**Case 1 - Call Center Staffing Analytics - Detecting Favoritism using Machine Learning**

**Situation**: A call center operation was under close scrutiny for uneven performance, shoddy operations and low employee morale. There was a rumor floating around that the call center manager was engaging in favoritism, that certain employees were given unfairly easy working conditions, but no one was able to present a convincing case against the manager.

**Complication**: Some even went so far as to accuse the manager of nepotism -- implying that the workers being given extra sweet deals were those that were related to the manager. The case was before a judge who, given the seriousness of the case, asked for hard evidence.

**Key question**: Can we use machine learning to ‘objectively’ identify whether there is any hard evidence to prove that certain employees were being treated in a systematically different way than others. Can we do this using not one, but multiple dimensions, together?

**Solution approach**: We are going to use a very practically relevant unsupervised machine learning method called <u>anomaly detection</u> to find whether there is evidence, or lack thereof, of nepotism.

**Dataset**:
We have a dataset that records the following features for each call center worker. Here is a sample view of the initial rows of the table.

| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Employee ID | Avg Tix / Day | Customer rating | Tardies | Graveyard Shifts Taken | Weekend Shifts Taken | Sick Days Taken | % Sick Days Taken on Friday | Employee Dev. Hours | Shift Swaps Requested | Shift Swaps Offered |
| 2 | 144624 | 151.8 | 3.32 | 1 | 0 | 2 | 3 | 0 | 0 | 2 | 1 |
| 3 | 142619 | 155.2 | 3.16 | 1 | 3 | 1 | 1 | 0 | 12 | 1 | 2 |
| 4 | 142285 | 164.2 | 4 | 3 | 3 | 1 | 0 | 0 | 23 | 2 | 0 |
| 5 | 142158 | 159 | 2.77 | 0 | 3 | 1 | 2 | 50 | 13 | 1 | 0 |
| 6 | 141008 | 155.5 | 3.52 | 4 | 1 | 0 | 3 | 67 | 16 | 1 | 0 |
| 7 | 145082 | 153.8 | 3.9 | 3 | 2 | 1 | 3 | 100 | 5 | 1 | 0 |
| 8 | 139410 | 162.1 | 3.45 | 3 | 3 | 1 | 3 | 0 | 13 | 2 | 1 |
| 9 | 135014 | 154 | 3.67 | 0 | 3 | 1 | 1 | 0 | 18 | 1 | 2 |
| 10 | 139356 | 157.5 | 3.4 | 0 | 1 | 1 | 4 | 25 | 14 | 0 | 3 |
| 11 | 137368 | 160.8 | 3.3 | 1 | 3 | 1 | 0 | 0 | 33 | 2 | 4 |
| 12 | 141982 | 157.3 | 3.85 | 2 | 3 | 1 | 2 | 0 | 8 | 1 | 2 |
| 13 | 144753 | 164.1 | 2.75 | 1 | 2 | 0 | 0 | 0 | 5 | 0 | 2 |
| 14 | 132229 | 152.9 | 3.77 | 1 | 1 | 1 | 3 | 67 | 19 | 2 | 2 |
| 15 | 132744 | 158 | 2.74 | 1 | 2 | 0 | 0 | 0 | 8 | 0 | 0 |
| 16 | 131177 | 154.8 | 3.21 | 1 | 1 | 2 | 0 | 0 | 14 | 2 | 3 |
| 17 | 140074 | 153.3 | 3.13 | 1 | 3 | 1 | 0 | 0 | 18 | 1 | 3 |
| 18 | 135633 | 159.7 | 3.45 | 3 | 2 | 2 | 4 | 0 | 10 | 0 | 0 |
| 19 | 139582 | 155.7 | 3.19 | 2 | 2 | 1 | 5 | 0 | 9 | 1 | 0 |
| 20 | 135197 | 160.7 | 4.43 | 2 | 4 | 1 | 2 | 0 | 6 | 1 | 3 |
| 21 | 131975 | 143.1 | 4.37 | 0 | 3 | 1 | 3 | 33 | 0 | 2 | 3 |

We also can run some summary statistics:

```
-- Variable type: numeric ------------------------------------------------------------------------------------------
# A tibble: 10 x 11
   skim_variable              n_missing complete_rate    mean     sd     p0    p25    p50    p75   p100 hist
 * <chr>                          <int>         <dbl>   <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl> <chr>
 1 Avg.Tix...Day                      0             1 156.     4.42  143.   153.   156.   159.   169.   ▁▆▇▆▁
 2 Customer.rating                    0             1   3.50   0.461   2.07   3.21   3.50   3.81   4.81 ▁▂▇▅▁
 3 Tardies                            0             1   1.46   0.973   0      1      1      2      4    ▃▇▆▂▁
 4 Graveyard.Shifts.Taken             0             1   1.98   0.795   0      1      2      2      4    ▁▃▇▃▁
 5 Weekend.Shifts.Taken               0             1   0.952  0.549   0      1      1      1      2    ▂▁▇▁▂
 6 Sick.Days.Taken                    0             1   1.88   1.67    0      0      2      3      7    ▇▃▅▁▁
 7 X..Sick.Days.Taken.on.Friday       0             1  35.2   39.3     0      0     25     67    100    ▇▁▁▁▂
 8 Employee.Dev..Hours                0             1  12.0    7.47    0      6     12     17     34    ▆▇▇▂▁
 9 Shift.Swaps.Requested              0             1   1.45   1.00    0      1      1      2      5    ▇▅▂▁▁
10 Shift.Swaps.Offered                0             1   1.76   1.81    0      0      1      3      9    ▇▅▂▁▁
```

The average number of tickets per day handled by employees is 156, the median is also 156 and the 75% percentile is 159. The mean percentage of sick days taken on Friday is 35.2, but the median is 25, and so on. In addition, you can see min, max and the 25th percentile for each column. You can even see the histogram of each variable on the far right.

**Discussion questions:**

1. How do we know we have anomalies in the data?
2. How is this different from outlier detection?
3. Can you think of other use cases of anomaly detection?