МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное

образовательное учреждение высшего образования

«ЧЕРЕПОВЕЦКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Институт информационных технологий

*наименование института(факультета)*

Кафедра математического и программного обеспечения ЭВМ

*наименование кафедры*

Основы Data Science

*наименование дисциплины в соответствии с учебным планом*

ЛАБОРАТОРНАЯ РАБОТА №5

«Визуализация данных»

|  |  |
| --- | --- |
| Исполнитель |  |
| Студент | 1ПИб-02-3оп-22 |
|  | *группа* |
|  | Маркелов С. А. |
|  | *ФИО* |
| Руководитель | Юдина О.В. |
|  | *ФИО преподавателя* |
| Оценка |  |
| Подпись |  |

2024 год

Цель: Знакомство с этапом понимания данных стандарта CRISP-DM.

Задача этапа – представить набор данных, использованный в лабораторных работах. Описать основные закономерности, которые в них содержатся, связи между этими данными, фактами, знаниями, получить что-то новое, понять, что они могут дать и визуализировать свои представления

Задание

Вам нужно представить результаты изучения этих данных руководству предприятия, чтобы решить в каком направлении строить дальнейшую кадровую политику или проводить исследование. Представьте в диаграммах имеющиеся данные, на основании которых возможно построить исследование. Вам нужно показать:

* Закономерности
* Структуры
* Возможные связи
* Влияние факторов друг на друга
* Если вам удалось построить модели – покажите их

Все диаграммы должны иметь комментарии: что вы хотели ею показать, какие выводы сделали

Требования

Диаграммы должны:

* Иметь подписи по осям и заголовки, отражающие их суть
* Размерности должны быть нанесены на оси
* Быть разными
* Их тип должен быть пригодным для представления типа данных, которые вы показываете
* Они должны отражать ваши данные, их связи, закономерности, которые вам удалось обнаружить.

Для построения можно использовать любые пригодные программы – например, Excel. Если вам удалось найти доступные программы для визуализации данных, то можно использовать их.

В ходе проведения исследования был изучен набор данных о 99 студентах, включающий в себя информацию об их поле, успеваемости, количестве часов сна, количестве друзей, наличии депрессии и некоторые другие показатели. Основная задача исследования – выяснить, какие факторы влияют на успеваемость студентов. Данные об успеваемости студентов представлены на рис. 1.

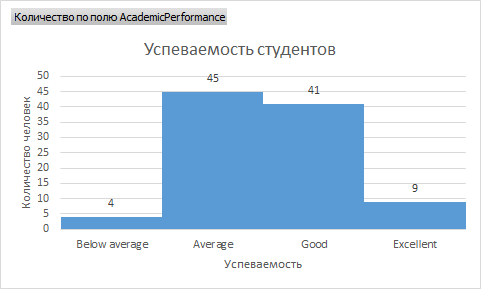


Рис. 1. Успеваемость студентов

Из представленной выше гистограммы следует, что большинство студентов учатся средне (Average) и хорошо (Good). Лишь некоторые учатся на отлично (Excellent) или, напротив, ниже среднего (Below average).

В ходе исследования был выдвинут ряд гипотез, чтобы выявить взаимосвязи между переменными и сделать выводы о влиянии тех или иных факторов на успеваемость студентов.

1.

H0 – У людей без проблем по учебе среднее количество часов сна такое же, как и у людей с проблемами

H1 – У людей без проблем по учебе среднее количество часов сна отличается по сравнению с людьми с проблемами

Перед проверкой гипотезы ознакомимся с данными. На рис. 2 представлена диаграмма, показывающая количество студентов, испытывающих проблемы с учебой.



Рис. 2. Количество студентов, которые испытывают проблемы с учебой

Количество студентов, испытывающих проблемы постоянно, иногда, или не испытывающих проблемы вовсе, распределилось почти поровну. Но тех, кто испытывает проблемы постоянно, немного больше (38%).

На рис. 3 представлена диаграмма, показывающая, сколько спят студенты (рис. 3).

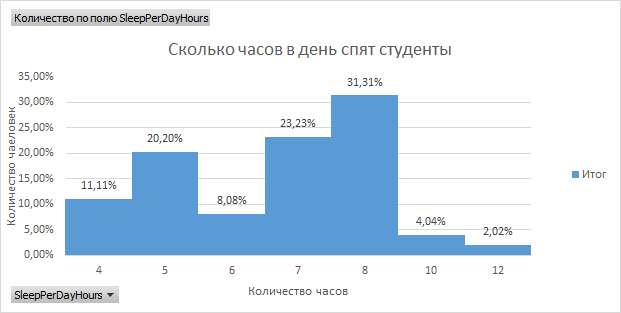


Рис. 3. Количество часов сна у студентов

Данная гистограмма показывает, сколько часов в день спят студенты. Почти треть студентов (31,31%) спит 8 часов в день. Лишь 2,02% студентов спят 12 часов в день.

Для проверки гипотезы средствами Excel был проведен 2-выборочный t-тест с одинаковыми дисперсиями. По итогам теста данная гипотеза подтвердилась.

Мы можем сделать вывод, что люди без проблем с учебой спят столько же, сколько и люди с проблемами. То есть взаимосвязи между количеством часов сна и проблемами с учебой нет. На графике (рис. 5) продемонстрированы средние значения количества часов сна.



Рис. 5. Зависимость наличия проблем с учебой от количества часов сна

На графике видно, что средние значения количества часов сна для всех категорий наличия проблем с учебой находятся в близком диапазоне (от 6,35 до 6,94 часа), это указывает на отсутствие значимой зависимости между количеством часов сна и наличием проблем с учебой.

2.

H0 – Не существует связь между успеваемостью и наличием депрессии

H1 – Существует связь между успеваемостью и наличием депрессии

Перед проверкой данной гипотезы ознакомимся с данными о наличии депрессии у студентов (рис. 10)

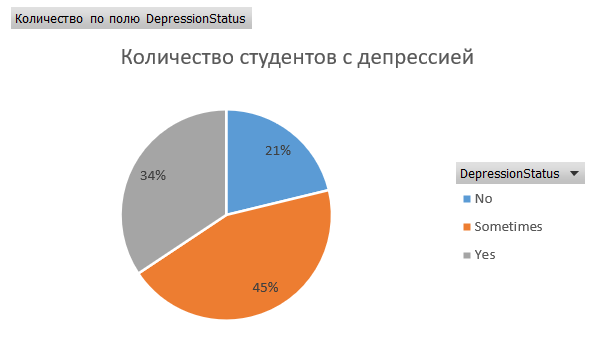


Рис. 10 Количество студентов с депрессией

Почти половина студентов (45%) испытывают депрессию иногда, еще 34% испытывают ее постоянно. Лишь 21% студентов депрессией не страдает.

Для проверки гипотезы средствами Excel был проведен тест хи-квадрат. По результатам теста гипотеза подтвердилась.

Мы можем сделать вывод, что зависимости между успеваемостью и наличием депрессии нет, т. е. депрессия не влияет на успеваемость. Зависимость успеваемости от количества часов сна визуализирована на диаграмме (рис. 12).

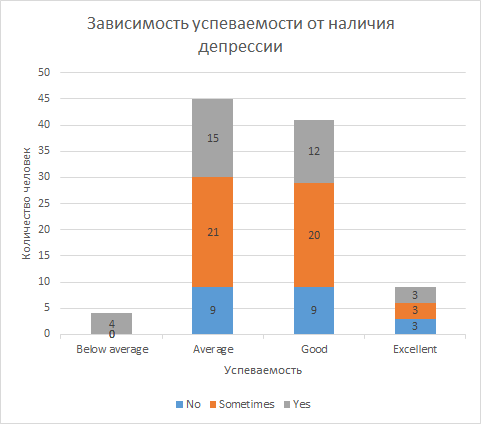


Рис. 12. Зависимость успеваемости от наличия депрессии

На данной диаграмме распределение людей с депрессией и без депрессии в категориях «Average» и «Good» довольно равномерное, что указывает на отсутствие явной зависимости между успеваемостью и наличием депрессии. В категориях «Below average» и «Excellent» небольшое количество людей, что делает выводы менее надежными, но также не показывает явной зависимости.

3.

H0 – Не существует связь между успеваемостью и полом

H1 – Существует связь между успеваемостью и полом

Процентное соотношение мужчин и женщин среди студентов в выборе представлено на рис. 13.



Рис. 13. Количество мужчин и женщин среди студентов

Для проверки гипотезы средствами Excel был проведен тест хи-квадрат. По результатам теста гипотеза подтвердилась.

Мы можем сделать вывод, что зависимости между успеваемостью и полом нет, т. е. мужчины и женщины учатся примерно одинаково. Зависимость успеваемости от пола визуализирована на диаграмме (рис. 15).

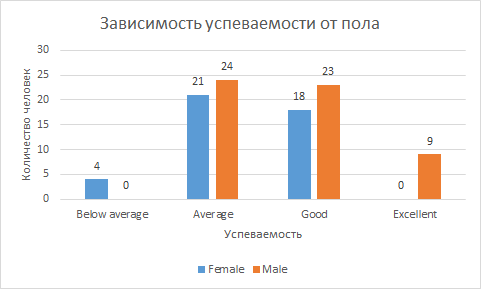


Рис. 15. Зависимость успеваемости от пола

На диаграмме в категории «Average» мы видим почти равное количество женщин (21) и мужчин (24), что указывает на отсутствие значительной зависимости между успеваемостью на данном уровне и полом. В категории «Good» женщин (18) чуть меньше, чем мужчин (23), но разница незначительная и не указывает на явную зависимость. По остальным категориям сложно сделать значимые выводы, т. к. количество людей невелико и данных недостаточно.

4.

H0 – Не существует связь между успеваемостью и ведением конспектов

H1 – Существует связь между успеваемостью и ведением конспектов

Данные о ведении конспектов студентами представлены на рис. 16.

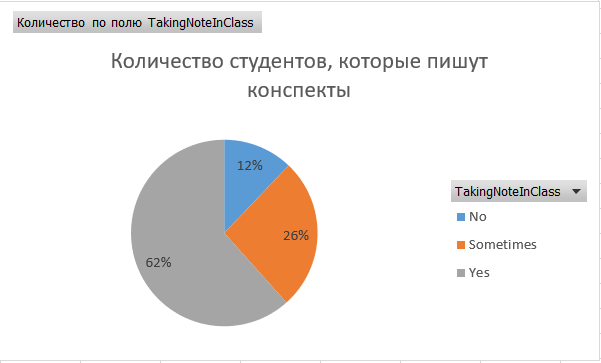


Рис. 16. Количество студентов, которые ведут конспекты

Абсолютное большинство студентов (62%) конспекты ведут, еще 26% делают это иногда. Лишь 12% вообще не ведут конспекты.

Для проверки гипотезы средствами Excel был проведен тест хи-квадрат. По результатам теста гипотеза подтвердилась.

Мы можем сделать вывод, что зависимости между успеваемостью и ведением конспектов нет. Зависимость успеваемости от ведения конспектов визуализирована на диаграмме (рис. 18).

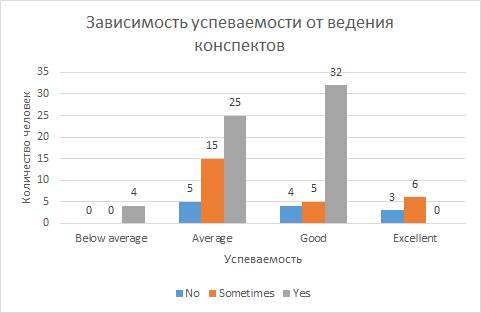


Рис. 18. Зависимость успеваемости от ведения конспектов

На графике видно, что в категориях «Average» и «Good» представлены все три группы: не ведут конспекты 5 и 4 человека соответственно, иногда ведут 15 и 5, всегда ведут 25 и 32. Это прямо указывает на то, что ведение конспектов не является определяющим фактором для успеваемости. В остальных категориях количество людей невелико, поэтому сделать значимые выводы сложно.

Вывод:

По итогам исследования мы можем сделать следующие выводы:

* У людей без проблем по учебе среднее количество часов сна такое же, как и у людей с проблемами
* Не существует связь между успеваемостью и наличием депрессии
* Не существует связь между успеваемостью и полом
* Не существует связь между успеваемостью и наличием конспектов

Эти выводы были получены путем проверки гипотез с помощью t-, z- и хи2-тестов.

Для большей наглядности и упрощения восприятия данные и результаты тестов были визуализированы в виде диаграмм и графиков.

Контрольные вопросы:

1. Приведите примеры диаграмм, подходящих для качественных переменных.

Для качественных переменных наиболее подходящими диаграммами являются столбчатая диаграмма, круговая диаграмма, гистограмма, точечная диаграмма, ящик с усами (box plot)

1. Какие диаграммы можно отнести к диаграммам в пиксельной технике?

К диаграммам в пиксельной технике можно отнести графики, диаграммы рассеяния, хитмапы и карты.

1. В чем разница между визуализацией информации и визуализацией данных?

Визуализация информации помогает представить сложную информацию в более доступном и понятном виде, а визуализация данных используется для исследования и анализа этих данных, чтобы выявить скрытые закономерности и зависимости.

1. Какие диаграммы основаны на геометрической технике?

Диаграммы, основанные на геометрической технике, включают в себя такие диаграммы, как диаграммы рассеяния, радарные диаграммы, трехмерные диаграммы, пузырьковые диаграммы и диаграммы Ганта.

1. Что можно визуализировать с помощью иконок? Приведите пример.

С помощью иконок можно визуализировать различные данные, например, количество продаж, оценки, рейтинги и т.д. Примером такой диаграммы могут служить графики с использованием иконок вместо столбиков или линий.