1 Предварительный анализ собранных данных

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

1.1.1 Анализ количественных переменных

Здесь необходимо построить и проанализировать гистограммы для всех количественных (интервальных и относительных) переменных в анализе. Необходимо охарактеризовать вид распределения по отношению к нормальному распределению — асимметрию, эксцесс, полимодальность. Для этого следует привести график гистограммы совместно с графиком плотности нормального распределения, а также таблицу основных статистик.

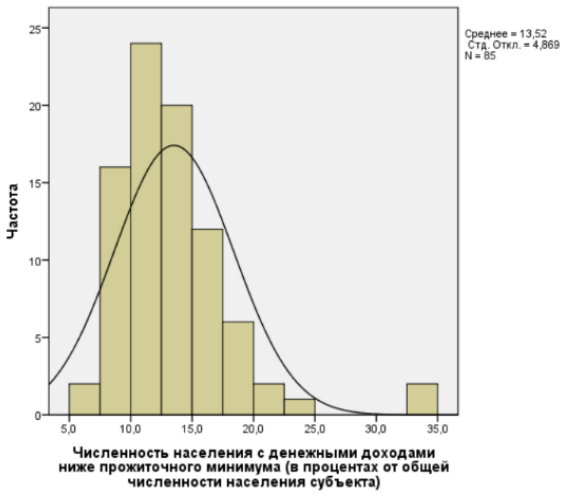

Рисунок 1. Численность населения ….

Статистика	Значение
Среднее	
Медиана	
Стандартное отклонение	
Межквартильный размах	
Верхняя квартиль	
Нижняя квартиль	
Коэффициент асимметрии	
Коэффициент эксцесса	
Количество наблюдений	
Количество пропущенных значений	

Таблица 2. Статистические свойства количественных факторов.

Необходимо дать интерпретацию статистических свойств количественных переменных в контексте предметной области. Например, (см. Рисунок 1), на основании гистограммы и числовых характеристик распределения можно сделать вывод о наличии небольшого количества субъектов федерации с очень большой долей бедного населения.

Также, для целевой переменной следует проанализировать наличие выбросов на основании правила «трех-сигм». Следует отметить в базе все выбросы и на основании сравнения соответствующих значений объясняющих переменных с их средними/медианными значениями объяснить, почему эти наблюдения могут интерпретироваться как выбросы.

1.1.2. Анализ качественных переменных.

Здесь следует привести столбчатые диаграммы, которые отражают количество измерений с разными уровнями для данной переменной.

		Частота	Проценты	Процент допустимых	Накопленный процент
Допустимо	Большая	13	15,3	15,3	15,3
	Выше среднего	23	27,1	27,1	42,4
	Маленькая	28	32,9	32,9	75,3
	Средняя	21	24,7	24,7	100,0
	Всего	85	100,0	100,0	

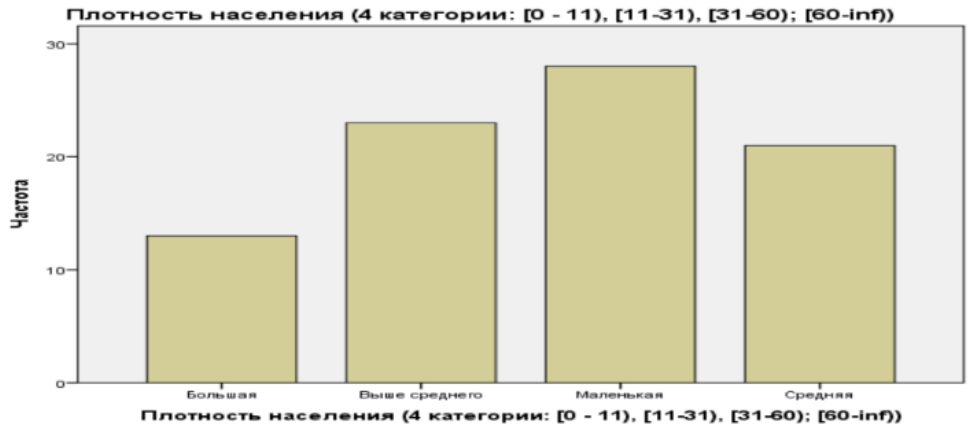

Рисунок 2. Плотность населения.

Необходимо проанализировать степень представленности всех уровней и при необходимости (наличии уровней с долей менее 5%) произвести укрупнение уровней. Результат привести на новых диаграммах. Принцип укрупнения пояснить.

1.2 Анализ статистической связи.

1.2.1 Графический анализ пары «целевая переменная – качественная объясняющая переменная».

Здесь для каждой пары {количественная зависимая переменная – качественная независимая переменная} необходимо построить категорированную диаграмму Бокса-Уискера (Box-Whisker).

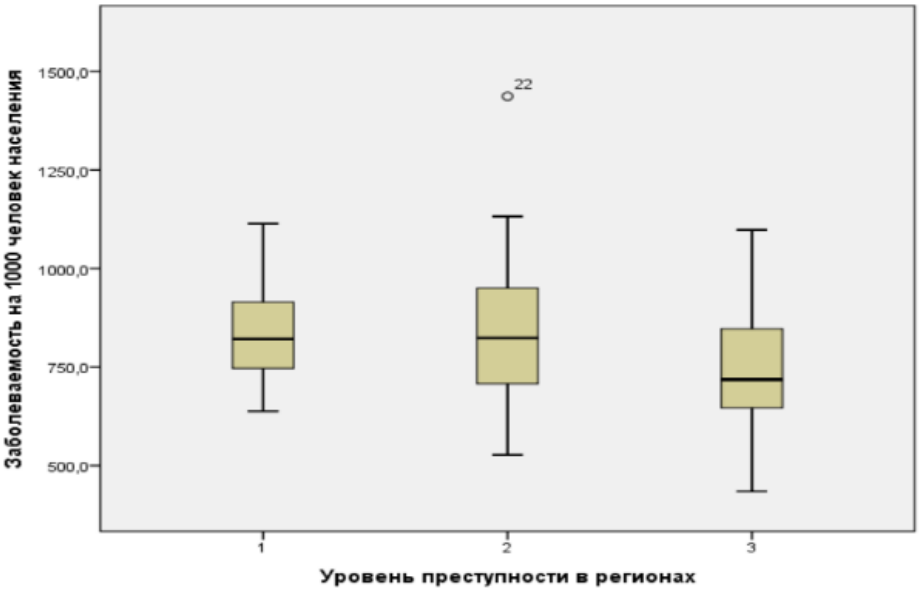

Рисунок 3. Зависимость заболеваемости от уровня преступности.

На основании анализа диаграммы следует охарактеризовать связь среднего значения и разброса количественной зависимой переменной с уровнями качественной независимой переменной. Интерпретацию дать в контексте предметной области.

1.2.2 Графический анализ пары «числовая зависимая переменная – числовая независимая переменная».

Здесь для каждой пары {количественная зависимая переменная – количественная независимая переменная} необходимо построить диаграммы рассеивания (Scatter plot).

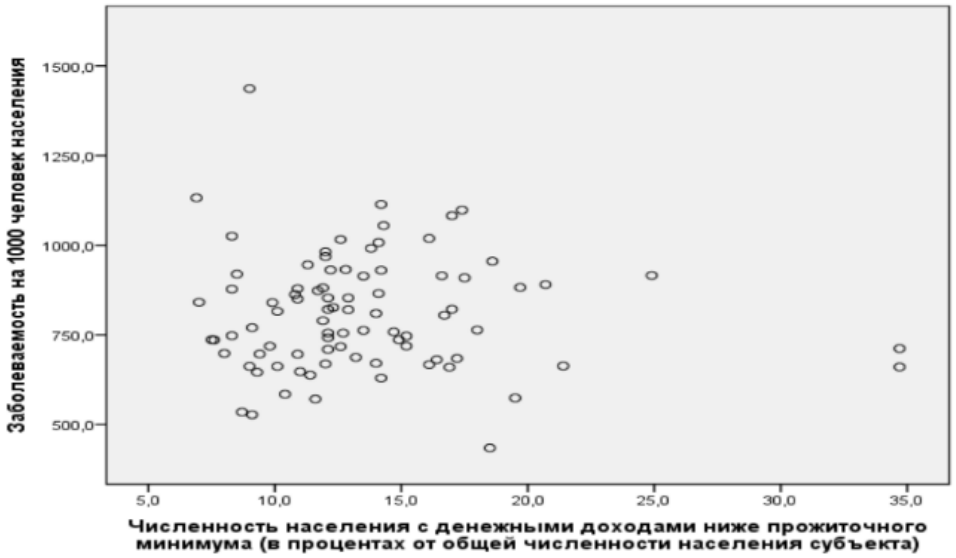

Рисунок 4. Зависимость заболеваемости от …

На основании визуального анализа диаграммы следует сделать предположение о наличии и характере статистической взаимосвязи. Интерпретацию результатов дать в контексте предметной области.

Для формальной проверки гипотезы о наличии связи следует подсчитать коэффициенты корреляции Пирсона и Спирмена и привести результаты проверки их значимости.

1.2.3 Анализ статистической взаимосвязи между независимыми переменными.

Следует проанализировать силу связи между независимыми переменными, используя инструменты пп. 2.2.1 и 2.2.2. Для анализа силы связи между качественными переменными следует использовать анализ таблиц: необходимо привести таблицу кросстабуляции, значения статистики хи-квадрат и V-Крамера.

1.2.4 Предварительная проверка гипотез

Здесь необходимо рассказать о результатах качественной проверки гипотез из п.1.3 на основании проведенного предварительного анализа данных.

Данные

Анализ ДТП в Москве 2017 год

Гипотезы

1. Количество погибших пешеходов и велосипедистов в поздневесенний, раннеосенний и летний периоды выше чем в позднеосенние, ранневесенние и зимние месяцы.
2. Количество произошедших ДТП в часы пик выше чем в остальные часы дня.
3. Количество погибших в пятницу и выходные дни выше чем в остальные дни.