Отчёт об анализе данных об отсутствии на работе

Динара Руслановна Файзуллина

12.10.2018

Описание данных

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.

Описательные статистики

Характеристики метрических переменных исследуемого набора данных представлены в таблице:

Переменная	Минимум	Среднее	Медиана	Максимум
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

Так, Pet изменяется от 0 до 8 со средним значением равным 0.7459 и медианой равной 0, что говорит о наличии выбросов в данных.

Аналогично, Absenteeism.time.in.hours изменяется от 0 до 120, тогда как среднее значение равно 6.924, а медиана равна 3. Откуда следует, что значение данной

переменной является небольшим для большинства записей, так как даже выброс в виде максимального значения не сильно повлиял на среднее.

Для описания набора данных также использовались следующие переменные:

- Идентификационная переменная – id сотрудника (ID): от 1 до 36
- Переменная категории – причина отсутствия (Reason.for.absence)[1]: от 1 до 28
- Переменная категории – месяц (Month.of.absence): от 1 до 12
- Переменная категории – день недели (Day.of.the.week)[2]: от 2 до до 6
- Переменная категории – время года (Seasons): от 1 до 4
- Бинарная переменная – дисциплинарное взыскание (Disciplinary.failure)
- Переменная категории – образование (Education)[3]
- Бинарная переменная – употребление алкоголя (Social.drinker)
- Бинарная переменная – курение (Social.smoker)

Распределение выборки по показателю Month.of.absence:
1 – 6.7%, 2 – 9.7%, 3 – 11.7%, 4 – 7.16%, 5 – 8.6%, 6 – 7.3%, 7 – 9%, 8 – 7.3%, 9 – 7.2%, 10 – 9.6%, 11 – 8.5%, 12 – 6.6%

Распределение выборки по показателю Day.of.the.week:
2 – 21.7%, 3 – 20.8%, 4 – 21.1%, 5 – 16.9%, 6 – 19.5%

Распределение выборки по показателю Seasons:
1 – 22.9%, 2 –25.9%, 3 – 24.7%, 4 – 26.3%

Распределение выборки по показателю Education:
1 – 82.5%, 2 – 6.2%, 3 – 10.6%, 4 – 0.5%

Распределение выборки по показателю Reason.for.absence:
1 – 2.1%, 2 – 0.1%, 3 – 0.1%, 4 – 0.2%, 5 – 0.4%, 6 – 1.0%, 7 – 2.0%, 8 – 0.8%, 9 – 0.5%, 10 – 3.3%, 11 – 3.5%, 12 – 1.0%, 13 – 7.4%, 14 – 2.5%, 15 – 0,2%, 16 – 0.4%, 17 – 0.1%, 18 – 2.8%, 19 – 5.4%, 20 – 0%, 21 – 0.8%, 22 – 5.1%, 23 – 20.1%, 24 – 0.4%, 25 – 4.1%, 26 – 4.4%, 27 – 9.3%, 28 – 15.1%

[1]Причины отсутствия 1-21 были засвидетельствованы Международной классификацией болезней (МСБ), причины 22-28 без МСБ

[2]Понедельник (2), Вторник (3), Среда (4), Четверг (5), Пятница (6)

[3]High school (1), graduate (2), postgraduate (3), master and doctor (4)

T-тест

Рассмотрим различия в уровне Absenteeism.time.in.hours по показателю Social.driker. Согласно критерию Стьюдента (t-test) не выявлены статистически значимые различия между группой 0 и группой 1 (p = 0.07396, t = -1.7895, 95 confidence interval от -3.6693353 до 0.1699306, средние значения 5.931250 и 7.680952 соответственно).

Теперь рассмотрим различия в уровне Absenteeism.time.in.hours по показателю Social smoker. Для них также не были выявлены статистически значимые различия между группой 0 и группой 1 (p = 0.756, t = 0.31204, 95 percent confidence interval от -2.468636 до 3.384088, средние значения 6.957726 и 6.500000 соответственно).

Однако по показателю Disciplinary.failure нам удалось выявить статистически значимые различия между группой 0 и группой 1 в уровне Absenteeism.time.in.hours (p < 2.2e-16, t = 14.239, средние значения 7.32 и 0.00 соответственно).

Хи-квадрат

Для выявления взаимосвязей между номинальными переменными используется критерий хи-квадрат.

Так, в рассматриваемых данных показано наличие статистически значимых взаимосвязей между следующими номинальными переменными:

- Day.of.the.week и Education (p = 0.5485, X = 10.773, df = 12)
- Day.of.the.week и Seasons (p = 0.1954, X = 15.91, df = 12)
- Day.of.the.week и Month.of.absence (p = 0.5622, X = 45.829, df = 48)
- Day.of.the.week и Reason.for.absence (p = 0.05806, X = 132.02, df = 108)

Однако для между номинальными переменными, представленными ниже, где казалось бы должна быть взаимосвязь, было показано отсутствие статистически значимой взаимосвязи:

- Reason.for.absence и Month.of.absence (p < 2.2e-16, X = 599.52, df = 324)
- Reason.for.absence и Seasons (p < 2.2e-16, X = 267.86, df = 81)

Корреляционный анализ

Корреляционный анализ позволяет определить взаимосвязь между метрическими переменными. Значения коэффициентов корреляции представлены в таблице, статистически значимые взаимосвязи выделены полужирным шрифтом.[4]

	Transport	Distance	Service	Age	Work	Son	Pet	Weight	Height	Index	Absent
Transport	1.00	**0.26**	**-0.35**	**-0.23**	0.01	**0.38**	**0.40**	**-0.21**	**-0.19**	**-0.14**	0.03
Distance	**0.26**	1.00	**0.13**	**-0.15**	-0.07	0.05	**0.21**	-0.05	**-0.35**	**0.11**	**-0.09**
Service	**-0.35**	**0.13**	1.00	**0.67**	-0.00	-0.05	**-0.44**	**0.46**	-0.05	**0.50**	0.02
Age	**-0.23**	**-0.15**	0.67	1.00	-0.04	0.06	**-0.23**	**0.42**	-0.06	**0.47**	0.07
Work	0.01	-0.07	-0.00	-0.04	1.00	0.03	0.01	-0.04	**0.10**	**-0.09**	0.02
Son	**0.38**	0.05	-0.05	0.06	0.03	1.00	**0.11**	**-0.14**	-0.01	**-0.14**	**0.11**
Pet	**0.40**	**0.21**	**-0.44**	**-0.23**	0.01	**0.11**	1.00	**-0.10**	**-0.10**	**-0.08**	-0.03
Weight	**-0.21**	-0.05	**0.46**	**0.42**	-0.04	**-0.14**	**-0.10**	1.00	**0.31**	**0.90**	0.02
Height	**-0.19**	**-0.35**	-0.05	-0.06	**0.10**	-0.01	**-0.10**	**0.31**	1.00	**-0.12**	**0.14**
Index	**-0.14**	**0.11**	**0.50**	**0.47**	**-0.09**	**-0.14**	**-0.08**	**0.90**	**-0.12**	1.00	-0.05
Absent	0.03	**-0.09**	0.02	0.07	0.02	**0.11**	-0.03	0.02	**0.14**	-0.05	1.00

В частности, Service.time прямо связана с Age, причем связь достаточно сильная ($r = 0.6709789$, $p < 2.2e\text{-}16$). Аналогично, Service.time связано достаточно сильно с Body.mass.index ($r = 0.499718$, $p < 2.2e\text{-}16$).

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

Тогда как связь между Absenteeism.time.in.hours с Transportation.expense, Distance.from.Residence.to.Work, Son оказалась достаточно слабой ($r = 0.02758463$, $p = 0.4537$; $r = -0.08836282$, $p = 0.0162$; $r = 0.1137565$, $p = 0.001939$).

Корреляционная плеяда

Результаты корреляционного анализа можно визуализировать в виде корреляционной плеяды, представленной на Рис. 1.

Регрессионный анализ

Регрессионный анализ позволяет определить зависимость между Service.time(y) и такими переменными, как Age(x_1), Weight(x_2), Body.mass.index(x_3) и Pet(x_4). Значения коэффициентов регрессионного уравнения и уровни значимости представлены в таблице ниже.

[4] Имена переменных были сокращены с сохранением семантики, а размер шрифта уменьшен с целью лучшей визуализации статистических данных

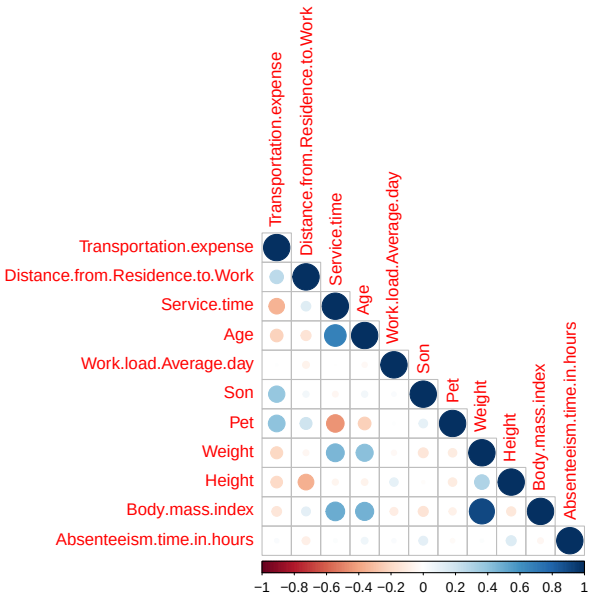

Рис. 1: Корреляционная плеяда для набора данных Absenteeism_at_work

	Estimate	Std.Error	t-value	p-value	Signif. level
(Intercept)	-5.32296	0.76735	-6.937	8.80e-12	0
Age	0.32599	0.01872	17.414	$<$ 2e-16	0
Weight	-0.00634	0.01902	-0.333	0.739	1
Body.mass.index	0.27239	0.05888	4.626	4.41e-06	0
Pet	-1.03315	0.08167	-12.650	$<$ 2e-16	0

Уравнение регрессии выглядит следующим образом:

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

Отметим, что данная модель объясняет 58% изменчивости данных (Multiple R-squared: 0.5845, Adjusted R-squared: 0.5822)

Кластеризация

Процедура кластерного анализа позволяет упорядочить объекты выборки в сравнительно однородные группы на основе информации о наборе данных.

Применим один из наиболее популярных методов кластеризации — метод k-средних.

Для определения оптимального количества кластеров использовался elbow method. Из Рис. 2 видно, что число кластеров $k = 5$ можно принять за оптимальное, так

как разница между суммами квадратов для большего числа кластеров изменяется не сильно.

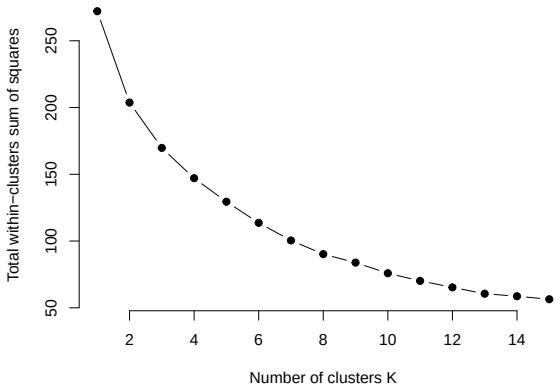

Рис. 2: Диаграмма суммы квадратов внутри кластеров для k.max = 15

После разделения набора данных на 5 кластеров было проведено построение clusplot, изображенное на Рис. 3, где каждый из кластеров обозначен уникальным цветом. Визуализация представлена по двум компонентам, по которым данные наиболее сильно изменяются.

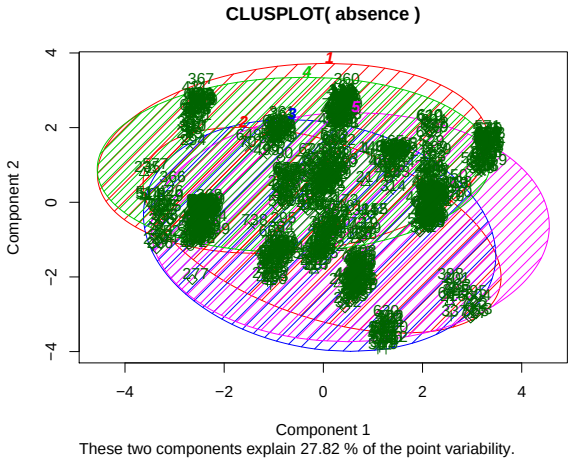

Рис. 3: Clusplot для двух компонент

Так, например, на Рис. 4 изображено распределение данных на кластеры при сравнении значений переменных Absenteeism.time.in.hours и Transportation.expense на наборе данных.

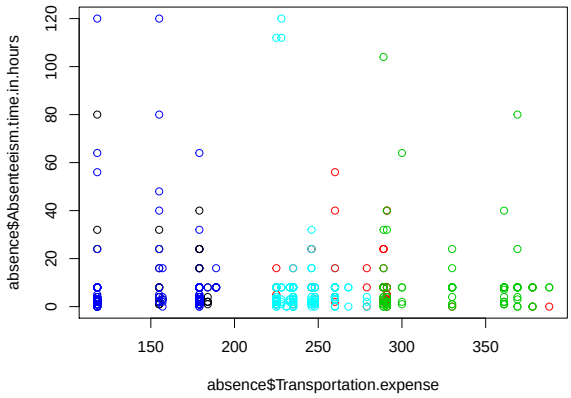

Рис. 4: Clusplot для двух компонент

Вывод

Для анализа данных об отстуствии сотрудников фирмы на рабочем месте были собраны описательные статистики, которые дали общее представление о распределении значений переменных по выборке.

После проведения Т-теста мы увидели, что число часов отстуствия не сильно отличалось для групп курящих/не курящих и пьщих/не пьющих сотрудников, однако оно достаточно сильно варьировалось для сотрудников с и без дисциплираных взысканий.

Критерий Хи-квадрат помог нам выявить взаимосвязь между днём недели отсутствия сотрудника и его образованием, временем года, месяцем, причиной отстуствия.

При корреляционном анализе и построении корреляционной плеяды стало очевидно, существует достаточно сильная связь между временем работы в компании и возрастом, весом, индексом массы тела, количеством домашних питомцев.

Регрессионный анализ позволил определить предсказанную зависимость между описанными выше характеристиками и построить уравение регрессии.

Результатом процедуры кластеризации оказались пять кластеров, на которые был разделен набор данных.