Министерство образования и науки Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ **ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ**

КАФЕДРА **СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ**

Отчет по лабораторной работе № 1
«Создание "истории о данных" (Data Storytelling)»
по курсу “Методы машинного обучения”

Исполнитель:
Студент группы ИУ5-22М
Желанкина А.С.
11.02.2021

Москва, 2021

Задание лабораторной работы

Выбрать набор данных (датасет). Вы можете найти список свободно распространяемых датасетов здесь. Для лабораторных работ не рекомендуется выбирать датасеты очень большого размера.

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

- История должна содержать не менее 5 шагов (где 5 - рекомендуемое количество шагов). Каждый шаг содержит график и его текстовую интерпретацию.
- На каждом шаге наряду с удачным итоговым графиком рекомендуется в юпитер-ноутбуке оставлять результаты предварительных "неудачных" графиков.
- Не рекомендуется повторять виды графиков, желательно создать 5 графиков различных видов.
- 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.
- История должна содержать итоговые выводы. В реальных "историях о данных" именно эти выводы представляют собой основную ценность для предприятия.
- Сформировать отчет и разместить его в своем репозитории на github.

Описание датасета

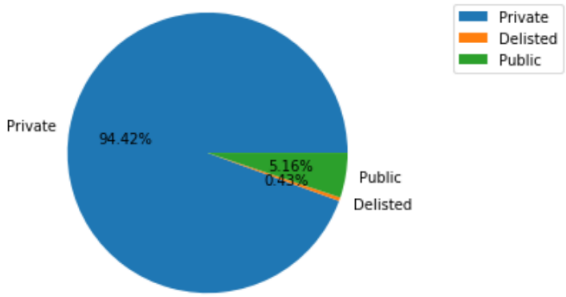
Рассмотрим статистику стартапов, которые были созданы в промежутке между 2011 и 2012 годами. Выбор этого периода объясняется тем, что за это время часть исследуемых стартапов с большой вероятностью достигла поставленных целей. В то время как стартапы основанные после 2013 года рассматривать рано, так как многие из них еще не успели достичь правильно интерпретируемых результатов. Для построения модели была использована база стартапов Crunchbase. Из неё был сформирован датасет, состоящий из 3987 строк и 19 столбцов.

Экранные формы с текстом программы и примерами её выполнения

Для начала было решено просмотреть состав двух переменных, из которых собирается целевая. Первой была рассмотрена переменная 'IPO Status'.

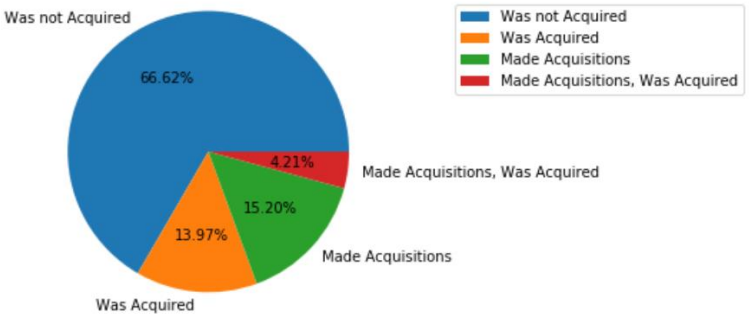
IPO может принимать такие значения, как «Private», «Delisted» и «Public». В случае успешного IPO повышается рыночная стоимость компании. Рассмотрим подробнее принимаемые переменной значения. «Public» статус означает, что стартапу удалось успешно разместить акции на бирже, «Private» – компания ещё не успела провести IPO и до этого момента считается частной. «Delisted» – компания перестала выставляться как публичная, возможно ей не удалось выйти на IPO или же она предпочла вид частного капитала. Обзор распределения в представленных данных показал, что большую часть рынка (94,42%) занимают компании с «Private» статусом, то есть еще не разместившие свои акции на бирже (рис. 1). Публичных компаний значительно меньше – 5,16%, в то время как стартапов со статусом «Delisted» всего 0,43%. Успешными стартапами в данном случае будут считаться компании, вышедшие на IPO («Public»).

In [20]:
```
fig1, ax1 = plt.subplots()

wedges, texts, autotexts = ax1.pie([3771, 17, 206], labels=labels, autopct='%1.2f%%')
ax1.axis('equal')
ax1.legend(loc='upper left', bbox_to_anchor=(1.0, 1.0))
plt.show()
```



Вторая переменная, Acquisition Status, обозначает статус приобретения стартапа и так же имеет четыре значения: не была продана (“Was not Acquired”), была продана (“Was Acquired”), приобрела другую компанию (“Made Acquisitions”), приобрела другую компанию и была куплена (“Made Acquisitions, Was Acquired”). Большую часть рынка (66,62%) занимают стартапы, которые еще не были приобретены (рис. 2). Приобретенные компании составляют 13,97% от общего числа. Стартапы, совершившие покупку других компаний составляют 15,20%, а стартапы с обеими операциями насчитывают всего лишь 4,21%. Статус «Was Acquired» используется в случае, если компания была продана, что является одним из параметров оценки успешности стартапа. Также можно считать успешной компанию, чей статус равен “Made Acquisitions, Was Acquired”, так как это означает, что компания была продана и при этом успела приобрести стартап. Статус «Made Acquisitions» как правило связан с покупкой другого стартапа.

```
In [22]:
fig1, ax1 = plt.subplots()

wedges, texts, autotexts = ax1.pie([2661, 558, 607, 168], labels=labels, autopct='%1.2f%%')
ax1.axis('equal')
ax1.legend(loc='upper left', bbox_to_anchor=(1.0, 1.0))
plt.show()
```



Переменные IPO Status и Acquisition Status будут рассмотрены в паре. Поэтому из них будет создана целевая переменная ‘target’. Посмотрим, есть ли явная корреляция целевой переменной с какой-либо другой из набора. Можно заметить, что целевая переменная ни с одной другой не имеет сильной связи. Однако сильно взаимосвязаны оказались число раундов инвестиций и количество инвесторов, а также зарегистрированные торговые марки и патенты, которыми владеет компания.

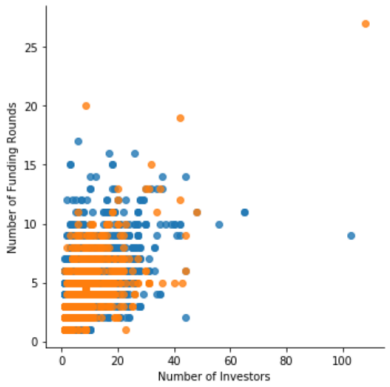
```
In [31]:
sns.heatmap(corr)
```

```
Out[31]: <matplotlib.axes._subplots.AxesSubplot at 0x1ca394ef7b8>
```

Рассмотрим найденные корреляции поподробнее. Большинство успешных стартапов имеет число раундов инвестиций не более 10 и количество инвесторов до 20.

```
In [36]:
#sns.regplot(x=data['Number of Investors'], y=data['Number of Funding Rounds'])
sns.lmplot(x='Number of Investors', y='Number of Funding Rounds', data=data, fit_reg=False, hue='target', legend=False)
```

```
Out[36]: <seaborn.axisgrid.FacetGrid at 0x14c7ec88160>
```



Целевая переменная имеет только два значения: 0 – неуспешный стартап, 1 – успешный стартап. С помощью следующего графика можно проиллюстрировать, что успешных стартапов в несколько раз меньше.

```
In [29]:
sns.distplot(data['target'], hist=True, kde=False, rug=False)
```

```
Out[29]: <matplotlib.axes._subplots.AxesSubplot at 0x14c7f4bc7f0>
```

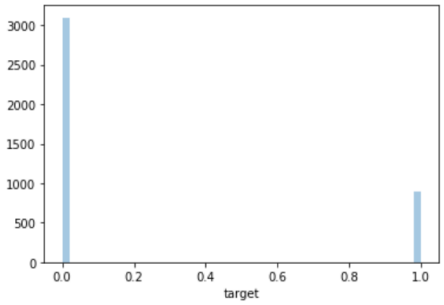
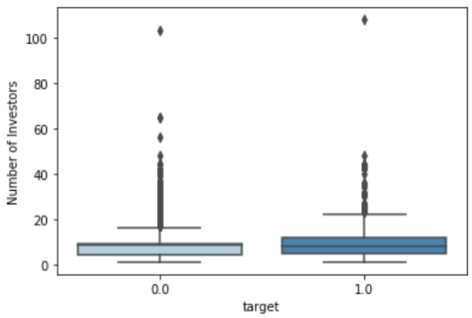


Диаграмма ящик с усами в удобной форме показывает медиану, нижний и верхний квартили, минимальное и максимальное значение выборки и выбросы. Рассмотрим ящики для целевой переменной по числу раундов инвестиций и количеству инвесторов. Медиана успешных стартапов по числу инвесторов находится в районе 10, а количеству инвесторов – 4. В обоих случаях имеются выбросы вверх, что требует дальнейшего изучения.

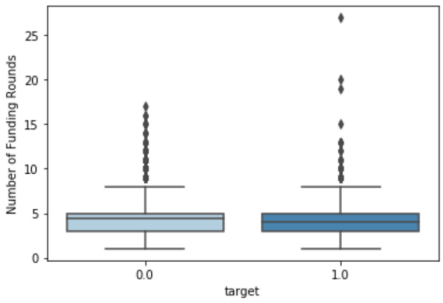
```
In [42]:
#sns.boxplot( x=data['target'], y=label_enc.inverse_transform(data['Funding Status']), palette="Blues")
sns.boxplot(x=data['target'], y=data['Number of Investors'], palette="Blues")
```

```
Out[42]: <matplotlib.axes._subplots.AxesSubplot at 0x14c01d86550>
```



In [43]:
```
#sns.boxplot( x=data['target'], y=label_enc.inverse_transform(data['Funding Status']), palette="Blues")
sns.boxplot(x=data['target'], y=data['Number of Funding Rounds'], palette="Blues")
```

Out[43]: <matplotlib.axes._subplots.AxesSubplot at 0x14c01dfc320>



Отношение успешных компаний по целевой переменной к индустриям и группам. На график выведем десятку лидеров. Лидирующую позицию с большим отрывом от остальных занимают стартапы, которые занимаются разработкой приложений.

In [37]:
```
x = most_fr.keys()
y = most_fr.values()
fig, ax = plt.subplots()
ax.bar(x, y, linewidth = 3)

#  Устанавливаем интервал основных делений:
ax.xaxis.set_major_locator(ticker.MultipleLocator(1))
#  Устанавливаем интервал вспомогательных делений:
ax.xaxis.set_minor_locator(ticker.MultipleLocator(1))

fig.set_figwidth(30)
fig.set_figheight(5)

plt.show()
```