**Intermediate Report of Your Project**

**Due Sept 30, 2021**

Please document everything you did so far to preprocess your data. You need to reason your actions.

If you used Python, please include the code.

Also submit your training dataset and test dataset. (Put under the shared google drive folder or just share with me [syilmaz@fcpsschools.net](mailto:syilmaz@fcpsschools.net).

Remove these ones:

Survey Year

Date is irrelevant

Timestamp

Date is irrelevant

Postal Code

International, Postal code is irrelevant

Other Databases

Mixed data with multiple select

Database Servers

Mixed data with multiple select

Counter

Useless

CompanyEmployeesOverall

Data is not fully numeric, 76% of data is missing

EducationIsComputerRelated

46% of attribute is missing

NewestVersionInProduction

Mixed data with multiple select

PopulationOfLargestCityWithin20Miles

58% of attribute is missing

Attribute Cleaning

**SalaryUSD(Class)**

* **We will discretize by rounding it to the nearest 25000 and normalize it from 0 to 1**

**Country**

* **Will stay as country names**

**Primary Database**

| Microsoft SQL Server | M |
| --- | --- |
| PostgreSQL | P |
| Other |  |
| Azure SQL DB | A |
| Microsoft Access | MA |
| Oracle | O |
| MySQL/MariaDB | MS |
| Amazon RDS (any flavor) | RDS |
| MongoDB | MDB |
| DB2 | DB2 |
| Cassandra | C |
| SQLite | SQLI |
| SAP | SAP |
| Teradata | T |
| Elasticsearch | E |

**Years With this Database**

* **Median Value: 10**
* **Replace unrealistic outliers with median**
  + **53716**
  + **30331**
  + **1050**
* **Fix years with real values**
  + **2000 = 21**
  + **2003 = 18**
  + **2020 = 1**
  + **2017 = 4**
  + **1997 = 24**

**Employment Status**

| **Full time employee** | **E** |
| --- | --- |
| **Full time employee of a consulting/contracting company** | **EC** |
| **Independent consultant, contractor, freelancer, or company owner** | **IC** |
| **Part time** | **P** |
| **Independent or freelancer or company owner** | **I** |

**Job Title**

| **Developer: Business Intelligence (SSRS, PowerBI, etc)** | **D** |
| --- | --- |
| **DBA (Production Focus - build & troubleshoot servers, HA/DR)** | **DBAP** |
| **DBA (General - splits time evenly between writing & tuning queries AND building & troubleshooting servers)** | **DBAG** |
| **Manager** | **M** |
| **Developer: App code (C#, JS, etc)** | **DA** |
| **Developer: T-SQL** | **DT** |
| **Architect** | **A** |
| **DBA (Development Focus - tunes queries, indexes, does deployments)** | **DBAD** |
| **Engineer** | **E** |
| **Analyst** | **AN** |
| **Other** |  |
| **Data Scientist** | **DS** |
| **DBA** | **DBA** |
| **Principal database engineer** | **PDE** |
| **DevOps, Sr Software Engineer DBA** | **DO** |
| **Technician** | **T** |
| **Database Specialist** | **DBS** |
| **Consultant** | **C** |
| **Systems Administrator** | **SA** |
| **Sales** | **S** |
| **DBA / BI Developer** | **DBABI** |
| **Sr Consultant** | **SRC** |
| **Analytics consultant** | **AC** |

**Manage Staff**

* **Yes/No → Y/N**

**Years with this type of job**

**How many Companies?**

* **Converted numbers + comments to numbers**

| **1 (this is the only company where I've had this kind of position)** | **1** |
| --- | --- |
| **5** | **5** |
| **4** | **4** |
| **2 (I worked at another similar position elsewhere before this one)** | **2** |
| **3** | **3** |
| **6 or more** | **6** |
| **Not Asked** |  |

**How many other people are on your team?**

| **1** | **1** |
| --- | --- |
| **5** | **5** |
| **4** | **4** |
| **2** | **2** |
| **3** | **3** |
| **5 or more** | **6** |
| **None** | **0** |

**How many employees does your company have overall?**

* **Median: 20**
* **Replace weird values like 1e+8 with median**

**Employment**

| **Private business** | **PB** |
| --- | --- |
| **Education (K-12, college, university)** | **E** |
| **State/province government** | **SG** |
| **Local government** | **LG** |
| **Non-profit** | **NP** |
| **Federal government** | **FG** |
| **Student** | **T** |

**Gender**

| **Male** | **M** |
| --- | --- |
| **Female** | **F** |
| **Prefer not to say/None** |  |
| **NonBinary** | **NB** |

**CareerPlan**

| **Stay with the same employer, same role** | **S** |
| --- | --- |
| **Prefer not to say** |  |
| **Change both employers and roles** | **C** |
| **Stay with the same role, but change employers** | **SR** |
| **Stay with the same employer, but change roles** | **SE** |

**Years with this kind of job?**

* **Same as years with database except no outliers**

**Education**

* **Unchanged**

**Certs?**

* **No/YesValid/YesExpired**

**Hours**

* **Numerical**

**Telecommunication hours per week**

* **zero/one/two/three/four/five/six <- (six or more)**

Fixing the Class Attribute

1. SalaryUSD is numerical, needs to be discretized, we used quantile bins with pandas to cut our data into 5 equal depth bins

import pandas as pd

df = pd.read\_csv("responses\_clean.csv", sep="|")

labels = ["Low", "Mid", "High", "Very High"]

print(pd.qcut(df['SalaryUSD'], q = 4, labels=labels, retbins=True))

Name: SalaryUSD, Length: 10339, dtype: category

Categories (5, object): ['Very Low' < 'Low' < 'Mid' < 'High' < 'Very High'], array([13.0, 60000.0, 81000.0, 100000.0, 122000.0, 1850000.0])

1. Here are our bins:

| Very Low | [13, 60000) |
| --- | --- |
| Low | [60000, 81000) |
| Mid | [81000, 100000) |
| High | [100000, 122000) |
| Very High | [122000, 1850000) |