Exploratory Analysis

In [13]:
```
import pandas as pd
```

In [14]:
```
data = pd.read_csv("diabeticVision.csv", index_col=0)
```

Look at summary statistics

In [15]:
```
data.describe()
```

Out[15]:

	id	age	trt	futime	status	risk	group
count	394.000000	394.000000	394.000000	394.000000	394.000000	394.000000	394.000000
mean	873.203046	20.781726	0.500000	35.579289	0.393401	9.697970	1.507614
std	495.523410	14.812074	0.500636	21.355896	0.489126	1.475033	1.119430
min	5.000000	1.000000	0.000000	0.300000	0.000000	6.000000	0.000000
25%	480.000000	10.000000	0.000000	13.977500	0.000000	9.000000	1.000000
50%	834.000000	16.000000	0.500000	38.800000	0.000000	10.000000	1.500000
75%	1296.000000	30.000000	1.000000	54.252500	1.000000	11.000000	3.000000
max	1749.000000	58.000000	1.000000	74.970000	1.000000	12.000000	3.000000

In [26]:
```
data[data["trt"] == 1]["status"].value_counts(normalize=0)
```

Out[26]:
```
0    143
1     54
Name: status, dtype: int64
```

In [25]:
```
data[data["trt"] == 0]["status"].value_counts(normalize=0)
```

Out[25]:
```
1    101
0     96
Name: status, dtype: int64
```

In [20]:
```
data["type"].value_counts()
```

Out[20]:
```
juvenile    228
adult       166
Name: type, dtype: int64
```

In [4]:
```
data.isna().sum()
```

Out[4]:
```
id        0
laser     0
eye       0
age       0
```

```
type      0
trt       0
futime    0
status    0
risk      0
group     0
dtype: int64
```

In [6]:
```
data.groupby(["trt", "status"])["futime"].mean()
```

Out[6]:
```
trt  status
0    0         46.321771
     1         18.948515
1    0         46.668112
     1         18.222407
Name: futime, dtype: float64
```

In [5]:
```
data.groupby(["trt", "status", "laser"])["futime"].mean()
```

Out[5]:
```
trt  status  laser
0    0       argon    43.247234
             xenon    49.270816
     1       argon    16.125000
             xenon    21.716667
1    0       argon    45.927500
             xenon    47.339600
     1       argon    20.004828
             xenon    16.154800
Name: futime, dtype: float64
```

In [48]:
```
data.pivot(index=["id", "trt"], columns=["status"])["futime"][0].dropna()
```

Out[48]:
```
id    trt
5     0      46.23
      1      46.23
14    1      42.50
16    0      42.27
      1      42.27
             ...
1717  1      51.60
1727  1      49.97
1746  1      45.90
1749  0      41.93
      1      41.93
Name: 0, Length: 239, dtype: float64
```

Look at various sub-groups

In [31]:
```
treat = data[data["trt"] == 1]
ctrl = data[data["trt"] == 0]
treat1 = treat[treat["laser"] == "argon"]
treat2 = treat[treat["laser"] == "xenon"]
```

In [ ]:
```
treat[treat["status"] == 1]
```

In [15]:
```
data.groupby("id")["eye"].sum().unique()
```

Out[15]:
```
array(['leftleft', 'rightright'], dtype=object)
```

In [17]:
```
data[data["futime"] == 46.23]
```

Out[17]:

	id	laser	eye	age	type	trt	futime	status	risk
1	5	argon	left	28	adult	1	46.23	0	9
2	5	argon	left	28	adult	0	46.23	0	9
195	832	argon	right	5	juvenile	1	46.23	0	12
196	832	argon	right	5	juvenile	0	46.23	0	12

In [24]:
```
ctrl.describe()
```

Out[24]:

	id	age	trt	futime	status	risk
count	101.000000	101.000000	101.0	101.000000	101.0	101.000000
mean	801.207921	23.079208	0.0	18.948515	1.0	9.970297
std	481.609869	15.532342	0.0	15.735833	0.0	1.465984
min	14.000000	1.000000	0.0	0.300000	1.0	6.000000
25%	409.000000	11.000000	0.0	6.530000	1.0	9.000000
50%	722.000000	19.000000	0.0	13.900000	1.0	10.000000
75%	1205.000000	37.000000	0.0	26.470000	1.0	11.000000
max	1746.000000	56.000000	0.0	61.830000	1.0	12.000000

In [25]:
```
treat1.describe()
```

Out[25]:

	id	age	trt	futime	status	risk
count	29.000000	29.000000	29.0	29.000000	29.0	29.000000
mean	822.344828	18.206897	1.0	20.004828	1.0	9.931034
std	497.132368	14.639426	0.0	17.418952	0.0	1.251600
min	100.000000	1.000000	1.0	1.500000	1.0	6.000000
25%	357.000000	9.000000	1.0	5.770000	1.0	9.000000
50%	866.000000	13.000000	1.0	13.330000	1.0	10.000000
75%	1184.000000	23.000000	1.0	34.370000	1.0	10.000000
max	1649.000000	53.000000	1.0	63.330000	1.0	12.000000

In [26]:
```
treat2.describe()
```

Out[26]:

	id	age	trt	futime	status	risk

	id	age	trt	futime	status	risk
count	25.000000	25.000000	25.0	25.000000	25.0	25.000000
mean	812.560000	18.600000	1.0	16.154800	1.0	9.720000
std	438.838531	13.044795	0.0	10.425035	0.0	1.369915
min	127.000000	3.000000	1.0	1.770000	1.0	6.000000
25%	503.000000	10.000000	1.0	7.070000	1.0	9.000000
50%	778.000000	13.000000	1.0	13.830000	1.0	10.000000
75%	1017.000000	25.000000	1.0	25.630000	1.0	11.000000
max	1688.000000	50.000000	1.0	42.430000	1.0	12.000000

In [10]:

```
def assign_group(row):
    if row["trt"] == 0:
        if row["laser"] == "argon":
            return 0
        return 1
    elif row["trt"] == 1:
        if row["laser"] == "argon":
            return 2
        return 3
```

Identify and label Laser and Treatment as an interaction term

In [11]:

```
data["group"] = data[["laser", "trt"]].apply(assign_group, axis=1).astype("categ
data
```

Out[11]:

	id	laser	eye	age	type	trt	futime	status	risk	group
1	5	argon	left	28	adult	1	46.23	0	9	2
2	5	argon	left	28	adult	0	46.23	0	9	0
3	14	argon	right	12	juvenile	1	42.50	0	8	2
4	14	argon	right	12	juvenile	0	31.30	1	6	0
5	16	xenon	right	9	juvenile	1	42.27	0	11	3
...	...	...	...	...	...	...	...	...	...	...
390	1727	argon	right	33	adult	0	2.90	1	10	0
391	1746	argon	right	3	juvenile	1	45.90	0	10	2
392	1746	argon	right	3	juvenile	0	1.43	1	10	0
393	1749	argon	right	32	adult	1	41.93	0	9	2
394	1749	argon	right	32	adult	0	41.93	0	9	0

394 rows × 10 columns

In [24]:
```
data["group"].describe()
```

Out[24]:
```
count     394
unique      3
top         2
freq      197
Name: group, dtype: int64
```

In [12]:
```
data.to_csv("diabeticVision.csv")
```

In [ ]:
```

```