Министерство науки и высшего образования 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 и искусственный интеллект

КАФЕДРА Системы обработки информации и управления

Лабораторная работа №1

По курсу
«Методы машинного обучения»
На тему:
«Создание "истории о данных"»

Подготовил:

Студент группы

ИУ5-25М Клюкин Н. А.

27.03.2024

Проверил:

Гапанюк Ю.Е.

2024 г.

- **Цель лабораторной работы**: изучение различных методов визуализация данных и создание истории на основе данных. 4
- **Краткое описание**. Построение графиков, помогающих понять структуру данных, и их интерпретация.

Задание

- Выбрать набор данных (датасет);

- 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.

Подключение библиотек

In [2]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

Ход работы

Выбор и описание датасета

- Этот набор данных содержит информацию о взаимодействии с клиентами, продажах и возможностях из системы CRM (Customer Relationship Management) вымышленной компании.
- В рамках данной работы изучим таблицу этого датасета, содержащую информацию о клиентах - юридических лицах. Имеются следующие атрибуты:
  - account. Название клиента;
  - sector. Сектор работы;
  - year_established. Год основания компании;
  - revenue. Выручка;
  - employees. Кол-во сотрудников;
  - office_location. Страна размещения офиса;
  - subsidiary_of. Если является дочерней компаний, то здесь указывается родительская.

In [3]:
```
# Импорт датасета
df = pd.read_csv('datasets/accounts.csv')

# Вывод первых 5 строк
df.head(5)
```

Out[3]:

	account	sector	year_established	revenue	employees	office_location	subsidiary
0	Acme Corporation	technolgy	1996	1100.04	2822	United States	N
1	Betasoloin	medical	1999	251.41	495	United States	N
2	Betatech	medical	1986	647.18	1185	Kenya	N
3	Bioholding	medical	2012	587.34	1356	Philipines	N
4	Bioplex	medical	1991	326.82	1016	United States	N

Создание истории о данных

Введение

- В рамках этой работы попробуем через графики составить охарактеризовать нашего клиента: в каком секторе работает, сколько получает, когда появился и т.д.

Глава 1. В каком секторе работают наши клиенты?

- Для создания данной диаграммы воспользуемся столбчатой диаграммой (barplot), поскольку в подходе *Data to Viz* для данных, содержащих один категориальный атрибут, используют такой тип графика
- Подход из книги *Storytelling with data* и в *Data to Viz* рекомендуют отображать столбцы в отсортированном порядке
- Помимо прочего я решил обозначтить топ-3 сектора другими цветами (топ-1, топ 2-3, соответственно)

Ниже представлен вариант с круговой диаграммой (неудачный). В данном случае её можно было бы использовать, но следовало выделить цветом ТОП-3, а остальных сделать одноцветными (либо сделать градацию цвета от 1 до 10)

In [4]:
```
sector_counts = df['sector'].value_counts()
plt.pie(sector_counts, labels=sector_counts.index, autopct='%1.1f%%')
plt.title('Распределение компаний по секторам работы')
plt.show()
```

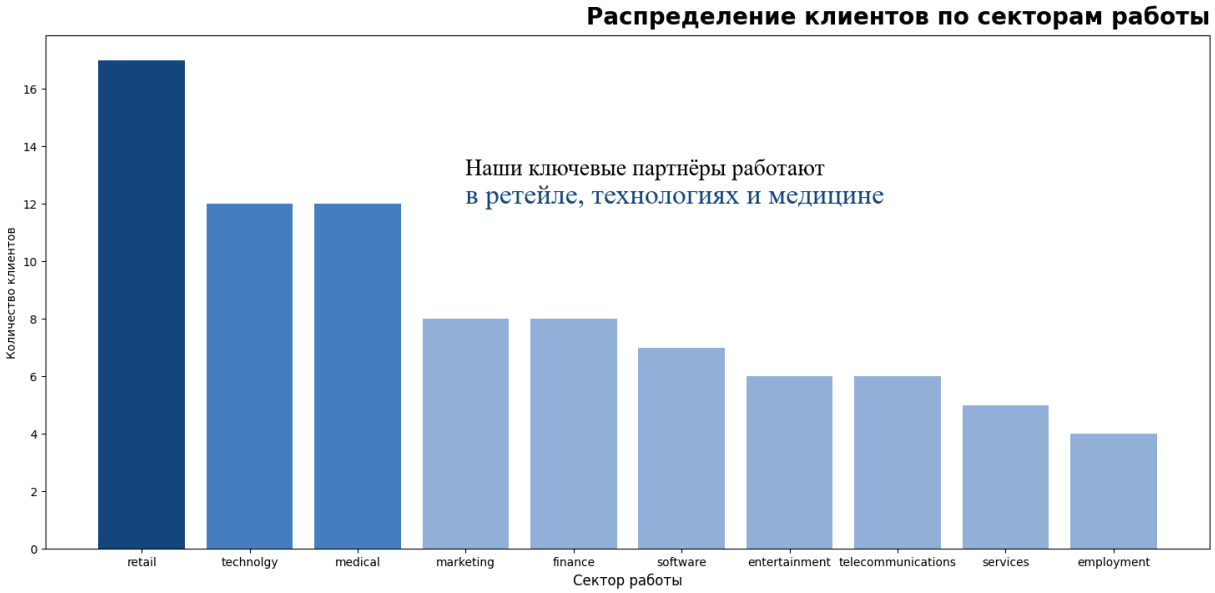
In [5]:
```
# Группировка данных по сектору работы и подсчет количества клиентов в каждом секто
sector_counts = df['sector'].value_counts()

# Задача цвета
colors = ['#174A7E', '#4A81BF', '#4A81BF','#94B2D7','#94B2D7','#94B2D7','#94B2D7','

# Создание графика
plt.figure(figsize=(18, 8))
plt.bar(sector_counts.index, sector_counts.values, color=colors)
plt.xlabel('Сектор работы', fontsize=12)
plt.ylabel('Количество клиентов')
plt.title('Распределение клиентов по секторам работы', fontsize=20, fontweight='bol

# Добавление объясняющего текста
plt.text(3, 13, 'Наши ключевые партнёры работают', fontsize=20, fontname='Times New
plt.text(3, 12, 'в ретейле, технологиях и медицине', fontsize=24, fontname='Times N

# Отображение графика
plt.show()
```



Глава 2. Какова средня прибыль клиентов по секторам?

- Для создания данной диаграммы воспользуемся гистограммой (histogram), поскольку в подходе *Data to Viz* для данных, содержащих один количественный и один категориальный атрибут, используют такой тип графика
- Согласно рекомендациям необходимо:
  - Поиграться с размером столбца
  - Не использовать с более чем 5ю атрибутами
  - Избегать безсмысленного окрашевания
- В нашем случае представлено 10 атрибутов, что значительно больше. Однако, мы используем этот тип диаграммы

In [6]:
```
# Высчитывем среднее
avg_revenue = df.groupby('sector')['revenue'].mean().sort_values(ascending=True)

# Построение гистограммы
colors = ['#94B2D7','#94B2D7','#94B2D7','#94B2D7','#94B2D7','#94B2D7','#94B2D7', '#
plt.barh(avg_revenue.index, avg_revenue.values, color=colors)


# Настройка осей и заголовка
plt.xlabel('Сектор')
plt.ylabel('Среднее значение')
plt.title('Среднее значение выручки по секторам')

# Добавление объясняющего текста
plt.text(10, 12, 'Самые выгодные для нас клиенты\nработают в секторе:', fontsize=15
plt.text(10, 11, 'Разработки, телекоммуникаций, технологий', fontsize=12, fontname=

# Отображение гистограммы
plt.show()
```

Самые выгодные для нас клиенты
работают в секторе:

Разработки, телекоммуникаций, технологий

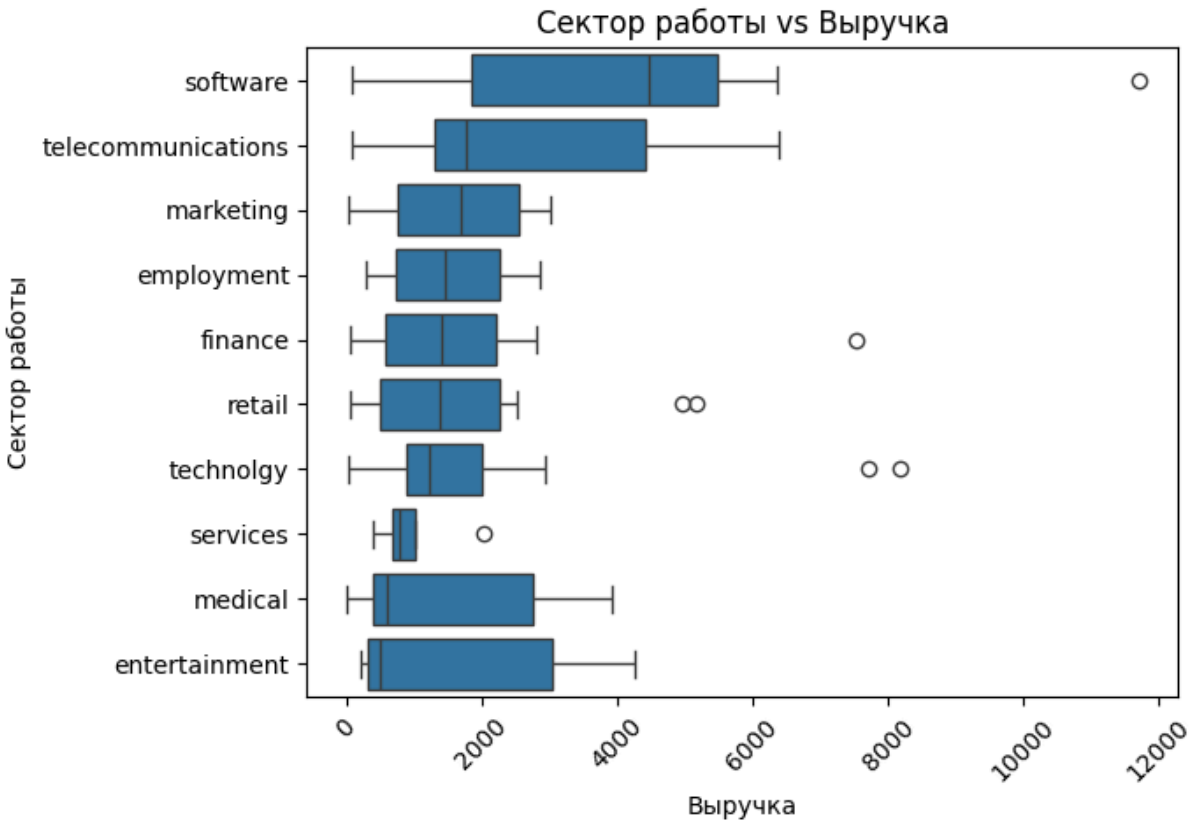
Глава 3. А каков раброс этой выручки?

- Для отображения разброса выручки по секторам воспользуемся диаграммой "коробочек", т.к. на ней можно наблюдать средний интервал для сектора и предел разброса.
- Согласно *Data to Viz* при использовании диаграммы рекомендуется:
  - Упорядочивать коробочки
  -

In [83]:
```
order = df.groupby('sector')['revenue'].median().sort_values(ascending=False).index

sns.boxplot(y=df['sector'], x=df['revenue'], order=order)
plt.ylabel('Сектор работы')
plt.xlabel('Выручка')
plt.title('Сектор работы vs Выручка')
plt.xticks(rotation=45)

plt.show()
```



Глава 4. Кто наши топовые клиенты?

- Поскольку на этой диаграмме мы хотим представить наших ключевых партёров с их долей в нашем бизнесе, то мы используем круговую диаграмму
- Я решил её использовать, поскольку на ней будет чётко видно, что ТОП-4 наших партнеров обладают выручкой более 50% от всей выручки ТОП-10, что выделяет их
- Подход *Data to Viz* рекомендует не использовать легенду, 3д отображение, и не тспользовать их подряд

In [7]:
```
top_10_revenue = df.nlargest(10, 'revenue')
top_10_revenue
```

Out[7]:

	account	sector	year_established	revenue	employees	office_location	s
41	Kan-code	software	1982	11698.03	34288	United States	
35	Hottechi	technolgy	1997	8170.38	16499	Korea	
43	Konex	technolgy	1980	7708.38	13756	United States	
76	Xx-holding	finance	1993	7537.24	20293	United States	
36	Initech	telecommunications	1994	6395.05	20275	United States	
60	Scotfind	software	1996	6354.87	16780	United States	
72	Treequote	telecommunications	1988	5266.09	8595	United States	
25	Ganjaflex	retail	1995	5158.71	17479	Japan	
20	Fasehatice	retail	1990	4968.91	7523	United States	
18	Dontechi	software	1982	4618.00	10083	United States	

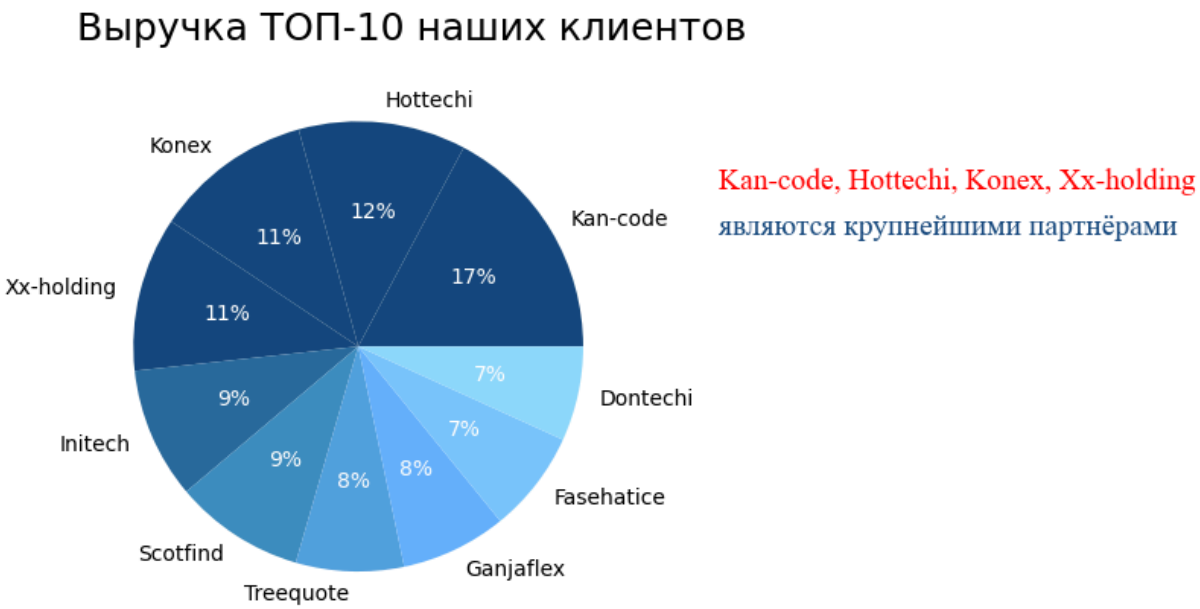
In [81]:
```
colors = ['#174A7E', '#174A7E', '#174A7E', '#174A7E', '#2B6B9E', '#3F8DBE', '#53A0D

labels_4 = top_10_revenue['account'].head(4)

_, _, text = plt.pie(top_10_revenue['revenue'], labels=top_10_revenue['account'], c
plt.setp(text, color='white')
plt.title('Выручка ТОП-10 наших клиентов', loc='left', pad=12, fontsize=18)

# Добавление объясняющего текста
plt.text(1.6, 0.7, 'Kan-code, Hottechi, Konex, Xx-holding', fontsize=14, fontname='
plt.text(1.6, 0.5, 'являются крупнейшими партнёрами', fontsize=14, fontname='Times

plt.show()
```



Глава 5. Связь между годом основания и выручкой

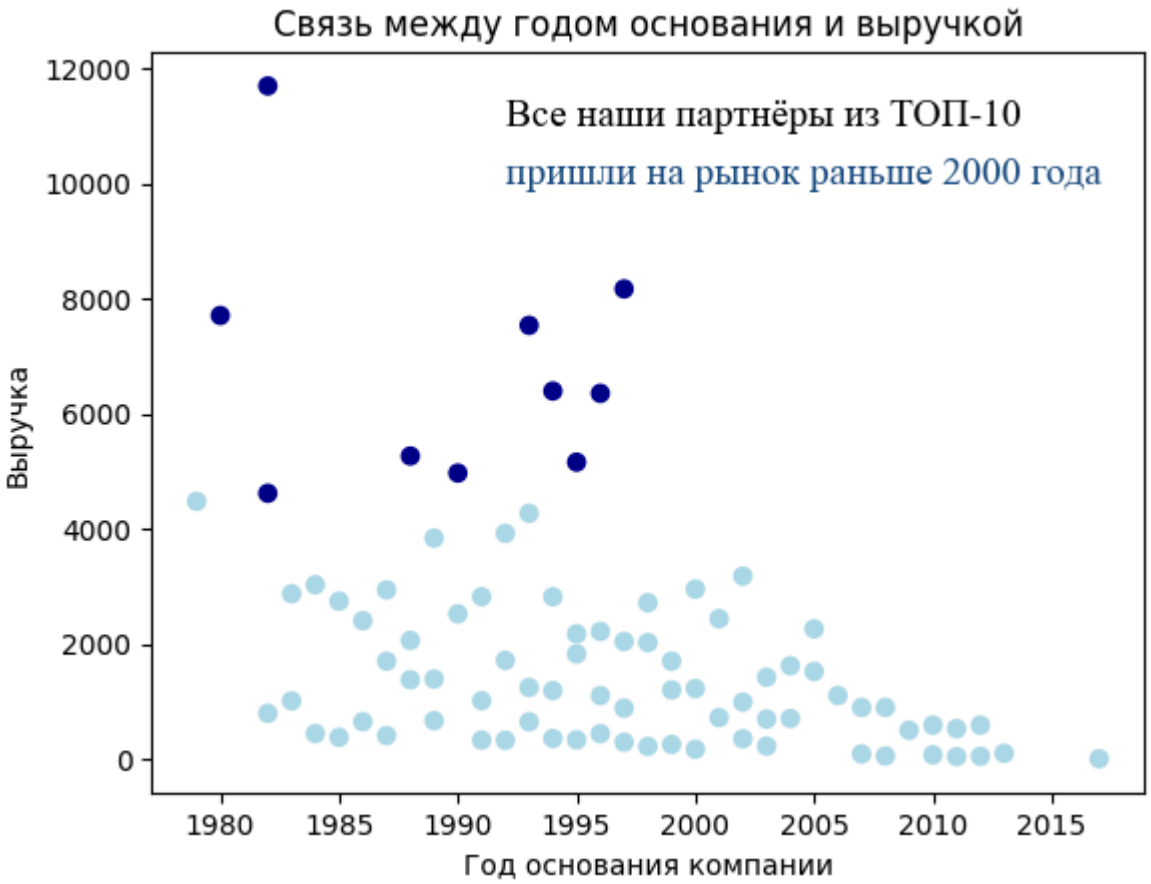
- Для показа связи между годом основания и выручкой воспользуемся графиком распределения
- Согласно подходу *Data to Viz* необходимо:
  - Избегать перенасыщения точками
  - Выделять подгруппы.

In [43]:
```
revenue = df['revenue']
colors = ['lightblue' if rev <= 4617 else 'darkblue' for rev in revenue]

plt.scatter(df['year_established'], revenue, c=colors)
plt.xlabel('Год основания компании')
plt.ylabel('Выручка')
plt.title('Связь между годом основания и выручкой')

# Добавление объясняющего текста
plt.text(1992, 11000, 'Все наши партнёры из ТОП-10', fontsize=14, fontname='Times N
plt.text(1992, 10000, 'пришли на рынок раньше 2000 года', fontsize=14, fontname='Ti

plt.show()
```



Итоги истории

- Таким образом, в графиках мы рассмотрели наших клиентов.
- Как оказалось, ключевыми секторами с которыми мы взаимодействуем являются продажи, технологи и медицина
- При этом, наиболее выгодными и перспективными в плане объёма проектов являются компании, работающие в сеторах: разработки, телекоммуникаций, технологий
- Ключевыми нашими партнёрами являются: Kan-code, Hottechi, Konex, Xx-holding
- Все наши партнёры из ТОП-10 пришли на рынок раньше 2000 года