МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
им. Н.Э. Баумана

Кафедра «Систем обработки информации и управления»

Лабораторная работа №1
по курсу «Методы машинного обучения»

ИСПОЛНИТЕЛЬНИЦА: Абросимова Н.Г.
ИУ5-24М

_____
подпись

"__"_____2021 г.

ПРЕПОДАВАТЕЛЬ: _____
ФИО

_____
подпись

"__"_____2021 г.

Москва - 2021

Задание

- Выбрать набор данных (датасет).

- Создать "историю о данных" в виде юпитер-ноутбука, с учетом следующих требований:

1. История должна содержать не менее 5 шагов (где 5 - рекомендуемое количество шагов). Каждый шаг содержит график и его текстовую интерпретацию.
2. На каждом шаге наряду с удачным итоговым графиком рекомендуется в юпитер-ноутбуке оставлять результаты предварительных "неудачных" графиков.
3. Не рекомендуется повторять виды графиков, желательно создать 5 графиков различных видов.
4. Выбор графиков должен быть обоснован использованием методологии data-to-viz. Рекомендуется учитывать типичные ошибки построения выбранного вида графика по методологии data-to-viz. Если методология Вами отвергается, то просьба обосновать Ваше решение по выбору графика.
5. История должна содержать итоговые выводы. В реальных "историях о данных" именно эти выводы представляют собой основную ценность для предприятия.

Текст программы и экранные формы

Набор данных – датасет с характеристиками экзопланет.

Наименование показателя	Идентификатор
Идентификатор планеты	PlanetIdentifier
Масса планеты [масса Юпитера]	PlanetaryMassJpt
Радиус [радиусы Юпитера]	RadiusJpt
Период [дни]	PeriodDays
Большая полуось [астрономические единицы]	SemiMajorAxisAU
Температура поверхности [K]	SurfaceTempK
Год открытия	DiscoveryYear
Расстояние от Солнца [парсек]	DistFromSunParsec
Масса родительской звезды [масса Солнца]	HostStarMassSlrMass
Радиус родительской звезды [радиусы Солнца]	HostStarRadiusSlrRad
Температура родительской звезды [к]	HostStarTempK
Большая полуось планеты	semimajoraxis

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

data=pd.read_csv('oec.csv', sep=",")
data.head()
```

Out[3]:

	PlanetIdentifier	TypeFlag	PlanetaryMassJpt	RadiusJpt	PeriodDays	SemiMajorAxisAU	Eccentricity	PeriastronDeg	LongitudeDeg	AscendingNodeDeg
0	HD 143761 b	0	1.0450	NaN	39.845800	0.2196	0.037	270.6	NaN	NaN
1	HD 143761 c	0	0.0790	NaN	102.540000	0.4123	0.050	190.0	NaN	NaN
2	KOI-1843.03	0	0.0014	0.054	0.176891	0.0048	NaN	NaN	NaN	NaN
3	KOI-1843.01	0	NaN	0.114	4.194525	0.0390	NaN	NaN	NaN	NaN
4	KOI-1843.02	0	NaN	0.071	6.356006	0.0520	NaN	NaN	NaN	NaN

5 rows × 25 columns

```
#размер датасета
data.shape
```

(3584, 25)

```
data.dtypes
```

Out[5]:

```
PlanetIdentifier         object
TypeFlag                  int64
PlanetaryMassJpt        float64
RadiusJpt               float64
PeriodDays              float64
SemiMajorAxisAU         float64
Eccentricity            float64
PeriastronDeg           float64
LongitudeDeg            float64
AscendingNodeDeg        float64
InclinationDeg          float64
SurfaceTempK            float64
AgeGyr                  float64
DiscoveryMethod          object
DiscoveryYear           float64
LastUpdated              object
RightAscension           object
Declination              object
DistFromSunParsec       float64
HostStarMassSlrMass     float64
HostStarRadiusSlrRad    float64
HostStarMetallicity     float64
HostStarTempK           float64
HostStarAgeGyr          float64
ListsPlanetIsOn          object
dtype: object
```

In [6]:

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

Out[6]:

```
PlanetIdentifier           0
TypeFlag                   0
PlanetaryMassJpt        2271
RadiusJpt                810
PeriodDays                99
SemiMajorAxisAU         2178
Eccentricity            2476
PeriastronDeg           3256
LongitudeDeg            3541
```

```
AscendingNodeDeg        3538
InclinationDeg          2919
SurfaceTempK            2843
AgeGyr                  3582
DiscoveryMethod           63
DiscoveryYear             10
LastUpdated                8
RightAscension            10
Declination               10
DistFromSunParsec       1451
HostStarMassSlrMass      168
HostStarRadiusSlrRad     321
HostStarMetallicity     1075
HostStarTempK            129
HostStarAgeGyr          3067
ListsPlanetIsOn            0
dtype: int64
```

In [7]:

```
# Основные статистические характеристки набора данных
data.describe()
```

Out[7]:

	TypeFlag	PlanetaryMassJpt	RadiusJpt	PeriodDays	SemiMajorAxisAU	Eccentricity	PeriastronDeg	LongitudeDeg	AscendingNodeDeg	Inc
count	3584.000000	1313.000000	2774.000000	3485.000000	1406.000000	1108.000000	328.000000	43.000000	46.000000	665
mean	0.097656	2.890944	0.371190	537.248317	2.000170	0.166910	150.363823	144.200847	90.624476	82.
std	0.424554	10.204485	0.416871	7509.660676	19.352699	0.189760	117.859945	127.865952	93.047968	21.
min	0.000000	0.000008	0.002300	0.090706	0.004420	0.000000	-233.000000	-174.640000	-5.112604	-0.(
25%	0.000000	0.150000	0.141062	4.757940	0.053000	0.020000	66.750000	37.167396	1.509500	85.(
50%	0.000000	0.940000	0.209600	13.071630	0.169500	0.100000	139.700000	162.280000	69.821251	87.(
75%	0.000000	2.500000	0.321518	49.514000	1.250000	0.247282	243.000000	252.625834	169.175000	89.
max	3.000000	263.000000	6.000000	320000.000000	662.000000	0.956000	791.000000	339.300000	320.800000	305

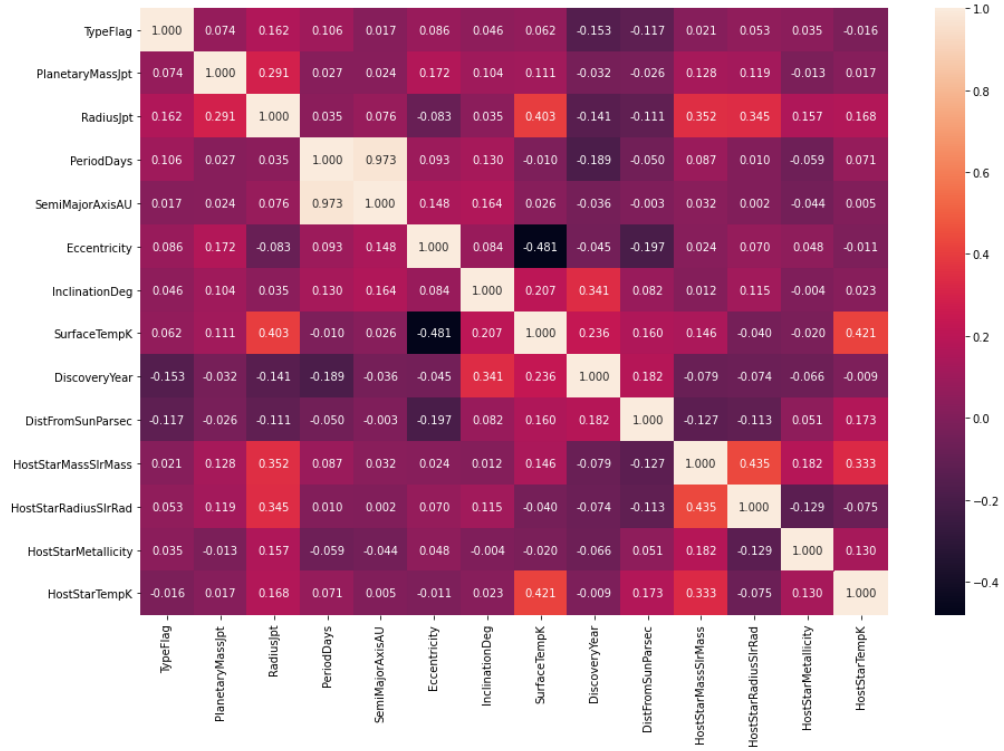
In [8]:

```
data=data.drop(['AgeGyr','LongitudeDeg','LongitudeDeg', 'PeriastronDeg', 'Asc
endingNodeDeg', 'HostStarAgeGyr'], axis='columns')
```

In [9]:

```
plt.figure(figsize=(15,10))
sns.heatmap(data.corr(), annot=True, fmt='.3f')
```

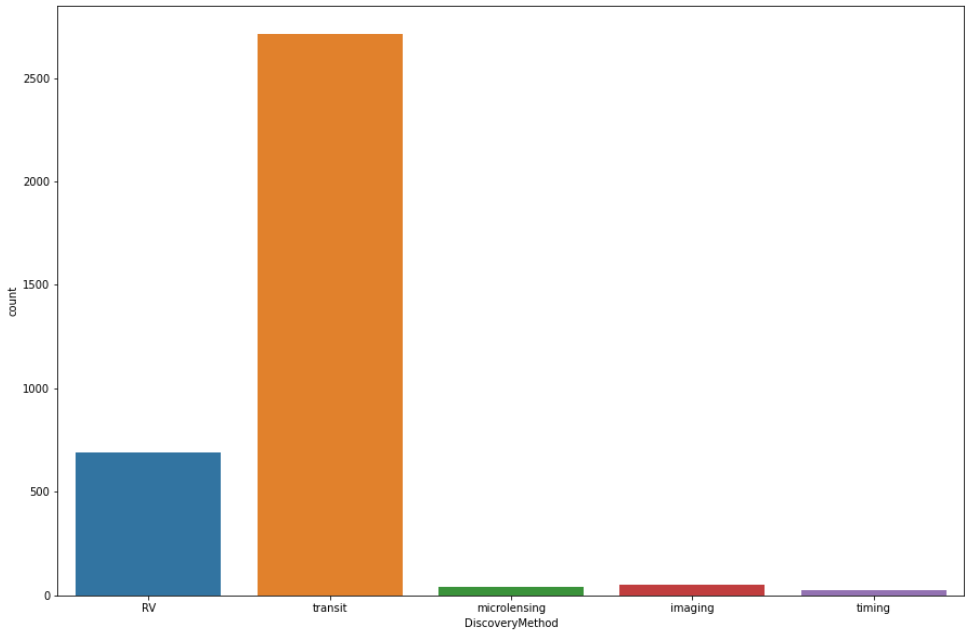
Out[9]: <AxesSubplot:>



На основе полученной таблицы можно сделать вывод, что наиболее связанными являются пары показателей SurfaceTempK и PlanetaryMassJpt, HostStarMassSlrMass и PlanetaryMassJpt, HostStarRadiusSlrRad и PlanetaryMassJpt, SurfaceTempK и RadiusJpt, HostStarRadiusSlrRad и RadiusJpt, SurfaceTempK и DistFromSunParsec, SurfaceTempK и HostStarTempK, DistFromSunParsec и HostStarTempK, HostStarMassSlrMass и HostStarRadiusSlrRad. То есть, можно отметить, что наибольшая связь наблюдается между температурными и размерными характеристиками экзопланет и их звёзд.

```
plt.figure(figsize=(15,10))
sns.countplot(x="DiscoveryMethod", data=data)
```

Out[10]: <AxesSubplot:xlabel='DiscoveryMethod', ylabel='count'>

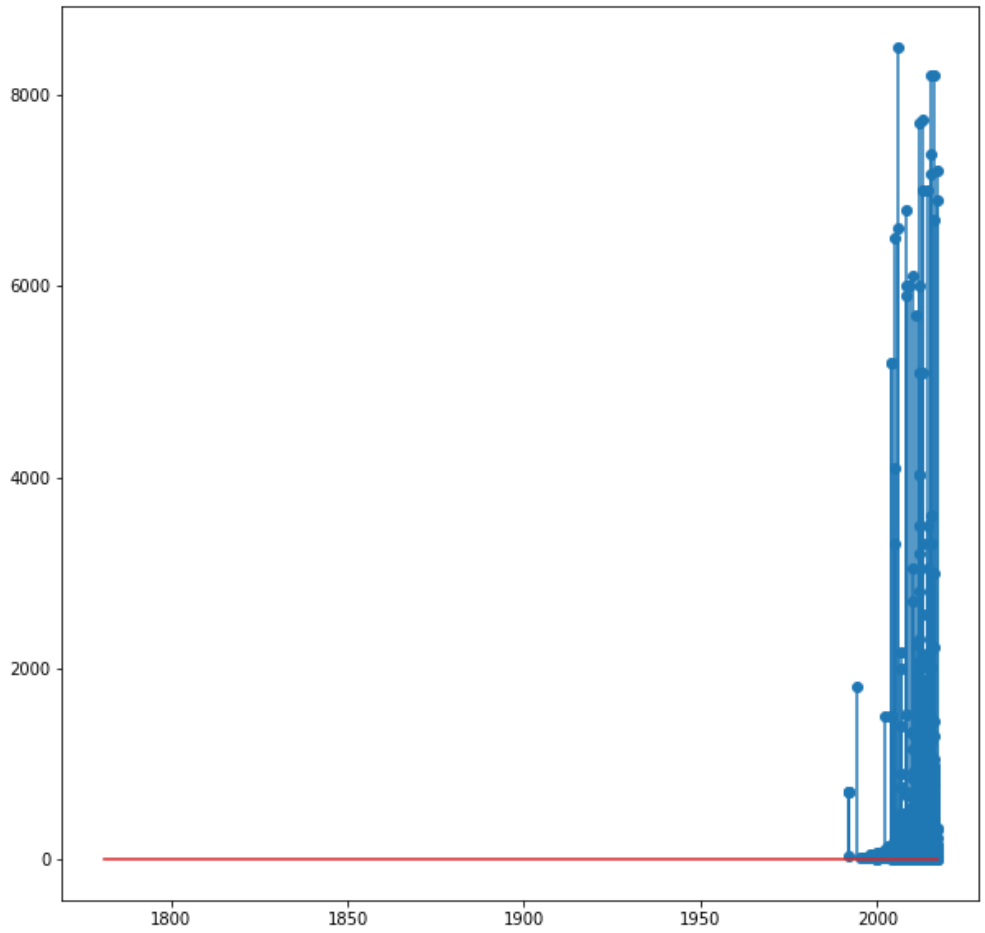


По этому графику можно понять, что самым распространённым методом обнаружения планет является транзитный метод (наблюдение за прохождением планеты на фоне

звезды). Следующим по распространённости является метод радиальных скоростей.

```
plt.figure(figsize=(10,10)) plt.stem(data['DiscoveryYear'],
data['DistFromSunParsec'])
```

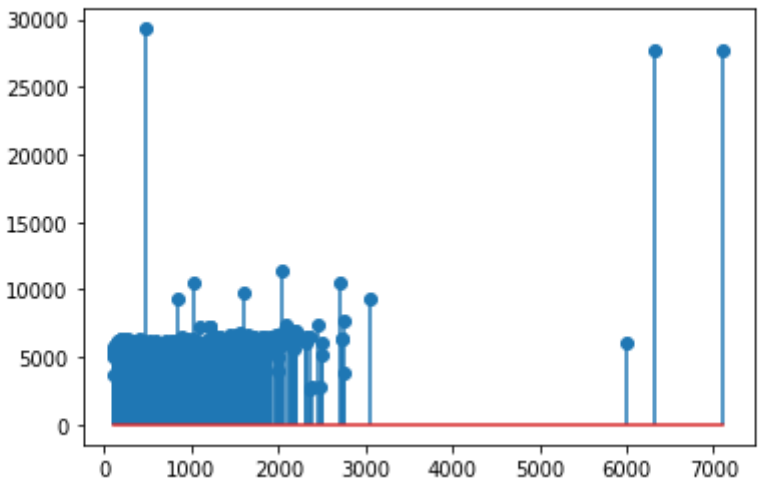
```
Out[11]: <StemContainer object of 3 artists>
```



Связь между расстоянием от солнечной системы и годом обнаружения. В основном, более далёкие планеты были обнаружены позже, но не всегда.

```
x=data['SurfaceTempK']
y = data['HostStarTempK']
plt.stem(x, y)
```

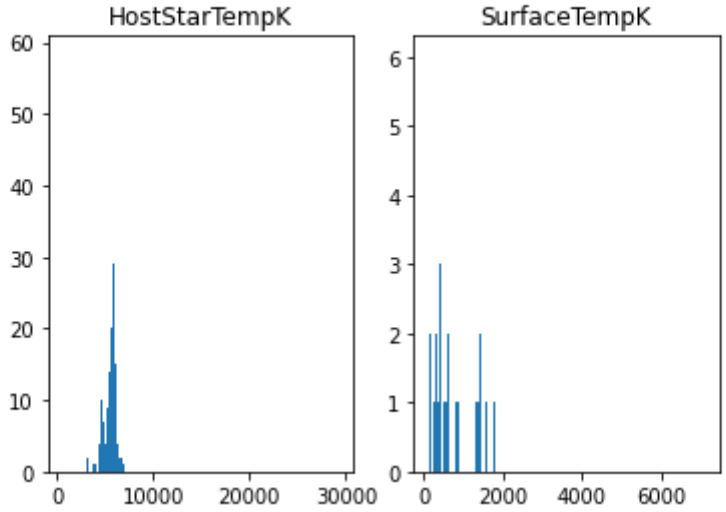
```
Out[13]: <StemContainer object of 3 artists>
```



Попытка визуализировать зависимость температур

```
fig, axs=plt.subplots(1,2) n=len(data) axs[0].hist(data['HostStarTempK'],
bins=n) axs[0].set_title('HostStarTempK')
axs[1].hist(data['SurfaceTempK'], bins=n) axs[1].set_title('SurfaceTempK')
```

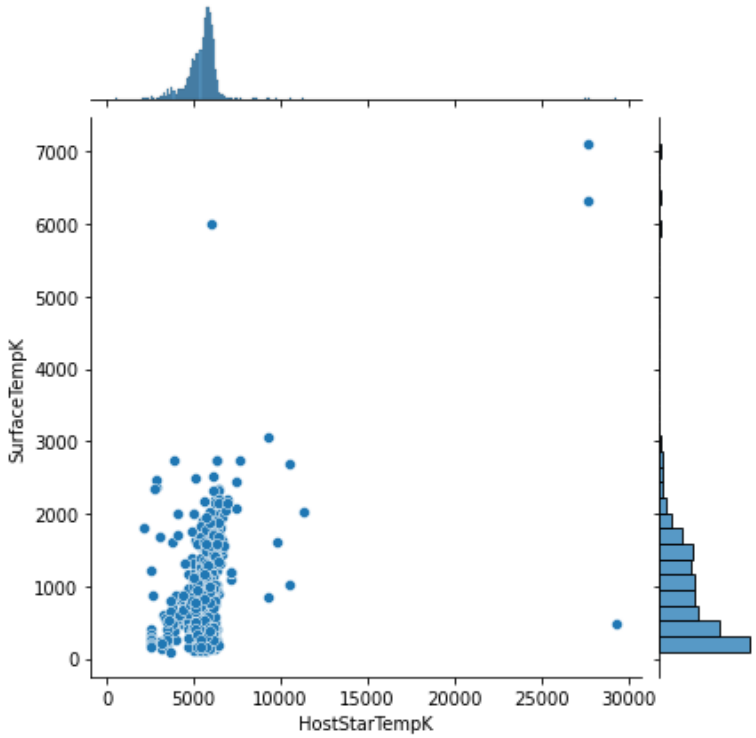
```
Out[14]: Text(0.5, 1.0, 'SurfaceTempK')
```



На гистограмме показаны распределения значений температуры звезды и температуры поверхности

```
plot=sns.jointplot(x=data['HostStarTempK'], y=data['SurfaceTempK'])
plt.show
```

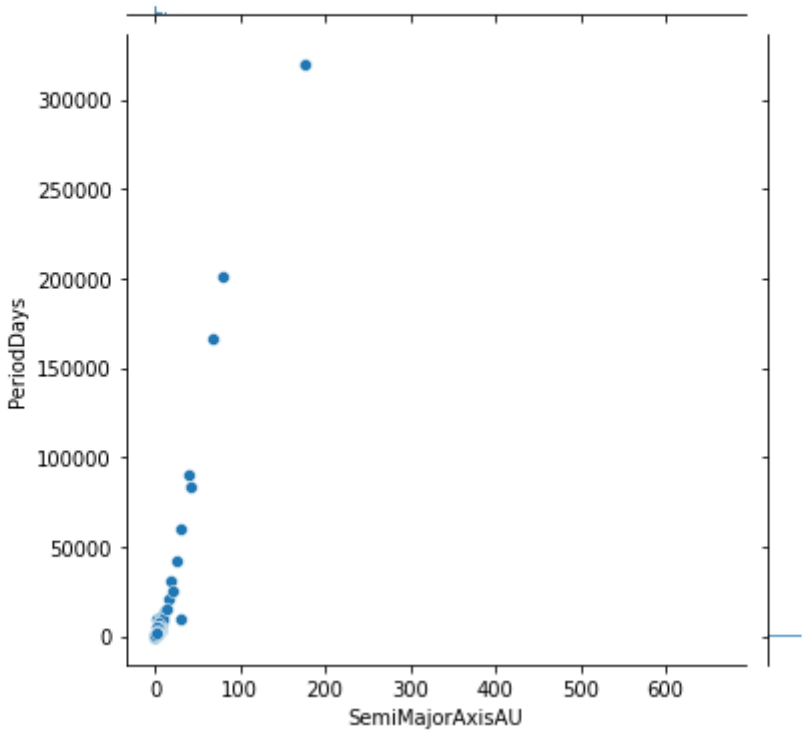
```
Out[15]: <function matplotlib.pyplot.show(close=None, block=None)>
```



Теперь взаимосвязь температур отображается более наглядно - прямая зависимость.

```
plot=sns.jointplot(x=data['SemiMajorAxisAU'], y=data['PeriodDays'])
plt.show
```

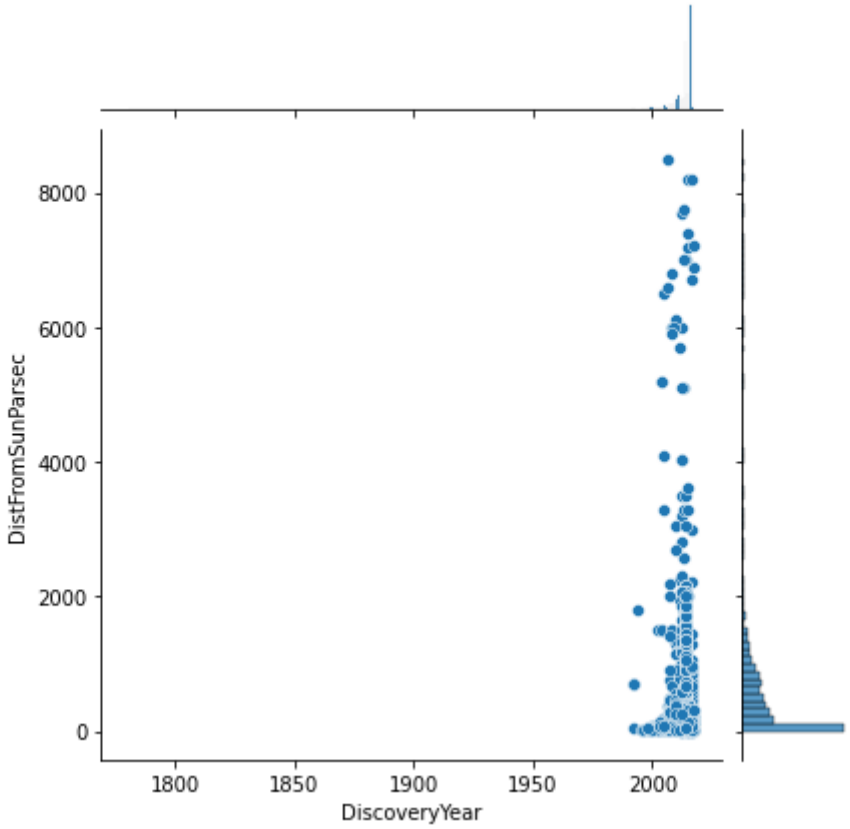
```
Out[16]: <function matplotlib.pyplot.show(close=None, block=None)>
```



Зависимость между периодом и большой полуосью планеты - чем больше расстояние от звезды, чем дольше период

```
plot=sns.jointplot(x=data['DiscoveryYear'], y=data['DistFromSunParsec'])
plt.show
```

```
Out[17]: <function matplotlib.pyplot.show(close=None, block=None)>
```

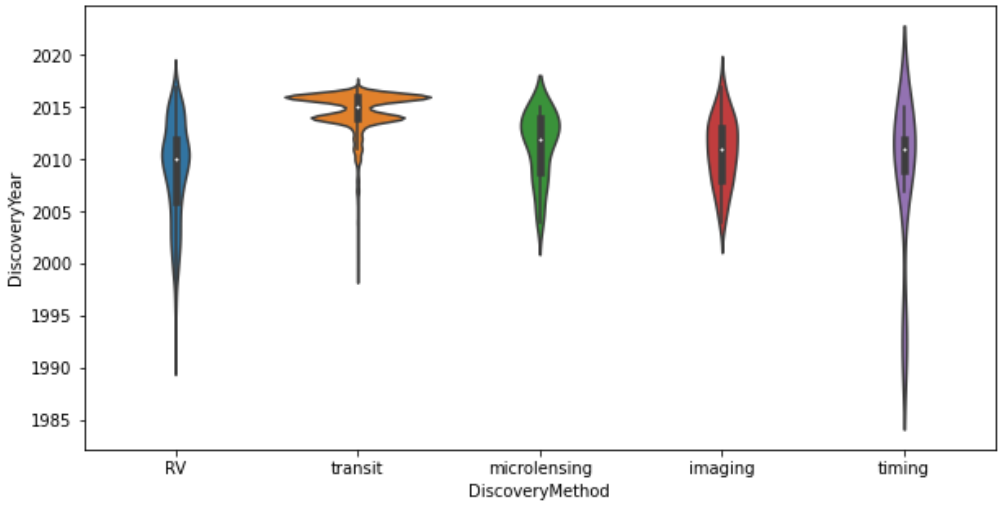



```
plt.figure(figsize=(10,5))
x=data['DiscoveryMethod']

y=data['DiscoveryYear']

sns.violinplot(x, y)
```

```
Out[18]: <AxesSubplot:xlabel='DiscoveryMethod', ylabel='DiscoveryYear'>
```



Зависимость методов исследований и года открытия

Итоги

Были выделены наиболее связанные между собой характеристики – наиболее связанными оказались температуры поверхности планеты и звезды, а также размерные характеристики планет и звёзд, а также подтвердилось предположение, что время обнаружения планеты и дальность их от солнца также связаны; рассмотрена связь между методами исследования и годами.

Знание и изучение закономерностей между характеристиками планет может помочь в случаях, если нужны предположения о тех данных исследования, которые пока являются неполными.