QUESTION 1: Take a look at `labeled_data.csv`. Write the functional dependencies implied by the data.

The functional dependencies are listed below:

Input_id is functionally dependent upon (labeldem, labelGOP, labeldjt)

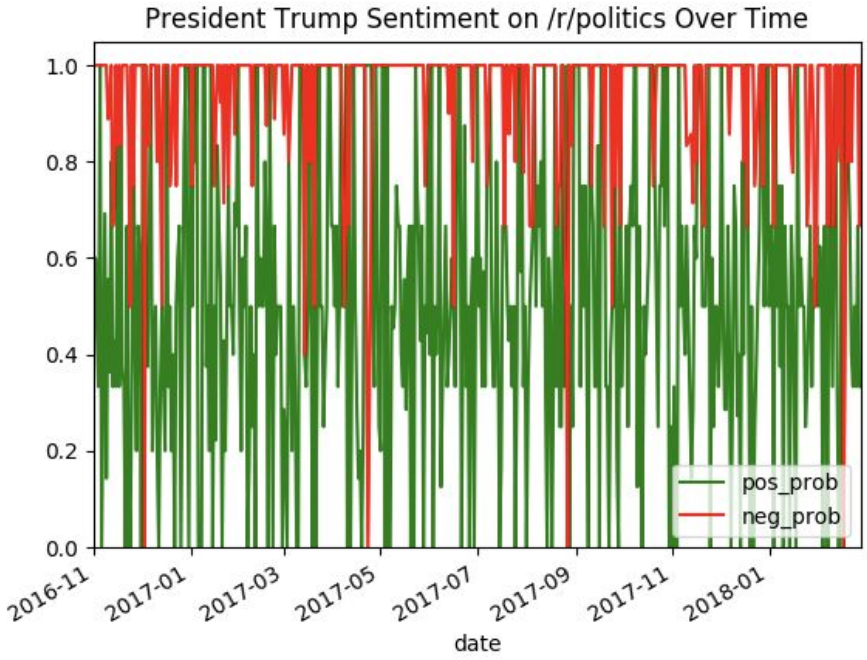
QUESTION 2: Take a look at the schema for the comments dataframe. Forget BCNF and 3NF. Does the data frame *look* normalized? In other words, is the data frame free of redundancies that might affect insert/update integrity? If not, how would we decompose it? Why do you believe the collector of the data stored it in this way?

The schema doesn't look normalized. There are redundancies that can be seen throughout the schemas, such as having multiple copies of different ids across many different fields. They can be broken down into their own tables and be referred to, which would remove the redundancies. One example of such a redundancy is the information about the author. We should not have all of this information in one comments table, as it is unnecessary and does not make sense. Instead, having a separate table for this information and referencing it with an id is better. In general, it makes more sense to split such redundant data into different tables where they would make logical sense and reference them in the comments table using some id like author_id.

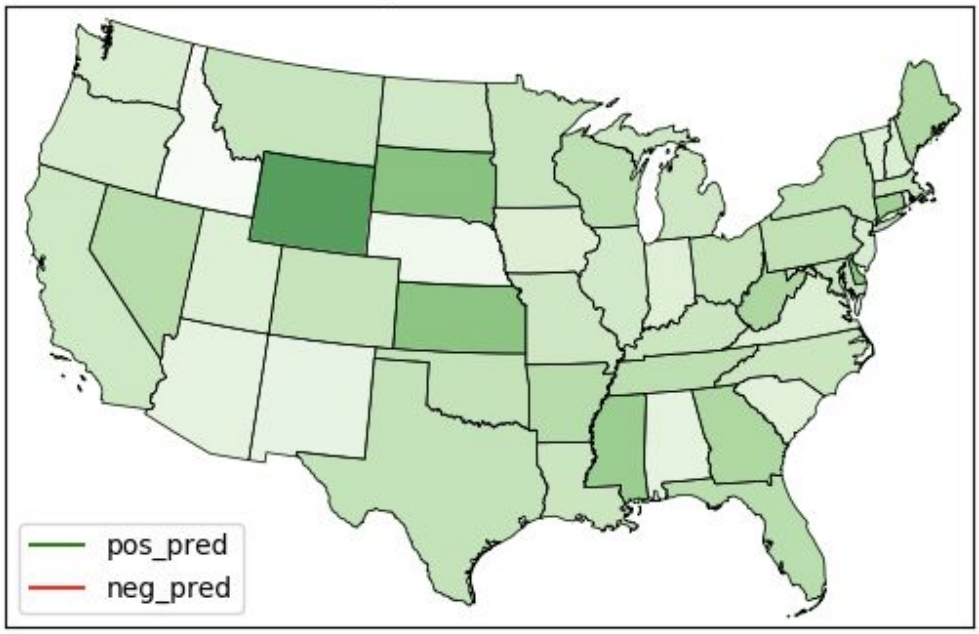
QUESTION 3:

```
== Physical Plan ==
*(2) Project [id#3465, body#3455, labeldem#3512 AS Dem#3520, labelgop#3513 AS GOP#3521,
labeldjt#3514 AS Trump#3522]
+- *(2) BroadcastHashJoin [Input_id#3511], [id#3465], Inner, BuildLeft
   :- BroadcastExchange HashedRelationBroadcastMode(List(input[0, string, true]))
   :  +- *(1) Project [Input_id#3511, labeldem#3512, labelgop#3513, labeldjt#3514]
   :     +- *(1) Filter isnotnull(Input_id#3511)
   :        +- *(1) FileScan csv [Input_id#3511,labeldem#3512,labelgop#3513,labeldjt#3514] Batched:
false, Format: CSV, Location:
InMemoryFileIndex[file:/media/sf_vm-shared/143proj2/labeled_data.csv], PartitionFilters: [],
PushedFilters: [IsNotNull(Input_id)], ReadSchema:
struct<Input_id:string,labeldem:string,labelgop:string,labeldjt:string>
   +- *(2) Project [body#3455, id#3465]
      +- *(2) Filter isnotnull(id#3465)
         +- *(2) FileScan parquet [body#3455,id#3465] Batched: true, Format: Parquet, Location:
InMemoryFileIndex[file:/media/sf_vm-shared/143proj2/comments.parquet], PartitionFilters: [],
PushedFilters: [IsNotNull(id)], ReadSchema: struct<body:string,id:string>
```

Spark seems to be using a Broadcast Hash Join to filter and query the data to create the final joined result. The code first scans the comments parquet file and filters the null out so we only have the values we want that are not null. Then we do a Broadcast Hash Join to join the data from the commnents.parquet and label_data.

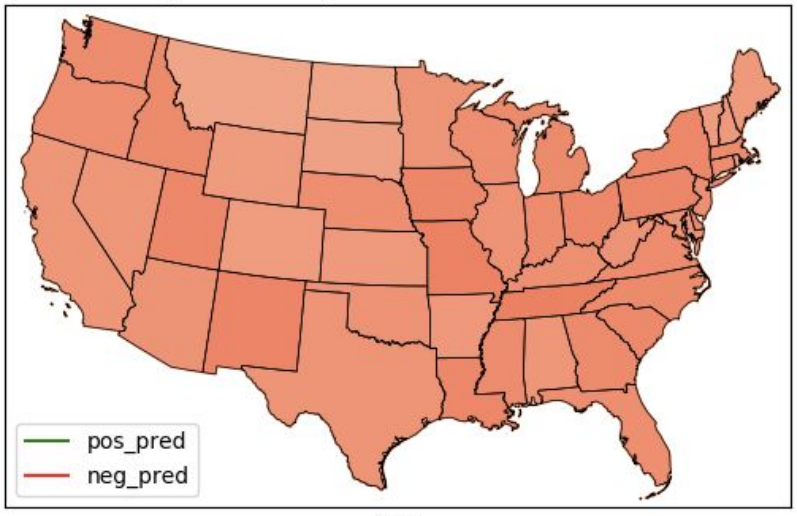

President Trump Sentiment on /r/politics Over Time
1.0
0.8
0.6
0.4
0.2
0.0
pos_prob
neg_prob
2016-11
2017-01
2017-03
2017-05
2017-07
2017-09
2017-11
2018-01
date

Positive Trump Sentiment Across the US

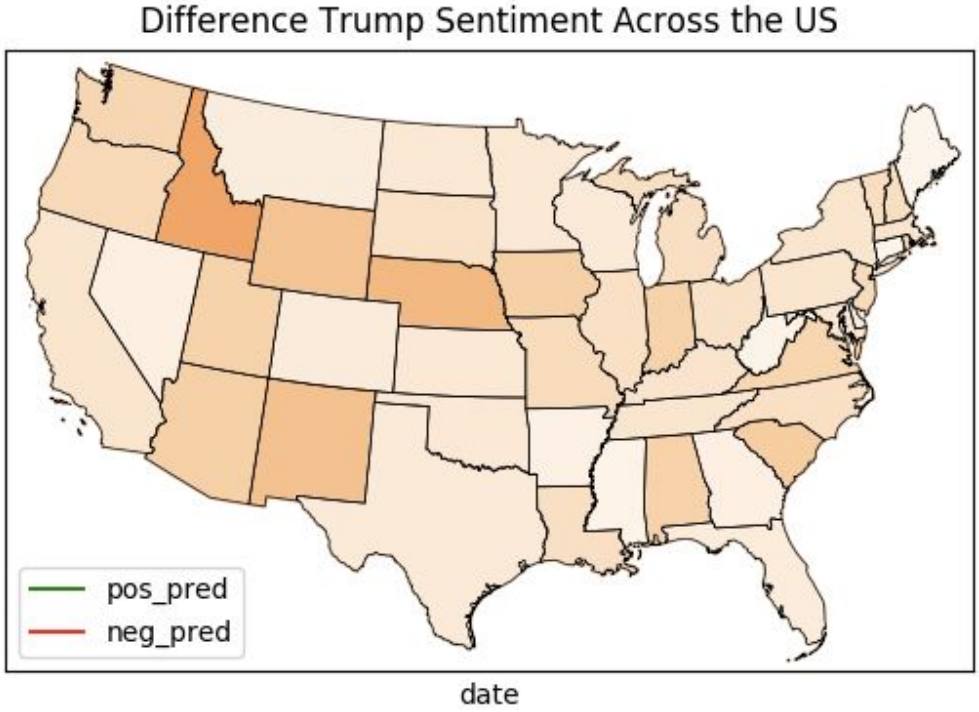


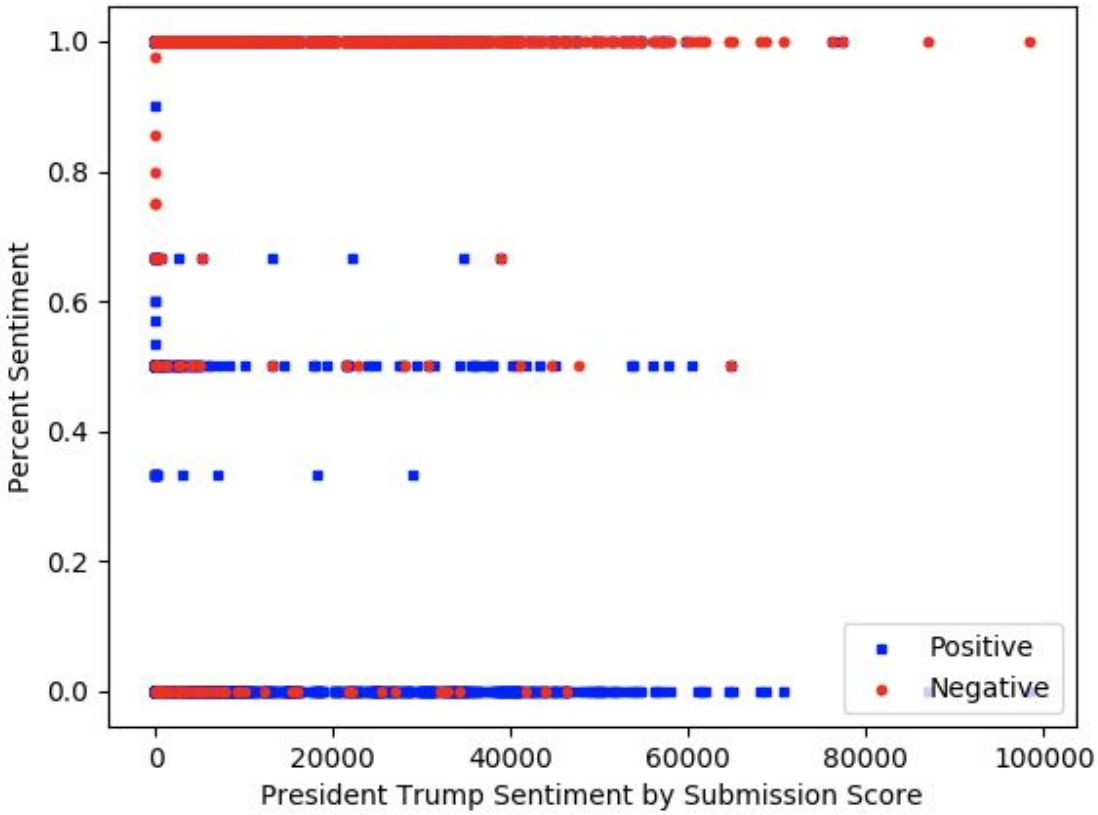
date

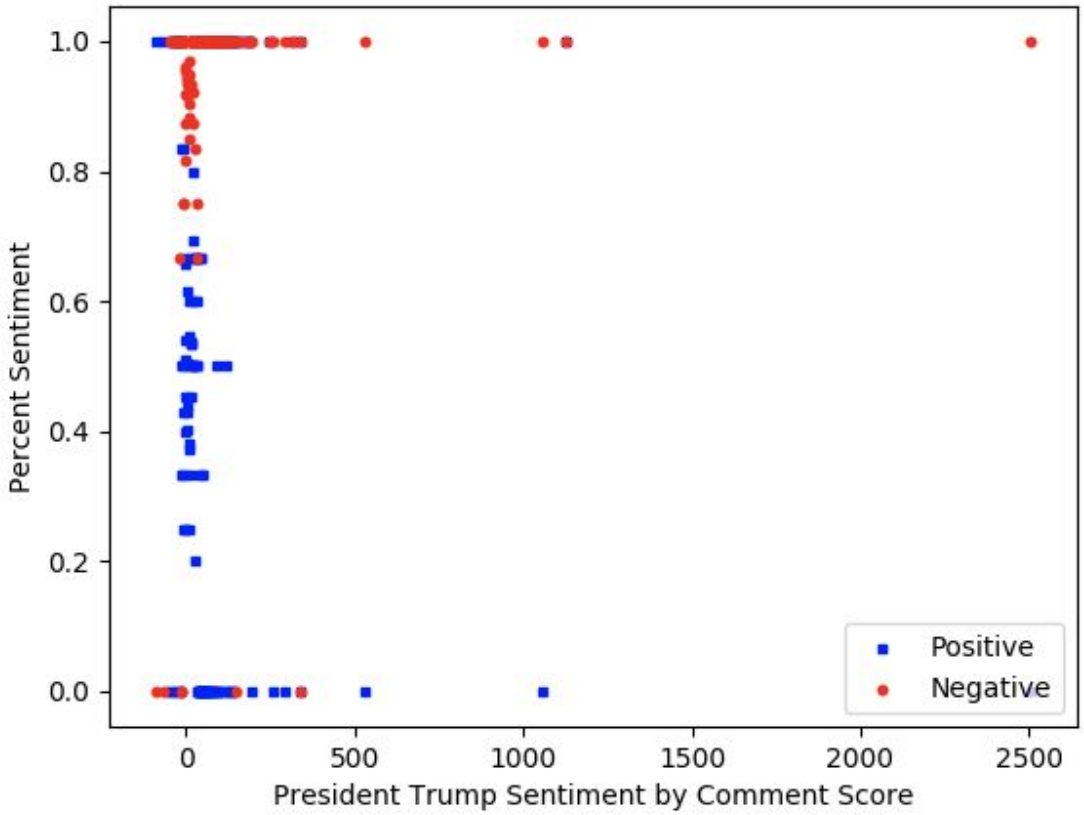
Negative Trump Sentiment Across the US



date


Difference Trump Sentiment Across the US
pos_pred
neg_pred
date


1.0
0.8
0.6
0.4
0.2
0.0
Percent Sentiment
0
20000
40000
60000
80000
100000
President Trump Sentiment by Submission Score
Positive
Negative



4th Graph

Top 10 Positive Stories:

- Pelosi Prevails: The Democratic leader beat back a challenge from Ohio Rep. Tim Ryan to remain minority leader
- BBC News: Ex-Trump aide Mike Flynn 'offered $15m by Turkey for Gulen
- Kirsten Gillibrand has voted against almost all of Donald Trump's nominees. 2020, anyone?
- Scientists have discovered two simple psychological differences that make you liberal or conservative
- Tapper calls out Trump on conspiracy theories
- Biden: 'I regret that I am not president' but 'it was the right decision' for family
- 6 senators voted to repeal Obamacare in 2015, but voted against repeal today
- TRUMP: I'm a 'smart person,' don't need intelligence briefings every single day
- Letters threatening genocide against Muslims and praising Trump sent to multiple California mosques
- House Republicans gut their own oversight

Top 10 Negative Stories:

- Jared Kushner failed to get a 'half-billion dollar investment' from a Qatari billionaire. Now he's 'hardening' America's stance towards Qatar.

- Walmart says it will raise age restriction to 21 for gun purchases, remove items resembling assault-style rifles from website
- 'Grab Him by the Mid-terms': Women's marches push power of the vote
- Equifax compromised half of the country’s information. Trump’s CFPB isn’t looking into it.
- Who is you 2020 presidential election nominee?
- Leaked communications show Reddit admin bias
- The most thorough, profound and moving defense of Hillary Clinton I have ever seen.
- AP source: Flynn agrees to provide documents to Senate panel
- The Evidence Inside The “House Intel Committee Four Page FISA Memo"
- 'It’s Outrageous': NYT Public Editor Liz Spayd Goes to Fox News, Chides Paper’s Liberal Bias

6.

From the
From Plot 5a, we can deduce that people who think negative about Trump have a high level of sentiment and people who think positive have a lower level of sentiment. This trend is alsot visible in the date plot with the positive and negative trends.