Product Analyst Challenge

Kamilla Valeeva

Task 1. Explorative Analysis

I used Python (**pandas** library) to get familiar with the dataset and perform some data cleaning:

```
import pandas as pd

xls = pd.ExcelFile(r"C:\Users\never\Downloads\claims_provider_data_anonymized.xlsx")
claims = pd.read_excel(xls, "data_claims_anonymized")
provider = pd.read_excel(xls, "data_provider_anonymized")
```

`shape` – attribute, returns a tuple representing the dimensionality of the DataFrame.

`info()` – method, prints a concise summary of a DataFrame.

`describe()` – method, generates descriptive statistics that summarize the central tendency, dispersion and shape of a dataset's distribution, excluding NaN values, analyzes both numeric and object series

Task 1. Explorative Analysis

Claims dataset

```
claims.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 55462 entries, 0 to 55461
Data columns (total 19 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   row             55462 non-null  int64
 1   treatment_id    55462 non-null  object
 2   claim_id        55462 non-null  object
 3   claim_amount    55431 non-null  float64
 4   claim_status    55462 non-null  object
 5   payment_type    55462 non-null  object
 6   date_treatment  55462 non-null  datetime64[ns]
 7   date_submitted  46668 non-null  datetime64[ns]
 8   date_modified   55046 non-null  datetime64[ns]
 9   modified_by     55462 non-null  object
 10  provider_name   55462 non-null  object
 11  patient_ref_id  55462 non-null  object
 12  patient_gender  55450 non-null  object
 13  patient_age     55456 non-null  float64
 14  diagnosis_code  55462 non-null  object
 15  item_name       55462 non-null  object
 16  item_quantity   52386 non-null  float64
 17  item_amount     52386 non-null  float64
 18  item_status     52388 non-null  object
dtypes: datetime64[ns](3), float64(4), int64(1), object(11)
memory usage: 8.0+ MB
```

```
print(claims.describe())
```

	row	claim_amount	patient_age	item_quantity	item_amount
count	55462.0	55431.0	55456.0	52386.0	52386.0
mean	27730.5	1923.0	28.6	43.4	13538.4
std	16010.6	243093.1	19.6	6179.7	3024769.9
min	0.0	-128.0	-997.0	0.0	0.0
25%	13865.2	0.0	16.0	1.0	0.0
50%	27730.5	50.0	29.0	1.0	10.0
75%	41595.8	500.0	38.0	1.0	100.0
max	55461.0	55071840.0	138.0	1000000.0	692307693.0

All `claims_amount` negative values were only 0,02% of all values and all were with `claim_status` = ‘Rejected’ so I did not do anything with that.

```
print(claims.shape)
```

```
(55462, 19)
```

Task 1. Explorative Analysis

Claims dataset cleaning

I found 12 missing values in column `patient_gender`, so decided to replace it with “No Gender” string value to avoid nulls in data using `fillna` function:

```
claims["patient_gender"].fillna
("No Gender", inplace=True)
```

There were also found 6 missing values in column `patient_age`, so those were replaced by the median value using `fillna` function:

```
claims["patient_age"].fillna
(claims["patient_age"].median(), inplace=True)
```

Finally, I have found some negative values in column `patient_age`, so those were replaced by the median value using `lambda-expression`:

```
claims['patient_age'].apply(lambda x: x if x > 0
else claims["patient_age"].median())
```

```
claims.isnull().sum()
```

row	0
treatment_id	0
claim_id	0
claim_amount	31
claim_status	0
payment_type	0
date_treatment	0
date_submitted	8794
date_modified	416
modified_by	0
provider_name	0
patient_ref_id	0
patient_gender	12
patient_age	6
diagnosis_code	0
item_name	0
item_quantity	3076
item_amount	3076
item_status	3074
dtype: int64	

row	0
treatment_id	0
claim_id	0
claim_amount	31
claim_status	0
payment_type	0
date_treatment	0
date_submitted	8794
date_modified	416
modified_by	0
provider_name	0
patient_ref_id	0
patient_gender	0
patient_age	0
diagnosis_code	0
item_name	0
item_quantity	3076
item_amount	3076
item_status	3074
dtype: int64	

Task 1. Explorative Analysis

Provider dataset

```
provider.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 396 entries, 0 to 395
Data columns (total 4 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   provider_name         396 non-null    object
 1   provider_type         1 non-null      object
 2   provider_country      396 non-null    object
 3   provider_star_rating  396 non-null    int64
dtypes: int64(1), object(3)
memory usage: 12.5+ KB
```

```
print(provider.describe())
```

```
       provider_star_rating
count                 396.0
mean                    2.6
std                     1.2
min                     1.0
25%                     2.0
50%                     2.0
75%                     4.0
max                     5.0
```

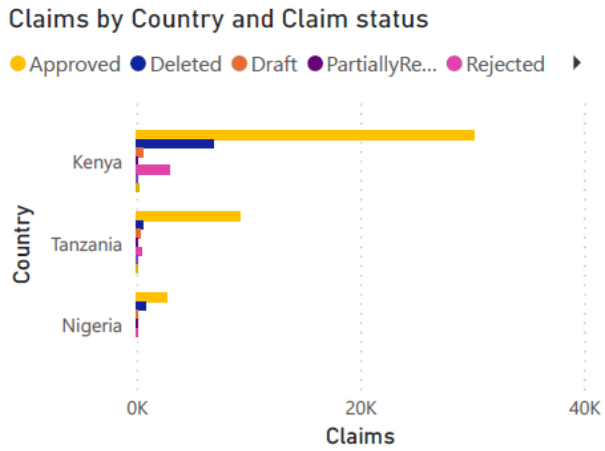
```
print(provider.shape)
```

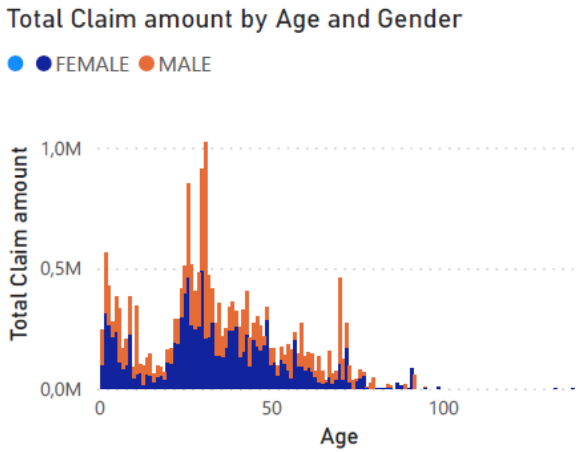
```
(396, 4)
```

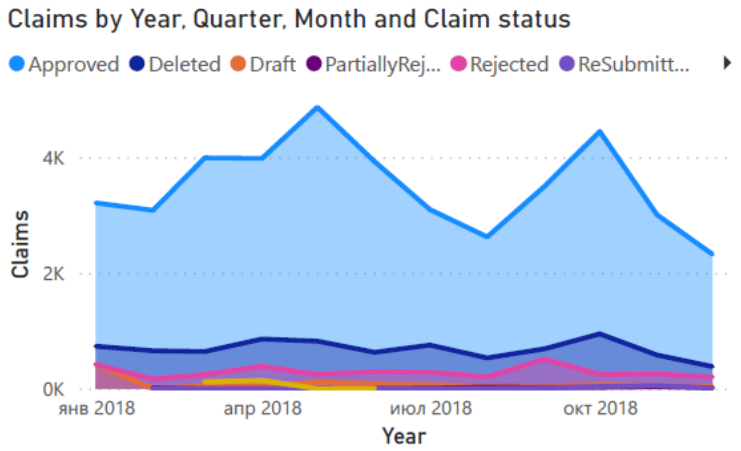
Task 1. Explorative Analysis

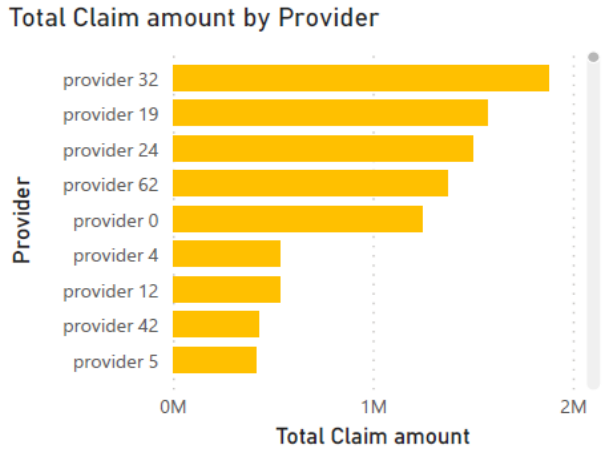
Analytics visualization

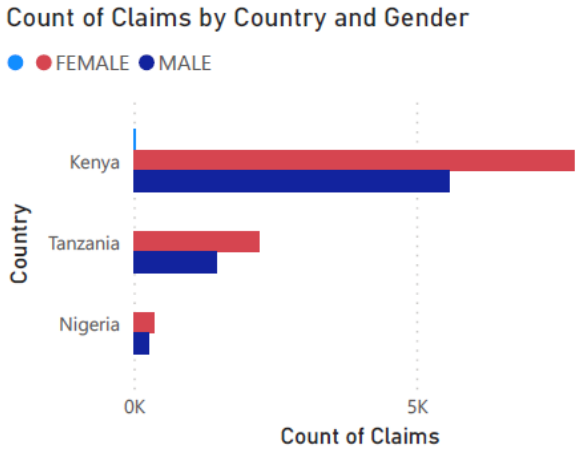
I have prepared an analytics dashboard in Power BI and used different visuals/metrics to find any data patterns, correlations and anomalies

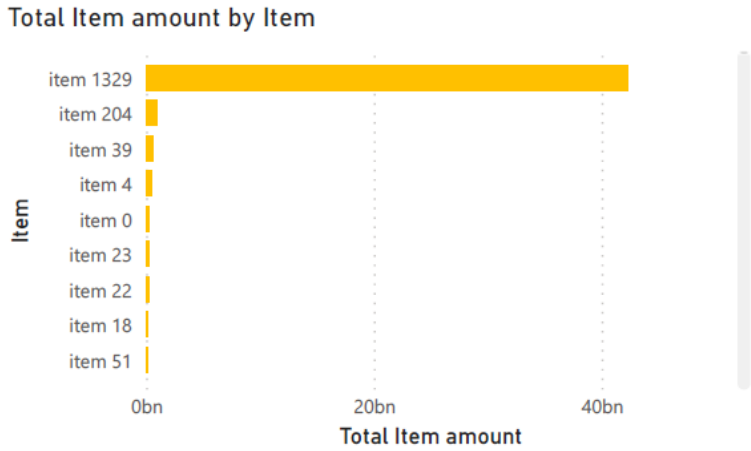






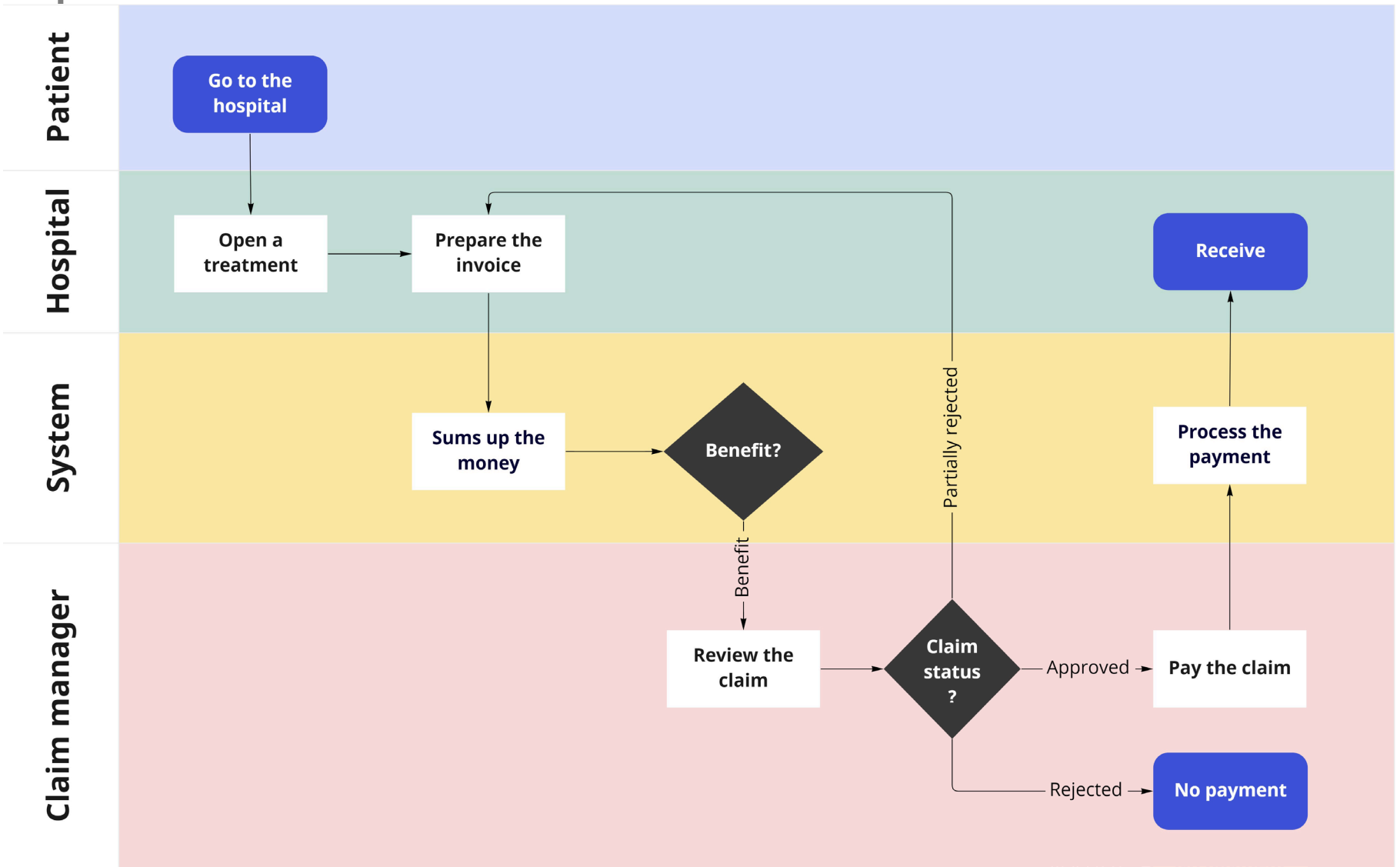






Task 1. Explorative Analysis

1b. Claim process visualization



Task 1. Explorative Analysis

1c. SQL Queries

What are the top 5 most expensive items (on average) for Kenyan providers?

```
SELECT top 5 item_name,
             average_amount = round(avg(item_amount*item_quantity), 1)
FROM dbo.data_claims c
JOIN dbo.data_provider p ON c.provider_name = p.provider_name
AND p.provider_country = 'Kenya'
WHERE item_status not in ('DELETED')
  AND item_status IS NOT NULL
  AND item_quantity > 0
  AND claim_status not in ('Deleted')
GROUP BY item_name
ORDER BY avg(item_amount*item_quantity) DESC
```

Task 1. Explorative Analysis

1c. SQL Queries

How many claims have been approved by 5-star rated Nigerian providers?

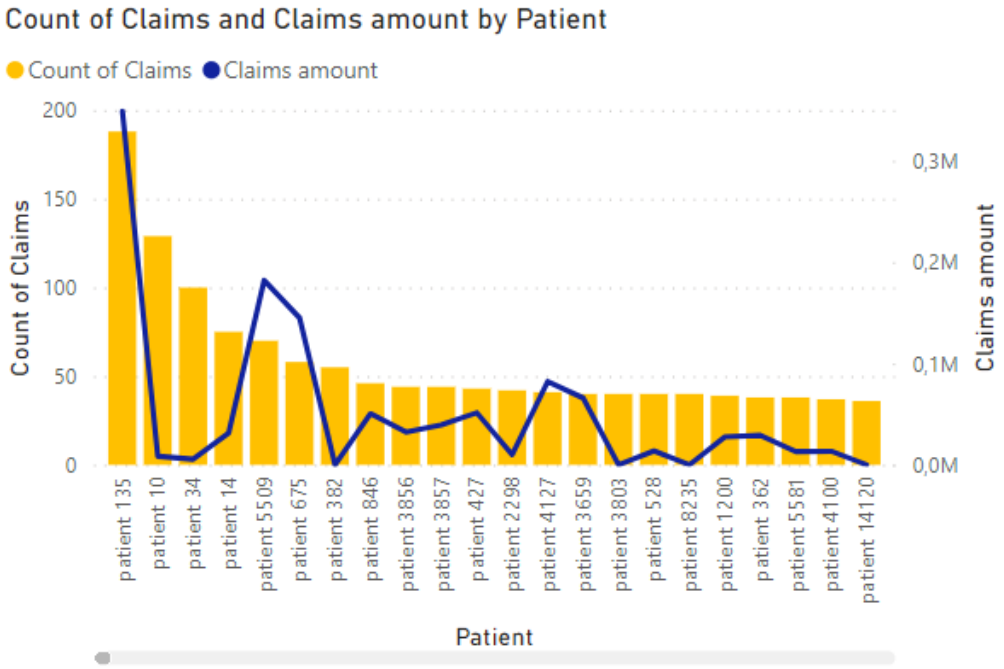
```sql
SELECT count(DISTINCT claim_id)
FROM dbo.data_claims c
JOIN dbo.data_provider p ON c.provider_name = p.provider_name
AND p.provider_country = 'Nigeria'
AND provider_star_rating = 5
WHERE claim_status = 'Approved'
```

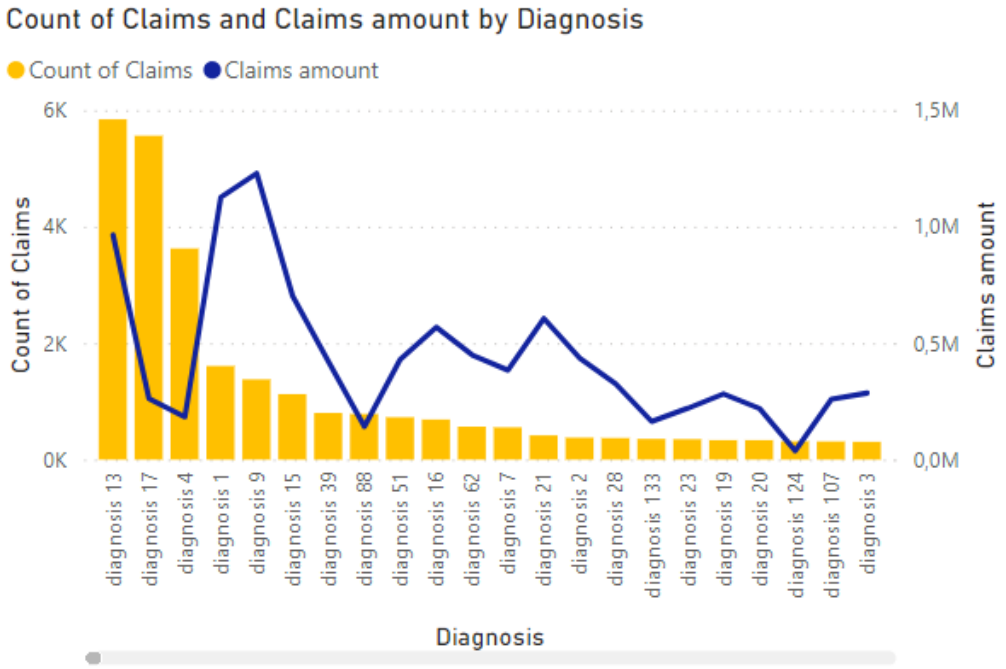
Task 2. Payer suggestions

While creating different visuals in Power BI, I noticed that the most expensive patient for the Benefit payer was patient 135, also patient 10 and patient 5509 were quite expensive as well.

Almost 200 treatment cases per one year looks quite suspicious, so I would recommend to check this case if there is a fraud.

Also, I have noticed that most popular approved cases were related to some specific diagnoses: diagnosis 13, 17, 4, 1, 9.

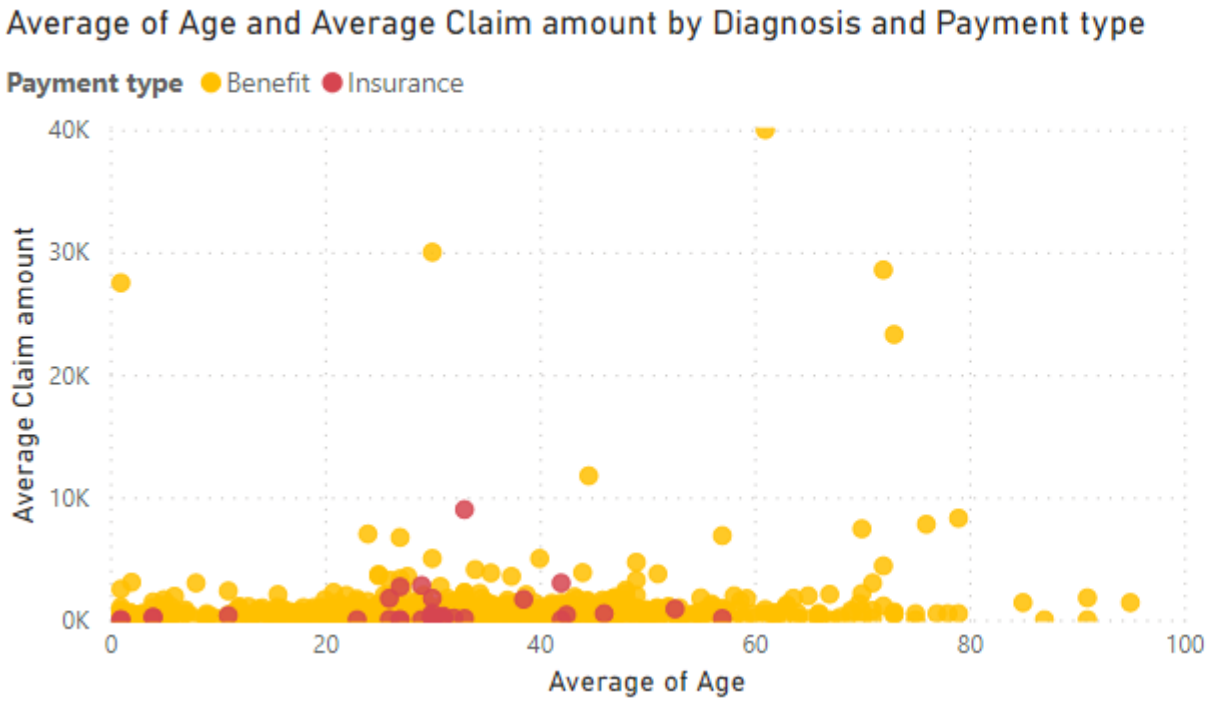


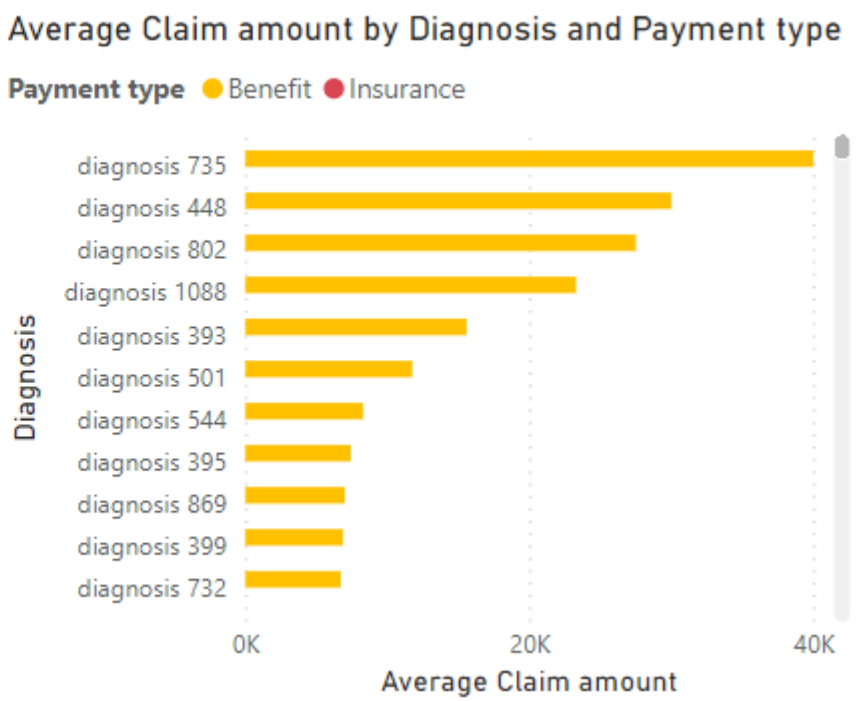


Task 2. Payer suggestions

However, there are also most expensive diagnoses (average claim > 20K): diagnosis 735, 448, 802, 1088.

Here I would also suggest to check on these cases, does it really take that much treatments and prescriptions what causes so much expenditure.

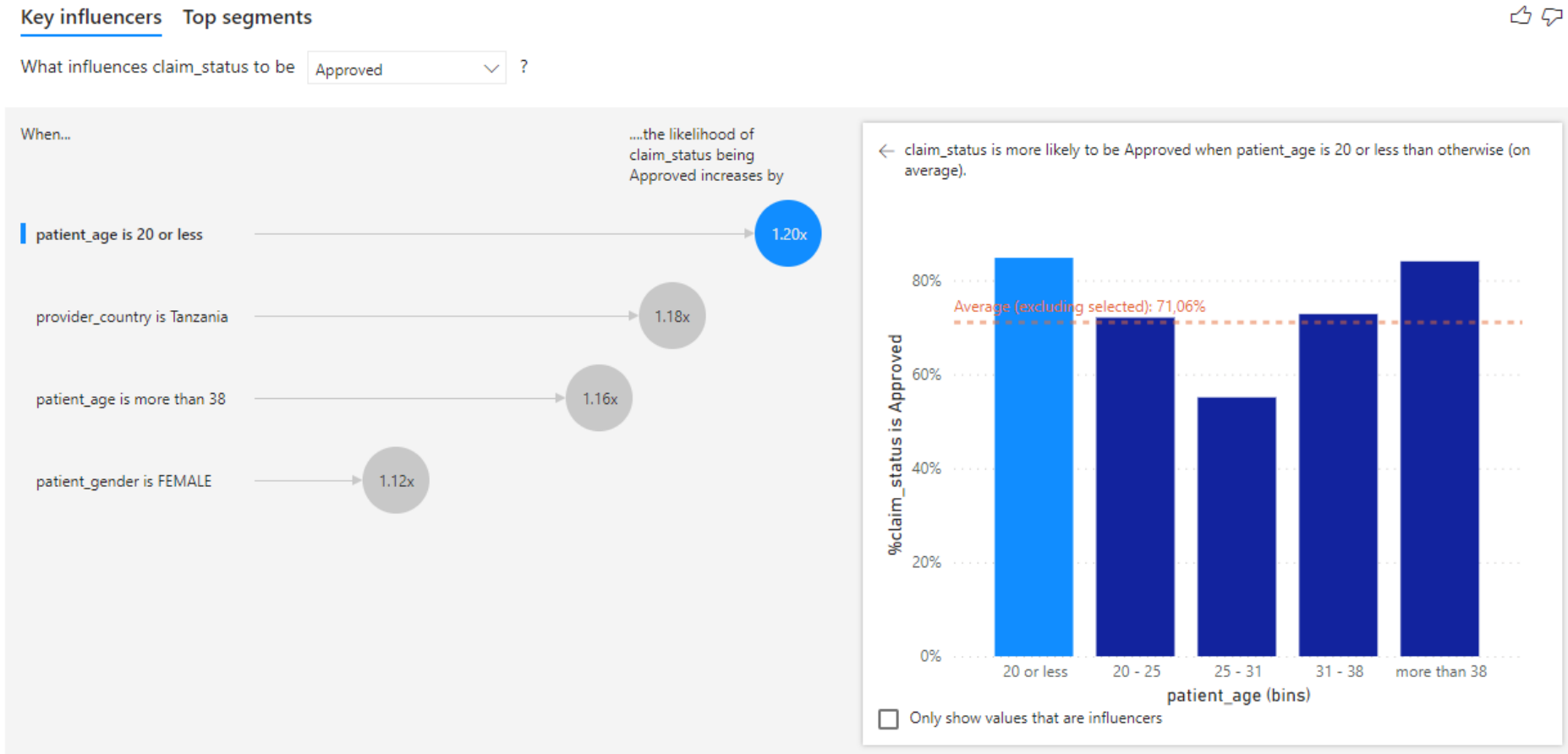




Task 3. CarePay suggestions

Correlations

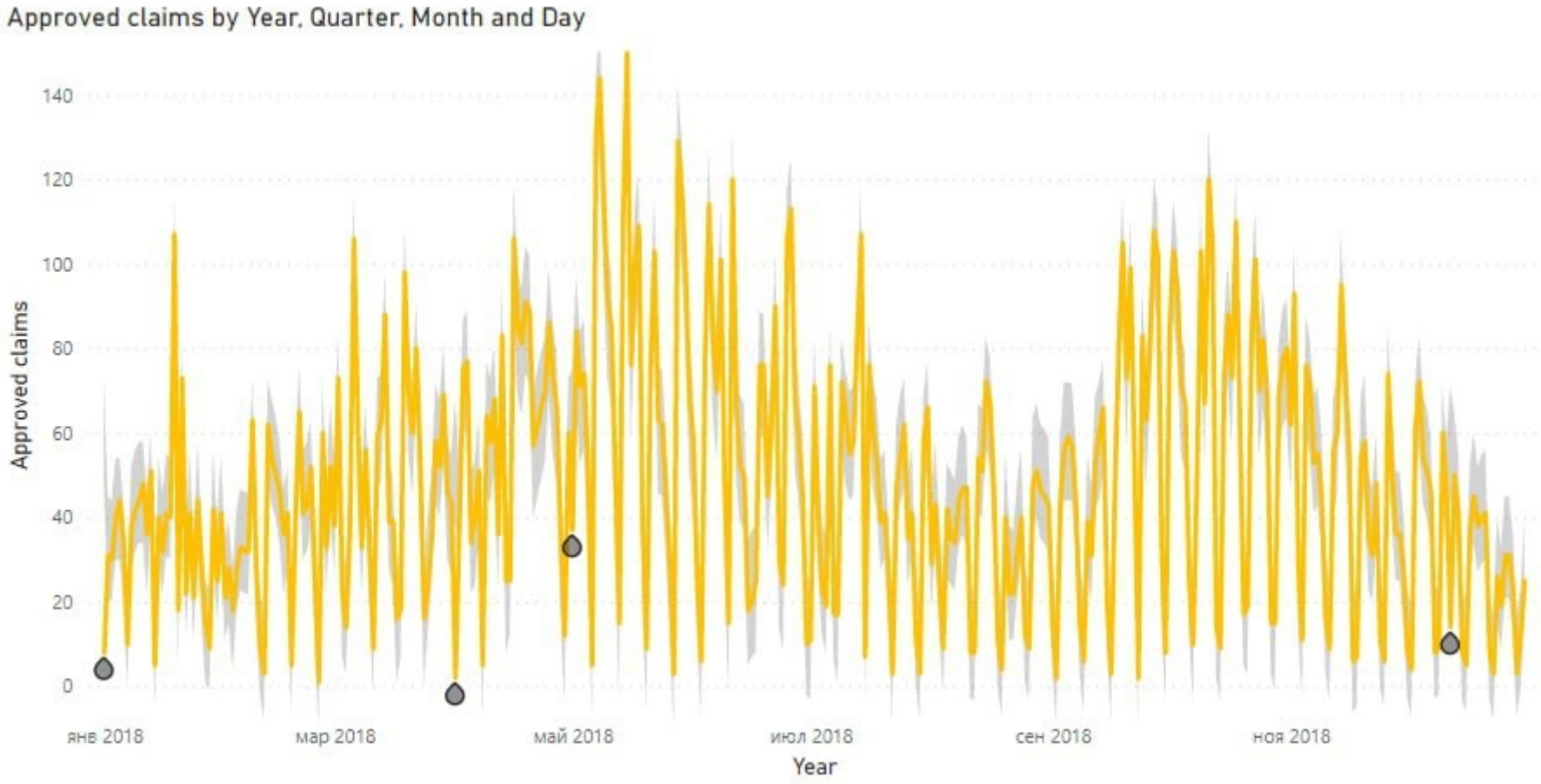
Below I have built a key influencers visual, which shows us what influences the claim to be approved. So, I found that the claims are mostly approved when the patient is female, younger than 20 or older than 38 and living in Tanzania.



Task 3. CarePay suggestions

Anomalies

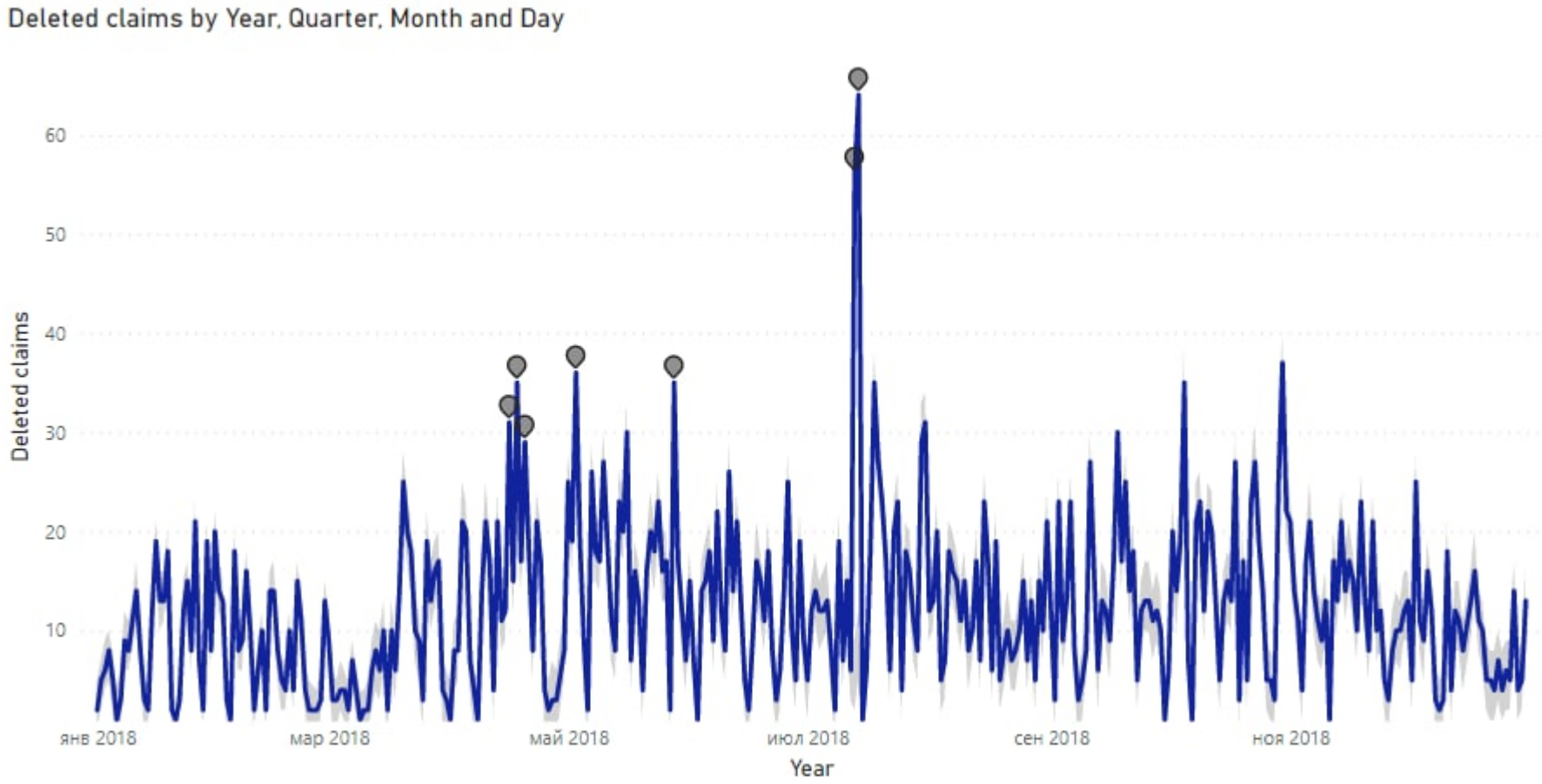
Here I found anomaly in claims count on January 1st, April 2nd, May 1st and December 12th, but after check I realized that it is related to public holidays.



Task 3. CarePay suggestions

Anomalies

On July 13th and 14th there were a huge count of deleted claims.
The reason is not determined, but probably it could relate to any application/server/network failure.



Task 3. CarePay suggestions

To improve the analysis process, I suggest the following:

1) Ensure the needed data is being collected and is correct, and that the data is ready and available for the analysis.
2) Define meaningless data to reduce waste time on looking there.
3) Track the system performance and outages to ensure that weak or incorrect data is related to technical issues, and not look for other reasons.
4) Bring more external data to the analysis by integrations: add public health data, medical data, pricing data, etc.

Task 4. Work plan

System improvements

1) Correct data: add the initial data checks to the system to avoid input errors and typos in the data.

2) More data:
   a) collect and analyze more data, such as a prescription categories;
   b) add more data to the analysis, such as public health data, comorbidities data, patient history.

3) Automatize the decision-making process, which could reduce human mistakes in the process and improve productivity:
   a) Collect a huge data pool with different diagnoses and treatment plans to check if the prescription was correct or any treatment was added intentionally or by mistake;
   b) Collect or integrate pricing references for the amount check, if the item is correct or too overpriced;
   c) Prepare an algorithm which could predict the decision based on points a) and b)

4) Use predictive analytics to help people pick their providers. This could help to generate for patient a list of clinics/doctors with the specialty they need.

Task 4. Work plan

Questions to claim manager

1) Describe the process of you reviewing the claim? How much time could it take?
2) How is the document management organized? Is there any digital documents storage/content management system?
3) What mostly makes you to reject the claim?
4) Are there any diagnoses references or pricing references you use for the decision?
5) What happens if the hospital is not agree with the decision? Can they then appeal this decision?