Задание – Дипломный проект по курсу «Python для анализа данных»

с комментариями Зельберг Ирины

Дан файл HR.csv с данными по опросу уровня удовлетворенности сотрудниками работой.

Файл доступен тут - https://drive.google.com/file/d/1INgo03nal-vwFJe7Lec5vOUtOwfJdUr1/view?usp=sharing

Признаки:

1. satisfaction_level - Уровень удовлетворенности работой
2. Last_evaluation - Время с момента последней оценки в годах
3. number_projects - Количество проектов, выполненных за время работы
4. average_monthly_hours - Среднее количество часов на рабочем месте в месяц
5. time_spend_company - Стаж работы в компании в годах
6. work_accident - Происходили ли несчастные случаи на рабочем месте с сотрудником
7. left - уволился ли сотрудник
8. promotion_last_5years - повышался ли сотрудник за последние пять лет
9. department - отдел в котором работает сотрудник
10. salary - относительный уровень зарплаты

Требуется выполнить следующее задание:

# Задание Баллы	Комментарии (И. Зельберг)
1 Загрузите файл HR.csv в pandas dataframe 5	Выполнено
2 Рассчитайте основные статистики для переменных (среднее, медиана, мода, мин/макс,сред.отклонение). 10	Выведены статистики способами: ▪ Методом describe из pandas ▪ Рассчитан каждый показатель отдельно методами numpy или pandas, результат статистики по столбам перевернут в сводную таблицу, затем результаты всех статистик добавлены в один датафрейм ▪ Рассчитан каждый показатель и транспонирован в горизонтальную таблицу, затем добавлены построчно в один датафрейм
3 Рассчитайте и визуализировать корреляционную матрицу для количественных переменных. Определите две самые скоррелированные и две наименее скоррелированные переменные. 10	Выполнено. Величины в представленных данных плохо коррелированы. Коэф-ты корреляции не выше 0,3. Данные были разделены на выборки с поиском более высоких корреляций внутри отдельных выборок (более узких и обобщенных логическим параметром): ▪ работники, с которыми произошел несчастный случай; ▪ уволившиеся сотрудники;

	▪ продвинувшиеся по службе за последние пять лет; ▪ с высокими зарплатами ▪ со средними зарплатами ▪ с низкими зарплатами Внутри каждой группы существенных корреляций между параметрами не выявлено. Вероятно, это отражает, что на данные по персоналу влияет много субъективных факторов, поэтому прямых зависимостей между переменными нет, на значение показателей и поведение людей влияет много и разных причин в любом сочетании.
4 Рассчитайте сколько сотрудников работает в каждом департаменте. 5	Выполнено: ▪ Методом value_count ▪ Группировкой в датафрейме
5 Показать распределение сотрудников по зарплатам. 5	Выполнено
6 Показать распределение сотрудников по зарплатам в каждом департаменте по отдельности 5	Выполнено
7 Проверить гипотезу, что сотрудники с высоким окладом проводят на работе больше времени, чем сотрудники с низким окладом 10	Применена функция stats.ttest_1samp из scipy. Результаты показаны, выводы сделать затруднительно. Способы проверки гипотез на курсе не раскрыты.
8 Рассчитать следующие показатели среди уволившихся и не уволившихся сотрудников (по отдельности): 10 ● Доля сотрудников с повышением за последние 5 лет ● Средняя степень удовлетворенности ● Среднее количество проектов	Рассчитано, отдельно по всем показателям
9 Разделить данные на тестовую и обучающую выборки Построить модель LDA, предсказывающую уволился ли сотрудник на основе имеющихся факторов (кроме department и salary) Оценить качество модели на тестовой выборки 20	Выполнено. Точность прогнозной модели – 75% Разделение на два класса (уволился или нет) не очень хорошо делается для этих данных, нечеткость и отсутствие прямых связей между показателями.
10 Загрузить jupyter notebook с решение на github и прислать ссылку 5	

Итого - максимум 85 баллов Для зачета необходимо набрать минимум 55