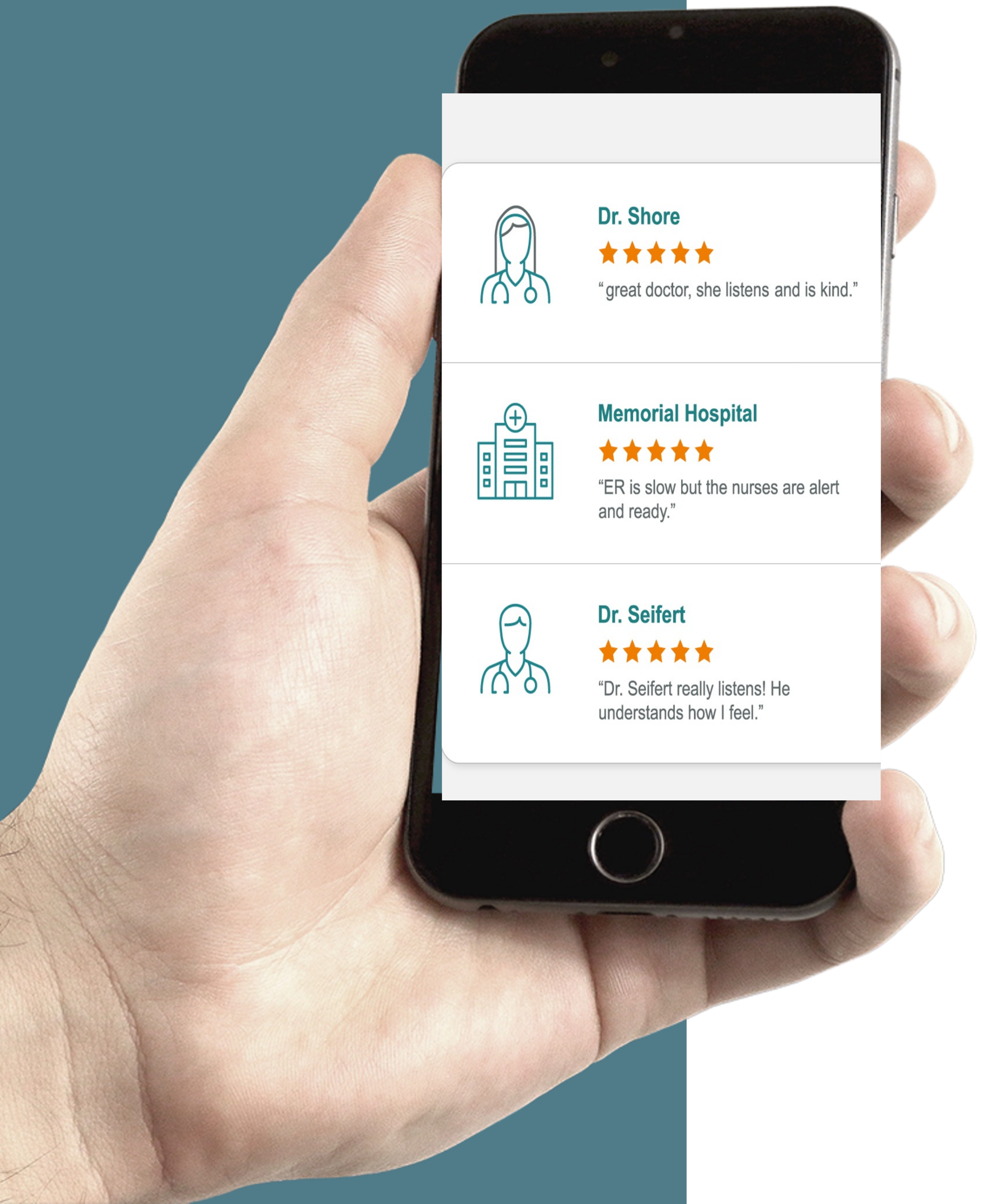
Do Patient Reviews Have Any Impacts On the Physician Sanctions?
Insights From Text-Mining Physician Reviews

<u>Group 5 Members</u>

Hoang Nha Nguyen (671491808)
Narayani Vadhuka (674778689)
Preethi Srinivasan (663981973)
Vinyl Anipe (670042568)



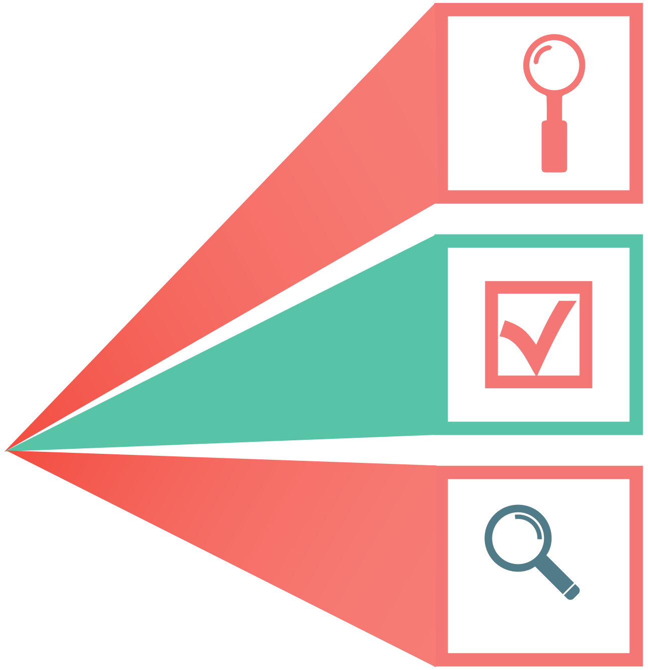
Agenda

1 Datasets
2 Text-Mining Methods
3 Sentiment Analysis - VADER
4 Topic Modeling - COREX
5 Topic Modelling - Zero Shot
6 Data Analysis and Modeling
7. Summary/ Insights

Project Goals

Problem Statement

Doctors can be disciplined for criminal convictions, medical negligence, wrongly prescribing controlled substances and other wrongdoing. However, many doctors typically have no more than a few reviews on a site, and the reviews may not accurately reflect the doctor's qualifications, personality or the patient experience.



To investigate the connection between the doctor sanctions and patient reviews.

To determine the sentiment of all the reviews and to identify the impact of each bucket on the doctor's ratings and their average sentiment scores

To examine the relationship between the ratings given to Medical Physicians on RateMDs and their patient reviews.

Project Workflow

Data Collection
Compile a list of fraudulent doctors in Illinois and Michigan states with the reasons for their sanction.

847 Doctors

Data Filtering
Consider only doctors who can be tracked by ID on RateMds (with URL).

449 Doctors

Data Scraping
Obtain doctors' ratings and every single review from each patient from RateMDs.

1712 Reviews

Sentiment Analysis
Remove null values.
Divide texts into positive and negative buckets; with the computed sentiment scores using VADER

1662 Reviews

Text Pre-process
Remove duplicate doctors after grouping by doctor and reviews. Then apply text preprocessing techniques.

202 Doctors

Topic Modeling
Obtain the topic strengths of the reviews. Also, perform Topic Modeling on the sanctioned reason using zeroshot classification

202 Doctors

EDA
Performed ANOVA and T test on significant variables and examine any relevant relationships.

202 Doctors

Data Modeling
Performed linear regression model

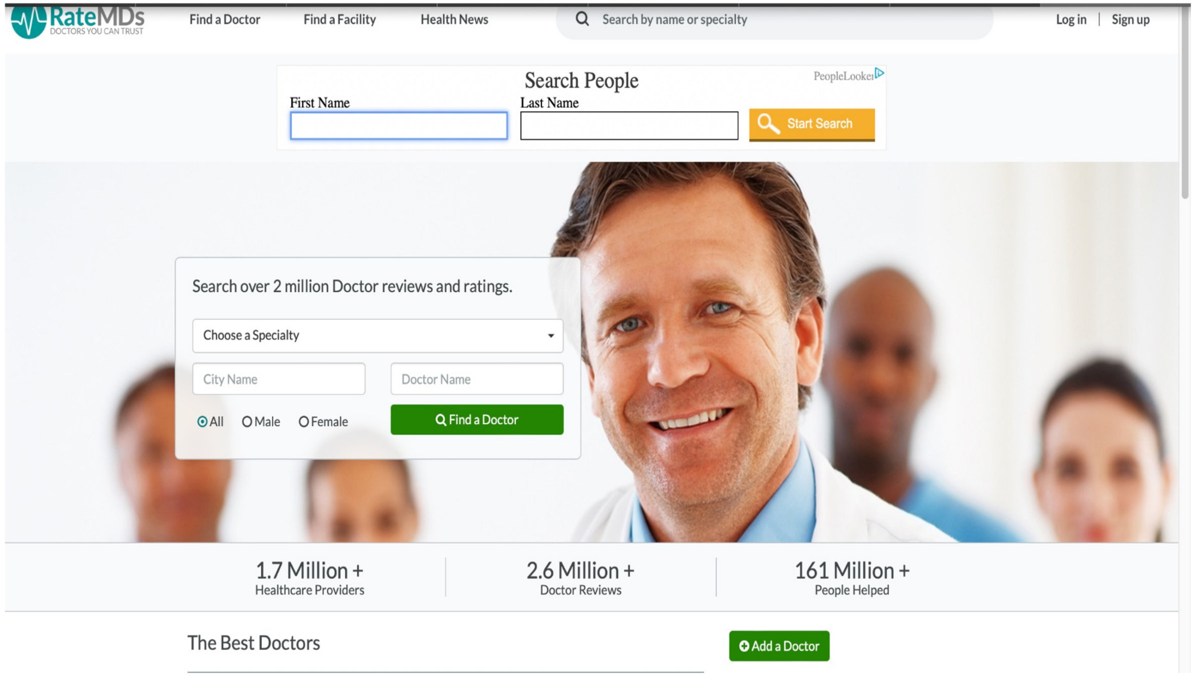
202 Doctors

Background Information

❖ ProPublica
- ➢ Provides helpful resources for researching doctors, and that we are informed that a doctor has been sanctioned by a professional licensing board
- ➢ To find information about a doctor in the state medical board license lookup page

State (click to look up a doctor)	Full disciplinary history available?	How to contact state medical boards	Lists of recent board actions
Alabama	Yes	For more information, contact Alabama's medical licensure commission at 334-242-4153 or bme@albme.org.	List of recent disciplinary actions
Alaska	No	Disciplinary documents are not available online. To access doctors' records, contact Alaska's medical licensing department at 907-269-8163 or debora.stovern@alaska.gov.	Lists of disciplinary actions since 1985
Arizona	Yes	For additional information, contact Arizona's medical board at 480-551-2700 or Arizona's Board of Osteopathic Examiners at 480-657-7703. For recent actions from the Board of Osteopathic Examiners, visit their central database.	List of recent disciplinary actions
Arkansas	No	Disciplinary documents are not available on Arkansas' medical board website. To access the documents, contact look up Arkansas' medical board at 501-296-1802.	List of recent disciplinary actions
California	Yes	Disciplinary documents for recent violations are available on each doctor's license verification page. For older violations, contact California's medical board at 916-263-2525 or complete the online request form. For recent actions from the California's Osteopathic Examiners Board, visit their central database.	List of disciplined doctors
Colorado	Yes	Most disciplinary documents are available on each doctor's license verification page. If a document is missing, contact Colorado's medical board at 303-894-7690 or DORA_MedicalBoard@state.co.us.	Lists of disciplinary actions

❖ RateMDs
- ➢ Online physician rating websites provide an open-end textual platform for patients to evaluate physicians based on their qualifications, personality and experiences.



❖ Many doctors continue to practice, some even changing states to do so, regardless of the fraudulent disciplinary record.

❖ One doctor can be listed more than once due to multiple sanction reasons.

Data Cozllection *PHASE 1*

- ❖ Review past disciplinary action from the **Fiscal Year 21/22** reports for **Illinois** and **Michigan** states according to ProPublica.
- ❖ The reports consisted of PDF documents that contained lists of all sanctioned physicians.
- ❖ We are targeting our research to physicians under Medicine Profession

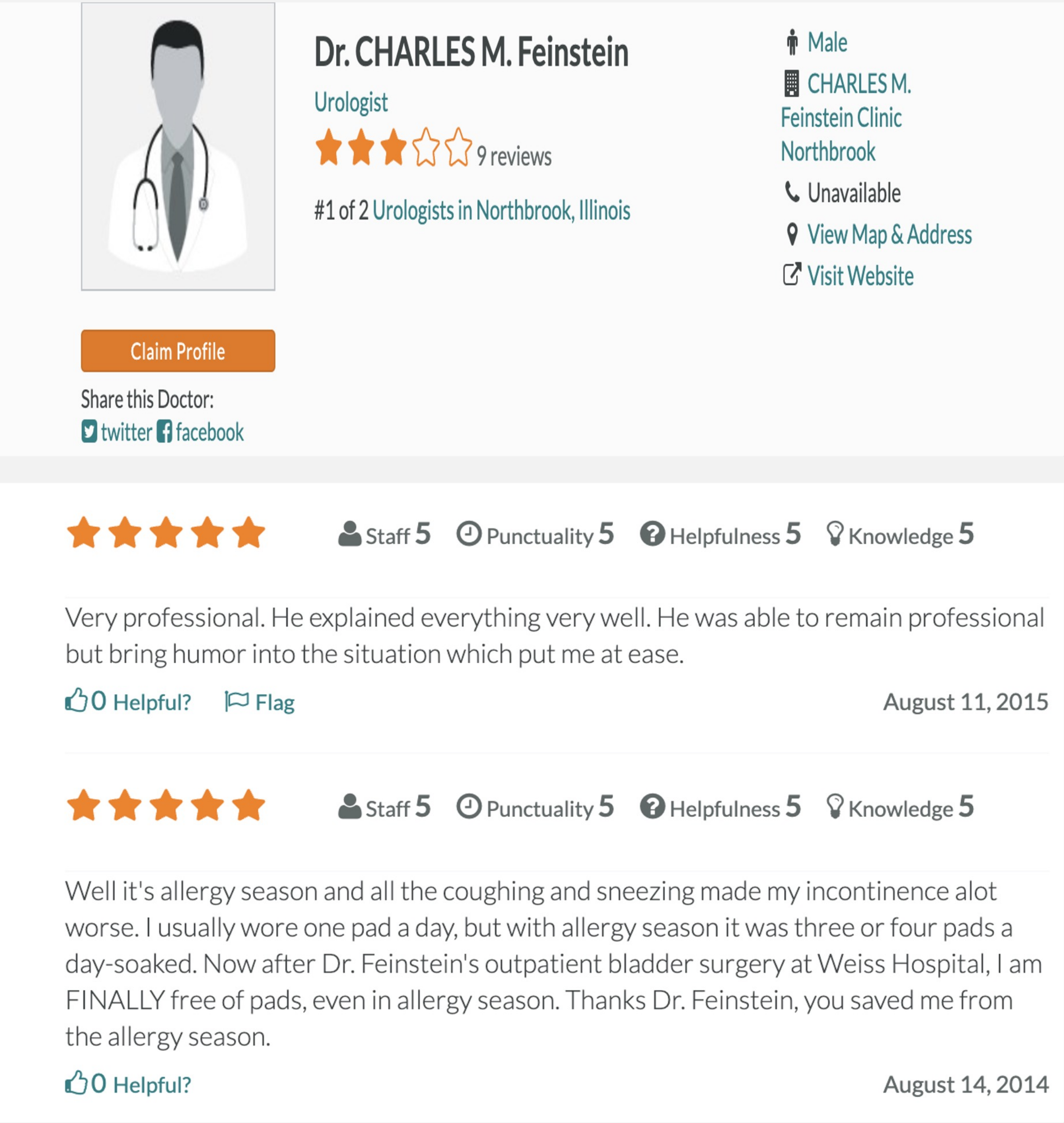
Profession/Name	License Number	Effective Date/Action	Basis for Action
Medicine ELLEN M JANSYN, M.D. Holland, MI	4301091388	08/20/2021 Fine Imposed; Probation	Continuing Education
Medicine NIDAL JI JBOOR, M.D. Warren, MI	4301082688	08/20/2021 Fine Imposed; Probation	Negligence-Incompetence; Technical Violation of the Michigan PHC
Medicine VICTOR JIMENEZ, M.D. Grand Rapids, MI	4301086868	07/21/2021 Fine Imposed; Summary Suspension Dissolved; Suspended	Sexual Misconduct; Incompetence; Lack of Good Moral Character; Negligence-Impaired Conduct, Practice, or Condition
Medicine ACHIN KIM, M.D. Rochester Hills, MI	4301055629	08/21/2021 Fine Imposed; Probation	Continuing Education

Data Collection *PHASE 2*

- ❖ A list of sanctioned doctors and their corresponding RateMDs URL links were manually recorded (**449 records**).
  - ➢ Among 237 records from IL in 2021, 79 records are not in rateMDs, **158** records have URL on rateMDs.
  - ➢ Among 221 records from IL in 2022, 65 records are not in rateMDs, **156** records have URL on rateMDs.
  - ➢ Among 208 records from MI in 2021, 122 records are not in rateMDs, **86** records have URL on rateMDs.
  - ➢ Among 182 records from MI in 2022, 133 records are not in rateMDs, **49** records have URL on rateMDs.
- ❖ **A total of 1662 rows of every review text for each doctor were scrapped, which included reviews and ratings of the sanctioned doctors listed in the manually created list.**

  Findings:
- ❖ 1662 records with every review of each doctor
- ❖ The information scraped from RateMDs include the following components: doctor name, gender, specialty, overall rating, number of reviews, review date, text of review.

Web Scraping



reviews and ratings from RateMds

- Import required libraries:
  - requests - used for requesting the content of the webpage.
  - BeautifulSoup - to parse the data from HTML files
  - Selenium - automate the process of accessing web browser via Python
- Create a dataframe used to store the scraped data
- Enter the URL of the page to be scrapped
- Set the number of pages to be scrapped.
- Scraped the following components:
  - Doctor name
  - Specialty
  - Overall Rating
  - Number of reviews
  - Review text
  - Review date
  - Gender

Natural Language Processing

Web Scraping Sample Codes

```python
# options.add_argument("--headless")
driver = uc.Chrome(options=options, executable_path=path)
# url=input('Please enter the URL of the page you want to scrape: ')
# n=input('How many pages do you want to Scrape? ')
df=pd.DataFrame(columns=['Doctor Name','Specialty','Overall Rating','No. of Reviews','Review', 'Review_Date','Review Rating'])
df2=pd.DataFrame(columns=['Doctor Name','Specialty','Overall Rating','No. of Reviews','Reviews', 'Review Rating'])
#url=input('Please enter the URL of the page you want to scrape: ')
#n=input('How many pages do you want to Scrape? ')
for url in data1['URL']:
    j=0
    reviews=[]
    review_dates=[]

#reviews=[]
#review_dates=[]
#https://www.ratemds.com/doctor-ratings/2288973/Dr-RICHARD+A.-FEELY-Chicago-IL.html/?page=
    for k in range(1,4):
        driver.get(url+"?page="+str(k))
        x=0
        while(x<1):
            time.sleep(2)
            try:
                parent_tags = driver.find_elements(By.CLASS_NAME, "rating-comment")
                for parent_tag in parent_tags:
                    try:
                        review_text = parent_tag.find_element(By.CSS_SELECTOR, "div.rating-comment-body").text
                    except (NoSuchElementException, StaleElementReferenceException):
                        review_text = ""
                    try:
                        review_date = parent_tag.find_element(By.CSS_SELECTOR, "p.rating-comment-created.pull-right").text.strip()
                    except (NoSuchElementException, StaleElementReferenceException):
                        review_date = ""
                    reviews.append(review_text)
                    review_dates.append(review_date)
            except StaleElementReferenceException:
                pass
            x+=1
```

```python
df3 = df2.assign(Review=reviews,Review_Date=review_dates)
soup=BeautifulSoup(driver.page_source,'lxml')
#doctor_name=soup.find('h1',{'itemprop':'name'})
#specialty=soup.find('div',{'class':'search-item-info'})
star_rating=soup.find('span',{'class':'star-rating'})
no_of_reviews=soup.find('span',{'itemprop':'ratingCount'})



#df3['Doctor Name']=doctor_name.text
#df3['Specialty']=specialty.text
try:
    specialty=soup.find('div',{'class':'search-item-info'})
    df3['Specialty']=specialty.text
    gender=soup.find('a',{'doctordetail':'.1.1.0.0.2.0.0.1'})
    df3['Gender']=gender.text
    no_of_reviews=soup.find('span',{'itemprop':'ratingCount'})
    df3['No. of Reviews']=no_of_reviews.text
    doctor_name=soup.find('h1',{'itemprop':'name'})
    df3['Doctor Name']=doctor_name.text
except AttributeError:
    df3['No. of Reviews']=0
    df3['Doctor Name']=0
    df3['Specialty']=0
df3['Overall Rating']=star_rating['title']
df=pd.concat([df, df3], axis=0)
print(df)
```

Dataset 1

- Dataset of 1712 rows of reviews obtained after scraping 449 sanctioned doctors with their corresponding RateMDs URL links.

```
df = pd.read_csv("FinalDataset.csv")
df
```

	Doctor Name	Specialty	Overall Rating	No. of Reviews	Review	Review_Date	Review Rating	Reviews	Gender
0	Dr. Emmanuel Paintsil	Internist / Geriatrician	1.25	1	DOES NOT CARE FOR THE PATIENTS NOR LONGEVITY O...	12-Mar-09	NaN	NaN	Male
1	Dr. Pramila Kasturi	Internist / Geriatrician	2.83	6	I had to switch doctors this year due to emplo...	3-Mar-16	NaN	NaN	Female
2	Dr. Pramila Kasturi	Internist / Geriatrician	2.83	6	She is the really bad dr don't know what she i...	26-Jun-15	NaN	NaN	Female
3	Dr. Pramila Kasturi	Internist / Geriatrician	2.83	6	Dr.Kasturi was very helpful and answered all m...	26-Jan-15	NaN	NaN	Female
4	Dr. Pramila Kasturi	Internist / Geriatrician	2.83	6	Office very friendly ,able to get an appointme...	17-Jun-14	NaN	NaN	Female
...	...	...	...	...	...	...	...	...	...
1707	Dr. Shakeeb Chinoy	Pediatrician	5.00	5	superb physician. thorough in exam, always rig...	8-Oct-17	NaN	NaN	Male
1708	Dr. Shakeeb Chinoy	Pediatrician	5.00	5	polite, gentle and caring. our kids like seein...	4-Mar-17	NaN	NaN	Male
1709	Dr. Shakeeb Chinoy	Pediatrician	5.00	5	we are very pleased with Dr. Chinoy. He has an...	19-Feb-17	NaN	NaN	Male
1710	Dr. Shakeeb Chinoy	Pediatrician	5.00	5	excellent doctor. my kid love him. highly reco...	18-Feb-17	NaN	NaN	Male
1711	Dr. Sergio Ponze	Nephrologist	5.00	1	Dr. Ponze is wonderful. He is a good listener ...	13-May-16	NaN	NaN	Male

1712 rows × 9 columns

Dataset 2

- Remove null rows
- Dataset 2 is used for sentiment analysis:
  - Pre-processing texts at this stage may change the meaning of the review texts

```
df.drop(columns=['Review Rating', 'Reviews'], inplace = True)
```

```
df.isnull().sum()
```

```
Doctor Name        0
Specialty          0
Overall Rating     0
No. of Reviews     0
Review            36
Review_Date        1
Gender            15
dtype: int64
```

```
df.dropna(inplace = True)
```

```
df.isnull().sum()
df
```

	Doctor Name	Specialty	Overall Rating	No. of Reviews	Review	Review_Date	Gender
0	Dr. Emmanuel Paintsil	Internist / Geriatrician	1.25	1	DOES NOT CARE FOR THE PATIENTS NOR LONGEVITY O...	12-Mar-09	Male
1	Dr. Pramila Kasturi	Internist / Geriatrician	2.83	6	I had to switch doctors this year due to emplo...	3-Mar-16	Female
2	Dr. Pramila Kasturi	Internist / Geriatrician	2.83	6	She is the really bad dr don't know what she i...	26-Jun-15	Female
3	Dr. Pramila Kasturi	Internist / Geriatrician	2.83	6	Dr.Kasturi was very helpful and answered all m...	26-Jan-15	Female
4	Dr. Pramila Kasturi	Internist / Geriatrician	2.83	6	Office very friendly ,able to get an appointme...	17-Jun-14	Female
...	...	...	...	...	...	...	...
1707	Dr. Shakeeb Chinoy	Pediatrician	5.00	5	superb physician. thorough in exam, always rig...	8-Oct-17	Male
1708	Dr. Shakeeb Chinoy	Pediatrician	5.00	5	polite, gentle and caring. our kids like seein...	4-Mar-17	Male
1709	Dr. Shakeeb Chinoy	Pediatrician	5.00	5	we are very pleased with Dr. Chinoy. He has an...	19-Feb-17	Male
1710	Dr. Shakeeb Chinoy	Pediatrician	5.00	5	excellent doctor. my kid love him. highly reco...	18-Feb-17	Male
1711	Dr. Sergio Ponze	Nephrologist	5.00	1	Dr. Ponze is wonderful. He is a good listener ...	13-May-16	Male

1662 rows × 7 columns

VADER

- Sentiment analysis was performed on the reviews text of doctors.
- VADER was used as the sentiment analysis tool.
- It is a class of sentiment analysis methods that rely on lexicons of sentiment-related words.
- Each word in the lexicon is rated and provided with four scores for each text: positive, negative, neutral, and compound.
- The compound score is an aggregated score of the first three scores and is used to measure the overall sentiment of the reviews.
- Further based on the compound score we classified the reviews into positive and negative (>= 0.05 - positive and < 0.05 - negative).

Text Pre-processing

- Performed pre-processing of the reviews text by removing stop words, lemmatization, special characters, punctuations, and tokenization.

```
import pandas as pd
from nltk.sentiment.vader import SentimentIntensityAnalyzer
sent = SentimentIntensityAnalyzer()
polarity = [round(sent.polarity_scores(i)['compound'], 2) for i in df['Review']]
df['sentiment_score'] = polarity
df
```

```
def format_output(output_dict):
  polarity = "neutral"
  if(output_dict['compound']>= 0.05):
    polarity = "positive"
  elif(output_dict['compound']< 0.05):
    polarity = "negative"
  return polarity
def predict_sentiment(text):
  output_dict =  sia.polarity_scores(text)
  return format_output(output_dict)
df["vader_prediction"] = df["Review"].apply(predict_sentiment)
df
```

Compute and Classify Sentiment Compound Scores

Sentiment Analysis

❖ Dataset after computing VADER Scores

Overall Rating	No. of Reviews	Review	Review_Date	Gender	vader_prediction	sentiment_score
1.25	1	DOES NOT CARE FOR THE PATIENTS NOR LONGEVITY O...	12-Mar-09	Male	negative	-0.03
2.83	6	I had to switch doctors this year due to emplo...	3-Mar-16	Female	positive	0.88
2.83	6	She is the really bad dr don't know what she i...	26-Jun-15	Female	negative	-0.58
2.83	6	Dr.Kasturi was very helpful and answered all m...	26-Jan-15	Female	positive	0.88
2.83	6	Office very friendly ,able to get an appointme...	17-Jun-14	Female	positive	0.54
...	...	...	...	...	...	...
5.00	5	superb physician. thorough in exam, always rig...	8-Oct-17	Male	positive	0.74
5.00	5	polite, gentle and caring. our kids like seein...	4-Mar-17	Male	positive	0.87
5.00	5	we are very pleased with Dr. Chinoy. He has an...	19-Feb-17	Male	positive	0.92
5.00	5	excellent doctor. my kid love him. highly reco...	18-Feb-17	Male	positive	0.89
5.00	1	Dr. Ponze is wonderful. He is a good listener ...	13-May-16	Male	positive	0.87

- After obtaining each vader sentiment score of the reviews, we aggregate reviews by positive and negative for each doctor.
- So, each doctor will have either one bucket of positive reviews or negative ones. They can also have both buckets

```
# group the DataFrame by Doctor and Review
grouped = df.groupby(['Doctor Name', 'vader_prediction'])

# aggregate the Comment column using the join() function to combine all comments into one row
aggregated = grouped.agg({'Review': lambda x: ' '.join(x),'vader_prediction': 'count'})

# reshape the DataFrame using the unstack() function to create separate rows for positive and negative re
reshaped = aggregated.unstack()

# rename the columns to remove the multi-index
result = pd.DataFrame(reshaped.to_records()).reset_index(drop=True)


result = result.merge(df[['Doctor Name', 'Specialty','Gender','No. of Reviews']], on='Doctor Name')
```

```
result
```

	Doctor Name	('Review', 'negative')	('Review', 'positive')	('vader_prediction', 'negative')	('vader_prediction', 'positive')	Specialty	Gender	No. of Reviews
0	Dr. ABDULMASSIH Abdulmassih	NaN	Dr. Abdulmasshi is the best doctor out there. ...	NaN	3.0	Family Doctor / G.P.	Male	3
1	Dr. ABDULMASSIH Abdulmassih	NaN	Dr. Abdulmasshi is the best doctor out there. ...	NaN	3.0	Family Doctor / G.P.	Male	3
2	Dr. ABDULMASSIH Abdulmassih	NaN	Dr. Abdulmasshi is the best doctor out there. ...	NaN	3.0	Family Doctor / G.P.	Male	3
3	Dr. Ahmad A. Shaher	NaN	Dr shaher is an awesome human being and a doct...	NaN	2.0	Family Doctor / G.P.	Male	2
4	Dr. Ahmad A. Shaher	NaN	Dr shaher is an awesome human being and a doct...	NaN	2.0	Family Doctor / G.P.	Male	2
...	...	...	...	...	...	...	...	...
1657	Dr. Zachary Solomon	NaN	Dr. Solomon is a fantastic doctor who been str...	NaN	1.0	Psychiatrist	Male	1
1658	Dr. Zulfiqar Ahmed	I waited months to get into him. I wrote down ...	Best treatment for bipolar and depression. I f...	2.0	2.0	Psychiatrist	Male	2
1659	Dr. Zulfiqar Ahmed	I waited months to get into him. I wrote down ...	Best treatment for bipolar and depression. I f...	2.0	2.0	Psychiatrist	Male	2
1660	Dr. Zulfiqar Ahmed	I waited months to get into him. I wrote down ...	Best treatment for bipolar and depression. I f...	2.0	2.0	Psychiatrist	Male	2
1661	Dr. Zulfiqar Ahmed	I waited months to get into him. I wrote down ...	Best treatment for bipolar and depression. I f...	2.0	2.0	Psychiatrist	Male	2

1662 rows × 8 columns

❖ Example of one CONCATENATED positive bucket for Dr. Zulfiqar Ahmed.

DR.V AND STAFF ARE VERY KIND, CARING ,YOU DON'T WAIT LONG TO SEE DOCTOR El Paso is lucky to have this doctor. He is so good at what he does. A true expert in the field. Dr V is a true human being. professional, kind and caring. Dr. Bratislav Velimirovich is one of the best doctors I have known in my 66 years of life. His primary goal is to relieve his patients of their pain. Not only is he knowledgeable and skillful, but he also sincerely cares about his patients.
I experienced chronic pain in my legs and back for several months. Dr. V. looked at my CT scan and recognized the problem immediately. He performed L4-L5 fusion in my lower back; this surgery eliminated my pain and restored my mobility.
I will always be grateful to Dr. V. for his expertise in making the correct diagnosis so quickly and performing such intricate surgery successfully.
Thank you, Dr. V.
With much gratitude,
Anne P. best doctors experience ever. he is extremely talented and staff is going out of their way He is very professional and courteous his staff is very helpful The best neurosurgeon in El Paso. He does minimally invasive spine surgery and minimally invasive brain aneurysm coiling My grandson (age 18) fell on trampoline and fox C5-6. Had surgery immediately in am. Had a very good recovery and has been great since. I feel his training and service that he has done in Chicago has brought new techniques to him and the patients have benefitted. Great doctor he is the best trained neurosurgeon in town. He has training in stroke as well Dr. V is the best neurosurgeon in town and has done an excellent job on me and my mom previously. However, ever since his former NP and medical assistant left, his east side office is not the same. My experience so far with the new people there is uncomfortable and wish they hadn‚Äôt left. Not the same. Dr. Velimirovic is one of the most intelligent and compassionate individuals I have been blessed to cross paths with. He performed surgery on me in June 2014. He is by far one of the best neurosurgeons in USA if not the globe. Dr. Velimirovic is passionate about his job and genuinely cares for his patients! I was so lucky to have him as my doctor! Thanks Dr. Velimirovic.
Sincerely,

❖ Example of one CONCATENATED negative bucket for Dr. Edward Paloyan

This doctor wants to do surgery, adn will charge you twice what the insurance company will pay.

He does not care about your health. I had the WORST experience with this doctor. After discussing my symptoms, he said he‚Äôd test me for Hashimoto disease but my symptoms were exactly opposite of Hashimoto disease. Was he thinking about dinner or whatever as he obviously wasn‚Äôt listening. I have been a patient for 14 yrs. I was happy until this year. I was not feeling well and was loosing lots of hair. I called the office to ask if I could get checked. I was told it was my age (46) and turned away. Another one of my doctors did blood work changed my meds. I felt better and my hair loss stopped. I went in for my annual ultrasound and check up. He was upset with the meds my other doctor prescribed and said that I am lucky I am not in a morgue from cardiac arrest. He insisted very egotisically that my problems were from my age and when I asked questions he told me the thyroid is too complicated for me to understand. That was my last appointment with this sexist, egotistical, hard of hearing doctor. He is the one who is old. I have been a patient of Dr. Paloyan since early 2000 and was very pleased with the outcome of my diagnosis with him that was mis-diagnosed by a previous Dr. and I can not express how grateful I am to him for that. I do have to agree with some of the comments about him not listening to what you say about how you feel. I am tired of him telling me I am getting older. The thyroid is a tiny organ that controls so much and without one and relying on a pill I don't always feel is enough. My levels are fine but what about what my body is saying?? Is there something else that I should be doing:(I have been tired and have had no energy and suffer from depression since I had my thyroid removed. I wish there was a transplant!! I stopped going to Dr. Paloyan because I felt he was not listening to my symptoms and issues I had related to my hypothyroidism. I was very disappointed seeing that he got such good reviews here. The office staff is really unpleasant. After you see the doctor you are sent to see the billing department at which time they try to get money from you. I had a big issue with incorrect billing to insurance I brought in all the insurance paperwork but in the end I ended up paying what they wanted, which was incorrect. I had to take blood tests every few months which gets to be really expensive even with insurance. The thyroid medicine the doctor gave me is expensive and hard to get. The pharmacist even questioned why he would be prescribing this particular medicine. I decided not to return after approximately a year and now am monitored by my regular internal medicine doctor. Sad is my major comment - i read all these others but i did not have this experience and i've been going for several years. He completely ignores anything i say about how i feel, he only checks my levels once a year even if i change meds or dosage - never ever checked bone density and i'm in menopause. he refused to discuss how i have no energy and my bones hurt, i started getting broken bones, my eyes lids were swelling, weight gain...he said the levels were fine and i didn't need anything else. SOOOOOO DISAPPOINTED!!!!!! DURING MY TREATMENT FOR AUTOIMMUNE HASHIMOTO/'S THYROIDITIS, TUMORS BEGAN TO ENLARGE AND BLOOD WORK AND BIOPSY INDICATED CHANGES. PARTIAL THYROIDECTOMY WAS RECOMMENDED. I WENT TO ANOTHER ENDO DR AND TRIED TO GET A 2ND OPINION. THE 2ND DR WAS VERBALLY ABUSIVE TOWARD DR. PALOYAN AND TOLD ME HE WAS TOO AGGRESSIVE IN TREATMENT. HE REVIEWED MY ENTIRE MEDICAL FILE AND STILL DID NOT CONCUR RE SURGERY. HE WANTED ME TO OBTAIN FOR HIM THE ACTUAL SLIDE OF THE BIOPSY (WE ALL KNOW HOW POSSIBLE THAT IS). I CALLED 2ND DR/'S OFFICE SEVERAL TIMES AND HE NEVER GOT BACK TO ME. DATE OF SURGERY ARRIVED AND I UNDERWENT A SUCCESSFUL PROCEDURE. THERE WAS CANCER FOUND. IT HAD NOT SPREAD. TODAY I AM STILL SEEING DR. PALOYAN AND VERY HAPPY WITH ALL FACETS OF HIS TREATMENT.

Dataset 3

- Removed all rows with duplicate doctor names after concatenating the reviews to obtain the final dataset of **202 unique doctors.**

```
[ ] result = result.drop_duplicates()
```

```
[ ] result
```

	Doctor Name	('Review', 'negative')	('Review', 'positive')	('vader_prediction', 'negative')	('vader_prediction', 'positive')	Specialty	Gender	No. of Reviews
0	Dr. ABDULMASSIH Abdulmassih	NaN	Dr. Abdulmasshi is the best doctor out there. ...	NaN	3.0	Family Doctor / G.P.	Male	3
3	Dr. Ahmad A. Shaher	NaN	Dr shaher is an awesome human being and a doct...	NaN	2.0	Family Doctor / G.P.	Male	2
5	Dr. Alan L. Sisson	I am not sure how how doctor stays in business...	NaN	1.0	NaN	Emergency Room Doctor	Male	1
6	Dr. Ali Kafi	Stay away from this doctor (and the term is us...	NaN	2.0	NaN	Cardiothoracic Surgeon	Male	1
8	Dr. Allan L. Tompkins	NaN	Dr. Tompkins is a great surgeon. He replaced m...	NaN	4.0	Orthopedic Surgeon	Male	3
...	...	...	...	...	...	...	...	...
1639	Dr. William Mikaitis	Prior to an incident in 2005, Dr. Mikaitis was...	Dr. William Mikaitis has been my family physic...	3.0	5.0	Family Doctor / G.P.	Male	8
1647	Dr. Won Sam Yi	Dr Yi treated me for breast cancer. I was seve...	Dr. Yi supervised my whole breast radiation at...	3.0	3.0	Radiologist	Male	6
1653	Dr. Yaseen K. Odeh	NaN	Simply one of the BEST DOCTORS IN THE NATION. ...	NaN	4.0	Internist / Geriatrician	Male	4
1657	Dr. Zachary Solomon	NaN	Dr. Solomon is a fantastic doctor who been str...	NaN	1.0	Psychiatrist	Male	1
1658	Dr. Zulfiqar Ahmed	I waited months to get into him. I wrote down ...	Best treatment for bipolar and depression. I f...	2.0	2.0	Psychiatrist	Male	2

202 rows × 8 columns

Topic Modeling - CorEX

- ➔ Correlation Explanation (CorEx) provides a flexible framework for learning topics that are maximally informative about a corpus of text.
- ➔ The CorEx model allows the incorporation of domain knowledge through user-specific anchor words which guide the model towards the topics of interest.
- ➔ This enables the model to represent those topics that do not naturally emerge and provides the ability to separate keywords allowing distinct topics to be identified.
- ➔ The CorEx model also has a strength parameter that defines the bias of the topics generated towards the anchor keywords.
- ➔ The anchor keywords are sets of keywords assigned to each topic for e.g in our project we have used anchor keywords like “Attitude”, “Patience”, “Appointment”, “Schedule”, “Late”, “Bills” etc.
- ➔ This value should always be above 1 and higher values indicate a stronger bias towards the anchor keywords.

Topic Modeling - CorEX

- COREX topic modeling approach was used to analyze reviews for each physician
- Proportion of each topic in the reviews for each physician was calculated
- Dependent variables such as Average Positive Score and Average Negative Score were calculated
- Topic modeling approach allows for identification of common themes in reviews for each physician

Topics

- Bedside Manners
- Communication
- Knowledge
- Office Environment
- Scheduling
- Wait Times
- Cost and Insurance

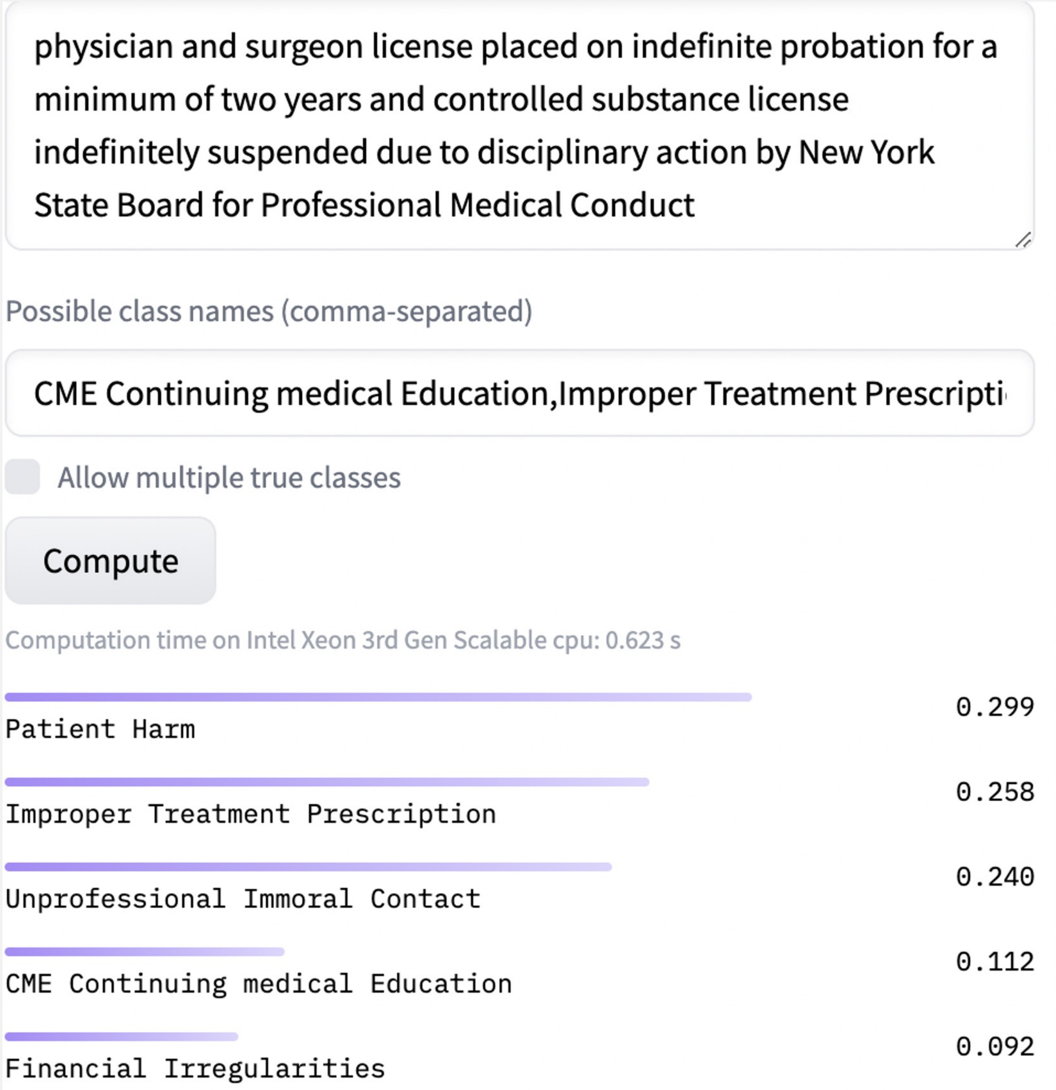
Zero-shot Classification

- Zero-shot text classification is a task in natural language processing where a model is trained on a set of labeled examples but is then able to classify new examples from previously unseen classes.
- Reasons for physician sanctions were identified using zero-shot classification.
- The technique allowed for automatic categorization of the reasons into predefined classes without any prior training or labeled data.
- The dataset includes information on the proportion of each reason for the sanctions.

<u>Topics</u>

- Improper treatment,
- Unprofessional Immoral conduct
- failure to pay income taxes
- CME Continuing Medical Education
- Patient Harm

- Converted the scored obtained into 0 and 1 by setting threshold value = 0.2.



Possible Sanction Reasons and computed scores

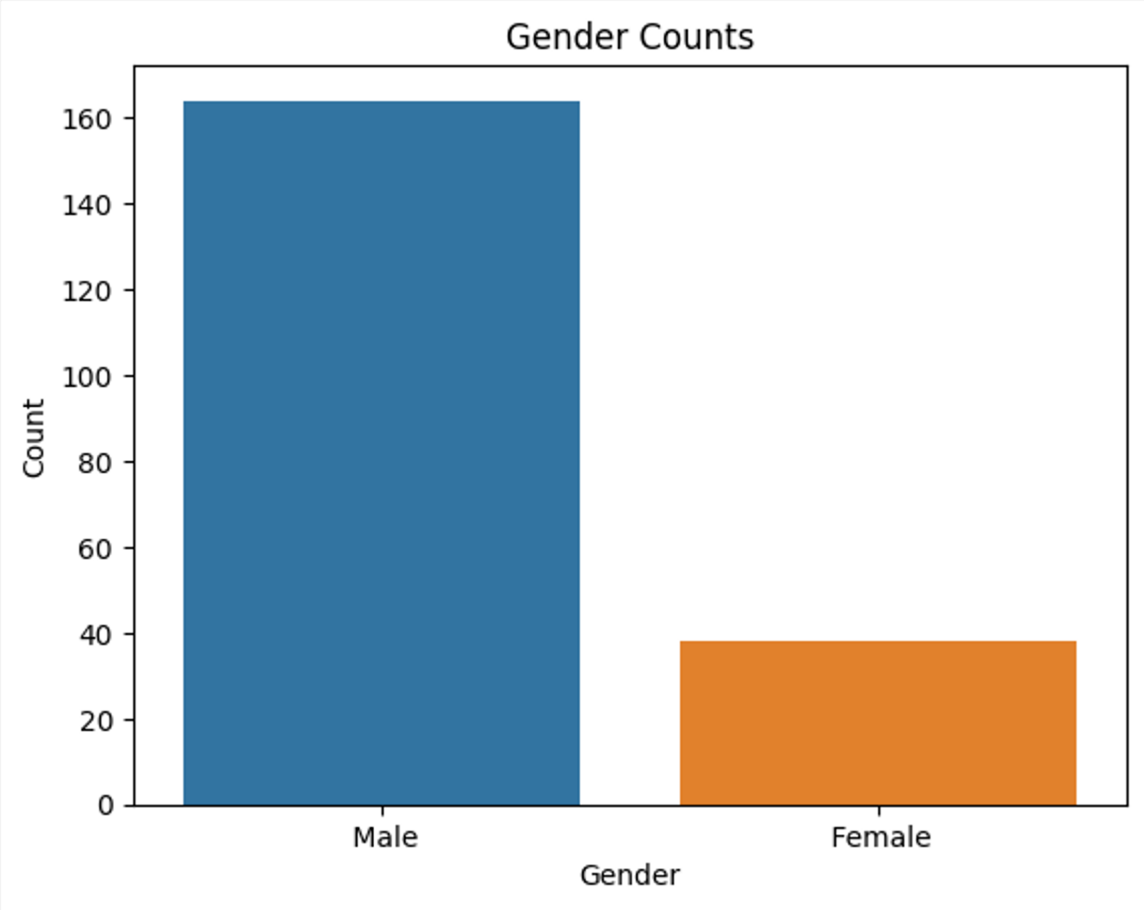
Topic Modeling

Final Dataset

- The final dataset obtained after sentiment analysis, topic modeling, and zero-shot classification contains **35 attributes and 202 data points**.
- Feature engineering was performed to create four additional attributes, contributing to the total number of attributes.
- The 35 attributes in the dataset represent a combination of sentiment scores, topic proportions, and demographic information for each data point.
- The four new attributes created through feature engineering include
  - **avg_pos_score**
  - **avg_neg_score**
  - **avg_count_pos_reviews**
  - **avg_count_neg_reviews**
- The final dataset is a comprehensive representation of the analyzed data and can be used for further analysis or modeling purposes.

```
> str(data)
'data.frame':	202 obs. of  36 variables:
 $ X                            : int  0 1 2 3 4 5 6 7 8 9 ...
 $ Doctor.Name                  : chr  "Dr. ABDULMASSIH Abdulmassih" "Dr. Ahmad A. Shaher" "Dr. Alan L. Sisson
i" ...
 $ State                        : chr  "Illinois" "Illinois" "Illinois" "California" ...
 $ Text_of_sanction             : chr  "physician and surgeon license reprimanded and fined $5,000 and must ta
five topic areas of Ethic"| __truncated__ "physician and surgeon license placed on indefinite probation for a min
rs," "physician and surgeon license and controlled substance license placed on permanent relinquished status afte
ted__ "Continuing Education" ...
 $ Gender                       : num  0 0 0 0 0 0 1 0 0 0 ...
 $ Overall_Rating               : num  4.58 4.75 1 1 3.83 1.84 5 3.75 3.65 3.63 ...
 $ Negative_Reviews             : chr  "" "" "I am not sure how how doctor stays in business. With manners lik
be a truck driver. No offense to the truckers." "Stay away from this doctor (and the term is used loosely. He per
y on my dear mother and he did t"| __truncated__ ...
 $ Positive_Reviews             : chr  "Dr. Abdulmasshi is the best doctor out there. Amazed. I have been goin
ssih for years ever since "| __truncated__ "Dr shaher is an awesome human being and a doctor. He is good teacher
tor. I have learnt so much "| __truncated__ "" "" ...
 $ No_of_Negative_Reviews       : num  NA NA 1 2 NA 7 NA 1 4 8 ...
 $ No_of_Positive_Reviews       : num  3 2 NA NA 4 2 1 NA 14 18 ...
 $ Total_Reviews                : int  3 2 1 1 3 9 1 2 18 13 ...
 $ PPNegative_1                 : chr  "" "nn" "sure doctor stay business manner ike truck iver offense trucke
octor term used loosely performed surgery dear mother incorrect surgery life miserable death answer "| __truncate
 $ PPPositive_1                 : chr  " best doctor amazed going   year ever since partner passed away inheri
now polite friendly import"| __truncated__ "wesome humn doctor good techer ptient doctor lernt much truly dess ev
ient strives provide best t"| __truncated__ "" "" ...
 $ n_score_bed_side             : num  1.55e-05 1.55e-05 1.55e-05 2.30e-04 1.55e-05 ...
 $ n_score_Knowledge            : num  6.11e-06 6.11e-06 6.11e-06 1.51e-04 6.11e-06 ...
 $ n_score_communication        : num  1.47e-05 1.47e-05 3.22e-04 3.52e-03 1.47e-05 ...
 $ n_score_environment          : num  2.48e-05 2.48e-05 2.48e-05 2.64e-05 2.48e-05 ...
 $ n_score_scheduling           : num  0.000112 0.000112 0.000112 0.000112 0.000112 ...
 $ n_score_wait_time            : num  0.000162 0.000162 0.00188 0.006773 0.000162 ...
 $ n_score_costs                : num  0.000001 0.002465 0.002466 0.002466 0.000001 ...
 $ avg_neg_score                : num  0.000048 0.0004 0.000689 0.0019 0.000048 1 0.000048 0.00567 0.145 1 ...
 $ p_score_bed_side             : num  1.0e-06 1.2e-06 1.0e-06 1.0e-06 1.0e-06 ...
 $ p_score_Knowledge            : num  2.88e-05 1.74e-06 1.00e-06 1.00e-06 2.04e-03 ...
 $ p_score_communication        : num  0.720607 0.000001 0.000001 0.000001 0.000143 ...
 $ p_score_environment          : num  9.50e-05 6.99e-03 1.00e-06 1.00e-06 3.01e-06 ...
 $ p_score_scheduling           : num  2.81e-02 8.47e-06 1.00e-06 1.00e-06 2.22e-04 ...
 $ p_score_wait_time            : num  7.40e-04 4.54e-04 2.73e-06 2.73e-06 3.10e-04 ...
 $ p_score_costs                : num  1.21e-03 1.08e-04 5.77e-06 5.77e-06 4.22e-04 ...
 $ avg_pos_score                : num  1.07e-01 1.08e-03 1.93e-06 1.93e-06 4.49e-04 7.19e-01 2.87e-01 1.93e-06
...
 $ CME.Continuing.medical.Education: int  1 1 1 1 1 1 1 1 1 1 ...
 $ Improper.Treatment.Prescription : int  1 1 1 0 0 1 1 1 1 0 ...
 $ Patient.Harm                 : int  1 1 1 0 0 1 1 0 1 0 ...
 $ Unprofessional.Immoral.Contact  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Financial.Irregularities     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ avg_count_pos_rev            : num  1 1 0 0 1.33 ...
 $ avg_count_neg_rev            : num  0 0 1 2 0 ...
```

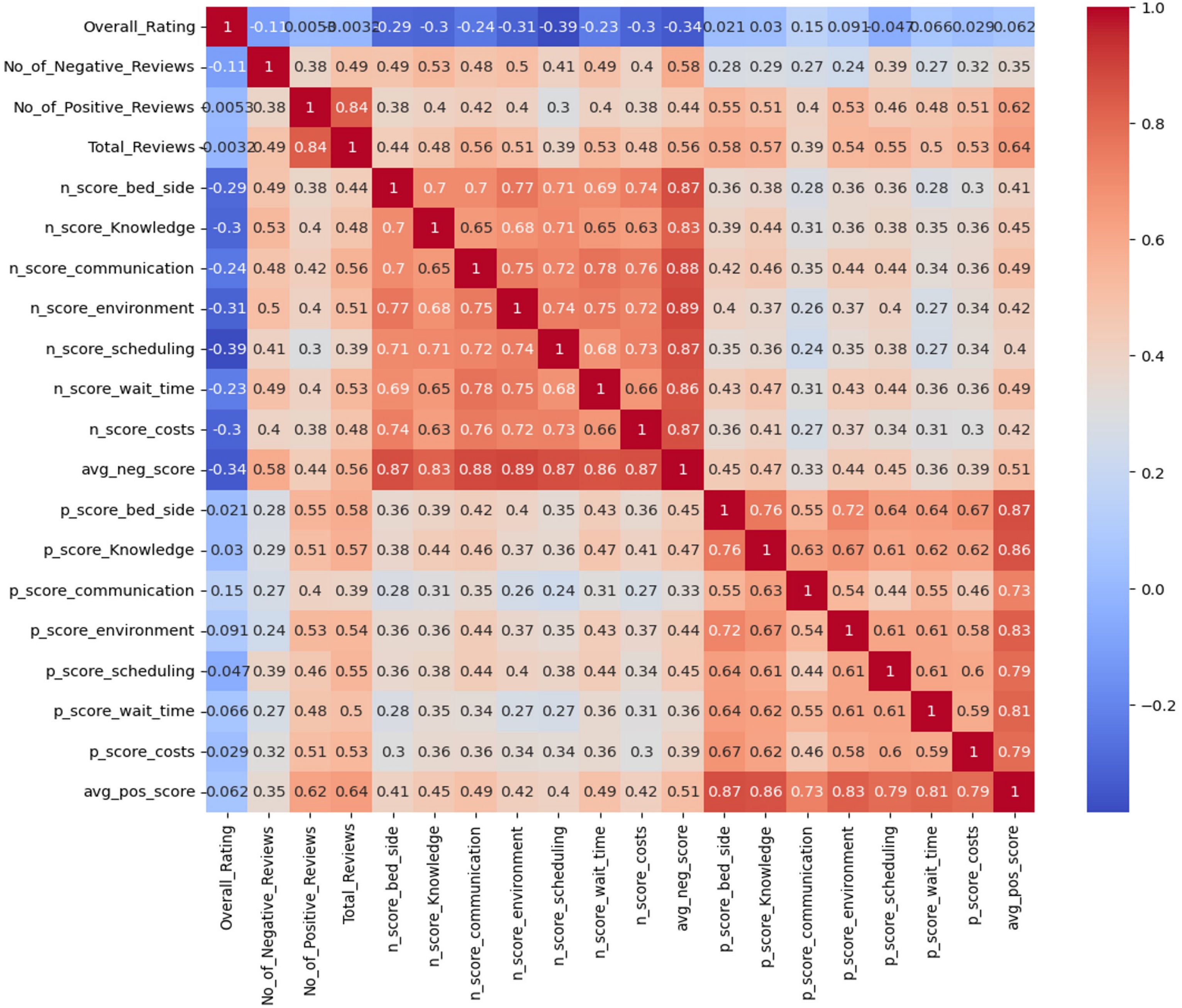
Univariate Analysis



- We have more Male physicians than female who have been sanctioned.

Correlation Matrix

- For each physician, the overall rating has relatively low negative correlations with all negative scores for common themes in reviews.
- Also, for each physician, the overall rating has relatively low positive correlations with all positive scores for common themes in reviews.



ANOVA (p=.05)

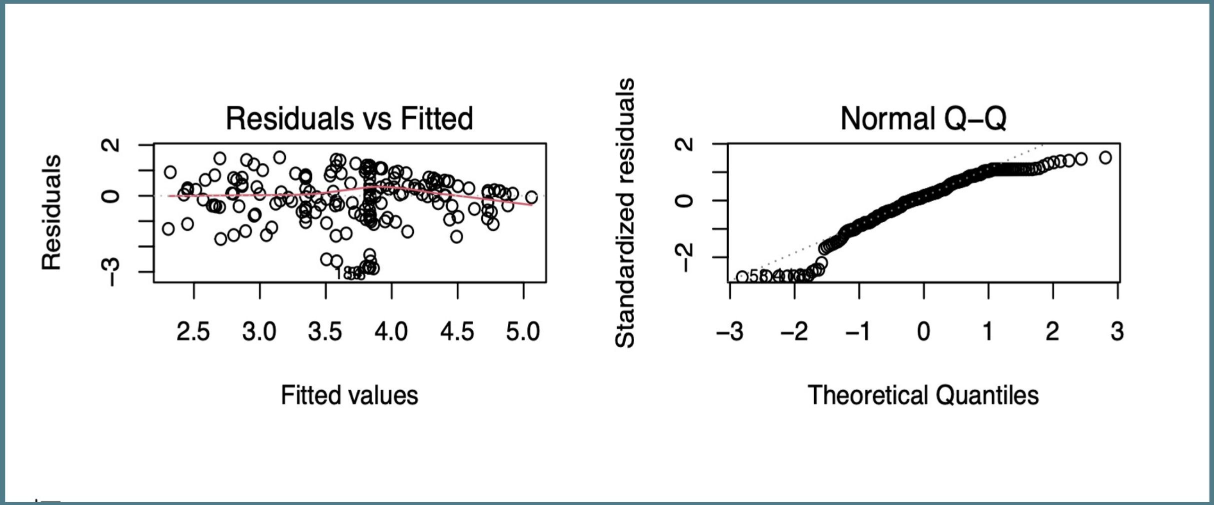
- Target variable is “**Overall_Rating**”
- We are examining the relationship between the physician overall ratings and sanction reasons including Improper treatment, Unprofessional Immoral conduct, CME Continuing Medical Education, Patient Harm, and Financial Irregularities
- At a significance level of 5%, there are no statistically significant differences between the means of the groups.

➔ No significant association between physicians’ sanction reasons and their overall ratings.

	df	Sum Square	Mean Square	F Value	Pr(>F)
Continuing Education Residuals	1 200	1.56 284.05	1.561 1.420	1.099	0.296
Improper Prescription Residuals	1 200	0.0 285.6	0.0038 1.4280	0.003	0.959
Patient Harm Residuals	1 200	0.51 285.10	0.5105 1.42l 55	0.358	0.55
Immoral Contact Residuals	1 200	0.09 285.91	0.0941 1.4276	0.066	0.798
Financial Irregularities Residuals	1 200	1.11 284.50	1.106 1.423	0.778	0.379

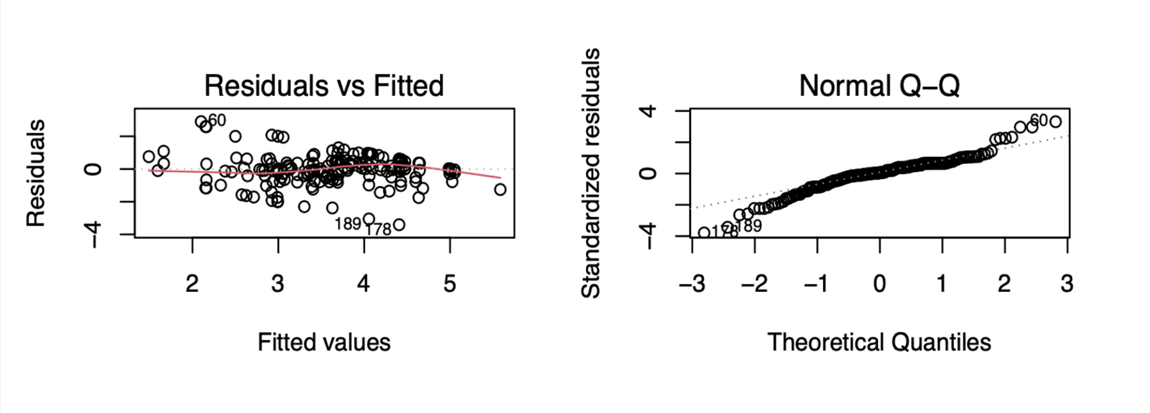
```
> overall<-subset(data, select = c("n_score_bed_side","n_score_Knowledge",
+                                  "n_score_communication","n_score_environment",
+                                  "n_score_scheduling","n_score_wait_time",
+                                  "n_score_costs","p_score_bed_side",
+                                  "p_score_Knowledge","p_score_communication",
+                                  "p_score_environment","p_score_scheduling",
+                                  "p_score_wait_time","p_score_costs",
+                                  "Overall_Rating"))
> overall_model = lm(overall$Overall_Rating~., data = overall)
> summary(overall_model)

Call:
lm(formula = overall$Overall_Rating ~ ., data = overall)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.8616 -0.5600  0.1545  0.8081  1.5101 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)            3.83354    0.11371  33.713  < 2e-16 ***
n_score_bed_side      -0.03506    0.29094  -0.121  0.90420    
n_score_Knowledge     -0.34095    0.26863  -1.269  0.20594    
n_score_communication  0.09458    0.30698   0.308  0.75836    
n_score_environment   -0.19684    0.31252  -0.630  0.52957    
n_score_scheduling    -0.78101    0.28486  -2.742  0.00671 ** 
n_score_wait_time      0.13212    0.29622   0.446  0.65610    
n_score_costs         -0.25478    0.28460  -0.895  0.37183    
p_score_bed_side      -0.15796    0.29858  -0.529  0.59741    
p_score_Knowledge      0.01208    0.28404   0.043  0.96613    
p_score_communication  0.48217    0.20772   2.321  0.02135 *  
p_score_environment    0.50580    0.25202   2.007  0.04619 *  
p_score_scheduling    -0.21563    0.23988  -0.899  0.36987    
p_score_wait_time      0.08366    0.23397   0.358  0.72106    
p_score_costs          0.18583    0.23625   0.787  0.43254    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.066 on 187 degrees of freedom
Multiple R-squared:  0.2553,	Adjusted R-squared:  0.1996 
F-statistic:  4.58 on 14 and 187 DF,  p-value: 4.163e-07
```

Linear Regression ~ Baseline Model 1



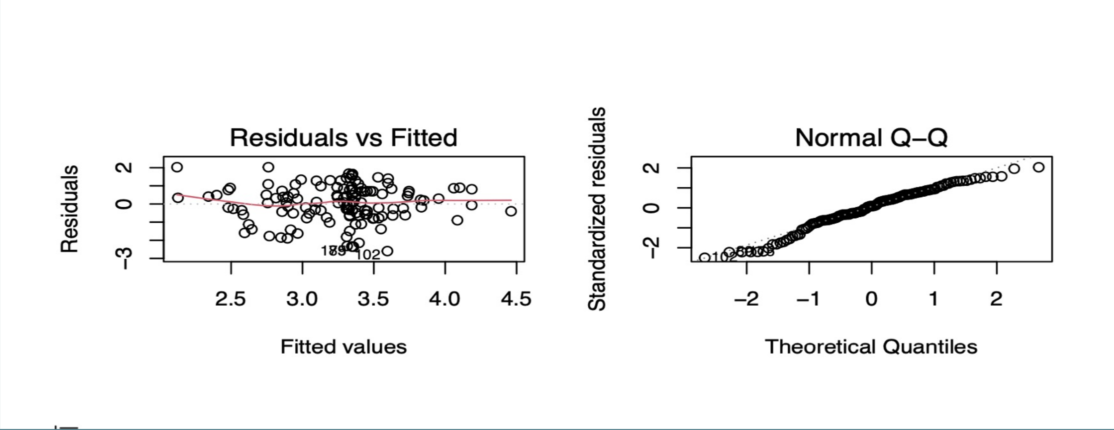
Linear Regression ~ Model 2

```
> data1<-subset(data, select = c("avg_count_pos_rev","avg_count_neg_rev","avg_neg_score"
+                               ,"avg_pos_score","Overall_Rating","Gender"))
> data1_model = lm(data1$Overall_Rating~., data = data1)
> summary(data1_model)

Call:
lm(formula = data1$Overall_Rating ~ ., data = data1)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.4067 -0.3820  0.0510  0.5441  2.9002 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)        3.82720    0.19564  19.562  < 2e-16 ***
avg_count_pos_rev  0.57951    0.14535   3.987 9.44e-05 ***
avg_count_neg_rev -0.83230    0.16419  -5.069 9.22e-07 ***
avg_neg_score     -0.49810    0.20996  -2.372   0.0186 *  
avg_pos_score      0.06843    0.21528   0.318   0.7509    
Gender            -0.35168    0.16405  -2.144   0.0333 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9048 on 196 degrees of freedom
Multiple R-squared:  0.4381,	Adjusted R-squared:  0.4238 
F-statistic: 30.57 on 5 and 196 DF,  p-value: < 2.2e-16
```



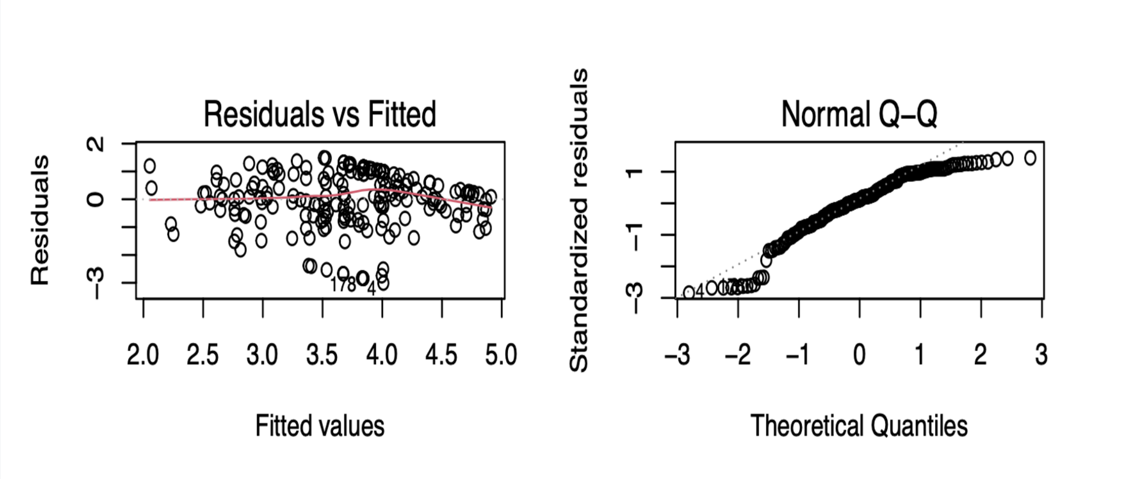
```
> data_neg<-subset(data, select = c("n_score_bed_side","n_score_Knowledge",
+                                   "n_score_communication","n_score_environment",
+                                   "n_score_scheduling","n_score_wait_time",
+                                   "n_score_costs","Overall_Rating",
+                                   "No_of_Negative_Reviews","Gender"))
> data_neg_model = lm(data_neg$Overall_Rating~., data = data_neg)
> summary(data_neg_model)

Call:
lm(formula = data_neg$Overall_Rating ~ ., data = data_neg)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.59496 -0.59225  0.08961  0.73439  2.02875 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)             3.38934    0.16954  19.992   <2e-16 ***
n_score_bed_side        0.10594    0.29549   0.359   0.7206    
n_score_Knowledge       0.01728    0.27421   0.063   0.9498    
n_score_communication   0.60369    0.30999   1.947   0.0538 .  
n_score_environment    -0.29287    0.31137  -0.941   0.3488    
n_score_scheduling     -0.73260    0.28492  -2.571   0.0113 *  
n_score_wait_time       0.27363    0.29868   0.916   0.3614    
n_score_costs           0.25055    0.29272   0.856   0.3937    
No_of_Negative_Reviews -0.04053    0.02462  -1.646   0.1023    
Gender                 -0.51440    0.25250  -2.037   0.0438 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.074 on 122 degrees of freedom
  (70 observations deleted due to missingness)
Multiple R-squared:  0.1357,	Adjusted R-squared:  0.07198 
F-statistic: 2.129 on 9 and 122 DF,  p-value: 0.03187
```

Linear Regression ~ Model 3



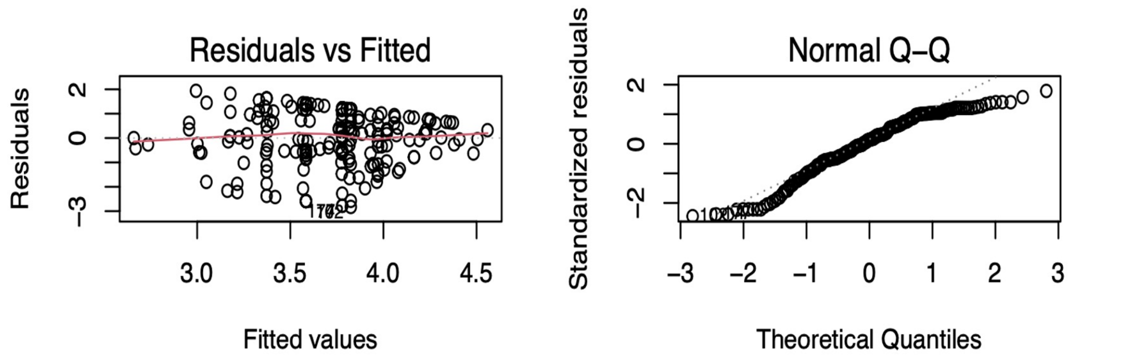
Linear Regression ~ Model 4

```
> all<-subset(data, select = c("Gender","n_score_bed_side","n_score_Knowledge",
+                              "n_score_communication","n_score_environment",
+                              "n_score_scheduling","n_score_wait_time",
+                              "n_score_costs","p_score_bed_side",
+                              "p_score_Knowledge","p_score_communication",
+                              "p_score_environment","p_score_scheduling",
+                              "p_score_wait_time","p_score_costs",
+                              "Overall_Rating","CME.Continuing.medical.Education",
+                              "Improper.Treatment.Prescription",
+                              "Patient.Harm","Unprofessional.Immoral.Contact",
+                              "Financial.Irregularities"))
> all_model = lm(all$Overall_Rating~., data = all)
> summary(all_model)

Call:
lm(formula = all$Overall_Rating ~ ., data = all)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.0089 -0.5579  0.1093  0.9000  1.4894 

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)                       3.572856   0.419383   8.519 6.11e-15 ***
Gender                           -0.331932   0.206199  -1.610   0.1092    
n_score_bed_side                  0.004221   0.301191   0.014   0.9888    
n_score_Knowledge                -0.303788   0.273106  -1.112   0.2675    
n_score_communication             0.172069   0.318172   0.541   0.5893    
n_score_environment              -0.266630   0.316918  -0.841   0.4013    
n_score_scheduling               -0.747944   0.288564  -2.592   0.0103 *  
n_score_wait_time                 0.116392   0.306423   0.380   0.7045    
n_score_costs                    -0.316569   0.287638  -1.101   0.2725    
p_score_bed_side                 -0.147704   0.299979  -0.492   0.6230    
p_score_Knowledge                 0.054449   0.290831   0.187   0.8517    
p_score_communication             0.416908   0.211970   1.967   0.0507 .  
p_score_environment               0.474495   0.256933   1.847   0.0664 .  
p_score_scheduling               -0.292952   0.245387  -1.194   0.2341    
p_score_wait_time                 0.168102   0.240874   0.698   0.4861    
p_score_costs                     0.188435   0.239046   0.788   0.4316    
CME.Continuing.medical.Education  0.435591   0.423709   1.028   0.3053    
Improper.Treatment.Prescription  -0.025681   0.250914  -0.102   0.9186    
Patient.Harm                     -0.140244   0.242518  -0.578   0.5638    
Unprofessional.Immoral.Contact   -0.104885   0.282120  -0.372   0.7105    
Financial.Irregularities          0.264608   1.125837   0.235   0.8144    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.071 on 181 degrees of freedom
Multiple R-squared:  0.2734,	Adjusted R-squared:  0.1931 
F-statistic: 3.405 on 20 and 181 DF,  p-value: 5.8e-06
```



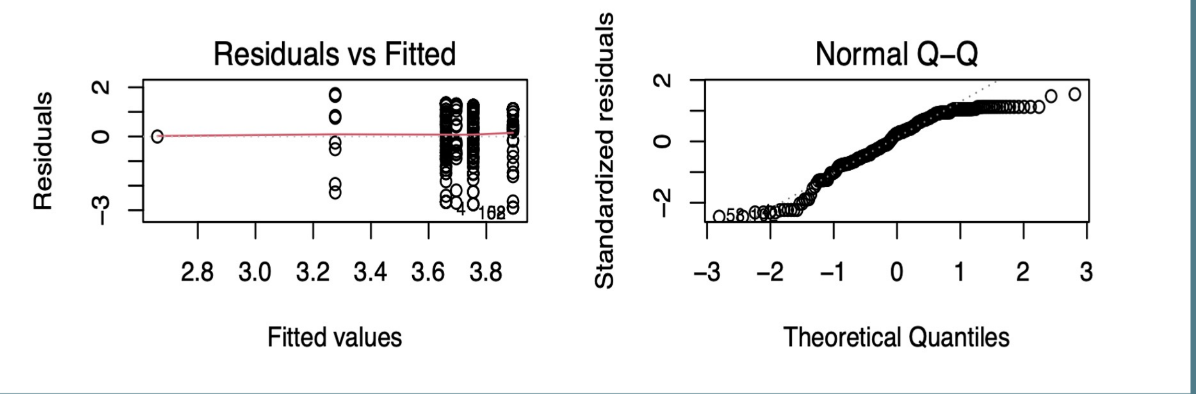
Linear Regression ~ Model 5

```
> df1<-subset(data, select = c("Gender","p_score_bed_side","p_score_Knowledge",
+                              "p_score_communication","p_score_environment",
+                              "p_score_scheduling","p_score_wait_time",
+                              "p_score_costs",
+                              "Overall_Rating","CME.Continuing.medical.Education",
+                              "Improper.Treatment.Prescription",
+                              "Patient.Harm","Unprofessional.Immoral.Contact",
+                              "Financial.Irregularities"))
> df1_model = lm(df1$Overall_Rating~., data = df1)
> summary(df1_model)

Call:
lm(formula = df1$Overall_Rating ~ ., data = df1)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.8249 -0.6184  0.1507  0.9614  1.9361 

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                       3.21519    0.44252   7.266 9.62e-12 ***
Gender                           -0.40587    0.22031  -1.842   0.0670 .  
p_score_bed_side                 -0.19384    0.32700  -0.593   0.5540    
p_score_Knowledge                -0.17272    0.30834  -0.560   0.5760    
p_score_communication             0.41381    0.22990   1.800   0.0735 .  
p_score_environment               0.35905    0.27989   1.283   0.2011    
p_score_scheduling               -0.53360    0.26266  -2.032   0.0436 *  
p_score_wait_time                 0.23307    0.26119   0.892   0.3733    
p_score_costs                     0.08673    0.26065   0.333   0.7397    
CME.Continuing.medical.Education  0.56361    0.45175   1.248   0.2137    
Improper.Treatment.Prescription   0.04605    0.26695   0.173   0.8632    
Patient.Harm                     -0.24082    0.25859  -0.931   0.3529    
Unprofessional.Immoral.Contact   -0.01563    0.30631  -0.051   0.9594    
Financial.Irregularities         -0.90841    1.21393  -0.748   0.4552    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.178 on 188 degrees of freedom
Multiple R-squared:  0.08666,	Adjusted R-squared:  0.0235 
F-statistic: 1.372 on 13 and 188 DF,  p-value: 0.1761
```



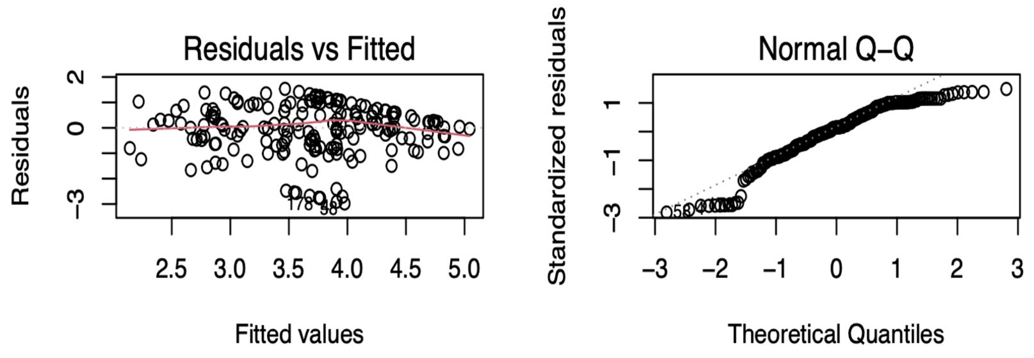
Linear Regression ~ Model 6

```
> #sanc
> df2<-subset(data, select = c(
+                             "Overall_Rating","CME.Continuing.medical.Education",
+                             "Improper.Treatment.Prescription",
+                             "Patient.Harm","Unprofessional.Immoral.Contact",
+                             "Financial.Irregularities"))
> df2_model = lm(df2$Overall_Rating~., data = df2)
> summary(df2_model)

Call:
lm(formula = df2$Overall_Rating ~ ., data = df2)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.8923 -0.7665  0.2927  1.0899  1.7237 

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      3.27625    0.42384   7.730 5.48e-13 ***
CME.Continuing.medical.Education 0.47823    0.44843   1.066    0.288    
Improper.Treatment.Prescription  0.13786    0.26335   0.523    0.601    
Patient.Harm                    -0.23220    0.25801  -0.900    0.369    
Unprofessional.Immoral.Contact   0.03566    0.30709   0.116    0.908    
Financial.Irregularities        -1.03579    1.22994  -0.842    0.401    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.199 on 196 degrees of freedom
Multiple R-squared:  0.01377,	Adjusted R-squared:  -0.01139 
F-statistic: 0.5473 on 5 and 196 DF,  p-value: 0.7403
```



Linear Regression ~ Model 7

```
Call:
lm(formula = full$Overall_Rating ~ ., data = full)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.9759 -0.5333  0.1401  0.8639  1.5333 

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)                        3.50660    0.42249   8.300  2.4e-14 ***
n_score_bed_side                 -32.14084   70.88029  -0.453    0.651    
n_score_Knowledge                -32.42557   70.86321  -0.458    0.648    
n_score_communication            -32.03400   70.91043  -0.452    0.652    
n_score_environment              -32.34223   70.90089  -0.456    0.649    
n_score_scheduling               -32.82578   70.79110  -0.464    0.643    
n_score_wait_time                -31.92369   70.87093  -0.450    0.653    
n_score_costs                    -32.37433   70.83450  -0.457    0.648    
p_score_bed_side                 -12.32544   61.26434  -0.201    0.841    
p_score_Knowledge                -12.17156   61.20361  -0.199    0.843    
p_score_communication            -11.72705   61.29186  -0.191    0.848    
p_score_environment              -11.65792   61.21669  -0.190    0.849    
p_score_scheduling               -12.44891   61.29044  -0.203    0.839    
p_score_wait_time                -12.05156   61.25610  -0.197    0.844    
p_score_costs                    -12.00439   61.25240  -0.196    0.845    
avg_neg_score                    224.70117  496.02831   0.453    0.651    
avg_pos_score                     85.30605  428.76859   0.199    0.843    
CME.Continuing.medical.Education   0.40076    0.42783   0.937    0.350    
Improper.Treatment.Prescription    0.03539    0.25065   0.141    0.888    
Patient.Harm                      -0.17577    0.24412  -0.720    0.472    
Unprofessional.Immoral.Contact    -0.08901    0.28528  -0.312    0.755    
Financial.Irregularities           0.34706    1.13638   0.305    0.760    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.081 on 180 degrees of freedom
Multiple R-squared:  0.264,	Adjusted R-squared:  0.1782 
F-statistic: 3.075 on 21 and 180 DF,  p-value: 2.563e-05

> 
```



Model Outputs Summary

	Residual SE	R-Squared	Adjusted R-squared	F-stat	p-value
Model 1	1.066 187 df	0.2553	0.1996	4.58	4.163e-07
Model 2	0.9048 196 df	0.4381	0.4238	30.57	2.2e-16
Model 3	1.074 122	0.1357	0.07198	2.129	0.03187
Model 4	1.071 181	0.2734	0.1931	3.405	5.8e-06
Model 5	1.178 188	0.08666	0.0235	1.372	0.1761
Model 6	1.199 196	0.01377	-0.01139	0.5473	0.7403
Model 7	1.081 180	0.264	0.1782	3.075	2.563e-05

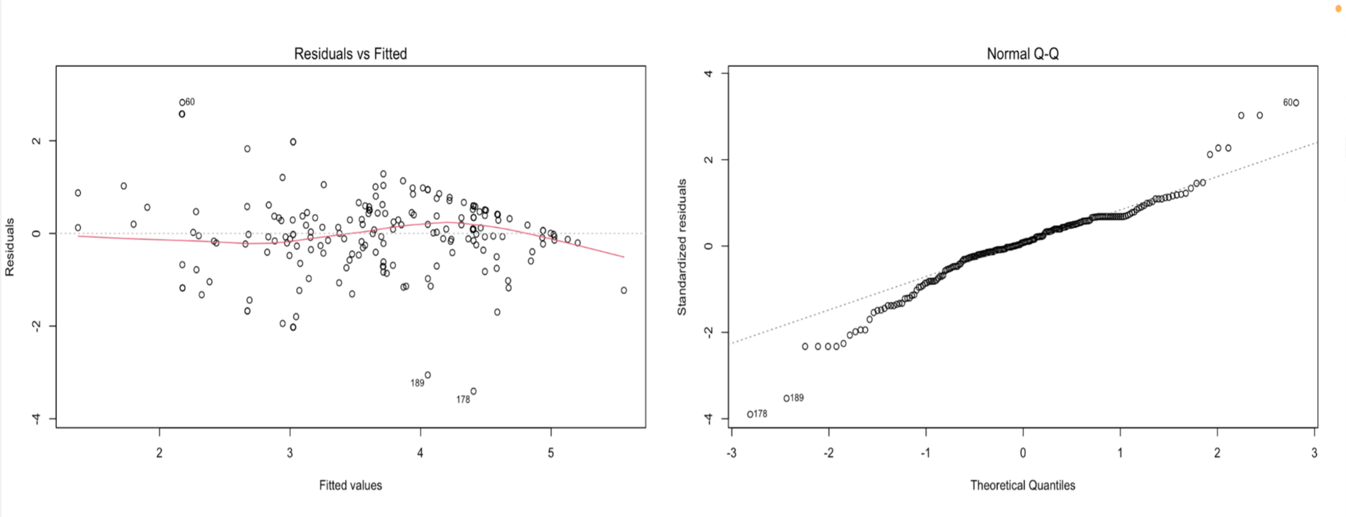
Linear Regression

We have used following variables

- ❖ p_score_communication
- ❖ n_score_communication
- ❖ p_score_environment
- ❖ n_score_scheduling
- ❖ p_score_scheduling
- ❖ avg_count_pos_rev
- ❖ avg_count_neg_rev
- ❖ Gender

These variables are significant in predicting the overall rating.

```
> summary(df3_model)

Call:
lm(formula = df3$Overall_Rating ~ ., data = df3)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.4056 -0.3871  0.0688  0.5076  2.8261 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)             3.8733     0.1928  20.087  < 2e-16 ***
p_score_communication   0.0871     0.1590   0.548 0.584465    
n_score_communication   0.2956     0.2041   1.448 0.149209    
p_score_environment     0.1893     0.1850   1.023 0.307408    
n_score_scheduling     -0.7439     0.1940  -3.836 0.000170 ***
p_score_scheduling     -0.2822     0.1770  -1.594 0.112593    
avg_count_pos_rev       0.5324     0.1435   3.710 0.000271 ***
avg_count_neg_rev      -0.8496     0.1576  -5.390 2.04e-07 ***
Gender                 -0.3505     0.1618  -2.166 0.031552 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8817 on 193 degrees of freedom
Multiple R-squared:  0.4746,	Adjusted R-squared:  0.4528 
F-statistic: 21.79 on 8 and 193 DF,  p-value: < 2.2e-16
```

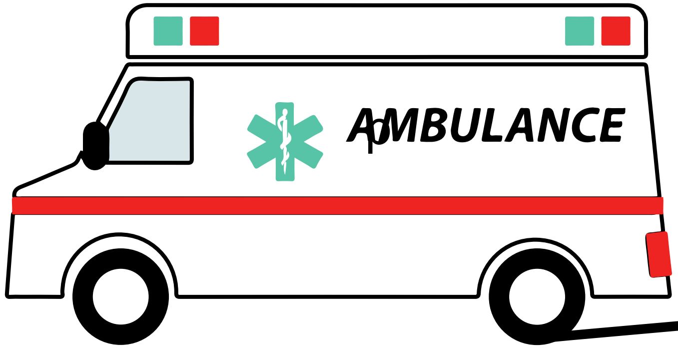


Significant Models - Summary

	Models’ input variables	r - Squared	RMSE	Significant Variables
1	All Positive and negative topic scores of reviews -14 var	0.25	1.02	n_scheduling, p_communication p_environment
2	Avg Review Count Gender	0.43	1.01	p_avg_review_count n_avg_review_count Gender
3	Gender, All Positive & negative topic scores of reviews,Sanctioned	0.27	1.01	n_scheduling, p_communication p_environment
4	Negative review topic scores+Gender	0.13	1.01	n_score_communication n_scheduling, Gender
5	Gender, Sanctioned Topic scores,Positive review Scores	0.19	1.23	Gender p_communication p_scheduling

Summary

- There is a weak association between the patient reviews and the physician sanctions.
- Communication, scheduling, and office staff are the most significant factors driving star ratings, while bedside manners, medical expertise, and cost are less important.
- The study also found that doctors who had been sanctioned for multiple reasons tended to have lower overall ratings, indicating a negative perception among the public. However, the analysis did not find any statistical significance between the sanctioned reasons and doctors' overall ratings
- Financial aspects were initially thought to be an important reason for physician sanctions, but this was not supported by the data visualization.
- Gender is a significant dependent variable, but it is not actionable for physicians to improve their service.



Thank You