003

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 пакете `nycflights13`, для доступа к ним нужно установить иимпортировать пакет.

Проанализировать встроенные в пакет `nycflights13` наборы данных с помощью языка R и ответить на вопросы:

1. Сколько встроенных в пакет `nycflights13` датафреймов?
2. Сколько строк в каждом датафрейме?
3. Сколько столбцов в каждом датафрейме?
4. Как просмотреть примерный вид датафрейма?
5. Сколько компаний-перевозчиков (carrier) учитывают эти наборы данных (представлено в наборах дан- ных)?
6. Сколько рейсов принял аэропорт John F Kennedy Intl в мае?
7. Какой самый северный аэропорт?
8. Какой аэропорт самый высокогорный (находится выше всех над уровнем моря)?
9. Какие бортовые номера у самых старых самолетов?
10. Какая средняя температура воздуха была в сентябре в аэропорту John F Kennedy Intl (в градусах Цельсия).
11. Самолеты какой авиакомпании совершили больше всего вылетов в июне?

12. Самолеты какой авиакомпании задерживались чаще других в 2013 году?

13. Оформить отчет в соответствии с шаблоном

Рекомендации по выполнению работы

1. Для установки пакета используйте `install.packages('nycflights13')`
2. Для загрузки библиотеки используйте `library(nycflights13)`
3. Для просмотра объектов в пакете используйте `nycflights13::` в консоли R.
4. Для просмотра структуры датафрейма используйте glimpse() из состава пакета `dplyr`.
5. Для создания конвейера применения функций можно использовать либо встроенный в R 4.1 оператор `|>` или оператор пайпа из состава dplyr -- `%>%`. Для перевода из шкалы Фаренгейта в шкалу Цельсия используйте формулу $t_C = \frac{5}{9} \times (t_F - 32)$.
6. 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