Московский Авиационный Институт (Национальный Исследовательский Университет)

Факультет информационных технологий и прикладной математики

Кафедра вычислительной математики и программирования

Лабораторная работа №0 по курсу “Машинное обучение”

Студент: Потапов Д.И.

Преподаватель: Ахмед Самир Халид

Группа: М80-406Б

Дата:

Оценка:

Подпись:

Лабораторная работа №0

**Тема:** Получение и обработка данных

**Задача:** Требуется сформировать/получить два набора данных соответствующие следующим критериям:

1) Один из датасетов должен представлять собой корпус документов. Язык, источник и тематика произвольна

2) Второй датасет должен содержать категориальные, количественные признаки. Для данного датасета определить предсказываемые признаки (для задачи регрессии и классификации). Если такого признака нет, спроектировать

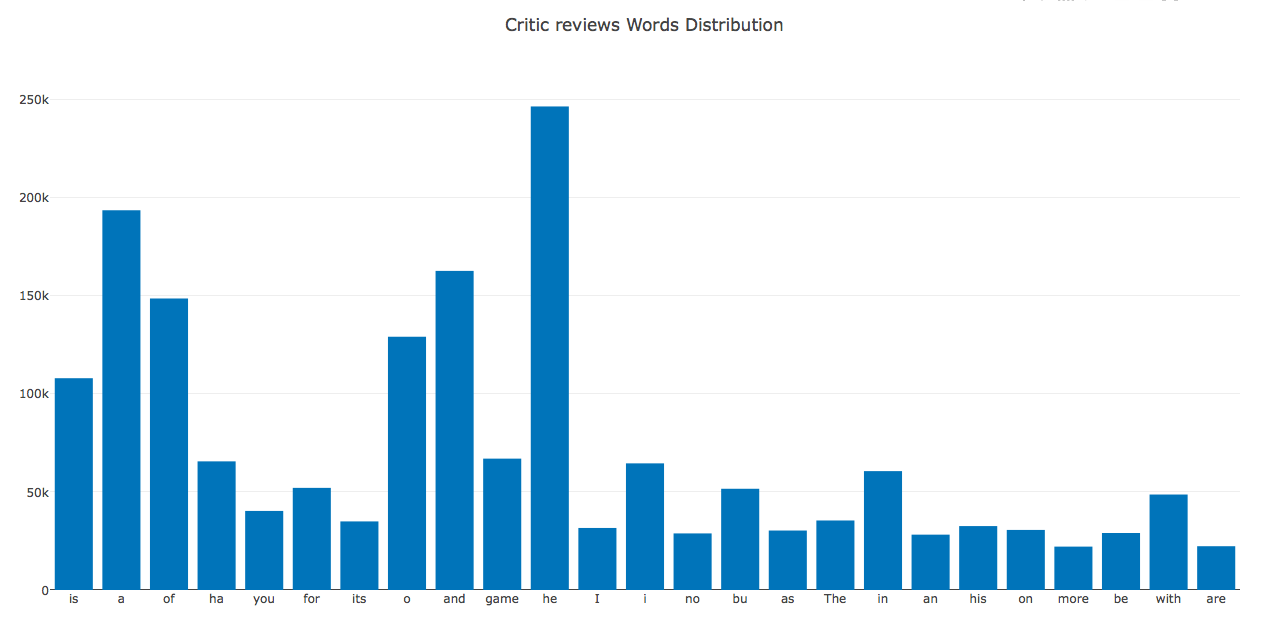
По каждому датасету построить распределения признаков (в случае корпуса документов – построить распределение слов) и объяснить имеющуюся картину. Вычислить статистические характеристики признаков. Обнаружить и решить возможные проблемы с данными. Если решить данную проблему невозможно, объяснить почему.

**Датасеты:**

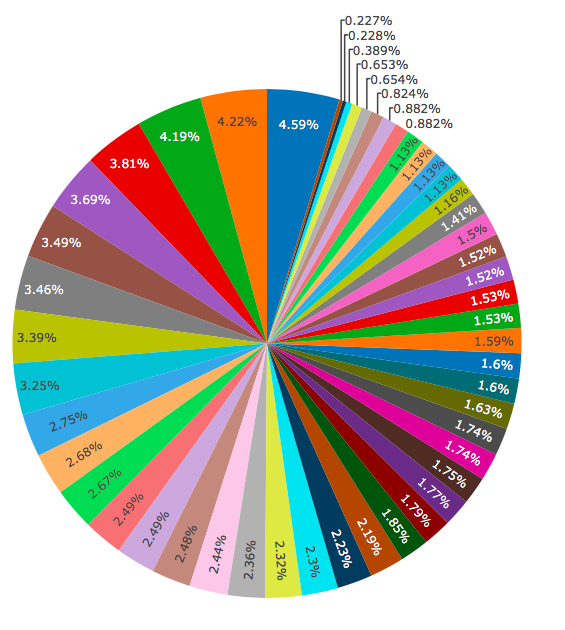
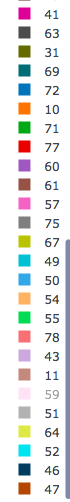
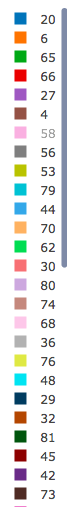
1. metacritic\_critic\_reviews.csv
2. pmsm\_temperature\_data.csv

Описание выполнения работы

1. Построение распределения слов в текстах из датасета metacritic\_critic\_reviews.csv для анализа частоты встречаемости слов. Для этого мы создали словарь, где ключом является слово, а значением — частота встречаемости. После чего построили график, в который включили слова, частота которых более 2000 раз. Для обработки использовалась библиотека pandas.

Программный код:

Результат:

1. Построение распределений для датасета pmsm\_temperature\_data.csv. Рассмотрим следующие признаки: математическое ожидание, дисперсия, мода, максимальное и минимальное значения, среднее значение, размах вариации.  
     
   **Распределение по profile\_id:**

В результате подсчетов статистических величин получаем следующий словарь:

{4: {'ambient': {'Dispersion': 0.001705412440534961,

'Expectation': -0.0016754929240343475,

'Max': 0.6420018,

'Median': -0.26035035,

'Min': -1.6584799,

'Mode': -0.64379674},

'coolant': {'Dispersion': 0.0,

'Expectation': -1.0,

'Max': -1.0127968,

'Median': -1.0724049999999998,

'Min': -1.1565747,

'Mode': -1.0586003000000002},

'i\_d': {'Dispersion': 0.6749543727373366,

'Expectation': 0.415073452413009,

'Max': 1.0328449,

'Median': 1.029112,

'Min': -1.8449407,

'Mode': 1.0291337},

'i\_q': {'Dispersion': 0.2962331328725728,

'Expectation': 0.20291416090715975,

'Max': 1.8940113999999997,

'Median': -0.24571137,

'Min': -0.257432,

'Mode': 1.1367687},

'motor\_speed': {'Dispersion': 0.86976034467283,

'Expectation': 0.4773958052837866,

'Max': 2.0241637,

'Median': -0.14024676,

'Min': -1.2224338,

'Mode': 2.024118},

'pm': {'Dispersion': 1.340932890524489,

'Expectation': -0.4887951410705203,

'Max': 2.2916512000000004,

'Median': -1.0865631000000002,

'Min': -2.524219,

'Mode': -1.032267},

'profile\_id': {'Dispersion': 0.0,

'Expectation': 4.0,

'Max': 4,

'Median': 4,

'Min': 4,

'Mode': 4},

'stator\_tooth': {'Dispersion': 1.2538670975077042,

'Expectation': -0.5769380366813273,

'Max': 1.0784453999999999,

'Median': -1.2120568,

'Min': -2.0661428,

'Mode': -1.9857601999999999},

'stator\_winding': {'Dispersion': 1.3628040570864375,

'Expectation': -0.49139813900607365,

'Max': 1.5885264,

'Median': -1.2172097,

'Min': -2.0199726000000005,

'Mode': 1.261606},

'stator\_yoke': {'Dispersion': 1.2916554468479788,

'Expectation': -0.6703467671962421,

'Max': 0.19008735,

'Median': -1.2336924999999999,

'Min': -1.8346876,

'Mode': -1.7462936999999998},

'torque': {'Dispersion': 0.2960835352900697,

'Expectation': 0.20279448284115728,

'Max': 1.9670967000000001,

'Median': -0.25617734,

'Min': -0.4046426,

'Mode': -0.25929403},

'u\_d': {'Dispersion': 0.31008586901235674,

'Expectation': -0.23076923076923078,

'Max': 0.33655748,

'Median': 0.29020193,

'Min': -1.574856,

'Mode': 0.30065578},

'u\_q': {'Dispersion': 0.2007599557191156,

'Expectation': 0.09939263381503755,

'Max': 1.7628089,

'Median': 0.6537933000000001,

'Min': -1.3066176999999999,

'Mode': 0.64065874}},

6: {'ambient': {'Dispersion': 0.2111075346027187,

'Expectation': 0.033302795453982716,

'Max': 2.7125096,

'Median': 0.42648387,

'Min': -3.6988986000000006,

'Mode': 0.6886087},

'coolant': {'Dispersion': 0.833436501844653,

'Expectation': -0.6076707851536386,

'Max': 1.1101636000000001,

'Median': -1.061172,

'Min': -1.2886666000000002,

'Mode': 0.38199493},

'i\_d': {'Dispersion': 0.9393121549013297,

'Expectation': -0.3793795033055191,

'Max': 1.0297006,

'Median': -0.99861956,

'Min': -3.2458737,

'Mode': 1.029144},

'i\_q': {'Dispersion': 0.4534132270284993,

'Expectation': -0.15720405080842845,

'Max': 2.005306,

'Median': 0.30529833,

'Min': -2.3980353,

'Mode': 0.48692548},

'motor\_speed': {'Dispersion': 1.2565429469878922,

'Expectation': 0.42392353975289077,

'Max': 1.4830375,

'Median': 1.4830242,

'Min': -1.2225861999999998,

'Mode': 1.4830275},

'pm': {'Dispersion': 0.5940525416594449,

'Expectation': 0.2308911283333746,

'Max': 1.9211223000000002,

'Median': 0.33030948,

'Min': -2.4344864,

'Mode': -0.023947535},

'profile\_id': {'Dispersion': 0.0,

'Expectation': 6.0,

'Max': 6,

'Median': 6,

'Min': 6,

'Mode': 6},

'stator\_tooth': {'Dispersion': 0.6488721618342536,

'Expectation': -0.08398742169509991,

'Max': 1.4419761,

'Median': -0.1958906,

'Min': -2.0595632,

'Mode': -1.9857601999999999},

'stator\_winding': {'Dispersion': 0.6166835863025231,

'Expectation': -0.04238987793101741,

'Max': 1.8799341,

'Median': 0.017978428,

'Min': -2.0029435,

'Mode': 1.261606},

'stator\_yoke': {'Dispersion': 0.5663703667021567,

'Expectation': -0.10726223784881274,

'Max': 1.3310958000000002,

'Median': -0.6144783,

'Min': -1.8328238000000001,

'Mode': -0.73472506},

'torque': {'Dispersion': 0.48537895857577934,

'Expectation': -0.13588530962933618,

'Max': 2.262785,

'Median': 0.34313664,

'Min': -2.8086636,

'Mode': 0.35150728},

'u\_d': {'Dispersion': 1.2456235917498204,

'Expectation': -0.2954416024958526,

'Max': 2.2747342999999995,

'Median': -1.1250342,

'Min': -1.6553733,

'Mode': 2.005292},

'u\_q': {'Dispersion': 0.26723945824151335,

'Expectation': -0.1455171218461386,

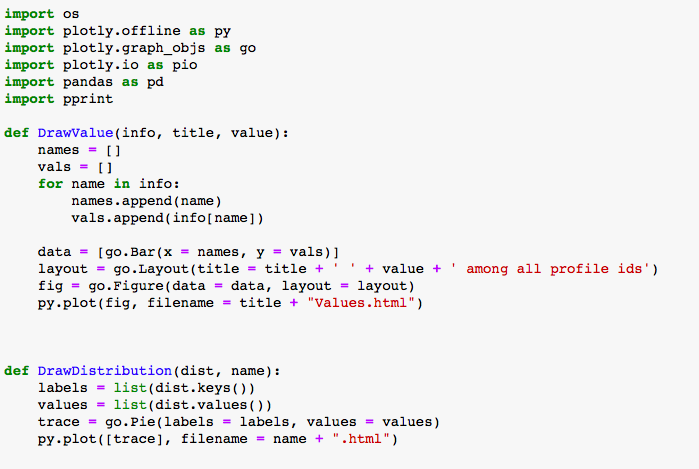
'Max': 1.7653299999999998,

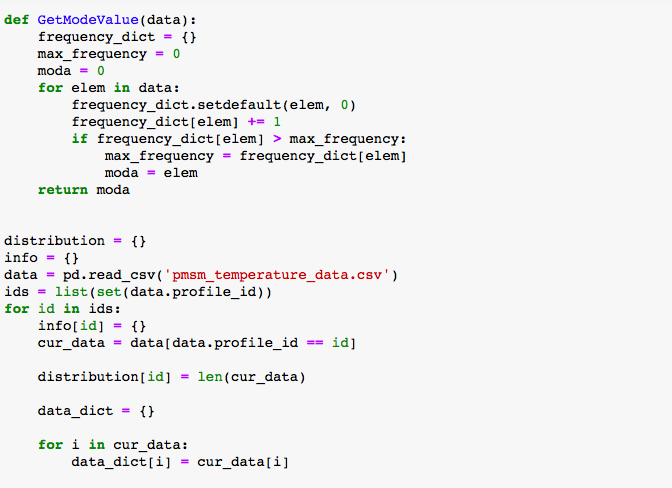
'Median': 0.7090653,

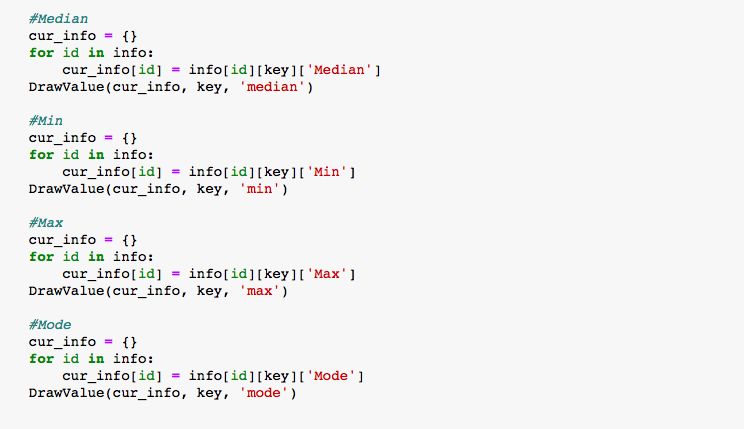
'Min': -1.8614631999999998,

'Mode': -1.3032839}},…

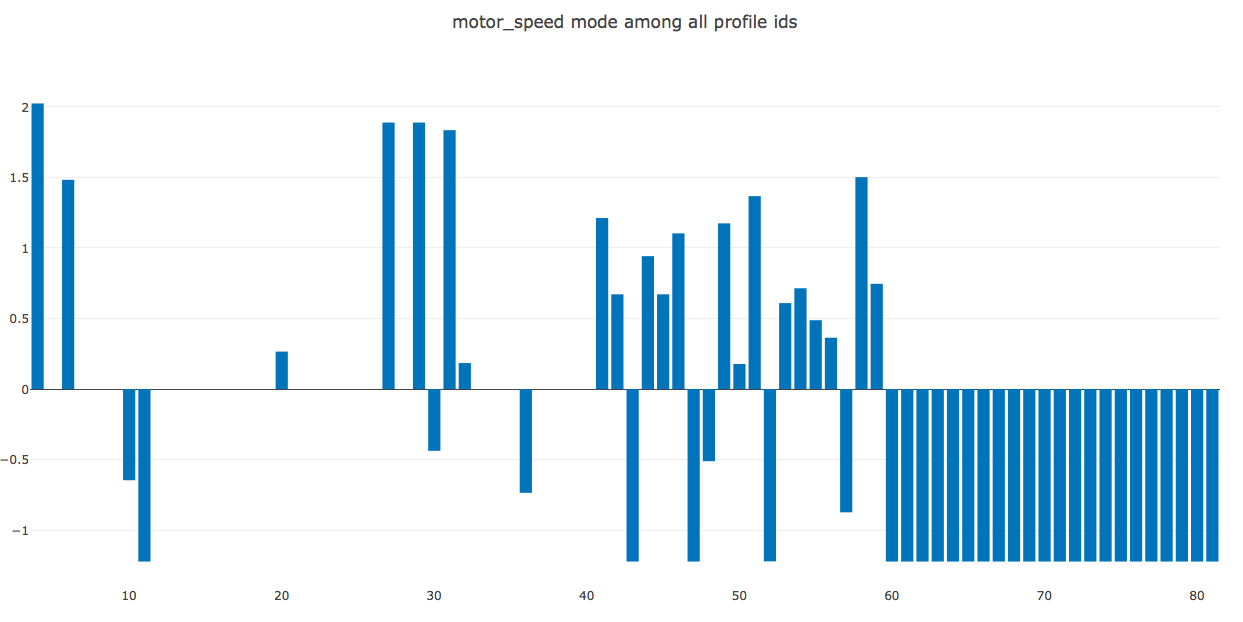
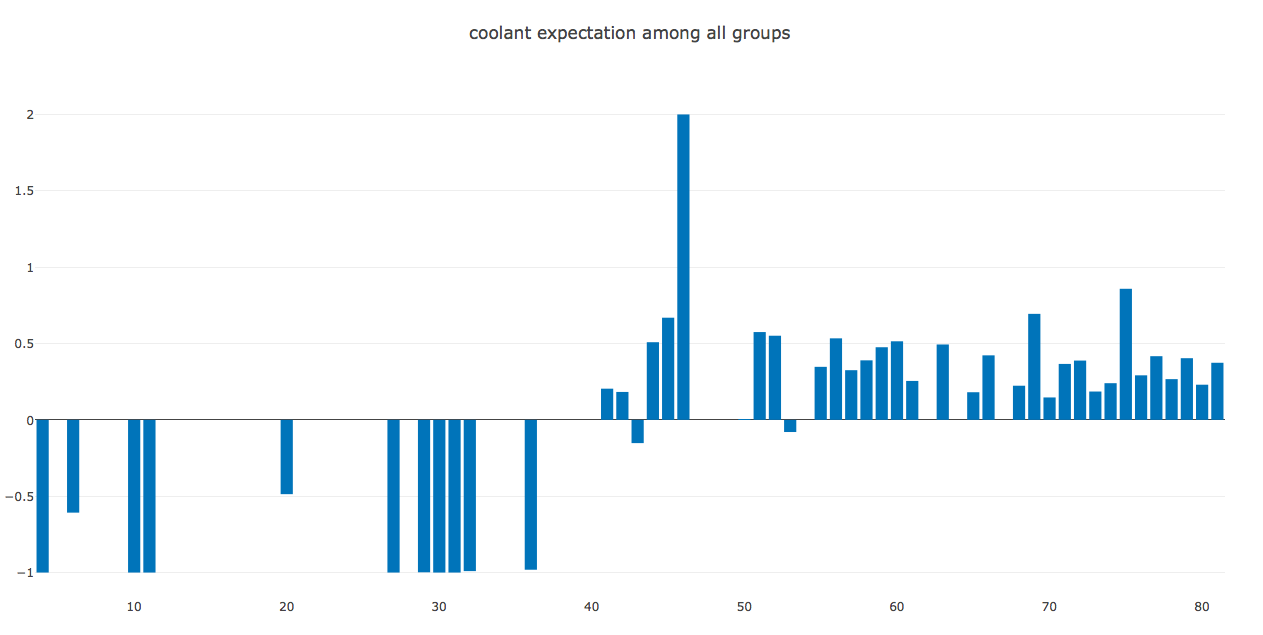
и тд для каждого profile\_id.

Программный код:





По данным мы построили графики для каждого статистического параметра и каждого критерия.

Примеры графиков:

Вывод

При выполнении данной лабораторной работы я познакомился с принципами сбора, анализа и визуализации данных их характеристик при помощи библиотек pandas и plotly, а также повторил основные математические и статистические понятия.