Beckah Chan - Python Project - UC Berkeley

California Police Scorecard

This dataset combines data from If Can't Wait and Campaign Zero on Californian police departments. It has been cleaned for your convenience: all missing values have been removed, and low-quality observations and variables have been filtered out. A brief descriptive summary of the dataset is provided below.

Summary

> The **California Police Scorecard** features data from 2016-2018 regarding police department performance and policies in all 58 California counties (with information about multiple police departments per county). We split the data into three datasets: **an arrests dataset, an accountability dataset, and a demographic dataset**.
>
> The **arrests data** (157 rows, 26 columns) quantifies police shootings, uses of force, arrests (with information about race) and homicides, budget, and more. The **accountability data** (157 rows, 17 columns) marks what policies are in place (1: yes, 0: no) for police departments to hold their police officers accountable. The **demographic data** (157 rows, 16 columns) provides race and economic information about the citizens in a police district to contextualize police behavior in their respective communities.
>
> The dataset has also recently expanded to include information about Sheriffs' Departments, the operator of county jails, which has the same metrics used to evaluate police departments.

For a quick glance of how each police department is performing, **Campaign Zero has rated the departments in three areas of policing: police violence, police accountability, and approach to policing**. They have averaged these scores to provide an **overall score** (the higher, the better) that summarizes the departments' performance in those three areas. Notably, **the majority of police departments have received an "F" grade**, indicating excess levels of lethal force, injured civilians, incidence of racial bias, over policing, complaints sustained, etc. You can read more about the methodology and formula used to compute the scores here: https://policescorecard.org/about.

Lastly, **we recommend exploring the Campaign Zero analysis - https://policescorecard.org/ - of the scorecard data to see what work has already been done and serve as inspiration for your own data project**. Campaign Zero's mission is to end police brutality in America by implementing research-backed policy solutions. From Campaign Zero: "The scorecard is designed to help communities, researchers, police leaders and policy-makers take informed action to reduce police use of force and improve accountability and public safety in their jurisdictions." Furthermore, this is an opportunity to apply the skills you've learned in Data 8 to social justice topics if you've felt moved to tackle systemic racism and police brutality in America.

Data Description

This dataset consists of three tables stored in the `data` folder:

1. `police-demographic`: provides basic demographic and economic information about the people living within each police department's jurisdiction
2. `police-arrests` provides information about the number of arrests and shootings that occured within ech police department
3. `police_accountability` provides information about the level of accountability for each police department as a result of county laws or police union agreements

A description of each table's variables is provided below:

`police-demographic`:

- `Region of California`: Region of California (Northern, Central, Southern) where the police district is located
- `Total Population`: Total number of people living in the police district
- `White Population`: Total number identified as White in the police district
- `Black Population`: Total number identified as Black in the police district
- `Hispanic Population`: Total number identified as Hispanic in the police district
- `Native American Population`: Total number identified as Native American in the police district
- `Asian Population`: Total number identified as Asian in the police district
- `Pacific Islander Population`: Total number identified as Pacific Islander in the police district
- `Other Population`: Total number not identified as any of the races above in the police district
- `Multiracial Population`: Total number identified as multiracial in the police district
- `Percent HS Graduates in Jurisdiction`: Percent of the population in the police district who graduated from high school
- `Unemployment Rate`: Unemployment rate in 2018
- `Median Income`: Median income in 2018
- `Poverty Rate`: What percent of the population living in the police district is under the US poverty line (earning below $12,490 per year) in 2018?

`police-arrests`:

- `Overall Score`: Average of Police Violence Score, Accountability Score, Approach to Policing. More information can be found in the links below.
- `Police Violence Score`: Average of Percentile Less Lethal Force Used per Arrest, Percentile Deadly Force Used per Arrest, Percentile Unarmed Civilians Killed or Seriously Injured, Percentile Racial Bias in Arrests and Deadly Force. More information can be found in the links below.
- `Police Accountability Score`: $\frac{2}{3}$ Percentile Civilian Complaints Sustained + $\frac{1}{6}$ Percent Discrimination and Excessive Force Complaints Sustained + $\frac{1}{6}$ Percent Criminal Complaints Sustained. More information can be found in the links below.
- `Approach to Policing Score`: Average of Percentile Misdemeanor Arrests per Population, Percent Homicides Cleared. More information can be found in the links below.
- `People Killed or Seriously Injured by Police, 2016-2018`: Total number of people killed or seriously injured by police between 2016 to 2018
- `Percent who Did Not Reportedly Have a Gun`: Of all incidents were people were killed or seriously injured by police between 2016 to 2018, what percent reportedly did not have a gun?
- `Percent Who were Confirmed Unarmed`: Of all incidents were people were killed or seriously injured by police between 2016 to 2018, what percent were confirmed unarmed?
- `People Deadly Force Used Against Who were Perceived to Have a Gun`: Of all incidents where deadly force was used, how many people were perceived to have a gun?
- `People Deadly Force Used Against Who were Confirmed with a Gun`: Of all incidents where deadly force was used, how many people were confirmed to have a gun?
- `2016 Police Shootings`: Number of police shootings in 2016
- `2017 Police Shootings`: Number of police shootings in 2017
- `2018 Police Shootings`: Number of police shootings in 2018
- `Total Arrests, 2016-2018`: Total number of arrests made between 2016 to 2018
- `Homicides (2013-2018)`: Total number of homicides from 2013 to 2018
- `percent_police_budget`: The percent of the county's total budget that goes towards the police
- `Log of police_budget`: The police's budget from 2018 in US$, after being log transformed. The actual police budget can be retrieved using the following formula: $\exp(\text{Log of police\_budget})$
- `Asian/Pacific Islander Drug Possession Arrests, 2016`: Total number of Asian/Pacific Islander drug possession arrests in 2016
- `Black Drug Possession Arrests, 2016`: Total number of Black drug possession arrests in 2016
- `Hispanic Drug Possession Arrests, 2016`: Total number of Hispanic drug possession arrests in 2016
- `Unknown Race Drug Possession Arrests, 2016`: Total number of drug possession arrests in 2016 by people of an unknown race
- `Other Race Drug Possession Arrests, 2016`: Total number of drug possession arrests in 2016 by people of other race
- `White Drug Possession Arrests, 2016`: Total number of White drug possession arrests in 2016

`police-accountability`:

- `disqualifies_complaints`: Does language in the police union contract disqualify misconduct complaints that are submitted too many days after an incident occurs or if an investigation takes too long to complete?
- `restricts_delays_interrogations`: Does language in the police union contract prevent officers from being interrogated immediately after being involved in an incident or otherwise restricts how, when or where they can be interrogated?
- `gives_officers_unfair_access_to_information`: Does language in the police union contract grant officers access to information civilians don't get prior to being interrogated?
- `limits_oversight_discipline`: Does language in the police union contract limit disciplinary consequences or otherwise hinder the capacity of civilian oversight structures or the media to hold police accountable?
- `requires_city_pay_for_misconduct`: Does language in the police union contract require cities to pay costs related to police misconduct? This includes giving officers paid leave while under investigation or paying legal fees and/or the cost of settlements.
- `erases_misconduct_records`: Does language in the police union contract prevent information on past misconduct investigations from being recorded or retained in an officer's personal file?
- `requires_deescalation`: Does language in the police's use of force policies require officers to de-escalate situations by communicating with subjects, maintaining distance or otherwise eliminating the need to use force?
- `bans_chokeholds_and_strangleholds`: Does language in the police's use of force policies ban the use of chokeholds and strangleholds against civilians?
- `duty_to_intervene`: Does language in the police's use of force policies require officers to intervene and stop excessive force used by other officers and reports these incidents to their supervisor?
- `requires_warning_before_shooting`: Does language in the police's use of force policies require officers to give a verbal warning, when possible, before shooting a civilian?
- `restricts_shooting_at_moving_vehicles`: Does language in the police's use of force policies restrict officers from shooting at moving vehicles?
- `requires_comprehensive_reporting`: Does language in the police's use of force policies require officers to report each time they use force or threaten to use force against civilians?
- `requires_exhaust_all_other_means_before_shooting`: Does language in the police's use of force policies require officers to exhaust all other reasonable means before resorting to deadly force?
- `has_use_of_force_continuum`: Does language in the police's use of force policies have a force continuum that limits the types of force and/or weapons that can be used to respond to specific types of resistance?

Additionally, here are some important definitions referenced in the variable descriptions above:

- Less Lethal Force: The total number of uses of tasers, batons, projectiles, pepper spray, other weapons and strangleholds against civilians.
- Civilian complaints: the total number of complaints, by type of complaint, reported by civilians against law enforcement personnel from 2016-2018.
- Deadly Force: All firearms discharges and all use of force incidents resulting in the death or serious injury of a civilian.

Last, the sources used for the data and descriptions. It is recommended that if you are confused about any of the descriptions, then you should take a look at the links below:

- http://useofforceproject.org/?sa=D&ust=1553928831132000#project
- https://policescorecard.org/about
- https://www.checkthepolice.org/?sa=D&ust=1553928831132000#project

Inspiration

A variety of exploratory analyses, hypothesis tests, and prediction problems can be tackled with this data. Here are a few ideas to get you started:

1. What is the trend over time for police shootings?
2. Are there geographic trends regarding violence and force used by police or sheriffs' departments?
3. Is there a statistically significant difference in the distribution of drug arrests for Blacks, Hispanics, Asians, Whites, etc?
4. What, if any, correlations exist between police policies/budget and civilians seriously injured or shot by officers in police departments?
5. What is the impact of graduation and poverty rates on the amount or types of crime in a city?
6. Can you predict the overall policing score from police budgets, arrests, or amounts of deadly force used?

The data sources websites may also provide some inspiration:

- If Can't Wait
- Campaign Zero

The analysis of police data has fueled media campaigns advocating for policies that increase police accountability and reduce police violence. We're excited to see what creativity you can bring to this growing discussion.

Don't forget to review the Final Project Guidelines for a complete list of requirements.

Preview

The tables are loaded in the code cells below. Take some time to explore them!

In [102]:

```
#Load the arrests data.
arrests = Table().read_table("data/police-arrests.csv")
print("The arrests dataset has {} rows and {} columns.".format(arrests.num_rows, arrests.num_columns))
arrests.show(5)
```

```
The arrests dataset has 157 rows and 26 columns.
```

Agency Name	Type of Agency	County	Region	Overall Score	Police Violence Score	Police Accountability Score	Approach to Policing Score	People Killed or Seriously Injured by Police, 2016-2018	Percent who Did Not Reportedly Have a Gun	Percen Who wer Confirme Unarme
Alameda County Sheriff's Department	Sheriff's Department	Alameda County	Northern	0.35	0.08	0.63	0.32	26	0.93	0.8
Alameda Police Department	Police Department	Alameda County	Northern	0.74	0.72	0.61	0.9	1	1	
Alhambra Police Department	Police Department	Los Angeles County	Southern	0.56	0.67	0.43	0.78	2	0.5	
Alpine County Sheriff's Department	Sheriff's Department	Alpine County	Northern	0.82	0.75	1	0.71	0	0	
Amador County Sheriff's Department	Sheriff's Department	Amador County	Northern	0.79	0.85	0.79	0.73	0	0	

... (152 rows omitted)

In [103]:

```
#Load the accountability data
accountability = Table().read_table("data/police-accountability.csv")
print("The accountability dataset has {} rows and {} columns.".format(accountability.num_rows, accountability.num_columns))
accountability.show(5)
```

```
The accountability dataset has 157 rows and 17 columns.
```

Agency Name	Type of Agency	County	disqualifies_complaints	restricts_delays_interrogations	gives_officers_unfair_access_to_
Alameda County Sheriff's Department	Sheriff's Department	Alameda County	1	1	
Alameda Police Department	Police Department	Alameda County	1	1	
Alhambra Police Department	Police Department	Los Angeles County	1	1	
Alpine County Sheriff's Department	Sheriff's Department	Alpine County	1	1	
Amador County Sheriff's Department	Sheriff's Department	Amador County	1	1	

... (152 rows omitted)

In [104]:

```
#Load the demographic data
demographic = Table().read_table("data/police-demographic.csv")
print("The demographic dataset has {} rows and {} columns.".format(demographic.num_rows, demographic.num_columns))
demographic.show(5)
```

```
The demographic dataset has 157 rows and 16 columns.
```

Agency Name	Type of Agency	County	Total Population of Jurisdiction	White Population	Black Population	Hispanic Population	Native American Population	Asian Population	Pacific Islander Population
Alameda Police Department	Police Department	Alameda County	79654	10066	33238	6078	29	24500	780
Alhambra Police Department	Police Department	Los Angeles County	86477	32234	9056	1450	136	42012	236
Anaheim Police Department	Police Department	Orange County	354891	189491	91910	6675	181	59712	987
Antioch Police Department	Police Department	Contra Costa County	112630	40000	31792	22257	391	10958	721
Bakersfield Police Department	Police Department	Kern County	379741	186871	124775	30373	1557	28112	717

... (152 rows omitted)

Research Report

Introduction

The data in this report were collected from Campaign Zero's Police Scorecard. The data are from 157 police departments in California and includes drug possession arrests in 2016 and demographic information for those arrests as well as for the population of the jurisdiction for each police department. The data also include other information, including the Police Scorecards that were created to rate police departments based on a variety of factors, such as Use of Force policies, police union contract, use of force, and other factors. The data are divided into three tables: `police-arrests`, `police-demographic` and `police_accountability`.

For the purpose of this report, we will be examining the relationship between the racial proportions of the population and the racial proportions of drug possession arrests. Therefore, we will be using only the `police-arrests` table and the `police-demographic` table.

The `police-arrests` table provides information on the total number of each race drug possession arrests in 2016 for each police department in California. On the other hand, the `police-demographic` table provides information on the population of each race of the jurisdiction for each police department, the unemployment rate in 2018, as well as the percentage of the population living in the police district that is under the US poverty line. These racial categories include White, Black, Hispanic, Asian/ Pacific Islander, other race, etc.

Note: Upon our analyses we discovered the demographic data were not reflective of populations in each county. A GSI told us during office hours the White, Hispanic, and Black populations were likely switched, so to remediate this we replaced the White population with Black, the Black population with Hispanic, and the Hispanic population with White.

Hypothesis Testing and Prediction Questions

Please bold your hypothesis testing and prediction questions.

Many current events have put a spotlight on police brutality and systematic racism. This led us to question if there is a significant difference in drug possession arrest race proportions and jurisdiction population race proportions for each police department.

Hence, **our hypothesis testing question is: Is there a significant difference between the proportions of race for jurisdiction population and for the drug possession arrests?** For this hypothesis testing question, our null hypothesis is that any differences between the racial proportions of the population and of drug possession arrests are due to chance. Our alternative hypothesis is that there is a significant difference between the racial proportions of the population and of drug possession arrests, and they are unlikely due to chance.

Many factors contribute to poverty. While some research has found a correlation between poverty and drug use, police may also have been targeting low-income neighborhoods for easier arrests that would net them more federal funding. Hence, **our prediction question is: Can we predict poverty rate if we know the unemployment rate?** To answer this prediction question, we apply a linear regression model and determine whether there is a linear association between the poverty rate and the unemployment rate.

Exploratory Data Analysis

You may change the order of the plots and tables.

Table Requiring a Join Operation:

First, we create two important tables by joining the data in `police-arrests` and `police-demographic`. The `pops_and_dpas` table contains the amount of people by race in the jurisdiction for each police agency and the amount of people arrested by race for each police agency, and the `with_dpa_pop_props` table contains the proportions for each race of the jurisdictions and of the drug possession arrests for 2016 for each police agency. Notice that we do not include unknown race in our tables, since it may impact the effectiveness of our data analysis.

In [105]:

```
#Selecting only the Agency Names and Population information from demographics table
dpops = demographic.select(0, 3, 4, 5, 6, 7, 8, 9, 10, 11)

# Combining Asian and Pacific Islander Populations for dpops to match arrest data
api_pop = dpops.column('Asian Population')+dpops.column('Pacific Islander Population')
dpops = dpops.drop('Asian Population', 'Pacific Islander Population').with_column('Asian/Pacific Islander Population', api_pop)

# Combining Multiracial, Native American, and Other populations into Other Column
other_pop = dpops.column('Native American Population')+dpops.column('Other Population')+dpops.column('Multiracial Population')
dpops = dpops.drop('Native American Population', 'Other Population', 'Multiracial Population').with_column('Other Population', other_pop)

dpas = arrests.select(0, 25, 21, 22, 20, 24)
# Adding a column to dpas of the total drug arrests in 2016
dpa_sum = 0
all_dpa_sums = make_array()
for i in np.arange(dpas.num_rows):
    for n in np.arange(1, 6):
        dpa_sum = dpa_sum + dpas.take(i).column(n).item(0)
    all_dpa_sums = np.append(all_dpa_sums, dpa_sum)
    dpa_sum = 0
all_dpa_sums
dpas = dpas.with_column('Total Drug Possession Arrests, 2016', all_dpa_sums).select(0,6,1,2,3,4,5)
dpas
```

Out[105]:

Agency Name	Total Drug Possession Arrests, 2016	White Drug Possession Arrests, 2016	Black Drug Possession Arrests, 2016	Hispanic Drug Possession Arrests, 2016	Asian/Pacific Islander Drug Possession Arrests, 2016	Other Race Drug Possession Arrests, 2016
Alameda County Sheriff's Department	2022	956	329	628	93	16
Alameda Police Department	107	49	17	25	9	7
Alhambra Police Department	279	34	12	214	11	8
Alpine County Sheriff's Department	5	4	0	0	0	1
Amador County Sheriff's Department	50	39	1	6	0	4
Anaheim Police Department	2179	652	128	1253	56	90
Antioch Police Department	397	169	153	65	3	7
Bakersfield Police Department	5062	2656	754	1560	44	48
Berkeley Police Department	554	194	255	76	12	17
Beverly Hills Police Department	123	52	41	25	2	3

... (147 rows omitted)

In [106]:

```
# joining dpops and dpas to a table with population demographic and arrest demographic data
pops_and_dpas = dpops.join('Agency Name', dpas, 'Agency Name')
pops_and_dpas
```

Out[106]:

Agency Name	Total Population of Jurisdiction	White Population	Black Population	Hispanic Population	Asian/Pacific Islander Population	Other Population	Total Drug Possession Arrests, 2016	White Drug Possession Arrests, 2016	Black D Possess Arre 2
Alameda County Sheriff's Department	179465	46333	63663	16766	42391	10312	2022	956	
Alameda Police Department	79654	10066	33238	6078	25280	4992	107	49	
Alhambra Police Department	86477	32234	9056	1450	42248	1489	279	34	
Alpine County Sheriff's Department	1203	117	777	18	8	283	5	4	
Amador County Sheriff's Department	24089	3565	18141	764	542	1077	50	39	
Anaheim Police Department	354891	189491	91910	6675	60699	6116	2179	652	
Antioch Police Department	112630	40000	31792	22257	11679	6902	397	169	
Bakersfield Police Department	379741	186871	124775	30373	28829	8893	5062	2656	
Berkeley Police Department	122188	14659	69677	6438	24731	6683	554	194	
Beverly Hills Police Department	35217	2649	27404	391	3327	1446	123	52	

... (147 rows omitted)

Notice that the White, Black, and Hispanic proportions for the jurisdiction population are way off from the reality. Hence, we replace the White Population with Hispanic Population, Black Population with White Population, and Hispanic Population with Black.

In [107]:

```
# Fixing labels for demographics for population
pops_and_dpas = pops_and_dpas.select(0, 1, 3, 4, 2, 5, 6, 7, 8, 9, 10, 11, 12)
pops_and_dpas = pops_and_dpas.relabel("Black Population", "x")
pops_and_dpas = pops_and_dpas.relabel("White Population", "y")
pops_and_dpas = pops_and_dpas.relabel("Hispanic Population", "Black Population")
pops_and_dpas = pops_and_dpas.relabel("x", "White Population")
pops_and_dpas = pops_and_dpas.relabel("y", "Hispanic Population")
pops_and_dpas
```

Out[107]:

Agency Name	Total Population of Jurisdiction	White Population	Black Population	Hispanic Population	Asian/Pacific Islander Population	Other Population	Total Drug Possession Arrests, 2016	White Drug Possession Arrests, 2016	Black D Possess Arre 2
Alameda County Sheriff's Department	179465	63663	16766	46333	42391	10312	2022	956	
Alameda Police Department	79654	33238	6078	10066	25280	4992	107	49	
Alhambra Police Department	86477	9056	1450	32234	42248	1489	279	34	
Alpine County Sheriff's Department	1203	777	18	117	8	283	5	4	
Amador County Sheriff's Department	24089	18141	764	3565	542	1077	50	39	
Anaheim Police Department	354891	91910	6675	189491	60699	6116	2179	652	
Antioch Police Department	112630	31792	22257	40000	11679	6902	397	169	
Bakersfield Police Department	379741	124775	30373	186871	28829	8893	5062	2656	
Berkeley Police Department	122188	69677	6438	14659	24731	6683	554	194	
Beverly Hills Police Department	35217	27404	391	2649	3327	1446	123	52	

... (147 rows omitted)

In [108]:

```
#Proportions by race for population and drug possession arrests
with_dpa_pop_props = Table().with_columns('Agency Name', pops_and_dpas.column(0),
                                          'White Population Proportion', pops_and_dpas.column(2)/pops_and_dpas.column(1),
                                          'Black Population Proportion', pops_and_dpas.column(3)/pops_and_dpas.column(1),
                                          'Hispanic Population Proportion', pops_and_dpas.column(4)/pops_and_dpas.column(1),
                                          'Asian/Pacific Islander Population Proportion', pops_and_dpas.column(5)/pops_and_dpas.column(1),
                                          'Other Population Proportion', pops_and_dpas.column(6)/pops_and_dpas.column(1),
                                          'Proportion White Drug Possession Arrests, 2016', pops_and_dpas.column(8)/pops_and_dpas.column(7),
                                          'Proportion Black Drug Possession Arrests, 2016', pops_and_dpas.column(9)/pops_and_dpas.column(7),
                                          'Proportion Hispanic Drug Possession Arrests, 2016', pops_and_dpas.column(10)/pops_and_dpas.column(7),
                                          'Proportion Asian/Pacific Islander Drug Possession Arrests, 2016', pops_and_dpas.column(11)/pops_and_dpas.column(7),
                                          'Proportion Other Drug Possession Arrests, 2016', pops_and_dpas.column(12)/pops_and_dpas.column(7),
                                          )
with_dpa_pop_props
```

Out[108]:

Agency Name	White Population Proportion	Black Population Proportion	Hispanic Population Proportion	Asian/Pacific Islander Population Proportion	Other Population Proportion	Proportion White Drug Possession Arrests, 2016	Proportion Black Drug Possession Arrests, 2016	Proportion Hispanic Drug Possession Arrests, 2016	Propo Asian/Pa Isla Posses Arr
Alameda County Sheriff's Department	0.354738	0.0934221	0.258175	0.236208	0.0574597	0.472799	0.16271	0.310584	0.045
Alameda Police Department	0.41728	0.076305	0.126372	0.317373	0.0626711	0.457944	0.158879	0.233645	0.084
Alhambra Police Department	0.104724	0.0167679	0.372755	0.488557	0.0172189	0.121864	0.0430108	0.767025	0.039
Alpine County Sheriff's Department	0.645885	0.0149626	0.0972569	0.00665004	0.235245	0.8	0	0	
Amador County Sheriff's Department	0.753082	0.0317157	0.147993	0.0224999	0.0447092	0.78	0.02	0.12	
Anaheim Police Department	0.258981	0.0188086	0.533941	0.171036	0.0172335	0.29922	0.0587425	0.575034	0.025
Antioch Police Department	0.282269	0.197612	0.355145	0.103694	0.0612805	0.425693	0.38539	0.163728	0.0075
Bakersfield Police Department	0.328579	0.0799835	0.492101	0.0759176	0.0234187	0.524694	0.148953	0.308179	0.0086
Berkeley Police Department	0.570236	0.0526893	0.119971	0.202401	0.0546944	0.350181	0.460289	0.137184	0.021
Beverly Hills Police Department	0.778147	0.0111026	0.0752196	0.0944714	0.0410597	0.422764	0.333333	0.203252	0.016

... (147 rows omitted)

Aggregated Data Table:

Next, we have calculated the proportion of each race by the population of all jurisdiction populations combined and of all drug possession arrests combined.

In [109]:

```
#Creating a proportion for entire state/all of the police agencies in the tables
total_pop = sum(pops_and_dpas.column(1))
pop_proportions = make_array()

for i in np.arange(2, 7):
    pop_proportions = np.append(pop_proportions, sum(pops_and_dpas.column(i)/total_pop))

total_arrests = sum(pops_and_dpas.column(7))
arrest_props = make_array()

for i in np.arange(8, 13):
    arrest_props = np.append(arrest_props, sum(pops_and_dpas.column(i))/total_arrests)
state_props = Table().with_columns('Race', make_array('White', 'Black', 'Hispanic', 'Asian/Pacific Islander', 'Other'),
                    'Population', pop_proportions,
                    'Drug Possession Arrests, 2016', arrest_props)
state_props
```

Out[109]:

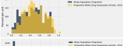
Race	Population	Drug Possession Arrests, 2016
White	0.376692	0.440982
Black	0.0599002	0.117081
Hispanic	0.382018	0.389206
Asian/Pacific Islander	0.147451	0.0250537
Other	0.0339391	0.0276776

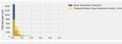
Quantitative Plot:

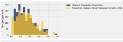
Next, we create some plots to provide a visual representation of our data distribution. In the plots below, we notice that the arrest versus population proportions for each race show that White, Black, and Hispanic people were arrested at larger proportions than of their population, and that Asian/Pacific Islander were arrested for drug possession significantly less than their population proportion and compared to any other race.

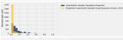
In [110]:

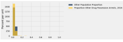
```
# Histograms for each race proportions by demographic vs by drug possession arrest
bins = np.arange(0, 1.1, 0.05)
with_dpa_pop_props.hist(1,6, bins=bins)
with_dpa_pop_props.hist(2, 7, bins=bins)
with_dpa_pop_props.hist(3, 8, bins=bins)
with_dpa_pop_props.hist(4, 9, bins=bins)
with_dpa_pop_props.hist(5, 10, bins=bins)
```







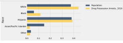




Qualitative Plot:

In [111]:

```
# Bar graph of state proportions comparison
state_props.barh('Race', overlay=True)
```



Hypothesis Testing

Do not copy code from demo notebooks or homeworks! You may split portions of your code into distinct cells. Also, be sure to set a random seed so that your results are reproducible.

In [112]:

```
# set the random seed so that results are reproducible
random.seed(1231)
# Calculating observed TVD
props_diffs = state_props.column(1)-state_props.column(2)
var = abs(props_diffs)

ca_prop_diffs = state_props.with_columns('Difference', props_diffs,
                                         'Variance', var)
tvd = sum(ca_prop_diffs.column('Variance'))/2
np.round(tvd, 3)
```

Out[112]: 0.13

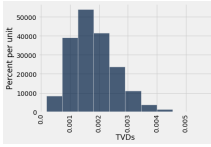
Our total variation distance for the drug possession arrests and population proportions by race is about 0.127.

In [113]:

```
# Generating random proportions based on population demographics
random.seed(1231)
pop_props = state_props.column('Population')
samp_size = sum(pops_and_dpas.column('Total Drug Possession Arrests, 2016'))
rand_samp = sample_proportions(samp_size, pop_props)
w_r_samp = state_props.with_column('Random Sample Proportion', rand_samp)
w_r_samp.barh(0)
```

In [114]:
```
# Setting Up some functions
random.seed(1231)
def tot_var_dist(arr1, arr2):
    '''Calculates total variation distance'''
    return sum(abs(arr1-arr2))/2
def rand_sampler():
    '''Generates random sample proportions'''
    return sample_proportions(samp_size, pop_props)
```

In [115]:
```
# Creating a histogram of TVDs from random sample proportions
random.seed(1231)
totvds = make_array()
n = 10000
for i in np.arange(n):
    tvd1 = tot_var_dist(pop_props, rand_sampler())
    totvds = np.append(totvds, tvd1)
tvds = Table().with_column('TVDs', totvds)
tvds.hist()
```



In [116]:
```
# Generating p-value of the population and arrest proportions' TVD and the random sampled TVDs
p = np.count_nonzero(tvds.column(0) >= tvd1)/n
p
```

Out[116]: 0.0

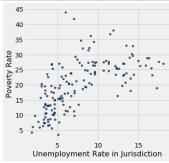
After random sampling and findling the TVD for 10,000 random samples, we find a p-value of 0.0. That means out of the 10,000 trials we ran, not once did we find a TVD of 0.127 or higher. Based on the p-value cut-off of 1% or 0.01, we reject the null.

Prediction

Be sure to set a random seed so that your results are reproducible.

In [117]:
```
def su(x):
    '''Converts to Standard Units'''
    x_avg = np.mean(x)
    x_std = np.std(x)
    x_su = (x-x_avg)/x_std
    return x_su

def correlation(x, y):
    '''Returns correlation'''
    r = np.mean(su(x)*su(y))
    return r

def parameters(x, y):
    '''Returns array of slope and intercept for regression line'''
    slope = correlation(x, y)*(np.std(y)/np.std(x))
    intercept = np.mean(y) - (slope*np.mean(x))
    params = make_array(slope, intercept)
    return params

def prediction(x_guess, x, y):
    '''Gives a prediction given an input x and slope and intercept in array params'''
    params = parameters(x, y)
    return (x_guess*params.item(0)) + params.item(1)

def residual(x_guess, x, y):
    '''produces residuals'''
    return y - prediction(x_guess, x, y)
```

In [118]:
```
pov_unempl = demographic.select('Unemployment Rate in Jurisdiction', 'Poverty Rate')
pov_unempl.scatter(0, 1)
```



In [119]:
```
correlation(pov_unempl.column(0), pov_unempl.column(1))
```

Out[119]: 0.6168406298754893

In [120]:
```
prediction(15, pov_unempl.column(0), pov_unempl.column(1))
```

Out[120]: 29.182955932045644

In [121]:
```
predictions_tbl = pov_unempl.with_column('Predictions', prediction(pov_unempl.column(0), pov_unempl
.column(0), pov_unempl.column(1)))
predictions_tbl.show(3)
```

Unemployment Rate in Jurisdiction	Poverty Rate	Predictions
4.4	10.2	15.9244
3.5	18.2	14.7987
4.5	17.4	16.0495

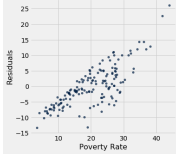
... (154 rows omitted)

In [122]:
```
residuals = predictions_tbl.with_column('Residuals', predictions_tbl.column(1) - predictions_tbl.co
lumn(2))
residuals.show(3)
```

Unemployment Rate in Jurisdiction	Poverty Rate	Predictions	Residuals
4.4	10.2	15.9244	-5.72445
3.5	18.2	14.7987	3.40128
4.5	17.4	16.0495	1.35047

... (154 rows omitted)

In [123]:
```
residuals.scatter(1, 3)
```



Conclusion

We rejected the null hypothesis: difference in drug arrest and population race proportions not likely due to chance. More, based on the predicted poverty rate given the unemployment rate, residuals showed that curved line would be a better predictor.