Московский Авиационный Институт

(Национальный Исследовательский Университет)

# Отчет по лабораторной работе 1

## по курсу “Искусственный интеллект”

**На тему: “Анализ и визуализация данных”**

Студент: Буркевич В. Д.

Группа: М80-307Б-17

Преподаватель: Самир Халид

Москва, 2020

**Постановка задачи**

Необходимо сформировать два набора данных для приложений машинного обучения. Первый датасет должен представлять из себя табличный набор данных для задачи классификации. Второй датасет должен быть отличен от первого, и может представлять из себя набор изображений, корпус документов, другой табличный датасет или датасет из соревнования Kaggle, предназначенный для решения интересующей вас задачи машинного обучения. Необходимо провести анализ обоих наборов данных, поставить решаемую вами задачу, определить признаки необходимые для решения задачи, в случае необходимости заняться генерацией новых признаков, устранением проблем в данных, визуализировать распределение и зависимость целевого признака от выбранных признаков. В отчете описать все проблемы, с которыми вы столкнулись и выбранные подходы к их решению.

**Датасеты**

Diamonds <https://www.kaggle.com/shivam2503/diamonds>

# Rain in Australia <https://www.kaggle.com/jsphyg/weather-dataset-rattle-package>

**Описание**

Diamonds – Predict price

**Входные данные**

Index counter-индекс

Carat-карат

Cut-качество огранки алмаза. Качество в порядке возрастания: Удовлетворительно, Хорошо, Очень Хорошо, Премиум, Идеально

Color-цвет бриллианта, где D - лучший, а J - худший

clarity-насколько очевидные включения в алмазе: (от лучшего к худшему, FL = безупречный, I3 = включения уровня 3) FL, IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1, I2, I3

depth-высота алмаза, измеренная от кулета до стола, деленная на средний диаметр пояса

table-ширина алмаза, выраженная в процентах от его среднего диаметра

x- длина мм

y- ширина мм

z- глубина мм

price- Целевая переменная. Цена бриллианта

**Количество строк и колонок, типы данных и размер занимаемой памяти**

RangeIndex: 53940 entries, 0 to 53939

Data columns (total 11 columns):

Unnamed: 0 53940 non-null int64

carat 53940 non-null float64

cut 53940 non-null object

color 53940 non-null object

clarity 53940 non-null object

depth 53940 non-null float64

table 53940 non-null float64

price 53940 non-null int64

x 53940 non-null float64

y 53940 non-null float64

z 53940 non-null float64

dtypes: float64(6), int64(2), object(3)

memory usage: 4.5+ MB

**Признаки, выбранные для решения**

| carat | cut | table | price | x | y | z |
| --- | --- | --- | --- | --- | --- | --- |

**Выбросы**

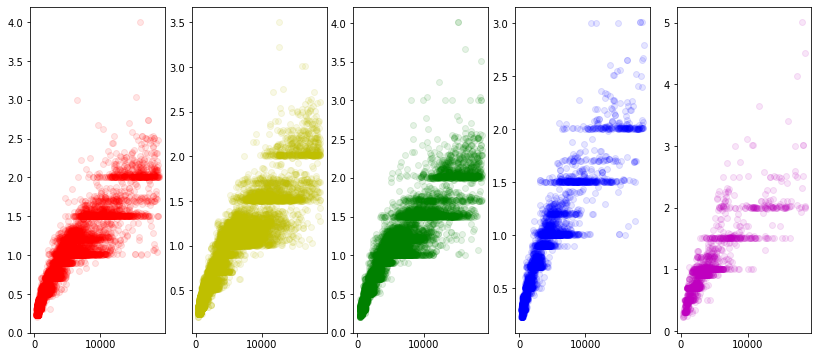
В данном датасете выбросов были единицы, которые можно было убрать и выявить вручную

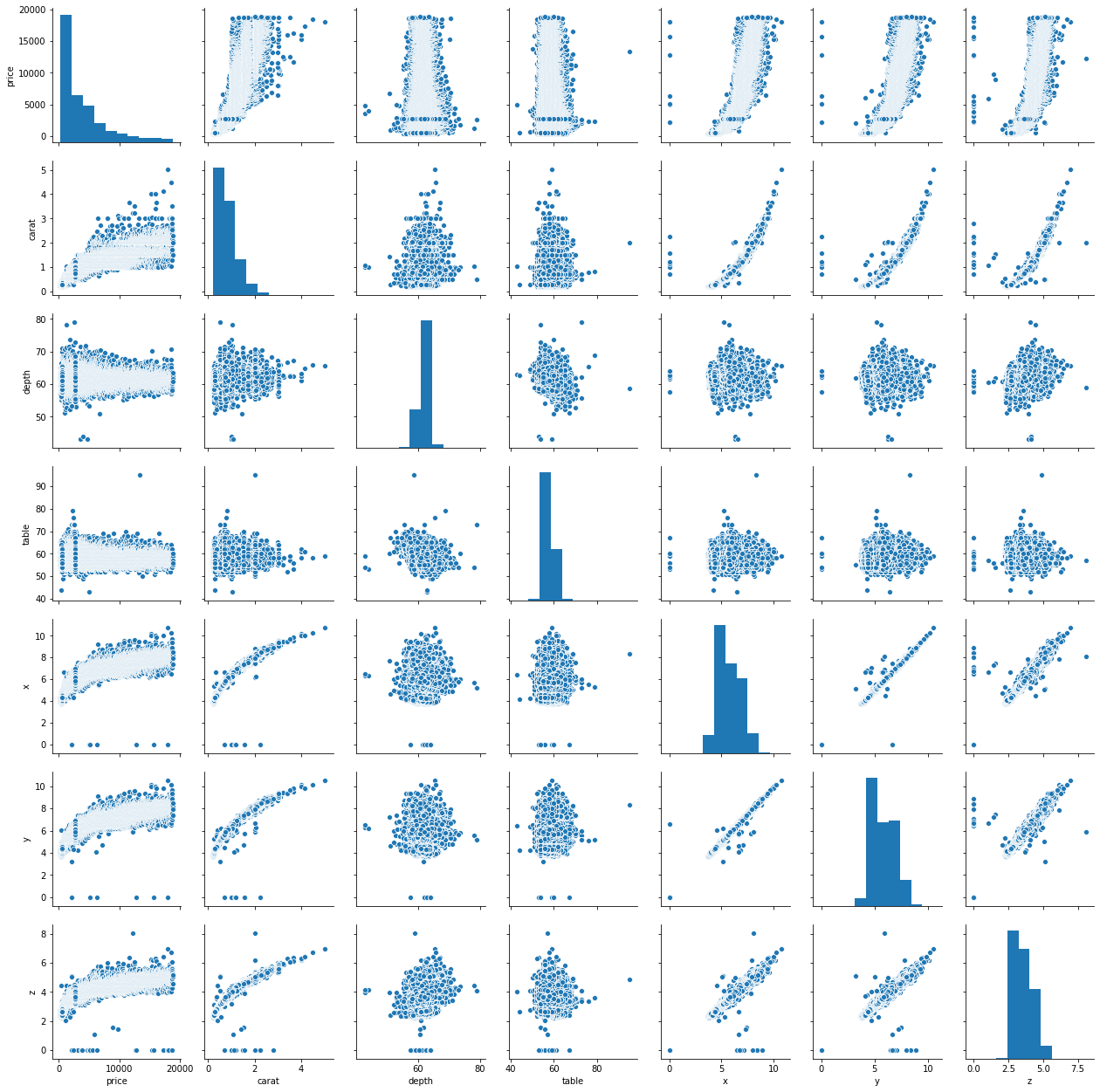
**Пропуски**

В данном датасете отсутствуют NaN-ы

**Визуализация**

При анализе данных были использованы различные варианты визуализации данных. Наиболее интересные из них:





Rain in Australia - Predict rain tomorrow in Australia:

**Входные данные**

Date-дата

Location-локация

MinTemp-минимальная температура в градусах Цельсия

MaxTemp-максимальная температура в градусах Цельсия

Rainfall-осадки в мм

Evaporation-испарение за сутки (до 9 часов утра)

Sunshine-количество часов яркого солнца в день

WindGustDir-направление сильнейшего порыва ветра за 24 часа до полуночи

WindGustSpeed- скорость (км / ч) сильнейшего порыва ветра за 24 часа до полуночи

WindDir9am- направление ветра в 9 утра

WindDir3pm-направление ветра в 15 часов

WindSpeed9am- скорость ветра (км / ч) в среднем за 10 минут до 9 утра

WindSpeed3pm- скорость ветра (км / ч) в среднем за 10 минут до 15 часов

Humidity9am- влажность (процент) в 9 утра

Humidity3pm- влажность (процент) в 3 часа дня

Pressure9am- атмосферное давление (hpa) снижается до среднего уровня моря в 9 часов утра

Pressure3pm- атмосферное давление (hpa) снижается до среднего уровня моря в 3 часа дня

Cloud9am- фракция неба затенена облаками в 9 утра. Это измеряется в «октах», которые являются единицей восьмых. Это записывает, сколько восьмых неба скрыто облаком. Мера 0 указывает на полностью чистое небо, а цифра 8 указывает на то, что оно полностью облачно.

Cloud3pm- фракция неба, затененная облаками (в «октах»: восьмые) в 3 часа дня.

Temp9am- температура в градусах Цельсия в 9 утра

Temp3pm- температура в градусах Цельсия в 15 часов

RainToday- 1, если количество осадков (мм) за 24 часа до 9 часов утра превышает 1 мм, в противном случае 0

RISK\_MM- Количество осадков следующего дня в мм. Используется для создания ответной переменной RainT Tom Этакая мера «риска»

RainTomorrow- Целевая переменная. Завтра будет дождь?

**Количество строк и колонок, типы данных и размер занимаемой памяти**

RangeIndex: 142193 entries, 0 to 142192

Data columns (total 24 columns):

Date 142193 non-null object

Location 142193 non-null object

MinTemp 141556 non-null float64

MaxTemp 141871 non-null float64

Rainfall 140787 non-null float64

Evaporation 81350 non-null float64

Sunshine 74377 non-null float64

WindGustDir 132863 non-null object

WindGustSpeed 132923 non-null float64

WindDir9am 132180 non-null object

WindDir3pm 138415 non-null object

WindSpeed9am 140845 non-null float64

WindSpeed3pm 139563 non-null float64

Humidity9am 140419 non-null float64

Humidity3pm 138583 non-null float64

Pressure9am 128179 non-null float64

Pressure3pm 128212 non-null float64

Cloud9am 88536 non-null float64

Cloud3pm 85099 non-null float64

Temp9am 141289 non-null float64

Temp3pm 139467 non-null float64

RainToday 140787 non-null object

RISK\_MM 142193 non-null float64

RainTomorrow 142193 non-null object

dtypes: float64(17), object(7)

memory usage: 26.0+ MB

**Генерация новых признаков**

Из признака Date были сгенерированы новые признаки, которые использовались в дальнейшем

year, quarter, month, day

**Признаки, выбранные для решения**

MinTemp, MaxTemp, Rainfall, WindGustDir, WindGustSpeed, WindDir3pm, WindSpeed9am, WindSpeed3pm, Humidity9am, Humidity3pm, Cloud9am, Cloud3pm, Temp9am, RainToday, RISK\_MM, RainTomorrow, year, quarter, month, day

**Выбросы**

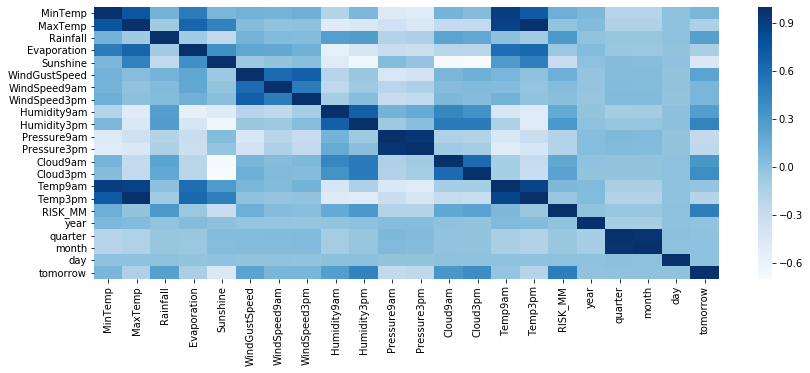
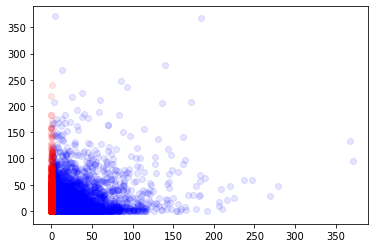
В данном датасете выбросы не найдены

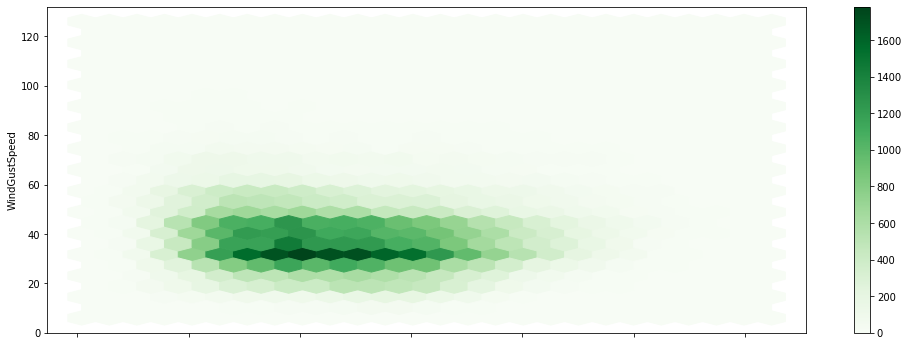
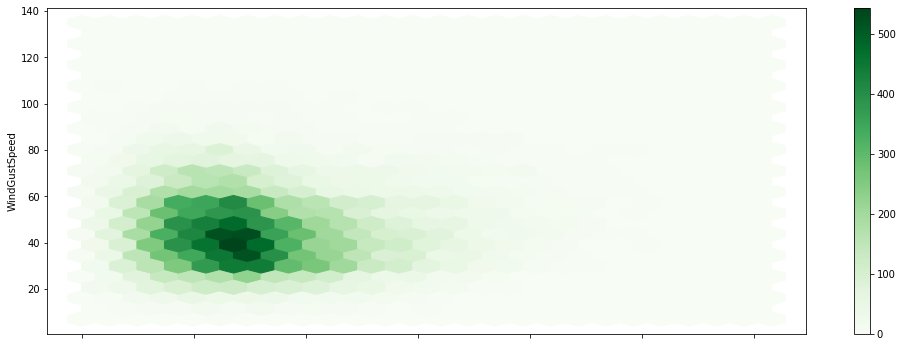
**Пропуски**

В данном датасете большое количество NaN-ов. Их можно заполнить средним значением, близким с наиболее похожим объектом и т.д. Но ввиду того что датасет достаточно объемный, его размер был сокращен за счет удаления объектов с пропусками.

**Визуализация**

При анализе данных были использованы различные варианты визуализации данных. Наиболее интересные из них:



**Вывод:**

В ходе выполнения данной лабораторной работы для каждого из двух датасетов были проанализированы и визуализированы данные, найдены очевидные зависимости. Например, из второго датасета признак RISK\_MM показывает, что только с его помощью уже можно точно предсказать значение целевой переменной.

Более того, для дальнейшей работы были изменены типы данных, что значительно уменьшило размер занимаемой памяти.