию самих данных.

2.1.1 Анализ количественных переменных

В исследуемый набор данных входят 4 количественные переменные: возраст (age), ИМТ (bmi), кол-во иждивенцев (children) и стоимость МС (charges). Для выявления и последующего интерпретирования статистических свойств всех количественных переменных были сформированы таблица основных статистических свойств (таблица 2) и гистограммы распределения (рисунок 1).

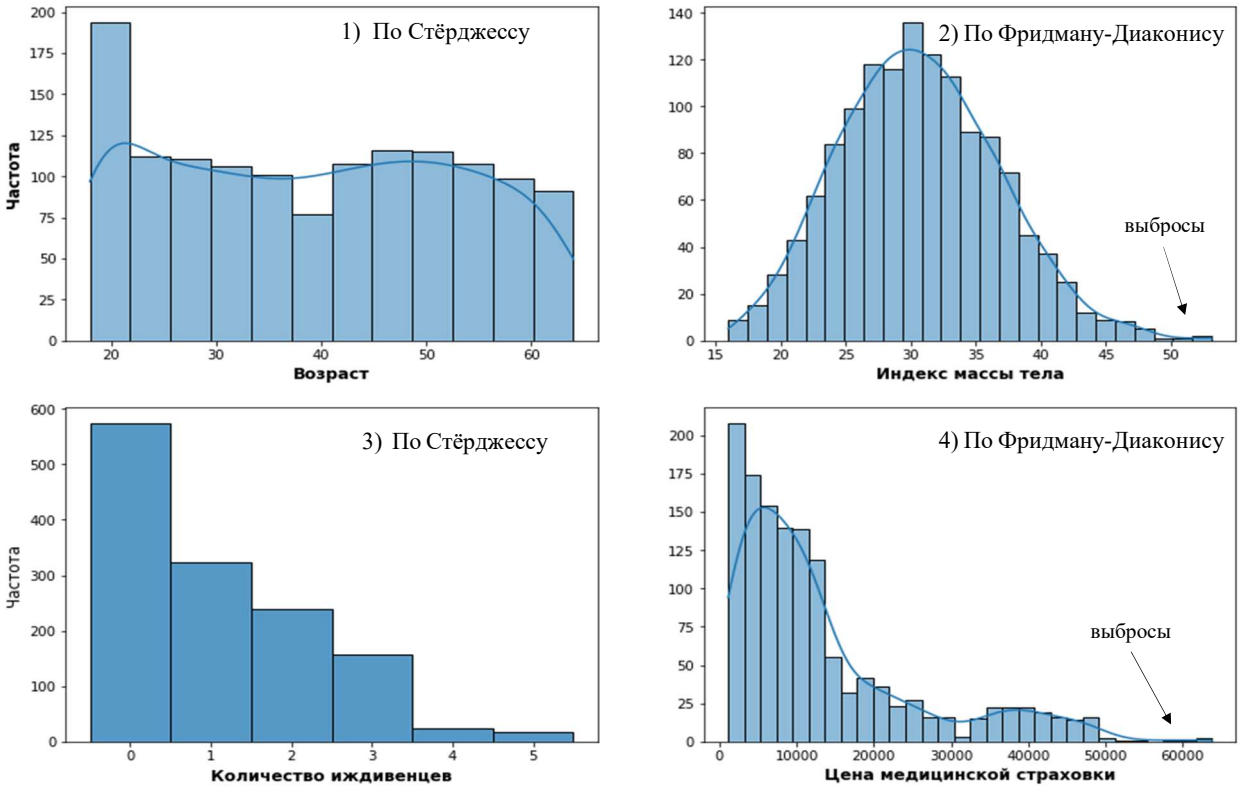

Рисунок 1 – Гистограммы распределения для количественных переменных (1 – по возрасту, 2 – по ИМТ, 3 – по кол-ву иждивенцев, 4 – по стоимости МС)

Статистические показатели по гистограммам с рисунка 1 сведем в таблицу 2:

Таблица 2 – Статистические свойства количественных факторов

	Возраст	ИМТ	Кол-во иждивенцев	Стоимость МС
Среднее	39.207	30.663	1.095	13270.422
Медиана	39.000	30.400	1.000	9382.033
Стандартное отклонение (σ)	14.050	6.098	1.206	12110.011
Межквартильный размах	24.000	8.398	2.000	11899.625
Верхняя квартиль	51.000	34.694	2.000	16639.913
Нижняя квартиль	27.000	26.296	0.000	4740.287
Коэффициент асимметрии	0.056	0.2837	0.937	1.514
Коэффициент эксцесса	-1.245	-0.055	0.197	1.596
Количество наблюдений	1338			
Кол-во пропущенных значений	0			

1) Гистограмма распределения возраста.

На графике наблюдается небольшая асимметрия, что подтверждается разницей между средним и медианой (39,2 > 39,0), а её коэффициент близок к нулю, но межквартильный размах почти равен значению нижней квартили, а отрицательный коэффициент эксцесса свидетельствует о том, что данное распределение плосковершинное. Поэтому данное распределение не является нормальным.

По этим данным, в совокупности с формой графика, можно установить, что распределение является равномерным. Это объясняется тем, что в США нет бесплатной медицины, соответственно, медицинская страховка актуальна для людей всех возрастов.

2) Гистограмма распределения ИМТ.

На этом графике также наблюдается крайне малая асимметрия (30,663 > 30,400), её коэффициент уже больше нуля, но всё еще небольшой, а межквартильный размах в три раза меньше значения нижней квартили. Коэффициент эксцесса отрицателен, но приближен к нулю, следовательно, распределение немного плосковершинное. Визуально, распределение ИМТ приближено к нормальному.

Нормальным показателем ИМТ является диапазон от 18,5 до 25, и поскольку среднее и медиана этого набора данных примерно равны 30, данный вид распределения может быть связан с повышенной степенью ожирения по стране.

Анализ на нормальность распределения по тесту Шапиро -Уилка показал, что коэффициент уровня значимости меньше 0.05, значит можно отвергнуть нулевую гипотезу о том, что распределение является нормальным.

Воспользовавшись правилом «трёх сигм», можем проверить график на наличие выбросов: вероятность отклонения за пределы диапазона [-3σ, 3σ], относительно медианы, равна ≈ 0,003. Таким образом, все значения вне диапазона [12.369; 48.958] являются выбросами. Таких значений – 4, что составляет 0,3% выборки.

3) Гистограмма распределения количества иждивенцев.

На графике явно наблюдается асимметрия (1,095 > 1,000), а коэффициент асимметрии этого массива данных уже приближен к 1, а поскольку в этом массиве слишком много нулевых значений (то есть у человека нет иждивенцев с медицинской страховкой), нижняя квартиль равна 0, а межквартильный размах невелик. Это подтверждает, что данное распределение не является нормальным. Коэффициент эксцесса положителен и невелик, следовательно, распределение немного остроконечно. По форме графика и данным таблицы, делаем вывод, что вид распределения экспоненциальный.

4) Гистограмма распределения стоимости МС.

Для данной гистограммы коэффициент асимметрии приблизительно равен 1.5, а межквартильный размах почти в два раза больше значения нижней квартили. На гистограмме наблюдаются выбросы, что объясняется положительным коэффициентом эксцесса, равным почти 1.6. Данное распределение не является нормальным. Можно сделать вывод, что для данной гистограммы наблюдается экспоненциальное распределение.

Это можно объяснить экономически, поскольку меньшее количество людей может финансово позволить себе дорогое медицинское страхование.

По тесту Шапиро-Уилка, нулевая гипотеза о нормальности распределения отвергается, так как полученное значение уровня значимости меньше 0,05. Аналогично предыдущей гистограмме, применяем правило «трёх сигм»: вне диапазона [0; 49600.456] обнаружено 7 значений, что составляет 0,5% от выборки.

2.1.2 Анализ качественных переменных

В исследуемый набор данных входят 3 качественные переменные: пол(sex), наличие курения(smoker) и регион(region). Их гистограммы распределения приведены на рисунке 3.

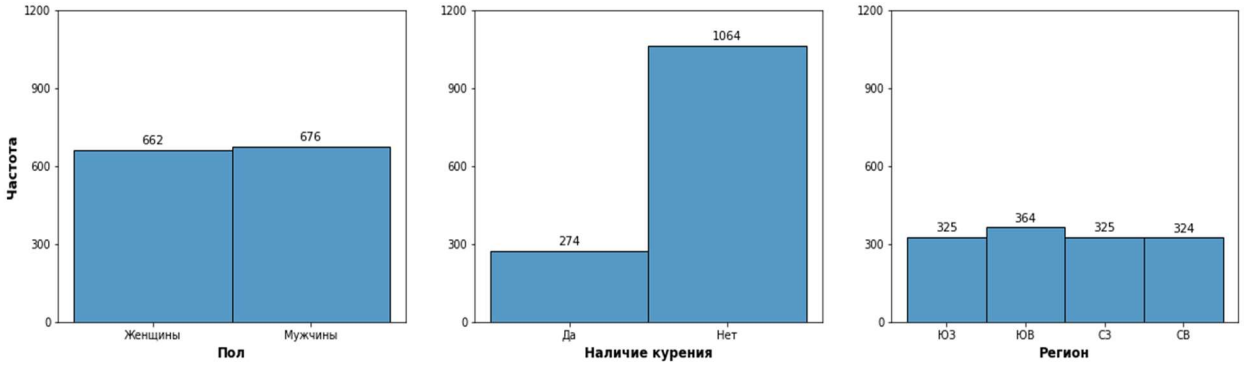

Рисунок 3 – гистограммы распределения качественных переменных. (1 – по полу, 2 – по наличию курения, 3 – по регионам)

1) Гистограмма распределения по полу.

Для данной переменной обозначены две категории – женщины и мужчины. Как видно из гистограммы, выборка по полу произведена почти равномерно, из чего следует, что мужчины и женщины одинаково заинтересованы в приобретении медицинской страховки, как и ожидалось.

2) Гистограмма распределения по наличию курения.

Исходя из гистограммы, видно, что наиболее представленным уровнем являются некурящие люди. Таким образом, в нашей выборке присутствует смещение в сторону некурящих людей, однако с учётом общего количества наблюдений она может отражать

свойства генеральной совокупности. Это может быть связано с малой востребованностью имения медицинской страховки у курящих людей.

3) Гистограмма распределения по регионам.

Данная гистограмма демонстрирует почти полную равномерность, как и распределение по полу. Большим показателем выделяется только юго-восточный регион, что может объясняться относительно высокой плотностью населения.

Уровней с долей менее 5% не было выявлено, необходимости в укрупнении нет.

2.1.3 Анализ репрезентативности выборки

Исходя из полученных заключений, можно сделать вывод о том, что данная выборка соответствует генеральной совокупности и репрезентативна, однако присутствует смещение в сторону некурящих людей, описанное в п. 2.1.2.

2.2 Анализ статистической связи

2.2.1 Графический анализ пары «числовая зависимая переменная – качественная независимая переменная».

Проведем анализ распределения стоимости МС по половому признаку, по признаку наличия курения и по регионам. На рисунке 4 представлены диаграммы Бокса-Уискера.

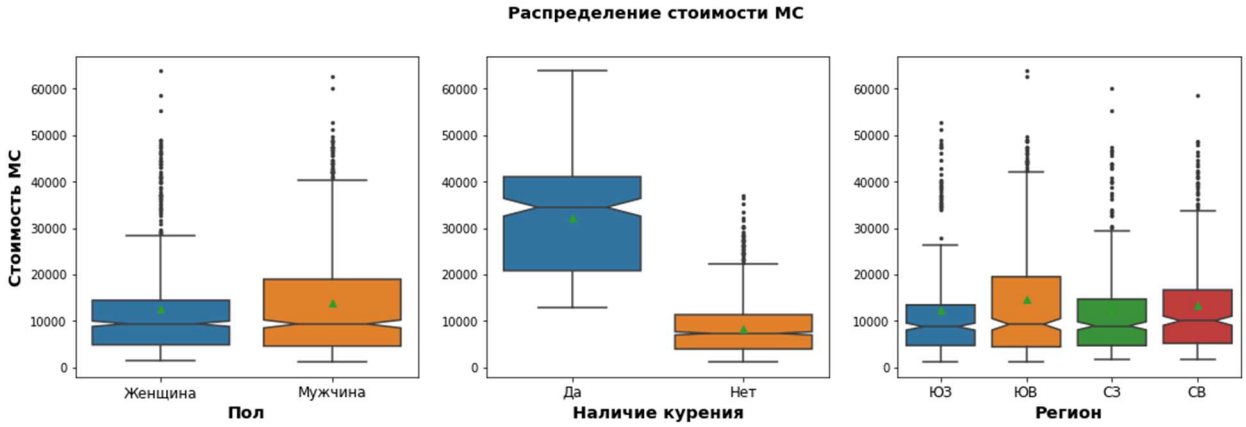

Рисунок 4 – Распределение стоимости МС по качественным признакам

1) Стоимость МС - Пол:

При анализе среднего значения и медианы стоимости МС у женщин и мужчин, которые находятся примерно на одном уровне для обоих категорий, можно сделать вывод о том, что существенных различий в стоимости МС у мужчин и женщин нет, но межквартильный размах и верхняя граница у мужчин больше, чем у женщин, что может свидетельствовать о том, что мужчины чаще склонны иметь большую стоимость МС, чем женщины. Это может быть связано, например, с большим спектром опасных работ и профессий у мужчин.

2) Стоимость МС - Наличие курения:

Из диаграммы отлично видно, что курящие люди почти всегда имеют стоимость МС куда выше, чем некурящие люди. Среднее и медианное значение сильно разнится у обоих

категорий, у категории "курящий" медиана и среднее значение более чем в 3 раза больше, чем у категории "некурящий". Межквартильный размах и верхняя граница у категории "курящий" также больше, что может быть связано с тем, что курящий человек обычно имеет больший спектр болезней, вызванных наличием курения, из-за чего стоимость МС куда выше, чем у некурящих людей.

3) Стоимость МС - Регион:

При анализе средних и медианных значений по диаграмме можно заметить, что средние и медианные значения у сторон света ЮЗ-СЗ и ЮВ-СВ попарно примерно равны, но при этом средние и медианные значения у ЮВ-СВ выше, чем у ЮЗ-СЗ, это может говорить о том, что на Востоке на медицинское страхование тратят в среднем немного больше, чем на Западе. Межквартильный размах и верхняя граница наибольшая у ЮВ региона, что может говорить о большей нестабильности этого региона, по отношению к другим регионам. Возможно, там хуже ситуация по болезням в целом или, например, этот регион занимается большим спектром опасных работ.

Для формальной проверки гипотезы о наличии статистической связи был выполнен непараметрический дисперсионный анализ с помощью критерия Краскела-Уоллиса. Результат анализа сведен в таблицу 3.

Таблица 3 – Результат непараметрического дисперсионного анализа стоимости МС по качественными переменным

	Значение критерия, H	Значимость критерия, p
Стоимость МС - Пол	2070.990	0.00
Стоимость МС - Наличие курения	2143.275	0.00
Стоимость МС - Регион	2022.174	0.00

P значение для каждого из наборов меньше 0.05, значит, с уровнем доверия 0.95 во всех трех случаях отвергаем нулевую гипотезу, которая говорит о том, что связь между параметрами отсутствует.

2.2.2 Графический анализ пары «числовая зависимая переменная – числовая независимая переменная»

Проведем анализ распределения стоимости МС по возрасту, ИМТ и количеству иждивенцев.

Формальная проверка гипотезы о наличии связи выполнена при помощи подсчета коэффициентов корреляции Пирсона, Спирмена и Кендалла. Также были посчитаны их уровни значимости и проведено сравнение уровня значимости коэффициентов с t критерием Стьюдента при 1336 степенях свободы и 0.95 уровне доверия. На основании этой проверки будет либо принята, либо отвергнута нулевая гипотезу о том, что коэффициенты корреляции генеральных совокупностей данных выборок будут равны 0.

2.2.2.1 Стоимость МС – Возраст

На рисунке 5 представлена диаграмма рассеивания стоимости МС по возрасту.

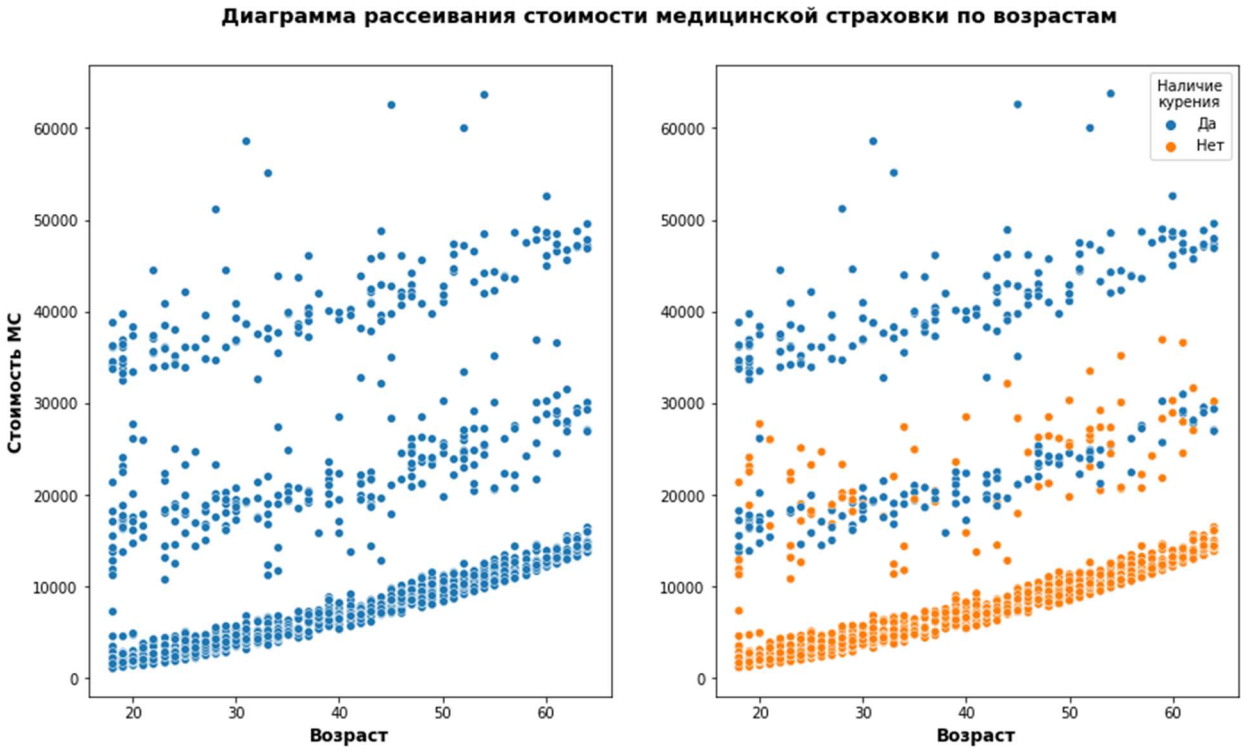

Рисунок 5 – Диаграмма рассеивания стоимости МС по возрастам

Из диаграммы рассеивания видно, что с увеличением возраста растет стоимость МС. Это связано с тем, что с возрастом возникает больше болезней, из-за чего растут расходы на медицинское страхование. Также, из диаграммы видно, что нижние значения по стоимости МС распределены кучнее друг к другу и, для наглядности, был представлен второй такой же график, но сгруппированный по курящим/некурящим людям. Из него отчетливо видно, что нижнюю полосу кучно друг к другу расположенных значений занимают именно некурящие люди.

Результаты формальной проверки гипотезы о наличии связи стоимости МС и возраста сведены в таблицу 4.

Таблица 4 – Результаты проверки гипотезы о наличии связи стоимости МС с возрастом

	Значение коэффициента, K	Уровень значимости критерия, p
Коэффициент Пирсона	0.30	11.453
Коэффициент Спирмена	0.53	23.109
Коэффициент Кендалла	0.48	19.746
Значение t критерия Стьюдента	1.646	

Коэффициенты корреляции говорят о том, что возраст и стоимость МС слабо связаны между собой.

В виду того, что уровни значимости критериев больше, чем критическое значения t критерия Стьюдента, то с уровнем доверия в 0.95 мы отклоняем нулевую гипотезу.

2.2.2.2 Стоимость МС – ИМТ

На рисунке 6 представлена диаграмма рассеивания стоимости МС по ИМТ.

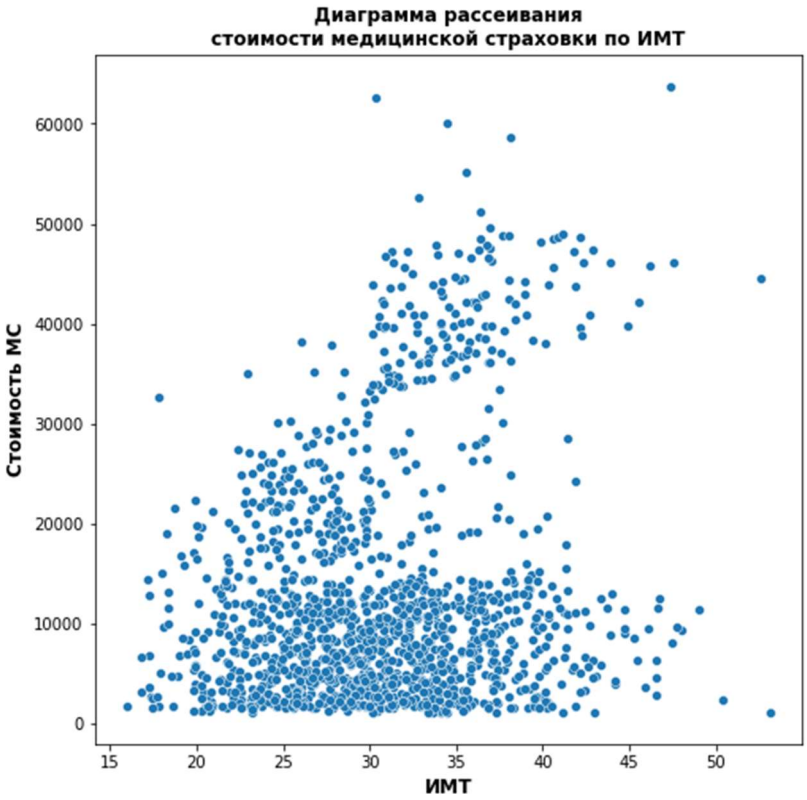

Рисунок 6 – Диаграмма рассеивания стоимости МС по ИМТ

Из диаграммы рассеивания видно, что с увеличением ИМТ стоимость МС также растет. Визуально уже сложнее оценить на сколько сильно стоимость МС и ИМТ связаны, но все еще можно наблюдать слабую прямо пропорциональную зависимость стоимости МС от ИМТ. Это можно объяснить тем, что ИМТ показывает связь роста и веса человека, то есть при большем индексе ИМТ у человека могут наблюдаться большие проблемы с ожирением и лишним весом, что отрицательно сказывается на его здоровье, а в виду чего будет большая стоимость МС.

Результаты формальной проверки гипотезы о наличии связи стоимости МС с ИМТ сведены в таблицу 5.

Таблица 5 – Результаты проверки гипотезы о наличии связи стоимости МС с ИМТ

	Значение коэффициента, K	Уровень значимости критерия, p
Коэффициент Пирсона	0.20	7.397
Коэффициент Спирмена	0.12	4.396
Коэффициент Кендалла	0.08	3.027
Значение t критерия Стьюдента	1.646	

Коэффициенты корреляции говорят о том, что ИМТ и стоимость МС очень слабо связаны между собой.

В виду того, что уровни значимости критериев больше, чем критическое значения t критерия Стьюдента, то с уровнем доверия в 0.95 мы отклоняем нулевую гипотезу.

2.2.2.3 Стоимость МС – Количество иждивенцев

На рисунке 7 представлена диаграмма рассеивания стоимости МС по количеству иждивенцев.

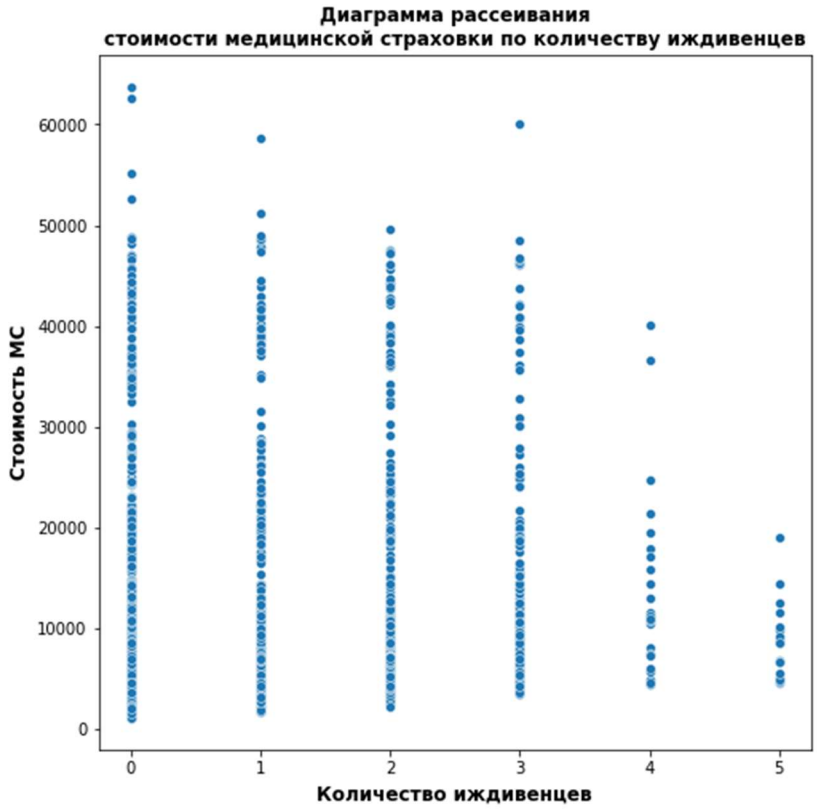

Рисунок 7 – Диаграмма рассеивания стоимости МС по количеству иждивенцев

Диаграмма рассеивания показывает слабозаметную обратно пропорциональную зависимость стоимости МС от количества иждивенцев. Все же объяснить это можно, например, наличием государственных льгот за количество иждивенцев в семье. Или поощрений в виде скидок от страховой компании, так как в конкретном наборе данных идет учет лишь тех иждивенцев в семье, которые тоже имеют медицинское страхование, а как правило, члены одной семьи имеют медицинское страхование в одной компании, а она, в свою очередь, может предоставлять скидки таким семьям.

Результаты формальной проверки гипотезы о наличии связи стоимости МС и количества иждивенцев сведены в таблицу 6.

Таблица 6 – Результаты проверки гипотезы о наличии связи стоимости МС с количеством иждивенцев

	Значение коэффициента, K	Уровень значимости критерия, p
Коэффициент Пирсона	0.07	2.491
Коэффициент Спирмена	0.13	4.918
Коэффициент Кендалла	0.10	3.789
Значение t критерия Стьюдента	1.646	

Коэффициенты корреляции говорят о том, что количество иждивенцев и стоимость МС очень слабо связаны между собой.

В виду того, что уровни значимости критериев больше, чем критическое значения t критерия Стьюдента, то с уровнем доверия в 0.95 мы отклоняем нулевую гипотезу.

2.2.3 Анализ наличия корреляции между независимыми переменными

2.2.3.1 Анализ наличия корреляции между независимыми качественными переменными

Для анализа силы связи между качественными переменными использовались таблицы кросс-табуляции, значения статистики Хи-квадрат и V-Крамера.

Критическое значение статистики Хи-квадрат для 1337 степеней свободы при уровне доверия 0.95 равно 1423.178. С этим значением сравнивались значения статистики Хи-квадрат для каждых из пар. Если полученное значение статистики Хи-квадрат меньше, чем критическое значение Хи-квадрат, то принимается нулевая гипотеза о том, что переменные являются независимыми, в ином случае – нулевая гипотеза отвергается

1) Пол – Наличие курения

Таблица кросс-табуляции представлена в таблице 7.

Таблица 7 – Кросс-табуляция признаков «Пол» и «Наличие курения»

Пол \ Наличие курения	Нет	Да	Всего
Женский	547	115	662
Мужской	517	159	676
Всего	1064	274	1338

Из таблицы видно, что курящих мужчин немного больше, чем курящих женщин, а относительно всех курящих людей в выборке - мужчин в полтора раза больше, чем женщин. При этом количество женщин в выборке примерно равно количеству мужчин.

Значение статистики Хи-квадрат: 168.166

V-Крамер: 0.069

Так как полученное значение статистики Хи-квадрат меньше, чем критическое значение статистики Хи-квадрат, то нулевая гипотеза принимается.

2) Пол – Регион

Таблица кросс-табуляции представлена в таблице 8.

Таблица 8 – Кросс-табуляция признаков «Пол» и «Регион»

Пол \ Регион	ЮВ	ЮЗ	СВ	СЗ	Всего
Женский	161	164	175	162	662
Мужской	163	161	189	163	676
Всего	324	325	364	325	1338

Из таблицы видно, что чуть большее количество людей проживает на ЮВ, а в остальных регионах проживает одинаковое количество людей. Соотношение мужчин и женщин в каждом регионе тоже примерно одинаковое.

Значение статистики Хи-квадрат: 415.88

V-Крамер: 0

Так как полученное значение статистики Хи-квадрат меньше, чем критическое значение статистики Хи-квадрат, то нулевая гипотеза принимается.

3) Наличие курения – Регион

Таблица кросс-табуляции представлена в таблице 9.

Таблица 9 – Кросс-табуляция признаков «Пол» и «Регион»

Наличие курения \ Регион	ЮВ	ЮЗ	СВ	СЗ	Всего
Нет	257	267	273	267	1064
Да	67	58	91	58	274
Всего	324	325	364	325	1338

Из таблицы видно, что количество курящих людей примерно одинаковое во всех регионах, немного большее значение лишь в регионе ЮВ, но это может быть связано с тем, что в этом регионе в принципе проживает немного большее количество людей по сравнению с другими регионами.

Значение статистики Хи-квадрат: 345.314

V-Крамер: 0.057

Так как полученное значение статистики Хи-квадрат меньше, чем критическое значение статистики Хи-квадрат, то нулевая гипотеза принимается.

2.2.3.2 Анализ наличия корреляции между независимыми количественными переменными

Формальная проверка гипотезы о наличии связи выполнена аналогично п.2.2.2.

1) Возраст – ИМТ

На рисунке 8 представлена диаграмма рассеивания стоимости МС по количеству иждивенцев.

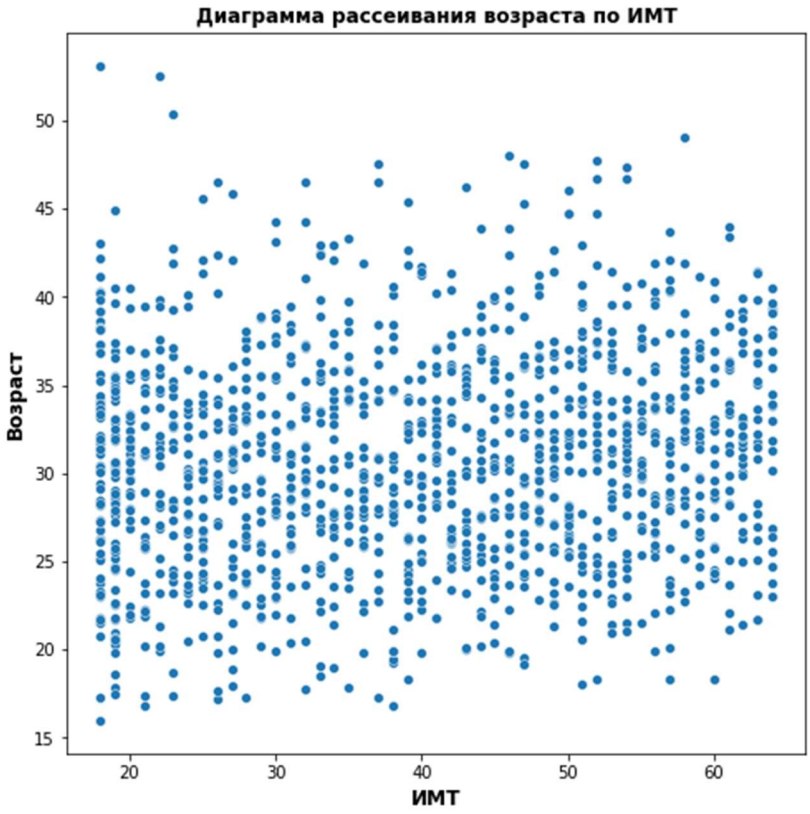

Рисунок 8 – Диаграмма рассеивания возраста по ИМТ

По диаграмме рассеивания очень трудно установить какие-либо тенденции возраста относительно ИМТ, присутствует лишь очень слабозаметная зависимость, что чем меньше возраст, тем больше вероятность встретить низкий показатель ИМТ. На практике это вполне нормальная ситуация, но это сложно назвать какой-либо закономерностью. Можно предположить, что чем младше человек, тем, скорее всего, вероятнее встретить у него меньший ИМТ. Это может быть связано с возрастными заболеваниями, которые могут сказываться на весе человека, например с такими как диабет.

Результаты формальной проверки гипотезы о наличии связи возраста и ИМТ сведены в таблицу 10.

Таблица 10 – Результаты проверки гипотезы о наличии связи возраста с ИМТ

	Значение коэффициента, K	Уровень значимости критерия, p
Коэффициент Пирсона	0.11	4.018
Коэффициент Спирмена	0.11	3.961
Коэффициент Кендалла	0.07	2.685
Значение t критерия Стьюдента	1.646	

Коэффициенты корреляции говорят о том, что ИМТ и возраст очень слабо связаны между собой.

В виду того, что уровни значимости критериев больше, чем критическое значения t критерия Стьюдента, то с уровнем доверия в 0.95 мы отклоняем нулевую гипотезу.

2) Возраст – Количество иждивенцев

На рисунке 9 представлена диаграмма рассеивания стоимости МС по количеству иждивенцев.

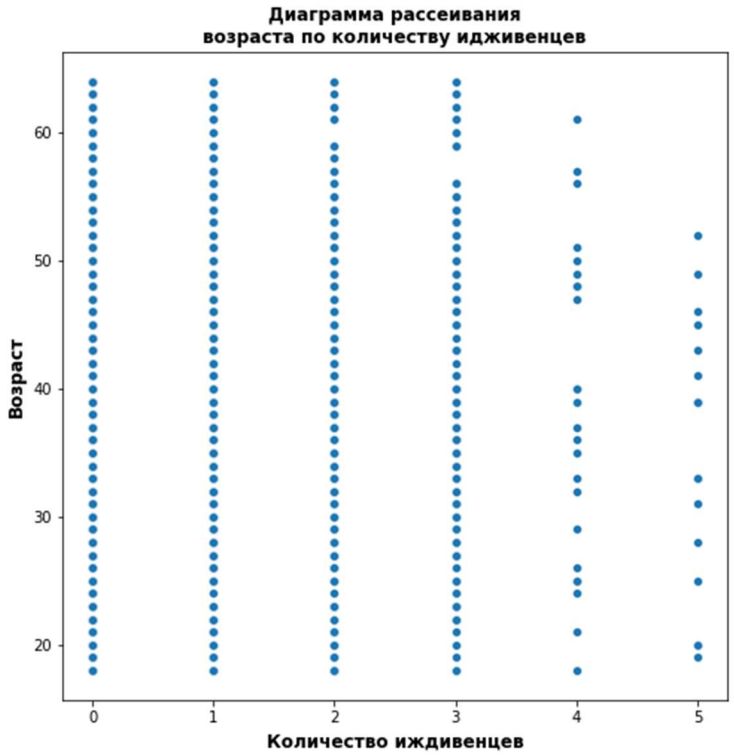

Рисунок 9 – Диаграмма рассеивания возраста по количеству иждивенцев

Из диаграммы рассеивания можно установить очень слабую обратно пропорциональную связь возраста с количеством иждивенцев. Также из диаграммы видно, что чем больше у тебя иждивенцев, то, исходя из данных, тем меньше должен быть твой возраст. Например, у людей, количество иждивенцев которых равно 5, почти не встречается возраст старше 50 лет, чего нельзя сказать о людях, с количеством иждивенцев меньше, чем 5. Но все равно эта связь крайне слабая, и зная возраст человека мы достаточно со слабой вероятностью может предсказать его количество иждивенцев, так как на практике эти величины действительно слабо связаны. По крайней мере очень трудно сказать, что чем ты старше, тем большее количество иждивенцев у тебя должно быть.

Результаты формальной проверки гипотезы о наличии связи возраста и количества иждивенцев сведены в таблицу 11.

Таблица 11 – Результаты проверки гипотезы о наличии связи возраста с количеством иждивенцев

	Значение коэффициента, K	Уровень значимости критерия, p
Коэффициент Пирсона	0.04	1.554
Коэффициент Спирмена	0.06	2.087
Коэффициент Кендалла	0.04	1.582
Значение t критерия Стьюдента	1.646	

Коэффициенты корреляции говорят о том, что количество иждивенцев и возраст очень слабо связаны между собой.

В виду того, что уровни значимости критериев Пирсона и Кендалла меньше, чем критическое значения t критерия Стьюдента, то с уровнем доверия в 0.95 мы принимаем нулевую гипотезу, но для критерия Спирмена с уровнем доверия 0.95 мы отклоняем эту нулевую гипотезу.

3) ИМТ – Количество иждивенцев

На рисунке 10 представлена диаграмма рассеивания стоимости МС по количеству иждивенцев.

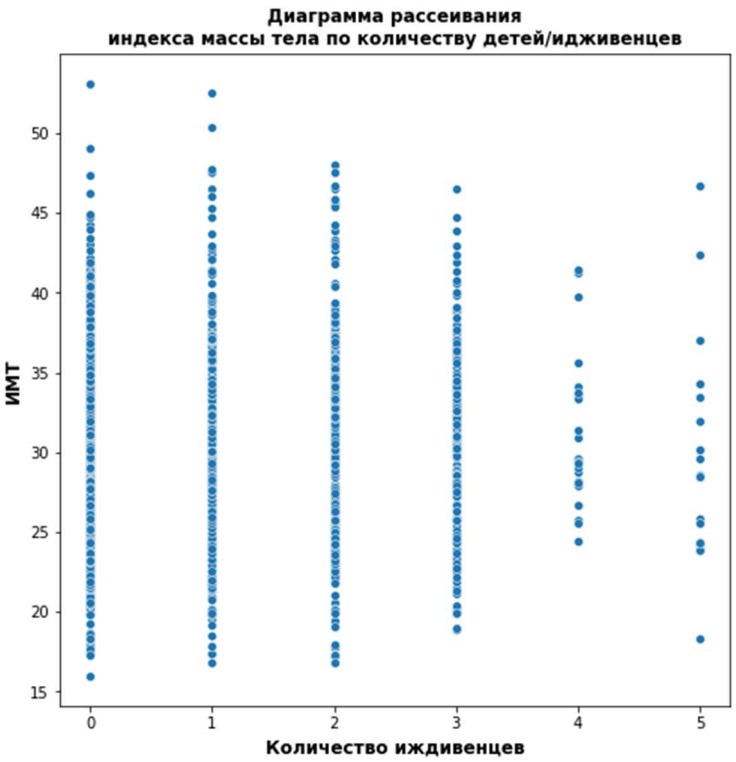

Рисунок 10 – Диаграмма рассеивания ИМТ по количеству иждивенцев

Из диаграммы рассеивания почти нельзя установить никакие тенденции. Все что можно сказать, что при некоторых значениях количества иждивенцев могут почти не

встречаться многие значения ИМТ. Так, например, для количества иждивенцев равное 4 почти нет значений ИМТ меньше 25 или больше 45. На практике нет вообще никаких зависимостей ИМТ от количества иждивенцев. Нельзя зная пропорцию веса и роста человека (именно это означает ИМТ) установить его количество иждивенцев.

Результаты формальной проверки гипотезы о наличии связи возраста и количества иждивенцев сведены в таблицу 12.

Таблица 12 – Результаты проверки гипотезы о наличии связи ИМТ с количеством иждивенцев

	Значение коэффициента, K	Уровень значимости критерия, p
Коэффициент Пирсона	0.01	0.466
Коэффициент Спирмена	0.02	0.571
Коэффициент Кендалла	0.01	0.423
Значение t критерия Стьюдента	1.646	

Коэффициенты корреляции говорят о том, что количество иждивенцев и ИМТ очень слабо связаны между собой.

В виду того, что уровни значимости критериев меньше, чем критическое значения t критерия Стьюдента, то с уровнем доверия в 0.95 мы принимаем нулевую гипотезу.

2.2.3.3 Анализ наличия корреляции между независимыми количественными и качественными переменными

Анализ наличия корреляции будет проводиться по аналогии с п.2.2.1.

1) Распределение количества иждивенцев по половому признаку, наличию курения и регионам

Диаграммы распределения количества иждивенцев по признакам представлены в виде Бокса-Уискера на рисунке 11.

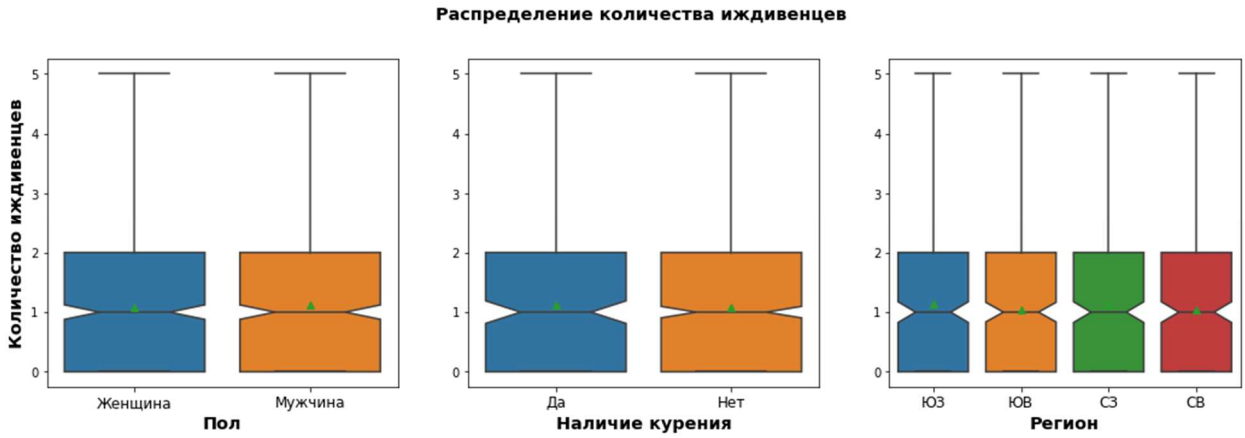

Рисунок 11 – Распределение количества иждивенцев по качественным переменным

- Количество иждивенцев – Пол

Из анализа диаграммы можно сделать вывод, что статистически важных различий между полами относительно количества иждивенцев почти не наблюдается, есть лишь небольшая разница в среднем значении у мужчин и женщин. Но так как выборка дискретная, то эта разница совсем несущественна.

- Количество иждивенцев – Наличие курения

Исходя из диаграммы можно сделать вывод, что статистически важных различий между курящими и некурящими людьми относительно количества иждивенцев не наблюдается.

- Количество иждивенцев – Регион

Почти аналогичные результаты наблюдаются при анализе количества иждивенцев относительно регионов. Единственное, что можно заметить – это небольшие различия в уровне средних значений у регионов.

Отсутствие каких-либо сильных различий по рассмотренным параметрам обусловлено тем, что на практике действительно не наблюдается какой-либо корреляции количества безработных членов семьи, имеющих медицинскую страховку от пола, региона проживания или наличия курения.

Для формальной проверки гипотезы о наличии статистической связи был выполнен непараметрический дисперсионный анализ с помощью критерия Краскела-Уоллиса. Результат анализа сведен в таблицу 13.

Таблица 13 – Результат непараметрического дисперсионного анализа количества иждивенцев по качественным переменным

	Значение критерия, H	Значимость критерия, p
Количество иждивенцев - Пол	218.602	1.82e-49
Количество иждивенцев - Наличие курения	82.722	9.44e-20
Количество иждивенцев - Регион	778.036	3.22e-171

P значение для каждого из наборов меньше 0.05, значит, с уровнем доверия 0.95 во всех трех случаях отвергаем нулевую гипотезу, которая говорит о том, что связь между параметрами отсутствует.

2) Распределение ИМТ по половому признаку, наличию курения и регионам

Диаграммы распределения ИМТ по признакам представлены в виде Бокса-Уискера на рисунке 12.

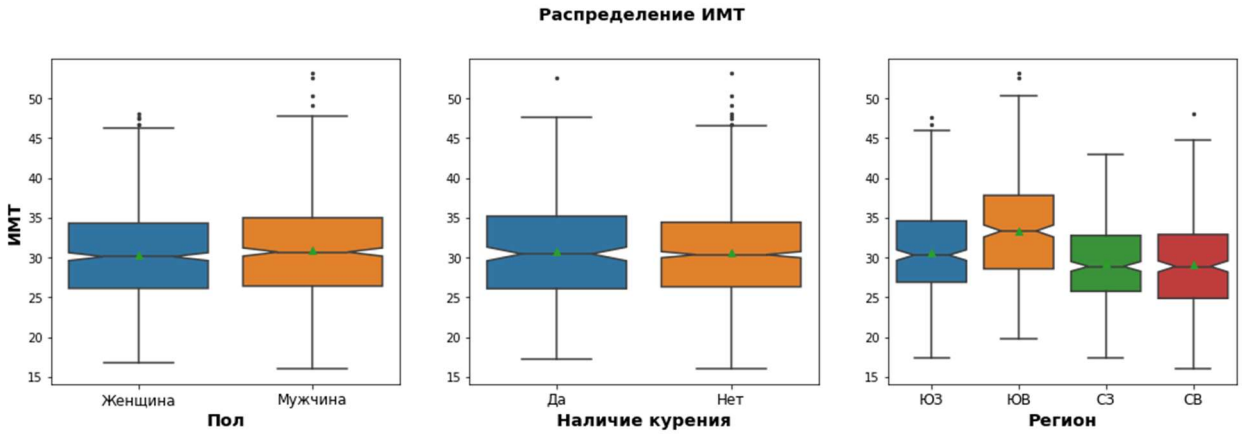

Рисунок 12 – Распределение ИМТ по качественным переменным

- ИМТ – Пол

Из анализа диаграммы видны небольшие различия ИМТ относительно пола людей. У мужчин чуть больше среднее и медианное значение, также чуть больше межквартильный размах, верхняя граница выше, чем верхняя граница у женщин, а нижняя граница ниже, чем нижняя граница у женщин. В виду этого можно сказать, что мужчины чуть больше подвержены излишнему весу или ожирению, чем женщины. Чем-либо обусловить это на практике почти не представляется возможным, но в целом, различия по ИМТ между полами пренебрежимо малы.

- ИМТ – Наличие курения

Аналогичная ситуация с наличием курения у мужчин и женщин. По диаграмме можно наблюдать совсем небольшие различия между мужчинами и женщинами, только в данном случае женщины имеют совсем чуть большие показатели по сравнению с мужчинами, а именно: чуть большие значения медианы и среднего значения, чуть больший межквартильный размах. Эти различия также пренебрежимо малы и на практике не имеют прямой связи.

- ИМТ – Регион

Чуть большие расхождения по ИМТ заметны по регионам. Например, в среднем, в ЮВ регионе ИМТ куда больше, чем в остальных. Из наблюдений, произведенных ранее, было сделано предположение, что в ЮВ регионе, возможно, хуже обстоят дела со здоровьем людей, о чем также может свидетельствовать индекс ИМТ (чем выше ИМТ - тем хуже здоровье). На 2 месте по ИМТ в среднем будет ЮЗ регион, а на 3 и 4 СЗ и СВ. Можно даже в общем сказать, что на Юге дела с ИМТ обстоят хуже, чем на Севере. Это также может быть связано с климатом.

Для формальной проверки гипотезы о наличии статистической связи был выполнен непараметрический дисперсионный анализ с помощью критерия Краскела-Уоллиса. Результат анализа сведен в таблицу 14.

Таблица 14 – Результат непараметрического дисперсионного анализа количества иждивенцев по качественным переменным

	Значение критерия, H	Значимость критерия, p
ИМТ - Пол	2070.992	0.00
ИМТ - Наличие курения	2143.277	0.00
ИМТ - Регион	2022.176	0.00

P значение для каждого из наборов меньше 0.05, значит, с уровнем доверия 0.95 во всех трех случаях отвергаем нулевую гипотезу, которая говорит о том, что связь между параметрами отсутствует.

3) Распределение возраста по половому признаку, наличию курения и регионам

Диаграммы распределения возраста по признакам представлены в виде Бокса-Уискера на рисунке 13.

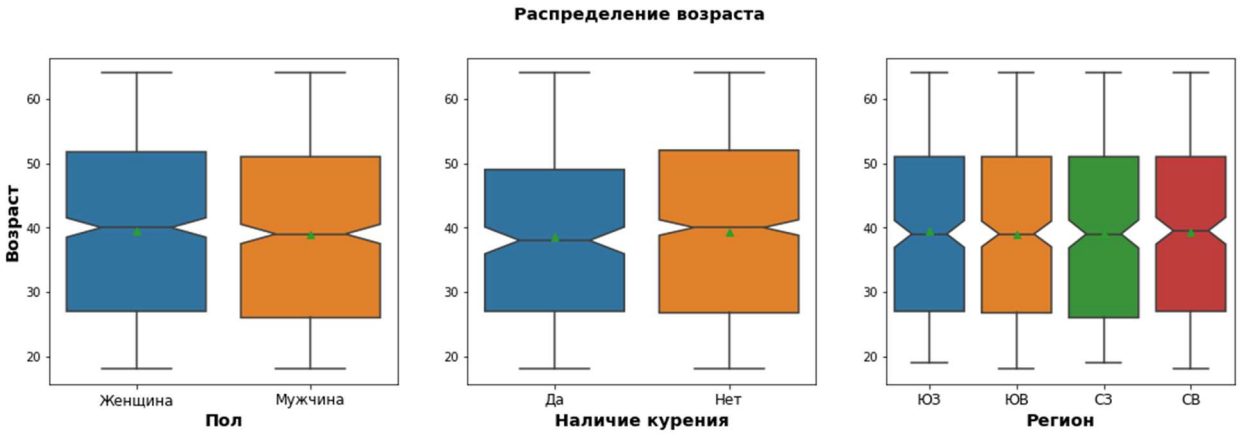

Рисунок 13 – Распределение возраст по качественным переменным

- Возраст – Пол

По диаграмме заметны крайне малые различия возраста относительно полов, что не является абсурдным, так как в выборке речь идет о текущем возрасте, а не о продолжительности жизни. Присутствуют совсем небольшие различия в межквартильном размахе и в медианном и среднем значении, но эти различия пренебрежимо малы.

- Возраст – Наличие курения

Достаточно похожая ситуация с наличием курения. Единственное различие - межквартильный размах у некурящих шире, а уровень верхней квартили - выше, что также не является абсурдным, интерпретировать это можно следующим образом: некурящий человек с большей вероятностью встретится в возрасте старше 50 лет, чем курящий человек. Возможно, это может быть связано с тем, что курящие люди живут меньше, чем некурящие.

- Возраст – Регион

Из диаграммы видно, что по регионам почти никакой разницы относительно возрастов нет. Уровень медианы и среднего значения во всех регионах примерно одинаковый, есть совсем несущественные различия по межквартильному размаху, но в целом, вполне

адекватная ситуация, что человек одного и того же возраста равновероятно может встретиться во всех регионах.

Для формальной проверки гипотезы о наличии статистической связи был выполнен непараметрический дисперсионный анализ с помощью критерия Краскела-Уоллиса. Результат анализа сведен в таблицу 15.

Таблица 15 – Результат непараметрического дисперсионного анализа возраста по качественным переменным

	Значение критерия, H	Значимость критерия, p
Возраст - Пол	2071.158	0.00
Возраст - Наличие курения	2143.455	0.00
Возраст - Регион	2022.334	0.00

P значение для каждого из наборов меньше 0.05, значит, с уровнем доверия 0.95 во всех трех случаях отвергаем нулевую гипотезу, которая говорит о том, что связь между параметрами отсутствует.