МИНОБРНАУКИ РОССИИ

ФГБОУ ВО Череповецкий государственный университет

Институт информационных технологий

Кафедра: МПО ЭВМ

Дисциплина: Основы Data Science

ЛАБОРАТОРНАЯ РАБОТА № 5

Тема: Визуализация данных

Выполнила:

студентка гр. 1ПИб-02-1оп-22 Рябкова А.Л.

Проверил: доцент, Юдина Ольга Вадимовна

Череповец, 2024 г.

Цель: Знакомство с этапом понимание 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

Требования:

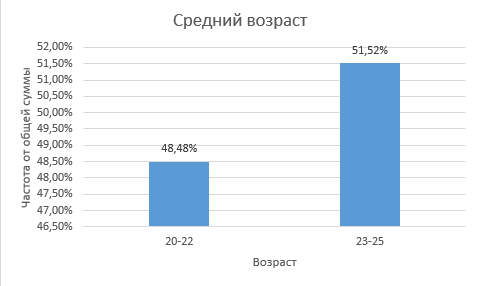
* Диаграммы должны иметь подписи по осям и заголовки, отражающие их суть
* Должны быть разными
* Их тип должен быть пригодным для представления данных
* Они должны отражать ваши данные, их связи, закономерности, которые вам удалось обнаружить.
* Для построения можно использовать любые пригодные программы – Excel, если вам удалось найти доступные программы для визуализации данных, то можно использовать их.

Ход работы:

В начале исследования я ознакомилась с предложенными для анализа данными, после чего был сделан ряд предположений, которые проверялись с помощью различных тестов.

Анализ количественных признаков:

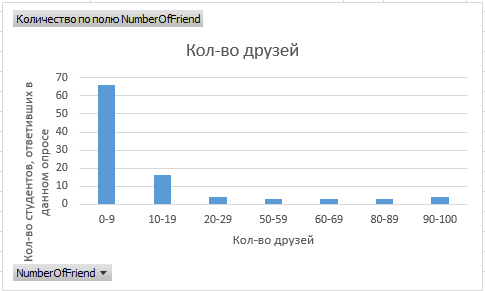
* 1. Можно предположить, что на данный момент в институте количество студентов старшего курса больше, чем количество студентов младшего курса



* 1. В среднем студенты спят по 7-8 часов в день

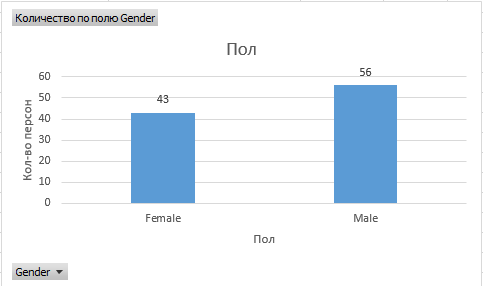


* 1. В среднем студенты имеют до 9 друзей, на 2 место от 10 до 19 друзей



Анализ качественных признаков

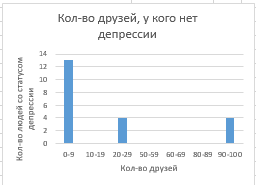
1. В университете учится больше мальчиков, чем девочек



1. Люди с депрессией спят меньше, чем люди у которых депрессии нет



1. Кол-во друзей не зависит от того, есть ли у студента депрессия или нет



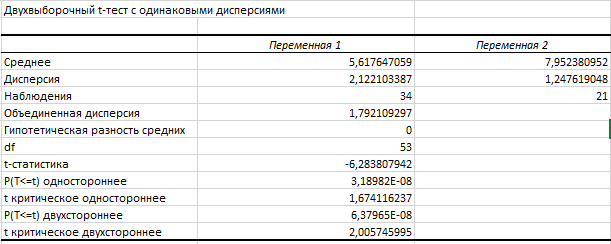
Далее проводились испытания гипотез. Для начала необходимо было сформулировать гипотезы.

Гипотеза №1

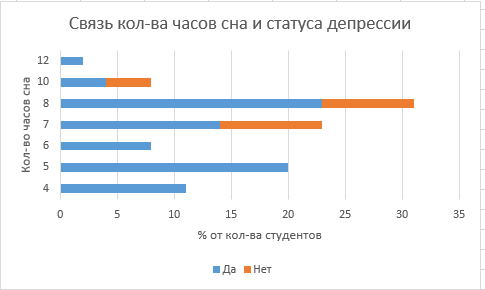
Предположим, что люди с депрессией спят столько же, сколько и люди без депрессии, тогда:

Н0: Те, у кого есть депрессия спят столько же, сколько и люди без депрессии

Н1: Те, у кого есть депрессия спят больше/меньше, чем люди без депрессии



Исходя из исследований (|t-статистика| > t-критического), гипотеза Н0 отвергается. Значит те, у кого есть депрессия спят больше/меньше, чем люди без депрессии



Гипотеза №2

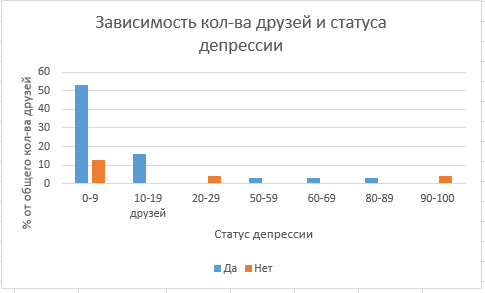
Предположим, что у тех, у кого депрессия друзей больше/меньше, по сравнению с людьми без депрессии

Н0 У тех, у кого депрессия друзей больше/меньше, по сравнению с людьми без депрессии

Н1 У тех у кого депрессия друзей столько же, сколько и у людей без депрессии



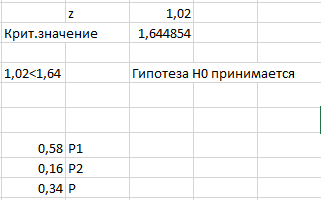
Исходя из исследований (t-статистика < t-критического), гипотеза Н0 принимается. Значит у тех, у кого депрессия, друзей больше/меньше, по сравнению с людьми без депрессии



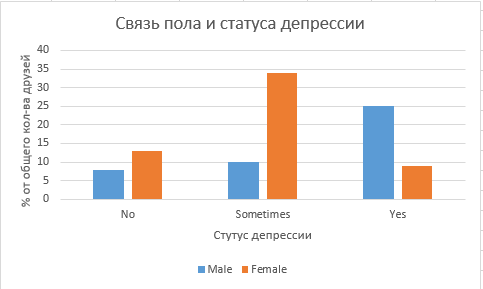
Гипотеза №3

H0: Девочки чаще страдают депрессией

H1: Мальчики чаще страдают депрессией



Исходя из исследований, мы принимаем гипотезу Н0, а это значит, что девочки чаще страдают депрессией

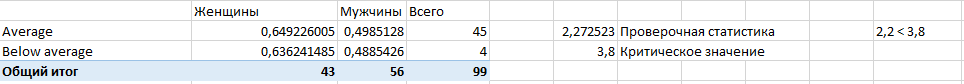


Далее я исследовала влияние различных факторов на успеваемость студентов.

1) Влияние пола на успеваемость студента

H0 Не существует связи между полом и оценками

Н1 Пол влияет на успеваемость студента

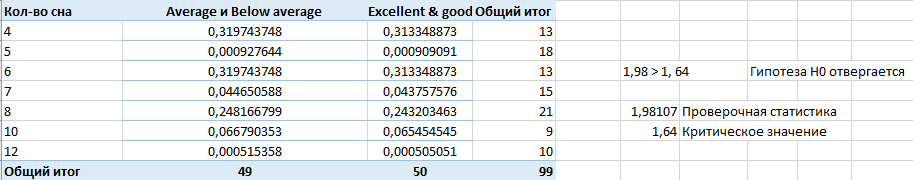


Исходя из исследования, т.к. гипотеза Н0 принимается, можно сказать, что гендер не влияет на обучение в университете

2) Влияние кол-ва сна на успеваемость

H0 Не существует связи между кол-вом сна и успеваемостью

Н1 Связь существует

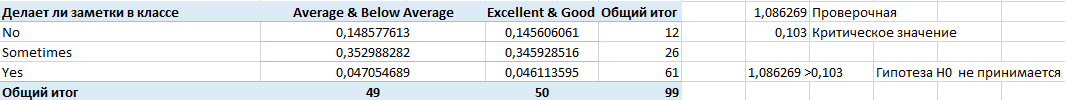


Т.к. проверочная статистика > критического значения, гипотеза Н0 отвергается, значит связь между кол-вом сна и успеваемостью студента существует

3) Влияние заметок в классе на успеваемость

Н0 Успеваемость не зависит от того, делают ли люди заметки в классе

Н1 Связь есть

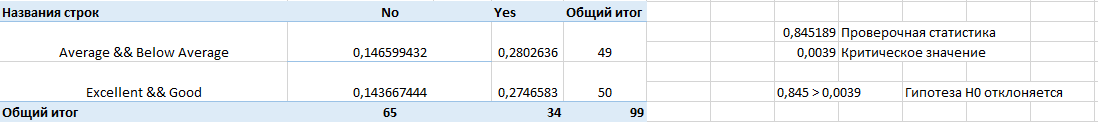


Т.к. проверочная статистика > критического значения, гипотеза Н0 отвергается, значит связь между тем, делают ли студенты заметки в классе, и успеваемостью студента существует

4) Влияние статуса депрессии на успеваемость

Н0 Не существует связь между успеваемостью и статусом депрессии

Н1 Связь существует

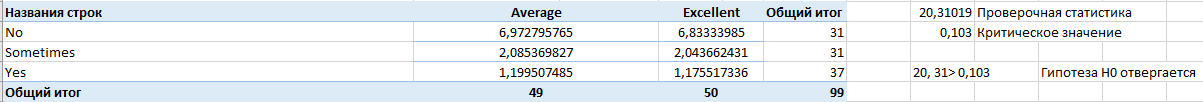


Т.к. проверочная статистика > критического значения, гипотеза Н0 отвергается, значит связь между успеваемостью и статусом депрессии существует

5) Влияние трудолюбия на успеваемость

Н0 Не существует связь между успеваемостью и трудолюбием

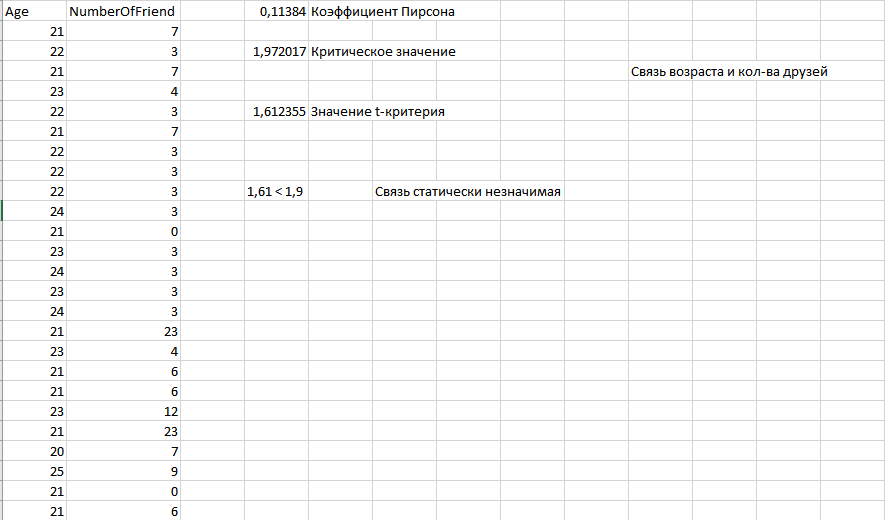
Н1 Существует связь



Т.к. проверочная статистика > критического значения, гипотеза Н0 отвергается, значит связь между трудолюбием студентов и их успеваемостью существует

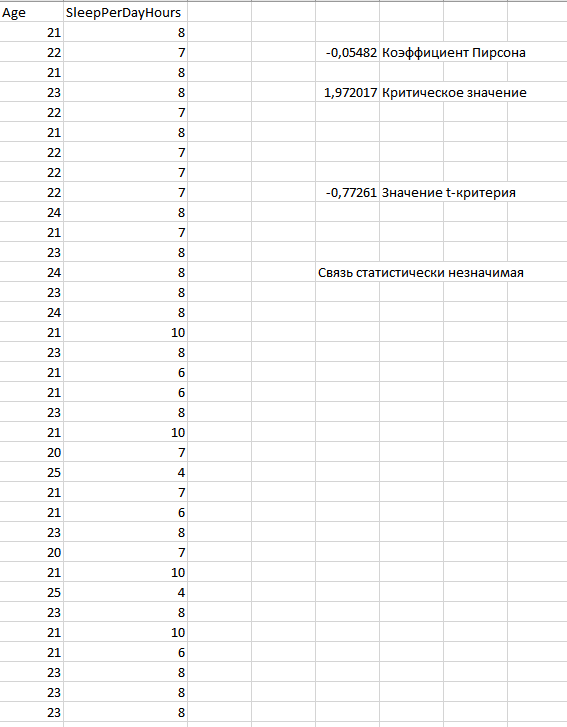
Далее мы ищем зависимость между переменными при помощи коэффициента корреляции Пирсона и Спирмена. Так как у нас мало входных количественных данных, разнообразных проверок зависимостей не получится.

1) Проверка связи возраста и кол-ва друзей (коэффициент Пирсона)



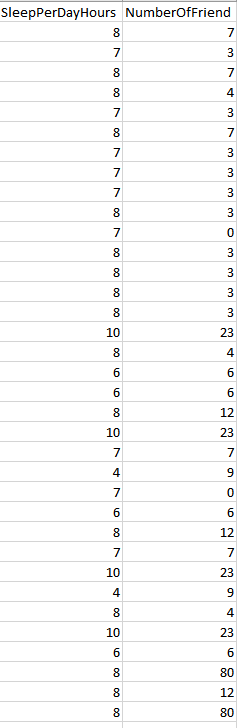
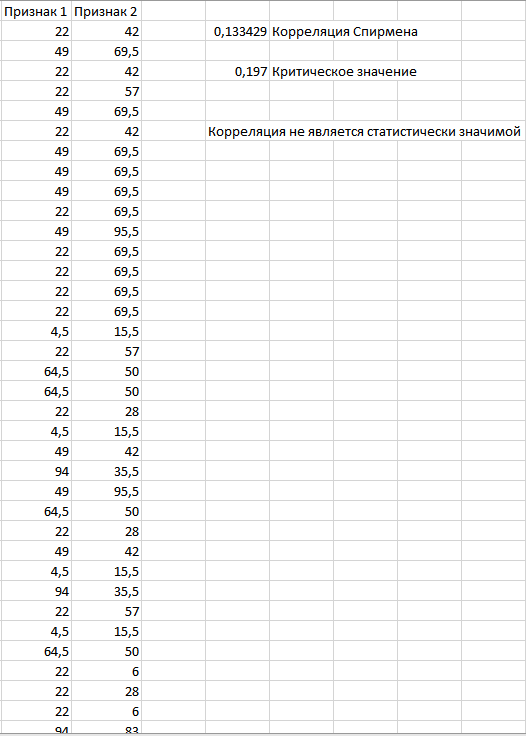
Мы нашли коэффициент Пирсона, значение t-критерия и критическое значение. Сверив его с критическим значением, оказывается, что связь возраста и кол-ва друзей статистически незначимая

2) Связь возраста и кол-ва часов сна (коэффициент Пирсона)



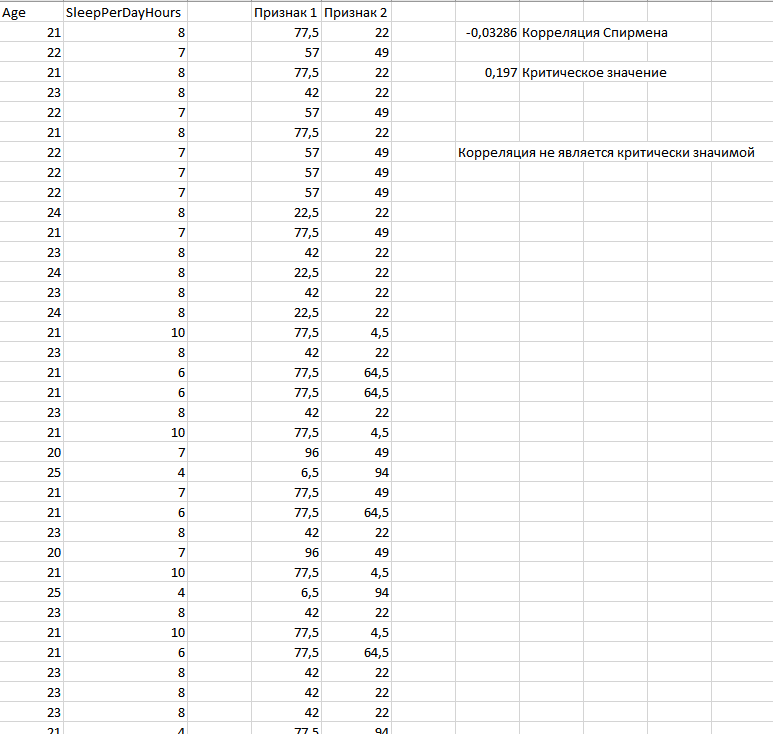
Мы нашли коэффициент Пирсона, значение t-критерия и критическое значение. Сверив его с критическим значением, оказывается, что связь возраста и кол-ва часов сна статистически незначимая

3) Связь кол-ва часов сна и кол-ва друзей (коэффициент Спирмена)

Мы ранжировали наши признаки, нашли корреляцию Спирмена, сравнили его с критическим значением, т.к. значение корреляции меньше критического значения, корреляция не является статистически значимой.

4) Связь возраста и кол-ва часов сна



Мы ранжировали наши признаки, нашли корреляцию Спирмена, сравнили его с критическим значением, т.к. значение корреляции меньше критического значения, корреляция не является статистически значимой.

Выводы:

Исходя из вышеперечисленных исследований, можно сказать, что для более качественных результатов нужно больше входных данных, из-за недостатка информции часто выходят неточные выводы.

Контрольные вопросы:

1. Приведите примеры диаграмм, подходящих для качественных переменных.

Диаграммы, которые подходят для качественных переменных: столбчатая и круговая диаграммы.

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

Примеры диаграмм в пиксельной технике: тепловые карты, точечные диаграммы.

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

Разница между визуализацией информации и визуализацией данных заключается в том, что визуализация информации призвана донести мысль/идею, сформулированные из уже изученных данных, а визуализация данных представляет собой способ изучения данных.

1. Какие диаграммы основаны на геометрической технике?

На геометрической технике основан ряд диаграмм, а именно: круговые, лепестковые, каркасные.

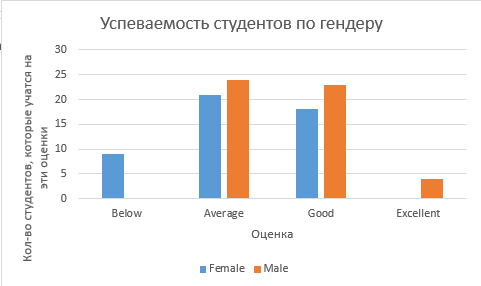
1. Что можно визуализировать с помощью иконок – приведите пример.

Одним из методов визуализации данных при помощи иконок являются «лица Чернова», они позволяют отображать многомерные данные в виде человеческого лица, каждая часть которого может отображать значение определённого параметра. Ярким примером того, что можно визуализировать при помощи иконок является выраженность того или иного типа правонарушения в РФ.

Исследование: Почему большинство студентов учатся на оценки «средне» и «ниже среднего», нежели чем «хорошо» и «отлично»



Проверим, чем зависит ли такая успеваемость среди студентов от других параметров.



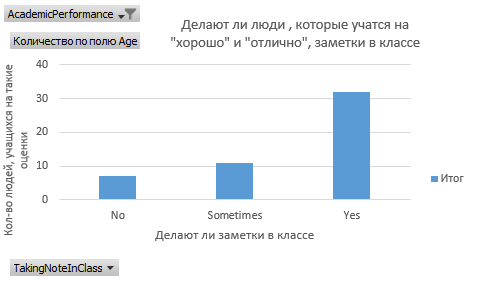
Как показывает график, большинство студентов учатся на средние оценки, причем парни учатся чуть лучше, чем девушки. И только девушки получают оценки «ниже среднего», при этом только парни получают оценки «отлично». Эти данные можно интерпретировать как то, что в данном университете девушки учатся чуть хуже, чем парни.



Исходя из данного графика можно сказать, что у тех студентов, которые учатся на «средне» и «ниже среднего» чаще всего наблюдается депрессия, либо она периодически проявляется. Что можно считать одним из факторов их плохих оценок.

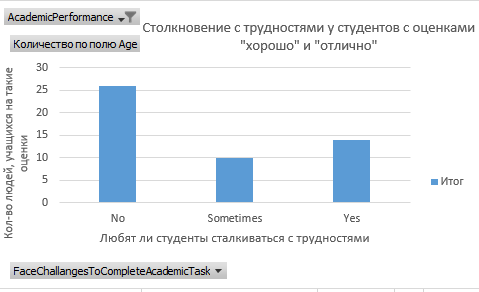


Как показывает таблица, люди с плохой успеваемостью почти всегда делают заметки в классе, для сравнения я решила проверить, делают ли заметки люди с более высокой успеваемостью.



Оказалось, что фактор того, делают ли студенты заметки в классе или не делают их не влияет на их учебу.





Далее я решила проверить, фактор «Столкновение с трудностями при учёбе», и оказалось, что люди с плохими оценками намного чаще сталкиваются с проблемами, чем люди с хорошими оценками. Из этого можно сделать вывод, что, возможно, людям с плохой успеваемостью просто сложнее преодолеть сложности по учебе, и из-за этого их оценки такие.