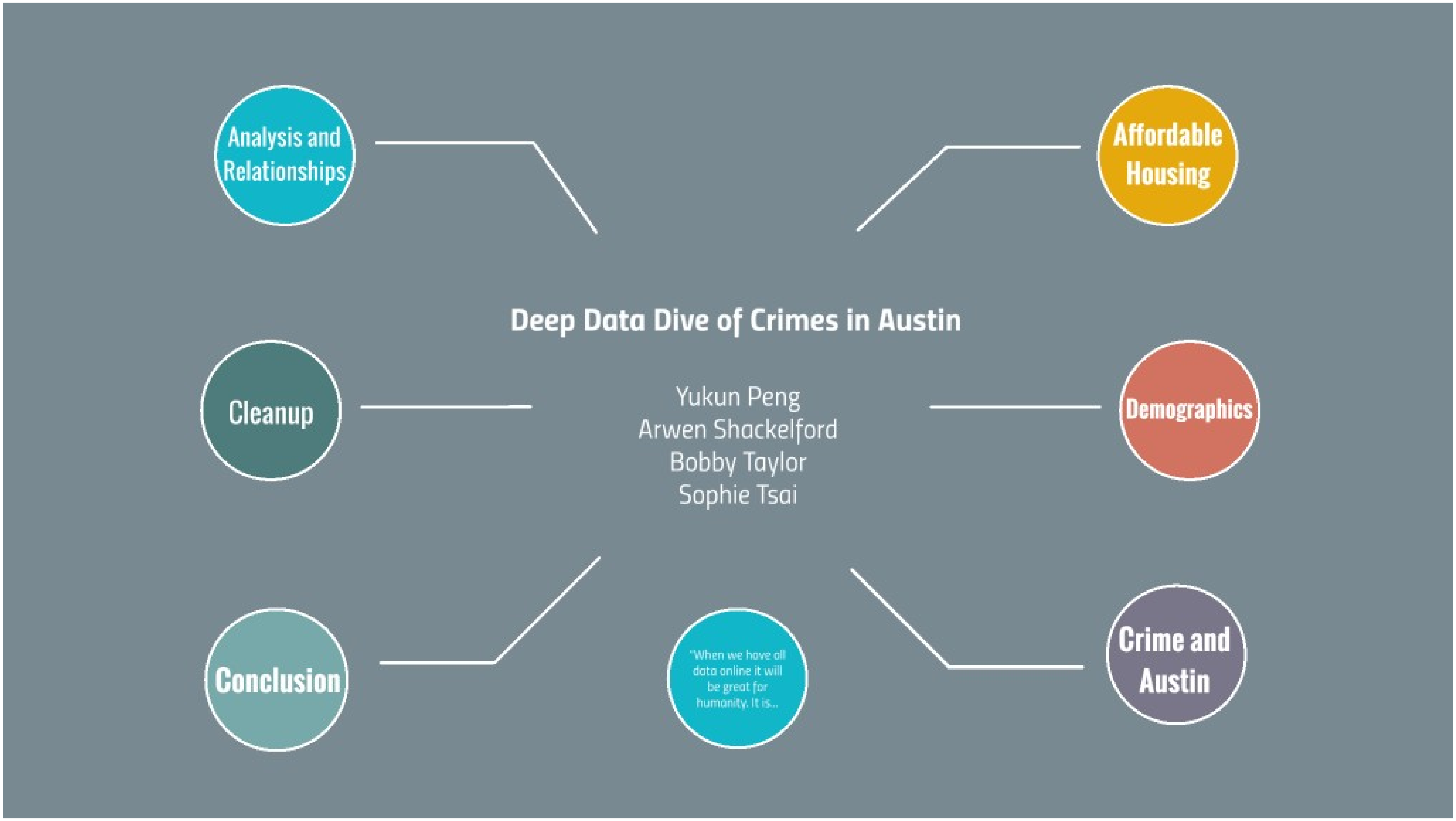
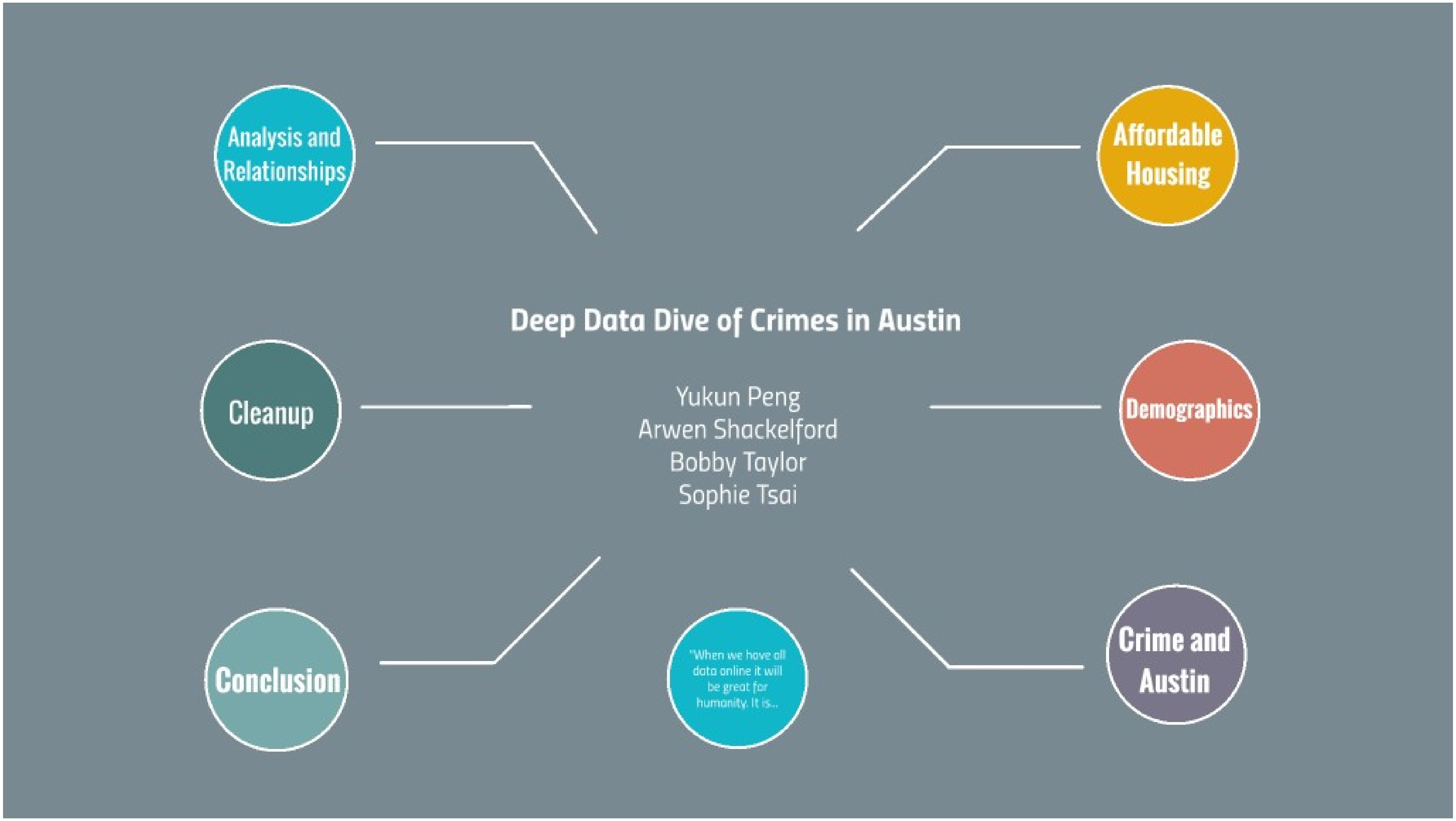
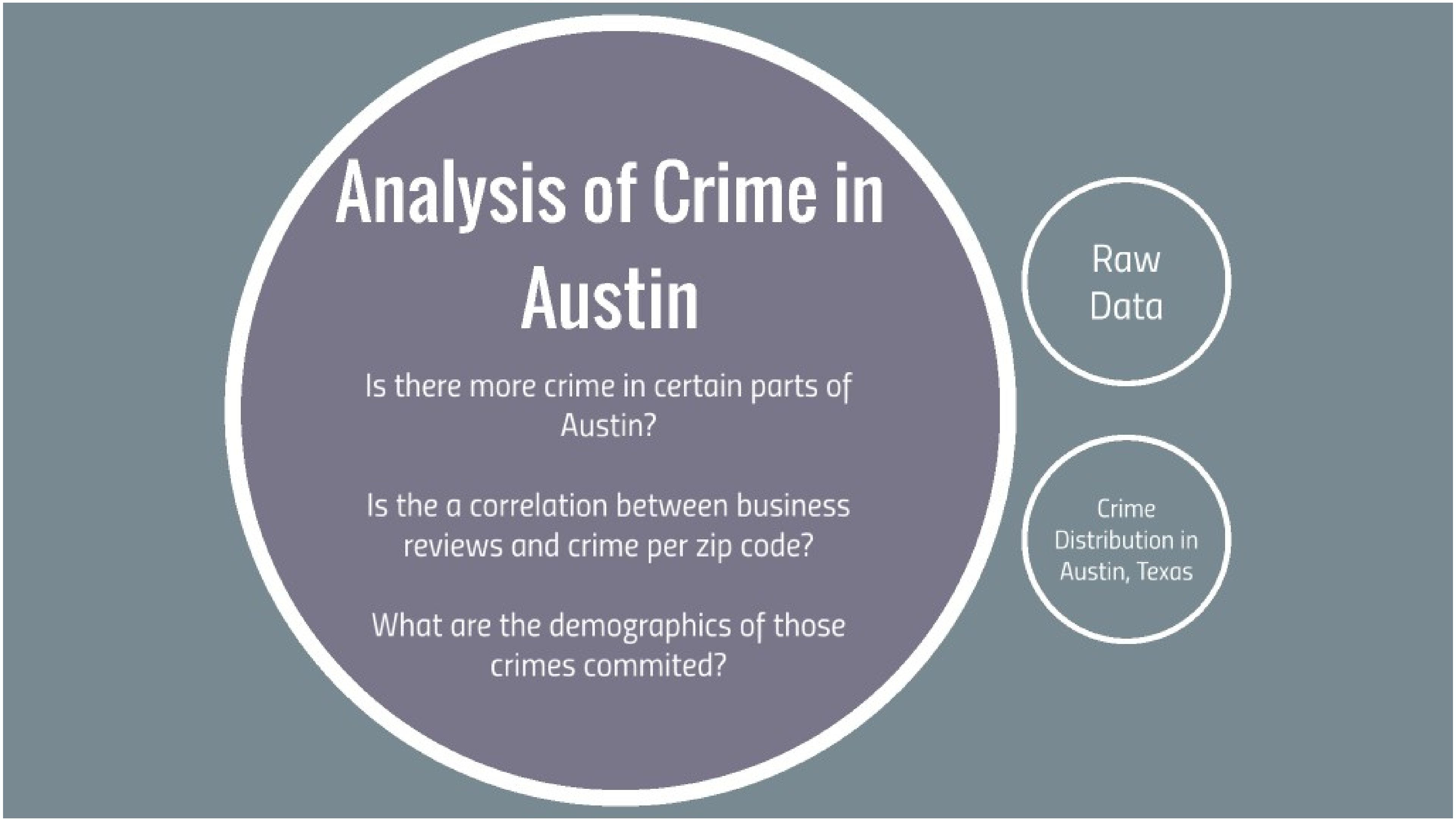

Analysis and Relationships
Affordable Housing
Deep Data Dive of Crimes in Austin
Cleanup
Yukun Peng
Arwen Shackelford
Bobby Taylor
Sophie Tsai
Demographics
Conclusion
"When we have all data online it will be great for humanity. It is...
Crime and Austin

"When we have all data online it will be great for humanity. It is a prerequisite to solving many problems that humankind faces." – Robert Cailliau, Belgian informatics engineer and computer scientist who, together with Tim Berners-Lee, developed the World Wide Web.


Deep Data Dive of Crimes in Austin
Yukun Peng
Arwen Shackelford
Bobby Taylor
Sophie Tsai
Analysis and Relationships
Affordable Housing
Cleanup
Demographics
Conclusion
"When we have all data online it will be great for humanity. It is...
Crime and Austin

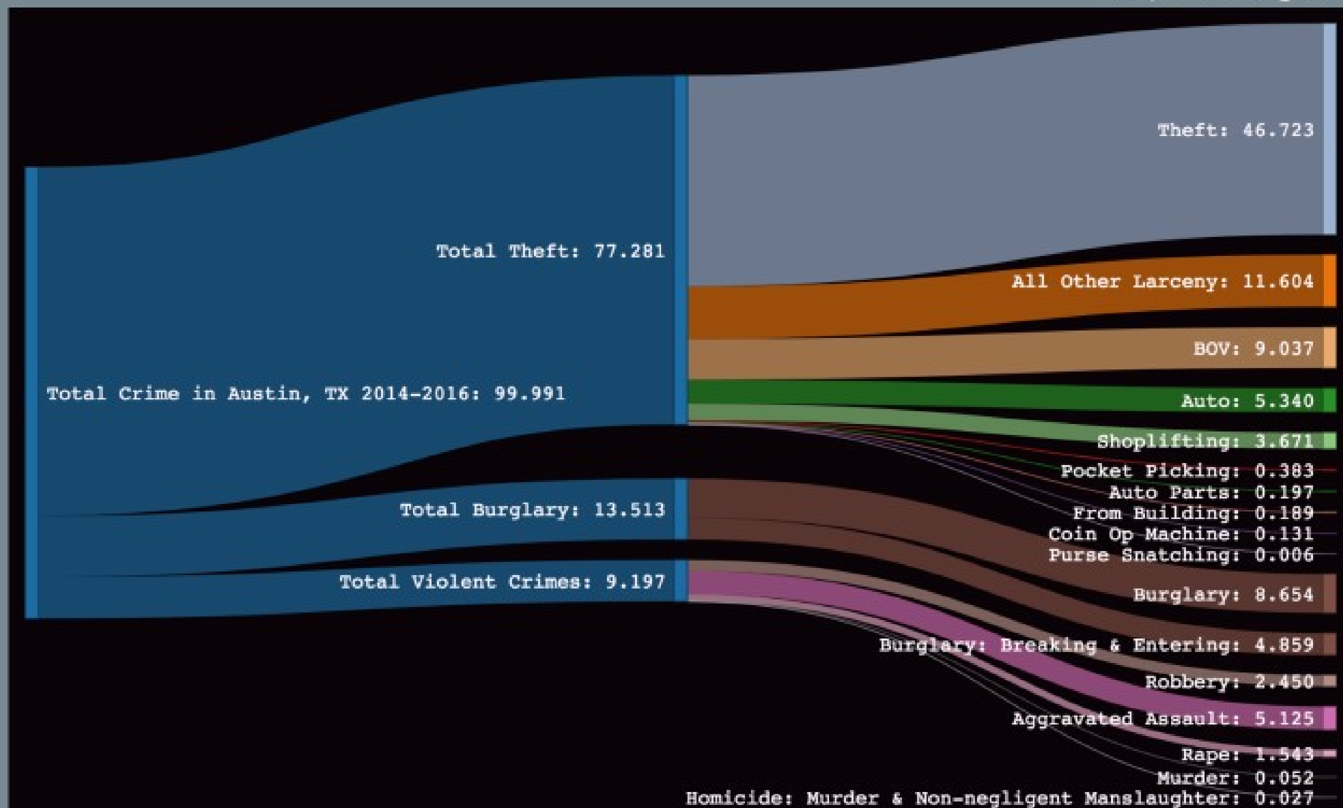

Analysis of Crime in Austin
Is there more crime in certain parts of Austin?
Is the a correlation between business reviews and crime per zip code?
What are the demographics of those crimes commited?
Raw Data
Crime Distribution in Austin, Texas

Raw Data

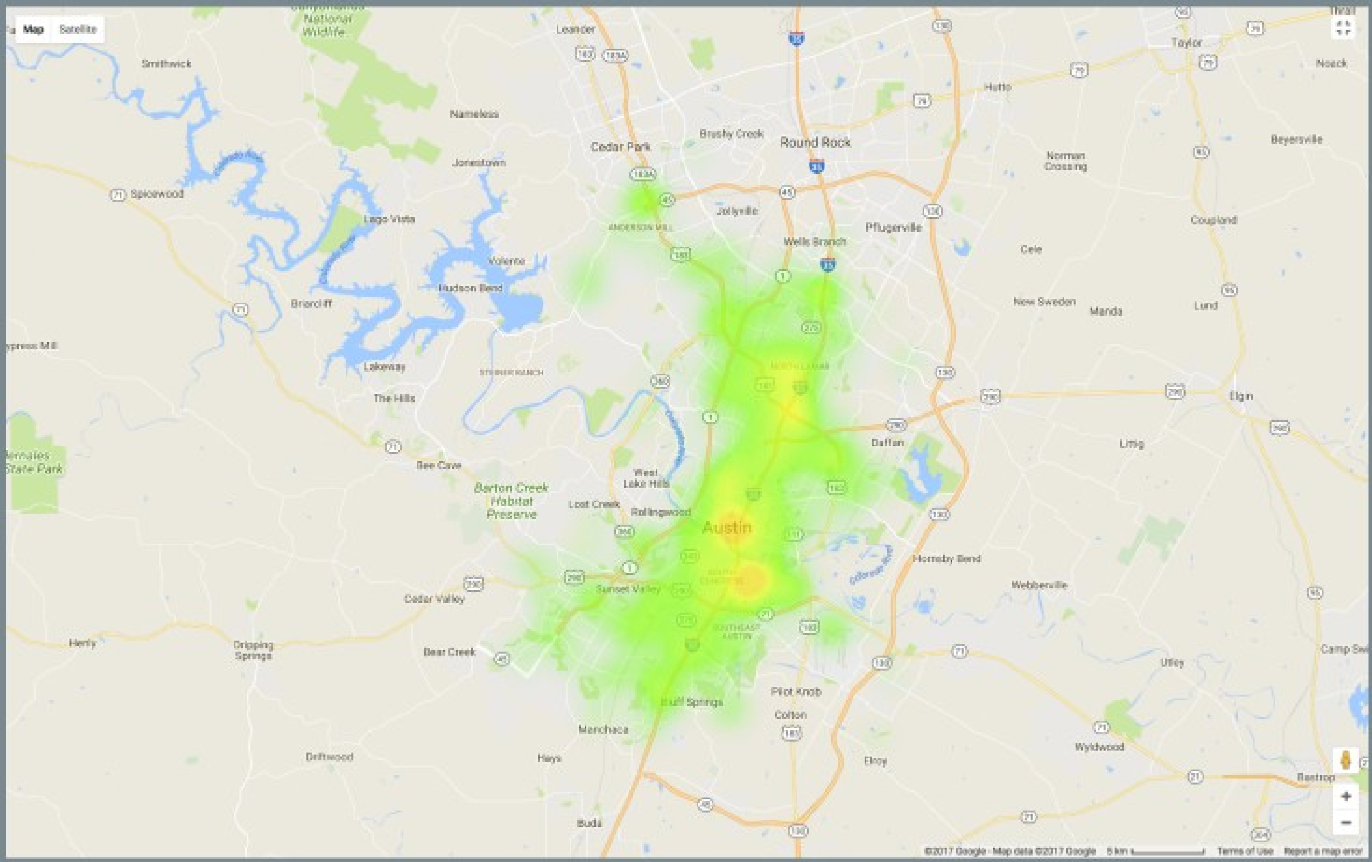
- 2014-2016 annual crime data retrieved from data.austintexas.gov
- Local business data retrieved from Yelp API
- Crime demographics retrieved from data.austintexas.gov
- Affordable Housing Project data from data.austintexas.gov

Crime Distribution in Austin, Texas

(in percentages)

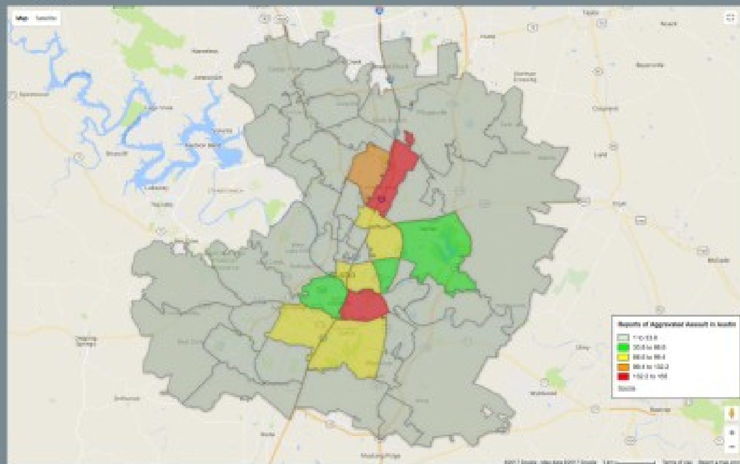


Total Crime Distribution in Austin

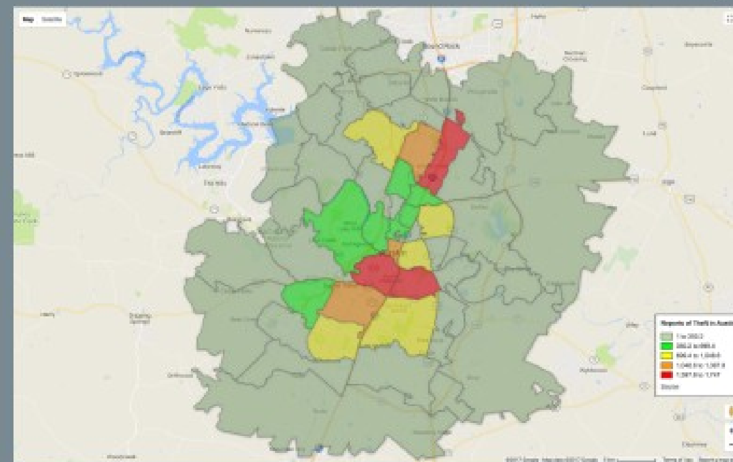


Crime Distribution by Type

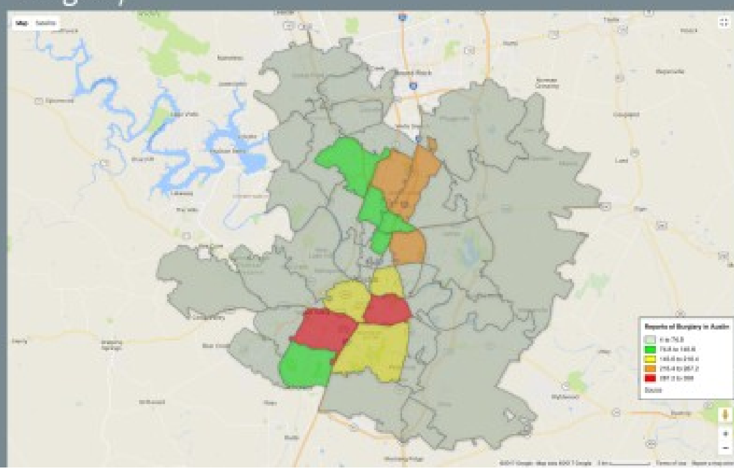
aggravated assault



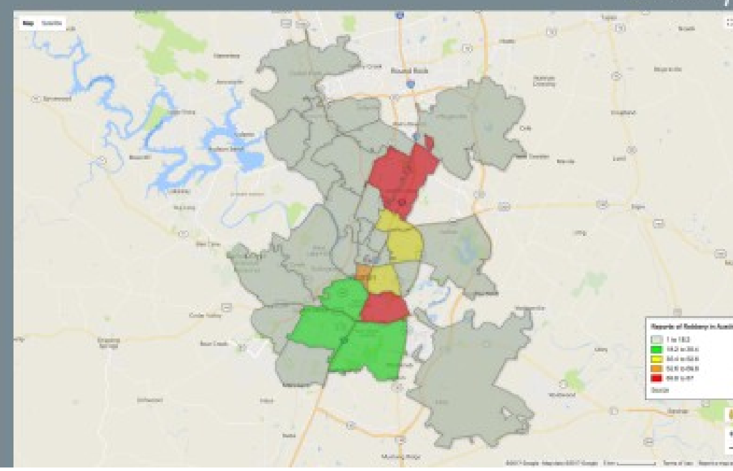
theft



burglary

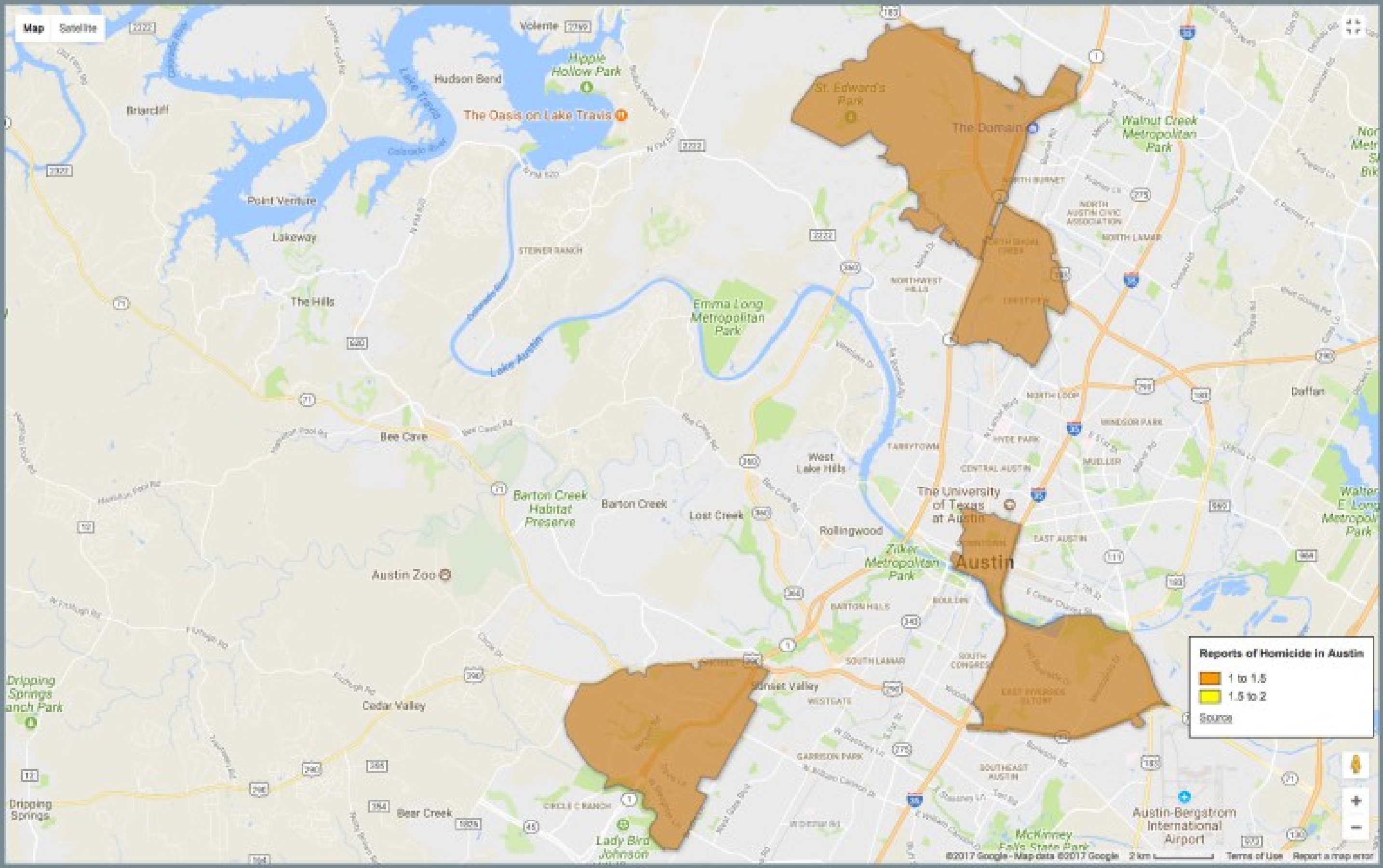


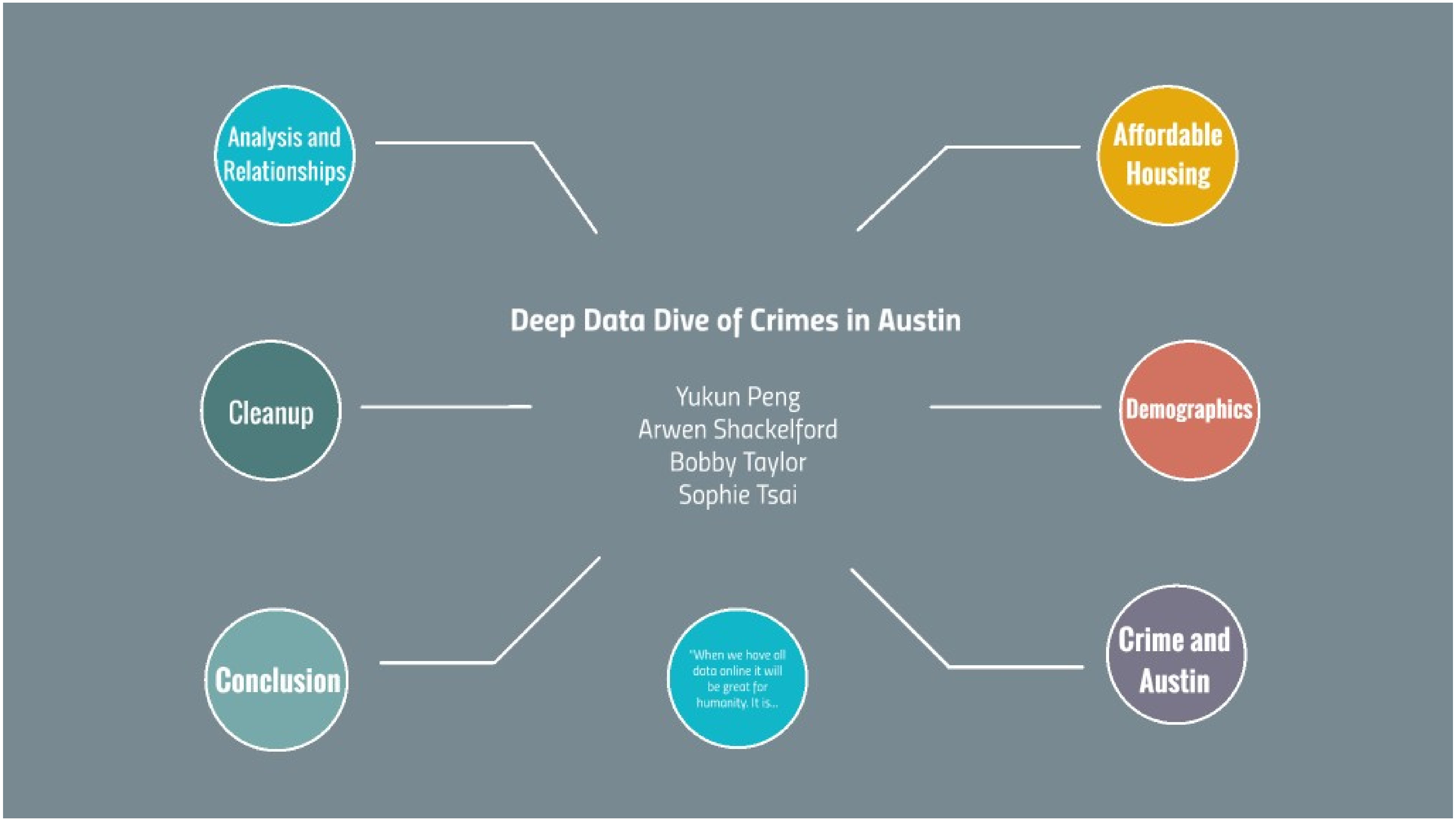
robbery



Crime Distribution by Type

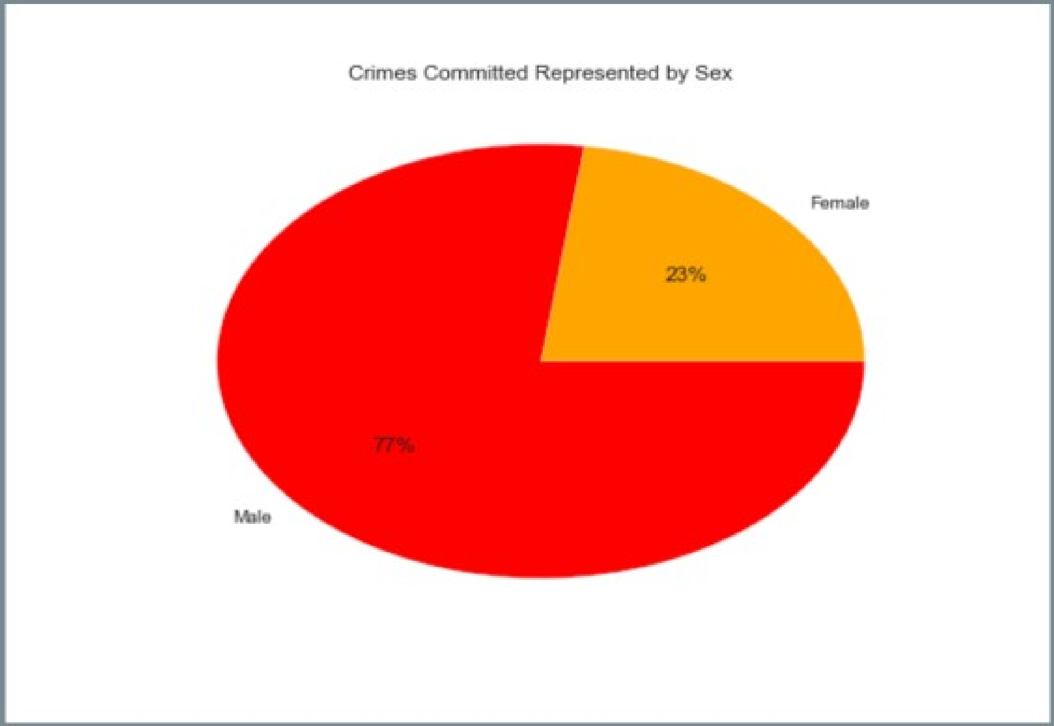
homicide




Analysis and Relationships
Affordable Housing
Deep Data Dive of Crimes in Austin
Cleanup
Yukun Peng
Arwen Shackelford
Bobby Taylor
Sophie Tsai
Demographics
Conclusion
"When we have all data online it will be great for humanity. It is...
Crime and Austin

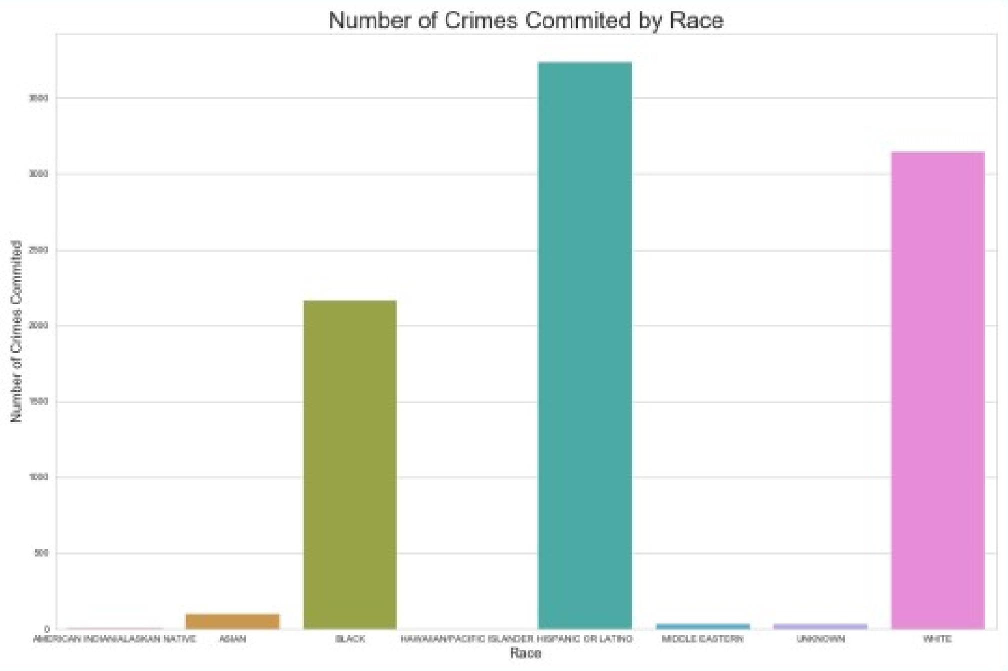

What is behind Austin's crime?
Gender Demographics
Racial Demograhpics

Gender Demographics

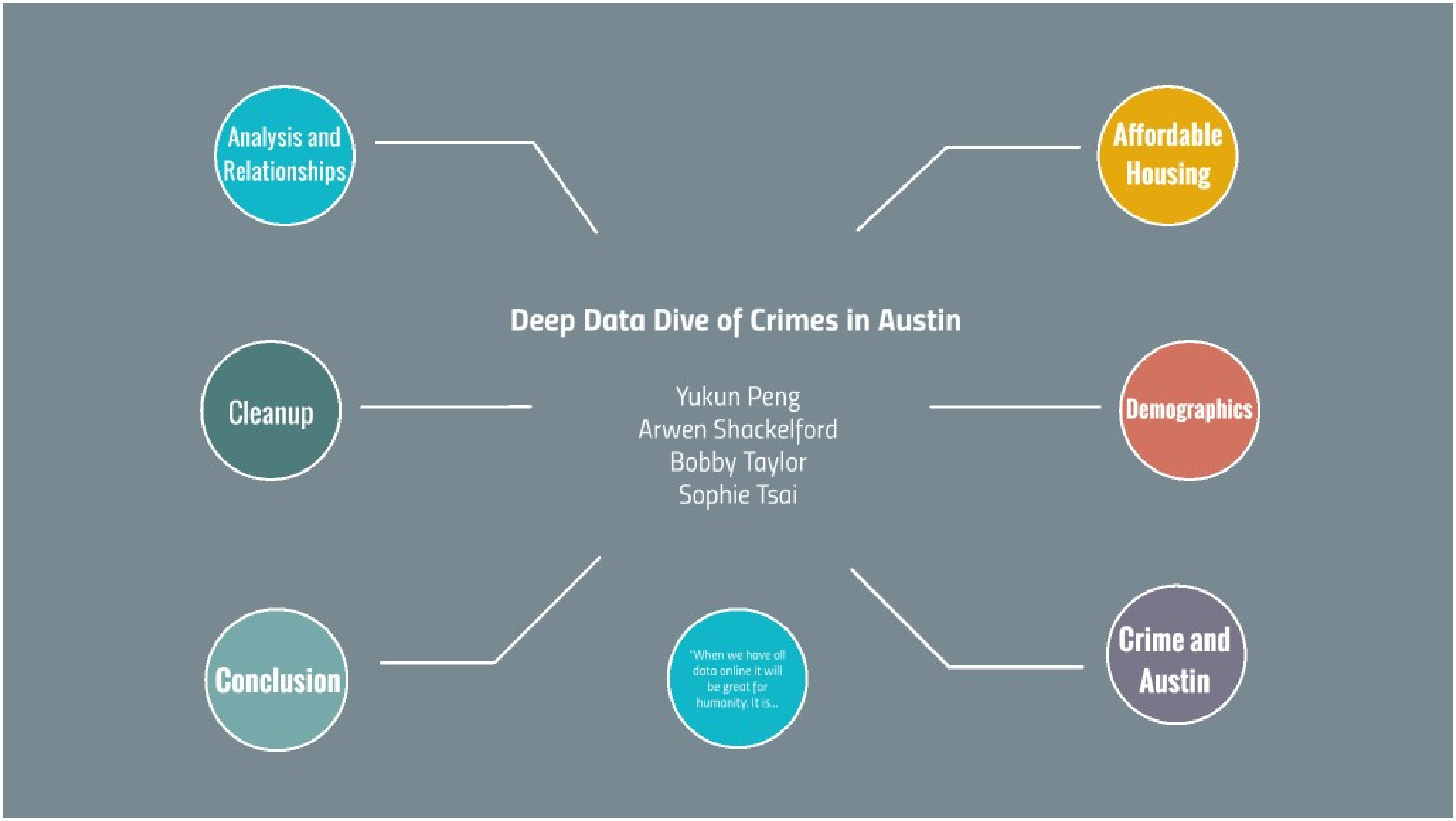


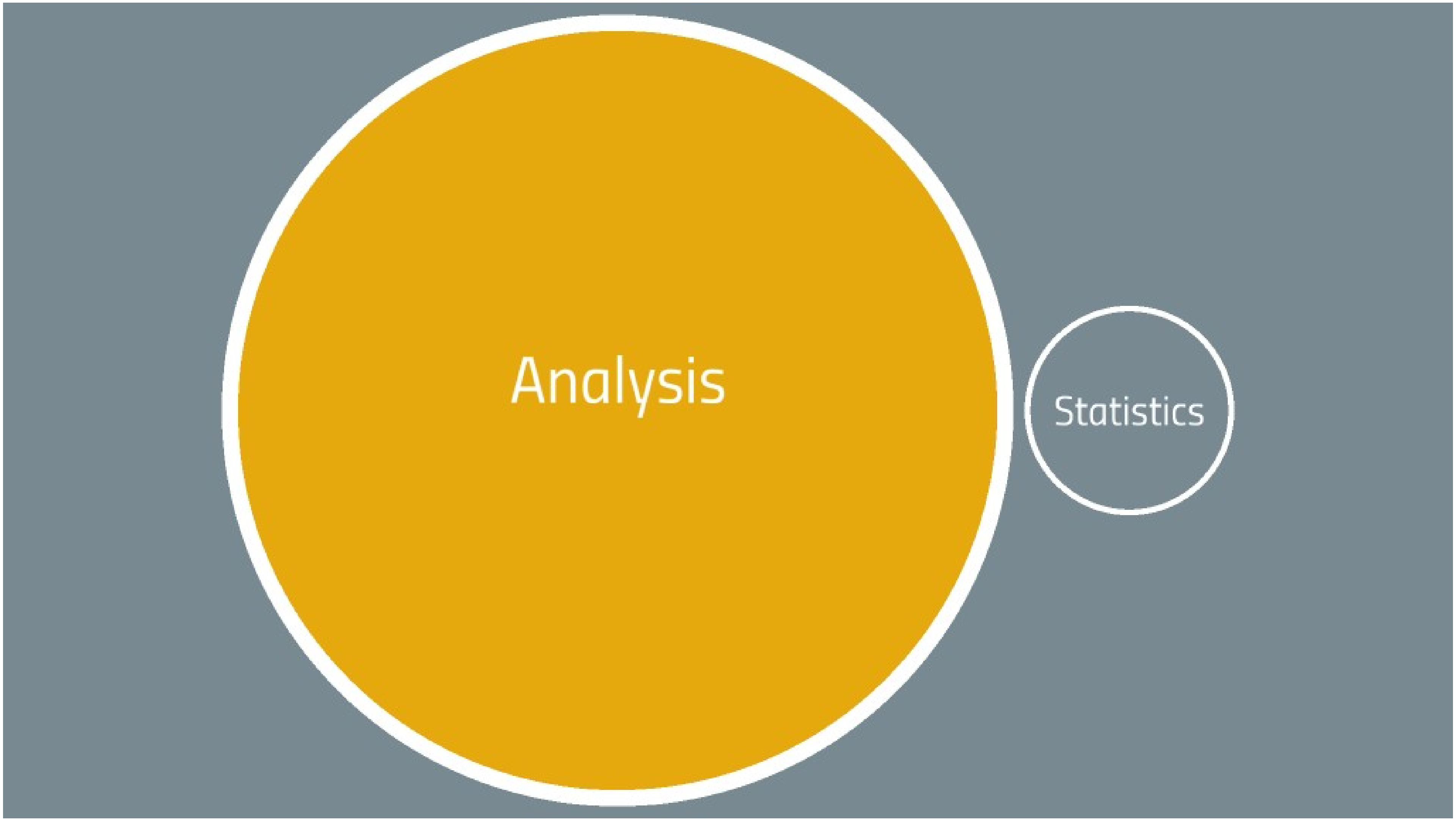
Sex	Crimes Committed	Percent of Crimes Commited
F	2105	22.858074
M	7104	77.141926

Racial Demograhpics

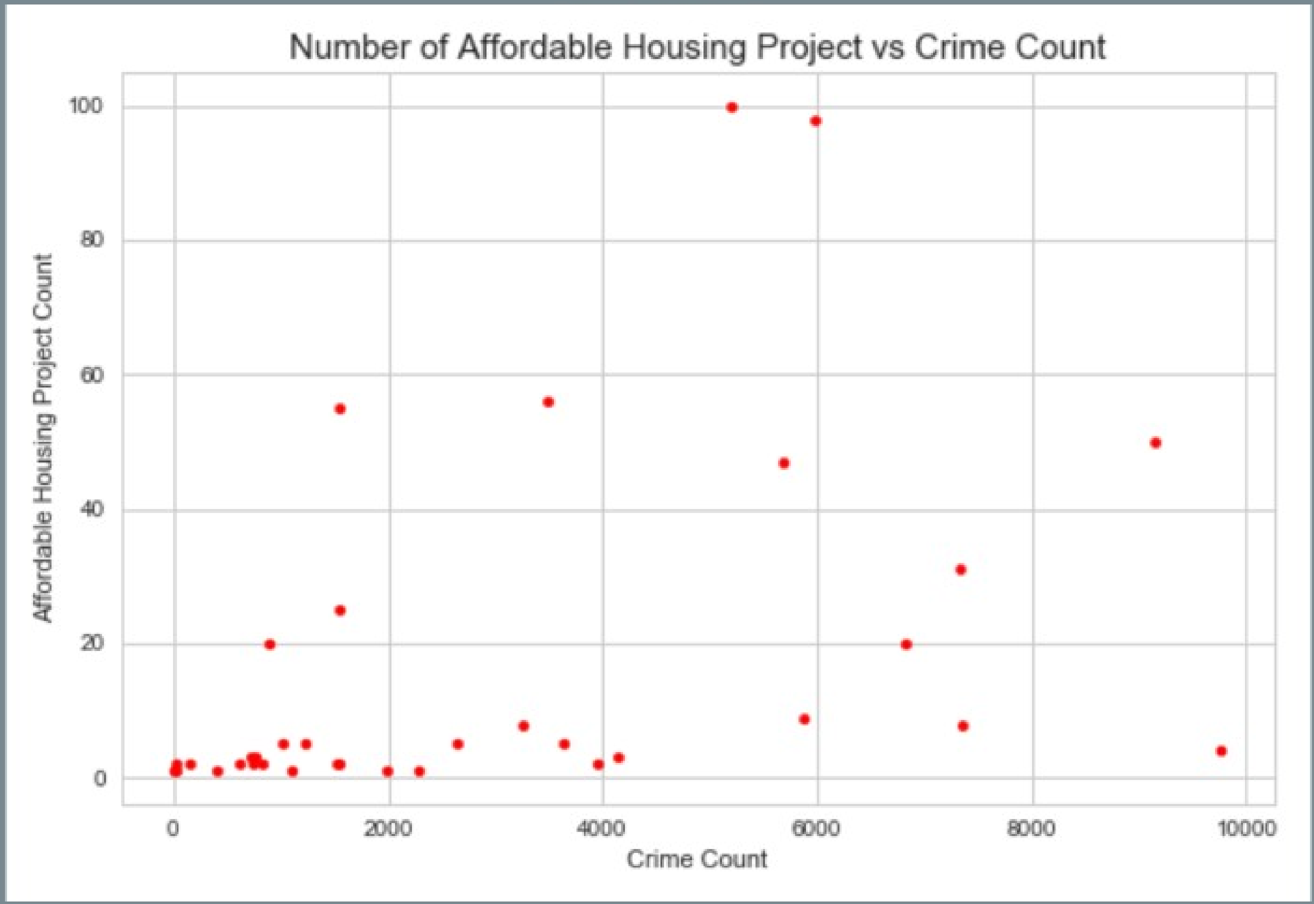


Race	Crimes Committed	Percent of Crimes Commited
AMERICAN INDIAN/ALASKAN NATIVE	5	0.054295
ASIAN	99	1.075035
BLACK	2164	23.498751
HAWAIIAN/PACIFIC ISLANDER	2	0.021718
HISPANIC OR LATINO	3735	40.558150
MIDDLE EASTERN	28	0.304050
UNKNOWN	29	0.314909
WHITE	3147	34.173092


Analysis and Relationships
Affordable Housing
Deep Data Dive of Crimes in Austin
Yukun Peng
Arwen Shackelford
Bobby Taylor
Sophie Tsai
Cleanup
Demographics
Conclusion
"When we have all data online it will be great for humanity. It is...
Crime and Austin

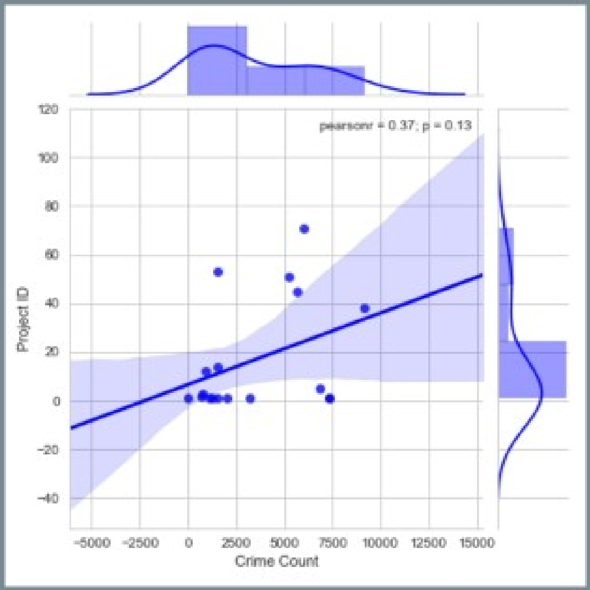

Analysis
Statistics

Statistics

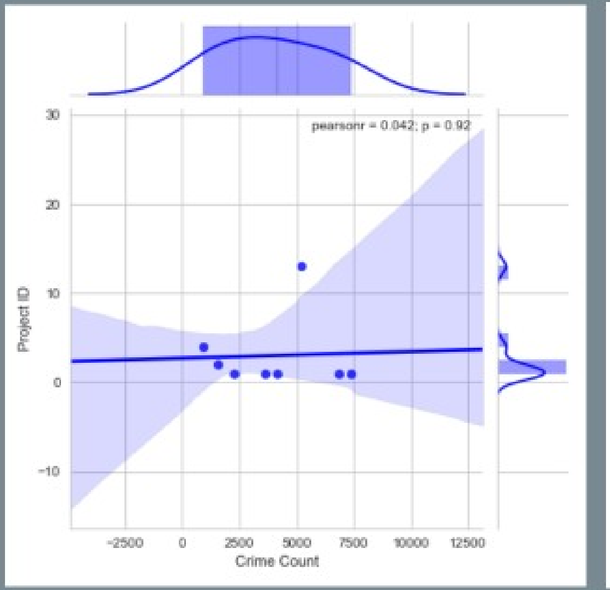


Pearson test results: (0.43411929023708457, 0.0091705761752694619)

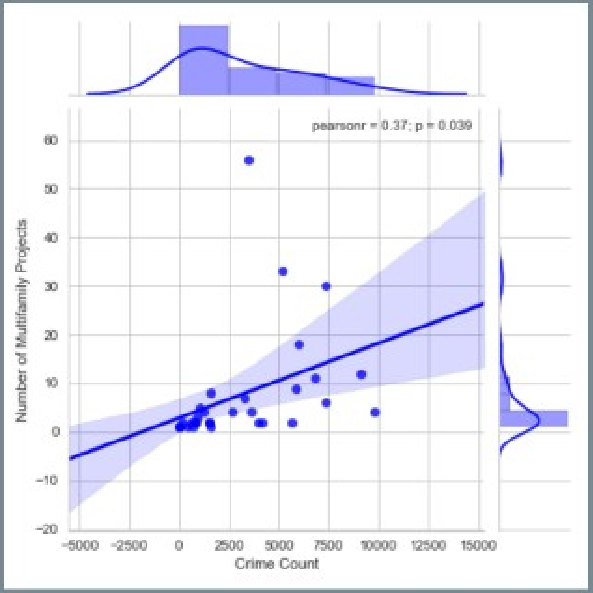
crime vs single family project

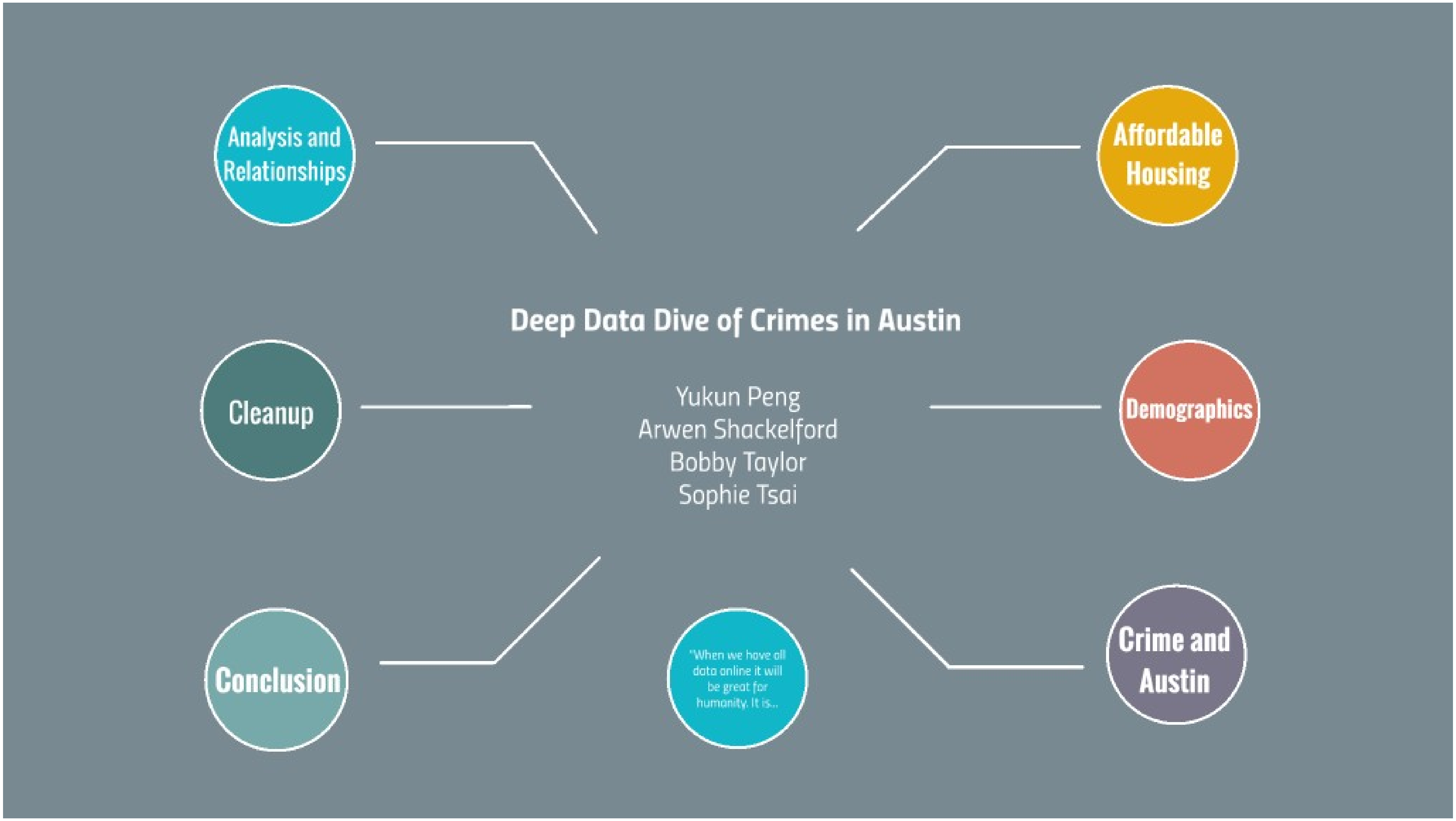


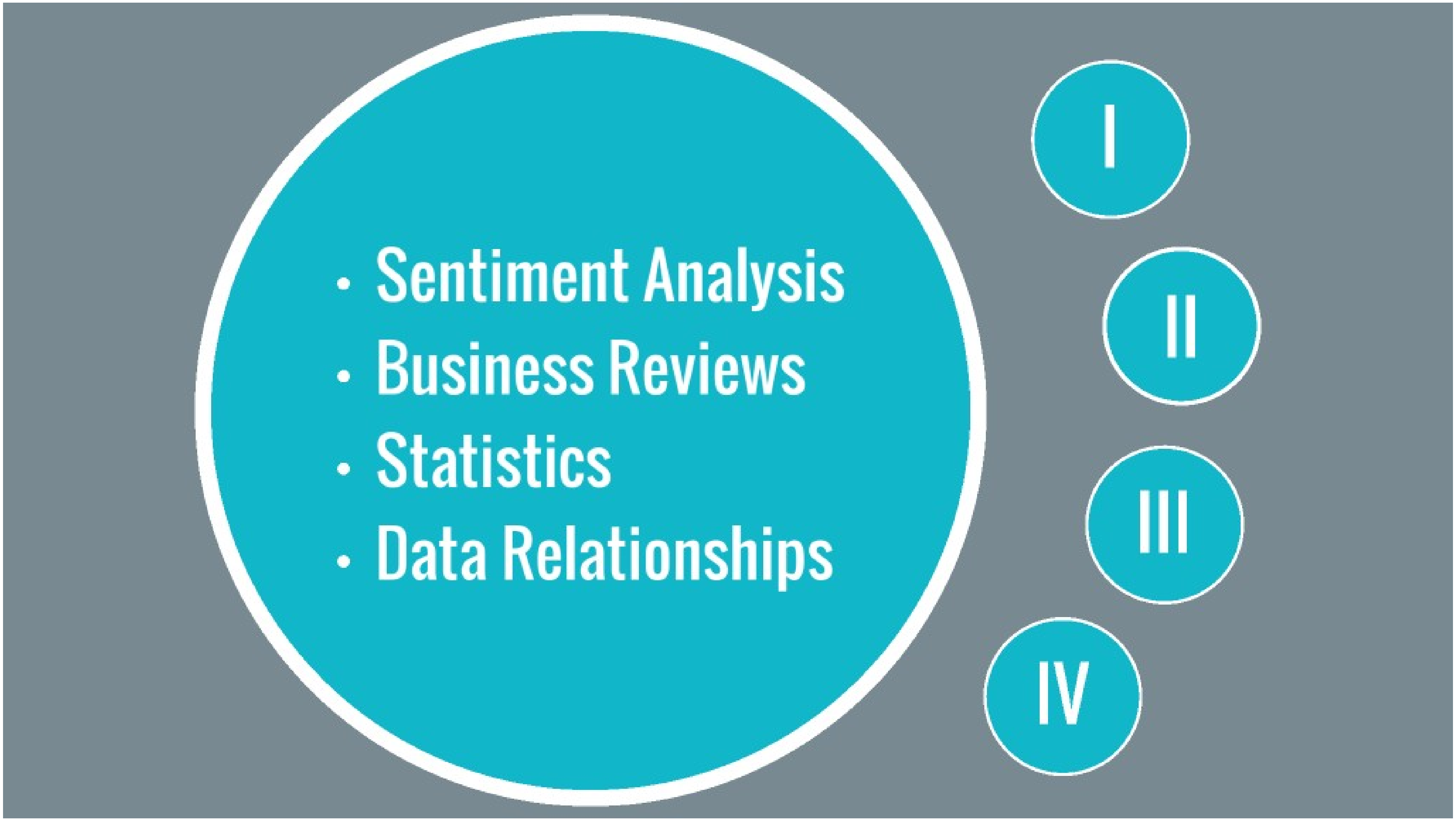
crime vs duplex

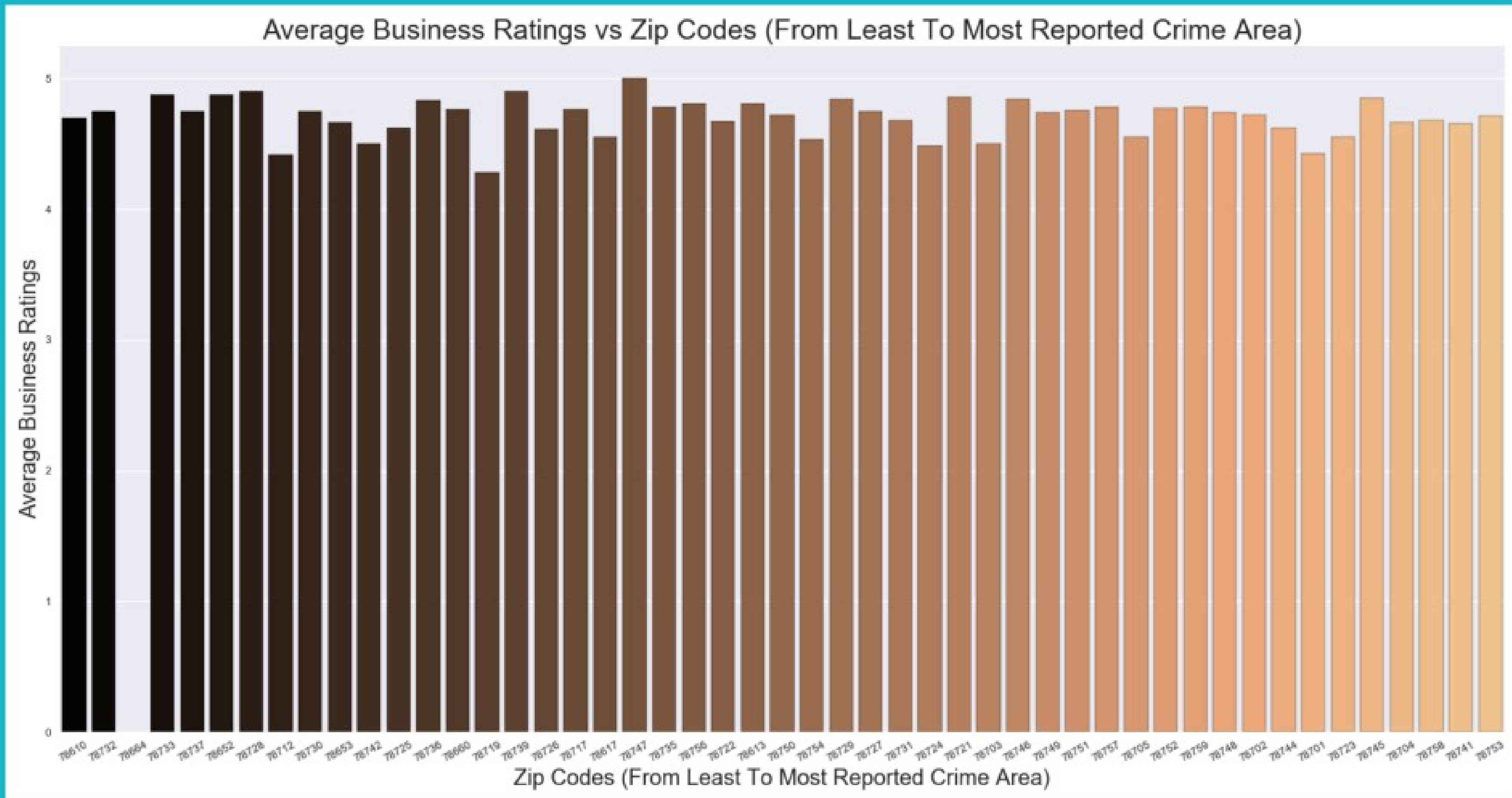


crime vs multifamily project

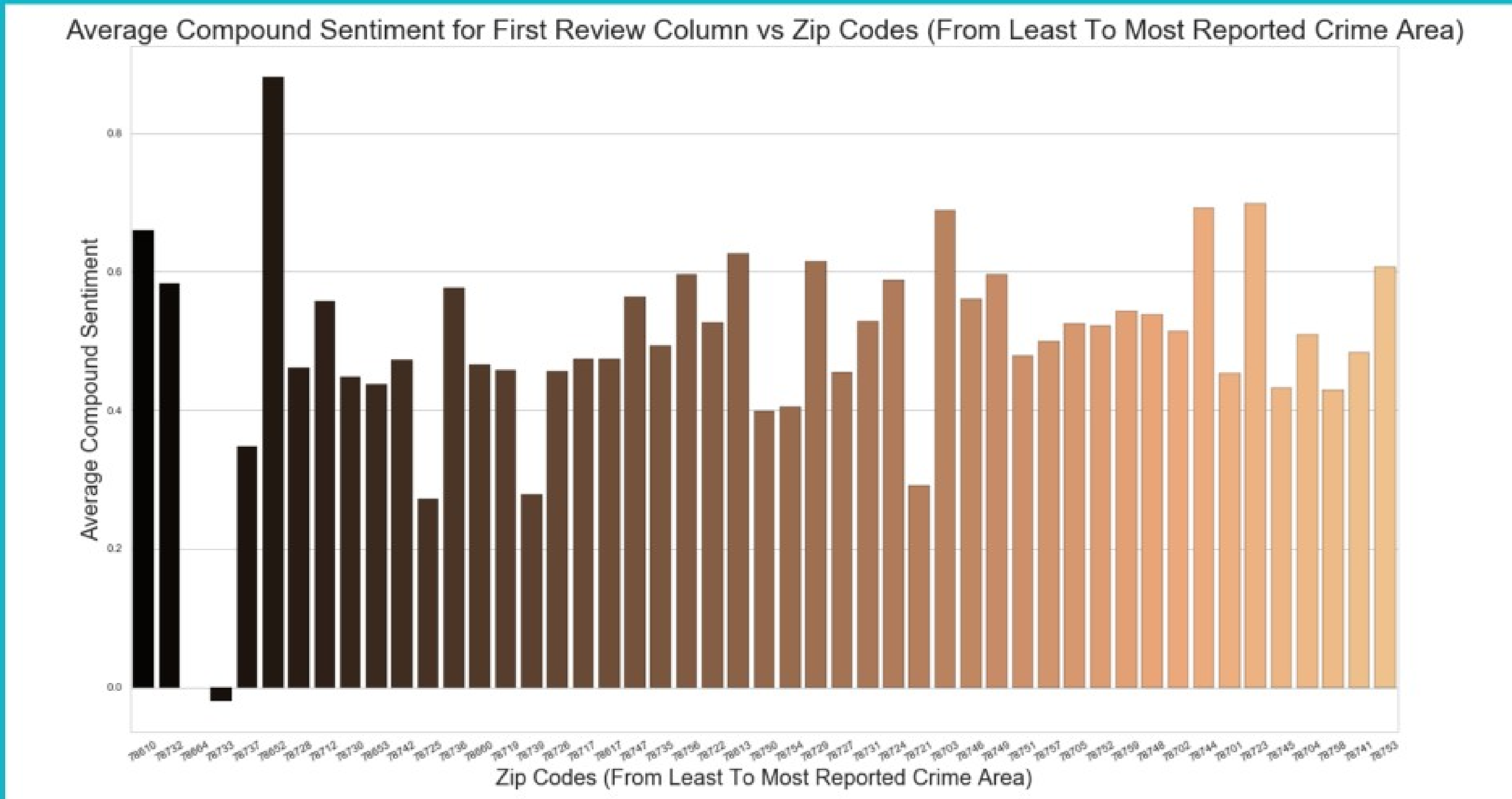



Deep Data Dive of Crimes in Austin
Yukun Peng
Arwen Shackelford
Bobby Taylor
Sophie Tsai
Analysis and Relationships
Affordable Housing
Cleanup
Demographics
Conclusion
"When we have all data online it will be great for humanity. It is...
Crime and Austin

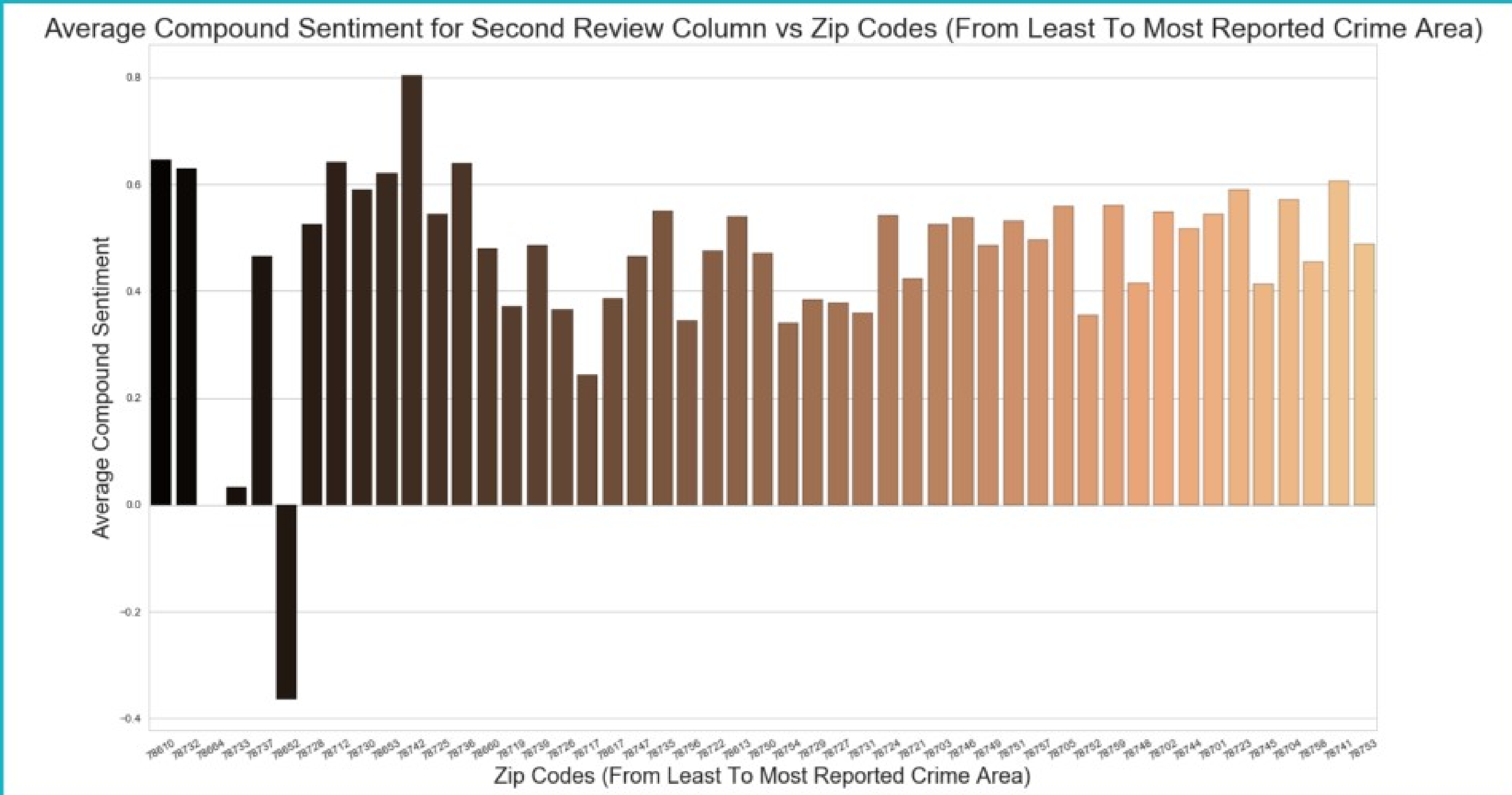

• Sentiment Analysis
• Business Reviews
• Statistics
• Data Relationships
I
II
III
IV


Average Business Ratings vs Zip Codes (From Least To Most Reported Crime Area)
Average Business Ratings
0
1
2
3
4
5
78610
78732
78664
78733
78737
78652
78728
78712
78730
78653
78742
78725
78736
78660
78719
78739
78726
78717
78617
78747
78735
78756
78722
78613
78750
78754
78729
78727
78731
78724
78721
78703
78746
78749
78751
78757
78705
78752
78759
78748
78702
78744
78701
78723
78745
78704
78758
78741
78753
Zip Codes (From Least To Most Reported Crime Area)

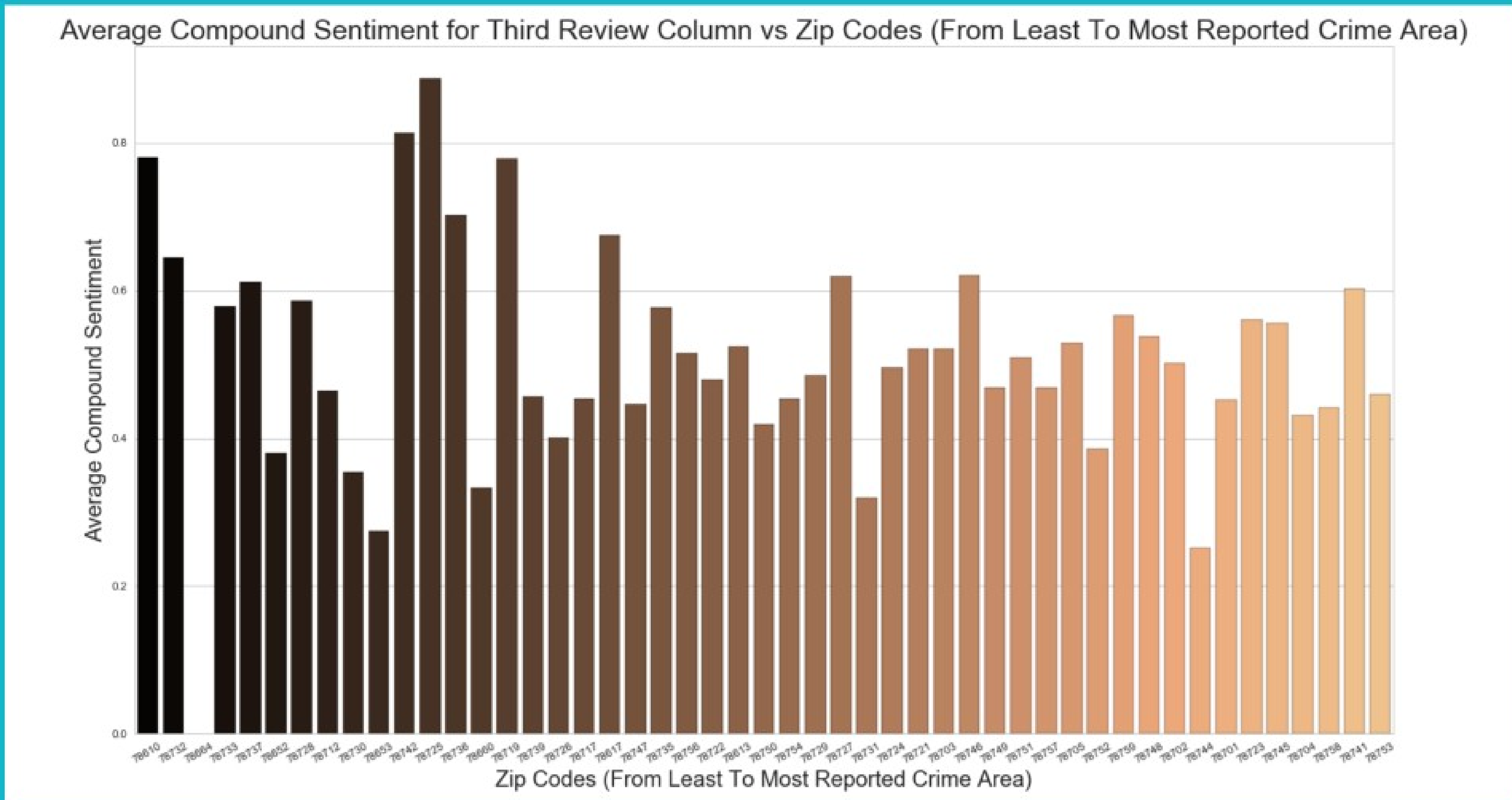
For the first review column, the zip code in area with low reported crimes has a higher than normal compound sentiment rating


Average Compound Sentiment for First Review Column vs Zip Codes (From Least To Most Reported Crime Area)
Average Compound Sentiment
0.0
0.2
0.4
0.6
0.8
78610 78732 78664 78733 78737 78652 78728 78712 78730 78653 78742 78725 78736 78660 78719 78739 78726 78717 78617 78747 78735 78756 78722 78613 78750 78754 78729 78727 78731 78724 78721 78703 78746 78749 78751 78757 78705 78752 78759 78748 78702 78744 78701 78723 78745 78704 78758 78741 78753
Zip Codes (From Least To Most Reported Crime Area)

Second review column actually has a negative compound sentiment for an area with low reported crimes

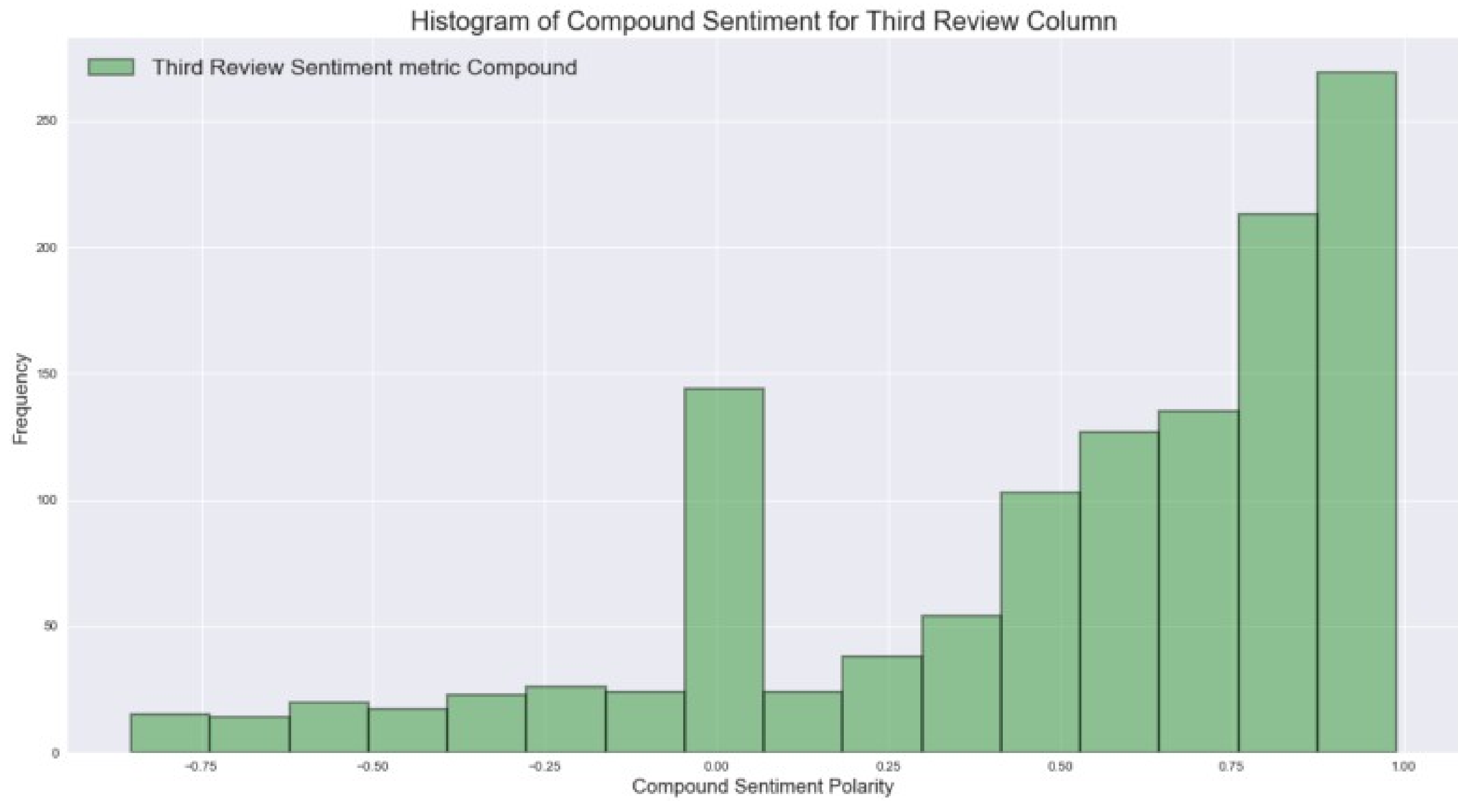

Average Compound Sentiment for Second Review Column vs Zip Codes (From Least To Most Reported Crime Area)
Average Compound Sentiment
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
78610
78732
78664
78733
78737
78652
78728
78712
78730
78653
78742
78725
78736
78660
78719
78739
78726
78717
78617
78747
78735
78756
78722
78613
78750
78754
78729
78727
78731
78724
78721
78703
78746
78749
78751
78757
78705
78752
78759
78748
78702
78744
78701
78723
78745
78704
78758
78741
78753
Zip Codes (From Least To Most Reported Crime Area)

The average compound sentiment tend to be higher for the third review column for reviews in areas with least reported crimes

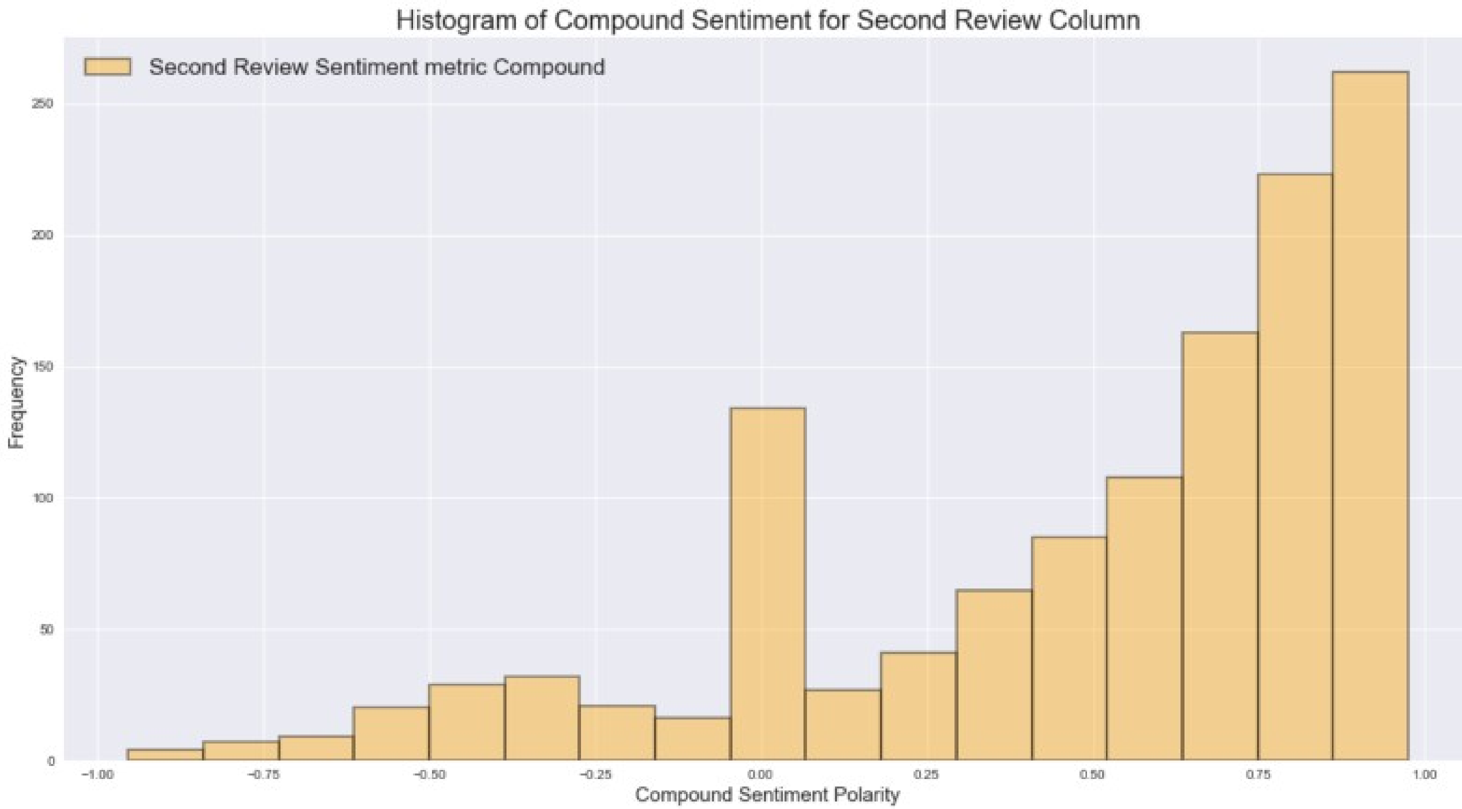


Average Compound Sentiment for Third Review Column vs Zip Codes (From Least To Most Reported Crime Area)
Average Compound Sentiment
0.0
0.2
0.4
0.6
0.8
78610
78732
78664
78733
78737
78652
78728
78712
78730
78653
78742
78725
78736
78660
78719
78739
78726
78717
78617
78747
78735
78756
78722
78613
78750
78754
78729
78727
78731
78724
78721
78703
78746
78749
78751
78757
78705
78752
78759
78748
78702
78744
78701
78723
78745
78704
78758
78741
78753
Zip Codes (From Least To Most Reported Crime Area)

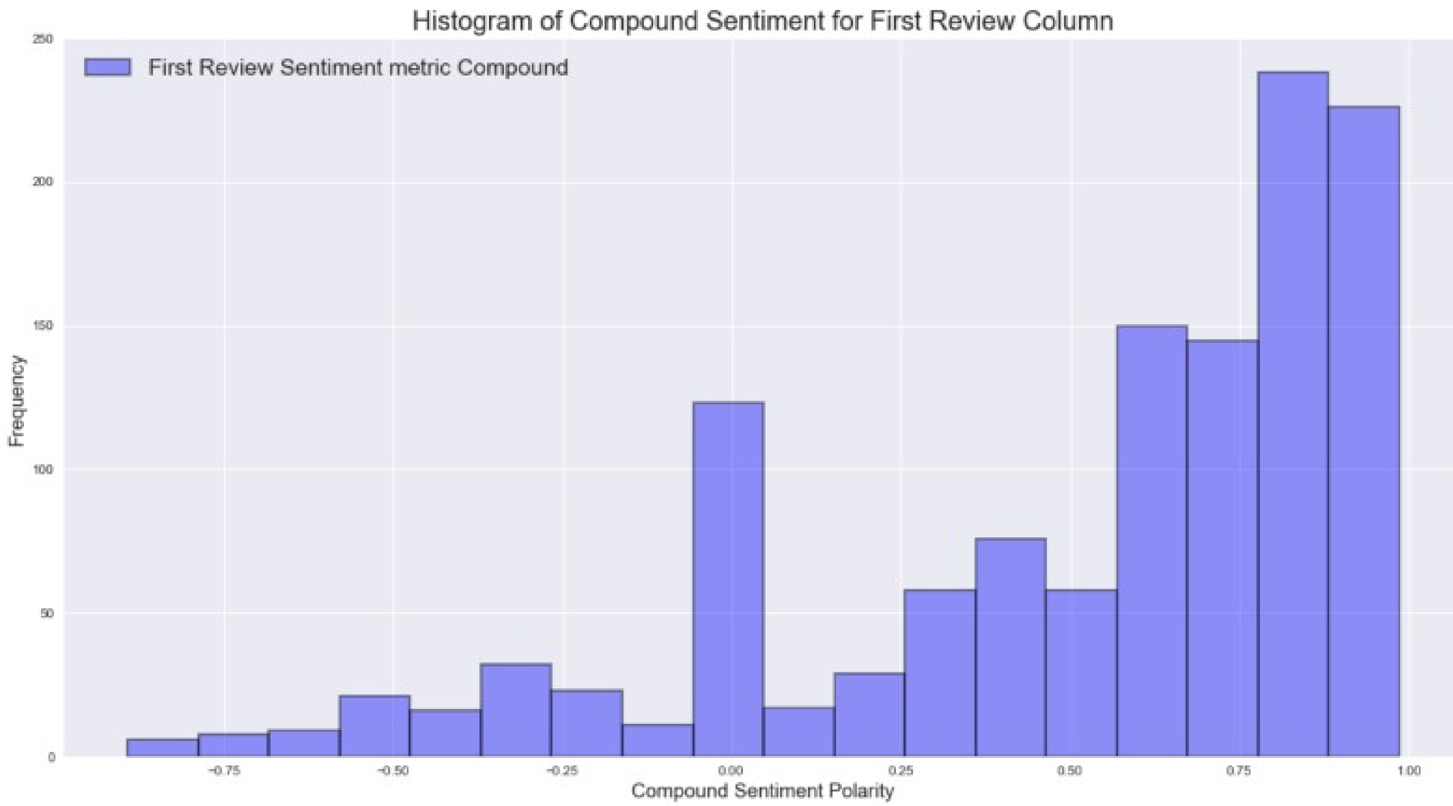
Sentiment Visualization



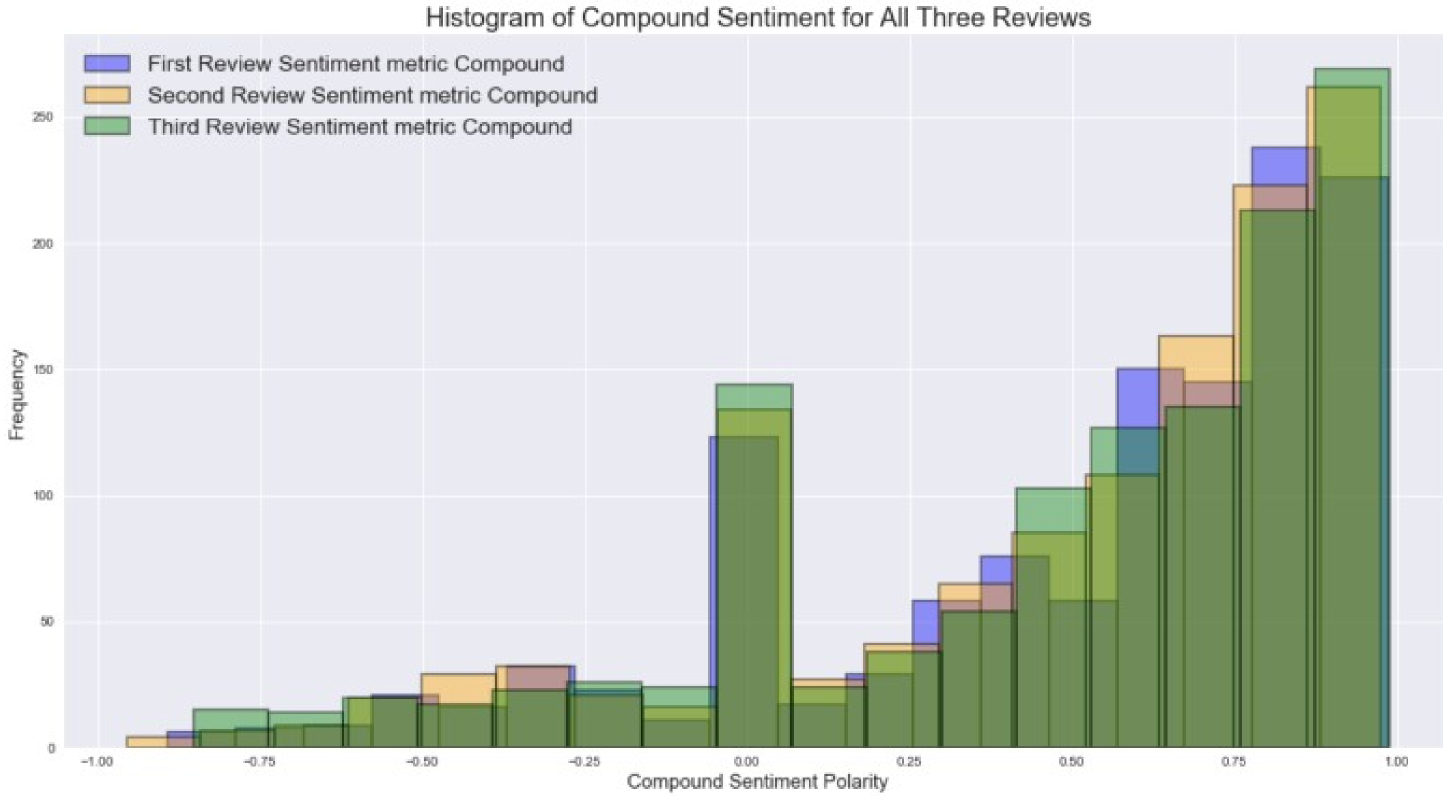
Sentiment Visualization



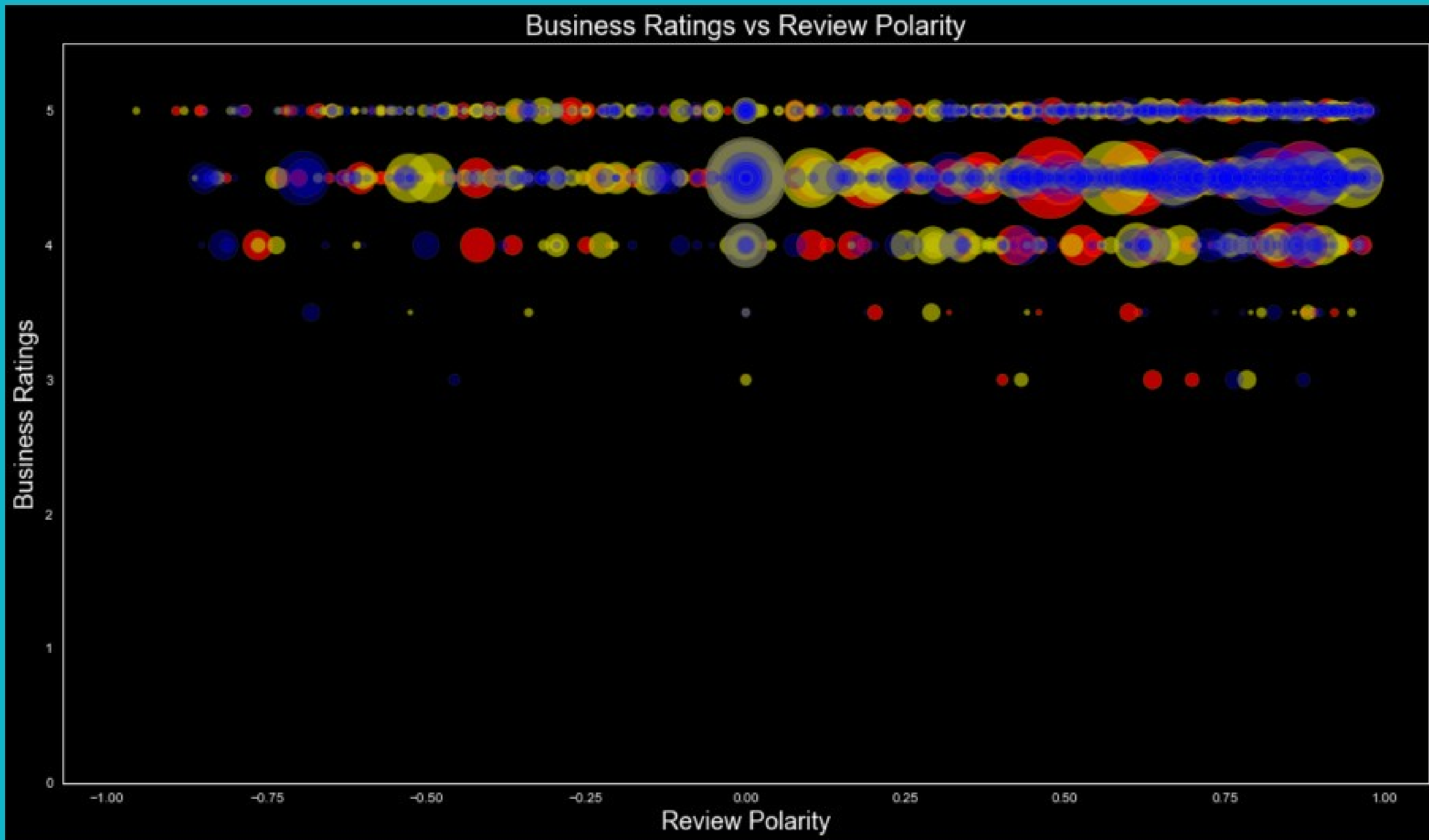
Sentiment Visualization



Sentiment Visualization



Indiviuals in Austin tend to have a Positive Compound Sentiment when giving business reviews


Business Ratings vs Review Polarity
Business Ratings
Review Polarity
0
1
2
3
4
5
-1.00
-0.75
-0.50
-0.25
0.00
0.25
0.50
0.75
1.00

Insights

Just because the compound sentiment is negative doesn't mean the business rating will be low

```
--------

Review 1: When I'm trapped at school without having packed a lunch or dinner (d'oh) it sends me into a crisis mode. W
here can a girl get food around campus where the...

--------

Review 2: I was in an horrific accident September 20th. I was stopped in traffic on I35. And was hit from behind by a
car going 65. Then our A/C went out. My husband...

--------

Review 3: All good Juice/Smoothies are pricey so no complaints about that! I've never had a drink I didn't enjoy her
e. Definitely check it out!

I realize now by...

--------

Review 4: I hate the dentist. Absolutely hate them. I have extreme dental anxiety - like get out of my mouth - get aw
ay from me - don't touch me or I will scream and...

--------

Review 5: I have SEVERE dental anxiety, I usually cry during X-rays. I've had bad experiences in the past when I was
little (accidental teeth pulling that resulted in...

--------

Review 6: I cannot say enough good things about this place. I came to Next Level after having horrible back and neck
pain post baby #2. I've seen both Dr. Sam and Dr....
```

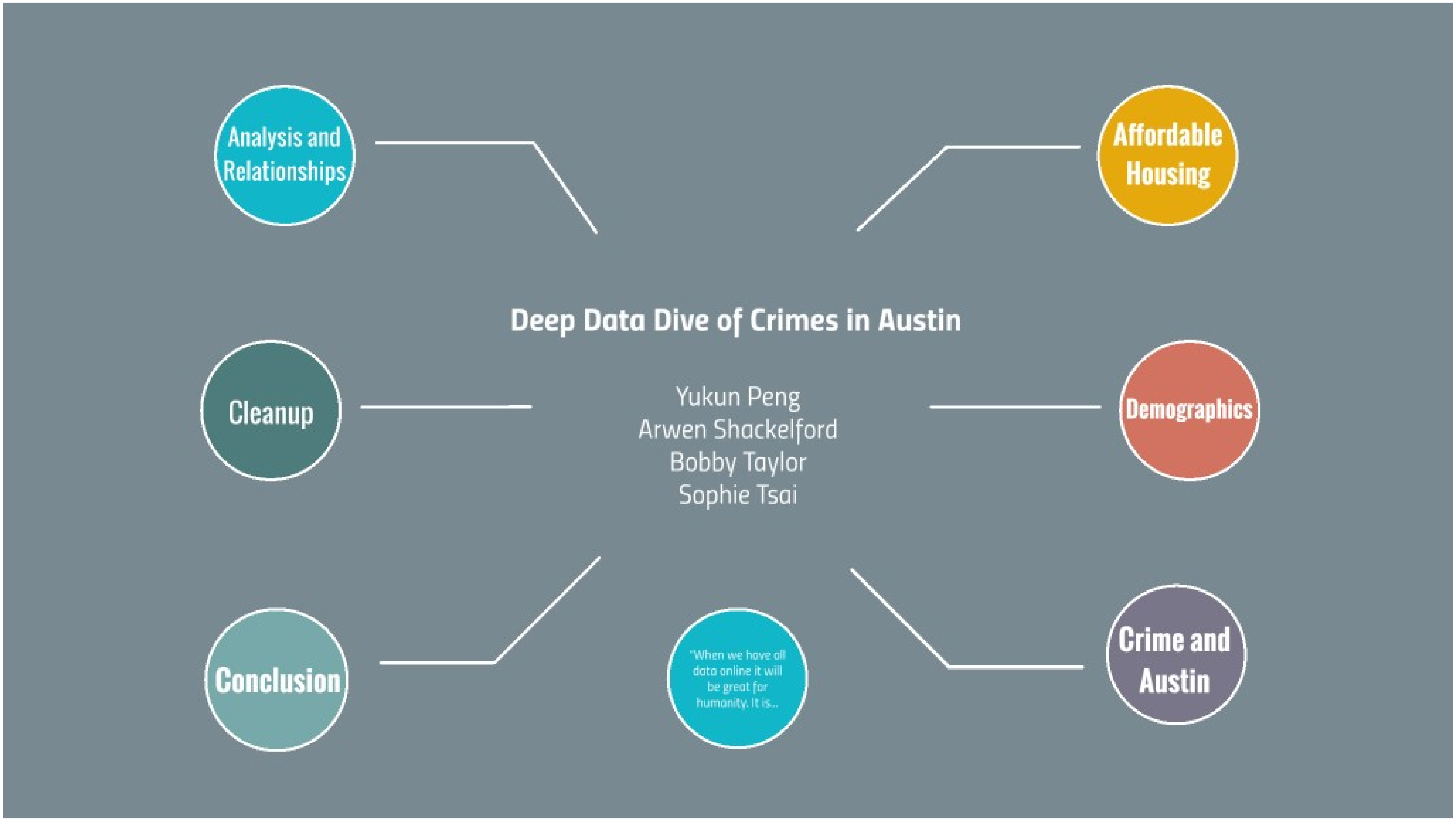
yelp

Statistics

Pearson test results: (-0.20351575062263114, 0.17491773186343634)

Data Relationships

- Most crimes in central-south Austin area
- No correlation between Yelp Reviews and crime data
- Areas with higher crime also have more affordable housing projects


Analysis and Relationships
Affordable Housing
Deep Data Dive of Crimes in Austin
Yukun Peng
Arwen Shackelford
Bobby Taylor
Sophie Tsai
Cleanup
Demographics
Conclusion
"When we have all data online it will be great for humanity. It is...
Crime and Austin


Data Cleaning
Raw Data

Raw Data

After importing dependencies and reading in the csv containing raw data...

```
# Check the columns to see what we're working with
crime_2014.columns
```

```
Index(['GO Primary Key', 'Council District', 'GO Highest Offense Desc',
       'Highest NIBRS/UCR Offense Description', 'GO Report Date',
       'GO Location', 'Clearance Status', 'Clearance Date', 'GO District',
       'GO Location Zip', 'GO Census Tract', 'GO X Coordinate',
       'GO Y Coordinate', 'Location_1'],
      dtype='object')
```

```
# Create a new dataframe with the columns we want to keep
crime_2014_sample = crime_2014[['GO Highest Offense Desc',
       'Highest NIBRS/UCR Offense Description', 'GO Report Date',
       'GO Location', 'Clearance Date',
       'GO Location Zip']]
crime_2014_sample.head()
```

	GO Highest Offense Desc	Highest NIBRS/UCR Offense Description	GO Report Date	GO Location	Clearan
0	AGG ROBBERY/DEADLY	Robbery	04/17/2014	12151 N IH 35 SVRD	04/

Pull a random sample from each csv file

```
crime_2014_sample = crime_2014_sample.sample(frac=0.25, replace=True)
crime_2014_sample.to_csv("2014_Crime_Data.csv", index=False)
```

After reading in the newly created annual crime data csv sample files...

```
# Getting row and column info for each file
print("There are " + str(sample_2014.shape) + " rows and columns for 2014 data")
print("There are " + str(sample_2015.shape) + " rows and columns for 2015 data")
print("There are " + str(sample_2016.shape) + " rows and columns for 2016 data")
```

```
There are (10160, 6) rows and columns for 2014 data
There are (9643, 6) rows and columns for 2015 data
There are (9365, 6) rows and columns for 2016 data
```

Begin the cleanup on the new, larger DataFrame

```
# Check for duplicates
sample_14_15_16.duplicated().sum()
```

```
3605
```

```
# how large will our DataFrame be after removing duplicates?
sample_14_15_16.shape[0] - sample_14_15_16.duplicated().sum()
```

```
25563
```

Drop the duplicated rows

```
sample_14_15_16 = sample_14_15_16.drop_duplicates()
```

Double check the row length

```
sample_14_15_16.shape[0]
```

```
25563
```

Check for rows that have null values and calculate how many in each area

```
sample_14_15_16.isnull().sum()
```

```
GO Highest Offense Desc                    0
Highest NIBRS/UCR Offense Description      0
GO Report Date                             0
GO Location                              445
GO Location Zip                          101
dtype: int64
```

Remove the null values from the DataFrame

```
sample_14_15_16 = sample_14_15_16.dropna()
```

Double check the null values

```
sample_14_15_16.isnull().sum()
```

```
GO Highest Offense Desc                  0
Highest NIBRS/UCR Offense Description    0
GO Report Date                           0
GO Location                              0
GO Location Zip                          0
dtype: int64
```

Checking to see what type of data we are working with

```
sample_14_15_16.dtypes
```

```
GO Highest Offense Desc                  objec
Highest NIBRS/UCR Offense Description    objec
```

Preview the DataFrame

```
sample_14_15_16.head()
```

	GO Highest Offense Desc	Highest NIBRS/UCR Offense Description	GO Report Date	GO Location	GO Location Zip
0	THEFT	Theft: All Other Larceny	02/24/2014 12:00:00 AM	3101 GUADALUPE ST ...	78705.0

The Zip Code column looks odd with that .0 at the end...

We duplicated the zip code column, only we renamed it and turned the values into integers

```
sample_14_15_16['Zip'] = sample_14_15_16['GO Location Zip'].astype(int)
```

```
sample_14_15_16.head()
```

	GO Highest Offense Desc	Highest NIBRS/UCR Offense Description	GO Report Date	GO Location	GO Location Zip	Zip
0	THEFT	Theft: All Other Larceny	02/24/2014 12:00:00 AM	3101 GUADALUPE ST ...	78705.0	78705
1	THEFT BY SHOPLIFTING	Theft: Shoplifting	08/18/2014 12:00:00 AM	1000 E 41ST ST ...	78751.0	78751
2	BURGLARY OF VEHICLE	Theft: BOV	08/22/2014 12:00:00 AM	117 W WILLIAM CANNON DR ...	78745.0	78745
3	AUTO THEFT	Auto Theft	07/19/2014 12:00:00 AM	6407 SPRINGDALE RD ...	78723.0	78723
4	THEFT	Theft: All Other Larceny	07/14/2014 12:00:00 AM	7201 BLESSING AVE ...	78752.0	78752

Changed the 'Zip' column values back into a string

```
sample_14_15_16['Zip'] = sample_14_15_16['Zip'].astype(str)
```

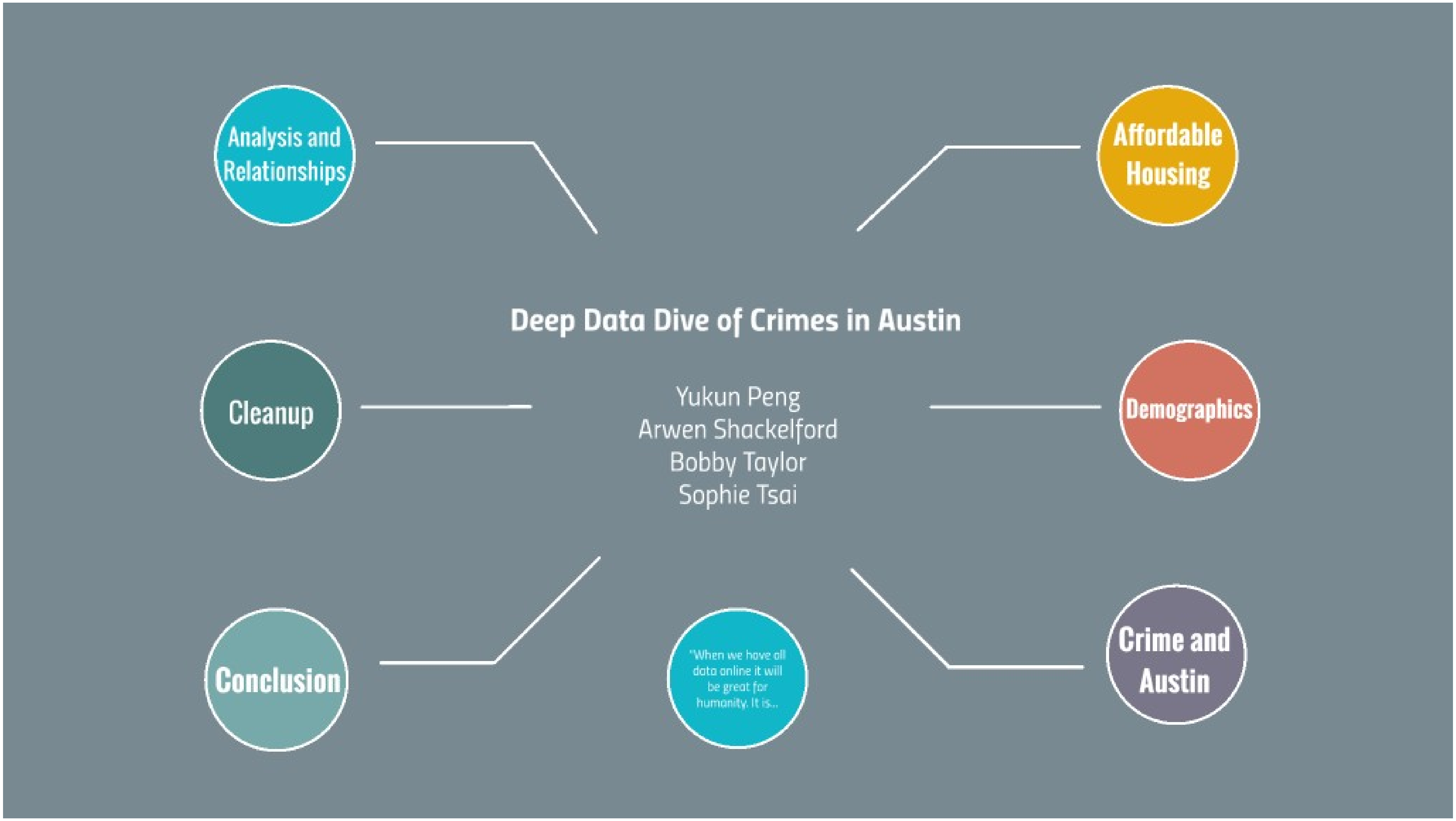
Changed the 'Zip' column values back into a string

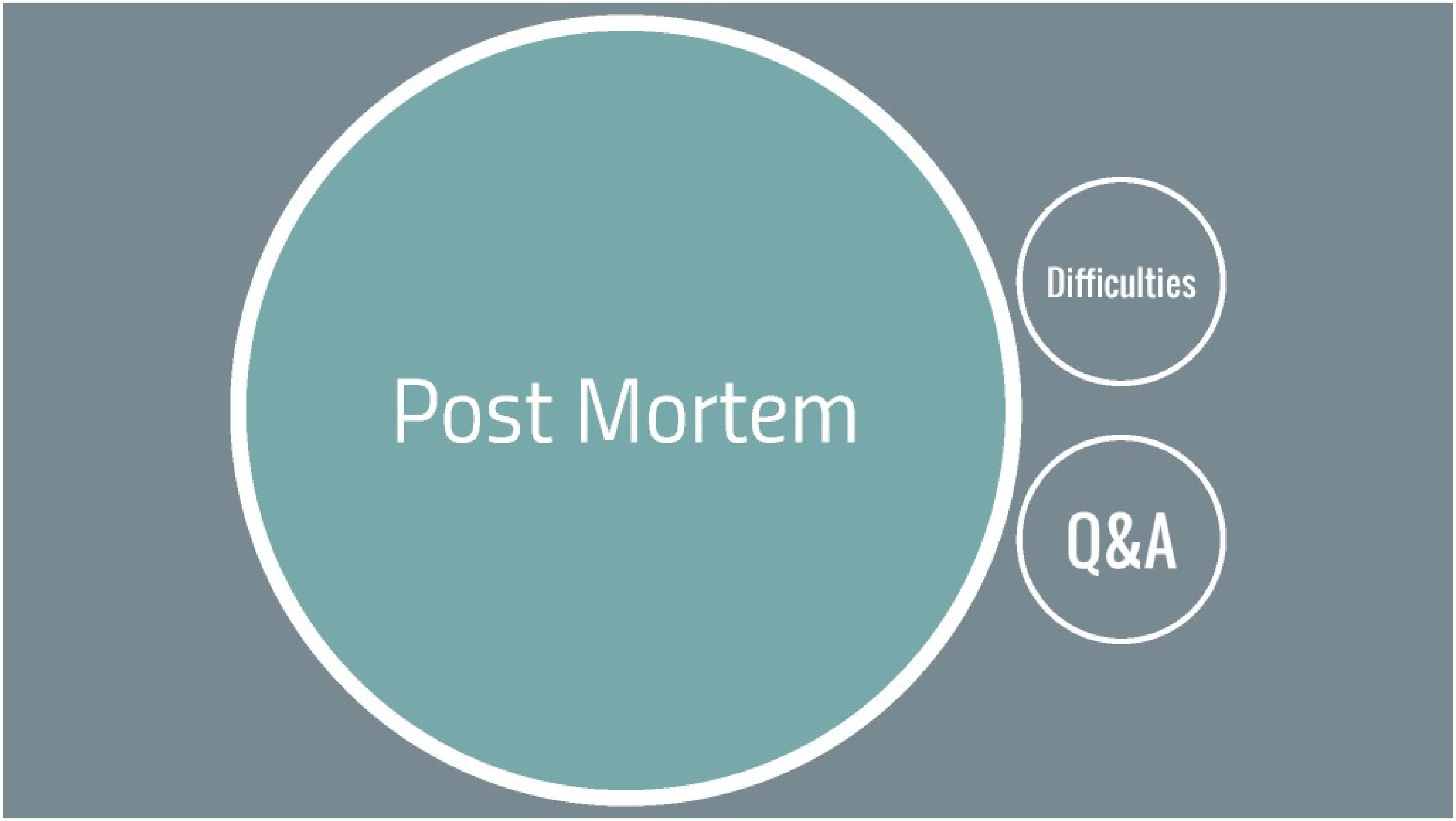
```
sample_14_15_16['Zip'] = sample_14_15_16['Zip'].astype(str)
```

Cleaning values in the DataFrame using .map

```
m = {'Agg Assault': 'Agg Assault',
     'Theft': 'Theft',
     'Robbery': 'Robbery',
     'Burglary': 'Burglary',
     'Auto Theft': 'Theft: Auto Theft',
     'Murder' : 'Murder',
     'Burglary / \nBreaking & Entering': 'Burglary: Breaking & E
     'Homicide: Murder & Nonnegligent Manslaughter': 'Homicide: 
     'Aggravated Assault': 'Aggravated Assault',
     'Theft: Pocket Picking': 'Theft: Pocket Picking',
     'Theft: Purse Snatching': 'Theft: Purse Snatching',
     'Theft: Shoplifting': 'Theft: Shoplifting',
     'Theft: from Building': 'Theft: from Building',
     'Theft: Coin Op Machine': 'Theft: Coin Op Machine',
     'Theft: BOV': 'Theft: BOV',
     'Theft: Auto Parts': 'Theft: Auto Parts',
     'Theft: All Other Larceny': 'Theft: All Other Larceny'}
```

```
sample_14_15_16['Offense_Description'] = sample_14_15_16.Offense_Description.map(m)
```


Analysis and Relationships
Affordable Housing
Deep Data Dive of Crimes in Austin
Yukun Peng
Arwen Shackelford
Bobby Taylor
Sophie Tsai
Cleanup
Demographics
Conclusion
"When we have all data online it will be great for humanity. It is...
Crime and Austin


Post Mortem
Difficulties
Q&A

Difficulties

API Limitations:

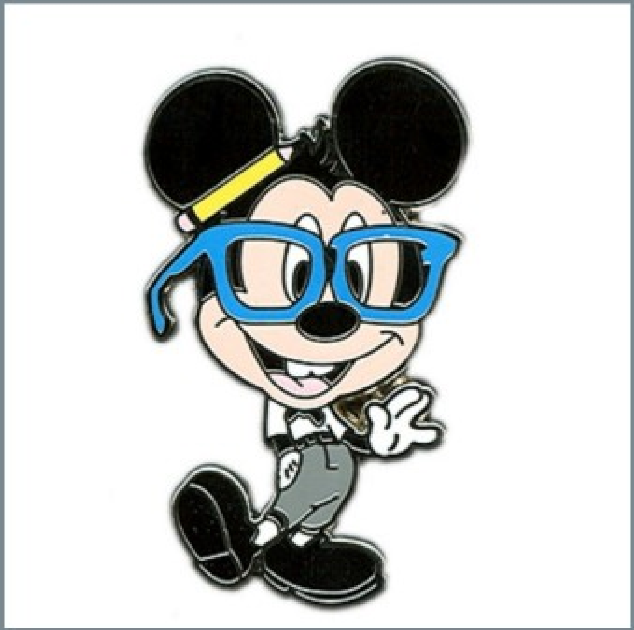
- Yelp API limits number of reviews pulled
- Google API query minute/daily query limit

Time (not enough!)

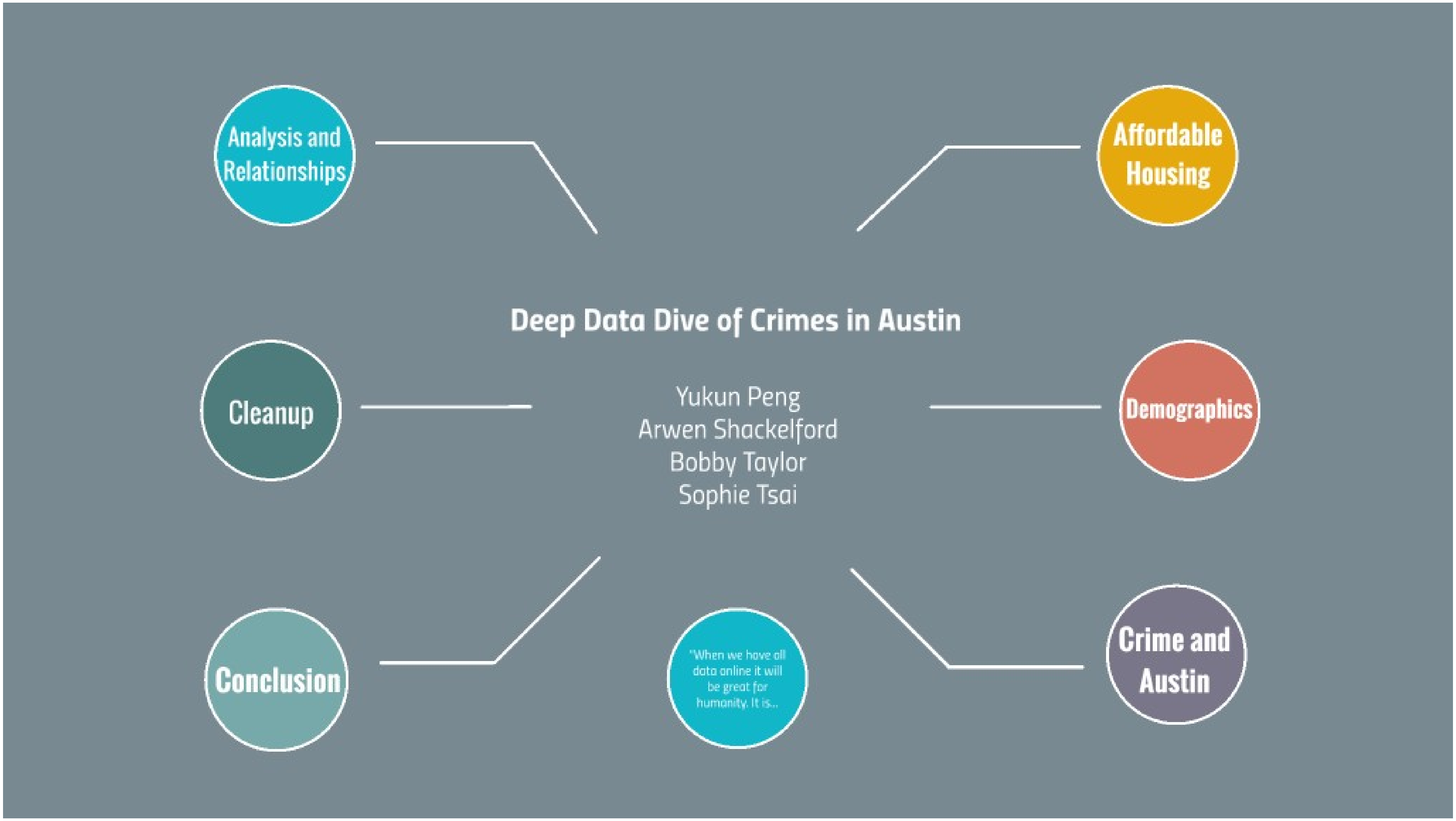
Quality of Yelp reviews - possibly biased

Additional Questions

What other type of data can be analyzed in each zip code?
.....median housing prices
.....family size
.....median household income
.....schools
.....total racial demographics of Austin



Q&A


Analysis and Relationships
Affordable Housing
Deep Data Dive of Crimes in Austin
Cleanup
Yukun Peng
Arwen Shackelford
Bobby Taylor
Sophie Tsai
Demographics
Conclusion
"When we have all data online it will be great for humanity. It is...
Crime and Austin