Checkpoint 3: Workflow Analytics **GROUP 5**
MSAI 339 **Jack R**
Prof. Jennie Rogers **Quincia H**
November 13, 2018 **Ikhlas A**

Purpose

The purpose of this checkpoint is to extrapolate from the previous checkpoint's cleaned tables. This will be carried out through using a Spark cluster to carry out more rigorous queries that may require more processing power.

Databricks

Databricks is a platform that acts as a universal terminal for data engineering, SQL querying, and data science. By uploading tables to a database provided by Databricks, we are able to take advantage of Spark's powerful processing power for quick querying and analysis. To answer the questions described in the project proposal for this checkpoint, we will upload the necessary tables to Databricks and build a notebook that provides sufficient information to build corresponding answers.

Workflow Analytics

Is there a specific district that consistently sees a higher percentage of complaints and crimes?

This was a necessary, fundamental question to analyze for further use in spark. District 8, Chicago Lawn, had the most complaints and the most crimes. We continued off of this by calculating the correlation coefficient between crimes and complaints, resulting in 0.66. This coefficient shows a positive correlation and that it it could be fairly predictable in the future.

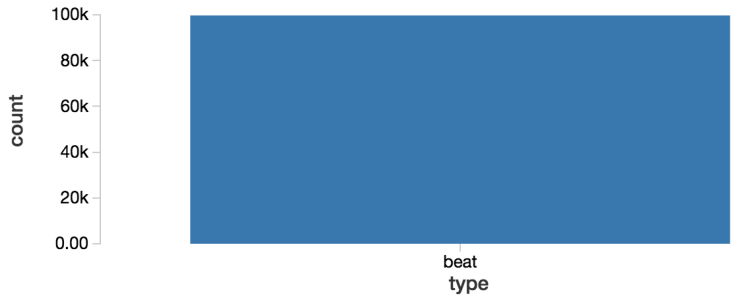
```
1 complaints = spark.sql("SELECT District, Count as Complaints from complaints_per_district")
2 crimes = spark.sql("SELECT District, Crime_Count as Crimes from crime_per_district")
3 district = crimes.join(complaints, "District")
```

District	Crimes	Complaints
2	340788	3440
3	362402	6925
4	404926	9248
5	315937	10117
6	411492	8682
7	422815	10299
8	486888	13095
9	353192	9371
11	453863	7904
12	348443	11259
15	312067	6529
16	236689	7628
17	207034	3536
18	312161	4012
19	317213	6675
20	123799	3928
21	164	4010
22	234083	3783
24	212819	4482
25	412287	6077

Is there a relationship between the area type and clusters of officers?

The purpose of this question is to attempt to identify personal vendettas against officers in areas. To approach this, we used the beat attribute to identify the area of an allegation, and connect the result to officers. After pursuing the answer of this question, we realized there is little information gain from connecting the beat attribute of allegations that we wouldn't already gain from using the district information. As the area table has very little information, it may be better to pursue a different relationship.

```
area = spark.sql("select id, type from area")
allegations_per_beat = allegations.drop("date").groupby("beat").count()
area = area.join(allegations_per_beat, allegations_per_beat["beat"] == area["id"])
```



What is the correlation coefficient of the relationship between police officer salary and the number of complaints received? What about the relationship between police officer rank and number of complaints received?

To identify the relationship between salary, rank, and number of complaints, we grouped by the column in question to get the number of complaints and used the correlation coefficient package in Pyspark. With the resulting correlation of -0.2 between salary and the number of complaints, there is a subtle negative correlation. With rank, nearly all of the complaints were against police officers, with a handful of allegations distributed between the rest of the ranks.

After trying to bucket the salaries into buckets of every $2500, $5000, and $10000, we came to a correlation coefficient between number of complaints against and officer and salary to be 6.433360349144116e-05, basically 0. I think the negative correlation stemmed from not bucketing the salaries. Because the allegations against officers are largely ranked "Police Officers" they all have similar salaries and a wide variety of complaint counts. The initial negative correlation meant as number of complaints increased, there was a minor correlation showing salary would decrease. If the negative correlation was taken at face value, it would show higher paid officers would likely have less complaints against them. After bucketizing the data, though it showed no correlation. This means that there isn't a distinct relationship between the number of complaints and the salary of an officer. A potential takeaway is that this shows there is no evidence that underpaid officers are more likely to receive more complaints. A potential followup question could be trying this same approach but on complaints against solely "Police Officer" ranked officers. If an officer feels as they are underpaid for their rank, maybe there is some further correlation.

```
allegations_officers = spark.sql("SELECT allegation_id,officer_id from officer_allegation_csv")
allegations = spark.sql("select ID, incident_date as date, beat_id as beat from allegations_csv")
rank_salary = spark.sql("select id, rank, salary from salary_csv")
tempA = allegations_officers.join(rank_salary,allegations_officers['officer_id'] == rank_salary['id']).drop('id')
df = tempA.join(allegations,tempA['allegation_id'] == allegations['ID']).drop('ID')
rank_salary = spark.sql("select rank, salary from salary_csv")
rank_salary = rank_salary.groupby("rank").agg({'salary':'mean'})
```

```
complaints_by_salary = df.groupBy('salary').count()
print(complaints_by_salary.corr("salary","count"))
```

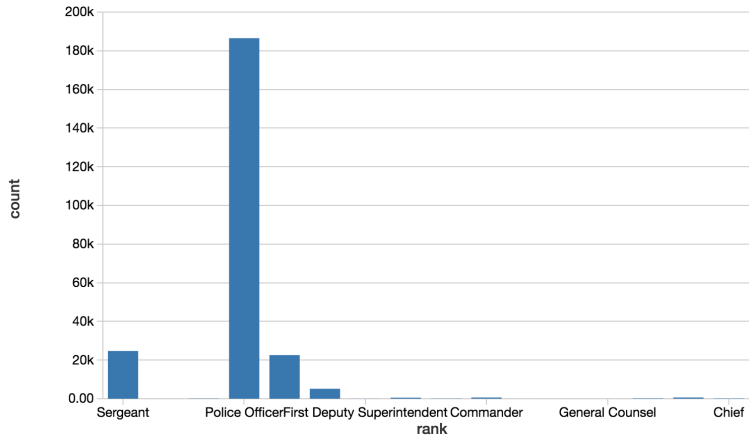
▸ (1) Spark Jobs

▾ complaints_by_salary: pyspark.sql.dataframe.DataFrame

salary: integer

count: long

```
-0.26332457391826597
```



What is an officer's track record of complaints over a period of time?

The purpose of this question is to see which officers have sudden and a large number of complaints from an otherwise "quieter" behavior. To answer this question, we decided to find the earliest and latest dates of allegations against an officer. We then divided the resulting date range by the number of allegations against the officer. This ratio provides a rough overview of how often an officer receives an allegation. While this only handles global maximums and minimums rather than local max and min dates for officers with numerous allegations, it still can be extended to the machine learning checkpoint.

```
from pyspark.sql.functions import datediff, to_date, lit, round
officer_complaints = df.groupby("officer_id").count()
maxDates = df.groupby("officer_id").agg({"date":"max"})
minDates = df.groupby("officer_id").agg({"date": "min"})
dateRanges = minDates.join(maxDates,"officer_id").join(officer_complaints,"officer_id").withColumn("Date Range", datediff("max(date)", "min(date)"))
dateRanges2 = dateRanges.withColumn("Ratio", round(dateRanges["Date Range"] / dateRanges["count"], 2))
```

officer_id	min(date)	max(date)	count	Date Range	Ratio
148	1992-11-11T00:00:00.000+0000	2008-12-31T00:00:00.000+0000	4	5894	1473.5
471	1981-06-20T00:00:00.000+0000	1985-04-29T00:00:00.000+0000	3	1409	469.67
833	1989-05-16T00:00:00.000+0000	2009-08-20T00:00:00.000+0000	44	7401	168.2
1088	1984-07-13T00:00:00.000+0000	1984-07-13T00:00:00.000+0000	1	0	0
1645	1992-06-16T00:00:00.000+0000	1995-07-30T00:00:00.000+0000	4	1139	284.75
1829	1998-10-23T00:00:00.000+0000	2004-05-25T00:00:00.000+0000	4	2041	510.25
1959	1990-10-01T00:00:00.000+0000	1995-02-04T00:00:00.000+0000	5	1587	317.4
2142	2016-03-08T00:00:00.000+0000	2016-03-08T00:00:00.000+0000	1	0	0

Results and Analysis

Apart from the area type question providing the be less fruitful, the results received on the other three questions were beneficial. It was interesting to see a lower negative correlation coefficient between salary and complaints. Given more time it may be beneficial to bin the salaries into paygrades and check for a correlation coefficient after. ***The results from our first question were definitely easier to interpret after using Spark's scattergram plot. Ultimately it showed as crimes increase, complaints increase as well (as seen with the 0.66 correlation coefficient)***. Ultimately we have built a few tables that will be beneficial in the future checkpoints, especially with heuristics being built off the date range table.