数据挖掘作业一

殷赫 生物医学工程 2220170871

数据集1: NFL Play-by-Play 2009-2017

**实验结果**

**1 数据可视化和摘要**

**（1）摘要**

**a.drive**

程序：stats\_str(data['Drive'])

结果：

1: 19261, 2: 17084, 3: 17425, 4: 17875, 5: 17822, 6: 17925, 7: 18088, 8: 17691, 9: 17033, 10: 17193, 11: 16898, 12: 16724, 13: 16745, 14: 16881, 15: 17188, 16: 17525, 17: 17459, 18: 17339, 19: 16928, 20: 15622, 21: 13998, 22: 12071, 23: 9933, 24: 7959, 25: 5842, 26: 4200, 27: 2815, 28: 1658, 29: 1077, 30: 675, 31: 366, 32: 237, 33: 97, 34: 45, 35: 9

**b.sp**

程序：stats\_str(data['sp'])

结果：

{0: 378398, 1: 29290}

**c.qtr**

程序：stats\_str(data['qtr'])

结果：

1: 89176, 2: 112317, 3: 90682, 4: 112641, 5: 2872

**d.down**

程序：stats\_num(data['down'])

结果：

MAX 4.000000

MIN 1.000000

MEAN 2.002476

MEDIAN 2.000000

1/4 1.000000

2/4 2.000000

3/4 3.000000

NULLNUMBER 61154.000000

**e. yrdln**

程序：stats\_num(data['yrdln'])

结果：

MAX 50.000000

MIN 1.000000

MEAN 28.488327

MEDIAN 30.000000

1/4 20.000000

2/4 30.000000

3/4 39.000000

NULLNUMBER 840.000000

**f. ydstogo**

程序：stats\_str(data['ydstogo'])

结果：

0: 61154, 1: 18315, 2: 13267, 3: 13998, 4: 15159, 5: 17894, 6: 17403, 7: 17395, 8: 16662, 9: 14487, 10: 158235, 11: 7819, 12: 5586, 13: 4395, 14: 3466, 15: 6315, 16: 2441, 17: 2254, 18: 1981, 19: 1627, 20: 4039, 21: 789, 22: 591, 23: 473, 24: 372, 25: 600, 26: 212, 27: 156, 28: 132, 29: 122, 30: 137, 31: 49, 32: 39, 33: 29, 34: 26, 35: 28, 36: 11, 37: 9, 38: 3, 39: 2, 40: 6, 42: 2, 43: 2, 44: 2, 46: 1, 48: 2, 50: 1

**g. ydsnet**

程序：stats\_str(data['ydsnet'])

结果：

-87: 1, -84: 1, -83: 1, -77: 3, -73: 2, -69: 2, -68: 1, -59: 2, -56: 1, -55: 2, -54: 1, -52: 1, -48: 2, -46: 1, -45: 2, -43: 1, -42: 3, -40: 2, -39: 2, -37: 1, -35: 2, -34: 1, -33: 4, -32: 3, -30: 1, -29: 6, -28: 3, -27: 8, -26: 7, -25: 7, -24: 15, -23: 18, -22: 12, -21: 28, -20: 48, -19: 49, -18: 83, -17: 73, -16: 130, -15: 225, -14: 182, -13: 220, -12: 326, -11: 435, -10: 1555, -9: 883, -8: 1156, -7: 1262, -6: 1459, -5: 2925, -4: 2176, -3: 2951, -2: 3786, -1: 6650, 0: 36899, 1: 7452, 2: 8831, 3: 9003, 4: 9266, 5: 9718, 6: 8420, 7: 8092, 8: 7690, 9: 10154, 10: 4843, 11: 8004, 12: 6950, 13: 6645, 14: 6455, 15: 6297, 16: 6070, 17: 6041, 18: 5692, 19: 5754, 20: 5764, 21: 5160, 22: 5172, 23: 5084, 24: 4973, 25: 4927, 26: 4663, 27: 4496, 28: 4310, 29: 4397, 30: 4413, 31: 4347, 32: 4041, 33: 3930, 34: 3916, 35: 4034, 36: 3792, 37: 3915, 38: 3601, 39: 3500, 40: 3625, 41: 3404, 42: 3270, 43: 3288, 44: 3310, 45: 3367, 46: 3130, 47: 3106, 48: 3081, 49: 2971, 50: 2919, 51: 2949, 52: 2910, 53: 2781, 54: 2705, 55: 2646, 56: 2742, 57: 2663, 58: 2698, 59: 2574, 60: 2678, 61: 2376, 62: 2341, 63: 2330, 64: 2202, 65: 2297, 66: 2117, 67: 2030, 68: 2077, 69: 2005, 70: 2041, 71: 1933, 72: 1946, 73: 1840, 74: 1967, 75: 3382, 76: 1530, 77: 1575, 78: 1540, 79: 1592, 80: 4717, 81: 885, 82: 809, 83: 698, 84: 678, 85: 666, 86: 542, 87: 455, 88: 475, 89: 404, 90: 494, 91: 320, 92: 293, 93: 225, 94: 190, 95: 132, 96: 106, 97: 75, 98: 90, 99: 64

**h. posteam**

程序：stats\_str(data['posteam'])

结果：

'PIT': 11859, 'TEN': 11511, nan: 24992, 'CLE': 11722, 'MIN': 11774, 'NO': 12408, 'DET': 12271, 'DAL': 11784, 'TB': 11696, 'HOU': 12315, 'NYJ': 12076, 'IND': 12156, 'JAC': 9253, 'DEN': 12398, 'CIN': 11931, 'CAR': 11870, 'PHI': 12397, 'KC': 11808, 'BAL': 12241, 'ATL': 12004, 'MIA': 11690, 'ARI': 11967, 'SF': 11591, 'STL': 9113, 'SEA': 11764, 'NYG': 12021, 'WAS': 11905, 'GB': 12052, 'CHI': 11566, 'NE': 12520, 'BUF': 11767, 'OAK': 12002, 'SD': 10661, 'LA': 2599, 'JAX': 2692, 'LAC': 1312

**i. AirYards**

程序：stats\_str(data['AirYards'])

结果：

-70: 1, -66: 1, -52: 1, -50: 1, -48: 2, -47: 1, -46: 1, -43: 1, -41: 1, -38: 4, -37: 1, -36: 1, -32: 1, -31: 1, -30: 2, -29: 3, -27: 1, -25: 1, -24: 2, -22: 4, -21: 3, -20: 2, -18: 3, -17: 3, -16: 1, -15: 15, -14: 20, -13: 22, -12: 29, -11: 33, -10: 56, -9: 110, -8: 261, -7: 568, -6: 1174, -5: 2304, -4: 3070, -3: 3937, -2: 4276, -1: 4200, 0: 259295, 1: 6687, 2: 8302, 3: 9633, 4: 10947, 5: 11911, 6: 8904, 7: 6878, 8: 5633, 9: 5060, 10: 5054, 11: 4286, 12: 4023, 13: 3868, 14: 3656, 15: 3856, 16: 3108, 17: 2936, 18: 2522, 19: 2064, 20: 2057, 21: 1574, 22: 1309, 23: 1147, 24: 1055, 25: 1090, 26: 849, 27: 781, 28: 737, 29: 661, 30: 693, 31: 565, 32: 510, 33: 510, 34: 478, 35: 484, 36: 406, 37: 349, 38: 388, 39: 354, 40: 370, 41: 305, 42: 299, 43: 273, 44: 233, 45: 257, 46: 206, 47: 184, 48: 185, 49: 128, 50: 120, 51: 98, 52: 65, 53: 63, 54: 52, 55: 29, 56: 15, 57: 23, 58: 13, 59: 5, 60: 12, 61: 5, 63: 2, 64: 2, 65: 2, 66: 1, 68: 1, 71: 1, 84: 1

**j. YardsAfterCatch**

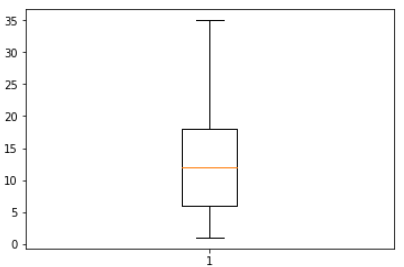
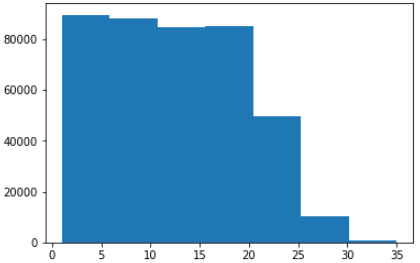
程序：stats\_str(data['YardsAfterCatch'])

结果：

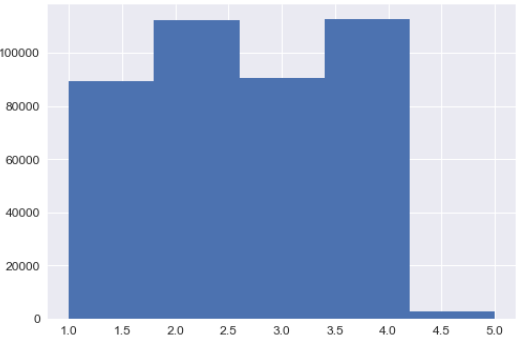
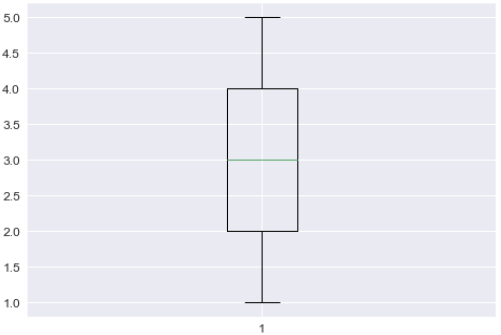
-81: 1, -44: 1, -37: 1, -32: 1, -30: 1, -25: 1, -24: 1, -23: 1, -20: 1, -19: 1, -17: 1, -16: 2, -14: 5, -13: 2, -12: 4, -11: 8, -10: 8, -9: 7, -8: 14, -7: 9, -6: 19, -5: 33, -4: 65, -3: 108, -2: 238, -1: 908, 0: 332408, 1: 10987, 2: 9745, 3: 7905, 4: 6780, 5: 5834, 6: 4942, 7: 4343, 8: 3519, 9: 3108, 10: 2612, 11: 2138, 12: 1715, 13: 1368, 14: 1164, 15: 1030, 16: 853, 17: 716, 18: 587, 19: 500, 20: 484, 21: 396, 22: 286, 23: 282, 24: 232, 25: 224, 26: 188, 27: 154, 28: 136, 29: 128, 30: 133, 31: 91, 32: 78, 33: 84, 34: 78, 35: 82, 36: 65, 37: 62, 38: 61, 39: 47, 40: 65, 41: 31, 42: 41, 43: 37, 44: 37, 45: 62, 46: 22, 47: 21, 48: 20, 49: 26, 50: 32, 51: 18, 52: 23, 53: 16, 54: 15, 55: 23, 56: 14, 57: 16, 58: 10, 59: 7, 60: 12, 61: 12, 62: 12, 63: 9, 64: 9, 65: 16, 66: 9, 67: 7, 68: 6, 69: 8, 70: 11, 71: 2, 72: 6, 73: 4, 74: 2, 75: 5, 76: 6, 77: 2, 78: 5, 79: 1, 80: 4, 81: 2, 82: 2, 83: 2, 85: 3, 87: 3, 88: 2, 89: 2, 90: 2

**（2）可视化**

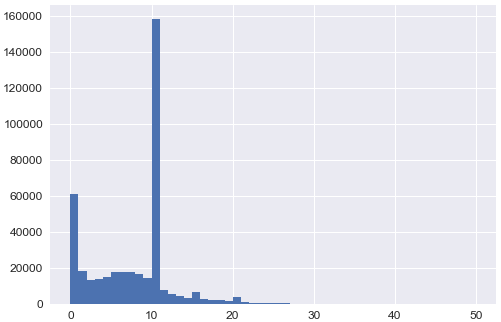
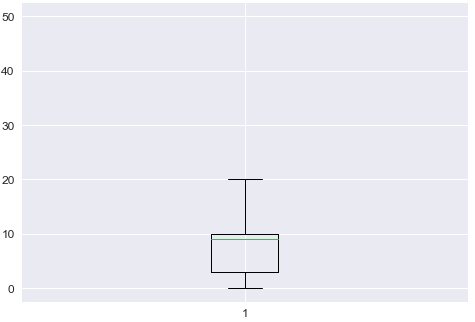
**a.drive直方图、盒图**



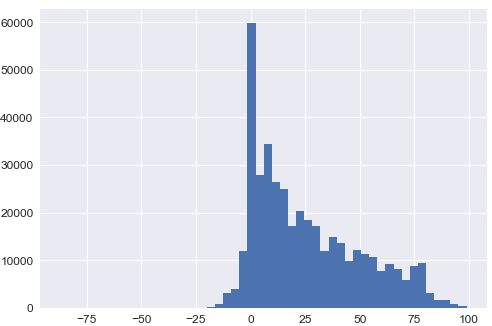
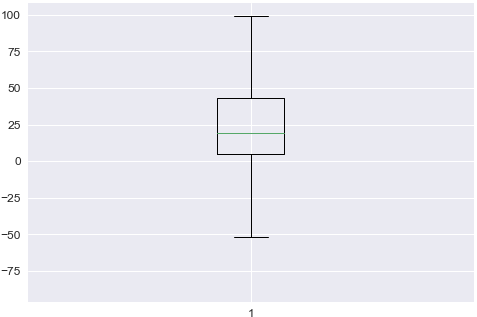
**b.qtr直方图、盒图**

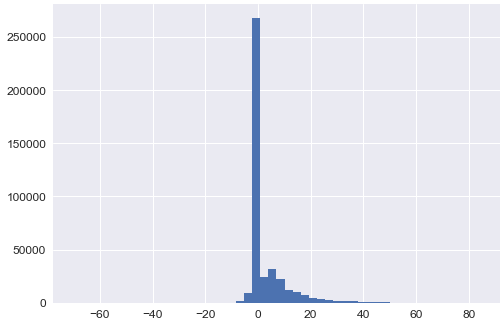
**c. ydstogo直方图、盒图**

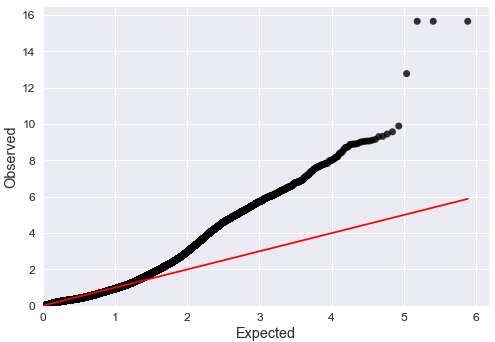
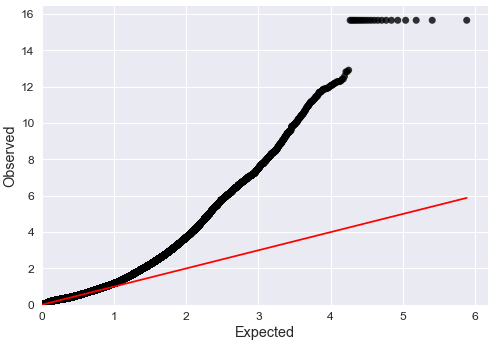
**d. ydsnet直方图、盒图**

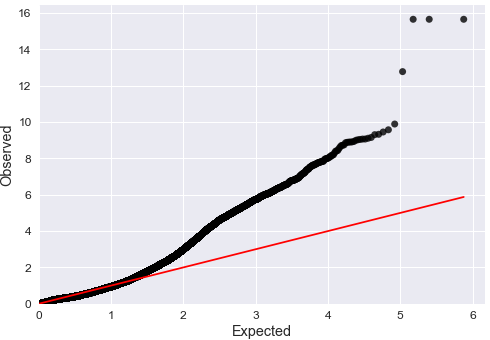
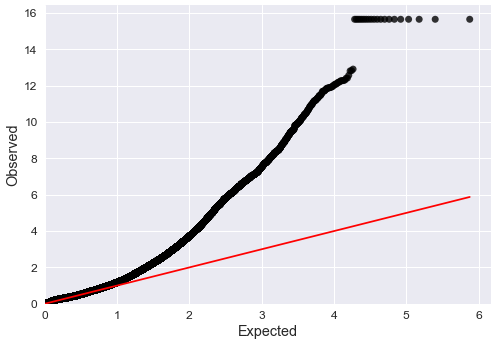
**e. AirYards直方图、盒图**

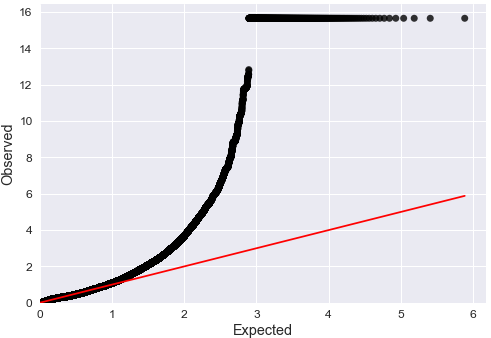
**f. Home\_WP\_pre、Away\_WP\_pre qq图**

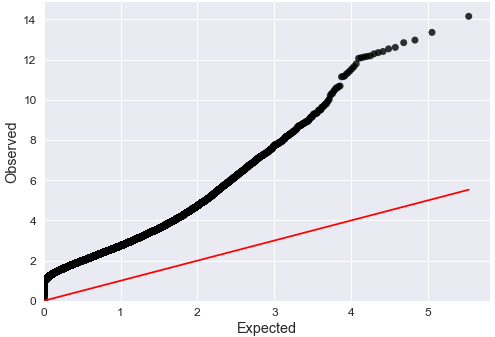
**g. Home\_WP\_post、Away\_WP\_post qq图**

**h. Win\_Prob qq图**



**i. WPA qq图**

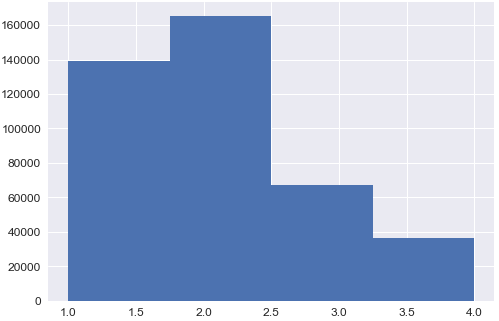
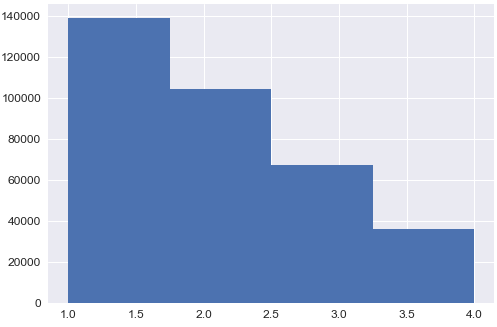


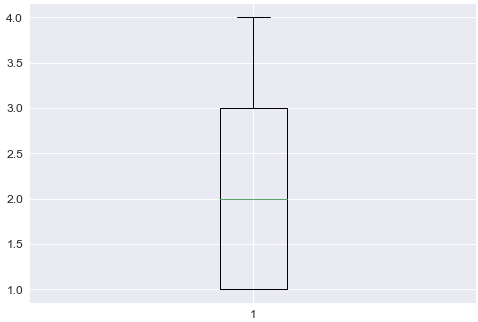
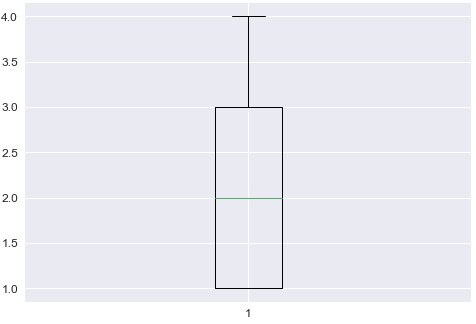
**2 数据缺失的处理**

**（1）将缺失部分剔除**

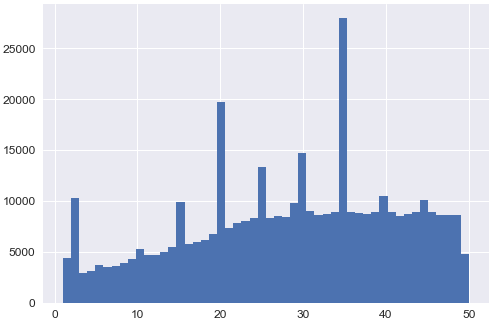
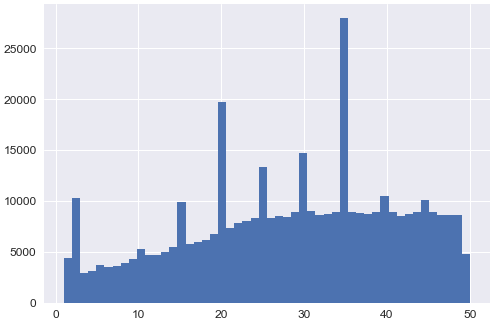
**a.down**

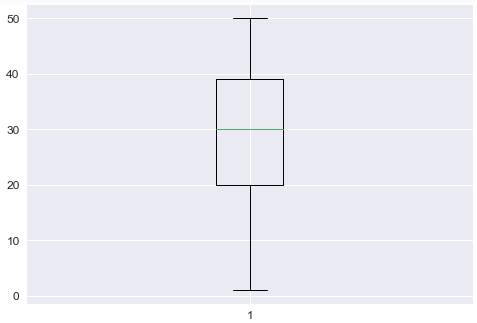
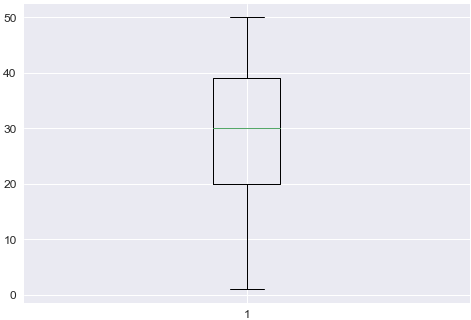
data\_drop(data['down'])

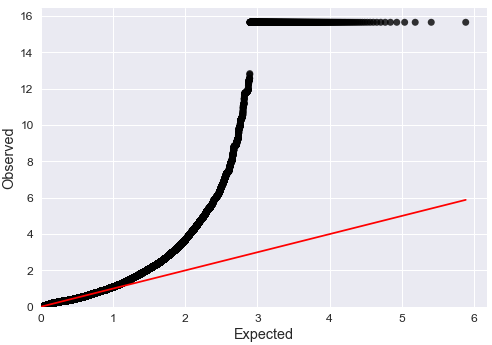
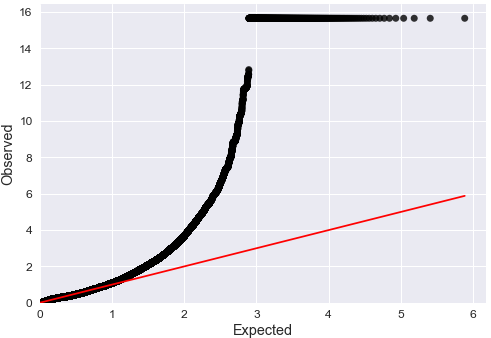
 

**b.yrdln**

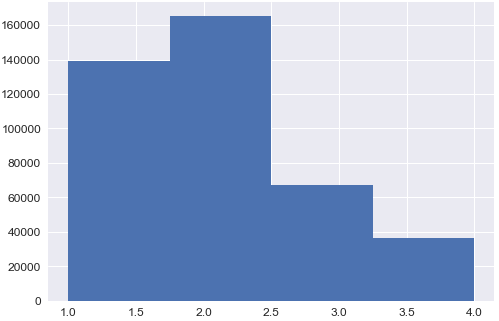
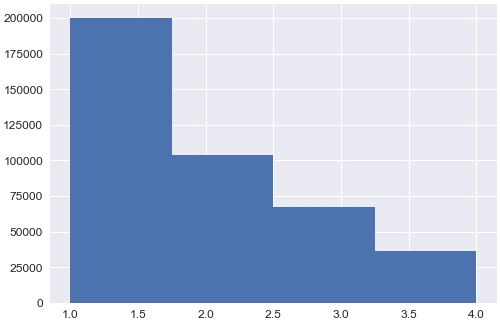
**c.win\_prob**

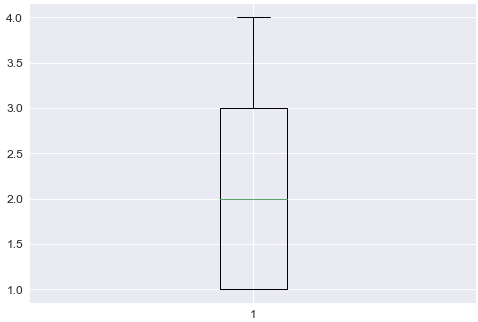
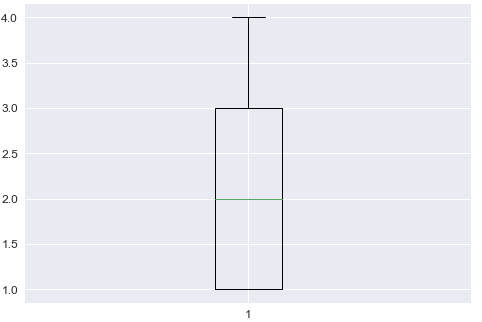


**（2）用最高频率值来填补缺失值**

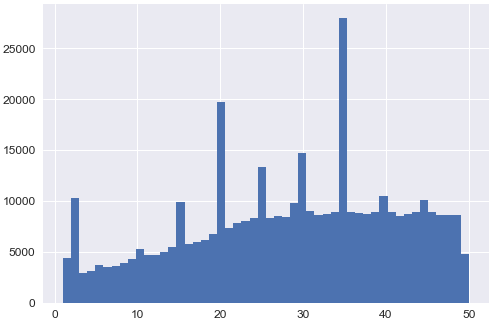
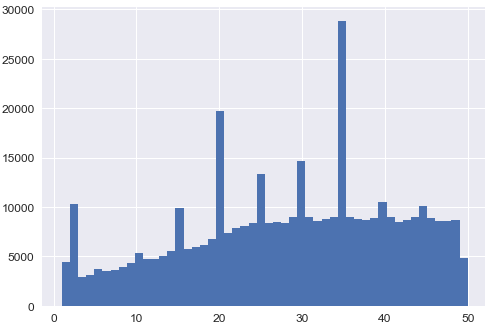
**a.down**

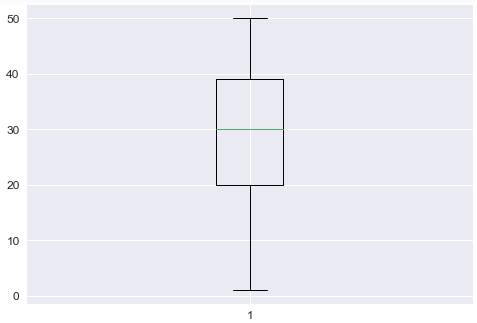
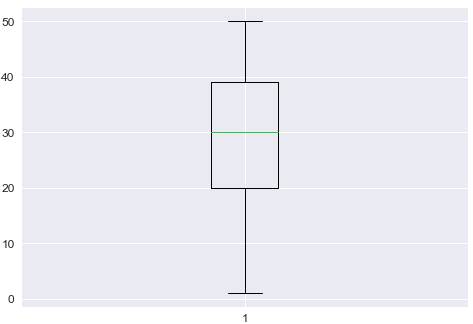
data\_drop(data['down'])

**b.yrdln**

**c.win\_prob**

