Import standard Libraries ¶

In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

Import Data

In [2]:

```
df = pd.read_csv('cancer_tumor_data_features.csv')
```

In [3]:

```
df.head()
```

Out[3]:

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mear symmetry
0	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	0.241
1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	0.181
2	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	0.206
3	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	0.259
4	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	0.180

5 rows × 30 columns

Scaling the Data with StandardScalar

In [4]:

```
from sklearn.preprocessing import StandardScaler
```

In [5]:

```
scaler = StandardScaler()
```

In [6]:

```
scaled_X = scaler.fit_transform(df)
```

In [7]:

```
scaled_X
```

Out[7]:

```
array([[ 1.09706398, -2.07333501,  1.26993369, ...,  2.29607613,
         2.75062224,  1.93701461],
       [ 1.82982061, -0.35363241,  1.68595471, ...,  1.0870843 ,
        -0.24388967,  0.28118999],
       [ 1.57988811,  0.45618695,  1.56650313, ...,  1.95500035,
         1.152255  ,  0.20139121],
       ...,
       [ 0.70228425,  2.0455738 ,  0.67267578, ...,  0.41406869,
        -1.10454895, -0.31840916],
       [ 1.83834103,  2.33645719,  1.98252415, ...,  2.28998549,
         1.91908301,  2.21963528],
       [-1.80840125,  1.22179204, -1.81438851, ..., -1.74506282,
        -0.04813821, -0.75120669]])
```

Now, lets see how variance of data differ when we reduce the data to different number rof dimensions

In [8]:

```
from sklearn.decomposition import PCA
```

In [9]:

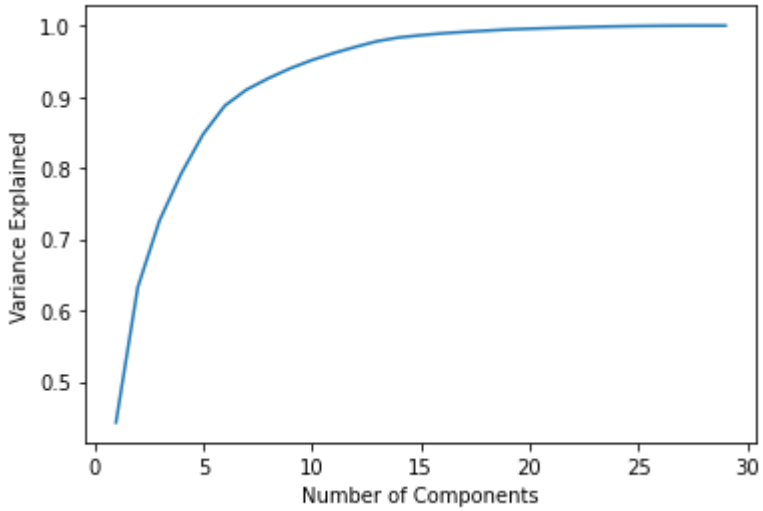
```
explained_variance = []

for n in range(1,30):
    pca = PCA(n_components=n)
    pca.fit(scaled_X)

    explained_variance.append(np.sum(pca.explained_variance_ratio_))
```

In [10]:

```
plt.plot(range(1,30),explained_variance)
plt.xlabel("Number of Components")
plt.ylabel("Variance Explained");
```



Lets reduce the dimensions where variance is 95% of the original data

In [11]:

```
pca = PCA(.95)
pca.fit(scaled_X)
```

Out[11]:

```
PCA(n_components=0.95)
```

In [12]:

```
pca.n_components_
```

Out[12]:

```
10
```

We see that we need to reduce the dimension to 10 components

In [13]:

```
principal_components = pca.fit_transform(scaled_X)
```

In [14]:

```
pca.components_
```

Out[14]:

```
array([[ 2.18902444e-01,  1.03724578e-01,  2.27537293e-01,
         2.20994985e-01,  1.42589694e-01,  2.39285354e-01,
         2.58400481e-01,  2.60853758e-01,  1.38166959e-01,
         6.43633464e-02,  2.05978776e-01,  1.74280281e-02,
         2.11325916e-01,  2.02869635e-01,  1.45314521e-02,
         1.70393451e-01,  1.53589790e-01,  1.83417397e-01,
         4.24984216e-02,  1.02568322e-01,  2.27996634e-01,
         1.04469325e-01,  2.36639681e-01,  2.24870533e-01,
         1.27952561e-01,  2.10095880e-01,  2.28767533e-01,
         2.50885971e-01,  1.22904556e-01,  1.31783943e-01],
       [-2.33857132e-01, -5.97060883e-02, -2.15181361e-01,
        -2.31076711e-01,  1.86113023e-01,  1.51891610e-01,
         6.01653628e-02, -3.47675005e-02,  1.90348770e-01,
         3.66575471e-01, -1.05552152e-01,  8.99796818e-02,
        -8.94572342e-02, -1.52292628e-01,  2.04430453e-01,
         2.32715896e-01,  1.97207283e-01,  1.30321560e-01,
         1.83848000e-01,  2.80092027e-01, -2.19866379e-01,
        -4.54672983e-02, -1.99878428e-01, -2.19351858e-01,
         1.72304352e-01,  1.43593173e-01,  9.79641143e-02,
        -8.25723507e-03,  1.41883349e-01,  2.75339469e-01],
       [-8.53124284e-03,  6.45499033e-02, -9.31421972e-03,
         2.86995259e-02, -1.04291904e-01, -7.40915709e-02,
         2.73383798e-03, -2.55635406e-02, -4.02399363e-02,
        -2.25740897e-02,  2.68481387e-01,  3.74633665e-01,
         2.66645367e-01,  2.16006528e-01,  3.08838979e-01,
         1.54779718e-01,  1.76463743e-01,  2.24657567e-01,
         2.88584292e-01,  2.11503764e-01, -4.75069900e-02,
        -4.22978228e-02, -4.85465083e-02, -1.19023182e-02,
        -2.59797613e-01, -2.36075625e-01, -1.73057335e-01,
        -1.70344076e-01, -2.71312642e-01, -2.32791313e-01],
       [ 4.14089623e-02, -6.03050001e-01,  4.19830991e-02,
         5.34337955e-02,  1.59382765e-01,  3.17945811e-02,
         1.91227535e-02,  6.53359443e-02,  6.71249840e-02,
         4.85867649e-02,  9.79412418e-02, -3.59855528e-01,
         8.89924146e-02,  1.08205039e-01,  4.46641797e-02,
        -2.74693632e-02,  1.31687997e-03,  7.40673350e-02,
         4.40733510e-02,  1.53047496e-02,  1.54172396e-02,
        -6.32807885e-01,  1.38027944e-02,  2.58947492e-02,
         1.76522161e-02, -9.13284153e-02, -7.39511797e-02,
         6.00699571e-03, -3.62506947e-02, -7.70534703e-02],
       [ 3.77863538e-02, -4.94688505e-02,  3.73746632e-02,
         1.03312514e-02, -3.65088528e-01,  1.17039713e-02,
         8.63754118e-02, -4.38610252e-02, -3.05941428e-01,
        -4.44243602e-02, -1.54456496e-01, -1.91650506e-01,
        -1.20990220e-01, -1.27574432e-01, -2.32065676e-01,
         2.79968156e-01,  3.53982091e-01,  1.95548089e-01,
        -2.52868765e-01,  2.63297438e-01, -4.40659209e-03,
        -9.28834001e-02,  7.45415100e-03, -2.73909030e-02,
        -3.24435445e-01,  1.21804107e-01,  1.88518727e-01,
         4.33320687e-02, -2.44558663e-01,  9.44233510e-02],
       [ 1.87407904e-02, -3.21788366e-02,  1.73084449e-02,
        -1.88774796e-03, -2.86374497e-01, -1.41309489e-02,
        -9.34418089e-03, -5.20499505e-02,  3.56458461e-01,
        -1.19430668e-01, -2.56032561e-02, -2.87473145e-02,
         1.81071500e-03, -4.28639079e-02, -3.42917393e-01,
```

```
         6.91975186e-02,  5.63432386e-02, -3.12244482e-02,
         4.90245643e-01, -5.31952674e-02, -2.90684919e-04,
        -5.00080613e-02,  8.50098715e-03, -2.51643821e-02,
        -3.69255370e-01,  4.77057929e-02,  2.83792555e-02,
        -3.08734498e-02,  4.98926784e-01, -8.02235245e-02],
       [-1.24088340e-01,  1.13995382e-02, -1.14477057e-01,
        -5.16534275e-02, -1.40668993e-01,  3.09184960e-02,
        -1.07520443e-01, -1.50482214e-01, -9.38911345e-02,
         2.95760024e-01,  3.12490037e-01, -9.07553556e-02,
         3.14640390e-01,  3.46679003e-01, -2.44024056e-01,
         2.34635340e-02, -2.08823790e-01, -3.69645937e-01,
        -8.03822539e-02,  1.91394973e-01, -9.70993602e-03,
         9.87074388e-03, -4.45726717e-04,  6.78316595e-02,
        -1.08830886e-01,  1.40472938e-01, -6.04880561e-02,
        -1.67966619e-01, -1.84906298e-02,  3.74657626e-01],
       [-7.45229622e-03,  1.30674825e-01, -1.86872582e-02,
         3.46736038e-02, -2.88974575e-01, -1.51396350e-01,
        -7.28272853e-02, -1.52322414e-01, -2.31530989e-01,
        -1.77121441e-01,  2.25399674e-02, -4.75413139e-01,
        -1.18966905e-02,  8.58051345e-02,  5.73410232e-01,
         1.17460157e-01,  6.05665008e-02, -1.08319309e-01,
         2.20149279e-01,  1.11681884e-02,  4.26194163e-02,
         3.62516360e-02,  3.05585340e-02,  7.93942456e-02,
         2.05852191e-01,  8.40196588e-02,  7.24678714e-02,
        -3.61707954e-02,  2.28225053e-01,  4.83606666e-02],
       [-2.23109764e-01,  1.12699390e-01, -2.23739213e-01,
        -1.95586014e-01,  6.42472194e-03, -1.67841425e-01,
         4.05910064e-02, -1.11971106e-01,  2.56040084e-01,
        -1.23740789e-01,  2.49985002e-01, -2.46645397e-01,
         2.27154024e-01,  2.29160015e-01, -1.41924890e-01,
        -1.45322810e-01,  3.58107079e-01,  2.72519886e-01,
        -3.04077200e-01, -2.13722716e-01, -1.12141463e-01,
         1.03341204e-01, -1.09614364e-01, -8.07324609e-02,
         1.12315904e-01, -1.00677822e-01,  1.61908621e-01,
         6.04884615e-02,  6.46378061e-02, -1.34174175e-01],
       [ 9.54864432e-02,  2.40934066e-01,  8.63856150e-02,
         7.49564886e-02, -6.92926813e-02,  1.29362000e-02,
        -1.35602298e-01,  8.05452775e-03,  5.72069479e-01,
         8.11032072e-02, -4.95475941e-02, -2.89142742e-01,
        -1.14508236e-01, -9.19278886e-02,  1.60884609e-01,
         4.35048658e-02, -1.41276243e-01,  8.62408470e-02,
        -3.16529830e-01,  3.67541918e-01,  7.73616428e-02,
         2.95509413e-02,  5.05083335e-02,  6.99211523e-02,
        -1.28304659e-01, -1.72133632e-01, -3.11638520e-01,
        -7.66482910e-02, -2.95630751e-02,  1.26095791e-02]])
```

In [15]:

```
df_comp = pd.DataFrame(pca.components_,index=['PC1','PC2','PC3','PC4','PC5','PC6','PC7','PC
```

In [16]:

```
df_comp
```

Out[16]:

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	me conca poir
PC1	0.218902	0.103725	0.227537	0.220995	0.142590	0.239285	0.258400	0.2608
PC2	-0.233857	-0.059706	-0.215181	-0.231077	0.186113	0.151892	0.060165	-0.0347
PC3	-0.008531	0.064550	-0.009314	0.028700	-0.104292	-0.074092	0.002734	-0.0255
PC4	0.041409	-0.603050	0.041983	0.053434	0.159383	0.031795	0.019123	0.0653
PC5	0.037786	-0.049469	0.037375	0.010331	-0.365089	0.011704	0.086375	-0.0438
PC6	0.018741	-0.032179	0.017308	-0.001888	-0.286374	-0.014131	-0.009344	-0.0520
PC7	-0.124088	0.011400	-0.114477	-0.051653	-0.140669	0.030918	-0.107520	-0.1504
PC8	-0.007452	0.130675	-0.018687	0.034674	-0.288975	-0.151396	-0.072827	-0.1523
PC9	-0.223110	0.112699	-0.223739	-0.195586	0.006425	-0.167841	0.040591	-0.1119
PC10	0.095486	0.240934	0.086386	0.074956	-0.069293	0.012936	-0.135602	0.0080

10 rows × 30 columns

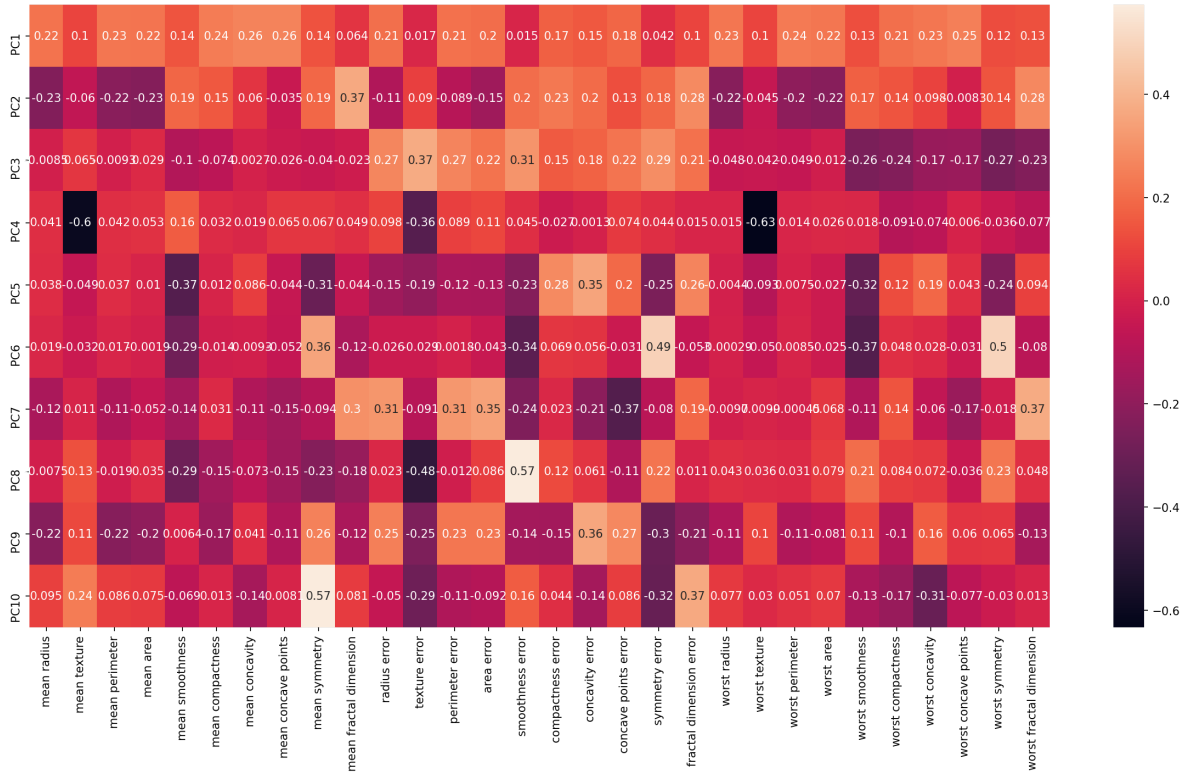
Lets, visualize the relationship of Original features to each principle components

In [17]:

```
plt.figure(figsize=(20,10),dpi=150)
sns.heatmap(df_comp,annot=True)
```

Out[17]:

<AxesSubplot:>



In [18]:

```
print("Explained variance ratio: ", np.sum(pca.explained_variance_ratio_))
```

```
Explained variance ratio:  0.9515688143366666
```

In [ ]:

```

```