Total 2357

Cinsiyet 1579 +

Il 2130 +

Alerjilerim 1873 +

KronikHastaliklarim 1965 +

BabaKronikHastaliklari +

AnneKronikHastaliklari +

Kiz Kardes Kronik Hastalılari +

Erkek Kardes Kronik Hastaliklari +

Kan Grubu

Kilo 2064 mean:80 +

Boy 2243 mean:174 +

65 kişinin cinsiyeti yok

Cinsiyet belirtilmemiş diycegiz 3 kategori M, F, Unkown

20 kişinin il bilgisi yok

41 kişinin Alerji bilgisi yok, 29 farklı alerji var + 1 NaN

Alerjisi olmayanlar NaN girilmiş

31 kişinin kronik hastalık bilgisi yok

Toplam 21 tane kronik hastalık var

13 kişinin baba kronik hastalikları yok

17 kişinin anne kronik hastalıkları yok

8 kişinin kız kardeş

11 kişinin erkek kardeş

28 kişinin kan grubu bilgisi yok

10 kişinin boy bilgisi yok

26 kişinin kilo bilgisi yok

Boyu olmayanların kilosu var, boyu olmayanların kilosu aynı olanlarla boy ortalaması

Kilosu olmayanların da boyu var

Null Values

Cinsiyet 778

Alerjilerim 484

Kronik Hastaliklarim 392

Kan Grubu 347

Kilo 293

Il 227

Anne Kronik Hastaliklari 217

Baba Kronik Hastaliklari 156

Erkek Kardes Kronik Hastaliklari 121

Boy 114

Kiz Kardes Kronik Hastaliklari 97

Kullanici\_id 0

Yan\_Etki 0

Ilac\_Bitis\_Tarihi 0

Ilac\_Baslangic\_Tarihi 0

Ilac\_Adi 0

Uyruk 0

Dogum\_Tarihi 0

Yan\_Etki\_Bildirim\_Tarihi 0

dtype: int64

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 lower english words. Attributes of this dataset and each attributes types are like following;

|  |  |
| --- | --- |
| **Variable** | **Type** |
| Kullanıcı\_id | Categorical |
| Cinsiyet | Categorical |
| Doğum Tarihi | Categorical |
| Uyruk | Categorical |
| Il | Categorical |
| Ilac\_adi | Categorical |
| Ilac\_baslangic\_tarihi | Categorical |
| Ilac\_bitis\_tarihi | Categorical |
| Yan\_etki | Categorical |
| Yan\_etki\_bildirim\_tarihi | Categorical |
| 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 of these features will be important from two perspectives, one is variable is ordered or not, two the number of classes to decide which encoder style will be used.

Let’s look at each attribute closely:

Not: From now on, we will use naming convention like mentioned above.

**user\_id**

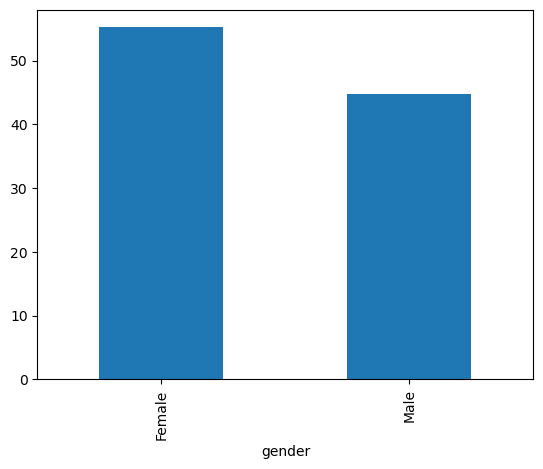
There are total 196 unique users and each user has 12 entries on average. Total spread of user counts is like a normal distribution. Some users has more than 20 entries and some others has less than 5. However, most of them is on the average.

There is no missing data on this field.

**gender**

There are 2 genders Male and Female and 54% of these users 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.

872 female, 707 male, 778 NaN



**birthdate**

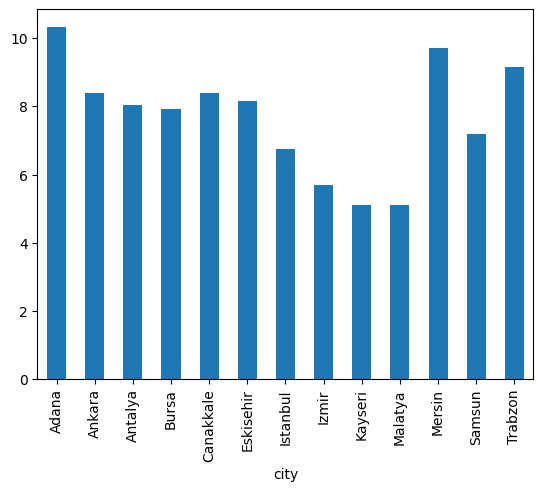
This is a datetime feature and related to age of the person. Changing this feature to age might be useful. Other than that i do no think datetime attributes of this feature is useful.

**origin**

This feature contains only one class so dropping this field is sensible.

**city**

Adana has most data points with 10% and least cities are İzmir, Malatya and Kayseri with 5%.

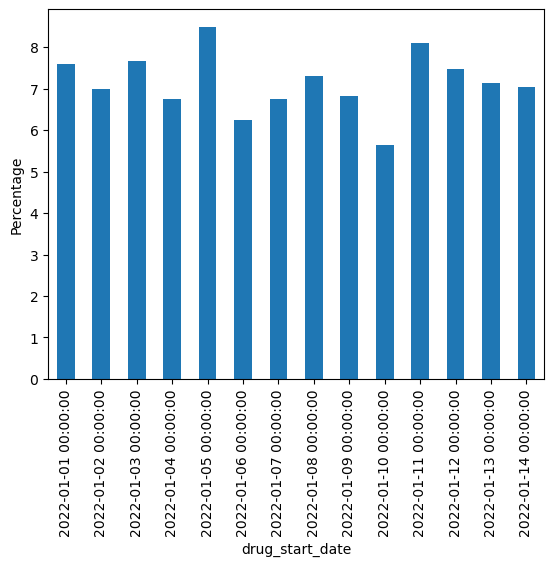


**drug\_name**

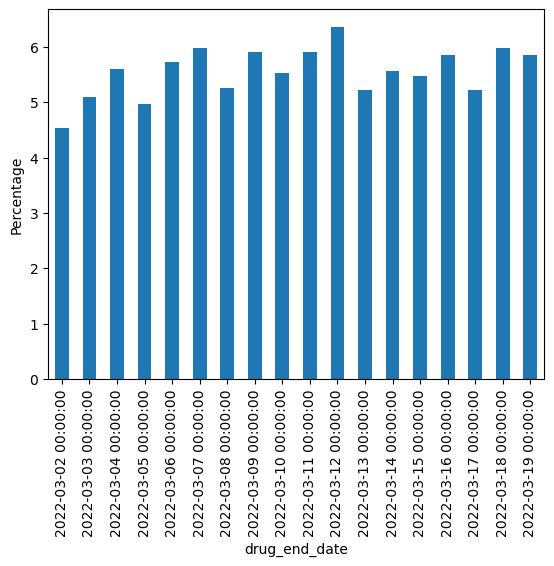
There are 151 different drugs.

**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.

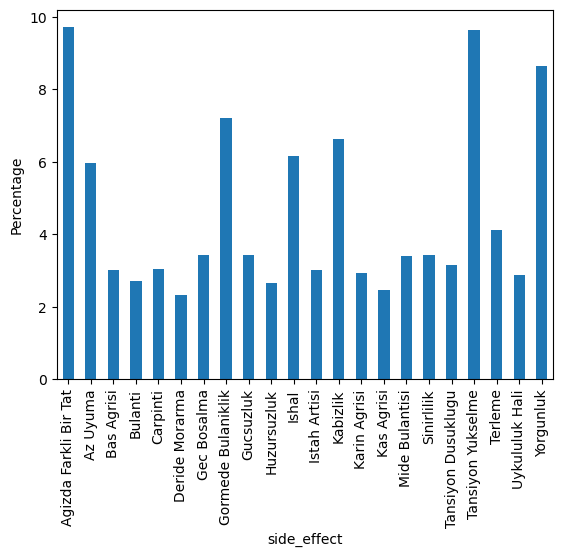


Drug end date begins from 2 March 2022 and ends in 19 March 2022. Each day contains approximately 130 entries.



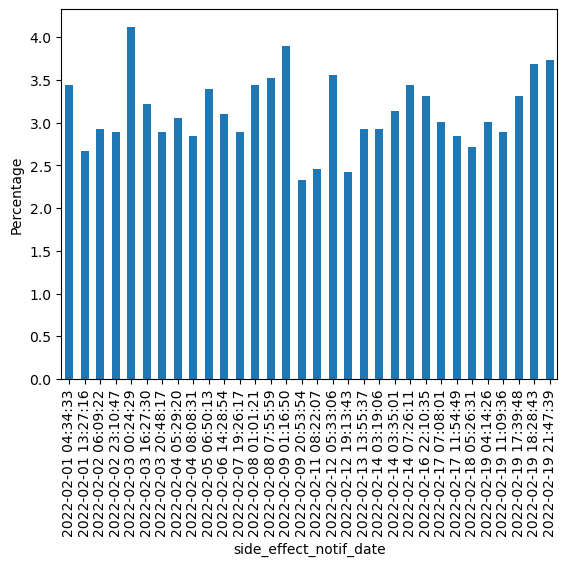
**Side\_effect**

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



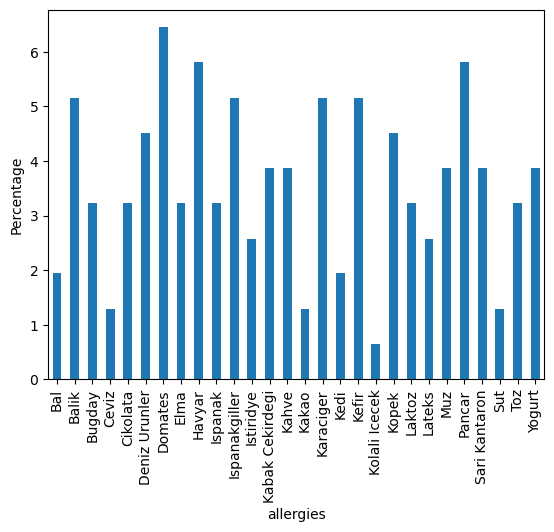
**Side\_effect\_notif\_date**

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**

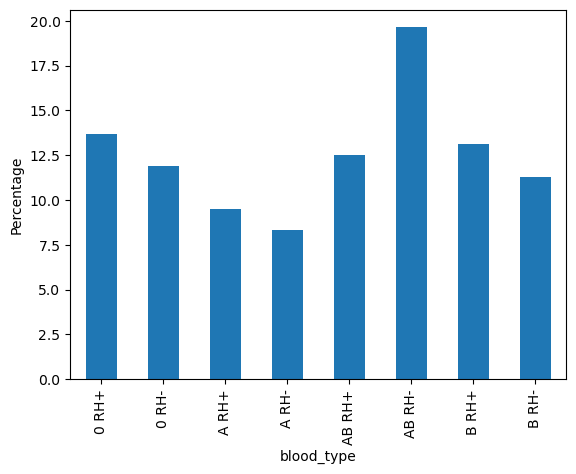
There are 80 different chronic diseases in the first glance, but when we look at the column. Each entry contains two disease with column separated. Some people contains one disease, some two, and some none. This is not an ordinal categorical value, so OrdinalEncoder will not give good results in this data because there is no order among the classes. OneHotEncoder will contain 80 length vectors for each entry. Splitting this feature in the two different features chronic\_diseases\_1 and chronic\_diseases\_2 and encoding with OneHotEncoder will work best.

**Family\_chronic\_diseases**

These fields will be considered together and above problem and solution is also on these fields.

**Blood\_type**

There are 8 different blood type. Most of them AB Rh- and least of them A Rh-. 0 Rh- expected to be least amon them but there are some missing values maybe this can be corrected by filling these values.



**Weight & Height**

These fields are numerical. 26 of people does not have weight and 10 of them does not have height information. However, people who do not have weight information does have height information and vice versa. I decided to use some threshold for weight information like +- 10 and filtered user dataframe by using this. After that i averaged people’s height on this dataframe that have same gender with the person who has no height value. Same principle applied for weight.

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

After filling these values Body Mass Index might be calculated since there are clear relation with bmi and chronic diseases.

**Datetime Features**

Datetime 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.

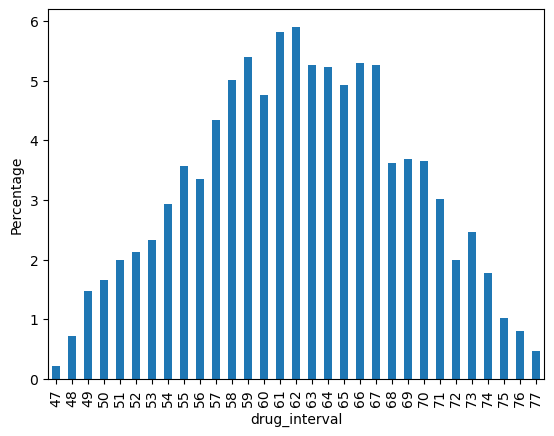
Features drug\_start\_day and drug\_end\_day are created, but there are 2 months interval between this dates. Therefore, creating drug\_interval feature might be more useful than these two fields. For comparison purposes we will not delete these two features, yet we do not plan to use in the model.

**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.

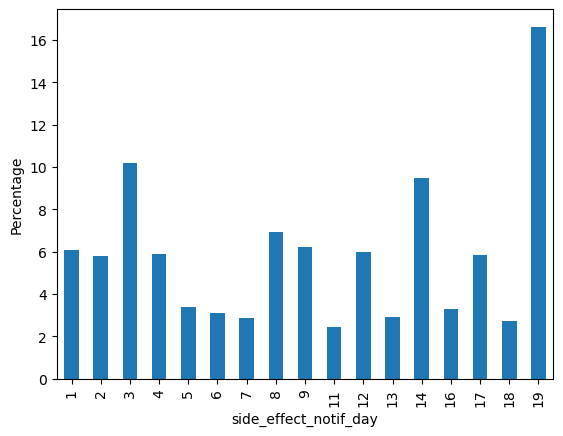
This features’ shape is similar to normal distribution, so it can be a good predictor for the model.



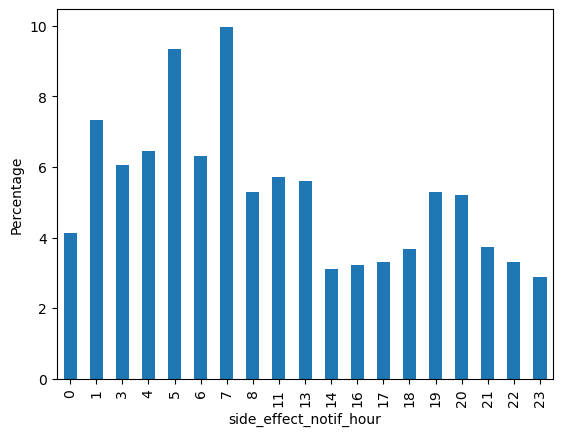
**side\_effect\_notif\_day and side\_effect\_notif\_hour**

The only different part in the side\_effect\_notif\_date field accross the dataset day and hour of this datetime.

Most bildirilen day is by far on 19th day of February. If we inspect drug\_start\_date, it is neither skewed to left nor to right. This is not about start\_date of the drug.



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



These features are again considered to be numeric.

**Histogram of Numeric Fields**

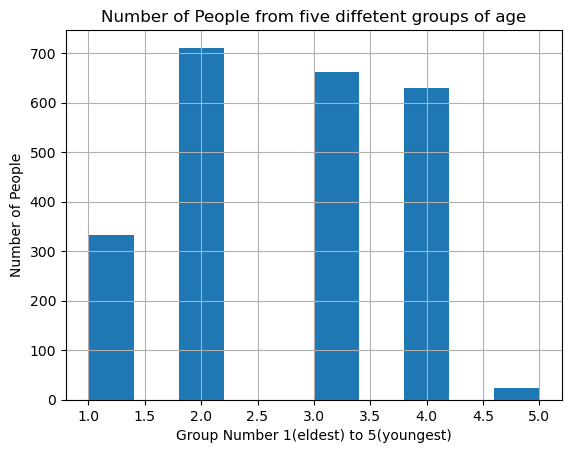
****

**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. There are 5 different groups;

1. 1930 – 1960 born
2. 1960 – 1980 born
3. 1980 – 2000 born
4. 2000 – 2010 born
5. 2010 – now born

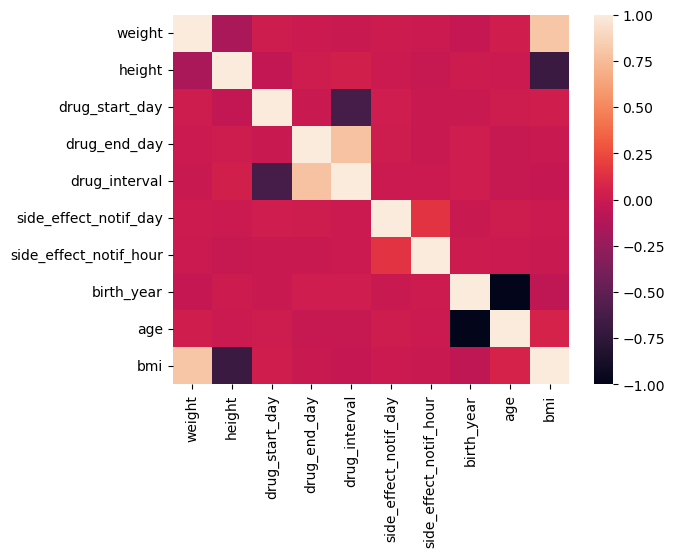


This feature comes in handy while splitting dataset into training and test.

**Correlation of Features**

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

1. Weight and height highly correlated with bmi this is expected since bmi is engineered from them.
2. Like above drug\_start\_day and drug\_end\_day is highly correlated with drug\_interval. This proves our point about dropping this features.
3. Birth year and age is the same.
4. 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 is used called StratifiedKSplit and age\_group feature will be the feature split criteria based on.

1. **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: Test Set:**

1. 14.6% 1) 11.9%
2. 29.6% 2) 32.4%
3. 27.9% 3) 28.8%
4. 26.9% 4) 25.6%
5. 0.01% 5) 0.01%
6. **Stratified Split**

Again train and test set contains 1885 and 472 samples, yet their sample ratio different and based on age of the people. Train and test much more similar in that sense. The question comes in our mind. Why age? In this kind of problem age of people is important so age criteria is chosen for splitting. However, other features can be tested and be more practical.

**Train Set: Test Set:**

1. 14.1% 1) 14.2%
2. 30.2% 2) 30.1%
3. 28.1% 3) 28.0%
4. 26.7% 4) 26.7%
5. 0.01% 5) 0.01%

**Missing Values**

**Irrelevant Features**

**Feature Importance**

**Outliers**

**Clustering and Dimensiality Reduction for Visualization**