Вариант задания

Номер варинта - 9

Номер задачи №1 - 9

Номер задачи №2 - 29

Дополнительные требования по группам:

Для студентов групп ИУ5-23М, ИУ5И-23М - для произвольной колонки данных построить график "Ящик с усами (boxplot)".

Каждая задача предполагает использование набора данных. Набор данных выбирается Вами произвольно с учетом следующих условий:

- Вы можете использовать один набор данных для решения всех задач, или решать каждую задачу на своем наборе данных.
- Набор данных должен отличаться от набора данных, который использовался в лекции для решения рассматриваемой задачи.
- Вы можете выбрать произвольный набор данных (например тот, который Вы использовали в лабораторных работах) или создать собственный набор данных (что актуально для некоторых задач, например, для задач удаления псевдоконстантных или повторяющихся признаков).
- Выбранный или созданный Вами набор данных должен удовлетворять условиям поставленной задачи. Например, если решается задача устранения пропусков, то набор данных должен содержать пропуски.

Условия задач

Задача № 9

Для набора данных проведите устранение пропусков для одного (произвольного) числового признака с использованием метода заполнения "хвостом распределения".

Задача № 29

Для набора данных проведите удаление константных и псевдоконстантных признаков.

Текстовое описание датасета

В качестве датасета будем использовать набор данных, содержащий данные с информацией об автомобиле.Данный набор доступен по адресу:https://www.kaggle.com/datasets/tawfikelmetwally/automobile-dataset

Набор данных имеет следующие атрибуты:

- Name: Уникальный идентификатор для каждого автомобиля.
- MPG: Эффективность использования топлива измеряется в милях на галлон.
- Cylinders: количество цилиндров в двигателе.
- Displacement: объем двигателя с указанием его размера или мощности.
- Horsepower: Выходная мощность двигателя.
- Weight: Вес автомобиля.
- Acceleration: Возможность увеличения скорости, измеряемая в секундах.
- Model Year: год выпуска модели автомобиля.
- Origin: Страна или регион происхождения каждого автомобиля.

Импорт библиотек

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style="ticks")
```

Загрузка данных

```
data = pd.read_csv('Automobile.csv')
```

Первичный анализ данных

Выведем первые 5 строк датасета:

```
data.head()
```

	name	mpg	cylinders	displacement	horsepower	weight	acceleration	model_year	origin
0	chevrolet chevelle malibu	18.0	8	307.0	130.0	3504	12.0	70	usa
1	buick skylark 320	15.0	8	350.0	165.0	3693	11.5	70	usa
2	plymouth satellite	18.0	8	318.0	150.0	3436	11.0	70	usa
3	amc rebel sst	16.0	8	304.0	150.0	3433	12.0	70	usa
4	ford torino	17.0	8	302.0	140.0	3449	10.5	70	usa

Определим размер датасета:

```
data.shape
```

```
(398, 9)
```

```
data.dtypes
```

```
name             object
mpg             float64
cylinders         int64
displacement    float64
horsepower      float64
weight            int64
acceleration    float64
model_year        int64
origin           object
dtype: object
```

Проверим наличие пропусков:

```
data.isnull().sum()
```

```
name            0
mpg             0
cylinders       0
displacement    0
horsepower      6
weight          0
acceleration    0
model_year      0
origin          0
dtype: int64
```

Задача № 9

Удалим колонки, содержащие пустые значения:

```
data_new_1 = data.dropna(axis=1, how='any')
(data.shape, data_new_1.shape)
```

```
((398, 9), (398, 8))
```

Выведем первые строки датасета на экран:

```
data_new_1
```

	name	mpg	cylinders	displacement	weight	acceleration	model_year	origin
0	chevrolet chevelle malibu	18.0	8	307.0	3504	12.0	70	usa
1	buick skylark 320	15.0	8	350.0	3693	11.5	70	usa
2	plymouth satellite	18.0	8	318.0	3436	11.0	70	usa
3	amc rebel sst	16.0	8	304.0	3433	12.0	70	usa
4	ford torino	17.0	8	302.0	3449	10.5	70	usa
...	...	...	...	...	...	...	...	...
393	ford mustang gl	27.0	4	140.0	2790	15.6	82	usa
394	vw pickup	44.0	4	97.0	2130	24.6	82	europe
395	dodge rampage	32.0	4	135.0	2295	11.6	82	usa
396	ford ranger	28.0	4	120.0	2625	18.6	82	usa
397	chevy s-10	31.0	4	119.0	2720	19.4	82	usa

398 rows × 8 columns

Удалим строки, содержащие пустые значения:

```
data_new_2 = data.dropna(axis=0, how='any')
(data.shape, data_new_2.shape)
```

```
((398, 9), (392, 9))
```

```
data_new_2.head()
```

	name	mpg	cylinders	displacement	horsepower	weight	acceleration	model_year	origin
0	chevrolet chevelle malibu	18.0	8	307.0	130.0	3504	12.0	70	usa
1	buick skylark 320	15.0	8	350.0	165.0	3693	11.5	70	usa
2	plymouth satellite	18.0	8	318.0	150.0	3436	11.0	70	usa
3	amc rebel sst	16.0	8	304.0	150.0	3433	12.0	70	usa
4	ford torino	17.0	8	302.0	140.0	3449	10.5	70	usa

Найдем значение квантиля для заполнения пропущенных значений:

```
quantile_value = data_new_3['horsepower'].quantile(0.95)
```

Замена пропущенных значений на значение из хвоста распределения:

```
data_new_3['horsepower'] = data_new_3['horsepower'].fillna(quantile_value)
```

Выведем на экран:

```
data_new_3.head()
```

	name	mpg	cylinders	displacement	horsepower	weight	acceleration	model_year	origin
0	chevrolet chevelle malibu	18.0	8	307.0	130.0	3504	12.0	70	usa
1	buick skylark 320	15.0	8	350.0	165.0	3693	11.5	70	usa
2	plymouth satellite	18.0	8	318.0	150.0	3436	11.0	70	usa
3	amc rebel sst	16.0	8	304.0	150.0	3433	12.0	70	usa
4	ford torino	17.0	8	302.0	140.0	3449	10.5	70	usa

```
data_new_3.isnull().sum()
```

```
name            0
mpg             0
cylinders       0
displacement    0
horsepower      0
weight          0
acceleration    0
model_year      0
origin          0
dtype: int64
```

Задача № 29

Выведем первые 20 строк:

```
data.head()
```

	name	mpg	cylinders	displacement	horsepower	weight	acceleration	model_year	origin
0	chevrolet chevelle malibu	18.0	8	307.0	130.0	3504	12.0	70	usa
1	buick skylark 320	15.0	8	350.0	165.0	3693	11.5	70	usa
2	plymouth satellite	18.0	8	318.0	150.0	3436	11.0	70	usa
3	amc rebel sst	16.0	8	304.0	150.0	3433	12.0	70	usa
4	ford torino	17.0	8	302.0	140.0	3449	10.5	70	usa

Анализ константных и псевдоконстантных признаков:

```
constant_features = [feat for feat in data.columns if data[feat].nunique() == 1]
pseudo_constant_features = [feat for feat in data.columns if data[feat].value_counts(normalize=True).values[0] > 0.9
```

Удаление константных и псевдоконстантных признаков:

```
data.drop(columns=constant_features + pseudo_constant_features, inplace=True)
```

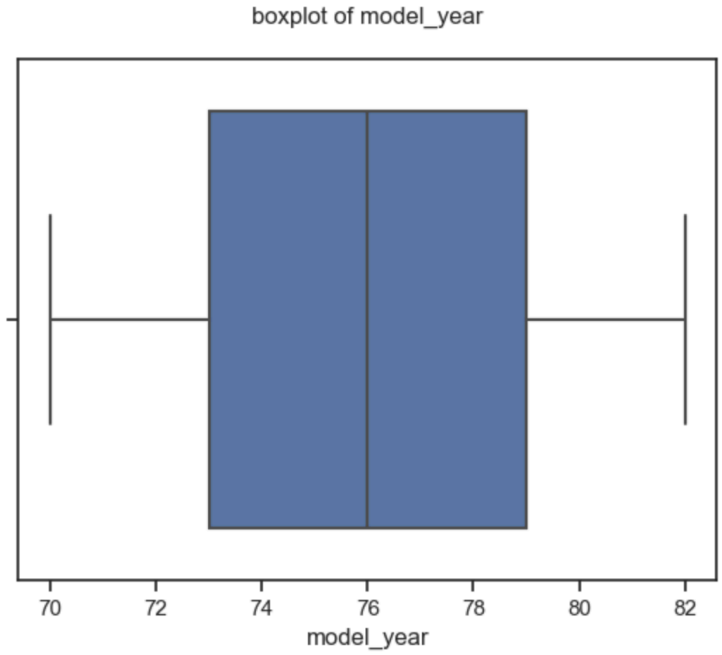
Выведем полученный результат:

```
data.head()
```

	name	mpg	cylinders	displacement	horsepower	weight	acceleration	model_year	origin
0	chevrolet chevelle malibu	18.0	8	307.0	130.0	3504	12.0	70	usa
1	buick skylark 320	15.0	8	350.0	165.0	3693	11.5	70	usa
2	plymouth satellite	18.0	8	318.0	150.0	3436	11.0	70	usa
3	amc rebel sst	16.0	8	304.0	150.0	3433	12.0	70	usa
4	ford torino	17.0	8	302.0	140.0	3449	10.5	70	usa

Отображение в виде "Ящика с усами":

```
sns.boxplot(data=data,x='model_year')
plt.title('boxplot of model_year\n')
plt.show()
```




```
sns.boxplot(x='cylinders', y='horsepower', data=data)
```

```
<Axes: xlabel='cylinders', ylabel='horsepower'>
```