002

Практическая работа

Информационно-аналитические технологии поиска угроз инорфмационной безопасности

Основы обработки данных с помощью R и Dplyr

Цель работы

1. Развить практические навыки использования языка программирования R для обработки данных
2. Закрепить знания базовых типов данных языка R
3. Развить практические навыки использования функций обработки данных пакета `dplyr` – функции `select(), filter(), mutate(), arrange(), group_by()`

Общая ситуация

Используя R и среду разработки RstudioIDE, выполнить задания

> ⚠ **Данные**
>
> Данные хранятся (встроены) в пакет `dplyr`, для доступа к ним нужно импортировать пакет.

Проанализировать встроенный в пакет `dplyr` набор данных `starwars` с помощью языка R и ответить на вопросы:

1. Сколько строк в датафрейме?

```
starwars %>% nrow()
```

2. Сколько столбцов в датафрейме?

```
starwars %>% ncol()
```

3. Как просмотреть примерный вид датафрейма?

```
starwars %>% glimpse()
```

4. Сколько уникальных рас персонажей (species) представлено в данных?
5. Найти самого высокого персонажа.
6. Найти всех персонажей ниже 170
7. Подсчитать ИМТ (индекс массы тела) для всех персонажей. ИМТ подсчитать по формуле

$$I = \frac{m}{h^2}$$

, где m – масса (weight), а h – рост (height).

8. Найти 10 самых “вытянутых” персонажей. “Вытянутость” оценить по отношению массы (mass) к росту (height) персонажей.
9. Найти средний возраст персонажей каждой расы вселенной Звездных войн.
10. Найти самый распространенный цвет глаз персонажей вселенной Звездных войн.
11. Подсчитать среднюю длину имени в каждой расе вселенной Звездных войн.
12. Оформить отчет в соответствии с шаблоном

Рекомендации по выполнению работы

1. Для загрузки библиотеки используйте

```
library(dplyr)
```

2. Для создания конвейера применения функций можно использовать либо встроенный в R 4.1 оператор `|>` или оператор пайпа из состава `dplyr` – `%>%`.
3. Для 4 задания можно использовать функцию `unique()`.
4. Символ `NA` – пропуск данных, означает что по каким-либо причинам данных нет. Часто, для корректных вычислений над данными, пропуски приходиться учитывать по разному – или пропускать такие данные, или замещать похожими.

> **Tip**
>
> Дополнительные материалы можно найти в Telegram https://t.me/datadrivencybersec

Отчет

Для оформления отчета используйте следующие материалы:

1. https://i2z1.ddslab.ru/posts/lab_recommendations/
2. https://i2z1.quarto.pub/checklab/criteria.html
3. https://github.com/i2z1/Report_template

Сайт курса

https://i2z1.ddslab.ru/IAMCTH