Part 1:

1. COGNIZANT TECHNOLOGY SOLUTIONS US CORPORATION

```
(Pdb) df.groupby('employer_name')['case_number'].count().sort_values(ascending=False)
employer_name
COGNIZANT TECHNOLOGY SOLUTIONS US CORPORATION    5441
INTEL CORPORATION                                1611
CISCO SYSTEMS, INC.                              1028
GOOGLE INC.                                       914
QUALCOMM TECHNOLOGIES INC.                        862
                                                 ...
LICKING MEMORIAL HEALTH SYSTEMS                     1
LIEBHERR MINING & CONSTRUCTION EQUIPMENT INC.       1
LIFE ALERT EMERGENCY RESPONSE, INC.                 1
LIFE IS MOTION LLC                                  1
xpedx, LLC                                          1
Name: case_number, Length: 17985, dtype: int64
```

2. MICROSOFT CORPORATION

```
(Pdb) df[df.case_status=='Certified-Expired'].groupby('employer_name')['case_number'].count().sort_values(ascending=False)
employer_name
MICROSOFT CORPORATION                     521
CISCO SYSTEMS, INC.                       398
QUALCOMM TECHNOLOGIES INC.                269
GOOGLE INC.                               253
AMAZON CORPORATE LLC                      231
                                         ...
HERBAL POWERS CORP D/B/A HP INGREDIENTS     1
HENNEMAN ENGINEERING INC                    1
HENKEL CORPORATION                          1
HENDRICKSON USA, LLC                        1
ZYXEL COMMUNICATIONS, INC.                  1
Name: case_number, Length: 6501, dtype: int64
```

3. 18
4. Comment in line 28

Part 2:

- Variable dropped = ['case_number','case_received_date','decision_date','case_status','job_experience_num_month,s']
- Job_experience_num_month dropped because of insufficient data
- Number of employee outlier (1400000) reduced to 700000.
- Created variable duration (decision date - received date)
- Job Education was one-hot-encoded (since ordinal), others were cat-coded (since not ordinal)

Top 5 as per variable Importance:

1. job education (x2)
2. Employer_num_employees
3. job state
4. employer name
5. Job_level

Model: Decision Regression Tree
Good with outliers, good with cat. data

CV SCORE 10 FOLD: Negative Mean-Abs-Error
array([-17191.52714435, -14019.15211106, -12256.82915734, -10721.09900318,
-20010.74330329, -19823.53217116, -20056.20842145, -19202.96785878,
-17164.58803806, -16864.36733838])

Not good! To improve, consider creating models for different wage units.