Osman Bayrak

212062465

Part A)

1) If a player has high number of turn overs then this player has high number of free throws and points.

I thought that if a player is try turn overs many times, he get faults and make free throws also he make points as well. So I wonder that my claim can be proven statistically in the data set.

2) My variables are turn overs “turnovers” column, successful free throws “ftMade” column and points ‘points’ column.

I decided to shrink my data set. In order to make it more specific I filter my data set by league name and season year also I did not take post season stats.

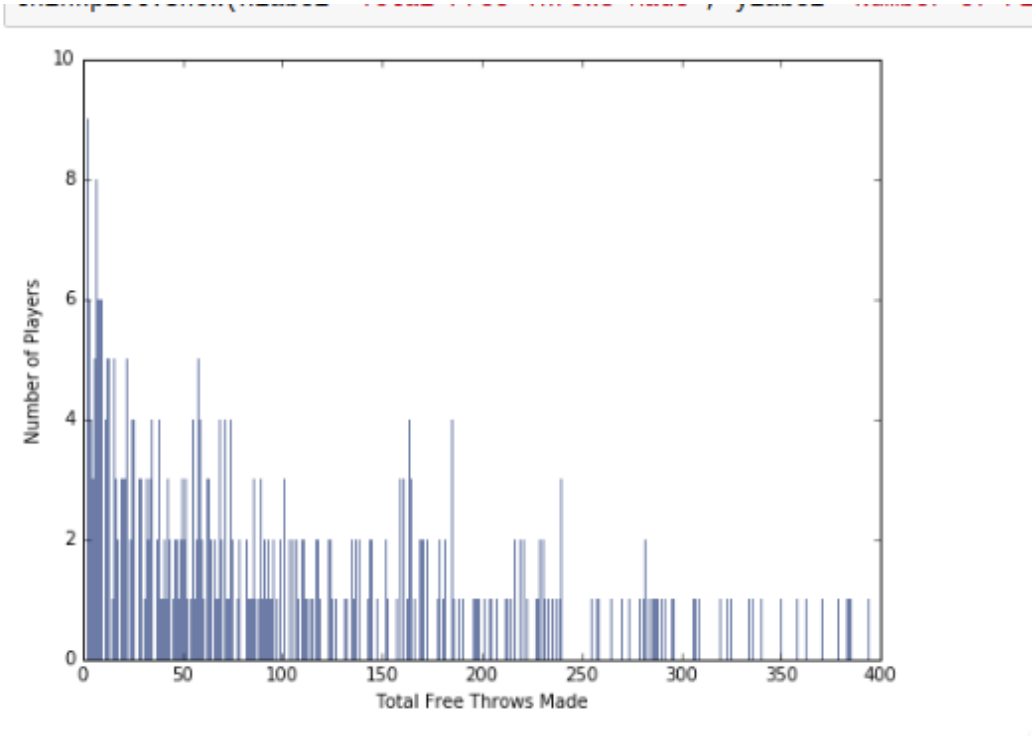
After that in order to clean my data set, I drop the NaN and inconsistent values from my data set. Results are make more sense after this cleaning.

```
In [162]: #Question 2
          nbaDf = df.loc[df['lgID'] == 'NBA']
          nbaDf =nbaDf.loc[nbaDf['year'] == 1990]

          turnOvers = nbaDf['turnovers']
          points = nbaDf['points']
          ftMade = nbaDf['ftMade']
          turnOvers.dropna(inplace=True)
          ftMade.dropna(inplace=True)
          points.dropna(inplace=True)

          print points.size
          print ftMade.size
          print turnOvers.size
          415
          415
          415
```

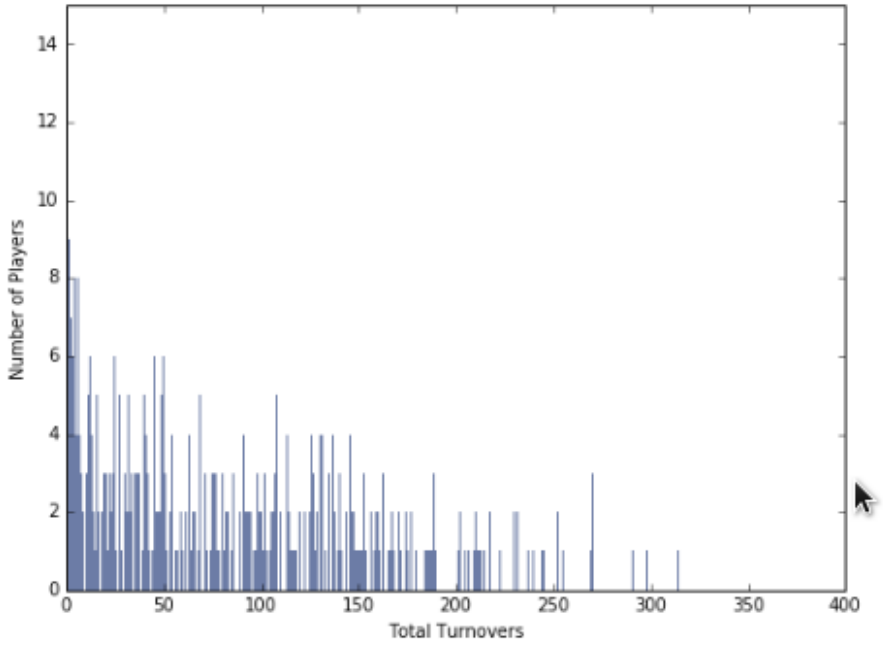
3) I calculate the standart deviations, means and spearman corralation between them.

In [167]:
```
print turnOvers.std()
print ftMade.std()
print turnOvers.mean()
print ftMade.mean()
print SpearmanCorr(turnOvers, ftMade)
```

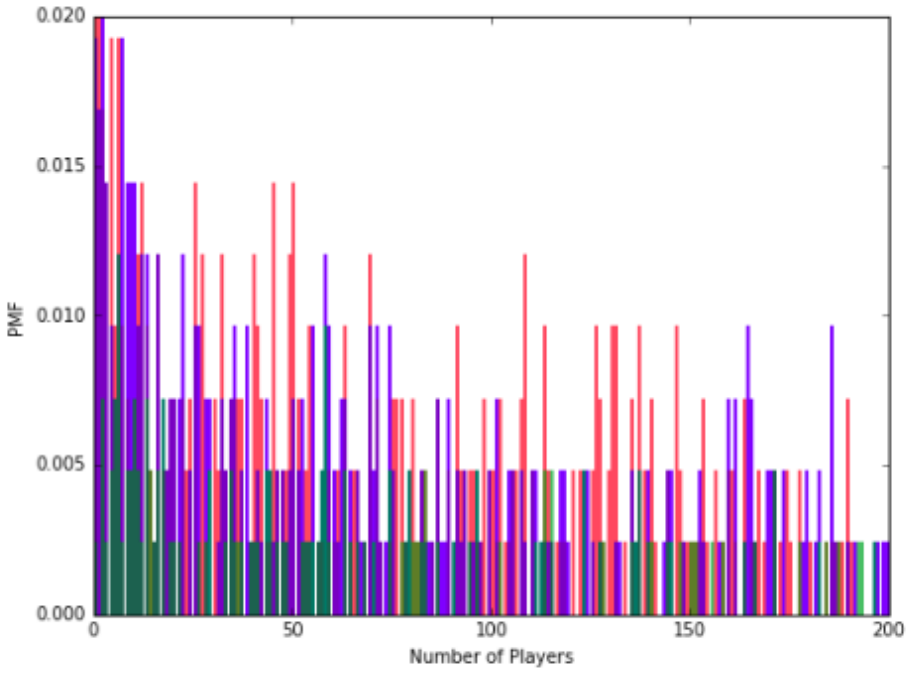
```
69.85582006864077
121.22415385376968
83.25060240963856
113.76626506024097
0.9291008082359609
```



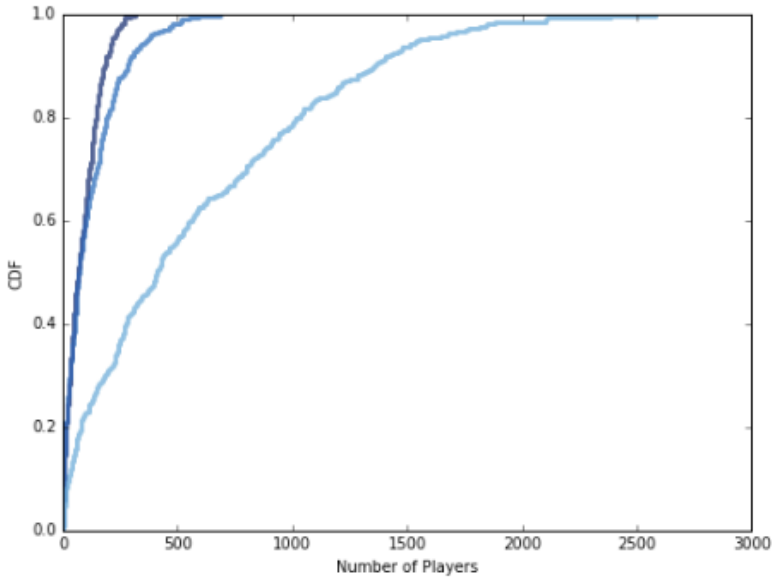
This is histogram of total free throws made in a season.



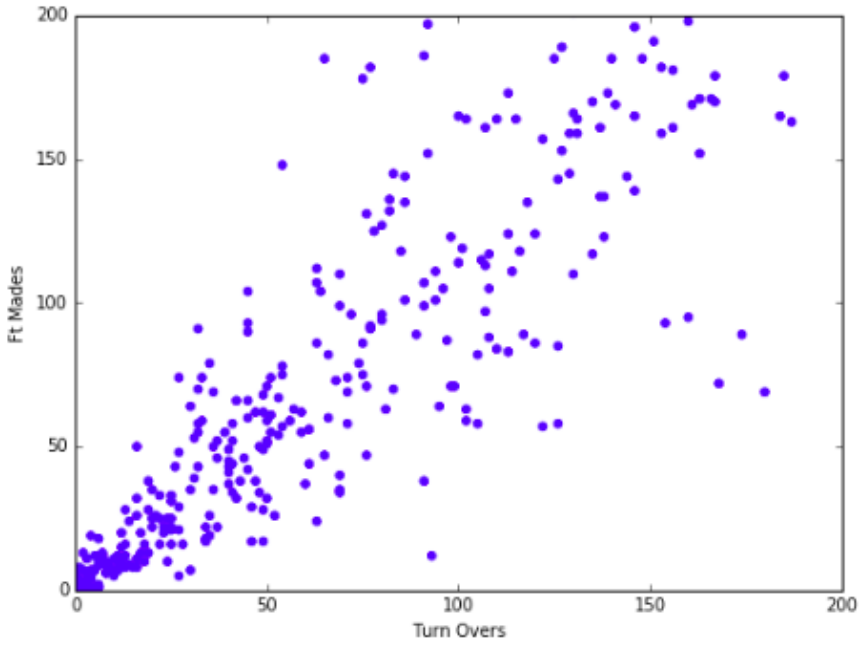
This is histogram of Total turn overs made in a season in NBA.



This is the PMF graph of Free Throws, Turn overs and Points made by players. Most of the increase and decreased locations are near each other. It means these variables are strongly related.



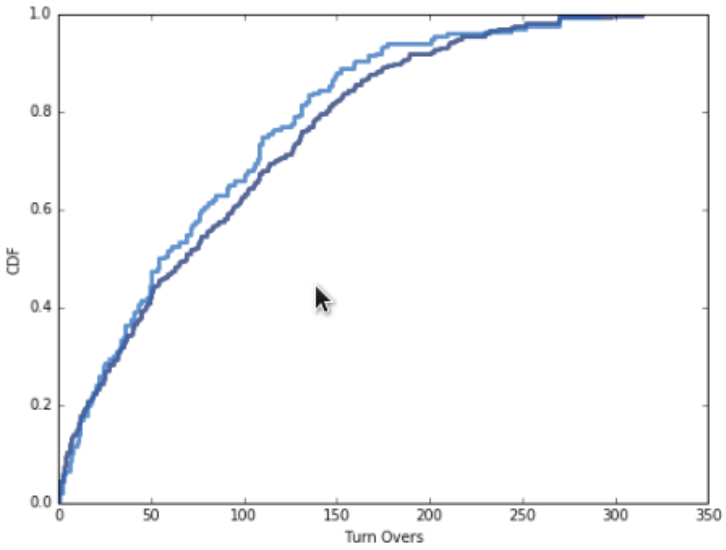
This is the CDF of them. Especially turn overs and free throws are very close. It means they are effecting each other.



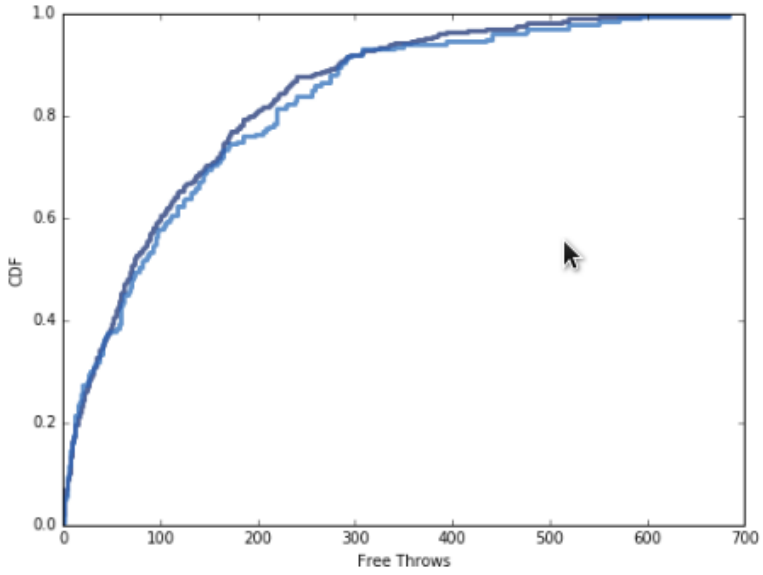
This is scatter plot graph of turn overs and free throws made by players. Another proven statistic graph for turn over and free throw variables.

4)

In [148]:
```
Turnovers_Resample = turnOvers_CDF.Sample(200)
thinkplot.Cdf(turnOvers_CDF)
thinkplot.Cdf(thinkstats2.Cdf(Turnovers_Resample, label='resample'))
thinkplot.Show(xlabel='Turn Overs', ylabel='CDF')
```



In [149]:
```
ft_Resample = ft_CDF.Sample(200)
thinkplot.Cdf(ft_CDF)
thinkplot.Cdf(thinkstats2.Cdf(ft_Resample, label='resample'))
thinkplot.Show(xlabel='Free Throws', ylabel='CDF')
```



I use cumulative distribution. Take samples from variables as subset and take the percentile ranks and they fit in my model.

In this step I calculate the percentile ranks of each variable then create a sample subset from this variables. Then calculate the percentile ranks and fit them in my model in order to see accuracy.

5)

I use spearman corralation and it shows me that my variables turn overs and free throws are strongly related with each other. Because spearman corr. Output should be in range of -1 and 1. If the result is between 0.5 and 1 it means that they are strongly related. And here is my results:

```
def SpearmanCorr(xs, ys):
    xranks = pd.Series(xs).rank()
    yranks = pd.Series(ys).rank()
    return Corr(xranks, yranks)


print SpearmanCorr(turnOvers, ftMade)
print SpearmanCorr(turnOvers, points)
```

```
0.9291008082359609
0.9519572930454144
```

6)

```
In [157]: class CorrelationPermute(thinkstats2.HypothesisTest):

              def TestStatistic(self, data):
                  xs, ys = data
                  test_stat = abs(SpearmanCorr(xs, ys))
                  return test_stat

              def RunModel(self):
                  xs, ys = self.data
                  xs = np.random.permutation(xs)
                  return xs, ys
```

```
In [160]: data = turnOvers.values, ftMade.values
          ht = CorrelationPermute(data)
          pvalue = ht.PValue()
          pvalue
```

```
Out[160]: 0.0
```

```
In [161]: ht.actual, ht.MaxTestStat()
```

```
Out[161]: (0.9291008082359609, 0.16925842651350545)
```

I tried to test my correlation but I could not find the p-value very well. It should be very near to 0 because this correlation is very strong. If the p-value would be less than 0.5 then it would be mean that variables correlation is strongly related.

7)

I found that my hypothesis if true. I claim that the turnovers leads faults and it means players will be made free throws very much. In the results their relation is very strong and their scatter plot graph, PMF and CDF graphs has very similar values in same locations. My p-value would be very close to 0 and it measn it would be smaller than 0.5 which is thresh hold and it means my results are significant.