How can I inderstand my dataset?

Ukážkový príklad ako sa pýtať tak, aby sme vám efektívne pomohli

datalys

Jakub Ševcech

21st September 2017

PeWe@FIIT
personalized web group

Pomenovanie problémovej oblasti

Stručne povedzte v akej doméne pracujete a čo je vašou úlohou

- Väčšinou netreba prácne vysvetľovať detaily o doméne
- Treba vybrať len to relevantné
- Dôležité je byť stručný, jasný a vybrať tie relevantné časti

Klasifikácia zachránených pri potopení Titaniku

- Mám údaje o pasažieroch lode Titanic a rôzne atribúty o nich
- Mojou úlohou je vytvoriť model, ktorý bude klasifikovať, či človek prežil alebo nie

Zdroj dát: https://www.kaggle.com/c/titanic

Dáta vyzerajú nejak takto

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
8	9	1	3	Johnson, Mrs. Oscar W (Eli								
9	10	1	2	Nasser, Mrs.								

Data Dictionary

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Formulovanie konkrétnej otázky

- Otázka by nemala byť príliš široká (napr: Čo mám teraz robiť? Máte nejaký nápad? ...)
- Pomenujte čo vás trápi a môžeme o tom začať diskutovať

Otázka

Aké nástroje exploratívnej analýzy by som mohol použiť na to, aby som sa oboznámil s takýmto datasetom?

Povedzte čo ste už spravili

- Ukážte čo ste už spravili a čo ste sa dozvedeli
- Na aké problémy ste narazili?
- Pomôže to smerovať diskusiu
- Ukážete, že ste už spravili niečo a prejavili ste nejakú snahu

Základné vlastnosti datasetu

Pozrel som sa na základné charakteristiky numerických premenných.

Sú tam nejaké chýbajúce hodnoty.

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

Základné vlastnosti datasetu

Pozrel som sa na početnosti kategorických atribútov

```
train['Embarked'].value_counts()
```

```
S    644
C    168
Q     77
Name: Embarked, dtype: int64
```

```
train['Sex'].value_counts()
```

```
male      577
female    314
Name: Sex, dtype: int64
```

```
train['Cabin'].value_counts()
```

```
B96 B98          4
C23 C25 C27      4
G6               4
C22 C26          3
E101             3
F2               3
D                3
F33              3
D36              2
B18              2
C92              2
```

Základné vlastnosti datasetu

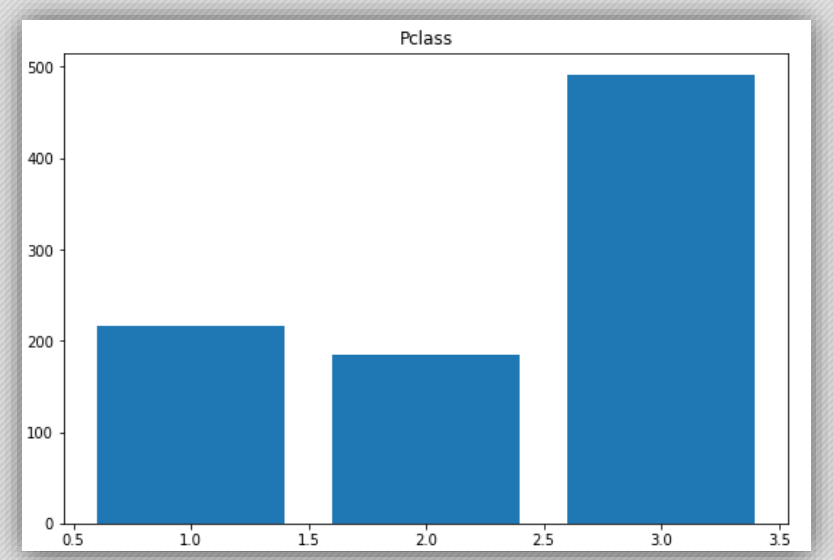
Niektoré v sebe toho ale schovávajú viac a neviem čo s tým

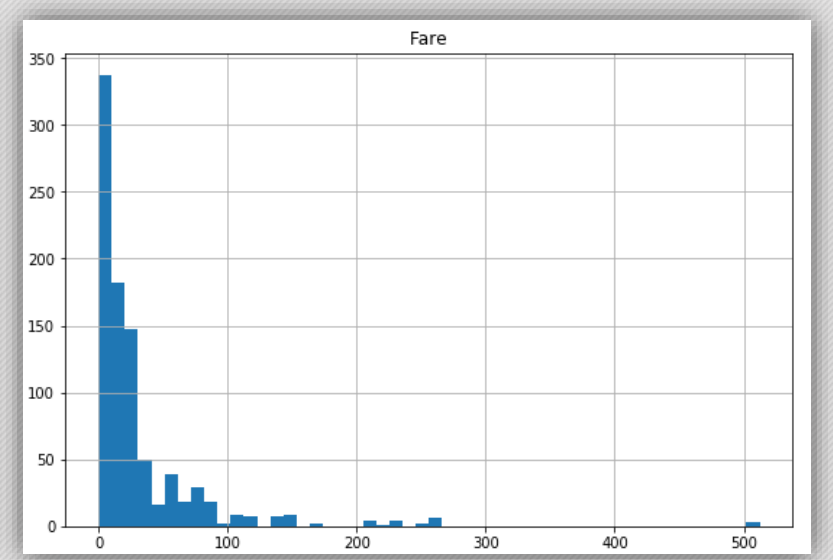
```
train['Name'].value_counts()
```

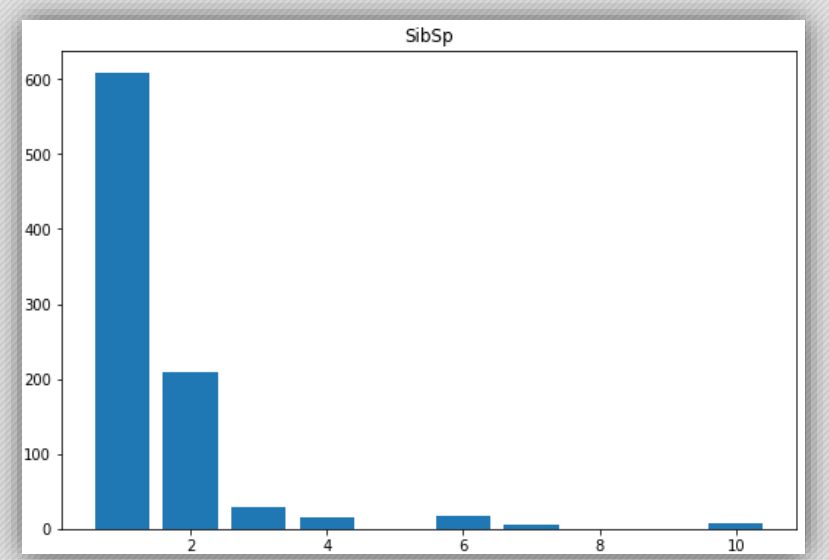
```
Berglund, Mr. Karl Ivar Sven                1
Kimball, Mr. Edwin Nelson Jr                1
Bishop, Mr. Dickinson H                     1
Anderson, Mr. Harry                         1
Dahl, Mr. Karl Edwart                       1
Ball, Mrs. (Ada E Hall)                     1
Beane, Mrs. Edward (Ethel Clarke)           1
Fortune, Miss. Alice Elizabeth              1
Carlsson, Mr. Frans Olof                    1
```

Základné vlastnosti datasetu

Spravil som nejaké histogramy

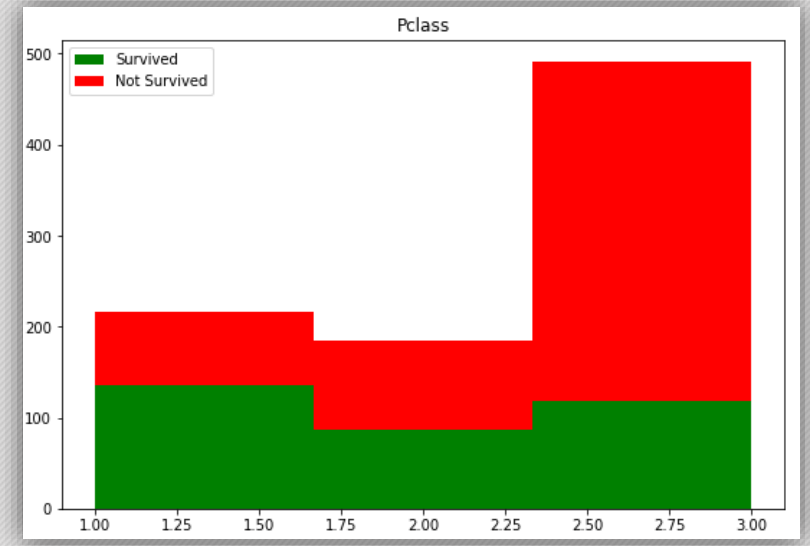


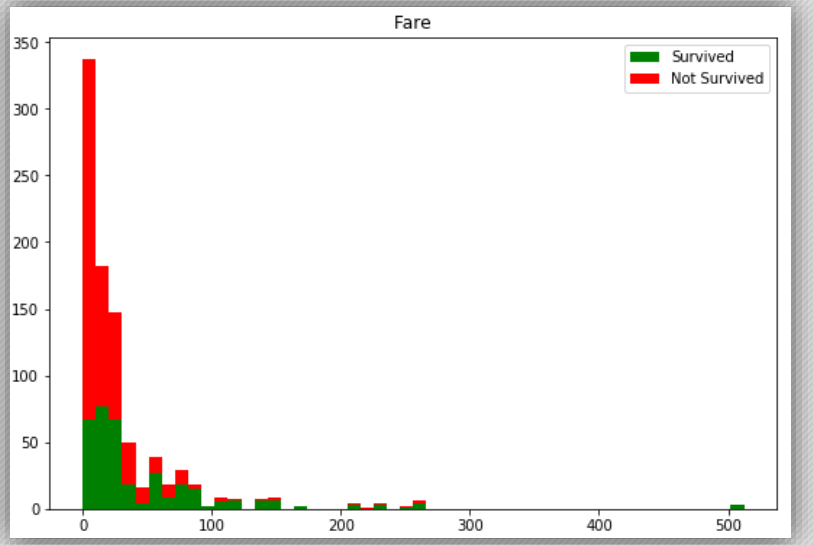


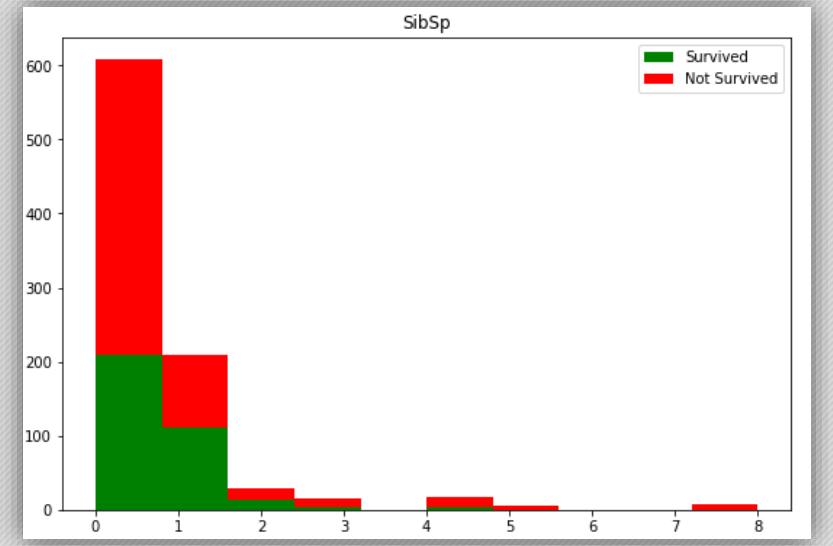


Základné vlastnosti datasetu

Skúsil som ich rozdeliť podľa toho, kto prežil







Aké ďalšie nástroje exploratívnej analýzy by som mohol použiť na mojich dátach?

Ako by som sa mohol inak pozrieť na ten dataset?

Ako by som ho mohol upraviť?

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NaN	C

Pre tých, čo to zaujíma - tutoriály nad týmto datasetom

- https://www.kaggle.com/c/titanic/details/tutorials
- https://www.kaggle.com/startupsci/titanic-data-science-solutions

Pár príkladov ďalších otázok, ktoré by som sa mohol pýtať

- Mám takéto veľa rozmerné údaje, ako by som ich mohol vizualizovať?
- Poznáte nejaké dobré nástroje na podporu reprodukovateľného výskumu v Pythne/R?
- Ako môžem spraviť zhlukovanie na zmiešaných kategorických a numerických atribútoch?
- Aká atribúty by som mohol vytvoriť z dátumu?
- Ako previesť kategorické atribúty na numerické? (alebo naopak)
- Aký klasifikátor sa hodí na tento môj problém?
- Aký typ predspracovania by sa hodil na tento typ údajov?
- Ako môžem anonymizovať môj dataset?
- …