Практическая работа №6
Манипулирование данными. Набор пакетов “tidyverse”

Цель: научиться использовать набор пакетов “tidyverse” для манипуляции данными.

Задания

1. Установить пакет **tidyverse**.
2. В среду **R** загрузить пакет **readr**.
3. В переменную **company** загрузить данные из таблицы **company2.csv,** созданной на практической работе №1, используя функции пакета **readr**.
4. Вывести в отдельном окне редактора кода данные переменной **company**.
5. Посмотреть тип данных переменной **company**.
6. Установить и загрузить пакет **reshape2**.
7. Присвоить новой переменной **company1** "расплавленные" в "длинный" формат данные, указав в качестве идентификационных переменных **возраст** и **общий стаж работы**.
8. Посмотреть первые 7 строк таблицы **company1**.
9. Сохранить в переменную **company2** "собранные" в "широкий" формат данные переменной **company1**.
10. Посмотреть первые 7 строк таблицы **company2**.
11. Сравнить наборы данных **company1** и **company2**.
12. Загрузить пакет **dplyr**.
13. Работа с оператором **pipe %>%**.
    - 13.1. Вывести на экран с 4 по 8 строки набора данных **company** с помощью команды **head()**.
    - 13.2. Посмотреть с 4 по 8 строки набора данных **company,** используя оператор **pipe %>%**.
    - 13.3. Сравнить полученные результаты, используя логическое равенство.
14. Работа с функцией фильтрации строк: **filter().**
    - 14.1. В качестве переменной, по которой необходимо выполнить фильтрацию, взять стаж работы в данной компании (в наборе данных **company** пять вариантов стажа: 1-5, выбрать по варианту,

данному преподавателем).

14.2. Отфильтровать таблицу **company** двумя способами (с использованием и без использования оператора **pipe %>%**). Сравнить являются ли, полученные в результате фильтрации, таблицы одинаковыми.

14.3. Отфильтровать набор данных **company** используя 2 условия: стаж работы сотрудников компании и возраст (критерий для возраста выбрать самостоятельно). Показать первые 4 строки полученной таблицы данных.

14.4. Выбрать из таблицы **company** со 2 по 6 строки с помощью оператора **pipe %>%** и функции **slice()**.

15. Работа с функцией упорядочения строк: **arrange().**

15.1. Упорядочить строки набора данных **company** по двум, самостоятельно выбранным, переменным, используя функцию **arrange()** и оператор **pipe %>%**.

15.2. Упорядочить строки таблицы **company** по убыванию переменной (переменную выбрать самостоятельно), используя функцию **desc()**.

16. Работа с функцией выбора колонок: **select().**

16.1. Выбрать самостоятельно любые 3 колонки из таблицы данных **company**.

16.2. Изучить вспомогательные функции **contains(), ends_with()**, **starts_with()** и **matches()**, привести примеры работы с ними.

17. Работа с функцией создания новых колонок: **mutate()**.

17.1. Используя функцию **mutate()**, создать в наборе данных **company** новый столбец, который будет содержать процент стажа работы в данной компании от общего стажа работы.

18. Установить пакет **gapminder** и загрузить его в среду программирования **R**.

19. Присвоить новой переменной **vvp** данные таблицы **gapminder** (входит в пакет **gapminder**).

20. Выбрать из таблицы **vvp** данные за 1952 год (переменная **year**) со средней продолжительностью жизни больше 60 лет (переменная **lifeExp**). Сохранить эти данные в таблицу **vvp1952**.

21. Данные таблицы **vvp1952** отсортировать по убыванию населения (переменная **pop**).
22. Из таблицы данных **vvp** выбрать переменные: ВВП (**gdpPercap**), континент (**continent**), год (**year**).
23. В набор данных **vvp** добавить столбец с продолжительностью жизни, указанной в днях (принять год равным 365 дней).
24. Для каждой страны (**country**) посчитать среднее значение ВВП. Отсортировать результаты по убыванию.
25. Разбить набор данных **vvp** на две таблицы: до 1980 г., после 1980 г.
26. Выбрать переменные **lifeExp**, **country**.
27. Вычислить для каждой страны среднюю продолжительность жизни.
28. Используя **inner_join** объединить две таблицы по стране.

Контрольные вопросы:

1. Перечислить функции пакета **reshape2**.
2. "**tidyverse**" – что это?
3. Перечислите особенности пакета **readr**.
4. Что такое **tbl_df**?
5. Что представляет собой "**tibble**"?
6. Используя какую функцию, можно "собрать" данные из длинного в широкий формат?
7. Назовите этапы работы пакета **dplyr**.
8. Назовите главное предназначение оператора **pipe %>%**?
9. Перечислите основные функции манипулирования данными, дайте краткую справку по каждой из них.

Домашнее задание

1. Оформить отчет по практической работе (структуру отчета взять из практической работы №3)
2. Изучить работу: **base::transform(), dplyr::mutate(), plyr::mutate()**.
3. Изучить операции связывания таблиц (**left_join**(), **full_join**(), **inner_join**(), **right_join**(), **anti_join**(), **semi_join**()).
4. Создать две таблицы на тематику, связанную с бизнесом.
5. Используя функцию **inner_join**, связать таблицы в одну.