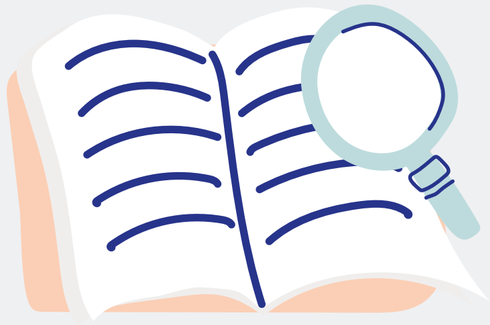
Drug Analysis with Algorithm in Sentiment Analysis

View on Github

Background



"The review can help to identify any medicines that are no longer needed or any that need the dosage changed"

Based on National Institute for Health and Care Excellence UK : Medicine Optimisation

direct review

online review

Problem Statement

WebMD is an American corporation which publishes online about drugs and is an important healthcare information website and it is the most popular consumer oriented health site.

Question : **Identifying which drug tend to have more negative reviews? Can you determine if a review is positive or negative?**

Goal : **To analyzing drug reviews. Furthermore hopefully that it can help drug companies to improve the quality of their products and helping paramedic in making decisions regarding patient treatment.**

Data Preprocessing

Dataset Overview

Source : https://www.kaggle.com/datasets/rohanharode07/webmd-drug-reviews-dataset?rvi=1

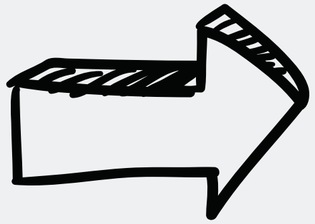
Data Cleaning

- Missing Values
- Duplicates
- Feature Engineering
- Handling Outliers

Data Preprocessing

#	Column
0	Age
1	Condition
2	Date
3	Drug
4	DrugId
5	EaseofUse
6	Effectiveness
7	Reviews
8	Satisfaction
9	Sex
10	Sides
11	UsefulCount

362806 entries



#	Column
0	Age
1	Condition
2	Drug
3	EaseofUse
4	Effectiveness
5	Satisfaction
6	Sex
7	Reviews
8	UsefulCount
9	Sides
10	Date
11	Year
12	Month
13	day_of_week

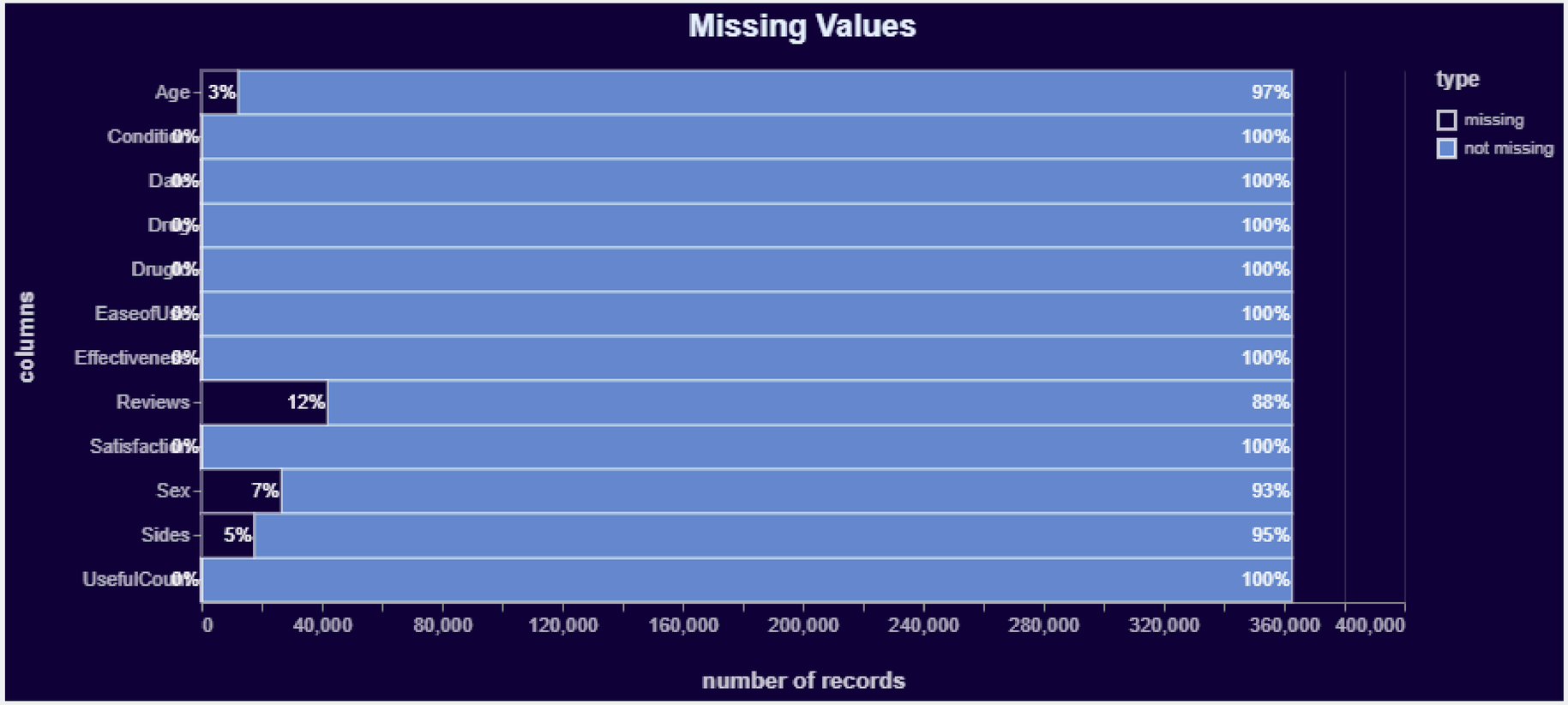
237049 entries

This data was taken from September 2007 to March 2020

There are 944 duplicated values and 27% Missing Values

There is a new feature namely day, month, year as a fraction of the date feature

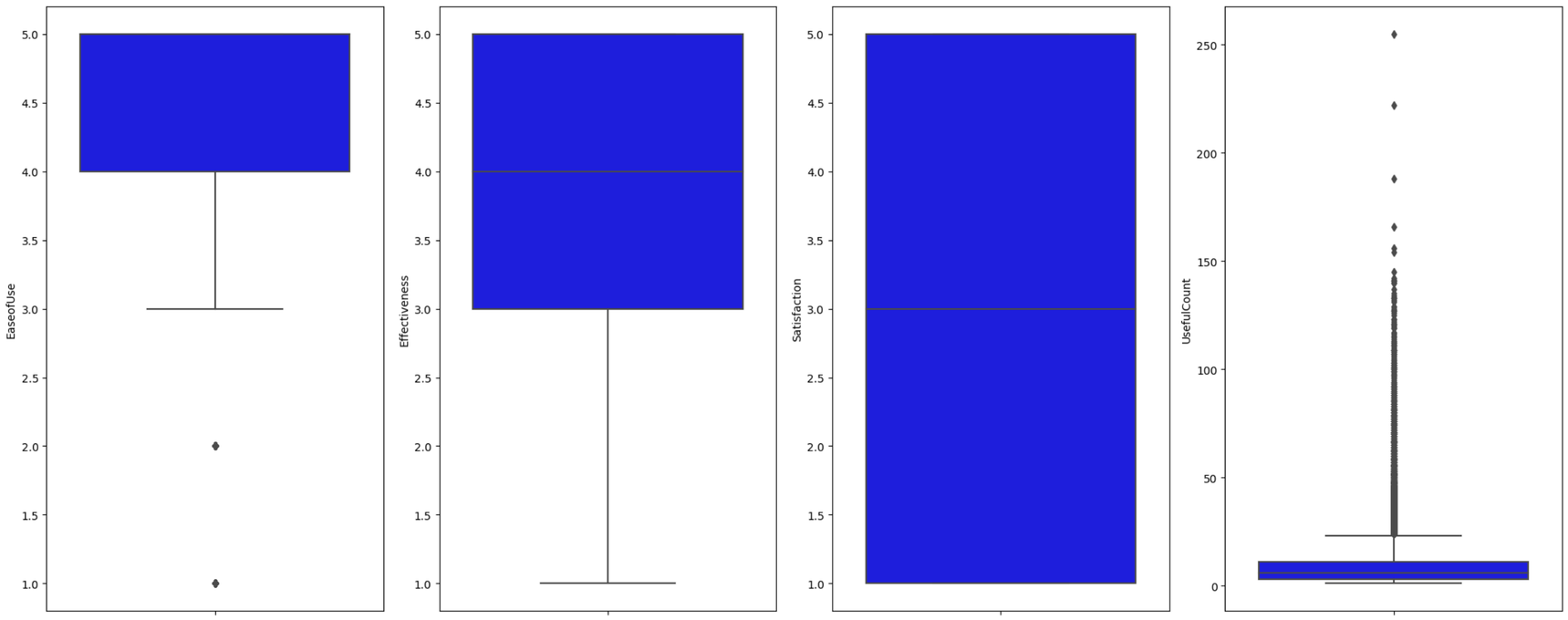
Data Preprocessing



27% Missing Values consisting of Reviews 12%, Sex 7%, Age 3%. and Sides 5%

Data Preprocessing

Outliers Checking



Because of an outlier still make sense, we will not remove all Useful count of drug that are more than 25 since it is only 5% from the total data.

Feature Engineering

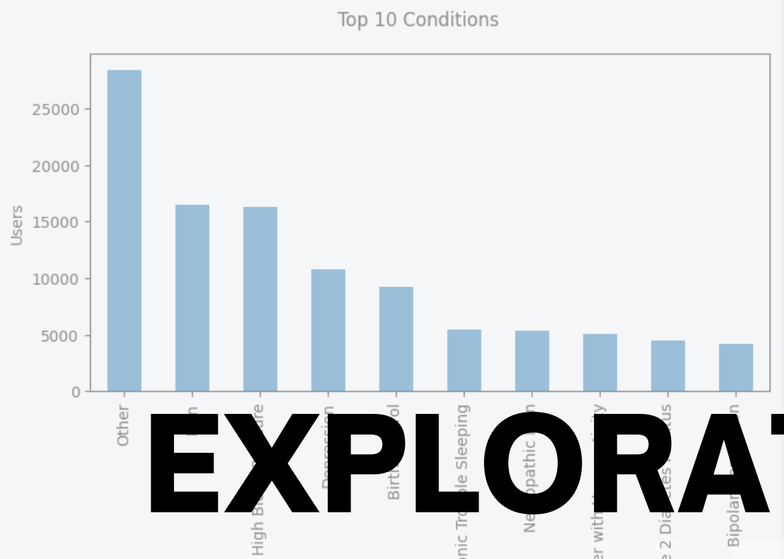
Drug vs DrugId

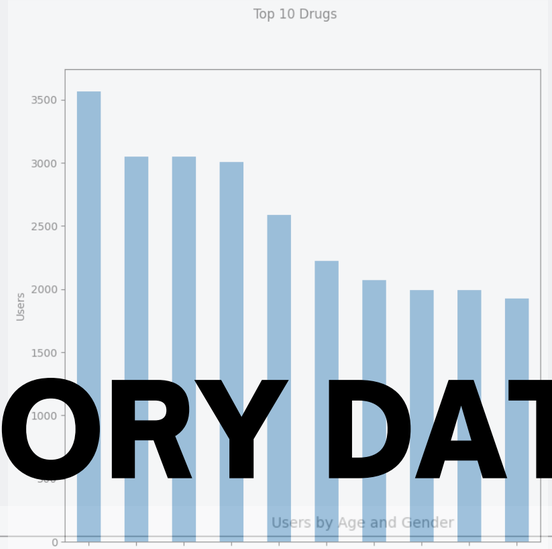
Column	Num of unique values
Drug	7093
DrugID	6572

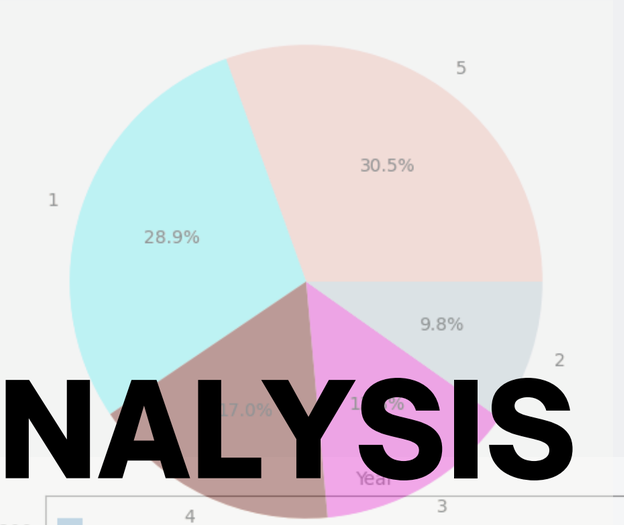
```
167493: ['lynparza tablet', 'lynparza']
 11594: ['lovastatin', 'lovastatin tablet, extended release 24 hr']
  7286: ['loteprednol etabonate ointment', 'loteprednol etabonate drops, gel']
  7319: ['lotemax drops, gel', 'lotemax ointment']
164437: ['lorcaserin tablet', 'lorcaserin tablet, extended release 24 hr']
  8892: ['lorazepam concentrate', 'lorazepam']
 76548: ['loratadine-d', 'loratadine d']
    73: ['loratadine', 'loratadine tablet,disintegrating']
  8555: ['loprox cream', 'loprox suspension, topical']
  4789: ['loperamide liquid', 'loperamide']
```

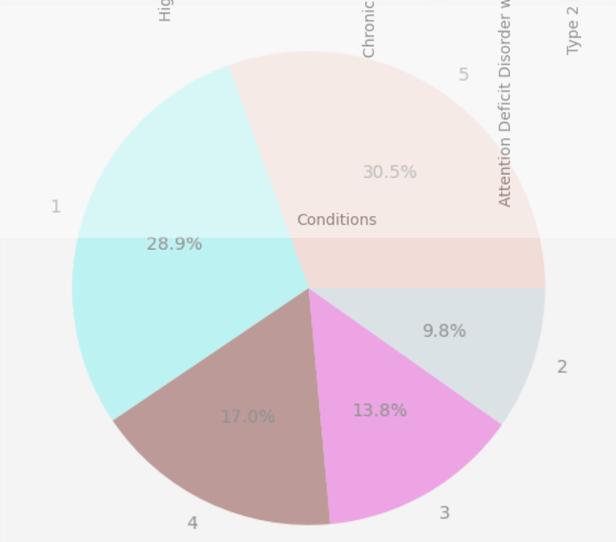
The reason why Drug has more unique values than DrugId is that some drugs have same ID number, like loprox cream, lorazepam, gel, etc. So we will only use Drug column instead of DrugID

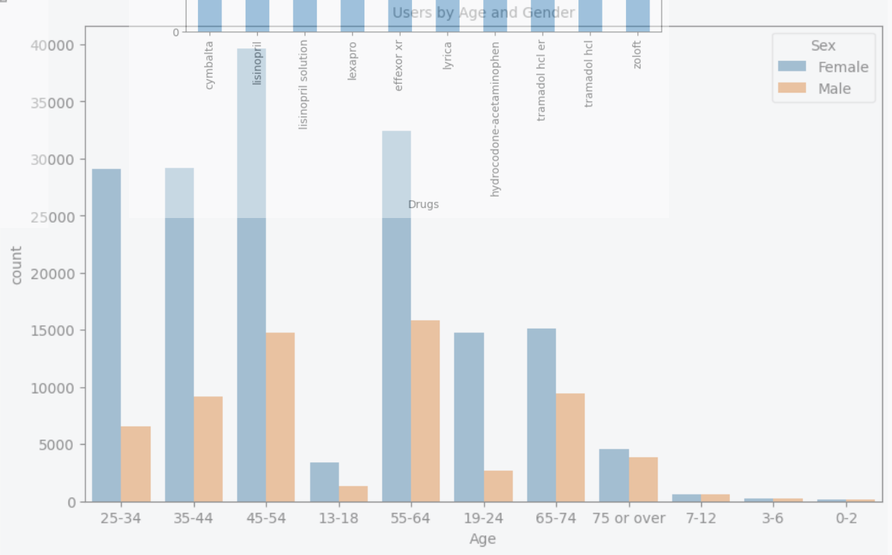
EXPLORATORY DATA ANALYSIS

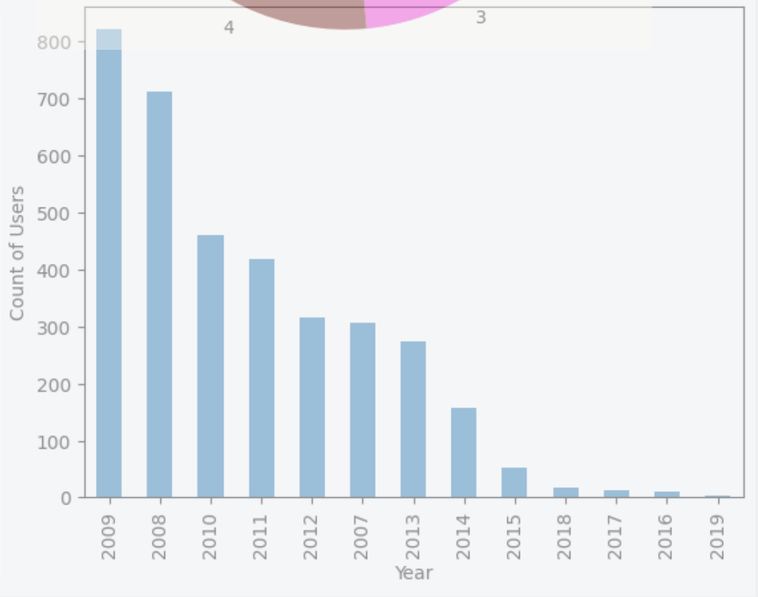


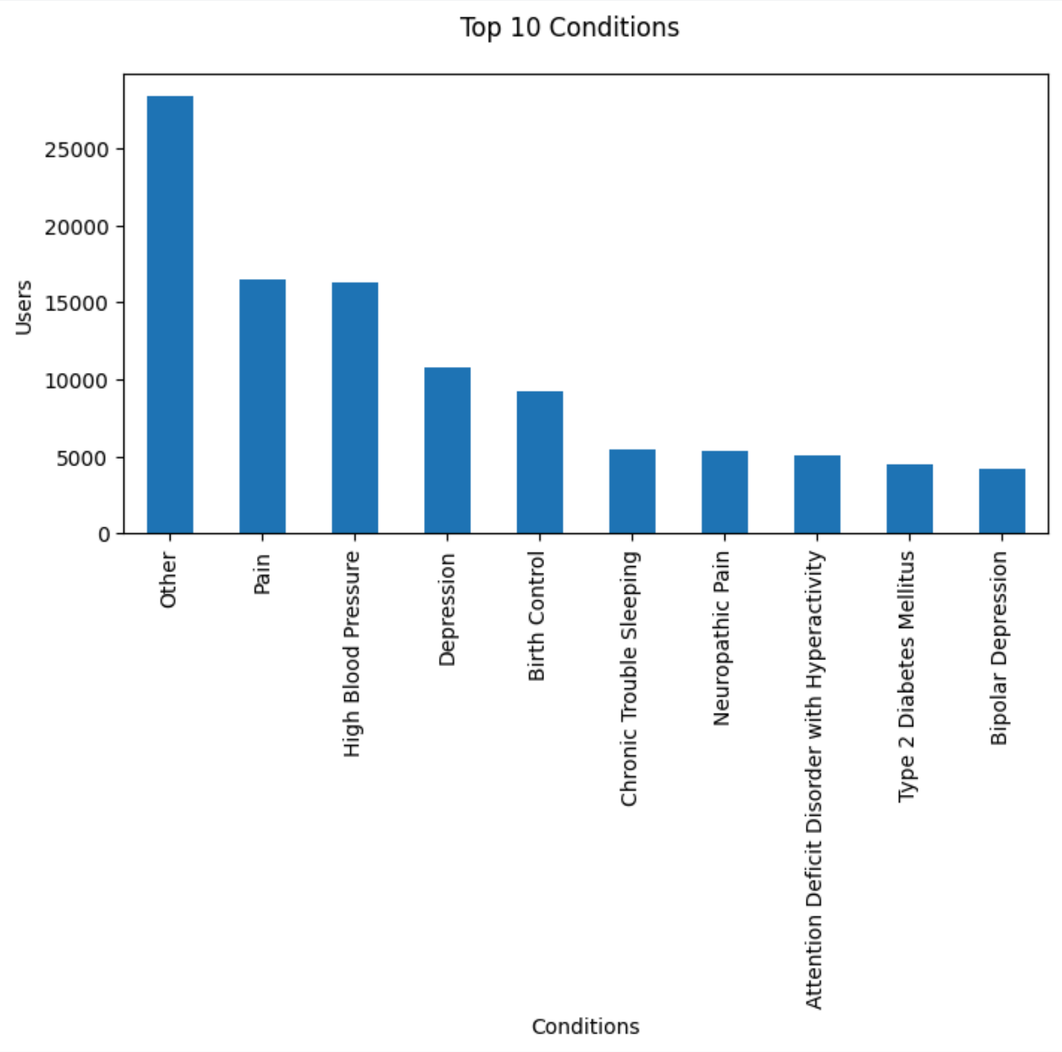










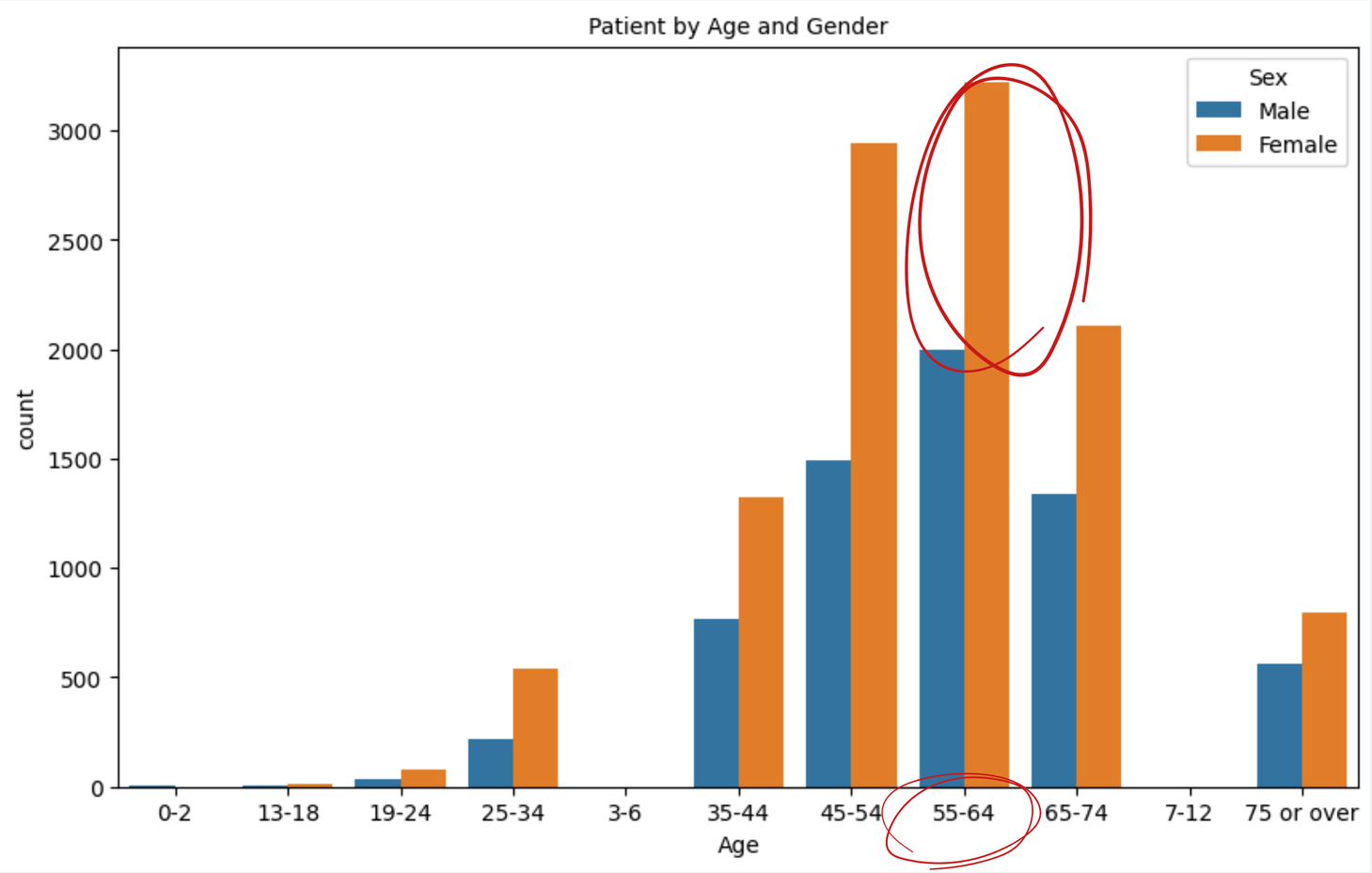


Top 10 Conditions among the patients

What is the common conditions among the patients?

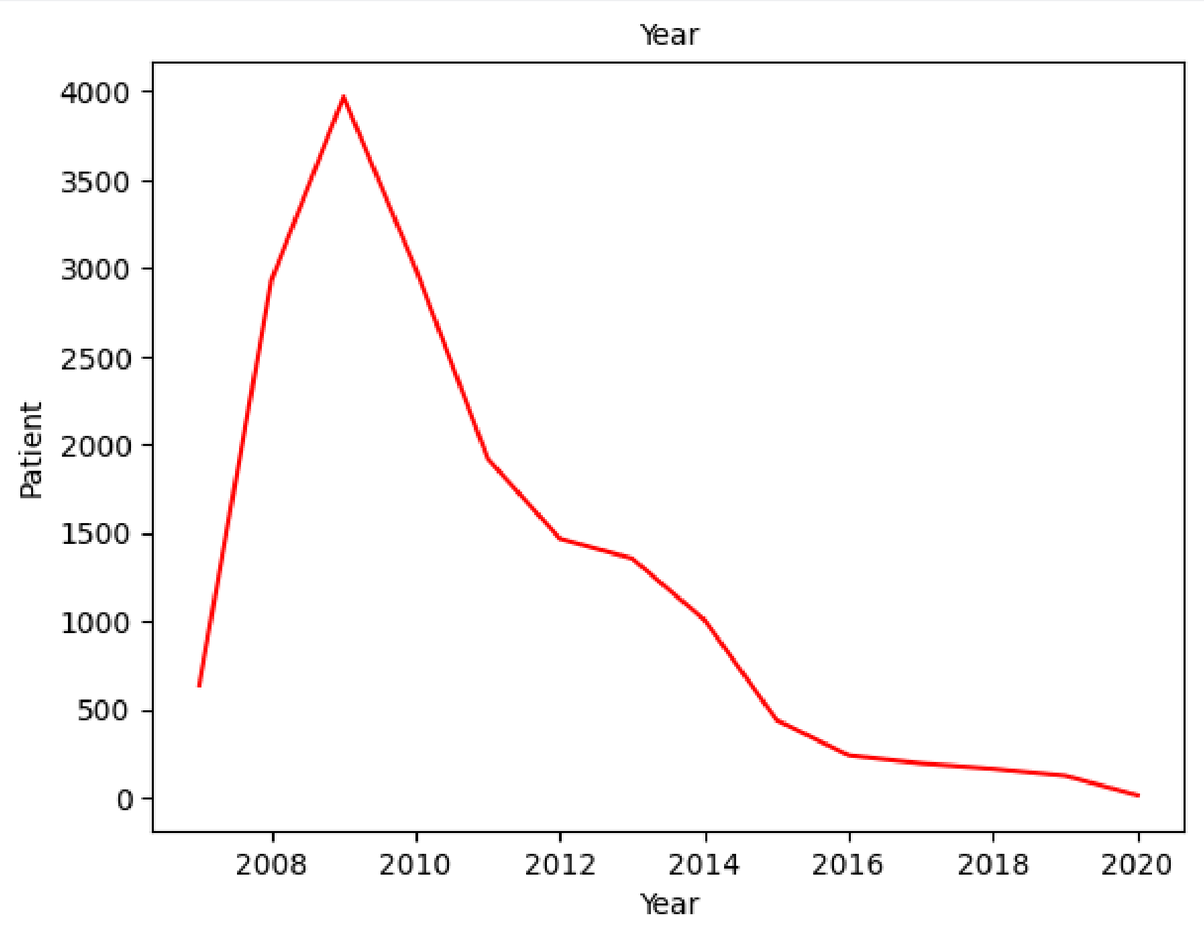
Common conditions among the patients during 13 years are high blood pressure, pain, and other conditions.

Lets find out more about high blood pressure!



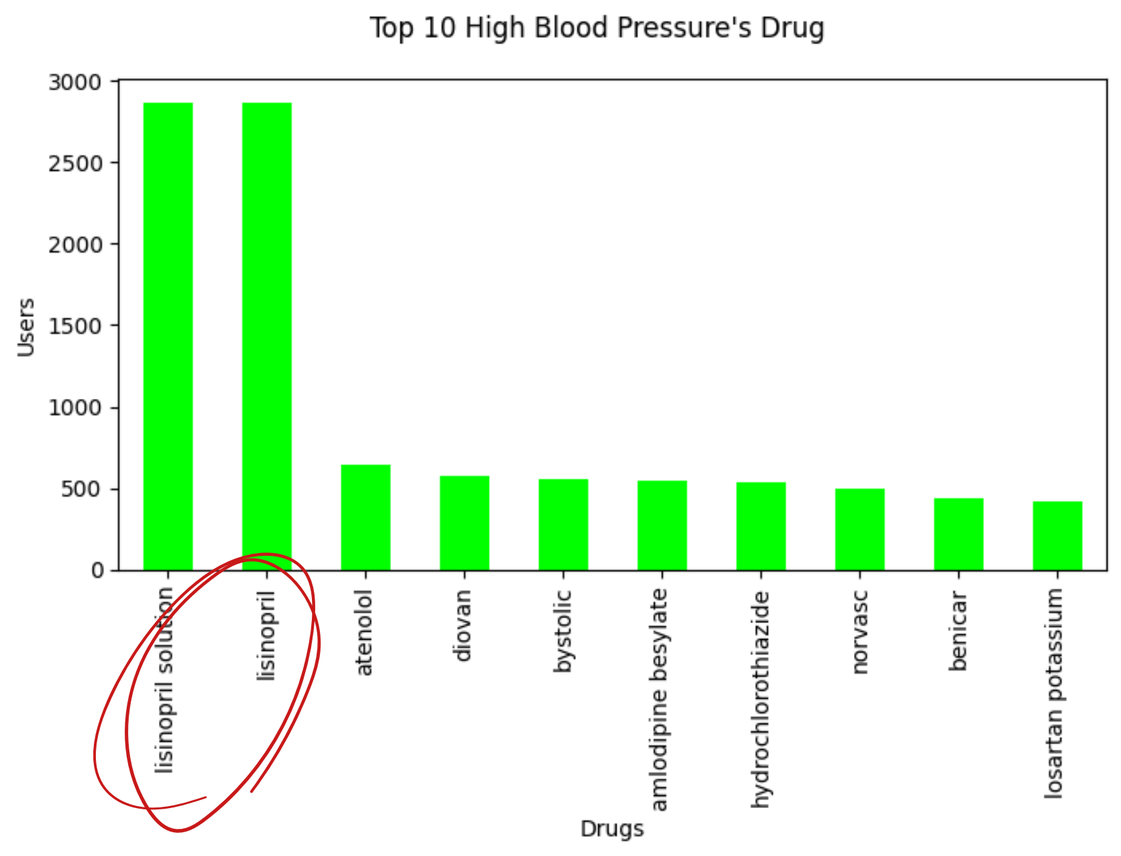
We can see that high blood pressure occurs frequently in age range **55-64 and is dominated by female**. The **causes of hypertension** are closely related to genetic factors such as **age and gender**, as well as lifestyle and diet (AHA, 2014)

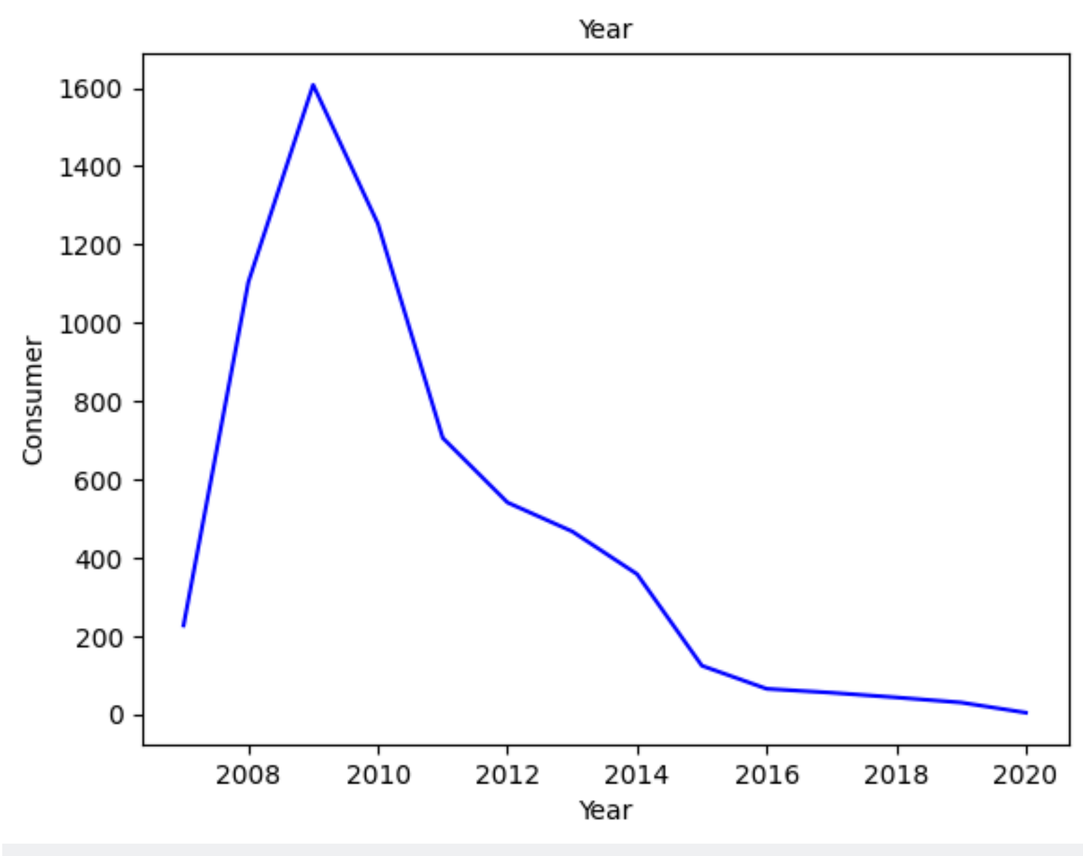
High blood pressure by Year



Hypertension was increase from 2007 to 2009. According to various sources, starting in 2007 there was an economic crisis in the world until several years later. This event has an impact, one of them is triggering **stress, this can be a factor in the occurrence of hypertension.**

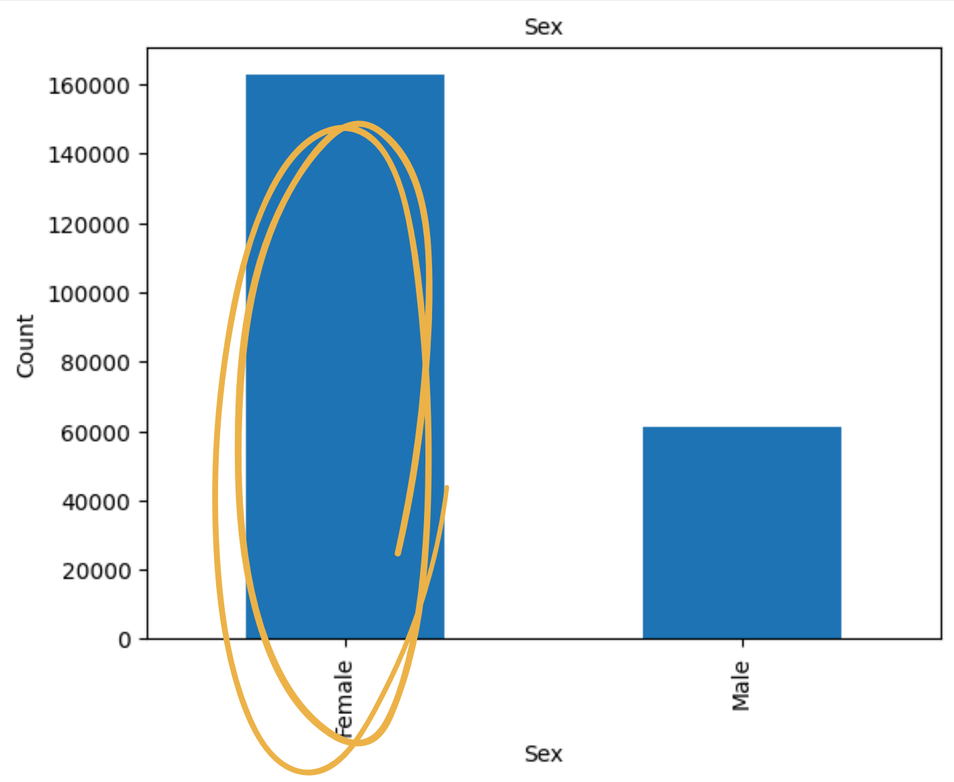
High Blood Pressure's Drug

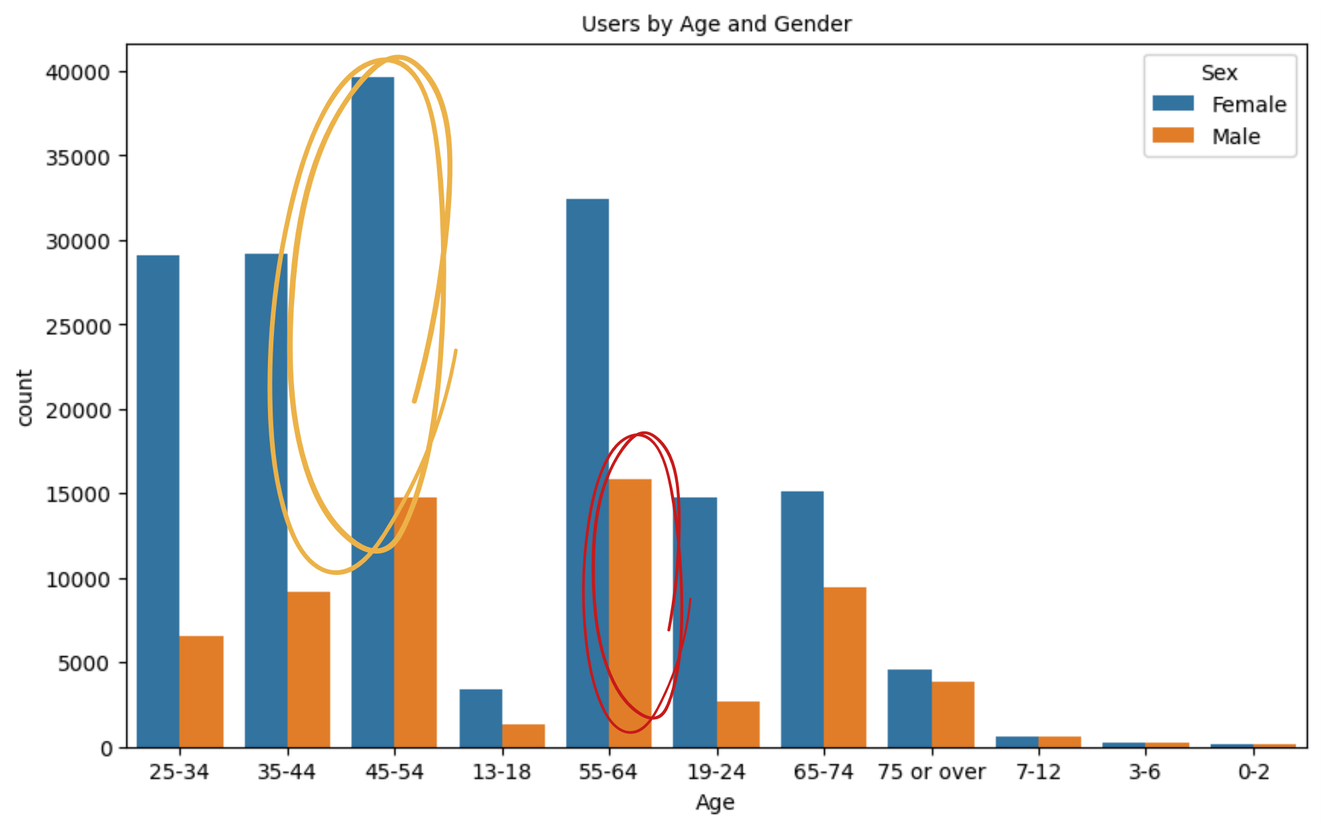




We can see lisinopril is the most consume as high blood pressure drug. Lisinopril was widely consumed from 2007 to 2009

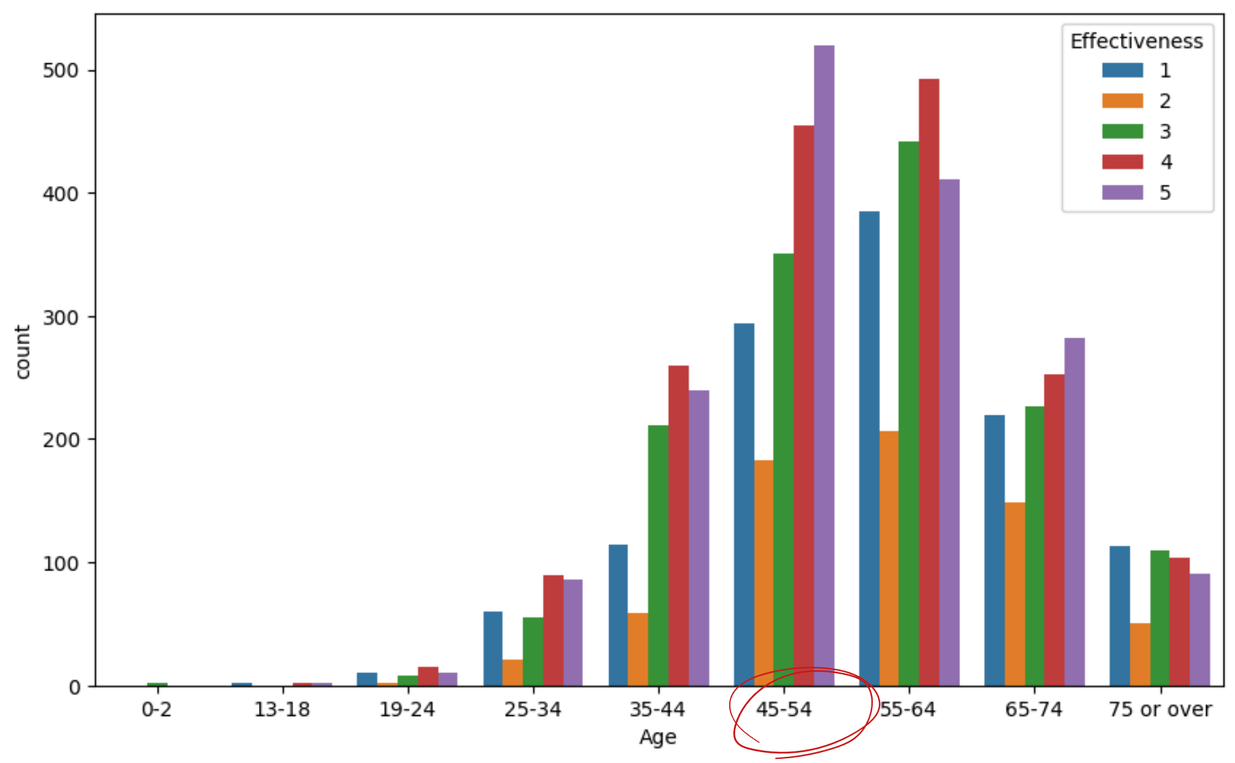
Let's find out more about Lisinopril!

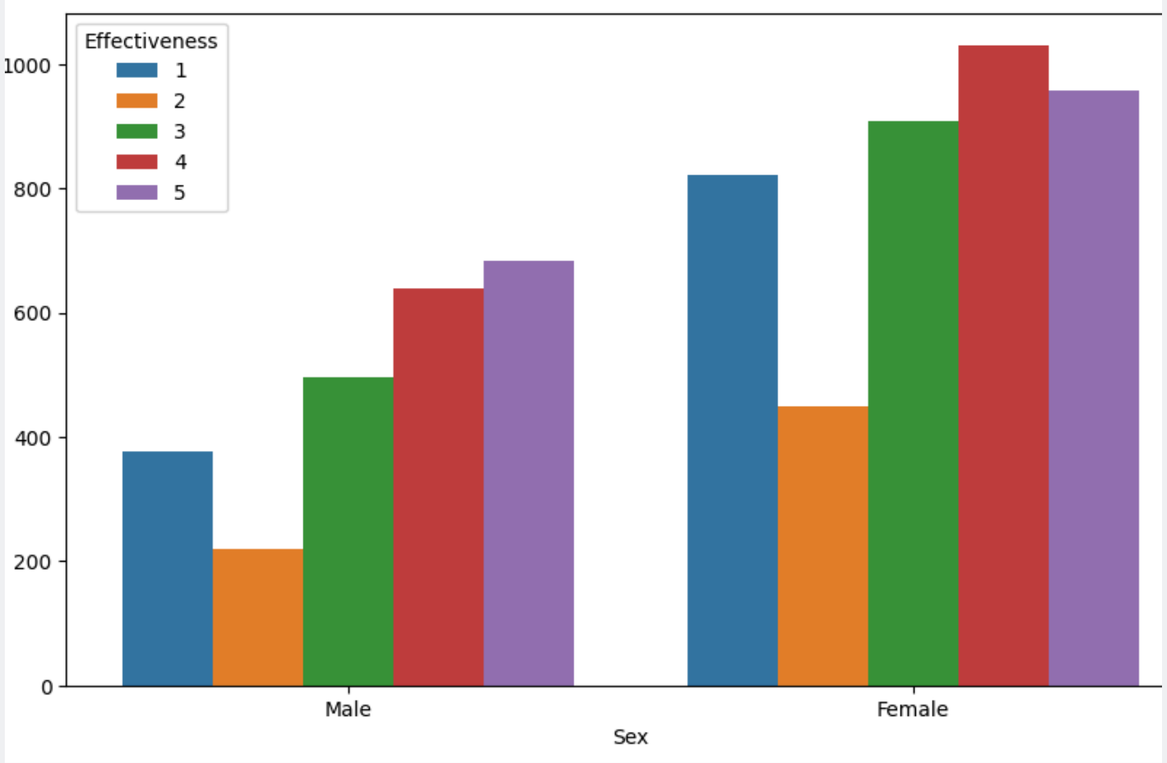




From all the patients, the majority were women. However, according to the age of most female patients in the age range 45 - 54 and most male patients in the age range 55 - 64

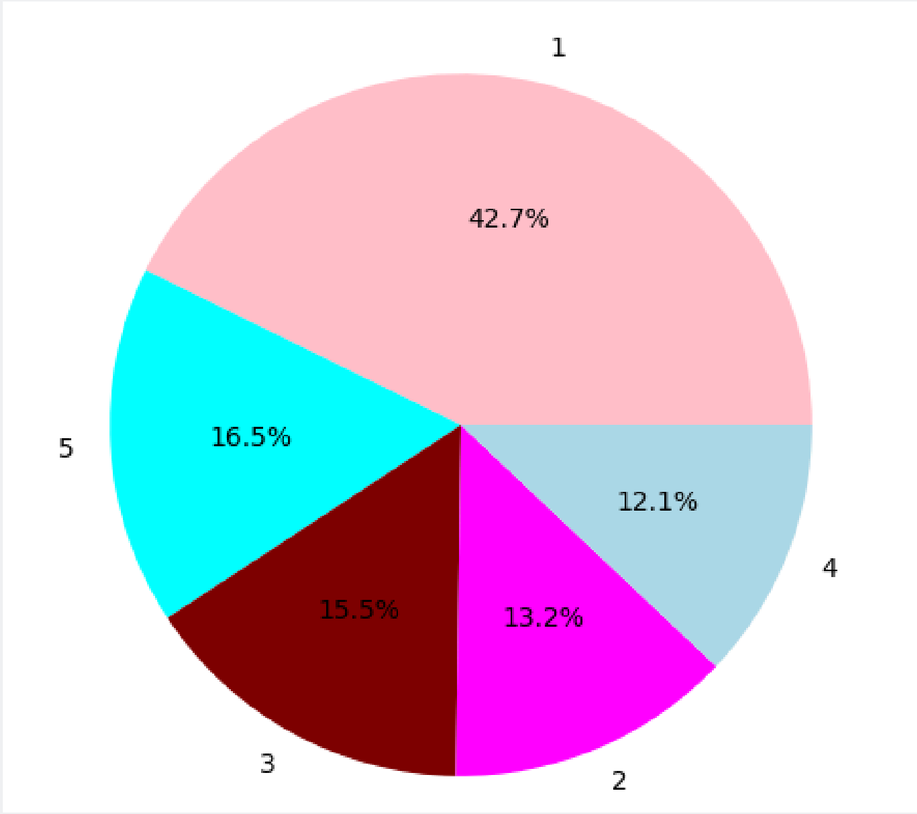
Effectiveness Lisinopril

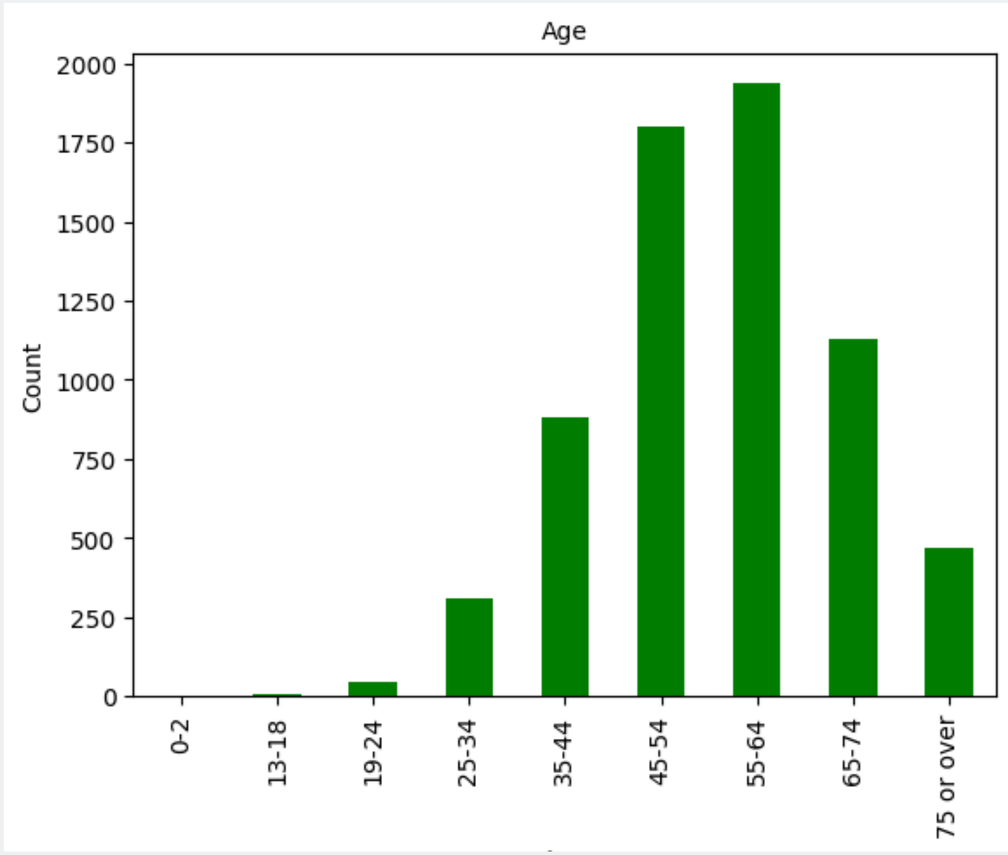




The level of effectiveness of the drug works well in mostly on female and the age range of 45 - 54. Some body functions may not work with the same efficiency as at a younger age, and this may affect the response to the drug.

Satisfaction's Rating of Lisinopril



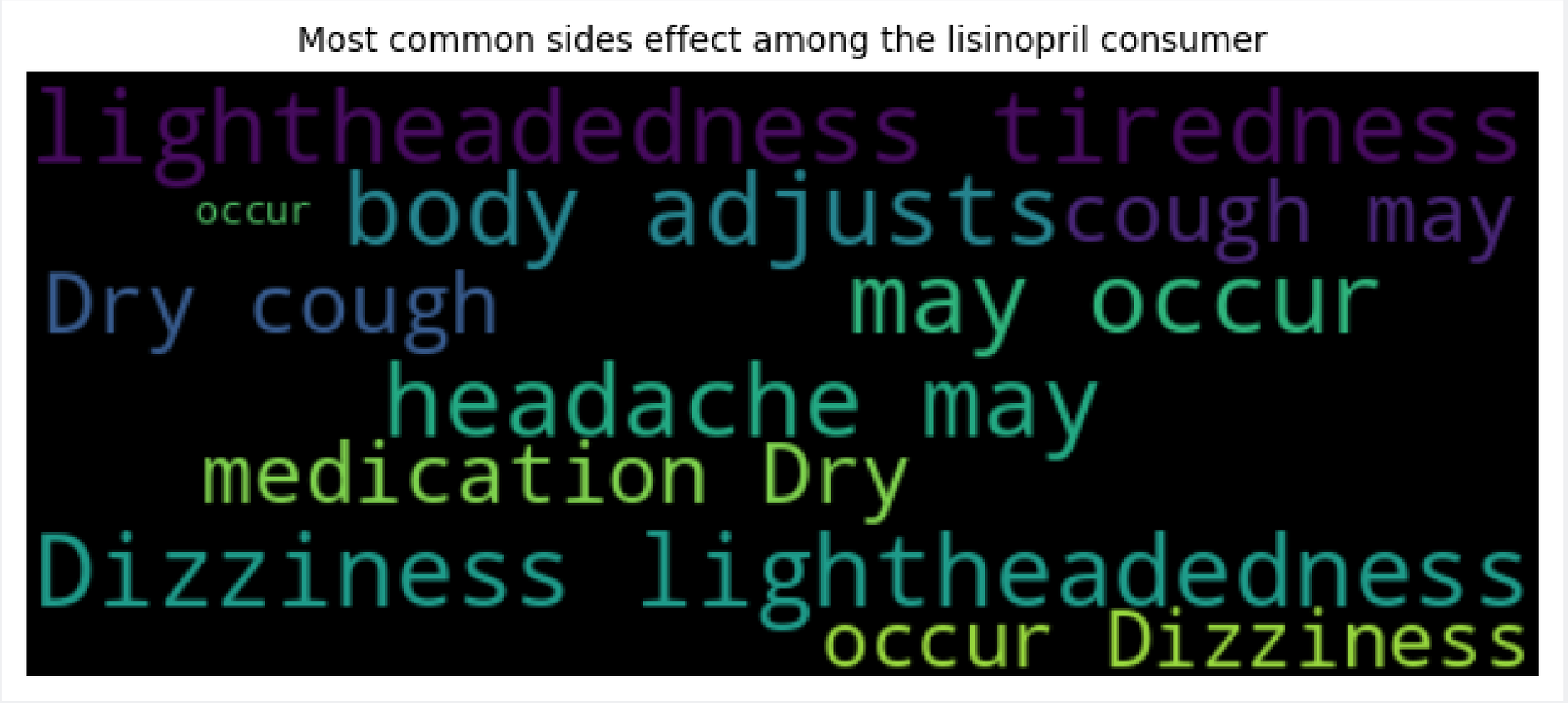


Because of the drug is mostly effective in people under 55 , while hypertension sufferers mostly from people over 55 , so the drug gets a low satisfaction rating.

Review's of Lisinopril

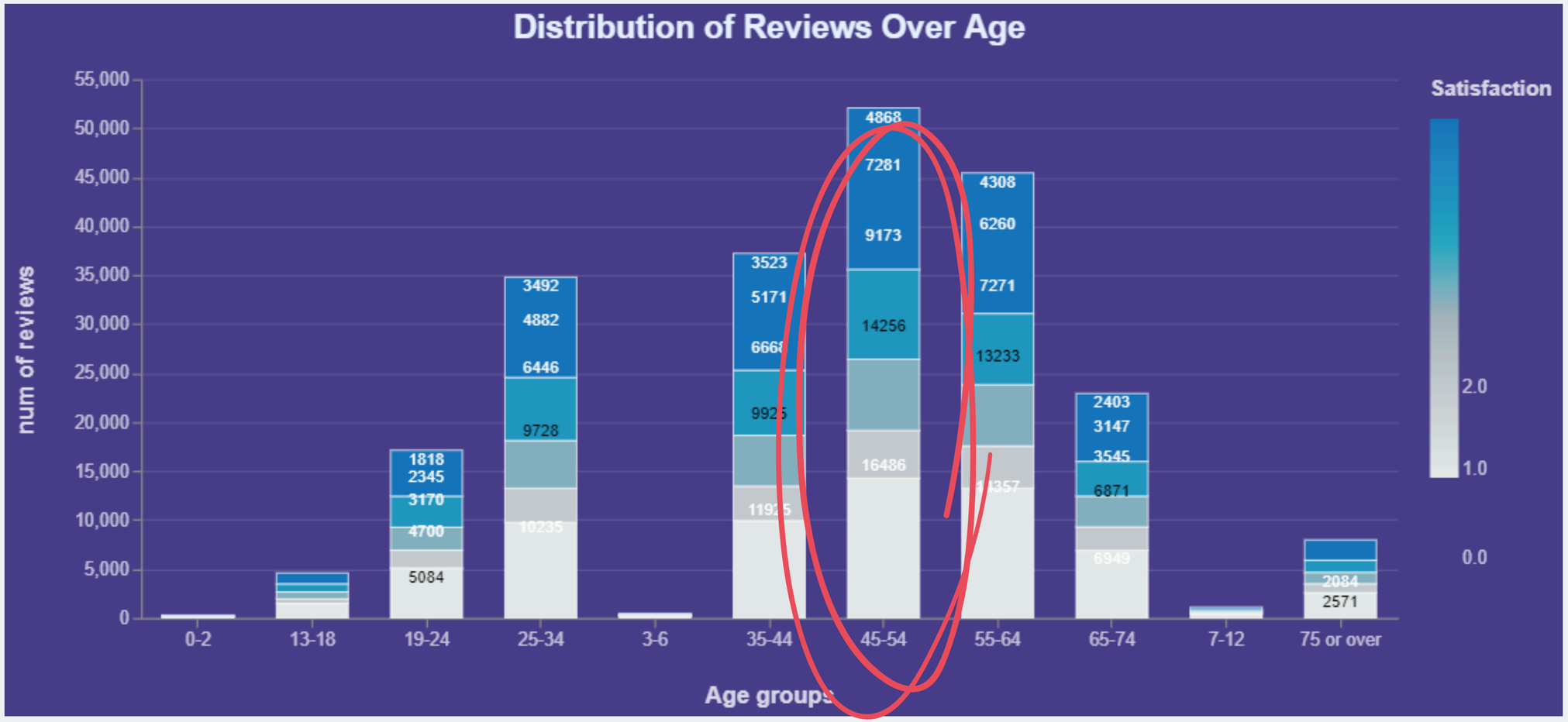
Based on these reviews, lisinopril is used to medication of high blood pressure. Besides that, lisinopril has side effects.

Sides Effect of Lisinopril

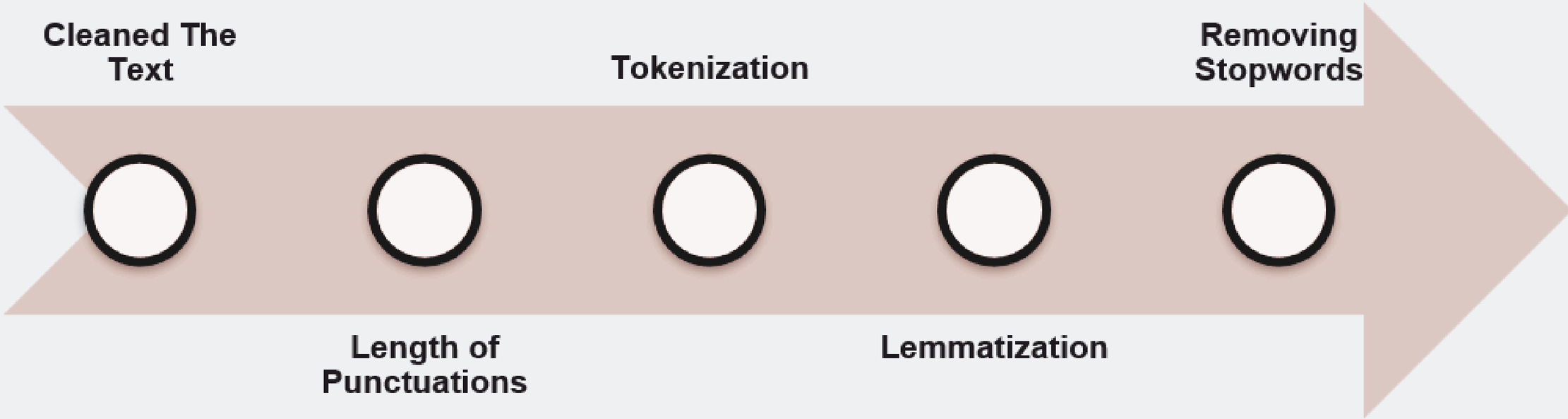


There are various side effects of lisinopril, one of which is lightheadedness

Distribution of Reviews by Age



Data Preprocessing Sentiment



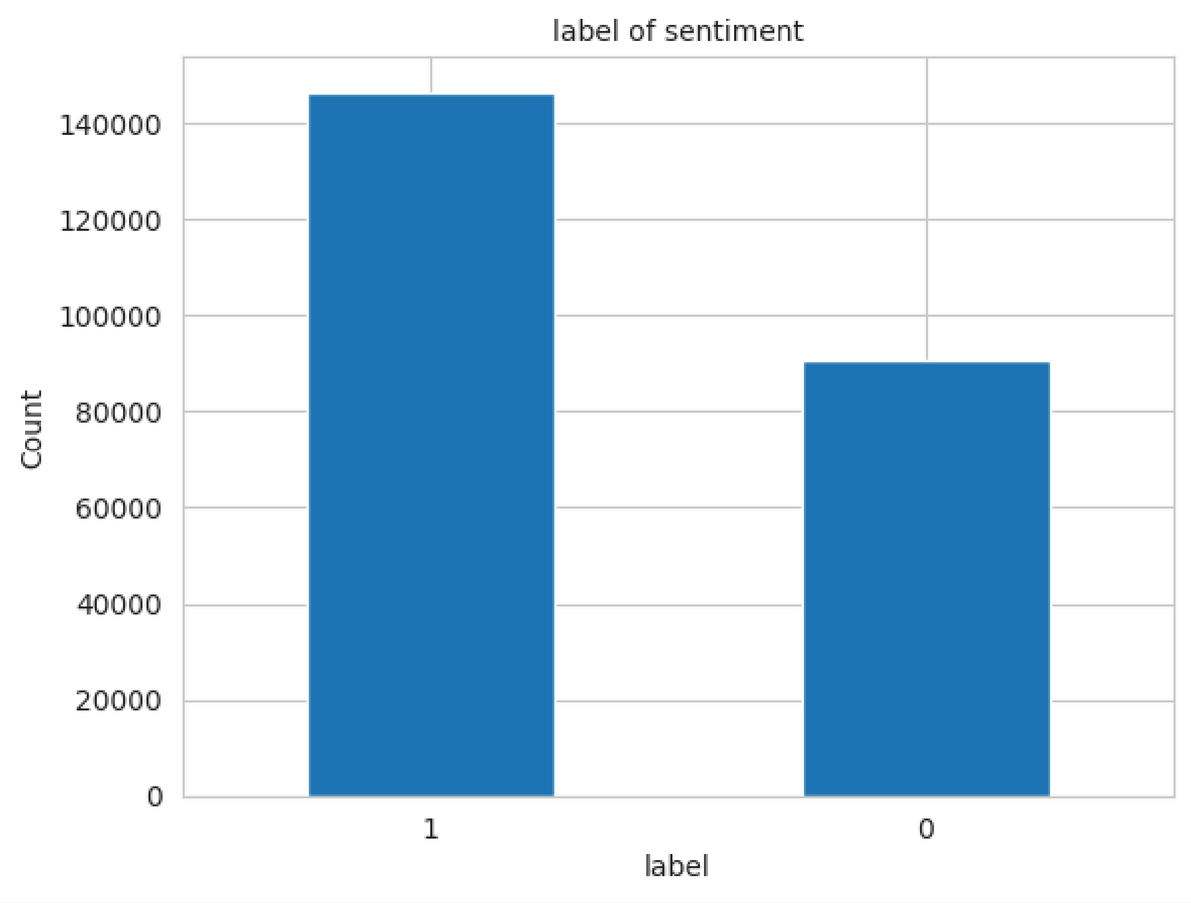
Data Preprocessing Sentiment

cleaned_text	label	Review_len	punct	tokens	lemmatized_review
cleared me right up even with my throat hurtin...	1	71	0.0	[cleared, me, right, up, even, with, my, throa...	cleared right even throat hurting went away ta...
lyza birth control these are the worst birth c...	0	394	4.8	[lyza, birth, control, these, are, the, worst,...	lyza birth control worst birth control pill ev...
i have been taking lyza for two months now i ...	1	288	3.5	[i, have, been, taking, lyza, for, two, months...	taking lyza two month anxious going back pill ...
i have been on this pill for a little over two...	0	302	3.3	[i, have, been, on, this, pill, for, a, little...	pill little two month experience horrible know...
my ob gyn placed me on this pill because i was...	1	595	3.9	[my, ob, gyn, placed, me, on, this, pill, beca...	ob gyn placed pill risk stroke estrogen birth ...

- **Label 1 = Satisfaction with 3, 4, 5 Rating**
- **Label 0 = Satisfaction with 1 & 2 Rating**

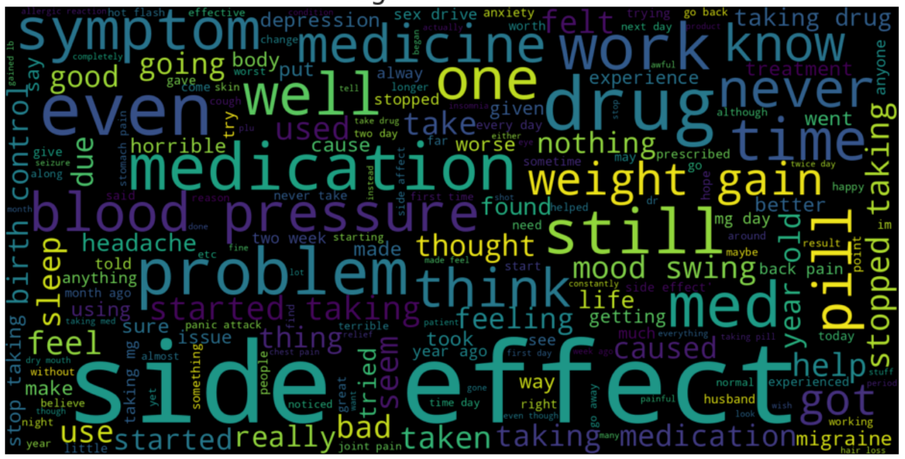
- **Label 0 is Negative Sentiment**
- **Label 1 is Positive Sentiment**

Exploratory Data Analysis Sentiment



Positive Reviews

Negative Reviews



Feature Extraction From Text

```
[ ]  1 from sklearn.feature_extraction.text import TfidfVectorizer
     2
     3 vectorizer = TfidfVectorizer()
     4 X_train = vectorizer.fit_transform(train_set["lemmatized_review"]).toarray()
     5 X_test = vectorizer.transform(test_set["lemmatized_review"]).toarray()
     6 y_train = train_set["label"].values
     7 y_test = test_set["label"].values
```

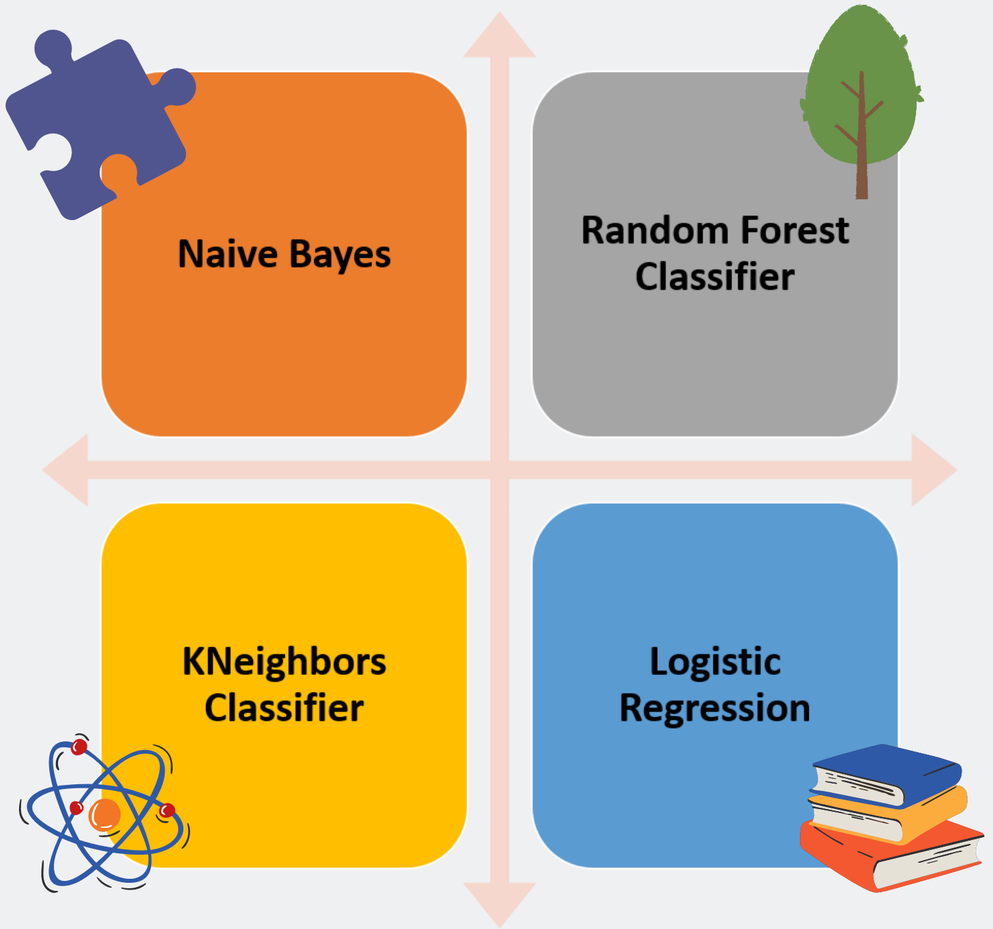
TF - IDF Vectorizer

helps represent text in the form of numeric vectors by assigning weights to each word in a document based on how important that word is in the document.

Lemmatize Text

This aims to reduce word variations to their basic form in order to identify relationships between words that have the same meaning.

Classification Models



Metric Evaluation

Train Set

No	Model	Metric			
		Precision	Recall	F1 - Score	Accuracy
1	Naive Bayes	0.92	0.88	0.90	0.88
2	Random Forest Classifier	0.62	1.00	0.76	0.62
3	KNeighbors Classifier	0.74	0.95	0.83	0.77
4	Logistic Regression	0.92	0.95	0.93	0.92

Because of the datasets are imbalance, then we can use accuracy for predicting. We still have another metric evaluation such as precision for measures the extent of a model's positive predictions are actually positive.

Metric Evaluation

Test Set

No	Model	Metric			
		Precision	Recall	F1 - Score	Accuracy
1	Naive Bayes	0.80	0.76	0.78	0.74
2	Random Forest Classifier	0.62	1.00	0.76	0.62
3	KNeighbors Classifier	0.67	0.88	0.76	0.66
4	Logistic Regression	0.80	0.84	0.82	0.77

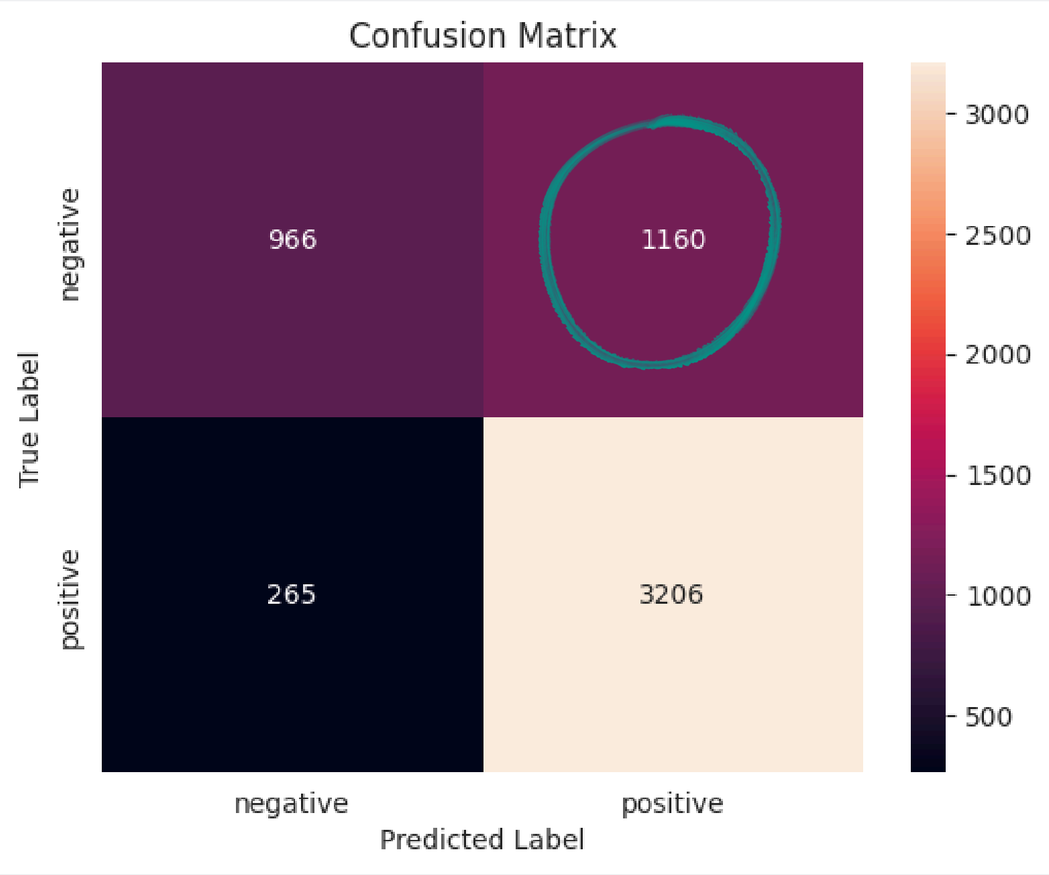
Because of the datasets are imbalance, then we can use accuracy for predicting. We still have another metric evaluation such as precision for measures the extent of a model's positive predictions are actually positive.

K-FOLD CROSS VALIDATION

K-Fold Cross-Validation to measure model performance more reliably by providing more consistent performance estimates

No	Model	Precision
1	Naive Bayes	0.807
2	Random Forest Classifier	0.617
3	KNeighbors Classifier	0.681
4	Logistic Regression	0.786

Random Forest Classifier



The stable value in precision among train and test set could be considered a good indication.

- FP : 1160 means 1160 negative reviews that predicted as positive reviews.
- TN : 966 means 966 reviews that correctly predicted as negative reviews
- FN : 265 means 265 positive reviews that predicted as negative reviews
- TP : 3206 means 3206 reviews that correctly predicted as positive reviews.

Precision is good to use when the false positive rate (classifying what should be negative as positive) is high.

Summary

Based on analyzing, **high blood pressure** is the most commonly condition among the patient. It occurs **widely in 2007-2008.** The occurrence of an economic crisis can trigger **stress, which can become a factor in high blood pressure.**

Most high blood pressure patients come from **age range 55-64 and Lisinopril is the most consume as high blood pressure's drug.** However the drug is **more effective in people under 55,** so it lead the **drug have low satisfaction rating**

Random Forest Classifier is the best model with a stable precision value among train and test set. This could be considered a good indication

Recommendation

Modeling

Perform imbalance data handling. compare models with the best evaluation metrics

Business

Improving the quality of drugs and develop strategies for product based on comparing sentiment trend. Furthermore communicating investor about positive sentiment to see the potential product.

Medication

Paramedic reconsider providing treatment by looking at its effectiveness for age and also the side effects it causes

Monitoring

Monitoring on various online platforms to carry out early detection if there are problems with the product, so it can be handled immediately. This is such an attempt to maintain public trust.

Thank you!

Reach me through :

jsuhmand@gmail.com

linkedin.com/in/amandaartyan/

https://github.com/amandartyan