Рынок заведений общественного питания Москвы

Описание проекта

Инвесторы из фонда «Shut Up and Take My Money» решили попробовать себя в новой области и открыть заведение общественного питания в Москве. Заказчики ещё не знают, что это будет за место: кафе, ресторан, пиццерия, паб или бар, — и какими будут расположение, меню и цены.

Для начала они просят вас — аналитика — подготовить исследование рынка Москвы, найти интересные особенности и презентовать полученные результаты, которые в будущем помогут в выборе подходящего инвесторам места.

Вам доступен датасет с заведениями общественного питания Москвы, составленный на основе данных сервисов Яндекс Карты и Яндекс Бизнес на лето 2022 года. Информация, размещённая в сервисе Яндекс Бизнес, могла быть добавлена пользователями или найдена в общедоступных источниках. Она носит исключительно справочный характер.

Основателям фонда «Shut Up and Take My Money» не даёт покоя успех сериала «Друзья». Их мечта — открыть такую же крутую и доступную, как «Central Perk», кофейню в Москве. Будем считать, что заказчики не боятся конкуренции в этой сфере, ведь кофеен в больших городах уже достаточно. Попробуйте определить, осуществима ли мечта клиентов.

Цель проекта: Исследовать рынок заведений общественного питания в Москве на предмет выявления особенностей работы заведений, их месторасположений, средних рейтингов и чеков, выявление взаимосвязи и влияния друг на друга, с целью учесть в последующем анализе "позитивный" и "негативный" опыт конкурентов при открытии собственного заведения.

План проекта:

1. Загрузка данных, получение общей информации и предобработка данных, будет рассмотрено в рамках раздела:
    - структуру представленных данных,
    - корректность названия столбцов,
    - размер датафрейма,
    - проверка на наличие пропусков,
    - анализ типов данных в датафрейме,
    - проверка на дубликаты в датафрейме,
    - добавление новых столбцов (при необходимости)
    - вывод по разделу.
2. Анализ рынка заведений общественного питания Москвы:

    2.1. Исследование объектов общественного питания по категориям

    2.2. Исследование количества посадочных мест в заведениях по категориям

    2.3. Анализ ТОП-15 популярных сетей в Москве

    2.4. Анализ административных районов Москвы

    2.5. Анализ распределения средних рейтингов

    2.6. Анализ распределения количества заведений и их категорий по улицам

    2.7. Анализ средних чеков заведений в рамках каждого района Москвы и построение фоновой картограммы (на основе полученных данных) с целью определения зависимости цены в заведениях от удаленности от центра

    2.8. Общий вывод по разделу

3. Открытие кофейни по мотивам сериала «Друзья»:

    3.1. Анализ распределения кофеен по регионам

    3.2. Анализ среднего рейтинга кофеен по регионам

    3.3. Анализ средней стоимости кружки капучино по регионам

    3.4. Анализ среднего чека кофеен по регионам

    3.5. Формулировка общего вывода по разделу.

Презентация: https://disk.yandex.ru/d/5zW-oMKTulCXdA

1. Загрузка данных и подготовка их к анализу

In [1]:
```
# импортируем библиотеки
import pandas as pd # импортируем pandas
from matplotlib import pyplot as plt # импортируем matplotlib
import plotly.express as px # для создания более сложныз графиков
from plotly import graph_objects as go # для воронки и круговой диаграммы
from plotly.subplots import make_subplots # для того, чтобы графики были
import seaborn as sns # импортируем seaborn
import json # подключаем модуль для работы с JSON-форматом
import missingno as msno
```

In [3]:
```
import folium
from folium import Marker, Map, Choropleth # импортируем маркер, карту и
from folium.plugins import MarkerCluster # импортируем кластер
from folium.features import CustomIcon # импортируем собственные иконки
```

In [6]:
```
df.head()
```

Out[6]:

	name	category	address	district	hours	lat	
0	WoWфли	кафе	Москва, улица Дыбенко, 7/1	Северный административный округ	ежедневно, 10:00–22:00	55.878494	37.4
1	Четыре комнаты	ресторан	Москва, улица Дыбенко, 36, корп. 1	Северный административный округ	ежедневно, 10:00–22:00	55.875801	37.4
2	Хазри	кафе	Москва, Клязьминская улица, 15	Северный административный округ	пн-чт 11:00–02:00; пт,сб 11:00–05:00; вс 11:00...	55.889146	37.5
3	Dormouse Coffee Shop	кофейня	Москва, улица Маршала Федоренко, 12	Северный административный округ	ежедневно, 09:00–22:00	55.881608	37.4
4	Иль Марко	пиццерия	Москва, Правобережная улица, 1Б	Северный административный округ	ежедневно, 10:00–22:00	55.881166	37.4

Структура данных:

- **name** — название заведения;
- **address** — адрес заведения;
- **category** — категория заведения, например «кафе», «пиццерия» или «кофейня»;
- **hours** — информация о днях и часах работы;
- **lat** — широта географической точки, в которой находится заведение;
- **lng** — долгота географической точки, в которой находится заведение;
- **rating** — рейтинг заведения по оценкам пользователей в Яндекс Картах (высшая оценка — 5.0);
- **price** — категория цен в заведении, например «средние», «ниже среднего», «выше среднего» и так далее;
- **avg_bill**** — строка, которая хранит среднюю стоимость заказа в виде диапазона, например:
  - «Средний счёт: 1000–1500 ₽»;
  - «Цена чашки капучино: 130–220 ₽»;
  - «Цена бокала пива: 400–600 ₽».
  - и так далее;
- **middle_avg_bill** — число с оценкой среднего чека, которое указано только для значений из столбца avg_bill, начинающихся с подстроки «Средний счёт»:
  - Если в строке указан ценовой диапазон из двух значений, в столбец войдёт медиана этих двух значений.
  - Если в строке указано одно число — цена без диапазона, то в столбец войдёт это число.
  - Если значения нет или оно не начинается с подстроки «Средний счёт», то в столбец ничего не войдёт.
- **middle_coffee_cup** — число с оценкой одной чашки капучино, которое указано только для значений из столбца avg_bill, начинающихся с подстроки «Цена одной чашки капучино»:
  - Если в строке указан ценовой диапазон из двух значений, в столбец войдёт медиана этих двух значений.
  - Если в строке указано одно число — цена без диапазона, то в столбец войдёт это число.
  - Если значения нет или оно не начинается с подстроки «Цена одной чашки капучино», то в столбец ничего не войдёт.
- **chain** — число, выраженное 0 или 1, которое показывает, является ли заведение сетевым (для маленьких сетей могут встречаться ошибки):
  - 0 — заведение не является сетевым,
  - 1 — заведение является сетевым.
- **district** — административный район, в котором находится заведение, например Центральный административный округ;
- **seats** — количество посадочных мест.

In [7]:
```
print(
    'Датафрейм с информацией о заведениях общественного питания Москвы со
    .format(df.shape[0], df.shape[1])
)
```

```
Датафрейм с информацией о заведениях общественного питания Москвы состои
т из 8406 строк и 14 столбцов
```

In [8]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8406 entries, 0 to 8405
Data columns (total 14 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   name               8406 non-null   object 
 1   category           8406 non-null   object 
 2   address            8406 non-null   object 
 3   district           8406 non-null   object 
 4   hours              7870 non-null   object 
 5   lat                8406 non-null   float64
 6   lng                8406 non-null   float64
 7   rating             8406 non-null   float64
 8   price              3315 non-null   object 
 9   avg_bill           3816 non-null   object 
 10  middle_avg_bill    3149 non-null   float64
 11  middle_coffee_cup  535 non-null    float64
 12  chain              8406 non-null   int64  
 13  seats              4795 non-null   float64
dtypes: float64(6), int64(1), object(7)
memory usage: 919.5+ KB
```

При анализе типов данных хранящихся в каждом столбце видно, что:

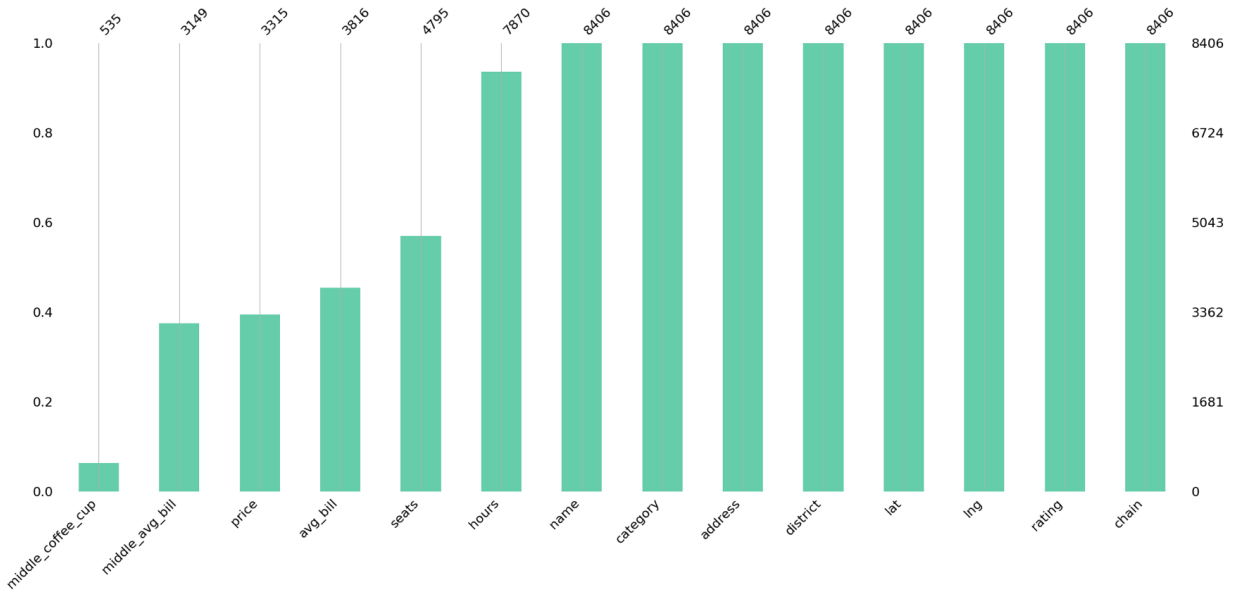
1. Количество мест в заведении хранится в типе float64 -дробного числа. Переведем значения данного столбца в целочисленное значение.
2. Значения в столбце принадлежности к сети представелены в виде целого числа, заменим на значения bool (true/false)

In [9]:
```
df = df.astype({'seats': 'Int64'})
```

In [10]:
```
def is_chain(value):
    if value == 1:
        return True
    elif value == 0:
        return False
```

In [11]:
```
df['chain'] = df['chain'].apply(is_chain)
```

In [12]:
```
# для удобства визуально отобразим количество пропусков по каждому столбц
msno.bar(df, fontsize=16, sort='ascending', color='mediumaquamarine')
plt.grid()
plt.show()

(
    pd.DataFrame(round(df.isna().mean()*100, 2))
    .sort_values(by=0)
    .style.background_gradient('BuGn')
)
```



Out[12]:

	0
name	0.000000
category	0.000000
address	0.000000
district	0.000000
lat	0.000000
lng	0.000000
rating	0.000000
chain	0.000000
hours	6.380000
seats	42.960000
avg_bill	54.600000
price	60.560000
middle_avg_bill	62.540000
middle_coffee_cup	93.640000

Из представленной информации можно сделать следующие выводы:

- имеется существенное количество пропущенных значений в столбцах:
  - middle_coffee_cup - 93,6% пропуков,
  - middle_avg_bill - 62,5% пропусков,
  - price - 60,5% пропусков,
  - avg_bill - 54,6% пропусков,
  - seats - 42,9% пропусков.
- имеется небольшое количество (535 строк или 6,37%) пропущенных значений в столбце hours.

Обработать пропуски в столбцах middle_coffee_cup, middle_avg_bill, price, avg_bill, seats не представляется возможным в связи с большим количеством отсутствующей информации и на столь малых доступных данных заполнение пропусками средним/медианой будет не корректно, что исказит дальнейший анализ.

Посмотрим детальнее на заведения с меньшим процентом пропусков (в столбце hours), проанализируем возможно ли заполнить данные.

In [13]:
```
df[df['hours'].isna()]
```

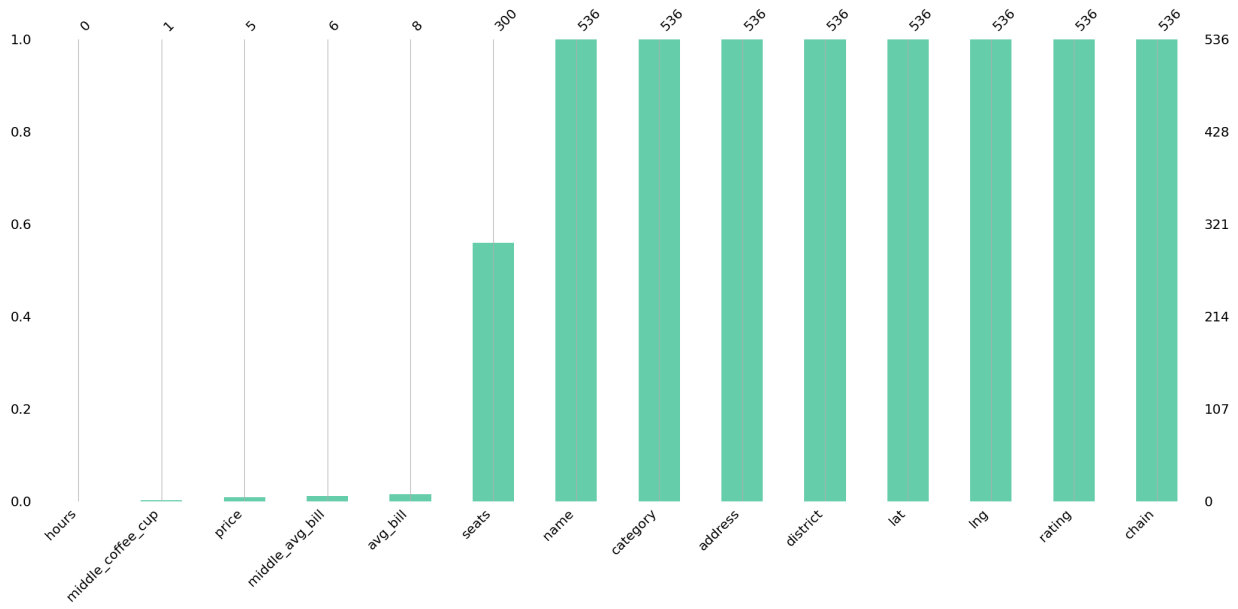
Out[13]:

	name	category	address	district	hours	lat	
38	Ижора	булочная	Москва, Ижорский проезд, 5А	Северный административный округ	NaN	55.888366	3
40	Кафе	кафе	Москва, Ижорская улица, 18, стр. 1	Северный административный округ	NaN	55.895115	3
44	Кафетерий	кафе	Москва, Ангарская улица, 24А	Северный административный округ	NaN	55.876289	3
56	Рыба из тандыра	быстрое питание	Москва, Коровинское шоссе, 46, стр. 5	Северный административный округ	NaN	55.888010	3
108	Кафе	бар,паб	Москва, МКАД, 82-й километр, вл18	Северо-Восточный административный округ	NaN	55.908930	3
...	...	...	...	...	...	...	
8236	1у	кафе	Москва, Нагатинская набережная, 40/1к1	Южный административный округ	NaN	55.685528	3
8375	Улица Гурьянова 55	кафе	Москва, улица Гурьянова, 55	Юго-Восточный административный округ	NaN	55.679981	3
8378	Восточно-грузинская кухня	быстрое питание	Москва, Зеленодольская улица, 32, корп. 3	Юго-Восточный административный округ	NaN	55.710540	3
8381	Аэлита	кафе	Москва, Ферганская улица, 8, корп. 2, стр. 1	Юго-Восточный административный округ	NaN	55.708871	3
8395	Истира Запрафка	кафе	Москва, Юго-Восточный административный округ, ...	Юго-Восточный административный округ	NaN	55.724686	3

536 rows × 14 columns

In [14]:

```
msno.bar(df[df['hours'].isna()], fontsize=16, sort='ascending', color='me
plt.grid()
plt.show()
```



При просмотре первых и последних 5 строк в срезе по заведениям, где отсутствуют данные о часах работы видно, что помимо этого столбца так же отсутствуют значения и в других столбцах.

Не исключено, что при выгрузке данных мог произойти сбой при котором данные выгрузились не полностью, поэтому в реальном кейсе необходимо уточнять этот момент и при возможности исправлять, т.к. исследование может быть не корректным.

Теоретически и в крайних случаях можно поискать информацию в интернете о представленных заведениях и проставить информацию вручную, но это займет время, как на поиск и заполнение информации, а также может вызвать сомнения в качестве выложенных данных. В этом случае необходимо определить мнению какого ресурса мы можем и готовы доверять.

Перед тем как осуществить проверку на полные дубликаты приведем названия заведений и улицу к нижнему регистру, заменим ё на е и уберем пробелы.

In [15]:
```
df['name'] = (
    df['name']
    .str.lower()
    .str.strip()
    .str.replace('ё', 'е')
)
df['address'] = (
    df['address']
    .str.lower()
    .str.strip()
    .str.replace('ё', 'е')
)
```

In [16]:
```
print(
    ' Количество полных дубликатов в датафрейме – {}'.format(df.duplicate
    '\n',
    'Количество дубликатов по названию и адресу заведения – {}'
    .format(df[['name', 'address']].duplicated().sum()),
    '\n',
    'Количество дубликатов по названию заведения, категории, адресу, широ
    .format(df[['name', 'address', 'category', 'lat', 'lng']].duplicated(
)
```

```
 Количество полных дубликатов в датафрейме – 0
 Количество дубликатов по названию и адресу заведения – 4
 Количество дубликатов по названию заведения, категории, адресу, широте
и долготе  – 1
```

Определим, дублирующие заведения и посмотрим на них детальнее.

In [17]:
```
df[df.duplicated(['name', 'address'])]
```

Out[17]:

	name	category	address	district	hours	lat
215	кафе	кафе	москва, парк ангарские пруды	Северный административный округ	ежедневно, 10:00–22:00	55.881438
1511	more poke	ресторан	москва, волоколамское шоссе, 11, стр. 2	Северный административный округ	пн-чт 09:00–18:00; пт,сб 09:00–21:00; вс 09:00...	55.806307
2420	раковарня клешни и хвосты	бар,паб	москва, проспект мира, 118	Северо-Восточный административный округ	пн-чт 12:00–00:00; пт,сб 12:00–01:00; вс 12:00...	55.810677
3109	хлеб да выпечка	кафе	москва, ярцевская улица, 19	Западный административный округ	NaN	55.738449

In [18]:
```
df[df.duplicated(['name', 'address', 'category', 'lat', 'lng'])]
```

Out[18]:

	name	category	address	district	hours	lat	lng
1511	more poke	ресторан	москва, волоколамское шоссе, 11, стр. 2	Северный административный округ	пн-чт 09:00–18:00; пт,сб 09:00–21:00; вс 09:00...	55.806307	37.497566

Дубликаты обнаружены по названию и адресу в 4-х заведениях:

- "кафе" по адресу москва, парк ангарские пруды,
- "more poke" по адресу москва, волоколамское шоссе, 11, стр. 2,
- "раковарня клешни и хвосты" по адресу москва, проспект мира, 118,
- "хлеб да выпечка" по адресу москва, ярцевская улица, 19.

И один дубликат ключевым опознавательным знакам (название, категория, адрес, широта и долгота) - "more poke" по адресу москва, волоколамское шоссе, 11, стр. 2

Рассмотрим дубликаты детальнее, чтобы принять решение об удалении или не удалении их из датафрейма.

In [19]:
```
df.query('name == "раковарня клешни и хвосты"')
```

Out[19]:

	name	category	address	district	hours	lat	
2211	раковарня клешни и хвосты	ресторан	москва, проспект мира, 118	Северо-Восточный административный округ	ежедневно, 12:00–00:00	55.810553	
2420	раковарня клешни и хвосты	бар,паб	москва, проспект мира, 118	Северо-Восточный административный округ	пн-чт 12:00–00:00; пт,сб 12:00–01:00; вс 12:00...	55.810677	3
7270	раковарня клешни и хвосты	бар,паб	москва, братиславская улица, 12	Юго-Восточный административный округ	пн-чт 12:00–00:00; пт,сб 12:00–01:00; вс 12:00...	55.659744	3

При детальном рассмотрении заведения "раковарня клешни и хвосты" имеется дубликат в названии и адресе, но важно отметить, что эти 2 заведения имеют разную категорию (ресторан и бар/паб) и широту и долготу местонахождения, при этом ресторан не относится к сети, а бар/паб имеет пренадлежность к сети.

Такая особенность могла возникнуть по следубщим причинам:

- у данного заведения основная категория бар/паб, при этом на основе одного заведения могло произойти расширение бизнеса и дополнительное создание ресторна с отдельным входом/помещением об этом нам говорит разница между значениями широта и долгота местонахождения, а так же бар/паб открыт дольше с пт по вс включительно.
- различие. значениях в принадлежности к сети могли возникнуть в связи с тем, что бар/паб мог отрыться по франшизе, а расширение бизнеса (создание другой категории заведения) на территории (пусть и под тем же брендом) может не входить в эту сеть. Все нюансы на основе договоренностей руководства заведения и франчази. Изменение этих данных будет не совсем корректно в связи с отсутствием информации об условиях сотрудничества.

Соответственно, удаление этого дубликата будет не совсем корректно, так как, вероятно, это 2 разных заведения.

In [20]:
```
df.query('name == "more poke"')
```

Out[20]:

	name	category	address	district	hours	lat	
1430	more poke	ресторан	москва, волоколамское шоссе, 11, стр. 2	Северный административный округ	ежедневно, 09:00–21:00	55.806307	37.49
1511	more poke	ресторан	москва, волоколамское шоссе, 11, стр. 2	Северный административный округ	пн-чт 09:00–18:00; пт,сб 09:00–21:00; вс 09:00...	55.806307	37.49
6088	more poke	ресторан	москва, духовской переулок, 19	Южный административный округ	ежедневно, 10:00–22:00	55.704177	37.61

При датальном рассмотрении заведения "more poke" видно, что имеется полное соответствие в названии, категории, адресе и широте и долготе местонахождения. При этом обнаружено так же различие в принаждежности к сети, это могло возникнуть в связи с ошибками в выгрузке или при заполнении файла.

Принимаю решение удалить данный дубликат с принадлежностью к сети - False.

In [21]:
```
df = df.drop(index=1430)
```

In [22]:
```
df.query('name == "кафе" and address == "москва, парк ангарские пруды"')
```

Out[22]:

	name	category	address	district	hours	lat	lng
189	кафе	кафе	москва, парк ангарские пруды	Северный административный округ	ежедневно, 09:00–23:00	55.880327	37.530786
215	кафе	кафе	москва, парк ангарские пруды	Северный административный округ	ежедневно, 10:00–22:00	55.881438	37.531848

При рассмотрении дубликата заведения "кафе" обнаружено, что:

- адрес заведения указан не совсем точно, располагается где-то в парке
- для того, что определить, где в парке необходимо ориентироваться на широту и долготу - они различаются,
- время работы так же раличается,
- оба заведения не являются сетью.

Считаю удаление данного дубликата не корректным, скорее всего это 2 разных заведения.

In [23]:
```
df.query('name == "хлеб да выпечка"')
```

Out[23]:

	name	category	address	district	hours	lat	
3091	хлеб да выпечка	булочная	москва, ярцевская улица, 19	Западный административный округ	ежедневно, 09:00–22:00	55.738886	37.4116
3109	хлеб да выпечка	кафе	москва, ярцевская улица, 19	Западный административный округ	NaN	55.738449	37.4109
7937	хлеб да выпечка	кофейня	москва, каширское шоссе, 61г	Южный административный округ	ежедневно, 09:00–22:00	55.621379	37.7141

Ситуация очень схожа с заведением "раковарня клешни и хвосты":

- категории различаются у дубликатов,
- широта и долгота местонахождения так же.

Вероятно на основе кафе была создана булочная, с отдельным входом (соединяющаяся в кафе) и клиент при посещении может присесть в кафе. Рвзличия в принадлежности к сети могли возникнуть по тем же причинам что и у "раковарня клешни и хвосты". Изменять принадлежность к сети не буду в связи с отсутствием полной уверенности в предоставленных данных.

In [24]:
```
print(
    ' Всего заведений датафрейме – {}'
    .format(len(df['name'])),
    '\n',
    'Количество уникальных названий заведений в датафрейме – {}'
    .format(len(df['name'].unique()))
)
```

```
 Всего заведений датафрейме – 8405
 Количество уникальных названий заведений в датафрейме – 5506
```

Создадим столбец street с названиями улиц из столбца с адресом.

In [25]:
```
df['address'][0].split(sep=',')[1].strip()
```

Out[25]:
```
'улица дыбенко'
```

In [26]:
```
df['street'] = df['address'].apply(lambda x: x.split(sep=',')[1].strip())
```

Создадим столбец is_24/7 с обозначением, что заведение работает ежедневно и круглосуточно (24/7):

- логическое значение True — если заведение работает ежедневно и круглосуточно;
- логическое значение False — в противоположном случае.

In [27]:
```
df['is_24/7'] = df['hours'].str.contains('ежедневно, круглосуточно')
```

In [28]:
```
df.head()
```

Out[28]:

	name	category	address	district	hours	lat	
0	wowфли	кафе	москва, улица дыбенко, 7/1	Северный административный округ	ежедневно, 10:00–22:00	55.878494	37.4
1	четыре комнаты	ресторан	москва, улица дыбенко, 36, корп. 1	Северный административный округ	ежедневно, 10:00–22:00	55.875801	37.4
2	хазри	кафе	москва, клязьминская улица, 15	Северный административный округ	пн-чт 11:00–02:00; пт,сб 11:00–05:00; вс 11:00...	55.889146	37.5
3	dormouse coffee shop	кофейня	москва, улица маршала федоренко, 12	Северный административный округ	ежедневно, 09:00–22:00	55.881608	37.4
4	иль марко	пиццерия	москва, правобережная улица, 1б	Северный административный округ	ежедневно, 10:00–22:00	55.881166	37.4

In [29]:
```
print(
    ' Размер датафрейма после предобработки: {} строк, {} столбцов'
    .format(df.shape[0], df.shape[1]),
    '\n',
    'Всего заведений датафрейме – {}'
    .format(len(df['name'])),
    '\n',
    'Количество уникальных названий заведений в датафрейме – {}'
    .format(len(df['name'].unique()))
)
```

```
Размер датафрейма после предобработки: 8405 строк, 16 столбцов
Всего заведений датафрейме – 8405
Количество уникальных названий заведений в датафрейме – 5506
```

Общий вывод по разделу:

1. Предоставлен датафрейм с информацией о заведениях общественного питания Москвы состоящий из 8406 строк и 14 столбцов,

2. Информация о пропусках:

   - имеется существенное количество пропущенных значений в столбцах:
     - middle_coffee_cup - 93,6% пропуков,
     - middle_avg_bill - 62,5% пропусков,
     - price - 60,5% пропусков,
     - avg_bill - 54,6% пропусков,
     - seats - 42,9% пропусков.
   - имеется небольшое количество (535 строк или 6,37%) пропущенных значений в столбце hours.

3. Пропуски не были заменены на какие-либо значения, чтобы не искозить

дальнейшее исследование, т.к. слишком большой % отсутствующих данных.

Не исключено, что при выгрузке данных мог произойти сбой при котором данные выгрузились не полностью, поэтому в реальном кейсе необходимо уточнять этот момент и при возможности исправлять, т.к. исследование может быть не корректным.

Теоретически и в крайних случаях можно поискать информацию в интернете о представленных заведениях и проставить информацию вручную, но это займет время, как на поиск и заполнение информации, а также может вызвать сомнения в качестве выложенных данных. В этом случае необходимо определить мнению какого ресурса мы можем и готовы доверять.

4. Для возможности качественного анализа данных было осуществлено:
    - приведены названия заведений и улица к нижнему регистру, а также заменено ё на е, убераны пробелы,
    - количество мест в заведении хранилось в виде дробного числа, изменено на целочисленное значение,
    - значения в столбце принадлежности к сети представелены в виде целого числа, замено на значения типа bool (true/false)
5. Информация о дубликатах:
    - Количество полных дубликатов в датафрейме - 0
    - Количество дубликатов по названию и адресу заведения - 4
    - Количество дубликатов по названию заведения, категории, адресу, широте и долготе - 1
6. Удалено дубликатов - 1 строка. У заведения "more poke" было выявлено, что имеется полное соответствие в названии, категории, адресе и широте и долготе местонахождения. При этом обнаружено так же различие в принаждежности к сети, это могло возникнуть в связи с ошибками в выгрузке или при заполнении файла. Остальные дубликаты заведений не были удалены в связи с тем, что заведения имеют разную категорию, широту/долготу местонахождения и принаждежность к сети.
7. Всего заведений датафрейме - 8405. Количество уникальных названий заведений в датафрейме - 5506.
8. Были созданы 2 дополнительных столбца:
    - s_24/7 с обозначением, что заведение работает ежедневно и круглосуточно (24/7),
    - street с названиями улиц из столбца с адресом.
9. Размер датафрейма после предобработки - 8405 строк, 16 столбцов.
    - Всего заведений датафрейме после предобработки - 8405.
    - Количество уникальных названий заведений в датафрейме после

предобработки - 5506

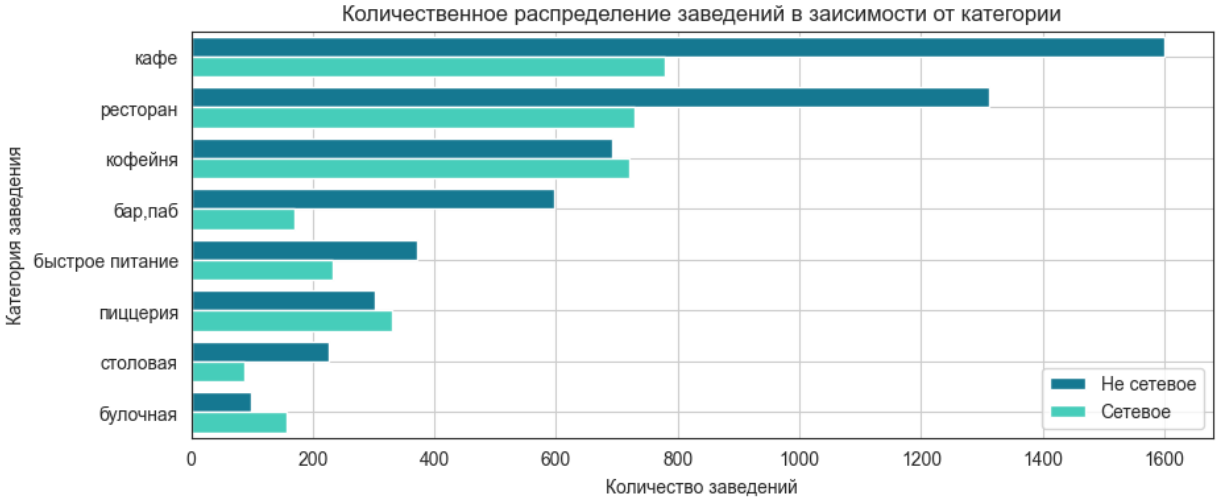
2. Анализ рынка заведений общественного питания Москвы

2.1. Исследование объектов общественного питания по категориям

Проанализируем распределение категорий заведений с целью определения наиболее популярных категории у сетевых и не сетевых организаций, а так же соотношение сетевых и не сетевых категорий заведений в датасете.

Рассмотрим количество заведений как в количественных значениях так и относительных.

In [30]:
```
df_category = (
    df
    .replace({'chain' : {True : 'Сетевое', False: 'Не сетевое'}})
    .pivot_table(index=['category', 'chain'], values='name', aggfunc='cou
    .rename(columns={'name':'total'})
    .sort_values(by='total', ascending=False)
    .reset_index()
)

sns.set_style('white')
plt.figure(figsize=(10, 4))

sns.barplot(x='total', y='category', data=df_category, hue='chain', palet

plt.title('Количественное распределение заведений в заисимости от категор
plt.xlabel('Количество заведений')
plt.ylabel('Категория заведения')
plt.legend(loc='lower right', fontsize=10)
plt.grid()
plt.show()
```



Как видно из диаграммы, наибольшее количество приходится на категорию "кафе" как у сетевых, так и у не сетевых организаций, при этом не сетевых заведений более чем в 2 раза больше чем сетевых.

На втором месте по популярности открытия - рестораны, при этом не сетевые рестораны превосходят в количестве сетевых почти в 2 раза.

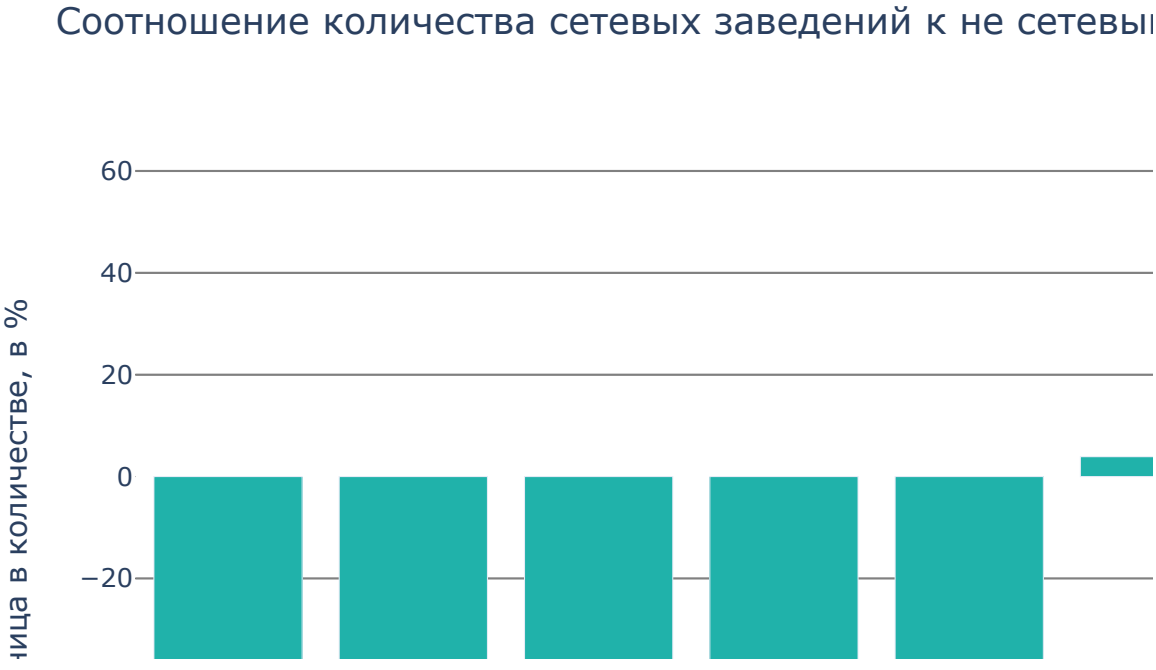
На третьем месте по популярности - кофейни, стоит отметить, то у каждой категории количество сетевых заведений немного превосходит количество не сетевых.

А вот четвертое место по количеству заведений у сетевых приходится на пиццерии, а у не сетевых заведений на бар/паб.

Проанализируем доли заведений в сетевых и не сетевых организаций.

In [31]:
```
df_category = (
    df
    .pivot_table(index='category', columns='chain', values='name', aggfun
    .rename(columns={True: 'Сеть', False: 'НЕ сеть'})
)
df_category['delta'] = round(((df_category['Сеть'] / df_category['НЕ сеть
df_category['total'] = df_category['Сеть'] + df_category['НЕ сеть']
df_category = df_category.sort_values(by='total', ascending=False).reset_
```

In [32]:
```
fig = go.Figure()
fig.add_trace(go.Bar(x = df_category['category'],
                     y = df_category['delta'].sort_values(),
                     textposition='outside',marker=dict(color="LightSeaGr
fig.update_layout(legend_orientation="h",
                  title="Соотношение количества сетевых заведений к не се
                  yaxis_title="Разница в количестве, в %", plot_bgcolor="
                 xaxis_title="Категория заведения")
fig.update_yaxes(gridcolor='grey')
fig.show()
```



Исходя из представленных данных можно сделать следующие выводы:

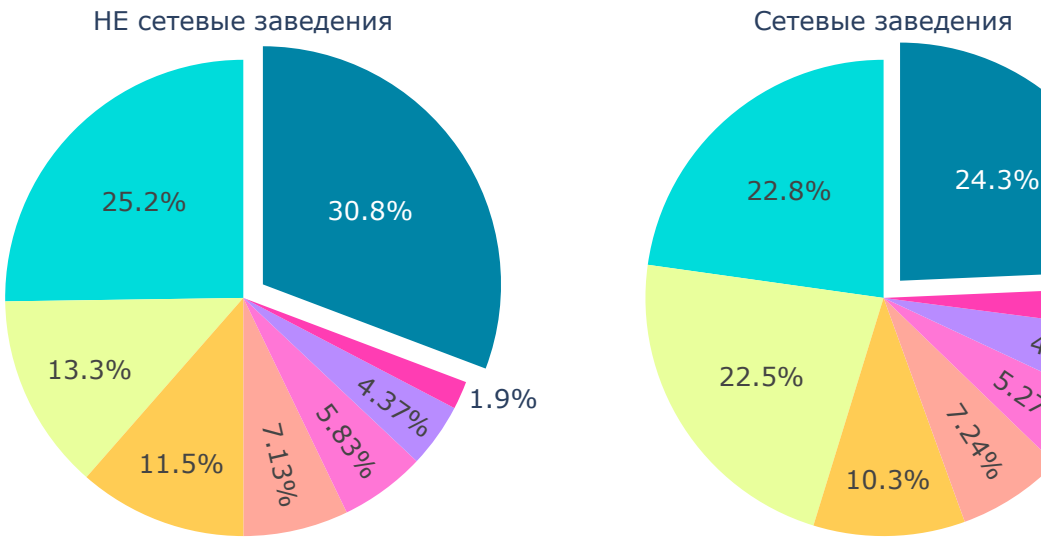
- у сетевых заведений имеется большее в количество (сетевые/не сетевые):
  - столовых разница около +60%,
  - ресторанов разница около +10%,
  - пиццерий разница около +5%
- у не сетевых заведений имеется большее в количество (не сетевые/ сетевые):
  - баров/пабов разница около +66%,
  - булочные разница около +70%,
  - быстрое питание разница около +50%,
  - кафе разница около +45%,
  - кофейни разница около +35%.

In [33]:

```
fig = make_subplots(rows=1, cols=2, specs=[[{"type": "pie"}, {"type": "pi

colors = ['#0084a6', '#00dcdb', '#e9ff9c', '#ffcc54', '#ffa89b', '#ff76d6

fig.add_trace(go.Pie(labels=df_category['category'],
                             values=df_category['НЕ сеть'].sort_values(asc
                             pull = [0.1, 0],
                             title='НЕ сетевые заведения'), row=1, col=1)

fig.add_trace(go.Pie(labels=df_category['category'],
                              values=df_category['Сеть'].sort_values(ascen
                              pull = [0.1, 0],
                             title='Сетевые заведения'), row=1, col=2)
fig.update_layout(title='Распределение объектов общественного питания по
                  width=800,
                  height=600,
                  annotations=[dict(x=1.12,
                                    y=1.05,
                                    text='Категория заведения',
                                    showarrow=False)])

fig.update_traces(marker=dict(colors=colors))
fig.show()
```

Распределение объектов общественного питания по категор

Кат



При анализе доли категорий заведений видно, что как у сетевых, так и у не сетевых организаций, имеется схожая тенденция и можно проранжировать категории от наиболее популярных до менее популярных.

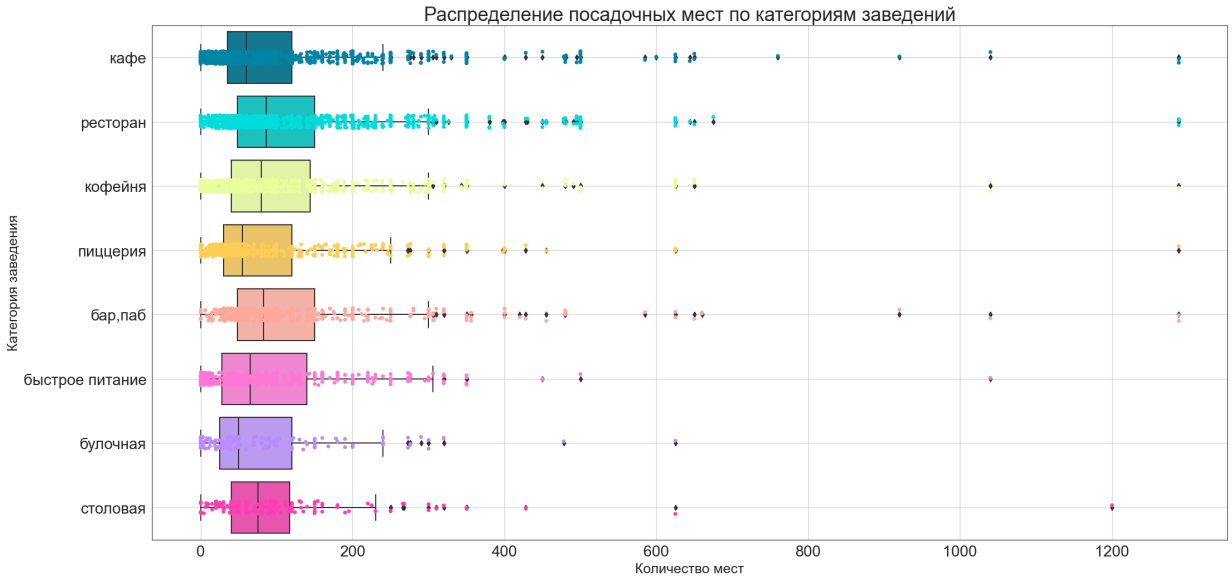
Проранжируем доли заведений в порядке убывания:

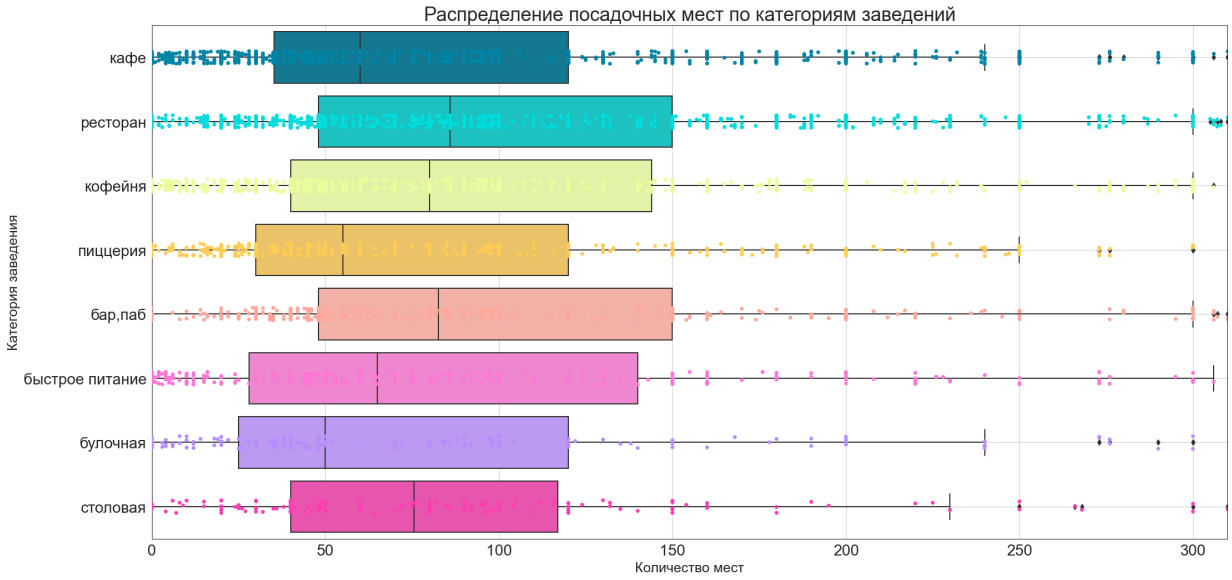
1. Кафе - занимают наибольшую долю среди всех заведений, на них приходится 30,8% у сетевых заведений, 24,3% - не сетевых.
2. Ресторан - доля среди не сетевых - 25,2%, сетевых - 22,8%,
3. Кофейня - доля среди не сетевых - 13,3%, сетевых - 22,5%,
4. Паб/бар - доля среди не сетевых - 11,5%, сетевых - 10,3%,
5. Пиццерия - доля среди не сетевых - 7,13%, сетевых - 7,24%,
6. Быстрое питание - доля среди не сетевых - 5,83%, сетевых - 5,27%,
7. Столовая - доля среди не сетевых - 4,37%, сетевых - 4,9%,
8. Булочная - доля среди не сетевых - 1,9%, сетевых - 2,75%.

2.2. Исследование количества посадочных мест в заведениях по категориям

Перед тем как проанализировать количество посадочных мест каждой категории заведения проверим наличие аномальных значений и выбросов.

In [34]:
```
fig, ax = plt.subplots(figsize=(25, 12))

ax = sns.boxplot(x=df['seats'].astype({'seats': 'float64'}), y=df['catego
ax.tick_params(rotation=0, labelsize=20)

ax = sns.stripplot(x=df['seats']
                   .astype({'seats': 'float64'}), y=df['category'], palet
plt.xlim()
plt.xlabel('Количество мест', size=18)
plt.ylabel('Категория заведения', size=18)
plt.title('Распределение посадочных мест по категориям заведений', size=2
plt.grid()

fig, ax = plt.subplots(figsize=(25, 12))

ax = sns.boxplot(x=df['seats'].astype({'seats': 'float64'}), y=df['catego
ax.tick_params(rotation=0, labelsize=20)

ax = sns.stripplot(x=df['seats'].astype({'seats': 'float64'}), y=df['cate
plt.xlim(0, 310)
plt.xlabel('Количество мест', size=18)
plt.ylabel('Категория заведения', size=18)
plt.title('Распределение посадочных мест по категориям заведений', size=2
plt.grid()
plt.show()
```

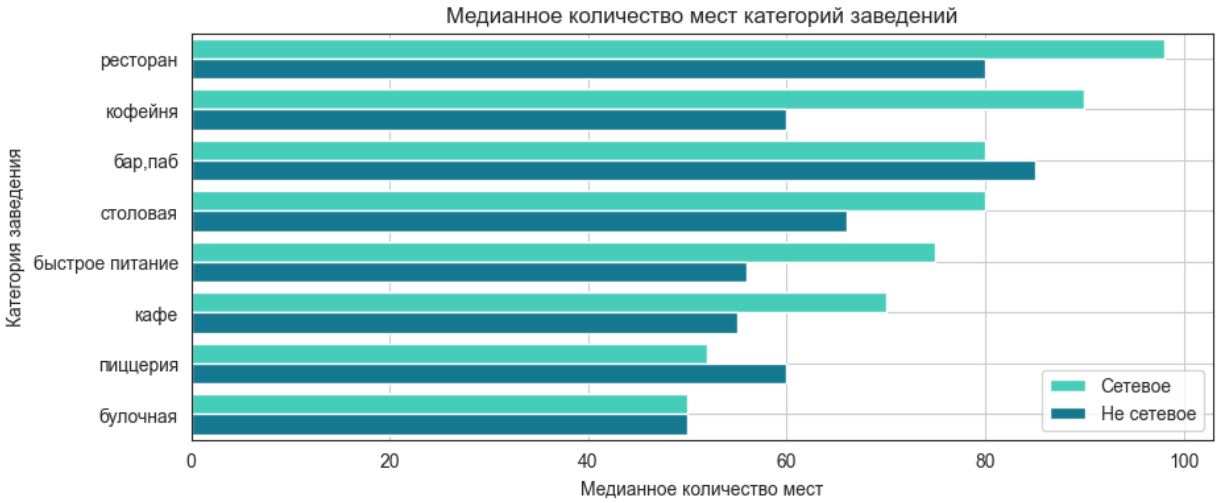




Из диаграммы ящика с усами и нанесенными на него значениями видно, что имеются выбросы и аномальные значения, соответственно, если рассчитывать среднее значение мест, то оно будет искажено, поэтому посмотрим медианное значение количество мест заведений.

In [35]:
```
df_seats = (
    df
    .replace({'chain' : {True : 'Сетевое', False: 'Не сетевое'}})
    .pivot_table(index=['category', 'chain'], values='seats', aggfunc='me
    .reset_index()
)
df_seats.columns = ['category', 'chain', 'median_seats']
df_seats = df_seats.sort_values('median_seats', ascending=False)
```

In [36]:
```
sns.set_style('white')
plt.figure(figsize=(10, 4))

sns.barplot(x='median_seats', y='category', data=df_seats, hue='chain', p

plt.title('Медианное количество мест категорий заведений')
plt.xlabel('Медианное количество мест')
plt.ylabel('Категория заведения')
plt.legend(loc='lower right', fontsize=10)
plt.grid()
plt.show()
```



В сетевых заведениях наибольшее количество мест у ресторанов и кофеен. Медианное количество мест среди категорий сетевых заведений:

- ресторан - чуть менее 100 мест,
- кофейня около 90 мест,
- в баре/пабе и столовой - 80 мест,
- заведение быстрого питания - 75 мест,
- кафе - 70 мест,
- пиццерия - 52 места,
- булочная - 50 мест.

У не сетевых заведений наибольшее количество мест у паб/бар и ресторанов. Медианное количество мест среди категорий сетевых заведений:

- в баре/пабе и столовой - 85 мест,
- ресторан - 80 мест,
- столовая - 66 мест,
- пиццерия и кофейня - по 60 мест,
- заведение быстрого питания - 56 мест,
- кафе - 55 мест,
- булочная - 50 мест.

Как видно из анализа, сетевые заведения обладают большим количеством мест по сравнению с не сетевыми заведениями, исключения составляют лишь пабы/бары и пиццерии у которых количество мест у не сетевых заведений больше.

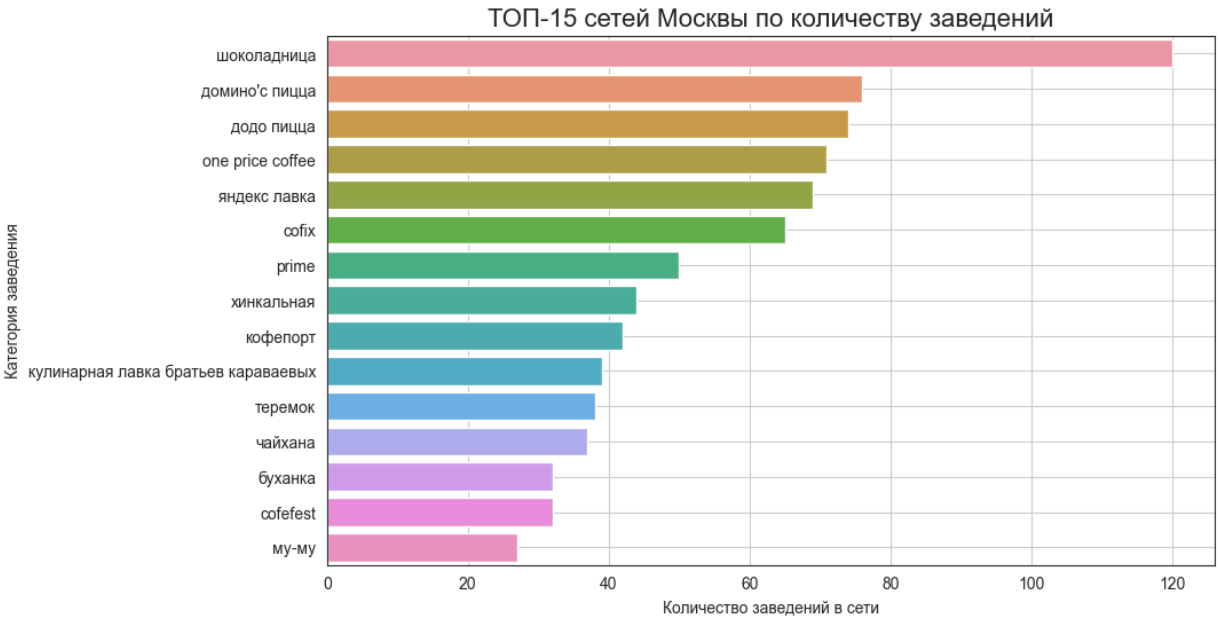
2.3. Анализ ТОП-15 популярных сетей в Москве

In [37]:
```
df_top = (
    df.query('chain == True')
    .pivot_table(index=['name'], values='chain', aggfunc='count')
    .sort_values('chain', ascending=False)[:15]
    .reset_index()
    .rename(columns={'name': 'name', 'chain': 'name_cnt'})
)
df_top
```

Out[37]:

	name	name_cnt
0	шоколадница	120
1	домино'с пицца	76
2	додо пицца	74
3	one price coffee	71
4	яндекс лавка	69
5	cofix	65
6	prime	50
7	хинкальная	44
8	кофепорт	42
9	кулинарная лавка братьев караваевых	39
10	теремок	38
11	чайхана	37
12	буханка	32
13	cofefest	32
14	му-му	27

In [38]:
```
sns.set_style('white')
plt.figure(figsize=(10, 6))

sns.barplot(x='name_cnt', y='name', data=df_top)

plt.title('ТОП-15 сетей Москвы по количеству заведений', size=16)
plt.xlabel('Количество заведений в сети')
plt.ylabel('Категория заведения')
plt.grid()
plt.show()
```



In [39]:
```
df_category_top = (
    df
    .query('name in @df_top["name"]')
    .pivot_table(index='name', columns='category', values='district', agg
)
df_category_top = df_category_top.fillna(0)
df_category_top['total'] = df_category_top.sum(axis=1)
df_category_top = df_category_top.sort_values(by='total').reset_index()
df_category_top = df_category_top.drop(columns = 'total')
df_category_top
```

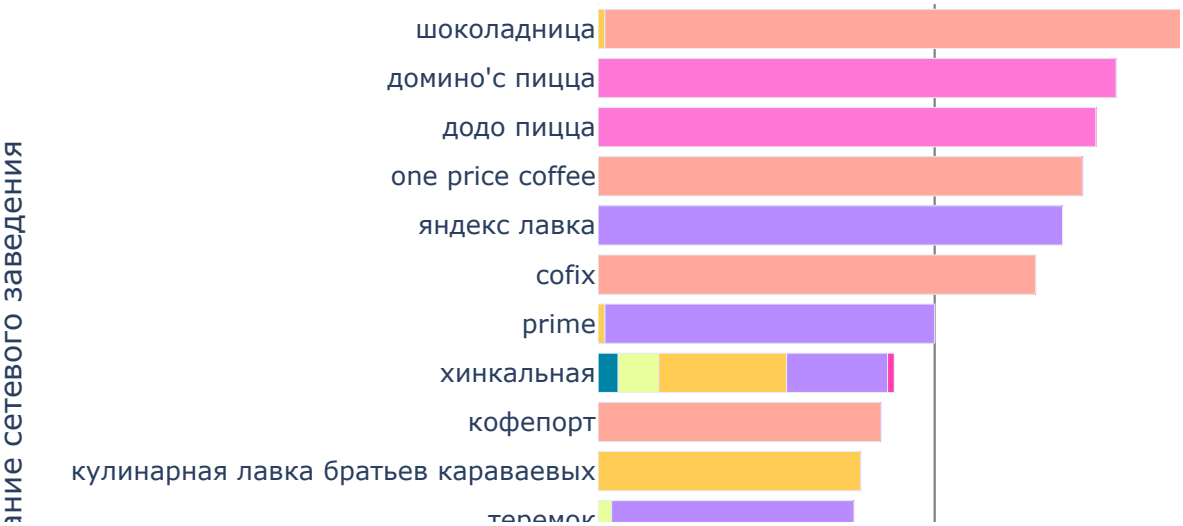
Out[39]:

category	name	бар,паб	булочная	быстрое питание	кафе	кофейня	пиццерия	рест
0	му-му	1.0	0.0	2.0	12.0	2.0	1.0	
1	cofefest	0.0	0.0	0.0	1.0	31.0	0.0	
2	буханка	0.0	25.0	0.0	1.0	6.0	0.0	
3	чайхана	0.0	0.0	2.0	26.0	0.0	0.0	
4	теремок	0.0	0.0	2.0	0.0	0.0	0.0	
5	кулинарная лавка братьев караваевых	0.0	0.0	0.0	39.0	0.0	0.0	
6	кофепорт	0.0	0.0	0.0	0.0	42.0	0.0	
7	хинкальная	3.0	0.0	6.0	19.0	0.0	0.0	
8	prime	0.0	0.0	0.0	1.0	0.0	0.0	
9	cofix	0.0	0.0	0.0	0.0	65.0	0.0	
10	яндекс лавка	0.0	0.0	0.0	0.0	0.0	0.0	
11	one price coffee	0.0	0.0	0.0	0.0	72.0	0.0	
12	додо пицца	0.0	0.0	0.0	0.0	0.0	74.0	
13	домино'с пицца	0.0	0.0	0.0	0.0	0.0	77.0	
14	шоколадница	0.0	0.0	0.0	1.0	119.0	0.0	

In [40]:

```
fig = px.bar(df_category_top, x=df_category_top.columns, y=df_category_to
       title='Категории заведений ТОП-15',
             color_discrete_sequence=colors)

fig.update_layout(yaxis_title='Название сетевого заведения', xaxis_title=
                  plot_bgcolor="white")

fig.update_xaxes(gridcolor='grey')
fig.show()
```

Категории заведений ТОП-15



Из представленной диаграммы видно, что в ТОП-15 входят почти все заведения, работающие по франшизе, исключение составляет cofefest по данной сети не было обнаружено наличие возможности приобрести франшизу для открытия.

При определении категорий заведений к которым относятся заведения сети было выявлено, что имеются ошибки в определении категории заведения это может быть вызвано как ошибкой при внесении информации, а так же особенностями работы и обслуживания в заведении.

Например, Кафе в сравнении с рестораном имеет меньший ассортимент блюд, при этом пиццерия — это тоже кафе, но главной позицией его меню является пицца.

Рейтинг заведений Москвы по количеству заведений:

1. Шоколадница - кофейня
2. Домино'с пицца - пиццерия
3. Додо пицца - пиццерия
4. One price coffee - кофейня,
5. Яндекс лавка - значится как ресторан, но не работающий как самостоятельное заведение, а осуществляющий доставку из других ресторанов потребителям,
6. Cofix - кофейня,
7. Prime - ресторан, но на оф. сайте организации обозначается как кафе,
8. Хинкальная - кафе / ресторан, вероятно имеется разница в ассортименте блюд и обслуживания в заведении,
9. Кофепорт - кофейня,
10. Кулинарная лавка братьев караваевых - кафе,
11. Теремок - ресторан / быстрое питание (вероятно аналог категории "ресторан" как у Макдоналдс),
12. Чайхана - кафе,
13. Буханка - булочная,
14. Cofefest - кофейня,
15. Му-му - кафе, кофейня

Наиболее популярными сетями заведений в Москве являются - кофейни и пиццерии.

2.4. Анализ административных районов Москвы

Перед тем как отобразить все районы Москвы и поместить на них заведения отобразим Москву.

Для этого ранее импортировали библиотеку и создадим объект, который будет хранить карту с центром в указанных координатах:

In [41]:
```
moscow_lat, moscow_lng = 55.751244, 37.618423

m = folium.Map(location=[moscow_lat, moscow_lng], zoom_start=10)
```

In [42]:
```
# загружаем JSON-файл с границами округов Москвы
try:
    state_geo = path + 'admin_level_geomap.geojson'
except:
    state_geo = 'https://code.s3.yandex.net/data-analyst/admin_level_geom
```

In [43]:
```
print('В датасете представлена информация о заведениях Москвы следующих р
      .format(list(df['district'].unique())))
```

```
В датасете представлена информация о заведениях Москвы следующих районов
['Северный административный округ', 'Северо-Восточный административный о
круг', 'Северо-Западный административный округ', 'Западный административ
ный округ', 'Центральный административный округ', 'Восточный администрат
ивный округ', 'Юго-Восточный административный округ', 'Южный администрат
ивный округ', 'Юго-Западный административный округ']
```

Проанализируем общее количество заведений и категории к которым относятся заведения в разбивке по районам.

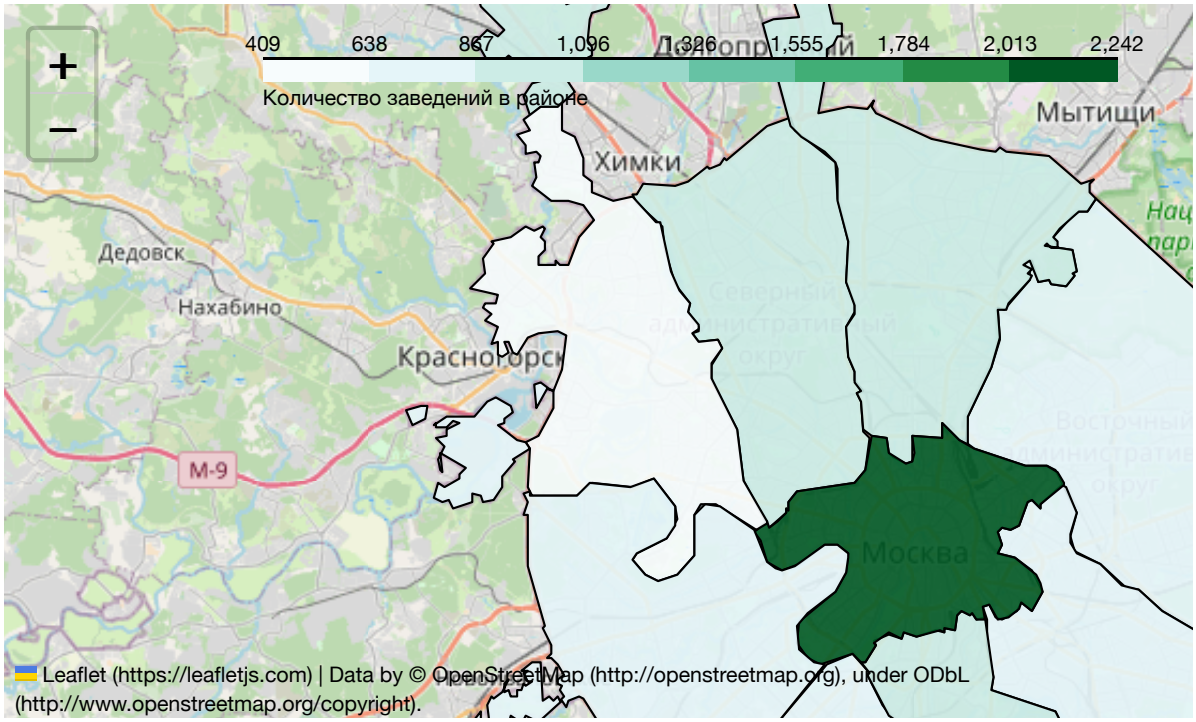
In [44]:
```
df_district = (
                df
                .pivot_table(index='district', values='name', aggfunc='co
                .reset_index()
                .rename(columns={'district': 'district', 'name': 'rest_cn
                .sort_values(by='rest_cnt',ascending=False)
)
df_district
```

Out[44]:

	district	rest_cnt
5	Центральный административный округ	2242
2	Северный административный округ	899
8	Южный административный округ	892
3	Северо-Восточный административный округ	891
1	Западный административный округ	851
0	Восточный административный округ	798
6	Юго-Восточный административный округ	714
7	Юго-Западный административный округ	709
4	Северо-Западный административный округ	409

In [45]:
```
fig = Choropleth(
    geo_data=state_geo,
    data=df_district,
    columns=['district', 'rest_cnt'],
    key_on='feature.name',
    bins = 8,
    fill_color='BuGn',
    fill_opacity=0.9,
    legend_name='Количество заведений в районе',
    name='Количество заведений в районе'
).add_to(m)

fig.geojson.add_child(
folium.features.GeoJsonTooltip(fields=['name'],
                               aliases=['округ:'],
                               labels=True,
                               localize=True,
                               sticky=False)
)

m
```

Out[45]:



Из фоновой картограммы видно, что наибольшее количество заведений расположено в Центральном административном округе Москвы - 2242 заведения.

Наименьшее количество заведений открыто в Северо-Западном административном округе - 409 заведений.

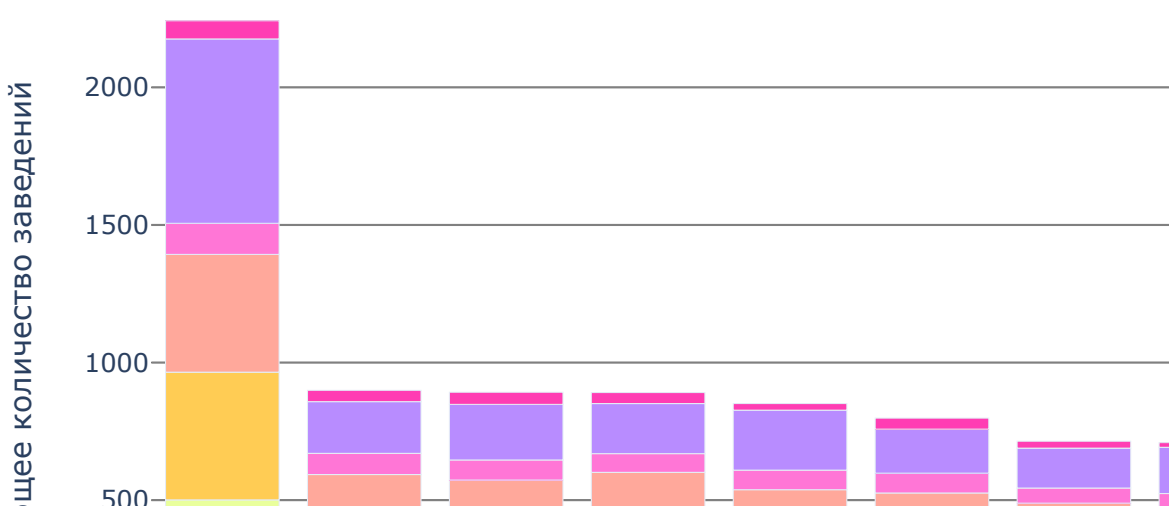
Проанализируем количество заведений каждой категории в каждом районе и посмотрим тепловую карту.

In [46]:
```
df_district = (
                df
                .replace({'district': {'Восточный административный округ'
                         'Западный административный округ': 'Западный АО'
                          'Северный административный округ': 'Северный АО
                          'Северо-Восточный административный округ': 'Сев
                           'Северо-Западный административный округ': 'Сев
                           'Центральный административный округ': 'Централ
                           'Юго-Восточный административный округ': 'Юго-В
                           'Юго-Западный административный округ': 'Юго-За
                           'Южный административный округ': 'Южный АО'}})
                .pivot_table(index='district', columns='category', values
)
df_district['total'] = df_district.sum(axis=1)
df_district = df_district.sort_values(by='total', ascending=False).reset_
df_district = df_district.drop(columns = 'total')
df_district
```

Out[46]:

category	district	бар,паб	булочная	быстрое питание	кафе	кофейня	пиццерия	рест
0	Центральный АО	364	50	87	464	428	113	
1	Северный АО	68	39	58	235	193	77	
2	Южный АО	68	25	85	264	131	73	
3	Северо-Восточный АО	63	28	82	269	159	68	
4	Западный АО	50	37	62	239	150	71	
5	Восточный АО	53	25	71	272	105	72	
6	Юго-Восточный АО	38	13	67	282	89	55	
7	Юго-Западный АО	38	27	61	238	96	64	
8	Северо-Западный АО	23	12	30	115	62	40	

In [47]:
```
fig = px.bar(df_district, x=df_district['district'], y=df_district.column
       title='Количество и категории заведений в каждом районе',
             color_discrete_sequence=colors)

fig.update_layout(yaxis_title='Общее количество заведений', xaxis_title='
                  plot_bgcolor="white")

fig.update_yaxes(gridcolor='grey')
fig.show()
```



In [48]:
```
sns.set_style('white')
plt.figure(figsize=(10, 4))

sns.heatmap(
    df
    .replace({'district': {'Восточный административный округ': 'Восточный
                           'Западный административный округ': 'Западный АО'
                            'Северный административный округ': 'Северный АО
                            'Северо-Восточный административный округ': 'Сев
                             'Северо-Западный административный округ': 'Сев
                             'Центральный административный округ': 'Централ
                             'Юго-Восточный административный округ': 'Юго-В
                             'Юго-Западный административный округ': 'Юго-За
                             'Южный административный округ': 'Южный АО'}})
    .pivot_table(index=['district'], columns='category', values='name', a
    annot = True, cmap='BuGn', fmt='g'
)

plt.title('Тепловая карта количества заведений каждой категории в районе'
plt.xlabel('Категория заведения')
plt.xticks(rotation=30)
plt.ylabel('Район')
plt.grid()
plt.show()
```


Тепловая карта количества заведений каждой категории в районе
Район
Восточный АО 53 25 71 272 105 72 160 40
Западный АО 50 37 62 239 150 71 218 24
Северный АО 68 39 58 235 193 77 188 41
Северо-Восточный АО 63 28 82 269 159 68 182 40
Северо-Западный АО 23 12 30 115 62 40 109 18
Центральный АО 364 50 87 464 428 113 670 66
Юго-Восточный АО 38 13 67 282 89 55 145 25
Юго-Западный АО 38 27 61 238 96 64 168 17
Южный АО 68 25 85 264 131 73 202 44
бар,паб
булочная
быстрое питание
кафе
кофейня
пиццерия
ресторан
столовая
Категория заведения
600
500
400
300
200
100

Из представленных диаграмм видно, что в Центральном АО более чем в 2 раза больше заведений по сравнению с другими районами, также можно выделить наиболее часто встречающиеся заведения в каждом районе:

- **Восточный АО** - больше всего заведений категории кафе - 272 заведения, на втором месте - рестораны (160) и следом за ними - кофейни (105).
- **Западный АО** - больше всего заведений категорий кафе (239) и ресторан (218), на втором месте - кофейни (150).
- **Северный АО** - больше всего заведений категории кафе - 235, на втором месте - кофейни (193) и рестораны (188).
- **Северо-Восточный АО** - больше всего заведений категории кафе - 269 заведения, на втором месте - рестораны (160) и следом за ними - кофейни (105).
- **Северо-Западный АО** - для данного АО характерно наименьшее количество заведений, наиболее часто встречающиеся заведения - кафе (115) и рестораны (109).
- **Центральный АО** - обладает наибольшим общим количеством заведений в АО, наиболее часто встречающиеся заведения - рестораны (670), следом идут кафе (464), кофейни (428), пабы/бары (364).
- **Юго-Восточный** - больше всего заведений категории кафе - 282 заведения, на втором месте - рестораны (145). Остальные категории так же не сильно представлены в районе.
- **Юго-Западный** - больше всего заведений категории кафе - 238 заведения, далее рестораны (168) и кофейни (96).
- **Южный АО** - больше всего заведений категории кафе - 264 заведения, далее рестораны (202) и кофейни (131).

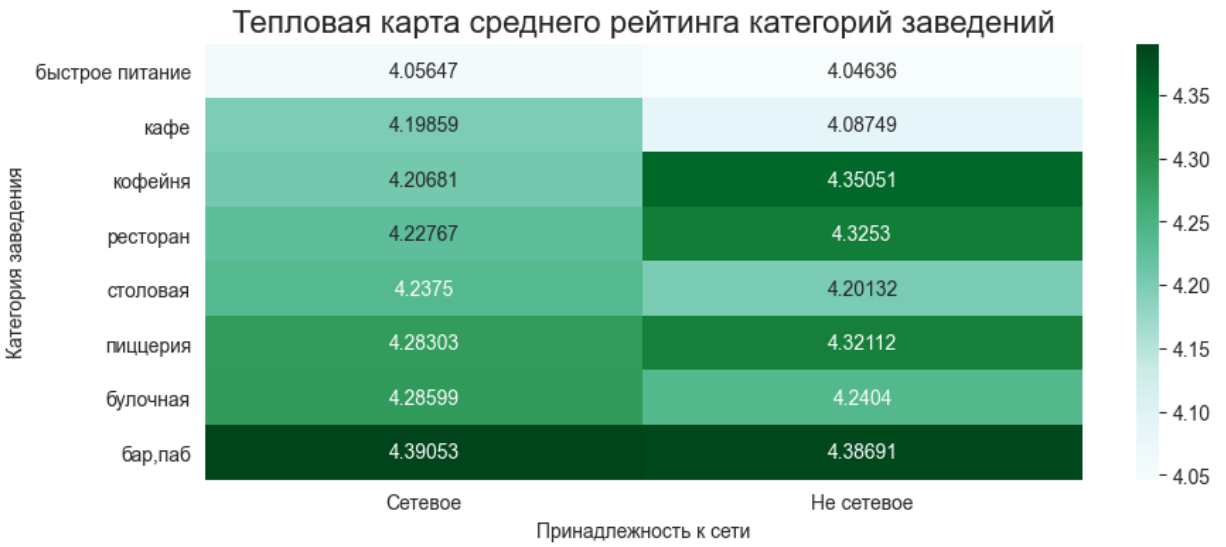
Таким образом, наиболее часто встречающиеся заведение в районе - кафе, за исключением Центрального АО, у которого часто встречающиеся заведение - ресторан, а уже следом идет уже кофе и кофейни.

Так же стоит отметить, что пабы/бары чаще открываются в Центральном АО, а в других районах количество данных категорий заведений существенно меньше.

2.5. Анализ распределения средних рейтингов

Сгруппируем категории заведений и определим средний рейтинг заведений. В связи с тем, что у значения с рейтингом имеются границы минимальной и максимальной оценки, таким образом, аномальные значения не возможны. На основе этих данных изобразим тепловую карту рейтинга.

In [49]:
```
sns.set_style('white')
plt.figure(figsize=(10, 4))

sns.heatmap(
    df
    .replace({'chain' : {True : 'Сетевое', False: 'Не сетевое'}})
    .pivot_table(index='category', columns='chain', values='rating', aggf
    .sort_values(by=['Сетевое', 'Не сетевое']),
    annot = True, cmap='BuGn', fmt='g'
)
plt.title('Тепловая карта среднего рейтинга категорий заведений', size=16
plt.xlabel('Принадлежность к сети')
plt.ylabel('Категория заведения')
plt.show()
```



Из предоставленной диаграммы видно, что:

- у Сетевых заведений средний рейтинг пабов/баров является наибольшим среди других категорий, аналогичная ситуация и у Не сетевых заведений
- на втором месте по среднему рейтингу у сетевых - булочная и пиццерия, а у не сетевых - кофейни, рестораны и пиццерии
- наименьший рейтинг у сетевых организаций - быстрое питание, а у не сетевых организаций - быстрое питание и кафе.

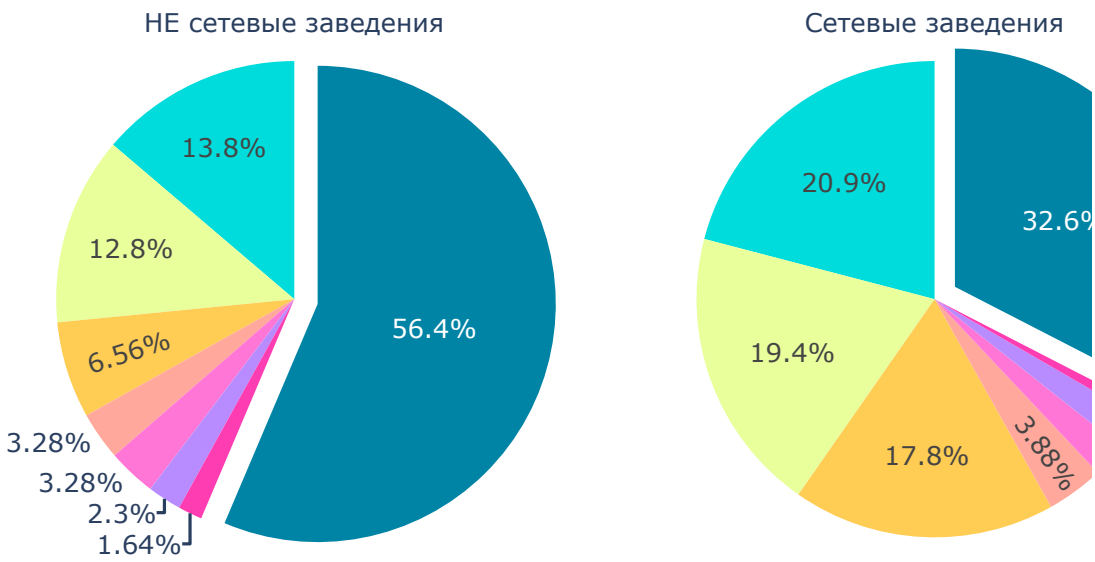
Стоит отметить, что средние рейтинги, что у сетевых и у не сетевых организаций не сильно различаются (разница между рейтингами не превышает 0,35 баллов (или 7% (0,35/5 * 100)).

In [50]:
```
df_category = (
    df
    .query('rating < 3.5')
    .pivot_table(index='category', columns='chain', values='name', aggfun
    .rename(columns={True: 'Сеть', False: 'НЕ сеть'})
)
df_category['total'] = df_category.sum(axis=1)
df_category = df_category.sort_values(by='total', ascending=False).reset_
df_category = df_category.drop(columns='total')
```

In [51]:
```
fig = make_subplots(rows=1, cols=2, specs=[[{"type": "pie"}, {"type": "pi

fig.add_trace(go.Pie(labels=df_category['category'],
                             values=df_category['НЕ сеть'].sort_values(asc
                             pull = [0.1, 0], title='НЕ сетевые заведения'
                             row=1, col=1)

fig.add_trace(go.Pie(labels=df_category['category'],
                             values=df_category['Сеть'].sort_values(ascend
                             pull = [0.1, 0], title='Сетевые заведения'),
                             row=1, col=2)

fig.update_layout(title='Доля категорий заведений с рейтингом ниже 3.5 ба
                  width=800,
                  height=600,
                  annotations=[dict(x=1.12,
                                    y=1.05,
                                    text='Категория заведения',
                                    showarrow=False)])

fig.update_traces(marker=dict(colors=colors))
fig.show()
```

Доля категорий заведений с рейтингом ниже 3.5 баллов

Кат



При анализе заведений с рейтингом меньше 3.5 баллов видно, что у сетевых и не у сетевых заведений наибольшую долю с низким рейтингом занимает категория Кафе, при этом у не сетевых заведений таких кафе практически вдвое больше, чем у сетевых.

Далее идут рестораны (у сетевых доля заведений с низкой оценкой больше на 7%, чем у не сетевых), следом за ними - быстрое питание (у сетевых заведений доля с низкой оценкой выше чем у не сетевых).

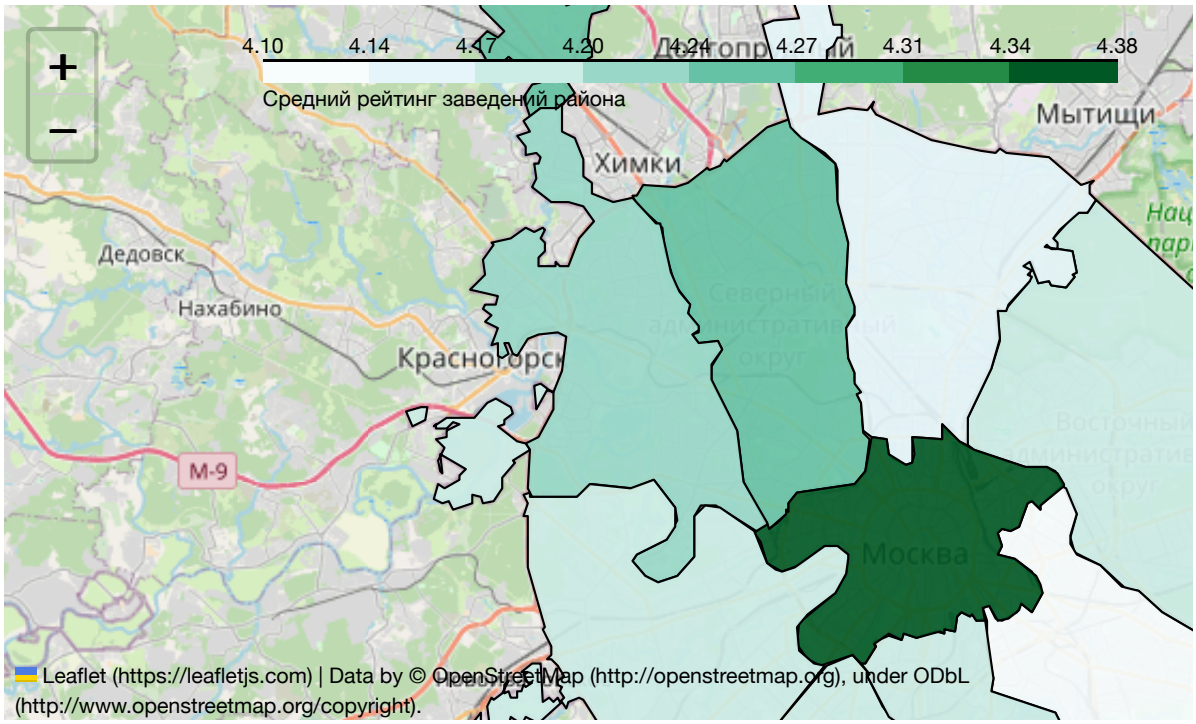
Стоит отметить, что доля кофеен с низким рейтингом у не сетевых организаций практически в трое ниже, чем у сетевых.

Теперь посмотрим имеется ли различие между средними рейтингами в районах.

In [52]:
```
df_district = (
                df
                .pivot_table(index='district', values='rating', aggfunc='m
                .reset_index()
                .rename(columns={'district': 'district', 'rating': 'mean_r
)
```

In [53]:
```
m = folium.Map(location=[moscow_lat, moscow_lng], zoom_start=10)
fig = Choropleth(
   geo_data=state_geo,
   data=df_district,
   columns=['district', 'mean_rating'],
   key_on='feature.name',
   bins = 8,
   fill_color='BuGn',
   fill_opacity=0.9,
   legend_name='Средний рейтинг заведений района',
   name = 'Средний рейтинг заведений района'
).add_to(m)

fig.geojson.add_child(
folium.features.GeoJsonTooltip(fields=['name'],
                               aliases=['округ:'],
                               labels=True,
                               localize=True,
                               sticky=False)
)

m
```

Out[53]:



Из фоновой картограммы видно, что различие в средних рейтингах заведений между районами не существенное - не более 0,3 баллов.

Центральный АО имеет наибольший средний рейтинг среди районов - в среднем 4,38 балла Вероятно это вызвано более строгому соответствию уровня клиентского сервиса и качества готовых блюд, т.к. Москву ежегодно посещает большое количество туристов как внутренних, так и врешних, которые, несомненно, посещают центральную часть города.

На втором месте по уровню рейтинга заведений - **Северный АО** (в среднем 4,24 балла). В САО расположен всемирно известный цыганский театр «Ромэн», театр классического балета имени Касаткиной и Васильева, музей русского импрессионизма, театр «Вернисаж», Петровский путевой дворец, Химкинское водохранилище (излюбленное место отдыха москвичей), Северный речной вокзал (популярное место для прогулок жителей севера столицы). Вероятно такое скопление достопремичательностей в данном районе накладывает и некую ответственность к ведению бизнеса, соответственно, и уровню сервиса и качеству блюд.

На третьем месте по уровню среднего рейтинга **Северо-Западный АО** - в среднем 4,2 балла (важно отметить, что у данного района наименьшее общее количество заведений по сравнению с другими районами). Более 46 % площади занимают природные ландшафты — лесопарковые массивы, водоёмы, заповедные зоны. СЗАО считается самым экологичным округом Москвы. Вероятно этим и вызвано малое количество заведений, а так же, вероятно, сложностями для получения разрешения на открытие заведения

Наименьший рейтинг (в среднем 4,1 балла) у **Юго-Восточного АО**, которому характерно - одна часть района находится в исторической застройке, другая — в бывших и нынешних промзонах. Здесь сосредоточен большой промышленный потенциал: Московский нефтеперерабатывающий завод, автозавод «Москвич» (АЗЛК) и технополис «Москва» и пр.. В связи с этим, а также с традиционной для Москвы западной розой ветров, Юго-Восточный округ считается некоторыми специалистами экологически неблагоприятным. Вероятно в связи с этим и наблюдается меньшее количество заведений и более низкий рейтинг, а те заведения, которые имеются, обслуживают работающее и/или живущее в этом районе население.

2.6. Анализ распределения количества заведений и их категорий по улицам

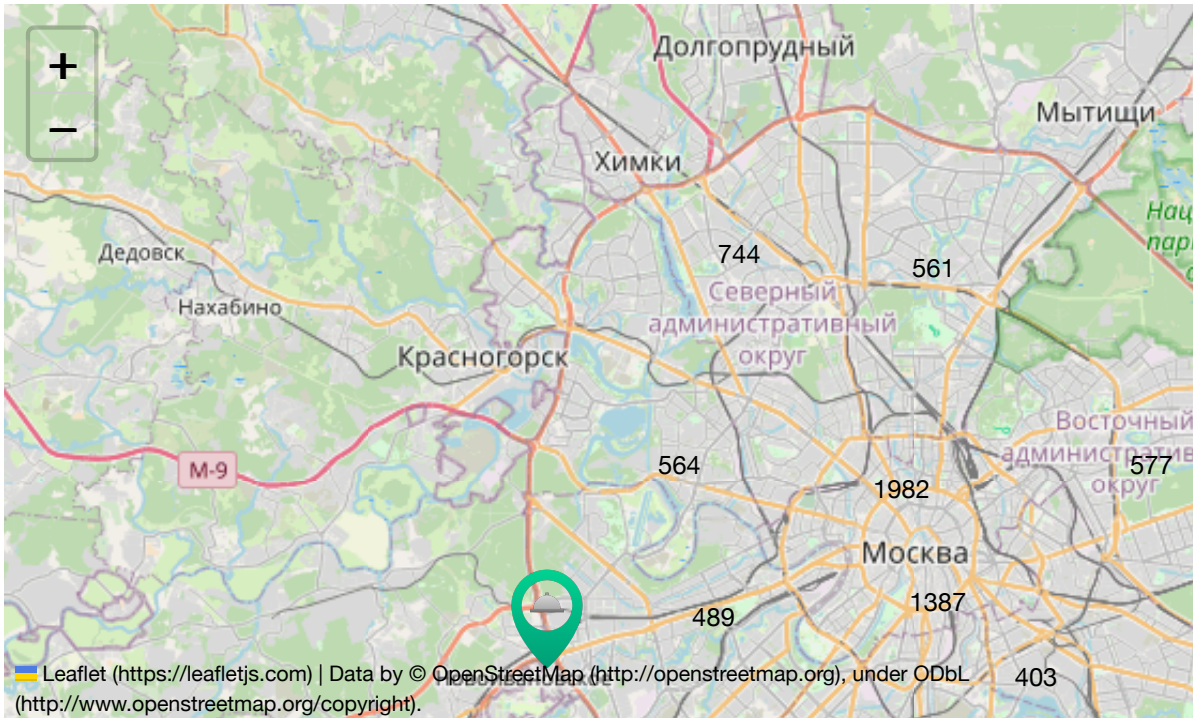
In [54]:
```
m = folium.Map(location=[moscow_lat, moscow_lng], zoom_start=10)
marker_cluster = MarkerCluster().add_to(m)
```

Отобразим все заведения на карте с помощью кластеров.

In [55]:
```
# пишем функцию, которая принимает строку датафрейма,
# создаёт маркер в текущей точке и добавляет его на карту

def create_marker(row):
    icon_url = 'https://img.icons8.com/external-flat-gradient-andi-nur-ab
    icon = CustomIcon(icon_url, icon_size=(50, 50))

    Marker(
        [row['lat'], row['lng']],
        popup=f"{row['name']} {row['rating']}",
        icon=icon,
    ).add_to(marker_cluster)
```

In [56]:
```
# применяем функцию для создания кластеров к каждой строке датафрейма
df.apply(create_marker, axis=1)

# выводим карту
m
```

Out[56]:

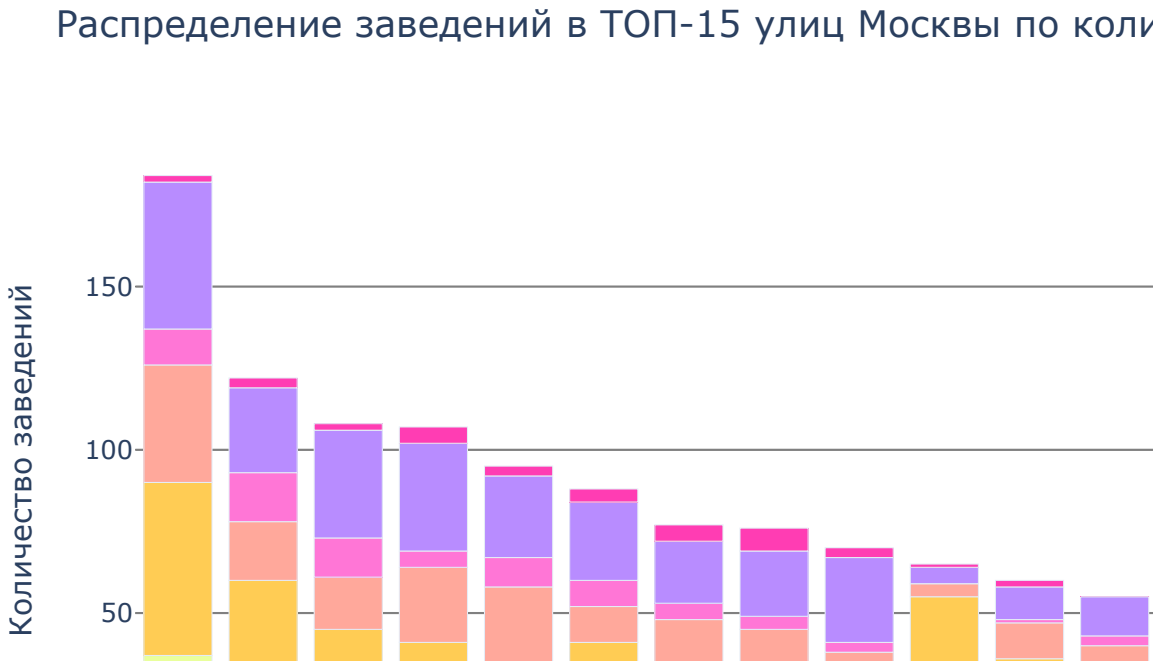


Найдите топ-15 улиц по количеству заведений и построим график распределения количества заведений и их категорий по этим улицам.

In [57]:
```
top_street = (
    df
    .pivot_table(index='street', values='name', aggfunc='count')
    .sort_values(by='name', ascending=False)[:15]
    .reset_index()
)
```

In [58]:
```
df_top_street = (
    df
    .query('street in @top_street["street"]')
    .pivot_table(index='street', columns='category', values='name', aggfu
    .sort_values(by='ресторан', ascending=False)
)
df_top_street = df_top_street.fillna(0)
df_top_street['total'] = df_top_street.sum(axis=1)
df_top_street = df_top_street.sort_values(by='total', ascending=False).re
df_top_street = df_top_street.drop(columns='total')
```

In [59]:
```
fig = px.bar(df_top_street, x=df_top_street['street'], y=df_top_street.co
       title='Количество и категории заведений в каждом районе',
            color_discrete_sequence=colors)

fig.update_layout(title="Распределение заведений в ТОП-15 улиц Москвы по
                  yaxis_title='Количество заведений', xaxis_title='Назван
                  plot_bgcolor="white")

fig.update_yaxes(gridcolor='grey')

fig.show()
```



Из представленной диаграммы видно, что на проспекте Мира сосредоточено наибольшее количество заведений. Так же можно отметить особенности и различия в регионах:

- *МКАД* обладает наибольшим количеством кафе по сравнению с другими улицами из ТОП-15, что в принципе логично, так как это довольно протяженная дорога и водителям зачастую нужен отдых и место, где можно полноценно перекусить, открытие отдельных кофеен не совсем актуально, т.к. приобрести кофе можно в том числе и на заправках.
- *Ленинградский проспект* обладает наибольшим количеством открытых пабов/баров по сравнению с другими улицами ТОПа-15. Следом за ним идет Проспект Мира по количеству открытых пабов/баров.
- Булочные отсутствуют на следующих улицах - *Каширское шоссе*, *Варшавское шоссе*, *МКАД*, *Любытинская улица*, *улица Миклухо-Маклая*.

В ТОП-15 улиц входят довольно протяженные улицы ведущие в центр Москвы из пригородов и удаленных районов, а так же МКАД. В связи с большим клиентопотоком на этих дорогах строится большое количество заведений общественного питания. Большой трафик - большое количество заведений.

Проанализируем количество заведений по уникальным названиям улиц и проверим имеются ли улицы, на которых всего одно заведение.

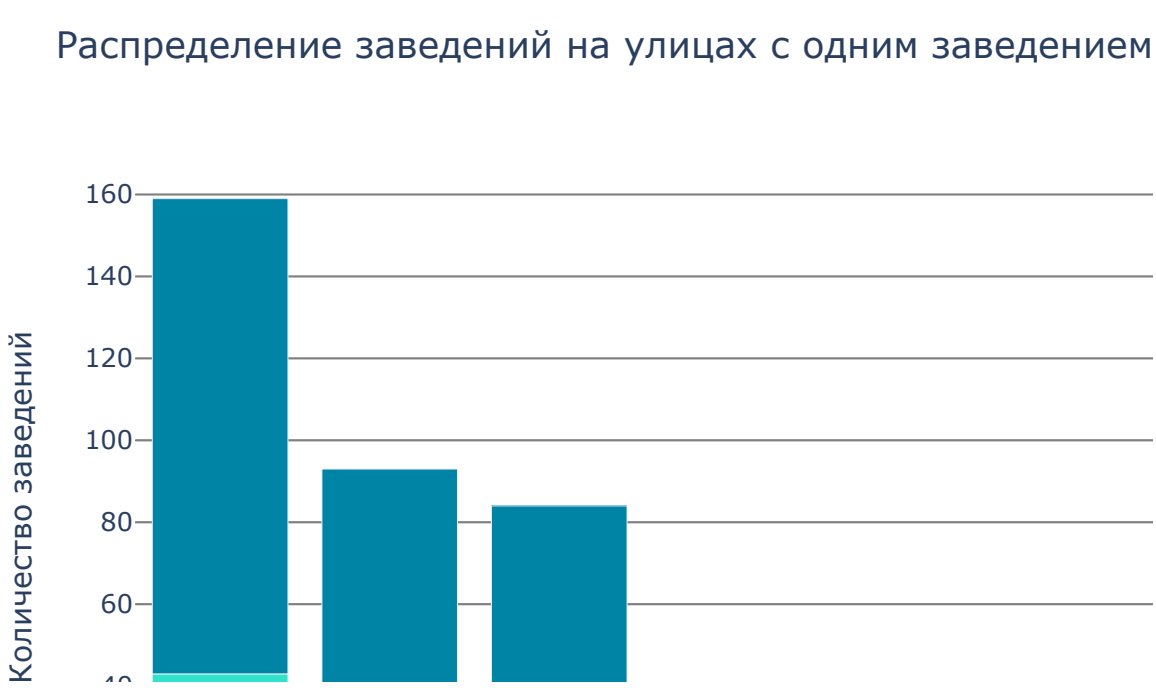
In [60]:
```
one_cafe = (
    df
    .pivot_table(index=['street'], values='name', aggfunc='count')
    .sort_values(by='name')
    .query('name == 1')
)
one_street = (
   df
    .query('street in @one_cafe.index')
)
```

In [61]:
```
print(
    ' Количество улиц с одним заведением на улице - {}'.format(len(one_st
    '\n    из них: \n',
    '       * количество улиц с одним сетевым заведением - {}'.format(len
    '\n',
    '       * количество улиц с одним не сетевым заведением - {}'.format(
)
```

```
 Количество улиц с одним заведением на улице - 457
    из них:
        * количество улиц с одним сетевым заведением - 133
        * количество улиц с одним не сетевым заведением - 324
```

In [62]:
```
category_one_street = (
    one_street
    .replace({'chain' : {True : 'Сетевое', False: 'Не сетевое'}})
    .pivot_table(index='category', columns='chain', values='name', aggfun
)
category_one_street['total'] = category_one_street.sum(axis=1)
category_one_street = category_one_street.sort_values(by='total',ascendin
category_one_street = category_one_street.drop(columns = 'total')
```

In [63]:
```
fig = px.bar(category_one_street, x='category',
             y=category_one_street.columns, color_discrete_sequence=['#30
             title='Количество и категории заведений в каждом районе', )

fig.update_layout(title="Распределение заведений на улицах с одним заведе
                  yaxis_title='Количество заведений', xaxis_title='Катего
                  plot_bgcolor="white")

fig.update_yaxes(gridcolor='grey')

fig.show()
```



При рассмотрении категорий заведений видно, что большую часть заведений относится к категории кафе.

Стоит отметить, что сетевые заведения почти в 3 раза реже открываются на улицах, где до этого не было другого заведения.

Не сетевые заведения относятся в большинстве к категории кафе, на втором месте - рестораны, далее кофейни и бары/пабы.

2.7. Анализ средних чеков заведений

Перед анализом среднего чека заведений проанализируем основные покзаатели столбца (минимальное и максимальное значение, среднее и медиану) после чего определимся какое значение (среднее или медианное) брать при анализе среднего чека региона.

In [64]:
```
df['middle_avg_bill'].describe()
```

Out[64]:
```
count     3149.000000
mean       958.053668
std       1009.732845
min          0.000000
25%        375.000000
50%        750.000000
75%       1250.000000
max      35000.000000
Name: middle_avg_bill, dtype: float64
```

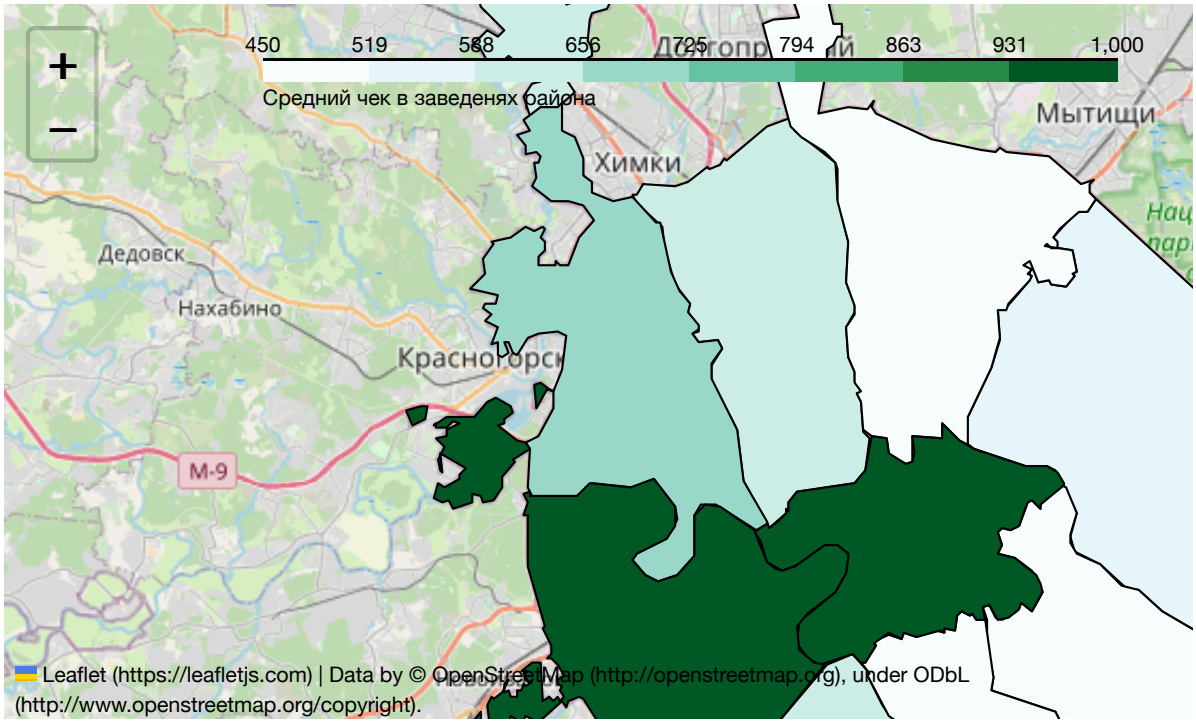
Видно, что максимальное значение чека - 35 000 руб., а минимальное 0 при этом среднее - 958, а медиана 750 это все говорит нам о том, что имебтся аномальные значения, которые искажают среднее значение. При анализе среднего чека регионов будем использовать медиану.

Проанализируем медианное значение среднего чека каждого района и отобразим это на фоновой картограмме.

In [65]:
```
avg_bill = (
    df
    .groupby('district', as_index=False)['middle_avg_bill'].agg('median')
)
```

In [66]:
```
m = folium.Map(location=[moscow_lat, moscow_lng], zoom_start=10)
fig = Choropleth(
   geo_data=state_geo,
   data=avg_bill,
   columns=['district', 'middle_avg_bill'],
   key_on='feature.name',
   bins = 8,
   fill_color='BuGn',
   fill_opacity=1,
   legend_name='Средний чек в заведенях района',
   name = 'Средний чек в заведенях района'
).add_to(m)

fig.geojson.add_child(
folium.features.GeoJsonTooltip(fields=['name'],
                               aliases=['округ:'],
                               labels=True,
                               localize=True,
                               sticky=False)
)

m
```

Out[66]:



Из фоновой картограммы видно, что наибольший средний чек (1 000 руб.) в заведениях Центрального АО и Западного АО.

Очевидный факт, что в центре цены дороже, а, соответственно, и средние чеки выше. На средний чек влияет несомненно влияет расположение к центру города, но помимо этого сказывается также и инфраструктура района.

Расположение **Западного АО** — одно из самых удачных: здесь много парков и мало промышленных предприятий, Качество недвижимости: мало ветхого жилья, много новых домов и крепких «сталинок», что соответственно сказывается на стоимости жилья в данном районе, а соответственно, и на

контингенте населения, которые могут себе позволить жить в этом районе. Респектабельный район привлекает так же крупные компании. Кутузовский проспект и проспект Вернадского усеяны бизнес-центрами и офисными зданиями. В Западном АО находятся офисы как отечественных мастодонтов вроде «Газпрома» и «Росгосстраха», так и российские представительства зарубежных компаний.

Наименьший средний чек у Юго-Восточного АО (450 руб.), Северо-Восточного АО (500 руб.), Южный АО (500 руб.).

- **Юго-Восточный АО** особенности:
  - имеет наименьший средний рейтинг (ранее выяснили),
  - часть района находится в исторической застройке, другая — в бывших и нынешних промзонах,
  - сосредоточен большой промышленный потенциал,
  - считается некоторыми специалистами экологически неблагоприятным.
- **Северо-Восточный АО** особенности:
  - самый густонаселенный округ Москвы, в нем проживает более 1,4 млн. человек,
  - значительная часть жилого фонда — старые «хрущевки» и панельные дома,
  - относительно доступные цены на жилье, в том числе и на аренду,
  - ощущается нехватка парковочных мест,
  - некоторые улицы и переулки считаются не совсем безопасными, особенно в вечернее время.
- **Южный АО** особенности:
  - закреплен статус промышленно-спального,
  - своеобразная буферная зона между престижным юго-западом и промышленно-неблагополучным юго-востоком,

Вероятно в связи с этими особенностями и наблюдается меньший средний чек, т.к. заведения районах скорее всего принимают местное население или людей, работающих в этих районах.

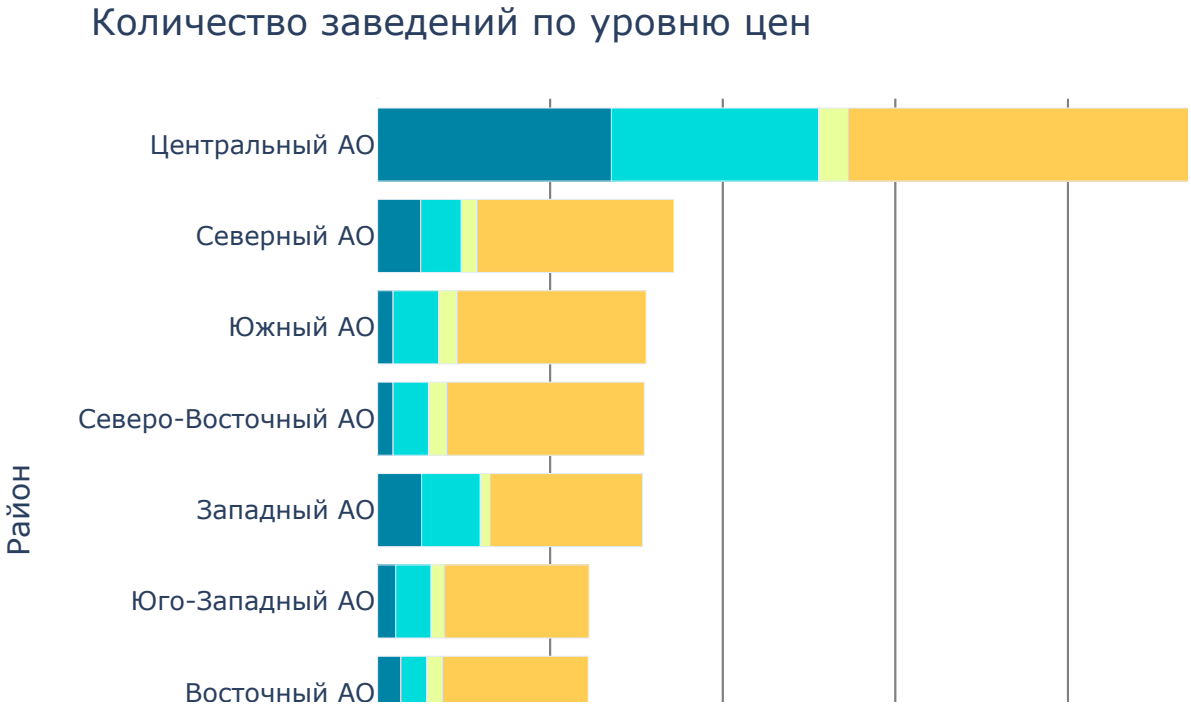
Стоит так же отметить, что данные могут быть ориентировочными по следующим причинам:

1. 62,5% пропусков имеется в столбце с информацией о среднем чеке, таким образом большей части данных у нас нет и ориентироваться только на предоставленные данные не корректно;
2. У некоторых заведений средний счет указан с пометкой "от", соответственно, указанная граница будет являться минимальным средним чеком.

Попробуем посмотреть количество заведений в разрезе установленных в них цен (низкие / средние / выше среднего/ высокие).

In [67]:
```
avg_bill = (
    df
    .replace({'district': {'Восточный административный округ': 'Восточный
                           'Западный административный округ': 'Западный АО'
                            'Северный административный округ': 'Северный АО
                            'Северо-Восточный административный округ': 'Сев
                             'Северо-Западный административный округ': 'Сев
                             'Центральный административный округ': 'Централ
                             'Юго-Восточный административный округ': 'Юго-В
                             'Юго-Западный административный округ': 'Юго-За
                             'Южный административный округ': 'Южный АО'}})
    .pivot_table(index='district', columns='price', values='name', aggfun
)
avg_bill['total'] = avg_bill.sum(axis=1)
avg_bill = avg_bill.sort_values(by='total').reset_index()
avg_bill = avg_bill.drop(columns='total')
```

In [68]:
```
fig = px.bar(avg_bill, x=avg_bill.columns, y=avg_bill['district'],
             color_discrete_sequence=colors)

fig.update_layout(title='Количество заведений по уровню цен',
                  yaxis_title='Район', xaxis_title='Количество заведений'
                  plot_bgcolor="white")

fig.update_xaxes(gridcolor='grey')

fig.show()
```



Во всей районах преобладают заведения со средним уровнем цен.

Заведения с высоким уровнем цен сосредоточены в основном в Центральном АО, Западном АО, Северном АО. В остальных районах они так же имеются но в малых количествах.

2.8. Общий вывод по разделу

Популярные категории заведений

Наибольшее количество приходится на категорию "кафе" как у сетевых, так и у не сетевых организаций, при этом не сетевых заведений более чем в 2 раза больше чем сетевых. На втором месте по популярности открытия - рестораны, при этом не сетевые рестораны превосходят в количестве сетевых почти в 2 раза.

На третьем месте по популярности - кофейни, стоит отметить, то у каждой категории количество сетевых заведений немного превосходит количество не сетевых.

А вот четвертое место по количеству заведений у сетевых приходится на

пиццерии, а у не сетевых заведений на бар/паб.

- у сетевых заведений имеется большее (в количестве):
  - столовых разница около +60%,
  - ресторанов разница около +10%,
  - пиццерий разница около +5%
- у не сетевых заведений имеется большее (в количестве):
  - баров/пабов разница около +66%,
  - булочные разница около +70%,
  - быстрое питание разница около +50%,
  - кафе разница около +45%,
  - кофейни разница около +35%.

При анализе доли категорий заведений выявлено, что у сетевых и не сетевых организаций, имеется схожая тенденция и можно проранжировать категории от наиболее популярных до менее популярных.

Проранжируем доли заведений в порядке убывания:

1. Кафе - занимают наибольшую долю среди всех заведений, на них приходится 30,8% у сетевых заведений, 24,3% - не сетевых.
2. Ресторан - доля среди не сетевых - 25,2%, сетевых - 22,8%,
3. Кофейня - доля среди не сетевых - 13,3%, сетевых - 22,5%,
4. Паб/бар - доля среди не сетевых - 11,5%, сетевых - 10,3%,
5. Пиццерия - доля среди не сетевых - 7,13%, сетевых - 7,24%,
6. Быстрое питание - доля среди не сетевых - 5,83%, сетевых - 5,27%,
7. Столовая - доля среди не сетевых - 4,37%, сетевых - 4,9%,
8. Булочная - доля среди не сетевых - 1,9%, сетевых - 2,75%.

- у Сетевых заведений средний рейтинг пабов/баров является наибольшим среди других категорий, аналогичная ситуация и у Не сетевых заведений,
- на втором месте по среднему рейтингу у сетевых - булочная и пиццерия, а у не сетевых - кофейни, рестораны и пиццерии,
- наименьший рейтинг у сетевых организаций - быстрое питание, а у не сетевых организаций - быстрое питание и кафе.

Стоит отметить, что средние рейтинги, что у сетевых и у не сетевых организаций не сильно различаются (разница между рейтингами не превышает 0,35 баллов (или 7% (0,35/5 * 100)).

При анализе заведений с рейтингом меньше 3.5 баллов видно, что у сетевых и не у сетевых заведений наибольшую долю с низким рейтингом занимает категория Кафе, при этом у не сетевых заведений таких кафе практически вдвое больше, чем у сетевых.

Далее идут рестораны (у сетевых доля заведений с низкой оценкой больше на 7%, чем у не сетевых), следом за ними - быстрое питание (у сетевых заведений

доля с низкой оценкой выше чем у не сетевых).

Стоит отметить, что доля кофеен с низким рейтингом у не сетевых организаций практически в трое ниже, чем у сетевых.

Количество мест в заведениях

Медианное количество мест среди категорий сетевых заведений:

- в баре/пабе и столовой - 85 мест,
- ресторан - 80 мест,
- столовая - 66 мест,
- пиццерия и кофейня - по 60 мест,
- заведение быстрого питания - 56 мест,
- кафе - 55 мест,
- булочная - 50 мест.

При этом сетевые заведения обладают большим количеством мест по сравнению с не сетевыми заведениями, исключения составляют лишь пабы/бары и пиццерии, у которых количество мест у не сетевых заведений больше.

ТОП-15 популярных сетей в Москве

в ТОП-15 входят почти все заведения, работающие по франшизе, исключение составляет cofefest по данной сети не было обнаружено наличие возможности приобрести франшизу для открытия. При определении категорий заведений к которым относятся заведения сети было выявлено, что имеются ошибки в определении категории заведения это может быть вызвано как ошибкой при внесении информации, а так же особенностями работы и обслуживания в заведении. Например, Кафе в сравнении с рестораном имеет меньший ассортимент блюд, при этом пиццерия — это тоже кафе, но главной позицией его меню является пицца.

Рейтинг заведений Москвы по количеству заведений:

1. Шоколадница - кофейня
2. Домино'с пицца - пиццерия
3. Додо пицца - пиццерия
4. One price coffee - кофейня,
5. Яндекс лавка - значится как ресторан, но не работающий как самостоятельное заведение, а осуществляющий доставку из других ресторанов потребителям,
6. Cofix - кофейня,
7. Prime - ресторан, но на оф. сайте организации обозначается как кафе,
8. Хинкальная - кафе / ресторан, вероятно имеется разница в ассортименте блюд и обслуживания в заведении,
9. Кофепорт - кофейня,

10. Кулинарная лавка братьев караваевых - кафе,
11. Теремок - ресторан / быстрое питание (вероятно аналог категории "ресторан" как у Макдоналдс),
12. Чайхана - кафе,
13. Буханка - булочная,
14. Cofefest - кофейня,
15. Му-му - кафе, кофейня

Наиболее популярными сетями заведений в Москве являются - кофейни и пиццерии.

Анализ административных районов Москвы

Наибольшее количество заведений расположено в Центральном административном округе Москвы - 2242 заведения. Наименьшее количество заведений открыто в Северо-Западном административном округе - 409 заведений.

Основные выявленные особенности каждого района:

- **Восточный АО**:
  - больше всего заведений категории кафе - 272 заведения, на втором месте - рестораны (160) и следом за ними - кофейни (105).
  - средний рейтинг заведений района - 4,17 балла,
  - средний чек - 575 руб.
- **Западный АО**:
  - больше всего заведений категорий кафе (239) и ресторан (218), на втором месте - кофейни (150).
  - имеет наибольший средний чек (1 000 руб.) вероятно из-за самого удачного расположения: здесь много парков и мало промышленных предприятий, Качество недвижимости: мало ветхого жилья, много новых домов и крепких «сталинок», что соответственно сказывается на стоимости жилья в данном районе, а соответственно, и на контингенте населения, которые могут себе позволить жить в этом районе. Респектабельный район привлекает так же крупные компании. Кутузовский проспект и проспект Вернадского усеяны бизнес-центрами и офисными зданиями. В Западном АО находятся офисы как отечественных мастодонтов вроде «Газпрома» и «Росгосстраха», так и российские представительства зарубежных компаний.
- **Северный АО**:
  - больше всего заведений категории кафе - 235, на втором месте - кофейни (193) и рестораны (188).
  - на втором месте по уровню рейтинга заведений. В САО расположен всемирно известный цыганский театр «Ромэн», театр классического балета имени Касаткиной и Васильева, музей русского импрессионизма, театр «Вернисаж», Петровский путевой дворец,

Химкинское водохранилище (излюбленное место отдыха москвичей), Северный речной вокзал (популярное место для прогулок жителей севера столицы). Вероятно такое скопление достопремичательностей в данном районе накладывает и некую ответственность к ведению бизнеса, соответственно, и уровню сервиса и качеству блюд.

- **Северо-Восточный АО**:
  - больше всего заведений категории кафе - 269 заведения, на втором месте - рестораны (160) и следом за ними - кофейни (105),
  - наименьший средний чек (500 руб.) верочтно в связи с тем, что это самый густонаселенный округ Москвы, в нем проживает более 1,4 млн. человек, значительная часть жилого фонда — старые «хрущевки» и панельные дома, относительно доступные цены на жилье, в том числе и на аренду, ощущается нехватка парковочных мест, некоторые улицы и переулки считаются не совсем безопасными, особенно в вечернее время.
- **Северо-Западный АО**:
  - для данного АО характерно наименьшее количество заведений, наиболее часто встречающиеся заведения - кафе (115) и рестораны (109).
  - на третьем месте по уровню среднего рейтинга - в среднем 4,2 балла (важно отметить, что у данного района наименьшее общее количество заведений по сравнению с другими районами). Более 46 % площади занимают природные ландшафты — лесопарковые массивы, водоёмы, заповедные зоны. СЗАО считается самым экологичным округом Москвы. Вероятно этим и вызвано малое количество заведений, а так же, вероятно, сложностями для получения разрешения на открытие заведения.
- **Центральный АО**:
  - обладает наибольшим общим количеством заведений в АО, наиболее часто встречающиеся заведения - рестораны (670), следом идут кафе (464), кофейни (428), пабы/бары (364).
  - пабы/бары здесь открываются чаще, а в других районах количество данных категорий заведений существенно меньше.
  - Имеет наибольшый средний рейтинг среди районов - в среднем 4,38 балла Вероятно это вызвано более строгому соответствию уровня клиентского сервиса и качества готовых блюд, т.к. Москву ежегодно посещает большое количество туристов как внутренних, так и врешних, которые, несомненно, посещают центральную часть города.
  - имеет наибольший средний чек (1 000 руб.)
- **Юго-Восточный**:
  - больше всего заведений категории кафе - 282 заведения, на втором месте - рестораны (145).
  - наименьший средний рейтинг заведений (в среднем 4,1 балла).
  - наименьший средний чек (450 руб.), вероятно из-за того, что часть

района находится в исторической застройке, другая — в бывших и нынешних промзонах, сосредоточен большой промышленный потенциал, считается некоторыми специалистами экологически неблагоприятным.
    - одна часть района находится в исторической застройке, другая — в бывших и нынешних промзонах. Здесь сосредоточен большой промышленный потенциал: Московский нефтеперерабатывающий завод, автозавод «Москвич» (АЗЛК) и технополис «Москва» и пр.. В связи с этим, а также с традиционной для Москвы западной розой ветров, Юго-Восточный округ считается некоторыми специалистами экологически неблагоприятным. Вероятно в связи с этим и наблюдается меньшее количество заведений и более низкий рейтинг, а те заведения, которые имеются, обслуживают работающее и/или живущее в этом районе население.
- **Юго-Западный**:
    - больше всего заведений категории кафе - 238 заведения, далее рестораны (168) и кофейни (96).
    - средний рейтинг заведений района - 4,17 балла,
    - средний чек - 600 руб.,
- **Южный АО**:
    - больше всего заведений категории кафе - 264 заведения, далее рестораны (202) и кофейни (131),
    - наименьший средний чек (500 руб.) вероятно в связи с тем, что за ним закреплен статус промышленно-спального, своеобразная буферная зона между престижным юго-западом и промышленно-неблагополучным юго-востоком.

Важно отметить, что данные по среднему чеку могут быть ориентировочными по следующим причинам:

1. 62,5% пропусков имеется в столбце с информацией о среднем чеке, таким образом большей части данных у нас нет и ориентироваться только на предоставленные данные не корректно;
2. У некоторых заведений средний счет указан с пометкой "от", соответственно, указанная граница будет являться минимальным средним чеком.

При анализе уровня цен завений видно, чт во всех районах преобладают заведения со средним уровнем цен.

Заведения с высоким уровнем цен сосредоточены в основном в Центральном АО, Западном АО, Северном АО. В остальных районах они так же имеются но в малых количествах.

ТОП-15 улиц Москвы

На проспекте Мира сосредоточено наибольшее количество заведений. Так же можно отметить особенности и различия в регионах:

- МКАД обладает наибольшим количеством кафе по сравнению с другими улицами из ТОП-15, что в принципе логично, так как это довольно протяженная дорога и водителям зачастую нужен отдых и место, где можно полноценно перекусить, открытие отдельных кофеен не совсем актуально, т.к. приобрести кофе можно в том числе и на заправках.
- Ленинградский проспект обладает наибольшим количеством открытых пабов/баров по сравнению с другими улицами ТОПа-15. Следом за ним идет Проспект Мира по количеству открытых пабов/баров.
- Булочные отсутствуют на следующих улицах - Каширское шоссе, Варшавское шоссе, МКАД, Любытинская улица, улица Миклухо-Маклая.

В ТОП-15 улиц входят довольно протяженные улицы ведущие в центр Москвы из пригородов и удаленных районов, а так же МКАД. В связи с большим клиентопотоком на этих дорогах строится большое количество заведений общественного питания. Большой трафик - большое количество заведений.

На средний чек влияет несомненно влияет расположение к центру города, но помимо этого сказывается также и инфраструктура района.

Улицы с одним заведением

Количество улиц с одним заведением на улице - 457, из них:

- количество улиц с одним сетевым заведением - 133
- количество улиц с одним не сетевым заведением - 324

Большая часть заведений относится к категории кафе.

Сетевые заведения почти в 3 раза реже открываются на улицах, где до этого не было другого заведения.

Не сетевые заведения относятся в большинстве к категории кафе, на втором месте - рестораны, далее кофейни и бары/пабы.

3. Открытие кофейни по мотивам сериала «Друзья»

Основателям фонда «Shut Up and Take My Money» не даёт покоя успех сериала «Друзья». Их мечта — открыть такую же крутую и доступную, как «Central Perk», кофейню в Москве. Будем считать, что заказчики не боятся конкуренции в этой сфере, ведь кофеен в больших городах уже достаточно. Попробуем определить, осуществима ли мечта клиентов.

Проанализируем:

1. Сколько всего кофеен в датасете? В каких районах их больше всего, каковы особенности их расположения?
2. Есть ли круглосуточные кофейни?
3. Какие у кофеен рейтинги? Как они распределяются по районам?
4. На какую стоимость чашки капучино стоит ориентироваться при открытии и почему?

3.1. Анализ распределения кофеен по регионам

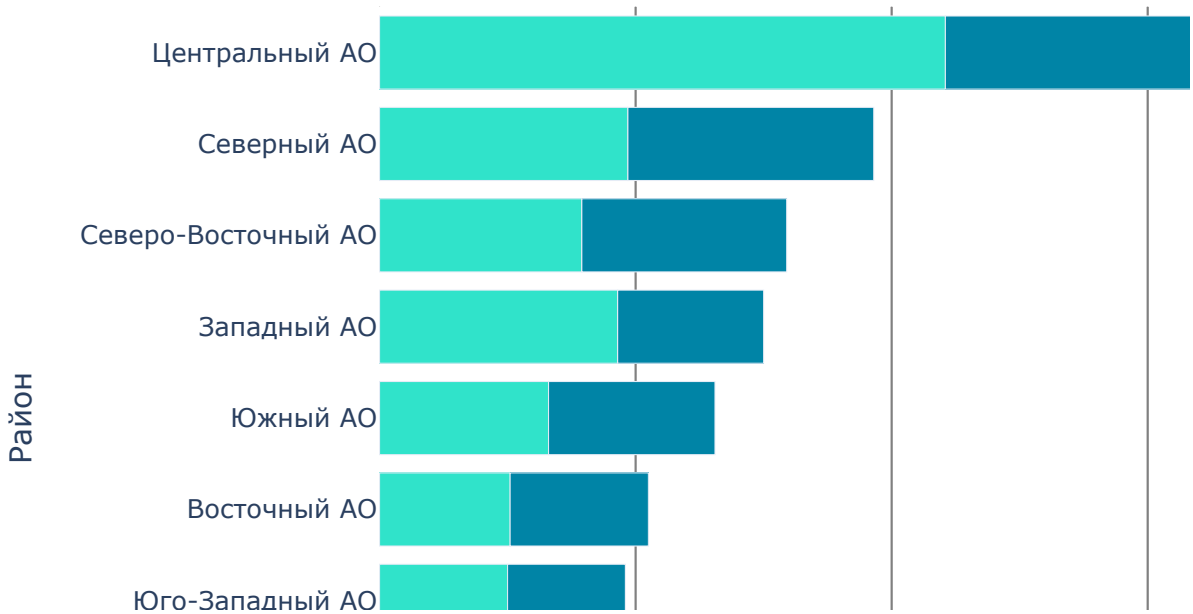
In [69]:
```
data = df.query('category == "кофейня"')
```

In [70]:
```
print(
    ' Всего кофеен в датасете – {} или {}% от общего количества заведений
    .format(len(data), round(len(data) / len(df)*100, 2)),
    '\n',
    '   из них:',
    '\n',
    '    * сетевых кофеен – {} или {}% от общего количества кофеен'
    .format(len(data.query('chain == True')), round(len(data.query('chain
    '\n',
    '    * не сетевых кофеен – {} или {}% от общего количества кофеен'
    .format(len(data.query('chain == False')), round(len(data.query('chai
    '\n',
    '\n',
    'Количество круглосуточных кофеен всего – {} или {}% от общего количе
    .format(data['is_24/7'].value_counts()[1], round(data['is_24/7'].valu
    '\n',
    '   из них:',
    '\n',
    '    * сетевых круглосуточных кофеен – {} или {}% от общего количеств
    .format(data.query('chain == True')['is_24/7'].value_counts()[1], rou
    '\n'
    '     * не сетевых круглосуточных кофеен – {} или {}% от общего колич
    .format(data.query('chain == False')['is_24/7'].value_counts()[1], ro
    '\n'
)
```

```
 Всего кофеен в датасете – 1413 или 16.81% от общего количества заведени
й
    из них:
     * сетевых кофеен – 720 или 8.57% от общего количества кофеен
     * не сетевых кофеен – 693 или 8.25% от общего количества кофеен

 Количество круглосуточных кофеен всего – 59 или 0.7% от общего количест
ва кофеен
    из них:
     * сетевых круглосуточных кофеен – 50 или 0.59% от общего количества
круглосуточных кофеен
     * не сетевых круглосуточных кофеен – 9 или 0.11% от общего количест
ва круглосуточных кофеен
```

In [71]:
```
df_district = (
    data
    .replace({'district': {'Восточный административный округ': 'Восточный
                           'Западный административный округ': 'Западный АО'
                            'Северный административный округ': 'Северный АО
                            'Северо-Восточный административный округ': 'Сев
                             'Северо-Западный административный округ': 'Сев
                             'Центральный административный округ': 'Централ
                             'Юго-Восточный административный округ': 'Юго-В
                             'Юго-Западный административный округ': 'Юго-За
                             'Южный административный округ': 'Южный АО'},
              'chain' : {True : 'Сетевое', False: 'Не сетевое'}})
    .pivot_table(index='district', columns='chain', values='name', aggfun
)
df_district['total'] = df_district.sum(axis=1)
df_district = df_district.sort_values(by='total').reset_index()
df_district = df_district.drop(columns='total')
```

In [72]:
```
fig = px.bar(df_district, x=df_district.columns,
             y='district', color_discrete_sequence=['#30E3CA', '#0084a6']

fig.update_layout(title="Распределение кофеен в регионах Москвы",
                  yaxis_title='Район', xaxis_title='Количество заведений'
                  plot_bgcolor="white")

fig.update_xaxes(gridcolor='grey')

fig.show()
```



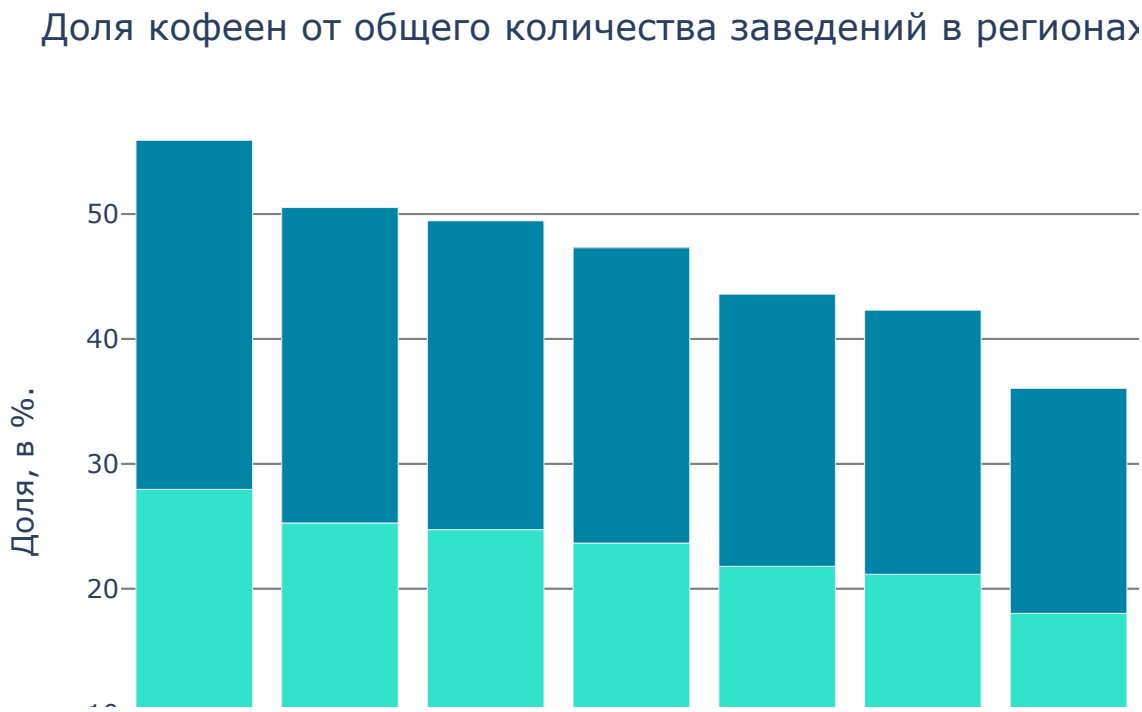
Из диаграммы видно, что соотношение сетевых к не сетевым кофейням в регионах +- равное.

Наибольшее количество кофеен открыто в Центральном АО, на втором месте Северный АО далее Северо-Восточный АО и Западный АО.

In [73]:
```
df_total_cafe = (
    df
    .replace({'district': {'Восточный административный округ': 'Восточный
                           'Западный административный округ': 'Западный АО'
                            'Северный административный округ': 'Северный АО
                            'Северо-Восточный административный округ': 'Сев
                             'Северо-Западный административный округ': 'Сев
                             'Центральный административный округ': 'Централ
                             'Юго-Восточный административный округ': 'Юго-В
                             'Юго-Западный административный округ': 'Юго-За
                             'Южный административный округ': 'Южный АО'},
              'chain' : {True : 'Сетевое', False: 'Не сетевое'}})
    .pivot_table(index='district', columns='chain', values='name', aggfun
)
```

In [74]:
```
df_district = df_district.merge(df_total_cafe, on='district', how='left')
df_district['prct_chain'] = round(df_district['Сетевое_x'] / df_district[
df_district['prct_not_chain'] = round(df_district['Сетевое_x'] / df_distr
df_district['total'] = df_district['prct_chain'] + df_district['prct_not_
df_district = df_district.sort_values(by='total', ascending=False)
```

In [75]:
```
fig = px.bar(df_district, x='district', y=['prct_chain', 'prct_not_chain'
             color_discrete_sequence=['#30E3CA', '#0084a6'])

fig.update_layout(title="Доля кофеен от общего количества заведений в рег
                  yaxis_title='Доля, в %.', xaxis_title='Район',
                  plot_bgcolor="white")

fig.update_yaxes(gridcolor='grey')

fig.show()
```

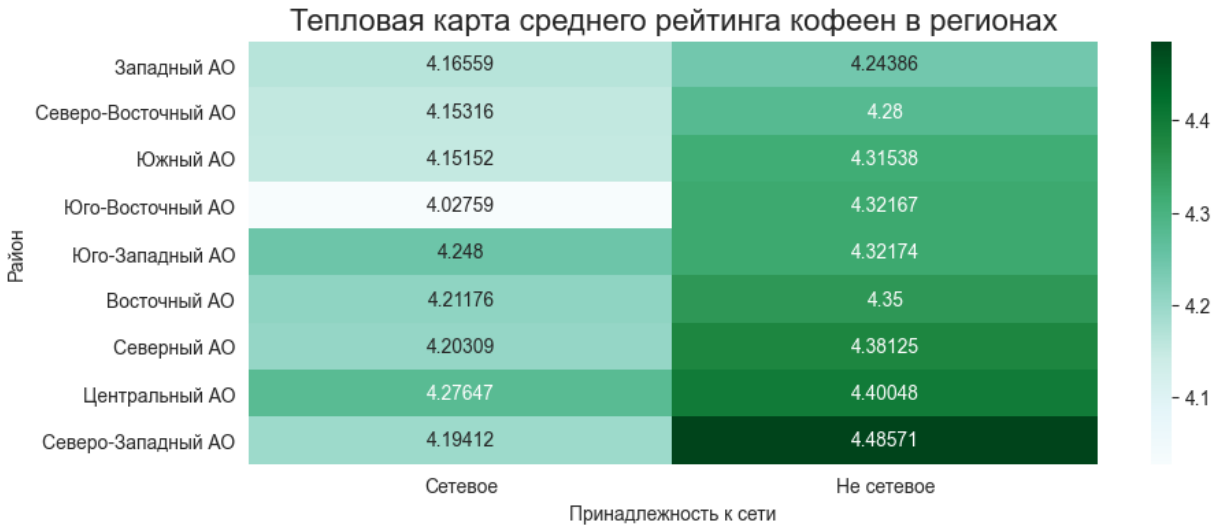


Из представленной диаграммы видно, что кофейни в большинстве регионов в совокупности сетевые + не сетевые занимают долю от 40% и более.

Меньшая доля кофеен обнаружена в Юго-Восточном АО, Юго-Западном АО, Восточном АО.

3.2. Анализ среднего рейтинга кофеен по регионам

In [76]:
```
sns.set_style('white')
plt.figure(figsize=(10, 4))

sns.heatmap(
    data
    .replace({'district': {'Восточный административный округ': 'Восточный
                           'Западный административный округ': 'Западный АО'
                            'Северный административный округ': 'Северный АО
                            'Северо-Восточный административный округ': 'Сев
                             'Северо-Западный административный округ': 'Сев
                             'Центральный административный округ': 'Централ
                             'Юго-Восточный административный округ': 'Юго-В
                             'Юго-Западный административный округ': 'Юго-За
                             'Южный административный округ': 'Южный АО'},
              'chain' : {True : 'Сетевое', False: 'Не сетевое'}})
    .pivot_table(index='district', columns='chain', values='rating', aggf
    .sort_values(by=['Не сетевое', 'Сетевое']),
    annot = True, cmap='BuGn', fmt='g'
)
plt.title('Тепловая карта среднего рейтинга кофеен в регионах', size=16)
plt.xlabel('Принадлежность к сети')
plt.ylabel('Район')
plt.show()
```



Из тепловой карты видно, что не сетевые кофейни обладают большим средним рейтингом.

Средний рейтинг сетевых кофеер варьируется от 4,02 балла до 4,27 баллов от региона к региону, когда у не сетевых средний рейтинг варьируется от 4,28 баллов до 4,48 баллов.

Регионы с высоким рейтингом у кофеен:

- Сетевые заведения - Центральный АО и Юго-Западный АО.
- Не сетевые заведения - Северо-Западный АО, Центральный АО, Северный АО.

Таким образом, при решении открыть кофейню необходимо ориентироваться на качество сервиса и качества продукции на не сетевые кофейни, так как они обладают большим рейтингом, и если этого не делать потребители будут уходить к конкурентам.

Перед тем как анализировать среднюю стоимость чашки капучино по регионам посмотрим минимальную и максимальную цену, среднее и медианное значение.

3.3. Анализ средней стоимости кружки капучино по регионам

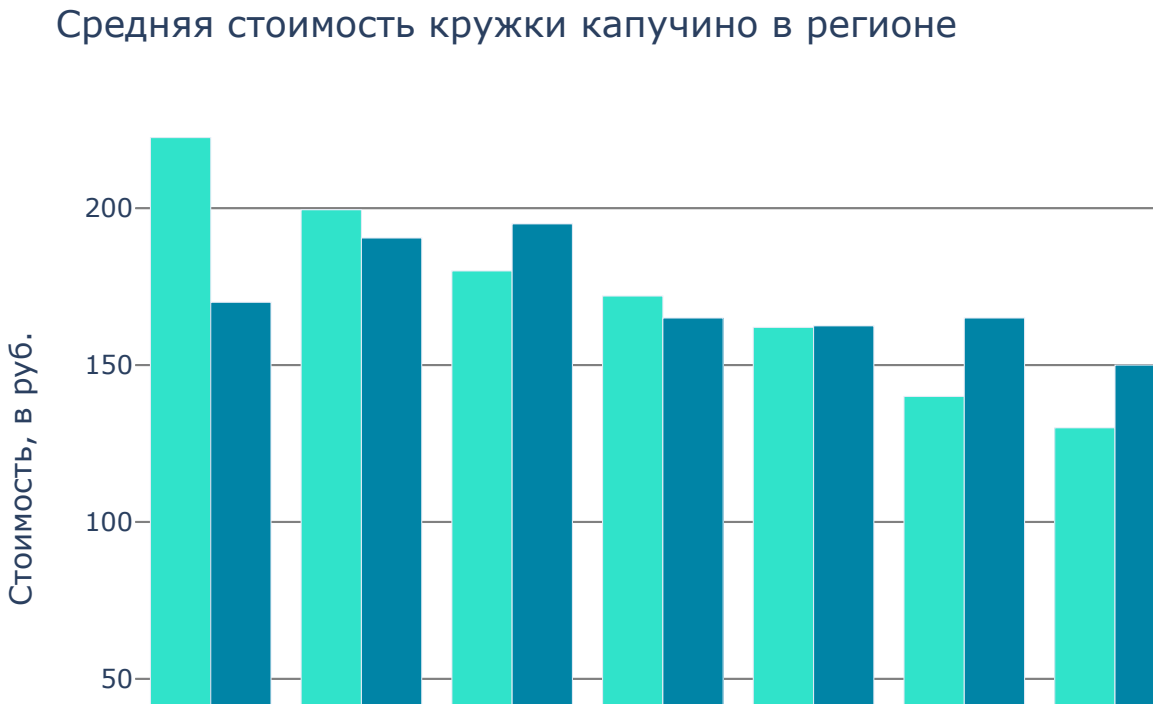
In [77]:
```
data['middle_coffee_cup'].describe()
```

Out[77]:
```
count     521.000000
mean      175.055662
std        89.753009
min        60.000000
25%       124.000000
50%       170.000000
75%       225.000000
max      1568.000000
Name: middle_coffee_cup, dtype: float64
```

Можно сделать вывод, что имеются аномальные значения такие как 1 568 руб. за кружку капучино max и 60 руб. min. Таким образом, в дальнейшем будем использовать медианное значение при анализе.

In [78]:
```
df_district = (
    data
    .replace({'district': {'Восточный административный округ': 'Восточный
                           'Западный административный округ': 'Западный АО'
                            'Северный административный округ': 'Северный АО
                            'Северо-Восточный административный округ': 'Сев
                             'Северо-Западный административный округ': 'Сев
                             'Центральный административный округ': 'Централ
                             'Юго-Восточный административный округ': 'Юго-В
                             'Юго-Западный административный округ': 'Юго-За
                             'Южный административный округ': 'Южный АО'},
              'chain' : {True : 'Сетевое', False: 'Не сетевое'}})
    .pivot_table(index='district', columns='chain', values='middle_coffee
)
df_district['total'] = df_district.sum(axis=1)
df_district = df_district.sort_values(by='total', ascending=False)
df_district = df_district.drop(columns='total')
```

In [79]:
```
fig = px.bar(df_district, x=df_district.index, y=df_district.columns,
             color='chain', barmode='group', color_discrete_sequence=['#3

fig.update_layout(title='Средняя стоимость кружки капучино в регионе',
                  yaxis_title='Стоимость, в руб.', xaxis_title='Район',
                  plot_bgcolor="white")

fig.update_yaxes(gridcolor='grey')

fig.show()
```



Из представленной диаграммы можно сделать следующие выводы:

- Сетевые заведения имеют цену выше за чашку капучино только в трех районах - Западный АО, Юго-Западный АО, Северо-Западный АО.
- Не сетевые заведения имеют цену выше за чашку капучино в следующих районах - Центральный АО, Северный АО, Юго-Восточный АО, Южный АО.
- в Северо-Восточном АО и Восточном АО цены одинаковы что у сетевых, что у не сетевых заведений.

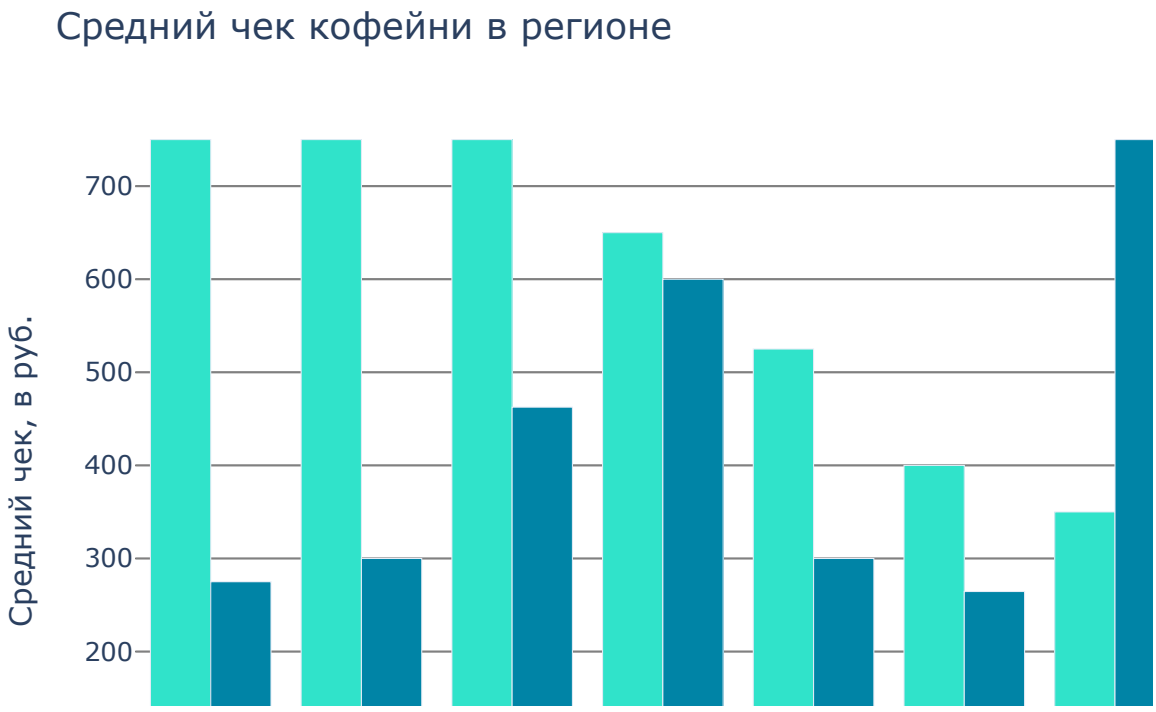
Средняя стоимость кружки капучино в основном варьируется от 160 до 200 руб. за кружку. При определении цены необходимо ориентироваться на среднюю цену стоимости кружки в регионе.

Проанализируем так же средний чек кофейни в каждом регионе.

3.4. Анализ среднего чека кофеен по регионам

In [80]:
```
df_district = (
    data
    .replace({'district': {'Восточный административный округ': 'Восточный
                           'Западный административный округ': 'Западный АО'
                            'Северный административный округ': 'Северный АО
                            'Северо-Восточный административный округ': 'Сев
                             'Северо-Западный административный округ': 'Сев
                             'Центральный административный округ': 'Централ
                             'Юго-Восточный административный округ': 'Юго-В
                             'Юго-Западный административный округ': 'Юго-За
                             'Южный административный округ': 'Южный АО'},
              'chain' : {True : 'Сетевое', False: 'Не сетевое'}})
    .pivot_table(index='district', columns='chain', values='middle_avg_bi
)
```

In [81]:
```
fig = px.bar(df_district.sort_values(by='Сетевое', ascending=False), x=df
             y=df_district.columns, color='chain', barmode='group',
             color_discrete_sequence=['#30E3CA', '#0084a6'])

fig.update_layout(title='Средний чек кофейни в регионе',
                  yaxis_title='Средний чек, в руб.', xaxis_title='Район',
                  plot_bgcolor="white")

fig.update_yaxes(gridcolor='grey')

fig.show()
```



В большинстве регионов средний чек сетевой кофейни выше, чем у не сетевой.

В Восточном АО и Западном АО средний чек сетевой кофейни почти в 3 раза выше, чем у не сетевой, и в 1,5 раза выше у Северного АО и Северо-Западного АО.

3.5. Общий вывод по разделу

Исходя из проделанного анализа можно предложить 3 района для возможного открытия:

1. **Центральный АО:**
   - большой трафик, который достигается в том числе туристами, как внешними, так и внутренними,
   - средний показатель среднего чека,
   - предлагаю установить стоимость одной кружки капучино - 195 руб. (на уровне не сетевых организаций),
   - рейтинги других кофеен - 4,4 балла не сетевые (лучший показатель среди регионов) и 4,27 балла у сетевых (лучший показатель среди регионов),

- регион с довольно высокой конкуренцией - уже открыто около 430 кофеен

2. **Западный АО:**
    - перспективный район, здесь много парков и мало промышленных предприятий, много новых домов, что соответственно сказывается на стоимости жилья в данном районе, а соответственно, и на контингенте населения, которые могут себе позволить жить в этом районе. Респектабельный район привлекает так же крупные компании, тут находятся офисы как отечественных мастодонтов вроде «Газпрома», так и российские представительства зарубежных компаний.
    - у сетевых кофеен довольно высокий средний чек, не исключено за счет громкого имени и узнаваемости бренда,
    - рейтинги других кофеен - 4,24 балла не сетевые (наихудший показатель среди регионов) и 4,16 балла у сетевых.
    - предлагаю установить стоимость одной кружки капучино - 190 руб. (среднее значение между сетевыми и не сетевыми кофейнями). За счет популярности сериала и концепции заведения привлечем большое количество фанатов, в последующем это может вывести заведение на уровень сетевых организаций (так как сетевые зачастую привлекают посетителей из-за узнаваемости бренда),
    - наполняемость региона конкурентами - средняя, имеется открытых 150 кофеен.
3. **Северо-Восточный АО:**
    - самый густонаселенный округ Москвы, в нем проживает более 1,4 млн. человек
    - довольно высокий средний чек у кофеен, при этом разница у сетевых и не сетевых не кофеен не велика,
    - предлагаю установить стоимость одной кружки капучино - 160 руб. - на уровне конкурентов,
    - рейтинги других кофеен - 4,28 балла не сетевые и 4,15 балла у сетевых,
    - наполняемость региона конкурентами - средняя, имеется открытых 159 кофеен,
    - некоторые улицы и переулки считаются не совсем безопасными, особенно в вечернее время поэтому нужно внимательнее подойти к выбору места (если останавливаться на этой локации), а так же к выбору времени работы.

При открытии детальнее проанализировать конкурентов из не сетевых кофеен с целью подчеркнуть их положительный опыт, т.к. они обладают лучшими рейтингами на довольно высоком уровне.

Так же рекомендуется проанализировать опыт похожего заведения категории кафе в г. Санкт-Петербург с целью почерпнуть их положительный опыт и в последующем возмножно расширить ассортимент блюд и составить конкуренцию категориям кафе.

Оптимальным для такого рода заведения график работы считаю - с пн по вс с 8:00 до 22:00. При таком графике работы в утренние часы посетители могут успевать забегать за кофе перед работой или позавтракать, а в вечернее время - снизим вероятность скопления людей с целью не совсем культурного времяпрепровождения.

Отмела остальные регионы по следующим причинам:

1. Концепция заведения подразумевает "дружеские посиделки" или возможность перекусить/ посидеть в комфортной атмосфере, без суеты, с последующим возможным прицелом на семейное кофе.
2. Нужны регионы с высоким потенциалом и/или высоким трафиком. Будет трафик - будут продажи, работь нужно будет в этом случае уже над качеством продукции/сервиса и конверсией.
3. Юго-Восточный АО привлекателен высоким средним чеком у несетевых кофеен, но так же этому округу характерно - не совсем удачное расположение и наличие промышленных предприятий, промзоны занимают больше трети территории
4. Юго-Западный - привлекателен в потенциале, но на данным момент там наблюдается довольно низкий средний чек у кофейни при довольно высоком уровне стоимости кружки кофе, а значит кросс-продажи идут с трудом.
5. Северо-Западный АО - у данного района наименьшее общее количество заведений по сравнению с другими районами), при этом доля кофеен больше 40%. Более 46 % площади занимают природные ландшафты — лесопарковые массивы, водоёмы, заповедные зоны. Могут возникнуть сложности при открытии и согласовании площадей. При этом средний чек не сетевых кофеен минимален при средней цены кружки капучино. Кофейни этого региона имеют высокий рейтинг - конкуренция может быть довольно жесткая.
6. Северный АО - доля кофеен в этом регионе свыше 50% - конкуренция довольно жесткая, но при этом видимо есть трафик, довольно высокий средний рейтинг у кофеен, средняя цена кружки капучино, высокий средний чек у сетевых кофеен, а у не сетеыых в в 1,5 раза меньше.
7. Южный АО - оочень низкий средний чек кофеен, вероятно в связи с тем, что за ним закреплен статус промышленно-спального, своеобразная буферная зона между престижным юго-западом и промышленно-неблагополучным юго-востоком. Стоимость кружки капучино - у сетевых выше среднего, а у не сетевых - ниже сренего, доля кофеен в регионе чуть выше 40%, рейтинги кофеен средние если смотреть среди всех рейтингов кофеен по регионам.
8. Восточный АО - видна существенная разница среднем чеке кофеен сетевых к не сетевым (у сетевых значительно превышает средний чек, более чем в 2 раза), средняя стоимость кружки капучино - ниже среднего, довольно высокие рейтинги кофеен, доля кофеен в регионе 35%.