Observations - Crime Trail Blazers

Satvik Ajmera, John Byun, Shaymus McTeague, Kenneth Tanaka

1 INTRODUCTION

In coming up with a topic, we thought it would be interesting to merge two datasets that didn't have any obvious correlations to one another on the surface, in hopes that some interesting and perhaps novel trends might emerge from the data. For our project group, professional sports was a natural starting point since it's an interest we all had in common and a subject matter for which complete and accurate data was easily accessible. Since sports data comes out-of-box delineated geographically by city, it made sense for us to seek out some form of city data as the companion data set to be paired with our sports data. Having already encountered city crime datasets in some of our class examples, it was known territory for us, which made for a good pairing candidate. Furthermore, we thought that sports and city crime correlations could potentially be interesting, so we went forward with it.

With sports and crime as our chosen areas of interest, we started first by asking the two most basic high-level questions. Do professional sporting events reduce or increase crime? We saw a lot of possibilities there, so we decided to expand on that. It was easy to conceive several common-sense reasons why this could be the case. Perhaps watching games would distract criminals enough that it would keep them off the streets and out of trouble. Or, conversely, watching games leads to an increase in alcohol consumption; lowering people's inhibitions and judgement, ultimately ending in crime. Maybe good game outcomes would put people in a good mood, making them less likely to commit crime. And maybe bad game outcomes would put them in a bad mood, making them more likely to commit crimes. At the root of all these reasons was the same ***core question****: does the occurrence of a professional sporting event influence the rate of crime in a city?*

With this core question, we then went on to decide what sport(s) and team(s) would enter the study. The possibility of there being confounding factors that would bias our data meant that we were going to have to be selective in choosing a city, sport and team for our study. If a city had many professional sports teams, the dataset would be skewed by the various sporting events. To minimize external factors we decided to go with one sport and one team for this study. Our criteria required us to choose a sport with plenty of games in a season to make a strong sample, along with a city with enough crime statistics. Ultimately, the NBA Portland Trailblazers were a perfect pick because the NBA has a relatively long (82 game) season, Portland is one of the few major professional sports cities with only a single major professional sports team, and Portland has one of the highest crime rates in the country per capita. This brought us to the ***Hypothesis***: If a Portland Trailblazers game takes place on a given day, then there will be a decrease in the rate of crime on that day.

As you will see later in this study, our results were inconclusive, but promising in showing a correlation between crime rates as related to Trail Blazers game occurrences.

2 Questions & Data

We analyzed the data for answers to the following questions in order to prove our main hypothesis:

- What trends of crime do we see during the full NBA season, specifically, on game days?
- What is the total crime count per year in Portland?
- How do crime rates on non-game days compare to those of game days?
- Does it make a difference if those games are played home or away?

3 Data Cleanup & Exploration

Fortunately for us, the data we needed was quite readily available as CSV files. Portland Trail Blazers data was gathered from Kaggle (NBA Team Game Stats from 2014 to 2018, 2018):

https://www.kaggle.com/ionaskel/nba-games-stats-from-2014-to-2018 (one CSV file)

And Portland crime data was pulled directly from the Portland crime bureau website (Police Bureau, n.d.):

https://www.portlandoregon.gov/police/71978 (five CSV files from 2015-2019)

3.1 Cleanup process

The NBA csv file was complete and contained data that was not pertinent to our project. The first steps required excluding all NBA teams except for Portland (POR). Date was reformatted using datetime, which allowed us to create a new column for the days of the week (DOW). A new data frame was created to remove data that was not pertinent to the project (ex: free throws, rebounds, assists, turnover, et al.). The result was a file of dates, home and away games, wins and losses for the Portland Trailblazers.

We had initially tried to merge two crime files and quickly learned that was not correct when we were left with a csv file with 57 million rows. After speaking with Koren, we were able to use concatenate to combine the crime files. We next needed to decide which of the two date columns, ReportDate and OccurDate, to use. After some analysis, the ReportDate included crimes in previous years and so we chose to use OccurDate and reformatted using datetime. The crime files also contained columns that were not pertinent to our project (ex: case number, address, latitude, longitude, et.al.). Five csv files for the period 2015 - 2019 were combined. We restricted the crime dates for the project to January 1, 2015 - December 31, 2018. (The 2019 file was included in order to capture any crimes that had occurred in prior years but were not reported till 2019.) A new data frame was created using the columns needed for the project; dates, neighborhood, offense category, DOW. (We had initially included neighborhood but did not use this data in our final project.)

After creating a simple line of the number of crimes (y axis) by month (x axis) for all years, the chart showed a big spike in crime in April. It was not until after creating a line graph for each year, did we see that the spike was only in 2015. Upon analyzing the 2015 data, we were surprised to see missing dates. Doesn't crime occur every day?

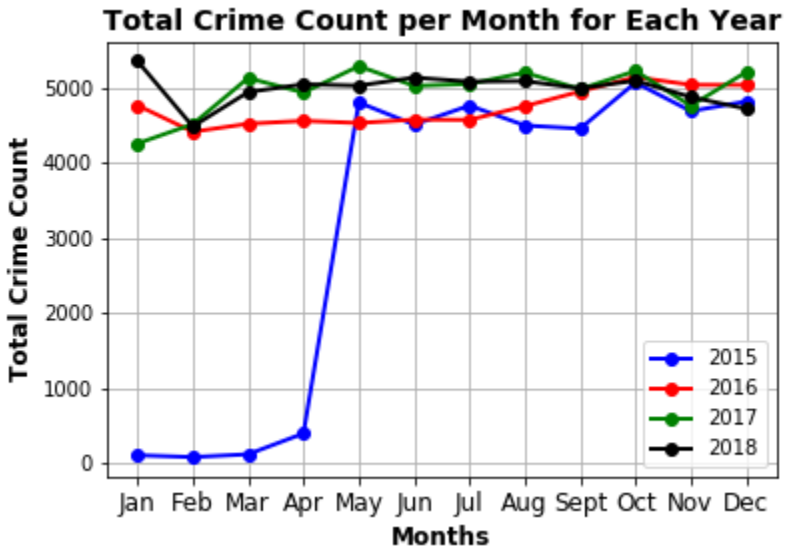
Further research into the Portland Police website, we found the cause:

"Prior to April 2015, Portland Police Bureau "PPB" reported statistics using the Uniform Crime Reporting (UCR) Program's Summary Reporting System (SRS). The SRS used a hierarchy to decide which offense to report. In the hierarchy system, only the highest or most serious offense is reported on each case." (Police Bureau, n.d.)

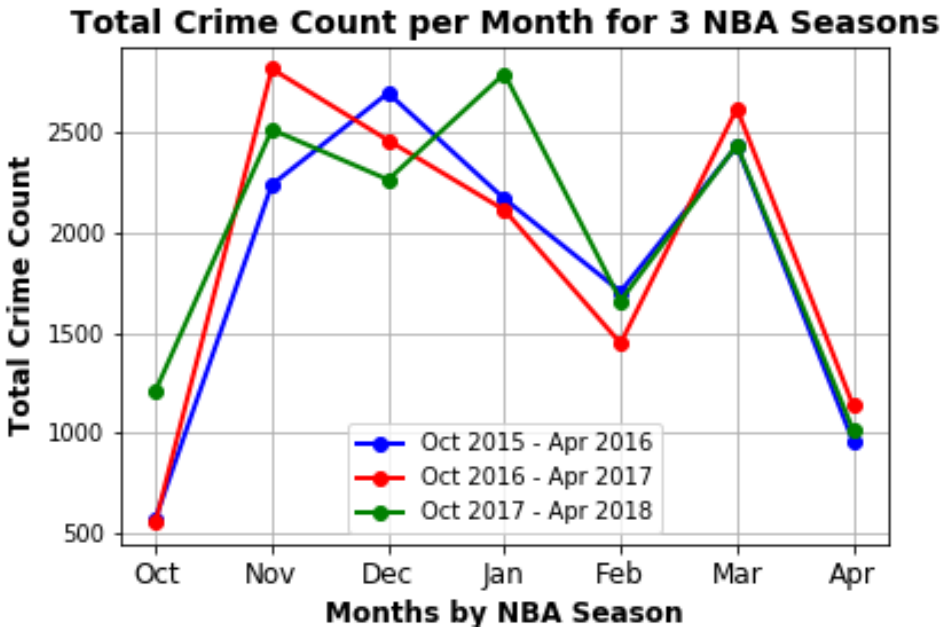
In April 2015, the PPB was recruited to participate in the National Crime Statistics Exchange, a program designed to compile a national sample of incident-based crime data using the National Incident-Based Reporting System (NIBRS). As a result, this changed the way reported crime statistics are calculated and presented. This new system NIBRS reports all unique offenses within an incident rather than just the most serious offenses.

4 Data Analysis

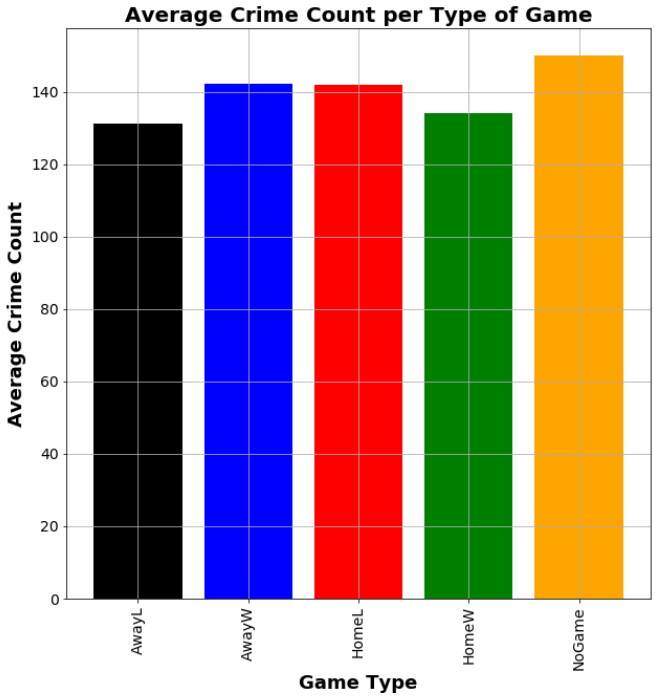
4.1 Visualizations



This visualization showed the total crime count per month during the years 2015, 2016, 2017 and 2018. In fact, this visualization showed that there was no visible trend in the crime data. During every year, the total crime counts per month varied. During 2015, January, February, March and April showed an upwards trend in crimes. However, this is due to the Portland crime reporting measures during this time.



For this visualization, we wanted to see if there are any trends in the total crime count per month for each NBA season. The visualization eliminated the bad data during Jan, Feb, Mar and Apr 2015. Even though our data did not show any obvious trends, it was interesting to at least be able to observe the peaks and troughs of the total crime count during this time.



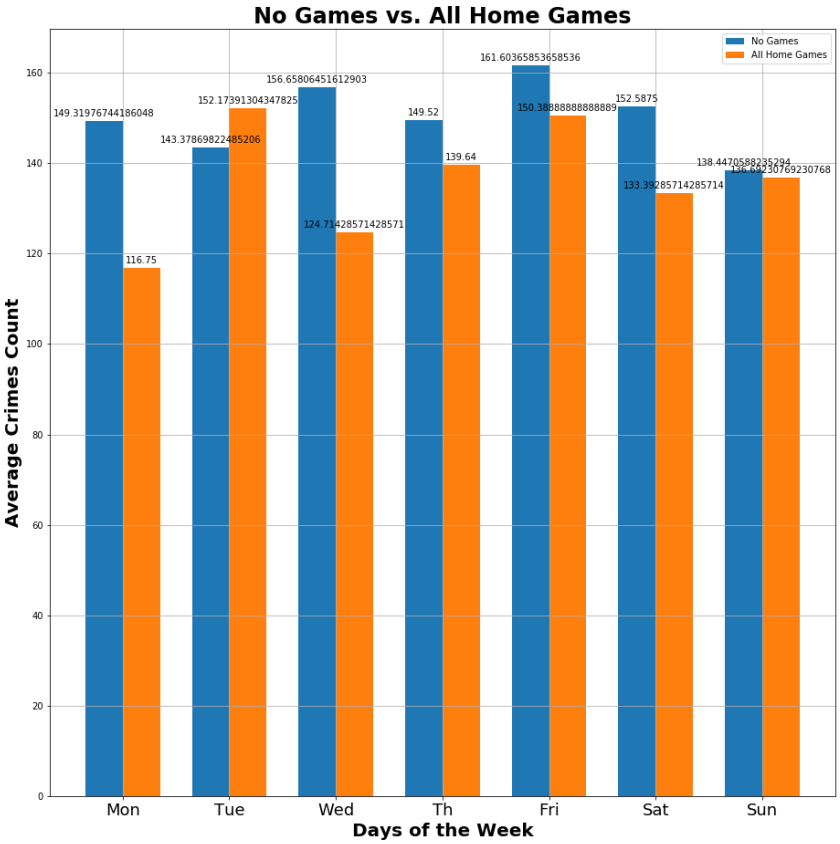
In this bar chart we have the y-axis as the average crime count and the x-axis the game type (AwayL, AwayW, HomeL, HomeW and NoGame). The AwayL bar shows the average crime count for every away game the Portland Trail Blazers lost. The AwayW bar shows the average crime count for every away game that the Trail Blazers won. The HomeL and HomeW bars show the average crime count for home losses and home wins in Portland. Lastly, the NoGame bar shows the average crime count for every non-game day. Furthermore, this is using crime data from 2015, 2016, 2017 and 2018. It is true that our data could be skewed just due to the occurrences of dates for the given game types. As you can see below:

```
In [33]: total_table
```

Out[33]:

GameStatus	OffenseCategory	Date	AverageCrime
AwayL	11284	86	131.209302
AwayW	8687	61	142.409836
HomeL	6953	49	141.897959
HomeW	13150	98	134.183673
NoGame	174815	1165	150.055794

For example, no game days have 1165 occurrences. On the other hand, HomeL has significantly less date occurrences and offense category counts. So, for further study, we would have to figure out a way to normalize the average crime count values because of this discrepancy.



4)

For this visualization, we compared all home games to non-game days. Using the years 2015, 2016, 2017 and 2018, we figured out the average number of crimes committed for each day of the week. As we can see, the average crimes count on non-game days exceed the average crimes count for home games except for Tuesdays. On Tuesdays, the average crime count is greater than all no game Tuesdays. I found this visualization interesting and of more substance, because it leads to conclusive evidence.

4.2 Statistical Testing

4.2.1 The Chi Square Test

We took the average crime for each category: Away Loss, Away Win, Home Loss, Home Win, No Game as the observed values. The expected average crime was calculated by taking the total Offense Category / total dates.

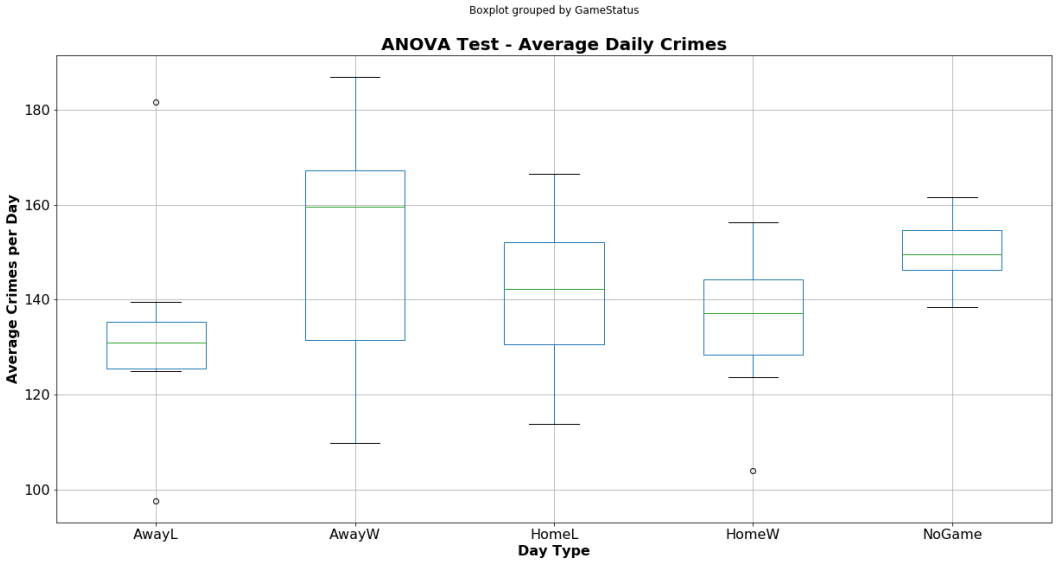
Unfortunately, the chi square value was less than the critical value, and so, we concluded from this that the results were not statistically significant.

- Critical Value = 9.48772
- Chi Square = 3.33059
- p-value = 0.50409

4.2.2 ANOVA Test

Though the Chi Squared Test didn't produce significant results while comparing the game types to *expected*, we figured we'd see if there was a significance between the game types in general, using an ANOVA Test. And despite seeing some hopeful variation within a box plot, the p-value lacked significance.

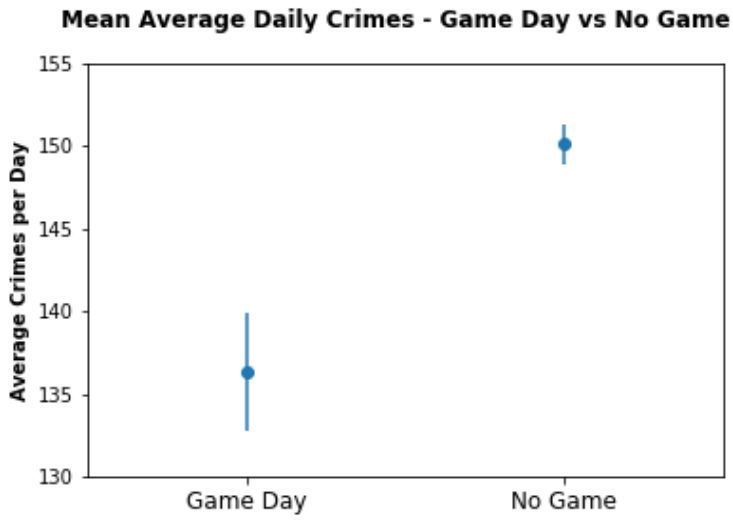
- p-value = 0.3593



4.2.3 Independent T Test

At this point, we began to feel that rejecting the null hypothesis was impossible; but by just looking at the box plot created for the ANOVA test gave us reason to believe looking at all game type variables separately doesn't really answer the inherit question of our hypothesis: "Do Trail Blazer games – in general – reduce crime?" With this new viewpoint an independent T Test was our final test for statistical significance to test general 'game days' vs 'non-game days'.

Starting by plotting an error bar graph, we can see a clear separation of the independent variables with no error overlap.



From here, we run the independent T Test and yield our result. Which in this case, provided a p-value of 0.00032. Which far beyond the 0.05 threshold we were looking to fall below. Great! We found significance, we found negative correlation. Again… Great! But also, so what? Does this correlation really mean causation? Let's think about that a little further.

5 CONCLUSIONS

So, in the end we found that game days vs non-game days had a statistically significant negative correlation. Unfortunately, thought, this still leave our study inconclusive. From this data alone we cannot reject the null hypothesis – meaning we have not proved our hypothesis. What we have done is provide that our hypothesis and is strong enough to encourage further research; which we can explore in the next section, the Postmortem.

6 POSTMORTEM

6.1 TWO MORE WEEKS

If we had two more weeks to work on this experiment, how could we best utilize that time to draw a definitive conclusion? First, we would re-examine our data sets possibly omit the four months of bad data in 2015, where the data had incomplete, pre-NIBRS information data (Police Bureau, n.d.) or omit 2015 all together to maintain a complete 12 month crime year across the board.

Second, finding negative correlation within Portland is nice, but being able to find similar results in other cities would be great. With two more weeks, we would have benefited by running the same tests in other NBA cities to see if there is a universal trend.

6.2 DIFFICULTIES

6.2.1 Data

Our biggest hindrance within this study had to have been the data gap in the 2015 crime packet. Four months out of four years of bad data when you are working with 48 total months will inevitably have an impact on your study. Over 8% of our crime data is problematic, which puts the whole outcome into question; which we believe would have been best resolved if we omitted 2015 all together and expanded the cities and teams. Unfortunately, time was too inhibitive to complete that task, but was a great learning experience.

6.2.2 Human

One last difficulty to mention, though it is not strictly data related, is the human factor. Our group worked great together, but we had a later start than we wanted because we found ourselves changing the project scope all the way until one week before the presentation date. We can attribute this to many things (varied interests, different time schedules, scope creep, desire to impress our teacher and TAs, etc.), but in review it appears that we were a project team that spent a lot of time "storming". If you are familiar with Tuckman's stages of group development (Tuckman, 1965) there are five stages of how a team develops. Using Tuckman's stages, here is how our group functioned:

- Forming: Quick to form group, get along great
- Storming: Took a long time to agree upon topic, then to agree upon scope of that topic
- Norming: We really started to feel like a normalizing group during the last week of the project
- Performing: Each team member's skills and roles were becoming obvious in the closing of the project, three days before project was due.

(Agile Scrum Guide, 2018)

Ultimately, this was a great team and a fascinating project to really dig our teeth into and build our skills within python with pandas and matplotlib. We had a great time working together and look forward to the next challenge that comes along.

7 REFERENCES

Agile Scrum Guide. (2018, February 5). Retrieved from Use Tuckman's Model of Team Dynamics: https://agilescrumguide.com/blog/files/Use-Tuckmans-Model-of-Team-Dynamics.html

NBA Team Game Stats from 2014 to 2018. (2018). Retrieved from Kaggle: https://www.kaggle.com/ionaskel/nba-games-stats-from-2014-to-2018

Police Bureau. (n.d.). Retrieved from The City of Portland Oregon: https://www.portlandoregon.gov/police/article/618535

Tuckman, B. (1965). Development sequence in small groups. *Psychological Bulletin*, 384-99.