Module 2 - Data collection, validation and privacy

Assignment overview

In this assignment, you will be exploring various aspects related to collecting data and identifying bias in datasets. You will also be asked to consider issues of data privacy and governance.

For this assignment, it is possible to work in **groups of up to 2 students**.

Group members

Leave blanks if group has less than 2 members:

- Student 1: Jingyuan Liu (S.N. 69763183)
- Student 2: Nicholas Tam (S.N. 45695970)

Learning Goals:

After completing this week's lecture and tutorial work, you will be able to:

1. Discuss the implications of data governance and data ownership in data science
2. Argue the advantages and disadvantages of collecting individuals' data online
3. Distinguish between a sample and a population, what attributes make a representative sample and the possible ethical implications of a non-representative sample
4. Explain the elements of experimental design
5. Identify possible sources of bias in datasets (such as historical, measurement, and representation bias)
6. Describe the ethical implications of variable choice in data science (e.g., use of proxies, use of gender and race as variables)
7. Apply good practices for minimizing errors in data cleaning
8. Apply methods for improving privacy and anonymity in stored data and data analysis, such as k-anonymity and randomized response
9. Explain the notion of differential privacy

Part 1: Data collection, sampling and bias

In class, we discussed different sources of bias that can affect the data we want to use for our Data Science applications. Here is a summary:

1. Historical bias

Historical bias: bias that exists in society and is reflected in the data. It is the most insidious because it arises even if we are able to perfectly sample from the existing population. Most often, it affects groups that are historically disadvantaged.

E.g. In 2018, 5% of Fortune 500 CEOs were women. Historically, women have less frequently made it to a CEO position. A classifier trained to predict the best choice for a new CEO may learn this pattern and determine that being a woman makes one less qualified to be a CEO.

2. Representation bias

Representation bias: the sample underrepresents part(s) of the population and fails to generalize well. This may happen for different reasons:

1. The sampling methods only reached a portion of the population. E.g. Data collected via smartphone apps can under-represent lower incomes or older groups, who may be less likely to own smartphones.

2. The population of interest has changed or is distinct from the sample used during model training. E.g. Data that is representative of Vancouver may not be representative if used to analyze the population in Toronto. Similarly, data representative of Vancouver 100 years ago may not reflect today's population.

3. Measurement bias

Measurement bias: it occurs when choosing features that fail to correctly represent the problem, or when there are issues with the data collection. Fore example:

1. The measurement processes varies across groups. E.g. one group of workers is monitored more closely and thus more errors are observed in that group.

2. The quality of data varies across groups. E.g. women often self-report less pain than men and are therefore less likely to receive certain diagnoses

3. The defined classification task or one of the features used is an oversimplification. E.g. We are designing a model to predict whether a student will be successful in college. We choose to predict the final GPA as metric of success. This, however, ignores other indicators of success.

Question 1

Consider a crowd-sourcing project called Street Bump aimed at helping improve neighbourhood streets in Boston from 2011 to 2014. Volunteers used a smartphone app, which captured GPS location and reported back to the city everytime the driver hit a pothole. The data was provided to governments so they could use the data to fix any road issues.

Can you think of any sources of bias in the scenario above? Explain them.

There is a risk of representation bias, as the data will likely under-represent lower-income or older groups that are less likely to have smartphones, on top of the population of people that would be interested in volunteering potentially not being representative of the overall population. There is also a risk of measurement bias, as road quality is determined by more attributes than potholes alone, such as effective drainage and traffic management. The frequency of drivers hitting potholes is also determined by other factors, such as the proficiency of the drivers themselves, or the location of the potholes. In other words, the feature used to determine the road quality is an oversimplification.

Refined answer:

There is a risk of representation bias, as the data will likely under-represent lower-income or older groups that are less likely to have smartphones, on top of the population of people that would be interested in volunteering potentially not being representative of the overall population. In addition, people who drive cannot represent all citizens in Boston, particularly in areas where people have more access to public transportation. The data underrepresents the lower-income people who cannot afford a car and those who prefer to take public transportation. For instance, people who have cars and can drive are more likely to volunteer, as the project is more directly relevant to them than otherwise, while those that prefer to not use vehicular transport would be less likely to volunteer. Besides, the distribution of traffic in the area could also influence how frequently drivers report a bump and how often people use vehicular transport in the area, such as those preferring to walk doing so due to the inconvenience of driving through that area. And the streets with more traffic tend to receive more bump reports from drivers because there are more vehicles. There is also a risk of measurement bias, as road quality is determined by more attributes than potholes alone, such as effective drainage and traffic management. The frequency of drivers hitting potholes is also determined by other factors, such as the proficiency of the drivers themselves, or the location of the potholes. In other words, the feature used to determine the road quality is an oversimplification.

Observational and experimental studies

- **Observational study:** study where there is no deliberate human intervention regarding the variable under investigation. Observational studies are ones where researchers observe the effect of a treatment/intervention without trying to change who is or isn’t exposed to it. In an observational study, the subjects are assigned or assign themselves to the exposure group they belong to.
- **Experimental study:** : study that involves planned intervention on the exposure to a condition. In an experiment, subjects are assigned to a condition by the researcher and thus one can establish a cause-and-effect relationship when we see a difference in the outcome between the experimental groups. Randomizing study subjects balances any differences between treatment groups with respect to all variables except the condition of exposure.

A/B testing

A/B testing can be considered the most basic kind of randomized controlled experiment.

Complete the following reading, then answer the comprehension questions below:
https://hbr.org/2017/06/a-refresher-on-ab-testing

Question 2

In the following table, select which statements are true or false:

Statement	True	False
A/B testing is an example of experimental study.	✔	
Observational studies require subjects to not be informed that they are being studied.		✔
Ethical experimental studies require genuine uncertainty about the benefits/harms of treatment or exposure (equipoise)	✔	
A researcher is interested in studying the effects of certain dietary habits. They recruite people and, through a survey, they ask them to disclose their current dietary habits, on which bases they will be assigned to treatment or control group. This is an example of experimantal study.		✔
The control group and the exposed group must include different individuals.		✔
One of the main advantages of experimental studies is that they allow for better randomization.	✔	

Question 3

Explain the role of blocking in A/B testing.

Blocking is defined as splitting the data by similarity in a factor that is of less interest, but will still heavily influence the success metric of our interest. For example, from the article, whether or not someone views a website on mobile or desktop will influence the click rates on both versions of a website, but the groups of interest in the study are the two versions of our website, not the devices of users. In this case, we should first divide the users into blocks for each type of device used, then randomly assign users to each version within each block. Blocking in A/B testing allows for a more accurate reflection of the distinctions between the methods of interest.

Refined answer:

Blocking is defined as splitting the data by similarity in a factor that is of less interest, but will still heavily influence the success metric of our interest. For example, from the article, whether or not someone views a website on mobile or desktop will influence the click rates on both versions of a website, but the groups of interest in the study are the two versions of our website, not the devices of users. In this case, we should first divide the users into blocks for each type of device used, then randomly assign users to each version within each block. Blocking in A/B testing allows for a more accurate reflection of the distinctions between the methods of interest by reducing the influence of irrelevant factors, especially those that cannot be mitigated by randomization, and thus emphasizing the effect of each method on the results. In other words, it controls the variation in the data due to the blocking variable by replacing it with the variation within each block. It makes it easier to detect the real effect of the groups of interest.

Question 4

The authors warn about observing too many metrics when running an A/B test. Why is that the case? What could happen if I ignore this warning?

Observing too many metrics runs the risk of observing "spurious correlations", where multiple variables are only seemingly correlated without being causally related. The more metrics we observe, the more likely we will see some statistically significant results that only happen by chance, which is as what Fund described as "random fluctuation". Ignoring the warning will lead to some incorrect or misleading conclusions, making the interpretation of results difficult due to too many metrics influencing changes in data all at once. For example, you may want to switch to the new version of the product because you found some metrics significant from the A/B testing. But if you have too many metrics, it is more likely that some significant metrics occur only by chance. In this case, if you make a decision to switch the product to the new version based on this result, the new version may at best not be any more effective than the original one.

Question 5

You want to determine the size of the subscribe button on your website. You plan to evaluate the performance by the number of visitors who click on the button. To run the test, you show one set of users one version and collect information about the number of visitors who click on the button. One month later you show users another version where the only thing different is the size of the button. Based on this test, you determine that the second version had a higher number of visitors who clicked on the button. Can you conclude that this version of the website leads to a higher number of visitors clicking on the button? Briefly explain.

I would argue that we cannot conclude that this version of the website leads to a higher number of visitors clicking on the button. There is no statistic provided to indicate that the difference in button clicks is statistically significant enough to reject the null hypothesis that the number of clicks for both websites is the same. More importantly, as the test was conducted in two different periods, there might be some other variables that could potentially influence the results also changing over time (i.e. users' mood, seasonal effect, etc). The data we collected for each version of the website may also be representative of different populations due to the difference in time frames, leading to representation bias. Therefore, we should conduct this test simultaneously by randomly assigning users to one of the versions, minimizing the effect of other variables on the result.

Ethical A/B testing

Ethical A/B testing still requires all the ethical considerations of any experimental study, such as informed consent or possibility to opt out. A notorious case of a company failing to meet ethics requirement in A/B testing is the infamous Facebook "social contagion experiment", in which almost 700,000 users were showed, for a week, only positive or only negative content, to see how this variation impacted their online behaviour. The selected users were not informed and could not opt out. Furthermore, their emotional state was affected. Facebook defended itself by saying that Facebook's Data Use Policy warns users that Facebook “may use the information we receive about you…for internal operations, including troubleshooting, data analysis, testing, research and service improvement”. This defense was largely rejected by the scientific community, which still considered the study as unethical. You can read more about this incident in this article.

Case Study: National Institute of Justice's (NIJ) Recidivism Dataset

We will now look at the NIJ's Recidivism data set, which contains data on 26,000 individuals from the State of Georgia released from prison on parole (early release from prison where the person agrees to abide by certain conditions) between January 1, 2013 and December 31, 2015. **Recidivism** is the act of committing another crime.

This dataset is split into two sets, training and test, 70% of the data is in the training dataset and 30% in the test dataset. The training set contains four variables that measure recidivism: whether an individual recidivated within three years of the supervision start date and whether they recidivated in year 1, year 2, or year 3. In this data set, recidivism is defined as being arrested for a new crime during this three-year period. The test set does not include these four variables.

The data was provided by the Georgia Department of Community Supervision (GDCS) and the Georgia Bureau of Investigation.

Source: https://data.ojp.usdoj.gov/stories/s/daxx-hznc

Let's start by familiarizing with the dataset source. The website includes a lot of information on the dataset and a detailed description of each of its columns (look for Appendix 2: Codebook).

Question 6 Think about how the data set was collected and what we are trying to predict. Are there any potential sources of bias (historical, representation, measurement)? Explain your answer.

- Historical: The historical bias against some certain racial groups (i.e. Black people) can affect the performance of the model nowadays, reflecting past inequalities and unfairness.
- Representation: The population of individuals used for model training would have changed over time, such as the proportions of people at certain ages on release, and thus may not be reflective of the current population of people on parole. In addition, as the data only collects individuals from the State of Georgia, people released from prison on parole from other states or countries may be underrepresented. The data was also collected between 2013 and 2015, which may not be representative of the population of

interests nowadays, due to social and political events and changes over time, such as COVID-19.

- Measurement: In this study, recidivism is defined as being arrested for a new crime during this three-year period. By contrast, recidivism is generally defined as the act of relapsing into criminal behaviour, often after intervention for a previous crime. As such, the study's definition excludes criminals that have reoffended but have not been arrested for such. Factors such as race and age may also lead to unfairness in measurements and quality of data between groups of individuals, such as stricter monitoring between those in the same supervision level due to racism or prioritization of criminal acts. In other words, the feature used to determine recidivism is an oversimplification.

Question 7: Exploratory Data Analysis (EDA)

We are now going to perform some Exploratory Data Analysis on the NIJ's Recidivism Training set. This will serve 2 purposes:

- it will help us familiarize with the dataset
- it will help us spot possible imbalances or sources of bias in the dataset

You are free to use tools and functions of your choice to complete the EDA. Your goal is to answer the following questions:

1. Does the dataset include protected characteristics? We recommend using the BC Human Rights Code for reference.
2. If the dataset includes protected characteristic, do you think they are necessary to perform the predictive task? Why or why not?
3. If we were to remove the columns including protected characteristics, do you think it would still be possible to retrieve that information through other features (proxies)? Explain how.
4. Is the target variable balanced? If not, what could happen?
5. Is the target variable balanced *across protected segments of the population?* What could happen if this is not the case?
6. Are there features with missing values? Do you suspect that they may be Missing Not At Random (MNAR), and if so, how would it be best to fill this information?

Notes:

- Bar charts and other plots are helpful to visually spot imbalances
- You are encouraged to talk to the instructor and TA to discuss your EDA strategy and if you need suggestions with the code

```
# Your solution here. You may add more code/markdown cells as needed.
import pandas as pd

train_df =
pd.read_csv("NIJ_s_Recidivism_Challenge_Training_Dataset.csv")
train_df.head()

   ID Gender   Race Age_at_Release  Residence_PUMA Gang_Affiliated  \
0   1      M  BLACK          43-47              16           False
```

```
1   2      M  BLACK          33-37                16           False
2   3      M  BLACK    48 or older                24           False
3   4      M  WHITE          38-42                16           False
4   5      M  WHITE          33-37                16           False

   Supervision_Risk_Score_First Supervision_Level_First  \
0                           3.0                Standard
1                           6.0             Specialized
2                           7.0                    High
3                           7.0                    High
4                           4.0             Specialized

         Education_Level Dependents  ... DrugTests_Cocaine_Positive  \
0  At least some college  3 or more  ...                        0.0
1   Less than HS diploma          1  ...                        0.0
2  At least some college  3 or more  ...                        0.0
3   Less than HS diploma          1  ...                        0.0
4   Less than HS diploma  3 or more  ...                        0.0

  DrugTests_Meth_Positive DrugTests_Other_Positive
Percent_Days_Employed  \
0                0.000000                      0.0
0.488562
1                0.000000                      0.0
0.425234
2                0.166667                      0.0
0.000000
3                0.000000                      0.0
1.000000
4                0.058824                      0.0
0.203562

  Jobs_Per_Year Employment_Exempt Recidivism_Within_3years  \
0      0.447610             False                    False
1      2.000000             False                     True
2      0.000000             False                     True
3      0.718996             False                    False
4      0.929389             False                     True

  Recidivism_Arrest_Year1  Recidivism_Arrest_Year2
Recidivism_Arrest_Year3
0                   False                    False
False
1                   False                    False
True
2                   False                     True
False
3                   False                    False
False
4                    True                    False
```

```
False

[5 rows x 53 columns]

display(train_df.describe())

                 ID  Residence_PUMA  Supervision_Risk_Score_First  \
count  18028.000000    18028.000000                  17698.000000
mean   13386.065343       12.307577                      6.064753
std     7721.451992        7.143255                      2.382811
min        1.000000        1.000000                      1.000000
25%     6702.750000        6.000000                      4.000000
50%    13405.500000       12.000000                      6.000000
75%    20081.250000       18.000000                      8.000000
max    26761.000000       25.000000                     10.000000

       Avg_Days_per_DrugTest  DrugTests_THC_Positive  \
count           13768.000000            14396.000000
mean               93.585860                0.063120
std               117.561341                0.138357
min                 0.500000                0.000000
25%                28.666667                0.000000
50%                55.000000                0.000000
75%               110.000000                0.068242
max              1087.000000                1.000000

       DrugTests_Cocaine_Positive  DrugTests_Meth_Positive  \
count                14396.000000             14396.000000
mean                     0.014173                 0.012768
std                      0.063473                 0.059572
min                      0.000000                 0.000000
25%                      0.000000                 0.000000
50%                      0.000000                 0.000000
75%                      0.000000                 0.000000
max                      1.000000                 1.000000

       DrugTests_Other_Positive  Percent_Days_Employed  Jobs_Per_Year

count              14396.000000           17721.000000   17494.000000

mean                   0.007681               0.480035       0.766423

std                    0.042224               0.424396       0.813474

min                    0.000000               0.000000       0.000000

25%                    0.000000               0.000000       0.000000

50%                    0.000000               0.466543       0.636324

75%                    0.000000               0.966184       1.000000
```

```
max                    1.000000               1.000000      8.000000
```

```
# 1, 2: Protected charactersitics
display(train_df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 18028 entries, 0 to 18027
Data columns (total 53 columns):
 #   Column                                             Non-Null Count
Dtype
---  ------                                             --------------
-----
 0   ID                                                 18028 non-null
int64
 1   Gender                                             18028 non-null
object
 2   Race                                               18028 non-null
object
 3   Age_at_Release                                     18028 non-null
object
 4   Residence_PUMA                                     18028 non-null
int64
 5   Gang_Affiliated                                    15811 non-null
object
 6   Supervision_Risk_Score_First                       17698 non-null
float64
 7   Supervision_Level_First                            16816 non-null
object
 8   Education_Level                                    18028 non-null
object
 9   Dependents                                         18028 non-null
object
 10  Prison_Offense                                     15707 non-null
object
 11  Prison_Years                                       18028 non-null
object
 12  Prior_Arrest_Episodes_Felony                       18028 non-null
object
 13  Prior_Arrest_Episodes_Misd                         18028 non-null
object
 14  Prior_Arrest_Episodes_Violent                      18028 non-null
object
 15  Prior_Arrest_Episodes_Property                     18028 non-null
object
 16  Prior_Arrest_Episodes_Drug                         18028 non-null
object
 17  Prior_Arrest_Episodes_PPViolationCharges           18028 non-null
object
```

```
 18  Prior_Arrest_Episodes_DVCharges                    18028 non-null
bool
 19  Prior_Arrest_Episodes_GunCharges                   18028 non-null
bool
 20  Prior_Conviction_Episodes_Felony                   18028 non-null
object
 21  Prior_Conviction_Episodes_Misd                     18028 non-null
object
 22  Prior_Conviction_Episodes_Viol                     18028 non-null
bool
 23  Prior_Conviction_Episodes_Prop                     18028 non-null
object
 24  Prior_Conviction_Episodes_Drug                     18028 non-null
object
 25  Prior_Conviction_Episodes_PPViolationCharges       18028 non-null
bool
 26  Prior_Conviction_Episodes_DomesticViolenceCharges  18028 non-null
bool
 27  Prior_Conviction_Episodes_GunCharges               18028 non-null
bool
 28  Prior_Revocations_Parole                           18028 non-null
bool
 29  Prior_Revocations_Probation                        18028 non-null
bool
 30  Condition_MH_SA                                    18028 non-null
bool
 31  Condition_Cog_Ed                                   18028 non-null
bool
 32  Condition_Other                                    18028 non-null
bool
 33  Violations_ElectronicMonitoring                    18028 non-null
bool
 34  Violations_Instruction                             18028 non-null
bool
 35  Violations_FailToReport                            18028 non-null
bool
 36  Violations_MoveWithoutPermission                   18028 non-null
bool
 37  Delinquency_Reports                                18028 non-null
object
 38  Program_Attendances                                18028 non-null
object
 39  Program_UnexcusedAbsences                          18028 non-null
object
 40  Residence_Changes                                  18028 non-null
object
 41  Avg_Days_per_DrugTest                              13768 non-null
float64
 42  DrugTests_THC_Positive                             14396 non-null
```

```
float64
 43  DrugTests_Cocaine_Positive                                          14396 non-null
float64
 44  DrugTests_Meth_Positive                                             14396 non-null
float64
 45  DrugTests_Other_Positive                                            14396 non-null
float64
 46  Percent_Days_Employed                                               17721 non-null
float64
 47  Jobs_Per_Year                                                       17494 non-null
float64
 48  Employment_Exempt                                                   18028 non-null
bool
 49  Recidivism_Within_3years                                            18028 non-null
bool
 50  Recidivism_Arrest_Year1                                             18028 non-null
bool
 51  Recidivism_Arrest_Year2                                             18028 non-null
bool
 52  Recidivism_Arrest_Year3                                             18028 non-null
bool
dtypes: bool(20), float64(8), int64(2), object(23)
memory usage: 4.9+ MB

None
```

Q7.1: From the list of columns provided by `train_df.info()`, the columns that likely include protected characteristics are `Gender`, `Race`, `Age_at_Release`, `Dependents`, and the characteristics involving prior arrests, convictions and revocations (`Prior_Arrest_Episodes_Felony`, `Prior_Arrest_Episodes_Misdemeanor`, `Prior_Arrest_Episodes_Violent`, `Prior_Arrest_Episodes_Property`, `Prior_Arrest_Episodes_Drug`, `Prior_Arrest_Episodes_PPViolationCharges`, `Prior_Arrest_Episodes_DomesticViolenceCharges`, `Prior_Arrest_Episodes_GunCharges`, `Prior_Conviction_Episodes_Felony`, `Prior_Conviction_Episodes_Misdemeanor`, `Prior_Conviction_Episodes_Violent`, `Prior_Conviction_Episodes_Property`, `Prior_Conviction_Episodes_Drug`, `Prior_Conviction_Episodes_PPViolationCharges`, `Prior_Conviction_Episodes_DomesticViolenceCharges`, `Prior_Conviction_Episodes_GunCharges`, `Prior_Revocations_Parole`, and `Prior_Revocations_Probation`). Q7.2: The `Gender`, `Race`, `Age_at_Release`, and `Dependents` characteristics are unnecessary as they appear to only be indirectly correlated to their probability of a person undergoing recidivism. The characteristics relating to prior arrests, convictions, and revocations appear to be more directly related to a person recommitting crimes and thus may be necessary for the predictive task. Q7.3: It could be possible to retieve an individual's `Age_at_Release`, `Prior_Arrest_Episodes_Drug`, and `Prior_Conviction_Episodes_Drug`. The former characteristic could be inferred from characteristics that would likely tie to someone's work such as `Percent_Days_Employed` and `Jobs_Per_Year`, under the assumption that a younger individual would likely have lower values for both of those characteristics. The latter two characteristics are potentially associated with other characteristics involving drug testing, as individuals with a history of drug abuse would likely be closely monitored for potential relapses.

Refined answer:

Q7.1: From the list of columns provided by `train_df.info()`, the columns that likely include protected characteristics are `Gender`, `Race`, `Age_at_Release`, `Dependents`, the characteristics involving prior arrests, convictions and revocations (`Prior_Arrest_Episodes_Felony`, `Prior_Arrest_Episodes_Misdemeanor`, `Prior_Arrest_Episodes_Violent`, `Prior_Arrest_Episodes_Property`, `Prior_Arrest_Episodes_Drug`, `Prior_Arrest_Episodes_PPViolationCharges`, `Prior_Arrest_Episodes_DomesticViolenceCharges`, `Prior_Arrest_Episodes_GunCharges`, `Prior_Conviction_Episodes_Felony`, `Prior_Conviction_Episodes_Misdemeanor`, `Prior_Conviction_Episodes_Violent`, `Prior_Conviction_Episodes_Property`, `Prior_Conviction_Episodes_Drug`, `Prior_Conviction_Episodes_PPViolationCharges`, `Prior_Conviction_Episodes_DomesticViolenceCharges`, `Prior_Conviction_Episodes_GunCharges`, `Prior_Revocations_Parole`, and `Prior_Revocations_Probation`) and the conditions of supervision (`Condition_MH_SA`, `Condition_Cog_Ed`, `Condition_Other`). Q7.2: The `Gender`, `Race`, `Age_at_Release`, and `Dependents` characteristics, along with the conditions of supervision, are unnecessary as they appear to only be indirectly related to a person's probability of undergoing recidivism. For instance, a person with a higher value for `Dependents` may recommit crimes to help sustain those under their care. The characteristics relating to prior arrests, convictions, and revocations appear to be more directly related to a person recommitting crimes and thus may be necessary for the predictive task. Q7.3: It could be possible to retrieve an individual's `Race`, `Age_at_Release`, `Education_level`, `Dependents`, `Prior_Arrest_Episodes_Drug`, and `Prior_Conviction_Episodes_Drug`. `Race` could be retrieved from a combination of characteristics `Residence_PUMA*` and `Gang_Affiliated` because the residential area may be able to represent the demographic information of a person such as race because people from the same racial group tend to resident under the same area. And whether the person is gang-affiliated can also help because some racial groups may be more likely to form a gang. `Age_at_Release` could be inferred from other characteristics that would likely tie to someone's work, such as `Percent_Days_Employed` and `Jobs_Per_Year`, under the assumption that a younger individual would likely have lower values for both of those characteristics. Similarly, `Education_level` and `Dependents` would also influence work-related characteristics, such as those with a higher `Education_level` remaining employed for longer, and those with higher `Dependents` values having more `Jobs_Per_Year`. The latter two characteristics are potentially associated with other characteristics involving drug testing, as individuals with a history of drug abuse would likely be closely monitored for potential relapses.

```python
# 4: Check if the class distribution is balanced
display(train_df["Recidivism_Within_3years"].value_counts(normalize=True),
```

```
train_df["Recidivism_Arrest_Year1"].value_counts(normalize=True),

train_df["Recidivism_Arrest_Year2"].value_counts(normalize=True),

train_df["Recidivism_Arrest_Year3"].value_counts(normalize=True))
```

```
Recidivism_Within_3years
True     0.578045
False    0.421955
Name: proportion, dtype: float64

Recidivism_Arrest_Year1
False    0.701742
True     0.298258
Name: proportion, dtype: float64

Recidivism_Arrest_Year2
False    0.819558
True     0.180442
Name: proportion, dtype: float64

Recidivism_Arrest_Year3
False    0.900655
True     0.099345
Name: proportion, dtype: float64
```

Q7.4: The target variable `Recidivism_Within_3years` is not balanced, with 57.8% of samples having `Recidivism_Within_3years == True`, and 42.2% of samples having `Recidivism_Within_3years == False`. The proportions of recidivism in each of the trees are not balanced either. As a result, the model may have biased predictive results in favor of the more frequent class of the target variable.

```
# 5: Check if class distribution is balanced within protected segments
for gender in train_df["Gender"].unique():
    print("Recidivism_Within_3years for gender:" + gender)
    display(train_df[train_df["Gender"] == gender]
["Recidivism_Within_3years"].value_counts(normalize=True))

for race in train_df["Race"].unique():
    print("Recidivism_Within_3years for race:" + race)
    display(train_df[train_df["Race"] == race]
["Recidivism_Within_3years"].value_counts(normalize=True))

for age in train_df["Age_at_Release"].unique():
    print("Recidivism_Within_3years for age group:" + age)
    display(train_df[train_df["Age_at_Release"] == age]
["Recidivism_Within_3years"].value_counts(normalize=True))

for dep in train_df["Dependents"].unique():
    print("Recidivism_Within_3years for dependent groups:" + dep)
```

```
    display(train_df[train_df["Dependents"] == dep]
["Recidivism_Within_3years"].value_counts(normalize=True))
```

```
Recidivism_Within_3years for gender:M

Recidivism_Within_3years
True     0.595155
False    0.404845
Name: proportion, dtype: float64

Recidivism_Within_3years for gender:F

Recidivism_Within_3years
False    0.543978
True     0.456022
Name: proportion, dtype: float64

Recidivism_Within_3years for race:BLACK

Recidivism_Within_3years
True     0.589159
False    0.410841
Name: proportion, dtype: float64

Recidivism_Within_3years for race:WHITE

Recidivism_Within_3years
True     0.563189
False    0.436811
Name: proportion, dtype: float64

Recidivism_Within_3years for age group:43-47

Recidivism_Within_3years
True     0.503229
False    0.496771
Name: proportion, dtype: float64

Recidivism_Within_3years for age group:33-37

Recidivism_Within_3years
True     0.57479
False    0.42521
Name: proportion, dtype: float64

Recidivism_Within_3years for age group:48 or older

Recidivism_Within_3years
False    0.587656
True     0.412344
Name: proportion, dtype: float64

Recidivism_Within_3years for age group:38-42
```

```
Recidivism_Within_3years
True     0.537745
False    0.462255
Name: proportion, dtype: float64

Recidivism_Within_3years for age group:18-22

Recidivism_Within_3years
True     0.719395
False    0.280605
Name: proportion, dtype: float64

Recidivism_Within_3years for age group:23-27

Recidivism_Within_3years
True     0.666574
False    0.333426
Name: proportion, dtype: float64

Recidivism_Within_3years for age group:28-32

Recidivism_Within_3years
True     0.6196
False    0.3804
Name: proportion, dtype: float64

Recidivism_Within_3years for dependent groups:3 or more

Recidivism_Within_3years
True     0.54828
False    0.45172
Name: proportion, dtype: float64

Recidivism_Within_3years for dependent groups:1

Recidivism_Within_3years
True     0.605972
False    0.394028
Name: proportion, dtype: float64

Recidivism_Within_3years for dependent groups:2

Recidivism_Within_3years
True     0.582845
False    0.417155
Name: proportion, dtype: float64

Recidivism_Within_3years for dependent groups:0

Recidivism_Within_3years
True     0.585462
```

```
False    0.414538
Name: proportion, dtype: float64
```

Q7.5: The target variable `Recidivism_Within_3years` is not balanced across most protected segments, nor are the distributions of each `Recidivism_Within_3years` category equal across each level of protected segments. For instance, the proportion of `Recidivism_Within_3years` being true is 59.5% among male individuals and 54.4% among female individuals; the proportion of `Recidivism_Within_3years` being true is 72% among age group:18-22 individuals and 50% among age group:43-47 individuals. This runs the risk of differential treatment and measurement of recidivism between categories of protected characteristics and increases the predictive bias against certain groups under protected characteristics

```
# 6: Presence of NaN
# https://stackoverflow.com/questions/36226083/how-to-find-which-columns-contain-any-nan-value-in-pandas-dataframe
display(train_df.isna().any())
```

```
ID                                                   False
Gender                                               False
Race                                                 False
Age_at_Release                                       False
Residence_PUMA                                       False
Gang_Affiliated                                       True
Supervision_Risk_Score_First                          True
Supervision_Level_First                               True
Education_Level                                      False
Dependents                                           False
Prison_Offense                                        True
Prison_Years                                         False
Prior_Arrest_Episodes_Felony                         False
Prior_Arrest_Episodes_Misd                           False
Prior_Arrest_Episodes_Violent                        False
Prior_Arrest_Episodes_Property                       False
Prior_Arrest_Episodes_Drug                           False
Prior_Arrest_Episodes_PPViolationCharges             False
Prior_Arrest_Episodes_DVCharges                      False
Prior_Arrest_Episodes_GunCharges                     False
Prior_Conviction_Episodes_Felony                     False
Prior_Conviction_Episodes_Misd                       False
Prior_Conviction_Episodes_Viol                       False
Prior_Conviction_Episodes_Prop                       False
Prior_Conviction_Episodes_Drug                       False
Prior_Conviction_Episodes_PPViolationCharges         False
Prior_Conviction_Episodes_DomesticViolenceCharges    False
Prior_Conviction_Episodes_GunCharges                 False
Prior_Revocations_Parole                             False
Prior_Revocations_Probation                          False
Condition_MH_SA                                      False
```

```
Condition_Cog_Ed                                        False
Condition_Other                                         False
Violations_ElectronicMonitoring                         False
Violations_Instruction                                  False
Violations_FailToReport                                 False
Violations_MoveWithoutPermission                        False
Delinquency_Reports                                     False
Program_Attendances                                     False
Program_UnexcusedAbsences                               False
Residence_Changes                                       False
Avg_Days_per_DrugTest                                    True
DrugTests_THC_Positive                                   True
DrugTests_Cocaine_Positive                               True
DrugTests_Meth_Positive                                  True
DrugTests_Other_Positive                                 True
Percent_Days_Employed                                    True
Jobs_Per_Year                                            True
Employment_Exempt                                       False
Recidivism_Within_3years                                False
Recidivism_Arrest_Year1                                 False
Recidivism_Arrest_Year2                                 False
Recidivism_Arrest_Year3                                 False
dtype: bool
```

Q7.6: The columns `Gang_Affiliated`, `Supervision_Risk_Score_First`, `Supervision_Level_First`, `Prison_Offense`, `Avg_Days_per_DrugTest`, `DrugTests_THC_Positive`, `DrugTests_Cocaine_Positive`, `DrugTests_Meth_Positive`, `DrugTests_Other_Positive`, `Percent_Days_Employed`, and `Jobs_Per_Year` contain missing values. Of these characteristics, `Gang_Affiliated`, `Supervision_Risk_Score_First`, `Supervision_Level_First`, and `Prison_Offense` are categorical, while the rest are numerical. The variables `Gang_Affiliated`, `Avg_Days_per_DrugTest`, and `Jobs_Per_Year` may be MNAR, as they may not applicable to the individual (e.g. `Avg_Days_per_DrugTest` for someone that never got tested for drugs in the first place), or actively refused to disclose such information (e.g. `Gang_Affiliated`). `Gang_Affiliated` and `Prison_Offense` can have their information filled by creating/using a separate "Other" category, while `Avg_Days_per_DrugTest` can be filled with a default value of 0 to indicate a lack of drug testing in the first place.

Part 2: Privacy

When collecting data for a study, privacy is almost always a primary concern. Our data set may include information that makes it possible to identify an individual, including:

- **Direct identifiers**, which are the ones that can be used to uniquely identify an individual or a household in a dataset, such as a record ID number, patient number, social insurance number, full address, etc. Usually, name is also considered a direct identifier (although several people can have the same name). Other features such as age, date of birth, or postal code are not sufficient on their own to uniquely identify an individual and would not be considered direct identifiers.

- **Indirect (or quasi) identifiers**, which are the columns that do not themselves identify any individual or household, but can do so when combined with other indirect-identifiers. For example, postal code and date of birth are often indirect identifiers, because it is very likely that within a zip code only one individual has this particular birth date. The more indirect identifiers that you have, the more likely it is that individuals become identifiable because there are more possible unique combinations of identifying features.

Question 8

1. Which columns in the NIJ dataset are direct identifiers? Briefly motivate your answer.
2. Which of the remaining columns make good candidates for indirect identifiers? Which ones do not?

Hint: It can be useful to use the `nunique()` and `value_counts()` dataframe methods to get an idea of how many distinct values a feature has.

```
# Your answer here (code portion)
display(train_df.nunique())
display(train_df.shape)
```

```
ID                                                    18028
Gender                                                    2
Race                                                      2
Age_at_Release                                            7
Residence_PUMA                                           25
Gang_Affiliated                                           2
Supervision_Risk_Score_First                             10
Supervision_Level_First                                   3
Education_Level                                           3
Dependents                                                4
Prison_Offense                                            5
Prison_Years                                              4
Prior_Arrest_Episodes_Felony                             11
Prior_Arrest_Episodes_Misd                                7
Prior_Arrest_Episodes_Violent                             4
Prior_Arrest_Episodes_Property                            6
Prior_Arrest_Episodes_Drug                                6
Prior_Arrest_Episodes_PPViolationCharges                  6
Prior_Arrest_Episodes_DVCharges                           2
Prior_Arrest_Episodes_GunCharges                          2
Prior_Conviction_Episodes_Felony                          4
Prior_Conviction_Episodes_Misd                            5
Prior_Conviction_Episodes_Viol                            2
Prior_Conviction_Episodes_Prop                            4
Prior_Conviction_Episodes_Drug                            3
Prior_Conviction_Episodes_PPViolationCharges              2
Prior_Conviction_Episodes_DomesticViolenceCharges         2
Prior_Conviction_Episodes_GunCharges                      2
```

```
Prior_Revocations_Parole                                 2
Prior_Revocations_Probation                              2
Condition_MH_SA                                          2
Condition_Cog_Ed                                         2
Condition_Other                                          2
Violations_ElectronicMonitoring                          2
Violations_Instruction                                   2
Violations_FailToReport                                  2
Violations_MoveWithoutPermission                         2
Delinquency_Reports                                      5
Program_Attendances                                     11
Program_UnexcusedAbsences                                4
Residence_Changes                                        4
Avg_Days_per_DrugTest                                 7654
DrugTests_THC_Positive                                 311
DrugTests_Cocaine_Positive                             203
DrugTests_Meth_Positive                                201
DrugTests_Other_Positive                               197
Percent_Days_Employed                                 7915
Jobs_Per_Year                                         3044
Employment_Exempt                                        2
Recidivism_Within_3years                                 2
Recidivism_Arrest_Year1                                  2
Recidivism_Arrest_Year2                                  2
Recidivism_Arrest_Year3                                  2
dtype: int64

(18028, 53)
```

- Q8.1: `ID` is the only column in the NIJ dataset that is a direct identifier, as the number of unique values in the training dataset is equal to the number of individuals in the dataset, which is 18028.

- Q8.2: `Gender`, `Race`, `Age_at_Release`, `Residence_PUMA`, `Education_Level`, and `Dependents` are effective as indirect identifiers, as they are unlikely to change drastically over extended periods of time and can be used to narrow down the individuals of interest; we can use a combination of these features to identify individuals of interest. The characteristics relating to supervision activities, from `Violations_ElectronicMonitoring` to `Employment_Exempt`, would make for poor candidates for indirect identifiers, as they are directly measured during parole and thus are unlikely to be matched with other anonymous data.

Refined answer:

- Q8.2: `Gender`, `Race`, `Age_at_Release`, `Residence_PUMA`, `Education_Level`, and `Dependents` are effective as indirect identifiers, as they are unlikely to change drastically over extended periods of time and can be used to narrow down the individuals of interest; we can use a combination of these features to identify individuals of interest. `Residence_PUMA` is particularly effective for identification, as among the listed identifiers, it has the greatest number of unique values and is relatively unique to the individuals living within each area, so the list of people of interest will be drastically narrowed down given `Residence_PUMA`. The characteristics relating to supervision activities, from `Violations_ElectronicMonitoring` to `Employment_Exempt`, would make for poor candidates for indirect identifiers, as they are directly measured during parole and thus are unlikely to be matched with other anonymous data.

De-identification of structured data

To safeguard the privacy of the individuals in our dataset, we need to make sure that they are not identifiable, either directly or indirectly. There are three main strategies to achieve this: suppression, pseudonymization, and generalization.

Suppression

Suppression is an effective way to get rid of a direct identifier by simply removing the entire column.

Question 9: using the appropriate dataframe methods, suppress all direct identifier in the NIJ training set. Save the result in a new dataframe called `suppressed_df`

```
# Your answer here
direct_id = ["ID"]
suppressed_df = train_df.drop(columns=direct_id)
suppressed_df.head()
```

```
  Gender   Race Age_at_Release  Residence_PUMA Gang_Affiliated  \
0      M  BLACK          43-47              16           False
1      M  BLACK          33-37              16           False
2      M  BLACK    48 or older              24           False
3      M  WHITE          38-42              16           False
4      M  WHITE          33-37              16           False

   Supervision_Risk_Score_First Supervision_Level_First  \
0                           3.0                Standard
1                           6.0             Specialized
2                           7.0                    High
3                           7.0                    High
4                           4.0             Specialized

          Education_Level Dependents   Prison_Offense  ...  \
0  At least some college  3 or more             Drug  ...
1    Less than HS diploma          1  Violent/Non-Sex  ...
2  At least some college  3 or more             Drug  ...
3    Less than HS diploma          1         Property  ...
4    Less than HS diploma  3 or more  Violent/Non-Sex  ...

  DrugTests_Cocaine_Positive DrugTests_Meth_Positive
DrugTests_Other_Positive  \
0                        0.0                0.000000
0.0
```

```
1                        0.0               0.000000
0.0
2                        0.0               0.166667
0.0
3                        0.0               0.000000
0.0
4                        0.0               0.058824
0.0

  Percent_Days_Employed Jobs_Per_Year Employment_Exempt  \
0              0.488562      0.447610             False
1              0.425234      2.000000             False
2              0.000000      0.000000             False
3              1.000000      0.718996             False
4              0.203562      0.929389             False

  Recidivism_Within_3years  Recidivism_Arrest_Year1
Recidivism_Arrest_Year2  \
0                    False                    False
False
1                     True                    False
False
2                     True                    False
True
3                    False                    False
False
4                     True                     True
False

  Recidivism_Arrest_Year3
0                   False
1                    True
2                   False
3                   False
4                   False

[5 rows x 52 columns]
```

Pseudonymization

A big issue with suppression of direct identifier is that it is not reversible. If at some point we need to identify an individual in our dataset, we would be out of luck. If you have reasons to believe that re-identification may be required, pseudonymization would be a better option to handle direct identifiers. Pseudonymization replaces one or more direct identifiers with a unique but less meaningful value. Usually when we pseudonymize an identifier, there is a possibility of re-identification if required (but it would not be available to the general public).

Question 10: pseudomyze the ID column of the NIJ training set and save the result in a new dataframe called `pseudo_df`. In a different code cell, show that it is possible to re-identify the samples by converting them back to the original ID number.

There are different ways to achieve this you may want to explore:

- Write your own pseudonymization function. You should write at least 2 functions: one to pseudomyze, and another to re-identify. The function does not have to be exceedingly complex but it should not be obvious either (e.g. only basic arithmetic involved).
- Use an extisting library, such as `cryptography`.

Q10 with `cryptography`

```python
# Your answer here (you may add more cells as needed)
from cryptography.fernet import Fernet

# define the pseudomyze function:
def psuedo_encry(col):
    key = Fernet.generate_key()
    f = Fernet(key)
    result1 = col.apply(lambda x: x.to_bytes(2, byteorder='big'))
    result2 = result1.apply(lambda x: f.encrypt(x))
    print("Data encrypted")
    return result2, f

# define the re-identify function:
def psuedo_decry(col, f):
    result1 = col.apply(lambda x: f.decrypt(x))
    result2 = result1.apply(lambda x: int.from_bytes(x,
byteorder='big'))
    print("Data decrypted")
    return result2

# Pseudomyzation
pseudo_df = train_df.copy()
pseudo_df["ID"], f = psuedo_encry(train_df["ID"])
pseudo_df.head()
```

```
Data encrypted
```

```
                                                 ID Gender   Race  \
0  b'gAAAAABm_XYqmYVnvucjE2NcBjMWTA3Un6XkvdlPxoo0...      M  BLACK
1  b'gAAAAABm_XYq9O9cphmcA2ues8IiOlNR3Zlo06t8Bsdh...      M  BLACK
2  b'gAAAAABm_XYqt0DbWEqwM20sDHbXmP8ehdRV6ANlf_gm...      M  BLACK
3  b'gAAAAABm_XYqiiGywjCh4Pd8-L-xrg-Xm4CtBZFlkSBi...      M  WHITE
4  b'gAAAAABm_XYq6sfLACN_zlFMZ6WYD9LjaAsFcKG6b_if...      M  WHITE

  Age_at_Release  Residence_PUMA Gang_Affiliated  \
0          43-47              16           False
1          33-37              16           False
2    48 or older              24           False
3          38-42              16           False
4          33-37              16           False

   Supervision_Risk_Score_First Supervision_Level_First  \
```

```
0                           3.0             Standard
1                           6.0          Specialized
2                           7.0                 High
3                           7.0                 High
4                           4.0          Specialized

         Education_Level Dependents  ... DrugTests_Cocaine_Positive  \
0  At least some college  3 or more  ...                        0.0
1   Less than HS diploma          1  ...                        0.0
2  At least some college  3 or more  ...                        0.0
3   Less than HS diploma          1  ...                        0.0
4   Less than HS diploma  3 or more  ...                        0.0

  DrugTests_Meth_Positive DrugTests_Other_Positive
Percent_Days_Employed  \
0                0.000000                      0.0
0.488562
1                0.000000                      0.0
0.425234
2                0.166667                      0.0
0.000000
3                0.000000                      0.0
1.000000
4                0.058824                      0.0
0.203562

  Jobs_Per_Year Employment_Exempt Recidivism_Within_3years  \
0      0.447610             False                    False
1      2.000000             False                     True
2      0.000000             False                     True
3      0.718996             False                    False
4      0.929389             False                     True

  Recidivism_Arrest_Year1  Recidivism_Arrest_Year2
Recidivism_Arrest_Year3
0                   False                    False
False
1                   False                    False
True
2                   False                     True
False
3                   False                    False
False
4                    True                    False
False

[5 rows x 53 columns]
```

```python
# Reidentification
pseudo_df["ID"] = psuedo_decry(pseudo_df["ID"], f)
pseudo_df.head()
```

```
Data decrypted

   ID Gender   Race Age_at_Release  Residence_PUMA Gang_Affiliated  \
0   1      M  BLACK          43-47              16           False
1   2      M  BLACK          33-37              16           False
2   3      M  BLACK    48 or older              24           False
3   4      M  WHITE          38-42              16           False
4   5      M  WHITE          33-37              16           False

   Supervision_Risk_Score_First Supervision_Level_First  \
0                           3.0                Standard
1                           6.0             Specialized
2                           7.0                    High
3                           7.0                    High
4                           4.0             Specialized

         Education_Level Dependents  ... DrugTests_Cocaine_Positive  \
0  At least some college  3 or more  ...                        0.0
1   Less than HS diploma          1  ...                        0.0
2  At least some college  3 or more  ...                        0.0
3   Less than HS diploma          1  ...                        0.0
4   Less than HS diploma  3 or more  ...                        0.0

  DrugTests_Meth_Positive DrugTests_Other_Positive
Percent_Days_Employed  \
0                0.000000                      0.0
0.488562
1                0.000000                      0.0
0.425234
2                0.166667                      0.0
0.000000
3                0.000000                      0.0
1.000000
4                0.058824                      0.0
0.203562

  Jobs_Per_Year Employment_Exempt Recidivism_Within_3years  \
0      0.447610             False                    False
1      2.000000             False                     True
2      0.000000             False                     True
3      0.718996             False                    False
4      0.929389             False                     True

  Recidivism_Arrest_Year1  Recidivism_Arrest_Year2
Recidivism_Arrest_Year3
0                   False                    False
```

```
False
1                    False                    False
True
2                    False                     True
False
3                    False                    False
False
4                     True                    False
False

[5 rows x 53 columns]
```

Generalization

Generalization is a commonly used technique in anonymization, which involves reducing the precision of a column. For example, the date of birth or the date of a doctor's visit can be generalized to a month and year, to a year, or to a five-year interval. Generalization can help achieving k-anonymity.

To check for k-anonymity, we will use the `pycanon` library. You can install this library in your virtual environment by running the command:

```
pip install pycanon
```

Question 11: `pycanon` includes several functions (feel free to explore them in the related documentation), but we will only be using `k-anonimity`. Look at the documentation, then use `k-anonimity` to determine the k-anonymity of the following groups of variables:

- k-anonymity of Gender and Race features: 743
- k-anonymity of Gender, Race, and Age_at_Release features: 44
- k-anonymity of Gender, Race, Age_at_Release and Residence_PUMA features: 1

```
# !pip install pycanon
from pycanon import anonymity

# Your answer here
k1 = anonymity.alpha_k_anonymity(train_df, quasi_ident = ["Gender",
"Race"], sens_att = ["Gender", "Race"])[1]
k2 = anonymity.alpha_k_anonymity(train_df, quasi_ident = ["Gender",
"Race", "Age_at_Release"], sens_att = ["Gender", "Race",
"Age_at_Release"])[1]
k3 = anonymity.alpha_k_anonymity(train_df, quasi_ident = ["Gender",
"Race", "Age_at_Release", "Residence_PUMA"], sens_att = ["Gender",
"Race", "Age_at_Release"])[1]
print("k-anonymity of Gender and Race features: " + str(k1))
print("k-anonymity of Gender, Race, and Age_at_Release features: " +
str(k2))
print("k-anonymity of Gender, Race, Age_at_Release and Residence_PUMA
features: " + str(k3))
```

```
k-anonymity of Gender and Race features: 743
k-anonymity of Gender, Race, and Age_at_Release features: 44
k-anonymity of Gender, Race, Age_at_Release and Residence_PUMA
features: 1
```

The k-anonymity of the combination of Gender, Race, Age_at_Release and Residence_PUMA is clearly problematic! It would be very easy to identify someone if we knew these 4 pieces of information about them.

Question 12: can you bin the Residence_PUMA feature to achieve 4-anonymity for this set of features? Add the new column to the existing dataframe, using the name `Binned_PUMA`.

For this task, you may want to look into the `cut()` and `qcut()` functions of the pandas library.

Remember that now, when checking for k-anonymity, you should be looking at the new column `Binned_PUMA`, not at `Residence_PUMA`.

```
# Your answer here
import warnings
warnings.filterwarnings("ignore", category=FutureWarning)

for bin in range(20):
    train_df["Binned_PUMA"] = pd.qcut(train_df["Residence_PUMA"], bin)
    k = anonymity.alpha_k_anonymity(train_df, quasi_ident = ["Gender",
"Race", "Age_at_Release", "Binned_PUMA"], sens_att = ["Gender",
"Race", "Age_at_Release"])[1]
    if k==4:
        print("Minimum optimal number of bins: " + str(bin))
        print("k-anonymity of Gender, Race, Age_at_Release and
Binned_PUMA features: " + str(k))
        display(train_df.head())
        break
```

```
Minimum optimal number of bins: 5
k-anonymity of Gender, Race, Age_at_Release and Binned_PUMA features:
4
```

	ID	Gender	Race	Age_at_Release	Residence_PUMA	Gang_Affiliated	\
0	1	M	BLACK	43-47	16	False	
1	2	M	BLACK	33-37	16	False	
2	3	M	BLACK	48 or older	24	False	
3	4	M	WHITE	38-42	16	False	
4	5	M	WHITE	33-37	16	False	

	Supervision_Risk_Score_First	Supervision_Level_First	\
0	3.0	Standard	
1	6.0	Specialized	
2	7.0	High	
3	7.0	High	
4	4.0	Specialized	

```
         Education_Level Dependents  ... DrugTests_Meth_Positive  \
0  At least some college  3 or more  ...                0.000000
1   Less than HS diploma          1  ...                0.000000
2  At least some college  3 or more  ...                0.166667
3   Less than HS diploma          1  ...                0.000000
4   Less than HS diploma  3 or more  ...                0.058824

  DrugTests_Other_Positive Percent_Days_Employed Jobs_Per_Year  \
0                      0.0              0.488562      0.447610
1                      0.0              0.425234      2.000000
2                      0.0              0.000000      0.000000
3                      0.0              1.000000      0.718996
4                      0.0              0.203562      0.929389

  Employment_Exempt Recidivism_Within_3years
Recidivism_Arrest_Year1  \
0             False                    False                    False

1             False                     True                    False

2             False                     True                    False

3             False                    False                    False

4             False                     True                     True


  Recidivism_Arrest_Year2  Recidivism_Arrest_Year3   Binned_PUMA
0                   False                    False  (15.0, 20.0]
1                   False                     True  (15.0, 20.0]
2                    True                    False  (20.0, 25.0]
3                   False                    False  (15.0, 20.0]
4                   False                    False  (15.0, 20.0]

[5 rows x 54 columns]
```

With 4-anonymity for these set of features, we can rest assured that there are at least 4 individuals sharing the same combination, making it more difficult to identify someone by knowing only these 4 pieces of information. However, let's not ignore the following issues:

- We did not test k-anonymity for other combinations of features, so it is very likely that our dataset is still not anonymized.
- 4-anonymity is not very strong; if I can narrow down my search to 4 people, I can still learn a lot about a person (at least approximatively).
- We may lose k-anonymity by adding more information.

Differential Privacy

As discussed in class, differential privacy is a stronger, mathematically robust definition of privacy for an algorithm. You can learn more about it by watching this video from Minute Physics: Protecting Privacy with MATH

After watching this video, try answering the following questions:

1. If you have two differentially private datasets, one with and one without your data, what does differential privacy guarantee regarding your privacy?
2. An algorithm has differential privacy ϵ = 2, another one ϵ = 4. Which one provides a higher level of privacy? Explain your answer.
3. The video highlights at least two of the main challenges with differential privacy. Summarize them.

- Q12.1: Differential privacy guarantees that the change in output of the algorithm between the two datasets will be minimal, and thus it would be difficult to deduce which dataset your data is present in. In other words, whether your data is in a dataset or not, the change in the output will be limited, so if the data is published, other people cannot easily detect your presence in the data.
- Q12.2: ϵ is a measure of the privacy loss as a result from differential changes in data, such as the addition or removal of new entries. As such, the algorithm with ϵ = 2 indicates that the change in the output will be small with or without the presence of your data, and thus has a higher level of privacy.
- Q12.3: There is a tradeoff between differential privacy and informational accuracy, so people will need to figure out the minimum amount of noise required to maximise both privacy and accuracy. The publication of multiple jittered statistics also runs the risk of being combined to reconstruct the data that was meant to be hidden, so their publication need to be future-proofed to prevent such.

Refined answer:

- Q12.3: There is a tradeoff between differential privacy and informational accuracy, so people will need to figure out the minimum amount of noise required to maximise both privacy and accuracy. The publication of multiple jittered statistics also runs the risk of being combined to reconstruct the data that was meant to be hidden, so their publication needs to be future-proofed to prevent such. There will be difficulties in convincing the public to consent to data collection, specifically by communicating the idea that the data will be protected effectively in a mathematically robust manner.

Randomized response

In class, we described randomized polling as a way to conduct interviews including sensitive questions, while protecting individuals' privacy.

Question 13: imagine that UBC has been surveying students to understand how many of them have been cheating in a final exam. Because the information is very sensitive and students will most likely not want to share this information, they use the randomized polling protocol described in class. If 1000 students have been surveyed, and 300 of them responded "yes", what is the actual percentage of students who cheated in a final?

Let x be the actual percentage of students who cheated in the final.

$$\frac{x*3}{4}+\frac{(1-x)*1}{4}=\frac{300}{1000}$$

$$\frac{1}{4}+\frac{x}{2}=\frac{3}{10}$$

$$x=\frac{1}{10}=10\%$$

Therefore, we conclude that $x = 10% $ is the actual percentage of students who cheated in the final.

Part 3: Data Governance

Data governance refers to the set of policies, procedures and standards that companies and organization must adopt to ensure quality, sacurity and usability of the data in their possession.

To gain a better understanding of what data governance is, why it is important and what common mistakes affect it, please read the following articles:

- https://www.egnyte.com/guides/governance/data-ownership
- https://atlan.com/data-governance-mistakes/#what-is-data-governance

As you can see, the issue of data governance is complex and multifaceted. A group of experts with a variety of experties is necessary to design and implement a robust data governance plan. Still, we can train ourselves to spot the most common mistakes when we see them. Take, for example, the following fictitional scenario (co-authored in collaboration with ChatGPT)

"SleekTech Solutions" is a cutting-edge technology company specializes in technologies related to artificial intelligence and data analytics. Their services include data analytics, big data processing, cloud computing, and Internet of Things (IoT). They offer their services to various industries, such as healthcare, finance, retail, manufacturing.

The company is young, only founded in 2021, and has rapidly expanded. At their inception, they used to accumulate data in a vast digital repository known as the "Data Lake." Initially, this seemed like a cost-effective solution to store all types of data, and they have not changed this strategy to this date.

To increase agility, SleekTech's different divisions have significant autonomy over their data. This means that the same data may be recorded by different department using different standards and metrics. SleekTech also encourages a culture of openness. Employees have access to vast amounts of data, including sensitive customer information, to complete the tasks they are assigned to.

SleekTech has been expanding rapidly. Founded in Canada, is now looking to expand into new markets including US and Europe.

Question 14: using the readings as reference, outline at least 4 distinct mistakes that SleekTech Solutions is likely to commit because of their data governance strategy.

- Unrestricted access privilages: In the question description, it says that "SleekTech also encourages a culture of openness. Employees have access to vast amounts of data, including sensitive customer information, to complete the tasks they are assigned to". All employees having little to no restriction on access privileges, including access to sensitive customer information, leads to a severe risk of security breaches and thus damage to the company's reputation.

- Inadequate communication: In the question description, it says that "The company is young, only founded in 2021, and has rapidly expanded". The rapid growth and significant autonomy would require significant amounts of communication between departments, otherwise "data governance initiatives may be misunderstood or improperly implemented" (atlan, 2023). The significant autonomy over the data within each department provides greater risks of potential misunderstandings and redundancies, improper implementation, and unnecessary confusion over the same data.

- Neglected data quality: SleekTech Solutions' data lake contains data on various industries with no indication that they are segregated for more efficient organization, which can lead to unnecessary difficulties in relevant operations involving seeking relevant data. Different departments are also likely to use differing standards and metrics on the same data, which if not communicated well can lead to misunderstandings that eventually cause degraded data quality and consistency, unnecessary maintenance costs and failures to comply with data regulations.

- Failure to evolve and adapt: The lack of adaptation to more effective data storage options since they were founded in 2021 may lead to their data management tools becoming redundant and irrelevant compared to their competition. By using the same strategy for more than three years without accommodating for changes in new technology, the tools for their system may not be compatible with said technology, driving away potential and existing customers, as well as more effort and costs required to update their system to remain competitive with similar companies.

Refined answer:

The four mistakes listed above can still be used for the refined answer. We add a couple of new distinct mistakes in addition.

- Lack of clear ownership and accountability: In this scenario, it is unclear who is responsible for data management in this organization. We can see that SleekTech's different divisions have significant autonomy over their data. However, without proper data ownership, accountability, and data management, they can end up with the inappropriate operation of the organization and poor data privacy and security. They should clearly define roles and responsibilities such as who the data owner is and who should be accountable for the mistakes in data operation.

Final thoughts

1) If you have completed this assignment in a group, please write a detailed description of how you divided the work and how you helped each other completing it:

- Jingyuan's response: We worked on the assignment separately, then collaborated to form our final assignment submission.
- Nicholas' response: We worked on the assignment separately, each taking turns answering all parts and modifying the responses down the line.

2) Have you used ChatGPT or a similar Large Language Model (LLM) to complete this homework? Please describe how you used the tool. **We will never deduct points for using LLMs for completing homework assignments,** but this helps us understand how you are using the tool and advise you in case we believe you are using it incorrectly.

- Jingyuan's response: I used ChatGPT to help debug the codes for pseudonymization and re-identification from `pycanon`.
- Nicholas' response: I have used Poe to assist in accessing the `pycanon` module, as well as the encoding in Q10 with both the pseudonymization function idea and using `cryptography`.

3) Have you struggled with some parts (or all) of this homework? Do you have pending questions you would like to ask? Write them down here!

- Jingyuan's response: Pending questions: what is the mathematical definition of ϵ-differential privacy? How do we interpret ϵ?
- Nicholas' response: Encoding ideas for Q10, computing ϵ for differential privacy.