NYPD Civilian Complaints

This project contains data on 12,000 civilian complaints filed against New York City police officers. Interesting questions to consider include:

- Does the length that the complaint is open depend on ethnicity/age/gender?
- Are white-officer vs non-white complaintant cases more likely to go against the complainant?
- Are allegations more severe for cases in which the officer and complaintant are not the same ethnicity?
- Are the complaints of women more succesful than men (for the same allegations?)

There are a lot of questions that can be asked from this data, so be creative! You are not limited to the sample questions above.

Getting the Data

The data and its corresponding data dictionary is downloadable here.

Note: you don't need to provide any information to obtain the data. Just agree to the terms of use and click "submit."

Cleaning and EDA

- Clean the data.
  - Certain fields have "missing" data that isn't labeled as missing. For example, there are fields with the value "Unknown." Do some exploration to find those values and convert them to null values.
  - You may also want to combine the date columns to create a `datetime` column for time-series exploration.
- Understand the data in ways relevant to your question using univariate and bivariate analysis of the data as well as aggregations.

Assessment of Missingness

- Assess the missingness per the requirements in `project03.ipynb`

Hypothesis Test / Permutation Test

Find a hypothesis test or permutation test to perform. You can use the questions at the top of the notebook for inspiration.

Summary of Findings

Introduction

This report is going to analyse a dataset from NYPD about civilians in New York City complaints against local police officers, which includes 31 kinds of data collected on 12000 people. Including ethnicity/age/gender of complaintants and officers, as well as the severity of violation, the type of violation, the time it happened and closed. This report is going to study on if the length that the complaint is open depend on the gender of complainants. This study also includes numbers of steps of data cleaning and exploring, as well as the assessment of missingness, to find the result of the thesis. The conclusion is that there is relationship between length that the complaint is open and the gender of complainants.

Cleaning and EDA

Cleaning:

1. First we notwice that there are many 'Unknown' in the data columns, so we really place these 'Unknown' with null values.

2. In the complainant age column, we find numbers of values with -4301.0, -1.0, and 0.0, we assume that they are invalid ages that can call police, hence we replace them as null value too.

3. We create a column to store case receiving time and case solving time into datetime form, and calculate the time period between them for further use in the following study

EDA:

Univariate Analysis:

1. We examine the complainant_gender column and using groupby and plot to see the distribution of the ages. From the graph we can see that people ages from 20-30 are the biggest part of complainants, which takes about #32.5% of the whole complainants population.

2. We want to explore if there is possible associations between the number of reported incidents and the time. The first analysis is between the number of reported incidents and the years (from 1985 to 2020). We used groupby and plot the result into a histogram and find that there is a positive relation between them, as the year increasing, the number of incidents compliant also increase.

3. The second analysis is between the number of reported incidents and the time within the year (from January to December). We used groupby and plot the result into a histogram and find that the number of reported incidents are generally higher in the time range from November to Feburary. Hence it shows that there is a relation between number of reported incidents and the month (perhaps the seasons)

Bivariate Analysis:

1. We also want to explore if there is any associations between the ethnicity of complainants and ethnicity of officers who deal with the case, we create a pivot table and take ethnicity of complainants as index, ethnicity of officer as columns, and incidents count as value. We find that all complainants tend to have highest possibility to get a white officer, the second is Hispanic. It might infer the proportion of police officers in NYPD, leaves for further study. And for all officers, they are most likely to receive a incident from black complainant, the second is Hispanic. It might also caused by the proportion population in that place, but since we do not have population data, it also leaves for further study. Therefore, the current data cannot provide evidences that there is correlation between ethnicity of complainants and ethnicity of officers.

2. We want to assess if the NYPD works efficiently, in other word, if they solve incidents quickly. We want to check the relationship between incidents' time_received and time_closed. First we make a scatter plot between year_received and year_closed, and we see that most of cases are solved within one year, but the later the time, the more the incidents that get solved after reported one year or two.

3. But does it mean that the incidents are solved more inefficiently today than in the past? We also plot the relationship between the time incidents happened and the length of time it takes to solve, and we find that it is true the later the incidents happened, the more time it takes to solve, which means there is a positive realtionship between the time incidents happened and the length of time it takes to solve.

Interesting Aggregates:

1. We look at the complainant_gender column and using groupby to see the distribution and frequency of the data. We find that male is the largest part of complainants, which takes about 82.4% of the whole complainants population.

2. We grouped genders and analyse if there are any findings on the aggregate data. We take the mean of complainant ages, officer ages, and time to solve the indicents. We find that on average, female complainants have age of 36.01, which is higher than male who is 31.94, and officers ages also slightly higher in female than male (33.5 > 32.2), and for the time to solve the indicents, incidents reported by male complainants takes 314.2 days to solve on average, and incidents reported by female complainants takes 298.83 days to solve on average, which shows that male complainants' incidents seems to take longer to solve than female complainants' incidents.

Assessment of Missingness

Possible NMAR: complainant_ethnicity

Reason: The missing of it might because the person is afraid to tell, or the person cannot tell at that condition.

Additional data: The serverity of the violation and the place of inaccidents happen, because that will infer how serevre the violation is, if it is enough to scare the person, and the place may infer the community of the person.

Case 1: If missing of ethnicity of complainants is MAR relate to age of complaints.

Result: p-value=0.00, It is MAR. It shows that the missing of ethnicity of complainant is relate to age of complainants. For instance, people with older ages might be more conservative to telling officers their personal information like ethnicity.

Case 2: If missing of ethnicity of complainants is MAR relate to unique id of officer.

Result: p-value=1.00, It is not MAR hence perhaps MCAR. It shows that the missing of ethnicity of complainants is not relate to id of officer.

Hypothesis Test

Hypothesis Test Question:

Is the length that the complaint is open depend on the gender of complainants?

Null hypothesis: Yes. The length of period of time the complaint is open are the same in male complainants and female complainants

Alternative hypothesis: No. The length of period of time the complaint is open are not the same in male complainants and female complainants. The speed of NYPD solving complaints complained by a gender is generally faster than by the other gender.

Test statistic : KS-statistic

Significant level : 0.05

Reason to choose these stats: KS-statistic measures the difference between two distributions. In this question, our 'length of period of time' is the (time closed) minus (time recived), hence it is a quantitative data, and we can put this data into a distribution where x-axis is the numerical value of the days, y-axis is the number of incidents solved, two distributions are incidents reported by male and incidents reported by female. Therefore, our object is to see that if 'time to solve incidents complained from male' and 'time to solve incidents complained from female' are two samples from the same distribution. KS is good choice for this goal. Significant level is to set at 0.05 because the most of test are using Significant level between 0.01, 0.05, and 0.1. So we choose the normal one

p-value: 0.00

Result: We have 0% confifence that our result will be like this under the null hypothesis, hence we reject the null hypothesis that the length of period of time taken to solve incidents are the same in male complainants and female complainants. The data shows that there are NOT likely to have no relationship between the time to solve incidents and complainants gender.

Further possible improvement:

We have numbers of missing data in the dataset, which might biased the result a little, since the missing of the data may also depends on something, and it might infers some reason. The date of incidents happened and solved are not accurate enough, they only have year and month, not exact day, which also makes this study less convinced. The method of data cleaning should also imporved, we will try our best to predict the missing value of the crucial data to get a better result in the future study.

Code

In [3]:

```
import matplotlib.pyplot as plt
import numpy as np
import os
import pandas as pd
import seaborn as sns
from scipy.stats import ks_2samp
%matplotlib inline
%config InlineBackend.figure_format = 'retina'  # Higher resolution figures
```

Cleaning and EDA

In [55]:

```
org_table=pd.read_csv('a.csv')
org_table.head()
#some missing data are represent as Unknown, we clean them to the np.nan form
df=org_table.replace('Unknown', np.nan)
#
df['complainant_age_incident']=df['complainant_age_incident'].replace((-4301.0,-1.0, 0.0), np.nan)
#combine year and month to dates columns for further use
df['dates_received']=df['year_received'].astype(str)+'-'+df['month_received'].astype(str)
df['dates_received']=pd.to_datetime(df['dates_received'])
df['dates_closed']=df['year_closed'].astype(str)+'-'+df['month_closed'].astype(str)
df['dates_closed']=pd.to_datetime(df['dates_closed'])
df['incidents_period']= df['dates_closed']-df['dates_received']
df['incidents_period_indays']=df['incidents_period'].dt.days
df.head()
```

Out[55]:

	unique_mos_id	first_name	last_name	command_now	shield_no	complaint_id	month_received	year_received	month_closed	year_closed	...	fado_type	allegation	precinct	conta
0	10004	Jonathan	Ruiz	078 PCT	8409	42835	7	2019	5	2020	...	Abuse of Authority	Failure to provide RTKA card	78.0	Repor
1	10007	John	Sears	078 PCT	5952	24601	11	2011	8	2012	...	Discourtesy	Action	67.0	Movir
2	10007	John	Sears	078 PCT	5952	24601	11	2011	8	2012	...	Offensive Language	Race	67.0	Movir
3	10007	John	Sears	078 PCT	5952	26146	7	2012	9	2013	...	Abuse of Authority	Question	67.0	PD violat
4	10009	Noemi	Sierra	078 PCT	24058	40253	8	2018	2	2019	...	Force	Physical force	67.0	Rep

5 rows × 31 columns

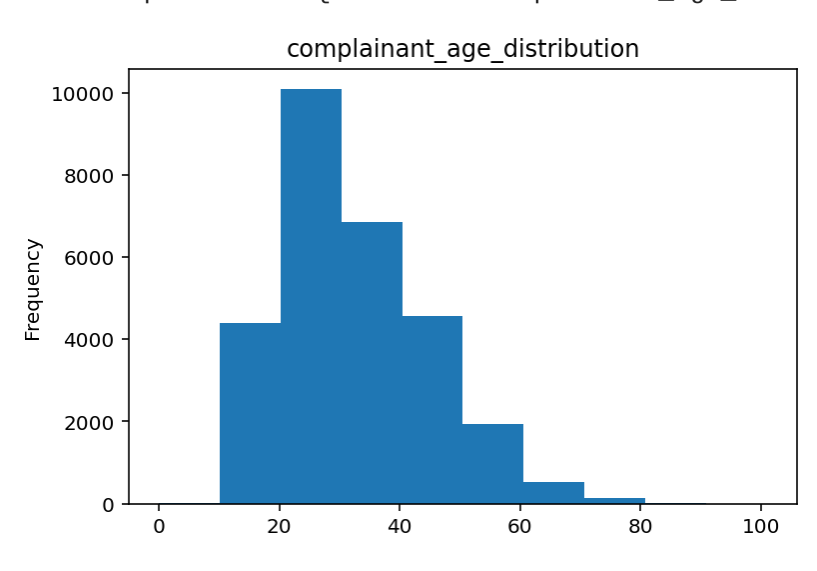
In [56]:

```
# Univariate Analysis
```

In [50]:

```
# Below is the histogram of the complainant age distributions
# From the graph we can see that people ages from 20-30 are the biggest part of complainants

complainant_age_distribution = df.groupby(['complainant_age_incident'], as_index=False).size()
df['complainant_age_incident'].plot(kind='hist', title='complainant_age_distribution')
```

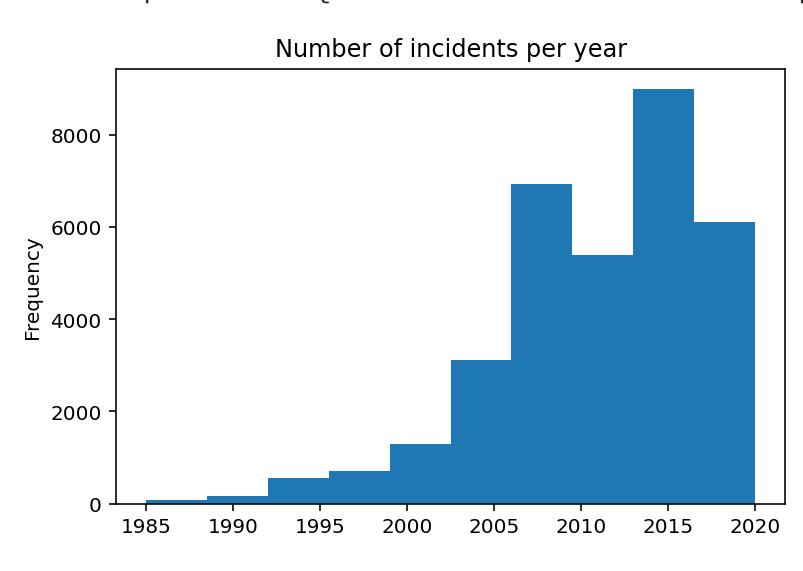
Out[50]: `<AxesSubplot:title={'center':'complainant_age_distribution'}, ylabel='Frequency'>`



In [58]:

```
# Below is the histogram of the number of incidents recieved in years, we are trying to find
# if there is relation between years and frequency of incidents recieved
# From the result we can see that it appears roughly the later the years, the higher the indicents frequency,
# and there are two unusual peaks in 2007 and 2015

df['year_received'].plot(kind='hist', title='Number of incidents per year')
```

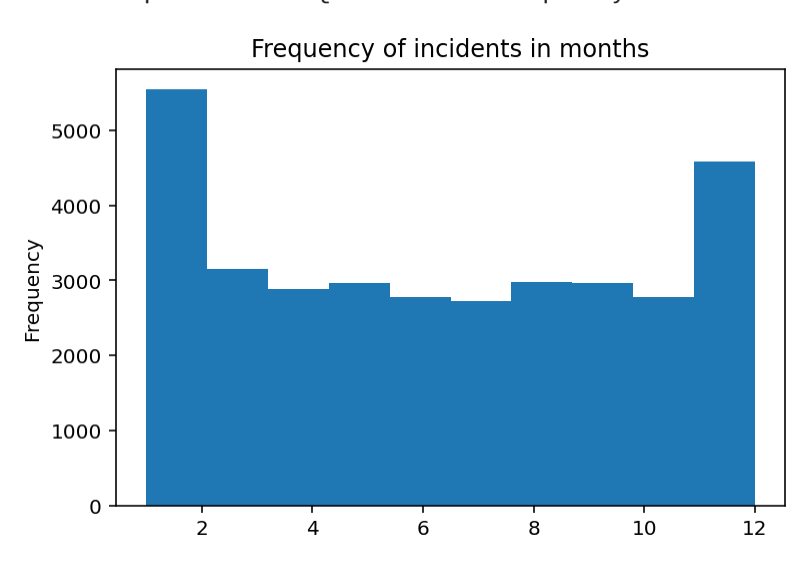
Out[58]: `<AxesSubplot:title={'center':'Number of incidents per year'}, ylabel='Frequency'>`



In [59]:

```
# Below is the histogram of the number of incidents recieved in months (include all years), we are trying to find
# if there is relation between months(seasons) and frequency of incidents recieved
# From the result we can see that it appears the time between November and February has higher indicents frequency,
# which might shows that there is higher frqeuncy of incidents happen in winter than in other seasons.

df['month_received'].plot(kind='hist', title='Frequency of incidents in months')
```

Out[59]: `<AxesSubplot:title={'center':'Frequency of incidents in months'}, ylabel='Frequency'>`



In [63]:

```
# Below is the result of summarizing contact_reason in the table, we can see that
# PD suspected C/V of violation/crime - street takes the most amount of reasons
# The second largerst is 'Other', which means that there are many cases don't have a clear classification

contact_reason_distribution = df.groupby(['contact_reason']).size()
contact_reason_distribution.sort_values(ascending=False).head()
```

Out[63]:

```
contact_reason
PD suspected C/V of violation/crime - street    10078
Other                                            4104
PD suspected C/V of violation/crime - auto       2981
PD suspected C/V of violation/crime - bldg       2542
Moving violation                                 1983
dtype: int64
```

In [ ]:

```
#Bivariate Analysis
```

In [21]:

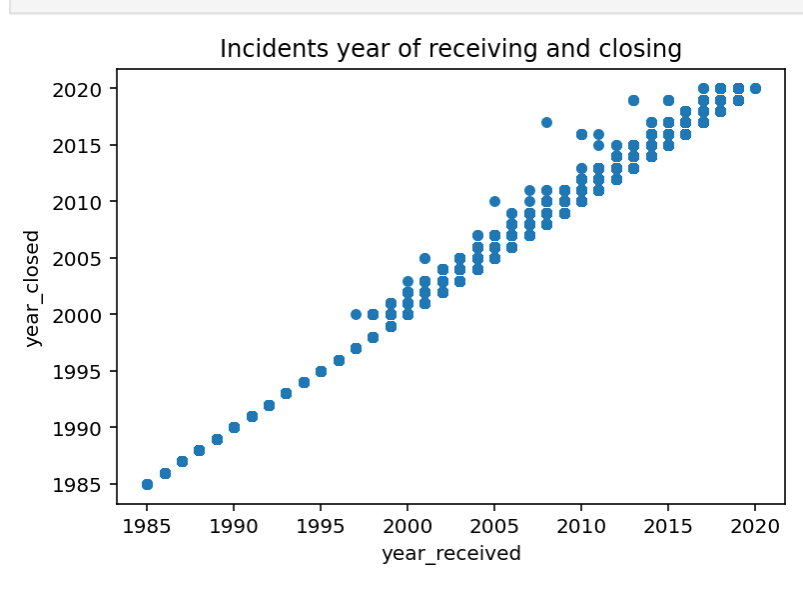
```
# Below is the pivot_table to examine the relation between complainant's ethnicity and officer's ethnicity

table = pd.pivot_table(df, values= ['unique_mos_id'], index=['complainant_ethnicity'],columns=['mos_ethnicity'], aggfunc='count', fill_value=0)
table
```

Out[21]:

	unique_mos_id				
mos_ethnicity	American Indian	Asian	Black	Hispanic	White
complainant_ethnicity					
American Indian	0	2	4	8	50
Asian	0	61	58	96	317
Black	18	558	2846	4722	8970
Hispanic	2	236	709	2215	3262
Other Race	6	19	98	187	367
Refused	0	18	38	58	145
White	2	129	380	625	1647

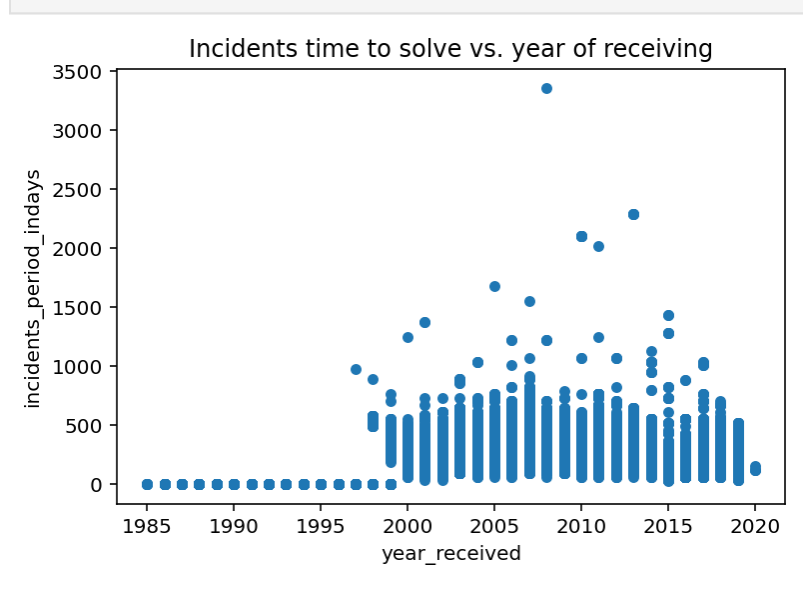
In [74]:

```
# Below is the scatterplot of incidents' year_received and year_closed

incidents_solving_inoneyear = df.plot.scatter(x='year_received', y='year_closed', title='Incidents year of receiving vs closing')
```



In [75]:

```
# Below is the scatterplot of incidents' year_received and incidents_period_indays

incidents_solving_time = df.plot.scatter(x='year_received', y='incidents_period_indays', title='Incidents time to solve vs. year of receiving')
```



In [ ]:

```
# Interesting Aggregates
```

In [26]:

```
# Below is the table of the number of incidents compliant by different genders
# From the table we can see that Males tend to have the most of complaints

(df.complainant_gender.value_counts())
```

Out[26]:

```
Male                      24058
Female                     5021
Not described                57
Transwoman (MTF)             20
Transman (FTM)                5
Gender non-conforming         2
Name: complainant_gender, dtype: int64
```

In [29]:

```
# Below is the series of the proportion of incidents compliant by different genders
# From the table we can see that Males takes about 82.4% of total compliants

complainant_gender_proportion = df.groupby(['complainant_gender']).size()
df.complainant_gender.value_counts()/(df.complainant_gender.value_counts()).sum()
```

Out[29]:

```
Male                      0.824949
Female                    0.172170
Not described             0.001955
Transwoman (MTF)          0.000686
Transman (FTM)            0.000171
Gender non-conforming     0.000069
Name: complainant_gender, dtype: float64
```

In [58]:

```
group_gender = df.groupby('complainant_gender').agg({'complainant_age_incident': ['mean'], 'mos_age_incident': ['mean'], 'incidents_period_indays':
group_gender
```

Out[58]:

	complainant_age_incident	mos_age_incident	incidents_period_indays
	mean	mean	mean
complainant_gender			
Female	36.019451	33.505477	298.834694
Gender non-conforming	35.000000	43.000000	304.000000
Male	31.943181	32.163314	314.202427
Not described	29.696970	36.666667	289.210526
Transman (FTM)	40.400000	28.400000	274.000000
Transwoman (MTF)	39.400000	38.750000	264.800000

In [ ]:

Assessment of Missingness

In [59]:

```
# If missing of ethinitity is relate to age
ks_list = []
#Copy original df and assign 'isnull' column
dfa=(df.assign(**{'is_null': df['complainant_ethnicity'].isnull()}))
grouped = dfa.groupby('is_null')['complainant_age_incident']
#KS for original
obs=ks_2samp(grouped.get_group(True), grouped.get_group(False)).statistic

for _ in range(100):
    # shuffle the complainant_ethnicity column
    shuffled_col = (dfa['complainant_ethnicity'].sample(replace=False, frac=1).reset_index(drop=True))
    shuffled = (dfa.assign(**{'complainant_ethnicity': shuffled_col}))
    shuffled = (shuffled.assign(**{'is_null': shuffled['complainant_ethnicity'].isnull()}))
    grps = shuffled.groupby('is_null')['complainant_age_incident']
    #KS for shuffled
    ks = ks_2samp(grps.get_group(True), grps.get_group(False)).statistic
    ks_list.append(ks)

pval = np.mean(np.array(ks_list) > obs)
pval
```

Out[59]: 0.0

In [61]:

```
# If missing of ethinitity is relate to age
ks_list = []
#Copy original df and assign 'isnull' column
dfa=(df.assign(**{'is_null': df['complainant_ethnicity'].isnull()}))
grouped = dfa.groupby('is_null')['unique_mos_id']
#KS for original
obs=ks_2samp(grouped.get_group(True), grouped.get_group(False)).statistic

for _ in range(100):
    # shuffle the complainant_ethnicity column
    shuffled_col = (dfa['complainant_ethnicity'].sample(replace=False, frac=1).reset_index(drop=True))
    shuffled = (dfa.assign(**{'complainant_ethnicity': shuffled_col}))
    shuffled = (shuffled.assign(**{'is_null': shuffled['complainant_ethnicity'].isnull()}))
    grps = shuffled.groupby('is_null')['unique_mos_id']
    #KS for shuffled
    ks = ks_2samp(grps.get_group(True), grps.get_group(False)).statistic
    ks_list.append(ks)

pval = np.mean(np.array(ks_list) > obs)
pval
```

Out[61]: 1.0

Hypothesis Test

In [47]:

```
# If complainant_gender is relate to incidents_period

#Copy original df
dfa=df
ks_list = []
#get period for two genders
gpA = dfa.loc[dfa['complainant_gender'] == 'Male', 'incidents_period']
gpB = dfa.loc[dfa['complainant_gender'] == 'Female', 'incidents_period']
#KS for original
obs = ks_2samp(gpA, gpB).statistic

for _ in range(100):
    # shuffle the complainant_gender column
    shuffled_col = (dfa['complainant_gender'].sample(replace=False, frac=1).reset_index(drop=True))
    shuffled = (dfa.assign(**{'complainant_gender': shuffled_col}))
    grps = shuffled.groupby('complainant_gender')['incidents_period']
    #KS for shuffled
    ks = ks_2samp(grps.get_group('Male'), grps.get_group('Female')).statistic
    ks_list.append(ks)

pval = np.mean(np.array(ks_list) > obs)
pval
```

Out[47]: 0.0

In [ ]: