In [1]:

```
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd
import scipy
import plotly.graph_objects as go
import plotly.express as px
from plotly.subplots import make_subplots

sns.set(font_scale=1.6, palette='summer')
```

Исследование предпочтений молодежи в отношении банков

Целью моего исследования было понять, какими банковскими услугами и банками пользуются люди в возрасте от 18 до 26 лет и, в итоге, составить портрет типичного представителя для этой возрастной группы. Сбор данных осуществлялся через гугл формы. Опрос содержал вопросы о возрасте, источниках дохода, используемом банке, использовании кредитных карт и количестве используемых карт.

Этот опрос может быть полезен, очевидно, банкам при создании новых продуктов или продвижении уже существующих.

Всего было получено около 240 ответов, но часть из них была нерелевантными и была отброшена.

In [2]:

```
survey = pd.read_csv('data/poll.csv').drop(columns=['Unnamed: 0'])
survey.columns = ['time', 'age', 'income', 'bank_name', 'use_credit', 'card_cnt']

survey.head()
```

Out[2]:

	time	age	income	bank_name	use_credit	card_cnt
0	2020/03/24 11:00:28 PM GMT+3	21-23	Зарплата;Финансовая поддержка от родителей, др...	Сбербанк;ВТБ	Пользуюсь, буду продолжать	3.0
1	2020/03/24 11:06:12 PM GMT+3	18-20	Зарплата;Финансовая поддержка от родителей, др...	Сбербанк	Пользуюсь, буду продолжать	1.0
2	2020/03/24 11:11:20 PM GMT+3	24-26	Зарплата;Доход от инвестиций (вклады, ценные б...	Сбербанк;ВТБ;Альфа-Банк;РокетБанк	Пользуюсь, буду продолжать	4.0
3	2020/03/24 11:18:01 PM GMT+3	18-20	Доход от инвестиций (вклады, ценные бумаги и т...	Сбербанк;Тинькофф;ВТБ	Не пользуюсь, не хочу открывать	5.0
4	2020/03/24 11:25:07 PM GMT+3	18-20	Финансовая поддержка от родителей, других родс...	Сбербанк	Не пользуюсь, не хочу открывать	0.0

In [3]:

```
survey.isna().sum()
```

Out[3]:

```
time          0
age           2
income        2
bank_name     2
use_credit    2
card_cnt      2
dtype: int64
```

Приведем данные в приемлемый вид. Просто уберем Nan, тк их немного, отформатируем ячейки, удалим людей не из интересуемой возрастной группы и тд. Добавим новые признаки

In [4]:

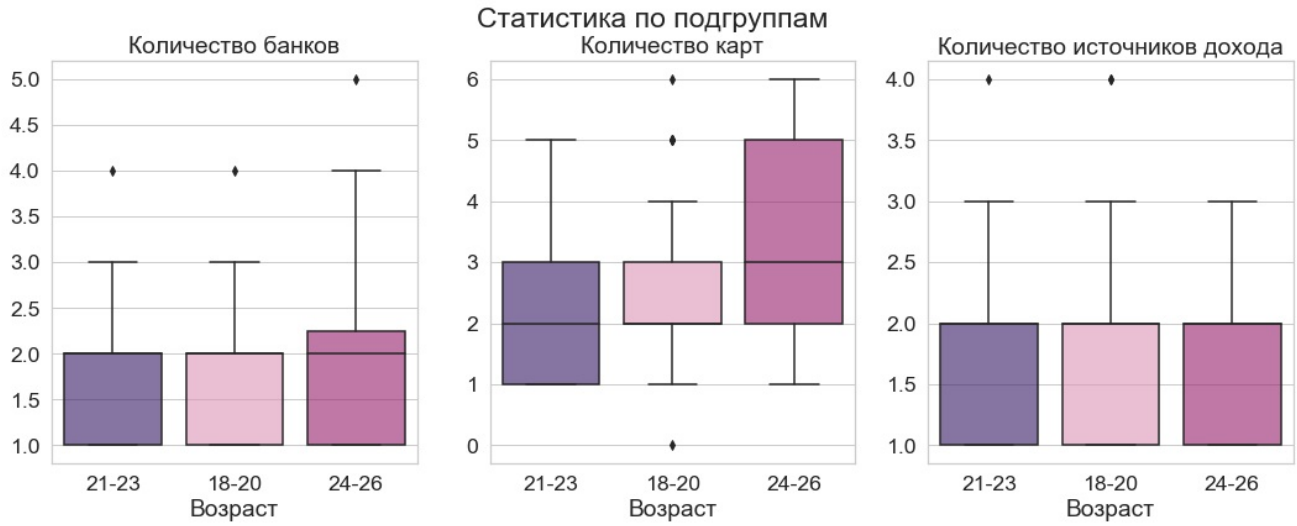
```
survey.dropna(inplace=True)

survey = survey[survey['age'].apply(lambda x: x not in ['больше', 'меньше'])]
survey['card_cnt'] = survey['card_cnt'].astype('int')

survey['bank_cnt'] = survey['bank_name'].apply(lambda x: len(x.split(';')))
survey['income_cnt'] = survey['income'].apply(lambda x: len(x.split(';')))

survey.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 186 entries, 0 to 236
Data columns (total 8 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   time        186 non-null    object
 1   age         186 non-null    object
 2   income      186 non-null    object
 3   bank_name   186 non-null    object
 4   use_credit  186 non-null    object
 5   card_cnt    186 non-null    int32 
 6   bank_cnt    186 non-null    int64 
 7   income_cnt  186 non-null    int64 
dtypes: int32(1), int64(2), object(5)
memory usage: 12.4+ KB
```

Посмотрим на статистические метрики для разных столбцов

In [5]:

```
plt.figure(figsize=(18, 6))

names = ("Количество банков","Количество карт", "Количество источников дохода")
cat_names = ['bank_cnt', 'card_cnt', 'income_cnt']

colors = ['#330C73', '#EB89B5', '#AA0C73']

plt.suptitle('Статистика по подгруппам')

for ax, category, name in zip(range(1, 4), cat_names, names):
    with sns.axes_style("whitegrid"):
        plt.subplot('13{}'.format(ax))
        ax = sns.boxplot(x='age', y=category, data=survey, palette=sns.color_palette(colors))
        for patch in ax.artists:
            r, g, b, a = patch.get_facecolor()
            patch.set_facecolor((r, g, b, .6))
        plt.title(name)
        plt.xlabel('Возраст')
        plt.ylabel('')
```



In [6]:

```
fig = px.histogram(
    survey,
    x="use_credit",
    title='Отношение к кредитным картам',
    color='age',
    color_discrete_map={'18-20': '#330C73','21-23': '#EB89B5','24-26': '#AA0C73'},
    opacity=0.75
)
fig.show()
```

Вывод:

Большинство молодых людей не хотят использовать кредитные карты, а те, кто уже используют, не собираются отказываться. Следовательно, люди не часто переходят из одной категории в другую.

Всего получилось 186 валидных участников. Исследуем данные

In [7]:

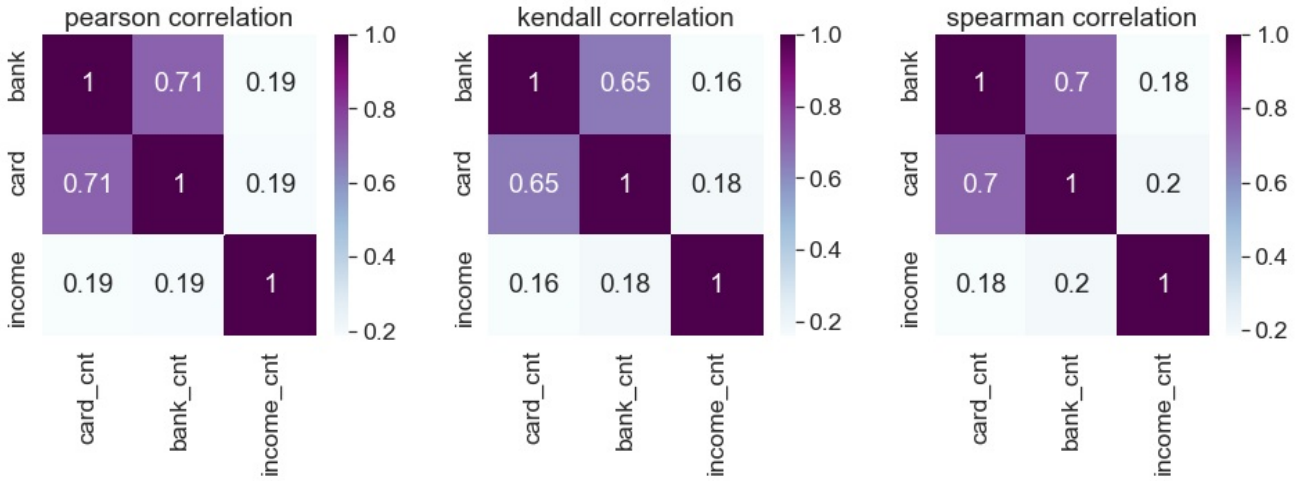
```
fig = make_subplots(
    rows=1, cols=3,
    subplot_titles=names,
    shared_yaxes=True,
)

colors = ['#330C73', '#EB89B5', '#AA0C73']

for idx, name in zip(range(1, 4), ['bank_cnt', 'card_cnt', 'income_cnt']):
    for age, color in zip(sorted(set(survey.age)), colors):
        if name != 'bank_cnt':
            show_leg = False
        else:
            show_leg=True
        fig.add_trace(go.Histogram(
            x=survey[survey['age'] == age][name],
            histnorm='percent',
            name=age,
            xbins=dict(start=-1, end=6, size=1),
            marker_color=color,
            opacity=0.75,
            showlegend=show_leg
        ),
        1, idx)

fig.update_layout(
    height=400, width=900,
    legend_title_text='Возраст',
    yaxis_title_text='Проценты',
    bargap=0.2,
    bargroupgap=0.1,
    xaxis = dict(tickmode = 'linear', tick0 = -1, dtick = 1),
)

fig.show()
```

Вывод

Видим, что возрастные группы похожи друг на друга, за исключением количесства используемых карт. Более старшая группа чаще использует больше карт.

In [8]:

```
plt.figure(figsize=(16,4))

#names = ("Количество банков","Количество карт", "Количество источников дохода")
names = ('pearson', 'kendall', 'spearman')
cat_names = ['bank', 'card', 'income']


plt.subplots_adjust(wspace=0.3, hspace=0.7)

for ax, category, name in zip(range(1, 4), cat_names, names):
    plt.subplot('13{}'.format(ax))

    cor = survey.corr(method=name)

    sns.heatmap(cor, annot=True, cmap="BuPu")
    plt.title('{} correlation'.format(name))
    locs, labels = plt.yticks()            # Get locations and labels

    plt.yticks(locs - 0.3, cat_names)

plt.show()
```



Вывод:

Видим, что количество источников дохода не сильно коррелирует с количеством карт или банков.

Количесвто клиентов, которые обслуживаются в одном банке и имеют более одной карты

In [9]:

```
tmp = survey[survey['bank_cnt'] == 1]
tmp[tmp['card_cnt'] > 1]['bank_name'].value_counts()
```

Out[9]:

```
Сбербанк    24
Другой      12
Тинькофф     1
Name: bank_name, dtype: int64
```

Вывод

Хммм, почему-то часто бывает, что человек имеет несколько карт сбербанка. Вероятно это вроисходит из-за того, что Сбербанк часто выдает степендиальные/зарплатные карты, которые являются вторыми, третьими и тд.

Перейдем непосредственно к составлению портрета

Сделаем категориальные признаки, чтобы понять самые популярные варианты

Достанем названия банков и источник дохода

In [10]:

```
income_cat = sorted(set(';'.join(survey['income']).split(';')))
col_income = ['investments', 'other', 'salary', 'scolarship', 'parents']

for cat_name, cat in zip(col_income, income_cat):
    survey[cat_name] = survey['income'].apply(lambda x: cat in x)

banks = sorted(set(';'.join(survey['bank_name']).split(';')))
col_banks = ['alpha', 'vtb', 'other_bank', 'raiffeisen', 'rocket', 'sberbank', 'tinkoff']

for bank_name, bank in zip(col_banks, banks):
    survey[bank_name] = survey['bank_name'].apply(lambda x: bank in x)

agregated = survey.groupby('age').sum()[col_income + col_banks]

agregated
```

Out[10]:

age	investments	other	salary	scolarship	parents	alpha	vtb	other_bank	raiffeisen	rocket	sberbank	tinkoff
18-20	9.0	18.0	33.0	77.0	100.0	5.0	42.0	54.0	3.0	3.0	89.0	28.0
21-23	1.0	5.0	27.0	20.0	23.0	4.0	5.0	16.0	1.0	1.0	35.0	9.0
24-26	7.0	2.0	19.0	1.0	5.0	6.0	4.0	4.0	1.0	2.0	17.0	8.0

In [12]:

```
agregated['credit_use'] = np.nan
agregated['card_cnt'] = np.nan

for age in agregated.index:
    agregated.loc[age, 'credit_use'] = survey[survey['age'] == age]['use_credit'].value_counts().index[0]
    agregated.loc[age, 'card_cnt'] = survey[survey['age'] == age]['card_cnt'].value_counts().index[0]
```

In [30]:

```
fig = make_subplots(rows=1, cols=3, specs=[[
    {'type':'domain'}, {'type':'domain'}, {'type':'domain'}
]])

for age, idx in zip(agregated.index, range(1, 4)):
    fig.add_trace(
        go.Pie(labels=agregated.loc[age][col_banks].index, values=agregated.loc[age][col_banks].values, name=age)
,
        1, idx
    )

fig.update_traces(hole=.4, hoverinfo="label+percent+name")

fig.update_layout(
    height=400, width=900,
    title_text="Популярность банков среди разных подгрупп",
    annotations=[dict(text='18-20', x=0.1, y=0.5, font_size=20, showarrow=False),
                 dict(text='21-23', x=0.5, y=0.5, font_size=20, showarrow=False),
                 dict(text='24-26', x=0.905, y=0.5, font_size=20, showarrow=False)
                ])
fig.show()
```

Вывод:

Как и ожидалось, самый популярный банк это Сбербанк. Также достаточно популярны Тинькофф, ВТБ. В подгруппе 24-26 14.3% пользуются Альфа-банком

In [27]:

```
agregated.loc['21-23'][col_income]
```

Out[27]:

```
investments     1
other           5
salary         27
scolarship     20
parents        23
Name: 21-23, dtype: object
```

In [28]:

```
agregated.loc['21-23'][col_income].index.values
```

Out[28]:

```
array(['investments', 'other', 'salary', 'scolarship', 'parents'],
      dtype=object)
```

In [25]:

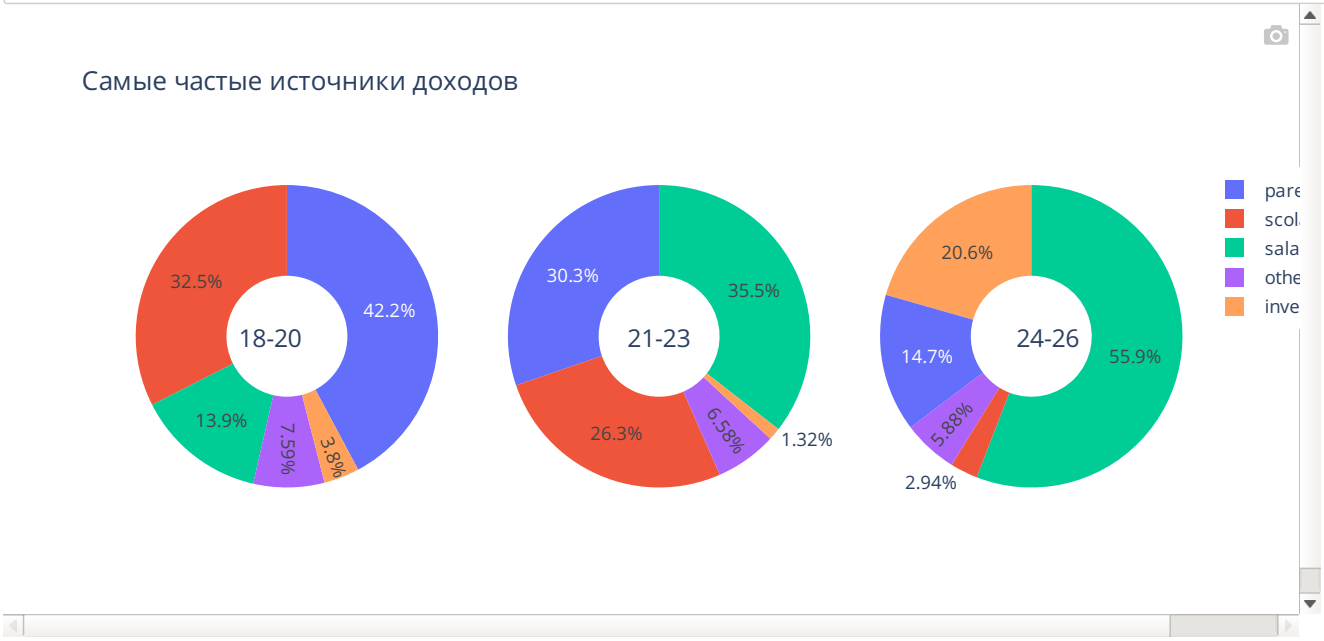
```
income_cat
```

Out[25]:

```
['Зарплата',
 'Финансирование от родителей',
 'Стипендия',
 'Инвестиции',
 'Другое']
```

In [39]:

```
fig = make_subplots(rows=1, cols=3, specs=[[
    {'type':'domain'}, {'type':'domain'}, {'type':'domain'}
]])

income_cat = ['Зарплата', 'Финансирование от родителей', 'Стипендия', 'Инвестиции', 'Другое']

for age, idx in zip(agregated.index, range(1, 4)):
    fig.add_trace(
        go.Pie(values=agregated.loc[age][col_income].values, labels=agregated.loc[age][col_income].index,
               name=age),
        1, idx
    )

fig.update_traces(hole=.4, hoverinfo="label+percent+name")

fig.update_layout(
    height=400, width=900,
    title_text="Самые частые источники доходов",
    annotations=[dict(text='18-20', x=0.095, y=0.5, font_size=16, showarrow=False),
                 dict(text='21-23', x=0.5, y=0.5, font_size=16, showarrow=False),
                 dict(text='24-26', x=0.905, y=0.5, font_size=16, showarrow=False)
                ])
fig.show()
```



Вывод

Видно, что чем старше человек, тем реже он получает деньги от родителей, и тем чаще он получает зарплату.

In [20]:

```
result = pd.concat(
    (
        agregated[['credit_use', 'card_cnt']],
        agregated[col_income].idxmax(axis=1),
        agregated[col_banks].idxmax(axis=1),
        agregated[set(col_banks) - set(['sberbank', 'other_bank'])].idxmax(axis=1),
    ), axis=1)

result.columns = [
    'Использование кредитной кары',
    'Количество карт',
    'Источник дохода',
    'Самый популярный банк',
    'Второй самый популярный банк'
]

result
```

Out[20]:

age	Использование кредитной кары	Количество карт	Источник дохода	Самый популярный банк	Второй самый популярный банк
18-20	Не пользуюсь, не хочу открывать	2.0	parents	sberbank	vtb
21-23	Не пользуюсь, не хочу открывать	2.0	salary	sberbank	tinkoff
24-26	Не пользуюсь, не хочу открывать	3.0	salary	sberbank	tinkoff

Результат исследования

Получили портрет типичного представителя по разным подгруппам. В целом все результаты закономерны.

- Можно заметить, что ни одна подгруппа не хочет пользоваться кредитными картами. Скорее всего так происходит, потому что пока ни у кого нет необходимости в кредитных деньгах или же нет стадибльного заработка, чтобы эту кредитную карту получить.
- Также обычно молодые люди используют 2 или 3 карты. Вероятно, это можно объяснить тем, что многим выдавались степендивальные карты МИР, или некоторые оформляли социальную карту.
- Самый популярный банк очевидно Сбербанк, но если посмотреть на второй по популярности, то для группы 18-20 это ВТБ. Так могло произойти из-за социальной карты, её можно получить только в ВТБ и Банке Москвы. В более старших группах виден приоритет Тинькофф банка. Скорее всего люди стали клиентами этого банка по собственному желанию отличие от ситуаций, описанных выше. Вероятно этот банк лучше понимает, что необходимо молодежи, чем остальные.
- С источниками дохода всё понятно. На младших курсах у многих нет возможности совмещать учебу и работу.