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

... (152 rows omitted)

In [4]:

```
#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	Other Population	Multi Popu
Alameda Police Department	Police Department	Alameda County	79654	10066	33238	6078	29	24500	780	562	
Alhambra Police Department	Police Department	Los Angeles County	86475	32234	9056	1450	136	42012	236	322	
Anaheim Police Department	Police Department	Orange County	354891	189491	91910	6675	181	59712	987	160	
Antioch Police Department	Police Department	Contra Costa County	112630	40000	31792	22257	391	10958	721	85	
Bakersfield Police Department	Police Department	Kern County	379741	186871	124775	30373	1557	28112	717	291	

... (152 rows omitted)

Research Report

Introduction

For our final project, we chose to explore and analyze the California Police Scorecard, which shows police department performance and policies in 58 California counties. This data is a combination of data from two organizations: 8 Can't Wait and Campaign Zero, two research-based campaigns with an aim to bring immediate change to policing in America. The California Police Scorecard dataset is split into three different tables: police-demographic, police-arrests, and police_accountability.

The police-demographic table consists of basic demographic and economic information regarding the constituents of each county

police department's jurisdiction. Since there are 58 counties in California, this dataset is a combination of all 58 police departments. This information helps us contextualize the behavior of a police department within different demographics. Thus, this table consists of important variables such as the distribution of races in the California population, as well as the median income. This distribution of races is the most important for our analysis purposes as it deals directly with our hypothesis test. We observed that in most large counties, the proportion of Whites is highest, with Blacks, Hispanics, Asians, Native Americans following (in that order). The police-arrests table gives us important information about different types of arrests, police shootings, use of force, etc. For our analysis purpose, we are focusing on the distribution of drug arrests within different races, so that we can determine if the distribution of race is similar to the distribution of drug arrests within the California population. From a preliminary analysis, we observed that counties with a higher poverty rate tend to have a higher amount of drug arrests, with the proportion of minorities being disproportionately higher. The police_accountability table describes the different levels of accountability for each police departments. For this table, we assigned each police department an overall score, which is made up from the individual criteria.

Hypothesis Testing and Prediction Questions

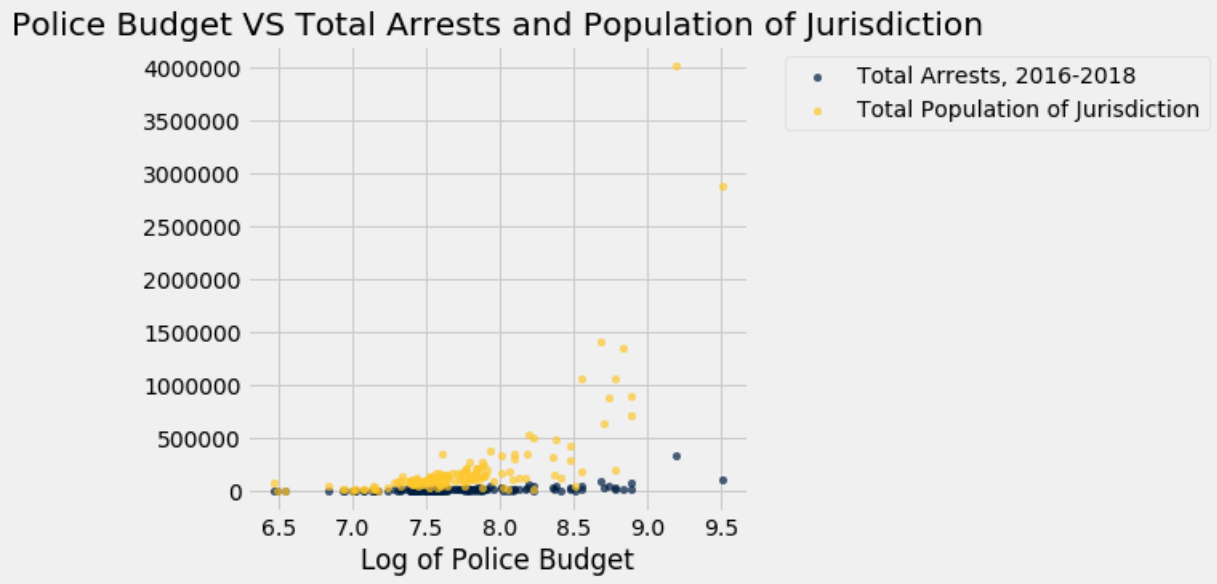
We plan on combining the information in the arrests and demographic tables to study the relationships (if there is one) between race and drug arrests. More specficially, we want to determine if the distribution of drug arrests by race is significantly different from the distribution of races of the California population, which would imply that one race is more likely to be arrested for drug-related crimes than another. **The corresponding null hypothesis is that the proportions of races being arrested for drug-related crimes in California is the same as the distribution of the Californian population. The alternate hypothesis is that the proportions of races being arrested for drug-related crimes in California is not the same to the distribution of the Californian population.**

Exploratory Data Analysis

Quantitative Plot:

In [5]:

```
budget_v_arrests = arrests.join('Agency Name', demographic, "Agency Name").select('Log of Police Budget', "Total Arrests, 2016-2018", "Total Population of Jurisdiction")
budget_v_arrests.scatter('Log of Police Budget')
plt.title("Police Budget VS Total Arrests and Population of Jurisdiction");
```

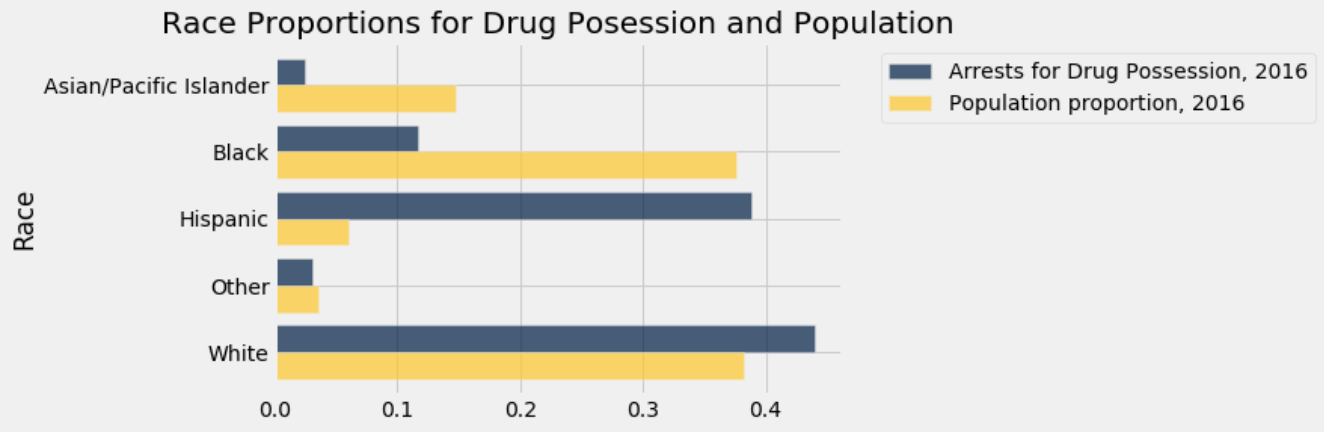


With this first graph, we wanted to explore the arrests vs the budget, as well as explore if Agencies who had higher funding arrested people more often, since they would have the money and means to. We also found it was important to compare the population as well, since we know for schools, population determines funding, so we wanted to explore that possibility as well. This graph tells us yes, there seems to be a correlation with how many individuals are being arrested, however, it is also evident that as our populations of jurisdictions increase in size, so does the police budget allowance. We determined from this graph that this information, though useful to know the assosiation, would not benefit our hypothesis in the long term, and turned towards looking specifically at the demographics themselves.

Qualitative Plot:

In [6]:

```
#data table for ultimately used for hypothesis testing
data_combo = arrests.join('Agency Name', demographic, 'Agency Name').select('Asian/Pacific Islander
Drug Possession Arrests, 2016', 'Black Drug Possession Arrests, 2016','Hispanic Drug Possession Ar
rests, 2016', 'Unknown Race Drug Possession Arrests, 2016','Other Race Drug Possession Arrests, 20
16', 'White Drug Possession Arrests, 2016', 'Total Population of Jurisdiction', 'White Population',
'Black Population','Hispanic Population','Native American Population','Asian Population','Pacific
Islander Population','Other Population','Multiracial Population')
cleaned_data = Table().with_column('Asian/Pacific Islander Drug Possession Arrests, 2016',
sum(data_combo.column('Asian/Pacific Islander Drug Possession Arrests, 2016'))).with_column('Black
Drug Possession Arrests, 2016', sum(data_combo.column('Black Drug Possession Arrests,
2016'))).with_column('Hispanic Drug Possession Arrests, 2016', sum(data_combo.column('Hispanic
Drug Possession Arrests, 2016'))).with_column('Unknown Race Drug Possession Arrests, 2016', sum(da
ta_combo.column('Unknown Race Drug Possession Arrests, 2016'))).with_column('Other Race Drug
Possession Arrests, 2016', sum(data_combo.column('Other Race Drug Possession Arrests,
2016'))).with_column('White Drug Possession Arrests, 2016', sum(data_combo.column('White Drug Posse
ssion Arrests, 2016'))).with_column('Total Population of Jurisdiction', sum(data_combo.column('Tota
l Population of Jurisdiction'))).with_column('White Population', sum(data_combo.column('White Popul
ation'))).with_column('Black Population', sum(data_combo.column('Black Population'))).with_column('
Hispanic Population', sum(data_combo.column('Hispanic Population'))).with_column('Pacific
Islander/Asian Population', sum(data_combo.column('Pacific Islander Population')+data_combo.column
('Asian Population'))).with_column('Other Population', sum(data_combo.column('Other Population')+da
ta_combo.column('Multiracial Population')+data_combo.column('Native American Population')))
props_prep = cleaned_data.with_column("Total Drug Possession Arrests",
cleaned_data.column('Asian/Pacific Islander Drug Possession Arrests, 2016')+cleaned_data.column('B
lack Drug Possession Arrests, 2016')+cleaned_data.column('Hispanic Drug Possession Arrests, 2016')
+cleaned_data.column('Unknown Race Drug Possession Arrests, 2016') + cleaned_data.column('Other
Race Drug Possession Arrests, 2016')+cleaned_data.column('White Drug Possession Arrests, 2016'))
columns_of_races = make_array('White','Asian/Pacific Islander', 'Black', 'Hispanic', 'Other')
#combined unknown and other and native american together as 'other'

#set up the Drug Arrests rates Table
column_of_arrests = make_array(props_prep.column('White Drug Possession Arrests, 2016').item(0),
props_prep.column('Asian/Pacific Islander Drug Possession Arrests, 2016').item(0),
props_prep.column('Black Drug Possession Arrests, 2016').item(0), props_prep.column('Hispanic Drug
Possession Arrests, 2016').item(0), props_prep.column('Other Race Drug Possession Arrests, 2016').
item(0)+props_prep.column('Unknown Race Drug Possession Arrests, 2016').item(0))/props_prep.column(
'Total Drug Possession Arrests').item(0)
arrests_props = Table().with_column('Race', columns_of_races).with_column('Arrests for Drug Possess
ion, 2016', column_of_arrests)
#set up the Population Rates table
column_of_population = make_array(props_prep.column('White Population').item(0), props_prep.column(
'Pacific Islander/Asian Population').item(0), props_prep.column('Black Population').item(0),
props_prep.column('Hispanic Population').item(0), props_prep.column('Other
Population').item(0))/props_prep.column('Total Population of Jurisdiction').item(0)
population_prop = Table().with_column('Race', columns_of_races).with_column('Population proportion,
2016', column_of_population)

#joined the two tables so we could look at them side to side on a bar graph
population_arrests_prop = arrests_props.join('Race', population_prop, 'Race')
population_arrests_prop.barh('Race')
plt.title("Race Proportions for Drug Posession and Population");
```



This bar chart, which we ultimately used for our hypothesis test, gave us key information that allowed us to look into the possibility that in fact our population distribution is in fact different than our arrests for drug possession among different races. Althought the Other and White populations are fairly close to their arrests for drug possession rates, for many POC, such as Asian/Pacific Islander, Black, and Hispanic, have drastically different proportions, with Hispanic being the largest difference, from roughly 6% of the states

population, to making up almost 39% of drug possession arrests in 2016. By seeing these side to side, we clearly note that there is some variation and we chose to explore it further to really see what these numbers mean.

Aggregated Data Table:

In [7]:

```
#this is the aggregated data we will be using in our prediction tests
total_drug_arrests = arrests.column('Asian/Pacific Islander Drug Possession Arrests,
2016')+arrests.column('Black Drug Possession Arrests, 2016')+arrests.column('Hispanic Drug
Possession Arrests, 2016')+arrests.column('Other Race Drug Possession Arrests,
2016')+arrests.column('Unknown Race Drug Possession Arrests, 2016')

lin_table = arrests.join("Agency Name", demographic, "Agency Name").with_column('Total POC Drug
Arrests, 2016', total_drug_arrests)
lin_table = lin_table.with_column("Proportion of Drug Arrests by Population",
lin_table.column("Total POC Drug Arrests, 2016")/lin_table.column("Total Population of
Jurisdiction"))
lin_table = lin_table.with_column("Proportion of Hispanics by Population",
lin_table.column("Hispanic Population")/lin_table.column("Total Population of Jurisdiction"))
lin_table = lin_table.select("Agency Name",'Total Population of Jurisdiction',"Total POC Drug
Arrests, 2016",  "White Drug Possession Arrests, 2016")
lin_table
```

Out[7]:

Agency Name	Total Population of Jurisdiction	Total POC Drug Arrests, 2016	White Drug Possession Arrests, 2016
Alameda County Sheriff's Department	179465	1069	956
Alameda Police Department	79654	60	49
Alhambra Police Department	86475	246	34
Alpine County Sheriff's Department	1203	3	4
Amador County Sheriff's Department	24089	12	39
Anaheim Police Department	354891	1528	652
Antioch Police Department	112630	228	169
Bakersfield Police Department	379741	2407	2656
Berkeley Police Department	122188	364	194
Beverly Hills Police Department	35217	71	52

... (147 rows omitted)

In this table, we put together all our data regarding every POC drug arrest (which added all that were not White) to a list with the total population as well, just to compare these numbers within each agency. As our data from a previous graph tells us, the White population is roughly 38% of our data, so it makes sense for a majority of these White drug arrests are smaller than POC (which is the rest of our data). We chose to ultimately use this as our model for prediction, as we felt it would be helpful to compare how similarly it is in correlated value and if its high enough, be able to roughly predict based off of real data how many arrest per White people that POC experience.

Table Requiring a Join Operation:

In [8]:

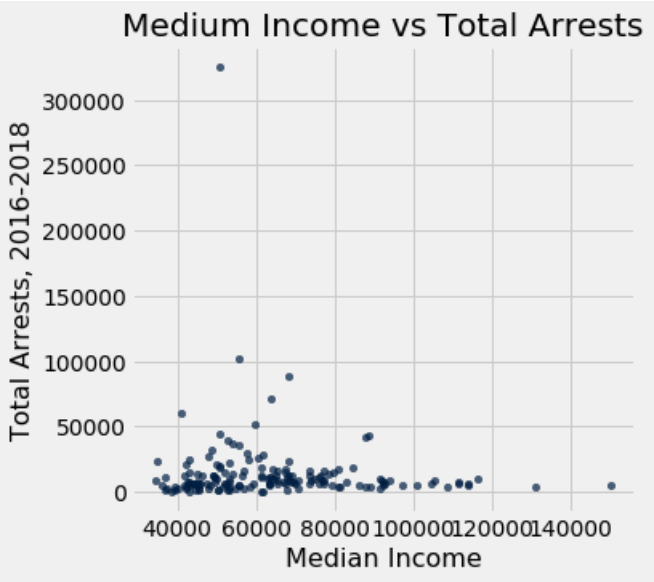
```
data_join = arrests.join('Agency Name', demographic, 'Agency Name').select('Agency Name', "Median I
ncome", "Total Arrests, 2016-2018")
data_join.scatter("Median Income", "Total Arrests, 2016-2018")
plt.title("Medium Income vs Total Arrests");
data_join
```

Out[8]:

Agency Name	Median Income	Total Arrests, 2016-2018
Alameda County Sheriff's Department	73775	15573
Alameda Police Department	81375	3581

Agency Name	Median Income	Total Arrests, 2016-2018
Alhambra Police Department	53214	2918
Alpine County Sheriff's Department	61343	37
Amador County Sheriff's Department	52964	1476
Anaheim Police Department	61877	28200
Antioch Police Department	63164	10115
Bakersfield Police Department	59925	51337
Berkeley Police Department	68969	7487
Beverly Hills Police Department	108736	4031

... (147 rows omitted)



Lastly, before starting our hypothesis test, we also looked towards income, and looked to see if that had any affect at all on the amount of arrests. Visually, we can see that the median income seems to play a very minor role, as there are many data points on our graph that have relatively low arrests with lower income medians, however, a greater concentration of them that seem to range a good amount are below 100k, but those are attributed to being outliers. They dont hold any meaning to our data as they are too extreme to be seen as part of the general trend. Another factor that may be causing this look is that some populations are higher, in term arresting more individuals.

Hypothesis Testing

First, we want to reiterate our null and alternate hypothesis, along with our test statistic and our significance level.

Null hypothesis: The proportions of races being arrested for drug-related crimes in California is **the same as the distribution of the Californian population.**

Alternate hypothesis: The proportion of races being arrested for drug-related crimes in California are **not the same as the distribution of the Californian population.**

Test Statistic: Here we are going to use Total Variation Difference because we have categorial data and it will give us an accurate visual on the variation among the population compare it to the arrest rates.

In [9]:

```
arrests_test_statistic = sum(abs(arrests_props.column('Arrests for Drug Possession, 2016') - popula
tion_prop.column('Population proportion, 2016')))/2
print("Our calculated test statistic for arrests is:", arrests_test_statistic)
```

```
Our calculated test statistic for arrests is: 0.38621420798486605
```

We will perform a regular hypothesis test and compare the data using TVD at a 5% significance level.

Creating a Simulation Test Stat and Histogram

This is the simulation our population will run through so that there is a clear distribution in the ways that our population proportion will display itself. Here we will run the test 5000 times to create a somewhat normal distribution as per the Central Limit Theorem describes (the more samples, the more normal the distibution looks. Included, when we take our sample proportion, we set it equal to the total amount of arrests to paint a more accurate view of a real life simulation.

In [10]:

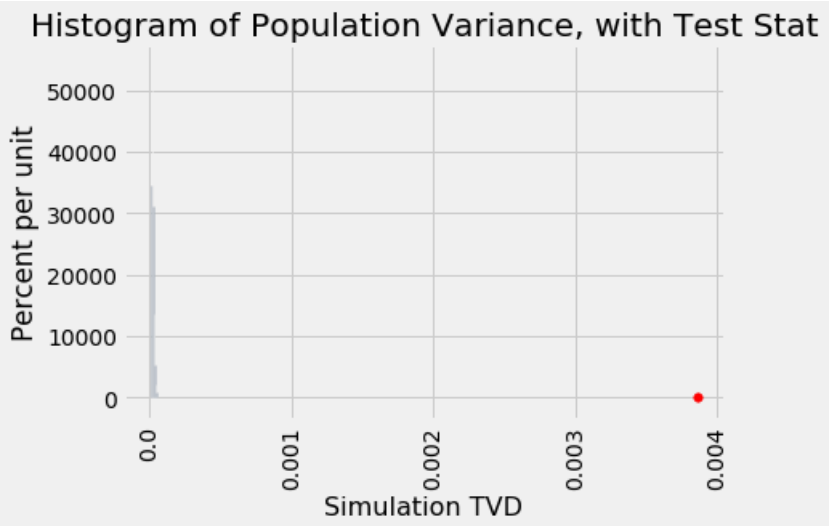
```
#one simulation statistic
def sim_test_stat(table0):
    simulated_prop = sample_proportions(149010, table0.column(1))
    return sum(abs(table0.column(1)-simulated_prop))/2
```

In [11]:

```
random.seed(1231)
#5000 simulations
simulations = make_array()
for test in np.arange(5000):
    new_stat = sim_test_stat(population_prop)
    simulations = np.append(simulations, new_stat)
```

In [12]:

```
import matplotlib.pyplot as plt
plt.style.use('fivethirtyeight')

Table().with_column('Simulation TVD', simulations).hist()
plt.scatter(arrests_test_statistic, 0, color='red', s=30)
plt.title("Histogram of Population Variance, with Test Stat");
```



Calculating our P-Value and comparing it to our significance level

By observation, we can definitely assume that we are going to fail to reject the null, however we need to be confident this is the case by **calculating the p-value**. To calculate our p-value based off our test statsitic, we are going to count the amount of true instances that a value in our simulations array is under our test statistic, then divide that by the length of the arrays, to get the percentage level. This number will give us a definitive value in which we can compare overall to our significance level and reject or fail to reject the null hypothesis.

In [13]:

```
p_value = np.count_nonzero(simulations > arrests_test_statistic)/len(simulations)
reject_null = True
print("Our calculated p-value is", p_value, 'and our signficance level is 5% (0.05). Do we reject
the null?', reject_null)
```

```
Our calculated p-value is 0.0 and our signficance level is 5% (0.05). Do we reject the null? True
```

After a comprehensive hypothesis test, we can claim with significant confidence that there is no possible way that our proportions of drug arrests per race is representative of our population proportions by race. We can attest to this to many possible reasons that may collude or arrests data. First, we take into account racial bias. There are times where people have biases against each other, and it can cause improper numbers to be counted. Another note, the total arrests in general might include people who are not part of the population as well, since the population is found through the census, which may not include every single individual who has visited, moved to CA, or are illegally in the state, but still have the same chance as everyone else to be arrested if they have drugs on them.

Although we don't know the true reason for the data being so unlike the proportion of the population, we can conclude that it does have to do with an outside or underlying reason, such as those mentioned before. We decided to explore in our prediction how many POC in general will be arrested for drug possession crimes in comparison to White, as seen on a scatterplot. We hope this will help us not only come to see what is going on, but to also weed out reasoning such as racial bias.

Prediction

In [14]:

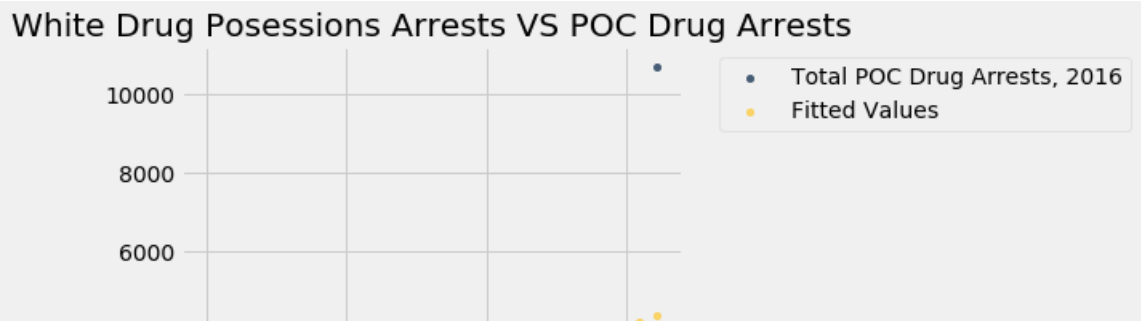
```
# set the random seed so that results are reproducible
random.seed(1231)

#standardize our variables
def standardize(arr):
    return (arr - np.mean(arr)) / np.std(arr)

#find the correlation between two
def correlation(tbl, var1, var2):
    return np.mean(standardize(tbl.column(var1)) * standardize(tbl.column(var2)))

#compute the slope
def slope(tbl, x, y):
    return correlation(tbl, x, y) * np.std(tbl.column(y)) / np.std(tbl.column(x))

#compute the intercept
def intercept(tbl, x, y):
    return np.mean(tbl.column(y)) - slope(tbl, x, y) * np.mean(tbl.column(x))

#find the line of best fit
def linear_fit(tbl, x, y):
    return slope(tbl, x, y) * tbl.column(x) + intercept(tbl, x, y)

correlation(lin_table, 'White Drug Possession Arrests, 2016', 'Total POC Drug Arrests, 2016')
```

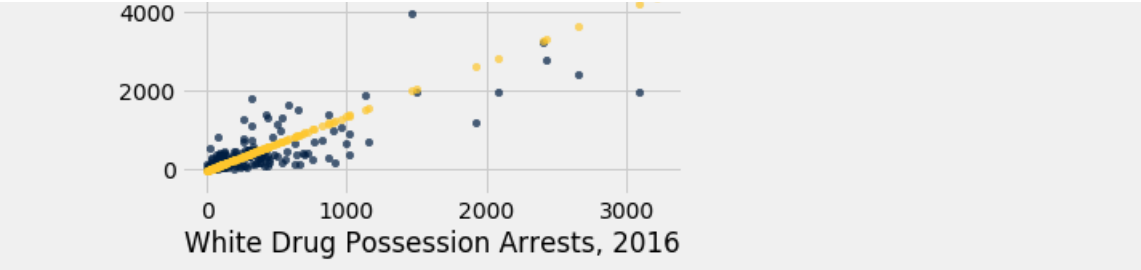
Out[14]:

```
0.7443017142159151
```

In [15]:

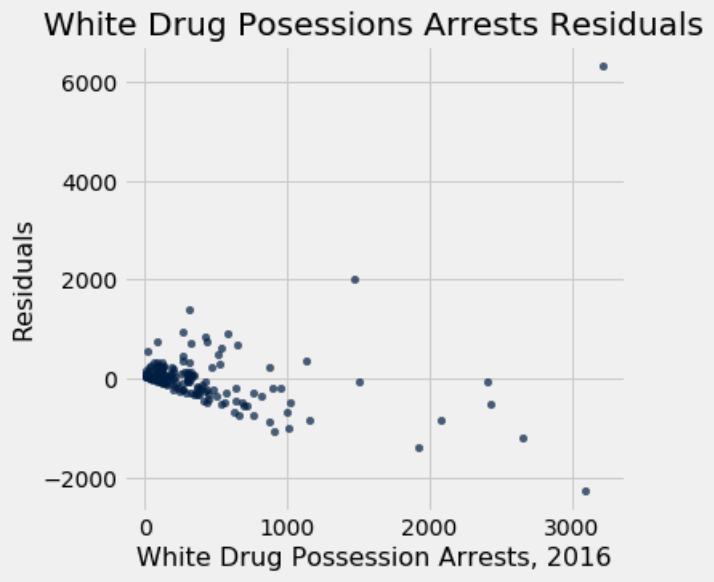
```
results = Table().with_columns(
    'White Drug Possession Arrests, 2016', lin_table.column('White Drug Possession Arrests,
2016'),
    'Total POC Drug Arrests, 2016',lin_table.column('Total POC Drug Arrests, 2016'),
    'Fitted Values', linear_fit(lin_table, 'White Drug Possession Arrests, 2016', 'Total POC Drug
Arrests, 2016')
)
results.scatter('White Drug Possession Arrests, 2016')
plt.title("White Drug Posessions Arrests VS POC Drug Arrests");
```





In [16]:

```
results = results.with_column(
    'Residuals', results.column(1) - results.column(2)
)
results.scatter('White Drug Possession Arrests, 2016', 'Residuals')
plt.title("White Drug Posessions Arrests Residuals");
```



The correlation is fairly strong in our test, which is to be expected, all people are being arrested at fairly similar rates. If we have 1000 White Drug Possession Arrests, we can assume that our POC Drug Possession Arrests are going to be roughly 1500, which is accurate to the population distribution categories we have created (White makes up about 40% of CA population while POC make up about 60%).

This data came as a suprise, however was expected since our hypthosis test failed. Our data, assuming we were perfectly on the line of best fit, claims that all the data points should be extremely close to that, however we have a fairly strong assosiation, but is lacking. Regardless of what races we chose, we should have gotten exremely close to 1, as if the rates were perfect with each other, they would rise and fall similarly.

We take note of several factors that may be causing this result. First, the proportion of individual races might be different among different Agency's jurisdictions, and may be causing a slight turn in drug arrests. Another, is racial bias and how racism can affect expected to be skewed in one direction or the other. Included, the poverty levels and median income may not be suifficent enough to survive, so selling drugs can be something one can go to make some money. More often than not, a county with a high median income would have a lesser total arrest rate overall, as seen from our 'Table Requiring a Join' section. Although these factors don't change the data, it does help shed light on how the data can show meaning.

Conclusion

We hypothesized that the proportion of drug arrests by race is different from the distribution of California population by race. We ran a hypothesis test using Total Variation Distance at a 5% significance level, as it would gives us variation among the population and compare it to the arrest rates. We determined that the proportion of drug arrests by race is different from the distribution of races in California, implying that there is a higher chance for someone to be arrested for drug-related matters based on their race. There were some confounding variables that may have affected our analysis, but overall, we saw a clear correlation between the White Drug Possession Arrests and POC Drug Possession Arrests. With all these factors considered, we can confidently say that there is a unfair bias towards arresting POC over Whites in California.

Presentation

In this section, you'll need to provide a link to your video presentation. If you've uploaded your presentation to YouTube, you can include the URL in the code below. We've provided an example to show you how to do this. Otherwise, provide the link in a markdown cell.

Link: *Replace this text with a link to your video presentation*

In [18]:

```
# Full Link: https://www.youtube.com/watch?v=BKgdDLrSC5s&feature=emb_logo
# Plug in string between "v=" and ""&feature":
YouTubeVideo('cTeJeQJwBvk')
```

Out[18]:

Submission

Just as with the other assignments in this course, please submit your research notebook to Okpy. We suggest that you submit often so that your progress is saved.

In [ ]:

```
# Run this line to submit your work
_ = ok.submit()
```

In [ ]:

```

```