Stat 427 Report

Racial Disparity in Champaign County's Criminal Justice System

Spring 2017

Stat 427:
Jack Yutong Li
Zeyu Zhang
Haoxian Zhong

Stat-Com:
Hongfei Li
Huanhuan Yue
Jingyi Zeng

Table of Contents

1. Introduction

Racial disparity is defined as existing in the criminal justice system when "the proportion of a racial/ethnic group within the control of the system is greater than the proportion of such groups in the general population" and has been an ongoing and hotly debated issue throughout the United States. The occurrence of racial disparity can be caused by a multitude or reasons, and they may or may not be related to racial discrimination.

The Champaign County Racial Justice Task Force has tasked us with the mission to find if there is any (or lack of) statistical evidence of racial disparity within the justice system of Champaign County using statistical tools. This report is thus a summary of the analysis that were carried out and results that were found from these analyses.

This analysis is conducted jointly by Jack Li, Zeyu Zhang, and Haoxian Zhong from the Stat 427 consulting class, and Hongfei Li, Huanhuan Yue, and Jingyi Zeng from the Stat-Com (Statistics and the Community) professional club.

We will begin this report with an overview of the objectives before diving into the bulk of the analysis. For the data analysis part, we will be working with three data sets, provided to us Dr. Wilson and Yuyang Huang from the Urban Planning department. The data sets that we will be considering includes the **Arrestees** data set dealing with information about arrestees, the **BookRJTF** data set dealing with the booking information of arrestees, and the **Circuit Clerk** data set which is focused on the court cases of people that were accused. Each data set will have a dedicated section to address the conducted analysis in details. After going through each data set, we will wrap up our findings in the **Conclusion** section. Due to various missing or inaccurate information within each data set, there were numerous test that we were like to carry out but were unable to do so. We will address these issues in the **Limitations** section.

No statistical knowledge or background is required to understand the analysis. We will be using laymen terms to describe the visualization and statistical test we conducted. If anyone is interested in the mathematical details and jargon of how each test works, we have provided links to this information at the **Online Sources for Statistical Test** section of this report.

2. Objectives

There are two main objectives of this analysis.

1) Conduct statistical analysis to see if there's statistical evidence of racial disparity within the justice system.

2) Assist Yuyan to further polish her web portal to help promote the understanding of these data to our community.
The current data portal that Yuyan created can be accessed through this link (http://141.142.170.106/VizTools/). She's planning to add a component called **"data story"** and this is where our work will go in.

For our analysis, we considered three different data set to look at different aspect of the criminal system and raised several interesting questions that will benefit our understanding of the issue of racial disparity.

- **Arrestees**: This dataset contains more than 100000 observations of the arrests conducted throughout Champaign county from 1/1/2010 to 9/15/2016.
- **BookRJTF**: Booking information of arrestees from 2010 to 2016.
- **Circuit Clerk**: Court case information throughout the year of 2016.

After going through these three data sets, we came up with 5 questions that will be meaningful to answer, shown below. These five question will be the backbone of this analysis and should be kept in mind when reading through the report.

1. **What is the proportion of arrestees under different race compared to census data? (Arrestees data set)**
2. **Given that someone is already arrested, is race a significant factor that would influence the outcome of whether an arrestee would be taken to jail? (Arrestees data set)**
3. **Is the waiting time, release time, and bond amount different for people who are booked due to race? (Booking data set)**
4. **Are African Americans charged with felonies more likely to be imprisoned? (Circuit Clerk data set)**
5. **Does race cause a significant difference on the length of the jail time given the same type of sentence? (Circuit Clerk data set)**

With these five questions in mind, let's dive into the analysis. We begin with the **Arrestees** dataset.

3. Arrestees Dataset

In this section, we will begin with the description and visualization of the Arrestees dataset.

3.1 Dataset Introduction

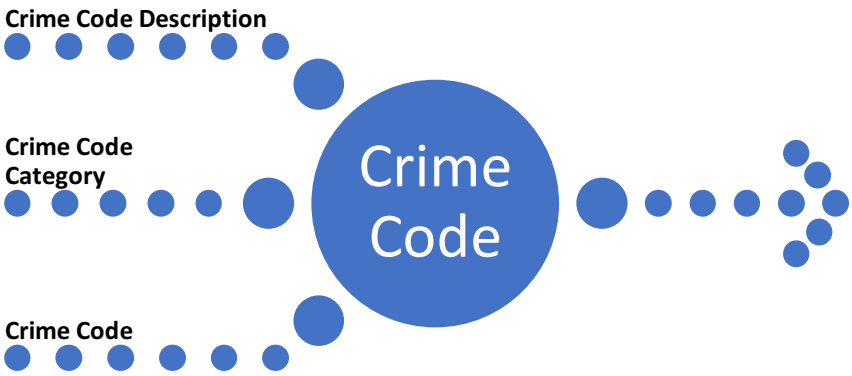
ID	DATE_OF_ARREST	TIME_OF_ARREST	ARREST_CODE	LOCATION_OF_ARREST	CRIME CODE	CRIME_CODE_DESCRIPTION	CRIME_CODE_CAT
1	01.01.2010	0:01	A10-00001	618 E DANIEL	1330	TRESPASS TO LAND/REAL PROPERTY	C17
2	01.01.2010	0:10	A10-00006	613 E GREEN	8367	DAMAGING PROPERTY	N
3	01.01.2010	1:26	A10-00007	618 E DANIEL	8120	MIP	N1
4	01.01.2010	2:06	A10-00003	502 S LIERMAN	3730	OBSTRUCTING JUSTICE	C32
5	01.01.2010	2:52	A10-00004	502 UNION	460	BATTERY	C05
5	01.01.2010	2:52	A10-00005	502 UNION	1365	TRESPASS-RESIDENCE	C17
6	01.01.2010	2:52	A10-00074	UNION/NEW	2461	OPERATE UNINSURED MOTOR VEHICLE	C29
6	01.01.2010	2:52	A10-00075	UNION/NEW	2470	NO DRIVERS LICENSE	C29
7	01.01.2010	3:28	A10-00048	1700 S NEIL	2480	SUSPEND REVOKED DRIVERS LICENSE	C29
8	01.01.2010	10:18	A10-00099	DOGWOOD AND W JOHN	6673	EXPIRED REGISTRATION	C29
8	01.01.2010	10:18	A10-00100	DOGWOOD AND W JOHN	2461	OPERATE UNINSURED MOTOR VEHICLE	C29
8	01.01.2010	10:18	A10-00101	DOGWOOD AND W JOHN	2480	SUSPEND REVOKED DRIVERS LICENSE	C29
9	01.01.2010	12:38	A10-00008	1512 N MATTIS	625	BURGLARY RESIDENTIAL	C09
10	01.01.2010	12:38	A10-00009	1512 N MATTIS	625	BURGLARY RESIDENTIAL	C09
11	01.01.2010	20:14	A10-00107	WINDSOR AND FIELDS SOUTH	6673	EXPIRED REGISTRATION	C29
12	01.01.2010	20:51	A10-00091	MATTIS/BRADLEY	2480	SUSPEND REVOKED DRIVERS LICENSE	C29
12	01.01.2010	20:51	A10-00092	MATTIS/BRADLEY	2461	OPERATE UNINSURED MOTOR VEHICLE	C29
13	01.01.2010	20:55	A10-00047	BRADLEY/REDWOOD	6673	EXPIRED REGISTRATION	C29
14	01.01.2010	21:13	A10-00093	MATTIS/BRADLEY	5081	WARRANT-IN STATE	C37
15	01.01.2010	22:51	A10-00049	1100 DORSEY	2470	NO DRIVERS LICENSE	C29
15	01.01.2010	22:51	A10-00050	1100 DORSEY	2461	OPERATE UNINSURED MOTOR VEHICLE	C29
16	01.01.2010	23:23	A10-00025	SPRINGFIELD/FOURTH	2461	OPERATE UNINSURED MOTOR VEHICLE	C29
16	01.01.2010	23:23	A10-00027	SPRINGFIELD/FOURTH	2480	SUSPEND REVOKED DRIVERS LICENSE	C29
16	01.01.2010	23:23	A10-00029	SPRINGFIELD/FOURTH	2410	DRIVING UNDER THE INFL-ALCOHOL	C28
16	01.01.2010	23:23	A10-00031	SPRINGFIELD/FOURTH	2430	ILLEGAL TRANSPORTATION OF LIQUOR	C29
16	01.01.2010	23:23	A10-00032	SPRINGFIELD/FOURTH	6621	FAILURE TO REDUCE SPEED	C29
17	01.02.2010	1:03	A10-00016	706 S FIFTH	8120	MIP	N1
18	01.02.2010	1:16	A10-00095	MCKINLEY/TREMONT	2470	NO DRIVERS LICENSE	C29
18	01.02.2010	1:16	A10-00096	MCKINLEY/TREMONT	2461	OPERATE UNINSURED MOTOR VEHICLE	C29
18	01.02.2010	1:16	A10-00097	MCKINLEY/TREMONT	5081	WARRANT-IN STATE	C37
19	01.02.2010	1:32	A10-00038	KIRBY/NEIL	2485	SEAT BELT-DRIVER & PASSENGER	C29
19	01.02.2010	1:32	A10-00039	KIRBY/NEIL	2470	NO DRIVERS LICENSE	C29
20	01.02.2010	8:53	A10-00189	204 FOXWELL	5081	WARRANT-IN STATE	C37
21	01.02.2010	10:07	A10-00260	700 W WINDSOR	6601	SPEEDING (RADAR)	C29

This dataset tells us about the complete arrests information from 2010 to 2015. For the year 2016, we only have part of the data. As it is shown in the figure above, every row indicates a single accusation of crime for a person. Therefore, if a person was charged of multiple crimes, there will be multiple rows for the same person. The columns give us the information related to this arrest including date of arrest, time of arrest, race of this person, sex and age of this person, etc.

The first data manipulation we did is to combine multiple rows into one, since we do not want to count the same person multiple times. Because we have so many variables, we need to identify which variables are important and which are not. Due to the limitation of both time and the data available, we will focus on the variables that we think are more important and relevant to the concerns of our clients.

3.2 Unimportant Variables

- ID: In the dataset, the ID helps us to identify each individual person. Each row of our cleaned data represents a single person; therefore, this variable is meaningless in our analysis and we can drop this variable.
- Variables related to crime code: We have three variables related to the type of crime the person was accused of.



However, having all three of them is redundant for our analysis. We choose one variable to best represents all three. The Crime Code variable categorize the crime using different code that has been documented by the state of Illinois. There are 35 of these crime codes, so we create 35 new variables to indicate which crime the person was accused of.

- Arrest Code/Weapon Code: We just drop these two variables because they are not relevant to our analysis.

3.3 Important Variables

- **Race**: Race is the variable that of our and our clients' priority concern. We need to focus on this variable.
- **Date/Time of arrests**: We think the date and time of arrests may also provide us some information so we want to include this in our analysis.
- **Age/Sex**: These two variables are important because we want to conduct analysis on the interaction between these demographic variables and race.
- **Result (Jail vs. Not Jail):** This is the result variable of our analysis. This variable tells us what happened after the person was arrested. This information is crucial in our analysis as it directly linked to our **second question: "Given that someone is already arrested, is race a significant factor that would influence the outcome of whether an arrestee would be taken to jail?"**.

3.4 Important Variables but Unable to Work With

- E**mployment**: We think employment status of a person may also be an underlying factor. Employment reflects the person's income and education level. We want to take this into

consideration, but the data we gotten has 13 different employment type but only 5 on them is documented, therefore we are unable to work with the data as of now.

- **Location of Arrests**: The location of the arrests is also an important part of our dataset. We can use it to identify whether there is discrimination within certain areas. However, we only have the street name of the arrest and nothing else. It will be great if we can map these streets into certain regions so that we can look at criminal activities on a broader geographic scope.

3.5 Comparison with Census Data

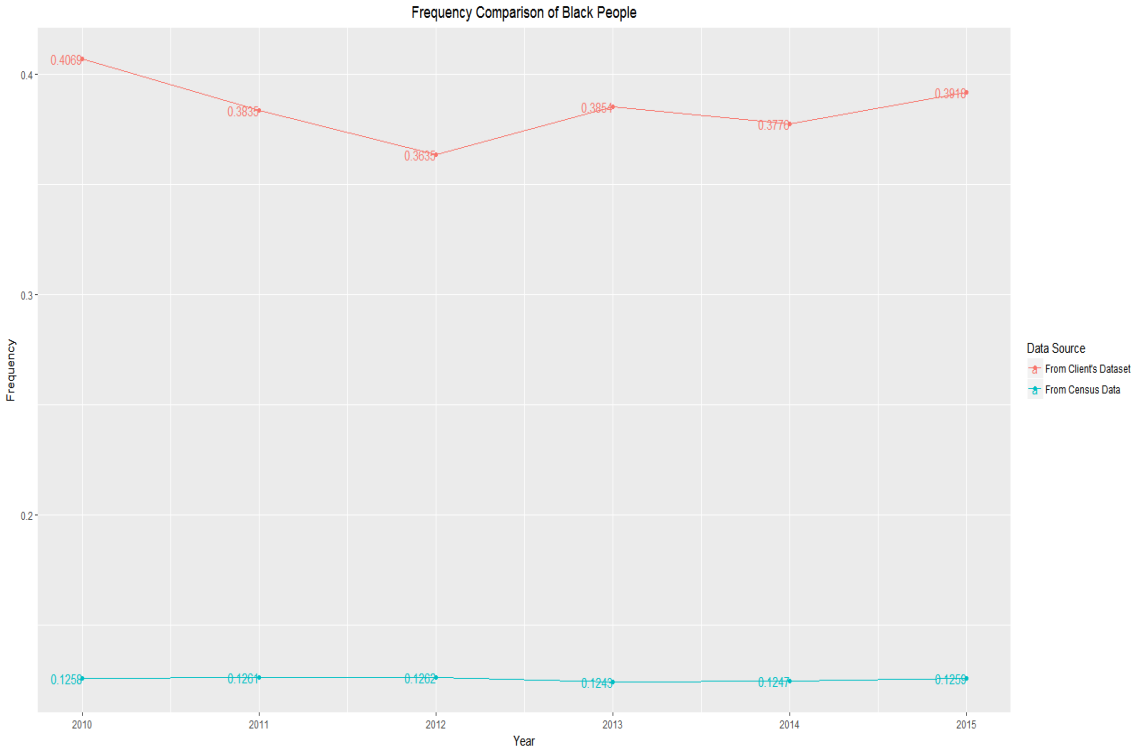
We obtain some online source about the census data in Champaign County to compare the proportion of different race within the census with that of the arrestee data. This information can be found at http://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml.This step is to address the **first question: "What is the proportion of arrestees under different race compared to census data?"**

.

Race Description of Dataset	Race Description of Census Data
• White	• White
• Black	• Black or African American
• Hispanic	• Asian
• Asian	• Two or More Races
• Asian/Pacific Island	• Native Hawaiian
• American	

The figure above shows the difference of race description in our dataset and the race description in our census data. We can see that the race categories are slightly different in the two different datasets. To avoid misclassification, we tend to focus on black people and white people since it is also our clients' main concern. In our later analysis, we will also focus our study on these two races.

50% Arrestees vs. 75% Champaign Population (White People)



38% Arrestees vs. 12% Champaign Population (Black People)

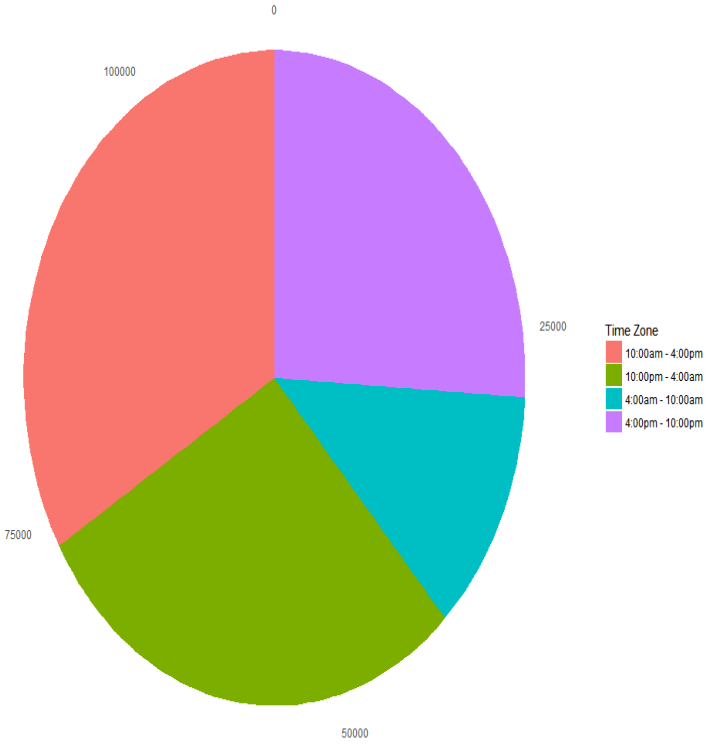
The two figures above show the trend and the comparison of the proportion of white people and black people within our data set and Champaign county. The red line is the trend for

the percentage of people in our **Arrestee** dataset. The blue line is the trend for the percentage of people in the census data. Since our data is not complete for 2016, we will only use data from 2010 to 2015 to do the comparison.

As we can see from the figures above, there are significant differences between white people and black people. The black people consists around 13% of the population but consists around 38% of the entire arrestees within a given year, while white people consists 75% of the population but only 50% of the arrestees.

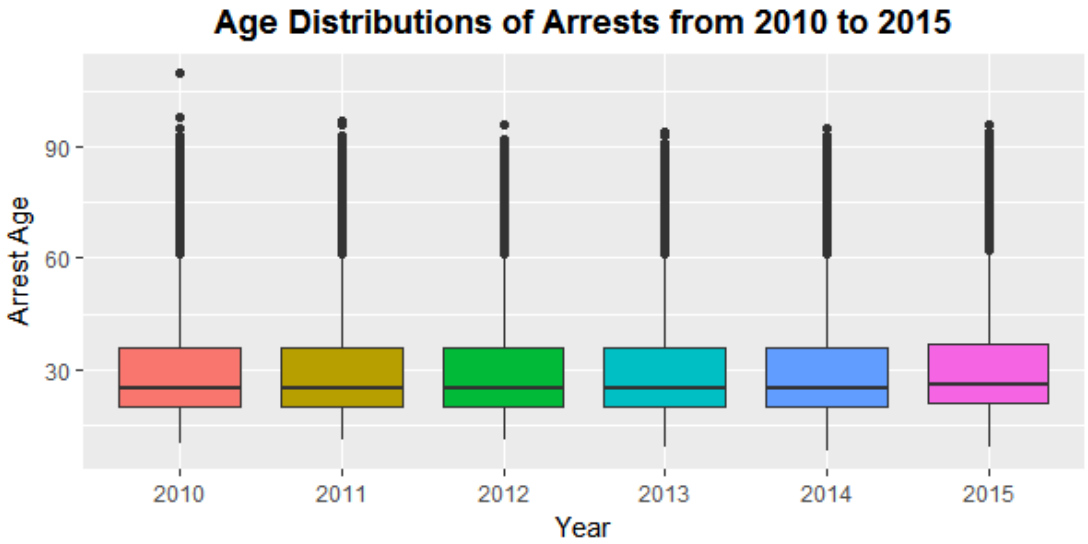
Through very simple visualization of the data we have, there are differences between the population of people in Champaign county to population of people who are accused of crime.

3.6 Time Distribution of Arrest Data



The pie chart above shows the distribution of crime with respect to the time of arrest. We divide the time of arrest into four different time intervals. 10:00 am to 4:00 pm, 4:00 pm to 10:00 pm, 10:00 pm to 4:00 am and 4:00 am to 10 :00 am. We can see from the pie chart that the red and purple are the two larger portion. It indicates that more crimes happen in the between 10:00 am to 10:00 pm. This fits our intuition as there are more human activities during this time.

3.7 Age Distribution of Arrest Data



Boxplot is a convenient way of graphically drawing groups of numerical data by their quartiles. The middle 50% of the data lies within the box. The lower edge of the box indicates the 25th percentile of the data and the upper edge box indicates the 75th percentile. the black line within the box is the median, or the 50th percentile. The black dots in each group are the outliers. These are the observations that are distant from all the other observations and does not fit into the overall distribution.

Age is another factor we are interested in. We want to find out the trend of arrest in different age group and the distribution of age group. The box plot above shows that during 2010 to 2015, there is not much change in the age distribution. Another information we can get is that most people who are accused of crime are from 20 years old to 40 years old. We group the age into 5 different subsets for our later analysis.

- Group 1: 0 - 18 years' old
- Group 2: 19 - 30 years’ old
- Group 3: 30 - 50 years' old
- Group 4: 50 - 75 years' old
- Group 5: 75+ years' old

3.8 Analysis on the Arrestee dataset

3.8.1 Focus only on the majority

After the data managements and initial visualization mentioned in the previous part, we are now going to focus on the analysis. In the **Arrestee** dataset, **90% of the observations are either African Americans or White Americans,** while the remain 10% belongs to Hispanic, Asian or Pacific Islanders. Therefore, we only focus on **African Americans and White Americans.**

3.8.2 Testing on the 35 crimes together

Goodness of fit test

We will be using a ***Sequential Goodness of Fit Test*** on a ***Logistic Regression Model*** to answer the above question. It basically tests the significance of a variable within a model. Shortly speaking, if adding this variable returns a low p-value, that mean this variable is **significant.** In other words, you can think of the variable as being an important predictor of whether the arrestee will go to jail. Otherwise, if the corresponding p-value of this variable is higher than 5%, then this variable is **not significant** and should not be included in the final model. Below are the results of this test:

Models

Demographic Models

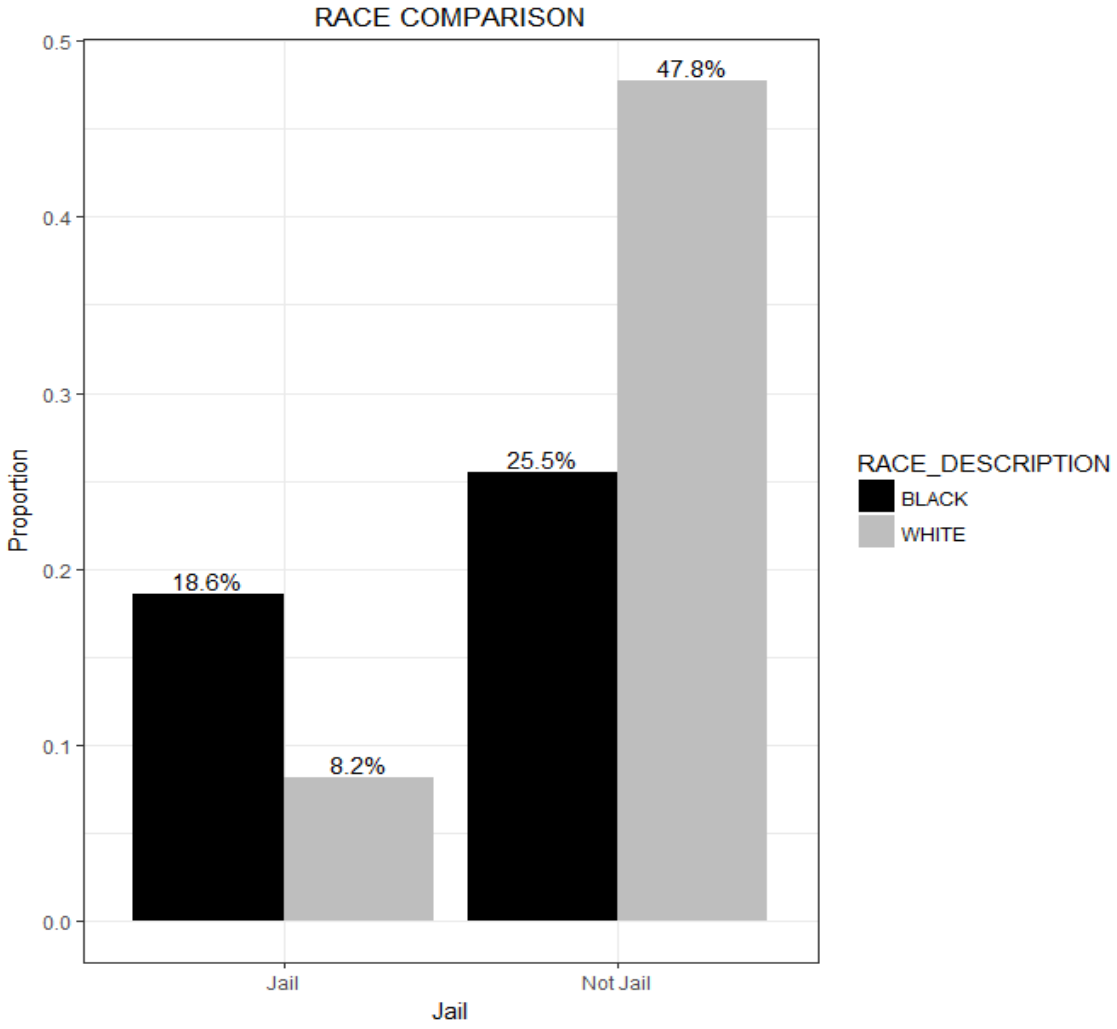
To find out if the three different demographic factors are significant variables, we will perform the goodness of fit test as mentioned above, with crimes being the initial model. The other demographic variables will be added and tested sequentially.

- Result ~ all 35 crimes (null / base model)
- Result ~ all 35 crimes + race
- Result ~ all 35 crimes + race + sex
- Result ~ all 35 crimes + race + sex + age group

Here, the null model indicates that whether a person got arrested will be put into jail is **only determined by the crime(s) he was accused of**. And for the second model listed above, it implies that besides the crimes a person was accused of, race is also another factor that contributing to whether this person will be put into jail. We can then **extend it to sex and age group and finally get 3 different models**.

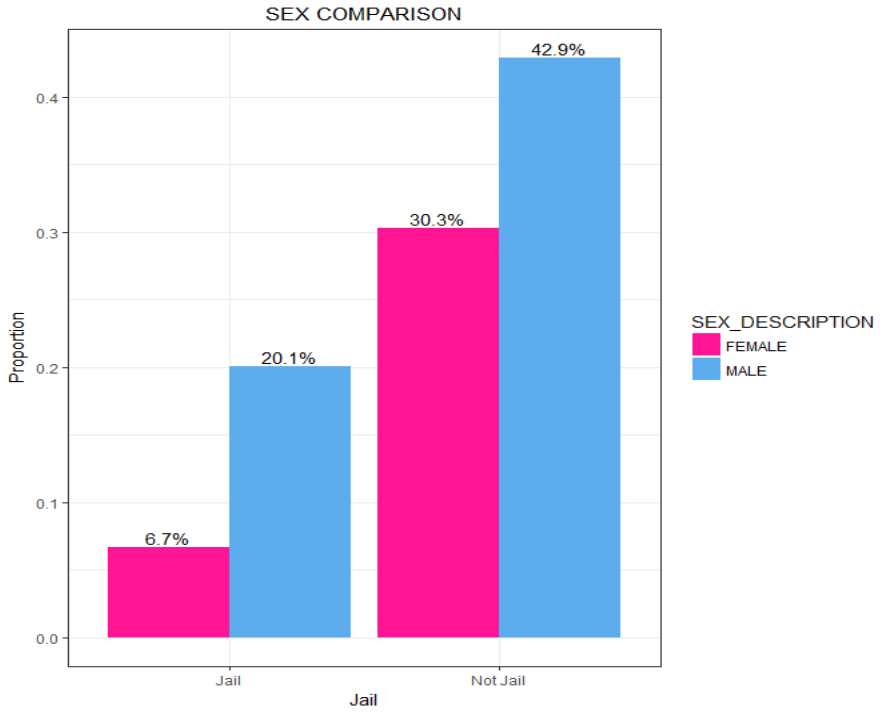
Results

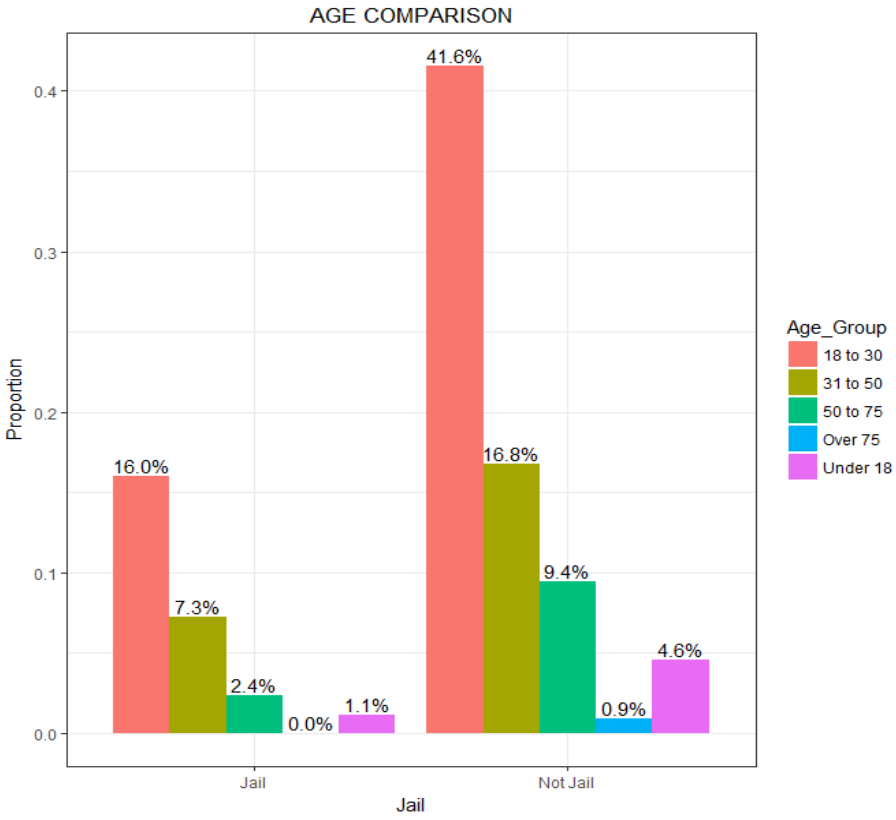
For the second model, the p-value is **less than 0.00001**, which is significantly small, indicating that **race** is a **significant predictor** in the model to determine whether a person would be put into jail. And for the second and the third full models listed above, both two p-values are **less than 0.00001**, therefore sex and age group are also significant predictor in the model. More importantly, the estimated coefficient for race in the model is 1.37, which means that given all others are the same, **the probability for black people to be put into jail is about 4 times the probability for white people to be put into jail**. The 4 here comes from e^(1.36) which is approximately 3.94, where e here is **Euler's number,** which is around 2.7183.



From the graph shown above, we can find that for the people not put into jail, the number of black people is about half of the white; however, when it comes to people put into jail, the number of black people is about twice to the while, which is a **significant change in distribution pattern**.

As a reference, we can also have a look at the distribution for people put into/not put into jail under the comparison of sex and age groups. Keep in mind that our test result suggests gender disparity and age disparity. However, this is not the focus of this project, thus we will not go into details





Demographic interaction Models

Besides the previous demographic predictors addictive models, we also considered a little bit more to see if there is any kind of **interaction between the demographic predictors**. As all three demographic predictors are significant in the additive model, we would use the fourth model we mentioned above as the null model, which is

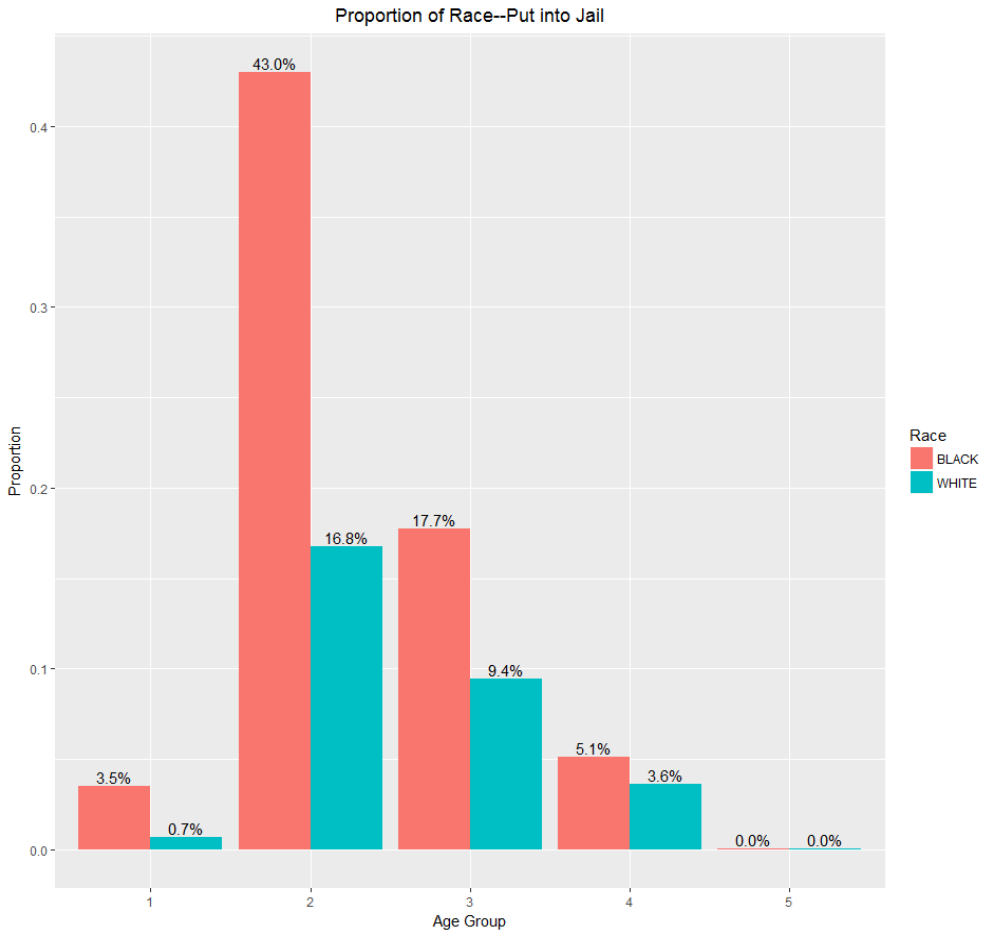
Result ~ all 35 crimes + race + sex + age group

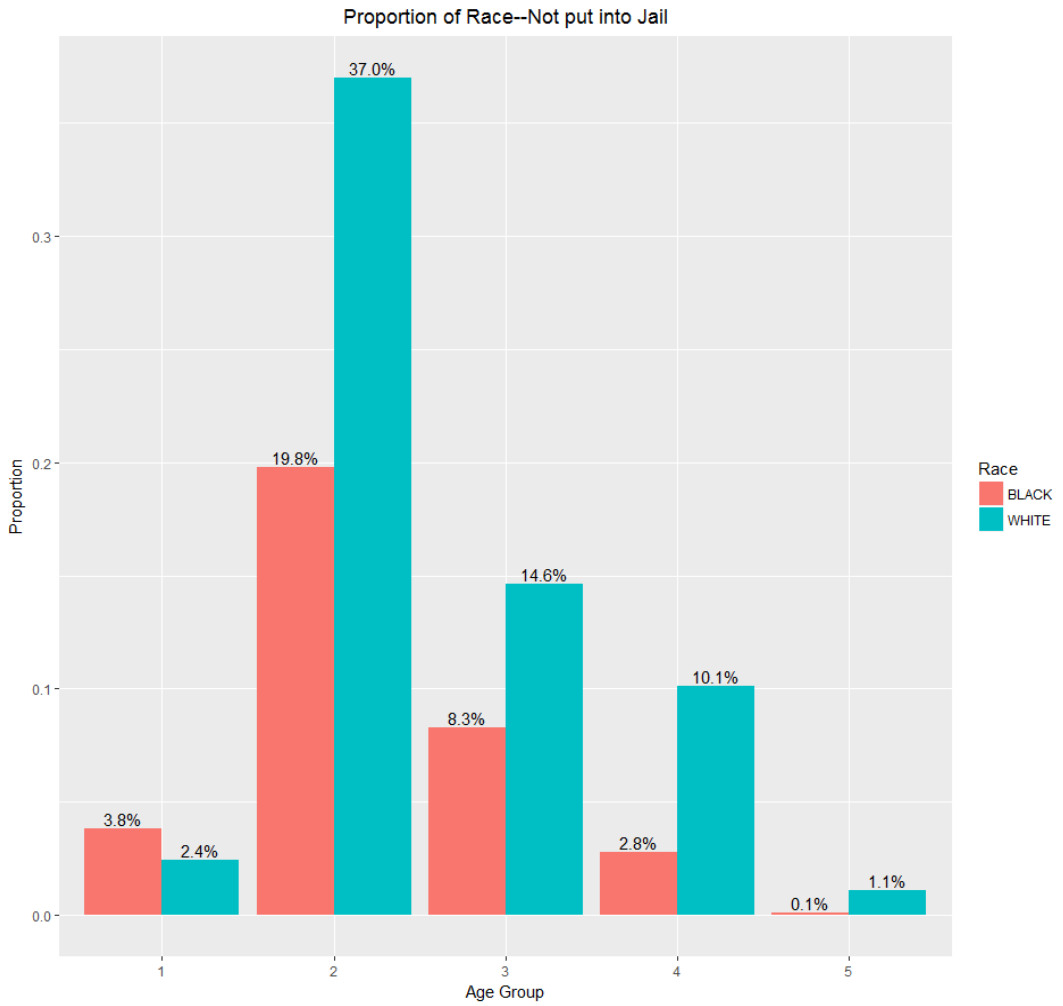
Then we add two race-interaction terms into this model to see if the interaction between the factors is significant.

- Result ~ all crimes + race + sex + age group + age group * race
- Result ~ all crimes + race + sex + age group + sex * race

Result

For the interaction terms, p-value for **age group-race interacted term is significant** with **p-value less than 0.0001**; however, the other interaction term which is sex-age interaction, is not significant as its **p-value is 0.1017**. And we can also see the corresponding graphs for the significant interaction term:





(For the age group here in the graph, 1-5 correspondingly indicate under18, 18 to 30, 30 to 50, 50 to 75 and over 75)

There is a **significant change in the distribution pattern** for **Age Group 2, 3, 4**, which supports our model testing results that the **interaction between race and age groups is significant.**

Individual Crime Models

As some of the observations in the data set are with multiple crimes accused in the same time, and most of the observations (more than 90%, 86866/96821) are accused of only one crime. Thus, for such single-crime-accused observations, we are going to **pull out the observations for each crime separately according to the crime they were accused of**, and test the race-significance using goodness of fit test under such different crime categories.

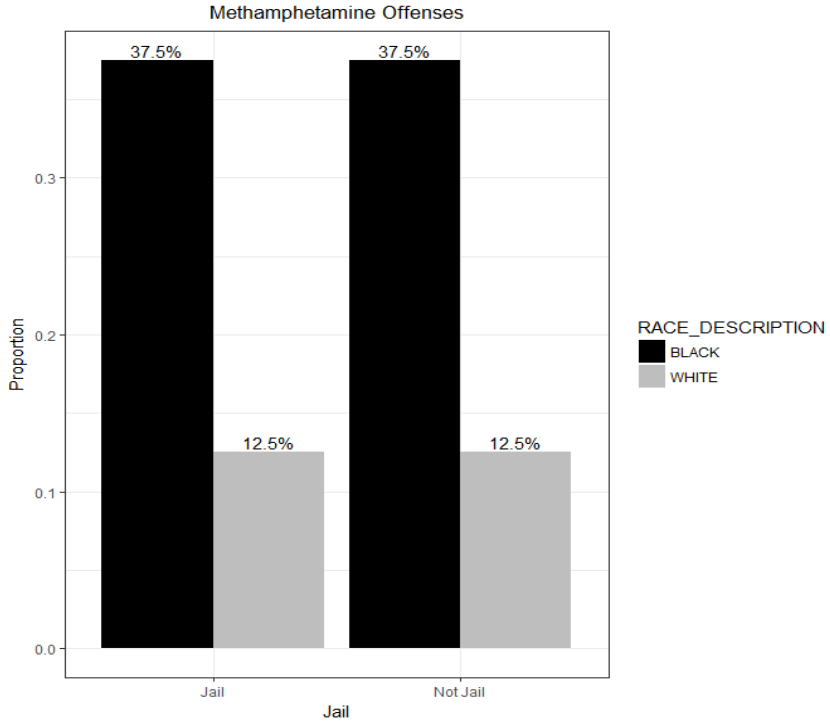
Results

After splitting the single-crime-accused observations into their corresponding categories, **28 out of 35 crimes are race-significant**; also, 11 and 19 out of 35 are age and sex significant.

For race-significant crimes, the significant categories are (with p-value threshold 0.05):

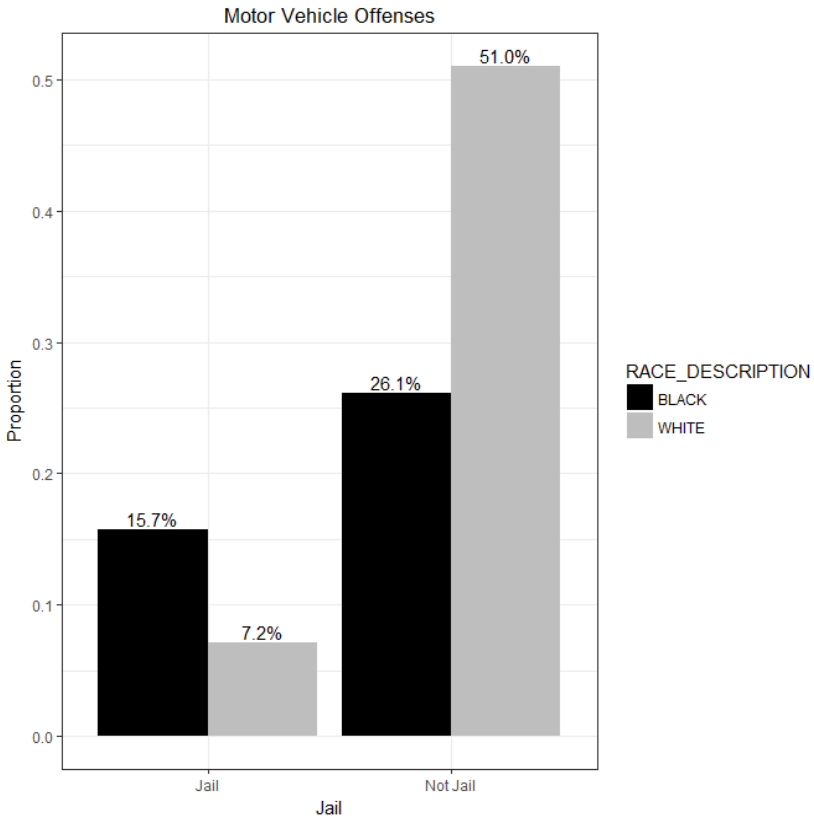
Homicide, Criminal Sexual Assault, Robbery, Battery, Assault, Burglary, Burglary or Theft from Vehicle, Theft, Motor Vehicle Theft, Deceptive Practices, Criminal Damage & Trespass to Prop, Deadly Weapons, Sex Offenses, Cannabis Control Act,
Controlled Substance Act, Drug Paraphernalia Act, Liquor Control Act Violations, Intoxicating Compounds, Motor Vehicle Offenses, Disorderly Conduct, Interference w/Public Officers,
Viol of Criminal Registry Laws, Other Offenses, Uncategorized, MIP, Unlawful use of I.D.,
Bicyclist violation, Noise

Furthermore, we can have a closer look at the graph to distribution for black and white people for non-race-significant crimes and race-significant crimes; first we look at the crime of "**Methamphetamine Offenses**", which is not race-significant:



As we can see, for people put into jail and the ones not put into jail, distributions for black and white people are the same, which indicates that race is **not an influencing factor** in for this crime.

In comparison, let's take another look at the crime of "**Motor Vehicle Offenses**", which takes up about more than half of the observations, and a crime type that the client cares about.



In the MVO crime, we can easily find that for people accused of MVO and put into jail, black people is about twice to the white; on the other side, for people not put into jail, black people is only about half for the white people, in which suggests an inconsistence in race-distribution between the two categories.

Conclusion for Arrestees Dataset :

Around 70% of Champaign County population are White Americans, compared to only 50% of the arrestees are White Americans for a given year. 12% of Champaign County's population are African Americans, but they constitute 38% of the arrests for a given year based on the data that we have. This directly answers the **first question**, and aligns with the definition of racial disparity. **Therefore, there is statistical evidence that there is racial disparity among arrestees in Champaign County.**

We also found out that race is significant in determining whether an arrestee will be put into jail, as well as age group. African Americans in age group 2, 3, and 4 are more likely to be sent into jail than White Americans under the same category. Among the 35 crime types that we considered in this analysis, race was a significant factor in 28 of them. **Therefore, given the same type of offense, race is a significant factor that would influence the outcome of whether an arrestee would be taken to jail.** To see which of the crime type has race as a significant factor, please refer to page 16.

4. BookRJTF Dataset

Introduction

The BOOKRJTF.xlsx data is the jail booking records of Champaign County from 2010 to 2016 received by FOIA request. It has 85372 records corresponding to 50995 Individuals. The variables in this dataset fall into four groups:

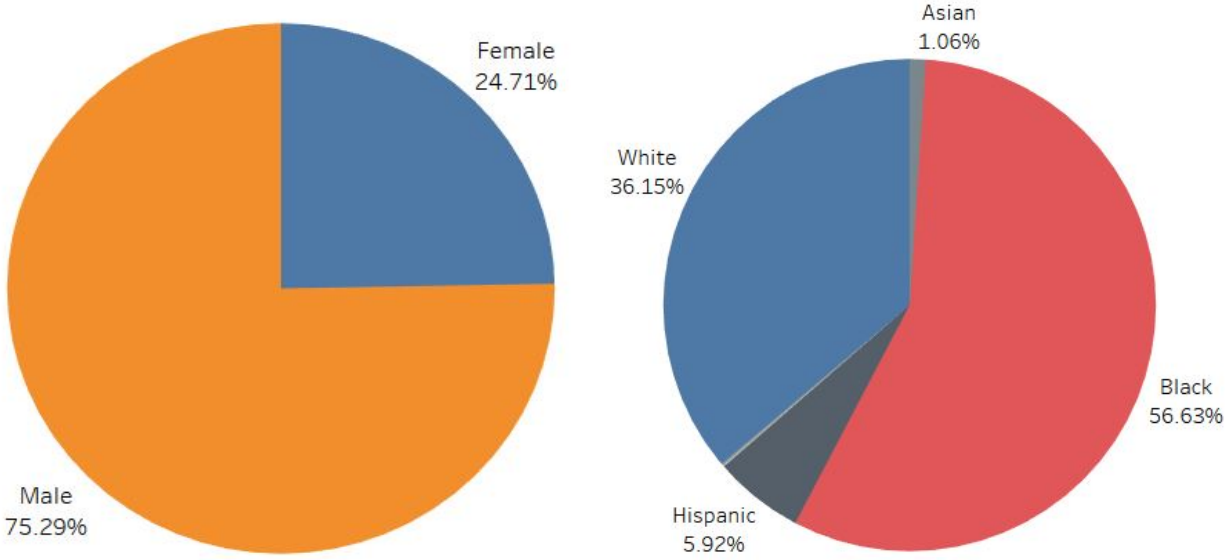
1. Demographic information: Gender, Date of Birth, Race, Address
2. Offenses information: Offenses Description, Statute Number
3. Time information: Booked Date and Time, Court Date, Release Date
4. Bond Amount

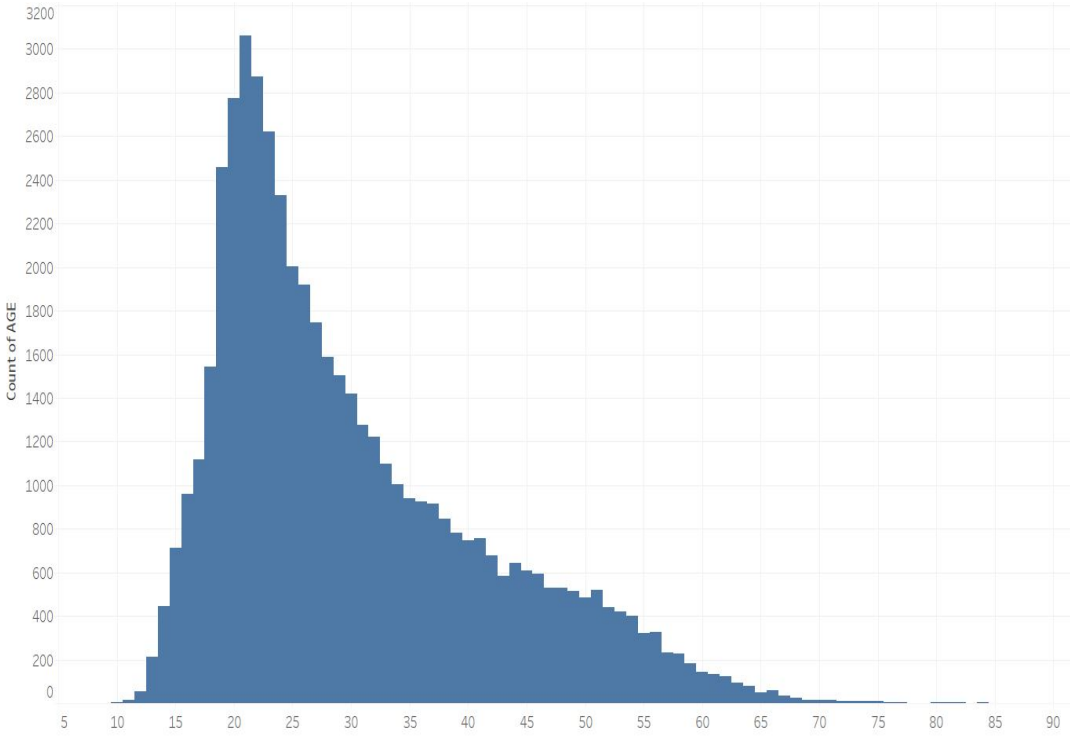
We are going to examine these variables one by one. Eventually, we hope to answer the following question: **Given same offenses, people receive different bond amounts and wait time days for court day and release day. Is this difference related to their race?**

We start with a quick visualization of the demographics variables that we have in our dataset on the next page.

Age, Sex, Race of Inmates

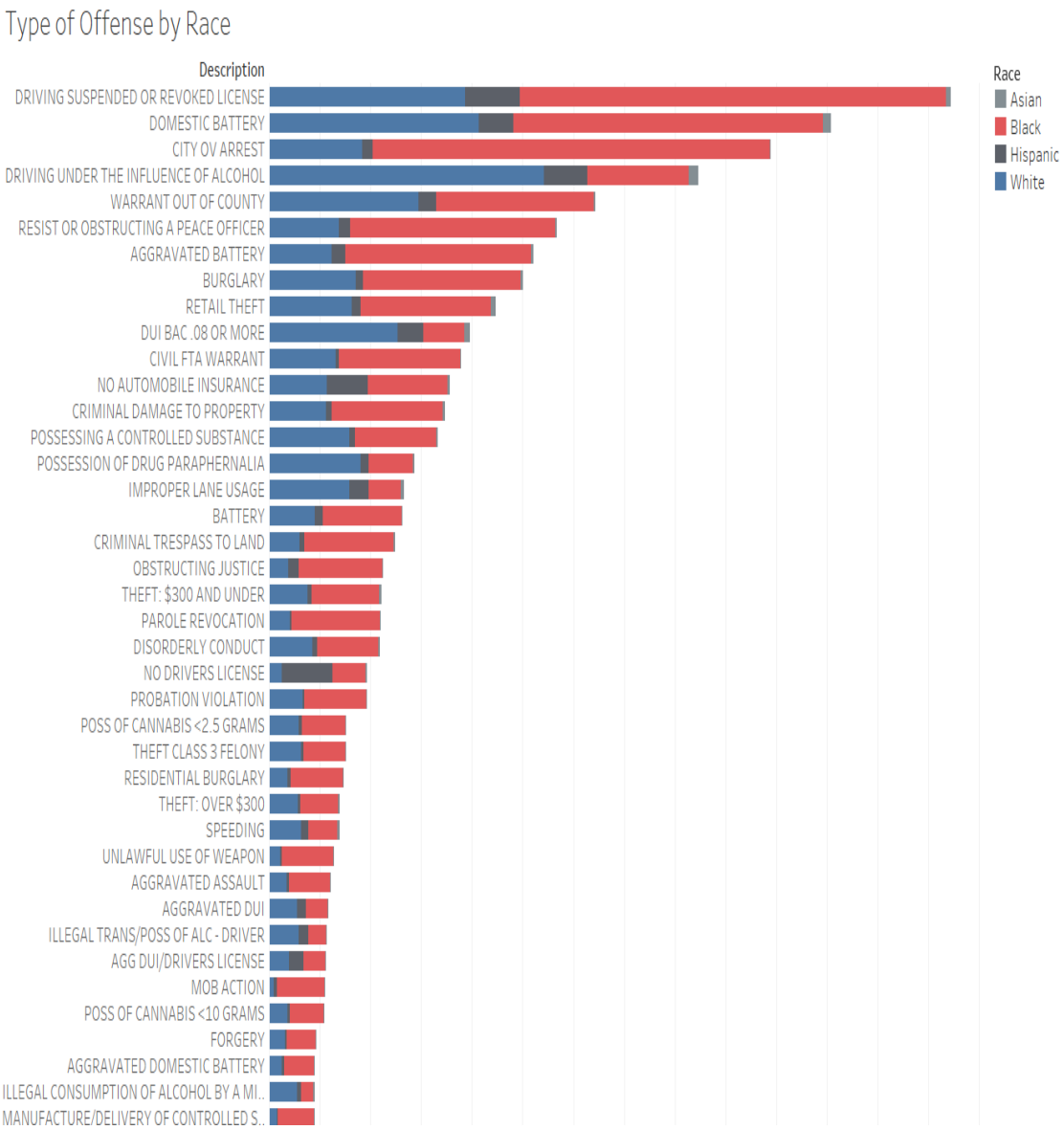
For all people in the booking records, approximately three fourth of them are male. 56% of them are black people while 36% are white people. Most people are young adults in their twenties.





Type of Offense by Race

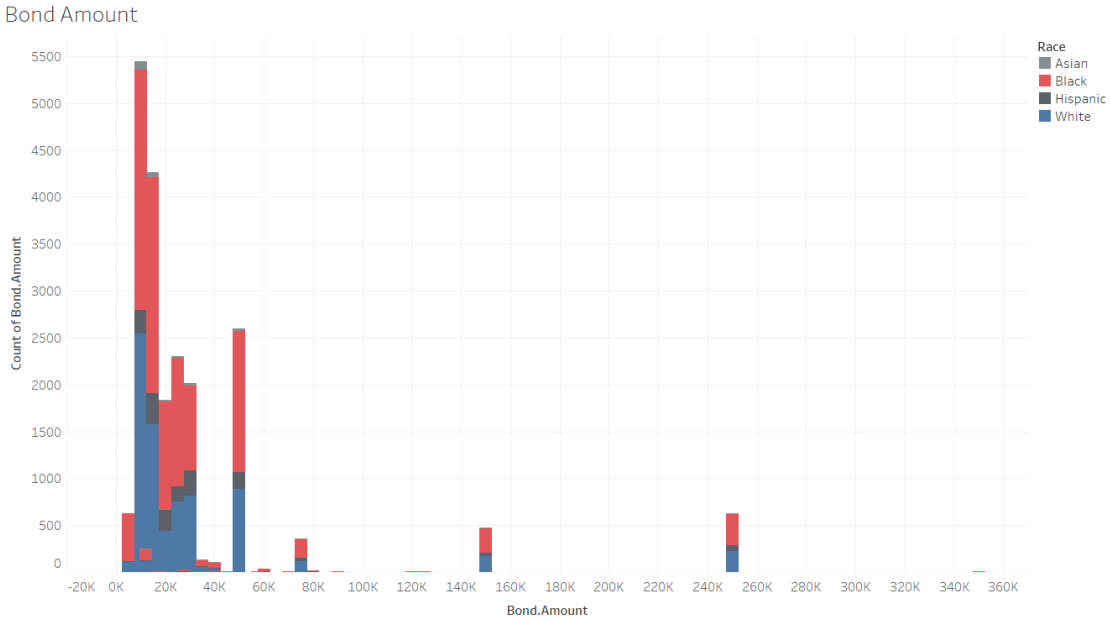
There are 499 unique descriptions of offenses in this dataset. The figure below shows the top 40 most frequent offenses. The red part stands for black people while the blue part stands for white people. The proportion of Black people versus White people is not the same for all offenses. For example, when we look at the third one, **CITY OV ARREST** (violating a city ordinance), we notice that the majority are black people. By contrast, most people who were arrested because of the fourth one, **DRIVING UNDER THE INFLUENCE OF ALCOHOL**, are white people.



Bond amount

In this dataset the majority of bond amount are missing values. We are not sure why they are missing. There are situations that you don't get a bond amount because you do not have the bailout option. However, the missing values are too much to be explained by these extreme cases.

The following figure shows the distribution of bond amount excluding extreme values. Most people have their bond amount set between 0 and 80,000 dollars.



The plot of count of Bond.Amount for Bond.Amount. Color shows details about Race. The view is filtered on Bond.Amount and Race. The Bond.Amount filter ranges from 100 to 350000 and keeps Null values. The Race filter keeps Asian, Black, Hispanic and White.

Given the same offenses, is there a difference between the average bond amounts of Black and White people? To answer this question statistically, we perform t-test.

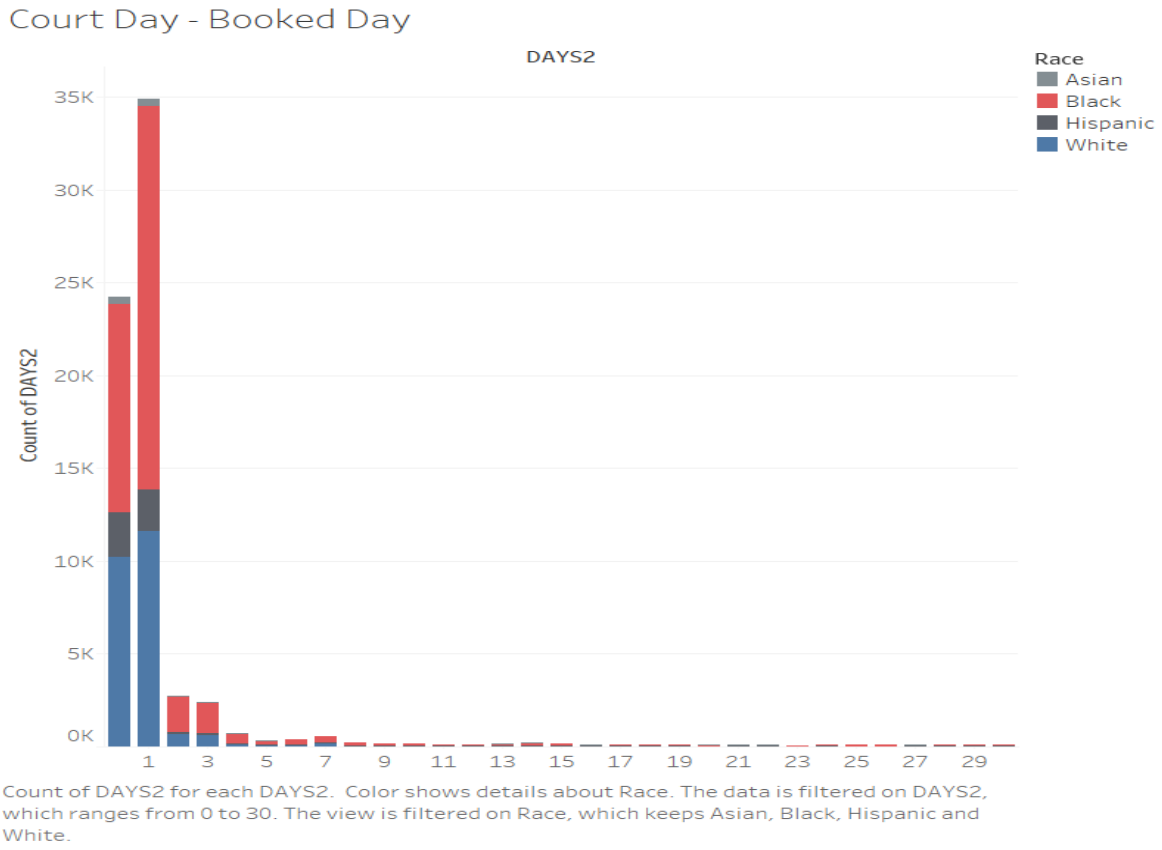
T-test is commonly used to determine if two sets of data are significantly different from each other. In our case, the two groups are the bond amount of Black people and White people. By computing t statistics and corresponding p-values, we reject or do not reject our null hypothesis. A small p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis, so you will reject the null hypothesis. In our case, the null hypothesis is that the average bond amount of Black people and White people are equal.

We perform t-test on the top 40 offenses and it turns out that there is no significant difference between bond amounts except one particular case: UNLAWFUL USE OF WEAPON.

In this case, the average bond amount of White people is $35167 while the average of Black people is $135,889. The p-value is less than 0.05, in other words, we are more than 95% confident that the Black people and White people have different bond amount when they are charged with "unlawful use of weapon".

Time from Booking to Court Date

Most people's court date is the same day as or one day after their booking date. However, there is a long tail to the right. In other words, some people, rare though, wait for several months for their court date. The longest waiting time for court day is 1462 days.

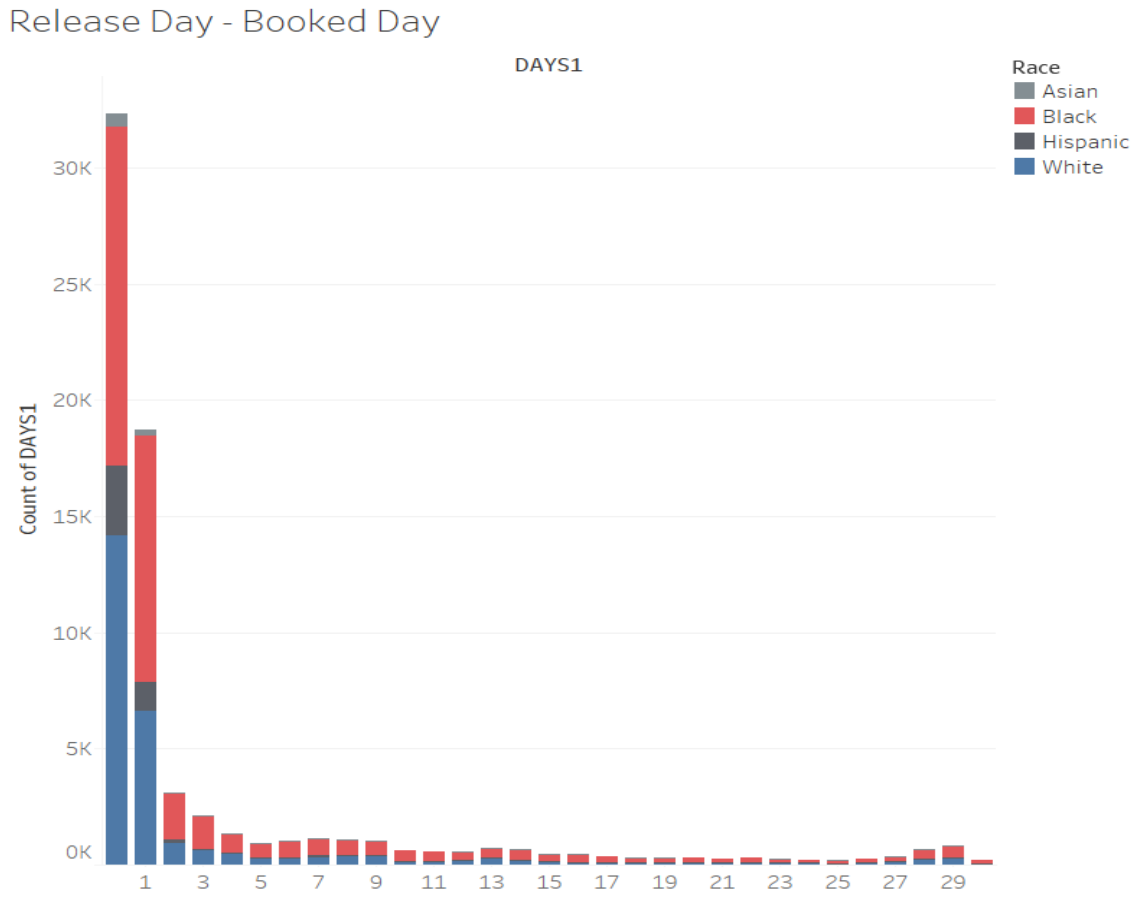


Here we ask a similar question: **Given the same type of offense, is there a difference between the average waiting time of Black and White people? (Booking to Court Date)** We perform t-test on top 40 offenses and find that, there are 5 cases where Black people wait longer for their court day.

Average waiting time for court day		
Offenses	Black people	White people
MAN\DEL CANNABIS 10-30 GR	4.695	2.881
OBSTRUCTING JUSTICE	3.1935	2.2022
Retail Theft	5.2638	4.5333
THEFT: $300 AND UNDER	3.9586	3.6084
UNLAWFUL USE OF WEAPON	8.747	4.099

Time from Booking to Release Date

Similarly, most people's release date is the same day as or one day after their booking date. There is also a long tail to the right. The maximum waiting days for release is 862 days.



Count of DAYS1 for each DAYS1. Color shows details about Race. The data is filtered on DAYS1, which ranges from 0 to 30. The view is filtered on Race, which keeps Asian, Black, Hispanic and White.

Again, **given the same offenses, is there a difference between the average waiting time of Black and White people? (Booking to Release Date)** After performing t-test on top 40 offenses, there are 3 cases where Black people wait longer for their release.

Average waiting time for release		
Offenses	Black people	White people
DUI: DRUGS OR ALC INTOX COMPOUND	25.06	3.08
MAN\DEL CANNABIS 10-30 GR	18.65	6.9
RESIST OR OBSTRUCTING A PEACE OFFICER	20.418	10.855

Conclusion

Although in some cases Black people do pay higher bond amounts or waiting longer time in jail, these special cases are rare if we consider the number of all offenses. Besides, bond amount and waiting time are influenced by many factors that are not essentially in our dataset. For example, we don't know if a person is a first-time offender. We cannot ignore this lack of information.

Thus, we do not think there is sufficient evidence for racial disparities in this dataset.

5. Circuit Clerk Dataset

In this section, we will discuss about the visualization and analysis we conducted regarding the circuit clerk dataset.

5.1 Dataset

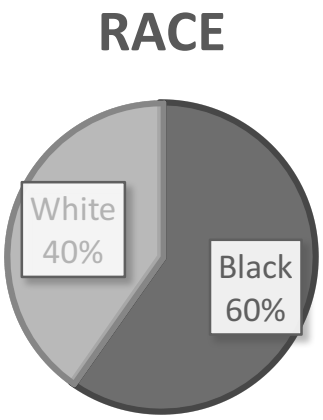
This dataset from the Champaign County Clerk Dataset. In this dataset, we have the information of the court cases throughout the year of 2016 with 3550 observations in total. However, a lot of individuals who were accused of multiple crimes will also have multiple rows within the data set. Therefore, we need to merge the multiple observations of the same person together to prevent duplicate counting. There are also a lot of observations with missing data, thus, we also need to remove the rows that consists of missing data to carry out analysis that is meaningful. After merging and cleaning, we have 873 observations in our data set that corresponds to a unique person.

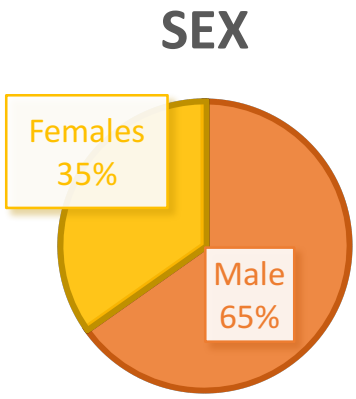
The variables that we will be considering for this data set are the following:

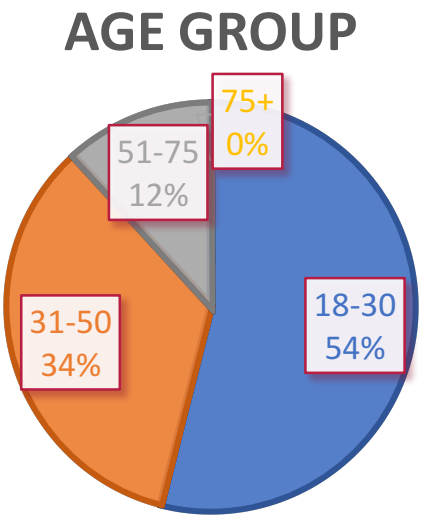
- Demographics:
  - Sex: Male / Female
  - Race: Black / White
  - Age Group: 0 - 18 / 18 - 30/ 30 - 50 / 50 - 75 / 75+
- Charge Agency: Agency where charge was issued
- Charge Type: Felony Class A / Felony Class B / Felony Class C / Felony Class D / Felony Class M / Felony Class N / Felony Class X /Misdemeanor Class A
- Sentence: Prison / Jail

5.2 Visualization

Before we conduct any analysis, let's have an initial idea of what the data looks like. Below, we provided some graphs and plots to visualize the information we currently have.







The first pie chart shows that there are more African Americans that White Americans in our data set with a ratio of 6:4. The majority for gender are male with a percentage of 65% opposed to 35% for females. Most of these people that showed up to court are in their youth, from 18 - 30 years old. People that are older than 30 years old takes up the other half of the sample.

5.3 Questions Considered

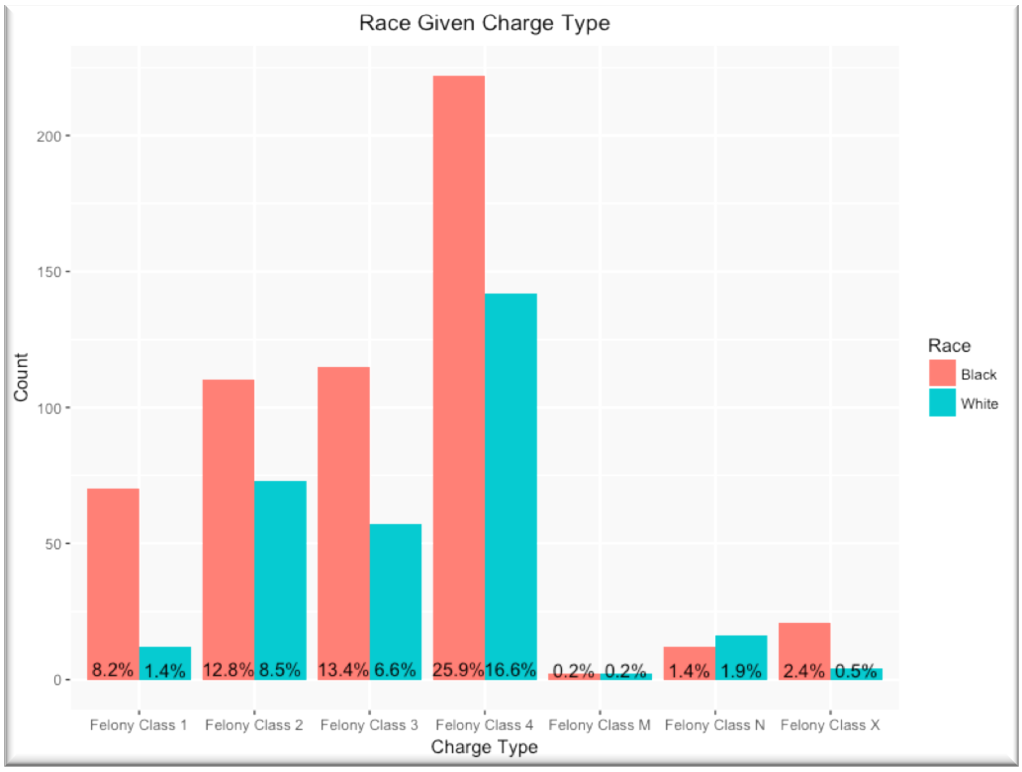
To test if there is any evidence of racial disparity within the court, we primarily look at the following two specific questions:

1. Are African Americans charged with felonies more likely to be imprisoned?
2. Does race cause a significant difference on the length of jail time given the same type of accusation?

5.4 Question 1

We first look at the first question regarding this data set: **"Are African Americans charged with felonies more likely to be imprisoned?"**.

For this matter, we will solely look at 859 samples that were charged with felonies, Below is a bar graph of the number of people within this newly subset samples:



The bar chart above depicts the relationship between race and charge type. We see that out of these 859 samples, most of them fall within the first four felony class type. As we look at the bars of each race, we see that the number of African Americans far exceeds that of White Americans, ranging from 1.5 times (Felony Class 2) to more than 4 times (Felony Class 1), with the only two exceptions being Felony Class M and Felony Class N. It is quite clear visually that there is a significant racial disparity within the Felony Type, as there are much more African Americans charged under felony.

5.4.1 The Entire Data set

We first analyzed the entire data set. That is, we want to build a model that best describe the relationship between the variables within our data set. In other words, we want to find out which of the variables mentioned above has a significant influence on whether a person go to jail or not. Since we need to know the person's sentence to conduct this analysis, we cannot include the samples with this information as missing. After removing these observations with missing data, we have 282 African Americans versus 160 White Americans in this analysis.

Method and Results:

We will be using a ***Sequential Goodness of Fit Test*** on a ***Logistic Regression Model*** to answer the above question. Not wanting to get into too much of the mathematical details of this test, it's basically testing the significance of a variable within a model. Shortly speaking, if adding this variable returns a low p-value, that mean this variable is **significant.** In other words, you can

think of the variable as being an important predictor of whether the person will go to prison. Otherwise, if the corresponding p-value of this variable is higher than 5%, then this variable is **not significant** and should not be included in the final model. Below are the results of this test:

Model	Tested Variable	P-value	Significant?
Prison ~ Agency	Agency	0.9139	No
Prison ~ Agency + Crime	Crime	< 0.0001	Yes
Prison ~ Agency + Crime + Age Group	Age Group	0.6881	No
Prison ~ Agency + Crime + Age Group + Sex	Sex	0.0696	No
Prison ~ Agency + Crime + Age Group + Sex + Race	**Race**	**0.3359**	**No**

The interpretation of the above is rather straightforward. The first column are the models that we considered. On the left hand side of the **tilde (~)** sign is the response: whether the person will go to prison or not. On the right-hand side are the variables that could possibly have significant influence on the response that needs to be tested. As you can see, the models are built upon the ones before it, each time adding a variable into the model to be tested, which is why we call it a sequential test.

1. In the beginning, we start with only testing the relationship of prison and the charge agency, which is equivalent to testing the significance of the variable Agency. The p-value for this test is 0.9319, which is way too large to be considered significant, therefore we have insufficient evidence to claim that the charge agency is a significant influence towards the response. We will later ignore this variable when we construct our final model.

2. Next, in the second model, we add the charge type of the person into the model. We can see that the p-value is extremely small, far smaller than 0.05. Therefore, the variable charge type is very significant towards prison, which also fits our intuitive sense. You are more likely to go into prison if you stab a person rather than just steal his wallet, thus there is no surprises here.

3. As you now have a sense of how this table works, you can see that the demographics variables Age Group, Sex, and most importantly, **Race** are **not significant**. We don't have sufficient evidence to claim that either of three demographics factor has significant influences on the response. Hence, these variables will not be considered in the final model.

Combining the above results, our final model for this part of the analysis is

Prison ~ Charge Type

It seems that no evidence of racial disparity has been found here. However, we were unconvinced of our results and decided to dig deeper. Instead of looking at all the felony charge types together, we will look at each individual felony charge.

5.4.2 Individual Felony Charge Type (Felony Class 1/2/3/4)

Due to the limited observations, we will only look at the four felony charge types with a sufficiently large sample size.

Method:

For this section, we will conduct two different tests with similar goals depending on the sample size. Namely, they are called **Pearson Chi-square Test for Independence** and **Fisher's Exact Test** respectively. The two tests share a main purpose, which is to test if two variables are independent or not. If the p-value is **larger than 0.05**, the test result is insignificant and the two variables are **independent**. Otherwise, if the p-value is **less than 0.05**, then the test is significant and the two variables are **dependent**. The main difference that they have is the **Chi-square Test** is suitable for a larger sample size while the **Fisher's Exact Test** is more accurate when we have a smaller sample size. We will conduct these two tests on the felony charge types accordingly.

However, before we can conduct these two test, we need to construct the relationship of prison and race into a **contingency table.** We will introduce the interpretation of this table below:

	Prison	No Prison	Total
Black	C11	C12	C1+
White	C21	C22	C2+
Total	C+1	C+2	C++

The table above is what you call a **contingency table.** It shows the relationship between the row factor (Race) and the column factor (Prison). Our two tests that we mentioned above will be applied on this type of table to carry out the hypothesis testing.

C11, C12, C21, C22 are called joint cells. The represent the number of people that falls under both their corresponding row category and the column category. For example, **C11** represents the **number of people that were both black and were sent to prison**. Conversely, **C22** represents the **number of people that were white and were not sent to prison**.

C+1, C+2, C1+, C2+ are called marginal cells, which indicate the number of people that either falls under the column factor (C+1, C+2) or the row factor (C1+, C2+). For example, **C1+** represents the **number of people that were black** while **C+2** in**dicates the number of people that were sent to prison**.

C++ indicates the total number of people within this sample.

Results:

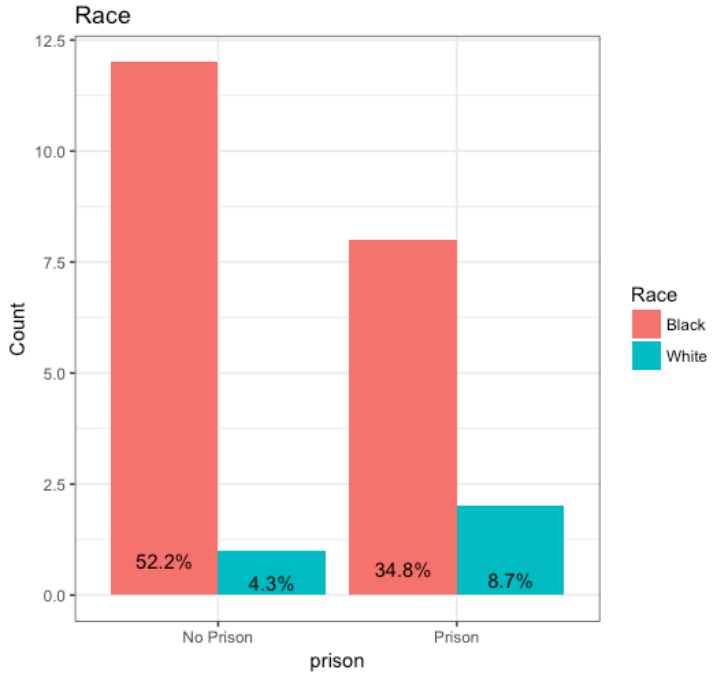
Felony Class 1:

In this sub-sample, we have 20 African Americans to 3 White Americans and 10 prisons to 13 non-prisons. We formulate this information into a **contingency** table as shown below:

	Prison	No Prison	Total
Black	8	12	20
White	2	1	3
Total	10	13	23

From the above table, we see that we have a small and unbalanced sample with respect to race. (23 total, 20 Black, 3 White). Since we have a relatively small sample size, we applied the Fisher's Exact Test to the table and gotten a p-value of **0.5596**, which is not significant. Therefore, we conclude that the outcome of whether a person is going to prison or not is **independent** from the race factor for Felony Class 1.

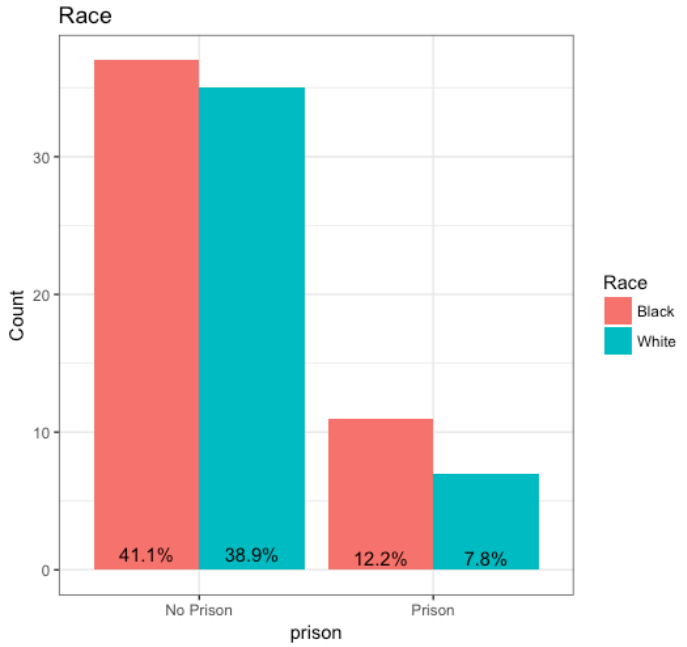
To visualize this result, please look at the bar chart below:



As you can see from the bar graph above, we see that the ratio between Black and White who were sentenced to prison are similar to those who were not sent into prison. This further supports the result of our significance test. Even though there were more Blacks that were sent into prison, there were also more Blacks that were not sent into prison with a very similar ratio. **This means that there is no racial disparity regarding whether the person would be sent to prison.**

Felony Class 2

Similar to the previous felony class, we start off with a bar graph to visualize the situations.



We construct another contingency table to illustrate the data for this felony charge type.

	Prison	No Prison	Total
Black	11	37	48
White	7	35	42
Total	18	72	90

Since we have a large sample size here, we will perform the **Chi-square Test for Independence.** Similar to what we tested for class 1, we are trying to test the dependence relationship between the factor Race and Prison. We obtained a p-value of **0.6345**, which means that we failed to find any dependence relationship. Therefore, we achieve a similar conclusion as in Felony Class 1, that we did not find any evidence of race will influence the outcome of prison.

Felony Class 3

Bar Graph:

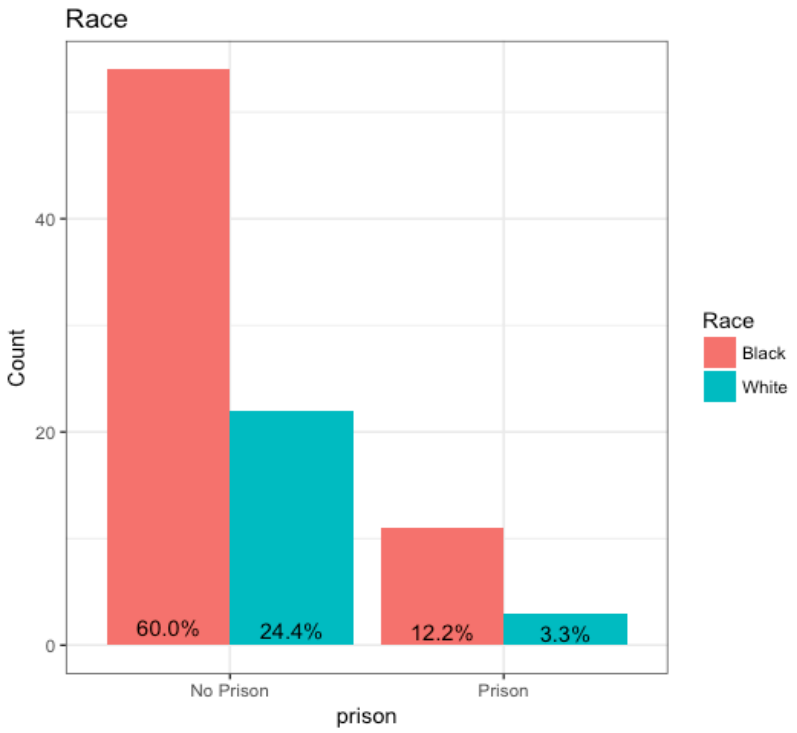


Table:

	Prison	No Prison	Total
Black	9	121	130
White	3	76	79
Total	12	197	209

Test: **Chi-square Independence Test**
P-value: **0.8006**
Conclusion: **P-value is not significant. Not enough evidence to claim that prison is dependent on race.**

Felony Class 4:
Bar Graph:

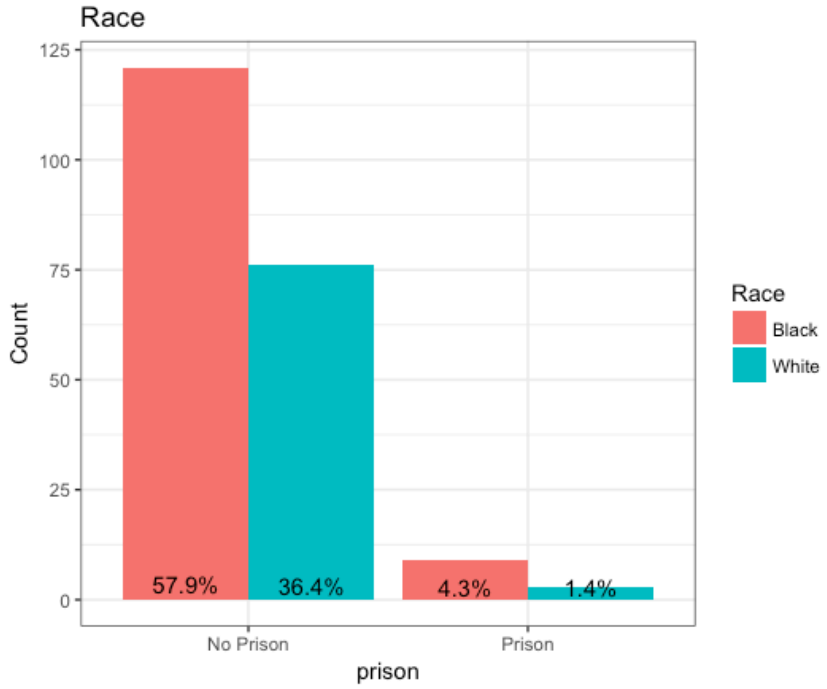


Table:

	Prison	No Prison	Total
Black	12	97	109
White	7	66	73
Total	19	163	182

Test: **Chi-square Independence Test**
P-value: **0.5253**
Conclusion: **P-value is not significant. Not enough evidence to claim that prison is dependent on race.**

5.4.3 Conclusion to Q1

We've reached consistent conclusion from the above conducted analysis and hypothesis testing. Initially, if we look at the entire data set at once and test the importance of each variable sequentially, the only important variable will be the type of charge, while **Race** and the other variables are not significant. We then look at each individual felony charge to find additional evidence to support the previous claim, and found out that **Prison** is **independent** from **Race** for all four types of felony charges considered.

We failed to find any evidence that African Americans charged with felonies are more likely to be imprisoned than their white counterparts. There is no evidence of racial disparity in this part of the analysis.

5.5 Q2 2

In this part of the analysis we are trying to figure out if **race will cause a significant difference in the amount of jail time given the same accusation.** Note that we are looking at the length of jail time and not the length of prison time, therefore this part of the analysis has nothing to do with what we analyze in the previous problem. Again, due to the need of a reasonably large sample size to carry out our test, we will be looking at the following five different charges.

- Felony Class 2/3/4
- Misdemeanor Class A

As always, let's have a quick visualization of the race distribution within these above selected charges.

Method:

Similar to what we've done in the booking data set, we will be conducting a 2 sample test to see if there's a significant difference between the length of jail time that Black and White convicts receive. Again, if we have a small p-value, that will be evidence that there is a significant difference between the length of jail time. If not, then we failed to find any evidence that the difference between jail time is statistically significant.

Results:

Charge Type	Average of Black Jail Time (Days)	Average of White Jail Time(Days)	P-value	Significant?
Felony Class 2	47.00	35.07	0.5022	No
Felony Class 3	35.05	37.94	0.7907	No
Felony Class 4	48.40	47.44	0.9093	No
Misdemeanor Class A	18.21	22.04	0.3445	No

As we see from the table above, the average jail time of Black and White has some obvious difference between each other. For example, for **Felony Class 2**, the average jail time for Blacks is 47 days compared to the average jail time of Whites which is only 35 days. However, we can also see that the average jail time for Blacks is also surprisingly less than Whites for all the other types of felonies. Therefore, is the difference between these two quantities due to chance or are they statistically significant?

If we look at the p-value of these four comparisons, we can see that none of the p-values are significant, since none of them are below 0.05. Therefore, we did not find any statistical evidence that the difference between the jail time of the two race is significant.

Conclusion to Q2

From the table above, we see that none of the p-values of the difference is significant. Therefore, **we did not find any evidence of racial disparity in the length of jail time, at least from this data set.**

6. Conclusions

6.1 Conclusions of Arrestee Dataset

- Around 70% of Champaign County population are White Americans, compared to only 50% of the arrestees are White Americans for a given year. 12% of Champaign County's population are African Americans, but they constitute 38% of the arrests based on the data that we have.

- The race term is significant using the sequential goodness of fit test on the entire dataset. African Americans are 4 times more likely to be put into jail than White Americans given all the other variables are held fixed.

- When examining different crimes separately, 28 out of the 35 crimes are race-significant, which echoes the result for the entire dataset

6.2 Conclusion of BookRJTF Dataset

- At the 95% confidence level, black people have higher average bond amount or waiting time for both court date and release date in several offenses. (Refer to Pg. 20 - 22) However, these numbers cannot be taken literally as there are too many underlying confounders. More data is needed to reach a more concrete conclusion)

6.3 Conclusion of Circuit Clerk Dataset

- The race term is not significant when using the sequential goodness of fit test. Therefore, African Americans that are charged with felonies are not more likely to be sent to prison according to our analysis.

- According to the results of the t-test, there is no significant difference in the length of jail time under the same charge type due to race. The difference that are observed in the sample are merely due to chance.

7. Limitations

There were many tests that we thought of conducting during our analysis but were unable to do so due to the lack or invalidity of the data that we were provided. We address these issues here so that the RJTF will know what additional data to collect in order to obtain more accurate statistical results.

6.1 General

- Names of arrestees: This information is extremely important. Knowing this information will allow us to merge multiple datasets together for stronger statistical power in our analysis, since we will have more information to work with for a single person. Currently, all three datasets are disjoint and we can only analyze them individually. We will also be able to know if a person has committed any crimes in the past, which was a huge issue that we encountered during our analysis for the **BookRJTF** dataset.

- Missing Data: There were many missing data in all three datasets, especially in the **BookRJTF dataset** and **Circuit dataset**. We must remove observations with missing value in our analysis, which reduces the effective sample size and thus statistical power.

6.2 Arrestee Dataset

- Employment Code: There were 13 different employment code within our dataset, but only 5 of them were documented. We were unable to utilize this information as we don't know what most of the code means.

- Location of Arrest: Currently we only have the street names of the arrests. It would be great if we can map these street names into a larger region so we can look at the arrests from a larger scope to pin-point the areas of Champaign County that have a higher volume of criminal activities. A naive way is to map these streets manually by human power, but these is a very tedious and time-consuming job for us as there are thousands of street names. An automated software that can deal with this issue will be ideal.

6.3 Booking Dataset

- Prior Criminal Records: Since there are many factors that influence the bond amount of the arrestees, a lot more information is needed than just the charge type of the accused. Since bond amount is dependent on whether the accused has prior criminal records, this information will allow our analysis of bond amount to be a lot more accurate.

- Failed to Pay Bond: If a person did not pay his bond amount, he might have to wait a lot longer for release. However, the lack of this information will result in the inaccuracy of our test on waiting time and release date.

8. Online Source for Statistical Test

- Logistic Regression and Goodness of Fit Test: https://onlinecourses.science.psu.edu/stat504/node/216

- Fisher's Exact Test: http://www.biostathandbook.com/fishers.html

- Pearson Chi-square Test of Independence: http://stattrek.com/chi-square-test/independence.aspx?Tutorial=AP

- Two Sample T-test: https://www.socialresearchmethods.net/kb/stat_t.php