

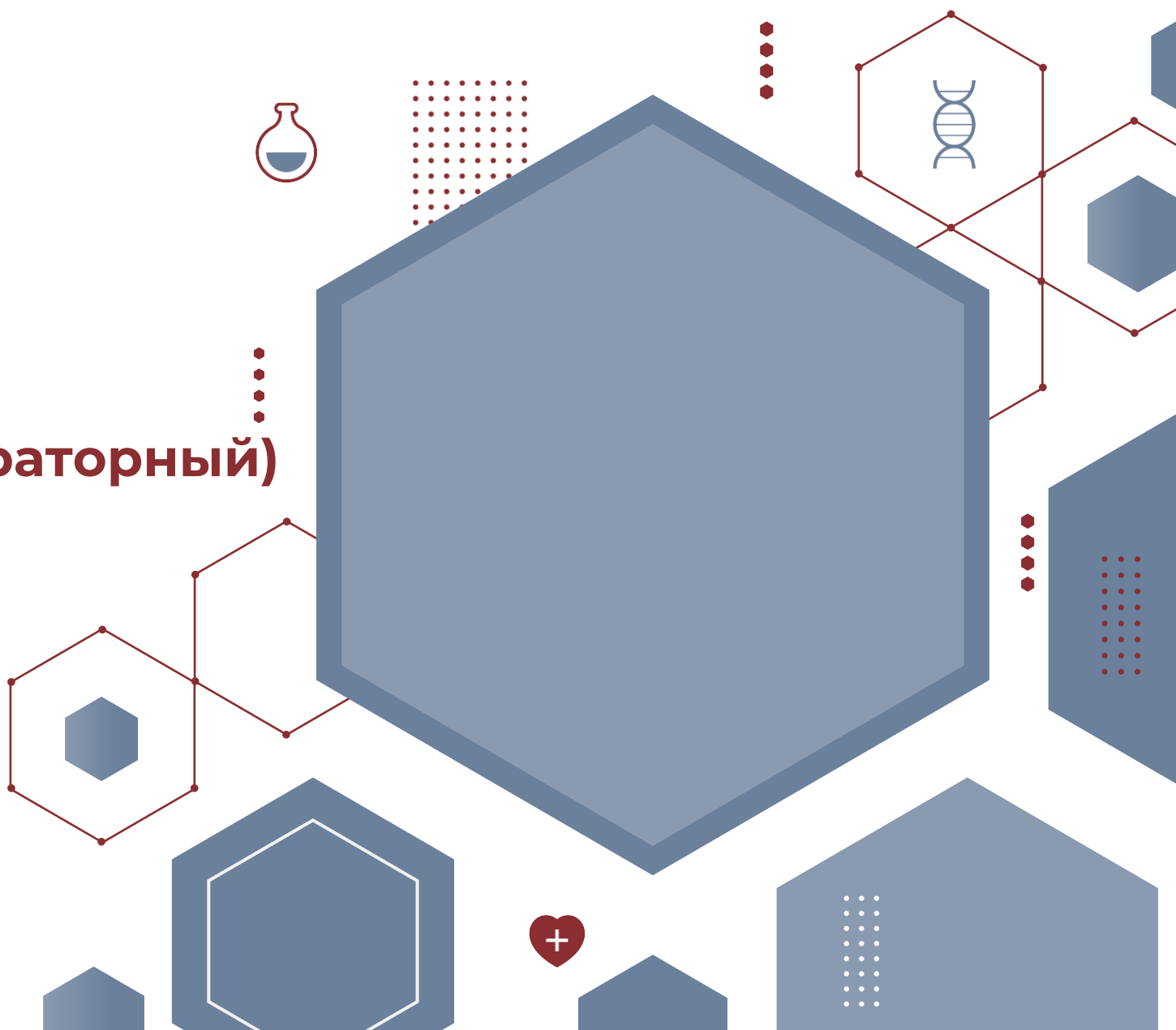
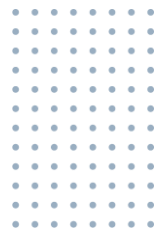


Казанский
Государственный
Медицинский
Университет

Предварительный (эксплораторный) анализ данных в R

Лакман Ирина Александровна

Казань, 2025



План лекции



Что такое эксплораторный анализ данных и зачем он нужен



Пакет эксплораторного анализа: finalfit



Пакет эксплораторного анализа: dlookr



Что такое эксплораторный анализ данных и зачем он нужен



Эксплораторный анализ – комплекс методов по предварительному анализу «сырой» (первичной) информации.

Эксплораторный анализ позволяет:

- определиться с дизайном исследования,
- понять какой набор инструментов моделирования можно применить к собранному набору данных,
- оценить достаточность собранной информации,
- выделить необходимость в дополнительном наборе недостающей информации

Эксплораторный анализ включает этапы:

- проверка полноты данных (отсутствия пропусков)
- проверка наличия аномальных наблюдений
- графический анализ
- выявление зависимостей между признаками (корреляционный анализ)

Пакет эксплораторного анализа: finalfit

```
library(finalfit)
ff_glimpse(melanoma)
#> $Continuous
#>      label var_type  n missing_n missing_percent  mean  sd  min
#> time      time    <dbl> 205      0          0.0 2152.8 1122.1 10.0
#> status    status  <dbl> 205      0          0.0   1.8   0.6   1.0
#> sex       sex     <dbl> 205      0          0.0   0.4   0.5   0.0
#> age       age     <dbl> 205      0          0.0   52.5  16.7   4.0
#> year      year    <dbl> 205      0          0.0 1969.9   2.6 1962.0
#> thickness thickness <dbl> 205      0          0.0   2.9   3.0   0.1
#> ulcer     ulcer   <dbl> 205      0          0.0   0.4   0.5   0.0
#>      quartile_25 median quartile_75  max
#> time      1525.0 2005.0      3042.0 5565.0
#> status      1.0   2.0         2.0   3.0
#> sex         0.0   0.0         1.0   1.0
#> age         42.0  54.0        65.0  95.0
#> year      1968.0 1970.0      1972.0 1977.0
#> thickness   1.0   1.9         3.6  17.4
#> ulcer       0.0   0.0         1.0   1.0
#>
#> $Categorical
#> data frame with 0 columns and 205 rows
```



Библиотека **finalfit** подходит для подготовки практически любого набора данных

Функция **ff_glimpse()** обеспечивает удобный обзор всех данных в виде таблицы или фрейма данных.

ff_glimpse() разделяет переменные на непрерывные и категориальные.

Появление переменных в том порядке, в котором они появляются во фрейме данных или в таблице осуществляется с помощью команд **missing_glimpse()** или **tibble::glimpse()**

```
missing_glimpse(melanoma)
#>      label var_type  n missing_n missing_percent
#> time      time    <dbl> 205      0          0.0
#> status    status  <dbl> 205      0          0.0
#> sex       sex     <dbl> 205      0          0.0
#> age       age     <dbl> 205      0          0.0
#> year      year    <dbl> 205      0          0.0
#> thickness thickness <dbl> 205      0          0.0
#> ulcer     ulcer   <dbl> 205      0          0.0
```

Пакет эксплораторного анализа: **finalfit**

В ходе анализа пропущенных значений у зависимых и независимых переменных, функция **missing_compare()** (библиотека **finalfit**) позволяет определить статистическую значимость различий в количестве пропущенных значений у зависимой переменной по всем объясняющим переменным.

```
missing_compare(data, dependent, explanatory, p = TRUE, na_include = FALSE)
```

Набор данных

Переменная, с помощью которой проверяют пропуски других переменных

Переменные, по которым будут проверяться пропуски



Пакет эксплораторного анализа: finalfit

Также библиотека позволяет провести структурные и аналитические группировки (в том числе многофакторные) и вычислить автоматически между признаками различия согласно статистическим тестам.

```
explanatory = c("age.factor", "sex.factor", "obstruct.factor")
```

```
dependent = 'mort_5yr'colon_s %>%
```

```
summary_factorlist(dependent, explanatory, p=TRUE,
```

```
add_dependent_label=TRUE) -> t1
```



Dependent: Mortality 5 year		Alive	Died	pvalue
Age	<40 years	31 (46.3)	36 (53.7)	0.020
	40-59 years	208 (61.4)	131 (38.6)	
	60+ years	272 (53.4)	237 (46.6)	
Sex	Female	243 (55.6)	194 (44.4)	0.889
	Male	268 (56.1)	210 (43.9)	
Obstruction	No	408 (56.7)	312 (43.3)	0.189
	Yes	89 (51.1)	85 (48.9)	
Perforation	No	497 (56.0)	391 (44.0)	0.671
	Yes	14 (51.9)	13 (48.1)	

Пакет эксплораторного анализа: dlookr

Диагностика, исследование и преобразование данных с помощью библиотеки эксплораторного анализа **dlookr**.

Функции:

Диагностика качества данных.

Нахождение подходящих сценариев для проведения последующего анализа путем изучения и понимания данных.

Получение новых переменных или выполнение преобразования переменных.

Автоматическое создание отчетов для трех вышеуказанных задач.

- **diagnose()** позволяет диагностировать переменные в фрейме данных. Как и в любых других dplyr функциях, первым аргументом является тиббл (или фрейм данных). Второй и последующие аргументы относятся к переменным внутри фрейма данных.



Пакет эксплораторного анализа: dlookr

Переменные объекта, `tbl_df` возвращаемые, `diagnose()` следующие.

- `variables`: имена переменных
- `types`: тип данных переменных
- `missing_count`: количество пропущенных значений
- `missing_percent`: процент пропущенных значений
- `unique_count`: количество уникальных значений
- `unique_rate`: ставка уникального значения. $\text{unique_count} / \text{количество наблюдений}$



Пакет эксплораторного анализа: dlookr

diagnose_numeric() диагностирует числовые (непрерывные и дискретные) переменные в кадре данных. Использование такое же diagnose(), но возвращает больше диагностической информации. Однако если вы укажете нечисловую переменную во втором и последующих списках аргументов, эта переменная автоматически игнорируется.

Переменные объекта, `tbl_df`возвращаемые, diagnose_numeric()следующие.

- `min`: минимальное значение
- `Q1`: 1/4 квартиля, 25-й процентиль
- `mean`: среднее арифметическое
- `median`: медиана, 50-й процентиль

diagnose_category() диагностирует категориальные (факторные, упорядоченные, символьные) переменные набора данных. Использование аналогично diagnose(), но возвращает больше диагностической информации.

функции **diagnose_outlier()** комплексного эксплораторного анализа позволяет производить диагностику выбросов





**Спасибо за
внимание!**