DATA VISUALIZATION FOR ANALYTICS

DATA VISUALIZATION OF US CRIME DATA

TABLE OF CONTENTS

Deliverables 1

<u>**Introduction to Tableau (0 points)**</u>

Tableau software is one of the leading software tools available for data visualization and exploratory data analysis. In this course, *Tableau will be used as the primary tool* for visualizing one-dimensional data (e.g., temporal data); two-dimensional data (e.g., geospatial data); multidimensional data (e.g., mapping relational data in n-dimensional space); as well as hierarchical tree data. Tableau provides a powerful toolset for identifying patterns and trends in complex datasets as well as creating spectacular visualizations of data from variety of domains (e.g., science, business, engineering, social media, etc.). In addition, Tableau has an intuitive user interface; it is compatible with a multitude of platforms; and easily handles large scale datasets.

While your hands-on work using Tableau will begin in future weeks, your assignment for this week is to download and install Tableau Desktop and begin watching the Tableau training videos to familiarize yourself with the basic features and functionality of Tableau. The Tableau software company provides an extensive suite of tutorial resources that you may find valuable throughout the course. Prior to future weeks assignments, please make sure that you have viewed the training videos listed below.

Instructions for Downloading and Installing Tableau Desktop

Each student should go to the landing page (see link below) to download the most recent version of Tableau Desktop and enter the product key noted below. This key will activate Tableau for the duration of the course.   Note: if you already have the latest version of Tableau Desktop loaded on your computer and an active license, there is no need to re-install Tableau Desktop.

1. [Download the latest version of Tableau Desktop and Tableau Prep Builder here](#)
2. Click on the link above and select "Download Tableau Desktop" and "Download Tableau Prep Builder". On the form, enter your school email address for Business E-mail and enter the name of your school for Organization.
3. Activate with your product key: TCYJ-FA5B-BAA0-F115-E53F
4. Already have a copy of Tableau Desktop installed. Update your license in the application: Help menu → Manage Product Keys

Note from Tableau: Students can continue using Tableau after the class is over by individually requesting their own one-year license through the [Tableau for Students program here](#) . Tableau has also provided you with access to their 'Tableau Prep' software, which is a powerful tool for data processing and cleaning. Although we will not use Tableau Prep in this course, you are welcome to explore the features and functionality of this powerful software tool on your own. If you choose to do so, the activation key above can also be used to activate Tableau Prep.

View the following Tableau training videos at:

Getting Started section, please watch the following 3 videos
(http://www.tableau.com/learn/training#getting-started):

- o Getting Started (25 min)
- o The Tableau Interface (4 min)
- o Distributing and publishing (4 min)

- **Under Connecting to Data, please watch the following video:**
    - o Data Prep with Text and Excel Files (5 min)

Deliverables 2

10 points

Goal of the course project: Leveraging the design principles and techniques that you will be learning in this course, you have to create an interactive, exploratory dashboard that includes at least three well-designed data visualizations as well as filter controls for key variables of interest that will allow the user to explore the data at a deeper level. Note: the three visualizations and filter controls should all be integrated into a SINGLE dashboard design. Storyboards that link a sequence of separate worksheet visualizations using a narrative are not acceptable for this project.

You may use any data set (other than the pre-packaged datasets that come with Tableau) that you find interesting and intriguing to create a dashboard that tells a compelling story. You have complete freedom in choosing the data domain as well as the visualization techniques but make sure you produce a high quality, interactive visualizations for your exploratory dashboard.

This goal can be achieved over the period of 5 weeks (week 2 till week 6).

In order to archive the goal of the course project we will perform the following steps

1. Define a research/business goal.
2. Describe the research/business queries related to the goal.
3. Identify the dataset for the research.
4. Explore the dataset and the visualization techniques.
5. Create an interactive and exploratory dashboard.

In this week you have to complete the following tasks

1. Define your research/business goal.

    - To analyze which agency solved crimes in different regions of the US

2. Briefly describe the queries (research/business questions) you want to address. Limit the number of queries to a maximum of four.

    - Weapons used for crimes that were solved by various agency types
    - The number of crimes committed by different age groups, gender, and race.
    - Types of weapons used by the perpetrator and their relationship with the victim.
    - Crimes committed over the years, since 1976 among multiple states.

3. Briefly discuss the source(s) from which you are planning to collect the data to address your research/business queries.

    Will be using crime data from one of these sources-

    1. [FBI's Supplementary Homicide Report from 1976 to 2014 on KAGGLE](#)
    2. [FBI's Hate Crime Statistics](#)

Deliverables 3

10 points

In this week your goal will be to identify the data set that will help you address your business/research goals. In addition to that you have to briefly discuss about the dataset. The following data repositories can be explored to identify a dataset to address your business/research goals

- Data Science Central
- National Science Fund
- Quora
- Scribble

Please remember this is not an exhaustive list of data repositories.

Once you have identified the dataset you have to complete the following tasks:

1) Provide a brief description of the dataset. The description could involve listing the following: the attributes, the data types of the attributes, number of instances, attribute values, missing values, other types of issues in the dataset if any, etc. You can follow the format for the data description provided in the Car Evaluation database (archive.ics.uci.edu/ml/machine-learning-databases/car/car.names)

- **Title:** FBI's Supplementary Homicide Report (1980 -2014)

- **Source:** FBI's Supplementary Homicide Report from 1976 to 2014 on KAGGLE

- **Past Usage:** The dataset incorporates murders from the FBI's Supplementary Homicide Report from 1980 to 2014 and the Freedom of Information Act. This dataset incorporates the age, race, sex, nationality of casualties and culprits, the connection between the person in question and culprit, and the weapon utilized. The information was accumulated and made accessible by the Murder Accountability Project, established by Thomas Hargrove

- **Number of Instances:** 638454
- **Number of Attributes**: 24

- **Attribute Values and data types:**
  - ➤ Agency Type: (character) 'Municipal Police' 'County Police' 'State Police' 'Sheriff' 'Special Police' 'Regional Police' 'Tribal Police'.
  - ➤ Crime Type: (character) 'Murder or Manslaughter' 'Manslaughter by Negligence'
  - ➤ Crime solved: (character) 'Yes', 'No'
  - ➤ Victim Sex: (character) 'Male' 'Female' 'Unknown'
  - ➤ Victim Race: (character)'Native American/Alaska Native' 'White' 'Black' 'Unknown'
  - ➤ 'Asian/Pacific Islander'
  - ➤ Victim Ethnicity: (character) 'Unknown' 'Not Hispanic' 'Hispanic'
  - ➤ Perpetrator Sex: (character) 'Male' 'Unknown' 'Female'
  - ➤ Perpetrator Race: (character) 'Native American/Alaska Native' 'White' 'Unknown' 'Black', 'Asian/Pacific Islander'

- ➢ Perpetrator Ethnicity: (character) 'Unknown' 'Not Hispanic' 'Hispanic'
- ➢ Record Source: (character)'FBI' 'FOIA'
- ➢ Agency Code: (character)12003 different codes
- ➢ Agency Name: (character) 9216 different names
- ➢ City: (character)1782 different cities
- ➢ State: (character)51 different states
- ➢ Year: (integer) 35 years
- ➢ Month: (character)12 months
- ➢ Incident: (character)1000 incidents
- ➢ Victim Age: (Integer)101 different ages
- ➢ Relationship: (character)28 different relationships
- ➢ Weapon: (character)16 different weapons

- **Missing Attribute Values**: none

- **Class distribution:**

Agency Type:
Municipal Police   0.772219
Sheriff   0.164964
County Police   0.035544
State Police   0.022296
Special Police   0.004525
Regional Police   0.000368
Tribal Police   0.000085

Crime Type:
Murder or Manslaughter   0.985722
Manslaughter by Negligence   0.014278

Crime solved:
Yes   0.701964
No   0.298036

Victim Sex:
Male   0.773940
Female   0.224519
Unknown   0.001541

Victim Race:
White   0.497173
Black   0.469727
Asian/Pacific Islander   0.015491
Unknown   0.010457
Native American/Alaska Native   0.007153

Victim Ethnicity:
Unknown   0.576867
Not Hispanic   0.309339
Hispanic   0.113794

Perpetrator Sex:

    Male   0.625794
    Unknown   0.298166
    Female   0.076040

Perpetrator Race:

    White   0.341830
    Black   0.335993
    Unknown   0.307065
    Asian/Pacific Islander   0.009470
    Native American/Alaska Native   0.005642

Perpetrator Ethnicity:

    Unknown   0.699205
    Not Hispanic   0.227381
    Hispanic   0.073415

Record Source:

    FBI   0.965844
    FOIA   0.034156

2) Describe the types of data and measurement scales (nominal, ordinal, interval, and ratio) across each attribute. For additional details, please view Types of Data and Measurement Scales

- **Attributes definitions and measurement scales:**

  ❖ Record ID: (Nominal) Unique crime ID for each crime.
  ❖ Agency Code: (Nominal) Unique for each agency
  ❖ Agency Name: (Nominal) Agencies involved in solving crime
  ❖ Agency Type: (Ordinal) Different types of law enforcement agencies.
  ❖ City: (Nominal) City where the crime happened
  ❖ State: (Nominal) state in which the city is located
  ❖ Year: (Interval) year when the crime happened
  ❖ Month: (Interval) In which month the crime happened.
  ❖ Crime Type: (Nominal) Type of crime
  ❖ Crime Solved: (Nominal) If the crime has been solved or not
  ❖ Victim Sex: (Nominal) Gender of victim
  ❖ Victim Age: (Ratio)Age of victim
  ❖ Victim Race: (Nominal) Victim's race
  ❖ Victim Ethnicity: (Nominal) Victim's Ethnicity
  ❖ Perpetrator Sex: (Nominal) Person's gender who committed the crime
  ❖ Perpetrator Age: (Ratio)Age of the culprit
  ❖ Perpetrator Race: (Nominal) race of the perpetrator

- ❖ Perpetrator Ethnicity: (Nominal) ethnicity of perpetrator
- ❖ Relationship: (Nominal) relationship between victim and perpetrator
- ❖ Weapon: (Nominal) Weapon used to commit the crime
- ❖ Victim Count: (Ratio) no of victim's deaths
- ❖ Perpetrator Count: (Ratio) no of culprit's deaths
- ❖ Record Source: (Nominal) crime record's source

Deliverables 4

10 points

In this week your task will be to provide details about the purpose of the dashboard, who its users will be, what data will drive each visualization in the dashboard, a list of analytic questions and queries that a person should be able to answer using the dashboard visualizations.

Purpose of the Dashboard:

- To increase the crime-solving rate by analyzing the pattern among different types of crime/ relationship between victim and perpetrator across multiple regions.

Potential Users:

- Government bodies can use the interactive dashboard to analyze various agencies that successfully solved crimes in different regions of the US.
- It can also be used by different agencies to observe the pattern between different crimes.
- Agencies that record the crime data (FBI/FOIA) can use this dashboard to increase the crime-solving rates by assigning crimes to agencies who have solved similar crimes in past with a high success rate.

Analytic questions and sub-queries and the attributes used to answer them:

1)Weapons used for crimes that were solved by various agency types:

a) Success rate of crimes solved by various agencies.

Attributes used: Agency Type, weapons, Percentage of success rate (calculated field)

2) The number of crimes committed by different age groups, gender, and race.

a) Can be used to analyze which age groups and gender are responsible for the majority of crime from a particular race.

Attributes used: Perpetrator age, Perpetrator race, Perpetrator sex

3)Types of weapons used by the perpetrator and their relationship with the victim.

a) Which weapons do perpetrators use to commit crimes when they have a certain relationship with the victim.

Attributes used: Relationship, weapon, crime type

4)Crimes committed over the years, since 1976 among multiple states.

a) Trend of crimes reported in different states over 30 years timespan

b) Most unsafe states where crimes are committed and dive to granular level by analyzing most crimes committed in a city/postal code.

Attributes used: Year, Month, City, State, Number of crimes solved, and number of crimes unsolved (calculated fields)

Deliverables 5

10 points

In this week your task will be to provide the rationale for the different design principles and techniques you will be using in creating the interactive dashboard visualizations.

1. **Visualization 1:Bar Graph Side-by-Side**

   The purpose of this graph is to visualize crimes solved by various agencies based on the weapons used for committing the crime. In the dataset, only states are given, so we are creating groups(east, west, central, north, south) for the state variable, so that it will be easier for the user to infer data dynamically in different regions of the USA. Creating groups also helps us reduce clutter in the final dashboard. Weapon & Agency type lies on Y-axis. Percentage of crime solved(calculated field) on X-axis. The state is used in filters to make the visualization interactive & the Agency type is used on color in marks card. The length of the bar represents the percentage of crimes solved. Agency type being categorical attribute can be best perceived by color hue. We can also observe that the visualization follows the Gestalt principle of enclosure.

2. **Visualization 2: Bullet Graph**

   The purpose of this visualization is to find out the relationship between the perpetrator's age and the victim's age based on their gender and race. After creating few calculated fields we calculate the average age of perpetrators and group them according to different age groups (teens 9-18, Young adults 18-35, adults- 35-60, senior 60-100 years). Based on these perpetrators' age groups the length of bars in the chart represent the average age of groups from a particular race of a particular gender and the black line represents an average of victims age killed by that category of the perpetrator. Through this, we can observe if an age group of a race and gender commits a crime on people younger or older to them. Here the x-axis represents the average perpetrator age categorized into 4 age groups, the y-axis represents the perpetrator's race and gender. Filters are used to remove unknown perpetrator sex and race. Perpetrator race is shown by color from marks card and average victim age is present in the detail of marks card. This helps the chart to be more visually attractive. Gestalt principle of similarity is used to group perpetrator's race by color and principle of enclosure to enclose different age groups (teen, adult, young adult, senior)

3. **Visualization 3: Packed Bubble Chart**

   This visualization aims to analyze the types of weapons used by the perpetrator and their relationship with the victim. For this visualization, we will use a bubble chart, where the number of crimes done is calculated by counting the record ids. This quantitative value is depicted using the size of the bubble. The colors of the bubble are used to distinguish between the weapons of choice which is a qualitative attribute. The bubble charts make an excellent visualization as 3-dimensional attributes can be used at a time. Both the weapon and relationship attributes are used as filters, this will allow us to visualize how a perpetrator's relationship can be used to find the weapon of crime. We can observe the gestalts principle of similarity here as the same weapons have the same colors.

4. **Visualization 4: Line Graph**
   This visualization helps us analyze the trend of crimes commited since 1976. Groups-(north,east,west, central,south) are created from the state attribute to visualize which part of USA has more crimes over 30 years time span. Line graphs are used to compare changes over the same period of time for more than one group. Also, since it's a time-series data line chart is best to use. Year attribute is on Y-Axis. State attribute and count of ID is on X-axis. ID attribute, being a quantitative data, is represented by position. The higher the crimes commited the more is the count of data point. We are also differentiating each state by color hue. We can observe the Gestalt Principles of visual perception in the chart. Since there's a continuity in data, we can see gestalt principle of continuity and connection .Since it is a time series data, to view crime rate annually over the years, year attribute will be added to pages which makes the dashboard animated.

   The dashboard will be exploratory dashboard since the users will be able to analyze the visualizations region and year wise. This will allow a granual level visualization as one can observe types of crimes solved by agencies, age group responsible for the crime, and the relationship between criminals and victims in different regions and also observe the trend over the years. The dashboard will also contain few animations since trend of each visualization can be checked over years and some insightful observations can be made from them.

Deliverables 6

10 points

In this week you have to submit your interactive and exploratory dashboard that you have created as part of your course project. Please submit a single packaged workbook (.twbx) that contains all the resources needed for users to run the dashboard in the canvas dropbox.

Submitted via E-mail
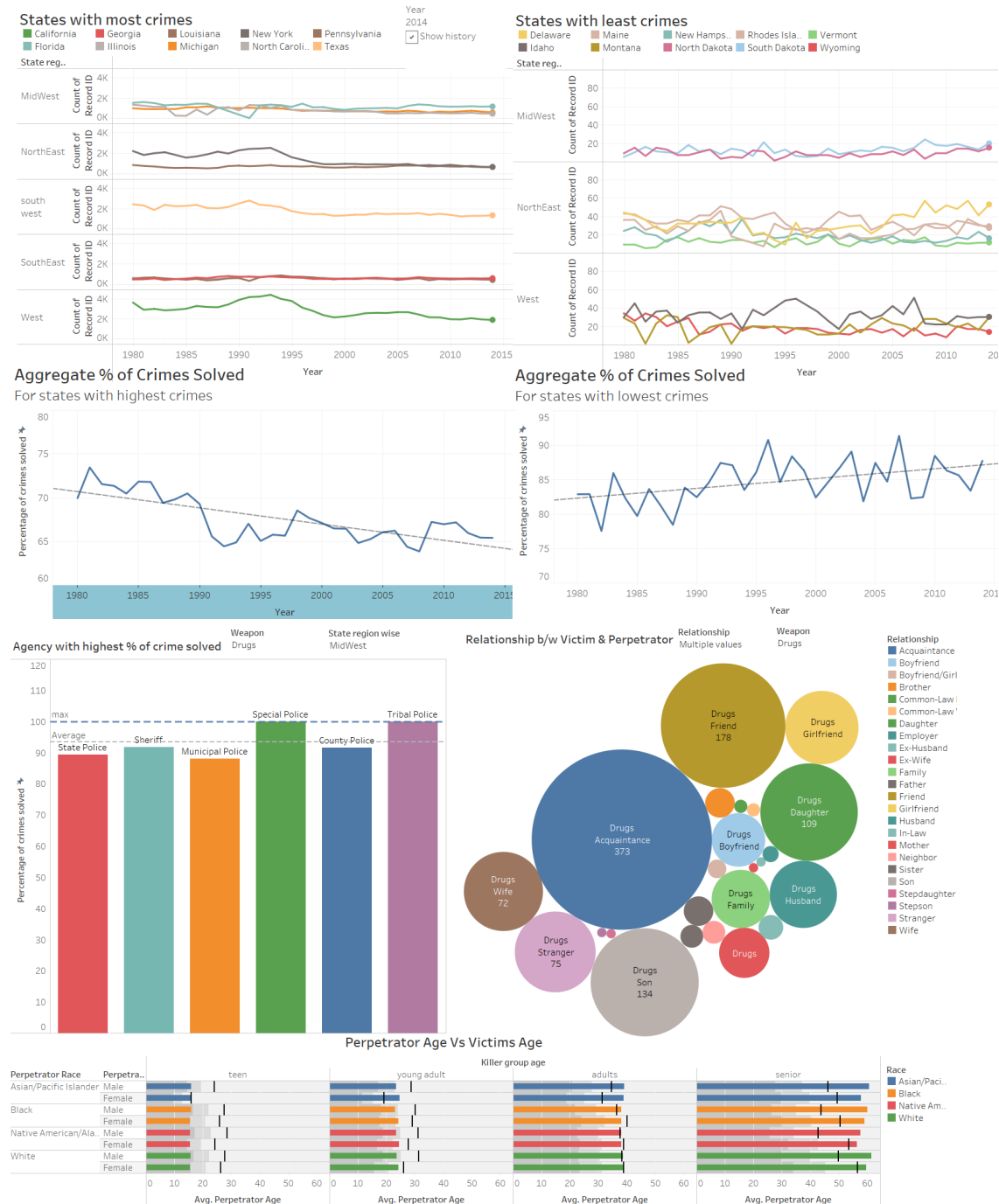
Deliverables 7

(10 + 5) points

Prepare a series of screenshots along with informative captions to illustrate the capabilities of your dashboard and guide a viewer through a particular use case.

Suggestion: Use as many screenshots as necessary to adequately communicate the dashboard capabilities and example use of your dashboard.
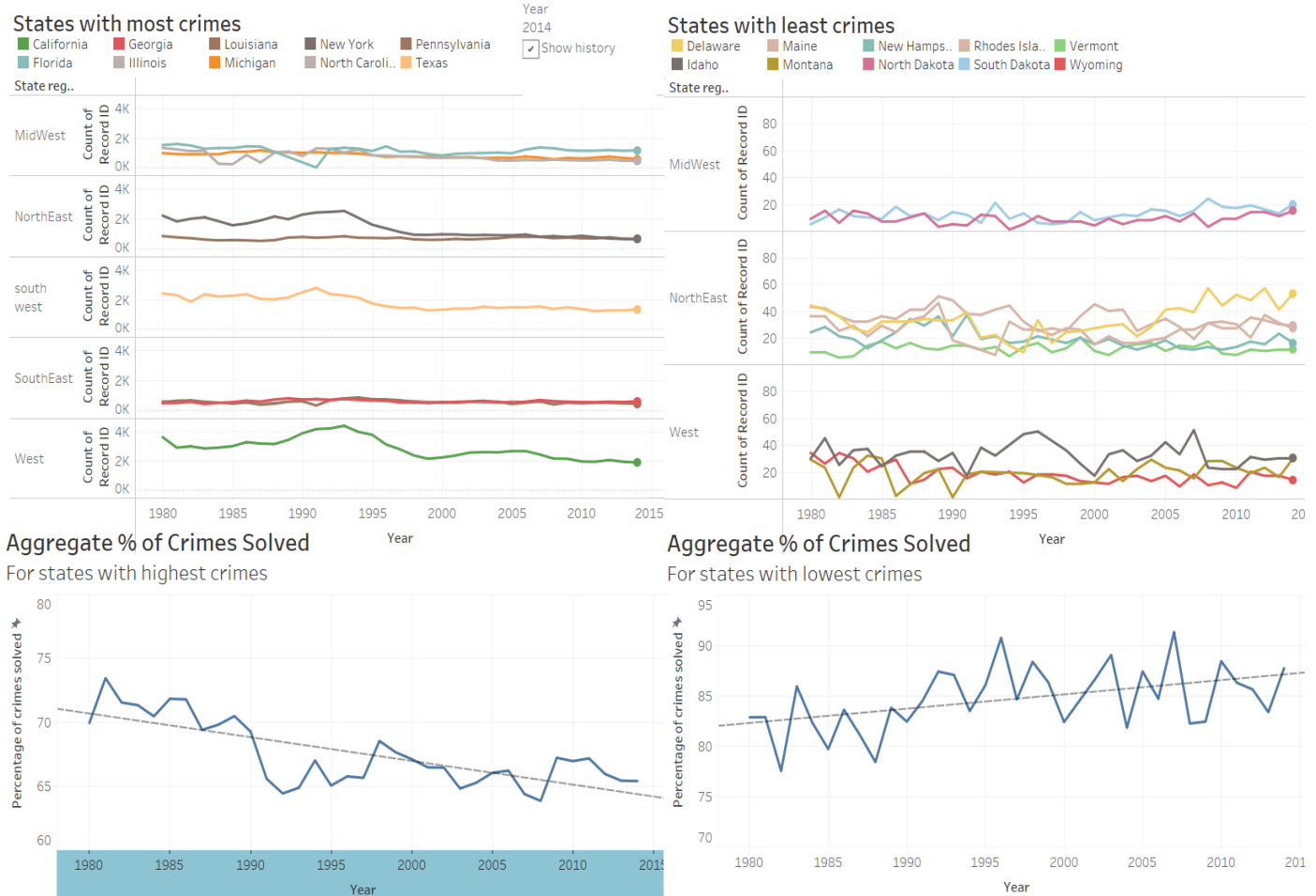
In addition to that, the team will have an opportunity to present their interactive and exploratory dashboard in the classroom for about 10-15 minutes.

**General view of two dashboards:**

Dashboard 1:
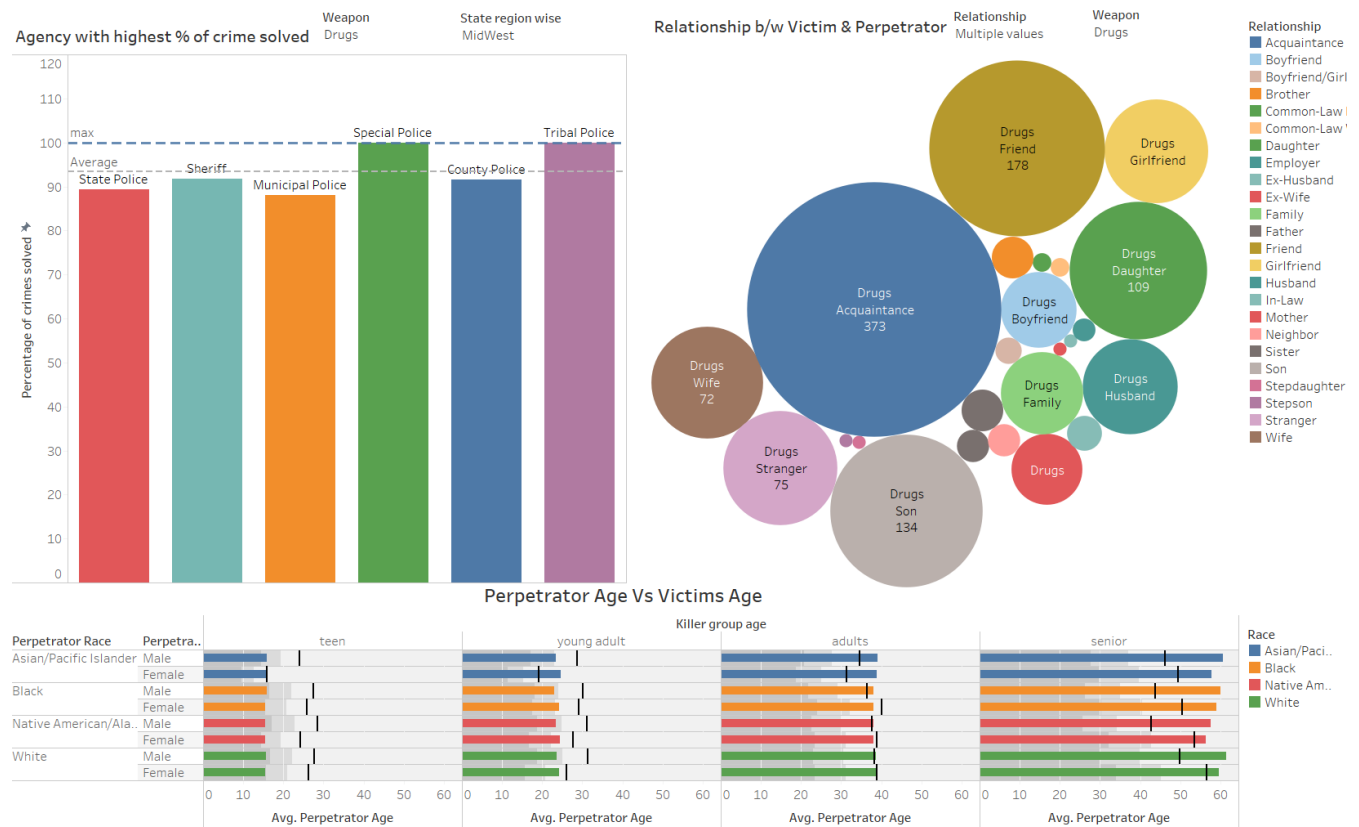
Dashboard 2: Trend over 30 years (1980-2014):

1. Visualization 1:Bar Graph Side-by-Side for agencies with highest % of crimes solved

The purpose of this graph is to visualize crimes solved by various agencies based on the weapons used for committing the crime. The end-user of this visualization can be the government agency that has the responsibility to allocate crimes to an agency. After knowing which weapon was used to commit the crime in a particular region, they can see different agencies and their percentage of successful crimes solved. With this data, they can assign the criminal case to the agency which has the highest crime solve ratio in that region and for that particular weapon.
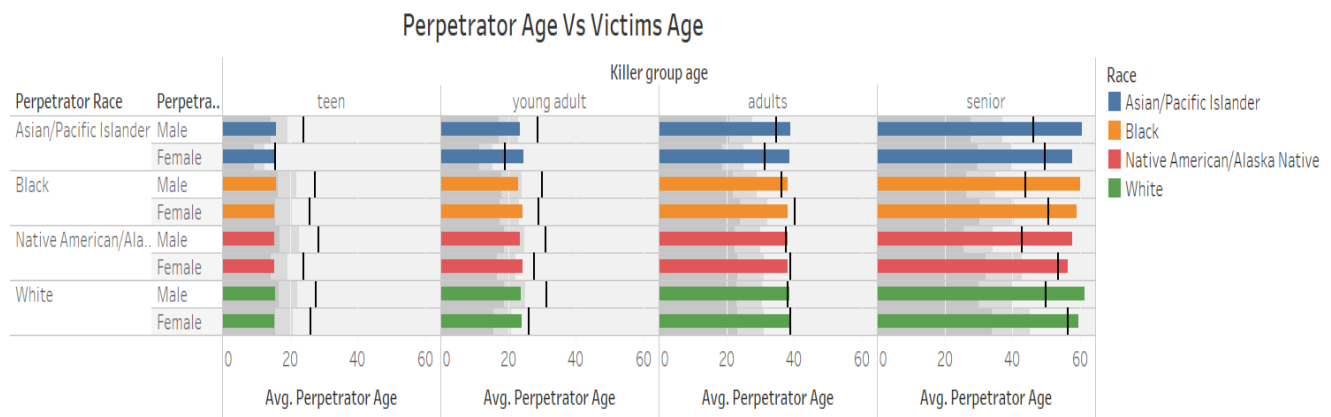
Agency with highest % of crime solved
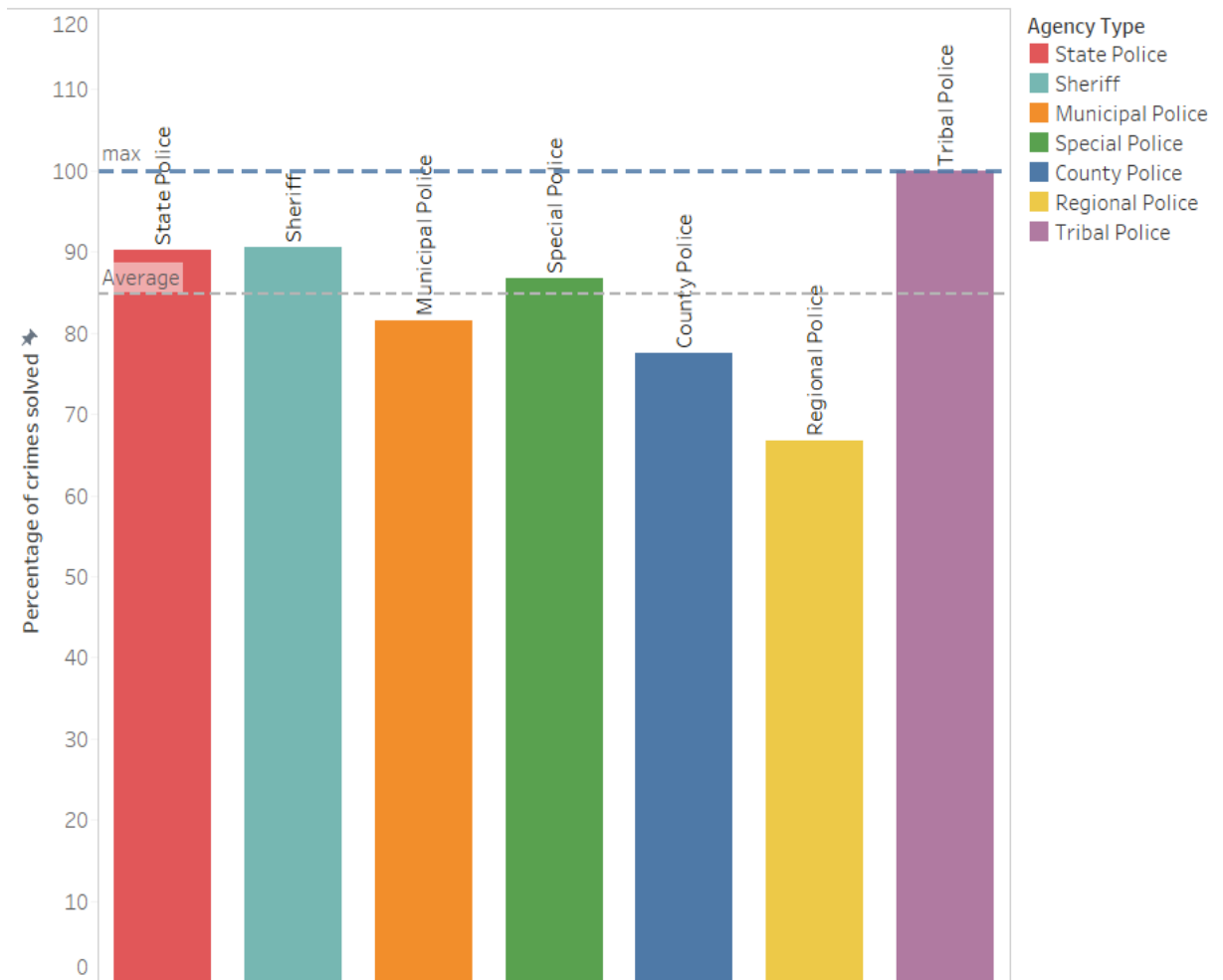
Percentage of crimes solved for each Agency Type broken down by Weapon. Color shows details about Agency Type. The marks are labeled by Agency Type. The data is filtered on State region wise, which keeps MidWest. The view is filtered on Weapon, which keeps Shotgun.

For example- When they know the crime is committed by the use of a shotgun and it took place in the Midwest Region of the United States. They would be able to filter accordingly and then observe that Tribal Police has the highest crime-solving success rate when shotgun crimes are handed over to them in the Midwest region. The government agencies can also see the average percentage of crimes solved for a particular region and a particular weapon. By doing so, they will be able to increase the crime-solving percentage of each region over the years, as they would be repeatedly handing over the crimes to agencies who have a high crime-solving rate.

2. Visualization 2: Bullet Graph for Perpetrator Age Vs Victims Age

The purpose of this visualization is to find out the relationship between the perpetrator's age and the victim's age based on their gender and race. After creating few calculated fields we calculate the average age of perpetrators and group them according to different age groups (teens 9-18, Young adults 18-35, adults- 35-60, senior 60-100 years). Based on these perpetrators' age groups the length of bars in the chart represent the average age of groups from a particular race of a particular gender and the black line represents an average of victims age killed by that category of the perpetrator. Through this, we can observe if an age group of a race and gender commits a crime on people younger or older to them.

Average of Perpetrator Age for each Perpetrator Sex broken down by Killer group age vs. Perpetrator Race. Color shows details about Perpetrator Race. The data is filtered on True age, which keeps True. The view is filtered on Perpetrator Race and Perpetrator Sex. The Perpetrator Race filter keeps Asian/Pacific Islander, Black, Native American/Alaska Native and White. The Perpetrator Sex filter keeps Female and Male.

For example, we can observe that in general criminals who are teens tend to commit crimes on people who are way older than them, whereas young adult and adult age group criminals tend to kill people similar to their age group, and senior-aged criminals usually commit crimes on young people. Through this we can also observe trends by gender and race, for instance, we can see that all age groups of Asian females criminals commit crimes on people younger than them. Taking another example, we can see that senior male American criminals tend to commit a crime on much younger citizens (with an average age difference of 16 years between perpetrator and victim).