МИНОБРАНУКИ РОССИИ

федеральное государственное бюджетное

образовательное учреждение высшего образования

«ЧЕРЕПОВЕЦКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Институт информационных технологий

*наименование института (факультета)*

Математическое и программное обеспечение ЭВМ

*наименование кафедры*

Основы Data Science

*наименование дисциплины в соответствии с учебным планом*

ЛАБОРАТОРНАЯ РАБОТА №1

Тема: Описание структуры данных

|  |  |
| --- | --- |
| Руководитель | Юдина О.В. |
|  | ФИО преподавателя |
| Исполнитель |  |
| студент | 1ПИб-02-1оп-22 |
|  | группа |
|  | Маслов В.А. |
|  | Фамилия, имя, отчество |
| Оценка |  |
| Подпись |  |

2025 год

Цель работы: обзор основных стандартов работы с данными. Знакомство с этапом понимания данных стандарта CRISP-DM.

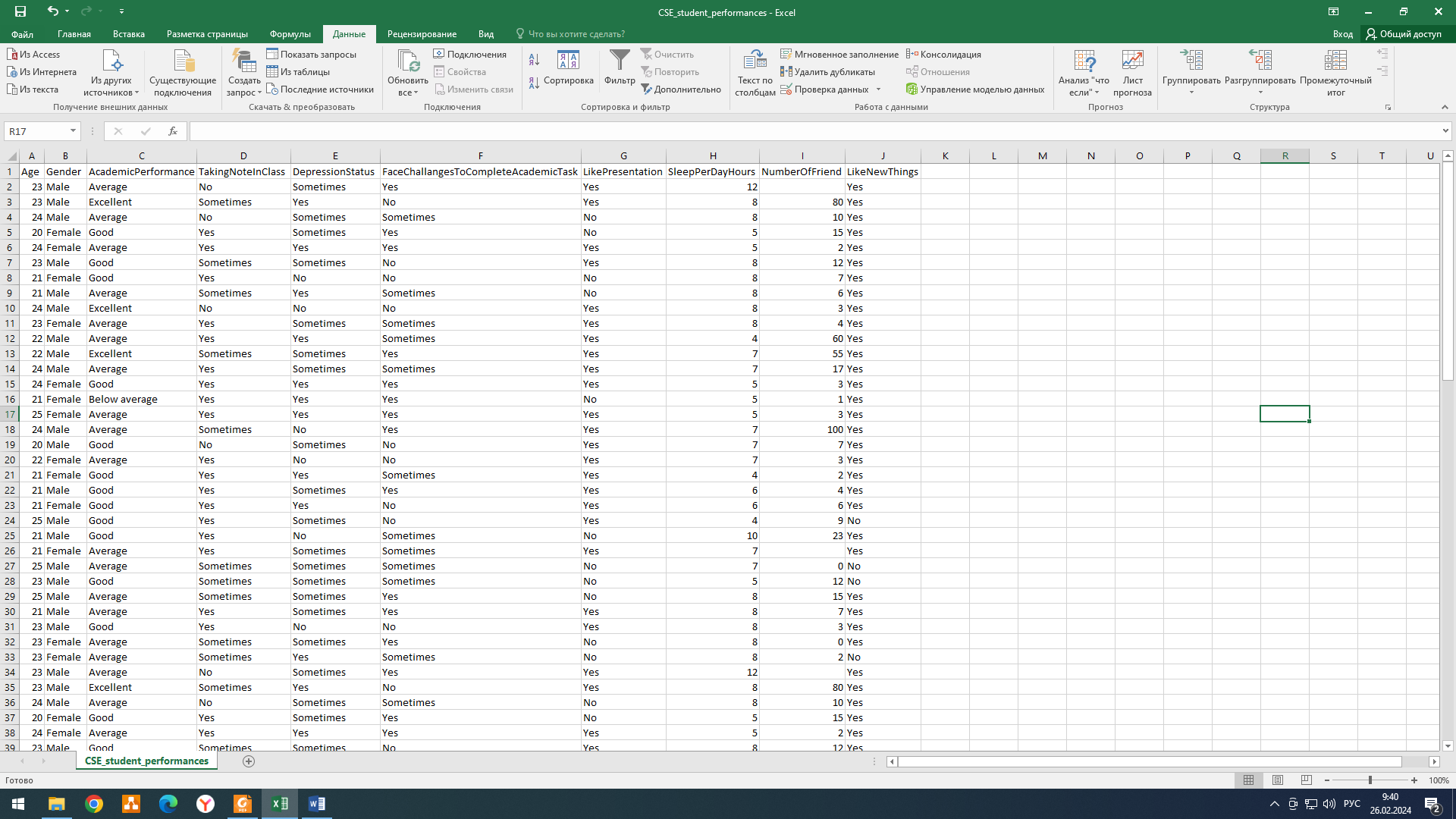
Задача: понять, что именно содержится в этих данных, какие факты они описывают; понять, что они могут дать.

Задание:

1. Откройте файл с данными. Приведите его к формату, подходящему для обработки. Он содержит данные о студентах одного из учебных заведений.
2. Выясните структуру данных. Для этого:
3. Для каждой переменной определите возможные шкалы, выберите шкалу, которая даст возможность получить больше информации. (Для категориальных переменных укажите – номинальными или порядковыми они являются. Для количественных уточните, по какой именно из числовых шкал они могут быть измерены.).
4. Используя Excel, постройте сводные таблицы для категориальных переменных и таблицы частот для количественных переменных.
5. Постройте диаграммы, которые иллюстрируют структуру, для качественных переменных (4-6 штук, с заголовками, подписями осей и пояснениями, что они показывают).
6. Для количественных переменных постройте гистограммы и графики (кумулятивный, частотный, простой, диаграмма рассеяния - какие будут уместны) которые иллюстрируют распределения (4-6 штук, с заголовками, подписями осей и пояснениями, что они показывают).
7. Сформулируйте, какие задачи могут решаться с применением этих данных. Приведите несколько формулировок.

Ход работы

1. Откроем файл с данными в Excel. Приведем его к формату, подходящему для обработки, с помощью функции «Текст по столбцам»

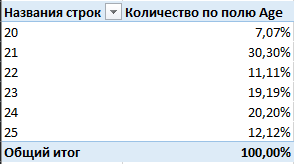
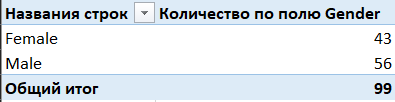


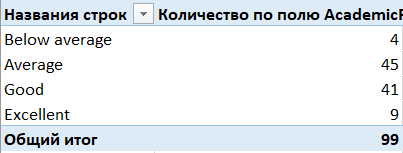
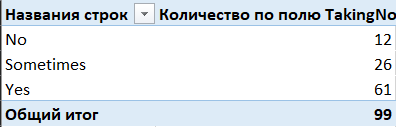
Данная таблица содержит следующие переменные:

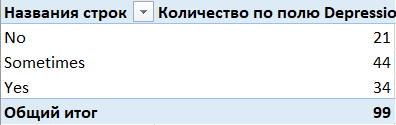
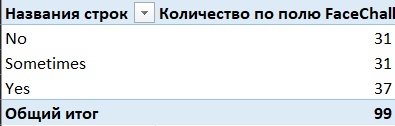
* **Age** – возраст, количественный признак, порядковая шкала.
* **Gender** – пол, качественный номинальный признак.
* **AcademicPerfomance** – успеваемость, качественный порядковый признак
* **TakingNoteInClass** – ведет ли студент конспекты, качественный номинальный признак.
* **DepressionStatus** – наличие депрессии, качественный номинальный признак.
* **FaceChallengesToCompleteAcademicTask** – сталкивается ли студент с проблемами в учебе, качественный номинальный признак.
* **LikePresentation** – любит ли студент публичные выступления, качественный номинальный признак.
* **SleepPerDayHours** – количество часов сна, количественный признак, порядковая шкала.
* **NumberOfFriend** – количество друзей, количественный признак, порядковая шкала.
* **LikeNewThings** – любит ли студент что-то новое, качественный номинальный признак.

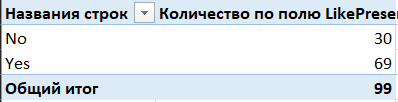
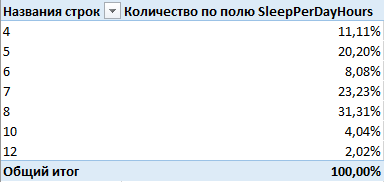
1. Построим сводные таблицы для качественных признаков и таблицы частот для количественных признаков.

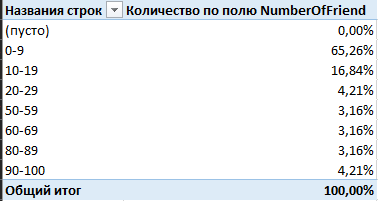
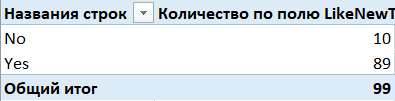
Значения с % – количество по полю с дополнительным вычислением % от общей суммы, остальные без вычислений.

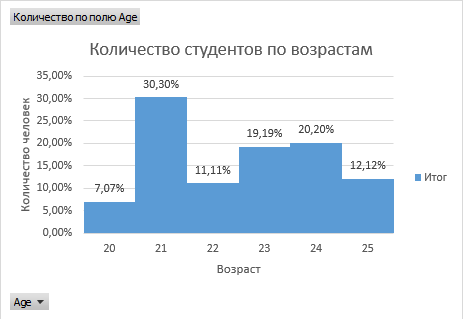
 

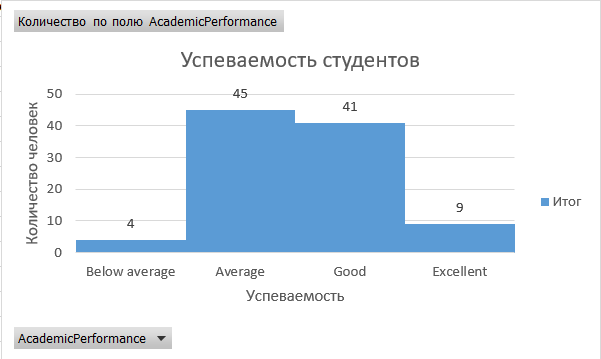
1. Построим гистограммы, диаграммы и графики для переменных.



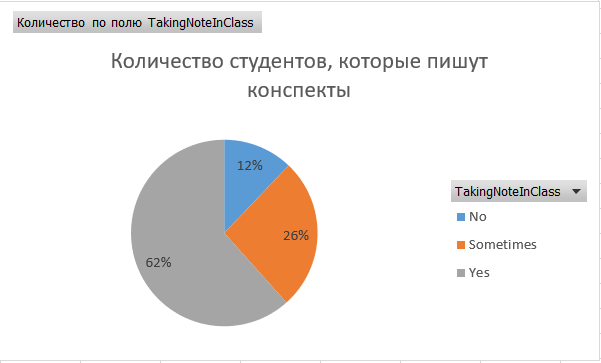
Данная гистограмма отображает количество студентов разных возрастов. По ней мы можем сделать вывод, что больше всего студентов, которым 21 год (30,30%), а меньше всего 20-летних (7,07%).



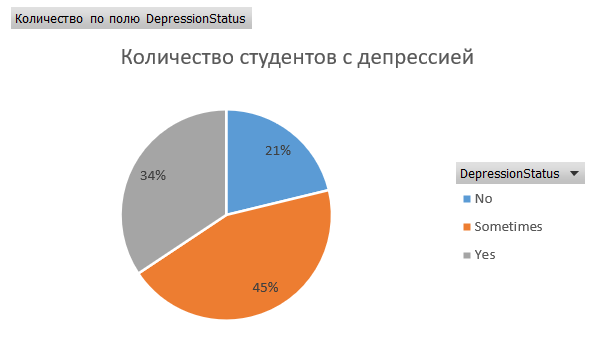
Данная диаграмма показывает процентное соотношение мужчин и женщин среди студентов. Мужчин больше.



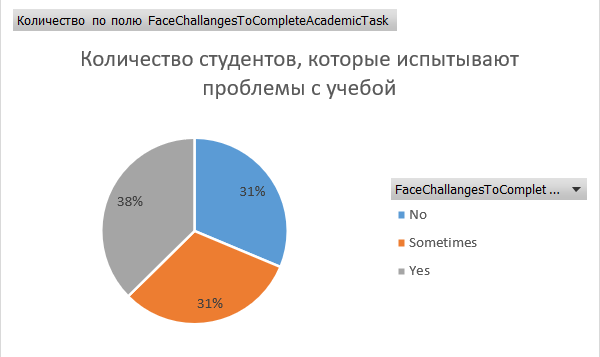
Данная гистограмма показывает успеваемость студентов. Из нее следует, что большинство студентов учатся средне (Average) и хорошо (Good). Лишь некоторые учатся на отлично (Excellent) или, напротив, ниже среднего (Below average).



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



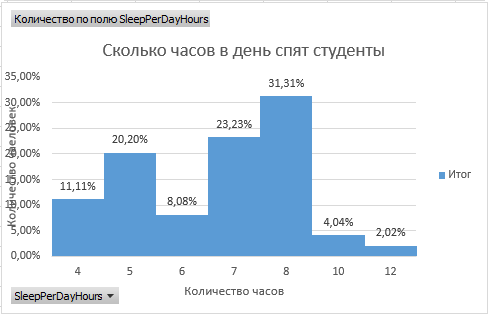
Данная диаграмма показывает процент студентов, которые испытывают депрессию. Почти половина студентов (45%) испытывают ее иногда, еще 34% испытывают ее постоянно. Лишь 21% студентов депрессией не страдает.



Данная диаграмма показывает процент студентов, испытывающих 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%).



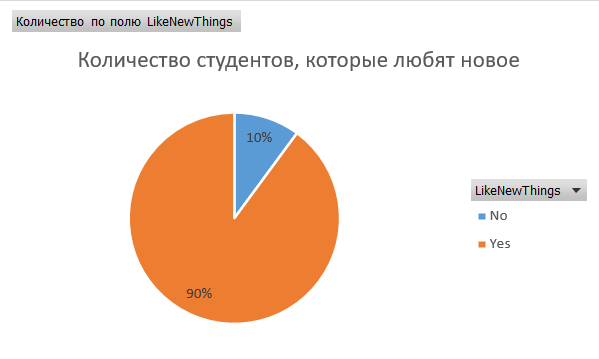
Данная диаграмма показывает процент студентов, которые любят публичные выступления. Абсолютное большинство (70%) публичные выступления любят, еще примерно треть (30%) их не любят.



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.



Данная гистограмма показывает количество друзей у студентов. Почти две трети студентов (65,26%) имеют от 0 до 9 друзей.



Данная диаграмма показывает процент студентов, которые любят новое. Таких студентов абсолютное большинство (90%). Лишь 10% новое не любят.

Вывод

С применением данных, изученных в ходе лабораторной работы, могут решаться следующие задачи:

1. Анализ успеваемости и факторов, влияющих на нее.

Используя данные, можно выявить корреляции между представленными факторами и успехами в учебе.

1. Прогнозирование риска депрессии или проблем с учебой.

На основе данных о студентах, можно создать модель для прогнозирования риска развития депрессии или проблем с учебой.

1. Изучение влияния образа жизни на успехи в учебе.

Анализ данных образа жизни студентов может помочь понять, какие привычки и поведенческие факторы связаны с лучшими учебными результатами.

1. Разработка персонализированных подходов к обучению и поддержке студентов.

На основе данных о студентах можно создать индивидуальные программы обучения и поддержки, учитывающие их индивидуальные потребности, интересы и особенности.

1. Исследование социальных взаимосвязей и их влияния на учебный процесс.

Используя данные, можно изучать, как социальные взаимосвязи влияют на успехи в учебе и общую учебную атмосферу.

1. Оптимизация программ поддержки студентов.

Анализ данных о студентах может помочь университетам и школам оптимизировать программы поддержки студентов, направленные на помощь в решении их проблем и повышении их учебной мотивации.