Exploratory Data Analysis Challenge

Task

In this exercise, you are given a dataset with claims data. You need to perform an Exploratory Data Analysis and present your results to business users (e.g. interactive dashboard, notebook, or some other tool of your choice).

Business users are interested to see trends and anomalies in the data as well as projections for the upcoming 6 months.

Note: Claims are expenses that insurance companies have to pay for medical services provided to patients.

Dataset

This dataset is a sampled aggregated data for the period of 2018/01 - 2020/07 (numbers are fictional).

The dataset contains the following columns:

- MONTH - a month claims were lodged
- SERVICE_CATEGORY - a department that provided services to patients
- CLAIM_SPECIALTY - a type of medical services by an official classification system
- PAYER - an insurance company
- PAID_AMOUNT - sum of expenses (claims), $

Requirements

- Python 3+
- Open-source libraries
- [Optional] Use Docker

Hints

- Kaggle EDA's as a reference
- Flask / FastAPI / plotly / bokeh / dash / etc.
- Perform any necessary data clean up
- For additional interactivity, you can use filtering data on the fly, pivoting data, etc.

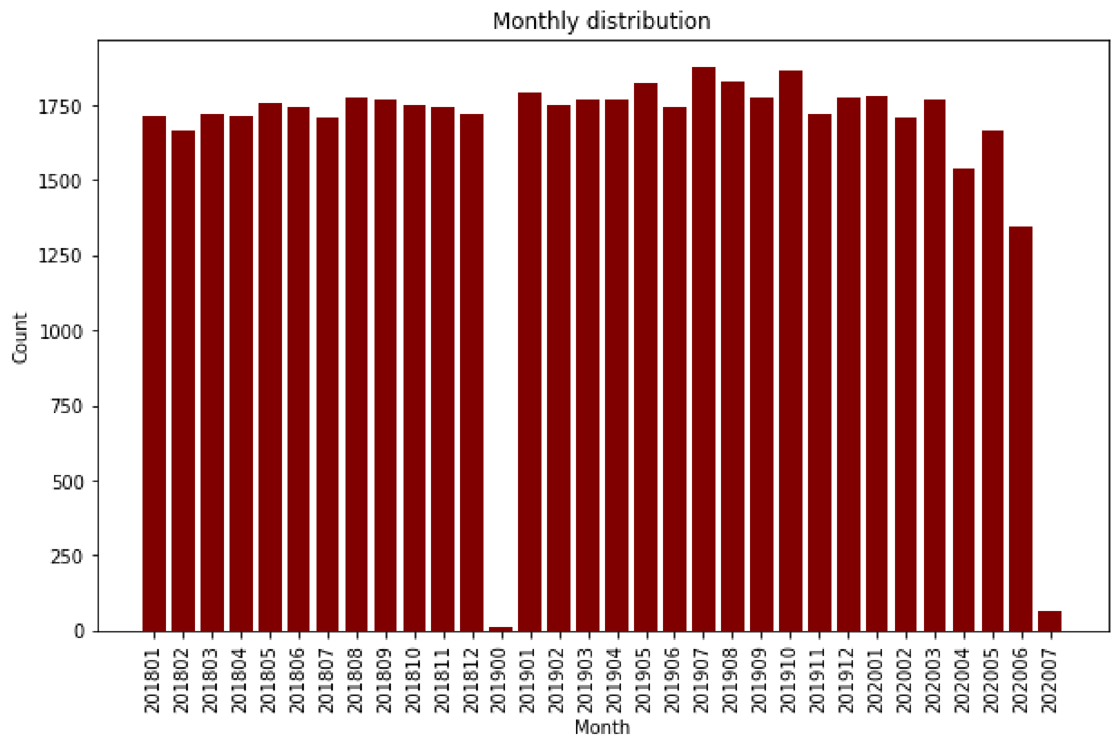
How to submit

Please upload the code for this project to GitHub, post a link to your repository below and give instructions on how to set up a local environment and run your code.

General properties of the dataset

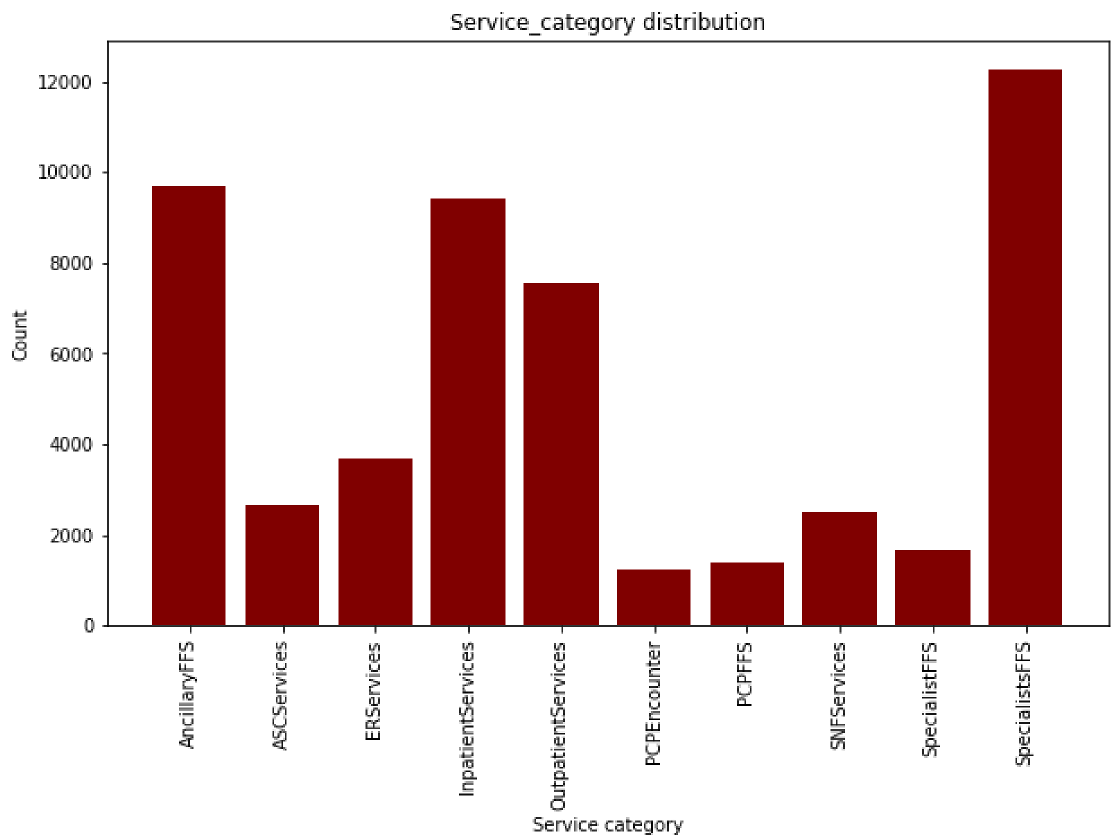
```
Number of rows:  52152

Minimum date:  201801

Maximum date:  202007
```

Counting the number of payments by month



Warining! We can see the strange month '201900'. That could be some mistake of writing the data. And we have small amount of payments on '202007'. It seems the month was not ended. We have too small number of payments on these two months so we can just remove them.

Let's compute all different Service categories

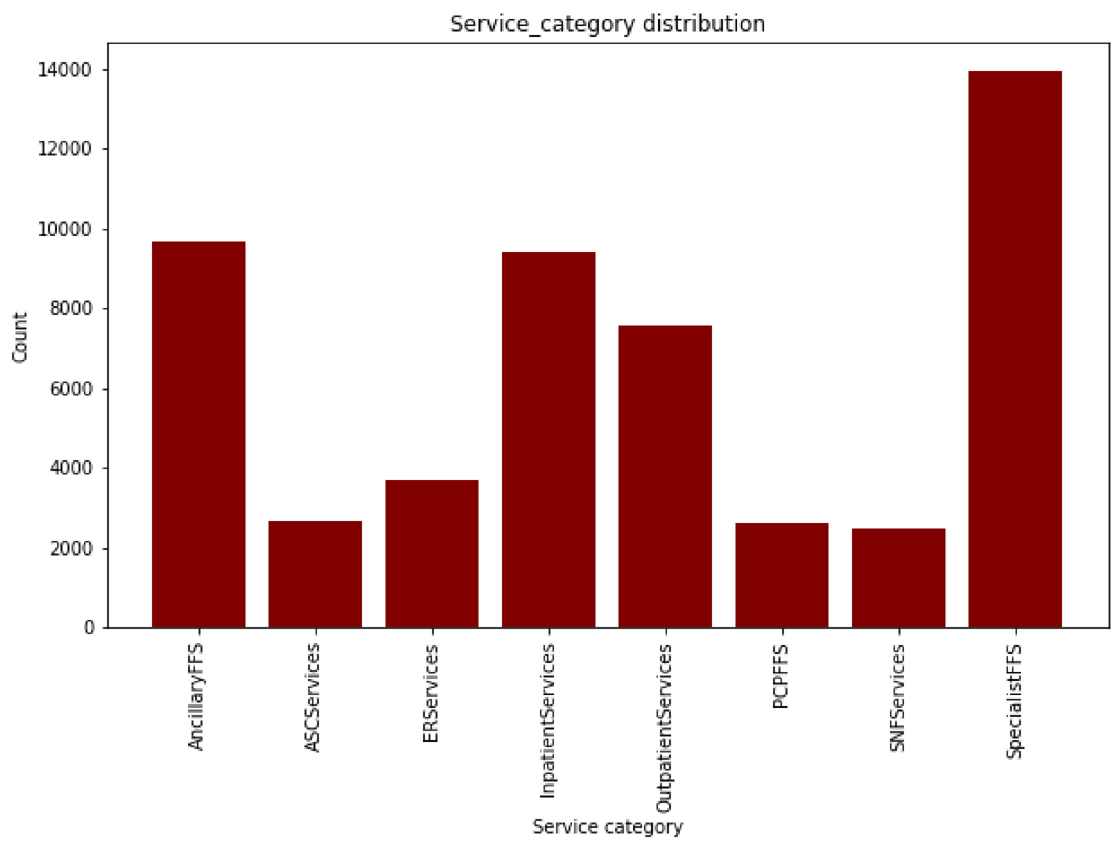


Description (from Internet)

- Fee-for-service (FFS) is a payment model in which doctors, hospitals, and medical practices charge separately for each service they perform. In this model, the patient or insurance company is responsible for paying whatever amount the healthcare provider charges for the service.
- Ambulatory Surgical Centers (ASC) Center | CMS
- Ancillary services are supportive or diagnostic measures that supplement and support a primary physician, nurse, or other healthcare provider in treating a patient. Some examples of ancillary services include: Imaging tests (e.g., X-rays, MRI, CT scan, ultrasound) Lab tests. Pharmacies.
- ER (emergency room) - This is probably because the letter E in the hospital setting typically refers to “emergency” (e.g., “ER” for emergency room and “ED” for emergency department).
- A primary care physician (PCP), or primary care provider, is a health care professional who practices general medicine. PCPs are our first stop for medical care. Most PCPs are doctors, but nurse practitioners and physician assistants can sometimes also be PCPs.

- What is an inpatient? In the most basic sense, this term refers to someone admitted to the hospital to stay overnight, whether briefly or for an extended period of time. Physicians keep these patients at the hospital to monitor them more closely.
- With this in mind, what is outpatient care? Also called ambulatory care, this term defines any service or treatment that doesn't require hospitalization. An annual exam with your primary care physician is an example of outpatient care, but so are emergent cases where the patient leaves the emergency department the same day they arrive. Any appointment at a clinic or specialty facility outside the hospital is considered outpatient care as well.
- Skilled Nursing Facility (SNF) is the full form of SNF medical abbreviation. It's a rehabilitation center where hospital patients are transferred after leaving the hospital. Seniors usually stay at an SNF for up to 100 days, and licensed medical professionals take care of them while they are there.

Grouping categories. Categories 'SpecialistFFS' and 'SpecialistFFS' seems to be the same category. 'PCPFFS' and 'PCPEncounter' are small categories but as they have the same prefix, we should group them also.



Let's check the data with zero and negative payments.

```
        MONTH SERVICE_CATEGORY      CLAIM_SPECIALTY     PAYER  PAID_AMOUNT
26      201801     AncillaryFFS          Cardiology  Payer UN         -250
46      201801     AncillaryFFS         Dermatology  Payer UN         -154
102     201801     AncillaryFFS   Geriatric Medicine  Payer UN         -201
103     201801     AncillaryFFS          GERIATRICS   Payer B         -203
144     201801     AncillaryFFS    Internal Medicine  Payer UN         -530
...        ...              ...                  ...       ...          ...
49357   202005     AncillaryFFS  Physician Assistant  Payer UN         -196
49414   202005     AncillaryFFS        Rheumatology  Payer UN        -1374
50158   202005           PCPFFS      FAMILY PRACTICE  Payer CA         -245
50165   202005           PCPFFS     GENERAL PRACTICE  Payer CA         -142
50172   202005           PCPFFS    INTERNAL MEDICINE  Payer CA        -3412

[318 rows x 5 columns]
```

Negative payments could be some kind of refund by court order.

```
Negative payments proportion 0.006106694319622076

Zero payments proportion 0.07385643507316511

Positive payments proportion 0.9200368706072128
```

Preprocessing of Claim Speciality

```
Proportion of nans in Claim speciality 0.004820063755424972

[nan 'ACH' 'Advanced Registered Nurse Prac' 'ADVANCED RN PRACT'
 'AMBULANCE' 'Ambulance' 'AMBULANCE SERVICE' 'AMBULATORY SURGICAL CENTER'
 'Ambulatory Transportation Services' 'ANATOM' 'ANESTHESIOLOGY' 'ARNP'
 'BEHAVIORAL HEALTH COUNSELING' 'CARD ELECTROPHYSIOLO'
 'CARDIAC ELECTROPHYSIOLOGY' 'CARDIOLOGY' 'Cardiology'
 'Cardiology/Cardiovascular Disease' 'Cardiovascular Medicine'
 'CARDIOVASCULAR SURGERY' 'CCS' 'Chiropractic Medicine'
 'Clinical Medical Laboratory\t' 'Colon and Rectal Surgery'
 'COMMUNITY MENTAL HEALTH CENTER'
 'COMMUNITY MENTAL HEALTH CENTER/OTHER REHAB CENTER' 'CONV CARE CLINIC'
 'COVERING PHYSICIAN' 'CRITICAL ACCESS' 'CSW' 'DEFAULT' 'Dermatology'
 'DERMATOLOGY' 'DERMATOPATHOLOGY' 'Dermatopathology' 'DIABETES EDUCATOR'
 'DIAG X-RAY CLINIC' 'DIAGNOSTIC RADIOLOGY' 'Diagnostics' 'DIAL'
 'DIALYSIS' 'Dialysis' 'DIALYSIS CENTER' 'DISEASE MANAGEMENT'
 'DME & Medical Supplies, Oxygen Equip & Supplies' 'Doctor of Psychology'
 'DURABLE MED EQUIPMENT' 'DURABLE MEDICAL EQUIPMENT'
 'Durable Medical Equipment' 'Durable Medical Supply']
...
```

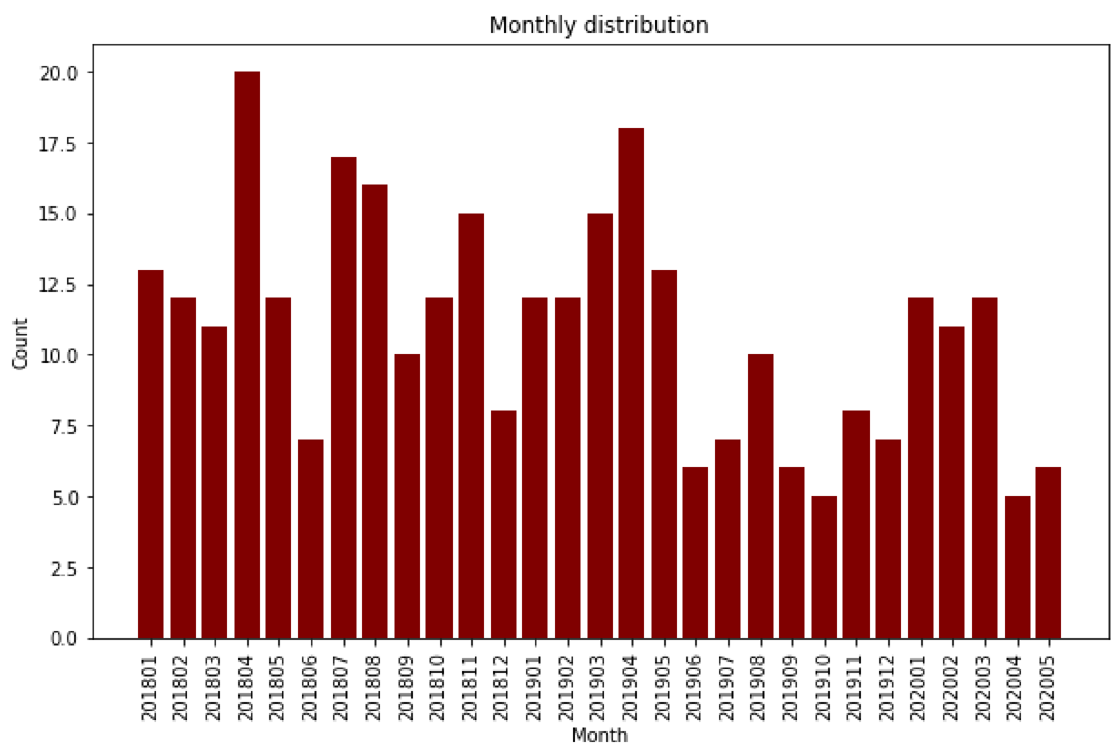
As we can see different records have different format and contain misprints and abbreviations so we should preprocess the field and indicate popular groups by hand

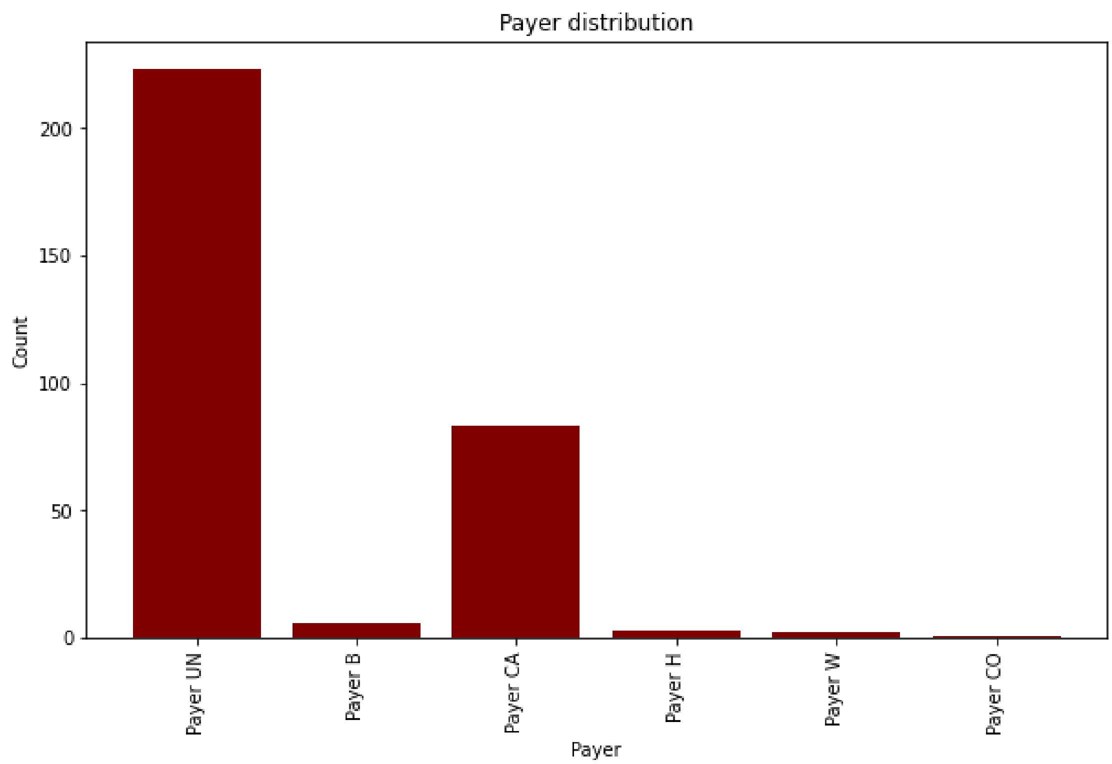
```
Preprocessed list:

['abulatory surgical center', 'accup', 'ach', 'acupu', 'acupuncturist',
'acute care hospital', 'acute short term hospital', 'addmed', 'addpsy',
'adlmed', 'adolescent medicine', 'advanced heart failure and transplant
cardiology', 'advanced registered nurse prac', 'advanced registered nurse
practitioner', 'advanced rn pract', 'agencies', 'ahftc', 'allergy', 'allergy
immunology', 'ambulance', 'ambulance emergency land', 'ambulance land',
'ambulance service', 'ambulance service provider', 'ambulatory health care',
'ambulatory surgery center', 'ambulatory surgical center', 'ambulatory
surgical centers', 'ambulatory surgical facility', 'ambulatory transportation
services', 'analytical labs', 'anast', 'anatom', 'anatomic and clinical
pathology', 'anatomic path clinical path', 'anatomic pathology', 'anes
assist', 'anesthesia', 'anesthesia tee', 'anesthesiologist',
'anesthesiologist assistant', 'anesthesiology', 'anesthesiology pain
medicine', 'anesthetist nurse', 'aprn', 'arnp', 'asc', 'audio',
'audiologist', 'audiology']
...
```

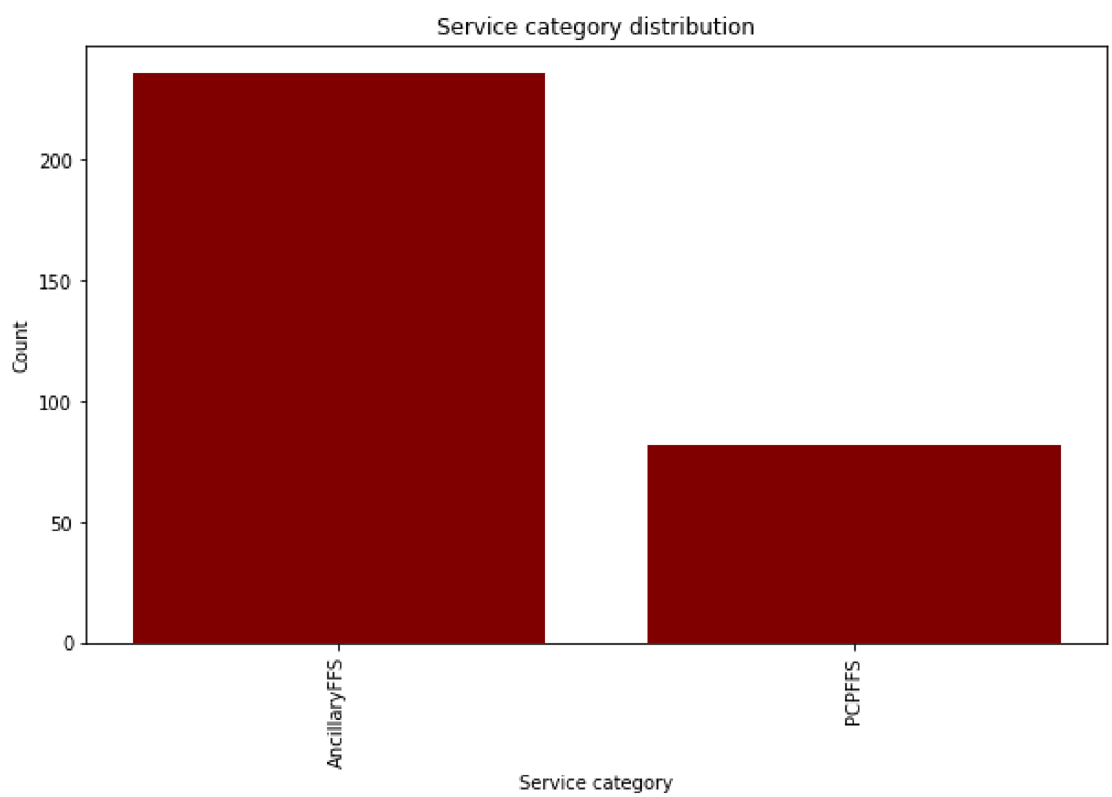
Making new features from grouped specialities

Exploring negative data

```
Number of rows: 318
```



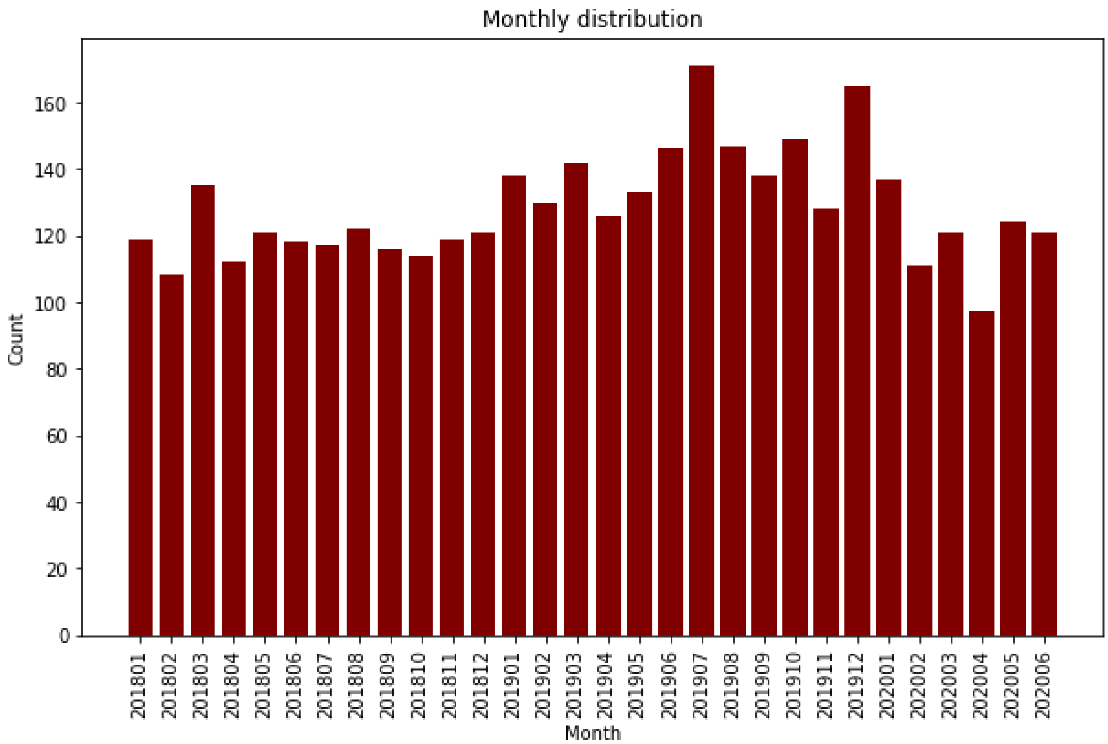


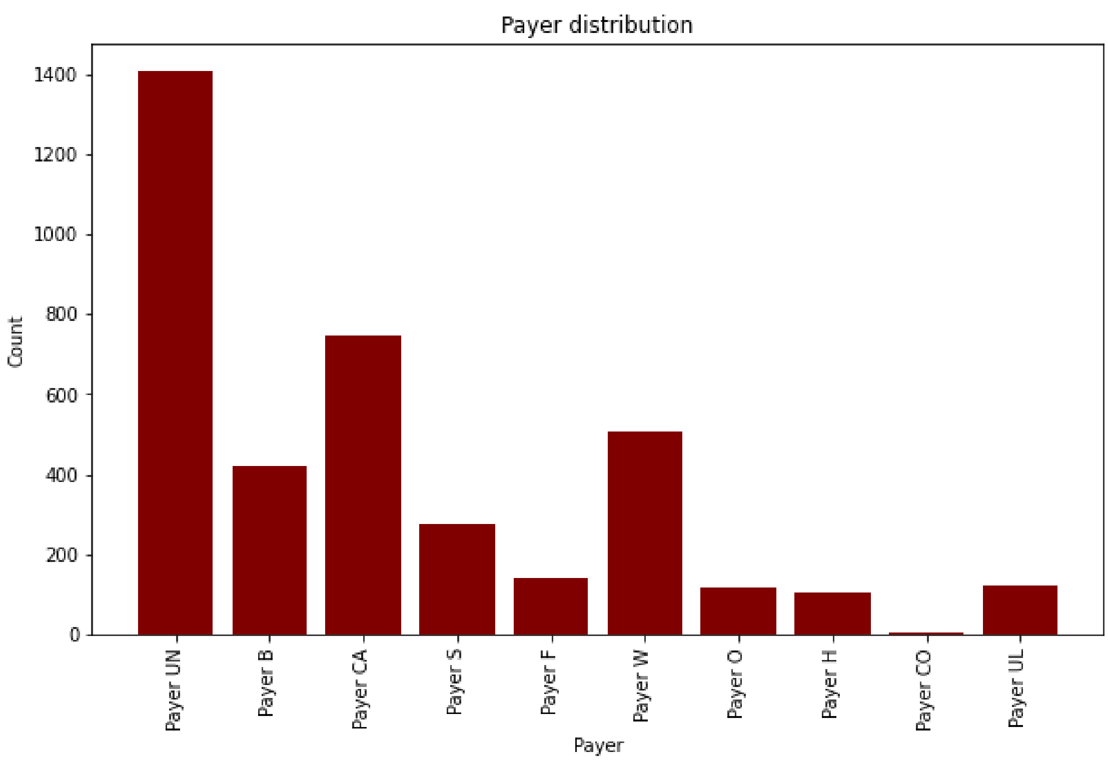


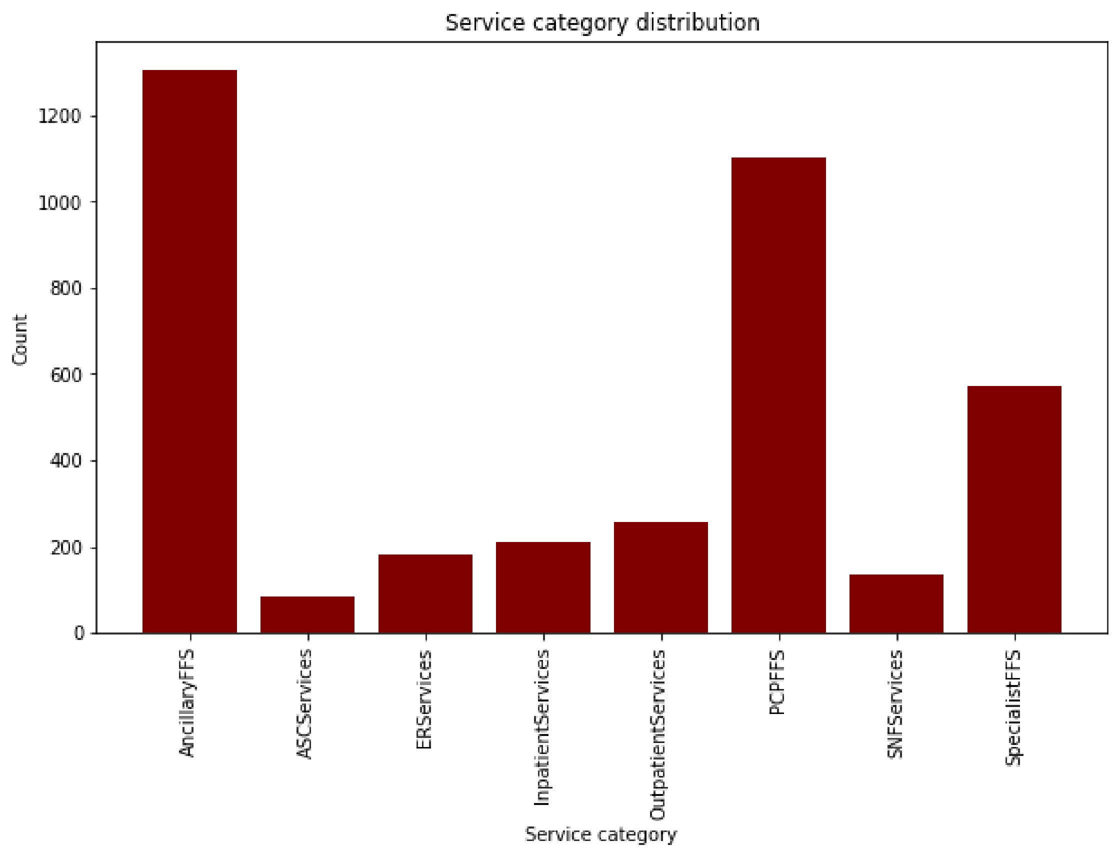
As we can see the most part of there payments belongs to 'Payer NU' and 'Payer CA' and two service categories - 'AncillaryFFS' and 'PCPFFS'

Exploring null data

```
Number of rows: 3846
```







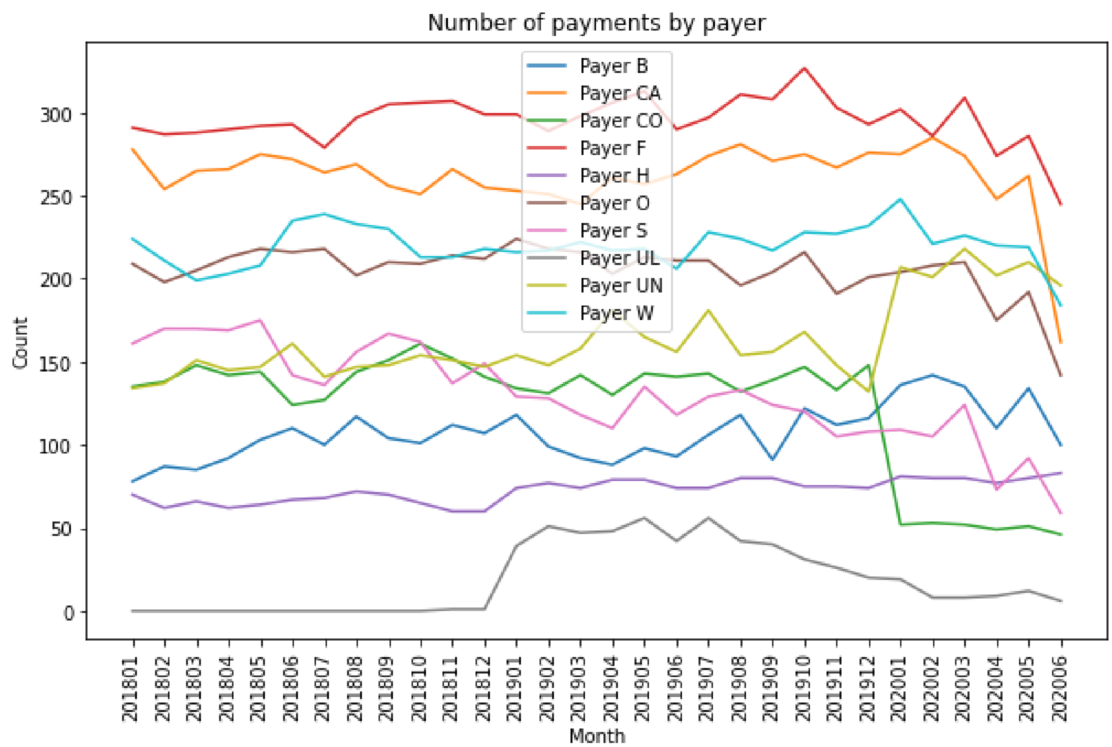
For zero payments one can see variety of companies and Service categories

Exploring positive data

This is the main part because positive payments if a majority of data (over 92%)

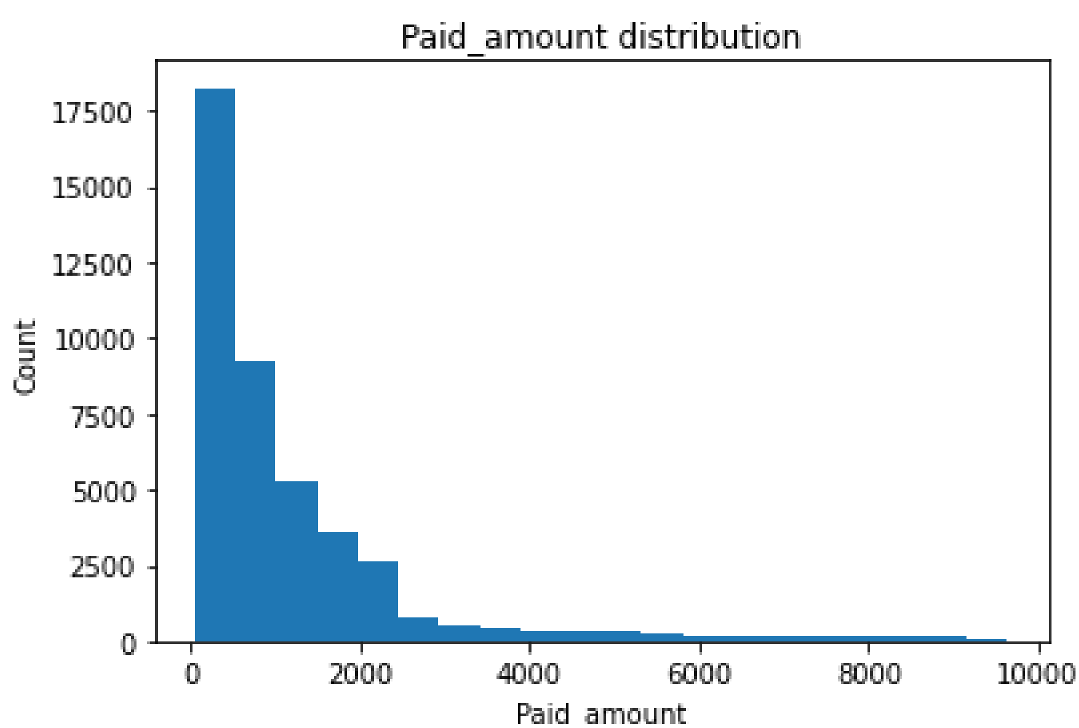
Pivot table for number of payments per company per month

	count									
MONTH	201801	201802	201803	201804	201805	201806	201807	201808	201809	...
PAYER										
Payer B	78.0	87.0	85.0	92.0	103.0	110.0	100.0	117.0	104.0	...
Payer CA	278.0	254.0	265.0	266.0	275.0	272.0	264.0	269.0	256.0	...
Payer CO	135.0	138.0	148.0	142.0	144.0	124.0	127.0	144.0	151.0	...
Payer F	291.0	287.0	288.0	290.0	292.0	293.0	279.0	297.0	305.0	...
Payer H	70.0	62.0	66.0	62.0	64.0	67.0	68.0	72.0	70.0	...
Payer O	209.0	198.0	205.0	213.0	218.0	216.0	218.0	202.0	210.0	...
Payer S	161.0	170.0	170.0	169.0	175.0	142.0	136.0	156.0	167.0	...
Payer UL	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
Payer UN	134.0	137.0	151.0	145.0	147.0	161.0	141.0	147.0	148.0	...
Payer W	224.0	211.0	199.0	203.0	208.0	235.0	239.0	233.0	230.0	...

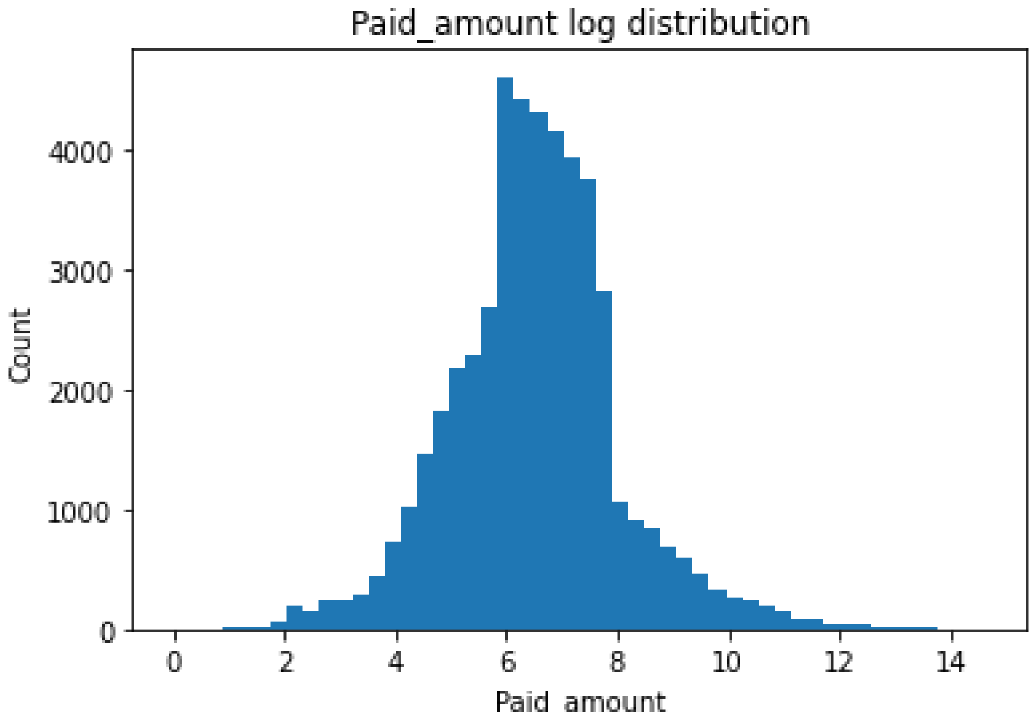


As one can see 'Payer UL' has less transactions than other payers and the first transaction appears on 201811

Paid_amount distribution



Probably paid_amount has log_normal distribution. Let's check it




```
Average of paid_amount: 4042.0415988311415
```

```
Standard deviation of paid_amount: 35971.74086297727
```

```
Standard deviation of paid_amount without outliers: 1464.6181974158987
```

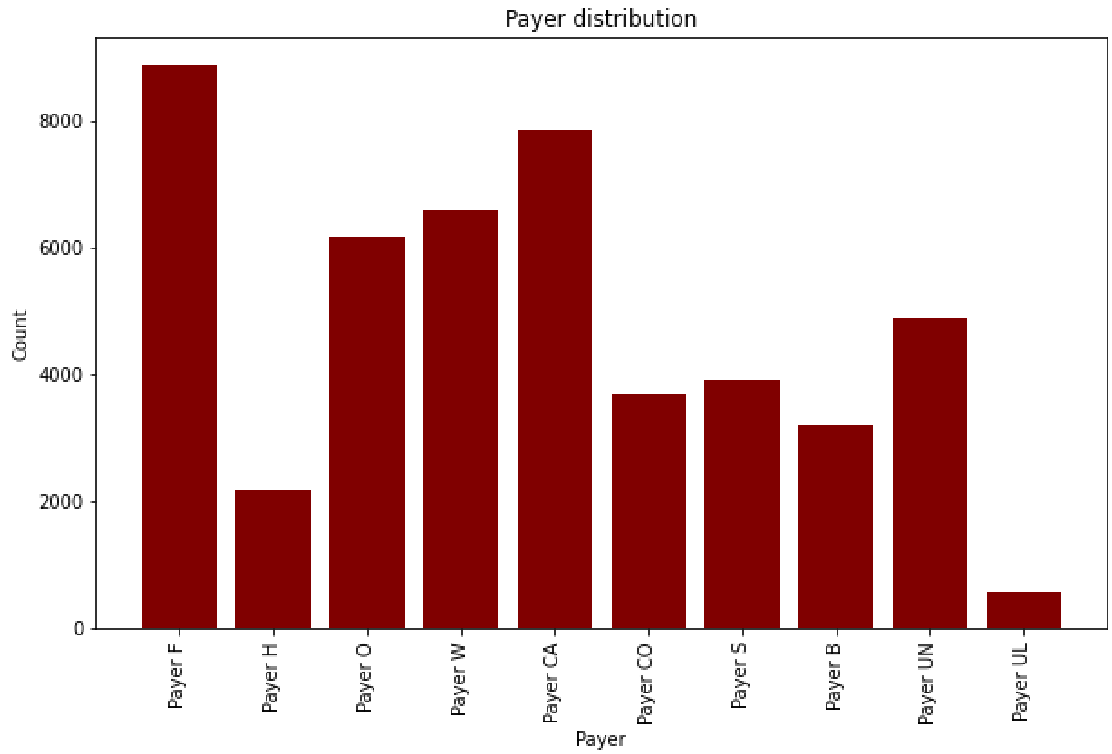
In the case when the distribution is log-normal, it's better to use median score instead of average because median is not too sensitive to the outliers

```
Median paid_amount: 676.0
```

The median is very close to the average of log-normal distribution

```
Log-normal average of paid_amount: 683.0967826328668
```

Payer distribution



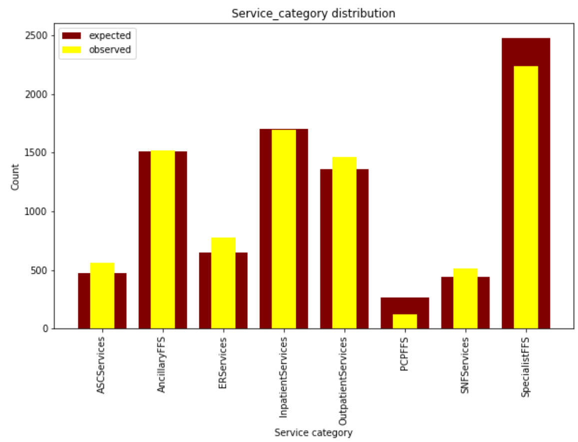
Computing how many different Service categories each payer has

```
Payer F : 8
Payer H : 8
Payer O : 8
Payer W : 8
Payer CA : 8
Payer CO : 8
Payer S : 8
Payer B : 8
Payer UN : 8
Payer UL : 8
```

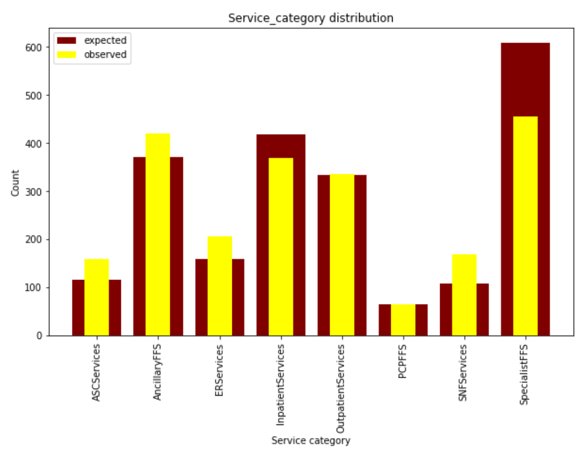
All 10 payers have all categories

GOODNESS OF FIT Test

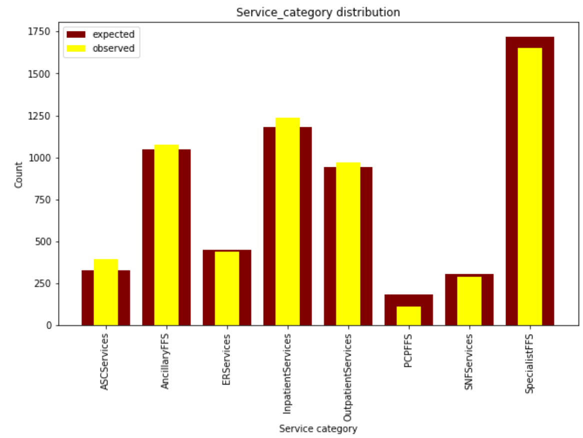
Comparing the distribution of service categories by payer to overall distribution



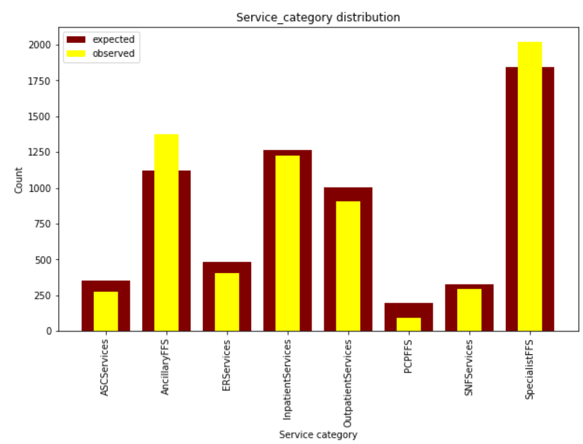

```
payer: Payer F
chi_square_test_statistic is :
159.11604550451807
p_value : 4.922792044589307e-31
12.591587243743977
```



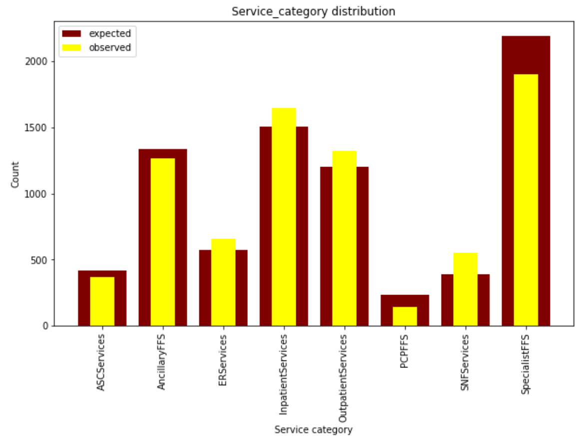

```
payer: Payer H
chi_square_test_statistic is :
115.792171352185
p_value : 5.753410270783393e-22
12.591587243743977
```



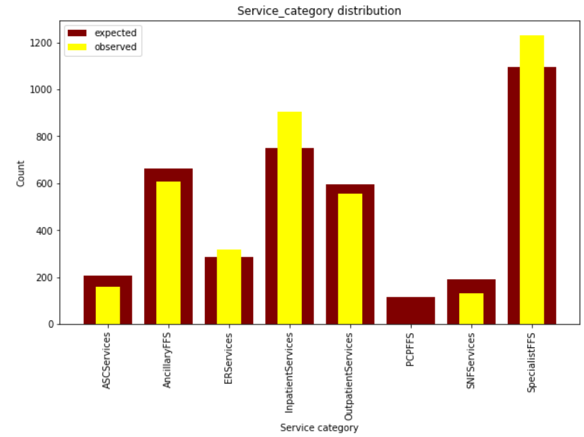

```
payer: Payer O
chi_square_test_statistic is :
52.34938375375981
p_value : 4.982251437196621e-09
12.591587243743977
```



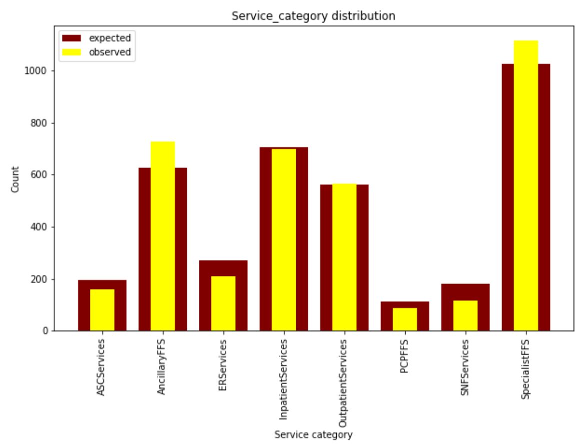

```
payer: Payer W
chi_square_test_statistic is :
176.0289455086634
p_value : 1.3426989397139578e-34
12.591587243743977
```



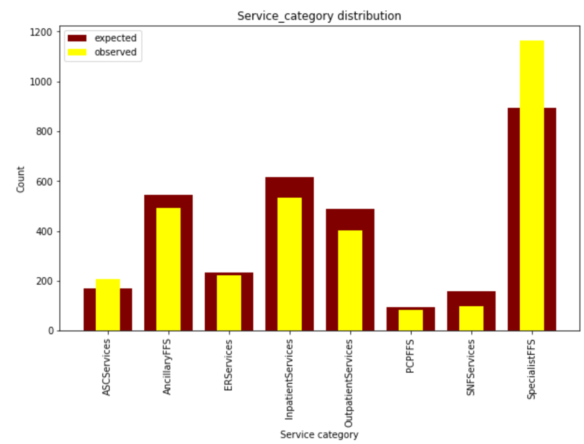
payer: Payer CA
chi_square_test_statistic is :
197.54232751920188
p_value : 3.804057411879403e-39
12.591587243743977



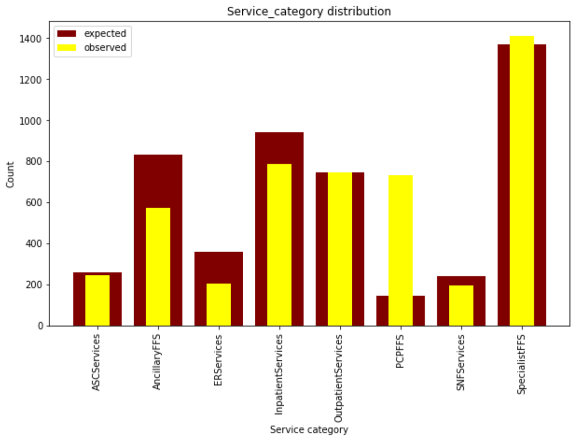
payer: Payer S
chi_square_test_statistic is :
205.22116354002327
p_value : 8.990849221299443e-41
12.591587243743977

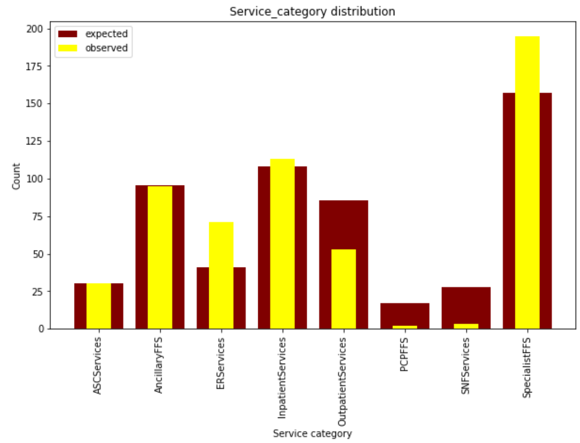


payer: Payer CO
chi_square_test_statistic is :
74.13116577018542
p_value : 2.1525009320621745e-13
12.591587243743977



payer: Payer B
chi_square_test_statistic is :
146.63526282809346
p_value : 2.0647085595041257e-28
12.591587243743977





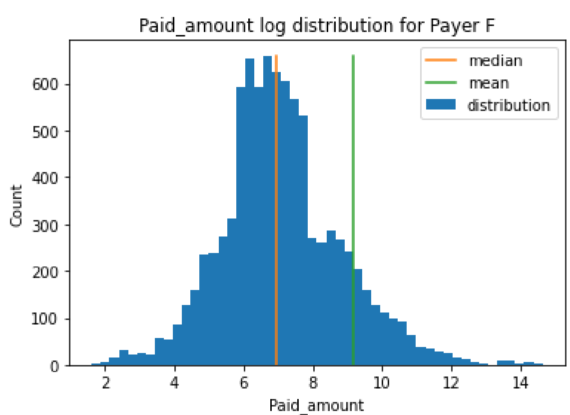
payer: Payer UN
chi_square_test_statistic is :
2508.030453893707
p_value : 0.0
12.591587243743977

payer: Payer UL
chi_square_test_statistic is :
78.7670210369333
p_value : 2.4569468480583376e-14
12.591587243743977

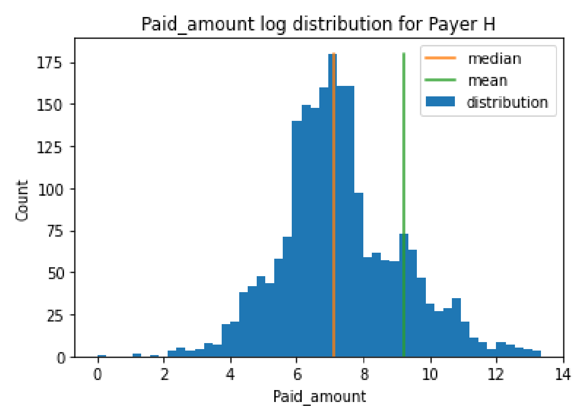
In each case we have significant difference but the most visible difference is for 'Payer UN' which has unproportional number of PCP payments

Payment distribution by payer

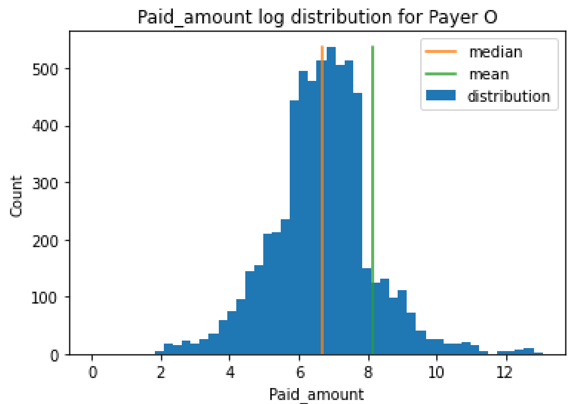
Payer: Payer F
Number of rows: 8870
Median paid_amount: 1035.0
Mean paid_amount: 9782.131792559188



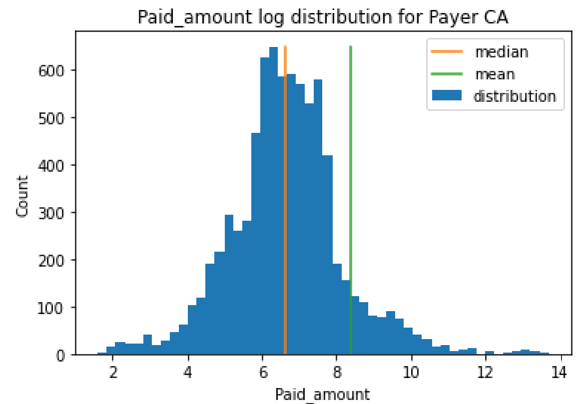
Payer: Payer H
Number of rows: 2182
Median paid_amount: 1245.5
Mean paid_amount: 9956.466086159488



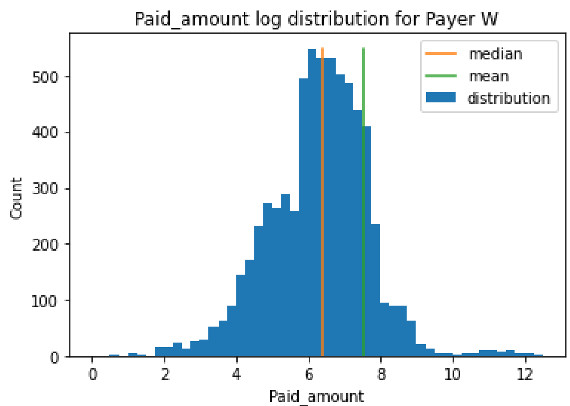
```
Payer: Payer O
Number of rows: 6159
Median paid_amount: 810.0
Mean paid_amount: 3477.723981165
```



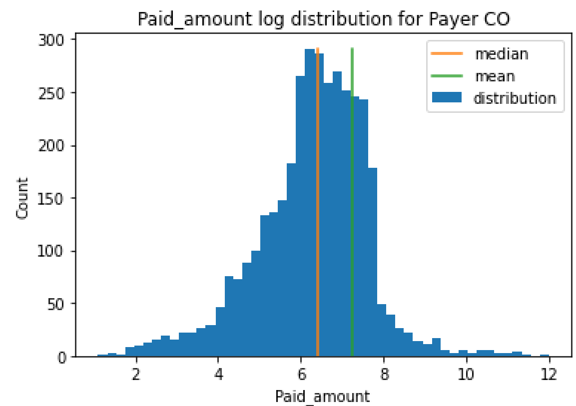

```
Payer: Payer CA
Number of rows: 7851
Median paid_amount: 757.0
Mean paid_amount: 4291.237804101388
```



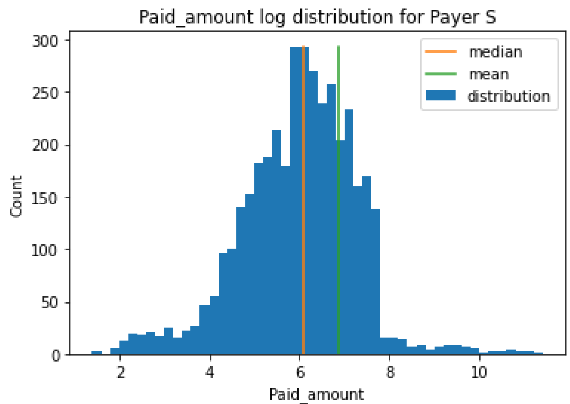

```
Payer: Payer W
Number of rows: 6596
Median paid_amount: 588.0
Mean paid_amount: 1846.734990903578
```



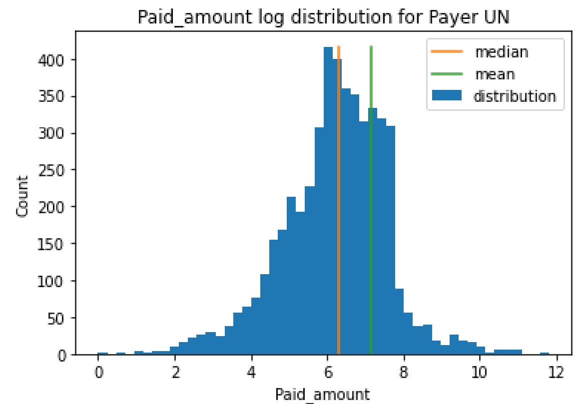

```
Payer: Payer CO
Number of rows: 3673
Median paid_amount: 607.0
Mean paid_amount: 1380.007078682276
```



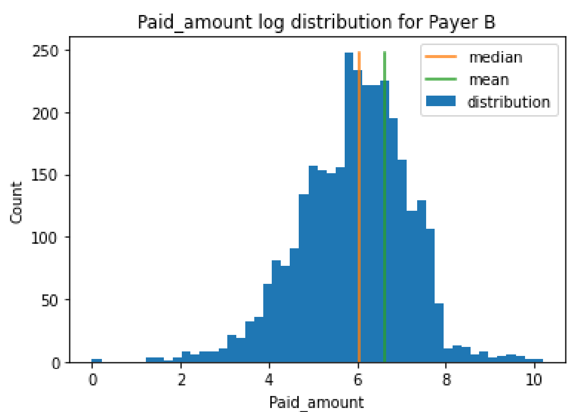
```
Payer: Payer S
Number of rows: 3913
Median paid_amount: 438.0
Mean paid_amount: 973.8686429849221
```



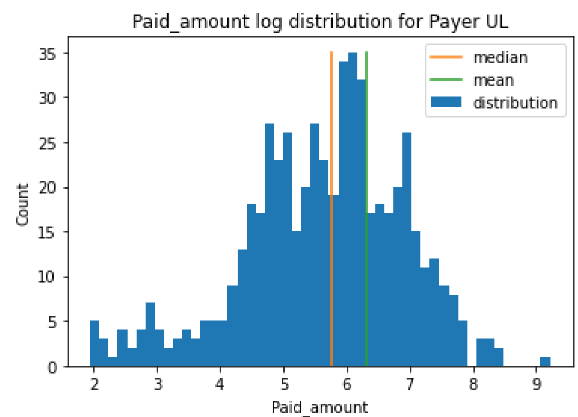

```
Payer: Payer UN
Number of rows: 4898
Median paid_amount: 544.0
Mean paid_amount: 1280.56472029399
```



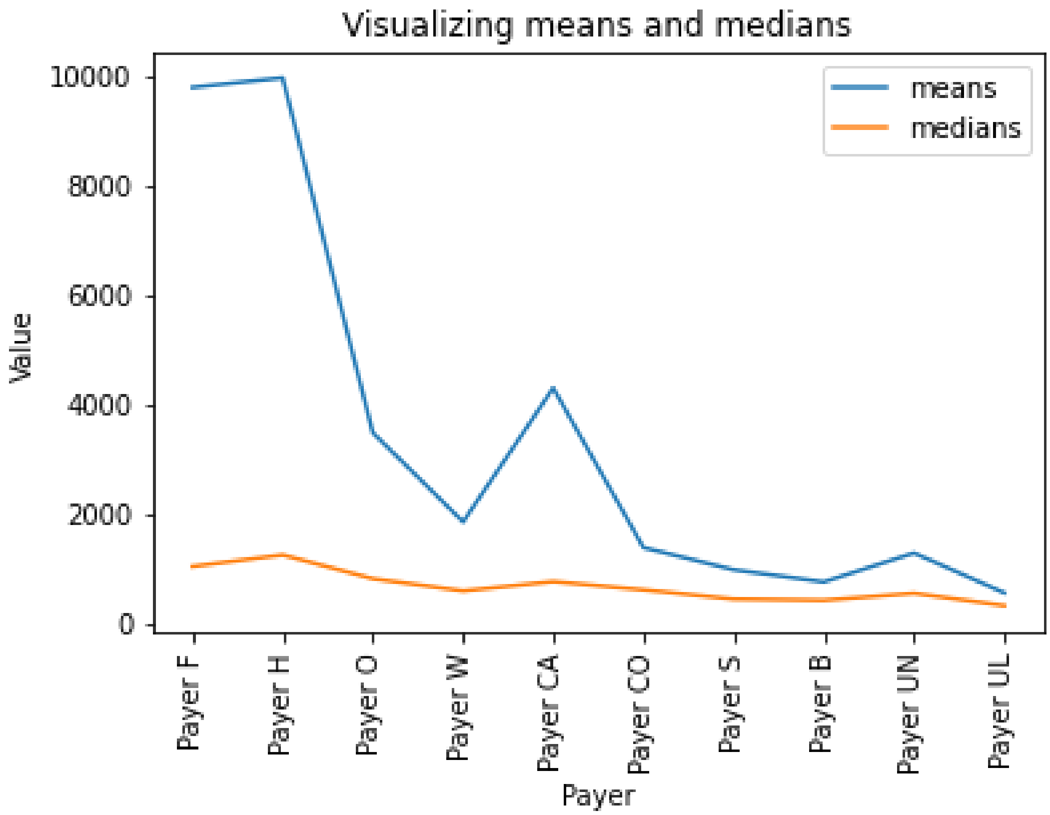

```
Payer: Payer B
Number of rows: 3206
Median paid_amount: 416.0
Mean paid_amount: 751.426699937617
```




```
Payer: Payer UL
Number of rows: 562
Median paid_amount: 318.5
Mean paid_amount: 551.6494661921708
```



Showing that median is more stable statistic



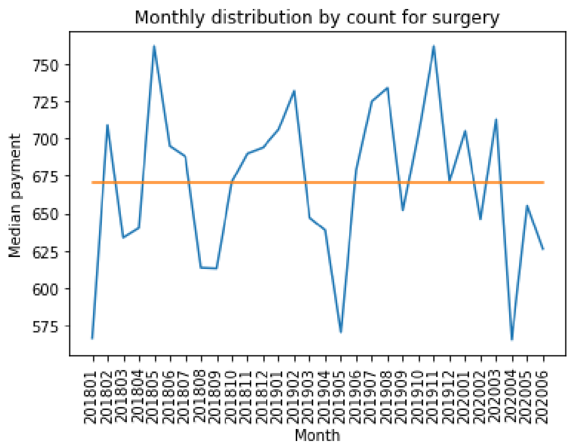

```
Correlarion between means and medians: 0.9510783412495436
P-value for correlation coefficient: 2.3618698414366877e-05
```

Exploring top-specialities

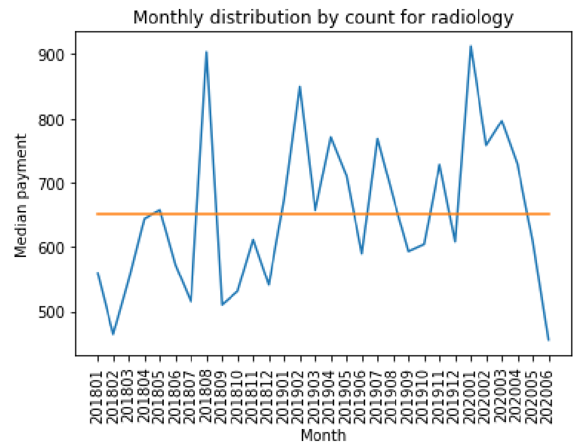
Top categories by count

	Speciality	Count
0	surgery	6047
1	otolaryngology	5490
2	radiology	3701
3	cardiology	3119
4	oncology	2164
5	general	2150
6	physical	1954
7	nurse	1773
8	internal	1739
9	urology	1711

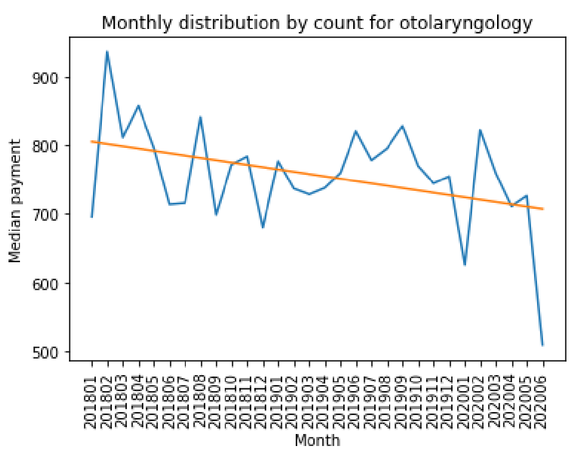
surgery -->
corr_coef: 0.032931277007990596 -->
p-value: 0.8628445382422639



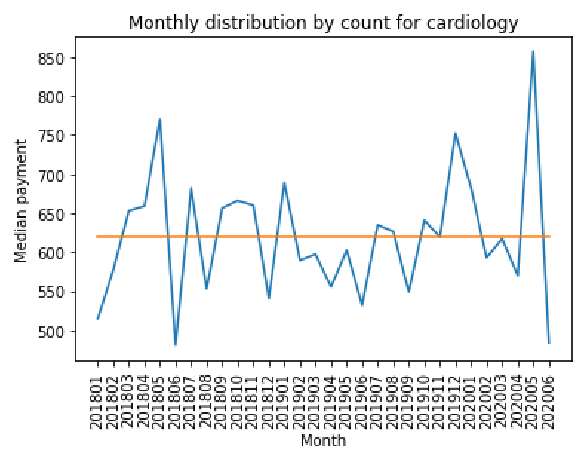
radiology -->
corr_coef: 0.319460027017729 -->
p-value: 0.085282512162



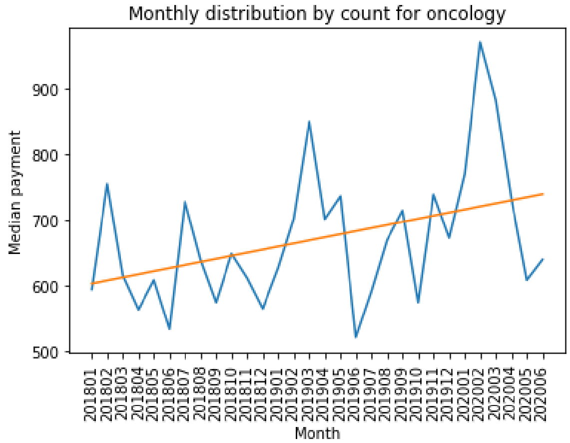
otolaryngology -->
corr_coef: -0.3872070810381735 -->
p-value: 0.03451882284



cardiology -->
corr_coef: 0.09391273122173291 -->
p-value: 0.6215762120784231



```
oncology -->
corr_coef: 0.3956229330988751 -->
p-value: 0.030463878999197228
```



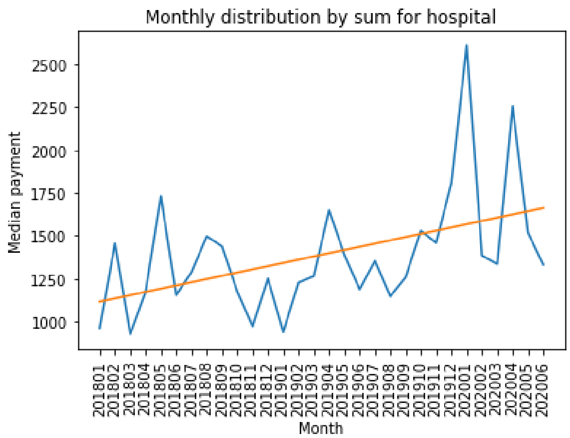
For the top specialities by count we can see the significant (0.05 threshold) trend for:

- otolaryngology (correlation: -0.39, p_value: 0.03)
- oncology (correlation: 0.39, p_value: 0.03)

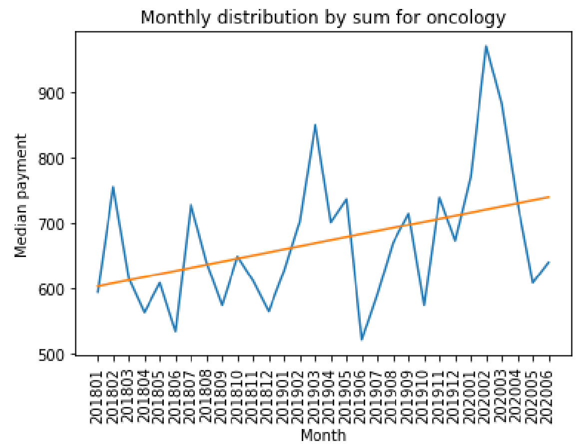
Top categories by paid sum

	Speciality	Sum_paid
0	hospital	41863179
1	otolaryngology	10368734
2	oncology	9718748
3	surgery	9039850
4	radiology	8190466
5	cardiology	7504275
6	hematology	6291089
7	internal	4966067
8	ambulance service	3664576
9	nurse	3554506

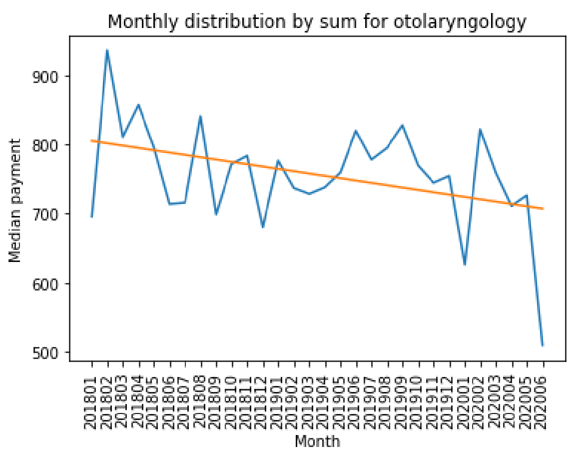
```
hospital -->
corr_coef: 0.45885957745161543 -->
p-value: 0.010757221768815231
```



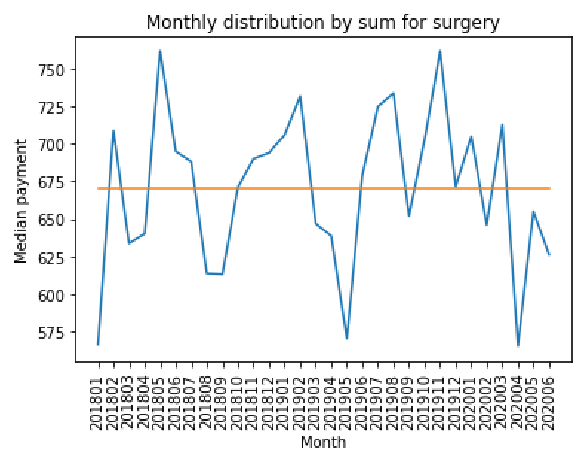

```
oncology -->
corr_coef: 0.3956229330988751 -->
p-value: 0.030463878999197228
```



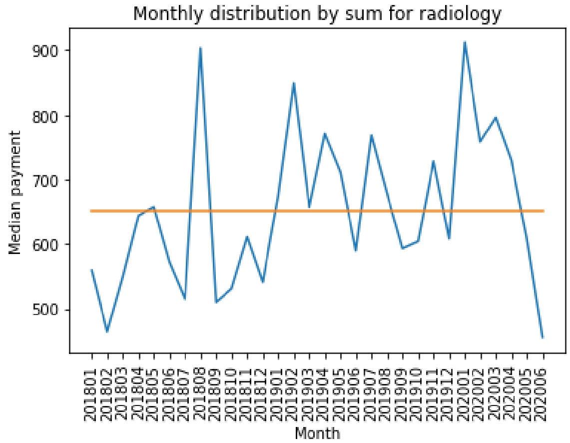

```
otolaryngology -->
corr_coef: -0.3872070810381735 -->
p-value: 0.03451882284
```




```
surgery -->
corr_coef: 0.032931277007990596 -->
p-value: 0.8628445382422639
```



```
radiology -->
corr_coef: 0.319460027017729 -->
p-value: 0.08528251216296788
```



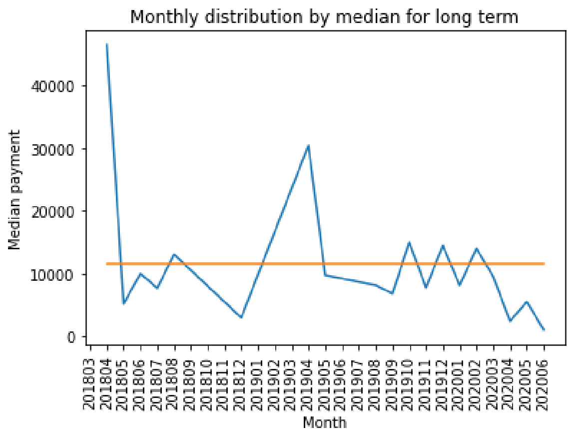
For the top specialities by sum we can see the significant (0.05 threshold) trend for:

- hospital (correlation: 0.46, p_value: 0.01)
- otolaryngology (correlation: -0.39, p_value: 0.03)
- oncology (correlation: 0.40, p_value: 0.03)

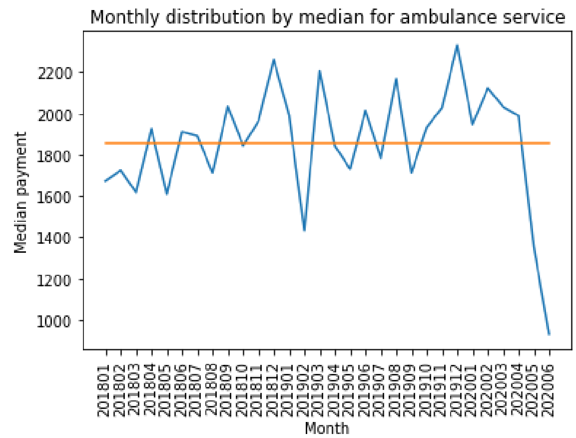
Top specialities by median

	Speciality	Median_payment
0	long term	8759.0
1	dialysis	6642.0
2	ambulance service	1846.0
3	home	1610.0
4	infusion	1345.5
5	hospital	1344.5
6	durable	1147.0
7	dental	1075.0
8	dermatology	951.0
9	gastroenterology	917.0

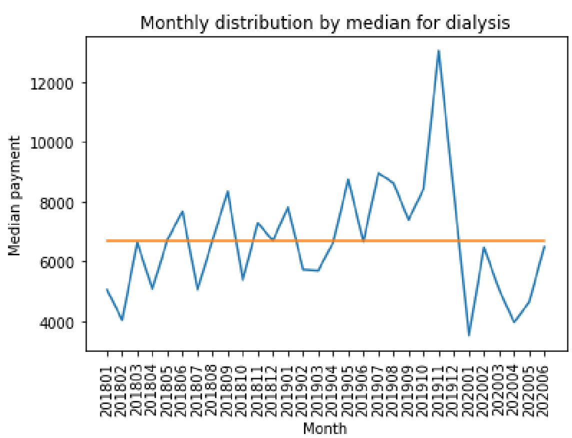
```
long term -->
corr_coef: -0.38852250243749 -->
p-value: 0.10019989880254503
```



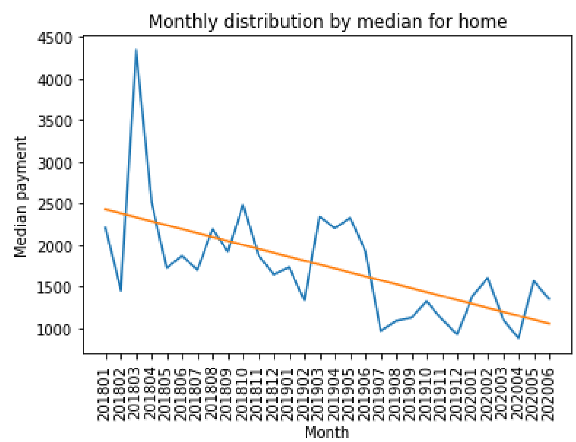

```
ambulance service -->
corr_coef: 0.00197792695692095 -->
p-value: 0.99172352
```



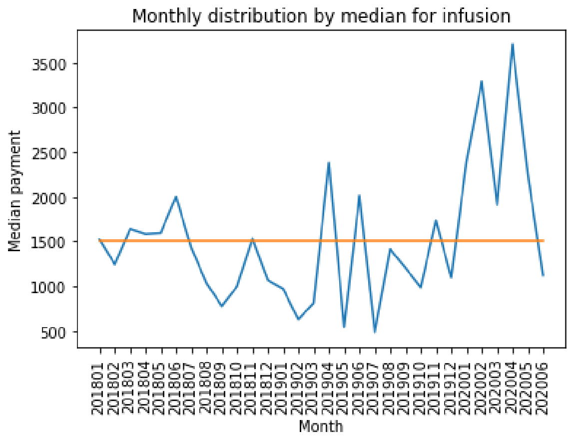

```
dialysis -->
corr_coef: 0.11341195049337842 -->
p-value: 0.5506954964399359
```




```
home -->
corr_coef: -0.6047007937506313 -->
p-value: 0.00040090201551917613
```



```
infusion -->
corr_coef: 0.33621129408152717 -->
p-value: 0.06928594951186588
```

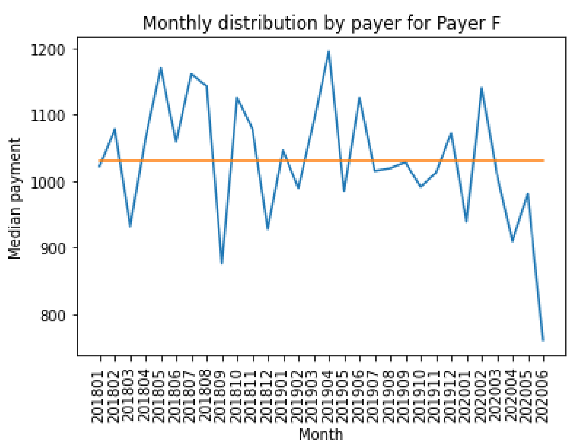


For the top specialities by median payment we can see the significant (0.05 threshold) trend for:

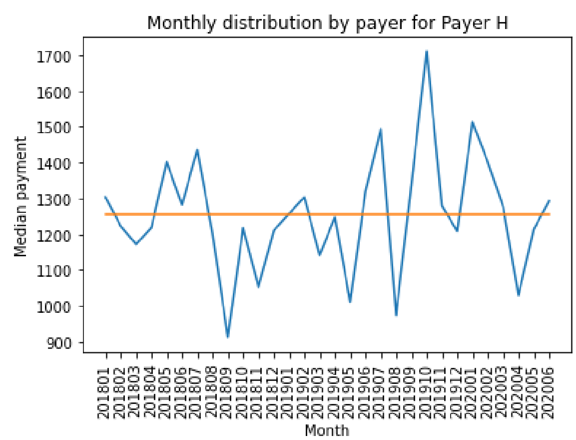
- home (correlation: -0.60, p_value: 0.0004)
- hospital (correlation: 0.46, p_value: 0.01)

Payment trends for payers

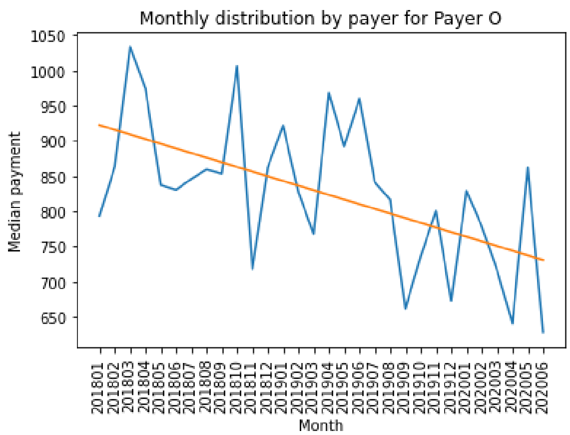
```
Payer F -->
corr_coef: -0.354724184895404086 -->
p-value: 0.05442710261042485
```



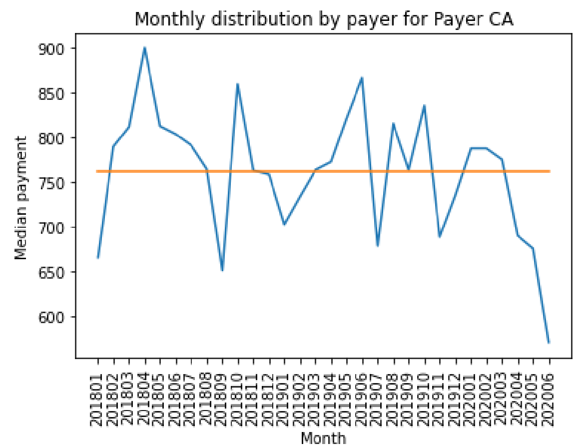

```
Payer H -->
corr_coef: 0.1342354948437787 -->
p-value: 0.47943758557549215
```



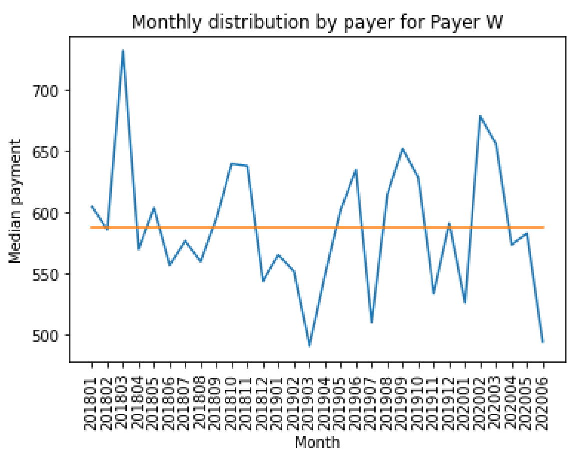
```
Payer O -->
corr_coef: -0.5537476680881521 -->
p-value: 0.0015006959768256332
```



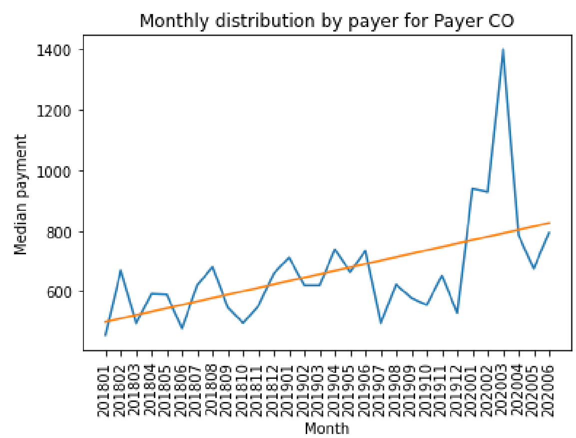

```
Payer CA -->
corr_coef: -0.33589995639861936 -->
p-value: 0.06956029961543066
```



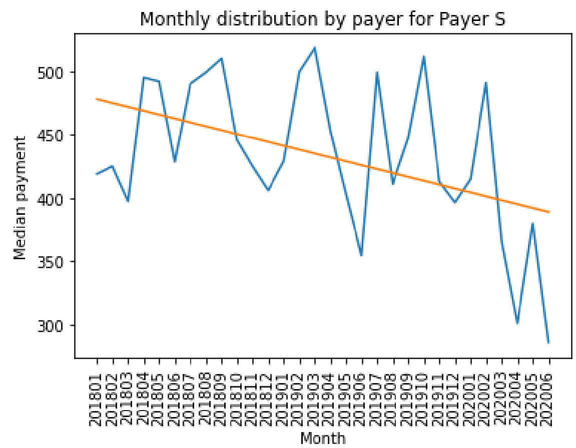

```
Payer W -->
corr_coef: -0.1340543304851863 -->
p-value: 0.4800362644441857
```



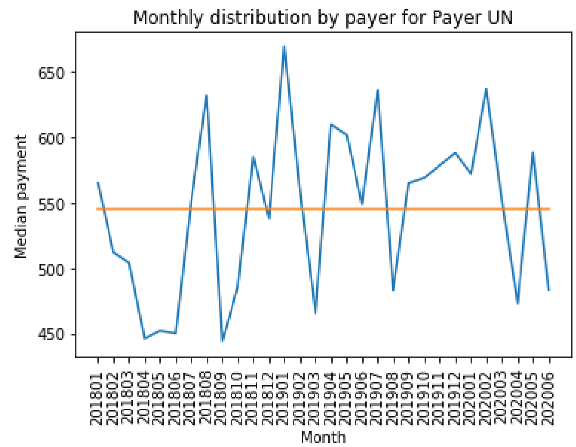

```
Payer CO -->
corr_coef: 0.5394459004591824 -->
p-value: 0.002095670825129852
```



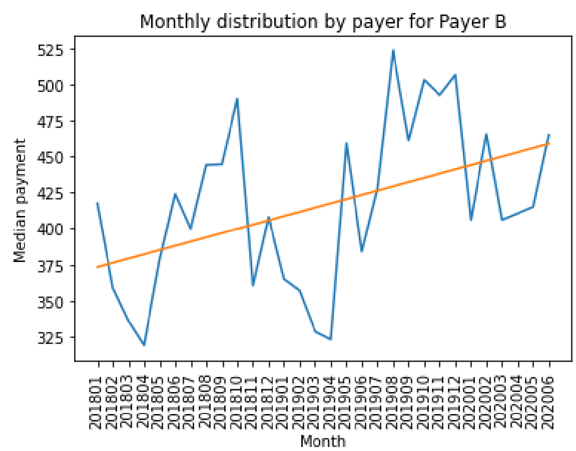
```
Payer S -->
corr_coef: -0.4474623809781648 -->
p-value: 0.013162094617099218
```



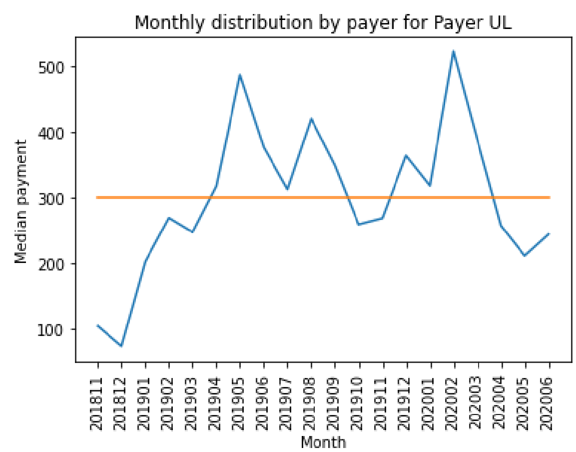

```
Payer UN -->
corr_coef: 0.2846193125394907 -->
p-value: 0.12740626211597295
```




```
Payer B -->
corr_coef: 0.44893165150710146 -->
p-value: 0.01282889657371473
```




```
Payer UL -->
corr_coef: 0.34863808611306984 -->
p-value: 0.1319400938083967
```



For the top specialities by sum we can see the significant (0.05 threshold) trend for:

- Payer O (correlation: -0.55, p_value: 0.001)
- Payer CO (correlation: 0.54, p_value: 0.002)
- Payer S (correlation: -0.44, p_value: 0.01)
- Payer B (correlation: 0.45, p_value: 0.01)

Looking most frequent specialities by payer

```
Payer F: surgery
Payer H: nurse
Payer O: surgery
Payer W: otolaryngology
Payer CA: surgery
Payer CO: surgery
Payer S: otolaryngology
Payer B: otolaryngology
Payer UN: surgery
Payer UL: physical
```

Looking most median payment specialities by payer

```
Payer F: dialysis
Payer H: hospital
Payer O: home
Payer W: long term
Payer CA: hospital
Payer CO: durable
Payer S: long term
Payer B: long term
Payer UN: long term
Payer UL: audiology
```

Looking most frequent specialities goes with surgery

cats	sums
surgery	6047
general	1087
vascular	914
orthopedic	909
thoracic	615
cardiology	583
plastic surgery	504
neurology & neurophysiology	491
otolaryngology	392
colon & rectal	333

Looking most frequent concurrence between specialities

```
ambulance service --> surgery --> 288
ambulance service --> otolaryngology --> 282
anatomic --> pathology --> 205
anesthesia --> nurse --> 264
cardiology --> vascular --> 634
cardiology --> otolaryngology --> 602
cardiology --> interventional --> 602
cardiology --> surgery --> 583
cardiology --> physiology --> 412
```

cardiology --> thoracic --> 303
colon & rectal --> surgery --> 333
critical care --> surgery --> 163
dental --> otolaryngology --> 437
diagnostic --> radiology --> 862
diagnostic --> sleep --> 139
durable --> otolaryngology --> 185
family --> nurse --> 196
gastroenterology --> otolaryngology --> 1077
general --> surgery --> 1087
gynecology --> oncology --> 165
hematology --> oncology --> 758
interventional --> otolaryngology --> 1155
interventional --> radiology --> 511
interventional --> vascular --> 166
mental --> otolaryngology --> 140
neurology & neurophysiology --> urology --> 854
neurology & neurophysiology --> surgery --> 491
neurology & neurophysiology --> radiology --> 158
nuclear --> radiology --> 218
occupational --> therapy --> 118
oncology --> radiology --> 435
oncology --> surgery --> 250
orthopedic --> surgery --> 909
orthopedic --> prosthetic orthotics --> 218
orthopedic --> sports --> 136
otolaryngology --> pain medicine --> 744
otolaryngology --> radiology --> 556
otolaryngology --> surgery --> 392
otolaryngology --> urgent --> 320
otolaryngology --> vascular --> 166
otolaryngology --> sleep --> 147
otolaryngology --> rehabilitation --> 132
physical --> therapy --> 398
physical --> rehabilitation --> 370
plastic surgery --> surgery --> 504
radiology --> vascular --> 141
surgery --> vascular --> 914
surgery --> thoracic --> 615

Summary

Preprocessing

- Analyzed the structure of the dataset
- Filtered strange rows with not positive payments
- Corrected errors by creating new speciality list and features

Analyzed

- Payers by paid_amount distribution
- Service category distributions by payer
- Payment distributions by payer

Analyzed trends

- Top specialities by count
- Top specialities by sum paid
- Top specialities by median payment
- Median payment by payer by month
- Correlations between specialities

Discovered:

- The data collecting process may have errors
- Service speciality list have overlaps
- Claim speciality list is not structured and have mistypes and missed data
- 'Payer UL' has lowest median paid_amount
- 'Payer H' has highest median paid_amount
- Top-5 specialities by count: surgery, otolaryngology, radiology, cardiology, oncology
- Top-5 specialities by sum: hospital, otolaryngology, oncology, surgery, radiology
- Top-5 specialities by median payment: long term, ambulance service, home, infusion, hospital

For the top specialities by count we can see the significant (0.05 threshold) trend for:

- otolaryngology (correlation: -0.39, p_value: 0.03)
- oncology (correlation: 0.39, p_value: 0.03)

For the top specialities by sum we can see the significant (0.05 threshold) trend for:

- hospital (correlation: 0.46, p_value: 0.01)
- otolaryngology (correlation: -0.39, p_value: 0.03)
- oncology (correlation: 0.40, p_value: 0.03)

For the top categories by median payment we can see the significant (0.05 threshold) trend for:

- home (correlation: -0.60, p_value: 0.0004)
- hospital (correlation: 0.46, p_value: 0.01)

Most frequent categories for payers are

- Payer F: surgery
- Payer H: nurse
- Payer O: surgery
- Payer W: otolaryngology
- Payer CA: surgery
- Payer CO: surgery
- Payer S: otolaryngology
- Payer B: otolaryngology
- Payer UN: surgery
- Payer UL: physical

Most expensive categories for payers are

- Payer F: home
- Payer H: hospital
- Payer O: home
- Payer W: long term
- Payer CA: hospital
- Payer CO: durable
- Payer S: long term
- Payer B: long term
- Payer UN: long term
- Payer UL: audiology

Most frequent categories for surgery are:

- general
- vascular
- orthopedic
- thoracic
- cardiology
- plastic surgery
- neurology & neurophysiology
- otolaryngology
- colon & rectal

Also discovered most frequent correlations between categories

Information could be useful for the Ministry of Health, hospitals as well as for insurance companies