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 (https://www.propublica.org/datastore/dataset/civilian-complaints-against-new-york-city-police-officers). The data dictionary is in the project03 folder.

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

The dataset implemented in this project is the data of civilian complaining against New York City police officers from *New York City's Civilian Complaint Review Board*. Crucial information such as the ethnicity of the police officer and the rank of him/her when the incident happened are recorded, which are useful to our investigation of whether there exists certain association between a police officer's ethnicity and his/her rank during the incident.

Cleaning and EDA

Assessment of Missingness

We started off by finding all the columns that contain missing values. Among these columns, we decided to select `complainant_ethnicity` as the column for missingness assessment, and this column's missingness is assessed through the likelihood of its dependence on `mos_ethnicity` and `rank_abbrev_incident`, the two key features of our investigation. Before implementing algorithms, we reached our assumption that missingness in `complainant_ethnicity` is not NMAR. Before cleaning, many of the missing data are displayed as "Unknown" or "Not described", which is impossible to relate to the police officer's rank or ethnicity, thus the missing does not depend on the value itself.

We chose a significance level of 0.05, as it is the most common level for most of the data analysis.

As categorical type data, we used total variance distance to assess the missingness between `complainant_ethnicity` and `rank_abbrev_incident`, and got a p-value of 0.647. This p-value suggests that missingness in complainant's ethnicity is not-at-all dependent on the rank of the officer. On the other hand, the missingness assessment between `complainant_ethnicity` and `mos_ethnicity` produced a p-value of 0, which suggests that the missingness in complainant's ethnicity is dependent on the officer's ethnicity, thus it is an MAR missingness via `mos_ethnicity`.

Hypothesis Test

Our hypothesis test information are as follows:

- Null Hypothesis: the ethnicity of the police officer is independent of the police officers' rank.
- Alternative Hypothesis: the ethnicity of the police officer is not independent of the police officers' rank.
- test statistics: Since the variables we are testing are both categorical variables, we used total variance distance for this hypothesis test.
- A significance level of 0.05 as the most common significance level is maintained during this part of investigation
- Conclusion: we reject our null hypothesis with a p-value of 0. The ethnicity of the police officer is not independent of the police officers' rank.

Code

In [244]:
```
import matplotlib.pyplot as plt
import numpy as np
import os
import pandas as pd
import seaborn as sns
%matplotlib inline
%config InlineBackend.figure_format = 'retina'  # Higher resolution figures
```

In [245]:
```
#read csv
df = pd.read_csv('nypd.csv')
cleaned = df.copy()
cleaned["time_received"] = (pd.to_datetime(cleaned['year_received'].astype(
cleaned["time_closed"] = (pd.to_datetime(cleaned['year_closed'].astype(str)
cleaned = cleaned.drop(['year_received', 'month_received', 'year_closed', '
cleaned
```

Out[245]:

cident	rank_abbrev_now	rank_now	...	complainant_gender	complainant_age_incident	fado_type	a
POM	POM	Police Officer	...	Female	38.0	Abuse of Authority	F
POM	POM	Police Officer	...	Male	26.0	Discourtesy	
POM	POM	Police Officer	...	Male	26.0	Offensive Language	
POM	POM	Police Officer	...	Male	45.0	Abuse of Authority	
POF	POF	Police Officer	...	NaN	16.0	Force	
...	...	...	...	...	...	...	
POM	SGT	Sergeant	...	Male	21.0	Discourtesy	
POM	SGT	Sergeant	...	Male	21.0	Abuse of Authority	In
POM	SGT	Sergeant	...	Male	21.0	Abuse of Authority	
POM	SGT	Sergeant	...	Male	21.0	Abuse of Authority	
POM	SGT	Sergeant	...	Male	21.0	Abuse of Authority	

Assessment of Missingness

To begin with, we started off by finding all the columns with missing values and the proportion of the missing values in the columns

In [246]:
```
null_columns = cleaned.columns[cleaned.isnull().any()]
null_columns = cleaned[null_columns].isnull().sum()/cleaned.shape[0]
null_columns
```

Out[246]:
```
command_at_incident          0.046286
complainant_ethnicity        0.133821
complainant_gender           0.125757
complainant_age_incident     0.144253
allegation                   0.000030
precinct                     0.000719
contact_reason               0.005966
outcome_description          0.001679
dtype: float64
```

In [ ]:

In [ ]:

In [ ]:

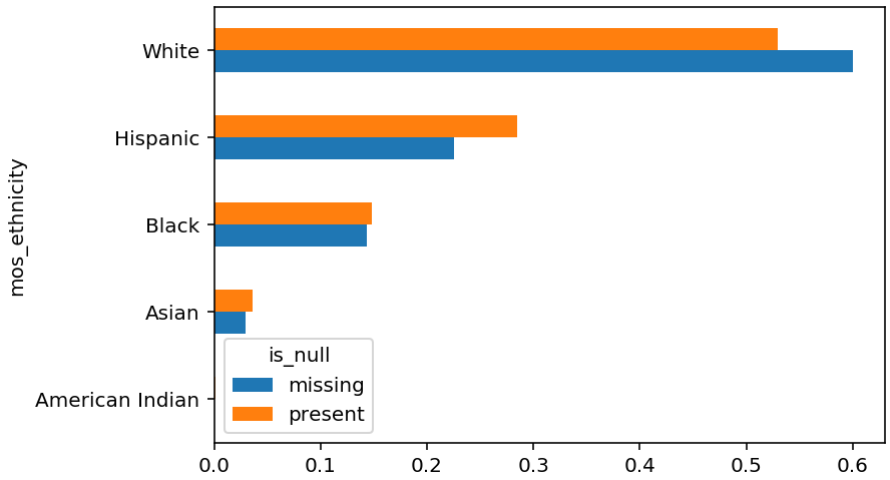
In [250]:
```
#missing keywords: complainant_gender: Not described, ethnicity: [Unknown,
cleaned['complainant_ethnicity'] = cleaned['complainant_ethnicity'].\
replace(['Unknown','Refused'], np.NaN)
cleaned['complainant_gender'] = cleaned['complainant_gender'].replace(['Not
```

In [ ]:

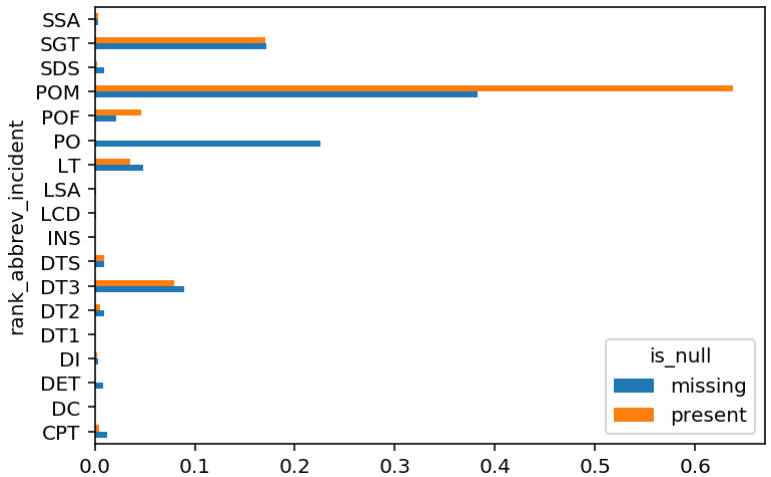
In [ ]:

In [ ]:

In [338]:
```
def nan_plot(data, col1, col2, plot = 'barh'):

    is_null = (
        data[col1]
        .isnull()
        .replace({True: 'missing', False: 'present'})
    )

    distrs = (
        data
        .assign(is_null=is_null)
        .pivot_table(index=col2, columns='is_null', aggfunc='size')
        .apply(lambda x:x/x.sum())
    )

    distrs.plot(kind=plot);
nan_plot(cleaned, 'complainant_ethnicity', 'mos_ethnicity')
```



In [339]:
```
nan_plot(cleaned, 'complainant_ethnicity', 'rank_abbrev_incident')
```



As shown above, bar plots between variables are drown. From the plots, it is obvious that `complainant_ethnicity` and `mos_ethnicity` have a very similar distribution, while `complainant_ethnicity` and `rank_abbrev_incident` have a different distribution.

In [ ]:

We created two extra columns for assessments using permutation tests

In [254]:
```
cleaned['age_null'] = cleaned['complainant_age_incident'].isnull()
cleaned['ethn_null'] = cleaned['complainant_ethnicity'].isnull()
```

The functions for different test statistics, including total variance distance, mean difference, ks-statistics, and sampling and p-value calculations are created.

In [333]:
```
def tvd(data, col, group_col):
    tvd = (
        data
        .pivot_table(
            index=col,
            columns=group_col,
            aggfunc='size',
            fill_value=0
        )
        .apply(lambda x: x / x.sum())
        .diff(axis=1).iloc[:, -1].abs().sum() / 2
        )

    return tvd
def diff_of_means(data, col, groupby):
    data_copy = data.copy()
    data_copy = data_copy.groupby(groupby)[col].mean()
    diff_mean = abs(data_copy.get(key=True) - data_copy.get(key=False))
    return diff_mean
def simulate_null(data, col, groupby, func):

    data_copy = data.copy()
    shuffled = (
        data_copy[col].sample(replace = False, frac=1).reset_index(drop=Tru
    )
    data_copy[col] = shuffled

    return func(data_copy, col, groupby)
def pval(data, col, groupby, func, rep=1000):
    diff = []
    for i in range(rep):
        result = simulate_null(data, col, groupby, func)
        diff.append(result)
    return np.count_nonzero(diff>np.float64(func(data, col, groupby)))/rep


def ks(data, col, groupby):
    from scipy.stats import ks_2samp
    v1 = data[groupby].unique()[0]
    v2 = data[groupby].unique()[1]
    ks_result = ks_2samp(data.loc[data[groupby]==v1, col],data.loc[data[grc
    return ks_result[0]
```

We got a p-value of 0.0 for the missingness in complainant's ethnicity and the officer's ethnicity, which is less than our significance level, so we determined that the missingness in complainant's ethnicity is dependent of the officer's ethnicity

In [334]:
```
pval(cleaned, 'ethn_null', 'mos_ethnicity',tvd)
```

Out[334]: 0.0

In [ ]:

We got a p-value of 0.647 for the missingness in complainant's ethnicity and the rank of the officer, which is greater than our significance level, so we determined that the missingness in the complainant's ethnicity is independent of the rank of the officer

In [335]:
```
pval(cleaned, 'ethn_null', 'rank_abbrev_incident',tvd)
```

Out[335]: 0.647

In [ ]:

In [ ]:

Cleaning and EDA

In [ ]:
```
# TODO
```

Assessment of Missingness

In [ ]:
```
# TODO
```

Hypothesis Test

In [336]:
```
pval(cleaned, 'rank_abbrev_incident', 'mos_ethnicity', tvd) #alternative hy
```

Out[336]: 0.0

The detailed process is written in the finding summary. In conclusion, we reject our null hypothesis with a p-value of 0. The ethnicity of the officer is dependent on the officer's rank

In [ ]:

In [ ]: