

# Содержание

- 1 Описание проекта
- 2 Загрузка данных и подготовка их к анализу
  - 2.1 Проверка на дубли
  - 2.2 Категоризация данных
- 3 Анализ Данных
  - 3.1 Исследование соотношения видов объектов общественного питания по количеству. Построить график.
  - 3.2 Исследование соотношения сетевых и не сетевых заведений по количеству. Построить график.
  - 3.3 Для какого вида объекта общественного питания характерно сетевое распространение?
  - 3.4 Что характерно для сетевых заведений: много заведений с небольшим числом посадочных мест в каждом или мало заведений с большим количеством посадочных мест?
  - 3.5 Для каждого вида объекта общественного питания опишите среднее количество посадочных мест. Какой вид предоставляет в среднем самое большое количество посадочных мест? Постройте графики.
  - 3.6 Выделить в отдельный столбец информацию об улице из столбца address
  - 3.7 Построить график топ-10 улиц по количеству объектов общественного питания. Воспользуйтесь внешней информацией и ответьте на вопрос — в каких районах Москвы находятся эти улицы?
  - 3.8 Найти число улиц с одним объектом общественного питания. Воспользуйтесь внешней информацией и ответьте на вопрос — в каких районах Москвы находятся эти улицы?
  - 3.9 Посмотрите на распределение количества посадочных мест для улиц с большим количеством объектов общественного питания. Какие закономерности можно выявить?
- 4 Общий вывод

## Рынок заведений общественного питания Москвы

### Описание проекта

Подготовить исследование рынка общественного питания Москвы на основании открытых данных. Дать ответ на следующие вопросы

- Исследование соотношения видов объектов общественного питания по количеству. Построить график.
- Исследование соотношения сетевых и не сетевых заведений по количеству. Построить график.

- Для какого вида объекта общественного питания характерно сетевое распространение?
- Что характерно для сетевых заведений: много заведений с небольшим числом посадочных мест в каждом или мало заведений с большим количеством посадочных мест?
- Для каждого вида объекта общественного питания описать среднее количество посадочных мест. Какой вид предоставляет в среднем самое большое количество посадочных мест? Построить графики.
- Выделить в отдельный столбец информацию об улице из столбца `address`.
- Построить график топ-10 улиц по количеству объектов общественного питания. Воспользуйтесь внешней информацией и ответьте на вопрос — в каких районах Москвы находятся эти улицы?
- Найти число улиц с одним объектом общественного питания. Воспользуйтесь внешней информацией и ответьте на вопрос — в каких районах Москвы находятся эти улицы?
- Посмотреть на распределение количества посадочных мест для улиц с большим количеством объектов общественного питания. Какие закономерности можно выявить?

Данные находятся в файле `rest_data.csv`

#### Описание данных

Таблица `rest_data`:

- `id` — идентификатор объекта;
- `object_name` — название объекта общественного питания;
- `chain` — сетевой ресторан;
- `object_type` — тип объекта общественного питания;
- `address` — адрес;
- `number` — количество посадочных мест.

## Загрузка данных и подготовка их к анализу

```
In [1]: import pandas as pd
from matplotlib import pyplot as plt
import seaborn as sns
import plotly.express as px
import warnings

warnings.simplefilter('ignore')

#путь к данным
path = '/home/andrey/anaconda3/data/rest_data.csv'
df = pd.read_csv(path)
display(df.head(5))
df.info()
```

	id	object_name	chain	object_type	address	number
0	151635	СМЕТАНА	нет	кафе	город Москва, улица Егора Абакумова, дом 9	48
1	77874	Родник	нет	кафе	город Москва, улица Талалихина, дом 2/1, корпус 1	35

	id	object_name	chain	object_type	address	number
2	24309	Кафе «Академия»	нет	кафе	город Москва, Абельмановская улица, дом 6	95
3	21894	ПИЦЦЕТОРИЯ	да	кафе	город Москва, Абрамцевская улица, дом 1	40
4	119365	Кафе «Вишневая метель»	нет	кафе	город Москва, Абрамцевская улица, дом 9, корпус 1	50

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 15366 entries, 0 to 15365
Data columns (total 6 columns):
#   Column          Non-Null Count  Dtype
---  -
0    id              15366 non-null  int64
1    object_name     15366 non-null  object
2    chain           15366 non-null  object
3    object_type     15366 non-null  object
4    address         15366 non-null  object
5    number         15366 non-null  int64
dtypes: int64(2), object(4)
memory usage: 720.4+ KB
```

Как видно из информационной таблице пропусков в данных нет. Необходимо проверить на дубли по столбцу `id`. Столбец `chain` должен иметь дискретное значение `нет/да` и столбец `object_type` тоже

## Проверка на дубли

```
In [2]: display(f'Количество дублирующих значений: {df["id"].duplicated().sum()}')
```

'Количество дублирующих значений: 0'

Дублирующихся значений в данном датасете нет.

## Категоризация данных

```
In [3]: for x in ('chain', 'object_type'): display(f'значения колоки {x}:{df[x].unique()}'
```

```
"значения колоки chain:['нет' 'да']"
"значения колоки object_type:['кафе' 'столовая' 'закусочная' 'предприятие б
ыстрого обслуживания'\n 'ресторан' 'кафетерий' 'буфет' 'бар' 'магазин (отде
л кулинарии)']"
```

Колонки `chain` и `object_type` имеют четкие дискретные значения и не нуждаются в предобработки

### Вывод:

`id` - имеет числовые уникальные значения (дублей нет) в предобработке не нуждается.  
`object_name` - проверка не требуется. Может содержать любое значение. Названия могут дублироваться и состоять из любых символов(теоретически)  
`chain` - данные могут принимать два значения (да/нет).Возможно в процессе анализа необходима будет проверка на пропуски  
`object_type` - категория. Имеет четкие определенные значения. Возможно в процессе анализа необходима будет проверка на пропуски  
`address` - проверка не требуется. Может содержать любое значение.  
`number` - принимает целочисленные значения

# Анализ Данных

## Исследование соотношения видов объектов общественного питания по количеству. Построить график.

Лучше всего отражает соотношение долей - круговая диаграмма. Воспользуемся библиотекой `plotly`. Данные графики можно сразу включать в презентацию, а так же они являются интерактивными, что в свою очередь облегчает работу если данные анализируются в Jupyter Notebook

In [4]:

```
#сгруппируем данные по количеству заведений
data_for_graf=df.groupby('object_type')['id'].count().sort_values(ascending=F
display(data_for_graf)
#построим круговую диаграмму
fig = px.pie(data_for_graf,
              values='id',
              names='object_type',
              title='Соотношение видов общественного питания по количеству'
              )
fig.show()
```

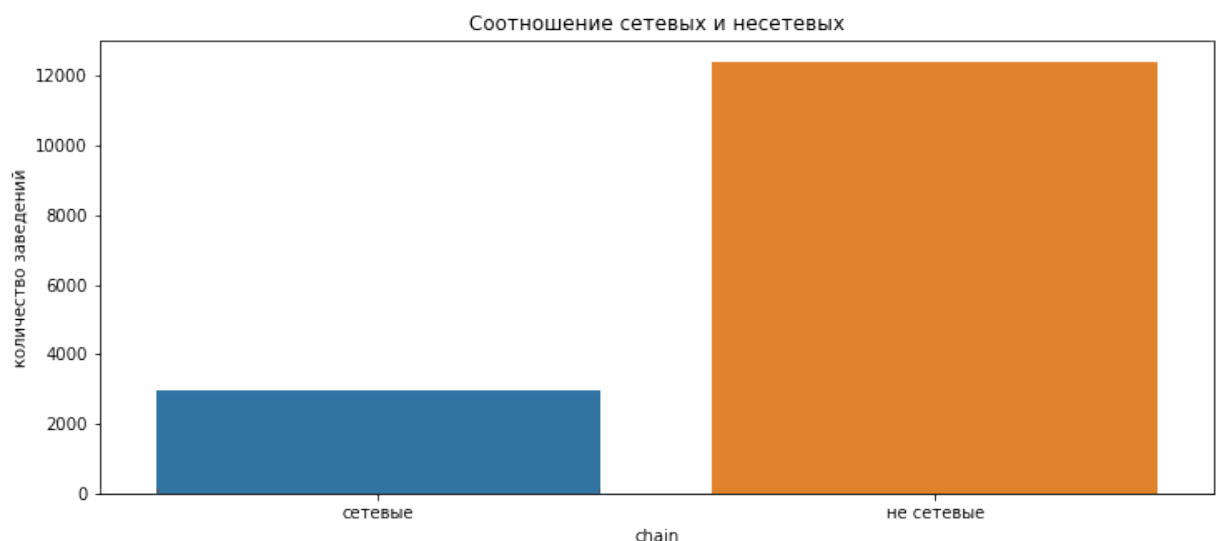
	object_type	id
0	кафе	6099
1	столовая	2587
2	ресторан	2285
3	предприятие быстрого обслуживания	1923
4	бар	856
5	буфет	585
6	кафетерий	398
7	закусочная	360
8	магазин (отдел кулинарии)	273

Соотношение видов общественного питания по количеству

**Вывод:** Больше всего заведений общественного питания в Москве - кафе, далее приблизительно с равными долями столовые с ресторанами

Исследование соотношения сетевых и несетевых заведений по количеству. Построить график.

```
In [5]: data_for_graf = df.groupby(['chain', 'object_type'])['id'].count().reset_index()
data_for_graf = data_for_graf.replace("да", "сетевые").replace("нет", "не сетевые")
plt.figure(figsize=(12, 5))
plt.title('Соотношение сетевых и несетевых')
sns.barplot(x='chain',
            y='id',
            data=data_for_graf,
            estimator=sum,
            ci=None).set_ylabel('количество заведений')
plt.show()
```



**Вывод** Как видно из графика. Основную массу заведений общественного питания составляют не сетевые заведения

Для какого вида объекта общественного питания характерно сетевое распространение?

```
In [6]: data_for_graf.query('chain == "сетевые"]').sort_values(by='id').tail(2)
```

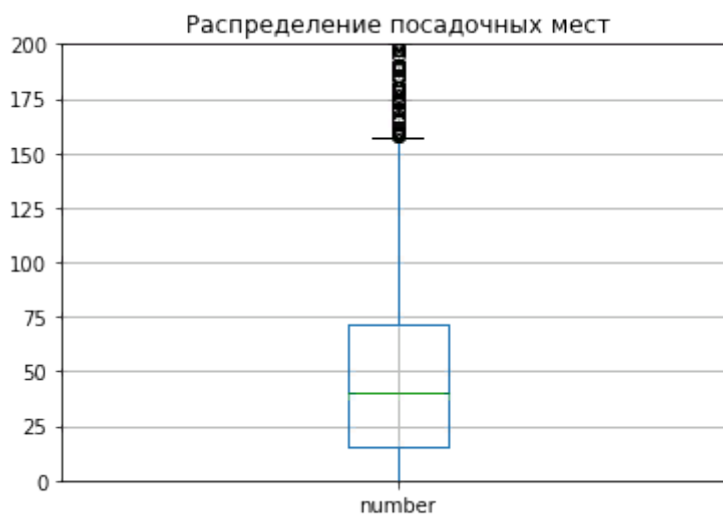
```
Out[6]:
```

	chain	object_type	id
6	сетевые	предприятие быстрого обслуживания	791
3	сетевые	кафе	1396

**Вывод по пунктам 3.2 и 3.3** Подавляющее количество в Москве не сетевых точек общепита, около 12000 на момент формирования данных. Если рассмотреть сетевые заведения, то больше всего кафе, а на втором месте по количеству предприятия быстрого обслуживания (FastFood)

Что характерно для сетевых заведений: много заведений с небольшим числом посадочных мест в каждом или мало заведений с большим количеством посадочных мест?

```
In [7]: data=df.query('chain=="да"')
plt.ylim(0, 200)
data.boxplot(column='number')
plt.title('Распределение посадочных мест')
plt.show()
data['number'].describe()
```



```
Out[7]: count    2968.000000
mean       52.919811
std        57.140644
min         0.000000
25%        15.000000
50%        40.000000
75%        72.000000
max       580.000000
Name: number, dtype: float64
```

Как видно из графика, чаще всего количество мест не превышает 150. При этом медианное значение 40.

Предлагаю разбить на 4-е категории по значениям квантилей:

- **мало** менее 15 посадочных мест
- **средне** от 15 до 40 посадочных мест
- **выше среднего** от 40 до 72 посадочных мест
- **много** выше 72 посадочных мест

```
In [8]: def group_by_number(row):
if row['number'] < 15: return 'мало'
if 15 <= row['number'] < 40: return 'средне'
if 40 <= row['number'] <= 72: return 'выше среднего'
else: return 'много'
```

```
data['category_by_number'] = data.apply(group_by_number, axis=1)
data = data.groupby('category_by_number')['id'].count()
plt.figure(figsize=(12, 5))
sns.barplot(x=data.index, y=data).set_ylabel('количество заведений')
plt.title('Распределение точек общепита по кол-ву посадочных мест')
plt.show()
```



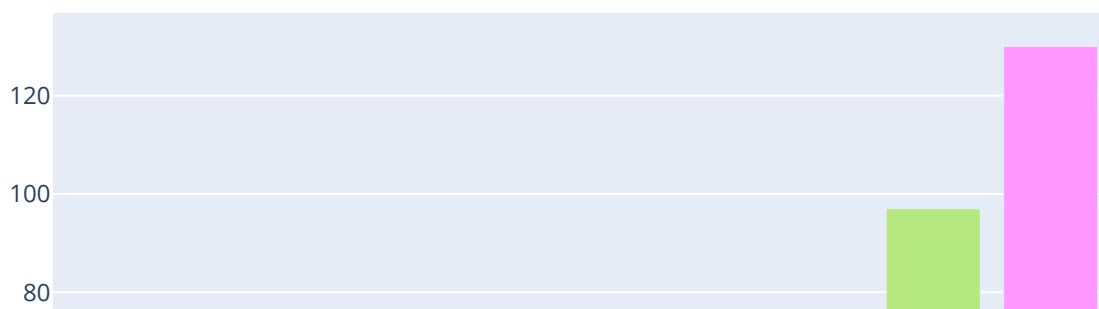
**Вывод:** Как видно из графика выше, если анализировать только сетевые заведения общепита г. Москва, то по количеству посадочных мест они распределены примерно одинаково. Т.е. можно сказать, что все заведения востребованы примерно в равной степени и при открытии новых заведений количество посадочных мест не будет оказывать существенного влияния, если исходить из приведенных выше данных

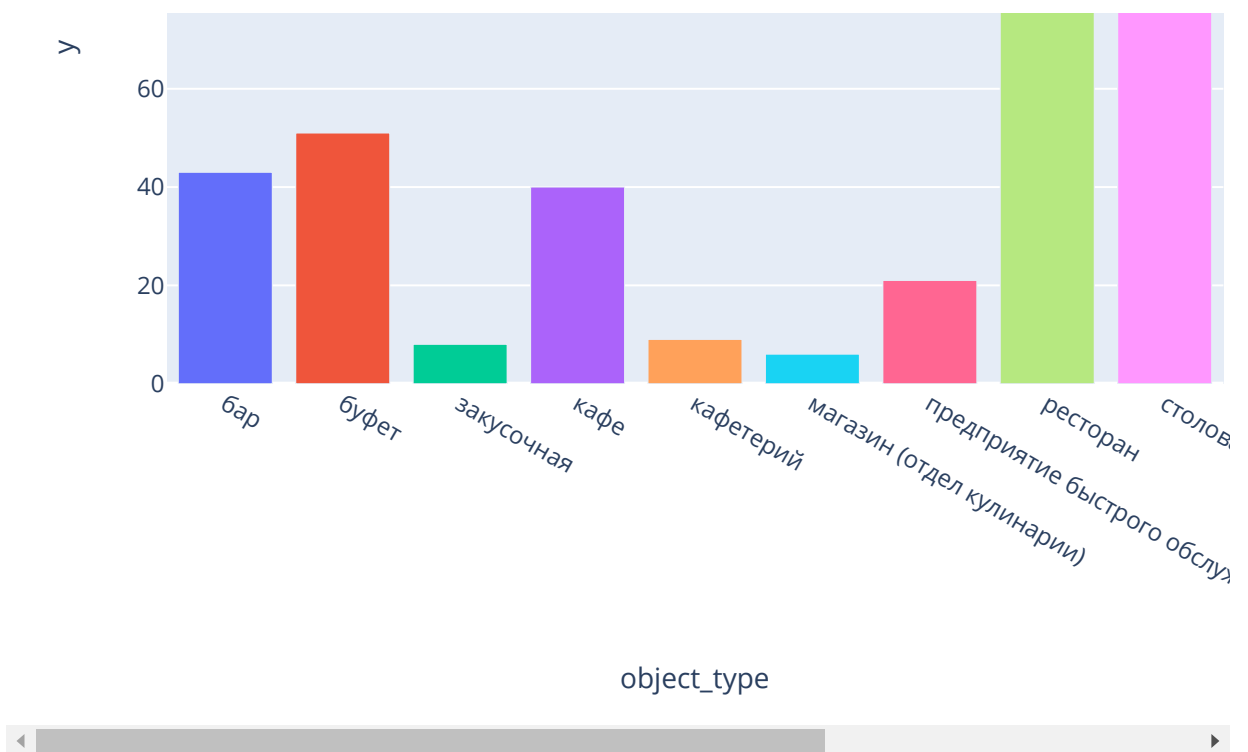
Для каждого вида объекта общественного питания опишите среднее количество посадочных мест. Какой вид предоставляет в среднем самое большое количество посадочных мест? Постройте графики.

In [9]:

```
data_for_graf = df.groupby('object_type')['number'].mean().round()
fig=px.bar(data_for_graf,
            x=data_for_graf.index,
            y=data_for_graf,
            title='Типы общепита по количеству посадочных мест',
            color=data_for_graf.index,
            height=600, width=900)
fig.show()
```

Типы общепита по количеству посадочных мест





**Вывод** Данный график отображает среднее количество посадочных мест. При выборе типа заведения можно опереться на эти данные для определения количества мест. Так например **столовая** должна обладать большим количеством мест, для того чтобы вместить всех желающих в обеденный перерыв, в то же время для кафетерия или закусочной 8-9 мест будет вполне достаточно

**Выделить в отдельный столбец информацию об улице из столбца address**

```
In [10]: # отделяем название города от названия улиц
df['streetname'] = df['address'].apply(lambda x: x.split(',')[1])
# удаляем лишние пробелы
df['streetname'] = df['streetname'].str.strip()
df.head(5)
```

	id	object_name	chain	object_type	address	number	streetname
0	151635	СМЕТАНА	нет	кафе	город Москва, улица Егора Абакумова, дом 9	48	улица Егора Абакумова
1	77874	Родник	нет	кафе	город Москва, улица Талалихина, дом 2/1, корпус 1	35	улица Талалихина
2	24309	Кафе «Академия»	нет	кафе	город Москва, Абельмановская улица, дом 6	95	Абельмановская улица
3	21894	ПИЦЦЕТОРИЯ	да	кафе	город Москва, Абрамцевская улица, дом 1	40	Абрамцевская улица
4	119365	Кафе «Вишневая метель»	нет	кафе	город Москва, Абрамцевская улица, дом 9, корпус 1	50	Абрамцевская улица



Построить график топ-10 улиц по количеству объектов общественного питания. Воспользуйтесь внешней информацией и ответьте на вопрос — в каких районах Москвы находятся эти улицы?

```
In [11]: data_street = df.pivot_table(index='streetname',
                                     values='id',
                                     aggfunc='count').sort_values(by='id', ascending=
data_street
```

```
Out[11]:
```

streetname	id
город Зеленоград	232
проспект Мира	204
Профсоюзная улица	182
Ленинградский проспект	171
Пресненская набережная	167
Варшавское шоссе	161
Ленинский проспект	147
поселение Сосенское	138
проспект Вернадского	132
Кутузовский проспект	114

В полученной выборке есть `город Зеленоград` и `поселение Сосенское`. Удалим эти данные

```
In [12]: data_street = df.query('streetname not in ["город Зеленоград", "поселение Сосенское"]')
data_street = data_street.sort_values(by='id', ascending=False)
```

```
Out[12]:
```

	streetname	id
0	проспект Мира	204
1	Профсоюзная улица	182
2	Ленинградский проспект	171
3	Пресненская набережная	167
4	Варшавское шоссе	161
5	Ленинский проспект	147
6	проспект Вернадского	132
7	Кутузовский проспект	114
8	Каширское шоссе	112
9	Кировоградская улица	110

Построим график отображающий количество заведений общепита на каждой улице из приведенной выше таблицы

```
In [13]: fig=px.bar(data_street,
                  x='streetname',
                  y='id',
                  title='10 лучших улиц по количеству общепита',
                  text='id',
                  color='streetname')
fig.show()
```

### 10 лучших улиц по количеству общепита



```
In [14]: from io import BytesIO
import requests
spreadsheet_id = '1f5FoKhd_WyYzLB3pxs92ht9gdZgSF_jgFkdFW01_z9g'
file_name='https://docs.google.com/spreadsheets/d/{}/export?format=csv'.format(spreadsheet_id)
r=requests.get(file_name)
data_st=pd.read_csv(BytesIO(r.content))
data_st
```

```
Out[14]:
```

	streetname	areaid	okrug	area
0	Выставочный переулок	17	ЦАО	Пресненский район
1	улица Гашека	17	ЦАО	Пресненский район
2	Большая Никитская улица	17	ЦАО	Пресненский район
3	Глубокий переулок	17	ЦАО	Пресненский район
4	Большой Гнездниковский переулок	17	ЦАО	Пресненский район
...	...	...	...	...

	streetname	areaid	okrug	area
4393	Вознесенский проезд	17	ЦАО	Пресненский район
4394	Волков переулок	17	ЦАО	Пресненский район
4395	Поварская улица	17	ЦАО	Пресненский район
4396	Кудринская площадь	17	ЦАО	Пресненский район
4397	Вспольный переулок	17	ЦАО	Пресненский район

4398 rows × 4 columns

```
In [15]: data_street_top = data_street.merge(data_st[['streetname', 'area']], on='stre
```

```
In [16]: #приведем таблицу к более "читаемому" виду
data_street_top = data_street_top.pivot_table(index=['streetname', 'area'])
data_street_top
```

```
Out[16]:
```

	streetname	area	id
Варшавское шоссе		Донской район	161
		Нагорный район	161
		Район Нагатино-Садовники	161
		Район Северное Бутово	161
		Район Чертаново Северное	161
		Район Чертаново Центральное	161
		Район Чертаново Южное	161
		Район Южное Бутово	161
Каширское шоссе		Район Москворечье-Сабурово	112
		Район Нагатино-Садовники	112
		Район Орехово-Борисово Северное	112
		Район Орехово-Борисово Южное	112
Кировоградская улица		Район Чертаново Северное	110
		Район Чертаново Центральное	110
		Район Чертаново Южное	110
Кутузовский проспект		Район Дорогомилово	114
		Район Фили-Давыдково	114
Ленинградский проспект		Район Аэропорт	171
		Район Беговой	171
		Район Сокол	171
		Хорошевский район	171
Ленинский проспект		Донской район	147
		Ломоносовский район	147
		Обручевский район	147

streetname	area	id
	Район Гагаринский	147
	Район Проспект Вернадского	147
	Район Теплый Стан	147
	Район Тропарево-Никулино	147
	Район Якиманка	147
Пресненская набережная	Пресненский район	167
Профсоюзная улица	Академический район	182
	Обручевский район	182
	Район Коньково	182
	Район Теплый Стан	182
	Район Черемушки	182
проспект Вернадского	Район Ясенево	182
	Ломоносовский район	132
	Район Гагаринский	132
	Район Проспект Вернадского	132
	Район Раменки	132
проспект Мира	Район Тропарево-Никулино	132
	Алексеевский район	204
	Мещанский район	204
	Останкинский район	204
	Район Марьино	204
	Район Ростокино	204
	Район Свиблово	204
	Ярославский Район	204

**Вывод** Мы получили таблицу для инвесторов с информацией на каких улицах и в каком районе наиболее сосредоточены точки общепита. Также выяснили по каким районам Москвы проходят данные улицы. Если предполагаемая точка имеет какое-либо существенное конкурентное преимущество, то ее выгоднее будет открыть на оживленных улицах, где есть спрос. Скорее всего на этих улицах уже находятся много открытых точек (потенциальных конкурентов). Далее остается выбрать только наиболее удобный район.

Найти число улиц с одним объектом общественного питания. Воспользуйтесь внешней информацией и ответьте на вопрос — в каких районах Москвы находятся эти улицы?

```
In [17]: data_street_one = df.pivot_table(index='streetname',
                                     values='id',
                                     aggfunc='count').query('id == 1').reset_index()
display(f'Число улиц с одним заведением общепита: {data_street_one.shape[0]}')
```

'Число улиц с одним заведением общепита: 579'

In [18]:

```
data_street_one = data_street_one.merge(data_st[['streetname', 'area']], on='streetname')
display(data_street_one)
data_street_one_are = data_street_one.groupby('area')['streetname'].count().sort_values(ascending=False)
```

	streetname	id	area
0	1-й Балтийский переулок	1	Район Аэропорт
1	1-й Басманный переулок	1	Басманный район
2	1-й Басманный переулок	1	Красносельский район
3	1-й Ботанический проезд	1	Район Свиблово
4	1-й Вешняковский проезд	1	Рязанский район
...	...	...	...
617	улица Шеногина	1	Район Хорошево-Мневники
618	улица Шумилова	1	Район Кузьминки
619	улица Шумкина	1	Район Сокольники
620	улица Щорса	1	Район Солнцево
621	шоссе Фрезер	1	Нижегородский район

622 rows × 3 columns

Мы получили таблицу содержащую 622 записи. Некоторые районы включают в себя несколько улиц с 1 заведением общепита. Инвестору скорее всего пригодится список районов с большим количеством "проблемных" улиц. Для выбора порога отсечения по количеству улиц, посмотрим на распределение данного количества

In [19]:

```
data_street_one_are[['streetname']].describe()
```

Out[19]:

```
count    99.000000
mean      5.656566
std       5.637270
min       1.000000
25%       2.000000
50%       4.000000
75%       6.500000
max      27.000000
Name: streetname, dtype: float64
```

In [20]:

```
# 4 улицы является медианной. Предлагаю взять этот показатель за порог отсечения
data_street_one_are.query('streetname > 4')['area'].unique()
```

Out[20]:

```
array(['Таганский район', 'Район Хамовники', 'Басманный район',
      'Тверской район', 'Пресненский район', 'Район Марьино',
      'Район Арбат', 'Мещанский район', 'Район Сокольники',
      'Район Замоскворечье', 'Район Лефортово', 'Даниловский район',
      'Район Соколиная Гора', 'Район Якиманка', 'Красносельский район',
      'Район Богородское', 'Нижегородский район', 'Район Люблино',
      'Можайский район', 'Район Южное Бутово', 'Район Аэропорт',
      'Район Преображенское', 'Тимирязевский район',
      'Алексеевский район', 'Лосиноостровский район',
      'Дмитровский район', 'Район Измайлово', 'Район Внуково',
      'Донской район', 'Район Хорошево-Мневники', 'Район Кунцево',
      'Савеловский район', 'Район Солнцево', 'Район Щукино',
```

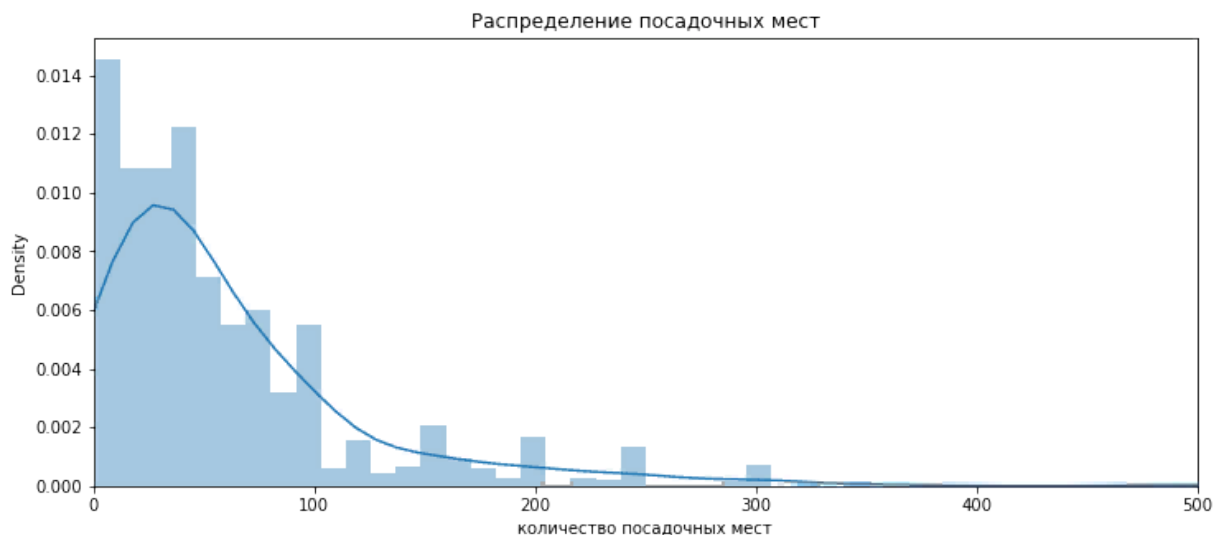
```
'Головинский район', 'Район Южное Тушино', 'Район Ясенево',  
'Рязанский район', 'Район Ростокино', 'Бабушкинский район',  
'Район Выхино-Жулебино', 'Хорошевский район', 'Южнопортовый Район'],  
dtype=object)
```

**Вывод** В ходе исследования мы получили список районов в которых есть большое количество улиц на которых размещается 1 объект общепита. При планировании открытия точки в таком районе, необходимо тщательно проанализировать место и причины малого количества конкурентов.

Посмотрите на распределение количества посадочных мест для улиц с большим количеством объектов общественного питания. Какие закономерности можно выявить?

In [21]:

```
# data_street['streetname'] - содержит информацию о 10 улицах с самым большим  
filt = df['streetname'].isin(data_street['streetname'])  
# создадим датафрейм с улицами из data_street['streetname'] и удалим данные с  
df_filt = df[filt].query('number != 0')  
plt.figure(figsize=(12, 5))  
plt.xlim(0, 500)  
plt.title('Распределение посадочных мест')  
sns.distplot(df_filt['number'], bins=150).set_xlabel('количество посадочных мест')  
plt.show()
```



**Вывод** По графику распределения видно, что больше всего в городе Москва заведений с количеством мест 0 - 10. Также довольно много заведений от 10 до 50 посадочных мест. Если ориентироваться на конкурентов (уже открытые заведения), то стоит обратить внимание на заведения от 0 - 50 посадочных мест, судя по всему они являются наиболее экономически эффективными

## Общий вывод

После анализа информации об заведениях общественного питания г. Москва, можно сделать следующий вывод. Но перед выводами нужно учитывать, что мы анализировали конкурентов (уже открытые заведения) и в выводах оперируем на поведение среднестатистического владельца заведения. Если проект будущей точки не типовой и имеет ряд конкурентных преимуществ (выделяется из общей массы объектов общепита) то приведенные выводы носят рекомендательный характер и могут оказать/не оказать влияние на работу точки

1. Основным типом заведения является **кафе** . На него приходится почти 40% количества заведений. На втором и третьем месте с приблизительно с одинаковым количеством **столовая** и **ресторан**
2. Основную массу составляют не сетевые заведения. Что обусловлено скорее всего суммой начальных инвестиций. Большинство кафе и столовые относятся к категории малого бизнеса с относительно небольшими стартовыми вложениями
3. Если рассматривать сетевые заведения, то основную массу составляют кафе и предприятия быстрого питания. Если при открытии точки рассматривается возможность расширения до сети, то оптимальный вариант это кафе. Данный тип имеет большую долю как в сетевых так и не сетевых заведениях
4. Анализ сетевых точек по количеству посадочных мест (также этот анализ можно рассмотреть как анализ площади заведения) показывает, что сети мало ориентируются на площадь зала. В Москве представлены сети приблизительно с одинаковым количеством заведений по посадочным местам. Из доли приблизительно одинаковы. Т.е. если начинать открывать свою сеть, то количество посадочных мест не будет иметь ключевое значение.
5. Больше всего площади и количества посадочных мест необходимо **столовым** . для них характерно большое количество посетителей в обеденные перерывы. Если рассматривать **кафе** , то оптимальным будет 9 мест (по анализе конкурентов)
6. В исследовании получен список улиц на которых располагаются больше всего заведений общепита и приведена таблица разделения по районам. Зачастую это длинные улицы и проходят в нескольких районах. Открыть на такой улице точку - хорошая идея, т.к. судя по всему они обладают хорошей проходимостью, но надо быть готовым к значительной конкуренции и необходимо проанализировать все районы по которым проходит улица. Распределение точек общепита может быть не равномерным, а тяготеть к местам наибольшего скопления людей. (ТЦ, ТРЦ, Метро и т.п.)
7. Также получен список наименее популярных улиц. На них находятся только по 1-му заведению. Для удобства выбора расположения получен список районов, в которые входят несколько таких улиц. При желании открыть точку в таком районе нужно более тщательно анализировать риски и месторасположение нового заведения
8. Наиболее часто в Москве встречаются заведения с 10 посадочными местами **Резюме**  
Исходя из выше перечисленного рекомендую открывать **кафе** . Это заведения потребует незначительных финансовых вложений на старте, требует небольшую площадь (0 - 10 посадочных мест), может быть перепрофилировано в сетевое заведение и пользуется значительной популярностью (открыто больше всего подобных заведений). Расположение можно выбрать опираясь на пункты 3.7 и 3.8 исследований.  
При наличии более большого стартового капитала можно рассмотреть **столовую** или **ресторан** .

Презентация: [https://drive.google.com/file/d/1b6wLMxvx0WHEET8SEnyK3SntjIEK8yb\\_/view?usp=sharing](https://drive.google.com/file/d/1b6wLMxvx0WHEET8SEnyK3SntjIEK8yb_/view?usp=sharing)