Data Analysis Report

Yusuf Mert Çelikarslan

Contact Info: yusufmertcelikarslan@gmail.com

Table of Contents

There are 2357 entries in the entire dataset. In order to easily access each attribute and for visual purposes columns of the dataset renamed to lowercase english words. Attributes of this dataset and each attributes types are like following;

Variable	Type
Kullanıcı_id	Categorical
Cinsiyet	Categorical
Doğum Tarihi	Datetime
Uyruk	Categorical
Il	Categorical
Ilac_adi	Categorical
Ilac_baslangic_tarihi	Datetime
Ilac_bitis_tarihi	Datetime
Yan_etki	Categorical
Yan_etki_bildirim_tarihi	Datetime
Alerjilerim	Categorical
KronikHastaliklarim	Categorical
BabaKronikHastaliklari	Categorical
AnneKronikHastaliklari	Categorical
Kiz Kardes Kronik Hastalılari	Categorical
Erkek Kardes Kronik Hastaliklari	Categorical
Kan Grubu	Categorical
Kilo	Numerical
Boy	Numerical

When the table is inspected, it can be seen 17 of 19 attributes are categorical and only 2 of them are numerical. Therefore, the encoding style of these features should be decided looking at two perspectives, one is variable is ordered or not, two the number of unique classes.

A closer look at each feature:

Note: From now on,I will use naming convention like mentioned above.

Feature Types

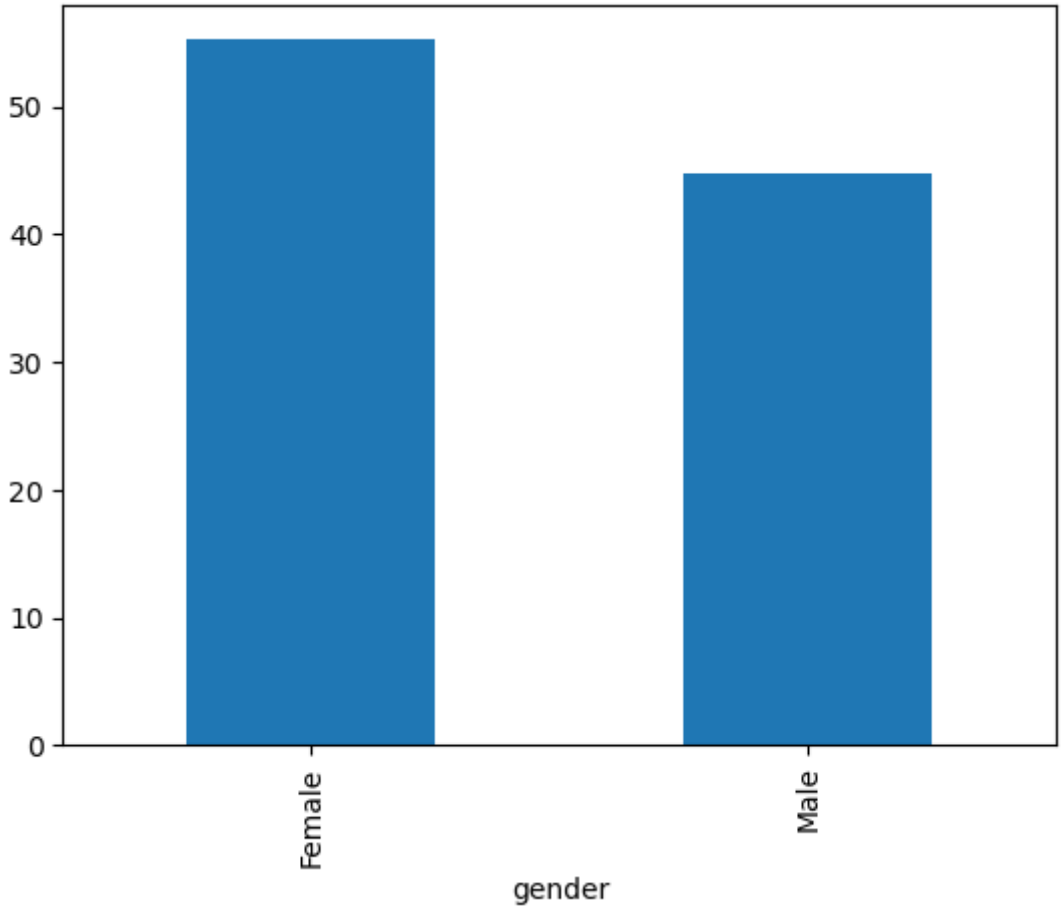
user_id

There are a total of 196 unique users and each user has an average of 12 entries. The overall distribution of the number of users is like a normal distribution. Some users have more than 20 entries and some have less than 5, but most users have a similar number of instances.

There is no missing data in this field.

gender

There are 872 female, 707 male, 778 NaN samples and 54% of these samples are female and others are male. Male – female ratio in the dataset reflects the real world data but there are missing entries, so our way of dealing with them should not effect this ratio.



birthdate

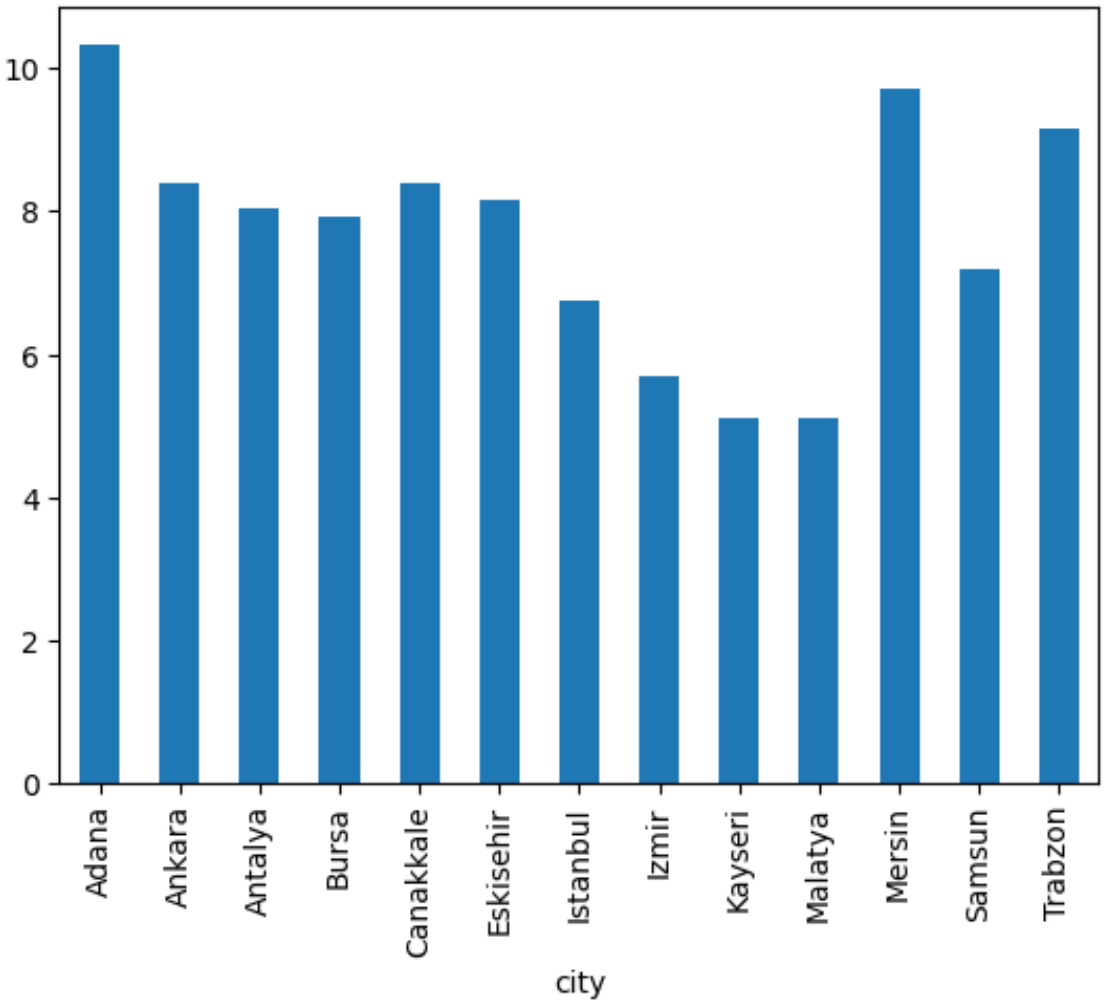
This is a datetime feature and related to age of the person. Changing this feature to age might be useful. Other than that i do no think date-time attributes of this feature is useful.

origin

Since this property contains only one class, it makes sense to omit this field.

city

The province with the highest number of data points is Adana with 10%, while the provinces with the lowest number of data points are Izmir, Malatya and Kayseri with 5%.

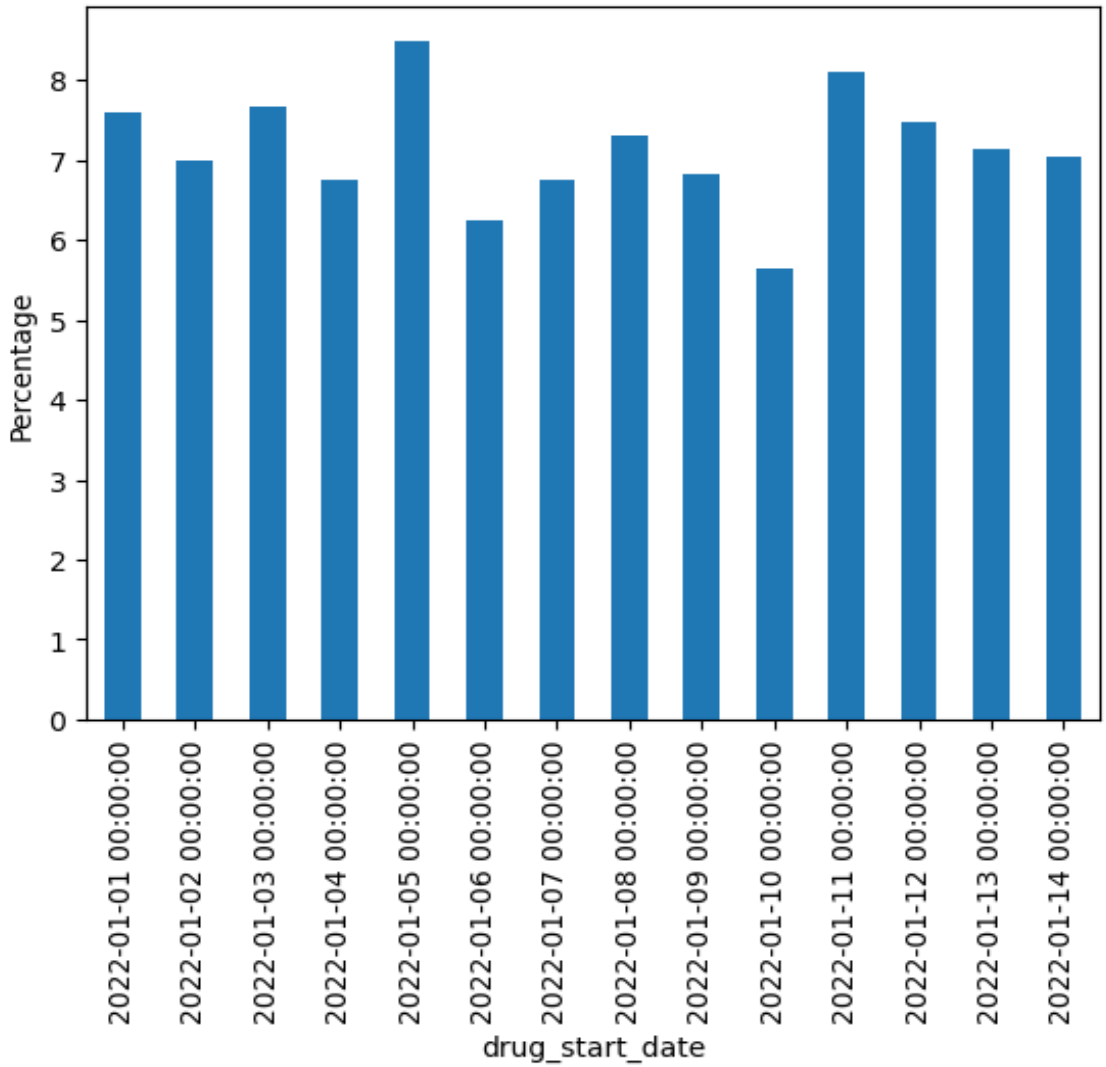


drug_name

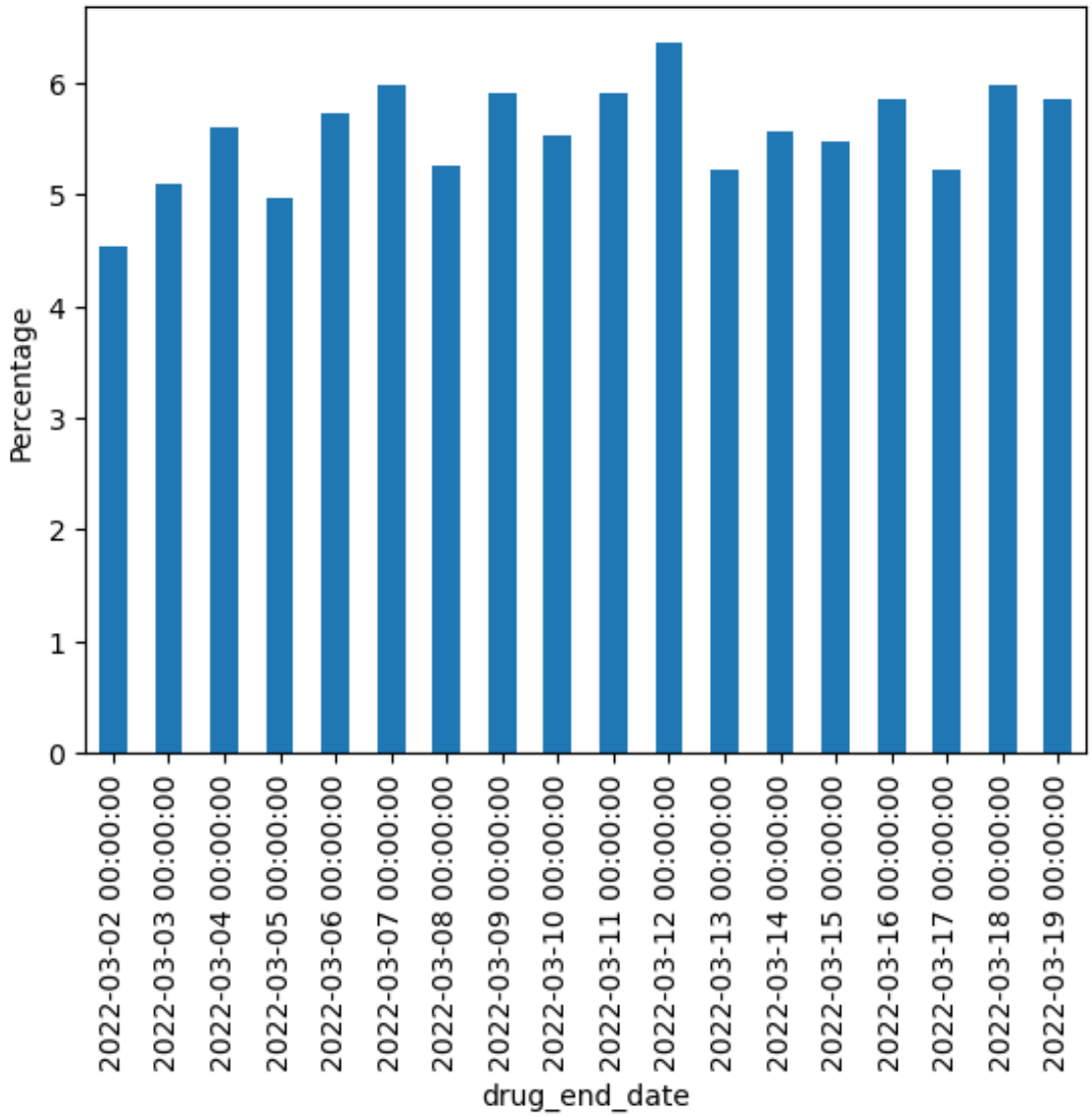
There are 151 different drugs.

drug_start_date & drug_end_date

Drug start date begins from 1 Jan 2022 and ends in 14 Jan 2022. Each day contains approximately 170 entries.

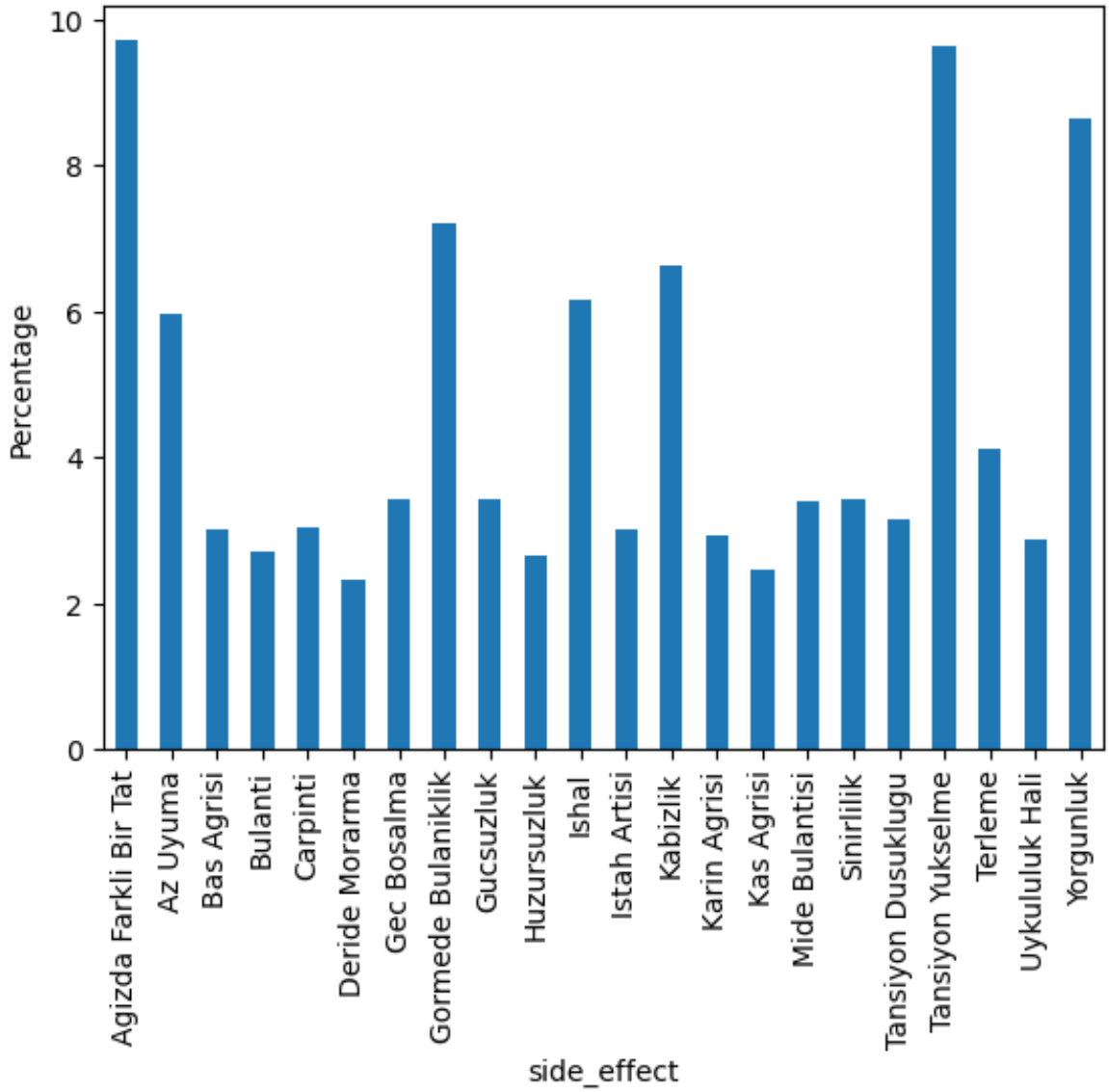


Drug end date begins from 2 March 2022 and ends in 19 March 2022. Each day contains approximately 130 entries.



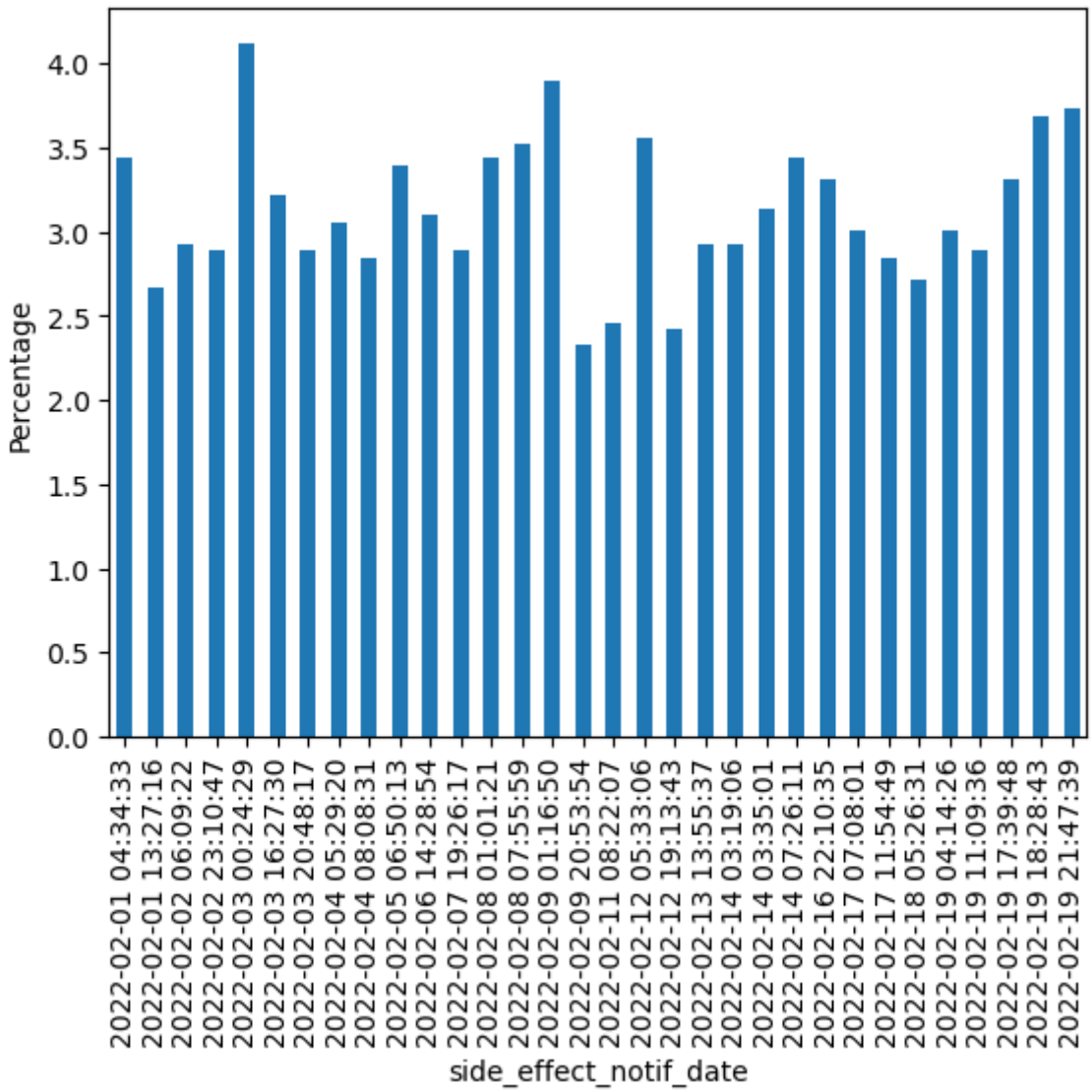
side_effect

There are 22 different side effects. Most seen side effects are “Agizda farkli bir tat”, “tansiyon” and “yorgunluk”.



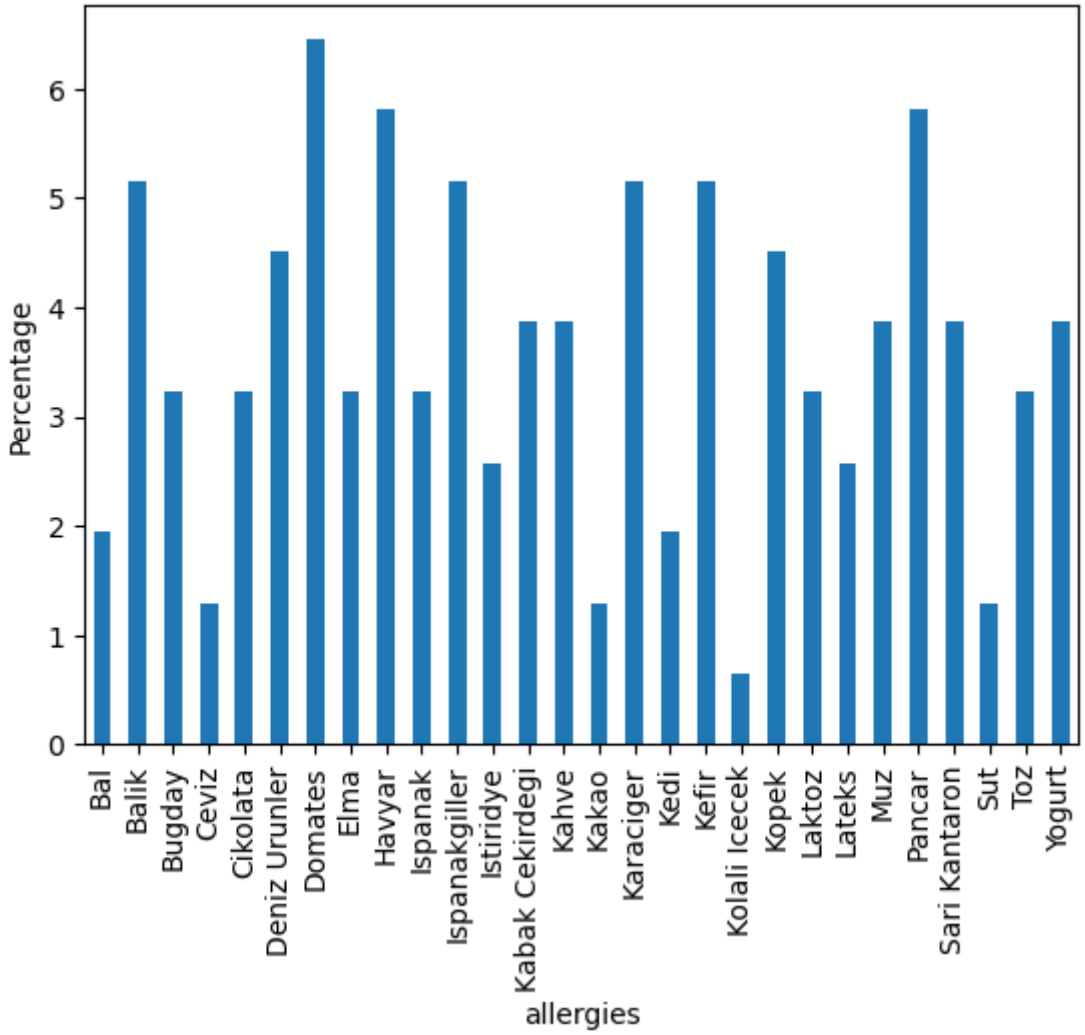
side_effect_notif_date

Side effect notification begins from 1 February 2022 and ends in 19 Februaray 2022. Each day contains approximately 70 entries. In some days there are more than one notifications.



allergies

There are 28 different allergies. Some users' entries are missing, yet there is no "no allergy" class. Therefore we will consider missing values as a none allergy class.



chronic_diseases

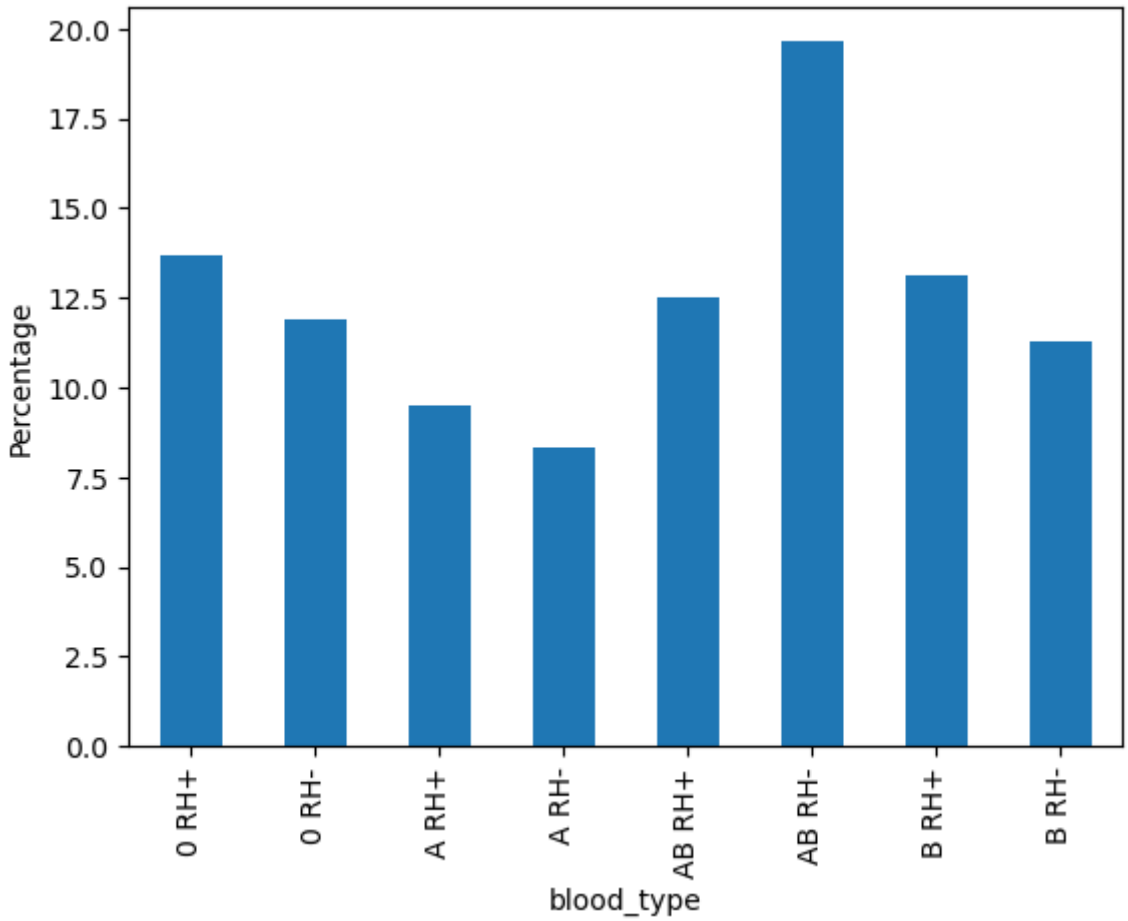
At first glance, there are 80 different chronic diseases. But looking at the column, each entry contains two diseases separated by commas. Some people have one disease, some have two, some have none. In addition, this is not an ordinal categorical value, so OrdinalEncoder will not work well with this data because there is no order between the classes. When this property is encoded with OneHotEncoder, it will turn into a vector of length 80 for each entry. It is best to split this property into two different properties, **chronic_diseases_1** and **chronic_diseases_2**, and encode them with OneHotEncoder.

family_chronic_diseases

These fields will be considered together, above solution will be applied to them as well.

blood_type

There are 8 different blood groups, of which AB Rh- is the most common and A Rh- the least common. The least common blood group in normal life is 0 Rh-, but there are some missing values, perhaps this imbalance can be corrected by filling in the missing values.



weight & height

These fields are numeric. 26 of people do not have weight and 10 of them do not have height information. However, people without weight information do have height information and vice versa. It was decided to use some threshold of +- 10 for weight information and using this, the user dataframe was filtered. Then, the height values of people without height values in this dataset were filled with the height values of the filtered data frame by matching gender. The same principle was applied for weight.

There is no meaningful relation with these and other numeric fields.

Once these values are filled in, the Body Mass Index can be calculated because there is a clear relationship between BMI and chronic diseases.

Date Features

Date features of this dataset are;

- drug_start_date [format: (YYYY-MM-DD)]
- drug_end_date [format: (YYYY-MM-DD)]
- side_effect_notif_date [format: (YYYY-MM-DD hh:mm:ss)]
- birthdate [format: (YYYY-MM-DD)]

Because month and year are the same for drug_start_date they are not meaningful features but day of this dates might be meaningful since they differ.

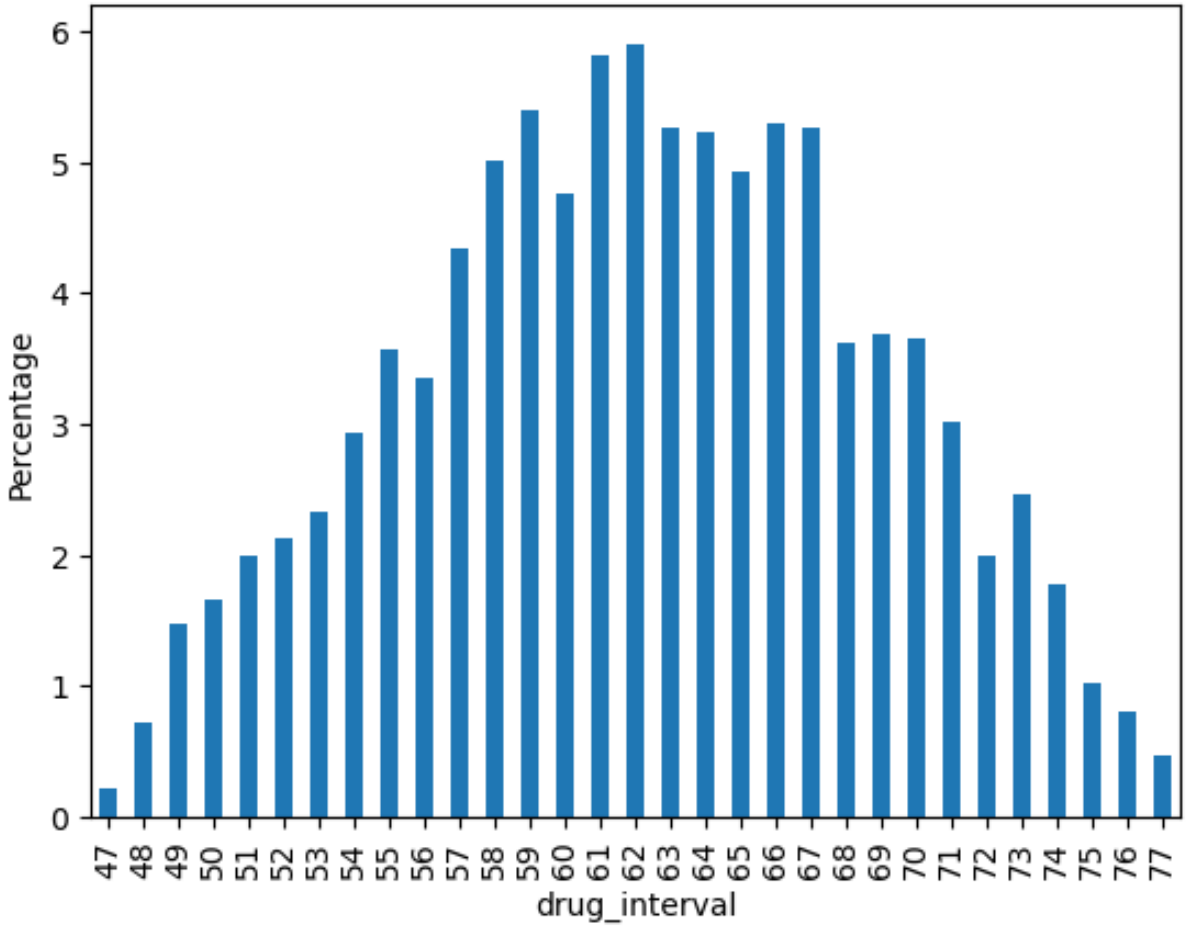
Features drug_start_day and drug_end_day are created, but there are 2 months interval between this dates. Therefore, creating drug_interval feature might be more useful than these two fields. For comparison purposes these two features will not be deleted, yet they are not planned to use in the last model.

Note: **drug_start_day** and **drug_end_day** plots has same distribution with **drug_start_date.**

drug_interval

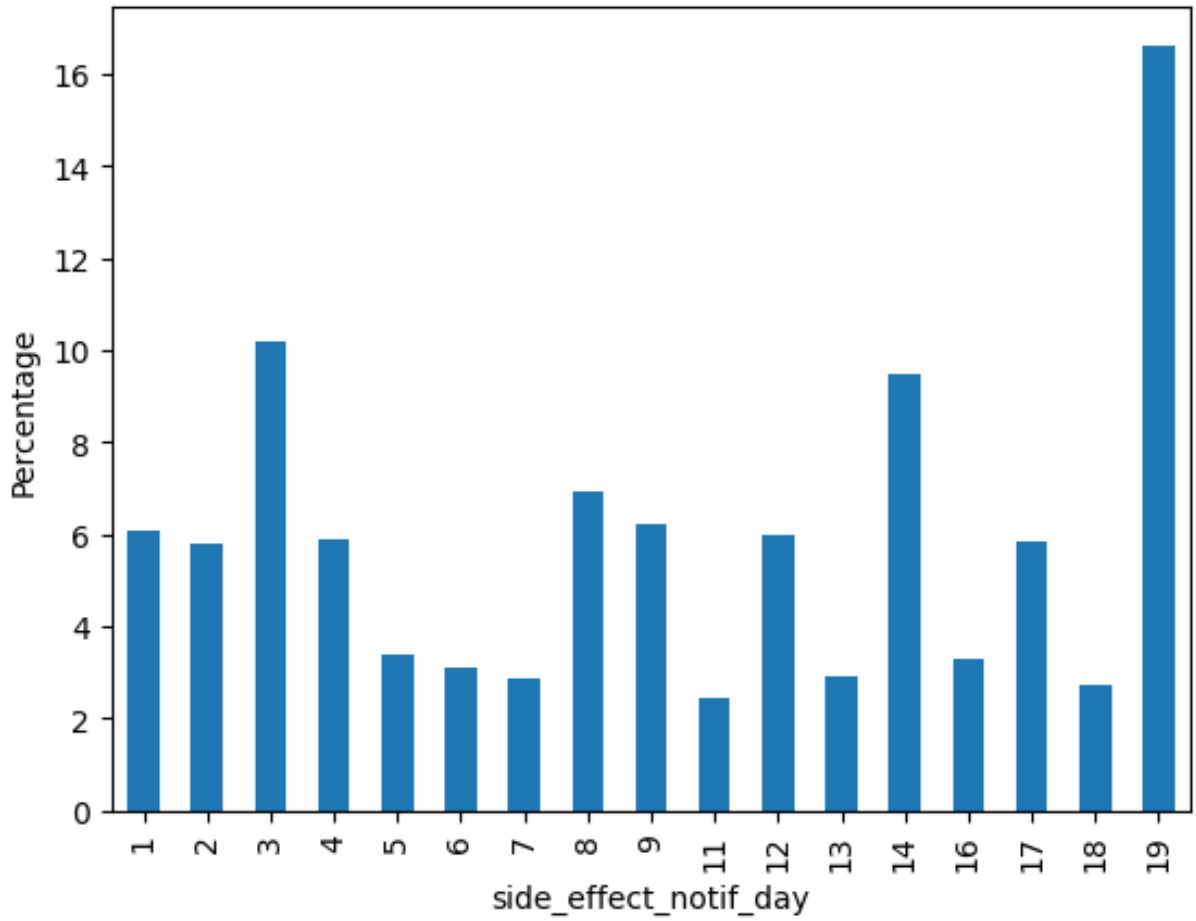
This is considered to be a numerical feature and since this is created from other features it is high correlated with them.

This features' shape is similar to normal distribution, so it can be a good predictor for the model.

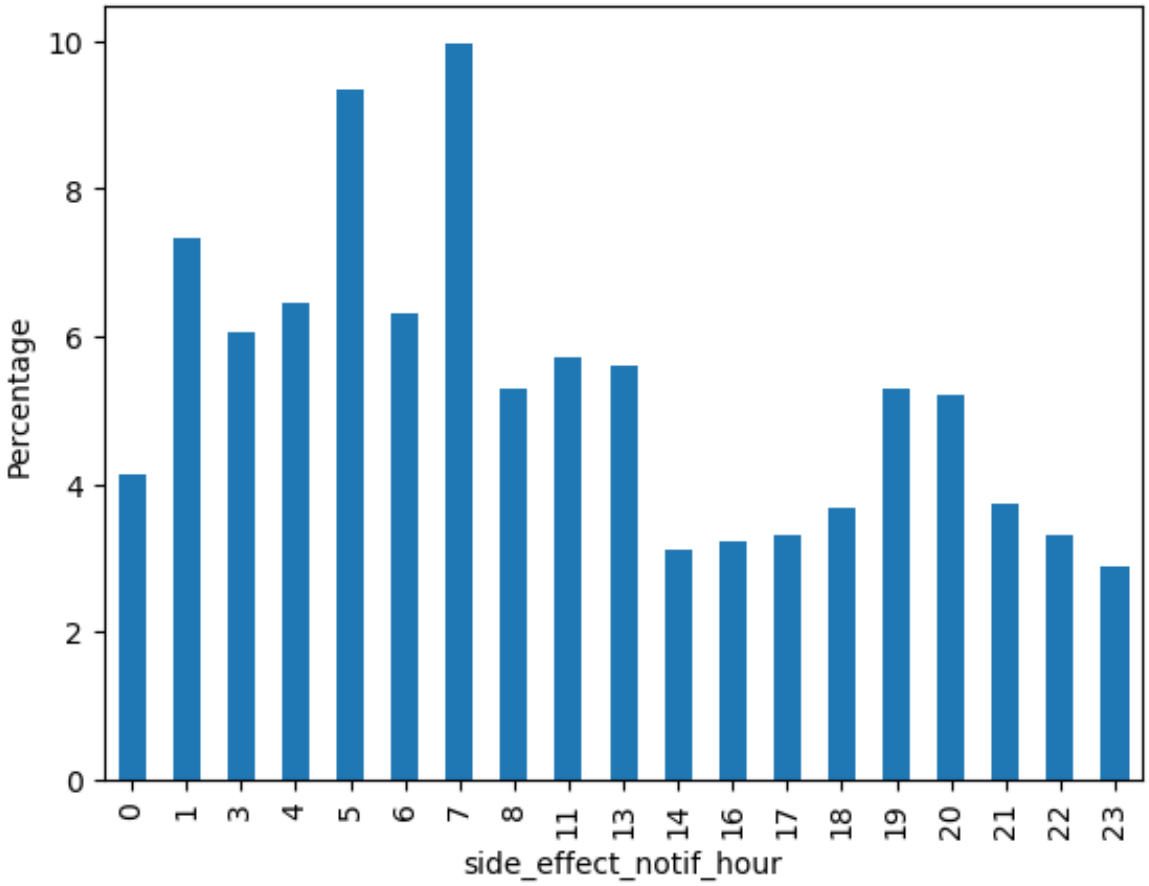


side_effect_notif_day and side_effect_notif_hour

The only meaningful attribute in the side_effect_notif_date field is the day and time of the value of this date. The day with the most notifications is by far the 19th day of February. If the drug_start_date field is examined, it can be seen that it is neither skewed left nor right. This is not related to the start_date of the drug.

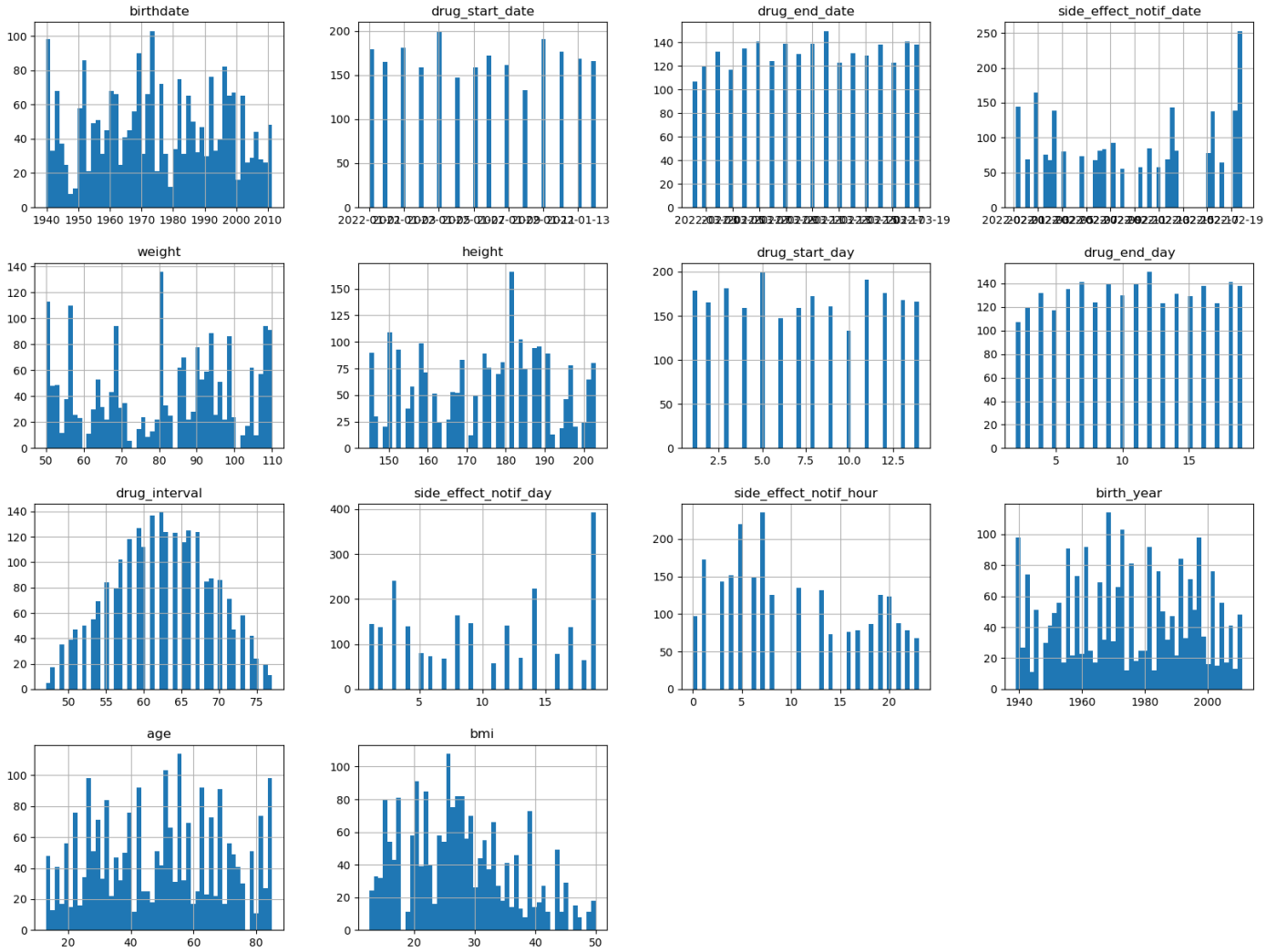


Before noon and after midnight there are more notifications. It is an expected result.



These features are considered to be numeric similar to other datetime features.

Histogram of Numeric Fields

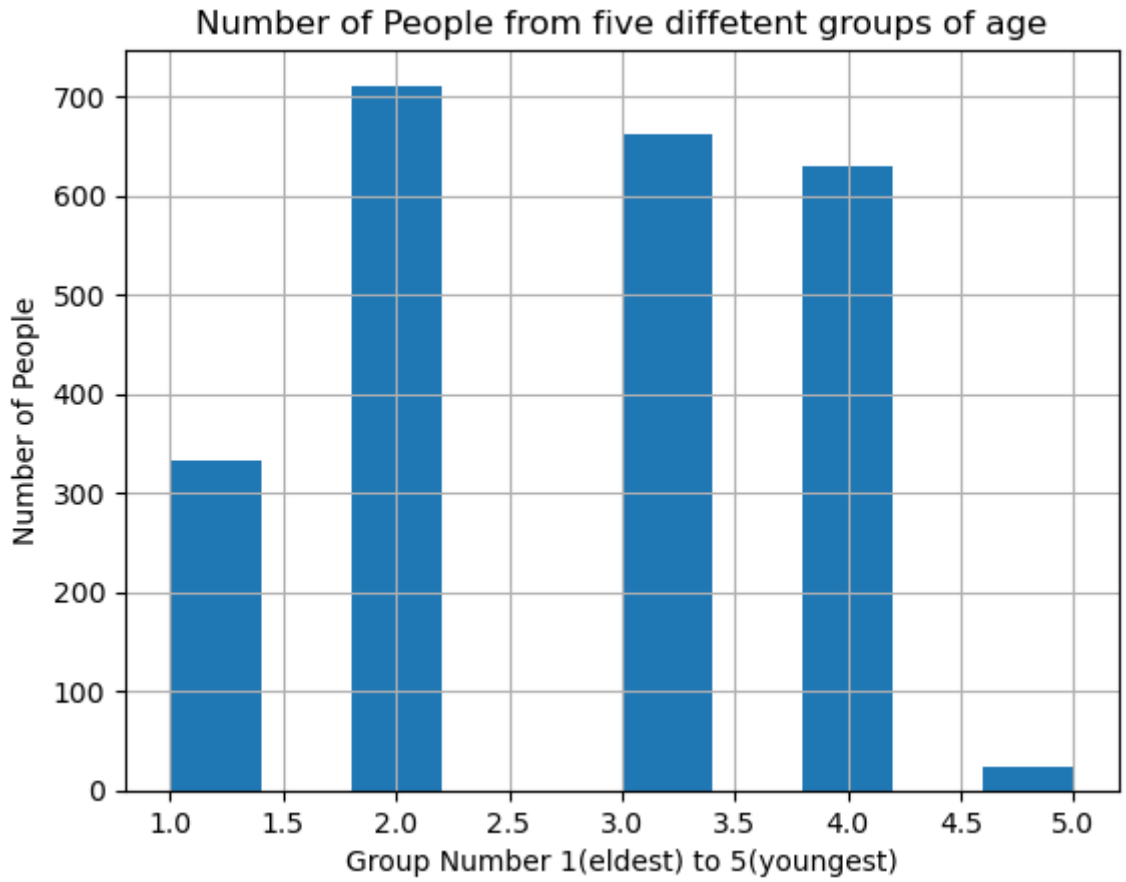


Feature Engineering

The **birth_year**, **age** and **bmi** features were created in addition to the engineered datetime features. Birth year and age are highly correlated like expected, age would be a useful predictor in that case.

Moreover, **age_group** feature were created and birth year are divided into 5 different groups;

1) 1930 – 1960 born
2) 1960 – 1980 born
3) 1980 – 2000 born
4) 2000 – 2010 born
5) 2010 – this year born

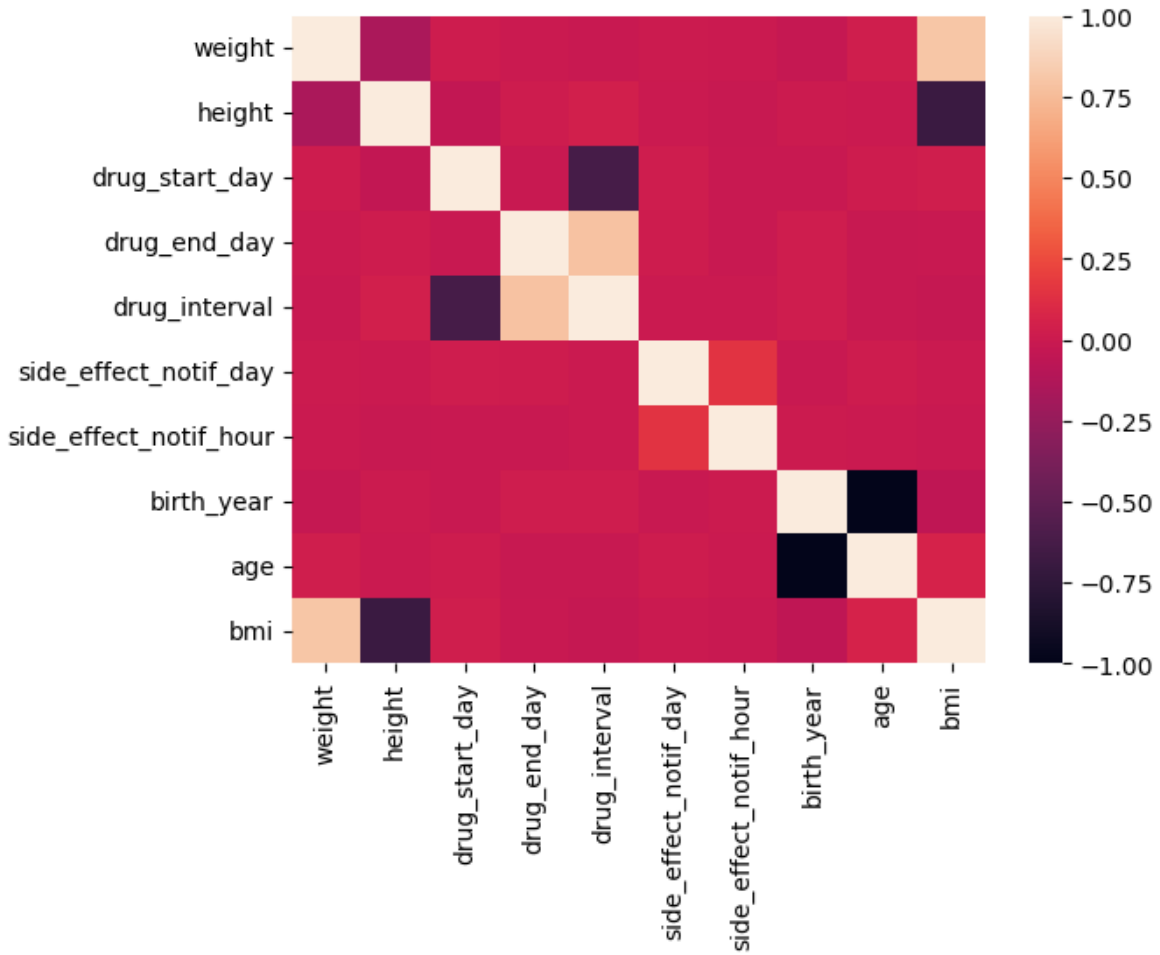


This feature comes in handy while splitting dataset into training and test.

Correlation of Features

There are total 10 numeric features with the added ones. When we look at this heatmap, several conlusions glance our eyes.

I. Weight and height highly correlated with bmi this is expected since bmi is engineered from them.
II. Like above drug_start_day and drug_end_day is highly correlated with drug_interval. This result proves our point about dropping this features.
III. Birth year and age are the same.
IV. There is no meaningful correlations between numeric fields except engineered ones.



Training and Test Split

There are 2357 entries in this dataset. This is very few compared to classic Machine Learning datasets. Normally, splitting dataset randomly 80% for train and 20% for test would be enough. Because this dataset is way smaller than normal standards, another way of splitting will be used called StratifiedKSplit and age_group feature will be the criteria of split.

A. Normal Split

Train set contains 1885 samples and test set contains 472 samples randomly sampled from dataset. If we calculate each groups' ratio in this train and test sets, it can be seen they have different ratios.

Train Set:

1) 14.6%
2) 29.6%
3) 27.9%
4) 26.9%
5) 0.01%

Test Set:

1) 11.9%
2) 32.4%
3) 28.8%
4) 25.6%
5) 0.01%

B. Stratified Split

Again train and test set contains 1885 and 472 samples, yet their sample proportions are different and based on age of the individuals. Trainining and testing set are much more similar in that sense. The question comes to mind. Why age? In this kind of problem, people's age affects their health status, which is why age criteria was chosen for splitting. However, other features can be tested and may be more practical.

Train Set:

1) 14.1%
2) 30.2%
3) 28.1%
4) 26.7%
5) 0.01%

Test Set:

1) 14.2%
2) 30.1%
3) 28.0%
4) 26.7%
5) 0.01%

Missing Values

For gender, allergies, chronic_diseases, blood_type, city and family_chronic_diseases; new class called "empty" created and filled those missing values with this value.

For weight and height, dataframe is filtered and averaged like mentioned in **Weight and Height** section. BMI is engineered from these two features so there is no need to fill that.

```
gender                     778
allergies                  484
bmi                        407
chronic_diseases           392
blood_type                 347
weight                     293
city                       227
mother_chronic_diseases    217
father_chronic_diseases    156
brother_chronic_diseases   121
height                     114
sister_chronic_diseases     97
side_effect_notif_date       0
side_effect                  0
drug_end_date                0
drug_start_date              0
drug_name                    0
origin                       0
birthdate                    0
drug_start_day               0
drug_end_day                 0
drug_interval                0
side_effect_notif_day        0
side_effect_notif_hour       0
birth_year                   0
age_group                    0
age                          0
user_id                      0
```

Additionally,

- Blood_type can be filled with some other technique.
- Empty City values considered to be people who come from abroad or not citizen of Turkey.
- Chronic_diseases is going to be filled according to bmi. If bmi is higher than specified threshold, empty value of person will be filled with chronic diseases of those people. Otherwise, it will be filled with “empty” class.
- Family chronic diseases can be filled with other families chronic diseases. However, i think that it will increase imbalance on the dataset, so it is not implemented.

Irrelevant Features

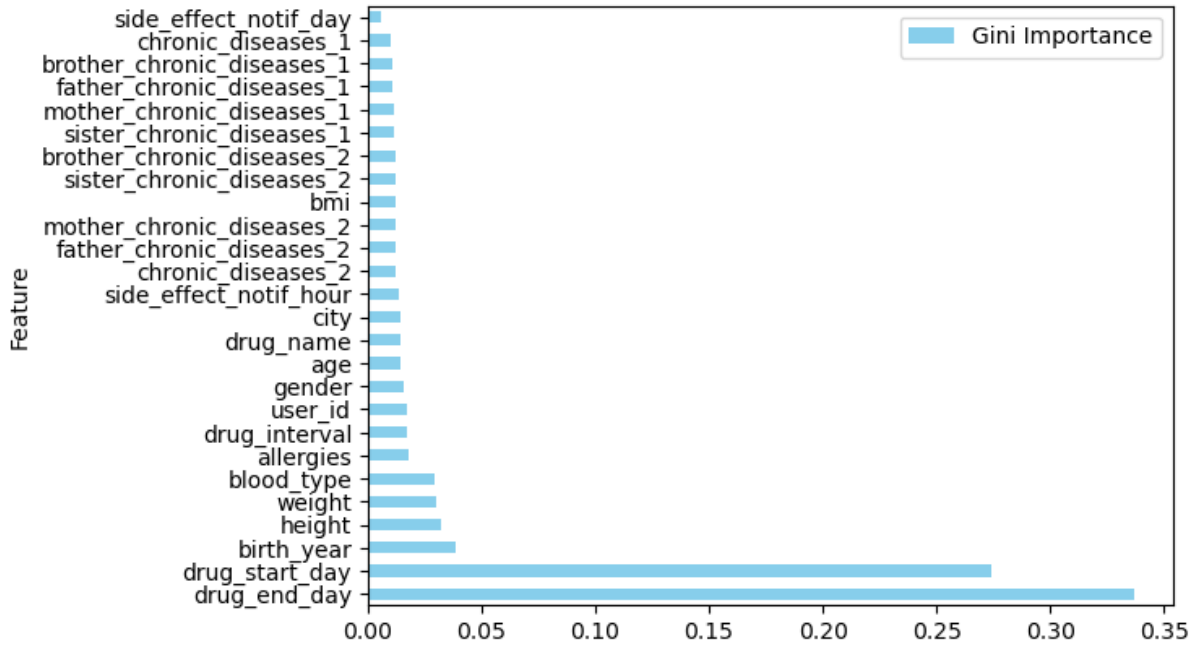
The following features have either been transformed into other features or do not contain important information;

'birthdate', 'origin', 'drug_start_date', 'drug_end_date','side_effect_notif_date'

so they will be discarded from dataset.

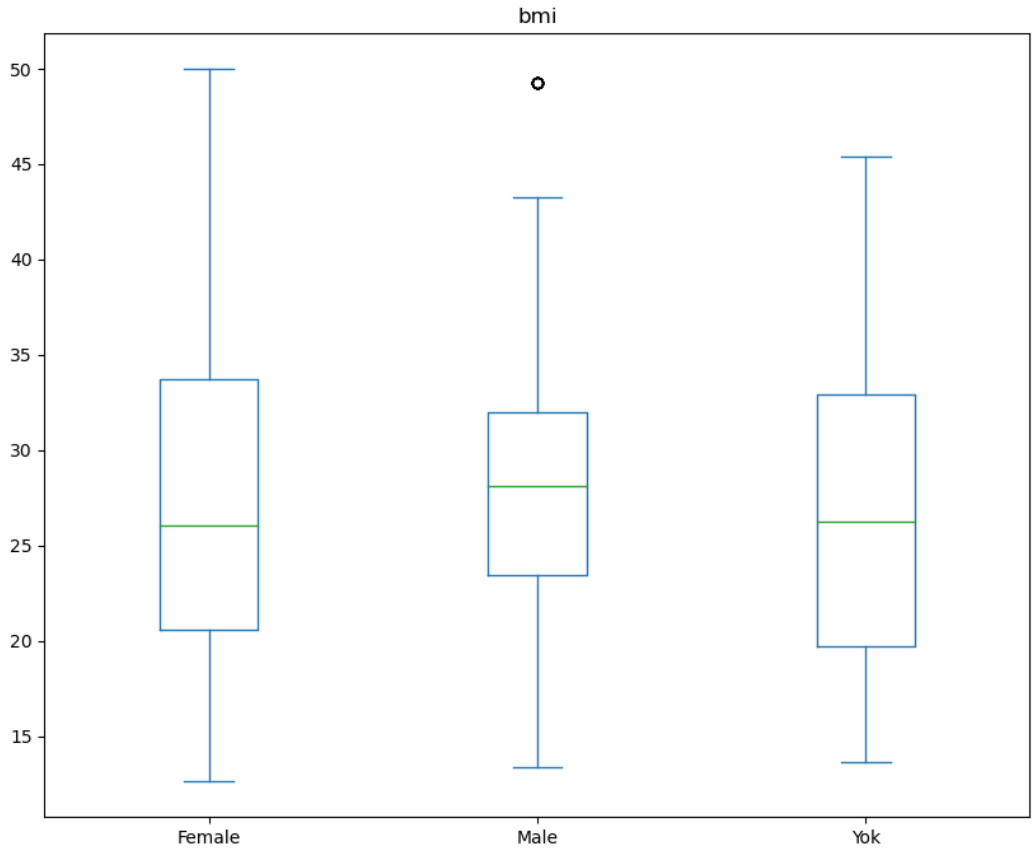
Feature Importance

There is no specified target variable that is expected to be predicted. However, knowing importance of each feature is insightful for our purposes. Side Effect field seems suitable fort his task. Predicting side effect is a classification task, so RandomForestClassifier was used since they are not prone to overfitting compared to other models. If below graph is inspected, drug_start_day and drug_end_day seems to be most important features for predicting **side_effect**, and side_effect_notif_day is least important feature.



Outliers

Firstly, boxplot of each features were plotted. All the data is in the interval of Q3 – Q1.In order to analyze outliers more carefully, Skewness, lower and upper bound of each quartile calculated and values bigger or smaller than these quartiles discarded from the dataset.



	Skewness
weight	-0.109103
height	-0.148114
birth_year	-0.035445
age	0.035445
bmi	0.496764
drug_start_day	0.036617
drug_end_day	-0.032704
drug_interval	-0.056719
side_effect_notif_day	0.029600
side_effect_notif_hour	0.335697

Clustering and Dimensionality Reduction for Visualization

Learning feature importance of each variable reveal some insights about dataset. In this regard each categorical data was labeled with OrdinalEncoder, but in this dataset most of the features have nominal type meaning there is no order among values. Therefore, for clustering OrdinalEncoder is changed with OneHotEncoder.

As a result, our feature dimension increased to 552. Then, data is divided into 3 clusters. Visualizing this 552 dimensions is not possible for 3 dimensional world that we live, so with the help of PCA data reduced to 2 dimension and 3 clusters were plotted as the below graph.

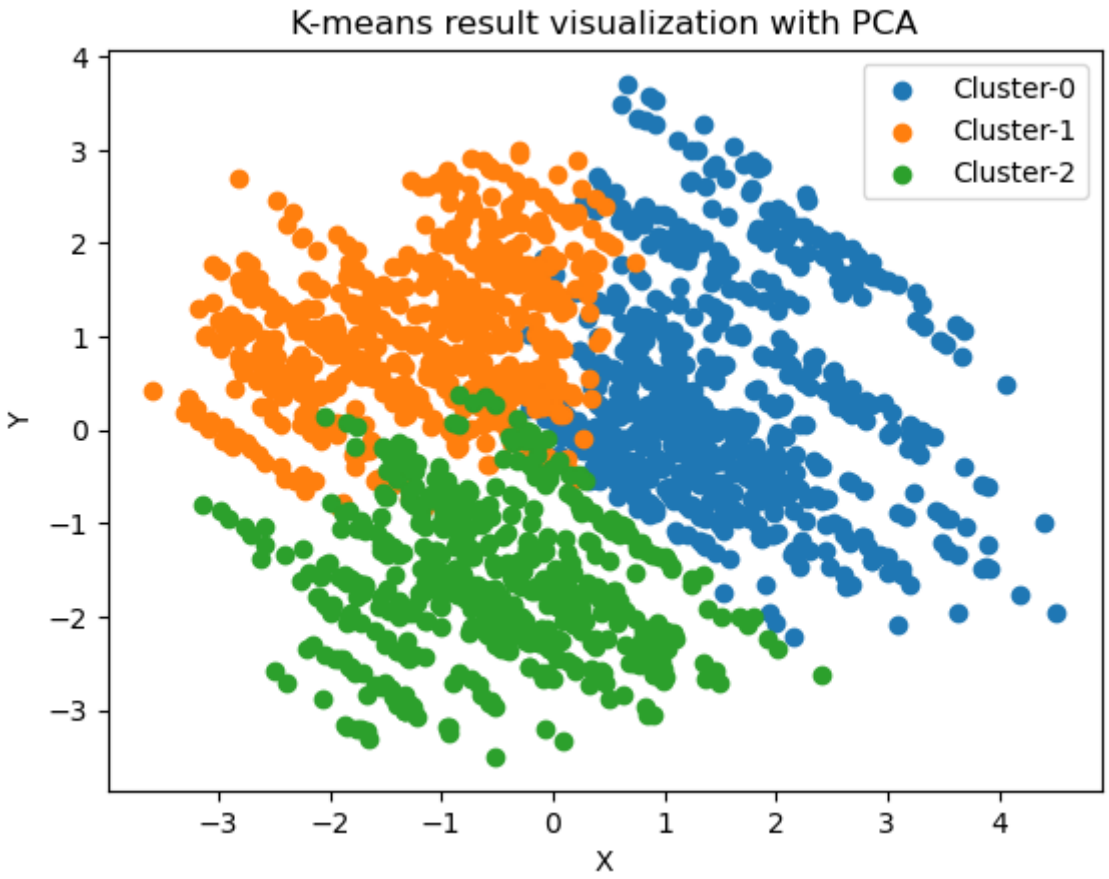
There are 3 distinct clusters in the below graph and each of them overlap in some points. In first glance, there are no clear meaning about these clusters. However, when data of each cluster was inspected. It can be seen that they diverge at some points.

Cluster 0 age column's median is 51.

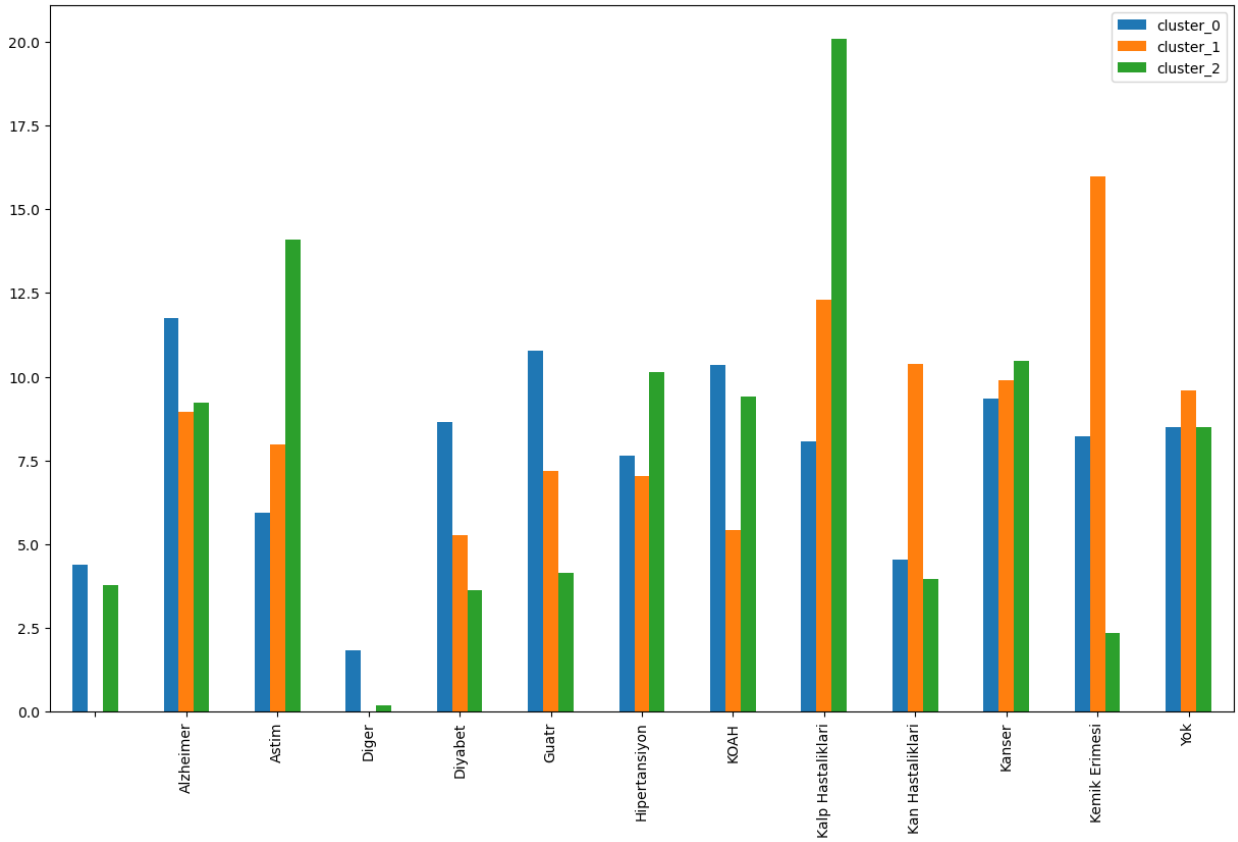
Cluster 1 age column's median is 29.

Cluster 2 age column's median is 72.

Hence, they are from different age groups.



Moreover, they have different distrubitions regarding family chronic diseases (father, mother).



Conclusion

Random forest gives 100% accuracy score on the train data and 85% on the test data. It clearly overfits. Different fine-tuning techniques have been tried, yet it couldn’t improve the model. This is mostly because of the small dataset we have. This can effect future predictive models, hence it should be taken into account.