Исследование зависимостей между характеристиками сотрудника в контексте отсутствия на работе

Файзуллина Динара

СПбГУ, 342 группа

Введение

Совокупные годовые затраты, связанные с потерей производительности, составляют $84 млрд. По данным опроса, ежегодные расходы, связанные с прогулами варьируются в зависимости от отрасли.

- Высококвалифицированные рабочие — 24,2
- Менеджеры/руководители — 15,7
- Работники сферы услуг — 8,5
- Клерки — 8,1
- Продавцы — 6,8
- Школьные учителя — 5,6
- Медсестры — 3,6

План работы

Цель

Выявление зависимостей между характеристиками сотрудников, отсутствующих на рабочем месте, с целью сокращения убытков компаний.

Задачи

Провести анализ данных с помощью следующих методов:

- Описательные статистики
- Сравнение групп (t-test, chi-test)
- Корреляция
- Регрессия
- Кластеризация

Данные

Анализировались данные об отсутствии сотрудников курьерной фирмы в Бразилии на рабочем месте, которые собирались с июля 2007 года по июль 2010 года.

Данные содержат 740 записей и 21 переменную.

Источник данных: UCI Machine Learning Repository, были собраны для исследования в Universidade Nove de Julho, Sao Paulo, Brazil

Описательные статистики

Характеристики метрических переменных исследуемого набора данных представлены в таблице:

Переменная	Мин	Среднее	Медиана	Макс
Transportation.expense	118.0	221.3	225.0	388.0
Distance.from.Residence.to.Work	5.00	29.63	26.00	52.00
Service.time	1.00	12.55	13.00	29.00
Age	27.00	36.45	37.00	58.00
Work.load.Average.day	205.9	271.5	264.2	378.9
Son	0.000	1.019	1.000	4.000
Pet	0.0000	0.7459	0.0000	8.0000
Weight	56.00	79.04	83.00	108.00
Height	163.0	172.1	170.0	196.0
Body.mass.index	19.00	26.68	25.00	38.00
Absenteeism.time.in.hours	0.000	6.924	3.000	120.000

Описательные статистики

Для описания набора данных также использовались следующие переменные:

- Идентификационная переменная – id сотрудника (ID): от 1 до 36
- Переменная категории – причина (Reason.for.absence): от 1 до 28
- Переменная категории – месяц (Month.of.absence): от 1 до 12
- Переменная категории – день недели(Day.of.the.week): от 2 до до 6
- Переменная категории – время года (Seasons): от 1 до 4
- Бинарная переменная – дисциплинарное взыскание(Disciplinary.failure)
- Переменная категории – образование (Education)
- Бинарная переменная – употребление алкоголя (Social.drinker)
- Бинарная переменная – курение (Social.smoker)

Сравнение групп

T-тест

- Не выявлены различия в Absenteeism.time.in.hours по Social.driker
  p = 0.07396, t = -1.7895, интервал [-3.67; 0.17], средние (5.93, 7.68)
- Не выявлены различия в Absenteeism.time.in.hours по Social smoker
  p = 0.756, t = 0.31204, интервал [-2.46; 3.38], средние (6.95, 6.50)
- Выявлены различия в Absenteeism.time.in.hour по Disciplinary.failure
  p < 2.2e-16, t = 14.239, интервал [6.31; 8.32], средние (7.32, 0.00)

Хи-квадрат

Выявлено наличие взаимосвязей между переменными:

- Day.of.the.week и Education ($p = 0.54$, $X = 10.773$, $df = 12$)
- Day.of.the.week и Seasons ($p = 0.19$, $X = 15.91$, $df = 12$)
- Day.of.the.week и Month.of.absence ($p = 0.56$, $X = 45.829$, $df = 48$)
- Day.of.the.week и Reason.for.absence ($p = 0.06$, $X = 132.02$, $df = 108$)

Корреляция

Очень сильная связь

Body.mass.index с Weight ($p = 0.9041169$, $p < 2.2e\text{-}16$)

Достаточно сильная связь

- Service.time с Age
  $r = 0.6709789$, $p < 2.2e\text{-}16$
- Service.time с Body.mass.index
  $r = 0.499718$, $p < 2.2e\text{-}16$

Слабая связь

- Absenteeism.time.in.hours с Distance.from.Residence.to.Work
  $r = -0.08836282$, $p = 0.0162$
- Absenteeism.time.in.hours с Son
  $r = 0.1137565$, $p = 0.001939$

Корреляционная плеяда

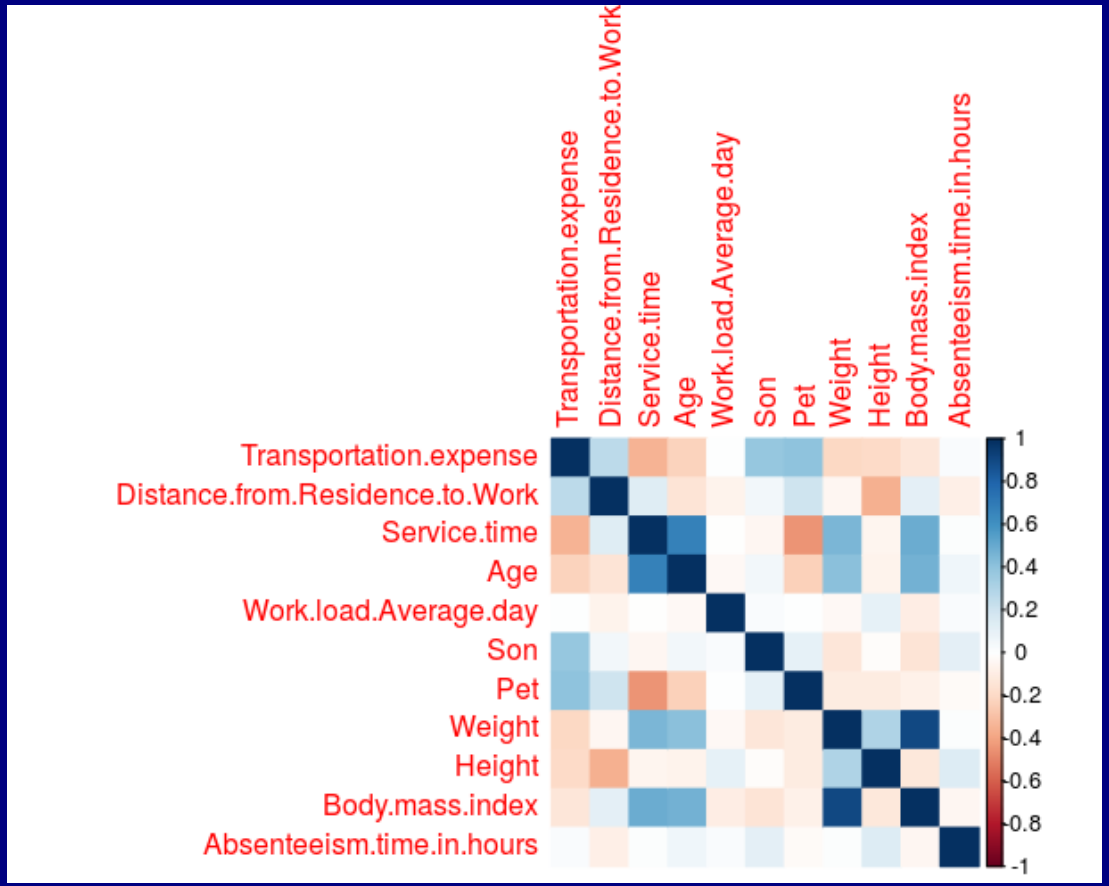


Рис. 1: Корреляционная плеяда для набора данных Absenteeism_at_work

Регрессия

Зависимость между Service.time(y) и переменными Age(x_1), Weight(x_2), Body.mass.index(x_3) и Pet(x_4) выражается уравнением регрессии:

$$y = 0.32599 * x_1 - 0.00634 * x_2 + 0.27239 * x_3 - 1.03315 * x_4 - 5.32296$$

Кластеризация

Для применения метода k-средних определим оптимальное число кластеров:

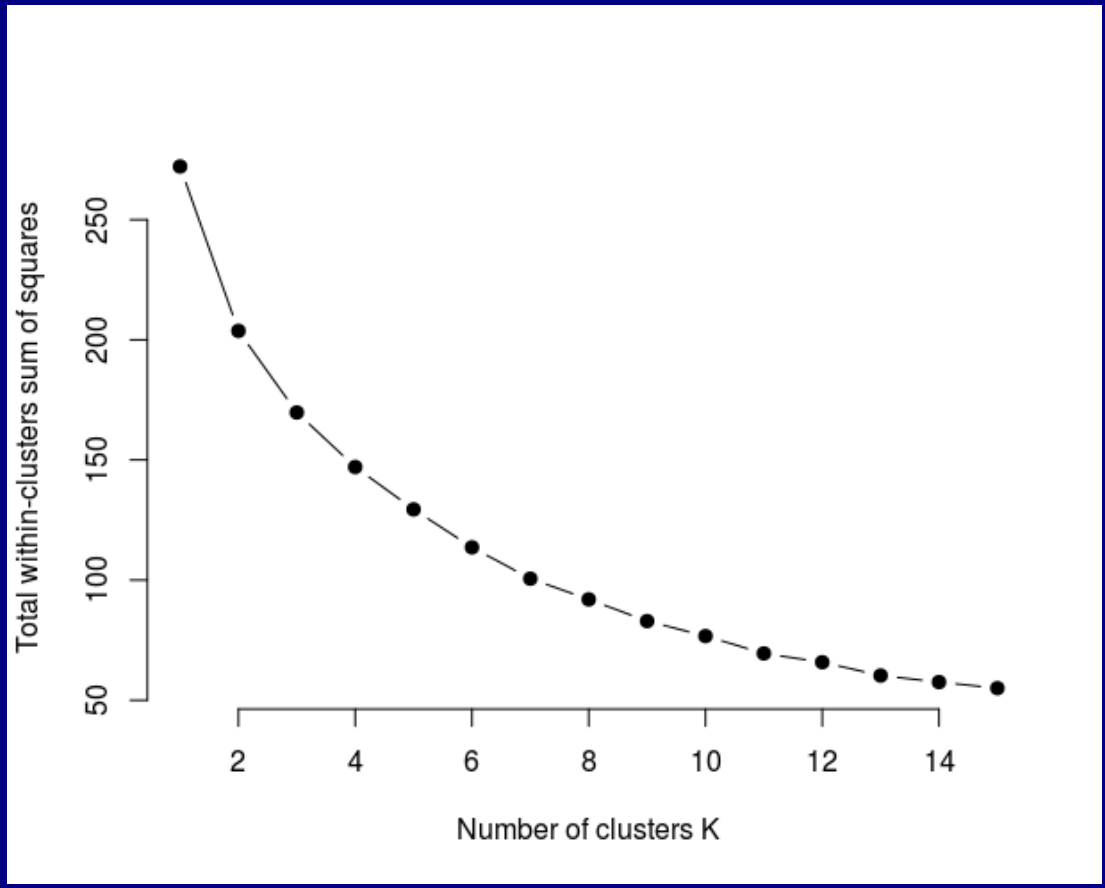
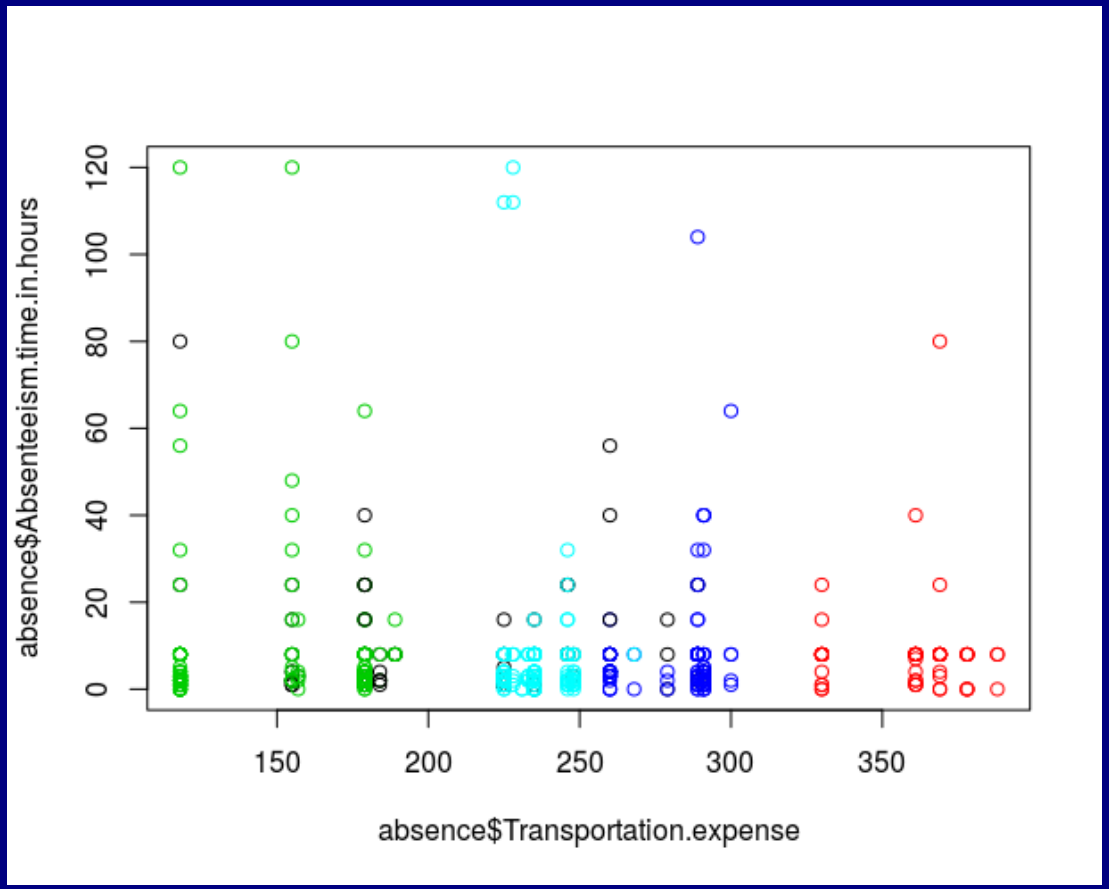


Рис. 2: Диаграмма суммы квадратов внутри кластеров для k.max = 15

Кластеризация

Распределение данных на кластеры на диаграмме рассеивания для характеристик Absenteeism.time.in.hours и Transportation.expense на наборе данных.



Исследование зависимостей между характеристиками сотрудника в контексте отсутствия на работе

Файзуллина Динара

СПбГУ, 342 группа