Задание 1.

1. Необходимо выбрать метрику и привести аргументацию.

По условию задачи целевая метрика равна какой-то из колонок, и тест заключается в том, чтобы проверить увеличение дохода. За доход как раз отвечает колонка NPV, поэтому метрика — среднее NPV.

In [ ]:
```
import csv
import pandas as pd
df = pd.read_csv('hist_telesales.csv',
                 index_col='ID',
                       sep=',')
df
```

Out[ ]:

ID	Флаг дозвона	Флаг продажи	Расходы	PV	NPV
0	1	0	90	0	-90
1	0	0	5	0	-5
2	0	0	68	0	-68
3	1	0	22	0	-22
4	1	0	22	0	-22
...	...	...	...	...	...
72156	1	1	577	1346	769
72157	0	0	8	0	-8
72158	0	0	23	0	-23
72159	0	0	4	0	-4
72160	1	1	132	1385	1253

72161 rows × 5 columns

2. Альтернатива в критерии.

Альтернатива H_1 — уменьшение цены продукта позволит суммарно увеличить доходность продукта. То есть $H_1 : NPV_{\text{тест}} > NPV_{\text{контроль}}$

3.1. Каков размер выборки? Привести аргументацию и написать как получилось то или иное число.

In [ ]:
```
from statistics import pvariance
import scipy.stats

a = 0.05
b = 0.2
sigma2 = pvariance(df.NPV.tolist())
z_1_a = scipy.stats.norm.ppf(1 - a)
z_b = scipy.stats.norm.ppf(b)
k = 0.5
MDE = 0.05 * df.NPV.mean()
N = sigma2 * (z_1_a - z_b)**2 / (k * MDE**2)
N
```

Out[ ]:
```
28671.429203969576
```

следовательно, минимальное N равно 28671

4. Принятие решения. Расписать подробно с аргументами.

Так как нужно сравнить средние, стоит выбор между применением критерия Стьюдента и критерием Манна-Уитни. Из документации к scipy.stats.mannwhitneyu: *The Mann-Whitney U test is a non-parametric version of the t-test for independent samples. When the means of samples from the populations are normally distributed, consider scipy.stats.ttest_ind.*

Поэтому проверим, являются ли обе выборки нормально распределёнными. Посмотрим на график:

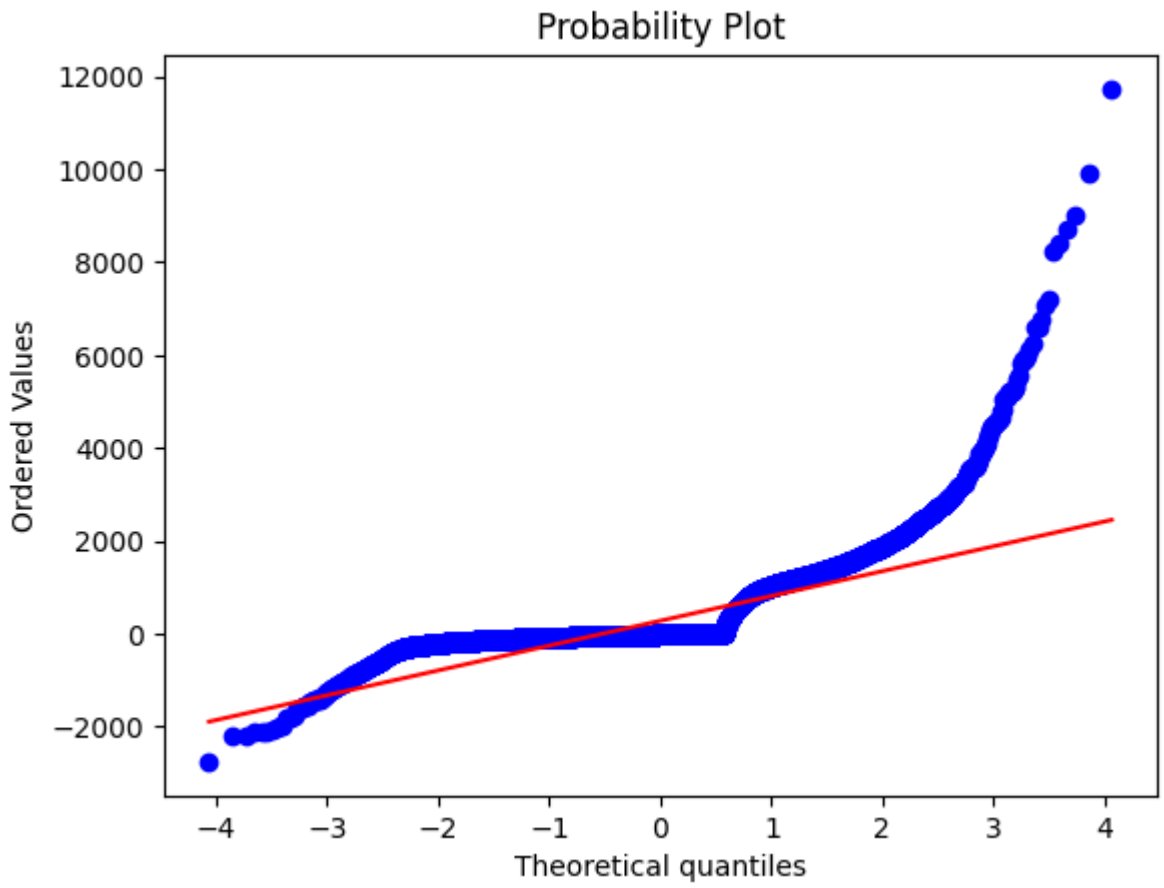
In [ ]:
```
import pylab

df_control = pd.read_csv('Контроль.csv',
                 index_col='ID',
                       sep=',')

df_test= pd.read_csv('Тест.csv',
                 index_col='ID',
```

```
                        sep=',')

scipy.stats.probplot(df_test.NPV.to_numpy(), dist="norm", plot=pylab)
pylab.show()
```



Действительно, тестовая выборка не имеет нормальное распределение, поэтому используем критерий Манна Уитни

In [ ]:
```
from scipy.stats import mannwhitneyu

mannwhitneyu(df_control.NPV.tolist(), df_test.NPV.tolist(),
             alternative = 'greater').pvalue
```

Out[ ]:
```
0.917100825637646
```

pvalue > 0.05, значит, мы не можем отвергнуть нулевую гипотезу теста. Следовательно, мы не можем утверждать, что уменьшение цены продукта позволит суммарно увеличить доходность продукта.

Задание 2.

1. Необходимо выбрать метрику и привести аргументацию.

Целевая метрика равна какой-то из колонок по условию задачи. Целью теста является увеличение доходности услуги, поэтому целевой метрикой является NPV.

In [ ]:
```
df = pd.read_csv('hist_credit_card.csv',
                 index_col='ID',
                       sep=',')
df
```

Out[ ]:

ID	Возраст	Доход клиента	Вероятность банкротства	Флаг утилизации счёта	Расходы	PV КК	PV услуги	NPV
0	19	21620.835463	0.138061	0	102	0	0	-102
1	27	24897.990863	0.035508	1	409	11686	1754	13031
2	50	23989.526947	0.098793	0	16	0	0	-16
3	18	38442.409756	0.365661	1	788	13738	1578	14528
4	24	21291.521612	0.036909	1	1048	6594	2213	7759
...	...	...	...	...	...	...	...	...
123250	32	26099.633927	0.110756	0	47	0	0	-47
123251	20	24579.749275	0.113920	1	594	14268	1672	15346
123252	36	34062.902531	0.247122	1	77	5950	2017	7890
123253	67	24609.838522	0.020752	1	279	6278	1847	7846
123254	36	23378.281337	0.092221	1	186	7093	6109	13016

123255 rows × 8 columns

2. Альтернатива в критерии.

Альтернатива H_1 — увеличение стоимости продукта позволит суммарно увеличить доходность продукта. То есть $H_1 : NPV_{\text{тест}} > NPV_{\text{контроль}}$

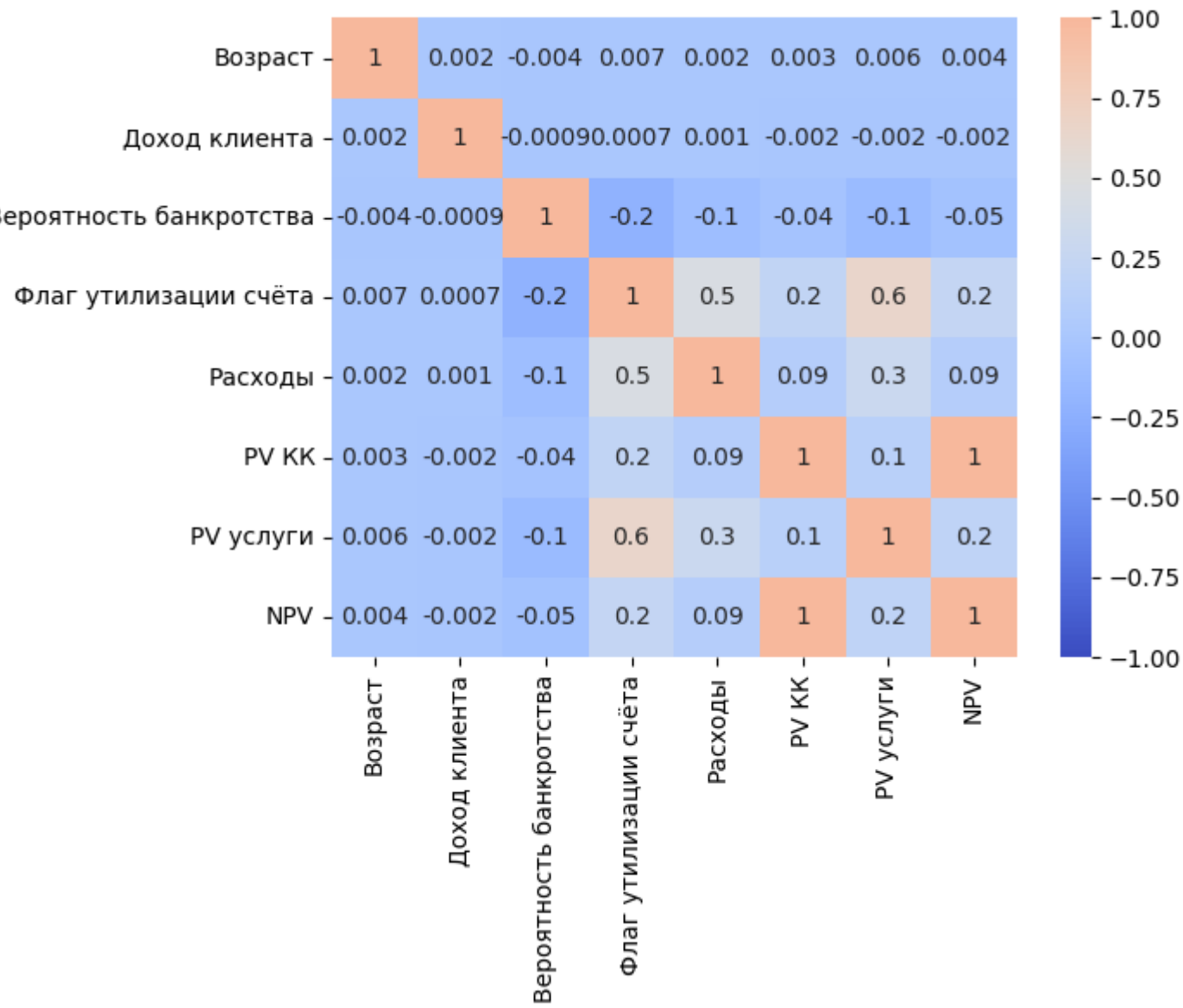
3. Нужно выбрать параметр(-ы), влияющий(-ие) на целевую метрику. Привести аргументацию.

Посчитаем коэффициенты корреляции:

In [ ]:
```
import seaborn as sns

sns.heatmap(df.corr(), annot=True, fmt='.1g', vmin=-1, vmax=1, center=0.5,
            cmap='coolwarm')
```

Out[ ]:
```
<Axes: >
```



Коэффициент корреляции для NPV и всех параметров, кроме PV KK, близок к 0, а значит эти параметры на целевую метрику практически не влияют. Корреляция между NPV и PV KK равна 1, поэтому на целевую метрику влияет параметр PV KK.

4.1. Каков размер выборки? Привести аргументацию и написать как получилось то или иное число.

Воспользуемся формулой из лекции и условием задачи:

In [ ]:
```
a = 0.05
b = 0.1
sigma2 = pvariance(df.NPV.tolist())
z_1_a = scipy.stats.norm.ppf(1 - a)
z_b = scipy.stats.norm.ppf(b)
k = 0.5
MDE = 0.08 * df.NPV.mean()
N = sigma2 * (z_1_a - z_b)**2 / (k * MDE**2)
N
```

Out[ ]:
```
17290.572625567944
```

Следовательно, минимально допустимый размер выборки равен 17290

5. Проверка на однородность, применение критерия. Принятие решения. Расписать подробно с аргументами.

Для проверки на однородность можем выбрать метрику, коррелирующую с целевой метрикой — среднее PV KK. Нужно проверить, имеют ли выборки на тесте и на контроле одинаковое распределение. Используем критерий Андерсона:

In [ ]:
```
from scipy import stats

df_test = pd.read_csv('Тест1.csv',
                      index_col='ID',
                            sep=',')
df_control = pd.read_csv('Тест1.csv',
                      index_col='ID',
                            sep=',')
stats.anderson_ksamp([df_test['PV KK'],
                      df_control['PV KK']]).pvalue
```

```
<ipython-input-88-5a6d9c62a787>:9: UserWarning: p-value capped: true value larger than 0.25
  stats.anderson_ksamp([df_test['PV KK'],
```

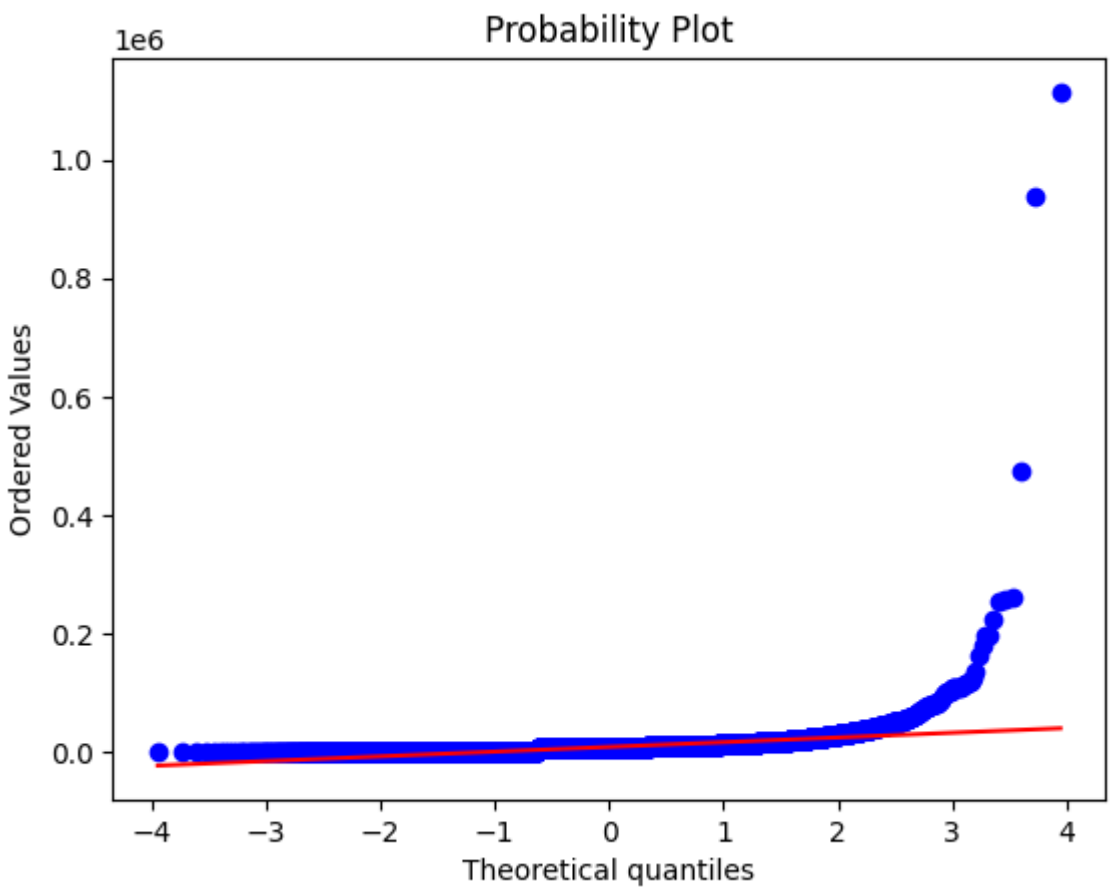
Out[ ]:
```
0.25
```

Поскольку проверка однородности выбранного параметра осуществляется с уровнем значимости 2%, мы не можем отклонить нулевую гипотезу, согласно которой выборки имеют одинаковое распределение. Следовательно, проводим тест для целевой метрики. Поскольку целевая метрика — среднее значение, то мы выбираем между критерием Стьюдента и критерием Манна Уитни.

Из документации к scipy.stats.mannwhitneyu: *The Mann-Whitney U test is a non-parametric version of the t-test for independent samples. When the means of samples from the populations are normally distributed, consider scipy.stats.ttest_ind.*

Посмотрим, имеют ли обе выборки нормальное распределение:

In [ ]:
```
scipy.stats.probplot(df_test.NPV.to_numpy(), dist="norm", plot=pylab)
pylab.show()
```



Для убедительности проведём тест

In [ ]:
```
from scipy.stats import kstest

kstest(df_test.NPV.to_numpy(), 'norm').pvalue
```

Out[ ]:
```
0.0
```

Здесь уже очевидно, что тестовая выборка не имеет нормальное распределение. Поэтому применим критерий Манна Уитни:

In [ ]:
```
mannwhitneyu(df_control.NPV, df_test.NPV, alternative = 'greater').pvalue
```

Out[ ]:
```
0.500000214910426
```

pvalue > 0.05, значит, мы не можем отвергнуть нулевую гипотезу теста. Следовательно, мы не можем утверждать, что увеличение стоимости продукта позволит суммарно увеличить доходность продукта.